High Performance Computing and Big Data with RDMA-enabled High Speed Interconnects – Delivering Science at SDSC for a Decade

Amit Majumdar
San Diego Supercomputer Center
University of California San Diego

MVAPICH User Group Meeting, Columbus, OH, August 6-8, 2018

Outline of Talk

1. Introduction
2. Trestles – A high-productivity HPC system targeted to modest-scale and gateway users 2011-2014
3. Gordon – An innovative data-intensive supercomputer (2009) 2012-2017
4. Comet – HPC for the long tail of science 2015 - 2021
5. NOWLAB/MVAPICH – Impact on science

1. Introduction

- Underlying theme is impact of MVAPICH and many other systems software coming out of the NOWLAB, OSU
  - Impact on HPC resources worldwide
  - Specifically their impact on science – in addition to latency/BW/scaling

- NOWLAB collaboration with researchers and teams (like us at SDSC) and others (like TACC) in the US and worldwide
  - To implement, optimize, research HPC system software on production resources
  - Uniqueness of research and impact on production HPC machines

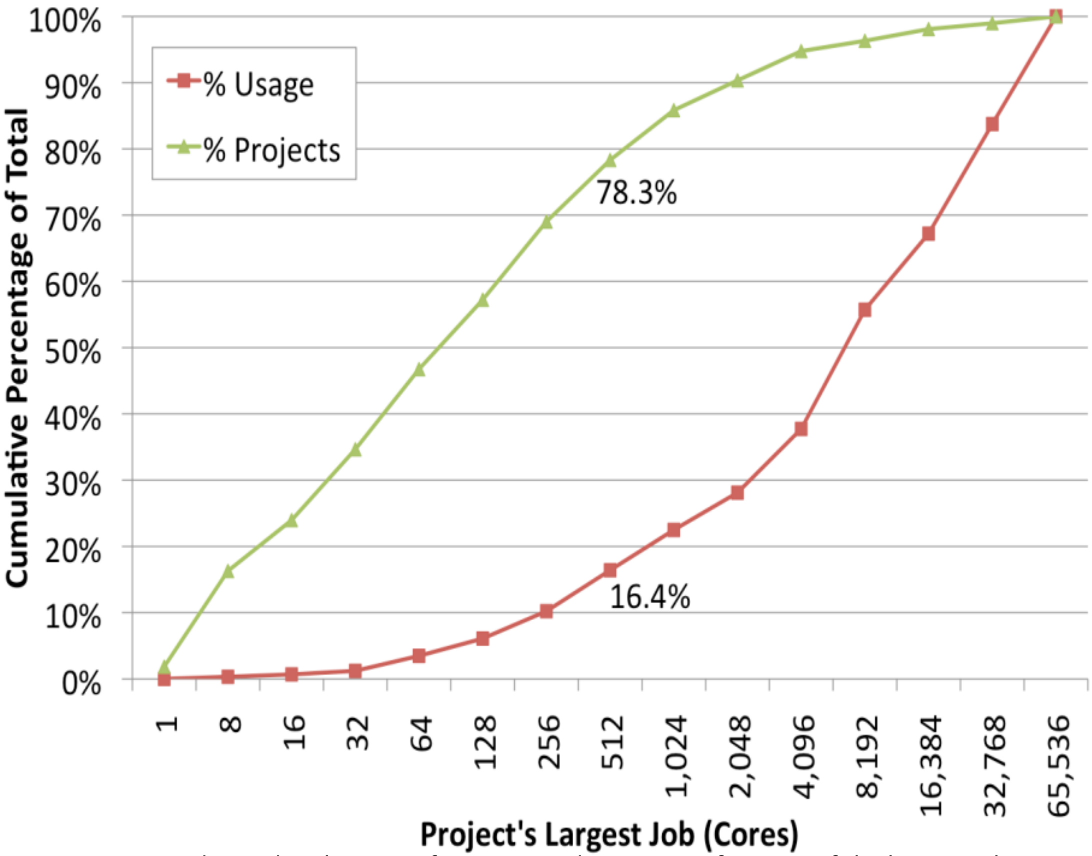
- Will try to present in the context of first three primarily NOWLAB system software (MVAPICH) powered NSF funded production HPC machines at SDSC – over a decade
  - Trestles: 2011 – 2014
  - Gordon: (2009) 2012 – 2017
  - Comet: 2015 - 2021

2. Trestles at SDSC: 2011 – 2014

A high-productivity HPC system targeted to modest-scale and gateway users

- *Designed for modest scale, high throughput and science gateway jobs*
- *Researchers from diverse areas who need access to a fully supported supercomputer with shorter turnaround times*
- *User requirements for more flexible access modes - enabled pre-emptive on-demand queues for applications which require urgent access in response to unpredictable natural or manmade events*
- *10,368 processor cores, a peak speed of 100 teraflop/s, 20 terabytes memory, and 39 terabytes of flash memory (pioneering use of flash)*
- *Large memory (64 GB) and core count (32) per node*
- *Local flash drives available as fast scratch space*

The Majority of TeraGrid/XD Projects Had Modest-Scale Resource Needs

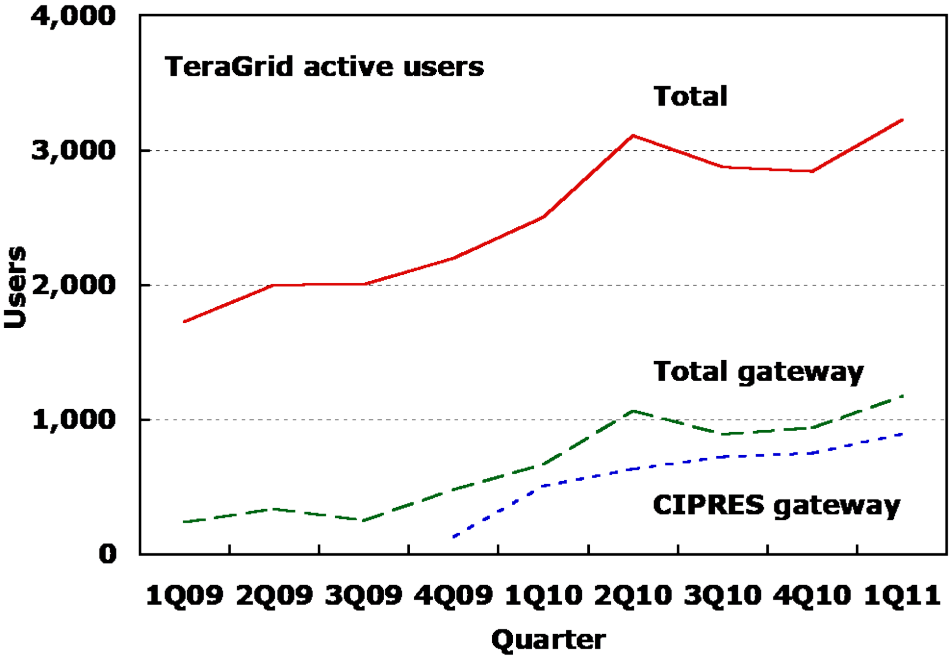


Exceedance distributions of projects and usage as a function of the largest job (core count) run by a project over a full year (FY2009)

- *"80/20" rule around 512 cores*
  - *~80% of projects only run jobs smaller than this …*
  - *And use <20% of resources*
- *Only ~1% of projects run jobs as large as 16K cores and consume >30% of resources*
- *Many projects/users only need modest-scale jobs/resources*
- *And a modest-size resource can provide the resources for a large number of these projects/users*

Trestles Targeted to Modest-Scale Users and Gateway Projects

- *Gateways - an emerging usage mode within TeraGrid/XD*
  - *Many more communities*
- *Growth in the number of TeraGrid users is largely driven by gateway users*
- *An effective system can off-load many users/jobs, including gateway users from capability systems … a win-win for everyone*

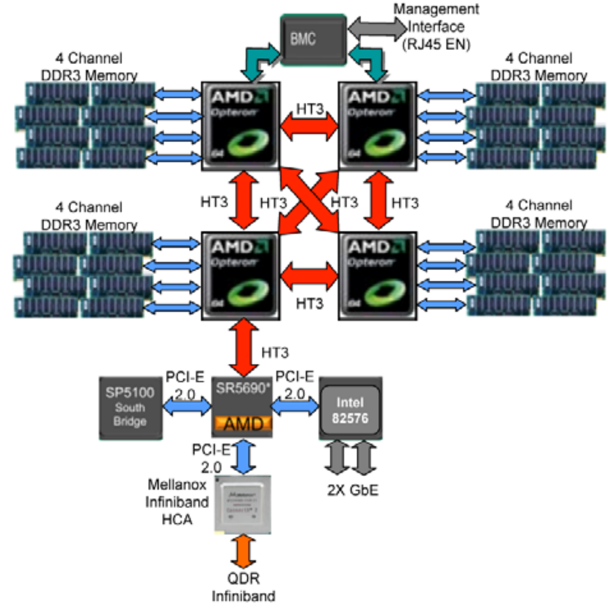


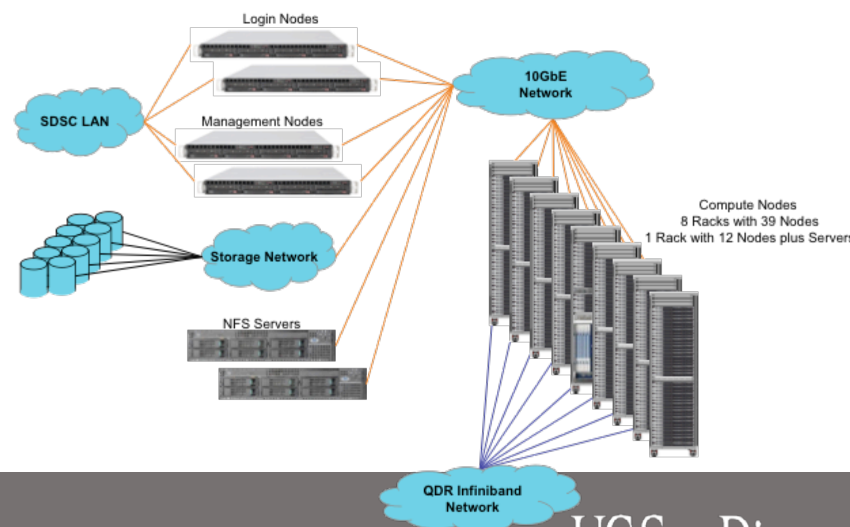
Number of active TeraGrid users by quarter

- *Many users cite queue wait times as primary drawback of TeraGrid/XD systems*
- *For a targeted base of modest-scale users, design the system for productivity, including fast turnaround!*

Trestles is a 100TF system with 324 nodes
(Each node 4 socket*8-core/64GB DRAM/120GB flash, AMD Magny-Cours)

System Component	Configuration
AMD MAGNY-COURS COMPUTE NODE	
Sockets	4
Cores	32
Clock Speed	2.4 GHz
Flop Speed	307 Gflop/s
Memory capacity	64 GB
Memory bandwidth	171 GB/s
STREAM Triad bandwidth	100 GB/s
Flash memory (SSD)	120 GB
FULL SYSTEM	
Total compute nodes	324
Total compute cores	10,368
Peak performance	100 Tflop/s
Total memory	20.7 TB
Total memory bandwidth	55.4 TB/s
Total flash memory	39 TB
QDR INFINIBAND INTERCONNECT	
Topology	Fat tree
Link bandwidth	8 GB/s (bidirectional)
Peak bisection bandwidth	5.2 TB/s (bidirectional)
MPI latency	1.3 us
DISK I/O SUBSYSTEM	
File systems	NFS, Lustre
Storage capacity (usable)	150 TB: Dec 2010 2PB : August 2011 4PB: July 2012
I/O bandwidth	50 GB/s





Trestles Focused on Productivity of its Users Rather than System Utilization

- *The system with a different focus than has been typical of TeraGrid/XD systems*
- Short queue waits are key to productivity
  - Primary system metric is expansion factor = 1 + (wait time/run time)
- Long-running job queues (48 hours std, up to 2 weeks)
- Shared nodes for interactive, accessible computing
- User-settable advance reservations
- Automatic on-demand access for urgent applications
- Robust suite of applications software

Trestles – Actively managed the system to achieve its objectives

- *Target modest-scale users*
  - Limit job size to 1024 cores (32 nodes)
- *Serve a large number of users*
  - Cap allocation per project at 1.5M SUs/year (~2.5% of annual total)
  - Gateways are an exception because they represent large # of users
- *Maintain fast turnaround time*
  - Allocate ~70% of the theoretically available SUs (may be revised as we collect data)
  - Limiting projects to small fractional allocations also should reduce queue waits
  - Configure queues and scheduler to manage to short waits and lower expansion factors
- *Be responsive to user's requirements*
  - Robust software suite
  - Unique capabilities like on-demand access and user-settable reservations
- *Bring in new users/communities*
  - Welcome small jobs/allocations, start-up requests up to 50,000 SUs, gateway-friendly

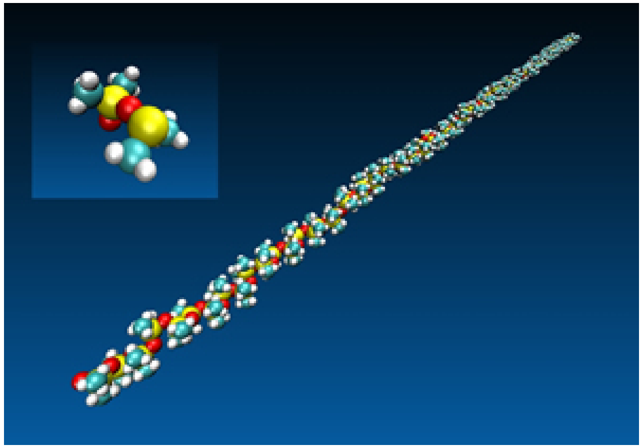
Queue Structure and Scheduler Policies

- Torque resource manager, Catalina external scheduler
- Limit of 32 nodes (1K cores) on all jobs
- Default job time limit 48 hrs, but allowed up to 2 weeks
- Nodes can be requested as exclusive or shared
- Node reservations ensured access for shorter jobs
  - 32 nodes for jobs <48 hrs, 2 nodes <30 min, 3 nodes for shared jobs
- Users can make their own reservations for nodes at specific times to ensure access and enable workflows
  - Individual reservations default to a limit of 2 reservations per allocations account, each with at most 4 nodes, each with at most 4 hours duration
  - Policy limit on user-settable reservations during any 24 hour period of 32 node-hours
- Approved users also have on-demand access for urgent computing

We managed to expansion factors as well as utilization

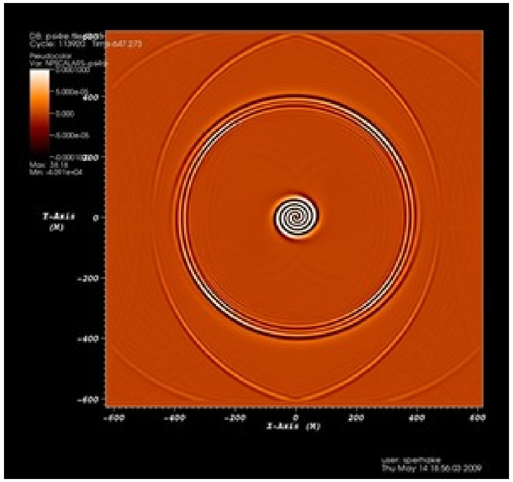
- Utilization – core-hours used/core-hours available - has always been a key metric *for HPC system operators*
- But productivity *for users* depends on wait time
- Expansion factor used to compare waits for different length jobs
  - "Scheduler-based" exp factor = 1 + wait time/requested time
  - "User-based" expansion factor = 1 + wait time/run time
- Objective is to keep expansion factors near unity while maintaining good utilization
  - Premise is that with a modest decrease in utilization, can achieve a significant improvement in turnaround => more productivity and scientific impact
- Active process to minimize user wait time while maximizing utilization

Trestles – couple of science highlights



Simulation shows atomic structure of a chain of polydimethylsiloxane (PDMS) a silicon-based polymer widely used in thermal management – microelectronics . T. Lou, MIT; April 2011, Journal of Applied Physics

- *Large number of smaller jobs simultaneously*
- *Larger simulation of 512 cores – tens of thousands of atoms*
- *Long jobs – running for ~two weeks*
- *I/O intensive first principle – benefitted from local flash*
- *64 GB – large memory for memory demanding FP*



Gravitational wave ripples generated during a high-energy collision of two black holes shot at each other at ~75% speed of light.. X and Y measures the horizontal/vertical distances from the center of mass in unites of the black hole's radius. U. Sperhake, CalTech. May 2011, Physical Review D

- *Used 100s of cores*
- *Total of 100s of GBs of memory*

Trestles Summary

- Modest-scale & gateway users were an evolving important user base
- Monitored and managed Trestles allocations, scheduler and queues to optimize turnaround time AND utilization
  - Complex coupling between utilization and expansion factors – depends on job duration (and size), user workflow, etc.
  - Tuned and modified Catalina scheduler as we gained experience
  - Phases where utilization/expansion factors were high/high and low/low, but also high/low (good) and low/high
  - Allocating ~70% of theoretically available cycles was the primary knob we could turn for tuning utilization, but it's hard to gauge impact as there is a lot of variability in day-to-day utilization and expansion factors
- Pre-emptive on-demand capability in production, with associated "run-at-risk" queue
- Also had user-settable reservations

Where is Trestles (2011 – 2014) now ? Since 2015 it is at ..

UNIVERSITY OF ARKANSAS

Research & Innovation
Arkansas High Performance Computing Center

Search this site

ABOUT RESEARCH RESOURCES SUPPORT TRAINING

Resources

Computational Resources
External Resources

STAY CONNECTED

CITE HPC RESOURCES IN YOUR RESEARCH

Click to Copy

This research is supported by the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission.

Resources

U of A / AHPCC / Resources

High Performance Computing is the resource companion for computational science which is now considered the third leg of science along with experimental and theoretical science.

HPC is for anyone who needs to solve very large numerical problems, process large data sets, or perform advanced simulations. Information technology, and high performance computing in particular, are essential tools in modern research and enable discovery in many disciplines.

Computationally intensive research has dominated the physical sciences such as physics and chemistry for decades and is now becoming prominent in biology, sociology, medical research, agriculture, and a growing list of other fields.

COMPUTATIONAL RESOURCES

Trestles

Trestles, an NSF XSEDE resource was acquired from San Diego Supercomputer Center and from the National Science Foundation in May of 2015. The Trestles cluster comprises 256 compute nodes, each with quad octa-core AMD Opteron 2.4 GHz 6194 processors at 2.4 GHz, 64 GB of memory, and 120 GB of SSD local disk. Trestles is interconnected with a 324-port QDR 40 Gbps nonblocking Mellanox InfiniBand switch, and is connected to shared Lustre file systems with 16TB of scratch space and 350TB of main storage.

Razor

3. Gordon at SDSC: (2009) 2012-2017

An innovative data intensive supercomputer

Designed for data and memory intensive applications that don't run well on traditional distributed memory machines

- *Large shared memory requirements*
- *Serial or threaded (OpenMP, Pthreads)*
- *Limited scalability*
- *High performance data base applications*
- *Random I/O combined with very large data sets*
- *Large scratch files*

Gordon – An Innovative Data Intensive Supercomputer

- Designed to accelerate access to massive amounts of data in areas of genomics, earth science, engineering, medicine, and others
- Emphasizes memory and IO over FLOPS.
- Appro (later Cray) integrated 1,024 node Sandy Bridge cluster
- 300 TB of high performance Intel flash
- Large memory supernodes via vSMP Foundation from ScaleMP
- 3D torus interconnect from Mellanox
- Production operation - February 2012
- Funded by the NSF and provided through the NSF Extreme Science and Engineering Discovery Environment program (XSEDE)

Gordon System Specification

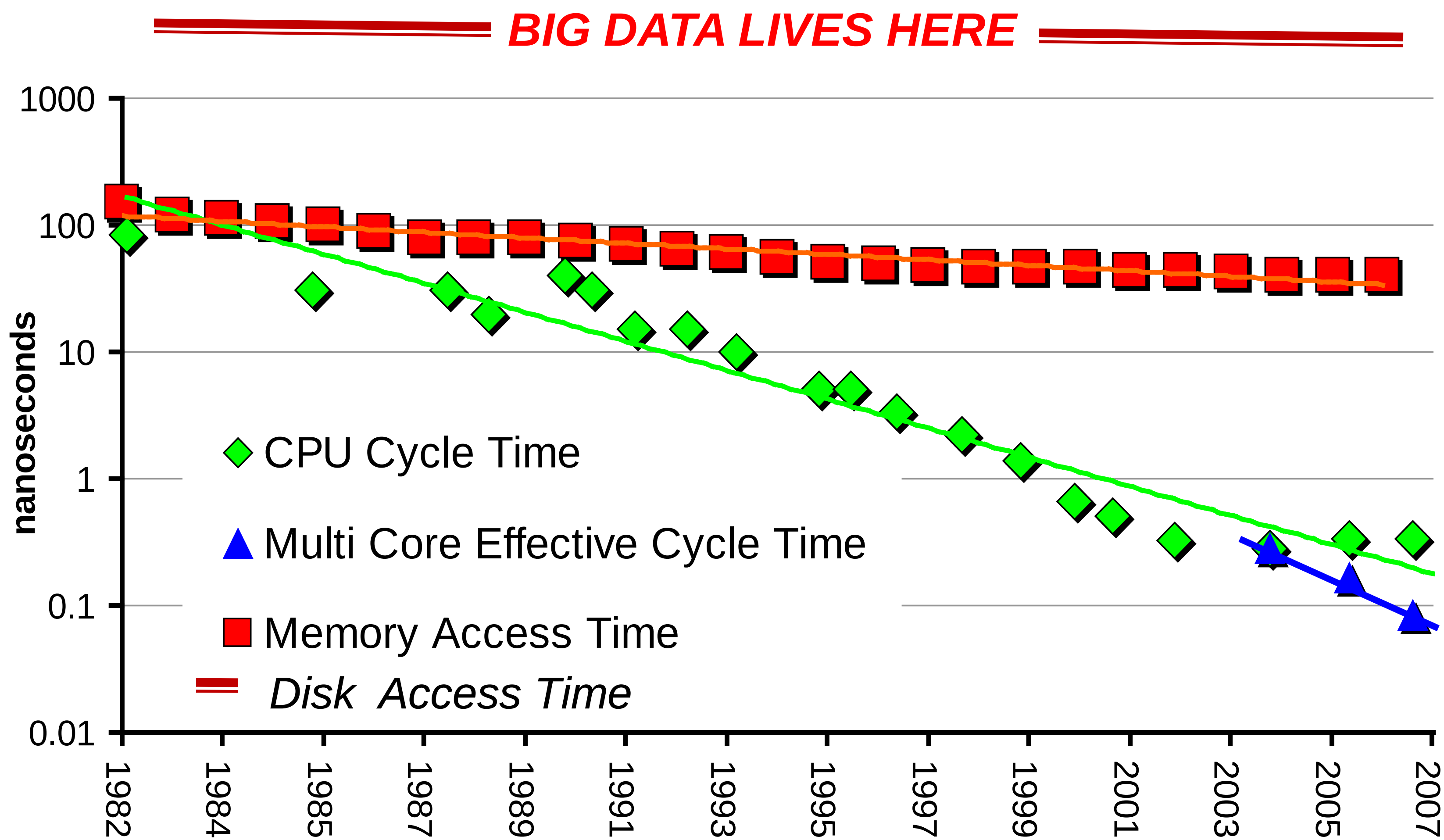
INTEL SANDY BRIDGE COMPUTE NODE	
Sockets, Cores	2 / 16
Clock speed	2.6
DRAM capacity	64 GB
SSD	80 GB
INTEL FLASH I/O NODE	
NAND flash SSD drives	16
SSD capacity per drive/Capacity per node/total	300 GB / 4.8 TB / 300 TB
Flash bandwidth per drive (read/write) **IOPS**	270 MB/s / 210 MB/s 38,000 / 2,300
SMP SUPER-NODE	
Compute nodes	32
I/O nodes	2
Addressable DRAM	2 TB
Addressable memory including flash	12TB
FULL SYSTEM	
Compute Nodes, Cores	1,024 / 16,394
Peak performance	341TF
Aggregate memory	64 TB
INFINIBAND INTERCONNECT	
Aggregate torus BW	9.2 TB/s
Type	Dual-Rail QDR InfiniBand
Link Bandwidth	8 GB/s (bidirectional)
Latency (min-max)	1.25 µs – 2.5 µs
DATA OASIS LUSTRE FILE SYSTEM	
Total storage	4.5 PB (raw)
I/O bandwidth	100 GB/s

Gordon Design: Two Driving Ideas

- Observation #1: Data keeps getting further away from processor cores ("red shift")
  - Do we need a new level in the memory hierarchy?

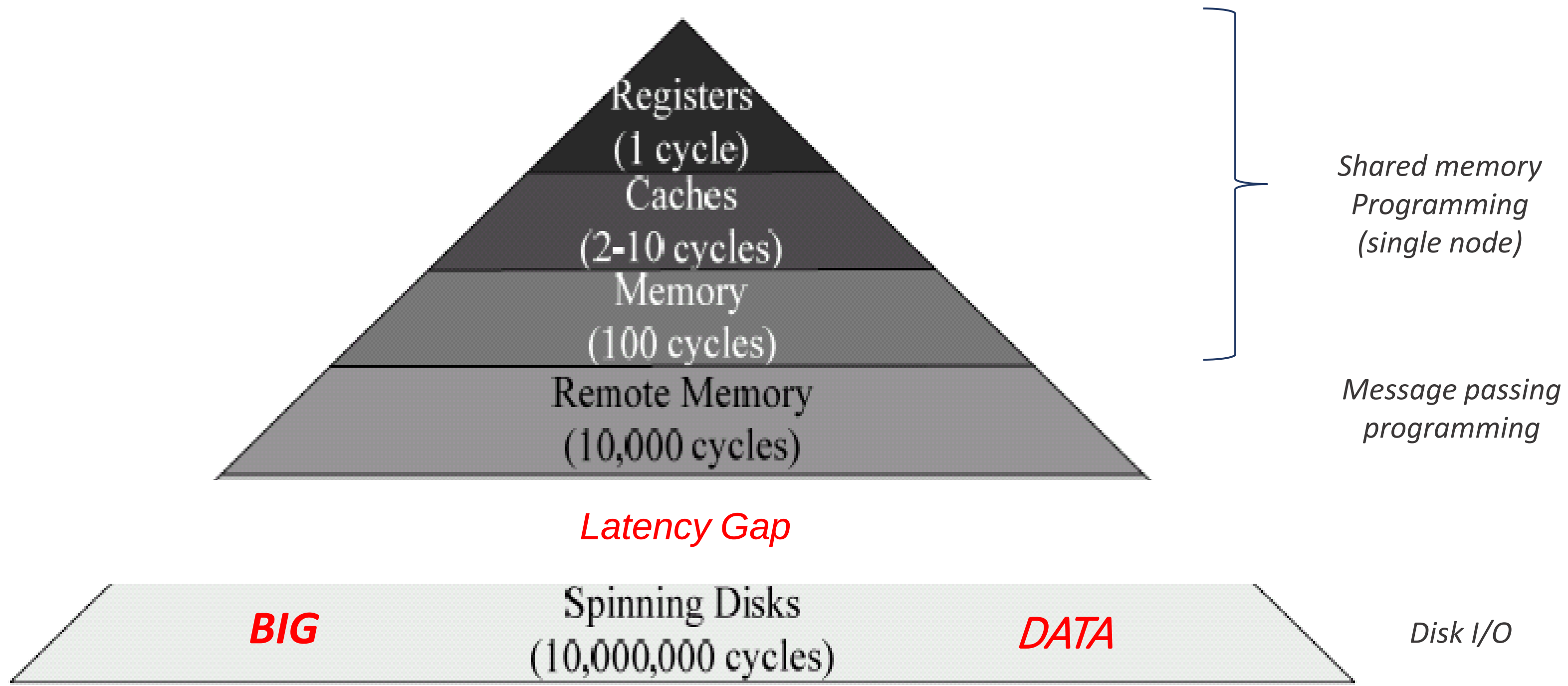
- Observation #2: Many data-intensive applications are serial and difficult to parallelize
  - Would a large, shared memory machine be better from the standpoint of researcher productivity for some of these?
  - → Rapid prototyping of new approaches to data analysis

Red Shift: Data keeps moving further away from the CPU with every turn of Moore's Law

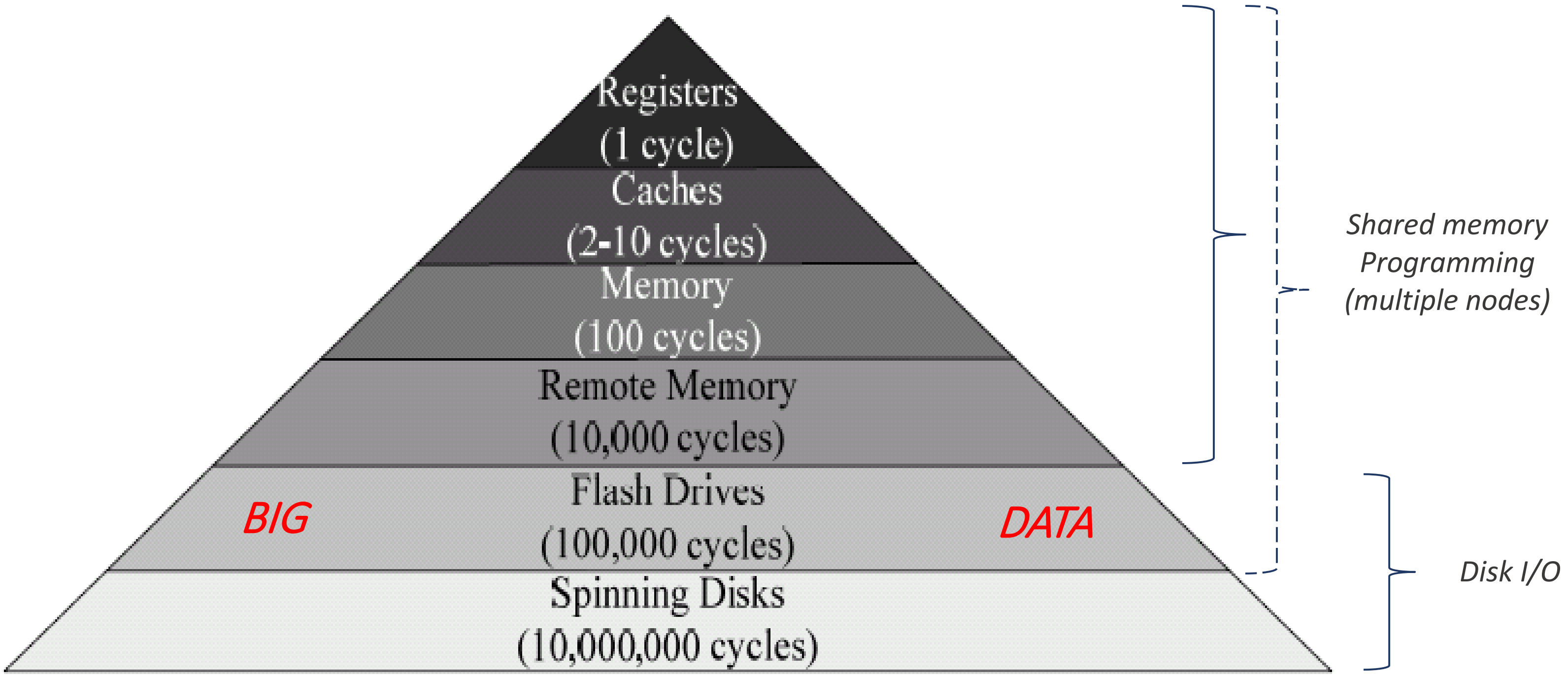


Source Dean Klein, Micron

The Memory Hierarchy of a Typical Supercomputer



The Memory Hierarchy of Gordon

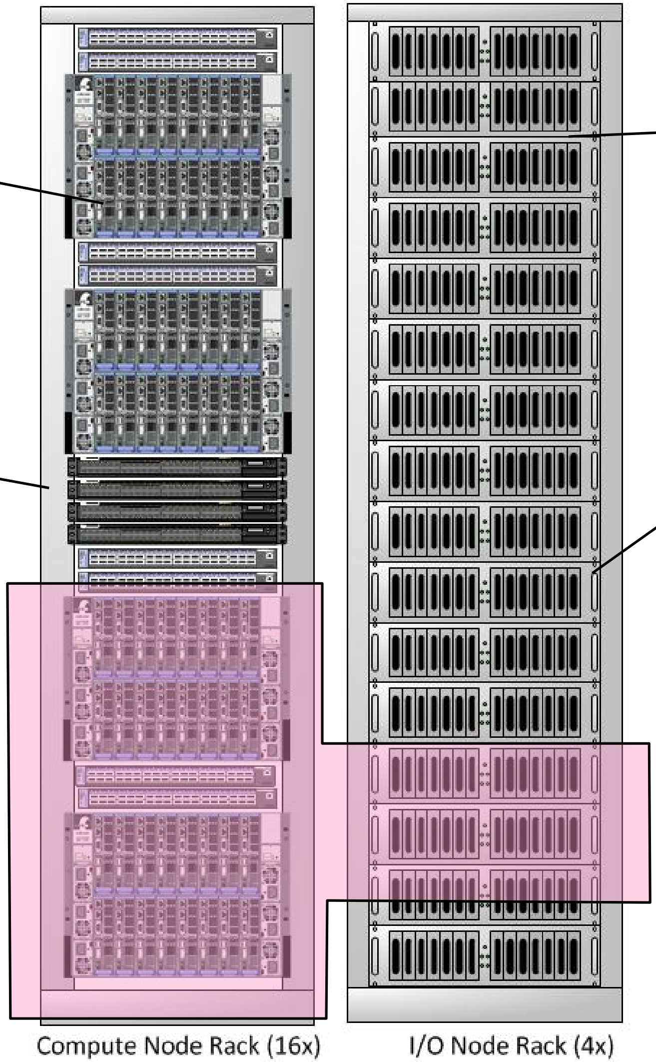


Gordon Design Highlights

- *1,024 2S Xeon E5 (Sandy Bridge) nodes*
- *16 cores, 64 GB/node*
- Intel Jefferson Pass *mobo*
- *PCI Gen3*

- *3D Torus*
- Dual rail QDR

- *Large Memory vSMP Supernodes*
- 2TB DRAM
- *10 TB Flash*

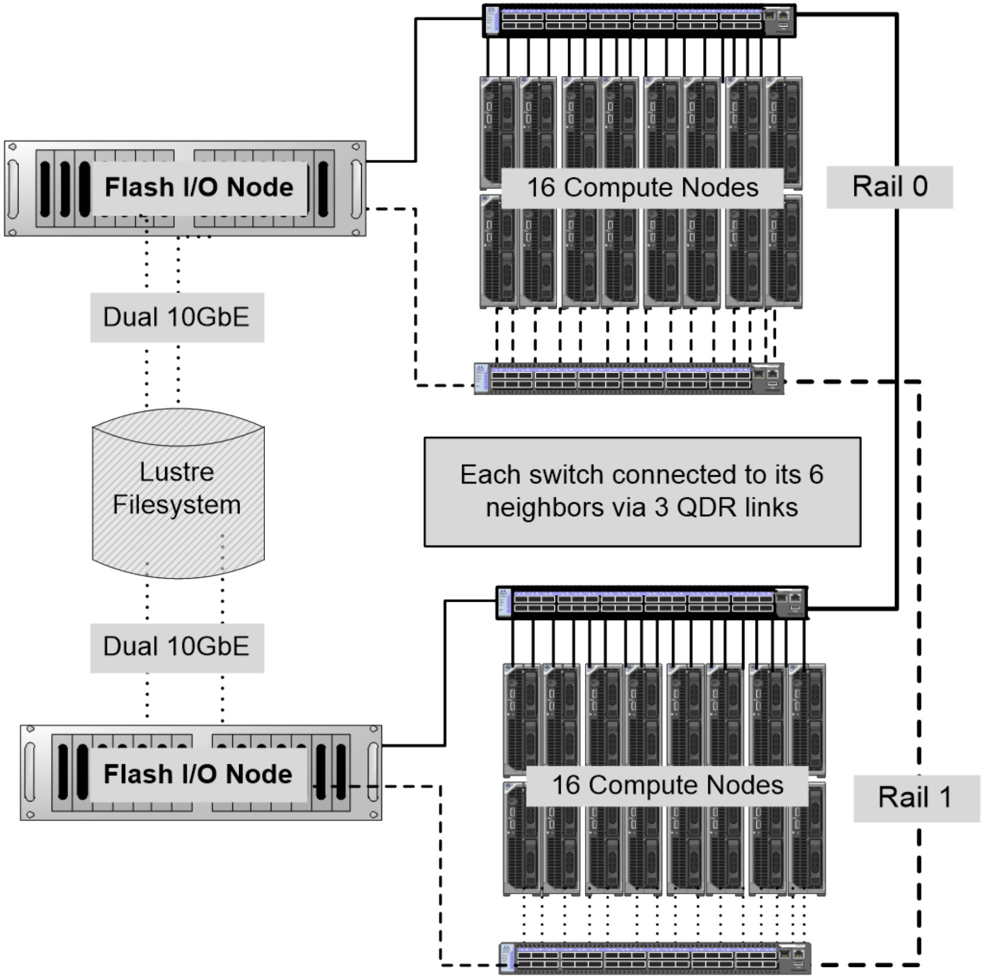


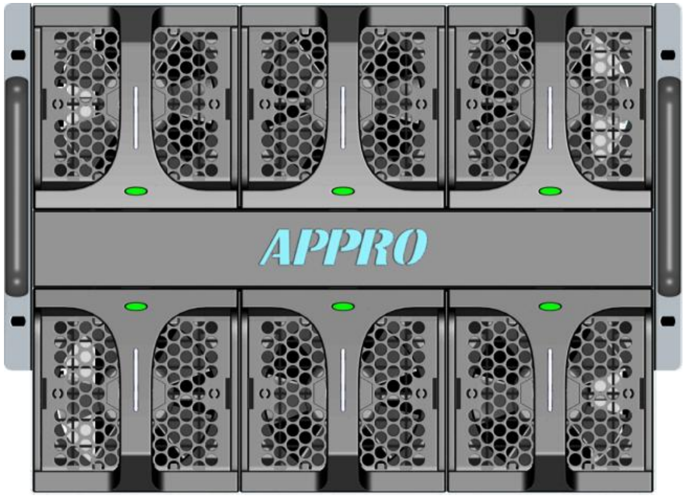
- 300 GB Intel 710 eMLC SSDs
- *300 TB aggregate*

- *64, 2S Westmere I/O nodes*
- 12 core, *48 GB/node*
- *4 LSI controllers*
- *16 SSDs*
- *Dual 10GbE*
- SuperMicro mobo
- *PCI Gen2*

"Data Oasis"
Lustre PFS
100 GB/sec, 4 PB

Subrack and Cabling Design Detail

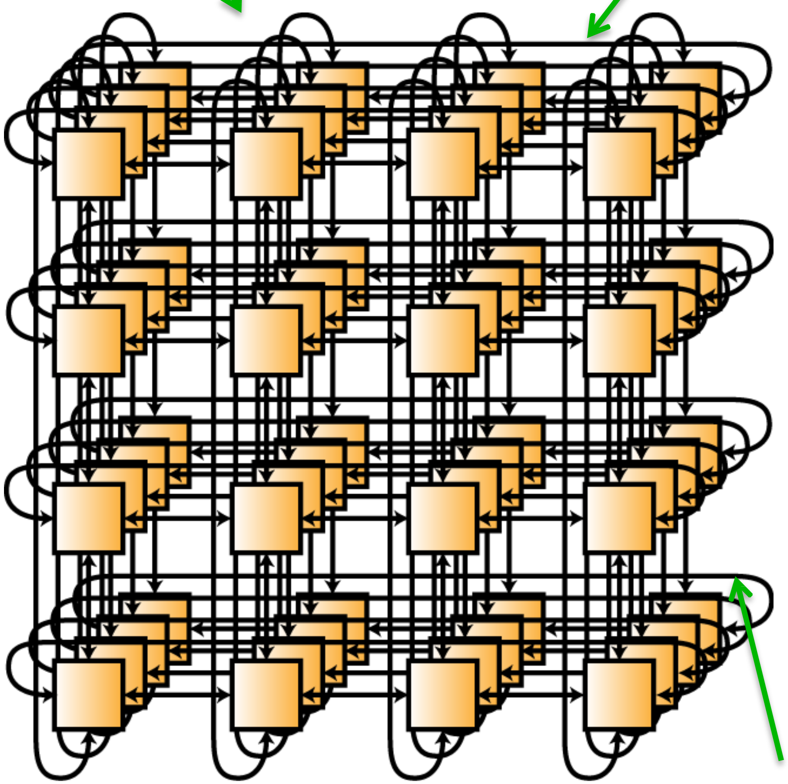




Gordon Architecture:3D Torus of Switches

Each node is switch

Connectivity wraps around

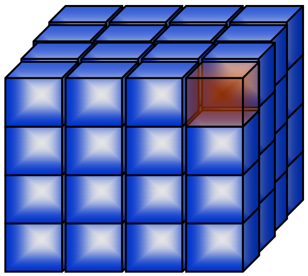


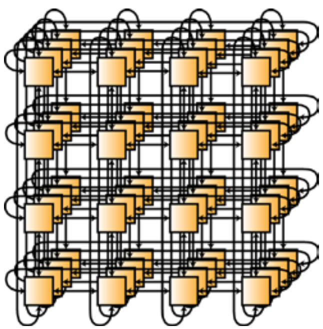
- *Switches are connected in 4x4x4 3D torus*
- *Linearly expandable*
- *Short Cables- Fiber Optic cables generally not required*
- *Lower Cost :40% as many switches, 25% to 50% fewer cables*
- *Works well for localized communication*
- *Fault Tolerant within the mesh with 2QoS Alternate Routing*
- *Fault Tolerant with Dual-Rails for all routing algorithms*
- *Two rails – i.e., two complete tori with 64switch nodes in each torus*
- *Maximum of 6 hops*

Switches are interconnected by 3 links in each +/- x, y, z direction

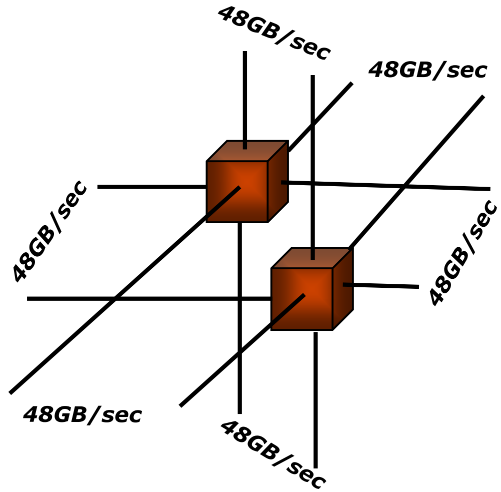
Gordon 3D Torus Interconnect Fabric
4x4x4 3D Torus Topology

4X4X4 Mesh
Ends are folded on all three
Dimensions to form a 3DTorus

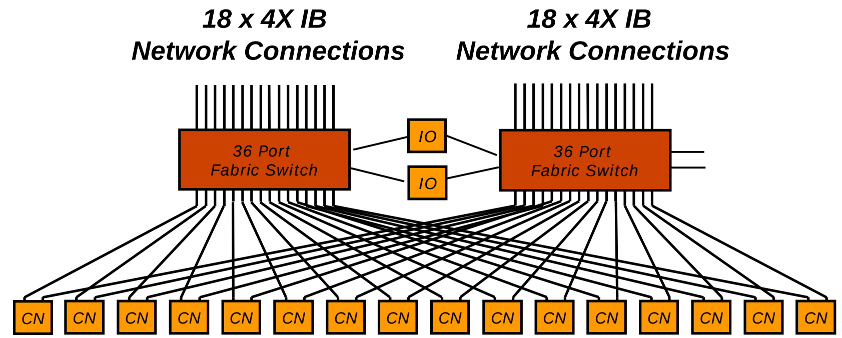




Dual-Rail Network
increased Bandwidth & Redundancy



Single Connection to each Network
16 Compute Nodes, 2 IO Nodes



Gordon Systems Software Stack

Cluster management	Rocks 5.4.3
Operating System	CentOS5.6 – modified for AVX
InfiniBand	OFED 1.5.3 Mellanox subnet manager
MPI	MVAPICH2 (Native) MPICH2 (vSMP)
Shared Memory	vSMP Foundation v4
Flash	iSCSI over RDMA (iSER) Target daemon (tgtd) XFS, OCFS, et al
User Environment	Modules; Rocks Rolls
Parallel File System	Lustre 1.8.7
Job scheduling	Torque (PBS), Catalina Local enhancements for topology aware scheduling

MVAPICH2 on Gordon

- *MVAPICH2 {version is 1.9] was the default MPI implementation on Gordon. (now version 2.1)*

- *Compiled with --enable-3dtorus-support flag. Multi-rail support.*

- *LIMIC2 [Version on system was 0.5.6]*

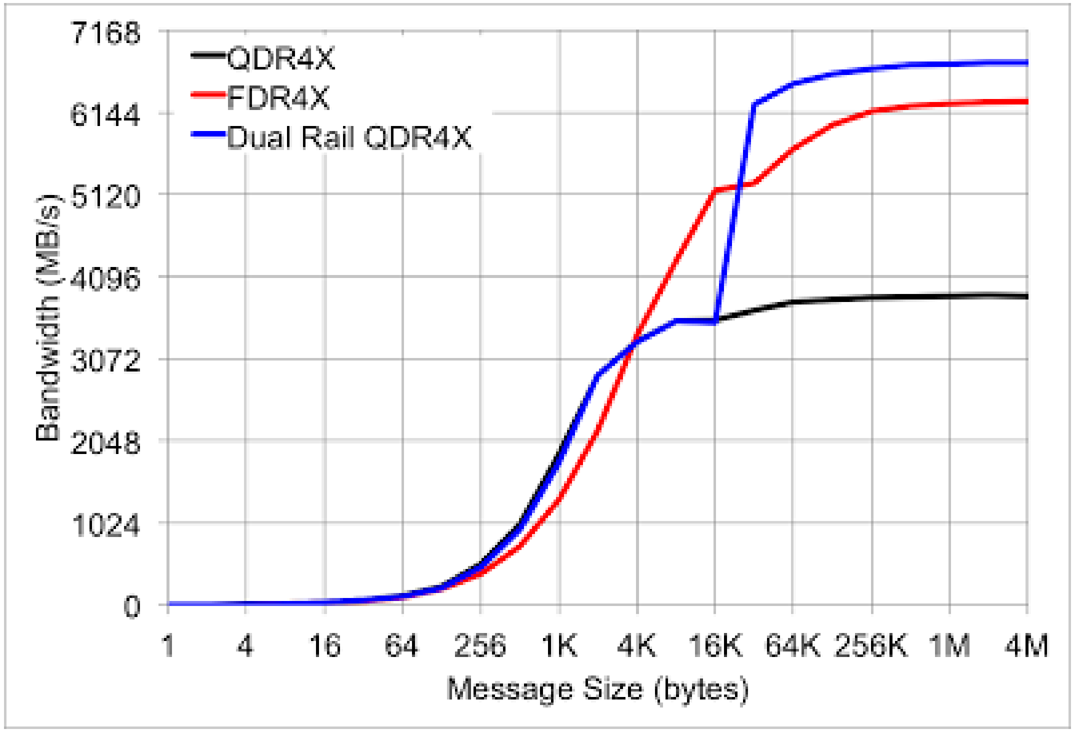
- *SSDs on Gordon are in I/O nodes. Exported to the compute nodes via iSER. Rail 1 (mlx4_1) is used for this part.*

- *I/O nodes also serve as lustre routers. Again I/O traffic is going on rail 1 (mlx4_1).*

- *Given I/O traffic, both to lustre and SSDs (local scratch) can saturate rail 1, default recommendation is to run MVAPICH2 with one rail [MV2_IBA_HCA=mlx4_0, MV2_NUM_HCAS=1]*

Dual Rail QDR vs FDR OSU Bandwidth Test

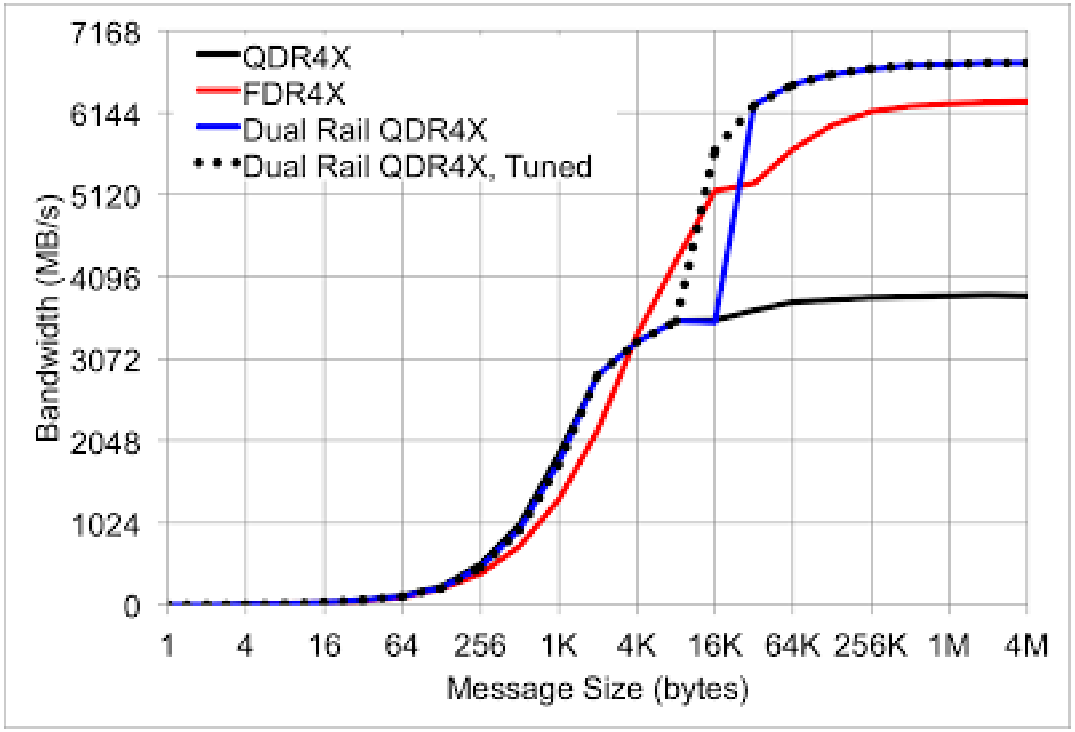
- ***MVAPICH2 out of the box without any tuning***



**Tests done by Glenn Lockwood (then at SDSC; now NERSC)*

Dual Rail QDR vs FDR OSU Bandwidth Test

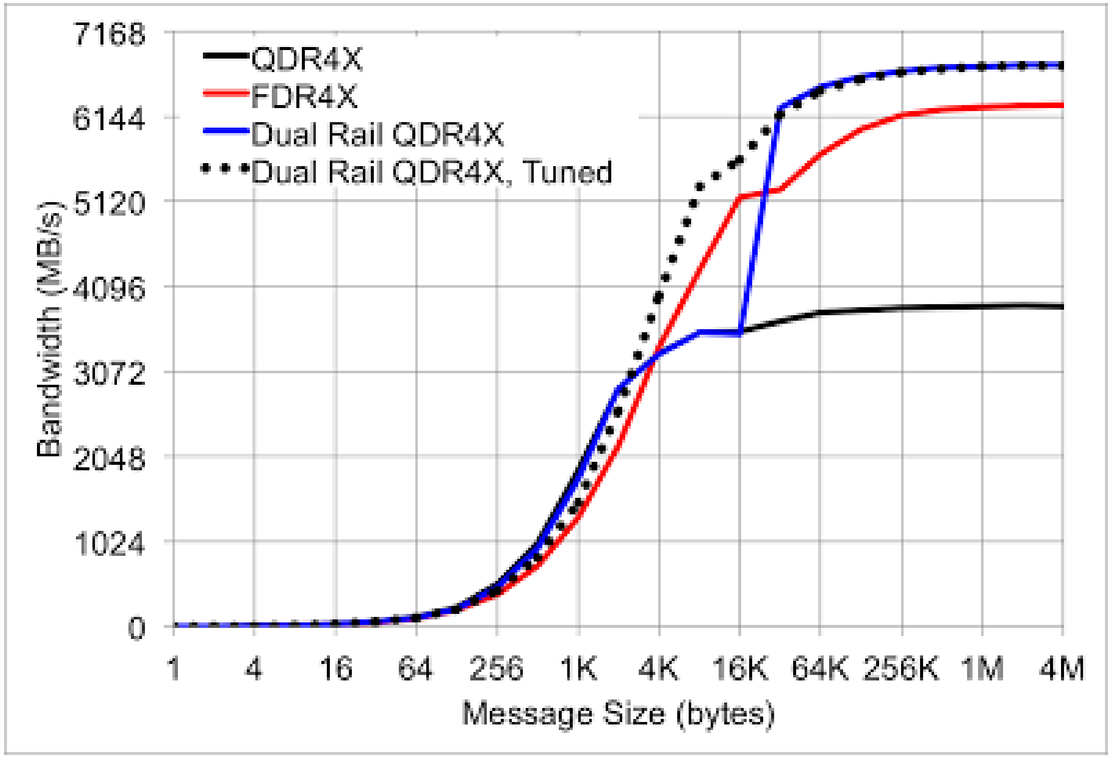
- ***MV2_RAIL_SHARING_LARGE_MSG_THRESHOLD=8k***



***Tests done by Glenn Lockwood (then at SDSC; now NERSC)**

Dual Rail QDR vs FDR OSU Bandwidth Test

- ***MV2_SM_SCHEDULING=ROUND_ROBIN***
- ***In new version this is MV2_RAIL_SHARING_POLICY, default***



**Tests done by Glenn Lockwood (then at SDSC; now NERSC)*

Production Gordon stack featured MVAPICH2 w/ --enable-3dtorus- support flag and dual rail support

- *Dual rail QDR performance competitive with FDR performance.*
  - *MVAPICH2 environment variables such as MV2_RAIL_SHARING_LARGE_MSG_THRESHOLD and MV2_RAIL_SHARING_POLICY (earlier MV2_SM_SCHEDULING) can be used to tune performance.*

- *Gordon has oversubscription of switch to switch links. Spreading tasks to reduce contention can improve performance.*

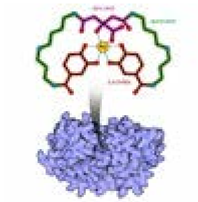
- *Big Thanks to Dr. Panda's group! Gordon was the first production dual rail InfiniBand 3-D torus machine and the MVAPICH2 deployment was flawless out of the box.*

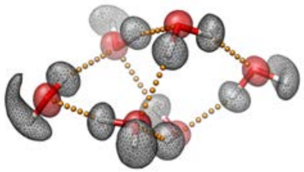
3D Torus Experiences

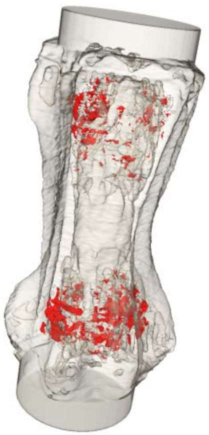
- First dual rail, 3D torus deployed (that we're aware of)
- Early engineering work on a 4x4x2 3D torus with Appro and Mellanox was an important risk mitigator
- Low-level performance benchmarks are excellent
  - 1.44 – 2.5 us latency
  - 3.2-3.8 GB/s link bandwidth (half duplex, single rail)
- Operations was a non-issue
  - Running 2 subnet managers (SM) – one for each rail
  - Have had zero failures of the SM
  - No issues with vSMP operations. Switches participate in both native and vSMP environments.
- Zero tolerance for errors in cabling
- Deployed configuration
  - Rail 0 is user MPI traffic
  - Rail 1 is for I/O traffic to I/O nodes (both flash and Lustre)
  - Research work with DK Panda's team to fully leverage the capabilities of the torus for failover, bandwidth.

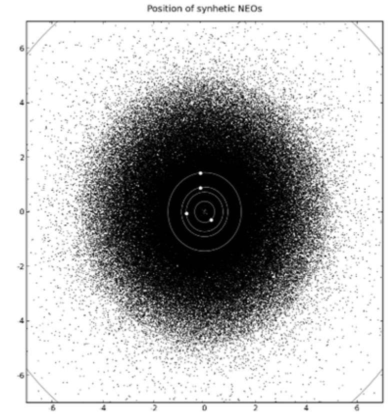
Gordon Science Highlights

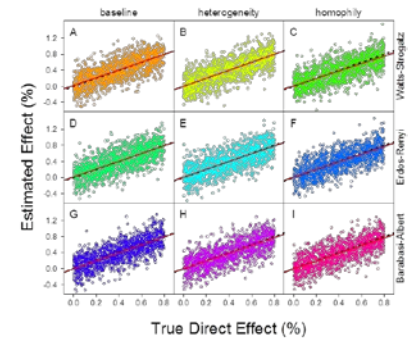
RCSB PDB
PROTEIN DATA BANK

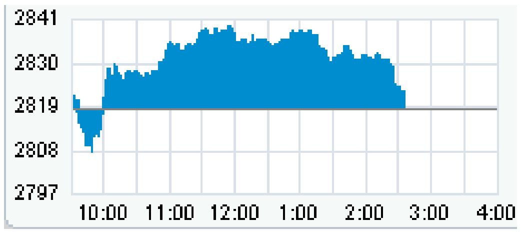




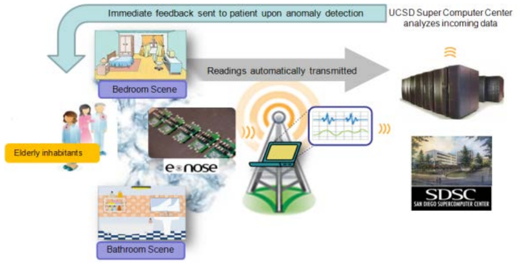








CIPRES



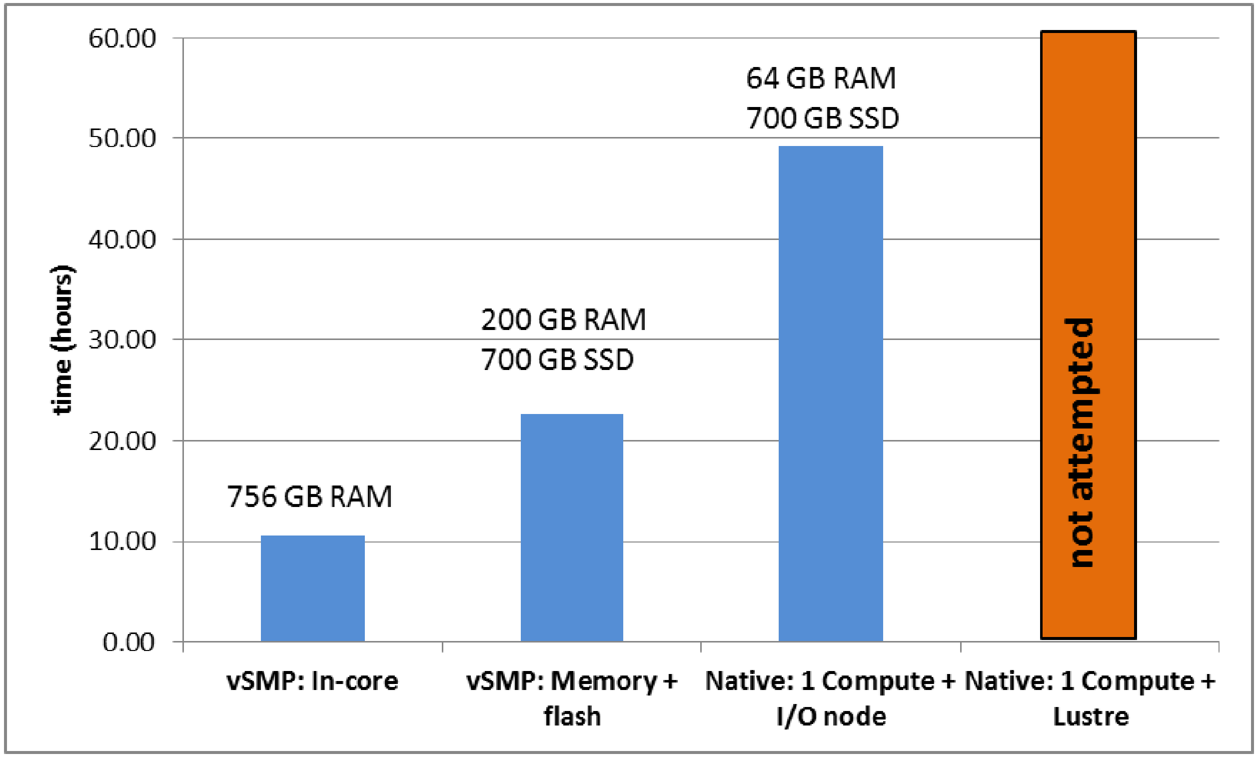
Computational Style Code
Answering the question: Why Gordon?

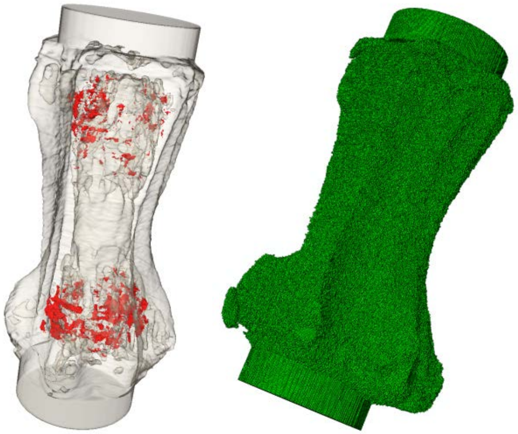
V	*M*	*F*
C	*T*	*L*

V: Uses vSMP
C: Computationally intensive, leverages Sandy Bridge architecture
M:Uses larger Memory/core on Gordon (4GB/core)
T: Threaded
F: Uses Flash
L: Lustre I/O intensive

Axial compression of caudal rat vertebra using Abaqus and vSMP

The goal of the simulations is to analyze how small variances in boundary conditions effect high strain regions in the model. The research goal is to understand the response of trabecular bone to mechanical stimuli. This has relevance for paleontologists to infer habitual locomotion of ancient people and animals, and in treatment strategies for populations with fragile bones such as the elderly.





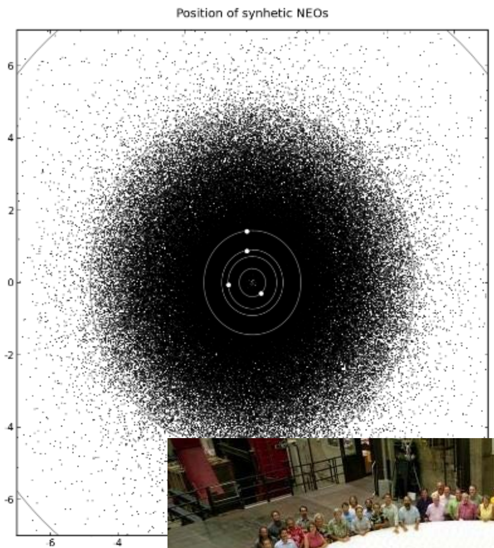
- *5 million quadratic, 8 noded elements*
- *Model created with custom Matlab application that converts 25^3 micro CT images into voxel-based finite element models*

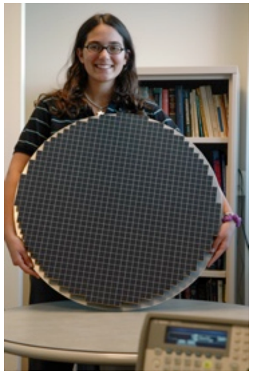
Source: *Matthew Goff, Chris Hernandez. Cornell University. Used by permission. 2012*

V	M	F
C	T	L

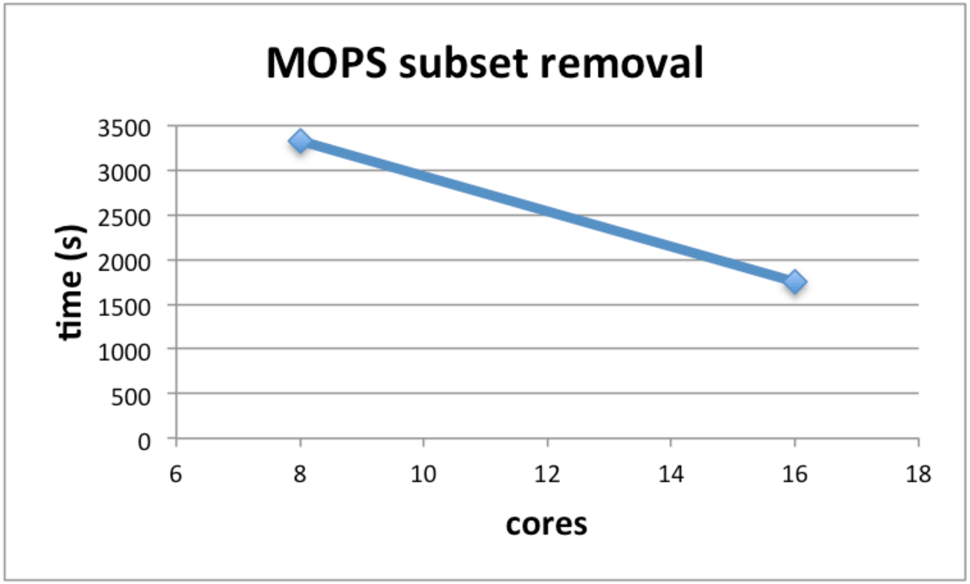
LSST – Moving Object Pipeline System

Images collected by the Large Synoptic Survey Telescope (LSST) will be be processed using the Moving Object Pipeline System (MOPS). Detections from consecutive nights are grouped together into tracks that potentially represent small portions of the asteroids' sky-plane motion





Run time for subset removal algorithm scales almost linearly out to 16 cores



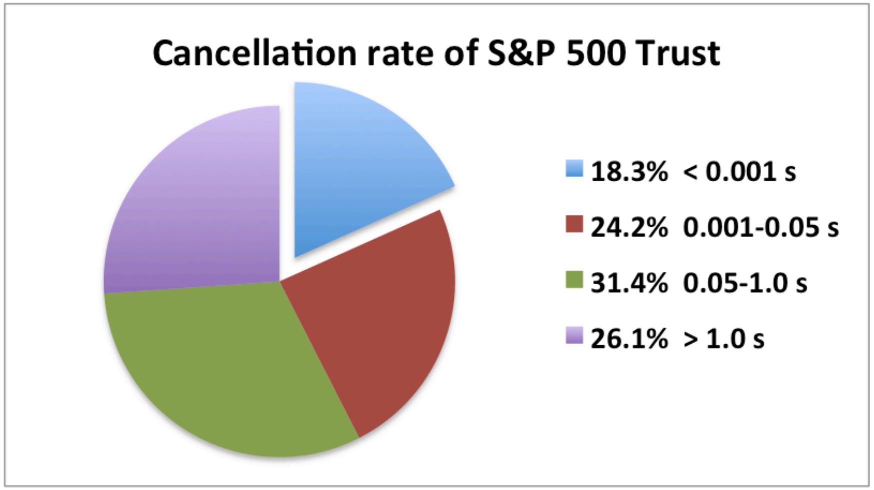
V	M	F
C	T	L

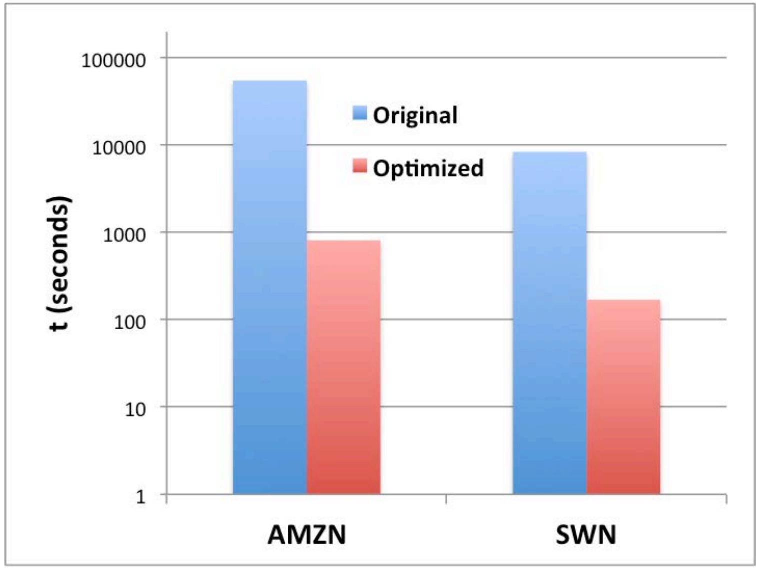
Source: *Jonathan Myers, LSST Used by permission. 6/4/2012*

Impact of high-frequency trading on financial markets

To determine the impact of high-frequency trading activity on financial markets, it is necessary to construct nanosecond resolution limit order books – records of all unexecuted orders to buy/sell stock at a specified price. Analysis provides evidence of quote stuffing: a manipulative practice that involves submitting a large number of orders with immediate cancellation to generate congestion

Time to construct limit order books now under 15 minutes for threaded application using 16 cores on single Gordon compute node



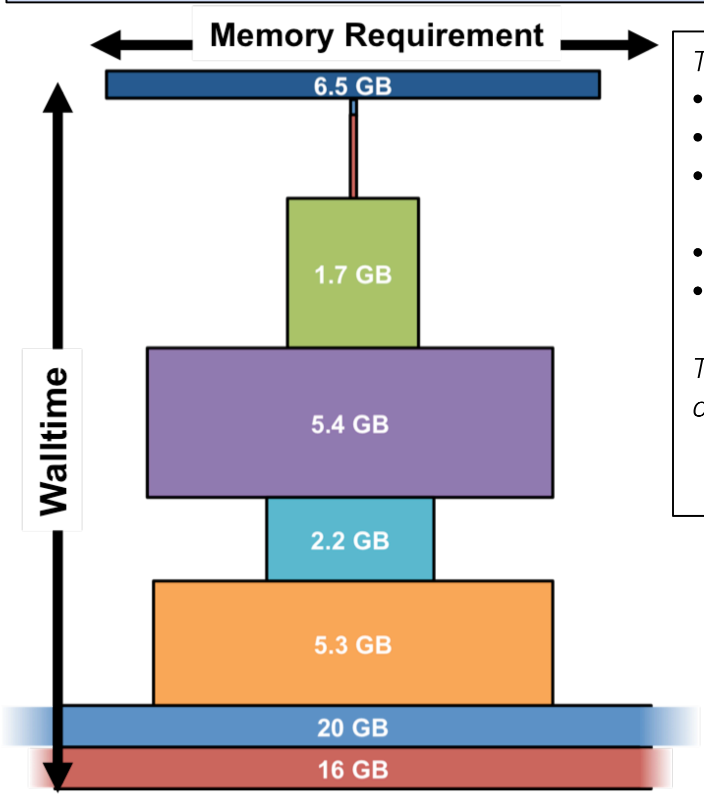


V	M	F
C	T	L

Source: *Mao Ye, Dept. of Finance, U. Illinois. Used by permission. 6/1/2012*

Large-scale pharmacogenomic analysis

Janssen R&D, a Johnson & Johnson company, has been using whole-genome sequencing in clinical trials of new drug therapies to correlate response or non-response with genetic variants. Janssen has partnered with the Scripps Translational Science Institute (STSI) to perform cutting-edge analyses on hundreds of full human genomes which presents many dimensions of data-intensive challenges.

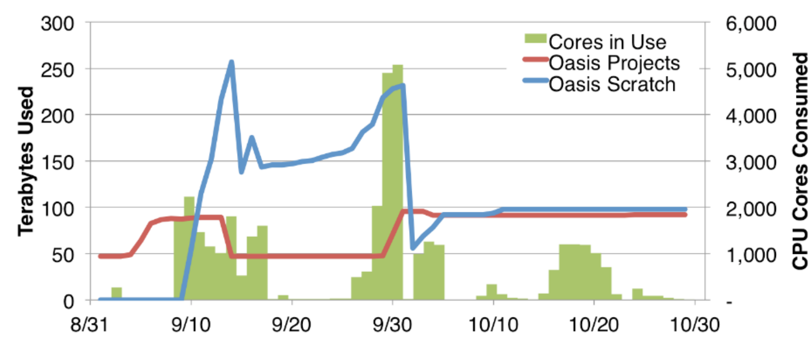


Each step of the 9-stage read-mapping pipeline had very different resource requirements

To analyze 438 human genomes, this project needed

- ***16-threads per node and hundreds of nodes*** *to achieve massive parallelism*
- *at least of* ***40 GB of RAM per node*** *for some pipeline stages*
- ***over 3 TB of flash storage per node*** *via "big flash" nodes at a metadata-IOPS rate not sustainable by Lustre*
- ***over 1.6 TB of input data per node*** *at some pipeline stages*
- ***1 GB/s read rate from Lustre per node***

This project accomplished in 5 weeks on Gordon what would have taken 2.5 years of 24/7 compute on a single, 8-core workstation with 32 GB of RAM.

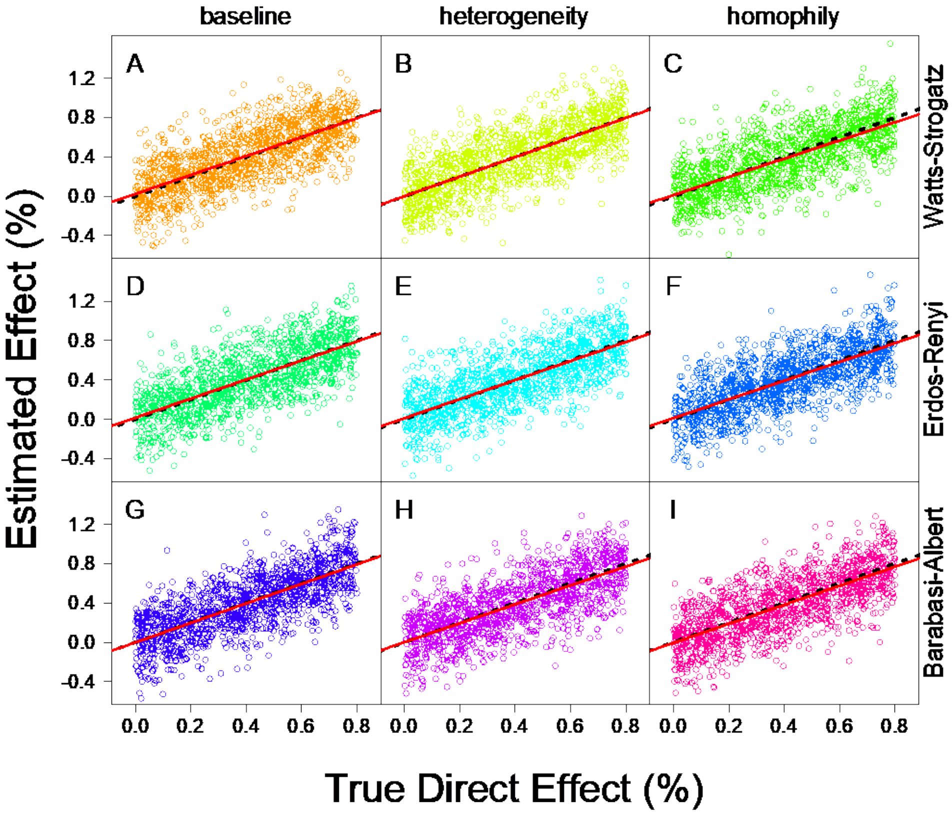


Peak footprint:

- *257 TB of Oasis Scratch*
- *5,000 cores in use (30% of Gordon's total capacity)*

Monte Carlo network permutation simulations

Monte Carlo simulations are run to test whether a network permutation method used for estimating causal effects is biased by structure in the population. The simulation generates a 5,000,000 person network 1,000 times to estimate the parameters of interest



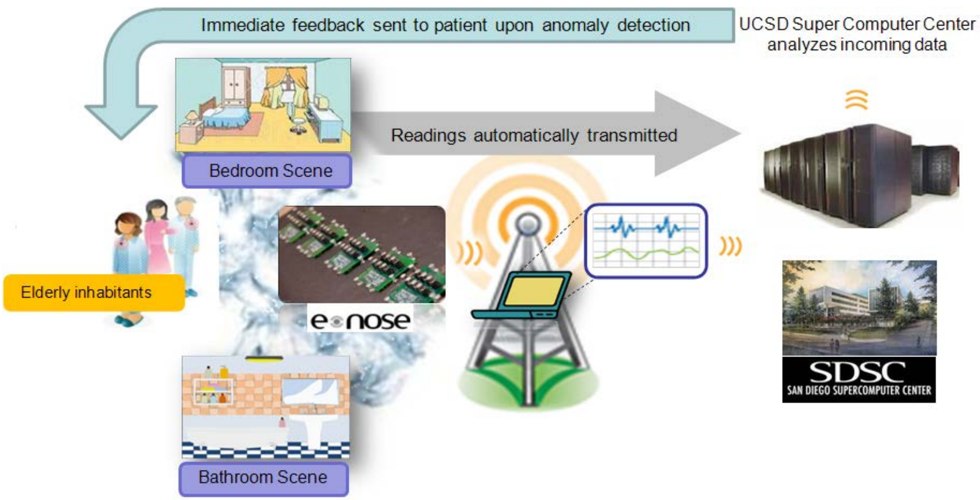
*Simulations performed using R on Gordon. Code was optimized to take advantage of vectorization and reduced run time from **one hour** for 1M node simulation to **one minute** for 5M node run*

V	M	F
C	T	L

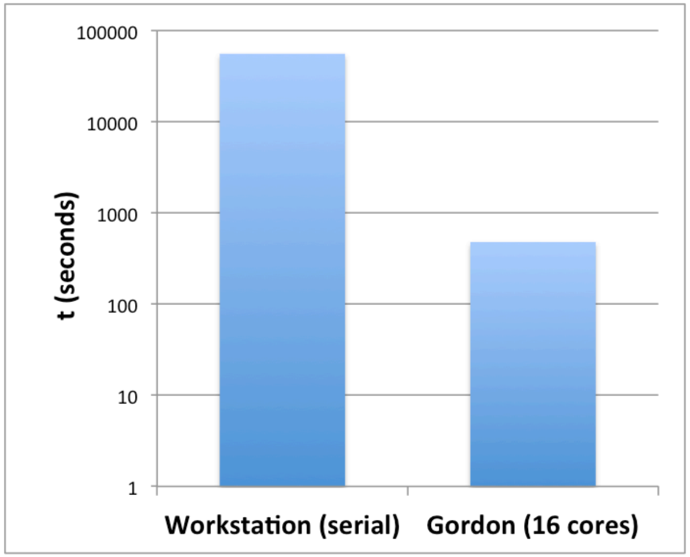
Source: *Chris Fariss, UCSD Dept. of Political Science Used by permission. 6/1/2012*

Classification of sensor time series data

Chemical sensors (e-noses) will be placed in the homes of elderly participants in an effort to continuously and non-intrusively monitor their living environments. Time series classification algorithms will then be applied to the sensor data to detect anomalous behavior that may suggest a change in health status.



After optimizing code, linking Intel's MKL and porting to Gordon, runtime reduced from 15.5 hours to 8 minutes



V	M	F
C	T	L

Source: *Ramon Huerta, UCSD Bio Circuits Institute Used by permission 6/1/2012*

Gordon Summary

- The nature of scientific research - more *data-intensive*, requiring new kinds of high-performance computer architectures and data management systems
- Gordon was (is) an innovated system that addressed a range of challenges associated with data intensive computing.
- A prototype system and significant testing mitigated the challenges in deploying a large memory system like Gordon
- Gordon supported a wide range of applications: large memory, MPI applications, and dedicated I/O node
- Outreach to new user communities was big part of Gordon project
- Productive data intensive computing was done on Gordon

Where is Gordon (2012-2017) now ? Since 2017 it is at..

Home News This Week Subscribe Contact More

UC San Diego News Center

March 14, 2017 | By Jan Zverina

Simons Foundation's Flatiron Institute to Repurpose SDSC's 'Gordon' Supercomputer

Resource to be used for Astrophysics, Biology, Materials Research

SDSC's data-intensive 'Gordon' supercomputer. Photo by Erik Jepsen/UC San Diego Publications

The San Diego Supercomputer Center (SDSC) at the University of California San Diego and the Simons Foundation's Flatiron Institute in New York have reached an agreement under which the majority of SDSC's data-intensive *Gordon* supercomputer will be used by Simons for ongoing research following completion of the system's tenure as a National Science Foundation (NSF) resource on March 31.

Under the agreement, SDSC will provide high-performance computing (HPC) resources and services on *Gordon* for the Flatiron Institute to conduct computationally-based

MEDIA CONTACT

Jan Zverina
858-534-5111
jzverina@sdsc.edu

RELATED LINKS

San Diego Supercomputer Center
Simons Foundation
Flatiron Institute
National Science Foundation

FILED UNDER

Research
Science & Engineering
Faculty
Biological Sciences
Physical Sciences

4. Comet at SDSC – "HPC for the long tail of science"

iPhone panorama photograph of 1 of 2 server rows

Comet Built to Serve the 99%

CHALLENGES OUR PROPOSAL ADDRESSES

- ✓ Attract new users and communities
- ✓ Support diverse applications with complex workflows
- ✓ Ensure responsiveness for thousands of users
- ✓ Transfer, store, analyze, and share massive data sets
- ✓ Integrate with XSEDE

COMET COMPUTE SYSTEM

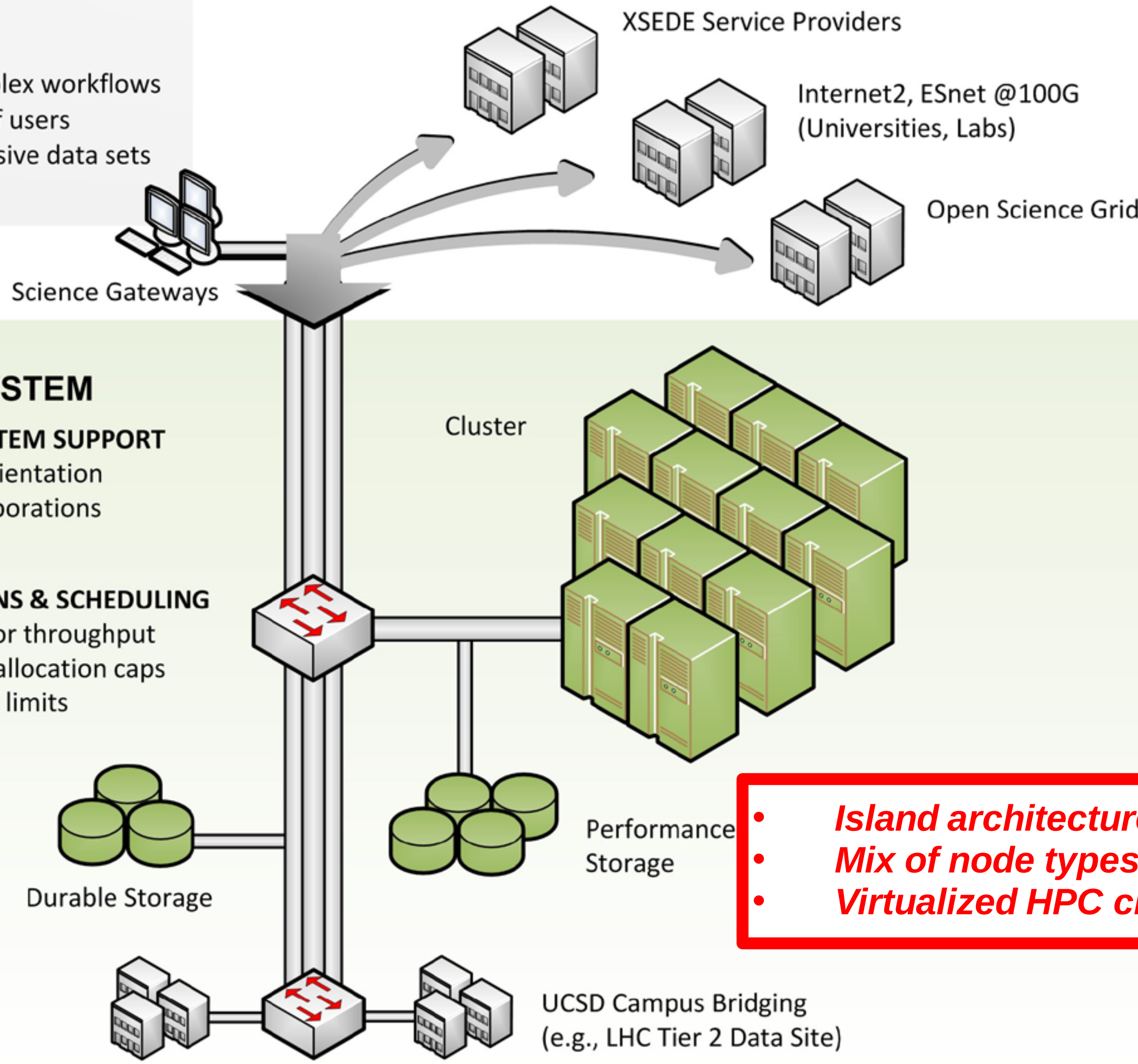
Cluster architecture
Fast standard nodes
Large-memory nodes
GPU-accelerated nodes
FDR InfiniBand

Storage architecture
Performance Storage
Durable Storage

Software
Science Gateways
Rich base of installed apps
Virtualization

USER & SYSTEM SUPPORT
New user orientation
XSEDE collaborations
FutureGrid

ALLOCATIONS & SCHEDULING
Optimized for throughput
Per-project allocation caps
Per-job core limits

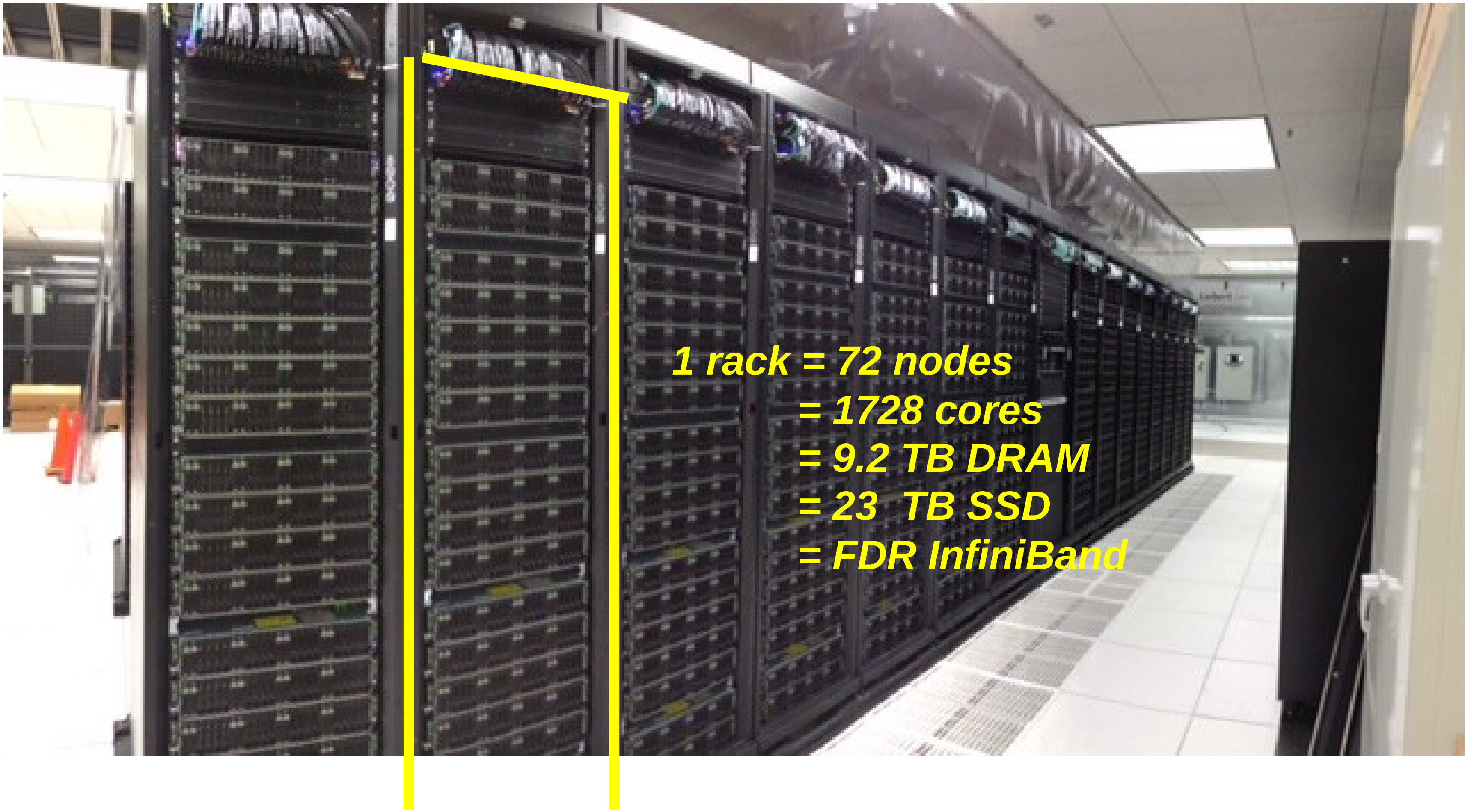


- *Island architecture*
- *Mix of node types*
- *Virtualized HPC clusters*

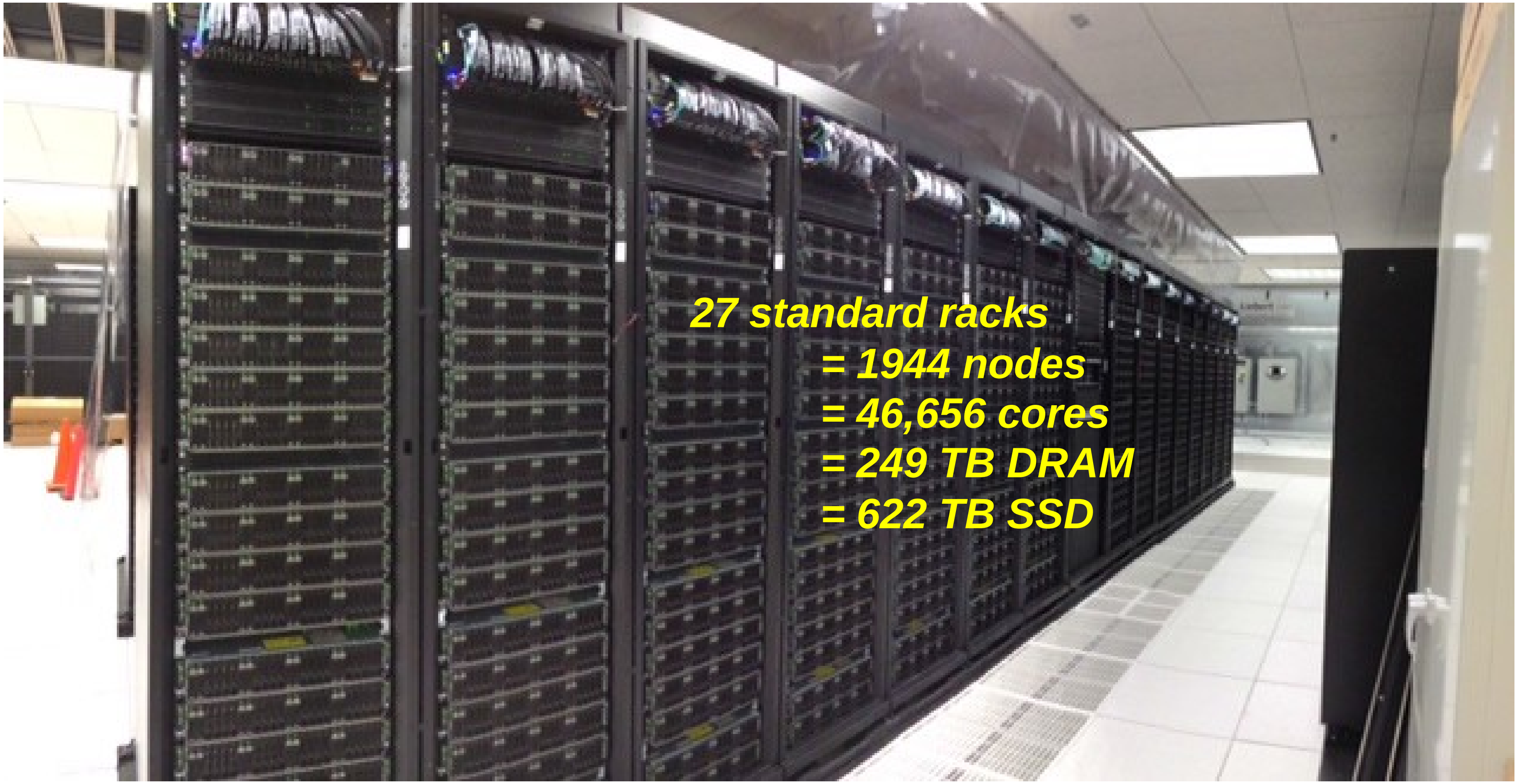
Comet: System Characteristics

- ***Total peak flops ~2.1 PF***
- ***Dell primary integrator***
  - *Intel Haswell processors w/ AVX2*
  - *Mellanox FDR InfiniBand*
- ***1,944 standard compute nodes (46,656 cores)***
  - *Dual CPUs, each 12-core, 2.5 GHz*
  - *128 GB DDR4 2133 MHz DRAM*
  - *2*160GB GB SSDs (local disk)*
- ***72 GPU nodes***
  - *36 nodes same as standard nodes plus Two NVIDIA K80 cards, each with dual Kepler3 GPUs*
  - *36 nodes with 2 14-core Intel Broadwell CPUs plus 4 NVIDIA P100 GPUs*
- ***4 large-memory nodes***
  - *1.5 TB DDR4 1866 MHz DRAM*
  - *Four Haswell processors/node*
  - *64 cores/node*
- ***Hybrid fat-tree topology***
  - *FDR (56 Gbps) InfiniBand*
  - *Rack-level (72 nodes, 1,728 cores) full bisection bandwidth*
  - *4:1 oversubscription cross-rack*
- ***Performance Storage (Aeon)***
  - *7.6 PB, 200 GB/s; Lustre*
  - *Scratch & Persistent Storage segments*
- ***Durable Storage (Aeon)***
  - *6 PB, 100 GB/s; Lustre*
  - *Automatic backups of critical data*
  - *Home grown*
- ***Home directory storage***
- ***Gateway hosting nodes***
- ***100 Gbps external connectivity to Internet2 & ESNet***

~67 TF supercomputer in a rack

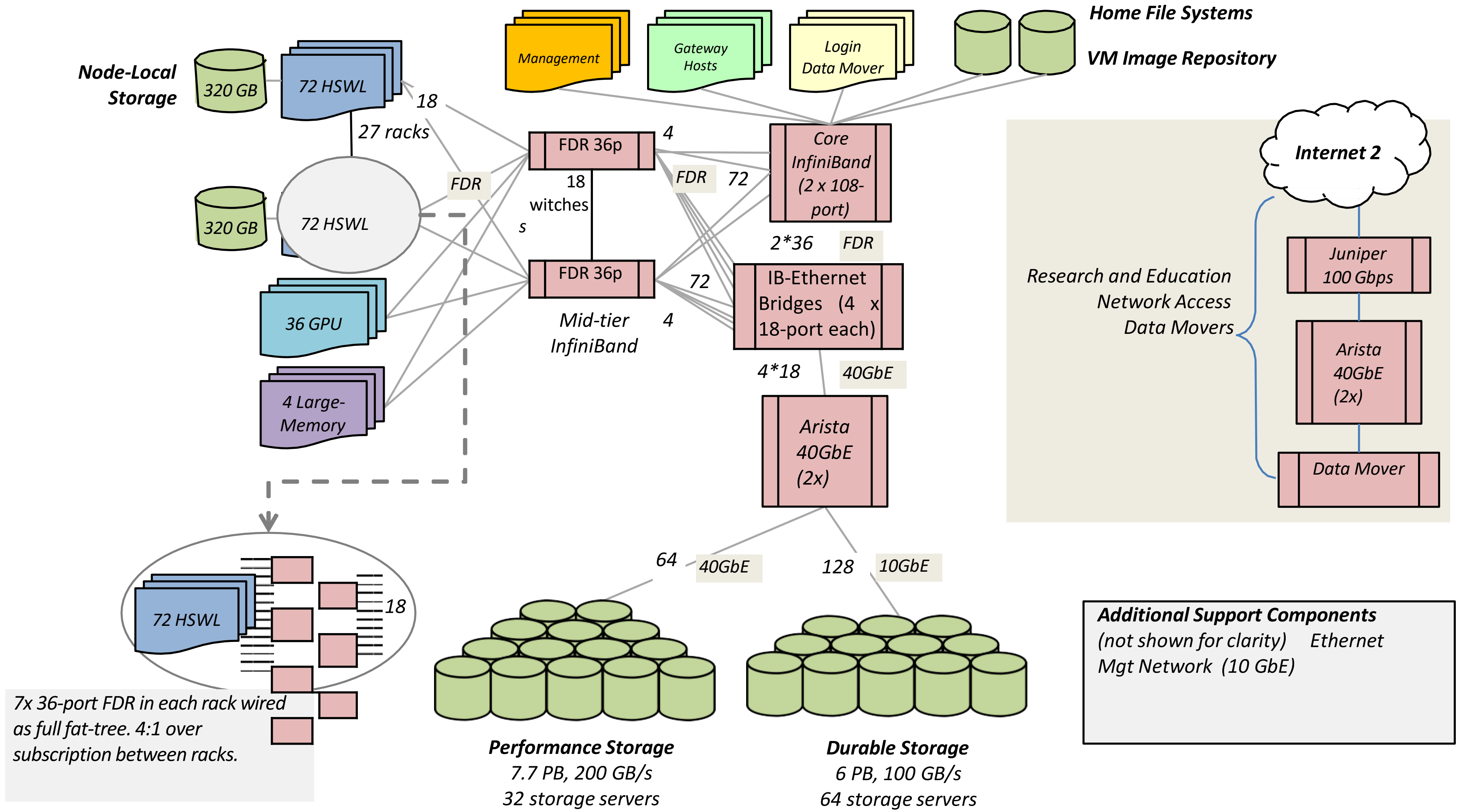


And 27 single-rack supercomputers



Comet Network Architecture

InfiniBand compute, Ethernet Storage

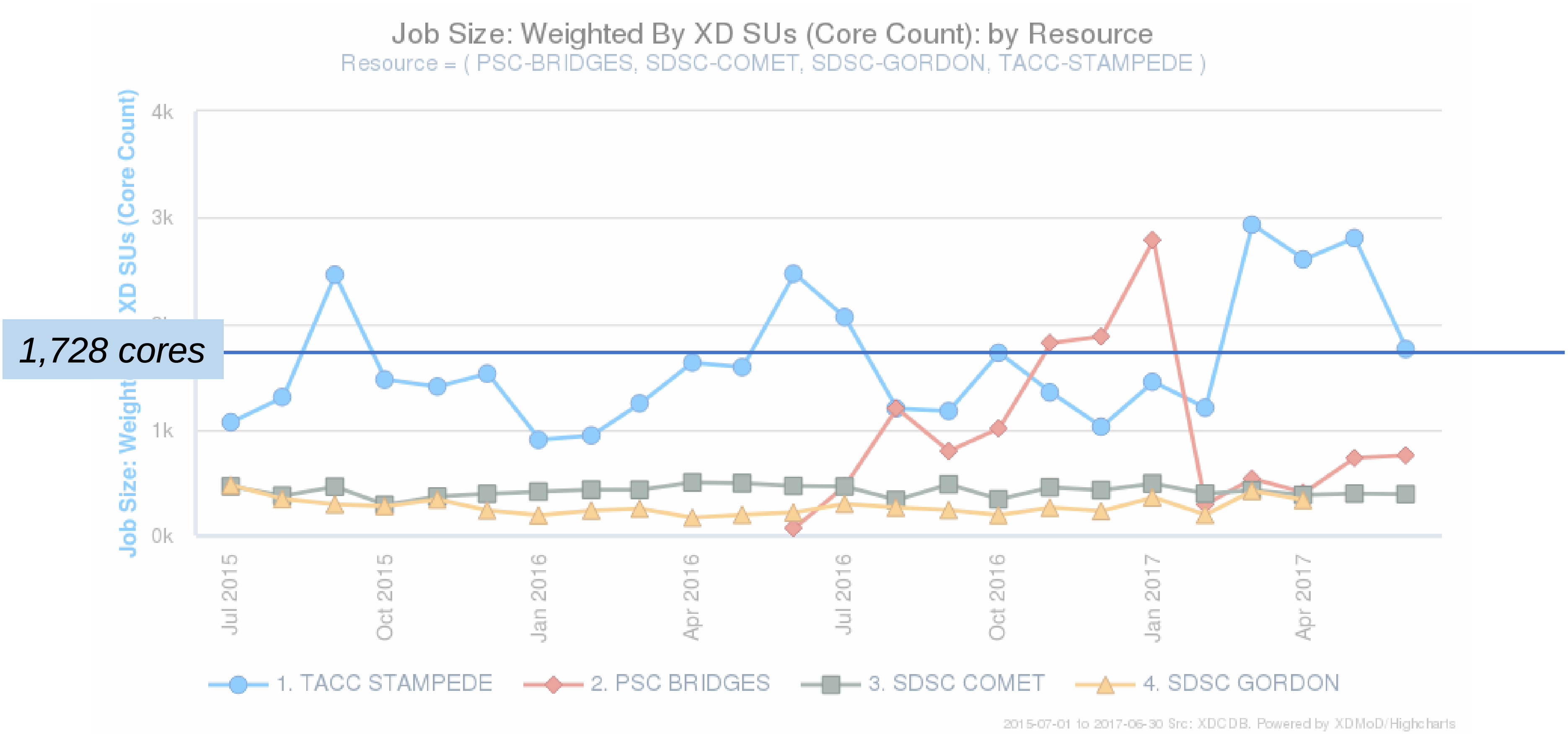


Comet Flexibility Addresses Diverse Needs

- ***Wide range of hardware options***
  - *Large number of regular compute nodes (**1,944**) with **128GB** of memory and **210GB of local flash**.*
  - *Subset of compute nodes have **1.5TB of local flash***
  - *4 large memory (**1.5TB RAM**) nodes*
  - ***72 GPU nodes (36 with K80s and 36 with P100s)** with 4 GPUs each.*
- ***Flexible Software Environment***
  - ***Rich set of applications** (>100) in regular compute environment*
  - ***Hadoop/Spark capability** can be enabled within regular scheduler environment.*
  - *Supports **Singularity based containerization** to enable other Linux based environments (for example Ubuntu). Users can upload their own images!*
  - *Virtual Clusters (VC) – see operational bullet below.*
- ***Flexible Operations***
  - ***Flexible scheduler environment** – shared and exclusive queues, long running jobs, focus on quick turn around time*
  - *Research Groups/communities, who have people in their group with **expert system administration skills**, can build their **own virtual clusters** with a custom OS and custom operational setup.*

One rack of Comet provides full bisection bandwidth up to 1,728 cores

(average job size across XSEDE < 2000 cores)



Comet's operational polices and software are designed to support long tail users

- Allocations
  - Individual PIs limited to 10M SU
  - Gateways can request more than 10M SUs
  - Gateways exempt from "reconciliation" cuts
- Optimized for throughput
  - Job limits are set at jobs of 1,728 cores or less (a single rack)
  - Support for shared node jobs is a boon for high throughput computing and utilization
  - Comet "Trial Accounts" provide 1000 SU accounts within one day
- Science gateways reach large communities
  - There 13 gateways on Comet, reaching thousands of users through easy to use web portals
- Virtual Clusters (VC) support well-formed communities
  - Near native IB performance
  - Project-controlled resources and software environments
  - Requires the allocation team possess systems administration expertise

Comet: MPI options, RDMA enabled software

MVAPICH2 v2.1 is the default MPI on Comet. v2.2 and v2.3 also available
Intel MPI and OpenMPI also available.

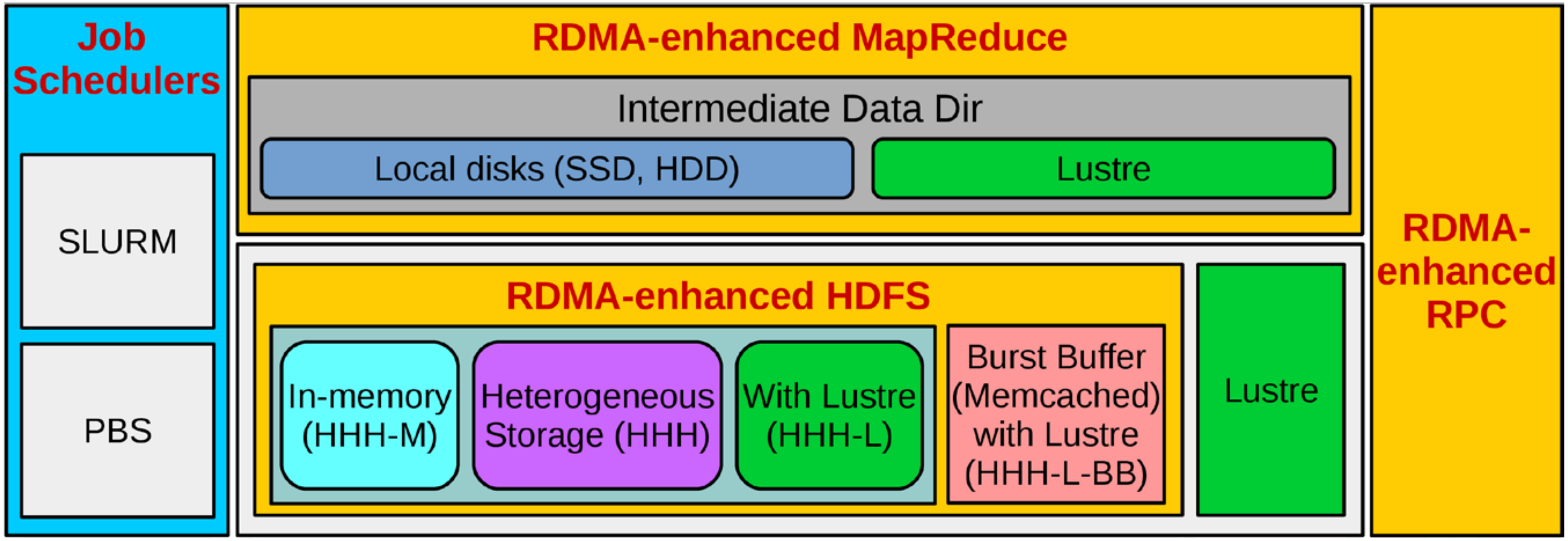
MVAPICH2-X v2.2a *to provide unified high-performance runtime supporting both MPI and PGAS programming models.*

MVAPICH2-GDR (v2.2) on the GPU nodes featuring NVIDIA K80s and P100s. (Tuesday presentation by Mahidhar on Benchmark and application performance)

RDMA-Hadoop (2x-1.1.0), RDMA-Spark (0.9.5) (from Dr. Panda's HiBD lab) also available.

RDMA-Hadoop, Spark *NOWLAB research*

- *Exploit performance on modern clusters with RDMA-enabled interconnects for Big Data applications.*
- *Hybrid design with in-memory and heterogeneous storage (HDD, SSDs, Lustre).*
- *Keep compliance with standard distributions from Apache.*



RDMA-Hadoop and RDMA-Spark

Network-Based Computing Lab, Ohio State University

- ***HDFS, MapReduce, and RPC over native InfiniBand and RDMA over Converged Ethernet (RoCE).***
- ***Based on Apache distributions of Hadoop and Spark.***
- ***Version RDMA-Apache-Hadoop-2.x 1.1.0 (based on Apache Hadoop 2.6.0) available on Comet***
- ***Version RDMA-Spark 0.9.5 (based on Apache Spark 2.1.0) available on Comet.***
- ***More details on the RDMA-Hadoop and RDMA-Spark projects at:***
  - http://hibd.cse.ohio-state.edu/

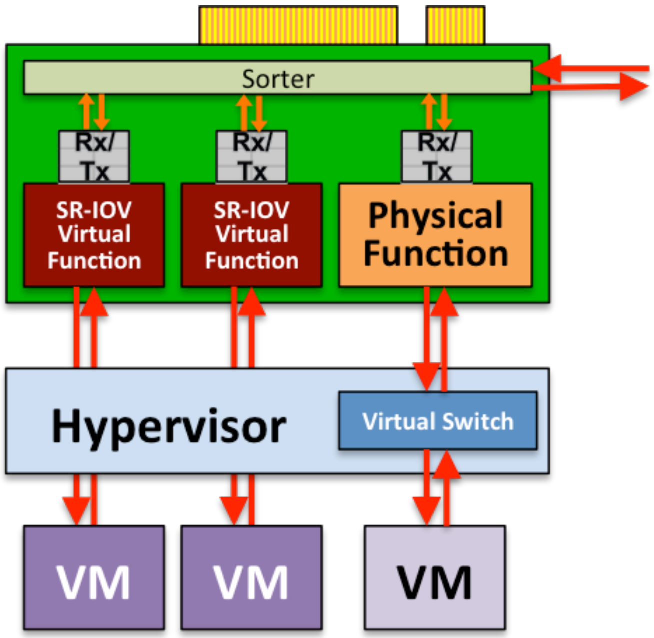
Motivation for Virtual Clusters

- OS and software requirements are diversifying. *Growing number of user communities that can't work in traditional HPC software environment.*
- Communities that have expertise and ability to utilize large clusters but *need hardware*.
- Institutions that have *bursty or intermittent need* for computational resources.

Goal: Provide near bare metal HPC performance and management experience for groups that can manage their own clusters.

Key for Performance: *Single Root I/O Virtualization (SR-IOV)*

- Problem: Virtualization generally has resulted in significant I/O performance degradation (e.g., excessive DMA interrupts)
- Solution: SR-IOV and Mellanox ConnectX-3 InfiniBand host channel adapters
    - One physical function → multiple virtual functions, each light weight but with its own DMA streams, memory space, interrupts
    - Allows DMA to bypass hypervisor to VMs
- *SRIOV enables virtual HPC cluster w/ near-native InfiniBand latency/bandwidth and minimal overhead*

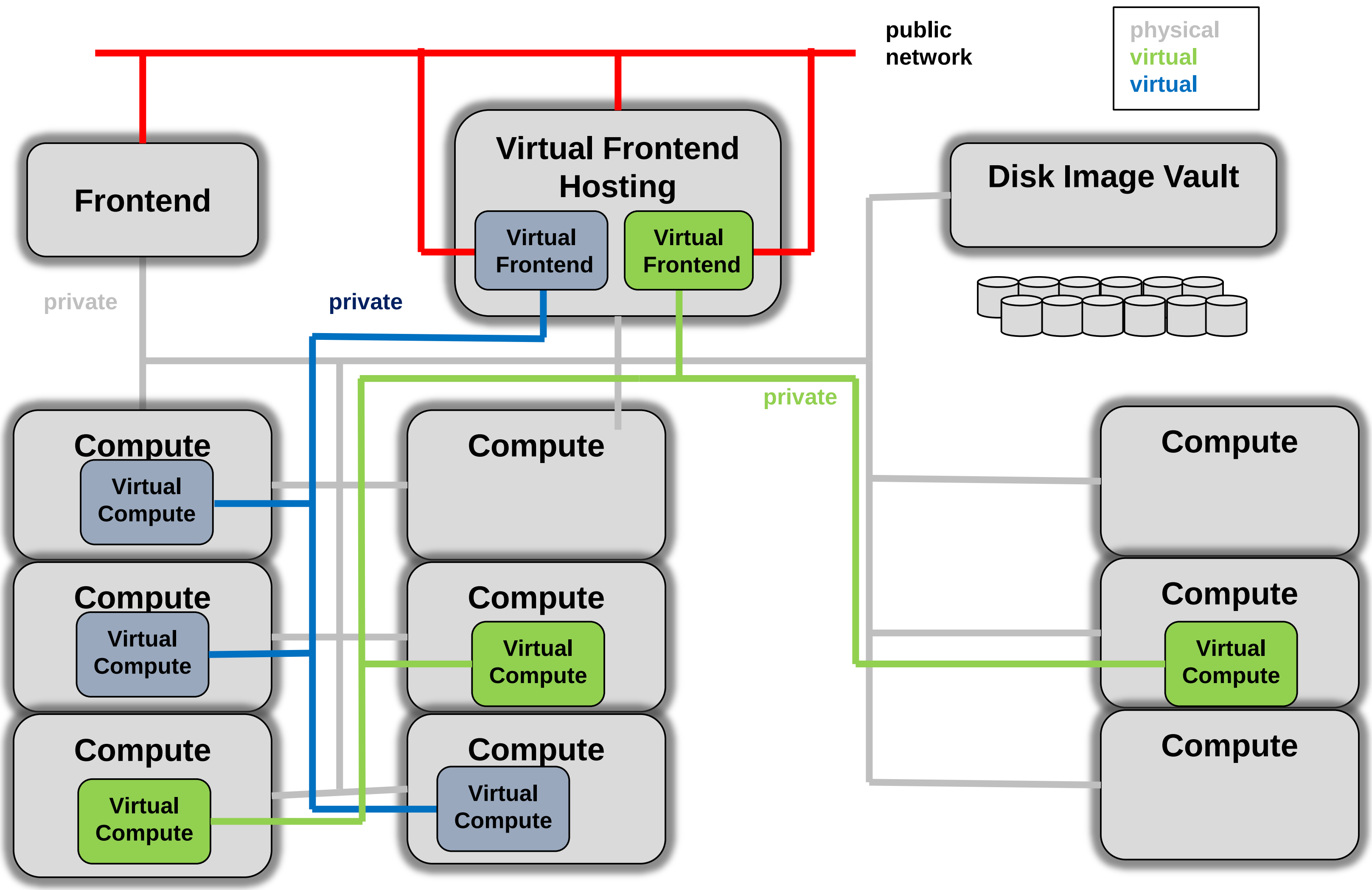


Mellanox TECHNOLOGIES

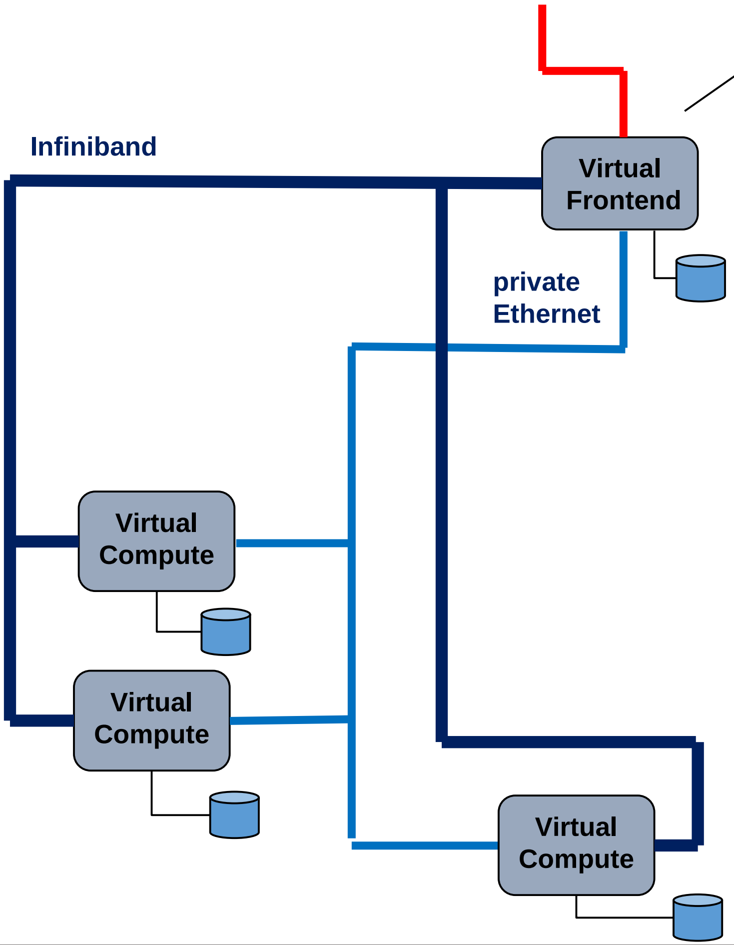
Overview of Virtual Clusters on Comet

- Projects have persistent VM for cluster management
  - Modest: single core, 1-2 GB of RAM
- Standard compute nodes will be scheduled as containers via batch system
  - One virtual compute node per container
- Virtual disk images stored as ZFS datasets
  - Migrated to and from containers at job start and end
- VM use allocated and tracked like regular computing

User-Customized HPC



High Performance Virtual Cluster Characteristics



Comet: Providing Virtualized HPC for XSEDE

Infiniband Virtualization
- 8% latency overhead.
- Nominal bandwidth overhead

All nodes have
- Private Ethernet
- Infiniband
- Local Disk Storage

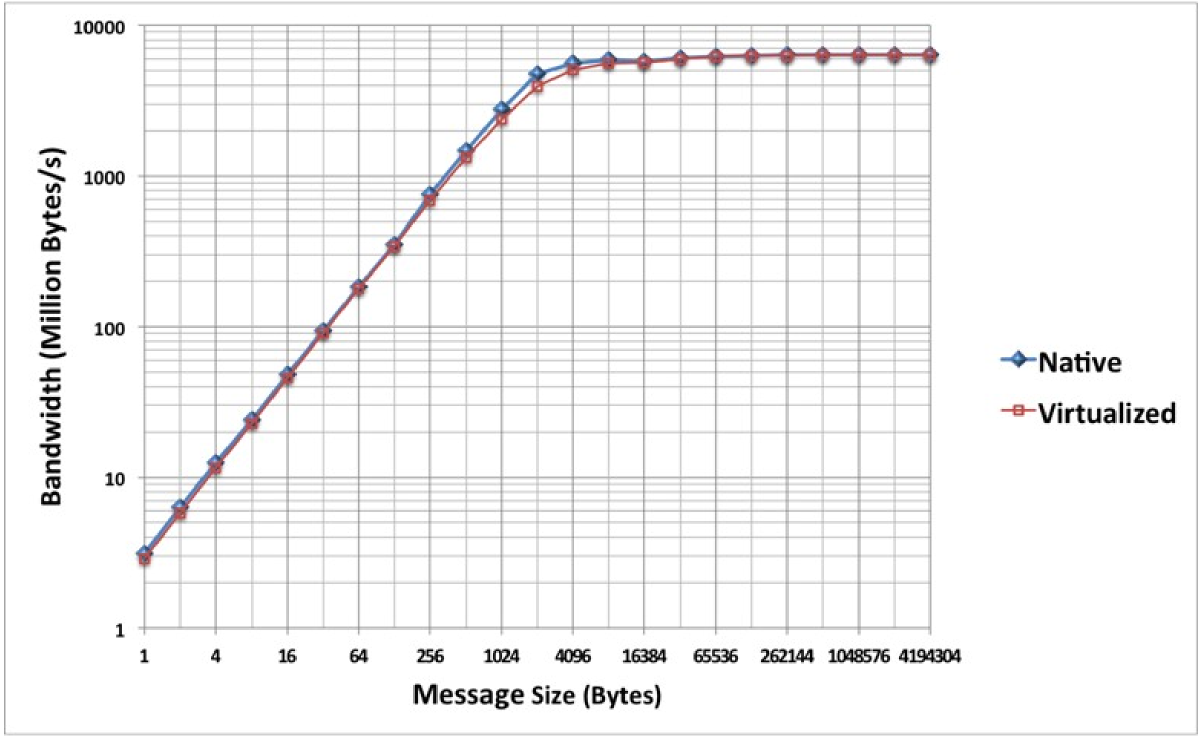
Virtual Compute Nodes can Network boot (PXE) from its virtual frontend

All Disks retain state
- *keep user configuration between boots*

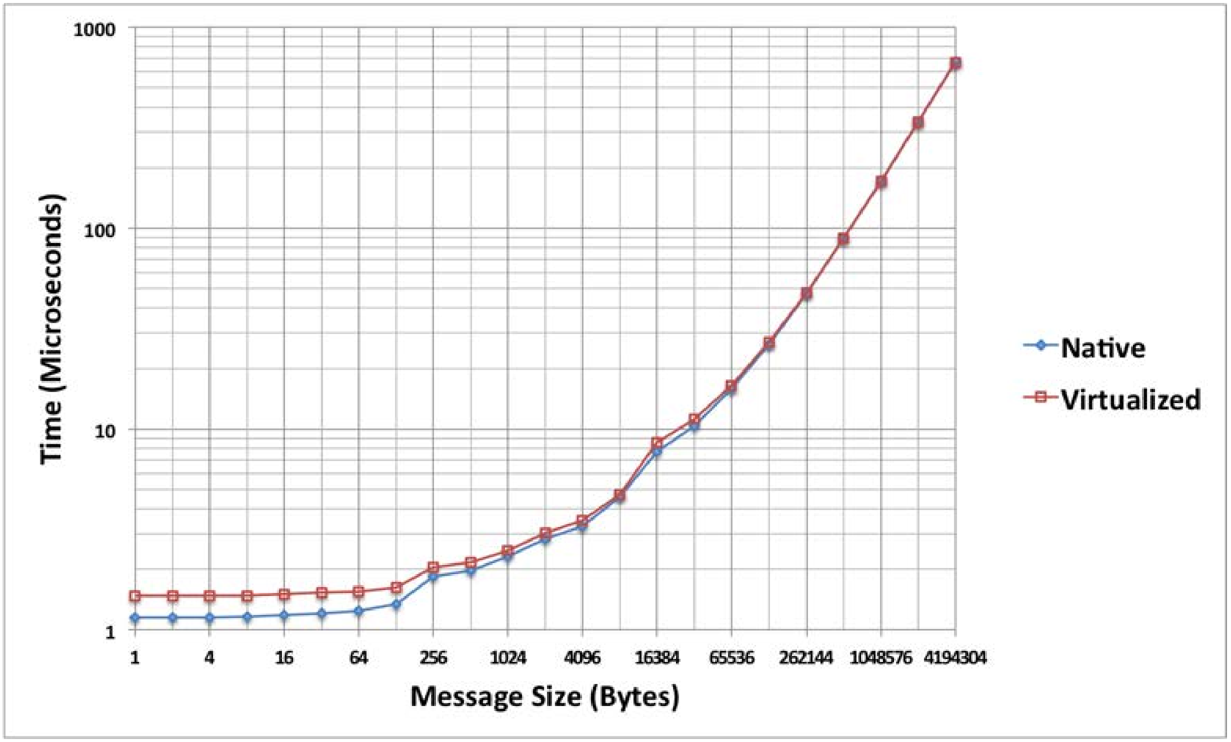
Data Storage/Filesystems

- Local SSD storage on each compute node
- Limited number of large SSD nodes (1.4TB) for large VM images
- Local (SDSC) network access same as compute nodes
- Modest (TB) storage available via NFS now

MPI bandwidth slowdown from SR-IOV is at most 1.21 for medium- sized messages & negligible for small & large ones

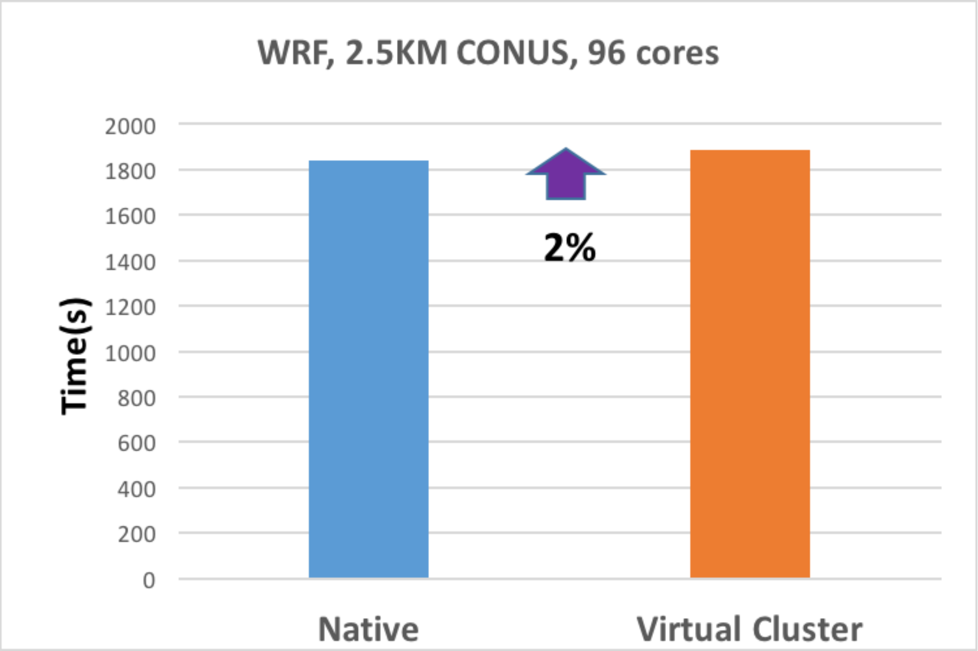


MPI latency slowdown from SR-IOV is at most 1.32 for small messages & negligible for large ones



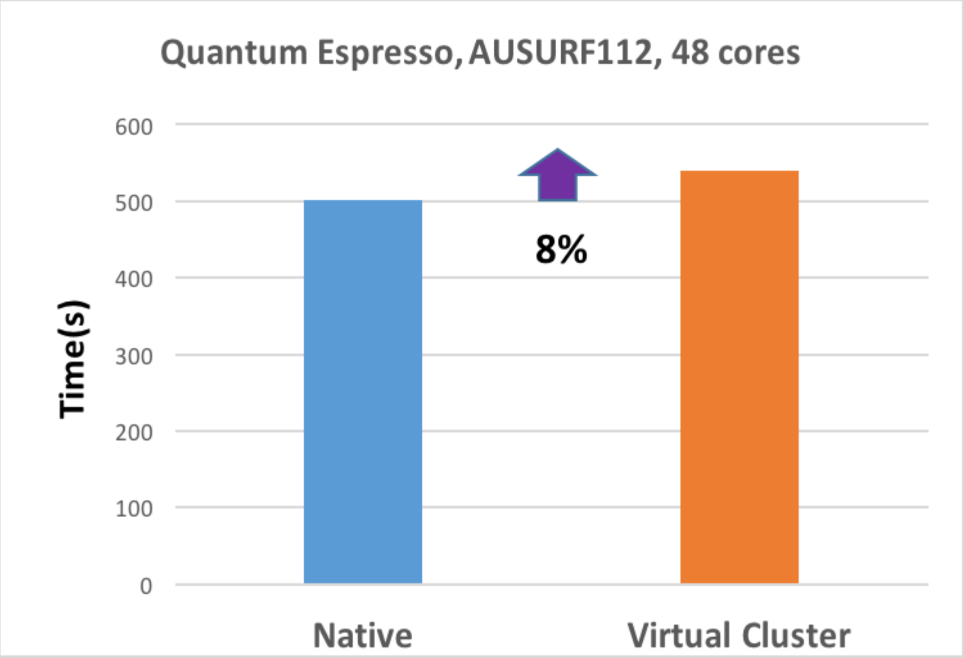
WRF Weather Modeling

- 96-core (4-node) calculation
- Nearest-neighbor communication
- Test Case: 3hr Forecast, 2.5km resolution of Continental US (CONUS).
- Scalable algorithms
- 2% slower w/ SR-IOV vs native IB.



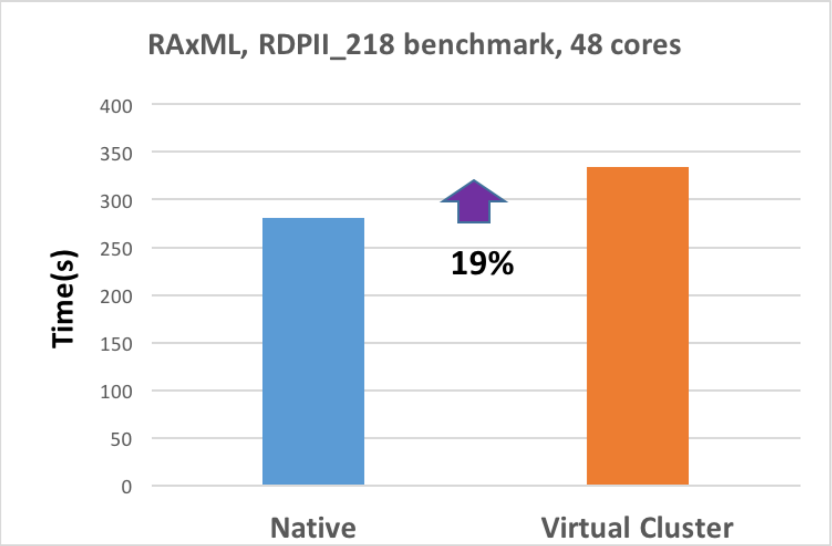
Quantum ESPRESSO

- *48-core (3 node) calculation*
- *CG matrix inversion - irregular communication*
- *3D FFT matrix transposes (all- to-all communication)*
- *Test Case: DEISA AUSURF 112 benchmark.*
- *8% slower w/ SR-IOV vs native IB.*



RAxML: Code for Maximum Likelihood-based inference of large phylogenetic trees.

- *Widely used, including by CIPRES gateway.*
- *48-core (2 node) calculation Hybrid MPI/Pthreads Code.*
- *12 MPI tasks, 4 threads per task. Compilers: gcc + mvapich2 v2.2, AVX options.*
- *Test Case: Comprehensive analysis,*
- *218 taxa, 2,294 characters, 1,846 patterns, 100 bootstraps specife d. 19% slower w/ SR-IOV vs native IB.*



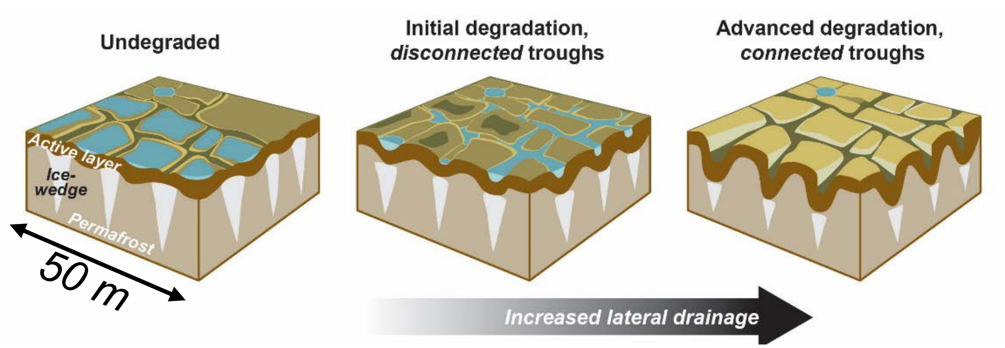
Research case studies

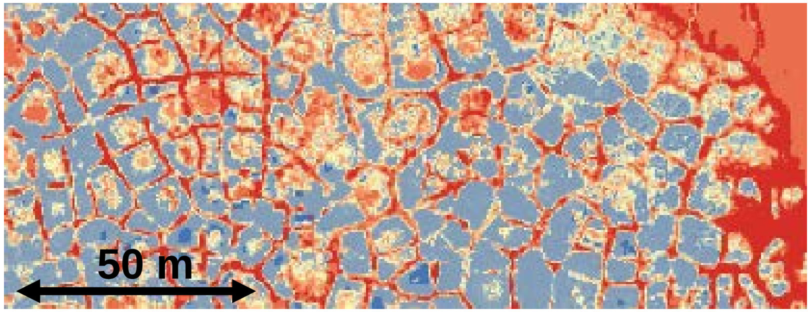
- The following slides illustrate the many ways in which researchers from a wide range of domains have used Comet
- We now have many projects that are outside of the “traditional” supercomputing domains
  - finance, linguistics, artificial intelligence, art, music and ecology.
- Usage by the life sciences has grown significantly and now accounts for half of compute cycles consumed on Comet.

Tundra drains as permafrost thaw

ECSS
Traditional HPC

Anna Liljedahl (Univ. of Alaska Fairbanks) has been using Comet to study Arctic hydrology and permafrost. This work has direct relevance to greenhouse gas emissions. No climate projections to date include permafrost thaw with differential ground subsidence at the <1 m scale, which drains the tundra.



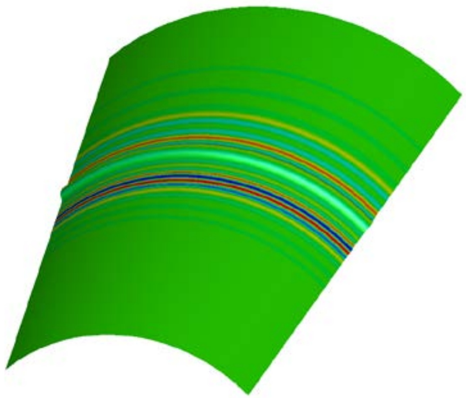


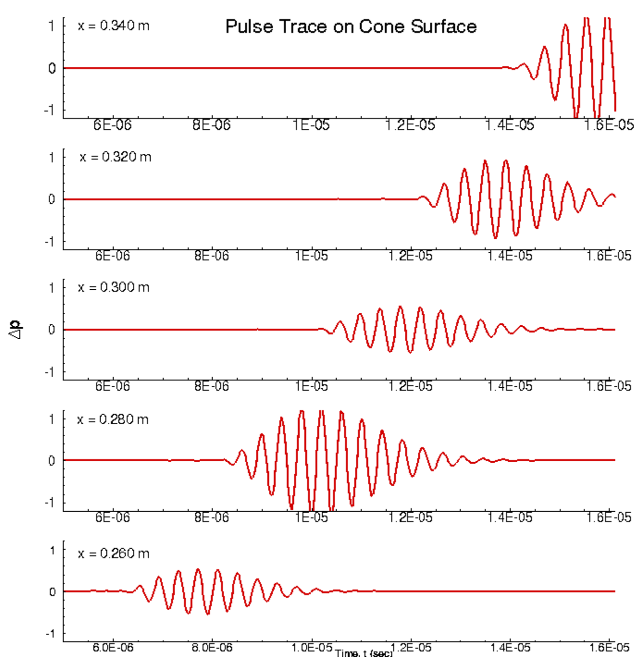
Soil temperature

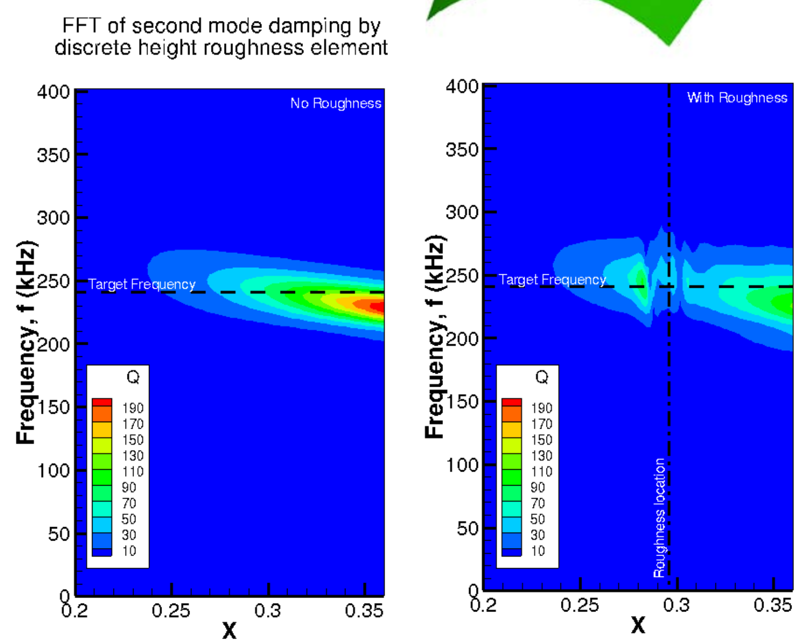
Liljedahl et al., Nature Geoscience 2016

Hypersonic Laminar-Turbulent Transition

As part of Dr. Xiaolin Zhong's research group (UCLA), Carleton Knisely and Christopher Haley use Comet to study boundary layer transition in hypersonic flows. Strategic placement of discrete roughness elements can dampen second mode instability waves, leading to a delay in transition to turbulence. Delaying transition can reduce the heat and drag on a hypersonic vehicle, allowing for heavier payloads and greater fuel efficiency.



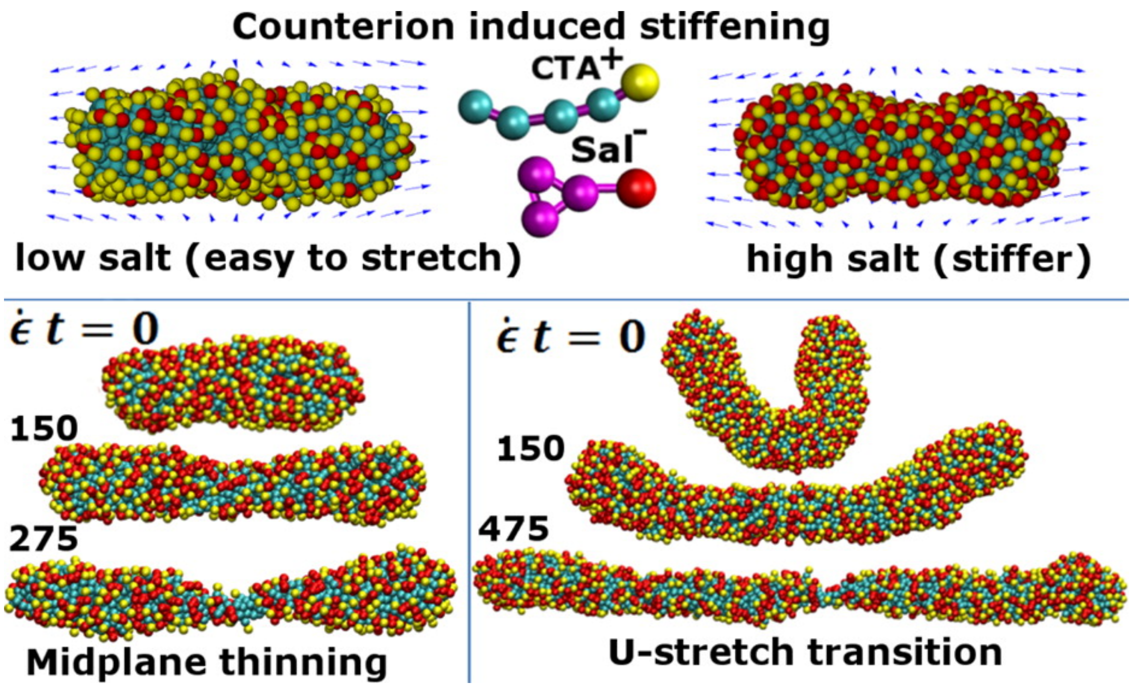




Fluid drag reduction

Comet GPU

Surfactants are often used to reduce fluid drag in a variety of applications (oil pipelines, firefighting equipment). Subas Dhakal (Syracuse) used Comet's GPU nodes to perform molecular dynamics simulations showing how micelles formed from these surfactants deform during the flow.

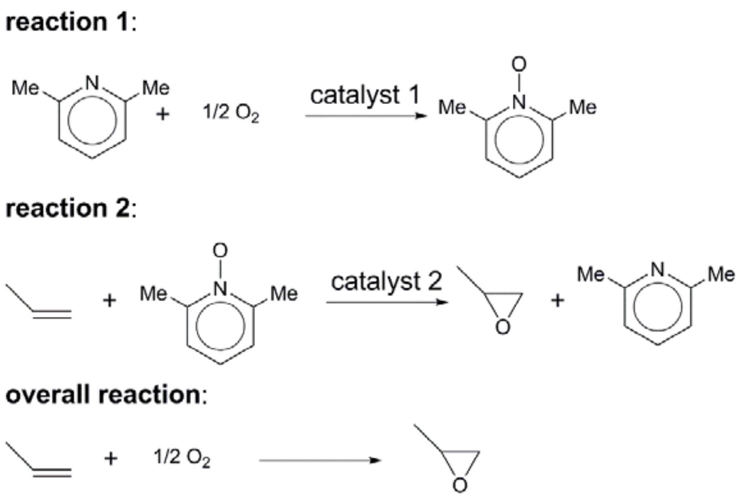


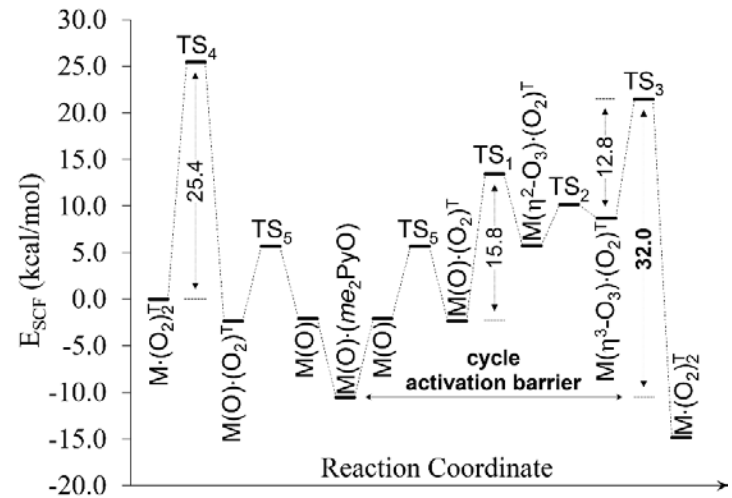
Schematic of the elongational flow simulation illustrating two types of deformations. The velocity field is shown by arrows. Color scheme: red (salicylate anions), yellow (hydrophilic part of the surfactant), cyan (hydrocarbon tail), green (Cl^-), pink (Na^+)

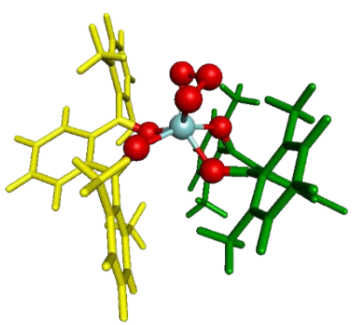
Organometallic Catalyst Design

Underserved community

Thomas Manz and Bo Yang (NMSU) have been using Comet to design new catalysts that utilize molecular O_2 as oxidant for organic selective oxidation reactions without requiring any co-reductant. The newly designed oxidation route could reduce energy consumption and environmental wastes.



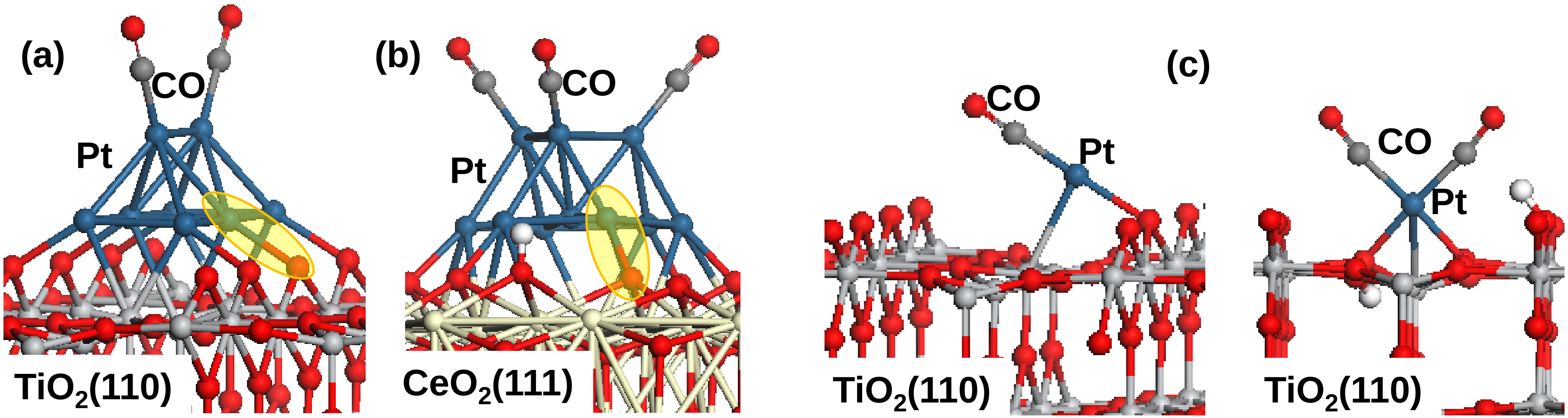




Top left: New 2-step selective oxidation scheme; catalyst 1 is a newly designed Zr-based organometallic catalyst, catalyst 2 is a Ru-porphyrin catalyst. Bottom left: the key intermediate of catalyst 1. Top right: computed energy profile for one of the catalytic cycles of reaction 1.

Multiscale Modeling of Bifunctional Catalysis

Andreas Heyden (U. South Carolina) has been using Comet to investigate bifunctional heterogeneous catalysis occurring at the three-phase boundary (TPB) of a gas-phase, a reducible oxide surface, and a noble metal cluster to understand the origin of the unique activity of these catalysts for the water-gas shift.



TPB models of Pt cluster supported on TiO_2 and CeO_2

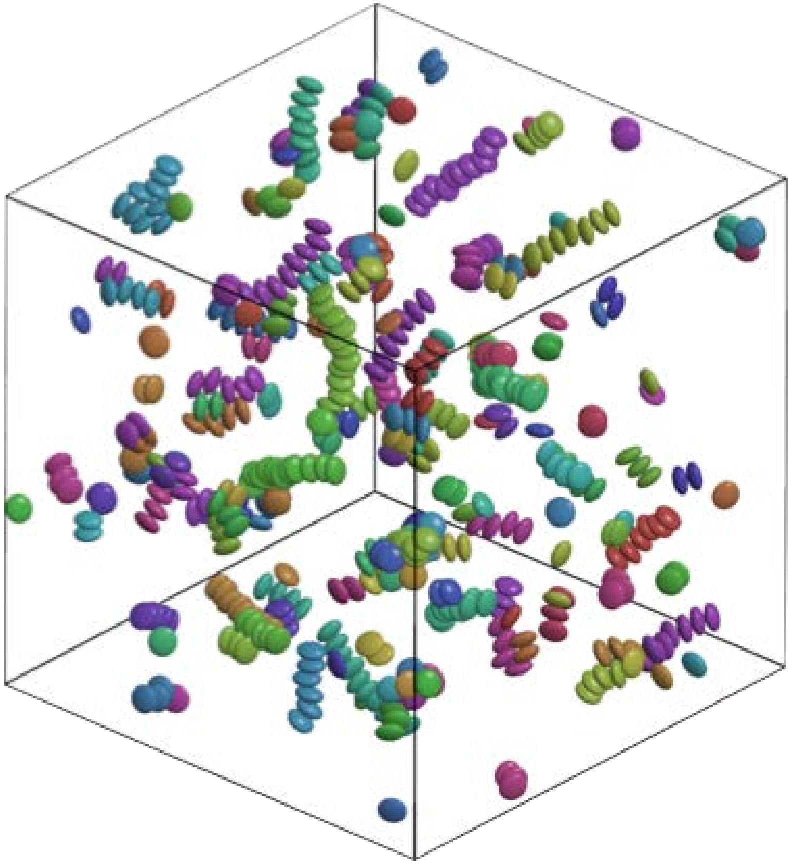
TPB models of single Pt metal supported on TiO_2

Insights obtained from this study can be applied to various chemical reactions catalyzed by reducible oxide supported noble metals. Understanding single metal catalysis will be beneficial for the design of novel heterogeneous catalysts which offer maximum atom efficiency with minimal amount of metal loading.

Colloids and self-assembling systems

GPU
High throughput

Sharon Glotzer (U. Michigan) uses Comet to simulate colloids of hard particles, including spheres, spheres cut by planes, ellipsoids, convex polyhedra, convex spheropolyhedra, and general polyhedra.



Glotzer's work can lead to the design of better materials, including surfactants. liquid crystals and nanoparticles that spontaneously assemble into sheets, tubes, wires or other geometries

Workload involves large numbers of small jobs – 147K ran on single core, 219K on single node

ECSS
Novel community

Image Analysis of Rural Photography

Elizabeth Wuerffel (Valparaiso U) and the IARP team are running computer vision techniques on Comet to analyze and tag 171,021 images from the Farm Security Administration – Office of War Information Photography Collection (1935-1944) at Library of Congress.

Feature extraction to database to interface to visual data mining

Title: "Barber and shop"
Location: Omaha, Nebraska
Photographer: John Vachon.
Date Created: 1938.

Image Gray Scale:

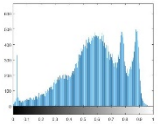
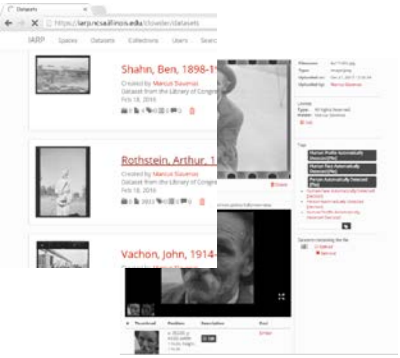


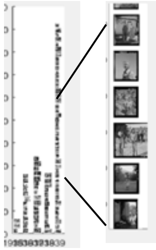
Image Content
OCR : BARBER SHOP; ENGLISH
FACES : 1
Metadata
SEMANTICS:
<shop::business;structure;entity>
<barber::worker;person>
GEO : 41.2°N, 95.9°W
etc…

IARP Database



SQL with visualizations

Data Mining on American life, visual rhetoric, and aesthetics.

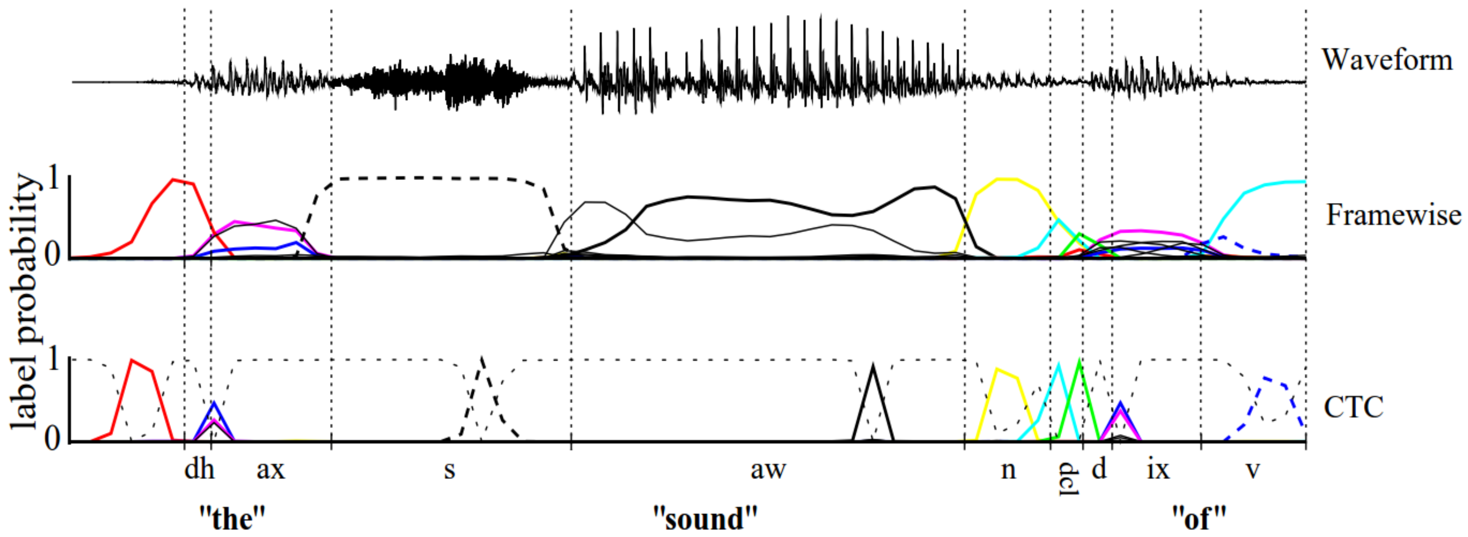


End-to-End Speech Recognition

GPU
Novel community

Florian Metze (Carnegie Mellon) uses Comet to study a new deep learning approach to automatic speech recognition where all components are neural networks and optimized as jointly. The results are available as open source.

Comparison of the novel “Connectionist Temporal Classification” approach to speech recognition with conventional frame-based approaches. Also see https://github.com/srvk/eesen.



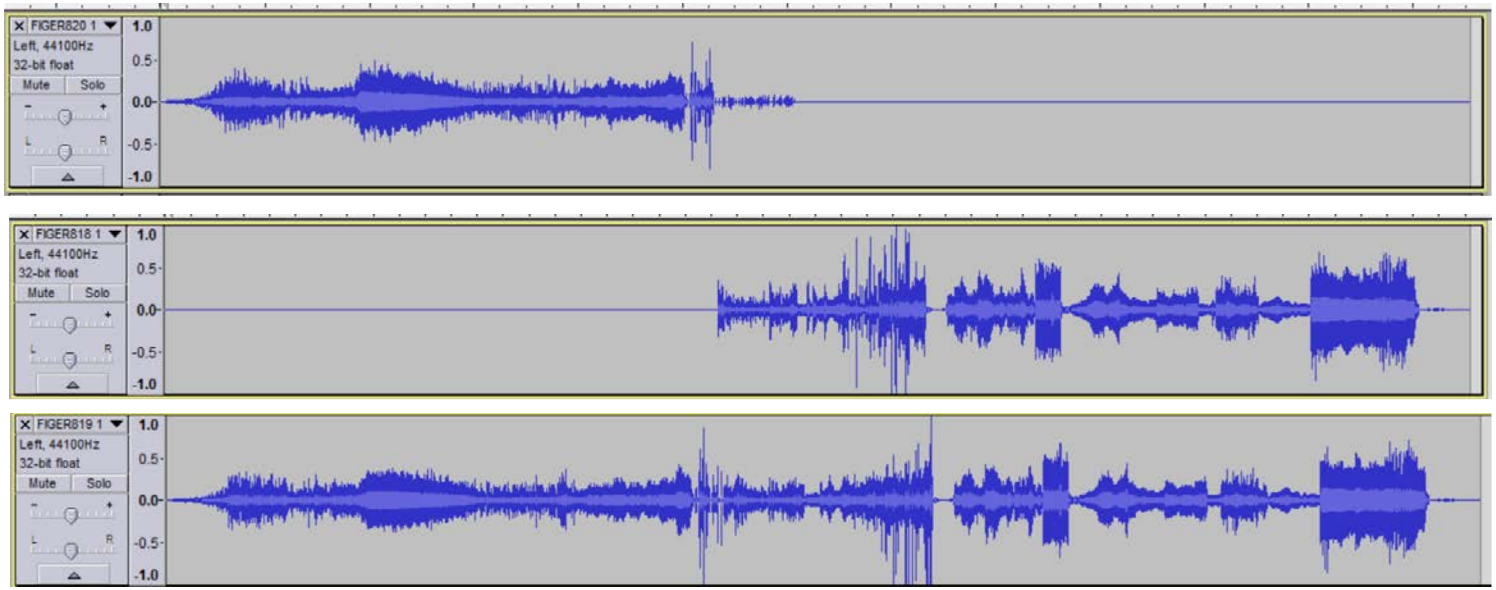
Digital instrument for sound synthesis and composition

Sever Tipei (Illinois at Urbana-Champaign) with XSEDE support is using Comet to implement a multi-node parallel version of his DISSCO tool for Computer-assisted Sound Composition.

Compute Node 1

Compute Node 2

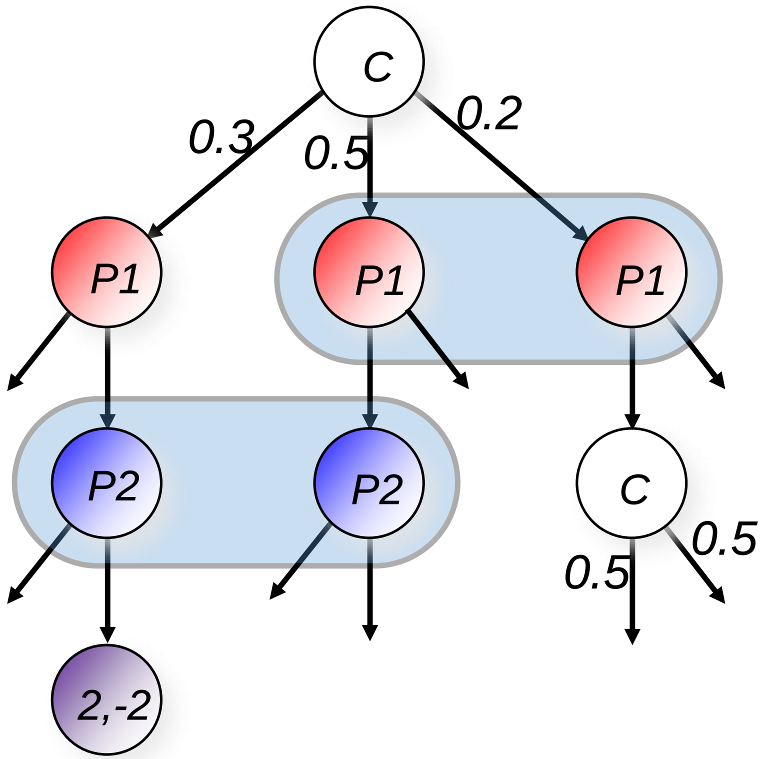
Master Node Combines



Speeding up processing toward real-time sound generation and musical scores.

Equilibrium Finding in Large Games

Prof. Tuomas Sandholm and PhD student Noam Brown (CMU) used Comet to compute near Nash-equilibrium strategies in very large imperfect-information games. A distributed, iterative regret-minimization algorithm traverses the game tree and converges to a solution over time.



The work can be applied to automated negotiation, cybersecurity, national defense, medicine (e.g., steering evolution and biological adaptation, and treatment planning) and other strategic interactions involving hidden information.

Winner AAAI Computer Poker Competition
Bankroll event of the Association for the Advancement of Artificial Intelligence (AIAA) 2016 Annual Computer Poker Competition.

ECSS
Novel community

Consumer purchase behavior

Federico Bumbaca (University of California – Irvine) uses Comet to develop highly scalable methods to study how consumers make purchase decisions. Since the number of purchases for most consumers is very small (typically 5-10), the analysis employs a Bayesian hierarchical model that allows for partial pooling between similar individuals.

Comet reduced the time to estimate the purchase behavior of 100 million consumers from 6 months to several hours

Hackathon enabled by virtualization

Virtualization
Novel community

Hosted at SDSC, February 2016. Theme was live measurements and monitoring of the global Internet routing system (BGP). Total of 90 attendees: 50 competing participants (30 graduate students), and 25 non-competing experts. Mix of academia, Industry, Institutions.

Comet provided compute resources to participants, including several virtualized nodes, that were essential to the event

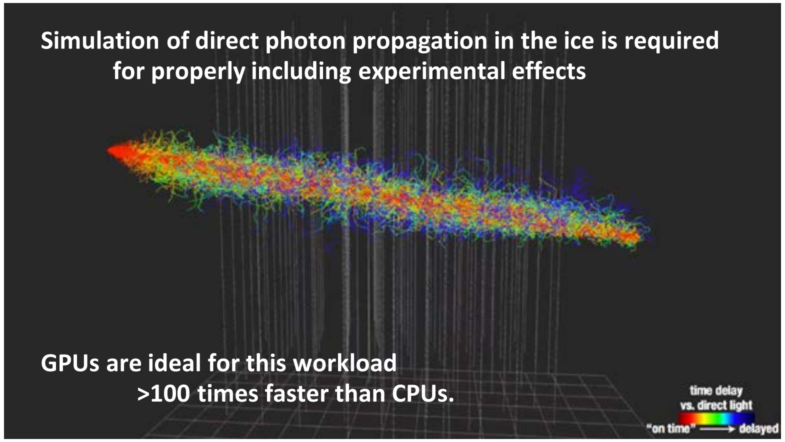
Over 15,000 SUs were used during 24 hours of hacking.

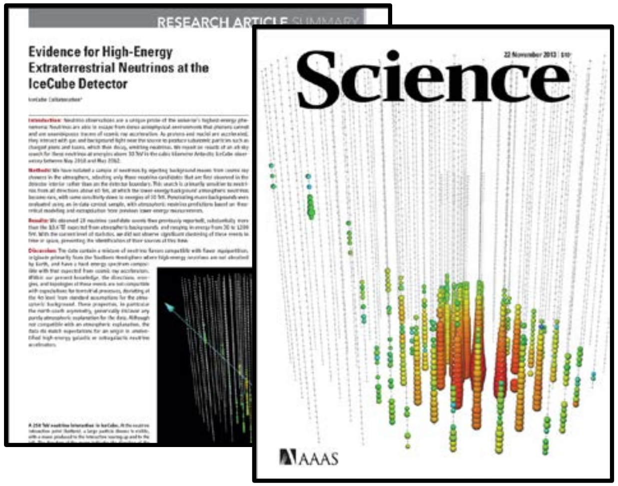
https://www.caida.org/workshops/bgp-hackathon/1602/

IceCube Neutrino Observatory

GPU

IceCube found the first evidence for astrophysical neutrinos in 2013 and is extending the search to lower energy neutrinos. The main challenge is to keep the background low and a precise simulation of signal and background is crucial to the analysis.



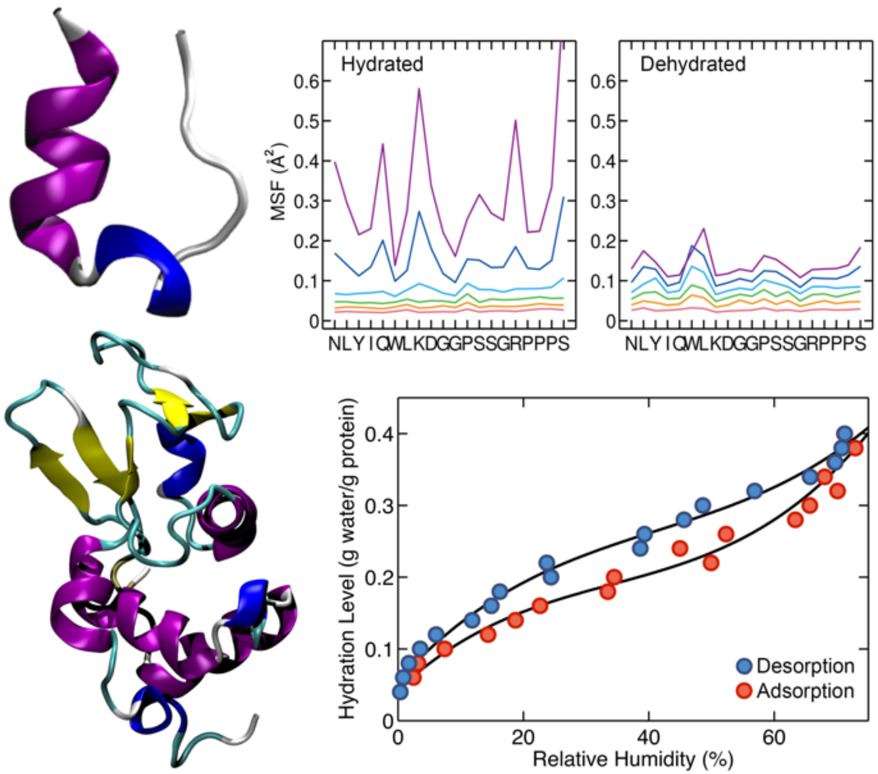


Comet's GPU nodes are a valuable resource for IceCube and integration with the experiment workload management system was very smooth thanks to previous OSG work on Comet

Protein lyophilization (freeze-drying)

Traditional HPC

Pablo Debenedetti (GaTech) uses Comet to study lyophilization (freeze-drying), a standard technique used to increase the storage life of labile biochemical, including therapeutic proteins, by the pharmaceutical industry

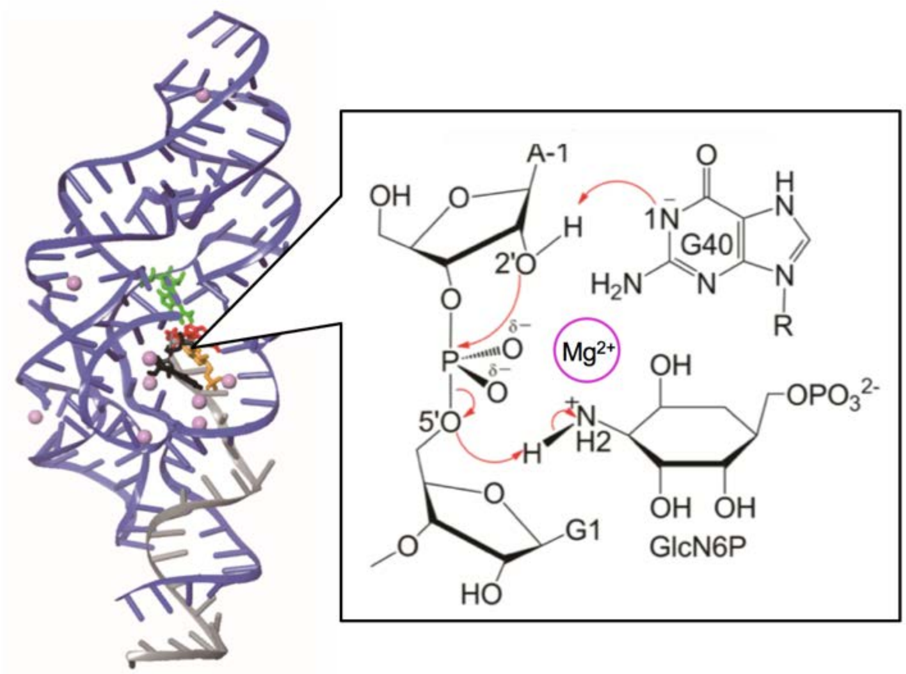


Top left: Trp-cage miniprotein structure. Top right: Mean-squared fluctuation for each residue in Trp-cage for the hydrated and dehydrated powder system. Bottom left: Lysozyme protein structure. Bottom right: Water sorption isotherm for lysozyme.

Self-cleavage of bacterial ribozyme

GPU

Sharon Hammes-Schiffer (UIUC) uses Comet's GPU nodes to perform molecular dynamics simulations of the self-cleavage reaction of the *glmS* ribozyme, which is essential for hydrocarbon synthesis in gram positive bacteria

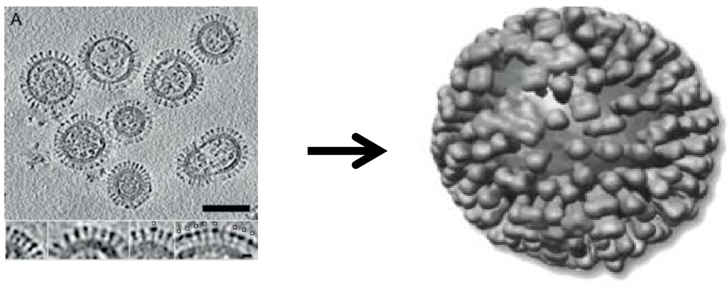


The glmS ribozyme structure shown and its active site. Red arrows indicate reaction directions: A-1(O2') is deprotonated by a general base and attacks the scissile phosphate, while G1 is protonated by the cofactor GlcN6P and eventually dissociates.

Studying flu at the molecular scale

GPU

Rommie Amaro (UCSD) uses Comet to to understand how molecular structure of the flu virus affects infectivity

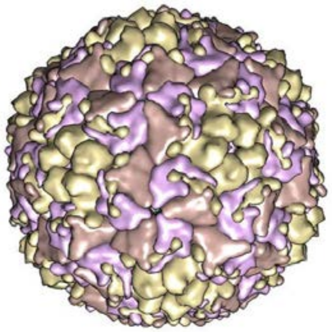


Alasdair Steven, NIH

Atomic model built from from experimentally determined structure. Brownian dynamics then used to understand how glycoprotein stalk height impacts substrate binding

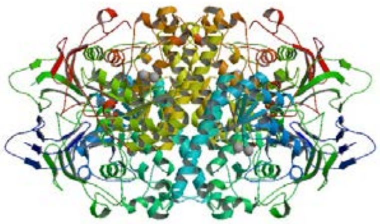
PDB_REDO – more accurate X-ray structures

The PDB_REDO project (Anastassis Perrakis, Netherlands Cancer Institute, with support from Janssen) aims to periodically update all X-ray structures in the PDB using more accurate algorithms. The calculations on the larger structures require Singularity and access to Comet's large memory nodes.



PDB ID 3ZFE: Human enterovirus 71 in complex with capsid binding inhibitor WIN51711

PDB ID 1O04 Cys302Ser mutant of human mitochondrial aldehyde dehydrogenase complexed with NAD+ and Mg2+



PDB ID 1GKP D-hydantoinase (Dihydropyrimidinas E) from Thermus Sp.

UC San Diego News Center

February 02, 2017 | By Jan Zverina

SDSC's 'Comet' Supercomputer Surpasses '10,000 Users' Milestone

Comet, the petascale supercomputer at the San Diego Supercomputer Center (SDSC), an Organized Research Unit of UC San Diego, has easily surpassed its target of serving at least 10,000 researchers across a diverse range of science disciplines, from astrophysics to redrawing the "tree of life".

In fact, about 15,000 users have used *Comet* to run science gateways jobs alone since the system went into production less than two years ago. A science gateway is a community-developed set of tools, applications, and data services and collections that are integrated through a web-based portal or suite of applications. Another 2,600 users have accessed the high-performance computing (HPC) resource via traditional runs. The target was established by SDSC as part of its cooperative agreement with the National Science Foundation (NSF), which awarded funding for *Comet* in late 2013.

Comet ~3 years in operation - Summary

- Users from total number of institutions 550+
- Total number of allocations 1,700+
- Number of unique standard users 4,700+
- Number of unique gateway users 33,000+

- Gateway friendliness impacting thousands of users

- GPU nodes making significant impact – some examples - analysis of data from large instruments(ICECUBE), MD packages (AMBER, LAMMPS), CIPRES gateway (BEAST), ML tools

- HPC Virtualization attracting users

Where is Comet (2015-2021) now ?

Answer: It is still at SDSC……till 2021…..

5. NOWLAB Impact on Science

Comet - ~3 years of operation
# of unique standard users 4,700+
# of unique gateway users 33,000+

Trestles
# of unique standard users 1,600+
Gateway users ~many thousands

Gordon
# of unique standard users 2,100+
Gateway users ~many thousands

SDSC alone ~2011 – now: 42,000 + users

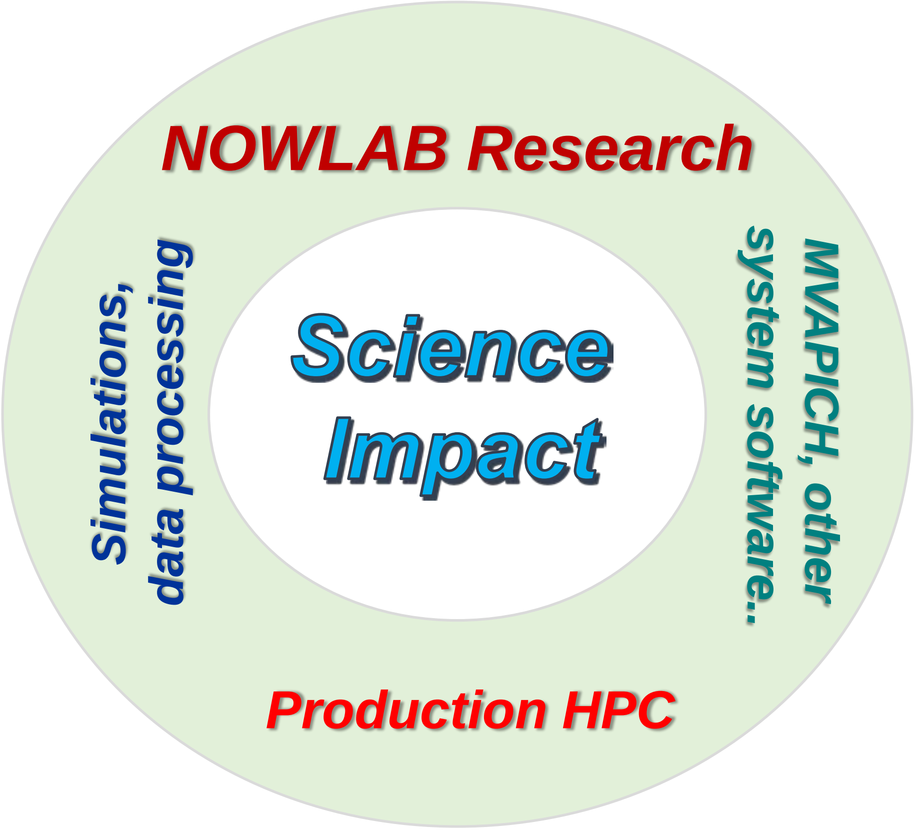
You add TACC machines – Ranger, Stampede etc.
Many more thousands of users

Personally Speaking

- *Wonderful and intellectually stimulating to collaborate with NOWLAB researchers/students*
- *Had (one ongoing) multiple NSF awards 3-way with OSU/OSC, TACC, SDSC over a decade now*
- *Worked with very smart people from OSU, TACC, SDSC for the past decade – many in the room today – hope to continue…..*

Think of total number of publications, Ph.D/MS thesis work

- **20% / 40% / 60% of the user publish**
- **However you look at it – thousands of research publications, Ph.D/MS thesis**
- **Impact of NOWLAB system software**



MUG13

MUG14

I got little younger!!

Karl got fashion - of course

MUG15

MUG16

Adam not wearing red OSU shirt

DK same blue shirt, jacket

MUG17