

大型計算機センター法制化50周年記念シンポジウム

High Performance Computing and Big Data: Challenges for the Future

Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory
University of Manchester

7/10/19

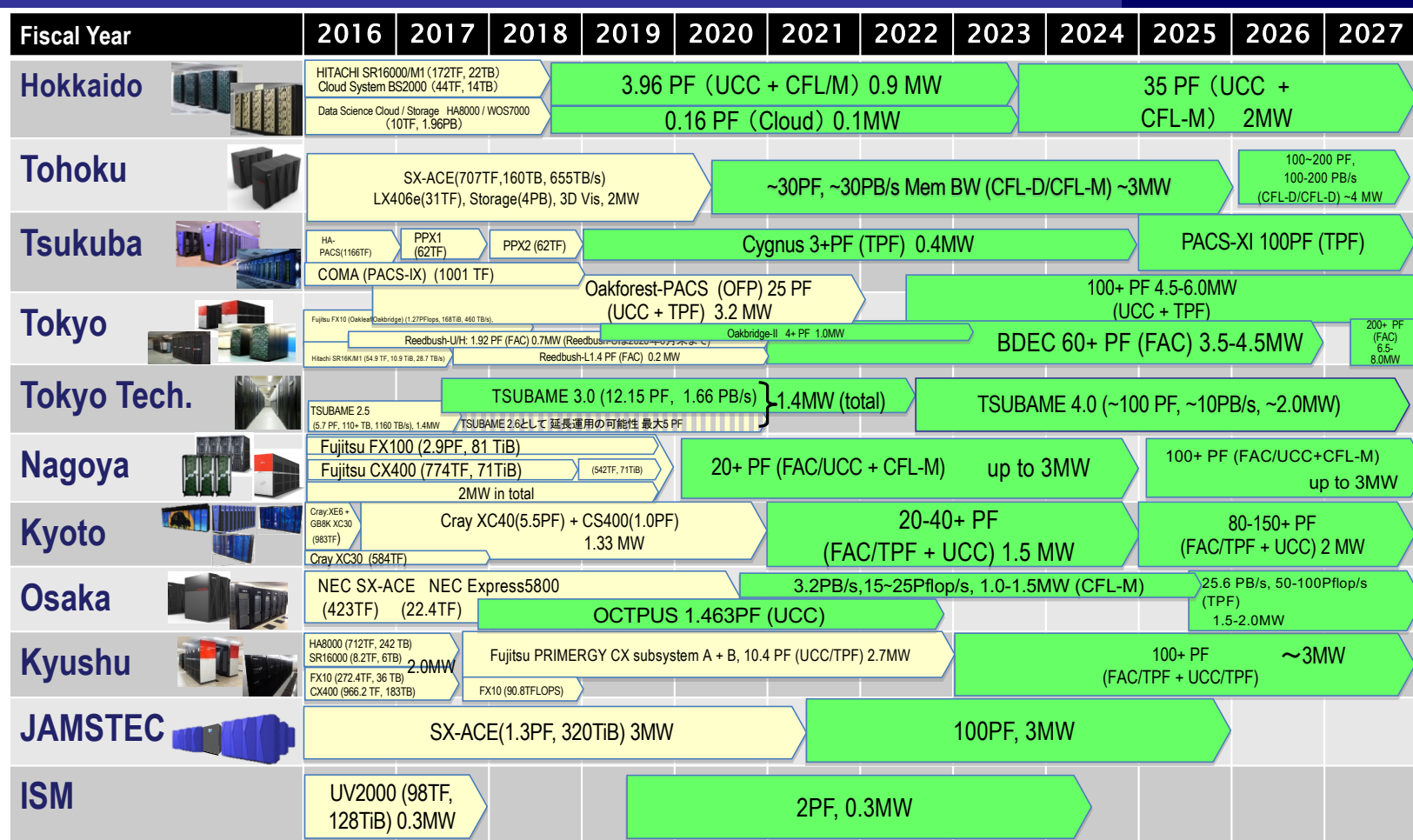
1



Outline

- Overview of High Performance Computing
- Look at issues on convergence of HPC and Big Data
- Killer apps for HPC, BD, and Machine Learning.

HPCI Tier 2 Systems Roadmap (As of Nov. 2018)



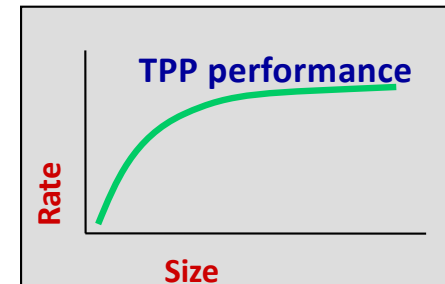
Power is the maximum consumption including cooling

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**



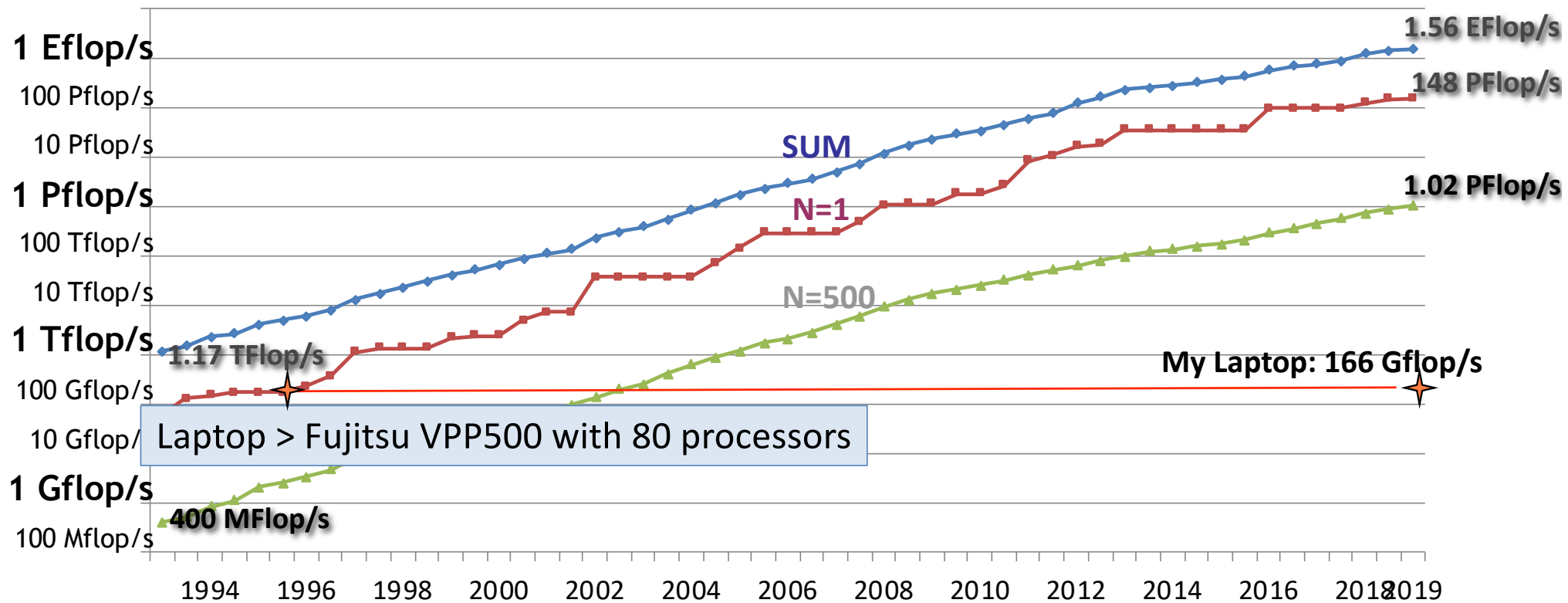
State of Supercomputing in 2019

- 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) (133 systems)
 - Special purpose lightweight cores (e.g. IBM BG, Knights Landing, TaihuLight, ARM (1 system))
- Interest in supercomputing is now worldwide, and growing in many new markets (~50% of Top500 computers are in industry).
- Intel processors largest share, 96% followed by AMD, .6%.
- Exascale (10^{18} Flop/s) projects exist in many countries and regions.

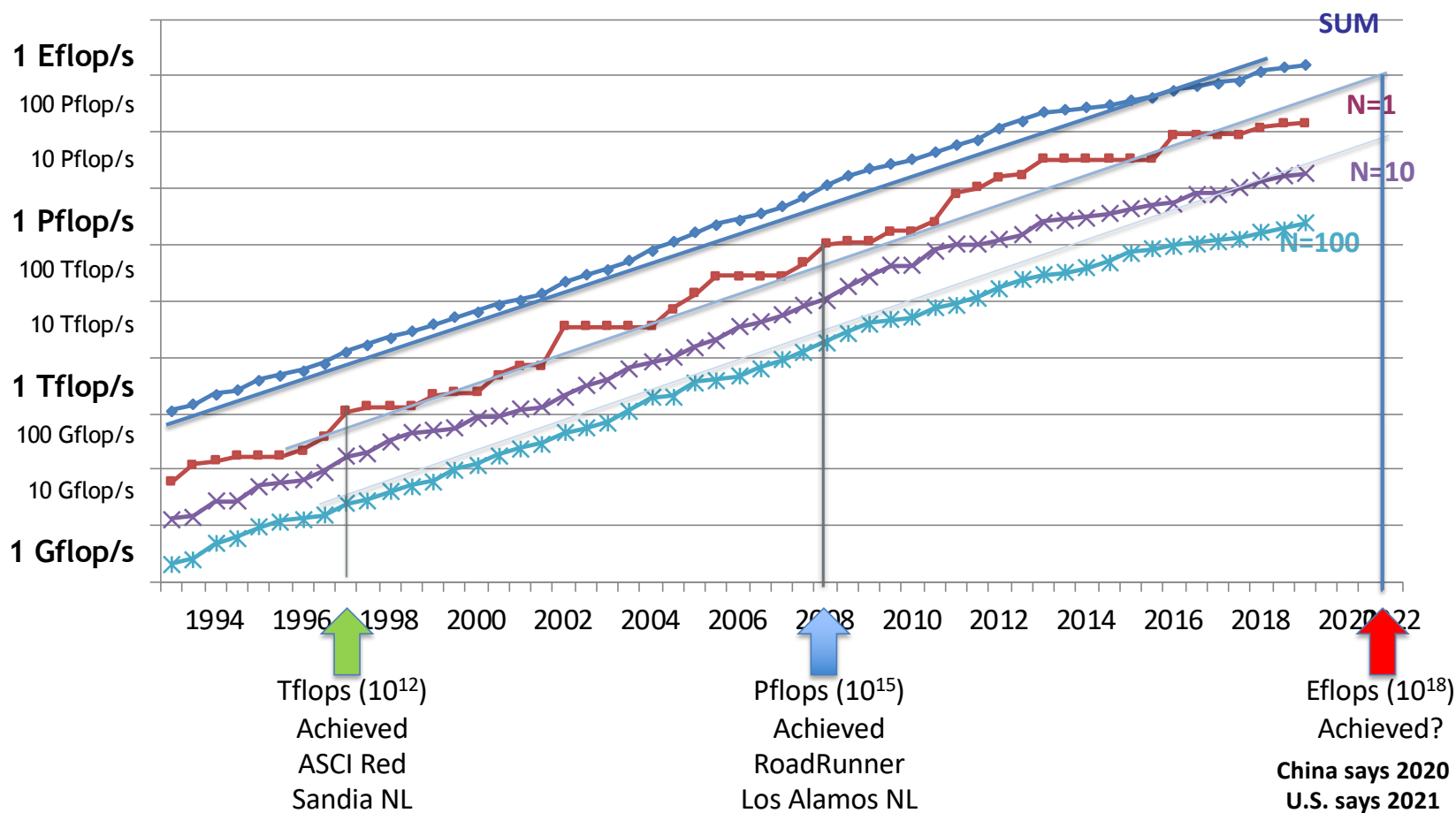
June 2019: The TOP 10 Systems (1/3 of the Total Performance)

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	GFlops/ Watt
1	DOE / OS Oak Ridge Nat Lab	Summit, IBM Power 9 (22C, 3.0GHz), <i>Nvidia GV100 (80C)</i> , Mellanox EDR		2,397,824	149.	74	11.1	14.7
2	DOE / NNSA Livermore Nat Lab	Sierra, IBM Power 9 (22C, 3.1GHz), <i>Nvidia GV100 (80C)</i> , Mellanox EDR		1,572,480	94.6	75	7.44	12.7
3	National Super Computer Center in Wuxi	Sunway TaihuLight, SW26010 (260C) + Custom		10,649,000	93.0	74	15.4	6.05
4	National Super Computer Center in Guangzhou	Tianhe-2A NUDT, Xeon (12C) + <i>MATRIX-2000</i> + Custom		4,981,760	61.4	61	18.5	3.32
5	Texas Advanced Computing Center / U of Texas	Frontera, Dell C6420, Xeon Platinum, 8280 28C 2.7 GHz, Mellanox HDR		448,448	23.5	61		
6	Swiss CSCS	Piz Daint, Cray XC50, Xeon (12C) + <i>Nvidia P100 (56C)</i> + Custom		387,872	21.2	78	2.38	8.90
7	DOE / NNSA Los Alamos & Sandia	Trinity, Cray XC40, Xeon Phi (68C) + Custom		979,968	20.2	49	7.58	2.66
8	Nat Inst of Advanced Indust Sci & Tech	AI Bridging Cloud Infrast (ABCI) Fujitsu Xeon (20C, 22.4GHz) <i>Nvidia V100 (80C)</i> IB-EDR		391,680	16.9	61	1.65	12.05
9	Leibniz Rechenzentrum	SuperMUC-NG, Lenovo, ThinkSystem SD530, Xeon Platinum 8174 24C 3.1GHz, Intel Omni-Path		311,040	19.5	72		
10	DOE / NNSA Livermore Nat Lab	Lassen, IBM Power System p9 22C 3.1 GHz, Mellanox EDR, <i>Nvidia V100 (80C)</i>		288,288	18.2	79		

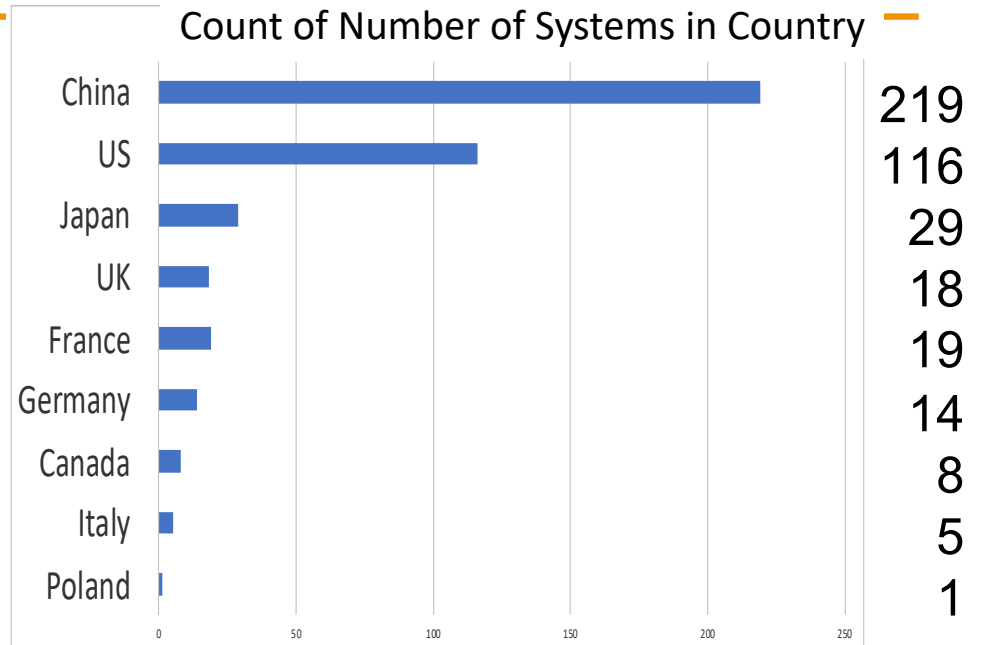
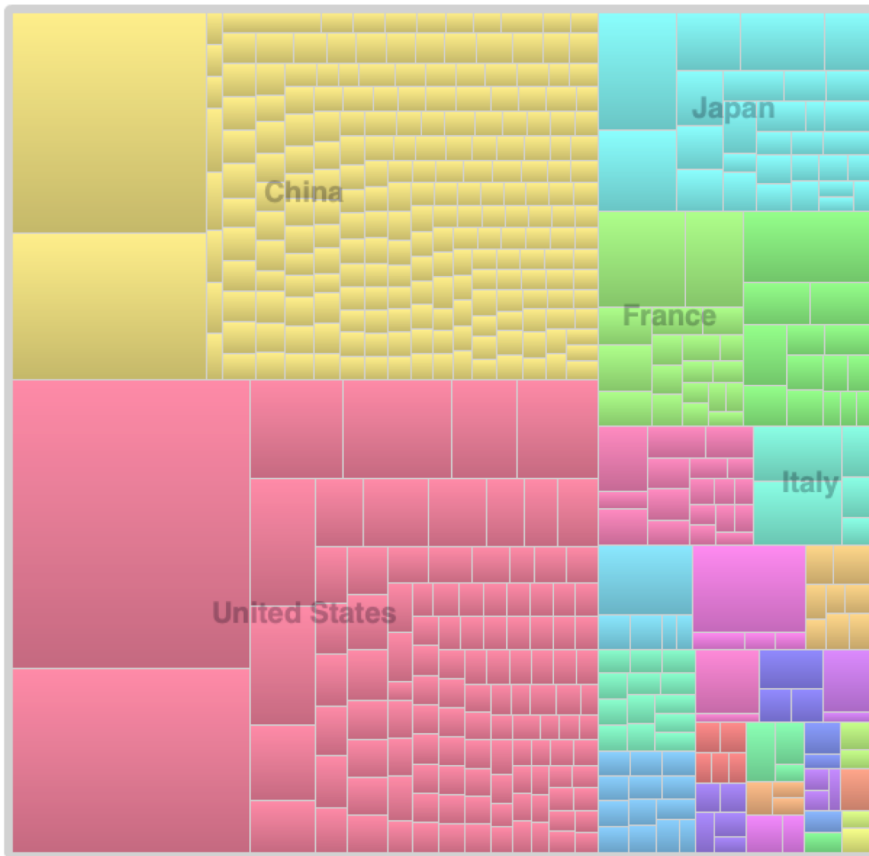
PERFORMANCE DEVELOPMENT OF HPC OVER THE LAST 26 YEARS FROM THE TOP500



PERFORMANCE DEVELOPMENT



COUNTRIES SHARE



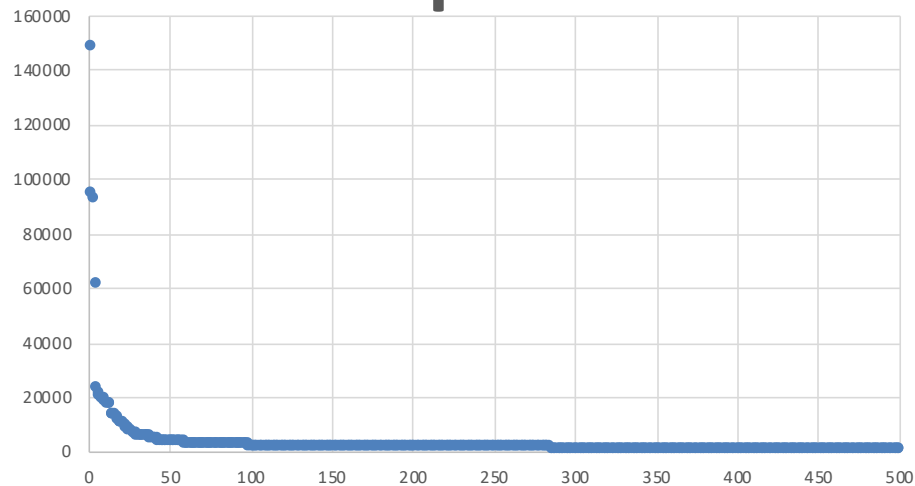
China has 42% of the systems
 US has 23% of the systems
 Japan has 5.8% of the systems

In terms of performance:
 US has 38%
 China has 30%
 Japan has 7.5%

Performance Distribution

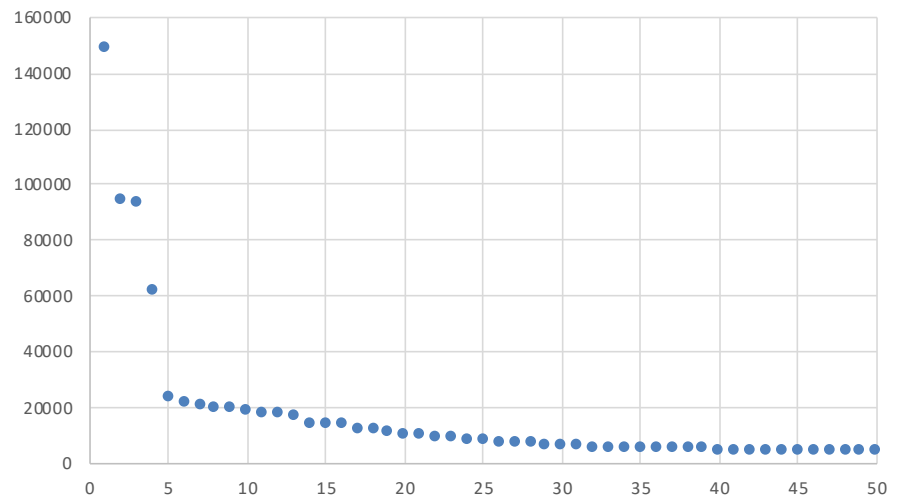


Top500



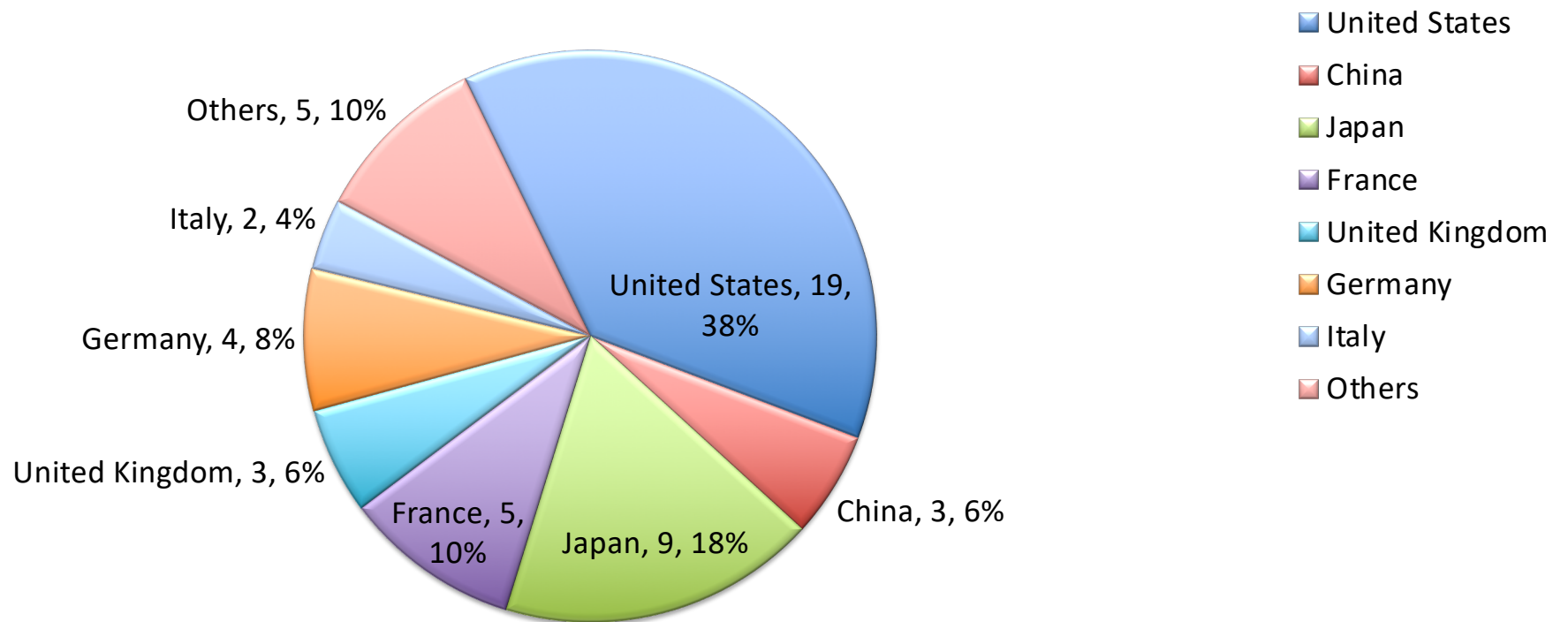
For all 500 systems
 $T_{\text{peak}_{500}} = 1.022 \text{ Pflop/s}$

Top50

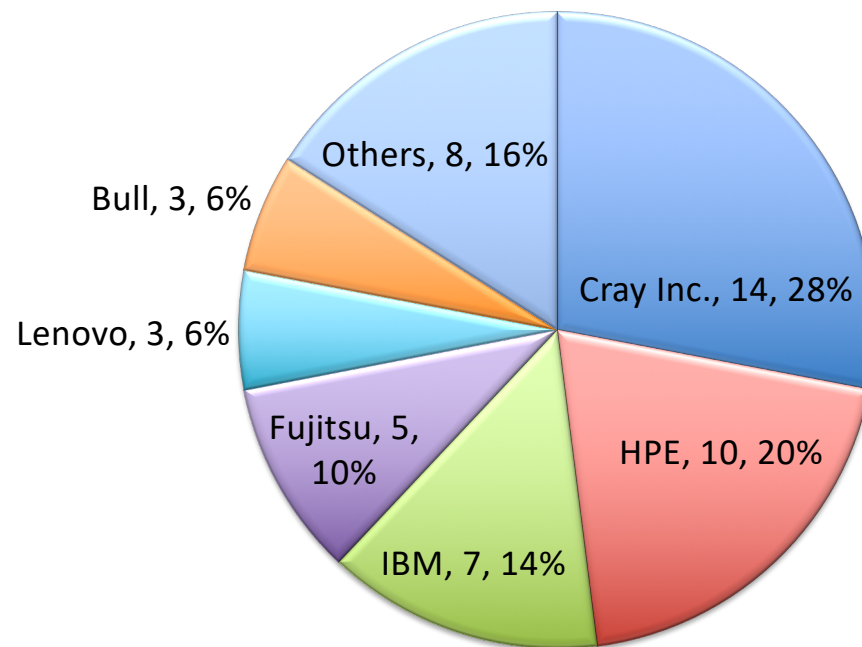


For top 50 systems
 $T_{\text{peak}_{50}} = 3.944 \text{ Pflop/s}$

COUNTRIES (TOP50) / SYSTEM SHARE

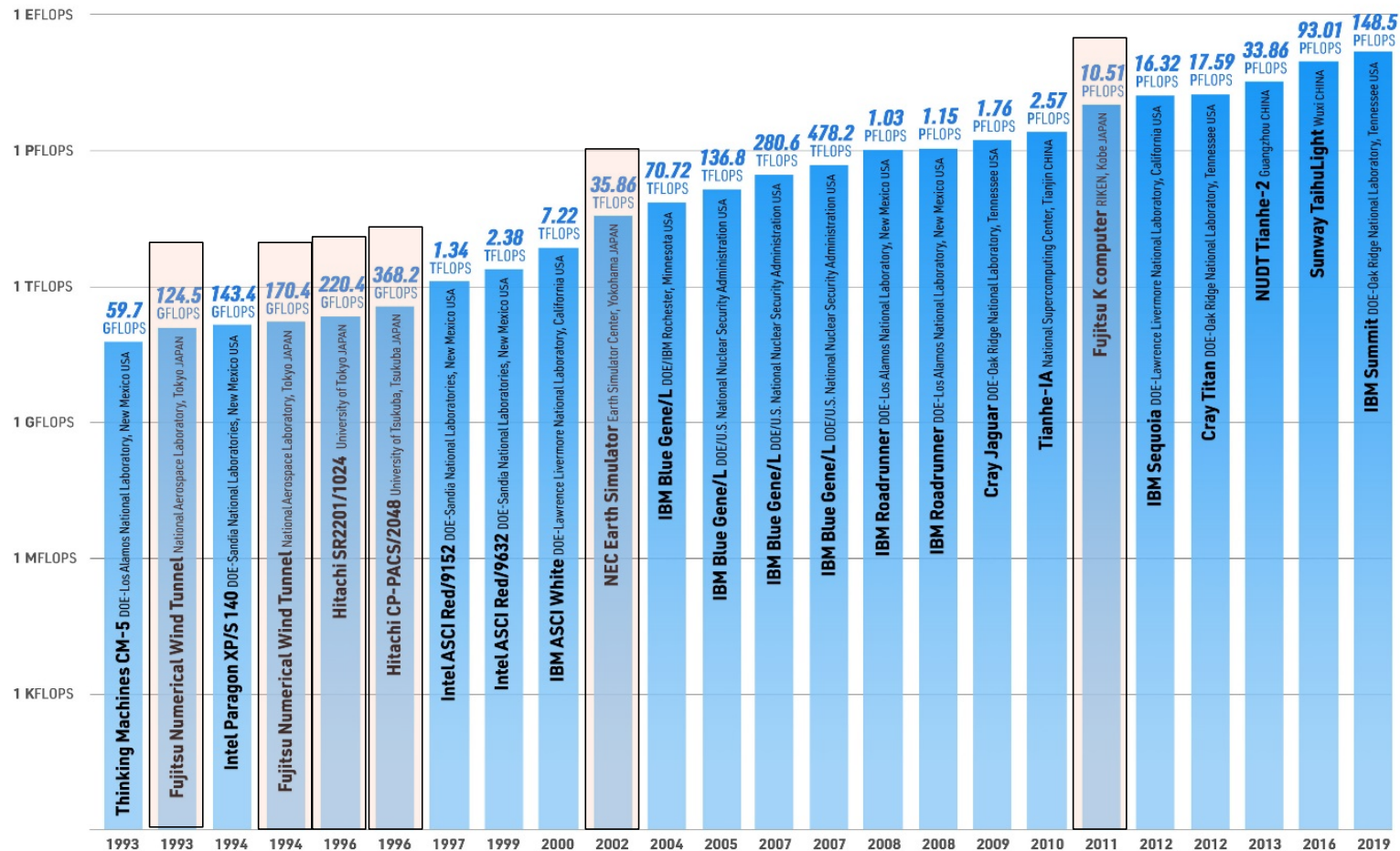


VENDORS (TOP50) / SYSTEM SHARE



- Cray Inc.
- HPE
- IBM
- Fujitsu
- Lenovo
- Bull
- Others

24 #1 Systems for the Top500



Current #1 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 \$15K each**
- Dual-rail Mellanox EDR InfiniBand network
- 250 PB IBM Spectrum Scale file system transferring data at 2.5 TB/s

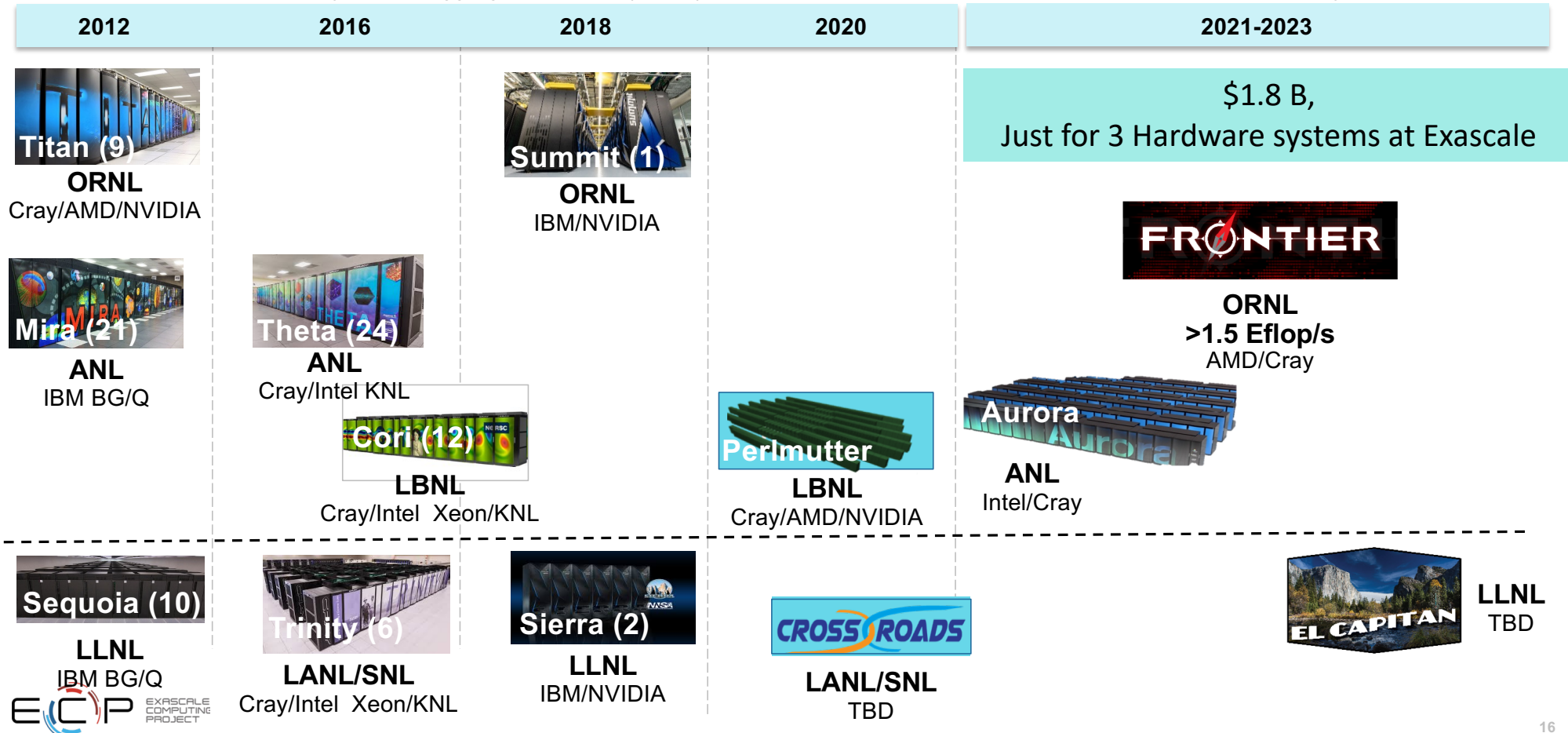


Department of Energy (DOE) Roadmap to Exascale Systems

An impressive, productive lineup of *accelerated node* systems supporting DOE's mission

Pre-Exascale Systems [Aggregate Linpack (Rmax) = 323 PF]

First U.S. Exascale Systems



As scientific research increasingly depends on both high-speed computing and data analytics, the potential interoperability and scaling convergence of these two ecosystems is crucial to the future.

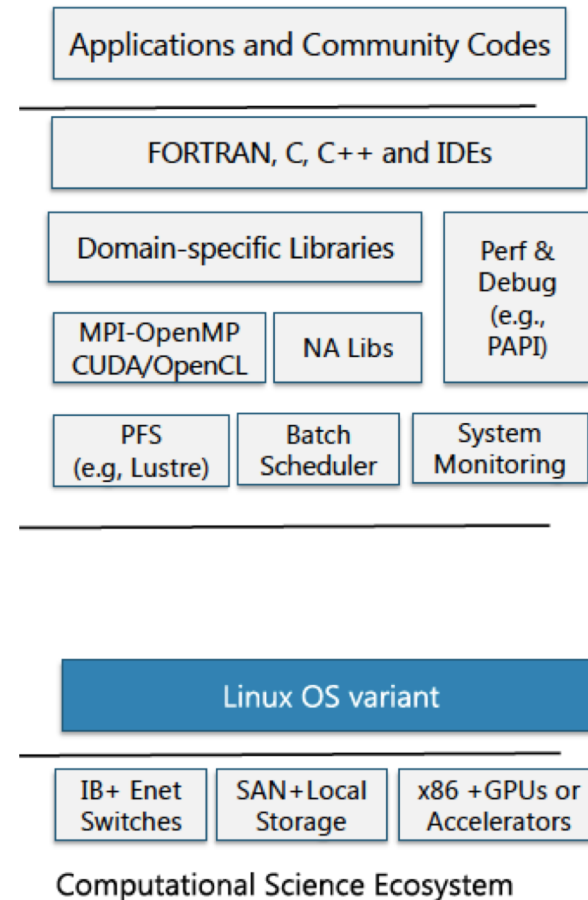
D. Reed & J. Dongarra, CACM 2015

DOI:10.1145/2699414

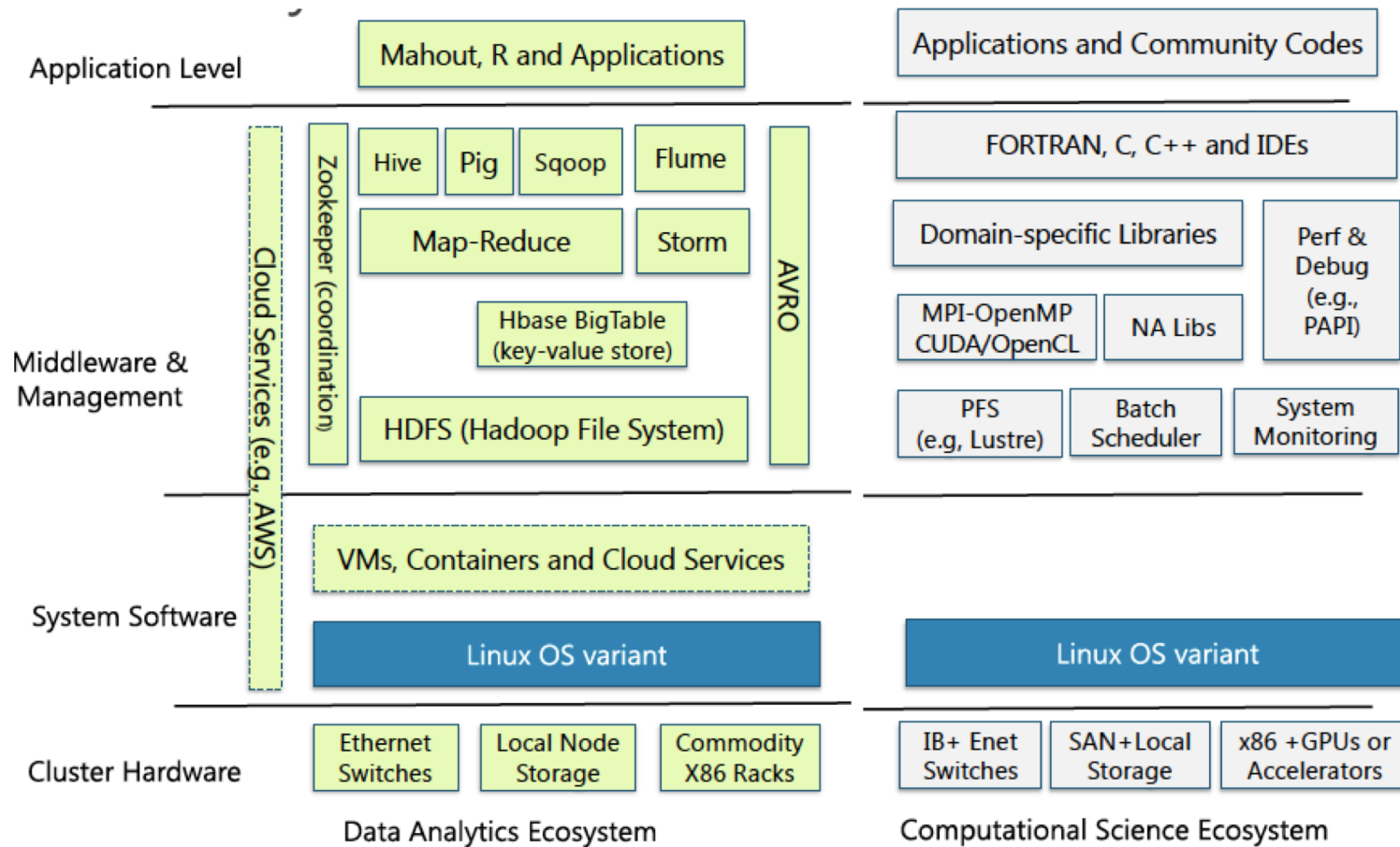
Scientific discovery and engineering innovation requires unifying traditionally separated high-performance computing and big data analytics.

BY DANIEL A. REED AND JACK DONGARRA

Exascale Computing and Big Data



As scientific research increasingly depends on both high-speed computing and data analytics, the potential interoperability and scaling convergence of these two ecosystems is crucial to the future.





Its All About Convergence

- High-end data analytics (big data) and HPC are both essential elements of an integrated computing research-and-development agenda; neither should be sacrificed or minimized to advance the other.
- Programming models and tools are perhaps the biggest point of divergence between the scientific-computing and big-data ecosystems.



Why Converge? (or Mostly Converge)

Independent paths: More Cost, Less Science,

- \$ multiple hardware software infrastructures
- \$ developing software for two communities
- \$ learning two computing models
- \$ smaller discovery community, fewer ideas
- Less science



Comparing Architecture

- **Scientific HP Computing**

- *Significant Cost* in memory and interconnect bandwidth
- *Significant Cost* in resilience hardware to reduce whole-system MTTI

- **Scientific Big Data**

- *? Cost* in memory and interconnect bandwidth
- *Little Cost* for hardware to support system-wide resilience



Comparing Operations

- **Scientific HP Computing**

- *New tightly integrated system purchased every 4 years*
- *Users charged for CPU hours, storage and networking is free*
- *Periodic access to compute resources via job submitted to scheduler and queue*
- *Space-shared compute resources for exclusive access during jobs*

- **Scientific Big Data**

- *New hardware capacity **purchased incrementally***
- *Users charged for all resources (storage, cpu, networking)*
- *Continuous access to long-lived “services” created by science community*
- *Time-shared access to elastic resources*



Comparing Data

- **Scientific HP Computing**

- Inputs *arrive infrequently*, buffering carefully managed
- Data often *reproducible* (repeat simulation)
- Data generated from simulation (*error: from simulation*)
- Data rate *limited by platform*

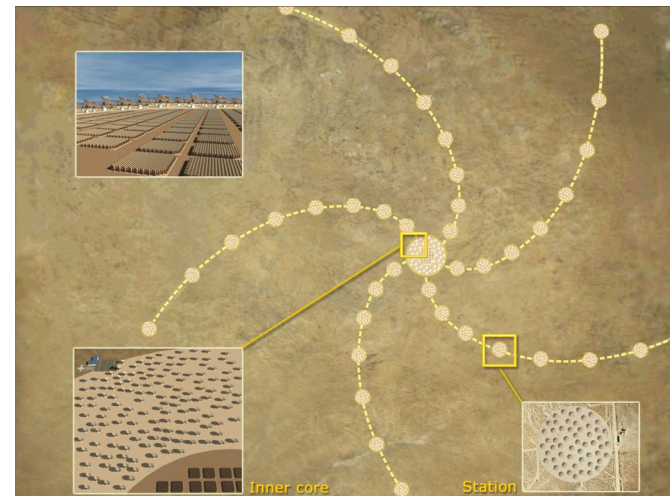
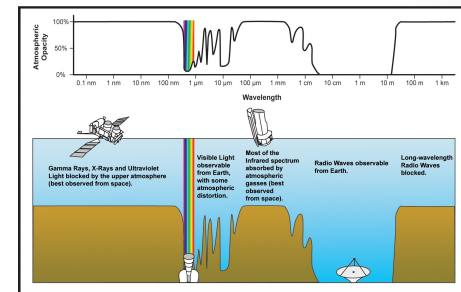
- **Scientific Big Data**

- Inputs *arrive continuously*, streaming workflows
- Data is *unrepeatable* snapshot in time
- Data generated by sensors (*error: from measurement*)
- Data rate *limited by sensors*

Square Kilometre Array



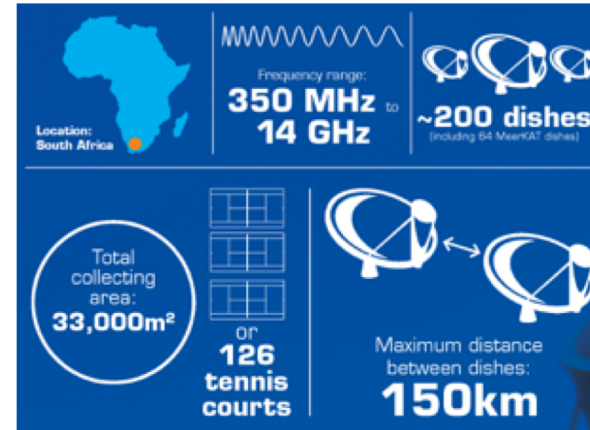
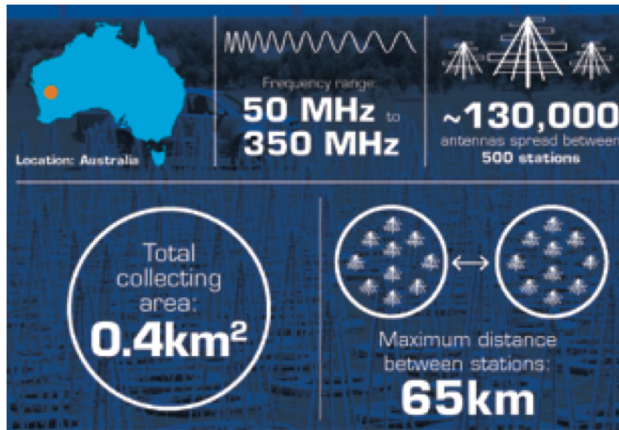
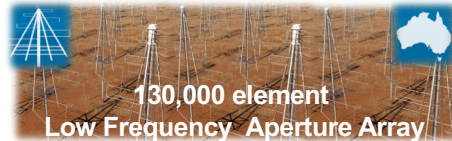
- Next Generation radio telescope – compared to best current instruments it is ...
 - ~100 times sensitivity
 - $\sim 10^6$ times faster imaging the sky
 - More than 1 square km of collecting area on baseline 150km
- Will address some of the key problems of astrophysics and cosmology (and physics)
- Builds on techniques developed in Europe
 - It is an interferometer
- Uses innovative technologies...
 - Major data/compute project
 - Need performance at low unit cost



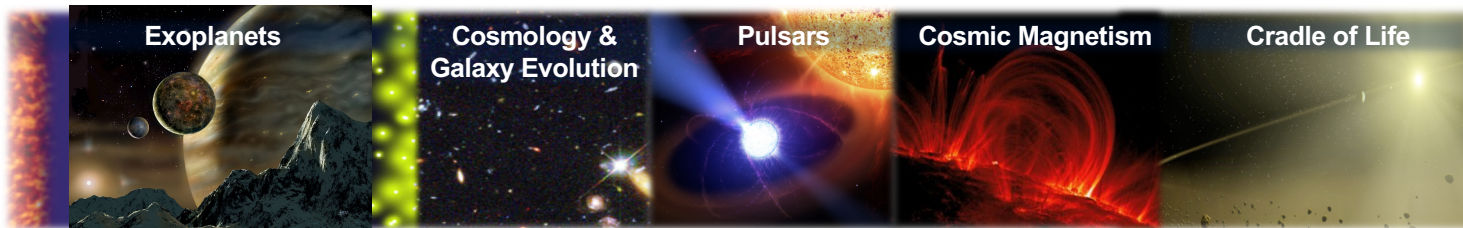
Exploring the Universe with the World's Largest Radio Telescope



Phase I : 2020



Science



50 MHz

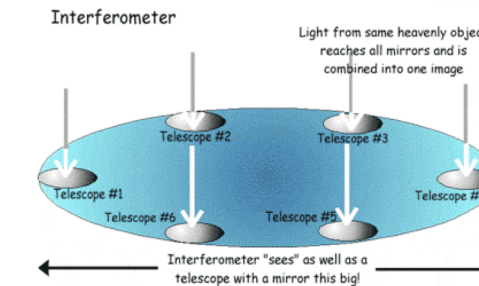
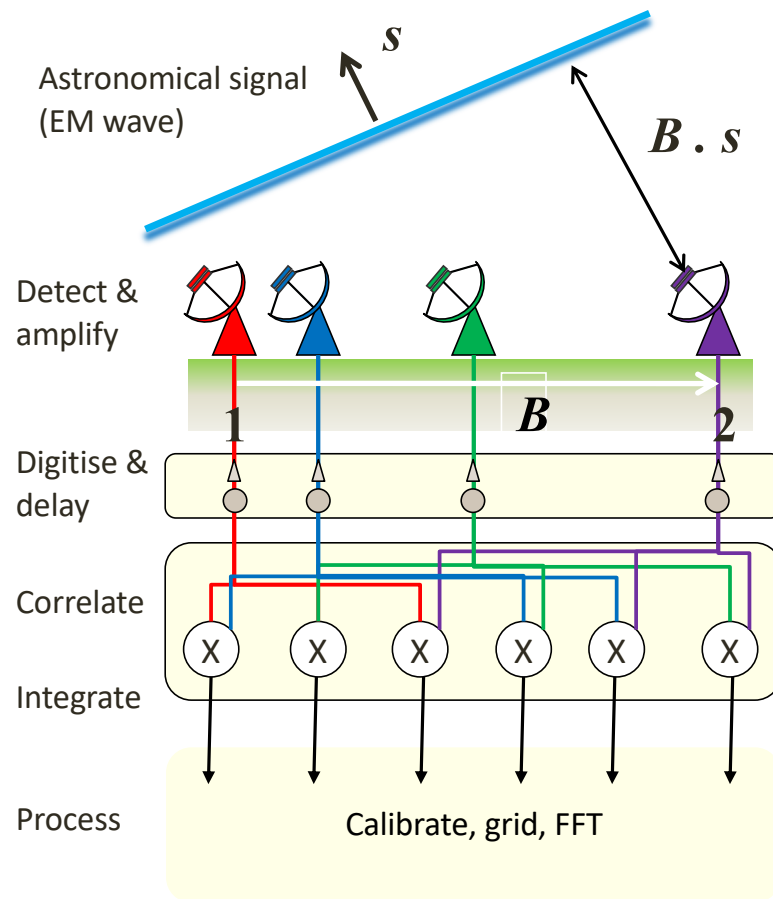
100 MHz

1 GHz

10 GHz

Exploring the

Standard Interferometer



Visibility:

$$V(B) = E_1 E_2^*$$

$$= I(s) \exp(i \omega B \cdot s / c)$$

- Resolution determined by maximum baseline

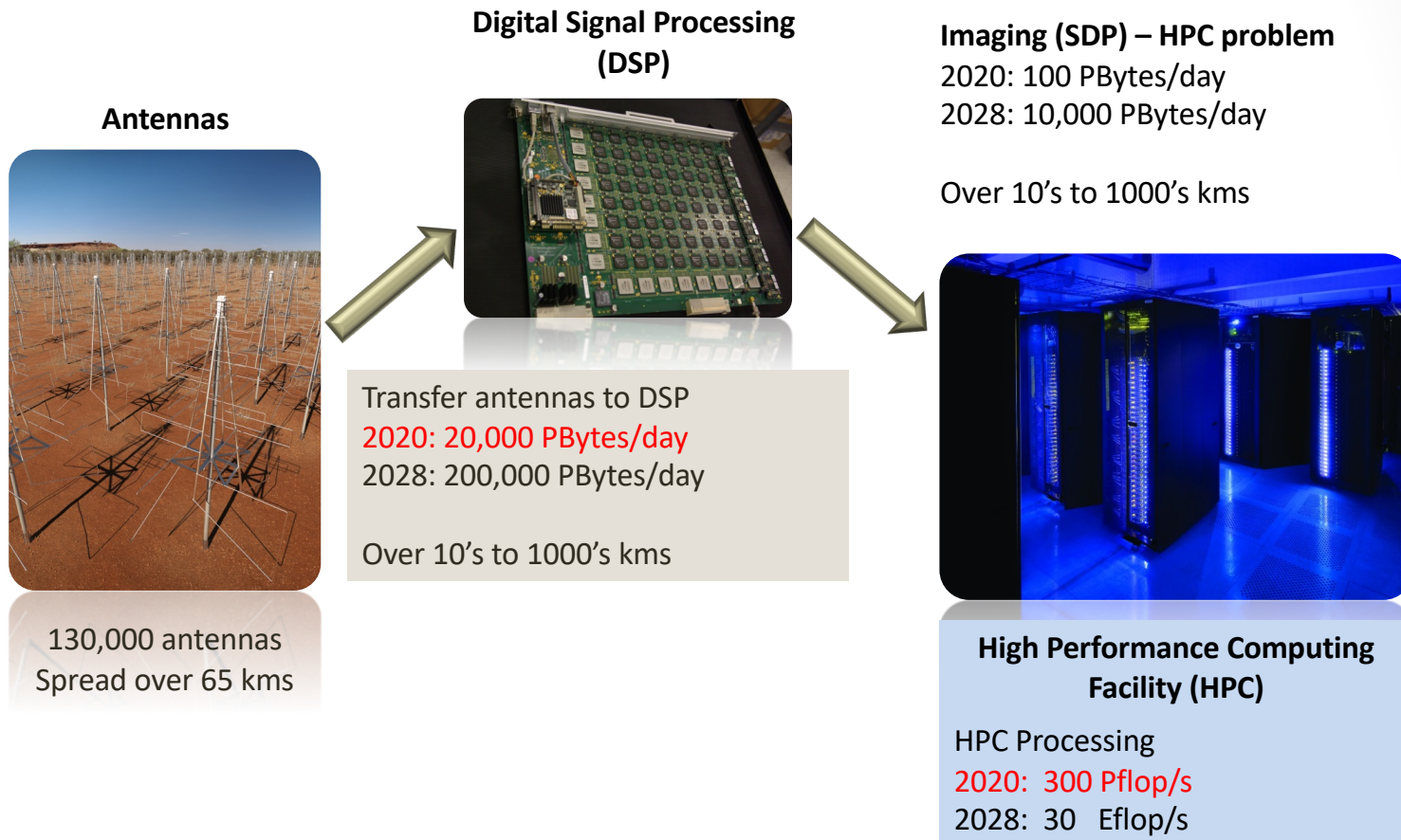
$$\theta_{\max} \sim \lambda / B_{\max}$$

- Field of View (FoV) determined by the size of each dish

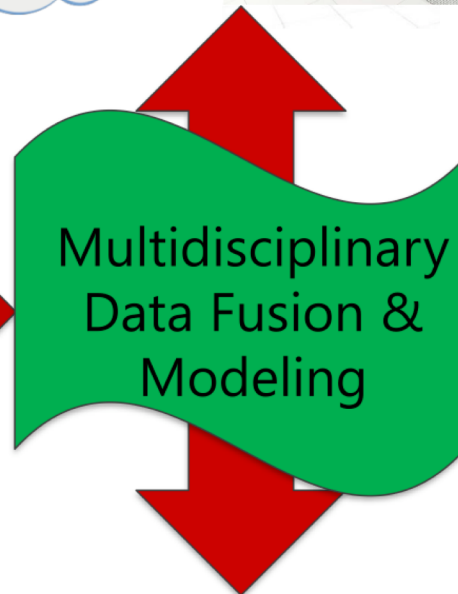
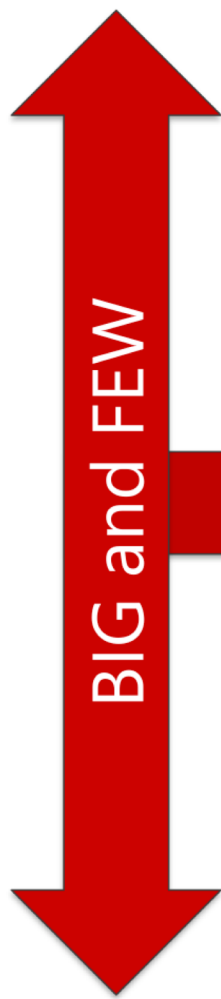
$$\theta_{\text{dish}} \sim \lambda / D$$

SKY Image

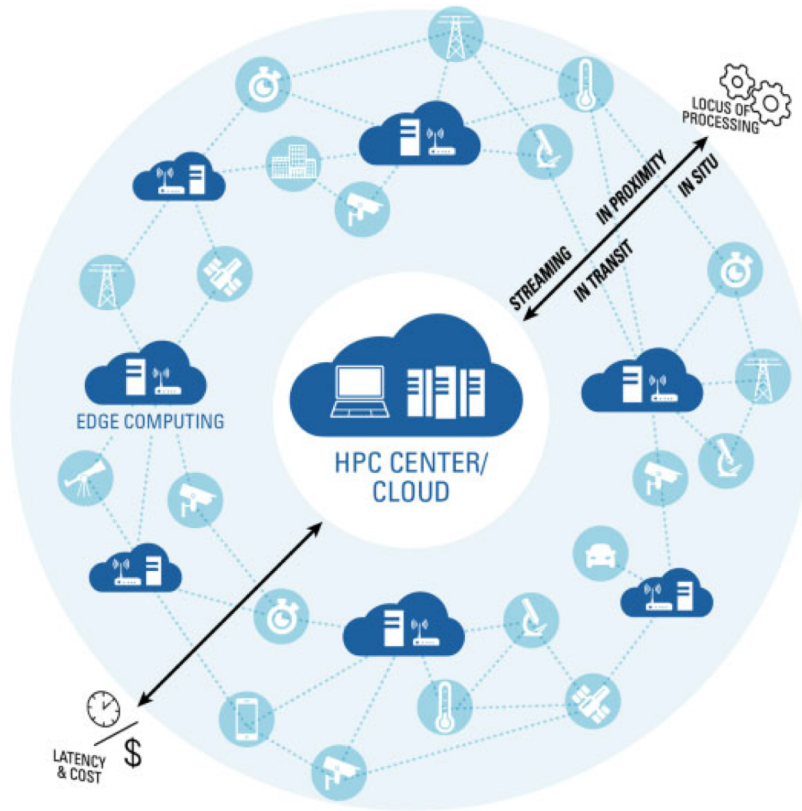
SKA: A Leading Big Data Challenge for 2020



Science Instrument Continuum

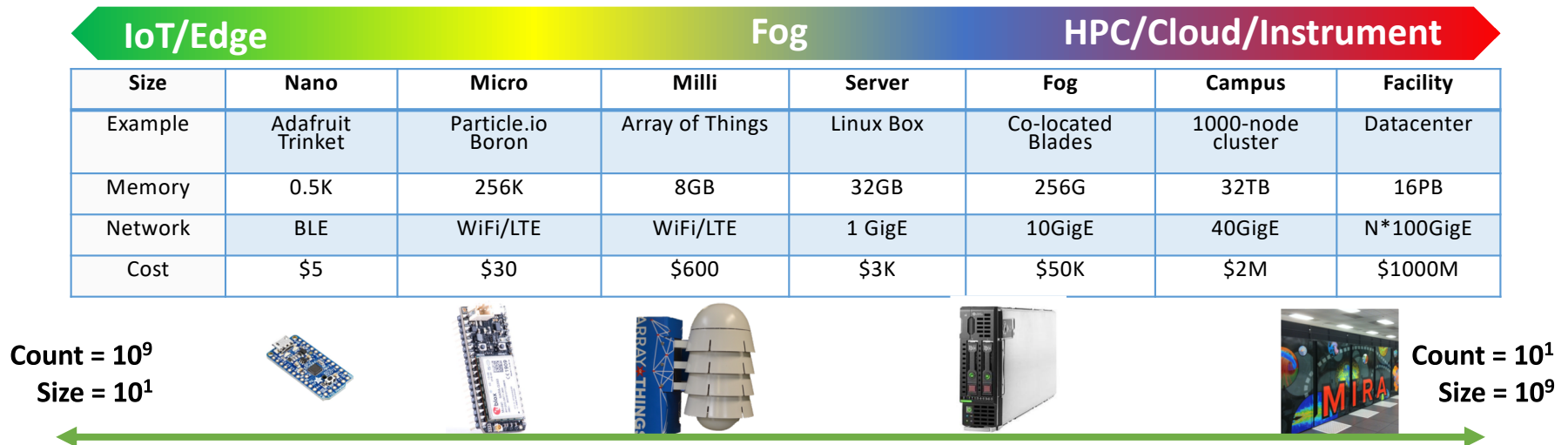


HPC Computing



- The highest concentrations of computing power and storage are in the “center”.
- Much of the rapid increase in data volumes and the dramatic proliferation of data generators is occurring in the edge, where the processing and storage infrastructure needed to cope with this rising flood of data is ad hoc and under provisioned at best.

The Computing Continuum



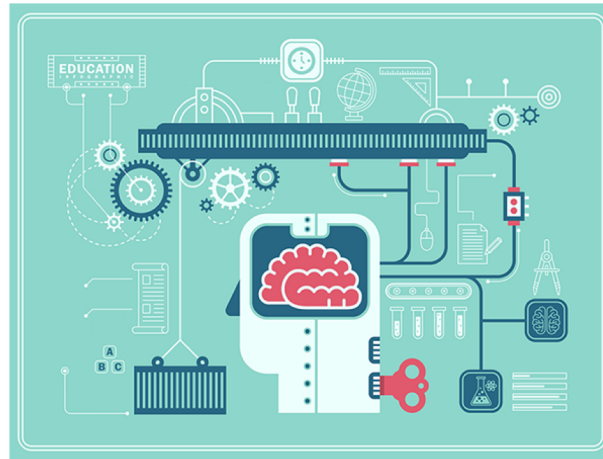
The number of network-connected devices—
sensors, actuators, instruments, computers, and data stores—
Now substantially exceeds the number of humans on this planet

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

Machine Learning in Computational Science

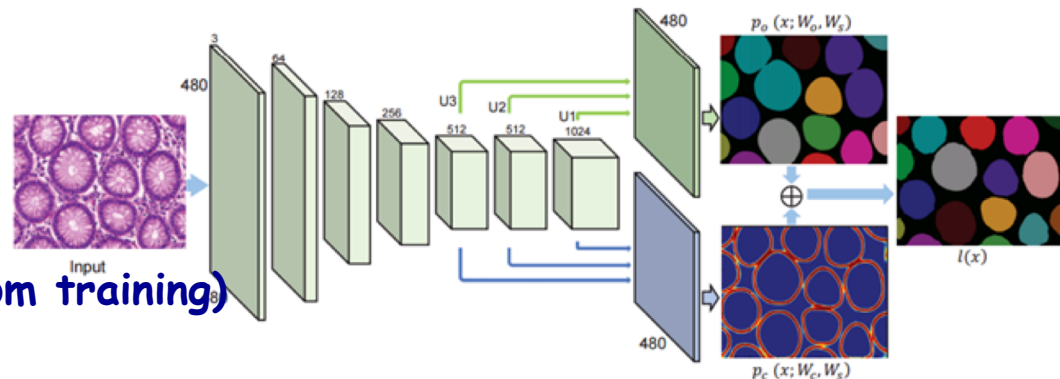
Many fields are beginning to adopt machine learning to augment modeling and simulation methods

- Climate
- Biology
- Drug Design
- Epidemiology
- 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

Deep Learning in Genomics (Deep Learning and Drug Discovery)

- Sta
- Su
- Dir
- Sta
- En
- Dig

Deep Learning



Automated

PNA
NAO

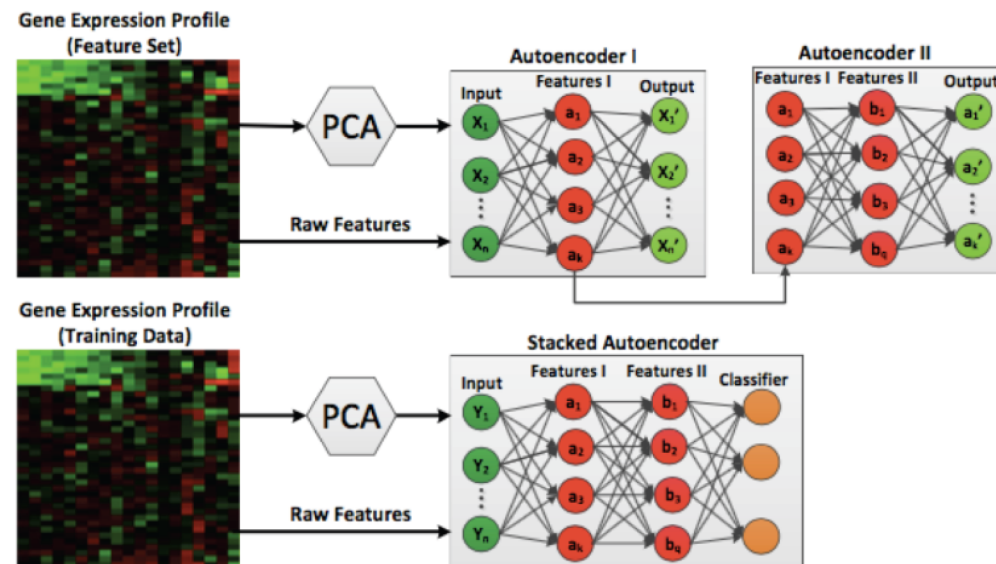
SOI
AAO



- Detect
- Most
- Some
- A new

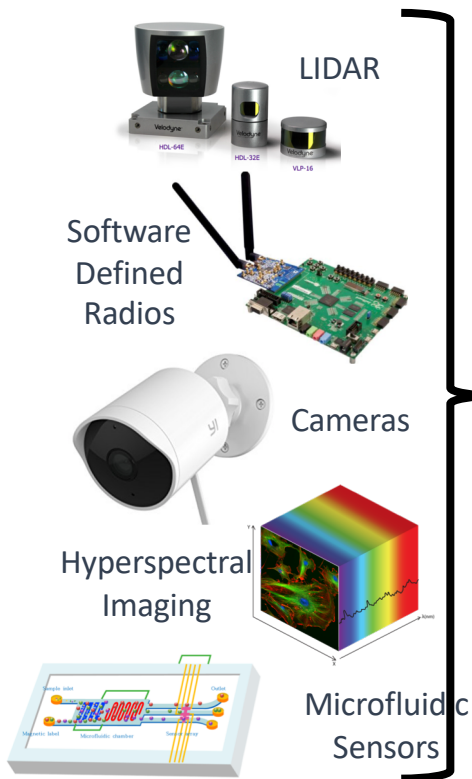
Prediction

Classification of Tumors

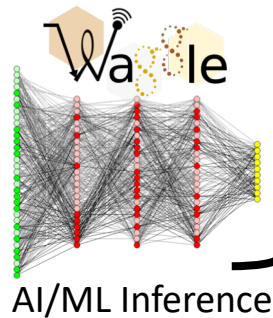


The Edge Computing Revolution

New Sensors



Powerful Parallel Edge Computing



Semantic Output

Reduced, Compressed data

New inference (program code)

Edge computing and deep learning with feedback for continuous improvement
HPC



Deep Learning Training



The Take Away

- **Three computer revolutions**
 - High performance computing
 - Deep learning
 - Edge & AI
- **Technical implications**
 - Fluid end-to-end cyberinfrastructure
 - Interdisciplinary data and infrastructure planning
- **The very small (edge/fog computing and sensors)**
- **The very large (clouds, exascale, and big data)**
- **Cultural implications**
 - Change management and strategic planning
 - Community collaboration

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