SoC FPGAs Fuel Next Generation of IoT, Data Center and Communications Infrastructure Applications through Power Efficient Processing

By Jag Bolaria
Principal Analyst

September 2015
SoC FPGAs Fuel Next Generation of IoT, Data Center and Communications Infrastructure Applications through Power Efficient Processing

By Jag Bolaria, Principal Analyst, The Linley Group

This paper was prepared by The Linley Group and sponsored by Altera, but the opinions and analysis are those of the author.

Executive Summary

This white paper describes how the Internet of Things (IoT), cloud computing and “big data” are changing the landscape of computing and how SoC FPGAs help meet the requirements of this rapidly changing environment. Escalating semiconductor development costs are forcing consolidation among traditional processor, DSP and ASSP vendors, leaving fewer implementation choices for system makers. Altera’s SoC FPGAs help meet requirements of this environment with customizable products for a wide variety of applications. Through the use of “fine-grained heterogeneous computing” – where custom accelerators are built into the configurable portion of the device - SoC FPGAs address increased performance requirements while maintaining an acceptable power and cost envelope. While the difficulty of hardware design languages has been a hurdle for FPGA implementations, new OpenCL programming support offers an efficient C-like language to program these devices. These trends are moving SoC FPGAs into the mainstream of computing for the foreseeable future.

Increasing Demands from Big Data

Far from being a monolithic market, the Internet of Things (“IoT”) is an umbrella term for a host of applications ranging from consumer appliances, to industrial controls, to automotive communications and wearable equipment as well as the communications infrastructure and cloud computing to support them. One of the side effects of all this “little data” from millions or even billions of small devices, is that it quickly becomes “big data” as it is aggregated and analyzed.

Big data and accompanying cloud computing are characterized by unique workloads, protocols, algorithms, and rapid change. This creates the need for software defined functionality. Along with software defined functionality, automation is required to manage and reduce costs for this large scale of data processing.

This increasing scale of data processing, transport, and storage also leads to an increased focus on power efficiency. By their very nature of running repeated standard instructions at high frequencies, general purpose processors are not the most cost efficient and power efficient option for all big data workloads. Similarly, while well suited to some large scale processing tasks, GPGPUs are required to run at high frequencies and are well known for their high power consumption requirements. At massive scale, these inefficiencies multiply to millions of dollars, forcing hyperscale customers to look at alternatives for lower cost and power. Though offering lower power by providing silicon tailored to a specific application, ASICs and ASSPs typically
take two years or longer to develop and changes to interface standards or protocols can make them obsolete after they are released. Increasing semiconductor development costs are making ASICs and ASSPs cost prohibitive for many markets as will be explained in the next section.

To summarize, a rapidly changing market has created a need to provide the performance necessitated by big data at the rate of change required by software defined functionality -- with better power efficiencies across numerous applications.

**Escalating Costs Push ASSP Vendor to Consolidation**

The ASSP vendor’s dilemma is that the development cost at each new process node is outpacing revenue growth expectations. For example, we expect development cost to double in the initial transition from 28nm to 14nm technology. Consequently, fewer and fewer markets can justify adoption of process technologies at 14nm and beyond. At the same time, after growing at an annual rate of ~10% or more over the last five years, the semiconductor industry is maturing and growth is projected to be around 5%¹. Corporations such as Broadcom, Intel, and Qualcomm are resetting expectations for around 5% growth.

![Diagram](image1)

**Figure 1. Growing Gap Between Cost and Average Revenue Growth.** The 14nm development costs are increasing faster than expected silicon revenue from the target markets. (Source: The Linley Group and industry research)

---

Figure 1 shows our estimate for development-cost increases over the next five-year period to be around 18% compound annual growth rate (CAGR), which is greater than the 5% average industry expansion and about the same as high growth markets such as data centers, where we expect double digit (16%) percentage growth. Where possible, chip vendors will stay on older process nodes to maintain low costs—but this is not possible for applications which demand higher performance and, thereby, smaller process nodes.

Striving to adapt to this new cost environment, vendors are adopting a new trajectory of focusing on key markets and consolidating costs. This high research and development cost environment is pushing consolidation even amongst the largest suppliers. Consolidation allows vendors to increase market share and maintain prices for current products—resulting in improved margins. In addition to consolidation, vendors are focusing on markets where they can be the number one or two player and at the same time exiting markets where the vendor cannot be a dominant player.

For example TI acquired National and is focusing on mixed signal products. At the same time TI appears to be reducing investment for processors and DSP products. Avago picked up LSI to focus on storage products and then divested out of non strategic products, which included wireless infrastructure processors. NXP is in the midst of acquiring Freescale to focus on IoT client products. The company may well divest Freescale’s embedded processor business. In addition, these vendors are reducing investment or just exiting shrinking markets—this includes standalone DSPs, which are being integrated into SoCs, or implemented in FPGAs or ASICs.

The story is different at market influencers such as Google or Amazon who can control the product requirements and simplify product development. More importantly, these players do not look at the components’ RoI (return on investment) in isolation but instead the leverage that a unique technology can deliver in winning market share from their competitors. In this case the higher cost of developing the silicon is amortized over the end business model, which could be selling more compute VMs (virtual machines) or generating greater advertisement revenue. These players may select an ASIC or FPGA to realize the component level functionality—the former is more efficient, but the latter provides a faster path to market and update flexibility.

Improvements to processor performance and power efficiency require leading process technology. ASSPs offer the best power efficiency but the market must be large enough or in the millions of units to justify the investment and stable enough to remain viable while the chip is in development. In some markets—including data centers, big-data analytics, IoT infrastructure, and wireless infrastructure—the requirements are changing so rapidly that it is virtually impossible to fully define an ASSP two years in advance.

For ASSP vendors, those opportunities are 3-5 years out. By then, the requirements will stabilize and the market will increase to $500 million to a $1 billion. Meanwhile, those markets will remain underserved by the major ASSP vendors. A leading alternative then becomes FPGAs and in particular SoC FPGAs that embed processors to offer a programmable and customizable solution for a broad range of applications.
**Roadmap of Power Efficiency**

Performance improvement for general purpose processors began with increasing clock frequencies then moved to adding more CPU cores. Today, several smart phone processor vendors are using heterogeneous computing – using a mix of different core architectures for different tasks - to get better performance improvements and power efficiency. This heterogeneous computing approach can also be performed at the system level by mixing different compute engines for tasks with different characteristics. Heterogeneous computing is good for workloads that can be parallelized and segmented to a specific core. It can improve power efficiency by up to 10 times compared with general purpose processing.

While multicore uses the identical CPU cores, heterogeneous computing uses different cores – such as ARM’s big.LITTLE, GPUs, DSPs, and FPGAs. Applications that need a lot of floating point operations per second (flops) will benefit from using GPUs, which offer better flops per watt than server processors and are lower priced. There are, however, two downsides with GPUs: devices such as Nvidia Tesla and Intel’s Xeon Phi dissipate high absolute power – typically in excess of 100W – and that works against the primary goal of improving power efficiency.

FPGAs, ASICs and ASSPs offer better gate utilization for a range of applications to further improve power efficiency. Many customers, however, want a software solution that can accommodate new features, protocols and customizations. Software-defined hardware can implement parallel functions of varying complexity in hardware, which then provides the next level of power efficiency improvements. Altera has coined the term “fine-grained” heterogeneous computing to describe a subset where custom hardware accelerators are developed and optimized for the underlying software code.

The next level of power efficiency will come from multiple acceleration tasks that must be configured, managed, monitored, and automated, which is best done by a processor. SoC FPGAs are particularly attractive for fine-grained heterogeneous processing because they have both configurable acceleration hardware and embedded CPUs.

**FPGAs Demonstrate Power Efficiency**

FPGAs can improve power efficiency by implementing only the required functionality in logic gates and by operating at slower clock frequencies than CPUs and GPUs. Slower clock speeds reduce power consumption, and any performance loss is more than offset by parallel execution in hardware.

In the past, the drawback of FPGAs was long and difficult development cycles. New FPGA development environments, such as OpenCL, are simplifying development and reducing development time to be closer to software cycles. OpenCL offers a C-like programming environment while supporting parallelization across multiple compute resources.

2 Source: Bob Broderson, Berkeley Wireless Group
Hyperscale customers have demonstrated power and cost efficiencies by offloading functions onto FPGAs. For example, Microsoft published a paper\(^3\) highlighting the benefits of using Altera’s Stratix V FPGA to offload ranking in big searches. By offloading the server processor with this FPGA, Microsoft doubled the throughput (for ranking) while increasing the cost by less than 30% power consumption by less than 25W (10% of the server’s total power). So compelling is this case that we believe Microsoft is deploying FPGAs in volume in its data centers. This is an emerging market, currently unserved by ASSPs.

FPGAs in general provide more hardware parallelism than is possible with GPUs or CPUs. Thus, tasks that can be parallelized and executed in hardware will be more efficient on an FPGA than in software on a GPU or CPU. For example, Altera ported a Monte-Carlo simulation (which computes risk analysis for different actions and can be applied to finance, oil, gas, insurance, or manufacturing) from Nvidia’s Kepler20 to a Stratix V FPGA. The result was performance improvement from 10 billion simulations/sec to 12 billion simulations/sec and at much lower power. The Stratix V consumed 45W in contrast to the Kepler20 that has a power rating of 212W.

Figure 2 shows another example that compares GZIP throughput for lossless data compression on a Stratix V FPGA with an Intel Core i5 650 processor running at 3.2 GHz. This example illustrates that the parallel arithmetic tasks in this compression algorithm execute at a greater rate in hardware than in software.

---


\(^4\) [http://www.eecg.utoronto.ca/~mohamed/iwocl14.pdf](http://www.eecg.utoronto.ca/~mohamed/iwocl14.pdf). 28-nm Stratix-V A7 FPGA device (25 W) while the CPU measurements were performed on a 32-nm Intel Core i5 650 processor (73 W for 2 cores) running at 3.2 GHz.

©2015 The Linley Group, Inc.
The high level language FPGA implementation using OpenCL at 2.84 GB/sec throughput is within 5.3% of the best known hand-coded Verilog FPGA implementation (“IBM Verilog FPGA”) at 3.0 GB/sec. Though slightly less performance, the OpenCL implementation required lower design effort and time. The OpenCL FPGA version is 12% faster than the fastest known commercial ASIC version (“AHA ASIC”) at 2.5 GB/sec, though the ASIC achieves a higher average compression ratio of 3.6x versus 2.43x for the OpenCL FPGA on the same benchmark set. Further optimizations could be applied to the FPGA code to improve this metric.

**SoC FPGAs Add Programmability**

SoC FPGAs are at the forefront of FPGA evolution from gate arrays to SoCs and are ideally suited for custom acceleration or fine-grained computing. Figure 3 shows this evolution over the last 15 years. At the end of the 1990s, FPGAs were mostly gate arrays with some embedded RAM. In the last decade FPGAs added more hardened IP functions, including transceivers, DSPs, DDR memory interfaces, PLLs, and programmable power technology (PPT). In this decade, FPGA vendors have added physical coding sublayer (PCS) such as Ethernet MACs, floating-point DSPs, and ARM processors. Altera calls these chips which contain embedded ARM processors “SoC FPGAs”.

![FPGA Evolution](image)

*Figure 3. FPGA Evolution. FPGAs have been adding hardened IP since 2000 and are now more SoC than FPGA (Source: Altera)*

Initially Altera SoC FPGAs integrated a dual-core ARM Cortex™-A9 processor. The obvious application is to consolidate designs that already used a processor and an FPGA. There the SoC FPGA eliminates a discrete processor and its associated memory subsystem as well as the interconnects between these components and the FPGA. The result is reduced board space, reduced power from removing interconnects, and potentially lower system cost. Example designs include embedded systems, communications, industrial, and automotive (ADAS) systems.

The quad-core ARM Cortex-A53 processor added to the Stratix 10 SoC gives users the flexibility to have a fast path with reconfigurable hardware and a management software
plane. With more than 4 million logic elements (MLEs), this device also provides plenty of logic for data center acceleration applications.

More interesting are new applications enabled by more powerful SoC FPGAs, which can accommodate system level changes. Customers can write code for the ARM processors, use standard interfaces, develop DSP algorithms, add new protocols using uncommitted programmable gates, and add custom features. For example a customer could use the DSPs and programmable logic for fast-path data and algorithms while using the ARM processor for control and management.

**Potential SoC FPGA Applications**

SoC FPGA applications include image signal processing (ISP) for 3D scanners, sensor fusion and object recognition in automotive Advanced Driver Assistance Systems (ADAS), motion sensing in industrial human machine interface (HMI) applications and distributed processing in large data centers. The SoC FPGA can accelerate specific algorithms and workloads while using the embedded processor for management.

In the near future, users will search the Internet using images as well as text. An SoC FPGA could be used for image classification, which requires a large number of parallel operations. Here, the DSPs and uncommitted gates would be applied to classify images and the ARM processor could be used for priority and orchestration.

Like cloud computing, wireless infrastructure is changing at a fast pace to accommodate more connected devices and traffic. Global operators are rolling out LTE basestations and enhancing coverage hot spots using small cells. To increase coverage and capacity, new deployment models are being added to traditional macrocellular infrastructure; these emerging technologies include more distributed cells (e.g. small cell and heterogeneous networks) as well as Centralized Radio Access Networks (CRAN). SoC FPGAs can be utilized for radio and/or baseband processing. In CRAN, for example, an SoC FPGA can implement the pooled baseband processing that is required to support multiple radio end-points. FPGAs also allow OEMs to rapidly add new proprietary protocols and products.

Though SoC FPGAs are overkill for most IoT wearable and basic sensor applications, they are suitable for IoT gateways and back-end services. An IoT gateway aggregates traffic from multiple end devices to the cloud. They may be used in multiple applications, including factory automation, oil and gas refineries, farms, and smart-cities. One trend, coined by Cisco as fog computing, is to push more computing to smart gateways at the network edge instead of sending all the information to the cloud and clogging up networks. These gateways need to support security, scalability, process management, network configuration and orchestration. Although the gateway will likely be built on servers to take advantage of existing software, an SoC FPGA can offload and accelerate tasks to improve power efficiency and reduce cost. Alternatively, the gateway can be built solely on an SoC FPGA for a fully customized approach.
Conclusions

One of the impacts of IoT is big data which requires performance, networking, and storage to improve at a faster clip than they do in traditional hardware cycles. Unique workloads, protocols, algorithms, and rapid change create the need for software defined functionality – the ability to add new functionality as quickly as it can be conceived and coded.

The high development costs of ASSPs on leading edge process nodes and the resulting industry consolidation are forcing chip companies to target $500+ million markets at the expense of smaller opportunities, leaving many underserved areas. In addition, performance per watt improvements from the heterogeneous computing approach are slowing down, forcing customers to look at custom accelerators and offloads.

These conditions are ideal for SoC FPGAs. Through the use of fine-grained heterogeneous computing - where custom accelerators are added to the FPGA and the ARM processor is used for control and management- SoC FPGAs offer flexibility, good power efficiency, and fast time to market. Unlike an ASSP that targets a large, specific market, an SoC FPGA targets a diverse range of applications, including emerging applications in cloud computing and the data center. Customers can use these economies of scale to get cost effective devices on leading edge process nodes even for smaller volume or emerging markets.

New development environments like OpenCL, enable developers to take advantage of performance parallelism and power efficiency using a software-style workflow to reduce development cycles and improve productivity.

Jag Bolaria is a principal analyst at The Linley Group, which offers the most comprehensive analysis of the microprocessor industry. We analyze not only the business strategy but also the internal technology. Our in-depth reports cover topics including embedded processors, mobile processors, network processors, base-station processors, and Ethernet chips. For more information, see our web site at www.linleygroup.com.