Publisher
Cory Cox

Managing Editor
Stuart Douglas

Content Architect
Biljana Badic

Program Manager
Stuart Douglas

Technical Editor
David Clark

Technical Illustrators
MPS Limited

Technical and Strategic Reviewers
Valerio Frascolla
Biljana Badic
Erfan Majed
Jan-Erik Mueller
Shilpa Talwar
Trevor Wieman
Luis Castedo Ribas
Kenneth Stewart
Dauna Schaus
John Aengus
Markus Brunnbauer
Steve Duffy
Marcos Katz
Pablo Puente

Intel Technology Journal

ISBN 978-1-934053-64-5, ISSN 1535-864X

Intel Technology Journal
Volume 18, Issue 3

Publisher: Cory Cox
Managing Editor: Stuart Douglas

Library of Congress Cataloging in Publication Data:

Printed in China
10 9 8 7 6 5 4 3 2 1

First printing: May 2014

4G WIRELESS COMMUNICATIONS: REAL WORLD ASPECTS AND TOOLS

Articles

Foreword

Before entering into the core of this technology journal, I would like to invite you on a short trip to the other side of the mirror. This journey is a reflection on where technology is today and where it is going tomorrow. I won't bother you further about well-known concepts like *faster, more mobility, anywhere, anytime*—all of these generally accepted technical marketing concepts shaping the future directions of wireless communications.

Today, I have spent my time walking around at Monterey Bay Aquarium, a world-renowned aquarium south of the Silicon Valley, to get inspiration for writing this foreword. You may be asking yourself why? Because it allows me to step back and reflect at the wonder of nature. Besides colorful fishes, sea otters, jellies, and other creatures, you see humans. I personally like to analyze human behavior, understand what generates curiosity, what creates emotion, what shapes interests. All of this with the intention to understand what excites the person (call him or her the end-user), meaning what excites you and me.

Now, "what do these humans do when visiting an aquarium?" the fish might ask on the other side of the glass. They look at me, take pictures or short movies of me with their smartphones of all kind of sizes and shapes, they upload them on Instagram, update their Facebook. In this architecturally interesting aquarium structure, mixing concrete and steel, and surrounded by over a million of gallons of water, you have an area where the cellular network coverage is not perfect, despite being augmented by free Wi-Fi*. And of course, it is just there, that one spot in front of this gorgeous scene of nature, at that instant that I tried skyping home to share this moment where... can't connect… no ring tone. I'm lost in front of the fish. Damn is he relaxed, staring at me!

(Source: ©Monterey Bay Aquarium/Randy Wilder)

The engineer inside me wakes up. Is it a software crash, a network coverage issue, an antenna issue, a network capacity issue? It does not work. Why?

Setting up wireless communication systems are real challenges for the complete industry ecosystem. If one fails, the entire ecosystem fails. Somebody will get the blame, but it may not be the one causing the failure. The end users do not care and they are right.

With 4G at our doorstep, we will further increase the high throughput and mobility anytime and anywhere. But that is not the end of the story: device diversity will further increase, leading to a larger variety of design considerations at the baseband, RF, and antenna level. More variables, more uncertainties, all of them lead to a new level of complexity that needs to be harnessed for success. New considerations need to be investigated to enhance performance of the LTE network, to ensure Wi-Fi backup. More and better tests, trials beyond just what the standard requires, are needed to ensure it works anywhere anytime. You can't transform every handset end user into an engineer to debug it.

This issue of the *Intel Technical Journal* presents eleven articles treating the real-world aspects and tools for developing 4G communication systems. What are the solutions to achieve better service, better development, better verification, better optimization, and better testing of this latest mobile standard? This is the question. You as a reader, engineer, technologist, will enjoy the beauty revealed in the articles. But keep in mind, all these tricky, complex techniques need to all work seamlessly to be valuable.

To take a lesson from the natural world, perhaps we can think of this another way: how do we as engineers transpose Darwin's theory to wireless communications? How do we develop an ecosystem that allows the next generation of wireless communications to evolve and flourish?

I'm now on my way home, sitting in a Wi-Fi-enabled airplane. I'm uploading the aquarium pictures and movies on the cloud. What a beautiful day… it works!

Enjoy the *Intel Technology Journal*.

Michael Dieudonné
European Research Projects Coordinator
Agilent Measurement Research Labs
Agilent Technologies

ANTENNA AND RF SUBSYSTEM INTEGRATION IN CELLULAR COMMUNICATIONS

Contributors

Pablo Herrero
Wireless Platform Research and Development, Intel Corporation

Pevand Bahramzy
Wireless Platform Research and Development, Intel Corporation

Simon Svendsen
Wireless Platform Research and Development, Intel Corporation

Alfonso Muñoz-Acevedo
Wireless Platform Research and Development, Intel Corporation

Boyan Yanakiev
Wireless Platform Research and Development, Intel Corporation

Tommaso Balercia
Wireless Platform Research and Development, Intel Corporation

Christian Rom
Wireless Platform Research and Development, Intel Corporation

We discuss in this article a number of techniques that can be used to improve the RF performance on a mobile device. All those techniques rely on tight antenna and modem subsystem codesign. In a short introduction, the article outlines the need of these techniques, based on the advent of new wireless standards and demanding power and size requirements. The first technique is based on integrating the antenna as part of the RF filtering chain to relax the requirements of current duplex filters up to 30dB off-band. We also outline a discussion on the different approaches for adaptive antenna matching, depending on the system's requirements. A particular antenna concept that implements two different wireless systems (LTE and Wi-Fi) is outlined in a subsequent section. Such a structure can reduce the volume required for antennas on a mobile device. The article concludes with an example of how the antenna design can directly impact the throughput of the whole system, showing the impact of the correlation coefficient on the signal to noise ratio.

Introduction

In the past, mobile communications and by extension cellular modem design has typically focused on delivering performance on a 50 Ohm connection reference, keeping only the RF modem into one "codesigned" box. As a further path for increasing RF performance, this article discusses possible solutions to also push the antenna system into the codesign box of the whole RF subsystem and device itself, including industrial design (ID) and air interface. This way, a whole end-to-end optimization from air to bit can be achieved, delivering better overall performance in an appealing form factor.

The frequency spectrum for mobile communication has increased dramatically due to the deployment of Long Term Evolution (LTE).[1] It is expected that mobile devices are able to operate at these frequencies without taking too much printed circuit board (PCB) volume or consuming a lot of power, impacting the size and the battery life of mobile communications products. In addition, different parts of the world have different frequency bands, making it quite challenging to support all the bands in a single stock keeping unit (SKU). It may require either more components (or larger ones) in the mobile device to be able to support multiple geographic configurations, thereby impacting the cost and reducing the benefit margin.

Integration has proven to be a technology success driver since electronics' early days. This trend materializes in the personal modems' segment as an aggregation of radio access technologies, hence their respective spectral allocation requirements.

This article explores possible device-antenna-RF codesign as an alternative way to cope with the problems described above, following a top-down approach. First, an implementation in which the filtering capabilities of the antenna system can positively impact the RF modem itself is investigated. A brief study of antenna control techniques to enable this filtering behavior under control will be exposed. The antenna concept and how it fits into the device design itself will follow to end up on the air interface with the antenna and its impact in throughput.

Different stages of codesign will focus on different improvements in performance, which will be quantified by simulation models and measurements.

The Antenna as Part of the RF Chain Filtering

The mobile platform has several radio frequency (RF) signal paths due to the many supported frequency bands. Each of these RF signal paths contains a duplex filter, a power amplifier, a low noise amplifier, and so on, resulting in various parallel paths. All this leads to a complicated RF front end (FE). Figure 1a illustrates a simplified version of such a multipath FE.

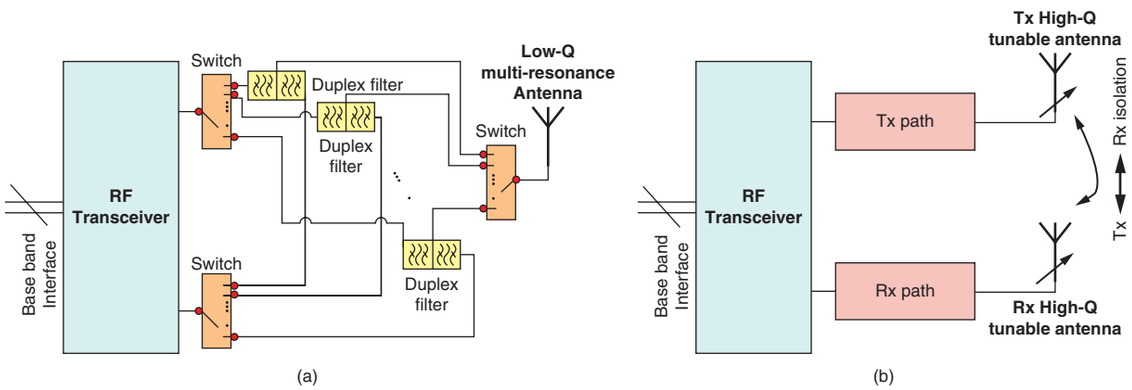


Figure 1: Simplified conventional FE (a), simplified proposed FE (b) (Source: Intel Corporation, 2014)

On the antenna side, this means many antenna subsystems or two very broadband antennas (for supporting MIMO and carrier aggregation). While broadband antennas are large and occupy more space in order to be efficient, multiband antennas with reasonable performance become a very cumbersome task to design in a small mobile form factor device due to the fundamental limitation of antennas.[2]

This can be addressed through codesign of the RF FE and the antenna system, resulting in a miniaturized FE and antenna, yet with coverage of the increased number of bands. Such an approach has separate transmitter and receiver paths throughout the FE, including the antennas.

Figure 1b presents a simplified block diagram of the architecture that integrates the antenna and the RF system.

While the traditional FE uses duplex filters and switches to separate Tx and Rx, the proposed FE relies on the filtering characteristics of the antennas together with tunable filters to separate the Rx from the Tx. In such a design, for each mode or application in the mobile terminal, the bandwidth is reduced to just the needed channel bandwidth, which for LTE is between 1.4 MHz and 20 MHz. Therefore, the Tx and Rx antennas can be designed to be quite narrowband. Tuning can be exploited to cover the required frequency range. This approach brings major advantages; some of them are listed below:

- It helps in reducing the PCB area for the RF engine. The Tx and Rx antennas exhibit high isolation because of the narrowband characteristic and frequency offset, providing filtering that can be used to replace some of the otherwise required filters from the FE. Also the space occupied by the antennas is reduced because the same elements can be used to cover all the requisite bands. This reduction in size will enable new and innovative industrial designs of mobile devices.
- It provides power and cost savings, because some of the otherwise necessary filters and switches can be removed from the FE, resulting in lower insertion loss, which lowers the power consumption in the total RF engine solution.
- It provides dynamic optimized overall FE efficiency by controlling the tunable antennas to mitigate the effects of the external environment, such as the handheld terminal user.

A practical example of the narrowband antennas and their advantages are presented in the following.

"The antennas have no PCB cutback and have a volume of only 0.3 cc each. With such small volumes, these tunable narrowband antennas are capable of covering a frequency range of 700 MHz through 1 GHz."

A PCB with mounted Tx and Rx antennas is depicted in Figure 2. The antennas have no PCB cutback and have a volume of only 0.3 cc each. With such small volumes, these tunable narrowband antennas are capable of covering a frequency range of 700 MHz through 1 GHz.

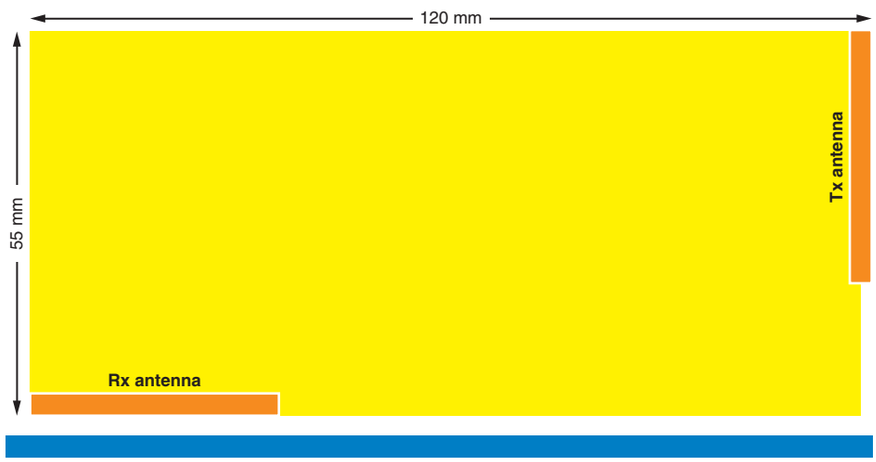


Figure 2: Narrowband Tx and Rx antennas on a PCB (Source: Intel Corporation, 2014)

The antenna tunability is illustrated in Figure 3. Coverage of the low band can be obtained by applying a capacitance range of 0.7 through 4 pF using 29 tuning states.

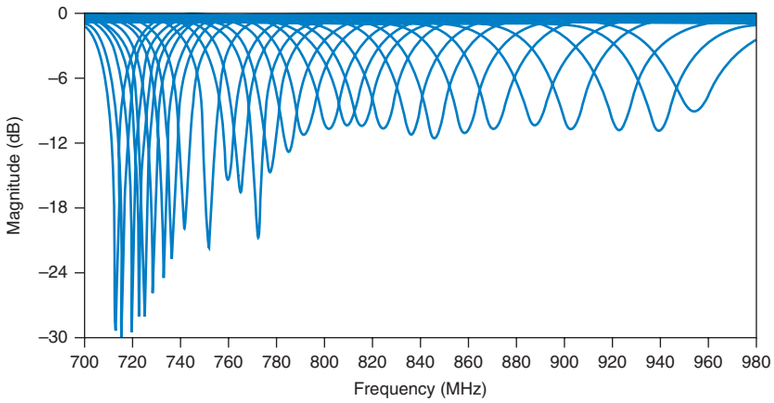


Figure 3: Narrowband antenna tuned to cover the low band
(Source: Intel Corporation, 2014)

Figure 4 shows that the Tx tunable antenna covers the frequency range 700–915 MHz with a bandwidth of 24 MHz at the highest operation frequency and 10 MHz at the lowest operation frequency. The Rx antenna can cover 729–975 MHz with a bandwidth of 21 MHz at the highest operation frequency and 9 MHz at the lowest operation frequency.

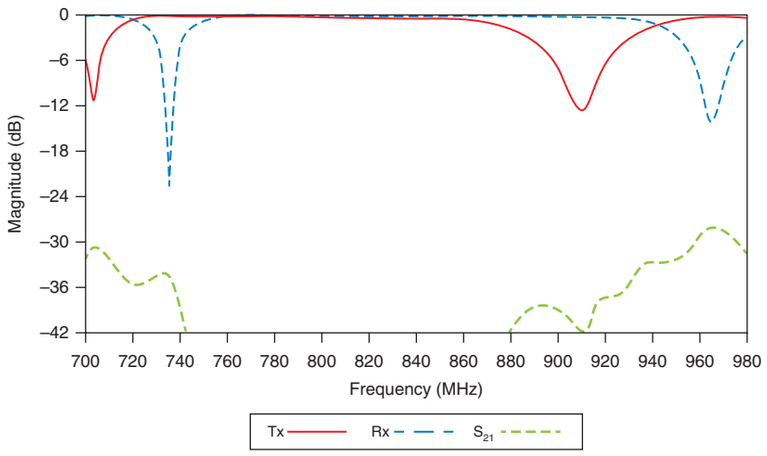


Figure 4: Narrowband Tx and Rx antennas reflection coefficients S_{11} (red), S_{22} (blue), and coupling coefficient S_{21} (green)
(Source: Intel Corporation, 2014)

Since the proposed FE relies on the filtering characteristic of the antennas, it is relevant to show the near field (antenna isolation/coupling) and far field filtering characteristic. The near field filtering or isolation between the antennas is better than -28 dB across the entire frequency range. Such a high isolation at

"Since the proposed FE relies on the filtering characteristic of the antennas, it is relevant to show the near field (antenna isolation/coupling) and far field filtering characteristic."

these low frequencies is achieved due to the narrowband nature of the antennas and frequency offset. Moreover, the antennas are not exciting the same PCB mode, which also contributes to the incredibly high isolation.

The far field filtering characteristic, inherently in the Tx antenna, is measured and shown in Figure 5. The antenna is tuned to 740 MHz and has an efficiency of -3.3 dB at that frequency. Outside the frequency band of interest the antenna acts as a filter. Above 1080 MHz and below 600 MHz, the antenna ensures 30 dB filtering. This is a very interesting result because filter requirements in the FE can be relaxed if the attenuation of unwanted RF blockers in the antenna is accounted for.

"Above 1080 MHz and below 600 MHz, the antenna ensures 30 dB filtering. This is a very interesting result because filter requirements in the FE can be relaxed if the attenuation of unwanted RF blockers in the antenna is accounted for."

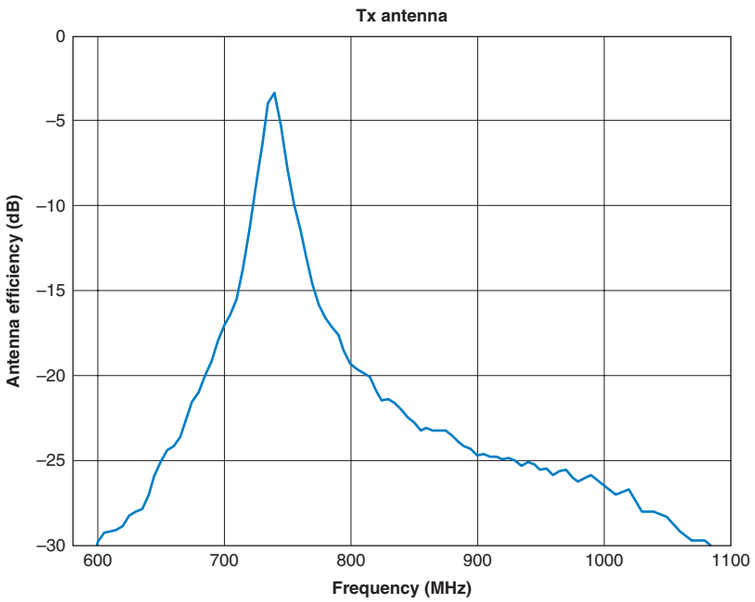


Figure 5: Narrowband Tx antenna efficiency measurement in an anechoic chamber
(Source: Intel Corporation, 2014)

The efficiencies are listed in Table 1, where it is noted that Tx antenna efficiency varies between −1.0 dB and −3.3 dB and Rx antenna efficiency varies between −1.2 dB and −3.6 dB.

	f_r [MHz]	eff. [dB]
Tx antenna	910	−1.0
	705	−3.3
Rx antenna	965	−1.2
	735	−3.6

Table 1: Total efficiencies at lowest and highest tuning frequencies of the Tx and Rx antennas.
(Source: Intel Corporation, 2014)

Antenna Control and Adaptive Matching

Classical antenna literature is however conclusive on the electromagnetic limits regarding radiating structures. It is clear that a proper concept for controlling the antenna bandwidth and resonant frequency of the element is required to achieve the "filtering effect" on the antenna. Out of the diverse approaches to control the antenna bandwidth, the reconfigurable antenna tuning solution is discussed in this section in more detail. The different concepts for implementation are presented also in conjunction with adaptive matching under user interaction scenarios.

"…a proper concept for controlling the antenna bandwidth and resonant frequency of the element is required to achieve the "filtering effect" on the antenna."

Hardware Integration: The World Indoors

Devices devoted to implement reconfigurability in antennas are diverse and so are the radiation mechanism reconfiguration techniques. A set of these is analyzed as in the Figure 6.

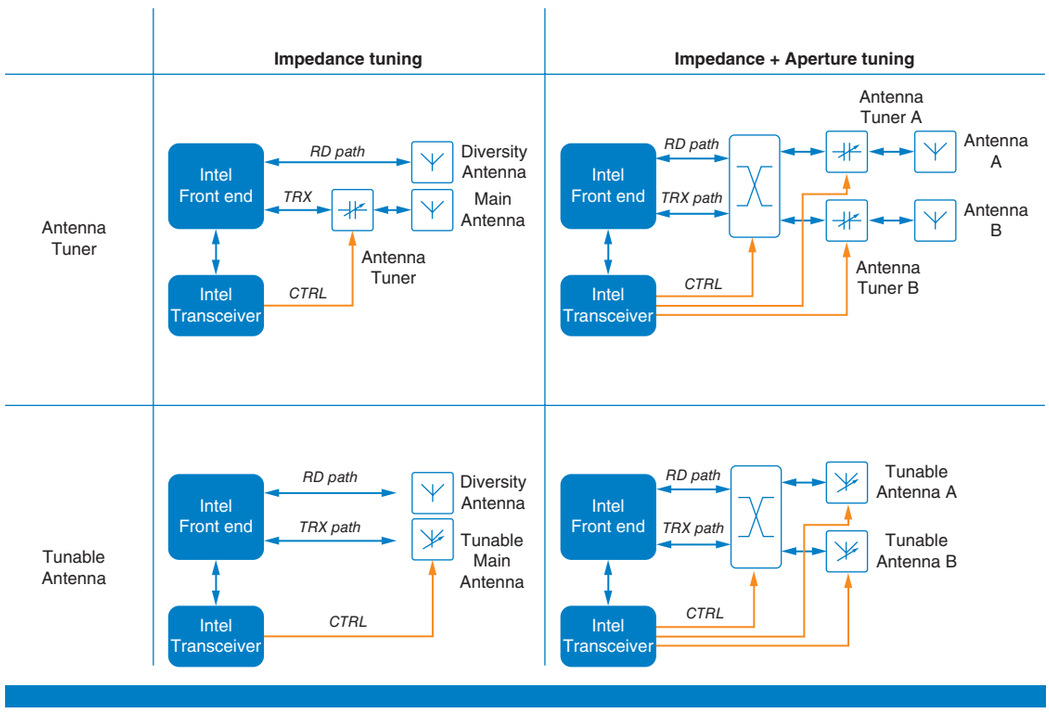


Figure 6: Reconfigurable antenna lineups (Source: Intel Corporation, 2014)

Rows stand for the integration trend. Antenna tuners act as reconfigurable impedance transformers, being two-port devices committed to interface an arbitrary impedance drifting antenna port with the front end's nominal 50 ohm port, according to a certain criterion. Gaining in integration is here linked to a loss in parameterization capabilities for the tuning device. Impedance transformation between two ports can be tabulated and then processed so a certain tuning criterion is met. Limitation arises for the one-port

device since that tabulation requires knowledge of further information, such as a user case interaction scheme, thus introducing an environment segmentation that stands for a nailed-down vision of the electromagnetic environment from where the antenna radiates. This inherent ill-conditioned nature of one-port devices tuning has as a tangible consequence a reformulation in the tuning criterion, leading to suboptimal tuning with respect to their two-port counterparts.

Radio Wave Radiation: The World Outdoors

The tuning criteria design follows an RF performance-increase commitment, linked to the circuital perspective of the radiating devices. Radiation-mechanism reconfiguration lays out of the action range of antenna tuners and its higher integration versions as well as the criteria they are operated with. Aperture tuning broadens this scenario by introducing further reconfiguration variables, which potentially diversify the radiating element's location, polarization, and feeding.

Most immediate approaches identify shadow conditions for certain antennas in the array thus redirecting RF power. This binary approach to signal routing may be refined with an upper layer—implemented either by means of antenna tuners or tunable antennas—whose reconfiguration criteria cooperates with the routing capabilities introduced by the switch.

Aperture tuning has a formal impact that materializes in a larger sphere of criteria. Antennas are seen beyond its circuit perspective, and coordination between binary routing and impedance matching criteria is desired so suboptimal reconfiguration can be avoided.

All in all, reconfigurable antenna designs have become a must in the mobile communication industry; however, performance advantages can only be expected out of this technology when tuning criteria are consistent for the level of hardware integration one is willing to set up in a mobile platform.

Antenna Sharing in Cellular Communications

"…due to the volume required for each antenna to achieve good performance. The performance of antennas in mobile phones is directly related to the volume allocated and the physical placement in the phone."

Once a proper (hopefully tunable) antenna concept has been selected, the designer faces the next problem when trying to build a mobile communications device: the number of antennas needed in modern smartphones is increasing, in order to support different bands, MIMO, carrier aggregation, WLAN, NFC and GPS, which is a major challenge due to the volume required for each antenna to achieve good performance. The performance of antennas in mobile phones is directly related to the volume allocated and the physical placement in the phone. Increasing the allocated volume for the antenna will in theory result in better antenna performance in terms of S_{11} and radiated efficiency. Also, the best performance of the antennas is obtained when they are placed at the circumference of the

phone. The width of the display and battery is often nearly as wide as the smartphone itself, whereby the available volume for antennas at the circumference near these components is very limited and in many cases not usable for antennas. Other components like the USB connector, audio jack, and different user control buttons are normally also placed at the circumference of the phone, thereby reducing the available volume for the antenna even more. This means that the space for good antenna performance on a modern smartphone is very limited. It would be a big advantage if some of the wireless systems and/or cellular bands could share the same antenna element and operate concurrently without a significant degradation of the performance. This would reduce the total number of antenna elements needed in the phone.

"…the space for good antenna performance on a modern smartphone is very limited."

This section shows a new and unique way of simultaneously coupling multiple wireless systems, such as one cellular band and two Wi-Fi* bands on to the same antenna element, without significantly degrading the performance of either systems.

Sharing of antenna elements between for example GPS and Wi-Fi has been done before by making a standard single-feed dual-resonance antenna and then feeding the GPS and Wi-Fi signal through a duplexer/diplexer to the antenna. This solution could in theory also be used for a cellular and Wi-Fi configuration if the added insertion loss of 1 dB to 2 dB of the duplex filter could be accepted.

A switched solution is also an option, where each system/band is switched on and off to the antenna element, so that only one band/system is coupled to the antenna element at a time. This solution does not support concurrent operation and also introduces loss, complexity, and an increase of price due to the additional switches.

Sharing of Cellular Low Band and WLAN

This example shows an antenna concept that couples the cellular low-band systems between 700 MHz and 960 MHz as well as the 2.4 GHz system onto the same antenna element, without increasing the volume of the antenna element itself needed to cover the low-band frequencies alone (700 MHz to 960 MHz).

The concept is based on indirect feeding, where the antenna is fed through a coupler and the element itself is either directly connected to ground, if it is resonating at the desired frequency, or forced to resonate at a desired frequency with an inductor, capacitor, or even a combination of these.

"The concept is based on indirect feeding, where the antenna is fed through a coupler and the element itself is either directly connected to ground or forced to resonate at a desired frequency…"

The original concept uses one coupler per single resonance element, whereby two elements and two couplers are needed for dual resonance operation. The antenna elements are often simple monopole type elements

as shown in Figure 7a. These types of elements can be relatively easily made into dual resonance elements by adding a slot in the element as shown in Figure 7b.

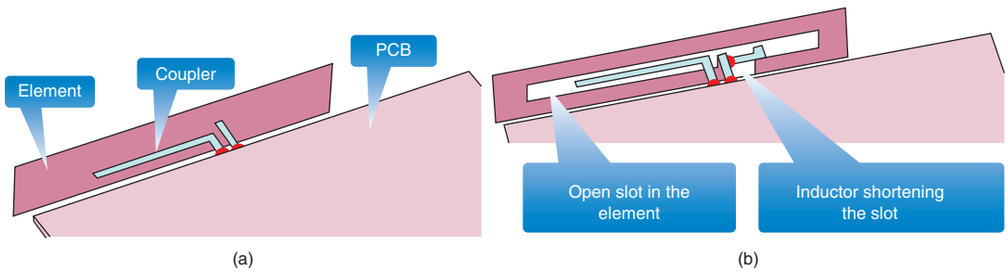


Figure 7: Indirect feeding technique; (a) original single resonance concept, (b) new dual resonance concept for antenna sharing
(Source: Intel Corporation, 2014)

The second element resonance can be "picked up" by adding a second coupler. The example concept shown in Figure 7b is designed to cover the cellular low-band frequency range from 704 MHz to 960 MHz and the WLAN 2.4 GHz frequency range from 2400 MHz to 2484 MHz. The impedances to the feedings are shown in Figure 8a and 8b. The simulation results show a good impedance matching over the bandwidth of interest.

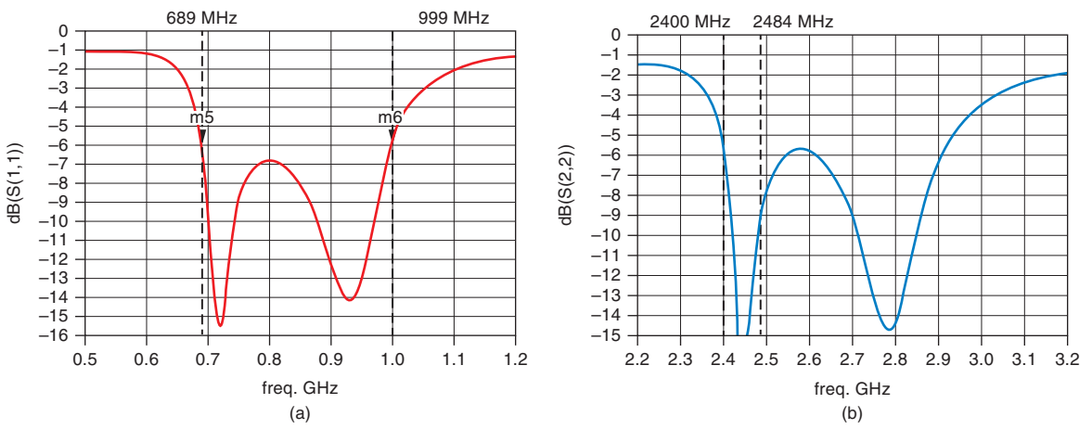


Figure 8: Impedances and isolation of the MCE concept; (a) S_{11} cellular low band, (b) S_{11} WLAN 2.4 GHz
(Source: Intel Corporation, 2014)

The use of these concepts can reduce the number of needed antenna elements in wireless devices, whereby more compact stack-ups can be made and/or more metal can be accepted on the phone, giving more degrees of freedom to the industrial designers.

MIMO and the New Challenge to Design for Throughput

The scope of antenna-device and RF codesign can be pushed to even higher levels in order to maximize overall performance of a mobile device. In previous sections, this performance improvement came by increasing filtering on the antenna (thus reducing the PCB area and/or current consumption) or by reducing the volume required for the antenna itself. In this section, throughput is discussed as another performance parameter that can be increased by a good antenna codesign.

The introduction of the spatial degree of freedom in modern communication systems introduces new opportunities and new challenges in the design of modern antenna systems. It offers in fact an expansion in the number of parameters to monitor and optimize for, the two most important ones being antenna correlation and power ratio (BPR). The importance of such quantities derives directly from the fact that these are the very parameters that limit the linear increase in capacity, and by extension throughput, to be expected from MIMO systems,[3] just as it has been shown previously for various combining techniques.[4]

In order to illustrate this aspect, here we focus on a simple 2x2 MIMO system and adopt the common approach that describes the antennas in terms of total efficiency per antenna, BPR, and correlation coefficient between the branches. As can be expected, these parameters are highly dependent on the propagation channel and its properties. Indeed, it is worth observing that the very definition of total efficiency cannot be given without assuming a specific channel—the isotropic one. Unless otherwise noted, we also use the assumption of isotropic incoming power. It must be stressed, however, that this assumption is by no means realistic, as pointed out in literature.[5][6][7] Here it is used just because of its mathematical simplicity.

The influence of the correlation coefficient and BPR at different SNR levels can be seen in Figure 9. The graphs in this figure are generated by computing the open loop channel capacity as the receiver correlation coefficient in the Kronecker model[8] sweeps between 0 and 1 (x-axis) and the BPR sweeps between 0 and −15 dB (y-axis). As for the power carried by the entire channel, it is worth mentioning that the channel matrices were always normalized to have unitary power. The normalization of the capacity was in turn done with respect to the maximum achievable capacity for the given SNR point, which is found when correlation and BPR are respectively equal to zero and 0 dB.

Figure 9 shows that the influence of BRP and correlation varies with the level of the SNR. At low SNRs, the BPR is more significant in determining the overall capacity. As the SNR increases, however, the correlation starts playing a noticeable role. In all cases, nonetheless, the capacity depends on both parameters nonlinearly. Before continuing, it is also worth observing that it is thanks to such analyses and considerations that the current industry "rule of thumb" numbers, between 0.5 and 0.7 for correlation and up to about 3 dB for the BPR, are considered a reasonable balance between what is realistically possible to implement and MIMO performance penalties.

"…the influence of BRP and correlation varies with the level of the SNR. At low SNRs, the BPR is more significant in determining the overall capacity. As the SNR increases, however, the correlation starts playing a noticeable role."

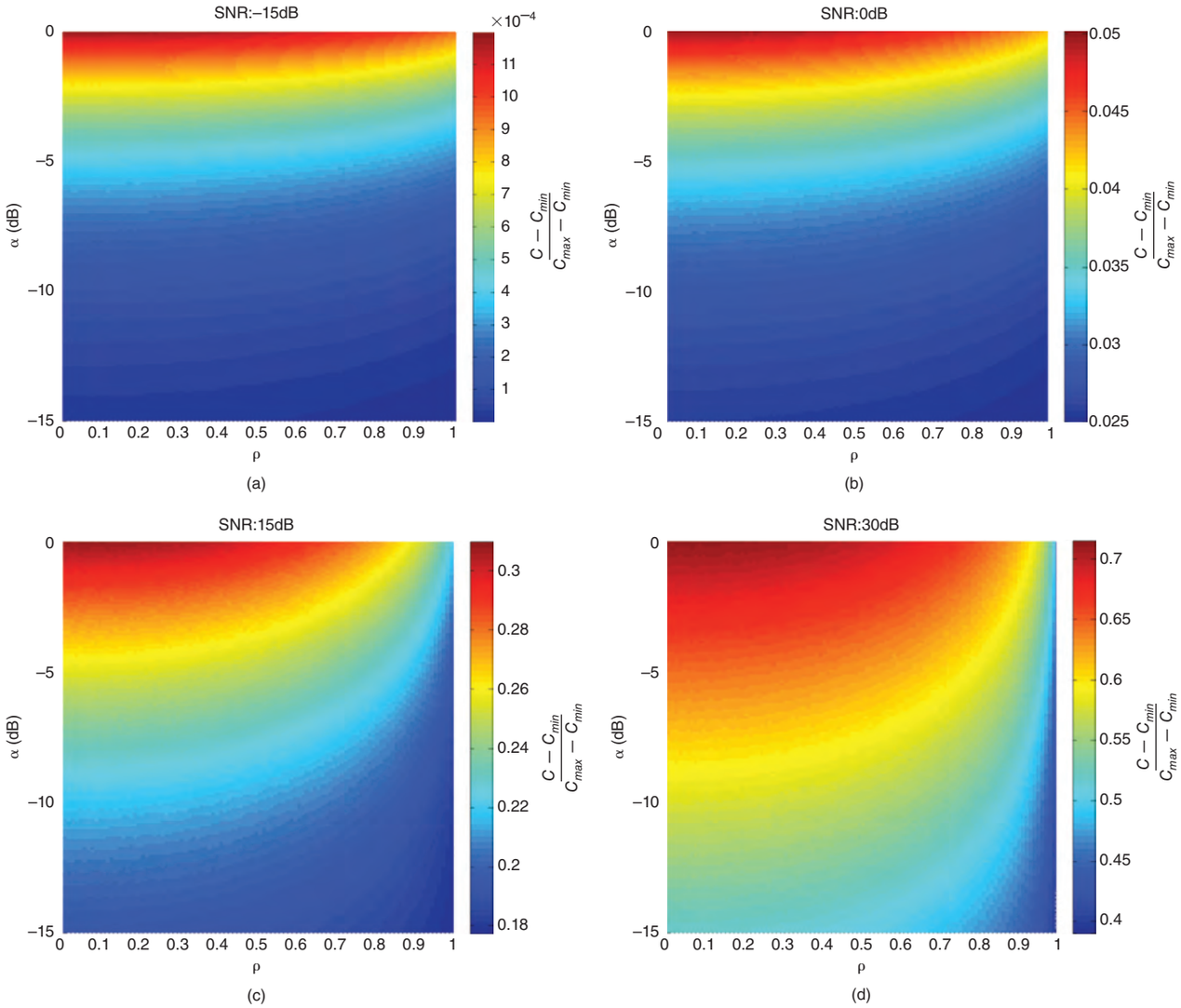


Figure 9: The influence of the correlation coefficient and BPR at different SNR levels (Source: Intel Corporation, 2014)

In looking at the aforementioned figures, another aspect to keep in mind is that correlation and BPR are, strictly speaking, not related—any combination between them is possible for the various antenna systems. In reality, however, it is almost impossible to decouple one from the other. Changes in efficiency lead in fact to changes in BPR and correlation. This is an aspect that plays a major role in the design of tunable antennas. For tunable antennas and RF front ends it is in fact possible to control these dependencies in order to optimize the achievable throughput for any given propagation condition.

Traditionally antennas are designed for the highest efficiency possible. Indeed, a number of publications demonstrate that channel capacity depends mostly on the total receiver power. This for instance can be seen in the work of Yanakiev et al.[9], which shows that designs optimized for low correlation can produce only marginally different performance under realistic channel and operation conditions. Whether or not such a traditional guideline should always be followed can, however, be challenged.

"Traditionally antennas are designed for the highest efficiency possible."

"Whether or not such a traditional guideline should always be followed can, however, be challenged."

The aspect can be understood by looking into the statistical sample described by Nielsen et al.[10], which represents real world measurement—3000 measurements in free space, with multiple users gripping the phone in various predefined positions as well as 10 different antenna systems.[10] Figure 10 shows a set of metrics derived using such a sample.

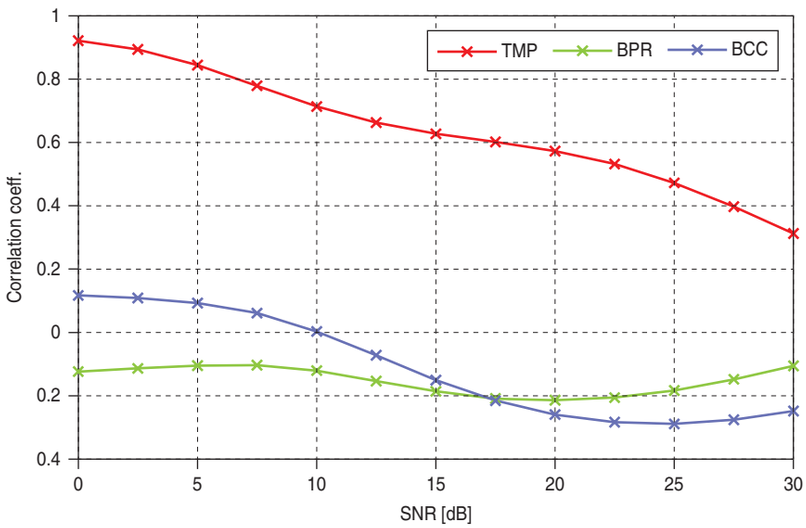


Figure 10: Correlation of capacity spread and TMP/BPR/BCC, vs. SNR (Source: Aalborg University, 2014)

In Figure 10, TMP stands for total mean power and BCC is the branch correlation coefficient. On the x-axis, various SNR operating points are shown, while the y-axis accounts for the correlation that the three plotted parameters have with the capacity spread. This last metric is defined as the difference between the 90th and 10th percentiles.

In looking at Figure 10, it is worth observing that the total power collected is crucial in determining the channel capacity. However, at certain SNR levels, the direct connection between power and capacity weakens and there is room for optimizing correlation and BPR with the aim of maximizing throughput. As mentioned above, this can be achieved with modern tunable antennas and smart RF front ends.

"…the total power collected is crucial in determining the channel capacity."

Conclusions

A number of techniques that can improve the RF performance and mobile device design have been discussed in this article. The approaches are based on taking into consideration the antenna design when building a mobile device: that is, when we "pull" the antenna design into the RF modem and the industrial design of the device itself. The key performance indicators (KPIs) that can be improved are current consumption (which will reflect on the battery life of the device) PCB area (impacting the size of the device) or throughput (which can translate into higher download rates and better user experience).

This allows for the development of better modems with superior performance, lower current consumption, and eventually, platforms that would require smaller area and lower cost, which can be more attractive to end users.

References

[1] Evolved Universal Terrestrial Radio Access (EUTRA); User Equipment (UE) radio transmission and reception, 3GPP Std. TS 36.101, http://www.3gpp.org/ftp/Specs/html-info/36101.htm.

[2] Chu, L. J., "Physical limitations of omni-directional antennas," *Journal of Applied Physics*, vol. 19, no. 12, pp. 1163–1175, 1948.

[3] Telatar, E., "Capacity of Multi-antenna Gaussian Channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, 1999.

[4] Pedersen, G. F., and J. B. Andersen, "Handset antennas for mobile communications: Integration diversity, and performance," *Rev. Radio Sci*. 1996–1999, pp. 119–137, 1999.

[5] Ertel, R. B., P. Cardieri, K. W. Sowerby, T. S. Rappaport, and J. H. Reed, "Overview of spatial channel models for antenna array communication systems," *Pers. Commun.* IEEE, vol. 5, no. 1, pp. 10–22, 1998.

[6] Martin, U., J. Fuhl, I. Gaspard, M. Haardt, A. Kuchar, C. Math, A. F. Molisch, and R. Thomä, "Model Scenarios for Direction-Selective Adaptive Antennas in Cellular Mobile Communication Systems – Scanning the Literature," *Wirel. Pers. Commun.*, vol. 11, no. 1, pp. 109–129, 1999.

[7] Bonek, E., M. Herdin, W. Weichselberger, and H. Ozcelik, "MIMO - study propagation first!," in *Signal Processing and Information Technology*, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium on, 2003, pp. 150–153.

[8] Costa, N. and S. Haykin, *Multiple-Input Multiple-Output Channel Models: Theory and Practice* (New York: Wiley, 2010).

[9] Yanakiev, B., J. Ødum Nielsen, M. Christensen, and G. F. Pedersen, "On Small Terminal Antenna Correlation and Impact on MIMO Channel Capacity," *Antennas Propag*. IEEE Trans., vol. 60, no. 2, pp. 689–699, Feb. 2012.

[10] Nielsen, J. Ø., B. Yanakiev, S. C. D. Barrio, and G. F. Pedersen, "Channel Statistics for MIMO Handsets in Data Mode," presented at the Accepted for publication at EuCAP 2014, The 8th European Conference on Antennas and Propagation, Hague, 2014.

Author Biographies

Pablo Herrero is head of the Antenna Systems department in the the Wireless Platform Research and Development division of Intel Corporation. His team focuses on smart reconfigurable architectures and antenna integration in mobile devices. He earned his Dipl.-Ing. in Electrical Engineering from the University of Zaragoza and his PhD degree in ultrafast wireless systems from the TU Braunschweig (Germany), having been awarded the IEEE Antennas and Propagation Society research award. Email: Pablo.herrero@intel.com.

Pevand Bahramzy received the BSc EE degree and the MSc EE degree in electrical engineering from the Danish Technical University (DTU), Copenhagen, Denmark, in 2006 and 2008, respectively. In 2008 he joined Molex Antenna Business Unit, where he worked with the design of integrated antennas for mobile devices. In 2013 he joined Intel Mobile Communications where he is working as an antenna engineer and pursuing a PhD in cooperation with Aalborg University, Denmark. His current research is focused on reconfigurable high-Q antennas for portable devices. Email: Pevand.bahramzy@intel.com.

Simon Svendsen received his MSc EE in Telecommunication in 1995 from Aalborg University. He joined Bang and Olufsen in 1996, where he worked with RF and antenna design for DECT phones. In 2000, he joined Maxon as an antenna designer for cellular mobile phones. He has worked as an antenna designer and mechanical engineering since then, for companies like Siemens mobile Phones, Motorola, and Molex. His current position is as a senior antenna designer at Intel Mobile Communications. Email: Simon.svendsen@intel.com.

Alfonso Muñoz-Acevedo holds a Dr. Engineer in Telecommunications degree, earned in 2012 from Universidad Politécnica de Madrid through his research on millimeter wavelength electromagnetism. Alfonso's scientific interests cover transversal topics spanning electromagnetic analysis algorithmia to applied physics and antennas. Email: alfonso.munoz-acevedo@intel.com.

Boyan Yanakiev received a BSc degree in physics from Sofia University, Bulgaria in 2006 and a MSc EE in Wireless Communication and a PhD from Aalborg University, Denmark in 2008 and 2011 respectively. His current position is as an antenna engineer at Intel Mobile Communications and an

industrial postdoctoral fellow at Aalborg University. His primary interests are in the area of small integrated mobile antennas, optical antenna measurement techniques, and radio channel measurements. He has been involved in the design and development of multiple RF-to-optical convertors for onboard handset measurements. Email: Boyan.Yanakiev@intel.com.

Tommaso Balercia received his MSc EE in Digital Signal Processing in 2007 from the Polytechnic University of Marche (Italy), and his PhD in 2013 from the Technical University of Braunschweig (Germany). In 2007, he joined the concept engineering group of Comneon GmbH as a graduate student. Since 2012 he has been a senior software engineer in the Wireless Platform Research and Development division of Intel Mobile Communications. His current research efforts are focused on channel and interference modeling. Email: Tommaso.Balercia@intel.com.

Christian Rom received his MSc EE in Digital Communication and his PhD from Aalborg University, respectively in 2003 and 2008. In 2003 he joined Infineon as a software engineer. In 2004, he began his industrial PhD under the tutelage of both Infineon and Aalborg University. He is currently with the Wireless Platform Research and Development division of Intel Mobile Communications as the technical group manager of 20 engineers working on several R&D projects. His research interests focus on baseband algorithms for wireless receivers and modem design. Email: Christian.Rom@intel.com.

SPECTRUM SHARING AS A MEANS TOWARDS MEETING FUTURE WIRELESS MOBILE BROADBAND CAPACITY DEMANDS

Contributors

Markus Mueck
Intel Mobile Communications Group
Intel Corporation

Reza Arefi
Intel Mobile Communications Group
Intel Corporation

Srikathyayani Srikanteswara
Intel Labs
Intel Corporation

Geoff Weaver
Intel Labs
Intel Corporation

Mohamed El-Refaey
Intel Labs
Intel Corporation

Graham MacDonald
Intel Global Public Policy Group
Intel Corporation

The mobile data explosion and the ensuing spectrum shortage for mobile operators have spurred discussions on various spectrum sharing paradigms. Recently the concept of Licensed Shared Access (LSA) was put forth by the European Commission, which provides a means for licensed spectrum holders to lease their spectrum to mobile operators. We present a historical overview of spectrum sharing and factors affecting it. We then describe the LSA concept and related activities in ETSI Reconfigurable Radio Systems Standardization. It will be shown how LSA allows mobile operators to maintain QoS in their networks during congestion periods and allows noncellular spectrum holders to economically benefit from sharing their spectrum. While spectrum sharing techniques have been discussed for decades, their implementation has met with practical limitations. LSA defines a framework that provides mutual benefit for both mobile operators and other spectrum holders, thus making spectrum sharing a practical reality.

Utilization of Radio Spectrum: A Historical Note

Radio spectrum is the lifeblood of all wireless communications. Since the early twentieth century when large-scale commercial use of radio spectrum started[1], utilization of radio waves has turned into a significant factor in many economies around the world. This contribution has been amplified with the advent of land mobile cellular systems in the latter part of the twentieth century and with proliferation of mobile Internet and smart devices in the last few years.[2]

In the meantime, new access technologies and device capabilities in recent years have resulted in an exponential surge in demand for wireless data communications.[3] This explosive demand has in many cases led to network congestion and outage. The wireless industry is developing the means to mitigate the problem through implementation of new techniques. However, without the introduction of additional suitable spectrum there cannot be a complete solution found and network operators, sooner or later, will face spectrum exhaustion unless they decelerate their growth.[4] But how can this additional spectrum be obtained?

"…without the introduction of additional suitable spectrum there cannot be a complete solution found and network operators, sooner or later, will face spectrum exhaustion…"

Spectrum Allocation and Use

Historically, radio waves have been the property of governments that could be leased to public or private entities for use for a limited time through a license. Regulations governing the licenses were designed to control harmful interference. There was also a need for reshuffling of spectrum assignments to make room for new applications. This process has been referred to as *spectrum refarming or repurposing*.

The majority of radio spectrum being utilized today is assigned through exclusive licenses; that is, in the geographical region of applicability of the license, only the licensee(s) is allowed to use the spectrum. There are, however, spectrum bands marked as license-exempt, set aside for applications such as Industrial-Scientific-Medical (ISM) to facilitate innovation through removing regulatory burdens related to obtaining licenses.

Mobile Spectrum

Spectrum below 6 GHz is typically considered suitable for implementation of cellular systems. However, only a fraction of this spectrum is currently being used by cellular systems. The remainder of this spectrum is being allocated to a variety of other services such as satellite and broadcasting.

The International Telecommunications Union - Radiocommunications (ITU-R) is an organization of the United Nations that develops recommendations and guidelines including Radiocommunications Regulations (RR), an international treaty governing spectrum allocations, obligatory on all ITU-R member administrations. A process for obtaining new mobile allocation is a necessary step in repurposing of spectrum for cellular systems in domestic regulations of countries.

Problems are, however, mounting for the repurposing process. Recently, the mobile industry has been facing difficulties in obtaining new mobile spectrum identified for International Mobile Telecommunications (IMT), commonly referred to as 3G/4G, a set of radio interface technologies contained in Recommendation ITU-R M.1457 (IMT-2000) and M.2012 (IMT-Advanced). Bands below 6 GHz are generally quite crowded due to favorable propagation characteristics for many services, and even more so in frequencies below 3 GHz suitable for implementation of low cost devices.

"On the one hand, the explosive demand for mobile services is pressing regulators to open more spectrum."

On the one hand, the explosive demand for mobile services is pressing regulators to open more spectrum. On the other hand, finding spectrum that could be repurposed with reasonable cost has become increasingly difficult. In particular the 2.3–2.4 GHz band is a promising candidate; concerned incumbents are using the band often only sporadically and/or over a limited geographic area. Access to this band for mobile services, however, requires a solution that guarantees an efficient and interference-free coexistence with those incumbents.

Spectrum Sharing and the Evolution of Dynamic Spectrum Access

"Spectrum sharing schemes are classified into two categories based on priority to use the band: equal priority, or hierarchical."

Spectrum bands are commonly shared by more than one service. Consideration of the fact that co-primary services in a band may not necessarily use the radio channel at the same time and in the same geographical area led to introduction of Dynamic Spectrum Access (DSA) and associated techniques. Spectrum sharing schemes are classified into two categories based on priority to use the band: equal priority, or hierarchical. Examples of the first category are

unlicensed spectrum or a spectrum pool where all users coexist and manage interference. In Wi-Fi*, this is achieved through Carrier Sense Multiple Access (CSMA), a network control scheme in which a node verifies the absence of other traffic before transmitting.[5] Basic detect and avoid mechanisms suffice in these situations. However, QoS guarantees are not possible in unlicensed spectrum since the number of users in the system, and hence the interference, is not controlled or coordinated. Examples of the second category are UWB, TVWS, and other overlay-underlay techniques based on detecting and avoiding the primary user.

Hierarchical spectrum sharing mechanisms can be broadly classified into:

- *Uncoordinated secondary usage (UWB, overlay-underlay, and so on):* The primary user has no knowledge of secondary user(s). However, the primary user is guaranteed to use the spectrum. The secondary user(s) can only use the channel if it is not in use by the primary. Mechanisms have been proposed to detect unused spectrum[6] which involve sensing and information gathering. While these mechanisms have spurred the largest amounts of interesting research, harmful interference has been the biggest concern.
- *Semicoordinated secondary use (database access, cognitive pilot channel):* The primary user is aware of the secondary users' existence. A database contains the spectrum that can be shared. The secondary user queries the database, as in the case of TVWS, or obtains information about available channels from a cognitive pilot channel.[7] The primary user is not aware of how many secondary users exist in a given location; however, they have the guarantee that none exist where the primary is using the spectrum.
- *Fully coordinated secondary use (such as licensed shared access):* These are evolving schemes that are described in more detail in the section on licensed shared access later in the article. The primary and secondary users fully coordinate on the use of a given spectrum. The secondary user obtains rights and guarantees to the primary user's spectrum during use. In addition, unlike the previous two schemes, the primary user is financially compensated for sharing their spectrum.

"Problems contributed so far to DSA's lack of success include lack of any incentive for the primary to share the band, accuracy of sensing and detection, relinquishing spectrum upon request, unwanted emissions of devices affecting adjacent services, business and security aspects, and regulatory obstacles."

Problems contributed so far to DSA's lack of success include lack of any incentive for the primary to share the band, accuracy of sensing and detection, relinquishing spectrum upon request, unwanted emissions of devices affecting adjacent services, business and security aspects, and regulatory obstacles. Exponential growth predictions for mobile data, however, have renewed interest in DSA. In the United States, NTIA believes that both governmental and nongovernmental users will need to adopt innovative sharing techniques to accommodate the growing demand for spectrum.[8] In Europe, a study mandated by the European Commission[9] concluded that up to 400 MHz of spectrum may become available to commercial wireless communication systems through efficient sharing. The next section provides more details.

Gronsund et al.[10] give a good overview of research in the area of spectrum trading. They conclude that a balance is needed between the cost of the

spectrum and the demand or load on the networks. While the end user prefers the oversupply resulting in low cost, too low of a cost is not sustainable. Caicedo and Weiss[11] describe spectrum trading between operators and conclude at least 6–10 participants are needed. Further, many articles discuss secondary spectrum traders who buy spectrum and resell to another mobile operator. While this makes for interesting academics, it is unlikely that such a scheme will gain commercial acceptance. Spectrum currently used by mobile operators is among the most highly utilized spectrum bands below 3 GHz. Given the projected growth, the situation could only worsen. As a result, sharing schemes involving operators' spectrum do not seem realistic. This has led to recent focus in industry on DSA involving mobile operators accessing spectrum of noncellular incumbents in other bands while providing compensation.

Current Regulation and Standardization Activities in the Field of Spectrum Sharing

In this section, most recent decisions and current opportunities for active participation, mainly for industry stakeholders, are outlined. These will help the reader in understanding how the future regulation and standardization framework for spectrum sharing can be influenced and shaped.

EC and CEPT Activities on Spectrum Sharing: Authorized Shared Access and Licensed Shared Access

The European Conference of Postal and Telecommunications Administrations (CEPT), the pan-European regulatory authority consisting of 47 countries, has discussed an input paper on the *Authorized Shared Access* (ASA) concept[12] since 2011. Under this concept, the existing spectrum user(s) share spectrum with one or several licensed ASA users. A key feature of ASA is to ensure a predictable quality of service for all spectrum rights of use holders, network operators, and for consumers. The more stable the incumbent's spectrum use, the better predictability there is for the QoS of the ASA licensees. The inherently guaranteed QoS is indeed a key argument for major stakeholders, in particular for cellular operators. As an inherent limitation, ASA will operate on shared and noninterference basis, subject to individual authorization (licenses), in bands allocated to the mobile service and identified for IMT by ITU-R.

"The inherently guaranteed QoS is indeed a key argument for major stakeholders, in particular for cellular operators."

"The EC's Radio Spectrum Policy Group (RSPG) has considered the regulatory aspects of the ASA approach and has used this as a basis to foster the potential to share spectrum,…"

On an EC (European Commission) level, ASA has also been considered and further developed. The EC's Radio Spectrum Policy Group (RSPG) has considered the regulatory aspects of the ASA approach and has used this as a basis to foster the potential to share spectrum, which is not only limited to the IMT bands, in a harmonizing manner within a license regime. The RSPG refers to this as *licensed shared access* (LSA).[13]

In October 2010, the CEPT Working Group on Frequency Management (CEPT WG FM) decided to further develop the required regulation

framework in order to enable the usage of LSA. For this purpose, two project teams have been set up:

- *Project Team 52 on 2300–2400 MHz band:* The objective is to develop a draft ECC Decision, aimed at the frequency band 2300–2400 MHz including regulatory provisions based on LSA.
- *Project Team 53 on Reconfigurable Radio Systems (RRS) and licensed shared access (LSA):* The objective is i) to enable the usage of white space devices, in particular in the UHF band (TVWSD) by exploiting geo-location database information, ii) to enable the introduction of LSA by studying the "level of guarantee" in spectrum access under LSA required by an operator for network investment, and iii) to further consider general objectives on reconfigurable radio systems (RRS).

"The 2.3–2.4 GHz band has been identified as the band for the first LSA implementation. "

The 2.3–2.4 GHz band has been identified as the band for the first LSA implementation. Additional bands are expected to follow.

Mandate for Reconfigurable Radio Systems

Beyond activities on the regulation level as indicated above, the EC has triggered related standardization activities by issuing a corresponding standardization mandate covering, among others, the following two directions:

- *Objective 1:* In the area of commercial applications, to enable the deployment and operation of cognitive radio systems (CRS) ... under Licensed Shared Access regime, dependent ... from geo-location databases (GLDB).
- *Objective 2:* To explore potential areas of synergy among commercial, civil security, and military applications.

The ETSI Reconfigurable Radio Systems (RRS) Technical Body is developing corresponding standards.

U.S. Activities

"…the Commerce Spectrum Management Advisory Committee (CSMAC) is specifically analyzing ways to implement the National Broadband Plan to make available 500 MHz of spectrum for wireless broadband."

In the United States, the National Telecommunications and Information Administration (NTIA), an agency of the U.S. Department of Commerce, oversees regulations pertaining to federal use of spectrum; the Commerce Spectrum Management Advisory Committee (CSMAC) is specifically analyzing ways to implement the National Broadband Plan to make available 500 MHz of spectrum for wireless broadband. It is acknowledged by the committee's reports that some of the bands under consideration cannot be made available for mobile use unless innovative sharing techniques are employed. The work is ongoing.

The Federal Communications Commission (FCC) has also put out a Notice of Proposed Rule-Making (NPRM) and public notices outlining the Spectrum Access System (SAS) for spectrum management as shown in Figure 1. They propose having three tiers of users: the incumbents, Priority Access (PA), and General Authorized Access (GAA) users in decreasing hierarchy. The PAs may have to accept interference from the incumbents but cannot cause interference to the incumbents and the GAAs have to accept interference from PAs and

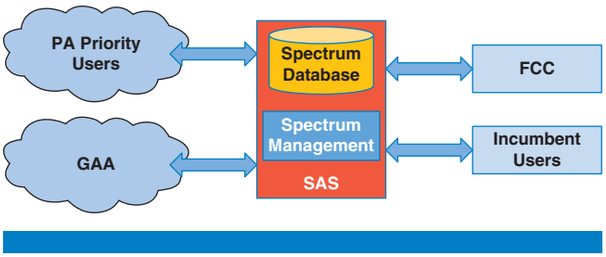


Figure 1: Spectrum Access System for 3.5 GHz (United States)
(Source: Redrawn by Intel Corporation from original provided by FCC, 2013)

the incumbents. The system can accommodate the LSA concept where the PA users are the LSA licensees and the GAA users are in a separate band.

Standardization Activities

As indicated above, the ETSI RRS Technical Body is expected to be the center of competence within ETSI for future spectrum sharing related standards. Currently, ETSI RRS is developing a System Reference Document (SRdoc)[14] that serves as a means for communicating officially with CEPT. Industry players are detailing the requirements on the setup of an LSA framework from their perspective and CEPT is subsequently chartered to consider corresponding regulation changes in order to make possible the entering to the market of such systems.

"…the ETSI RRS Technical Body is expected to be the center of competence within ETSI for future spectrum sharing related standards."

In the framework of the above-mentioned mandate, ETSI RRS is expected to furthermore develop harmonized standards (HS), which are a regulatory tool in Europe. The European Standards Organizations (ESOs), including ETSI, support European legislation in helping the implementation of the EC directives. European Standards developed in response to a mandate are called *harmonized standards.* Those harmonized standards are supporting EU directives and regulations and are essential for device certification.

Licensed Shared Access as a Use Case for Realizing Cognitive Radio

LSA is a technology enabling the secondary usage of spectrum based on a long-term license agreement between an LSA licensee (typically a cellular operator) and an incumbent (such as public safety). The European Commission's *Radio Spectrum Policy Group* (RSPG) currently defines LSA as follows:

"LSA is a technology enabling the secondary usage of spectrum based on a long-term license agreement between an LSA licensee (typically a cellular operator) and an incumbent (such as public safety)."

"A regulatory approach aiming to facilitate the introduction of radiocommunication systems operated by a limited number of licensees under an individual licensing regime in a frequency band already assigned or expected to be assigned to one or more incumbent users. Under the LSA framework, the additional users are allowed to use the spectrum (or part of the spectrum) in accordance with sharing rules included in their rights of use of spectrum, thereby allowing all the authorized users, including incumbents, to provide a certain QoS." [15]

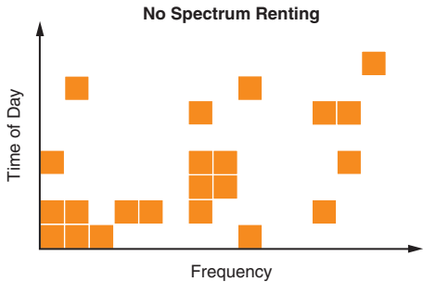


Figure 2: Usage of spectrum capacity without spectrum sharing (Source: Intel Corporation, 2013)

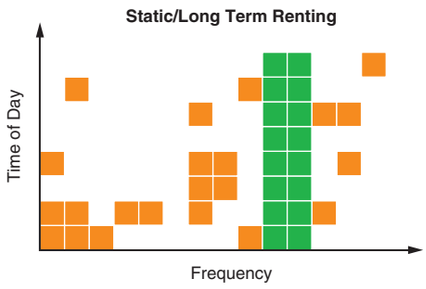


Figure 3: Usage of spectrum capacity with licensed shared access (Source: Intel Corporation, 2013)

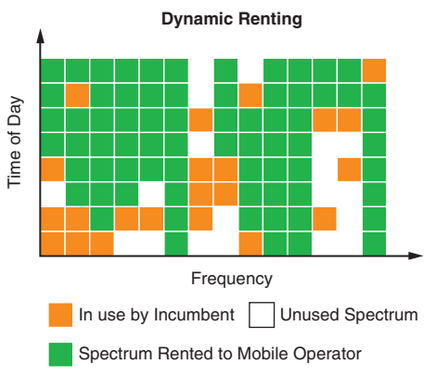


Figure 4: Usage of spectrum capacity with DSA (Source: Intel Corporation, 2013)

The licensing approach in combination with static or quasi-static availability of shared bands leads to a guaranteed level of quality of service (QoS) and a business case that is far more straightforward and obvious than the highly dynamic DSA case.

Indeed, LSA can provide additional resources to mobile operators and economic incentives to governments even if it is used for relatively static and long term spectrum sharing. Figures 2 through 4 illustrate the differences between the traditional, exclusively licensed, LSA- and DSA-based approaches. It becomes obvious that LSA will improve the exploitation of spectrum resources in the short- to mid-term but a quasi-optimum exploitation is expected to require a more dynamic DSA approach in the long term despite the technical, economic, and business feasibility hurdles.

As mentioned previously, the European Commission has recently issued EC Mandate M/512 with a specific request to develop standards enabling the implementation of LSA in Europe. ETSI is working on corresponding solutions in the ETSI RRS Technical Committee and has issued a first deliverable that details the intended first implementation in Europe in the 2.3-2.4 GHz band.[16] This band corresponds to 3GPP LTE Band 40, which was first made available to cellular communication in China. Thanks to this fact, the latest generations of mobile devices support this mode and are thus inherently "LSA ready" in the 2.3–2.4 GHz band—under the assumption, of course, that no further features are required in the mobile devices and the access to LSA bands is managed from the network infrastructure side. The high-level system design proposed by ETSI is further illustrated in Figure 5. In this context, two new functions are introduced into the wireless and mobile ecosystem:

- The LSA repository (that constitutes a geo-location database) interacts with incumbents in order to gain information on LSA band availabilities and access conditions for LSA licensees
- The LSA controller accesses the LSA repository in order to derive LSA spectrum access requirements for LSA licensees. Typically, the LSA controller interfaces with the network infrastructure via the cellular operators' OA&M system.

Building on the upper high-level, conceptual presentation of LSA, ETSI RRS currently further develops the LSA system specification. The definition of system requirements is currently ongoing in ETSI TS 103 154: "System requirements for LSA in 2300-2400 MHz."[17] This activity is expected to be followed by the detailed definition of an LSA system architecture and finally the definition of related interfaces.

In parallel to ETSI standardization activities, CEPT WG FM is working towards ensuring the readiness of LSA introduction to the market from a

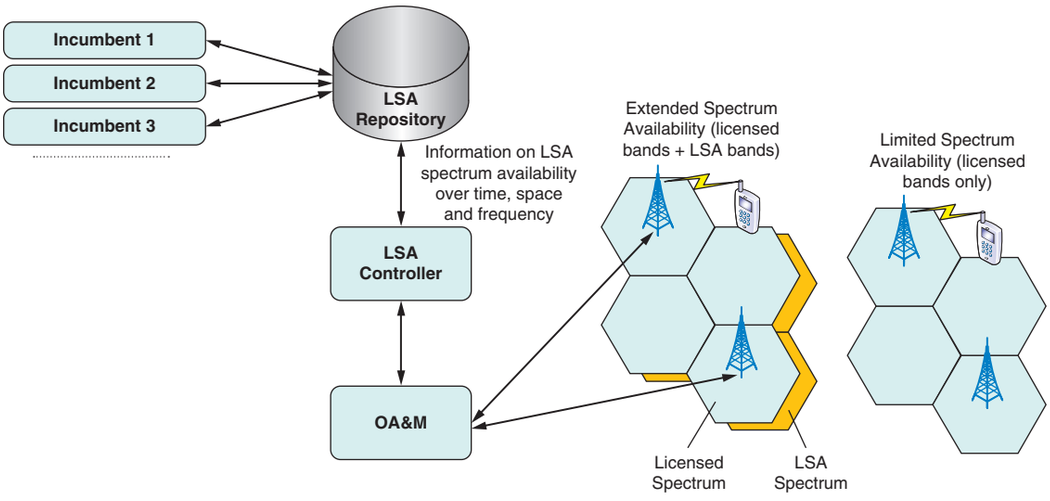


Figure 5: Licensed shared access approach (Source: Intel Corporation, 2014.)

regulation perspective. Indeed, two project teams are developing deliverables with the following scope:

- CEPT WG FM PT52: "The Project Team shall: develop a draft ECC Decision, aimed at harmonising implementation measures for MFCN (including broadband wireless access systems) in the frequency band 2300–2400 MHz."[18]
- CEPT WG FM PT53: "The Project Team shall handle the following tasks: … Develop an ECC Report on general conditions, including possible sharing arrangements and band-specific (if not dealt with by a specific project team) conditions for the implementation of the LSA that could be used as guidelines for CEPT administrations."[19] A corresponding ECC Report is about to be finalized under the title "ECC Report 205: Licensed Shared Access (LSA)."

The development of LSA is thus a brilliant example for an efficient interaction of the European Commission, CEPT, and ETSI driving the introduction of a new technology to the market.

Summary and Conclusions

As a result of extensive research in the field of cognitive radio (CR) over more than 10 years, the technology is finally reaching market readiness, in particular with the imminent market introduction of LSA in Europe in the 2.3–2.4 GHz band. This late adoption of the technology is due to the complex interrelations between concerned stakeholders and the disruptive economic and business

"…the technology is finally reaching market readiness, in particular with the imminent market introduction of LSA in Europe in the 2.3–2.4 GHz band."

considerations. Investment in network infrastructure is only justified if there are guarantees of available mobile and wireless spectrum capacity—this simple fact still represents a key hurdle for further CR solutions that have become mature from a technical perspective.

References

[1] Hazlett, Thomas W., "Optimal abolition of FCC spectrum allocation," *Journal of Economic Perspectives*, Vol. 22, No. 1, 2008.

[2] ITU-D, "Impact of Broadband on the Economy," April 2012.

[3] UMTS Forum Report 44, "Mobile traffic forecasts 2010–2020 report," January 2011, available at http://www.umts-forum.org/component/option,com_docman/task,doc_download/gid,2537/Itemid,213/.

[4] Goldman, David, "Sorry America: your wireless airwaves are full," CNN Money, http://money.cnn.com/2012/02/21/technology/spectrum_crunch/index.htm; February 21, 2012.

[5] National Communication Systems, Federal Standard 1037C, http://www.its.bldrdoc.gov/fs-1037/fs-1037c.htm.

[6] Yucek, T. and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Communications Surveys & Tutorials*, Vol. 11, Issue 1, 2009, pp. 116–130.

[7] Filo, M., A. Hossain, A. R. Biswas, and R. Piesiewicz, "Cognitive pilot channel: Enabler for radio systems coexistence," *Second International Workshop on Cognitive Radio and Advanced Spectrum Management*, CogART 2009, pp. 17–23.

[8] U.S. Department of Commerce, "An Assessment of the Viability of Accommodating Wireless Broadband in the 1755–1850 MHz Band," March 2012, available at http://www.ntia.doc.gov/files/ntia/publications/ntia_1755_1850_mhz_report_march2012.pdf.

[9] SCF Associates for the European Commission, "Perspectives on the value of shared spectrum access - Final Report for the European Commission," February 2012, available at http://ec.europa.eu/information_society/policy/ecomm/radio_spectrum/_document_storage/studies/shared_use_2012/scf_study_shared_spectrum_access_20120210.pdf.

[10] Gronsund, P., R. Mackenzie, P. H. Lehne, K. Briggs, O. Grondalen, P. E. Engelstad, and T. Tjelta, "Towards spectrum micro-trading," *Future Network & Mobile Summit* (FutureNetw), pp. 1–10.

[11] Caicedo, C. E. and M. B. H. Weiss, "The Viability of Spectrum Trading Markets," *2010 IEEE Symposium on New Frontiers in Dynamic Spectrum,* DYSPAN 2010, pp. 1–10.

[12] CEPT WG FM, Report on ASA concept, document FM(12)084 Annex 47.

[13] Radio Spectrum Policy Group, "Collective Use of Spectrum," 2011, (a report on CUS and other spectrum sharing approaches) available at http://rspg.ec.europa.eu/_documents/documents/meeting/rspg26/rspg11_392_report_CUS_other_approaches_final.pdf.

[14] ETSI, "Electromagnetic compatibility and Radio spectrum Matters (ERM); Mobile broadband services in the 2300 MHz – 2400 MHz frequency band under Licensed Shared Access regime," System Reference Document, 2012.

[15] Radio Spectrum Policy Group 2011, "Collective Use of Spectrum," November 2011 (a report on CUS and other spectrum sharing approaches), http://rspg-spectrum.eu/_documents/documents/meeting/rspg26/rspg11_392_report_CUS_other_approaches_final.pdf.

[16] ETSI, "Electromagnetic compatibility and Radio spectrum Matters (ERM); Mobile broadband services in the 2300 MHz – 2400 MHz frequency band under Licensed Shared Access regime," System Reference Document, 2012.

[17] ETSI TS 103 154, "Reconfigurable Radio Systems (RRS); System requirements for operation of Mobile Broadband Systems in the 2300 MHz - 2400 MHz band under Licensed Shared Access (LSA) regime," forthcoming.

[18] CEPT WG FM PT52, Terms of References, http://www.cept.org/ecc/groups/ecc/wg-fm/fm-52/page/terms-of-reference.

[19] CEPT WG FM PT53, Terms of References, http://www.cept.org/ecc/groups/ecc/wg-fm/fm-53/page/terms-of-reference.

Author Biographies

Markus Mueck (Markus.Dominik.Mueck@intel.com) received the Doctorate degree of the Ecole Nationale Supérieure des Télécommunications (ENST), Paris, in Communications. He is with Intel Mobile Communications and currently acts as General Chairman of ETSI RRS (SDR and Cognitive Radio Standardization) and is an adjunct professor at Macquarie University, Sydney, Australia. Dr. Mueck has filed over 90 patents, published over 80 scientific conference and journal papers, acts as TPC member of numerous conferences, and is involved as reviewer for the

evaluation of European Research projects in the 7th Framework Programme of the European Commission.

Reza Arefi (Reza.Arefi@intel.com) directs market-driven spectrum strategies for Intel's wireless products at Intel's Standards and Advanced Technology group in the Mobile Communications Group (MCG SAT) and regularly represents Intel in regional and international regulatory radiocommunication standards organizations including ITU-R.

Srikathyayani Srikanteswara (Kathyayani, Srikathyayani.Srikanteswara@intel.com) is a senior research scientist at Intel Labs. She leads research on spectrum sharing and has been involved in standardization efforts in the United States and Europe. With over a decade of experience, she has shaped research in cognitive radios, TVWS, simultaneous operation of radios, and developed and prototyped spectrum sensing and interference mitigation algorithms. Prior to joining Intel, she was with Navsys Corporation, and a research faculty member at Virginia Tech. She received her BS with Honors in Electrical Engineering from IT-BHU, India, and her MS and PhD from Virginia Tech.

Geoff Weaver (Geoff. Weaver@intel.com) is a portfolio strategist with Intel Labs with a focus on wireless and IoT (Internet of Things) topics. Geoff was previously a technology strategist for over five years with Intel Labs where he developed strategies to maximize Intel value for research in wireless, manageability, and security technologies. Geoff has held various strategy and business development positions over his 16 years at Intel. Prior to his career at Intel, Geoff held senior marketing and product management positions at software and semiconductor companies.

Mohamed El-Refaey (Mohamed.ElRefaey@intel.com) is a senior research scientist in Emerging Platform Solutions at Intel Labs, with 14+ years of experience in leadership positions with established public companies. He has diverse experience in cloud, virtualization, wireless systems and software design and development. He is an expert in cloud computing and an invited speaker at many national and international conferences on the subject. He has authored and published many technical papers and has contributing chapters to three books on cloud computing and cognitive radio. He has been awarded in recognition of innovation and thought leadership while working at EDS. He is also a key contributor to the IEEE P2302 Standard for Intercloud Interoperability and Federation (SIIF).

Graham MacDonald (Graham.MacDonald@intel.com) is Director for Europe, Middle East, and Africa (EMEA) Communications Policy within Intel's Global Public Policy (GPP). Graham joined Intel in 2002 and is based in the United Kingdom, working closely with Intel's Washington D.C. Office on communications policy initiatives. Graham's current focus includes technical and policy advocacy leading up to the World Radiocommunications Conference 2015 and beyond to secure access to additional spectrum for mobile broadband and Wi-Fi, influencing the European regulatory

framework on spectrum sharing under licensed shared access (LSA), and representing Intel public policy in the UK. Previously, Graham worked for Philips Electronics, UK Ministry of Defense, UK Radiocommunications Agency (now Ofcom), Nortel Networks, Intellect (renamed "techUK") and UMTS Forum.

VIRTUAL PROTOTYPING METHODOLOGY FOR MOBILE PLATFORMS

Contributors

Guido Stehr
Intel Mobile Communications
Intel Corporation

Josef Eckmüller
Intel Mobile Communications
Intel Corporation

Oliver Bell
Intel Mobile Communications
Intel Corporation

Thomas Wilde
Intel Mobile Communications
Intel Corporation

Ulrich Nageldinger
Intel Mobile Communications
Intel Corporation

Intel has driven innovation in computing over decades. Intel's goal is to bring the powerful x86 Intel® Architecture to virtually all computing devices. Rarely are those devices used in isolation. Instead, virtually everything that computes connects, be it wired or, increasingly, wireless. The proliferation of mobile devices puts a particular emphasis on cellular connectivity, bringing about an always-on, always-connected lifestyle. Those consumer-oriented devices are characterized by fast-paced innovation where more and more functionality is implemented in software. Fast-paced innovation is only possible if software and system verification can be done before the availability of silicon. In response to this challenge, development teams at Intel adopted virtual prototypes (VPs), which are simulation models of the entire system hardware and the verification environment. These VPs consist of transaction-level models, which are far more abstract than RTL representations. The chosen level of abstraction must properly balance features to be modelled on the one hand and simulation speed on the other. Simulation speedups of more than two orders of magnitude as compared to RTL enable the simulation of complex use cases. This allows pre-silicon system concept validation and production software development. This article describes the infrastructure and methodology used for transaction-level modeling at Intel and outlines a number of success stories for integrated mobile platforms.

Introduction

At Intel we deal with a wide variety of computing devices, from high-end servers down to tiny embedded systems. All these different device classes have their own characteristics and design challenges. Hence, there are a number of optimized variants of the general design flow. This article deals with parts of the SoC design flow for mobile communication platforms.

"Intel's goal is to deliver innovative mobile platform solutions into the OEM ecosystem that enable mobile devices staying connected everywhere, all the time…"

Mobile Platforms

Intel's goal is to deliver innovative mobile platform solutions into the OEM ecosystem that enable mobile devices staying connected everywhere, all the time:

- Entry platforms: cost-efficient 2G (GSM)/2.5G (EDGE) single-chip platforms for ultra-low-cost feature phones and entry-level smartphones
- Smart platforms: leading-edge 3G (UMTS), HSPA and 4G (LTE) platforms for smartphones and tablets for the high-end market, featuring high-speed Internet access

A typical mobile platform consists of these parts:

- RF front end: reception and transmission for multiple radio standards (2G, 2.5G, 3G, and 4G)
- Analog Baseband (ABB): modulation and demodulation
- Digital Baseband (DBB): digital signal processing (signal modulation, radio frequency shifting, encoding, and so on) and communication protocol stack for the different radio standards
- Power Management Unit (PMU): power supply and power saving
- Software (SW): wide variety of code ranging from low-level control and signal processing (often referred to as firmware, FW) to operating system and user applications
- Application processor: execution of higher-layer software subject to no or soft real-time requirements

One of our LTE offerings is the Intel® XMM™ 7260 platform, as shown in Figure 1. It is a slim modem platform with carrier aggregation. It was developed specifically for smartphone LTE architectures. It is based on the Intel® X-GOLD™ 726 baseband processor and the Intel® SMARTi™ 4.5 RF transceiver. The platform enables download speeds up to 300 Mbps and upload speeds up to 50 Mbps.

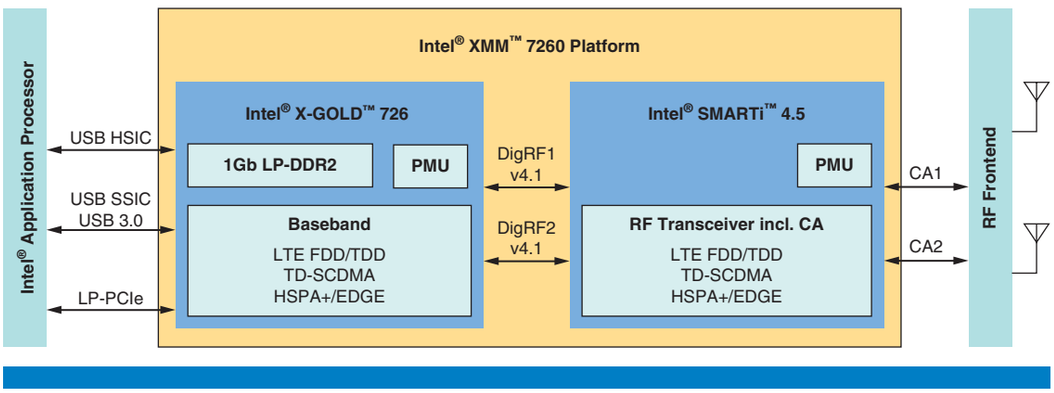


Figure 1: Intel® XMM™ 7260 Mobile Platform (Source: Intel Corporation, 2014)

Virtual Prototyping

The development of a mobile platform requires comprehensive know-how in the areas of RF, mixed signal, power management, monolithic integration, costs, and mobile phone software. Delivering into a consumer-driven market requires a yearly product innovation cadence and has to enable the mobile phone supplier to exactly hit the seasonal landing zones with their products. As levels of System on Chip (SoC) integration have increased and software development for silicon devices has mounted, software development cycles have increasingly imposed themselves on project critical paths. Without

"As levels of System on Chip (SoC) integration have increased and software development for silicon devices has mounted, software development cycles have increasingly imposed themselves on project critical paths."

concurrent engineering for chip and software it would not be possible to meet such schedule requirements.

A virtual prototype (VP) provides a software simulation model of an SoC device, including processors, peripherals, memories, interconnect, and connectivity. It may have multiple abstraction levels, such as functional, instruction-accurate, or cycle-accurate, depending on its application scenario. With a virtual prototype it is possible to run unmodified production software (application code, drivers, and so on), sometimes close to real-time.

System-Level Design Flow

In general, a design flow is the combination of electronic design tools along with the documentation of best known methods to support the complete design process of an electronic system. Virtual prototyping is one part of such a design flow. It is a design technique at a comparatively high level of abstraction, commonly referred to as system level.

Motivation

Virtual prototype development for mobile platforms is a challenge in itself. At Intel we found it useful to have a common system-level design flow across the sites and projects, extending the existing classical SoC RTL to GDSII hardware implementation flow. In our experience it is useful to have a focused expert team developing and supporting the common system-level design flow, whereas the virtual prototype development is executed within the project teams.

"The Intel system-level design flow provides an efficient framework for fast virtual prototype development based on the SystemC standards."

The Intel system-level design flow provides an efficient framework for fast virtual prototype development based on the SystemC[1][2] standards.[3] It runs both on Windows* and on Linux*. Reflecting our commitment, Intel led the definition of the TLM2 standard[2] and plays a significant role in further evolving SystemC.[4][5][6][7]

The use of a common system-level design flow helps to:

- Enable teamwork and model exchange across divisions
- Standardize workflows
- Provide tested infrastructure
- Let design teams focus on core competence instead of tool setup and maintenance

Application

In our system level design methodology, virtual prototypes are used for the following main use cases, as shown in Figure 2:

- *Architecture exploration, performance and power simulation:* In the early project phase the virtual prototype is used for system architecture exploration and feasibility analysis prior to start of implementation. This helps in finding the optimized system architecture for specific use cases. For example, the virtual prototype helps to optimize the cache sizes or to

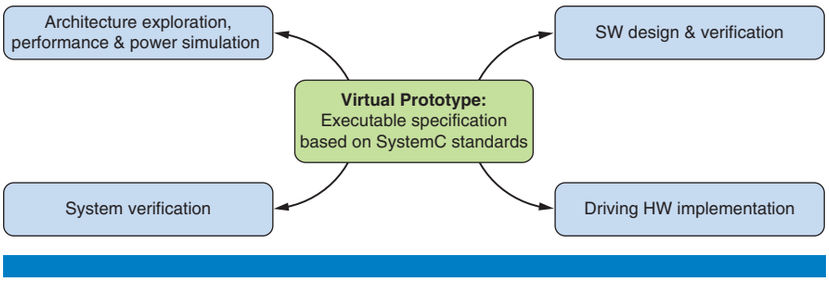


Figure 2: Main use cases for virtual prototypes
(Source: Intel Corporation, 2014)

analyze whether the hardware performance is sufficient to run specific use cases (see the section "Combination of Cores").

- *System verification:* Virtual prototypes help to prepare system test cases prior to silicon availability. This allows for fast system verification when silicon has arrived; the test case setup is available by then (see the sections "Hardware/Software Co-Debugging" and "Early Software Development").
- *Driving hardware implementation:* Virtual prototypes can be used as references for hardware implementation. A verification of RTL components at system level is possible through their instantiation in virtual prototype test benches (see the section "Fast Hardware Simulation"). In such a VP/RTL co-simulation, parts of the VP are replaced by their RTL representation. In this configuration the RTL component can be tested with system test cases. Moreover, the co-simulation also makes sense when an abstract model of the peripheral is not available or when a cycle-accurate model is required within the VP.
- *Software design and verification:* VPs model the system hardware and therefore they can be used to run embedded software on the cores, which are integrated in the virtual prototype. Early availability of executable models allows an earlier start of software development on the virtual prototype without silicon as outlined in Figure 3. This enables a reduction of the overall system design cycle as exemplified in the section "Early Software Development."

"…Virtual prototypes help to prepare system test cases prior to silicon availability. This allows for fast system verification when silicon has arrived…"

"Early availability of executable models allows an earlier start of software development on the virtual prototype without silicon…"

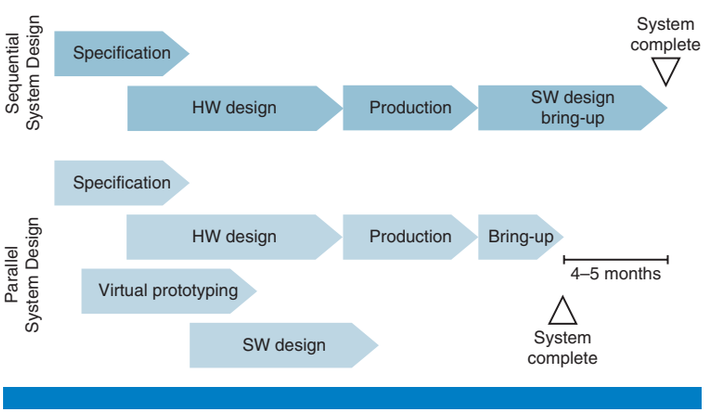


Figure 3: Accelerated system design
(Source: Intel Corporation, 2014)

Figure 4 gives an overview of the workflow used for virtual prototype development along with the titles of the sections where each of these steps is further discussed.

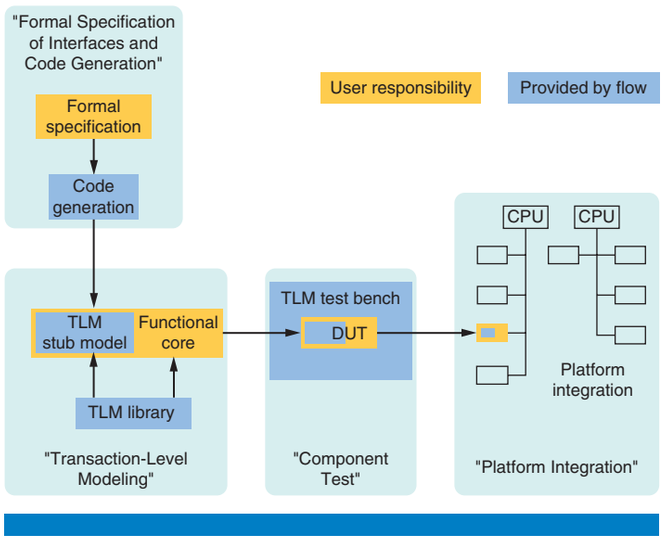


Figure 4: System-level design flow overview (Source: Intel Corporation, 2014)

The section "Formal Specification of Interfaces and Code Generation" describes how a formal specification allows to automatically generate code supporting the design and verification of the VP. In the following section, "Transaction-Level Modeling," we explain how the use of modeling libraries helps to reduce the effort for model development. The "Component Test" section focuses on the validation of the specification with a test bench infrastructure and how consistency between the abstract model and its RTL implementation can be verified. The section "Platform Integration" explains how the developed models are integrated in the virtual prototype of the entire platform and how the software can be executed.

Formal Specification of Interfaces and Code Generation

The first step in the development process for virtual prototypes is the automatic generation of the part of the code that is related to static design aspects (interfaces and registers) from a formal component description.[9][10]

"Creating a formal specification instead of written prose text implies extra effort for the system engineer, but provides overall benefits."

Motivation

Creating a formal specification instead of written prose text implies extra effort for the system engineer, but provides overall benefits. The term "formal" in this context means that such a specification is exact, unambiguous, and machine-processable, thus enabling automated derivation of target code.

The formal specification serves as the single source for the generation of code in several design steps, which multiplies the time-saving effects. These steps include algorithm development, virtual prototyping, software development, driver development, RTL design, component verification, and documentation generation. The benefits of this approach can be summarized as follows:

- The individual designer is relieved of manually writing those code parts that can be generated automatically. Even for components with only a few dozen registers the generation produces several thousand lines of SystemC code (register and memory descriptions, interface descriptions) and software code (access functions). Automatic code generation easily saves several person months of work in a project.
- In addition to the pure time saving the generated code is correct by construction, thus further avoiding time-costly bug fixes.
- Consistency is guaranteed throughout the different disciplines and development phases.
- If there is a change in the formal specification, this change can be propagated very quickly to all process steps, significantly reducing the cycle time for bug fixes.

Outline of the Code Generation Infrastructure

At the core of the code generator infrastructure is a flexible data model suitable for capturing static design aspects related to interfacing (registers, bus interfaces, and ports) of components as well as connectivity (component wiring) for systems. Behavioral aspects are not supported. The scope of the description is a superset of the standardized IP-XACT format.[10] We try to abstract from any implementation and allow for combination of both conceptual and implementation aspects in one description. The resulting descriptions are XML files following a specific schema.

The two pillars of the generator infrastructure are a flexible data model and tooling based thereon. This includes import/export filters, interactive editing, and automatic generation of code. Figure 5 shows the workflow.

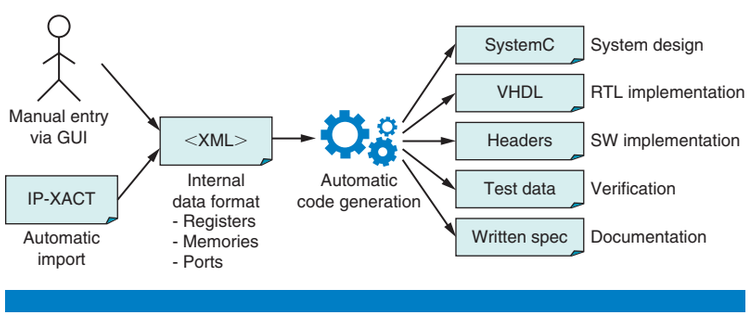


Figure 5: Code generation overview
(Source: Intel Corporation, 2014)

The central data model can be filled either manually using a respective GUI or automatically using import filters. These import filters can read numerous custom text formats (XML, CSV, XLS, TXT) or the IP-XACT format, which is available from many IP providers.

The generation tooling employs the open source template engine "Mako"[9][11]; the tool reads one or more XML files as well as a template for the specific target code to be generated. Making use of a template-based approach allows for easy development and maintenance of generators, since the static parts of the target code can be put into the template directly, with additional constructs for the retrieval of information from the data model to fill in the dynamic parts.

Application for Virtual Prototype Development

The relevant code for VP development that is derived from the formal specification comprises both SystemC code for the VP model as well as software functions.

The VP code contains register and memory classes including access methods for attributes like reset value and so on. The bit fields are transformed into corresponding field access methods in the register classes. Enumerations are mapped to (C language) enumerations of valid constants. The bus interfaces and ports are mapped to TLM sockets[2] and SystemC ports, with special handling of reset and clock ports.

The software code contains access macros for registers and bit fields as well as enumerations for reset values, offsets and valid constants for the bit fields.

Transaction-Level Modeling

"Many modern circuits have reached a size where they cannot efficiently be handled at the Register Transfer Level any more."

"In our virtual prototype methodology we use Transaction-level modeling for hiding hardware aspects that are not relevant at system level."

Digital electronic circuits have been designed at the Register Transfer Level (RTL) since the late 1980s. At this level of abstraction, digital circuits are described by data transfers between registers and the logic operations on this data. Many modern circuits have reached a size where they cannot efficiently be handled at the Register Transfer Level any more. As a solution, more abstract description styles have emerged, which are often embraced by the term Electronic System Level (ESL).[12] Transaction-level modeling (TLM) is one such ESL technique.

Introduction

Transaction-level modeling uses abstract representations for communication channels and payload data.[1][3][8] In our virtual prototype methodology we use Transaction-level modeling for hiding hardware aspects that are not relevant at system level. This leads to a high simulation speed. Typically, IP components making up a platform can be divided into two categories: standard IP (CPU/DSP cores, buses, memories) and custom platform-specific IP (hardware accelerators, peripherals, control blocks). In order to provide a complete VP, models for both IP types are required. While models of standard IP are usually commercially available those for custom IP components need to be developed within the project. In order to enable fast and convenient development of

custom models we offer a comprehensive TLM environment that serves as common, unified platform across projects.

The transaction-level modeling process is part of the overall VP development flow. It is based on input from SystemC code generation and the resulting models will be used within the VP of the complete platform. The TLM infrastructure consists of modeling libraries, build system, IDE, tools for debugging, code analysis, and documentation generation.

Structure of a Transaction-Level Model

A transaction-level model is usually separated into a model stub and the functional model core as shown in Figure 6.

The model stub represents the interfaces and structural details of a component. It can be generated by mapping the information from the generic component description into the SystemC modeling domain. A typical model stub contains bus and wire ports, type and number of resources (registers, embedded memory blocks), and the mapping of resources to bus port address offsets.

The functional model core contains details of the actual behavior of a hardware component. It can use the members of the model stub class, from which it is derived. Typical content of the model core are register access callbacks, functions for data processing, and control flow elements like state machines.

The separation of a model into stub and core allows a fresh generation of the formal part of the model whenever the specification changes without affecting the user-supplied details in the model core. This, of course, assumes that the specification changes do not relate to functionality reflected in the core already.

Reuse of Algorithmic Models

Modern mobile devices need to support a variety of complex communication standards and data formats. In order to execute various system use cases the VP also needs to offer all of this functionality. This makes model creation and verification potentially complex and expensive. By reusing existing algorithmic cores that were created during the earlier concept phase we can ensure consistency and increase modeling efficiency: the algorithmic core (for example from a stream-driven simulation tool), is wrapped with a TLM shell that implements the high-level control flow, data exchange, and timing.

Transaction-Level Modeling Library

On top of SystemC we built an in-house TLM library to fulfill the specific needs of our SoC development and in order to provide a competitive advantage. It is the foundation of our transaction-level models and aims at components with memory-mapped resources that communicate over system buses. It provides a common model structure and base functionality for all hardware component models along with a set of building blocks. It also enforces common rules and concepts for all models. It is designed to facilitate TLM component development and foster interoperability of models from different contributors during platform integration.

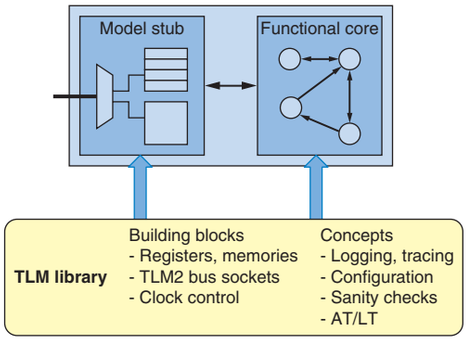


Figure 6: TLM component modeling (Source: Intel Corporation, 2014)

"On top of SystemC we built an in-house TLM library to fulfill the specific needs of our SoC development and in order to provide a competitive advantage."

The library mainly provides the following concepts and components to support the creation of transaction-level models at the AT (approximately timed) and LT (loosely timed) level:

- Resource models (registers, memories, address spaces)
- Bus access and resource decoding functionality
- Resource access callbacks
- Clocking and timing handling
- Reset and power domain handling
- Power state tracing
- Logging and tracing capabilities
- Parameter handling and configuration infrastructure
- Building blocks including interconnect, clock control, interrupt control, and a host-emulated CPU

"The benefit of our common transaction-level modeling library in conjunction with generated SystemC code became obvious again when we adopted the OSCI TLM 2.0 standard."

The benefit of our common transaction-level modeling library in conjunction with generated SystemC code became obvious again when we adopted the OSCI TLM 2.0 standard. Most of the legacy models could be migrated with minor effort: the SystemC generators were updated to use the modified library elements, which encapsulate many implementation details and offer the user a modeling convenience layer. Functional component cores could often remain untouched, since they mostly relied on the abstract interface.

"The TLM development process is supported by a build system which handles the details of building and combining all the required artifacts in the correct order."

Build Environment

The TLM development process is supported by a build system which handles the details of building and combining all the required artifacts in the correct order. The capabilities of the build system not only cover the generation of libraries and executables. It rather automates steps of the TLM workflow like code generation, component testing, code analysis, and documentation generation. This build process can be controlled by the user through customization of default makefile templates. This enables efficient generation of up-to-date, consistent, and compatible output data and thus is a major prerequisite for successful integration of numerous individual contributions into the overall platform VP.

Component Test

The main purposes of transaction-level modeling are architecture design, software development, and system verification. For these applications the TLM component has to correctly reflect its specification. In our VP development flow we ensure this following these design principles:

- The test cases used for specification validation can be reused for verifying the proper platform integration of the TLM component.

- The register and bit field macro functions used for developing the specification validation test cases are generated from a single source description (see the earlier section "Formal Specification of Interfaces and Code Generation"). The same access macros are used for the software accesses on the virtual prototype and on silicon.
- The TLM component test bench and the test cases are developed and compiled independently. This reflects the independent development of hardware and software at virtual prototype level.

To ensure that a TLM component correctly reflects its specification we use a generic specification validation test bench architecture with OSCI TLM 2.0 compliant interfaces. The target sockets of the TLM DUT are connected to a generic TLM interconnect model and the interrupt signals are connected to a generic interrupt control. The test cases are executed on a generic host-based CPU emulator[9], which provides an initiator socket connected to the interconnect model and which is able to handle the interrupts triggered by the DUT. The test bench architecture is shown in Figure 7.

"To ensure that a TLM component correctly reflects its specification we use a generic specification validation test bench architecture with OSCI TLM 2.0 compliant interfaces."

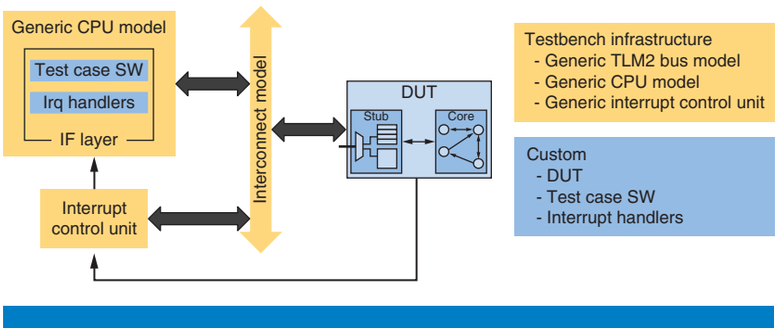


Figure 7: Specification validation test bench architecture (Source: Intel Corporation, 2014)

Generic TLM Interconnect Model

The OSCI TLM 2.0 compliant interconnect model supports any number of masters and slaves connected to its initiator and target sockets, respectively. It can be configured for an arbitrary data width. All OSCI TLM 2.0 transport schemes are supported: blocking, nonblocking, debug, and DMI calls. Hence, the interconnect model can be used for OSCI TLM2 AT and LT modeling style.[2] The interconnect models support a hub and a crossbar routing structure; round robin and priority arbitration schemes can be used. Routing structure and arbitration are defined at construction time. An interface function is provided for binding and address range registration of a target socket connected to the interconnect model. The bus model is developed using the TLM libraries described in the "Transaction-Level Modeling Library" section and therefore inherits the features supported by the library. This is especially valid for the logging and tracing mechanisms.

Generic Host-based CPU Emulator

A CPU emulator[9] executes the test case software, which is natively compiled for the host CPU. Hence, it is not an instruction set simulator, but the focus is on functionality and fast execution.

The test case function has to be registered with the CPU emulator and is executed after the reset is finished.

By default the execution of the software on the host-based CPU emulator is untimed. But there are two ways to model timing for software execution. First, the bus access timing can be configured. Second, the CPU emulator offers a *wait()* function that can be included in the software to model delays. This basically allows the annotation of timing that is typically consumed when the software executes the instructions on the CPU. In a production software build this *wait()* function is simply ignored.

For each interrupt signal the handler of an interrupt service routine can be registered along with its priority and execution time. Based on that, incoming interrupts are scheduled preemptively.

Test Case Infrastructure

The test case software uses the generated access macros for register and bit field accesses. Within these macros bus read and write accesses via OSCI TLM 2.0 initiator and target sockets are initiated which then trigger the bus access functions of the CPU emulator.

In order to enable independent development and compilation of the test cases and the test bench hardware model, the test case code is compiled into a dynamic library and loaded into the executable at runtime. This is achieved by exporting the following functions that constitute the interface between the test bench hardware model and the test cases:

- Software *main()* function that is triggered after a reset of the CPU emulator
- Interrupt service routines
- Bus read and write functions

These functions exist both in the test bench hardware model and in the test case code. The software *main()* function and the interrupt service routines are triggered from the CPU emulator and the bus read/write functions are triggered from the test case code.

Separate test cases are implemented for each DUT feature given in the specification. The simulation is started with the test case name as argument. This infrastructure can conveniently be used by regression runners.

"Once a virtual prototype behaves as intended, consistency must be ensured between the TLM implementation and its RTL counterpart."

Consistency between TLM and RTL

Once a virtual prototype behaves as intended, consistency must be ensured between the TLM implementation and its RTL counterpart. There are different measures to do this, which should be combined for best results.

First of all, generating regular code from a single specification automatically makes sure that the different disciplines have a consistent view on the interfaces of the components such as registers, memories, or ports. For the TLM and RTL domain we generate component stubs which are then filled with functionality manually. The software implementation is supported through the generation of register and memory access functions.

Beyond this it is comparatively easy to generate functional code in well-defined modeling domains. For example, we have a tool to transform UML descriptions of finite state machines to TLM and RTL. This is a first step towards high-level synthesis. If the project can meet the requirements on the input model, which must be clocked and wire accurate, then this can be a viable path from TLM to RTL.

Since any model transformation, be it manual or automatic, can produce errors, it is a good idea to formulate assertions that ensure essential properties at both levels of abstraction.[14]

Along this line, a common test bench can be used to test TLM and RTL models together as shown in Figure 8. There are two fundamentally different test techniques[15]: directed tests represent real-world application scenarios and try to validate the functionality from a feature point of view; constrained random approaches, in contrast, apply random test patterns in order to achieve a good coverage. In both cases it is important to bridge the gap between the RT level of abstraction and TLM appropriately:

"…a common test bench can be used to test TLM and RTL models together…"

- The translation between abstract transactions (TLM) and signal protocols (RTL) is done via state machines called transactors, also known as bus functional models.
- Incompatible signal types (for example, C enums vs. integers) can be mapped by a stateless type translator
- Elementary signal types (such as integer) can be connected directly.

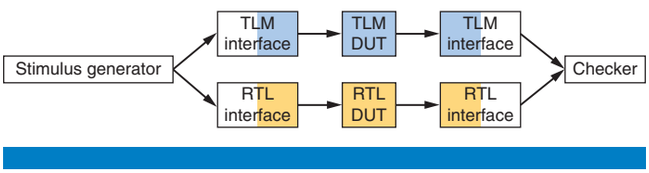


Figure 8: TLM/RTL co-simulation in common test bench (Source: Intel Corporation, 2014)

Platform Integration

After all TLM components have been verified according to the previous section they are brought together for the complete picture during platform integration. This is a challenging step because the entire system must be mastered, which is usually beyond the scope of the individual contributors. Through its build system (see the earlier section "Build Environment") the system level design flow makes sure that the code of all involved components is properly handled during platform

"Platform integration is a challenging step because the entire system must be mastered, which is usually beyond the scope of the individual contributors."

"Virtual prototypes provide good visibility and can thus help to solve integration problems and to find actual system bugs. We observe that virtual prototypes are frequently used for system debugging even after hardware has become available."

"Platform integration is a particular challenge because in this step components from numerous sources come together…"

build. The system integrator can then focus on the actual modeling challenges. Virtual prototypes provide good visibility and can thus help to solve integration problems and to find actual system bugs. We observe that virtual prototypes are frequently used for system debugging even after hardware has become available.

IP Instantiation

Platform integration is a particular challenge because in this step components from numerous sources come together, as shown in Figure 9. A few of these should be emphasized:

- Custom TLM components as described in the previous sections of this article. These models may serve different purposes: at component scope they can be used as references for RTL design. At the system level they may not need to exhibit all the implemented details. Instead, it improves the simulation efficiency to have a configurable level of detail and to offer implementation variants. By the same token the modeling library offers support for dynamically configurable debug levels.
- Peripherals, such as USB components, often come as commercial IP, which allows the focusing of in-house development on differentiators. In this context the importance of the TLM 2.0 standard with respect to model interoperability becomes evident.
- CPU models play a prominent role at the platform level. If functional correctness is the focus, then the use of generic CPU models as described in the section "Generic Host-based CPU Emulator" is a viable and very efficient approach. For performance analysis more detailed core models are required: instruction set simulators (ISS) reproduce the behavior of the target CPU core. They interpret the target code at runtime and emulate the core behavior on the actual host CPU that executes the simulation. For

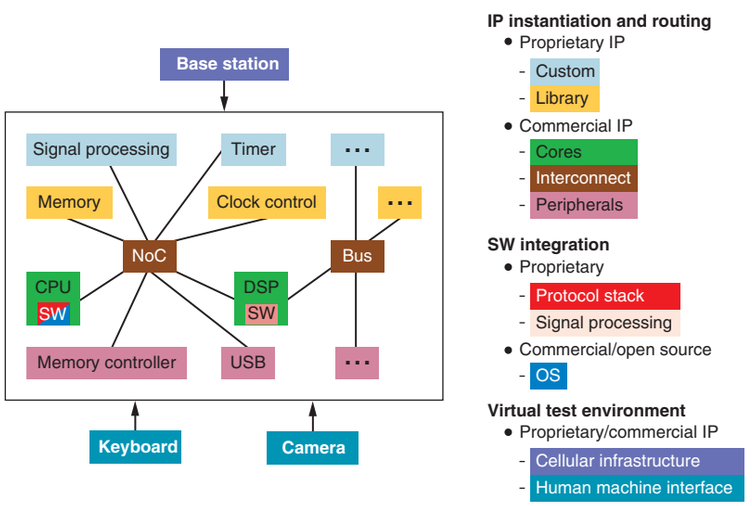


Figure 9: Platform integration
(Source: Intel Corporation, 2014)

a good simulation performance this overhead may be too high. There are core models available that essentially map the instructions of the simulated core on native instructions of the host CPU. This eliminates the dynamic translation step and significantly speeds up the simulation.

- Interconnect models allow the evaluation of different interconnect topologies, be it buses or networks on chip (NoCs). For example, it is easy to try out different arbitration schemes. Bottlenecks can be identified and resolved with good profiling, debug, and tracing features. The TLM 2.0 standard, however, was primarily targeted at bus modeling and does not support NoCs as naturally. Moreover, the resolution of bus transactions is inherently limited due to abstraction: a common resolution is to have 4-phase transactions in the TLM 2.0 approximately timed (AT) style.[2]

Top Level Routing

The top level routing, which establishes the overall topology, can be done in several ways:

- First, it is entirely possible to do the wiring textually using plain SystemC. While this can be laborious the benefit is that the resulting system description can be run in the SystemC simulator[13] immediately without any other tool in between.
- Second, it is possible to do the routing in a commercial platform assembly tool graphically. The individual components must first be imported to the model library of the tool so that it has knowledge of the ports and their properties. If the wiring it not too complex then this approach can be more intuitive and clearer than the textual approach. Prior to an actual simulation the tool must generate SystemC code from the routing information. In order to avoid discrepancies this generated code should not be put under version control because it is merely derived from an original tool-specific description. Hence, the platform assembly tool must be integrated into the simulation setup process and a check for potential routing updates has to be hooked into the build system.
- The section "Formal Specification of Interfaces and Code Generation" discussed why and how register descriptions for TLM and RTL are derived from a single source. By the same argument it is possible to describe the top level routing formally and to generate both the TLM and the RTL routing from that.

Software Execution

Already at the component level there was a generic CPU executing software in order to generate stimuli for testing (see the earlier section "Test Case Infrastructure"). At the platform level, there are usually detailed models of the actual target cores. These cores run software images specifically compiled for the target architecture. Again, software is separated from the hardware model and is compiled into a dynamic library, which is loaded at runtime. In a first step, the test cases from the component level can be reused as integration tests. After successful component integration the cores execute software that implements actual platform functionality. Depending on the modeling level of

"Depending on the modeling level of detail the execution times can well reach 1/20th of the actual hardware speed."

detail the execution times can well reach 1/20th of the actual hardware speed. Software bugs can be chased down with the usual target debuggers.

Application Examples

"Since abstract TLM components are two to three orders of magnitude faster than RTL they readily allow the simulation of complex test cases on the entire system."

This section covers examples that demonstrate application scenarios for virtual prototypes. Since abstract TLM components are two to three orders of magnitude faster than RTL they readily allow the simulation of complex test cases on the entire system. This supports different application scenarios as discussed in the following subsections:

- Check of correct system functionality through assertions, signal traces, event logs, or program traces
- Quantitative analyses through inspection of statistics collected by TLM components such as CPU cores, buses, or custom hardware
- Hardware/software co-debugging
- Early software development for project pull-in

Fast Hardware Simulation

"Using a validated VP as reference, the behavior of the corresponding RTL implementation can be verified in a TLM/RTL co-simulation."

Especially for interactive development, short simulation times are essential. Figure 10 conceptually shows a cellular 2G transmitter. Due to a dedicated hardware focus, a very detailed transaction-level model was written, featuring a cycle- and bit-true signal processing chain. The functionality was checked by means of timing and signal analysis of the processed input symbols. A typical performance simulation covered one 3.5 ms GSM frame. On the RTL model a simulation took half an hour, whereas the transaction-level simulation was done after 35 seconds, which is a speedup factor of 50. Using this validated VP as reference, the behavior of the corresponding RTL implementation was verified in a TLM/RTL co-simulation according to Figure 8.

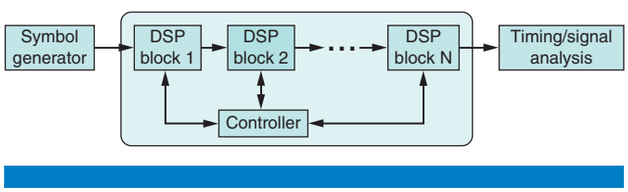


Figure 10: GSM transmit chain
(Source: Intel Corporation, 2014)

Combination of Cores

"The trend towards integration of formerly separate chips provides new potential for architecture optimization."

The trend towards integration of formerly separate chips provides new potential for architecture optimization. In the course of an integration project the plan was to replace two legacy CPU cores (one for 3.5G signal processing and one for the protocol stack) by a single core. Both signal processing and protocol stack were executed on a virtual prototype with a model of the new core, shown in Figure 11. Even a naïve addition of partial loads ignoring cache pollution and other higher-order effects led to a 113-percent overload estimation.

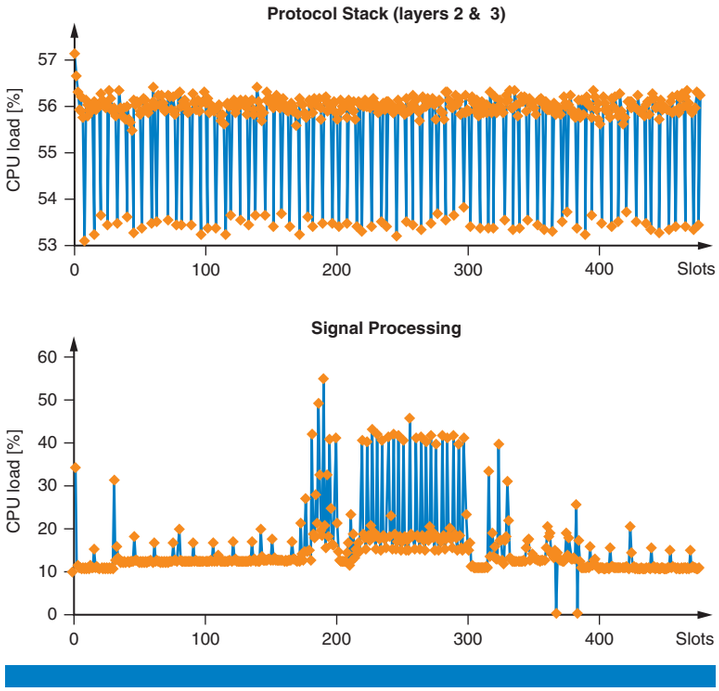


Figure 11: Analysis of CPU loads from legacy cores (Source: Intel Corporation, 2014)

Hardware/Software Co-Debugging

One of the prominent applications of virtual prototypes is the debugging of the hardware/software interaction while there is still time to correct hardware errors before tapeout.

Figure 12 shows how software and hardware components need to interact in searching a cell to camp on. It is evident that the system simulation includes scenarios that were not covered in the component tests: All the detected problems, as indicated by the dotted vertical arrows, relate to component interaction, mostly between software and hardware.

"One of the prominent applications of virtual prototypes is the debugging of the hardware/software interaction while there is still time to correct hardware errors before tapeout."

"…system simulation includes scenarios that were not covered in the component tests…"

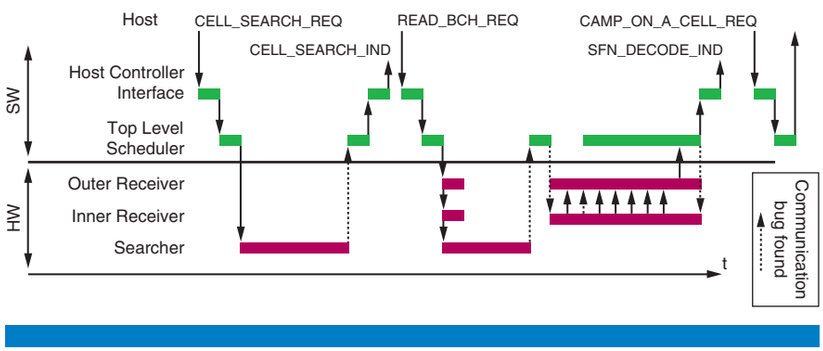


Figure 12: Hardware/software co-simulation (Source: Intel Corporation, 2014)

"With virtual prototyping we saw project schedules accelerated by up to five months."

Early Software Development

Virtual prototypes enable software development without the need for actual target hardware. In comparison to the traditional sequential development flow where software is developed exclusively on hardware samples, we saw project schedules accelerated by up to five months.

The following briefly sketches some application scenarios:

- Core upgrade: The decision was made to do a disruptive update of the application processor in one of our mobile phone platforms. It took only a week to update the VP. After three months the entire software including OS, protocol stack, and drivers, was up and running and a voice call could be simulated. Code freeze was thus achieved three months before silicon availability.

"The hardware bring-up of one of our phone platforms was done in one and a half days after sample availability."

- The hardware bring-up of another mobile phone platform was done in one and a half days after sample availability. This was possible because all the required software and test cases were ready and tested when the hardware entered the lab and there were no severe hardware bugs.
- Every new mobile communication generation is far more complex than the previous one. Moreover, the old infrastructure remains in operation for a long time and still needs to be supported. This gives rise to complex 2G/3G/4G handover scenarios that are expected to work without call drops and glitches. We have all communication generations combined in a single virtual prototype to shorten the learning phase in the field.

Summary

"The usage of virtual prototypes for the development of mobile platforms can no longer be seen as "nice to have." Instead, it is a key enabler."

The usage of virtual prototypes for the development of mobile platform Systems on Chip (SoCs) can no longer be seen as "nice to have." Instead it is a key enabler to handle the ever-growing system complexity and the tough requirements on the system design cycles in a rapidly changing wireless communication domain. Since virtual prototype development is on its way into the mainstream EDA, we need standardized design processes and workflows to efficiently provide virtual prototypes with short development time, predictable quality, and minimal costs. In this article we describe the infrastructure and methods implemented in the Intel system level design flow that allow us to achieve this goal.

We use a single source description of the memory map, the component interfaces, and the routing to efficiently and consistently generate code required in the different steps executed at system and RT level. With powerful TLM libraries we can minimize the manually written code of TLM components, support debugging, and enforce a standard coding style. A TLM component can be seen as executable specification, and to validate this specification we provide a TLM test bench infrastructure with a generic TLM bus model and a generic host-based CPU emulator. With the TLM components pre-verified in this way, the platform integration can be started and initial releases of the

virtual prototypes can already be done. To ensure the consistency between the TLM and RTL description, we reuse common test benches for RTL and TLM.

Around ten major virtual prototypes for wireless platforms have been developed with the flow described above. These VPs supported the engineers during

"Around ten major virtual prototypes for wireless platforms have been developed with this flow."

- architecture exploration including performance and power simulation
- software development and verification
- system verification
- hardware implementation

References

[1] IEEE Standard 1666, "SystemC Language Reference Manual," IEEE Computer Society, 2006.

[2] OSCI Standard, "OSCI TLM-2.0 Language Reference Manual," *Open SystemC Initiative*, 2009.

[3] Ghenassia, Frank, (Ed.) *Transaction-Level Modeling with SystemC: TLM Concepts and Applications for Embedded Systems* (Dordrecht, The Netherlands: Springer, 2005).

[4] Accellera SystemC Configuration, Control & Inspection (CCI) Working Group http://www.accellera.org/activities/committees/systemc-cci/

[5] Accellera SystemC Language Working Group (LWG) http://www.accellera.org/activities/committees/systemc-language/

[6] Accellera SystemC Synthesis Working Group (SWG) http://www.accellera.org/activities/committees/systemc-synthesis/

[7] Accellera SystemC Verification Working Group (VWG) http://www.accellera.org/activities/committees/systemc-verification

[8] Stehr, Guido and Josef Eckmüller, "Transaction Level Modeling in Practice: Motivation and Introduction," ICCAD 2010.

[9] Ecker, Wolfgang, Wolfgang Müller, and Rainer Dörmer (Eds.), *Hardware-dependent Software: Principle and Practice* (Springer, 2009).

[10] http://www.accellera.org/activities/committees/ip-xact

[11] http://www.makotemplates.org

[12] Bailey, Brian, Grant Martin and Andrew Piziali, *ESL Design and Verification* (Morgan Kaufmann, 2007)

[13] Accellera Systems Initiative: Download SystemC http://www.accellera.org/downloads/standards/systemc

[14] Esen, Volkan, "A New Assertion Language Covering Multiple Levels of Abstraction," PhD thesis, Technische Universität München (TUM), 2008.

[15] Bergeron, Janick, *Writing Testbenches Using SystemVerilog* (New York: Springer, 2006).

Author Biographies

Guido Stehr is a senior staff engineer in the System Design Methodology team at Intel Mobile Communications. He received a degree in Electrical Engineering (Dipl-Ing.) from the University of Karlsruhe in 1999. After obtaining a Dr.-Ing. degree in electrical engineering from the University of Technology in Munich (TUM) in 2005 he joined Infineon Technologies and worked as a system engineer before he shifted his focus to system-level design methodology in 2007. He had a leading role in establishing the TLM-based methodology in the RF transceiver development and in its dissemination after he joined Intel Mobile Communications in 2011. Email: guido.stehr@intel.com

Josef Eckmüller received his doctorate from the Institute of Computer-Aided Design at the University of Technology, Munich in 1998. From his start in the semiconductor industry 1997 until 2003, he was involved in the methodology development for the design and verification of analog and mixed-signal integrated circuits. Since 2003 he has been working in the system methodology and flow development for complex wireless communication SoCs. He joined Intel Mobile Communication in 2011 where he is responsible for the development, maintenance, and support of the system-level design flow, which includes the transaction-level modeling and virtual prototyping methodologies. Email: josef.eckmueller@intel.com

Oliver Bell joined Intel Mobile Communications in March 2011. Oliver received his Dipl.-Ing. engineering degree in microelectronics from the University of Applied Sciences Nuremberg (Germany). Since his start in the semiconductor industry in 1995, he has been involved in the development and application of advanced pre-silicon verification methodologies for complex SoC in communication and consumer applications. Oliver coordinates the system level and functional verification methodology and is involved in adopting virtual prototyping techniques throughout Intel. He also serves as Vice Chair of the Design and Verification Conference (DVCon) Europe. Email: oliver.bell@intel.com

Thomas Wilde is a senior engineer methodology development at Intel Mobile Communications. He received his degree (Dipl.-Ing.) in Electrical Engineering from University of Technology in Chemnitz and joined Siemens in 1996. For more than nine years he has been developing system-level design methodologies with a focus on SystemC-based transaction level modeling. He was involved in

the development of the TLM2 standard and contributed to the European SoC research project SPRINT. Email: thomas.wilde@intel.com

Ulrich Nageldinger earned his Dr.-Ing. degree in Computer Science from the Technical University of Kaiserslautern. He has been involved in single-source / code generation methodology for eight years, starting at Infineon Technologies and carrying on after joining Intel Mobile Communications in 2011. He currently coordinates code generation methodology in Intel Mobile Communications and is involved in common single-source efforts throughout Intel. Email: ulrich.nageldinger@intel.com

THE VIENNA MIMO TESTBED: EVALUATION OF FUTURE MOBILE COMMUNICATION TECHNIQUES

Contributors

Martin Lerch
Vienna University of Technology

Sebastian Caban
Vienna University of Technology

Martin Mayer
Vienna University of Technology

Markus Rupp
Vienna University of Technology

In order to evaluate current and upcoming mobile communications standards and to investigate new transmission as well as receiver techniques in a real-world environment, a very flexible testbed was set up at the Vienna University of Technology, comprised of multiple base stations, each equipped with several antennas. After providing an overview of this testbed and its capabilities, different kinds of measurements and their underlying methodologies are described in the context of 3GPP Long Term Evolution (LTE) transmissions. These are, on the one hand, point-to-point LTE Multiple-Input Multiple-Output (MIMO) throughput measurements, employing a single base station and, on the other hand, modern interference alignment measurements, utilizing up to three base stations simultaneously.

Introduction

The decades after Marconi's invention were filled with wireless experiments. Although we understand many physical phenomena of wireless propagations today much better than in the past, the channel models we use still capture only a part of the complex physical process. Nevertheless, in the last two decades, it has become a common method to entirely skip experimental validation and trust existing channel models when designing mobile communication systems. As the complexity of mobile communication standards also increases, simulation methods appear to be the Holy Grail to solve open design questions. While these methods deliver quantitative results in acceptable time, many important issues are simplified or not modeled at all, trading off timely results for accuracy. Converting new algorithmic ideas into hardware on the other hand is quite time consuming and often lacks flexibility so that experimental evaluation remains no longer an attractive choice. We show that with our testbed approach, we essentially combine the advantages of both worlds: design flexibility and timeliness under true physical conditions.

"As the complexity of mobile communication standards also increases, simulation methods appear to be the Holy Grail to solve open design questions."

"We show that with our testbed approach, we essentially combine the advantages of both worlds: design flexibility and timeliness under true physical conditions."

In this article, we explain briefly our testbed approach[1] and our optimized measurement methodology that allows the deducing of results based on a minimal number of sampling points in the following section, "The Vienna MIMO Testbed: A Marriage of Hardware and Software." Experimental examples based on 3GPP Long Term Evolution (LTE) are then provided in the section "LTE Measurements," and finally experimental results for the interference alignment (IA) transmission scheme[2] are presented in the section "Interference Alignment Measurements." The article briefly sums up the findings with a "Conclusion" section.

The Vienna MIMO Testbed: A Marriage of Hardware and Software

The Vienna MIMO testbed consists of various hardware components that couple data generating and capturing PCs, radio frequency front ends, antennas, and a suite of software tools, the so-called Vienna LTE Simulators.[3][4][5] While the LTE simulators help in designing optimal algorithms, the same signals can be fed into the testbed in order to transmit them over the air. The captured data are then input to the receiver part of the simulators and can be evaluated offline later on. It is important to note that the Vienna MIMO testbed is not limited to LTE transmissions. The LTE simulator can be easily replaced by software implementations of any desired communication system that meets the constraints of the testbed.

"While the LTE simulators help in designing optimal algorithms, the same signals can be fed into the testbed in order to transmit them over the air."

Hardware

Figure 1 exhibits the main hardware components required to convert *a priori* generated data into electromagnetic waves, transmitting them over the air and finally capturing them before storing them in digital form for further evaluation. The major hardware components are:

- Three rooftop transmitters supporting four antennas each. The digital signal samples are converted with a precision of 16 bits and are transmitted with adjustable power within a continuous range of about −35 dBm to 35 dBm per antenna.
- One indoor receiver with four channels that converts the received signals with a precision of 16 bits before the raw signal samples are saved to hard disk. The receive antennas are mounted on a positioning table, which allows for measurements at different positions within an area of about 1 m × 1 m.
- The carrier frequency, the sample clock, and the trigger signals are generated separately at each station utilizing GPS synchronized rubidium frequency standards. The synchronization of the triggers is based on exchanging timestamps in the form of UDP packets over a trigger network.[6] The precision of this trigger mechanism does not require any further

"The carrier frequency, the sample clock, and the trigger signals are generated separately at each station utilizing GPS synchronized rubidium frequency standards."

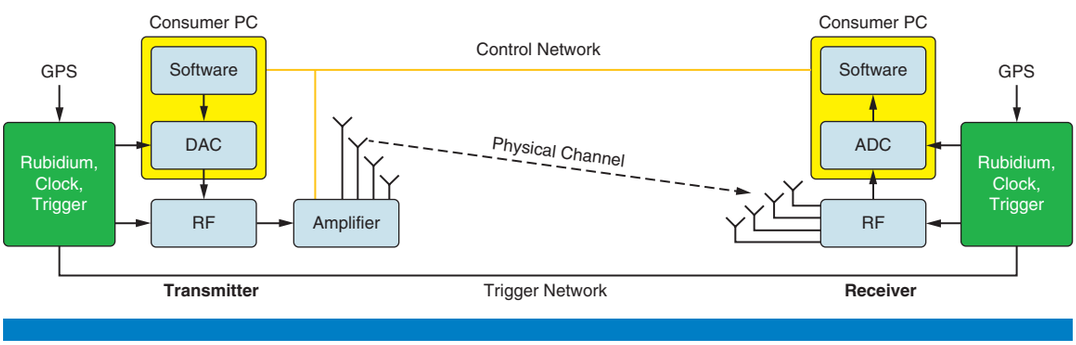


Figure 1: Testbed setup showing GPS-controlled rubidium clocks at both ends of the transmission chain (Source: Vienna University of Technology, 2014)

post-synchronization at the receiver. It is sufficient to measure the delay once and time-shift all signals according to the measured delay.

- A dedicated fiber-optic network is utilized to exchange synchronization commands as well as feedback information and general control commands.

The current setup supports a transmission bandwidth of up to 20 MHz at a center frequency of 2.503 GHz.

LTE Simulators

Along with the hardware setup, a suite of software-based simulators are employed. Currently we support:

- The Vienna LTE-A Downlink Link Level Simulator (DL-LL)
- The Vienna LTE Uplink Link Level Simulator (UL-LL)
- The Vienna LTE Downlink System Level Simulator (DL-SL)

"Our simulators together with published results were released under an open license agreement, free of charge for academic research."

Our simulators together with published results were released under an open license agreement, free of charge for academic research. Over the years, many thousand users have formed a Web-based exchange forum where open problems were posed and solutions discussed. Due to these efforts in reproducibility, our simulators not only increased in functionality but also gained substantial quality. Many companies are now also using the simulators because they offer a convenient platform to exchange results between partners. The DL-SL simulator is only listed here to provide a complete list; this simulator uses link-level abstractions and supports simulations with hundreds of users since such complexity would not lead to acceptable run times in link-level precision. Both link-level simulators can be used to generate inputs to the testbed and testbed outputs can be fed back into them, hence providing an LTE-compliant transmission chain whose data can be directed to the transmit antennas and captured at the receive antennas instead of running transmissions over simulated channels such as ITU or Winner. Although the transmission only allows a burst mode, we can continuously generate such data bursts and mimic accurately continuous transmissions. The received symbols are time-stamped and can later be fed back into the simulator chain for evaluation. By this we can directly compare simulations with measured results based on identical transmit data and identical receiver algorithms, allowing very rigorous research results.

Measurement Methodology

"As a rule of thumb, in a typical scenario approximately 30 measurements of different receive antenna positions are necessary to get sound results for an LTE signal with a bandwidth of 10 MHz."

In typical measurements, the transmission of desired signals, or rather signals generated according to parameters of interest, is repeated with different values of transmit power in order to obtain results for a certain range of receive signal-to-noise ratios (SNR). Furthermore, the transmission of such signals at all values of transmit power is repeated at different receive antenna positions in order to average over small-scale fading scenarios. As a rule of thumb, in a typical scenario approximately 30 measurements of different receive antenna positions are necessary to get sound results for an LTE signal with a bandwidth of 10 MHz. In order to check whether we have measured enough channel realizations, we always

include BCa bootstrap confidence intervals in our results (see the following section and Caban et al.[7]). While this process is usually the same for different kinds of measurements, they may differ in the way transmit signals are generated.

As illustrated in Figure 2, two different methodologies are utilized as detailed in the following:

- *Brute force measurements:* All signals of interest are pre-generated, transmitted over the physical channel, and saved as raw signal samples to hard disk. The received signals are then evaluated offline. This approach is only feasible as long as the time duration of all the different transmit signals is small compared to the channel variations so that successively transmitted data sets appear to be transmitted over the same channel.
- *Measurements with feedback:* The transmit signals are generated on the fly utilizing channel state information obtained via a preceding transmission of training symbols. While the processing and evaluation of the actual data symbols can be computed offline, the demodulation of the training symbols, evaluation, and decision about the generation of the next transmit signal has to be performed in (quasi-) real time.

"This approach is only feasible as long as the time duration of all the different transmit signals is small compared to the channel variations so that successively transmitted data sets appear to be transmitted over the same channel."

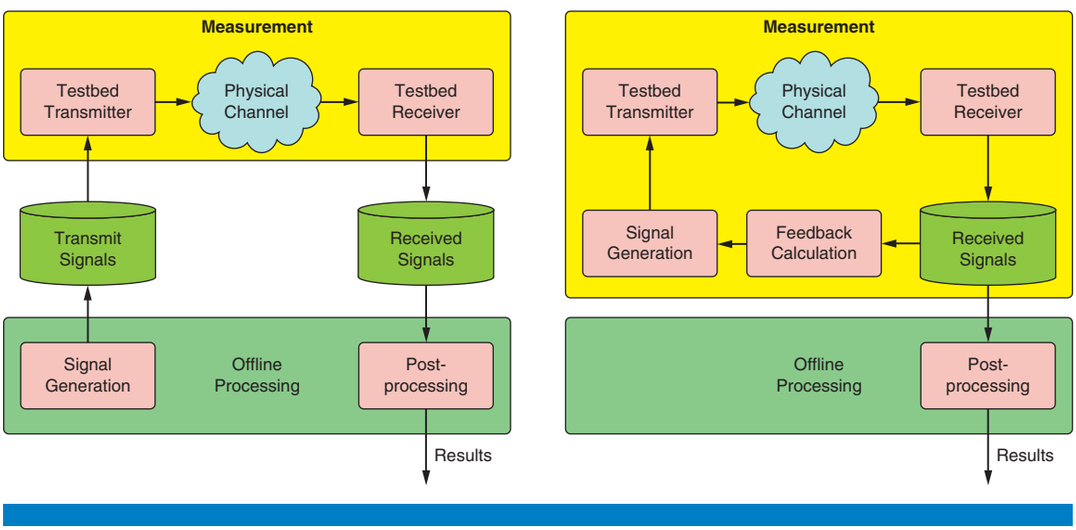


Figure 2: Measurement process without (left) and with feedback (right) (Source: Vienna University of Technology, 2014)

While brute force measurements typically take longer than the feedback approach and the number of different signals that need to be evaluated is much higher, results obtained by brute force measurements are typically more detailed and are certainly not contaminated by the quality of the feedback function. If the number of different transmit signals is not too large, a combination of both methodologies is possible. All signals of interest are pre-generated, but only those a feedback function decides for are transmitted. This approach reduces the number of signals that have to be evaluated and signals do not have to be generated during the measurement. Nevertheless, it should be noted that if the number of possible

transmit signals is rather large or infinite (for example, zero-forcing Multi-User MIMO mode), only a feedback approach is feasible.

LTE Measurements

"While LTE cellular systems are already being rolled out and operated in many countries around the world, there are still unresolved issues in transmission technology."

While LTE cellular systems are already being rolled out and operated in many countries around the world, there are still unresolved issues in transmission technology. Focusing on point-to-point single-user LTE transmissions, there exist many open questions that can be best tackled by LTE measurements:

- Comparison of different kinds of receivers (receiver algorithms)
- Performance of novel and modified transmission schemes following the LTE standard
- Performance measurements at extreme channels (for example, very high speed) for which channel models are very crude or even nonexistent
- Comparison of different penetration scenarios or different antenna configurations

In the following, we present two measurements comparing on the one hand two different transmit antenna configurations and two different scenarios on the other hand. For both measurements the brute force approach using the DL-LL simulator as software implementation of the base station and the user equipment was used. We chose the open loop spatial multiplexing transmission mode where no feedback of the preferred precoder is performed. Thus the number of different transmit signals is small enough to apply the brute force approach.

A Comparison of Different Antenna Configurations

We were interested in evaluating the influence of the transmit antenna configuration on the performance of the LTE MIMO downlink. Possible configurations for the case of two transmit antennas are cross-polarized antennas, as they are used in today's base station antennas, and equally polarized antennas. For the measurements presented below, we utilized an off-the-shelf double cross-polarized sector antenna (Kathrein 800 10543) whose four antenna elements were used for implementing both a cross-polarized antenna pair and two equally polarized antennas with a spacing of 1.24 wavelengths (see legend of Figure 3).

"…spatial multiplexing outperforms single-stream transmission only above a certain average SNR we observe that a cross-polarized configuration outperforms an equally polarized configuration."

Figure 3 shows the measured throughput of the LTE open-loop downlink. In the left plot the results are shown over measured average SNR for a fixed transmission rank (1 or 2) where for single-stream transmission (rank 1) both antenna configurations performed similarly. The results for two spatial streams (rank 2) show on the one hand that spatial multiplexing outperforms single-stream transmission only above a certain average SNR. On the other hand, we observe that a cross-polarized configuration outperforms an equally polarized configuration. In the right plot, the throughput was maximized over the number of spatial streams for every channel realization, resembling a feedback selection scheme. More details on measurements comparing different vertical and horizontal setups for 2x2 as well as for 4x4 transmissions are described by Lerch and Rupp.[8]

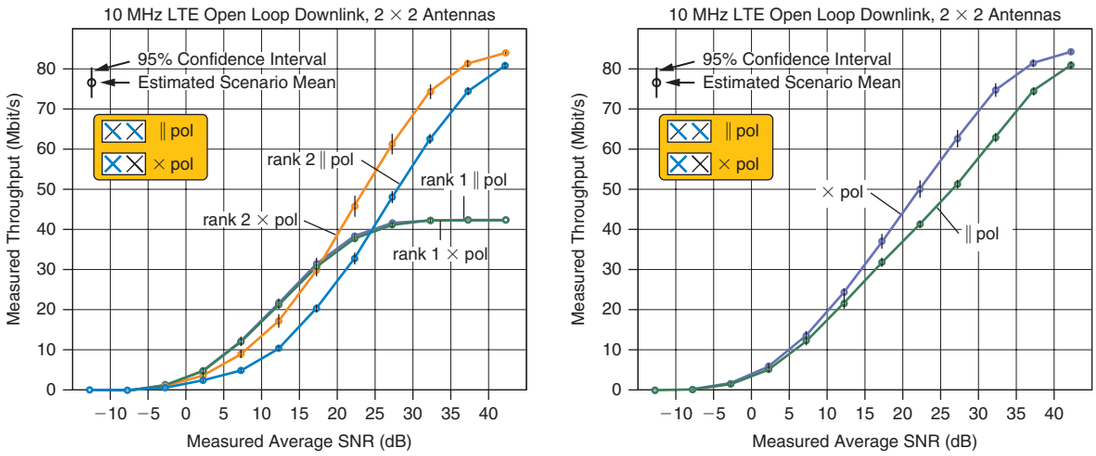


Figure 3: Comparison of cross-polarized and equally polarized transmit antennas in terms of LTE throughput (Source: Vienna University of Technology, 2014)

A Comparison of Different Scenarios

While the previous results were obtained in a certain scenario given by the location of the transmitter and the receiver, we were further interested in a comparison of different scenarios. Therefore the measurement was repeated placing a transmitter at a different location and keeping the receiver location. In the previous scenario there was no line-of-sight path between the transmitter and the receiver. We considered a second scenario with a significant line-of-sight path. These two scenarios will be referred to as Non-Light-of-Sight (NLOS) and Line-of-Sight (LOS).

The left plot of Figure 4 shows the results considering two transmit antennas. While the cross-polarized antennas perform similarly in both scenarios, an even worse performance is obtained when considering equally polarized antennas in the LOS scenario, although only in the higher SNR regions where data is transmitted over two spatial streams. At lower SNRs, where only a single data stream is transmitted, the performance is quite independent of the scenario and the transmit antenna configuration used. Finally, the right plot of Figure 4 shows a comparison of both scenarios considering four transmit antennas. The results are similar to those for two antennas. While in the lower SNR regions the difference is negligible; at higher SNRs, where data is transmitted over multiple spatial streams, the performance in the NLOS scenario is better than in the LOS scenario.

"While in the lower SNR regions the difference is negligible; at higher SNRs, where data is transmitted over multiple spatial streams, the performance in the NLOS scenario is better than in the LOS scenario."

Interference Alignment Measurements

Interference alignment (IA)[2] is an example of a system setup that utilizes all three transmitters simultaneously with feedback. IA measurements fully exploit all capabilities of the Vienna MIMO testbed.

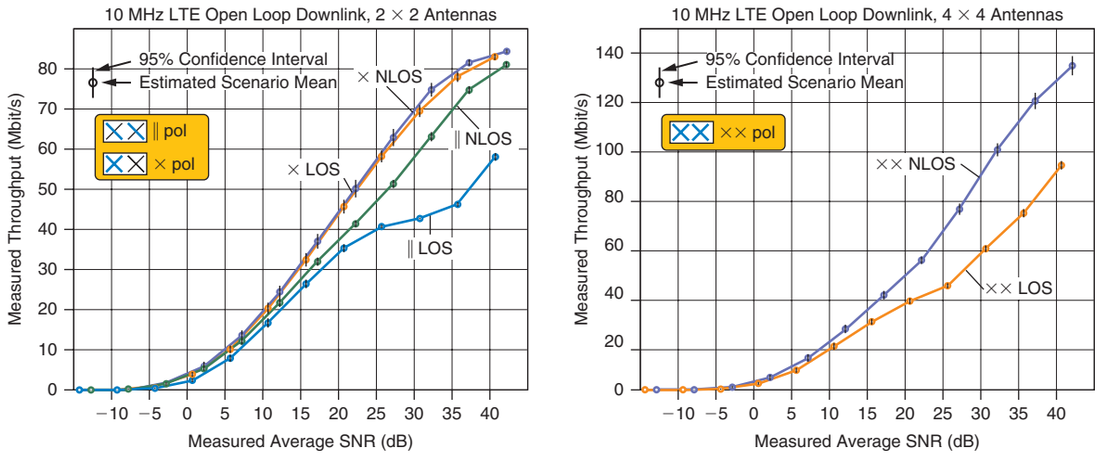


Figure 4: LTE throughput measurement results comparing LOS and NLOS scenarios for two and four transmit antennas.
(Source: Vienna University of Technology, 2014)

"All three users experience heavy interference from the other base stations approximately as strong as the desired downlink signal."

"In the ideal case, half of the capacity of the interference-free case with four data streams can be achieved, which is more than what is obtained by using resource sharing."

The Concept of Interference Alignment

Consider the cellular scenario shown in Figure 5 where three mobile users at the cell edges wish to communicate with a different base station. All three users experience heavy interference from the other base stations approximately as strong as the desired downlink signal. An emerging MIMO technique to cope with such a scenario is called *interference alignment.* Based on the knowledge of the channels between every user and every base station, a joint calculation of precoding matrices, V_i i = 1, 2, and 3 and receive filters, U_i i = 1, 2, and 3, is performed, by which half of the degrees of freedom of the MIMO channels are utilized for data transmission while the other half are exploited to align the interferences at the receiver. By applying the receive filters, interferences are eliminated and only the data signal of interest is retained. Thus, instead of transmitting the maximum number of four data streams over a 4x4 MIMO channel, only two data streams are transmitted. In the ideal case, half of the

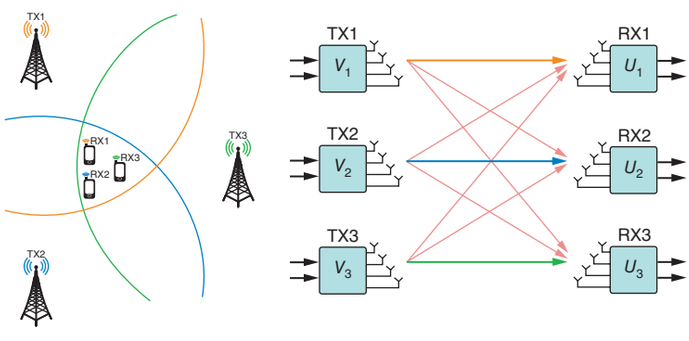


Figure 5: Interference alignment scenario and setup
(Source: Vienna University of Technology, 2014)

capacity of the interference-free case with four data streams can be achieved, which is more than what is obtained by using resource sharing. (Orthogonal multiple access schemes like time division multiple access can only offer a third of the capacity for each of the three users).

Measurements

Measurements with the Vienna MIMO testbed were carried out to show the feasibility of IA in a real-world scenario and to provide a basis for further investigations on the impact of different kinds of practical issues on the performance of IA. In order to evaluate IA on our testbed, we implemented the ideas below:

- While all three base stations are needed in order to transmit at the same time, one receiver that can act as any of the three receivers is indeed sufficient. The other two receivers can be virtual receivers and the respective channels can either be generated randomly or can be results of past channel measurements.
- Every mobile user has to estimate the channels to all transmitters. Therefore, not only must the training symbols from different antennas of each base station be orthogonal among themselves, but the training symbols of all different base stations must also be orthogonal. In the case of three base stations, each having four antennas, the channels from twelve transmit antennas have to be estimated simultaneously. Thus, standard compliant LTE, whose pilot structure only supports up to four transmit antennas, is not applicable, and therefore we implemented our own transmission scheme described below.
- Each transmit frame consists of a pilot preamble followed by the data payload. The pilot preamble is constructed to estimate the channels to all transmit antennas simultaneously and the data payload is designed to estimate the mutual information of the transmission. The precoders applied to the data payload are based on the channel estimates obtained from the respectively previous transmission. In order to keep the time between the channel measurement and the application of the respective precoders short (less than 20 ms), we took only a single subcarrier into account.

"IA requires all involved transmitters and receivers to be synchronous in terms of carrier frequency and time."

IA requires all involved transmitters and receivers to be synchronous in terms of carrier frequency and time. On our testbed, rubidium frequency standards at every station combined with a GPS based trigger network provide the required synchronicity. In order to receive the signals from different transmitters perfectly synchronous at the receiver, the transmit signals are time shifted according to the delays between the respective transmitter and the receiver. For a more detailed discussion of IA measurements on the Vienna MIMO testbed, the reader is referred to Mayer et al.[9]

Results

The left plot of Figure 6 shows the mutual information results of a measurement for a fixed signal-to-interference ratio (SIR) of −3 dB, that is, the signals from all three base stations are received equally strongly at the receivers. In order to compare the performance of IA to a non-cooperative scheme, a full rank

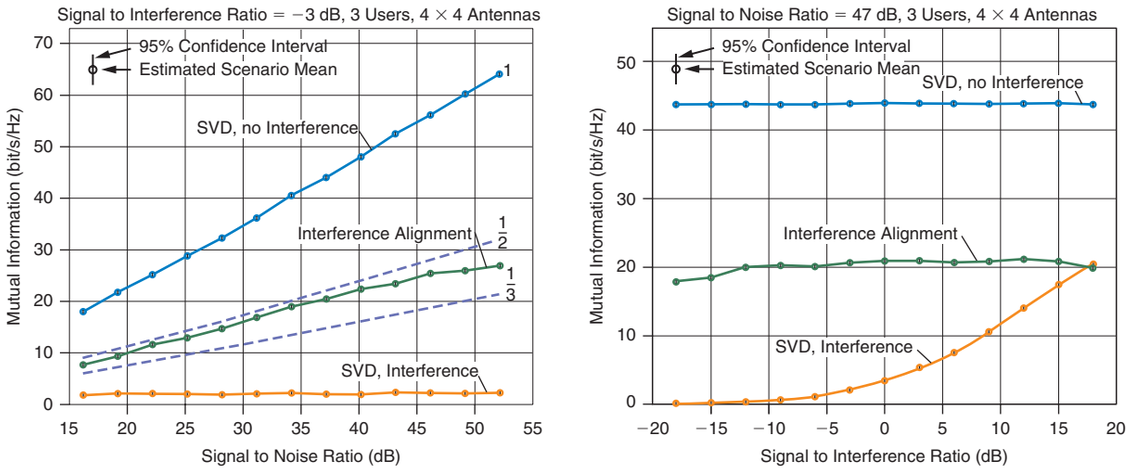


Figure 6: Performance of IA for a fixed SIR (left) and a fixed SNR (right). (Source: Vienna University of Technology, 2014)

"IA reaches, almost half of the mutual information of the interference-free full rank case IA is very sensitive to channel estimation errors…"

transmission based on the Singular Value Decomposition (SVD) of the channel utilizing all four available data streams was measured in the same scenario. In the case of interference, the performance of the SVD transmission (orange curve) is solely determined by the rather low SIR and therefore quite constant over the observed SNR region. IA (green curve) reaches, following the theory (dashed blue curve), almost half of the mutual information of the interference-free full rank case (blue continuous line). At high SNR, a saturation of mutual information is observed in the IA case. This is due to the fact that IA is very sensitive to channel estimation errors as mentioned by Garcia-Naya et al.[10] The precoders are always computed from the channel estimate of the previous transmission, and since our feedback is only finitely fast, the channels have time to change between transmissions. Thus, despite the high SNR, there will always be residual interference power due to imperfect alignment that limits mutual information, as long as the channel is not perfectly static.

"With increasing SIR (decreasing interference) the performance of the non-cooperative full rank transmission increases and outperforms IA…"

The results for a fixed signal-to-noise ratio of 47 dB are shown on the right plot of Figure 6. With increasing SIR (decreasing interference) the performance of the non-cooperative full rank transmission increases and outperforms IA at a certain level of SIR. Above this level, a mobile user should rather be scheduled for a different transmission scheme than IA.

Conclusion

The article describes a testbed methodology that combines the rapid development speed of software with the precise measurements results including physical wireless channels.

The capability of our testbed to measure over a wide range of transmit power within the same scenario allows for deep insights into the performance of

modern mobile communication systems. As an example, we demonstrated in terms of LTE throughput that the performance of MIMO techniques does not only depend on the signal-to-noise ratio. It also depends on the actual transmit antenna configuration and the scenario. Our measurement based evaluation of interference alignment provides viable information regarding its possible fields of application and the entailed constraints. The inherent precoder feedback delay and the extensive channel knowledge requirement narrow the field down to applications that comprise reliable feedback and coordination. Fully exploiting the capabilities of the Vienna MIMO Testbed, interference alignment was shown to be feasible, achieving results close to theory and outperforming orthogonal access schemes in case of intermediate to strong interference at fairly high SNR.

"…the performance of MIMO techniques does not only depend on the signal-to-noise ratio. It also depends on the actual transmit antenna configuration and the scenario."

References

[1] Caban, S., C. Mehlführer, M. Rupp, and M. Wrulich, *Evaluation of HSDPA and LTE: From Testbed Measurements to System Level Performance* (New York: John Wiley & Sons, 2012), p. 404.

[2] Cadambe, V. R. and S. A. Jafar, "Interference Alignment and Degrees of Freedom of the K-User Interference Channel," *IEEE Transactions on Information Theory,* vol. 54, no. 8, pp. 3425–3441, 2008.

[3] Mehlführer, C., J. Colom Ikuno, M. Simko, S. Schwarz, M. Wrulich, and M. Rupp, "The Vienna LTE Simulators - Enabling Reproducibility in Wireless Communications Research," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 1–13, 2011.

[4] Schwarz, S., J. Colom Ikuno, M. Simko, M. Taranetz, Q. Wang, and M. Rupp, "Pushing the Limits of LTE: A Survey on Research Enhancing the Standard," *IEEE Access,* vol. 1, pp. 51–62, 2013.

[5] "The Vienna LTE simulators," Institute of telecommunications, Vienna University of Technology, [Online]. Available: www.nt.tuwien.ac.at/ltesimulator.

[6] Caban, S., A. Disslbacher-Fink, J. A. Garcia Naya, and M. Rupp, "Synchronization of Wireless Radio Testbed Measurements," in *IEEE International Instrumentation and Measurement Technology Conference,* Binjiang, 2011.

[7] Caban, S., J. A. Garcia Naya, and M. Rupp, "Measuring the Physical Layer Performance of Wireless Communication Systems," *IEEE Instrumentation & Measurement Magazine, v*ol. 14, pp. 8–17, 2011.

[8] Lerch, M. and M. Rupp, "Measurement-Based Evaluation of the LTE MIMO Downlink at Different Antenna Configurations," in *17th International ITG Workshop on Smart Antennas*, Stuttgart, 2013.

[9] Mayer, M., G. Artner, G. Hannak, M. Lerch, and M. Guillaud, "Measurement Based Evaluation of Interference Alignment on the Vienna MIMO Testbed," in *The Tenth International Symposium on Wireless Communication Systems*, Ilmenau, 2013.

[10] Garcia-Naya, J. A., L. Castedo, O. Gonzalez, I. Ramirez, and I. Santamaria, "Experimental evaluation of Interference Alignment under imperfect channel state information," in *9th European Signal Processing Conference (EUSIPCO),* Barcelona, Spain, 2011.

Acknowledgements

The authors would like to thank the LTE research group and in particular Prof. Christoph Mecklenbräuker for continuous support. This work has been funded by the Christian Doppler Laboratory for Wireless Technologies for Sustainable Mobility, KATHREIN Werke KG, and A1 Telekom Austria AG. The financial support by the Austrian Federal Ministry of Economy, Family and Youth and the National Foundation for Research, Technology and Development is gratefully acknowledged.

Author Biographies

Martin Lerch studied electrical and communication engineering at the Vienna University of Technology where he received his master's degree in 2008. During this time and later, Martin developed several database, Web, and desktop applications before he returned to the Vienna University of Technology in 2011 to work on the Vienna MIMO testbed. Martin's work focuses on the development of new measurement methodologies for static and high mobility mobile communication scenarios. Contact him at mlerch@nt.tuwien.ac.at

Sebastian Caban works as a consultant for operational excellence next to being a postdoctoral researcher at the Vienna University of Technology. He holds a BSc, MSc, and PhD in Telecommunication Engineering as well as a BBA and MBA in business administration. His research focuses on developing measurement methodologies and building testbeds to quantify the *actual* performance of wireless communication systems. He can be contacted at scaban@nt.tuwien.ac.at

Martin Mayer is a research assistant at the Vienna University of Technology. He holds a BSc degree in Electrical Engineering and Information Technology and an MSc degree in Telecommunication Engineering, both received at the Vienna University of Technology in 2011 and 2013, respectively. For his master's degree,

his research focused on evaluating interference alignment on the Vienna MIMO testbed. In 2014, his field of interest shifted to advanced signal processing in RFID. Martin can be contacted at mmayer@nt.tuwien.ac.at

Markus Rupp holds the chair for signal processing in mobile communications at the Vienna University of Technology. His research is devoted to digital wireless communication systems with a focus on cellular communications but also addressing near field communications as well as traffic modelling. He can be contacted at mrupp@nt.tuwien.ac.at

OPERATOR GRADE WI-FI* AS A COMPLEMENTARY ACCESS MEDIA IN THE LTE ERA

Contributors

Gideon Prat
Wireless Platforms R&D
Intel Corporation

Penny Efraim-Sagi
Wireless Platforms R&D
Intel Corporation

Sharon Ben Porath
Wireless Platforms R&D
Intel Corporation

"The demand for being always connected in the best way possible, and the continuous growth of data consumption by mobile users, create a huge challenge for cellular operators."

This article discusses the cellular operators' need for offloading mobile traffic to Wi-Fi in order to deal with capacity crunch. It discusses the variety of standard solutions for achieving this "Wi-Fi offload." It further describes the fundamental differences between Wi-Fi networks and the cellular networks, the challenges resulting from these differences, and the existing solutions for overcoming them. The focus of the article is on Wi-Fi as a potential complementary access media technology serving the operator, with the ability to provide operator-grade service to the end users.

Introduction and Problem Statement

The demand for being always connected in the best way possible, and the continuous growth of data consumption by mobile users, create a huge challenge for cellular operators. In the past few years, mobile usages evolved far beyond Internet browsing and basic media consumption to include many services that consume significant amounts of data. These new mobile usages include (see Figure 1): full HD video and movie streaming, highly intensive and dynamic multiplayer gaming, always-aware personal assistance, video calls, augmented and virtual reality, and other image analysis applications. Such usages are driving demand for faster access to off-device (cloud) content, contiguous coverage, higher bandwidth, better responsiveness, and higher network capacity. The demand for more bandwidth greatly outpaces cellular network technology evolution.

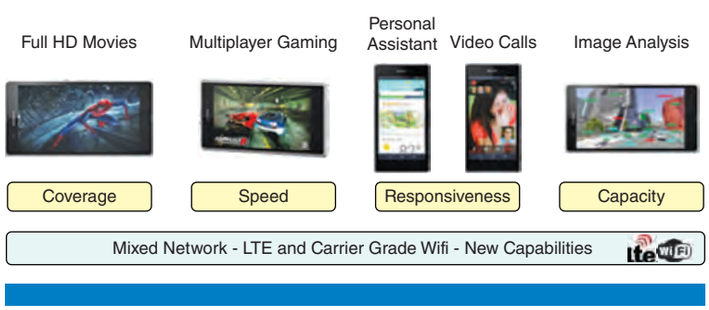


Figure 1: Mobile usages and required capabilities (Source: Intel Corporation, 2013)

The industry is exploring several directions for dealing with the above challenges and providing both short-term and long-term solutions for the capacity problem. One key method for operators to tackle the problem is their constant search for more spectrum and bandwidth. Some operators choose to increase the cell density in existing bands by using overlay network

deployments, including small cells, nano-cells and femto-cells. Others are exploring the use of extremely high frequencies, like 40 GHz and 60 GHz, as well as other bands, which are being released by the regulatory bodies across the globe (such as analog TV spectrum refarming). In parallel there are initiatives for introduction of unlicensed spectrum for the use of the operators, like an unlicensed version of Long Term Evolution, called LTE-U. All those can serve as solutions for the operators to deal with their networks data congestion. This is one of the essential problems we can expect the future cellular standards to be trying to deal with.

Offload to Wi-Fi, or Wireless LAN (WLAN), is the natural first choice. The operators' customer's devices already support Wi-Fi and they use it for more or less the same applications. Wi-Fi deployment continues to grow. According to a recent report by market research company MarketsandMarkets[1], for example, the outdoor Wi-Fi market is expected to grow from USD 15.41 billion in 2013 to USD 37.2 billion in 2018. The ecosystem is already in place.

"Offload to Wi-Fi, or Wireless LAN (WLAN), is the natural first choice. The operators' customer's devices already support Wi-Fi and they use it for more or less the same applications."

The cellular community (operators, standardization committees, and so on) has significantly changed their approach towards Wi-Fi since the early 2000s, when Wi-Fi was the "enemy," or at least irrelevant, and is now looking to use Wi-Fi to offload data traffic. Residences and workplaces account for a large majority of mobile device data consumption. These locations are traditionally dominated by "primitive" Wi-Fi offload of most data on mobile devices. A variety of reasons make these locations more suitable for Wi-Fi offload. In such locations Wi-Fi network performance is usually good with high availability and accessibility. The Wi-Fi network is usually free of charge, secured, and in many cases allows access to local devices (such as NAS or wireless printers). The use of Wi-Fi connectivity is typically also more power efficient and enables longer battery life.

Different studies show the trends of offloading mobile traffic from cellular networks to Wi-Fi networks and foresee an exponential growth. A recent forecast example is taken from Cisco VNI Forecast Update.[2] According to this forecast, by 2017, 46 percent of the mobile data traffic will be offloaded.

Wi-Fi offload can be made available and usable in a much wider variety of scenarios and locations by adding "smartness." Original equipment manufacturers (OEMs), OS vendors (OSVs), and independent software vendors (ISVs) are investing in improving the smartness of the connection management and the resulting user experience. Most of the mobile traffic offload to Wi-Fi is achieved today in "primitive" offload, or per 3GPP (Third Generation Partnership Project) nomenclature: Non Seamless WLAN Offload (NSWO).[3] Different degrees of nonstandard enhancements could be used to improve user experience. In this article we focus on the challenges that operators are facing when trying to connect Wi-Fi networks to their core networks in order to achieve more than is available with NSWO.

"Most of the mobile traffic offload to Wi-Fi is achieved today in "primitive" offload, or per 3GPP (Third Generation Partnership Project) nomenclature: Non Seamless WLAN Offload (NSWO)."

Cellular operators' target for Wi-Fi offload is to create transparent background capabilities that can kick in according to predefined policies. Of particular interest is offloading to Wi-Fi whenever the cellular mobile network is

overloaded, especially in dense areas like shopping centers, or to fill gaps in noncontiguous LTE coverage, such as in modern dense urban areas and indoors.

This article discusses key aspects in deployment of mixed networks of LTE and Operator Grade Wi-Fi, for supporting new capabilities and use cases, which can be divided into three main groups as follows (illustrated in Figure 2):

- Smart connection selection and seamless mobility
- Application based radio selection: traffic from different applications (running concurrently) routed to different radio connections
- Link aggregation: taking advantage of the aggregated bandwidth provided by the different radios for higher throughputs and lower latencies

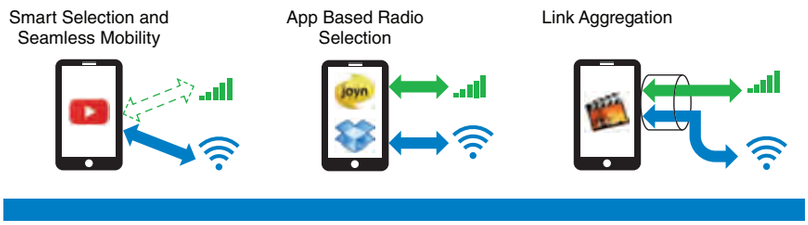


Figure 2: Main types of use cases for Wi-Fi offload (Source: Intel Corporation, 2013)

Wi-Fi offloading presents many challenges for operators. Operators are accustomed to design their network as trusted, controllable, and reliable resources. These networks are centralized systems where access to channels, security, capacity, and quality of service (QoS) are controlled. Legacy, 2G/3G networks are very limited in the QoS they provide, where the majority of data traffic is passed in a nonprioritized "best effort" manner. LTE networks have introduced finer QoS granularity and can provide high quality of service for data traffic. Voice over IP (VoIP) traffic, for example, can be provided over an LTE network, in an equivalent user experience to legacy (circuit-switch) cellular voice call services.

"Operators are expecting Wi-Fi networks to have similar properties to those available in cellular networks, in order for them to consider Wi-Fi as a true alternative or augmentation method to the cellular networks."

Operators are expecting Wi-Fi networks to have similar properties to those available in cellular networks, in order for them to consider Wi-Fi as a true alternative or augmentation method to the cellular networks. A priori, though, Wi-Fi is just the opposite. Wi-Fi operates in unlicensed ISM bands, which are intended for free access. Anyone can create a network by setting up a Wi-Fi access point on any channel. Out-of-the-box, Wi-Fi networks are distributed, unmanaged, and have limited support for quality of service. The Wi-Fi access point owners are responsible for setting their security level. Moreover, Wi-Fi is now increasingly used for device-to-device communication applications, without any access point involved.

In practice, though, Wi-Fi is maturing as well. The drive to replace Wired LAN (Ethernet) has created solutions, especially for enterprise usages, that place Wi-Fi on the right path for serving operators. Today we can find examples for QoS

delivery and for centralized control of authorizations. One example for such QoS enhancement for Wi-Fi is the Wi-Fi Alliance (WFA) "Voice-Enterprise"[4] certification program that relies on IEEE 802.11r[5] and 802.11k[6] amendments.

The WFA*, the GSMA (Global Standard for Mobile communication Association) and mainly the 3GPP are working to build a set of specifications, adopting those solutions and complementing them with Wi-Fi-cellular interworking specifications. There is also much room for vendor creativity beyond the implementation of the mechanisms defined in the standards.

This article discusses some of the challenges and opportunities related to the usage of Wi-Fi by cellular operators.

It is expected that with growing deployments of Wi-Fi and availability of enhanced Wi-Fi to cellular interworking, operators will find it easier to trust Wi-Fi networks for applications that require reliability, security, and QoS. It is therefore expected that operators will start offloading plain Internet traffic to Wi-Fi and eventually move up to offloading telephony services.

"It is expected that with growing deployments of Wi-Fi and availability of enhanced Wi-Fi to cellular interworking, operators will find it easier to trust Wi-Fi networks for applications that require reliability, security, and QoS."

Challenges and Solutions for Operator Grade Wi-Fi

In this section we explain some of the challenges in more detail, as seen from the cellular operator's perspective, as well as from the end-user point of view.

We start with a general simplified description of the cellular network, as defined in the 3GPP standard, focusing on the fourth generation LTE (Long Term Evolution). Figure 3 provides a high-level view.

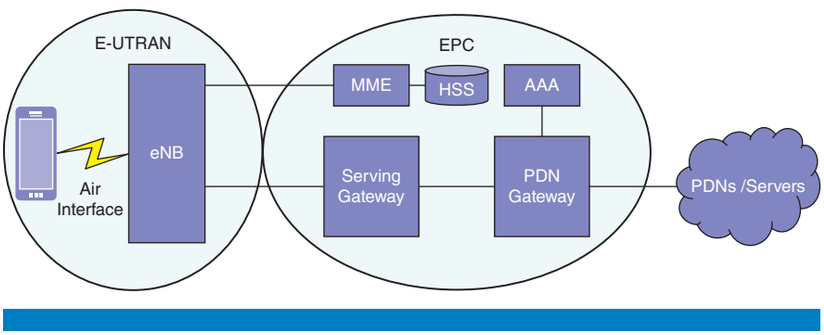


Figure 3: Basic LTE network block diagram
(Source: Intel Corporation, 2013)

Cellular networks are composed of the air interface, the connection between the user mobile device and the cellular base station, and the cellular operator data network, the latter being an interconnection of routers, gateways, and servers, owned by the operator and its partners. The air interface is called Radio Access Network (RAN) in 3G, or Evolved UMTS Terrestrial RAN (E-UTRAN) in LTE. The base station in LTE is called evolved Node B (eNB). The Data network is called Packet Core for 3G, or Evolved Packet Core (EPC) for LTE.

The Core network may be regarded as the operator's "private network," deployed and maintained by the operator. It includes the entities that manage the network access permission, such as the authenticating, authorizing, and accounting (AAA) server, entities that manage the user device mobility, including management of handover between different cells and radios, such as the Mobility Management Engine (MME), and databases that contain user-related and subscriber-related information, such as the Home Subscriber Server (HSS). The access to the Internet is handled by gateways in the core network like the Serving Gateway facing the RAN and the Packet Data Network (PDN) Gateway facing the connection to the Internet.

"The 3GPP standard includes specifications for integration of "non-3GPP IP access network" to the cellular core network. This applies to other communication technologies like Wi-Fi. LTE introduced for the first time integration of nontrusted non-3GPP IP access networks."

The 3GPP standard includes specifications for integration of "non-3GPP IP access network" to the cellular core network.[7] This applies to other communication technologies like Wi-Fi. LTE introduced for the first time integration of nontrusted non-3GPP IP access networks.

The requirements for Wi-Fi to be "operator grade" can be divided into the elements listed below.

Seamless discovery and trusted connections establishment:

- Automatic discovery of valid networks based on user preferences and operator preferences
- Mobile device provision, providing the end-user equipment with the information required for identifying and connecting to operator's Wi-Fi access points.
- Authentication of the Wi-Fi connection with the operator's authorization mechanisms, enabling the user device to connect to the operator network via security provisioned access points

Trust and control of data traffic:

- Provision of basic information about connection capabilities and usage policies of the connection, specific to preferred access points, user account, and operator's services. Such policies may specify the applications that should be using the Wi-Fi connection and those that should not. For example, allow the use of Wi-Fi for Internet browsing, while the cellular network should be used for operator QoS services, like voice or video calls.
- Authentication of the mobile device and establishment of a trusted data connection with data encryption. This ensures the privacy of the data that flows through the operator network, as well as over Wi-Fi connections, in case of offloading.
- Emulation of the cellular network model over Wi-Fi, enabling multiple data links over multiple PDNs, as done over the cellular network.

Advanced solutions for seamlessness:

- Seamless cellular-Wi-Fi usage: mechanisms that allow use of Wi-Fi or cellular and switch between them seamlessly, from the user perspective.

- Optimization for dense or otherwise challenging environments and for QoS. For example, solutions for enabling fast response to changing network loads and/or radio environments.
- Optimization for telephony seamlessness when Wi-Fi is the only available connection, without cellular as a backup.
- Optimization for usages in heterogeneous (hetnet) small cells, combining Wi-Fi and cellular.

The WFA and 3GPP have defined solutions for the first two challenges. Solutions are being developed to address the third challenge, yet opinions still vary whether all related problems are completely solvable. Substantial opportunities for system-wide differentiated solutions may exist.

Cellular networks are centrally managed. The Packet Core deploys a set of management entities (servers and gateways) for controlling network access permissions, security, mobility, network load balancing, billing, and QoS. The purpose of this centralized control is to ensure that the mobile device and the end user employing it will get the best possible service from the operator. As the cellular networks evolve, the number of control parameters is constantly growing. This growth complicates the centralized control approach. For dealing with this complication, LTE E-UTRAN added a distributed method for eNB management.

We shall go through the different basic properties of the LTE cellular network and consider what is required in order to implement Wi-Fi offloading.

Seamless Discovery and Trusted Connections Establishment

Cellular mobile networks use SIM (Subscriber Identity Module) to identify the subscriber. The AAA server authenticates the device.

Wi-Fi networks have their own access control mechanisms, handled by the Wi-Fi access point (AP). In order to be granted access to the operator's network to enjoy the operator's services, the operator should be able to identify and authenticate the subscriber connected through Wi-Fi. This requires an additional (second-level) authentication mechanism on top of the native Wi-Fi one.

The WFA has adopted the Extensible Authentication Protocol (EAP) EAP-SIM and EAP-AKA[8] mechanisms defined in 3GPP to support a second level of authorization by the operator AAA server. Those methods rely on SIM-based credentials, for non-SIM devices other EAP-based methods, such as EAP-TLS[8], can be used.

The key challenge is to achieve a 3GPP-like user experience of the Wi-Fi authentication and network access control. Seamless discovery also helps addressing the noncontiguous nature of both cellular and Wi-Fi networks. A set of specifications addresses this challenge. Wireless Internet Service Provider roaming[9] (or WISPr) is a draft protocol submitted to the WFA for allowing roaming between wireless Internet service providers (the WISPr paper has been withdrawn due to a technicality, but the concept is accepted and is being adopted). An additional specification called Hotspot 2.0 (also commercially

"The key challenge is to achieve a 3GPP-like user experience of the Wi-Fi authentication and network access control. Seamless discovery also helps addressing the noncontiguous nature of both cellular and Wi-Fi networks."

known as "Passpoint")[10], driven by Intel and other partners, defines the mechanism for efficient Wi-Fi service discovery by the mobile devices, without full connection establishment. The 3GPP Access Network Discovery and Selection Function (ANDSF)[11] complements the solution toolset by defining how the mobile device is provided with information about Wi-Fi networks availability per geographical region. The rules for supporting the seamless discovery and connection establishment in ANDSF are referred to as Inter System Mobility Policy (ISMP). Another set of ANDSF policies is defined in Operator Policies for IP Interface Selection (OPIIS).

Note: Currently there are still missing elements in ANDSF to support all Wi-Fi offload models. However, it is safe to assume that those gaps will be closed in subsequent releases of the specification.

"Device and user authentication is a first step for allowing the mobile device access to the network. In order for the operator to be able to protect its core network, it requires additional mechanisms to ensure the data traffic passing through the core network can be trusted."

Trust and Control of Data Traffic

Device and user authentication is a first step for allowing the mobile device access to the network. In order for the operator to be able to protect its core network, it requires additional mechanisms to ensure the data traffic passing through the core network can be trusted. This desired trust level could be achieved through two main models:

- *Operator controlled Wi-Fi ("Trusted model"):* The operator adds to the core network a new entity called Trusted WLAN Access Gateway (TWAG), as a Wi-Fi access interface to the network, and another entity called Trusted WLAN AAA Proxy. Those two entities make up the Trusted WLAN Access Network (TWAN).[12] The operator deploys the Wi-Fi network by itself or by agreement with trusted Wi-Fi partners. This way the operator can set the desired security level for making the Wi-Fi network "trustable" and hooking it directly to the cellular core network. The S2a Mobility based on GTP (SaMOG)[13] work group in 3GPP defines the details for the Trusted model. Trusted Wi-Fi connection to EPC is done via an interface marked in the standard as *S2a* and the link to the trusted WLAN AAA Proxy is marked *STa*, as shown in Figure 4. The Trusted model is thus referred to sometimes as the S2a model.
- *Tunnel over noncontrolled Wi-Fi ("Untrusted model"):* The operator adds to the core network a new entity called evolved Packet Data Gateway (ePDG)[12] and the mobile device is required to open a secure tunnel (IPsec tunnel) to the ePDG using operator-provided credentials. The link between the ePDG and the EPC is marked in the standard as "S2b" and the untrusted Wi-Fi network link to the 3GPP AAA Server is marked *SWa*, as shown in Figure 4. The untrusted model is thus referred to sometimes as the *S2b* model.

In LTE networks several data links are typically established between the mobile device and the network, where usually each and every one of the operator's services requires a different link. For example, one link can serve for Voice over LTE (VoLTE)[14] calls, another for Multimedia Messaging Service (MMS)[15] and a third one for standard Internet traffic (Web browsing and so on). A separate Packet Data Network (PDN) on the Packet Core network side is used for managing each of those data links.

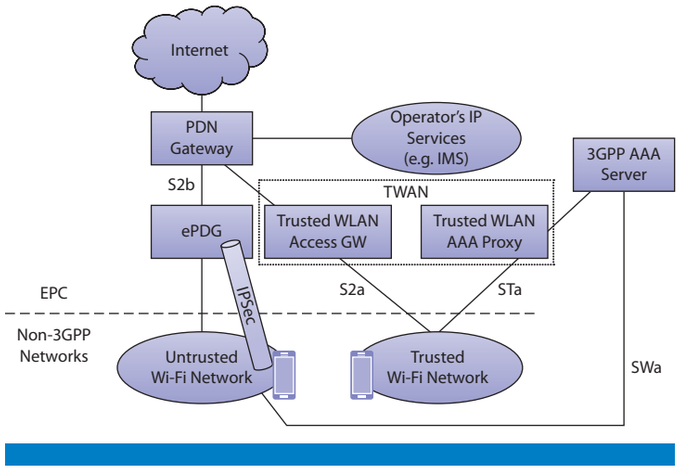


Figure 4: 3GPP EPC network architecture for supporting Wi-Fi offload
(Source: Intel Corporation, 2013)

Releases 10 and 11 of the 3GPP standard define handling of a single PDN offload over Wi-Fi, for both trusted and untrusted offloading. This can be used mainly for best effort service and Over The Top (OTT) Internet applications. Releases 12 and 13 of 3GPP add capabilities for handling the offload of multiple PDNs, in a similar manner to the way data links are handled in the LTE networks. This enhancement allows a voice call to be handed off from the LTE (VoLTE) to Voice over WLAN (VoWLAN or VoWifi), maintaining the required QoS, while in parallel best-effort traffic of OTT Internet applications can be transferred over a separate data link using a second PDN.

The ability to efficiently support such advanced use cases requires good control over both the cellular modem and the Wi-Fi client, providing full platform vendors an advantage.

In line with the centralized network approach, in order to deal with Wi-Fi offload, the operator should define a set of rules and policies, determining when and what type of traffic to offload onto the Wi-Fi network. The rules can be static, dynamic, or a combination of both types. An example for a static rule may be to always keep operator-provided services on the cellular network, namely IP Multimedia Sub-system (IMS)[14][15] traffic, and not to offload this type of traffic to Wi-Fi under any condition. An example for a dynamic rule may be to offload a specific IP flow when Wi-Fi link conditions reach a predefined threshold. As operators set the rules, they should be able to easily configure them over the air (OTA) and they should be OS and hardware agnostic.

As explained above, the ANDSF defines tools for provisioning mobile devices with operator's policies. ANDSF defines policies covering static rules, as well as simple thresholds for dynamic behavior, like data throughput. Definition of more advanced dynamic rules is still in process in 3GPP.

"Releases 10 and 11 of the 3GPP standard define handling of a single PDN offload over Wi-Fi, for both trusted and untrusted offloading. This can be used mainly for best effort service and Over The Top (OTT) Internet applications. Releases 12 and 13 of 3GPP add capabilities for handling the offload of multiple PDNs, in a similar manner to the way data links are handled in the LTE networks."

"Ideally, offload solutions would provide users with a seamless experience for all types of applications and services used on their devices. The mobile device should make intelligent decisions about keeping data flows on preferred networks (for example, keep some traffic, such as VoIP, on LTE even when Wi-Fi is available), while considering resource optimization, such as battery life."

Advanced Solutions for Seamlessness

Ideally, offload solutions would provide users with a seamless experience for all types of applications and services used on their devices. The mobile device should make intelligent decisions about keeping data flows on preferred networks (for example, keep some traffic, such as VoIP, on LTE even when Wi-Fi is available), while considering resource optimization, such as battery life.

Handover procedure is a fundamentally important procedure in the cellular mobile network, covering the mobile device as well as the network behavior, when the mobile device switches from one eNB to another, from one section of the network to another, but also from one radio technology to another (for example, from cellular network to Wi-Fi). The 3GPP standard defines different mechanisms for supporting handover between 3GPP and non-3GPP IP networks, like Wi-Fi.[16] Those mechanisms can be classified into two basic types:

- *Handover without IP preservation*. In this process the mobile device goes through an authentication and IP address allocation in the Wi-Fi network in the background while connected to the cellular network. The existing IP connection (TCP and UDP) can either continue on the existing 3GPP connection or switch to the Wi-Fi network. When switching, those connections must be reestablished using the new IP address.
- *Handover with IP preservation*. When switching to a Wi-Fi network, the same IP address that was used in the cellular network is allocated to the mobile device for use in the Wi-Fi network. It means that the IP connections (TCP and UDP) can be kept without interruption.

Basic handover without IP preservation can always be established. This is what practically happens on mobile devices in which no special mechanisms are used, and the connection switches between cellular network and Wi-Fi network due to user action (switching on Wi-Fi), roaming into an area with Wi-Fi coverage, or the like. There are proprietary mechanisms that can be used in order to provide seamless service and user experience, in spite of the IP address switch and connection break and reestablishment. Some software vendors are implementing such mechanisms (especially for client-server solutions that can handle the end-to-end connectivity). It can also be performed in software (middleware) that takes care of the seamlessness for unaware and unchanged applications, but such solutions should be tailored to each application (or application type) to be supported.

The 3GPP standard defines several mechanisms that enable IP preservation. Proxy Mobile IP (PMIPv6)[17] based handover takes care of IP address anchoring in the network side, which means the mobile device keeps using the same IP address it was allocated by the cellular network, and a network side entity is taking care of the IP address conversion. Although in this case the network handles the IP address preservation, there is still a need for some effort

from the mobile device side in order to support a seamless experience. This includes the proper use of two TCP/IP stack network interfaces, one for each of the two physical interfaces, with the ability to switch application sockets between the network interfaces.

In the case of untrusted model Wi-Fi, the link to the cellular network is established through an IPSec tunnel that goes with a dedicated IP address. The IP preservation can be achieved in this case, if the tunnel is maintained when the mobile device hands over between the cellular and the Wi-Fi networks. It is possible, but is implementation dependent.

The standard also defines a mechanism that relies on a client-based solution, implementing Dual-Stack Mobile IP (DSMIPv6)[18] protocol on the mobile device. Such an implementation on the mobile device side is heavy and is not likely to receive wide market acceptance.

ANDSF includes a set of rules for supporting mobile devices that can activate more than one connection in parallel, for seamless handover and traffic routing purposes. This set of rules is referred to as Inter System Routing Policy (ISRP). Additional relevant policies are defined in ANDSF under Data Identification (DIDA). WLAN Network Selection (WLAN_NS)[19] is another related 3GPP Release 12 Work Item, which defines the required Wi-Fi information (beyond SSID) for improving the selection process in complex scenarios like Wi-Fi roaming, multiple visitor public mobile networks, and more.

Implementation Considerations

Where to implement the Wi-Fi offload solution is not a trivial question. Traditionally Wi-Fi is connected to the application processor (host CPU) and managed by the operating system (OS), while all layers of the cellular protocol stack are embedded into the cellular modem solution and controlled by the cellular network. The offload solution is in between these two worlds, thus creating space for a variety of solutions.

Cellular operators would naturally prefer to have an embedded solution similar to a cellular modem solution, in order to increase the predictability and consistency of the solution and to achieve control over the decisions (network access, radio attach/detach and selection, traffic routing and so on). An embedded solution can also ease the certification process. In this way operators can force similar behavior for different mobile devices through certification requirements. On the other hand changes in the OS level, adding new communication features, or merely introducing changes to the OS architecture or implementation might change the embedded solution playing field. For example if the OS establishes concurrent connection on Wi-Fi and cellular like a certain leading smartphone OEM has recently done, adding in the OS level Multipath TCP (MPTCP)[20] for Voice assistance service, it is actually an offload solution in the application layer with no operator influence.

"Where to implement the Wi-Fi offload solution is not a trivial question. Traditionally Wi-Fi is connected to the application processor (host CPU) and managed by the operating system (OS), while all layers of the cellular protocol stack are embedded into the cellular modem solution and controlled by the cellular network. The offload solution is in between these two worlds, thus creating space for a variety of solutions."

"Operating system vendors (OSVs) and cellular operators might potentially pull in two different directions. An OSV, focused on the user experience, may desire to include within the OS the Wi-Fi offloading functionality, like network selection, network connection, and traffic routing. Cellular operators may desire such functionality to be embedded in the cellular modem and Wi-Fi client in a standard certifiable manner."

Operating system vendors (OSVs) and cellular operators might potentially pull in two different directions. An OSV, focused on the user experience, may desire to include within the OS the Wi-Fi offloading functionality, like network selection, network connection, and traffic routing. Cellular operators may desire such functionality to be embedded in the cellular modem and Wi-Fi client in a standard certifiable manner. Such embedding of the functionality would allow interoperability testing as part of the 3GPP and cellular operator certification process. Implementation of Wi-Fi offload solutions in an OS environment would cause undesired interdependency between OSV and cellular operator. It would be unacceptable for operators, as well as for mobile device vendors or OEMs, to redo certification of basic network features with each OS release. It is thus natural to expect operators to be looking for OS-independent solutions for the mobile devices, while the OSVs continue upgrading the connection management. One can imagine the OS handling native Wi-Fi usages and user-level connection management, while the operator policies and control of the Wi-Fi offloading capabilities is done in an embedded manner, isolated from the OS. The challenge for Intel and other mobile device platform and wireless communication vendors is achieving solutions that implement the desired characteristics of Wi-Fi offloading, outside the OS environment, in a way that complements the OS-level connection management yet brings the best user experience as well as OS independence. Intel is in an especially good position to seize this opportunity as a chip vendor for both Wi-Fi and cellular modems, as well as a provider of full mobile platforms.

Achieving a Practical Functional Parity with Wi-Fi

The safe bet for most operators is to use Wi-Fi as Internet data offload. Many of them intend to channel most Internet traffic outside the EPC anyway; they use a "local break-out" at the edge of the Packet Core network. Offloading Internet traffic to Wi-Fi leaves more bandwidth for operator services, especially voice and video telephony. The requirements for best-effort Internet traffic without guaranteed QoS are limited. Trust of connection and of data access with some level of policy management is sufficient for such cases.

Other operators are ready to be at the leading edge of Wi-Fi offload, implementing the upcoming specifications enabling multiple Packet Data Networks (PDNs) and IP preservation for QoS applications. They target either offloading all traffic or prioritizing offload of QoS services, like IMS. The motivation of those operators may be to solve the problems created in very dense environments with transient loads—stadiums, shopping centers, airports, train stations, and so on. Even before full solutions may be available, some type of best-effort Wi-Fi may be preferred over a potential total break of service.

In Figure 5, we have made an effort to summarize the pragmatic steps that can be taken in order to achieve an increasingly better experience through the use of the different Wi-Fi offload mechanisms described in this article.

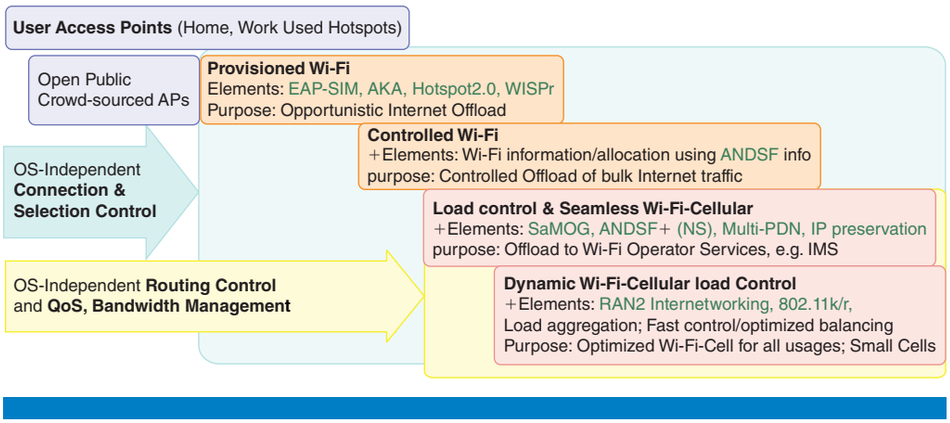


Figure 5: Steps in Wi-Fi offloading deployment
(Source: Intel Corporation, 2013)

The basic steps described in Figure 5, though not easily implemented in the field, are most likely to be insufficient to guarantee Wi-Fi as a reliable solution for all applications. This is especially true for QoS services, with telephony being the highest priority service.

The problem can be divided into two levels:

- *Wi-Fi as a cellular complement*: For most operators, the cellular mobile network is planned to remain the anchor for QoS.
- *Wi-Fi as a cellular replacement*: For cases when a cellular connection is unavailable, even in a localized area.

To ensure seamlessness in the first case (Wi-Fi complementing cellular), the cellular community recognizes that Wi-Fi-cellular handover decisions need to be very dynamic and based on real-time cellular and Wi-Fi quality metrics. Neither cellular nor Wi-Fi, by their nature as wireless networks, can offer constant, fully robust performance for the mobile user. However, real-time tracking of the conditions of both networks may allow the combined Wi-Fi/cellular system to evolve into a much more robust overall solution for the mobile user. The handover decision may also need to apply to granularities finer than Packet Data Networks (PDNs), possibly per individual IP flows or even splitting of flows (load aggregation). Combining time dynamics with fine granularity of allocation and application-specific load management has the potential to guarantee Wi-Fi plus cellular QoS delivery in almost all cases.

"Combining time dynamics with fine granularity of allocation and application-specific load management has the potential to guarantee Wi-Fi plus cellular QoS delivery in almost all cases."

Some operators are specifically interested in Wi-Fi solutions for small cell deployments, including heterogeneous networks (HetNets). Although the problems are the same for small cells and macro cells, the dynamics of small cells are quantifiably different, possibly leading to different optimizations and solutions.

For the second case (Wi-Fi as cellular replacement), guaranteeing QoS over Wi-Fi so that it is equivalent to cellular is a harder task. Without a cellular

network as a backup for extreme cases, Wi-Fi and usage of Wi-Fi bands may need to be much more strongly policed and managed.

These challenges are the next frontier in dealing with Wi-Fi for operators and are the subject of much work in the 3GPP, WFA, and other standardization bodies.

Summary

There is a clear need today for offloading traffic from cellular mobile networks to the widely available Wi-Fi networks.

"…critical differences between Wi-Fi networks and cellular mobile networks must also be bridged in order for Wi-Fi offload solutions to become efficient and useful for the operators and for the end users."

Wi-Fi networks include many of the required characteristics for this purpose; however, critical differences between Wi-Fi networks and cellular mobile networks must also be bridged in order for Wi-Fi offload solutions to become efficient and useful for the operators and for the end users.

We have made an effort in this article to scan the various problems operators and equipment vendors are facing in order to meet the challenge of Wi-Fi offload implementation. We reviewed many of the mechanisms already defined in the different standards and tried to propose a pragmatic step approach for deploying Wi-Fi offloading capabilities. Finally, we made an attempt to describe the next frontier that awaits us after we manage to solve the basic challenges reviewed—how to improve Wi-Fi and its usage in order to achieve a real parity with cellular and get a true complementary or even replacement access media to the cellular radio.

In Figure 6 we map the different mechanisms described in the article onto the three types of use cases for Wi-Fi offload that we have described in the beginning of the article.

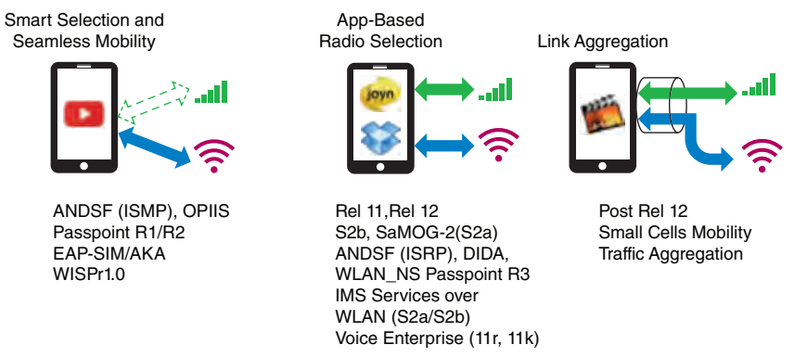


Figure 6: Steps in Wi-Fi offloading deployment (Source: Intel Corporation, 2013)

What keeps us excited and confident in the future is the belief that these challenges bring great opportunities for wireless communication vendors, especially for vendors like Intel that can provide a full mobile platform solution, with room for innovation, differentiation, and better user experience, while providing better margins for operators and OEMs.

References

[1] Markets and Markets market research company, Outdoor Wi-Fi Market: Global Advancements, Business Models, Worldwide Market Forecasts and Analysis (2013–2018), http://www.marketsandmarkets.com/Market-Reports/outdoor-wi-fi-market-945.html.

[2] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html.

[3] Architecture enhancements for non-3GPP accesses, Release 12,3GPP TS 23.402.

[4] Wi-Fi CERTIFIED™ Voice-Enterprise: Delivering Wi-Fi® voice to the enterprise, 2012.

[5] IEEE Std 802.11r™-2008: Fast Basic Service Set (BSS) Transition (Amendment 2).

[6] IEEE Std 802.11k™-2008: Radio Resource Measurement of Wireless LANs (Amendment 1).

[7] Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP access networks; Stage 3, Release 12, 3GPP TS 24.302.

[8] Extensible Authentication Protocol, or EAP, defined in IETF RFC 3748, 2004. EAP-SIM defined in RFC 4186, 2006, EAP-AKA in RFC 4187, 2006 and EAP-TLS in RFC 5216, 2008.

[9] WiSPR or Wireless Internet Service Provider roaming is a draft protocol submitted to the Wi-Fi Alliance. The WISPr 2.0 specification was published by the Wireless Broadband Alliance in March 2010.

[10] Hotspot 2.0 Specification, Phase 1; Wi-Fi Alliance Technical Committee Hotspot 2.0 Task Group, March 21, 2012.

[11] Access Network Discovery and Selection Function (ANDSF) Management Object (MO), (Release 12), 3GPP TS 24.312.

[12] 3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U) (Release 11). 3GPP TS 29.281

[13] Study on S2a Mobility based on GPRS Tunnelling Protocol (GTP) and Wireless Local Area Network (WLAN) access to the Enhanced Packet Core (EPC) network (SaMOG); Stage 2. 3GPP TR 23.852.

[14] IMS Profile for Voice and SMS; Version 7.0, 03 March 2013.

[15] Protocols for Advanced Networking (TISPAN); Support of SMS and MMS over NGN IMS subsystem; Stage 3 (Release 8), 3GPP TS 24.451.

[16] Architecture enhancements for non-3GPP accesses, Release 12, 3GPP TS 23.402.

[17] Proxy Mobile IPv6, IETF, August 2008.

[18] Mobile IPv6 Support for Dual Stack Hosts and Routers, IETF, June 2009.

[19] Study on Wireless Local Area Network (WLAN) network selection for 3GPP terminals; Stage 2 (Release 12), 3GPP TR 23.865.

[20] IETF TCP Extensions for Multipath Operation with Multiple Addresses, draft RFC 6824, 2014-01.

Author Biographies

Gideon Prat has been with Intel Corporation for 30 years. Today, Gideon is in the Intel Wireless Platforms R&D group. He worked on hardware design and testing of the early Intel Ethernet components and spent four years in various strategic planning and marketing roles. He headed the architecture of Intel LAN products, developing the first integrated LAN controller. Gideon is a senior architect and pathfinder in the area of wireless solutions, involved in several initiatives such as AMT, AOAC, WiDi and WiGig. Gideon holds a BSc EE degree from the Technion IIT, Haifa, Israel. Email: Gideon.prat@intel.com.

Penny Efraim-Sagi has more than 20 years of experience in standardizing wireless and cellular systems (Tetra, GSM, WB-CDMA, LTE, WiMAX) and proprietary system solutions, such as the LMDS system in 24 GHz. Penny was involved in Tetra and 3GPP standardization activities. Penny has held various positions as a system architect team leader, senior system architect, and projects manager. Penny has been working for five months at Intel and previously worked at Intel's cellular division from 2003 through 2007. She has a BSc EE degree from Ben Gurion University, Israel. Email: penny.efraim-sagi@intel.com.

Sharon Ben Porath is Wireless Systems Engineering Manager at Intel, with 20 years of practical experience in the development and bringing to market of different standard and proprietary wireless systems, including Wi-Fi, LTE, WiMAX, and multi-comm embedded systems at the defense industry, startup companies, and for the last 10 years, at Intel Corporation. Today, Sharon is responsible for wireless connectivity and multi-comm systems engineering in the Intel Wireless Platforms R&D group.

Sharon holds patents in multi-comm solutions, such as advanced collaborative coexistence schemes.

Sharon holds BSc EE and multidisciplinary ME degrees from the Technion IIT, Haifa, Israel and an MBA in Technology Management from Tel-Aviv University, Israel. Email: sharon.ben-porath@intel.com.

RF CHALLENGES OF LTE-ADVANCED

Contributor

Jan Whitacre
Agilent Technologies

With nearly 1.5 billion LTE subscriptions worldwide expected by 2018 according to ABI Research[1], mobile operators are scrambling to add the speed- and capacity-increasing features of LTE-Advanced to their almost-new LTE networks. Millions of smartphones, tablets, and other mobile devices are already devouring bandwidth on the networks.

The benefits of LTE-Advanced come at the cost of adding more complexity to an already complex wireless network environment. Developers of RF components and systems for network equipment and mobile devices have to deal with new architectures for carrier aggregation, 8x8 MIMO and other LTE-Advanced options, making the new technology work on multiple frequency bands and alongside other communication formats … all while maintaining or even improving the power efficiency of the previous generation of equipment.

"This article discusses the challenges faced by developers of the wireless devices that must operate in these systems along with the infrastructure they must interact with."

This article discusses the challenges faced by developers of the wireless devices that must operate in these systems along with the infrastructure they must interact with.

Introduction

At a high level, the challenges developers face for LTE-Advanced can be grouped into five categories:

- Designing user equipment (UEs) and base stations to operate in Het-Net environments
- Maximizing UE and BS power efficiency
- Managing the dramatic increase in the amount of design and verification effort
- Working on designs before conformance tests are fully defined
- Designing to pass operators' acceptance tests for end-user experience.

While none of these challenges is unique to LTE-Advanced, each intensifies the development effort, especially when added to the operators' higher data throughput, system capacity, and time-to-market needs. The remainder of this article touches on some of the underlying issues that have to be addressed for each of these challenges and techniques to overcome them.

With 1.5 billion LTE subscriptions worldwide expected by 2018 (ABI Research), one clear issue emerges for mobile operators: to find, given spectrum

scarcity, ways to increase speed and capacity for the LTE network. Without the upgrades, operators may not be able to deliver a reliable, consistent end-user experience if traffic loads continue to grow exponentially as predicted. The current usage demonstrates and reinforces this need.

"Without the upgrades, operators may not be able to deliver a reliable, consistent end-user experience if traffic loads continue to grow exponentially as predicted."

First on every operator's list of what to implement is carrier aggregation (CA), a feature that allows mobile operators to bundle diverse frequencies into a larger, single-channel bandwidth to achieve significantly higher data rates. This feature could be a game changer for those operators with limited spectrum and no new allotments on the horizon.

Other LTE-Advanced features that are being heavily investigated by operators include techniques for managing interference among large and small cells in heterogeneous networks (Het-Nets), and incorporation of higher order MIMO (multiple input, multiple output) antenna systems for higher data rates and better connections.

The benefits of LTE-Advanced come at the cost of adding more complexity to an already complex wireless network environment as shown in Figure 1. Functionality like carrier aggregation, 8x8 MIMO, and other LTE-Advanced options, making the new technology work on multiple frequency bands and alongside other communication formats increases the complexity of the implementation for the different engineering groups responsible to develop the new components. Besides this increase of complexity, R&D engineers needs to balance the power efficiency and keep the new more complex system as efficient as the previous equipment generation.

"Besides this increase of complexity, R&D engineers needs to balance the power efficiency and keep the new more complex system as efficient as the previous equipment generation."

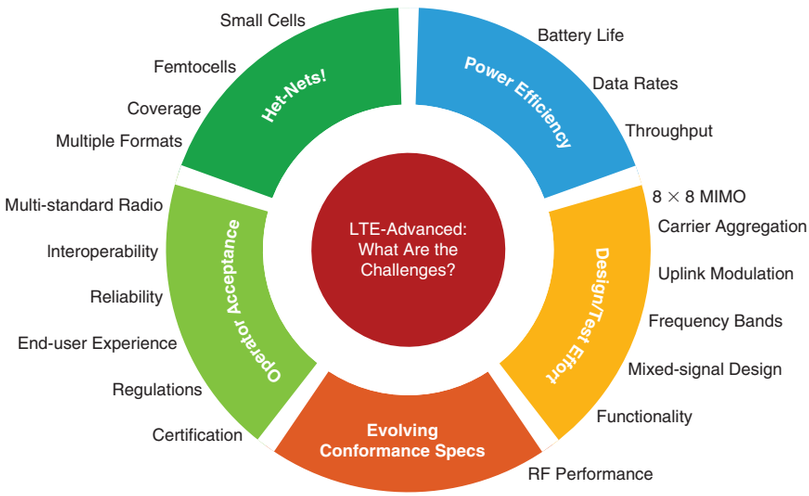


Figure 1: LTE-Advanced adds even more complexity to an already-challenging cellular environment
(Source: Agilent Technologies, 2013)

Designing User Equipment and Base Stations to Operate in Heterogeneous Network Environments

"…adding macrocells is not a good solution to increase capacity in an overloaded network."

"…small cells in Het-Nets are full-fledged local base stations with their own backhaul."

Advanced radio access techniques such as MIMO require near-ideal signal environments with high signal-to-noise ratio and power. These conditions are usually found close to the base station; however, as mobile devices get farther away and approach the cell edge, performance goes down. Adding more traditional base stations (macrocells) to improve coverage is expensive for many reasons: the difficulty of finding suitable locations, initial cost of the hardware, power requirements, and the installation and maintenance costs. For these same reasons adding macrocells is not a good solution to increase capacity in an overloaded network. LTE-Advanced therefore supports the use of relay nodes and small cells, which are much less expensive to acquire and operate and relatively easy to deploy.

As shown in Figure 2 small cells in Het-Nets are full-fledged local base stations with their own backhaul. The term includes microcells, picocells, and femtocells, with the latter getting the most attention. Femtocells are most often associated with the home base station defined in the LTE and LTE-Advanced specifications, but these small cells can be applied effectively in many situations from personal hotspots to the metrocells that enhance coverage in dense urban areas and indoor campuses. Elements of a Het-Net may encompass many radio access technologies from cellular (LTE, UMTS, GSM) to Wi-Fi*. Remote radio heads (RRH) and distributed antenna systems (DAS) may also be deployed.

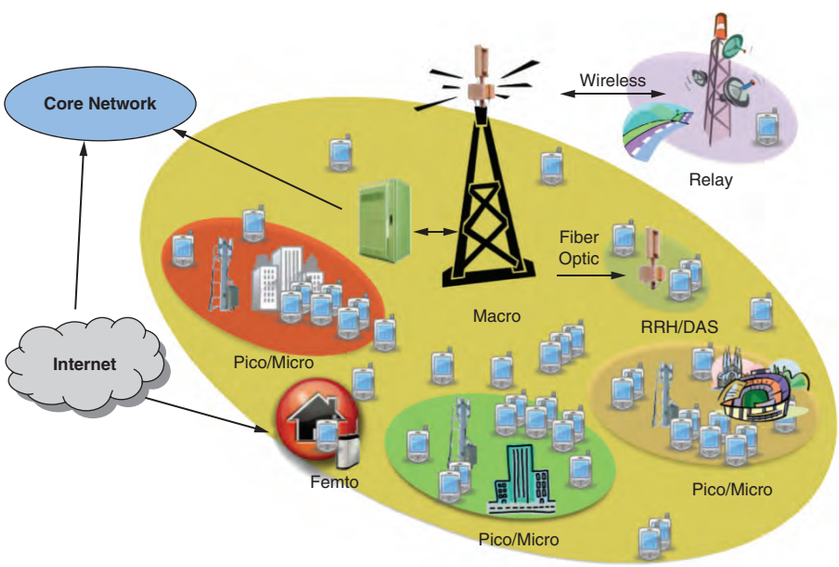


Figure 2: A heterogeneous network (Het-Net) supports the deployment of small cells and relay nodes, each optimized for different communication requirements. (Source: Agilent Technologies, 2013)

It's been estimated that in the next decade capital expenditures will shift from traditional base stations to small cells, accounting for USD 12 billion in 2018 according to Maravedis-Rethink.[2] By then 4.6 million LTE metrocells will have been deployed. At the same time much of the backhaul traffic may be offloaded to high capacity Wi-Fi networks. Clearly operators have a huge network management challenge ahead of them—not least of which is handling the interference that will be generated by the interactions of the multiple layers of cells and other RF-emitting devices occupying the same frequency. Contributing to the interference will be the multiple new transceivers in LTE-Advanced devices that are required for new speed and capability enhancements such as MIMO and dual-layer beamforming.

"…capital expenditures shift from traditional base stations to small cells, in 2018 4.6 million LTE metrocells will have been deployed."

In the Het-Net environment new co-channel interference scenarios arise that require new inter-cell interference coordination solutions. Two distinct forms of co-channel heterogeneous deployment each require a different approach to interference avoidance: the open subscriber group (OSG) and the closed subscriber group (CSG). OSG allows users to roam between the macro network and any local area base station deployed by the operator on the same frequency. In the area of the network where the strengths of the wide area and local area base stations are similar—typically a ring around the local area base station—interference is greatest and performance may be significantly degraded. Closer to the local area base station the interference becomes less problematic.

"Two distinct forms of co-channel heterogeneous deployment are open subscriber group (OSG) and the closed subscriber group (CSG)."

CSG limits local base station access to a fixed group of subscribers such as the occupants of a dwelling or employees of an enterprise. In the local base station coverage area, service for the CSG is good but all other users experience significant interference. This situation could be a major problem for macro network coverage in densely populated areas. The obvious solution is to assign different channels to the local base station and the macrocell. However, the solution is not available to operators with only a single channel. Some form of partial frequency reuse is also possible although there will still be interference in the control channels.

Given the difficulty of CSG, it is the focus of the initial LTE-Advanced standards work on enhanced interference mitigation in heterogeneous networks. Meticulous design of network devices and rigorous interference testing from design through deployment will be key to keeping this problem under control.

Maximizing User Equipment and Base Station Power Efficiency

As any smartphone user will tell you, battery life is a critical feature in a high-end mobile device. But making the battery larger to accommodate the extra transceivers required by higher order MIMO, for example, is not really an option. And operators also want base stations and small cells to operate as efficiently as possible, for both economic and ecological reasons. Developers

must therefore optimize power efficiency by considering new techniques for RF, baseband, and system-level designs.

"PAs are an essential component affecting the overall performance and throughput of wireless systems…"

Power amplifiers (PAs) account for a significant portion of both the energy consumed and heat generated by the RF front end. PAs are an essential component affecting the overall performance and throughput of wireless systems and are inherently nonlinear. Techniques to enable PAs to operate near saturation, where they are most efficient but also more nonlinear, are available now and will be more widely adopted over the next year or two.

Crest factor reduction (CFR) and digital pre-distortion (DPD) are two techniques that, particularly when used together, improve the linearity of a PA so that it may be operated at its high power-added efficiency (PAE) region, near saturation, without significant signal distortion. CFR pre-conditions a signal, reducing its high peak-to-average power ratio (PAPR) without causing significant additional distortion. DPD is a method of determining a PA's distortion characteristics, then applying the opposite effect to the baseband signal via a pre-distortion algorithm to improve linearity at the PA output. Both CFR and DPD are used by developers today.

"APT and ET will become more widely adopted over the next 12-18 months."

Average power tracking (APT) and envelope tracking (ET) are newer techniques to improve PA performance and efficiency. Both involve the control of the PA supply voltage as a function of the signal amplitude, an approach that has been used for many years. What is new is that APT and ET techniques work with modern PAs that offer switched high- and low-power operation rather than constant supply voltage. Thus, for example, envelope tracking can dynamically adjust the PA's supply voltage to track the magnitude of the envelope of the RF input signal. When the input signal envelope is low, the supply voltage can be reduced so the amplifier operates closer to its optimal efficiency point as shown in Figure 3. APT and ET will become more widely adopted over the next 12-18 months.

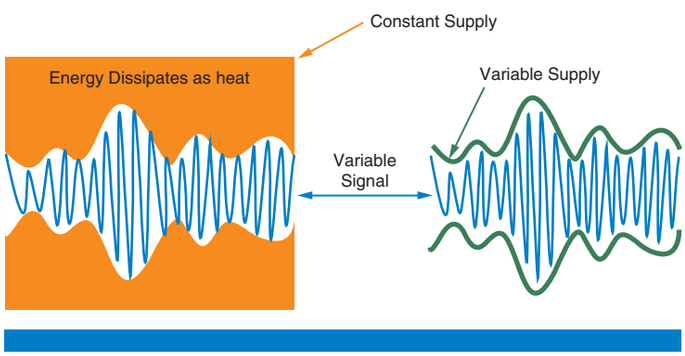


Figure 3: Envelope tracking is a technique that improves power amplifier performance by dynamically adjusting the supply voltage to track the magnitude of the RF input signal envelope.
(Source: Agilent Technologies, 2013)

In LTE-Advanced devices, it is not just the primary radio that requires power; power is required for multiband multi-RAT support, receive diversity, MIMO, interference cancellation, high data rates, Wi-Fi, Bluetooth*, FM radio, MP3, MP4, GPS, larger brighter displays, and—in the not so distant future—integrated video projection. Since battery life must be increased but not battery size, developers are directing an increasing amount of R&D effort towards designing, measuring, optimizing, and verifying UE current consumption in an ever wider set of use cases. Fortunately advanced battery-current drain measurement solutions are available now for analyzing current drain and validating and optimizing UE run times.

"battery life must be increased but not battery size, developers are directing an increasing amount of R&D effort towards verifying UE current consumption in an ever wider set of use cases."

Managing the Dramatic Increase in Design and Verification Effort

The schedule for LTE-Advanced development is aggressive given that LTE is just now enjoying widespread deployment. The newness and the complexity of both LTE and LTE-Advanced give rise to product development challenges, and not least is the fact that these are evolving standards, open to change and interpretation. From the technology perspective, new techniques add substantial complexity. For example, the use of multiple antenna configurations—with up to 8x8 MIMO supported in LTE-Advanced—makes the design of UEs more complicated, as does the new uplink modulation scheme introduced in LTE and the addition of carrier aggregation in LTE-Advanced.

"The "real-world" behavior of these enhancements is only now becoming understood and products optimized accordingly."

The "real-world" behavior of these enhancements is only now becoming understood and products optimized accordingly. Multiple channel bandwidths, while increasing the flexibility and capability of the cellular system, at the same time add to its overall complexity. Moreover, LTE-Advanced networks and devices have to operate with LTE and UMTS operating modes and as well as other wireless formats such as Wi-Fi and Bluetooth. Thus the ability to interwork seamlessly with other technologies is critical. Certain aspects of LTE-Advanced such as MIMO over-the-air (OTA) performance require entirely new test approaches, which are being defined in the 3GPP specifications.

The integration of the TD-SCDMA standard into the 3GPP specifications for LTE put a renewed emphasis on the development of systems with TDD capability and TD-LTE is emerging as a popular option. New components in the network architecture such as small cells and femtocells further complicate the picture.

Along with development challenges specific to LTE and LTE-Advanced are those generally associated with designing products for emerging wireless systems. Product designs tend to be mixed-signal in nature, consisting of baseband and RF sections. Overall system performance depends on the performance of the whole, yet each component type is associated with particular impairments—for example, nonlinearity and effective noise figure in an RF up-converter or down-converter; phase and amplitude distortion from a power amplifier; channel impairments such as multipath and fading; and impairments associated with the fixed bit-width of baseband hardware.

"Design simulation tools can address LTE-Advanced development challenges and verify their interpretations of the standard."

"…during the transition to hardware testing, a means of moving smoothly back and forth between design simulation and testing will ensure that engineers are not forced to redesign the product on the bench to get it to work."

With performance targets for LTE-Advanced set exceptionally high, developers have to allocate resources to cover each critical part of the transmit and receive chain. Astute decisions regarding system performance budgets will be key in meeting system-level specifications as well as time-to-market goals. Clearly managing the effort required in the design and verification process will be a major challenge to developers at every step of the product development lifecycle.

Design simulation tools can address LTE-Advanced development challenges and verify their interpretations of the standard. Models simulated at various levels of abstraction can support the progression from product concept through detailed design. Performance of both baseband and RF sections can be evaluated individually and together to minimize the problems and surprises encountered during system integration and other phases of the development cycle. Then, during the transition to hardware testing, a means of moving smoothly back and forth between design simulation and testing will ensure that engineers are not forced to redesign the product on the bench to get it to work.

Design and test integration provides even greater power and flexibility for hardware testing. For example using signal creation and analysis software in simulation along with logic analyzers, digital oscilloscopes, and RF signal analyzers provides as illustrated in Figure 4, a common test methodology with

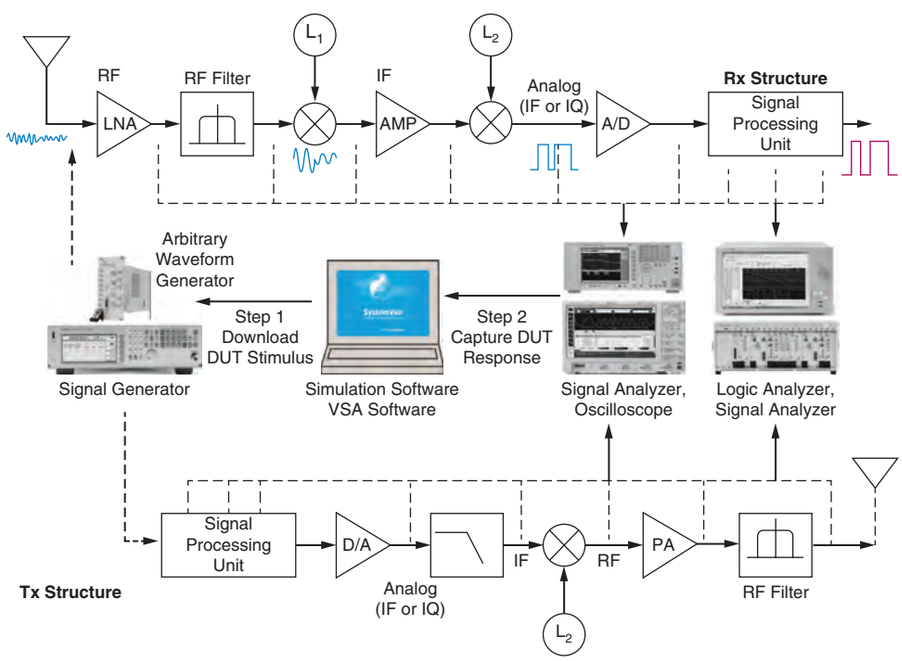


Figure 4: Combining simulation and test facilitates measurement and troubleshooting at various stages along the RF and mixed signal transmitter and receiver chain of a product design.
(Source: Agilent Technologies, 2013, *LTE and the Evolution to 4G Wireless, Design and Test Issues, Second Edition*. Moray Rumney, editor, Copyright 2013.)

a consistent user interface to help diagnose issues along the mixed-signal, RF transmitter and receiver chain (baseband, analog IQ, IF, and RF). This powerful capability can be used to identify potential issues earlier in the cycle, when they are easiest and least costly to fix.

Working on Designs before Conformance Tests are Fully Defined

The core specifications are needed to design any cellular product, and conformance tests provide the methods of measuring that product's compliance to the core specifications. For LTE and LTE-Advanced, specifications are defined by the 3rd Generation Partnership Program (3GPP). The conformance tests cover RF, radio resource management (RRM), and signaling (protocol) conformance. They are used by test labs in the process of certifying devices for the market, under the auspices of the GCF (Global Certification Forum) representing GSM and UMTS operators and the PTCRB (PCS Type Certification Review Board) in North America.

While the core specifications are essential for designing a product and, in the case of LTE, have been published at an astonishingly rapid pace, the conformance tests definitions tend to lag behind. That means manufacturers are often trying to bring products to market before their compliance can be officially confirmed—a risky proposition since finding compliance problems late in the design or production phase is very costly.

"…manufacturers are often trying to bring products to market before their compliance can be officially confirmed—a risky proposition since finding compliance problems late in the design or production phase is very costly."

Additionally, the number of frequency bands specified for LTE and LTE-Advanced along with the option for FDD- or TDD-based systems and the use of multiple subcarriers and multiple bandwidths creates a seemingly endless number of possible test configurations as shown in Figure 5. The specifications thus far are limited in the number of test scenarios available and of those, the certification groups have chosen a limited set of tests. Work continues on the test definitions, however, so manufacturers may find that the tests for a particular configuration do not yet exist or the tests change during the course of product development. Test equipment vendors who provide standards-compliant test platforms can be of help ahead of validated conformance testing by providing knowledge of the most important types of test and acceptable test procedures.

The list of conformance tests for LTE may seem large, but it's important to remember that passing all the tests on that list still assures only a minimum level of performance. Many other kinds of tests are still needed—in particular, design verification tests that provide a more thorough investigation of the product's performance margins, since the pass/fail nature of conformance testing does not reveal how close a product is to a particular limit. From the end users' perspective there is a need to test applications since the conformance tests are aimed at a lower level of capability to ensure mainly that the underlying transport mechanisms are in place to carry end-user services. Another step in the process of getting a product to market is operator

"The list of conformance tests for LTE may seem large, but it's important to remember that passing all the tests on that list still assures only a minimum level of performance. Many other kinds of tests are still needed…"

UE RF Transmitter Test Cases	UE RF Receiver Test Cases
UE maximum output UE maximum output power for intra-band contiguous carrier aggregation (CA) Maximum power reduction (MPR) Additional maximum power reduction (A-MPR) Additional maximum power reduction (A-MPR) for intra-band contiguous CA Configured UE transmitted output power Minimum output power General ON/OFF time mask PRACH time mask SRS time mask Power control absolute power tolerance Power control relative power tolerance Aggregate power control tolerance Frequency error Transmit modulation—error vector magnitude (EVM) Transmit modulation—PUSCH-EVM with exclusion period Transmit modulation—carrier leakage Transmit modulation—in-band emissions for non-allocated RB Transmit modulation—EVM equalizer spectrum flatness Occupied bandwidth 2 Out-of-band emission—spectrum emission mask Out-of-band emission—additional spectrum emission mask Out-of-band emission—adjacent channel leakage power ratio (ACLR) Transmitter spurious emissions Spurious emission band UE coexistence Spurious emission band UE coexistence (Release 9 and forward) Additional spurious emissions Transmit intermodulation	Reference sensitivity level Maximum input level Adjacent channel selectivity In-band blocking A In-band blocking for CA Out-of-band blocking Out-of-band blocking for CA Narrowband blocking Narrowband blocking for CA Spurious response Spurious response for CA Wideband intermodulation Spurious emission

Figure 5: Example of required RF tests for LTE UE transmitters and receivers—a subset of the tests required for industry certification. This list continues to evolve with each new release of the 3GPP specifications.
(Source: 3GPP Technical Specification 36.521-1,v11.1.0 (2013–06))

acceptance testing, which typically includes user-centric tests not otherwise covered. So while conformance testing is an essential step towards the successful deployment of a new system, it is by no means the beginning or end of the test process.

Designing to Pass Operators' Acceptance Tests for End-User Experience

The growth of smart devices has repositioned network operators from suppliers of "dumb pipes" through which wireless devices communicate to gatekeepers who put demanding performance, quality, and security metrics as entry qualifications to their network. Although type approval is the only formal step required of all devices before connecting to the network, operators frequently do more rigorous conformance and acceptance testing of devices before marketing them to customers.

"Unlike conformance tests, which are publicly available, the details of an operator's acceptance tests are usually confidential."

Acceptance tests are designed around specific characteristics and conditions of a given wireless network. Unlike conformance tests, which are publicly available, the details of an operator's acceptance tests are usually confidential. They cover

a multitude of scenarios that stress both the network and the UEs to ensure that problems are detected and resolved before they affect end users. Operators are particularly concerned about perceived quality of service, which can be greatly impacted by device performance. While new features or performance enhancements may strike a chord with customers, they add complexity to the UE and to network operation. Operators use acceptance tests to ensure that such enhancements do not degrade the quality or usability of devices.

"Operators are particularly concerned about perceived quality of service, which can be greatly impacted by device performance."

Typically acceptance tests cover the following areas:

- RF, EMC, and safety regulations as determined by the country
- RF and protocol conformance, usually test scenarios not included in certification testing
- interoperability, to ensure compatibility with other devices on the network
- functionality, to verify the operation of major features and function
- end-user experience, to measure the quality of the experience as perceived by the user
- reliability, to check that the device works properly under a range of different usage and environmental conditions

As more devices are designed and submitted to operators for testing, the strain on operator test facilities is increasing. Therefore test methods must be highly effective, efficient, reliable, and repeatable. Stress-testing a device in a lab under simulated environmental conditions is ideal to fully understand how a device is behaving and to analyze what has happened if something goes wrong. However, executing complex stress tests manually is time consuming, and operators may wish to automate as many of the test cases in a given test plan as possible.

"As more devices are designed and submitted to operators for testing, the strain on operator test facilities is increasing. Therefore test methods must be highly effective, efficient, reliable, and repeatable."

The process can become costly and time consuming when multiple UE vendors submit new devices for acceptance testing every few months. Therefore we expect that eventually UE vendors, rather than submit their new devices to the operator for acceptance testing, will themselves perform the acceptance tests as defined by the operator. Alternatively, the tests may be outsourced to operator-approved third-party labs. If a device does not pass acceptance testing at this stage, the UE vendor will be able to take immediate action to resolve the problem. Due to the complexity and proprietary nature of acceptance tests, operators will need to approve the exact implementation of the test systems used by UE vendors. An approved list of test systems will ensure that all tests carried out by other parties will be done to the same standards as used by the operator.

Finding the Right Solutions

New design and test methods are a critical part of what's needed to develop components, devices, and systems for communication technologies that must balance complex requirements such as the highest possible data throughput and system capacity with lowest possible power consumption.

"The best will offer more than a generic set of hardware and software tools; they will provide application-focused solutions for LTE and LTE-Advanced."

Partnering with the right design and test equipment vendor can increase the likelihood of success. The best will offer more than a generic set of hardware and software tools; they will provide application-focused solutions for LTE and LTE-Advanced. Here's what to look for:

- The vendor sells both design software and test solutions optimized for LTE-Advanced. Such vendors have to stay on the leading edge of standards requirements. Check out their level of participation in the key standards bodies.
- The vendor's LTE-Advanced design and test solutions provide best-in-class features and functions optimized for specific applications (for example, device design or network design). Check their track record for evolving over time to meet your application's changing needs.
- The vendor has a reputation for measurement integrity. Check their record for delivering consistent, traceable measurement results with accuracy appropriate to the required speed of the measurement.
- Every solution involves tradeoffs, so look for a vendor who can offer multiple solution options for both hardware and software. You're much more likely to find one that is right for your unique situation.
- Does the vendor provide easy access to measurement expertise related to your LTE-Advanced application? Check out the depth of their application notes, the frequency and quality of their webcasts, the value of their technical articles to you.
- Since the design of wireless devices and networks increasingly requires a collaboration between global teams, make sure your vendor has the resources to support your LTE-Advanced application wherever and whenever you need them.

References

[1] ABI Research: Press release, June 2013, https://www.abiresearch.com/press/lte-advanced-subscriptions-to-reach-500-million-by.

[2] Maravedis-Rethink: Press release, April 2013, http://archive.constantcontact.com/fs167/1103610692385/archive/1112994824383.html.

Author Biography

Jan Whitacre is LTE Program Lead at Agilent Technologies. Jan has over 30 years of engineering experience in cellular technologies. She received her BS degree in electrical engineering from the University of Wisconsin, Madison and then completed her MBA in Spokane, WA. Jan currently works on solutions from across Agilent and is currently focused on LTE, LTE-Advanced and WLAN. She coordinates communications and writes and presents training courses and articles and most recently was the project lead for both the first and second editions of the book *LTE and the Evolution to 4G Wireless*. Email: jan_whitacre@agilent.com.

HIGH PERFORMANCE CLUSTER COMPUTING AS A TOOL FOR 4G WIRELESS SYSTEM DEVELOPMENT

Contributors

Lars Thiele
Fraunhofer Heinrich Hertz Institute

Thomas Wirth
Fraunhofer Heinrich Hertz Institute

Michael Olbrich
Fraunhofer Heinrich Hertz Institute

Thomas Schierl
Fraunhofer Heinrich Hertz Institute

Thomas Haustein
Fraunhofer Heinrich Hertz Institute

Valerio Frascolla
Intel Mobile Communications

This article discusses the importance, benefits, and potential of electronic design automation (EDA) tools for the pre-development of features in the context of current and future wireless communication and other digital signal processing systems. The strengths, weaknesses, and overall added values will be assessed, looking forward to a more structured potential adoption into standardization bodies. Critical systems features of EDA systems such as scalability, expandability, flexibility, and ubiquitous application space are addressed in detail and illustrated with specific application examples.

Introduction

The full adoption of the Internet in every aspect of modern society creates huge opportunities for companies, institutions, and the public in general. The digital age we live in is characterized by a variety of digital representations of data, which can be sensor data from temperature sensors to web cams, documents, accounting data, and everything else that can be digitized. Furthermore, availability of such an enormous amount of data requires fast and reliable communication between places of origin and places where the data is processed in order to convert it into information valuable for enterprises, consumers, or processes. Such reliable communication links are the backbone of our industries and everyday lives and are facing a steady increase of data rates to access and transfer locally and/or over long distances.

"Future digital processing platforms need to comply with scalability, expandability,…"

With all systems for communication between systems and subsystems and systems for data acquisitions, processing, and analysis, we observe a steady rise in complexity that requires digital processing platforms with the following characteristics as an answer: scalability, expandability, flexibility, and ubiquitous application space.

The growing complexity of such systems arises from specific complex digital data or signal processing mechanisms inside. This can be the complexity of a signal processing algorithm in itself, meaning the number of multiplications and additions and memory accesses might be very high, or, as with many numerical approximation algorithms of iterative nature, the procedure is repeated again and again until a predefined stop criterion is reached. Another common increase in complexity comes from the amount of data to be processed either in parallel or in serial, one data block after another. Here, it is necessary to have the ability to partition the data analysis and processing such that as many pieces of data or algorithmic code can be accessed and processed in parallel. In many cases the data sets and processing chains have interrelated

inputs and outputs, meaning input and output data are a function of each other—another source of increased complexity.

In order to keep the overall processing time within a reasonable range, for instance for system evaluations of large and complex communication systems or in case of real-time requirements within affordable time constraints of the application, it is of utmost importance to keep the signal processing capacity and memory management scalable, flexible, and expendable in order to increase parallelization of computation tasks appropriately to the application at hand.

Another important aspect comes from the fact that in most cases next-generation communication devices require a well-balanced tradeoff between capital expenditure (CAPEX) and operational expenses (OPEX) platforms tailored to specific digital signal processing (DSP) tasks. This again requires a significant scalability in units sold within a certain timeframe such as user equipment. OPEX here is not to be understood just in terms of cost to operate the devices; often aspects of energy consumption will make a decisive difference in determining whether a device will be considered suitable for daily use, such as, for example, sufficient battery life for the application. Furthermore, most of the modules or subsystems should be the basis for further use in future systems or derivatives in the evolution of existing standards for cellular communication systems like 3GPP LTE and LTE-Advanced.

On the infrastructure side, the lifecycle of equipment is very much different due to high CAPEX during initial rollout of a new infrastructure and the constant need of functional updates for communication infrastructure optimization. Therefore, especially on the infrastructure side, software-defined-radio platforms are state of the art today. Nevertheless, the use of commercial off-the-shelf (COTS) hardware usually used for PCs or for data centers would make it possible to take advantage of the market scale and therefore reduce CAPEX and allow potential sharing of hardware improvement cycles and the resulting upgrade path for more signal processing speed and capabilities. On the other side, the real-time requirements for involved signal processing demand the design of new modular signal processing architectures to balance adaptively between universal computation capabilities (flexibility aspect) and specialized signal processing functional support by hardware accelerators (task-dependent high-performance aspect).

"…the use of commercial off-the-shelf (COTS) hardware usually used for PCs or for data centers would make it possible to take advantage of the market scale and therefore reduce CAPEX…"

This article is structured in the following way: the introductory section sets the scene with a general introduction of requirements for increasing signal processing capabilities. The section "An HPC Approach for SMEs and Research Labs" introduces the use of high-performance computing needs of small- and medium-sized enterprises (SMEs) and research labs that are not sufficiently supplied with PCs and simply can't afford to use big data centers' computing capabilities. We introduce the architecture and the modular design that allows the tailoring of high-performance computing (HPC) to customer needs. We show performance benchmarks with common applications used in the scientific community.

The section "High-Performance Computing Application Examples" illustrates the performance potential with some meaningful use cases from various application fields, including cell planning for large wireless deployments and wireless system performance analysis, and real-time wireless signal processing in the cloud RAN architecture context, which became very popular with mobile network operators just recently in order to handle the rising complexity in their networks in an adaptive and scalable manner.

The concluding section provides an outlook on relevant research issues to be addressed in future work.

An HPC Approach for SMEs and Research Labs

This section deals with the motivation and a prototypical setup for an expandable and scalable HPC architecture. Finally this section summarizes performance benchmarks using extensive mathematic computations.

Motivation

In the past the top reason not to deploy HPC at the enterprise level turned out to be both the hardware and application development costs.[1] On the other hand, Joseph et al.[1] highlight the fact that HPC solutions will play an important role that can dramatically gain importance at R&D labs.

"…Joseph et al. highlight the fact that HPC solutions will play an important role that can dramatically gain importance at R&D labs."

Therefore, it is of high interest to develop a scalable and expandable HPC architecture that can easily be extended once the computation complexity increases. In essence, buying an HPC system that is just big enough to handle today's processing tasks saves a lot of CAPEX now and OPEX over the years.

"…constraints such as privacy and know-how protection issues will endorse R&D labs to operate self-owned and self-maintained clusters on premises."

Other constraints such as privacy and know-how protection issues will endorse R&D labs to operate self-owned and self-maintained clusters on premises.

High-Performance Computing Architecture

The suggested HPC architecture is a collection of processing hardware such as standard x86 central processing units (CPU) or other highly specialized DSPs (see Figure 1). For some applications it may be worth spending the effort to integrate advanced graphical processing units (GPUs) or Intel's MIC[12] (many integrated cores) boards into such a computing architecture. The decision highly depends on the processing tasks case by case, and the overhead in transferring the data from the host CPU to the acceleration board has to be considered in detail.

"…simple NAS architectures do not provide sufficient data transfer bandwidth for high volume data and a large amount of data file access."

As a second important building block, the proposed architecture integrates a centralized storage system. The simplest solution is a network-attached storage device (NAS). However, in many HPC applications with a massive amount of data exchange, simple NAS architectures do not provide sufficient data transfer bandwidth for high volume data and a large amount of data file access. At this stage, we propose to add a cluster file system that can be distributed among several storage nodes.

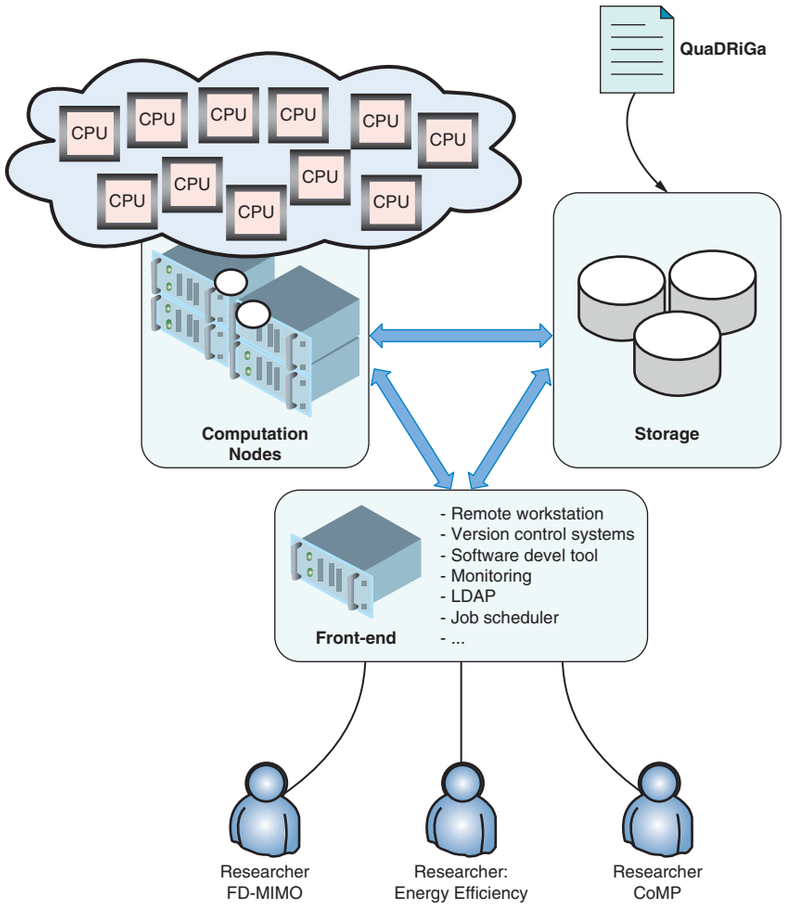


Figure 1: An expandable HPC architecture for R&D labs and SMEs. (Source: Intel Corporation, 2014)

Performance Benchmarking

For initial testing purposes, our HPC environment consists of 12 computing nodes, 2 nodes for centralized, shared storage of user data and a set of 15 virtualized servers. The parameters are summarized in Table 1. For high level performance benchmarking we use the standard Linpack package[3] for solving various systems of linear equations. In order to optimize the basic mathematic computations we utilize the OpenBLAS (Basic Linear Algebra Subprograms)[4], which can be tuned to the available system architecture. The chosen approach results in the following observations and benchmarks:

- At moderate efficiencies (around 60 percent): For the *Intel® Xeon® (X5690)*[13] processor, based on an enhanced *Nehalem* architecture called *Westmere-EP*[14], *NB =216, and hyperthreading enabled* we observe a computation peformance increase that is almost linear to the number of involved nodes (loss in efficiency is less than 3 percent).

"…we observe a computation peformance increase that is almost linear to the number of involved nodes…"

Computing nodes	12 nodes each with
CPUs	2 Intel® Xeon® X5690 3.47 GHz processors
cores	2 6
HT cores	2 12
Memory	96 GB
Interconnect	QDR InfiniBand*
OS	RHEL 6.4
Compiler	gcc 4.4.6
Math library	OpenBLAS 0.2.8
MPI	OpenMPI 1.6.4
Linpack	HPL 2.1
Rpeak	166.56 Gflops
Storage nodes	**Distributed file system over 2 nodes each with**
CPUs	2 Intel Xeon X5620 2.4 GHz processors
cores	2 6
HT cores	2 6
Memory	16 GB
Interconnect	QDR InfiniBand
OS	RHEL 6.4
HDDs	SATA Raid 6, 26 TB,

Table 1: High performance cluster computing configuration
(Source: Intel Corporation, 2014)

- At high efficiencies: The loss in weighted peak computing performance is less than 13 percent.
- Almost 90 percent efficiency is reached without hyperthreading and with thread pinning to the cores.
- The *Sandy Bridge (E5-2680)*[15] architecture doubles the overall computing performance of a *Westmere-EP (X5690)* architecture.
- In practice, we have to carefully select the right configuration for block size NB and the use of hyperthreading. For example, in applications that are not highly optimized and that also involve massive file I/Os and high volume data exchange, it may be beneficial to use these options, since time durations of CPU idles can be used by other processes.

"Almost 90 percent efficiency is reached without hyperthreading and with thread pinning to the cores."

"…in applications that are not highly optimized and that also involve massive file I/Os and high volume data exchange, it may be beneficial to use…"

These observations allow us to clearly argue that the suggested architecture is scalable in performance.

High-Performance Computing Application Examples

Within the following section we highlight the different applications for using HPC architecture in 4G wireless system development. It ranges from radio propagation modeling, large-scale radio network system-level analysis to real-time signal processing such as cloud-RAN or video re-encoding.

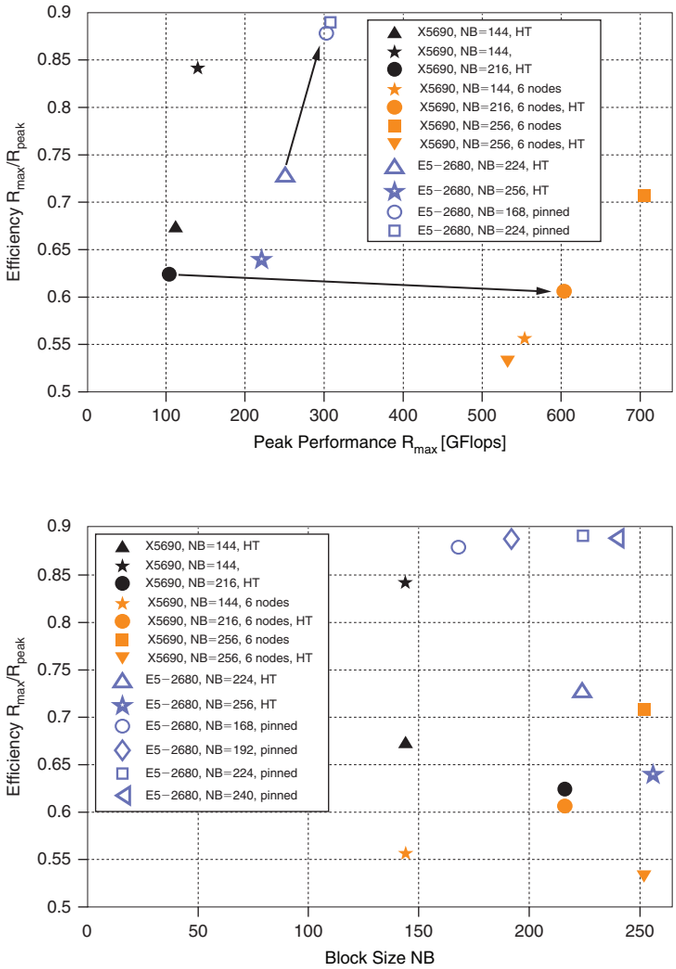


Figure 2: Measured Linpack performance using Open MPI 1.6.4 and Open BLAS 0.2.8 on both a Westmere-EP and a Sandy Bridge CPU architecture. Interconnection is done using QDR InfiniBand. (Source: Intel Corporation, 2014)

"Figure 2 shows highest efficiency without hyperthreading as well as a linear increase in peak performance with amount of nodes in the cluster."

Radio Propagation Modeling

Current and next-generation wireless communication systems need to be well planned prior to rollout. In fact it is key to choose for an optimized number of access nodes, so to provide the required coverage and initial capacity in the beginning. This also allows to evaluate migration paths towards cell densification or further capacity enhancing hardware and feature updates at the macro-cellular sites. In order to evaluate such complex communication systems, we need accurate channel models to capture the performance-relevant effects of the wireless transmission channel. Therefore, the geographic topology, positions of base stations, cell layout, frequency bands to be used, and the antenna configurations at the base station and terminal side will greatly influence the wireless system performance. In order to keep the

"…we need accurate channel models to capture the performance-relevant effects of the wireless transmission channel."

"…wireless channel models were derived and applied as performance benchmarking and feature performance comparison in standardization and for radio cell planning."

"Quadriga was recently published under GPL license and extends the family of the Winner channel models with several features."

system complexity manageable and the dominant physical effects sufficiently reflected in the evaluation, specific wireless channel models were derived and applied as performance benchmarking and feature performance comparison in standardization and for radio cell planning.

The guidelines of the 3GPP[16] spatial channel model (SCM)[9][10] introduce a ray-based bidirectional multilink model. The model was improved by the European project called Wireless World Initiative New Radio (WINNER)[8], to cover emerging requirements for 3GPP standardization of future cellular air interfaces such as Long Term Evolution (LTE) or LTE-Advanced (LTE-A). The fundamental idea of those channel models is to emulate the wireless channel with a set of rays, having a direct connection or being scattered at obstacles in the surrounding environment denoted as line-of-sight (LOS) and non-line-of-sight (NLOS), respectively. Each ray arrives at the receiver with a certain delay and power under a deterministic angle for the LOS connection. For the NLOS or multipath components (MPC), this angle is following certain geometry, yielding a multi-tap channel profile. The Fraunhofer Heinrich Hertz Institute developed a geometric-stochastic channel model denoted as Quasi-Deterministic Radio Channel Generator (Quadriga). *Quadriga*[5][6][7] was recently published under GPL license and extends the family of the Winner channel models with several features. The most prominent features are the 3D quasi-deterministic propagation assumptions, time evolution, and antenna representation with geometric polarization. Furthermore, *Quadriga* enables coherent generation of propagation conditions for different cell hierarchies such as macro cells, indoor and outdoor small cells, and satellite to ground communications, as shown in Figure 3.

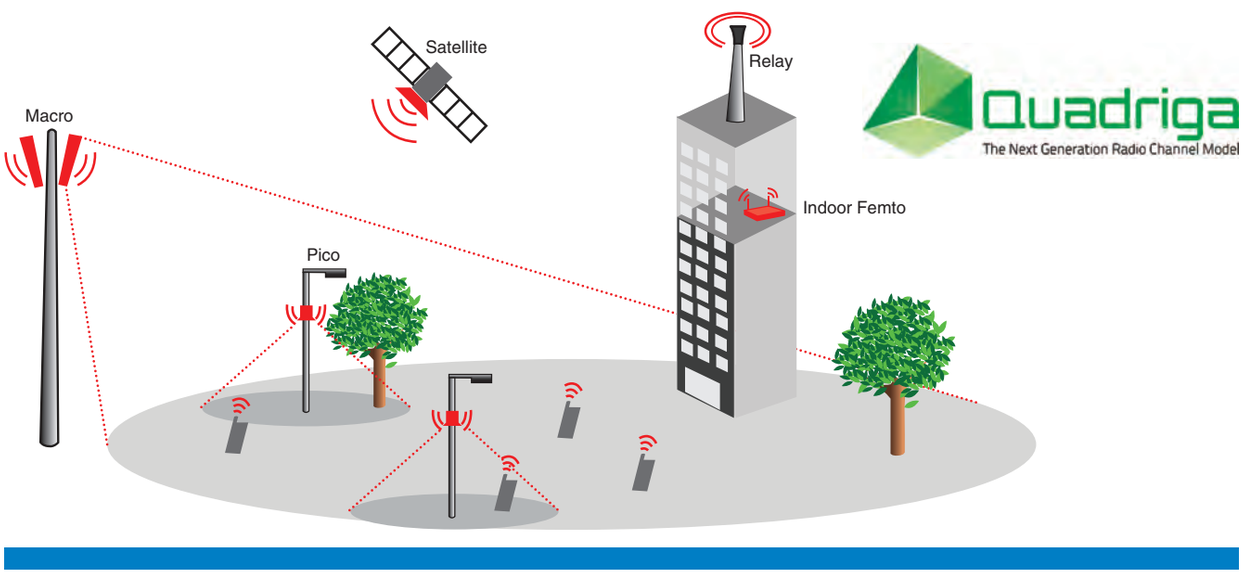


Figure 3: Wireless HetNet deployment scenario to be represented by Quadriga channel model (Source: Fraunhofer HHI, 2013[2])

Such wireless channel models contain statistical and deterministic processes that generate a very high volume of data to accurately model wireless system parameters such as channel time evolution and antenna polarization for hundreds of thousands of wireless links. An overview of the modeling steps is shown in Figure 4. The user provides the network layout, that is, the positions of the base stations, antenna configurations, antenna downtilts, the positions and trajectories of the mobile terminals (MTs), and the propagation scenarios. The channel coefficients are then calculated in seven steps that are described by Jaeckel et al.[5] in sections 3.2 and 3.3.

"…very high volume of data to accurately model wireless system parameters such as channel time evolution and antenna polarization…"

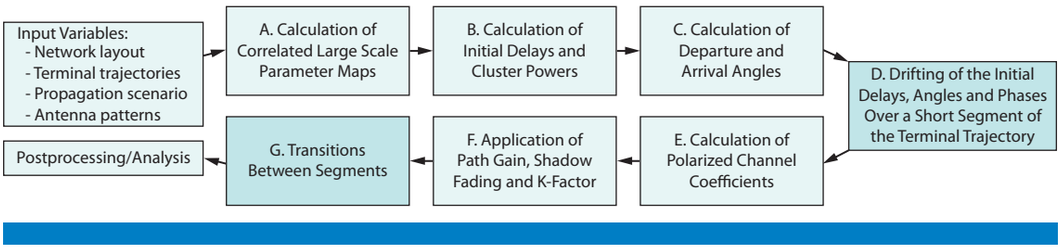


Figure 4: Block diagram from Quadriga Documentation[2] (Source: Fraunhofer HHI, 2013[5])

The computational complexity for reasonable cell layouts increases quickly and easily exceeds the processing and storage capabilities of a standalone desktop or laptop computer. Table 2 lists the required storage space depending on the desired deployment topology and amount of antennas per link and time samples (snapshots), as well as independent Monte-Carlo drops.

"The computational complexity for reasonable cell layouts increases quickly and easily exceeds the processing and storage capabilities of a standalone desktop or laptop computer."

Type	eNBs	ues	ant_eNB	ant_ue	taps	drops	Channel coefficients (1e9)	Needed Storage (GB)
Macros	21	630	4	2	20	500	2	47
Macros	57	630	4	2	20	500	6	128
Picos	42	630	2	2	16	500	2	38
Picos	114	630	2	2	16	500	5	103
LSAS	57	630	64	2	20	500	92	2,055
LSAS	57	630	128	2	20	500	184	4,110
LSAS	57	630	256	2	20	500	368	8,219
LSAS	57	630	512	2	20	500	735	16,438
LSAS	57	630	1024	2	20	500	1471	32,877

Table 2: Complexity of propagation modeling. For future wireless networks, such as large-scale antenna systems (LSAS), the number of channel coefficients is growing beyond billions. (Source: Fraunhofer HHI, 2013[2])

Table 3 summarizes the computation time for a 57 macro base station setup on a standard laptop using two hyperthreaded cores. The computation time can be reduced from 300 hours down to 9 hours when submitting the processing job to HPC with 140 hyper-threaded cores. The time saving scales linearly with

Quadriga simulation with 450 GB disk space	Laptop with Intel® Core™ i7 processor	HPCC
Involved Cores	4 @ 3.4 GHz	140 @ 3.4 GHz
Processing time	300 hours / 12.2 days	9 hours

Table 3: Cluster performance at the application level: it involves propagation modeling through Quadriga[2] and uses both extensive processing and storage capabilities.
(Source: Intel Corporation, 2014)

the amount of CPU cores. Note: the dimension for parallel processing is sufficiently large since we use an amount of 500 independent Monte-Carlo drops. Figure 5 depicts a typical coverage plot showing the signal-to-interference-and-noise ratio in a map layout. This metric is influenced by the path loss, the shadow fading, antenna pattern and transmit power per base station.

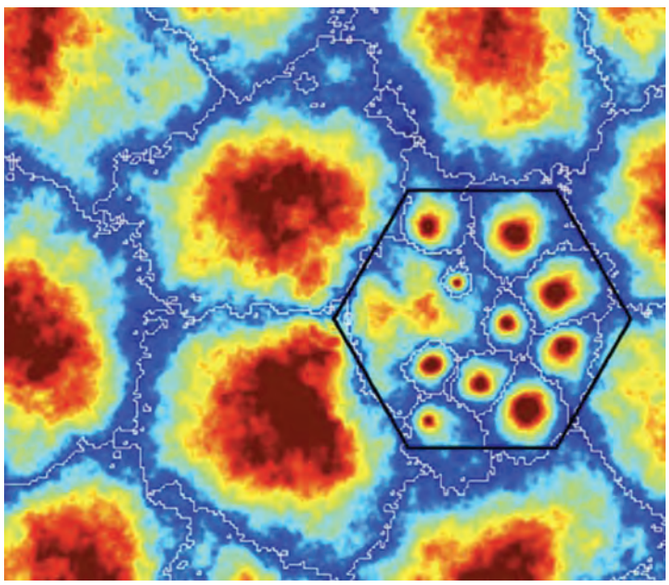


Figure 5: Wideband SINR coverage plot for a heterogeneous cellular deployment. This is usually considered as a high-level result from propagation modeling.
(Source: IEEE Asilomar Conference, 2013)

Other Applications for HPC Use

"…large-scale radio network system-level analysis for LTE/LTE-Advanced,…"

Other use cases for HPC usage are tasks requesting a heavy processing power, such as large-scale radio network system-level analysis for LTE/LTE-Advanced, data format conversion, or data analytics. A prominent example of data format conversion is video format re-encoding and/or rescaling in order to provide appropriate video formats to be transmitted over the Internet and to be displayed on a variety of screens. Since nowadays about half the Internet

traffic is video, the offering of an appropriate video size and compression/encoding format is of utmost importance. In fact it not only provides the basis for a good quality of experience for the end user, but also is needed to achieve a fast processing of the packets through the IP network, in order to allow fast download or streaming, and significant overall system energy savings. Different access bandwidths on the server and client side have to be supported, as well as a wide range of user equipment capabilities. The latest trends advocate cloud edge servers in order to reduce the data transport distance and therefore application-layer response time at least for predictable content to be consumed in the near future or by many people in the same area. As a consequence, jobs like data format conversion have to be done locally and distributed just where and when the need occurs. In our chosen example this means video re-encoding at the cloud edge if a video streaming or download on demand is requested from a nearby user. In an extreme case this might mean virtual real-time re-encoding of YouTube* videos for users employing poor bandwidth wireless connections to the Internet or overloaded cells. This provides a good motivation for localized signal processing using scalable HPC architectures that can be extended depending on needed and observed data traffic and signal processing requirements.

"…jobs like data format conversion have to be done locally and distributed just where and when the need occurs."

"…localized signal processing using scalable HPC architectures that can be extended depending on needed and observed data traffic and signal processing requirements."

Another prominent application is big data analysis. In this context we consider a big data analysis job to be performed where the huge amount of data to be analyzed can be distributed over many locations or is aggregated at a single location. In the former case, usually data analysis clients or agents (software modules) are executed at each location and analysis results or condensed information and data is forwarded and aggregated at a centralized point or forwarded to another location. These distributed agents do part of the analysis separately and forward the condensed part of the information to another more centralized data analysis instance. By doing so, data and preprocessed data from many locations can be analyzed in depth for correlations and other results of interest at the central location. For such applications HPC is the recommended architecture to do the job.

Advanced Real-Time Signal Processing

Besides large data processing for system behavior and system performance evaluation at the system level, many applications require true real-time processing for signal or data format conversion, signal or data transmission, or information extraction/collection for such things as pattern recognition. Real-time in this context means that the signal or data processing capability of the computing machine can keep pace with the speed and amount of incoming data or signals. Computation-induced delays must be limited to tolerable delays dictated by the application itself or the process that follows. For many applications these real-time requirements are stringent and in state-of-the-art solutions are often addressed with signal-processing task-matched hardware and hardware-based signal processing accelerators.

A prominent example is real-time video compression for live streaming of camera pictures, such as for live events like open air concerts or content

delivery from many cameras in a stadium to users or distribution centers. Since the application like a football game has to be streamed in real-time to the consumers at minimized latency, real-time processing is a must. The latest video standards such as HEVC aka ITU-T Rec. H.265 allow the specific compression rate to be a function of the input/output bit rate versus time. Such parameter space makes it possible to provide the appropriate video format for the tranmission bandwidth and the video display size and decoder capability requesting the streaming service. Also the number of videos and video formats encoded simultaneously in real-time demand scalable HPC architectures.

HEVC/H.265

In January 2013, and ten years after the widely-used ITU-T Rec. H.264/MPEG-4 AVC video coding standard was published, the High Efficiency Video Coding (HEVC) standard or ITU-T Rec. H.265 was finalized, published in June 2013 by ITU-T and in November 2013 by ISO/IEC. The coding efficiency of HEVC achieves a bit rate reduction of 50 percent for the same subjective quality compared to the best profile of H.264/MPEG-4 AVC, the High Profile. Thus it allows for doubling the number of services using HEVC compared to H.264 if using, for example, the same network resources.

"…coding efficiency of HEVC achieves a bit rate reduction of 50 percent…"

The coding efficiency gain of HEVC comes with an increased complexity on the encoder side, the coding complexity has increased by roughly a factor of 10 to 15 compared to H.264/MPEG-4 AVC. Therefore the HEVC coding standard was prepared to be executed on multiprocessor, high performance platforms.

"…the coding complexity has increased by roughly a factor of 10 to 15 compared to H.264/MPEG-4 AVC."

Unlike H.264/MPEG-4 AVC, where parallelism was an afterthought, the HEVC design contains several techniques making the codec better "parallelizable." H.264/MPEG-4 AVC supports slices, which were introduced mainly to prevent loss of quality in the case of transmission errors, but can also be used to parallelize the encoder. Employing slices for parallelism, however, introduces several problems such as a reduced video quality. The two main parallelization approaches included in the HEVC design are Tiles and Wavefront Parallel Processing (WPP). Both allow for creating picture partitions, but only the latter furthermore allows parallel processing without incurring any significant coding losses. Tiles not only target an effective parallelization of the codec, but also are optimized for conversational services.

"Each row is to be processed by a different core or processor in what is called a wave-front parallel fashion."

The idea of WPP is to partition the picture into rows of video processing blocks, which are called in HEVC Coding Tree Units (CTUs). Each row is to be processed by a different core or processor in what is called a wave-front parallel fashion. That is, while coding such blocks in a so-called raster scan order from the top to the bottom row plus from the very left to the very right block per row, coding such blocks is dependent on prior coded blocks, that is, the upper left one and the right one block. This dependency gives the name to the parallel processing procedure, which is not all processors start their rows to process at the same time, but by a shift, so that the upper left block is available from the predecessor row's coding process. The picture below shows the "shifted" and

row-wise processing fashion of WPP. The very new thing of WPP in HEVC is that the wave-front processing is not only applicable to the transform coding but also to the entropy coding parts of the hybrid video coding process. This means that in HEVC the full coding process can be executed in wave-front fashion, where for example in H.264/MPEG-4 AVC only the transform coding part was processable in a wave-front fashion. Figure 6 also shows the entropy coding dependencies as arrows for the very first left block of each row. This arrow indicates the initialization value for the entropy coding process, which originally was taken from the last block of the predecessor CTU row.

"In HEVC, both transform and entropy coding process can be executed in wave-front fashion."

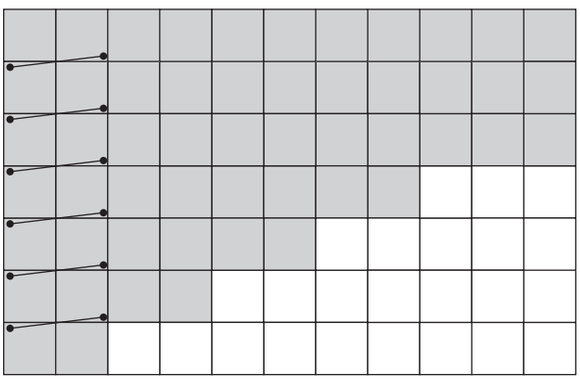


Figure 6: HEVC Wavefront Parallel Processing scheme (Source: IEEE Transactions on Circuits and Systems for Video Technology, 2012)

Since WPP's parallelization is limited by the number of CTU rows, WPP may even further scale on multi-core systems if a picture-overlapping approach were used, that is, the next picture is already processed, while the process of the current picture is still ongoing. We called that process Overlapping Wave-front Processing (OWP), as presented in [TCSVT-HEVC]. Figure 7 illustrates the OWP approach, where Thread T4 already processes the next picture, while Threads T1–T3 are still processing the current picture.

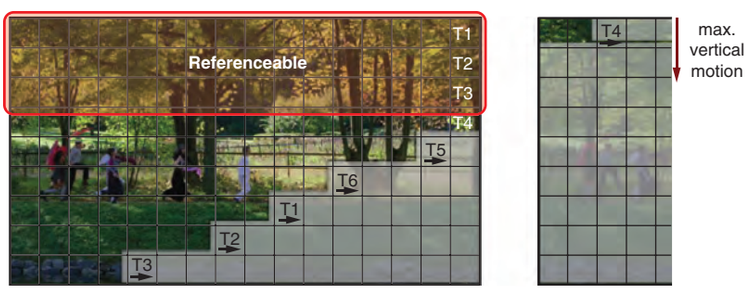


Figure 7: Overlapping Wavefront scheme (Source: IEEE Transactions on Circuits and Systems for Video Technology, 2012)

The parallelization techniques of HEVC may allow the extensive use of the processing performance available on multi-core/multiprocessor platforms as present in high performance clusters.

"The increasing capabilities of COTS computing hardware allows for virtualization of many signal processing functions, also known as network function virtualization (NFV),…"

Cloud-RAN Baseband Signal Processing

Another true real-time signal processing task is motivated by centralized signal processing or wireless transmission signals. The current trend of software-defined radio architectures in the radio access network (RAN) allow more and more software-based implementations of signal processing functions. The increasing capabilities of COTS computing hardware allows for virtualization of many signal processing functions, also known as network function virtualization (NFV), where the actual virtualization started originally from core network functions and is going step by step closer to the near antenna signal processing (see Figure 8).

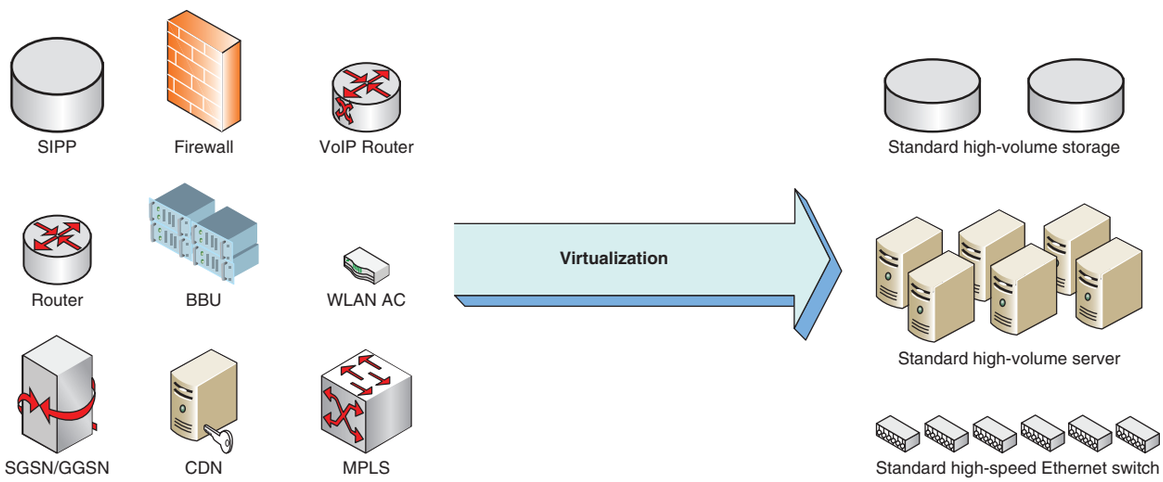


Figure 8: Cloud RAN architecture exploiting network function virtualization (Source: Intel Corporation, 2014)

The benefits expected from such RAN virtualization are:

- Multi-RAT support: each RAT can be implemented as one virtual machine on a single piece of computing hardware.
- Cost reduction and resource utilization improvement: multiple independent BBU (base band unit) entities fit into the same physical server (see Figure 9).
- Live migration to consolidate resources, so to save power.
- Resource sharing and consolidation according to traffic variance.

"…platforms require a real-time OS and hardware accelerators for radio-standard-specific high complexity algorithms…"

Two stringent requirements should be mentioned when considering standard servers. Most open platforms require a real-time OS and hardware accelerators for radio-standard-specific high complexity algorithms like matrix-vector multiplications, matrix inversion, turbo-decoding, or fast Fourier transforms (FFTs).

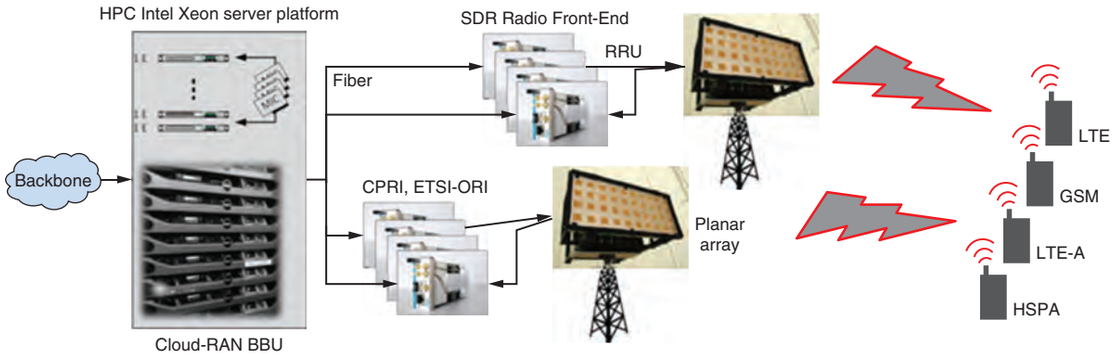


Figure 9: Server-based BBU (base band unit) and single RRU (remote radio unit) allowing multi-RAT processing in the same BBU.
(Source: Intel Corporation, 2014)

As an example, the signal processing complexity of the fourth generation radio standard LTE is shown, where the signal processing tasks can be shifted and split appropriately between BBU and RRU.

The red lines in Figure 10 depict splitting options between BBU and RRU in order to have a balance signal processing, either more centralized or more distributed at the remote radio unit (RRU).

"…splitting options between BBU and RRU in order to have a balance signal processing,…"

1. Soft-bit fronthaul (softbits plus control information).
2. Subframe data fronthaul (frequency domain I/Q plus control).
3. Subframe symbol fronthaul (frequency domain I/Q).
4. CPRI/OBSAI/ETSI-ORI fronthaul (time domain I/Q).
5. Compressed CPRI/OBSAI/ETSI-ORI fronthaul (time domain I/Q).

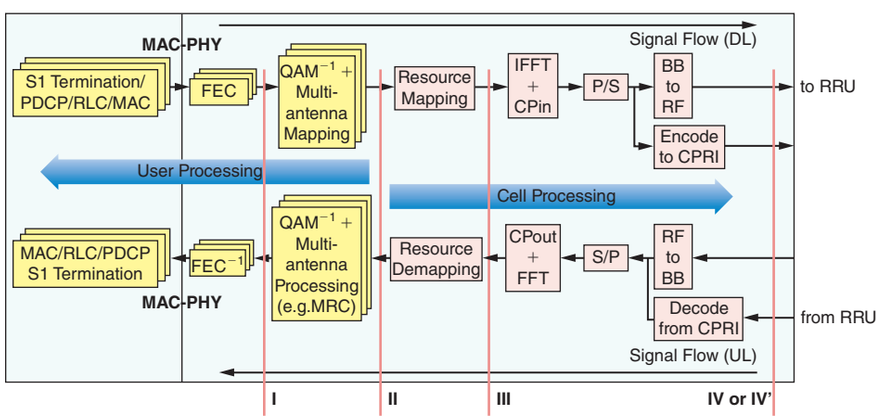


Figure 10: Server-based BBU and single RRU allowing multi-RAT processing in the same BBU.
(Source: Intel Corporation, 2014)

Furthermore, this has an impact on communication data rate on the interface between BBU and RRU. The Common Public Radio Interface (CPRI) as an example requires 2.5 Gbps for a 20 MHz LTE carrier and two transmit and two receive antennas per RRU.

Taking into account that a macro site consists of three sectors minimum, with up to eight transmit and receive antennas, the interface communication data rate multiplies by 12. Further capacity increase per site can be obtained by carrier aggregation techniques, which can allow bonding up to 5 20 MHz LTE channels (2.5 Gbps 12 5 = 150 Gbps of data rate would be needed just to drive the interface between BBU and the three multiband multi-antenna RRUs at one macro site). CPRI latency requirements are in the range of 100 μs depending on the particular signal processing task to be performed, for example coordinated multi-point (CoMP) over X2.

"…150 Gbps of data rate would be needed just to drive the interface between BBU and the three multiband multi-antenna RRUs at one macro site."

Assuming as the chosen scenario that several dozen macro eNBs are connected to the cloud RAN BBU plus a number of small cells (a 3GPP compliant implementation would imply up to 3–10 small cells per macro sector), this would rationalize the tremendous amount of signal processing power to perform real-time baseband signal processing and real-time interconnections between BBU and many RRUs.

To provide some latency constraints within the LTE framework we have to process all cell data in real-time at a processing latency end-to-end on IP level of about 4–5 ms one way, since hybrid automatic repeat request (H-ARQ) acknowledgments are pre-scheduled for retransmission requests. Parallel signal processing can be done at most stages but some algorithms like MIMO equalizers can scale cubically with the number of antennas to be processed jointly, which is especially challenging when going for advanced cooperative signal processing like CoMP, or the number of pilots involved for advanced channel estimation all to be done as fast as possible in order to keep signal processing latency to a minimum.

"…real-time at a processing latency end-to-end on IP level of about 4–5 ms…"

Figure 11 depicts the signal processing chain of an OFDM receiver for a multi-antenna single-component LTE carrier in the down link. Considering that MIMO channel estimation can be processed in parallel per channel coefficient, this means 1200 OFDM subcarriers per antenna pair of N transmit and M receive antennas. After channel estimation and interpolation per antenna pair, the channel equalizer has to be calculated considering the spatial structure between the antennas. Therefore the next step of equalizer calculation can be parallelized along the subcarrier axis. Then the equalizer is applied onto the receive signals, which can be computed in parallel by matrix times vector multiplications per subcarrier again.

In the next step, user-related data has to be separated from the allocated OFDM symbols and subcarriers defining physical resource blocks (PRBs) allocated to each user. These data symbols are to be demodulated and decoded separately for each user, a scalable option to work in parallel depending on the number of users to be processed.

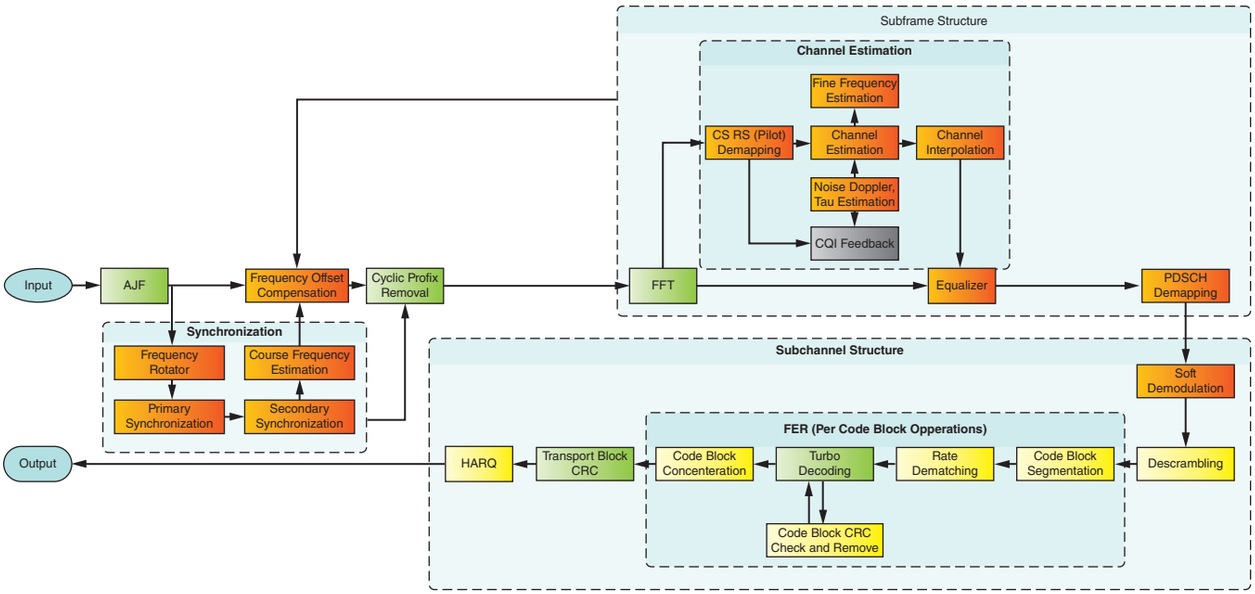


Figure 11: LTE real-time signal processing OFDM receiver chain (Source: Intel Corporation, 2014.)

If such processing is to be extended towards more transmit and receive antennas, then crosswise signal processing is possible for performance enhancements but at a price of increased signal processing complexity. If this is done all at the same location, in the same computing hardware using an HPC architecture, these feature enhancements and hardware extensions can be realized on demand.

In order to provide some reasoning of state of the art approaches, most eNBs are software defined radio (SDR) based but with dedicated DSPs for various reasons: real-time constraints require low-level hardware implementation of the signal processing routines; dedicated hardware macros are available on these platforms, such as FFTs, turbo-decoder, and matrix-vector multiplications. Furthermore, energy consumption per base station affects operating expenses, therefore these DSPs are tailored for the specific application in wireless communications. On the downside, feature upgrades have to be within the available hardware processing capability available at the distributed eNBs. Further hardware upgrades on remote locations can become quite costly and in many cases require extensive changes in the software partitioning over several boards on vendor-specific interfaces.

As a current trend, some vendors consider clustering of such dedicated signal processing hardware in centralized shelves (HPC) in order to have the benefits of centralized inter-eNB signal processing and the option of adding more dedicated hardware if needed.

Conclusion

"…ideally suited for SMEs and R&D organizations…"

In this article the concept of a small-scale HPC architecture was introduced, which is ideally suited for SMEs and R&D organizations that have to deal with high-complexity signal and data processing with scalability in amount of data and available computing time.

The proposed HPC architecture is shown to be scalable in terms of hardware components including server blades and attached memory banks on the one hand and on the other hand showed linear scalability regarding processing performance benchmarked with state of the art tools.

Furthermore we showed the applicability of such scalable HPC architectures for various application spaces and illustrated the capabilities and advantages with three particular examples taken from the wireless system evaluation and signal processing domain, wireless channel modeling, wireless multicellular system-level analysis, and cloud-RAN based real-time signal processing for LTE. These examples show that HPC allows flexibility in signal processing enhancements for evolutionary feature extensions and scalability in computing performance, if more of the same or more in-depth processing is to be done at a specific location.

"…HPC allows flexibility in signal processing enhancements for evolutionary feature extensions and scalability in computing performance,…"

"…small- to medium-size scalable HPC architectures providing new application space for SMEs and R&D centers…"

Summarizing the article highlights current trends, moving away from the two classic extremes of distributed small signal processing units and fully centralized big data centers towards small- to medium-size scalable HPC architectures providing new application space for SMEs and R&D centers in the wireless industry. The flexibility, scalability, and extendibility of such HPC architectures allow a new degree of freedom in CAPEX and OPEX optimization for a ubiquitous application space based on commercial-of-the-shelf hardware.

References

[1] Joseph, E., J. Wu, S. Conway, and S. Tichenor, "Benchmarking industrial use of high performance computing for innovation," White Paper, Council on Competitiveness, 2008.

[2] Quadriga, http://quadriga-channel-model.de

[3] http://www.netlib.org/Linpack/

[4] http://www.netlib.org/blas/

[5] Jaeckel, S., L. Raschkowski, K. Börner, L. Thiele, F. Burkhardt, and E. Eberlein, "QuaDRiGa - Quasi Deterministic Radio Channel Generator, User Manual and Documentation," Fraunhofer Heinrich Hertz Institute, Tech. Rep. v1.0.5–171, 2013.

[6] Jaeckel, S., K. Börner, L. Thiele, and V. Jungnickel, "A Geometric Polarization Rotation Model for the 3D Spatial Channel Model,"

IEEE Transactions on Antennas and Propagation, vol. 60, no. 12, pp. 5966–5977, December 2012.

[7] Jaeckel, S., L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D Multicell Channel Model with Time Evolution for Enabling Virtual Field Trials," submitted to *IEEE Transactions on Antennas and Propagation*, 2013.

[8] Kyösti, Pekka, Juha Meinilä, Lassi Hentilä, et al. IST-4–027756 WINNER II D1.1.2 v.1.1:WINNER II channel models. Technical report, 2007.

[9] Baum, D. S., J. Hansen, and J. Salo. "An interim channel model for beyond-3G systems," *Proc. IEEE VCT '05 Spring*, 5:3132–3136, 2005.

[10] 3GPP TR 25.996 V6.1.0, "Spatial channel model for multiple input multiple output (MIMO) simulations (Release 6)," Tech. Rep., Sep. 2003. Online Available: http://www.tkk.fi/Units/Radio/scm/.

[11] Chi, C. C., M. Alvarez-Mesa, B. Juurlink, G. Clare, F. Henry, S. Pateux, and T. Schierl: Parallel Scalability and Efficiency of HEVC Parallelization Approaches, IEEE Transactions on Circuits and Systems for Video Technology, IEEE TCSVT, Special Issue on Emerging Research and Standards in Next Generation Video Coding, vol. 22, issue 12, pp. 1827–1838, 2012.

[12] http://www.intel.com/content/www/us/en/architecture-and-technology/many-integrated-core/intel-many-integrated-core-architecture.html.

[13] http://www.intel.com/content/www/us/en/processor-comparison/processor-specifications.html?proc=52576.

[14] http://www.theregister.co.uk/2010/03/16/intel_xeon_5600_launch/

[15] http://ark.intel.com/products/64583.

[16] http://www.3gpp.org/.

Author Biographies

Lars Thiele (lars.thiele@hhi.fraunhofer.de) received the Dipl.-Ing. (M.S.) degree in electrical engineering from the Technische Universität Berlin in 2005. He joined the Fraunhofer Heinrich Hertz Institute (HHI) in September 2005. In 2013 he received the Dr.-Ing. (PhD) degree from the Technical University of Munich (TUM). He has contributed to receiver and transmitter optimization under limited feedback, performance analysis for MIMO transmission in cellular ODFM systems, fair-resource allocation, and CoMP

transmission under constrained CSIT. Lars has authored and coauthored about 50 conference and journal papers as well as a couple of book chapters in the area of mobile communications. He leads the System Level Innovation research group at Fraunhofer HHI and is actively participating in the GreenTouch Consortium.

Thomas Wirth (thomas.wirth@hhi.fraunhofer.de) received a Dipl.-Inform. (M.S.) degree in computer science from the Universität Würzburg, Germany, in 2004. In 2004 he joined Universität Bremen, Germany, where he worked in the field of robotics. In 2006 he joined HHI's WN department as senior researcher with the focus on real-time implementation for future wireless SDR prototypes. Since 2011, Thomas is head of the Software Defined Radio (SDR) group with the interest on algorithms for baseband processing including PHY and MAC as well as cross-layer design techniques for optimized video transmission over wireless systems.

Michael Olbrich (michael.olbrich@hhi.fraunhofer.de) is studying electrical engineering at the Technische Universität Berlin. Currently, he is working towards the Dipl.-Ing. degree (M.Sc.) at the Berlin Institute of Technology and the Fraunhofer Heinrich Hertz Institute (HHI). From 2003 until 2007, he was with Siemens Mobile (later Nokia Siemens Networks) and joined HHI in 2008. His research interests include performance evaluation of communication systems, MIMO-OFDM transmission, and high-performance computing.

Thomas Schierl (thomas.schierl@hhi.fraunhofer.de) received the Dr.-Ing. degree in Electrical Engineering (passed with distinction) from Berlin University of Technology (TUB) in October 2010. He is head of the research group Multimedia Communications in the Image Processing Department at Fraunhofer Heinrich Hertz Institute (HHI), Berlin. Thomas is the co-editor of various IETF RFCs and various MPEG standards. In 2007, Thomas visited the Image, Video, and Multimedia Systems group of Prof. Bernd Girod at Stanford University, CA, USA for different research activities. Thomas' research interests include system integration of video codecs, delivery of real-time media over mobile IP networks such as mobile media content delivery over HTTP, and real-time multimedia processing in cloud infrastructures.

Thomas Haustein (thomas.haustein@hhi.fraunhofer.de) received the Dr.-Ing. (Ph.D.) degree in mobile communications from the Technische Universität Berlin in 2006. In 1997, he joined HHI in Berlin, where he worked on wireless infrared systems and radio communications with multiple antennas and orthogonal frequency division multiplexing. He focused on real-time algorithms for baseband processing and advanced multiuser resource allocation. In 2006, he joined Nokia Siemens Networks, where he conducted research for LTE and LTE-Advanced. He is currently the head of the Wireless Communication and Networks Department at Fraunhofer HHI.

Valerio Frascolla earned his MSc in electrical engineering in 2001 and his PhD in electronics in 2004. He worked as research fellow at Ancona University, Italy, then moved to Germany, joining Comneon in 2006 and Infineon Technologies in 2010. Since 2011 he has been funding and innovation manager at Intel Mobile Communications, acting as facilitator of research collaborations using Agile methodologies and focusing on the program management of publicly funded projects and innovation activities. He is author of several peer-reviewed scientific publications and has been an invited speaker at international events. Email: valerio.frascolla@intel.com.

PACKAGING FOR MOBILE APPLICATIONS

Contributors

Thorsten Meyer
Intel Mobile Communications GmbH

Sven Albers
Intel Mobile Communications GmbH

Christian Geissler
Intel Mobile Communications GmbH

Gerald Ofner
Intel Mobile Communications GmbH

Klaus Reingruber
Intel Mobile Communications GmbH

Georg Seidemann
Intel Mobile Communications GmbH

Andreas Wolter
Intel Mobile Communications GmbH

This article gives an overview of mature and upcoming package technologies for mobile applications. It includes the Flip Chip package and the Wafer Level package as well as Embedding Die package technologies. The article introduces to the construction and main properties and it compares advantages and capabilities of the different packaging technologies. Finally the article shows the potential and possibilities for system integration in the different package platforms and proposes a selection guideline for choosing the right package technology for mobile applications.

Introduction

"Joel, this is Marty." With these words the mobile phone era started on April 3, 1973, when Martin Cooper from Motorola called his rival Joel Engel of Bell Labs. Many things have changed since then; phones have become smaller and functionality, like camera, text message service, mail service, and much more, has been added. Then mobile phones again became bigger with the introduction of the touch screen and additional features, like Internet and apps, paired with more and more sensors. In parallel, new and other mobile products were introduced. Personal digital assistants (PDAs), tablets, and the upcoming wearables, like glasses, wristwatches, or fitness trackers, are applications derived from mobile phones in other product classes.

It is common to all the mobile products that the functionality is increasing while dimensions shrink. This requires innovative technologies for displays, batteries, assembly technologies, semiconductors and, not least, packaging.

The most common material of choice for electronic integrated circuits is silicon. Advanced technology nodes are providing cost-effective and high performance integrated Circuits (ICs), which are getting smaller and smaller (Intel is ramping the 14-nm node right now). The printed circuit board (PCB) industry cannot follow the associated reduction in interconnect pitch at reasonable cost, which is one reason that these ICs cannot be directly soldered to a PCB in a phone or tablet. This "interconnect gap" has to be bridged by the packaging technology.

Besides, there are many other tasks the package has to fulfill, like hermetic sealing and mechanical protection of the IC in smallest dimensions, performance compatibility, thermal management, cost competitiveness, and reliability. These partly oppositional attributes and the growing importance of packaging for the future at the border to the end of Moore's Law is a good reason for a closer look into packaging.

Package History

The market of mobile communication and consumer electronics is determined by short product cycles, fast performance steps, and a hard cost competition due to many market players. For packaging technologies, this requires a continuous decrease of packaging and testing costs, reduction of package footprint and height, increase of electrical and thermal performance and, more and more importantly, the capability of system integration. In the early days of mobile packaging the (V)QFN ((Very thin) Quad Flat No lead) packages provided sufficient interconnects to the board with peripheral interconnect arrangement. With higher component integration more pins were needed. Due to these needs QFN-packages were replaced by ball grid array (BGA) packages. BGA packages are using an interposer and an area array of I/Os realized by solder balls. Chip pads were connected to the interposer by wire bonds. The area array allowed an area optimized chip and a PCB routing with relaxed design rules, which reduced the system cost. In the next step wire bond interconnects were replaced by solder bumps on the die. The bumped die was flipped and soldered on the substrate interposer, leading to the so called Flip Chip BGA packages (FCBGA).

"The market of mobile communication and consumer electronics is determined by short product cycles, fast performance steps, and a hard cost competition due to many market players."

In order to further reduce cost and package size, a thin film redistribution layer (RDL) replacing the substrate interposer and the solder bump was introduced with the Wafer Level Package Technology (WLB, Wafer Level Ball Grid Array). This can either be run on silicon wafers (fan-in WLB) or on artificial wafers (fan-out WLB). From a cost perspective for fan-out WLB it is beneficial to increase the degree of parallel processing. Here the transition from round wafer-like carriers to square or rectangular panels has already started.

The latest stage of mobile packaging technologies is now reached with 2.5D and 3D packages, where two or even more chips or packages are stacked onto each other. This is saving footprint and is adding performance by shortening the connection length between the active ICs.

"The latest stage of mobile packaging technologies is now reached with 2.5D and 3D packages, where two or even more chips or packages are stacked onto each other."

We will discuss the above-mentioned packaging platforms in detail in the following sections.

Flip Chip Packages

Flip chip technology was introduced by IBM in the early 1960s for use in their mainframe systems as so called "controlled collapse chip connection (C4)." The term *flip chip* refers to the chip being flipped from face-up to face-down orientation before placing it on the substrate.

Construction and Examples

Figures 1 and 2 show a typical flip chip package. It consists of a substrate with solder balls, a chip with bumps mounted on the substrate, an underfill, and a mold cap.

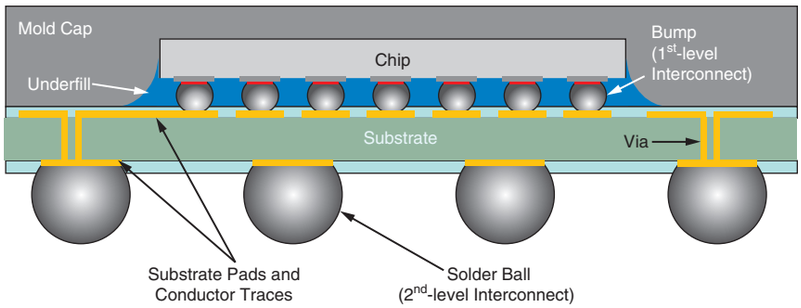


Figure 1: Schematic cross-sectional view of flip chip package (Source: Intel, 2014)

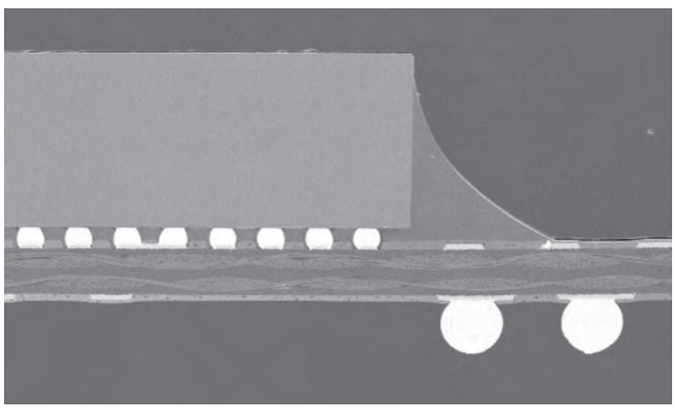


Figure 2: Cross section of flip chip package (Source: Intel, 2014)

"The bumps serve as first-level interconnect between chip and substrate."

"Routing layers and vias of the substrate connect top-side bump pads to bottom-side ball pads."

The bumps serve as first-level interconnect between chip and substrate. They are typically generated by reflowing deposits of electroplated solder or printed solder paste, nowadays drops of preformed bumps are also available. Solder balls attached to the bottom side of the substrate serve as second-level interconnects for mounting on a printed circuit board (PCB). Routing layers and vias of the substrate connect top-side bump pads to bottom-side ball pads. The underfill mechanically enforces the connection from chip to substrate while the mold cap protects the chip against environmental conditions. In the case of Figure 1 and Figure 2, it is a capillary type underfill, applied by dispensing along one or more chip edges. Alternatively, the chip can be underfilled by mold compound in the process step generating the mold cap. Underfill and mold compound for mechanical protection are applied in one molding process step and the mold compound takes over the role of the underfill. This simplifies the process flow because there is no longer a separate underfill step. Besides cost advantages, molded underfill also reduces keep-out areas for placement of the chips on the substrate.

The example in Figure 2 shows a two-layer substrate construction of PCB material with plated-through holes. The choice of the substrate is critical

because it is the dominant cost driver of the overall package. Besides the (cheap) two–layer construction with plated-through holes also (expensive) multilayer substrates, using laser-drilled micro vias, are available.

The routing layers of flip chip substrates allow the redistribution of the dense set of first-level interconnects to the chip to a larger set of second-level interconnects, suitable for mounting on a PCB. In many cases the chip area is too small for arranging all required second-level interconnects. For flip chip packages the substrate solves this problem by adding fan-out area, that is, by making the package larger than the chip. Flip chip packages thus belong accordingly to the class of fan-out packages, because the second-level interconnects occupy an area larger than the chip. In contrast, all second-level interconnects are arranged within the chip area for fan-in packages. Depending on the packaging platform, there are different options to realize the fan-out option, as we will discuss also in the following sections.

For a long time, solder bumps were the only first-level interconnect technology of flip chip packages. But the pitch of the first-level interconnect is limited by two factors: a minimum distance between the bumps and a minimum bump width. The distance between solder bumps has to be large enough to avoid solder bridging. A smaller bump, which would allow a tighter bump pitch, would also lead to smaller bump height and thereby to a reduced board-level reliability. Nevertheless, progress in frontend technology demands an ever-increasing density of first-level interconnects.

"For a long time, solder bumps were the only first-level interconnect technology of flip chip packages."

Smaller bump pitches can be achieved by Cu-pillar bumps because these can be produced with higher aspect ratios than solder bumps. Furthermore their cylindrical shape with only a solder cap allows for smaller space between the bumps without risking solder bridging. As an example Figure 3 shows the

"Smaller bump pitches can be achieved by Cu-pillar bumps because these can be produced with higher aspect ratios than solder bumps."

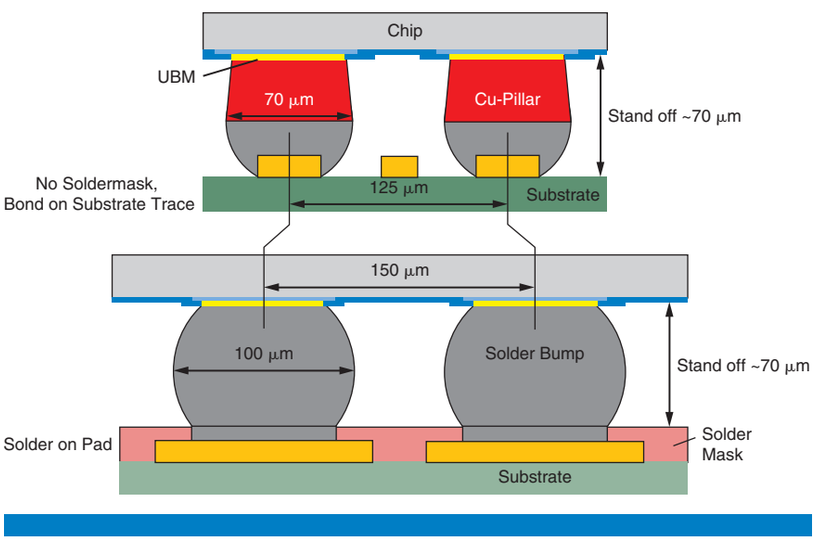


Figure 3: Cu-Pillar bump vs. solder bump
(Source: Intel, 2014)

reduction of the bump pitch from 150 µm with solder bumps to 125 µm with Cu-pillar bumps. The substrate used with Cu-pillar bumps has no solder resist and the Cu-pillar bumps are soldered directly onto the traces of the substrate. This allows the routing of a trace between two balls even at the reduced pitch. It also increases the flexibility of the interconnection due to the smaller pad dimensions on the board side and is thereby reducing the stress on the chip layers below. This is essential to avoid cracks in the brittle low-k materials used newest generation chip technologies.

"Reflow soldering is the common process for bonding flip chips to the substrate."

Reflow soldering is the common process for bonding flip chips to the substrate. The underfill material, which is reducing the thermo-mechanical stress in the first-level interconnect, is typically applied after reflow. During cool down from reflow, the mismatch of coefficient of thermal expansion (CTE) between chip (CTE ~3 ppm/°C) and substrate (CTE ~16 ppm/°C) leads to stress on the bumps and possible cracks in the chip layers below. Especially low-k-layer chips of the newest generation are prone to this failure mode. For Cu-pillar bumps, TCNCP bonding recently emerged as an alternative bonding method, reducing this stress (TCNCP stands for thermo-compression nonconductive paste). The TCNCP-process sequence (Figure 4) starts with dispensing the nonconductive paste, which later acts as an underfill on the substrate. Then the chip is placed into the paste and the solder-capped Cu-pillar bumps are pushed though the NCP. Bonding of the bumps to the substrate lands and curing of the NCP are achieved simultaneously by applying force and temperature. This way the NCP is already enforcing the first-level interconnect during cool-down after bonding. TCNCP allows for smaller bump pitches than reflow soldering because the chip position is well controlled during bonding by the bonding tool. Another advantage of TCNCP bonding is its thin chip capability. Even if bent, thin chips are forced into planar shape and good mechanical contact to the substrate by the applied bonding force.

"TCNCP allows for smaller bump pitches than reflow soldering because the chip position is well controlled during bonding by the bonding tool."

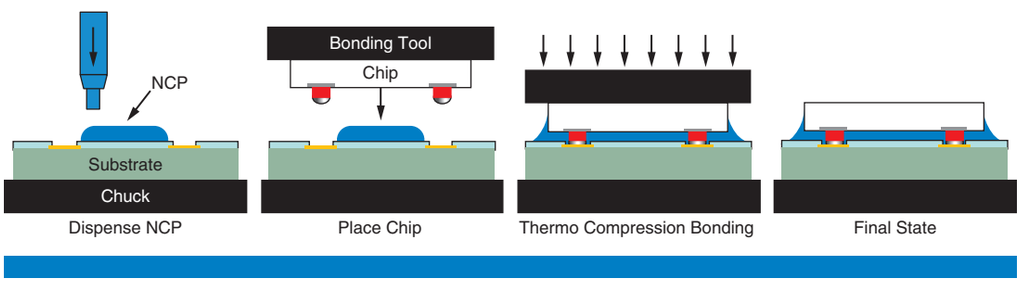


Figure 4: TCNCP process flow
(Source: Intel, 2014)

Advantages, Tradeoffs and Use

Generally, key performance indicators (KPIs) of packaging technologies for mobile applications are cost performance, electrical performance and interconnect density, thermal performance, System-in-Package performance, and board-level reliability. For comparison of the different packaging

technologies, we will discuss these KPIs in the respective sections starting here with flip chip packages:

- *Cost performance:* Flip chip technology offers good cost performance, especially for large packages.
- *Electrical performance and interconnect density:* Flip chip packages use an area array arrangement for the first-level interconnects. This allows a higher I/O count and shorter signal paths compared to, for example, wirebond packages.
- *Thermal performance:* Heat removal faces high thermal resistances, because the chip is encapsulated in mold compound and is separated from the board (heat sink) by the substrate. Both have a comparatively low thermal conductivity.
- *System-in-Package (SiP) performance:* Flip chip and wire bond stacking and side-by-side are well-established SiP solutions.
- *Board-level reliability:* Flip chip packages offer very good board-level reliability, because the substrate acts as a stress buffer. Furthermore for standard substrate materials the CTEs of substrate and PCB are similar.

"…Flip chip packages offer very good board-level reliability, because the substrate acts as a stress buffer."

A rating of these properties for flip chip packages is displayed in Figure 5.

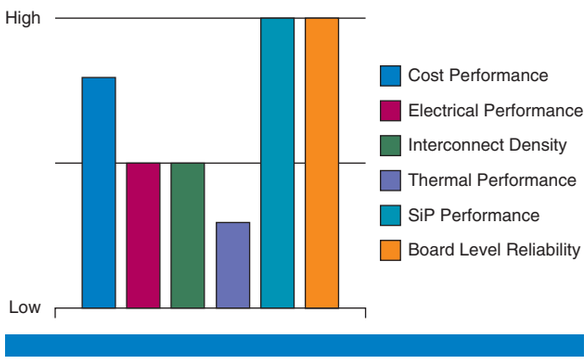


Figure 5: Performance rating of flip chip technology (higher performance indicates better values) (Source: Intel, 2014)

System Integration Capabilities

Flip chip technology is the basis for several approaches to integrate multiple chips into the same package, as shown in Figure 6.

Side-by-side integration and traditional stacking use well-known processes that allow for high yield and a broad supplier base. Disadvantages are the large footprint of the side by side construction and limited electrical performance associated with wire bonds in the case of chip stacking.

"Side-by-side integration and traditional stacking use well-known processes that allow for high yield and a broad supplier base."

The package-on-package approach allows the stacking of multiple and even different types of packages. The packages can be tested or subject to burn-in before assembly. Drawbacks are large height and footprint. Electrical performance is limited by long connections between the chips.

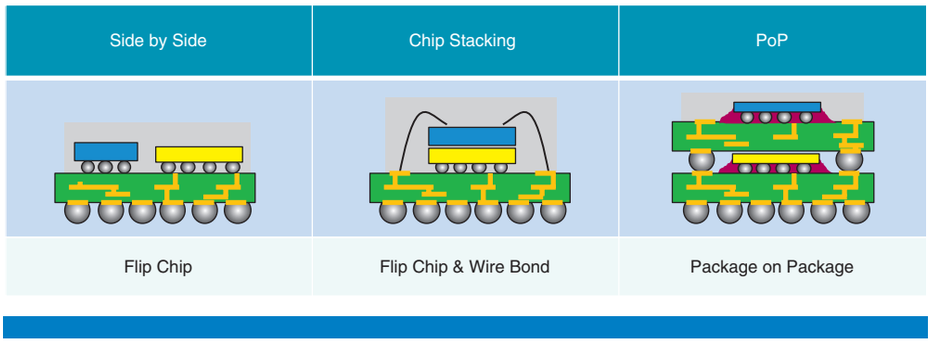


Figure 6: Flip-chip-based multichip packages (Source: Intel, 2013)

"Drivers for future developments are the need for finer first-level interconnect pitches, height reduction, system integration and cost reduction."

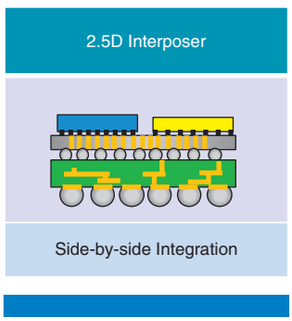


Figure 7: Flip chip package for side-by-side integration on 2.5D interposer (Source: Intel, 2013)

Future

Drivers for future developments are the need for finer first-level interconnect pitches, height reduction, system integration and cost reduction. Cu-pillar bumps are well established. Optimizations and the wider adoption of TCNCP with direct bonding on substrate traces (bump on lead technology) will allow for a considerable pitch reduction. Coreless substrates are a recent development, targeting cost-efficient generation of thinner substrates with finer line/space capability. In such a substrate all layers are built up sequentially and connected by micro vias. The trend towards thinner packages will lead to substantial warpage issues for flip chip packages. New substrate materials with lower CTE, higher Tg and Young's Modulus will help to solve this challenge.

The integration of passive components either side by side or into the substrate allows smaller overall solutions with improved electrical performance.

Also, flip chip technology will play a key role for packaging 2.5D interposers as shown in Figure 7. Side-by-side integration on 2.5D silicon interposers allows very fine pitches, high wiring densities, and a perfect CTE matching between chips and interposer.

Wafer-Level Packaging

For wafer-level packages (wafer level ball grid array, WLB), the substrate and the first-level interconnect of a flip chip or wire bond BGA melt into a thin-film redistribution layer (RDL). There are two classes of wafer-level packages: For fan-in wafer-level packages, all second-level interconnects are arranged within the chip area and the RDL is applied on the silicon wafer. For fan-out WLBs, the second-level interconnects require more area than just the chip area and therefore an artificial wafer is constructed prior to application of the RDL. The RDL then extends over the chip area. Both technologies will be discussed in the following sections.

Fan-in Wafer-Level Packages

Wafer-level packages (WLBs) have been around for almost 20 years. WLBs are increasing their market share at a rate in the low two-digit percentage range

year by year. WLB technology was introduced for small chips to reduce cost, dimensions, and to improve the electrical performance by reducing the length of conducting lines.

Construction and Examples

Two groups of fan-in wafer-level packages can be distinguished, WLB without and WLB with a redistribution layer, as illustrated in Figure 8.

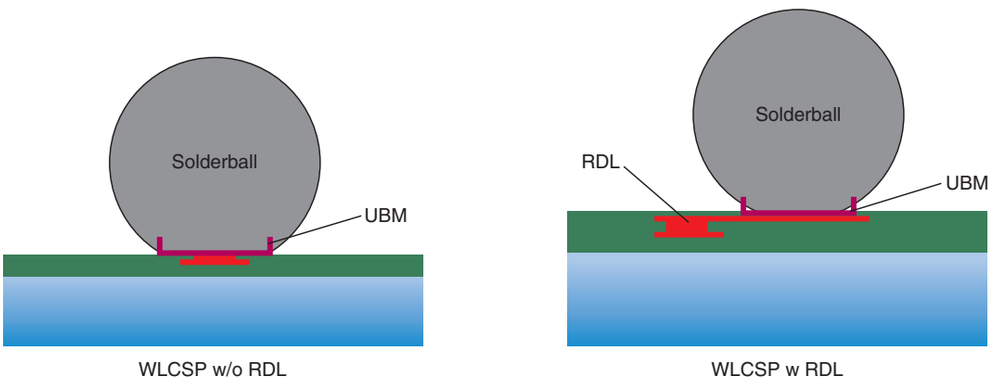


Figure 8: Schematic cross-sectional view of WLB technologies (Source: Intel, 2014)

WLB without RDL is used for small ICs with low ball counts, where the pad positions on the chip fit to the landing pad positions on the customer board, as shown in Figure 9.

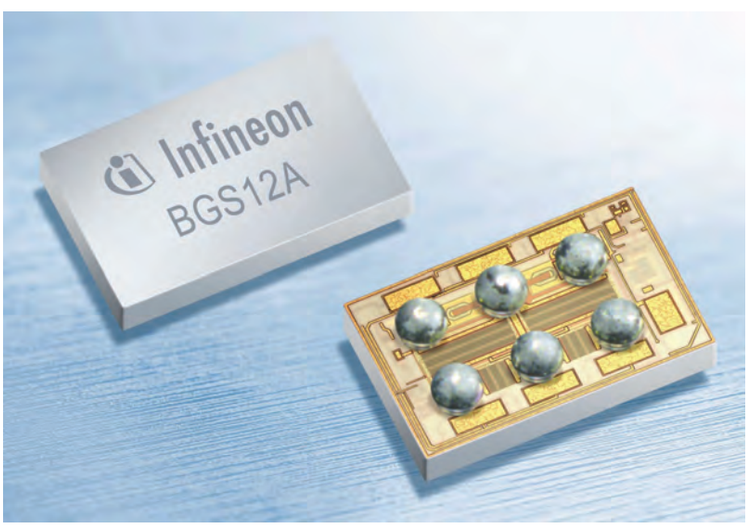


Figure 9: WLB without RDL (Source: Infineon, 2008)

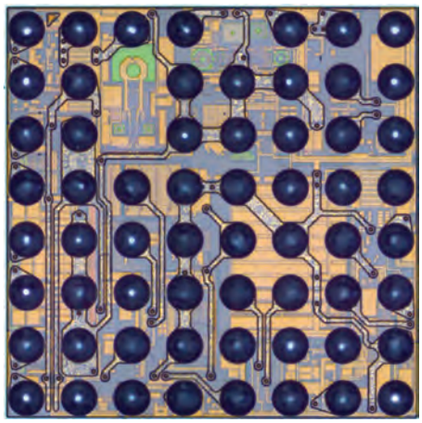


Figure 10: WLB with RDL
(Source: Infineon, 2011)

"If the pad pitch of the chip is not matching to the pad pitch on the customer board, a WLB with redistribution layer is used,…"

If the pad pitch of the chip is not matching to the pad pitch on the customer board, a WLB with redistribution layer is used, as shown in Figure 10. It allows the redistribution of the pad positions of the dense first-level interconnects to the positions of the second-level interconnects.

For both options, with and without RDL, a similar process flow, shown in Figure 11, is used. Packaging starts with a silicon wafer that has completed frontend fabrication. This wafer is used as substrate for the complete packaging process.

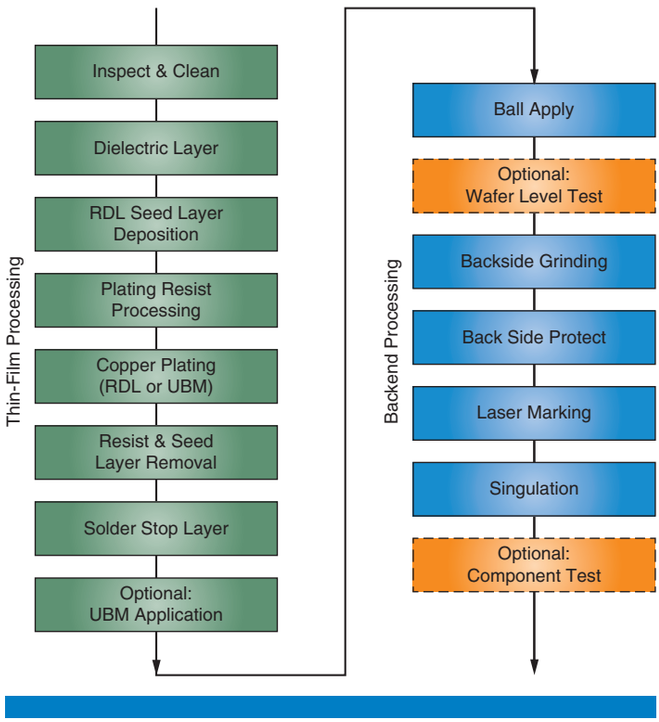


Figure 11: WLB process flow
(Source: Intel, 2014)

First, a polymer passivation is deposited on the wafer. The pads of the chip are opened. Then a seed layer, starting point for electroplating of the RDL, is applied. After deposition and structuring of a plating resist, the RDL is electroplated. If no RDL is required, only landing pads (under-bump metallization, UBM) for the solder balls are applied. The plating resist and seed layer are removed. A polymer solder stop layer is deposited and structured; typically the same material as for the polymer passivation is used again. For improved thermo-mechanical reliability, an additional UBM can now be applied. After the balls are applied, the dies can be tested on the wafer level. If a classical component test is preferred, the final steps of grinding, marking, and dicing are performed prior to the component test.

"Wafer-level packages typically are not underfilled after assembly on the customer board."

Wafer-level packages typically are not underfilled after assembly on the customer board.

Advantages, Tradeoffs, and Use

The WLB technology is suitable for small chips with low to medium I/O count. These are, for example, integrated passive devices, power management units, DC/DC converters, MEMS, RF filters, and connectivity chips for Wi-Fi*, Bluetooth*, or GPS.

"The WLB technology is suitable for small chips with low to medium I/O count."

- *Cost performance:* Fan-in wafer-level packages are low cost packages with minimum lateral dimensions and minimum package height. Due to the use of thin-film technology, a very high yield can be achieved.
- *Electrical performance:* The electrical performance is excellent due thin-film technology and shortest interconnect length.
- *Interconnect density:* Due to the high accuracy of the thin-film technology, very high interconnect density can be achieved. The use of the WLB package technology is only limited by the number of second-level interconnects being applicable to the chip.
- *Thermal performance:* This is characterized by low thermal resistances and short connections, therefore the WLB technology provides good thermal performance. On the other hand, heat transfer is restricted to the solder balls.
- *Board-level reliability and thermal performance:* WLBs show good reliability performance for small chips. Due to the mismatch in thermal expansion between chip and PCB, the maximum size of a reliable WLB is restricted.
- *System in Package:* The System-in-Package capabilities are limited, see also the following section.

"The electrical performance is excellent due thin-film technology and shortest interconnect length."

Figure 12 shows a general overview of the WLB package properties.

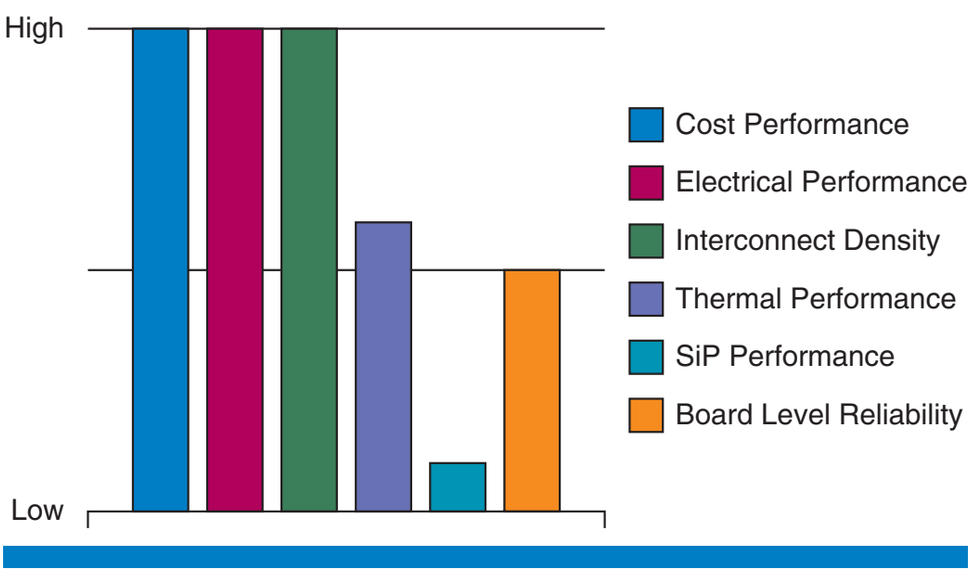


Figure 12: Performance rating of WLB technology (higher performance indicates better values) (Source: Intel, 2014)

"The WLB packages typically are small, offering no space for system integration."

System Integration Capabilities

The system integration capability of the fan-in WLB technology is limited. The WLB packages typically are small, offering no space for system integration.

Passives can be implemented in the RDL of the WLB. However, Quality-factors are limited due to the silicon substrate, and the achievable values for capacitance or inductance are fairly low.

Three-dimensional integration either requires through-silicon vias (TSVs), which add appreciably to cost, or is restricted to possum chips, attached face-to-face to the active side of the WLB.

Future

Mobile applications will need more functionality in the future and yet have to be thinner, lighter, faster, and cheaper. Fan-in WLBs offer the lowest cost and smallest form factor. Therefore, it is the best suitable packaging technology for applications with a low to moderate number of I/Os. The development of reliability enhancement features will open up the door to larger package sizes, allowing additional products to jump on the technology. And the further use of WLB packages is not limited to mobile phones and tablets. In wearables and medical devices as well as in automotive applications, WLB packages will also find future use.

"Fan-out wafer-level packages are the extension of WLCSP packages to higher I/O counts."

Fan-out Wafer-Level Packages

Fan-out wafer-level packages are the extension of WLCSP packages to higher I/O counts. Additional area around the die, applied in the packaging process prior to the application of the previously described thin-film layers, serves as fan-out area.

Fan-out wafer-level packages (FO-WLBs) were developed especially for the mobile application market. Infineon was the first company, introducing its FO-WLB technology, called eWLB (embedded wafer-level ball grid array) to the market in 2009. Today, eWLB is in high volume manufacturing at multiple suppliers. With the acquisition of the wireless business from Infineon, Intel has adopted the technology. Other companies are also offering their own fan-out wafer-level package solution. Most of them are still in the development phase; some are currently starting with production.

Construction and Examples

Based on the eWLB-package technology, the typical process blocks of FO-WLB technologies are visualized in Figure 13.

"The main innovation in the step from classical fan-in WLBs to FO-WLB was the introduction of an artificial wafer to generate the fan-out area."

The main innovation in the step from classical fan-in WLBs to FO-WLB was the introduction of an artificial wafer to generate the fan-out area. This process block is called *reconstitution*. Tested good dies are picked from a singulated silicon wafer and placed on a carrier with a larger, free arbitrary spacing. This spacing is defining the fan-out area of the package. By using a compression molding process the dies are embedded into mold compound, filling the gaps between the chips and forming the artificial wafer. After that,

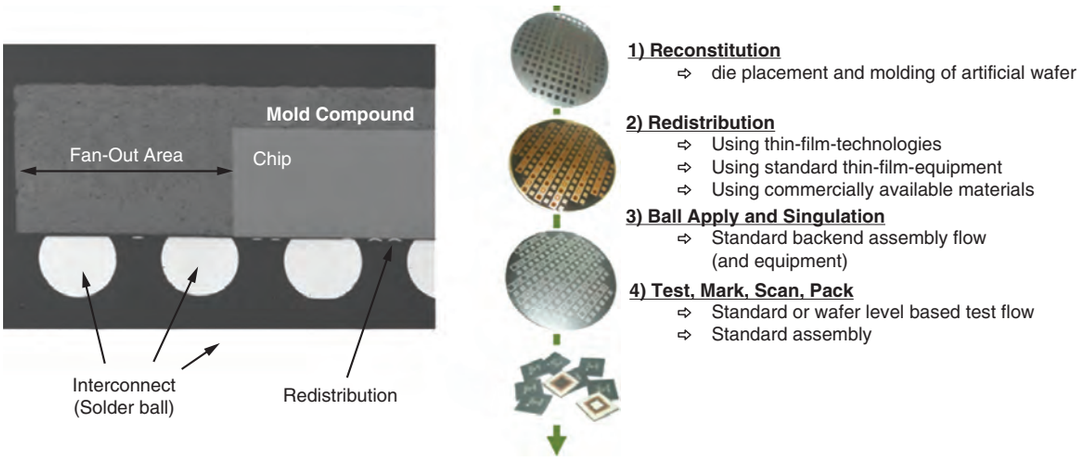


Figure 13: eWLB cross section and process overview (Source: Infineon, 2008)

the carrier is de-bonded from the carrier. The reconstituted wafer is now processed similar to the previously discussed fan-in WLB silicon wafers. A dielectric is applied and structured. Due to the limited temperature stability of the mold compound, a low temperature cure material is used. The redistribution layer is applied in thin-film technology and is typically extending laterally beyond the chip area. The solder stop layer covers the redistribution and defines the landing pads for solder balls. Final process steps in the backend process block include wafer thinning, laser marking, the ball application process, and artificial wafer dicing.

Advantages, Tradeoffs and Use

FO-WLBs can be used to meet the big challenge of z-height reduction in mobile applications. Package thicknesses below 0.3 mm are possible, for single die as well as for package stacking (PoP) options.

"FO-WLBs can be used to meet the big challenge of z-height reduction in mobile applications."

- *Cost performance:* FO-WLBs achieve an excellent packaging cost position, especially for small packages. Since it is a fairly new packaging technology on the market, additional opportunities for cost reduction are still available (such as substrate size increase, and so on).
- *Electrical performance:* Electrical performance is very good because of short, low resistant connections and very low parasitics.
- *Interconnect density:* Due to the high accuracy of the thin-film technology, very high interconnect density can be achieved.

"…Due to the high accuracy of the thin-film technology, very high interconnect density can be achieved."

- *Thermal performance:* Due to the slim stack up of FO-WLPs good thermal dissipation into the board is given.
- *Board-level reliability:* FO-WLBs have reliability limitations similar to fan-in WLBs. The CTE mismatch between PCB (~16 ppm) and silicon chip (~3 ppm) has to be buffered by the solder balls. TCOB lifetime improvement is a common development target of the industry. For customers it is also

"Lowest warpage changes over temperature are provided by eWLB."

very important to have a high package coplanarity up to the reflow temperature, in order to achieve a high board assembly yield, especially for PoP applications. Lowest warpage changes over temperature are provided by eWLB.

- *System in Package:* FO-WLBs are best suitable for system integration, as described in the next section.

Figure 14 shows a general overview of FO-WLB package properties.

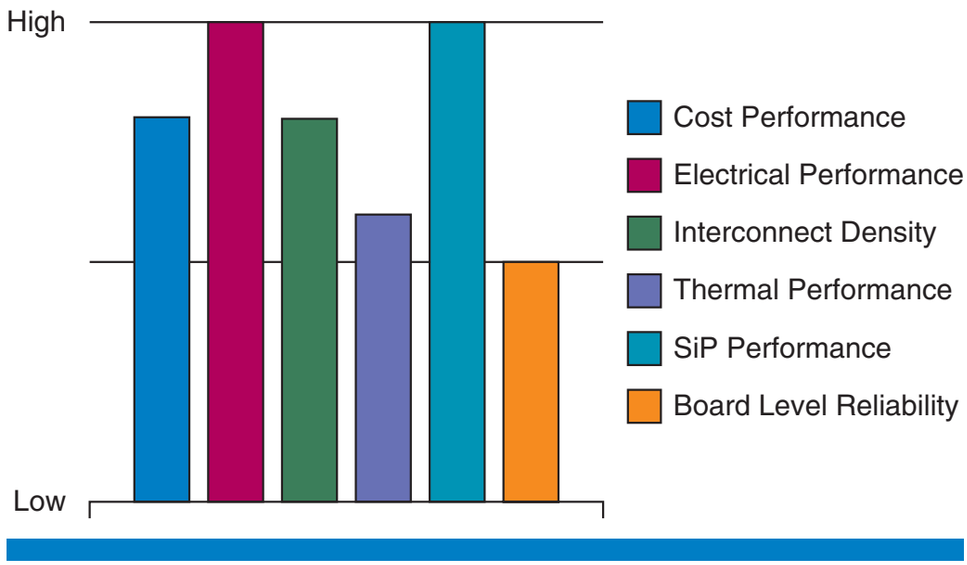


Figure 14: Performance rating of eWLB technology (higher performance indicates better values) (Source: Intel, 2014)

"FO-WLBs are best suited for System in Package (SiP) applications by placing two, three, or more (active) dies in one package,…"

System Integration Capabilities

FO-WLBs are best suited for System in Package (SiP) applications by placing two, three, or more (active) dies in one package, for example, baseband, power management unit, and an RF module. Integration of passive devices like SMDs and IPDs saves significant space on the application board and improves electrical system performance. Multiple components can be placed side by side above the die, or face to face on the active chip area without increasing z-height. Furthermore coils, capacitors, resistors, and especially antennas can be directly applied in the redistribution layers. Coils and antennas in particular can be realized in the fan-out area with an excellent quality factor.

FO-WLBs are also suitable to be used for package-on-package (PoP) stacking. Typically, a memory package is placed on the back side of the bottom package, which often contains a processor or another logic chip. Top to bottom connection can be realized by laser drilled through mold vias or by placing prefabricated contact bars in the fan-out area of the bottom die. Also, by implementing a backside RDL with landing pads and connection to the front

side, area array can be achieved. The very stable coplanarity over temperature is another big advantage of this technology for package stacking. Figure 15 shows a cross-section of an eWLB PoP package with a memory BGA stacked on top.

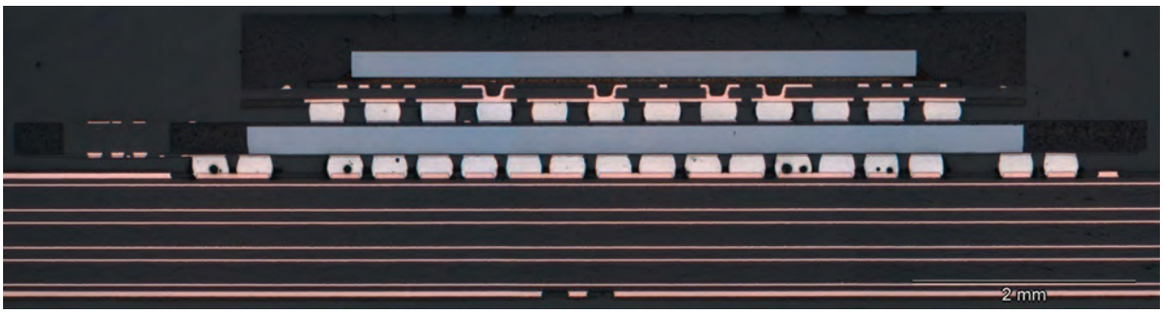


Figure 15: Cross-section of an eWLB PoP package (bottom) with top memory BGA (Source: Infineon, 2009)

Future

Packaging cost reduction is a major target for FO-WLB technologies in general. Therefore, the switch from round wafer format to rectangular panel format of the cost-effective PCB process technology will be a next step. Even with PCB-technology-related lower pattern density and a possible need for additional redistribution layers, the lower process costs and the much higher process parallelism will improve the costs per package significantly.

"…the switch from round wafer format to rectangular panel format of the cost-effective PCB process technology will be a next step."

Continuous shrinking of the silicon nodes leads either to reduced die sizes or to higher I/O counts due to increased functionality. In both cases, a need to switch from the die size limited fan-in WLBs to the much more flexible FO-WLB technology arises.

The future focus will move from the single-die package, using one redistribution layer, to highly integrated systems in packages (SiPs), using multiple routing layers on the package front side and, for package stacking technology (PoP), also additional routing layers on the package back side. This will speed up the integration of sensors and MEMS close to the application processor or at least to bundled modules.

Embedded Die Package

For a few years embedding in laminate has also played a growing role within the big variety of different packaging technologies. Nearly all OSATs (outsourced assembly and test) and/or laminate suppliers offer the embedding of passive components or active and passive components into laminate. Actually, chip embedding into laminate is in an early high volume production status with a few products.

"Actually, chip embedding into laminate is in an early high volume production status with a few products."

"Different principle flows with different cost structures and, most important, different design rules, are available on the market."

Construction and Examples

In general, there are two ways of embedding components into laminate: The first way includes the formation of a passive within the routing layers of the laminate itself. Resistors or capacitances are structured on a laminate area. In addition, the laminate metal layers can be used as chip shielding. No additional process steps but co-design preparation of chip and package are required.

The second way is to place and embed components into the core material of the laminate or between any other of the internal layers. The components will be partly or completely buried and contacted by laminate and redistribution layers applied later. Connections are mainly realized by laser drilling and electroplating or similar processes. Different principle flows with different cost structures and, most important, different design rules, are available on the market. The main difference between these flows is the required pad dimension and pitch of the chip (see Figures 16, 17, and 18).

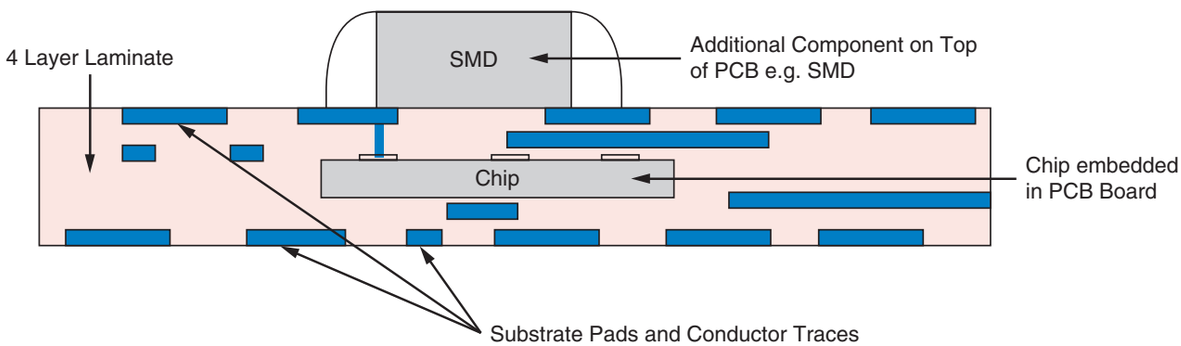


Figure 16: Schematic drawing of an embedded chip in PCB with SMD on top of PCB (Source: Intel, 2014)

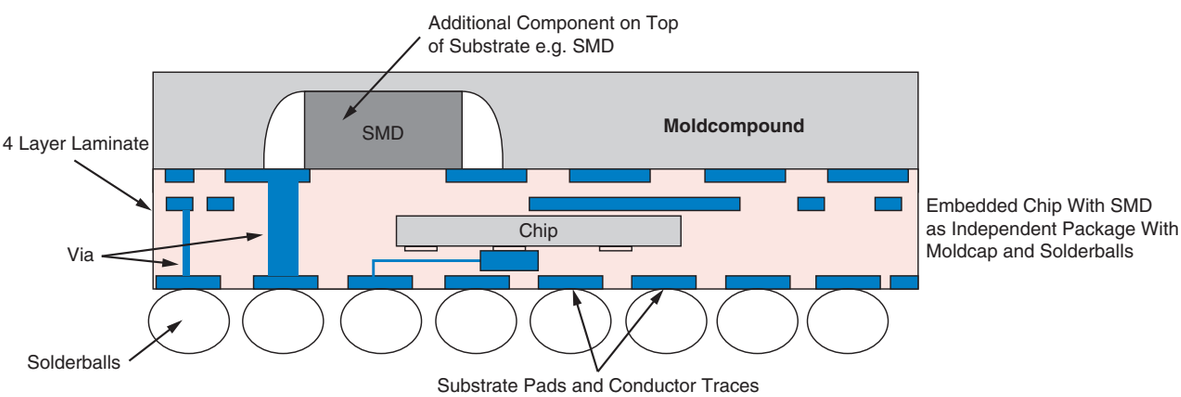


Figure 17: Schematic drawing of an embedded chip with SMD and mold cap (Source: Intel, 2014)

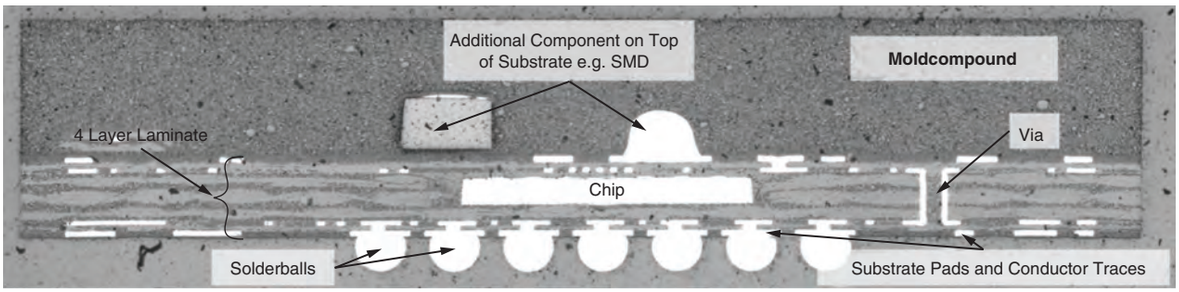


Figure 18: Cross-section image of an embedded chip in laminate (Source: Intel, 2013)

Figure 19 shows a schematic of the process flow and supply chain.

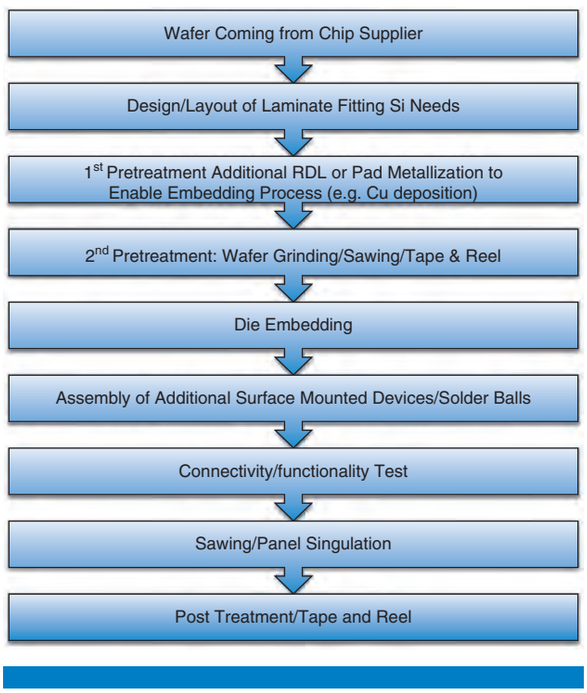


Figure 19: Schematic of the process flow and supply chain (Source: Intel, 2014)

Advantages, Tradeoffs and Use

Embedded products are already in use within some mobile devices like smartphones, wearables, and medical devices. The first products with embedded components on the market are low-cost products with a low I/O count and a small chip or a low number of embedded passives, for example a DC-to-DC converter.

"Embedded products are already in use within some mobile devices like smartphones, wearables, and medical devices."

- *Cost performance:* The packaging cost of an embedded die may be higher than for other platforms. Embedding of chips and passives require a complex

process flow, may not have established supply chains, and may have a PCB-technology-related higher yield loss. The yield loss will cause the scrap of good chips, increasing the package cost for large chips and modules significantly. But recent figures of major embedded die suppliers show already strongly increased yields for the technology. The embedded device has to have a specific pad termination to fit to the production flow of the laminateprocessing, which adds process steps and costs. Additionally, this requires a complex supply chain between chip supplier, laminate supplier, and OSAT houses.

- *Electrical performance:* The interconnection length between multiple components in the package can be reduced to a minimum. It is only determined by the thickness of the covering layers of the embedding material. These short connections can lead to less parasitic and high performance.

"…A major restriction of today's embedded die technology is the limited pad pitch and pad size suitable for embedded die technology."

- *Interconnect density:* A major restriction of today's embedded die technology is the limited pad pitch and pad size suitable for embedded die technology. The reason is the limited accuracy of the PCB technology. Depending on the given I/O count and pad pitch of the component the right embedding technology flow has to be used (see "History, Construction, and Examples" earlier in this section).
- *Thermal performance:* Embedded die packages have a medium thermal performance, since the die is embedded in polymer material with high thermal resistance. Being placed close to metal structures, the heat transfer may be improved.

"…PCB technology inherently offers solderable top and bottom interfaces as well as various types of vias. stacking capability is given by default."

- *System in Package (SiP) performance:* PCB technology inherently offers solderable top and bottom interfaces as well as various types of vias. Therefore stacking capability is given by default. The top and bottom laminate surfaces of an embedded die package are conventional laminate surfaces and can be equipped with additional active or passive components.
- *Board-level reliability:* Embedded die packages are showing a high reliability performance due to the protective effect of the embedding material surrounding chip and chip connections.

Figure 20 shows the general overview of embedded die package properties.

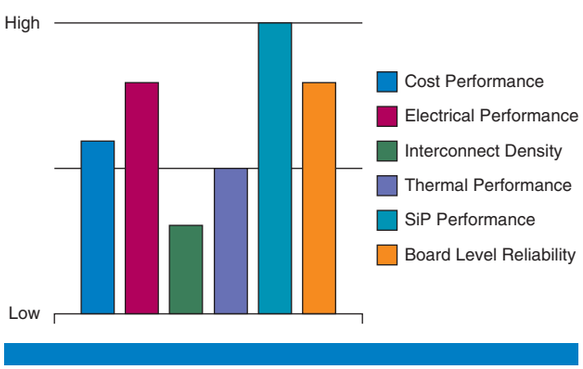


Figure 20: Performance rating of embedded die technology (higher performance indicates better values) (Source: Intel, 2014)

System Integration Capabilities

From system integration perspective, the embedded die technology is very interesting. PCB technology, which forms the basis of the embedded die technology, offers multiple routing layers, different via capabilities, and solderable top and bottom interfaces by default. By adding active and passive components within the laminate, module suppliers can design very compact and highly efficient modules. Standard components can be attached on top or bottom, having the shortest connections between the different parts. Passives and actives can be embedded in nearly any combination in, on, and below the laminate, including side-by-side placement of multiple components. In addition, passives can be implemented in any routing layer in the multilayer structure. Even stacking of laminates with embedded chips onto each other is possible.

"PCB technology, which forms the basis of the embedded die technology, offers multiple routing layers, different via capabilities, and solderable top and bottom interfaces by default."

Future

In the future the market for embedded dies will increase significantly. System integration will be a big topic of future mobile package products. Available PCB area in the products will decrease due to different product target dimensions (as in the case of wearables) or increased integration of functionality (as in the case of sensors).

Currently only a few passive or active components are embedded into one layer of the same laminate. The number of components per layer will increase and stacking within the different layers of one substrate will be realized in the future too. In addition, different laminates with embedded components can be stacked as different packages.

"The number of components per layer will increase and stacking within the different layers of one substrate will be realized in the future too."

One limiting factor these days is the large pad pitch and the pad size. Depending on the technology used, the values for pad pitch in volume production are limited. These values have to be reduced for use in highly sophisticated technology nodes. Also, developments are ongoing to avoid the special pad termination (today copper is required) on the chip.

Another important future target is heterogeneous integration of different frontend technology nodes into one package. Embedded die technology is well suited to this and will be one future packaging platform of choice.

System Integration on the Package Level

The increasing demand for new and more advanced electronic products with superior functionality and performance is driving the integration of functionality for future packaging technologies.

"The increasing demand for new and more advanced electronic products with superior functionality and performance is driving the integration of functionality for future packaging technologies."

Traditionally this was reached by downscaling the dimensions, enabling the integration of an increasing number of transistors on a single chip (Moore's law). However, many quantitative and functional requirements do not scale with Moore's law. "More than Moore" is required, such as, for example, the integration into systems. System integration generally is defined as the process of bringing together the component subsystems into one system and ensuring that the subsystems are functioning together as a system. There are general advantages

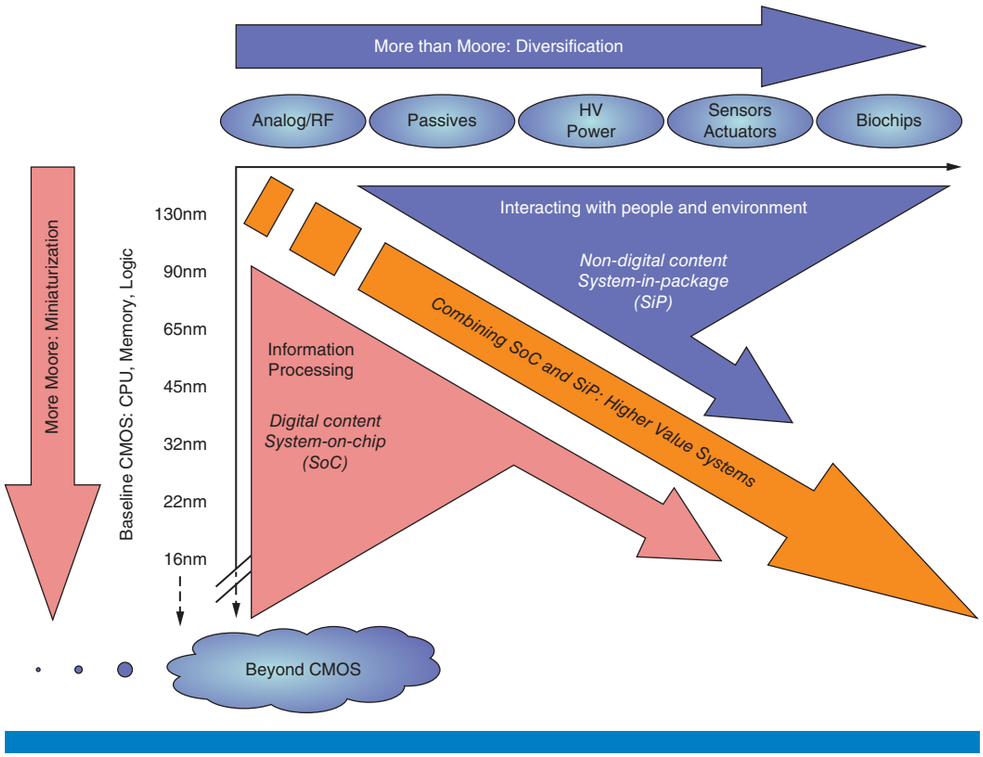


Figure 21: Moore's law and more
(Source: ITRS, 2013)

of system integration, like higher electrical performance, smaller form factor, or the possibility of heterogeneous integration. Also, time to market, one of the important criteria of the mobile market (see the "five Ps" discussed in the next section) can be reduced. And, the increasing number of passives, sensors, and new requirements for devices such as wearables will drive the system integration.

Technically, the possibilities of integration are manifold. Horizontal and vertical integration can be differentiated. Horizontal integration describes the in-plane arrangement of the subsystems. It is a cost-effective solution with the tradeoff of large lateral dimensions and long distance interconnections between the subsystems.

"There are multiple ways to generate a vertically stacked approach, differing by the interconnect method and the achievable dimensions and performance."

In order to provide shorter interconnection length, beneficial for electrical performance, vertical integration is used. There are multiple ways to generate a vertically stacked approach, differing by the interconnect method and the achievable dimensions and performance.

For the previously discussed package platforms, different integration solutions are available. Any of them has advantages in certain areas, for example cost, integration capability, or reliability. There is no general packaging solution; the product requirements always define the optimal solution. For example, die stacking can either be realized traditionally by flip chip and wire bond

interconnection or with TSV (and a flip chip connection). Figure 22 shows the two options. Both have different advantages and disadvantages (such as cost, performance, dimensions), which have to be considered for the decision of the integration technology.

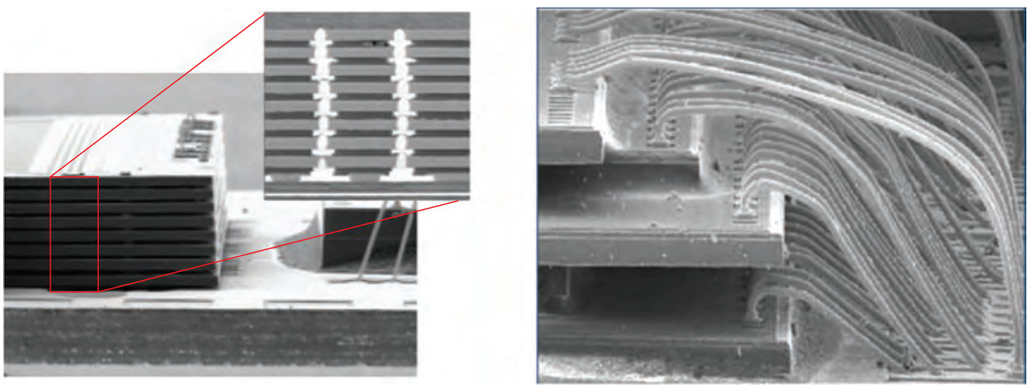


Figure 22: Samsung 16-Gb NAND stack with TSV and four-die wire-bond stack-up (Source: Samsung, Palomar, 2011)

Another example of different solutions for a similar integration problem is the 2.5D Interposer vs. an eWLB-FC (embedded WLB flip chip) package. Both address the need of a package solution for chips with very tight pad pitch. The silicon interposer offers the tight pitch on its top side for microbump connection of the chip. Through TSVs and further routing layers, the pitch at the bottom of the silicon interposer is enlarged and therefore suitable for standard flip chip connection on a BGA substrate. The solution is capable of very high accuracy, and the mechanical stress on the die is very low, but cost will be very high.

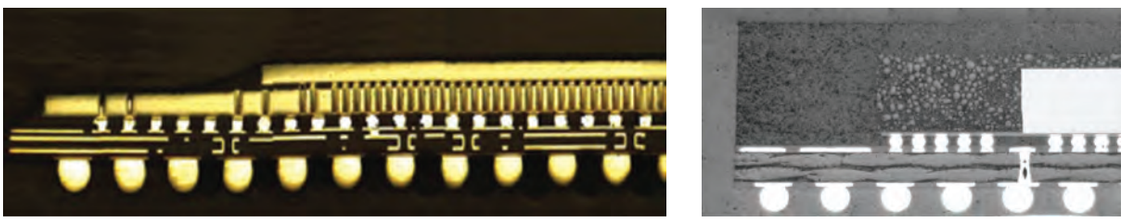


Figure 23: Silicon interposer and eWLB-FC (Source: Intel, 2013)

On the other side, the eWLB-FC package offers fan-out area for additional bumps allowing more relaxed design rules for the bump pitch. A RDL distributes the tight chip pad pitch into a suitable bump pitch for standard BGA substrate. The stress level on the die may be higher and realizable pad pitches may be lower than for the silicon interposer solution. But also the packaging cost will be much lower. So, again, different solutions are available for a similar problem.

Since the requirements of mobile products change very quickly, it is important to focus on the development of building blocks for system integration. Depending on the specific product requirements, the best suitable integration technology can be selected.

Figure 24 gives an overview of the system integration capabilities of the previously discussed packaging technologies.

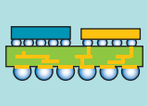
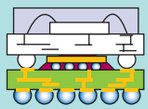
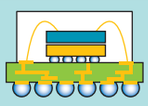
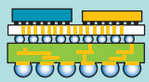
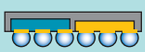
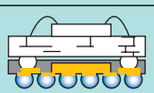
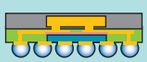
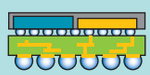
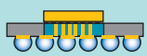
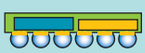
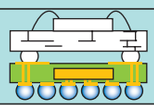
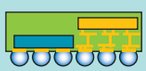
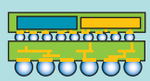
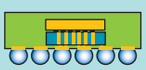
	Horizontal	Vertical			
Example	Side-by-side	Package on Package Stacking	Traditional Stacking	2. 5D Interposer	3D TSV Stack
Flip Chip Based					
WLB Based					
eWLB Based					
Embedded Die Based					
Description	Side-by-side Placement on/in Carrier	Stack of Multiple Finalized Packages	Die Stack With Wire Bond, Flip Chip or Thin-film Connections	2. 5D Side-by-side Integration on Substrate	Vertical Stacking With Connections Through Silicon Body
(Possible) Use for...	• Multi-chips With no Lateral Size Constraints	• Memory on Logic Stacks (mobile)	• Memory Stacks • Memory on Logic Stacks	• Complex Stacks With no Size and Cost Constraints • (Hi-end Front-end Technology)	• MEMS and Sensors • (Wide IO DRAM) • (Memory Cubes)

Figure 24: System integration capabilities
(Source: Intel, 2014)

Solutions for Mobile Applications

The existing mobile application space is not homogeneous; in fact this area is quite fragmented. Packaging has to support applications from low-end voice phone via tablets to high-end feature phones. The advanced features, realized in these devices, reflect the complex technology behind them.

Packaging Requirements for Mobile Applications

In recent years the mobile market has become more and more the driving factor for compact semiconductor packages with increased performance and higher complexity at the lowest possible cost. Figure 25 shows the relation of smartphone functionality and high density packaging.

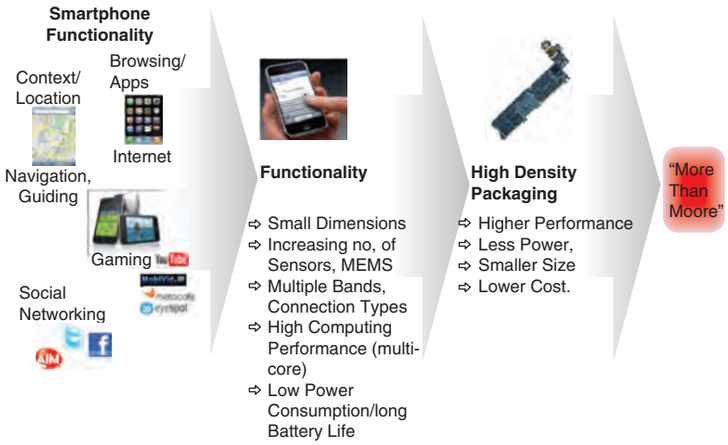


Figure 25: Complex technology enabling mobile features (Source: Intel, 2013)

On the other side, the gap between the speed of IC scaling (Moore's law) and the rather slow advancements in scaling of printed circuit boards (PCBs)—also known as the "interconnection gap"—needs to be addressed as well.

This environment of small form-factor (SFF) devices and the existing Interconnection Gap provide the unique opportunity for packaging to become a key differentiation factor in the whole product design flow. Figure 26 illustrates the interconnect gap.

"…the gap between the speed of IC scaling (Moore's law) and the rather slow advancements in scaling of printed circuit boards (PCBs)—also known as the "interconnection gap"—needs to be addressed…"

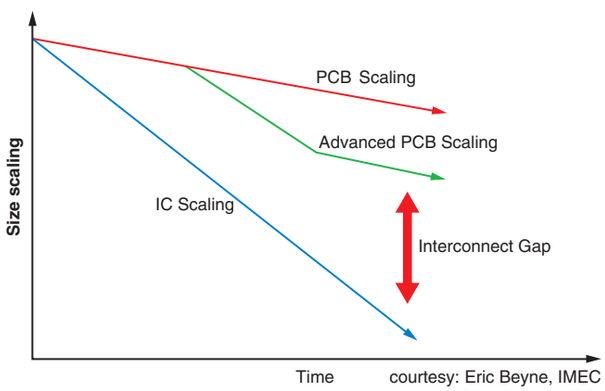


Figure 26: Interconnect gap (Source: ECE 407/ 507 University of Arizona (http://www.ece.arizona.edu/mailman/listinfo/ece407))

Key Challenges for the Next Generation Mobile Platforms

Besides the project execution (time to market), performance, power, PCB area, and price are the key challenges in this arena, (the so-called "5 Ps").

"There is no universal package solution…"

There is no universal package solution to get the optimum result out of these five challenges, rather a constant reiteration and close co-design approach, combined with a deep knowledge of system partitioning and system integration, will provide the optimum result for each application segment.

Although the technology plays an important role for enabling, it must be always kept in mind that price (cost) as the ultimate driver will dictate the primary direction most of the time.

Figure 27: Key challenges—the 5 Ps (Source: Intel, 2013)

"…there will always be a consideration between System-on-Chip (SoC) and System-in-Package (SiP)."

"From a packaging point of view, SoC provides the ultimate single-die package, where Moore's law will be the primary driving factor."

The Optimum Package Solution for the Mobile Space

Besides the key challenges there will always be a consideration between System-on-Chip (SoC) and System-in-Package (SiP).

Actual SiP or SoC package concepts include new developments like wafer-level and panel-level packages, embedded dies and through silicon/ mold via, but also classic technologies like flip chip packages and their capabilities in system integration (as described in previous sections).

From a packaging point of view, SoC provides the ultimate single-die package, where Moore's law will be the primary driving factor. SiP on the other side, opens up all possibilities in combining 2D/2.5D/3D package concepts to find the optimum solution for the specific application, which could be the smarter solution on the system level, but which also adds significantly to package complexity. This in turn always requires different paths on the available packaging roadmap, such as Moore's law or miniaturization, cost or price performance, and system integration and/or performance, as shown in Figure 28.

Therefore the optimum solution can never be a specific package platform; rather it's an application-specific smart choice of "bricks" to be combined.

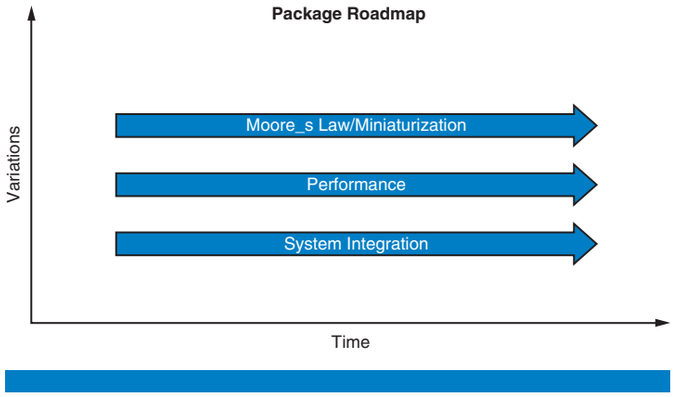


Figure 28: Required paths on the package roadmap
(Source: Intel, 2013)

In choosing the optimum package solution, it's mandatory to consider the five Ps as a basis for the decision as well. For the package decision, we will focus only on the following four criteria: price (respective cost), performance (including electrical, thermal, and reliability), PCB (footprint, needed board density), and project execution (or time to market). Even this depends on many other factors like reuse of IP, chip and package design, package roadmap, useable package portfolio, but of course also SiP and/or SoC capability and access to different technologies (as described in previous sections).

The criterion power will be not considered here, because it is not primarily driven by the package.

As a generic statement, there must always be a balance between performance and PCB area versus price, as these properties normally are in opposition.

Based on experience, the decision flow for the optimum package solution can be described efficiently in the flowchart shown in Figure 29.

Starting from the first product idea, customer requirement, product roadmap, and feature set, it will be decided which IP macros will be needed and which FE node will be suitable for it. This partitioning will lead to parameters like chip size, chip I/O count (power, ground, signals), and I/O density. With this first set of parameters, a selection of possible package platforms and package ideas can be initiated. To follow a structured approach here, the chart in Figure 30 can be used.

If the pre-selection is done for one or more possible options, the next step can be triggered: execution of feasibility studies and assessments.

This step includes work packages like performance check (thermal, electrical, RF), thickness/footprint requirements, reliability requirements, co-design (chip, package, PCB), test requirements, needed interfaces (such as memory, and so on) and, of course, a cost estimation for all options.

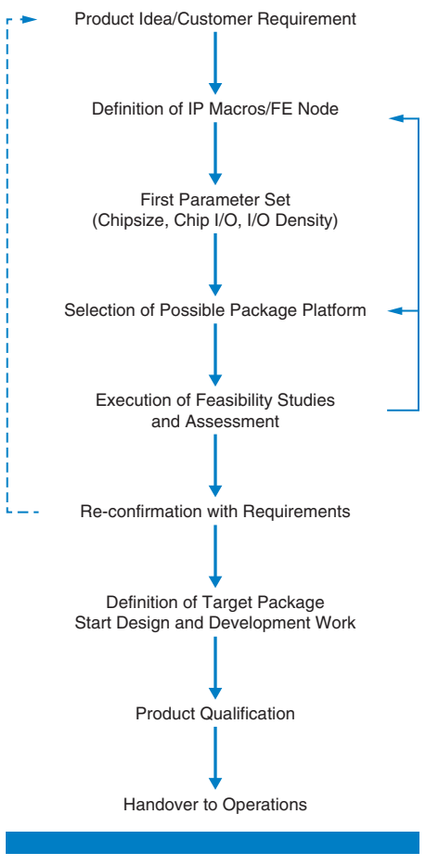


Figure 29: Package concept engineering flow
(Source: Intel, 2013)

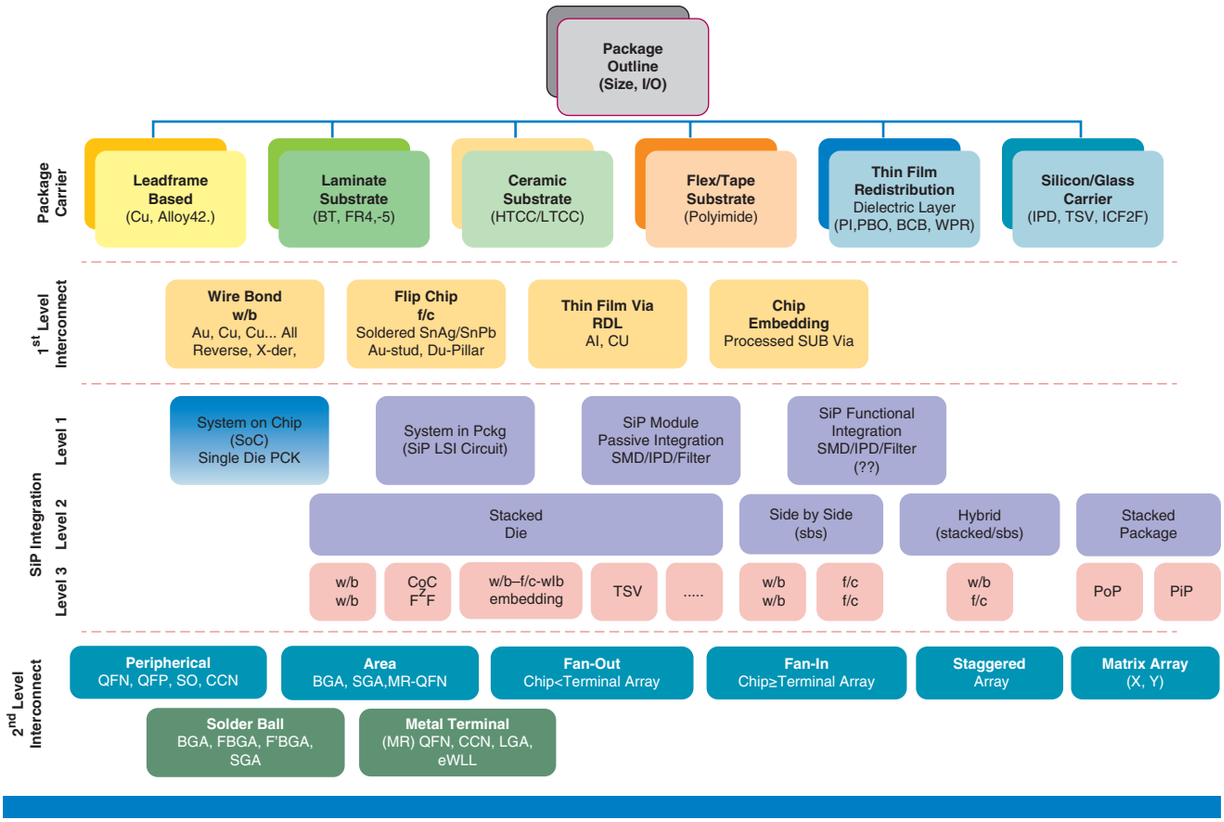


Figure 30: Morphological box for package selection process (Source: Intel, 2012)

A final decision is based on the five Ps (respective four Ps) and an overall risk assessment.

"If no optimum solution can be found, reiteration cycles have to be run until the optimum solution is finally reached."

If no optimum solution can be found, reiteration cycles have to be run until the optimum solution is finally reached. After reconfirmation of the requirements, the target package is defined and the design and development work can start.

Within the given time to market the product will be qualified and the project handover to operation can be done. With the given flexibility there will be an optimum solution, if the five Ps are constantly considered during the whole concept phase.

In summary, the key to a successful package technology selection can be described as follows:

- Evaluate real customer requirements
- Look for reuse/standardization
- Start co-design as soon as possible
- Focus from beginning on design to cost

The Future of Mobile Packaging

The pace of change in packaging technology today has accelerated to the highest rate in history. The technology cycles have tended to last ten years in the past, but this cycle seems to actually be shortening.

"The pace of change in packaging technology today has accelerated to the highest rate in history."

Mobile computing is a growing market; mobile electronic products are being deeply woven into our daily life. The number of mobile phones, tablets, and ultra-mobile products (thin and light clamshell PC designs and hybrid devices) is increasing, while PC sales are dropping (see Figure 31).

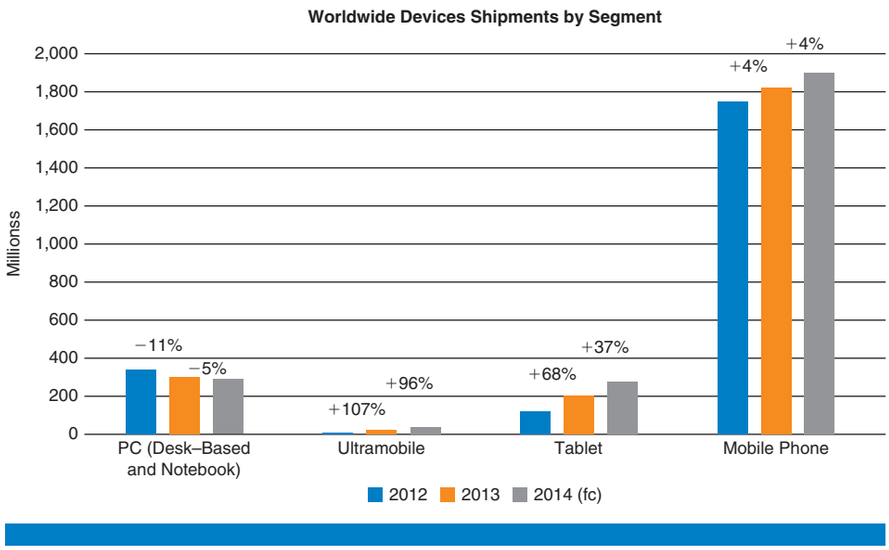


Figure 31: Worldwide device shipments (Source: Intel, 2014)

A clear trend for the future is that consumers want anytime-anywhere computing that allows them to consume and create content with ease, but also share and access that content from a different portfolio of products.

"Wearables may be the next stage of mobile devices."

Wearables may be the next stage of mobile devices. The idea is not new. Mechanical wearable devices have been around for a long time. But now, electronics find their way into rings, wristbands, and glasses (see Figure 32).

So far, the focus has been on fitness tracking and supporting, but soon the transition to medical support and implantables may also happen. Packaging requirements may differ due to special requirements of small, bendable, and twistable wearables. Large integrated SoCs may be split into smaller units, in order to be able to place small dies in defined areas of the wearable. New interconnect methods and flexible packaging will be required.

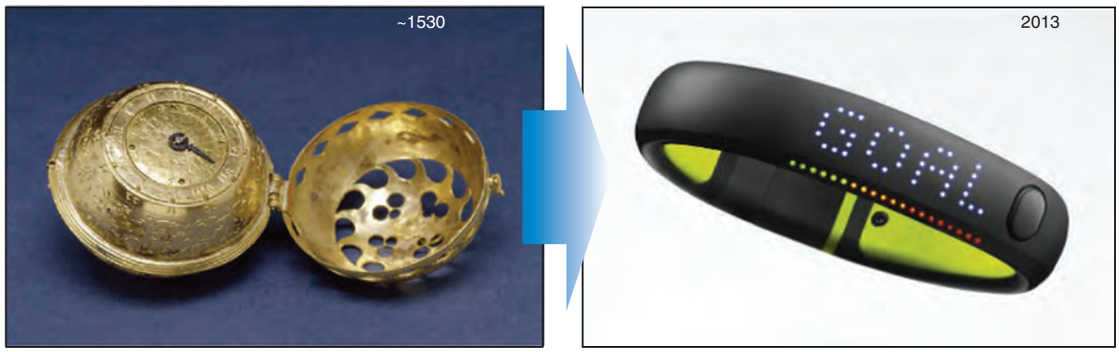


Figure 32: Wearables: An early mechanical watch (c. 1530) and Nike Fuelband* fitness wristband (Source: Wikipedia, Nike Fuelband, Company website, 2014)

Some general important future requirements of packaging for mobile products include:

- Continued cost reduction of chip and package
- Data bandwidth increase
- Increased battery life—a critical issue
- Small and smaller chip/package dimensions
- Increasing sensor integration

This will require continued development in the packaging technology area.

Author Biographies

Thorsten Meyer is IMC Principal Wafer Level, working on the next generation's packaging technologies for mobile applications in Regensburg, Germany. He earned his engineering degree (Dipl.-Ing.) in Production Engineering from the University of Erlangen-Nürnberg. Thorsten is the author of many publications in the area of advanced packaging and has received four best paper awards. He holds more than 130 patents and patent applications. Email: Thorsten.meyer@intel.com

Sven Albers is a project manager in backend technology development at Intel Mobile Communications. He received his engineering degree (Dipl.-Phys.Ing) at the University of Essen in 1997. Sven joined Infineon 1998 where he lead different frontend technology development projects before moving to Intel in 2011 where he is working on the next generation's packaging technologies for mobile applications in Regensburg, Germany. Email: Sven.albers@intel.com

Christian Geissler is a project manager in backend technology development at Intel Mobile Communications, working in the Package Technology and Innovation Group of IMC in Regensburg, Germany. He received his engineering degree (Dipl.-Phys.) at the University of Regensburg in 1997 and joined Siemens/Infineon in 1997. He worked as device expert and project

manager on different frontend and backend technology development projects until he moved over to Intel in 2011. Email: Christian.geissler@intel.com

Gerald Ofner is director for package technology and innovation within Intel's Mobile and Communications Group. Gerald received his engineering degree (Dipl. Ing.) from the University of Linz (Austria) and the ETH Zurich (Switzerland) and joined Siemens in 1997. He contributed on several engineering and managing positions at Siemens Semiconductor, Infineon Technologies, and Intel Mobile Communications on packaging solutions for mobile applications. Gerald has filed more than 100 patent applications in this field. Email: Gerald.Ofner@intel.com

Klaus Reingruber is staff engineer for package development, working in the Package Technology and Innovation Group of Intel Mobile Communications in Regensburg, Germany. He earned his engineering degree (Dipl. Phys.) in physics working in the field of III-V compound semiconductors from the Friedrich-Alexander University Erlangen Nürnberg in 1995. Klaus is coauthor of several papers on opto-electrical research in III-V compounds. Before he joined package development, he spent over ten years in silicon wafer technology as a process engineer and also as an integration engineer. Email: Klaus.reingruber@intel.com

Georg Seidemann is senior staff engineer for package development working in the Package Technology and Innovation Group of Intel Mobile Communications in Regensburg, Germany. He is a specialist on analytic topics. He earned his engineering degree (Dipl.Phys.) in physics working in the field of thin film electromigration investigation on TEM from the University Regensburg in 1995. Before he joined the package development group, he worked for 13 years in silicon wafer technology as a development and integration engineer in BEOL metallization. Email: Georg.seidemann@intel.com

Andreas Wolter is project leader in backend technology development at Intel Mobile Communications. He received his PhD in Physics at the University of Karlsruhe in 1995. After a period of postdoctoral research at the Commission à l'Energy Atomique in Grenoble (France), he moved to the PCB industry where he worked as delegate to the Fraunhofer Institute for Reliability and Microintegration IZM. Andreas joined Infineon in 2000. With the acquisition of Infineon's wireless business he changed to Intel in 2011. His projects were in the fields of flip chip on board, bare-die testing, through silicon via interconnection, and eWLB. Email: Andreas.wolter@intel.com

OVER-THE-AIR TESTING FOR 4G SYSTEMS

Contributors

Nicolas Obriot
Agilent Technologies

Moray Rumney
Agilent Technologies

Janus Faaborg
Agilent Technologies

Michael Dieudonne
Agilent Technologies

It is perhaps a surprising fact that the performance of the wireless devices that now dominate much of our connected lives has been measured using wired connections which bypass the device's antennas. For many years, such test methods have been simple and sufficient to predict how the device will operate in wireless "over the air" (OTA) conditions, but in recent years this test simplification is no longer appropriate to predict user experience in a real network. The first wireless devices operated in a single radio frequency band and with an external "whip" antenna. Such simple designs were almost guaranteed to work, which is why cabled tests were sufficient. However, in recent times antennas have become integrated and there has been a need to support multiple frequency bands. The consequence is that the volume available per antenna is far smaller than it used to be and this has often led to a loss of radiated performance that was not picked up by cabled testing. This article discusses the evolution of wireless test from traditional cabled methods to the newer radiated methods that are being developed to keep up with 4G standards incorporating MIMO and active antennas.

Introduction

During the last ten years, mobile communications and networks have seen dramatic growth. The mobile subscriber penetration rate is typically above 100 percent in most developed countries. Data traffic volumes continue to rise as users demand access to information anytime, anywhere.

To guarantee a good user experience, the radio transmission from the base station to the mobile device has to meet demanding requirements since any weakness will affect the user's experience. The role of the antennas in mobile devices, and in particular in data-hungry smartphones, is therefore very important and directly impacts user experience. In earlier times, the mobile phone antenna was external to the device. The integration of the antenna inside the phone was seen as progress from an industrial design perspective but this evolution has not been helpful from a radio performance perspective. For example, if the user places his or her hands on certain parts of the device, it will affect the antenna impedance and absorb some of the RF energy, which impacts the device performance. This dependence on hand proximity makes antenna design much more complex, especially with integrated antennas. Also, in the pursuit of enhanced user experience and higher data throughput, techniques such as the use of multiple frequency bands or the use of multiple antennas have further complicated antenna design.

The first OTA measurement procedures were developed by the Wireless Association (CTIA) in 2001 for devices with single antennas known as single-input single-output (SISO) devices. More recently, the use of multiple antennas at the base station and mobile device, called multiple-input multiple-output (MIMO), has become common. Consequently, MIMO OTA measurements and simulations for network and mobile device performance evaluation and prediction have become important research topics. MIMO is used here to refer to any multi-antenna technique including Rx and Tx diversity, beam steering, and spatial multiplexing.

Research into MIMO OTA for standardization purposes has been ongoing in the CTIA since 2007, the Third Generation Partnership Project (3GPP) since 2009, and the European Cooperation in Science and Technology (COST) since 2008. This work was motivated by the need to develop accurate, realistic, and cost-effective test methods for the MIMO-capable devices specified in the 3GPP Universal Mobile Telephone System (UMTS) and Long Term Evolution (LTE) standards.

Although many MIMO-capable networks have been deployed since 2010, the first stable MIMO test methods only emerged towards the end of 2013. The development of MIMO OTA test methods has proven to be much more complex than the much simpler methods developed for SISO systems available today.

"Although many MIMO-capable networks have been deployed since 2010, the first stable MIMO test methods only emerged towards the end of 2013."

Unlike SISO OTA, which was relatively straightforward and purely a function of the device, MIMO OTA is highly dependent on the interaction between the propagation characteristics of the radio channel and the parameters of the receive antennas of the mobile device.

"MIMO OTA is highly dependent on the interaction between the propagation characteristics of the radio channel and the parameters of the receive antennas of the mobile device."

Consequently, the existing SISO measurement techniques are unable to test the MIMO properties of mobile devices. Many different MIMO test methods have been proposed, which vary widely in their size, cost, and ability to emulate specific propagation channel characteristics.

The most recent standardization activities have investigated whether the candidate test methodologies provide the same results, since it is essential to clearly differentiate good and bad MIMO device performance independently of any approved test method.

The goal of this article is to provide a valuable source of information for the state of this important research area. The next section, "A Brief History of Conducted and Radiated Testing," gives a history of OTA testing, followed by "MIMO OTA Testing Challenges," a section that describes the latest challenges facing MIMO OTA. The candidate test methodologies recently approved by 3GPP for MIMO OTA are then introduced and described in "Selected Test Methodologies," and finally a comparison of follows in "Methods Benchmarking" before the conclusion.

A Brief History of Conducted and Radiated Testing

Previously, mobile device testing used to be done using the conducted method. This involved connecting a cable to what is called the "temporary antenna connector," which bypasses the device-under-test (DUT) antenna to provide direct access to the transceiver. The validity for such a technique is based on the assumption that the DUT antenna can be fairly represented by an isotropic antenna (equal gain in all directions) with 0 dB gain. This is known as a 0 dBi antenna. In the days when the antenna was a dipole tuned to a single band, this assumption was reasonable, thus avoiding the need for radiated (OTA) testing using a large and expensive anechoic chamber.

With the advent of multiband integrated antennas, the 0 dBi assumption is no longer valid. Using a cable connection bypasses the actual antenna, which may have characteristics very different from the 0 dBi assumption. In this case it is easy to see how the results from conducted tests for requirements such as reference sensitivity and maximum output power may no longer represent the radiated performance of the device in a real network.

"The advantage of cable-conducted measurements is that the measurements themselves are accurate and cost effective."

The advantage of cable-conducted measurements is that the measurements themselves are accurate and cost effective. The uncertainty is measured in tenths of a decibel but the problem is the results are not always representative of the radiated performance seen in the network. When switching to SISO OTA testing, the results are representative of real-life conditions but the uncertainty is higher being around ±2 dB.

Moving towards OTA Measurements

Radiated testing for regulatory purposes started with electromagnetic compatibility (EMC) testing for spurious emissions, and more recently a hearing aid compatibility test was added and a safety test called specific absorption ratio (SAR) to assess how much of the DUT radiated power is absorbed by a phantom head. However, these tests measure unwanted side effects and do not assess the desired radio performance of the DUT for the purpose of communication.

"OTA tests have now been developed to predict communication performance in a real network."

OTA tests have now been developed to predict communication performance in a real network. The normal process in standardization is to simulate the desired performance and then set performance targets for developers. Since OTA performance requirements were very much retrospective, simulation was not an option. Also, the complexity of trying to simulate a realistic performance target was considered out of scope. However, once an accurate test method had been agreed upon and developed, the performance requirements were instead developed through a series of measurement campaigns using real devices. This is clearly not ideal, but under the circumstances this was the only practical solution for SISO OTA testing.

SISO OTA Testing History

SISO OTA testing is conceptually simple, comprising one Figure of Merit (FoM) for the DUT transmitter called Total Radiated Power (TRP) and

another for the DUT receiver called Total Reference Sensitivity (TRS). For TRS, CTIA uses the term Total Isotropic Sensitivity (TIS). The TRP is defined as the integral of the power transmitted in different directions over the entire radiation sphere. TRS is a similar measure, but it represents the average reference sensitivity of the DUT receiver.

With these two FoMs agreed upon, the bulk of the standards work for SISO OTA was in defining the details of the method, the test system uncertainty, and finally the performance requirements. The first SISO OTA test method was developed by CTIA and was based on an anechoic chamber. Substantial theoretical analysis of the measurement uncertainty was performed by CTIA and the European COST273 project, resulting in an error model with over 20 terms. The requirement for overall test system uncertainty was calculated to be in the region of ±2 dB; This figure has since been validated using a golden radio by CTIA accredited labs. Further work on an alternative test method was done by 3GPP using a reverberation chamber and test results indicate a similar level of uncertainty.

"…the bulk of the standards work for SISO OTA was in defining the details of the method, the test system uncertainty, and finally the performance requirements."

Agreeing to minimum requirements in 3GPP was not straightforward. Network operators wanted to set high performance targets, while mobile vendors needed to protect the installed base of existing mobile devices and existing designs from becoming obsolete, and also to maintain design margins for the ever-decreasing size of future SISO devices. The end result was necessarily a compromise that took some three years to agree to due to the many factors that had to be considered such as primary frequency bands, mechanical modes, and operating positions.

The 3GPP normally relaxes the minimum DUT requirements by the full test system uncertainty, but since the OTA test uncertainty is quite high (±1.9 dB for TRP and ±2.3 dB for TRS), the relaxation is limited to about half of the allowed test system uncertainty. This choice prevents the requirements from becoming too relaxed, which would allow some bad DUTs to pass, but it does slightly increase the risk that a good DUT might fail the test.

"…since the OTA test uncertainty is quite high (±1.9 dB for TRP and ±2.3 dB for TRS), the relaxation is limited to about half of the allowed test system uncertainty."

Figures were agreed on for minimum performance that allowed the bulk of legacy devices to remain compliant. Device vendors also accepted tougher, though not mandatory, "recommended" average performance—typically 3 dB better—to give the industry something to aim for. Even then the recommended TRP performance is some 6 dB below the nominal power for the conducted test, suggesting that there is still considerable room for improvement in antenna performance, although with the continual downward pressure on the space available within devices for antennas, such improvement may not be realistic.

Device performance is measured at the low, middle, and high channels of all the frequency bands supported by the DUT and at two orthogonal RF polarizations, such as vertical and horizontal. The DUT has to be tested in its primary mechanical mode, which may involve sliding or folding open the DUT. Testing in other mechanical modes is not part of the minimum requirement, although

some network operators require all modes to be tested. The final consideration is the physical environment. Tests are carried out in free space or in the proximity of a specific anthropomorphic mannequin phantom head, better known as SAM. Tests are carried out on both the left and right sides of the head, which is filled with different liquids to match the frequency-dependent RF loading effect of the human head. In order to ensure repeatability across labs, detailed guidelines on how to align the DUT to the test environment are provided. The latest CTIA test plan has added a phantom hand, which can take four positions depending on the DUT design: monoblock, fold, narrow data, and PDA. The hand may be used on its own for "data" positions or in conjunction with the head to emulate a more realistic speech position than the head only tests. Currently, testing is also performed with both right and left hands separately since the interaction between device and hand can be highly asymmetric.[1] From this brief overview of the scope of SISO OTA testing it is easy to see that characterizing one multiband device requires thousands of measurements. Testing can take up to two weeks in an expensive anechoic facility.

"Currently, testing is also performed with both right and left hands separately since the interaction between device and hand can be highly asymmetric."

MIMO OTA Testing Challenges

With the exception of the phantom head and hand, the SISO measurements are independent of the external radio environment. The situation for MIMO OTA is quite different. MIMO is all about taking advantage of instantaneous spatial diversity in the radio channel, and thus the measured performance is tightly coupled to the radio propagation, noise, and interference conditions in which a device is tested. This extends to the closed loop behavior of the end-to-end system; that is, the real-time interaction between how the DUT measures the radio environment and the subsequent behavior of the base station scheduler, which can choose to reconfigure the downlink as frequently as 1000 times per second.

"When moving from SISO OTA to MIMO OTA, much of the discussion has been on methods for creating spatially diverse signals."

When moving from SISO OTA to MIMO OTA, much of the discussion has been on methods for creating spatially diverse signals. For MIMO OTA, the primary FoM is the downlink power required for a specific throughput. Other FoM continue to be studied such as the throughput achieved at a specific operating point.

As well as the dependence on the radio propagation conditions, a further challenge comes from the huge increase in the number of supported frequency bands, with hex band devices now being common. Each band imposes unique demands on the optimal receive and transmit antennas, resulting in the need for separate antennas for some bands. And although the focus here is cellular, there are further antenna demands for the support of Wi-Fi*, Bluetooth*, GPS, FM radio, DVB, and so on.

SISO OTA standardization had the advantage that every DUT on the planet was a potential measurement candidate. Even though MIMO devices and networks are now readily available, early MIMO OTA work was constrained by the lack of wide availability of a rich diversity of MIMO devices.

Everyone agrees that MIMO OTA testing needs to be no more complex, time consuming, or expensive than necessary, but even after five years of study, considerable work remains to agree on reference performance in specific conditions and the associated accuracy of the proposed methods of test.

Reference Antennas and Spatial Channel Models

CTIA and 3GPP have been collaborating on how to evaluate the most appropriate radio conditions in which to measure MIMO OTA performance as well as elaborating and evaluating the capabilities of the candidate methodologies to measure MIMO device performance.

In addition to the radio propagation conditions, the correlation between transmit antennas and receive antennas plays an important role in overall MIMO performance. Figure 1 shows the theoretical capacity of a 2x2 MIMO system, with α being the correlation between the transmitting antennas and β being the correlation between the receiving antennas.

"…the correlation between transmit antennas and receive antennas plays an important role in overall MIMO performance."

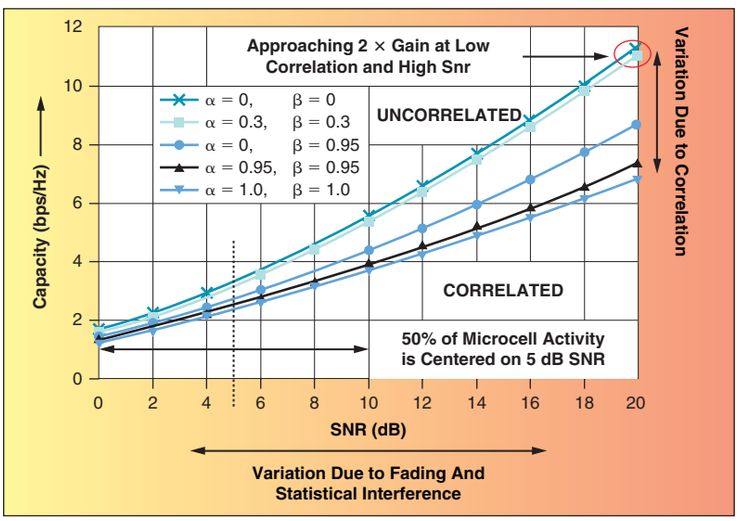


Figure 1: Shannon theoretical capacity for 2x2 MIMO system (Source: Agilent Technologies, 2010)

The overall system capacity is a function of the SNR and the correlation of the received signal. Figure 1 shows examples of different antenna correlations on the system capacity: the lower the correlation, the more the MIMO system will take advantage of the spatial diversity. (The additional correlation introduced by the radio propagation channel is not shown here.)

The spatial multiplexing gain potential provided by the decorrelation of the signal is explained in Figure 2. LTE's use of spatial multiplexing (a form of MIMO) tries to double the capacity by transmitting a parallel data stream per additional antenna. In the case where the propagation environment is uncorrelated, the receiver can separate and decode the two pre-coded data

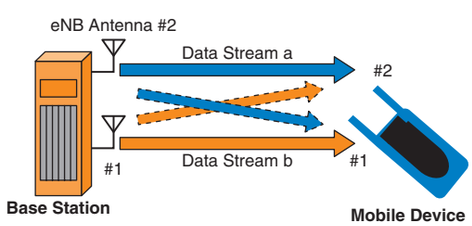


Figure 2: 2x2 MIMO spatial multiplexing (Source: Agilent Technologies, 2014)

streams successfully, therefore enabling twice as much throughput. When the channel is fully correlated, the receiver is unable to distinguish the two data streams and cannot take advantage of the multipath fading, which limits the capacity back to a SISO system.

To help with MIMO OTA test development, CTIA developed reference device antennas, which have been used both in simulation of expected performance in known channel conditions and actual measurements on real devices. These antennas were designed with low, medium, and high correlation in order to represent good, nominal, or bad performance respectively. Designs were created for low (700 MHz), medium (1800MHz), and high operating bands (2600 MHz). Reference antennas are shown in Figure 3:

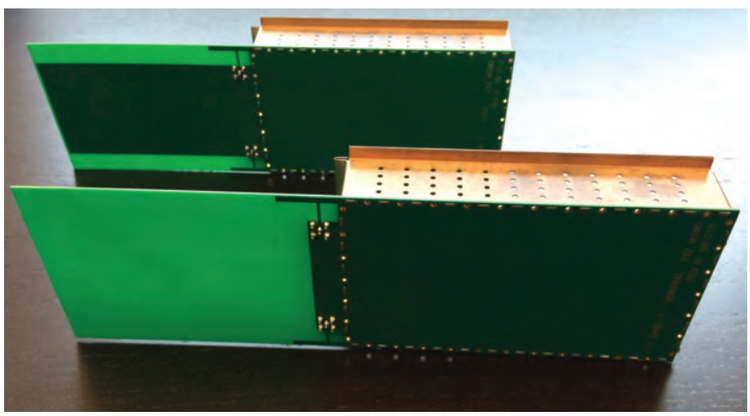


Figure 3: Band 7 (2655 MHz) good and nominal reference antennas. Note the large ground plane between the antenna elements for the good antenna.
(Source: Agilent Technologies, 2014)

The reference antennas connect to the temporary antenna connectors of the test device in order to provide the device with a known antenna pattern. To minimize the interaction between the reference antennas and the device, a screened box is used to isolate the device.

"The reference antennas provide the essential traceability required during the development of conformance test methods…"

The reference antennas provide the essential traceability required during the development of conformance test methods and future device minimum performance requirements.

Radio propagation channel models also had to be defined as part of the work to model the signal propagation conditions a device might see in a real network. Measurement campaigns carried out by COST resulted in the definition of two spatial channel models respectively called Spatial Channel Model Extended Urban Macro (SCME UMa) and SCME Urban Micro

(UMi). An implementation of SCME for MatLab* is publicly available. Although the SCME models are considered realistic, some of the simpler test methods are not able to fully reproduce the models. This complicates direct comparison of all the methods and will be discussed further in the following section.

CTIA also created a channel model validation procedure, which aims at verifying that the channel models presented to the device have the desired characteristics and properties between labs and methods. This is an essential step to minimize uncertainties in the results provided by different MIMO OTA test methods and labs.

"CTIA also created a channel model validation procedure, which aims at verifying that the channel models presented to the device have the desired characteristics and properties between labs and methods."

The combination of the reference antennas, the selected reference channel models, and other environmental considerations were all necessary for 3GPP and CTIA to make decisions for the selection of viable test methods.

Selected Test Methodologies

As part of the 3GPP MIMO OTA work item, eight different candidate tests methods were proposed in the Technical Report (TR) 37.977 for creating the necessary environment to test MIMO OTA performance. The test methods can be categorized into three groups:

- Anechoic chamber methods
- Reverberation chamber methods
- Multistage methods

The anechoic and reverberation methods take fundamentally different approaches towards the emulation of a spatially diverse radio channel. In the case of the anechoic chamber, multiple probes are used to launch signals at the DUT in order to create known angles of arrival, which map onto the required spatial channel model. This is a powerful approach, although in order to emulate arbitrarily complex channel models, large numbers of probes are required, which is costly and challenging to calibrate. For this reason only 2D channels have been studied so far. In the reverberation chamber methods, the spatial richness is provided by relying on the natural reflections within the chamber, which are further randomized by use of mode stirrers that oscillate to provide a rich spatial field. This process produces an intrinsically 3D field which, over long periods of time, approaches an isotropic (uniform) field. Isotropic fields are unable to be used for spatial multiplexing; however, the instantaneous spatial field is not isotropic, which means that the reverberation chamber can be used to measure spatial multiplexing gain through any uncorrelated antennas.

"The anechoic and reverberation methods take fundamentally different approaches towards the emulation of a spatially diverse radio channel."

Multistage methods disaggregate the measurement process into two or more simpler steps in order to emulate the different contributions of the overall MIMO system. Multistage methods use anechoic chambers and may also use a mixture of radiated and cable conducted measurements.

At the conclusion of the work item, four out of eight candidate methodologies were selected as having met defined criteria for standardization. These four methods are briefly described in the following subsections. For further details, refer to 3GPP TR 37.977.

Anechoic Ring of Probes Method

The ring of probes method is based on a symmetric ring of probe antennas equidistant around the DUT, which is placed at the center of the anechoic chamber. Each probe is fed by a channel emulator to generate the temporal characteristics of the desired channel model. The spatial components of the channel model are mapped onto the equally spaced probe antennas in such a way that an arbitrary number of clusters with associated angular spreads can be generated. This flexible approach enables any 2D spatial channel model to be generated without having to reposition (and recalibrate) the probe antennas. This method can be extended to 3D with several 2D rings for example. The number of antennas in the ring affects the accuracy with which the spatial dimension of the channel model can be implemented. A typical configuration is a 22.5 degree raster with vertical and horizontal polarization at each location giving a total of 32 probes, each independently driven by a channel emulator. It is also common practice to use horn antennas for the probes, due to their similar transmit characteristics across different bands. An example ring of probes setup is shown in Figure 4.

"A typical configuration is a 22.5 degree raster with vertical and horizontal polarization at each location giving a total of 32 probes,…"

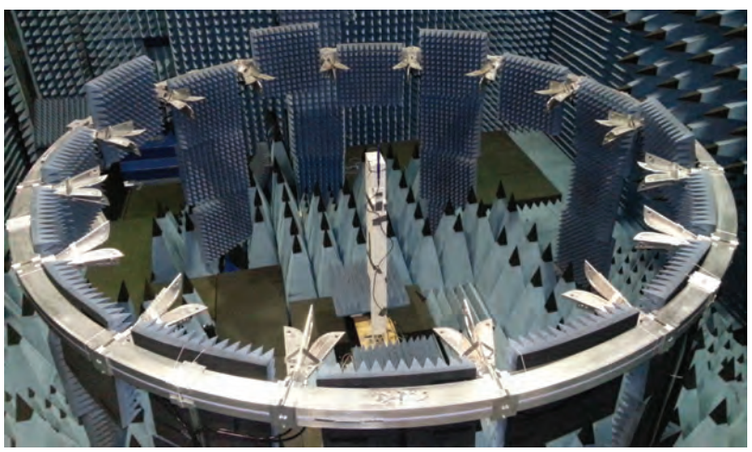


Figure 4: Anechoic ring of probe setup in Aalborg, Denmark (Source: Intel Corporation and Aalborg University[2], 2013)

Anechoic Two-Stage Method

The two-stage method takes an alternative approach to creating the necessary conditions to test MIMO performance.[3][4] The first stage of the method is illustrated in Figure 5. This stage involves the measurement of the antenna pattern of the DUT using an anechoic chamber of the size and type used for existing SISO tests. The measured pattern can be a simple 2D cut or a full

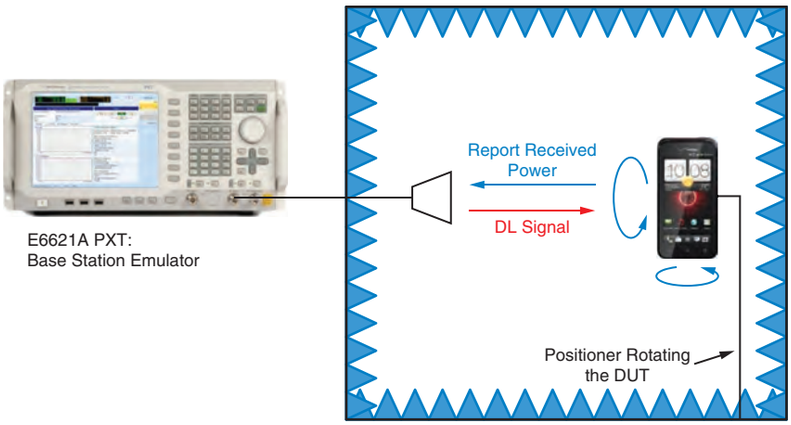


Figure 5: First stage of the anechoic two-stage method (Source: Agilent Technologies, December 2013)

3D measurement, depending on the type of evaluation being performed. The traditional way to measure antenna patterns uses RF-choked cables attached to the device. However, this is not considered acceptable for mobile device testing so a special antenna measurement test function has been developed to measure the antenna pattern "non-intrusively" using the DUT's own receiver with the help of special calibration routines. This special test function is required for enabling the two-stage method, which reports the received power per antenna and relative phase between antennas for a given received signal. From these measurements and further calibration procedures it is possible to reconstruct the DUT antenna patterns and phase responses between the antennas.

The second stage takes the measured antenna pattern from the first stage and convolves it with the desired channel model using a channel emulator. The output of the channel emulator then represents the faded downlink signal modified by the spatial properties of the DUT's antenna. This is the same signal that the DUT would have received had the antennas been placed in the desired radio field. Throughput measurements are then made using the two outputs of the channel emulator.

In essence, the two-stage method replaces the complexity of creating an arbitrary radio field in an anechoic chamber with the challenge of measuring the DUT antenna pattern.

"…the two-stage method replaces the complexity of creating an arbitrary radio field in an anechoic chamber with the challenge of measuring the DUT antenna pattern."

If the full 3D antenna pattern is measured in the first stage, the two-stage method can be used to emulate any 2D cut or any arbitrary 3D channel propagation condition. The rotation of the DUT relative to the channel model is accomplished by switching to the representative antenna pattern within the channel emulator.

There are two different methods of connecting the channel emulator output to the DUT. The simplest method is direct cable connection, known as the conducted two-stage method. This is convenient since throughput can be

measured using a simple screened enclosure. However, this approach does not capture the effect of radiated self-interference and so is limited to development purposes. For DUT conformance testing, an alternative radiated connection is established known as the radiated two-stage method. This approach does include the effect of DUT radiated self-interference. To establish the radiated connection, an anechoic chamber is required. Two orthogonal probes in the chamber are used to create the connection to the DUT.

A further calibration step is then required to isolate the transmissions to the DUT from each probe antenna. This is done by measuring the transmission loss and phase of each probe antenna to each DUT receiver at one position of the DUT antenna. From these measurements an inverse transmission matrix is computed, which, when applied to the transmit signals removes the effects of the anechoic chamber and, allows each probe to transmit over the air to one DUT receiver with less than 20 dB crosstalk. Furthermore, this calibration procedure, which is based on nulling techniques, effectively removes any sensitivity to the absolute accuracy of the pattern measured by the DUT in the first stage; only its shape is of consequence since gain errors will be calibrated out later.

"From these measurements an inverse transmission matrix is computed, which, allows each probe to transmit over the air to one DUT receiver with less than 20 dB crosstalk."

An illustration of the radiated second stage is given in Figure 6.

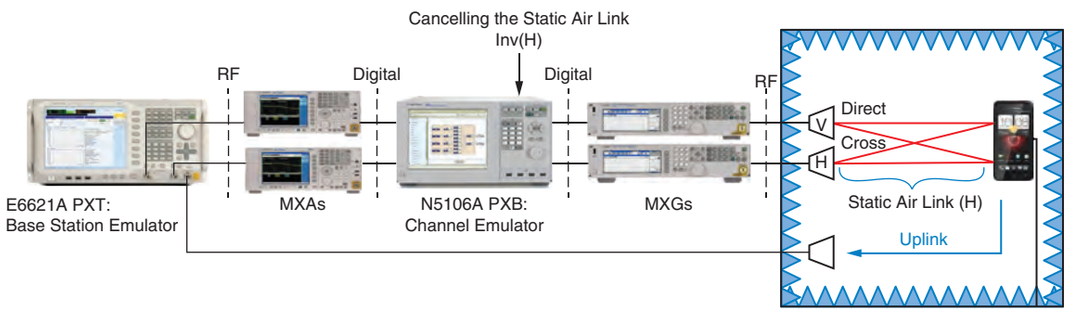


Figure 6: Second stage of the radiated two-stage method (Source: Agilent Technologies, December 2013)

Reverberation Chamber Method

The first of the reverberation-based methods uses the intrinsic reflective properties of the mode-stirred reverberation chamber to transform the downlink test signal into a rich 3D multipath signal. An illustration is shown in Figure 7.

"The spatial characteristics of the signal are random and over time can be shown to be isotropic, but when observed over the time period of a demodulated data symbol, they are known to be highly directional."

The spatial characteristics of the signal are random and over time can be shown to be isotropic, but when observed over the time period of a demodulated data symbol, they are known to be highly directional. This non-uniformity provides the DUT with diverse signals on each antenna thus enabling spatial multiplexing gain. The natural time domain response of the chamber can be modified with the use of small amounts of RF absorptive

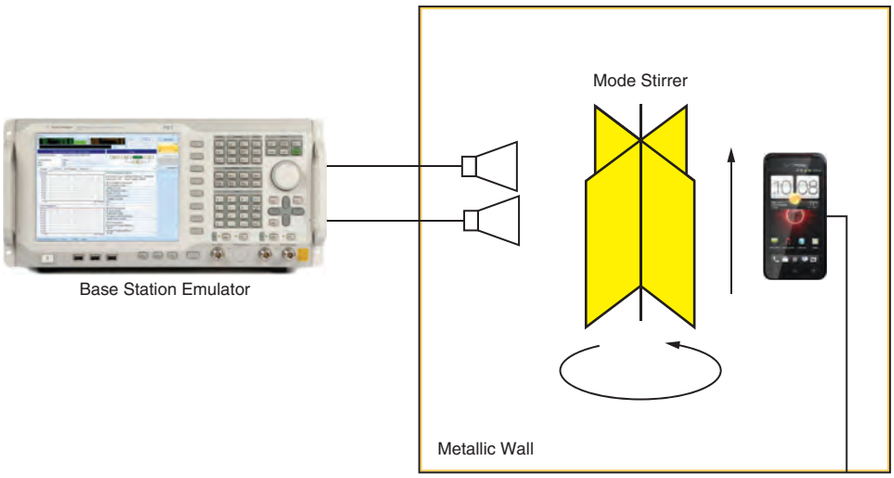


Figure 7: Reverberation chamber
(Source: Agilent Technologies, December 2013)

material. It is however not possible to emulate the SCME channel models with a reverberation chamber due to a lack of control of the instantaneous spatial characteristics. The basic reverberation chamber is limited to a single power delay profile and a relatively slow Doppler spectrum determined by the speed of the mode stirrer. Further control of the power delay profile and spatial aspects can be obtained by cascading two or more reverberation chambers, and there has also been research by the National Institute of Standards and Technology (NIST) using nested chambers and coupled chambers. In addition to the conventional Rayleigh 3D isotropic fading scenario emulated by single-cavity reverberation chambers, multicavity multisource mode-stirred reverberation chambers employ de-embedding algorithms for enhanced repeatability. They have also added capabilities to emulate different K-factors for Rician fading, different non-isotropic scenarios including single and multiple cluster with partial door opening, and standardized or arbitrary amplitude power delay profiles (for example, 802.11n, Nakagami-m, on-body, and user-defined) using sample selection techniques.

"Further control of the power delay profile and spatial aspects can be obtained by cascading two or more reverberation chambers,…"

Reverberation Chamber and Channel Emulator Method

This last method addresses the limitation of the basic or cascaded reverberation chamber by adding a channel emulator to the downlink prior to launching the signals into the chamber. Figure 8 shows the new setup with a channel emulator. This allows the temporal aspects of the desired channel model to be fully controlled. With the use of a channel emulator capable of negative time delay (inverse injection), multiple cavity mode-stirred reverberation chambers can accurately emulate the power delay profiles of 3GPP SCME channel models. The SCME spatial characteristics however remain uncontrolled.

"With the use of a channel emulator capable of negative time delay (inverse injection), multiple cavity mode-stirred reverberation chambers can accurately emulate the power delay profiles of 3GPP SCME channel models."

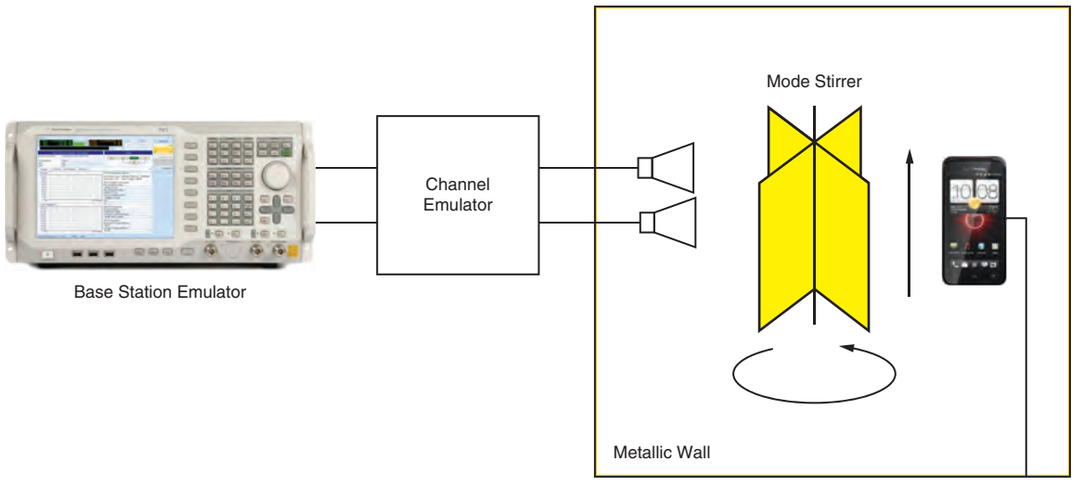


Figure 8: Reverberation chamber with channel emulator
(Source: Agilent Technologies, December 2013)

Methods Benchmarking

"…there is still the need to precisely assess the limitations and uncertainty of each method."

Even though four methods have been selected by 3GPP for characterizing MIMO performance, there is still the need to precisely assess the limitations and uncertainty of each method. The methods were selected on the basis that they could clearly discriminate DUT receiver performance when using the bad, nominal, and good MIMO reference antennas and met the uncertainty criteria of the legacy SISO OTA (TRS) of ±2.3 dB. However, this is only a provisional estimate of the actual system uncertainty and much more work needs to be done.

The DUT performance requirements are also to be defined. Work in on device characterization and measurement uncertainty is continuing in CTIA.

Each method has certain capabilities and limitations. To help understanding, a benchmarking table of the 3GPP selected methods was agreed and captured in 3GPP TR 37.977[5] Table 12.4-1.

"The ring of probes and radiated two-stage methods can generate the same channel models and provide nominally aligned results."

The ring of probes and radiated two-stage methods can generate the same channel models and provide nominally aligned results. The only exception is that the two-stage method is not currently able to test DUTs with active antennas that can change their pattern dynamically in response to the channel conditions.

The results from the reverberation chamber methods cannot be directly compared to the ring of probes and two-stage methods because the channel models that can be created are different. The results for nonpolarized antennas averaged in many orientations are similar between anechoic and reverberation methods; however this comparison does not hold well for devices designed with specific antenna polarization and DUT orientation such as laptops.

Conclusion

MIMO OTA testing is a revolutionary challenge for LTE and considerable work remains in order to provide accurate and efficient MIMO LTE testing. The currently selected 3GPP methods are not fully harmonized and discrepancies can be observed between them. Only once this is resolved will it be possible to move to the next stage and define performance requirements for devices. In the future, further complexities such as active antennas and carrier aggregation will need to be considered.

Acknowledgment

We would like to acknowledge the support from the Danish National Advanced Technology Foundation for the 4th Generation Mobile Communication and Test Platform (4GMCT) project which contributing to this article.

References

[1] CTIA, Test Plan for Wireless Device Over the-Air Performance, "Method of Measurement for Radiated RF Power and Receiver Performance" October 2013.

[2] Fan, Wei, Xavier Carreño, Jagjit S. Ashta, Jesper Ø. Nielsen, Gert F. Pedersen, and Mikael B. Knudsen, "Test Setup for Anechoic Room based MIMO OTA Testing of LTE Terminals," 7th European Conference On Antennas and Propagation (EuCAP), 2013.

[3] Rumney, Moray, Steve Duffy, Ya Jing, Zhu Wen, and Hongwei Kong, "Two-stage MIMO OTA method," COST 2100 TD(09)924, Vienna, Austria, September 2009.

[4] 3GPP R4-091361, Agilent Technologies, "MIMO OTA test methodology proposal" March 2009.

[5] 3GPP TR 37.977, "Verification of radiated multi-antenna reception performance of User Equipment (UE)", January 2014.

Author Biographies

Nicolas Obriot (niob@telenor.dk) joined Agilent in 2013 as a research engineer after receiving an MSc degree in Telecommunications from Aalborg University. He has been working on the validation and acceptance of the two-stage method.

Moray Rumney (moray_rumney@agilent.com) joined Hewlett-Packard/ Agilent Technologies in 1984 after receiving a BSc degree in Electronics from Heriot-Watt University in Edinburgh. His career has spanned manufacturing engineering, product development, applications engineering, and most recently

technical marketing. His main focus has been the development and system design of base station emulators used in the development and testing of cellular phones. Rumney joined ETSI in 1991 and 3GPP in 1999 where he was a significant contributor to the development of type approval tests for GSM and UMTS. He currently represents Agilent at 3GPP RAN WG4, developing the air interface for HSPA+ and LTE. His current focus is on radiated testing of MIMO devices. Rumney has published many technical articles and is a regular speaker and chairman at industry conferences. He was editor of Agilent's book *LTE and the Evolution to 4G Wireless*, whose second edition covering LTE-Advanced was published in February 2013.

Janus Faaborg (janus_faaborg@agilent.com) joined Agilent Technologies Denmark in 2009 as leader of a research program within 4G-based mobile devices and test infrastructure (4GMCT). He holds an MSc within Electrical Engineering from Aalborg University 2002 and gained broad experience from the wireless chipset industry—especially in the field of RF tests and measurements. He has been working for several years as R&D manager of hardware and software teams.

Michael Dieudonne (michael_dieudonne@agilent.com) joined Agilent in 2001 through the acquisition of Sirius Communications. He holds an MSc in Electronic Engineering (1999) and an MBA (2004). Over the last years, he has been managing different technology research groups in the corporate research lab in diverse geographies and fields such as telecommunications, nonlinear component characterization, and nanotechnologies. Beside the research management, he has been active in business development and European funded research programs from partner to project leader role.

DEVELOPMENT OF ADVANCED PHYSICAL LAYER SOLUTIONS USING A WIRELESS MIMO TESTBED

Contributors

Pavel Loskot
College of Engineering,
Swansea University, United Kingdom

Biljana Badic
Platform Engineering Group,
Intel Corporation

Timothy O'Farrell
Electronic and Electrical Engineering,
Sheffield University, United Kingdom

A wireless multiple-input multiple-output (MIMO) testbed with up to 6 GHz radio-frequency (RF) carrier and up to 40 MHz modulated signal bandwidth provides a unique opportunity for real-time realistic experimentation, demonstration, testing, validation, and general R&D, as well as education. Wireless device testing and development and verification of the advanced physical layer (PHY) solutions are of particular interest to cellular network operators and vendors of 4G equipment. This article portrays needs for hardware emulation of wireless systems and their deployment challenges. Furthermore, the article describes an affordable 2x2 MIMO testbed, its hardware and software components, and their connectivity. The testbed usage scenarios and experiments are illustrated through examples that are of great interest to the research community. Finally, likely future developments in the emulation of wireless communication systems are outlined.

The Need for Hardware Emulation of Wireless Systems

"The recent revolution in wireless communications has been fueled by the progress in emulation and simulation techniques…"

The recent revolution in wireless communications has been fueled by the progress in emulation and simulation techniques that are used in the R&D of these systems. One of the main aims of simulation and emulation is to evaluate the system's performance and verify its conformation to design specifications. Nevertheless, the ultimate motivation is to reduce the R&D costs as well as shorten development times despite the growing complexity of systems. In general, the system's performance prediction should be:

- *accurate* (unbiased)
- *efficient* (in terms of time and effort required)
- *verifiable* (reproducible)
- *trustable* (inspire some level of confidence)

"Whereas emulation mimics the system's behavior, simulations rely on mathematical modeling…"

Whereas emulation mimics the system's behavior by duplicating the system's inner workings, simulations rely on mathematical modeling of the system. The latter have a distinctive feature of readily supporting the simulation code debugging; however, the former is more likely to reproduce a broader range of the system's behaviors and do so more faithfully and in real time. Furthermore, simulations are often limited by the tractability of mathematical models, trading off the model complexity with the model accuracy, while the modeling is strongly dependent on assumptions adopted about the system. Consequently, real-time system emulations are gaining importance not only in the industrial environment, but also in academia.

The current transition from the 3G (Third Generation) to the 4G (Fourth Generation) mobile systems requires over-the-air (OTA) testing of wireless equipment with multiple antennas. In addition to power-delay profile (PDP), Doppler spread, and path-loss attenuation (PLA), the multiple-input multiple-output (MIMO) radio-channel modeling and emulation have to account for the antenna radiation patterns and polarization, spatial selectivity including angular spread, and multipath correlations.[1] More realistic but also more challenging radio-channel emulation assumes these channel characteristics under various types of mobility, often referred to as *virtual drive testing* (VDT). Moreover, testing of wireless equipment imposes at least three conditions on radio-channel emulation: well-defined, real-time, and reproducible testing conditions. For instance, the real-time channel emulation in hardware can be combined with testing inside an anechoic chamber such that the real antennas provide OTA connection between the channel emulator and the wireless equipment under test.

It should be also noted that channel models are constantly evolving. For instance, the 3GPP (Third-Generation Partnership Project)[2] has recently introduced more realistic nonstationary channel models including a moving propagation channel model that allows time-varying multipath delays, and a birth-death propagation channel model in order to capture erratic multipath delay changes due to mobility. Emulation as well as simulation of radio channels with changing multipath delays is more challenging; however, many hardware channel emulators available in the market support this option, including replaying the previously recorded propagation conditions such as Propsim*[3] and Spirent*[4] channel emulators. The concepts of flexible testing of wireless equipment and systems in all phases of the development lifecycle has been explained in a recent article.[5]

Perspectives of Hardware Emulation of Wireless Systems

From the cellular network operator point of view, offering the Quality-of-Service (QoS) guarantee is of paramount importance, since it is directly linked to the operational costs (OPEX) and potential revenues. In particular, the network operators are interested in defining the worst-case scenarios and radio propagation conditions in order to provide the desired levels of QoS and to mitigate the risks of poor performance for the wireless equipment considered. On the other hand, vendors of wireless equipment and network infrastructure providers are interested in defining unified tests in order to provide the performance guarantees and verify the standards compliance for their wireless equipment.

"From the cellular network operator point of view, offering the Quality-of-Service (QoS) guarantee is of paramount importance,…"

"…defining unified tests in order to provide the performance guarantees and verify the standards compliance…"

Another perspective on wireless system emulation is represented by the developers of signal processing algorithms working in the R&D labs in industry as well as in academia. Unlike general software development, signal processing algorithms employed at the transmitter and at the receiver in a wireless communication system have to be tested and validated for different random realizations of the radio channel. Such testing often focuses on studying rare events that may lead to algorithm failures, such as, for example, failures caused by an accumulated numerical instability or a failed convergence

in iterative processing. The random radio-channel realizations corresponding to these rare events can be identified, stored, and then replayed in future tests of modified or improved algorithms. The repeated tests require reproducible radio-channel realizations with the minimum error vector magnitude (EVM). More importantly, the entire design of all the transmitter and receiver algorithms has to be extensively evaluated prior to the field tests. However, this is a computationally very demanding task that can usually be only accomplished by emulation and not by simulations.

"…the demand for shorter and cheaper product development cycles are increasing, many universities and industrial labs operating the hardware wireless testbeds."

In summary, as the complexity of wireless systems and the demand for shorter and cheaper product development cycles are increasing, emulation of wireless systems plays an increasingly important role in the R&D process. It is also more affordable, with many universities and industrial labs operating the hardware wireless testbeds.

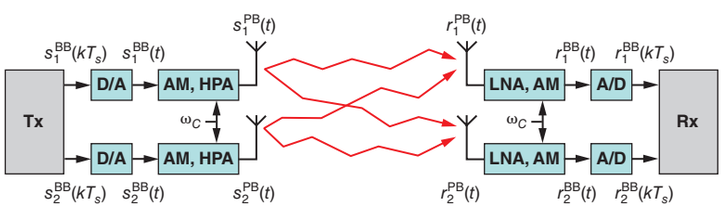
In this article, we first describe the MIMO testbed using mathematical models. This is followed by more detailed discussion of workings of the MIMO testbed, including the description of the hardware and software components and their connectivity. A large section is devoted to description of the MIMO testbed experiments in various scenarios for device testing and algorithm development. The last section concludes the article and indicates future developments in emulation of wireless communication systems.

A Mathematical Model of a Wireless MIMO Testbed

"Mathematical models are necessary for theoretical design and computer simulations, as well as for evaluations and predictions of the performance…"

Before explaining the workings of the MIMO testbed, we first describe a mathematical model representing processing and flows of the modulated signal in the testbed. Mathematical models are necessary for theoretical design and computer simulations, as well as for evaluations and predictions of the performance of the proposed schemes, provided that the model is analytically tractable.

The mathematical model of the wireless testbed considered is shown in Figure 1 assuming $N_T = 2$ transmitting and $N_R = 2$ receiving antennas.



$$r_1^{PB}(t) = \sum_{i=1}^{L} \left\{ g_{11}(i;t)\, s_1^{PB}\left(t - \tau_{11}(i;t)\right) + g_{21}(i;t)\, s_2^{PB}\left(t - \tau_{21}(i;t)\right) \right\} + w_1(t)$$

$$r_2^{PB}(t) = \sum_{i=1}^{L} \left\{ g_{12}(i;t)\, s_1^{PB}\left(t - \tau_{12}(i;t)\right) + g_{22}(i;t)\, s_2^{PB}\left(t - \tau_{22}(i;t)\right) \right\} + w_2(t)$$

Figure 1: Mathematical model of the 2x2 MIMO testbed (Source: Swansea University, 2014)

The signal flow from the transmitter to the receiver in the testbed undergoes conversions between the digital and analog signals, and between the baseband signals and the passband signals modulated at the carrier frequency. The fully digital transmitter (Tx) creates equally spaced samples of the digitally modulated signals $s_1^{BB}(kT_S)$ and $s_2^{BB}(kT_S)$ in the baseband (BB) where k is the index, and T_S is the sampling period. The samples of modulated signals are converted to the continuous-time baseband waveforms $s_1^{BB}(t)$ and $s_2^{BB}(t)$ using a digital-to-analog (D/A) converter. The baseband signals are up-converted to the (angular) carrier frequency ω_C before high-power amplification (HPA), and transmission of the passband (PB) signals $s_1^{PB}(t)$ and $s_2^{PB}(t)$ from the transmitter antennas. The transmitted signals are distorted by the wireless MIMO channel before they are received as signals $r_1^{PB}(t)$ and $r_2^{PB}(t)$ at the two receiving antennas, respectively.

"The signal flow from the transmitter to the receiver in the testbed undergoes conversions between the digital and analog signals, and between the baseband signals and the passband signals…"

The transformation of signals $s_1^{PB}(t)$ and $s_2^{PB}(t)$ into signals $r_1^{PB}(t)$ and $r_2^{PB}(t)$ is mathematically expressed as shown in Figure 1. In this transformation, $g_{uv}(i;t)$ and $\tau_{uv}(i;t)$ denote the channel attenuation (combining small-scale fading with large-scale shadowing and free-space path loss) and the time-varying delay of the i thpropagation multipath between the transmitter antenna $u \in \{1,2\}$ and the receiver antenna $v \in \{1,2\}$, respectively, and $w_1(t)$ and $w_2(t)$ are zero-mean independent white Gaussian processes representing additive noise. Hence, the received signals are a noisy superposition of the transmitted signals, which are distorted in time as well as in frequency by possibly nonstationary propagation channels, each channel consisting of up to L multipath components. For the purpose of emulation and simulation of the radio-propagation channel, this signal distortion can be equivalently implemented as filtering of the transmitted signal by a linear, possibly non-stationary filter having a (time-domain) finite impulse response

$$h_{uv}(t,\tau) = \sum_{i=1}^{L} g_{uv}(i;t)\ \delta\left(\tau - \tau_{uv}(i;t)\right)$$

In the frequency domain, the channel transfer function is calculated using the Discrete-time Fourier transform (DtFT), that is:

$$H_{uv}(t,f) = \sum_{i=1}^{L} g_{uv}(i;t)\ \mathrm{Exp}\left[2\pi f\ \tau_{uv}(t)\right]$$

The details of how to extend the channel impulse response to spatial domain can be found, for example, in Kyösti et al.[6]

The statistics of the random processes $g_{uv}(i;t)$ and $\tau_{uv}(i;t)$ have a dominating effect on the performance of the MIMO link[7], and they are the basis of the MIMO channel models.[8] The joint statistics are often partly described assuming the correlation or covariance matrices. The MIMO channel models reflect the spatiotemporal characteristics of the propagation environment, so they are often referred to as spatial channel models (SCMs). The geometry-based stochastic channel models consider the clusters of scatterers. These models have been studied in the European

"The MIMO channel models reflect the spatiotemporal characteristics of the propagation environment, spatial channel models (SCMs)."

framework projects such as COST#259, COST#273, and WINNER 1 and 2.[6] The recorded channel measurements obtained by channel sounding techniques or ray-tracing methods are often used to estimate the statistics of the channel impulse response.[9] A more deterministic approach to SCM assumes replaying the recorded channel impulse responses in the course of channel emulation in order to accurately reproduce the radio-propagation environment.[3][4] The SCMs have been standardized by the 3GPP[2] and the IEEE.[10]

"Linear multidimensional modulations such as MIMO-OFDM are used in many contemporary communication systems…"

The transmitter signaling $s_1^{PB}(t)$ and $s_2^{PB}(t)$ depends on the scenario considered. The determining factor is whether the transmitted signals are designed jointly as in the case of beamforming and space-time coding[11], or they can represent two independent data streams. Linear multidimensional modulations such as MIMO-OFDM (orthogonal frequency division multiplexing) are particularly attractive and are used in many contemporary communication systems due to their excellent spectral efficiency.[11] The receiver design to process signals $r_1^{PB}(t)$ and $r_2^{PB}(t)$ depends on the transmitter signaling, the channel model adopted, the allowable complexity of algorithms including any real-time processing requirements, and the amount of *a priori* knowledge. To explain the latter, the receiver usually requires full knowledge of all channel processes $g_{uv}(i;t)$ and $\tau_{uv}(i;t)$ in order to successfully recover the transmitted signals $s_1^{PB}(t)$ and $s_2^{PB}(t)$. Such knowledge can be obtained, for example, in offline receiver processing when realizations of the channel processes are recorded in the channel emulator during the transmission. However, in real-time (online) processing, for example, during the wireless devices tests, the receiver has to estimate the processes $g_{uv}(i;t)$ and $\tau_{uv}(i;t)$ prior to detecting the signals $s_1^{PB}(t)$ and $s_2^{PB}(t)$, which significantly increases the implementation complexity of the receiver. More detailed discussion about the setup of various tests of wireless devices and development and evaluations of the signal processing algorithms at the receiver and/or transmitter will be provided in the section "Scenarios and Case Studies."

Operation of the Wireless MIMO Testbed

"The testbed components have been chosen to allow real-time transmissions at the carrier frequencies up to 6 GHz,…"

We describe the components comprising the wireless MIMO testbed, their connectivity and the overall testbed operation. The specific hardware and software components comprising the wireless MIMO testbed are listed in Table 1 and Table 2, respectively, assuming the MIMO testbed with two transmitting and two receiving antennas (2x2 MIMO). The testbed components have been chosen to allow real-time transmissions at the carrier frequencies up to 6 GHz, which covers most of the wireless cellular and local area network systems including the attractive worldwide license-free bands at 2.4 GHz and 5.7 GHz. The basic concept of a wireless MIMO testbed is depicted in Figure 2, assuming a testbed set up for the algorithm development; other configurations of the testbed including real-time wireless device testing will be discussed in the next section.

Equipment	Pcs.	Description
Agilent E4438C	1	Vector Signal Generator
Agilent E8267C	1	Vector Signal Generator
Agilent 54855A	1	Infiniium Digital Oscilloscope
Agilent 89600S	2	Vector Signal Analyzer
Agilent E4418B	1	Power Meter
Agilent E8491B	1	IEEE 1394 Interconnect
Agilent 89605B	2	RF Input
Agilent E2731B	1	6-GHz RF Tuner
Agilent E1439C	2	Analog-to-Digital Converter
Agilent E82357B	3	GPIB-USB converter
Spirent SR5500	1	Channel Emulator
Spirent 6GHz	1	6 GHz converter for SR5500
Desktop PC	3	Personal desktop computers
RF cables	8	Cables and connectors
Agilent E82357B	3	GPIB-to-USB converter

Table 1: Hardware components of the MIMO testbed
(Source: Swansea University, 2014)

Software	Description
Matlab* ICToolbox	Instrument Control Toolbox
Agilent 89601B	Software for 89600 VSA
Spirent Testkit	SR5500 control and GUI
Device drivers: VISA-VXI, USB and GPIB, and C compiler	

Table 2: Software components of the MIMO testbed
(Source: Swansea University, 2014)

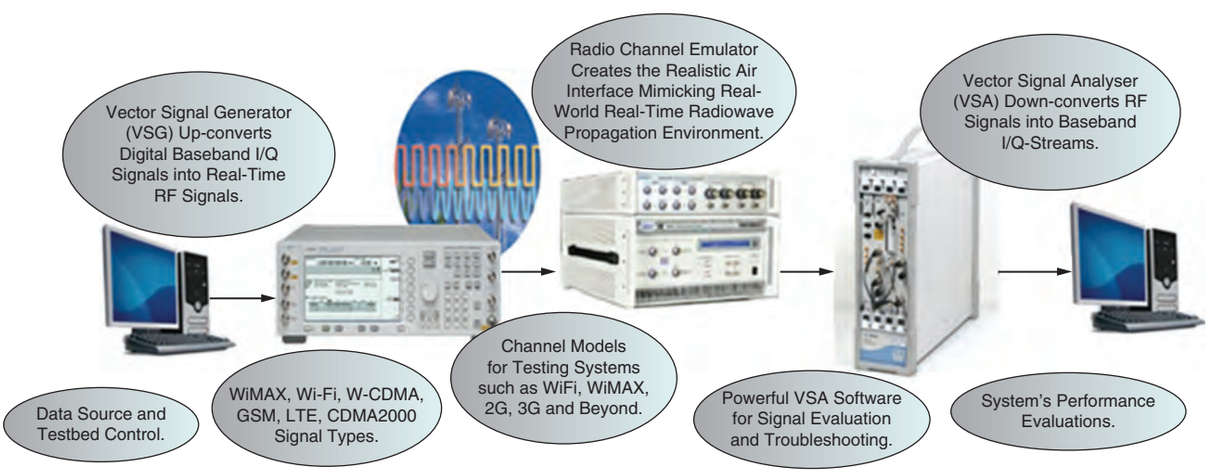


Figure 2: Conceptual structure of the MIMO testbed
(Source: Swansea University, 2009)

Hardware Components of the MIMO testbed

In Figure 2, samples of the modulated signals in the baseband $s_1^{BB}(kT_S)$ and $s_2^{BB}(kT_S)$ are generated on a personal computer (PC) using some suitable software such as MATLAB. The samples are stored in a file using a specific file format that is uploaded to the vector signal generator (VSG). While reading the file, the VSG interpolates the baseband samples using the DAC, and the baseband signal is possibly up-converted to the carrier frequency and sent to the VSG radio-frequency (RF) output port. We have chosen the E4438C and E8267C VSGs from Agilent Technologies[12] to generate two independent modulated signals. We also employ Agilent Technologies Digital Signal Oscilloscope (DSO) 54855A[12] to monitor the signals in both time and frequency domains with a range of data preprocessing options available including averaging, differentiation, addition, inversion, and lowpass and highpass filtering. The Infiniium Programmer's Reference describes the syntax, data types, status reports, and programming interface for communications including data transfers between the DSO and the connected PC.

"The core component of the wireless MIMO testbed is the hardware channel emulator."

The core component of the wireless MIMO testbed is the hardware channel emulator. It emulates radio-wave propagation conditions in real time. The input-output (I/O) signals of the channel emulator are either analog or digital signals in the baseband and, optionally, the analog RF signals at the carrier frequency. Internally, the channel filtering of modulated signals in the emulator is performed digitally in the baseband. Thus, the emulator input signals are first down-converted and sampled, whereas the output samples in the baseband are interpolated and possibly up-converted back to the carrier frequency. The digital channel filtering offers a full control over the virtual radio-propagation environment while enabling a number of unique features such as accurately setting the signal-to-noise ratio (SNR) or the carrier-to-noise ratio (CNR), compensating for the cable losses, calibrating and correcting phase offsets, repeating the same channel realizations, and replaying the measured or calculated (for example, using a ray-tracing) channel impulse responses. These features allow the identification of the performance issues of the equipment or algorithms early in the design cycle. In addition, the channel emulator offers many preset standardized channel models devised in the industry for testing and evaluation purposes. Notably, the use of a dedicated hardware channel emulator is necessitated by the significant computational resources required for the digital multipath channel filtering in MIMO systems, which may also include real-time generation of a relatively large number of pseudorandom channel processes; such computational resources are far beyond those available on a single PC. We have used the Spirent SR5500 channel emulator[4] with the 6 GHz RF option. SR5500 is a fully programmable 2x2 MIMO channel emulator with up to 24 independent multipaths for each pair of the transmitter-receiver antennas.

"…the channel emulator offers many preset standardized channel models devised in the industry for testing and evaluation purposes a dedicated hardware channel emulator is necessitated by the significant computational resources required…"

The channel emulator outputs are captured using a vector signal analyzer (VSA). The VSA performs down-conversion (usually in several stages using intermediate carrier frequencies), and sampling of the baseband signals. The samples are stored in the memory before being sent from the VSA to

a PC. This PC carries out processing of the baseband samples, and thus it represents the digital receiver, whereas the VSA represents the receiver frontend analog processing. Provided that a LAN is used to interconnect the testbed components (see next subsection), the PC for generating modulated signal samples and the PC for processing the received signal samples can be identical.

We have configured the modular VSA 89600S from Agilent Technologies[12] to obtain the samples of received signals in the baseband. In particular, the following modules are used: a resource and system controller and IEEE-1394 (FireWire*) interface module E8491B, two RF input and calibration modules E89605B, two RF tuner/down-converter modules E2731, and two ADC modules E1439.[12] The modules are interconnected via the VXI-bus (VME eXtensions for Instrumentation), which is an open standard industry platform recommended for high performance data acquisition applications. The E8491B module also serves as the central timing module, but it does not perform any signal processing or data acquisition. The 89605B module also generates the reference (RF synchronization) signals for the subsequent modules E2731 and E1439 modules. Provided that the received (input) signal is already in the baseband being fed from the corresponding output of the MIMO channel emulator, the E89605B module diverts the signal to the E1439 module bypassing E2731; otherwise, the E89605B output signal is routed through the E2731 module. The E2731 RF tuner down-converts the input signal in multiple stages using a set of local oscillators, amplifiers, and filters. The E1439 module samples the input signal at the rate of 95 MSa/s (Mega-samples per second). The input signal of E1439 is either already in the baseband with up to 36 MHz of bandwidth, or it can be a modulated signal at the intermediate frequency (IF) with up to 70 MHz of bandwidth.

Hence and importantly, the testbed configuration described above fully supports real-time transmissions of modulated signals with up to 6 GHz carrier frequency. However, the modulated signal bandwidth is limited by the E1439 module to 36 MHz. Such bandwidth is sufficient for testing most of the cellular systems (for example, typical bandwidth allocated to the 4G network operators is 10–15 MHz per transmission link), and some wireless LAN configurations (for example, IEEE 802.11a/b/g links require up to 20 MHz bandwidth).

Connectivity of the Testbed Components

The 2x2 MIMO testbed defines two modulated signal paths and also requires the control signals for its operation as shown in Figure 3.[13]

"The 2x2 MIMO testbed defines two modulated signal paths and also requires the control signals for its operation…"

This makes the testbed very rich in the number and types of signal interfaces. In particular, the analog modulated signals are carried over the RF coaxial cables with optional built-in attenuators. The RF terminals at the equipment are mostly Type-N connectors with the internal 50 Ohms impedances. The digital oscilloscope 54855A has four BNC-type input terminals to its four signal channels. The RF output of the signal generator E8267C is the SMA-type connector whereas the low-frequency outputs from E8267C and E4438C are the BNC-type connectors.

"…the testbed very rich in the number and types of signal interfaces."

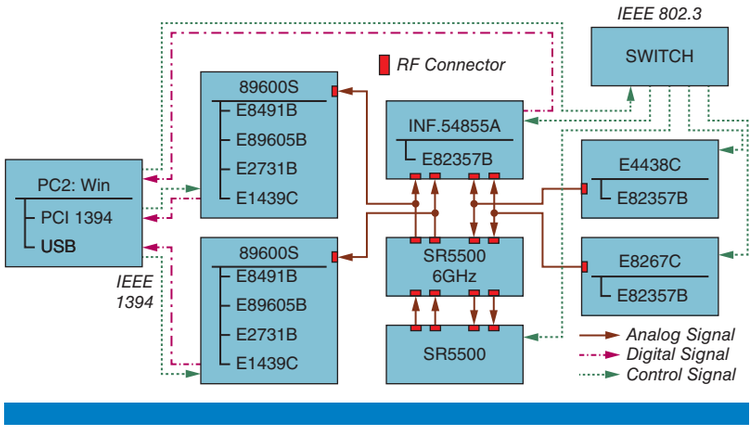


Figure 3: Block diagram of signal flows in the MIMO testbed (Source: Swansea University, 2012)

The control signals in the testbed can be carried over one of the following interfaces: the IEEE 488, better known as the GPIB (General Purpose Interface Bus), the IEEE 1394 (FireWire), and the IEEE 802.3 (Ethernet). The communications via FireWire interface require the installation of Agilent I/O (Input/Output) Libraries on the connected PC. The Agilent I/O Libraries provide a set of relevant drivers including Agilent VISA (Virtual Instrument Software Architecture) and Agilent SICL (Standard Instrument Control Library) drivers. The IEEE 488 popularity is decreasing especially for small-scale experiments with a few interconnected devices due the overall high cost of the GPIB cards and cables, and also for experiments requiring high command or data throughput. The GPIB-to-USB converters E82357B can be used to control the VSGs E4438C and E8267C and the DSO 54855A including the uploading and downloading of data files to and from the equipment, respectively. In the absence of the E82357B adaptor, the PC requires a rather expensive GPIB expansion card (such as the Agilent Technologies 82351A PCI card) and a GPIB cable. We note that some of the Agilent equipment also has the RS-232 interface; however, we did not find any benefits to consider this interface to control the testbed.

"The IEEE 488 popularity is decreasing especially for small-scale experiments with a few interconnected devices due the overall high cost…"

"The hardware interfaces are controlled by the appropriate software drivers that are supported within the particular operating system…"

The hardware interfaces are controlled by the appropriate software drivers that are supported within the particular operating system (OS). The IEEE 488 interface offers a very rich library of programming and control commands known as the Standard Commands for Programmable Instruments (SCPI). However, unlike the USB (Universal Serial Bus) and the FireWire interfaces that are Plug and Play (that is, automatically recognized by most operating systems), the instrument control interface IEEE 488 requires special drivers to be installed. These drivers and the corresponding communication protocols are defined as part of VISA. VISA is now considered to be the industry standard Application Programming Interface (API) for communications between the test and measurement (T&M) instruments and personal computers over, for example, the GPIB and VXI bus. The VISA standard ensures compatibility of applications due to unification of presentation and operation capabilities. VISA-based APIs have been created for many common and less common programming languages.[14]

Agilent Technologies generally provides very limited support for instrument communications under Linux* OS and instead develops proprietary applications for Windows* OS. Several open-source VISA drivers are available for Linux even though their development seems to be less active in recent years. Hence, a release of the new Linux kernel 3.x has created a number of incompatibility issues with the existing Linux VISA drivers. This is particularly true for the USBTMC driver (the VISA driver for T&M equipment over the USB). Nevertheless, we managed to recompile the Linux-GPIB driver under the new 3.x Linux kernel. Since VISA support in Linux is rather limited at present, it is recommended that to use the VISA-VXI drivers providing the API with data and control messages sent over TCP sockets. On the other hand, the VISA drivers for Windows may be preferred because these drivers are well maintained and their installation is straightforward.

Overall, we recommend using the VXI family of VISA drivers to interconnect the testbed components. These drivers are based on the widely used TCP/IP protocols, so they support the IEEE 802.3 (Ethernet) interfaces as depicted in Figure 3. In this case, the testbed control over the IEEE 802.3 network can be readily established using an unmanaged switch. This approach also has the advantage that it is mostly independent of the actual OS used.

"…the testbed control over the IEEE 802.3 network has the advantage that it is mostly independent of the actual OS used."

Software Components of the MIMO Testbed

The MIMO testbed consisting of the interconnected hardware cannot be operated without the appropriate software components. The software is used to configure and control the testbed either directly via the defined APIs, or indirectly via the user interfaces. Graphical user interfaces (GUIs) are Windows OS applications normally running on a PC; exceptions are the Propsim channel emulator and the Infiniium DSO implementing the GUI within their internal Windows OS. The Windows OS allows for the exploitation of other OS services such as folder sharing and running a web server to remotely configure the DSO from a web browser. The GUI often greatly extends the functions accessible by the hardware user interface consisting of buttons, switches, rotary buttons, and keypads. However, the functions accessible via the GUI are only a subset of those provided by the API.

All functionality of the SR5500 channel emulator is accessible via the instrument control software TestKit. This software provides the GUI and a collection of various tools and utilities. However, some of these tools and utilities can only be used if the channel emulator is controlled via the Ethernet network due to high data rate requirements. In particular, Test Assistant automatically configures the channel emulator for various device testing applications. The time-varying Power Delay Profile modeling can support evaluations of the adaptive (Modulation & Coding, AMC) schemes. A Channel Wizard configures the MIMO channel models. Interference Editor allows real-time changes of the additive background noise levels to emulate the interference. Channel Player is a key feature of the TestKit. It enables greater control over the channel emulation, because it provides functions similar to software debugging (set timers, break points, pause emulation, and so on).

"Standard Commands for Programming Instruments (SCPI) defines a universal comprehensive set of commands for programming the instruments…"

"The MATLAB ICT toolbox provides support for data transfer to and from the instrument, access to instrument's hardware resources, evaluation of the instrument's status,…"

Dynamic Environment Emulation (DEE) feature allows you to dynamically (that is, in runtime) change the parameters of the MIMO channel. Fading Lab can reply the channel data from real-world measurements (using, for example, channel sounder) or simulations (using, for example, ray tracing).

Standard Commands for Programming Instruments (SCPI) defines a universal comprehensive set of commands for programming the instruments using the API. The commands and the data are formatted and are device-independent as well as actual physical-interface-independent. In the case that the instrument connectivity is realized with Ethernet, the SCPI commands are sent as the application layer messages between the TCP sockets using the VXI driver. The SCPI commands can be readily sent to the instruments attached to a PC using a MATLAB interpreter, Instrument Control Toolbox (ICT), and a free Waveform Download Assistant developed by Agilent Technologies. The MATLAB ICT toolbox provides support for data transfer to and from the instrument, access to instrument's hardware resources, evaluation of the instrument's status, overall instrument control using a GUI, event-based communications, and a range of communication protocols including TCP, UDP, VESA, GPIB, USB, and VXI. Other alternatives for establishing communications and control of the instrument from an OS are higher programming languages supporting TCP socket programming such as C++, Java, and Python. The Agilent I/O Library Suite also supports a client-server model for communication and control of the T&M instruments. This model is especially useful for providing a remote access to the equipment. The clients make requests to the server, which is directly connected to the instrument of interest. In this setting, the instruments are assigned their own IP (Internet Protocol) address; however, the VISA-oriented protocols require that the instrument's IP address is in the same sub-network as the server.

Finally, the 89601B software from Agilent Technologies provides complex signal manipulations and analysis for the modular VSA 89600. All VXI-bus based devices with LAN, GPIB, USB, and FireWire connectivity are supported. The measurement receiver capabilities for modulation analysis are implemented as specific (separately priced) options for a wide range of wireless communication systems.

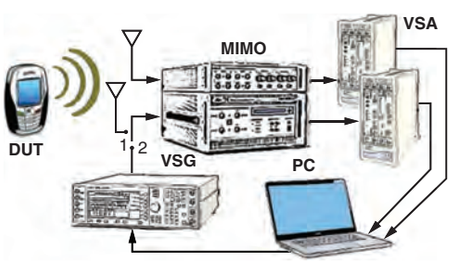


Figure 4: Device under transmitting test using the MIMO testbed
(Source: Swansea University, 2014)

Scenarios and Case Studies

In this section, we describe opportunities and limitations of the MIMO testbed in conducting various experiments in the R&D of wireless communication systems. These experiments can be broadly categorized as equipment testing and algorithm validation.

Device-under-Test (DUT) MIMO Testbed Experiments

Consider first the scenario involving a DUT. The MIMO testbed setup of hardware components is shown in Figure 4 and Figure 5 for the transmitting and receiving DUT, respectively.

In both scenarios, it is recommended to conduct the testing inside the antenna anechoic chamber (more precisely, only the antennas and the UE should be placed inside the anechoic chamber in order to avoid reflections off the T&M kit) in order for the link between the DUT and the receiving (in Figure 4) and the transmitting antenna (in Figure 5) to appear as a time-invariant frequency nonselective Gaussian channel. Such a channel neither distorts the RF signal in time nor in frequency, so the underlying wireless link can be considered ideal (albeit having some fixed attenuation and phase shift). The corresponding end-to-end channel is then determined, and thus, fully controlled by the MIMO hardware emulator.

The DUT experiment in Figure 4 can be set up in two uplink modes (that is, from the DUT to an emulated access point) depending on the position of the switch at the second input of the MIMO channel emulator. In the first switch position labeled as "1," the signal transmitted by the DUT is received at both receiving antennas at the two MIMO emulator inputs. Under the anechoic propagation conditions, possible attenuation imbalances and phase shifts between the received input signals can be automatically digitally compensated by the MIMO emulator[4] to ensure that the received signals are identical. The specific MIMO channel emulator configurations are discussed in the next subsection (see Figure 6a).

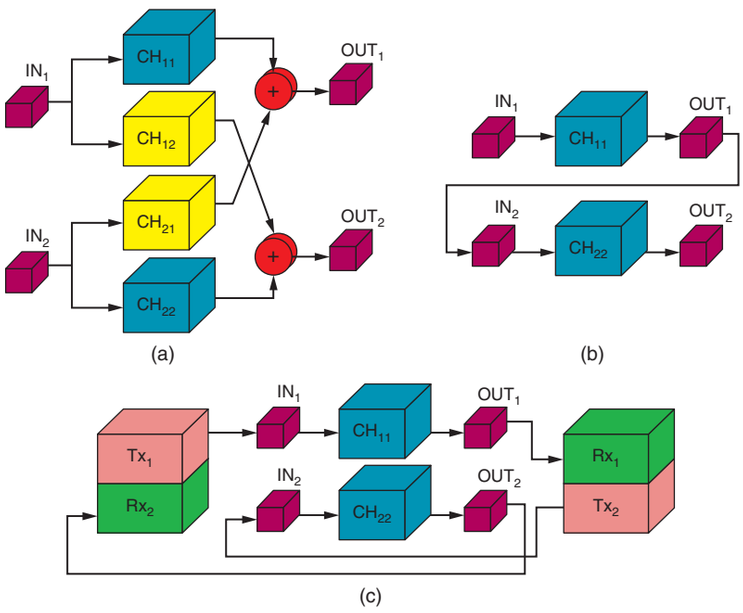


Figure 6: The MIMO testbed configurations: basic 2x2 MIMO channel emulator setup (a), amplify and forward relaying (b), and bidirectional SISO link (c).
(Source: Swansea University, 2013)

The switch in Figure 4 in the second position labeled "2" generates the second input signal to the MIMO emulator by the VSG. Thus, this setup enables

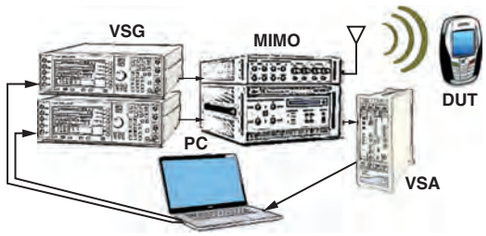


Figure 5: Device under receiving test using the MIMO testbed
(Source: Swansea University, 2014)

"…it is recommended to conduct the testing inside the antenna anechoic chamber…"

the creation of a fully controllable and reproducible interference channel. The interfering signal fed from the VSG can be a sum of several interferers to represent more than one source of interference. For the single interferer, the VSG generates a modulated signal that is distorted by the multipath channel between the second input and the first output of the MIMO emulator (the channel CH_{21} in Figure 6a). However, the case of multiple interferers requires that the sum of modulated signals and their corresponding distortions by multipath channels is already generated in the VSG while the channel CH_{21} (see Figure 6a) is turned off, that is, replaced by a pure connection. Such overall interference generation by the VSG represents combination of the analytical interference signal modeling with the hardware (real-time) channel emulation. A typical scenario where multiple interferers affect the transmitted signal is the co-channel interference (CCI), which is often present at the cell edges due to the transmissions of other users located in the nearby cells. Another type of interference that can be precisely generated via the VSG in Figure 4 is the adjacent channel interference (ACI), which often arises, for example, in Wi-Fi networks due to frequency overlapping of the transmission channels.

"A typical scenario where multiple interferers affect the transmitted signal is the co-channel interference (CCI). Another type of interference is the adjacent channel interference (ACI),…"

The MIMO testbed setup depicted in Figure 5 can be used for the DUT experiments in the downlink (that is, from an emulated access point to the DUT). Importantly, since the signals simultaneously transmitted from two antennas interfere at the receiving antenna, only a single transmitting antenna attached to one of the MIMO emulator (RF) outputs can be used, unlike in the experiment in Figure 4 where possibly two receiving antennas at the MIMO emulator inputs can be used. Recall also from the above discussion that, in the anechoic chamber, the wireless link in Figure 5 can be considered to be an ideal Gaussian channel that does not distort the transmitted signal. The modulated signals are generated at the two VSGs and fed to the MIMO channel emulator. These signals can be either be a so-called transmitter diversity scheme (to be discussed in next subsection), or one of these signals represents the interference (see the previous paragraph for the interference discussion in Figure 4). Capturing the received signal from the second output of the MIMO emulator via the VSA is optional, but it can serve as a reference. In particular, the MIMO channel emulator can be configured in such a way that its output signals are identical, so one can more readily assess the expected performance for the DUT, knowing exactly the channel realizations and the received signal. Another option is to attach antennas to both outputs of the MIMO channel emulator in Figure 5 and then switch between these antennas to emulate, for example, the antenna switching diversity and the antenna (access point) handovers (see the next subsection for more detail).

"Another option is to switch between these antennas to emulate, the antenna switching diversity and the antenna (access point) handovers…"

MIMO Testbed Experiments for Algorithm Development

One may argue that whereas the industry is focusing on DUT experiments for validity and conformity purposes, academia is more focused on carrying out experiments supporting algorithm development. In this subsection, we provide an overview of the transmission techniques relevant for algorithm development and the underlying configurations of the MIMO channel emulator.

Assuming the basic 2x2 configuration of the MIMO channel emulator (with or without the 6 GHz RF option), its internal block structure is shown in Figure 6a. We observe that the 2x2 MIMO channel consists of four single-input single-output (SISO) channels. The channels CH_{11} and CH_{22} directly interconnect the input and output ports of the MIMO emulator, whereas the channels CH_{12} and CH_{21} represent the cross-interference between the direct channels CH_{11} and CH_{22}, respectively. All SISO channels can be independently configured. For instance, the cross-channels CH_{12} and CH_{21} can be turned off to obtain two independent SISO links. In some scenarios, either only one input or only one output of the MIMO channel emulator are used, so some of the SISO channels can be turned off. The powers of the input and output signals of the SISO channels are digitally measured in order to normalize the signals and accurately set the resulting values of SNR or CNR at the output of the MIMO emulator. In addition, the phase shift errors between the input and output signals of the SISO channels due to real-time emulation of the RF signals can be tracked and accurately compensated, which is important for some transmission schemes such as antenna beamforming.

The transmission diversity schemes rely on the joint design of the transmitted modulated signals. For instance, the transmitter antenna beamforming exploits the phase-shifted but otherwise identical transmitted signals in order to constructively add the transmitted signals at the receiver antenna and to maximize the SNR. Knowledge of the channel phase shifts at the transmitter required by the antenna beamforming techniques generally requires a feedback channel in order to report the channel coefficients estimates from the receiver to the transmitter. In the MIMO testbed, it may be more efficient to exploit exact knowledge of the channel coefficients obtained from the MIMO channel emulator control outputs. Knowledge of the channel coefficients at the transmitter is also required for AMC schemes. The feedback signaling from the receiver to the transmitter is also required for other adaptive transmission techniques such as automatic repeat request (ARQ) retransmissions and its variants.

"The transmission diversity schemes rely on the joint design of the transmitted modulated signals."

The MIMO testbed configuration in Figure 5 can be used to evaluate cooperative transmission techniques such as so-called coordinated multipoint transmissions where two or more modulated signals are simultaneous transmitted from different nodes in the network (typically, but not necessarily, the base stations in the cellular network) in order to maximize the SNR and minimize the interference. The transmission schemes distributed across the nodes in the network are a very active area of the current research, and they include, for example, the design of distributed modulation and coding schemes.

"The transmission schemes distributed across the nodes in the network are a very active area of the current research,…"

As discussed above, the modulated signal at the second input port of the MIMO emulator in Figure 4 and Figure 5 can be used to create a controlled interference channel and design and validate the interference suppression schemes at the receiver. More generally, assuming the 2x2 MIMO emulator configuration in Figure 6a, we can emulate two independent but geographically co-located and thus mutually interfering transmission links.

"Multihop transmissions are of great interest to reduce the transmission distances between the source and destination nodes…"

Multihop transmissions are of great interest to reduce the transmission distances between the source and destination nodes in the network. It is straightforward to configure the MIMO channel not only as SISO, SIMO and MISO channels, but also as a multihop system. In particular, one can readily emulate the two-hop Amplify and Forward (A&F) relaying using the MIMO configuration shown in Figure 6b. Note that the two-hop Decode & Forward (D&F) relaying can be implemented in the arrangement in Figure 6c. Therein, the first output of the MIMO emulator is directly connected to its second input, possibly via an attenuator, or we can rely on the automatic scaling of the signals by the emulator. The first output of the MIMO emulator directly connected to the second input then represents a relay node performing the A&F relaying. Emulation of more sophisticated relaying schemes beyond A&F may require additional hardware resources.

The MIMO channel emulator configuration for emulation of a bidirectional SISO link is indicated in Figure 6c. The implementation of transmitters and receivers at both ends of the SISO link in Figure 6c determines the emulation strategies. In particular, employing the VSGs and VSAs for the transmitters and receivers (as in Figure 4 and Figure 5) will generally lead to a batch processing, that is, the signals are transmitted and processed as (possibly very long) frames. However, we have also successfully interconnected antenna terminals of the Wi-Fi development boards directly to the I/O ports of the MIMO emulator. The latter configuration of the MIMO testbed allows real-time continuous evaluations of the TCP/IP protocol stack under various radio propagation conditions.

"…real-time emulation of the wireless transmissions can be particularly attractive spectrum sensing."

Finally, we mention several other applications for which the MIMO testbed and real-time emulation of the wireless transmissions can be particularly attractive. The first such application is spectrum sensing. Spectrum sensing is critical for efficient allocation of scarce bandwidth and other radio resources. Even though spectrum sensing solutions can be developed using computer simulations, true validation of spectrum sensing algorithms is only possible with realistic channels that exhibit high levels of uncertainty, unpredictable interference, and with the actual RF hardware components at the receiver front-end. Another application of interest is evaluation of transmission nonlinearity, typically due to a high-power amplification (HPA) at the transmitter. The simplest method is to use a nonlinear transformation of samples of the modulated signal in the baseband prior to the VSG (see Figure 2). A more sophisticated approach would enable configuration of nonlinear transformations of the input signals of the MIMO emulator; however, to the best of our knowledge, such an option is not currently provided.

Conclusion and Discussion on the Future of Hardware Emulation

Here we summarize this article and outline the future of hardware emulation of wireless communication systems. In particular, in this article, we first highlighted a growing interest and need for emulation of wireless communication

systems to complement the more traditional approach employing computer simulations, and how it very positively impacts the cost and duration of the system development. This trend is also being fueled by diminishing costs of the hardware and software components for building the wireless testbed, which makes it an affordable R&D option even in many academic institutions. We then gave a high-level mathematical description of a MIMO communication link that is emulated by the hardware testbed and outlined modeling methodologies of spatiotemporal radio propagation channels. We provided a detailed description of the hardware and software components comprising the wireless MIMO testbed and focused on explaining how these components can be interconnected. Our recommendation is to use LAN (Ethernet) in lieu of somewhat obsolete GPIB interconnectivity. A significant part of the article was devoted to illustrate usage scenarios and experiments where the MIMO testbed can be used to support testing of wireless devices and development of algorithms. For the latter, we specifically pointed out a number of scenarios where the MIMO testbed is likely to be of current interest to the research community.

"Our recommendation is to use LAN (Ethernet) in lieu of somewhat obsolete GPIB interconnectivity."

Wireless MIMO Testbed in a Cloud

We conclude this article by depicting our vision of future developments in the hardware emulation. Future wireless communication systems will see coexistence of multiple technologies that not only complement but also actively cooperate within new complex network architectures based on mesh and centralized-distributed hybrid topologies in a highly opportunistic communications which will necessitate fundamentally new approaches to T&M of these systems. In addition, since the hardware emulation is first of all a (huge) computational task that is often constrained by the real-time requirements, it is useful to examine the current trends in computing and adapt the modern computing techniques for the emulation of communications systems.[15] Specifically, we consider virtualization and cloud computing as the two techniques that are likely to change how the emulation of communications systems is going to be implemented in the future.

"Future wireless communication systems will see coexistence of multiple technologies that not only complement but also actively cooperate…"

The wireless MIMO testbed fully interconnected via a LAN offers the possibility of remote access over the Internet. The users may be submitting their emulation requirements as jobs being scheduled for automatic execution on the testbed depending, for example, on the job priority and availability of the testbed resources. Our vision of such a platform being an excellent example of the so-called emulation-as-a-service (EaaS) is depicted in Figure 7.

The key feature of such a platform is providing a cost-effective as well as maintenance-effective solution for establishing a high-end wireless laboratory that can be shared by many users worldwide and that would otherwise be unaffordable by many of its potential users. The amount of data to be transmitted between the emulation cloud and the user is dependent on the scenario; for example, the user may need to upload the measured radio-channel data to the cloud prior to emulation. Furthermore, provided that sufficient computational resources are available, virtualization of emulation will enable the creation of almost arbitrary T&M instruments operated in real time.

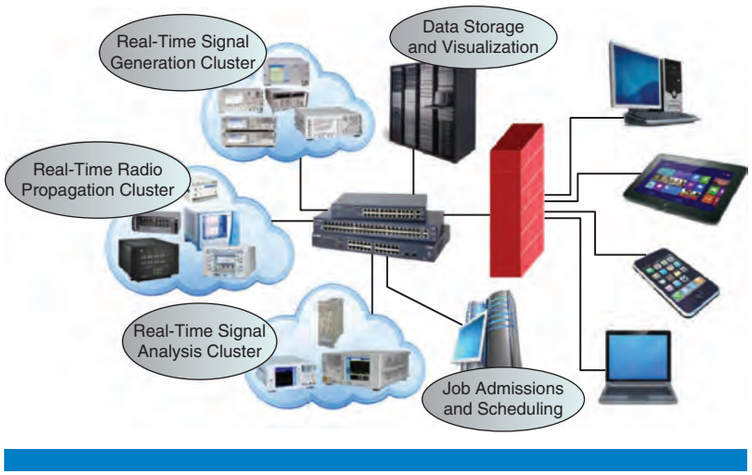


Figure 7: T&M Cloud offering emulation-as-a-service (EaaS). (Source: Swansea University, 2014)

"…cloud computing and T&M instrument virtualization is likely to redefine the current T&M industry companies may establish large professionally maintained clusters of the T&M equipment with worldwide access over the Internet."

In summary, cloud computing and T&M instrument virtualization is likely to redefine the current T&M industry and how the emulation is delivered to the users in academia and industry. For instance, Agilent Technologies or other such companies may establish large professionally maintained clusters of the T&M equipment with worldwide access over the Internet.

Acknowledgements

The authors want to acknowledge financial support of Knowledge Transfer Centre (KTC) HE 09 KTC 1002: "MIMO TestBed – Supporting the Wireless Economy," and the European Regional Development Fund (ERDF) from the Welsh European Funding Office (WEFO).

References

[1] Kermoal, J. P., L. Schumacher, K. I. Pedersen, P. E. Mogensen, and F. Frederiksen, "A stochastic MIMO radio channel model with experimental validation," *IEEE J. Selected Areas in Communications*, vol. 20, no. 6, pp.1211-1226, Aug 2002.

[2] 3GPP TS 25.104, Release 12, Version 12.1.0. Annex B: Propagation conditions.

[3] Anite Propsim Radio Channel Emulator. Technical Datasheet available at: http://www.anite.com/.

[4] Spirent SR5500 Wireless Channel Emulator. Documentation available at: http://www.spirent.com/Products/SR5500.

[5] Pinola, J., Perälä, J., Jurmu, P., Katz, M., Salonen, S., Piisilä, J., Sankala, J., and Tuuttila, P., "A systematic and flexible approach for

testing future mobile networks by exploiting a wraparound testing methodology," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 160-167, Mar. 2013.

[6] Kyösti, P., J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, C. Schneider, M. Narandzić, M. Milojević, A. Hong, J. Ylitalo, V.-M. Holappa, M. Alatossava, R. Bultitude, Y. de Jong, and T. Rautiainen, "WINNER II Channel Models, "IST-WINNER D1.1.2, ver. 1.1, September 2007. Available at: https://www.ist-winner .org/WINNER2-Deliverables/D1.1.2v1.1.pdf.

[7] Gesbert, D., H. Bolcskei, D. A. Gore, and A. J. Paulraj, "Outdoor MIMO wireless channels: models and performance prediction," *IEEE Trans. Communications*, vol. 50, no. 12, pp. 1926–1934, Dec. 2002.

[8] Costa, N. and S. Haykin, *Multiple-Input Multiple-Output Channel Models: Theory and Practice* (New York: Wiley-Blackwell, 2010).

[9] Radio Mobile propagation tool. Available at: http://www.cplus. org/rmw/english1.html.

[10] Erceg, V., L. Schumacher, P. Kyritsi, A. Molisch, D. S. Baum, A. Y. Gorokhov, C. Oestges, Q. Li, K. Yu, N. Tal, B. Dijkstra, A. Jagannatham, C. Lanzl, V. J. Rhodes, J. Medbo, D. Michelson, M. Webster, E. Jacobsen, D. Cheung, C. Prettie, M. Ho, . Howard, B. Bjerke, L. Jengx, H. Sampath, S. Catreux, S. Valle, A. Poloni, A. Forenza, and R. W. Heath, "TGn Channel Models," IEEE 802.11 document #11-03/0940r4.

[11] Paulraj, A. J., D. A. Gore, R. U. Nabar, and H. Bolcskei, "An overview of MIMO communications - a key to gigabit wireless," *Proc. of the IEEE*, vol. 92, no. 2, pp. 198–218, February 2004.

[12] Agilent Technologies, Electronic Test & Measurements. URL: http://www.agilent.com/

[13] Ziyenge, N. and P. Loskot, "Instrument Connectivity in a Wireless MIMO Testbed," *Smart Wireless Communications,* Manchester, May 2012.

[14] IVI Foundation Specifications. URL: http://www.ivifoundation .org/specifications.

[15] Amazon Elastic Compute Cloud (EC2). URL: http://aws.amazon .com/ec2/.

Author Biographies

Pavel Loskot is a Senior Lecturer at the College of Engineering, Swansea University. He has researched wireless communication systems since 1996. Currently, he is focusing on signal processing problems in telecommunications

as well as in the life sciences. He received the best article award from Chinacom 2012, has filed three patents in wireless communications, and one in knowledge mining. He is a Senior Member of the IEEE and holds a PhD from University of Alberta in Canada. Email: p.loskot@swan.ac.uk

Biljana Badic is a system engineer in the Product Engineering Group at Intel. She has been with Intel since 2010 and has participated in the development of 4G wireless modems with particular focus on the interference cancellation and mitigation algorithms. She is also actively involved in Intel's research activities on 4G-and-beyond systems. Biljana holds a PhD from Vienna University of Technology, Austria and has published over 50 scientific articles and filed over 10 patents on 4G technologies. Email: biljana.badic@intel.com

Timothy O'Farrell is a Chair Professor in Wireless Communications at the University of Sheffield specializing in physical layer signal processing, radio resource management and energy efficient network planning. He has led over 18 major research projects as PI, supervised to completion 17 PG theses and published over 280 research articles, including 8 granted patents. In the context of Mobile VCE (mVCE), he was Academic Coordinator of the Core 5 Green Radio program and currently is a member of the mVCE Steering Committee responsible for developing new research programs. He is a Chartered Engineer, and a Member of the IET and the IEEE. Email: T.OFarrell@sheffield.ac.uk

PRODUCT LINE SOFTWARE ARCHITECTURE FOR MOBILE PHONE PLATFORMS IN UML

Contributors

Alexander Fried
Platform Engineering Group,
Intel Mobile Communications

Thomas Finke
Platform Engineering Group,
Intel Mobile Communications

Valerio Frascolla
Platform Engineering Group,
Intel Mobile Communications

Pascal Lefèvre
Platform Engineering Group,
Intel Mobile Communications

"…A typical modem platform is not perceived to contain a high amount of software, but today the opposite is true."

"…We address our requirements by introducing a software product line, and formally modeling it as a component-based software architecture in UML."

We've developed and successfully introduced a software architecture methodology that uses a central UML model of our multimode LTE baseband software architecture common to all our products. This software product line model is used to define the architecture on different abstraction levels, to secure quality, to derive specific platform architecture specifications, and to guide and validate the implementation. To specify concrete products, we've applied a feature modeling approach, connected with variation points inside the architecture, to manage and define the relevant parts for each of our product platforms.

Introduction

Intel provides and further develops a number of similar but different mobile phone platforms, ranging from 2G entry-level phone to high performance 4G slim modem and smartphone platforms.

As an embedded device, a typical modem platform is not perceived to contain a high amount of software, but today the opposite is true. Counting the baseband software only, which contains for example the cellular protocol stack (responsible to communicate with the higher layers of the cellular network), it is made of more than 800 software components implemented in total in about 50,000 source files containing 4,000,000 lines of code.

Effectively managing large and complex software projects demands an abstracted view on the software. We achieve this by defining and maintaining a component-based software architecture of the baseband software.

Right from the architecture level, we want to make sure that the system realizes all necessary use cases and requirements, with special focus on nonfunctional requirements generally and the so called "-ilities"[5] specifically:

- reusability,
- extensibility,
- understandability, or
- maintainability.

In our solution we address these requirements by introducing a software product line[2], and formally modeling it as a component-based software[11] architecture in UML[1]. The product line is captured as a supserset of all products and is explicitly configurable for each of those. It further serves as a single source for all architecture work, allows us to calculate key performance

indicators (KPIs)[7], is used to generate specification documents and quality reports directly from the model, and finally provides a mechanism to automatically check the compliance of the implementation against the architecture model.

In the section "Software Architecture" we give a short introduction to software architecture in general and then describe in the section "Software Architecture Model" how defining a such a model in the Unified Modeling Language (UML)[8] helped us to define a single and consistent architecture, how we secure the architecture quality, and keep both architecture and implementation synchronized.

The section "Software Product Line Architecture" dives deeply into how to extended such a software architecture methodology from a single platform to a set of them, thereby fostering reuse of software components across different platforms. This is done in a formal way by explicitly modeling variation points in the architecture and connecting them to an external feature model.

Software Architecture

As said earlier, it is the purpose of software architecture to give an abstracted, high-level view on the software system. This abstract view is used both,

"…It is the purpose of software architecture to give an abstracted, high-level view on the software system."

- in the early development phases to define, review, and document important architectural design decisions before detailed design or the actual implementation are worked out, as well as
- after the implementation is ready to provide an abstracted view to ease the system understanding.

The creation of the architecture is based on the overall requirements the system needs to fulfill. In an interlinked fashion, the requirements are recursively broken down to the component and subcomponent level, thereby defining the requirements for each component.

Model Structure

One of the main elements of our software architecture is a tree of components. Starting with the root node of the tree, representing the complete software, complex components are broken down into smaller, less complex ones. This process, together with the requirements breakdown, is recursively applied until we reach a level where the components can be implemented.

Each component is seen as a black box, that is, the component is hiding its internals from the outside. To interact with other components on the same level, components expose interfaces that other components can use. Those interfaces denote the services a component provides and typically include a dynamic description (such as by message sequence charts) on how they are to be used.

"…Each component is seen as a black box, that is, the component is hiding its internals from the outside. "

The main elements of our architecture therefore are:

- A hierarchy of software components
- Software interfaces, provided by one component and used by another
- Dependencies from components to required interfaces
- Dynamic descriptions of the interaction between components and interfaces
- General textual documentation

Implementation

As the architecture provides the high-level view on the software system, it is expected that the actual implementation is according to the architecture. In our methodology, the actual implementation happens on the lowest component level, that is, on these components that are not further broken down into subcomponents. We call those the bottom-level components. All components on top of them are compound components that are defined by their subcomponents and the interaction between those. Those higher layer components can then explicitly define which interfaces are exposed to the next higher level, allowing for active control over the component boundary and providing the basis for modularity.

"For each bottom-level component it is expected that it is implemented in isolation, realizes the defined interfaces, and only makes use of the interfaces that are defined in the architecture."

For each bottom-level component it is expected that it is implemented in isolation, realizes the defined interfaces, and only makes use of the interfaces that are defined in the architecture.

Interfaces are entities in their own right as well and need to be defined in a distinct way at the implementation level, that is, in the case of the C programming language, by a set of header files implementing a specific architectural interface declared outside of both the realizing and the using components.

Software Architecture Model

"In the past the software architecture was purely specified in documents using non-standard ad-hoc diagrams, but as the complexity of the model increases, eventually the limits of it become apparent."

In the past the software architecture was purely specified in documents using non-standard ad-hoc diagrams; these diagrams were created in diagraming tools like Microsoft Visio* and as long as the architecture is simple enough, this approach works out well, but as the complexity of the model increases, eventually the limits of it become apparent:

- Inconsistencies between the diagrams, such as similar but different names given to the same element, to be detected only by manual reviews.
- No defined status of the overall architecture, as its definition is spread across multiple documents, potentially with different versioning schemes.
- Limited support for multiple users doing architectural work in parallel.
- Self-defined diagrams missing necessary explanation and a high likelihood of ambiguity.

- No means to calculate quality metrics automatically.
- Detecting divergence of underlying implementation is hard manual work, such as, for example, checking to see whether only allowed interfaces are used by a component implementation.

Introducing a formal model of our software architecture using UML we perceive several advantages, like

- inherent consistency,
- ability to calculate quality metrics and check compliance to implementation, and
- automatic generation of specification documents in both browsable HTML or printable PDF.

Modeling Architecture

A model is a simplification of reality[1] and is typically used to only focus on specific aspects of a complex problem. There are different kinds of models used in software engineering; some of the typical ones are[6]:

- *Analysis models:* Abstract view on the problem domain.
- *Architecture models:* High-level view on the overall structure of the software system.
- *Design models:* More fine grained than architecture models, focusing on component details.

As in the following we only focus on architecture models, the term "model" is interchangeable with "architecture model" throughout the text.

Applying UML to Architecture Modeling

UML is the de facto standard in software modeling and we chose to use it for the proposed software architecture model. From Booch et al.:

"UML is the de facto standard in software modeling and we chose to use it for the proposed software architecture model."

"The Unified Modeling Language (UML) is a general-purpose visual modeling language that is used to specify, visualize, construct, and document the artifacts of a software system."[1]

UML is standardized by the Object Management Group (OMG)[9], is used all over the world, specifies an expressive modeling language, and is defined via a meta-model.

Working with UML, we need to differentiate between the definition of the elements in the model, including their interrelationships, and the views on it in diagrams. This effectively means that the same element can be used in multiple diagrams and, the other way round, a diagram can only show contents defined in the model. This way, the model provides a means for keeping consistency across multiple diagrams. As a very simple example, structural changes like renaming a component or removing a usage dependency between a component and an interface would automatically be reflected in all relevant diagrams.

"Working with UML, we need to differentiate between the definition of the elements in the model and the views on it in diagrams."

"To implement a formal architecture model, we decided to restrict ourselves to a subset of elements offered by UML…"

"These restrictions allow us to focus on the core aspects of architecture Modeling…"

Architectural Elements in UML

UML was created to be a very versatile and expressive modeling language[8], usable for many different model types, including the ones named in the section "Modeling Architecture". To implement a formal architecture model, we decided to restrict ourselves to a subset of elements offered by UML:

- Structural elements that make up the core structure of the UML model, mainly consisting of the component hierarchy, the interfaces, and the interdependencies between those.
- Diagrams showing both static and dynamic views on the model.
- Documentation elements extending the model with additional textual descriptions.

These restrictions allow us to focus on the core aspects of architecture modeling, ease the introduction to a large software development organization not yet familiar with the details of UML, and finally define a coherent modeling language, an aspect visible in all the architecture diagrams and the generated documentation.

In the following the core elements are described in more detail.

Basic Structure

The main structure of the UML architecture model is based on the component hierarchy, where UML components map 1:1 to the software architecture components. The UML elements used by our methodology are reduced to probably the smallest meaningful set still fulfilling our requirements:

- *UML component:* Used to represent each of the components. Structured in a hierarchy of components, forming a tree.
- *UML interfaces:* Representing the interfaces exposed by components.
- *Dependency relationships:* Modeling usages of interfaces by components. Note that we don't allow direct dependencies between components, but only from components to interfaces.
- *Realization relationships:* A component is providing/implementing a specific interface.
- *Delegate:* The provision of an interface is actually delegated to a subcomponent.
- *Ports:* A communication point used to cross a parent boundary for both dependency and delegate relationships. Required to derive black-box or white-box component views in different diagrams.

An example diagram showing this structure can be seen in Figure 1. In this example the topmost component called Software contains two subcomponents, component A and component B, and each of those are further partitioned in smaller bottom-level components. The realization of interfaces is shown using the so-called *lollipop notation*.[3] Dashed arrows

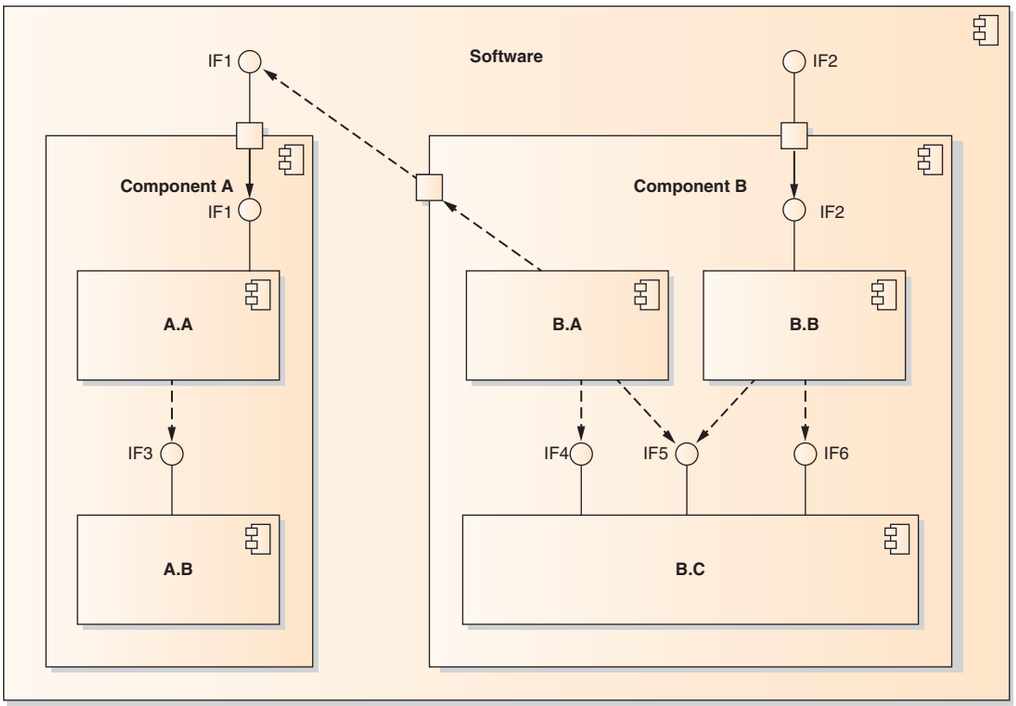


Figure 1: Complete example structure
(Source: Intel Corporation, 2014)

show dependencies and solid ones delegations. In cases of crossing the parent boundary, for example for the dependency from component B.A to IF1, a port is used, represented by a small square.

To better structure other contents of the model, we decided to create an UML package structure to house the component tree. In this package structure each component is represented by two UML items:

"We use an UML package structure to house the component tree."

- A UML component element
- A UML package with the same name as the UML component containing the UML component

In this respect, we deviate on purpose from the UML meta-model[8], where subcomponents would be placed directly beneath their parent components, but the practical reasons and advantages of model management were stronger than strict adherence to the UML standard, allowing the "component" package to contain diagrams, and other sub-packages. In our methodology, these sub-packages can be grouped in four categories:

"…Practical reasons and advantages of model management were stronger than strict adherence to the UML standard…"

- Component packages for subcomponents.
- Dedicated packages for the interfaces defined in the component's scope (see below).

- Diagram packages: These packages contain diagrams only and are used to group them together, and must not contain other architectural elements. See the "Diagrams" section below.
- Documentation packages: Special packages used to add chapters to generated documents. See the "Documentation Elements" section below.

Figure 2 shows how the structure of the example from Figure 1 is captured in the model.

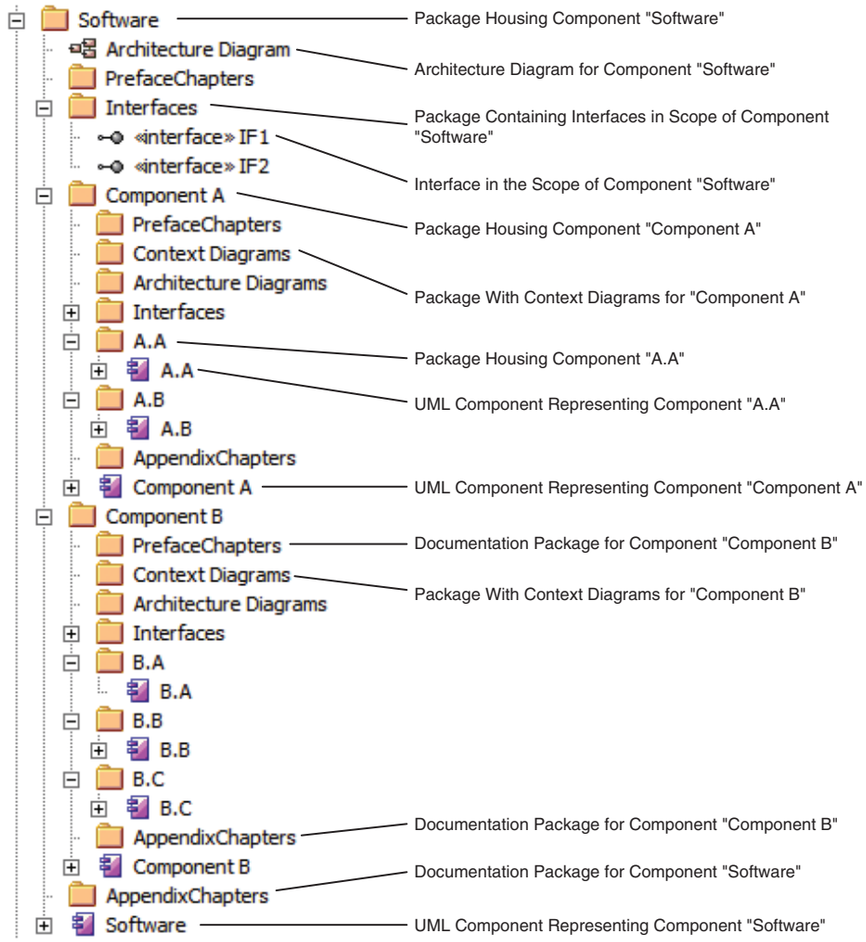


Figure 2: Interleaved package/component hierarchy (Source: Intel Corporation, 2014)

"For each parent component it can then be defined which interfaces of its subcomponents are to be accessed externally and which are private to it and therefore not visible to the outside."

Interfaces

As stated above, each component is treated as a black box and can contain subcomponents that are therefore hidden inside it. For each parent component it can then be defined which interfaces of its subcomponents are to be accessed externally and which are private to it and therefore not visible to the outside. For example in Figure 1, the interface IF3 is private to component A, as it is not explicitly exposed by it and therefore component B would not be allowed to use it.

Architectural interfaces are entities in their own right and modeled using the UML interface element[1], and they differ from their detailed design counterparts in that they focus on both the functionality they provide and on the logical dependency direction. For instance in case of the asynchronous callback, the detailed design would show two interfaces, one implemented by the serving component and one by the served one, whereas in the architecture both are combined into a single interface; the reason being that a callback interface is part of the contract defined by the original one, that is, as a consequence of using the interface, the client has to implement the callback (see Figure 3), resulting in a logical dependency from client to provider only.

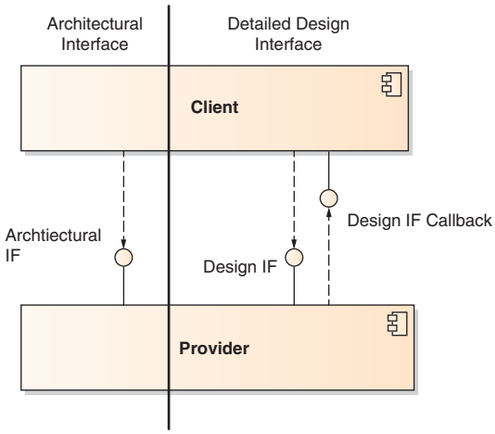


Figure 3: Architectural vs. detailed design interfaces
(Source: Intel Corporation, 2014)

Each component can

- provide interfaces, modeled via the UML realization relationship, or
- require/use an interface, modeled via the dependency relationship.

In case a component is not implemented directly, but instead further broken down into subcomponents, it needs to define which of its subcomponents serves a specific interface. This is done using a delegation relationship connected to a port. In Figure 1, both IF1 and IF2 are such cases.

Using this mechanism, we are effectively able to define modular components by managing explicitly how they are connected internally and what functionality they provide externally. This is of special importance in embedded software environments where the programming language C dominates the implementation[10] and doesn't provide any language mechanism to define visibility levels of interfaces.

"Defining modular components is of special importance in embedded software environments where the programming language C dominates and doesn't provide language mechanism to define visibility levels of interfaces."

Diagrams

The graphical diagram language is probably UML's most prominent feature, but in formally defining a software architecture model, we think diagrams are only of second priority in that they are only able to show what is already defined in the model. Their advantage lies in providing a graphical and potentially filtered view on it.

In our model we distinguish between structural diagrams and dynamic ones. On the structural side, most components will be described by:

- One or more context diagrams showing the component under definition and its surroundings, that is, the required interfaces and their providing components as well as the provided interfaces and their users. This is a black-box view on the component. Figure 4 shows the context diagram of component B from the example above. In simple cases the context diagram will be very similar to the parent component's architecture diagram, in more complex cases it allows the restricting of the view only to other components actually interacting with the component.
- One or more architecture diagrams showing the component's internal structure as a white box. Here we can see which subcomponent is actually

serving an interface. Figure 5 shows the architecture diagram of component B from the example above.

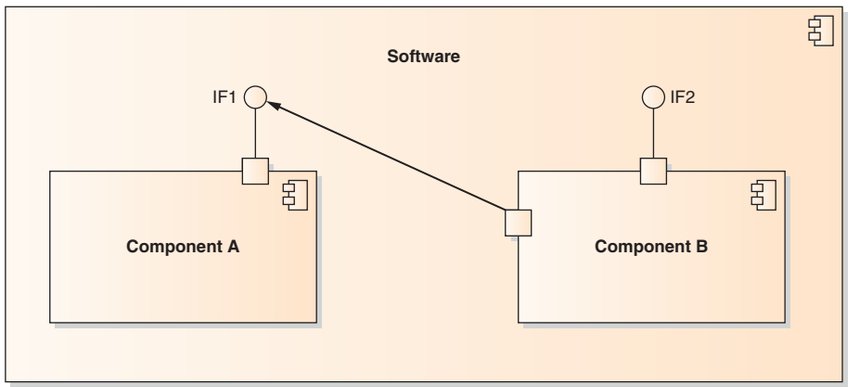


Figure 4: Context diagram
(Source: Intel Corporation, 2014)

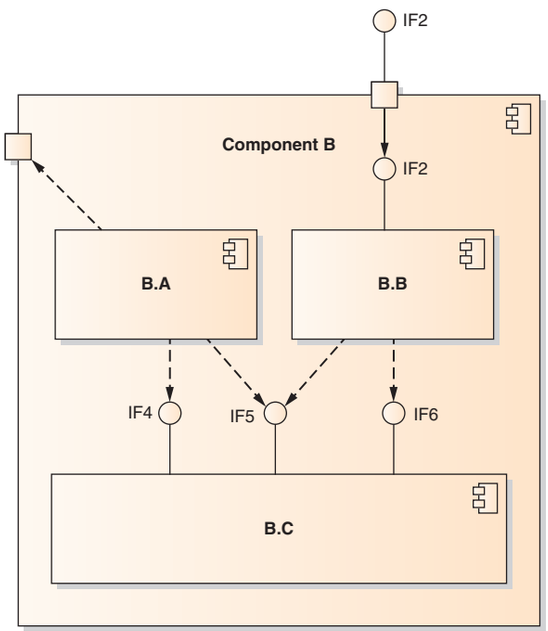


Figure 5: Architecture diagram
(Source: Intel Corporation, 2014)

In simple cases it might be feasible to put all the details in a single diagram, but we saw that in many cases it is likely for the diagram to be overloaded. We promote to split those diagrams up into multiple ones, each focusing on specific aspects (see the section "Diagrams and Subset Models").

"We promote to split those diagrams up into multiple ones, each focusing on specific aspects…"

On the dynamic side, the main diagram type used is the sequence diagram[1] used for both, defining the dynamics of an interface or a dynamic usage scenario

across several components. As we are still at a very high architectural level, specific function calls or messages to be exchanged are not yet defined, instead the focus lies on the principle interaction between the components, described in pseudocode and with basic properties like synchronous or asynchronous defined. Later in the design phase those interactions are worked out in more detail.

To easily distinguish between context, architecture, or dynamic diagrams, simple naming conventions are defined that allow our tooling to identify them correctly.

Documentation Elements

One of our goals is to use the architecture model as a single source for both, defining the architecture and allowing generating specifications and documentation from it. Although the model structure plays the most important part in this, textual documentation is still a necessity. Most UML modeling tools already provide basic support for such textual documentation and allow each model element to be described by a text property. With the exception of ports, we are using this mechanism for all the structural elements listed above.

"One of our goals is to use the architecture model as a single source for both, defining the architecture and allowing generating specifications and documentation from it."

Nevertheless, to complete the documentation and also allow for generation of readable documents, such as specifications, we enhance the architecture model with even more textual elements. For instance, to generate specifications we typically add introduction chapters, additional titled sections, or tables on abbreviations to the generated PDF documents. To capture those in the UML modeling tool, we are exploiting tooling-specific elements together with naming conventions to identify textual elements.

Quality Aspects

Assuring the quality of the software architecture is definitely one of our main goals for formally modeling it in UML. We see several advantages compared to a document-based approach:

- Having a central place where the complete architecture can be accessed by all stakeholders.
- By modeling, a certain consistency is automatically given.
- UML is providing standardized element types and diagrams.

In the end, architecture needs to be reviewed manually. To aid in this process, we are further exploiting the model by doing automatic checks and metrics on it. In this respect we divide quality into three distinct areas:

- Syntactical correctness
- Architectural quality
- Implementation compliance

Syntactical Correctness

With syntactical correctness we focus on the correct usage of UML itself, adherence to our own, more restrictive guidelines on how to use UML for architecture modeling, and on completeness of the model, for example in case an element is missing any documentation.

Examples of such correctness rules are:

- Provided textual documentation for all elements
- All required diagrams in place
- Proper usage of dependencies and delegates, that is, only to interfaces, and using a port to cross a parent boundary

Such syntactical rules are easily checked automatically. Violations get reported in multiple ways, for instance by daily created reports, by allowing users to create them on demand running tool plugins, or by automatically adding them to the generated documents during review phases.

"Architecture quality is much harder to measure than syntactical correctness."

Architecture Quality

Architecture quality is much harder to measure than syntactical correctness. Here we ask the question: "Is the architecture any good?" Unfortunately, this question can't be answered by an automatic formalism, because it depends a lot on the concrete requirements for the product or product line, or on external circumstances. Therefore, we don't even try to measure it directly. Instead, we defined a set of metrics that allow us to identify problematic areas (hot spots) that need to be analyzed manually, and false positives (usually due to a mismatch between what we want to have measured and what is/can actually be measured) are to be expected.

Examples for such metrics are:

- *Coherency:* Measuring the ratio between internal and external dependencies makes it possible to identify potential incoherent components. Those components might thereby reduce the modularity of the system and increase the complexity.
- *Interface scope and users:* The higher up in the component hierarchy an interface is exposed and the higher the number of actual using components, the more important the interface is from an architectural point of view. Any change to an interface that is highly exposed and/or has many clients should be controlled and changes reviewed.

In general it is noteworthy that for most of our metrics, the position in the hierarchy is an important factor, because issues deep down in the tree impact the quality less than on the top-level of the architecture.

"Without a feedback loop that ensures that the implementation is following the architecture, several issues have been experienced in the past…"

Implementation Compliance

Today's standard development environments are lacking feedback between architecture and implementation. Without a feedback loop that ensures that the implementation is following the architecture, several issues have been experienced in the past:

- Compliance between architecture and implementation is not visible. Analyses can only be performed manually by code review, which is a huge effort or even impossible

- Undiscovered deviations lead to outdated architectures not reflecting the actual implementation
- Communication between architecture and development team is reduced to a minimum
- Architectural flaws found during the implementation phase mostly are not propagated back to the architecture. As a result, follow-up projects suffer from already resolved issues

To mitigate this, we created an automatic check, called "Implementation Compliance" based on static code analysis, that allows us to visualize where the implementation deviates from the architecture, for example:

"…We created an automatic check based on static code analysis, that allows us to visualize where the implementation deviates from the architecture,…"

- A component is accessing an interface that it should not use.
- A component is accessing the private implementation of another component.
- Components or code files are present that should not be part of the given configuration.

To do this, the mechanism requires as input:

- The architecture model in UML
- The source code
- A mapping from architectural elements (components, interfaces) to their implementation counterparts (source files)

As a result, in our daily business we experience an increase in communication and feedback flow between architects and implementation teams. Besides closing the feedback loop, this concept also helps us to visualize code dependencies in a clear and interactive way.

"As a result, in our daily business we experience an increase in communication and feedback flow between architects and implementation teams."

Generating Specifications and Reports

Generating documents and reports directly from the UML model not only increases the efficiency, as all the work put into the model can be reused during the development process, it is also able to provide views on the model that cannot easily be seen otherwise. Today we generate:

"Generating documents and reports directly from the UML model not only increases the efficiency, it is also able to provide views on the model that cannot easily be seen otherwise"

- Architecture specifications in HTML and PDF
- Metric reports
- Delta reports

Architecture Specifications

Architecture specifications are an integral part of our development process and a mandatory deliverable from the architecture group. By generating PDF documents from the model, we create up-to-date specifications without much additional effort.

In addition, we generate easy browsable HTML specifications every night. Those allow the organization to access the architecture specification without the need

for a dedicated UML modeling tool. Furthermore, we are able to simplify the presentation compared to using a UML tool directly. For example we generate a single page per component where all the relevant information is gathered in one place. All the component's diagrams are shown, the additional textual documentation is presented, and the provided and required interfaces are listed.

For each interface, we list the providing components and the clients. As components and interfaces are linked, it is easy to jump from the providing component to the using one, thereby exploring the architecture.

"We saw that a browsable HTML export was key for spreading architectural knowledge across the organization, and very helpful to new team members"

We saw that such an export was key for spreading architectural knowledge across the organization, and very helpful to new team members.

Metric and Delta Reports

Quality metrics (described in the earlier section "Quality Aspects") are published in regular intervals as HTML-based reports, allowing the organization easy access to the current state of the system.

In addition, we generate a delta report that is able to compare two versions of the architecture, either over time or between two variants (see the section "Product Variants"). We see multiple uses for this mechanism today, such as controlling the change introduced in the architecture in a regular change control board or visualizing the differences between two configurations for better understanding.

Software Product Line Architecture

Although all our mobile phone platforms are defined by their own feature sets, they have a lot in common. Ignoring this fact and developing each of one of them on their own, either completely independently or forked from a predecessor (clone-and-own), is costly both in time and effort—something we had to learn the hard way.

"Fostering a culture of reuse is key, as many of our software components need to be shared across multiple products…"

"…We implemented a software product line approach, where we can explicitly control variation in the software, allowing for configurability where needed and keeping common parts stable."

Fostering a culture of reuse is key, as many of our software components need to be shared across multiple products, and it is expected that features developed on one platform should easily be put on another one. To fulfill this need we implemented a software product line approach, where we can explicitly control variation in the software, allowing for configurability where needed and keeping common parts stable.

Software product lines were first described by the Software Engineering Institute of Carnegie-Mellon University and are defined by Clemens and Northrop as follows:

"A software product line is a set of software-intensive systems sharing a common, managed set of features that satisfy the specific needs of a particular market segment or mission and that are developed from a common set of core asset in a prescribed way."[2]

In our case, following a software product line approach demanded a complete change in mindset: instead of defining each platform as an individual product,

we now have to think of all of our platforms together as being a single product (line). The actual products are then only variants of the software product line, created by configuring the line. The actual development happens on a single mainline, and using both compile-time and runtime configurations, specific builds for the different products are generated.

Superset Architecture Model

From an architecture and methodology point of view, several additional aspects need to be considered in such a scenario:

- Modeling the architecture of the complete product line instead of "just" a single product
- Defining optional parts in the architecture
- Deriving product variants from the product line architecture

Building on top of our existing UML methodology, as described in the "Software Architecture Model" section, we reflected the product line architecture by modeling it as a superset. Before, the architecture of each project was capture in its own model. Now, the product line architecture merges them all into a single architecture model and therefore in a single software architecture. In this superset model all components and interfaces from all the different platforms are related to each other. This approach is well suited for our product line, because our products have a lot of commonality and the majority of the elements in the architecture are applicable to most product variants.

"… The product line architecture is modeled as a superset."

"In this superset model all components and interfaces from all the different platforms are related to each other."

Having all architectural elements of the product line in a single model has several benefits:

- The architecture of the complete product line can be centrally defined and controlled.
- The impact of any change is visible in the full context of all product variants.
- Component variations can be visualized next to each other.

Product Variants

With our superset approach, deriving a product variant from the product line architecture requires the definition of a subset for it. This is done by logically removing those components, interfaces, and relationships that are not part of the specific variant.

This removal process can't be a simple pick-and-choose activity, because it would likely result in a meaningless architecture, missing required components and choosing incompatible ones. Instead, we use two connected mechanisms to explicitly specify the superset architecture and how it can be configured:

- Variation points define those places in the product line architecture where a variation can happen, for example by selecting a specific component form a set of options, or removing unneeded ones.

"Variation points define those places in the product line architecture where a variation can happen…"

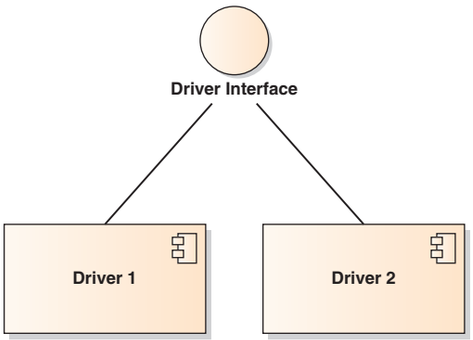


Figure 6: Optional components
(Source: Intel Corporation, 2014)

- Configuration features allow the selection of functionality that should be present in the architecture. Every feature controls one or, typically, more variation points.

An example of a typical pattern for a variation point in our cellular software architecture is the selection of alternative implementations for the same functionality, for instance if different nonfunctional requirements need to be fulfilled (such as memory vs. performance) or different hardware versions need to be supported. Figure 6 shows such an example where the same functionality is implemented by two different drivers. The drivers share a common interface towards their clients, thereby abstracting from the concrete implementation, and either of them can be selected.

Assuming that the example shown in Figure 1 represents a superset architecture, two concrete product variants could look like Figure 7 and Figure 8. In each of those, parts of the superset architecture have been removed to form the final product architecture.

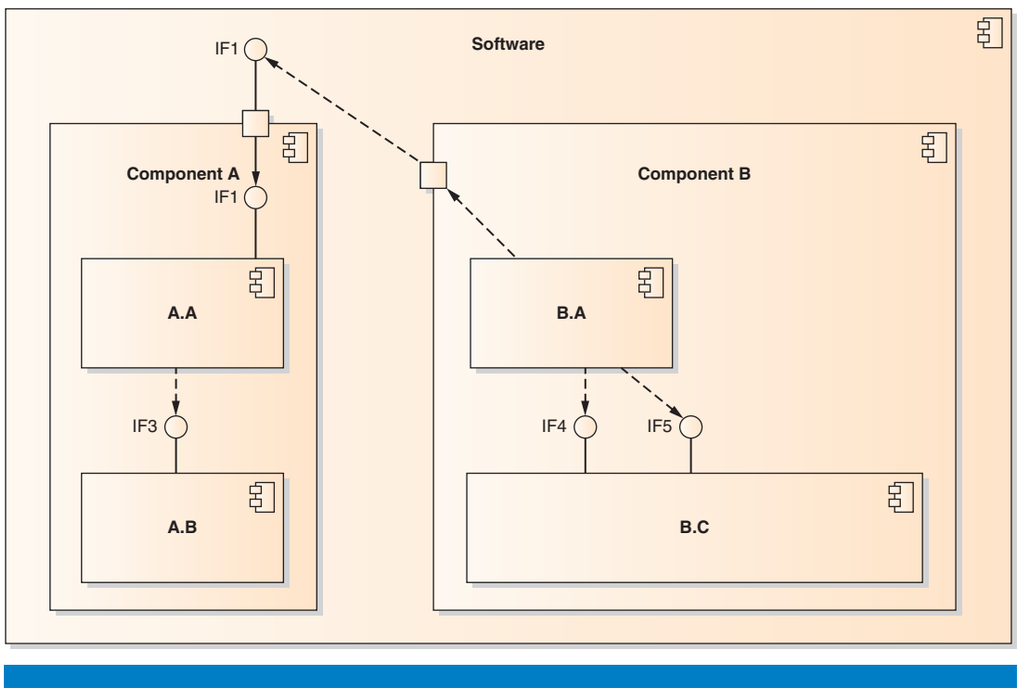


Figure 7: Example variant 1
(Source: Intel Corporation, 2014)

To derive those two product variants, at least two features are required. Let's call them feature X and Y (see Figure 9).

- Feature X is linked to the complete component A, component B.A, IF1, and IF4.
- Feature Y controls component B.B, IF2, and IF6.

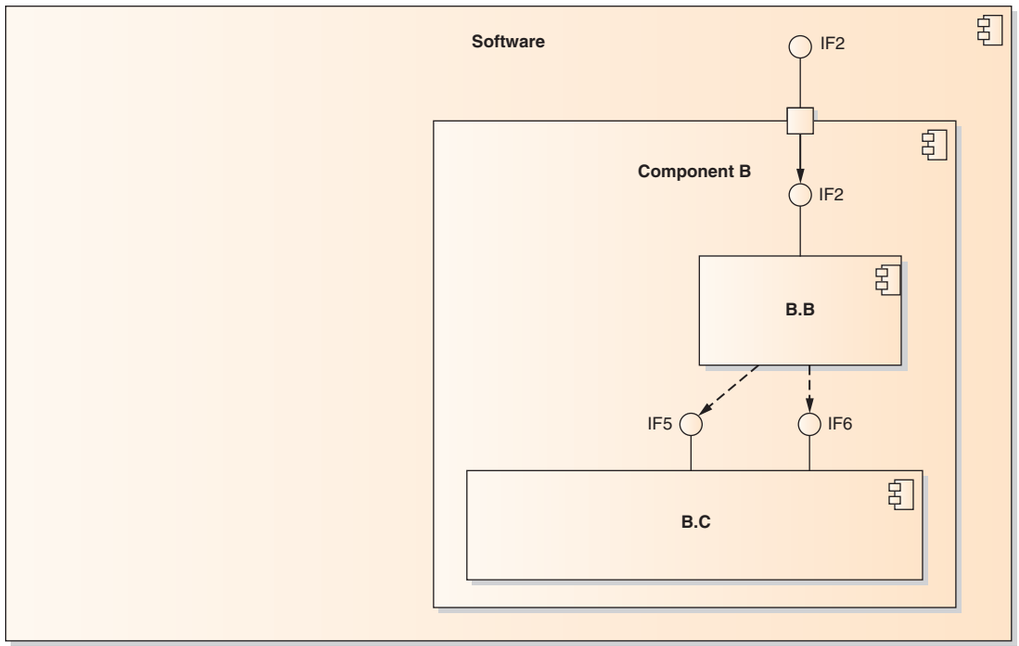


Figure 8: Example variant 2
(Source: Intel Corporation, 2014)

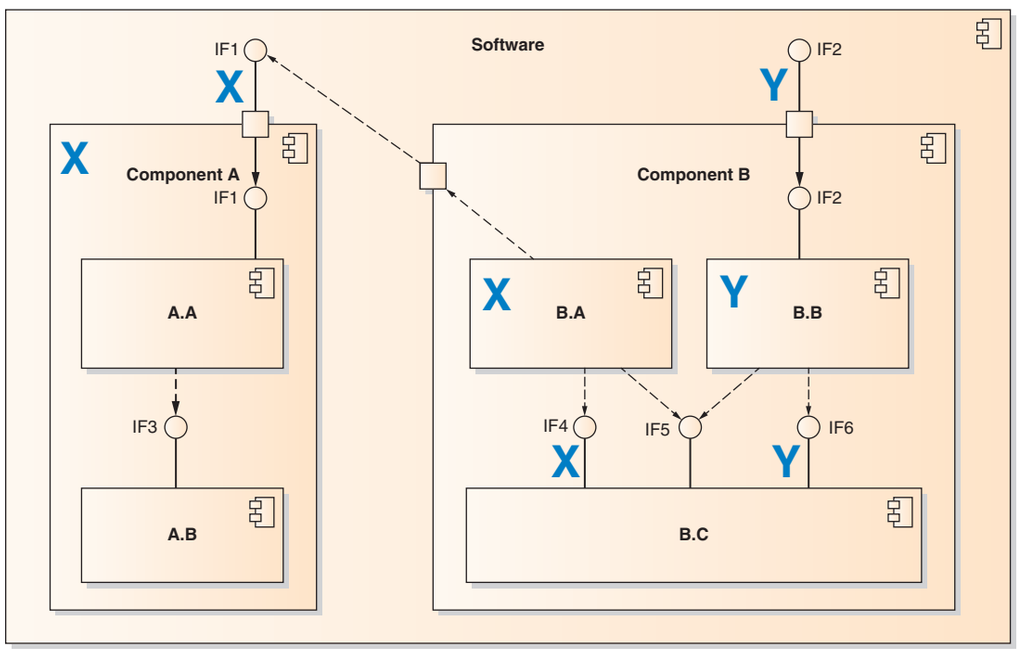


Figure 9: Complete example structure with feature annotations
(Source: Intel Corporation, 2014)

By selecting a configuration where only feature X is present, we would then end up with variant 1, whereas in a configuration only selecting feature Y, the variant 2 would be derived. And selecting both (all) features will give the complete superset.

Feature Modeling

With the introduction of a feature-based configuration mechanism for the product variants, we were able to reduce the complexity of deriving subset architecture, because multiple variation points could be controlled at once. But as the number of features increases, the complexity rises exponentially. For instance, a product line with only 100 features can be configured in 2^{100} different ways, but only a small fraction of those configurations will be commercially or technically meaningful.

To further reduce the number of supported configurations, both from a business and technical view, the space of supported configurations needs to be defined and managed.

We address this by applying the concept of feature modeling, together with the creation of a product line architecture model and its methodology.

"By explicitly and formally defining a feature model, it is possible to reason on its content as well as share and discuss it with others, for example the business lines or the platform teams…"

Feature models consist of[4]:

- *Feature diagram:* Hierarchical tree of features, including the indication of whether or not a feature is mandatory, alternative, or optional
- *Feature definitions:* Name and description of each feature
- *Composition rules for features (constraints):* Valid and invalid combinations of features
- *Rationale for features:* Reasons for choosing or not choosing a feature

By explicitly and formally defining a feature model, it is possible to reason on its content as well as share and discuss it with others, for example the business lines or the platform teams, and thereby bridging the gap between requirements and the technical solution. The feature model is defined using a feature modeler, a tool that allows the derivation of configurations from feature models where each feature is either selected or deselected.

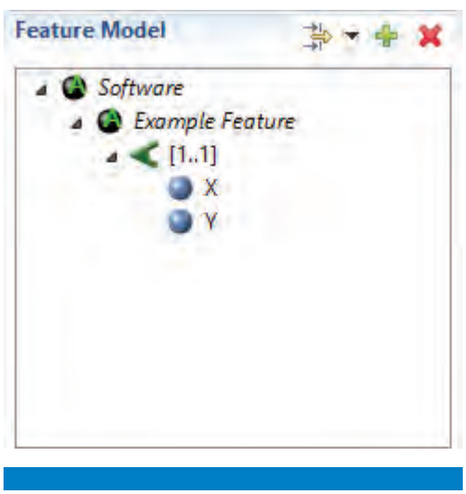


Figure 10: Example feature model (Source: Intel Corporation, 2014)

A screenshot of a small example feature model, based on the features from the "Product Variants" section can be seen in Figure 10. In this example, the features X and Y are put into a feature group, ruling out a configuration where both are enabled at the same time.

The user is guided through the creation of the configuration with automatic application of the feature model's rules, guaranteeing the correctness of the specific configuration in accordance with the definition. This configuration is then used to automatically derive a concrete software architecture based on the superset architecture.

Real-World Challenges

Applying architecture and feature modeling to a real-world development organization, we experienced several challenges that we had to overcome. In the following sections we give a brief overview of selected challenges and outline the applied mitigations.

Impact Visibility

Selecting or deselecting features is a straightforward process, but its complete impact can only be analyzed after selecting a complete configuration. Due to the size of the architecture, the impact of a single feature is then not easily visible.

By enhancing the feature modeling tool with a view that shows an immediate preview of the resulting architecture subset as well as allowing browsing between features and variation points, the user is given direct feedback on his actions. As a result, we could perceive a huge impact on productivity and comprehensibility.

"By enhancing the feature modeling tool with a view that shows an immediate preview of the resulting architecture the user is given direct feedback on his actions."

Figure 11 shows a configuration derived from the feature model presented above. In this configuration only feature Y is enabled. On the right side of the screenshot, the impact on the variation points of the architecture is visualized, showing the elements to be removed as crossed out. The preview of the architecture thereby matches the subset architecture shown in Figure 8.

Change Lifecycle

Today our product development cycles are overlapping. That is, some platforms are in early definition while others are already under heavy development or even approaching completion. As the products are all variants of the same product line, both the architecture and the feature model are continuously adapted.

In the classical feature modeling methodology, such changes invalidate all existing configurations and require them to be created from scratch again. By introducing a change lifecycle, realized via change annotations, we can completely automate the upgrade of a configuration to the latest feature model in most cases and give guidance in the few cases where this is not possible. That way, upgrading a configuration is easy and fast, and only makes feature modeling practical for us.

"By introducing a change lifecycle we can completely automate the upgrade of a configuration to the latest feature model in most cases…"

Diagrams and Subset Models

From a methodology point of view, formally deriving a subset architecture for a product variant based on a given feature configuration is a simple task, as long as only the structural model is concerned. Taking diagrams into account as well is much harder.

Diagrams are an integral part of the generated documentation and need therefore to be included in the product-variant–specific documentation. For

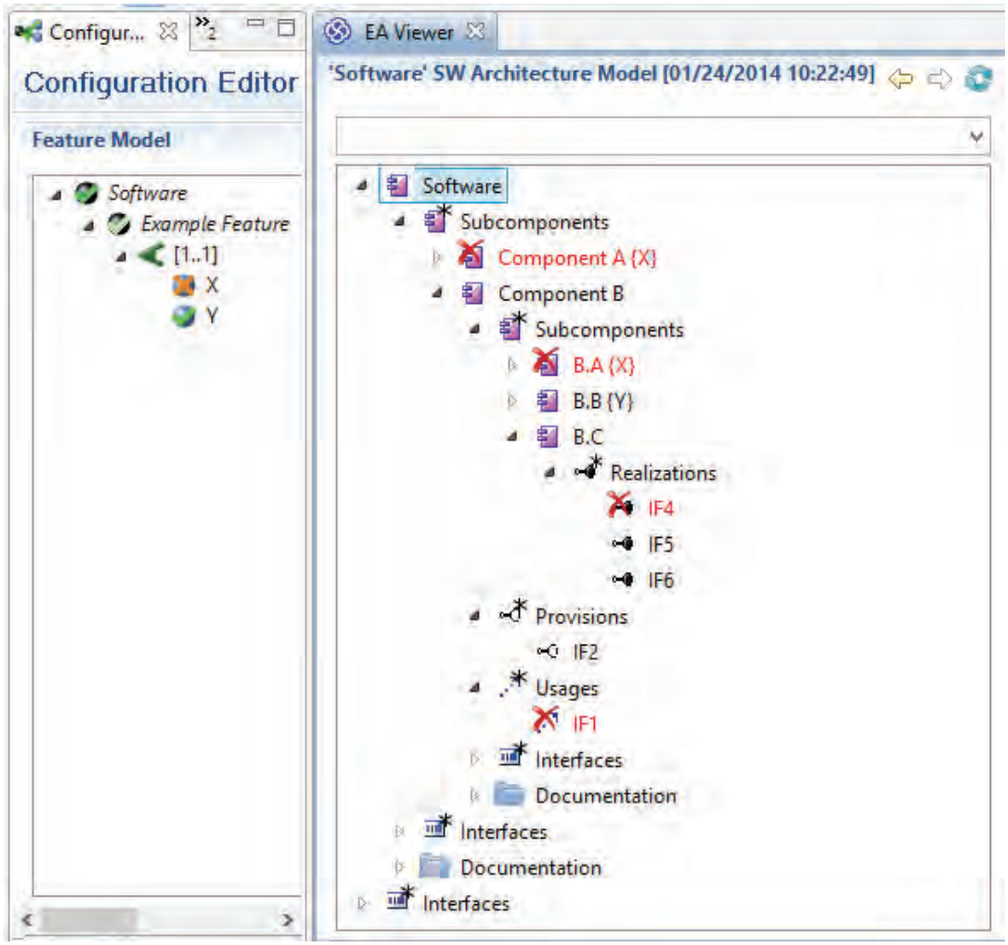


Figure 11: Configuration with architecture impact
(Source: Intel Corporation, 2014)

this documentation, first the structural subset is derived, and based on that the document is generated. To include diagrams we evaluated two potential options, but finally concluded that only one of those is practical:

- Automatic removal of elements from diagrams
- Automatic removal of complete diagrams

In theory it would be easy to automatically remove all those elements from the diagram that are not applicable to the subset. Unfortunately, doing so impacts the visual style and layout of the diagram, and, in extreme cases, completely changes the meaning of it, as core elements might be removed. In addition, it is not possible to adapt the description of the diagram to reflect the removed elements.

Our chosen alternative completely removes a diagram in case it shows at least one element not part of the specific variant. This extreme solution

requires large diagrams to be split up in smaller, aspect-based ones, or alternatively, the creation of multiple variants of the same diagram. This guarantees that major aspects of the architecture are visualized in the generated document.

Results and Outlook

With our work we have shown that a central software architecture model can not only be valuable for a single project, for example by generating specifications from a single consistent source or validating the implementation against the architecture, but such a model can also be successfully deployed in a software product line environment. The superset approach for software product lines makes possible the detection of conflicts between different products in the early architecture phase, because a single solution needs to be defined in the central model.

Our methodology has already been actively used for several years and:

- Several architects are concurrently working on a central UML model.
- All software architectural specifications are generated on a per-project/per-configuration basis out of the UML model.
- The quality of the model and its changes are monitored and published.
- The implementation is regularly checked for architecture compliance.

As a major next step we plan to integrate the feature modeling approach into the software build system. That way, the same mechanism used to derive product variant architecture specifications will then also be used to configure the actual build and ensure that both configurations are in sync.

References

[1] Booch, Grady, James Rumbaugh, and Ivar Jacobson. *The Unified Modeling Language User Guide* (Reading, MA: Addison-Wesley, 1999).

[2] Clemens, Paul and Linda M. Northrop. *Software Product Lines: Practices and Patterns* (Boston: Addison-Wesley, 2001).

[3] Clements, Paul, et al. *Documenting Software Architectures: Views and Beyond,* Second Edition (Boston: Addison-Wesley Professional, 2010).

[4] Czarnecki, Krzysztof and Ulrich Eisenecker. *Generative Programming: Methods, Tools, and Applications* (Boston: Addison-Wesley, 2000).

[5] Laplante, Phillip. *Requirements Engineering for Software and Systems* (Boca Raton: CRC Press, 2009).

[6] Larman, Craig. *Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and Iterative Development* (Upper Saddle River, NJ: Prentice Hall, 2004).

[7] Marr, Bernard. *Key Performance Indicators (KPI)* (Harlow, England: Financial Times/Prentice Hall, 2012).

[8] Object Management Group, Unified Modeling Language, http://www.uml.org.

[9] Object Management Group, http://www.omg.org.

[10] VDC Research, "2011 Embedded Engineer Survey Results – Programming languages used to develop software," http://blog.vdcresearch.com/embedded_sw/2011/06/2011-embedded-engineer-survey-results-programming-languages-used-to-develop-software.html.

[11] Vliet, Hans van. *Software Engineering: Principles and Practice*. (Chichester, England; Hoboken, NJ: John Wiley & Sons, 2008).

Author Biographies

Dr. Alexander Fried received his diploma and doctoral degree in computer science from the Johannes Kepler University Linz, Austria, where he also worked as a research assistant. Afterwards he joined the Infineon group in the mobile phone platform division working as a platform software architect. During this time he created and rolled out the UML-based software architecture methodology. Today he leads the software architecture framework group at Intel Mobile Computing (IMC), where he is both responsible for the software architecture methodology and the overarching software product line architecture. Email: alexander.fried@intel.com

Thomas Finke received his diploma in computer science from the University of Karlsruhe. He works as a software architect inside the system architecture framework group of wireless system engineering on methodologies and concepts for software architecture in the area of cellular platforms. His work includes defining and enhancing the methodology to represent software architecture in UML, maintaining and ensuring the quality of the current central modem platform architecture model, and developing code generators based on model-driven architecture. Email: thomas.finke@intel.com

Dr. Valerio Frascolla earned his MSc in electrical engineering in 2001 and his PhD in electronics in 2004. He worked as research fellow at Ancona University, Italy, and then moved to Germany, joining Comneon in 2006 and Infineon Technologies in 2010. Since 2011 he has been funding and innovation manager at IMC, acting as facilitator of research collaborations using Agile methodologies and focusing on the program management of publicly funded projects and innovation activities. He is author of several peer-reviewed

scientific publications and has been an invited speaker at international events. Email: valerio.frascolla@intel.com.

Pascal Lefèvre received his diploma from the University of Kaiserslautern. He has been working for more than 18 years in the mobile phone industry. He is a general manager of wireless system engineering for the Platform Engineering Group / Wireless Platform Research & Development working on general software architecture as well as related methodologies and concepts in the area of cellular platforms. He and his team are creating, enhancing, and maintaining a product line software architecture spanning from value feature phone to high-end slim modem product segments. Email: pascal.lefevre@intel.com