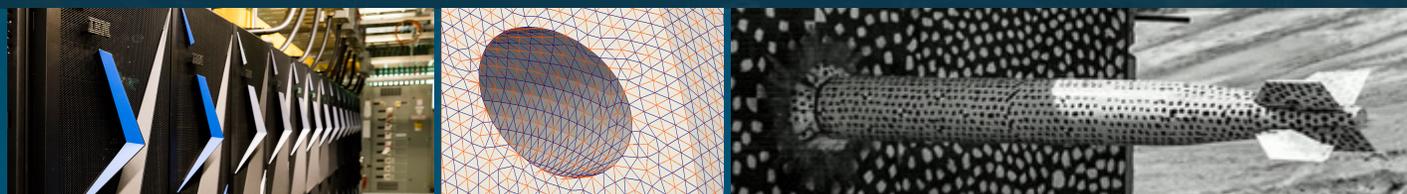


The Challenges of Exascale Computing



PRESENTED BY

Clay Hughes and Si Hammond



Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

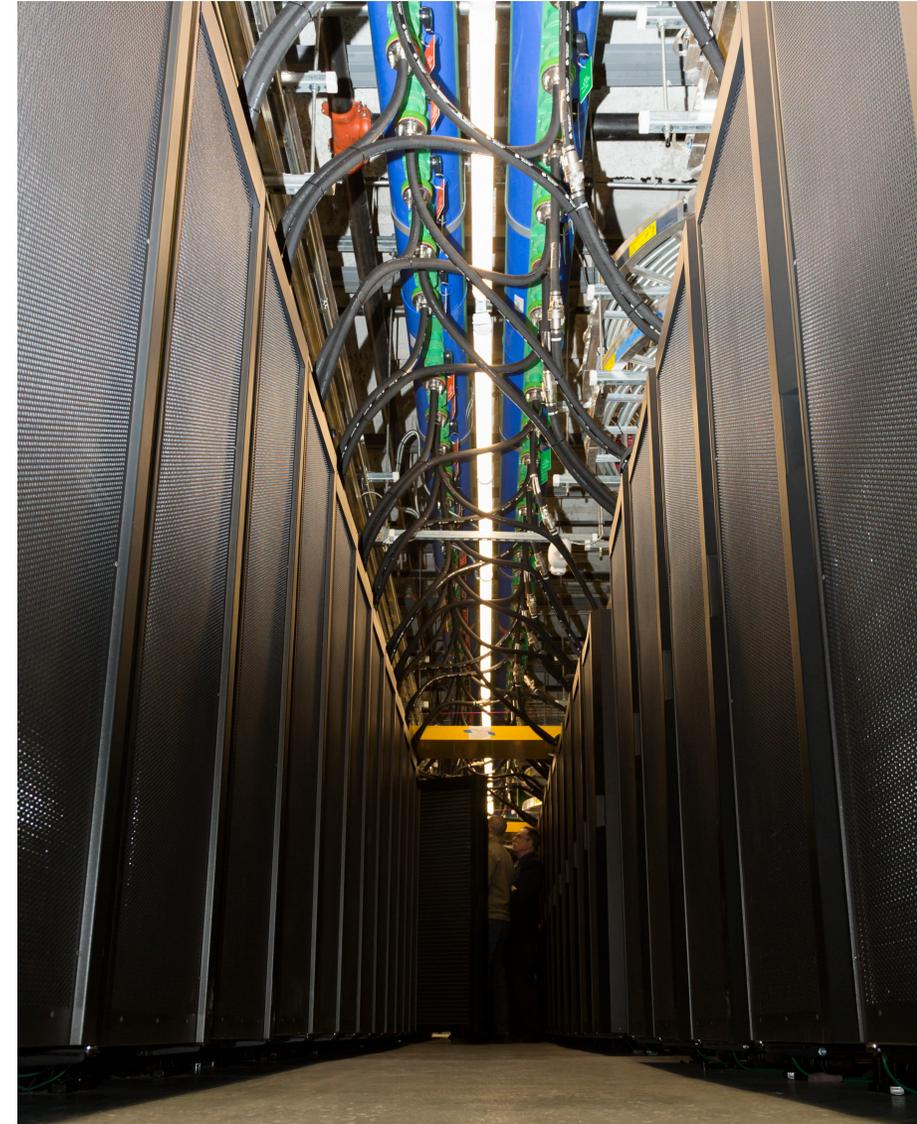
What is High Performance Computing?



Massively parallel systems, typically made of many small commodity servers networked together, performing a few large or many small tasks.

Why do we need it?

- People use supercomputers in their everyday lives
 - Google (capacity computing)
 - Weather forecasts (capability)
 - Amazon (capacity computing)
- People rely on research conducted with supercomputers
 - Oil discovery and extraction
 - Drug research and personalized medicine
 - Fundamental science (physics, chemistry, etc.)



Summit – Image Courtesy of ORNL

Sandia's Interest in High-Performance Computing



Sandia is a national security laboratory

World leading engineering capability

- Ensure safety of critical systems
- Ensure reliability of complex systems

Thread detection and modeling

Material science and modeling

Energy security

Robotics

Microsystems and microelectronics

Space Systems

Significant computational challenges in these areas





President Obama issued an executive order on July 29, 2015 to establish the National Strategic Computing Initiative (NSCI)

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. Deploy new HPC technologies for economic competitiveness and scientific discovery
2. Foster public-private collaboration, relying on the respective strengths of government, industry, and academia to maximize the benefits of HPC
3. Draw upon the strengths of and seek cooperation among all departments and agencies with HPC expertise while also collaborating with industry and academia
4. Develop a comprehensive technical and scientific approach to transition HPC research on hardware, system software, development tools, and applications efficiently into development and, ultimately, operations

Mission Need Defines the Application Strategy

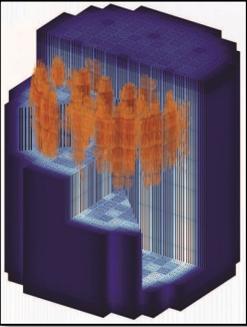
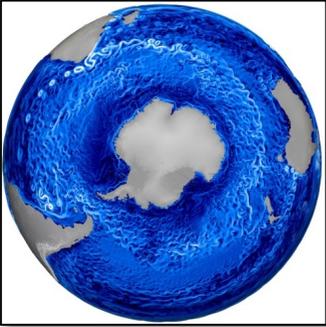
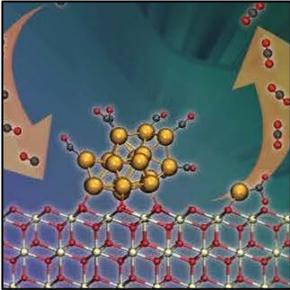
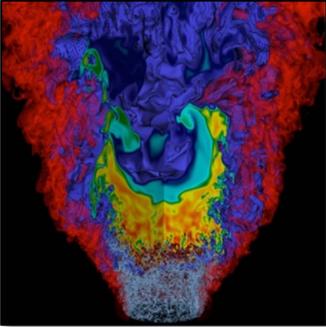


Support DOE science <i>and</i> energy missions	Meet national security needs	Key science and technology challenges to be addressed with exascale
<ul style="list-style-type: none">• Discover and characterize next-generation materials• Systematically understand and improve chemical processes• Analyze the extremely large datasets resulting from the next generation of particle physics experiments• Extract knowledge from systems-biology studies of the microbiome• Advance applied energy technologies (e.g., whole-device models of plasma-based fusion systems)	<ul style="list-style-type: none">• Stockpile Stewardship Annual Assessment and Significant Finding Investigations• Robust uncertainty quantification (UQ) techniques in support of stockpile lifetime extension programs• Understanding evolving nuclear threats posed by adversaries and in developing policies to mitigate these threats	<ul style="list-style-type: none">• Materials discovery and design• Climate science• Nuclear energy• Combustion science• Large-data applications• Fusion energy• National security• Additive manufacturing• Many others!

6 Exascale Applications Will Address National Challenges



Summary of current DOE Science & Energy application development projects

Nuclear Energy (NE)	Climate (BER)	Chemical Science (BES, BER)	Wind Energy (EERE)	Combustion (BES)
<p>Accelerate design and commercialization of next-generation small modular reactors*</p> <p>Climate Action Plan; SMR licensing support; GAIN</p>	<p>Accurate regional impact assessment of climate change*</p> <p>Climate Action Plan</p>	<p>Biofuel catalysts design; stress-resistant crops</p> <p>Climate Action Plan; MGI</p>	<p>Increase efficiency and reduce cost of turbine wind plants sited in complex terrains*</p> <p>Climate Action Plan</p>	<p>Design high-efficiency, low-emission combustion engines and gas turbines*</p> <p>2020 greenhouse gas and 2030 carbon emission goals</p>
				

* Scope includes a discernible data science component

The Exascale Computing Project (ECP)

A collaboration between two US Department of Energy (DOE) organizations:

- Office of Science (DOE-SC)
- National Nuclear Security Administration (NNSA)

A 7-year project to accelerate the development of *capable* exascale systems

- Led by DOE laboratories
- Executed in partnership with academia and industry



A *capable* exascale computing system will leverage a balanced ecosystem (software, hardware, applications)

Capable Exascale Computing

A capable exascale computing system requires an entire computational ecosystem that:

- Delivers 50x the performance of today's 20PF systems, supporting applications that deliver high-fidelity solutions in less time and address problems of greater complexity
- Operates in a power envelope of 20–30MW
- Is sufficiently resilient (average fault rate: $\leq 1/\text{week}$)
- Includes a software stack that supports a broad spectrum of applications and workloads



This ecosystem will be developed using a co-design approach to deliver new software, applications, platforms, and computational science capabilities

Why the Fuss?



Sierra – LLNL

Nodes	PPN	GPN	Node Peak (tFLOP/s)	System Peak (pFLOP/s)	Off-Node BW (GB/s)	Peak Power (MW)
4320	2	4	29.1	125	45.5	~12

Summit – ORNL

Nodes	PPN	GPN	Node Peak (tFLOP/s)	System Peak (pFLOP/s)	Off-Node BW (GB/s)	Peak Power (MW)
~4600	2	6	~40	~200	~50	~15

Why the Fuss?



Sierra – LLNL

Nodes	PPN	GPN	Node Peak (tFLOP/s)	System Peak (pFLOP/s)	Off-Node BW (GB/s)	Peak Power (MW)
4320	2	4	29.1	125	45.5	~12

Exascale Sierra (8x)

34560	2	4	29.1	1000	45.5?	~96
-------	---	---	------	------	-------	-----

Summit – ORNL

Nodes	PPN	GPN	Node Peak (tFLOP/s)	System Peak (pFLOP/s)	Off-Node BW (GB/s)	Peak Power (MW)
~4600	2	6	~40	~200	~50	~15

Exascale Summit (5x)

23000	2	6	~40	~1000	~50?	~75
-------	---	---	-----	-------	------	-----

Why the Fuss?



Sierra – LLNL

Nodes	PPN	GPN	Node Peak (tFLOP/s)	System Peak (pFLOP/s)	Off-Node BW (GB/s)	Peak Power (MW)
4320	2	4	29.1	125	45.5	~12

Exascale Sierra (8x)

34560	2	4	29.1	1000	45.5?	~96
-------	---	---	------	------	-------	-----

Summit – ORNL

Nodes	PPN	GPN	Node Peak (tFLOP/s)	System Peak (pFLOP/s)	Off-Node BW (GB/s)	Peak Power (MW)
~4600	2	6	~40	~200	~50	~15

Exascale Summit (5x)

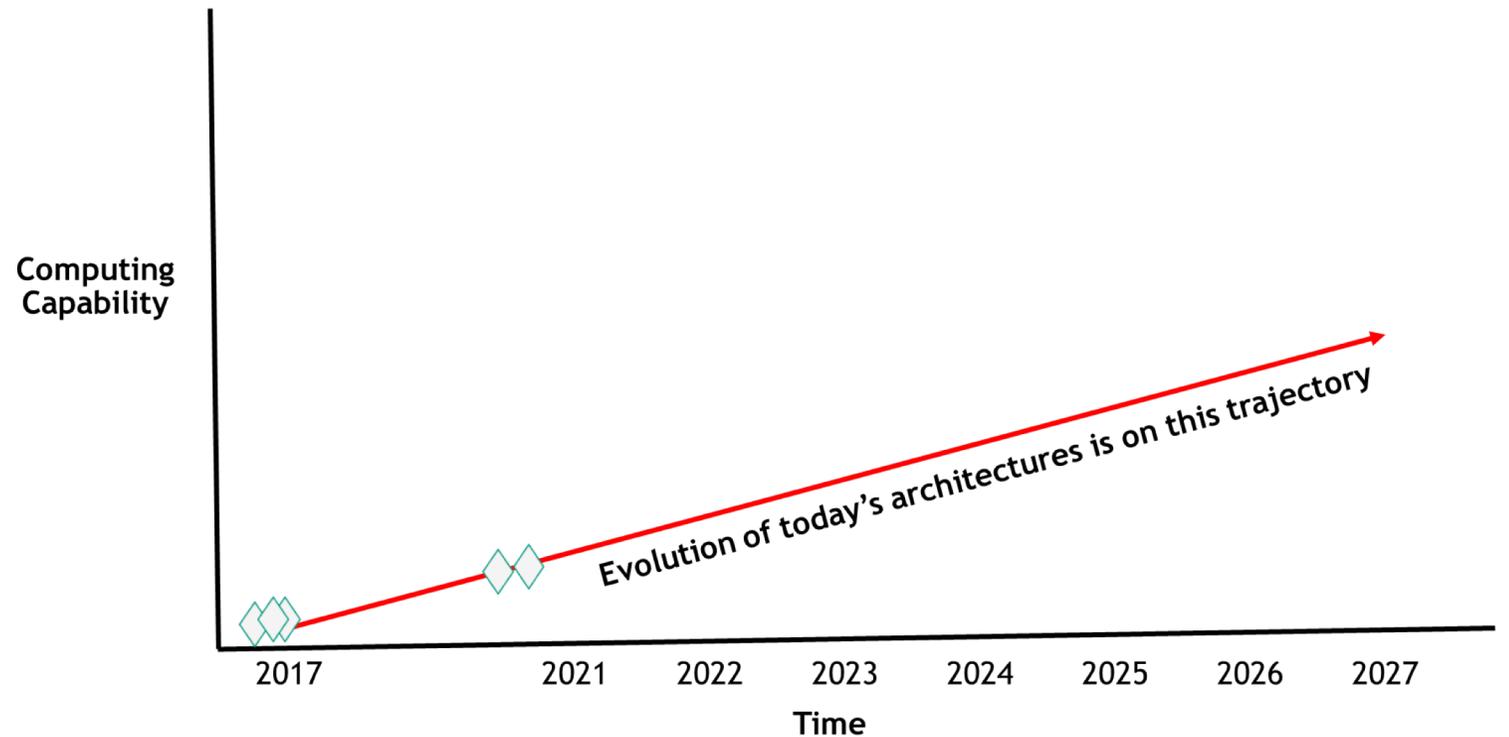
23000	2	6	~40	~1000	~50?	~75
-------	---	---	-----	-------	------	-----

1MW is about \$1M per year in electric cost



Transistor density has continued to scale well despite decreasing clock frequency's

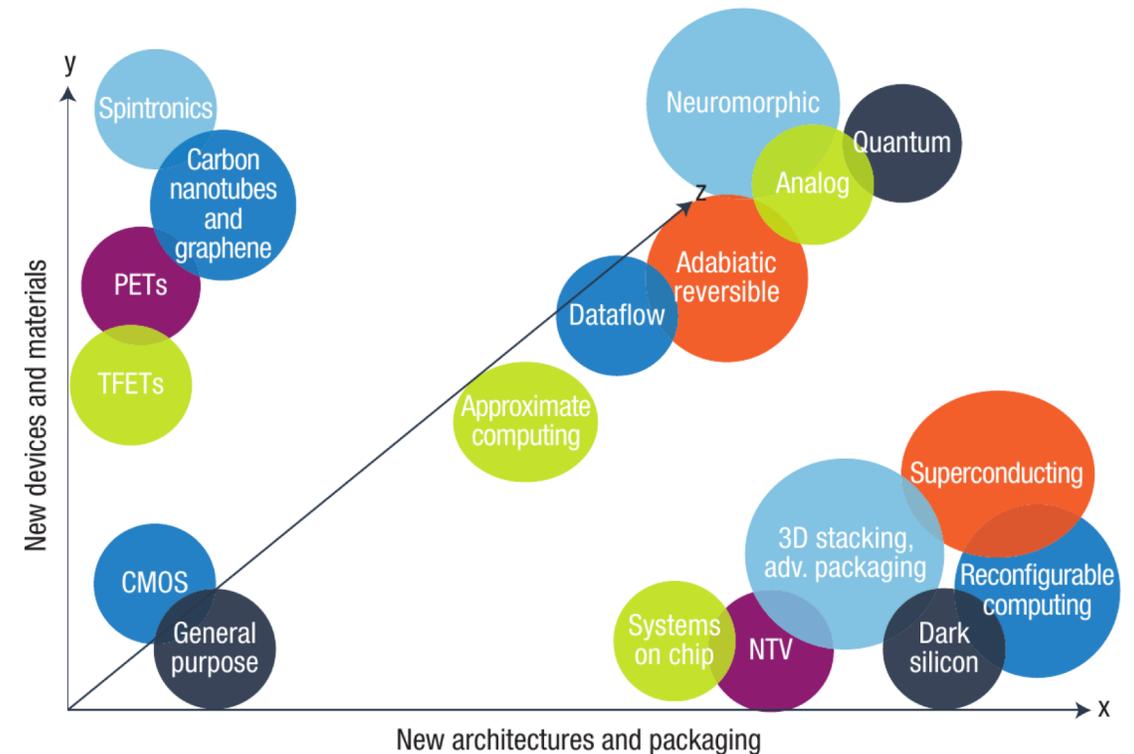
- Vendors provide more computation on a single chip
 - How do we use this computational capacity effectively?
- Scaling will probably end sometime before 2030
 - 2D lithography approaching atomic scale





Transistor density has continued to scale well despite decreasing clock frequency's

- Vendors provide more computation on a single chip
 - How do we use this computational capacity effectively?
- Scaling will probably end sometime before 2030
 - 2D lithography approaching atomic scale



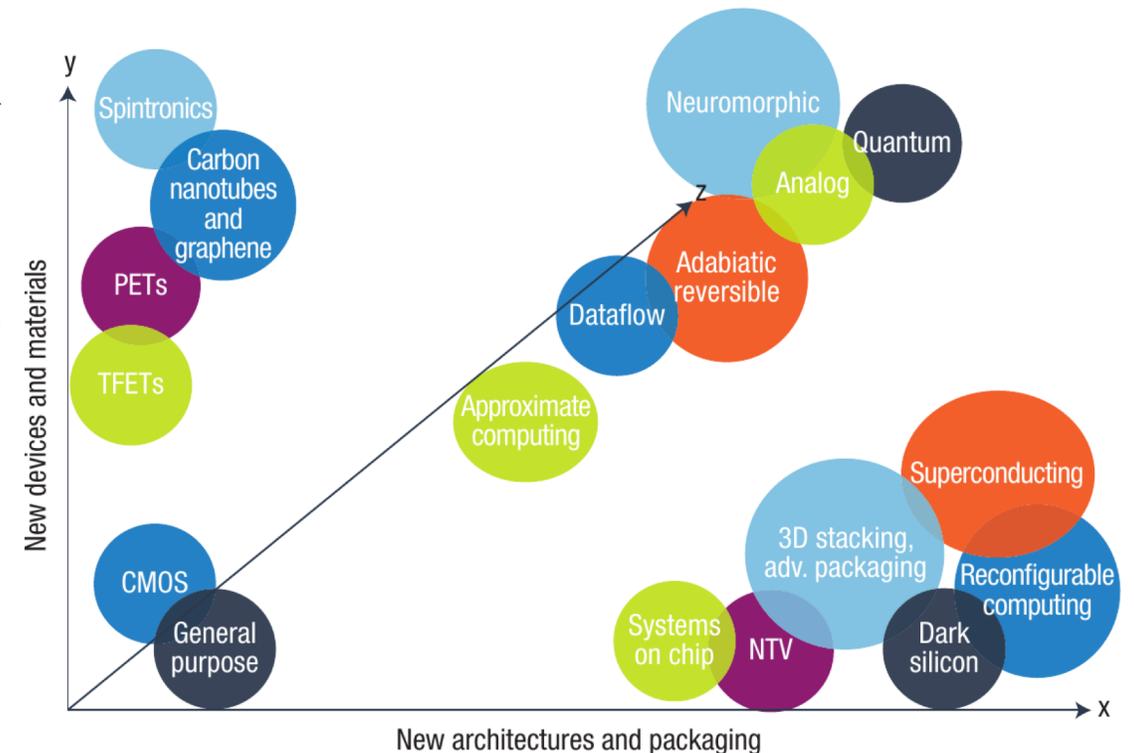


Transistor density has continued to scale well despite decreasing clock frequency's

- Vendors provide more computation on a single chip
 - How do we use this computational capacity effectively?
- Scaling will probably end sometime before 2030
 - 2D lithography approaching atomic scale

Regardless of transistor scaling, there is still the problem of resistance

- $E \propto bitRate * d$
- Pushes programming models toward more localized data movement





Data movement will be a key challenge at exascale [Kestor, 2014]

- $E \propto \text{bitRate} * d$
- Pushes programming models toward more localized data movement
- Memory becomes a major performance and energy bottleneck
- Fetching data for the cores becomes the dominant activity in terms of energy

Operation	Energy (nJ)
ADD	0.64
L1 → REG	1.11
L2 → REG	2.21
L3 → REG	9.80
MEM → REG	63.64
Prefetch	65.08

Kestor, et. al. – ModSim, 2014

Horizontal data movement poses an entirely different set of problems

- Multiple NUMA domains
- Topology-aware routing

Exascale Challenges – Power

Peak power is a major issue

- Constrained to 20-30 MW? How much can we fluctuate up/down?
- Issues for power generation facilities for large swings over short time periods

How do we cool this 20MW space heater?

- Traditional air cooling
- Cold water cooling
- Warm/hot water cooling (use the waste heat for heating buildings)
- Exotic cooling technologies (direct Freon, liquid gases)

Application phases defined by different classes of physics in a “Multiphysics” code



Compute Throughput



Clock rates have been (roughly) flat for the past several years

- Exascale systems will require $O(1b)$ concurrent operations
- Data movement must be streamlined to increase FLOP/B ratios

Technology transitions have often occurred when hardware is able to expose more concurrency and locality to algorithms and applications

- Vector computing (Cray 1)
 - First large body of applications that were designed to map to a specific compute model
- Massively Parallel Processing (distributed memory machines)
 - No easy path to port a performant vector application
 - New framework developed to support new compute model
 - Many codes ditched vectors in favor of optimizing for memory
- Threading and accelerators
 - Much easier to port since frameworks remained (mostly) valid



Cray-2, IT History

Exascale Challenges – FLOPs



Alternative architectures

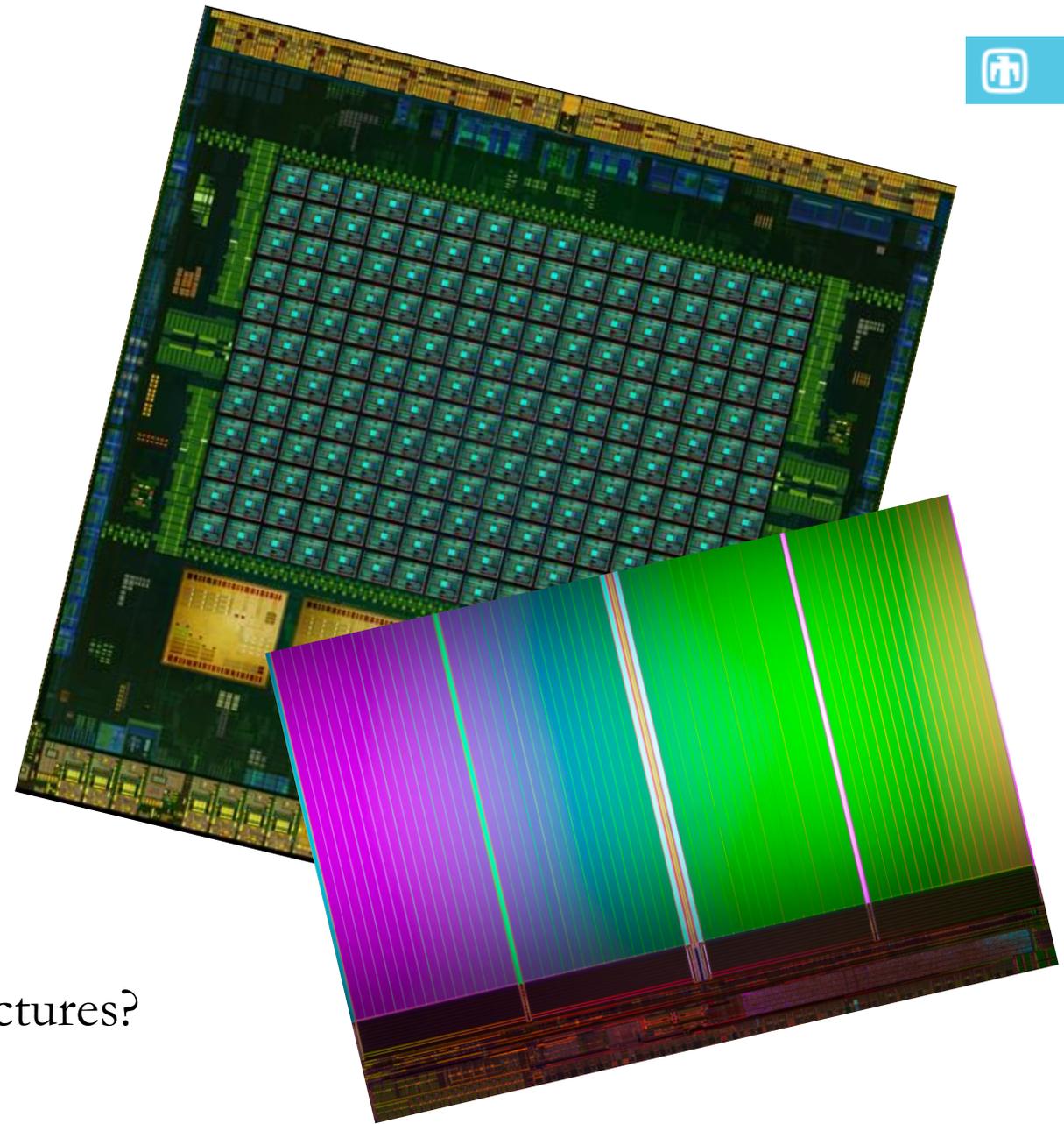
- Many core processors
- Custom accelerators
- GPUs
- FPGAs

Deep memory hierarchies

- NVM (maybe even over fabric)
- More levels of cache
- HBM/Hybrid memory cube

More capability leads to more complexity

- Can we take advantage of many-core architectures?
- What bottlenecks exist?
- What applications map well to alternative computational architectures?

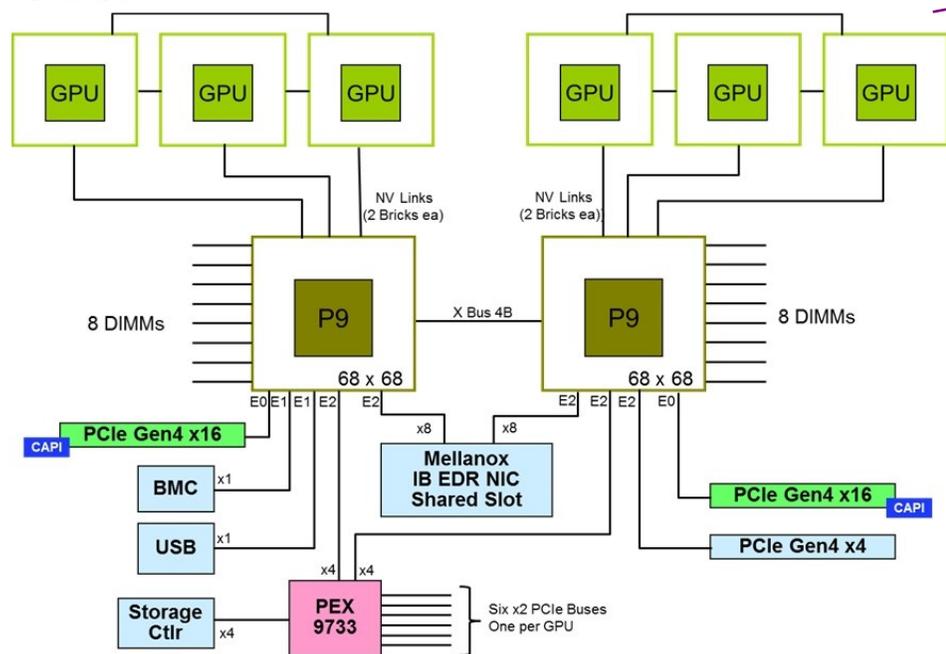




Summit

Nodes	PPN	GPN	Node Peak (tFLOP/s)	System Peak (pFLOP/s)	Off-Node BW (GB/s)	Peak Power (MW)
~4600	2	6	~40	~200	~50	~15

Node





How do that many threads/parallel execution streams communicate effectively?

Are we still going to use a message passing + X type programming model for communication?

- Too much money/time has been invested in MPI programs → It looks like we're stuck with MPI

How does the message passing model effectively scale to Exascale sizes?

- One leading interconnect, InfiniBand, has issues with scalability
- Can we even match messages in software anymore?

Consider that future nodes might have >1000 threads!

- Synchronization overhead is likely to be extremely expensive

Exascale Challenges – Applications



How can we develop applications that are portable across all of the different hardware models?

- Porting applications can take thousands of man-hours

Even if the DOE had the money to port all of the legacy applications, we do not have enough application and algorithm developers. 

The DOE is not monolithic; it has a large base of codes

Application Code Property	Modeling and Simulation	Large Scale Data Analytics
Spatial Locality	High	Low
Temporal Locality	Moderate	Low
Memory Footprint	Moderate	High
Computation	May be FP Dominated	INT Dominated
Input-Output Oriented	Output Dominated	Input Dominated



How can applications take advantage of billion-way parallelism?

- Hardware has changed at a greater rate than software
 - How do we adapt applications to take advantage of on-node memory and compute hierarchies?

What subset of applications are even viable at Exascale?

- Programming models – what to use where?
- Mutli-physics coupling
 - Integrating software components that use disparate approaches

How do we debug a billion thread program?

- We need some automated tool assistance
- Current parallel debuggers are not up to the task
- New debuggers must be fast enough to be tolerable
 - Cannot have multi-minute latency on a single debugger step/command



NNSA ASC's Advanced Technology Development and Mitigation (ATDM) strategy follows an Application-centric Co-design path:

All Architecture R&D is not equal

- Prioritize efforts that ease the application/algorithm developer burden

All Application Development is not equal

- Prioritize approaches with synergy to a common application development eco-system

System Software Investments that support these priorities are critical to DOE Exascale co-design

Have real impact and chart the path for other applications to follow

- Means an exploratory role but also a practical context – this technology *must* work

The ECP plan of record



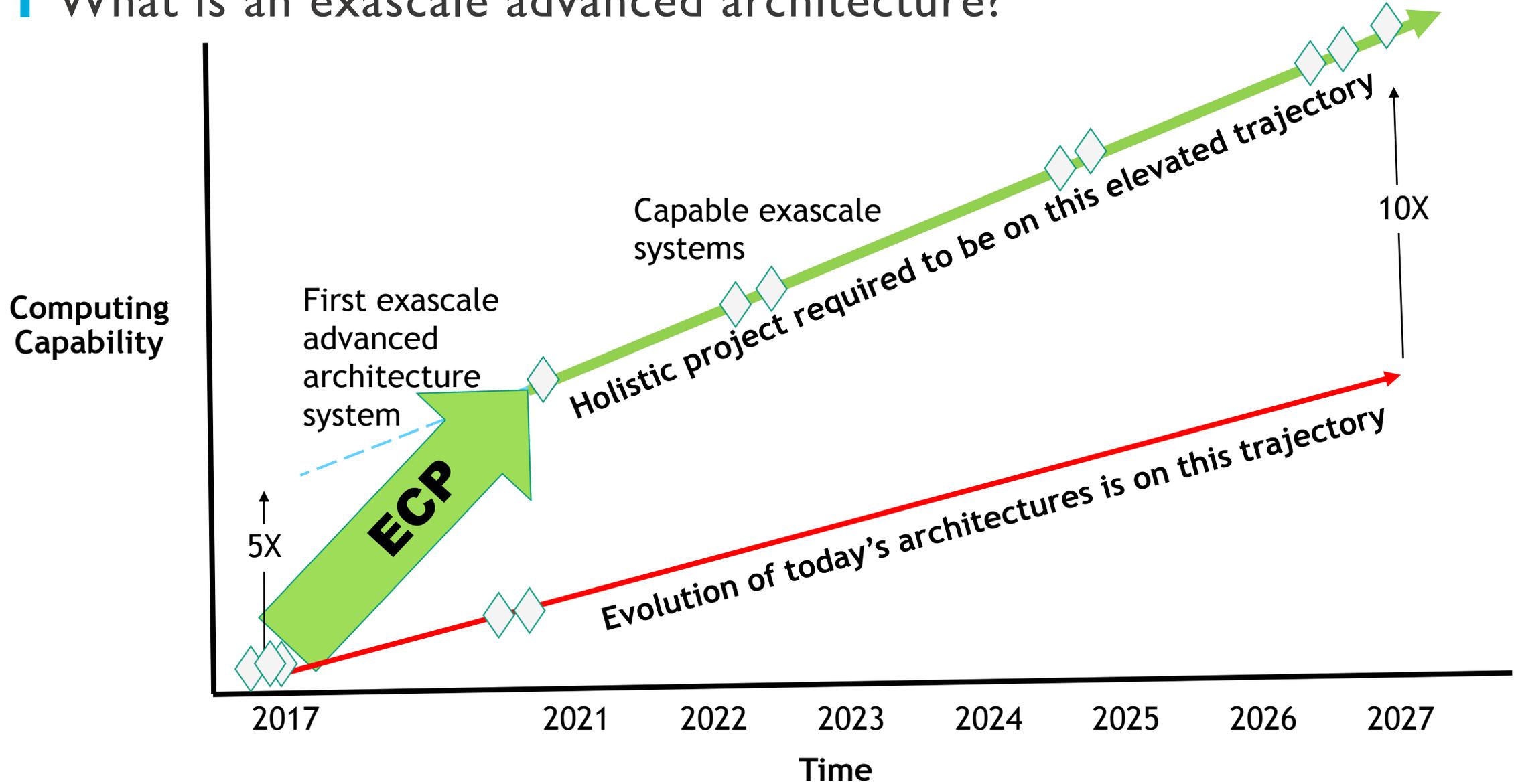
A 7-year project that follows the holistic/co-design approach, which runs through 2023 (including 12 months of schedule contingency)

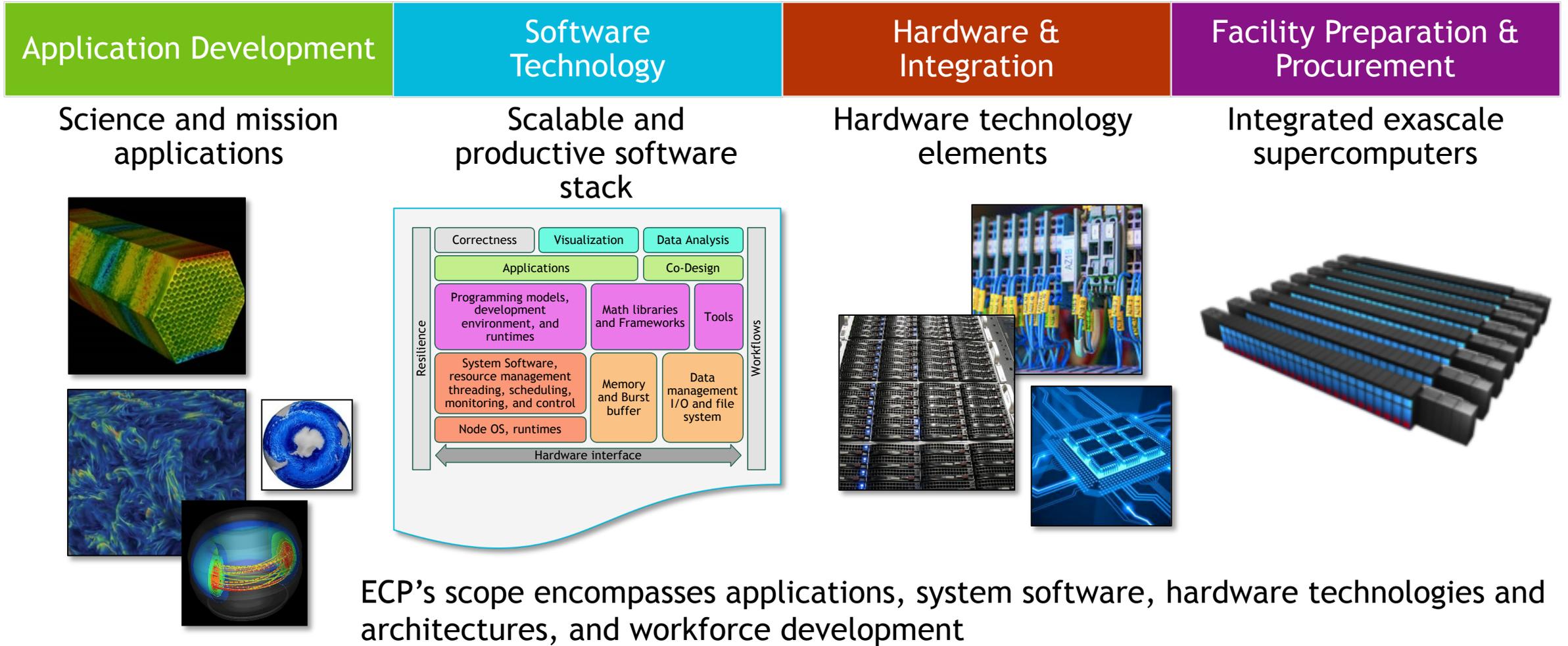
Enable initial exascale systems based on an advanced architectures and delivered in 2021

Enable capable exascale systems, based on ECP R&D, delivered in 2022 and deployed in 2023 as part of an NNSA and SC facility upgrades

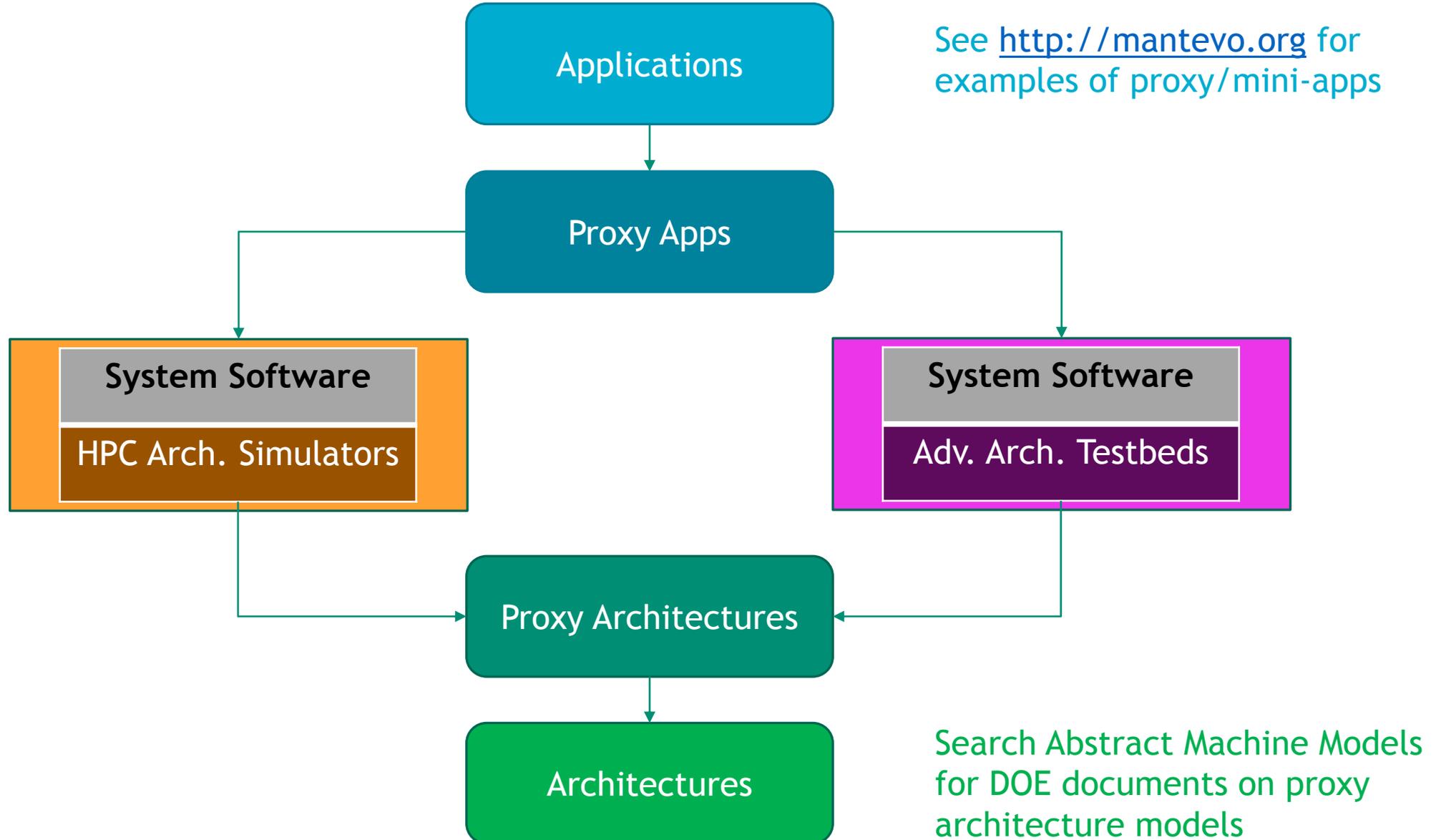
Acquisition of the exascale systems is outside of the ECP scope, will be carried out by DOE-SC and NNSA-ASC facilities

What is an exascale advanced architecture?

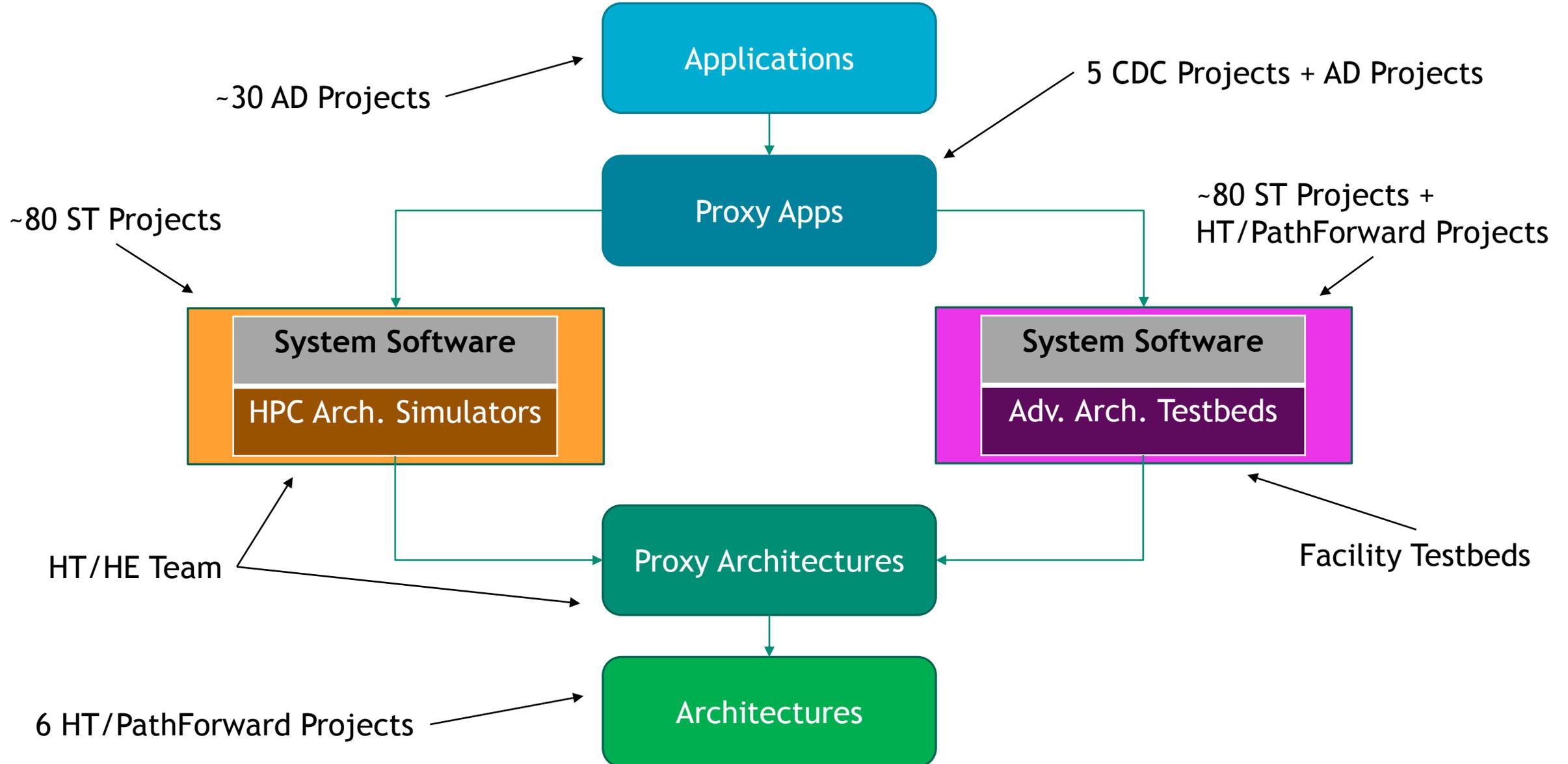




Holistic Approach – Co-design and Integration



Holistic Approach – Co-design and Integration





Exascale Software Technologies

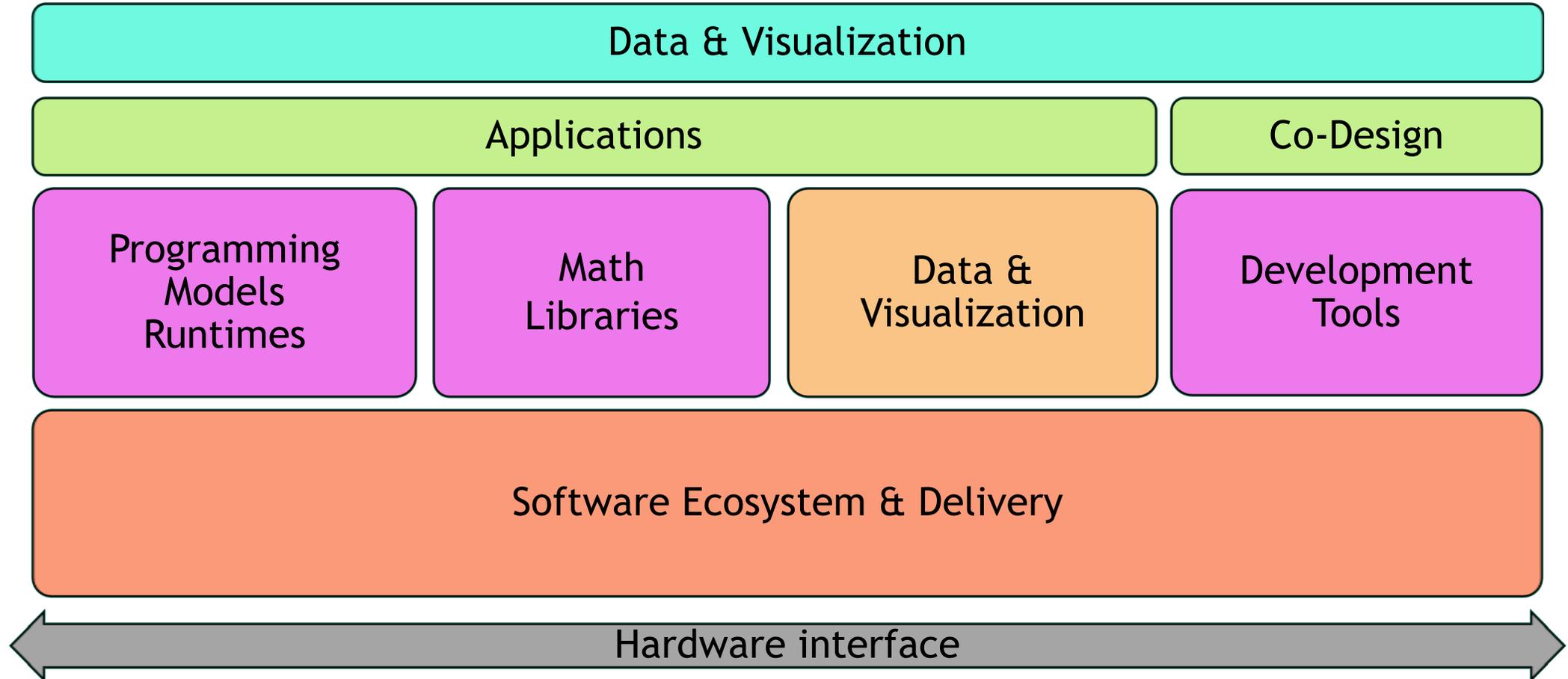


ECP will build a comprehensive and coherent software stack that will enable application developers to productively write highly parallel applications that can portably target diverse exascale architectures

ECP will accomplish this by:

- Extending current technologies to exascale where possible
- Performing R&D required to develop new approaches where necessary
- Coordinating with vendor efforts to develop and deploy high-quality and robust software products

Realistically will need plans to migrate legacy models as well as approaches for more modern techniques



Exascale Multi-Node Programming Models



Area of significant interest to DOE/ECP

Historical focus and significant use of MPI

- Deeply embedded in application portfolio and the design of algorithms
- Highly optimized software toolchains and build systems
- Broad developer expertise

But...an area open to change? New long-term strategies?

- Use of multi-node tasking runtimes?
- Greater use of one-sided operations (MPI, OpenSHMEM, PGAS, ...)
- Hardware support for global load/store semantics may change how runtimes are written



Exascale On-Node Programming Models



One of the greatest areas of change in DOE applications

- Most applications have limited use of on-node parallelism

Multiple options because of diverse hardware

- Directives (OpenMP, OpenACC, ...)
- Language runtimes (Fortran DO CONCURRENT, C++ Parallel STL, CUDA, ...)
- Template metaprogramming frameworks (Kokkos, RAJA, FleCSI, ...)
- Source-to-source translation (ROSE, Autotuners, ...)
- Task runtimes (Qthreads, Cilk, TBB, ...)

Really want to try to find common approaches across DOE portfolio

- Challenge of maintaining so many models is expensive and it is difficult to find skilled programmers



Exascale Hardware Technologies



Objective: Fund R&D to design hardware that meets ECP's Targets for application performance, power efficiency, and resilience

Establish *PathForward* (PF) Hardware Architecture R&D contracts that deliver:

- Conceptual component, node and system designs
- Analysis of performance improvement on conceptual exascale system designs
- Technology demonstrators to quantify performance gains over existing roadmaps
- Support for active industry engagement in ECP holistic co-design efforts

DOE labs engage to:

- Participate in evaluation and review of PathForward deliverables
- Lead Design Space Evaluation through Architectural Analysis, and Abstract Machine Models of PathForward designs for ECP's holistic co-design

PathForward Vendor Contracts



Competitive RFP released in June, 2016

DOE announced 6 contract awards on June 15, 2017

- Firm Fixed Price contracts with milestone deliverables and payments
- DOE Advance IP Waivers for vendors that provide $\geq 40\%$ cost share
- Project duration: 3 years
- Total DOE Investment: \$258M

Six awardees

- AMD
- Intel
- Cray
- Nvidia
- HPE
- IBM

Awarded 31 work packages across all projects

- 259 total milestones/deliverables

PathForward Summary



PathForward reduces the Technical Risk for NRE investments in the 2022/23 capable exascale system(s)

Establishes a foundation for architectural diversity in the HPC eco-system

Provides hardware technology expertise and analysis

Provides an opportunity for inter-agency collaboration

As of October, 2017, 31 Deliverables Completed

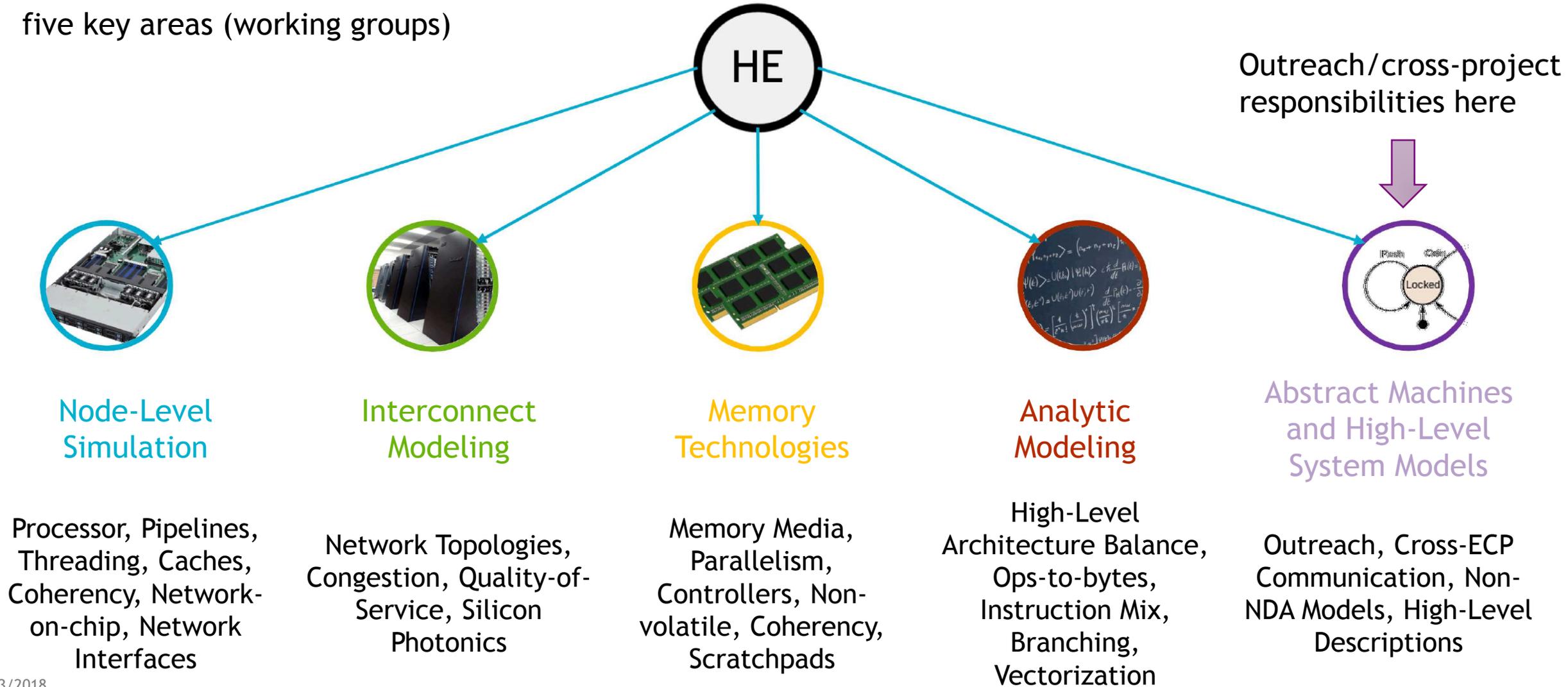
- 3 application analysis
- 5 processor design
- 8 memory
- 10 interconnect fabric
- 2 energy and power
- 3 system design and resilience



Hardware Evaluation (DOE Analysis Capability)



Hardware Evaluation covers five key areas (working groups)





Summary



Unique opportunity to do something special for the nation on a rapid time scale

- This is an exciting time to be in computing!

The advanced architecture system in 2021 affords the opportunity for

- More rapid advancement and scaling of mission and science applications
- More rapid advancement and scaling of an exascale software stack
- Rapid investments in vendor technologies and software needed for 2021 and 2023 systems
- More rapid progress in numerical methods and algorithms for advanced architectures
- Strong leveraging of and broader engagement with US computing capability



When ECP ends, we will have

- Run meaningful applications at exascale in 2021, producing useful results
- Prepared a full suite of mission and science applications for 2023 capable exascale systems
- Demonstrated integrated software stack components at exascale
- Invested in the engineering and development, and participated in acquisition and testing of 2023 capable exascale systems
- Prepared industry and critical applications for a more diverse and sophisticated set of computing technologies, carrying US supercomputing well into the future

We are always looking for new staff and new collaborations...

- Exascale requires that we draw from a diverse pool of talent across multiple areas!

Your
E C O P
EXASCALE COMPUTING PROJECT
Needs You!



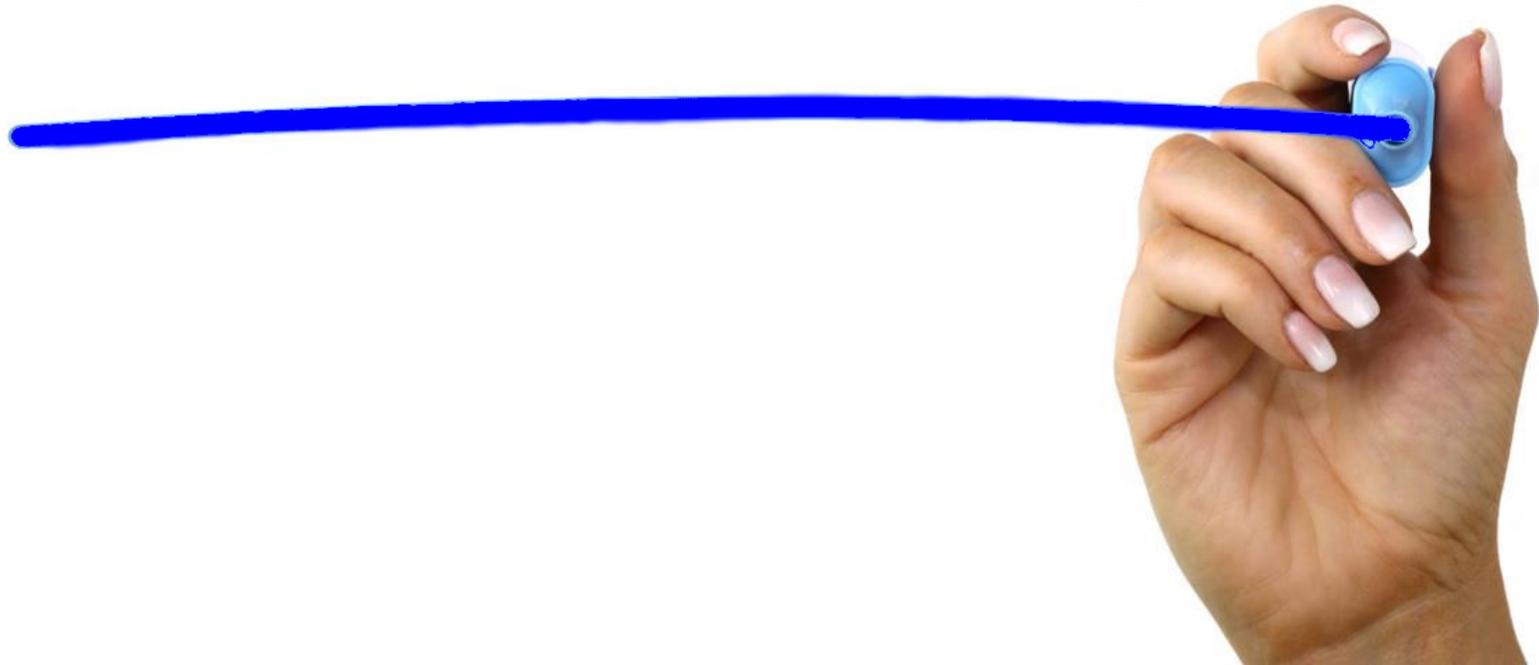


Many thanks to James Ang (SNL/PNNL) who supplied many of the slides derived from his various ECP talks

To all of the contributors of the technical reports:

- Extreme Computing: Pushing the Frontiers of Science, SAND2016-4296PE
- Exascale System and Node Architectures: The Summit and Beyond, SAND2016-5876C
- Computing Beyond Moore's Law, SAND2016-7422J
- The National Strategic Computing Initiative and Synergistic Opportunities for Massive-scale Scientific and Data-analytic Computing, SAND2017-0746PE
- The Exascale Computing Project: Strategy for System Development, SAND2017-11208PE

QUESTIONS





Backup Slides

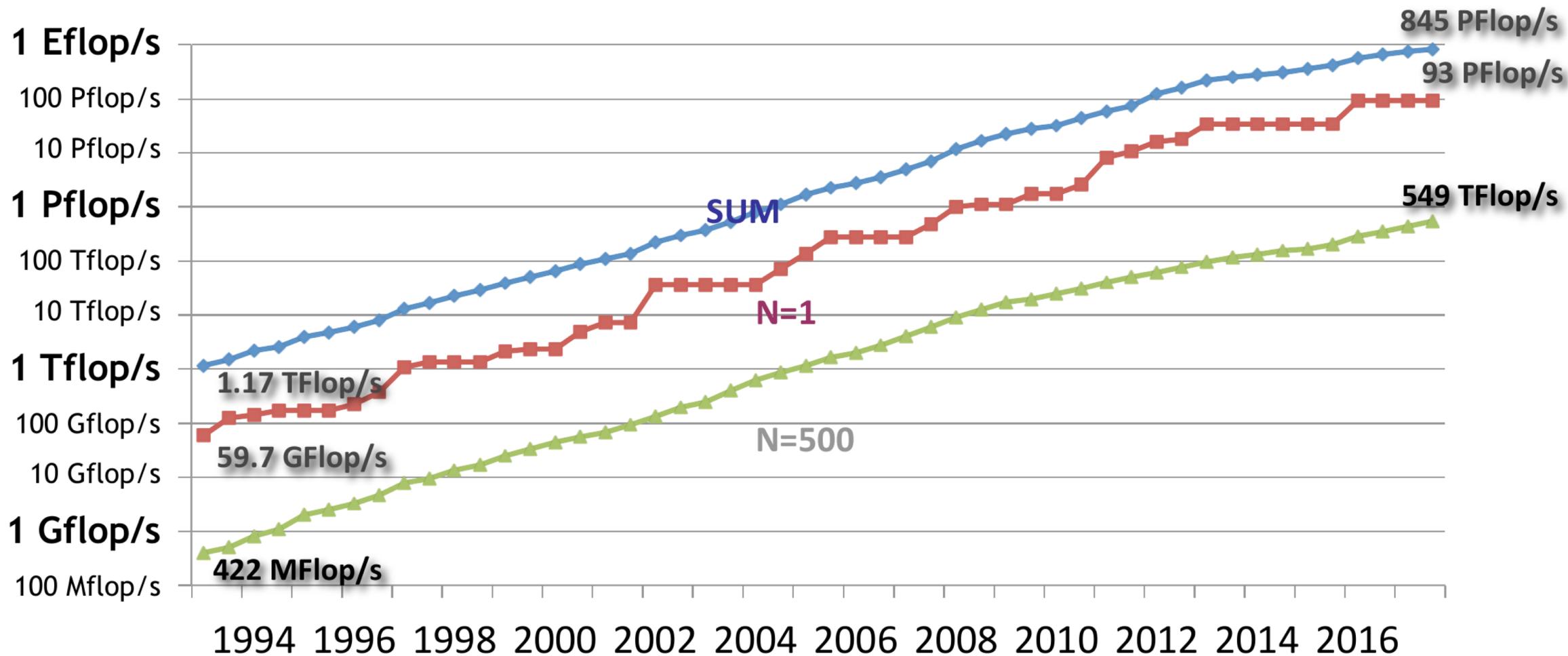


Image Courtesy of top500.org

Survey of Application Motifs



Application	Monte Carlo	Particles	Sparse Linear Algebra	Dense Linear Algebra	Spectral Methods	Unstructured Grid	Structured Grid	Comb. Logic	Graph Traversal
Cosmology									
Subsurface									
Materials (QMC)									
Additive Manufacturing									
Chemistry for Catalysts & Plants									
Climate Science									
Precision Medicine Machine Learning									
QCD for Standard Model Validation									
Accelerator Physics									
Nuclear Binding and Heavy Elements									
MD for Materials Discovery & Design									
Magnetically Confined Fusion									

Survey of Application Motifs



Application	Monte Carlo	Particles	Sparse Linear Algebra	Dense Linear Algebra	Spectral Methods	Unstructured Grid	Structured Grid	Comb. Logic	Graph Traversal
Combustion S&T									
Free Electron Laser Data Analytics									
Microbiome Analysis									
Catalyst Design									
Wind Plant Flow Physics									
SMR Core Physics									
Next-Gen Engine Design									
Urban Systems									
Seismic Hazard Assessment									
Systems Biology									
Biological Neutron Science									
Power Grid Dynamics									

Survey of Application Motifs



Application	Monte Carlo	Particles	Sparse Linear Algebra	Dense Linear Algebra	Spectral Methods	Unstructured Grid	Structured Grid	Comb. Logic	Graph Traversal
Stellar Explosions	Present	Present	Present	Present	Absent	Absent	Present	Absent	Absent
Excited State Material Properties	Absent	Absent	Absent	Present	Present	Absent	Absent	Absent	Absent
Light Sources	Absent	Absent	Present	Present	Present	Present	Absent	Absent	Absent
Materials for Energy Conversion/Storage	Absent	Present	Present	Present	Present	Absent	Absent	Absent	Absent
Hypersonic Vehicle Design	Present	Present	Present	Present	Absent	Present	Absent	Absent	Absent
Multiphase Energy Conversion Devices	Absent	Present	Present	Absent	Absent	Present	Absent	Absent	Absent

What is an exascale advanced architecture?

