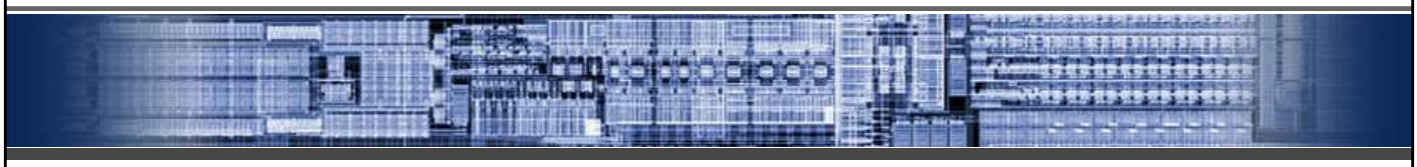
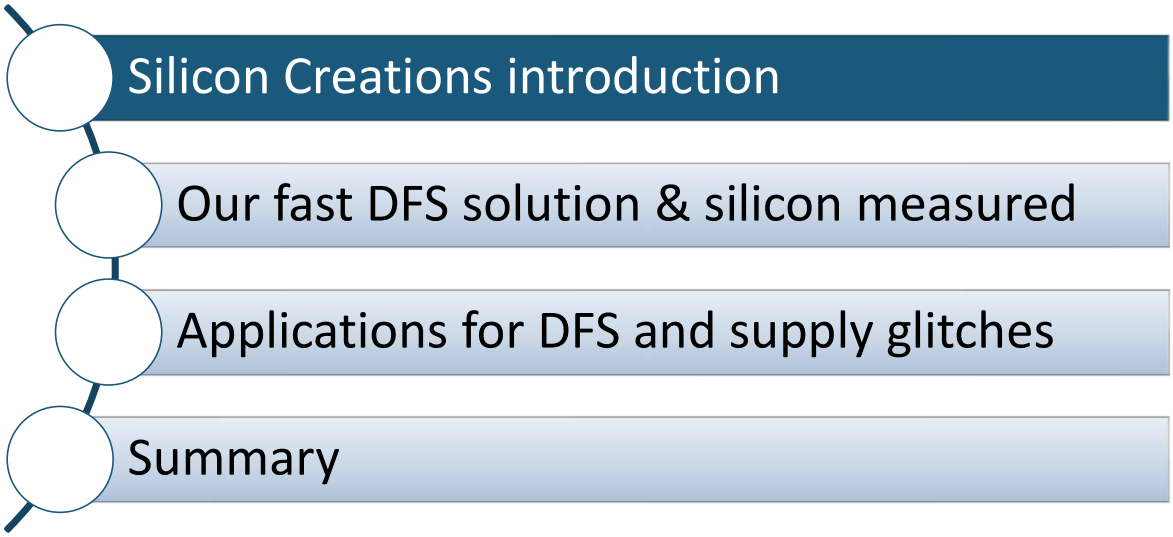


Maximizing SoC Bandwidth using Dynamic Voltage and Frequency Scaling

Randy Caplan, Andrew Cole, CC Chen

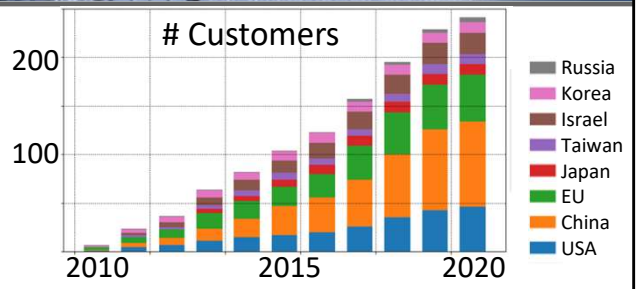
Outline



In this talk we will be introducing two new kinds of products that have been made at Silicon Creations for ultra-fast DFS and talking about some difficult problems that can be solved with these new products.

Silicon Creations Overview

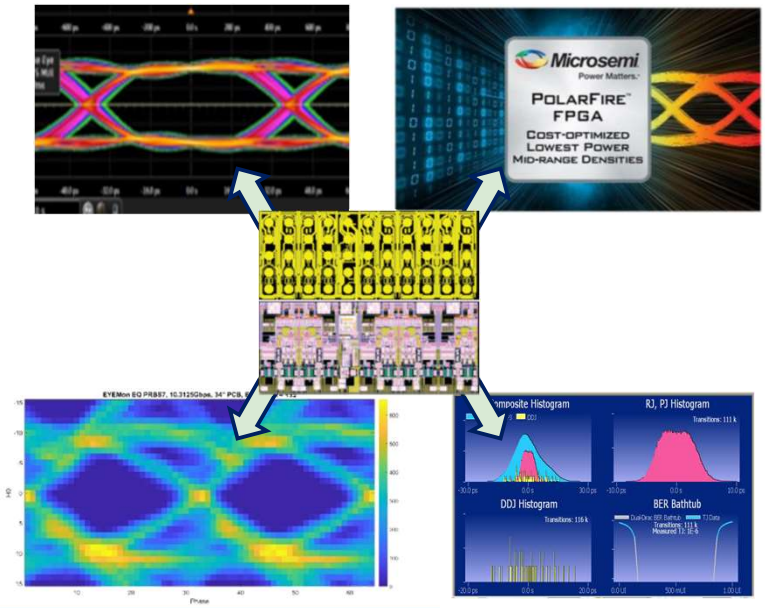
- IP provider: PLLs, Oscillators and SerDes
- Design: Atlanta, USA and Krakow, Poland
- TSMC 2017, 2018, 2019, 2020 and 2021 "Analog Mixed-Signal Partner of the Year"
- SMIC awards 2013 – 2016 including "Best Customer Support"
- ISO9001 certified
- Mass production from 5nm to 180nm
- 3nm PLL ready for design starts
- >330 customers, >120 in China



We have been providing clocking IPs to our fabless IC customers since 2006 and have been growing consistently since 2010 as you can see from the graph of customers by region and year. We have development centers in Atlanta, USA and Krakow Poland and our customers chips using our IPs are in volume production from 180nm down to 5nm FF. Our **3nm PLLs are silicon proven already**. Our design engineers understand their main mission is to help our customers get to market quickly and our quality and support has been recognized with multiple best-of awards including best AMS from TSMC for the past 5 years. We are also ISO9001 certified which is appreciated by our customers making automotive chips.

SerDes from Silicon Creations

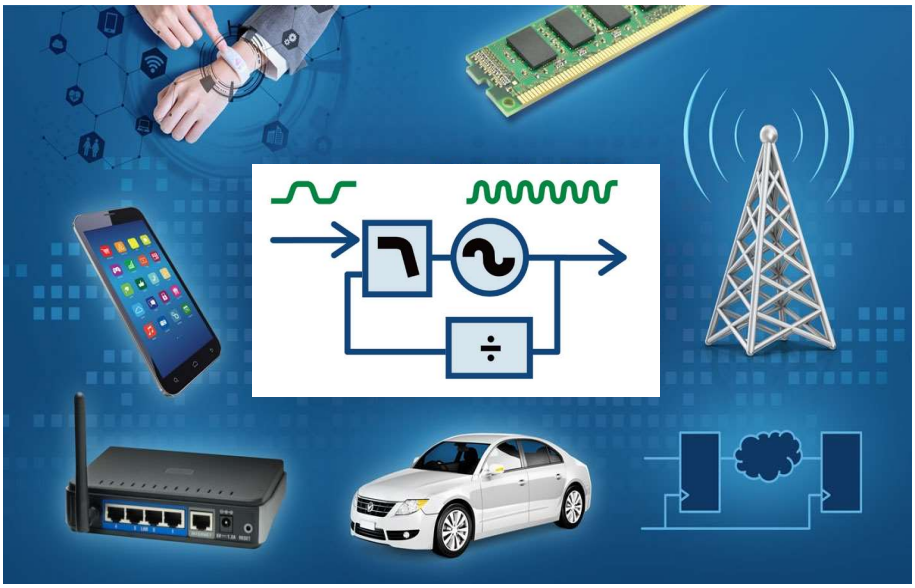
- Robust and in production from 12nm to 180nm and from <100Mbps to >25Gbps
- Multiprotocol in TSMC, GF, UMC 6nm, 12nm, 16nm, 28nm, 40nm
- PCIe5 available now
- Targeted protocols including SGMII, XAUI, RapidIO, V-by-1 HS/US, DP, FPDLink, OIF-CEI, JESD204, CPRI, PCIe1-5, 10G-KR, …



SerDes is one of our key product lines, and we focus primarily on the PMA – the high-speed analog part including the serializer, line driver, line receiver, equalization, CDR and deserialization. Our PLLs have enabled extremely good performance in the Microsemi FPGA, and also very low power. Over the past few years we've ported and proven this multiprotocol PMA to a number of mainline process nodes. In addition to our multiprotocol PMA, we have protocol-specific PMAs built for V-by-one, Display Port, JESD204, FPD-link and several custom interfaces as well as ultra-low latency PMAs.

PLLs from Silicon Creations

- Highest volume analog IPs
  … >10 million wafers 5nm to 180nm
  … robust design and good QA are essential
- PLL products include general purpose, fractional, low jitter AFE, µW IoT, Automotive



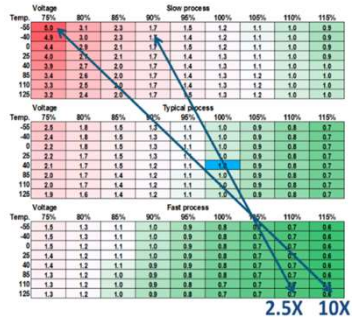
Around half our revenue still comes from PLLs. The core of this business is general purpose Fractional-synthesizer which so widely programmable it's been used in hundreds of chips and a wide variety of markets. This has resulted in some truly gigantic volumes. For example, in TSMC 28nm, over 140 production chips use this IP and in 12/16nm over 50 chips use it. This has resulted in many millions wafers delivered to customers with our PLL IP. These kinds of volumes are a testament to our robust, high-yield design and customer support. We've leveraged this high-volume design to make PLL variants that are optimized for low jitter, extremely low power and Deskew/clean-up PLLs for the Cadence DDR PHYs. Our 3nm silicon is being tested now and performance matches simulations.

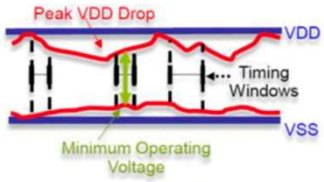
Outline

Now we will introduce you to a new kind of IP – a Frequency Generator or Frequency Locked Loop that is capable of switching frequencies within a few nsec making it ideal for new kinds of applications.

Why DFS?

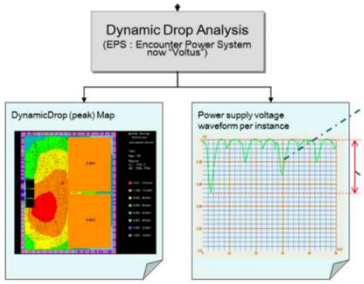
- Logic speed may vary 10X across PVT
  SoC designers must close timing at worst case (longest delay) corner, which limits speed in normal conditions
- **D**ynamic **F**requency **S**caling enables frequency to scale with supply voltage and temperature, giving optimal timing margin at any condition
  Supply can be set for process corner, but PVT sensors and PLLs cannot respond to rapid supply dips due to large events (e.g., enabling or resetting a large logic block)

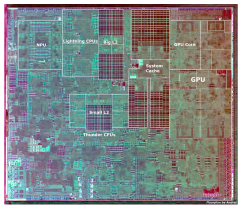




Effective Voltage Calculation [Tesla]

Courtesy of: cadence

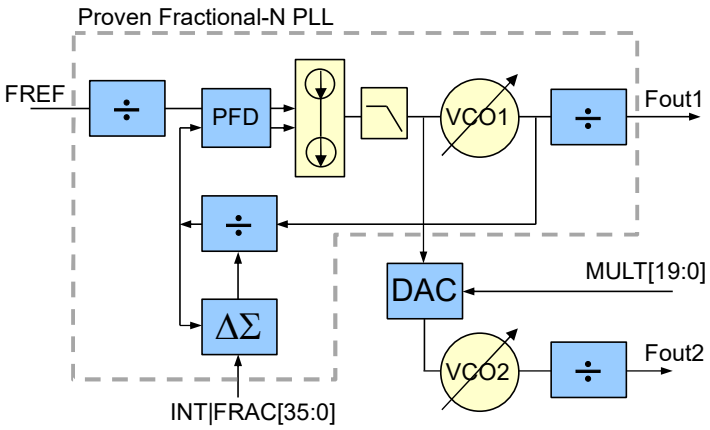


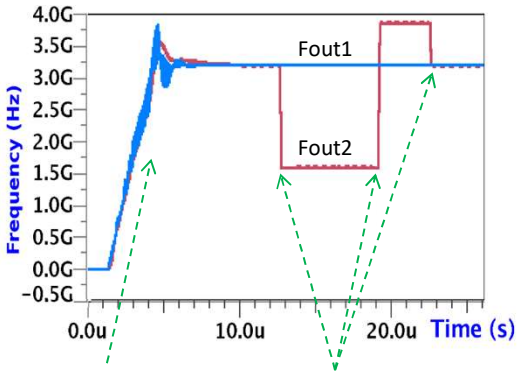


Example 5nm SoC: 8.5B transistors, many subsystems

Firstly, though we're going to talk about why designers use Dynamic Frequency Scaling. The matrix on RHS shows how logic gate speed varies with P, V & T. The variation is gigantic – 2.5x over normal voltage ranges, or up to 10x over the extended range commonly used in modern SoC's and especially in HP Compute chips and AI accelerators. Because logic timing must be met for the worst case, most chips in production are performing well below their potential speed. Supply voltage is shown on the horizontal axis of the tables, and it's obvious this effect dominates. This means that even if chips are binned into fast parts and slow parts, and the supply voltage is adjusted accordingly, dynamic changes in the supply voltage must be accounted for by allowing more timing margin. If the clock rate can be adjusted as the power supply level changes, the chips can run faster when the supply is high and slower when the supply drops.

Frequency Generator





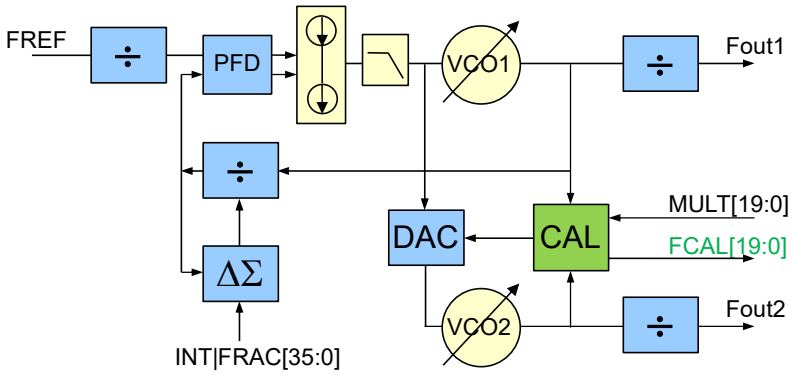
PLL Settling:
VCO2 follows VCO (~5us)

Frequency Hopping:
Asynchronous DAC switching
Rise/fall time ~12ns. No glitches, missing pulses or overshoot

We have built a new product that is designed specifically to generate the clock needed in the situation where supply voltages change rapidly. It's called a Frequency Generator. This product is built from a proven Fractional N PLL so is a low-risk derivative. We take a copy of the locked PLL VCO current and pass that through a DAC to drive a copy of the VCO. This provides a second output which is frequency locked to a fractional multiple of the reference clock and that can be programmed over a 2.5:1 frequency range before an asynchronous output divider. The DAC is designed to be asynchronously programmable and settle very quickly without overshoot. While PLLs initially settle on timescales of microseconds, and because they are higher order loops, they either overshoot or settle very slowly, the DAC does not overshoot at all and settles on a timescale of 12ns. This enables DFS with smaller timing margin and with frequency changes 50x to 100x faster than before. The frequency switching is even fast enough to detect a supply glitch and drop the logic clock ahead of the supply drop.

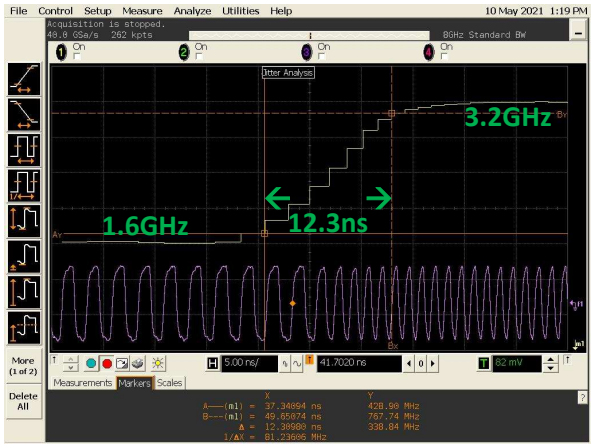
Self-calibration for mismatch

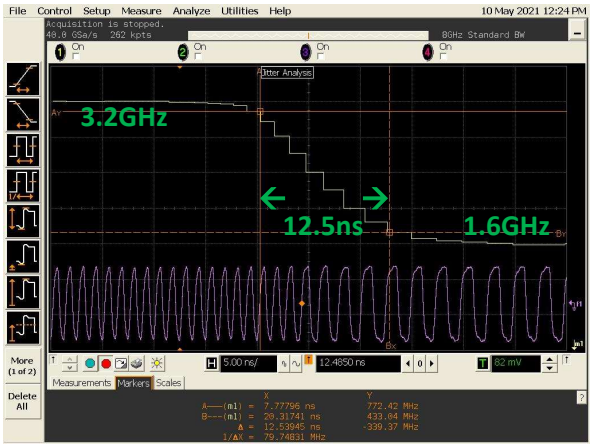
- Fout2 = Fout1 * MULT
  (0.6x to 1.5x with postdiv = 1)
- MULT is set asynchronously and Fout2 slews to new frequency with 12ns rise/fall time
- Due to device mismatches between VCOs, output may initially be up to 2% off target
- Slow digital loop continuously compares actual frequency and target, adjusts DAC for exact match, outputs calibrated MULT value (FCAL)
- FCAL can be stored in a LUT so Fout2 will be exact immediately



As we all know, transistors in the most advanced processes do not match very well. This means the output of the VCO2 will not be perfectly on target. In this design the copy can have a frequency error in the range of 2%. To deal with this, we have added a slow digital calibration loop that adjusts the DAC control word so that the actual ratio of the two VCO output frequencies matches the programmed ratio. Once this calibration is complete, the calibrated DAC control word can be read out, stored in a Look-up table in non-volatile memory and used for future transitions to the same frequency. With one point calibration the initial error on frequency jumps is below 0.5% and with multiple calibration points can be better than 0.01%. This calibration can be performed at production test, or at first boot of the system.

5nm silicon measurement





5nm FG silicon settles to new frequency with τ < 7ns and no overshoot

We developed our first FG in 5nm, and this slide shows some measurements with the DAC programmed to have the output switch from 1.6GHz to 3.2GHz and back. You can see the output clock (divided down to 800MHz to get it off chip) transitions rapidly and smoothly to the target frequency without glitches or overshoot. The steps are an artifact of updating the frequency after each clock period measured. The equivalent time constant we measure here is around 7ns which is likely faster than a supply droop in a chip with acceptable on-die decoupling.

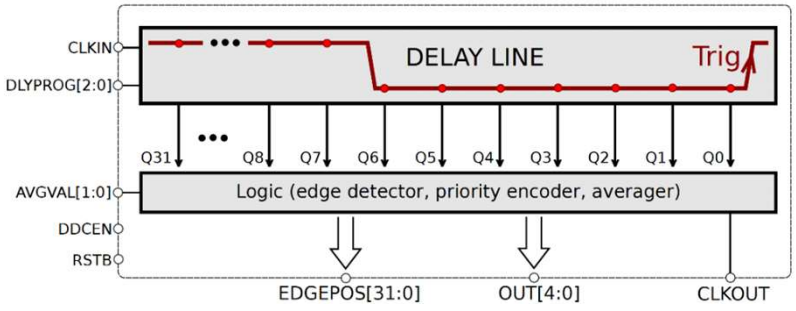
[It turns out we can make the edges a little faster using pre-emphasis! That's a complication that's not needed here – it would be too much information and distract from the basic message. If 7nm is too slow for someone they will tell us.]

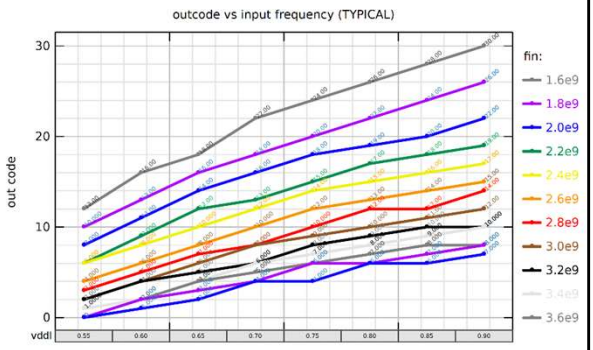
Outline

Now we will take a closer look at applications for the Frequency Generator and introduce a companion IP that we have developed.

Companion IP – the "DDC"

- The **D**elay to **D**igital **C**onverter is a delay line built from core logic cells
- The DDC measures logic delay on each clock input (as quickly as every 2ns) and outputs a code that scales with logic delay
- Needs one-time calibration for process variations

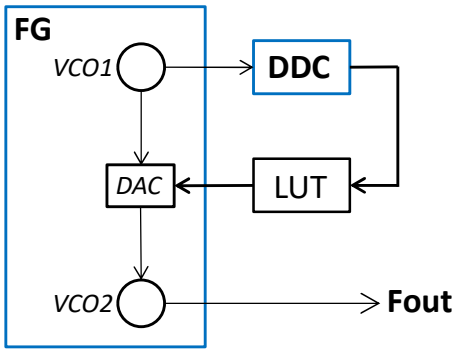




The companion IP is a new kind of circuit building block that we have named a DDC. This is a Delay to Digital Converter. Because the FG will be used to react to changes in logic circuit timing, we believe that it is better to directly sense the logic timing than to detect the three factors, process, voltage and temperature that impact timing, and attempt a complex calculation to predict what clock frequency will be acceptable. The DDC is clocked by a fixed frequency and generates an output word that represents the core logic delay at the moment of the clock. The DDC can also average this delay measurement over several clocks. The graph on the right shows how the DDC output code varies with supply voltage for different clock rates, chip process corners and junction temperatures. As expected, the output codes increase monotonically with supply voltage.

Typical DFS Application

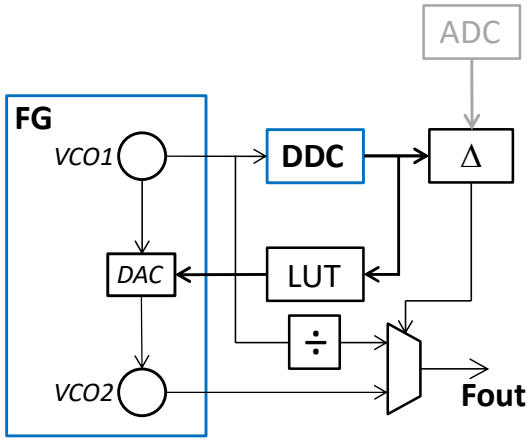
- Fractional-N PLL's VCO outputs a fixed clock for the DDC
- One-time calibration of DDC and VCO/DAC mismatches at production test or first boot of system, stored in LUT
- If changes in die temperature or supply voltage cause logic to slow down, DDC output will show this within 5ns to 10nsec
- VCO2 automatically outputs appropriate frequency based on DDC code and LUT so that logic can always run as fast as possible for each chip (process corner and current Vdd)
- DDC can be placed close to controlled logic



The FG and DDC can be used for automatic DFS using the simple application diagram on this slide. The fixed output frequency of VCO1 is used to clock the DDC generating codes that represent the instantaneous logic speed. These codes can be input to a LUT to generate a desired frequency ratio between the outputs of VCO1 and VCO2. The LUT needs to be calibrated once for each chip, and the actual LUT values will then account for mismatch effects in the DAC, VCOs and the DDC as well as process corner and supply voltage effects. This will enable each chip to be calibrated to always run at maximum possible safe frequency no matter what the current junction temperature and supply voltage are. Whether the chips have been speed-binned or not, this method enables chip designers to obtain significantly more performance without needing to move to more expensive process nodes.

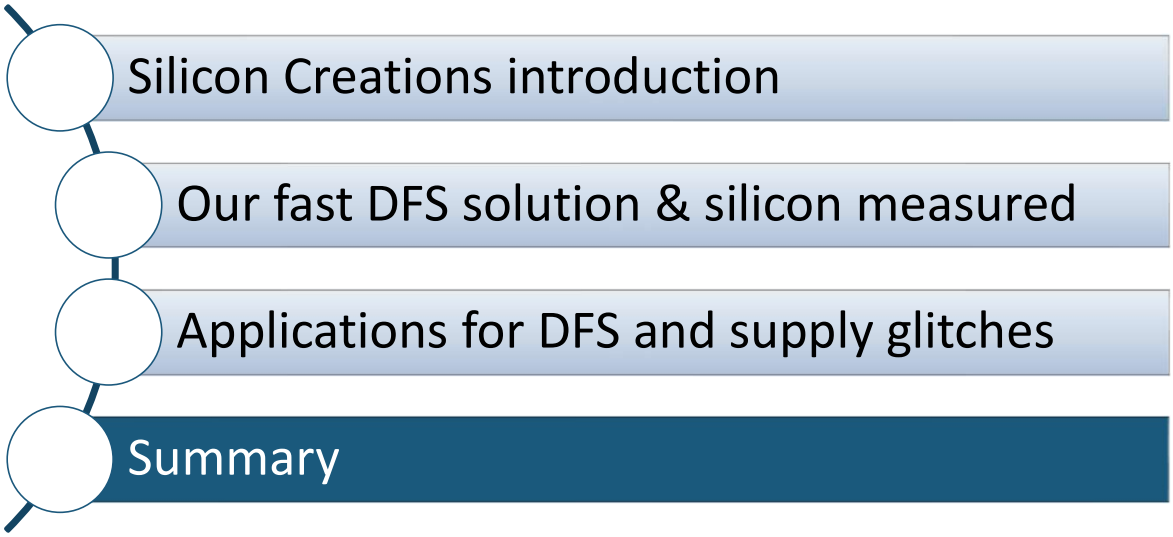
Reacting to Supply Dips

- In normal operation, DDC controls Fout as before
- Supply dip is detected (in "Δ" block) by rapid change in DDC output and clock MUX switches quickly to slower clock
- Switch back to VCO2 for controlled recovery to fastest logic rate as supply recovers
- Optional ADC (glitch detector) can also command MUX



With a small addition, the DDC and FG can also enable chips to automatically adjust logic clock rates when an event outside a logic circuit causes the supply to droop. This might be something like a reset of a set of memories or starting several processor cores at the same time. This kind of event can cause rapid increases in supply current, and the package and PCB inductance results in fast supply droops. In most cases, a rapid change in the DDC code can be asynchronously detected in a digital comparator to indicate the start of such a supply droop. The comparator can immediately switch the logic clock to a safe, low frequency. After a short time, the circuit can switch back to the VCO2 output which is controlled by the LUT and DAC. It is not necessary to have knowledge of, or design for the dynamic behavior of the on-die and PCB level supplies or how the package impacts their dynamic behavior for each unit built. The DDC+LUT loop will automatically adapt in each case. Some early adopters of our solution believe that a supply glitch detector will be needed in some chips as the supply dips will be too fast even for a DDC clocked at 2GHz. While we have not found this to be necessary the diagram shows how a supply voltage dip detector can provide input to the digital comparator to trigger the clock MUX.

Outline



Thank you for listening to our presentation. We trust it has been worth your time.

Summary

- Silicon Creations has been providing reliable, high-performance clocking and SerDes solutions since 2006.
- Our PLLs are in production from 180nm to 5nm … > 10 million of wafers. We are the lowest risk solution for your clocking needs. 3nm PLL are available now.
- Our Multiprotocol SerDes solutions are available to 32Gbps and equally low risk.
- New FG product for DFS was created from a proven Fractional-N PLL and generates output able to hop between frequencies without overshoot and in ~10ns
- New DDC product directly and rapidly measures logic delay. When combined with the FG this makes automatic DFS possible in ~10ns without complex, risky, slow and inaccurate Speed = f(process, voltage) calculations.
- In most cases this is even fast enough to respond to supply glitches, but a simple addition can make it even faster.

We began by giving you a high-level overview of our company and main product lines and alerted you to our 5 years of best-of awards from TSMC, and recent silicon success in 3nm. Following this we introduced you to a new kind of circuit that we call the FG. This IP is built from any of our proven Fractional-N PLLs but adds a second output whose frequency can be changed over 50x faster than a PLL can adjust and does this without glitches or overshoot.

We then introduced you to a second new kind of circuit that we call a Delay to Digital Converter which directly measures instantaneous logic delay.

Using the DDC together with the FG enables very DFS with response times of less than 10ns without the need to understand and predict the complex relationship between process, voltage and die temperature and the resulting maximum logic speed. This pair of IPs will help you get more processing cycles out of the same designs without needing to move to more expensive wafer processes.

We think that the DDC is fast enough in practical logic designs to even adjust for supply glitches but showed you how you can react even faster.

Once again, we thank you for your valuable time.