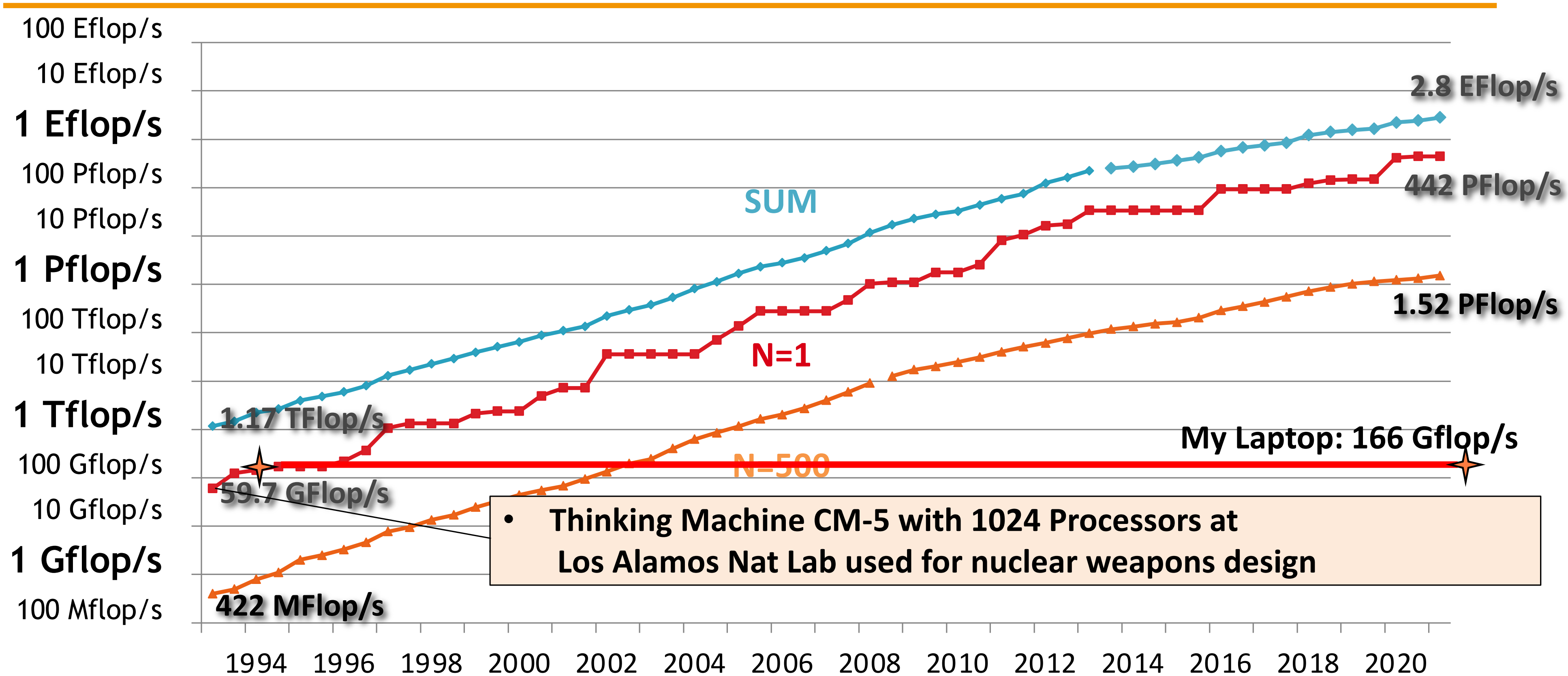
HPC: The Where We Are Today And A Look Into The Future

Jack Dongarra
University of Tennessee

PERFORMANCE DEVELOPMENT OF HPC OVER THE LAST 28 YEARS FROM THE TOP500



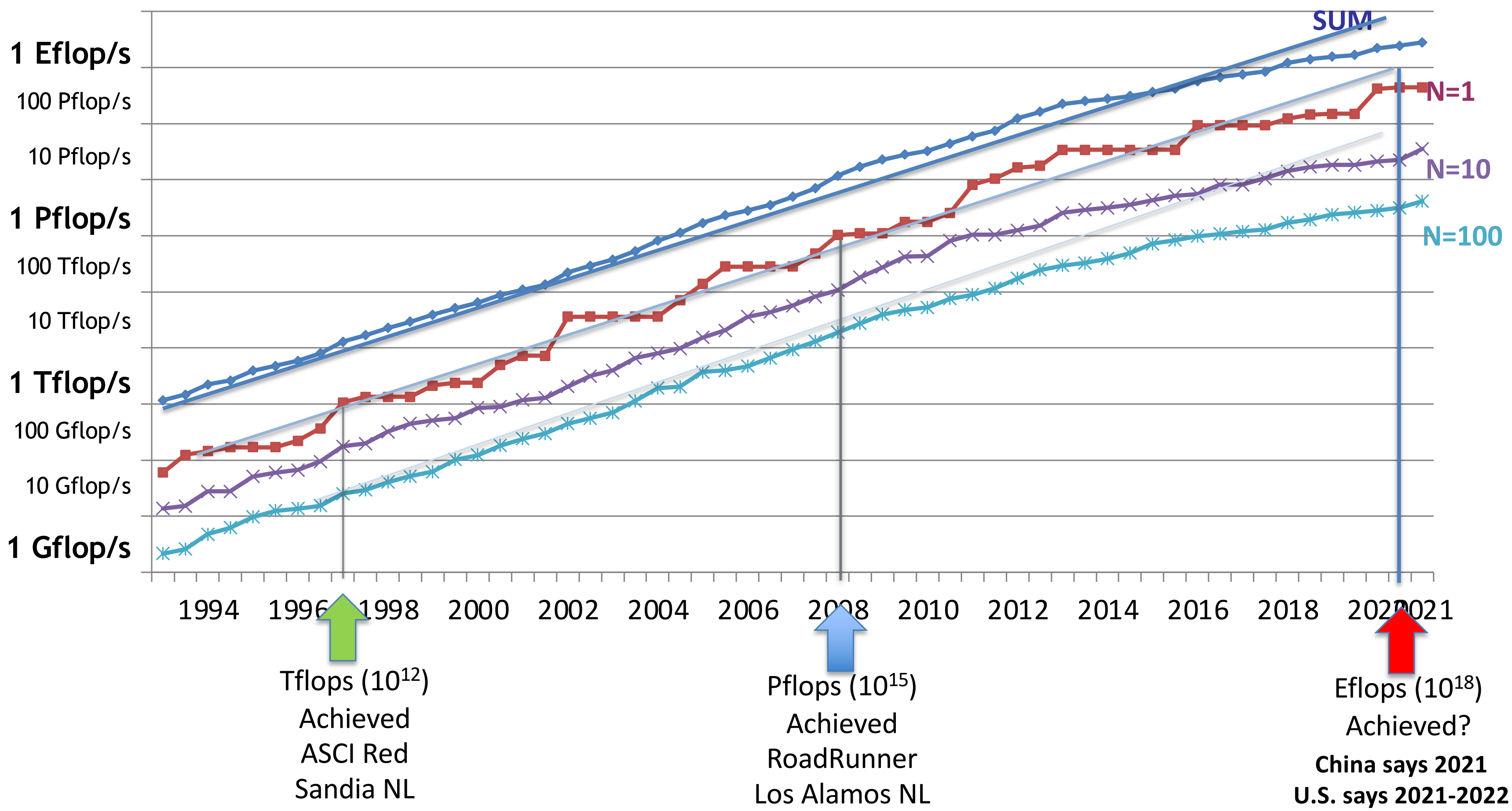
June 2021: The TOP 10 Systems (38% of the Total Performance of Top500)

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	GFlops/ Watt
1	RIKEN Center for Computational Science	Fugaku, ARM A64FX (48C, 2.2 GHz), Tofu D Interconnect	Japan	7,299,072	442.	82	29.9	14.8
2	DOE / OS Oak Ridge Nat Lab	Summit, IBM Power 9 (22C, 3.0 GHz), NVIDIA GV100 (80C), Mellonox EDR	USA	2,397,824	149.	74	10.1	14.7
3	DOE / NNSA L Livermore Nat Lab	Sierra, IBM Power 9 (22C, 3.1 GHz), NVIDIA GV100 (80C), Mellonox EDR	USA	1,572,480	94.6	75	7.44	12.7
4	National Super Computer Center in Wuxi	Sunway TaihuLight, SW26010 (260C) + Custom	China	10,649,000	93.0	74	15.4	6.05
5	DOE / OS NERSC - LBNL	Perlmutter HPE Cray EX235n, AMD EPYC 64C 2.45GHz, NVIDIA A100, Slingshot-10	USA	706,304	64.6	69	2.53	25.5
6	NVIDIA Corporation	Selene NVIDIA DGX A100, AMD EPYC 7742 (64C, 2.25GHz), NVIDIA A100 (108C), Mellanox HDR Infiniband	USA	555,520	63.4	80	2.64	23.9
7	National Super Computer Center in Guangzhou	Tianhe-2A NUDT, Xeon (12C) + MATRIX-2000 (128C) + Custom	China	4,981,760	61.4	61	18.5	3.32
8	JUWELS Booster Module	Bull Sequana XH2000 , AMD EPYC 7402 (24C, 2.8GHz), NVIDIA A100 (108C), Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite	Germany	448,280	44.1	62	1.76	25.0
9	Eni S.p.A in Italy	HPC5, Dell EMC PowerEdge C4140, Xeon (24C, 2.1 GHz) + NVIDIA V100 (80C), Mellonax HDR	Italy	669,760	35.5	69	2.25	15.8
10	Texas Advanced Computing Center / U of Texas	Frontera, Dell C6420, Xeon Platinum, 8280 (28C, 2.7 GHz), Mellanox HDR	USA	448,448	23.5	61		

ISC21 TOP500 Highlights

- Japanese's Fugaku continues as #1 in the TOP500
  - 16% of the TOP500 perform
  - It measured at over 2 Exaflop on the HPL-AI using mixed precision algorithm
- TOP10 has one new system, Perlmutter at LBNL from HPE/Cray, AMD, & NVIDIA
  - TOP10 has 38% of the Top500 performance
- The entry level to the list moved up to the 1.52 Pflop/s mark on the Linpack benchmark.
- The average concurrency level in the TOP500 is 154,246 cores per system, up from 144,932 six months ago.
- China: Top consumer and producer overall.
- Intel processors largest share, 86% followed by AMD, 10%.

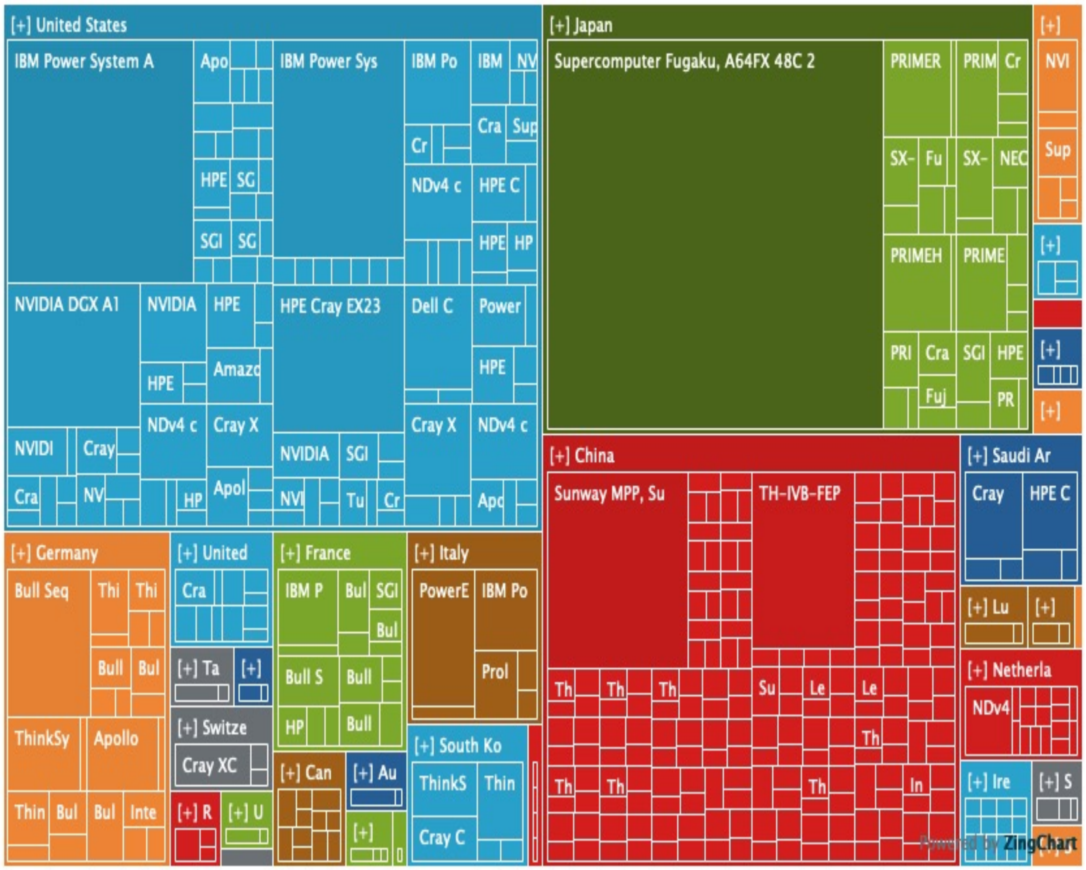
PERFORMANCE DEVELOPMENT

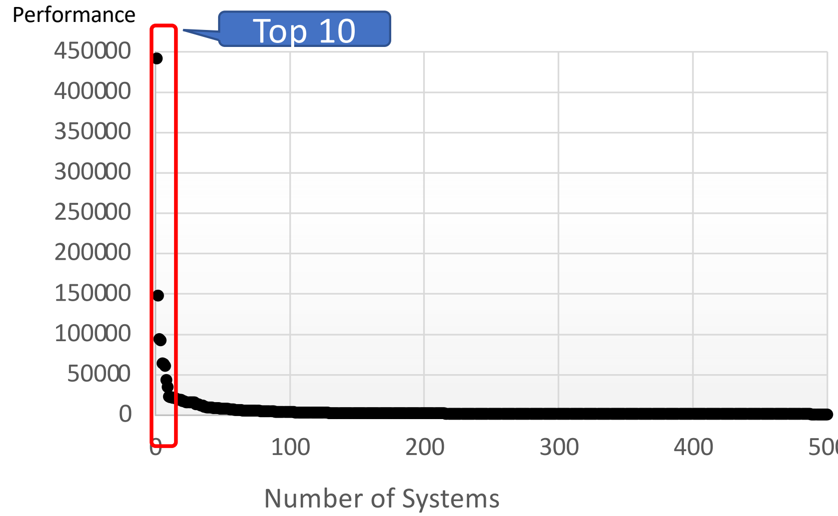


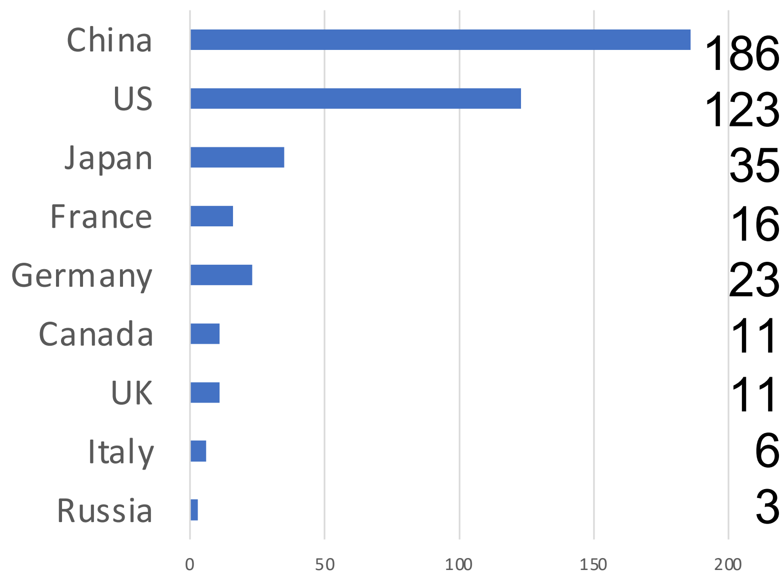
WHAT'S AN EXAFLOP?

- There are 365*24*60*60 seconds in a year
  - About 31 million seconds in a year
  - More precisely 31.536 X 10^6 seconds in one year.
- If you were to do 1 operation per second,
  - You would do 31.536 million operations in a year.
- To get to Exascale …
  - It would take you over 30 billion years to do what an exascale computer does in one second.
  - More precisely 31.71 X 10^9 years.
    - By the way, there has been only 13.8 billion years since the Big Bang.

Countries Share





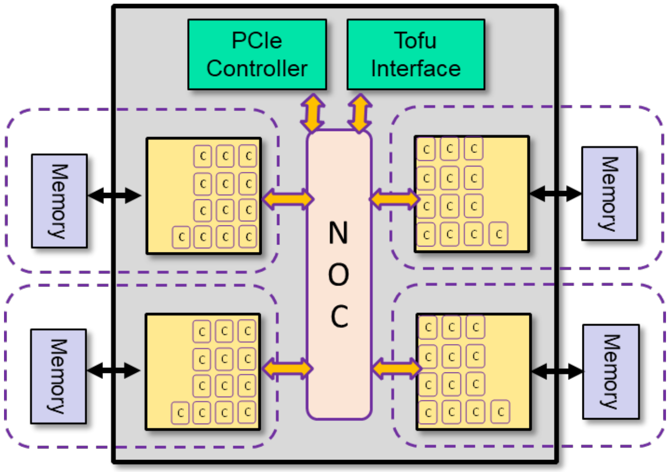


THREE RUSSIAN SYSTEMS ON TOP500

Rank	Name	Computer	Site	Manufacturer	Year	Segment	Total Cores	Accelerator /Co-Processor Cores	LINPACK Rmax [TFlop/s]	Rpeak [TFlop/s]	Accelerator /Co-Processor	Processor Generation
62	Christofari	NVIDIA DGX-2, Xeon Platinum 8168 24C 2.7GHz, Mellanox InfiniBand EDR, NVIDIA V100	SberCloud	Nvidia DGX-2	2019	Industry	99600	96000	6669	8790	NVIDIA Tesla V100	Xeon Platinum
200	Lomonosov 2	T-Platform A-Class Cluster, Xeon E5-2697v3 14C 2.6GHz,Intel Gold 6126, Infiniband FDR, Nvidia K40m/P-100	Moscow State University - Research Computing Center	T-Platforms A-Class Cluster	2014	Academic	64384	40960	2478	4945	NVIDIA Tesla K40m	Intel Xeon E5 (Haswell)
242	MTS GROM	NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100 40GB, Infiniband	#CloudMTS	Nvidia	2021	Industry	19840	17280	2258	3112	NVIDIA A100	AMD EPYC 7742

1 Fugaku's Fujitsu A64fx Processor is…

- A Many-Core ARM CPU…
  - 48 compute cores + 2 or 4 assistant (OS) cores
  - New core design
  - Near Xeon-Class Integer performance core
  - ARM V8 --- 64bit ARM ecosystem
  - Interconnect Tofu-D
  - 3.4 TFLOP/s Peak 64-bit performance

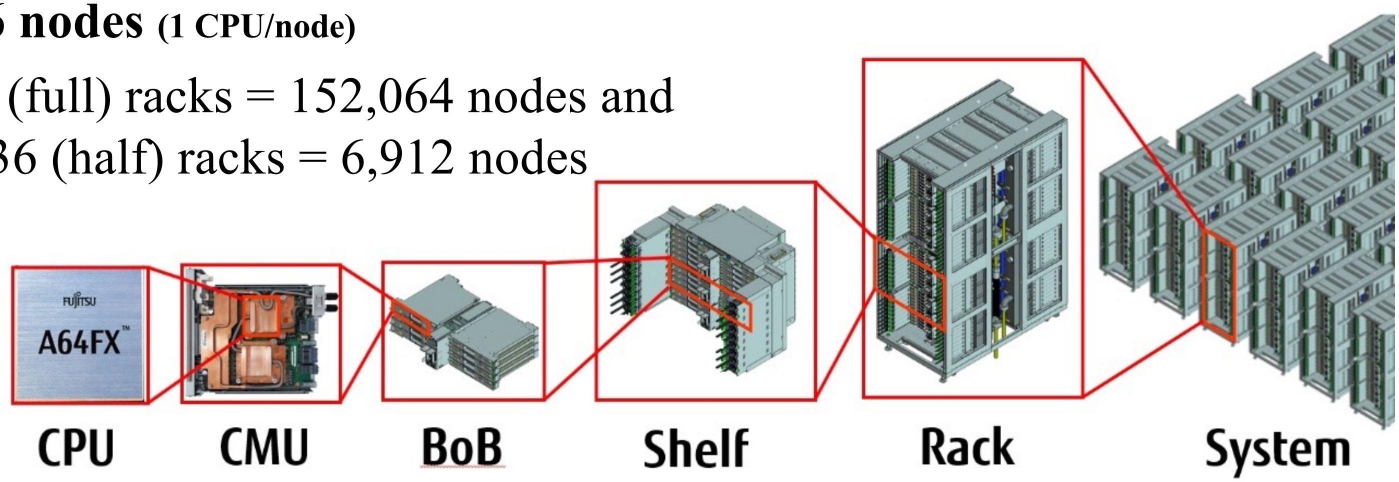


- …but also an accelerated GPU-like processor
  - SVE 512 bit x 2 vector extensions (ARM & Fujitsu)
    - Integer (1, 2, 4, 8 bytes) + Float (16, 32, 64 bytes)
  - Cache + memory localization (sector cache)
  - HBM2 on package memory – Massive Mem BW (Bytes/DPF ~0.4)
    - Streaming memory access, strided access, scatter/gather etc.
  - Intra-chip barrier synch. and other memory enhancing features

http://bit.ly/fugaku-report

Fugaku Total System Config & Performance

- **Total # Nodes: 158,976 nodes (1 CPU/node)**
  - 384 nodes/rack x 396 (full) racks = 152,064 nodes and 192 nodes/rack x 36 (half) racks = 6,912 nodes



Footprint: 1,920 m²

- **Theoretical Peak Compute Performances**
  - Normal Mode (CPU Frequency 2GHz)
    - **64 bit** Double Precision FP: **488 Petaflops**
    - **32 bit** Single Precision FP: **977 Petaflops**
    - **16 bit** Half Precision FP (AI training): **1.95 Exaflops**
    - **8 bit Integer** (AI Inference): **3.90 Exaops**
- **Theoretical Peak Memory BW: 163 Petabytes/s**

Fugaku represents 16% of all the other Top500 systems.

http://bit.ly/fugaku-report

Current #2 System Overview

System Performance

- Peak performance of 200 Pflop/s for modeling & simulation
- Peak performance of 3.3 Eflop/s for 16 bit floating point used in for data analytics, ML, and artificial intelligence

Each node has

- 2 IBM POWER9 processors
  - Each w/22 cores
  - 2.3% performance of system
- 6 NVIDIA Tesla V100 GPUs
  - Each w/80 SMs
  - 97.7% performance of system
- 608 GB of fast memory
- 1.6 TB of NVMe memory

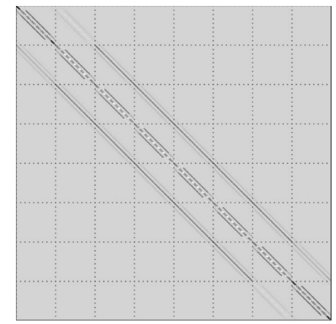
The system includes

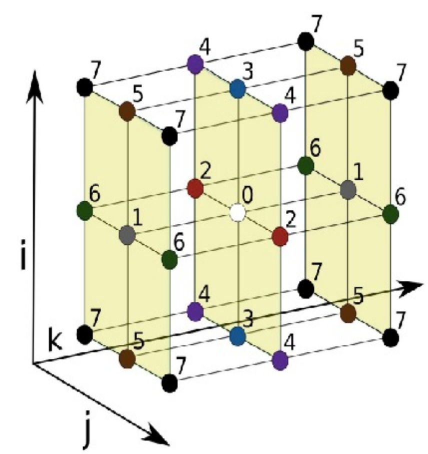
- 4608 nodes
  - 27,648 GPUs
  - Street value $10K each
- Dual-rail Mellanox EDR InfiniBand network
- 250 PB IBM Spectrum Scale
  file system transferring data at 2.5 TB/s

hpcg-benchmark.org

HPCG Results; The Other Benchmark

- High Performance Conjugate Gradients (HPCG).
- Solves *Ax=b, A* large, sparse, *b* known, *x* computed.
- An optimized implementation of PCG contains essential computational and communication patterns that are prevalent in a variety of methods for discretization and numerical solution of PDEs



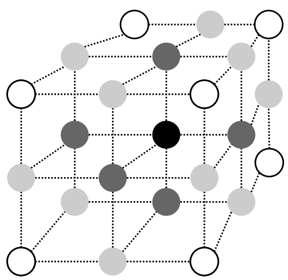


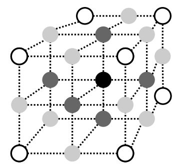
27-point stencil operator

- Patterns:
  - Dense and sparse computations.
  - Dense and sparse collectives.
  - Multi-scale execution of kernels via MG (truncated) V cycle.
  - Data-driven parallelism (unstructured sparse triangular solves).
- Strong verification (via spectral properties of PCG).

HPCG Details

3D Laplacian discretization

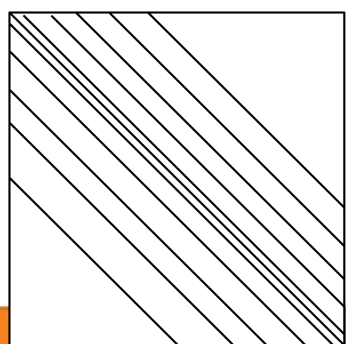




Multigrid

$L[u] \equiv \nabla^2 u = f$

Sparse matrix based on 27-point stencil



$Au = f$

Preconditioned Conjugate Gradient solver

$p_0 \leftarrow x_0,\ r_0 \leftarrow b - Ap_0$

for $i = 1, 2,$ to max_iterations **do**

$\quad z_i \leftarrow M^{-1} r_{i-1}$

$\quad$ **if** $i = 1$ **then** — Multigrid and Gauss-Seidel

$\quad\quad p_i \leftarrow z_i$

$\quad\quad \alpha_i \leftarrow \text{dot\_prod}(r_{i-1}, z_i)$

$\quad$ **else**

$\quad\quad \alpha_i \leftarrow \text{dot\_prod}(r_{i-1}, z_i)$

$\quad\quad \beta_i \leftarrow \alpha_i / \alpha_{i-1}$

$\quad\quad p_i \leftarrow \beta_i p_{i-1} + z_i$

$\quad$ **end if**

$\quad \alpha_i \leftarrow \text{dot\_prod}(r_{i-1}, z_i)/\text{dot\_prod}(p_i, Ap_i)$

$\quad x_{i+1} \leftarrow x_i + \alpha_i p_i$

$\quad r_i \leftarrow r_{i-1} - \alpha_i Ap_i$

$\quad$ **if** $\|r_i\|_2 <$ tolerance **then**

$\quad\quad$ STOP

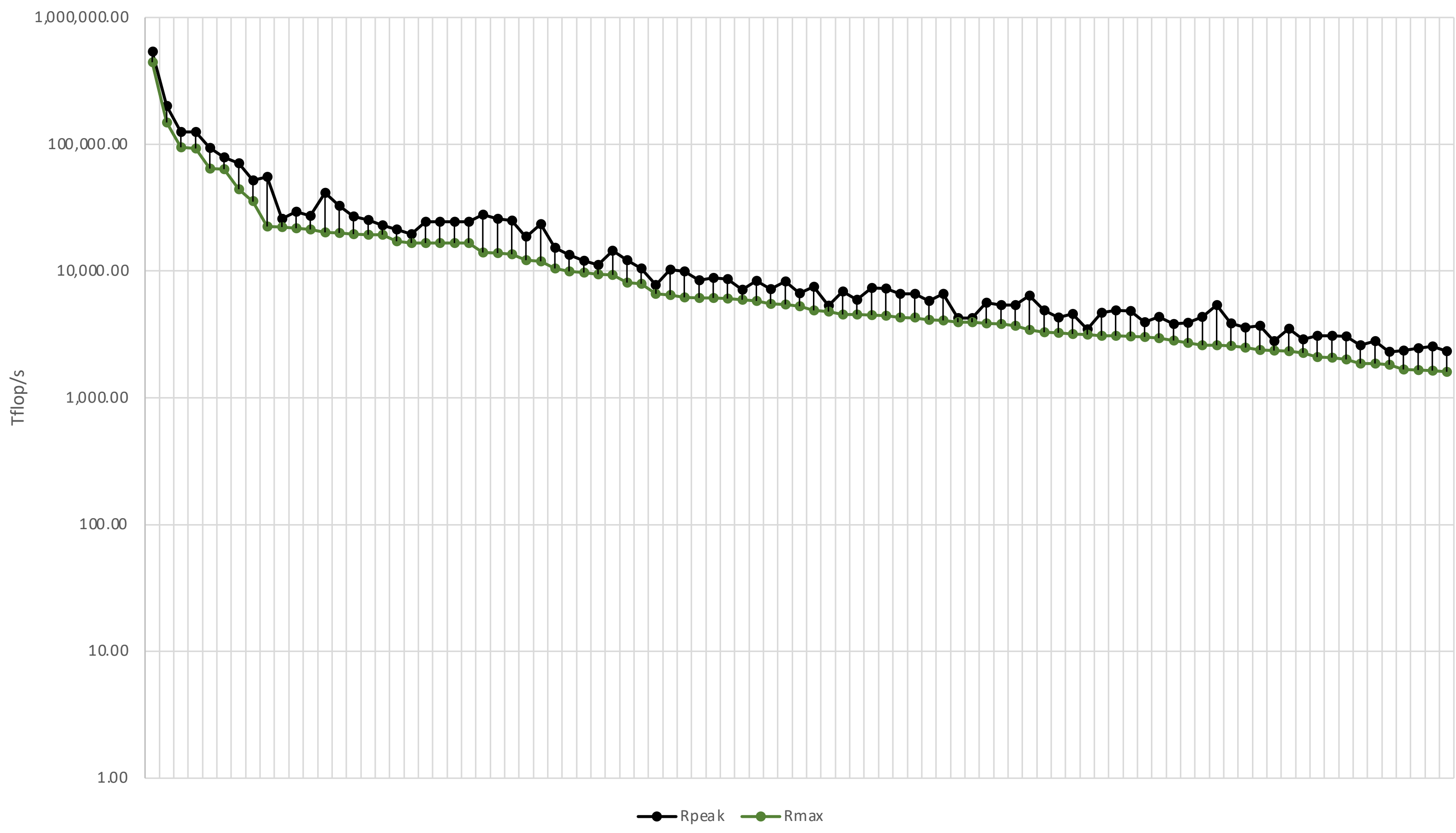
$\quad$ **end if**

end for

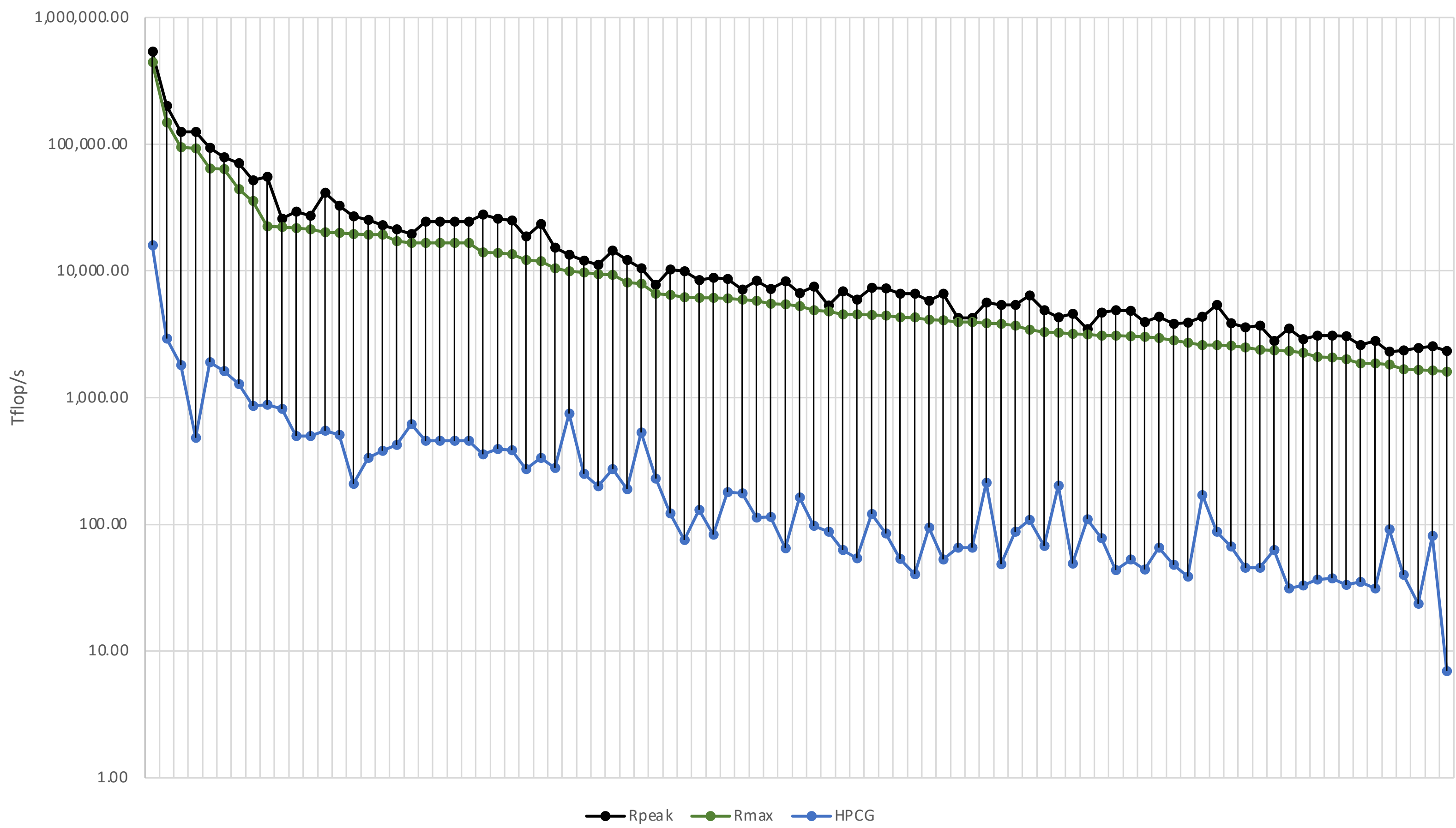
HPCG Top10, June 2021

Rank	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	Fraction of Peak
1	RIKEN Center for Computational Science Japan	**Fugaku**, Fujitsu A64FX 48C 2.2GHz, Tofu D, Fujitsu	7,630,848	442.0	1	16.0	3.0%
2	DOE/SC/ORNL USA	**Summit**, AC922, IBM POWER9 22C 3.7GHz, Dual-rail Mellanox FDR, NVIDIA Volta V100, IBM	2,414,592	148.6	2	2.93	1.5%
3	**DOE/SC/LBNL USA**	**Perlmutter**, HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10	761,856	64.6	5	1.91	2.0%
4	DOE/NNSA/LLNL USA	**Sierra**, S922LC, IBM POWER9 20C 3.1 GHz, Mellanox EDR, NVIDIA Volta V100, IBM	1,572,480	94.6	3	1.80	1.4%
5	NVIDIA USA	**Selene**, DGX SuperPOD, AMD EPYC 7742 64C 2.25 GHz, Mellanox HDR, NVIDIA Ampere A100	555,520	63.5	6	1.62	2.0%
6	Forschungszentrum Juelich (FZJ) Germany	**JUWELS Booster Module**, Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, Mellanox HDR InfiniBand, NVIDIA Ampere A100, Atos	449,280	44.1	8	1.28	1.8%
7	Saudi Aramco Saudi Arabia	**Dammam-7**, Cray CS-Storm, Xeon Gold 6248 20C 2.5GHz, InfiniBand HDR 100, NVIDIA Volta V100, HPE	672,520	22.4	11	0.88	1.6%
8	Eni S.p.A. Italy	**HPC5**, PowerEdge, C4140, Xeon Gold 6252 24C 2.1 GHz, Mellanox HDR, NVIDIA Volta V100, Dell	669,760	35.5	9	0.86	1.7%
9	Information Technology Center, The University of Tokyo, Japan	**Wisteria/BDEC-01 (Odyssey)**, PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu D	368,640	22.1	13	0.82	3.2%
10	Japan Agency for Marine-Earth Science and Technology	**Earth Simulator -SX-Aurora TSUBASA**, A412-8, Vector Engine Type20B 8C 1.6GHz, Infiniband HDR200	43,776	0.01	41	0.75	5.6%

Comparison between Peak and HPL for June 2021



Comparison between Peak, HPL, and HPCG for June 2021



Modern Hardware: Lower Precision for Deep Learning

- Hardware (company)
  - GPU Tensor Cores (NVIDIA)
  - TPU MXU (Google)
  - Zion (Facebook)
  - DaVinci (Huawei)
  - Dot-product engine (HPE)
  - Eyeriss (Amazon)
  - Wafer Scale Engine (Cerebras)
  - Nervana (Intel)
  - Deep Learning Boost (Intel AI)
  - Graph Core
  - ...

60+

- Lower-precision benchmarks
  - Baidu
  - Dawn
  - mlperf
  - Deep500
  - ...
  - HPL-AI

WHY MIXED PRECISION? (Less is Faster)

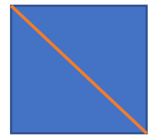
- There are many reasons to consider mixed precision in our algorithms…
  - Less Communication
    - Reduce memory traffic
    - Reduce network traffic
  - Reduce memory footprint
  - More Flop per second
    - Reduced energy consumption
    - Reduced time to compute
  - Accelerated hardware in current architecture.
  - Suitable numerical properties for some algorithms & problems.

J. Langou, J. Langou, P. Luszczek, J. Kurzak, A. Buttari, and J. J. Dongarra. Exploiting the performance of 32 bit floating point arithmetic in obtaining 64 bit accuracy. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, 2006.

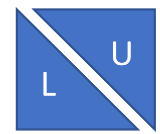
Mixed Precision: Hardware Motivation

IBM Cell Broadband Engine	Apple ARM Cortex-A9	NVIDIA Kepler K10, K20, K40, K80	NVIDIA Volta/Turing	NVIDIA Volta/Turing
14x	7x	3x	2x	16x
32 bits / 64 bits	32 bits / 64 bits	32 bits / 64 bits	32 bits / 64 bits	16 bits / 64 bits

HPL-AI Benchmark Utilizing 16-bit Arithmetic



1. Generate random linear system Ax=b
2. Represent the matrix A in low precision (16-bit floating point)



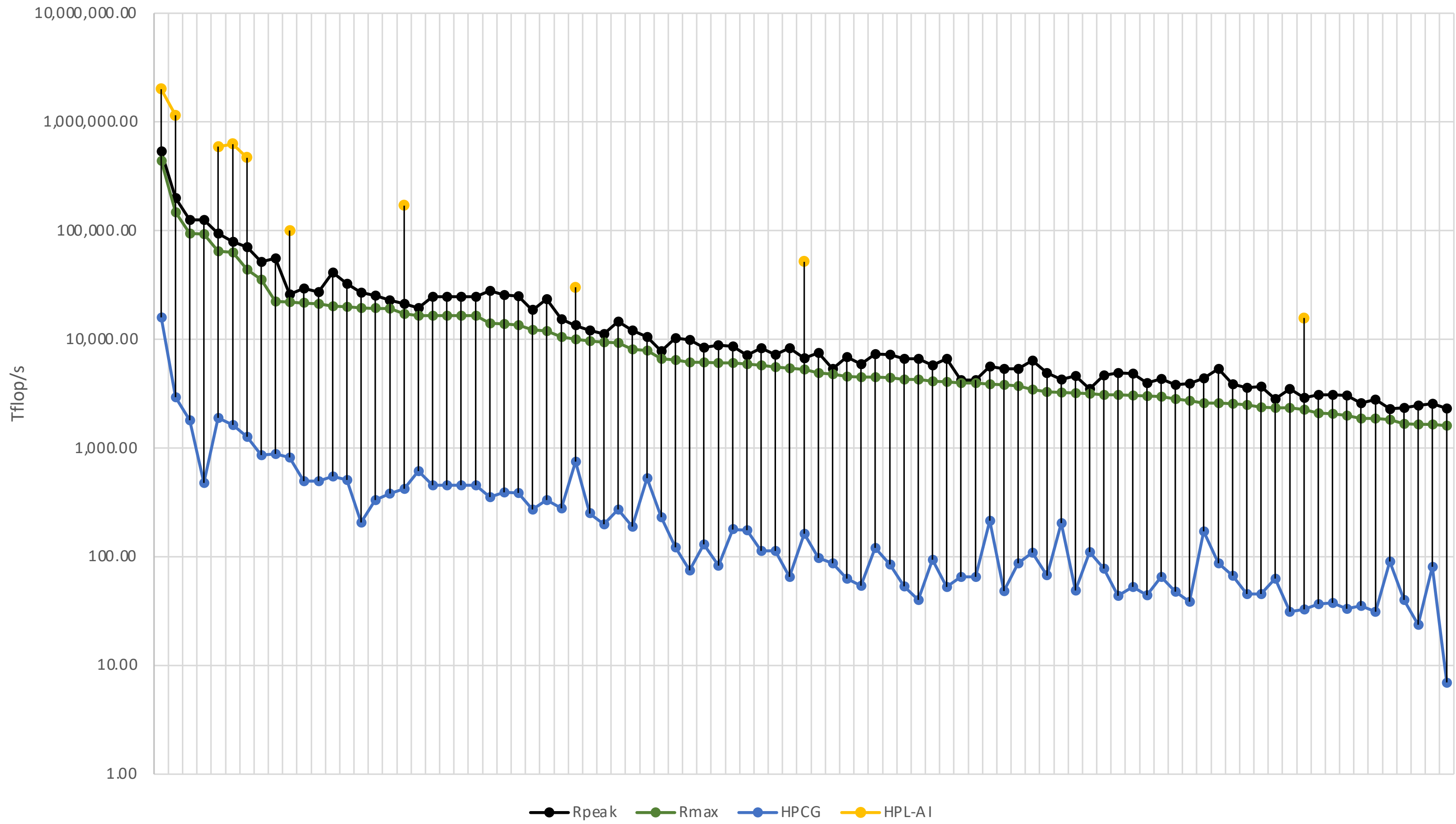
3. Factor A in lower precision into LU by Gaussian elimination
4. Compute approximate solution with LU factors in low precision
5. Perform up to 50 iterations of refinement, e.g., GMRES to get accuracy up to 64-bit floating point
6. Use LU factors for preconditioning
7. Validate the answer is correct: scaled residual small $\frac{||Ax-b||}{||A||||x||+||b||} \times \frac{1}{n\epsilon} \leq O(10)$
8. Compute performance rate as $\frac{2}{3} \times \frac{n^3}{\text{time}}$

Iterative refinement for dense systems, *Ax = b*, can work this way.		
L U = lu(A)	**lower precision**	**$O(n^3)$**
x = U\(L\b)	**lower precision**	**$O(n^2)$**
GMRes preconditioned by the LU to solve Ax=b	**FP64 precision**	**$O(n^2)$**

HPL-AI Top 10 for June 2021

Rank	Site	Computer	Cores	HPL Rmax (Eflop/s)	TOP500 Rank	HPL-AI (Eflop/s)	Speedup
1	RIKEN Center for Computational Science, Japan	**Fugaku**, Fujitsu A64FX, Tofu D	7,630,848	0.442	1	2.0	4.5
2	DOE/SC/ORNL USA	**Summit**, AC922 IBM POWER9, IB Dual-rail FDR, NVIDIA V100	2,414,592	0.149	2	1.15	7.7
3	NVIDIA USA	**Selene**, DGX SuperPOD, AMD EPYC 7742 64C 2.25 GHz, Mellanox HDR, NVIDIA A100	555,520	0.063	6	0.63	9.9
4	DOE/SC/LBNL/NERSC USA	**Perlmutter**, HPE Cray EX235n, AMD EPYC 7763 64C 2.45 GHz, Slingshot-10, NVIDIA A100	761,856	0.065	5	0.59	9.1
5	Forschungszentrum Juelich (FZJ) Germany	**JUWELS Booster Module**, Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, Mellanox HDR InfiniBand, NVIDIA A100, Atos	449,280	0.044	8	0.47	10
6	University of Florida USA	**HiPerGator**, NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Infiniband HDR	138,880	0.017	23	0.17	9.9
7	Information Technology Center, The University of Tokyo, Japan	**Wisteria/BDEC-01 (Odyssey)**, PRIMEHPC FX1000, A64FX 48C 2.2GHz, Tofu D, Fujitsu	368,640	0.022	13	0.10	4.5
8	National Supercomputer Centre (NSC), Sweden	**Berzelius**, NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, A100, Infiniband HDR, Atos	59,520	0.005	84	0.05	9.9
9	Information Technology Center, Nagoya University, Japan	**Flow Type II subsystem**, PRIMERGY CX2570 M5, Xeon Gold 6230 20C 2.1GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR	79,560	0.0049	87	0.03	4.3
10	#CloudMTS Russia	**MTS GROM**, NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, A100 40GB, Infiniband	19,840	0.0023	245	0.015	7

Comparison between HPL-AI, Peak, HPL, and HPCG for June 2021



2026 and 2030 Planning Targets

2026 – 10 Eflop/s (64 bit floating point) and
>**100 Eflop/s** (AI 16 bit floating point)

2030 – 50 Efflop/s (64 bit floating point) and
>1000 Eflop/s (AI 16 bit floating point)

A few questions:

- How achievable are these targets given the roadmaps and vendor plans?
- Will AI accelerators (distinct from GPUs) make sense to integrate into future nodes or as sub-clusters?
- When will quantum computing accelerators intersect mainstream supercomputing?

Zettascale System Metrics

Table 1 Zettascale metrics

Metric	Value
Peak performance	1 Zflops
Power consumption	100 MW
Power efficiency	10 Tflops/W
Peak performance per node	10 Pflops/node
Bandwidth between nodes	1.6 Tb/s
I/O bandwidth	10–100 PB/s
Storage capacity	1 ZB
Floor space	1000 m^2

Front Inform Technol Electron Eng

Chinese proposes Zettascale by 2035

- 600x ORNL Frontier
- 3.4x ORNL Frontier
- 200x ORNL Frontier
- 66x ORNL Frontier
- 16x ORNL Frontier
- 1000x ORNL Frontier
- 1000x ORNL Frontier
- 2x ORNL Frontier

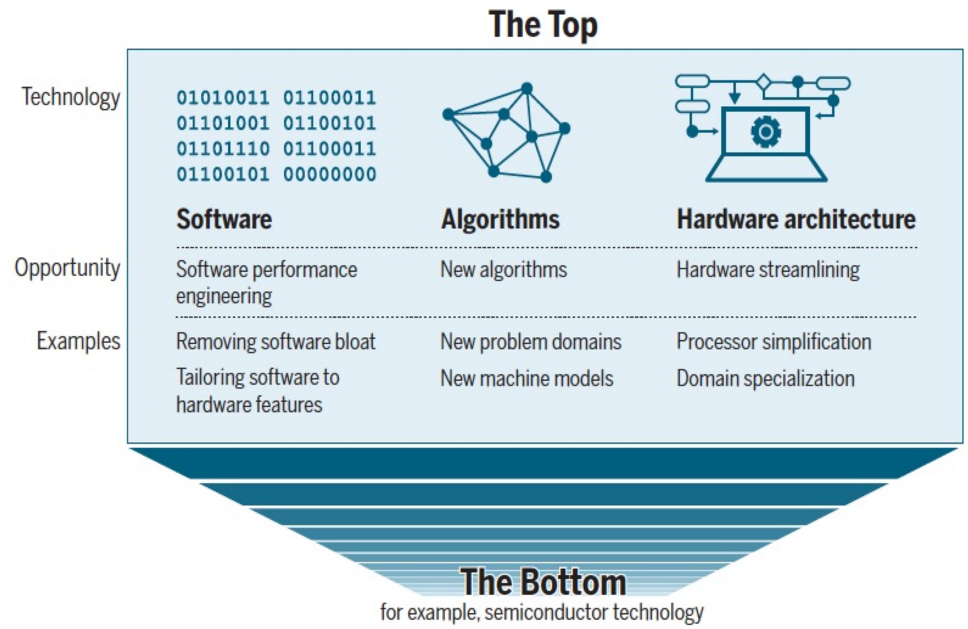
Moving from exascale to zettascale computing: challenges and techniques. Frontiers of Information Technology & Electronic Engineering, 19(10):1236-1244. https://doi.org/10.1631/FITEE.1800494 Front Inform Technol Electron Eng

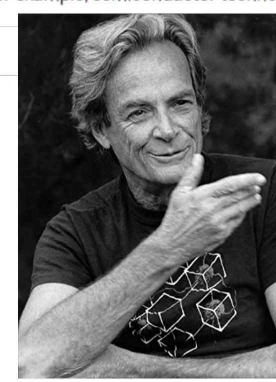
The Take Away

- HPC Hardware is Constantly Changing
  - Scalar
  - Vector
  - Distributed
  - Accelerated
  - Mixed precision
- Three computer revolutions
  - High performance computing
  - Deep learning
  - Edge & AI
- Algorithm / Software advances follows hardware.
  - And there is “plenty of room at the top”

“There’s plenty of room at the Top: What will drive computer performance after Moore’s law?”

Leiserson *et al.*, *Science* **368**, 1079 (2020)   5 June 2020





Feynman’s 1959 Lecture @ CalTech