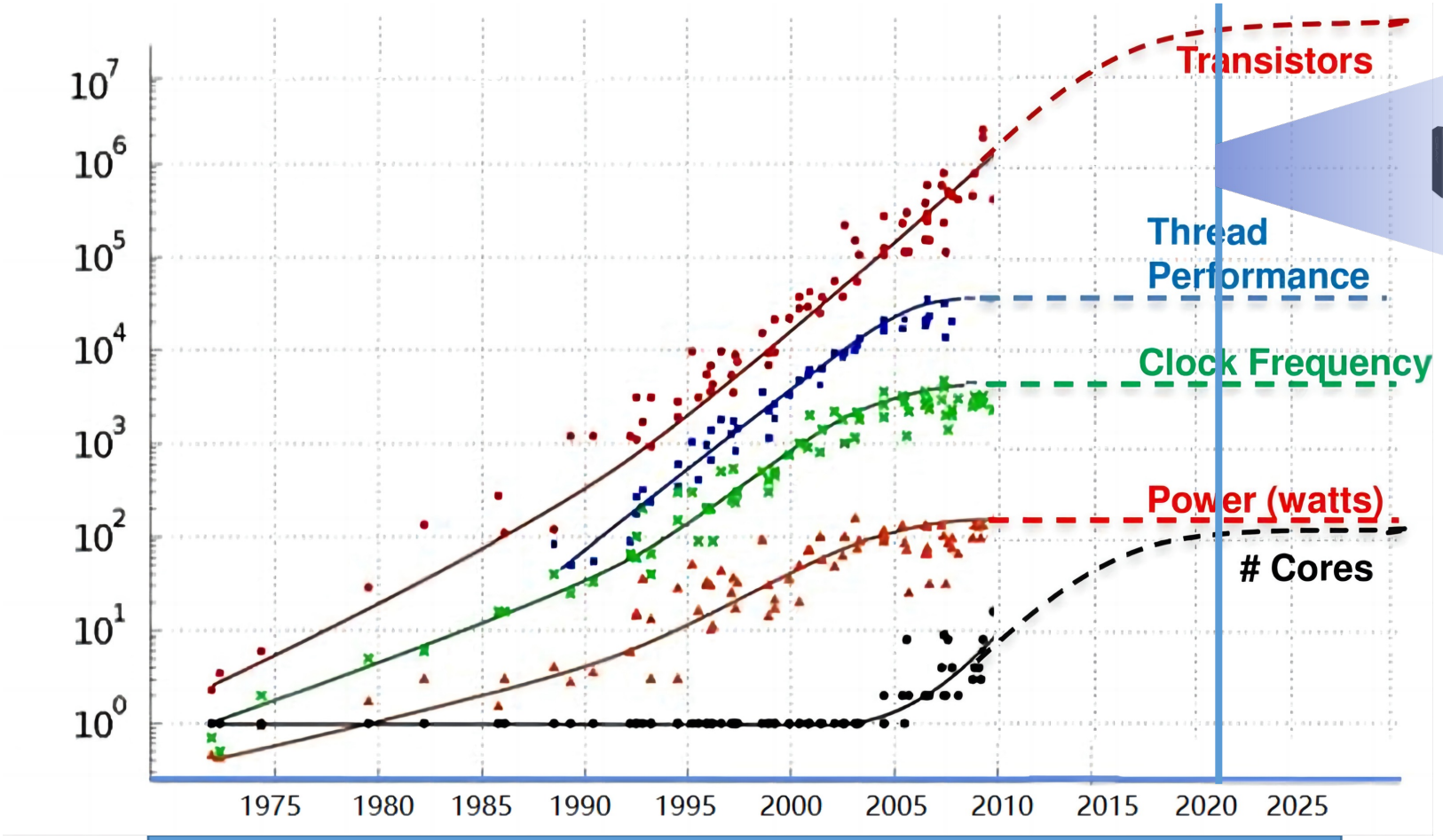
Architectural Design and Trade-Offs on the Path to Exascale

Ruibo Wang
National University of Defense Technology, China
Sept. 2023 Russian Supercomputing Days

Technology is not well scaling

Exascale achieved, how to get higher performance?

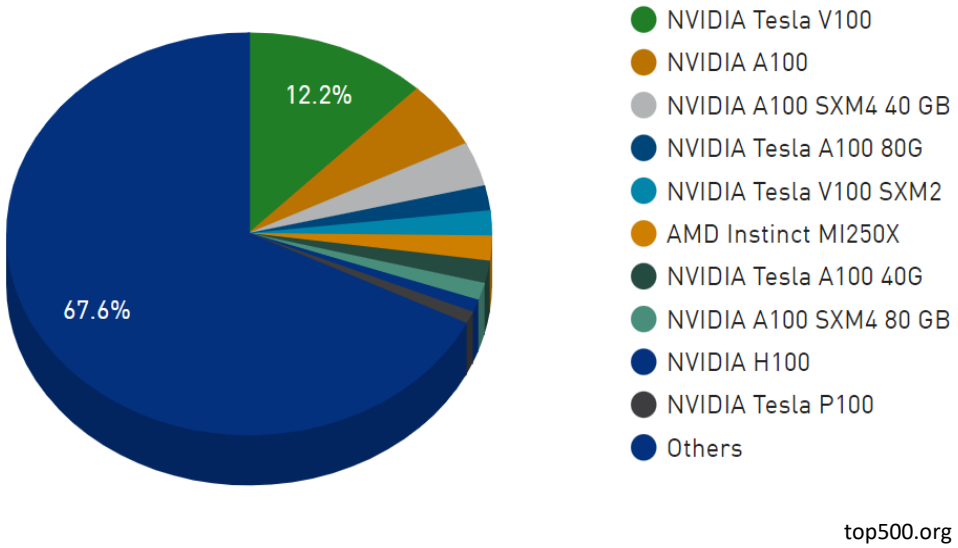


Running out of steam on every front
Call for architectural innovation!

1. Figure from Kunle Olukotun, Lance Hammond, Herb Sutter, Mark Horowitz

Homogeneous Architecture

Homogenous architecture refers to the systems which use only one type of processor or core (**mainly CPU**)



top500.org

The **majority** of systems on the TOP500 list , **particularly the smaller ones still adopt homogenous architecture** where only CPUs are used for computing.

Fugaku from RIKEN and Fujistu is so far the fastest homogenous supercomputer, **537Pflops** achieved with **~158K** compute nodes.

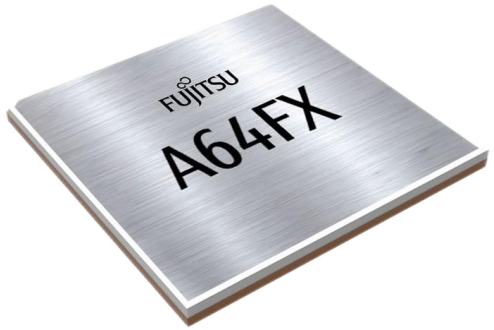
Using Homogenous Architecture for Exascale is challenging

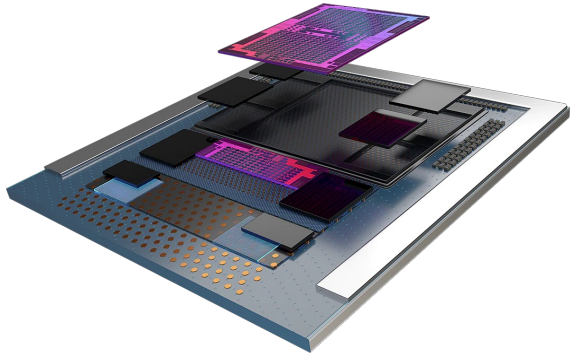
Power Problem

Assume a ~10x performance boost for each generation, how can we achieve that with a **reasonable power budget**?
CPUs alone is far from enough.
Most GPUs (wide vector processors) may even not be enough either.

Scalability Problem

Existing pre-exascale Fugaku has over 150K compute nodes.
Managing such a large num of nodes itself is very challenging.
Designing **scalable interconnect** is hard for such a scale





Heterogenous architecture seems to be a must for Exascale and beyond

The early days of heterogenous supercomputers

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Jaguar** - Cray XT5-HE Opteron 6-core 2.6 GHz, Cray/HPE DOE/SC/Oak Ridge National Laboratory United States	224,162	1,759.00	2,331.00	6,950
2	**Roadrunner** - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband, IBM DOE/NNSA/LANL United States	122,400	1,042.00	1,375.78	2,345
3	**Kraken XT5** - Cray XT5-HE Opteron 6-core 2.6 GHz, Cray/HPE National Institute for Computational Sciences/University of Tennessee United States	98,928	831.70	1,028.85	3,090
4	**JUGENE** - Blue Gene/P Solution, IBM Forschungszentrum Juelich (FZJ) Germany	294,912	825.50	1,002.70	2,268
5	**Tianhe-1** - NUDT TH-1 Cluster, Xeon E5540/E5450, ATI Radeon HD 4870 2, Infiniband, NUDT National SuperComputer Center in Tianjin/NUDT China	71,680	563.10	1,206.19	
6	**Pleiades** - SGI Altix ICE 8200EX, Xeon QC 3.0 GHz/Nehalem EP 2.93 Ghz, HPE NASA/Ames Research Center/NAS United States	56,320	544.30	673.26	2,348
7	**BlueGene/L** - eServer Blue Gene Solution, IBM DOE/NNSA/LLNL United States	212,992	478.20	596.38	2,329
8	**Intrepid** - Blue Gene/P Solution, IBM DOE/SC/Argonne National Laboratory United States	163,840	458.61	557.06	1,260
9	**Ranger** - SunBlade x6420, Opteron QC 2.3 Ghz, Infiniband, Oracle Texas Advanced Computing Center/Univ. of Texas United States	62,976	433.20	579.38	2,000
10	**Red Sky** - Sun Blade x6275, Xeon X55xx 2.93 Ghz, Infiniband, Oracle Sandia National Laboratories / National Renewable Energy Laboratory United States	41,616	423.90	487.74	

- **In the TOP500 list in Nov. 2009**
- Among the TOP10 systems in the list, only **two** use accelerators.
  - **Roadrunner** uses IBM PowerXCell processors together with AMD Opteron DC cores. PowerXCell acts similar to modern accelerators and contribute to the majority of performance. Opteron runs the OS
  - **Tianhe-1** uses Intel Xeon CPUs and ATI Readon GPUs. It is **the first** supercomputer (get into TOP5) to propose the heterogenous architecutre of **CPU+GPU**

The early days of heterogenous supercomputers

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,824,768	238.70	304.47	7,404
5	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
6	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
7	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371
8	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70.87	93.75	2,589
9	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63.46	79.22	2,646
10	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT National Super Computer Center in Guangzhou China	4,981,760	61.44	100.68	18,482

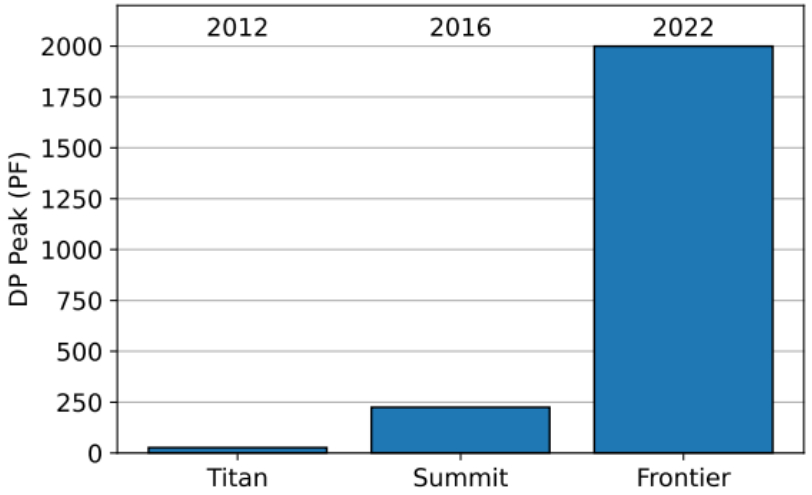
TOP500
The List.

- **In the TOP500 list in Jun. 2023**
- Out of the TOP10 systems in the list, **9 systems adopt the heterogenous architecture**.
  - i.e., Frontier, LUMI, Leonardo, Summit, Sierra, Sunway TaihuLight, Perlmutter, Selene and Tianhe-2A
  - Fugaku is the **only** system which belongs to the homogenous architecture
  - **Tianhe-1A CPU+GPU design** dominates (7/9 in the TOP10) the modern heterogenous supercomputers

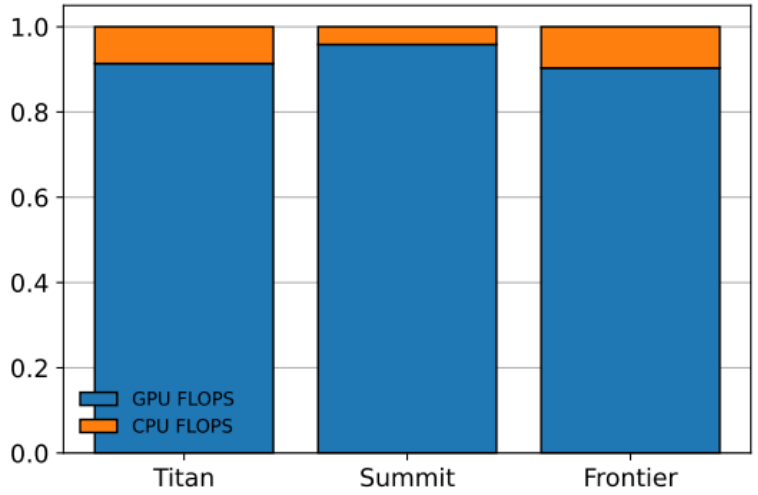
Heterogenous Architecture

- Since Tianhe-1A **first adopted** the CPU+GPU architecture (get No.1 in TOP500) in 2010, it is now the defacto standard for advanced high-performance computers.
- GPUs contribute to the **majority share of performance** for modern HPC systems.
- **Frontier** uses the same architecture to reach exascale.

Tianhe-1A supercomputer
Vendor: NUDT



peak performance[1]



performance share by device[1]

1. Tom Evans. Exascale Computing at ORNL Past, Current, and Future: Opportunities for High Energy Physics

Heterogenous Architecture

- Similar to Frontier, the other onging exascale US systems also use GPUs as the accelerators.

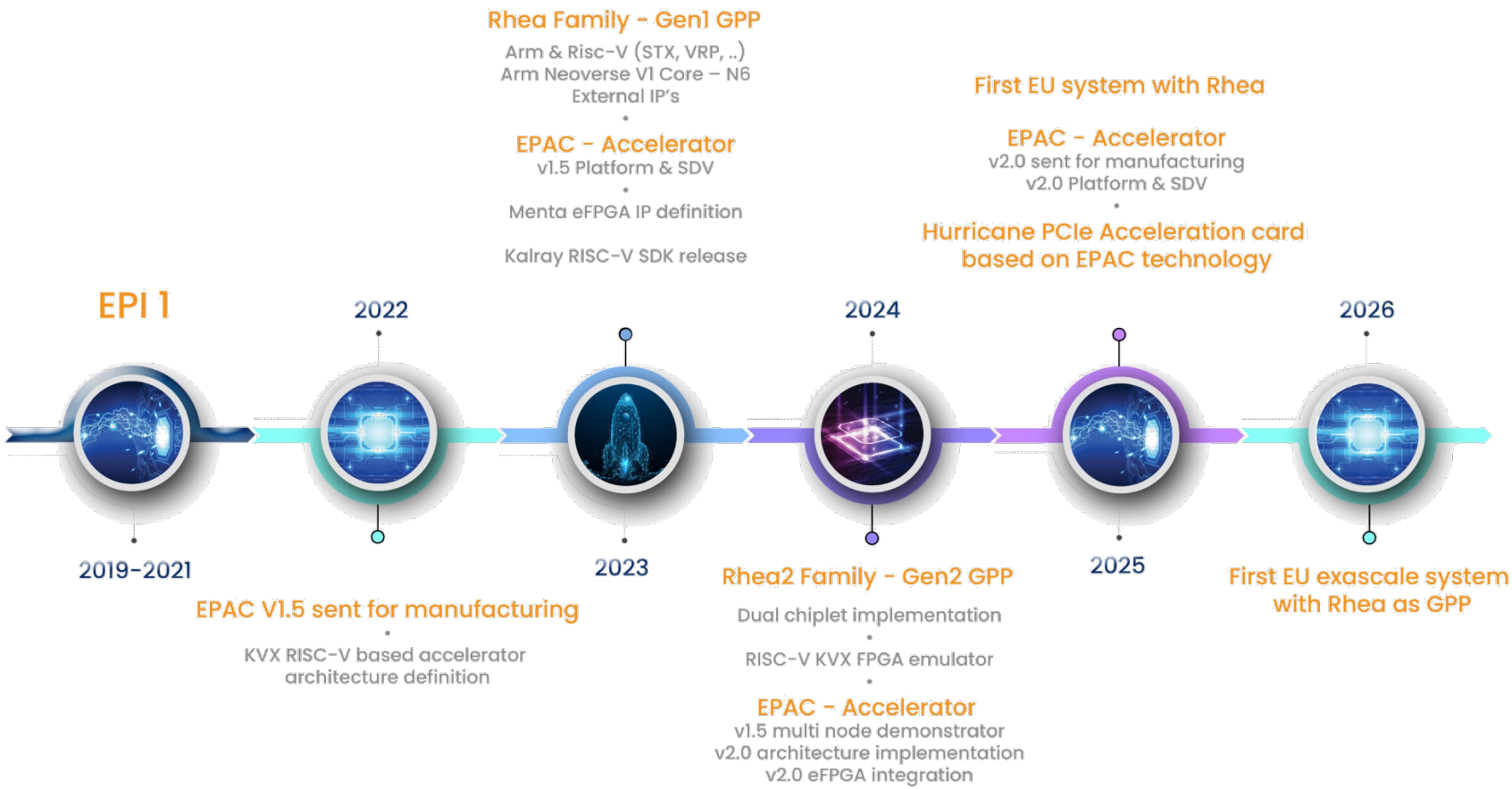
Frontier (AMD Instinct MI250X)

El Capitan (AMD Instinct MI300X)

Aurora(Intel Data Center GPU Max Series)

Heterogenous Architecture

- EuroHPC is leading the development of both ARM CPUs (Rhea series) and accelerators (RISC-V based) to enable Exascale computing for Europe.

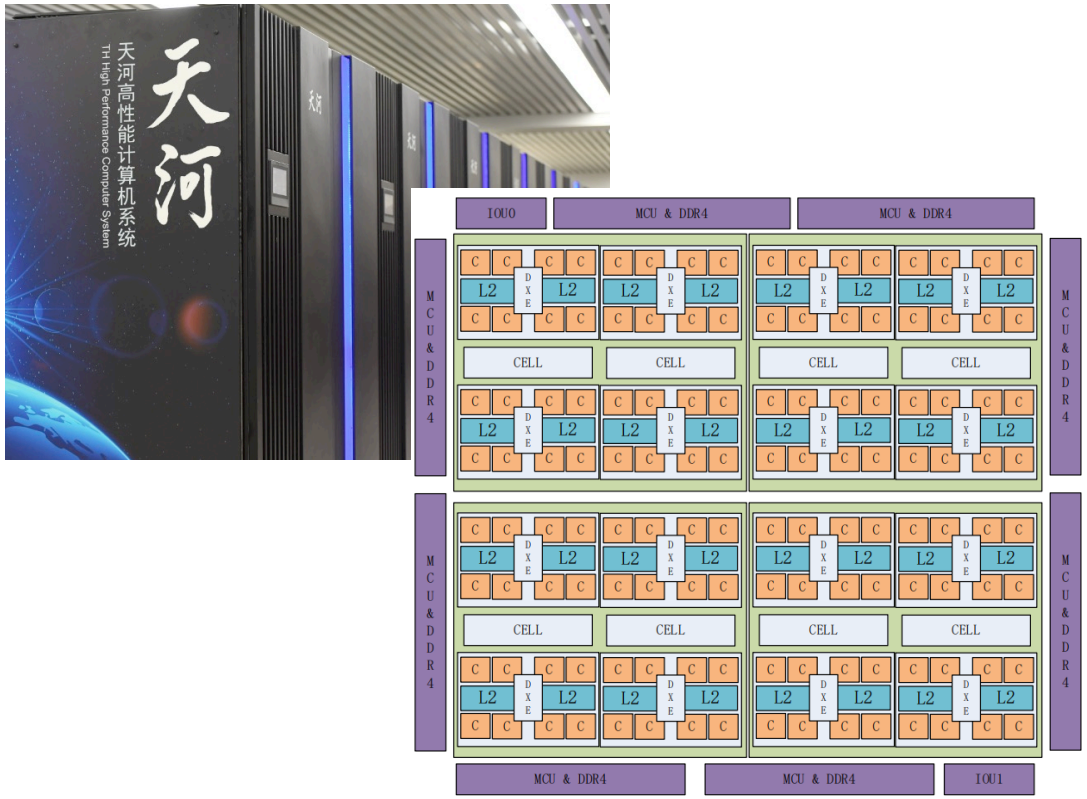


The European Processor Initiative (EPI) project roadmap[1]

1. https://www.european-processor-initiative.eu/project/epi/

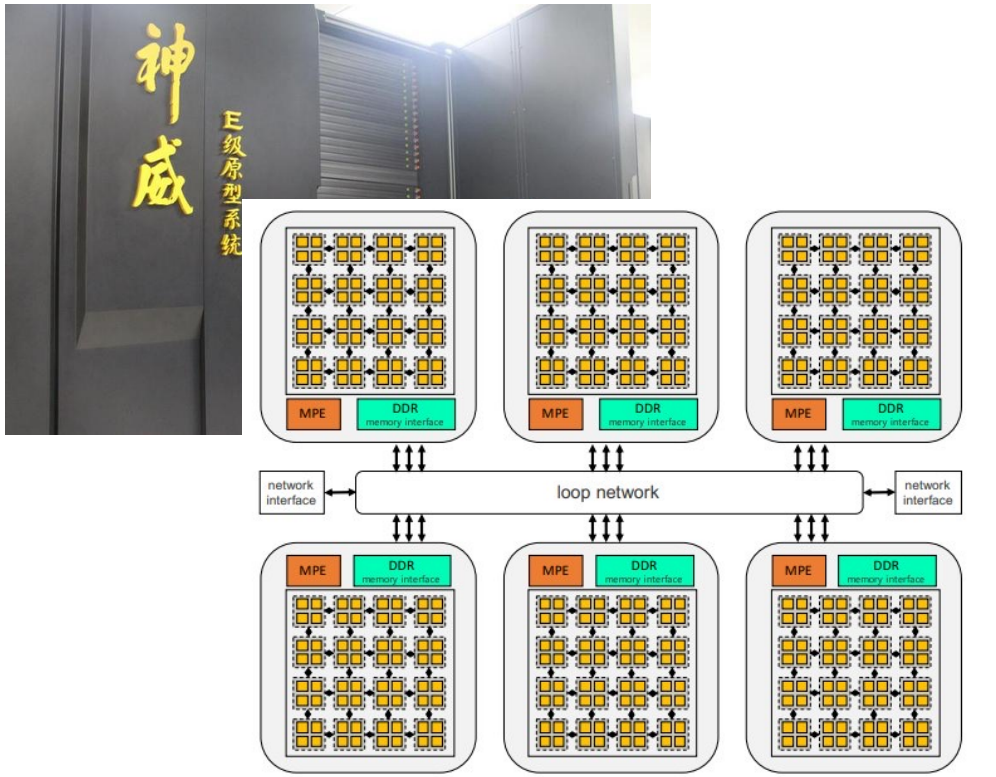
Heterogenous Architecture

- China is exploring new accelerators for Exascale computing.



MT-2000+ accelerator

Tianhe Exascale Prototype, Vendor: NUDT

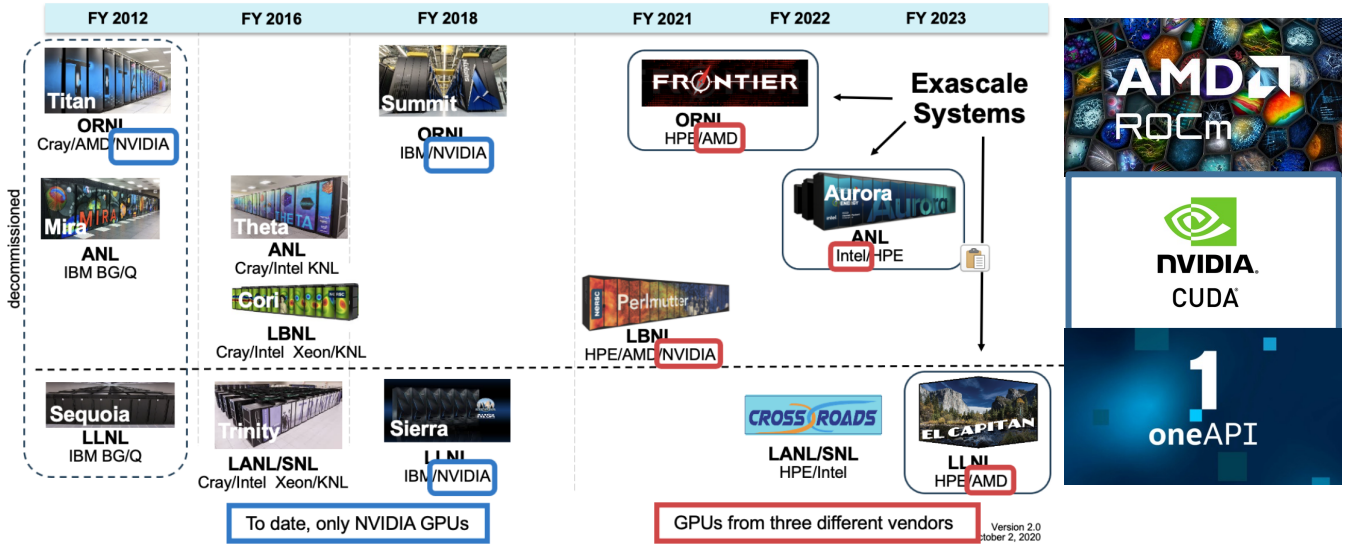


Sunway 26010-Pro compute engine

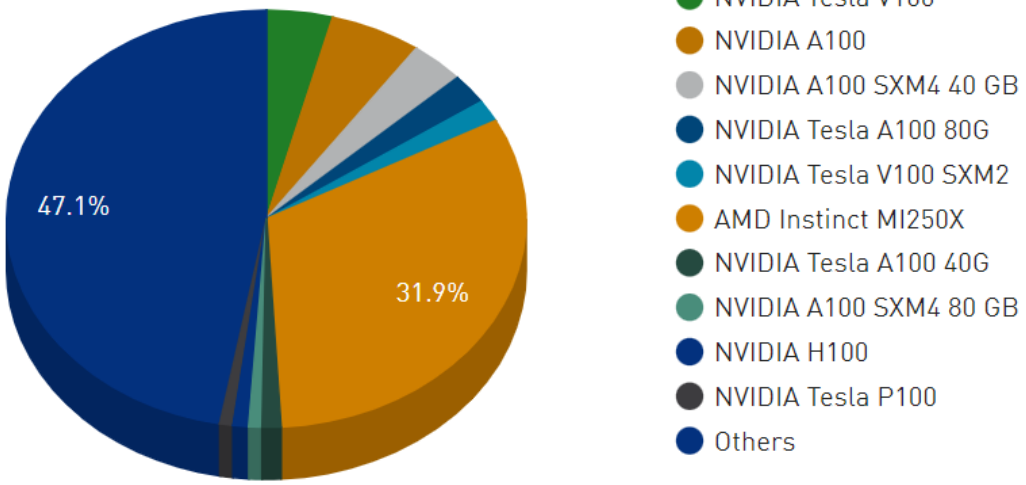
Sunway Exascale Prototype, Vendor: NRCPC

GPU ecosystem is splitting

- Each new generation of machines has been **significantly different** from previous ones
- We **used to have NVIDIA** GPUs. But now different GPU vendors appear in the HPC market
- **AMD** even **start to dominate** in terms of the performance share



GPU types for top HPC systems in US[1]

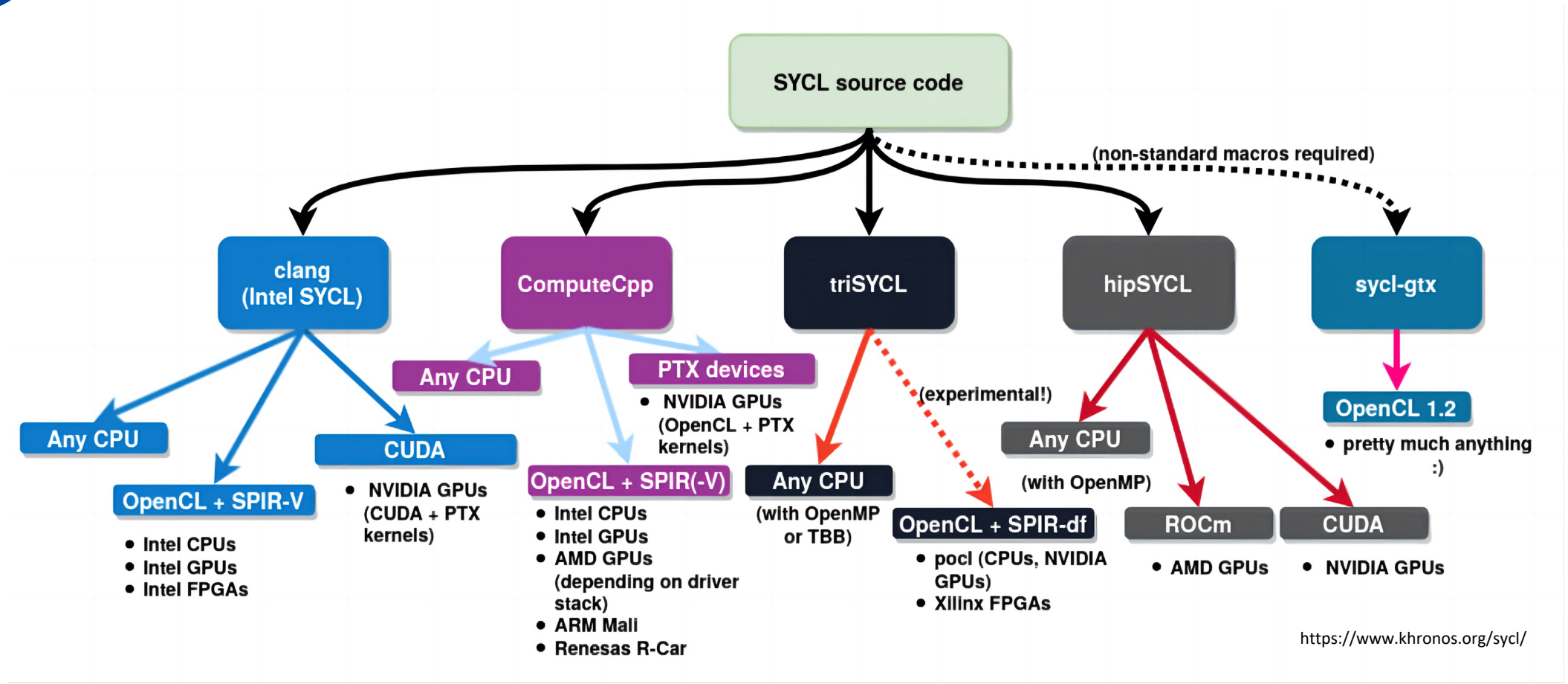


accelerator/co-processor performance share[2]

Portability becomes a critical issue now

1. A. Dubey et al. 2021
2. top500.org

There are indeed solutions to unify the ecosystem

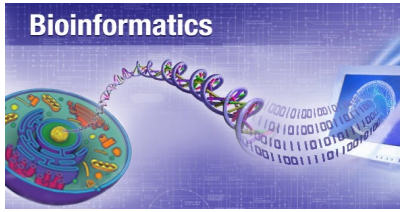


- One can write the code with unified programming model to obtain the application portability
- e.g., Intel OneAPI built upon SYCL
- Market dominators may not want to unify?

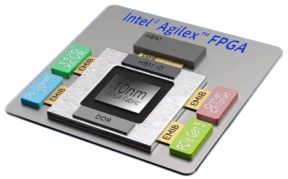
Heterogenous Architecture beyond GPUs

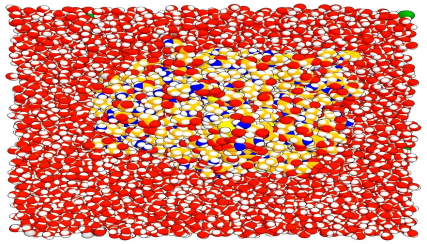
- GPUs are more specialized than CPUs, but still fit many applications. There are multiple levels of accelerator domains.

programmability increases

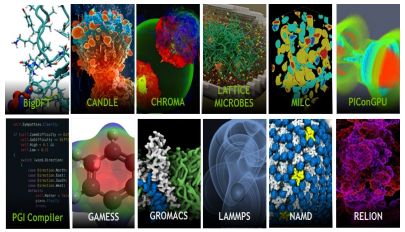


High-Level Synthesis
- Designed for a very specific task (C/C++ RTL)
- Examples: FPGA, CGRA





Application-Specific Architectures
- Co-designed hardware for single application/domain
- Examples: Anton, SIGMA, TPU, ASV

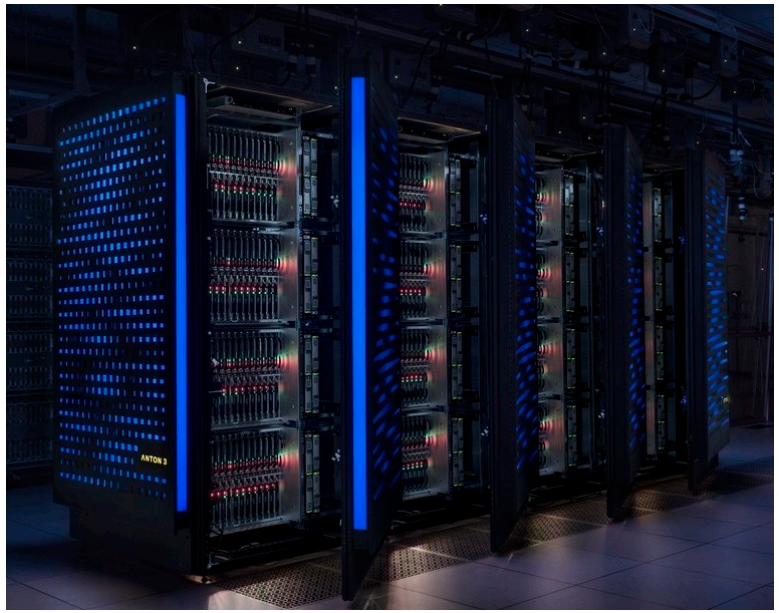


General Purpose
- Suitable for many applications/domains
- Examples: GPGPU, GPDSP

performance increases

Application-Specific Architectures

- Anton is way more effective for MD simulations compared with general HPC



	DHFR	ApoA1	ATPase	STMV	Ribosome	STMV 5×2×1	STMV 5×2×2	HIV-1 Capsid
# atoms	24K	92K	328K	1,067K	2,181K	10,666K	21,333K	72,404K
Anton 3 512-node (μs/day)	—	—	166.9	121.1	93.2	22.4	16.0	1.9
Anton 3 64-node (μs/day)	212.2	152.4	90.6	42.3	25.4	—	—	—
Anton 2 512-node (μs/day)	87.2	63.7	34.5	13.9	5.0	—	—	—
Non-Anton (μs/day)	1.70 (*a*) 1.46 (*b*) 0.63 (*c*)	0.57 (*b*) 1 (*d*)	0.20 (*b*)	0.064 (*b*) 0.055 (*e*)	0.204 (*f*) 0.031 (*b*)	0.12 (*g*)	0.12 (*h*)	0.008 (*i*)

Quadro RTX A6000 GPU

MDGRAPE -4A

V100 GPU

HYDRA cluster

Summit

Titan

Anton3 is ~100x faster than the general HPC (Summit) of the same period.

AI – Another hot topic for Application-Specific Architectures

	Frontier supercomputer (June 2020)	Human brain
Speed	1.102 exaFLOPS	~1 exaFLOPS (estimate)
Power requirements	21 MW	10–20 W
Dimensions	680 m^2 (7,300 sq ft)	1.3–1.4 kg (2.9–3.1 lb)
Cost	$600 million	Not applicable
Cabling	145 km (90 miles)	850,000 km (528,000 miles) of axons and dendrites
Memory	75 TB/s read; 35 TB/s write; 15 billion IOPS flash storage system, along with the 700 PB Orion site-wide Lustre file system	2.5 PB (petabyte)
Storage	58 billion transistors	125 trillion synapses, which can store 4.7 bits of information each

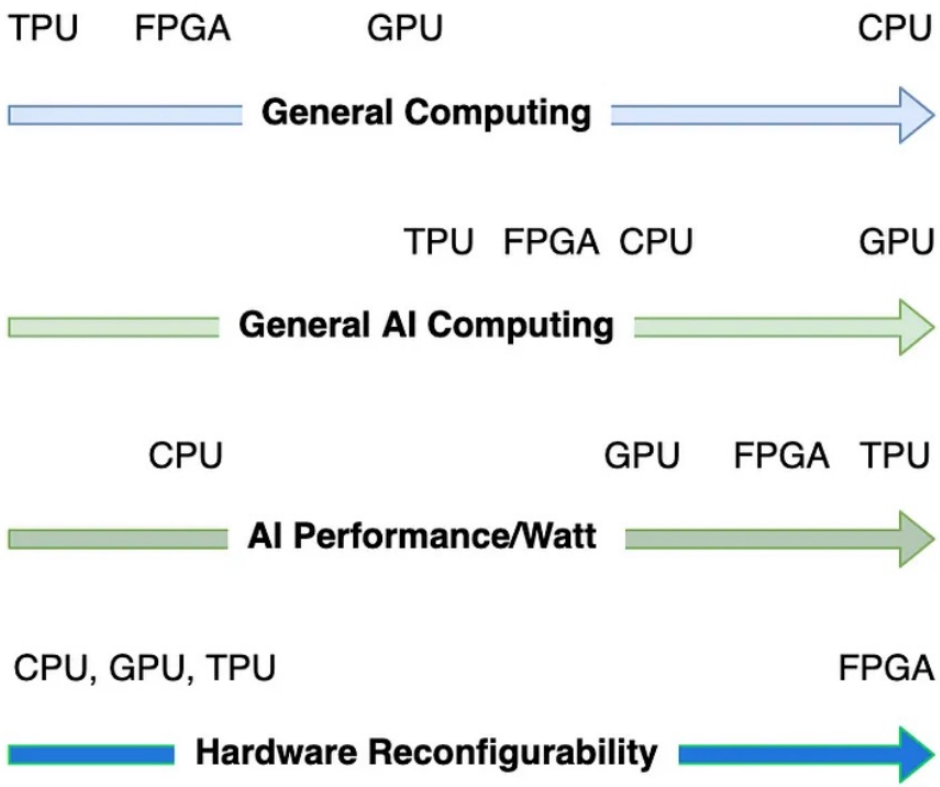
AI is to create intelligent systems that can perform tasks typically requiring **human intelligence**

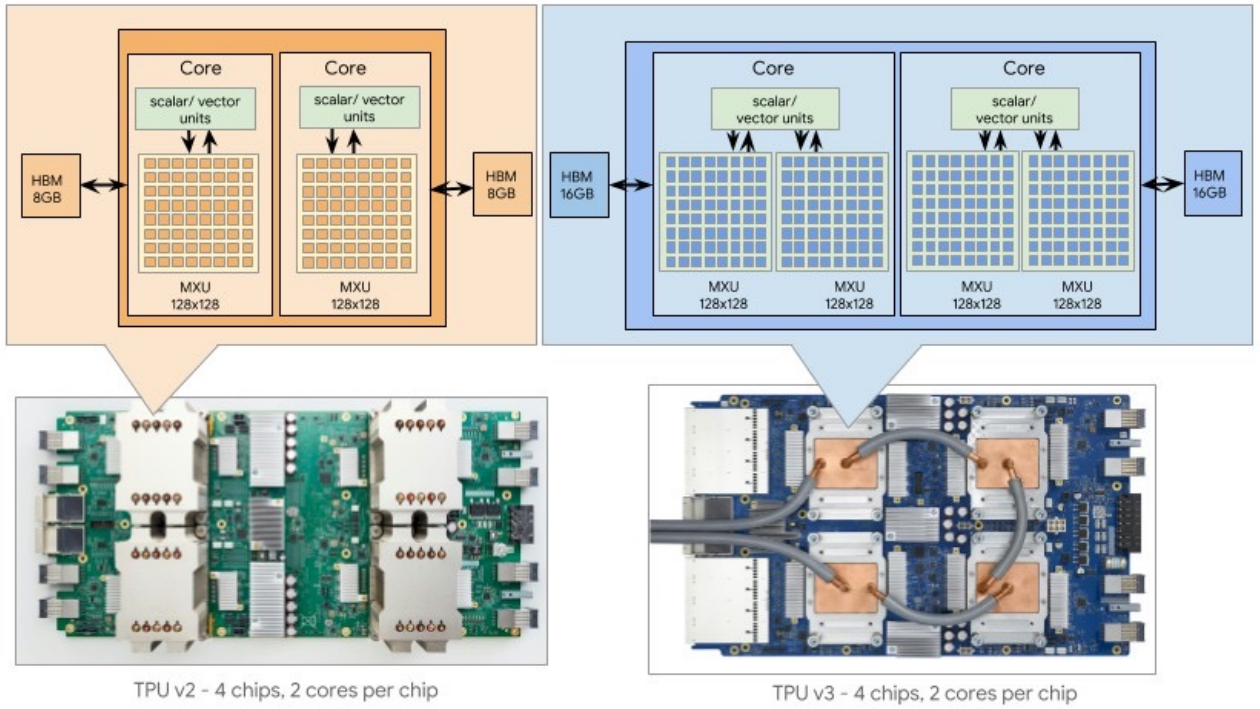
- Human brain has 100 billion **low-precision** neurons
- The power is **only 20 Watt**

Architecture for AI should achieve
high performance + high power efficiency

Smirnova, L et al. (2023). Organoid intelligence (OI): the new frontier in biocomputing and intelligence-in-a-dish. *Frontiers in Science*, 0.

AI – Another hot topic for Application-Specific Architectures





TPU v2 - 4 chips, 2 cores per chip

TPU v3 - 4 chips, 2 cores per chip

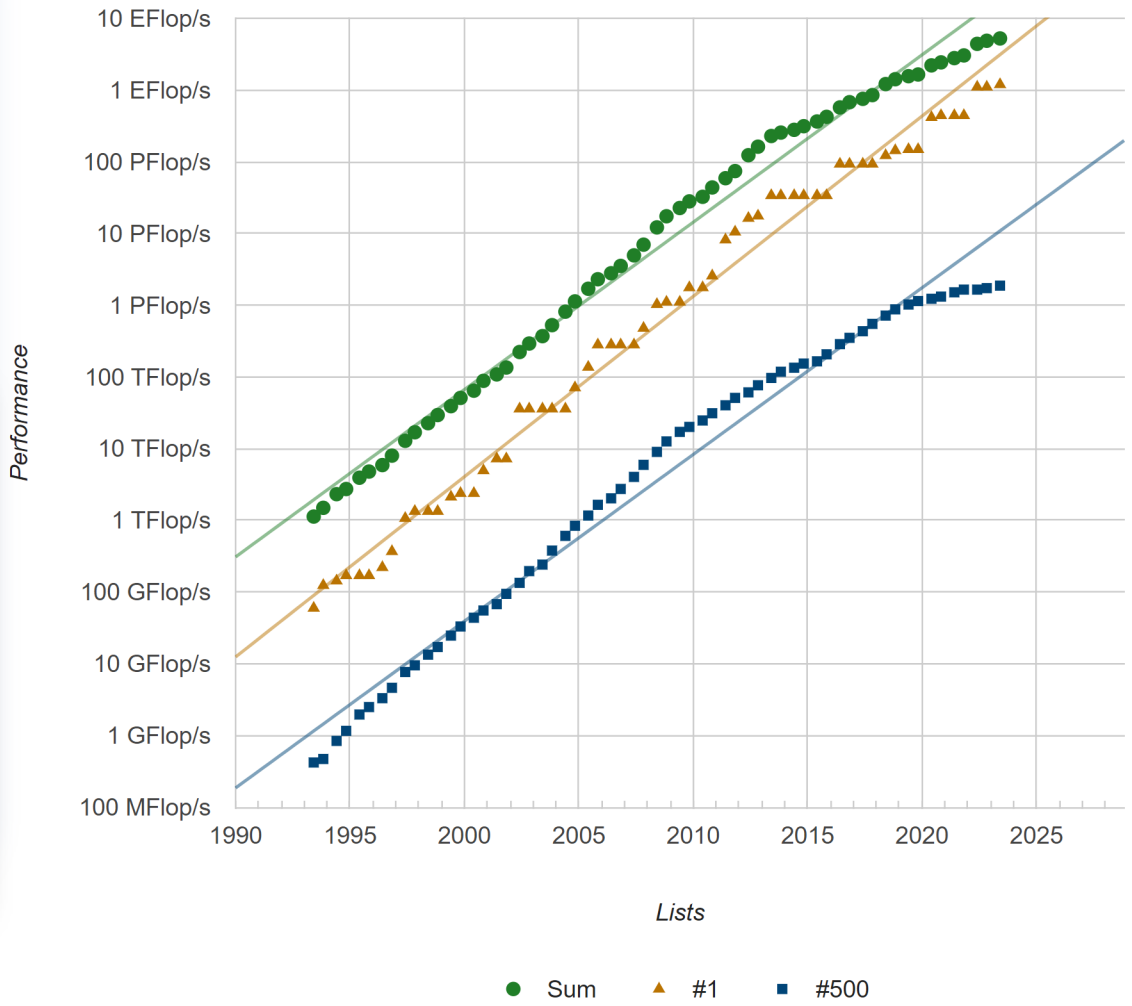
A general GPU provides the programmability that many projects may consider as "fat"

Google TPU, as well as other AI ASICs, sacrifice programmability for efficiency and performance

What for post-exascale computing?

- Performance development of HPC is **slowing down**.
- What engine should we rely on for building the **post-exascale** systems?

Projected performance development[1]



1. top500.org

Possible Directions

Extreme heterogeneity

More accelerators and platforms are required for complex HPC workflows.

New Computing Paradigms

New computing paradigms are promising to achieve higher performance for certain apps.

Disaggregated Architecture

Disaggregated architecture can provide better flexibility and resource utilization

Wafer Scale Integration

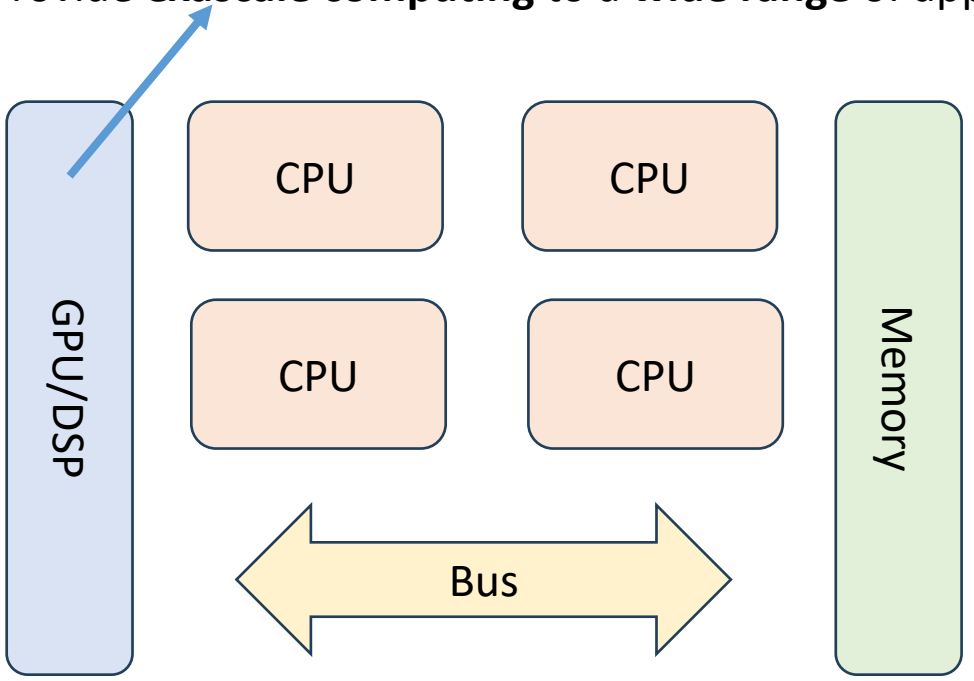
WSI can significantly improve the performance of memory access and communication

Extreme heterogeneity

FACT The spacing of circuits on an integrated circuit is reaching the scale of individual atoms, if we cannot put more transistors on the chip, we must **use the space wisely**.

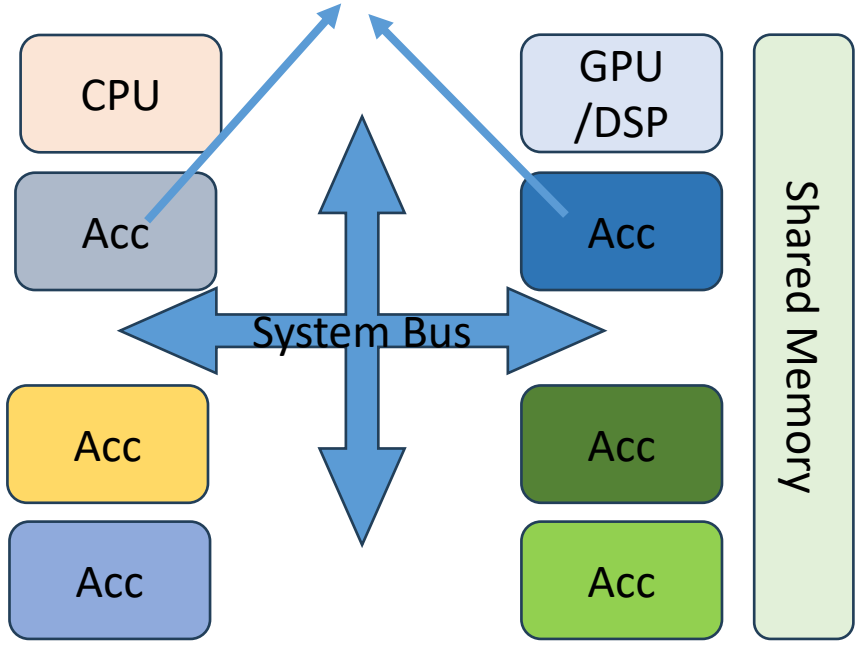
CONSEQUENCE Different accelerators use transistors more efficiently by **specializing the architecture to the target scientific problem**.

Provide **exascale computing** to a **wide range** of applications.



Now

Provide **post-exascale computing** to a **specific** application.



Future (maybe)

Extreme heterogeneity

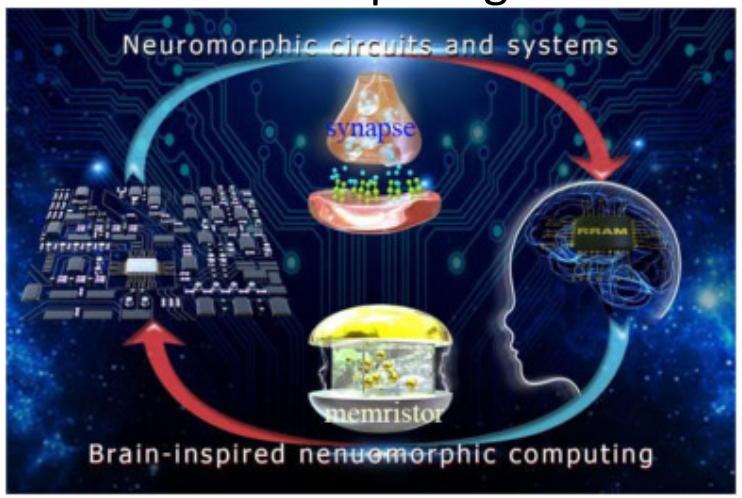
- Instead of putting multiple accelerators into the machine, a recent startup named Tachyum is designing a chip which **unifies CPU, GPU, TPU… INTO A SINGLE CHIP**.
- So far, their design is still under evaluation on FPGA.
- The key **programmability problem** does not change no matter if you put accelerators separately, or as a whole.

- Under EH architecture, it's **extremely unlikely** that code will be **performance portable across different platforms**.
- As architectural diversity grows for EH architectures, and complexity grows, the current approach to **write different code for each accelerators will become infeasible**.

The current programming models and monolithic OS need to evolve to adapt EH architecture
SW/HW codesign is becoming more important than ever

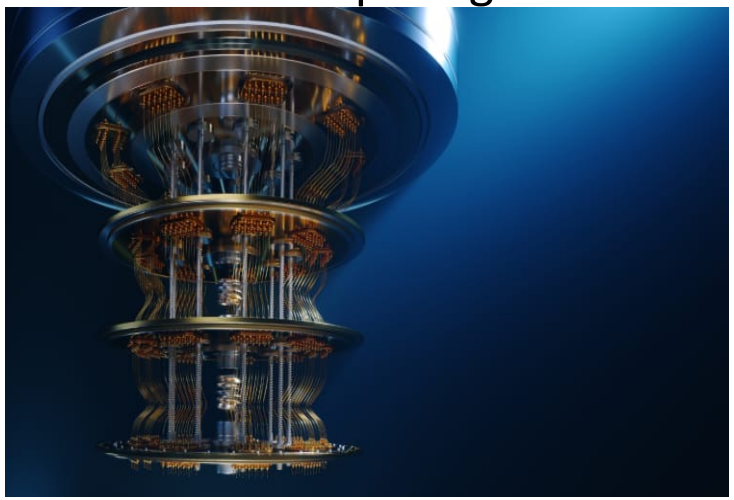
New Computing Paradigms

neuromorphic computing



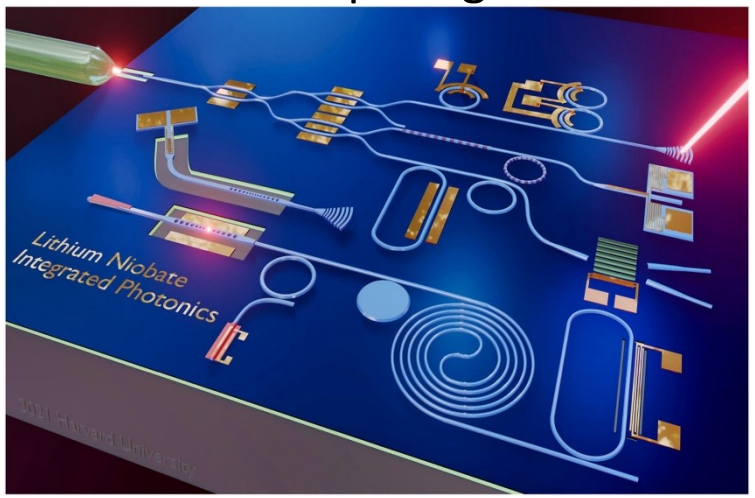
Neuromorphic architectures leverage **massive parallelism, sparse activity**, **and event-driven** computing. Suitable for machine learning, scientific computing as well as **modeling cognitive tasks**.

quantum computing



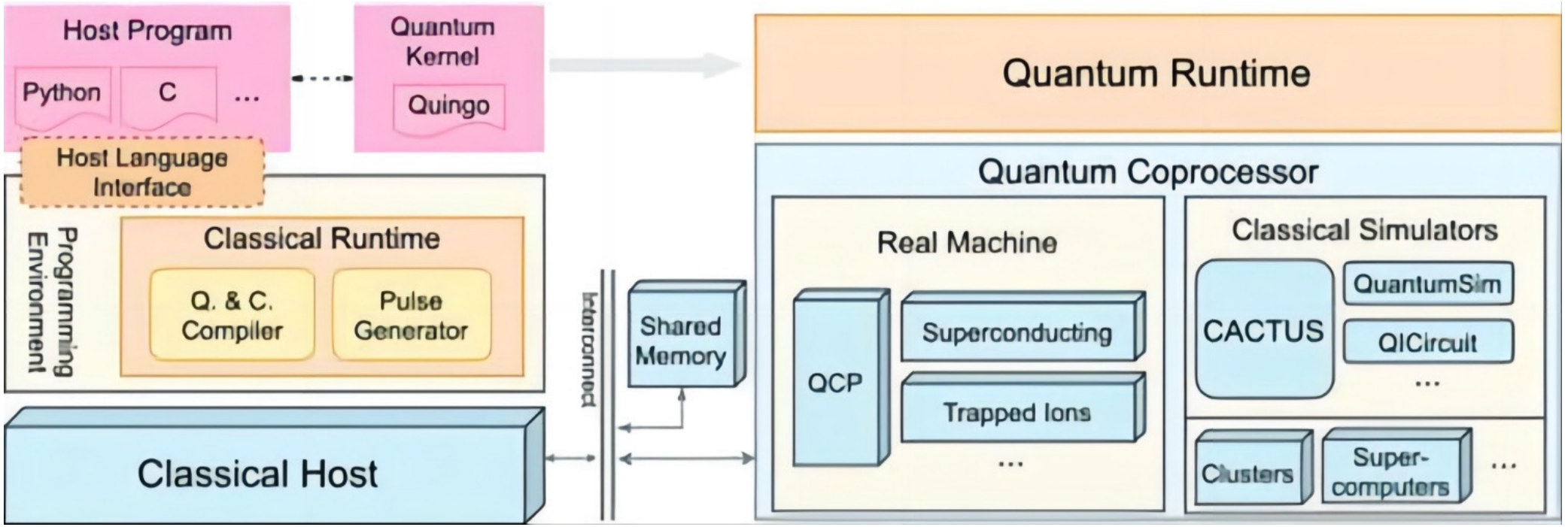
Quantum computing can overtake classical HPC in **certain tasks**. i.e., **quantum advantage**. Therefore, integrating quantum computing into HPC systems attracts much attention.

photonic computing



Photonic computing is carried out via **multi-polarization** channels, leading to an enhancement in **computing density** by several orders compared to that of conventional electronic chips.

New Computing Paradigms

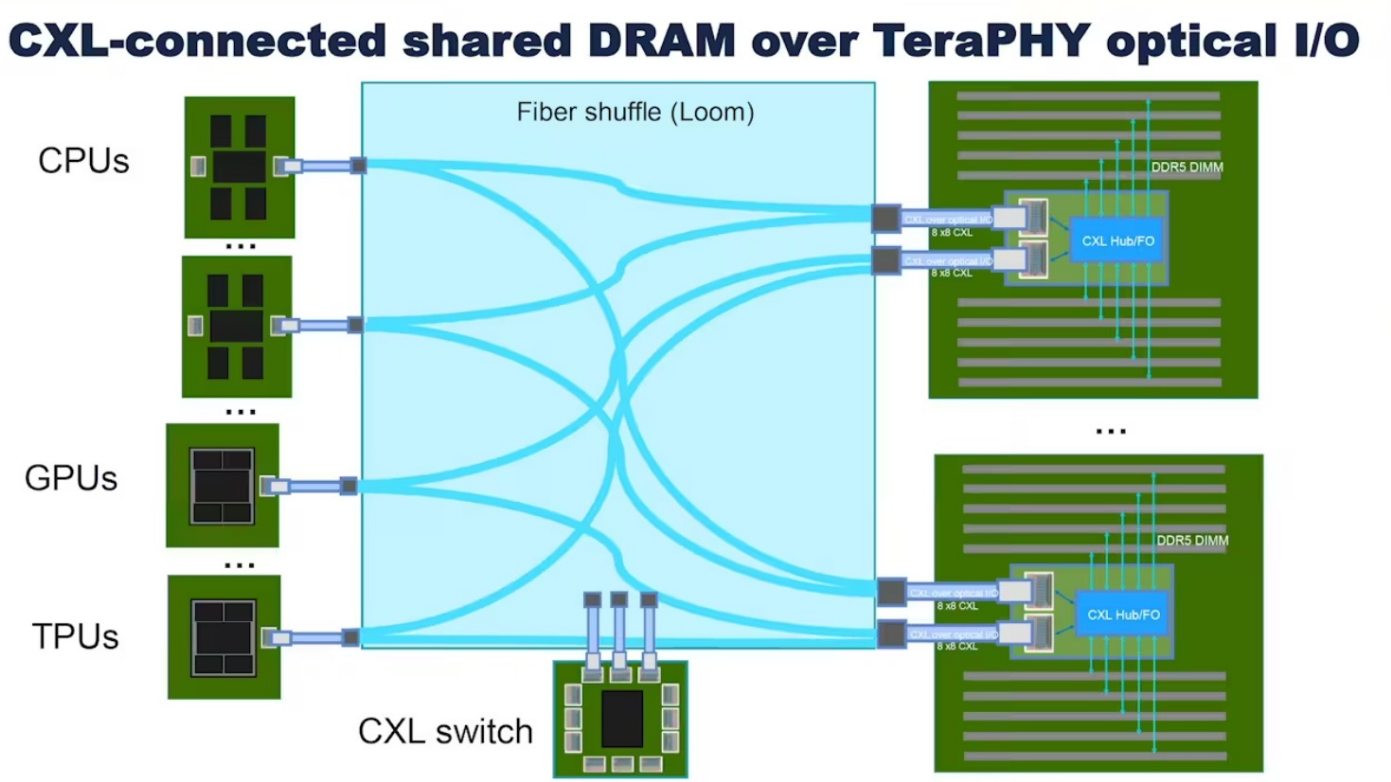


In NUDT, we're working on compilers, programming models to enable easier programming on quantum computers. We proposed Quingo, which is a Programming Framework for Heterogeneous Quantum-Classical Computing with NISQ Features.

Disaggregated Architecture

WHY? Current HPC systems consist of massive compute and memory resource that are tightly coupled in nodes.
BUT more than 90% of jobs utilize less than 15% of the node memory capacity

WHY? In the current HPC architectures, accelerators have isolated memory space.
SO THAT much energy and time are spent for **data movements**.



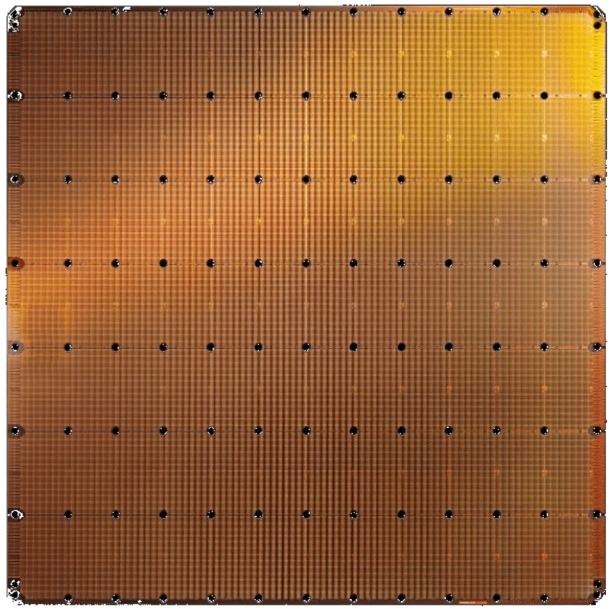
https://ayarlabs.com/webinar-disaggregated-system-architectures-for-next-generation-hpc-and-ai-workloads/

Wafer Scale Integration

WSI itself is not a new concept.

The major issues for application of WSI include generality, technical problems such as testing and yield statistics, and practical problems such as packaging, ruggedness, repairability, and system partitioning.

——Conference Notes of ISSCC 1984



Cerabras WSE2
2.6 Trillion transistors
46225 mm^2 silicon
850,000 cores

Largest GPU
54.2 Billion transistors
826 mm^2 silicon

Cerebras CS-2 system

Chips like these have only recently been truly produced, but with **very high COST** (millions of dollars for one chip).
It remains a question what applications can **benefit commercially** from such chips?

Thanks

Porting may be easy, but performance portability?

- There are projects, such as RAJA, which tries to provide better **performance portability**
- RAJA seeks to make a single-source code performance portable across heterogenous HPC architectures, through **parallelizing loops on different platforms**

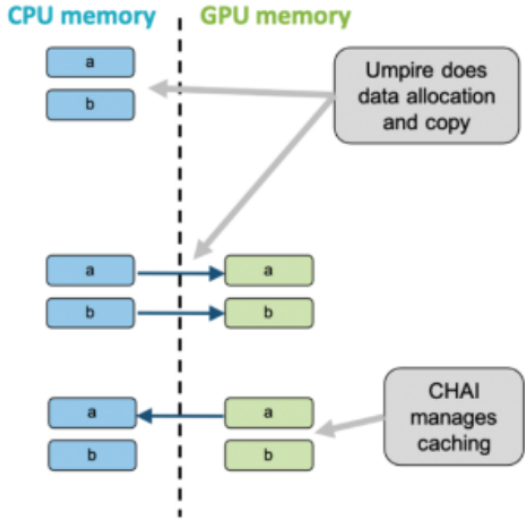
```
/*
  An example loop which adds two vectors, ported to RAJA and parallelized with OpenMP,
shown below (taken from the RAJA examples):
  RAJA::omp_parallel_for_exec - executes the forall loop using the
  #pragma omp parallel for directive
*/
RAJA::forall<RAJA::omp_parallel_for_exec>
  (RAJA::RangeSegment(0, N), [=](RAJA::Index_type i) {
    C[i] = A[i] + B[i];
  });
/*
where RangeSegment(0, N) generates a sequential list of numbers from 0 to N. The same lo
op parallelized and executed on a GPU with CUDA looks similar:
*/
RAJA::forall<RAJA::cuda_exec<CUDA_BLOCK_SIZE>>
  (RAJA::RangeSegment(0, N), [=] __device__(RAJA::Index_type i) {
    C[i] = A[i] + B[i];
  });
checkSolution(C, N);
```

```
chai::ManagedArray<float> a(100);
chai::ManagedArray<const float> b(100);

RAJA::RangeSegment range(0, 100);

// Run GPU kernel
RAJA::forall< RAJA::cuda_exec >(
  range, RAJA_LAMBDA (int i) {
    a[i] += b[i];
} );

// Run CPU kernel
RAJA::forall< RAJA::seq_exec >(
  range, RAJA_LAMBDA (int i) {
    std::cout << "a[i] = " << a[i] << "\n";
} );
```



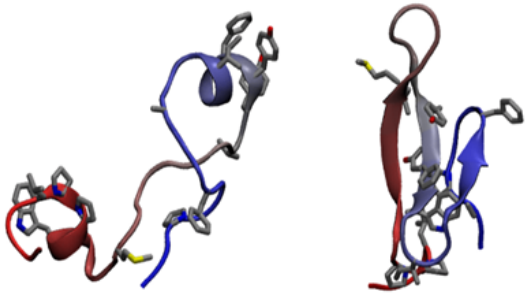
The **CHAI library** implements a **managed array** abstraction to automatically copy data

But essentially, making RAJA widely applicable is not much easier than manual porting

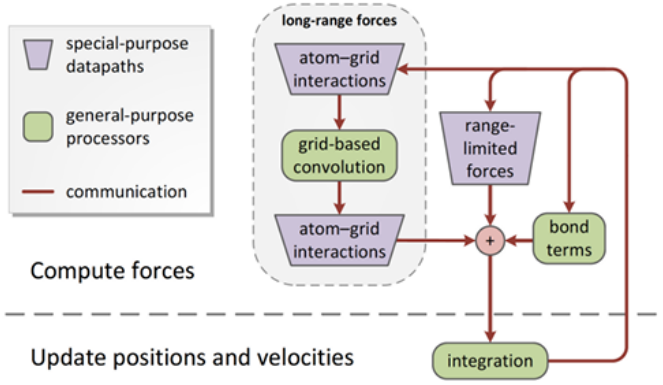
Application-Specific Architectures

- An **molecular dynamics simulation** models the **motion of** a set of **atoms** over a large number of discrete time steps. Why is Anton so fast for MD? Take Anton as an example

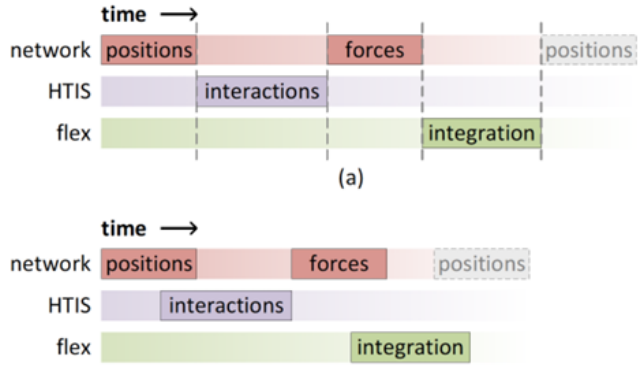
Molecular Dynamics



Special-purpose Hardware



Optimized Algorithms



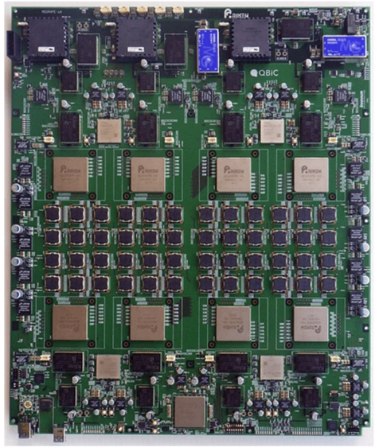
- ✓ Very **low end-to-end internode communication latency** for fine-grained messages
- ✓ **Application-specific compression** reduce the size of messages between nodes
- ✓ A **new** hardware **synchronization primitive** which supports fast fine-grained synchronization for parallel MD application

1. Anton2:Special-purpose computing for molecular dynamics simulations

Application-Specific Architectures

MDGRAPE-4A from RIKEN[1] is also designed for accelerating molecular dynamics

MDGRAPE-4A Board: 8 LSIs



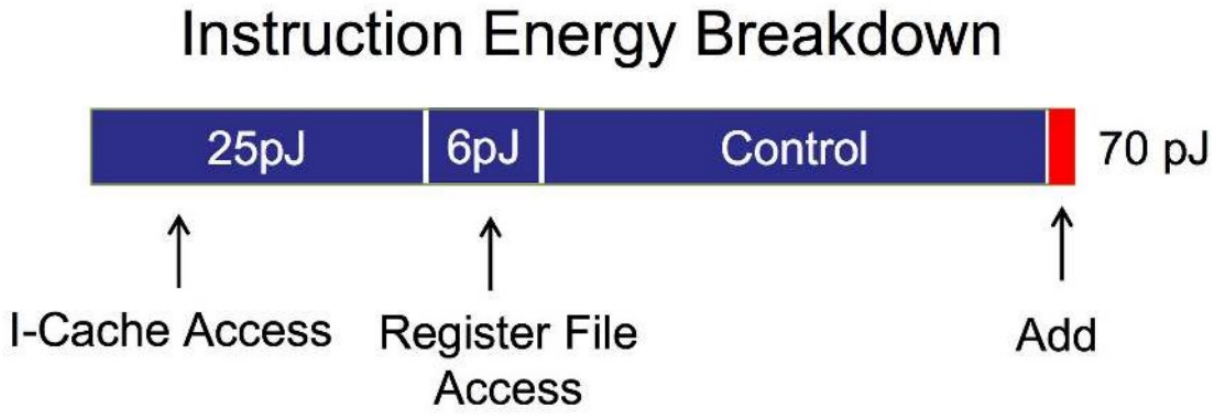
MDGRAPE-4A Full System: 64 Boards

Manufacturer: Tokyo Electron Device,Integran, HOKS, Fujikura
Power: 65kW
Cooling: air-flow
Cost: ~$6,500,000

Computer System	Elapsed time for single step (μs)	Performance (μs/day,dt=2.5fs)
MDGRAPE-4A	200	1
Commodity Cluster	1,000	0.2
GPU	2,000	0.1
Laptop	100,000	0.002

1. https://www.r-ccs.riken.jp/exhibit_contents/SC20/mdgrape-4a.html

How does TPU achieve its goal?

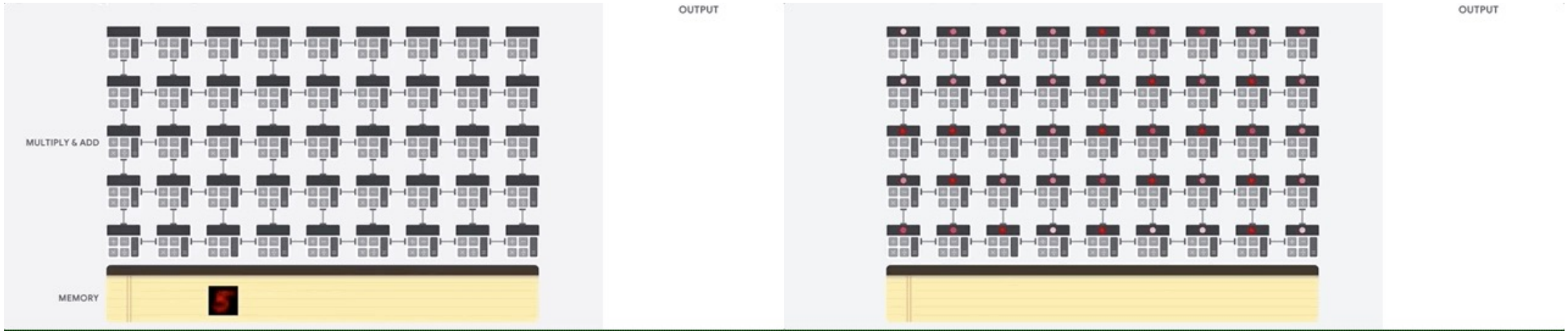


Operation	Energy (nJ)
ADD	0.64
L1 → REG	1.11
L2 → REG	2.21
L3 → REG	9.80
MEM → REG	63.64
Prefetch	65.08

[Kestor, 2014]

- **Data movement and memory become the major performance and energy bottleneck**
- **Pushes programming models to more localized data movement**

How does TPU achieve its goal?



Figures from the introduction of Cloud TPU by Google

GPU: Although **added many parallel compute units**, each still needs to **frequently access memory**

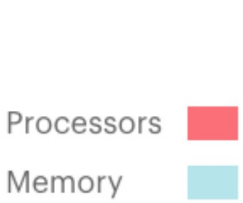
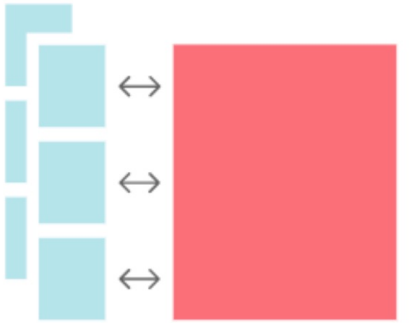
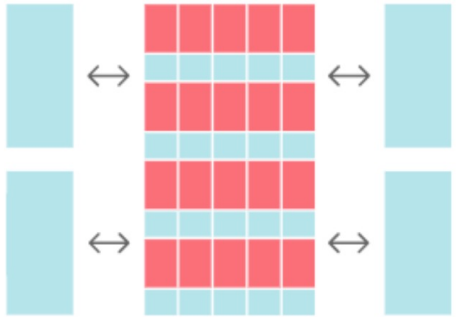
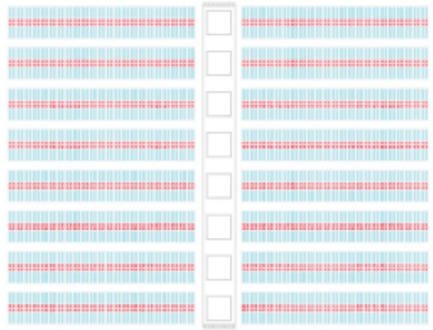
TPU: Rather than providing general programmability, focuses on the large-scale **addition and multiplication**

better fits neural networks

Each time multiplication is performed, the **result is passed to the next multiplier**.

no need to access memory

Graphcore (IPU) uses model parallelism

chip	CPU	GPU/TPU	IPU
parallelism	weak parallelism	data parallelism	model parallelism
				
memory access	Memory and chips are separated.	Model and data are separated	Model and data are tightly coupled

New Computing Paradigms

Although promising, it may still take a long way before we see real integration of these paradigms in HPC.

"Our quantum computing research really **focuses on quantum practicality and scalability**. We're trying to bring quantum out of the physics lab and into a commercial reality. "

Anne Matsuura, Intel

"I'm curious what it can bring, but I **don't see any benefits in the near-time**. Although I'm a skeptic, I believe there are certain applications where quantum computing will be probably helpful in the future. "

Natalia Vassilieva
Cerebras Systems director of product

"The unfortunate piece about this is that there is **only a handful of algorithms that provide speed-up (over classical)** and every other algorithm is basically composed out of those.. "

Torsten Hoefler
Professor， ETH Zürich