Recent Advanced Computing Infrastructure Developments at Penn State

a presentation for the
HPC User Forum

Imperial College, London
July 5-6, 2012

HLRS, University of Stuttgart
July 9-10, 2012

Vijay K. Agarwala
Senior Director, Research Computing and CI
Information Technology Services
The Pennsylvania State University
University Park, PA 16802 USA

PENNSTATE 1855

its@PennState
information technology services

Research Computing and Cyberinfrastructure (RCC) Services

At a high level, RCC services can be grouped into three categories:

1. Teaching workshops, seminars, and guest lectures in courses in the areas of large-scale computational techniques and systems
2. Providing one-on-one support and consulting for independent curiosity-driven research
3. Providing support for sponsored research

Services provided by RCC

- Provide systems services by researching current practices in operating system, file system, data storage, job scheduling as well as computational support related to compilers, parallel computations, libraries, and other software support. Also supports visualization of GIS data and large computed datasets to gain better insight from the results of simulations.
- Enable large-scale computations and data management by building and operating several state-of-the art computational engines.
- Consolidate and thus significantly increase the research computing resources available to each faculty participant. Faculty members can frequently exceed their share of the machine to meet peak computing needs.
- Provide support and expertise for using programming languages, libraries, and specialized data and software for several disciplines.
- Investigate emerging visual computing technologies and implement leading-edge solutions in a cost-effective manner to help faculty better integrate data visualization tools and immersive facilities in their research and instruction.
- Investigate emerging architectures for data and numerically intensive computations and work with early-stage companies in areas such as interconnects, networking, and graphics processors.
- Help build inter- and intra-institutional research communities using cyberinfrastructure technologies.
- Maintain close contacts with NSF, DoE, NIH, NASA and other federally funded centres, and help faculty members with porting and scaling of codes across different systems.

Programs, Libraries, and Application Codes in Support of Computational Research

- **Compilers and Debuggers:** DDT, GNU Compilers, Intel Compiler Suite, NVIDIA CUDA, PGI Compiler Suite, TotalView
- **Computational Biology:** BLAST, BLAST+, EEGLAB, FSL, MRIcro, MRICron, SPM5, SPM8, RepeatMasker, wuBlast
- **Computational Chemistry and Material Science:** Accelyrs Materials Studio, Amber, CCP4, CHARMM , CPMD, Gamess, Gaussian 03, Gaussian 09, GaussView, Gromacs, LAMMPS, NAMD, NWChem, Rosetta, Shrondinger Suite, TeraChem, ThermoCalc , VASP, WIEN2K, WxDragon
- **Finite Element Solvers:** ABAQUS, LS-DYNA, MD/Nastran and MD/Patran
- **Fluid Dynamics:** Fluent, GAMBIT, OpenFOAM, Pointwise
- **Mathematical and Statistical Libraries and Applications:** AMD ACML, ATLAS, BLAS, IMSL, LAPACK, GOTO, Intel MKL, Mathematica, MATLAB, Distributed MATLAB, NAG, PETSc, R, SAS, WSMP
- **Multiphysics:** ANSYS, Comsol
- **Optimization:** AMPL, CPLEX, GAMS, Matlab Optimization Toolbox, Matgams, OPL
- **Parallel Libraries:** OpenMPI, Parallel IMSL, ScaLAPACK
- **Visualization Software:** Avizo, Grace, IDL, Tecplot, VisIt, VMD

All software installations are driven by faculty. The software stack on every system is customized and entirely shaped by faculty needs.

Visualization Services and Facilities

Promote effective use of visualization and VR techniques in research and teaching via strategic deployment of facilities and related support. Staff members provide consulting, teach seminars, assist faculty and support facilities for visualization and VR.

Recent support areas include:

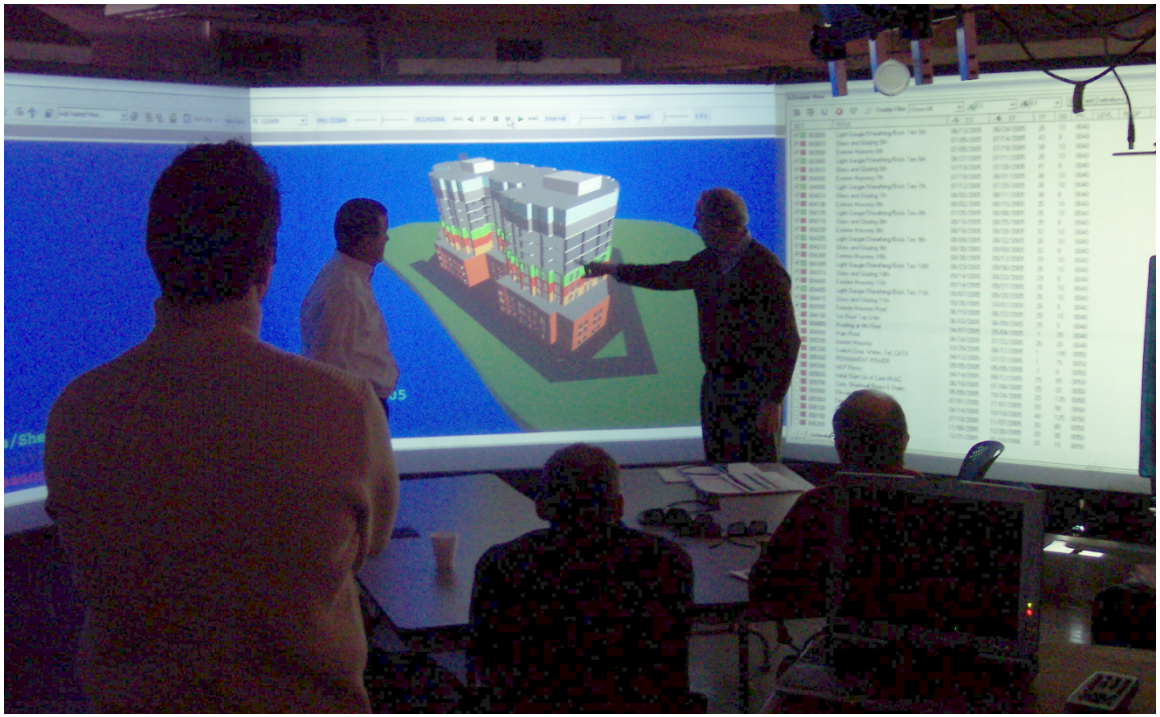
- Visualization and VR system design and deployment
- 3D Modeling applications: FormZ, Maya
- Data Visualization applications: OpenDX, VTK, VisIt
- 3D VR development and device libraries: VRPN, VMRL, JAVA3D, OpenGL, OpenSG, CaveLib
- Domain specific visualization tools: Avizo (Earth, Fire, Green, Wind), VMD, SCIRun
- Telecollaborative tools and facilities: Access Grid, inSORS, VNC
- Parallel graphics for scalable visualization: Paraview, DCV, VisIt
- Programming support for graphics (e.g. C/C++, JAVA3D, Tcl/Tk, Qt)

Successful installations:

- Immersive Environments Lab: in partnership with School of Architecture and Landscape Architecture
- Immersive Construction Lab: in partnership with Architecture Engineering
- Visualization/VR Lab: in partnership with Materials Simulation Center
- Visualization/VR Lab: in partnership with Computer Science and Engineering
- Sports Medicine VR Lab: a partnership between Kinesiology, Athletics and Hershey Medical Center

Immersive Construction Lab: Virtual Environments for Building Design, Construction Planning and Operations Management

Dr. John I. Messner
Associate Professor of Architectural Engineering
College of Engineering



its@PennState
information technology services

Immersive Environments Lab: Virtual Environments for Architectural Design Education and Telecollaborative Design Studio

Loukas Kalisperis
Professor of Architecture
College of Arts and Architecture

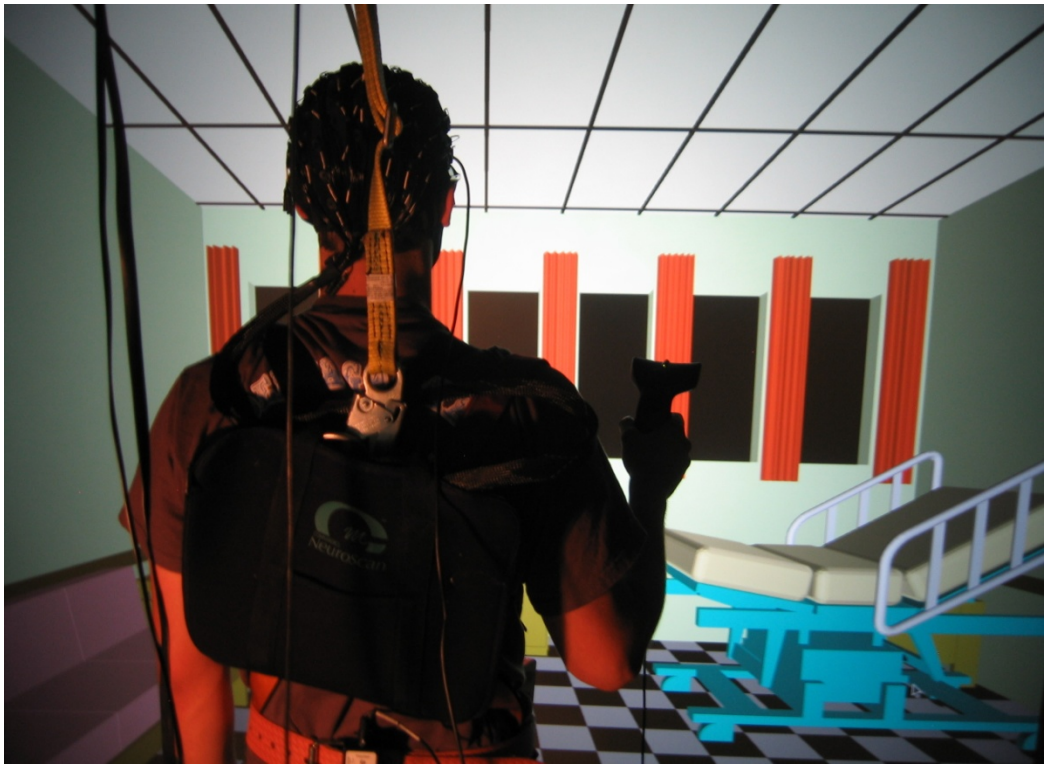
Katsuhiko Muramoto
Associate Professor of Architecture
College of Arts and Architecture

its@PennState
information technology services

Virtual Reality Lab: VR for Assessment of Traumatic Brain Injury

Dr. Semyon Slobounov
Professor of Kinesiology
College of Health and Human Development

Dr. Wayne Sebastianelli
Director of Sports Medicine
College of Medicine



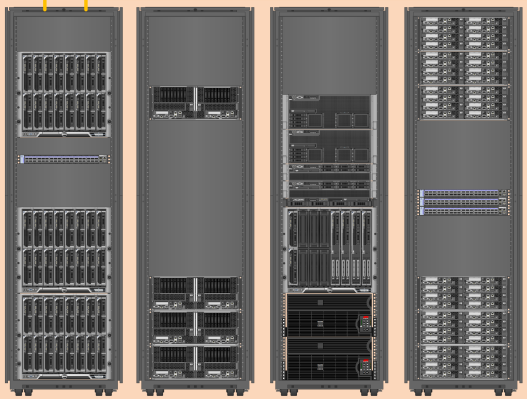
its@PennState
information technology services

RCC support of sponsored research programs

- Lion-X clusters are built with a combination of RCC/ITS funds and investments from research grants by participating faculty members.
- RCC partners with faculty to provide a guaranteed level of service in exchange for keeping Lion-X clusters open to the broader community at Penn State.
- Faculty partnerships enable researchers to access a larger shared system than they would be able to build on their own, an extensive and well-supported software stack, and access to computing expertise of RCC staff members.

Penn State RCC/ITS Distributed Data Center

Compute Capacity at Computer Building

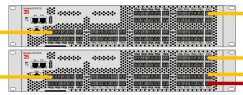


Lion-XG: 192 nodes with 16 cores (E5-2665), 64-128 GB of memory, FDR/14 Infiniband
Lion-XF: 184 nodes with 12 cores (X5675), 48-96 GB of memory, QDR Infiniband
Lion-GA: 8 nodes with 12 cores (X5675), 48 GB of memory, QDR IB, 8 Tesla M2090 GPUs
224 cores in machines with memory >= 256 GB (up to 1 TB) with 4 TB of Flash
288 cores in support of interactive jobs and visualization
2416 additional cores in general purpose Infiniband clusters

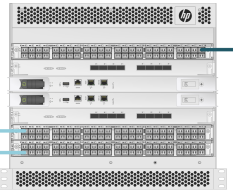
Legend

- 1 Gb/s Ethernet
- 10 Gb/s Ethernet
- 8 Gb/s FibreCh
- 16 Gb/s FibreCh

40 Gb/s TRILL agg.

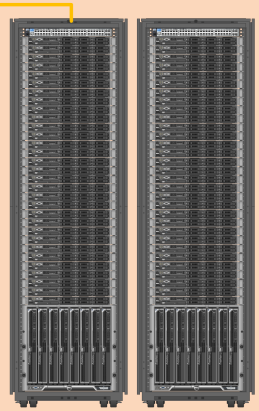


Two Brocade VDX 6720
60-port 10Gb/s Ethernet
switches



HP SN8000b
4-slot FibreChannel director
with 144 16Gb/s ports active

Compute Capacity at Pattee

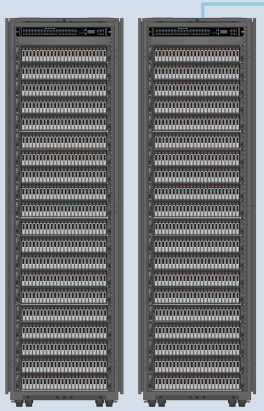


CyberStar
192 nodes: 8 cores, 24 GB of memory, QDR Infiniband
32 nodes: 16 cores, 96 GB of memory, QDR Infiniband

Compute Capacity at EES

Lion-XJ
128 nodes: 8 cores, 32 GB of memory, DDR Infiniband
16 nodes: 8 cores, 64 GB of memory, DDR Infiniband

Storage Capacity at Computer Building



1.2 PB 7.2K NL-SAS and 4 TB Flash

IBM TS3500 tape library with 2.4 PB capacity and 8 drives

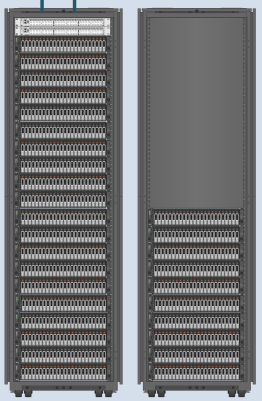
Compute at Hershey

256 core cluster co-located in Hershey, PA

Compute Capacity at MSC

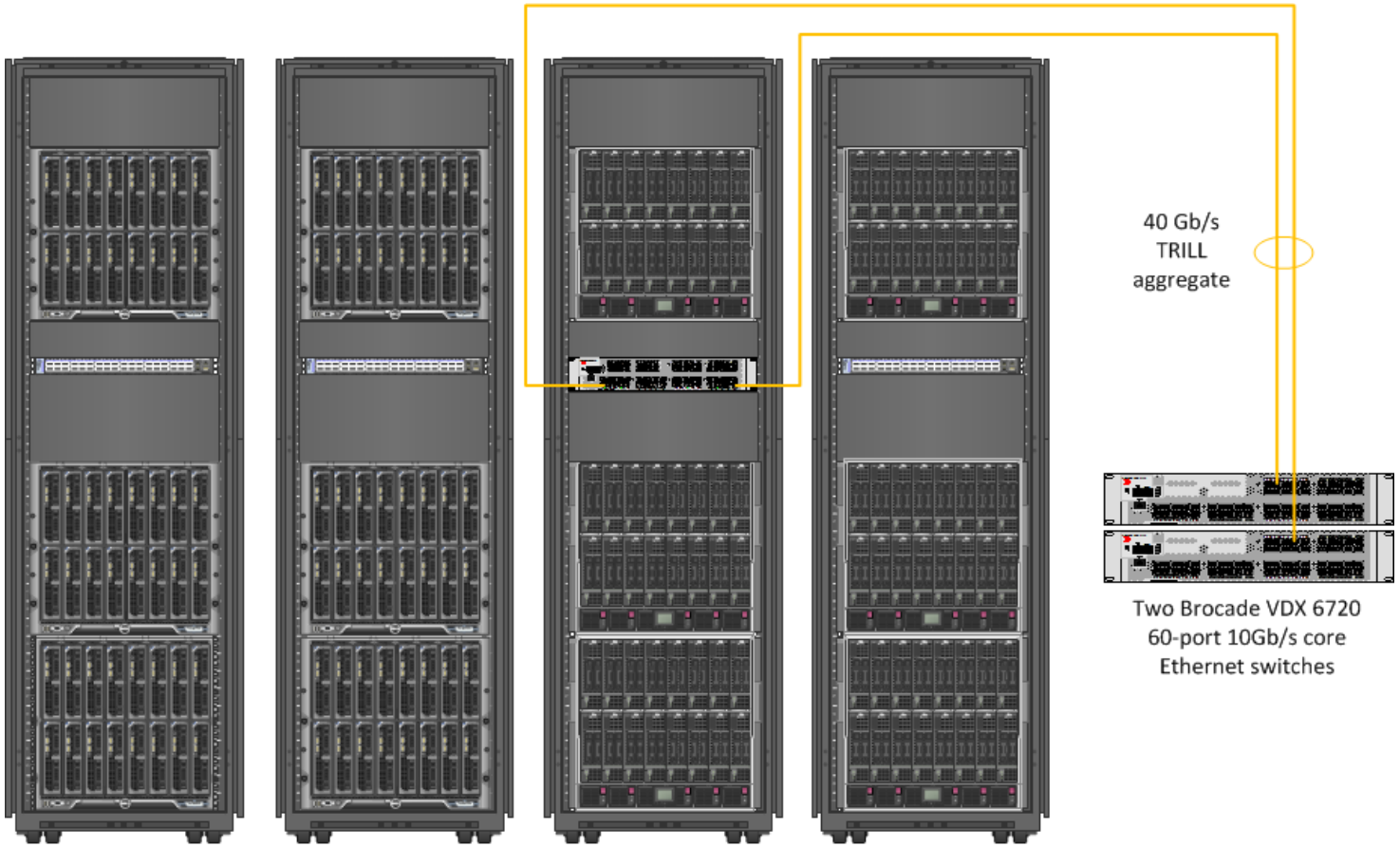
640 core BioStar cluster and
256 core CyberStar-II cluster with
8 to 16 GB of memory per core

Storage Capacity at MSC



644 TB of storage for BioStar and
350 TB of storage for CyberStar-II

Lion-XG – The latest compute engine at Penn State RCC/ITS



- 144 nodes with two 8-core Intel Xeon “Sandy Bridge” E5-2665 processors, 64 GB of memory
- 48 nodes with two 8-core Intel Xeon “Sandy Bridge” E5-2665 processors, 128 GB of memory
- 3,072 total cores and 15 TB of total memory. Estimated maximum performance of 50 TFLOPS.
- 10 Gb/s Ethernet and 56 Gb/s FDR InfiniBand to each node. The InfiniBand fabric has full bisection bandwidth within each 16-node chassis and half bisection bandwidth among the 12 chassis
- Brocade VDX6730 10 Gb/s Ethernet with 16 ports of 8 Gb/s FibreChannel to enable FCoE
- Transparent Interconnect of Lots of Links (TRILL) connection to RCC core switches using Brocade VCS technology

Performance Impact of Recent Advanced Computing Infrastructure Developments

1. Storage upgrade to FibreChannel attached arrays
2. GPFS metadata upgrade with FibreChannel attached Flash memory arrays
3. Large memory servers accelerate Big Data science
4. GPU-enabled clusters accelerate research across disciplines

Storage upgrade to FibreChannel attached arrays

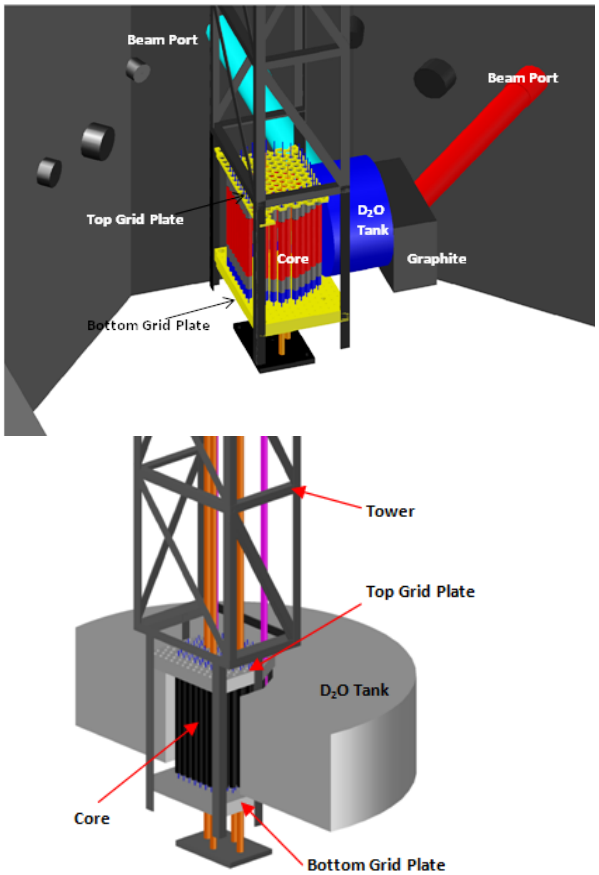
- Previous GPFS Storage Architecture
  - Most storage SAS attached to pairs of servers.
  - If two servers in a pair became unavailable, data became unavailable.
  - Access to most storage for GPFS NSD servers was through 10 Gbps Ethernet.
- Current GPFS Storage Architecture
  - All storage and GPFS NSD servers Fibre Channel attached to 16 Gbps Fibre Channel switch.
  - Any GPFS NSD server has direct block level access to all storage, increasing speed of system operations such as backup and GPFS operations such as migrating data to new disk.
  - Seven NSD servers in failover group for any filesystem - possible to lose up to six servers and still have access to data, assuming quorum isn't lost.
  - Allows creation of fast Data Mover hosts with direct Fibre Channel access to GPFS storage for users to upload or download data - no penalty of sending data over Ethernet to the NSD servers to get to storage.

GPFS metadata upgrade with FibreChannel attached Flash memory arrays

- GPFS Metadata issues faced
  - Metadata IO intensive system operations such as backup and GPFS policy scans bottlenecked by previous metadata disk performance.
  - GPFS file system responsiveness to user operations impacted during intensive system operations.
- Previous Metadata State
  - 100 15K RPM SAS drives in multiple drive enclosures, drawing about 2.5 KW, using 20U of rack space, providing 19,000 IOPS to five GPFS filesystems with 0.5 PB capacity.
- Current Metadata State
  - 2 Flash Memory Arrays, drawing about 500 watts, using 2U of rack space, **providing 400,000 IOPS to nine GPFS filesystems with 2 PB capacity**.
- Effect of Moving to Flash Memory Arrays for GPFS Metadata
  - Nightly incremental backups: Time to run nightly incremental backup on 212 million files reduced from **six hours to one hour**.
  - Able to run GPFS policy scan on **400 million files in five minutes**.
  - Responsiveness of file systems to users not impacted by metadata intensive system operations or general metadata load. **Time to open/save/create files greatly reduced.**

Large memory servers accelerate Big Data science

Thermal-Hydraulic Analysis of the New Penn State Breazeale Reactor Core using **ANSYS Fluent®** Code

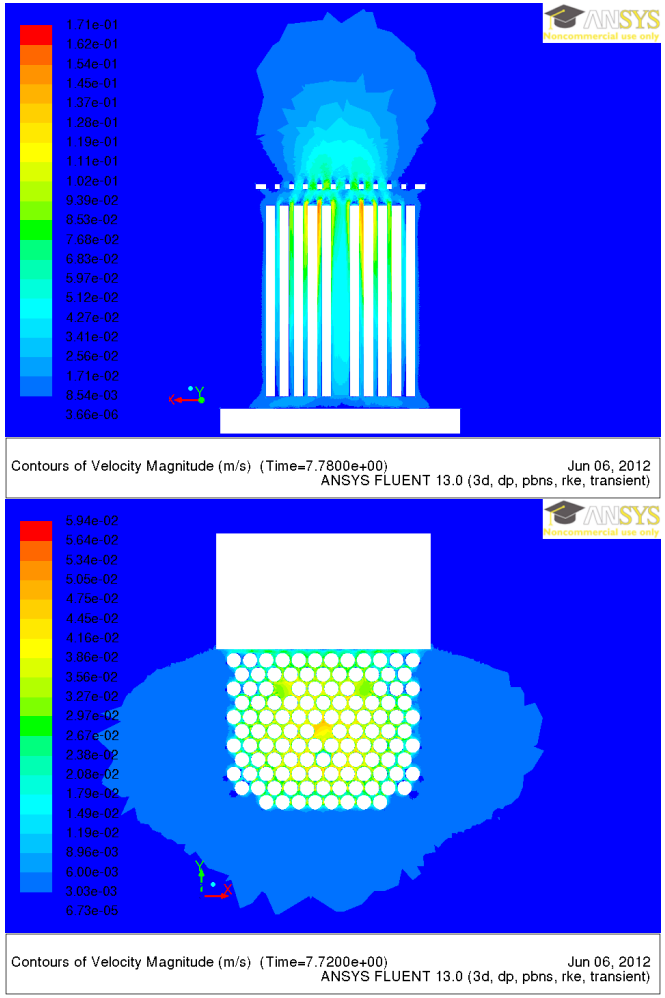


3D CAD drawings of existing and new PSBR designs

- Penn State Breazeale Reactor (PSBR) is a MARK-III type TRIGA design with 1 MWt nominal power
- Existing PSBR core has some inherent design problems which limits the experimental capability of the facility
- A thermal-hydraulics analysis of the new design was initiated to verify safe operation of the new core-moderator configuration
- The purpose of the study is to generate **ANSYS Fluent®** CFD models of existing and new PSBR designs to calculate the temperature and flow profile in core and around the core-moderator designs

Large memory servers accelerate Big Data science

Thermal-Hydraulic Analysis of the New Penn State Breazeale Reactor Core using **ANSYS Fluent** ® Code



Velocity profile calculated by Fluent in the center of existing PSBR design (Side View & Top View)

- Two CFD models of the existing and new PSBR designs were prepared using **ANSYS Gambit** ®, each with a total of ~30 million structured and unstructured 3D cells
- Simulation of either model requires ~300 GB of memory, something that can only be achieved on our large memory machine Lion-LSP
- Simulation runtime on 24 processors is about a day

its@PennState
information technology services

Large memory servers accelerate Big Data science

Climate Change Impacts on Household Location Choice

- This research employs a two-stage residential sorting model to examine climate change impacts on residential location choices in the US
- The main dataset used for estimation is the Integrated Public Use Microdata Sample (IPUMS), which provides demographic characteristics of approximately 2.4 million households located in 283 Metropolitan Statistical Areas (MSAs) of the US in the year 2000
- A two-stage random utility sorting model (RUM) is employed
- The first-stage discrete choice model employs a multinomial logit specification to recover heterogeneous parameters associated with MSA specific variables, migration costs, along with the mean indirect utility of each MSA
- In particular, the interaction terms of temperature extremes and individual-specific characteristics, such as one's birth region, age and educational attainment, are used to recover valuations of temperature extremes for different classes of people with potentially different preferences
- The second stage of this model decomposes the mean indirect utility obtained from the first stage into its MSA-specific attributes controlling for unobservables using region fixed effects.

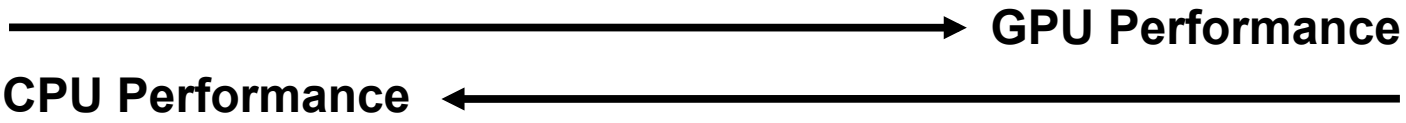
Large memory servers accelerate Big Data science

Climate Change Impacts on Household Location Choice

- The model is simulated as an optimization problem in **MATLAB ®** using FMINUNC
- The dependent variable, whether an individual resided in an MSA in the year 2000, is a matrix of size 2.4 million × 283
- In addition, there are 11 to 15 interaction variables for the interaction of the individual's demographic factors with extreme temperatures in the MSA, each a matrix of the same size as the dependent variable
- Matrix operations on such large datasets required more than 150 GB of memory, hence Lion-LSP was used
- This simulation required 2 to 3 days to complete on a single processor

GPU-enabled clusters accelerate research across disciplines

Finite Differences - Small BLAS/LAPACK - Large BLAS/LAPACK - Monte Carlo
Density Functional Theory - Sparse Linear Algebra - Quantum Chemistry
Reverse Time Migration - Kirchhoff Depth Migration - Kirchhoff Time Migration

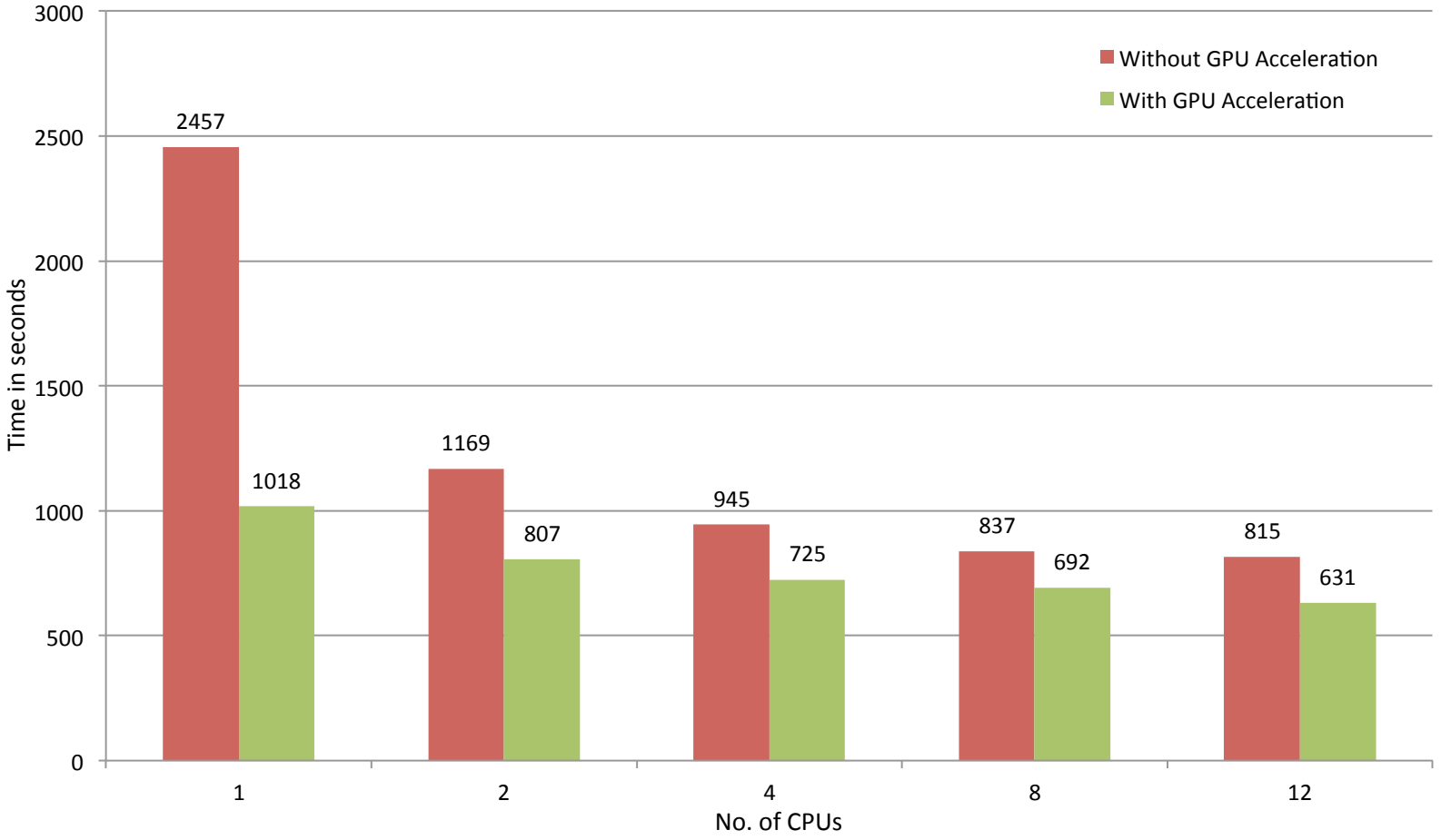


- Several key entry points to GPU acceleration, PSU engages at several levels :
  - Existing GPU applications eg., PetaChem, ANSYS, ABAQUS
  - Linking against libraries eg., cuBLAS
  - Using OpenACC compiler directives
  - Porting code to GPU using CUDA, OpenCL
- Depending on algorithm/application, realistically one experiences 2x-50x over optimized CPU code and GPU FLOPS/Watt >> CPU FLOPS/Watt
- Following examples compare Nvidia M2090 performance versus Intel Westmere; all applications compiled with -O2 optimization, SSE where applicable

GPU-enabled clusters accelerate research across disciplines

Accelerating Finite Element Analyses with GPUs

ANSYS 14.0, Sparse Solver: 376,000 DOFs

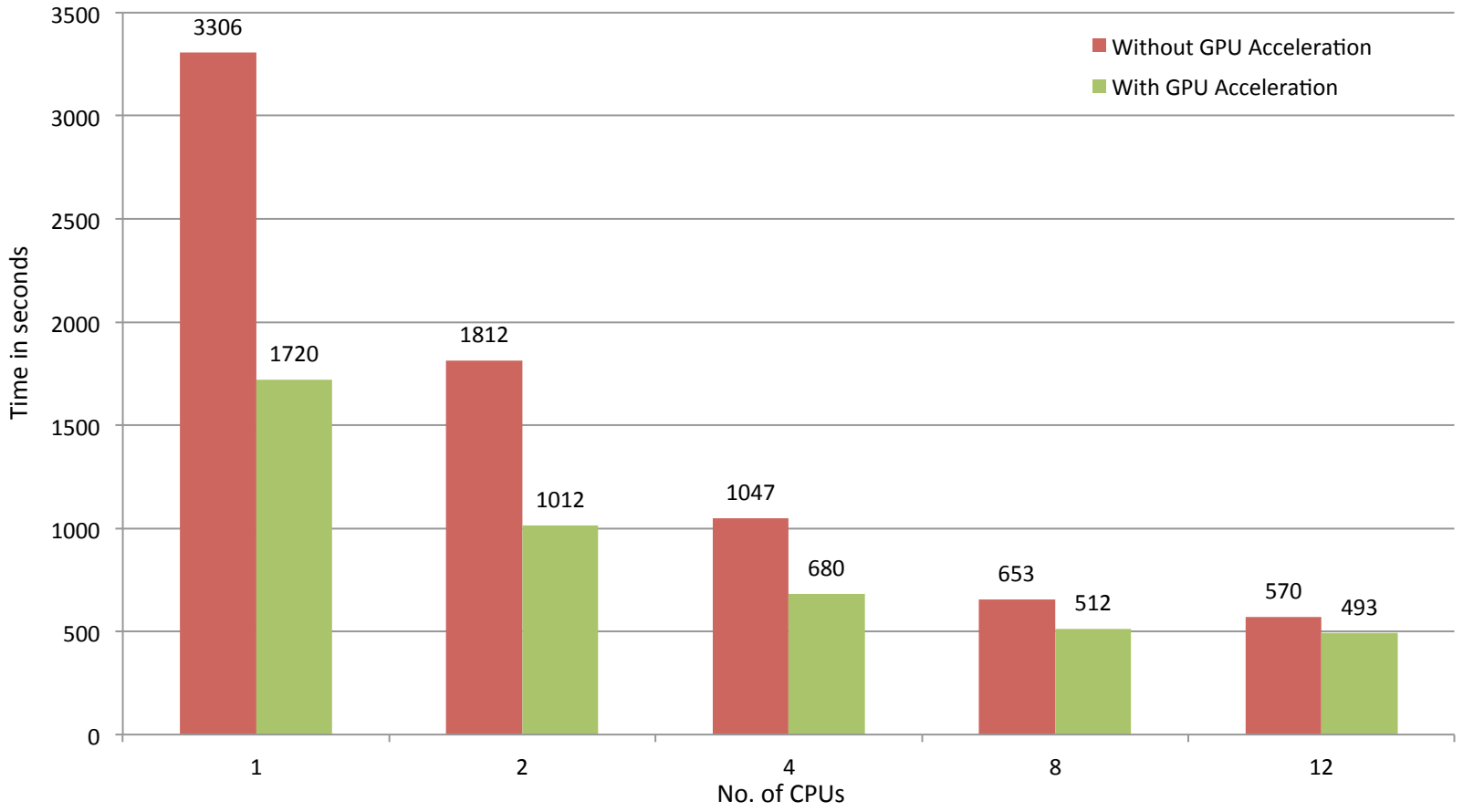


its@PennState
information technology services

GPU-enabled clusters accelerate research across disciplines

Accelerating Finite Element Analyses with GPUs

ABAQUS/Standard 6.11-1: 2.2 million DOFs



its@PennState
information technology services

GPU-enabled clusters accelerate research across disciplines

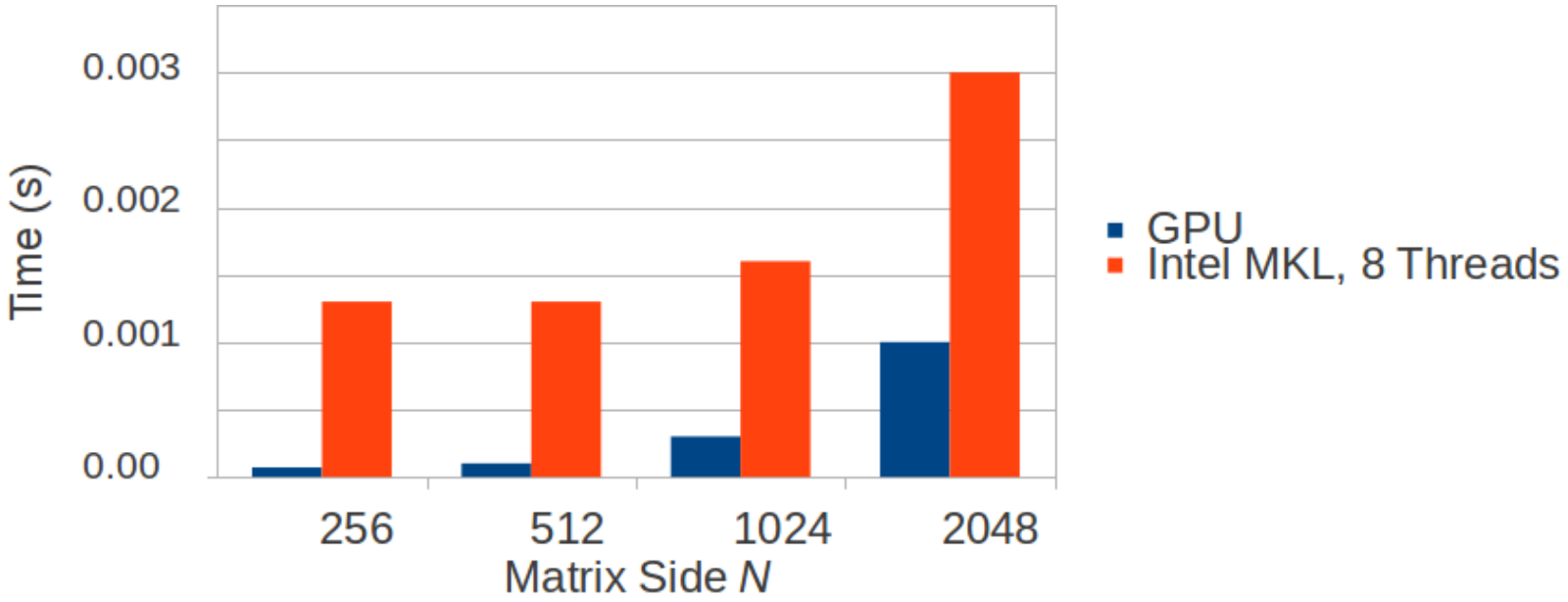
Fractional Quantum Hall Effect

- FQHE at filling fraction ν=5/2 is believed to support excitations that exhibit non-Abelian statistics, so far not observed in any other known systems
- Research involves numerical studies of competing variational wave functions Ψ that contain intricate correlations, requires $N!$ / $(N/2)!$ $(N/2)!$ *LU* decompositions of $(N/2)^2$ matrices
- GPU provides **24x** performance improvement over single CPU (4 times single socket/6 threads using OpenMP)
- ***Scaling is same when increasing batch size (small matrices < cache size)***
- For example, for *N*=11, computation time using 8 GPU devices (w/ MPI), 1024 Monte Carlo iterations is ~ 246 seconds from ~ 31488 single CPU

GPU-enabled clusters accelerate research across disciplines

Lanczos Algorithm

- Important algorithm for finding eigenvalues of large matrices eg., for latent semantic indexing in search
- Currently being ported to GPU for use in discovering topological exitons in FQHE (diagonalization of the Hamiltonian); matrices involved have side *N* > 1e6
- Algorithm relatively simple, although requires restart/re-orthogonalization throughout, uses for example `sgemv` routines from **MKL** and/or **cuBLAS**, initial results promising; **3x** speedup over 8 CPU threads



GPU-enabled clusters accelerate research across disciplines

Phylogenetics using BEAGLE

- BEAGLE is GPU/CPU library that performs the calculations (Markov processes) at the heart of most Bayesian and Maximum Likelihood phylogenetics packages
- At PSU both BEAST and mrBayes are linked against BEAGLE
- In general, Monte-Carlo based algorithms enjoy thread independence/ low communication overhead, very amenable to GPU
- Examples :
  - simulating genomic DNA with 16 taxa and 10000 site patterns (5 reps) with GPU provides ~ **10x** speedup over CPU
  - simulating genomic 64-state data with 10 taxa and 100 site patterns (5 reps) with GPU provides ~ **40x** speedup over CPU

GPU-enabled clusters accelerate research across disciplines

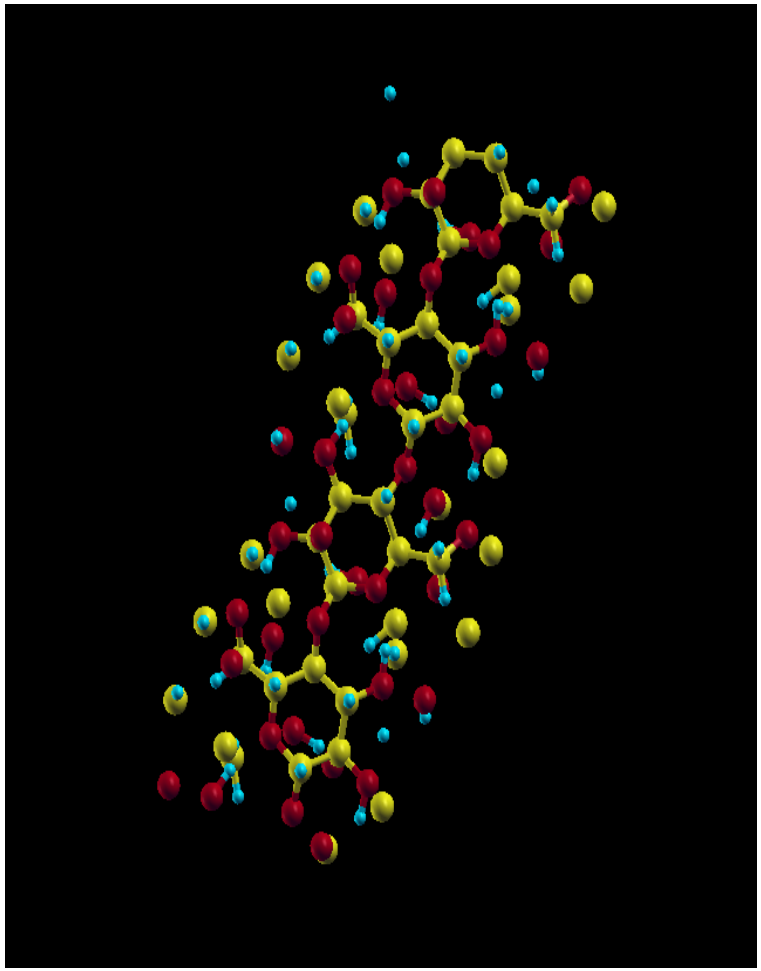
MD using Gromacs

- **Gromacs** performs classical molecular dynamics, usually employed for proteins and lipids, GPU acceleration by way of the **OpenMM** (https://simtk.org/home/openmm) library, a highly optimized set of molecular modeling routines
- Example : Amber99sb forcefield, homo-dimer with 381 residues and ~6100 total atoms, solvated in ~28,000 water molecules, with 14 chloride ions added for charge neutrality, PME electrostatics, NVT ensemble, Anderson thermostat; GPU gives approximately **7x** speedup over CPU
- Workflow used copious writes btwn host-device which effects performance, as data must traverse the PCIe bus

CPU	GPU	hrs/ns
1	1	14.657
1	0	105.357
2	0	46.846
4	0	24.784

GPU-enabled clusters accelerate research across disciplines

DFT study of Cellulose



Cellulose unit cell

- Density Functional Theory (DFT) has enjoyed huge growth in popularity owing to computational and numerical advancements; used widely in material science
- **Quantum Espresso** (QE) is an open source DFT package which has recently added GPU acceleration, largely through BLAS and FFT routines
- When building QE with **MAGMA** (UT/ORNL), one introduces heterogeneous linear algebra routines which use both CPU & GPU
- Throughout Self-Consistent Field (SCF) cycle in DFT, much communication between CPU and GPU is necessary, overall speedup **2x** speedup eg., for cellulose (168 atoms), PBE pseudopotentials, walltime is 7h29min (3 GPU's) vs 14h33min (3 CPU's)