A Tailored Approach to FPGA Process Selection

WP-01194-1.0 White Paper

This white paper examines three categories of process characteristics, relates them to the internal structure of modern FPGAs, and then, in turn, looks at the impact the FPGAs have on the systems that employ them. In particular, a focus on the deployment of so-called FinFET transistors shows how Altera is exploiting Intel's 14 nm Tri-Gate process to achieve a level of FPGA density, performance, and power efficiency not reachable at all on the planar FET roadmap.

Introduction

The characteristics that FPGAs exhibit in a system are determined, through complex relationships, by the characteristics of the semiconductor processes in which the chips are manufactured. In the past, all aspects of every process improved at each node, and the one best process choice for every new device was the newest process with the finest geometry. Today that is no longer true.

Instead, a programmable-logic supplier today must exploit a variety of process alternatives in order to serve the huge range of design requirements in which FPGAs are used. In this white paper, we will examine three categories of process characteristics, relate them to the internal structure of modern FPGAs, and then, in turn, look at the impact the FPGAs have on the systems that employ them. In particular, we will examine the revolution that is gathering momentum around deployment of so-called FinFET transistors, and how Altera is exploiting a unique FinFET process, specifically Intel's 14 nm Tri-Gate process, to achieve a level of FPGA density, performance, and power efficiency not reachable at all on the planar FET roadmap.

Process Characteristics

For the IC designers, there are three categories of process-determined characteristics that together provide a thumbnail sketch of the process. They are feature pitches, transistor behavior, and availability.

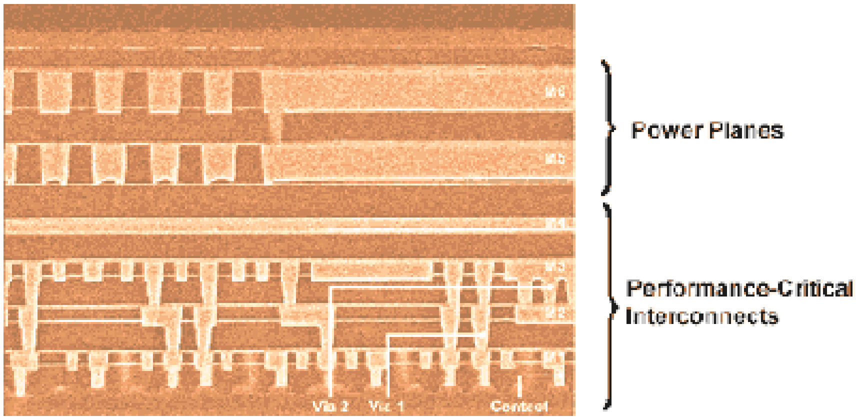
Pitches, the minimum spacings between like features on the finished IC, help determine die size and capacity and, less directly, circuit speed and power. Each layer of features on the die—transistors, local interconnect, contacts, and each consecutively higher metal layer—can have its own pitch. The pitches on these different layers are chosen by the process engineers on the basis of lithography limits and other process constraints, costs, and the way the process designers believe that customers will use the process. These pitches then interact to determine the actual density of transistors in a particular kind of circuit.

101 Innovation Drive
San Jose, CA 95134
www.altera.com

Let’s begin at the bottom. How closely you can pack transistors in a given circuit depends, roughly, on two issues: how closely you can fit the transistors together, and how much space the interconnect to hook them up requires. Either can be a limitation, depending on the circuit design and layout. How closely you can physically pack in the transistors depends, naturally, on their size and shape.

As you move above the local interconnect, contact layer, and the stack of higher metal layers (Figure 1) the pitch becomes rapidly larger. Typically, the local interconnect and lower metal layers make connections between nearby transistors and determine the density of carefully packed structures like standard cells or SRAMs. Higher metal layers connect circuits and, eventually, functional blocks together, implement bus wiring, and distribute power and clock connections. The number of higher layers and their pitches are also important to chip designers, because they can determine the bandwidth and power consumption of connections between portions of the chip.

Figure 1. Metal pitches increase as you move higher in the stack, as illustrated in this legacy Altera® CPLD.



Transistor Characteristics

At the simplest level, there are only three things digital designers care about in their transistors: how large they are, how fast the transistors can switch between ON and OFF, and how much power they consume. For years, the three numbers were locked in a happy partnership: every new process node gave us smaller transistors that switched faster and used less power.

But in recent process generations, as transistors became smaller, power separated into two distinct components: active power, incurred by the action of switching, and static, or leakage power, consumed by current that the transistor is unable to shut off. While speed and active power continue to work together somewhat, with speed going slowly up and switching power going slowly down at each new process node, static power has begun to rise at each new node. Today, if you want a transistor that switches fast, it will leak. If you want low leakage current, the transistor will be slower. As a consequence at the chip level, in some 28 nm SoCs half the power is static leakage power.

Process and circuit designers have fought back. Process engineers have provided chip designers with a variety of transistors with different speeds and leakage currents. Circuit designers have employed careful selection of transistors and their ability to turn down or off clocks and supply voltages to actively manage power consumption. These innovations allow designers to create cell-based digital blocks that exhibit both high peak performance and low leakage.

But the problems with planar FETs continue to become worse. It has become increasingly difficult to further reduce the operating voltage required by planar FETs in succeeding generations. And, today, many process engineers agree that beyond the 20 nm node it will no longer be possible to lower the delay-power product of planar FETs: the figure around which circuit designers must make their speed-power tradeoffs. The conventional planar FET, despite all of the efforts that have gone into extending its life, is running out of roadmap.

Today, many process designers believe that the future belongs to a new kind of transistor: the FinFET, or as Intel calls their version of the device, the Tri-Gate transistor. By, in effect, standing the transistor on its edge and wrapping the gate around the three exposed sides, the FinFET gives the gate much more powerful control over current through the channel, and thus achieves significantly lower leakage at a given speed (technically, a lower threshold voltage) than a planar FET of the same size. The lower leakage allows designers to achieve a given operating frequency at a lower voltage, slashing both active and leakage power, or to operate at a much higher speed at a given total power level.

In addition, because the transistor is standing on its side, the channel width, which influences the drive current among other things, no longer directly limits how closely the transistors can be packed together. So an array of FinFETs can be much denser than an array of planar FETs using the same channel width. Further, because of the geometry of the device, FinFETs can in principle have much greater uniformity than minimum-geometry planar FETs. Process engineers explain that the planar devices have literally become so small that it is possible to count the number of dopant atoms in the channel. An error of a few atoms one way or the other—impossible to control in a production process—makes a significant difference in the transistor’s threshold voltage, and therefore in its electrical performance. FinFETs have proved much less sensitive to this variation.

In several ways, then, the FinFET represents a complete departure from the growing troubles that beset planar FETs. FinFETs can be packed much closer together. They can have lower, and more consistent, threshold voltages, without allowing unacceptable leakage currents. The lower threshold voltage also allows either operation at a lower power-supply voltage for significantly lower power, or operation at a normal voltage for significantly greater speed.

When We Get It

The third vital process issue is not technical at all; it is availability. Specifically, the schedule of device simulation models, test chips, sample shuttles, and production from the foundry must line up with the requirements of the system designer's development schedule. Process models must be available early enough for FPGA designers to estimate the performance their chips will achieve and pass this information to system architects early in the system design flow. Samples and development kits must be available for hardware prototyping and the start of hardware-software integration. And, of course, volume production of FPGAs must be available in time for system production.

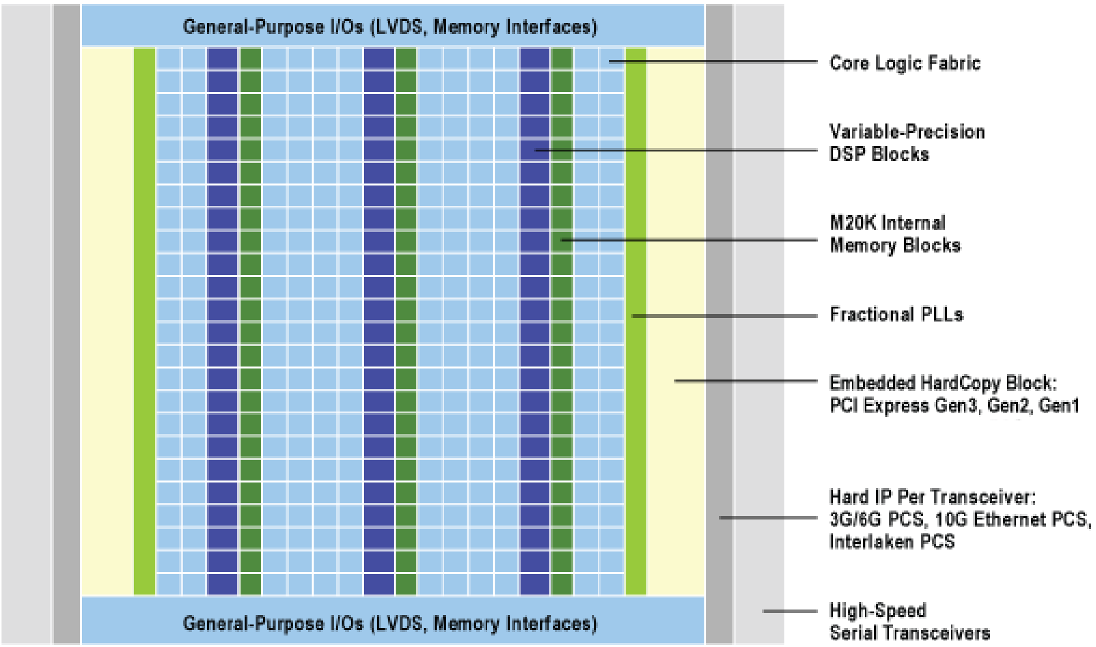
"Availability" is the main answer to the obvious question from the above discussion: why don't we just switch to building everything with FinFETs? But there is another answer as well. In many applications today, existing 28 nm or emerging 20 nm planar FET processes can meet the specific needs of the system design. These system designs don't need to wait for FinFET processes. To see how this works, we need to look at the individual types of structures within the modern FPGA.

Inside FPGAs

The technical characteristics of a semiconductor process translate into system behavior by influencing a number of different structures within the FPGA. A few generations ago, it was accurate to think of an FPGA as a uniform sheet of programmable logic fabric surrounded by a configurable ring of general-purpose I/Os. Today, that model is incorrect.

Modern FPGAs include four distinct categories of functional blocks: programmable logic fabric, cell-based digital intellectual property (IP), hand-crafted block RAM, and custom analog IP (Figure 2). Each of these can respond differently to the characteristics of a new process.

Figure 2. A modern FPGA is a blend of programmable logic, cell-based IP, and analog blocks, as shown in this Altera Stratix® V diagram.


General-Purpose I/Os (LVDS, Memory Interfaces)
Core Logic Fabric
Variable-Precision DSP Blocks
M20K Internal Memory Blocks
Fractional PLLs
Embedded HardCopy Block: PCI Express Gen3, Gen2, Gen1
Hard IP Per Transceiver: 3G/6G PCS, 10G Ethernet PCS, Interlaken PCS
High-Speed Serial Transceivers
General-Purpose I/Os (LVDS, Memory Interfaces)

The FPGA logic fabric is essentially a vast array of replicated, custom-designed logic elements (LEs)—tiny SRAMs, multiplexers, and registers—and switch boxes, overlaid with a highly complex, multilayer pattern of metal segments. Thus the design of the programmable fabric is an artful balancing act between how much functionality to put into a LE and how much interconnect the element will then require. For a given architecture, the overall density of the fabric is clearly sensitive to the pitch of the lower and intermediate metal layers. But because architects attempt to use all the available area under the interconnect stack, it is also sensitive to the packing density of the transistors in the LE. The speed and power-efficiency of circuits that users implement in the logic fabric will depend on the transistor characteristics, but also on fabric density, as well as on the interconnect RC products and transistor drive currents.

So, in general, a process that can offer finer metal pitches and more closely packed transistors will make possible denser logic fabric and both higher performance and lower power consumption for user circuits. Leakage current is a particular issue for the logic fabric because the chip designers, not knowing what the user will do with the programmable logic, has limited ability to use the circuit-level power-management techniques that help greatly with static power in cell-based designs.

Cell-based digital IP, in contrast, has critical paths dominated by fast transistors directly connecting to each other through short local-interconnect or low-level metal segments. This category of structures in modern FPGAs includes digital signal processing (DSP) blocks, I/O and memory controllers, hardened CPU cores, and the like. These IP blocks' size is heavily influenced by the density of the carefully packed standard-cell libraries, and by the variety of cells in the library. And unlike programmable fabric, in which the user can create any circuitry desired, cell-based hard IP is well defined ahead of time, so the chip designer can employ the full range of power-management techniques. Thus, hard digital IP benefits greatly from finer process geometry and greater transistor speed and, at the system level, can use power-management techniques to at least partially offset the higher leakage currents of planar FETs.

Block RAM is a special case of cell-based IP. It is usually built using foundry-supplied, hand-optimized SRAM cells, but the arrays are often tuned by the FPGA designer for optimum speed, density, and power over the range of uses to which the blocks will be put. Because of the great flexibility of the blocks, it can be quite difficult to implement power management strategies for FPGA RAMs. So probably no other structure in the FPGA is as sensitive to the full range of transistor characteristics.

These considerations mean that the best process choice for an FPGA for a specific system application depends on the relative stress that the system design will place on blocks implemented in the programmable fabric and in cell-based logic. The less dependent the overall system performance is on the behavior of blocks implemented in the fabric, the more likely that a midrange FPGA in a 28 nm or 20 nm process can deliver the necessary system performance at attractive cost and on an attractive schedule.

Finally, there is the question of high-performance analog IP, consisting today primarily of phase-locked loop (PLL) and serializer-deserializer (SerDes) circuits. Instead of using minimum spacings, these design commonly use a variety of transistor sizes, circuit layouts, and metal pitches, usually involving hand layout. They are very sensitive to transistor electrical behavior, including some parameters that are of little concern to digital engineers. Where digital designers simulate logic functions, analog designers simulate transistors. Another absolutely key issue for analog designers is uniformity: many standard circuits rely on pairs of closely matched transistors.

There has been a debate over the finFET. Some analog designers point out that you cannot select an arbitrary width for a FinFET. Since the transistors are standing on edge, meaning that the width is now measured vertically, they must all have the same width. You either use one minimum-width FinFET or, if you want more current, you use several in parallel. These designers fear that the new transistors will be difficult or impossible for analog designers to use in their familiar circuit topologies.

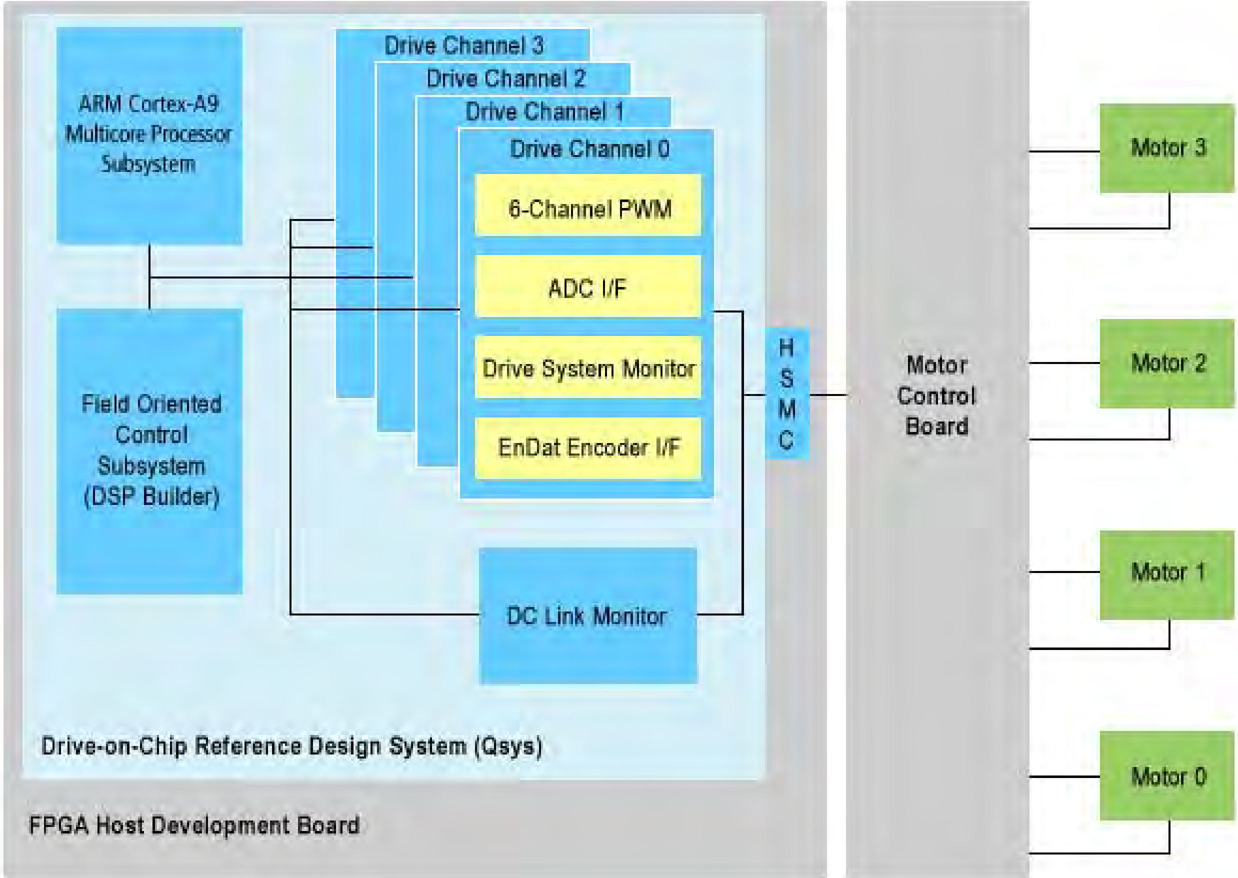
But other experienced analog designers have pointed out that the greater speed, stronger channel control, and—especially—greater uniformity of FinFETs are all strong positives for analog design, far outweighing the quantization of the transistor width. The debate continues, but evidence from Intel’s work with analog structures in the CPUs they have developed using their 22 nm Tri-Gate process indicates that Tri-Gate transistors have, in fact, advanced the performance envelope for precision analog design.

Tailoring Begins with the Right Process

We have seen that process characteristics affect different structures in the FPGA in different ways. Similarly, different applications stress these FPGA structures to different degrees. As a result, there is no one process technology that at a given moment can provide the best platform for a wide range of applications. Schedule, cost, and the performance requirements placed on specific structures in the FPGA all go into the mix, and consequently a tailored approach to FPGA design requires a range of process choices.

Three examples might illustrate this point. First, consider a single-chip motor control SoC (Figure 3). The chip receives position data from shaft sensors on four motors, all at high-kHz rates, and drives four driver boards at low-MHz rates. It connects to moderate-speed DDR2 DRAM for code and data storage, and to an Industrial Ethernet linking the SoC into a factory-floor control network.

Figure 3. The single-chip, multi-axis motor controller combines cell-based DSP circuitry for computing the FOC algorithm, programmable logic for encoding and decoding I/O signals, and a CPU for supervision and functional-safety algorithms.



Internally, the chip supports two main tasks. The first is a field-oriented control (FOC) algorithm—essentially an intense stream of matrix arithmetic for each motor—calculated in the FPGA's DSP blocks. I/O circuits in the programmable fabric decode position data and encode signals for the driver boards at relatively low rates and low energy. The second task is a functional safety package, a set of heuristics designed to protect the machine operators and the integrity of the equipment, running on the SoC FPGA's embedded ARM® Cortex™-A9 CPUs.

There are two primary challenges in this design. The first is that customers want continual improvement in energy efficiency, accuracy, and acoustic noise, all of which demand greater bandwidth and more complex algorithms for the FOC calculations. Thus the application is demanding of the hard DSP blocks and RAMs. The second serious challenge is cost.

Analyzing the situation, the most critical FPGA structures for this application are hard IP blocks, the block RAMs, and, as functional safety requirements tighten, the CPU cores. These blocks, in turn, demand of the semiconductor process good standard-cell libraries, decent SRAMs, and the best possible price. Today, Altera's Cyclone® V SoC products are employing TSMC's 28 nm Low Power (28LP) process to deliver the best combination of high-performance hard IP and memory, low cost, and immediate availability.

Helping Drivers Drive

A second example is a next-generation automotive driver assistance system (ADAS) design. This SoC will receive data from an automotive radar and several HD video cameras, interpret the vehicle's situation using image-processing routines and artificial-intelligence (AI) algorithms, drive two real-time displays, and send commands to vehicle control modules for steering, braking, and drive-train control. Most of this I/O traffic will pass through a redundant pair of 10G Ethernet ports. Due to an aggressive introduction schedule, architectural design on the system must begin in mid-2013.

The challenges in this system are the substantial video and radar signal processing needed to identify objects, the computing needs of the classification and AI routines, and the need for very large amounts of local and external memory at high bandwidth. These needs will fall primarily upon the programmable fabric and its access to DSP hard IP, block RAM, and external DRAM. Because the computing load is sporadic—little activity when the vehicle is motionless or moving slowly, and activity levels based on the complexity of the environment—aggressive power management is feasible. Such an FPGA will need a process with both metal pitches and transistor performance better than that of today's midrange FPGAs in order to meet the performance goals for both the programmable fabric and the hard IP. But the design will not initially need the speed-power advantage of FinFETs. Altera's 20 nm product family, based on TSMC's 20 nm system-on-a-chip (20SoC) planar process, promises the right combination of bandwidth, computing performance, and availability.

Finally, let's reach further into the near future. A coming generation of data centers will include not just dense clusters of server-class CPU chips, but also extremely capacious and fast FPGAs. These FPGAs will sit on the super-speed local networks along with the CPUs and shared caches, acting as virtualized, dynamically reconfigurable network packet engines and computational accelerators.

Such chips will require the best available transistor density and metal pitches for capacity, bandwidth within the chip, and—especially in light of the tight thermal and power limitations inside a server rack and the high duty-cycles, which limit the effectiveness of dynamic power management—a power-performance point well beyond what can be achieved with any proposed planar transistor. In addition, in order to connect to super-speed data networks and to support enormous bandwidth to external memory, these FPGAs will require integrated analog circuits at a level of performance beyond anything presently discussed for an FPGA. Such an application lies in the world for which Altera has selected Intel's 14 nm Tri-Gate process.

Conclusion

We have offered three scenarios, in each of which the combination of hard IP use, programmable fabric use, memory bandwidth, and I/O bandwidth is best served by a different semiconductor process. This process is the essence of Altera's tailored approach: to ensure that each category of applications has an FPGA with the performance, headroom, schedule, and cost that its intended system requires. The best fit gives the system developer a measurable advantage.

Further Information

- Innovations at 14 nm and 20 nm—the Next-Generation Advantage: www.altera.com/technology/system-tech/next-gen-technologies.html

Acknowledgements

- Ron Wilson, Editor in Chief, Altera Corporation

Document Revision History

Table 1 shows the revision history for this document.

Table 1. Document Revision History

Date	Version	Changes
February 2013	1.0	Initial release.