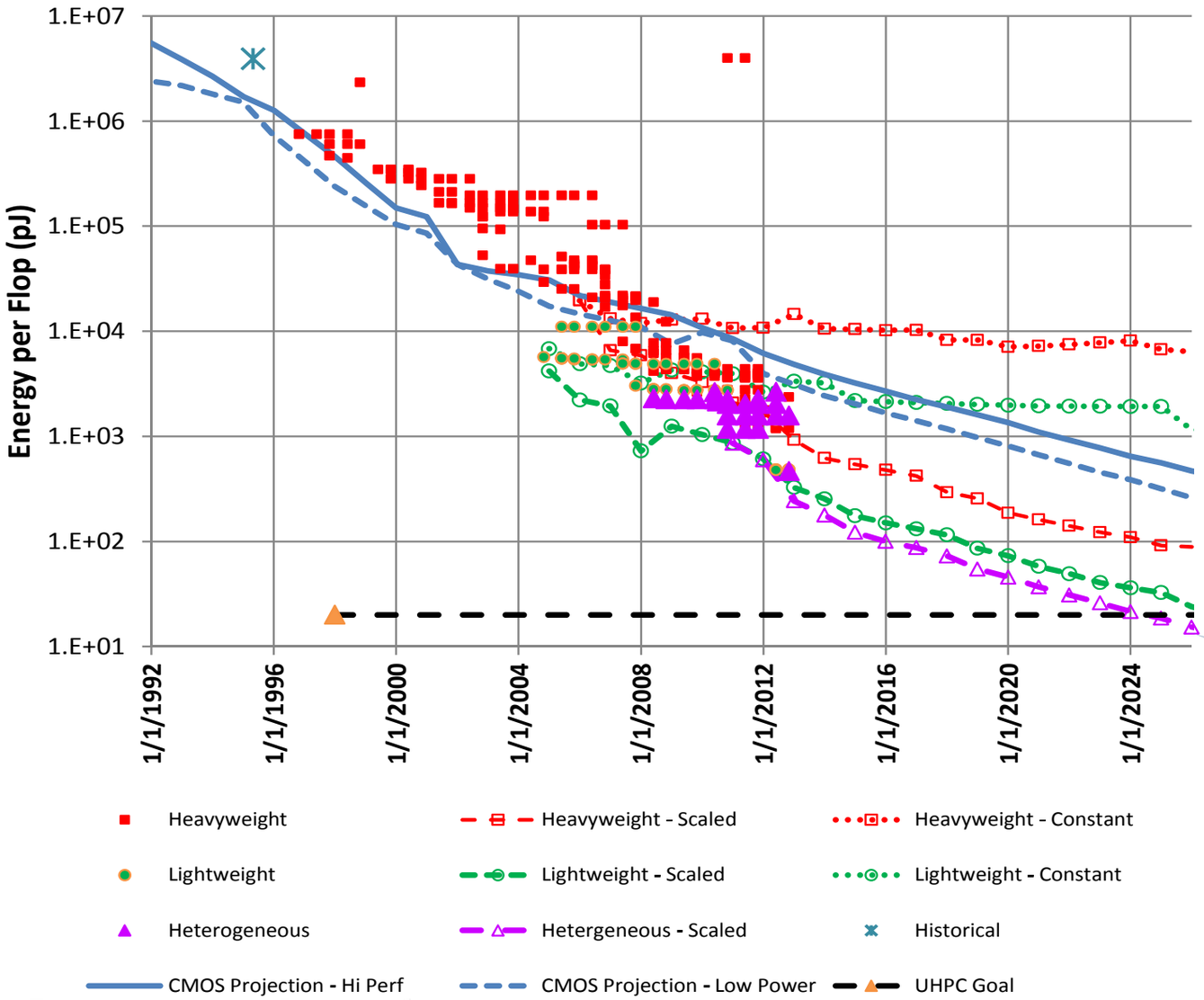
ASCR Facility Projections

Selected slides from presentations by Sudip Dosanjh (NERSC), Tim Williams (ALCF), and Jack Wells (OLCF) at the ASCR/HEP Exascale Requirements Review in June, 2015

Power as a Driver of Architecture

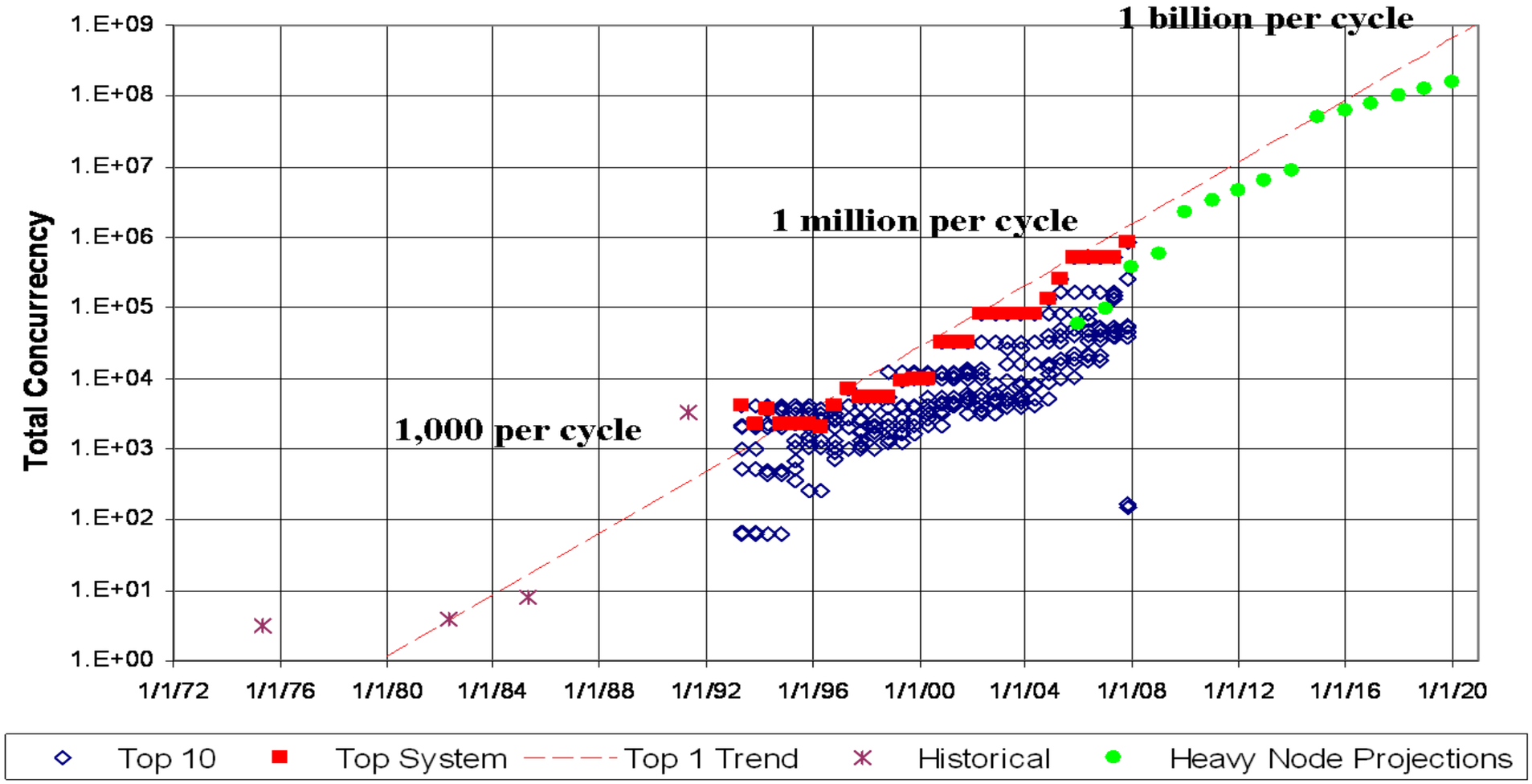
We need to transition to energy efficient architectures



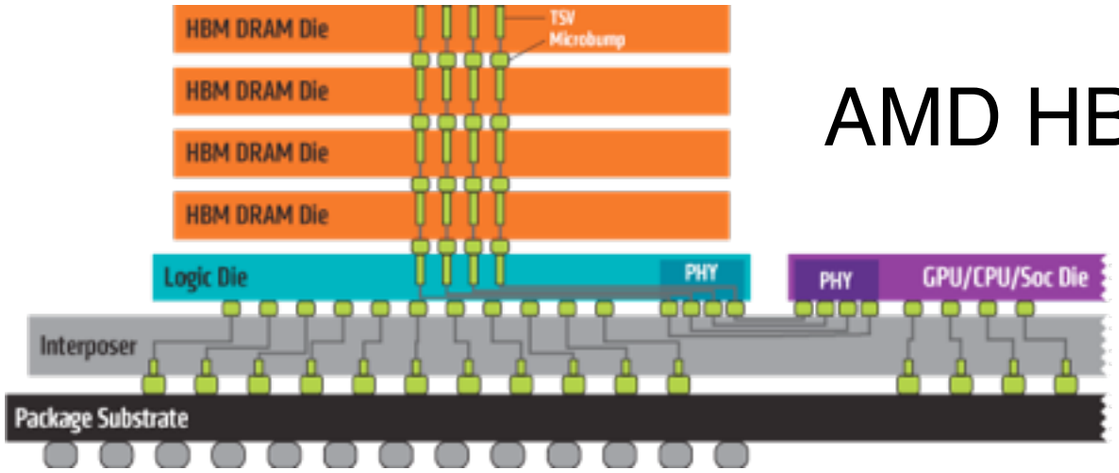
Manycore or Hybrid is the only approach that crosses the exascale finish line

Extreme Concurrency

Projected Parallelism for Exascale



Memory Bandwidth



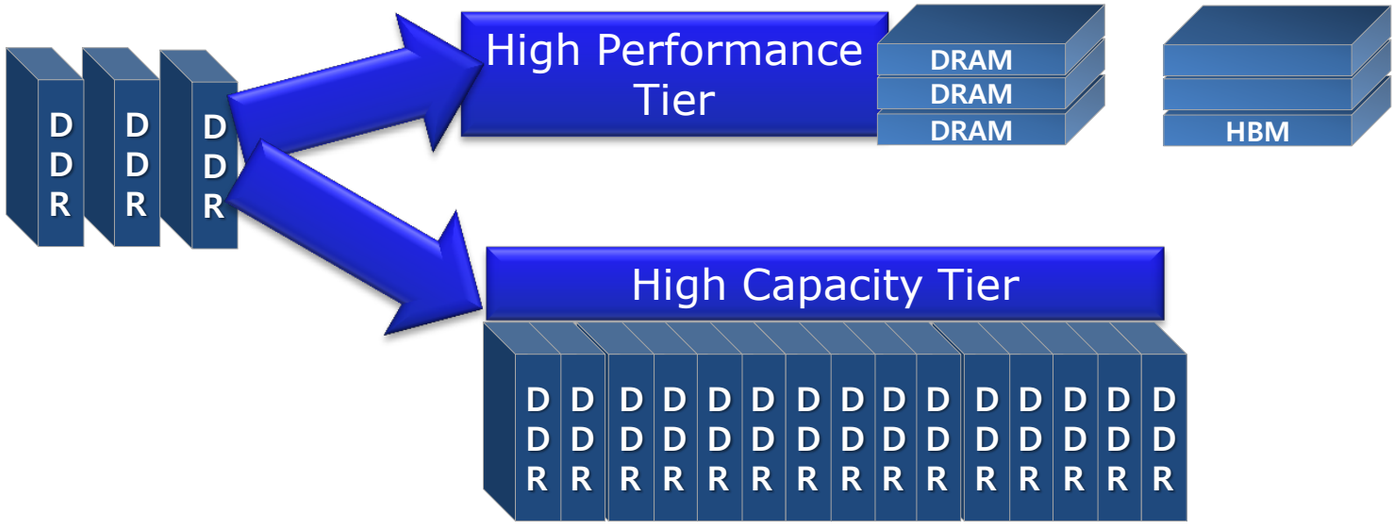
AMD HBM example

High-Speed interconnect

Cost (increases for higher capacity and cost/bit increases with bandwidth)

Power

Bandwidth\Capacity	16 GB	32 GB	64 GB	128 GB	256 GB	512 GB	1 TB
4 TB/s							
2 TB/s	Stack/PNM						
1 TB/s			Interposer				
512 GB/s				HMC organic			
256 GB/s					DIMM		
128 GB/s							NVRAM



Can get capacity OR bandwidth, but not both in the same technology

Two Swim lanes

Two Tracks for Future Large Systems

Tianhe-2 (NUDT): TH-IVB-FEP
Intel Xeon E5-2692 12 C 2.2 GHz
TH Express-2
Intel Xeon Phi 31S1P

Titan (Cray): Cray XK7
AMD Opteron 6274 16C 2.2 GHz
Cray Gemini
NVIDIA K20x

Sequoia (IBM): BlueGene/Q
Power BQC 16C 1.6 GHz

K computer (Fujitsu)
SPARC64 VIIIfx 2.0 GHz
Tofu

Mira (IBM): BlueGene/Q
PowerPC A2 16C 1.6 GHz

Piz Daint (Cray): Cray XC30
Intel Xeon E5-2670 8C 2.6 GHz
Cray Aries
NVIDIA K20x

Edison (Cray): Cray XC30
Intel Xeon E%-2695v2 12C 2.4 GHz
Aries

Many Core

- 10's of thousands of nodes with millions of cores
- Homogeneous cores
- Multiple levels of memory – on package, DDR, and non-volatile
- Unlike prior generations, future products are likely to be self hosted

Hybrid Multi-Core

- CPU / GPU Hybrid systems
- Likely to have multiple CPUs and GPUs per node
- Small number of very fat nodes
- Expect data movement issues to be much easier than previous systems – coherent shared memory within a node
- Multiple levels of memory – on package, DDR, and non-volatile

Cori at NERSC

- Self-hosted many-core system
- Intel/Cray
- 9300 single-socket nodes
- Intel® Xeon Phi™ Knights Landing (KNL)
- 16GB HBM, 64-128 GB DDR4
- Target delivery date: June, 2016

Summit at OLCF

- Hybrid CPU/GPU system
- IBM/NVIDIA
- 3400 multi-socket nodes
- POWER9/Volta
- More than 512 GB coherent memory per node
- Target delivery date: 2017

ALCF-3 at ALCF

- 3rd Generation Intel Xeon Phi (Knights Hill (KNH)
- > 50,000 compute nodes
- Target delivery date: 2018

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

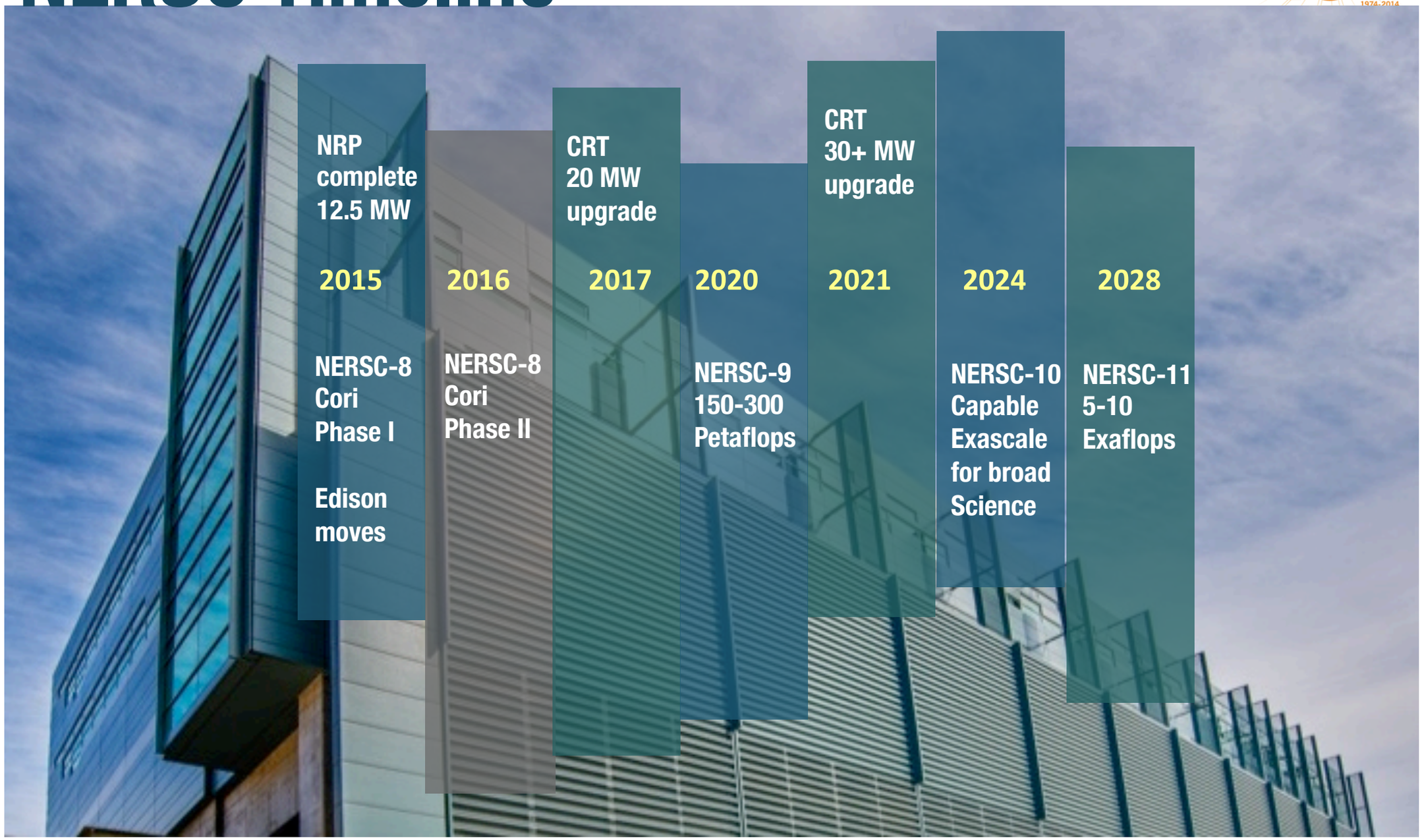
ASCR Computing

ASCR Computing At a Glance

now ← → future

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**	**Aurora 2018-2019**
System peak (PF)	2.6	27	10	> 30	150	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	10	1.7	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM) +1.5PB persistent memory	> 1.74 PB DDR4 + HBM + 2.8 PB persistent memory	>480 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	2nd gen Intel Xeon Phi processor (code name Knights Landing)	3rd gen Intel Xeon Phi processor (code name Knights Hill)
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	>2,500 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®

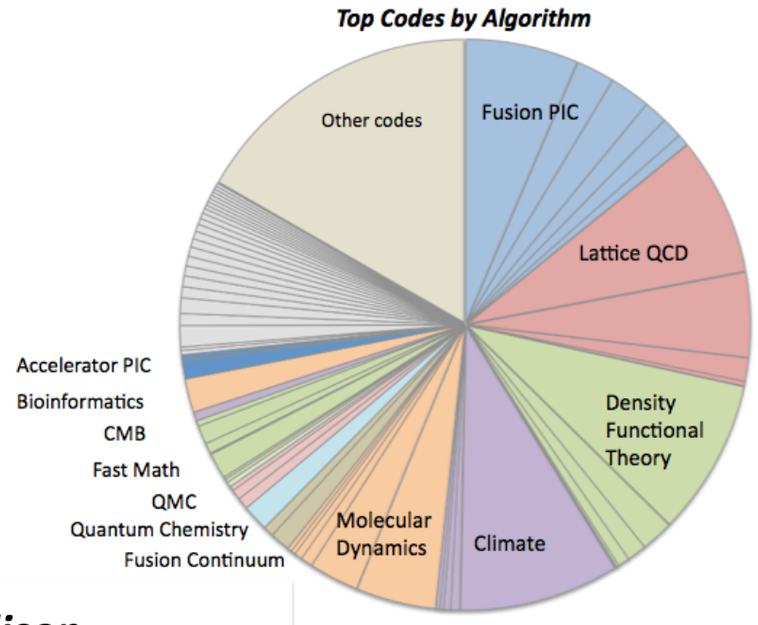
NERSC Timeline

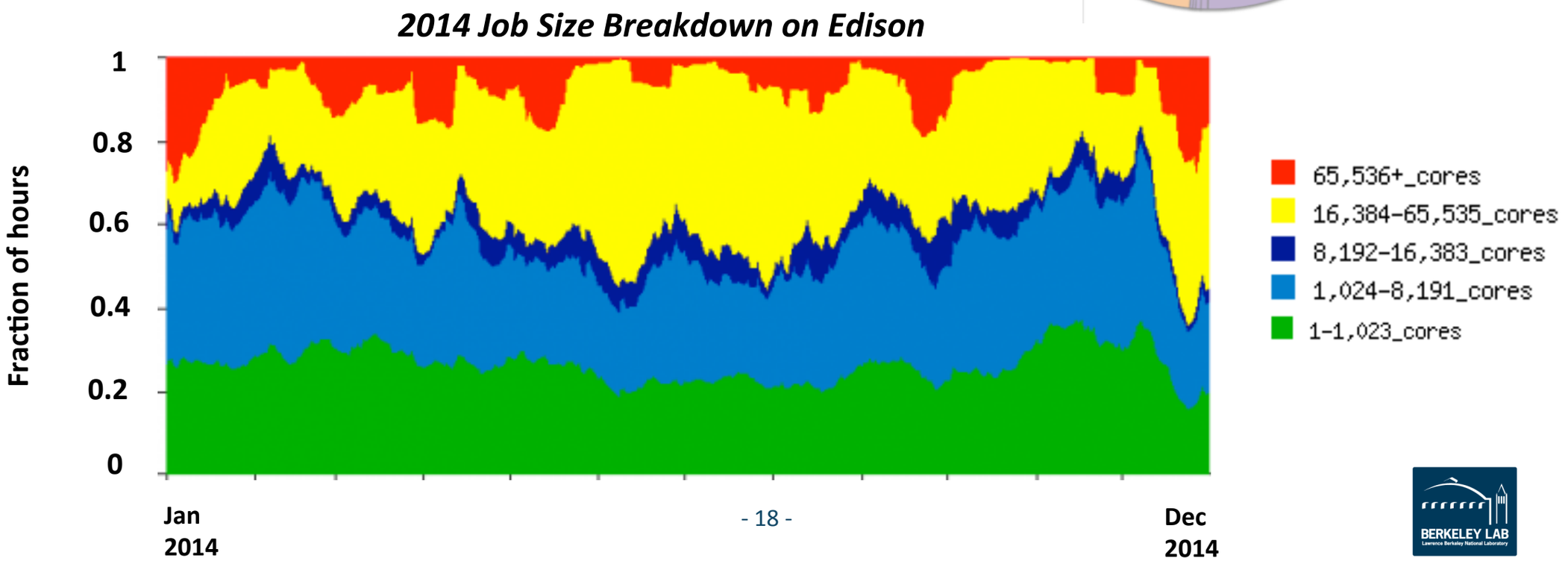


NERSC Workloads

We support a diverse workload

- **Many codes (600+) and algorithms**
- **Computing at scale and at high volume**





Cori@NERSC

The NERSC-8 System: Cori

- **Cori will support the broad Office of Science research community and begin to transition the workload to more energy efficient architectures**
- **Cray XC system with over 9300 Intel Knights Landing compute nodes –mid 2016**
  - Self-hosted, (not an accelerator) manycore processor with over 60 cores per node
  - On-package high-bandwidth memory
- **Data Intensive Science Support**
  - 10 Haswell processor cabinets to support data intensive applications – Summer 2015
  - NVRAM Burst Buffer to accelerate data intensive applications
  - 28 PB of disk, >700 GB/sec I/O bandwidth
- **Robust Application Readiness Plan**
  - Outreach and training for user community
  - Application deep dives with Intel and Cray
  - 8 post-docs integrated with key application teams

Image source: Wikipedia

System named after Gerty Cori, Biochemist and first American woman to receive the Nobel prize in science.

Intel "Knights Landing" Processor

- Next generation Xeon-Phi, >3TF peak
- Single socket processor - Self-hosted, not a co-processor, not an accelerator
- Greater than 60 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™
- Intel® "Silvermont" architecture enhanced for high performance computing
- 512b vector units (32 flops/clock – AVX 512)
- 3X single-thread performance over current generation Xeon-Phi co-processor
- High bandwidth on-package memory, up to 16GB capacity with bandwidth projected to be 5X that of DDR4 DRAM memory
- Higher performance per watt

KNL IPM

Knights Landing Integrated On-Package Memory

Cache Model

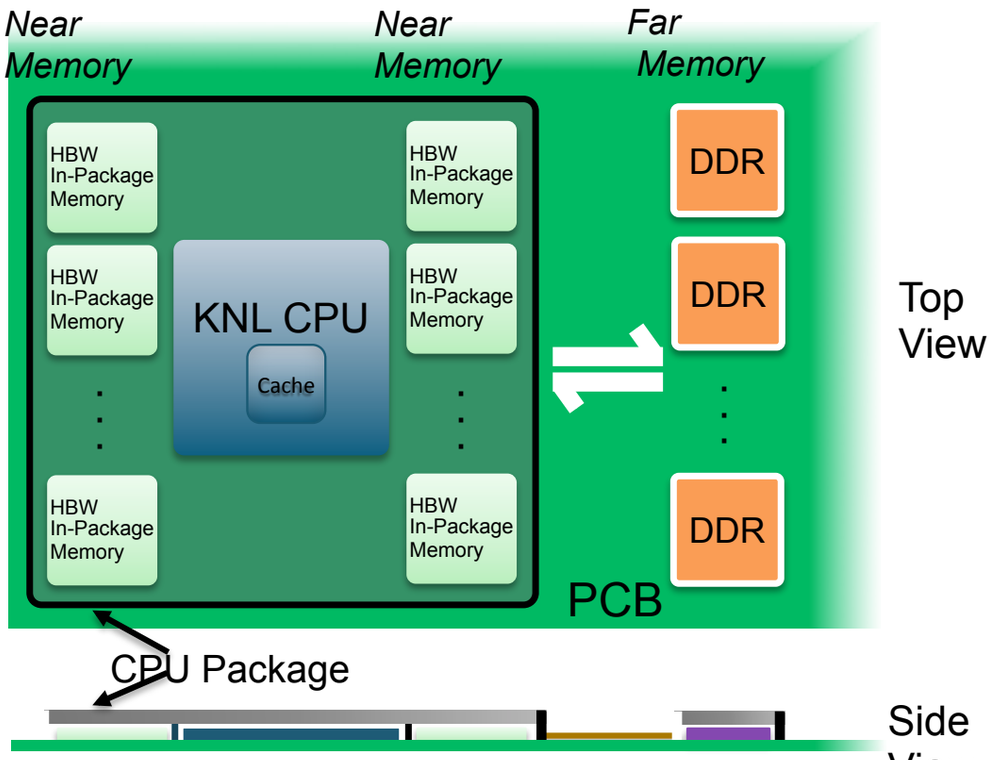
Let the hardware automatically manage the integrated on-package memory as an "L3" cache between KNL CPU and external DDR

Flat Model

Manually manage how your application uses the integrated on-package memory and external DDR for peak performance

Hybrid Model

Harness the benefits of both cache and flat models by segmenting the integrated on-package memory



Maximum performance through higher memory bandwidth and flexibility

NESAP

20 NESAP Tier-1 and Tier-2 codes

ASCR (2)

Almgren (LBNL) – **BoxLib AMR Framework** used in combustion, astrophysics

Trebotich (LBNL) – **Chombo-crunch** for subsurface flow

BES (5)

Kent (ORNL) – **Quantum Espresso**
Deslippe (NERSC) – **BerkeleyGW**
Chelikowsky (UT) – **PARSEC** for excited state materials
Bylaska (PNNL) – **NWChem**
Newman (LBNL) – **EMGeo** for geophysical modeling of Earth

BER (5)

Smith (ORNL) – **Gromacs** Molecular Dynamics
Yelick (LBNL) – **Meraculous** genomics
Ringler (LANL) – **MPAS-O** global ocean modeling
Johansen (LBNL) – **ACME** global climate
Dennis (NCAR) – **CESM**

HEP (3)

Vay (LBNL) – **WARP & Synergia** accelerator modeling
Toussaint (U Arizona) – **MILC** Lattice QCD
Habib (ANL) – **HACC** for *n*-Body cosmology

NP (3)

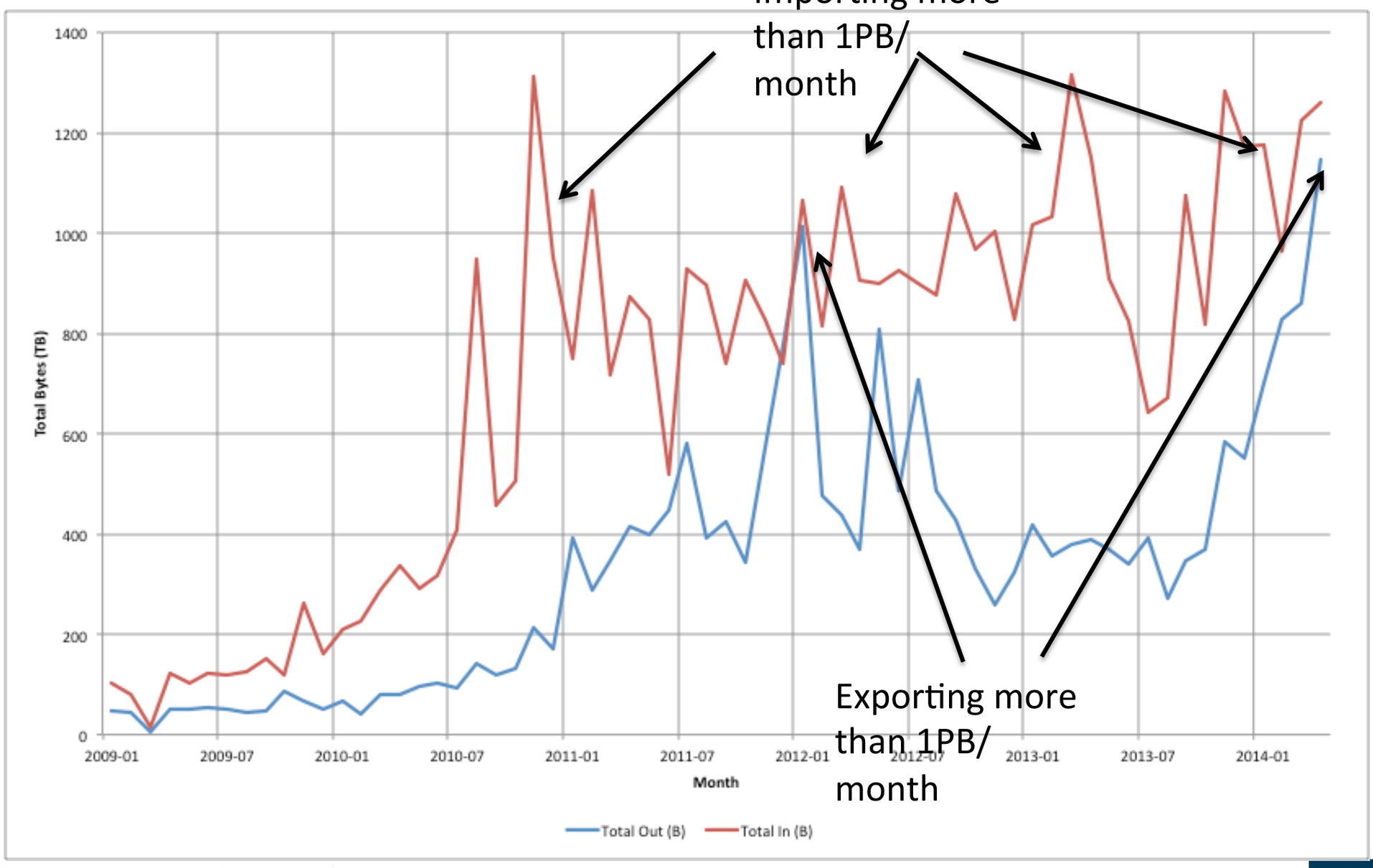
Maris (U. Iowa) – **MFDn** *ab initio* nuclear structure
Joo (JLAB) – **Chroma** Lattice QCD
Christ/Karsch (Columbia/ BNL) – **DWF/HISQ** Lattice QCD

FES (2)

Jardin (PPPL) – **M3D** continuum plasma physics
Chang (PPPL) – **XGC1** PIC plasma

Data Issues

NERSC users import more data than they export!



Cori and Data

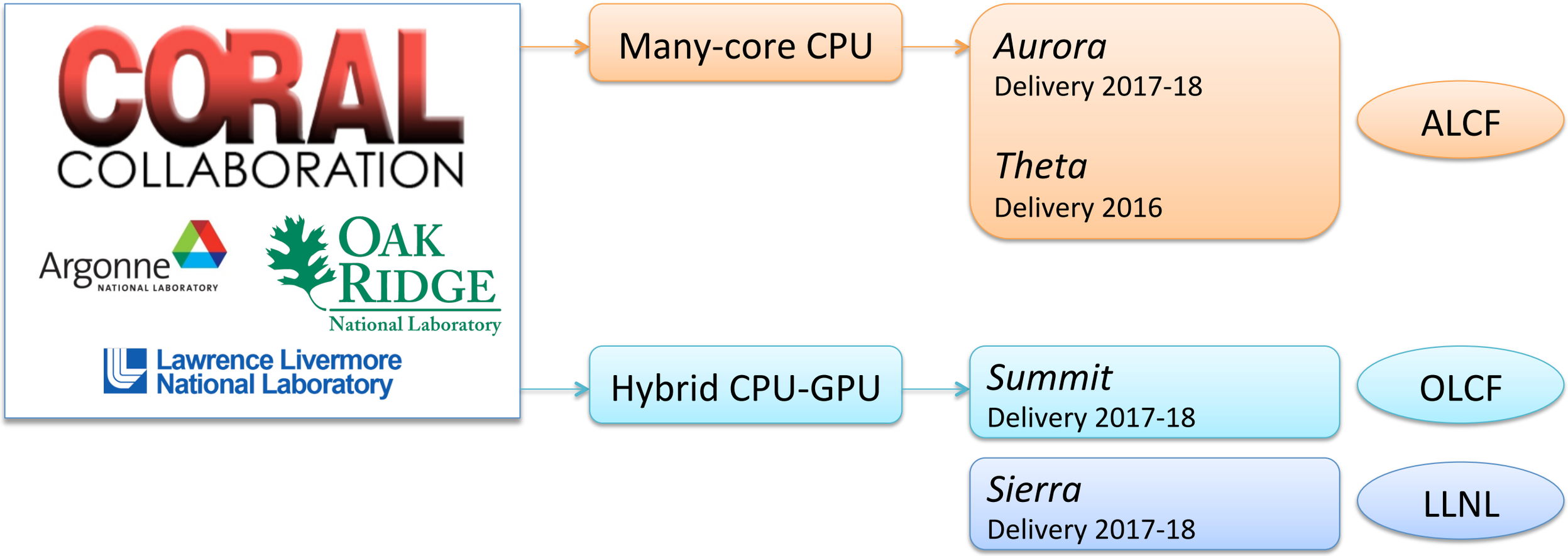
Cori Data Enhancements

- **Data partition with large memory nodes, software to enable data workflows (including user defined images)**
- **IO enhancements-- NVRAM nodes on the interconnect fabric for caching, software defined networking**
- **Larger disk system**

Goals are to enable the analysis of large experimental data sets and in-situ analysis coupled to Petascale simulations

CORAL

Next-Generation ALCF-3 System



Aurora, Summit are LCF's **pre-exascale** systems.

Argonne Leadership Computing Facility

Theta@ALCF

Theta Details

System Feature	*Theta*
Peak System performance (FLOPs)	>8.5 PetaFLOP/s
Processor	2nd Generation Intel® Xeon Phi™ processors (Code name: Knights Landing)
Number of Nodes	>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 On-Package Memory	Up to 16 Gigabytes per compute node
DDR4 Memory	192 Gigabytes
On-node storage	128 GB SSD
File System	Intel Lustre* File System
System Interconnect	Cray Aries* high speed Dragonfly* topology interconnect
File System Capacity (Initial)	10 Petabytes
File System Throughput (Initial)	210 Gigabytes/s
Peak Power Consumption	1.7 Megawatts
Delivery Timeline	Mid-2016

Aurora Details

System Feature	*Aurora*
Peak System performance (FLOPs)	180 - 450 PetaFLOPS
Processor	3rd Generation Intel® Xeon Phi™ processor (code name Knights Hill)
Number of Nodes	>50,000
Compute Platform	Cray Shasta next generation supercomputing platform
High Bandwidth On-Package Memory, Local Memory, and Persistent Memory	>7 Petabytes
System Interconnect	2nd Generation Intel® Omni-Path Architecture with silicon photonics
Interconnect interface	Integrated
Burst Storage Buffer	Intel® SSDs, 2nd Generation Intel® Omni-Path Architecture
File System	Intel Lustre* File System
File System Capacity	>150 Petabytes
File System Throughput	>1 Terabyte/s
Peak Power Consumption	13 Megawatts
FLOPS/watt	>13 GFLOPS/watt
Delivery Timeline	2018
Facility Area	~3,000 sq. ft.

Theta ESP

Tier 1 projects:

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

Large-Scale Simulation of Brain Tissue: Blue Brain Project, EPFL
PI: Fabien Delalondre, Ecole Federale Polytechnique de Lausanne

First-Principles Simulations of Functional Materials for Energy Conversion
PI: Giulia Galli, Argonne National Laboratory/University of Chicago

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

Direct Numerical Simulations of Flame Propagation in Hydrogen-Oxygen Mixtures in Closed Vessels
PI: Alexei Khokhlov, University of Chicago

Free Energy Landscapes of Membrane Transport Proteins
PI: Benoit Roux, Argonne National Laboratory/University of Chicago

Summit@OLCF

2017 OLCF Leadership System
Hybrid CPU/GPU architecture

SUMMIT
OAK RIDGE NATIONAL LABORATORY

OpenPOWER™

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

At least 5X Titan's Application Performance

Approximately 3,400 nodes, each with:

- Multiple IBM POWER9 CPUs and multiple NVIDIA Tesla® GPUs using the NVIDIA Volta architecture
- 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.

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

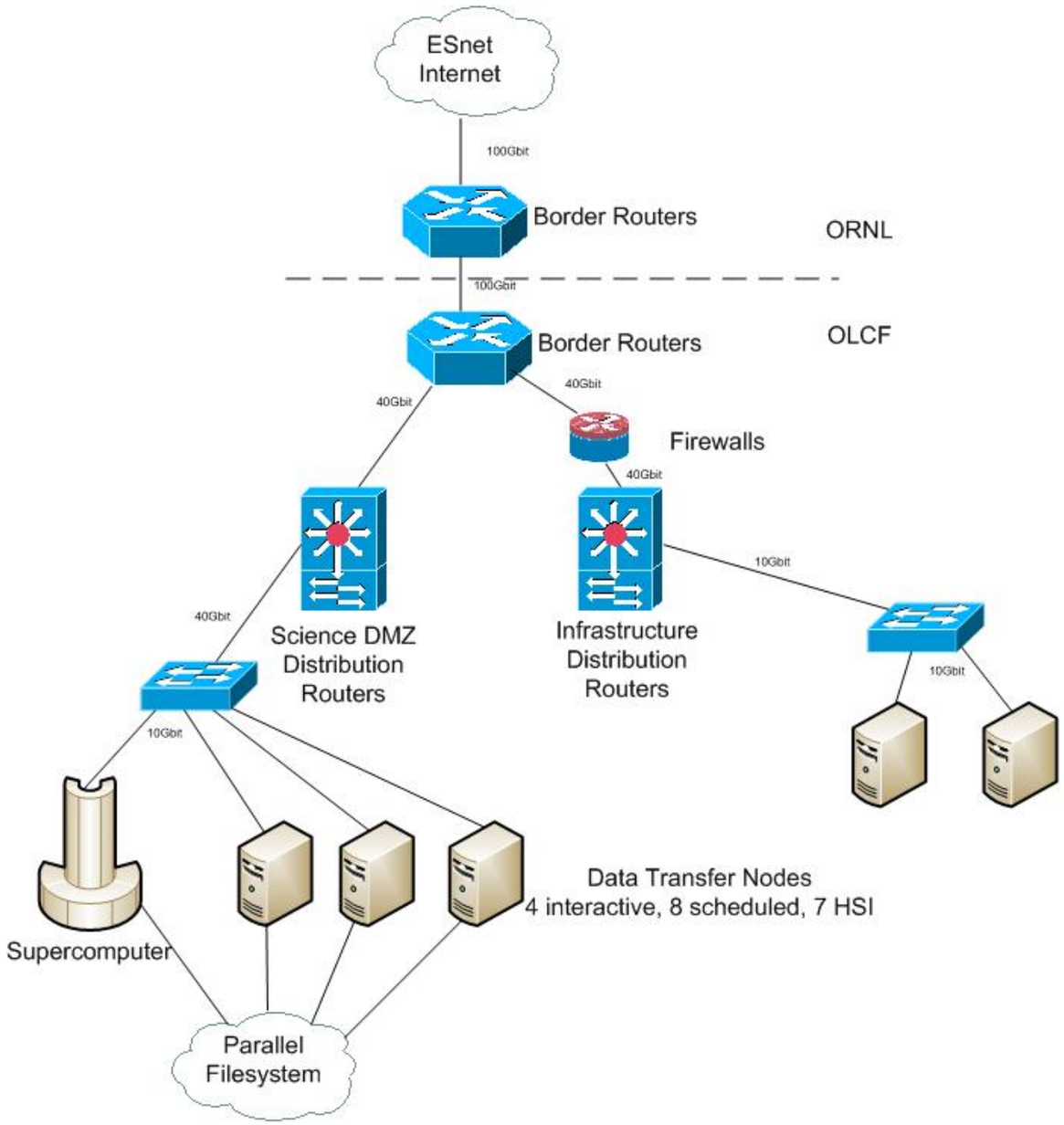
CAAR (ESP)

New Application Readiness Activities CAAR

Application	Domain	Principal Investigator	Institution
ACME (N)	*Climate Science*	David Bader	Lawrence Livermore National Laboratory
DIRAC	*Relativistic Chemistry*	Lucas Visscher	Free University of Amsterdam
FLASH	*Astrophysics*	Bronson Messer	Oak Ridge National Laboratory
GTC (NE)	*Plasma Physics*	Zhihong Lin	University of California – Irvine
HACC(N)	*Cosmology*	Salman Habib	Argonne National Laboratory
LSDALTON	*Chemistry*	Poul Jørgensen	Aarhus University
NAMD (NE)	*Biophysics*	Klaus Schulten	University of Illinois – Urbana Champaign
NUCCOR	*Nuclear Physics*	Gaute Hagen	Oak Ridge National Laboratory
NWCHEM (N)	*Chemistry*	Karol Kowalski	Pacific Northwest National Laboratory
QMCPACK	*Materials Science*	Paul Kent	Oak Ridge National Laboratory
RAPTOR	*Engineering*	Joseph Oefelein	Sandia National Laboratory
SPECFEM	*Seismic Science*	Jeroen Tromp	Princeton University
XGC (N)	*Plasma Physics*	CS Chang	Princeton Plasma Physics Laboratory

Science DMZ@OLCF

OLCF & ESNet are implementing the Science DMZ to enable high-performance access to ESNet WAN



IDGE boratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY