WP451 (v1.1) October 15, 2015

UltraScale Architecture Low Power Technology Overview

By: Srinivasa Kolluri

Continued power reduction and unprecedented architectural innovations make Xilinx® UltraScale™ devices the clear choice for many next-generation applications.

The massively interconnected global community continues to demand physically smaller logic devices that are capable of using more bandwidth at greater and greater speeds, while also demanding that these devices perform with higher levels of power efficiency — that is, they must move more data at higher speeds while using less power.

To accomplish this, whole new power-management technologies were conceived and implemented in the form of real-world UltraScale™ devices, capable of moving unprecedented volumes of data — terabits per second — at ultra-low, high-efficiency power levels that were inconceivable just a few years ago.

Designers are now able to meet revolutionary market demands with the enhanced power reduction technology built into the Kintex® UltraScale and Virtex® UltraScale families, based on the industry's first ASIC-class programmable architecture. Designed to scale from 20 nm planar technology through 16 nm FinFET and beyond, UltraScale architecture equips an already-successful architectural platform with numerous innovative power reduction techniques. Co-engineered with the Vivado® Design Suite, the UltraScale architecture enables developers to build smarter, more flexible, more power-efficient systems than what has been possible using current-generation technology and solutions.

This white paper explores the challenges of managing power efficiently, reducing device power requirements, and innovating new power solutions at the speed of Moore's law.

Introduction

The increased proliferation of affordable and powerful portable devices is driving the convergence of wireless and cloud services. This convergence is itself a key driver of traffic growth over the telecommunications networks of the future. *Cloud Computing* and the *Internet of Things* are being widely viewed as the next major evolutionary steps for the Internet and Internet-based services. Many markets and applications require a tremendous increase in system bandwidth and processing capability:

- Wired networking solutions are increasing from multiple links at 100 Gb/s through 400 Gb/s and up to 1 Tb/s
- Digital video applications are ramping from 1080p display through 4K (QuadHD) and up to 8K (Super Hi-Vision)
- Wireless networks are moving from 3G through LTE Advanced to NxN LTE Advanced

Power consumption in programmable devices has become a primary factor for device selection. Whether the concern is absolute power consumption, usable performance, battery life, thermal challenges, or reliability, power consumption is at the center of it all. UltraScale devices address all of these issues with multiple architectural enhancements and innovations, starting with routing, clocking, and logic structures.

Xilinx has continued to study and implement many different power reduction strategies, which span process changes and improvements, architecture changes, voltage-scalable products, and software power-optimization strategies. In planning UltraScale devices, all of these strategies were evaluated based on their impact on static power, dynamic power, and I/O power. Additional risk assessments were made of the new power technologies' impact on time to market for implementation, performance, software, and die area—all of which can be equated to cost.

This white paper describes several aspects of power related to UltraScale devices, including: 20SoC process chosen by Xilinx; benefits on power, its usefulness across all Xilinx product offerings, and the architectural innovations and features for power reduction across the dimensions of static power, dynamic power, and I/O power.

Static Power Reductions

20 nm Process Technology

Xilinx and TSMC have enjoyed a successful partnership and collaborated closely in creating the popular 28HPL process, which combines high performance and low power. The 28HPL process demonstrated significant advantages of the HKMG transistor technology for programmable device applications, and helped create scalable, optimized, architecture-based FPGAs. The 20SoC process employs second-generation gate-last HKMG and third-generation Silicon Germanium (SiGe) strain technology to deliver improved performance at lower power. The same design methodology was applied in choosing the 20SoC process as the successor to 28HPL at the 20 nm node. TSMC's 20 nm process technology can provide 30% higher speed at 1.9X the density, compared to its 28 nm technology.

The benefits of the 20SoC process are multi-pronged, since 20SoC is the highest-density process available today at the 20 nm node; this results in significant reductions in chip area and power. The process is engineered to support high yield by meeting the requirements of advanced manufacturing design, such as adhering to double-patterning layout rules. This is particularly conducive for high-volume markets that require high-performance FPGAs. Architectural and block-level innovations unique to Xilinx add to the power efficiency advantages at every level. In addition, to ensure a smooth production rollout, 20 nm UltraScale FPGAs provide credible power estimation and optimization, block-level power optimization features, and power management features.

Voltage Scaling

The large designed-in voltage headroom in the 20SoC process enables the choice of supplying V_{CC} at a wider range of values, thus allowing a more flexible power/performance strategy.

The power benefits of the 20SoC process include:

- **High-Performance Mode (V_{CC}= 0.95V):** In the range of typical performance targets for FPGAs, 20SoC offers better performance than 28HP/HPL at lower static power.
- **Low-Power Mode (V_{CC}= 0.9V):** 20SoC offers 65% lower static power than 28HP.

The V_{CC} headroom in 20SoC process devices allows Xilinx to select portions of the power distribution curve that perform well even at a reduced V_{CC} of 0.9V. Dynamic power is also reduced by ~10% at this lower voltage. Refer to Figure 1 for a graphical presentation of the effects of these V_{CC}-tolerant design optimizations.

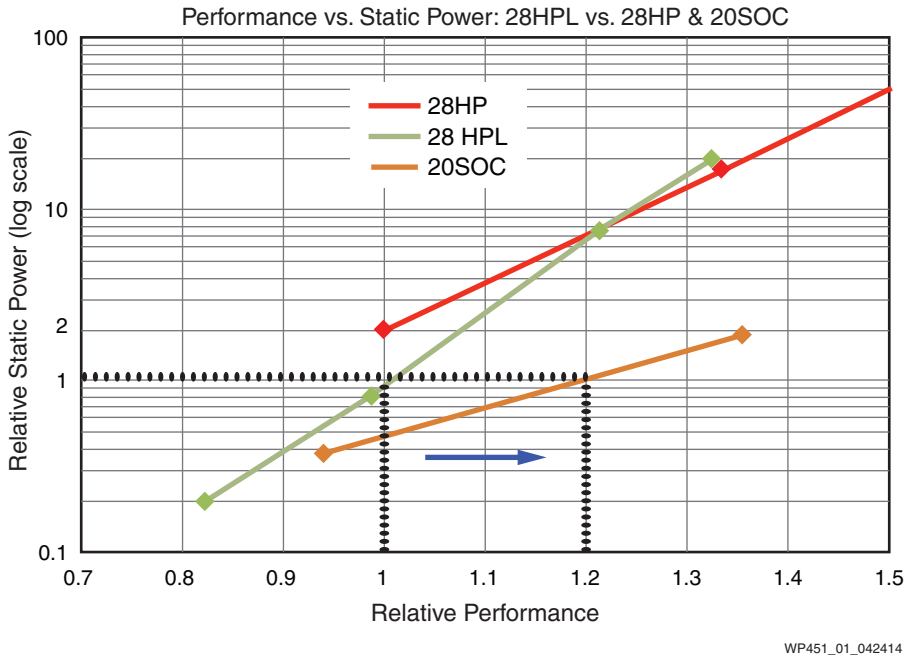

Figure 1: **Performance vs. Power**

UltraScale FPGAs can operate at two core voltages: 0.95V and 0.9V. This is similar to the power strategy first introduced in Virtex-7 devices. FPGAs that can run at 0.95V or 0.9V are designated

-1L, based on their speed grade at 0.95V. Their performance is identical to that of a -1 speed grade at 0.95V and similar to -1 when run at 0.9V, but the "L" signifies that the device is capable of low-voltage operation. At 0.9V, the voltage drop in these FPGAs offers a static power reduction of ~10%.

Normally, this voltage drop would reduce performance, but Xilinx screens these -1L devices for tighter speed and leakage specifications than standard FPGAs. In other words, Xilinx chooses only the lowest leakage and highest performance FPGAs to become -1L devices.

This screening method yields a 35% reduction in power at maximum (worst-case) process compared to the standard speed grade devices. A power distribution system needs to accommodate the maximum power draw to ensure that the power supplies are appropriately sourced for proper functionality of the design under maximum process conditions. Consequently, Xilinx has focused much of its efforts to reduce static power at the maximum process.

Stacked Silicon Interconnect Technology

FPGAs based on the UltraScale architecture will leverage second-generation SSI technology to push the performance and capability envelope to the next level.

As programmable devices get larger, maximum process leakage can become a serious problem because each transistor has a leakage component and some of the larger devices can be upwards of one billion transistors. The larger UltraScale FPGAs are created using Xilinx's second-generation stacked silicon interconnect (SSI) technology. Simply stated, this technology uses multiple Super Logic Regions (SLR) to create a single large device. One benefit of the SSI technology is the reduction in maximum static power compared to a similar size device with a standard monolithic die. See Figure 2.

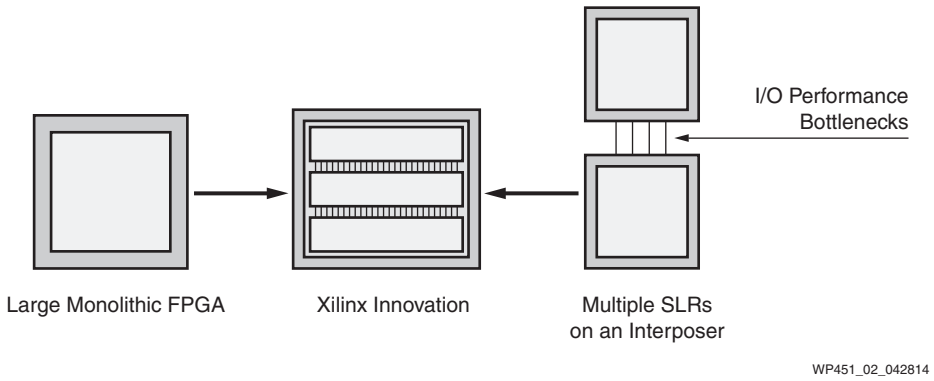

Figure 2: **Second Generation SSI Technology**

Xilinx actively decided never to place *all* maximum process leakage SLRs in a single device. One SLR might be close to maximum process leakage, but the other SLRs in the device will be closer to typical. The result, then, is a much lower maximum process leakage specification when compared to a single die with the same density.

Second-generation SSI technology also provides a significant reduction in I/O interconnect power. Delivering 5.5M System Logic Cells with smaller discrete FPGAs, rather than with Xilinx's SSI

technology devices, would require *thousands of I/O configurations* to connect the separate devices at a functional bandwidth. With SSI technology, the I/O interconnect power is 100X less (bandwidth/W) than an equivalent interface built with I/Os and transceivers. This dramatic reduction is due to all connections being built *on-chip* rather than using the power required to drive the signals *off-chip*. This design concept change enables incredibly high speed at low power.

Dynamic Power Reductions

A tremendous amount of technology and research is required to determine the correct low-power process. However, low power does not stop at the process level. Xilinx focused on power efficiency from every angle in the 20 nm node. Dozens of options were evaluated on the percentage of dynamic power reductions that each could yield, as well as their associated risks and time to implement. As always, each power reduction technique was also evaluated on impact to performance, cost, design flow methodology, and overall schedule. Many options were implemented in Xilinx 20 nm devices—and with this architecture strategy, the low-power features are available across all families built on the UltraScale architecture. See Figure 3.

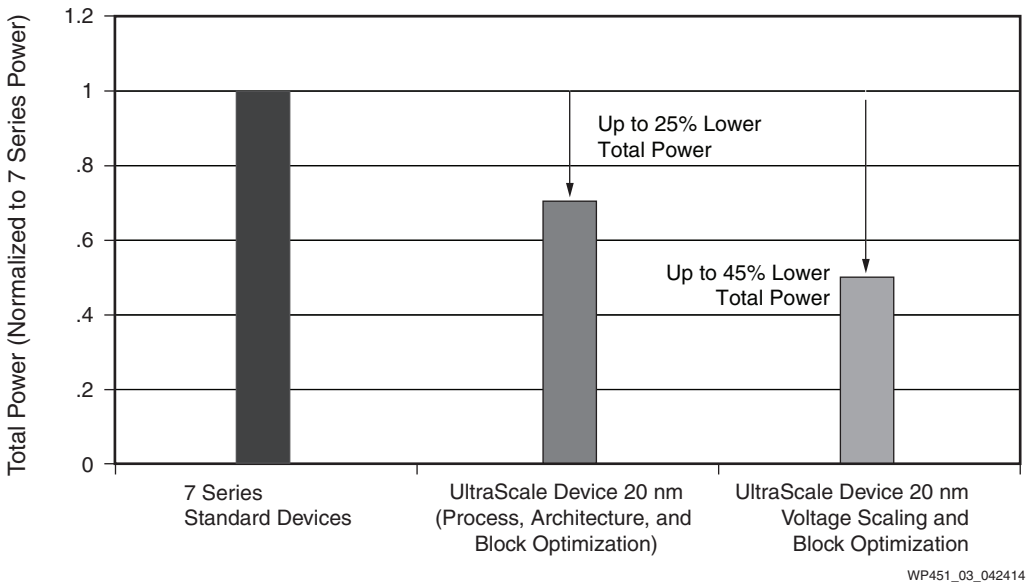

Figure 3: **Overall Power Reduction**

ASIC-Like Clocking Delivers Power Savings

The clock routing and buffers in the UltraScale architecture have been entirely redesigned to provide vastly greater flexibility than existing FPGA architectures. With an abundance of clock routing and clock distribution tracks in both horizontal and vertical directions, the UltraScale architecture also provides hundreds of global-capable clock buffers. The UltraScale architecture has more than 20X the number of global-capable clock buffers than previous architectures, with *thousands* of placement options. In essence, the "center" of the clock network, i.e., the locus from where clock skew starts to accumulate, can be placed in *any* clock region within an UltraScale FPGA. This enables clock networks to run only to where they are needed—the same as an ASIC! The UltraScale architecture provides clock networks with the lowest skew and fastest performance that consume only the power needed to get clock signals from their source to all their destinations.

Clock dynamic power can be further reduced by fine-grain clock gating. The clock drivers are dynamically gated OFF when logic is not in use. This feature can be asserted statically, for sections of circuitry that need to be ON or OFF on a coarse time basis, or dynamically, with a granularity of single clock cycles. In the largest Xilinx 20 nm devices, there are thousands of leaf-gateable clocks for customers to design with, in addition to the globally gateable clocks.

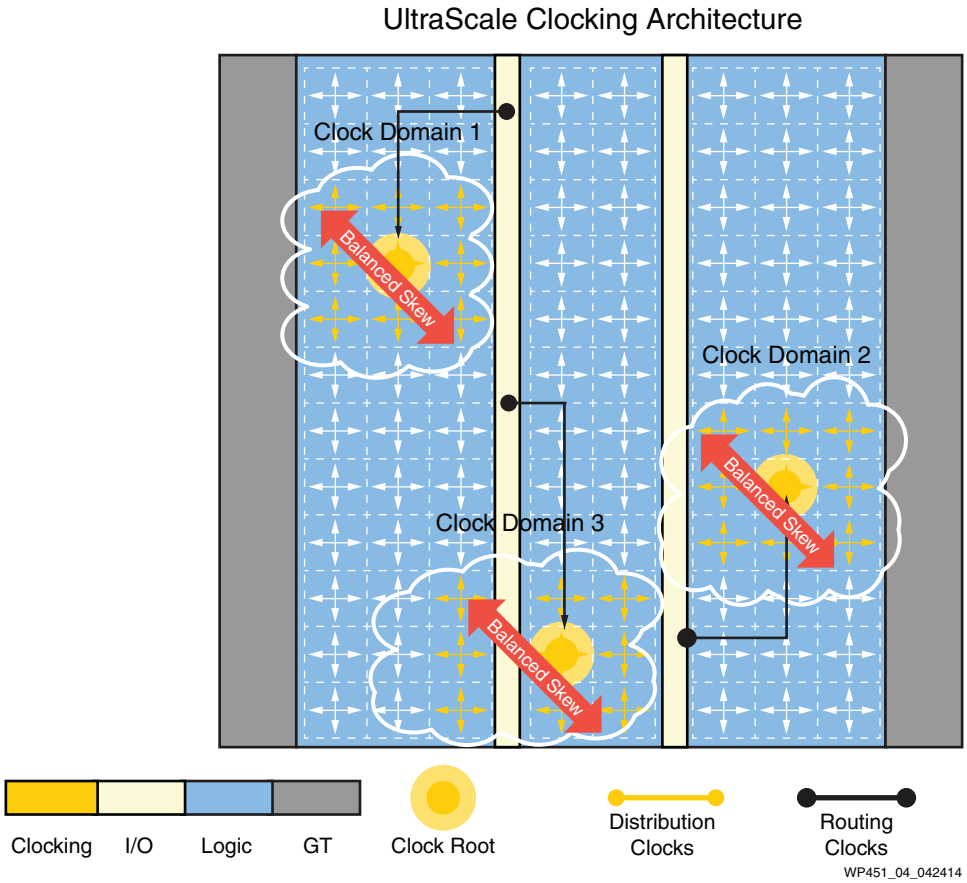

Figure 4: **ASIC-Like Clocking in UltraScale Devices**

Most of the clock tree power (CV^2f) is actually at the horizontal-buffer and leaf-clock levels because this is where thousands of loads (i.e., capacitance) need to be driven. Allowing this level of gating cuts dynamic power considerably. Reducing the fanout drops clock buffer power, because the buffer now sources only a few loads—and, more importantly, cuts power from the clock tree. With the larger quantity of gateable clocks, some designs can save 10–15% in clock tree power, depending on the enable rate.

Designs Use Fewer CLBs Resulting in Lower Power

After the clock and data signals arrive at the logic resources, the UltraScale architecture provides an enhanced Configurable Logic Block (CLB) to make the most efficient use of the available resources, with the goal of reducing total interconnect (i.e., "wire") length. During the UltraScale architecture design phase, every aspect of the existing CLB structure was analyzed to explore how the components could be used more efficiently. Collectively, the resulting enhancements enable the Vivado® Design Suite tools to place many more components (often functionally unrelated to each

other) in a single CLB to achieve a tightly packed design. Operating at high performance, such designs consume the lowest possible power by achieving the best overall device utilization.

Numerous changes within the CLB structure provide added flexibility to the possible packing options. Every 6-input LUT is combined with two flip-flops. Each flip-flop has dedicated inputs and outputs, enabling all the components to be used together or completely independently of one another. The flip-flops benefit from the increased quantity and flexibility of their control signals, with double the quantity of available clock-enable signals, optional “ignore” on the clock-enable and reset ports, optional reset inversion allowing both active-High and active-Low reset flip-flops in the same CLB, and an additional clock signal for shift registers and distributed RAM functions.

Together with the UltraScale architecture's increased quantity of routing resources and a highly flexible clocking architecture, dramatic increases in CLB connectivity enable tightly packed, high-performance designs, driving up device utilization. In conjunction with the additional routing, the denser packing ultimately results in less wire-length, and thus, less wire-capacitance—all of which contributes to lowering the total power. See Figure 5.

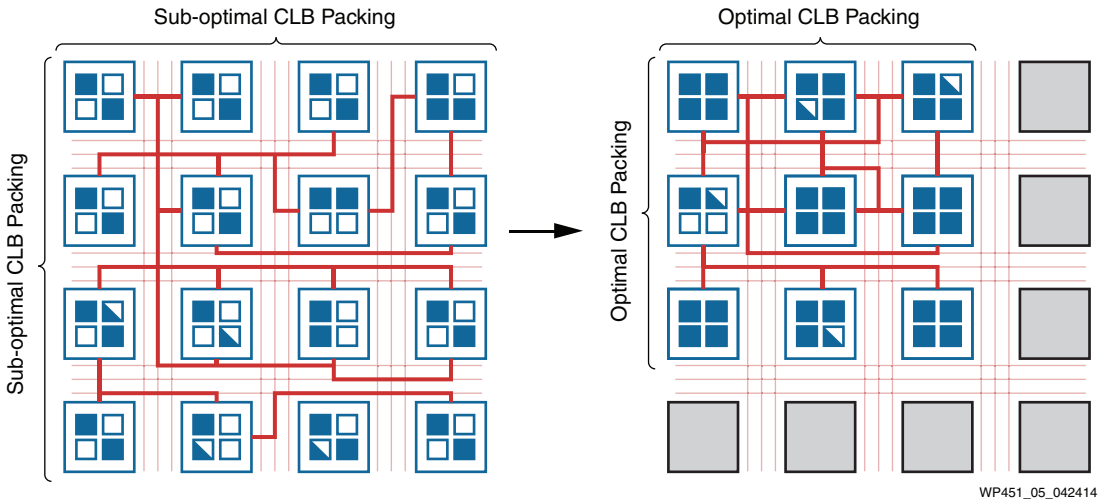

Figure 5: **Optimal Packing Conserves Power in UltraScale Devices**

Block RAM Power Reduction Features

In 28 nm devices, Xilinx supported power gating of unused block RAM; the UltraScale architecture also supports power gating of unused block RAM. Static leakage from block RAM contributes a major portion to overall device leakage, and any power gating of block RAM helps reduce leakage.

In 20 nm devices, block RAM leakage occurs only in blocks that are being used for that design—and *not* in *all* block RAM on the device. The software determines if the embedded memory is instantiated or not. When the design is loaded, power is routed only to the instantiated memories, disabling power to the unused block RAM.

The block RAM also supports a high-speed, memory-hardened cascade feature. Block RAM embedded memory blocks have dedicated data-cascade routing and output multiplexing, which enables faster large block RAM arrays to be built with dramatically lower dynamic power requirements. The cascade feature works by cascading data out from one 36 kb block RAM (RAMB36) into the next block RAM serially to make a deeper memory in a bottom-up fashion. The

data out cascading feature is supported for all RAMB36 port widths. See Figure 6.

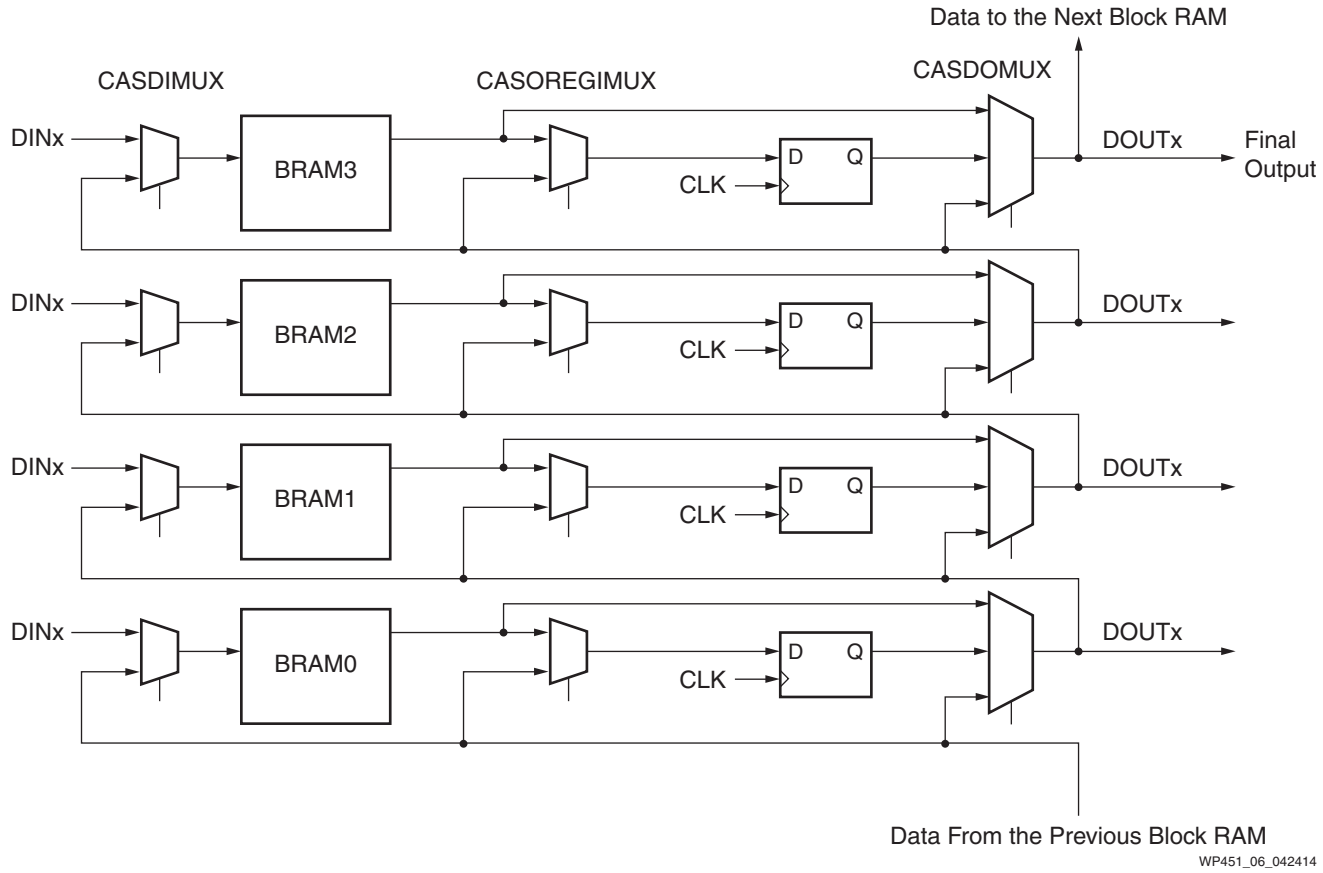

Figure 6: **Cascade Topology in UltraScale Block RAM Architecture**

Multiple block RAMs can be cascaded as required. In cascade mode, all common inputs across all cascaded block RAMs must be connected together. The block RAM provides flexibility to support many different implementations of the cascade feature. The block RAM architecture in UltraScale devices allows for multiplexers that can be used to select datapaths, pipeline registers, or cascade versus direct data input or output. In other words, the cascade feature works by minimizing the number of active block RAMs at any given instant, reducing dynamic power. The integrated cascade does not leave the block RAM column, conserving device interconnect for use in other applications. There is no impact on block RAM timing. See Figure 7.

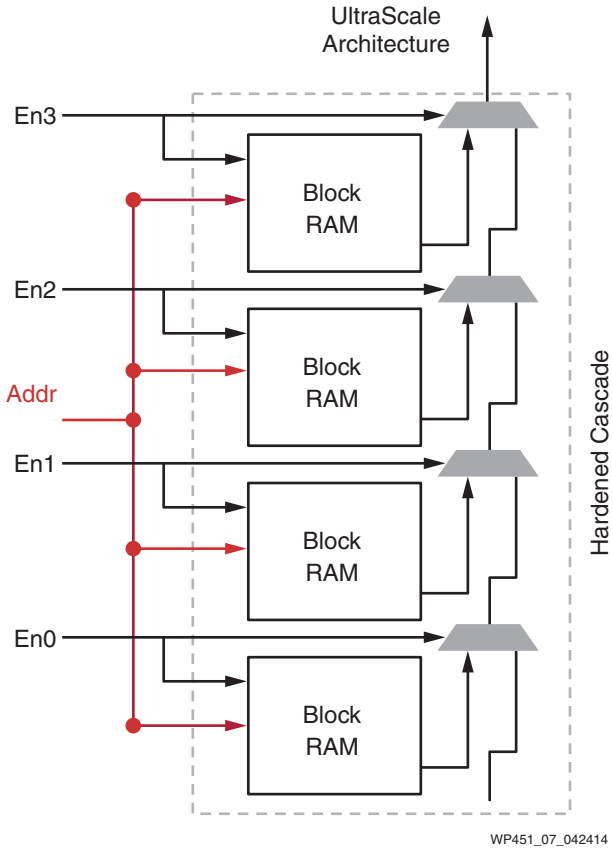

Figure 7: **Dedicated Block RAM Cascade in UltraScale Architecture**

Other block RAM architectural enhancements include:

- Dedicated cascading for FIFO and multiplexing functions
- Dynamic power gating, where block RAM is put into a "sleep" mode while still preserving data

Enhanced DSP Reduces Power Consumption

Xilinx has significantly enhanced the Virtex-7 FPGA's DSP slice, already the industry's performance leader, for the UltraScale architecture. This enhancement permits faster digital signal processing while consuming fewer routing or logic resources outside the DSP block. A number of innovations applied to the DSP slice improve multiplication and MACC operations, enhancing functional performance and reducing power consumption.

The UltraScale architecture DSP slice incorporates 27x18-bit multipliers that permit the mapping of larger functions into fewer DSP slices. For example, the UltraScale architecture DSP block, with its wider 27x18-bit multipliers, can implement IEEE Std 754 double-precision arithmetic, using two-thirds fewer DSP blocks compared with the same function implemented with the DSP blocks in Xilinx 7 series devices.

The DSP block in UltraScale FPGAs also has significant power reduction features implemented with the addition of Wide XOR and Wide MUX functions. These functions allow performing a 96-bit XOR, making the implementation of wired designs and wide multiplexing functions much more

efficient. Functions like Complex Multiply Accumulate can be implemented in half the number of DSP slices required by ordinary 7 series devices, running faster and more efficiently than equivalent implementations in logic, delivering significant power savings.

These enhancements increase performance, lower power consumption, and reduce the usage of CLBs, leaving more CLBs free for the implementation of other functions. See Figure 8.

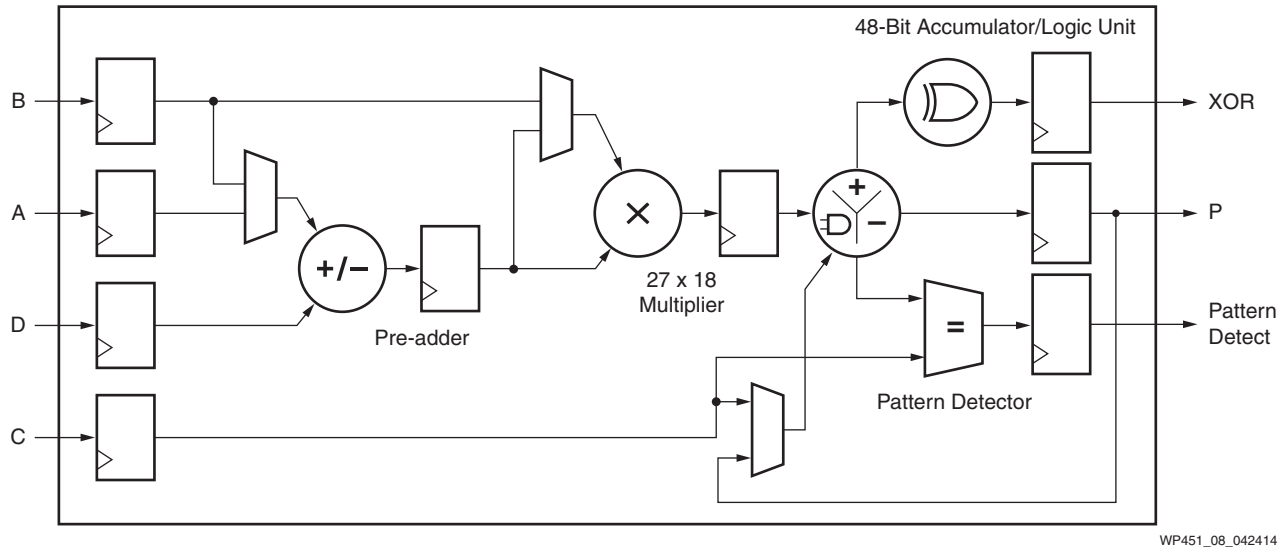

Figure 8: **Enhanced DSP in UltraScale Architecture**

Another major DSP block power-reduction enhancement is the ability to pack more multipliers into a single slice or tile, which results in a smaller footprint and reduced power. It is through the DSP block and similar innovations that the UltraScale architecture is able to simultaneously meet the application demands of the next generation: increased processing capabilities and lower power utilization.

I/O Power Reductions

I/O power has become a much more significant contributor to the total power requirement of a device. As programmable devices have evolved, core power has been greatly reduced. But until recently (with the advent of the Xilinx 7 series families), I/O power had not. Especially in memory-intensive applications, massive I/O requirements can consume up to 50% of a design's total power budget. Xilinx aggressively reduced I/O power in the 7 series FPGAs. UltraScale devices enable all of these power-saving features.

In addition to programmable slew rates and drive strength, special standards like HSLVDCI save considerable power from device to device and in lower-speed memory interfaces. See WP389, *Lowering Power at 28 nm with Xilinx 7 Series Devices*, for more details.

DDR4 Solution Delivers Massive Bandwidth with Lower Power

The UltraScale architecture takes memory interfacing to a new level by enabling multiple DDR3/4-capable SDRAM memory controllers and including integrated DDR physical-layer (PHY)

blocks on chip. See Figure 9.

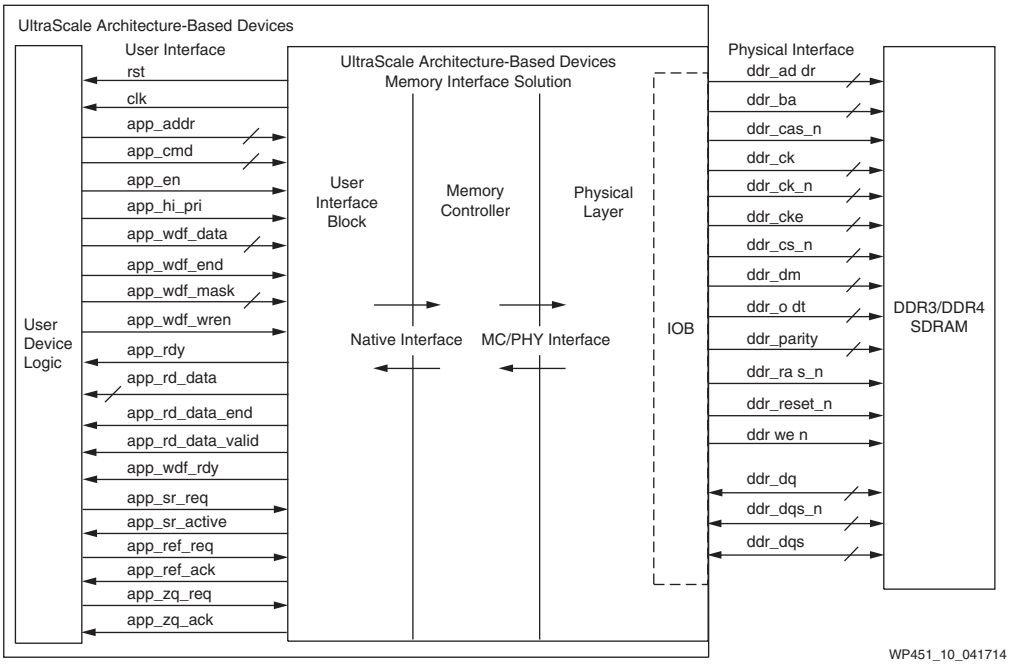

Figure 9: **DDR4 Controller in UltraScale Architecture**

The new DDR4 memory interface in UltraScale devices provides more than 1 Tb/s of memory bandwidth to handle the massive data flow, fast processing, and enormous memory requirements of leading-edge, next-generation system designs in key applications, such as video imaging and processing, traffic management, and high-performance computing. Applications see a reduction in read latency by 30% and significant power savings at the same data rate by going from a DDR3 to DDR4 interface.

DDR power has been as important as performance in many user designs. Customers will also see a 20% reduction in power when moving from DDR3 to DDR4, because DDR4 operates at a lower voltage of 1.2V. Further power savings are possible when leveraging I/O, DDR4, plus integrated PHY.

DDR4's new memory interface employs pseudo-open-drain (POD) termination, meaning that memory cells can store a logical 1 (High) *without consuming power*. POD relies on switchable, on-die termination instead of a separate pull-up resistor. This means that the DDR4 module consumes power only when the V_{DD} rail is pulled down to a logical 0 (Low).

Another new power-saving feature is data bus inversion, in which the logic decides whether the actual bus state *or its complement* involves the least switching and, therefore, the least overall power consumption, on each cycle.

Additionally, with higher-density parts and low-power features like POD I/O and data bus inversion, DDR4 is attractive because of its higher performance at lower power compared to DDR3 or DDR3L.

Transceiver Power

The transceivers in Xilinx 20 nm devices have been optimized for high performance and low jitter, offering several low-power operating features. To facilitate balancing power and performance trade-offs, each transceiver offers power features that enable the user to customize operational flexibility and granularity. The UltraScale architecture-based GTH transceiver has been redesigned to cut the total power requirement by 50% compared to the GTX and GTH transceivers in the 7 series FPGAs.

The transceivers in UltraScale devices provide a *low-power mode* of operation. Many non-backplane applications do not need decision feedback equalizer (DFE) circuitry, which burns extra power, so Xilinx gives designers a choice when servicing other applications. To save power, the designer can turn off the DFE circuitry and use the linear equalizer (LE) by itself. The LE uses much less power than the DFE because of its lower Rx gain and minimal circuitry.

Integrated blocks reduce static power by minimizing transistor count, but they can also have a big impact on dynamic power. Integrated blocks eliminate programmable interconnects and reduce trace lengths and logic levels, thereby shrinking footprint area and dynamic power. In all, replacing soft IP with an integrated block can result in up to a 10X reduction in power consumption. See Figure 10.

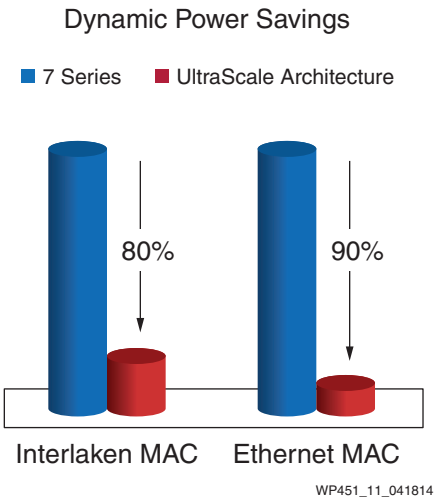

Figure 10: **Integrated MAC in UltraScale Architecture**

Xilinx implements an integrated Interlaken IP core that scales up to 150 Gb/s of chip-to-chip connectivity. The Xilinx IP core is based on the industry's leading and most widely deployed implementation. It is a flexible, high-performance, low-power implementation of the Interlaken interface protocol specification protocol rev 1.2, and supports 12.5 Gb/s and 25 Gb/s transceivers. Combined with the transceiver technology in the UltraScale architecture and a flexible protocol layer, the integrated IP core minimizes the pin and power overhead of chip-to-chip interconnect.

Integrating the Ethernet MAC and Interlaken IP in Virtex UltraScale FPGAs has several benefits to the system and to the user. Integrated IP has lower latency than an equivalent soft IP solution, enabling performance that has not previously been possible. The FPGA resources that would previously have been used to implement the soft MAC and Interlaken interface blocks are now

available to the designer to perform pre-packet processing, time stamping, and other functions at a lower power profile. Also, integrating multi-chip functionality into one UltraScale FPGA results in significant power reduction.

Power Management

For many generations, Xilinx has offered on-chip and off-chip monitoring via the System Monitor (SYSMON) and the XADC feature to support intelligent system and power management.

In the UltraScale architecture, Xilinx has enhanced SYSMON to better meet user power and system management requirements. The number of on-chip supply sensors has been expanded from three to seven with the addition of four new user-defined on-chip supply sensors. These sensors can be configured, for example, to access any of the V_{CCO} supply voltage present on the chip. In addition, any I/O bank can be used to support the sixteen auxiliary external analog inputs to SYSMON.

A new integrated I2C interface enables the power or system manager to access SYSMON data (both pre- and post-configuration) via an industry-standard management interface. This interface provides easy access to the vital real-time operating conditions of the FPGA, greatly enhancing the reliability, safety, and security of the system.

Vivado Design Suite and Power Optimization

First introduced for the Xilinx 7 series families, the Vivado® Design Suite is an SoC-strength design environment built from the ground up for the next decade of Xilinx All Programmable devices, including the UltraScale architecture. The Vivado Design Suite attacks the key design bottlenecks that arise in programmable systems integration and implementation, providing up to a four-fold productivity advantage over competing development environments. Vivado tools extend the silicon power savings with software-enabled power optimizations.

The Vivado Design Suite uses a multi-variable cost function to find the optimal placement, allowing the designer to quickly find a routable solution — even at device utilizations greater than 90%—with no performance degradation. Optimal placement also results in reduced interconnect power.

The Vivado Design Suite also supports many of the UltraScale architectural power reduction features. Vivado Design Suite power gates portions of the design by generating logic that drives leaf-clock buffer enables on the logic. The tool also automatically generates logic to support both static and dynamic power gating for block RAMs in the designer's code. In addition, the Vivado Design Suite can infer cascaded block RAMs to exploit the UltraScale architecture's efficiency.

Intelligent clock-gating optimizations are also performed automatically on the entire design by the Vivado Design Suite. These optimizations generate no changes to the existing logic or clocks that might alter the behavior of the design. The clock gating capability uses a set of innovative algorithms that perform an analysis on all portions of the design, including legacy and third-party IP blocks. The output logic of the sourcing registers is analyzed at each clock cycle; those that do not contribute to the logical result can be gated off by one of the plentiful clock enables available in the Xilinx 20 nm logic. Fine-grained clock gating or logic gating signals are thereby created that can be used to neutralize superfluous switching activity.

Conclusion

Xilinx targeted and achieved a significant reduction in power requirements for the UltraScale architecture. By applying a holistic approach to reducing FPGA and system power, Xilinx was able to achieve significant reductions in power consumption across multiple functional blocks, such as DSP, serial transceivers, block RAM, I/O, and perfect integration with the Vivado Design Suite and its robust development tools. The Xilinx 20 nm devices offer up to 45% total power reduction—and an even greater power reduction at maximum process when comparing the 20 nm devices' maximum process to the previous generation's equivalent. Power can be further reduced by leveraging integrated cascade functions, integrated high-bandwidth Ethernet MAC and Interlaken IP, memory interfaces, and advanced clock and logic gating software. Table 1 shows the UltraScale architecture power reduction strategies and the specific Xilinx innovations that made the achievement of these strategic goals a reality.

Table 1: **Summary of UltraScale Architecture Power Reduction Innovations**

Power Reduction Strategies	Architecture Innovations
Static Power	
Overall static power reduction	• 20 nm SoC process node
Binning	• C-grade vs. I-grade vs. -1IL-grade • SSI technology with static power binned slices
Turning off unused resources	• I/O and block RAM
Voltage scaling	• -1IL (0.95V, 0.9V)
Dynamic Power	
Intelligent clock gating	• Software-based
Hardware-based clock power saving	• Segmented clock network • Fine-grain clock gating
Block RAM power	• Integrated data cascade • Integrated address cascade • Dynamic power gating to save static power • 60% power reduction vs. 7 series
DSP power	• Better utilization enabled, reducing area and power • Wide XOR • 20% power reduction vs. 7 series
I/O and Memory	
I/O and memory interface power reduction	• DDR4 1.2V power savings • Pseudo-open drain • Turn off termination and receiver power during IDLE • DCI/low-power IBUF • 20% power reduction vs. 7 series (DDR3)
Serial Transceivers	
Low-power modes	• LPM mode
General transceiver power reduction	• Re-architected transceivers; power reduction vs. 7 series
Ethernet MAC/Interlaken MAC	• Integrated MACs save significant power

Additional Resources

1. WP436, *Leveraging Power Leadership at 28 nm with Xilinx 7 Series FPGAs*
2. WP389, *Lowering Power at 28 nm with Xilinx 7 Series Devices*
3. UltraScale Advantage Demo Video from Xilinx Youtube Channel
4. UG997, Vivado Power Analysis and Optimization Guide
5. Kintex UltraScale Architecture Power Estimator

Revision History

The following table shows the revision history for this document:

Date	Version	Description of Revisions
10/15/2015	1.1	Updated Voltage Scaling and Stacked Silicon Interconnect Technology.
07/09/2015	1.0.1	Minor typographical updates.
05/01/2014	1.0	Initial Xilinx release.

Disclaimer

Automotive Applications Disclaimer