High-Performance Computing Research at Oak Ridge National Laboratory

Dr. Christian Engelmann

Computer Science and Mathematics Division
Oak Ridge National Laboratory, USA

Largest Multipurpose Science Laboratory within the U.S. Department of Energy



- Privately managed for US DOE
- $1.1 billion budget
- 4200 employees total
- 3,000 research guests annually
- 30,000 visitors each year
- Total land area 58mi^2 (150km^2)

- Nation's largest energy laboratory
- Nation's largest science facility:
  - The $1.4 billion Spallation Neutron Source
- Nation's largest concentration of open source materials research
- Nation's largest open scientific computing facility

ORNL East Campus: Site of World Leading Computing and Computational Sciences



National Center for Computational Sciences



- 40,000 ft^2 (3700 m^2) computer center:
  - 36-in (~1m) raised floor, 18 ft (5.5 m) deck-to-deck
  - 12 MW of power with 4,800 t of redundant cooling
  - High-ceiling area for visualization lab:
    - 35 MPixel PowerWall, Access Grid, etc.

- 4 systems in the Top 500 List of Supercomputer Sites:
  - Jaguar: 2. Cray XT5, MPP with 37544 quad-core Processors ⇨ 1 381 TFlop
  - Jaguar: 8. Cray XT4, MPP with 7744 quad-core Processors ⇨ 260 TFlop
  - Kraken: 15. Cray XT4, MPP with 4489 quad-core Processors ⇨ 165 TFlop
  - Eugene: 130. IBM Blue Gene/P, MPP with 2048 quad-core Processors ⇨ 27 TFlop



Jaguar: World's most powerful computer— Designed for science from the ground up



Peak performance	**1.645 PF**
System memory	362 TB
Disk space	10.7 PB
Disk bandwidth	240+ GB/s
Interconnect bandwidth	532 TB/s

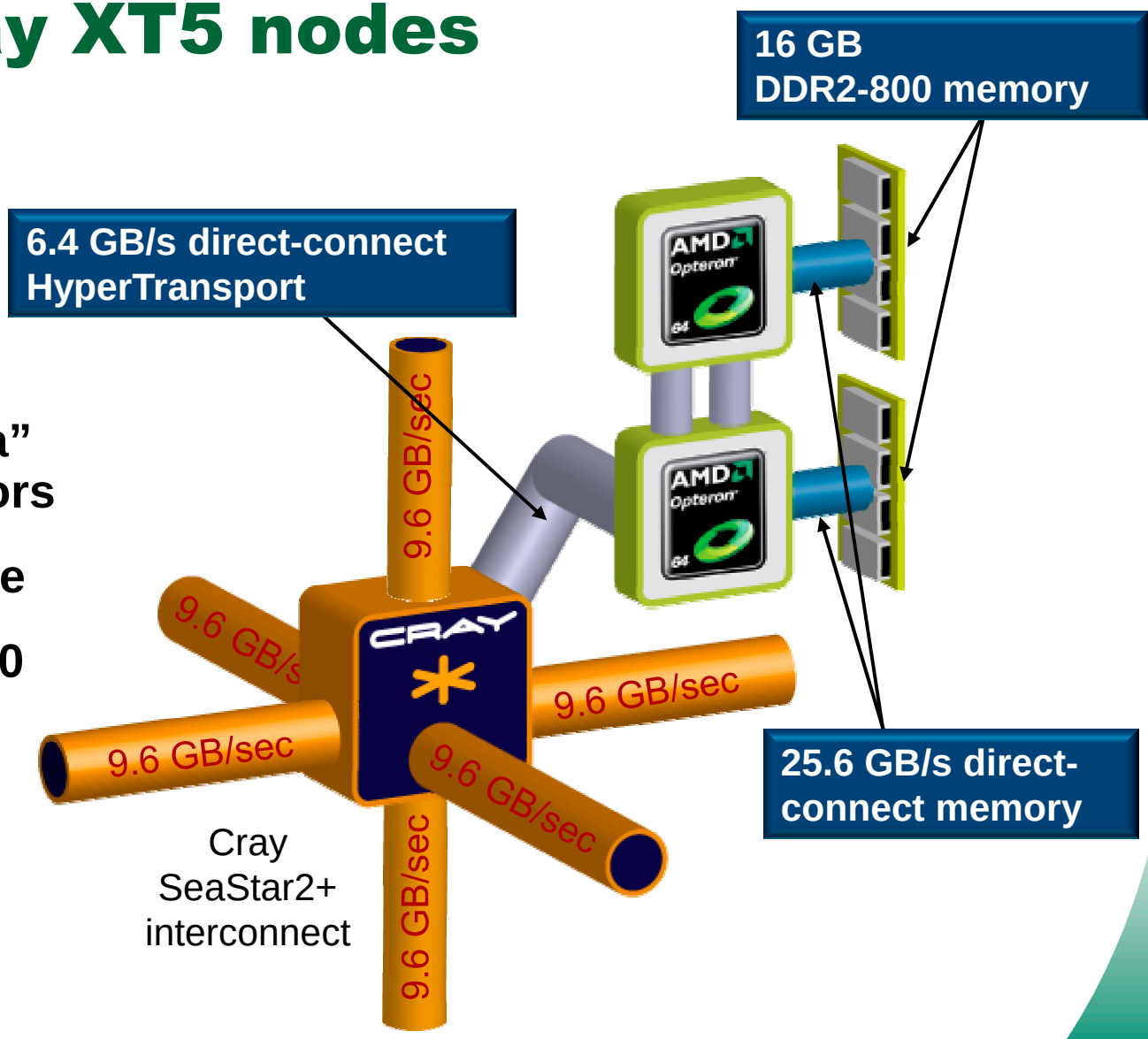
What does the system look like?

Jaguar combines the existing 263 TF Cray XT4 system at ORNL's NCCS with a new 1,382 TF Cray XT5 to create a 1.64 PF system

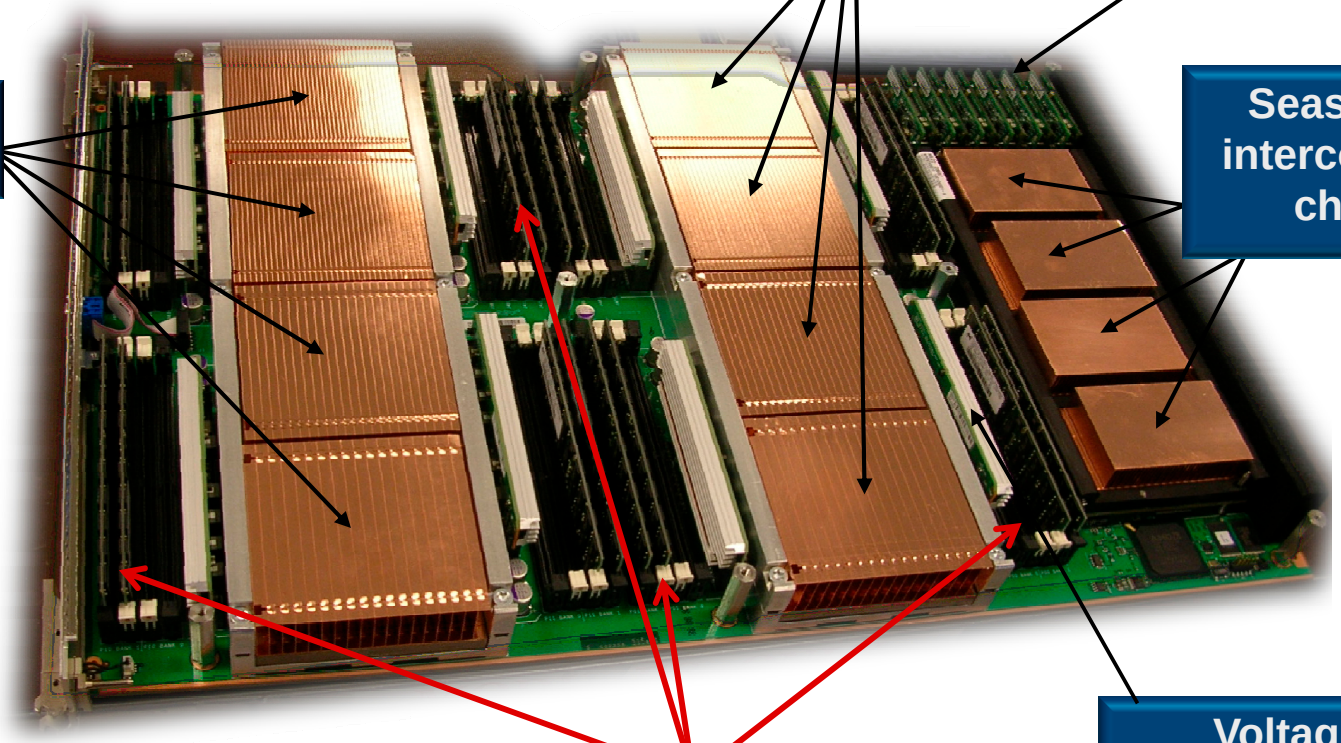
System attribute	XT5	XT4
Quad-core AMD Opteron processors	37,544	7,832
Node architecture	Dual socket SMP	Single Socket
Memory per core/node (GB)	2/16	2/8
Total system memory (TB)	300	62
Disk capacity (TB)	10,000	750
Disk bandwidth (GB/s)	240	44
Interconnect	SeaStar2+ 3D torus	SeaStar2+ 3D torus

Jaguar's Cray XT5 nodes

- Eight-core SMP
- Two 2.3 GHz AMD Opteron "Barcelona" quad-core processors
- 73.6 Gflops per node
- Four 4 GB DDR2-800 DIMMS with four empty DIMM slots per node
- OpenMP Support



Jaguar's Cray XT5 blades



- 4 nodes per blade
- 294.4 GF
- 64 GB
- 4 network connections

24 Cray XT5 blades make a Jaguar cabinet



One Jaguar XT5 cabinet:

- 7 TF
- 192 Opteron processors
- 776 Opteron cores
- 1,536 GB memory
- 96 nodes
- 1•4•24 section of 3D torus

200 cabinets complete the XT5 side of the system



Compute partition

- Nodes: 18,772
- Opteron processors: 37,544
- Processor cores: 150,176
- Peak TF: 1.38
- Memory: 300 TB
- Interconnect: 25•32•24 3D torus

Service and I/O partition

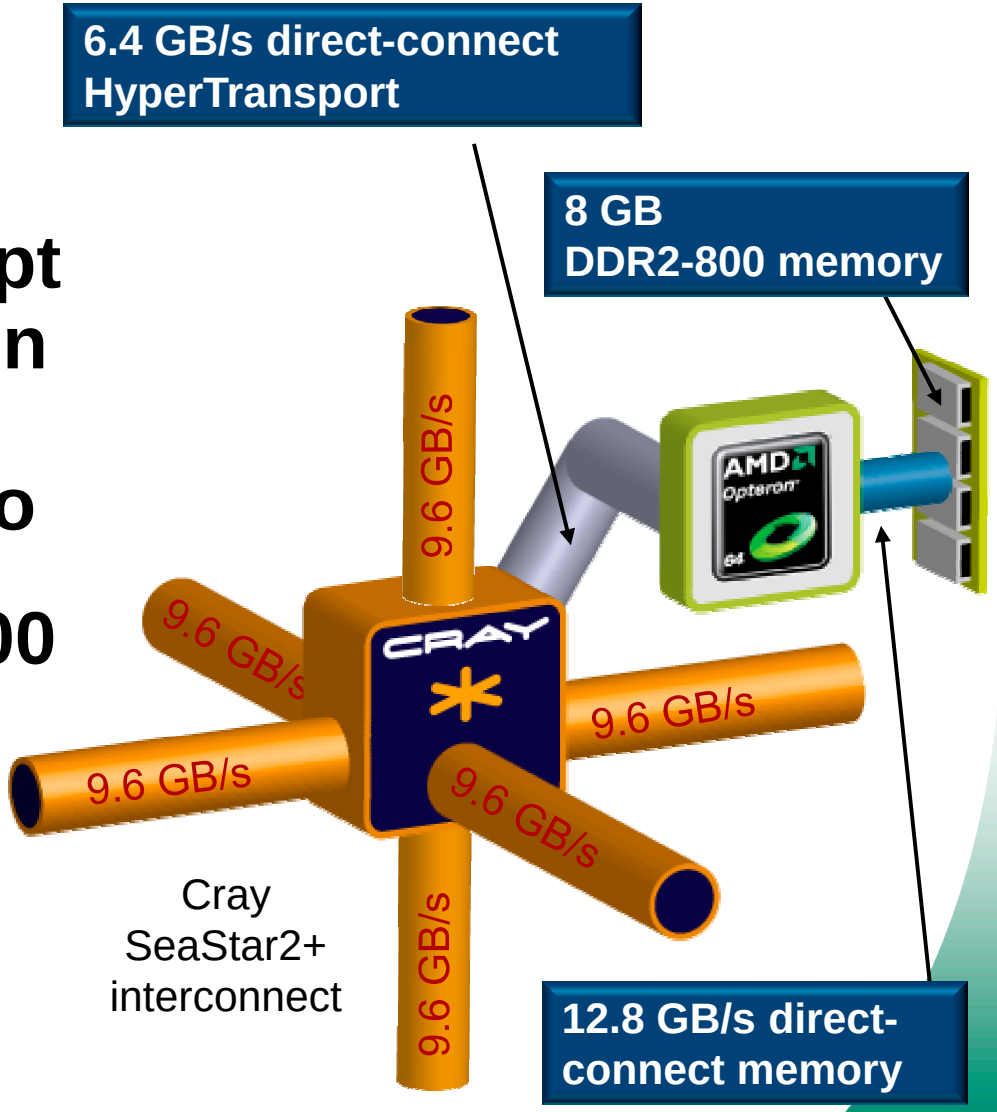
- Nodes: 214
- InfiniBand cards: 192 DDR
- Memory per SIO node: 8 GB
- 10 Gigabit Ethernet: 10

Power and cooling

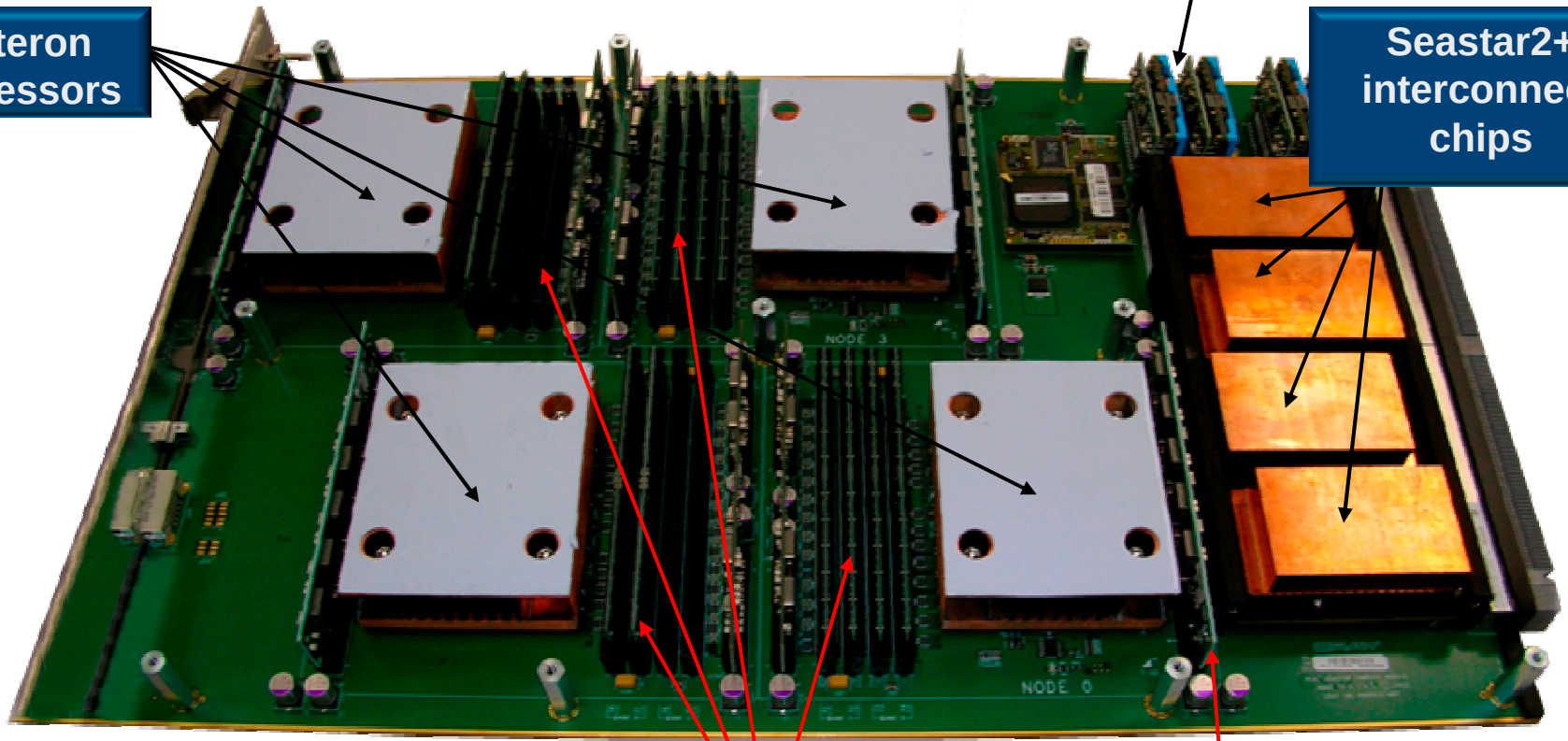
- 480 V power supplies
- R134a refrigerant
- Single large fan per cabinet

How is the XT4 side different?

- Basic node design is the same as XT5, except it has one AMD Opteron quad-core processor per node instead of two
- Same 2 GB of DDR2-800 memory per core
- Same Seastar2+ interconnect

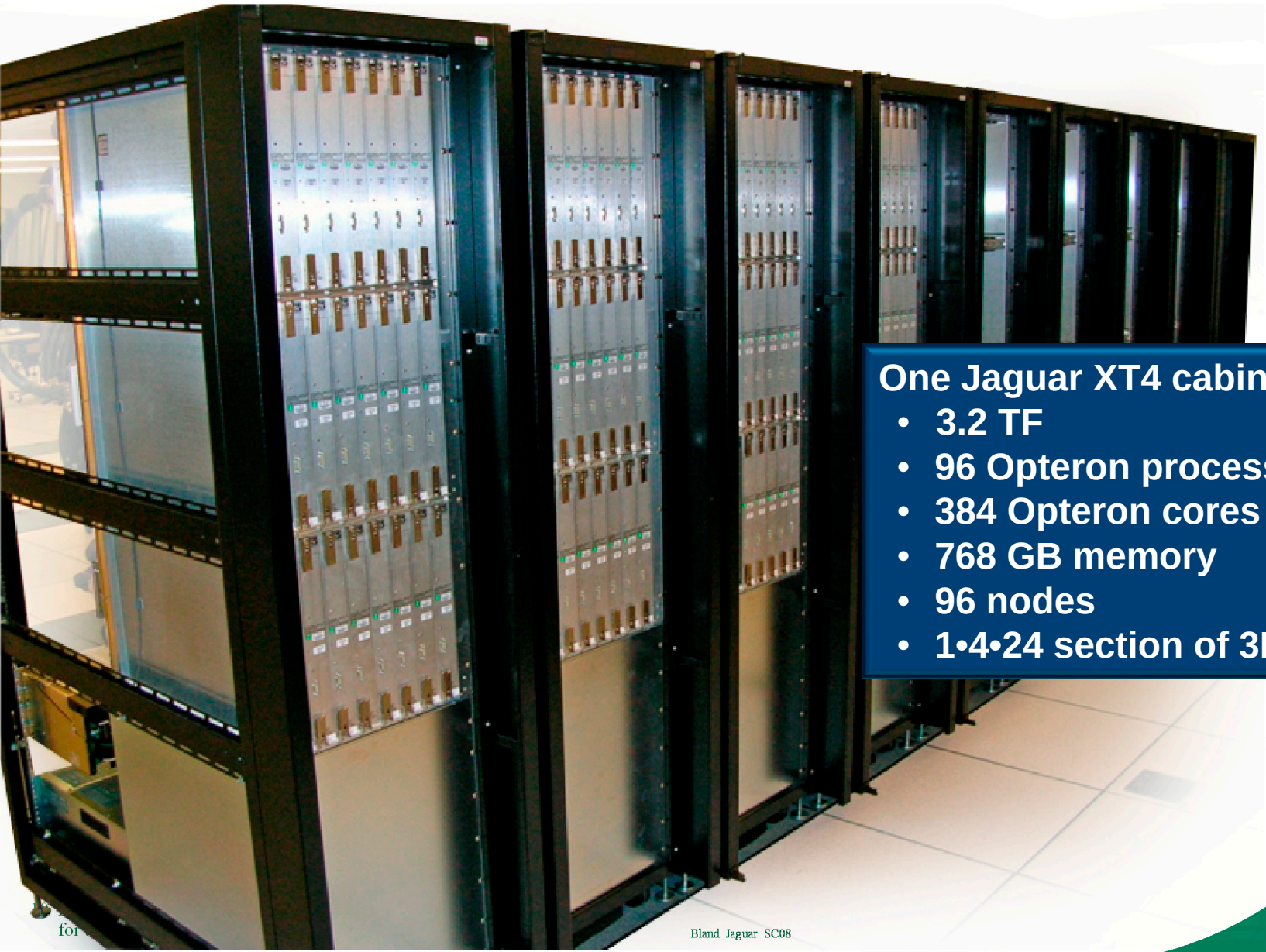


The XT4 board has four nodes



- 4 nodes per blade
- 134.4 GF
- 32 GB
- 4 network connections

And 24 Cray XT4 blades make a Jaguar cabinet



One Jaguar XT4 cabinet:

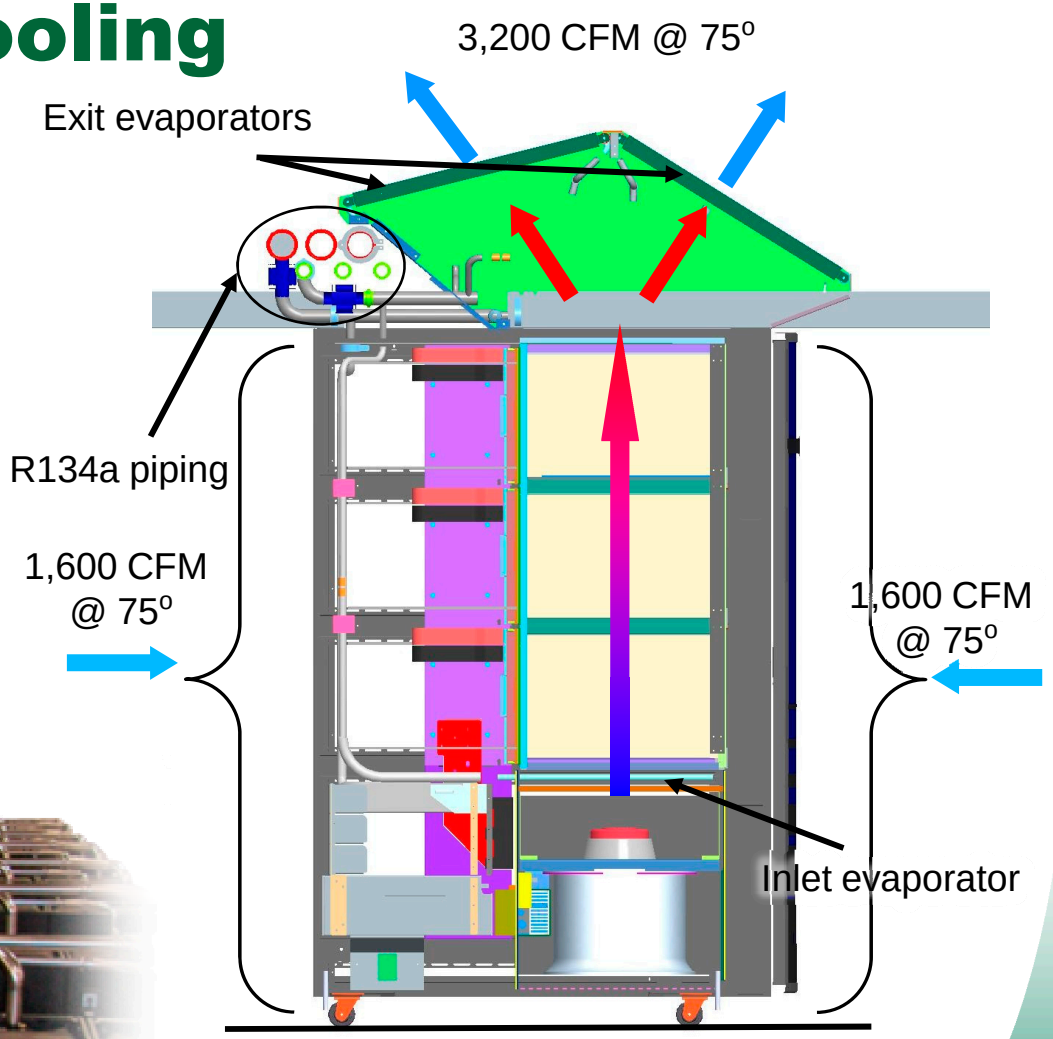
- 3.2 TF
- 96 Opteron processors
- 384 Opteron cores
- 768 GB memory
- 96 nodes
- 1•4•24 section of 3D torus

OAK RIDGE National Laboratory

ECOphlex liquid cooling

- **Liquid-cooled design exhausts heat to R134a before it leaves the cabinet. Replaces 100 CRAC units!**
- **Saves about 900 KW of power in air movement alone**





- **Phase change of liquid to gas removes heat much more efficiently**

OAK RIDGE National Laboratory

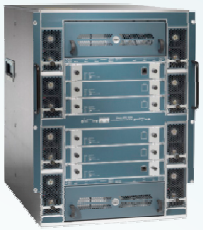
480 V power saves money

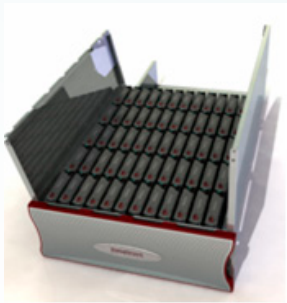


- Power supply efficiency
  - Worked with Cray to select highest-efficiency power supplies
- 480 V power to the cabinet saves the inefficiency of converting from 480 V to 208 V to 48 V
- Keeping the voltage high saved $1M in installation and $500K in electrical costs

Centerwide file system







- "Spider" is being installed to provide a shared, parallel file system for all systems
  - Based on Lustre file system
- Bandwidth of more than 240 GB/s
- More than 10 PB of RAID6 Capacity
  - 13,440 1 TB SATA drives
- 192 storage servers
  - 3 TB of memory
  - 14 TF
- Available from all systems via our high-performance scalable I/O network
  - More than 3,000 InfiniBand ports
  - More than 3 miles of cables
  - Scales as storage grows
- Engineered for high availability

Jaguar combines a new 1.38 PF Cray XT5 with the existing 263 TF Cray XT4





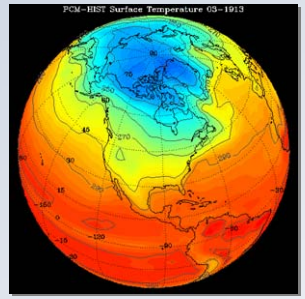


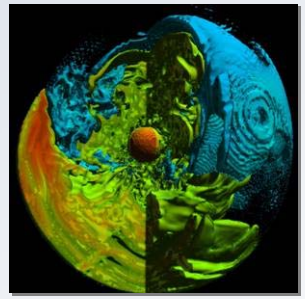


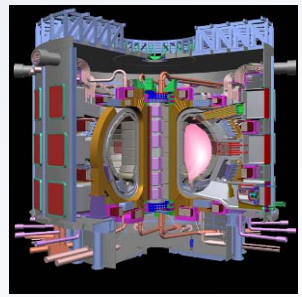
System components are linked by 4×-DDR InfiniBand (IB) using three Cisco 7024D switches

- **XT5 has 192 IB links**
- **XT4 has 48 IB links**
- **Spider has 192 IB links**

At Forefront in Scientific Computing and Simulation









- Leading partnership in developing the National Leadership Computing Facility
  - Leadership-class scientific computing capability
  - 1 PFlop/s in 2008 (upgrade recently finished)
  - 500 TFlop/s in 2009 (installation in progress)
  - 1 PFlop/s in 2009 (commitment made)
- Attacking key computational challenges
  - Climate change
  - Nuclear astrophysics
  - Fusion energy
  - Materials sciences
  - Biology
- Providing access to computational resources through high-speed networking

ORNL provides leadership computing to 2008 INCITE program

- The NCCS is providing leadership computing to 30 projects in 2008 under DOE's Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program
- Leading researchers from government, industry, and the academic world will use more than 75 million processor hours on the center's Cray leadership computers
- The center's Cray XT4 (Jaguar) and Cray X1E (Phoenix) systems provide more than 75% of the computing power allocated for the INCITE program



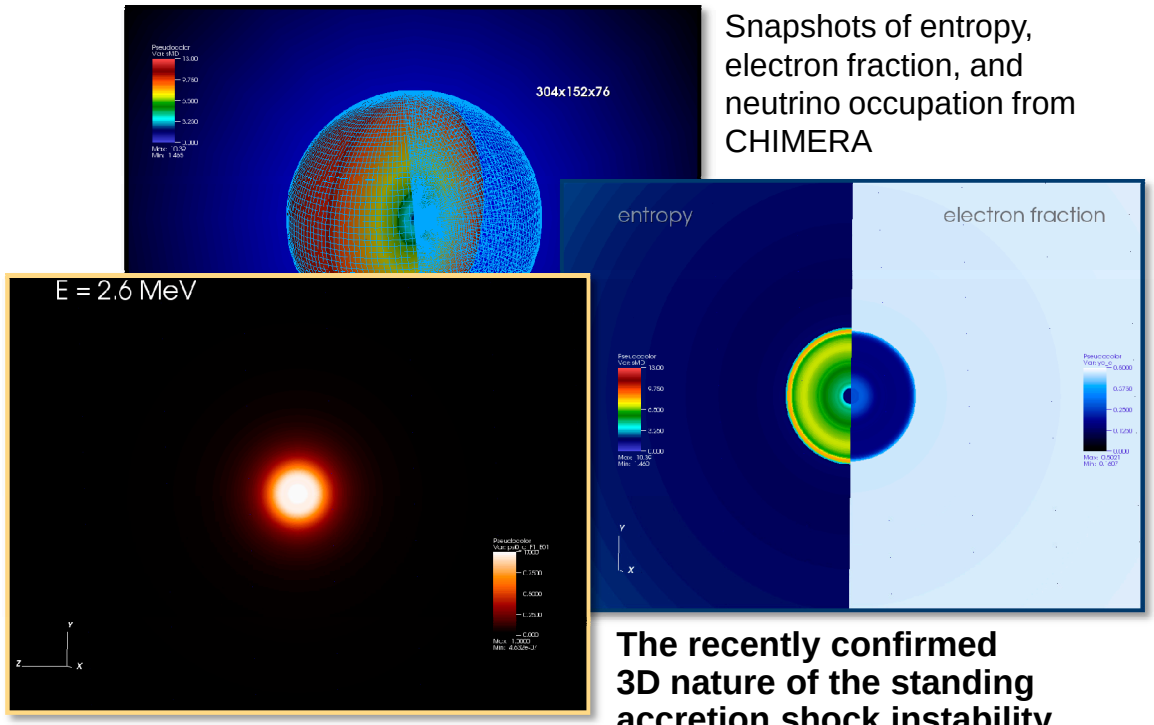
Project allocations: 145.3 million hours

Industrial allocations: 11.9 million hours

Astrophysics

Three-dimensional supernova explosions simulations with high fidelity physics

- The world's first 3D supernova simulations incorporating multifrequency neutrino transport are currently under way on Jaguar
- 304×152×76 spatial mesh running on 11,552 CPUs
- Initial results are promising, but full evolution of the explosion epoch will require millions of hours



Snapshots of entropy, electron fraction, and neutrino occupation from CHIMERA

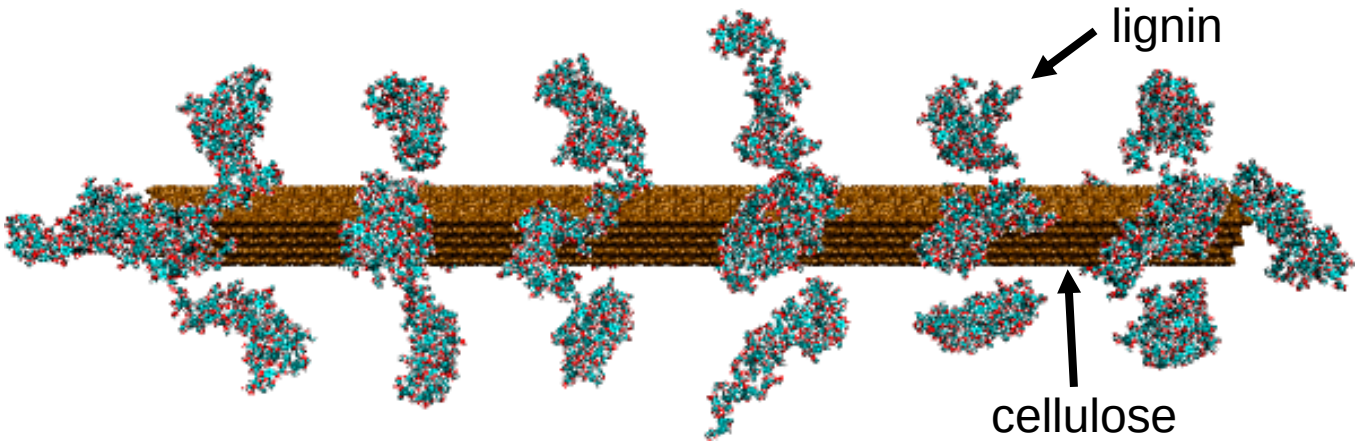
The recently confirmed 3D nature of the standing accretion shock instability and the known 3D nature of convective overturn make performing supernova simulations in 3D essential

LCF liaison contributions

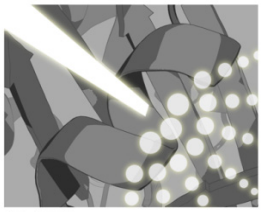
- Implementing efficient, collective I/O
- New visualization tools
- Improved software management

Biology

Cellulosic ethanol: Physical basis of recalcitrance to hydrolysis of lignocellulosic biomass



Simulations provide physical insight into physical origins of biomass recalcitrance and cellulosome organization and assembly



Center for Molecular Biophysics

ORNL

Loukas Petridis, Benjamin Lindner, and Jeremy C. Smith

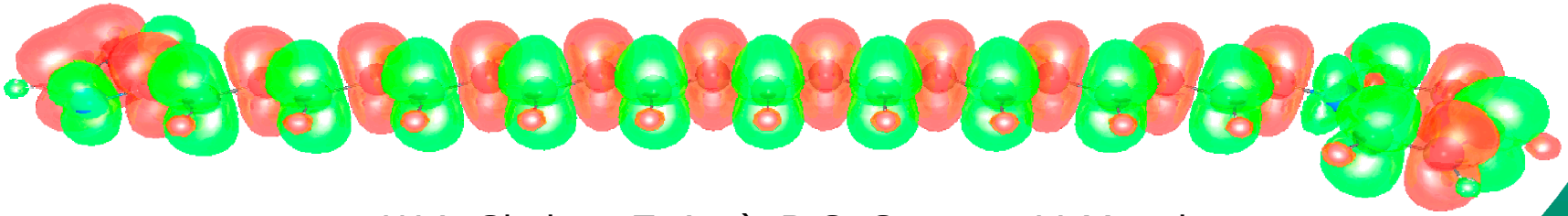
See L. Petridis and J.C. Smith, *J. Comp. Chem*. In press.

- **Understanding biomass structure is key to overcoming recalcitrance**
- **Large-scale molecular dynamics simulation (1−3M atoms)**
  - **Currently using NAMD code with 1.9M atoms for lignocellulose (on 6072 cores)**
- **Results to be used to interpret biophysical experiments**
  - **Lignin-cellulose simulations will be used to calculate small-angle neutron scattering intensities, which will be compared with experimental data obtained at the Spallation Neutron Source**

Chemistry
Organic molecular magnets from zwitterionic molecules

PI: Robert Harrison, ORNL

- **Hybrid DFT calculations of the electronic structure of a new class of zwitterionic molecules (betaine derivatives) reveals an unexpected magnetic ground state**
- **Nanochains with a maximum length of 10.4 nm and over 1700 basis functions. Over 500 Jaguar processors used. Scaling with respect to system size is $O(N^2)$**
- **Origin of the magnetism is quite complex and not well understood as compared to more conventional magnetic systems**
- **Magnetic materials that are purely organic are rather rare, and the betaine derivatives promise to offer a new class of molecular magnetic materials with diverse potential applications including organic spintronics, information storage, and nanoscale sensors**



W.A. Shelton, E. Aprà, B.G. Sumpter, V. Meunier

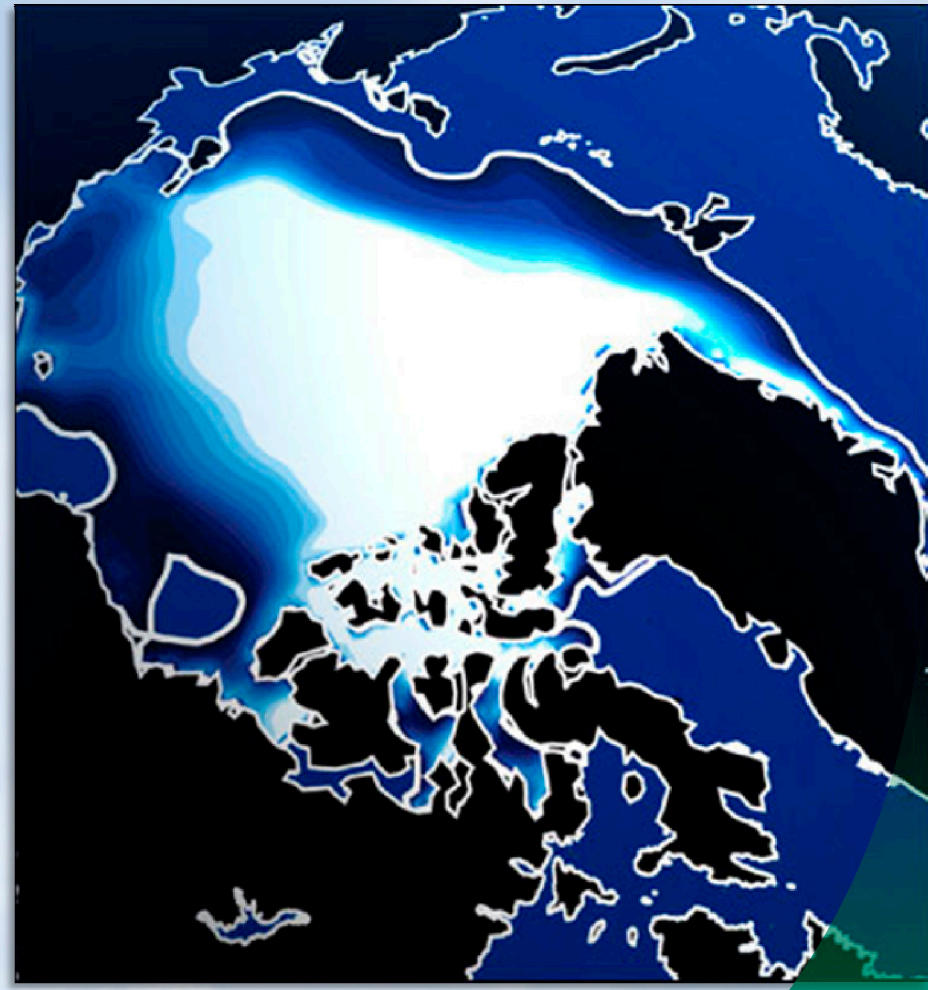
Climate

Modeling effects and agents of abrupt climate change

PI: Zhengyu Liu, University of Wisconsin

This project is addressing two fundamental questions on future climate change

- **What is the sensitivity of the climate system to the change of greenhouse gases, notably CO_2?**
- **How does the climate system exhibit abrupt changes on decadal-centennial timescales?**



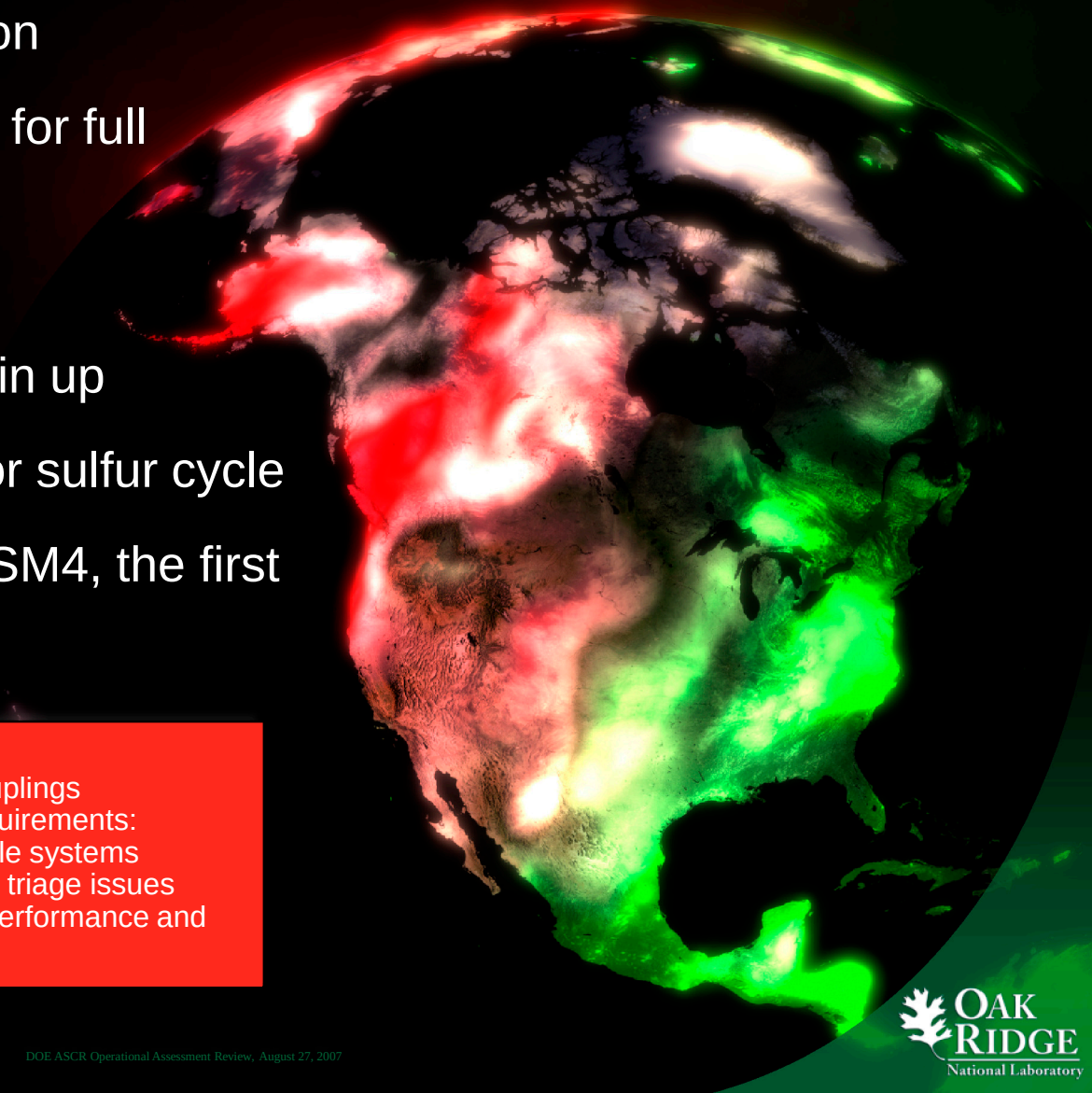
CCSM simulation of the 21,000-year last deglaciation

PI: Warren Washington, NCAR

Climate
Modeling the full Earth system

- New standard and low-emission climate scenarios at higher resolution
- Final stage of ocean spinup for full Earth System Model
- 100 years of ocean spin up
- 20 years of carbon-cycle spin up
- Approaching steady state for sulfur cycle
- Approaching release of CCSM4, the first Earth System Model

Simulated time evolution of the atmospheric CO_2 concentration originating from the land's surface



LCF liaison contributions
- Advanced visualizations of new climate couplings
- Tuning of LCF infrastructure for climate requirements: special queues, reservations, dedicated file systems
- Weekly conference calls with developers to triage issues
- Development of timing libraries to assess performance and balance components

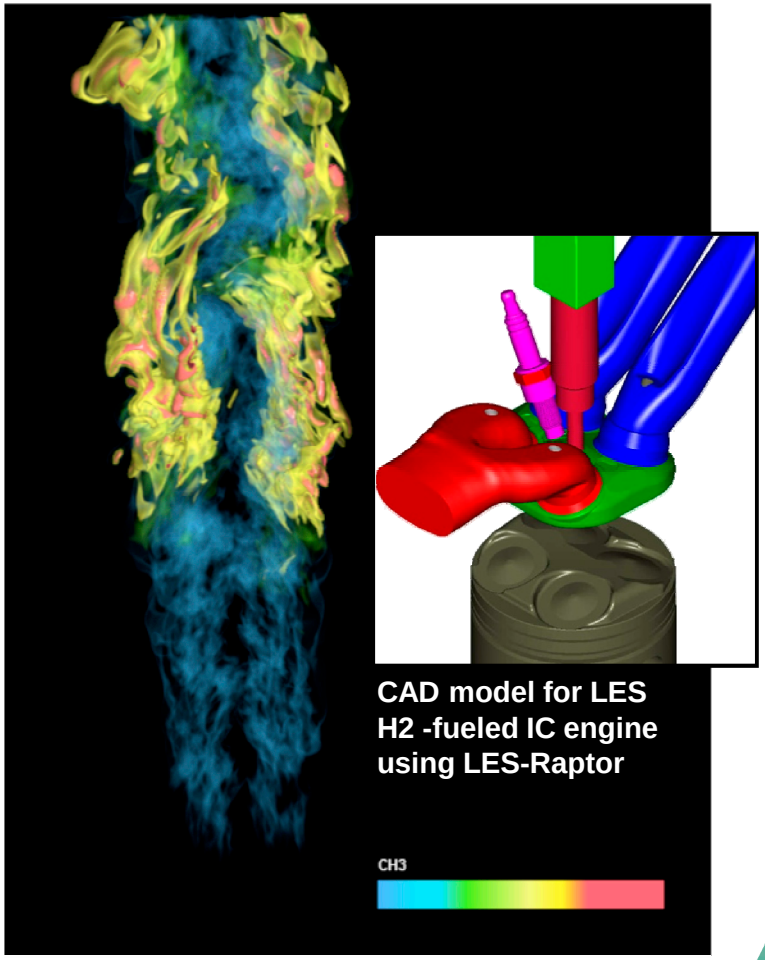
Engineering

Flame stabilization in a lifted hydrocarbon flame

- Lifted flames occur in diesel engines and gas turbine combustors
  - Flame stabilized against fuel jet and recirculating hot gases
- Source of flame stabilization was revealed in a simulation of a hydrogen-air lifted flame
  - Upstream autoignition
  - Vorticity generation at flame base due to baroclinic torque
- Direct numerical simulation of a 3D lifted autoigniting ethylene-air jet flame
  - Reynolds number = 10,000
  - 1.3B grid points and 22 chemical species
  - 50 TB of data generated
- Thirteen publications/presentations in just last quarter!

LCF liaison contributions

- Distributed Lagrangian particle-tracking algorithm and S/W
- Ethylene chemistry performance improvement

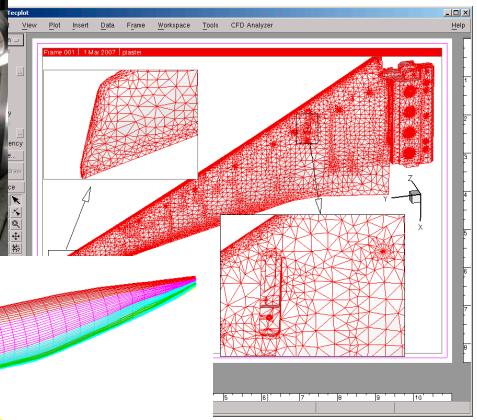


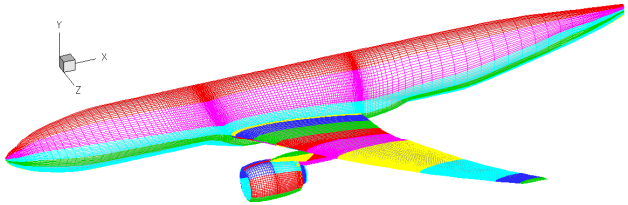
Instantaneous concentration of methyl radical. Low concentrations of methyl upstream of the high-temperature flame base serve as a marker of ignition induction chemistry

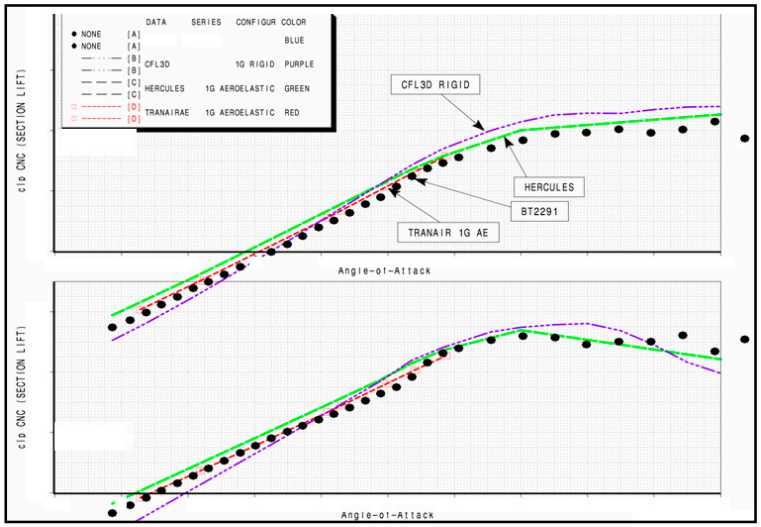
Engineering

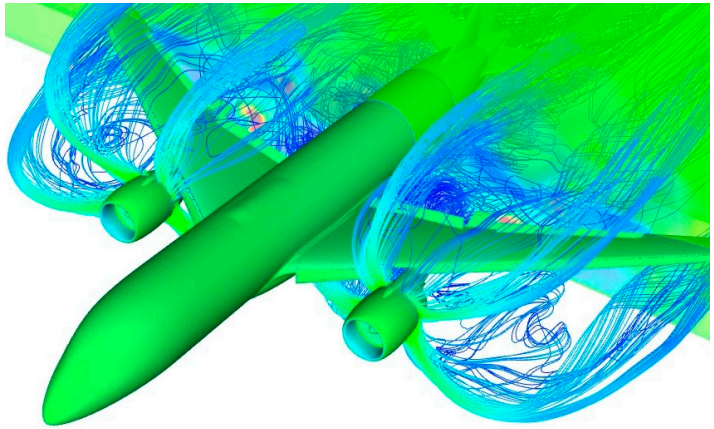
Development and correlations of large-scale computational tools for flight vehicles











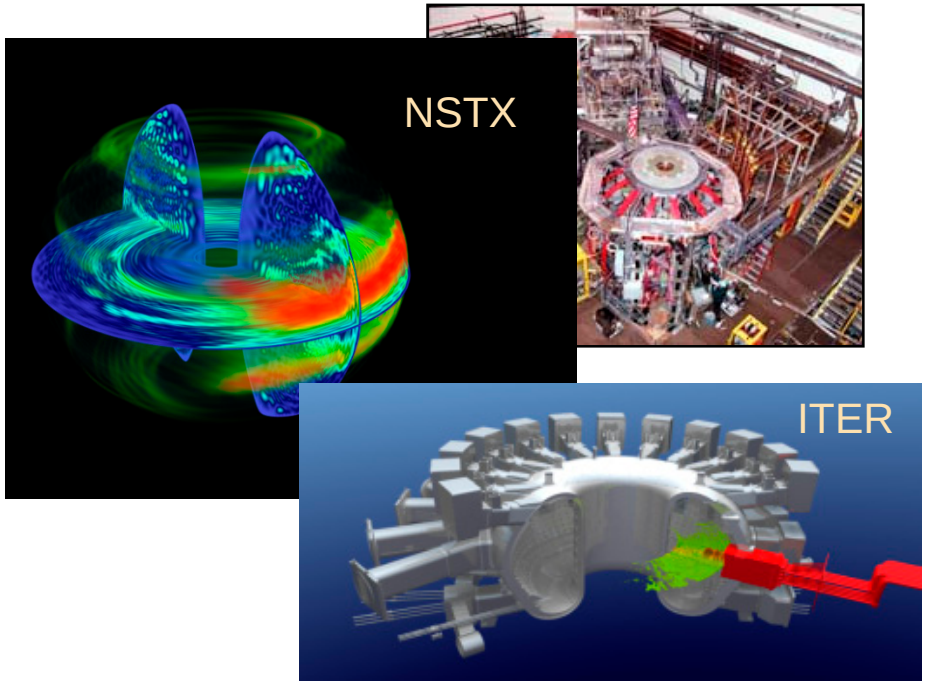
- Rare validation of unsteady Navier-Stokes tools for “simple” unsteady pitch oscillation motion
- Demonstration of the applicability and predictive accuracy of a CFD-based aeroelastic analysis process
- Confidence level improvement for rigid CFD results (toward its use in product downstream processes)
- Testing and development of new methods and methodologies (blade-out analysis process)
- Investigations of unsteady methods for better predictions (e.g. unsteady detached eddy simulations/DES)
- Evaluation of wind-tunnel effects

Fusion

Producing new insights for RF heating of ITER plasmas

Three-dimensional simulations of RF heating in the ITER fusion reactor as well as in present tokamaks (NSTX) shed new light on the behavior of superheated ionic gas

- When deuterium (D) or helium-3 (He^3) are used to damp the launched waves, they are accelerated to high energies, forming supra-thermal tails that significantly affect the wave propagation and absorption
- Energetic minority D ions enhance the fusion reaction rate
- Energetic He^3 ion tails form on both the tritium and He^3 distributions



- **3D simulations reveal new insights**
  - **"Hot spots" near antenna surface**
  - **"Parasitic" draining of heat to the plasma surface in smaller reactors**
- **Work pushing the boundaries of the system (28,900 cores, 154 TF) and demonstrating**
  - **Radial wave propagation and absorption**
  - **Efficient plasma heating**
- **AORSA's predictive capability can be coupled with Jaguar power to enhance fusion-reactor design and operation for an unlimited clean energy, source**

LCF liaison contributions

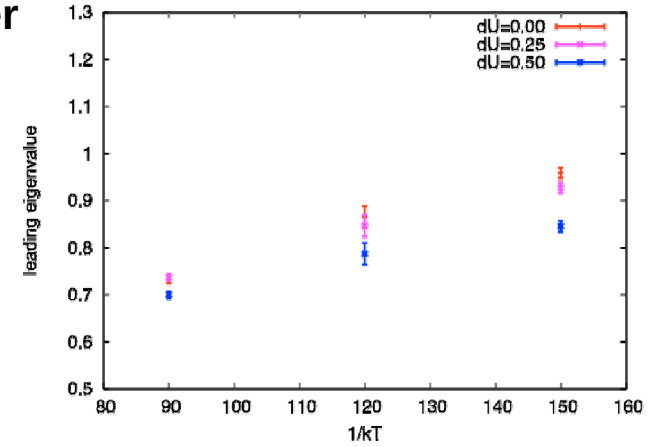
- **Converted HPL from double real to double complex and replaced ScaLAPACK**
- **Acquired new version of BLAS from TACC**
- **Net performance gain of a factor of two**

Materials and Nanoscience

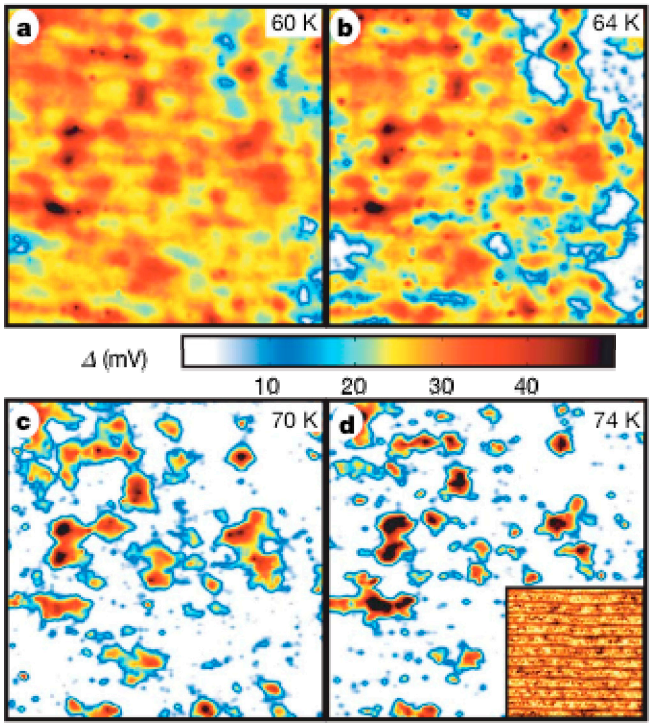
Role of nanoscale inhomogeneities in high-temperature superconductors

PI: Thomas Schulthess, ORNL

- Investigate role of inhomogeneities on pairing mechanisms and transition temperature—petascale computing problem
- Develop materials-specific extensions to the Hubbard model to understand variations of transition temperatures—develop materials design tool
- Simulations of 2D Hubbard model with disorder in local Coulomb interaction parameter (Ui = 1 +/- dU) indicate *Tc* is indeed suppressed due to disorder



- Further analysis of the role of inhomogeneities (substitutions, vacancies, etc.) is now possible in the simulations with 16- and 24-site clusters



Temperature evolution of the superconducting gap taken on a 300 A area of a cuprate with T_c = 65K [reproduced from Gomez et al., *Nature* **447**, 569−572 (2007)]. The gap varies spatially on a scale of 1−3 nm and persists in some regions to temperatures well above T_c, as can be seen from panels c and d

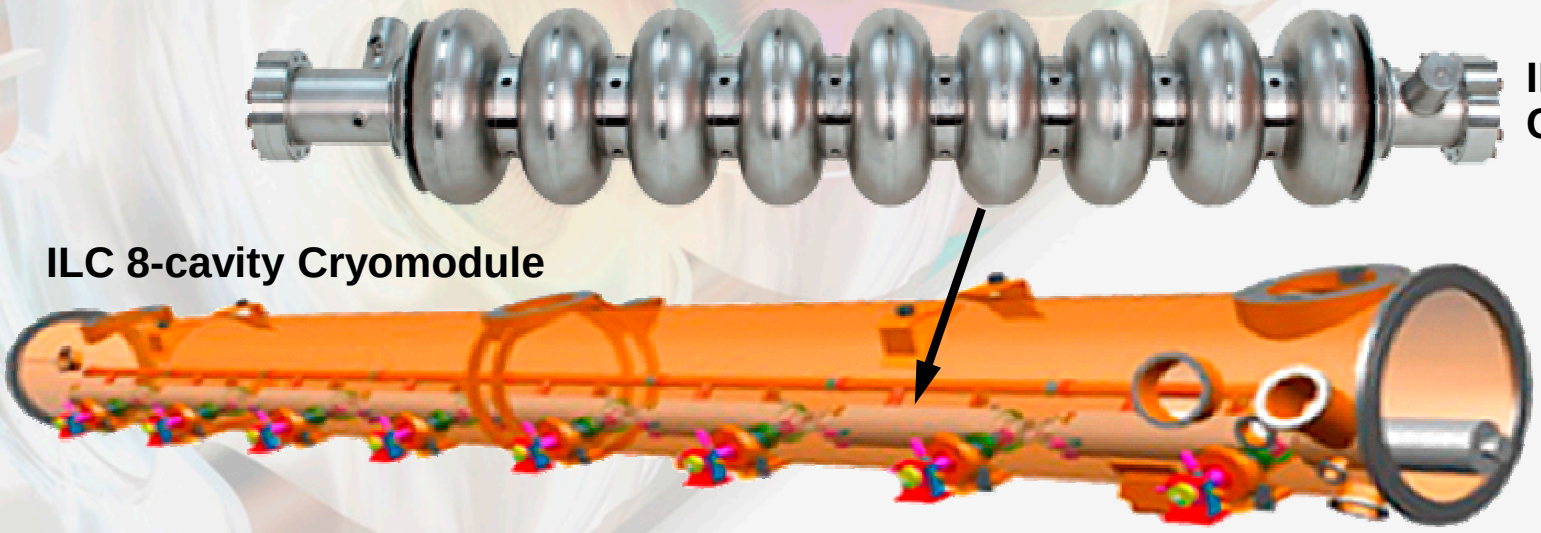
Simulations at 31K cores on Jaguar perform >50% of peak (Gordon Bell submission)

Petascale computing for terascale particle accelerator

International linear collider design and modeling

PI: Lie-Quan Lee, Stanford Linear Accelerator Center

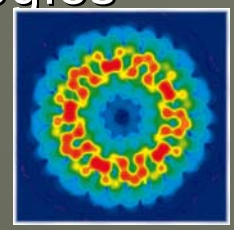
- The ILC is a particle accelerator aimed to address many of the most compelling physics questions about dark matter; dark energy; and the fundamental nature of matter, energy, space, and time
- Using the most advanced simulation tools (Omega3P, T3P), the computationally challenging problems associated with the ILC accelerator design will be tackled for improved machine performance, increased reliability, and reduced cost

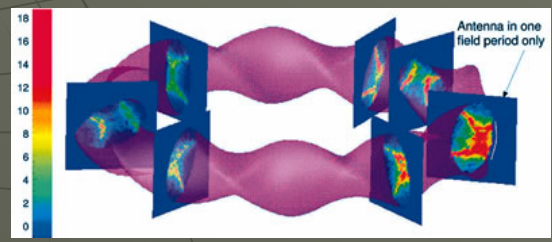


Computer Science Research

- Computer Science and Mathematics (CSM) Division.
  - Applied research focused on computational sciences, intelligent systems, and information technologies
- CSM Research Groups:
  - Climate Dynamics
  - Complex Systems
  - Computational Chemical Sciences
  - Computational Materials Science
  - Future Technologies
  - Statistics and Data Science
  - Computational Mathematics
  - *Computer Science Research*
    (~20 researchers, 3 postdocs/postmasters, students, …)





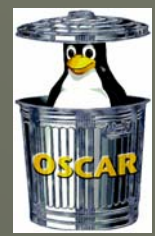


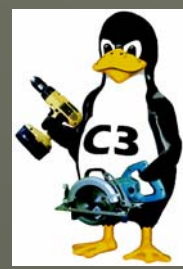
Computer Science Research Projects

- Parallel Virtual Machine (PVM)
- MPI Specification, FT-MPI and Open MPI
- Common Component Architecture (CCA)
- Open Source Cluster Application Resources (OSCAR)
- Scalable cluster tools (C3)
- Scalable Systems Software (SSS)
- Fault-tolerant metacomputing (HARNESS)
- High availability for HPC (FAST-OS MOLAR)
- System-level virtualization (VSE)
- Harness HPC workbench
- Resilience for HPC (FAST-OS RAS)



FT-MPI

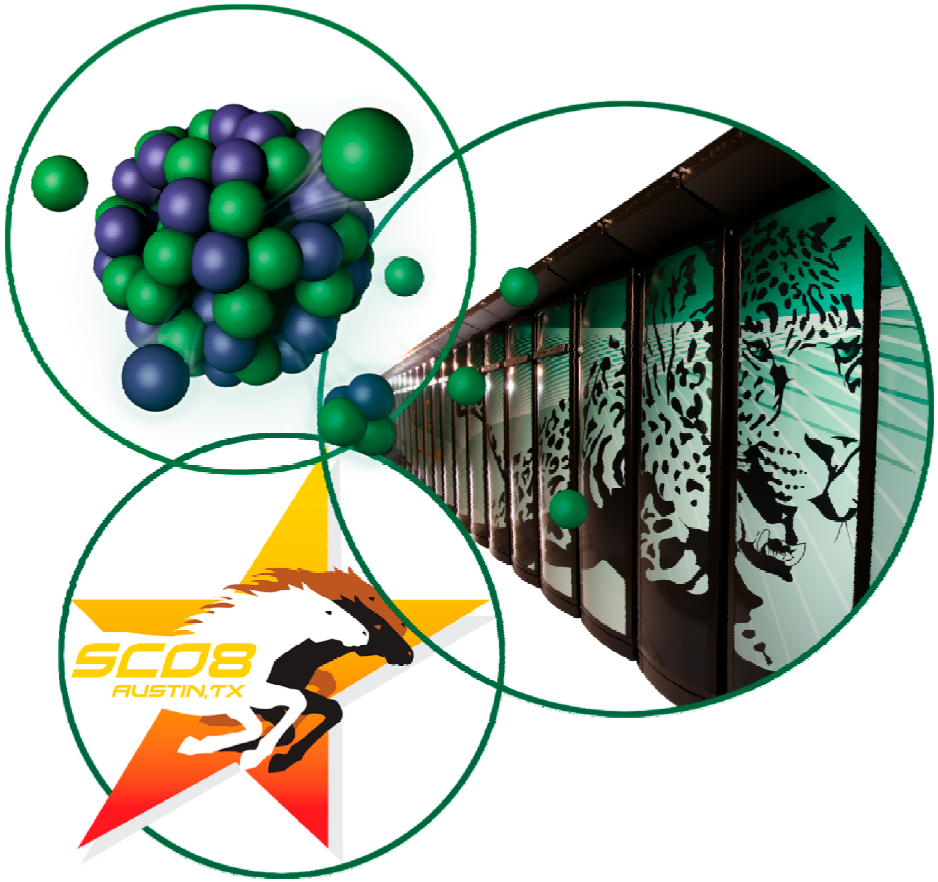




System Research Team

- SRT team
  - Stephen L. Scott
  - Christian Engelmann
  - Hong Ong
  - Geoffroy Vallée
  - Thomas Naughton
  - Anand Tikotekar
- R&D topics
  - High availability
  - Fault tolerance
  - Operating systems
  - Cluster computing
  - Virtualization technologies
  - Resource management
  - Tools

The Harness Workbench: Unified and Adaptive Access to Diverse HPC Platforms



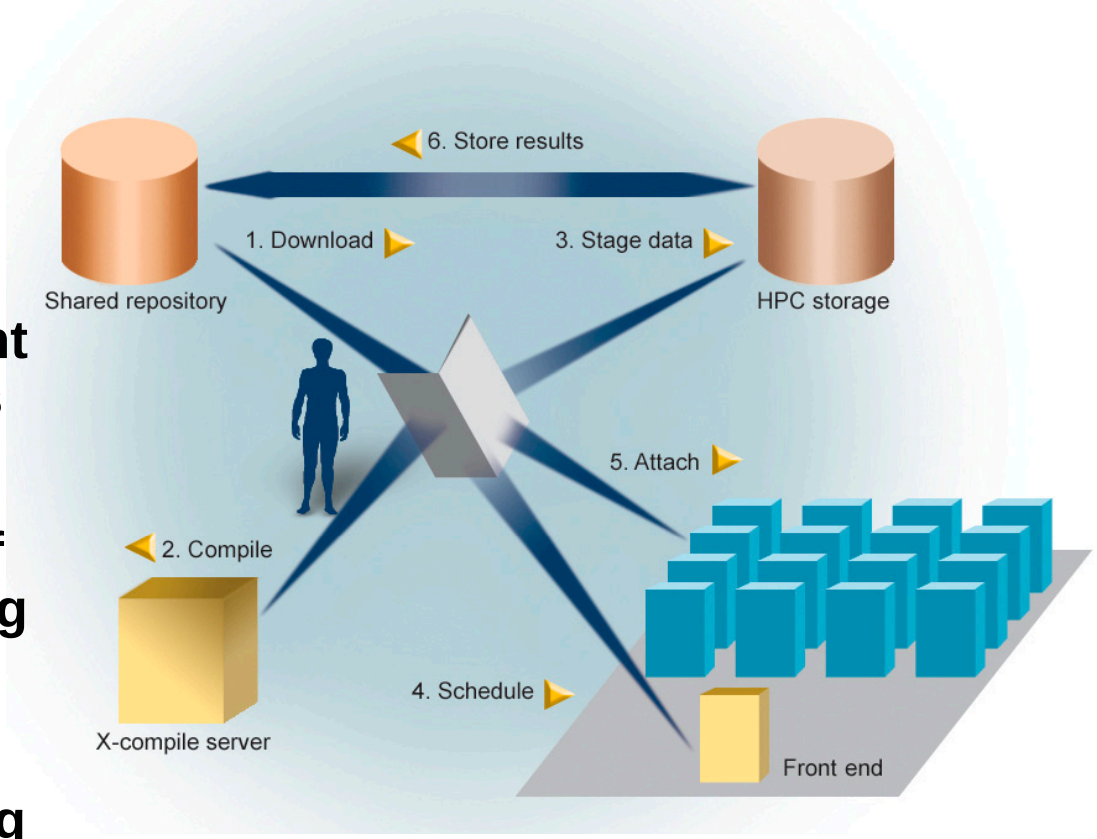
Presented by

Christian Engelmann

Computer Science Research Group
Computer Science and Mathematics Division
Oak Ridge National Laboratory

Research and development goals

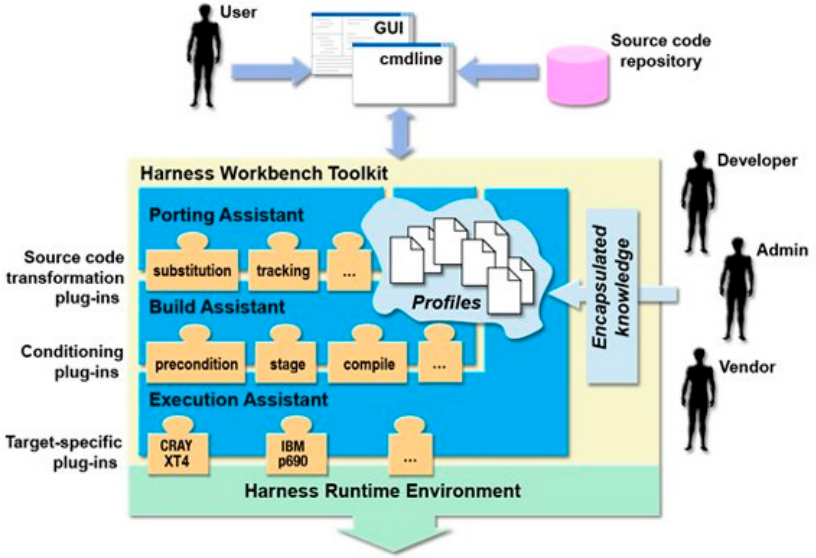
- Increasing the overall productivity of developing and executing computational codes
- Optimizing the development and deployment processes of scientific applications
- Simplifying the activities of application scientists, using uniform and adaptive solutions
- "Automagically" supporting the diversity of existing and emerging high–performance computing architectures

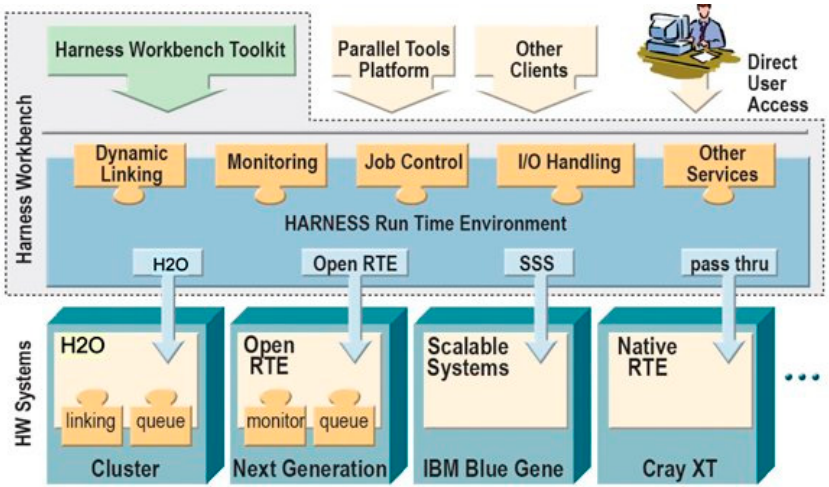


Typical scientific application development, deployment, and execution activities

Harness Workbench core components

- **Harness Workbench Toolkit**
  - **Unified development, deployment, and execution**
  - **Common view across diverse HPC platforms**
  - **User-space installation and virtual environments**

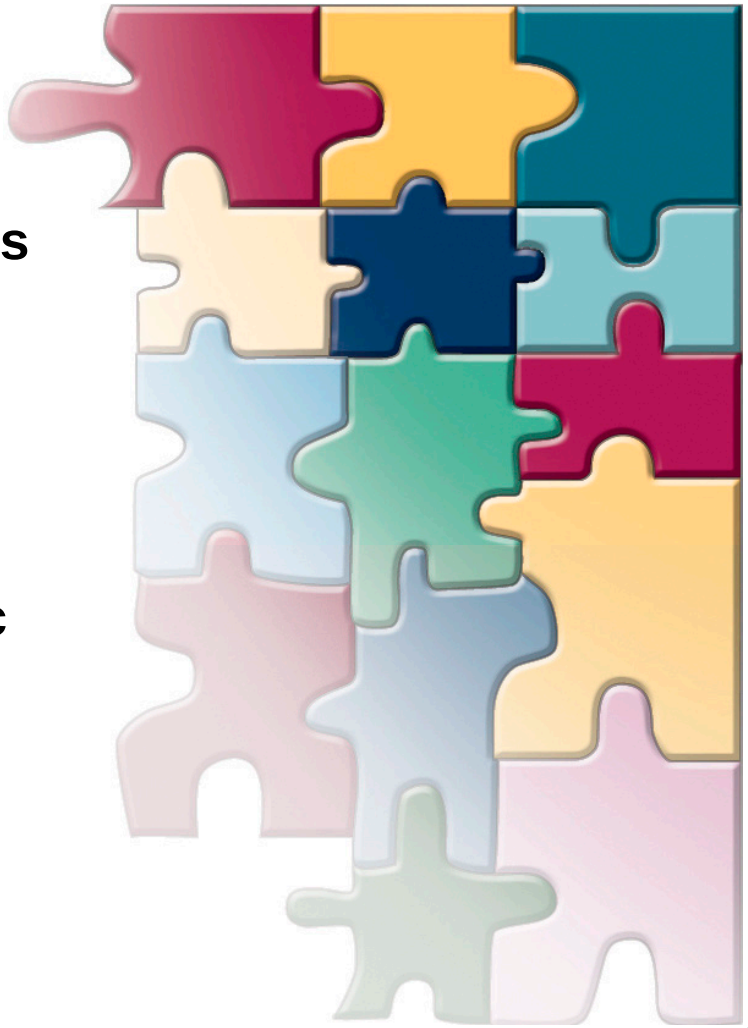




- **Next-generation runtime environment**
  - **Flexible, adaptive, lightweight framework**
  - **Management of runtime tasks**
  - **Support for diverse HPC platforms**

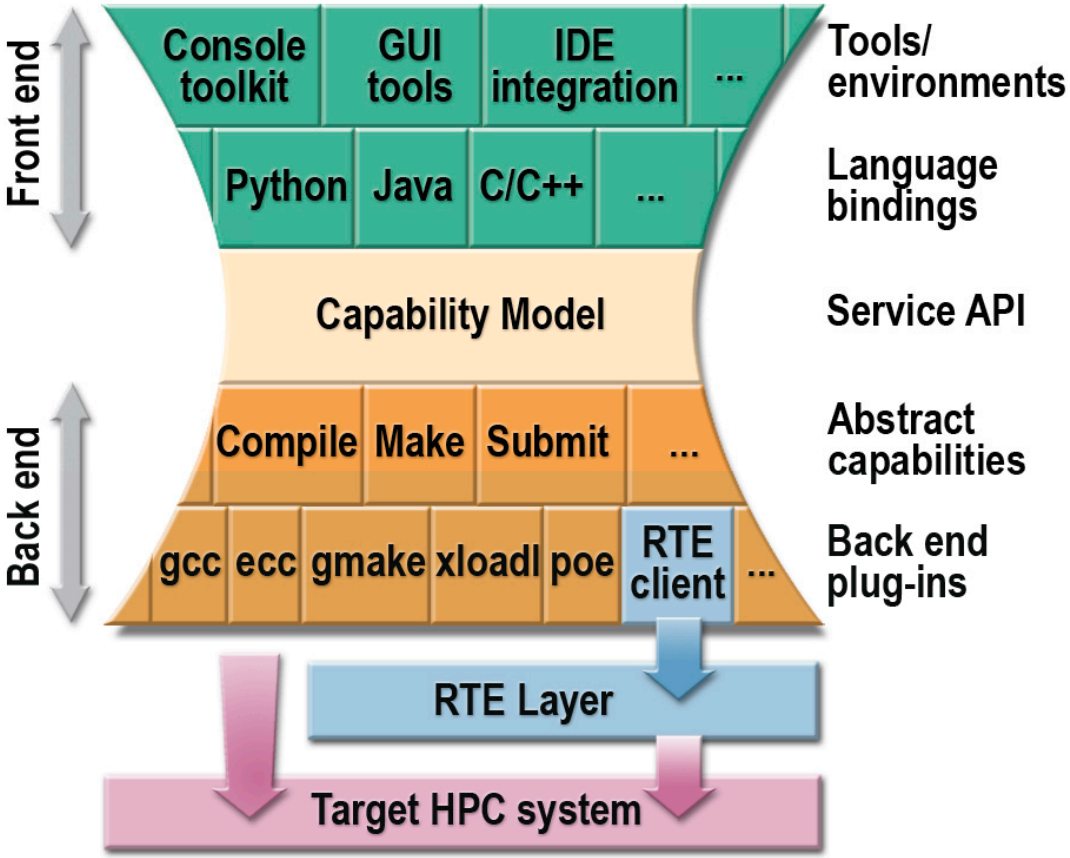
Harness Workbench core technologies

- **Automatic adaptation using pluggable modules**
  - **Harness Workbench Toolkit plug-ins**
  - **Runtime environment plug-ins**
- **Development environment and toolkit interfaces**
  - **Easy-to-use interfaces for scientific application development, deployment, and execution**



Common view across diverse platforms

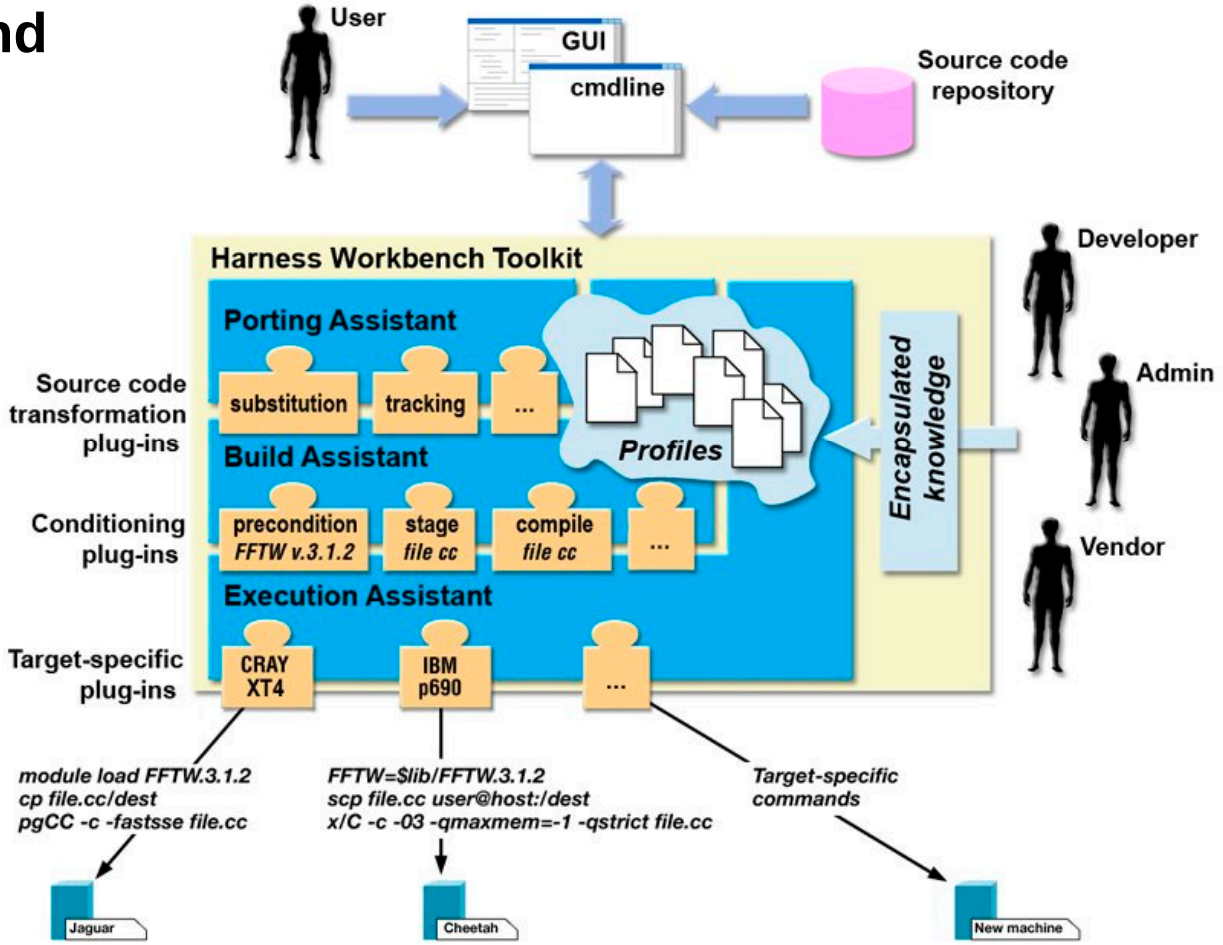
- Various interfaces and bindings to external development and deployment tools and environments
- Generalized model for unified access to common development and deployment activities
- Mapping of generalized activities onto platform-specific toolkits and runtime environments (RTEs) via pluggable modules



EMORY UNIVERSITY
OAK RIDGE National Laboratory

Harness Workbench Toolkit

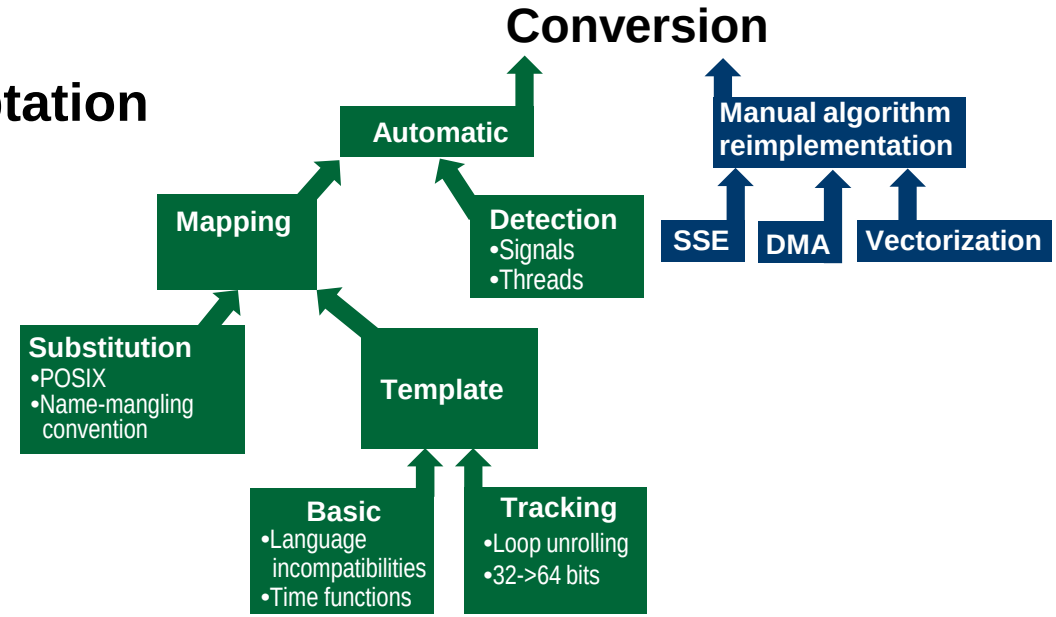
- Unifying abstraction over heterogeneous HPC resources
- Command line and GUI tools
- Translation into fine-tuned invocations of native toolkits
- Behavior encapsulated in plug-ins
- Configurable through profiles
- Tunable by end users

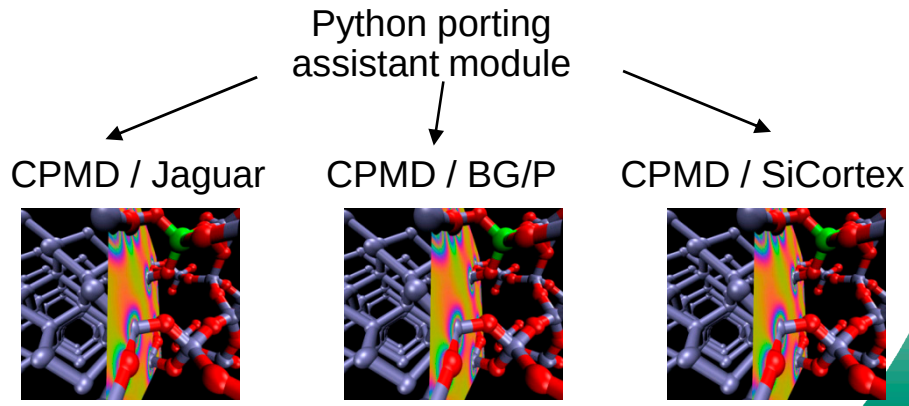


EMORY UNIVERSITY
OAK RIDGE National Laboratory

Harness Workbench Toolkit porting assistant

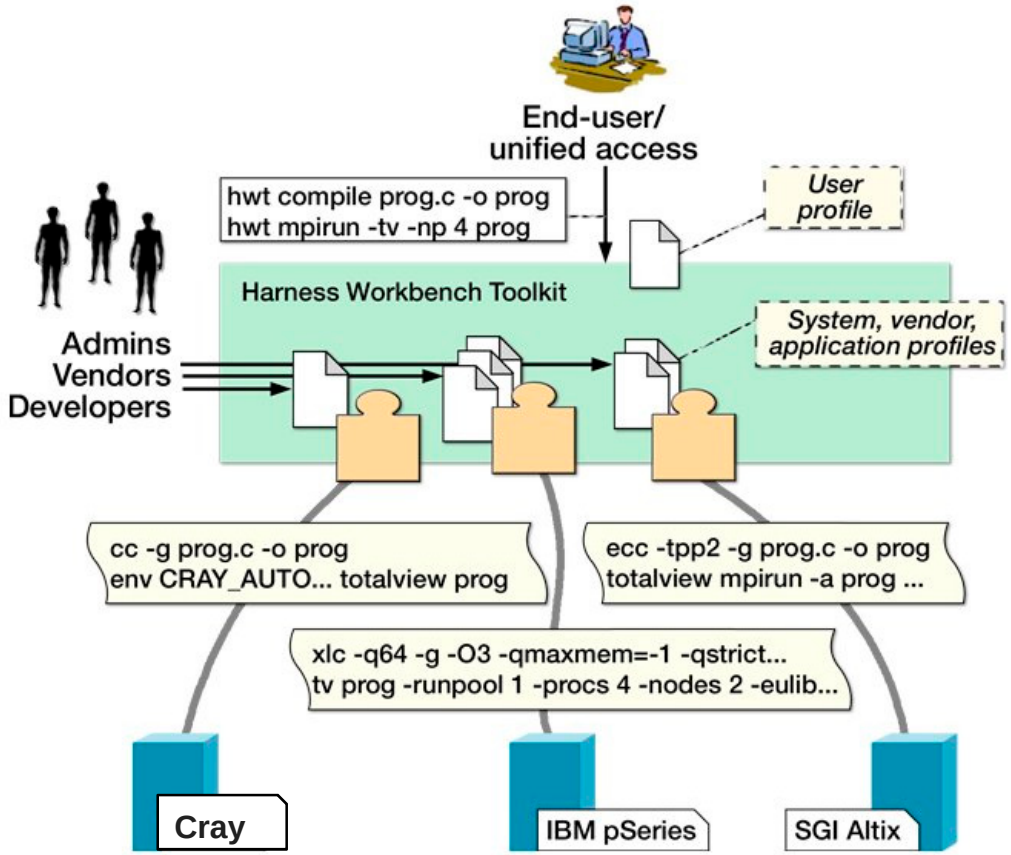
- Facilitates source code adaptation through specialized plug-ins
  - Suggests safe conversions
  - Highlights manual code replacement areas
  - Guided by situation-specific profiles
- Prototype
  - Plug-ins as Python scripts
  - Porting CPMD across
    - Jaguar Cray XT4
    - IBM Blue Gene/P
    - SiCortex
- Example conversions
  - Detection, function mappings, data type size changes





Harness Workbench Toolkit build and execution layer

- Abstraction layer between users and platform-specific compilers, linkers, libraries, testing and debugging software, launching systems, etc.
- End-users issue generic build commands that are processed to produce a target-specific set of commands
- Pluggable modules to deliver back-end functionality



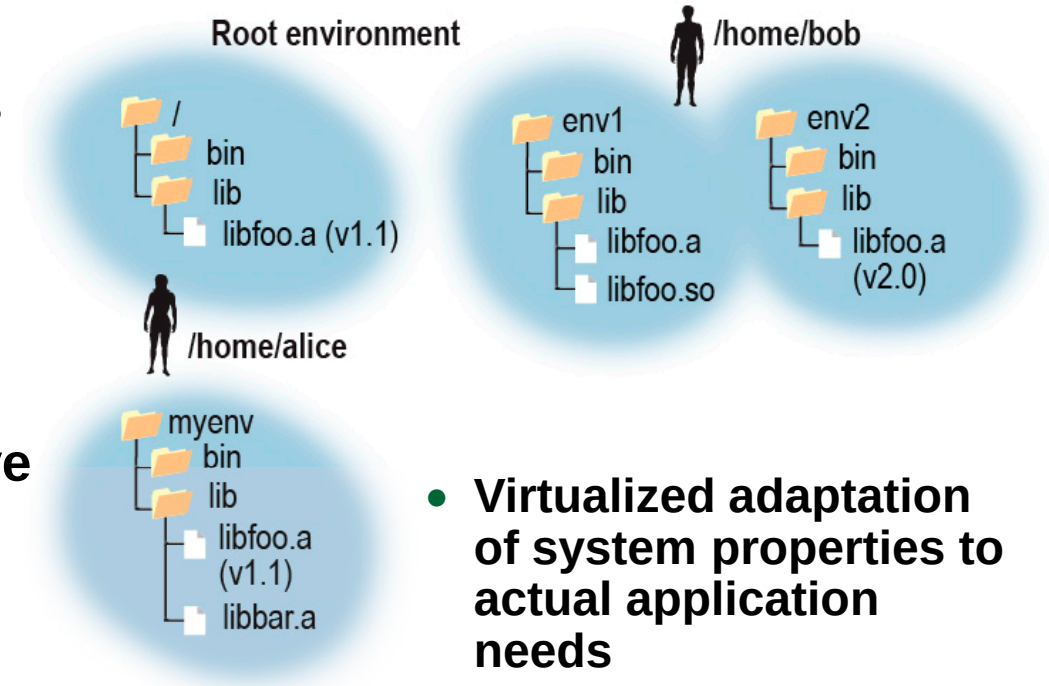
Harness runtime environment (HRTE)



- **Adaptability and flexibility via plug-ins**
  - **Lightweight pluggable component framework**
  - **Components for various HPC platforms and runtime environments**
- **Scalability and resilience with fully distributed approach**
  - **Binomial graph network topology**
- **Runtime fault tolerance through self-stabilization**
  - **Oracle for resource discovery**
  - **Failure detector to deal with crashed processes**
- **Runtime support for virtualized environments**

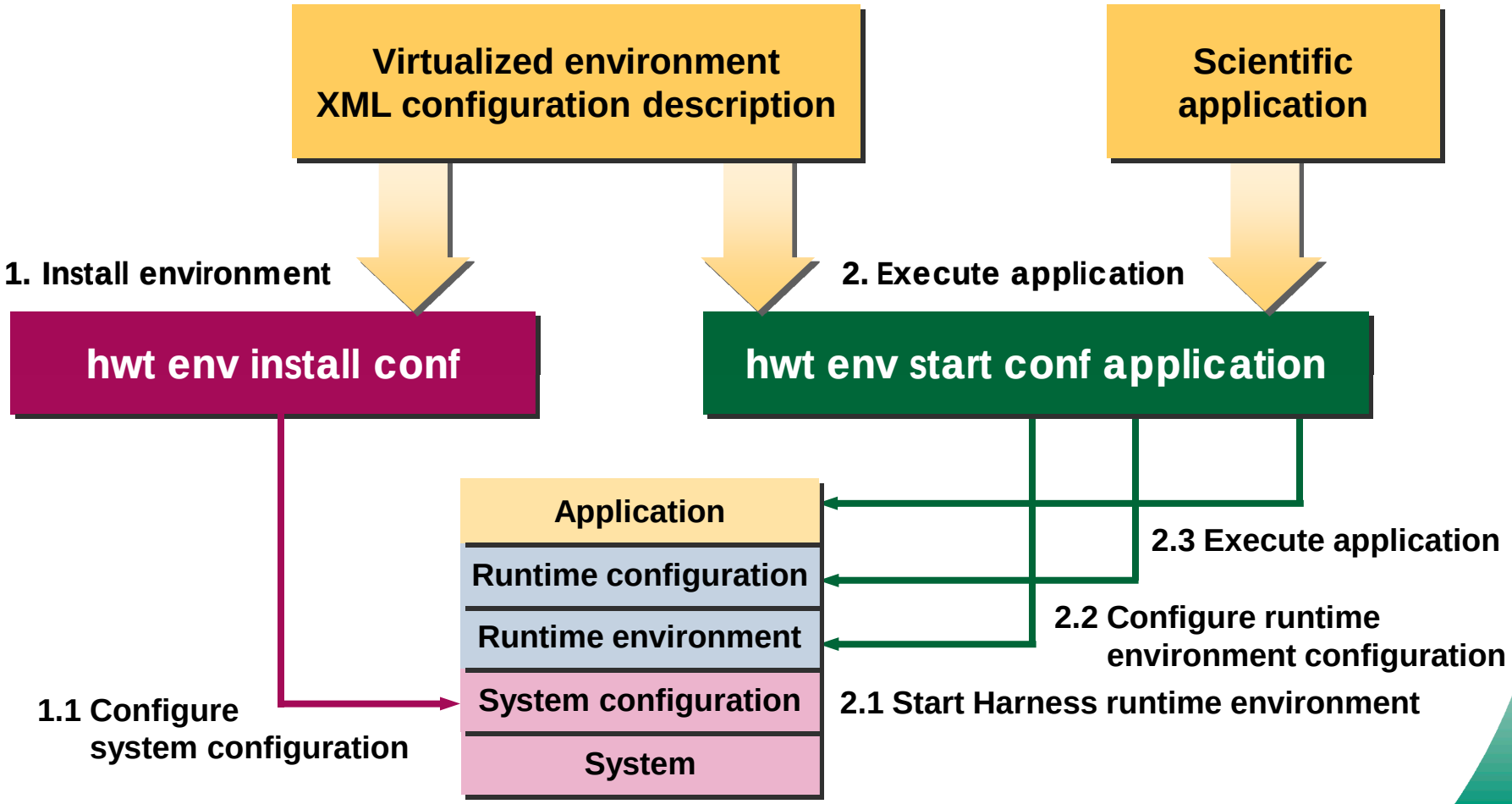
Harness Workbench Toolkit /HRTE virtualized environments

- **Problem**
  - Application dependencies may cause conflicts with system-wide installed libraries
- **Solution**
  - Use co-existing, alternative user-space installations
- **Approach**
  - Provide isolated installation environments (“sandboxes”)
  - These can inherit from one another to build nested hierarchies



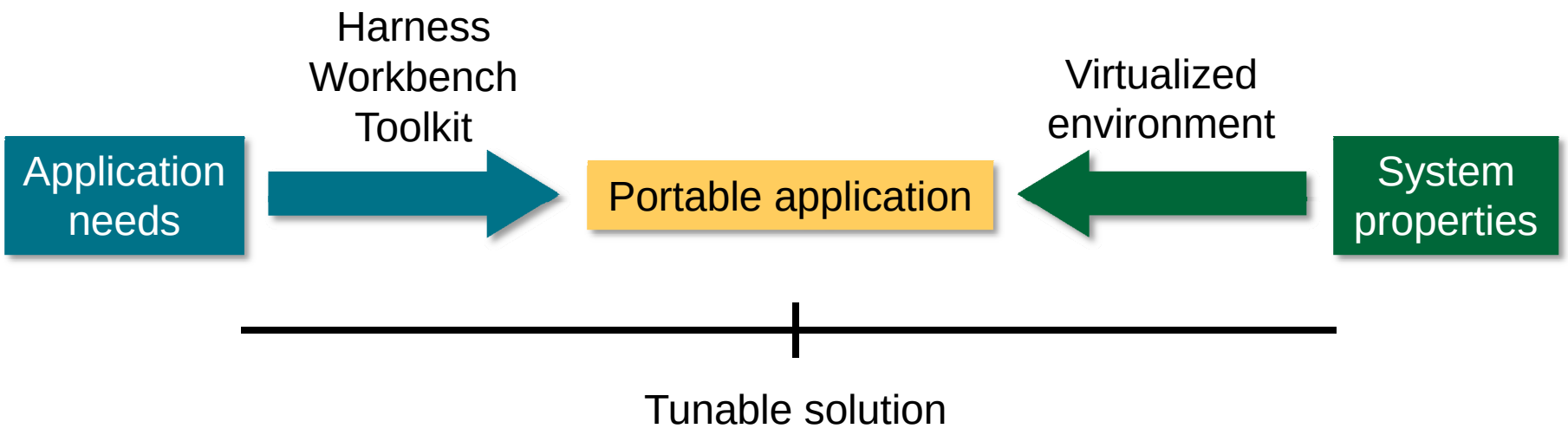
- Virtualized adaptation of system properties to actual application needs
- System and runtime environment virtualization

Configurable “sandboxes” for scientific applications



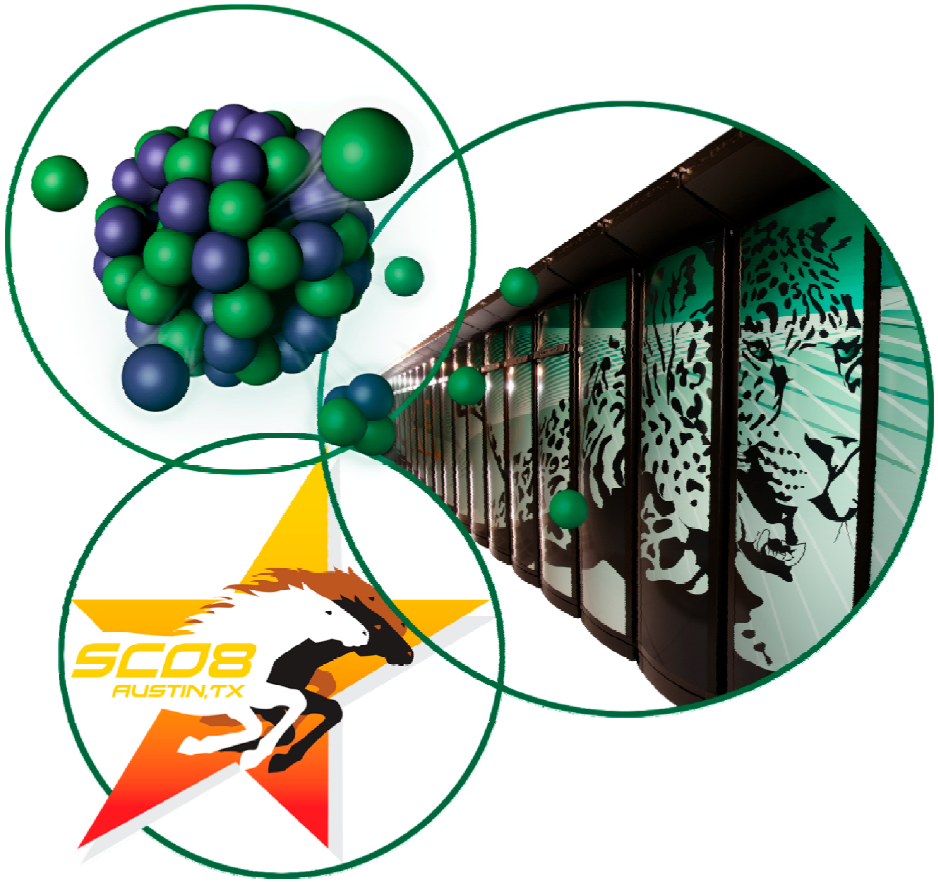
EMORY UNIVERSITY
OAK RIDGE National Laboratory

Tunable portability for day-one operation



- **Harness Workbench Toolkit for adaptation of application to HPC system properties**
- **Virtualized environment for adaptation of HPC system to application needs**
- **Tunable solution, where applications can move from one HPC system to another with initial minimal changes enabling day-one operation**
- **Continued adaptation to actual target platform properties ensures eventual performance gains by gradually taking advantage of new system features**

Reliability, Availability, and Serviceability (RAS) for High-Performance Computing (HPC)



Presented by

Stephen L. Scott
Christian Engelmann

Computer Science Research Group
Computer Science and Mathematics Division

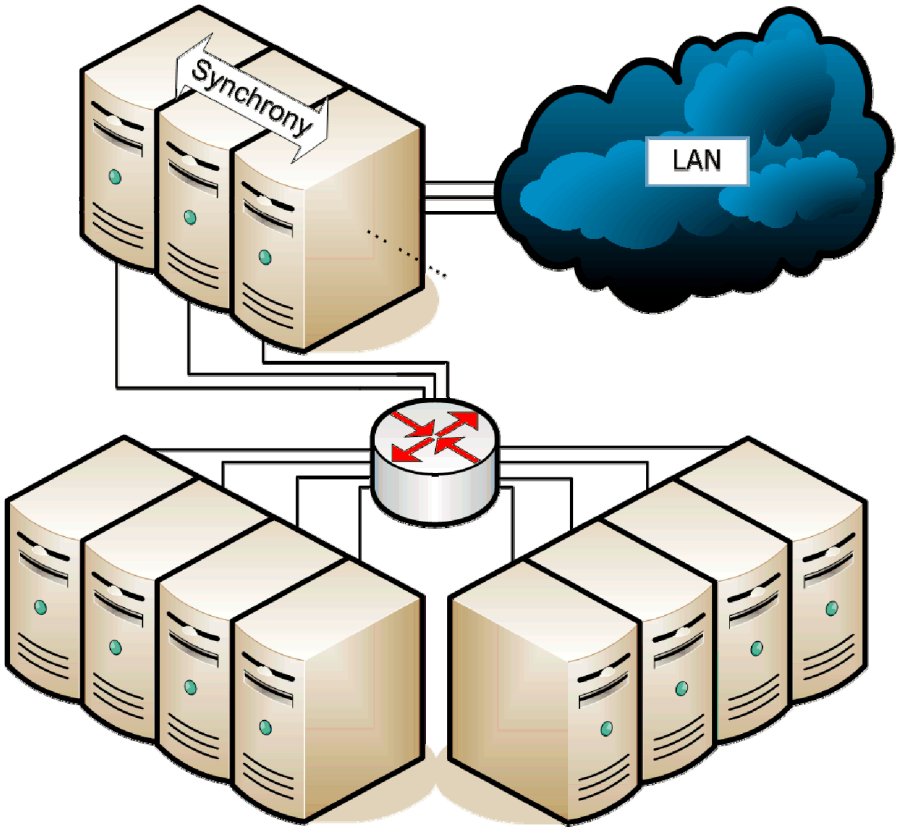
Research and development activities

- **Efficient redundancy strategies for HPC head/service nodes for high availability and high performance of critical services**
- **Reactive fault tolerance for HPC compute nodes utilizing the job pause approach and checkpoint placement adaptation**
- **Proactive fault tolerance using preemptive migration of computation away from compute nodes that are about to fail**
- **Reliability analysis for identifying pre-fault indicators, predicting failures, and modeling and monitoring reliability**
- **Holistic fault tolerance through combination of adaptive proactive and reactive fault tolerance mechanisms**

Office of Science U.S. DEPARTMENT OF ENERGY | OAK RIDGE National Laboratory | NC STATE UNIVERSITY | LOUISIANA TECH UNIVERSITY

Symmetric active/active redundancy

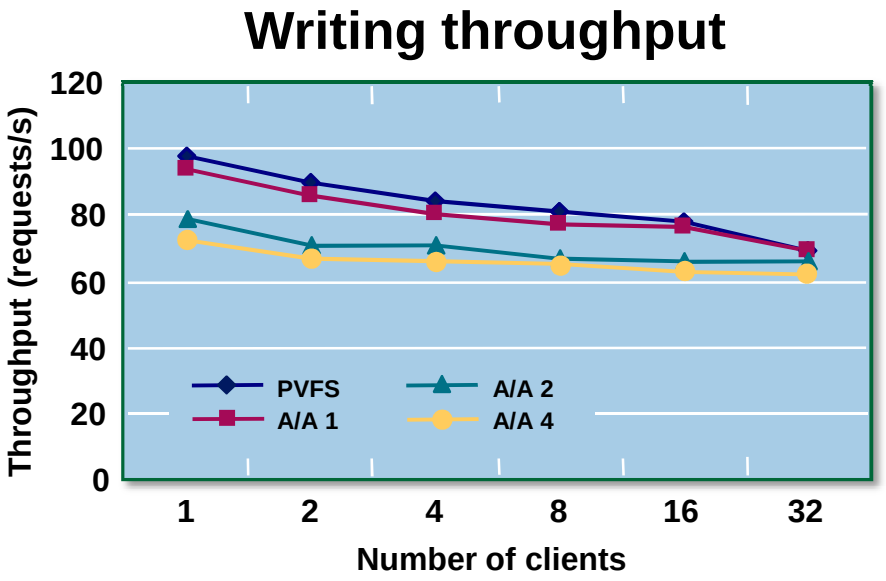
Active/active head nodes

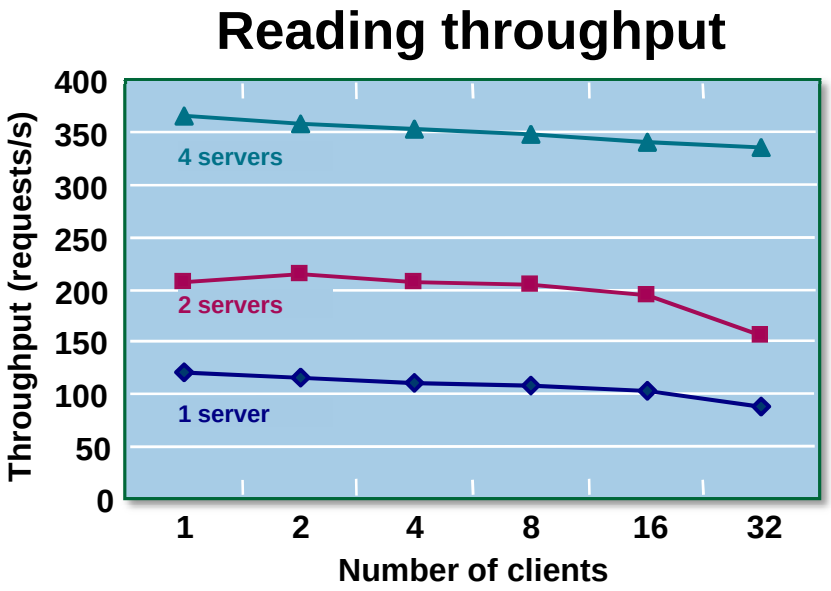


- Many active head nodes
- Workload distribution
- Symmetric replication between head nodes
- Continuous service
- Always up to date
- No fail-over necessary
- No restore-over necessary
- Virtual synchrony model
- Complex algorithms

➡ **Prototypes for Torque and Parallel Virtual File System metadata server**

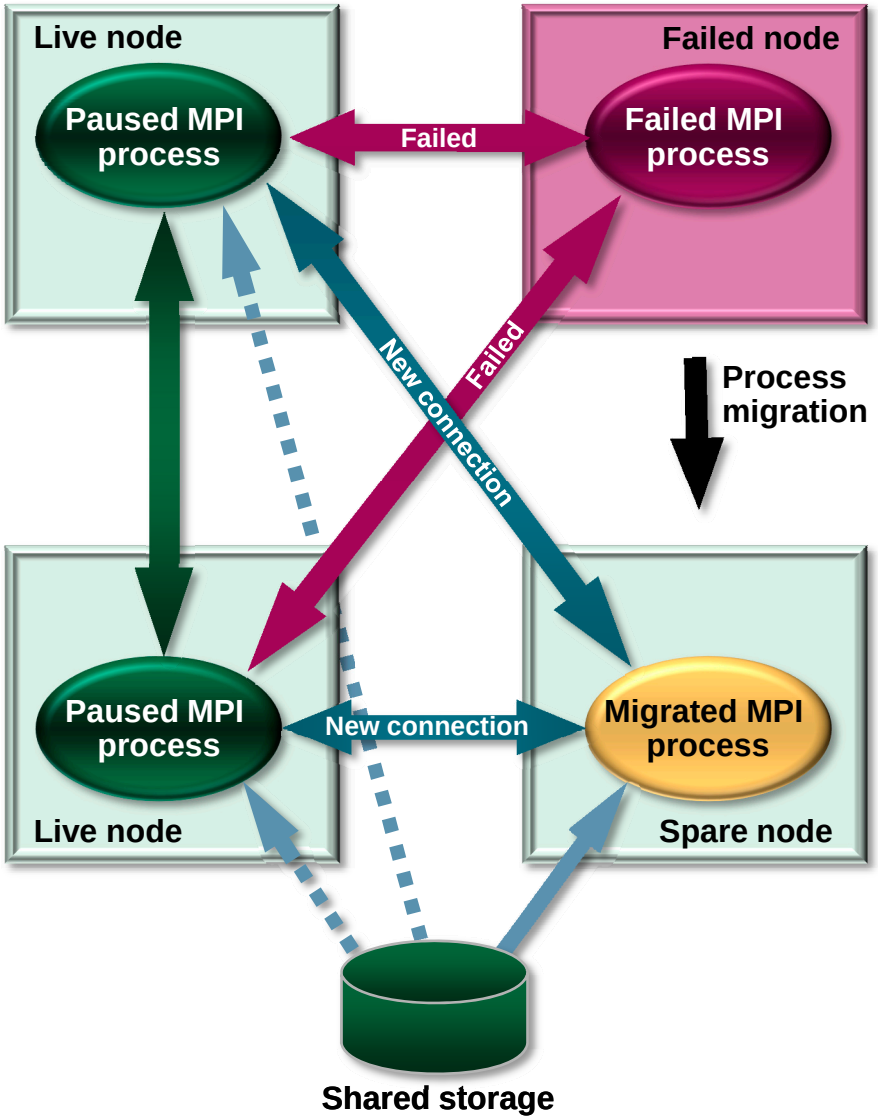
Symmetric active/active Parallel Virtual File System metadata server





Nodes	Availability	Est. annual downtime
1	98.58%	5d, 4h, 21m
2	99.97%	1h, 45m
3	99.9997%	1m, 30s

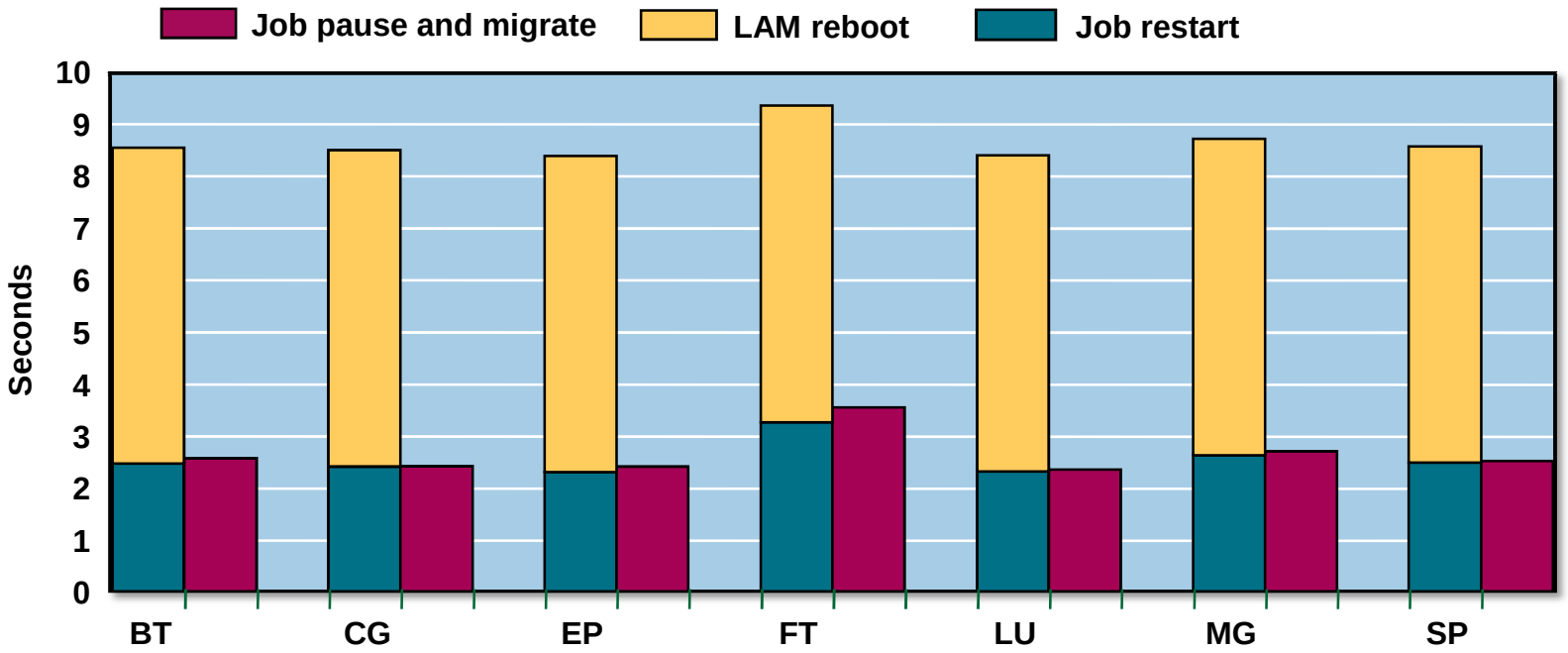
Reactive fault tolerance for HPC with LAM/MPI+BLCR job-pause mechanism



- Operational nodes: Pause
  - BLCR reuses existing processes
  - LAM/MPI reuses existing connections
  - Restore partial process state from checkpoint
- Failed nodes: Migrate
  - Restart process on new node from checkpoint
  - Reconnect with paused processes
- Scalable MPI membership management for low overhead

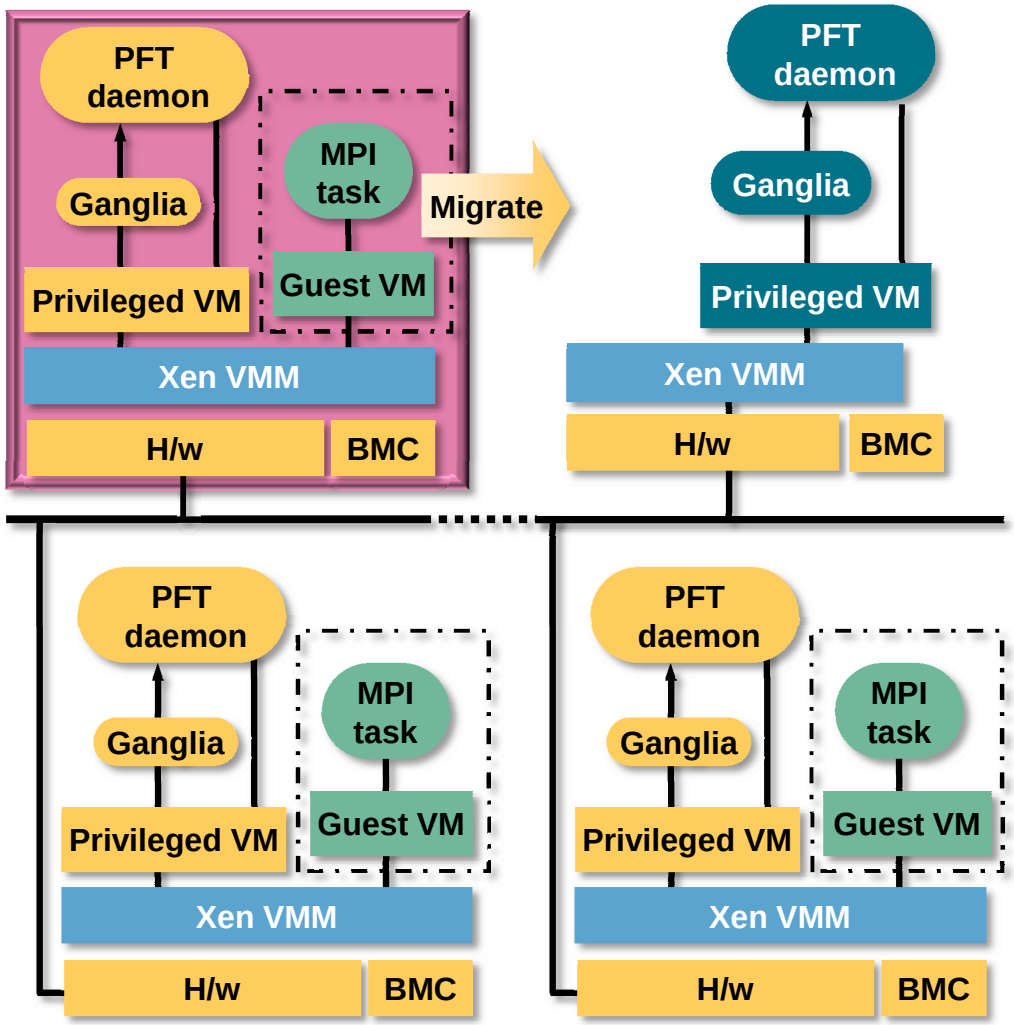
➡ **Efficient, transparent, and automatic failure recovery**

LAM/MPI+BLCR job pause performance



- 3.4% overhead over job restart, but
  - No LAM reboot overhead
  - Transparent continuation of execution
- No requeue penalty
- Less staging overhead

Proactive fault tolerance for HPC using Xen virtualization

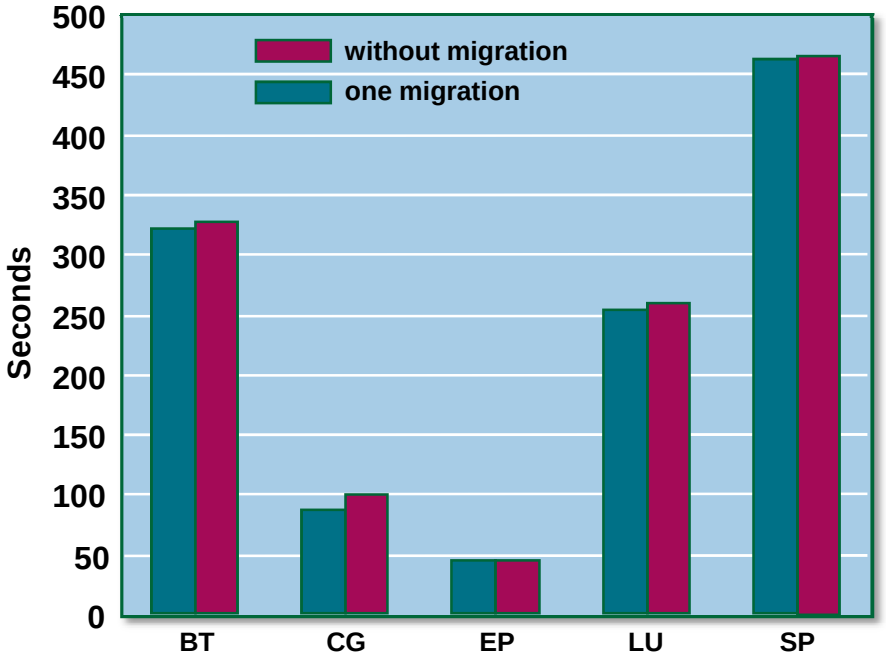


- Standby Xen host (spare node without guest VM)
- Deteriorating health
  - Migrate guest VM to spare node
- New host generates unsolicited ARP reply
  - Indicates that guest VM has moved
  - ARP tells peers to resend to new host

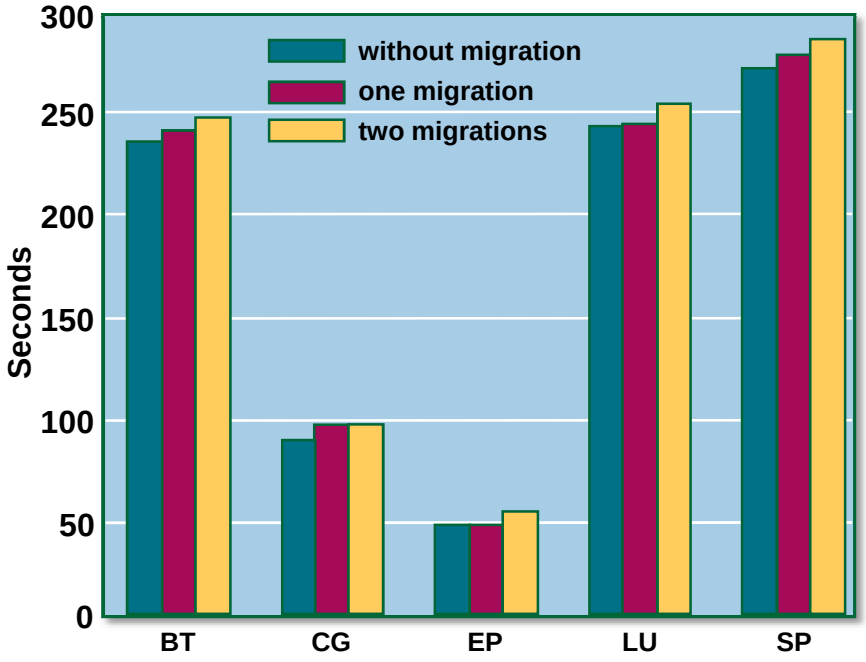
➡ **Novel fault-tolerance scheme that acts before a failure impacts a system**

VM migration performance impact

Single node failure



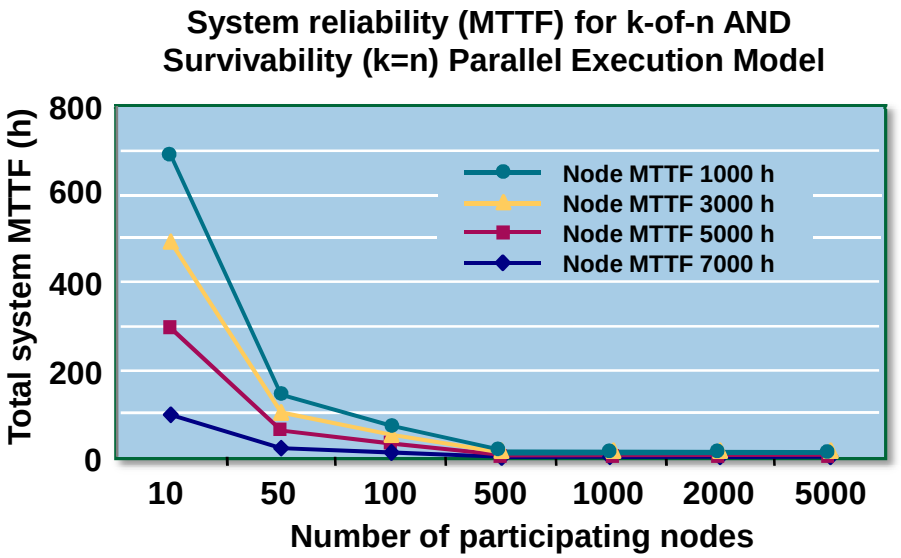
Double node failure

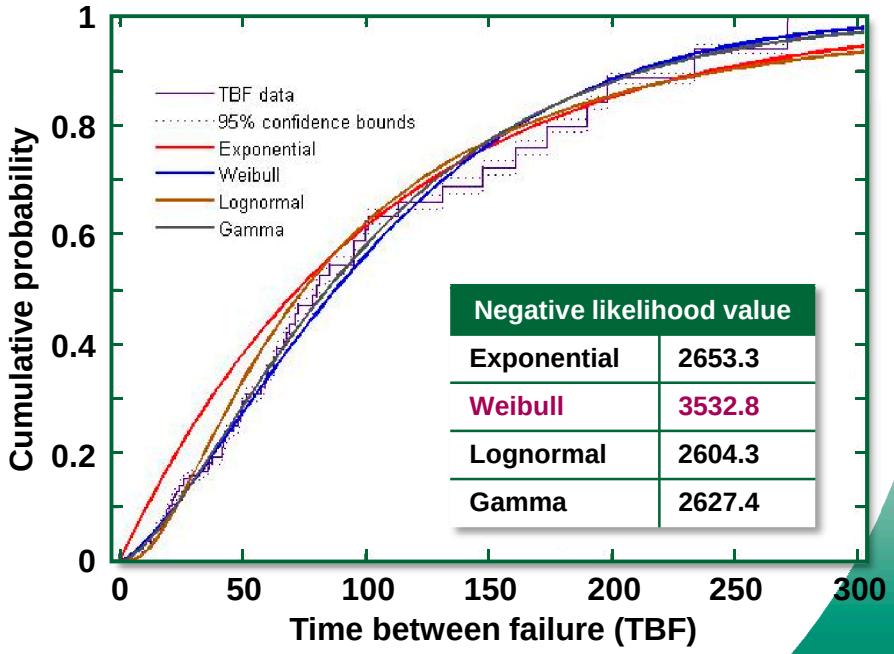


- **Single node failure: 0.5–5% additional cost over total wall clock time**
- **Double node failure: 2–8% additional cost over total wall clock time**

HPC reliability analysis and modeling

- Programming paradigm and system scale impact reliability
- Reliability analysis
- Estimate mean time to failure (MTTF)
- Obtain failure distribution: Exponential, Weibull, gamma, etc.
- Feedback into fault-tolerance schemes for adaptation

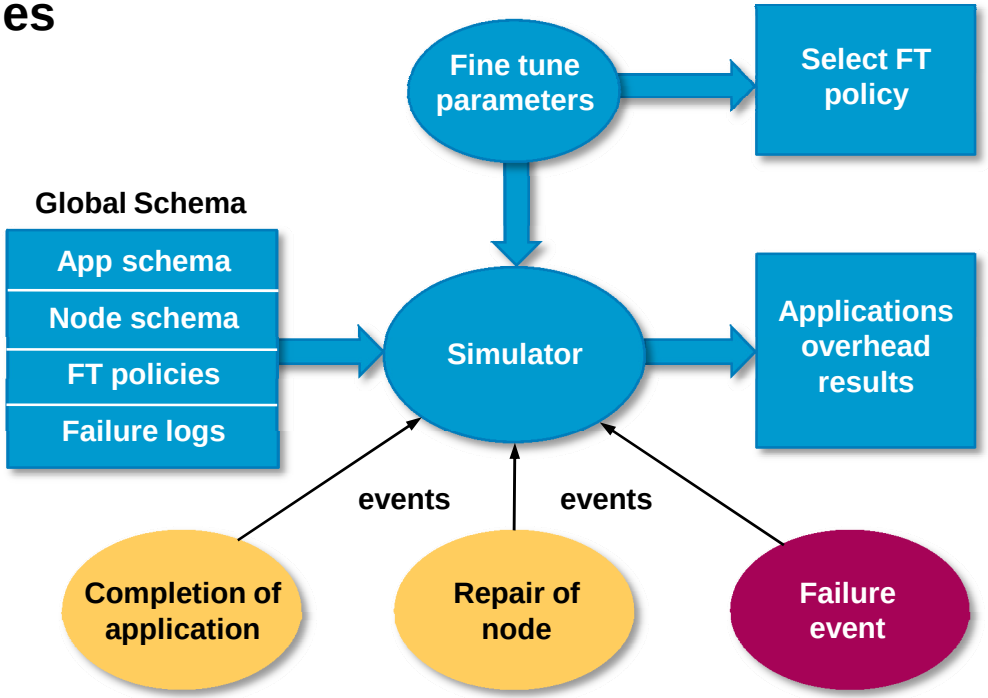




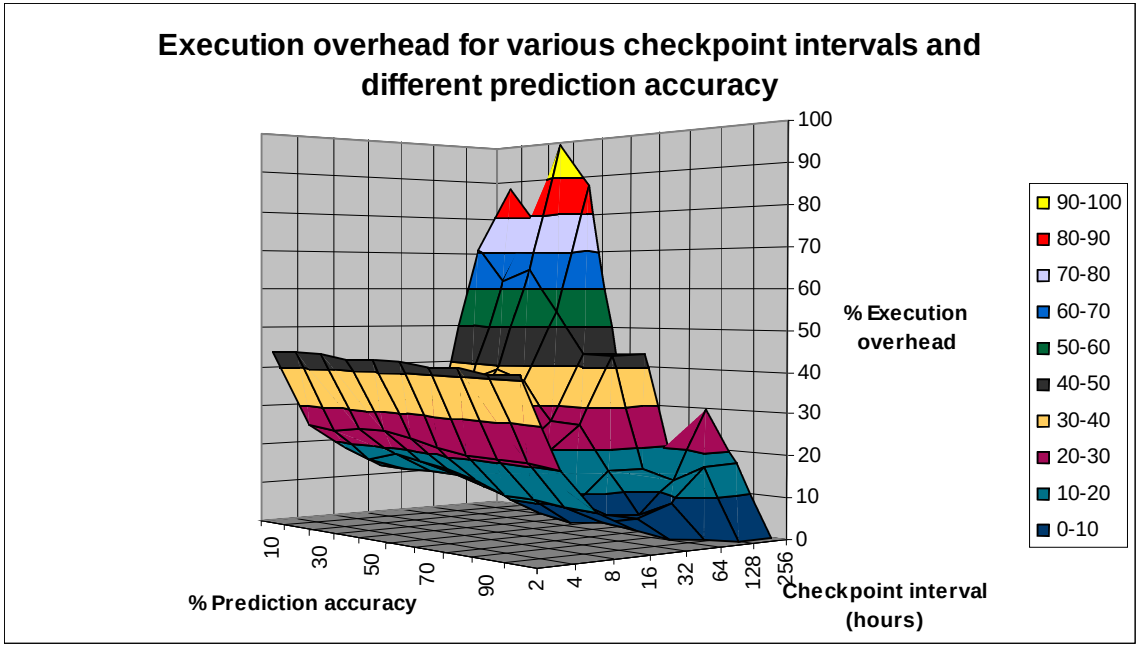
Negative likelihood value	
Exponential	2653.3
Weibull	3532.8
Lognormal	2604.3
Gamma	2627.4

Simulation framework for HPC fault-tolerance policies

- Evaluation of fault tolerance policies
  - Reactive only
  - Proactive only
  - Reactive/proactive combination
- Evaluation of fault tolerance parameters
  - Checkpoint interval
  - Prediction accuracy
- Event-based simulation framework using actual HPC system logs
- Customizable simulated environment
  - Number of active and spare nodes
  - Checkpoint and migration overheads



Combination of proactive and reactive fault tolerance: Simulation example

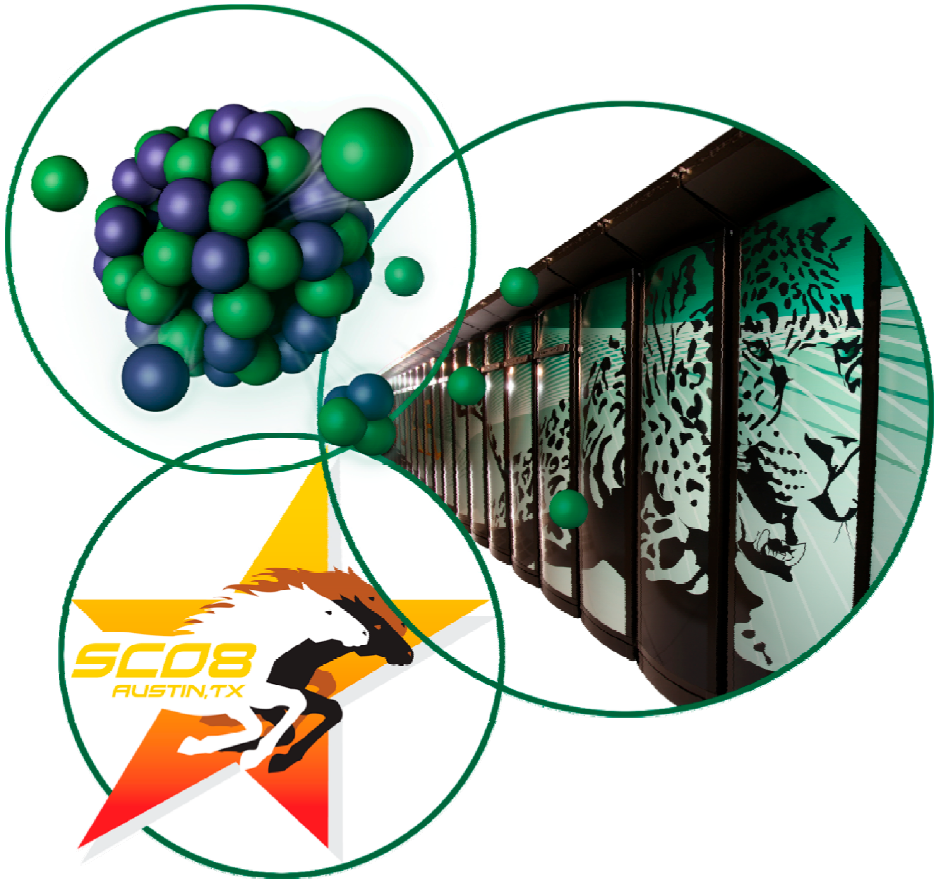


- **Best: Prediction accuracy >60% and checkpoint interval 16−32 hours**
- **Better than only proactive or only reactive**
- **Results for higher prediction accuracies and very low checkpoint intervals are worse than only proactive or only reactive**

Number of processes	**125**
Active nodes / Spare nodes	**125 / 12**
Checkpoint overhead	**50 min/checkpoint**
Migration overhead	**1 min/migration**

Simulation based on ASCI White system logs (nodes 1 – 125 and 500-512)

System-Level Virtualization and OSCAR-V



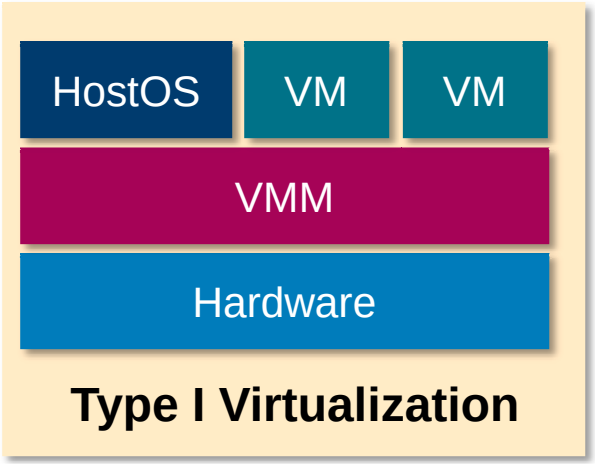
Presented by

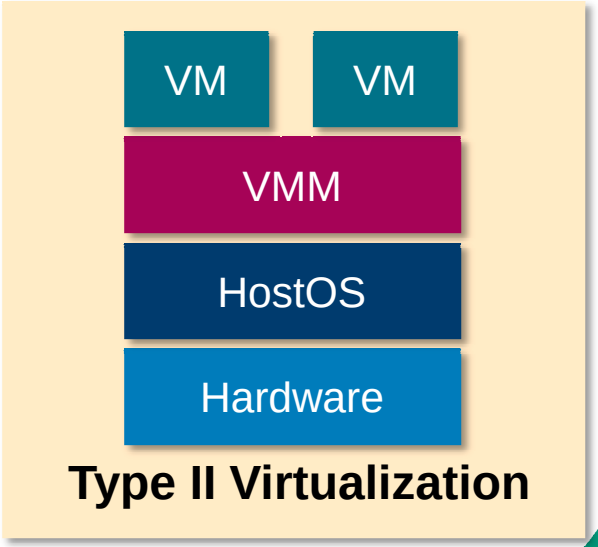
Stephen L. Scott
Thomas Naughton
Geoffroy Vallée

Computer Science Research Group
Computer Science and Mathematics Division

System-level virtualization

- First research in the domain, Goldberg—73
  - Type-I virtualization
  - Type-II virtualization
- Xen created a new real interest
  - Performance (paravirtualization)
  - Open source
  - Linux based
- Interest for high-performance computing (HPC)
  - VMM bypass
  - Network communication optimization
  - Etc.





Virtual machines

- **Basic terminology**
  - **Host OS: The OS running on a physical machine**
  - **Guest OS: The OS running on a virtual machine**
- **Today, different approaches**
  - **Full virtualization: Run an unmodified OS**
  - **Paravirtualization: Modification of OS for performance**
  - **Emulation: Host OS and Guest OS can have different architecture**
  - **Hardware support: Intel-VT, AMD-V**

Why virtualization in HPC?

- **Improved utilization**
  - **Users with differing OS requirements can be easily satisfied, e.g., Linux, Catamount, others in future**
  - **Enable early access to petascale software environment on existing smaller systems**
- **Improved manageability**
  - **OS upgrades can be staged across VMs and thus minimize downtime**
  - **OS/RTE can be reconfigured and deployed on demand**
- **Improved reliability**
  - **Application-level software failures can be isolated to the VMs in which they occur**
- **Improved workload isolation, consolidation, and migration**
  - **Seamless transition between application development and deployment using petascale software environment on development systems**
  - **Proactive fault tolerance (preemptive migration) transparent to OS, runtime, and application**

Why a virtualization specifically *for* HPC?

- **Networking**
  - **Bridges vs. zero copy (VMM bypass)**
  - **No RDMA support**
- **Memory: Important vs. minimal memory footprint**
- **Processor: Current solutions treat multicores as SMPs**
- **Tools: No tools available for the management of hundreds of VMs, hypervisors, and Host OSs**

Three approaches

Investigate the development of an HPC hypervisor

1. New hypervisor from scratch
2. New hypervisor using the microkernel Catamount
3. New hypervisor modifying and extending Xen

1 Hypervisor from scratch

- Develop a new hypervisor using GeekOS
- Current status: A minimal hypervisor has been developed supporting Intel-VT

Pros

- Only necessary features
- Very small system footprint

Cons

- Longer-term effort

2 Hypervisor based on Catamount

- **Extend Catamount**
  - **To be used as hypervisor**
  - **As Guest OS**
- **Current status: Catamount ported to XenoLinux**

Pros

- Very small system footprint
- Provide the XT environment within the VMs

Cons

- Still based on the Xen hypervisor

3 Xen-based hypervisor

- Remove unneeded Xen features
- Extend the hypervisor for adaptation (concept of modules)
  - Adaptation capability
  - Designed FY 2007
  - Implementation FY 2008
- Current status
  - Paravirtualization supported
  - Working toward full virtualization (Intel-VT, AMD-V)

Pros

- Quick prototyping
- Compatibility with emerging architectures

Cons

- No optimization (yet)

Reaping the benefit of virtualization: Proactive fault tolerance

- **Context**
  - **Large-scale systems are often subject to failures as result of the number of distributed components**
  - **Checkpoint/restart does not scale very well**
- **Provide capabilities for proactive fault tolerance**
  - **Failure prediction**
  - **Migrate application away from faulty node**
    - **Without stopping application**
    - **Without application code knowledge (or code modification)**

Proactive fault tolerance
(System and application resilience)

- **Modular framework**
  - **Support virtualization: Xen, VMM-HPC**
  - **Designed to support process-level checkpoint/restart and migration**
  - **Proactive fault-tolerance adaptation: Possible to implement new policies using our SDK**
- **Policy simulator**
  - **Ease the initial phase of study of new policies**
  - **Results from simulator match experimental virtualization results**

Virtual system environment

- Powerful abstraction concept that encapsulates OS, application runtime, and application
- Virtual parallel system instance running on a real HPC system using system-level virtualization
- Key issues addressed
  - Usability through virtual system management tools
  - Partitioning and reliability using adaptive runtime
  - Efficiency and reliability via proactive fault tolerance
  - Portability and efficiency through hypervisor + Linux/Catamount

OSCAR-V

Enhancements to support virtual clusters

- OSCAR-core modifications
- Create OSCAR Packages for virtualization solutions
- Integrate scripts for automatic installation and configuration
- Manage both Host OSs and VMs

Abstracts differences in virtualization solutions

- Must provide abstraction layer and tools—*libv3m/v2m*
- Enable easy switch between virtualization solutions
- High-level definition and management of VMs: Mem/cpu/etc., start/stop/pause

OSCAR-V: Image management

Host OS

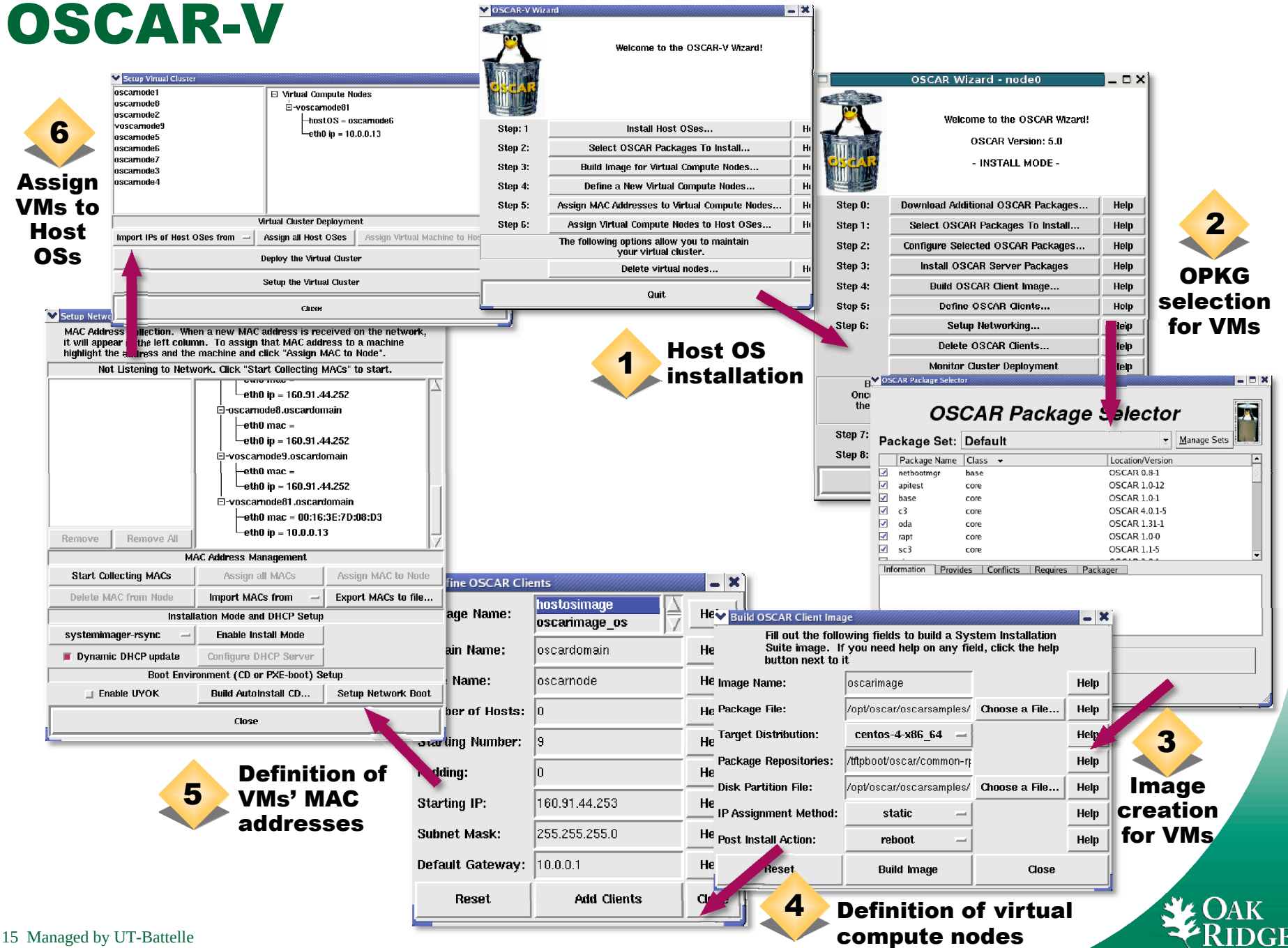
- OSCAR Packages (OPKG) are available
  - Xen case: Xen hypervisor, Xen kernels (dom0, domU), Xen tools
- Use the unmodified OPKG/OPD mechanism
  - Automatically add software components
  - Automatically set up the virtualization solution
- Current limitation
  - Only REHL, CentOS, Fedora Core are currently supported

Virtual machines

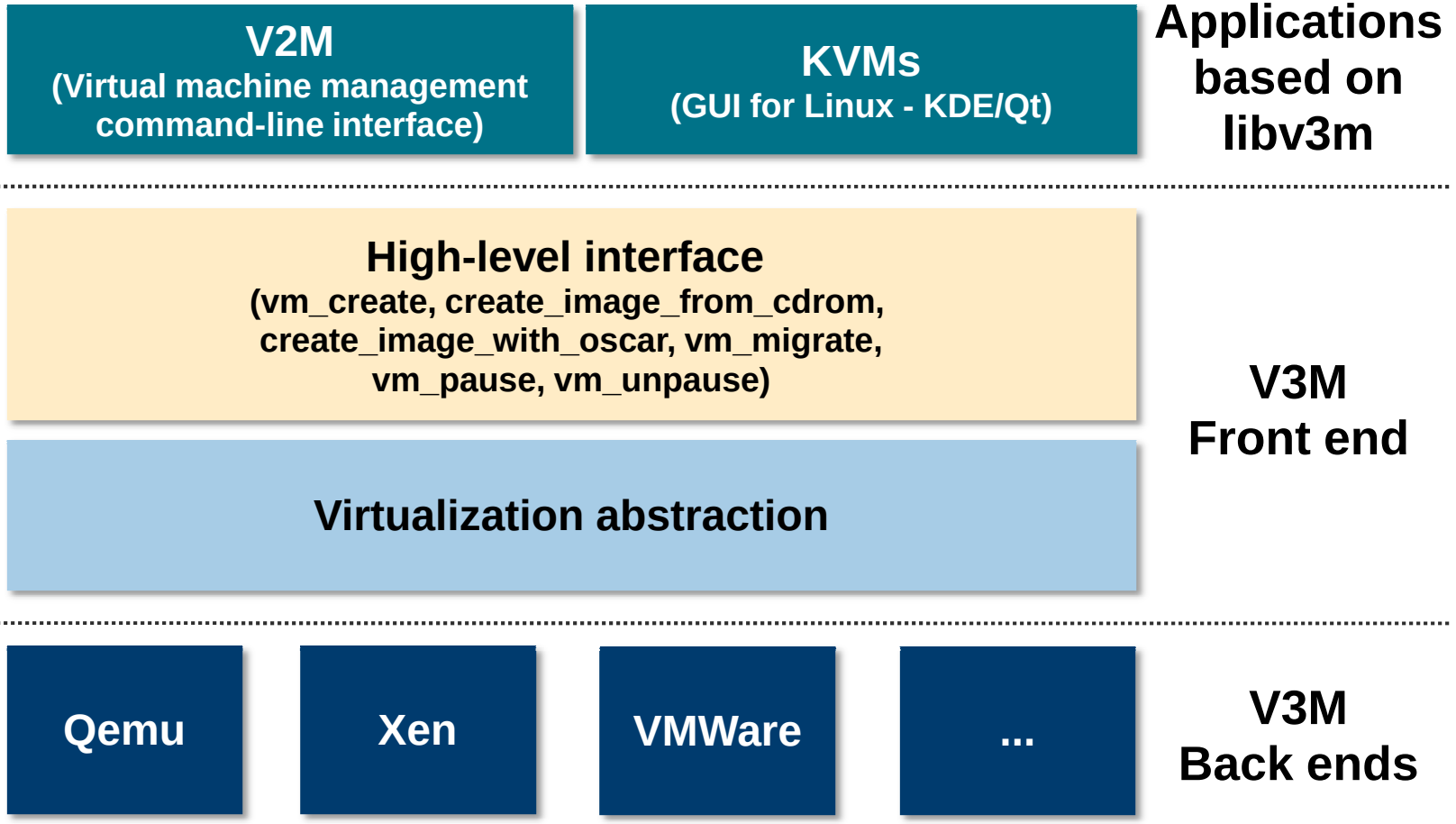
- One OSCAR Package is available
  - Automatically includes the kernel (optional)
  - Automatically sets up the environment
- OSCAR can be used to define VMs
  - Set up the number of VMs
  - MAC addresses
  - IPs

Virtual machines may be deployed

OSCAR-V


6
Assign VMs to Host OSs
Setup Virtual Cluster
oscarnode1
oscarnode8
oscarnode2
voscarnode9
oscarnode5
oscarnode6
oscarnode7
oscarnode3
oscarnode4
Virtual Compute Nodes
-voscarnode81
hostOS = oscarnode6
eth0 ip = 10.0.0.13
Virtual Cluster Deployment
Import IPs of Host OSes from
Assign all Host OSes
Deploy the Virtual Cluster
Setup the Virtual Cluster
Close
OSCAR-V Wizard
Welcome to the OSCAR-V Wizard!
Step: 1
Install Host OSes...
Step 2:
Select OSCAR Packages To Install...
Step 3:
Build Image for Virtual Compute Nodes...
Step 4:
Define a New Virtual Compute Nodes...
Step 5:
Assign MAC Addresses to Virtual Compute Nodes...
Step 6:
Assign Virtual Compute Nodes to Host OSes...
The following options allow you to maintain your virtual cluster.
Delete virtual nodes...
Quit
OSCAR Wizard - node0
Welcome to the OSCAR Wizard!
OSCAR Version: 5.0
- INSTALL MODE -
Step 0:
Download Additional OSCAR Packages...
Step 1:
Select OSCAR Packages To Install...
Step 2:
Configure Selected OSCAR Packages...
Step 3:
Install OSCAR Server Packages
Step 4:
Build OSCAR Client Image...
Step 5:
Define OSCAR Clients...
Step 6:
Setup Networking...
Delete OSCAR Clients...
Monitor Cluster Deployment
Help
2
OPKG selection for VMs
1
Host OS installation
Setup Network
MAC Address collection. When a new MAC address is received on the network, it will appear in the left column. To assign that MAC address to a machine highlight the address and the machine and click "Assign MAC to Node".
Not Listening to Network. Click "Start Collecting MACs" to start.
eth0 ip = 160.91.44.252
-oscarnode8.oscardomain
eth0 mac =
eth0 ip = 160.91.44.252
-voscarnode9.oscardomain
eth0 mac =
eth0 ip = 160.91.44.252
-voscarnode81.oscardomain
eth0 mac = 00:16:3E:7D:08:D3
eth0 ip = 10.0.0.13
Remove
Remove All
MAC Address Management
Start Collecting MACs
Assign all MACs
Assign MAC to Node
Delete MAC from Node
Import MACs from
Export MACs to file...
Installation Mode and DHCP Setup
systemimager-rsync
Enable Install Mode
Dynamic DHCP update
Configure DHCP Server
Boot Environment (CD or PXE-boot) Setup
Enable UYOK
Build AutoInstall CD...
Setup Network Boot
Close
OSCAR Package Selector
OSCAR Package Selector
Package Set: Default
Manage Sets
Package Name
Class
Location/Version
netbootmgr base OSCAR 0.8-1
apitest core OSCAR 1.0-12
base core OSCAR 1.0-1
c3 core OSCAR 4.0.1-5
oda core OSCAR 1.31-1
rapt core OSCAR 1.0-0
sc3 core OSCAR 1.1-5
Information
Provides
Conflicts
Requires
Packager
3
Image creation for VMs
Build OSCAR Client Image
Fill out the following fields to build a System Installation Suite image. If you need help on any field, click the help button next to it
Image Name: oscarimage
Package File: /opt/oscar/oscarsamples/
Choose a File...
Target Distribution: centos-4-x86_64
Package Repositories: /tftpboot/oscar/common-rp
Disk Partition File: /opt/oscar/oscarsamples/
Choose a File...
IP Assignment Method: static
Post Install Action: reboot
Reset
Build Image
Close
Help
fine OSCAR Clients
age Name: hostosimage
oscarimage_os
ain Name: oscardomain
Name: oscarnode
ber of Hosts: 0
Starting Number: 9
dding: 0
Starting IP: 160.91.44.253
Subnet Mask: 255.255.255.0
Default Gateway: 10.0.0.1
Reset
Add Clients
4
Definition of virtual compute nodes
5
Definition of VMs' MAC addresses

OSCAR-V: V2M—virtual machine management



OSCAR-V: V3M—supported features summary

Supported features	Xen (paravirtualization	Xen (full virtualization)	Qemu	VM ware
VM instantiation	Yes	Yes	Yes	Yes
VM image creation	Yes	Yes	Yes	No
Installation via CD-ROM	N/A	Yes	Yes	No
Installation via OSCAR	Yes	Yes	Yes	No
VM migration	Yes	Experimental	No	No
VM pause/unpause	Yes	Experimental	Experimental	Experimental
Virtual disk	Yes	Yes	Yes	Yes

Virtualization Collaboration Team



Lead the development of new hypervisor from scratch



Lead the development of new hypervisor based on Catamount



Lead the development of new hypervisor by modifying and extending Xen

Contacts



- **Dr. Christian Engelmann**

System Research Team
Computer Science Research Group
Computer Science and Mathematics Division
Oak Ridge National Laboratory

engelmannc@ornl.gov
www.csm.ornl.gov/~engelman

OAK RIDGE National Laboratory