

# Emerging Technologies and Silicon for HPC

**(Not all Cores are Equal)**

EasyBuild User Meeting, April 2023

Ian Cutress  
Chief Analyst, More Than Moore

# Silicon or Survive

- The New HPC Era
  - Types of Legacy Hardware: CPU, GPU, FPGA, ASIC
  - New Paradigms: Analog, Neuro, Quantum, Optical
  - The Push For Low Precision and Scale: AI Hardware
- AI Hardware
  - Established Players: NVIDIA GPU, Intel CPU
  - Startup Funding: A \$10B+ investment
  - Case Studies
  - Roadmaps
  - Software stacks - OneAPI, ROCm, vendor specific ones
- Q&A

# Ian Cutress

**More  
Than  
Moore.**



- Chief Analyst and Founder, More Than Moore
- Online Influencer and Educator, TechTechPotato



- Senior CPU Editor, AnandTech.com (2011-2022)
- PhD, Computational Chemistry, Oxford (2011)
- MChem, Computational Chemistry, Hull (2008)

# Silicon or Survive

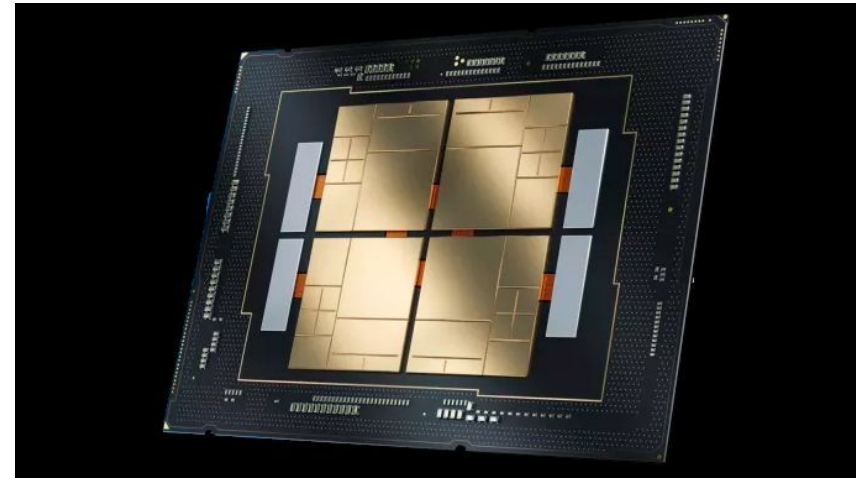
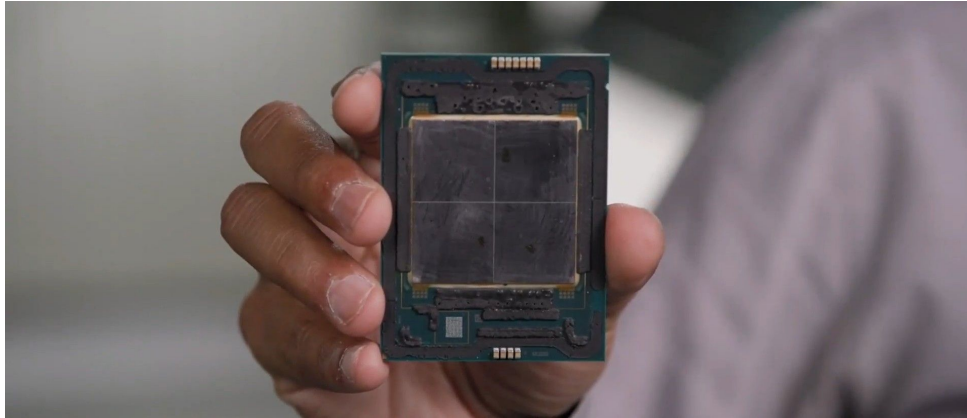
- A New HPC Era
  - Types of Legacy Hardware
  - New Paradigms
  - Push for Low Precision
- AI Hardware
  - Established Players
  - Current \$10B Market
  - Case Studies
  - Roadmaps
  - Software
- Q&A



# Silicon or Survive

- A New HPC Era
  - Types of Legacy Hardware
- CPU: x86, Arm, POWER
- GPU: NVIDIA, AMD
- FPGA: Intel (Altera), AMD (Xilinx)
- ASIC: Offload

# Intel CPU

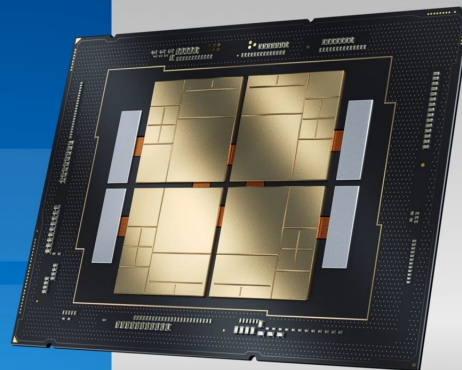


Super-Charged CPU

## Sapphire Rapids HBM

Sampling Today  
Production 2H '22

Up to 2.8x Perf  
over 3<sup>rd</sup> Gen Xeon<sup>1</sup>



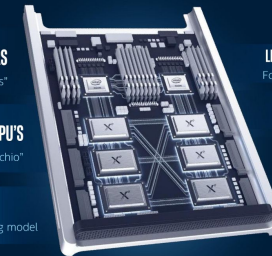
## Aurora: Bringing It All Together

**2** INTEL XEON SCALABLE PROCESSORS  
"Sapphire Rapids"

**6** XE ARCHITECTURE BASED GPU'S  
"Ponte Vecchio"

ONEAPI

Unified programming model



LEADERSHIP PERFORMANCE

For HPC, data analytics, AI

UNIFIED MEMORY ARCHITECTURE

Across CPU & GPU

ALL-TO-ALL CONNECTIVITY WITHIN NODE

Low latency, high bandwidth

UNPARALLELED I/O SCALABILITY ACROSS NODES

8 fabric endpoints per node, DAOS

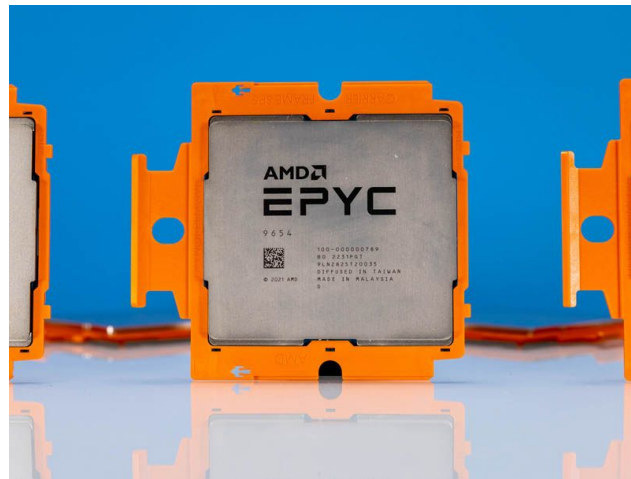
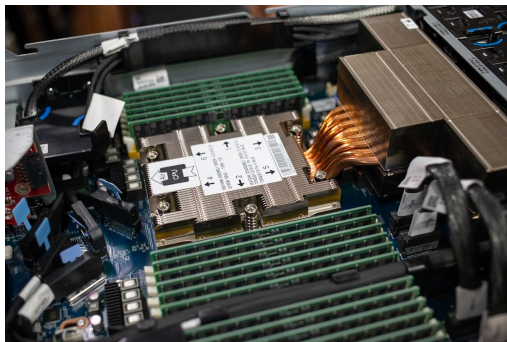
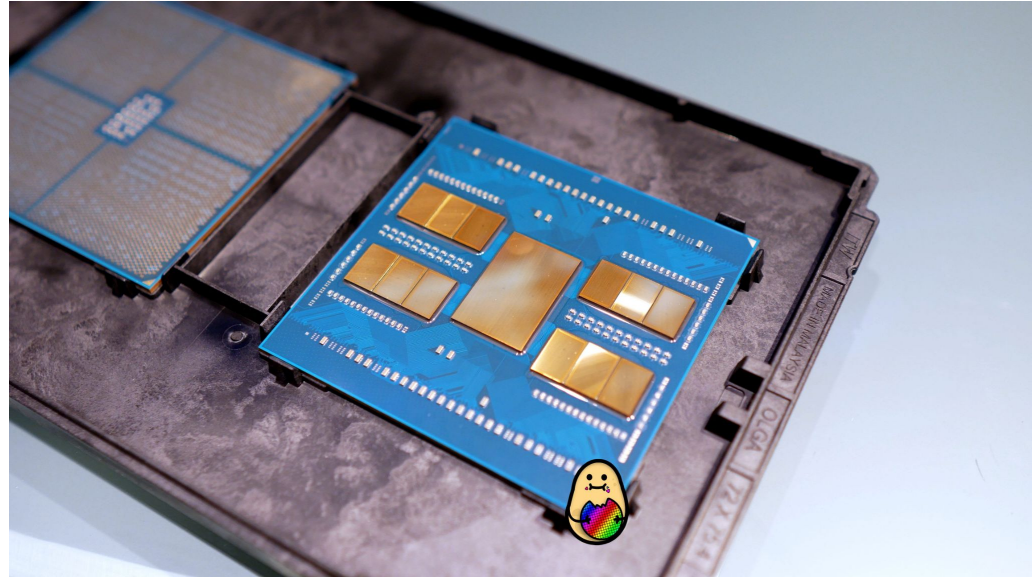
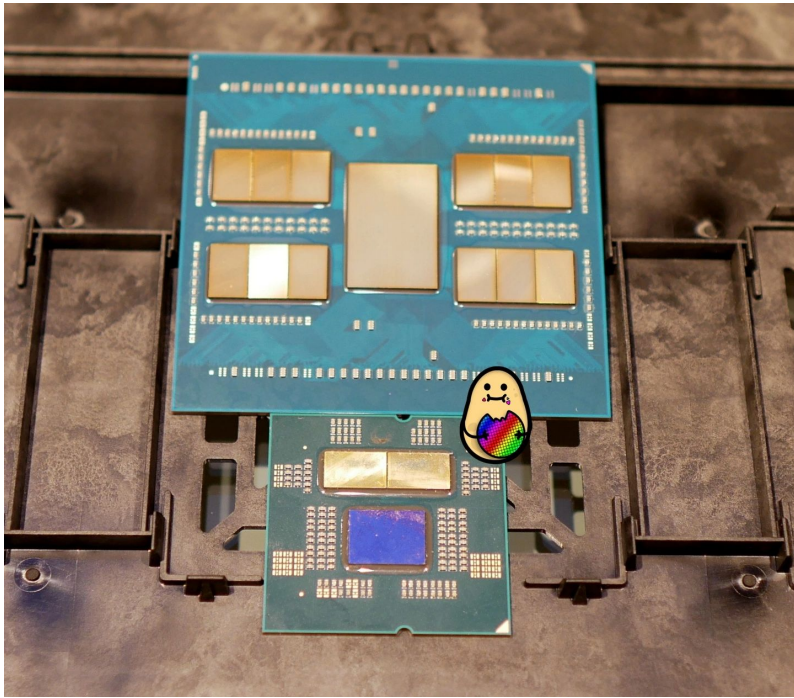
DELIVERED IN 2021



News Under Embargo: November 17, 2019 - 4:00 p.m. Pacific Time

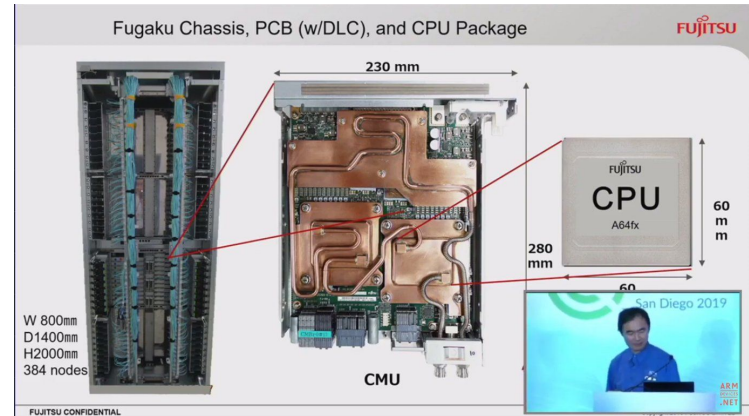
<sup>1</sup>Based on pre-production measurements. Learn more at [www.intel.com/PerformanceIndex](http://www.intel.com/PerformanceIndex). Results may vary.

# AMD Genoa



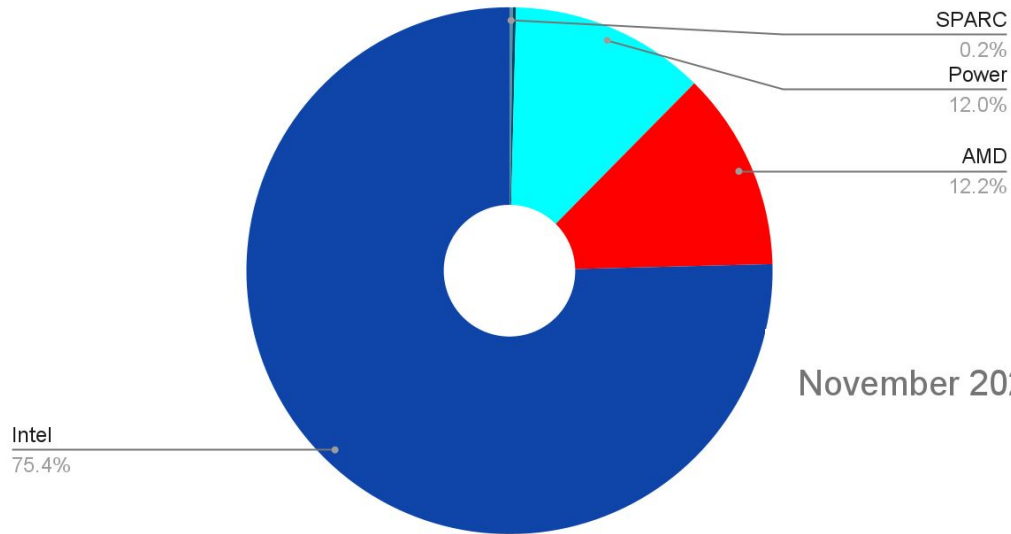


# Fujitsu A64FX (Arm) in Fugaku

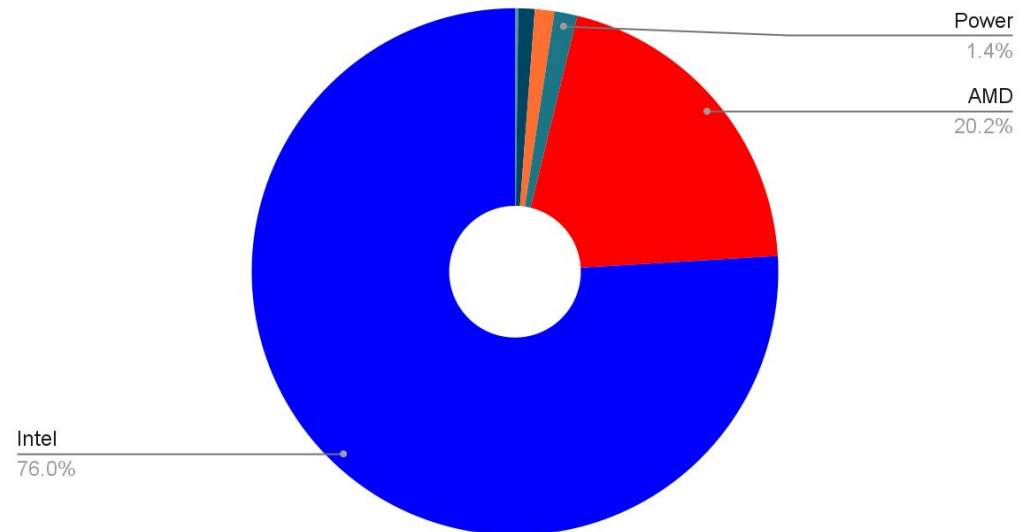


# TOP 500 CPU

November 2008



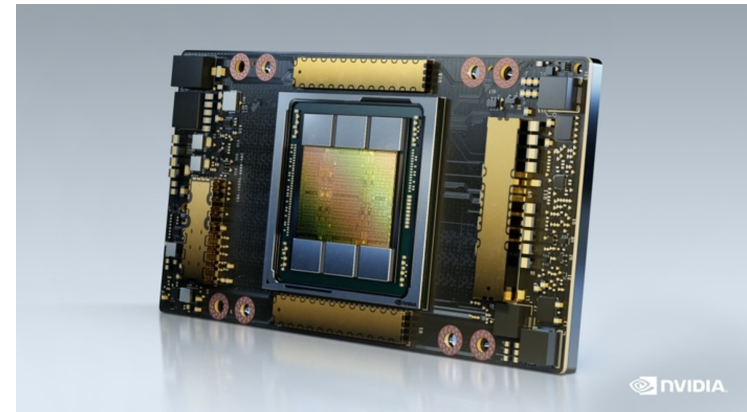
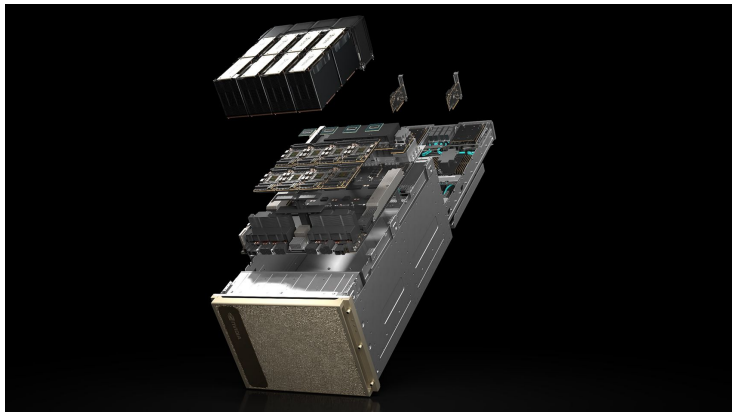
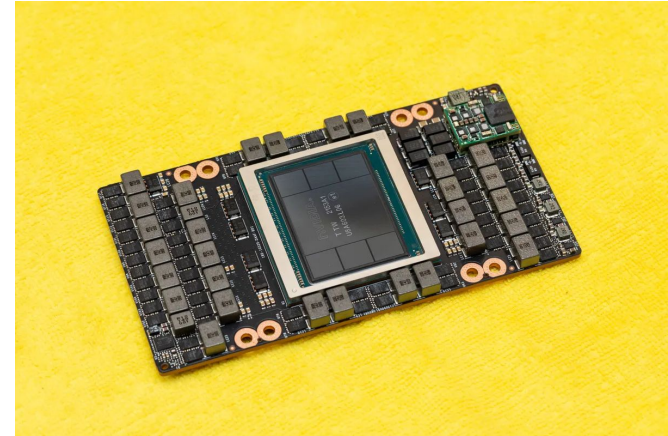
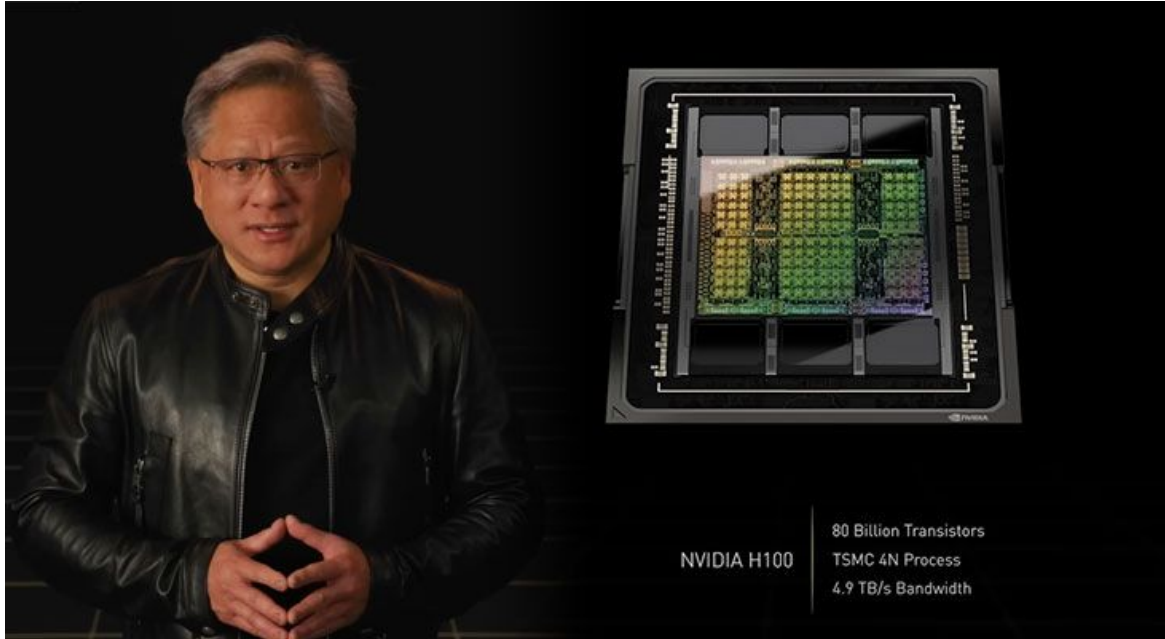
November 2022



# Silicon or Survive

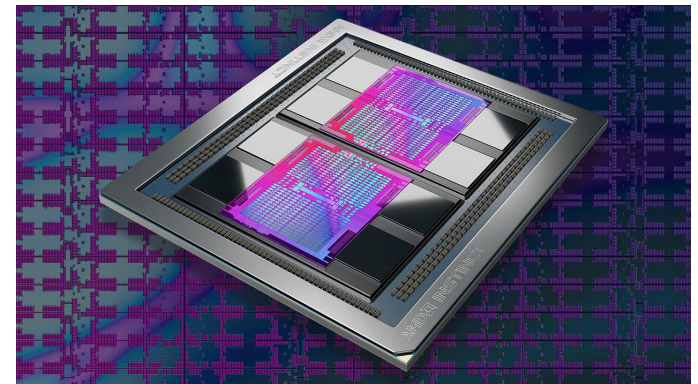
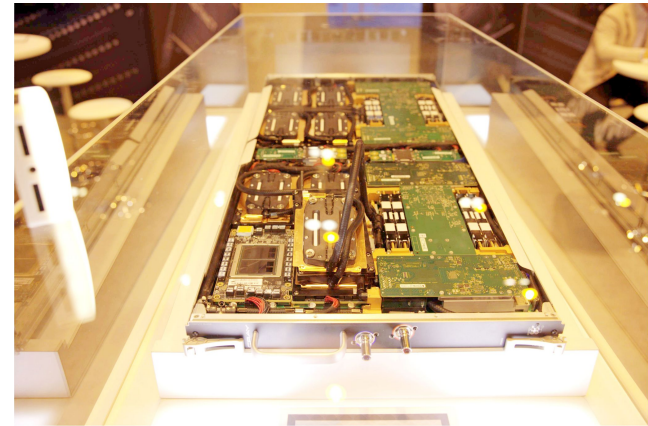
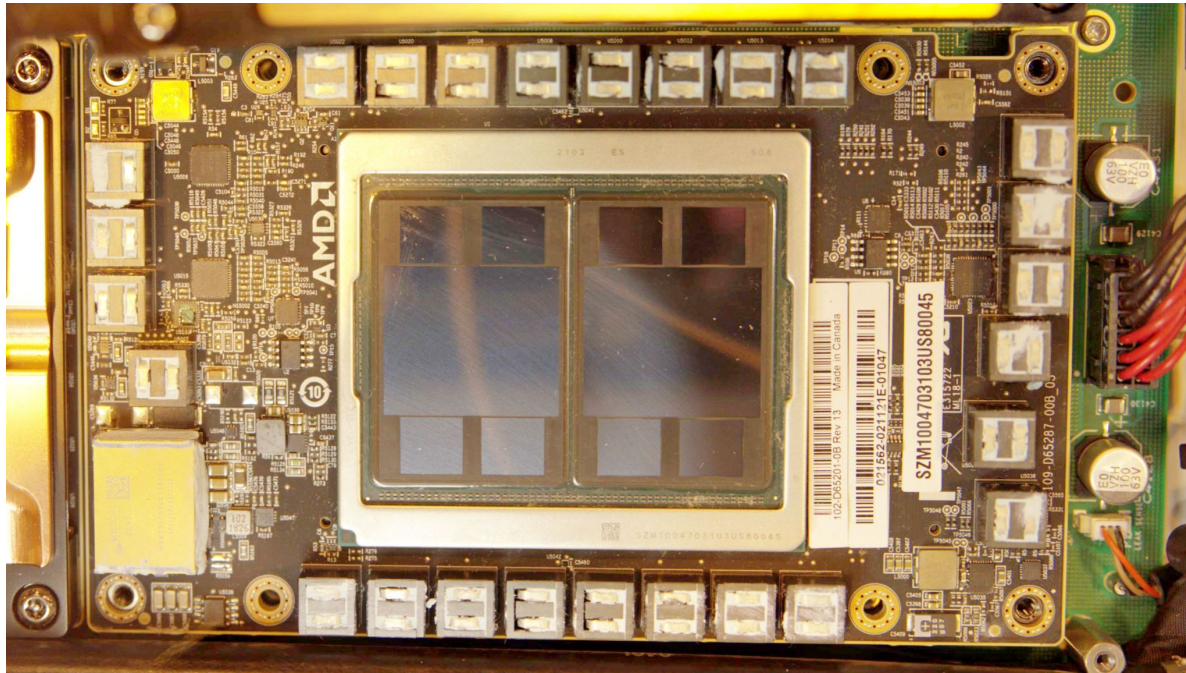
- A New HPC Era
  - Types of Legacy Hardware
- CPU: x86, Arm, POWER
- GPU: NVIDIA, AMD, Intel(?)
- FPGA: Intel (Altera), AMD (Xilinx)
- ASIC: Offload

# NVIDIA H100



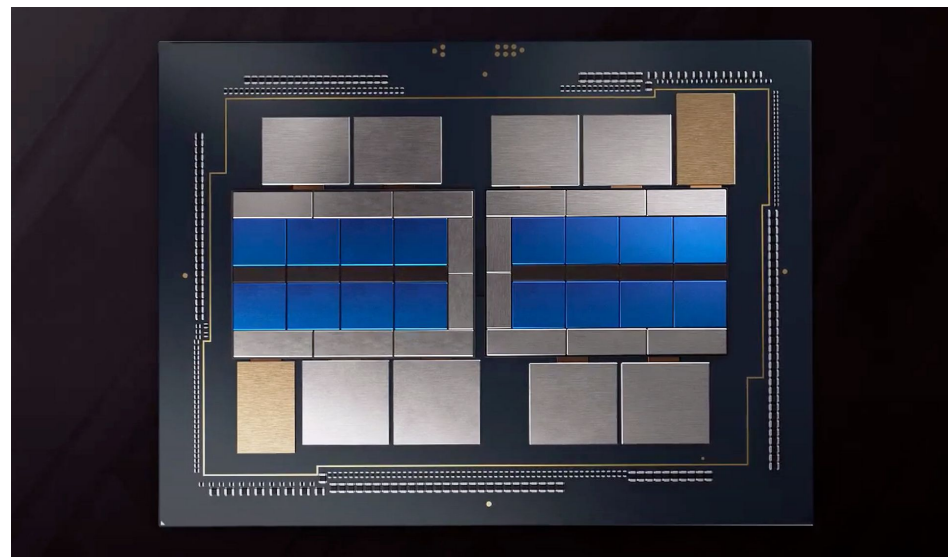
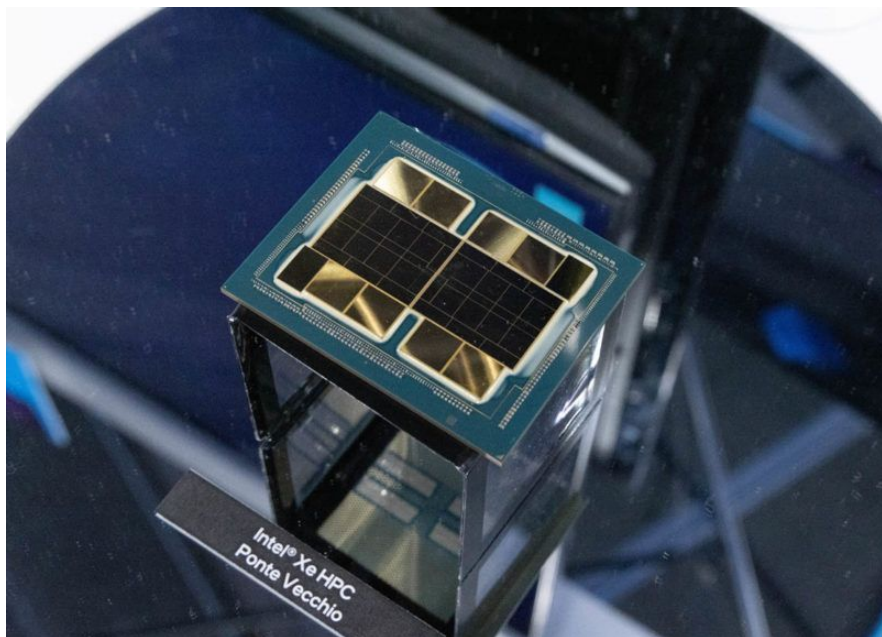


# AMD MI250X





# Intel Ponte Vecchio



Compute	Up to <b>128</b> Ray tracing Units	Highest Compute Density socket & node	<b>128 Xe Cores</b>
Memory	Up to <b>64MB</b> L1 cache in 2 Stacks	Up to <b>408MB</b> L2 Cache in 2 Stacks	<b>HBM2e</b>
I/O	Up to <b>8</b> Fully Connected GPUs	<b>PCIe Gen 5</b>	<b>Xe Link</b> High-Speed Coherent Unified Fabric
Technology	<b>EMIB</b>	<b>Foveros</b>	Intel 7 TSMC N5 TSMC N7

**Ponte Vecchio**  
Xe HPC based GPU

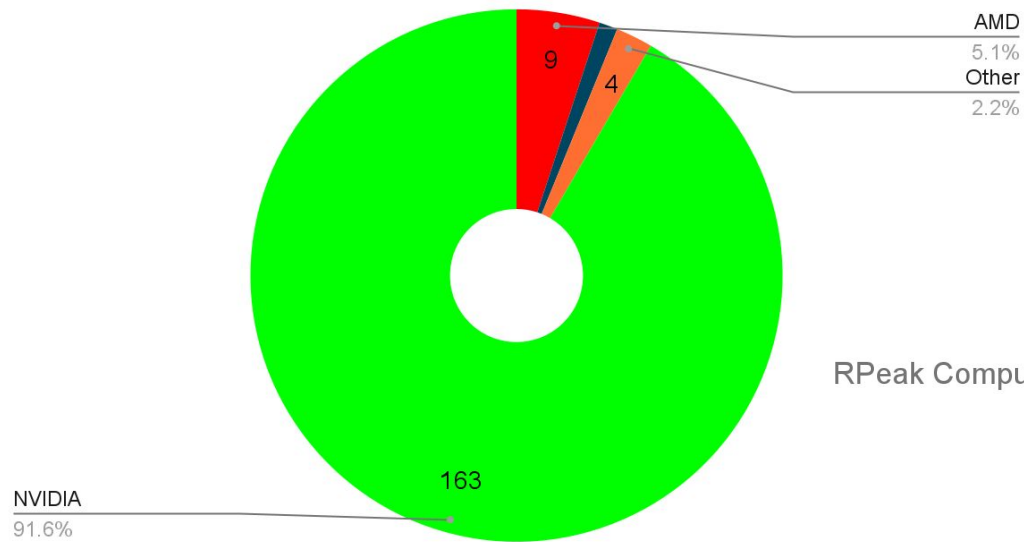
**Ponte Vecchio x4 Subsystem with Xe Links**  
+ 2S Sapphire Rapids

**Ponte Vecchio x4 Subsystem with Xe Links**

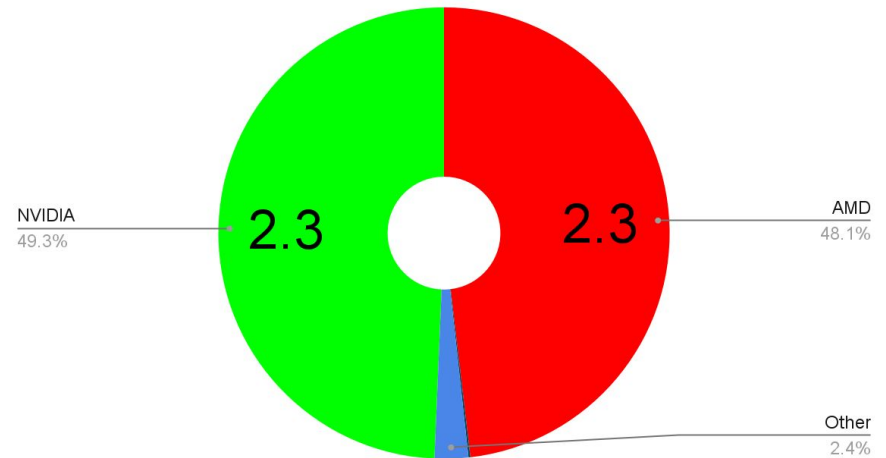
**Ponte Vecchio OAM**

# TOP 500 Accelerators

Systems

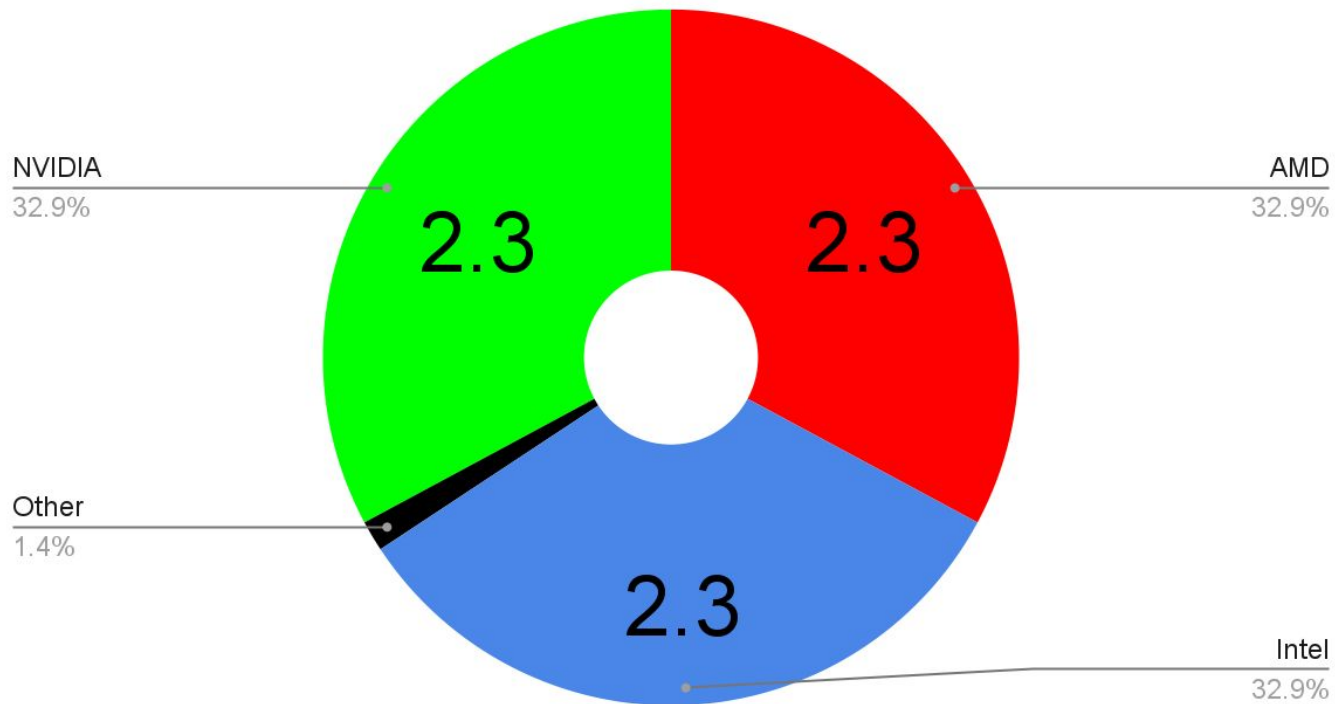


RPeak Compute in ExaFLOPs



# TOP 500 Accelerators + Aurora

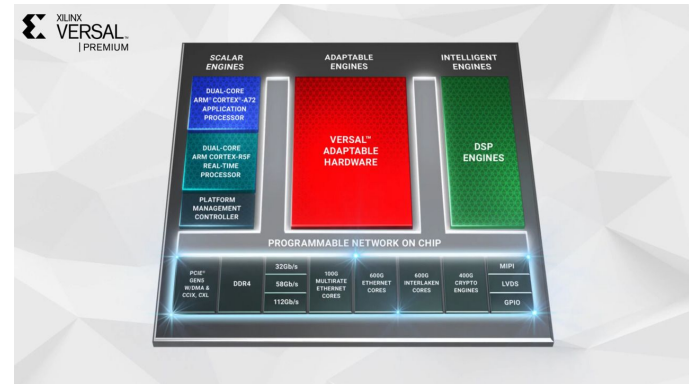
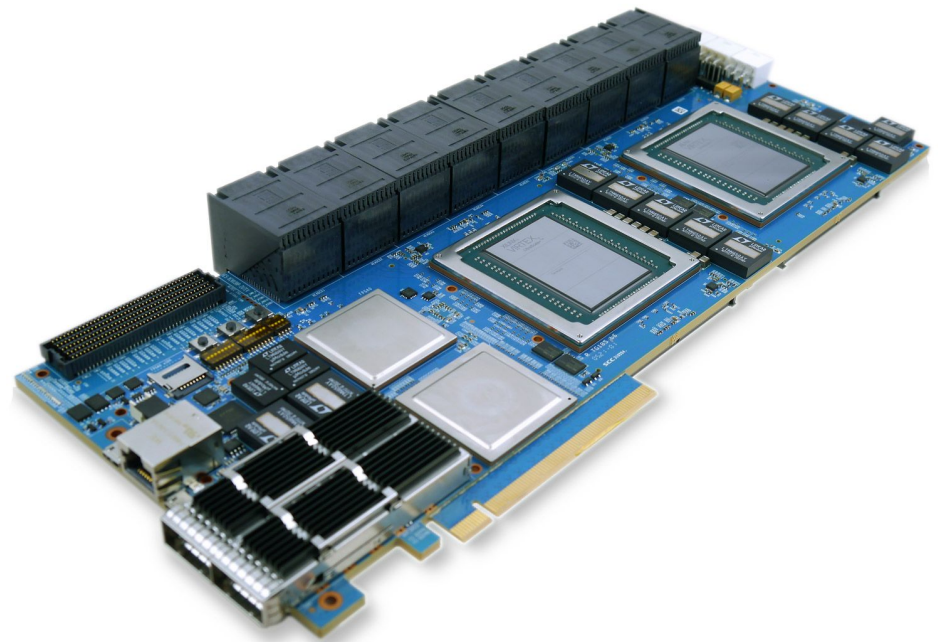
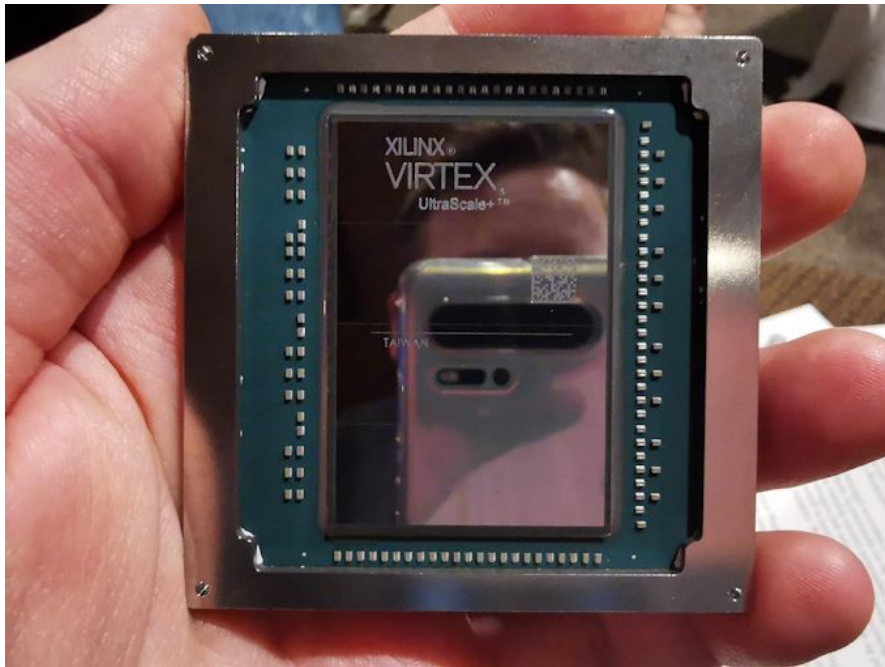
RPeak Compute in ExaFLOPs



# Silicon or Survive

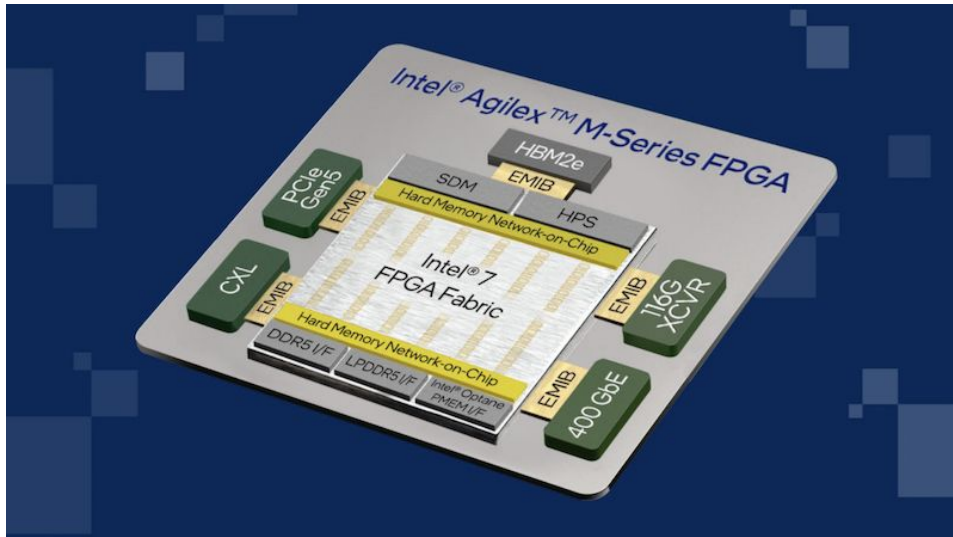
- A New HPC Era
  - Types of Legacy Hardware
- CPU: x86, Arm, POWER
- GPU: NVIDIA, AMD
- FPGA: Intel (Altera), AMD (Xilinx)
- ASIC: Offload

# AMD Xilinx Virtex + Versal

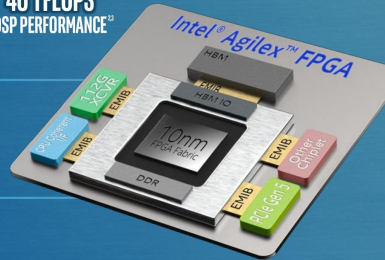




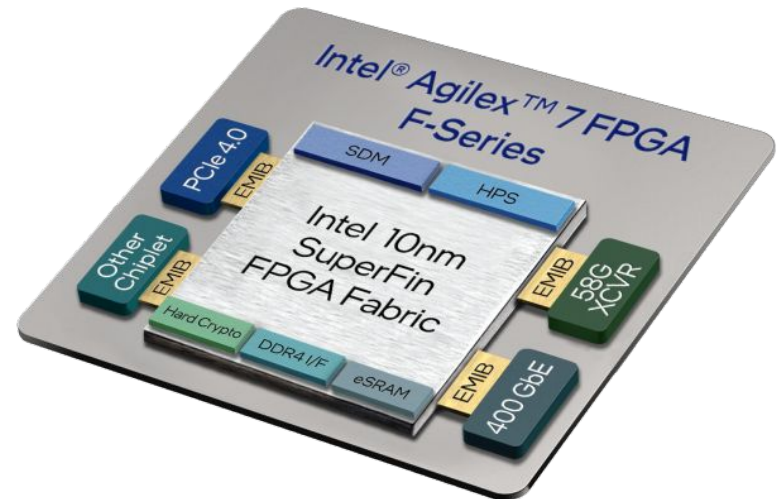
# Intel Altera Agilex



## The FPGA for the Data-Centric World

<b>PROCESS DATA</b>	2 <sup>ND</sup> GENERATION INTEL® HYPERFLEX™ ARCHITECTURE	UP TO 40% HIGHER PERFORMANCE*	UP TO 40% LOWER POWER**	UP TO 40 TFLOPS DSP PERFORMANCE**
<b>STORE DATA</b>	DDR5 & HBM	INTEL® OPTANE™ DC PERSISTENT MEMORY SUPPORT		
<b>MOVE DATA</b>	INTEL® XEON™ PROCESSOR COHERENT CONNECTIVITY & PCIe GEN5	112G TRANSCEIVER DATA RATES	<small>                 * compared to Intel® Stratix™ 10 FPGAs with FP16 configuration                  ** Based on current estimates, see slide 19 for details             </small>	

EMBARGO: APRIL 2, 2019 (10:00AM PACIFIC TIME)



# Silicon or Survive

- A New HPC Era
  - Types of Legacy Hardware
- CPU: x86, Arm, POWER
- GPU: NVIDIA, AMD
- FPGA: Intel (Altera), AMD (Xilinx)
- ASIC: Offload

# SmartNICs

## 2021 STH NIC Continuum



### Foundational

Basic interface for network connectivity (popular in the 10/100/1000 and some Nbase-T NICs)  
Less common at 100Gbps+ speeds due to packet processing demands on host CPUs

### Offload NIC

Offload for common network traffic functions (e.g., TCP/IP stack, limited virtualization features)

### SmartNIC

Offload functions with additional programmability to offload specific tasks from host systems (e.g., compression/decompression.)  
Designed to be a more flexible and expanded offload device

### DPU

Extended compute, offload, memory, and OS capabilities  
Designed to be an infrastructure endpoint that exposes resources to the data center and offloads key functionalities for data center scale computing (compute, storage, networking)  
Higher-levels of compute, offload, memory than SmartNICs

### Exotic

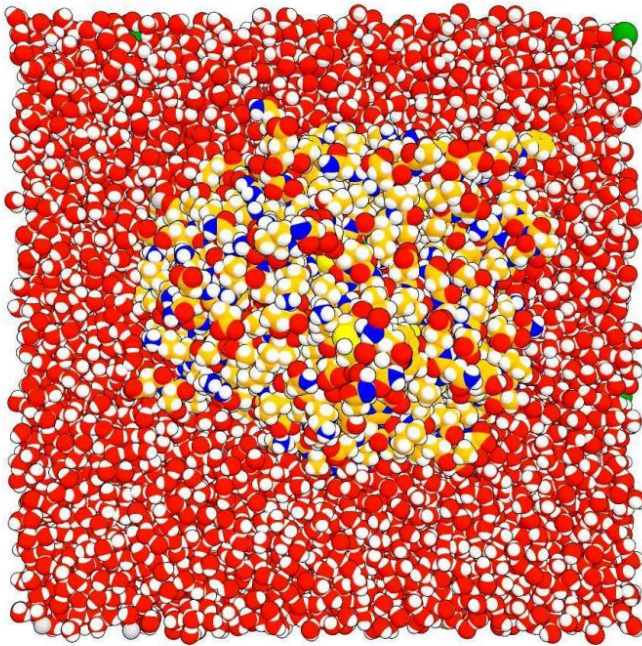
Usually, FPGA-based solutions that have fully customizable pipelines allowing for environment specific optimization  
Hardware also generally more specific to given deployment scenario

Increasing Cost, Complexity, Capabilities



# DEShaw's Anton 3

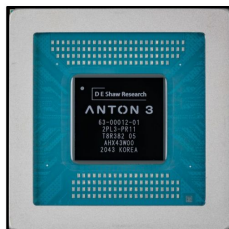
## Molecular dynamics (MD) simulation



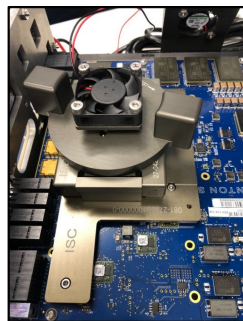
- Understand biomolecular systems through their motions
- Numerical integration of Newton's laws of motion
  - Model atoms as point masses
  - Compute forces on every atom based on current positions
  - Update atom velocities and positions in discrete time steps of a few femtoseconds
- Force computation described by a model: the force field

# DEShaw's Anton 3

## Baby pictures



29 September 2020: chips arrive  
MD running (water) < 9 h later

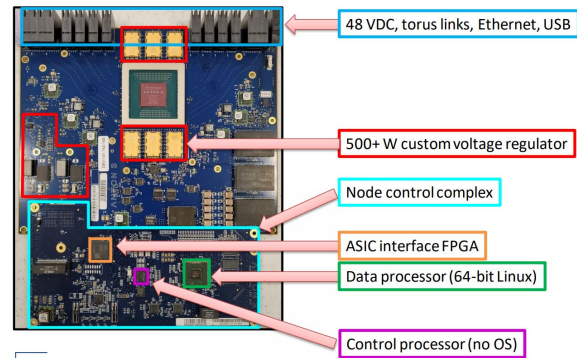


30 September 2020: 1<sup>st</sup> protein run  
Faster @ 250 MHz than Anton 2



31 October 2020: Multi-node

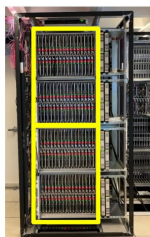
## Node board



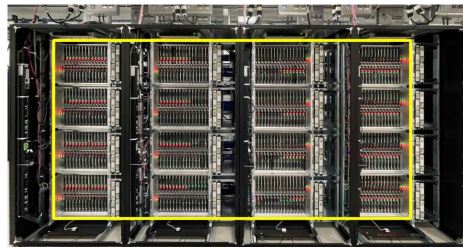
## Scale up



8x8 nodes



2x64 nodes



512 nodes

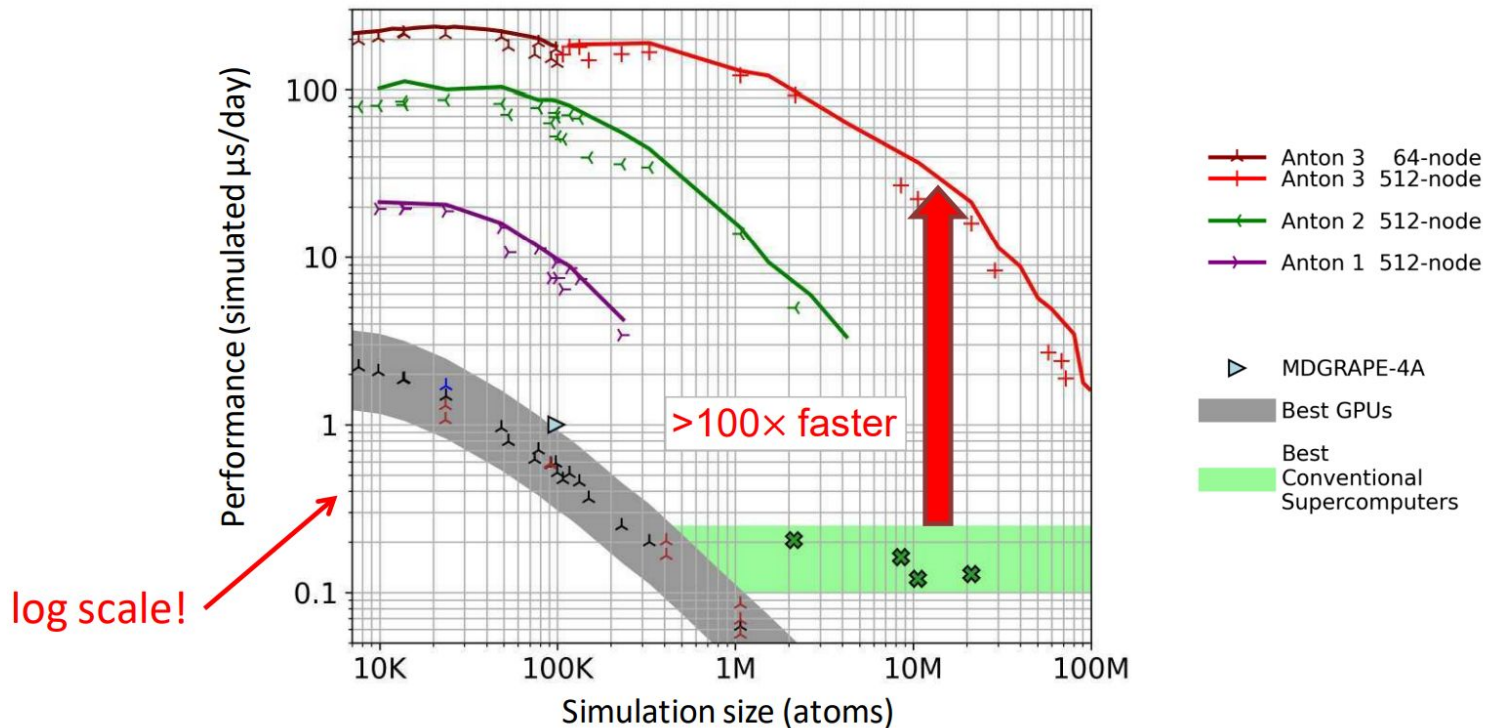
## The evolution of ANTON

	ANTON	ANTON 2	ANTON 3
Tape-out	2007	2012	2020
CPU cores	8+4+1	66	528*
PPIMs	32	76	528*
Flex SRAM	0.125 MiB	4 MiB	66 MiB*
Atoms / node	460	8,000	110,000*
Clock frequency	0.485/0.970 GHz	1.65 GHz	2.8+ GHz
Channel bandwidth	0.607 Tbps	2.7 Tbps	5.6+ Tbps
Process node	90 nm	40 nm	7 nm
Transistors	0.2 G	2.0 G	31.8 G
Die size	299 mm <sup>2</sup>	410 mm <sup>2</sup>	451 mm <sup>2</sup>
Power	30 W	190 W	360 W

\* 22/24 columns

# DEShaw's Anton 3

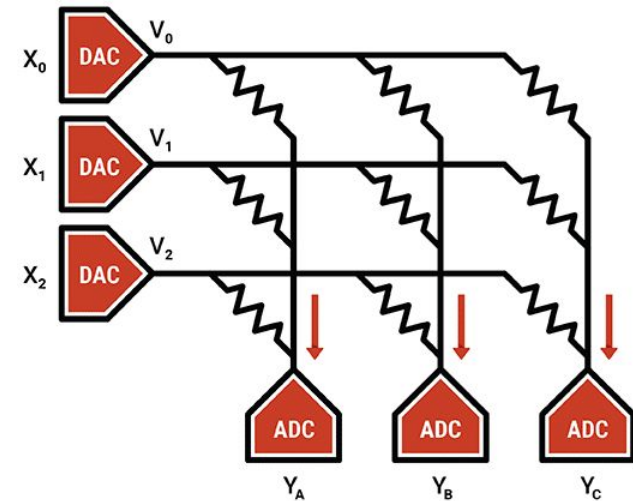
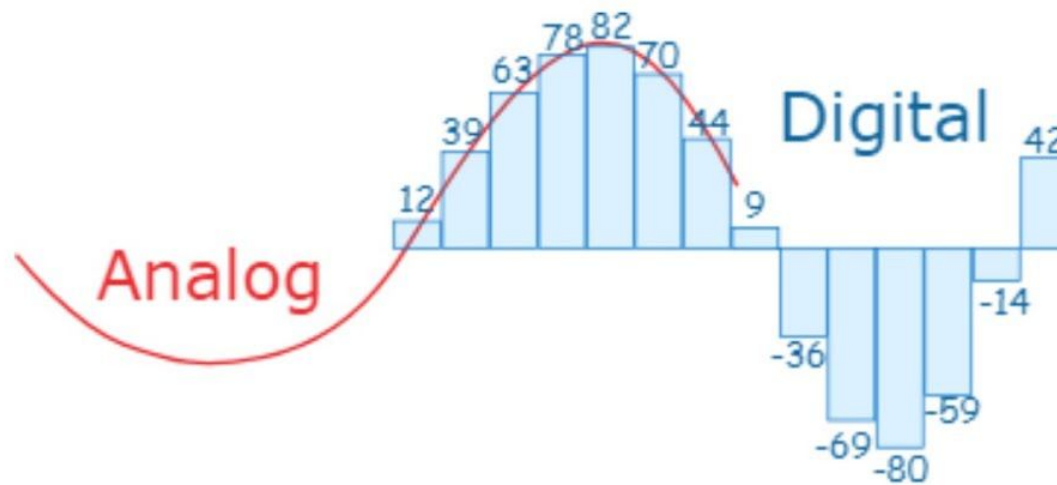
## MD performance



# Silicon or Survive

- A New HPC Era
  - Types of Legacy Hardware
  - New Paradigms: Analog, Neuro, Quantum, Optical
  - Push for Low Precision
- AI Hardware
  - Established Players
  - Current \$10B Market
  - Case Studies: Wafer Scale, Analog Edge
  - Roadmaps
  - Software
- Q&A

# Analog Computing



- ▲ Super Low Power
- ▲ Super Low Latency
- ▲ 'Any' Value Possible

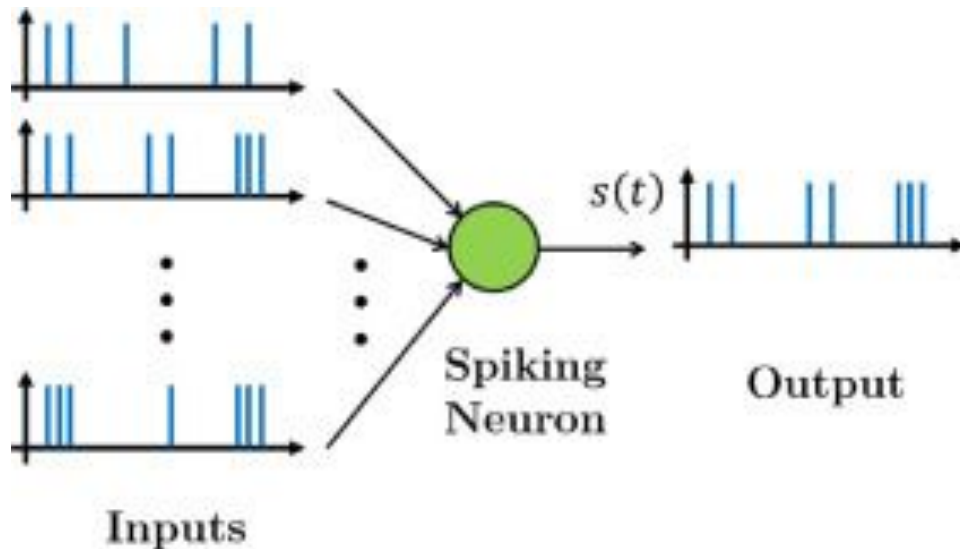
- ▼ Conversion Accuracy
- ▼ Non-Linear Response
- ▼ Scaling

Recent Key Players: Mythic AI, IBM, Aspinity



# Neuromorphic Computing

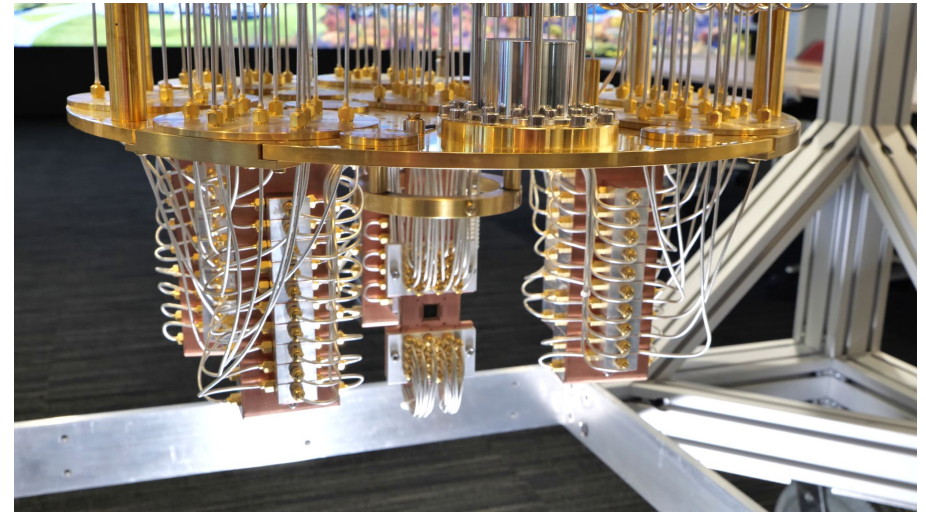
Beware! Some companies calling themselves 'Neuromorphic' aren't actually doing it



Recent Key Players: Intel Loihi 2, Spinnaker



# Quantum Computing



- ▲ Physics, Chem, Bio
- ▲ Math + Encryption
- ▲ Machine Learning

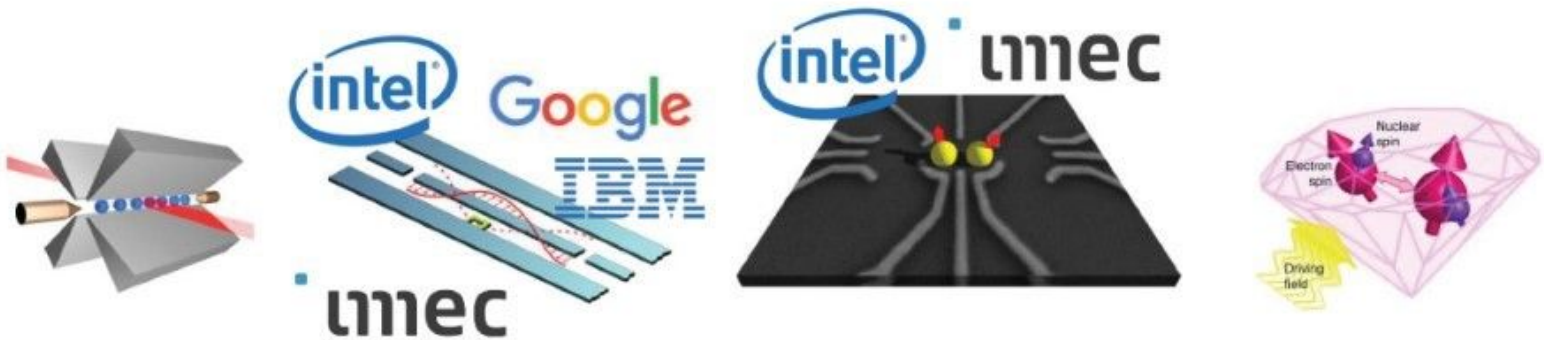
- ▼ Any other math
- ▼ High Barrier to Entry
- ▼ Need a billion qubits

Recent Key Players: Intel, IBM, Google, Microsoft, Amazon, DWave, Alibaba, IonQ



# Quantum Computing

- Several types of Qubits available:

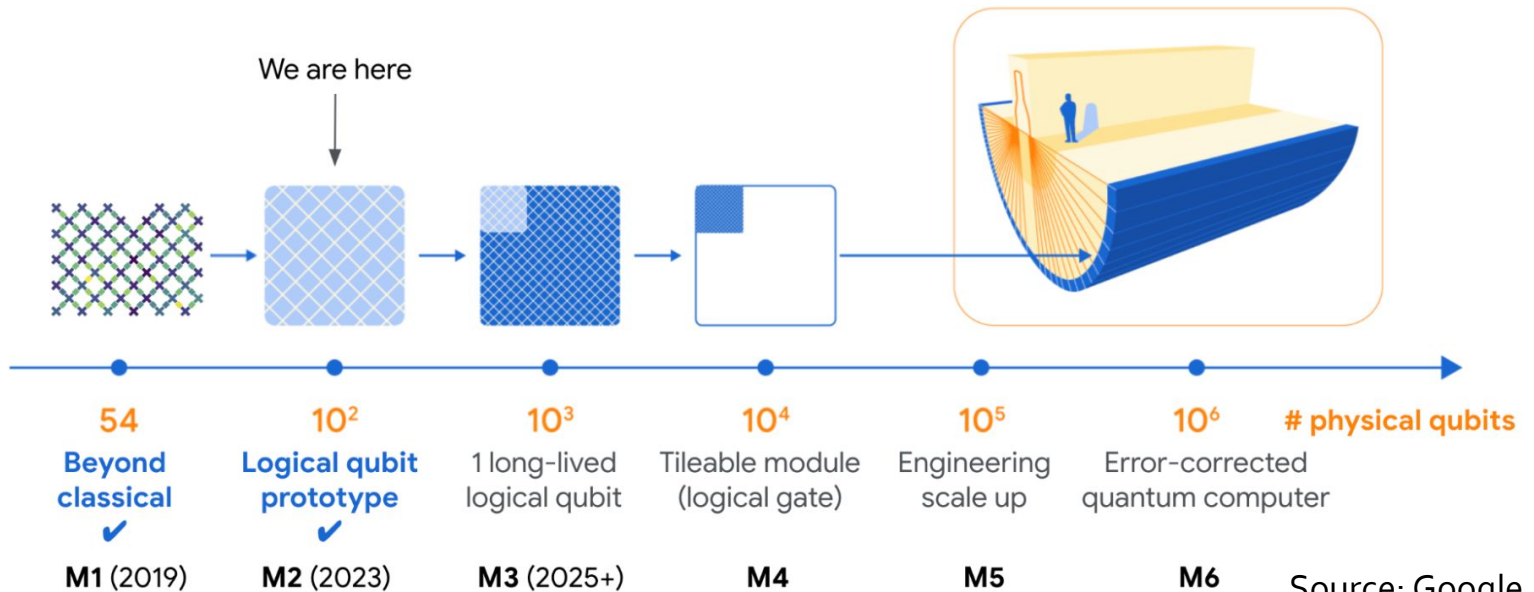


	Ion trap	Superconducting	Semiconducting	NV-centers
<b>Coherence time</b>	> 1 s	~ 90 $\mu$ s	~ 28 ms	~ 250 ns
<b># Qubits</b>	~ 10	17 (50)	~ 3	~ 3
<b>Materials</b>	$^9\text{Be}^+$ , $^{43}\text{Ca}^+$ , ...	Al, Nb, TiN, ...	Si, GaAs, ...	C, N
<b>Scalability</b>	-	+	+++	++

# Quantum Computing



## Error Correction and Qubit Scaling?

Quantum error correction	–	Enabled	At scale
# Physical qubits	10 – 100	100 – 1000	$10^4 – 10^6$
# Logical qubits	–	1	10 – 1000+
Logical error	$10^{-3}$	$10^{-2} – 10^{-6}$	$10^{-6} – 10^{-12}$



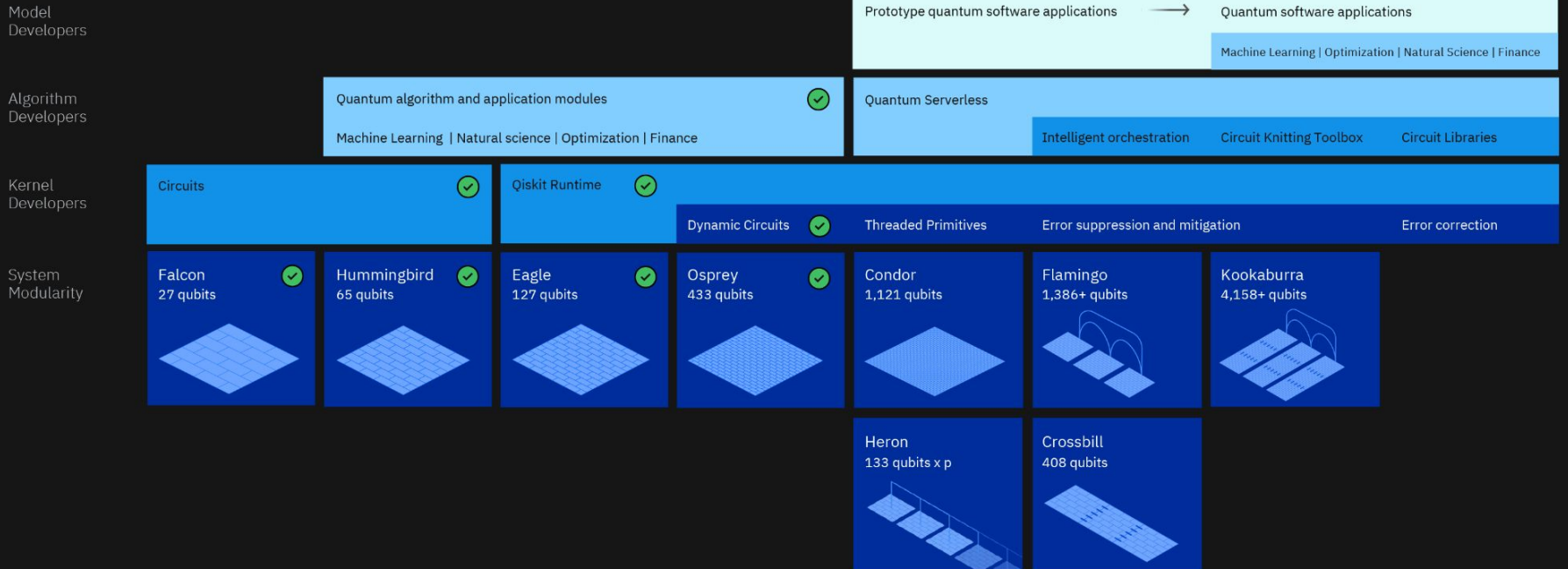
# Qubits Today

## Development Roadmap

Executed by IBM   
On target 

IBM Quantum

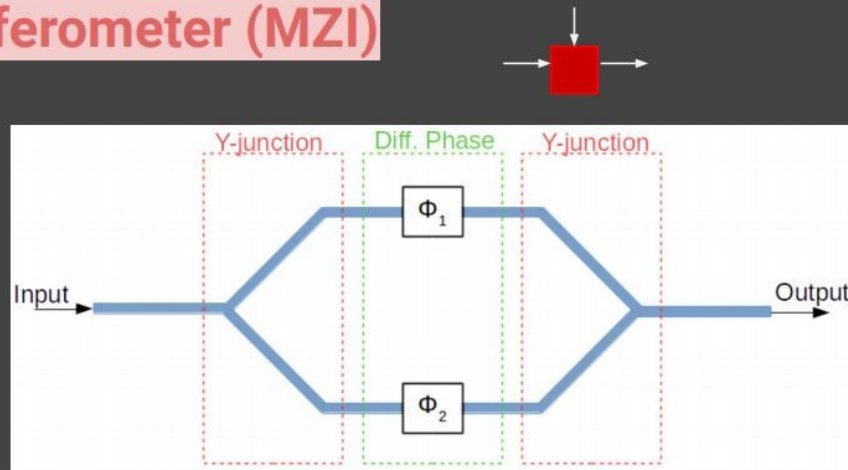
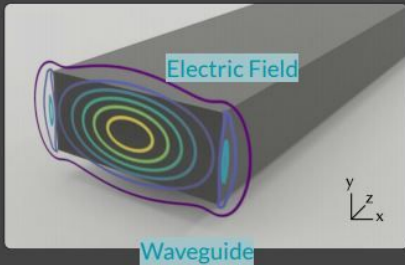
2019      2020      2021      2022      2023      2024      2025      Beyond 2026



# Optical Computing

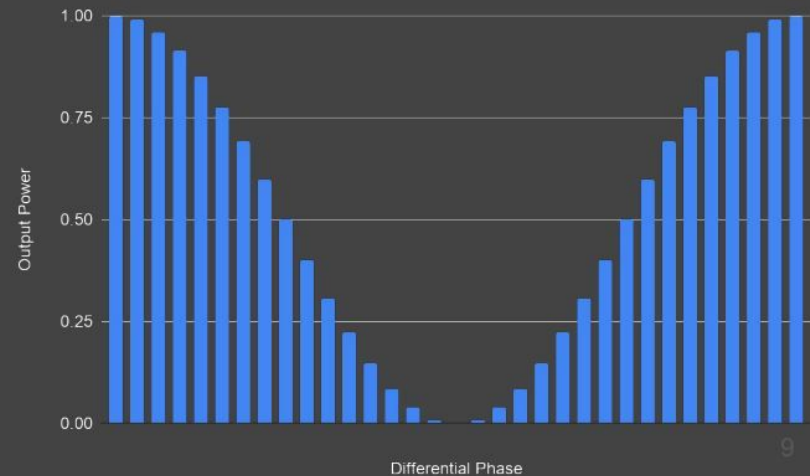
## Mach Zehnder Interferometer (MZI)

Compute tile

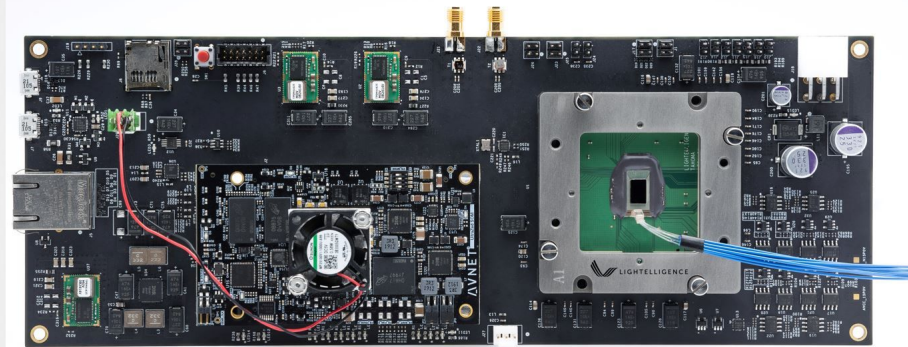


MZIs provide observation of phase shift through interference

These become useful when you modulate the phase on  $\Phi_1$  and  $\Phi_2$



# Optical Computing



- ▲ No Power
- ▲ Speed-of-light fast

- ▼ Scales Poorly
- ▼ Manufacturing

Recent Key Players: LightMatter, Lightelligence

# Optical Computing

The future of artificial intelligence.

Faster, lower energy and decoupled from Moore's Law.

## Photonic Compute Module

12LP ASIC

90WG PIC

64x64 matrix

1GHz vector rate

8-bit signed operands

200ps latency

150 mm<sup>2</sup>



**Optical computing.**



# Silicon or Survive

- A New HPC Era
  - Types of Legacy Hardware
  - New Paradigms: Analog, Neuro, Quantum, Optical
  - Push for Low Precision
- AI Hardware
  - Established Players
  - Current \$10B Market
  - Case Studies: Wafer Scale, Analog Edge
  - Roadmaps
  - Software
- Q&A

# Quantization

Using fewer bits saves power, and with the right architecture, can be sped up. But the tradeoff is range and accuracy

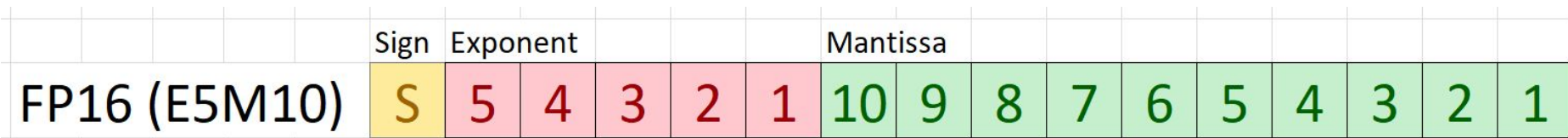
## FP16 vs FP32





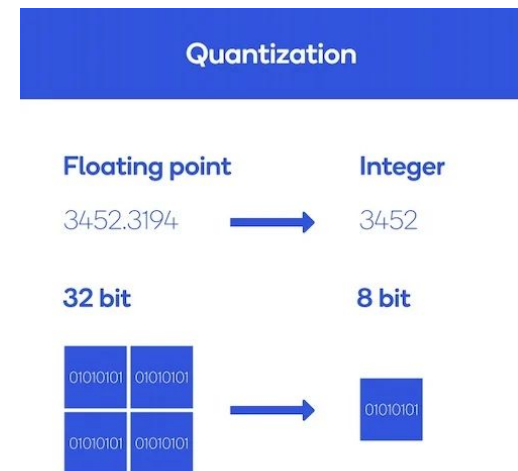
# Quantization

How numbers are represented matters:

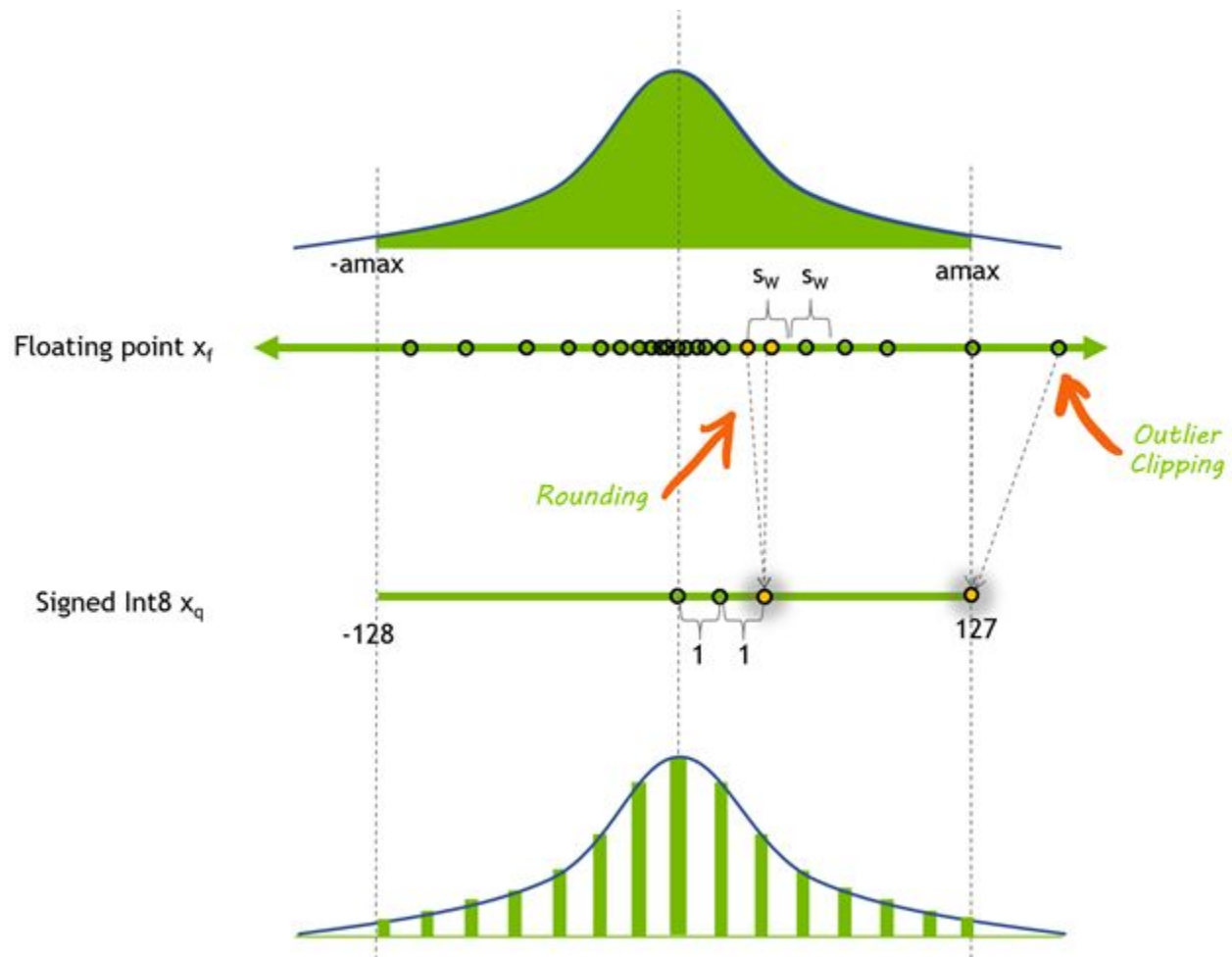


$$\text{Value} = (-1)^S \cdot \text{mantissa} \cdot 2^{\text{exponent}}$$

IEEE754 defines number standards - what to do with infinities, sub-normal values, etc. Lots of chip companies are now defining their own number types to improve performance.



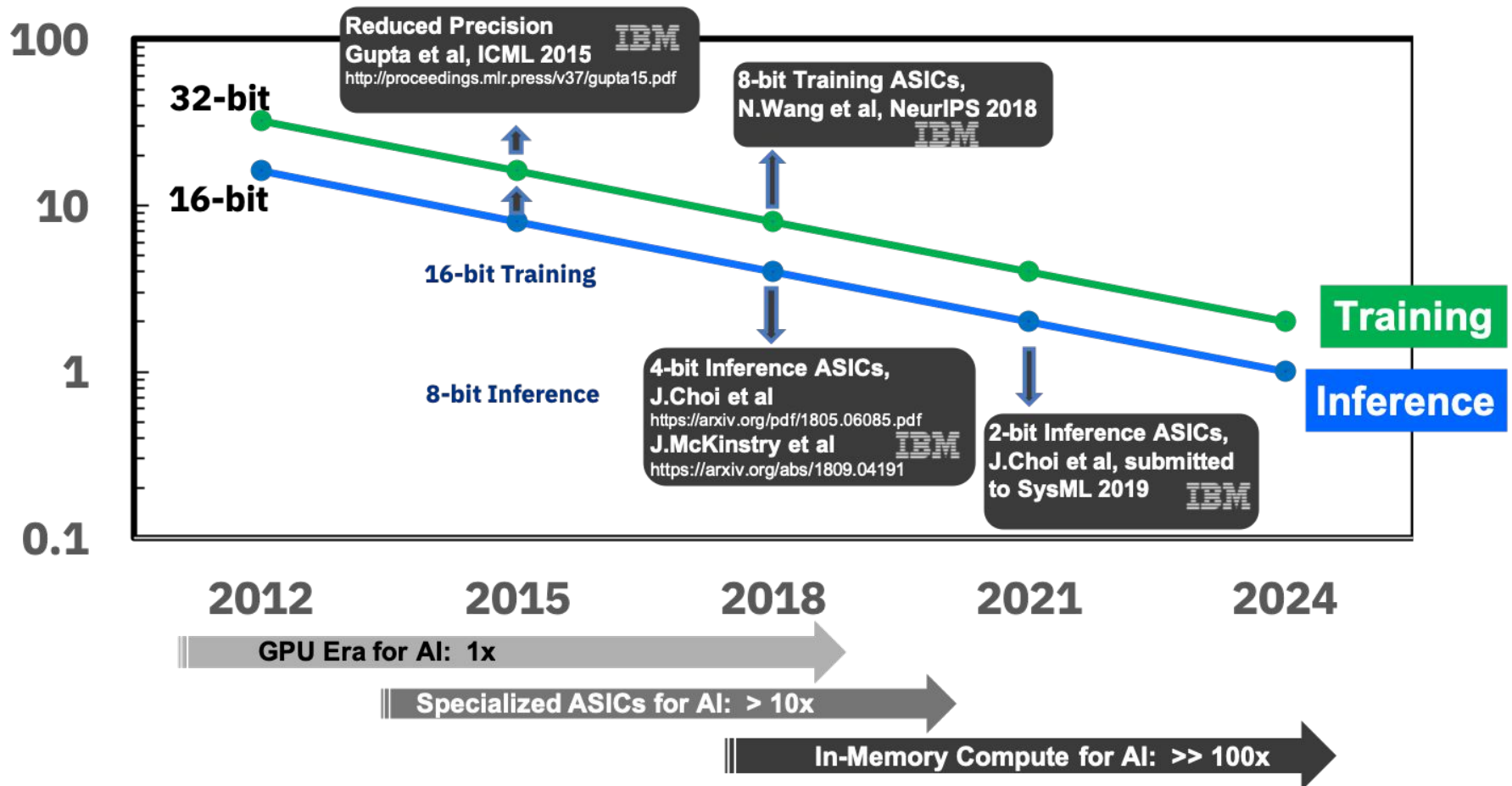
# Quantization





# Quantization

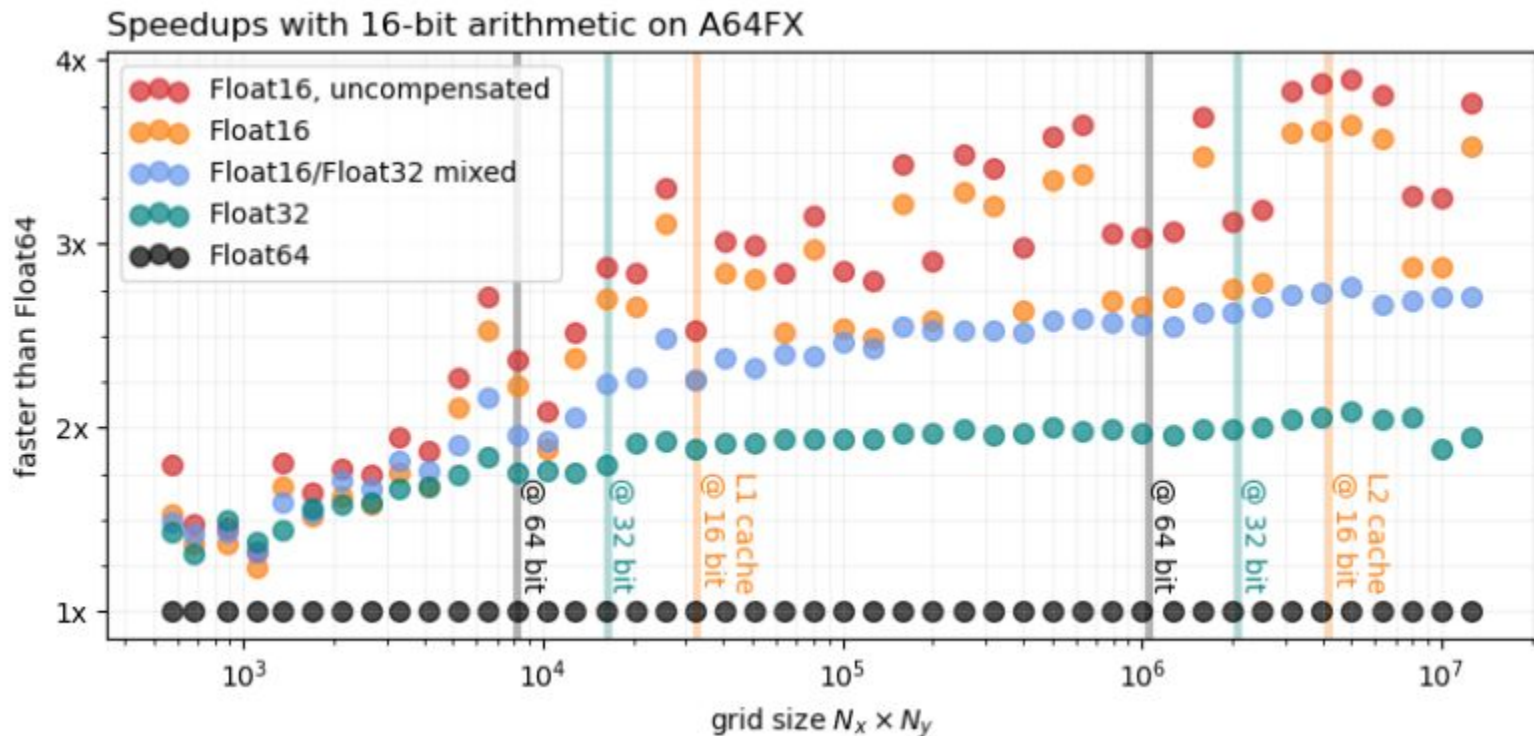
## IBM Research is Leading in Reduced Precision Scaling





# Quantization in HPC

But all HPC runs are FP64 or FP32, right?



Source: NextPlatform - FP16 on Climate and Weather on Isebard (Bristol)

# Quantization in HPC

Perhaps the solution is mixed precision: FP64 when you need it, FP16 when you don't

4:00 PM

4:00 PM 4:25 PM **A Mixed Precision Randomized Preconditioner for the LSQR Solver on GPUs**

Randomized preconditioners for large-scale regression problems have become extremely popular over the past decade. Such preconditioners...

Hall 4 - Ground Floor

Research Paper



**Vasileios Georgiou**  
Karlsruhe Institute of...

Mixed Precision Algorithms Numerical Libraries

9:00 AM

9:50 AM 10:15 AM **GPU-based Low-precision Detection Approach for Massive MIMO Systems**

Massive Multiple-Input Multiple-Output (M-MIMO) uses hundreds of antennas in mobile communications basestations to increase the amou...

Hall F - 2nd Floor

Research Paper



**Adel Dabah**  
KAUST

Mixed Precision Algorithms Industrial Use Cases of HPC, ML and QC

3:00 PM

3:00 PM 3:20 PM **High-Performance GMRES Multi-Precision Benchmark: Design, Performance, and Challenges**

We propose a new benchmark for high-performance (HP) computers that uses a variation of the Generalized Minimum RESidual method (GMRES)...

Hall Z - 3rd Floor

Focus Session



**Piotr Luszczek**  
University of Tennessee,...

Exascale Systems Mixed Precision Algorithms  
Performance Modeling and Tuning Emerging HPC Processors and Accelerators

3:20 PM 3:35 PM

**Q&A: Mixed Precision Algorithms**

Hall Z - 3rd Floor

Focus Session



**Hatem Ltaief**  
KAUST



**Piotr Luszczek**  
University of Tennessee,...

2:00 PM

2:00 PM 6:00 PM **Modern Mixed-Precision Methods: Hardware Perspectives, Algorithms, Kernels, and Solvers**

This tutorial will expose the audience to the rapidly expanding landscape of mixed- and multi- precision methods. The ongoing cross-pollination...

Hall Y6 - 2nd Floor

Tutorial



**Jack Dongarra**  
University of Tennessee,...



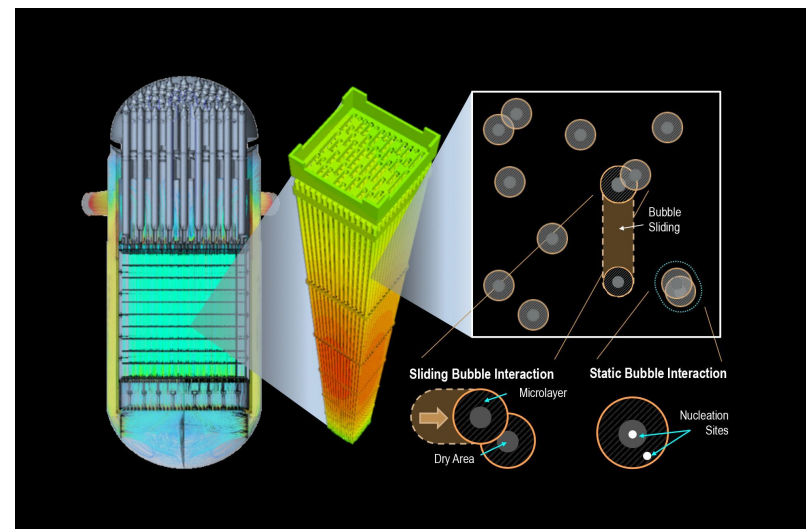
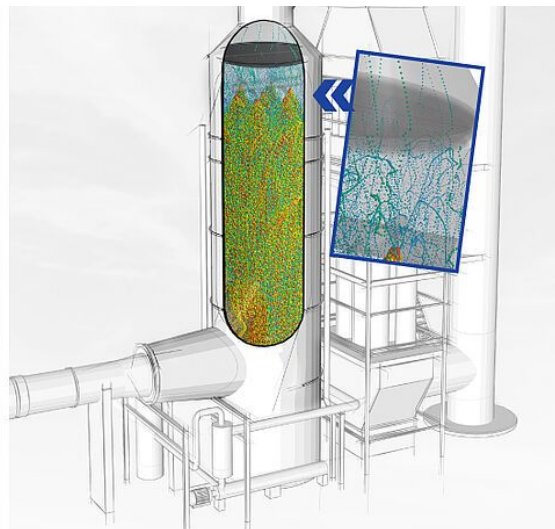
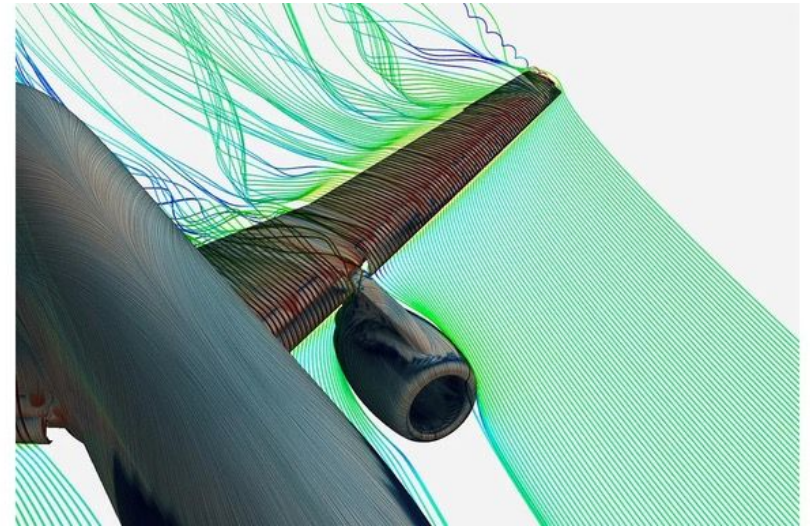
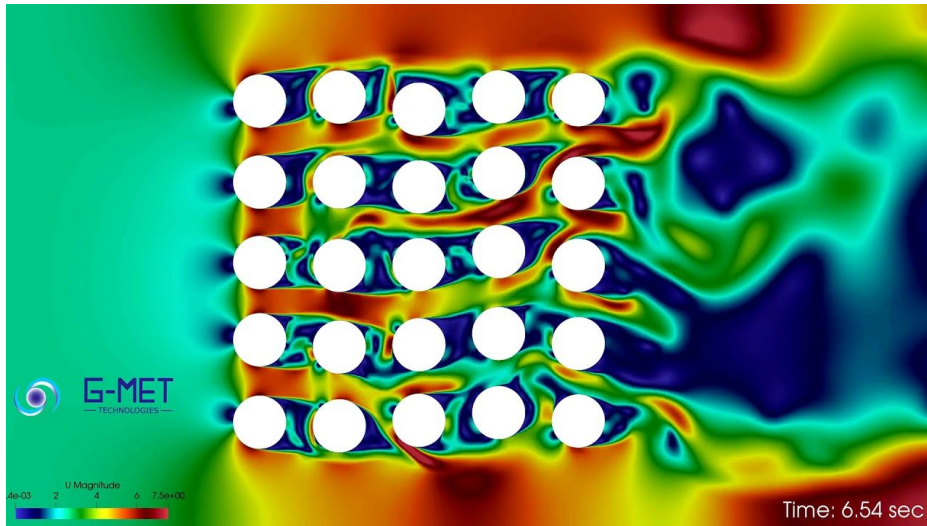
**Hartwig Anzt**  
University of Tennessee,...



**Piotr Luszczek**  
University of Tennessee,...

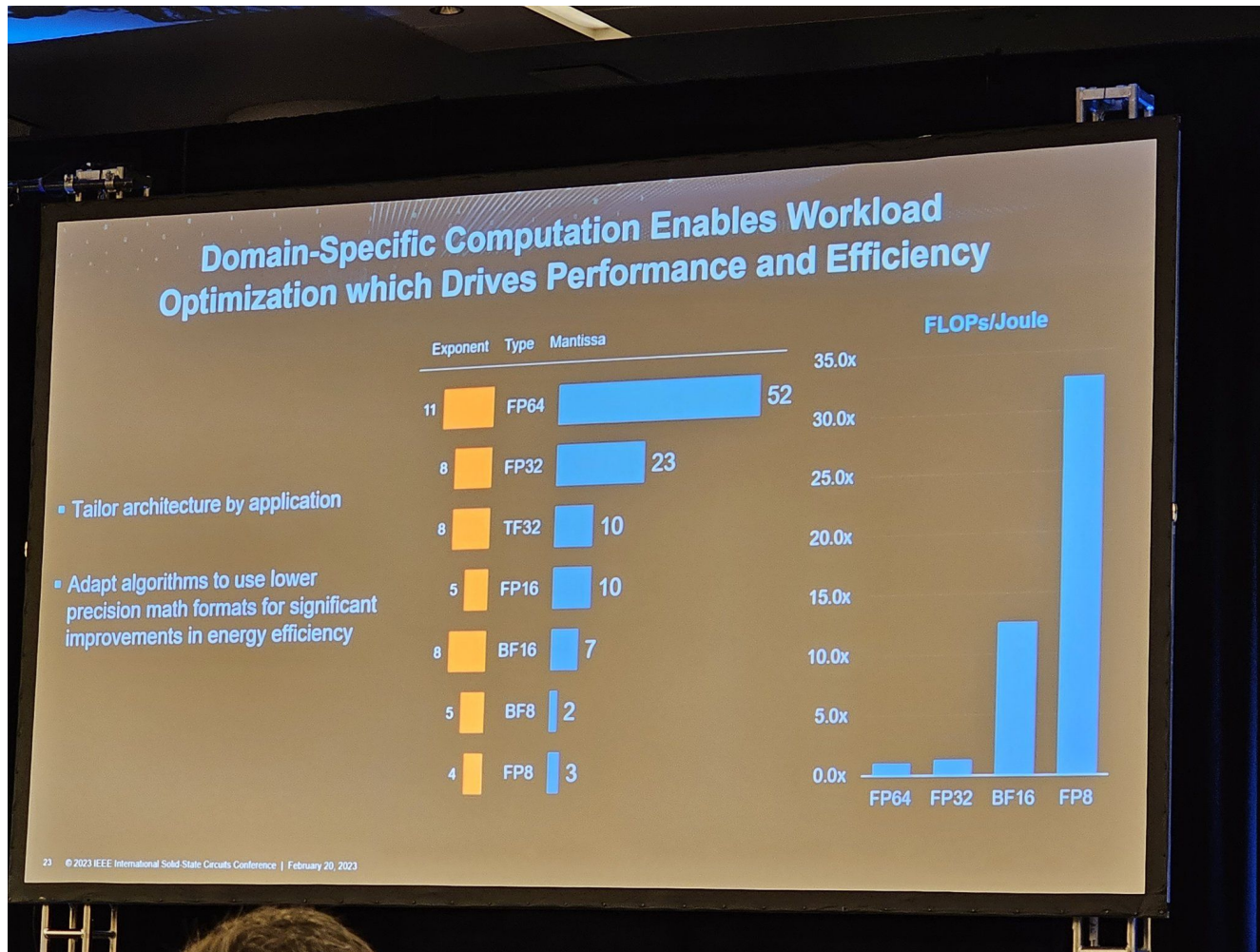
Mixed Precision Algorithms Numerical Libraries  
Performance Modeling and Tuning Emerging HPC Processors and Accelerators  
Sustainability and Energy Efficiency

# Quantization in HPC





# Push for Quantization



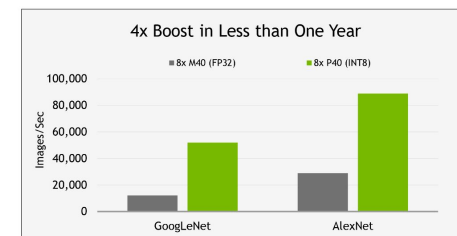


# Push for Quantization

## Push for reduced precision comes from AI

	Peak Performance
Transistor Count	54 billion
Die Size	826 mm <sup>2</sup>
FP64 CUDA Cores	3,456
FP32 CUDA Cores	6,912
Tensor Cores	432
Streaming Multiprocessors	108
FP64	9.7 teraFLOPS
FP64 Tensor Core	19.5 teraFLOPS
FP32	19.5 teraFLOPS
TF32 Tensor Core	156 teraFLOPS   312 teraFLOPS*
BFLOAT16 Tensor Core	312 teraFLOPS   624 teraFLOPS*
FP16 Tensor Core	312 teraFLOPS   624 teraFLOPS*
INT8 Tensor Core	624 TOPS   1,248 TOPS*
INT4 Tensor Core	1,248 TOPS   2,496 TOPS*
GPU Memory	40 GB
GPU Memory Bandwidth	1.6 TB/s
Interconnect	NVLink 600 GB/s PCIe Gen4 64 GB/s
Multi-Instance GPUs	Various Instance sizes with up to 7MIGs @5GB
Form Factor	4/8 SXM GPUs in HGX A100
Max Power	400W (SXM)

NVIDIA Architecture	CUDA Cores				Tensor Cores					
	FP64	FP32	FP16	INT8	FP64	TF32	FP16	INT8	INT4	INT1
Volta	32	64	128	256			512			
Turing	2	64	128	256			512	1024	2048	8192
Ampere (A100)	32	64	256	256	64	512	1024	2048	4096	16384
Ampere, sparse						1024	2048	4096	8192	



P40	
# of CUDA Cores	3840
Peak Single Precision	12 TeraFLOPS
Peak INT8	47 TOPS
Low Precision	4x 8-bit vector dot product with 32-bit accumulate
Video Engines	1x decode engine, 2x encode engines
GDDR5 Memory	24 GB @ 346 GB/s
Power	250W

GoogleNet, AlexNet, batch size = 128, CPU: Dual Socket Intel E5-2697v4

Source: NVIDIA A100

# Silicon or Survive

- A New HPC Era
  - Types of Legacy Hardware
  - New Paradigms: Analog, Neuro, Quantum, Optical
  - Push for Low Precision
- AI Hardware
  - Established Players
  - Current \$10B Market
  - Case Studies: Wafer Scale, Analog Edge
  - Roadmaps
  - Software
- Q&A

# AI Hardware Established Players

## Training

## Inference



H100  
A100  
V100



Trainium



Sapphire Rapids  
Ice Lake  
Skylake  
Greco



H100 / A100  
A10  
T4



TPU v4  
TPU v3



Ponte Vecchio  
Sapphire Rapids  
Ice Lake  
Habana Gaudiz  
Falcon Shores



Inferentia  
Graviton 3



TPU v4  
TPU v3  
Argos

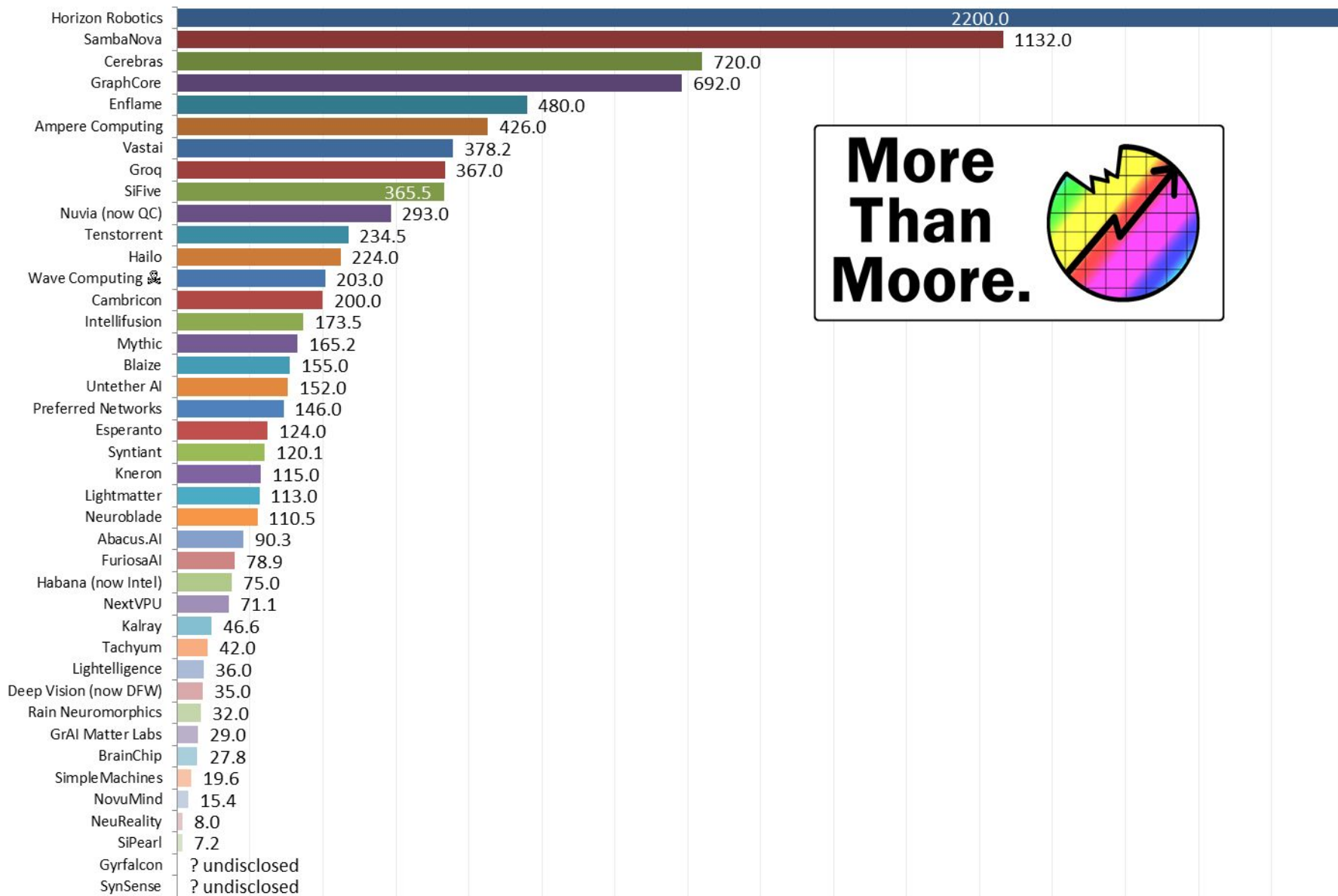
# AI Hardware Investment

Most of that hardware you buy, or is in the cloud. But there are 50+ startups creating AI hardware, some of which is already in HPC.

This market is \$10B+



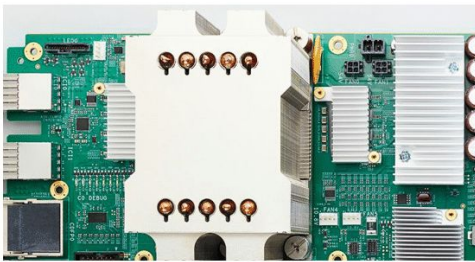
## AI Pure Play (+ others) Funding September 2022 - Values in USD \$m



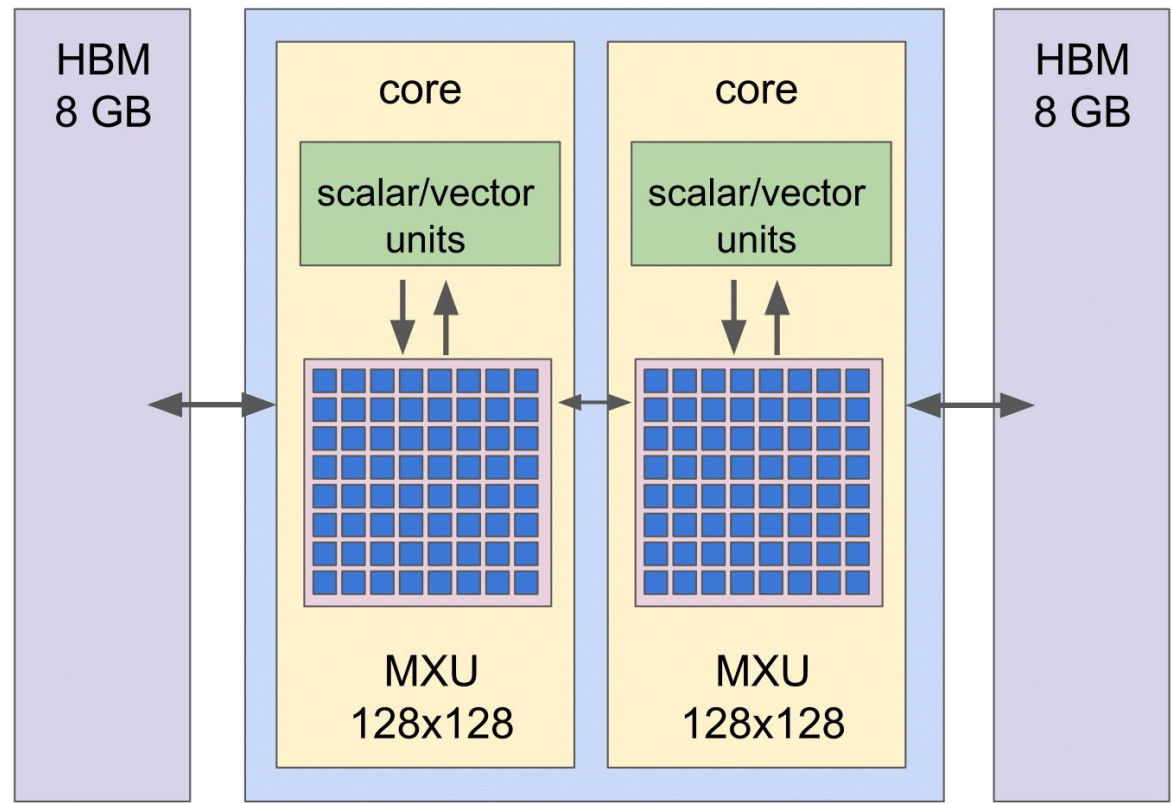
**More Than Moore.**

# Google TPU v2

## TPUv2 Chip

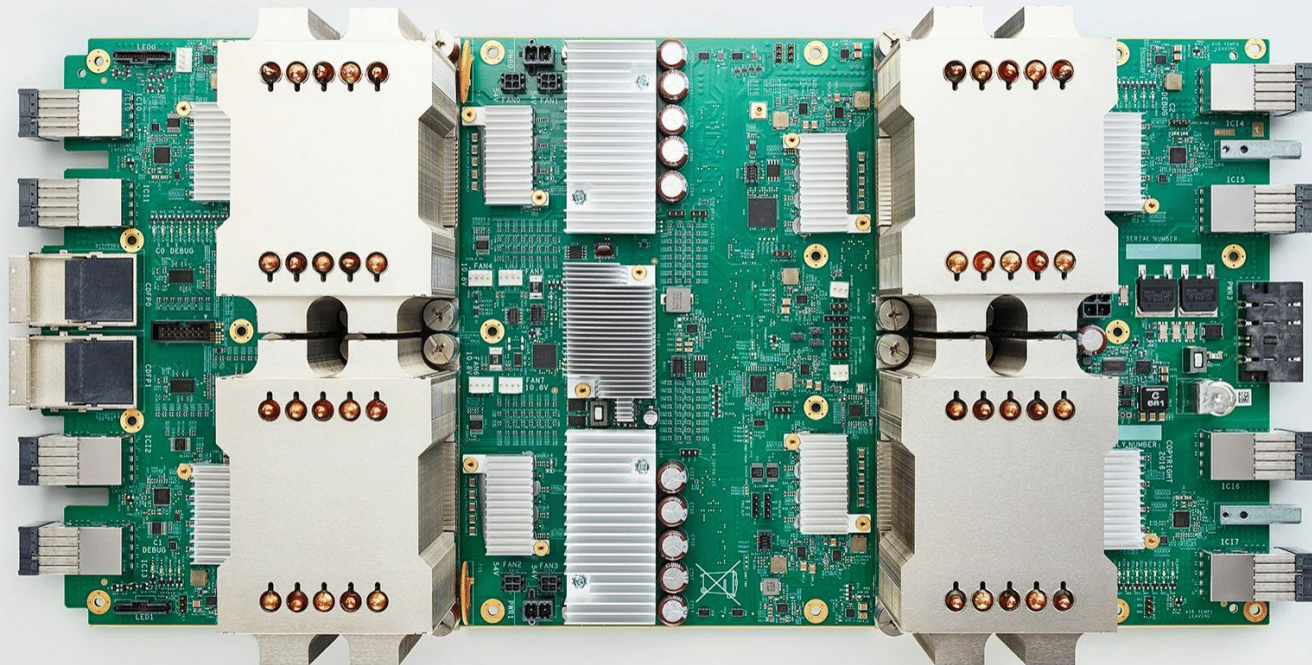


- 16 GB of HBM
- 600 GB/s mem BW
- Scalar/vector units: 32b float
- MXU: 32b float accumulation but reduced precision for multipliers
- 45 TFLOPS



# Google TPU (v2)

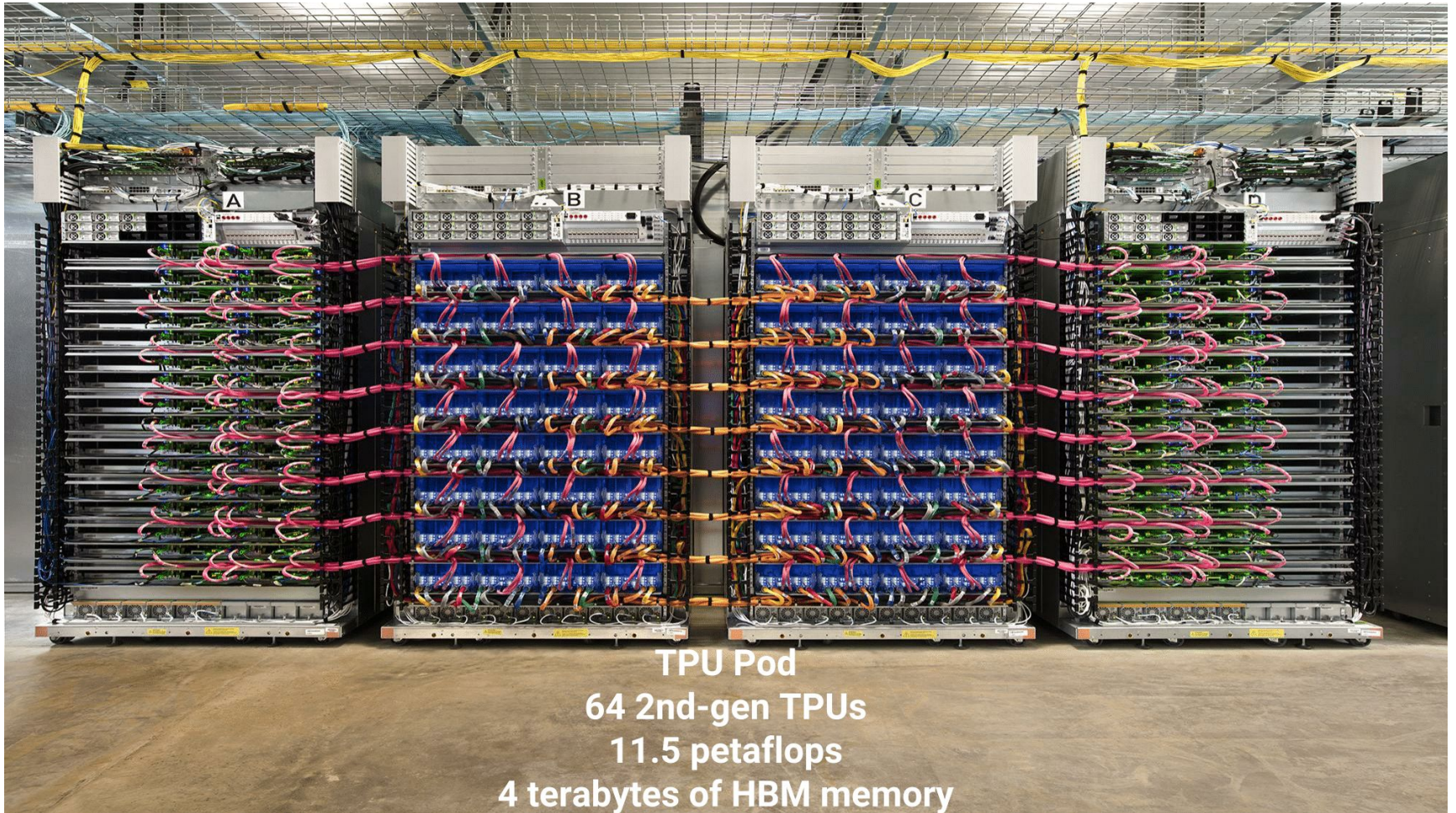
## Tensor Processing Unit v2



Google-designed device for neural net **training** and **inference**

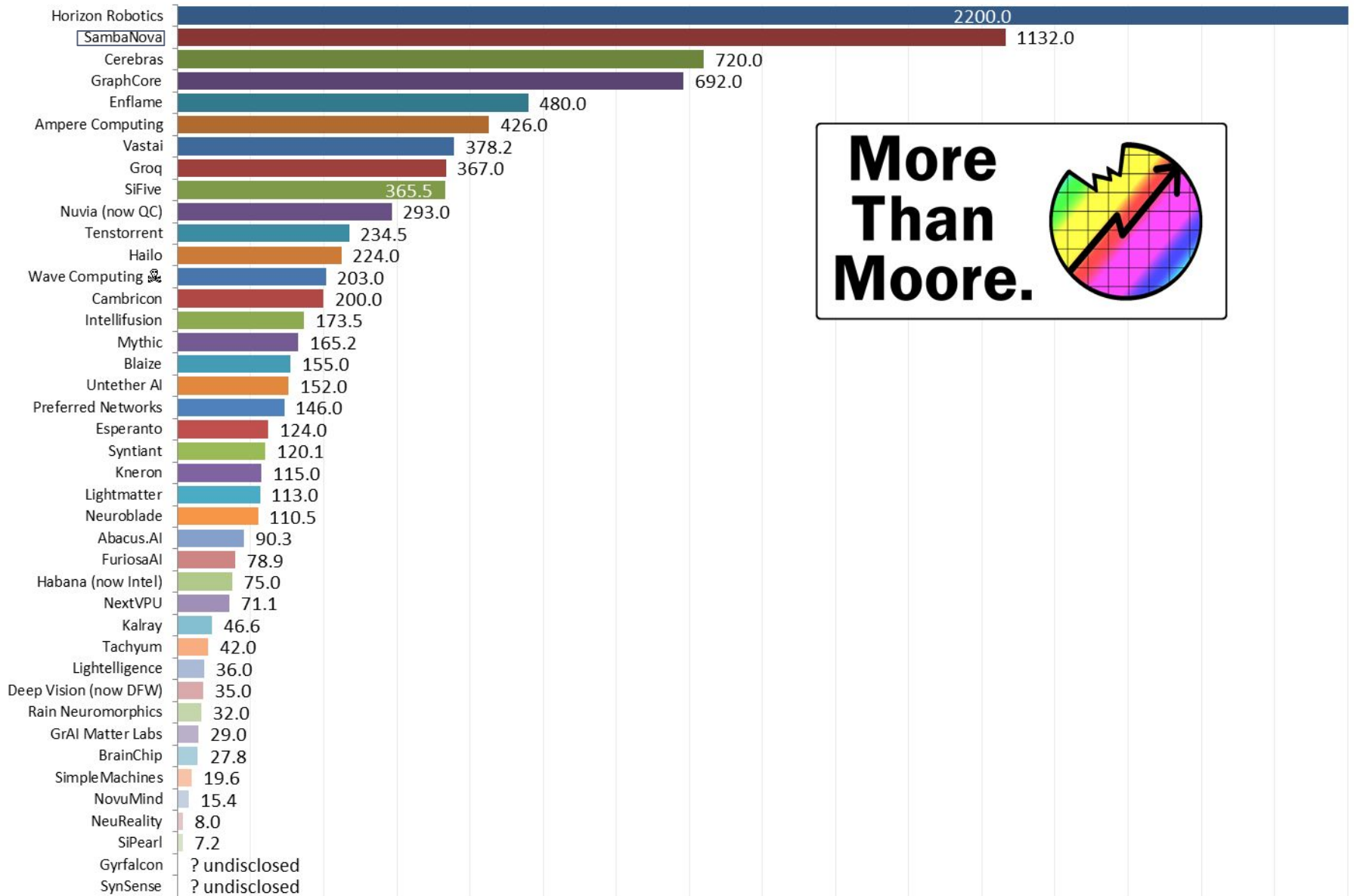


# Google TPU v2





# AI Pure Play (+ others) Funding September 2022 - Values in USD \$m



**More Than Moore.**

# SambaNova Cardinal



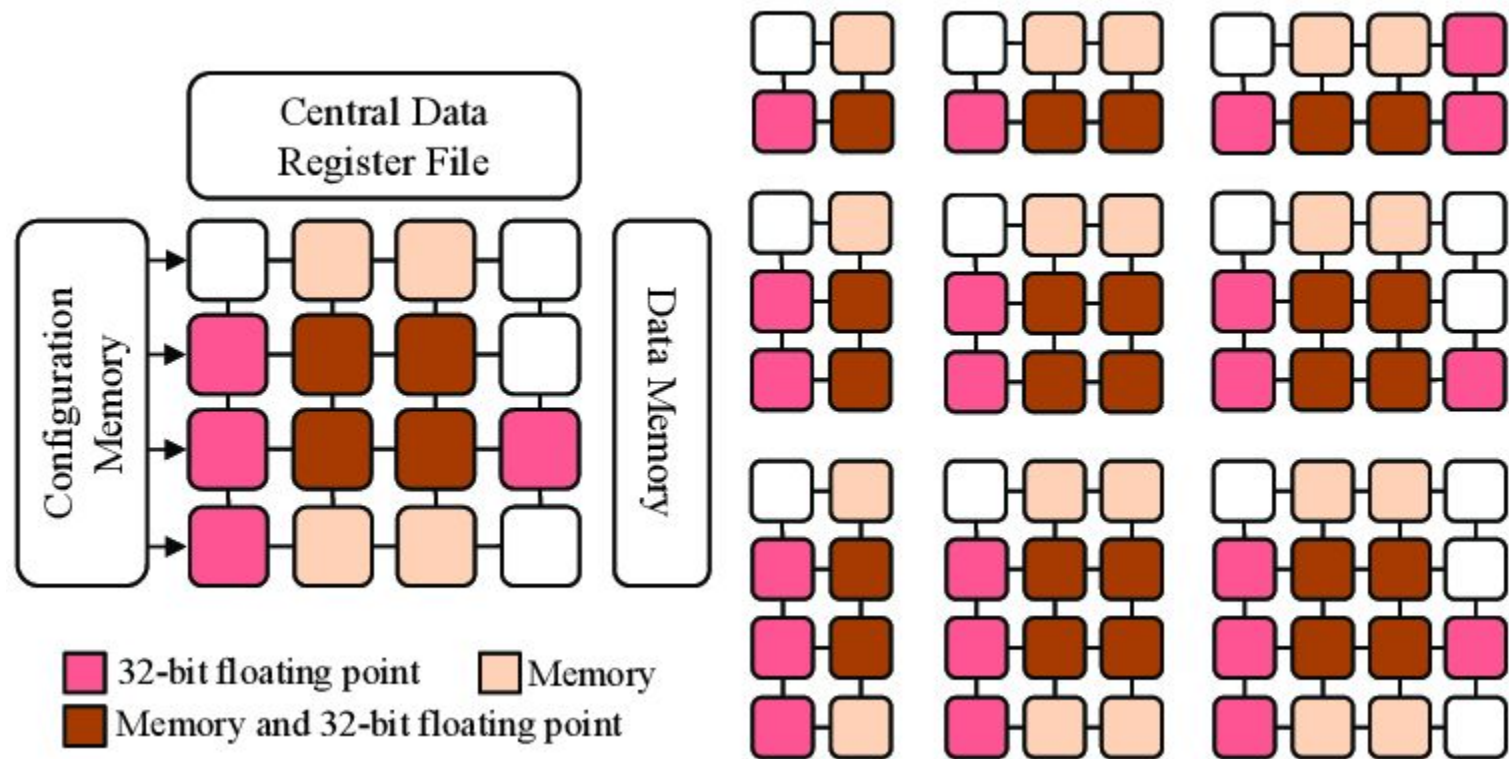
## SambaNova Systems® Cardinal SN10 RDU



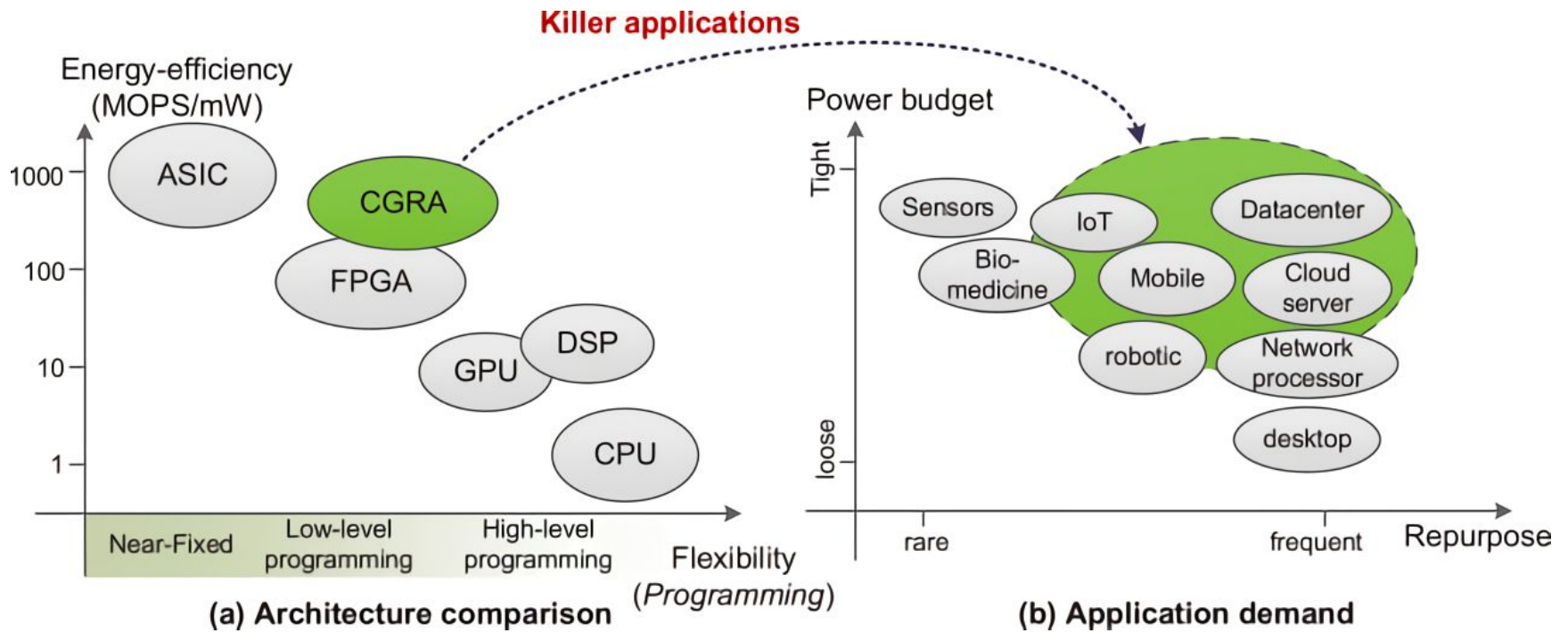
- First Reconfigurable Dataflow Unit (RDU)
- TSMC 7nm
  - Taped Out first half of 2019
  - 40B transistors, 50 Km of wire
- 640 Pattern Compute Units
  - >300 BF16 TFLOPs
  - BF16 with FP32 accumulation, stochastic rounding
  - Also supports FP32, Int32, Int16, Int8 data formats
- 640 Pattern Memory Units
  - >300 MB on-chip memory
  - 150 TB/s on-chip memory bandwidth
  - Memory transformation operations



# CGRA



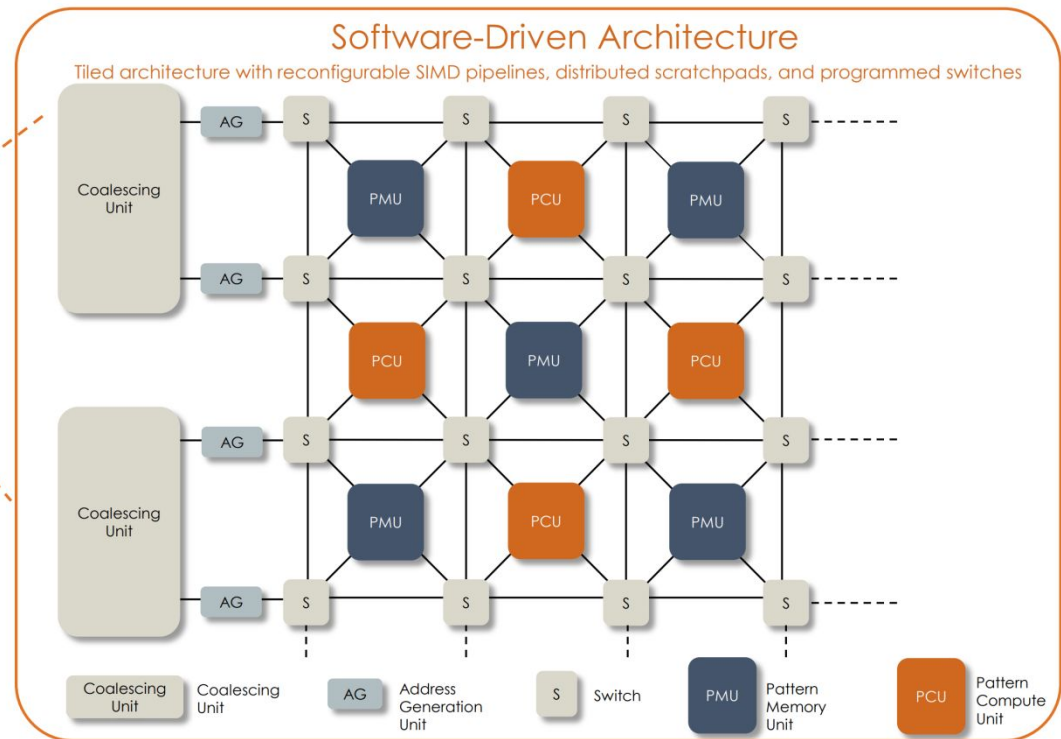
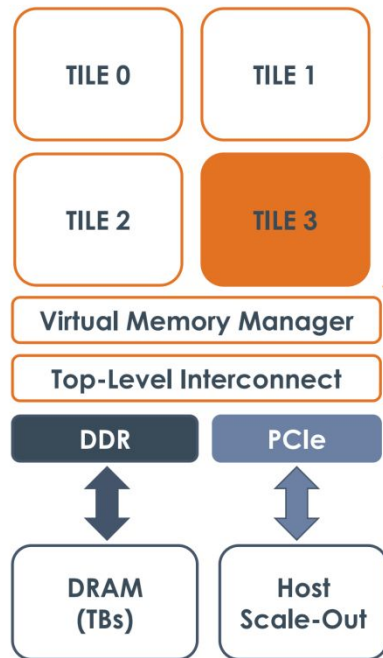
# CGRA





# SambaNova Cardinal SN10

## Cardinal SN10: Tile





# SambaNova Cardinal SN10

## Dataflow Architecture for Terabyte Sized Models



DataScale SN10-8R  
1/4 Rack System

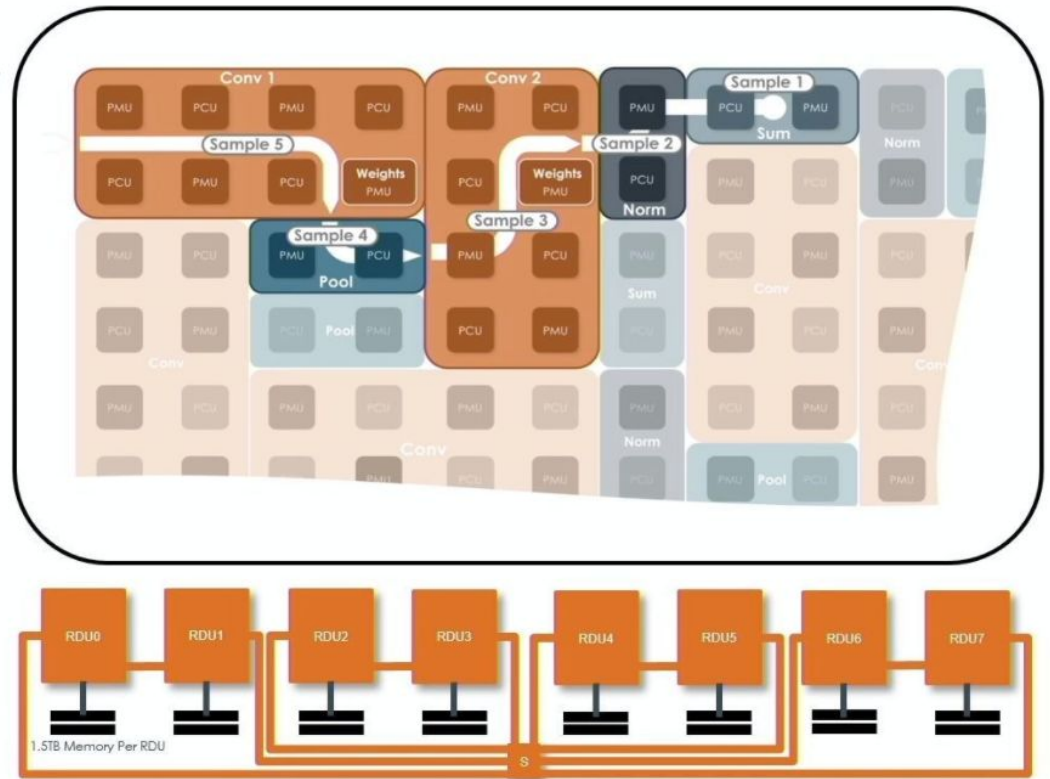
Dataflow Efficiency

+

Compute  
Capability

+

Large  
Memory Capacity



# SambaNova

## SambaNova system added to Fugaku supercomputer to boost AI performance

World's second fastest supercomputer gets a boost

March 06, 2023 By: [Sebastian Moss](#) [Have your say](#)



AI hardware-as-a-service company SambaNova Systems will deploy one of its supercomputers at the RIKEN Center for Computational Science (R-CCS) in Japan.

The DataScale system will be paired with Fugaku, the world's second fastest supercomputer.

"The provision of SambaNova's system resources to R-CCS provides a new option for accelerating the integration of HPC simulations and AI with Fugaku," Professor Satoshi Matsuoka, director of R-CCS, said.

"The new SambaNova system will boost research into the convergence of HPC and AI, including ultra-high-resolution computer vision for building a digital twin for the Society 5.0 era."



R-CCS researchers will use DataScale to develop surrogate models to improve the accuracy of ultra-high-resolution 3D computer vision, including for the inspection of social infrastructure such as highways, and to process ultra-high-resolution image datasets.

While Fugaku uses 152,064 Fujitsu's 48-core Arm-based A64FX processor chips, SambaNova has developed its own Reconfigurable Dataflow Unit (RDU) chip, which is only available within the wider DataScale package. The company claims that developing a single hardware and software package optimizes it for AI workloads.

It previously paired a DataScale system with Lawrence Livermore National Laboratory's Corona supercomputer, and another one of its systems [is being tested by Argonne National Laboratory](#).



### Resources

[More](#)



#### Uptime on the Line?

14 Apr 2023

#### Data Center Ecosystem Report 2023: The Irish Market

13 Apr 2023

#### Partnering to Create Data Centers of the Future

12 Apr 2023

#### Rethinking Data Center SSDs in context of real-world workloads and TCO

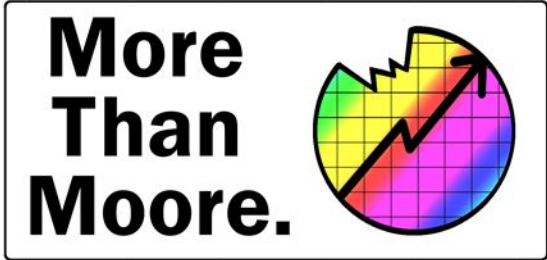
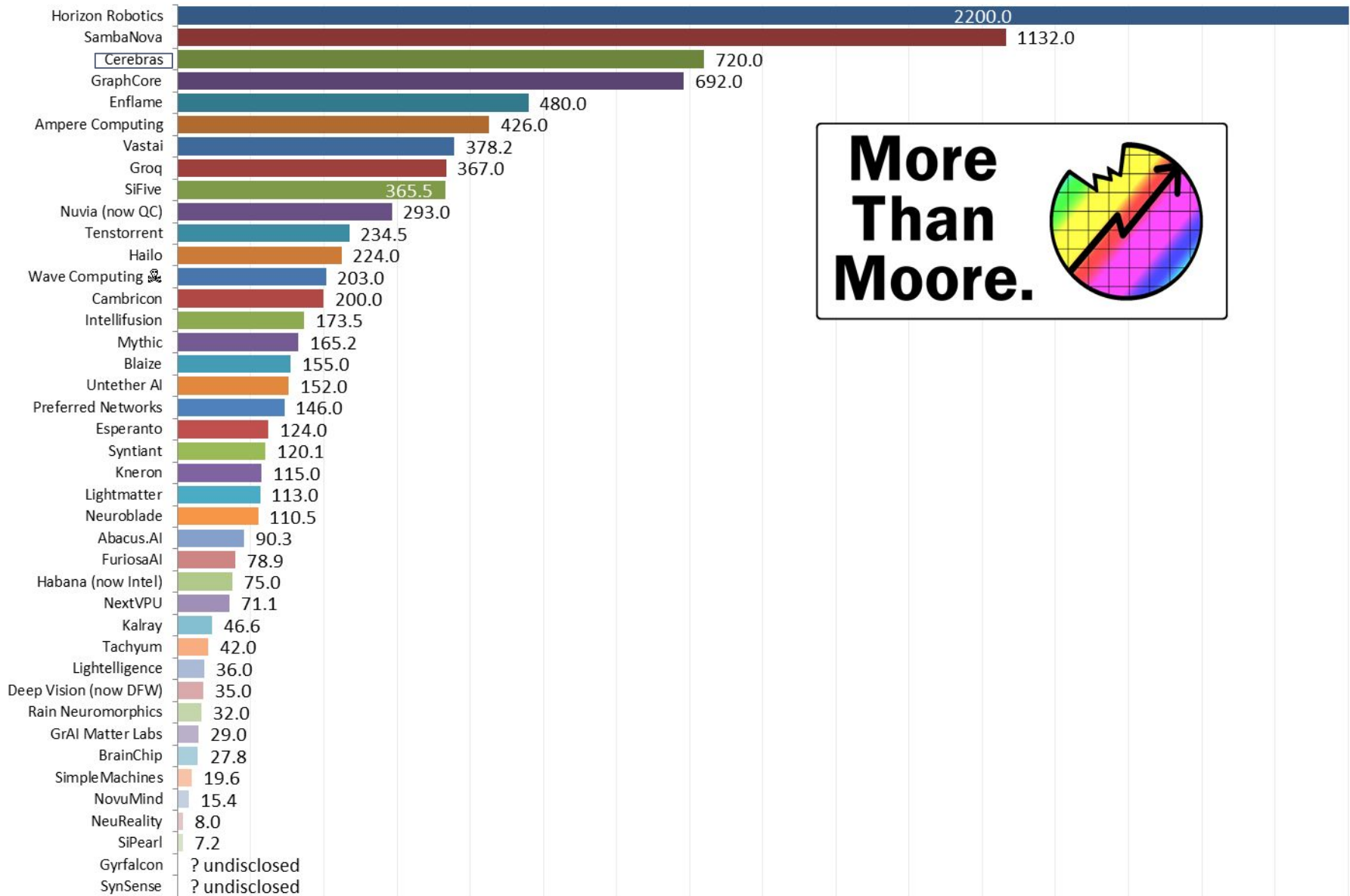
04 Apr 2023



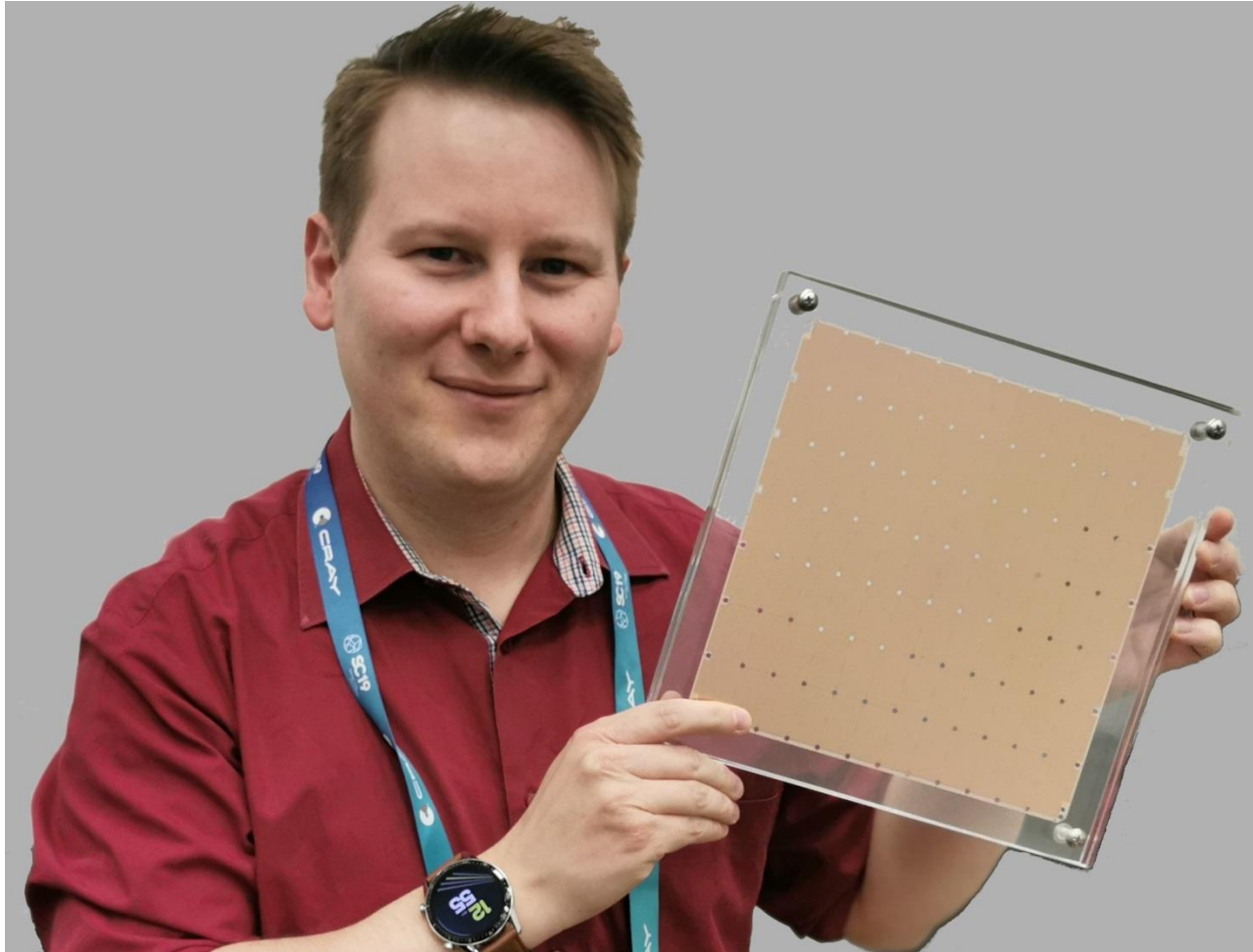
Watch now: [Inside NTT's Award-](#)

Source: DataCenter Dynamics

# AI Pure Play (+ others) Funding September 2022 - Values in USD \$m



# Cerebras Wafer Scale Engine



# Cerebras Wafer Scale Engine

## Cerebras WSE 2 The Largest Chip Ever Built

- 46,225 mm<sup>2</sup> silicon
- 2.6 trillion transistors
- 850,000 AI optimized cores
- 40 Gigabytes on chip memory
- 20 Petabytes memory bandwidth
- 220 Petabits fabric bandwidth
- TSMC 7nm





# Cerebras Wafer Scale Engine



# Cerebras Wafer Scale Engine



## New Cerebras Systems technology will double capacity, allow larger deep-learning models and data

The Neocortex high performance artificial intelligence (AI) computer at PSC has been upgraded with two new Cerebras [CS-2](#) systems, powered by the second-generation [wafer-scale engine](#) (WSE-2) processor. The WSE-2 doubles the system's cores and on-chip memory and offers a new execution mode with even greater advantages for extreme-scale deep-learning tasks, enabling faster training, larger models and larger input data.

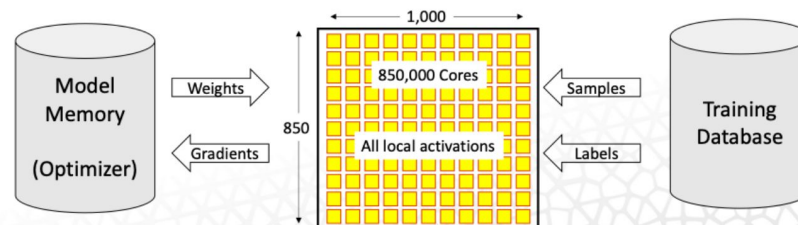
Neocortex, funded with \$11.25 million from the National Science Foundation to date, is supported under the NSF's Innovative HPC Program, meant to further the field of high performance computing (HPC) by funding new technologies with innovative approaches. The system now features a groundbreaking integration of two WSE-2's – an improved new technology that accelerates deep-learning AI with a unique chip architecture – with a powerful HPE Superdome Flex HPC server. By pairing the robustly provisioned HPE Superdome Flex server for massive data handling capability with the two WSE-2's, the system has unlocked new potential for rapidly training AI systems capable of learning from vast data sources.

"We are extremely excited to welcome the CS-2 servers into Neocortex," said Paola Buitrago, principal investigator of Neocortex and Director, Artificial Intelligence & Big Data at PSC. "This upgrade enhances support for new models, algorithms and research opportunities. We look forward to the breakthroughs that the now even greater capabilities of Neocortex would enable. We will continue working with the research community to help them take advantage of this technology that is orders of magnitude more powerful."

The CS-2 is based on the innovative WSE. The WSE-2 is the largest chip in existence and is the fastest AI processor. Whereas traditional processors are the size of postage stamps, the WSE-2 is the size of a dinner plate. In AI, big chips process information more quickly, producing answers in less time.

In deep learning, an AI program represents characteristics of a computational problem as layers, connected with each other by lines of inference. The AI first trains on data in which humans have labeled the "right answers," pruning or strengthening inference connections until it is predicting correctly. The researchers then test the AI against a dataset without such labels, to grade its performance. Finally, once the AI is performing adequately, it can be set to the task it has been designed to address.

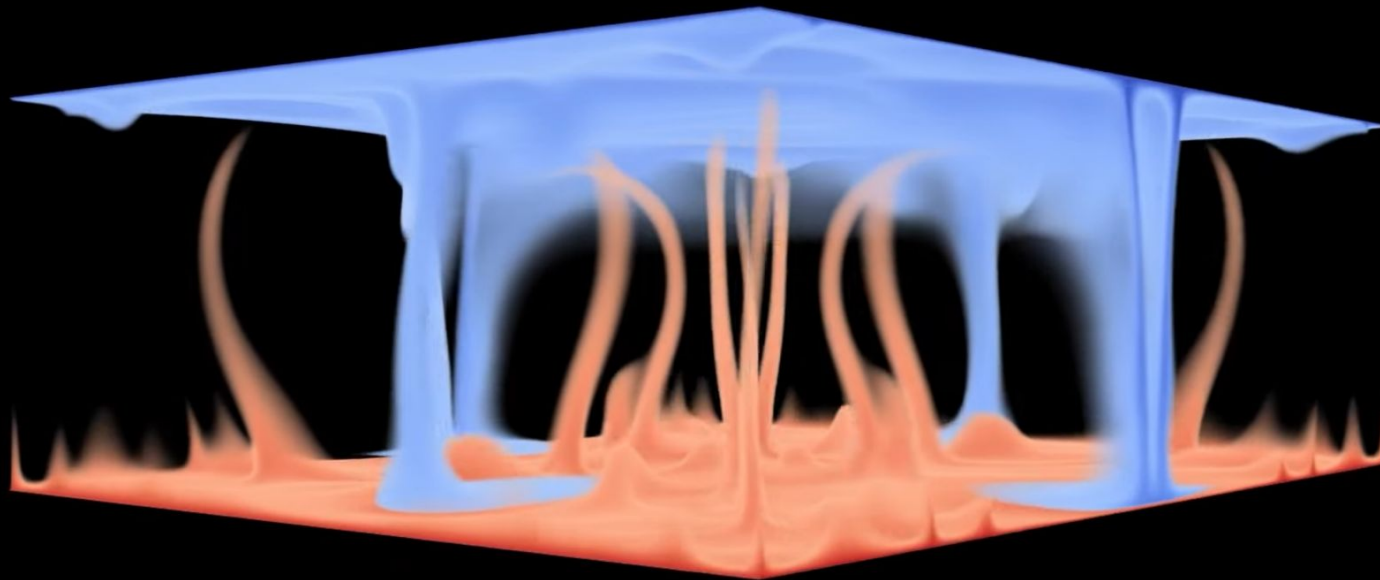
The two-dimensional grid of cores on the WSE-2 allows the system to route machine-learning tasks in physical space, essentially reproducing the layers of a deep-learning algorithm on different parts of the chip. By leveraging a 7-nm fabrication process, the CS-2 improves upon the CS-1's capabilities by expanding the number of cores from 400,000 to 850,000 and on-chip memory from 18 GB to 40 GB. The CS-2 does this with the same footprint, power, and cooling requirements as the CS-1.



# Cerebras Wafer Scale Engine



National Science Foundation



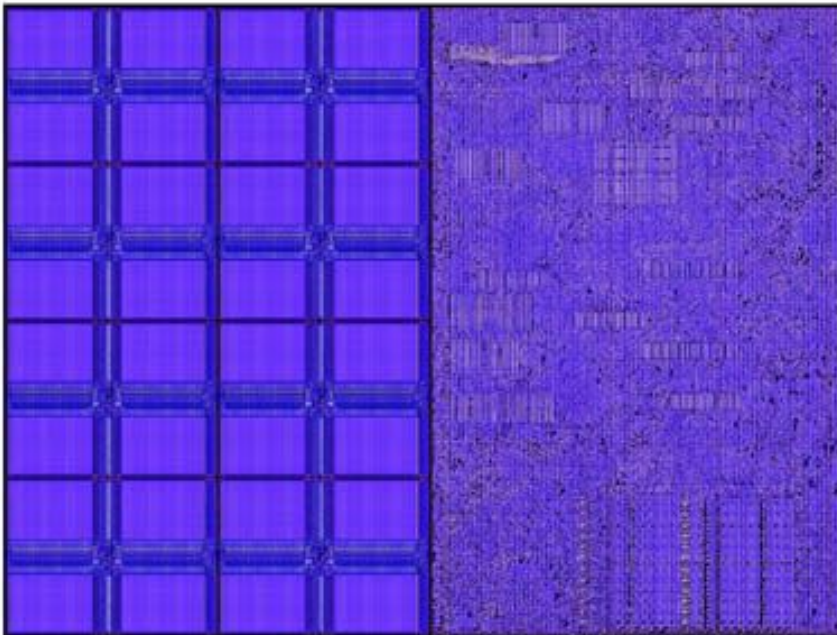
Simulation performed on a single Cerebras CS-2  
within the Neocortex system  
at the Pittsburgh Supercomputing Center

736 x 896 x 300 cells (198 million)  
Fluid volume of 23 x 28 x 9.4 meters  
Video playback rate is approximately at actual solution speed



# Cerebras Wafer Scale Engine

## Core Design



### Efficient small core design

- 228 $\mu\text{m}$  x 170 $\mu\text{m}$  core area
- TSMC N7

### Balanced logic and memory

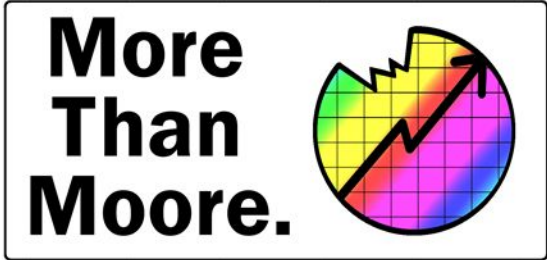
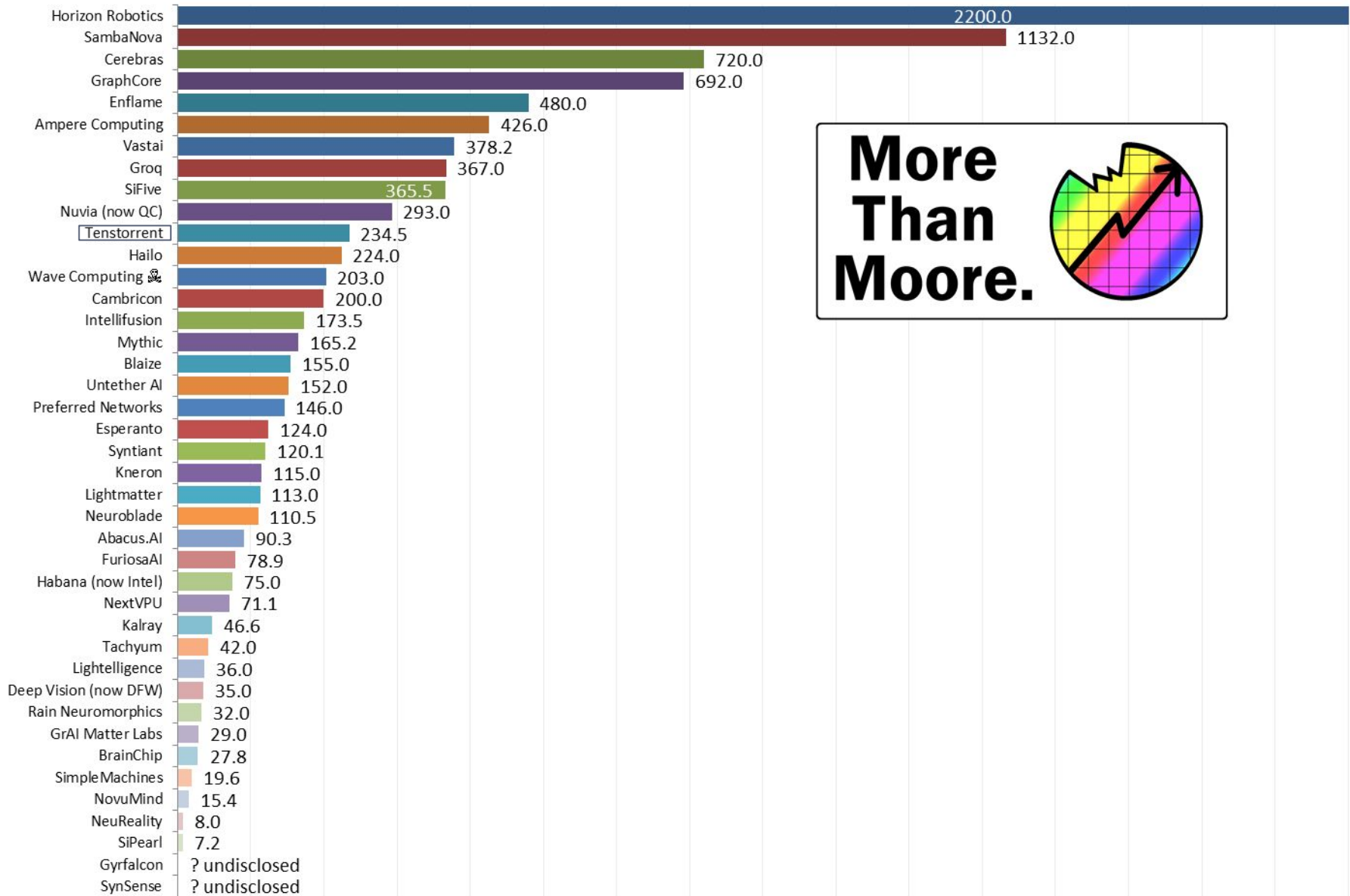
- 50:50 logic to SRAM area ratio
- 110,000 logic standard cells
- 48kB high density SRAM memory

### Power efficient design point

- 1.1GHz clock frequency
- 30mW peak power



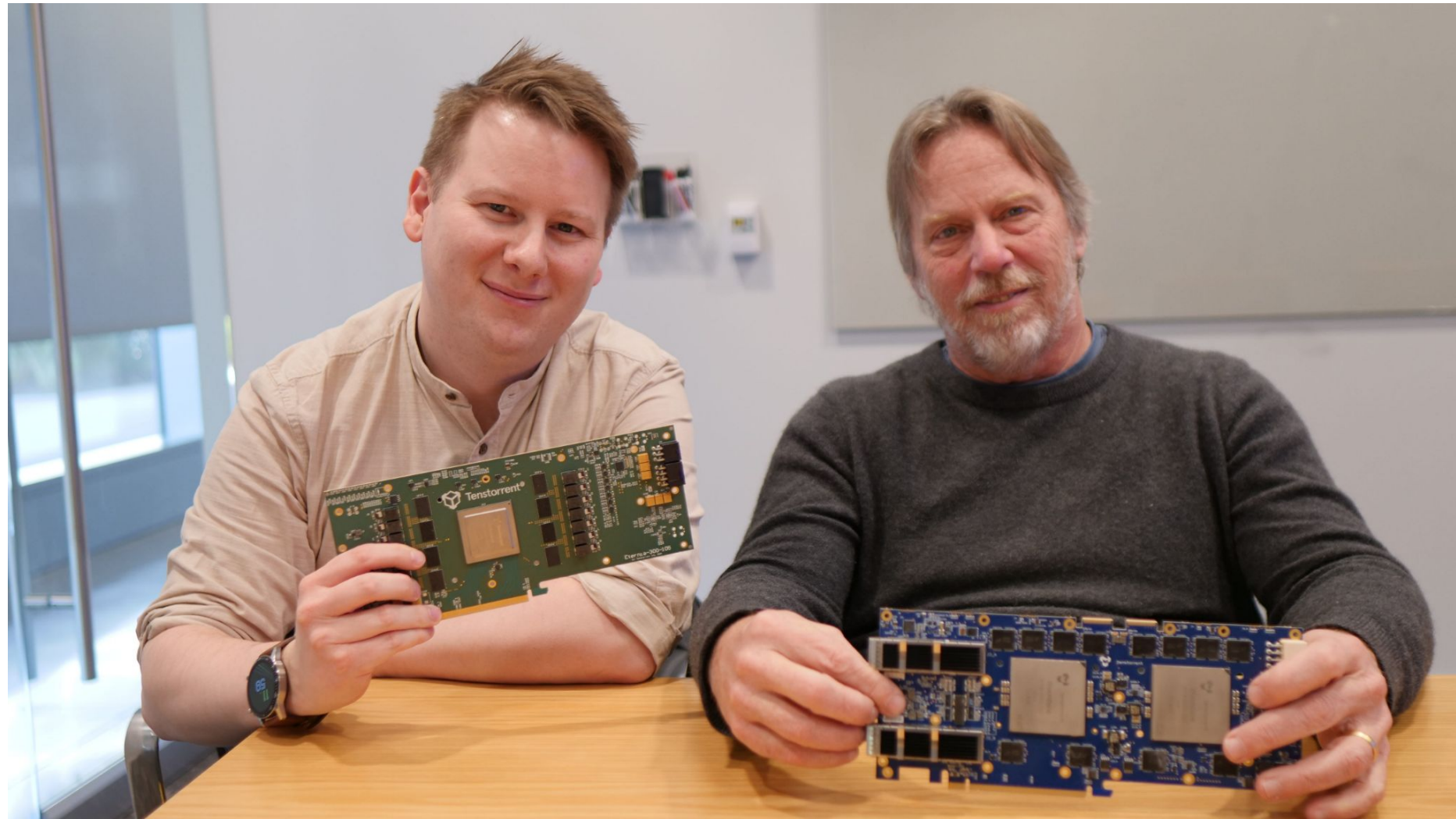
# AI Pure Play (+ others) Funding September 2022 - Values in USD \$m



# Tenstorrent



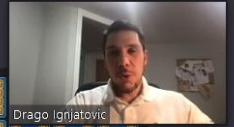
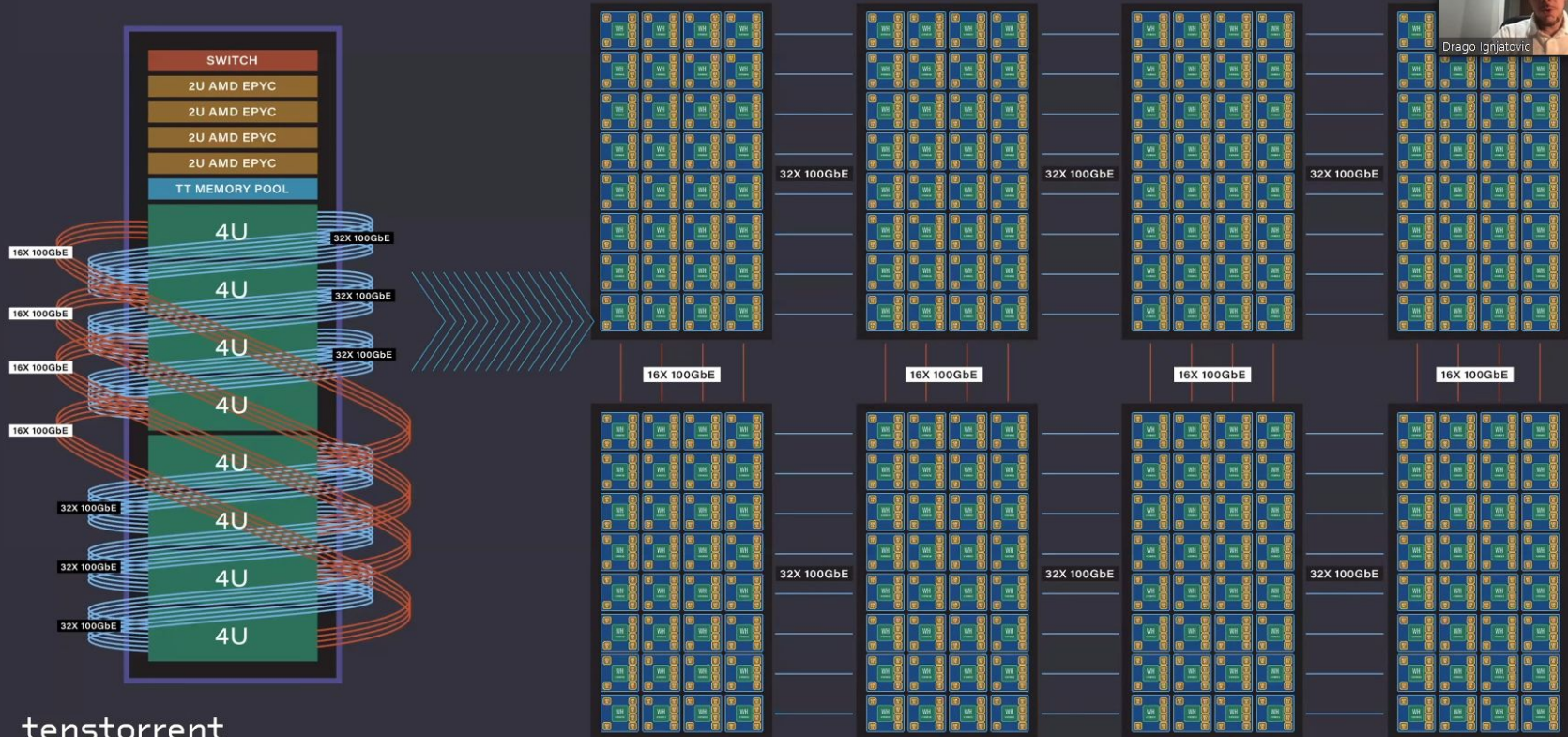
# Tenstorrent





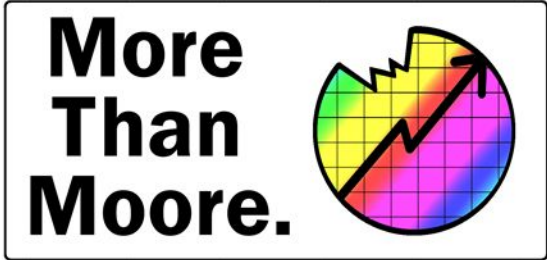
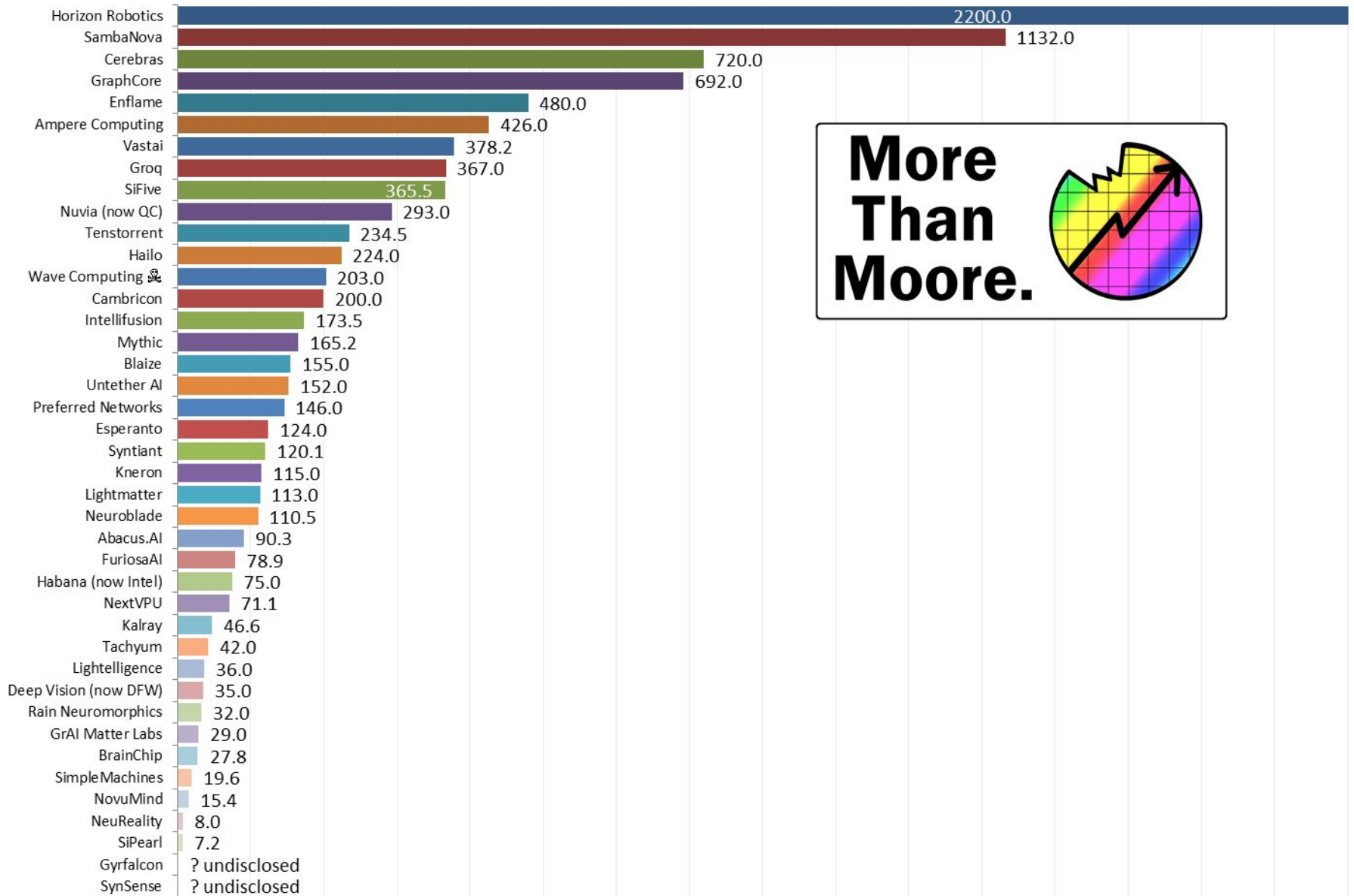
# Tenstorrent

## Galaxy – Supercomputer Topology





# AI Pure Play (+ others) Funding September 2022 - Values in USD \$m



# Silicon or Survive

- A New HPC Era
  - Types of Legacy Hardware
  - New Paradigms: Analog, Neuro, Quantum, Optical
  - Push for Low Precision
- AI Hardware
  - Established Players
  - Current \$10B Market
  - Case Studies: Wafer Scale, Analog Edge
  - Roadmaps
  - Software
- Q&A

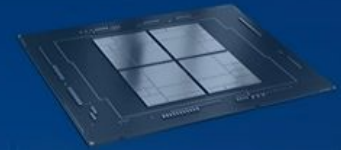
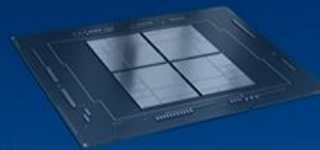
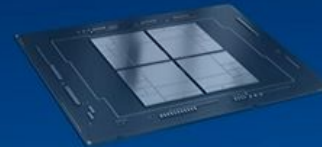
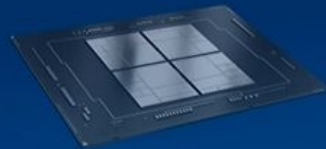
# Intel DCAI

## Executing on Our Xeon Roadmap

intel.  
XEON

CPU P-Core

CPU E-Core



4th Gen Intel® Xeon® Scalable processors

5th Gen Intel® Xeon® codenamed Emerald Rapids

Next-Gen Intel® Xeon® codenamed Sierra Forest

Next-Gen Intel® Xeon® codenamed Granite Rapids

Next-Gen Intel® Xeon® codenamed Clearwater Forest

Today

Q4 2023

2024  
(First Half)

2024  
(closely following  
Sierra Forest)

2025

# Intel DCAI

## DCAI Architecture Evolution

### CPU P-Core



4th Gen Intel® Xeon® Scalable processors

Intel® Xeon® CPU Max Series



5th Gen Intel® Xeon®  
codenamed Emerald Rapids



Intel® Xeon® Processors  
codenamed Granite Rapids

### CPU E-Core



Intel® Xeon® Processor  
codenamed Sierra Forest



Intel® Xeon® Processor  
codenamed Clearwater Forest

### GPU



Intel® Data Center GPU Flex Series  
codenamed Arctic Sound-M



Intel® Data Center GPU Max Series  
codenamed Ponte Vecchio



Intel® Data Center GPU Flex Series  
codenamed Melville Sound

Next-Generation Accelerator  
Architecture  
codenamed Falcon Shores

### Dedicated AI



Habana®  
Gaudi® 2



Habana®  
Gaudi® 3

Next-Generation Accelerator  
Architecture

### FPGA



15 new FPGAs on  
schedule to PRQ in 2023



Next Gen  
FPGAs

Roadmap: 2023-2025

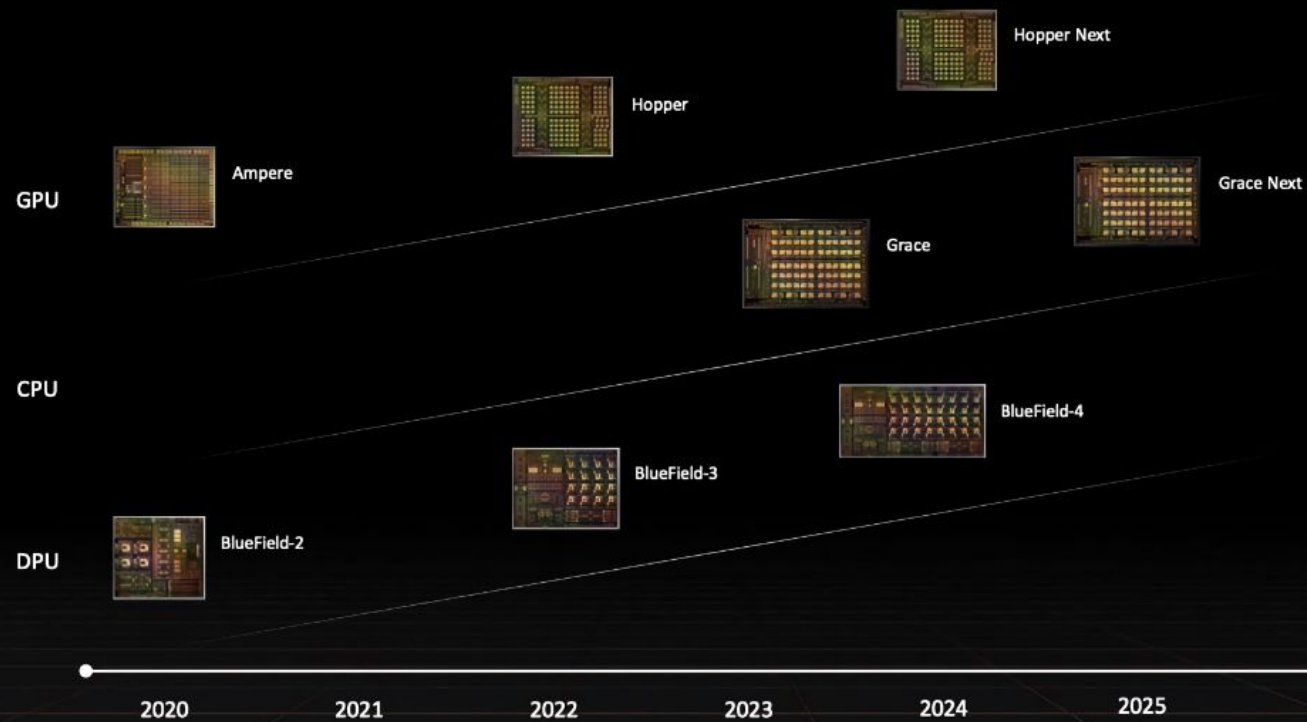


# Intel DCAI

Continuous, visible data points will provide confidence that we are **rebuilding our execution engine**



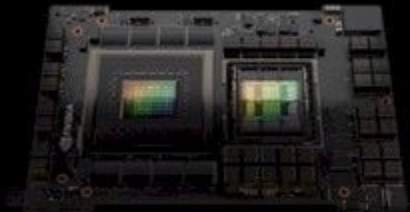
# NVIDIA



EXECUTING AT THE SPEED OF LIGHT

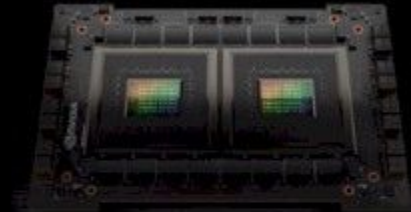
# NVIDIA

## GRACE HOPPER SUPERCHIP For Giant-Scale AI and HPC



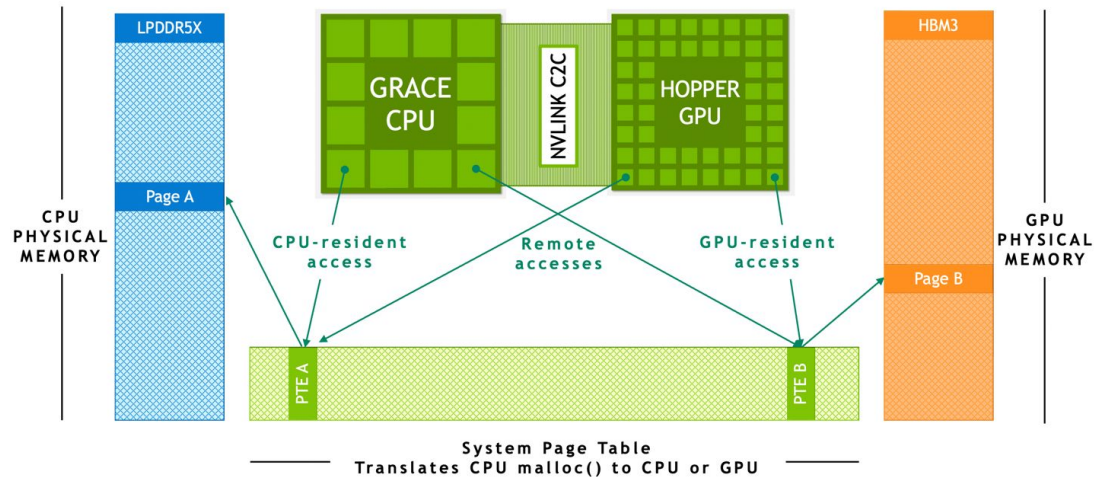
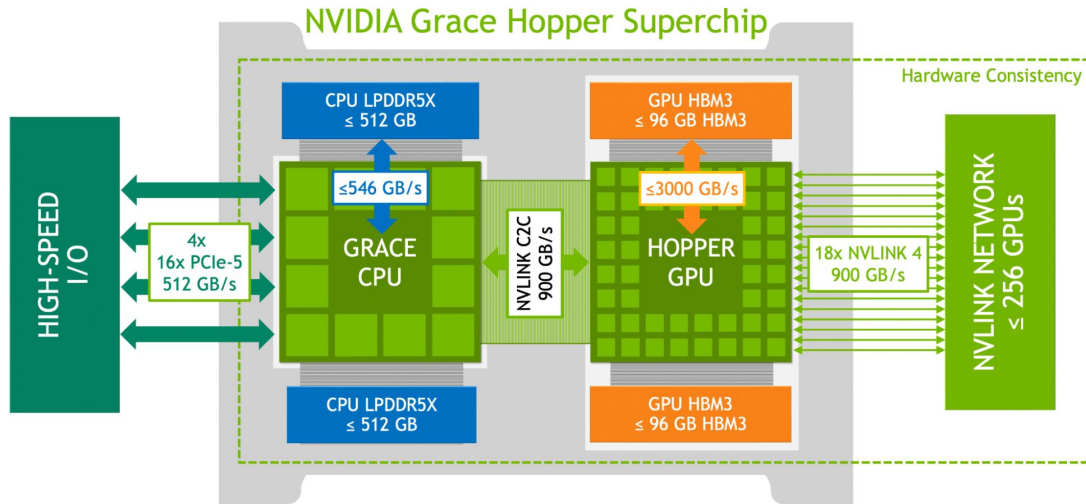
- CPU+GPU Designed For Giant-Scale AI and HPC
- 600GB Memory GPU for Giant Models
- New 900 GB/s Coherent Interface
- 30X Higher System Memory B/W to GPU in a Server
- Runs NVIDIA Computing Stacks
- Available 1H 2023

## GRACE CPU SUPERCHIP For HPC and AI Infrastructure



- High Performance CPU for HPC and AI
- 144 Cores | 740 SPECrate@2017\_int\_base est.
- First LPDDR5X Memory With ECC. 1TB/s Memory Bandwidth
- 2X Perf/Watt Over Traditional Servers
- Runs NVIDIA Computing Stacks
- Available 1H 2023

# NVIDIA





# AMD CPU

## INDUSTRY LEADING OPTIMIZED SILICON



# AMD GPU



AMD Instinct™ **MI100**  
AMD CDNA™

## Ecosystem Growth

First purpose-built GPU architecture for the data center



AMD Instinct™ **MI200**  
AMD CDNA™ 2

## Driving HPC and AI to a New Frontier

First multi-die data center GPU expands scientific discovery and brings choice to AI training



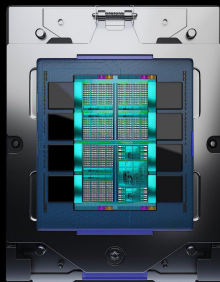
AMD Instinct™ **MI300**  
AMD CDNA™ 3

## Data Center APU

Breakthrough architecture designed for leadership efficiency and performance for HPC and AI

2020

2023



The world's first integrated data center CPU + GPU

AMD INSTINCT™  
**MI300**

Breakthrough architecture to power the exascale AI era

## AMD CDNA 3 THE JOURNEY CONTINUES

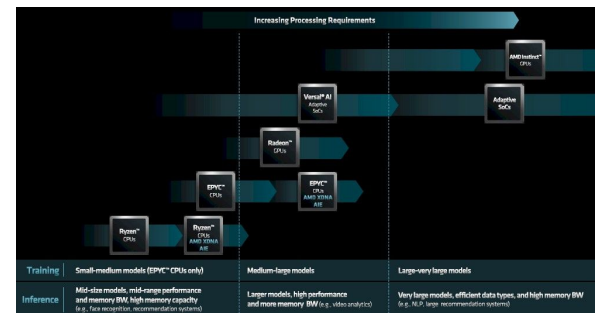
AI Performance/Watt Uplift

>5X

AMD CDNA 2 AMD CDNA 3

Expected performance-per-watt uplift through:

- 5nm Process and 3D Chiplet Packaging
- Next-Gen AMD Infinity Cache™
- 4th Gen Infinity Architecture
- Unified Memory APU Architecture
- New Math Formats



## Rapid Pace of Innovation



# Tenstorrent

## Chip Roadmap



ML Processor

Networked ML Processor

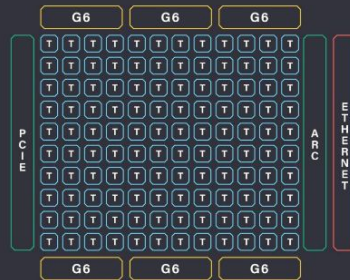
Standalone ML computer

Custom high-speed RISC V



**Grayskull**  
(12nm, 620mm<sup>2</sup>)

315 8b TFLOPS  
PCI E gen 4  
8 channels LPDDR4



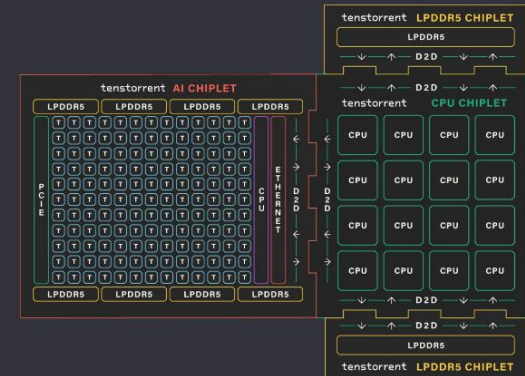
**Wormhole**  
(12nm, 650 mm<sup>2</sup>)

350 8b TFLOPS  
400GB/sec ethernet  
6 channels GDDR6  
16 lanes of PCI E gen 4



**Black Hole**  
(6nm, 600mm<sup>2</sup>)

1000 8b TFLOPS  
1200GB/sec ethernet  
8 channels of GDDR6X  
32 lanes of PCI E Gen5  
2TB/sec die to die interface



**Grendel**  
(3nm Ascalon Chiplet with  
6nm Blackhole Chiplet)

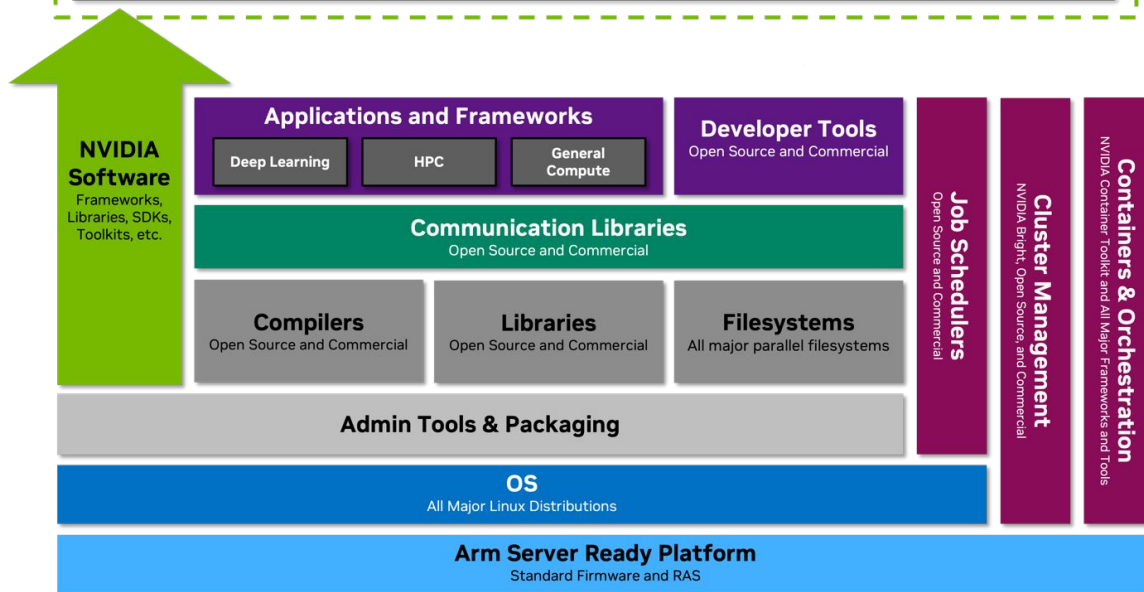
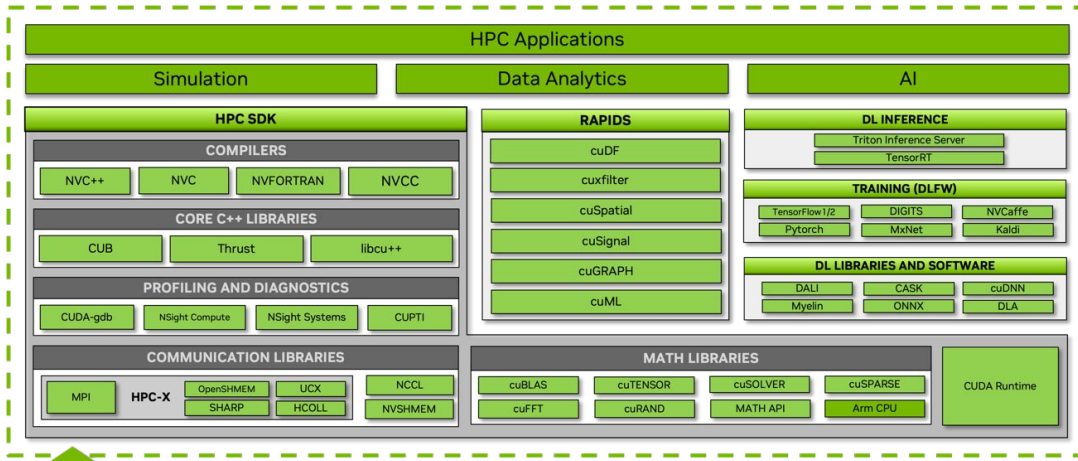
Black Hole chiplet  
3nm CPU chiplet with 128 RISC V cores  
Two chiplets with 4 channels of LPDDR5 each



tenstorrent

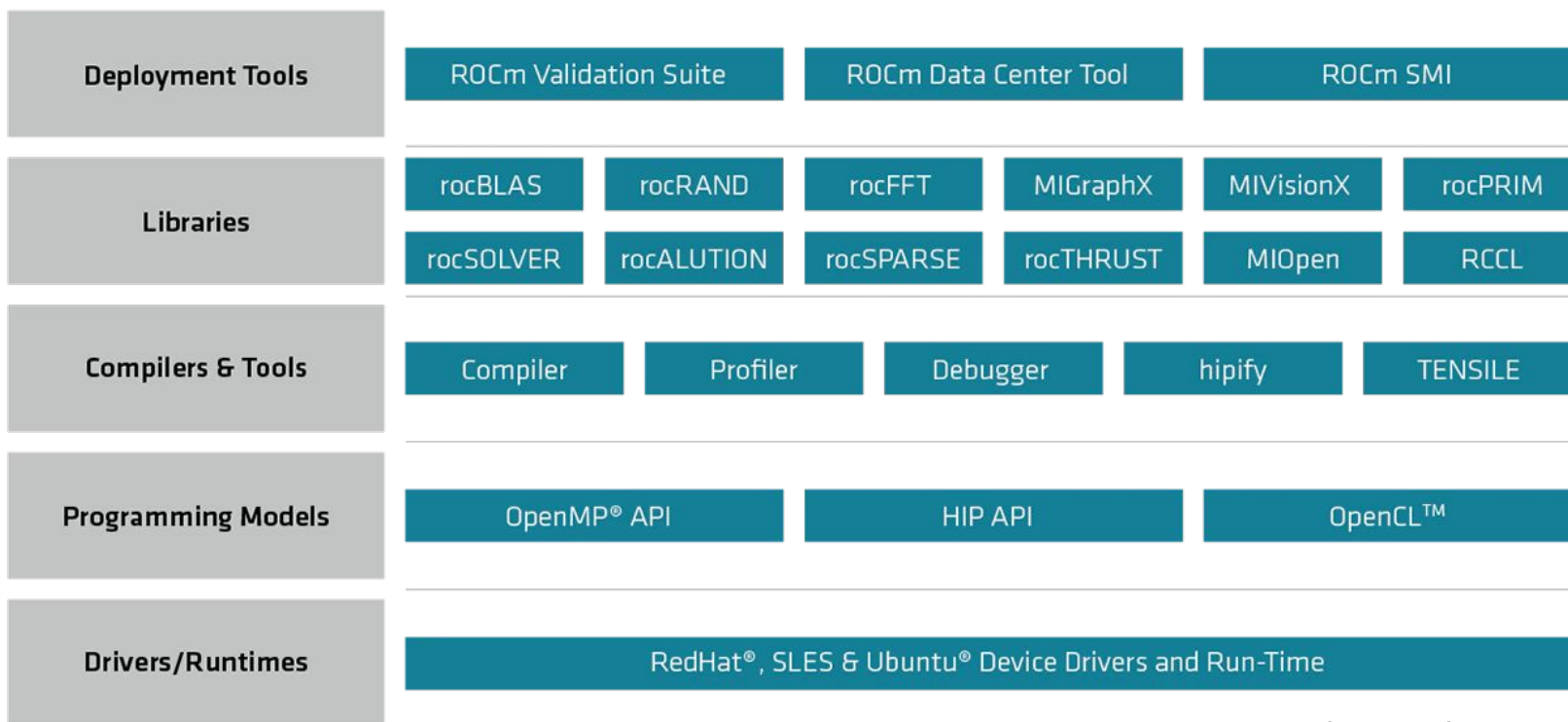


# Software: NVIDIA CUDA 12.1.0



- Full support for x86, ARM, even CPUs and Windows
- CUDA Math Libraries, cuDNN, cuFFT, cuBLAS, NAMD, CFD
- Multi-user virtualization support for HPC + Cloud

# Software: AMD ROCm 5.4



<https://docs.amd.com/>

- Translates CUDA code to AMD via hipify
- OpenMP/OpenCL support, Linux (RHEL/Ubuntu), ESXi 7/8
- Extensive release notes with each version

# Software: Intel OneAPI

## oneAPI Industry Initiative

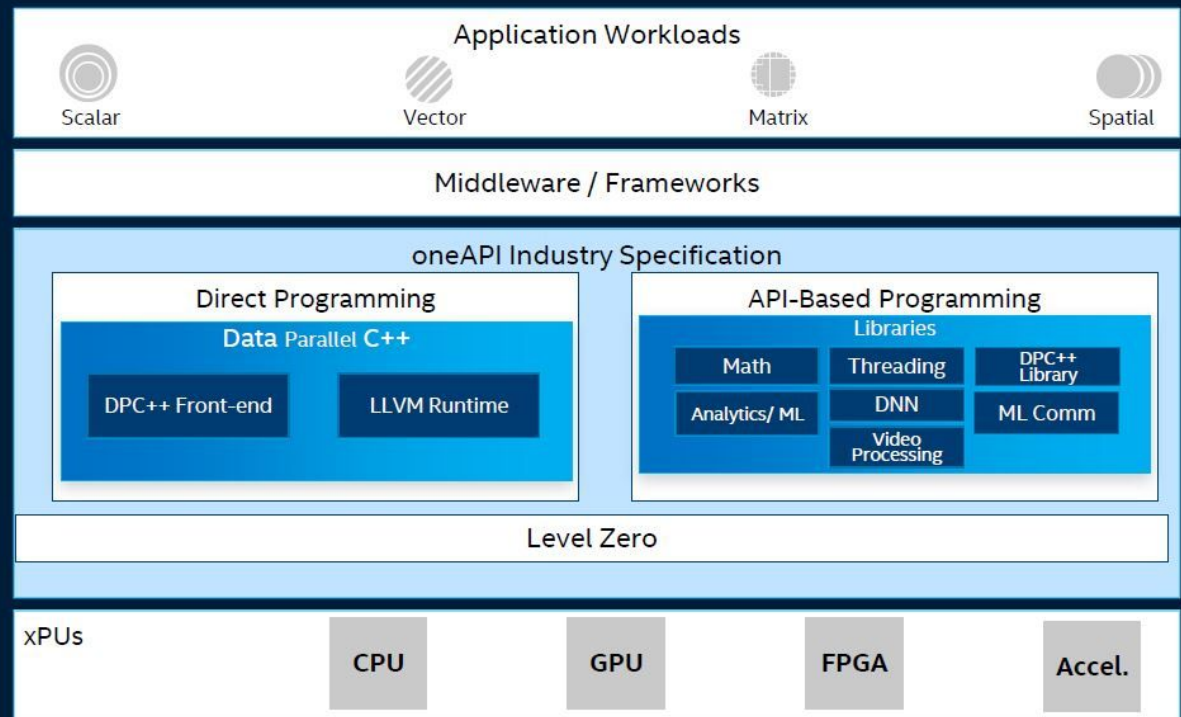
### Specifies:

- A standards based cross-architecture language, DPC++, based on C++ and SYCL
- Powerful APIs designed for acceleration of key domain-focused functions
- Low-level hardware interface to provide a hardware abstraction layer to vendors

Open standard to promote community and industry support

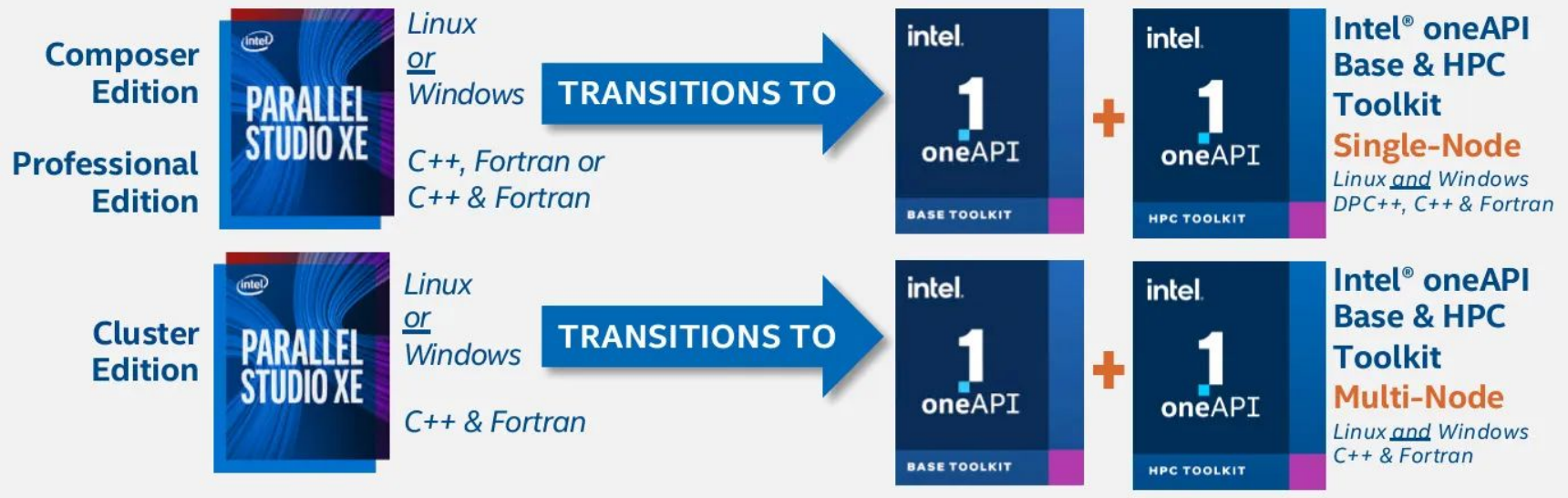
Enables code reuse across architectures and vendors

Available now



# Software: Intel oneAPI

## Transitioning to new Editions



- Attempt to unify industry around DPC++ / SYCL
- Write once, compile often? Hardware agnostic?
- A proper ground-up redesign of Intel's software support



# Software: Intel oneAPI

## Intel® oneAPI Base & HPC Toolkit

### Direct Programming

Intel® C++ Compiler Classic

Intel® Fortran Compiler (Beta)

Intel® Fortran Compiler Classic

Intel® oneAPI DPC++/C++ Compiler

Intel® DPC++ Compatibility Tool

Intel® Distribution for Python\*

Intel® FPGA Add-On for oneAPI Base Toolkit

### API-Based Programming

Intel® MPI Library

Intel® oneAPI DPC++ Library

Intel® oneAPI Math Kernel Library

Intel® oneAPI Data Analytics Library

Intel® oneAPI Threading Building Blocks

Intel® oneAPI Video Processing Library

Intel® oneAPI Collective Communications Library

Intel® oneAPI Deep Neural Network Library

Intel® Integrated Performance Primitives

### Analysis & Debug Tools

Intel® Inspector

Intel® Trace Analyzer & Collector

Intel® Cluster Checker

Intel® VTune™ Profiler

Intel® Advisor

Intel® Distribution for GDB\*

# What I didn't cover in this talk

- Interconnect
  - Infiniband vs Ethernet vs GPU-to-GPU
  - Integrated Optics/Photonics
  - New paradigms like CXL
- Storage
  - Rise and fall of Intel/Micron's Optane
  - HBM vs DDR vs Compute-in-Memory
- China
  - Did they really have the first exaflop supercomputer?

# The AI Hardware Show

## THE AI HARDWARE SHOW

2023 Episode 1



**Google TPU**  
**NVIDIA A100**  
**IBM AIU**  
**Biren BR100**  
**AMD MI250X**  
**EdgeQ**

# Silicon or Survive

- A New HPC Era
  - Types of Legacy Hardware
  - New Paradigms: Analog, Neuro, Quantum, Optical
  - Push for Low Precision
- AI Hardware
  - Established Players
  - Current \$10B Market
  - Case Studies: Wafer Scale, Analog Edge
  - Roadmaps
  - Software
- Q&A



# TechTechPotato Mugs



<http://merch.techtechpotato.com>

First 50 orders  
20% OFF with code **EUM23**