WP478 (v1.0) April 15, 2016

Breakthrough UltraScale+ Device Performance with SmartConnect Technology

Vivado Design Suite 2016.1 extends SmartConnect technology to solve the system interconnect bottleneck for high performance, multi-million system logic cell designs, without requiring any code rewrite or extra latency insertion.

ABSTRACT

System performance generally is not limited by the speed of local data processing, but rather by the choice of the interconnect network moving data between the processing blocks and the system interfaces.

Available in Vivado® Design Suite 2016.1, AXI SmartConnect IP is designed for low latency and high system throughput. Also in this release, Xilinx extends SmartConnect technology with optimization techniques, including useful skew optimization, time borrowing, automated retiming, and pipeline analysis, to identify and mitigate system performance bottlenecks without requiring heavy manual optimizations, and costly architecture changes.

Introduction

The Xilinx® UltraScale+™ portfolio, shipping since 2015, is the only 16nm FinFET programmable technology in the industry. Comprised of Zynq®, Kintex® and Virtex® UltraScale+ devices, the UltraScale+ portfolio delivers a 2–5X performance-per-watt improvement over 28nm offerings, enabling market-leading applications such as 5G wireless, Software-Defined Networks, and next-generation Advanced Driver Assistance Systems.

In the 2016.1 release, the Vivado Design Suite HLx Editions deliver AXI SmartConnect IP. The 2016.1 release also extends SmartConnect technology to solve the system interconnect bottleneck for high performance, multi-million system logic cells designs, delivering up to 2X higher performance than 28nm technology devices, without requiring redesign or extra latency insertion. In contrast, other solutions require heavy manual optimizations, and costly architecture choices to meet timing requirements for isolated IP designs.

When designing a complete system-on-a-chip on programmable devices, the system performance is generally not limited by the speed of local data processing, but rather by the choice of the interconnect network moving data between the processing blocks and the system interfaces, as well as by wire delays.

Trade-offs and optimizations can be made to reduce the overall system interconnect cost based on the characteristics of the data movement in the system. The UltraScale+ portfolio was co-optimized with the Vivado Design Suite, using SmartConnect technology, to provide designers with maximum performance per watt. SmartConnect technology includes a system interconnect IP designed for low latency and high system throughput, and optimization techniques (described in this white paper), enabled by architecture innovations in the UltraScale+ portfolio, to solve wire delay bottlenecks. These optimizations include useful skew optimization, time borrowing, automated retiming and pipeline analysis to identify the system bottlenecks.

High System Throughput with AXI SmartConnect IP

The architecture of a system interconnect is a critical consideration for high-performance designs. Typical interconnect networks include high-performance crossbars (coupled with data-width converters using FIFOs, protocol converters, clock-domain-crossing circuitry, and arbitration), which can have very high area utilization. Alternatively, a soft Network-on-a-Chip (NoC) can typically deliver lower area and latency at a higher clock frequency, resulting in higher system throughput.

The AXI SmartConnect IP, is the third generation of Xilinx's AXI Interconnect, based on the ARM® AMBA® AXI4 Interface protocol. The new IP resides in the Vivado IP Catalog, delivering the maximum system throughput at low latency by synthesizing a low area custom interconnect that is optimized for important interfaces.

SmartConnect technology boosts performance per watt of AXI interconnect by optimizing interconnect networks for performance and area, within the specific interconnectivity requirements inherent to the overall design. Benefiting the most from the new AXI SmartConnect IP are systems comprised of multiple IP, DMAs, and system interfaces, including high bandwidth interfaces, e.g., DDR4, connected by AXI interconnect. The example in Figure 1 shows a system described in the

Vivado IP Integrator, including a PCIe® DMA subsystem driving DDR4 and flash/SRAM interfaces through the AXI SmartConnect IP.

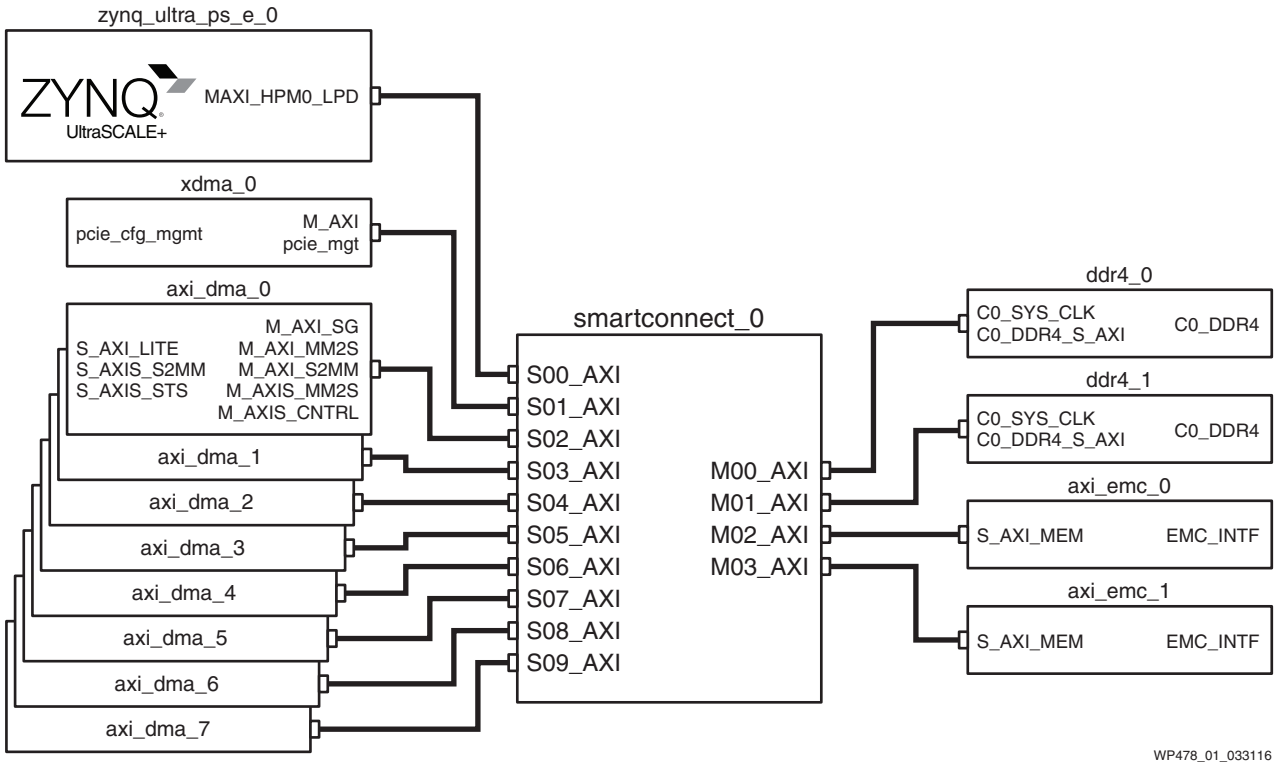


Figure 1: **System Leveraging AXI SmartConnect IP**

SmartConnect technology also introduces a number of ease-of-use automation features around clocking and system reset topologies. Arbitration has also been optimized to provide the highest throughput in systems with multiple master IP—such as DMA and processor subsystems—addressing high bandwidth interfaces (e.g.,DDR4). SmartConnect scales to very large systems and can be optimally pipelined to increase the clock frequency and further reduce the data transport bit-width.

Breakthrough Clock Frequencies Enabled by Useful Skew and Time Borrowing Optimizations

When designing custom hardware on a large device, clock frequencies are often limited by the excessive clock skew through the clock network, causing the data to be clocked too early or too late. The UltraScale+ portfolio provides an ASIC-like clock network that minimizes clock skew. Additionally, realizing that clock skew can also be beneficial if controlled, Xilinx has added a new leaf-clock delay feature that enables fine-grained control over clock delays throughout the clock network. New optimizations in the Vivado Design Suite take advantage of this feature to add useful skew as a means to compensate for the wire delays on the interconnect logic routing network, thus significantly increasing the operating clock frequencies. Inserting a delay element on the clock route to registers that capture the output of longer combinatorial paths allows more time for data to propagate through those longer combinatorial paths before being captured by the register.

The useful skew technique is illustrated in Figure 2. By employing a 0.5ns leaf-clock delay on the clock route feeding register B, there is a full 2.5ns between the first rising edge at register A and the next rising edge at register B, allowing data to fully propagate through the long combinatorial logic cone and be captured correctly. The leaf-clock delay introduces useful skew, thereby reducing the edge-to-edge delay from registers B to C, where the combinatorial logic cone requires only 1.5ns to fully propagate.

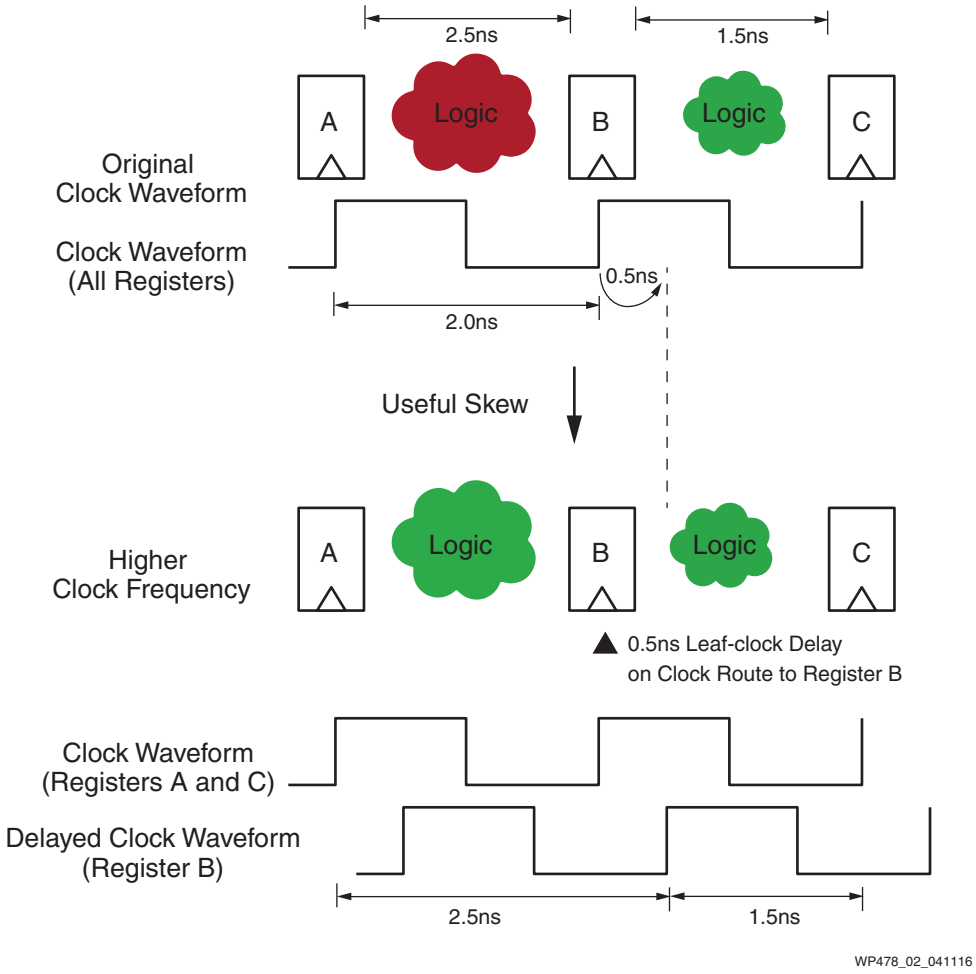


Figure 2: **Useful Skew Optimization–Adjusting Clock Waveforms to Maximize Frequencies**

Having fine-grained control over time in the clock network with leaf-clock delays is a powerful and inexpensive way to effectively reduce wire delay. A poor alternative to virtually retiming clock edges is to add millions of feature-reduced registers in the logic interconnect routing network, to physically retime, replicate, or even pipeline simple wires. While the retiming approach is genuinely useful for the longest critical path, its use along with the addition of millions of feature-reduced registers to reduce wire delay alone is expensive and ineffective—degrading area, power, and system latency—when compared to useful skew techniques.

Figure 3 shows a snapshot of the UltraScale+ device leaf-clock architecture.

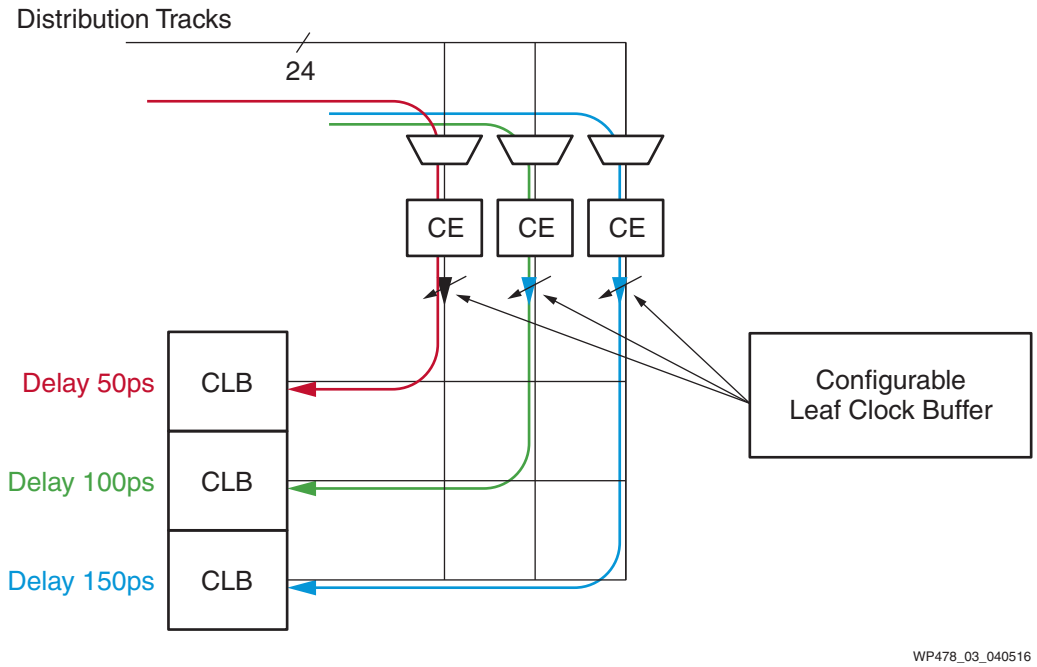


Figure 3: **UltraScale+ Device Leaf-Clock Buffer Delay Feature (Enabling Useful Skew Optimization)**

The UltraScale+ device clock network includes programmable leaf-clock delays. The leaf-clock buffers have five discrete delay-tap settings that allow the router to automatically optimize the leaf-clock delay setting for fixing setup and hold violations without the designer's intervention. The Vivado Design Suite determines the exact tap setting, which helps achieve timing closure. This architectural feature provides significant benefits with no effort from the designer.

Time borrowing is the second optimization technique enabled by UltraScale+ devices that can be employed to automatically meet performance requirements. Time borrowing in level-sensitive latches requires complicated analysis by the timing engine. The Vivado Design Suite performs this analysis out-of-the box without any designer intervention.

The device architecture also allows flip-flops in the configurable logic block (CLB) to be configured as a pulse latch by the Vivado Design Suite. Dedicated circuitry in the leaf-clock buffers allows the generation of a programmable clock pulse. This gives the Vivado tools the flexibility to substantially improve performance. The block diagram of the UltraScale+ device architecture with the programmable pulse generator and the configurable latch is shown in Figure 4.

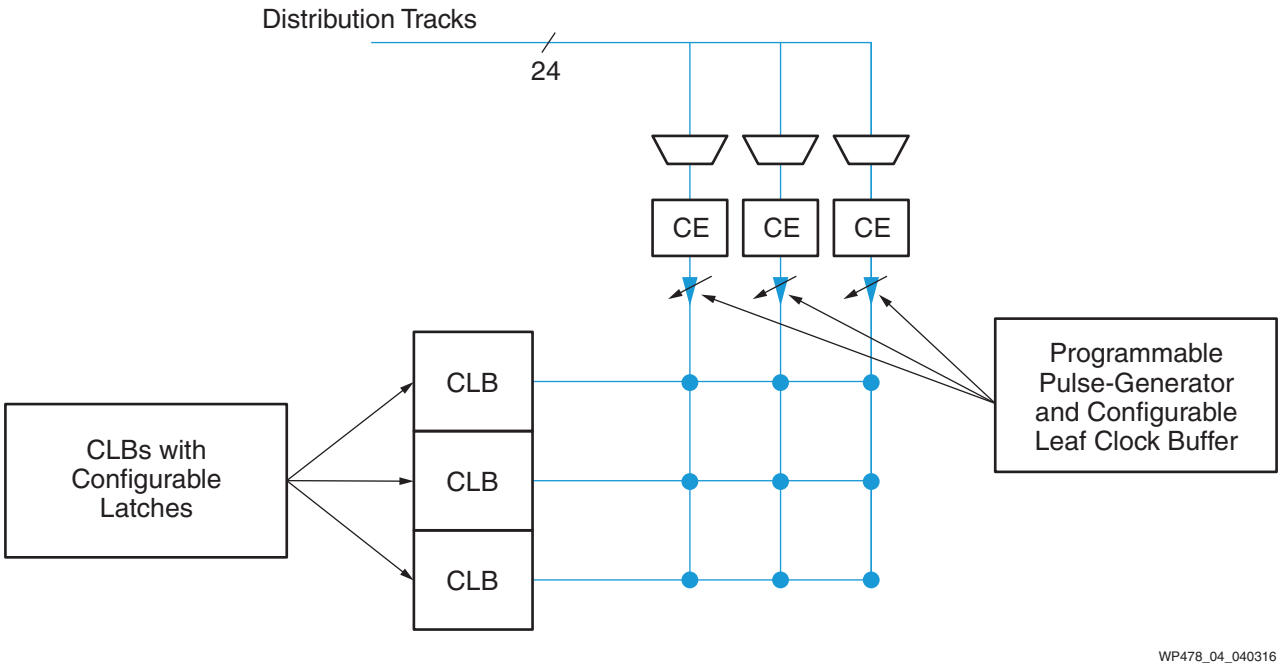


Figure 4: **UltraScale Architecture Support for a Pulsed Latch**

Achieving the Highest Possible Frequencies with Pipeline Analysis and Retiming

As performance requirements increase, architectural-level trade-offs have a far greater impact than tool options or simple design changes. One such trade-off involves sacrificing latency to increase clock frequency by inserting pipeline register stages to cut the longest critical paths into smaller, faster operating segments.

The Vivado Design Suite's pipeline analysis feature (`report_pipeline_analysis`) provides a unique insight into design bottlenecks and opportunities to improve the design's F_{MAX} by adding pipeline registers. Because pipelining changes the sequential behavior of the design and requires extra attention for verification, the focus is on providing accurate guidance rather than inserting pipeline stages automatically.

There are three steps to take advantage of this capability.

1. The `report_pipeline_analysis` feature analyzes the design to give a summary. Internally, the design is broken down into its feedback and feed-forward portions and pipeline analysis is performed only on the feed-forward sections.

2. Once broken down, a copy of the design is made for exploration, and then latency stages are added to the exploration model in feed-forward portions with a new, faster, critical path emerging after each iteration.

   Eventually, there are no further performance gains because either the new critical path is sufficiently fast and inserting pipeline registers cannot increase the F_{MAX}, or the new critical paths are part of a feedback loop. Pipeline analysis also considers the maximum possible

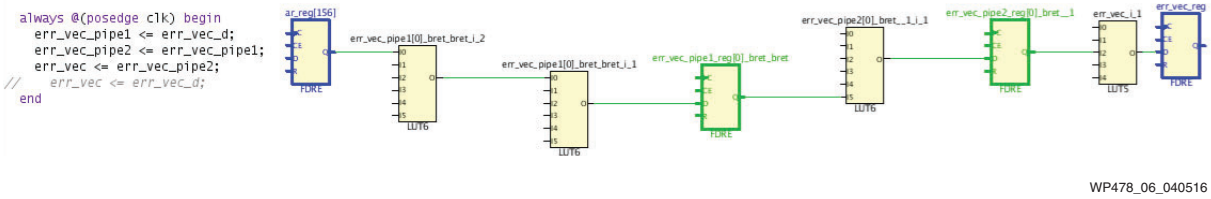
operating clock frequency, and will not recommend insertion that would result in an F_{MAX} that far exceeds the device capabilities. In the example report in Figure 5, the F_{MAX} can be improved from 295MHz to 710MHz by adding two pipeline stages, representing 107 total registers, to the specific feed-forward path in the design where the endpoints are listed. In this example, the feedback loop is significantly faster than the feed-forward paths and does not limit the performance gains achieved by pipelining. The pipeline analysis stopped after two stages of added latency. Inserting a third stage likely pushed the estimated F_{MAX} beyond the limits of both the slowest loop and the maximum F_{MAX} capability of the device.

```
2. Maximum improvements by stage insertion
-------------------------------------------

Intra-Clock Summary
+-------+---------------+-----------------+-------------------+-----------------+-----------+---------------+----------------+-----------------------------+-----------------------------+
| Clock | Added Latency | Ideal Fmax (MHz) | Ideal Delay (ns) | Requirement (ns) | WNS (ns)* | Added Pipe Reg | Total Pipe Reg | Pipeline Insertion Startpoint | Pipeline Insertion Endpoint |
+-------+---------------+-----------------+-------------------+-----------------+-----------+---------------+----------------+-----------------------------+-----------------------------+
| clk   |       0       |      295.48     |       3.384       |       1.550     |   -1.834  |      n/a      |        0       | err_vec_i_1/O               | err_vec_reg/D               |
| clk   |       1       |      548.14     |       1.824       |       1.550     |   -0.274  |       50      |       50       | tc_i_2/O                    | ctr[7]_i_1/I1               |
| clk   |       2       |      710.05     |       1.408       |       1.550     |    0.142  |       57      |      107       | ctr_reg[4]/Q                | ctr[5]_i_1/I0               |
+-------+---------------+-----------------+-------------------+-----------------+-----------+---------------+----------------+-----------------------------+-----------------------------+

3. Critical Loop Paths
----------------------

Summary
+------------+----------+-------------------+-----------------+-----------+-----------------------------+-----------------------------+
| From Clock | To Clock | Ideal Delay (ns) | Requirement (ns) | WNS (ns)* | Critical Path Startpoint | Critical Path Endpoint |
+------------+----------+-------------------+-----------------+-----------+-----------------------------+-----------------------------+
| clk        | clk      |       1.020       |       1.550     |    0.530  | ctr_reg[6]/Q                | ctr_reg[6]/D                |
+------------+----------+-------------------+-----------------+-----------+-----------------------------+-----------------------------+
```

WP478_05_040516

Figure 5: **report_pipeline_analysis Report**

3. The designer modifies the HDL code based on the report's recommendation, by adding the two pipeline register stages at the endpoints of the path. Synthesis is rerun, with the new retiming optimization enabled. These new register stages are automatically retimed into the feed-forward logic cone to balance the critical paths. Figure 6 shows these added pipeline registers, highlighted in green, retimed into the critical path between the two endpoints, highlighted in blue.



WP478_06_040516

Figure 6: **Registers Added in HDL at Path Endpoints, Retimed to Optimally Pipeline the Path**

The overall impact to placement and routing is minimal because the pipeline registers are effectively inserted between LUTs. As shown in Figure 7, each LUT is architecturally paired with two registers allowing the LUT outputs to be routed directly to a register at no cost. Additionally, the delay incurred by the wire segments driven by this register can be mitigated with time borrowing or useful skew techniques. Alternative approaches that add millions of registers to the interconnect have no measurable benefit over these techniques, but come at a cost to area, complexity, and power.

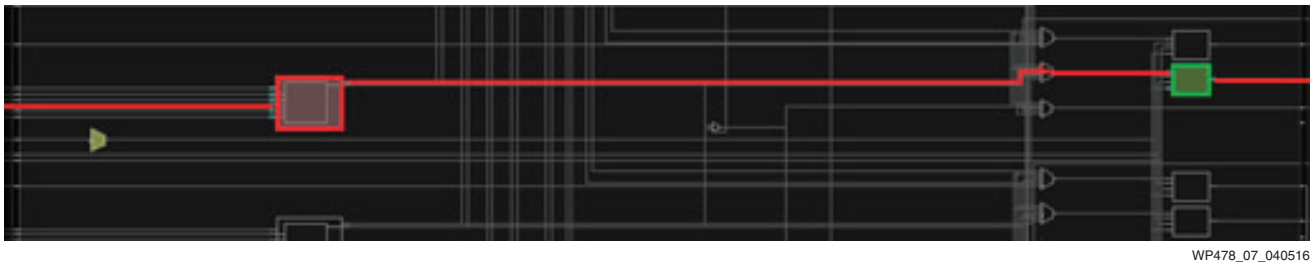


Figure 7: **Dedicated LUT-Register Route for Maximum Performance**

Inserting pipeline stages in sequential feedback loops is complicated, because it alters the design's functionality due to the dependencies on data from previous cycles. There are very limited possibilities for successfully pipelining feedback loops, and those are often accompanied with substantial area increases and reduced system throughput. Small loops can sometimes be transformed manually, but often large loops cannot be pipelined. Pipeline analysis identifies loops and their sizes to help the designer evaluate the practicality of transforming them. See Figure 8.

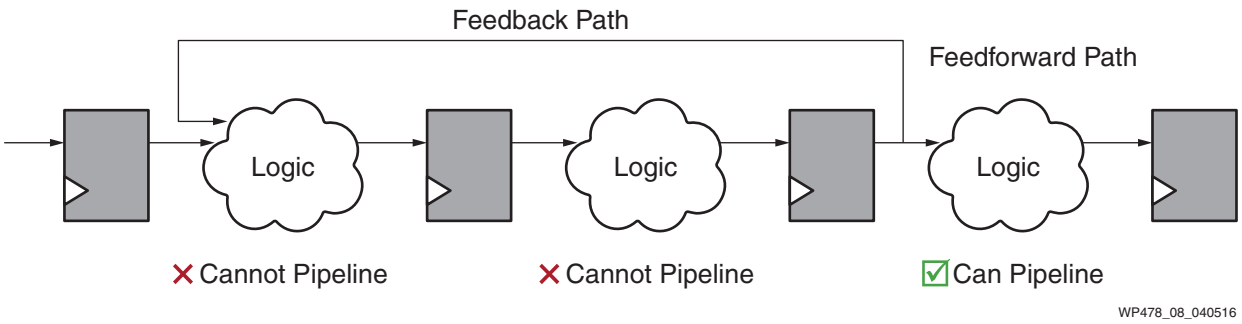


Figure 8: **Feedback Loop Cannot be Pipelined Without Changing Functionality and Degrading Throughput**

Practical Example: Complex Wireless Radio Design

The example shown in Figure 9 is a complex wireless radio design, well pipelined by design. It does not contain any critical feedback loops, uses 83% of the System Logic Cells in a Virtex UltraScale+ device (an XCVU9P—2.6M System Logic Cell device). Additionally, this design utilizes 76% of the DSP blocks and 53% of the block RAM.

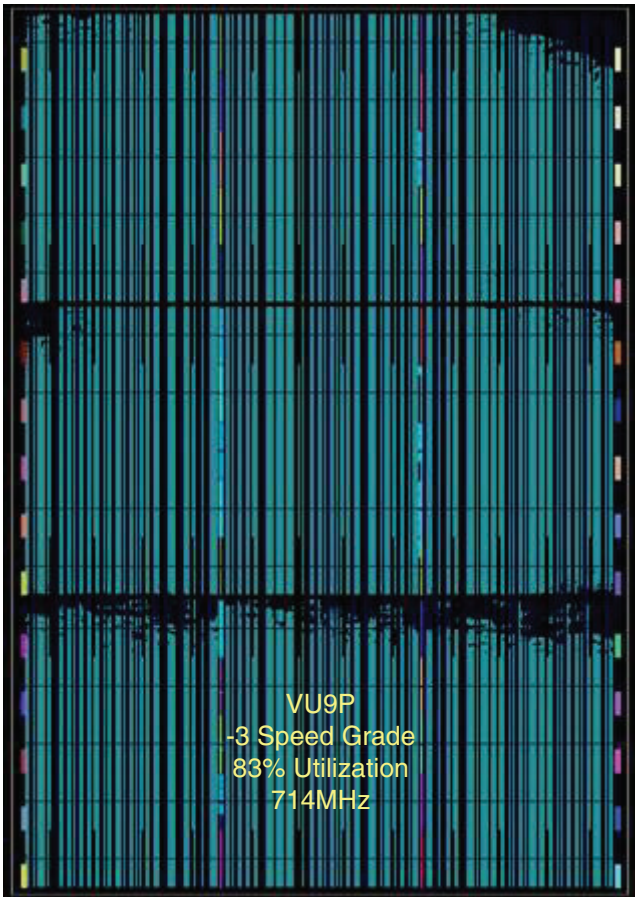


Figure 9: **Large Wireless Radio Design Implemented on an XCVU9P**

Using this skew technique alone, the Vivado Design Suite 2016.1 is able to close timing at an unprecedented 714MHz, as shown in Figure 10. The example shows how high-speed performance can be achieved on entire, fully-utilized programmable devices, and not just on a small example IP out of the context of a user design. Although the datapath delay is 1.5ns (or 667MHz), the requirement time of 1.4ns (or 714MHz) was met with a positive slack of 26ps, due to the automated useful skew optimization during implementation.

Summary

Name	Path 3
Slack	0.026ns
Source	fir_filte
Destination	fir_filte
Path Group	clk1
Path Type	Setup
Requirement	1.400ns
Data Path Delay	1.503ns

WP478_10_040516

Figure 10: **Timing Report for the Example Radio Design—Timing Closed at 1.4ns Clock Period (714MHz)**

To compare this performance to 28nm devices, the size of this design was reduced by two-thirds to fit it for lower utilization on 28nm devices. With the fastest speed grade, the Virtex®-7 FPGA and the Vivado Design Suite can achieve an F_{MAX} of 375MHz. Therefore, on this wireless radio design, the UltraScale+ device delivers 1.9X (714MHz/375MHz) higher performance out of the box, over the 28nm device.

In this well-pipelined filter design, long wire delays were mitigated with useful skew; no performance improvements can be achieved with more aggressive pipelining or retiming. However, additional improvements can be achieved when integrating such designs in a larger system, by using the AXI SmartConnect IP.

Summary

SmartConnect technology comprises AXI SmartConnect IP and SmartConnect optimizations, including useful skew, time borrowing, retiming, and pipelining recommendations. SmartConnect technology enables unprecedented high clock speed and high system throughput for high utilization designs.

The new AXI SmartConnect IP creates a custom interconnect architecture, maximizing system throughput for the specific designs.

Using the new UltraScale+ device clock network features, new SmartConnect optimizations in the Vivado Design Suite are able to compensate for large wire delays in the interconnect logic domain, by adding useful skew in the clock network. Generated by the Vivado Design Suite, the `report_pipeline_analysis` provides the guidance to reliably insert pipeline register stages in the interconnect logic while detecting sequential feedback loops causing performance bottlenecks. Lastly, retiming can be applied to balance path delays, especially after extra registers are added at the interfaces of the design during the pipelining process.

The Xilinx UltraScale+ Portfolio with SmartConnect technology solves the system interconnect bottlenecks for high performance, multi-million System Logic Cell designs.

Revision History

The following table shows the revision history for this document:

Date	Version	Description of Revisions
04/15/2016	v1.0	Initial Xilinx release.

Disclaimer

Automotive Applications Disclaimer