CORAL: the Nation's Leadership Computers for Open Science from 2017 through 2022

and beyond

Susan Coghlan

ALCF Project Director

Argonne National Laboratory
1 September 2015

DOE LEADERSHIP COMPUTING FACILITY

Leadership Class Resources

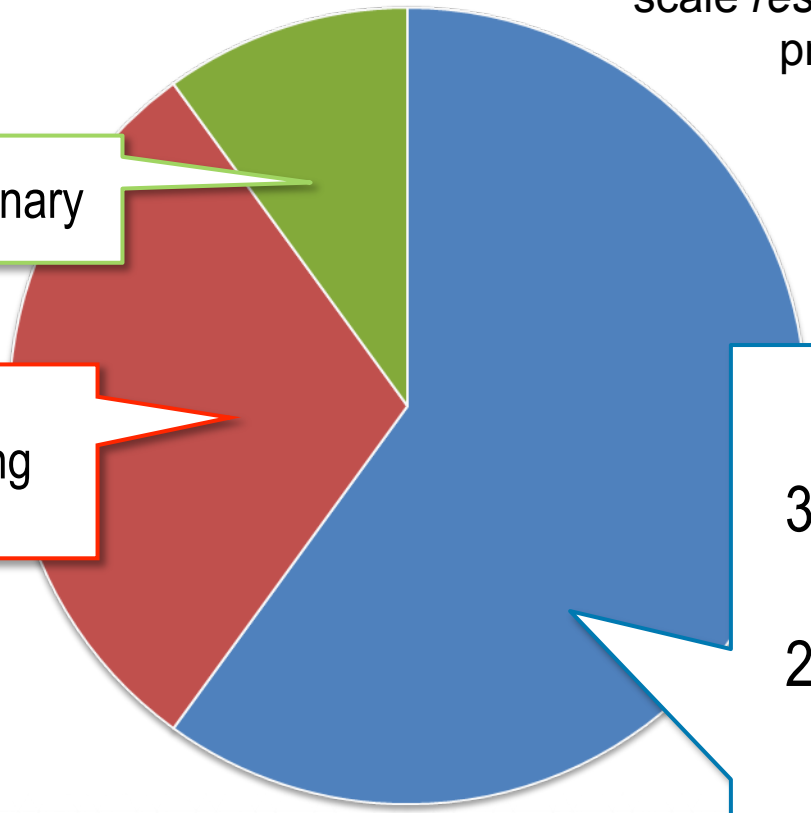
- Dedicated to high impact breakthrough open science
- Supported by DOE's Advanced Scientific Computing Research Program
- Two Centers – ALCF & OLCF
- Two of the world's most powerful supercomputers
- Two diverse architectures

THREE PRIMARY WAYS TO ACCESS LCF

DISTRIBUTION OF ALLOCABLE HOURS

Leadership-class computing

INCITE seeks computationally intensive, large-scale *research* and/or *development* projects with the potential to significantly advance key areas in science and engineering.



10% Director's Discretionary

Up to 30% ASCR Leadership Computing Challenge

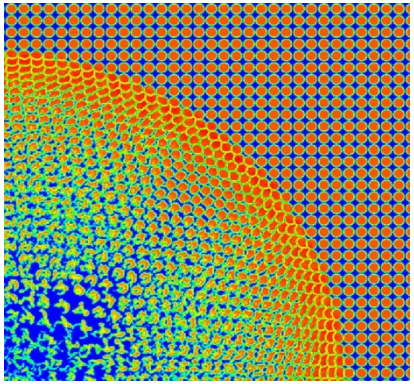
DOE/SC capability computing

60% INCITE

3.57 billion core-hours on Mira in CY2015

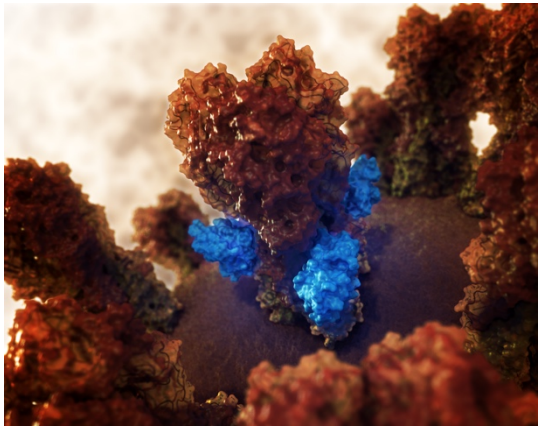
2.25 billion core-hours on Titan in CY2015

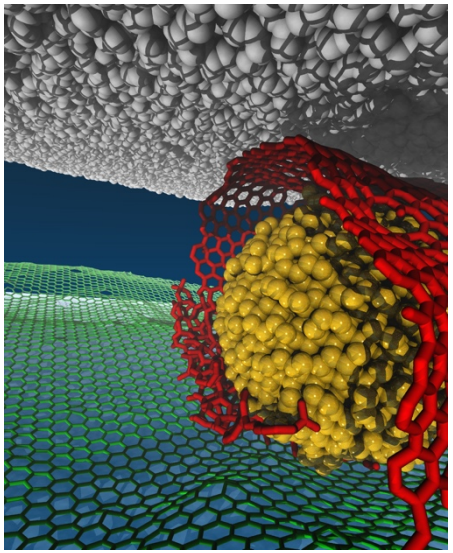
ALCF SCIENCE HIGHLIGHTS



Public Safety
Identifying the critical packing density of conventional explosives for safe transport.
M. Berzins, University of Utah

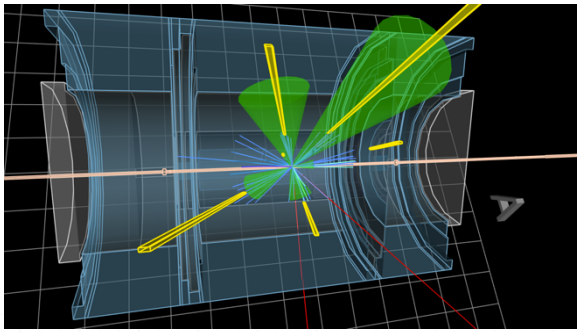
Designer Proteins
Enabling artificially designed proteins with targeted properties, including disease-fighting drugs.
D. Baker, University of Washington





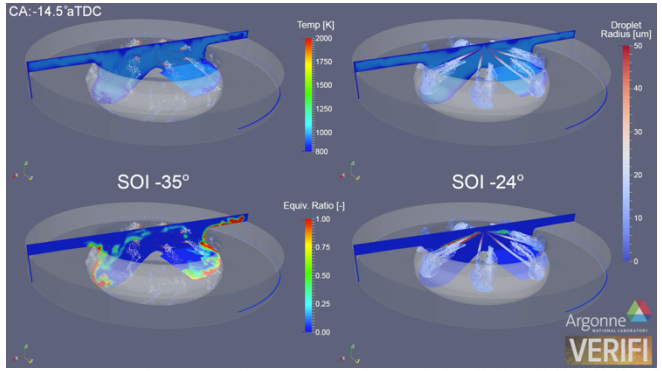
Materials Discovery
Observing a mechanism for eliminating friction at the macroscale with potential to achieve superlubricity for a wide range of mechanical applications.
S. Sankaranarayanan, Argonne National Laboratory

Superconductivity
Providing advanced calculations to identify the electronic mechanisms for high-temperature superconductivity in the cuprates.
L. Wagner, University of Illinois at Urbana-Champaign



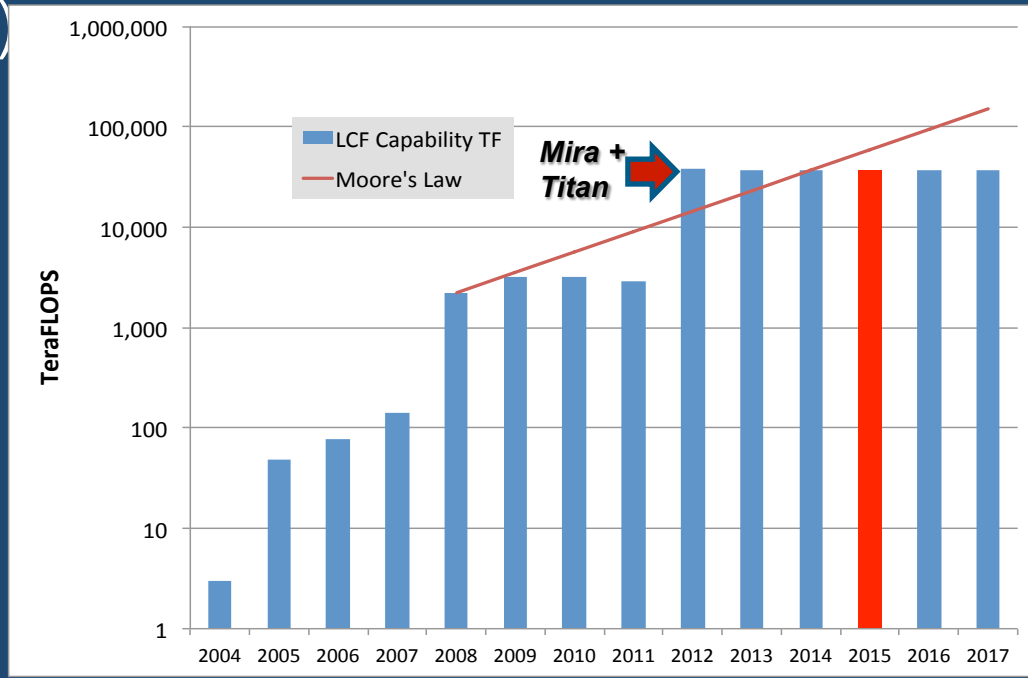
Supporting LHC
Helping to understand experiments at the LHC by routinely simulating millions of collision events in parallel.
T. LeCompte, Argonne National Laboratory

Better Engines
Partnering with industry to develop simulation capabilities that can improve the fuel economy of vehicles.
S. Som, Argonne National Laboratory

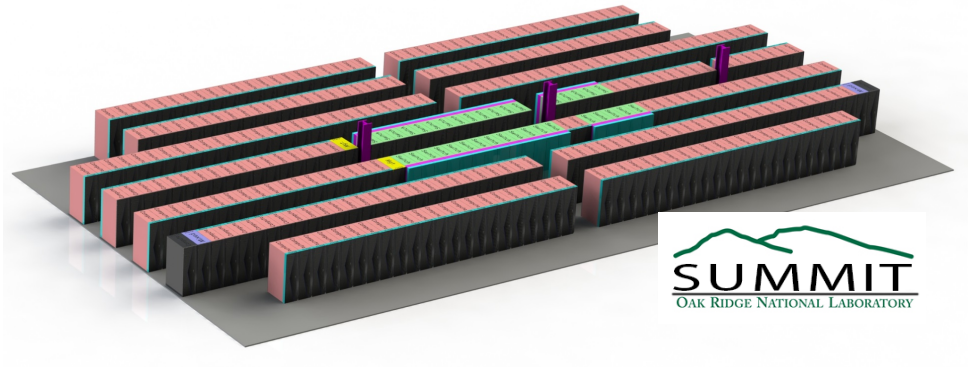


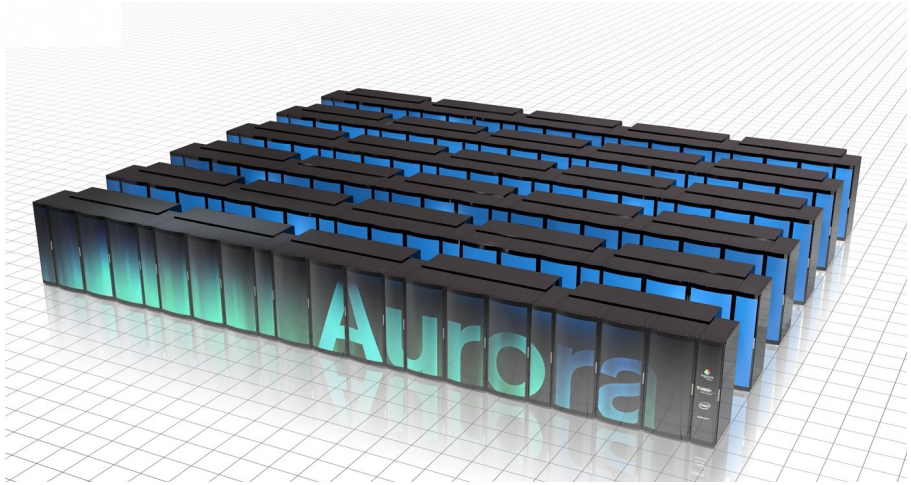
LCF SYSTEM CAPABILITY TO 2017

- Increased 10,000 x since LCF founding in 2004
- Capability use - LCF users typically utilize large fraction of system
  - 72% of hours are capability
  - 41% in largest bin (262,144 cores or more)
- Strong partnerships key to success
  - System designers
  - Science community



2018/2019 TO 2022 SYSTEMS





CORAL

COLLABORATION OF OAK RIDGE, ARGONNE, AND LIVERMORE

- Acquire DOE 2018 – 2022 Leadership Computing Capability
- Three leadership class systems – one each at ALCF, LLNL, OLCF
  - With arch diversity between ALCF and OLCF
- ALCF: Intel (Prime) Cray (Integrator)
- OLCF: IBM (Prime)
- LLNL: IBM (Prime)

APPLICATION PERFORMANCE REQUIREMENTS
HIGHEST PRIORITY TO CORAL

CORAL benchmark suite

- Representative of DOE workloads for Office of Science and NNSA
- Used to evaluate future technology options
- Provides flexibility to vendors for diverse architectures

"Figure of merit" (FOM) – average improvement over Titan and Sequoia

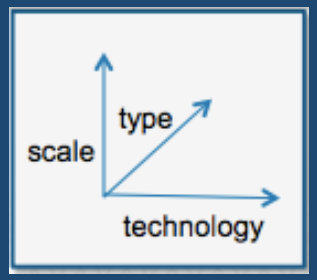
- Actual, predicted and/or extrapolated performance results for proposed system
- CORAL system performance
  - **Average FOM over four key scalable science apps >= 4.0**
  - **Average FOM over four key throughput apps >= 6.0**
  - Raw results for three **Data Centric apps** and five skeleton apps
- CORAL system extended performance
  - Proposals got a plus for presenting results for additional CORAL apps
- CORAL system micro-benchmarks performance
  - Present raw results for micro-benchmarks suite

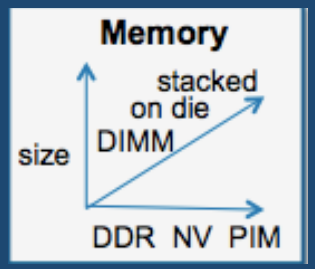
DIVERSITY DIGRESSION

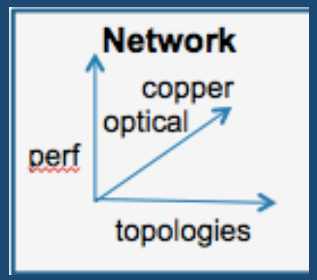
System diversity advantages

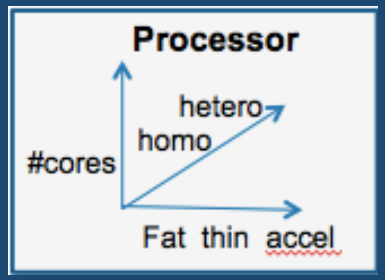
- Promotes price competition
- Promotes competition of ideas and technologies
- Promotes rich and healthy high performance computing ecosystem
- Leading HPC priority goal in DOE's strategic plan
- Reduces risk

- Ways systems can be diverse
  - Few big, many little
  - Technologies, ecosystems
- Diversity can be along different dimensions









foo

2017 OLCF LEADERSHIP SYSTEM

HYBRID CPU/GPU ARCHITECTURE

Vendor: IBM (Prime) / NVIDIA™ / Mellanox®

At least 5X Scalable Science FOM

150 PF+ peak

Approximately 3,400 nodes, each with:

- Multiple IBM POWER9™ CPUs and multiple NVIDIA Volta® GPUs. CPUs and GPUs completely connected with high speed **NVLink**
- Large coherent memory: over 512 GB (**HBM** + DDR4)
  - all directly addressable from the CPUs and GPUs
- An additional 800 GB of NVRAM, which can be configured as either a **burst buffer** or as extended memory
- over 40 TF peak performance

Dual-rail Mellanox® EDR-IB full, non-blocking fat-tree interconnect

IBM Elastic Storage (GPFS™) - 1TB/s I/O and 120 PB disk capacity.

2016 ALCF THETA SYSTEM

STEPPING STONE TO 2018 SYSTEM

System Features	Theta Details
Peak System Performance	>8.5 PetaFLOP/s
Compute Node CPU	2nd Generation Intel® Xeon Phi™ processors - **KNL**
Compute Node Count	>2,500 single socket nodes
Compute Platform	Cray XC supercomputing platform
Compute Node Peak Performance	>3 TeraFLOP/s per compute node
Cores Per Node	>60 cores
High Bandwidth Memory (HBM) On-Package	**Up to 16 Gigabytes per compute node**
DDR4 Memory	**192 Gigabytes per compute node**
SSD	**128 Gigabytes per compute node**
File System	Intel Lustre File System
File System Capacity (Initial)	10 Petabytes
File System throughput (Initial)	200 Gigabytes/s
System Interconnect	Cray Aries Dragonfly topology interconnect
Intel Arch (x86-64) Compatibility	Yes

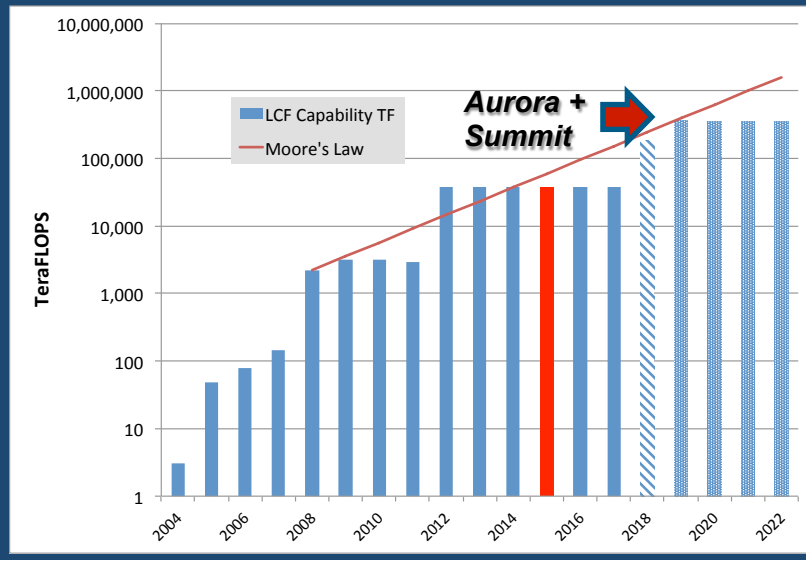
2018 ALCF LEADERSHIP SYSTEM

MANY-CORE HOMOGENOUS

System Feature	Aurora Details
Scalable Science FOM	**>7x (>13x Mira)**
Peak System performance (FLOPs)	180 - 450 PetaFLOPS
Processor	3rd gen Intel Xeon Phi processor - **KNH**
Number of Nodes	>50,000
Compute Platform	Cray Shasta next generation SC platform
HBM, local mem, persistent mem	**>7 PetaBytes**
HBM BW	**>30 PB/s**
System Interconnect	2nd gen Intel Omni-Path with silicon photonics
Interconnect Aggregate Node Link BW	>2.5 PB/s
Interconnect Bisection BW	>500 TB/s
Interconnect interface	Integrated
Burst Storage Buffer	Intel SSDs
File System	Intel Lustre File System
File System Capacity	>150 PetaBytes
File System Throughput	>1 TeraByte/s

TOP APPLICATION CHALLENGES

- Significant application challenges for these systems
  - Node concurrency increases
  - Data movement & memory scaling
  - Communication variability
- Approach
  - Partnerships with system designers – NRE funding, etc.
  - Partnerships with science community - LCF Early Science programs
  - Partnerships with library and tools community



LCF EARLY SCIENCE

PREPARING FOR SUMMIT AND AURORA SYSTEMS

LCF programs, one at ALCF and one at OLCF, to prepare the science community

- ALCF's Early Science Program (ESP)
- OLCF's Center for Accelerated Application Readiness (CAAR)

ESP and CAAR goals

- Ensure high impact science is ready to effectively use Aurora and Summit on day one
- Develop body of expert knowledge to share with the science community
- Form partnership teams of science teams, LCF staff, and system vendors to prepare science applications for optimal use of the systems

Both programs underway

- 13 CAAR Summit applications selected
- 8 Tier 1, 8 Tier 2 ESP Theta applications selected
- Aurora ESP call for proposal Q3 CY2016

ARCHITECTURE AND PERFORMANCE PORTABILITY

Application portability among ALCF, OLCF, and NERSC architectures is critical concern of ASCR

- Application developers target wide range of architectures
- Maintaining multiple code versions is difficult
- Porting to different architectures is time-consuming
- Many Principal Investigators have allocations on multiple resources
- Applications far outlive any computer system

Improve data locality and thread parallelism

- Many-core or GPU optimizations improve performance on all architectures
- Exposed fine grain parallelism transitions more easily between architectures
- Data locality optimized code design also improves portability

Use portable libraries

- Library developers deal with portability challenges
- Many libraries are DOE supported

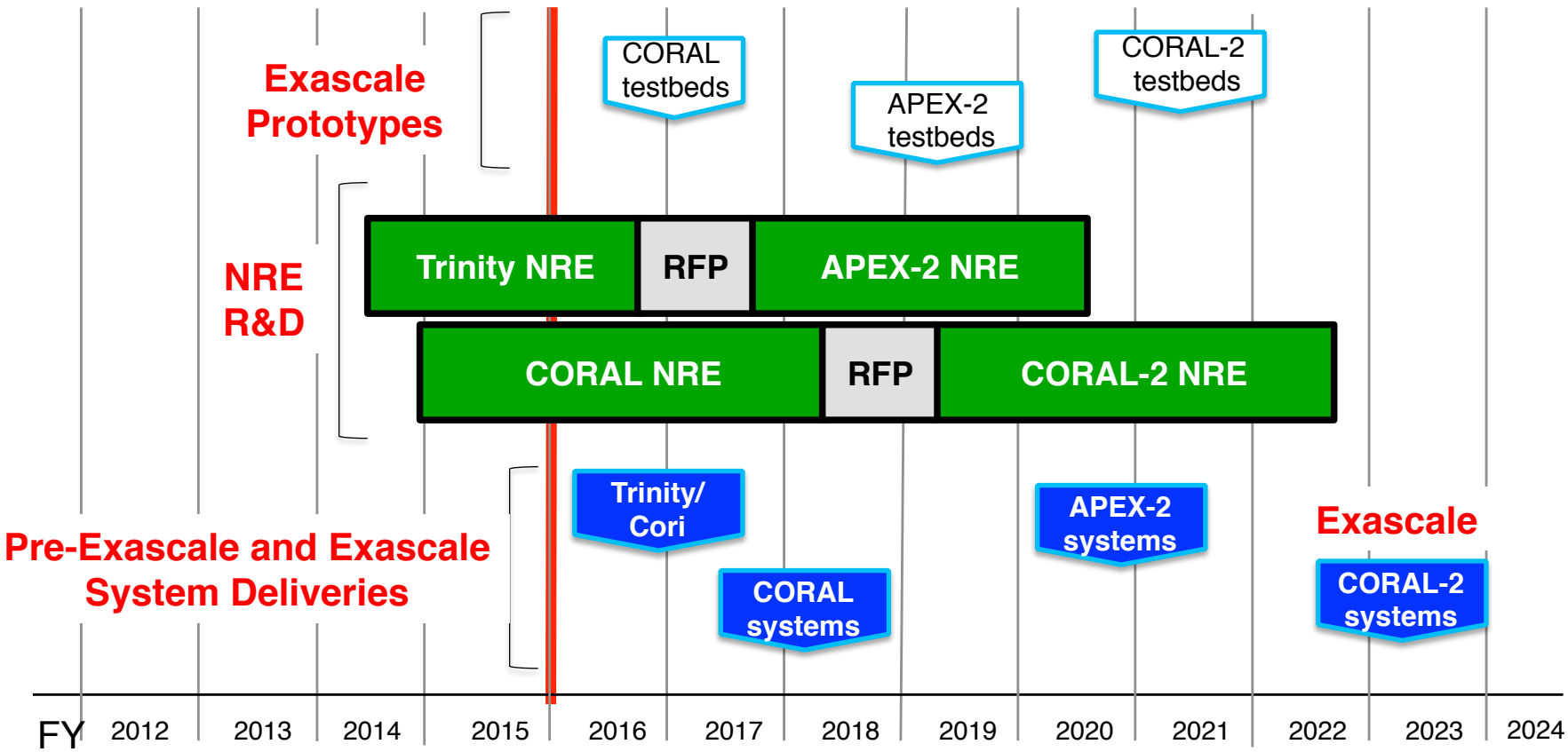
MPI+OpenMP 4.0 could emerge as common programming model

- Significant work is still necessary
- All ASCR centers are on the OpenMP standards committee

Encourage portable and flexible software development

- Use open and portable programming models
- Avoid architecture specific models such as Intel TBB, NVIDIA CUDA
- Use good coding practices: parameterized threading, flexible data structure allocation, task load balancing, etc.

EXASCALE PLATFORM DEPLOYMENT TIMELINE



CORAL = Collaboration ORNL, Argonne, LLNL

APEX = Collaboration LANL, NERSC, SNL

LOOKING TO EXASCALE – CORAL 2 TWO ARCHITECTURE PATHS

Power concerns for large supercomputers are driving the largest systems to either Hybrid or Many-core architectures

Many Core (like Aurora)

- CPU - homogeneous cores
- 10's of thousands of nodes with millions of cores
- Multiple levels of memory – on package, DDR, and non-volatile
- Coherent shared memory within node
- High-speed high-radix interconnect

Hybrid Multi-Core (like Summit)

- CPU / GPU hybrid - multiple CPUs and GPUs per node likely
- Small number of very fat nodes
- Multiple levels of memory – on package, DDR, and non-volatile
- Coherent shared memory within a node

THE BIGGEST CHALLENGES

Hardware

- Energy efficiency
- Interconnect technology
- Memory technology

Science & Software

- Scalable system software
- Programming environment
- Data management (I/O too)
- Resilience and correctness
- Exascale algorithms
- Algorithms for discovery, design, & decisions
- Scientific productivity

Top Ten Exascale Research Challenges,
DOE ASCAC Subcommittee Report, Feb 10,2014

CHANGING PRIORITIES
CULTIVATED EVOLUTION

Program Model Target	Past	Now & Future
Design Point	FLOPS (and data)	Power (and data)
Cost	FLOPS	Data movement
Memory Scaling	BW, byte/FLOP	Tiered, slower BW
Locality	MPI (+X) Off-node costs ok	MPI+X++(?) Off-node cost growing
Concurrency	Node count growth, Node concurrency slow	Node count shrinking, Node concurrency growing
Uniformity	Homogenous	Heterogeneous
Integration	Single vendor	CPU, GPU, Memory, NIC, SW, Integrator – multiple vendors
Reliability	Hardware Issue	Whose Issue is It?

WHAT'S EXASCALE LOOK LIKE?
(HYBRID CPU/GPU PATH)

				CORAL-2
Date	**2009**	**2012**	**2017**	**2023**
System	**Jaguar**	**Titan**	**Summit**	**Exascale**
System peak	2.3 Peta	27 Peta	150+ Peta	**1-2 Exa**
System memory	0.3 PB	0.7 PB	2-5 PB	**10-20 PB**
NVM per node	none	none	800 GB	**~2 TB**
Storage	15 PB	32 PB	120 PB	**~300 PB**
MTTI	days	days	days	**O(1 day)**
Power	7 MW	9 MW	10 MW	**~20 MW**
Node architecture	CPU 12 core	CPU + GPU	X CPU + Y GPU	**X loc + Y toc**
System size (nodes)	18,700	18,700	3,400	**How fat?**
Node performance	125 GF	1.5 TF	40 TF	**depends (X,Y)**
Node memory BW	25 GB/s	25 - 200 GB/s	100 – 1000 GB/s	**10x fast vs slow**
Interconnect BW	1.5 GB/s	6.4 GB/s	25 GB/s	**4x each gen**
IO Bandwidth	0.2 TB/s	1 TB/s	1 TB/s	**flat**

Al Geist: NITRD talk 4/16/2015

WHAT'S EXASCALE LOOK LIKE?
(MANY-CORE PATH)

				CORAL-2
Date	**2008**	**2012**	**2018**	**2023**
System	**Intrepid**	**Sequoia**	**Aurora**	**Exascale**
System peak	0.6 Peta	20 Peta	180+ Peta	**1-2 Exa**
System memory	0.08 PB	1.6 PB	>7 PB	**10-50 PB**
MTTI	weeks	weeks*	days	**O(1 day – 1 week)**
Peak Power	2 MW	9.6 MW	~13 MW	**~25 MW**
Node architecture	CPU 4 core	CPU 16 core	CPU >72 core	**CPU + ?**
System size (nodes)	40,960	98,304	>50,000	**~100K**
Node performance	13.6 GF	204.8 GF	-	**O(10 TF)**
Node memory BW	13.6 GB/s	42.5 GB/s	-	**50x fast vs slow**
Interconnect BW	5.1 GB/s	40 GB/s	-	**4x each gen**
Storage	6 PB	50 PB	>150 PB	**~1000 PB**
IO Bandwidth	80 GB/s	1TB/s	>1 TB/s	**flat**

* Mira MTTI

QUESTIONS?