CSE5351: Parallel Processing

Part 1B

State of the Art In Supercomputing

Several of the next slides (or modified) are the courtesy of Dr. Jack Dongarra, a distinguished professor of Computer Science at the University of Tennessee.

What is meant by performance?

- **What is a xflop/s?**
  - **xflop/s is a rate of execution, some number of floating point operations per second.**
    - **Whenever this term is used it will refer to 64 bit floating point operations and the operations will be either addition or multiplication.**
- **What is the theoretical peak performance?**
  - **The theoretical peak is based not on an actual performance from a benchmark run, but on a paper computation to determine the theoretical peak rate of execution of floating point operations for the machine.**
  - **The theoretical peak performance is determined by counting the number of floating-point additions and multiplications (in full precision) that can be completed during a period of time, usually the cycle time of the machine.**
  - **For example, an Intel Xeon 5570 quad core at 2.93 GHz can complete 4 floating point operations per cycle or a theoretical peak performance of 11.72 GFlop/s per core or 46.88 Gflop/s for the socket.**

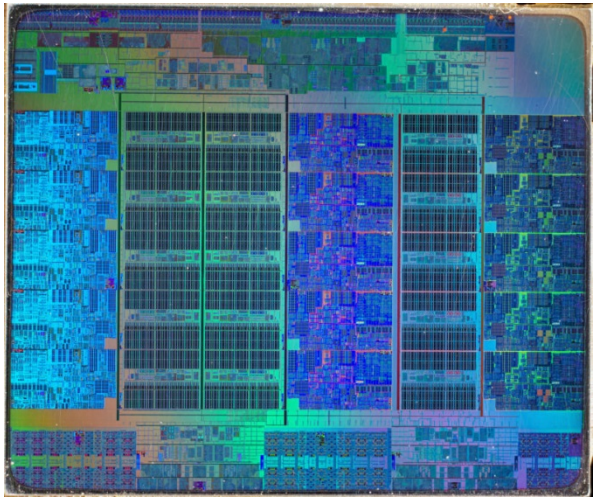
State of Supercomputing in 2021

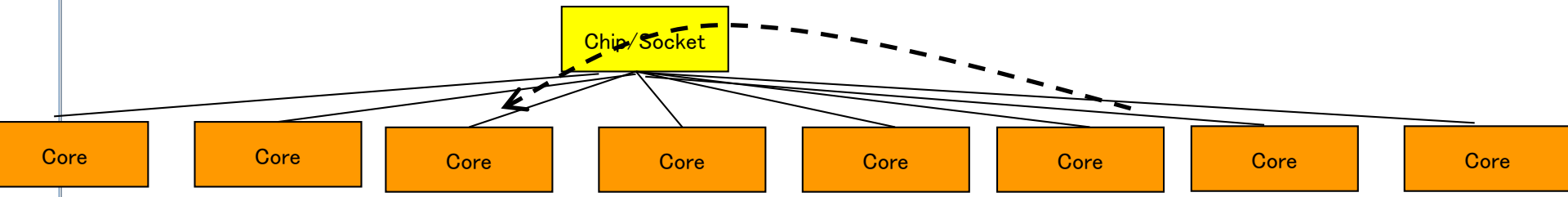
$$\text{FLOPS} = \text{cores} \times \text{clock} \times \frac{\text{FLOPs}}{\text{cycle}}$$

Computer performance

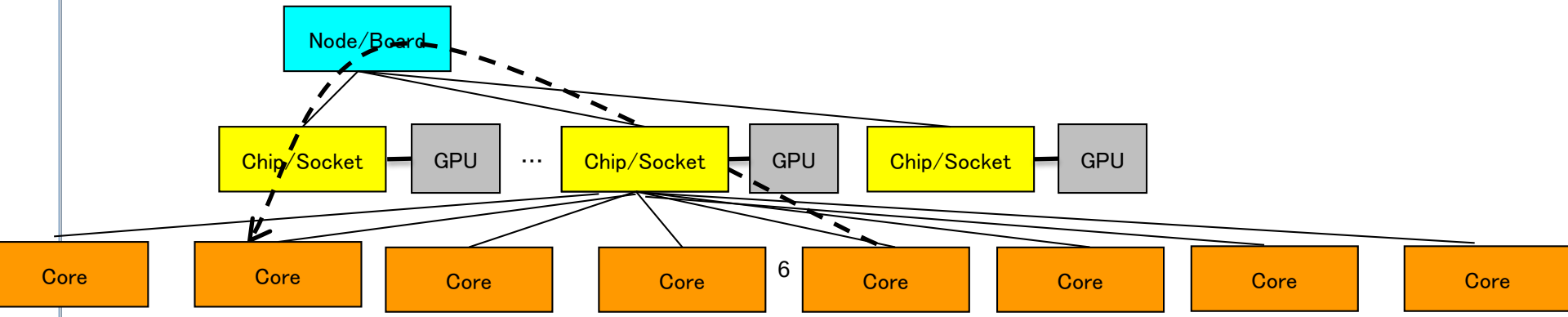
Name	FLOPS
yottaFLOPS	10^{24}
zettaFLOPS	10^{21}
exaFLOPS	10^{18}
petaFLOPS	10^{15}
teraFLOPS	10^{12}
gigaFLOPS	10^{9}
megaFLOPS	10^{6}
kiloFLOPS	10^{3}

Example of a Typical Supercomputer





Example of a Typical Supercomputer



Example of a Typical Supercomputer

Shared memory programming between processes on a board and
a combination of shared memory and distributed memory programming
between nodes and cabinets

Example of a Typical Supercomputer

Combination of shared memory and distributed memory programming

State of Supercomputing in 2021

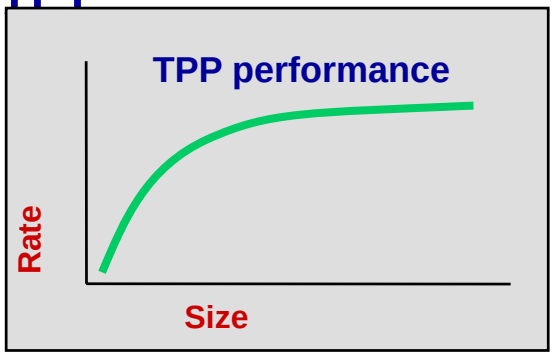
- **Pflops (> 10^{15} Flop/s) computing fully established with all 500 systems.**
- **Three technology architecture possibilities or "swim lanes" are thriving.**
  - **Commodity (e.g. Intel)**
  - **Commodity + accelerator (e.g. GPUs) (160 systems; 144 NVIDIA, 11 Intel Phi + 5)**
  - **Lightweight cores (e.g. IBM BG, Xeon Phi, TaihuLight, ARM (5 system))**
- **China: Top consumer and producer overall (but not for the TOP50 (yet))**
- **Interest in supercomputing is now worldwide, and growing in many new markets (~50% of Top500 computers are in industry).**
- **Intel processors largest share, 81% followed by AMD, 15%.**
- **Exascale (10^{18} Flop/s) projects exist in many countries and regions.**

Since 1993

H. Meuer, H. Simon, E. Strohmaier, & JD

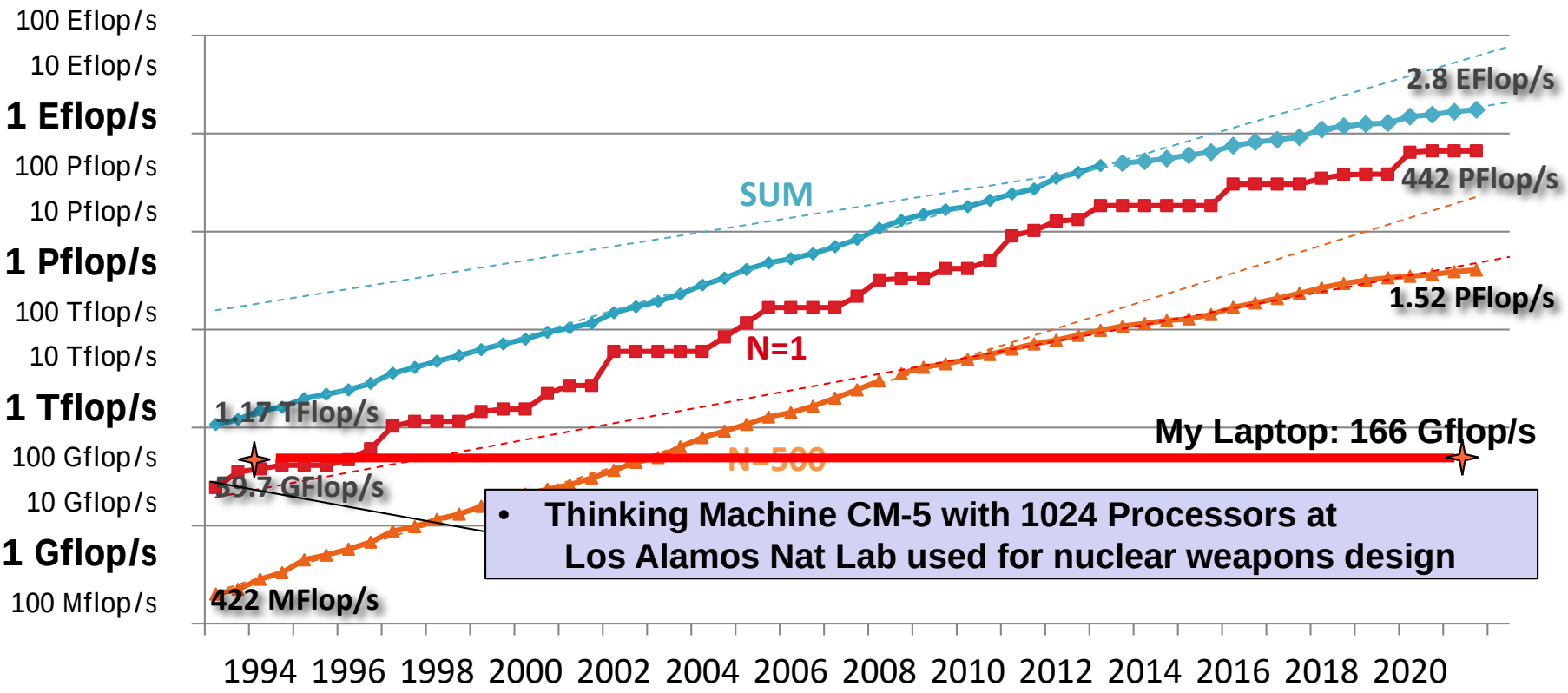
- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

  $Ax=b$, *dense problem*

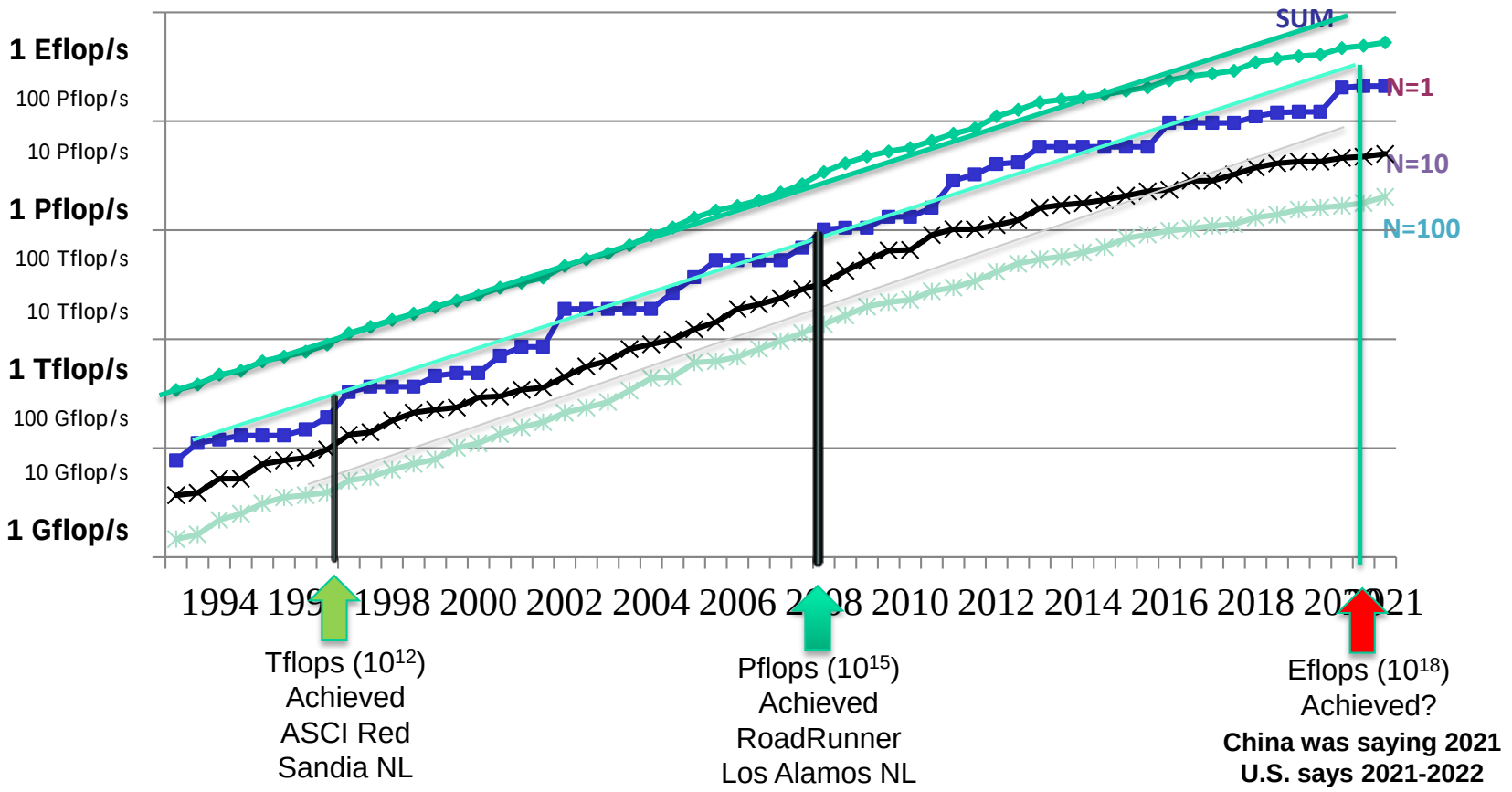


- Updated twice a year
  SC‘xy in the States in November
  Meeting in Germany in June
- All data available from **www.top500.org**

Performance Development of HPC over the Last 28 Years from the Top500



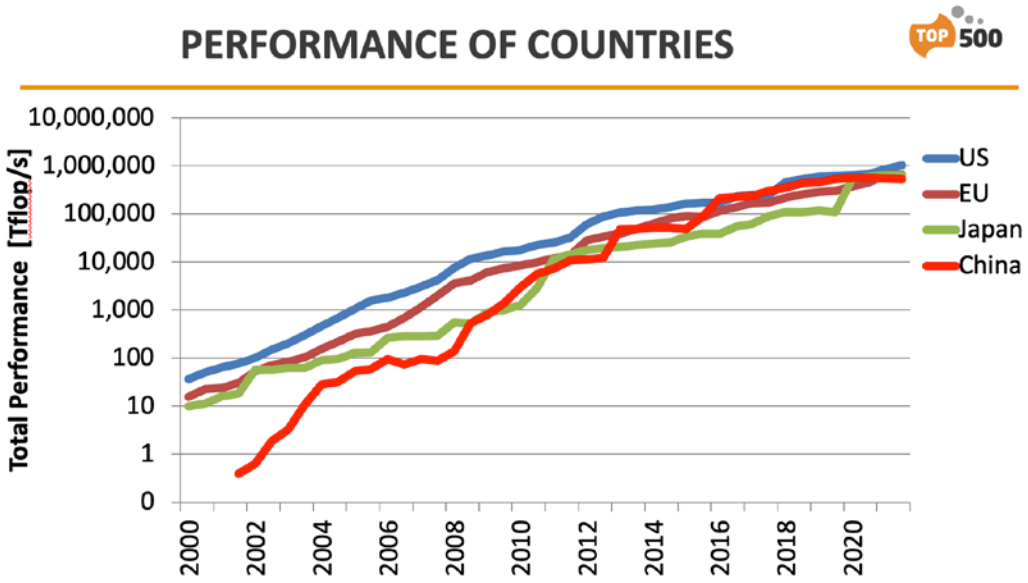
Performance Development



November 2021: The TOP 10 Systems (35% of the Total Performance of Top500)

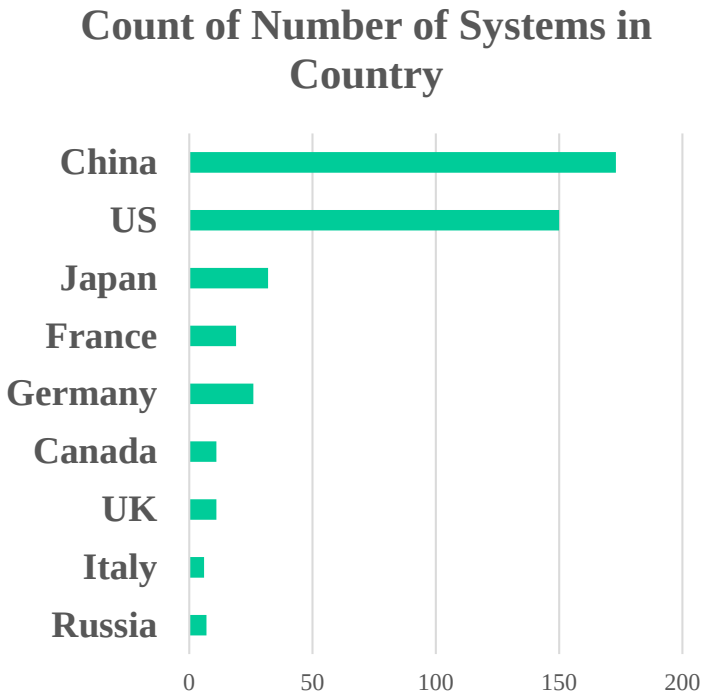
Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	GFlops/ Watt
1	RIKEN Center for Computational Science	Fugaku, ARM A64FX (48C, 2.2 GHz), Tofu D Interconnect	Japan	7,299,072	442.	82	29.9	14.8
2	DOE / OS Oak Ridge Nat Lab	Summit, IBM Power 9 (22C, 3.0 GHz), NVIDIA GV100 (80C), Mellonox 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	Azure East US 2 Microsoft Azure	Voyager-EUS2, ND96amsr_A100_v4, AMD EPYC (48C 2.45GHz), NVIDIA A100 (80C), Mell. HDR	USA	253,440	30.1	76		

Countries Share

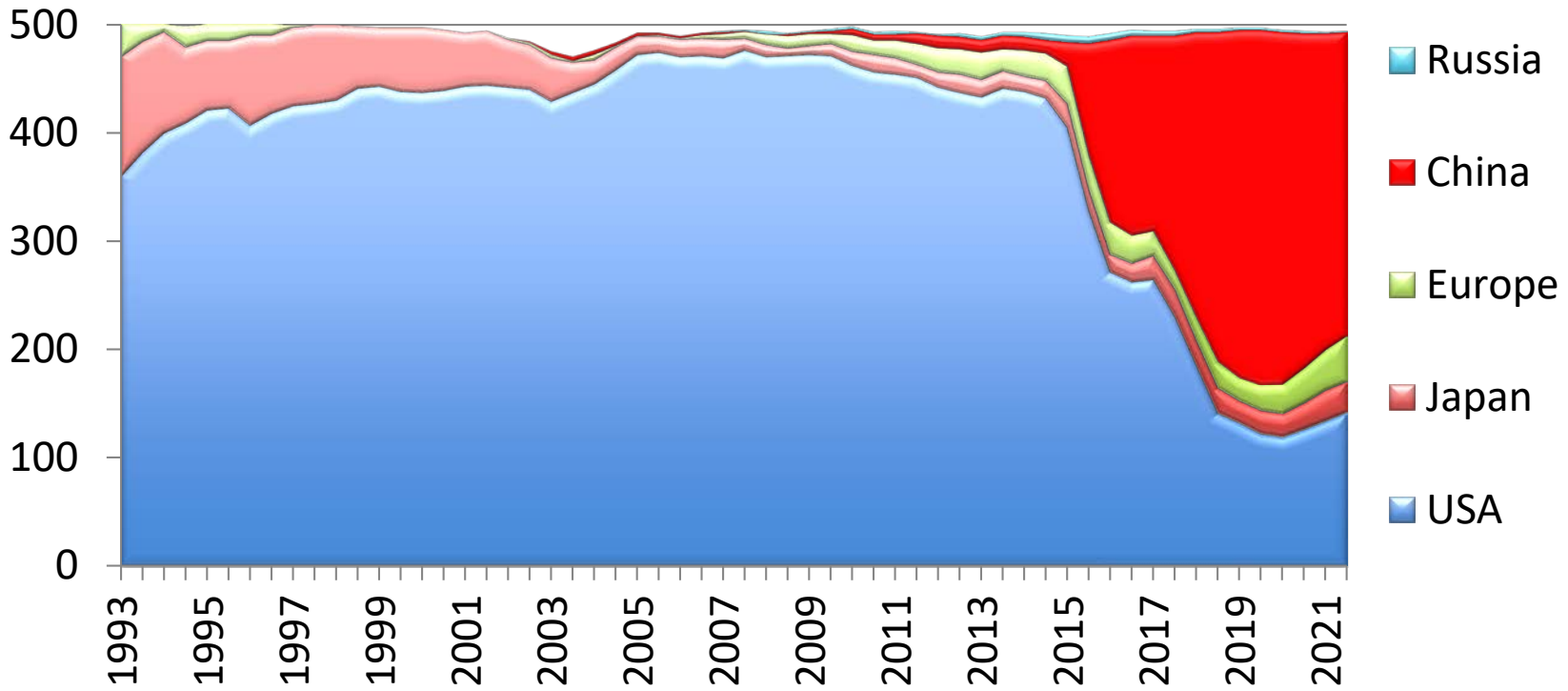


In terms of number of systems: China has 35% of the systems
US has 30% of the systems
Japan has 6% of the systems

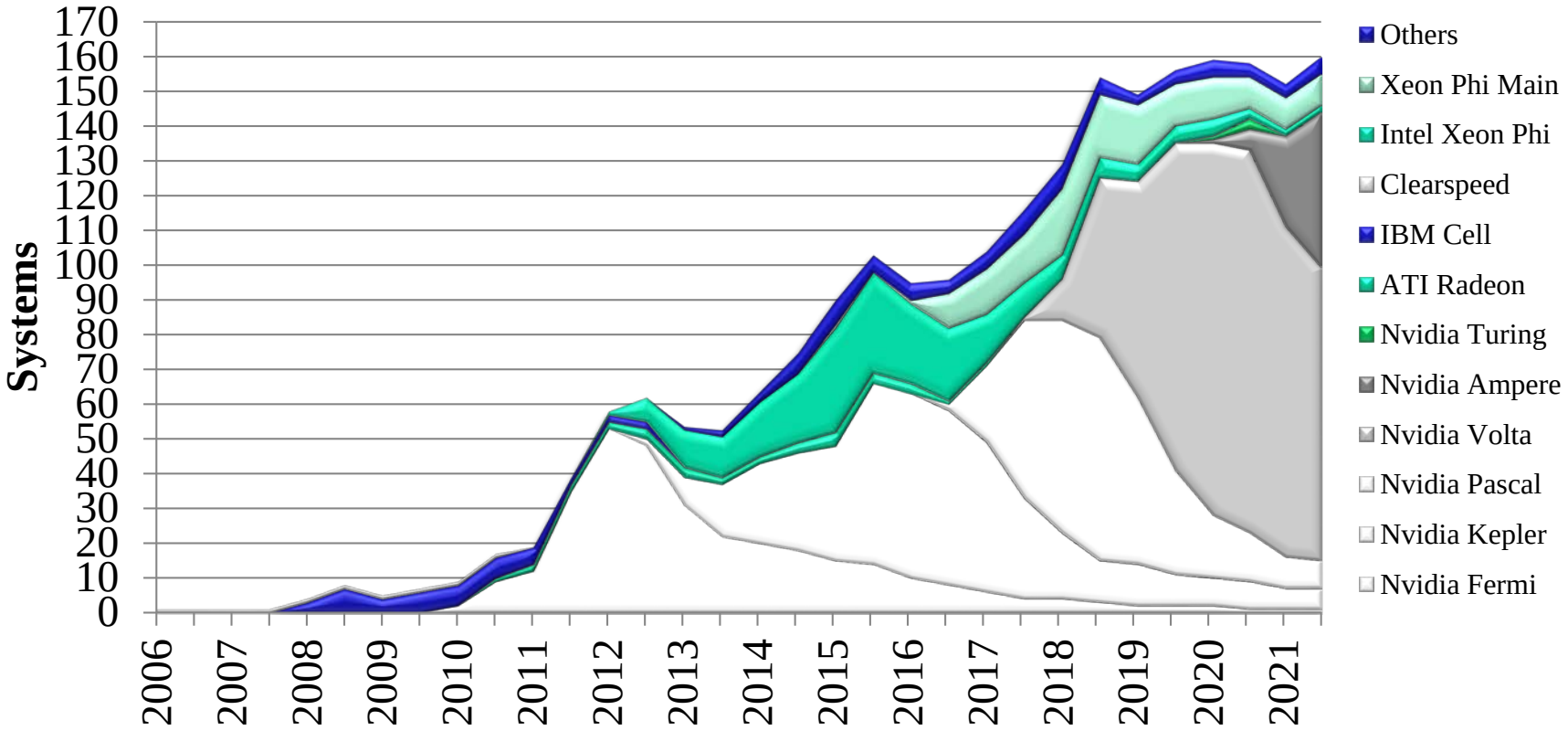
In terms of performance: US has 33%
Japan has 20%
China has 17%



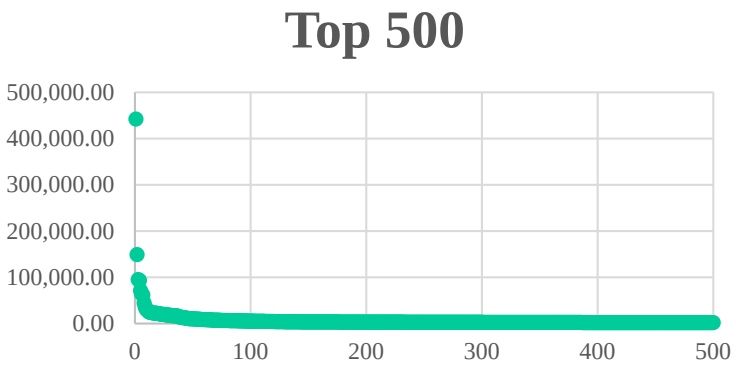
Producers

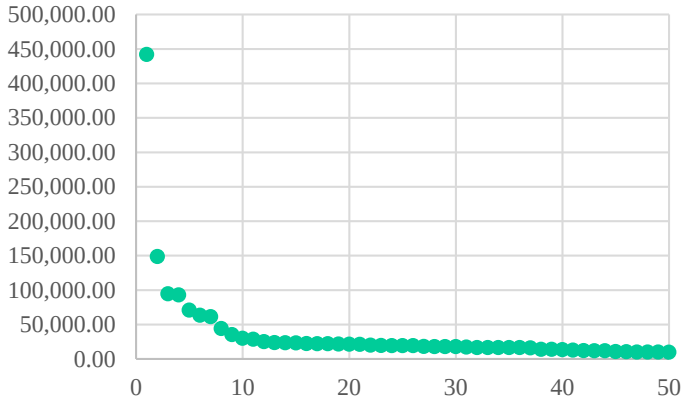


Accelerators



Performance Distribution





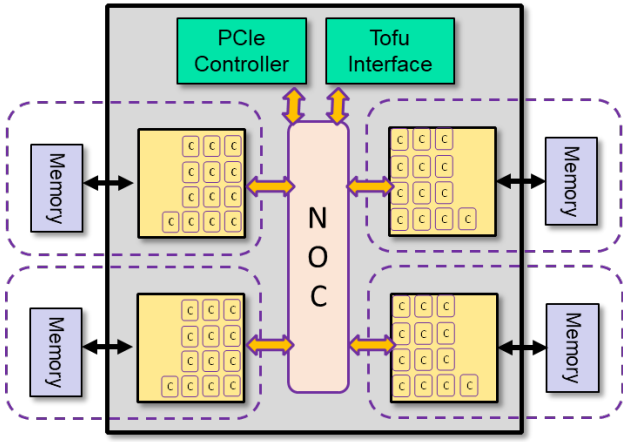
For all Top 500 systems
$Tpeak_{500}$ = 1.649 Pflop/s

For Top 50 systems
$Tpeak_{50}$ = 9.95 Pflop/s

58% of the total performance of Top500 in Top5

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

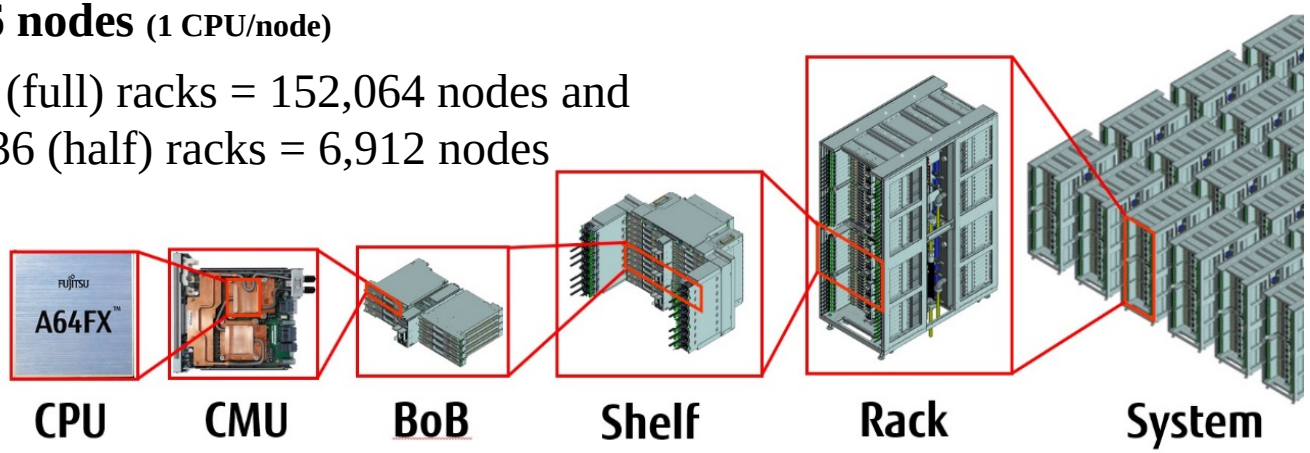


- …but also an accelerated GPU-like processor
  - SVE 512 bit x 2 vector extensions (ARM & Fujitsu)
    - Integer (1, 2, 4, 8 bytes) + Float (16, 32, 64 bytes)
  - Cache + memory localization (sector cache)
  - HBM2 on package memory – Massive Mem BW (Bytes/DPF ~0.4)
    - Streaming memory access, strided 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 m^2

- **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 14% of all the other Top500 systems.

http://bit.ly/fugaku-report

#2 System Overview

System Perf

- **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

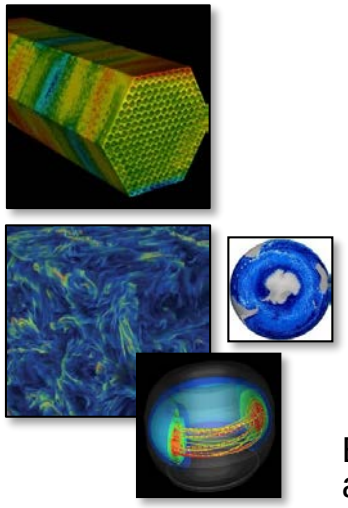
- **2 IBM POWER9 processors**
  - **Each w/22 cores**
  - **2.3% performance of system**
- **6 NVIDIA Tesla V100 GPUs**
  - **Each w/80 SMs**
  - **97.7% performance of system**
- **608 GB of fast memory**
- **1.6 TB of NVMe memory**

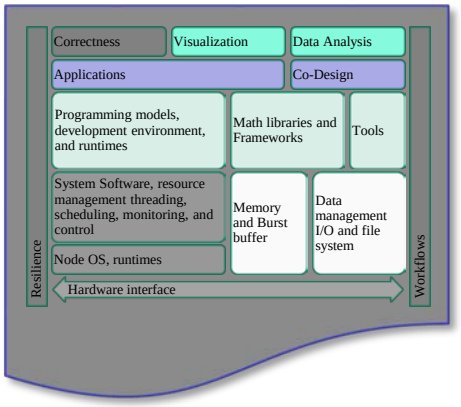
The system includes

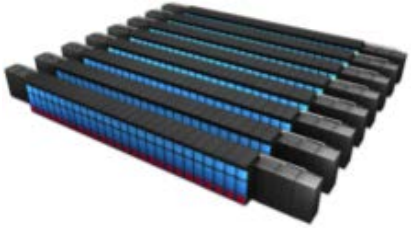
- **4608 nodes**
  - **27,648 GPUs**
  - **Street value $15K each**
- **Dual-rail Mellanox EDR InfiniBand network**
- **250 PB IBM Spectrum Scale file system transferring data at 2.5 TB/s**

US Department of Energy Exascale Computing Program has formulated a holistic approach that uses co-design and integration to achieve capable exascale

Application Development	Software Technology	Hardware Technology	Exascale Systems
Science and mission applications	Scalable and productive software stack	Hardware technology elements	Integrated exascale supercomputers

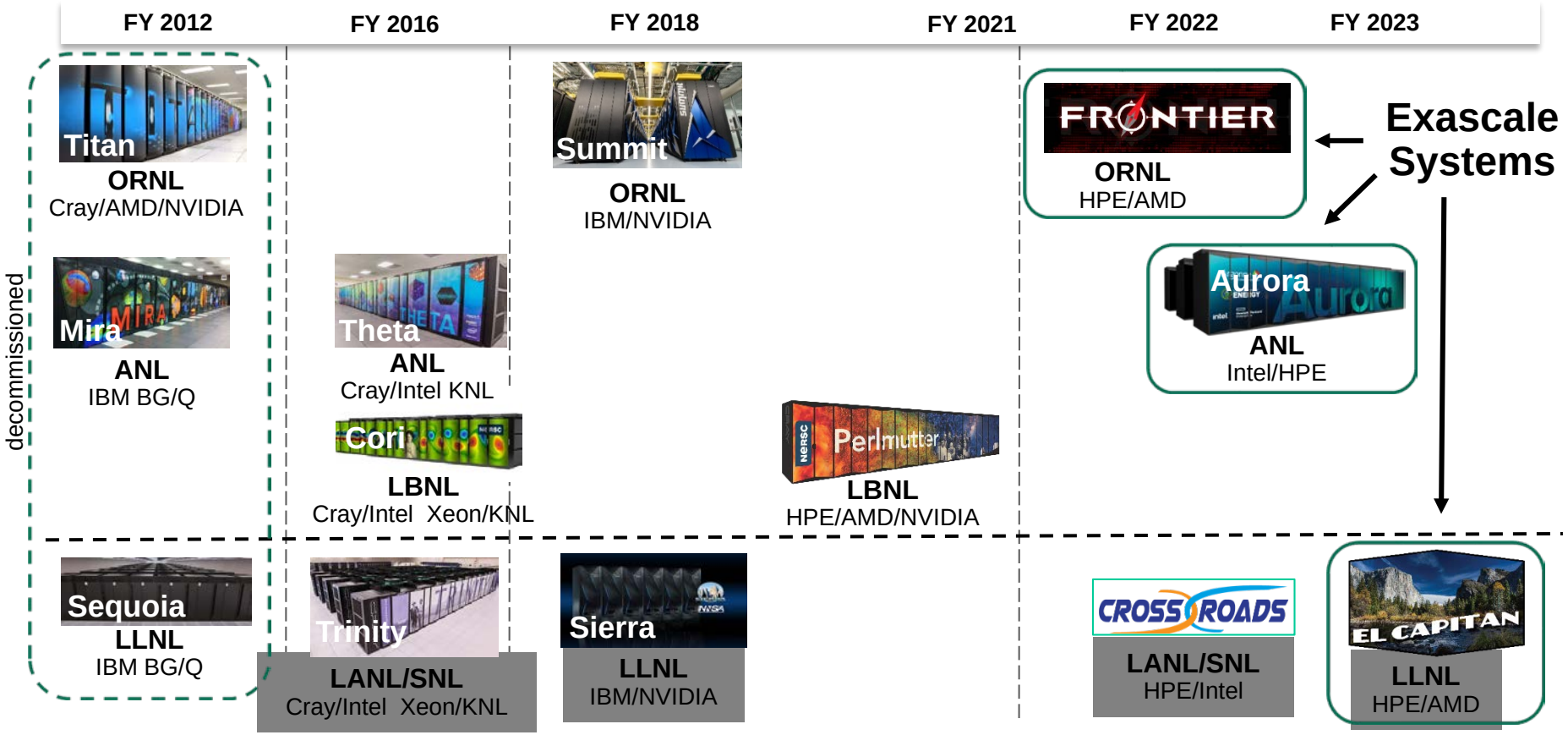






ECP's work encompasses applications, system software, hardware technologies and architectures, and workforce development

DOE HPC Roadmap to Exascale Systems



Exascale architectures will excel at HPC and HPC + AI problems

Exascale is costing DOE $3.6B in total, over 5 years

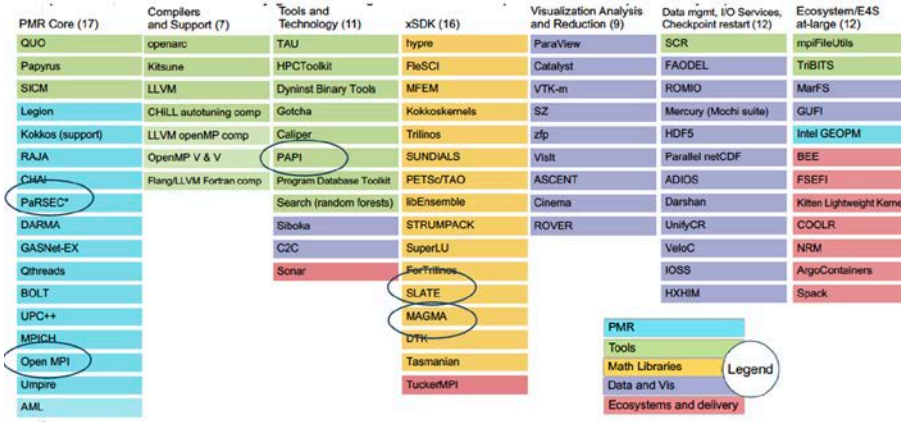
What do you get for $3.6B

- 3 computers
  - $600M each
- 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^{2-3} 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

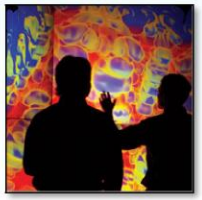
- A bunch of software (84 projects)

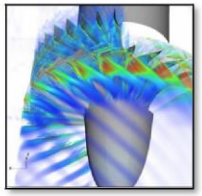


Department of Energy's Computational Science Problems of Interest

National security

- Stockpile stewardship
- Next-generation electromagnetics
- simulation of hostile environment and virtual flight testing for hypersonic re-entry vehicles



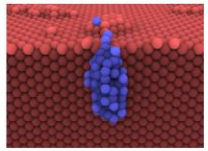


Energy

- Turbine wind plant efficiency
- High-efficiency, low-emission combustion engine and gas turbine design
- Materials design for extreme environments of nuclear fission and fusion reactors
- Design and commercialization of Small Modular Reactors
- Subsurface use for carbon capture, petroleum extraction, waste disposal
- Scale-up of clean fossil fuel combustion
- Biofuel catalyst design

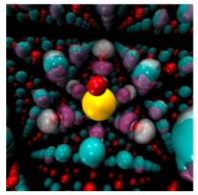
Economic security

- Additive manufacturing of qualifiable metal parts
- Reliable and efficient planning of the power grid
- Seismic hazard risk assessment
- Urban planning



Scientific discovery

- Find, predict, and control materials and properties
- Cosmological probe of the standard model of particle physics
- Validate fundamental laws of nature
- Demystify origin of chemical elements
- Light source-enabled analysis of protein and molecular structure and design
- Whole-device model of magnetically confined fusion plasmas

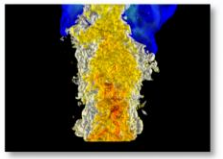


Earth system

- Accurate regional impact assessments in Earth system models
- Stress-resistant crop analysis and catalytic conversion of biomass-derived alcohols
- Metagenomics for analysis of biogeochemical cycles, climate change, environmental remediation

Health care

- Accelerate and translate cancer research



EXASCALE COMPUTING PROJECT

Gordon Bell Award

- Since 1987 the Gordon Bell Prize is awarded at the SC conference to recognize outstanding achievement in high-performance computing.
- The purpose of the award is to track the progress of parallel computing, with emphasis on rewarding innovation in applying HPC to applications.
- Financial support of the $10,000 award is provided by Gordon Bell, a pioneer in high-performance and parallel computing.
- Authors‘ mark their SC paper as a possible Gordon Bell Prize competitor.
- Gordon Bell committee reviews the papers and selects 6 papers for the competition.
- Presentations are made at SC and a winner is chosen.

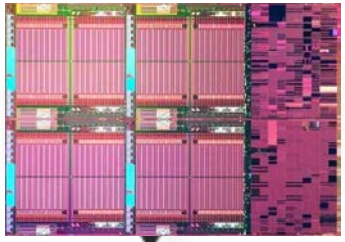
Commodity plus Accelerator Today

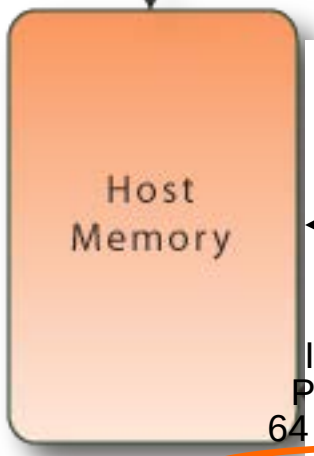
Commodity

Intel Xeon
8 cores
3 GHz
8*4 ops/cycle
96 Gflop/s (DP)

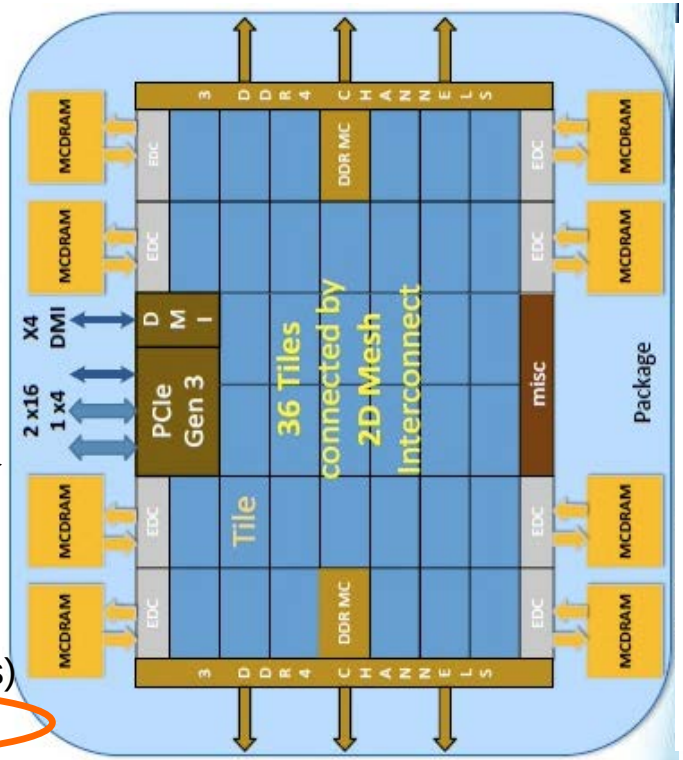
Accelerator/Co-Processor

Intel Xeon Phi (KNL)
72 “cores”
32 flops/cycle/core
1.4 GHz
72*1.4*32 ops/cycle
3.22 Tflop/s (DP) or 6.45 Tflop/s (SP)

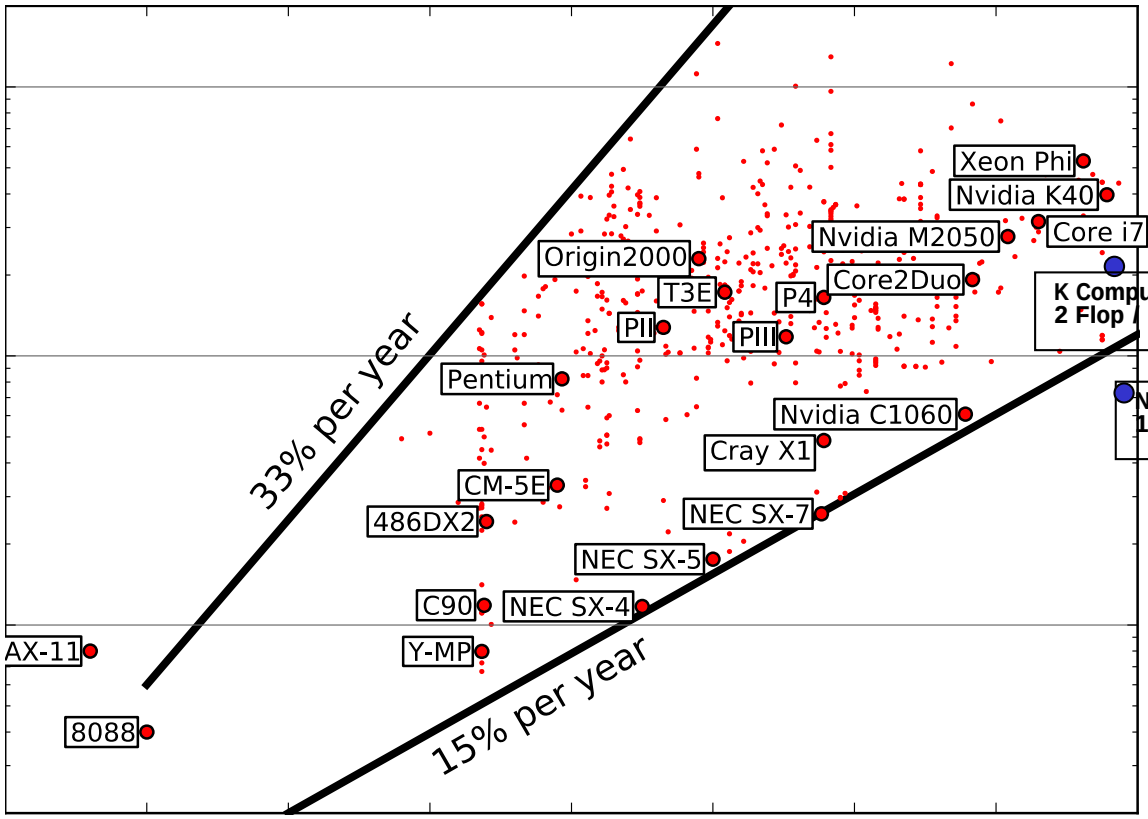




Interconnect
PCI-X 16 lane
64 Gb/s (8 GB/s)
1 GW/s



Ratio of CPU speed to memory bandwidth increases 15–33% yearly



- Flops "free," memory expensive
- Good for dense, BLAS-3 operations (matrix multiply)

- Flops & memory access balanced
- Good for sparse & vector operations

Data from Stream benchmark (McCalpin) and vendor information sheets

Conventional Wisdom is Changing

Old Conventional Wisdom

- **Peak clock frequency** *as primary limiter for performance improvement*
- **Cost:** *FLOPs are biggest cost for system: optimize for compute*
- **Concurrency:** *Modest growth of parallelism by adding nodes*
- **Memory scaling:** *maintain byte per flop capacity and bandwidth*
- **Uniformity:** *Assume uniform system performance*
- **Reliability:** *It's the hardware's problem*

New Conventional Wisdom

- **Power** *is primary design constraint for future HPC system design*
- **Cost:** *Data movement dominates optimize to minimize data movement*
- **Concurrency:** *Exponential growth of parallelism within chips*
- **Memory Scaling:** *Compute growing 2x faster than capacity or bandwidth*
- **Heterogeneity**: *Architectural and performance non-uniformity increase*
- **Reliability***: Cannot count on hardware protection alone*

Exascale System Architecture with a cap of $200M and 20MW

Systems	2019 Sunway Taihu	2019 Summit	2022 (may be 2023)	Difference Today & Exa
System peak	**125.4 Pflop/s**	**143.5 Pflop/s**	**1 Eflop/s**	**~10x**
Power	**15 MW** **(8 Gflops/W)**	**9.7 MW** **(14.9 Gflops/W)**	**~20 MW** **(50 Gflops/W)**	**O(1)** **~6x**
System memory	1.31 PB	2.8 PB	32 - 64 PB	~50x
Node performance	3.06 TF/s	42 TF/s	1.2 or 15TF/s	O(1)
Node concurrency	260 cores	520 cores	O(1k) or 10k	~5x - ~50x
Node Interconnect BW	16 GB/s	1600 GB/s	200-400GB/s	~25x
System size (nodes)	40,960	4,608	O(100,000) or O(1M)	~6x - ~60x
Total concurrency	10.6 M	2.4 M	O(billion)	~100x
MTTF	Few / day	Few / day	Many / day	O(?)

Technology Scaling Trends
Exascale in 2021… and then what?

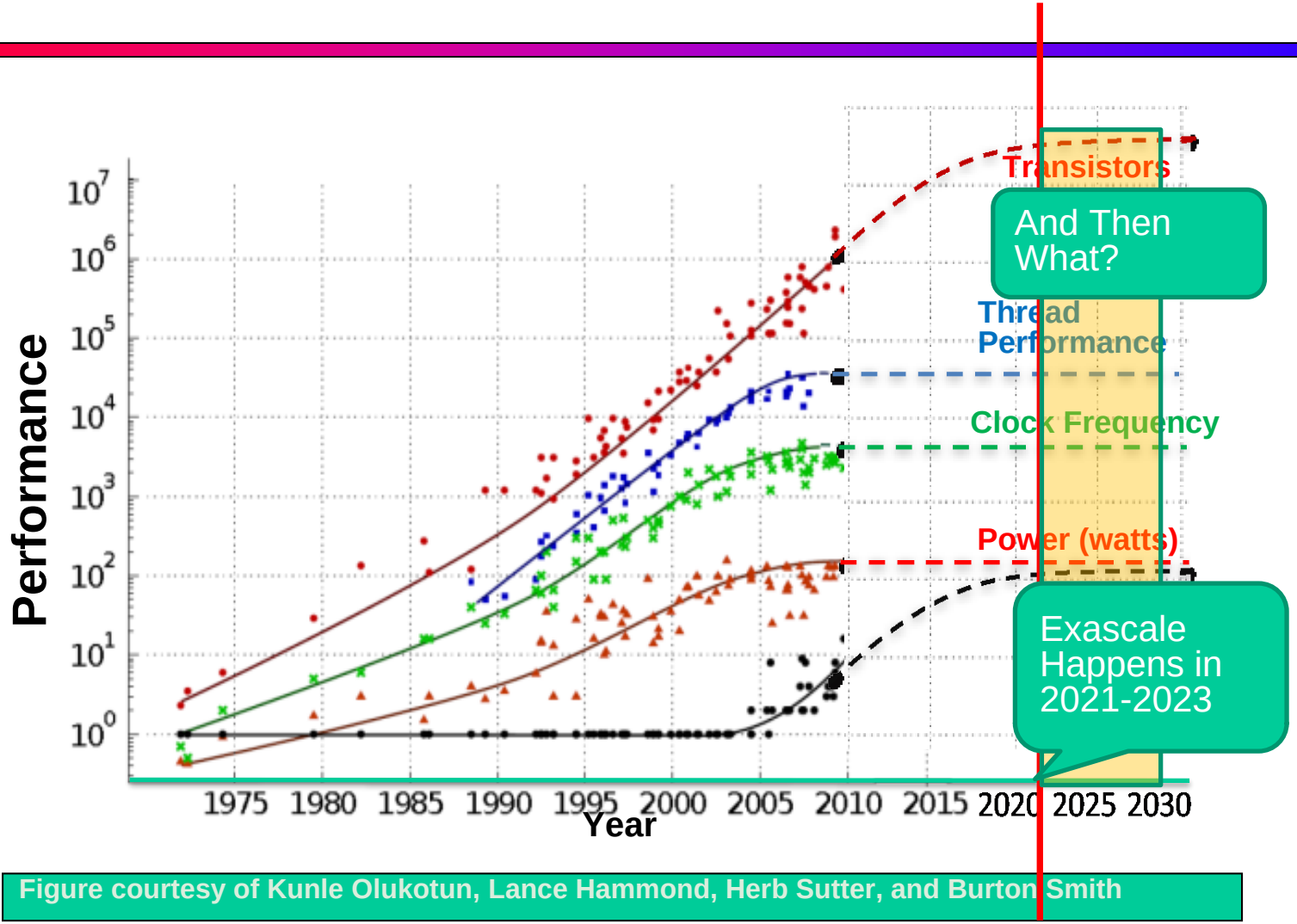
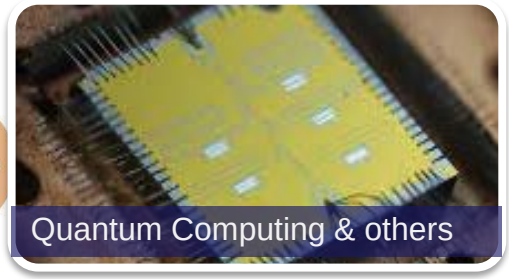
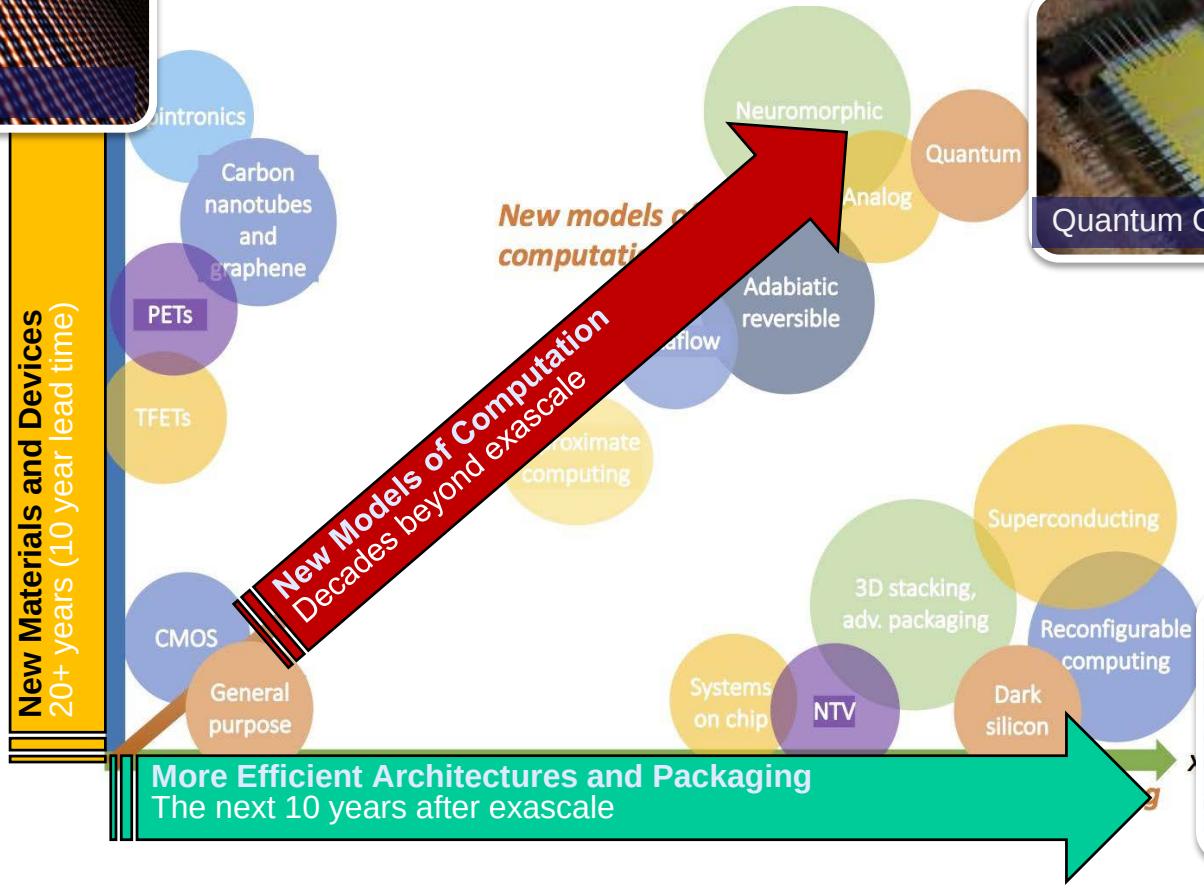


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Numerous Opportunities Exist to Continue Scaling of Computing Performance



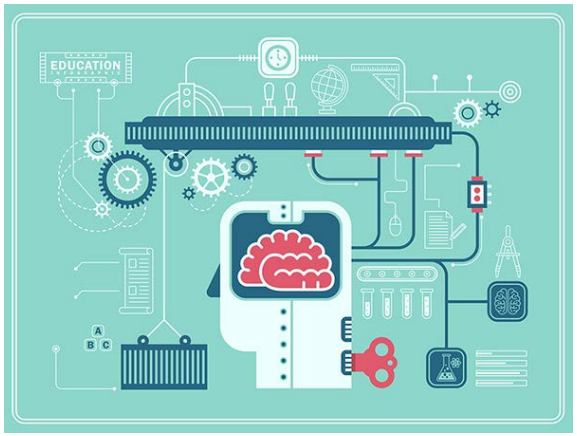


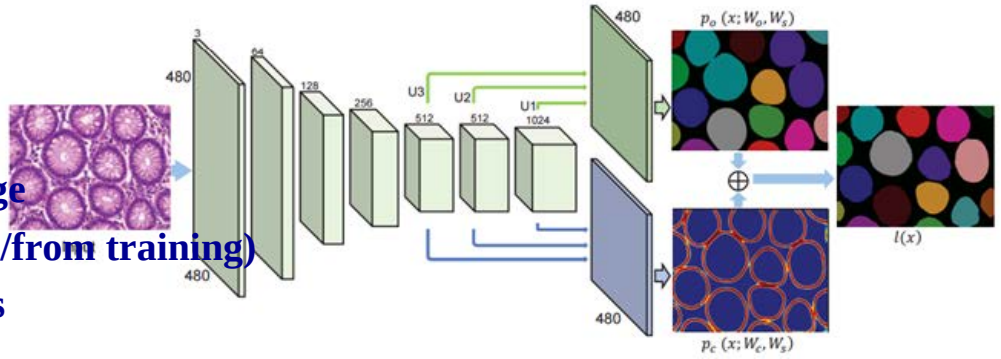
Many unproven candidates yet to be invested at scale. Most are disruptive to our current ecosystem.

Machine Learning in Computational Science

Many fields are beginning to adopt machine learning to augment modeling and simulation methods

- Climate
- Biology
- Drug Design
- Epidemology
- Materials
- Cosmology
- High-Energy Physics

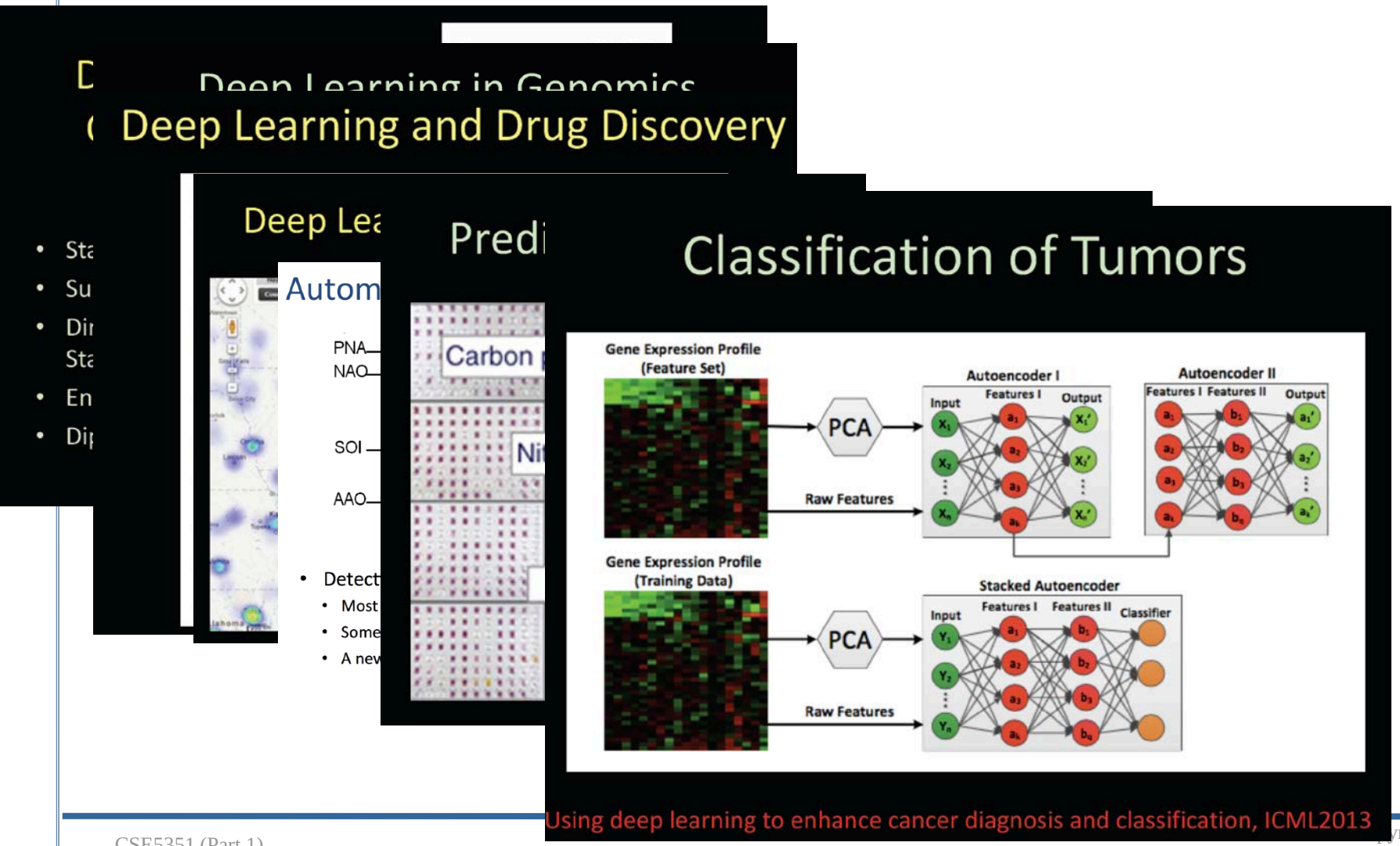




ML is changing Science

- HPC HW&SW must change
- Data will "slosh" (model to/from training)
- Scientists will share models

Some Deep Learning Applications



IN DEPTH | ARTIFICIAL INTELLIGENCE

Science
AAAS

DOE readies multibillion-dollar AI push

Robert F. Service

+ See all authors and affiliations

Science 01 Nov 2019:
Vol. 366, Issue 6465, pp. 559-560
DOI: 10.1126/science.366.6465.559

in a globally crowded field

By **Robert F. Service**, *in Washington, D.C.*

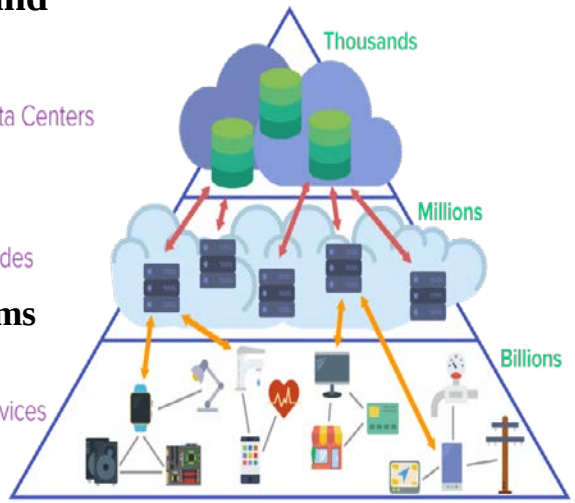
The U.S. Department of Energy (DOE) is planning a major initiative to use artificial intelligence (AI) to speed up scientific discoveries. At a meeting here last week, DOE officials said they will likely ask Congress for between $3 billion and $4 billion over 10 years, roughly the amount the agency is spending to build next-generation "exascale" supercomputers.

In Ten Years…

- **Learned Models Begin to Replace Data**
  - queryable, portable, pluggable, chainable, secure
- **Experimental Discovery Processes Dramatically Refactored**
  - models replace experiments, experiments improve models
- **Many Questions Pursued Semi-Autonomously at Scale**
  - searching for materials, molecules and pathways, new physics
- **Simulation and AI Approaches Merge**
  - deep integration of ML, numerical simulation and UQ
- **Theory Becomes Data for Next Generation AI**
  - AI begins to contribute to advancing theory
- **AI Becomes Common Part of Scientific Laboratory Activities**
  - Infuses scientific, engineering and operations

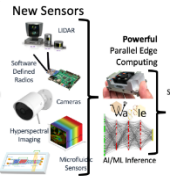
Scientific Computing is Changing

- **In the past, we moved experimental data to the centralized servers, which provided bulk storage and computational resources for analysis and simulation.**
- **Three things have changed:**
  - **CPU advances have enabled edge/IoT devices to be small parallel computers with real operating systems and multithreaded programming models (CUDA, OpenMP, TensorFlow, etc.).**
  - **Machine learning and AI has helped create a new class of algorithms that can sift through massive amounts of experimental data, and then pushing to the data center only the relevant results**
  - **Edge devices and scientific instruments are rapidly expanding, creating a new class of "edge software defined instrument" that must connect to cyberinfrastructure.**



- **These three changes are forcing us to rethink the central services model of HPC and embrace a new model where network infrastructure computes-along-the-way.**
- **To realize that goal, we need a new conceptual model for programming this new end-to-end infrastructure.**

IoT and Supercomputing



IoT/Edge → HPC/Cloud

Size	Nano	Micro	Milli	Server	Fog	Campus	Facility
Example	IoT	Smart Device	Sage Node	Linux Box	Co-located Blades	1000-node cluster	Datacenter
Memory	0.5K	256K	8GB	32GB	256G	32TB	16PB
Network	BLE	WiFi/LTE	WiFi/LTE	1 GigE	10GigE	40GigE	N*100GigE
Cost	$5	$30	$600	$3K	$50K	$2M	$1000M

Count = 10^9
Size = 10^1

Count = 10^1
Size = 10^9

- The number of network-connected devices—
  - sensors, actuators, instruments, computers, and data stores
- Now substantially exceeds the number of humans on this planet

- **Machine learning in the application.**
  - **for enhanced scientific discovery**
- **Machine learning in the computational infrastructure.**
  - **for improved performance**
- **Machine learning at the edge.**
  - **for managing data volume**

The Take Away

- **Three computer revolutions**
  - **High performance computing**
  - **Deep learning**
  - **Edge & AI**
- **The very small (edge/fog computing and sensors)**
- **The very large (clouds, exascale, and big data)**

- **Technical implications**
  - **Fluid end-to-end cyberinfrastructure**
  - **Interdisciplinary data and infrastructure planning**
- **Cultural implications**
  - **Change management and strategic planning**
  - **Community collaboration**

Harnessing the computing continuum will catalyze new consumer services, business processes, social services, and scientific discover

The Computing *Continuum*

IoT/Edge — Fog — HPC/Cloud/Instrument

Size	Nano	Micro	Milli	Server	Fog	Campus	Facility
Example	Adafruit Trinket	Particle.io Boron	Array of Things	Linux Box	Co-located Blades	1000-node cluster	Datacenter
Memory	0.5K	256K	8GB	32GB	256G	32TB	16PB
Network	BLE	WiFi/LTE	WiFi/LTE	1 GigE	10GigE	40GigE	N*100GigE
Cost	$5	$30	$600	$3K	$50K	$2M	$1000M

Count = 10^9
Size = 10^1

Count = 10^1
Size = 10^9

The number of network-connected devices—
sensors, actuators, instruments, computers, and data stores—
Now substantially exceeds the number of humans on this planet

We lack a programming and execution model that is inclusive and capable of harnessing the entire computing continuum to program our new intelligent world.

Conclusions

- **Exciting time for HPC**
- **For the last decade or more, the research investment strategy has been overwhelmingly biased in favor of hardware.**
- **This strategy needs to be rebalanced - barriers to progress are increasingly on the software side.**
- **High Performance Ecosystem out of balance**
  - **Hardware, OS, Compilers, Software, Algorithms, Applications**