U.S. DEPARTMENT OF ENERGY | Office of Science

Future Trends in Advanced Scientific Computing

Presented to the

U.S. Climate Modeling Summit
USGCRP/IGIM

by

Steve Binkley
DOE Office of Advanced Scientific Computing Research
Steve.Binkley@science.doe.gov

March 9, 2016

Leadership and Production Computing Facilities

Titan:

- Peak performance of 27.1 PF
- 18,688 Hybrid Compute Nodes
- 8.9 MW peak power

Mira:

- Peak performance of 10 PF
- 49,152 Compute Nodes
- 4.8 MW peak power

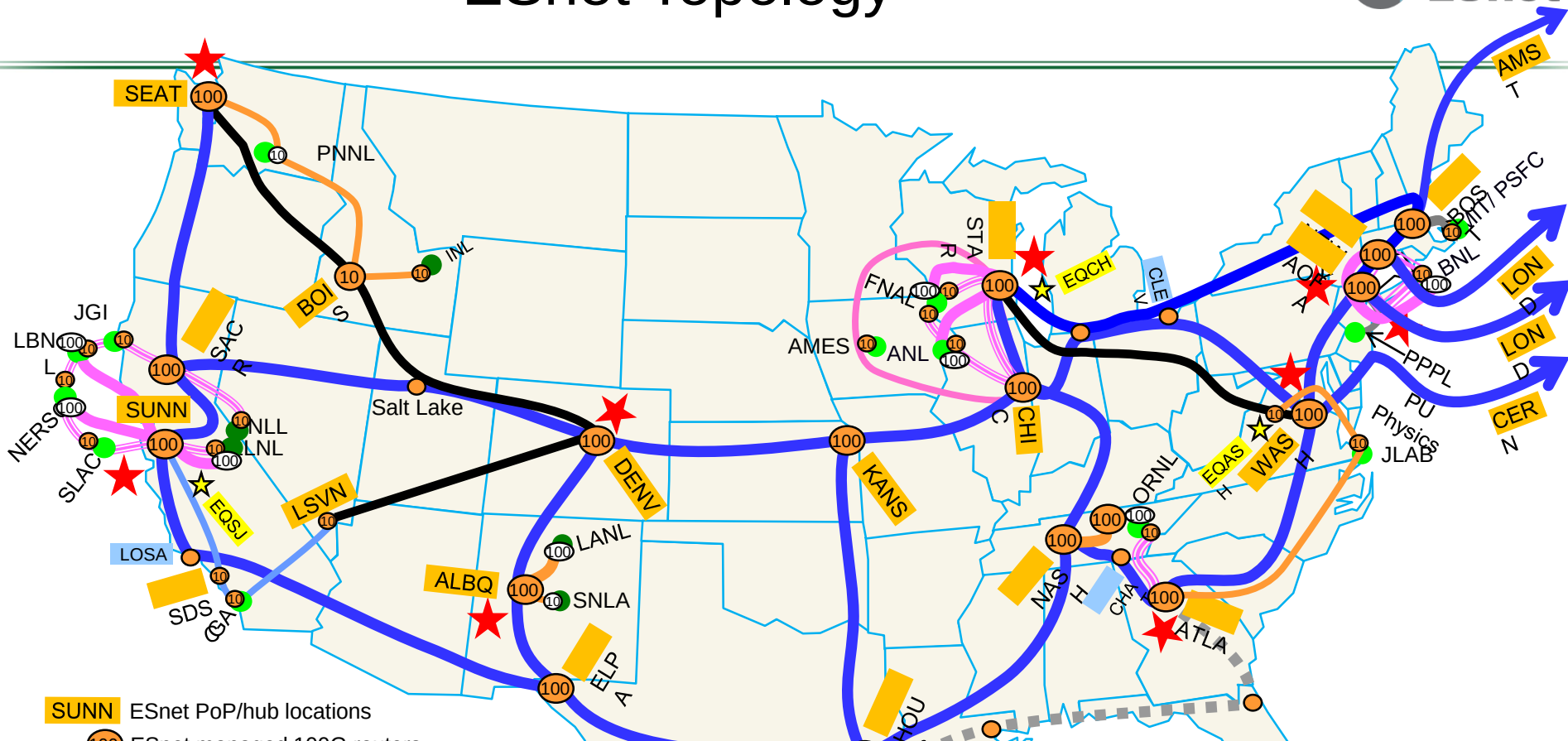
Edison XC30:

- Peak performance 2.4 PF
- 124,608 processing cores
- 2.1 MW peak power

ASCR Computing Upgrades At a Glance

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Name Planned Installation	**Edison**	**TITAN**	**MIRA**	**Cori 2016**	**Summit 2017-2018**	**Theta 2016**	**Aurora 2018-2019**
System peak (PF)	2.6	27	10	> 30	200	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	13.3	1.7	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM)+1.5PB persistent memory	> 2.4 PB DDR4 + HBM + 3.7 PB persistent memory	676 TB DDR4 + High Bandwidth Memory (HBM)	> 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory
Node performance (TF)	0.460	1.452	0.204	> 3	> 40	> 3	> 17 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	Intel Knights Landing Xeon Phi many core CPUs	Knights Hill Xeon Phi many core CPUs
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~4,600 nodes	>3,200 nodes	>50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	Aries	2nd Generation Intel Omni-Path Architecture
File System	7.6 PB 168 GB/s, Lustre®	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	120 PB 1 TB/s GPFS™	10PB, 210 GB/s Lustre initial	150 PB 1 TB/s Lustre®

ESnet Topology



SUNN ESnet PoP/hub locations
100 ESnet managed 100G routers
10 ESnet managed 10G router
10 100 Site managed routers
LOSA ESnet optical node locations (only some are shown)
commercial peering points
R&E network peering locations – US and international
LBNL Major Office of Science (SC) sites
LLNL Major non-SC DOE sites

Geography is only representational

Routed IP 100 Gb/s
Routed IP 40 Gb/s
Routed IP 4 X 10 Gb/s
3rd party 10Gb/s
Express / metro / regional 100 Gb/s
Express / metro 10G
Express multi path 10G
Lab supplied links
Other links
Tail circuits

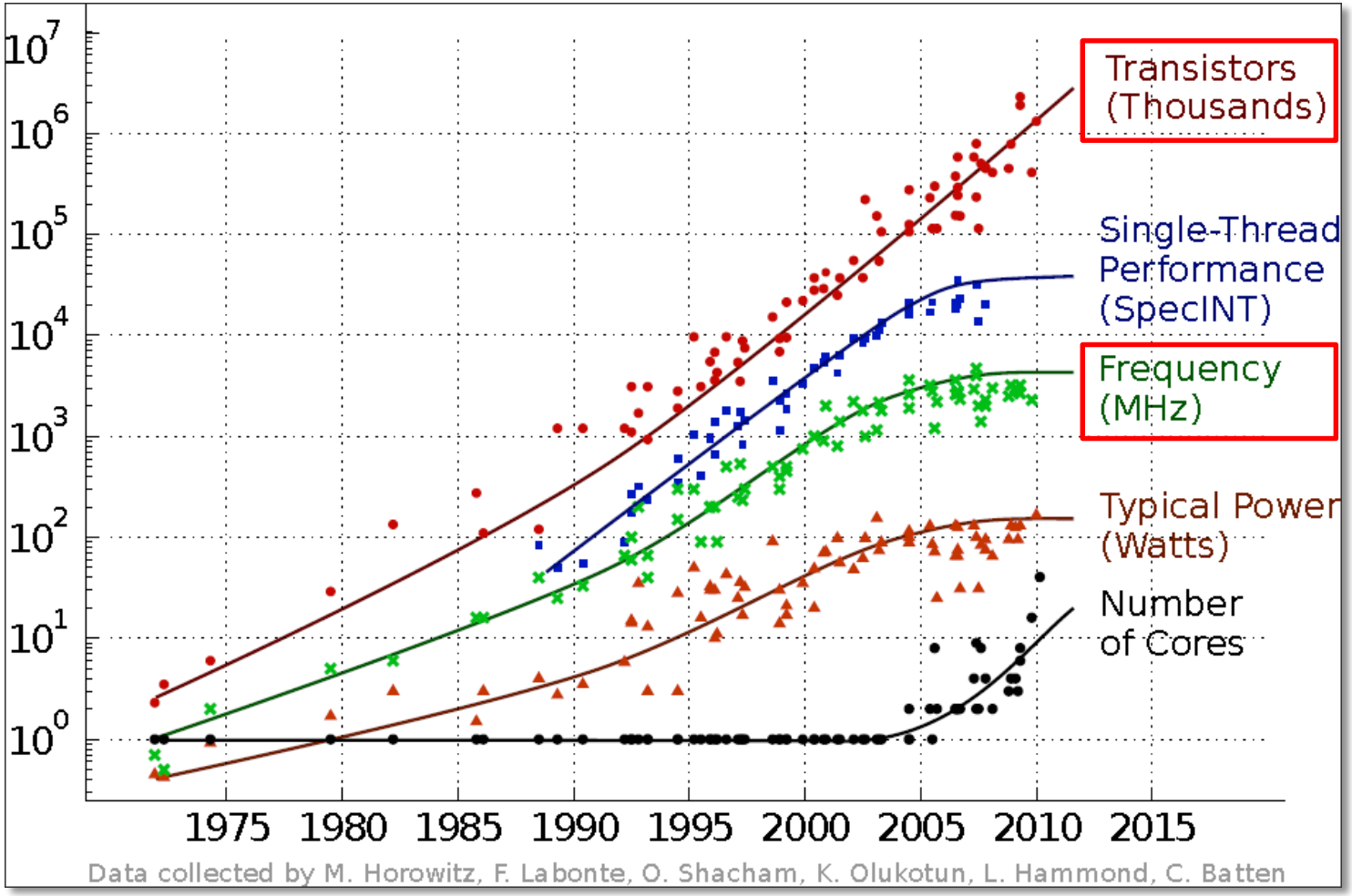
100G testbed

SF Bay Area — SUNN — Chicago — STAR — New York — AOFA

Trends Across the HPC Landscape

- **Increasing importance of computing and simulation within SC and DOE programs (across our missions: science, national security, and quest for cleaner energy)**
  - Continued full subscription of computational resources
  - Increasing importance of effective partnerships (domain + applied math + CS)
  - Difficulties persist in acquiring and retaining highly skilled workforce
- **Nexus of big data and powerful compute is an emerging frontier**
- **The drive toward exascale – Scientifically important and challenging questions await exascale**
  - Dennard scaling is driving further increases in concurrency → billion-way concurrency is coming
  - Post-CORAL computer architectures may be significantly different → significant re-coding
  - Data movement is increasingly costly → becoming the rate limiter
  - Power consumption remains an issue ($1M/megawatt-year)
  - A petaflop in a 19-inch rack
- **The post-Moore's Law epoch is drawing nearer – we need to start preparing**

Progress in CMOS CPU Technology



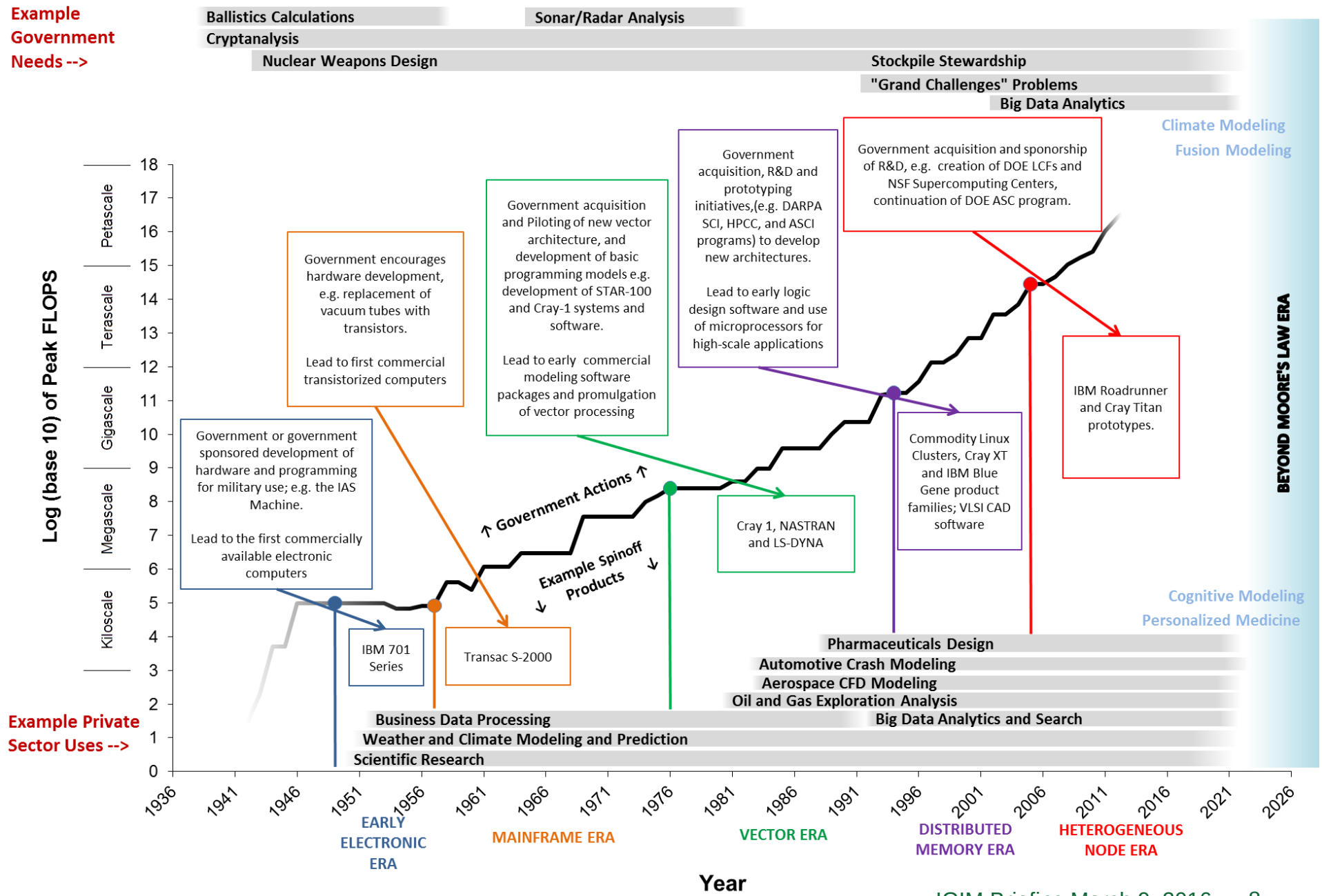
Moore's Law continues

- Transistor count still doubles every 24 months

Dennard scaling stalls

- Voltage
- Clock Speed
- Power
- Performance/clock

National Strategic Computing Initiative


Example Government Needs -->
Ballistics Calculations
Sonar/Radar Analysis
Cryptanalysis
Nuclear Weapons Design
Stockpile Stewardship
"Grand Challenges" Problems
Big Data Analytics
Climate Modeling
Fusion Modeling
Government acquisition and sponorship of R&D, e.g. creation of DOE LCFs and NSF Supercomputing Centers, continuation of DOE ASC program.
Government acquisition, R&D and prototyping initiatives,(e.g. DARPA SCI, HPCC, and ASCI programs) to develop new architectures.
Lead to early logic design software and use of microprocessors for high-scale applications
Government acquisition and Piloting of new vector architecture, and development of basic programming models e.g. development of STAR-100 and Cray-1 systems and software.
Lead to early commercial modeling software packages and promulgation of vector processing
Government encourages hardware development, e.g. replacement of vacuum tubes with transistors.
Lead to first commercial transistorized computers
Government or government sponsored development of hardware and programming for military use; e.g. the IAS Machine.
Lead to the first commercially available electronic computers
IBM Roadrunner and Cray Titan prototypes.
Commodity Linux Clusters, Cray XT and IBM Blue Gene product families; VLSI CAD software
Cray 1, NASTRAN and LS-DYNA
↑ Government Actions ↑
Example Spinoff Products
IBM 701 Series
Transac S-2000
Cognitive Modeling
Personalized Medicine
Pharmaceuticals Design
Automotive Crash Modeling
Aerospace CFD Modeling
Oil and Gas Exploration Analysis
Big Data Analytics and Search
Business Data Processing
Weather and Climate Modeling and Prediction
Scientific Research
Example Private Sector Uses -->
BEYOND MOORE'S LAW ERA
Log (base 10) of Peak FLOPS
Petascale
Terascale
Gigascale
Megascale
Kiloscale
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
1936
1941
1946
1951
1956
1961
1966
1971
1976
1981
1986
1991
1996
2001
2006
2011
2016
2021
2026
EARLY ELECTRONIC ERA
MAINFRAME ERA
VECTOR ERA
DISTRIBUTED MEMORY ERA
HETEROGENEOUS NODE ERA
Year

National Strategic Computing Initiative
Executive Order 13702 Signed July 29, 2015

- **National**
  - "Whole of government" approach
  - Public/private partnership with industry and academia
- **Strategic**
  - Leverage beyond individual programs (a key "platform" technology)
  - Long time horizon (decade or more)
- **Computing**
  - HPC = most advanced, capable computing technology available in a given era
  - Multiple styles of computing and all necessary infrastructure
  - Scope includes everything necessary for a fully integrated capability
    - Theory and practice, software and hardware
- **Initiative**
  - Above baseline effort
  - Link and lift efforts

Enhance U.S. strategic advantage in HPC for security, competitiveness, and discovery

National Strategic Computing Initiative – NSCI

- OSTP/NSTC Fast Track task force, established November 2013
  - Led by Pat Falcone, Associate Director for National Security
  - Charged to define USG approach to HPC for next decade, in time to inform FY 2016 budget
- Membership comprises all federal agencies with equity in HPC
  - EOP (OSTP, OMB)
  - DOE (SC, NNSA, Energy, Labs)
  - IC (IARPA, CIA, NSA)
  - DoD (HPCMod)
  - DOC (NIST, NOAA)
  - NASA
  - NIH
  - NSF
- Executive Order 13702 issued July 29, 2015
- NSCI Executive Council first meeting August 26, 2015
- NSCI “kick-off” workshop, October 20-21, 2015

National Strategic Computing Initiative (NSCI)

Executive Order 13702, July 29, 2015

It is the policy of the United States to sustain and enhance its scientific, technological, and economic leadership position in HPC research, development, and deployment through a coordinated Federal strategy guided by four principles:

1) The United States must deploy and apply new HPC technologies broadly for economic competitiveness and scientific discovery.
2) The United States must foster public-private collaboration, relying on the respective strengths of government, industry, and academia to maximize the benefits of HPC.
3) The United States must adopt a "whole-of government" approach that draws upon the strengths of and seek cooperation among all Federal departments and agencies with significant expertise or equities in HPC in concert with industry.
4) The United States must develop a comprehensive technical and scientific approach to efficiently transition HPC research on hardware, system software, development tools, and applications into development and, ultimately, operations.

This order establishes the NSCI to implement this whole-of-government strategy, in collaboration with industry and academia, for HPC research, development, and deployment.

NSCI Objectives

Executive departments, agencies, and offices participating in the NSCI shall pursue five strategic objectives:

1) Accelerating delivery of a capable exascale computing system that integrates hardware and software capability to deliver approximately 100 times the performance of current 10 petaflop systems across a range of applications representing government needs.
2) Increasing coherence between the technology base used for modeling and simulation and that used for data analytic computing.
3) Establishing over the next 15 years, a viable path forward for future HPC systems in the Post-Moore's-Law Era to advance beyond traditional lithographic scaling of devices.
4) Increasing the capacity and capability of an enduring national HPC ecosystem, employing a holistic approach that addresses relevant factors such as networking technology, workflow, downward scaling, foundational algorithms and software, and workforce development.
5) Developing an enduring public-private collaboration to ensure that the benefits of the research and development advances are, to the greatest extent, shared between the U.S. commercial, government, and academic sectors.

NSCI Agency Responsibilities

"whole-of government" approach

The order identifies lead agencies, foundational research and development agencies, and deployment agencies:

- Lead agencies (DOE, DoD, NSF) are charged with developing and delivering the next generation of integrated HPC capability and will engage in mutually supportive research and development in hardware, software, and workforce to support the objectives of the NSCI.
- Foundational research and development agencies (IARPA, NIST) are charged with fundamental scientific discovery work and associated advances in engineering necessary to support the NSCI objectives.
- Deployment agencies (NASA, FBI, NIH, DHS, NOAA) will develop mission-based HPC requirements to influence the early stages of design of new HPC systems and will seek viewpoints from the private sector and academia on target HPC requirements.

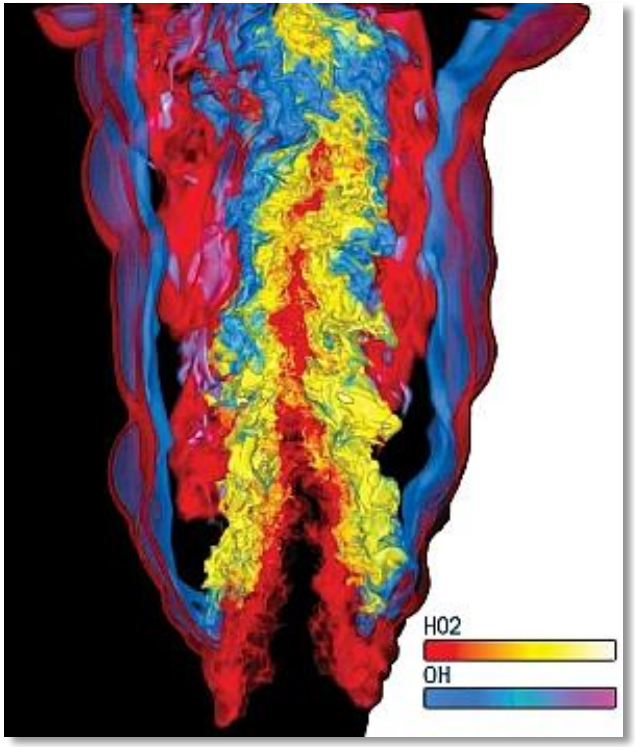
Future Hardware Technology Challenges

- Extending CMOS
  - Serious problems as transistor lengths approach 5nm
  - International Technology Roadmap for Semiconductors predicts 2028
- Other Possibilities
  - Memristors, carbon nanotubes, cryogenic operation, quantum computing
  - All show promise, but nowhere near commercial deployment
  - HPC technology must fit within much larger context of consumer products
- Other Impacts
  - New architectures
  - New programming models based on localized data
  - Need for fault tolerance at higher hardware and software layers
  - New algorithms, languages, software

Exascale

DOE's Exascale Computing Initiative: Next Generation of Scientific Innovation

- **Departmental Crosscut – In partnership with NNSA**
- **"All-in" approach: hardware, software, applications, large data, underpinning applied math and computer science**
- **Supports DOE's missions in national security and science:**
  - Stockpile stewardship – support annual assessment cycle
  - Discovery science – next-generation materials
  - Mission-focused basic science in energy – next-generation climate software
  - Use current Leadership Computing approach for users
- **The next generation of advancements will require Extreme Scale Computing**
  - 100-1,000X capabilities of today's computers with a similar physical size and power footprint
  - Significant challenges are power consumption, high parallelism, reliability
- **Extreme Scale Computing, cannot be achieved by a "business-as-usual," evolutionary approach**
  - Initiate partnerships with U.S. computer vendors to perform the required engineering, research and development for system architectures for capable exascale computing
  - Exascale systems will be based on marketable technology – Not a "one off" system
  - Productive system – Usable by scientists and engineers



Exascale Target System Characteristics

- 20 pJ per average operation
- Billion-way concurrency (current systems are O(million))
- Ecosystem to support new application development and collaborative work, enable transparent portability, accommodate legacy applications
- High reliability and resilience through self-diagnostics and self-healing
- Programming environments (high-level languages, tools, …) to increase scientific productivity

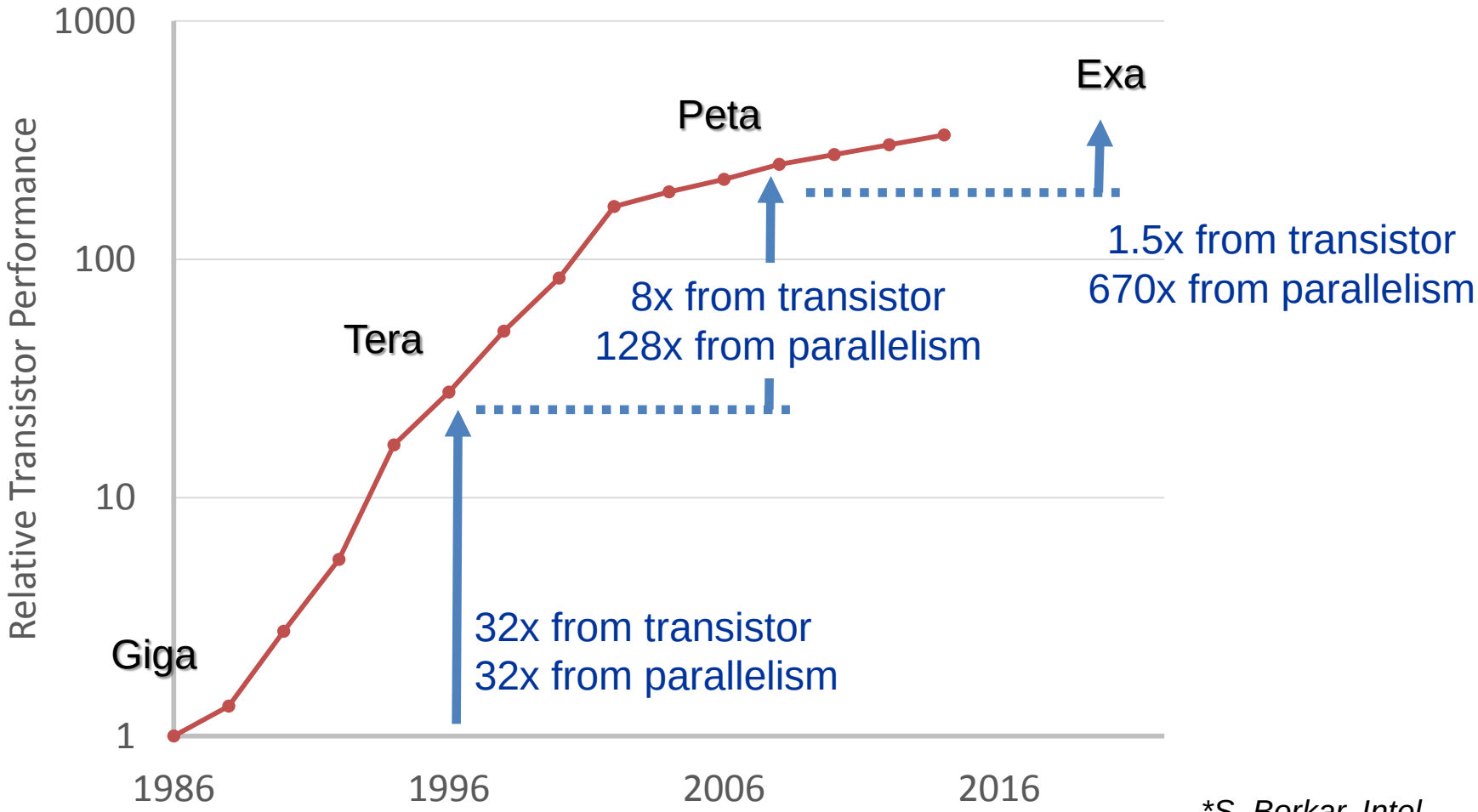
Top Ten Technical Challenges for Exascale

(Advanced Scientific Computing Advisory Committee Findings)

1. **Energy efficiency:** Creating more energy efficient circuit, power, and cooling technologies.
2. **Interconnect technology:** Increasing the performance and energy efficiency of data movement.
3. **Memory technology:** Integrating advanced memory technologies to improve both capacity and bandwidth.
4. ➔ **Scalable System Software:** Developing scalable system software that is power and resilience aware.
5. ➔ **Programming systems:** Inventing new programming environments that express massive parallelism, data locality, and resilience
6. ➔ **Data management:** Creating data management software that can handle the volume, velocity and diversity of data that is anticipated.
7. ➔ **Exascale algorithms:** Reformulating science problems and refactoring their solution algorithms for exascale systems.
8. ➔ **Algorithms for discovery, design, and decision:** Facilitating mathematical optimization and uncertainty quantification for exascale discovery, design, and decision making.
9. **Resilience and correctness:** Ensuring correct scientific computation in face of faults, reproducibility, and algorithm verification challenges.
10. ➔ **Scientific productivity:** Increasing the productivity of computational scientists with new software engineering tools and environments.

http://science.energy.gov/~/media/ascr/ascac/pdf/reports/2013/report.pdf

From Giga to Exa, via Tera & Peta*



*S. Borkar, Intel

Performance from parallelism

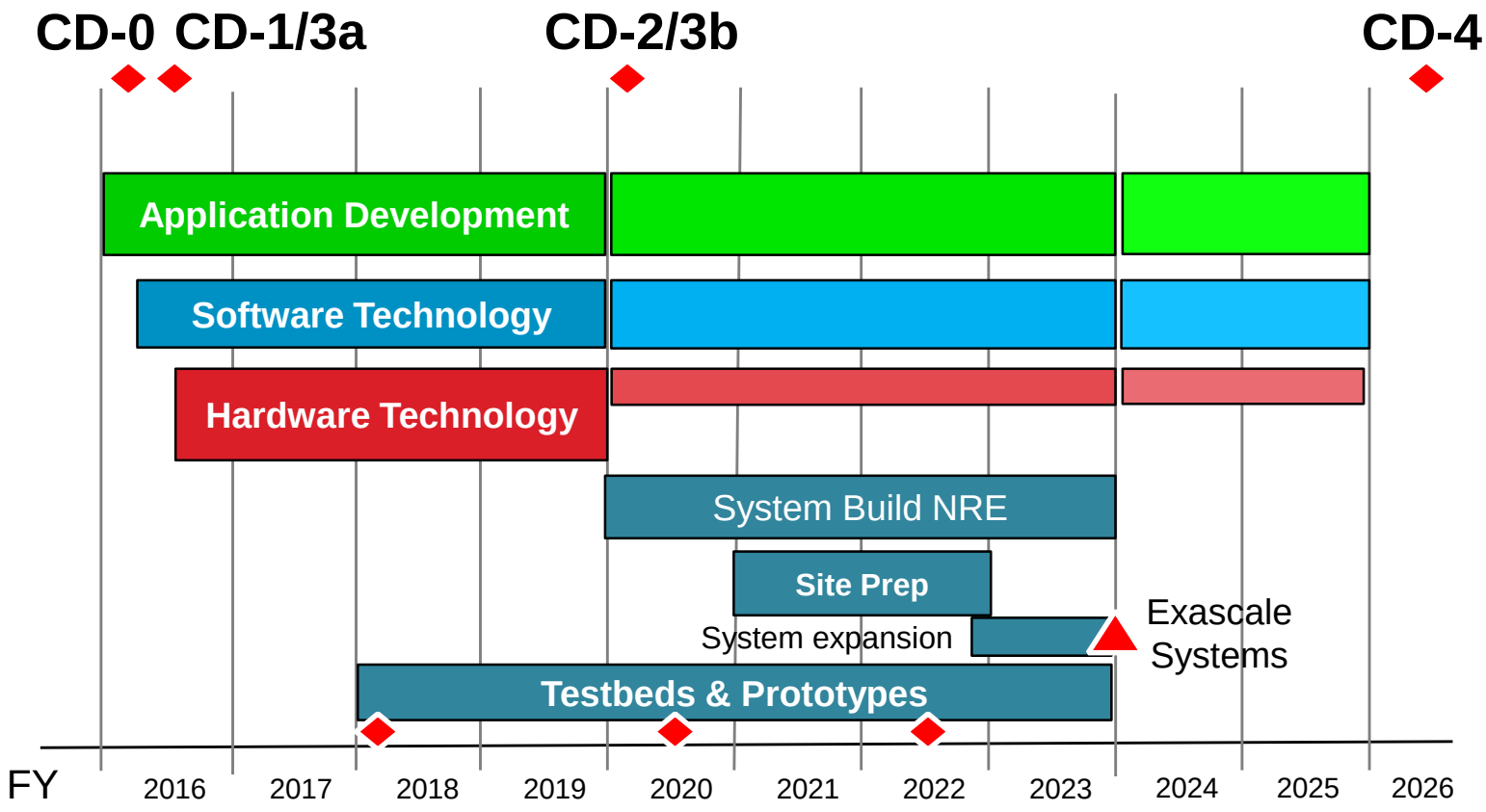
Broadening the Reach of Exascale Science Applications

- **Initial call for input from all 17 DOE Laboratories released on May 31, 2015 to identify potential applications that could deliver new science capabilities on exascale systems. 126 application white papers received. Input will be used by ASCR/ASC to define/refine exascale applications**
  - to identify additional key scientific areas for exascale discovery, and specific opportunities for new and existing scientific applications.
  - to provide broad input on the kinds of partnerships and investments required to address technical challenges of exascale applications.

- **NIH-NSF-DOE Request for Information to identify scientific research topics that need High Performance Computing (HPC) capabilities that extend 100 times beyond today's performance on scientific applications. Released 9/15/15; Responses due to NIH 11/23/15**
  - Information will be used to assist agencies to construct the NSCI roadmap that will guide development of an exascale ecosystem to support scientific research, and to inform the research, engineering and development process. It is likely that a range of advanced capabilities will need to be developed to respond to the varied computing needs across science disciplines.

ECP Timeline

DOE Milestones



CORAL Node/Rack Layout – ORNL Summit Computer

CORAL rack layout

- 18 nodes
- **779 TF**
- 11 TB RAM
- **55 KW**

CORAL/Summit System

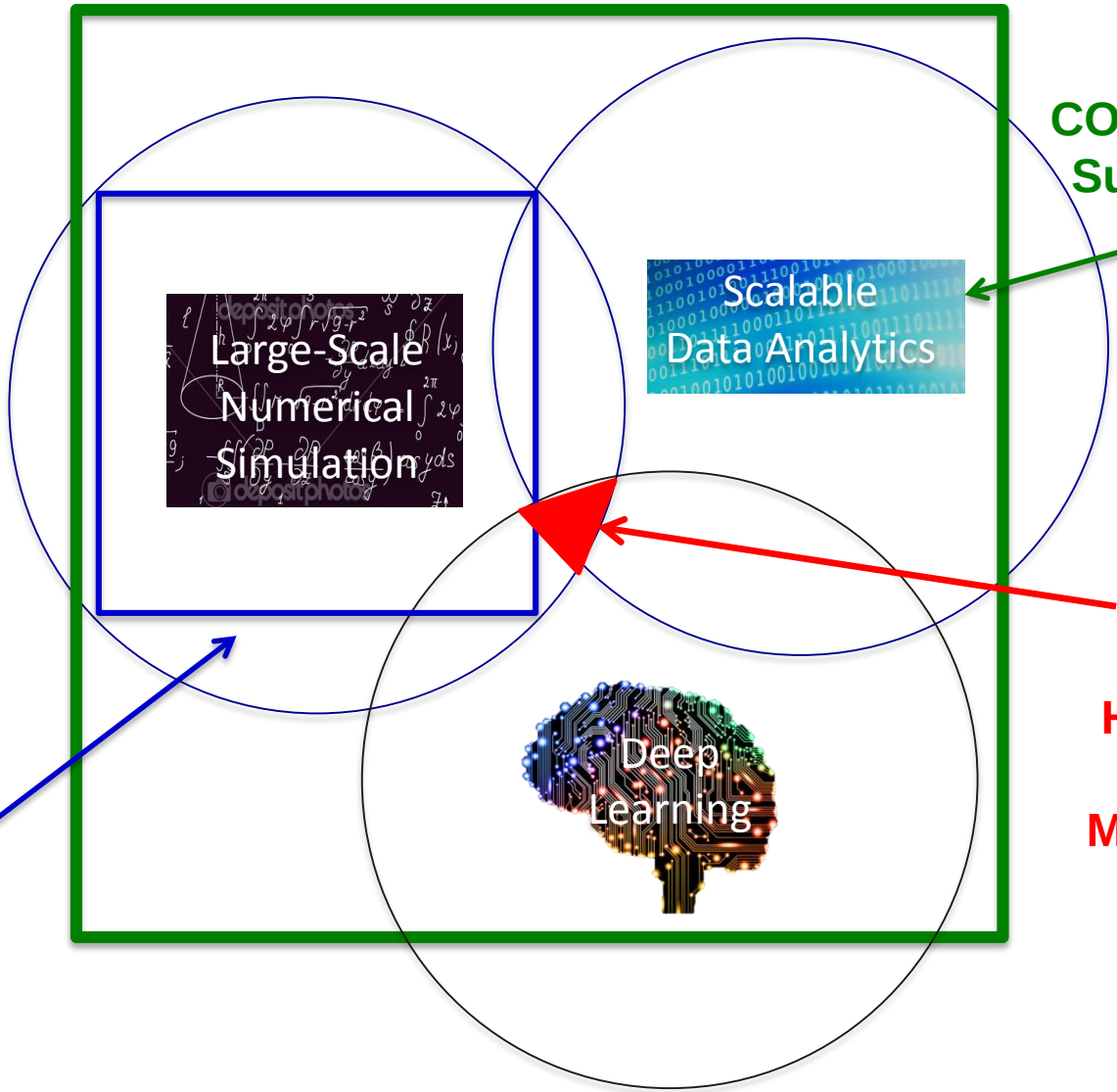
- ~200 racks

Post-Moore's Law Computing: What comes after exascale?

- CMOS lithographic feature sizes are approaching fundamental limits
  - Currently at 22 nm (both Intel and Nvidia)
  - 11 nm is projected for ~~2015~~ 2016 (both Intel and Nvidia)
    - However, gate lengths may be smaller than 6 nm – corresponding gate dielectric thickness may reach a monolayer or less
  - The industry roadmap reaches beyond 11 nm (7 nm and 5 nm) but may be unattainable
    - Non-silicon extensions of CMOS, e.g., using III-V materials or nanotubes/nanowires or non-CMOS technoligies, including molecular electronics, spin-based computing, single-electron devices, and graphene have been proposed
    - At scales of ~10 nm, quantum tunneling may become significant
  - Capital costs for tooling are increasing dramatically as feature sizes shrink
- Options:
  - Computing using superconducting technologies
  - Quantum computing/quantum information science
  - Neuromorphic computing
  - Probabilistic computing
  - ???

Considerable R&D required

Integrating Numerical Simulation, Data Analytics and Machine Learning



Planned CORAL/APEX/ECP Supercomputers

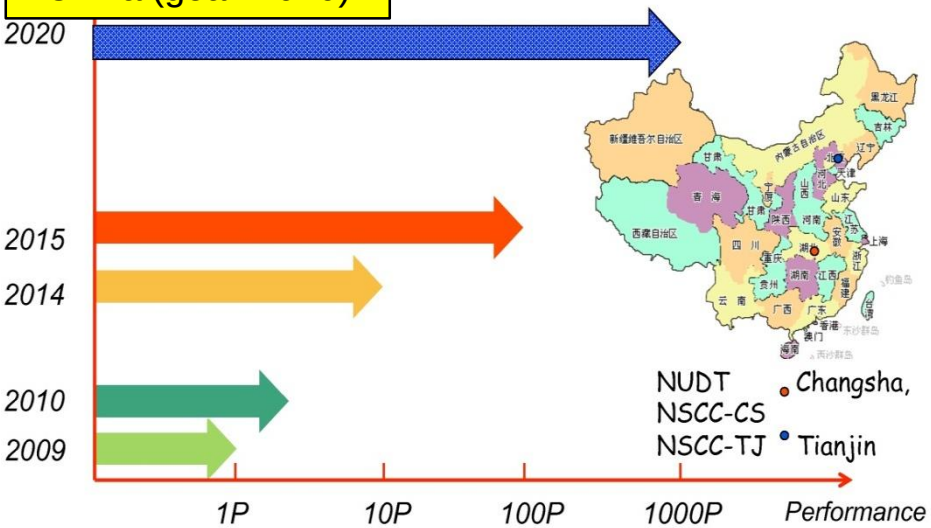
Convergence of HPC, Large-Data Analytics, Machine Learning

Traditional "Big-Iron" HPC Systems

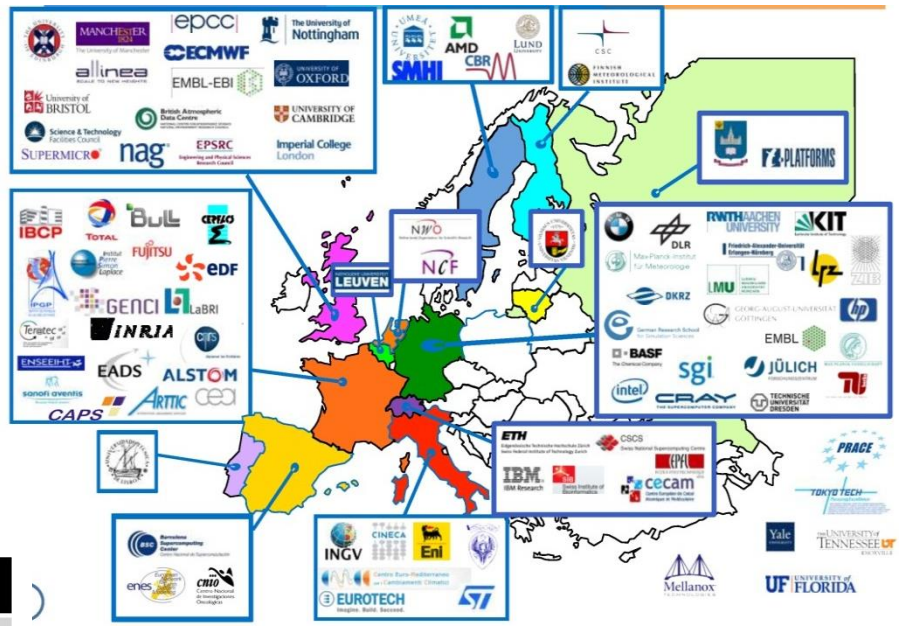
Questions?

International Competitors Have Ambitious Plans

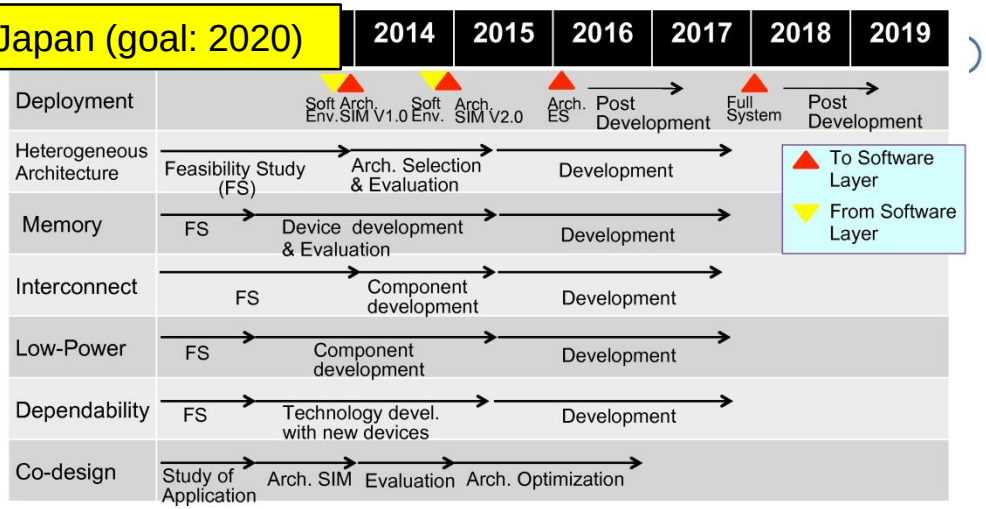
China (goal: 2020)



European Union (goal: 2020)



Japan (goal: 2020)



Other countries:
Russia
South Korea
India
Singapore
Australia

EU: J-Y. Berthou, IESP Workshop, Kobe, Japan, April 2012
China: Y. Lu, IESP Workshop, Kobe, Japan, April 2012
Japan: M. Kondo, IESP Workshop, Kobe, Japan, April 2012

Challenges in High Performance Computing (HPC) for Climate Prediction and Projection

Brian Gross
Acting Deputy Director, NOAA High Performance Computing and Communications

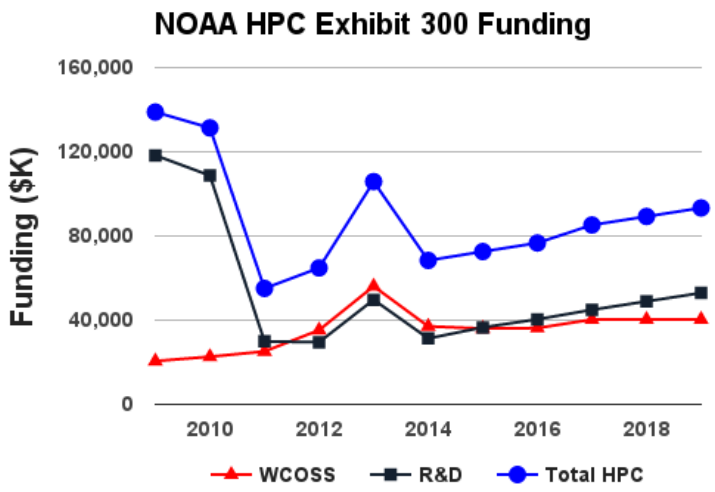
March 9, 2016

Acknowledgements: Bill Collins (DOE) and Bill Putman (NASA)

Driving Questions

- What horizontal and vertical resolutions are necessary to adequately resolve processes in the coupled system that drive both prediction error in short term forecasts and climate simulation bias?
- What is the computational cost of the key biogeochemical/physical processes must be included in models to address mission requirements?
- What is the ideal size of the ensemble needed for this effort both for prediction, for understanding coupled processes and biases, and quantifying uncertainty?
- What modeling improvements will most significantly impact computing and storage requirements (e.g., resolution, processes/complexity, ensemble members, etc) and system balance (between compute, networking, storage, etc)?

Funding Can be a Challenge



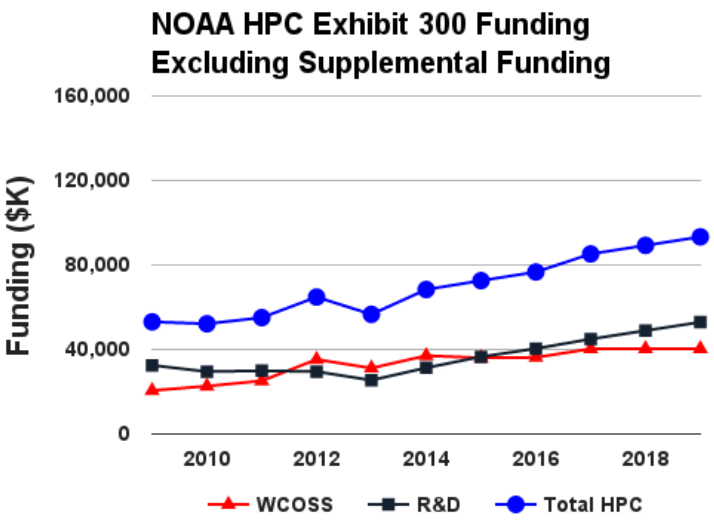
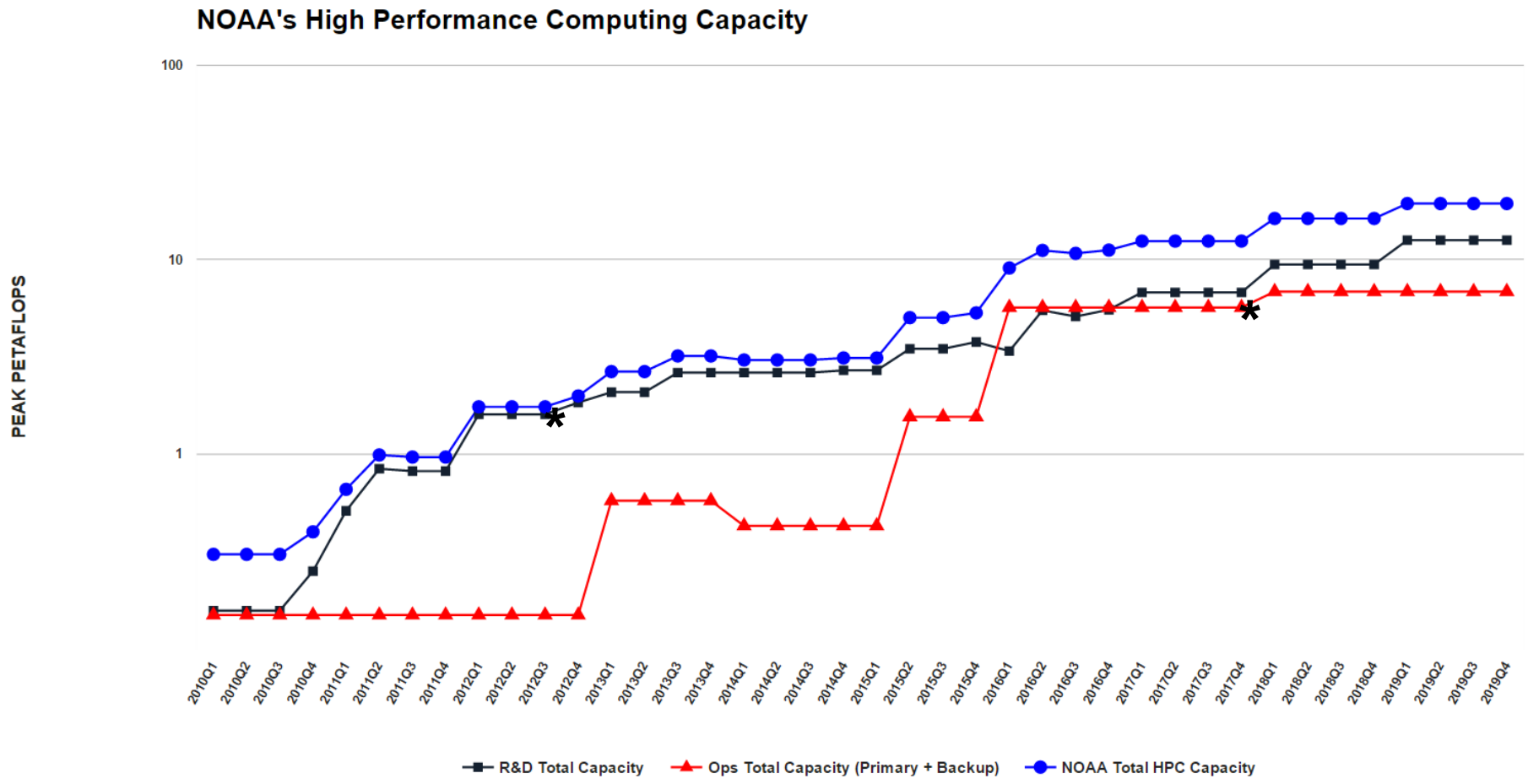


Chart includes FY16 PB increase profile for R&D HPCS

Chart includes FY16 PB increases profile for R&D HPCS

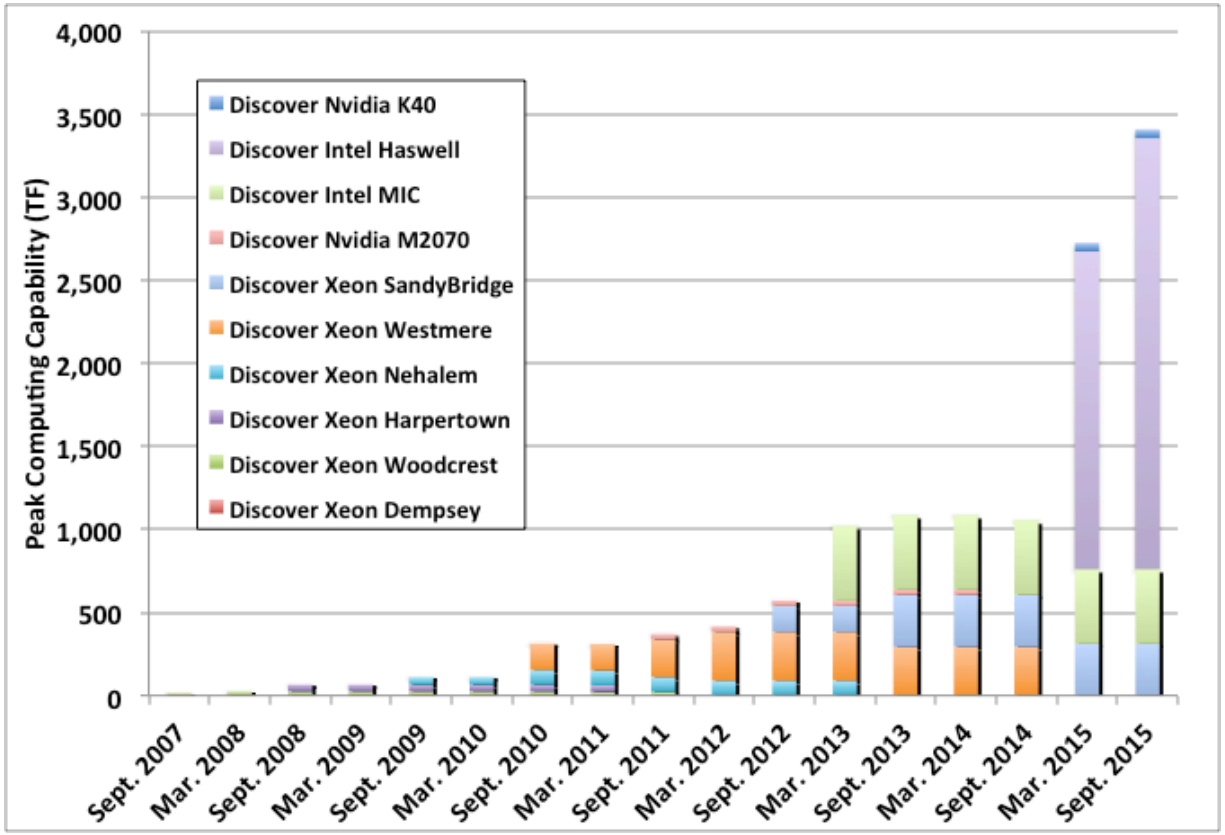
NOAA HPC Growth



NASA HPC Growth

This graph shows the growth of the total peak computing capability of Intel Xeon processors, GPUs, and Intel Phis at the NASA Center for Climate Simulation

- Grew by almost 3x over the past few years
- Over a 100x increase in compute capability over the past 8 years
- Comparable growth in spinning disk over the same time period

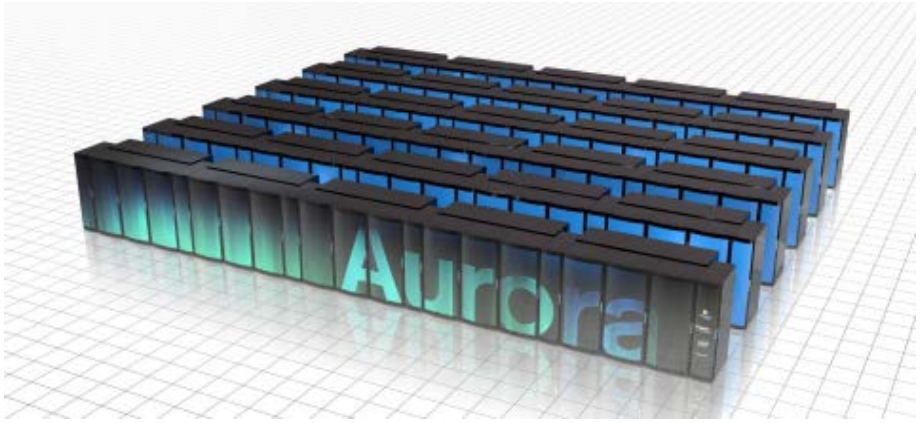


Today's DOE Leadership Systems

- NERSC Edison
  - 2.4PFlops
  - 24 cores (dual Intel Ivey Bridge)
- NERSC Cori (Phase I)
  - 1.9 Pflops
  - 16-core Intel Haswell
- OLCF Titan
  - 27 Pflops
  - 16 core AMD Opteron + NVIDIA GPU
- ALCF Mira
  - 10 PFlops
  - 16 core PowerPC

DOE Next Generation Machines

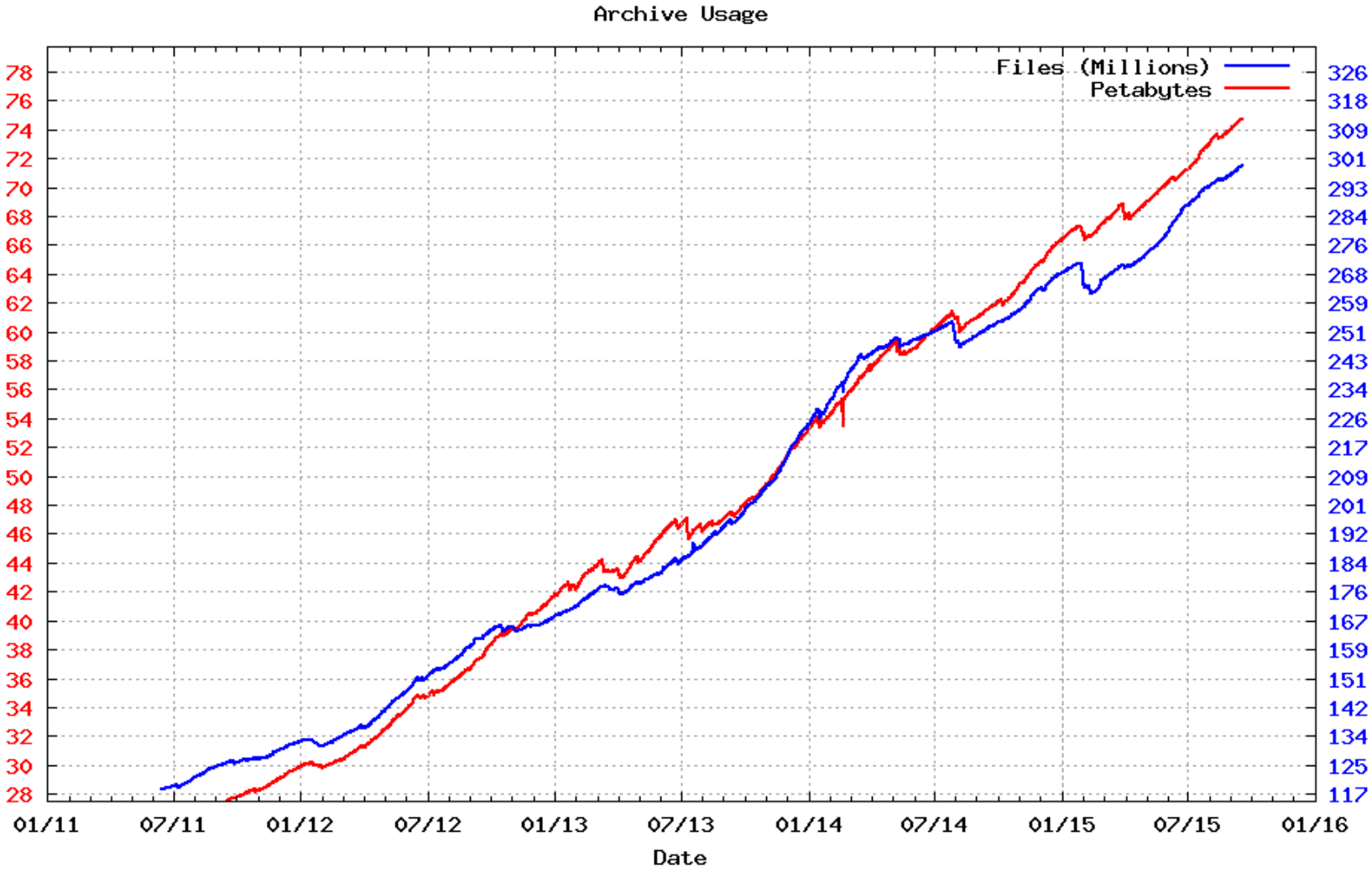
- Key goal of ACME: Run on next generation DOE machines:
- NERSC Cori (Phase II) Late 2016
  - ≥30 PFlops
  - Intel Phi KNL, >60 cores each
- OLCF Summit 2018
  - ~200 PFlops
  - Multiple IBM power9 and NVIDIA GPUs
- ALCF Aurora 2018
  - ~180 Pflops
  - 50K Nodes, 3rd gen Intel Phi



Data Challenges

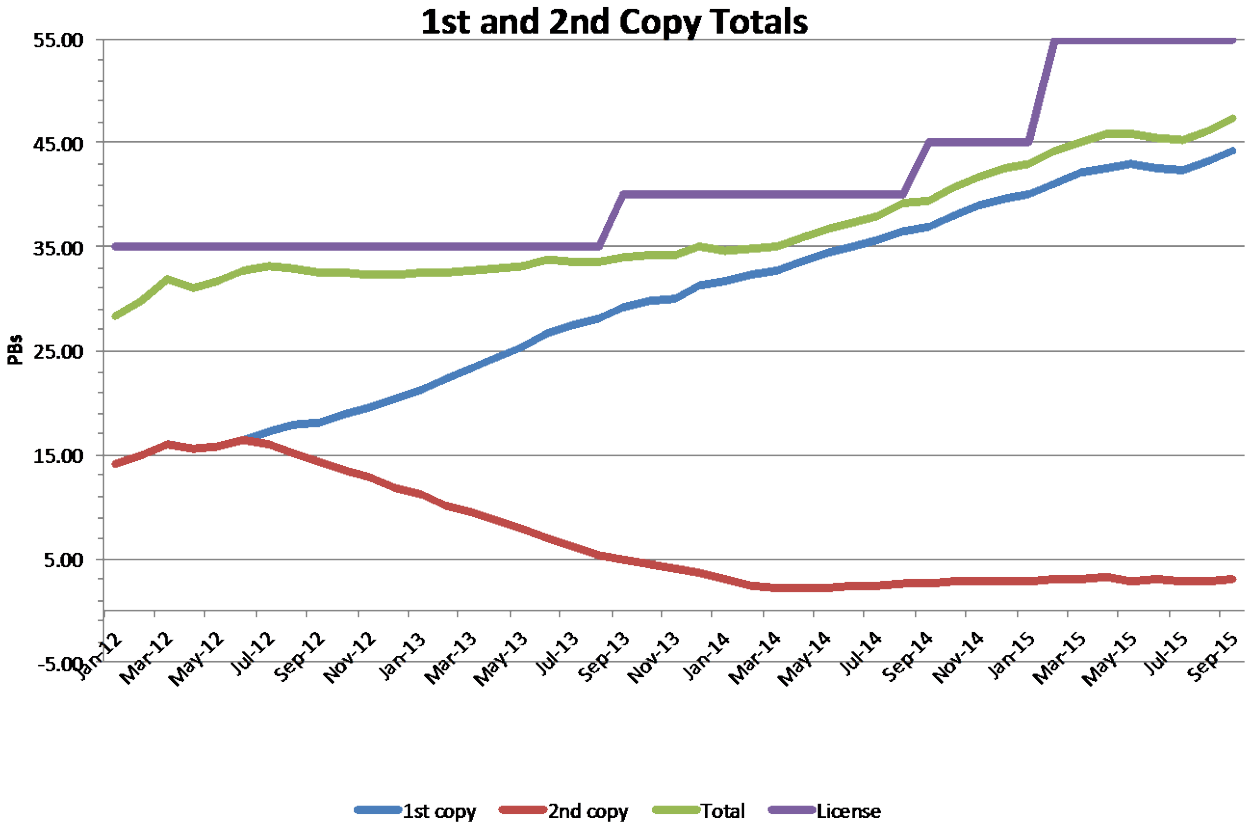
- How will increasingly high-resolution data be stored and shared for community research?
- As resolution increases, it becomes more difficult to save every bit to disk. How can we reduce the storage burden from coupled hi-res integrations?
- What must be analyzed at full resolution, and what can be evaluated at coarser spatial resolution?
- What aspects of your analysis can be in-lined during computation to reduce the required storage?
- How does increased horizontal resolution impact the necessary temporal resolution of your analysis and data storage?
- What new technologies, such as non-volatile random-access memory (NVRAM), provide the greatest potential to improve the scalability and efficiency of your coupled systems and particularly IO bottlenecks that are inevitable at high resolution?

NOAA Data Archive Growth

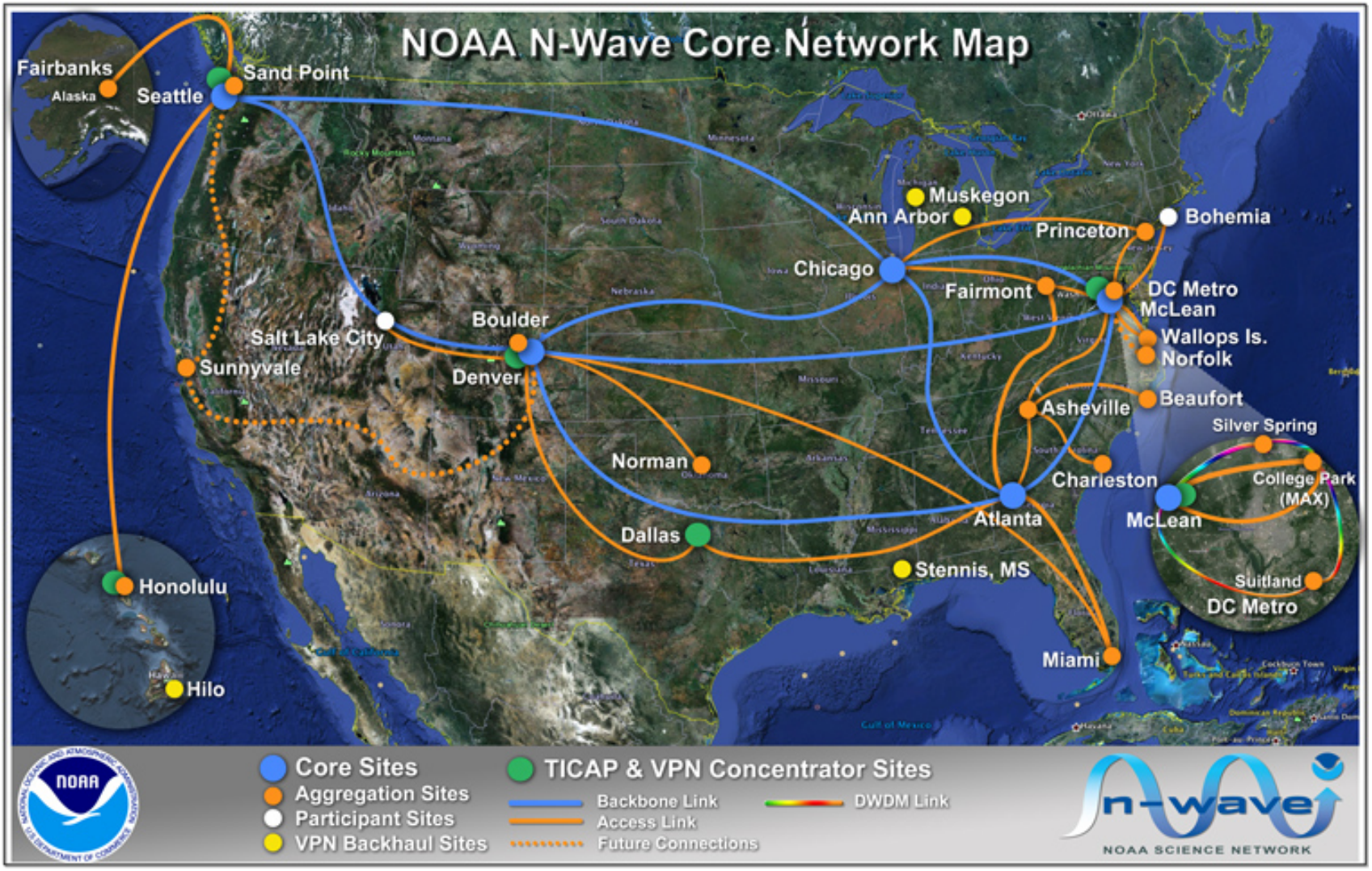


NASA Data Archive Growth

- Stopped making second copies of tape in 2012.
- Close to linear growth regardless of compute resources.
- Quotas have been established as of 2015.
- No sign of this slowing down.



NOAA's Science Network



On software development ….

- Experience to date with fine-grained architectures: kernels can sing (~40X), but complex multi-physics codes croak (~<2X)
- Approach: code revisions for performance on conventional architectures will get us a significant way toward performance on fine-grained systems.
  - Component Concurrency
  - Offload I/O, Diagnostics
  - Performance analysis tools
  - vectorization (requires interaction with compiler vendors)
  - wide halos (to reduce comms)
  - nonmalleable executables (aka static memory)
  - direct use of coarray

ACME Mini-App Strategy

- Recommendation of ACME/Exascale study group: *Identify key kernels/modules that are small enough so a single person can understand/refactor/rewrite to test new approaches, but that are large enough that successful results are meaningful for ACME.*
- Target: Transport mini-apps for both atmosphere and ocean to cover finite element and finite volume approaches used in ACME
- Atmosphere tracer transport is single most expensive ACME component. Ocean tracer transport is 30% of the ocean model

Summary/Discussion

- Driving questions center on what HPC capability and configuration will be need needed to address science priorities.
- Climate agency HPC is growing, but it's unlikely that it's growing fast enough.
- The interaction between simulation, data analytics, and storage needs to be constantly assessed.
- Software innovations to leverage anticipated HPC architectures are hard to implement. And necessary.
- Partnerships on hardware and software have accelerated us toward our goals.

Questions?

Backup

R&D HPCS Overview

Development HPC

Systems Integration Contract (CSC)

- May 2010-May 2019 / $317M / IDIQ
- 9 yrs with 4-yr base, 4-yr option, 1-yr transition

Systems Configuration

Theia - Fairmont, WV (Zeus Replacement)
- Short-term/seasonal/inter-annual predictions
- 1,024 teraflops - Cray

Jet - Boulder, CO (NOAA Skaggs Facility)
- Hurricane forecast improvement
- 421 teraflops - Aspen & Cray

Princeton, NJ (NOAA/GFDL)
- Climate post-processing and analysis
- 106 nodes (8 core Intel Xeon) – Dell

Performance Measures

- Minimum 96.0% System Availability
- Minimum 99.0% Data Availability

Research HPC

Interagency Agreement (DOE/ORNL)

- Aug 2009-Aug 2016 / $108M / Cost Reimbursable
- 5 year agreement extended 2 years
- New IA signed June 18, 2015, through FY2020

Systems Configuration

Gaea - Oak Ridge, TN (Oak Ridge National Lab)
- Climate change research and projections
- 1,100 teraflops Cray

Titan - Oak Ridge, TN (Oak Ridge National Lab)
- Applications for next generation architectures
- 500 teraflops (2.6M node-hours) allocation of 27,000 teraflops Cray using Nvidia Graphics Processing Units

Performance Measures

- Minimum 96.0% System Availability
- Minimum 99.0% Data Availability

WCOSS Overview

Contract Award

- Awarded to IBM on November 23, 2011
- ID/IQ w/ firm fixed price task orders
- Contract Value of $502 million
- Period of Performance
  - 5 year base period (Nov 2011 - Nov 2016)
  - 3 year option period
  - 2 year option period for transition

Task Orders

- Task Order 01
  - Initial project management task
- Task Order 02
  - Phase I Base system 170 TF (2012)
  - Phase 2 Midlife Upgrade 600 TF (2015)
- Task Order 03
  - Phase I enhancement 60 TF (2012)
- Task Order 04: Cray XC-40
  - 2,060 teraflops per site

Facility Locations

- Primary
  - Reston, VA (IBM provided facility)
- Backup
  - Orlando, FL (IBM provided facility)

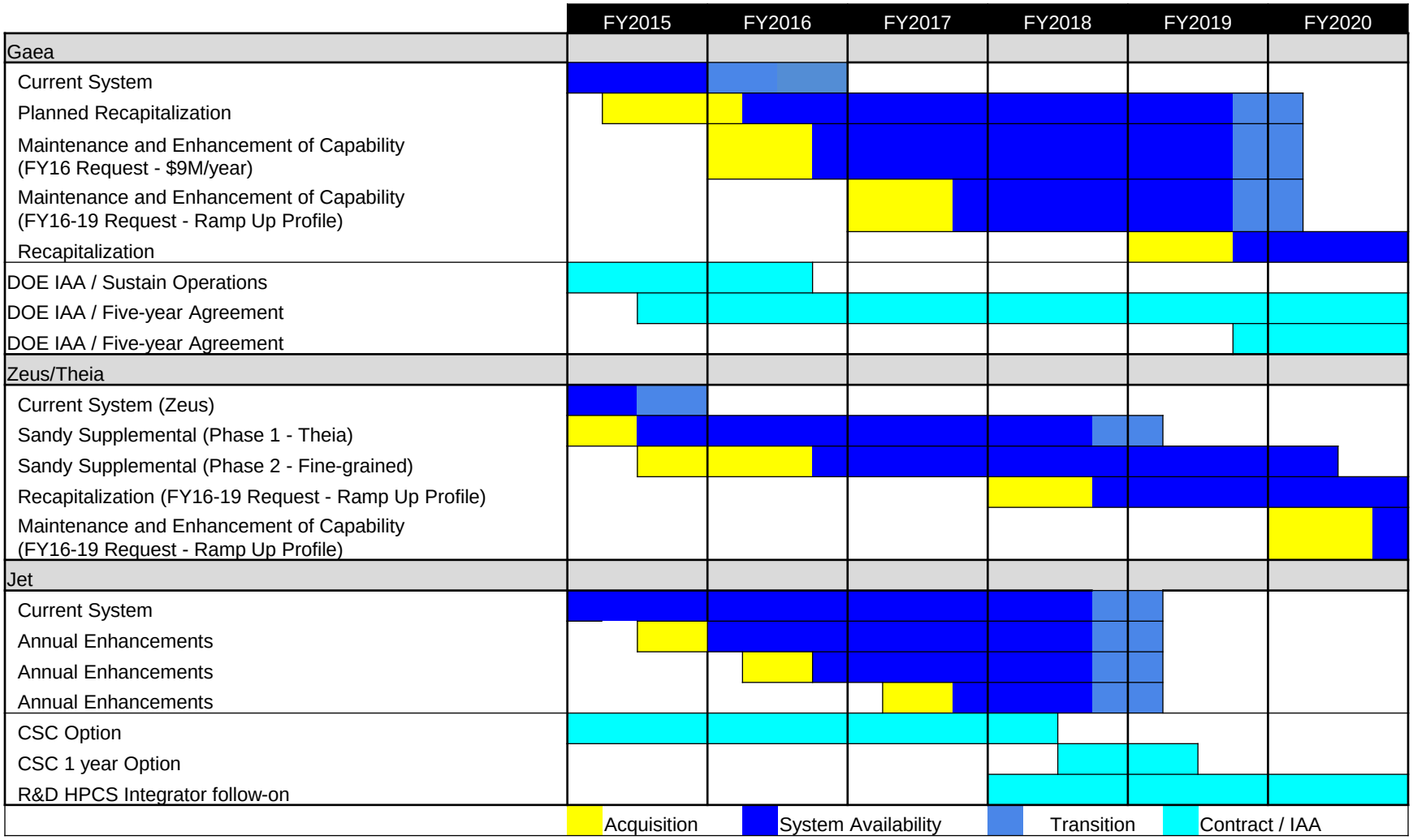
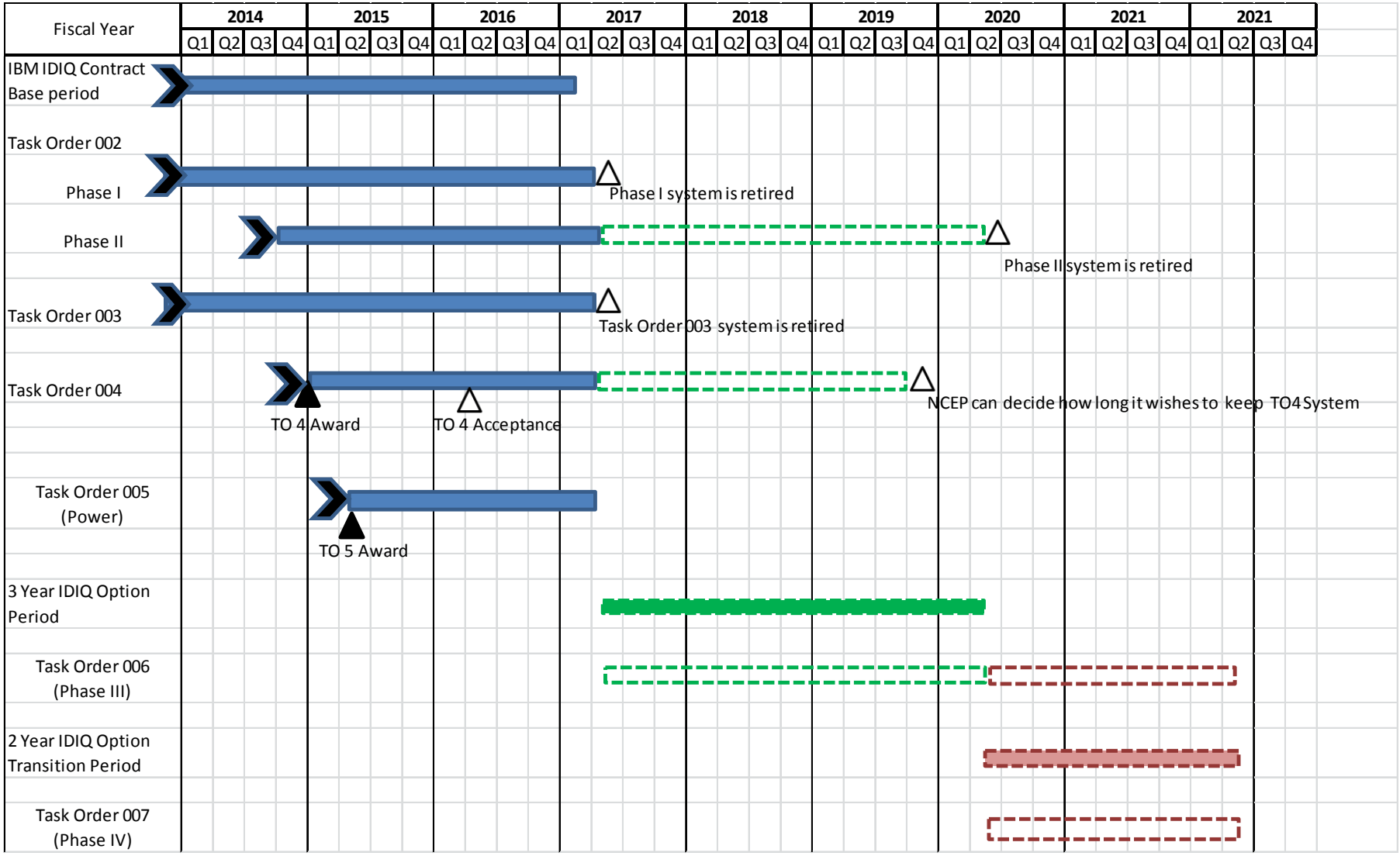
System Configuration

- Identical Systems (per site)
  - IBM iDataPlex (Sandybridge)/NextScale (IvyBridge)
  - 830 teraflops per site

Performance Requirements

- Minimum 99.9% Operational Use Time
- Minimum 99.0% On-time Product Generation
- Minimum 99.0% Development Use Time
- Minimum 99.0% System Availability
- Failover tested regularly

R&D HPCS Schedule



WCOSS Schedule