Chip Architecture Trends in 2022

Narrative from several sources…

Sharan Kalwani

Tuesday, June 14, 2022

IEEE-CNSV
Consultants' Network of Silicon Valley

About Me

- ✓ Senior Member of IEEE, SAE, ASME, ACM
- ✓ Currently Chair of the IEEE SE Michigan Section (since 12/2021)
- ✓ Editor of the Section's Monthly Newsletter ~ *Wavelengths*
- ✓ Published author & enjoys teaching, student mentoring
- ✓ Advisory Board member – CS Department @ Loyola University-Chicago
- ✓ Recipient of the IEEE MGA Achievement Award (2018)
- ✓ IEEE Region 4 Jack Sherman Outstanding Member Award (2021)
- ✓ IEEE Southeastern Michigan Robert Neff Section Award (2022)
- ✓ Started life in tech compute/Cray/SGI/Automotive sector/

LinkedIn ~ https://www.linkedin.com/in/sharankalwani

IEEE-CNSV
Consultants' Network of Silicon Valley

Report Outline

- Chip technology
- Major Areas
- New ideas to come...

Major Areas

- Technology Areas
- A few words on ML first
  - Connecting them all together (start off with this)
  - CPUS
  - Assisted processors
- Advent of AI for dedicated chips
- Using AI for CHIP design!

IEEE-CNSV
Consultants' Network of Silicon Valley

Current tech areas

- Theme: *ML Performance and Real World Applications*
- *Players:*
  - NVIDIA *(Modern Neural Networks and their Computational profiles)*
  - Google *(MLPerf™ Training and Inference) & Large TPU machines)*
  - Graphcore *(Software/hardware co-optimization on the IPU)*
  - Intel *(update)*
  - Facebook *(AI at Scale for the Modern Era)*
  - Amazon
  - Microsoft *(Comment on their work on Extreme Scale Deep Learning)*

IEEE-CNSV
Consultants' Network of Silicon Valley

Areas not covered

- *Advanced Packaging*
- *But these folks are very much into packaging….*
  - Intel *(Intel packaging technologies for chiplets and 3D) & (Case Study: Intel products built with 2.5D and 3D packaging)*
  - TSMC *(TSMC packaging technologies for chiplets and 3D)*
  - AMD
  - *SAMSUNG*
  - *IBM*

IEEE-CNSV
Consultants' Network of Silicon Valley

Start: All of them devices!!!

- CXL
- Compute Express Link is an **industry**-supported Cache-Coherent Interconnect for **Processors**, **Memory** Expansion and Accelerators.
- CXL is built on the PCI Express (PCIe) physical and electrical interface
  - PCIe-based block input/output protocol (***CXL.io***) and
  - Cache-coherent protocols for accessing system memory (***CXL.cache***) and
  - Device memory (***CXL.mem***)

IEEE-CNSV
Consultants' Network
of Silicon Valley

Devices: CXL

- History of CXL
- Intel, Microsoft, HPE, Huawei, Google, Facebook, DELL, CISCO, Alibaba, etc.
- *Later (2022): AMD, NVidia, Samsung and Xilinx joined the founders on the board of directors, while ARM, Broadcom, Ericsson, IBM, Keysight, Kioxia, Marvell, Mellanox, Microchip, Micron, Oracle, Qualcomm, Rambus, Renesas, Seagate, SK Hynix, Synopsys, and Western Digital, among others, joined as contributing members.*
- *Gen-Z and CXL have joined hands*
- *Several standards are bridging their protocols*

IEEE-CNSV
Consultants' Network of Silicon Valley

Devices: CXL

- Latest Developments
- *CXL 1.0 specs ~ 2019 (based on PICe 5.0)*
- *CXL 1.1 released (allows switching and several shared memory/storage configurations)*
- *CXL support ⇨ SAMSUNG 128 GB DDR-5, Intel "Sapphire Rapid", AMD Zen 4, …*
- *CXL 2.0 specs ~ released 11/2020*

- ***CXL 3.0 ~ expected release 2022 (based on PCIe 6.0)***

IEEE-CNSV
Consultants' Network of Silicon Valley

Day 2: Advances in Chip Architectures

- Several areas
  - General purpose CPUs
  - Different Chip Designs
  - Infrastructure & Data Processors
  - Future Enabling Technologies

IEEE-CNSV
Consultants' Network of Silicon Valley

CPUs

- CPUs (mainstream stuff)
  - Intel CPU Architecture(s)
  - AMD Next Gen Zen
  - IBM 5+ GHz Z processor chip

IEEE-CNSV
Consultants' Network of Silicon Valley

Intel

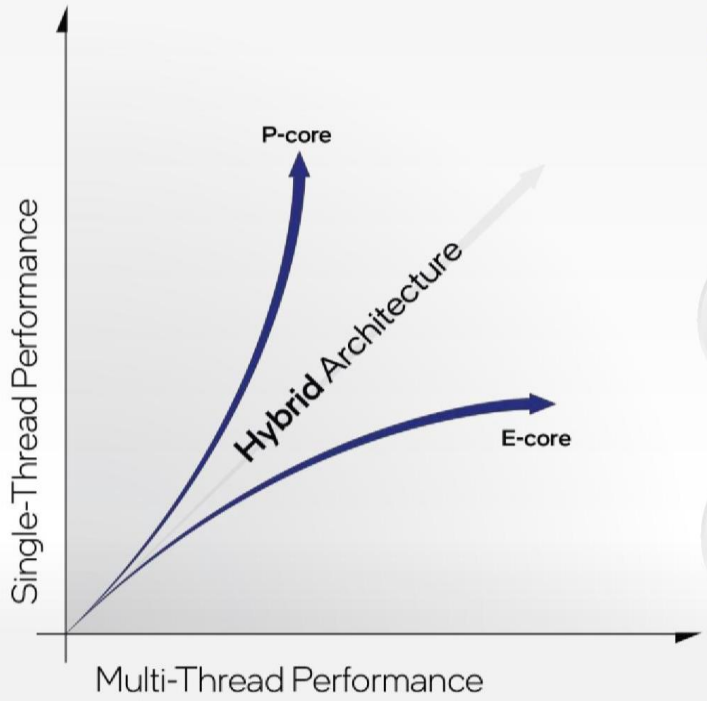
- Intel CPU Architecture Directions
- Hybrid combines “high-performance” and “energy-efficient” cores
- *Intel 7 process (aka 10 nm Enhanced SuperFIN)* 12th Gen Core, now 14th Gen Cores (more on this later)
- Example Alder Lake ~16 cores (8 p-cores + 8 e-cores)

IEEE-CNSV
Consultants' Network of Silicon Valley

Intel

- Intel e-cores and p-cores

Alder Lake Performance Hybrid



Same architecture
Different micro architecture
Different optimization and design point

P·core — Unleash the design to deliver even more ST performance

E·core — Created an efficient throughput machine

P-Core delivers higher Performance on single and lightly threaded scalable apps.

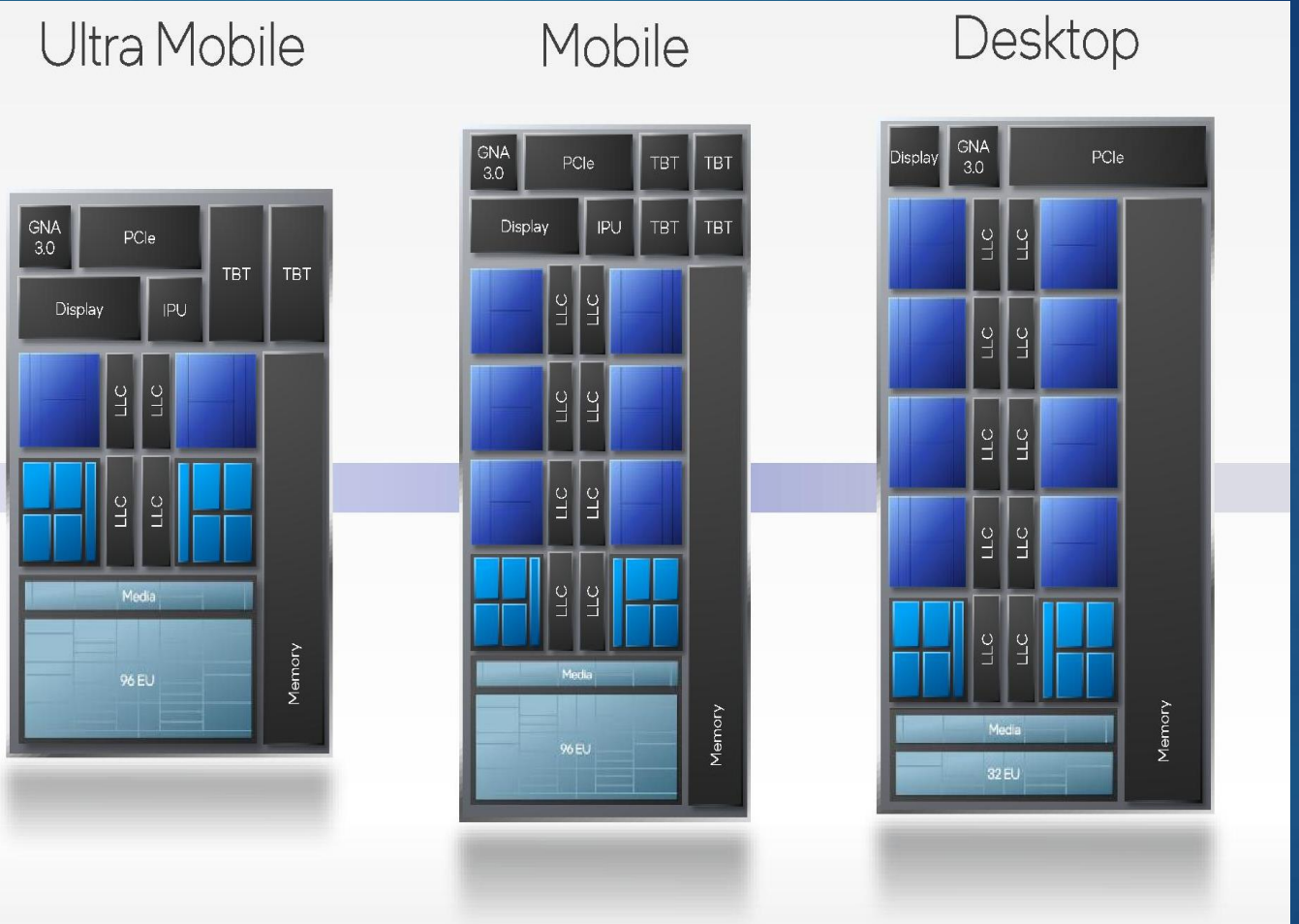
E-Core provide higher computational density under given physical constraints

IEEE-CNSV
Consultants' Network of Silicon Valley

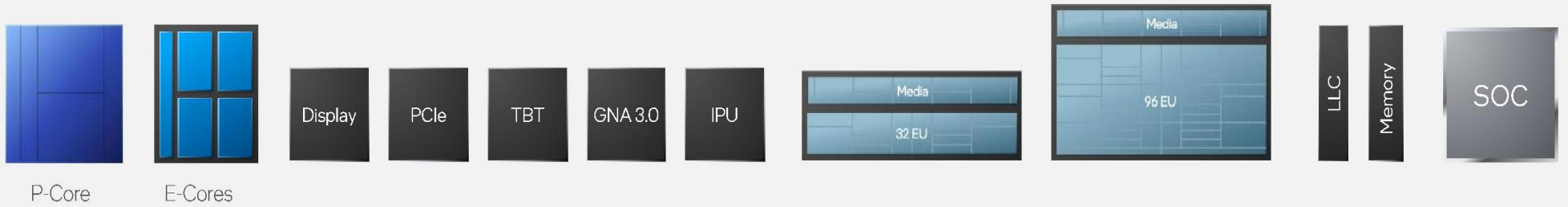
Intel

All cores are exposed to OS
Logical processors accessible

Hybrid topology enumerated
No hard coded information
Same OS and SW for all builds
HW support and management



Building Blocks



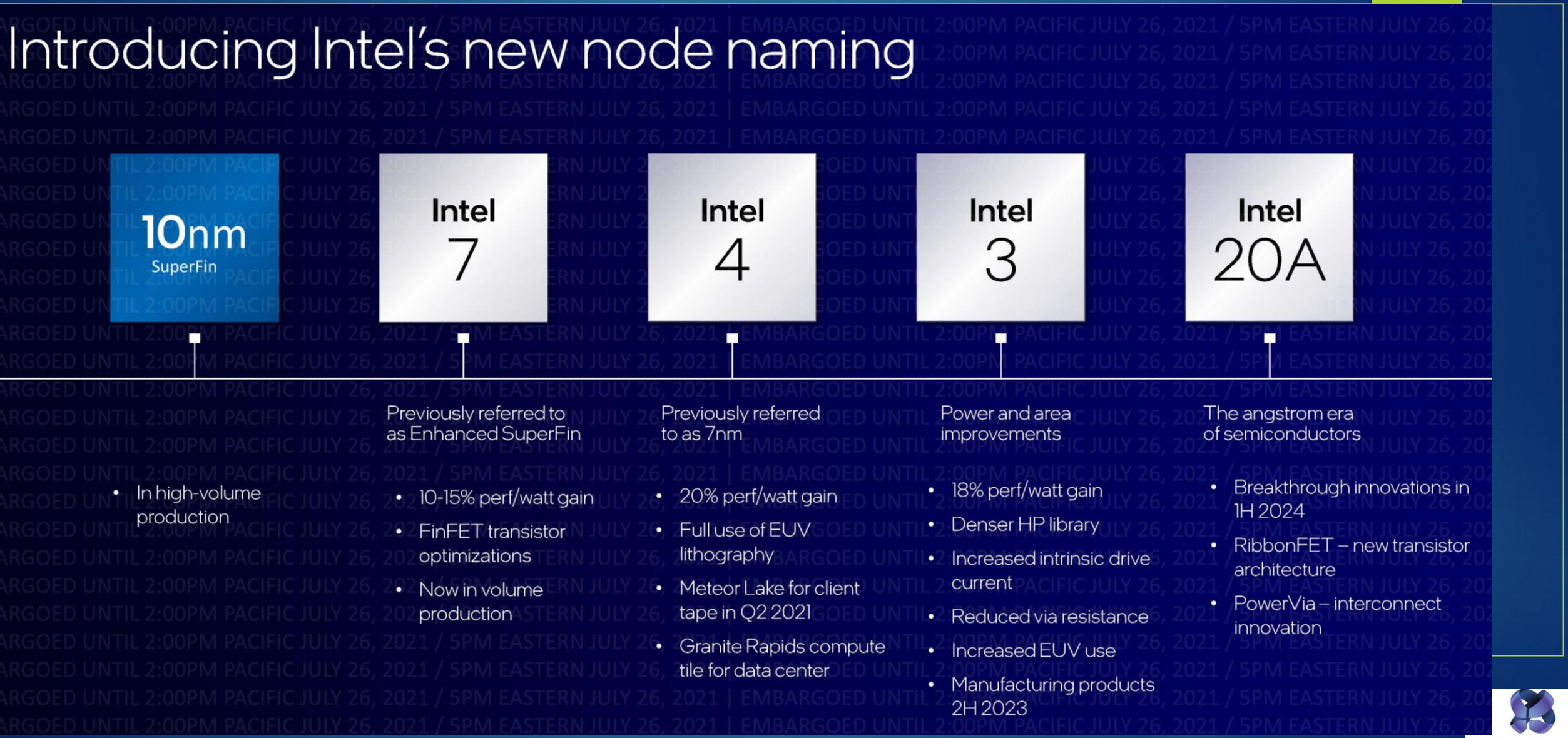
IEEE-CNSV
Consultants' Network of Silicon Valley

Intel

- Sapphire Rapid (intel 7)
- Emerald Rapids followed by
  - (p-cores Granite Rapids, e-cores Sierra Forest)
- (06/13/22)
- Intel technology nodes
  - 10 nm
  - Intel 7
  - Intel 4 (Meteor Lake)
  - Intel 3
  - Intel 20A

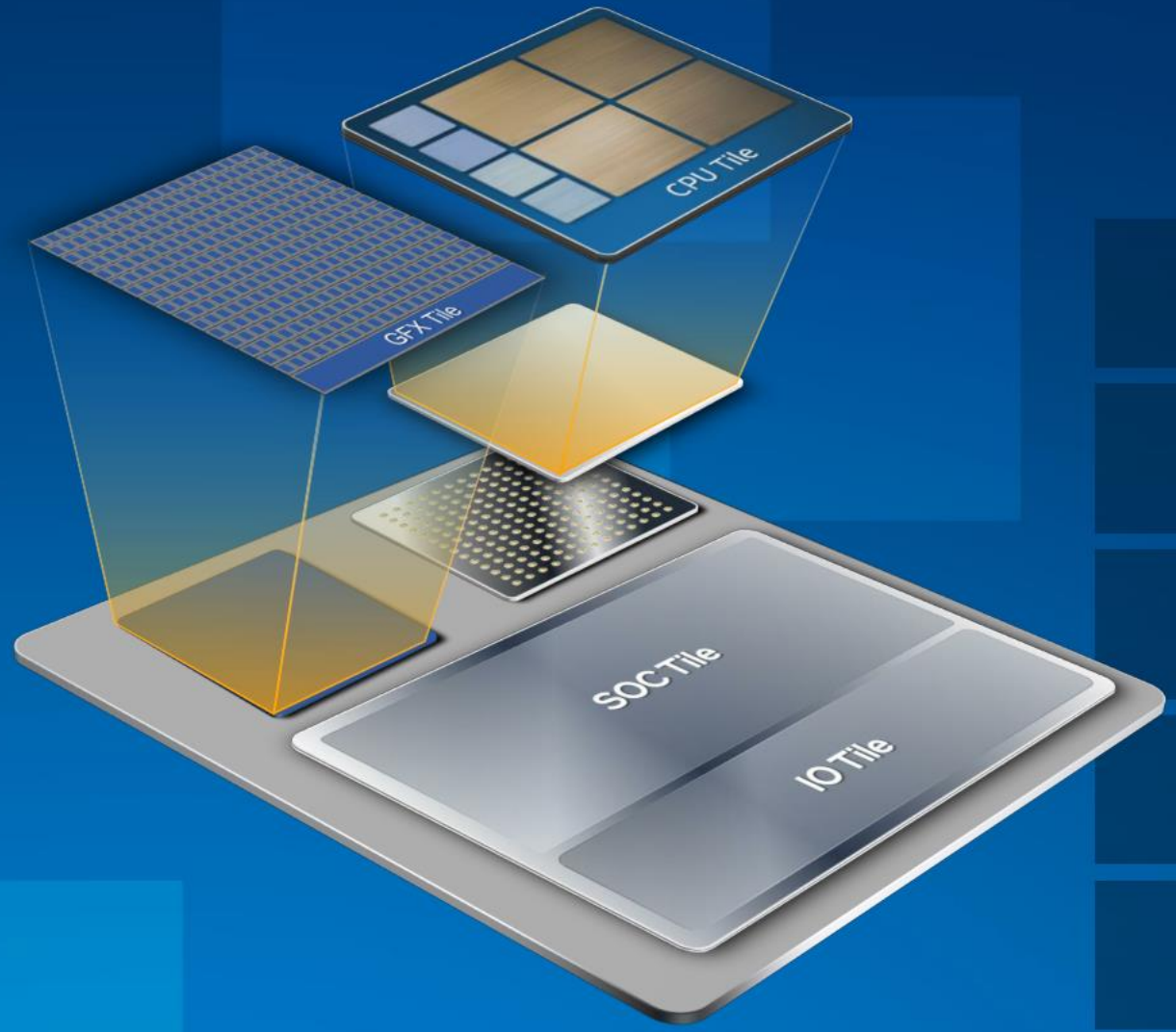
IEEE-CNSV
Consultants' Network of Silicon Valley

Intel (reminder to the audience here…)



IEEE-CNSV
Consultants' Network of Silicon Valley

Intel



Meteor Lake

built on Intel 4 process node

powering on in Q2'22, shipping in 2023

- New Flexible Tiled Architecture
- Hybrid Cores
- Lower Power
- Next Gen Graphics Engine: tGPU
- Integrated AI Acceleration

IEEE-CNSV
Consultants' Network of Silicon Valley

AMD

- AMD Zen-3
- Uses TSMC's 7 nm process for the chiplets and
- Global Foundries's 14 nm process for the I/O die on the server chips and 12 nm for desktop chips.
- Zen-3 powers Ryzen 5000 mainstream desktop processors and Epyc server processors (codenamed "**Milan**").
- Has a 19% higher instructions per cycle (IPC) on average than Zen 2.
- Zen-4 expected in late 2022, will use 5nm process

IEEE-CNSV
Consultants' Network of Silicon Valley

AMD

- AMD Zen-series

Gen	Year	Name	Cores
1st	2017	Naples	32 × Zen 1
2nd	2019	Rome	64 × Zen 2
3rd	2021	Milan	64 × Zen 3
4th	2022	Genoa	96 × Zen 4

IEEE-CNSV
Consultants' Network of Silicon Valley

AMD

- AMD Zen-4
- Late 2022
- TSMC 5nm process
- Ryzen 7000 (*"Raphael"*)
- Epyc 7004 (servers)
- 5.5 GHz clock
- 2x L2 cache/core

IEEE-CNSV
Consultants' Network of Silicon Valley

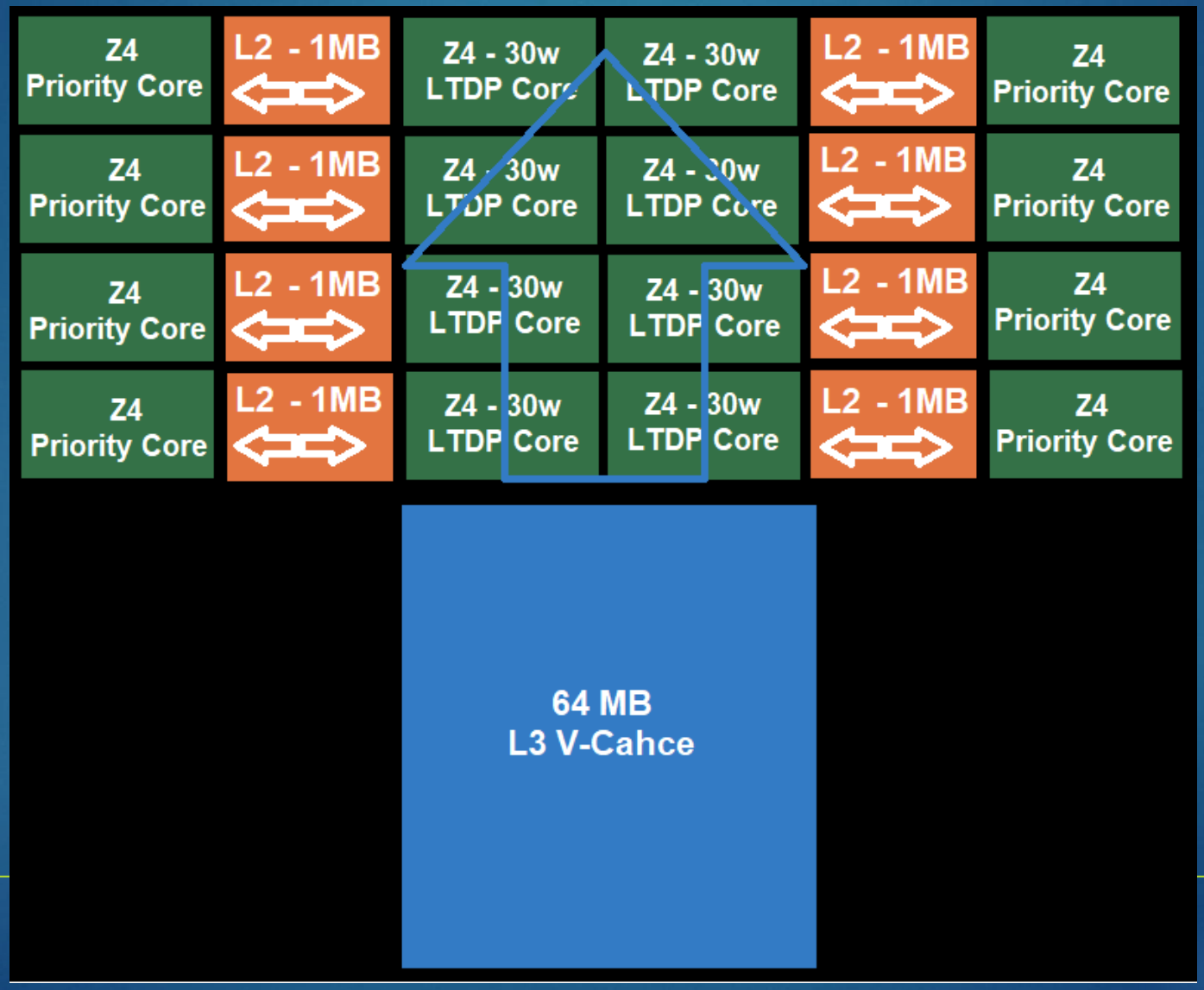
AMD

- AMD Zen-5
- Late 2024
- TSMC 3 nm process
- Ryzen 8000 (“Granite Ridge”!?)
- Epyc 7005 (“Turin”)

IEEE-CNSV
Consultants' Network of Silicon Valley

AMD

- AMD Zen-4



IEEE-CNSV
Consultants' Network of Silicon Valley

IBM

- IBM 5+ GHz Z processor chip
  - Codename: “Telum” for their IBM Z mainframe computers,
  - Telum is IBM's first processor that contains on-chip acceleration for AI inferencing while a transaction is taking place.
  - Telum-based system is planned for 2022
  - Based on Samsung 7 nm process
  - 540 mm^2 chip size
- https://spectrum.ieee.org/mainframe-ibm-z16-telum

IEEE-CNSV
Consultants' Network of Silicon Valley

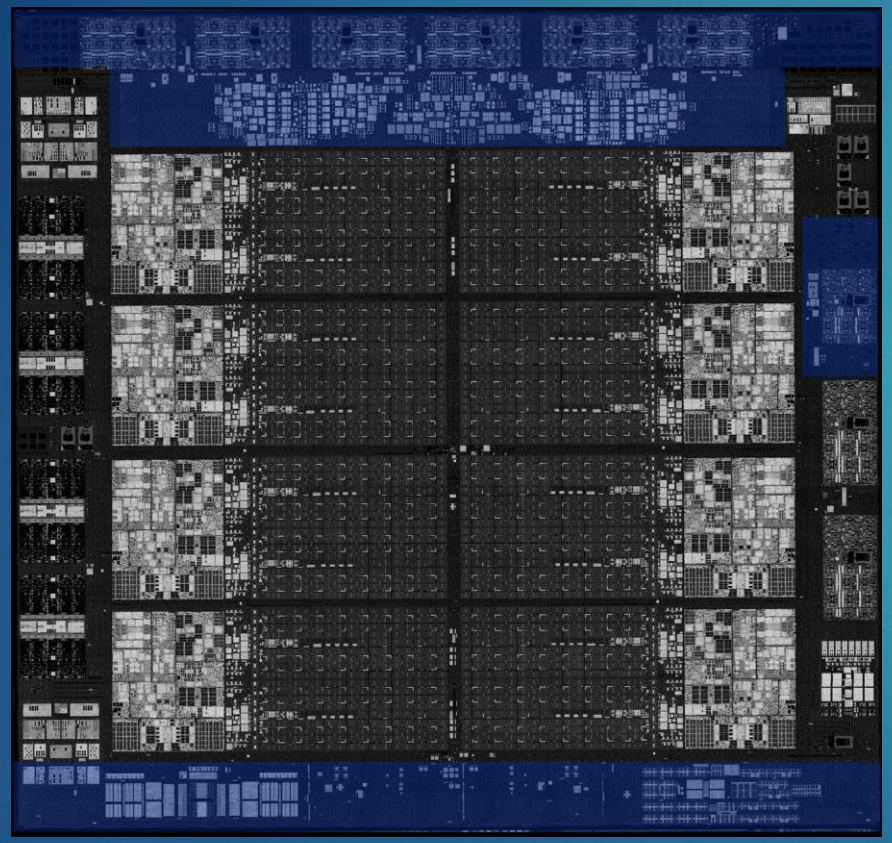
IBM

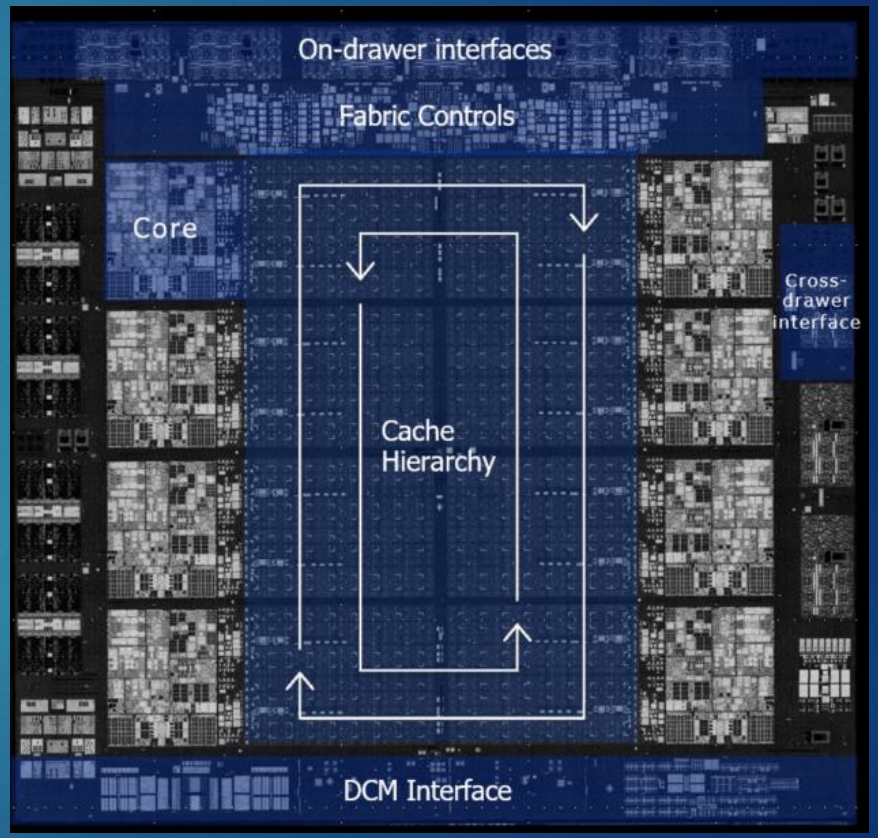
- IBM 5+ GHz Z processor chip
  - The chip contains 8 processor cores w/
  - Deep superscalar out-of-order pipeline,
  - All this at ***more than 5 GHz clock frequency***!
  - Completely redesigned $ & chip-interconnect infrastructure
  - Provides 32 MB $/core, and can scale to 32 Telum chips.
  - The dual-chip module design contains:
    - 22 billion transistors and
    - 19 miles of wire on
    - 17 metal layers

IEEE-CNSV
Consultants' Network of Silicon Valley

IBM

- IBM 5+ GHz Z processor chip

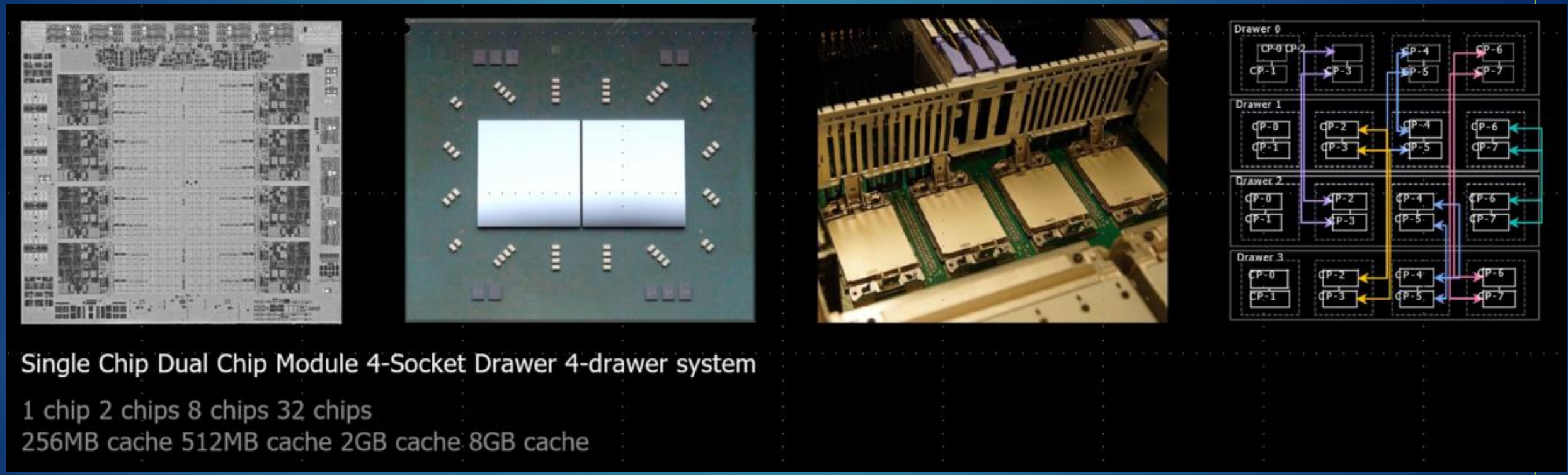




IEEE-CNSV
Consultants' Network of Silicon Valley

IBM

- IBM 5+ GHz Z processor chip



IEEE-CNSV
Consultants' Network of Silicon Valley

Intel revisited (stacking/bonding/multi tech integration)

- Next-Gen Intel Xeon CPU – Sapphire Rapids

IEEE-CNSV
Consultants' Network of Silicon Valley

Intel: Reminds me of 56 Leonard Street, NY

- Intel Xeon CPU – Sapphire Rapids
- Two key areas improved: Node & Data Center performance
- Node performance:
  - New micro-core architecture
  - Multiple integrated engines
  - Increased core counts
  - Larger private & shared $
  - DDR 5, next gen Optane support
  - PCIe 5.0
  - Embedded Silicon Bridge (EMIB), modular SoC on die fabric

IEEE-CNSV
Consultants' Network of Silicon Valley

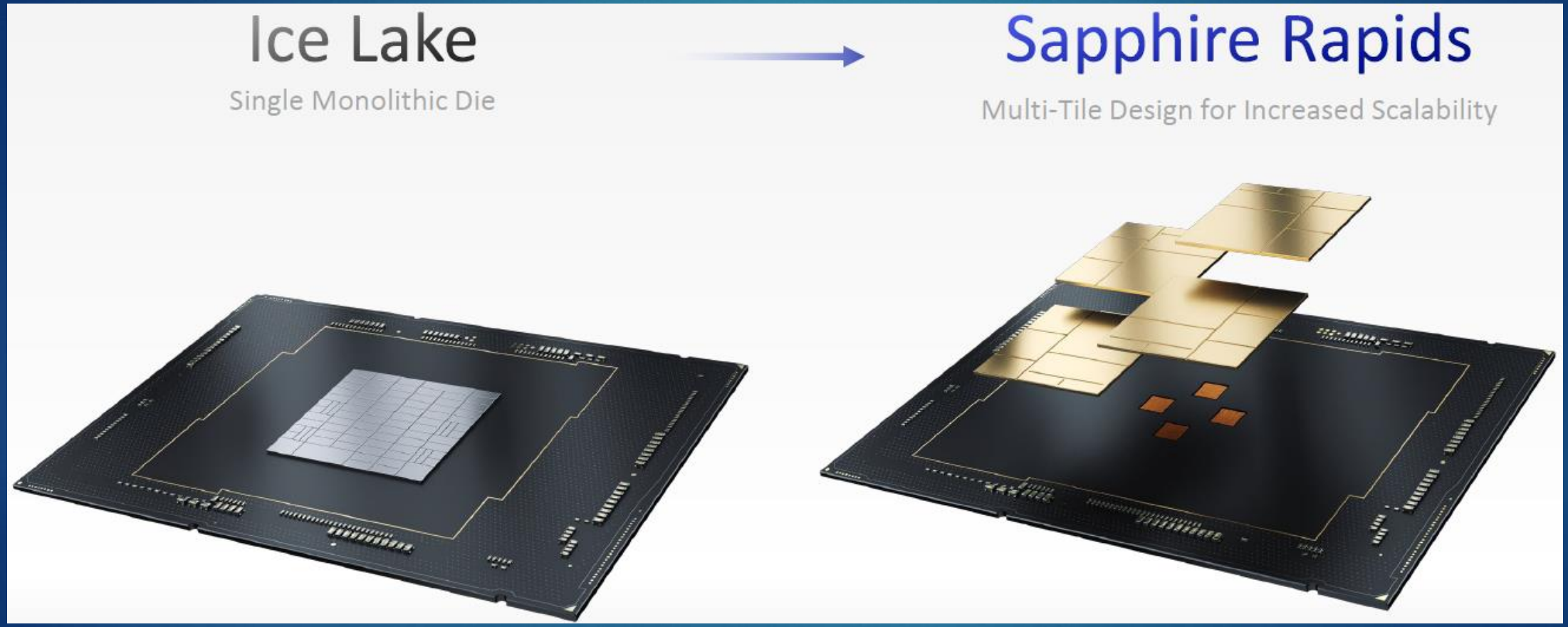
Intel (stacking/bonding/multi tech integration)

- Data center performance:
  - Faster VM migration, I/O virtualization (cloud folks)
  - Improved Security & RAS (HW helps SW)
  - More consistent $-ing and memory latency
  - Inter processor interrupt virtualization (many VMs)
  - More telemetry between dies (on chip stuff)

IEEE-CNSV
Consultants' Network of Silicon Valley

Day 2: Main conference (CPUs)

- Next-Gen Intel Xeon CPU – Sapphire Rapids



IEEE-CNSV
Consultants' Network of Silicon Valley

Leaving CPUs and CPU offloading: *sometimes called DPUs, IPUs, etc*

- Infrastructure & Data Processors:
  - Arm Neoverse N2: Arm's 2nd gen high performance infrastructure CPUs and system products
  - NVIDIA DATA Center Processing Unit (DPU) Architecture
  - Intel's Hyperscale-Ready SmartNIC for Infrastructure Processing

IEEE-CNSV
Consultants' Network of Silicon Valley

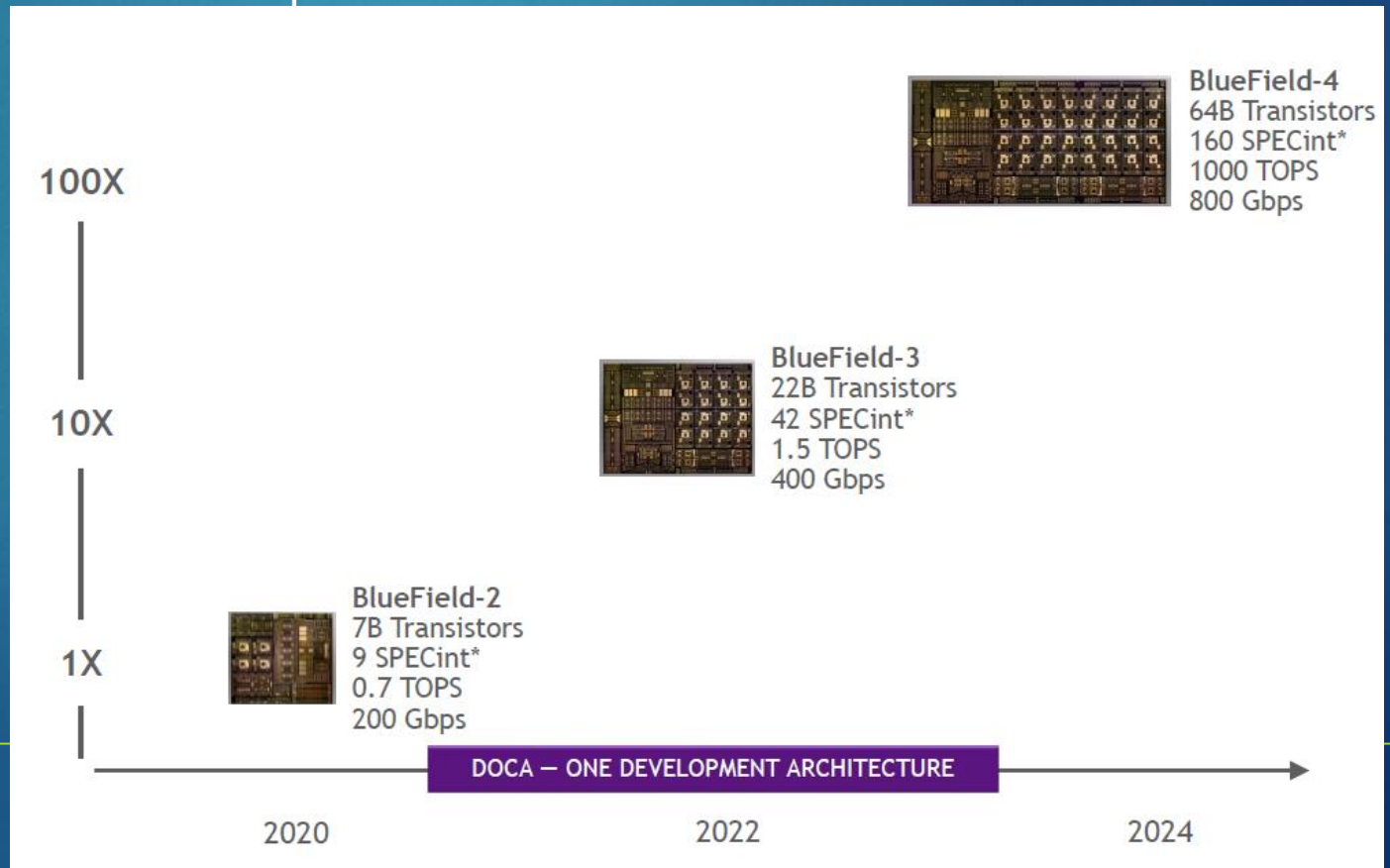
Leaving CPUs and CPU offloading: *sometimes called DPUs, IPUs, etc*

- Infrastructure & Data Processors:
  - Arm Neoverse N2: Arm's 2nd gen high performance infrastructure CPUs and system products
  - Cortex A710, ARM v9.0-A Instruction set
  - Internally Perseus platform , 5 nm, PCIe gen5, CXL 2
  - NVIDIA Grace CPU based on N2 – building the Alps Supercomputer 2023
  - N3 (late 2022 or 2023)
    - PCIE Gen6, CXL .....
    - Internally Poseidon platform, add ons are ML etc.

IEEE-CNSV
Consultants' Network of Silicon Valley

Leaving CPUs and CPU offloading: *sometimes called DPUs, IPUs, etc*

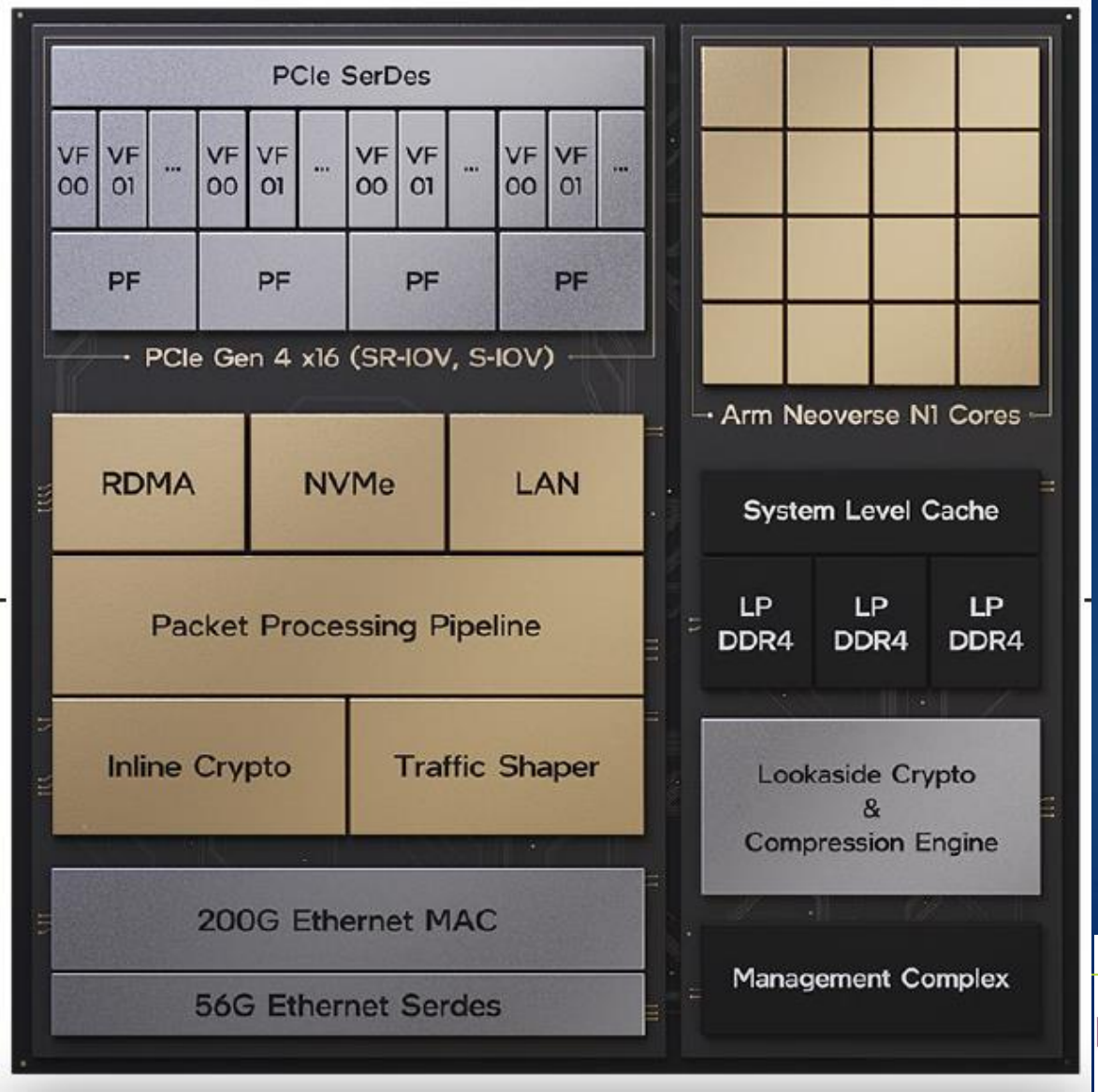
- Infrastructure & Data Processors:
  - NVIDIA DATA Center Processing Unit (DPU) Architecture
  - Courtesy of the Mellanox acquisition
  - Called BlueField



IEEE-CNSV
Consultants' Network of Silicon Valley

Leaving CPUs and CPU offloading: *sometimes called DPUs, IPUs, etc*

- Intel Hyperscale-Ready
- Mt Evans



IEEE-CNSV
Consultants' Network of Silicon Valley

Leaving CPUs and CPU offloading: *sometimes called DPUs, IPUs, etc*

- Enabling Technologies:
  - Heterogeneous computing to enable the highest level of safety in automotive systems, *Infineon*
  - Architecting an Open RISC-V 5G and AI SoC for Next Generation 5G Open Radio Access Network, *EdgeQ*
  - Aquabolt-XL: Samsung HBM2-PIM with in-memory processing for machine learning accelerators, *Samsung*

IEEE-CNSV
Consultants' Network of Silicon Valley

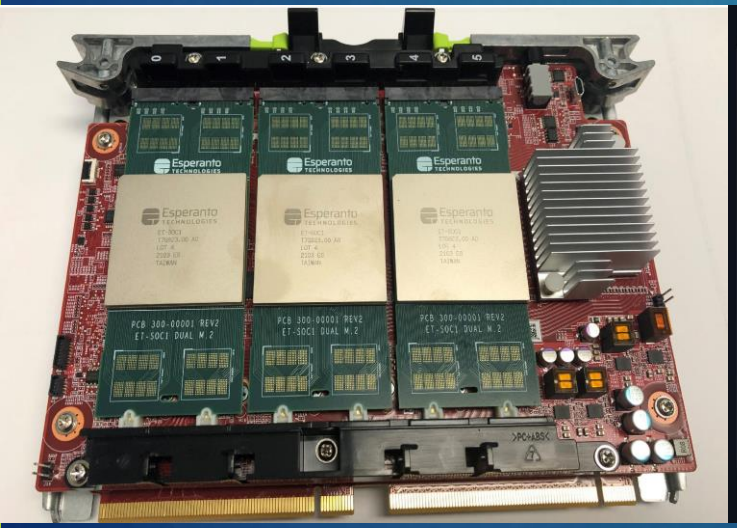
AI and the chip trends (*2)

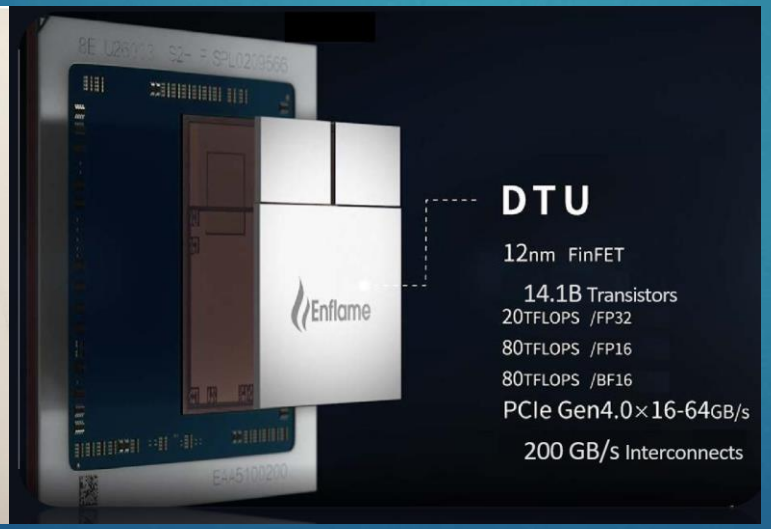
- ML Inference for the Cloud
- ML chips as Computational Platforms
- Graphics & Video
- New Technologies

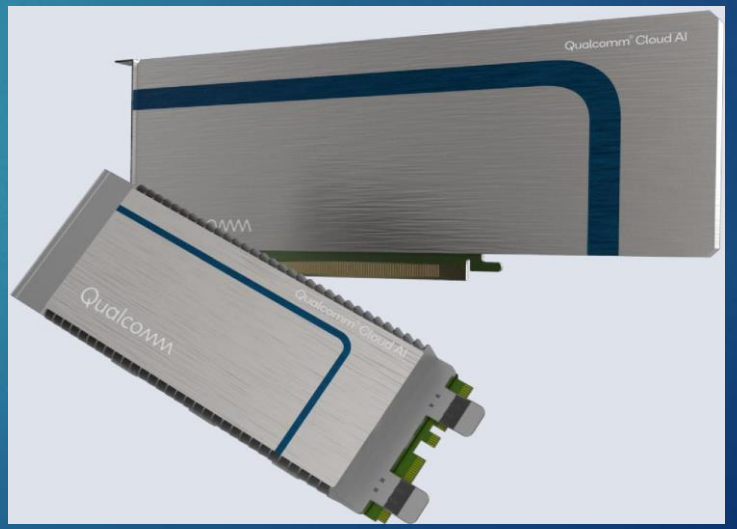
IEEE-CNSV
Consultants' Network of Silicon Valley

AI/ML chips for the cloud

- Accelerating ML Recommendation with over a Thousand RISC-V/Tensor Processors on *Esperanto Technologies*'s ET-SoC-1 Chip
- AI Compute Chip from *Enflame Technology*
- *Qualcomm* Cloud AI 100: 12 TOPs/W Scalable, High Performance and Low Latency Deep Learning Inference Accelerator







IEEE-CNSV
Consultants' Network of Silicon Valley

AI chips for the "edge" of the cloud……..

- Xilinx 7 nm Edge Processors

Versal™ AI Edge: Architecture Overview

XILINX VERSAL™ | AI CORE

XILINX VERSAL™ | PREMIUM

XILINX VERSAL™ | AI EDGE

ADAPTABLE TO MULTIPLE WORKLOADS

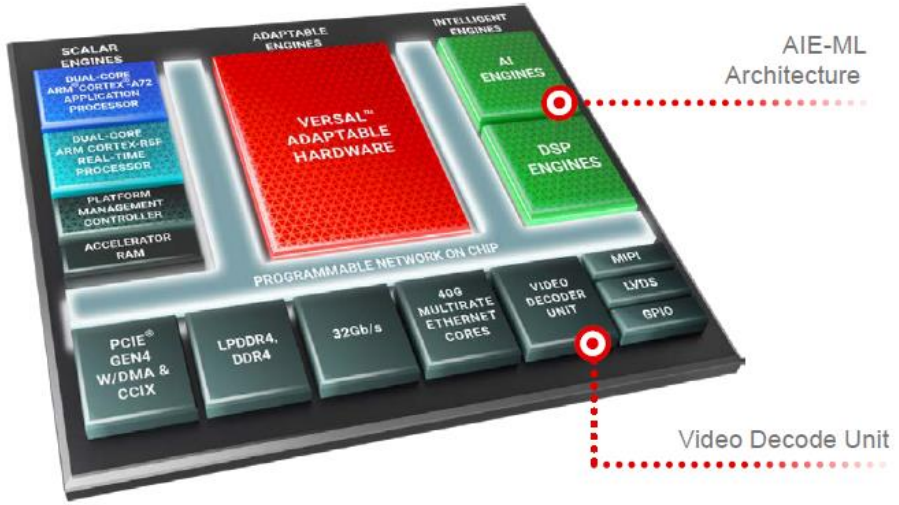
- Versal AI Core
- Versal Premium
- Versal AI Edge

COMPUTE ACCELERATION

- Scalar Engines
- Adaptable Engines
- Intelligent Engines

PLATFORM

- Software programmable NoC
- Platform Management Controller
- Dedicated Interfaces (e.g., PCIe, DDR)



Technology Node: TSMC 7nm FinFET

tsmc 7nm

XILINX

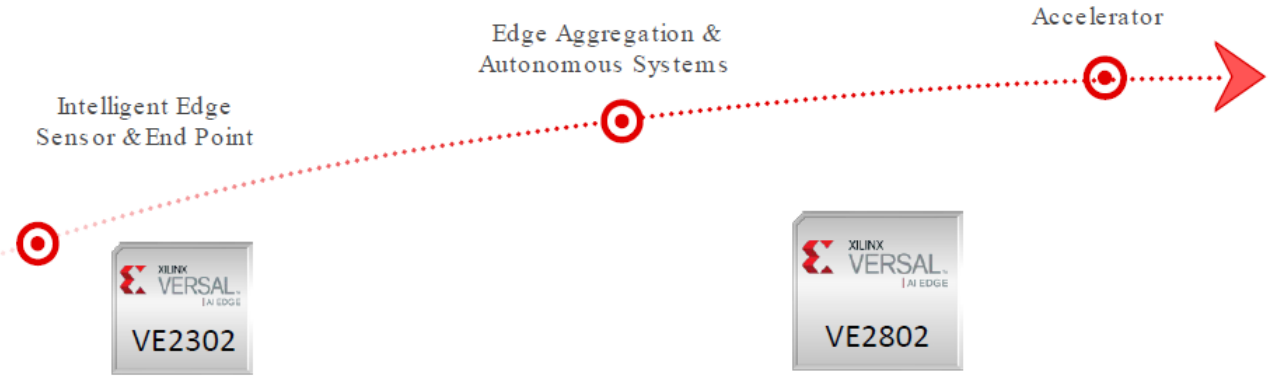
© Copyright 2021 Xilinx

IEEE-CNSV
Consultants' Network of Silicon Valley

AI chips for the "edge" of the cloud……..

- Xilinx 7 nm Edge Processors

Optimized Devices for AI/ML Edge Inference



		VE2302	VE2802
Engines	AI Compute (INT8x4)[1]	67 TOPS	479 TOPS
	AI Compute (INT8)[1]	31 TOPS	228 TOPS
	AIE-ML Tiles	34	304
	Adaptable Engines	150K LUTs	521K LUTs
	Processing Subsystem	Dual-Core Arm® Cortex®-A72 Application Processing Unit / Dual-Core Arm Cortex-R5F Real-Time Processing Unit	
RAM	Accelerator RAM (4MB)	✓	-
	Total Memory	172Mb	575Mb
	32G Transceivers	8	32
	PCIe®	✓	✓ (PCIe gen5 w/ DMA)
	Video Decode Unit (VDU)	-	✓
	Power [2]	15-20W	75W

1: Total AI compute includes AI Engines, DSP Engines, and Adaptable Engines
2: Power Projections

© Copyright 2021 Xilinx
XILINX.

IEEE-CNSV
Consultants' Network of Silicon Valley

Special chips geared just to do AI/ML!

- ML and Computation Platforms
  - Graphcore Colossus Mk2 IPU,
  - *Cerebras,* Multi-Million Core, Gigantic Chip
  - *SambaNova*, SN10 RDU

IEEE-CNSV
Consultants' Network of Silicon Valley

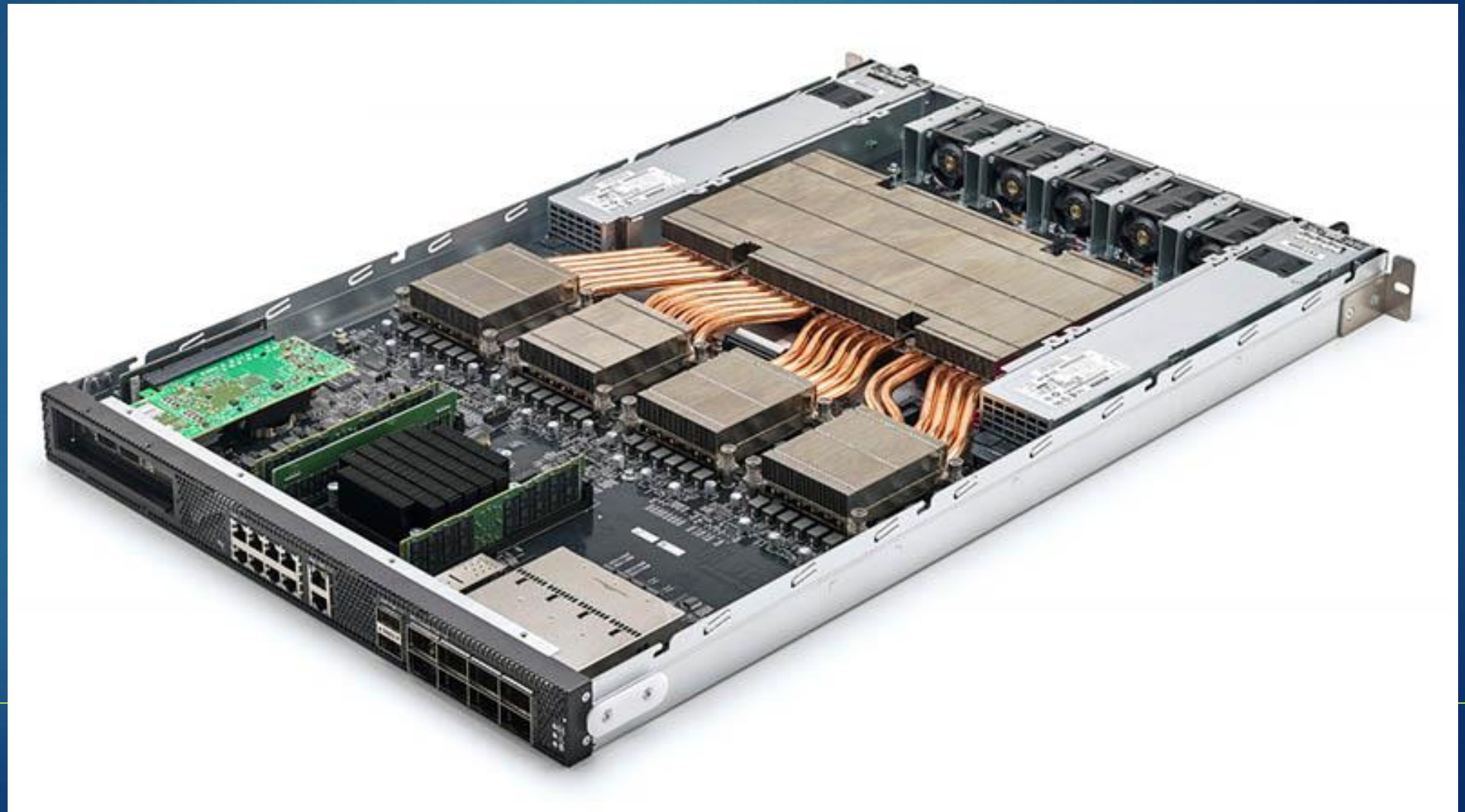
Special chips geared just to do AI/ML!

- *Graphcore* Colossus Mk2 IPU
- 59,334,610,787 active transistors in TSMC N7
- 1472 processor tiles @ 624KiB
- Total 250Tflop/s + 896MiB SRAM
- 62TB/s memory, 7.8TB/s inter-tile, 320GB/s inter-chip
-
- Plans for using TSMC N3 (circa ?)

IEEE-CNSV
Consultants' Network of Silicon Valley

Special chips geared just to do AI/ML!

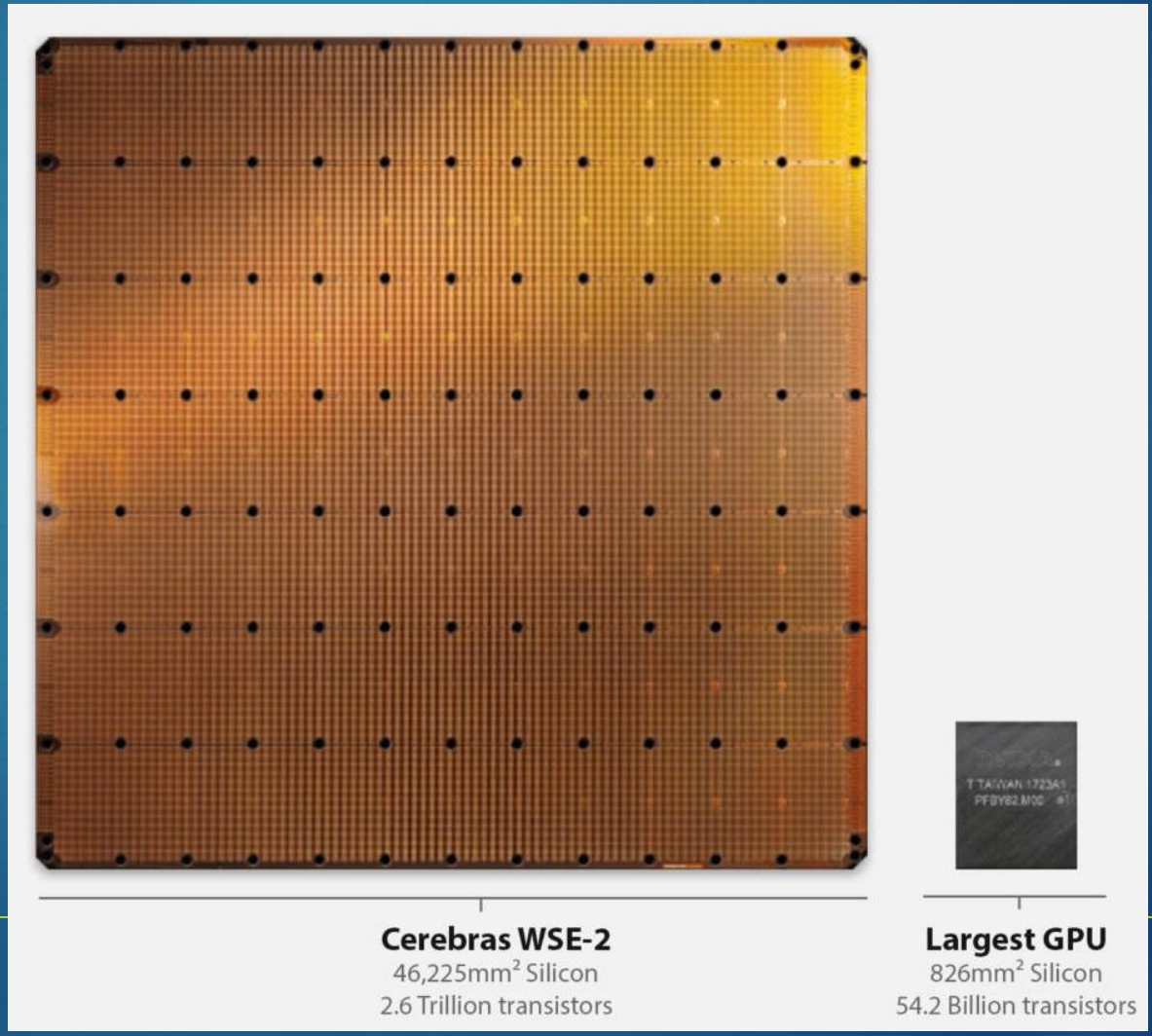
- *Graphcore* Colossus Mk2 IPU



IEEE-CNSV
Consultants' Network of Silicon Valley

Special chips geared just to do AI/ML!

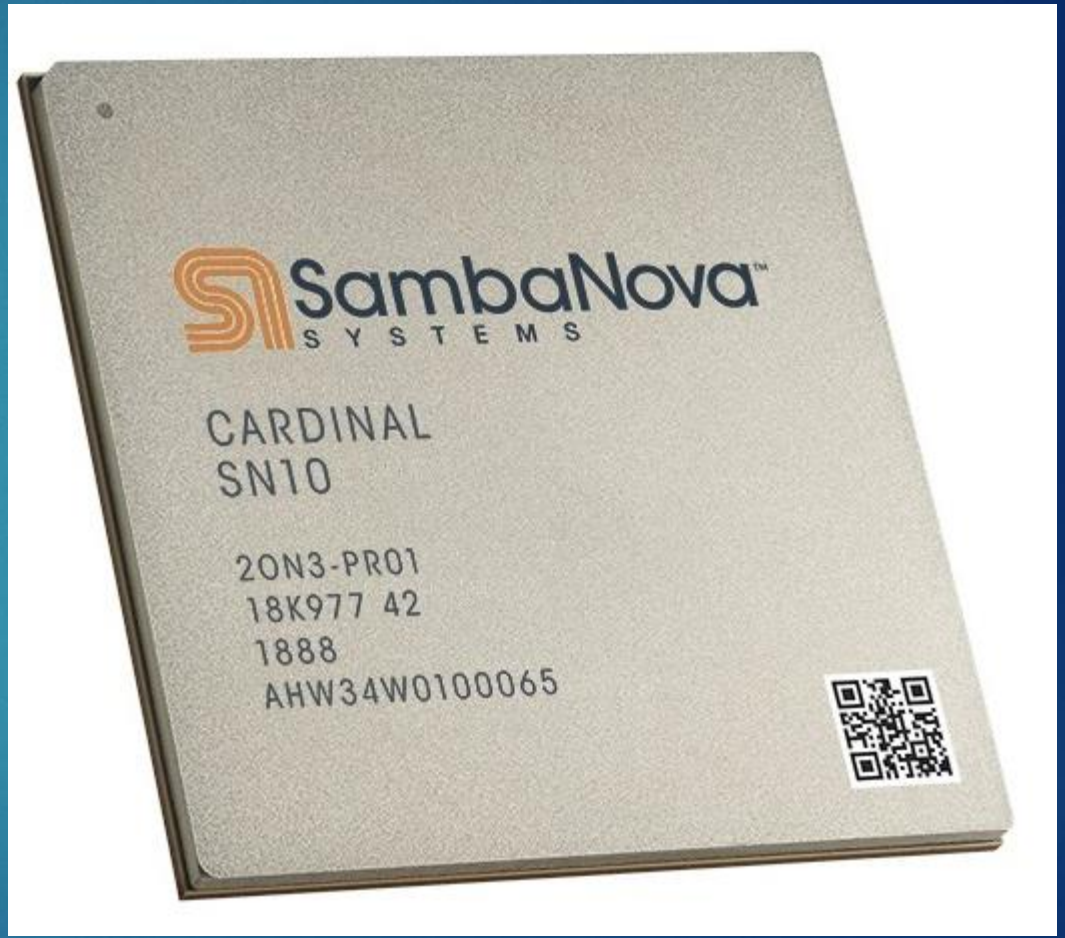
- *Cerebras,* Multi-Million Core, Gigantic Chip



IEEE-CNSV
Consultants' Network of Silicon Valley

Special chips geared just to do AI/ML!

- *SambaNova*, SN10 RDU
- *TSMC 7nm*
- 40B transistors, 50 Km of wire
- 640 Pattern Compute Units
- 640 Pattern Memory Units
- >300 MB on-chip memory
- 150 TB/s on-chip memory bandwidth
- Memory transformation operations



IEEE-CNSV
Consultants' Network
of Silicon Valley

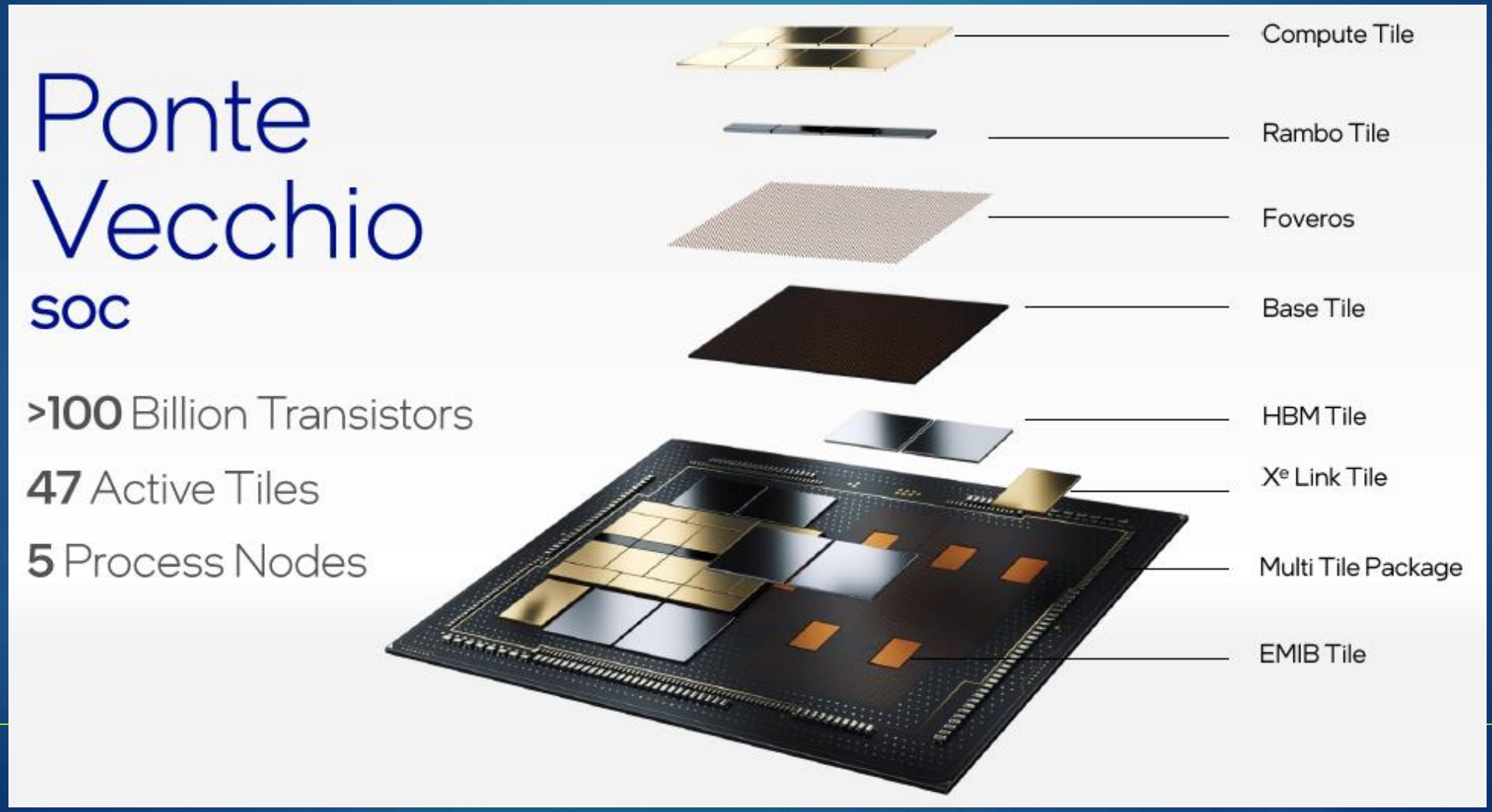
Chips just to do Graphics/Video…….

- Intel's Ponte Vecchio GPU Architecture, *Intel*
- AMD RDNA™ 2 Graphics architecture, *AMD*
- Google Video Coding Unit (VCU), *Google*

IEEE-CNSV
Consultants' Network of Silicon Valley

Chips just to do Graphics/Video…….

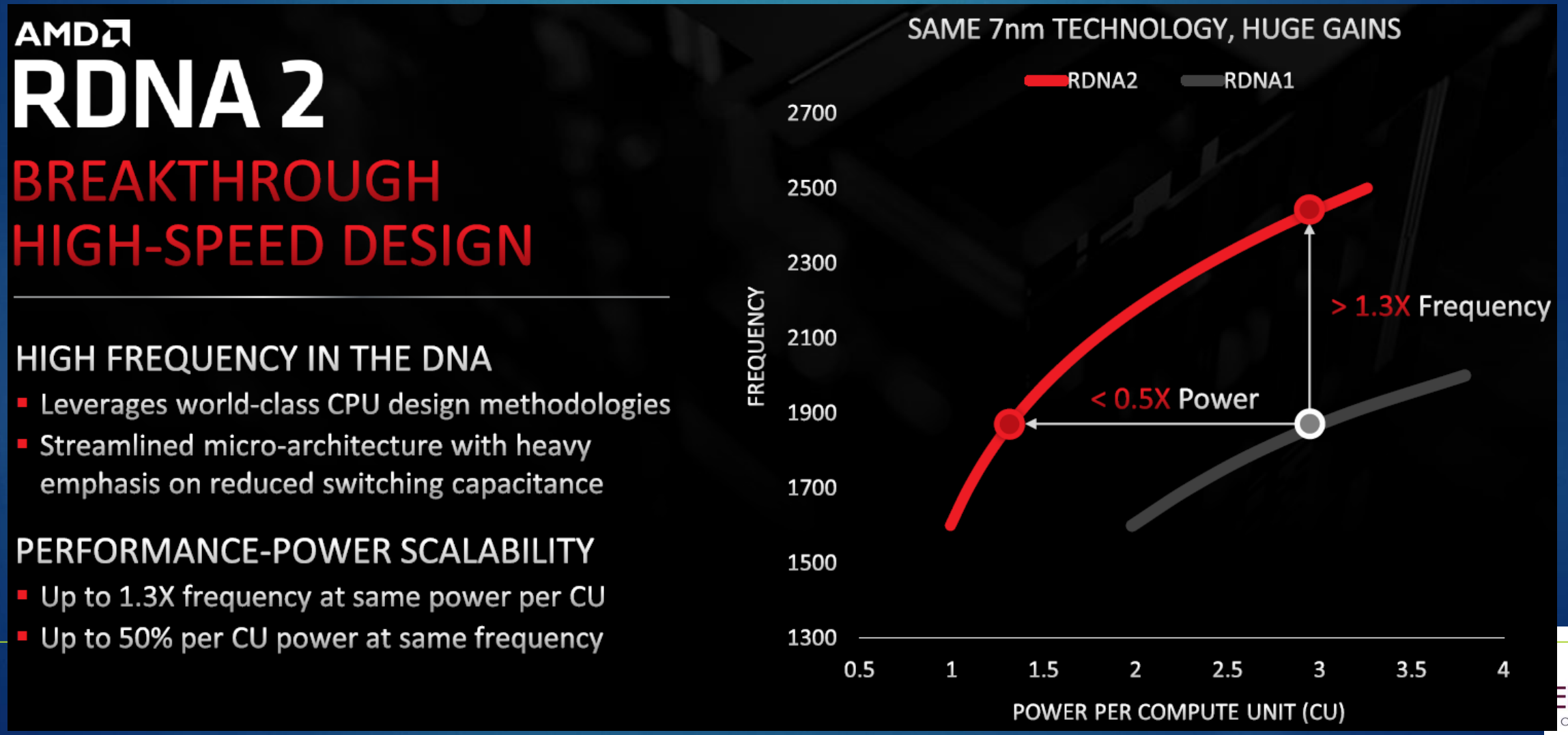
- Intel's Ponte Vecchio GPU Architecture



IEEE-CNSV
Consultants' Network of Silicon Valley

Chips just to do Graphics/Video…….

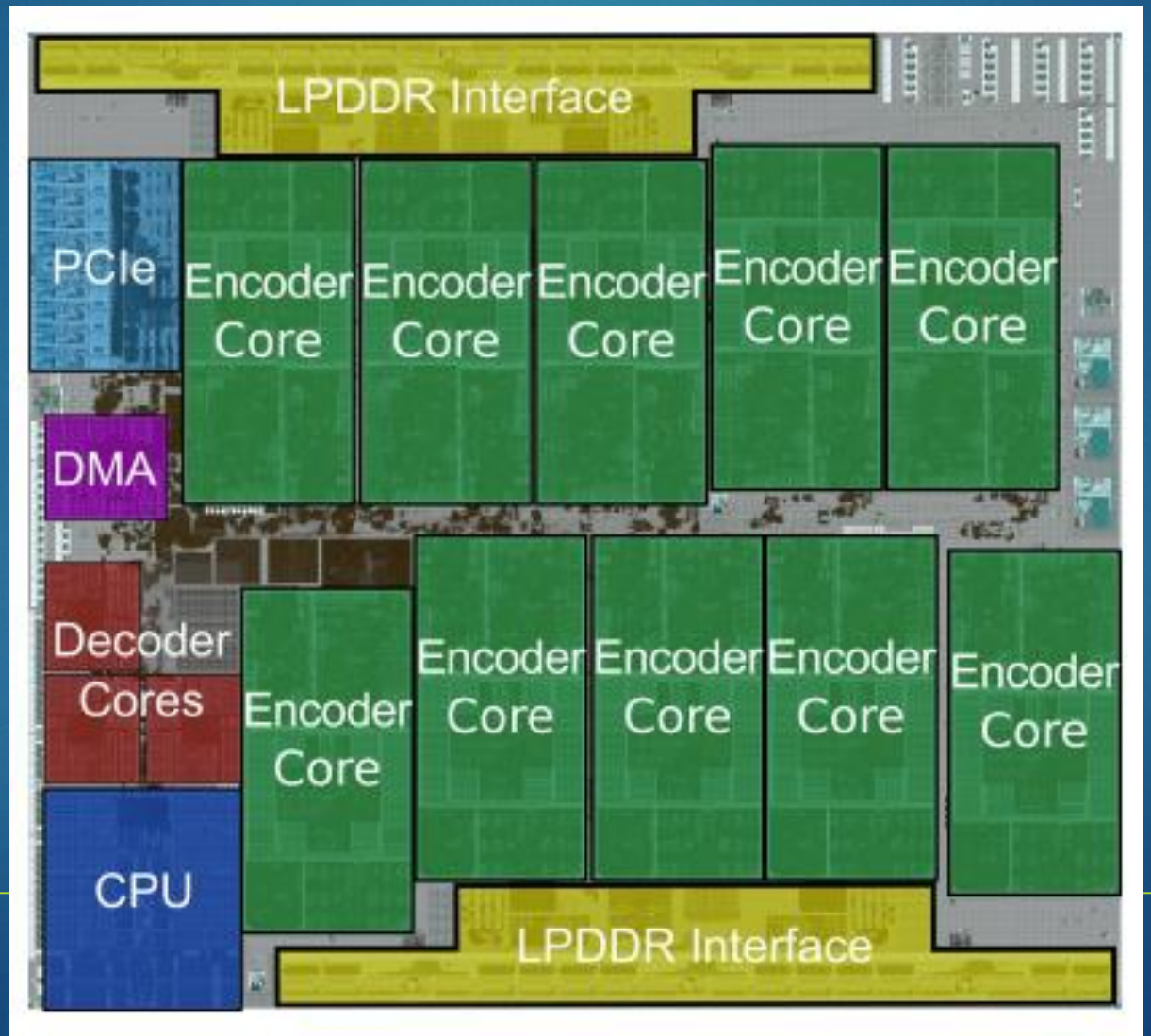
- AMD RDNA™ 2 Graphics architecture



EEE-CNSV
Consultants' Network of Silicon Valley

Chips just to do Graphics/Video…….

- Google Video Coding Unit (VCU)



IEEE-CNSV
Consultants' Network of Silicon Valley

Chips just to do Graphics/Video…….

- Google Video Coding Unit (VCU)

IEEE-CNSV
Consultants' Network of Silicon Valley

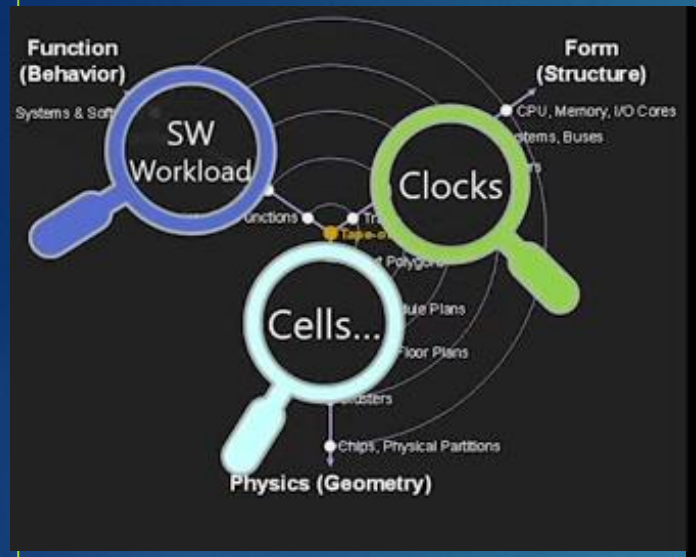
What about using AI for chip design itself?

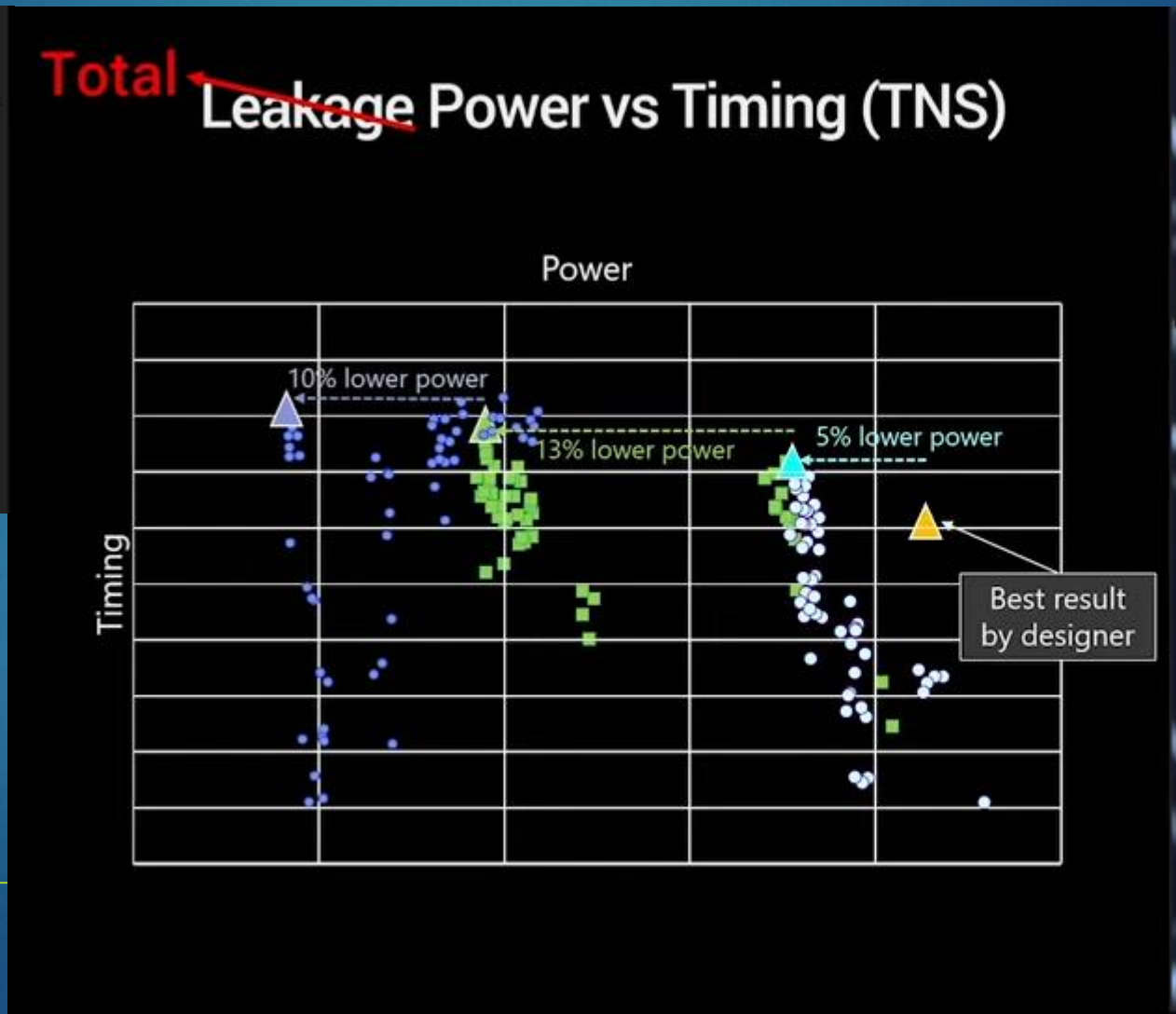
- Most exciting prospect of modern day chip architecture trends!!
- *Using ML on chip floorplan designs*
- *Optimize for power*
- *Optimize for cell layouts*
- *Optimize for workflows*
- **Aart de Geus, CEO, Synopsys**

IEEE-CNSV
Consultants' Network of Silicon Valley

What about using AI for chip design itself?

- Most exciting prospect of modern day chip architecture trends!!





Observations: AI chips for highly mobile use

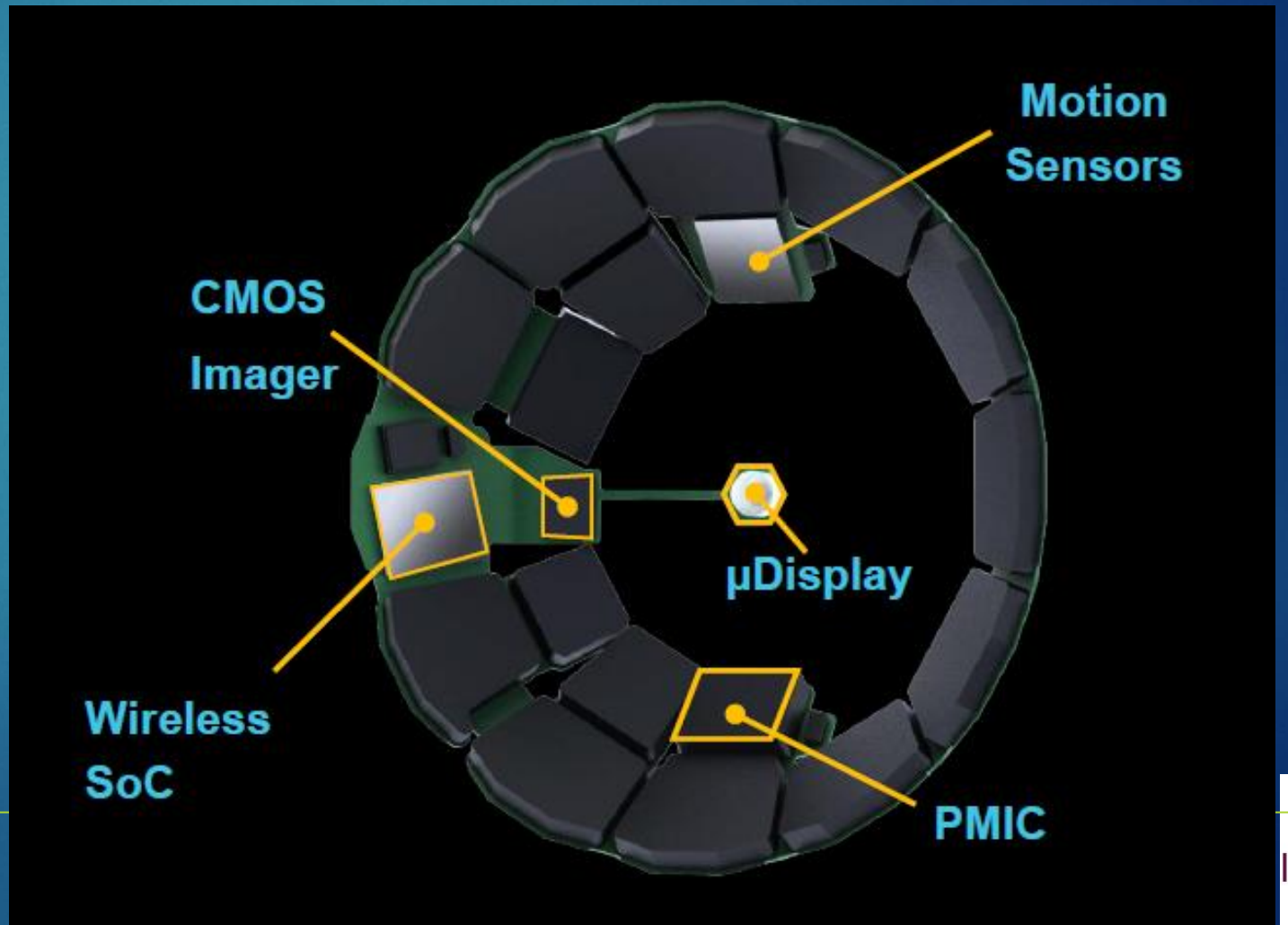
- Skydio Autonomy Engine: Enabling the Next Generation of Autonomous Flight, ***Abraham Bachrach, CTO, Skydio***

COMPUTE RESOURCE MAPPING

	CPU	ISP	GPU	DSP	NPU
HIGH LEVEL LOGIC	✓				
LOW LEVEL IMAGE PROCESSING		✓	✓	✓	
GEOMETRIC COMPUTATIONS			✓	✓	
DEEP LEARNING			✓	✓	✓
NONLINEAR OPTIMIZATION	✓				

Observations: AI chips for highly mobile use

- Mojo Vision
- ~0.5 mm μ display



IEEE-CNSV
Consultants' Network
of Silicon Valley

Thank You!

SHARAN.KALWANI@IEEE.ORG

LEAVE FEEDBACK AT:

LinkedIn ~ https://www.linkedin.com/in/sharankalwani

IEEE-CNSV
Consultants' Network of Silicon Valley

Future happenings

- Lots of CXL offerings for many different memories/devices, etc.
- SW for running on ML platforms, AI/ML circuit design
- I suppose more products from NVIDIA, AMD,
- Innovative packaging (TSMC, Samsung, etc)
- Many research projects slipping into the commercial attempts
- More assistive chips (DPU, IPUS, SmartNICs, etc)
- More AI chips for Edge Computing, etc.

IEEE-CNSV
Consultants' Network of Silicon Valley

Backup slides

IEEE-CNSV
Consultants' Network of Silicon Valley

Intel vs TSMC comparison

	Intel 4	Intel 7	TSMC N5	TSMC N3
HP Library Density	160 MTr/mm^2 (est.)	80 MTr/mm^2	130 MTr/mm^2 (est.)	208 MTr/mm^2 (est.)

IEEE-CNSV
Consultants' Network of Silicon Valley