BEYOND LEADING ORDER CALCULATIONS ON HPC

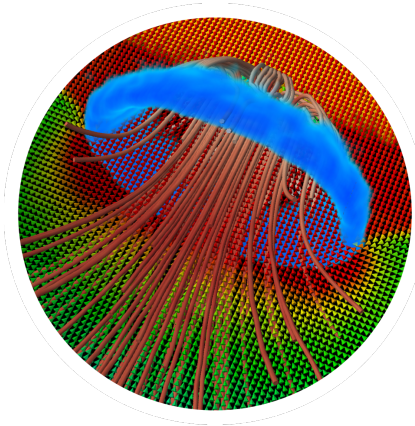
Argonne NATIONAL LABORATORY

EFFECTIVELY TARGETING THE ARGONNE LEADERSHIP COMPUTING FACILITY

KALYAN KUMARAN
Deputy Director of Science – Advanced Technologies & Data
Co-Lead, Joint Laboratory for System Evaluation (JLSE)

September 22, 2016

WHAT DOES ALCF DO?



We deliver cycles to support computational science

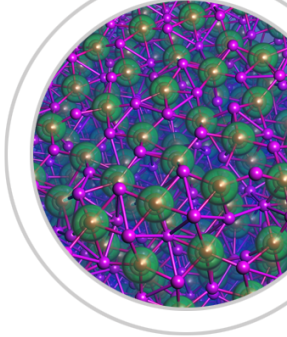
- Delivers billions of core hours of compute time
- Overall availability for the resource exceeds 96%

We partner with community to produce science

- ALCF provides expert computational scientists, called Catalysts, to assist the science teams to ready their codes to efficiently use the resources.
- ALCF provides performance engineering, full user support, and data analysis and visualization services to the science teams.
- Last year, ALCF-supported research resulted in 160+ refereed publications, in journals such as *Proceedings of the National Academy of Sciences*, *Nature*, and *Physical Review Letters*.

We partner with community on R&D in hardware and software

HOW TIME ON DOE LEADERSHIP COMPUTING SYSTEMS IS AWARDED



Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program

- Open to any researcher in the world
- Allocates time on ALCF's IBM BG/Q (Mira) and OLCF's Cray XK7 (Titan)
- Approximately **60 percent** of LCF resources are allocated through INCITE

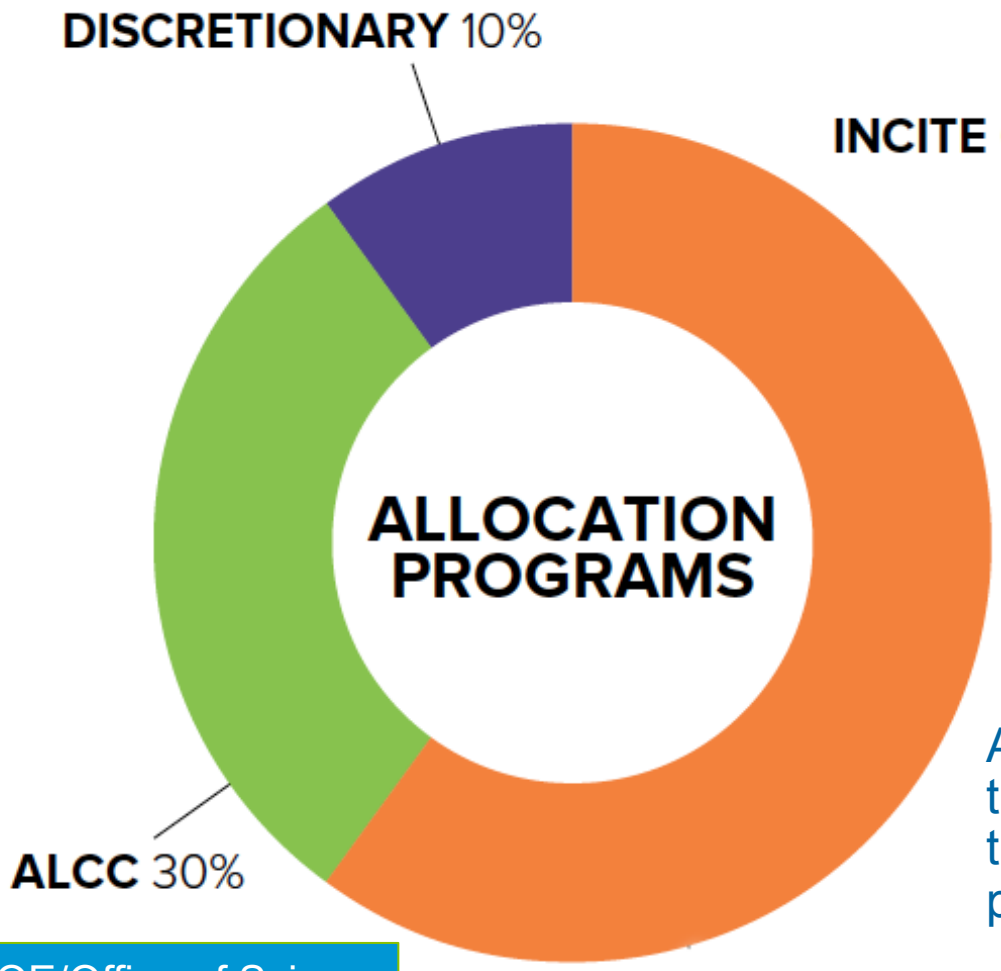
ASCR Leadership Computing Challenge (ALCC) program

- Emphasis on high-risk, high-reward simulations in areas directly related to DOE's energy mission, national emergencies, or for broadening the user base
- Allocates up to **30 percent** of the computational resources at ALCF, NERSC, and OLCF

Director's Discretionary

- Open to researchers in academia and industry
- Primarily a "first step" for projects working toward an INCITE or ALCC award
- Allocates up to **10 percent** of the computational resources at ALCF

LCF AWARD PROGRAM BREAKDOWN



5.9 billion core-hours delivered in CY2015
(3.53B delivered by ALCF)

A breakdown of how computing time on Mira is allotted among the three primary allocation programs.

DOE/Office of Science capability computing

Allocation Programs at the LCFs

LCF Allocation Programs	INCITE 60%		ALCC 30%		Director's Discretionary 10%	
Mission	High-risk, high-payoff science that requires LCF-scale resources		High-risk, high-payoff science aligned with DOE mission		Strategic LCF goals	
Call	1x/year – (Closes June)		1x/year – Closes February		Rolling	
Duration	1-3 years, yearly renewal		1 year		3m,6m,1 year	
Typical Size	30 - 40 projects	75M – 500M core-hours/yr.	10-20 projects	10M – 300+M core-hours/yr.	~100 of projects	.5M – 10M core-hours
Total Hours	~5 billion core-hours (~3.5B ALCF)		~2.5 billion core-hours (~1.75 ALCF)		~590 million ALCF	
Review Process	Scientific Peer-Review	Computational Readiness	Scientific Peer-Review	Computational Readiness	Strategic impact and feasibility	
Managed By	INCITE management committee (ALCF & OLCF)		DOE Office of Science		LCF management	
Readiness	High		Medium to High		Low to High	
Availability	Open to all scientific researchers and organizations *Capability > 131,072 cores (16.7% of Mira)*					

ALCF-2 PRODUCTION SYSTEMS

Mira – IBM BG/Q	Cetus – IBM BG/Q	Vesta – IBM BG/Q	Cooley - Cray/NVIDIA
– 49,152 nodes	– 4,096 nodes	– 2,048 nodes	– 126 nodes (Haswell)
– 786,432 cores	– 65,536 cores	– 32,768 cores	– 1512 cores
– 786 TB RAM	– 64 TB RAM	– 32 TB RAM	– 126 Tesla K80
– 10 PF	– 836 TF	– 419 TF	– 48 TB RAM (3 TB GPU)

Storage

HOME: 1.44 PB raw capacity
SCRATCH:

- fs0 - 26.88 PB raw, 19 PB usable; 240 GB/s sustained
- fs1 - 10 PB raw, 7 PB usable; 90 GB/s sustained
- fs2 (ESS) - 14 PB raw, 7.6 PB usable; 400 GB/s sustained (not in production yet)

TAPE: 21.25 PB of raw archival storage [17 PB in use]

ALCF TEAMS OF EXPERTS

Catalyst – work with the project teams to maximize and accelerate their research

Performance Engineering – optimize the science applications

Operations – support all aspects of HPC hardware and software

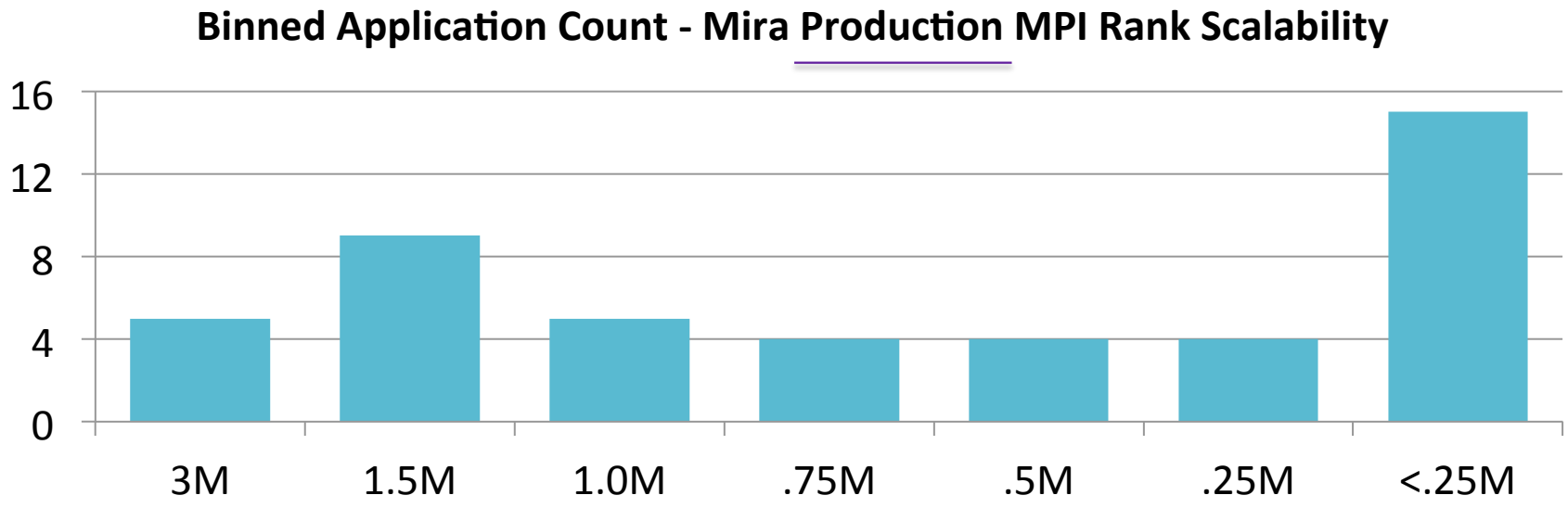
Data Analytics and Visualization – help to explore the results

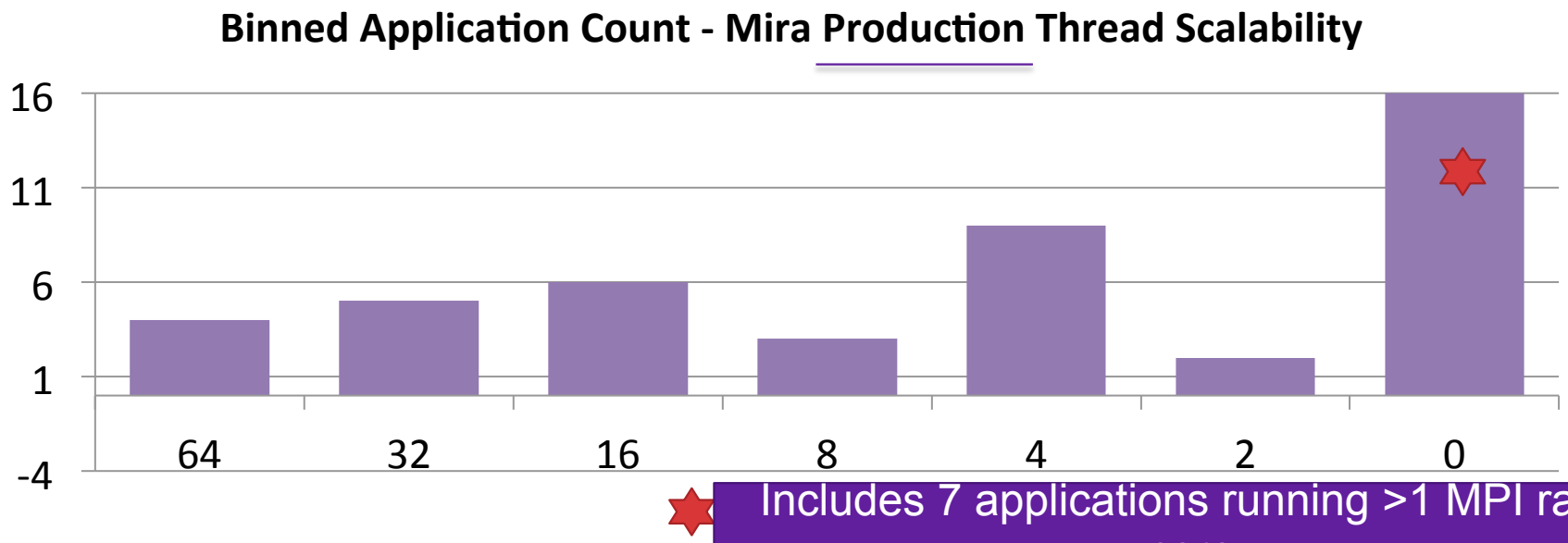
User Experience – support users and tell their stories

APPLICATIONS STATUS ON MIRA

- Summary: applications from all domains have been able to exploit high levels of concurrency for their productions runs, using MPI plus OpenMP for threading or MPI only
- Examined characteristics of 46 INCITE applications in the first 14 months since Mira production started
  - Approximately 59 INCITE codes; did not include some codes due to lack of complete information or because they played minor role in project
- Information gathered on codes configured for runs at production scale
  - Not benchmark runs or scaling exercises
- Reporting ‘threads’ generically on 30 apps - 65% of apps
  - 26 projects used OpenMP for threading
  - 4 used pthreads (at 32 or 64 threads/rank)

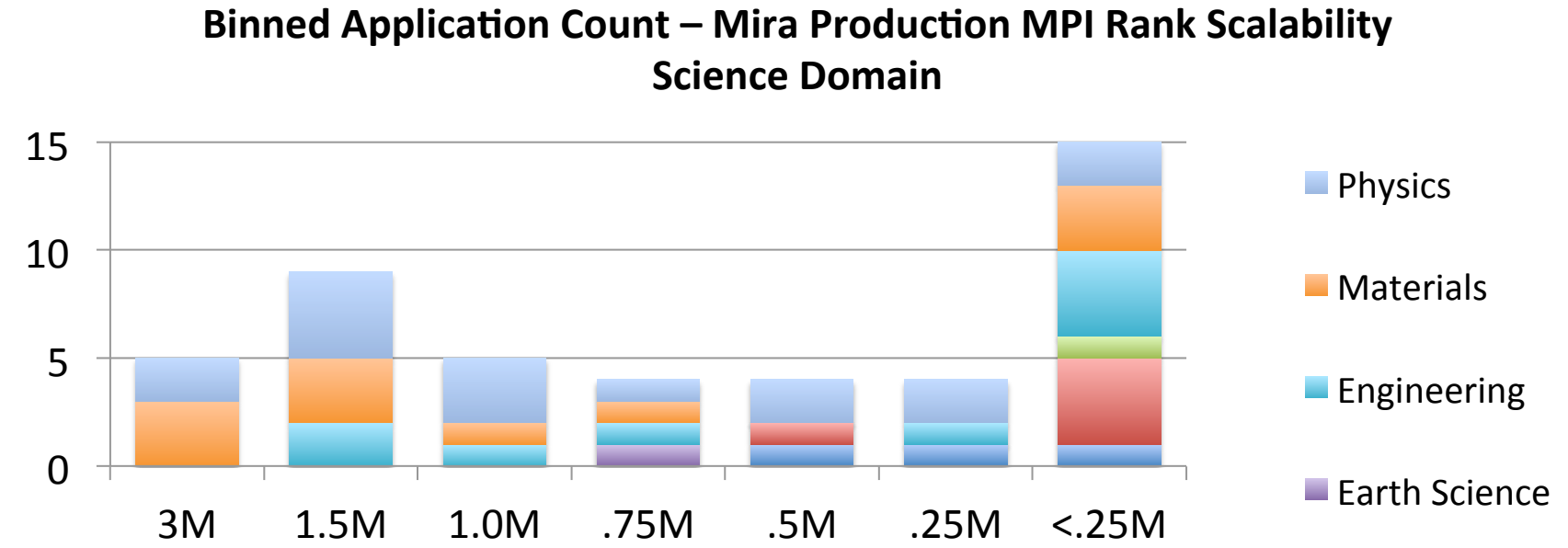
APPLICATIONS REVEAL A LOT OF SCALABILITY

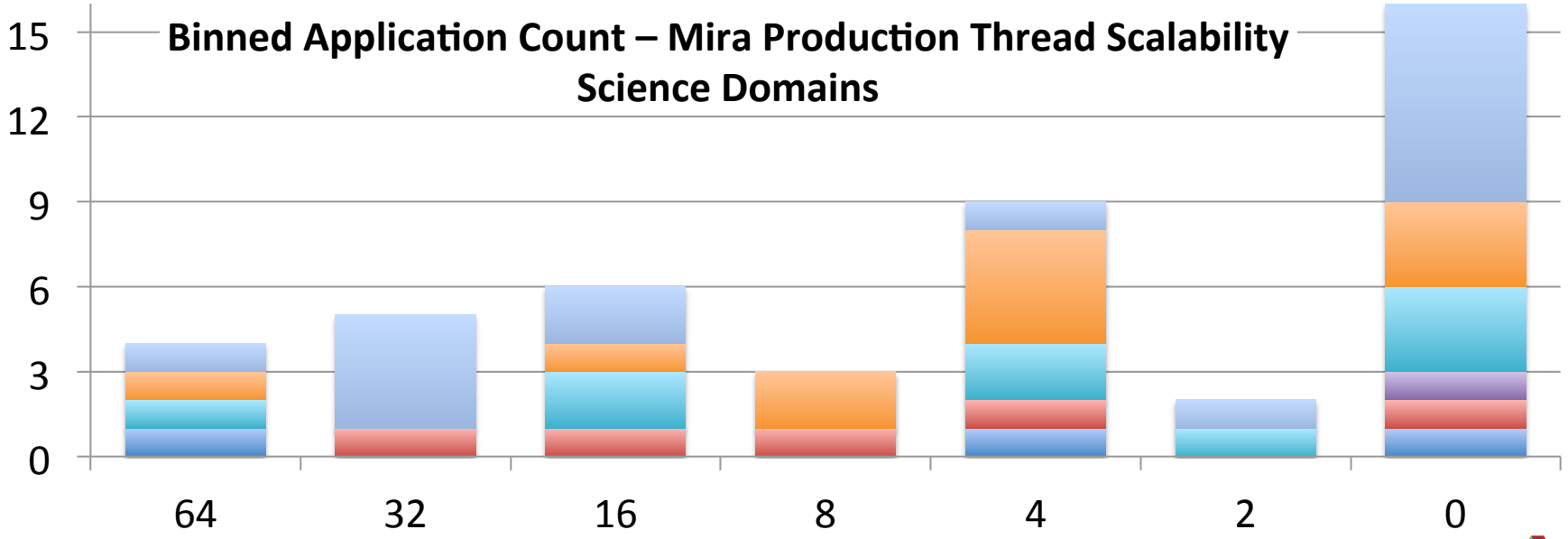




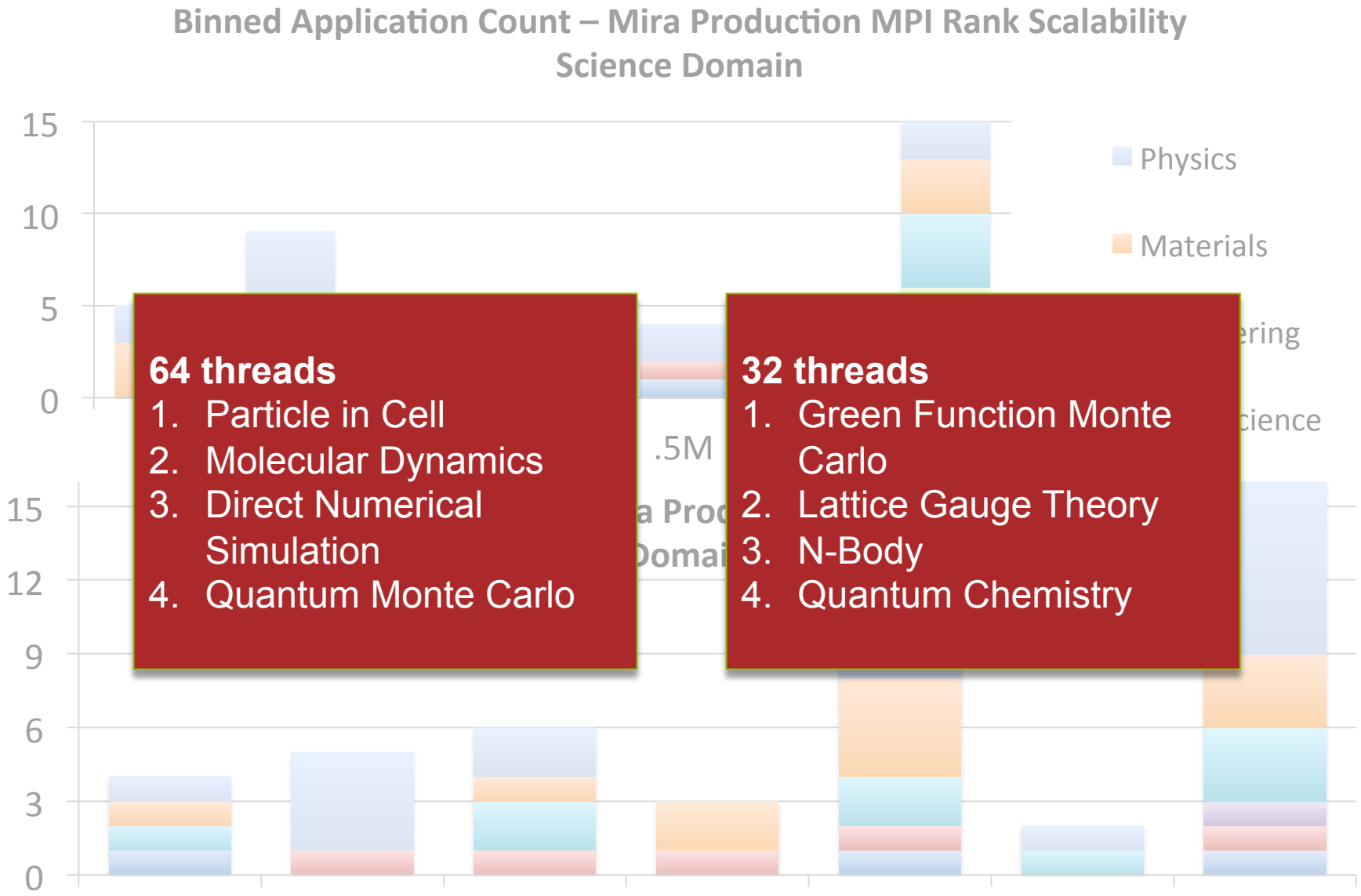
Includes 7 applications running >1 MPI rank/ core.

NO APPLICATION DOMAIN DOMINATES

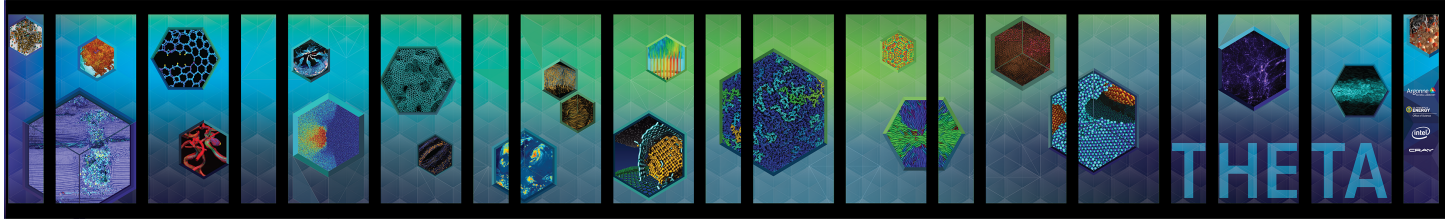




NO APPLICATION DOMAIN DOMINATES



THETA

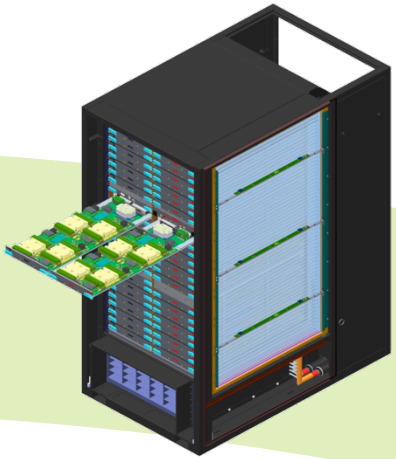


Theta is the first of two pre-exascale systems coming to Argonne

- Serves as a bridge between Mira and Aurora, transition and data analytics system
- Vendor: Intel (Prime) / Cray (Integrator)

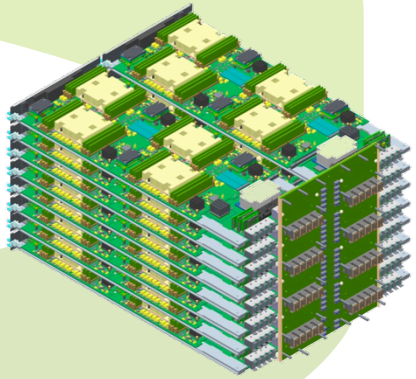
System	Config	Comments
KNL Compute Nodes	3240	7230 SKU, 64cores 1.3GHz
KNL DDR4 Memory	607.5 TB	192 GB/node DDR4-2400
KNL MCDRAM Memory	50.6 TB	16 GB/node
SSD Capacity	414.7 TB	128 GB/node
Aries Global Links	336 Optical Links	
Aries Bi-section BW	7.2 TB/s Bi-directional	
Racks	18	
Lustre LNET Nodes	30	
DVS – GPFS Nodes	60	

CRAY XC40 BUILDING BLOCKS



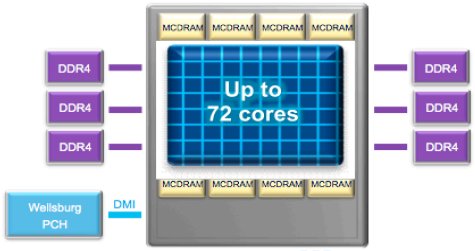
Cabinet: 3 Chassis, 129 kW liquid cooled
Local Group 384 Nodes
510.72 TF 3TB IPM, 36TB DRAM

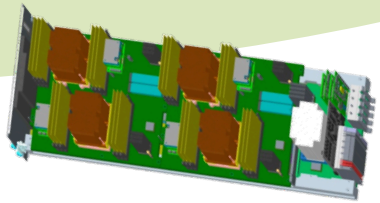
System: 18 Cabinets
3240 Nodes, 810 Switches
Dual-plane, 9 Groups, Dragonfly 7.2 TB/s Bi-Sec
8.62 PF 50.6 TB IPM, 607.5 TB DRAM



Chassis: 16 Blades, 16 Cards
64 Nodes, 16 Switches
192 TF 1TB IPM, 12TB DRAM

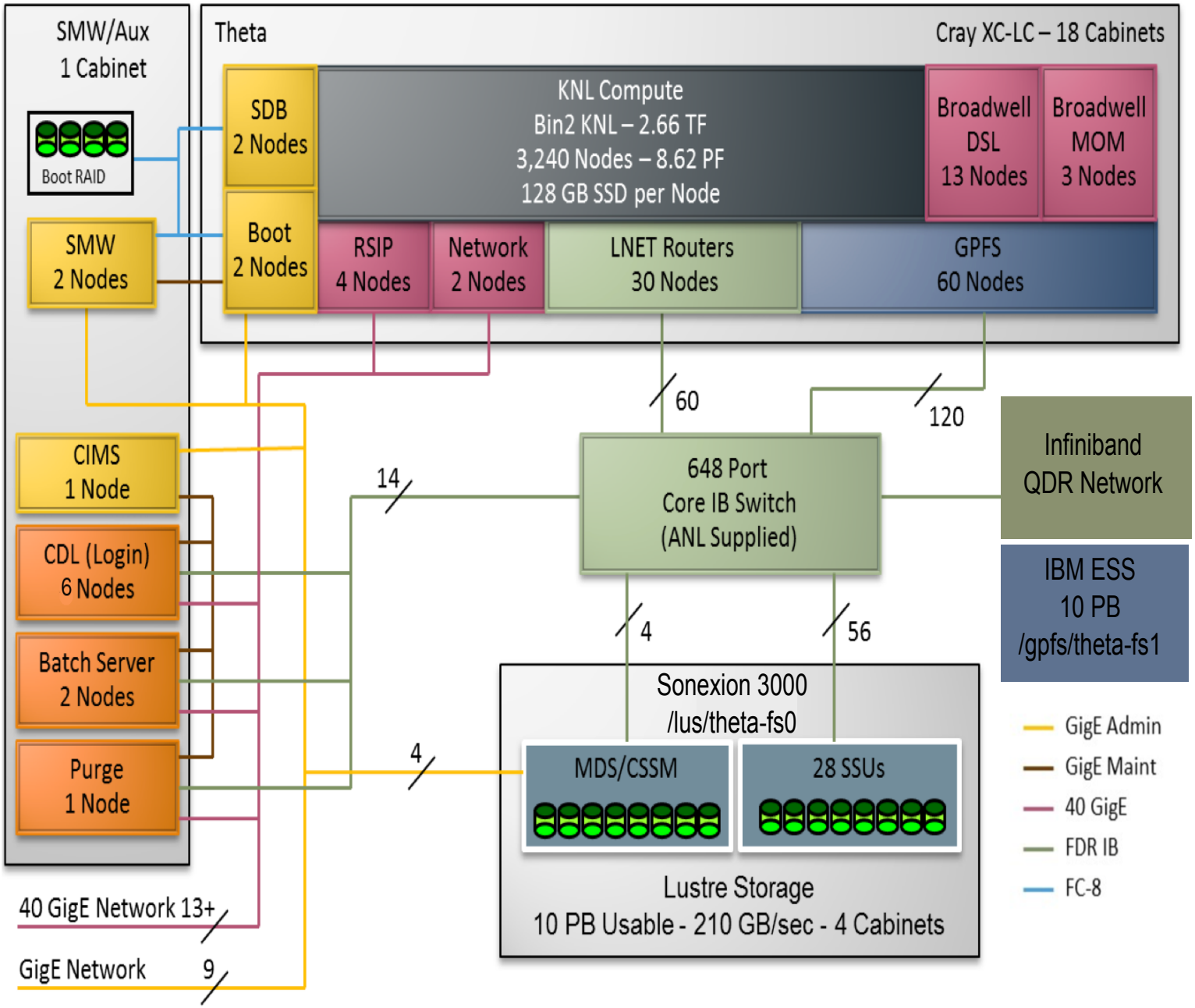
Compute Blade:
4 Nodes/Blade + Aries switch
128GB SSD
10.64 TF 64GB IPM, 768GB DRAM





Node: KNL Socket
192 GB DDR4 (6 channels)
2.66 TF 16GB IPM

THETA SYSTEM

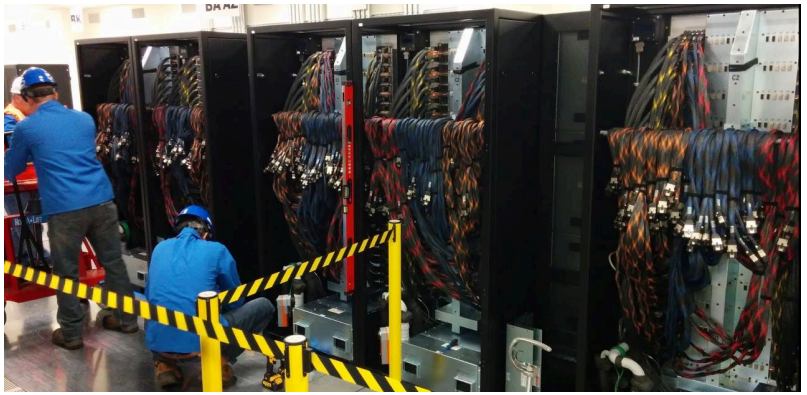


THETA SSD

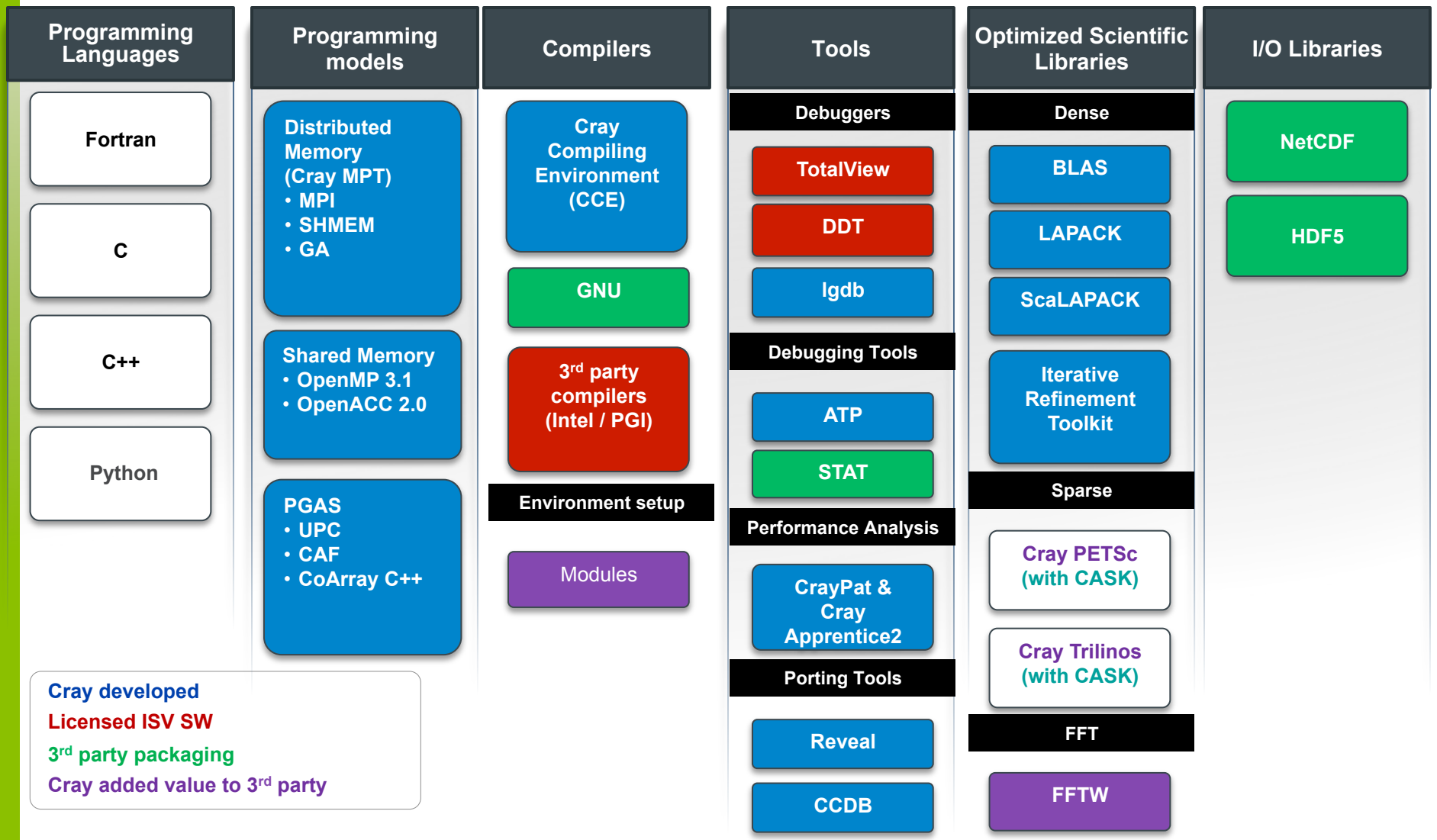
Node local storage

- Theta compute nodes contain a single SSD with a raw capacity of 128 GB
- A local volume is presented to the user as an ext3 system on top of an LVM volume
- Userspace applications can access the SSD via standard POSIX APIs
- The final capacity available to the end user is still TBD
- The ALCF is experimenting with different programming models for the SSD
- Example Use Cases
  - Temporary files used during execution but not saved
  - Dynamic Libraries
  - File system communication for multiple application workflows

THETA



CRAY PROGRAMMING ENVIRONMENT



PORTING FOR PERFORMANCE

Enable High Performance Computational Science

Debuggers

- Allinea *(C. January)*
- RogueWave *(J. DelSignore, C. Schneider, S. Lawrence)*

Performance Tools

- HPCToolkit *(J. Mellor-Crummey, Rice U.)*
- TAU *(S. Shende, ParaTools)*
- PAPI *(H. McCraw, UTK)*

Compilers

- *LLVM* *(Hal Finkel, ALCF)*

OS

- Argo *(K. Iskra, K. Yoshii, ANL)*

Libraries

- PetSC *(B. Smith, ANL)*
- Elemental *(J. Poulson, Stanford)*
- Intel MKL ScaLAPACK *(Intel)*
- LIBXSMM *(Intel)*
- ELPA *(Intel)*
- NWChem packages *(Intel)*

Programming Models

- MPICH *(P. Balaji, ANL)*
- BerkelyUPC *(P. Hargrove, Y. Zheng, LBNL)*
- ARMCI-MPI *(Intel)*
- CommAgent MPI *(Intel)*
- EP-lib *(Intel)*

I/O

- GLEAN *(V Vishwanath, ANL)*
- *ADIOS* *(S. Klasky, N. Podhorszki, Q. Liu, H. Abbasi, J. Choi, ORNL; M. Parashar, Rutgers)*
- HDF5 *(Q. Koziol , M. Chaarawi, J. Soumagne, S. Breitenfeld, N. Fortner, UIUC)*
- Mercury *(R. Ross, ANL)*
- MPI *(R. Latham, ANL)*
- Darshan *(P. Carns, ANL)*
- Parallel I/O library (PIO) *(J. Edwards, J. Dennis, NCAR; J. Krishna, ANL)*

Other

- Model Coupling Toolkit *(R. Jacob, ANL)*
- Common Infrastructure for Modeling the Earth (CIME) *(M. Vertenstein, J. Edwards, NCAR; R. Jacob, ANL)*
- Shifter *(S. Canon, D. Jacobsen, NERSC)*

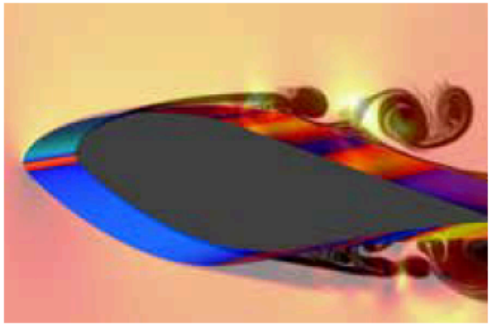
THETA EARLY SCIENCE PROGRAM

- Part of the process of bringing a new machine into production
- Based on the successful Mira ESP
- Brings together computational scientists, code developers, and computer hardware experts
- Optimizes key applications for Theta
- Solidifies libraries and infrastructure
- Enables breakthrough science and engineering research

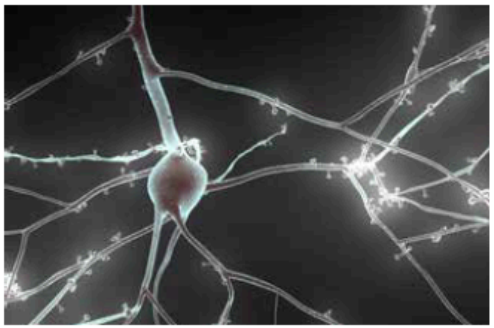
Six “Tier One” computational science projects

- Selected by peer review
- Represent a range of scientific areas and numerical methods.
- **Prepares Theta for science on day one!**

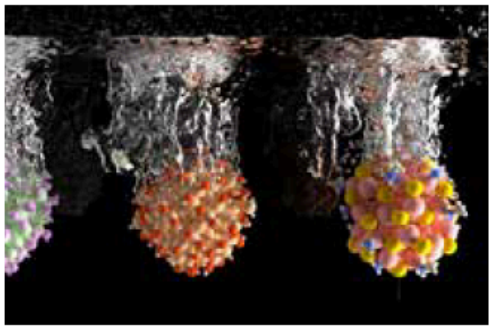
THETA'S EARLY SCIENCE PROJECTS



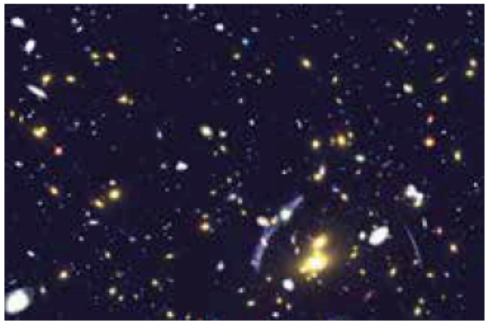
Scale-Resolving Simulations of Wind Turbines with SU2
Juan J. Alonso
Stanford University
Code: SU2



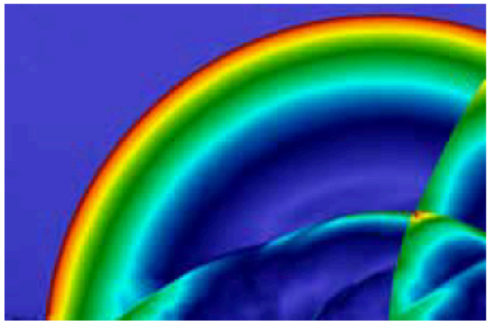
Large-Scale Simulation of Brain Tissue: Blue Brain Project, EPFL
Fabien Delalondre
EPFL
Code: CoreNeuron



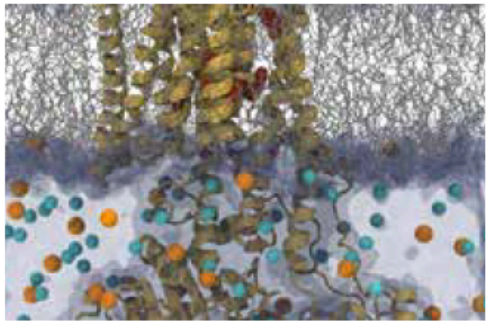
First-Principles Simulations of Functional Materials for Energy Conversion
Giulia Galli
The University of Chicago
Codes: Qbox, WEST



Next-Generation Cosmology Simulations with HACC: Challenges from Baryons
Katrin Heitmann
Argonne National Laboratory
Code: HACC



Direct Numerical Simulations of Flame Propagation in Hydrogen-Oxygen Mixtures in Closed Vessels
Alexei Khokhlov
The University of Chicago
Code: HSCD



Free Energy Landscapes of Membrane Transport Proteins
Benoît Roux
The University of Chicago
Code: NAMD

KNL 7210 node at JLSE

HACC FULL APPLICATION: 5X CORE-TO-CORE

320^3: 16 cores, 1 node of BG/Q, ¼ of the node of KNL

Cores	RPN	OMP	TH	BG/Q Time, s	KNL B0, Time, s	Ratio
16	4	4	16	4297	616.3308	6.98
16	4	8	32	2677	543.7294	4.92
16	4	16	64	2504	530.2267	4.72
16	8	2	16	4362	544.7519	8.00
16	8	4	32	2571	459.5265	5.59
16	8	8	64	**2278**	**437.2058**	5.18
16	16	4	64	2581	468.5037	5.50

512^3: 64 cores, 4 nodes of BG/Q, 1 node of KNL

Cores	RPN	OMP	TH	BG/Q Time, s	KNL B0, cache mode Time, s	KNL B0, flat mode Time, s	Ratio
64	16	4	64	4542	678.7571	678.2269	6.69
64	16	8	128	2823	606.1815	609.2007	4.66
64	16	16	256	2556	587.2716	587.4443	4.35
64	32	2	64	4747	620.7261	621.2356	7.65
64	32	4	128	2824	536.1650	534.9907	5.27
64	32	8	256	**2503**	**503.0927**	501.8637	4.98
64	64	4	256	2539	510.3745	506.7107	4.98

ALCF DATA SCIENCE PROGRAM (ADSP)

- "Big Data" science problems that require the scale and performance of leadership computing resource
- Projects will cover a wide variety of application domains that span computational, experimental and observational sciences
- Focus on data science techniques including but not limited to statistics, machine learning, deep learning, UQ, image processing, graph analytics, complex and interactive workflows
- Two-year proposal period and will be renewed annually. Proposals will target **science** and **software technology** scaling for data science
- Proposal Deadline: June 3, 2016 (Expected yearly call for proposals) https://www.alcf.anl.gov/alcf-data-science-program

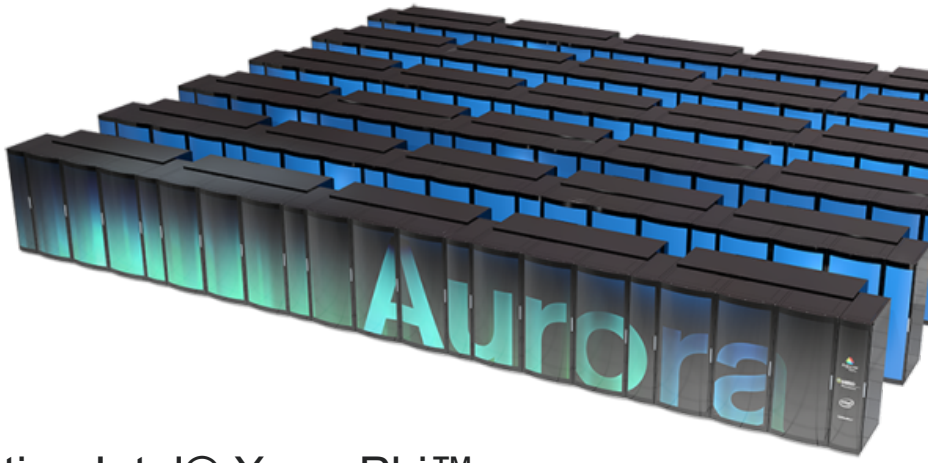
SUPPORT

- Funded postdoctoral appointee
- ALCF staff support

COMPUTE RESOURCES

- Theta
- Sage (Urika-GX Cluster)
- Cooley and Mira
- Aurora (Future)

AURORA – COMING 2018

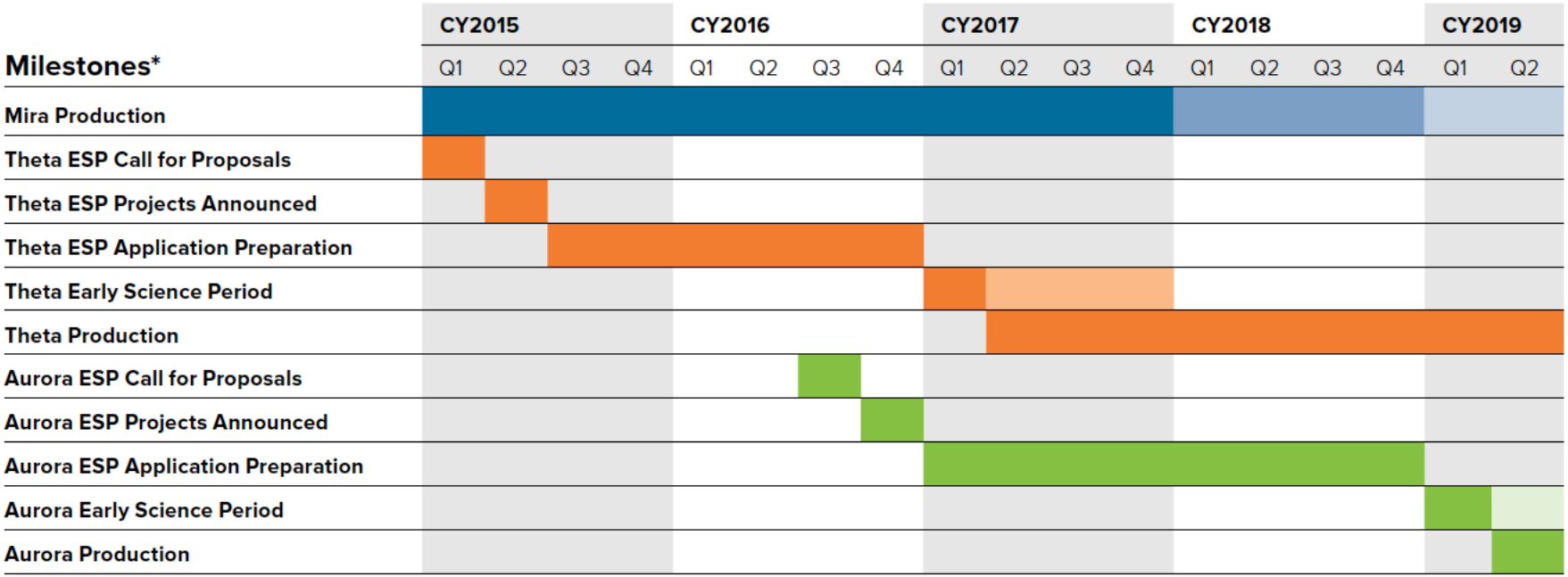


- Over 13X Mira's application performance
- Over 180 PF peak performance
- More than 50,000 nodes with 3rd Generation Intel® Xeon Phi™ processor
  - codename Knights Hill, > 60 cores
- Over 7 PB total system memory
  - High Bandwidth On-Package Memory, Local Memory, and Persistent Memory
- 2nd Generation Intel® Omni-Path Architecture with silicon photonics in a dragonfly topology
- More than 150 PB Lustre file system capacity with > 1 TB/s I/O performance

Call for Aurora ESP just closed at the beginning of the month!

MIRA TO AURORA TIMELINE

From Mira to Aurora



* Dates are subject to change.

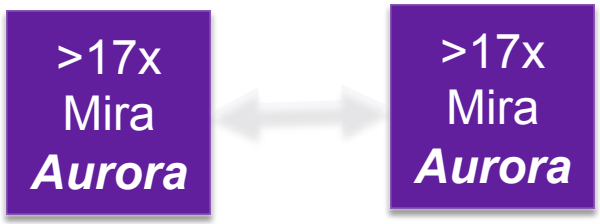
MIRA, THETA, AURORA TRANSITION

System Feature	Mira (2012)	Theta (2016)	Aurora (2018)
Peak performance	10 PF	8.6 PF	180 PF
Number of nodes	49,152	3240	>50,000
Aggregate high-bandwidth, on-package memory, local memory, and persistent memory	786 TB	650 TB	>7 PB
File system capacity	26 PB	10 PB	>150 PB
File system throughput	300 GB/s	200 GB/s	>1 TB/s
Peak power consumption	4.8 MW	1.7 MW	13 MW
GFLOPS/watt	2.1	>5	>13
Machine footprint	1,536 sq. ft.	~1,000 sq. ft.	~3,000 sq. ft.

NODE CAPABILITY GROWING FAST

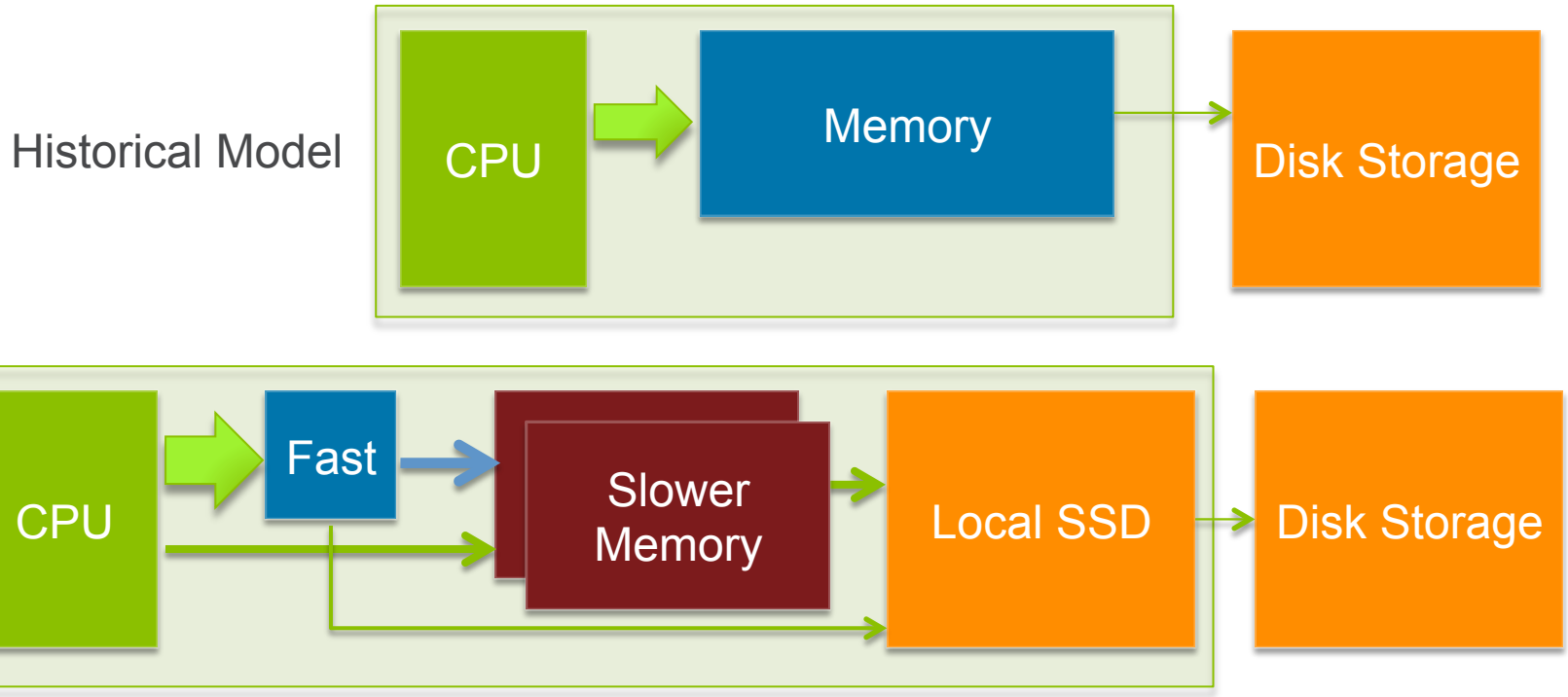
2016/2017 Systems

- Node level parallelism in the 100s now
- Node growth is not matched by interconnect
- Reliance on vectorization will continue increase



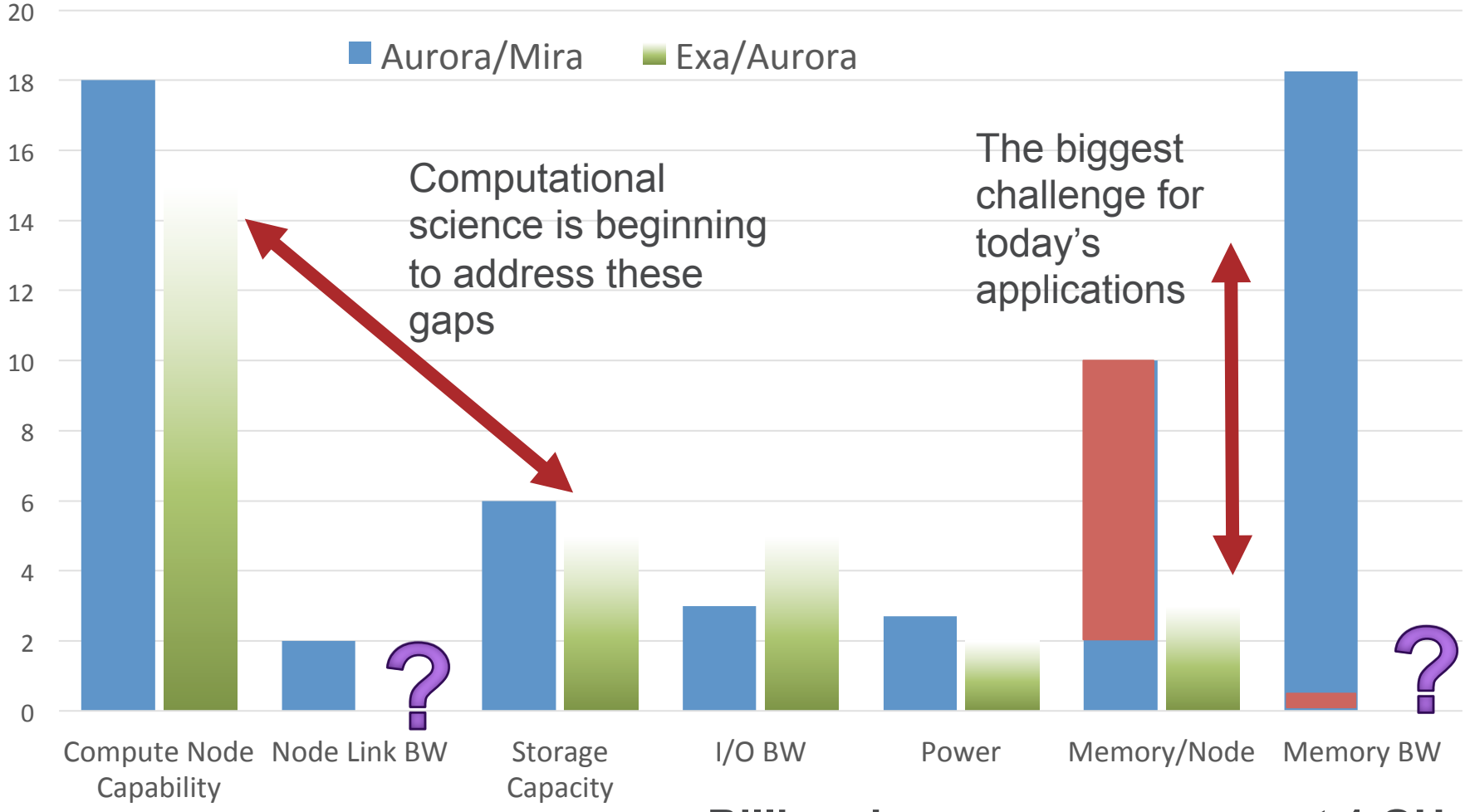
NODE CAPABILITY INCLUDES MEMORY HIERARCHY

- Model of application memory use has changed



New model for many core or accelerated

PROJECTED EXASCALE SYSTEMS (2022) RELATIVE TO AURORA (2018), *TREND ILLUSTRATION*



>Billion degree concurrency at 1 GHz

Mira ⇒ Exascale roughly 200x capability

JLSE: LCF-MCS JOINT LABORATORY FOR SYSTEM EVALUATION

OBJECTIVE: Evaluate future hardware and software platforms for computing and computational science activities at Argonne. Jointly managed by LCF and MCS.

GOALS:

- Improve science productivity on current & future LCF platforms
- Investigate alternative approaches to current & future deployments (hardware & software)
- Maintain a range of hardware & software test beds
- Help to drive standards in technologies, programming models, languages etc.

IMPLEMENTATION:

- Manage variety of test beds for hardware & software evaluation
- Coordinate interactions with vendors
- Engage in joint research on performance modeling, application optimization, programming models etc.

JLSE RESOURCES

System Name	Description	Hostname(s)	Status	Service Date
Sage	Cray uREKA-GX Analytics Cluster - 32x Haswell-EP E5-2697v3, 256GB RAM, 800GB NVMe SSD, Aries Interconnect Access Restricted	• sagelogin1 • sagelogin2 • sage00-27 • sagelnet1 • sagelnet2	Waiting configuration	
KNL	Qty: 10 Intel Adam's Pass S7200AP, (8 - KNL 7230 64c 1.3Ghz, 2 KNL 7250), 192GB RAM, OmniPath HCA Access restricted	• knl00-10	In Service	
It	Qty: 13 Intel HNS2600KPR, 2x E5-2699v4 22c 2.2Ghz, 128GB RAM, OmniPath HCA	• it00-12	In Service	2016-05-25
Gomez	Qty: 5 - SuperMicro 4048B-TR4FT, 4 Socket - Intel Haswell-EX E7-8867v3 16c 2.5Ghz. 1TB RAM, 2x Mellanox EDR, 1x OmniPath, 4RU.	• gomez00-04	In Service	2016-05-25
Pugsley	Qty: 6 - IBM x3650 M4, previously IBM GSS file servers, now for Lustre testbed Access Restricted	• pugsley-mds0-1 • pugsley-oss0-3	Waiting on provisioning	
Lurch	Qty: 8 - IBM x3650 M4, previously IBM GSS file servers, now for JLSE general use and high-end Filesystem Benchmark clients	• lurch00-07	Waiting on provisioning	
Firestone	Qty: 1 - IBM S822LC, OpenPower, Power8 10c 2.92Ghz, 1xEDR, Nvidia K80 GPU, 128GB RAM	• firestone	In Service	2016-04-11
Tubes2	Qty: 1 - IBM x3650 M4, 40GbE DTN Testbed. Access Restricted	• tubes2	In Service	2016-02-05
Petrel v2	Qty: 1 - IBM ESS GL6, 2PB raw, 1.6PB GPFS, 10xFDR10 Vesta Fabric connectivity Qty: 2 - Intel HNS2600KPF, E5-1660v3, 4-Nodes in 2RU chassis, 1x40GbE WAN, 1xQDR for GPFS. Access Restricted	• petrel1-2 • petrelmgmt1 • petreldtn1-8	In Service	2016-01-22
Fester	Qty: 1 - Intel R1208W2GS, production JLSE Monitoring server	• fester	Waiting configuration	
Mustang	Qty: 2 AppliedMicro X-C1 Development Kit Plus (X-Gene ARM64), https://www.apm.com/products/data-center/x-gene-family/x-c1-development-kits/x-c1-development-kit-plus/	• mustang00-01	In Service	2015-02-06
Thing	Qty: 8 - SuperMicro X10DRi, Intel Haswell-EP E5-2699v3, 4RU	• thing00-07	In Service, Thing07 dedicated to projects	2014-12-12

How to gain access?

Fill in your project request and request account on our website

http://www.jlse.anl.gov