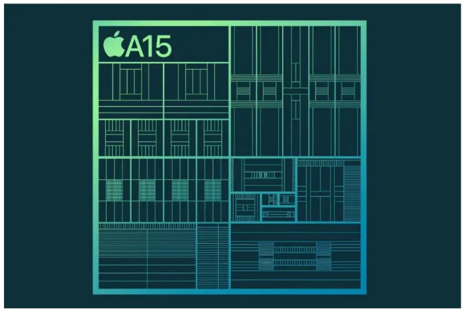
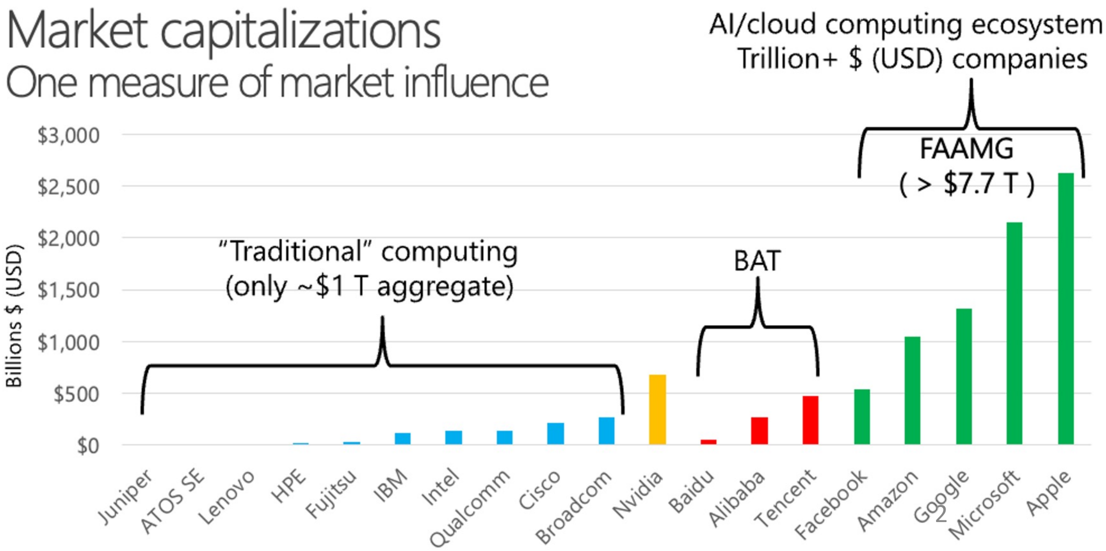
An Overview of High Performance Computing and Future Requirements

Jack Dongarra
University of Tennessee
Oak Ridge National Laboratory
University of Manchester

A Changing World

- Computing pervades all aspects of society
  - Socialization and communication
  - E-commerce and business
  - Research and development
- Apple, Samsung, and Google
  - Dominate the world of smartphones
  - Design their own silicon
- Google, Microsoft, Amazon, Apple
  - Dominate the NASDAQ (market cap > $1T each)
  - Baidu, Alibaba, and TenCent are not far behind
  - Also designing their own silicon

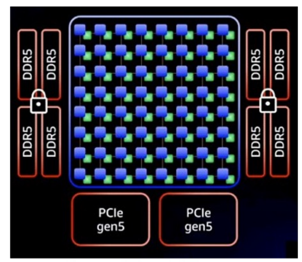




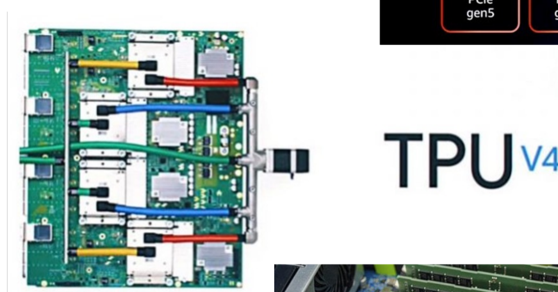
Cloud vendors

- Alibaba
  - CIPU, 128 core ARM based
  - Alibaba’s Elastic Compute Service

- AWS Graviton3
  - 64 ARM Neoverse V1 cores, chiplet design
  - 55 billion transistors, DDR5 memory



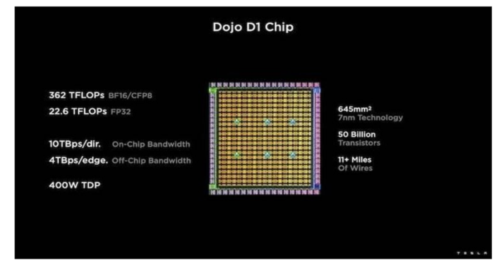
- Google TPU4
  - 2X TPU3 performance
  - 4096 units per “pod”
  - Reconfigurable optical interconnect



- Microsoft Azure
  - Ampere Alta ARM processors
  - Project Catapult/Brainewave

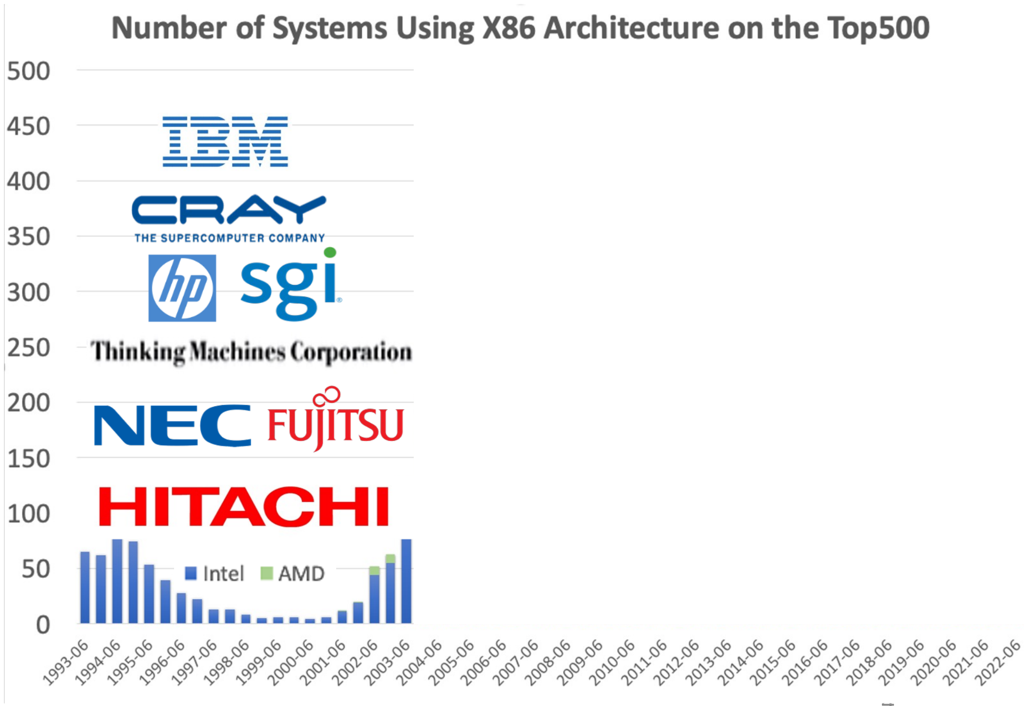
Even car makers

- Tesla



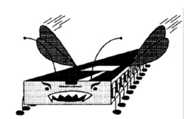
High Performance Computing is a Monoculture – Processors

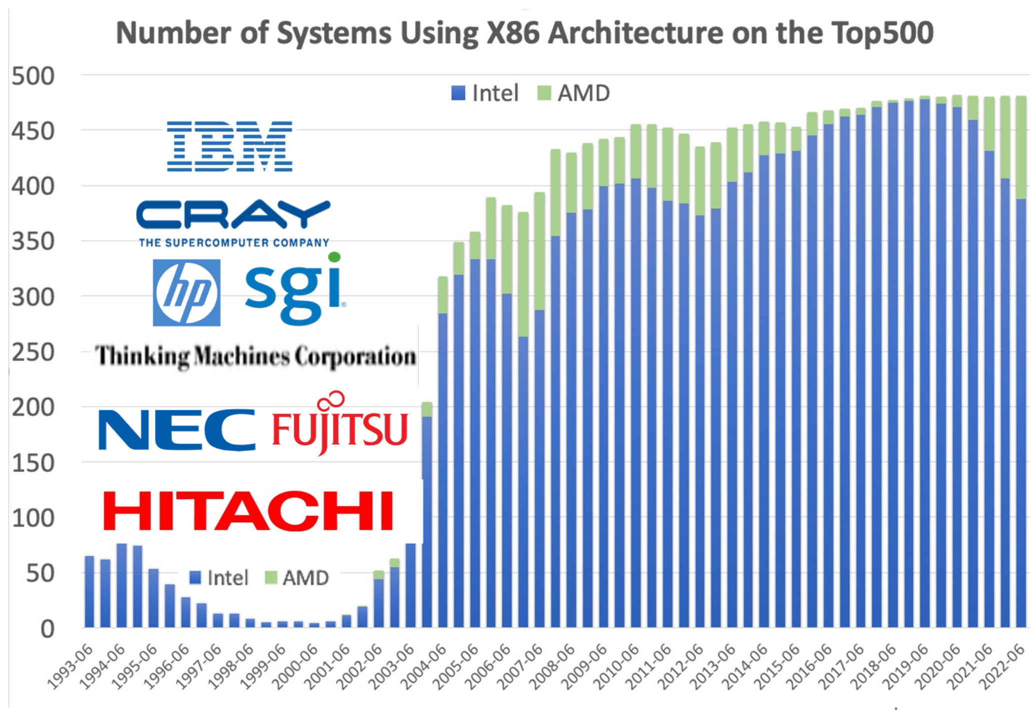
- TOP500 list began in 1993
  - 65 systems used Intel's i860 architecture
  - Remainder had specialized architectures, mainly vector based



High Performance Computing is a Monoculture – Processors

- TOP500 list began in 1993
  - 65 systems used Intel's i860 architecture
  - Remainder had specialized architectures, mainly vector based



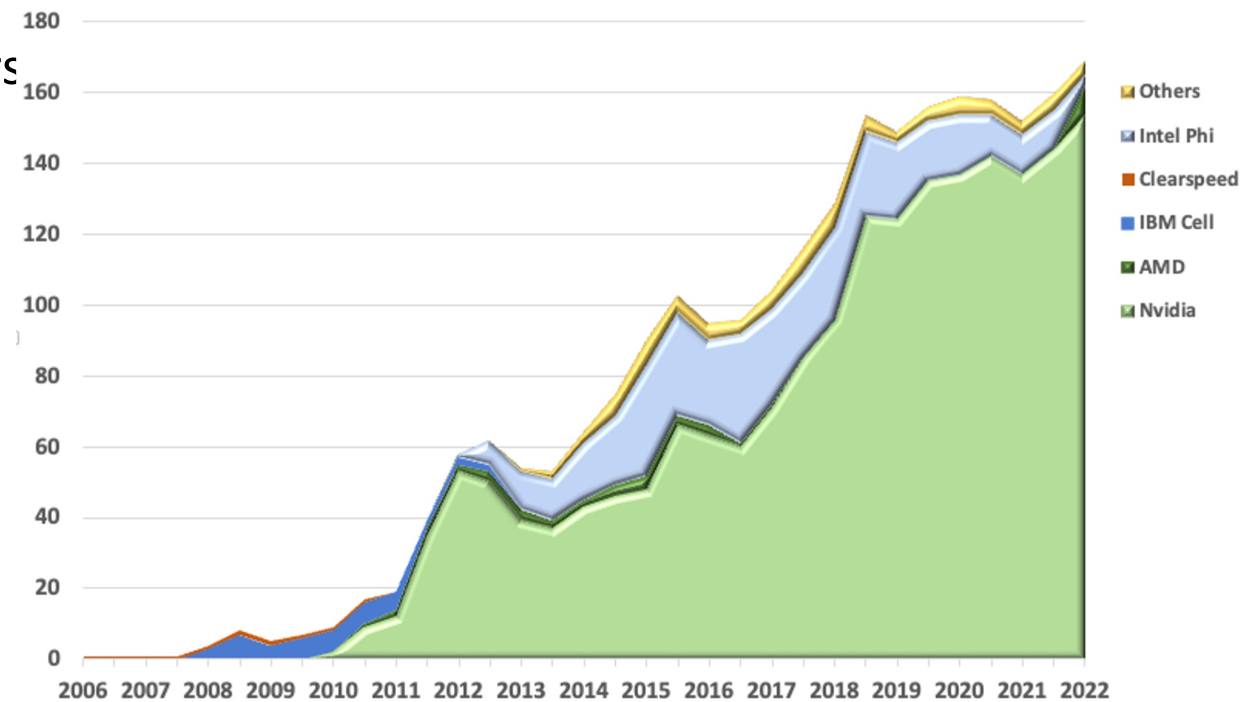


- Today's TOP500 list
  - 78% of systems used Intel processors
  - Another 19% used AMD processors

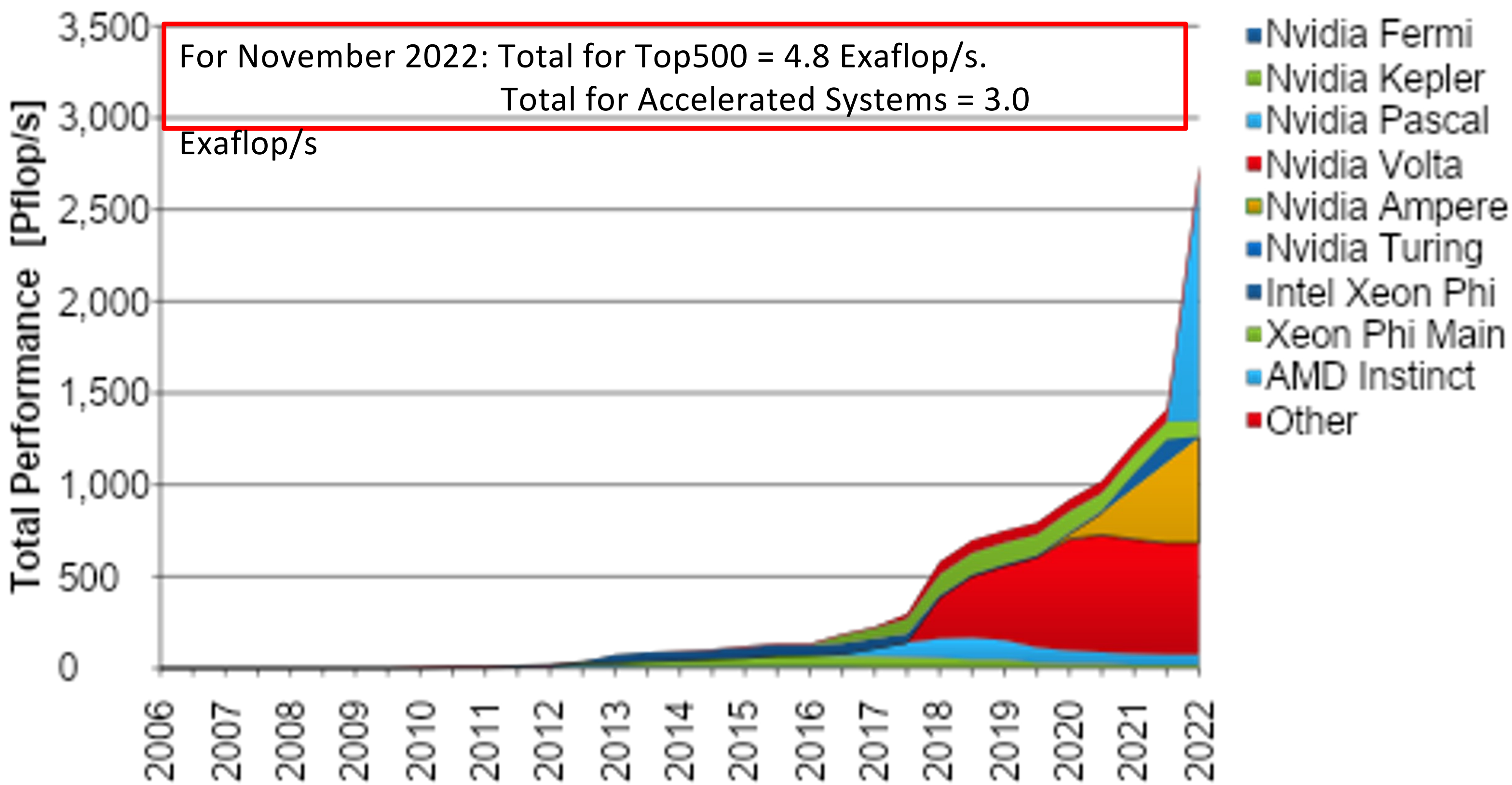
- **97% of the systems use x86-64 architecture**
  - Many use GPU accelerators

HPC Monoculture – Accelerators/Interconnects/OS

- Nvidia dominates accelerators
- Interconnects are mainly Ethernet/InfiniBand
  - 85% of the Top500
  - 426 systems
- Linux is standard everywhere



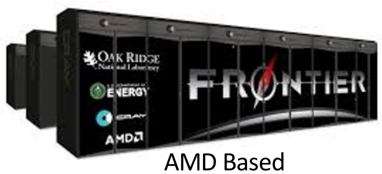
62% of Top500 Performance on Accelerators



DOE: Exascale investing > $4 B in total, over 7 years

What do you get for $4 B?

- 3 computers
  - $600M each
  - $400M to vendors for Design, Path, Fast - Forward



AMD Based
(Up & running)

Intel Based
(Being installed)

AMD Based
(Planned)

21 Applications

Domain*	Base Challenge Problem
Wind Energy	2x2 5 MW turbine array in 3x3x1 km^3 domain
Nuclear Energy	Small Modular Reactor with complete in-vessel coolant loop
Fossil Energy	Burn fossil fuels cleanly with CLRs
Combustion	Reactivity controlled compression ignition
Accelerator Design	TeV-class $10^{2\text{-}3}$ times cheaper & smaller
Magnetic Fusion	Coupled gyrokinetics for ITER in H-mode
Nuclear Physics: QCD	Use correct light quark masses for first principles light nuclei properties
Chemistry: GAMESS	Heterogeneous catalysis: MSN reactions
Chemistry: NWChemEx	Catalytic conversion of biomass
Extreme Materials	Microstructure evolution in nuclear matls
Additive Manufacturing	Born-qualified 3D printed metal alloys

Domain*	Challenge Problem
Quantum Materials	Predict & control matls @ quantum level
Astrophysics	Supernovae explosions, neutron star mergers
Cosmology	Extract "dark sector" physics from upcoming cosmological surveys
Earthquakes	Regional hazard and risk assessment
Geoscience	Well-scale fracture propagation in wellbore cement due to attack of CO_2-saturated fluid
Earth System	Assess regional impacts of climate change on the water cycle @ 5 SYPD
Power Grid	Large-scale planning under uncertainty; underfrequency response
Cancer Research	Scalable machine learning for predictive preclinical models and targeted therapy
Metagenomics	Discover and characterize microbial communities through genomic and proteomic analysis
FEL Light Source	Protein and molecular structure determination using streaming light source data

- A bunch of software (84 projects)

PMR Core (17)	Compilers and Support (7)	Tools and Technology (11)	xSDK (16)	Visualization Analysis and Reduction (9)	Data mgmt, I/O Services, Checkpoint restart (12)	Ecosystem/E4S at-large (12)
QUO	openarc	TAU	hypre	ParaView	SCR	mpiFileUtils
Papyrus	Kitsune	HPCToolkit	FleSCI	Catalyst	FAODEL	TriBITS
SICM	LLVM	Dyninst Binary Tools	MFEM	VTK-m	ROMIO	MarFS
Legion	CHiLL autotuning comp	Gotcha	Kokkoskernels	SZ	Mercury (Mochi suite)	GUFI
Kokkos (support)	LLVM openMP comp	Caliper	Trilinos	zfp	HDF5	Intel GEOPM
RAJA	OpenMP V & V	PAPI	SUNDIALS	VisIt	Parallel netCDF	BEE
CHAI	Flang/LLVM Fortran comp	Program Database Toolkit	PETSc/TAO	ASCENT	ADIOS	FSEFI
PaRSEC*		Search (random forests)	libEnsemble	Cinema	Darshan	Kitten Lightweight Kernel
DARMA		Siboka	STRUMPACK	ROVER	UnifyCR	COOLR
GASNet-EX		C2C	SuperLU		VeloC	NRM
Qthreads		Sonar	ForTrilinos		IOSS	ArgoContainers
BOLT			SLATE		HXHIM	Spack
UPC++			MAGMA			
MPICH			DTK			
Open MPI			Tasmanian			
Umpire			Ginkgo			
AML						

Legend
- PMR
- Tools
- Math Libraries
- Data and Vis
- Ecosystems and delivery

1000 people working on ECP, and the project will end in 8 months. There is no follow-on project at this scale!!

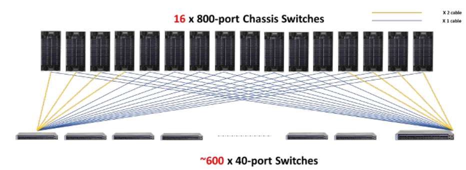
Today's HPC Environment for Scientific Computing

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

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

- 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	$10^{\pm 4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

The Fastest Supercomputers are at an Exaflop.

What's an Exaflop?

- 1 flop = Addition or Multiplication of 64-bit floating point numbers
- Exaflop is a billion-billion (10^{18}) floating point operations per second
- 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.

An Accidental Benchmarker

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

Top 23 List from 1977
Performance of solving *Ax=b* using LINPACK software

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

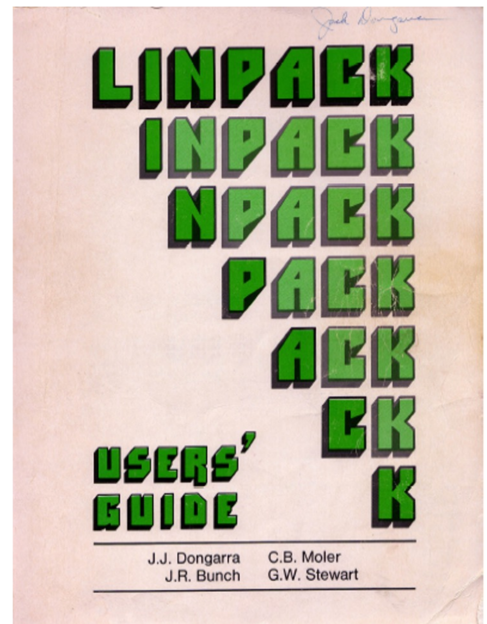
Facility	TIME N=100 secs.	UNIT micro-secs.	Computer	Type	Compiler
NCAR	.049	0.14	CRAY-1	S	CFT, Assembly BLAS
LASL	.148	0.43	CDC 7600	S	FTN, Assembly BLAS
NCAR	.192	0.56	CRAY-1	S	CFT
LASL	.210	0.61	CDC 7600	S	FTN
Argonne	.297	0.86	IBM 370/195	D	H
NCAR	.359	1.05	CDC 7600	S	Local
Argonne	.388	1.33	IBM 3033	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

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;

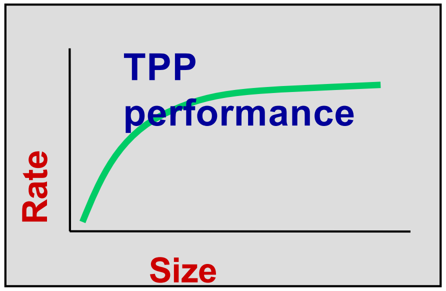
- Cray 1 to DEC PDP-10



Top500 Since 1993

- Hans Meuer and Erich Strohmaier had a list of fastest computers ranked by peak performance.
- I had a list of benchmark results and we put the two lists together.
- Listing of the 500 most powerful computers in the World.
- Yardstick: Performance for

  Ax=b, dense problem



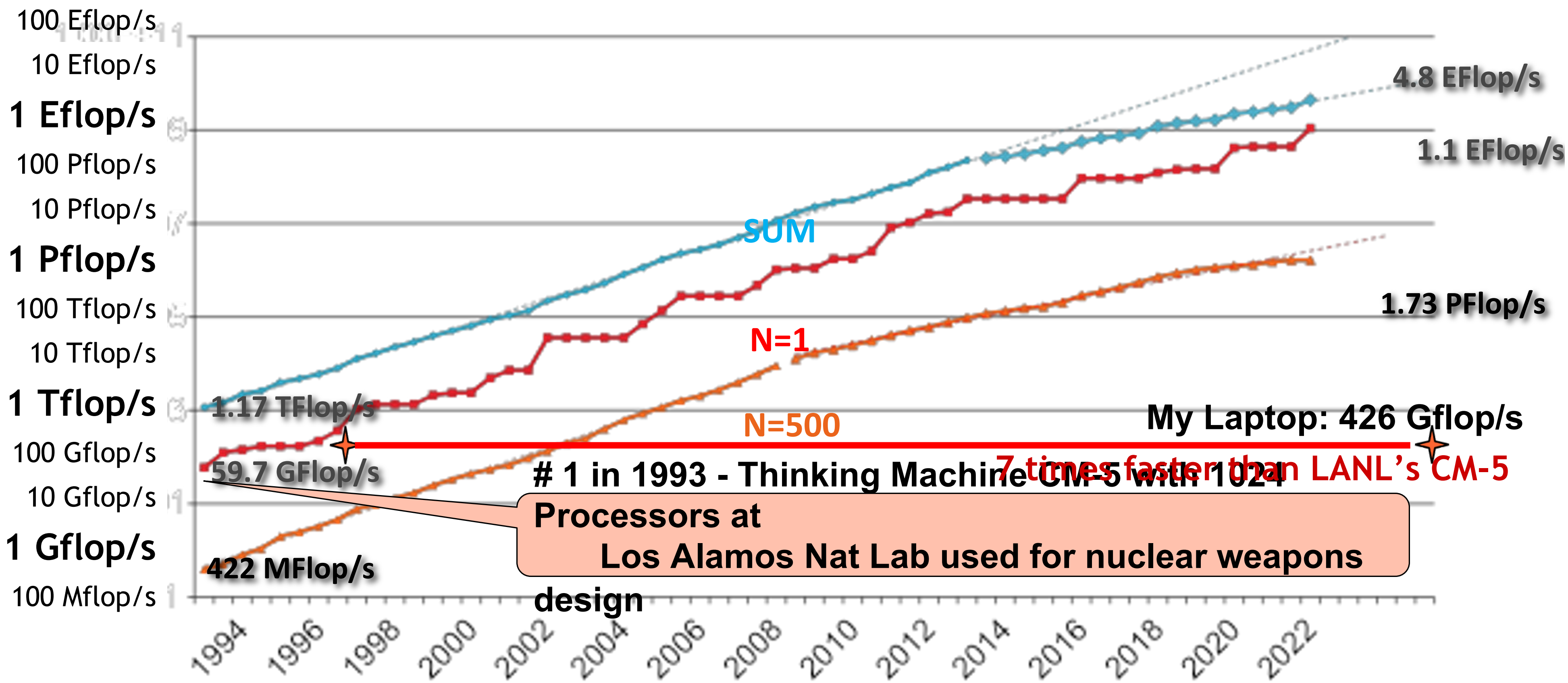
Maintained and updated twice a year:

SC'xy in the States in November

Meeting in Germany in June

TOP500 The List.

Performance Development of HPC over the Last 30 Years from the Top500



November 2022: The TOP 10 Systems (53% of the Total Performance of Top500)

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	GFlops/ Watt
1	DOE / OS Oak Ridge Nat Lab	Frontier, HPE Cray Ex235a, AMD 3rd EPYC (64C, 2 GHz), AMD Instinct MI250X, Slingshot 11	USA	7,733,248	1,102	65	21.1	52.2
2	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
3	EuroHPC /CSC	LUMI, HPE Cray EX235a, AMD 3rd EPYC (64C, 2 GHz), AMD Instinct MI250X, Slingshot 11	Finland	1,268,736	304.	72	2.94	52.3
4	EuroHPC/CINECA	BullSequana XH2000, Xeon Platinum 8358 (32C, 2.6GHz), NVIDIA A100 (108C), Quad-rail NVIDIA HDR100	Italy	1,463,616	175.	68	5.6	31.1
5	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
6	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
7	National Super Computer Center in Wuxi	Sunway TaihuLight, SW26010 (260C), Custom Interconnect	China	10,649,000	93.0	74	15.4	6.05
8	DOE / OS NERSC - LBNL	Perlmutter HPE Cray EX235n, AMD EPYC (64C, 2.45GHz), NVIDIA A100, Slingshot 10	USA	706,304	64.6	71	2.59	27.4
9	NVIDIA Corporation	Selene NVIDIA DGX A100, AMD EPYC 7742 (64C, 2.25GHz), NVIDIA A100 (108C), Mellanox HDR	USA	555,520	63.4	80	2.64	23.9
10	National Super Computer Center in Guangzhou	Tianhe-2A NUDT, Xeon (12C), MATRIX-2000 (128C) + Custom Interconnect	China	4,981,760	61.4	61	18.5	3.32

Current #1 System Overview

System Performance

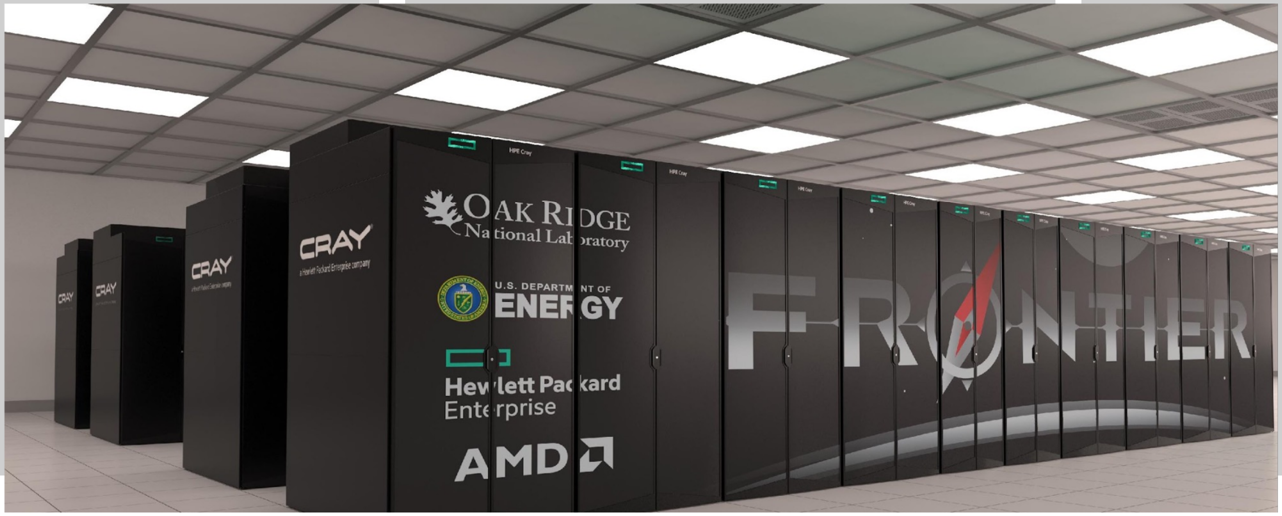
- Peak performance of 2 Eflop/s for modeling & simulation
- Power: 20+ MW
- Peak performance of 11.2 Eflop/s for 16 bit floating point used in for data analytics, ML, and artificial intelligence

Each node has

- 1-AMD EPYC 7A53 CPU w/64 cores (2 Tflop/s)
  < 1% performance of the system
- 4-AMD Instinct MI250X GPUs
  Each w/220 cores (4*53 Tflop/s)
  99% performance of the system
- 730 GB of fast memory
- 2 TB of NVMe memory

The system includes

- 9408 nodes
  37,632 GPUs
  8.88M Cores
- Cray Slingshot interconnect
  - 4 end points per node
- 706 PB Memory
  - (695 PB Disk + 11 PB SSD)



System Overview (Based on public data)

System Performance

- Peak performance of 3.34 Eflop/s for modeling & simulation @ 64 bit float pt
  - At 1.6 GHz (nominal, may be lower)
- Facility Power capacity 60 MW
- Peak performance of 53.5 Eflop/s for 16 bit floating point used in for data analytics, ML, and artificial intelligence

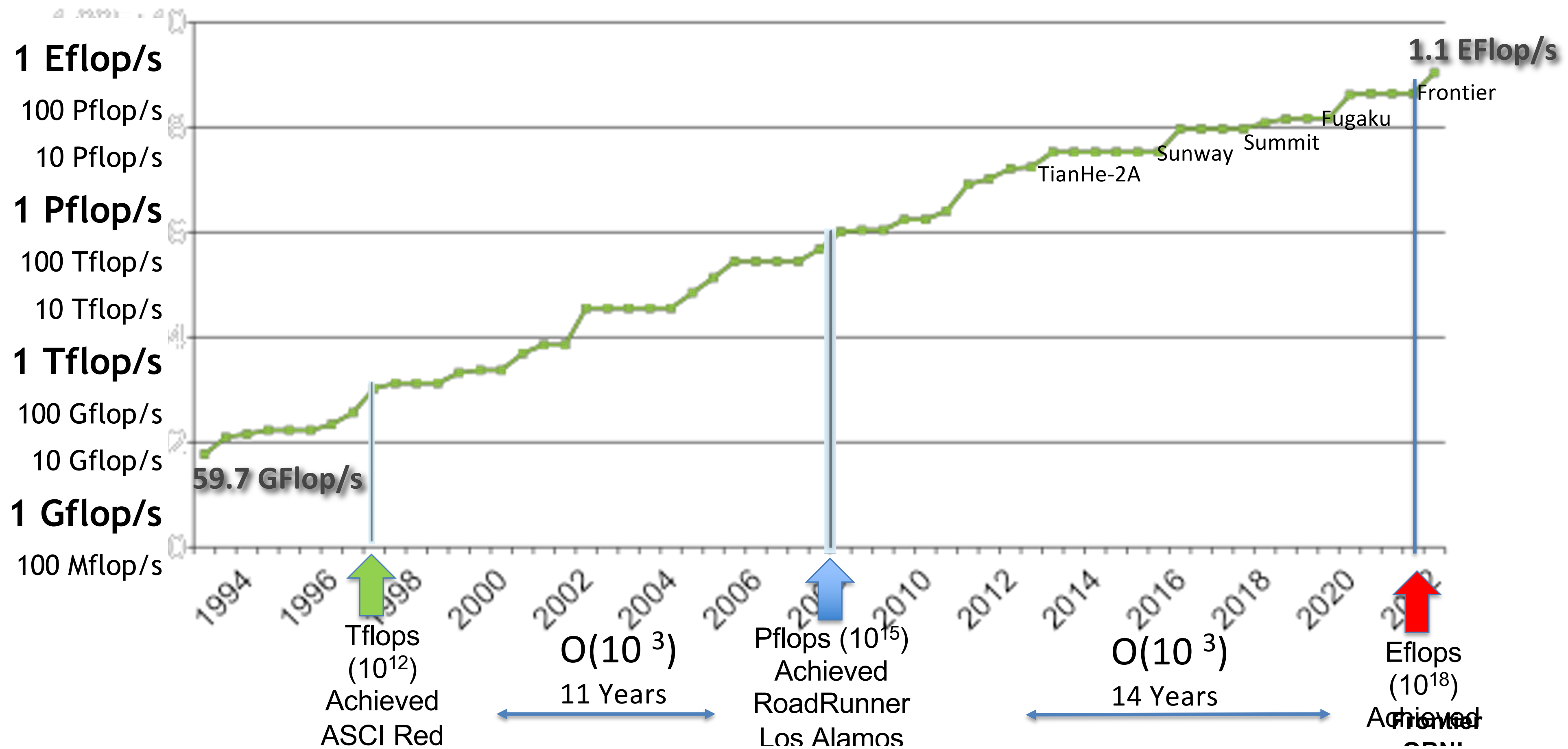
Each node has

- 2 - Intel Sapphire Rapids CPU processors; w/52 cores (5.3 Tflop/s)
  < 2% performance of the system
- 6 - Intel Xe Ponte Vecchio GPUs
- (6*52.4 Tflop/s = 314 Tflop/s)
  98% performance of the system
- 896 GB of HBM memory;
  Plus 1.02 TB DDR5 on the CPUs

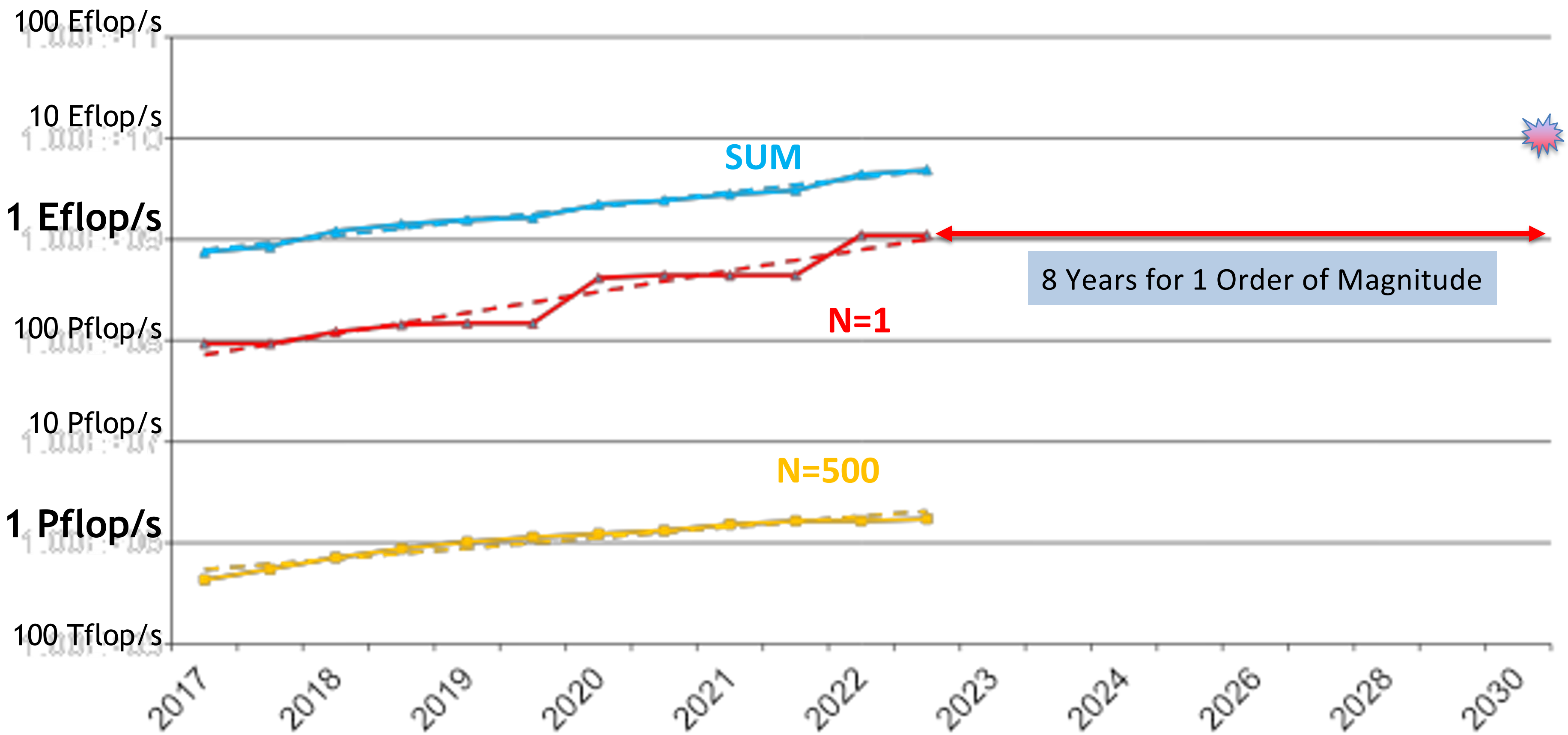
The system includes

- 10,624 nodes
  63,744 GPUs
  1.1M Cores
- Cray Slingshot interconnect
  - 8 end points per node
- 10.9 PB DDR Memory
- 9.52 PB HBM
  - (230 PB Intel Optane)
- 230 PB of NVMe memory total (DAOS servers)

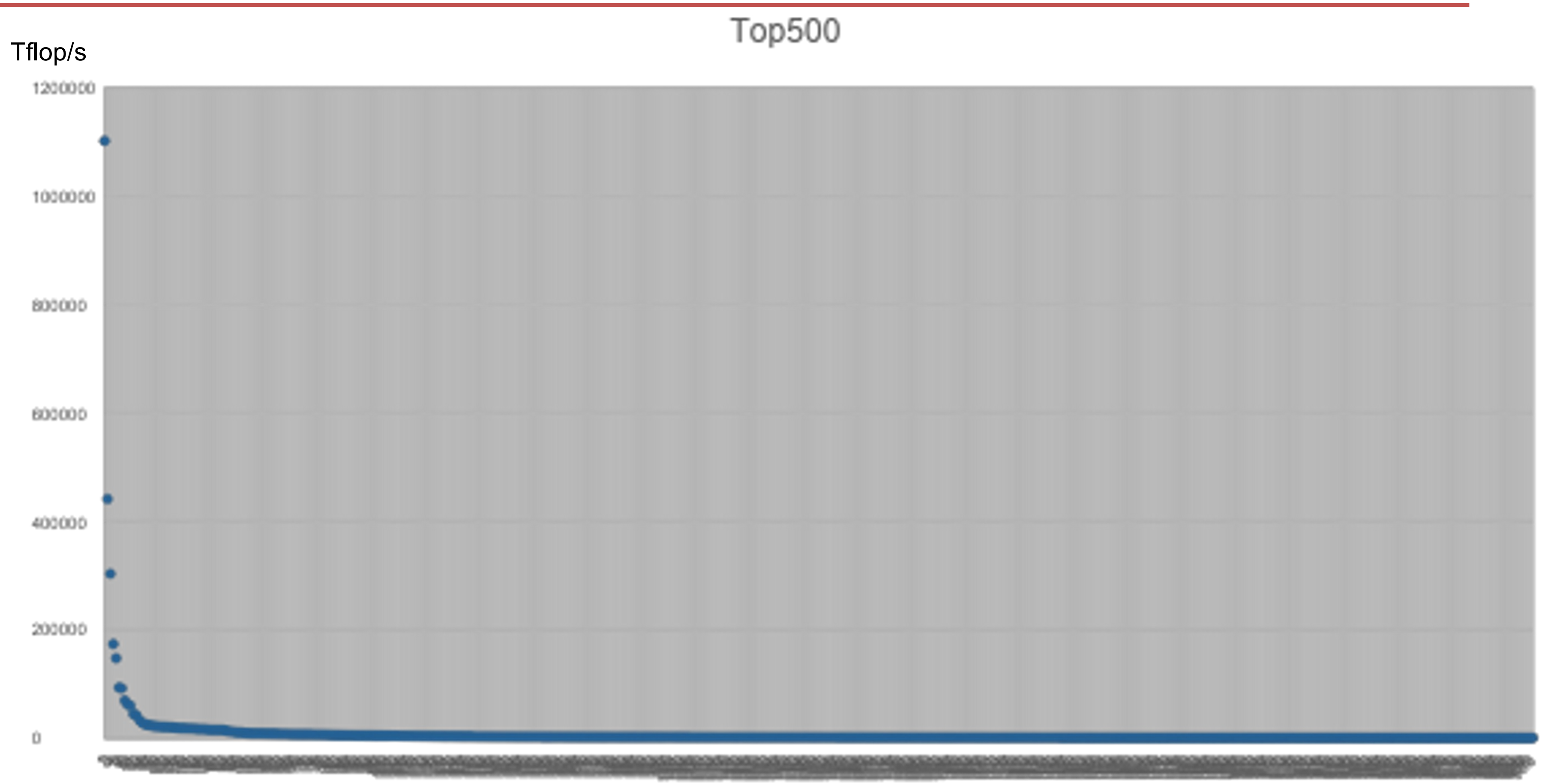
PERFORMANCE DEVELOPMENT



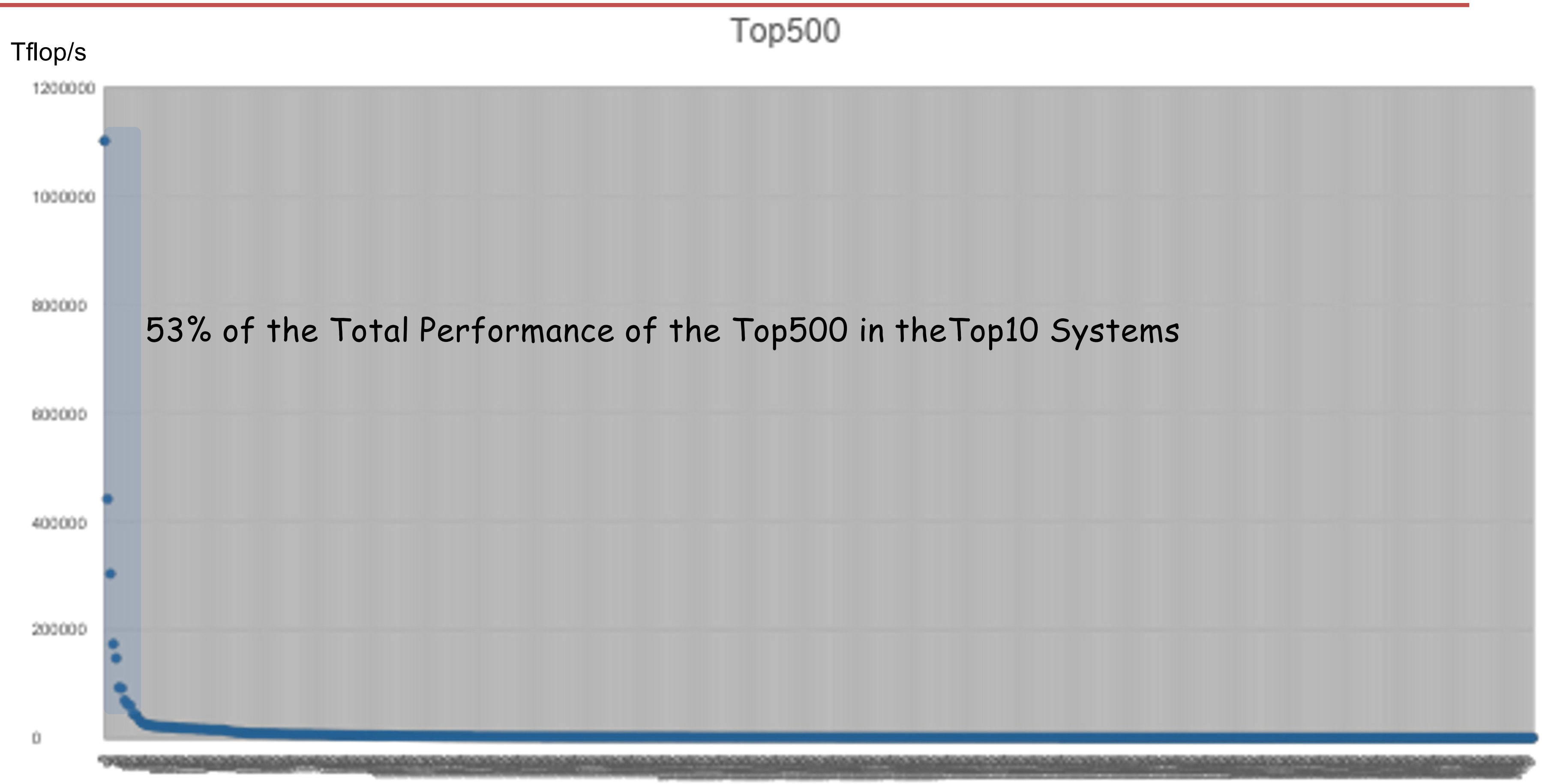
PROJECTED PERFORMANCE DEVELOPMENT



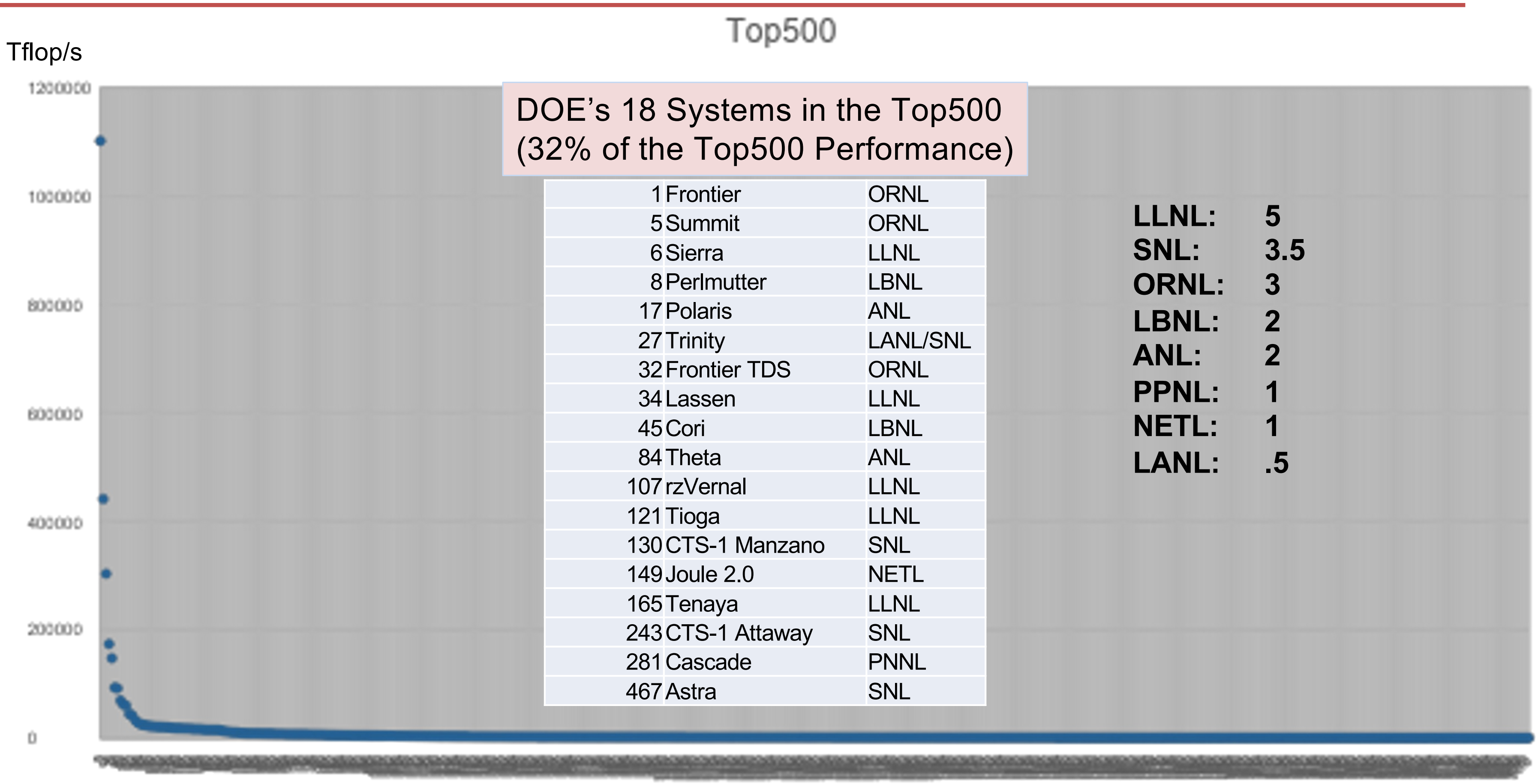
Plot of the Top500 Systems by Performance



Plot of the Top500 Systems by Performance



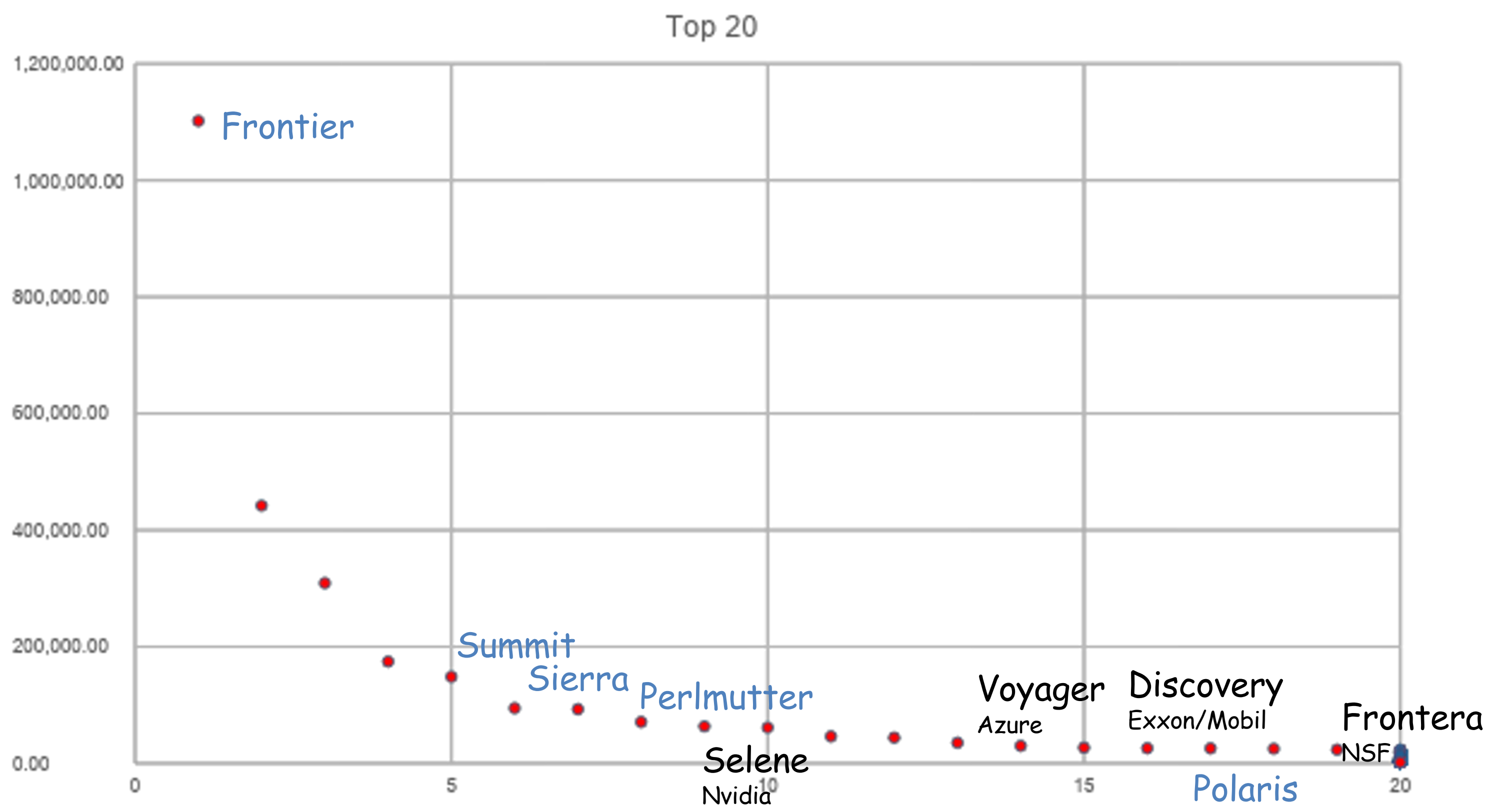
Plot of the Top500 Systems by Performance



1	Frontier	ORNL
5	Summit	ORNL
6	Sierra	LLNL
8	Perlmutter	LBNL
17	Polaris	ANL
27	Trinity	LANL/SNL
32	Frontier TDS	ORNL
34	Lassen	LLNL
45	Cori	LBNL
84	Theta	ANL
107	rzVernal	LLNL
121	Tioga	LLNL
130	CTS-1 Manzano	SNL
149	Joule 2.0	NETL
165	Tenaya	LLNL
243	CTS-1 Attaway	SNL
281	Cascade	PNNL
467	Astra	SNL

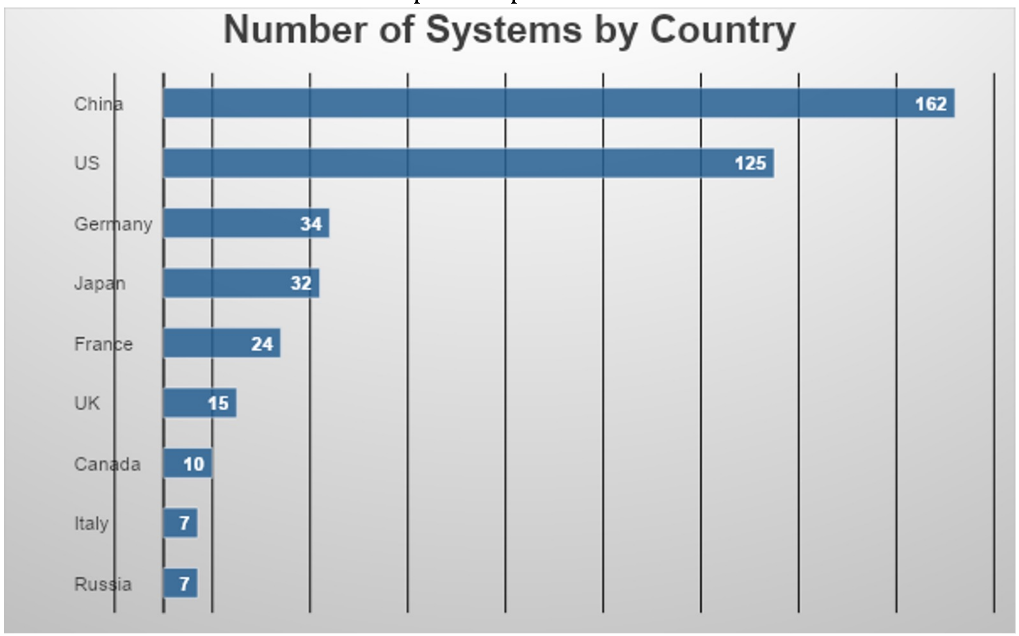
LLNL:	**5**
SNL:	**3.5**
ORNL:	**3**
LBNL:	**2**
ANL:	**2**
PPNL:	**1**
NETL:	**1**
LANL:	**.5**

Plot of the Top20 Systems by Performance



China

Supercomputers



China: Top consumer and producer overall.
5 main manufactures of HPC in China:
Lenovo, Sugon, Inspur, Huawei, NUDT

Rumored 2 Exascale Systems in Chinese

- Qingdao Marine Sunway Pro "OceanLight" (Shandong Prov)
  - Completed March 2021, 1.3 EFlops Rpeak and 1.05 EFlops Linpack
  - ShenWei post-Alpha CPU ISA architecture with large & small core structure
  - Est 96 cabinets x 1024 SW39010 390-core 35MW
  - Science on this machine won Gordon Bell Prize in 2021
- NSCC Tianjin Tianhe-3
  - Dual-chip FeiTeng ARM and Matrix accelerator node architecture
  - Est -1.7 EFlops Rpeak

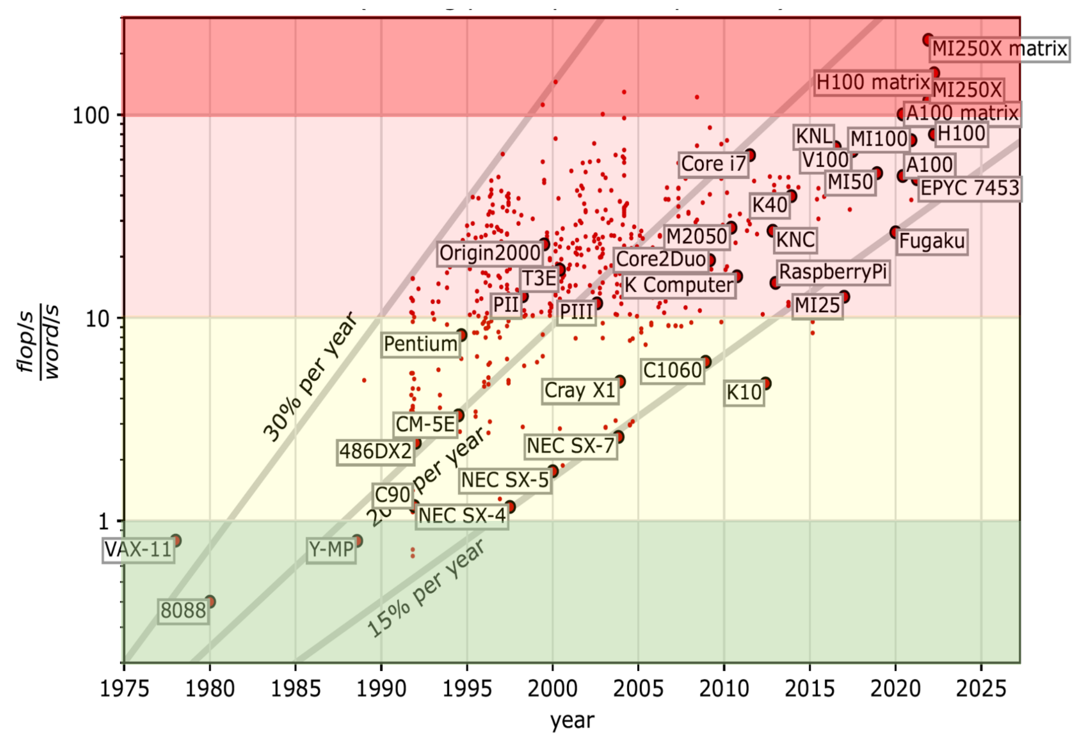
When We Look at Performance in Numerical Computations

…

- Data movement has a big impact
- Performance comes from balancing floating point execution (Flops/sec) with memory->CPU transfer rate (Words/sec)
  - “Best” balance would be 1 flop per word-transfered
- Today’s systems are close to 100 flops/sec per word-transferred
  - Imbalanced: Over provisioned for Flops

Machine Balance

Ratio of Fl Pt Ops per Data Movement over Time



Graph from Mark Gates

Plot for 64-bit floating point data movement & operations

(Bandwidth from CPU or GPU memory to registers)

Performance and Benchmarking Evaluation Tools

Linpack Benchmark - Longstanding benchmark started in 1979

- **Lots of positive features; easy to understand and run; shows trends**

However, much has changed since 1979

- **Arithmetic was expensive then and today it is over-provisioned and inexpensive**

Linpack performance of computer systems is no longer strongly correlated to real application performance

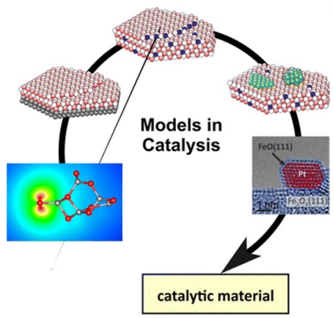
- **Linpack benchmark based on dense matrix multiplication**

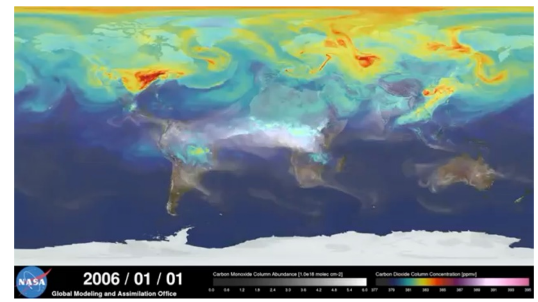
Designing a system for good Linpack performance can lead to design choices that are wrong for today's applications

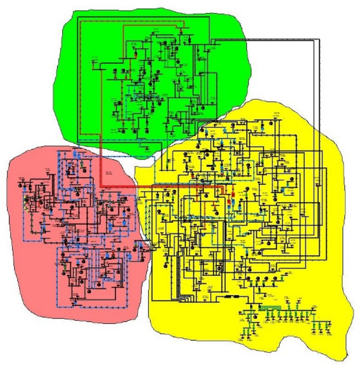
Today's Top HPC Systems Used to do Simulations

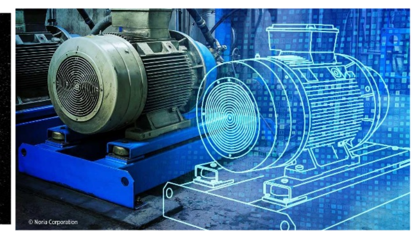
- *Climate*
- *Combustion*
- *Nuclear Reactors*
- *Catalysis*
- *Electric Grid*
- *Fusion*
- *Stockpile*
- *Supernovae*
- *Materials*
- *Digital Twins*
- *Accelerators*
- …

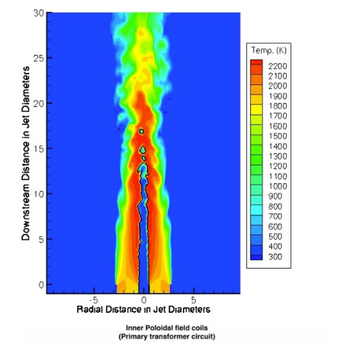
- Usually 3-D PDE's
  - Sparse matrix computations, not dense

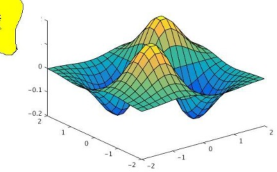


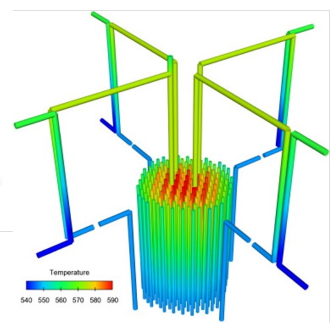


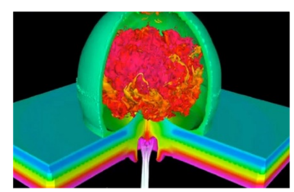


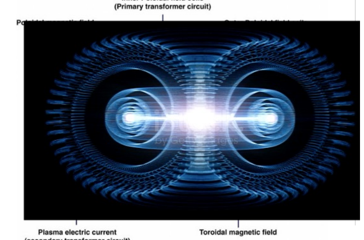








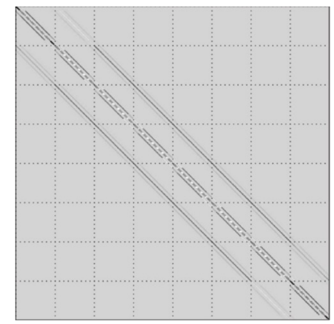




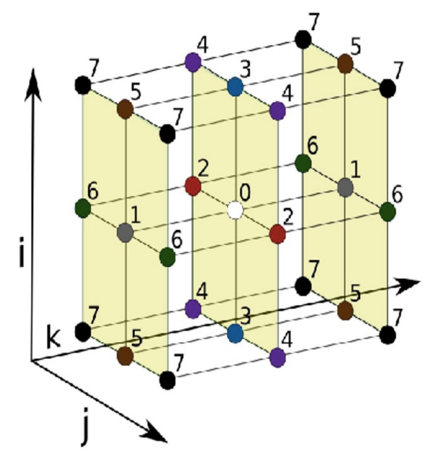
hpcg-benchmark.org With Piotr Luszczek and Mike Heroux

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



- 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).



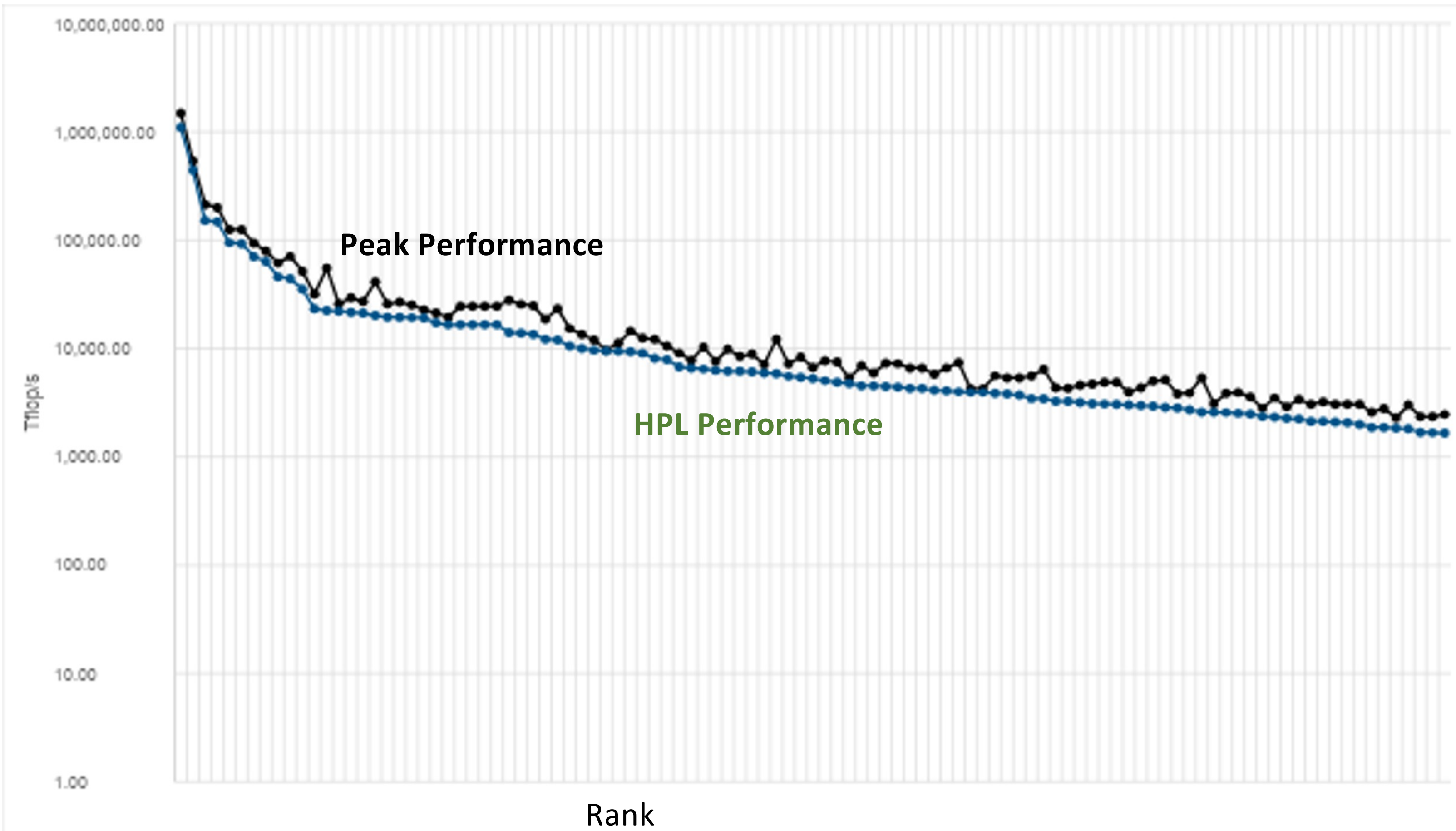
27-point stencil operator

HPCG Top 10, November 2022

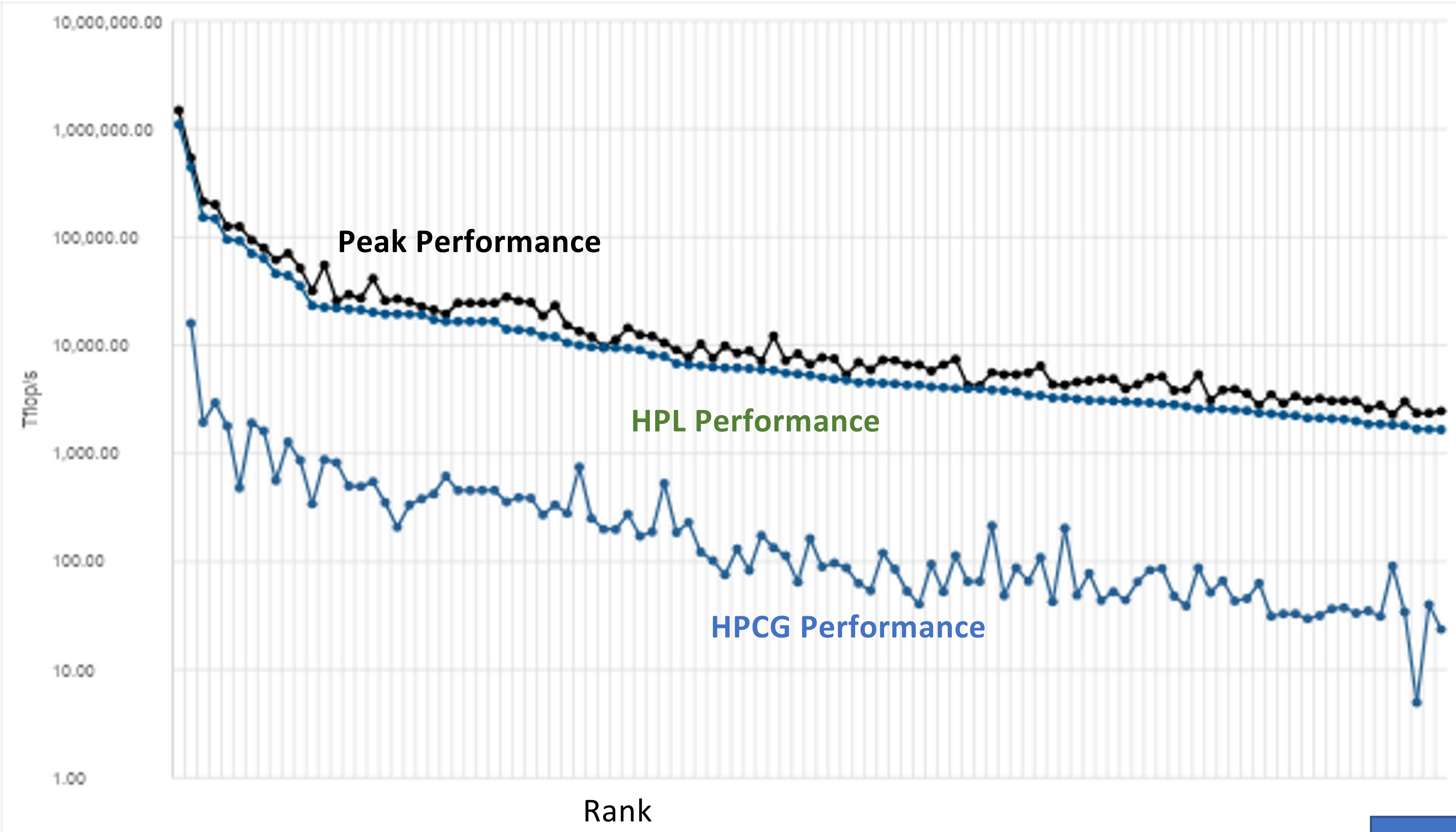
Rank	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	Fraction of Peak
1	RIKEN Center for Computational Science **Japan**	**Fugaku**, Fujitsu A64FX 48C 2.2GHz, Tofu D, Fujitsu	7,630,848	442	2	16.0	**3.0%**
2	DOE/SC/ORNL **USA**	**Frontier,** HPE Cray Ex235a, AMD 3rd EPYC 64C, 2 GHz, AMD Instinct MI250X, Slingshot 10	8,730,112	1,102	1	14.1	**0.8%**
3	EuroHPC/CSC **Finland**	**LUMI**, HPE Cray EX235a, AMD Zen-3 (Milan) 64C 2GHz, AMD MI250X, Slingshot-11	2,174,976	304	3	3.41	**0.8%**

Think of a race car that has the potential of 200 MPH but only goes 2 MPH!

5	EuroHPC/CINECA **Italy**	Leonardo, BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 40 GB, Quad-rail NVIDIA HDR100 Infiniband	1,463,616	175	4	2.57	**1.0%**
6	DOE/SC/LBNL **USA**	**Perlmutter**, HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10	761,856	70.9	8	1.91	**2.0%**
7	DOE/NNSA/LLNL **USA**	**Sierra**, S922LC, IBM POWER9 20C 3.1 GHz, Mellanox EDR, NVIDIA Volta V100, IBM	1,572,480	94.6	6	1.80	**1.4%**
8	NVIDIA **USA**	**Selene**, DGX SuperPOD, AMD EPYC 7742 64C 2.25 GHz, Mellanox HDR, NVIDIA Ampere A100	555,520	63.5	9	1.62	**2.0%**
9	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	12	1.28	**1.8%**
10	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	20	0.88	**1.6%**



10,000,000.00
1,000,000.00
100,000.00
10,000.00
1,000.00
100.00
10.00
1.00
Tflop/s
Peak Performance
HPL Performance
Rank


10,000,000.00
1,000,000.00
100,000.00
10,000.00
1,000.00
100.00
10.00
1.00
Tflop/s
Peak Performance
HPL Performance
HPCG Performance
Rank

From a talk: Distributed Training of Large Language Models on Fugaku
Rio Yokota, Tokyo Tech

GPT-4 would take 8 Years to train on Fugaku or 3 months on the OpenAI platform

How Long It Will Take to Train GPT

GPT-4: 3×10^{25} FLOPs (speculated)
GPT-3.5 (ChatGPT): 3×10^{24} FLOPs (speculated)
GPT-3: 3×10^{23} FLOPs

Fugaku:
FP32 6.76TFLOP/s x 158,976 = 1.07 EFLOP/s (theoretical peak)
GPT-4: 328 days x 10
GPT-3.5: 32 days x 10
GPT-3: 3.3 days x 10

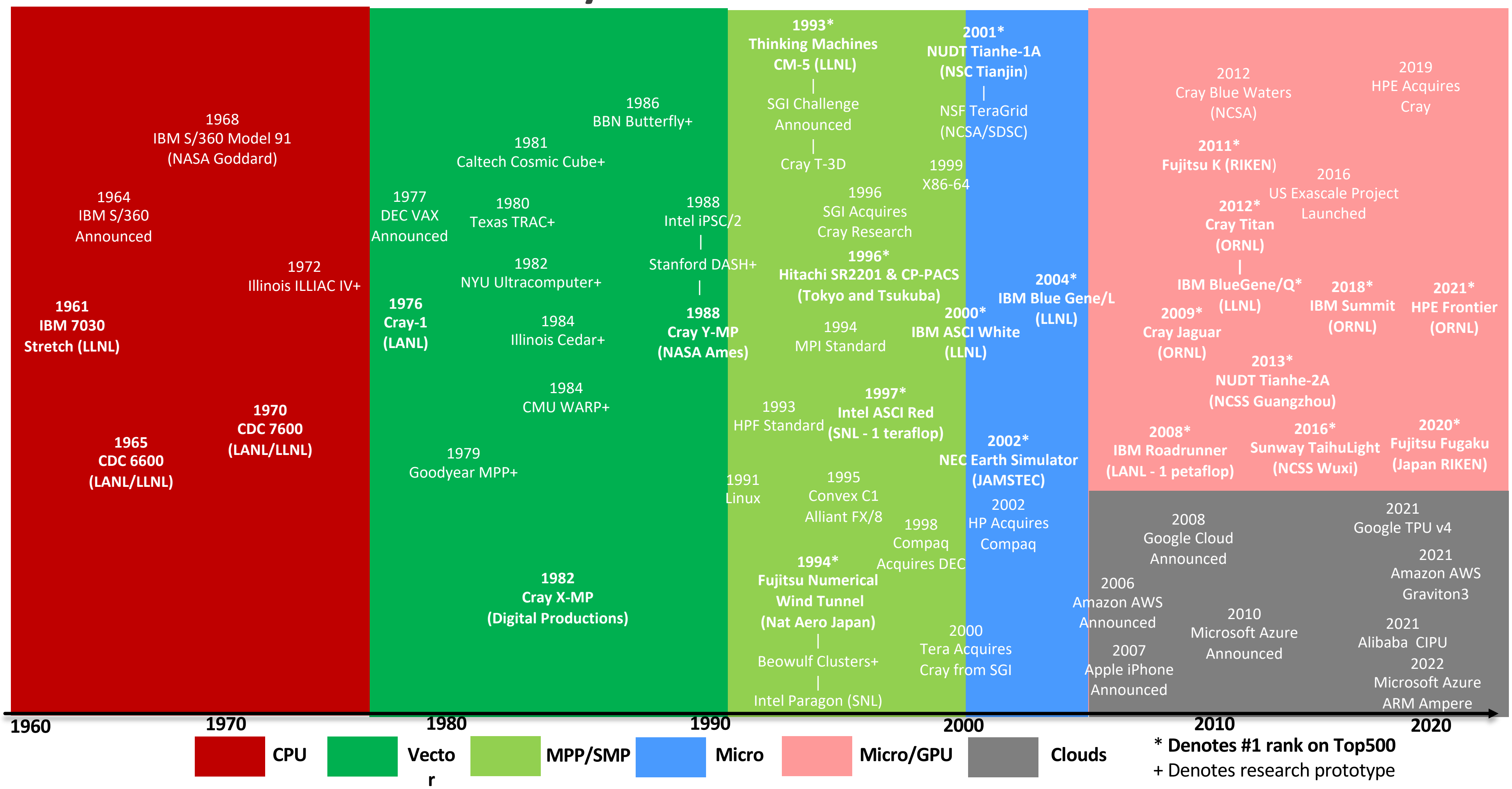
Assuming 10% efficiency
(No 16-bit Fl. Pt.)

OpenAI:
BF16 312 TFLOP/s x 25,000 = 7.8 EFLOP/s (theoretical peak)
GPT-4: 45 days x 2
GPT-3.5: 4.5 days x 2
GPT-3: 11 hours x 2

Assuming 50% efficiency

Actual Performance

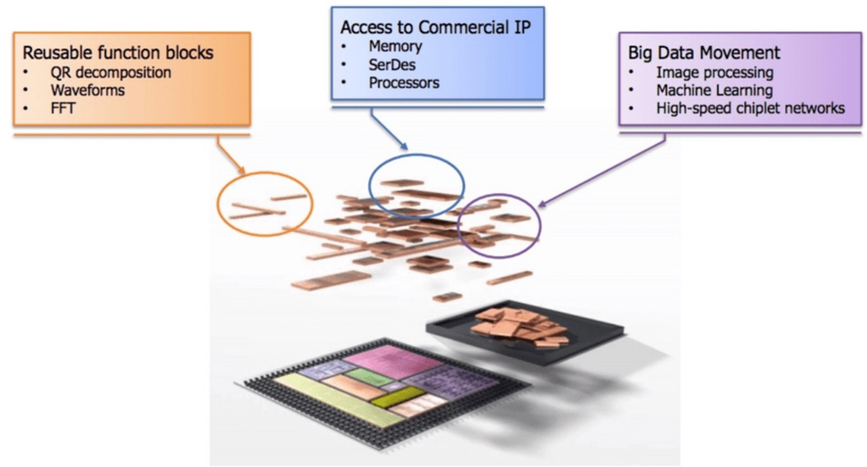
HPC: From Diversity to Monoculture



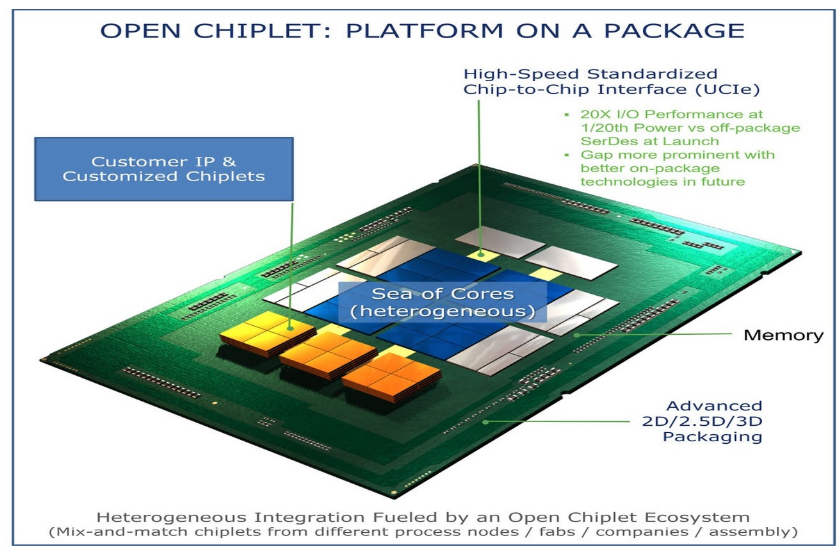
* Reed, Gannon, Dongarra, "HPC Forecast: Cloudy and Uncertain," *Communications of the ACM* (February 2023)

Chiplets: Integrating Multiple Functions

- Rather than fabricating a monolithic system-on-a-chip, chiplet technology combines multiple chips, each representing a portion of the desired functionality, possibly fabricated using different processes by different vendors and perhaps including IP from multiple sources
- Chiplet designs are part of the recent offerings from Intel and AMD
  - Amazon's Graviton3 also uses a chiplet design with seven different chip dies
  - Advanced Query Accelerator (AQUA) for AWS Redshift, Amazon's powerful and popular data warehouse service, relies on a package of custom ASICs and FPGA accelerators



CHIPS modularity targets the enabling of a wide range of custom solutions



Conclusions

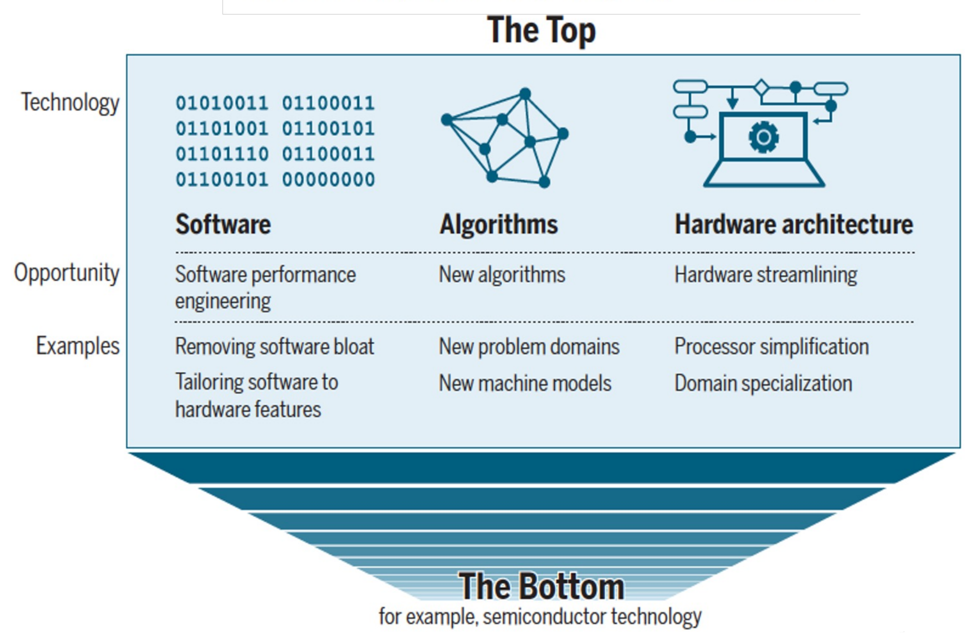
- The computing ecosystem is in enormous flux, creating both opportunities and challenges for the future of advanced scientific computing
- Looking forward, it seems increasingly unlikely that future high-end HPC systems will be procured and assembled solely by commercial integrators from only commodity components
- Advances will require embracing end-to-end design, testing and evaluating advanced prototypes, and partnering strategically … real co-design.
- Leading edge, HPC computing systems are increasingly similar to large-scale scientific instruments (LHC, LIGO, SKA) with limited economic incentives for commercial development

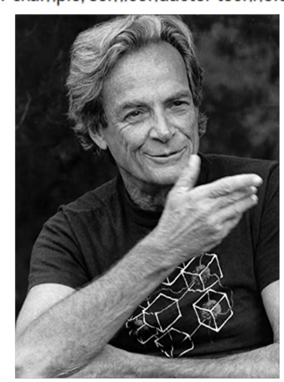
The Take Away

- 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

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





Feynman’s 1959
Lecture @