IEEE EDPS 2022

The Future of SoC Connectivity

The Dawn of Clockless Technology

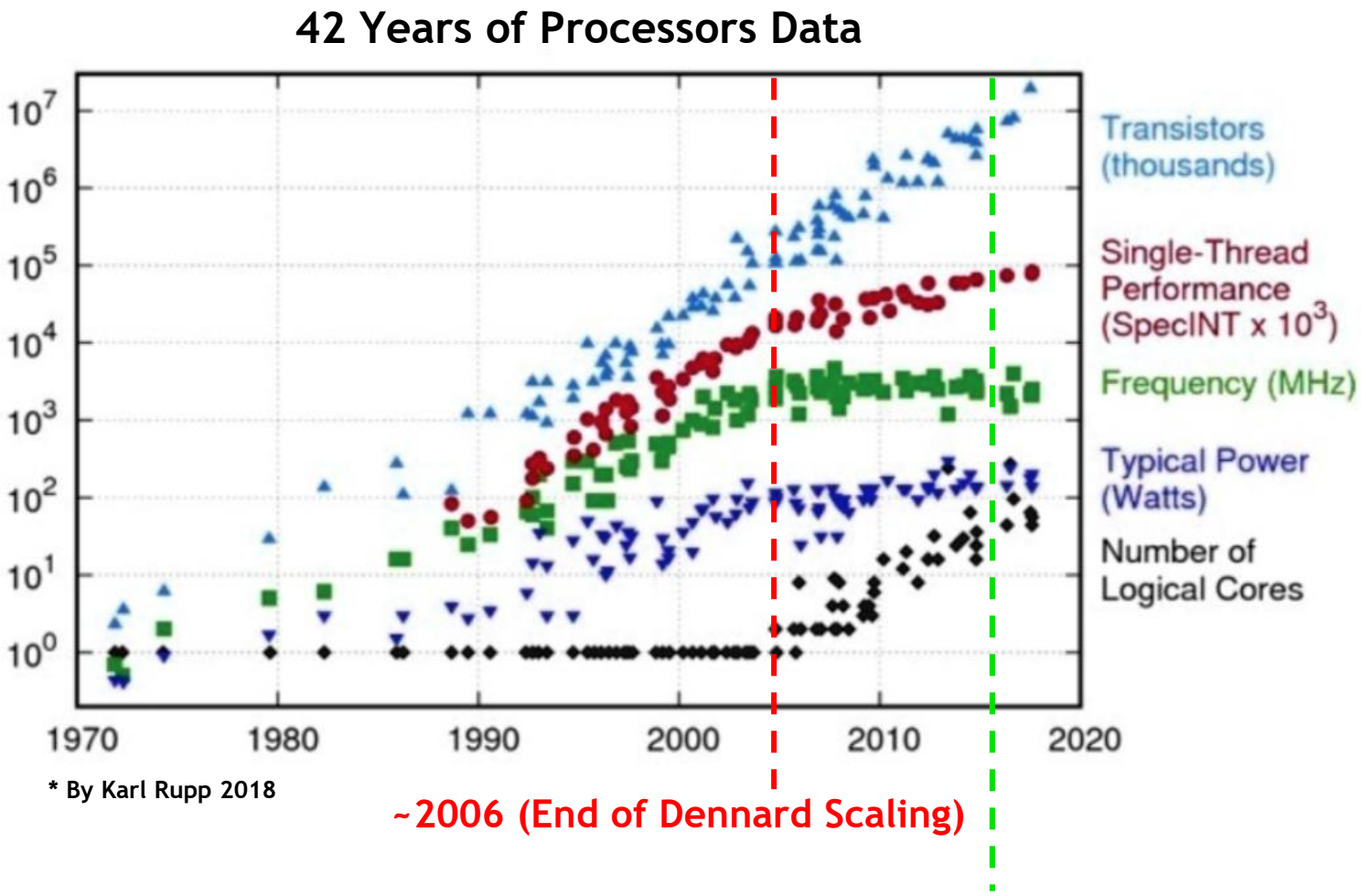
October 5th, 2022

Introduction

- The introduction of AI and ML to computing systems, paired with the widespread availability of High-Speed mobile devices, has put a lot of pressure on traditional computing architectures
- New applications require huge amount of data bandwidth both in the cloud and at the edge, directly challenging the interconnect infrastructure in every system
- The fundamental assumptions of traditional digital design are pushed to the limit requiring a different disruptive technology to sustain progress
- Fortunately, this technology already exists, and can enable high performance scalable solution for years to come

History of Computing Architectures

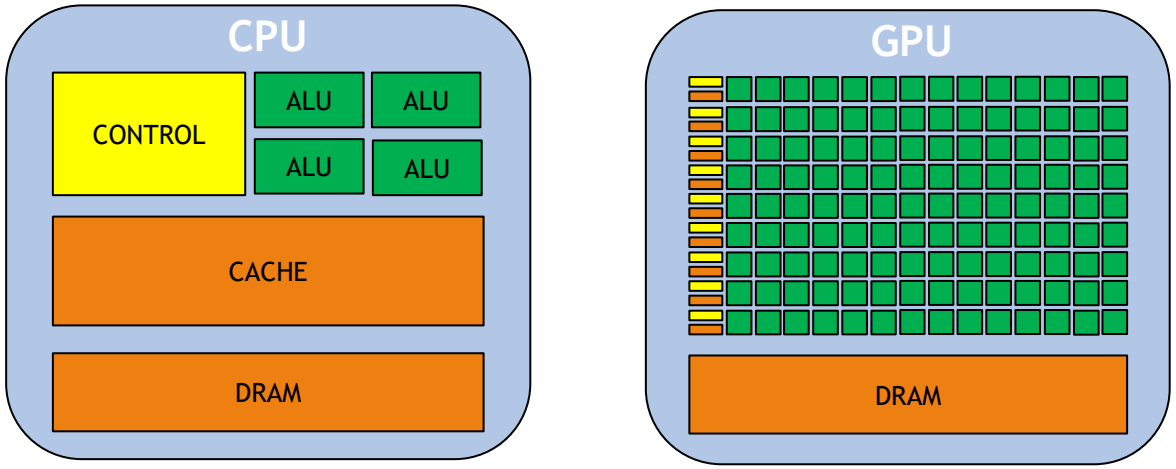
A Quest for Higher Performance



- **Moore's Law:** "The number of transistors in a dense IC doubles every 2 years"
- **Dennard Scaling:** "As transistors get smaller, their power density stays constant"
  - Pushed smaller geometries and higher clock frequencies for next generations
- **Amdahl's Law:** $S_{\text{latency}}(s) = \frac{1}{(1-p) + \frac{p}{s}}$
  - More cores as alternative to higher frequency to keep pushing performance within a limited power envelope

GPU architecture

CPU Vs GPU



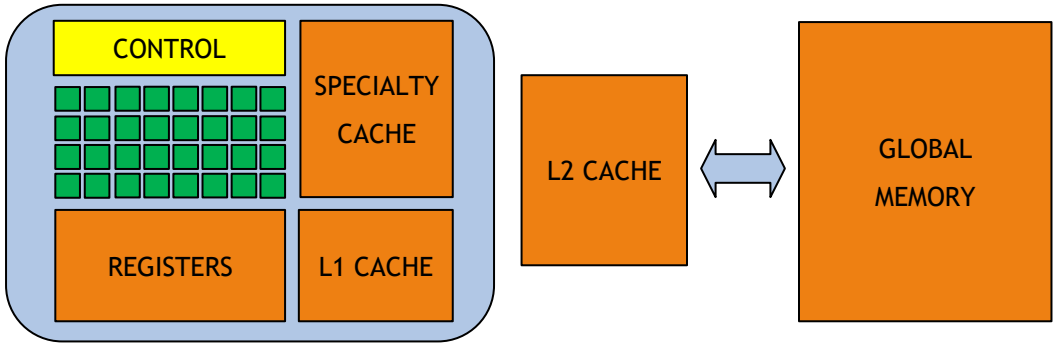
CPUs:

- Good at running complex tasks
- Easier to program (sequential machines)
- Cheaper than GPUs

GPUs:

- Good at running simple tasks on large data such as matrix multiplications
- Can hide memory latency if the data fits within the internal memory

Modern GPU memory hierarchy



- Registers are the fastest memory in the system followed by L1 and Specialty Cache, L2 Shared Cache and finally the Global Memory
- The faster is the memory the smaller is the size of the memory
- Large amount of memory are required by modern inference algorithms

Another Brick in "The Memory Wall"

EXAMPLE

- 32x32 matrix multiply
- 8 Stream Multiprocessors
- 8 warp per SM

D = A x B

OPERATION DATA

- **Global Memory access** (up to 48GB): ~200 cycles
- **Shared memory access** (up to 165kb per SM): ~20 cycles
- **Fused Multiply-Add** (FMA): 4 cycles
- **Tensor Core matrix multiply**: 1 cycle
- **Warp** (smallest unit of threads on a GPU): 32

Latency cost w/o Tensor Cores

Global Memory: 200 cycles + Shared Memory: 8 * 20 cycles + FMA: 8 * 4 cycles = **392 cycles**

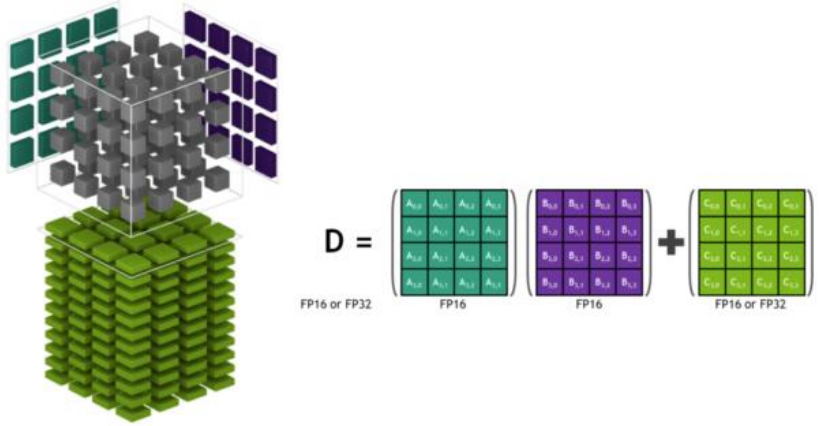
Each SM does 8x dot products

Latency cost w/ Tensor Cores

Global Memory: 200 cycles + Shared Memory: 20 cycles + Tensor Core: 1 cycle = **221 cycles**

No matter how much you improve the computing unit, the limiting factor is by far **MEMORY ACCESS**

Tensor Core operation



Interconnect is the Bottleneck

"One of the things that we have been seeing is the model sizes are exploding, no model fits in one node."

"While there is a lot of hype on deep learning accelerators, their utilization is super-low because we are busy moving the parameter data across the network because the 100-billion parameters don't fit."

Intel architecture day event, August 2020

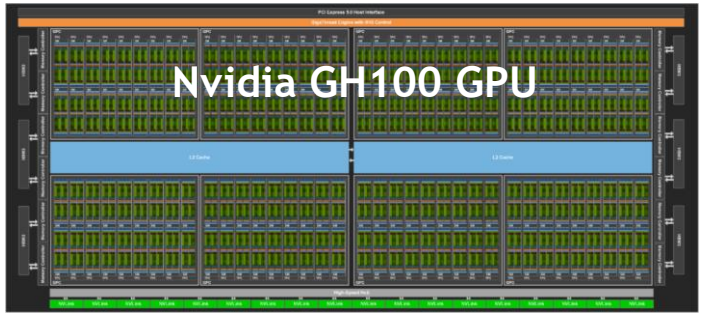
Raja Koduri
(Head of architecture @ Intel)

- Complexity of latest deep learning models is growing exponentially.
- The latest released Pathways Language Model (PaLM) from Google has 540-billion parameters
  - **weights that must be multiplied over each piece of input data**
- Future models for real time Imaging and Video recognition will grow even more

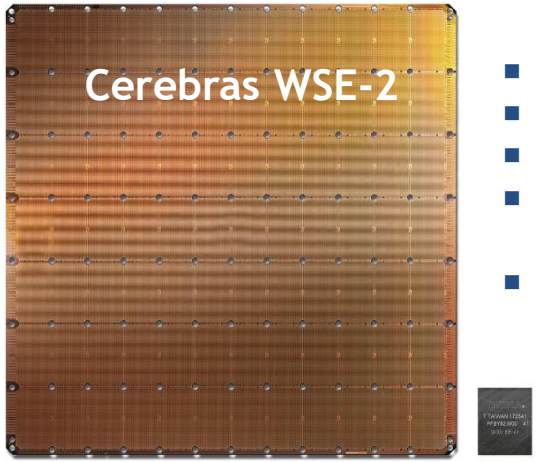
- Training of large models have Tensor Core utilization of ~30%
  - **~70% of the time is spent moving data around**
- Inference at the edge has even tighter requirements for latency performance
  - **Especially in real time applications**

Market Segments

Cloud AI: Training / Inference



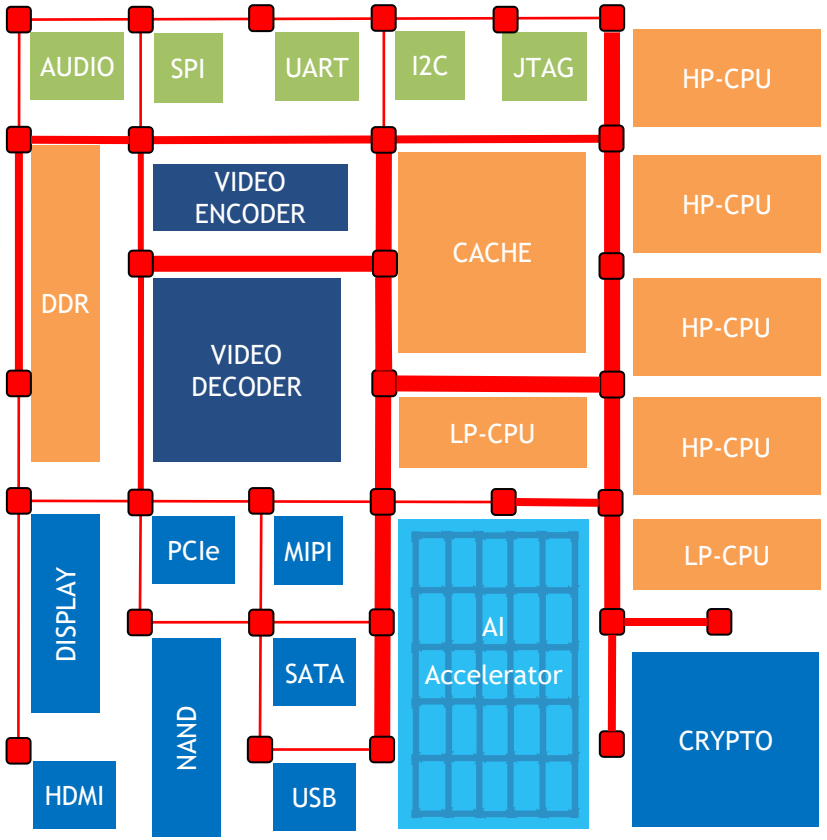
- TSMC 4nm 4N
- 80B Transistors
- 826 mm^2
- 80GB of HBM2e stacks
- 2 Banks of 25MB L2 Cache



- TSMC 7nm
- 2.6T Transistors
- 46,225 mm^2
- 40GB of on-chip Memory
- 850,000 Core

- **Larger silicon dies to reduce memory access**
- **Repetitive structure**
- **Industry trend for next generation chip:**
  - Bigger size than its predecessor more cores and more on chip memory
- **Cores connected by Clocked Distributed NoC**
  - Simplified router
  - Complex Clock distribution network
  - Challenging timing closure
  - Clock frequency determined by core area
    - Usually, no pipelining between routers
    - Core could run at higher speed
- **Hybrid Bonding**
  - Extend area beyond die size
  - Challenging timing across different silicon dies

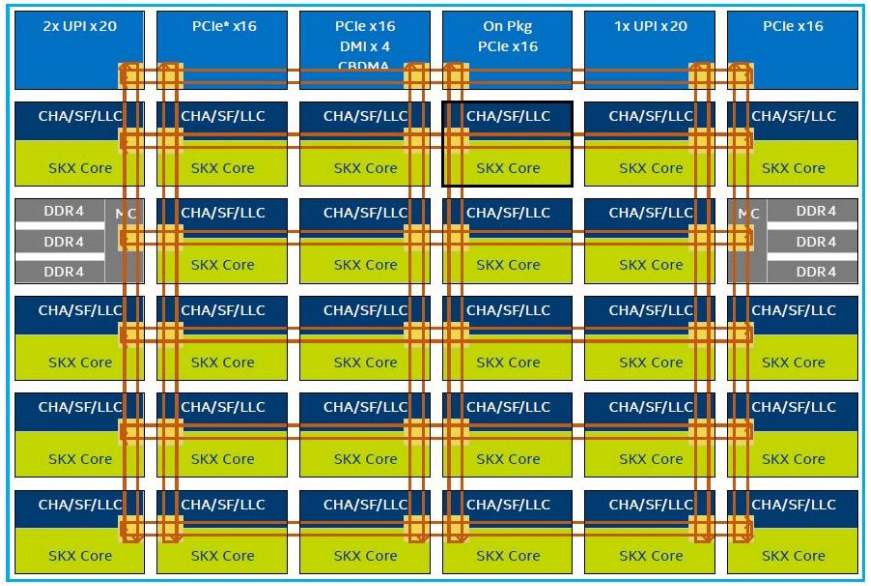
Mobile / Edge-Computing / Automotive



- Extremely Complex SoC
- Heterogeneous System
  - Many cores and sensors
- Security required
- Dedicated Computing Units for specific tasks
  - AI Acceleration
  - Crypto
- High Throughput Requirement
  - HD-Video - Gaming
- Low Latency Requirement
  - Real-Time applications - 5G
- IPs connected through Clocked Distributed NoCs
  - Very High-speed clocks (↑ Throughput, ↓ Latency)
  - Difficult to close timing at top level
  - Challenging Floorplan
  - Difficult to scale

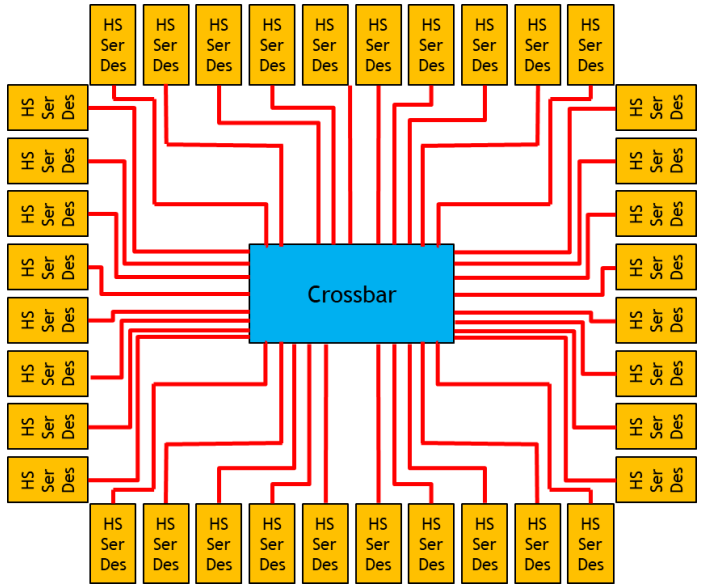
Data Center

Intel Skylake-SP 28-core



CHA – Caching and Home Agent ; SF – Snoop Filter; LLC – Last Level Cache;
SKX Core – Skylake Server Core; UPI – Intel® UltraPath Interconnect

HPC Switch



- **Server Chip**
  - High-Bandwidth High-Throughput chip
  - Mesh architecture
  - High Frequency
  - Long interconnect (Horizontal)
  - Challenging interconnect design

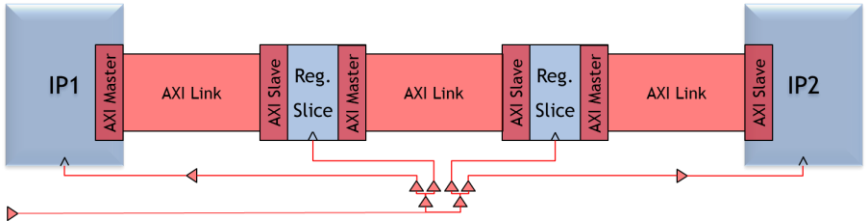
- **Die area is determined by other factors**
  - Die size close to reticle limit to maximize the # of High-Speed IOs at the boundary
  - Low Latency requirements
  - Very long interconnects

A dark time for Interconnect

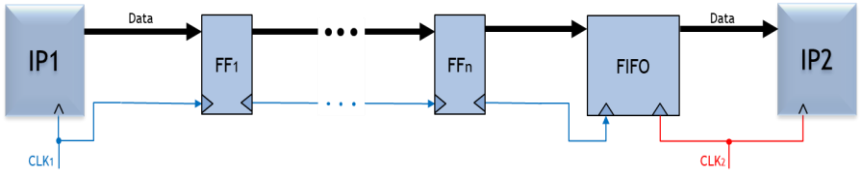
…but the sun is rising

Clocked Interconnect Challenges

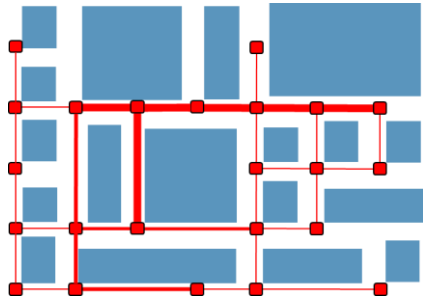
- **Traditional Synchronous Pipelining:**
  - Extremely complex clock distribution for long/wide busses
  - Cannot achieve the required low latency
  - Challenging timing across PVT and modes
  - Suffers from degradation for long distances and high freq.



- **Source-Synchronous:**
  - Difficult to deploy on large scale (Semi-Custom)
  - Suffers from degradation for long distances and high freq.
    - Clock Skew and Duty Cycle
  - Designed and operates in WC

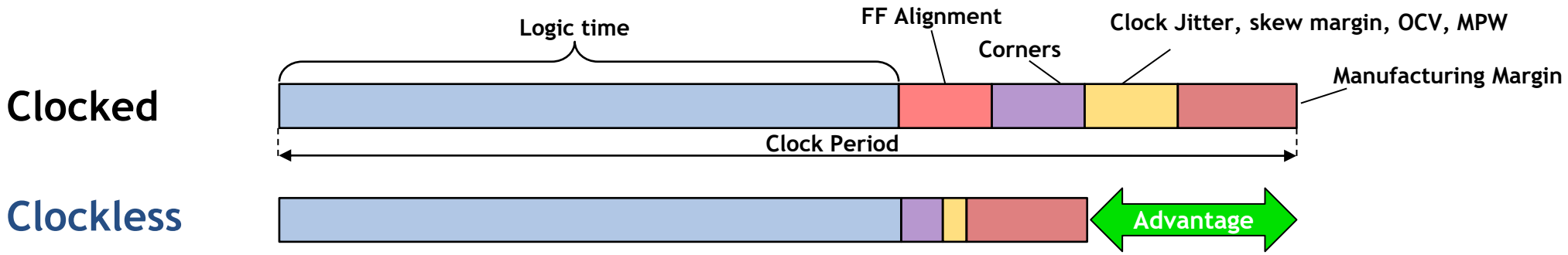


- **Clocked Distributed NoC Mesh**
  - Massive clock distribution
  - In order to achieve low latency requires high-speed clocks
  - Nightmare to close timing at top level
  - Latency and Throughput are correlated

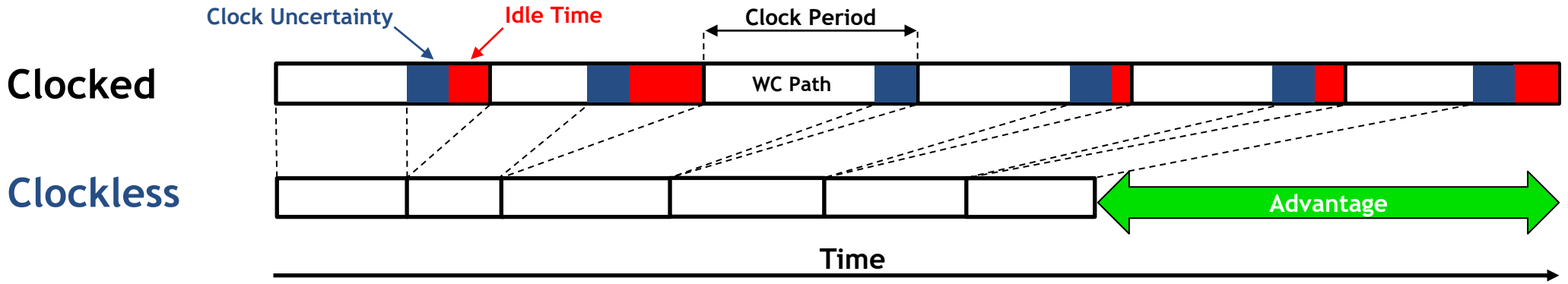


Clockless Data Transport Advantage

Throughput



Latency



Why Clockless Now?

Clockless technology benefits

- **Performance**
  - Fast (Low-Latency, High-Throughput)
  - Self clock gating (No clock distribution)
  - Reduced footprint (Serialization Technology)
  - Scalable (not limited in length)
- **More resilient**
  - Designed for WC, but operates on TYP
  - No Hold-Time Violation possible
  - Resilient to PVT
  - Low EMI
- **Ideal for Heterogeneous Systems and High-Performance Switches**

Historic shortfalls

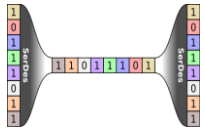
- **Deployment**
  - No Standard flow integration
  - Proprietary Custom Tools and Flow
  - Complex and error prone sign-off
- **Post silicon issues**
  - Custom Dynamic logic cells
  - Lack of test capabilities (Std. test and HVM)
- **Poor power and area**
  - QDI Encoding doubles the wires (larger Area)
  - Higher Power

Once shortfalls are eliminated -> Perfect match for current requirements

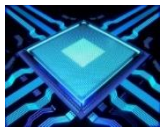
Technology and Product Offering

- **Chronos Technology is an advanced fabric built on the following pillars:**

Low Latency Communication



Clockless Serialization



Deployable with Standard Logic & Tools

Embedded On Die Telemetry

- **Products:**

Link

Clockless Point-to-Point Interconnect

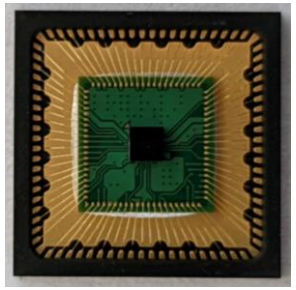
- **Competitive Advantage:**
  - Best in class latency & throughput
    - Advantage grows at low voltage
  - Bus width serialization technology
    - Enables tradeoff between area and perf
  - Implementation via industry standard tools
- **Readiness:** Available now
  - Internal TO (7nm TSMC) (Sangiovese)
  - Verified on Silicon with Tier1 Cust. TO (5nm)
    - Customer currently defining product intercept

AI NoC

Clockless NoC for AI Accelerators

- **Competitive Advantage:**
  - Addresses latency bottle-neck in AI
  - Each Compute Unit can truly be independent
    - Enables improved power/perf optimization
  - Bus width serialization technology
    - Enables possible mesh area reduction
  - Low Power
- **Readiness:** Available Q1/23
  - Development and verification in progress

Sangiovese Test-Chip

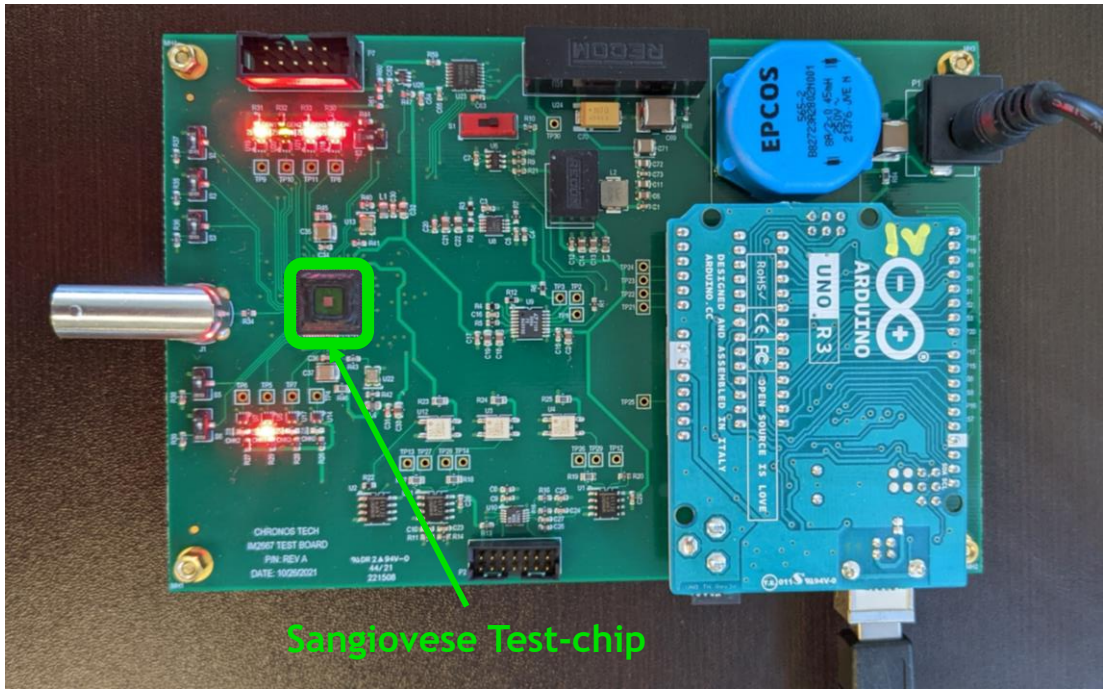


TSMC N7 CMOS LOGIC Fin FET ELK Cu, 1P13M, HKMG, 0.75/1.8V

H300 ULVT/LVT Chronos cells used in the design

Wafer: T8U996.00#01

Dies mounted on interposer and wire-bonded to standard QFN package

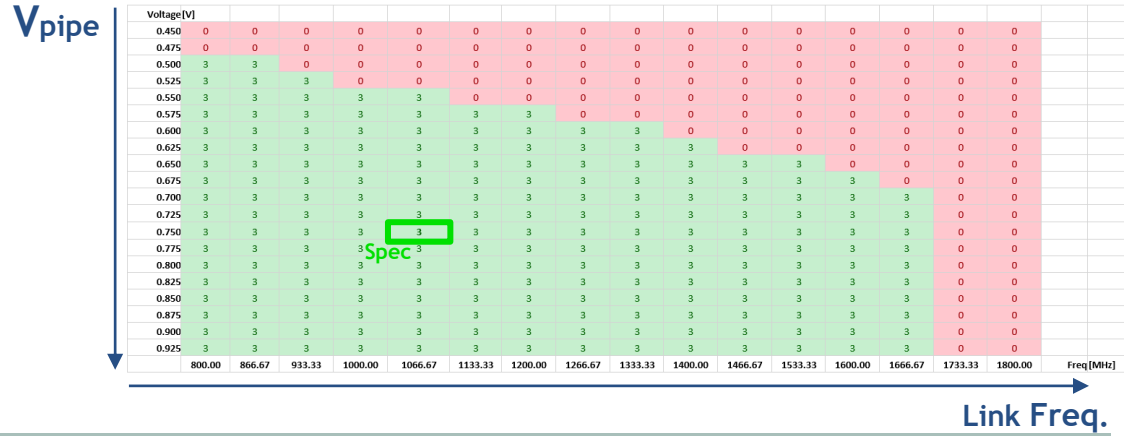


- Demo Package available (environment + sim + board)

Specifications

- Features: 2 PLLs, multiple voltage domains, programmable correlated and uncorrelated noise generators
- Generator and Checker able to produce pseudo-random and custom-pattern data and controls (Valid, Ready)
- Different Chronos links (with different PPA optimization)
- Channel lengths: 10mm
- Buses data width: 64bits

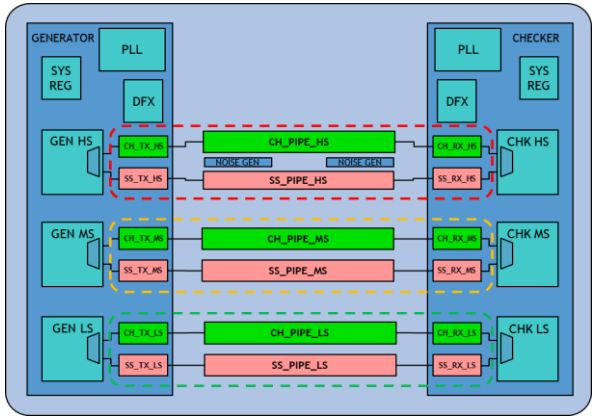
- Latency <6.5ns
- Throughput maintained down to 550mV in the pipeline, saving ~49% power
  - Power Efficiency: 0.047pJ/bit/mm *



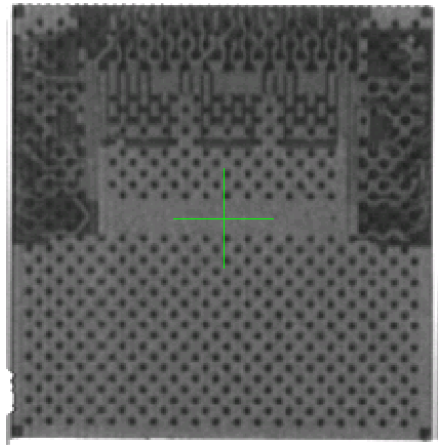
* LFSR31 Data Pattern, 100% bus utilization

Silicon Results: Chronos vs Source Synchronous

Test-Chip



Technology: 5nm FinFET



Specifications	High-Speed	Medium-Speed	Low-Speed
Frequency	**1.563Ghz**	**781MHz**	**195MHz**
Serialization Ratio	x2	x4	x16
Effective Channel Speed	**3.126GHz**	**3.124GHz**	**3.12GHz**
Link Length	10mm	10mm	10mm
Voltage Domains	5	5	5
Bus Data Width	64bit	64bit	64bit

Results (NN)	0.75V	0.58V	0.75V	0.58V	0.75V	0.58V
Throughput Advantage	**46.6%**	**115%**	**46.5%**	**99%**	**46.6%**	**98%**
Latency Advantage	**30%**	**40%**	**54%**	**63%**	**39%**	**58%**
Bus Width Reduction	**1.2%**		**48.1%**		**86.8%**	

- **Best in class Latency: 7.6ns (over 10mm @ 0.75V)**
  - Includes serialization, deserialization and synchronization

The Sun is Out

- Traditional clocked interconnect technology is facing significant challenges in modern computing architectures
- Large die area, high throughput and low latency requirements are demanding a flexible new solution
- Clockless technology is the ideal fit for new systems, enabling
  - High-Throughput
  - Low-Latency
  - PVT resilience
  - Low EMI
  - Easy of Deployment
  - Tool/Flow integration (No need to re-train the team)
  - Advanced testing capabilities
- EDA vendors and design houses can help accelerate the transition by
  - Expanding the std-cell libraries to include common clockless elements
  - Integrating Relative Timing Constraints (RTC) within the flow
  - Enabling timing analysis of logic loops within the data-path
  - Integrating test vector optimization with custom data encoding

Thank you

“The electric light did not come from the continuous improvement of candles”

(Prof. Oren Harari)

Join Us on the Clockless Revolution!