ORNL Center for Accelerated Application Readiness (CAAR)

Preparing today's applications for tomorrow's machines

OLCF
OAK RIDGE LEADERSHIP COMPUTING FACILITY

1st Hybrid Multicore Consortium Workshop
San Francisco, CA
20-22 January 2010

John A. Turner
Group Leader, Computational Engineering & Energy Sciences
Computer Science & Math Division

ORNL computational resources support a broad range of applications.

NCCS

National Center for Computational Sciences

NICS

National Institute for Computational Sciences

Applications

- Superconductivity
- Chemistry
- Fission
- Climate change
- Bioinformatics
- Materials
- Combustion
- Fusion
- Geosciences
- Biology
- Nanoscience
- Energy
- Molecular dynamics
- Groundwater
- Biophysics

Institute for Advanced Architecture and Algorithms
DOE-SC and NNSA

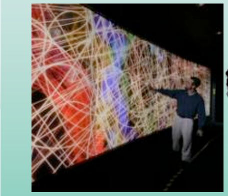
Extreme Scale Software Center
DOE-SC and DOD

HPC experience and operations

Facilities and infrastructure

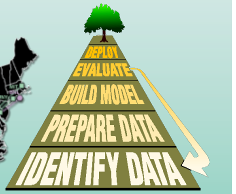
Data storage and file systems

Computer systems



Visualization

Networks



Knowledge discovery

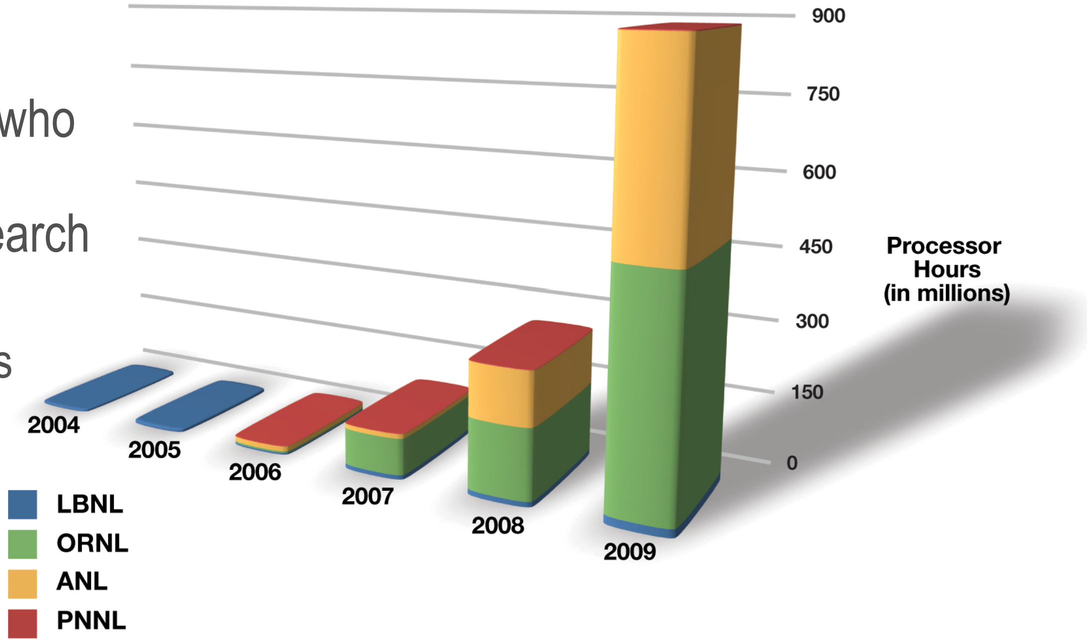
DOE’s INCITE Program

- The INCITE program grants large blocks of time to a few nationally important projects.
- Projects are selected through scientific peer review
- Call for proposals open to academic, industry, and government researchers
- NCCS also provides discretionary allocations for scientists and engineers who want to port, scale, and use their applications codes for cutting-edge research
  - For INCITE preparation, industrial collaboration, and strategic partnerships

Office of Science
U.S. DEPARTMENT OF ENERGY

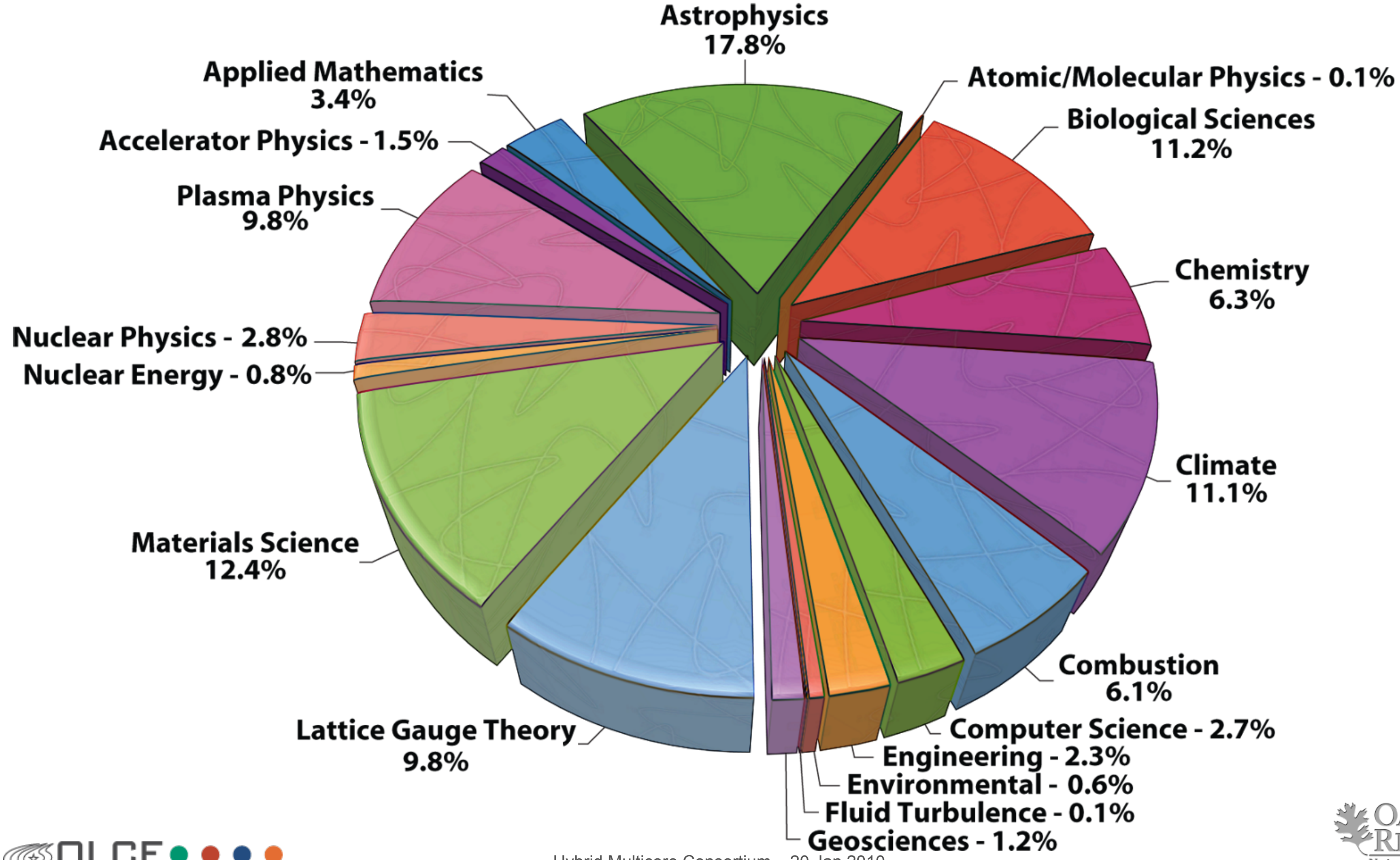
INCITE
Advancing America's Science
and Industrial Competitiveness

INCITE allocations (millions of CPU hours)



2009 INCITE Awards

nearly 900 million processor-hours were awarded
- 25 new projects and 41 renewal projects.



INCITE applications span a broad range of science challenges

Projects	2006	2007	2008	2009
Accelerator physics	1	1	1	1
Astrophysics	3	4	5	5
Chemistry	1	1	2	4
Climate change	3	3	4	5
Combustion	1	1	2	2
Computer science	1	1	1	1
Fluid Dynamics			1	1
Fusion	4	5	3	5
Geosciences		1	1	1
High energy physics		1	1	
Life sciences	2	2	2	4
Materials science	2	3	3	4
Nuclear physics	2	2	1	2
Industry	2	3	3	3
Total Projects:	**22**	**28**	**30**	**38**
CPU Hours:	**36,156,000**	**75,495,000**	**145,387,000**	**469,683,000**

Today's systems at ORNL represent significant advance.

2004

Cray X1
3 TF

2005

Cray XT3
Single-core
26 TF

2006

Cray XT3
Dual-core
54 TF

2007

Cray XT4
119 TF

2008

Cray XT4
Quad-core
166 TF

Cray XT4
Quad-core
263 TF

2009

Cray "Baker"
6-core, dual-socket SMP
~1000 TF
100TB, 2.5PB

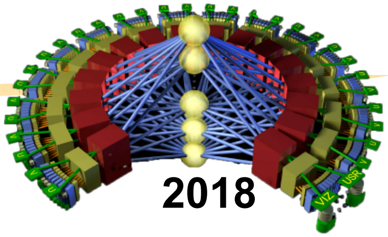
Cray "Baker"
6-core, dual-socket SMP
~2600TF
300 TB, 10PB

2011

OLCF-3 /
DARPA
20 PF

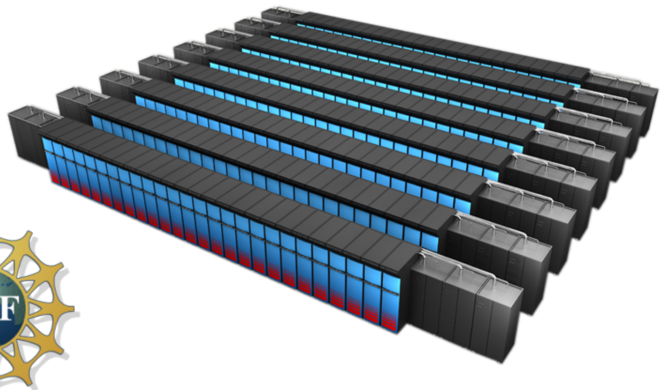
2015

Future
system
100–250 PF

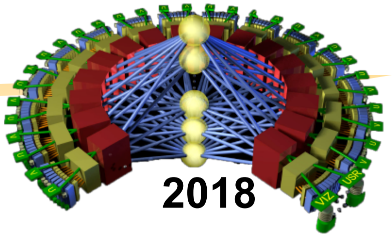


2018

Future
system
1000 PF
(1 EF)



Dramatic increases in computational capabilities will continue.



2004

Cray X1
3 TF

2005

Cray XT3
Single-core
26 TF

2006

Cray XT3
Dual-core
54 TF

2007

Cray XT4
119 TF

2008

Cray XT4
Quad-core
166 TF

Cray XT4
Quad-core
263 TF

2009

Cray “Baker”
6-core, dual-socket SMP
~1000 TF
100TB, 2.5PB

Cray “Baker”
6-core, dual-socket SMP
~2600TF
300 TB, 10PB

2011

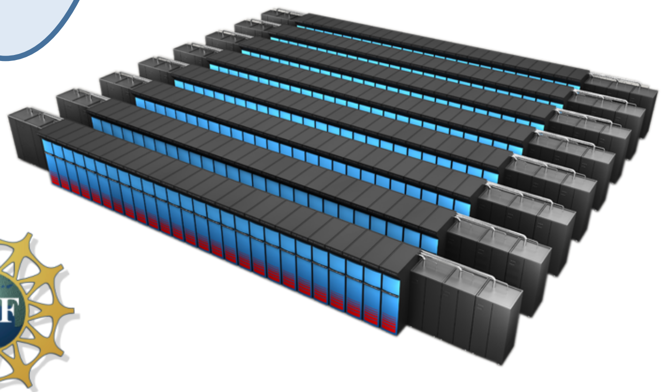
OLCF-3 /
DARPA
20 PF

2015

Future
system
100–250 PF

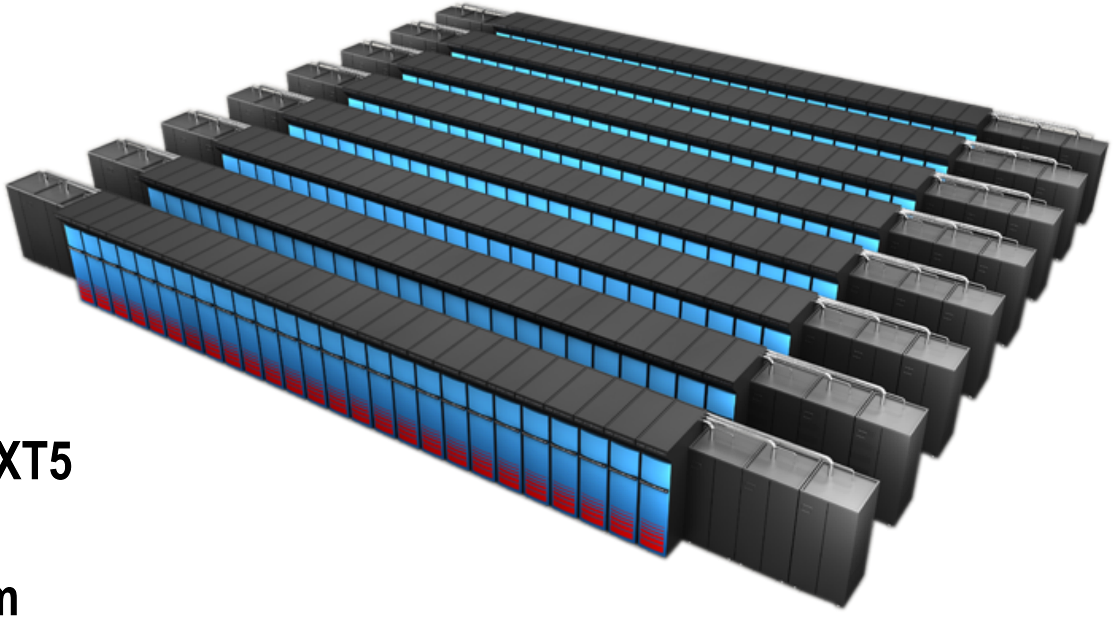
2018

Future
system
1000 PF
(1 EF)



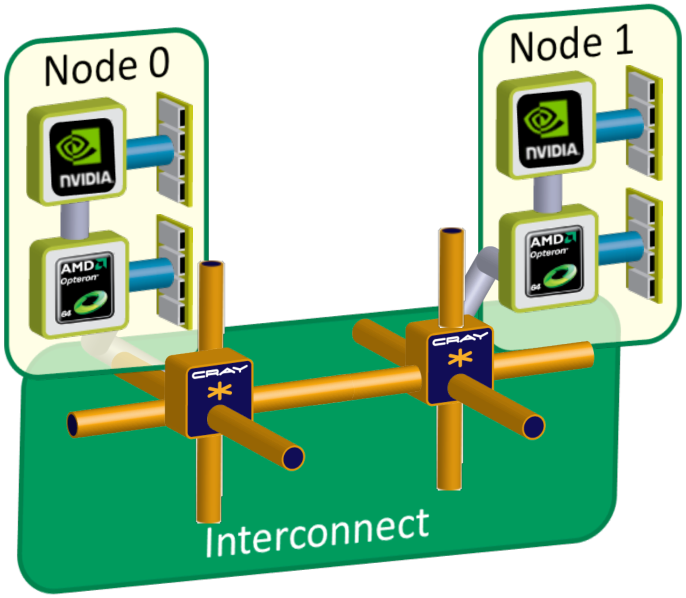
Potential OLCF-3 system description

- Same number of cabinets, cabinet design, and cooling as Jaguar
- Operating system upgrade of today's Cray Linux Environment
- New Gemini interconnect
    - 3-D Torus
    - Globally addressable memory
    - Advanced synchronization features
- New accelerated node design
- 10-20 PF peak performance
    - 15x performance of today's XT5
- Much larger memory
- 3x larger and 4x faster file system
- ≈ 10 MW of power

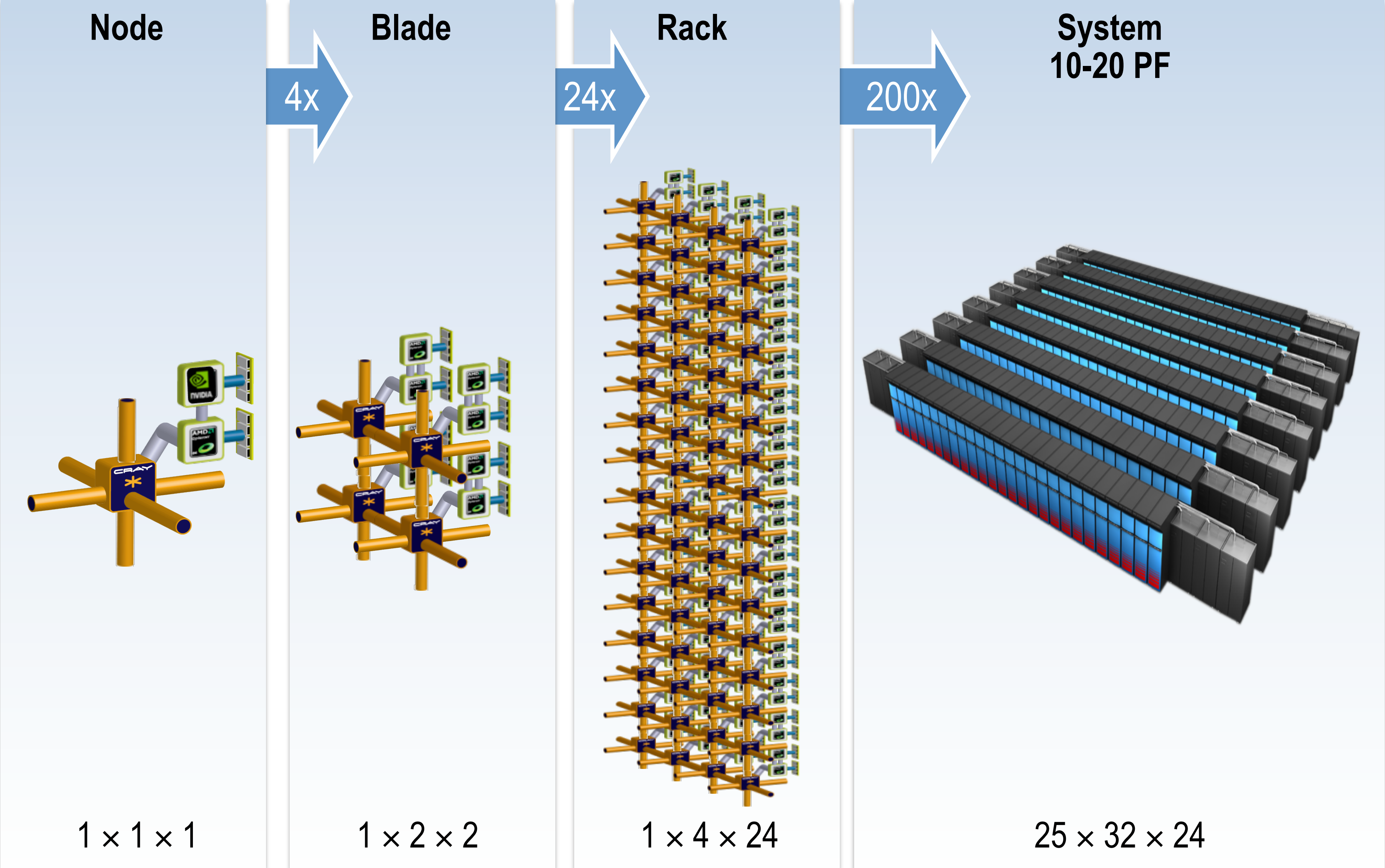


OLCF-3 node description

- **Accelerated Node Design**
  - Next generation interconnect
  - Next generation AMD processor
  - Future NVIDIA accelerator
- **Fat nodes**
  - 70 GB memory
  - Very high performance processors
  - Very high memory bandwidth



Building the OLCF-3 system



Node

4x

Blade

24x

Rack

200x

System
10-20 PF

1 × 1 × 1

1 × 2 × 2

1 × 4 × 24

25 × 32 × 24

X × Y × Z 3-D Torus Dimensions

NVIDIA's commitment to HPC

Features for computing on GPUs

- Added ***high-performance*** 64-bit arithmetic
- Adding ECC and parity
  that other GPU vendors have not added
  – Critical for a large system
- Larger memories
- Dual copy engines for
  simultaneous execution and copy
- S1070 has 4 GPUs exclusively
  for computing
  – No video out cables
- Development of CUDA
  and recently announced
  work with PGI on Fortran CUDA

May 04, 2009

NVIDIA Shifts GPU Clusters Into Second Gear

by Michael Feldman, HPCwire Editor

GPU-accelerated clusters are moving quickly from the "kick the tires" stage into production systems, and NVIDIA has positioned itself as the principal driver for this emer high
performance com
The
along
, are
rcial
rd)

Tesla Preconfigured Clusters

GPU COMPUTING SOLUTIONS
- Main
- What is GPU Computing
- Vertical Industry Solutions
- Tesla Personal Supercomputer
- Tesla Preconfigured Clusters
- Tesla S1070
- Tesla C1060
- Pre-purchase Assistance
- Where to Buy
- Contact US
- Sign Up for Tesla Alerts

SOFTWARE AND HARDWARE INFO
- Software Development Tools
- Compatible Platforms
- Product Literature
- Legacy Products
- Features and Benefits
- Tech Specs
- Build Your Own
  Learn how to build your own Tesla Personal Supercomputer today!

NEWS AND SUCCESS STORIES
- Webinars
- Tesla Videos on YouTube
- CUDA Software and Applications
- Success Stories
- News
- Testimonial Quotes
- Awards

GET YOUR OWN SUPERCOMPUTER

Supercharge your cluster with Tesla S1070 1U GPU computing systems.

Supercharge your cluster with Tesla S1070 1U GPU computing systems. Experience the supercomputing performance of a large cluster with just a small cluster of Tesla GPUs. Tesla-based clusters require lower power, lower space, and yet deliver up to 30 times the performance of equivalent CPU-only clusters. With 4 Tesla T10 GPUs, the NVIDIA® Tesla™ S1070 GPU computing delivers 4 Teraflops of performance in a 1U. Each GPU is based on the revolutionary NVIDIA® CUDA™ parallel computing architecture and powered by 240 parallel processing cores.

SUPERCOMPUTING PERFORMANCE
- Revolutionary CUDA Architecture
- 960 cores, 4 Teraflops in a 1U. Massively parallel processing
- 30x the compute performance of a standard cluster

LOW POWER AND LOW COST
- Superconputing performance from a small cluster
- Smaller cluster needs less power and less cooling

READY FOR DEPLOYMENT
- Turnkey, ready to power up cluster with software installed
- Program in C for Windows or Linux
- Available from resellers worldwide

BUY A Tesla Preconfigured Cluster TODAY

Look for Tesla Preconfigured Clusters from the following NVIDIA Tesla Preferred Providers.

"GPUs have evolved to the point where many real-world applications are easily implemented on them and run significantly faster than on multi-core systems. Future computing architectures will be hybrid systems with parallel-core GPUs working in tandem with multi-core CPUs."

Prof. Jack Dongarra
Director of the Innovative Computing Laboratory
The University of Tennessee Read more

Click here to contact a sales representative.
Click here to contact the Tesla GPU computing group.

GPGPU interest is widespread

- **Large GPU-based systems springing up everywhere**
  - NSF Track 2D in negotiation with Georgia Tech/ORNL
  - Japan: “Tsubame” at Tokyo Institute of Technology (upgrade planned)
  - CEA: 300 TF Nehalem/NVIDIA cluster located at CCRT
  - “Orbit” ORNL 100 TF NVIDIA testbed
  - Oil and gas industry deploying large clusters
- **Large and growing pool of people who know how to program accelerators and who will develop tools**
  - Every laptop has a processor and GPU
    - Macintosh, PC, Linux ports of CUDA available
  - Most computer science programs now teach GPU programming
- **At ORNL: We thoroughly understand the applications and have helped scale them out to 100,000+ cores on today’s Jaguar system**

300+ accelerated applications listed on NVIDIA’s web site: http://www.nvidia.com/object/cuda_home.html

DOE/ASCR Strategic Directions*

2010 and beyond

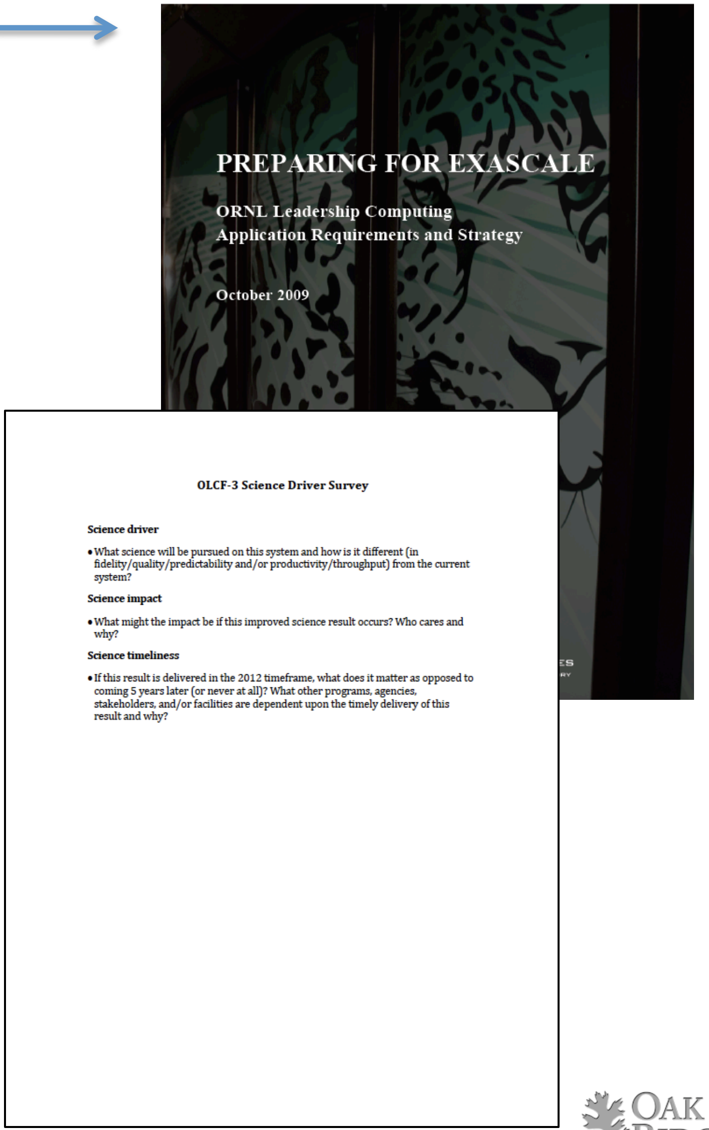
*Dr. William F. Brinkman, Director, Office of Science, ASCAC presentation, August 11, 2009

- Science for National Need – Delivering forefront scientific knowledge and state-of-the-art tools to serve the nation
- Many areas of research require leadership computing power for discovery
- Even with the newly harnessed petascale resources, many critical simulations are limited by computing power
- ASCR's strategic directions for 2010 and beyond continues to be:
  - Advance the state of the art in computational capability
  - Develop tools and methods for harnessing that capability
  - Bring this capability to bear on scientific questions with national need:
    - **Climate Modeling** – half of the runs for the U.S. contribution to the IPPC AR4 were done on ASCR computing systems; the Earth Systems Grid is the primary mechanism for sharing this data; and SciDAC supports the development of next generation codes
    - **Combustion** – advances in computing power move combustion simulations closer to real world conditions and provides new insights into how to improve fuel efficiency and reduce emissions
    - **Bioenergy** – working with DOE Bioenergy Research Centers, simulations of enzymes breaking down cellulose will help make cellulosic ethanol (energy from non-food crops) an economically viable option
    - **Nuclear Energy** – the worlds largest simulation of a reactor core was achieved at the ALCF and will help engineers to enhance safety and reduce waste in next generation reactors
    - **Fusion** – advances in computing power move fusion simulations closer to real world conditions and help engineers improve efficiency and design control systems
    - **Advanced Materials** – including the worlds first petascale application modeling superconductivity

OLCF-3 Applications Analysis

informed by two requirements surveys

- Project application requirements
  - Elicited, analyzed, and validated using a new comprehensive requirements questionnaire
  - Project overview, science motivation & impact, application models, algorithms, parallelization strategy, S/W, development process, SQA, V&V, usage workflow, performance
  - Results, analysis, and conclusions documented in 2009 OLCF application requirements document
- OLCF-3 baseline plan developed in consultation with 50+ leading scientists in many domains
  - What are the science goals and does OLCF-3 enable them?
  - What might the impact be if the improved science result occurs?
  - What does it matter if this result is delivered in the 2012 timeframe?



PREPARING FOR EXASCALE

ORNL Leadership Computing
Application Requirements and Strategy

October 2009

OLCF-3 Science Driver Survey

Science driver

- What science will be pursued on this system and how is it different (in fidelity/quality/predictability and/or productivity/throughput) from the current system?

Science impact

- What might the impact be if this improved science result occurs? Who cares and why?

Science timeliness

- If this result is delivered in the 2012 timeframe, what does it matter as opposed to coming 5 years later (or never at all)? What other programs, agencies, stakeholders, and/or facilities are dependent upon the timely delivery of this result and why?

OLCF-3 Applications Analysis

select to optimize readiness activities

- Look at application candidates from all relevant perspectives
  - Science
    - Science results, impact, timeliness
    - Alignment with DOE and US science mission (CD-0)
    - Broad coverage of science domains
  - Implementation (models, algorithms, S/W)
    - Broad coverage of relevant programming models, environment, languages, implementations
    - Broad coverage of relevant algorithms and data structures (motifs)
    - Broad coverage of scientific library requirements
  - User community (current and anticipated)
    - Broad institutional and developer/user involvement
    - Good representation of current and anticipated INCITE workload
  - Preparation for steady state ("INCITE ready") operations
    - Mix of low ("easy") and high ("hard") risk porting and readiness requirements
    - Availability of OLCF liaison with adequate skills/experience match to application
    - Availability of key code development personnel to engage in and guide readiness activities

Surveyed >50 scientists in 12 different science domains to understand application requirements

OLCF-3 Applications Analyzed

science outcomes were elicited from a broad range of applications

Application area	Application codes	Science target
Astrophysics	Chimera, Genasis	• Core-collapse supernovae simulation validation against observations of neutrino signatures, gravitational waves, and photon spectra.
	MPA-FT, Sne, MAESTRO, SEDONA	• Core-collapse supernovae • Full-start type 1a supernovae simulations of thermonuclear runaway with realistic subgrid models
Bioenergy	LAMMPS, GROMACS	• Cellulosic ethanol: dynamics of microbial enzyme action on biomass
Biology	LAMMPS	• Systems biology • Genomic structure
Chemistry	CP2K, CPMD	• Interfacial chemistry
	GAMESS	• Atmospheric aerosol chemistry • Fuels from ligno-cellulosic materials
Combustion	S3D	• Combustion flame front stability and propagation in power and propulsion engines
	RAPTOR	• Internal combustion design in power and propulsion engines: bridge the gap between device- and lab-scale combustion
Energy Storage	MADNESS	• Electrochemical processes at the interfaces; ionic diffusion during charge-discharge cycles

OLCF-3 Applications Analyzed

science outcomes were elicited from a broad range of applications

Application area	Application codes	Science target
Fusion	GTC	• Energetic particle turbulence and transport in ITER
	GTS	• Electron dynamics and magnetic perturbation (finite-beta) effects in a global code environment for realistic tokamak transport. • Improved understanding of confinement physics in tokamak experiments. • Address issues such as the formations of plasma critical gradients and transport barriers.
	XGC1	• First-principles gyrokinetic particle simulation of multiscale electromagnetic turbulence in whole-volume ITER plasmas with realistic diverted geometry
	AORSA, CQL3D	• Tokamak plasma heating and control
	FSP	• MHD scaling to realistic Reynolds numbers • Global gyrokinetic studies of core turbulence encompassing local & nonlocal phenomena and electromagnetic electron dynamics • Finite ion-orbit-width effects to realistically assess RF wave resonant ion interactions • Multi-scale integrated electromagnetic turbulence in realistic ITER geometry
	GYRO, TGYRO	• Predictive simulations of transport iterated to bring the plasma into steady-state power balance; radial transport balances power input
Geoscience	PFLOTRAN	• Stability and viability of large-scale CO_2 sequestration • Predictive contaminant ground water transport

OLCF-3 Applications Analyzed

science outcomes were elicited from a broad range of applications

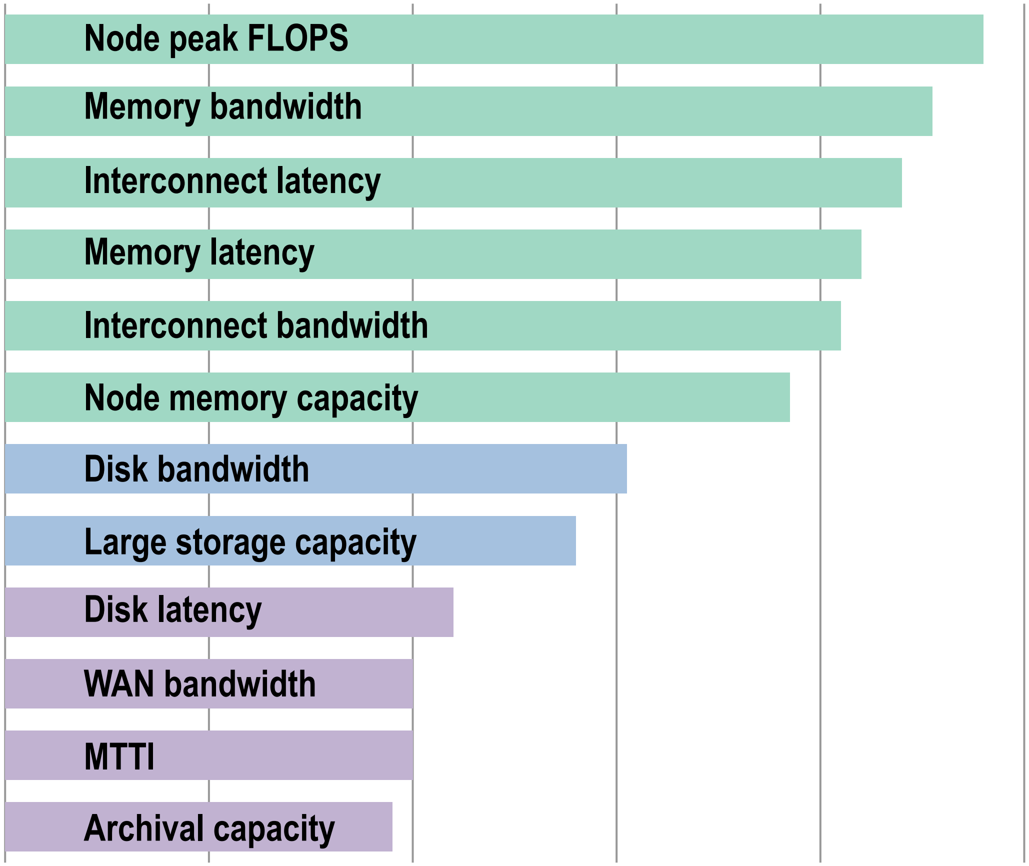
Application area	Application codes	Science target
Nanoscience	OMEN	• Electron-lattice interactions and energy loss in full nanoscale transistors
	LS3DF	• Full device simulation of a nanostructure solar cell
	DCA++	• Magnetic/superconducting phase diagrams including effects of disorder • Effect of impurity configurations on pairing and the high-T superconducting gap • High-T superconducting transition temperature materials dependence in cuprates
	WL-LSMS	• To what extent do thermodynamics and kinetics of magnetic transition and chemical reactions differ between nano and bulk? • What is the role of material disorder, statistics, and fluctuations in nanoscale materials and systems?
Nuclear energy	Denovo	• Predicting, with UQ, the behavior of existing and novel nuclear fuels and reactors in transient and nominal operation
	UNIQ	• Reduce uncertainties and biases in reactor design calculations by replacing existing multi-level homogenization techniques with more direct solution methods
Nuclear Physics	NUCCOR MFDn	• Limits of nuclear stability, static and transport properties of nucleonic matter • Predict half-lives, mass and kinetic energy distribution of fission fragments and fission cross sections
QCD	MILC,Chroma	• Achieving high precision in determining the fundamental parameters of the Standard Model (masses and mixing strengths of quarks)
Turbulence	DNS	• Stratified and unstratified turbulent mixing at simultaneous high Reynolds and Schmidt numbers
	Hybrid	• Nonlinear turbulence phenomena in multi-physics settings

What do the Science Codes Need?

What system features do the applications need to deliver the science?

- 20 PF in 2011–2012 time frame with 1 EF by end of the decade
- Applications want powerful nodes, not lots of weak nodes
  - Lots of FLOPS and OPS
  - Fast, low-latency memory
  - Memory capacity ≥ 2GB/core
- Strong interconnect

Application Team's priority ranking of most important system characteristics



Application team concerns

Algorithmic	Compilers and tools
• ability to identify and extract suitable kernels for acceleration – self-contained, with enough parallelism • ability to achieve sufficient concurrency to amortize overhead – data movement, thread creation, etc.	• ability to use particular language features (e.g., C++ templates) • acceleration of kernels in applications undergoing active development • ability to accelerate without generating a large amount of non-portable code

Coverage Table for 10 Sample Applications

App	Science Area	Algorithm(s)	Grid type	Programming Language(s)	Compiler(s) supported	Communication Libraries	Math Libraries	Acceleration status	Difficulty (trivial=1, impossible=5)
CAM-HOMME	climate	spectral finite elements, dense & sparse linear algebra, particles	structured	F90	PGI, Lahey, IBM	MPI	Trilinos	2 kernels identified for acceleration	4
DCA++	materials	quantum Monte Carlo, FFT, dense linear algebra	N/A	C++	GNU	MPI	DGEMM	identified kernel for acceleration, implemented test kernel in CUDA	2
LAMMPS	biology	molecular dynamics, FFT, particles	N/A	C++	GNU, PGI, IBM, Intel	MPI	FFT	Indentified kernals that account for 80% of time, needs major code restructuring to tale advantage of GPUs. Planning in progress.	3
MADNESS	chemistry, etc.	dense & sparse linear algebra, multiwavelet analysis	un-structured	C++	GNU	MPI, pthreads	DGEMM	identified kernel for acceleration, preliminary implementation in progress	3
S3D	combustion	Navier-Stokes, finite diff, dense & sparse linear algebra, particles	structured	F77, F90	PGI	MPI	None	1 CUDA kernel	3
Chimera	astrophysics	finite-volume hydro, flux-limited diffusion, dense linear solve for nuclear kinetics	structured	F77, F90	PGI, Cray, IBM, Intel	MPI	LAPACK	identified kernel for acceleration	3
Denovo	nuclear energy	wavefront sweep, GMRES	structured	C++, Fortran, Python	GNU, PGI, Cray, Intel	MPI	Trilinos, LAPACK, SuperLU, Metis	identified kernel accounting for 80% of time, developed plan for kernel accel, began CUDA implementation	3
AORSA	fusion	dense linear algebra, definite integral	structured	F90	PGI	MPI	FFT, ScaLAPACK, ZHPL	experimenting with integral eval, tracking external HPL efforts	3
gWL-LSMS	nanoscience	density functional theory, Monte Carlo	N/A	F77, F90, C, C++	PGI, GNU	MPI	ZGEMM	identified kernel for implementation	2
GTS	fusion	particle-in-cell	unstructured	F90, C, C++	PGI, Cray	MPI, OpenMP	PETSc, FFT	accelerating particle push	3

Application algorithm motifs and system attribute pressure points

support a more capable node and interconnect

Code	Structured grids	Unstructured grids	FFT	Dense linear algebra	Sparse linear algebra	Particles	Monte Carlo
S3D	X			X	X	X	
CAM	X		X	X	X	X	
gWL-LSMS			X	X			X
LAMMPS			X			X	
GTS		X	X		X	X	
Denovo	X			X	X	X	X

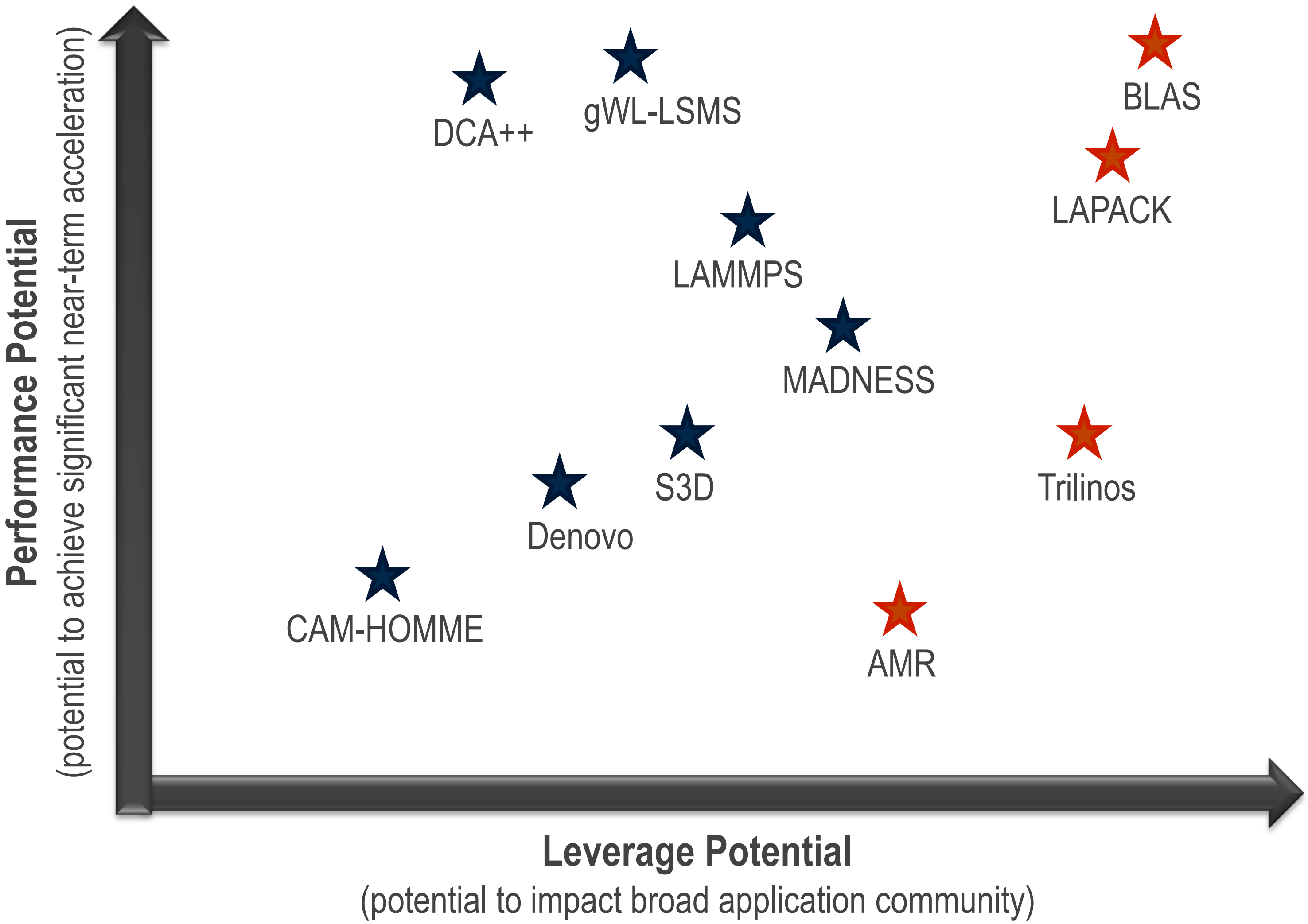
Motif	Peak node FLOPs	Memory capacity	Interconnect latency	Memory bandwidth
FFT	X		X	
Dense Linear Algebra	X	X		
Sparse Linear Algebra			X	X
Unstructured Grids			X	X
Particles		X	X	

Leveraging Philosophy / Approach

Making sure OLCF readiness activities & products are not "one offs"

- **interact with broader community**
  - Hybrid Multicore Consortium
  - OpenCL Consortium
  - compiler & tool development efforts
- **work closely with NVIDIA**
  - leverage expertise for near-term performance
  - leverage & influence via community connections
- **choose applications and libraries with broad relevance**
  - BLAS/LAPACK(MAGMA) already included via Tier 1 apps
  - widely-used Trilinos library
  - AMR
- **document and share acceleration experiences**
  - software/repositories, reports, lessons learned, workshops

Performance-Leverage Map



"INCITE Ready"

What does it mean & how do we get there

- **Today on Jaguar for *most* (but not all!) new applications**
  - New application shows up at "time zero" and it might compile in <1 week
  - 1 month: execute at scales consistent with current application performance
  - 2-3 months: execute at larger scales, but similar node performance
  - 5-6 months: execute at still larger scales, improved performance
  - 6-12 months: execute at leadership scales, "optimized" node performance
- **On OLCF-3 with a "brand new" GPU-ignorant application arriving "on day one"**
  - <1 week: compile and execute correctly (remember – we still have our "same old CPUs")
  - 1 month: compile and execute on CPUs at scales consistent with current application performance
  - 2-3 months: GPU utilization via available libraries and kernels
  - 5-6 months: new GPU kernels identified; some created and running
  - 6-12 months: GPU kernels identified to enable "fully" utilized acceleration
  - >12 months: more intrusive restructure/redesign to enable aggressive acceleration

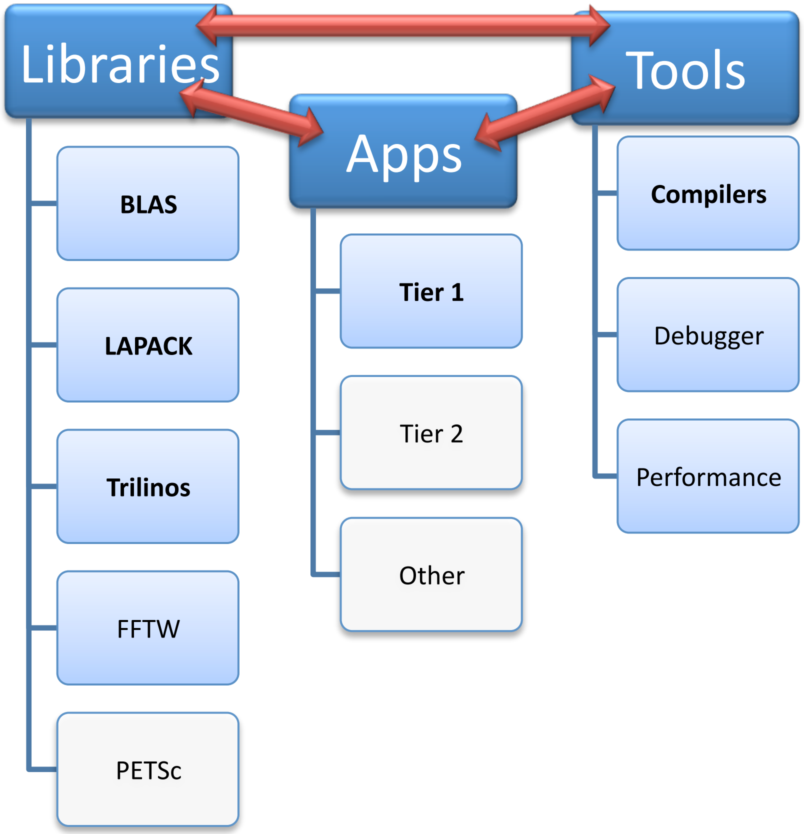
This paradigm is not really different than vectorization, cache coherency, SSE utilization, & other accelerators (e.g., "hummers")

We've done this before and have been doing this all along!

Center for Accelerated Application Readiness (CAAR) has been formed.

Three integrated teams are preparing applications for OLCF-3.

Applications	• 6 Tier 1 applications • 1 domain scientist, 1 computer / computational scientist for each application • multiple paths for each application – evolution – re-factor / re-design
Libraries	• staff focused on key libraries
Tools (including compilers)	• staff interface between applications and partners • identify common features and requirements across applications • fold recommendations into compiler / tools chain



Generic Application Readiness Process

- define test problem(s)
- baseline CPU performance
- identify kernel(s)
- understand kernel performance
- extract kernel(s)
- if appropriate…
  - create standalone driver(s)
  - define standalone test problem(s)
- provide kernel(s) to NVIDIA
- accelerate kernel(s)
  - e.g. implement kernel(s) in CUDA
- understand accelerated kernel performance & optimize
- implement accelerated kernel(s) in application
- understand accelerated application performance

Process customized for each application as appropriate

Acceleration paths

- applications assessing various paths to acceleration
- CAAR cannot impose path on app teams
- below is a subjective assessment of options
  - "1" is "best", ranking only

Path	Difficulty / Impact	C++ Support	Portability	When?	Performance Potential
CUDA	2	2	3 (NVIDIA)	1	1
Directives (incl. OpenMP)	1	depends	2 (compiler-specific)	2	2
OpenCL	3	host – yes accel – no	1	3	3

Summary

- The OLCF has a clear message from numerous sources about the science requirements and goals it must support
- The OLCF has implemented and carried out a detailed, bottoms-up applications analysis for OLCF-3
  - 50+ scientists surveyed and 30+ applications considered
- Applications yield excellent coverage from all relevant perspectives: science, models, algorithms, S/W, leveraging potential, institutions, difficulty
- The Center for Accelerated Application Readiness (CAAR) is chartered and executing
  - ~40 staff engaged from multiple institutions (NVIDIA, Cray, universities, DOE labs)
- We will use metrics to track progress and inform course corrections
  - How “ready” are the apps and how complex are they getting? What is the effort expended to date? What is the performance payoff?

Questions?

John Turner
Email: TurnerJA@ORNL.Gov

The research and activities described in this presentation were performed using the resources of the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC0500OR22725.