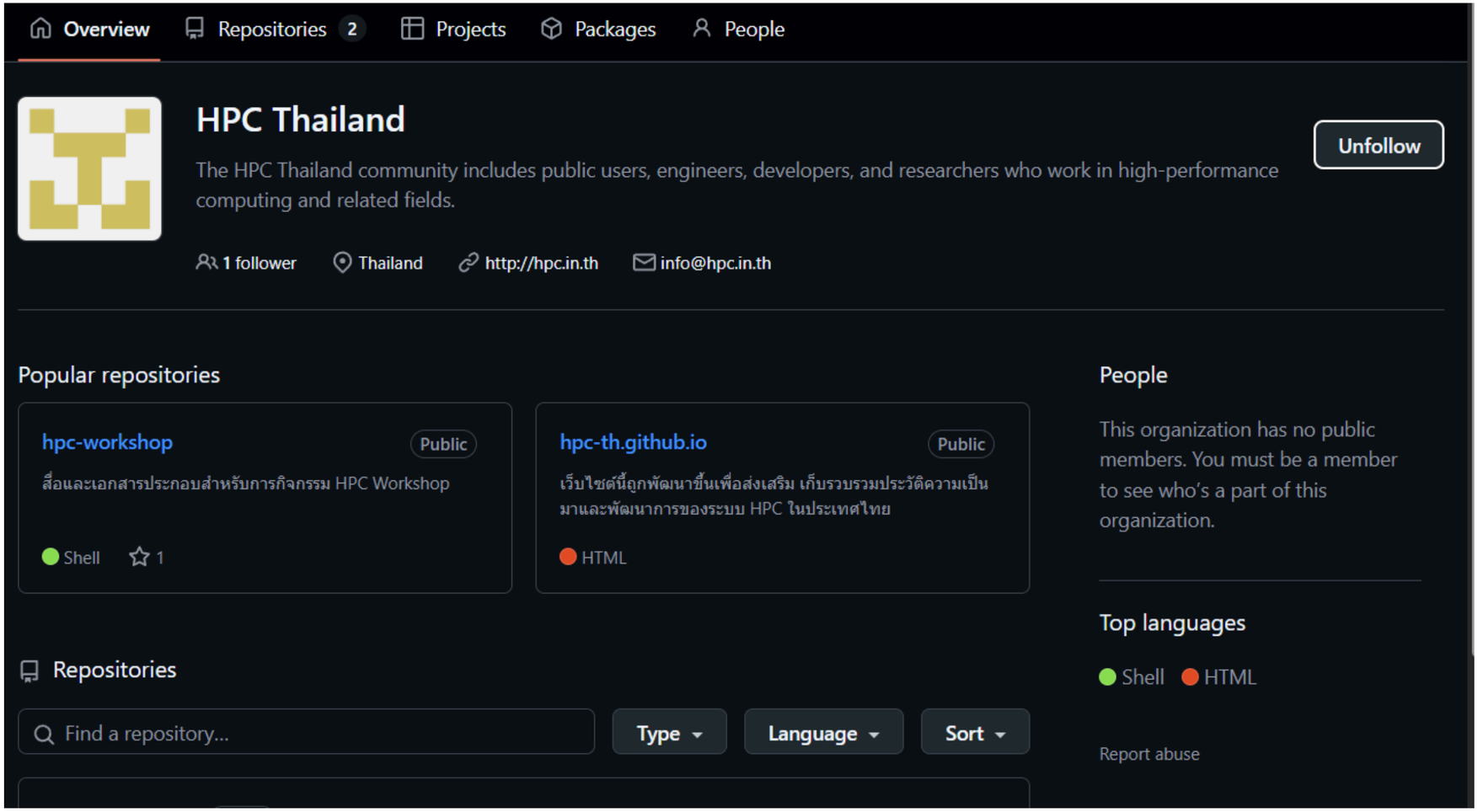
INTRODUCTION TO HIGH PERFORMANCE COMPUTING

Permpoom Boonyarit
System Engineer
AI Engineering Institute
CMKL University
permpoon.b@cmkl.ac.th



https://github.com/HPC-Thailand

Outline

Introduction

- History
- Parallel Computing
- Supercomputer
- Scale of supercomputer
- TOP500

HPC Architecture

- Case Study: Frontier

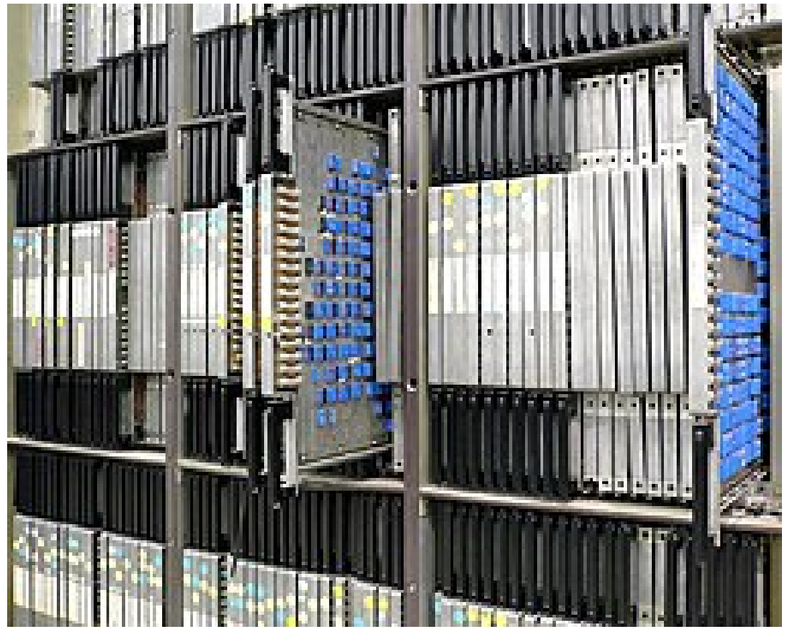
INTRODUCTION

History

Early HPC Systems 1970s

CDC 6600 10 MHz (100 NS Clock Cycle)

- Main memory 982 KB (60-bit words)
- Seymour Cray design
- Peak 3 MFLOPS
- **first successful Supercomputer**



ILLIAC IV 25 MHz (125 NS Clock Cycle)

- Main memory 1MB
- By design 1GFlop/s but due to budget constraints
- 50 MFLOPS

https://en.wikipedia.org/wiki/CDC_6600
https://en.wikipedia.org/wiki/ILLIAC_IV

https://www-users.york.ac.uk/~mijp1/teaching/4th_year_HPC/lecture_notes/History_of_Supercomputers.pdf

History

Early 1970s HPC Systems

CDC 7600 36.4 MHz (27.5 ns clock cycle)

- Primary memory 65 Kwords (60-bit words)
- Seymour Cray design
- Peak 36 Mflop/s
- Broke down at least once/day (often four or five times)

IBM 370/195 18.5 MHz (54 ns clock cycle)

- High degree of parallelism
- Up to 7 operations at a time
- Up to 4 MB of memory
- **Peak 55 Mflop/s**

Both systems had a high degree of instruction-level pipelining and parallelism.

https://www.hpcwire.com/2022/11/16/jack-dongarra-a-not-so-simple-matter-of-software/

History

What is FLOPS?

FLOPS - Floating point operation per second

FLOPS is a measure of a computer's performance based on the number of floating-point arithmetic calculations that the processor can perform within a second. Floating-point arithmetic is a term used in computing to describe a type of calculation carried out on floating-point representations of real numbers.

$$FLOPS = nodes \times \frac{cores}{nodes} \times \frac{cycles}{second} \times \frac{FLOPs}{cycle}$$

3 MegaFLOPS → 3 MFLOPS → 3 x 10^6 Floating point operation per second
1 GigaFLOPS → 1 GFLOPS → 1 x 10^9 Floating point operation per second

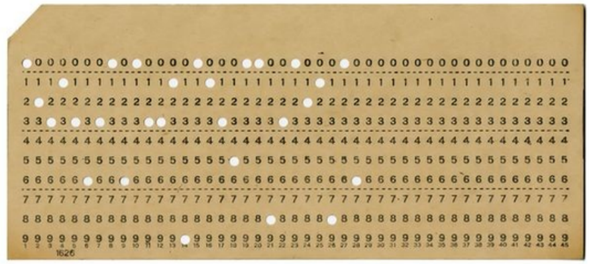
https://www.techtarget.com/whatis/definition/FLOPS-floating-point-operations-per-second

https://www.bu.edu/tech/files/2017/09/Intro_to_HPC.pdf

History

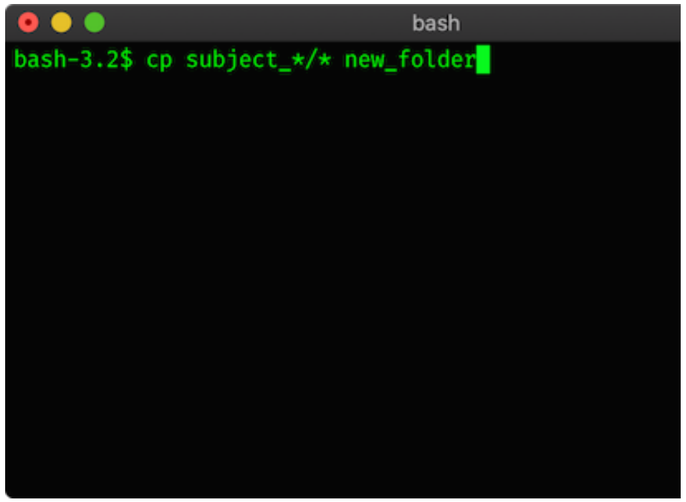
How do we use computer back then?

- Punched card
- Terminal



Terminal & Time Sharing Computer

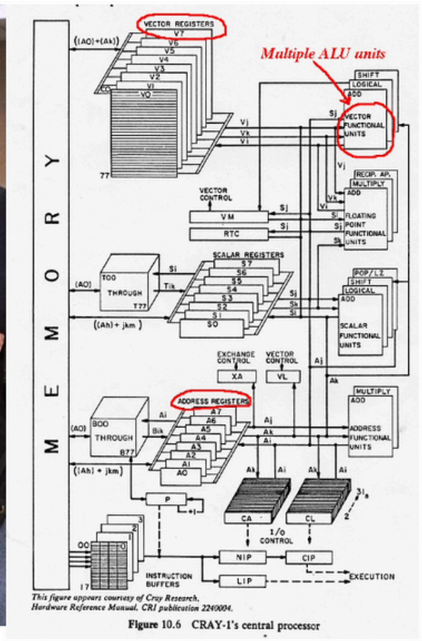
- Computer back then were expensive, not ideal for everyone to have computer at home
- people connected to a **central computer** using **terminals**—simple devices with a keyboard and screen, no computing power
- Since the central computer could only handle one task at a time, it would **"time-slice"** between users, giving the illusion that everyone had their own machine!
- Nowadays we have **terminal (emulator)** on our machine



https://www.cs.cornell.edu/wya/AcademicComputing/text/earlytimesharing.html
https://www.computerhistory.org/revolution/punched-cards/2

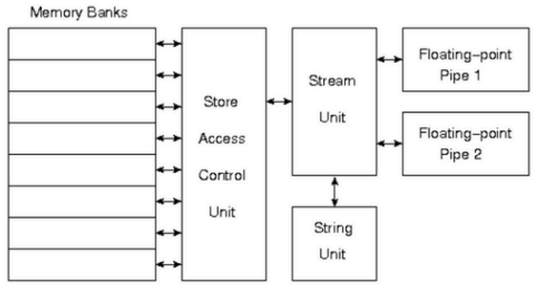
History

Vector Computers 1970s – 1980s



This figure appears courtesy of Cray Research, Hardware Reference Manual, CRI publication 2240004.

Figure 10.6 CRAY-1's central processor



Cray 1

- 80 MHz (12.5 ns), 1 MWords of memory
- 120 Mflop/s, with vector registers
- Over 100 Cray-1s were sold

CDC Star 100

- 25 MHz (40 ns), 1 MWords memory
- 100 MFLOPS, vector operations
- memory-to-memory operations

https://www.hpcwire.com/2022/11/16/jack-dongarra-a-not-so-simple-matter-of-software/

History

History (1970-1990)

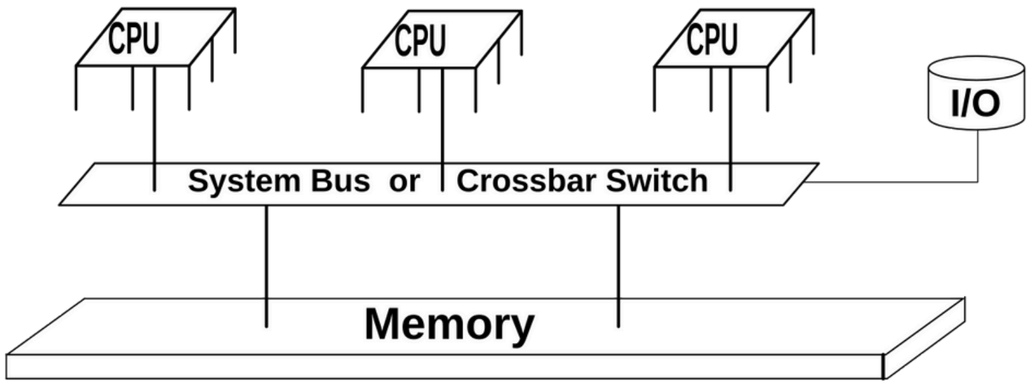
Year	Name	Peak speed	Location
1974	CDC Star-100	100 MFLOPS (vector) ~2 MFLOPS (scalar)	Lawrence Livermore Lab. USA
1975	Cray-1	80 MFLOPS (vector) 72 MFLOPS (scalar)	Los Alamos Lab. USA
1981	CDC Cyber-205	400 MFLOPS (vector) peak, avg much lower	
1983	Cray X-MP	500 MFLOPS (4 CPUs)	Los Alamos Lab. USA
1985	Cray-2	1.95 GFLOPS (4 CPUs) 3.9 GFLOPS (8 CPUs)	Lawrence Livermore Lab. USA
1989	ETA-10G	10.3 GFLOPS (vector) peak, avg much lower (8 CPUs)	
1990	Fujitsu Numerical Wind Tunnel	236 GFLOPS	National Aerospace Lab, Japan

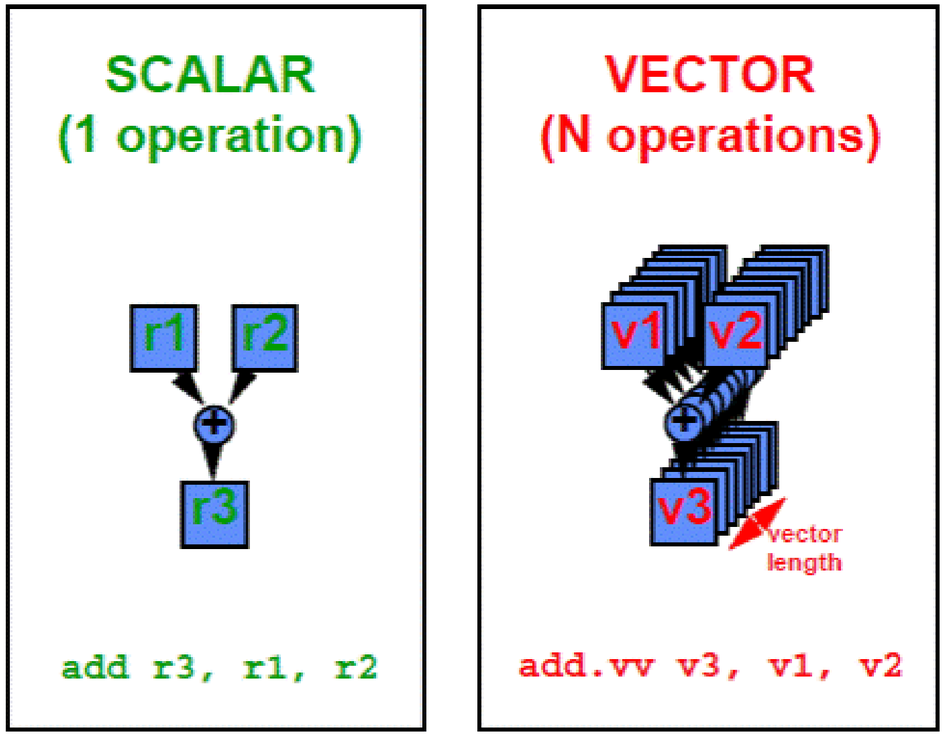
https://www.math-cs.gordon.edu/courses/cps343/presentations/History_and_Overview_of_HPC.pdf

History

Vector Processor and Shared Memory

- During 1980s, there are two strategy to achieve high performance for mainframe/supercomputer
- **Vector Processors**
  - pipeline architecture to rapidly perform a single floating point operation on a large amount of data
- **Shared Memory Multiprocessing**
  - a small number (up to 8) processors with access to the same memory space. Interprocess communication took place via the shared memory.



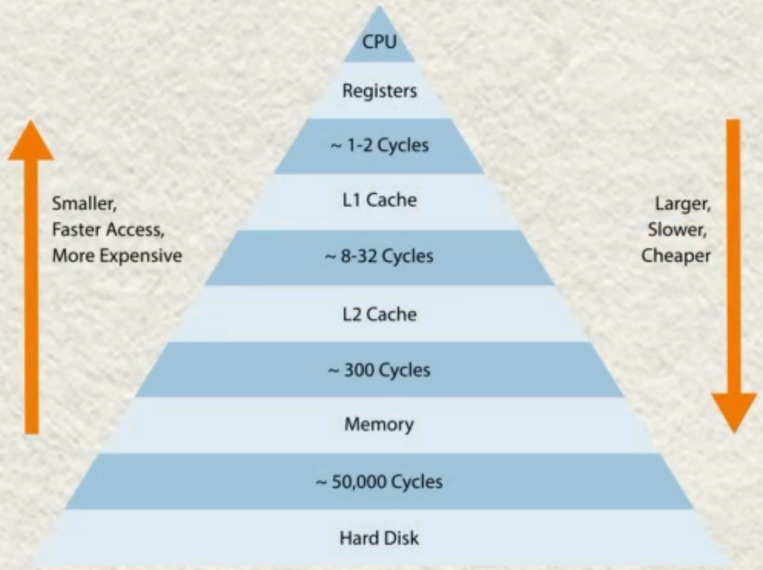


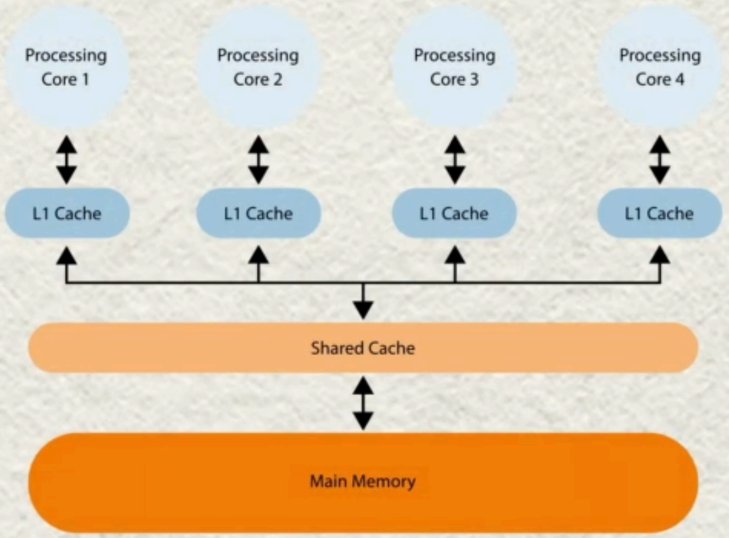
https://www.math-cs.gordon.edu/courses/cps343/presentations/History_and_Overview_of_HPC.pdf

History

1990s Cache & Shared Memory Parallel Systems

Several commercial companies sprang up to implement this architecture:





Multiflow
Supertex
Alliant
Parstec
Convex

Ardent
Kendell Square
Encore
Sequent
...

2021 AM Turing Award Recipient Jack Dongarra Turing Lecture: "A Not So Simple Matter of Software"

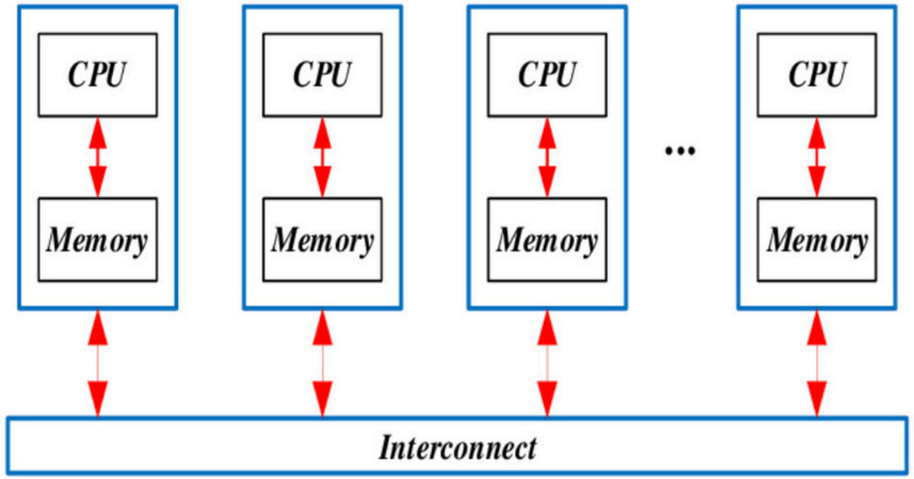
History

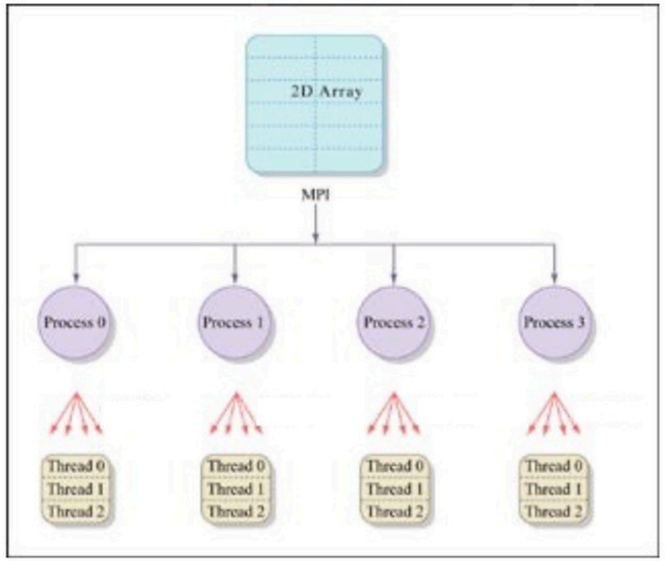
Event	Latency	Scaled
1 CPU cycle	0.3 ns	1 s
Level 1 cache access	0.9 ns	3 s
Level 2 cache access	3 ns	10 s
Level 3 cache access	10 ns	33 s
Main memory access (DRAM, from CPU)	100 ns	6 min
Solid-state disk I/O (flash memory)	10–100 µs	9–90 hours
Rotational disk I/O	1–10 ms	1–12 months
Internet: San Francisco to New York	40 ms	4 years
Internet: San Francisco to United Kingdom	81 ms	8 years
Lightweight hardware virtualization boot	100 ms	11 years
Internet: San Francisco to Australia	183 ms	19 years
OS virtualization system boot	< 1 s	105 years
TCP timer-based retransmit	1–3 s	105–317 years
SCSI command time-out	30 s	3 millennia
Hardware (HW) virtualization system boot	40 s	4 millennia
Physical system reboot	5 m	32 millennia

Systems Performance, 2nd Edition

History

Parallel Computing 2000s
Shared & Distributed Memory Systems





Again, a number of commercial companies created to develop systems:

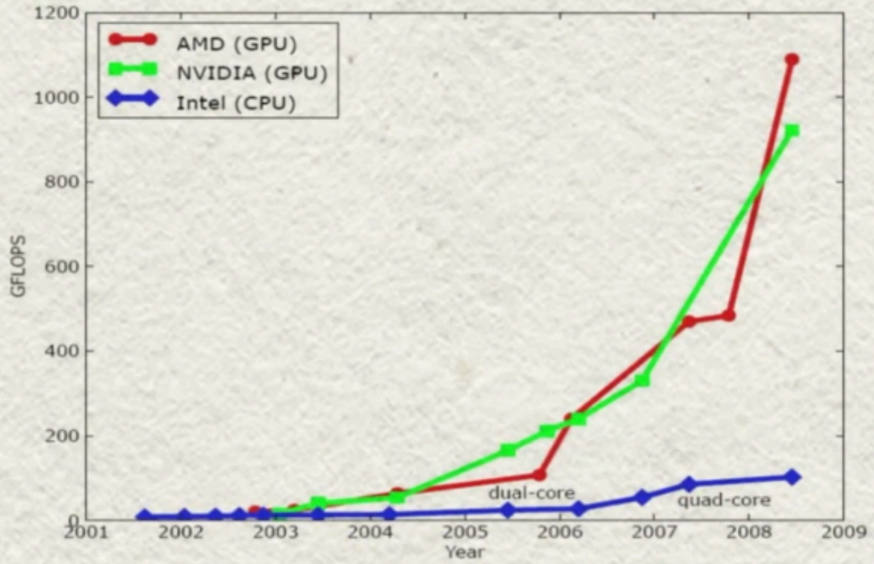
BBN, Elxsi, Myrias, Tera, Thinking Machines, Intel Sci Computer, Meiko, Maspar, nCube, AMT/DAP ...

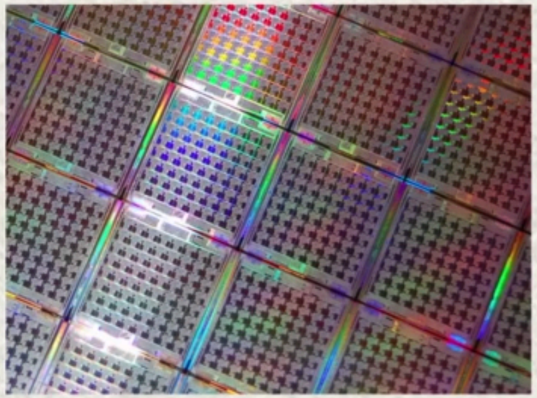
https://www.hpcwire.com/2022/11/16/jack-dongarra-a-not-so-simple-matter-of-software/

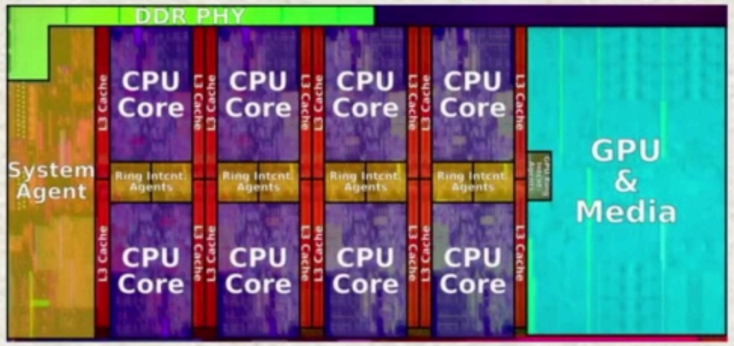
Message Passing

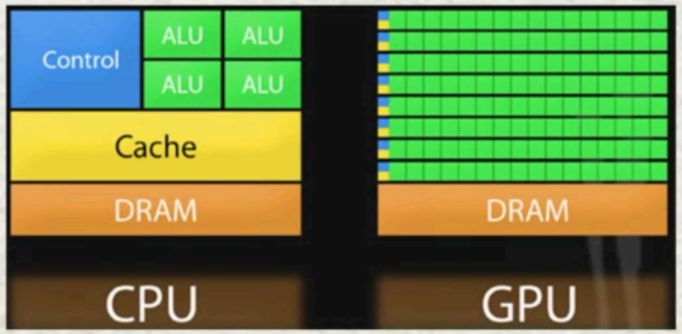
History

Multicore/Hybrid Architectures 2010s









2021 AM Turing Award Recipient Jack Dongarra Turing Lecture: "A Not So Simple Matter of Software"

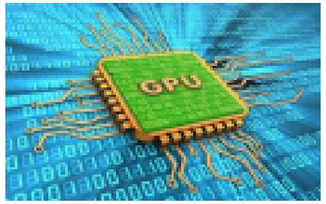
History

Today's HPC Environment for Scientific Computing

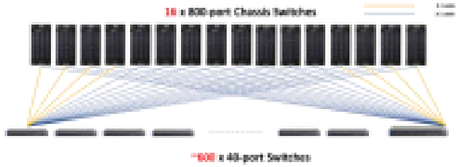
- Highly parallel
  - Distributed memory
  - MPI + Open-MP programming model

ORNL Frontier, 2 Eflop/s,
8.8 x 10^6 Cores, 9408 nodes, 30 MW
(node = 1-AMD CPU + 4-AMD GPUs)
> 98% of performance from GPUs

- Heterogeneous
  - Commodity processors + GPU accelerators



- Communication between parts very expensive compared to floating point ops



- Floating point hardware at 64, 32, 16, & 8 bit levels

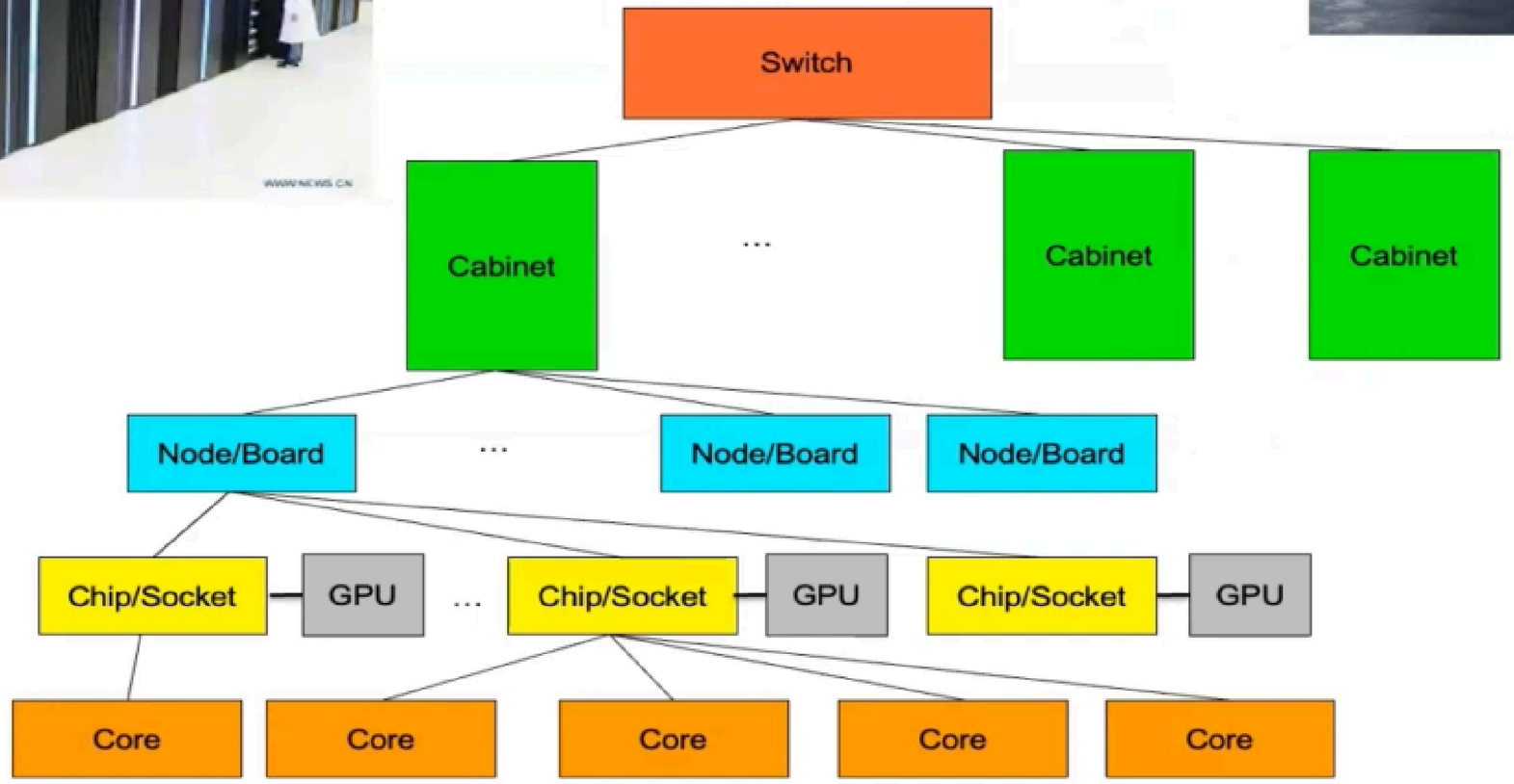
Type	Size	Range	$u = 2^{-t}$
half	16 bits	$10^{\pm 5}$	$2^{-11} \approx 4.9 \times 10^{-4}$
single	32 bits	$10^{\pm 38}$	$2^{-24} \approx 6.0 \times 10^{-8}$
double	64 bits	$10^{\pm 308}$	$2^{-53} \approx 1.1 \times 10^{-16}$
quadruple	128 bits	$10^{\pm 4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

https://www.fujitsu.com/us/imagesgig5/9_evening%20keynote%20Prof.%20Dongarra%20Final.pdf

History

Example of a Typical Supercomputer

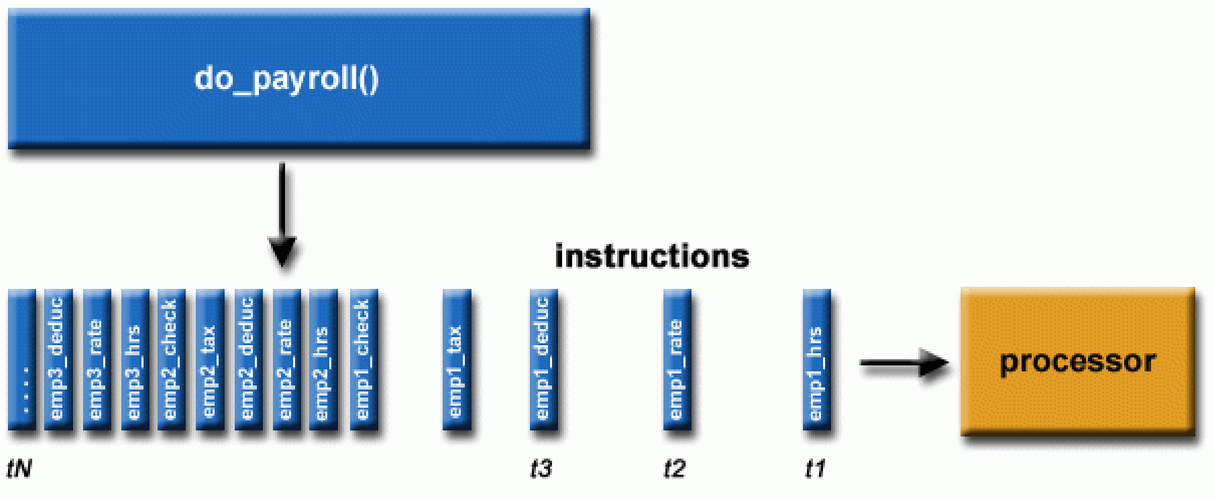
Combination of shared memory and distributed memory programming



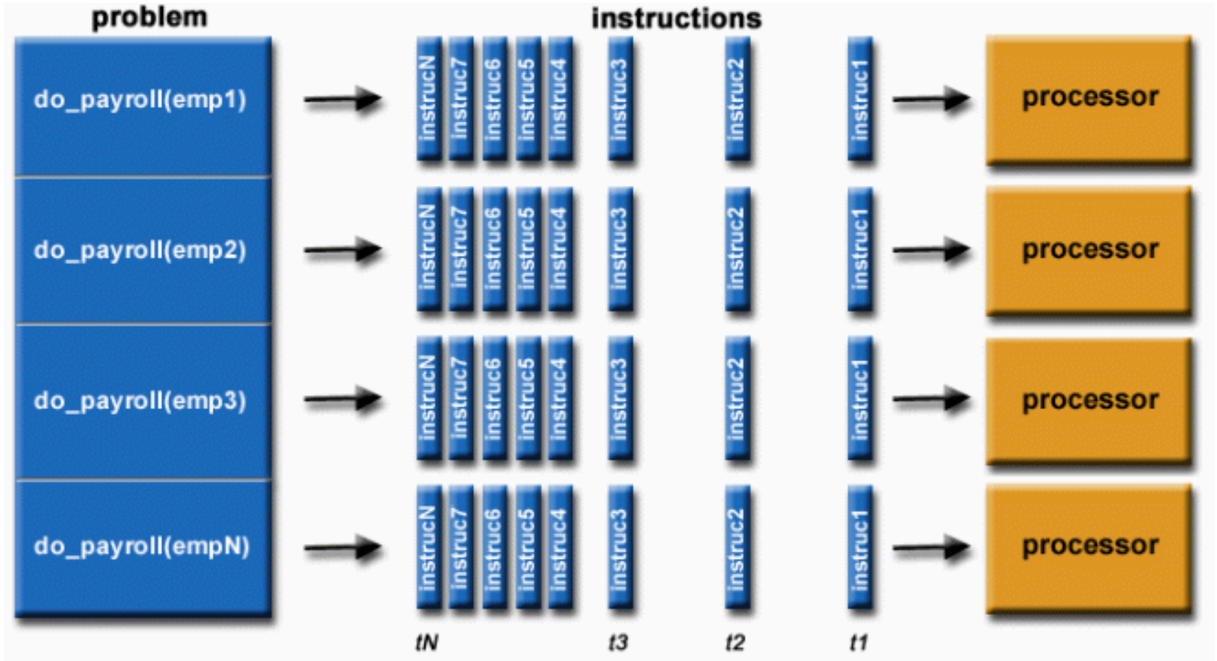
Jack Dongarra at GYSS 2023 – An Overview of High-Performance Computing

Parallel Computing

Serial Computing



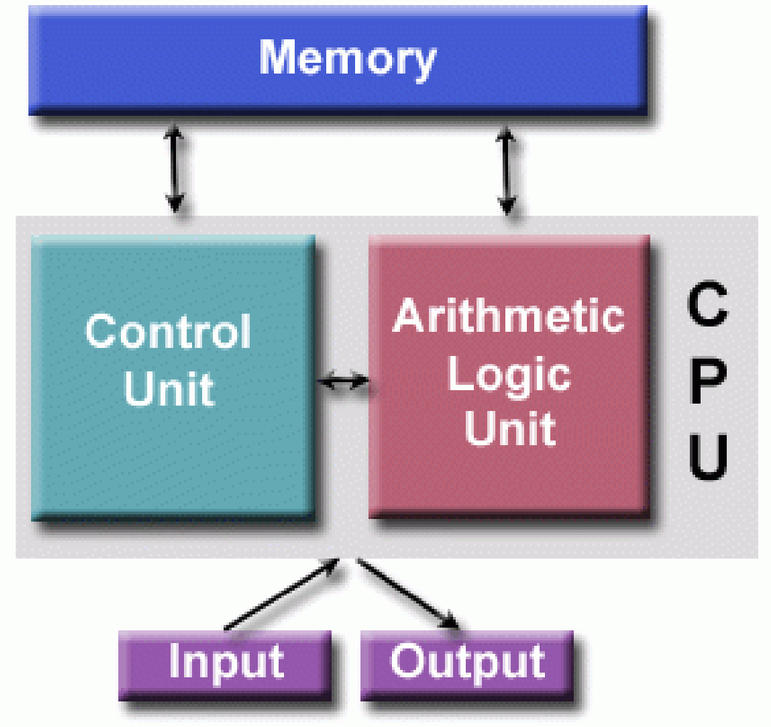
Parallel Computing



https://hpc.llnl.gov/documentation/tutorials/introduction-parallel-computing-tutorial

Parallel Computing

Von Neumann Computer Architecture



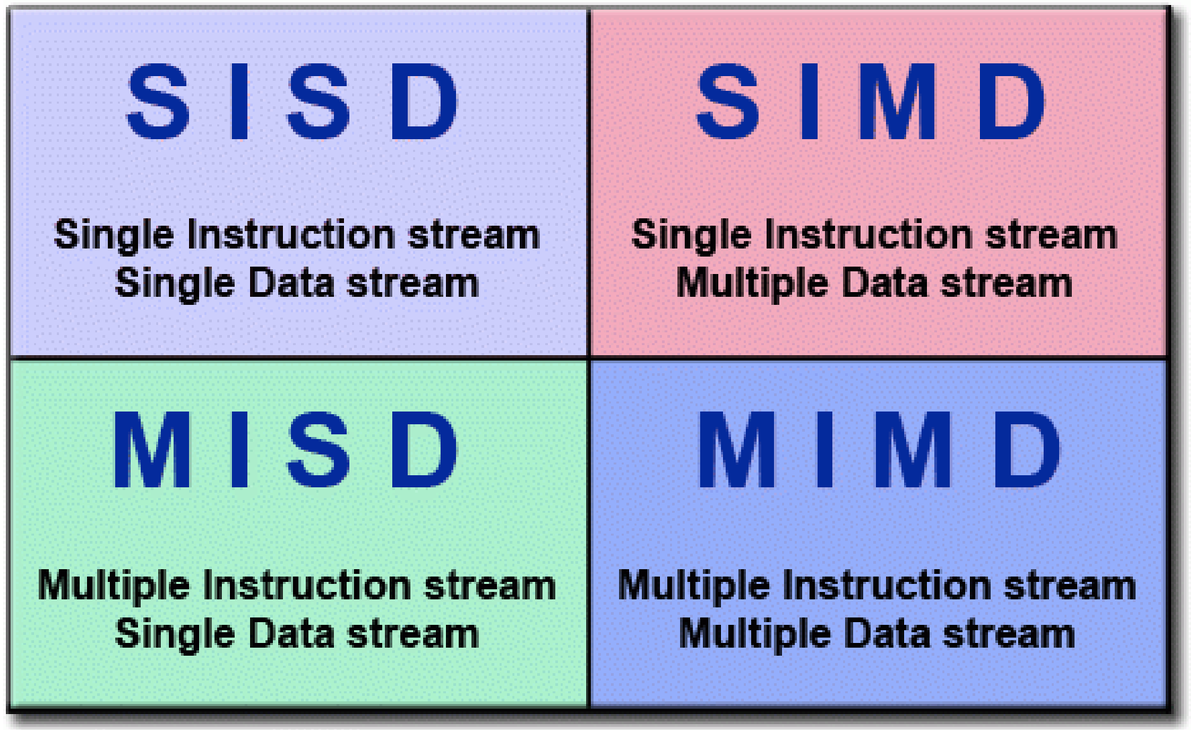
- Also known as "stored-program computer"
- Since then, virtually all computers have followed this basic design

- Read/write, random access memory is used to store both program instructions and data
- Program instructions are coded data which tell the computer to do something
- Data is simply information to be used by the program
- Control unit fetches instructions/data from memory, decodes the instructions and then **sequentially** coordinates operations to accomplish the programmed task.
- Arithmetic Unit performs basic arithmetic operations
- Input/Output is the interface to the human operator

https://hpc.llnl.gov/documentation/tutorials/introduction-parallel-computing-tutorial

Parallel Computing

Flynn's Classical Taxonomy



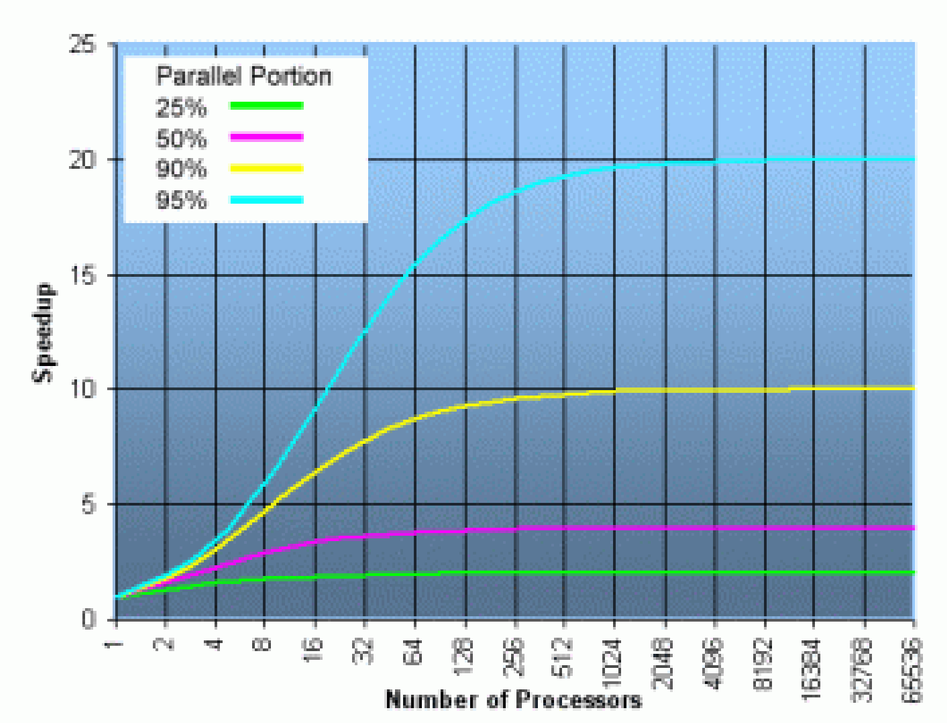
- Flynn's taxonomy distinguishes multi-processor computer architectures according to how they can be classified along the two independent dimensions of **Instruction Stream** and **Data Stream**. Each of these dimensions can have only one of two possible states: Single or Multiple.

SISD: older generation mainframes
SIMD: Most modern computers, particularly those with graphics processor units (GPUs)
MISD: multiple cryptography algorithms attempting to crack a single coded message.
MIMD: most current supercomputers

https://hpc.llnl.gov/documentation/tutorials/introduction-parallel-computing-tutorial

Parallel Computing

Amdahl's Law



- Amdahl's Law states that potential program speedup is defined by the fraction of code (P) that can be parallelized

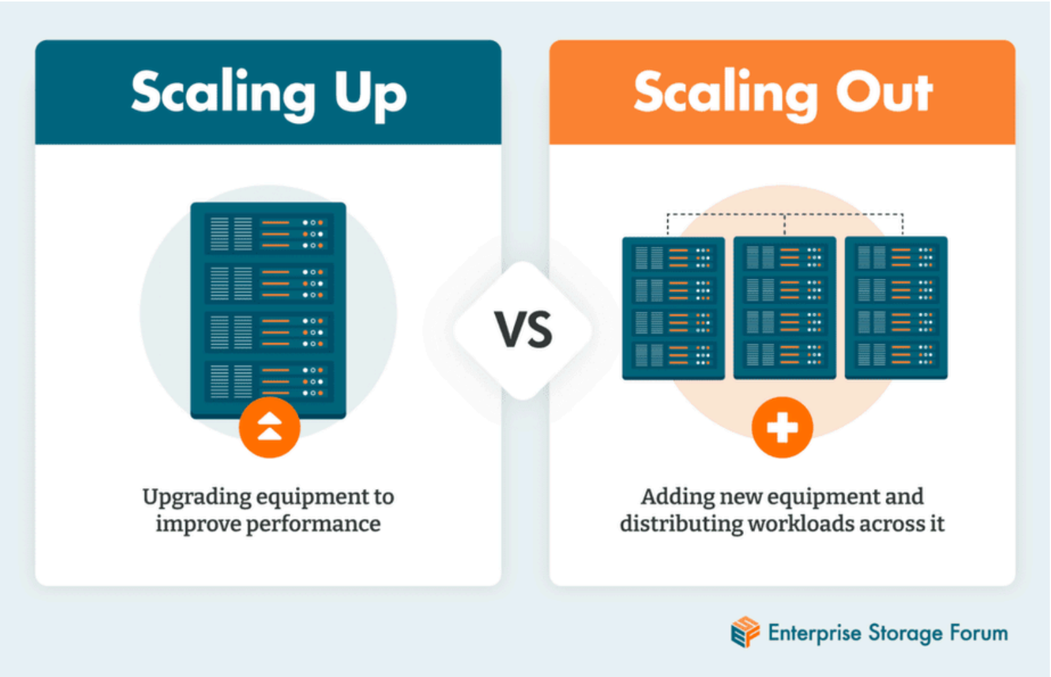
$$\text{speedup} = \frac{1}{\frac{P}{N} + S}$$

https://hpc.llnl.gov/documentation/tutorials/introduction-parallel-computing-tutorial

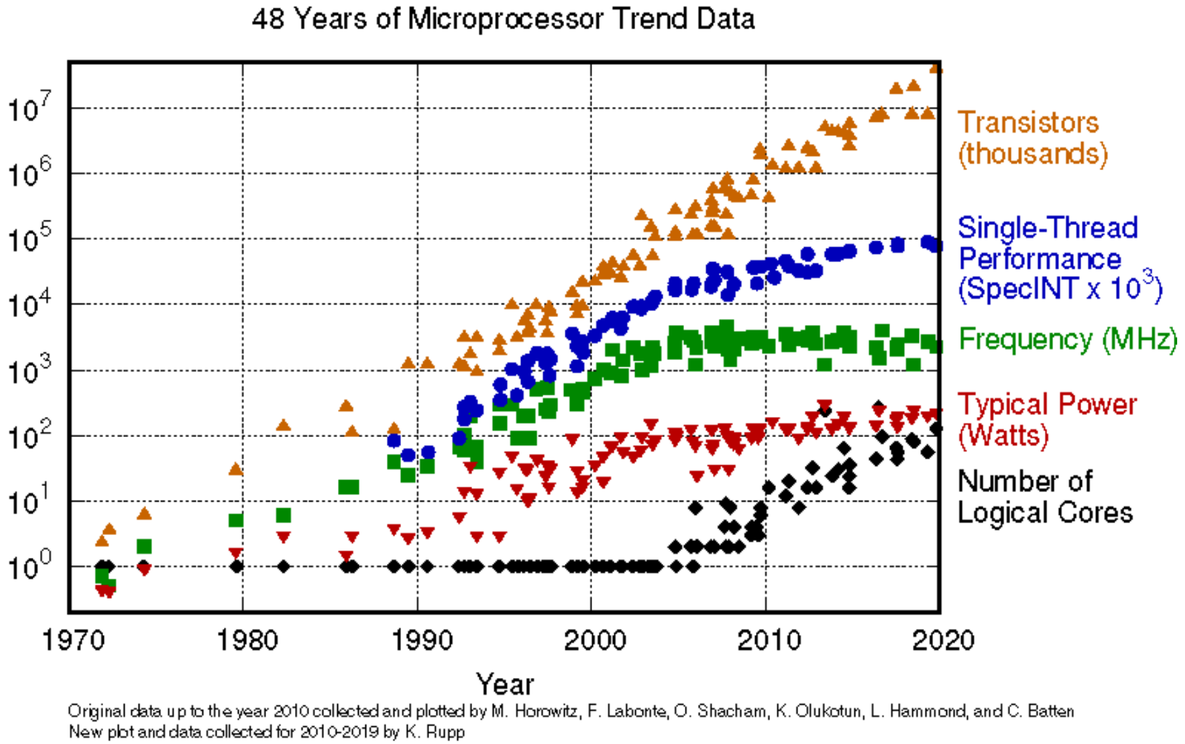
Parallel Computing

The death of Moore's Law?

- Moore's Law states that the number of transistors on a microchip **doubles about every two years** with a minimal cost increase.
- While Moore's law is still delivering exponential improvements, the results are being **achieved at a slower pace**.
- However, the pace of technology **innovation** is NOT slowing down.
  - hyperconnectivity, big data, and artificial intelligence
- As the ability to scale a single chip slows, the industry is finding other methods of innovation to maintain exponential growth



https://www.enterprisestorageforum.com/hardware/scale-up-vs-scale-out/



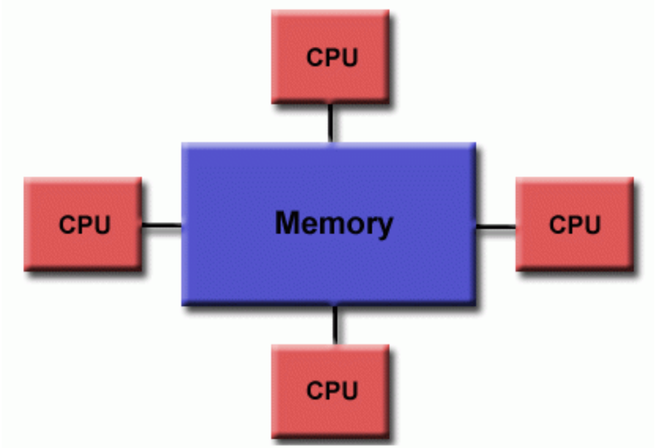
https://semianalysis.com/2023/02/04/a-century-of-moores-law/

https://www.synopsys.com/glossary/what-is-moores-law.html

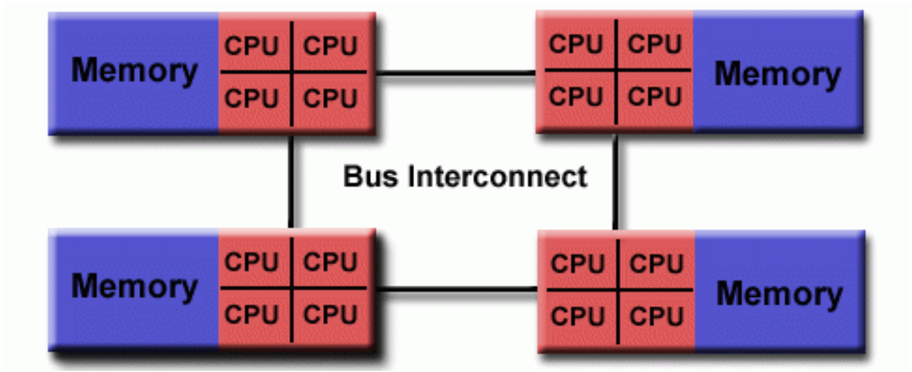
Parallel Computing

Parallel Computer Memory Architectures

Shared Memory

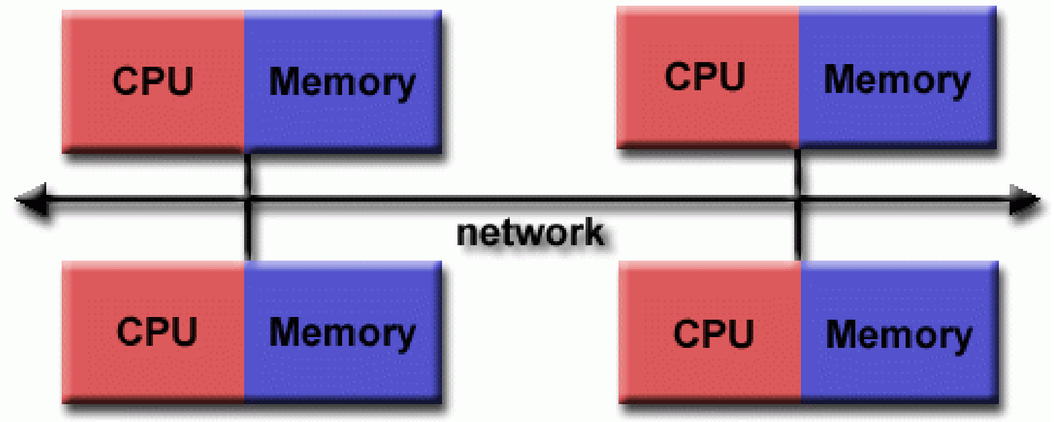


Uniform Memory Access



Non-Uniform Memory Access

Distributed Memory



https://hpc.llnl.gov/documentation/tutorials/introduction-parallel-computing-tutorial

Supercomputer

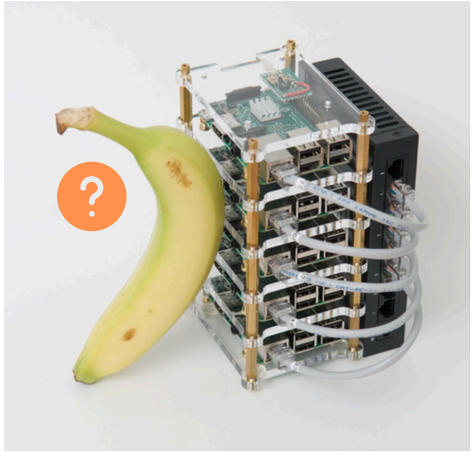
What actually is supercomputer?

https://thaisc.io/th/thaisc-resorces/lanta

- 346 node Heterogeneous HPE Cray EX cluster with a peak performance of 8.15 PFlop/s
- 176 GPU nodes with 704 NVIDIA A100 GPUs designed for GPU-intensive workload
- 160 CPU nodes with 20480 CPU-cores
- 10 High-memory nodes, each contains 4TB of memory.
- 10 PB of high-performance parallel storage
- High-performance interconnect using 200 Gbps

TOP500 rank 70 Nov 22

https://en.wikipedia.org/wiki/Frontier_%28supercomputer%29



https://www.jeffgeerling.com/blog/2021/why-build-raspberry-pi-cluster

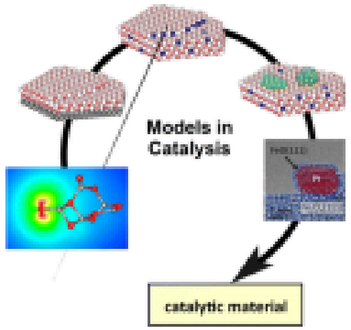
- RPI5 ~30GFlops
- To match Frontier performance we need around 43,333,333 RPI5
  - **If we don't care about networking** (XD)

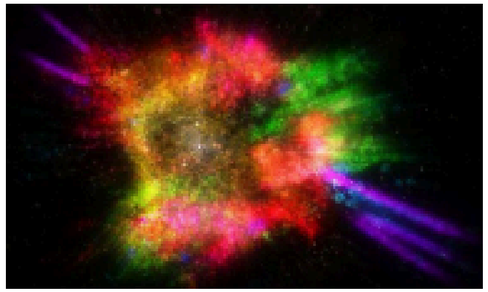
Supercomputer

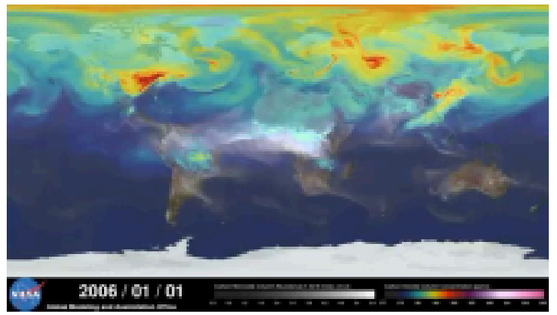
Why need Supercomputer?

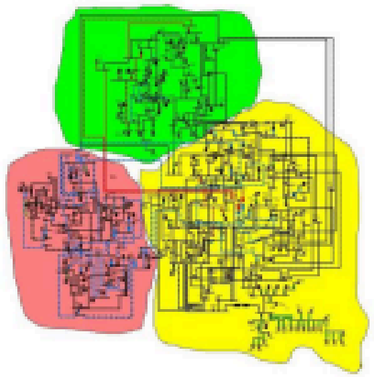
Today's Top HPC Systems Used to do Simulations

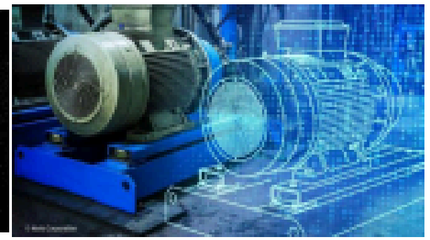
- *Climate*
- *Combustion*
- *Nuclear Reactors*
- *Catalysis*
- *Electric Grid*
- *Fusion*
- *Stockpile*
- *Supernovae*
- *Materials*
- *Digital Twins*
- *Accelerators*
- …

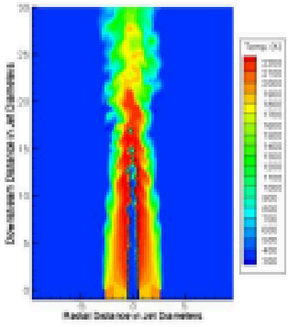


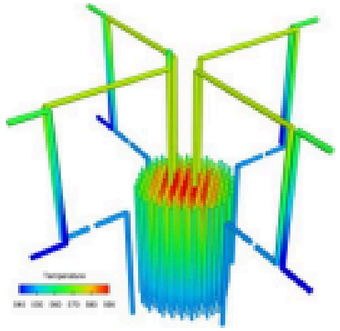


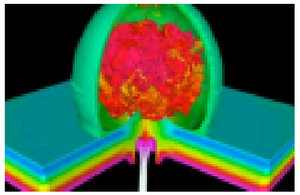


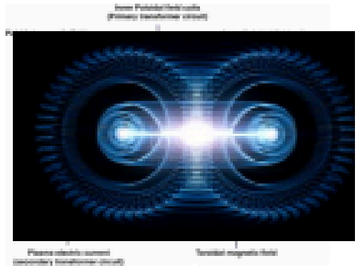






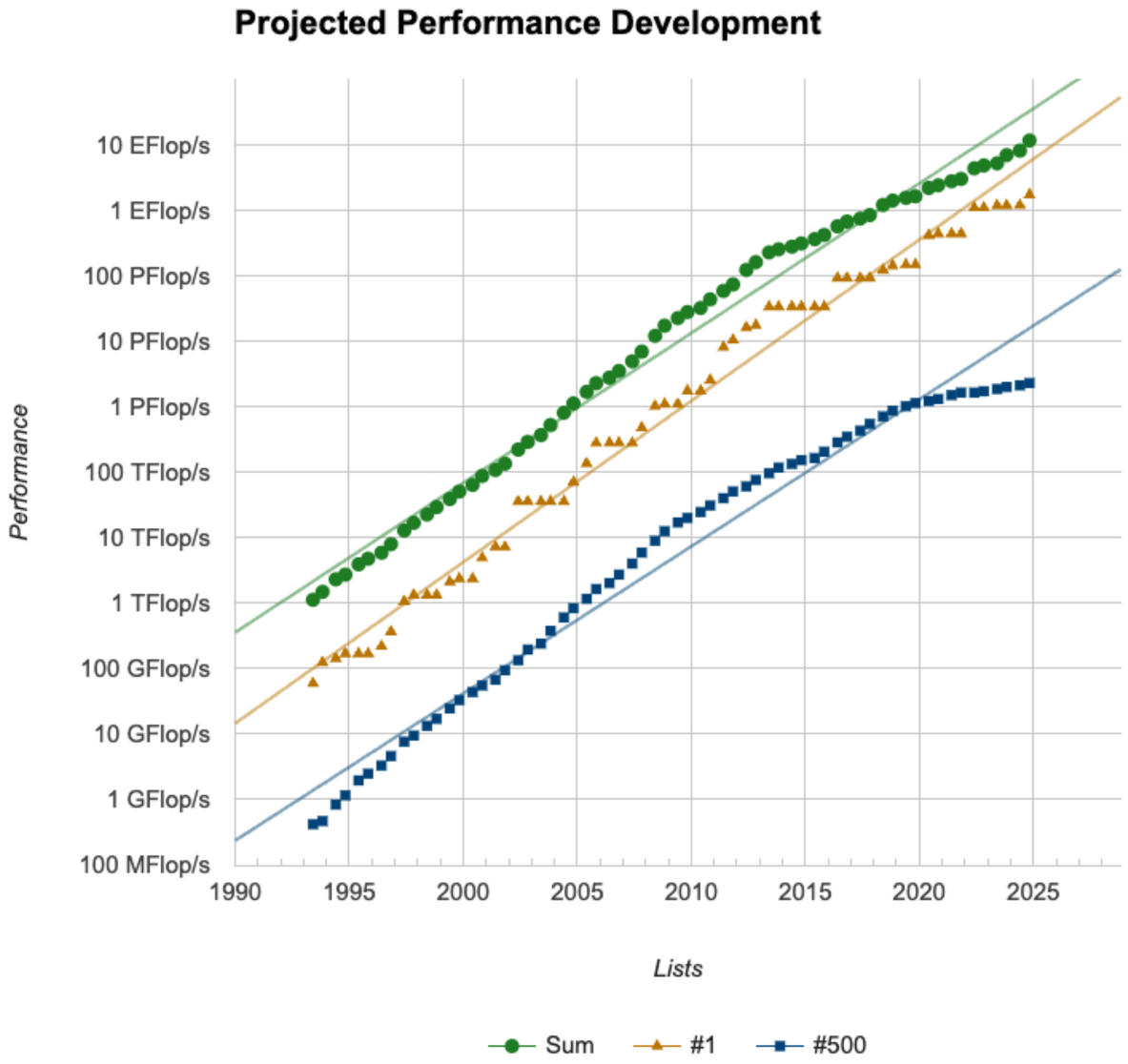






https://www.fujitsu.com/us/imagesgig5/9_evening%20keynote%20Prof.%20Dongarra%20Final.pdf

TOP500



https://top500.org/statistics/perfdevel/

An Accidental Benchmarker

LINPACK was an NSF Project w/ ANL, UNM, UM, & UCSD
We worked independently and came to Argonne in the summers

UNIT = 10**6 TIME/(1/3 100**3 + 100**2)

Facility	TIME N=100 secs.	UNIT micro-secs.	Computer	Type	Compiler
NCAR	.049	0.14	CRAY-1	S	CFT, Assembly BLAS
LASL	.148	0.43	CDC 7600	S	FTN, Assembly BLAS
NCAR	.192	0.56	CRAY-1	S	CFT
LASL	.210	0.61	CDC 7600	S	FTN
Argonne	.297	0.86	IBM 370/195	D	H
NCAR	.359	1.05	CDC 7600	S	Local
Argonne	.388	1.33	IBM 3033	D	H
NASA Langley	.489	1.42	CDC Cyber 175	S	FTN
U. Ill. Urbana	.506	1.47	CDC Cyber 175	S	Ext. 4.6
LLL	.554	1.61	CDC 7600	S	CHAT, No optimize
SLAC	.579	1.69	IBM 370/168	D	H Ext., Fast mult.
Michigan	.631	1.84	Amdahl 470/V6	D	H
Toronto	.890	2.59	IBM 370/165	D	H Ext., Fast mult.
Northwestern	1.44	4.20	CDC 6600	S	FTN
Texas	1.93*	5.63	CDC 6600	S	RUN
China Lake	1.95*	5.69	Univac 1110	S	V
Yale	2.59	7.53	DEC KL-20	S	F20
Bell Labs	3.46	10.1	Honeywell 6080	S	Y
Wisconsin	3.49	10.1	Univac 1110	S	V
Iowa State	3.54	10.2	Itel AS/5 mod3	D	H
U. Ill. Chicago	4.10	11.9	IBM 370/158	D	G1
Purdue	5.69	16.6	CDC 6500	S	FUN
U. C. San Diego	13.1	38.2	Burroughs 6700	S	H
Yale	17.1*	49.9	DEC KA-10	S	F40

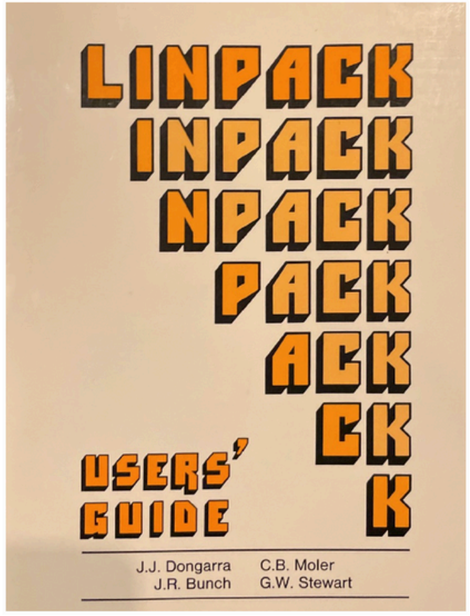
* TIME(100) = (100/75)**3 SGEFA(75) + (100/75)**2 SGESL(75)

Appendix B of the Linpack Users' Guide

- Designed to help users estimate the run time for solving systems of equation using the Linpack software.

First benchmark report from 1977

- Cray 1 to DEC PDP-10



https://www.hpcwire.com/2022/11/16/jack-dongarra-a-not-so-simple-matter-of-software/

TOP500

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
5	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461

HPL
Solving a large dense system of linear equations via Gaussian elimination

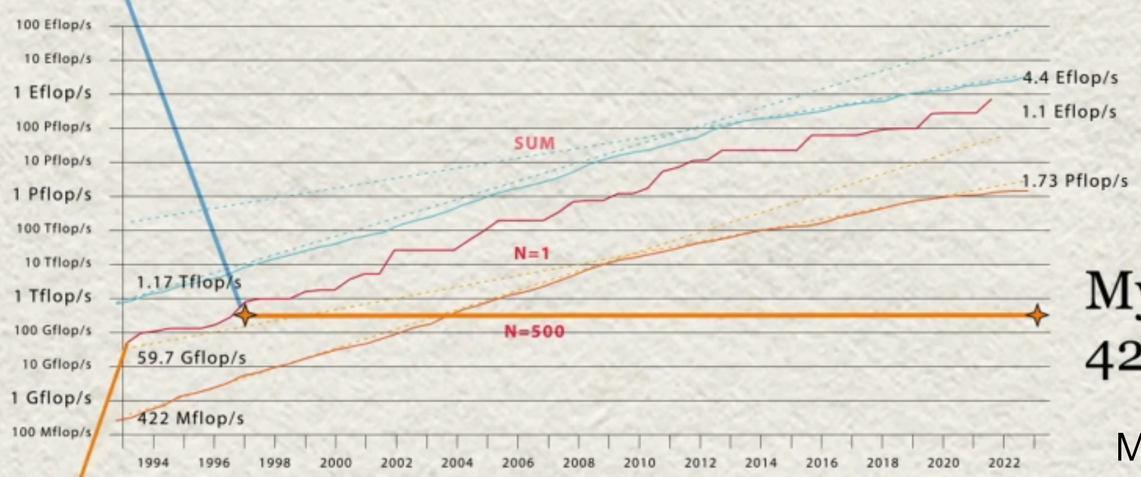
- Focuses on dense matrix computations, which are compute-bound.
- Primarily tests the peak floating-point performance.
- Not representative of the majority of scientific workloads, but useful for evaluating raw computing power.

https://top500.org/lists/top500/2024/11/

TOP500

Performance Development of HPC over the last 30 years from the Top 500

#1 in 1997 - Hitachi CP-PAC with 2048 Processors at Center for Computational Science, University of Tsukuba



My Laptop: 426 Gflop/s

MACBOOK Air M2

1 in 1993 - Thinking Machine CM-5 with 1024 Processors at Los Alamos Nat Lab used for nuclear weapons design

2021 AM Turing Award Recipient Jack Dongarra Turing Lecture: "A Not So Simple Matter of Software"

HPCG

High-Performance Conjugate Gradient (HPCG) Benchmark

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	HPCG (TFlop/s)
1	6	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	16004.50
2	2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	14054.00
3	3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	5612.60
4	8	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	4586.95
5	7	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	2,121,600	434.90	3671.32

HPCG

- Focuses on sparse matrix computations, which are memory-bound.
- Tests memory access patterns, network latency, and bandwidth.
- Represents workloads like computational fluid dynamics and finite element analysis.

https://hpcg-benchmark.org/

HPL-MxP

HPL-MxP Mixed-Precision Benchmark

November 2024

Rank	Site	Computer	Cores	HPL-MxP (Eflop/s)	TOP500 Rank	HPL Rmax (Eflop/s)	Speedup
1	DOE/SC/ANL	Aurora	8,159,232	11.643	3	1.0120	11.5
2	DOE/SC/ORNL	Frontier	8,560,640	11.390	2	1.3530	8.4
3	EuroHPC/CSC	LUMI	2,752,704	2.350	8	0.3797	6.2
4	RIKEN	Fugaku	7,630,848	2.000	6	0.4420	4.5
5	EuroHPC/CINECA	Leonardo	1,824,768	1.842	9	0.2412	7.6
6	CII, Institute of Science	TSUBAME4	172,800	0.641	47	0.0255	25.2
7	NVIDIA	Selene	555,520	0.630	23	0.0635	9.9
8	DOE/SC/LBNL	Perlmutter	888,832	0.590	19	0.0792	7.4
9	FZJ	JUWELS BM	449,280	0.470	33	0.0441	10.7
10	GENCI-CINES	Adastra	319,072	0.303	30	0.0461	6.6

HPL-MxP

- **Mixed-Precision Strategy**: Combines low-precision computations for speed with high-precision refinement for accuracy.
- **Hardware Utilization**: Leverages specialized hardware features, such as tensor cores in GPUs, to maximize performance.
- **Benchmarking Relevance**: Provides a more representative measure of system performance for modern workloads that blend **HPC and AI** tasks

Benchmark	Precision Focus	Workload Representation	Use Case
HPL	Double (FP64)	Traditional HPC	TOP500 rankings
HPCG	Double (FP64)	Memory-bound applications	Real-world HPC workloads
HPL-MxP	Mixed (e.g., FP16/FP32 + FP64)	AI and HPC convergence	AI-accelerated HPC performance

https://hpl-mxp.org/results.md

GREEN500

Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	222	**JEDI** - BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, ParTec/EVIDEN EuroHPC/FZJ Germany	19,584	4.50	67	72.733
2	122	**ROMEO-2025** - BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, Red Hat Enterprise Linux, EVIDEN ROMEO HPC Center - Champagne-Ardenne France	47,328	9.86	160	70.912
3	440	**Adastra 2** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, RHEL, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France	16,128	2.53	37	69.098
4	155	**Isambard-AI phase 1** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE University of Bristol United Kingdom	34,272	7.42	117	68.835
5	51	**Capella** - Lenovo ThinkSystem SD665-N V3, AMD EPYC 9334 32C 2.7GHz, Nvidia H100 SXM5 94Gb, Infiniband NDR200, AlmaLinux 9.4, MEGWARE TU Dresden, ZIH Germany	85,248	24.06	445	68.053

GREEN500

- Evaluates the energy efficiency of supercomputers
- Measures **performance per watt** using the TOP500 measure of high performance LINPACK benchmarks at double-precision floating-point format.

18	1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	29,581	58.889
22	2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	24,607	54.984
64	3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	38,698	26.151

https://top500.org/lists/green500/2024/11/

ARCHITECTURE

Components

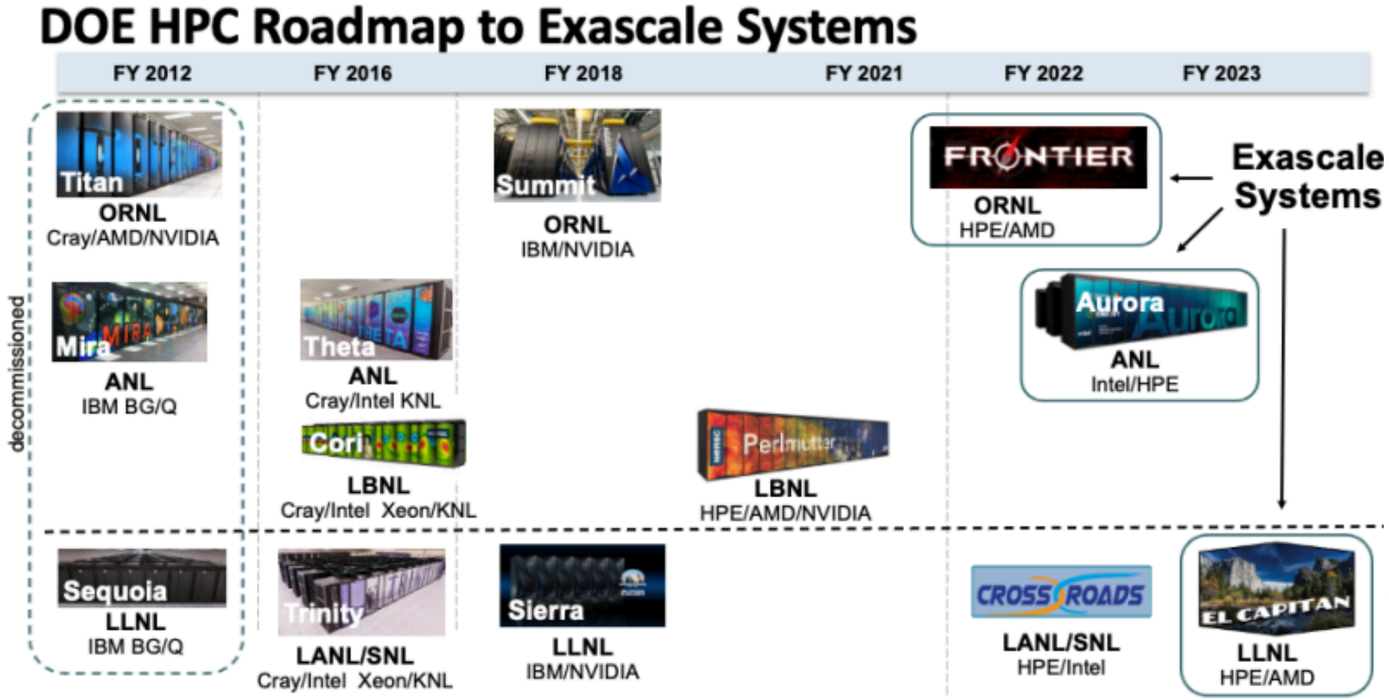


Figure 1: DOE HPC Roadmap to Exascale Systems

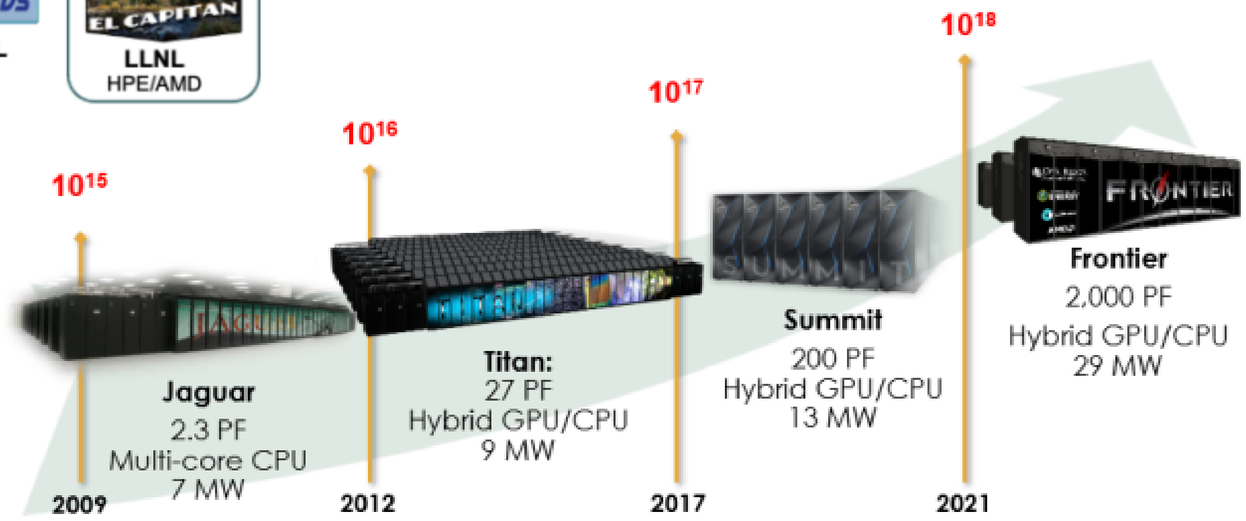


Figure 2: Four Generations of Supercomputers at ORNL

https://icl.utk.edu/files/publications/2022/icl-utk-1570-2022.pdf

Components

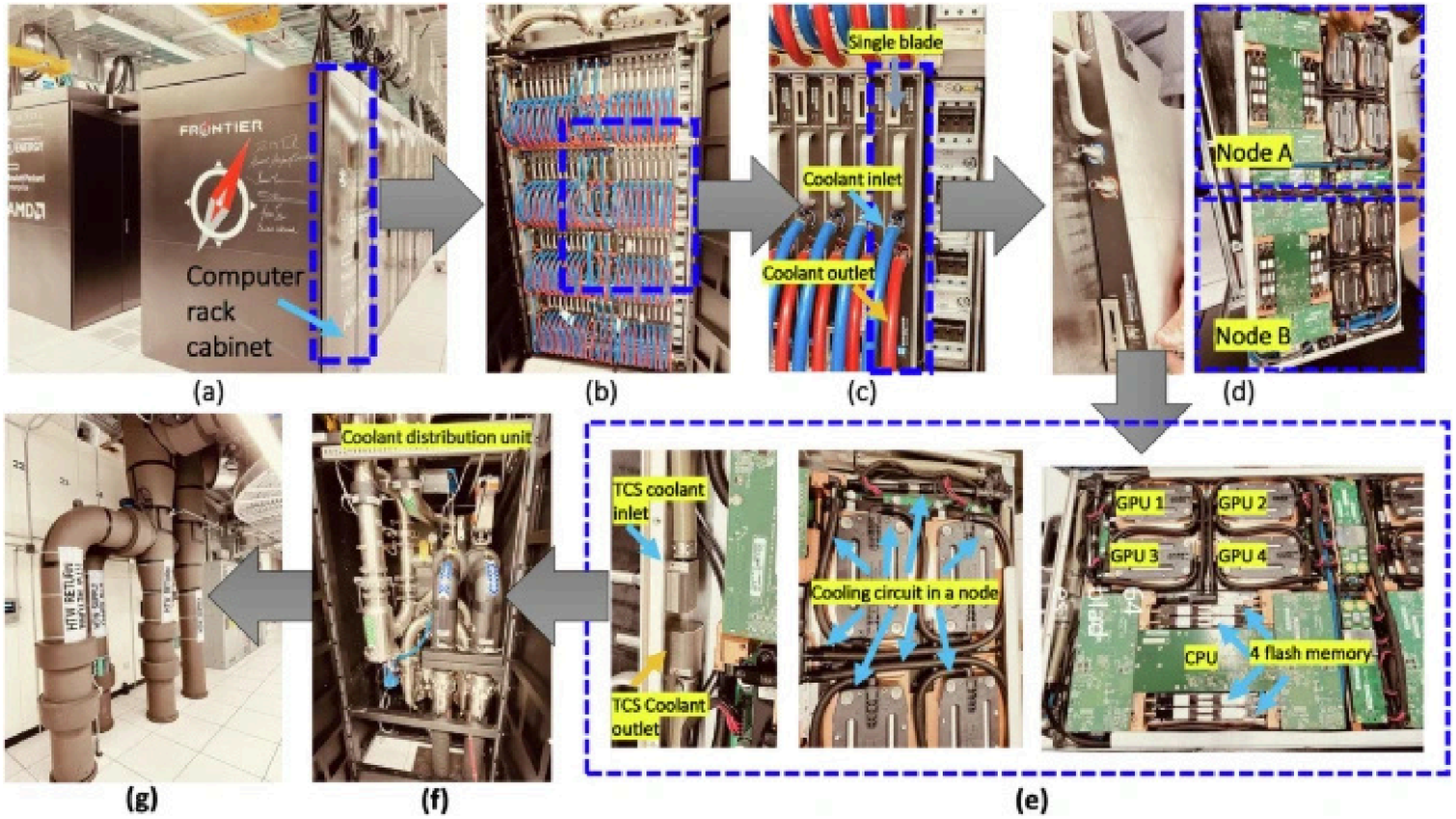
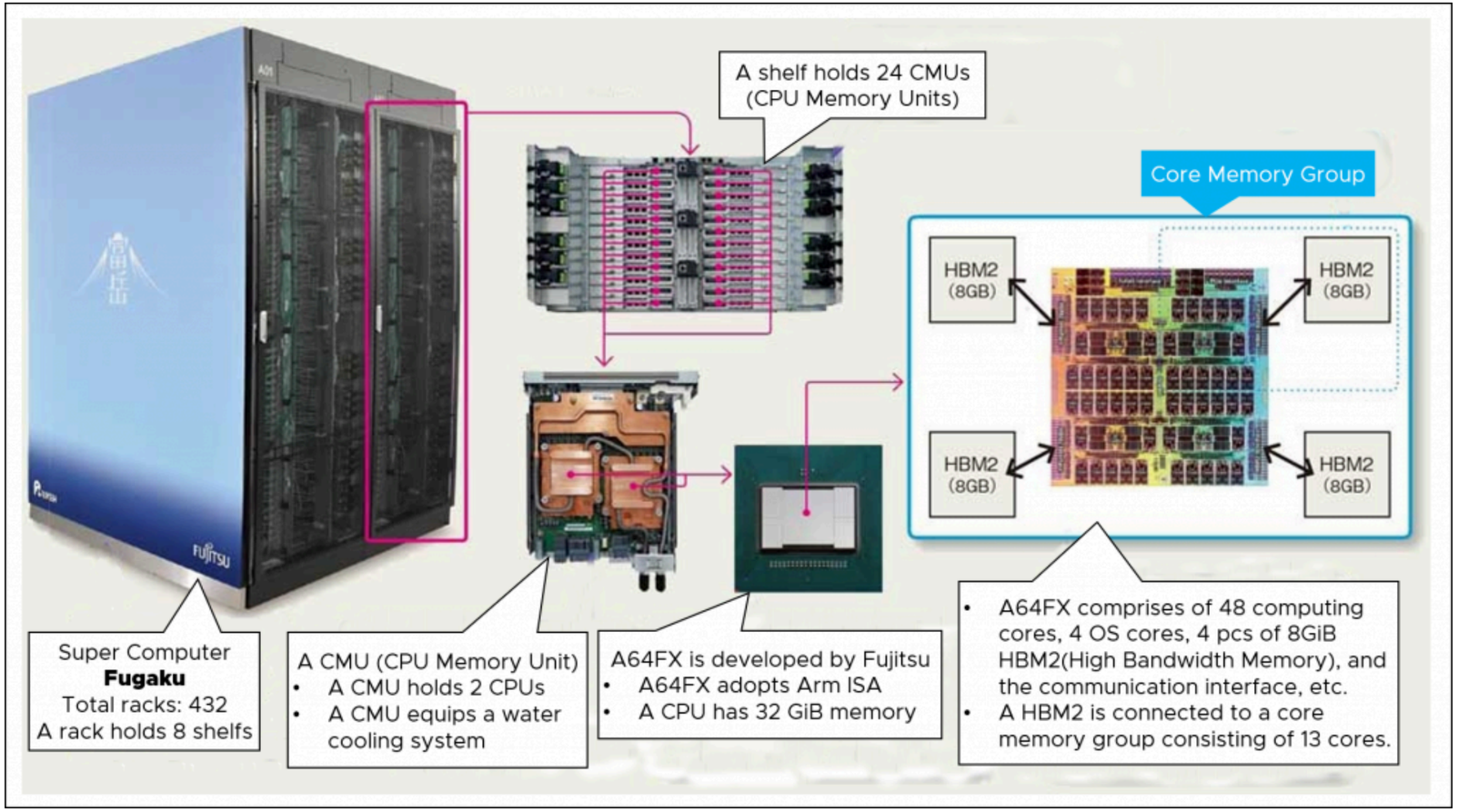
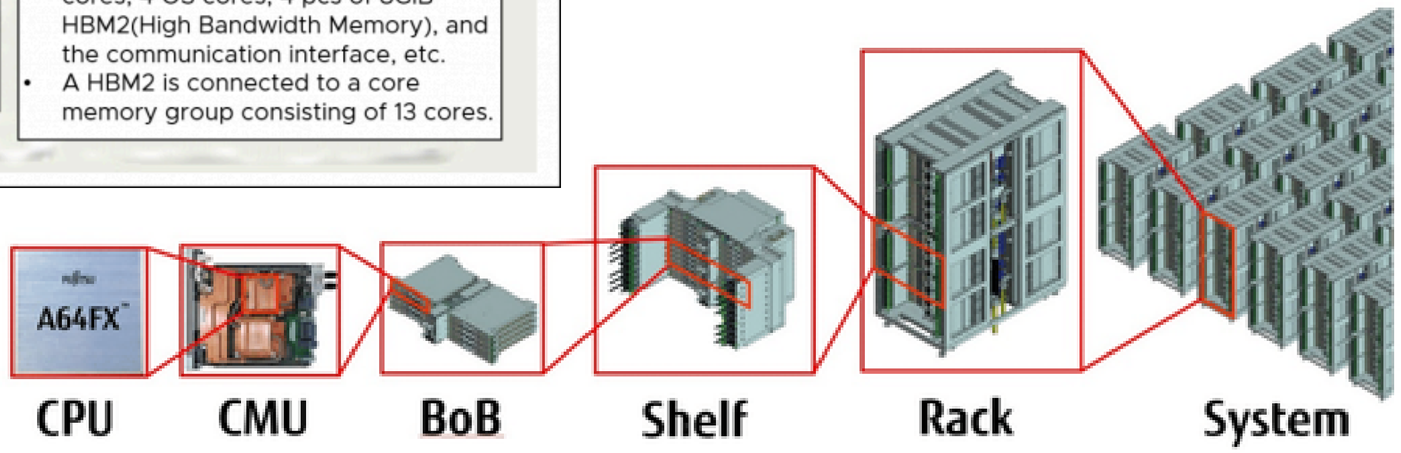


Figure14: One Rack Containing 64 Blades (128 nodes)

https://icl.utk.edu/files/publications/2022/icl-utk-1570-2022.pdf
https://www.nature.com/articles/s41597-024-03913-w

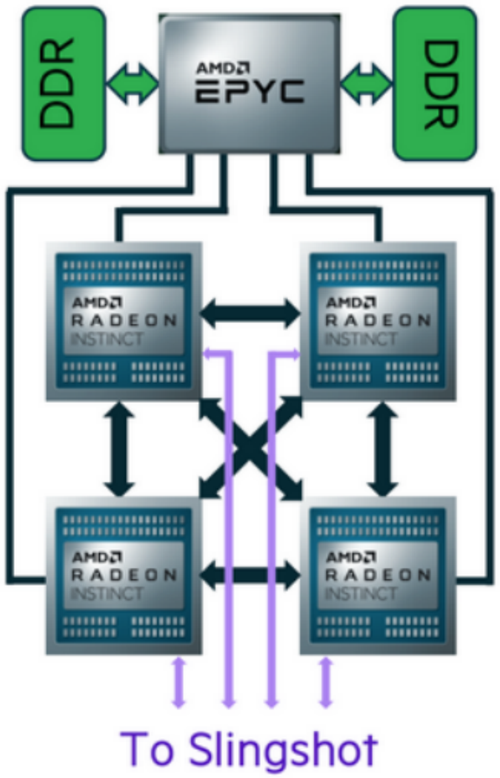
Components





https://hacarus.com/ai-lab/20210405-fugaku/
https://insidehpc.com/2020/06/arm-based-fugaku-stands-on-summit-of-new-top500/

Compute Node



Frontier Compute Blade (Two Nodes)

Front view showing thickness and
Removal handle, hot, and cold water inputs

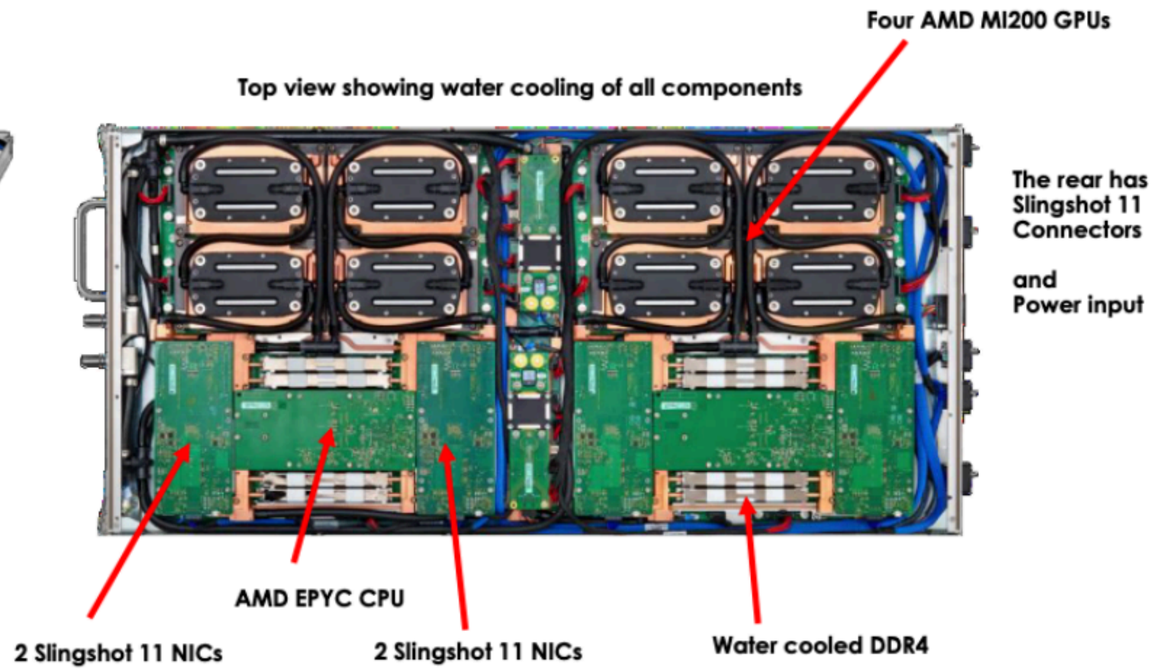
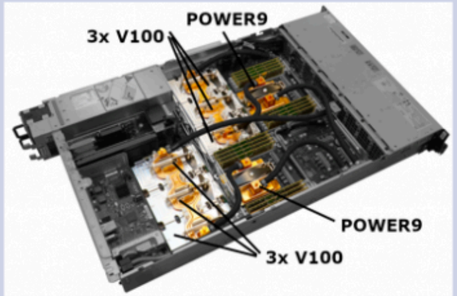
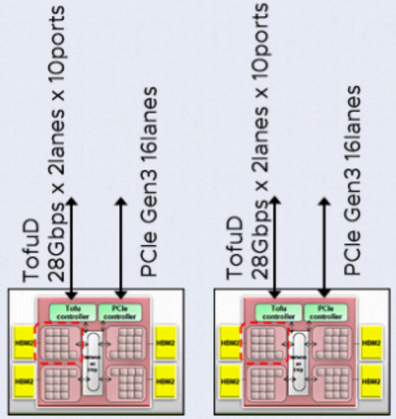
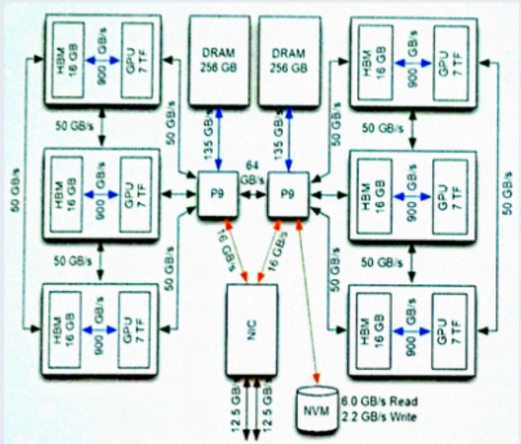


Figure13: Frontier Compute Blade made up of 2 nodes

https://icl.utk.edu/files/publications/2022/icl-utk-1570-2022.pdf

Compute Node

Node hardware comparison

	Fugaku	Summit
Node picture		
Circuit block diagram		
The number of major ICs on a PCB	2	10

- Fugaku: A white package includes an A64FX with 4 pcs of HBM2. Two white packages are counted as two major ICs. A PCB holds two nodes.

  Ref.: https://news.mynavi.jp/article/20191202-931937/

- Summit: 2 pcs of POWER9, 6 pcs of Nvidia's Tesla V100 with a HBM, 256GB DRAM, and 1 pc? of Mellanox EDR are counted as 10 major ICs. A PCB holds two nodes. Please refer to Figure 5.

  Ref.: https://news.mynavi.jp/article/summit-1/

- PCB stands for Printed Circuit Board.

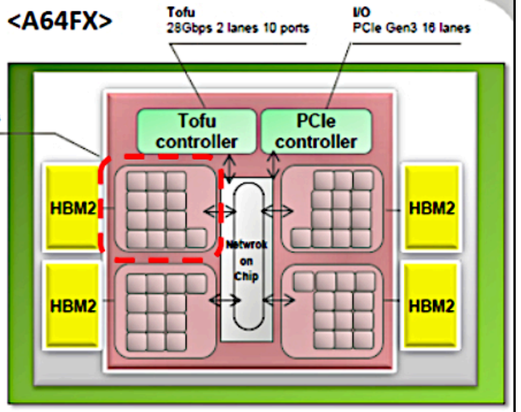
A64FX Chip Overview

FUJITSU

Architecture Features

- Armv8.2-A (AArch64 only)
- SVE 512-bit wide SIMD
- 48 computing cores + 4 assistant cores*
  *All the cores are identical
- HBM2 32GiB
- Tofu 6D Mesh/Torus
  28Gbps x 2 lanes x 10 ports
- PCIe Gen3 16 lanes

<A64FX>



7nm FinFET

- 8,786M transistors
- 594 package signal pins

Peak Performance (Efficiency)

- >2.7TFLOPS (>90%@DGEMM)
- Memory B/W 1024GB/s (>80%@Stream Triad)

	A64FX (Post-K)	SPARC64 XIfx (PRIMEHPC FX100)
ISA (Base)	Armv8.2-A	SPARC-V9
ISA (Extension)	SVE	HPC-ACE2
Process Node	7nm	20nm
Peak Performance	>2.7TFLOPS	1.1TFLOPS
SIMD	512-bit	256-bit
# of Cores	48+4	32+2
Memory	HBM2	HMC
Memory Peak B/W	1024GB/s	240GB/s x2 (in/out)

https://hacarus.com/ai-lab/20210405-fugaku/

Storage System

The storage system consists of three primary layers:

- Cache for global file system
- Temporary file systems
  - Local file system for compute node
  - Shared file system for a job
- Global file system

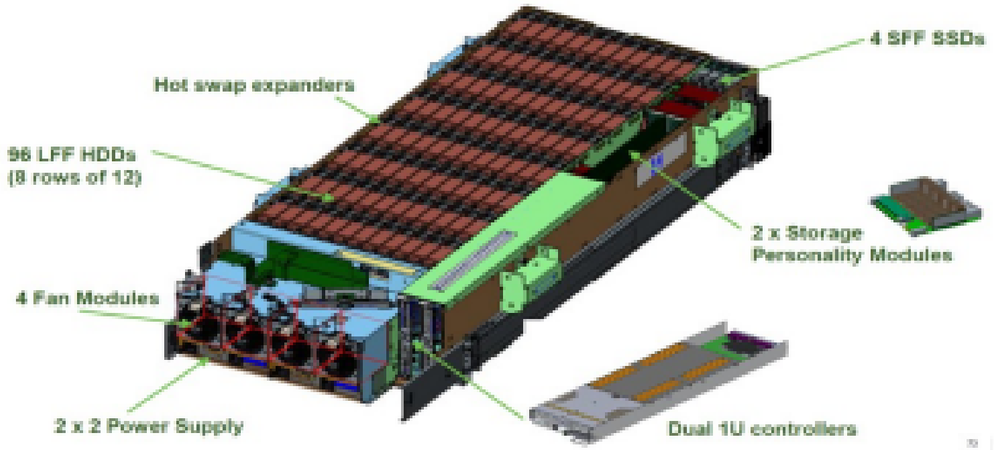


Figure 9: Moose HDD Storage board (Capacity Tier)

There are two - 2 TB SSD NVM per node of local storage (Flash).

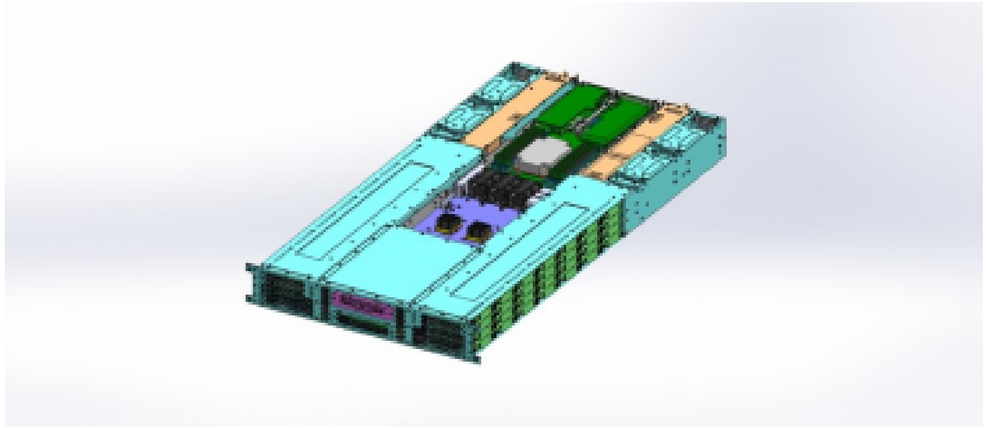


Figure 8: Gazelle SSD Storage board (Performance Tier and Metadata)

- DDN Exascaler
- DAOS
- BeeGFS
- SeaweedFS
- GlusterFS

https://icl.utk.edu/files/publications/2022/icl-utk-1570-2022.pdf

Networking System

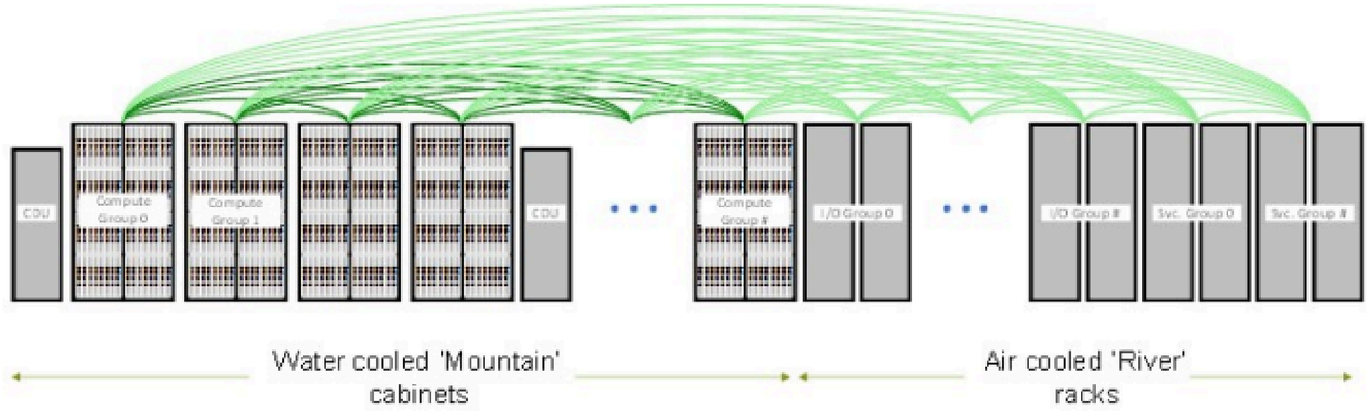


Figure7: Cabinet Interconnect

Figure 12: One row of the Frontier System as Installed at DOE's ORNL

- Inter-node communication
  - Compute ↔ Compute
  - Compute ↔ Storage
- Intra-node communication
  - CPU ↔ GPU
  - CPU ↔ DPU

- 400GB/s, 800GB/s

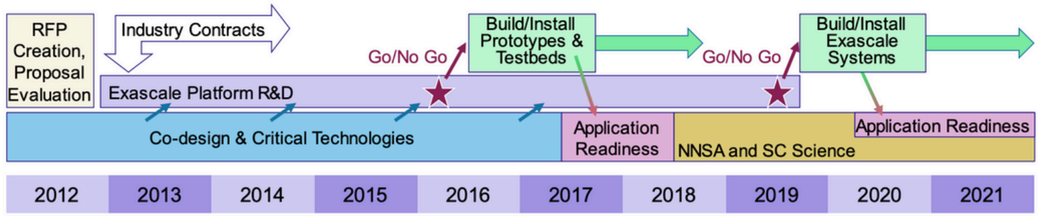
- NVLink https://www.nvidia.com/en-us/data-center/nvlink/
- UALink https://ualinkconsortium.org/

https://icl.utk.edu/files/publications/2022/icl-utk-1570-2022.pdf

Case Study: Frontier

Back in 2012

- Around 2008, we have entered **Petaflop** Era (https://ieeexplore.ieee.org/document/5217926)
- In 2011, DOE (Department of Energy) want to enable **Exascale** by the year 2019-2020 (?) and start a *Forward Project (star-Forward)
- **Exascale** → Hitting 1 EF or 1000PF in HPL Benchmark
- The requirement is simple yet challenging
  - Not surprising is the performance target of 1000PF (an exaflop), but it is worth noting that the **DOE also calls out an objective of 300PF** on (to be determined) other workloads that may not be as regular or computationally intense as LINPACK.
  - A machine designed only to win on a single benchmark **would not be acceptable** as it would have limited broader scientific utility.
  - **20MW** limit is only for the computational components and is not inclusive of storage systems or facility infrastructure such as power delivery and cooling)



(a)

Exascale System	Goal
Delivery Date	2019-2020
Performance	1000 PF LINPACK 300 PF on to-be-specified applications
Power Consumption	20 MW
Memory Capacity (incl. NVRAM)	128 PB
Node Memory Bandwidth	4 TB/s
Node Interconnect Bandwidth	400 GB/s

(b)

Figure 1. (a) Exascale timeline and (b) system objectives from the 2011 U.S. DOE exascale research and development Request for Information.

Exascale computing is a great achievement, but it is merely a milestone and not a destination. The technical challenges going forward in what is now the post-Exascale era are only getting more difficult, and we will collectively need to harness the entire community's creativity and innovation!

If each person on Earth completed 1 calculation per second, it would **take more than 4 years** to do what an Exascale computer can do in 1 second.

A Research Retrospective on AMD's Exascale Computing Journey

Case Study: Frontier

Known Technology

- AMD began its Exascale collaboration with the DOE in **2012**. We will know how they planned their design for decade ahead.
- The trend toward **general purpose GPU computing** was already underway in the industry, and so the inclusion of accelerated computing as a performance engine was a natural component to consider.
- At that time, AMD just release their APU (CPU+GPU in single chip) which provides a Unified memory
- Not until 2015 when AMD launched the AMD Radeon™ 300 series GPUs featuring silicon interposer technology and 3D high-bandwidth memory (HBM). Because of the requirements 4TB/s node memory bandwidth

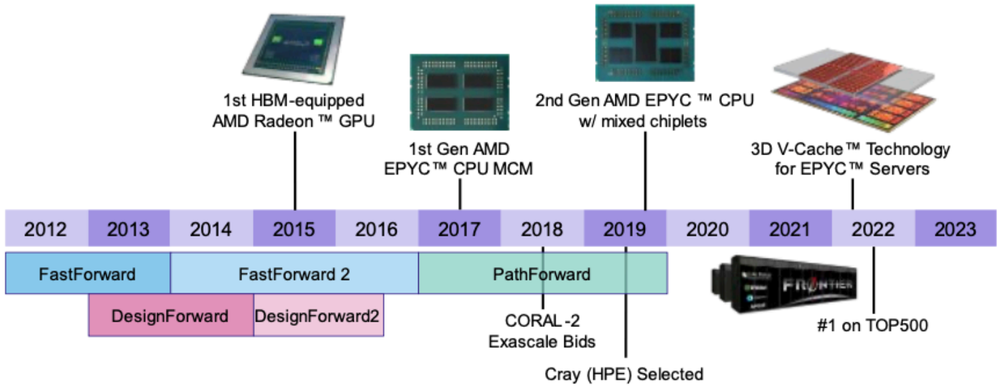
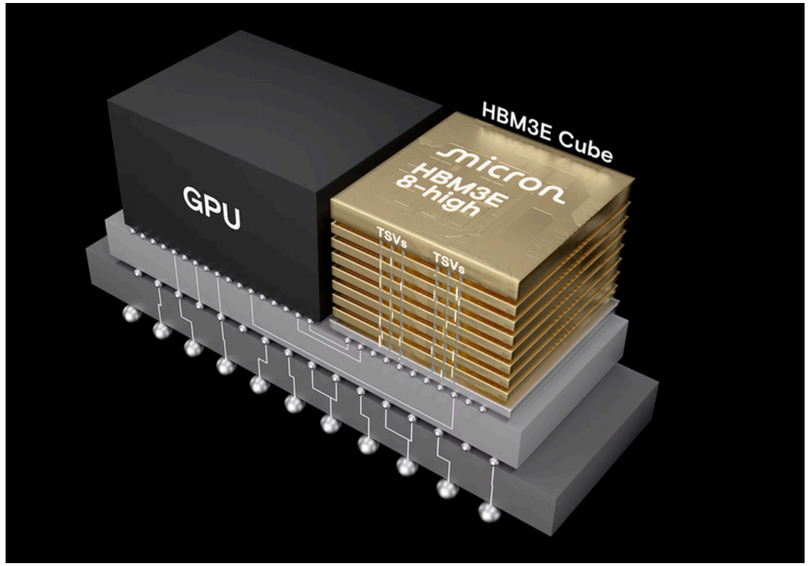


Figure 2. Timeline illustrating U.S. DOE exascale R&D programs and milestones (bottom) and key AMD technology introductions (top).

Unified memory → 2 or more processors (in this case CPU and GPU) share the same memory. Unified Memory reduce data movement. No need to copy back and forth
High-bandwidth Memory -> a type of computer memory that is optimized for fast data transfer and reduced power consumption.

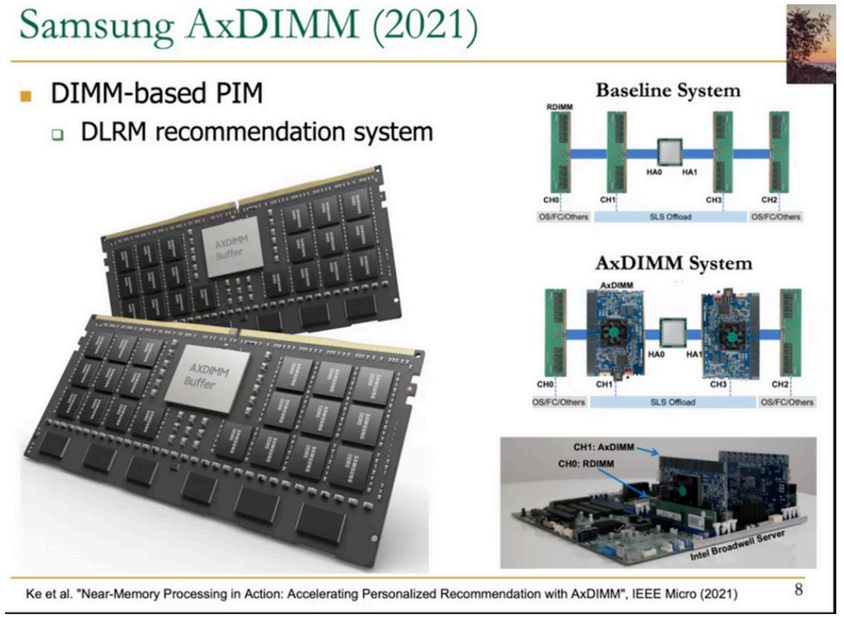


https://www.micron.com/products/memory/hbm
A Research Retrospective on AMD's Exascale Computing Journey

Case Study: Frontier

Future Technologies?

- **The end of Moore's Law** is not surprising, silicon technology would slow down for sure.
  - At the time (2012), AMD was shipping products in a **32nm** process technology. Projecting forward, we predicted that in 2019-2020 we would be in a **10nm** technology. (TSMC began commercial production of "10 nm" chips in 2016)
- AMD think that by the time of Exascale era, **3D stacking** technology should be a thing (not only stacking memory, but also processors)
  - Reduce bandwidth between processors or memory
- **Processing In Memory** (PIM)
  - Minimizing data movement was seen as an important aspect of improving bandwidth while simultaneously reducing power consumption.
- Not only processors that could be slow down in term of scaling in the future, **DRAM also face the same challenges**
  - **NVRAM** could potentially provide a path to achieving the DOE's system-level memory capacity targets in a more scalable and/or cost-effective manner.
- At that time, **Photonic interconnects** promise to significantly increase the amount of bandwidth



NVRAM → a type of computer memory that retains its stored information even when the power is turned off
Processing In Memory -> is the practice of taking action on data entirely in computer memory (e.g., in RAM)
Photonic interconnects → Photonics interconnects use light (typically via optical fibers or integrated photonics) to transmit data between components, rather than using electrical signals through copper wires

Real-world PIM: UPMEM PIM Architecture
https://opg.optica.org/jlt/abstract.cfm?uri=jlt-30-4-448
https://www.lenovo.com/th/en/glossary/nvram/
A Research Retrospective on AMD's Exascale Computing Journey
https://hazelcast.com/foundations/data-and-middleware-technologies/in-memory-processing/

Case Study: Frontier

Exascale Heterogeneous Processor

- Exascale Heterogeneous Processor (EHP) is what we see as a compute node
  - EHP has a lots of versions and revision
  - Each version has their own improvement compare to previous version
  - Some changes are aligned with commercial AMD products

EHP v1 (2012)

- EHP is a high-performance APU coupled with 3D DRAM (HBM) memory system.
- GPU (Vector Units) provides **12TF**
- APU (CPU+GPU) share a single unified memory system consisting **128GB** in-package DRAM with 4TB/s bandwidth
- Also connect to multiple **NVRAM** modules outside of the package.
- Integrated NIC with photonic interconnects
- EHP **stacked CPU die on top of 3 layers of GPU/Vector units**
  - Thought: 3D stacking is needed to maximize the compute in limited packaging
  - Also minimize cost of data movement between compute components
- Not only APU die are stacked,DRAM and NVRAM also stacked
- 3D Stacking bring new problem: **Thermal**
- Each EHPv1 has 12TF of peak compute power, they need **83,334** EHP to reach 1.0 exaflop.
  - It is unlikely for every node to operate at 100% Efficiency, so we need around **100K nodes** to sustain an exaflop computer.
  - With 100K nodes, we have an upper bound of **200W per node**. (20MW limit)
  - Networking for 100K nodes is also challenging

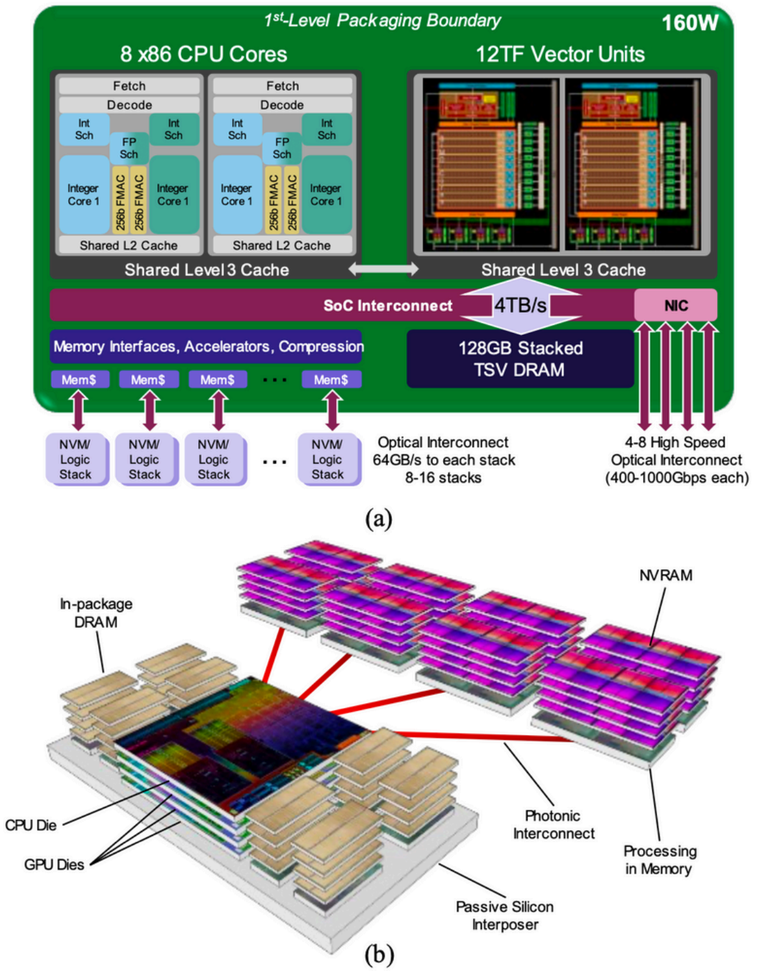


Figure 3. (a) Block diagram of the Exascale Heterogeneous Processor (EHP) concept from the original FastForward program circa 2012, (b) illustrative packaging view of the EHP.

A Research Retrospective on AMD's Exascale Computing Journey

Case Study: Frontier

EHP v2 (2014)

- With increasing silicon cost trend for leading-edge technology nodes, led AMD to pioneer **chiplet technology** for building processors. For example, **AMD EPYC processor**
- Chiplet technology allow AMD to **reuse** their silicon components in multiple product configurations.
- EHPv2 utilize both **CPU Chiplets (CCD) and the IO die (IOD**) from their products
- This decision try to backing away from aggressive 3D-stacked
  - with a cost of consuming a significant portion of the package
  - note that they don't stack CPU on top of GPU anymore
- To offset the loss of packaging area for 3D DRAM, they move the DRAM to stack on top of GPU instead
  - GPU also partitioned into chiplet
- To maintain the same total capacity of DRAM, they double the layer in DRAM stacks
  - previously 8 (though in figure is 4), now **16 per stack**
- **NVRAM was removed!**
  - DOE said that modifying their algorithm to make use of multi-tiered memory organization is challenging. (L1,L2,L3,DRAM,NVRAM)
  - This is not a technical problem with NVRAM
- **Issues with V2:**
- HBM only implemented eight-high stacks (aimed for 16)
- GPU has a problem with **thermal**
- AMD Packaging engineers concerns about the asymmetry of the overall package (CPU on one side and GPU on other).
- Routing I/O and other external memory would be challenging. IOD is not in center of package.

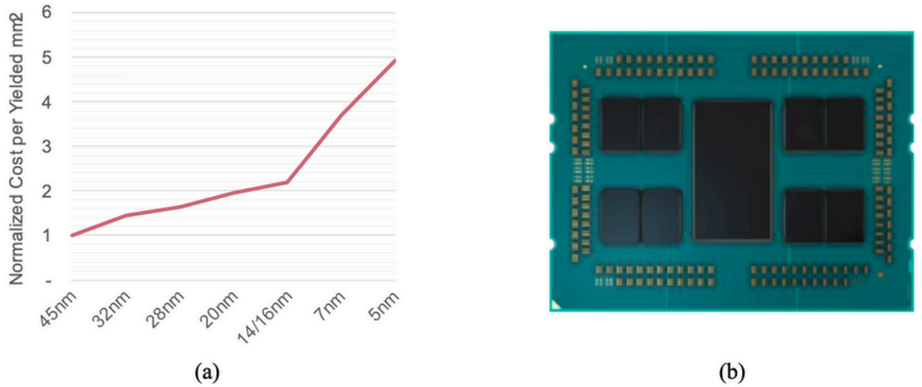


Figure 4. Silicon cost trends over time and (b) an AMD EPYC™ processor utilizing chiplets.

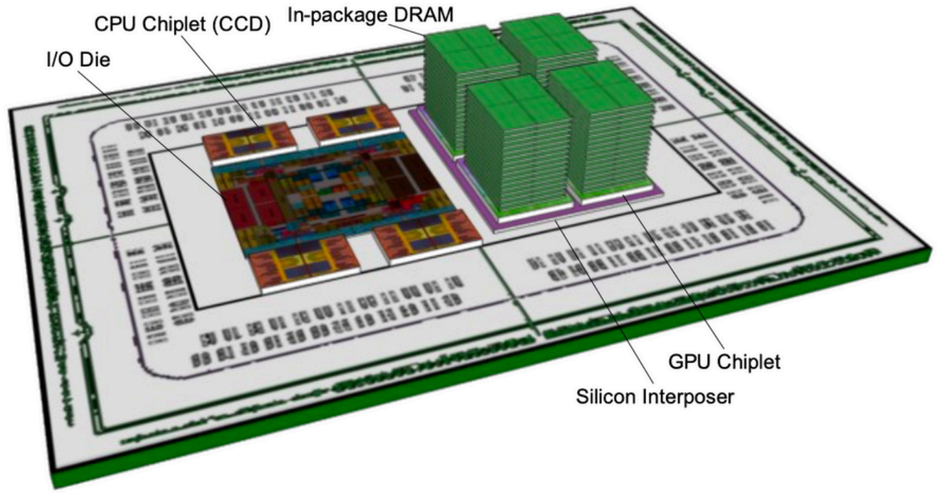


Figure 5. Refinement of the EHP (v2), circa 2014.

A Research Retrospective on AMD's Exascale Computing Journey

Case Study: Frontier

EHP v3 (2016)

- 16-high memory stacks would be unlikely in the target timeframe. So they revert to 8-high DRAM stacks.
- **Issues with V3:**
- power density of GPU regions still present thermal challenges

EHP v4 (2018)

- **New packaging** alternatives to make more room, previous iterations have suffered from density of chips and stacking memory
- Instead of 8 GPU chiplets **into 2 larger pieces of silicon**
  - smaller chiplets good for cost benefits
  - but required higher bandwidth to support data movement and work distribution among the GPU compute units would be far less efficient to route among chiplets
- **Rebalance CPU-to-GPU compute ratio**, reduce CCDs from 8 in v3 to 2
  - Allowed AMD to reuse IOD and CCDs from mainstream server parts
- Now we have 2 GPUs, how can we route GPU-to-GPU traffic between them?
  - one option is to route through IOD, utilizing IOD's existing Infinity Fabric (IF)
  - AMD choose to add **multiple IF links** directly across the package between the GPU chips
- **Issues with v4:**
- The remaining concerns were less about fundamental technologies but rather more about business.
  - EHP fixes the ratio of CPU-to-GPU compute resources for a given package, and customers with **different workload requirements might prefer different ratios**
  - some might need more GPU, some might not need GPU at all

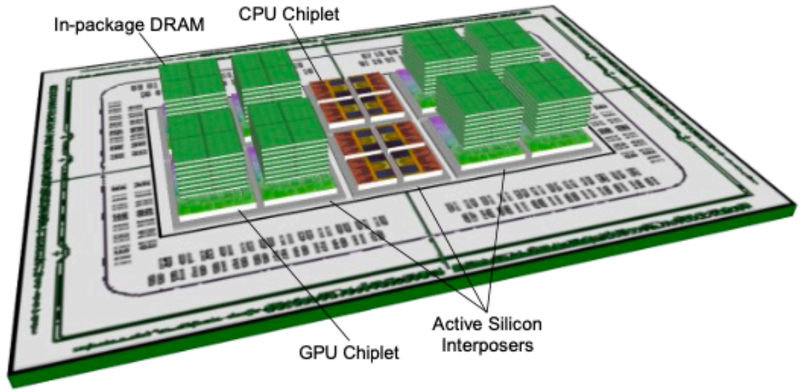


Figure 6. Refinement of the EHP (v3), circa 2016.

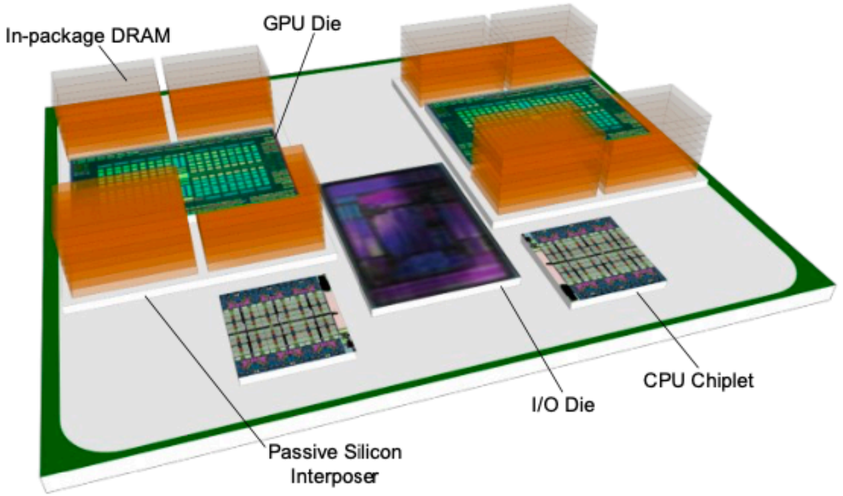


Figure 7. Refinement of the EHP (v4), circa 2018.

A Research Retrospective on AMD's Exascale Computing Journey

Case Study: Frontier

Discrete Node Architecture (DNA)

- The EHP target the APU architecutre to make it easier to program by **supporting unified memory and cache coherence between processor** types.
- Their research also considered discreate node architecture (DNA)
- DNA still support cache coherence and a flat physical address space like an APU, albiet at lower bandwidths and higher latencies
- DNA architecture was attractive to some partner in this project (*Forward)
  - separation of CPU and GPU components allowed more customization
  - Example; change to another GPU vendor

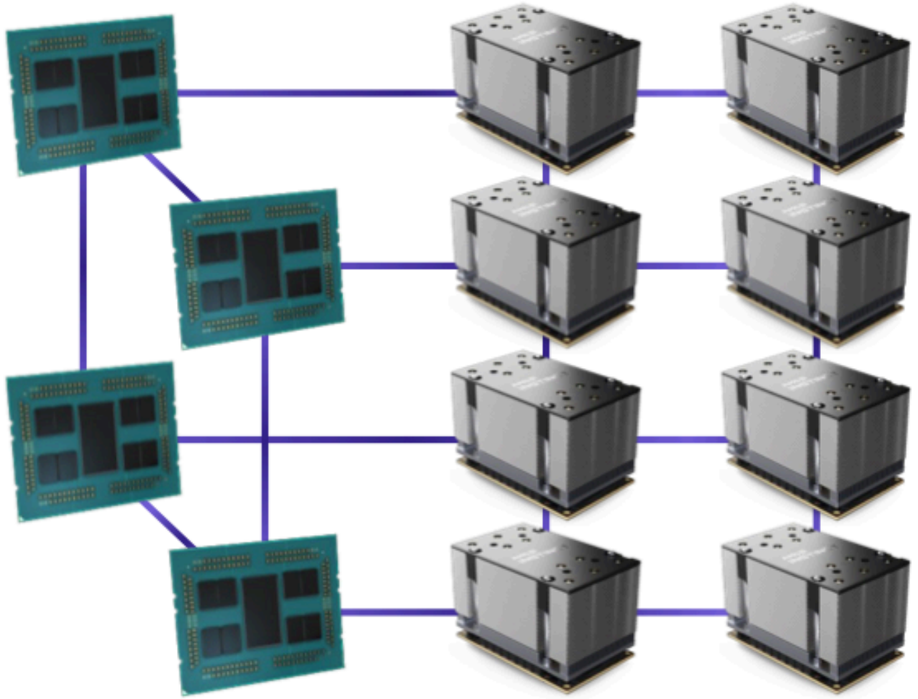


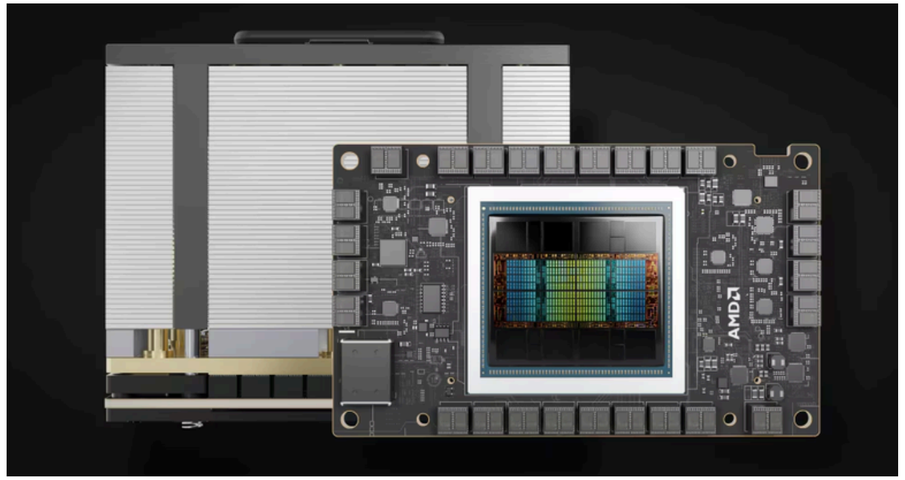
Figure 8. Discrete Node Architecture consisting of interconnected CPUs (left) and accelerators (right).

A Research Retrospective on AMD's Exascale Computing Journey

Case Study: Frontier - Research Topics

Compute-Optimized GPUs

- What could GPU look like if it did not actually have to worry about graphics?
  - GPU were originally designed for graphics rendering
- “Compute-Optimized GPU” is a GPU that **removed** all the specialized hardware that is only used for graphics rendering tasks. (color units, spline interpolation, depth processing)
- Current AMD’s GPU product lines
  - **RDNA** architecture targets gaming and graphics
  - **CDNA** architecture services HPC and machine learning



https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html

A Research Retrospective on AMD's Exascale Computing Journey

https://wccftech.com/amd-radeon-rx-8000-rdna-4-gpu-leak-56-compute-units-2-1-ghz-16-gb-vram/

Case Study: Frontier - Research Topics

CPU Core Microarchitecture

- researching techniques and enhancements for “traditional” CPU microarchitecture
- general topics such as branch prediction, instruction fetch, scheduling, caching, and prefetching
- Also focused on how to improve the CPU architecture specifically for the types of compute and memory patterns exhibited by the DOE’s workloads, which is not act like commercial benchmarks

Power-Performance Efficiency

- Achieving 1 Exaflop in 20MW (50GF/W) required a lot of improvements of current technologies
- improving efficiency of CPU and GPU microarchitecture, data movement and networks on chips, caches and memory, circuits, and software algorithms.

Programming Model and Software Optimization

- Need to porting DOE exascale proxy application into new system

Reliability

- The Reliability and Resiliency research explored three areas
  - understanding the nature of faults that occur in real HPC systems in the field
  - development of early-stage architectural fault modeling techniques and tools for EHP architecture and new fault modes
  - exploring low-cost pervasive fault detection techniques for GPUs

Multi-Level Memory

- With two-tier memory system (DRAM+NVRAM) in first version of EHP
- One key challenge is that many research on this topic work well on average, but they can still suffer from access patterns that cause performance to drop unacceptably
  - now programmer need to resolve that D:

A Research Retrospective on AMD's Exascale Computing Journey

Case Study: Frontier - Research Topics

Processing In Memory

- moving your code to data
- The technology readiness of PIM did not align with the exascale schedule, but multiple recent industry announcements about PIM could indicate that its time may soon be coming

Integrated Silicon Photonics

- Similar to PIM, the amount of recent work and startups developing prototypes and testbeds brings hope that integrated photonics might not be too far off.

Outcome

- In May 2019, the U.S. DOE announced that it had contracted with Cray to build the Frontier supercomputer to delivered to ORNL
- Frontier Node Architecture leverages concepts from both EHP and DNA approaches

From EHP to the Frontier Node Architecture

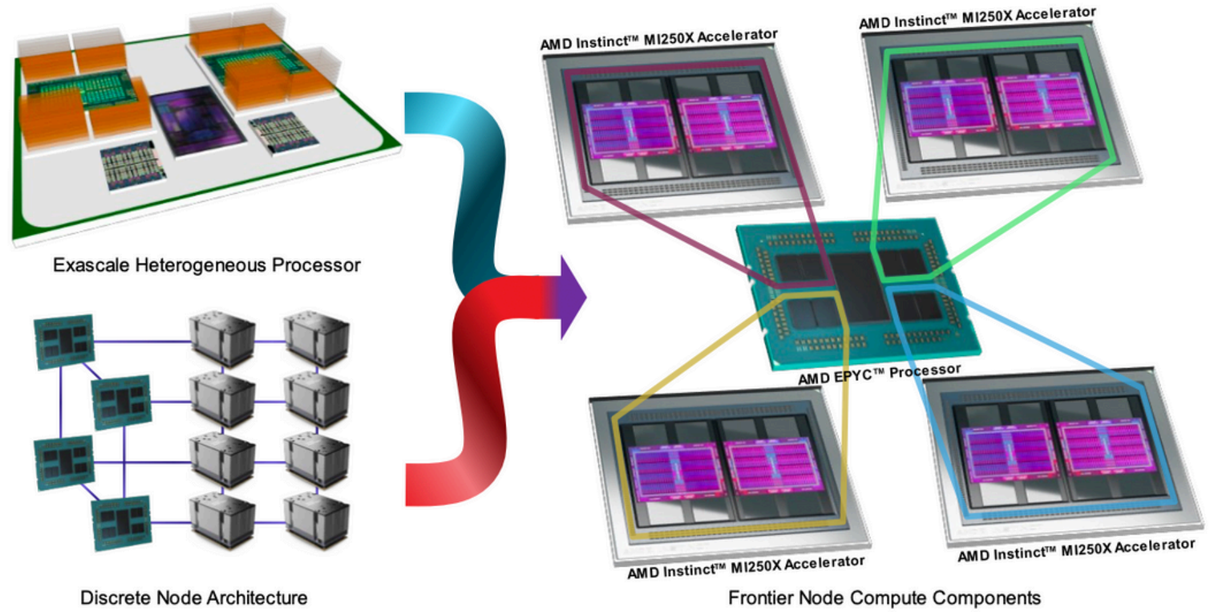


Figure 9. Synthesis of AMD exascale research concepts into the final components of the node architecture employed by the Frontier supercomputer.

A Research Retrospective on AMD's Exascale Computing Journey

Case Study: Frontier

Frontier Technical Details

- The overall machine provides **9,408 compute nodes** housed in 74 cabinets.
- Compute node has a 64-core **EPYC™ 7A53** “Optimized 3rd Gen EPYC™” CPU
- The **64-core** processor shares eight channels of **DDR4 memory with 512GB** of total capacity
- Each node also has **four AMD Instinct™ MI250X accelerators**, each with two GPU compute dies and eight stacks of HBM2E memory supplying **128GB of capacity** (64GB per GPU die).
- Each AMD Instinct™ MI250X accelerator can deliver a peak vector double-precision performance of 47.9 TF and a peak of 3.2 TB/s of bandwidth from the eight DRAM stacks
- The eight CPU chiplets can be partitioned into four non-uniform memory access (NUMA) domains

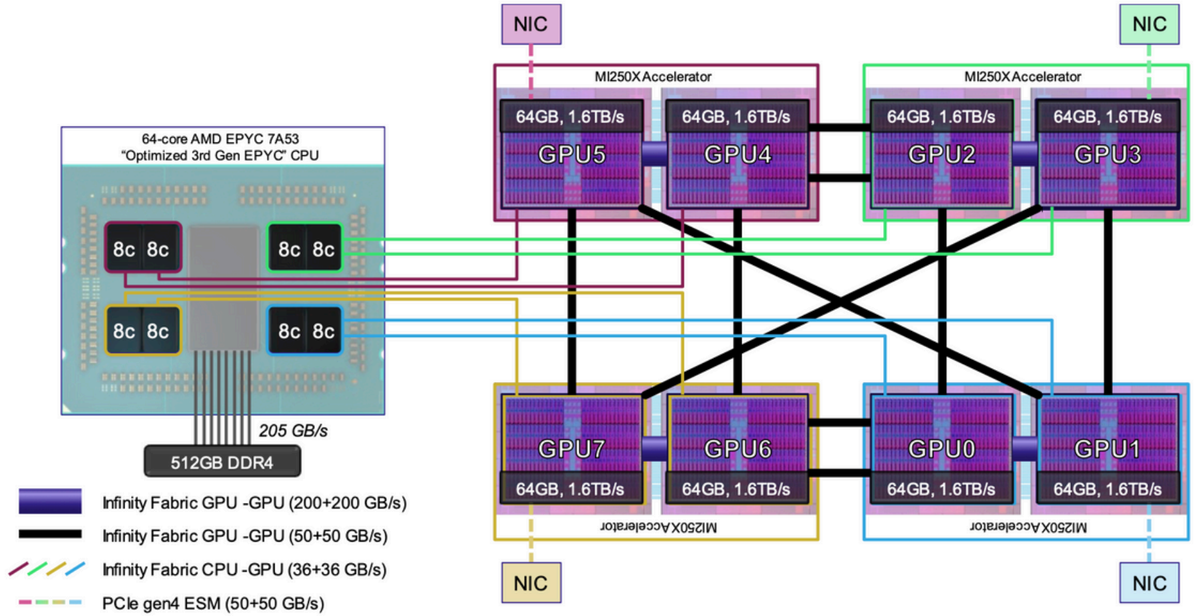


Figure 10. Block diagram of one Frontier Compute Node with peak theoretical memory and interconnect speeds. The “X+X GB/s” notation indicates X GB/s of bandwidth each for send and receive.

A Research Retrospective on AMD's Exascale Computing Journey

Case Study: Frontier

Initial Results and Impacts

- A key initial DOE exascale target was staying within a power budget of 20MW. At a delivered HPL performance of 1.1EF, Frontier consumes 21.1MW [96]. Normalizing this to 1.0EF, the power consumption is 19.2MW per EF.

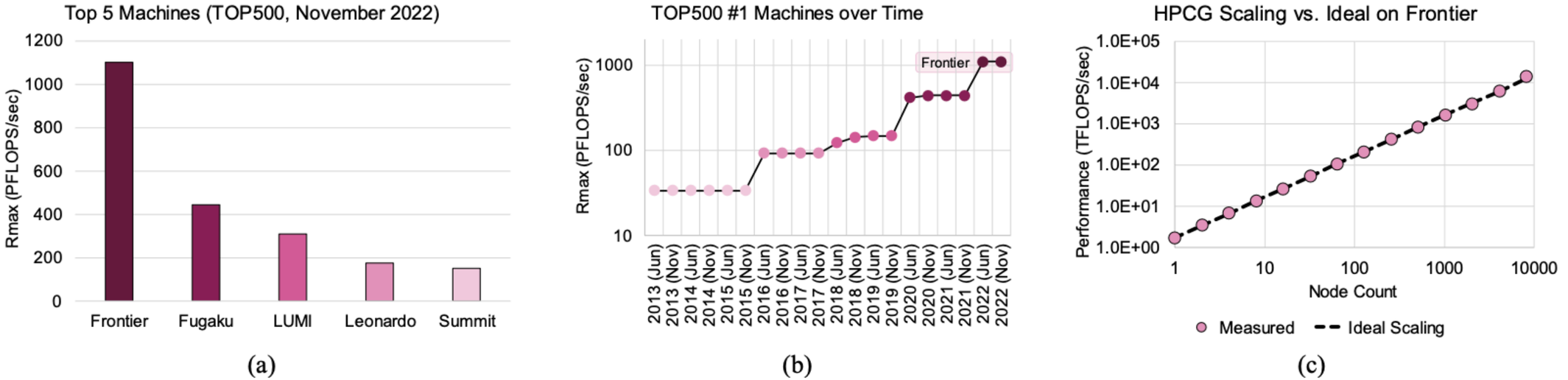


Figure 11. (a) LINPACK performance (R_{MAX}) of the top five supercomputers on the TOP500 list as of November 2022, (b) the R_{MAX} of the #1 supercomputer on the TOP500 list over the past decade, and (c) the scalability of the HPCG benchmark on Frontier (courtesy of HPE and Oak Ridge National Laboratory).

END