


A desert landscape with mountains and a road. The sky is blue with some clouds. The mountains are brown and rocky. The road is dark and runs across the bottom of the image.

# *Gordon*

## *A Data-Intensive Supercomputer*

*HPC User Forum*  
*April 16-18, 2012*

Robert Sinkovits  
Gordon Project Applications Lead  
San Diego Supercomputer Center



*Gordon*  
*A Data Monster (M. Vildibil - Appro)*

*HPC User Forum*  
*April 16-18, 2012*

Robert Sinkovits  
Gordon Project Applications Lead  
San Diego Supercomputer Center

---

# Why Gordon?



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 is a national resource made possible through a grant from the National Science Foundation. One of three OCI Track 2D awards for innovative systems

Available to all U.S. academic researchers on a competitive basis and on a limited basis for-fee to most non-academic users



Design  
Deployment  
Support



Processors  
Motherboards  
Flash drives



Integrator



vSMP Foundation



Funding  
OCI #0910847



3D Torus

---

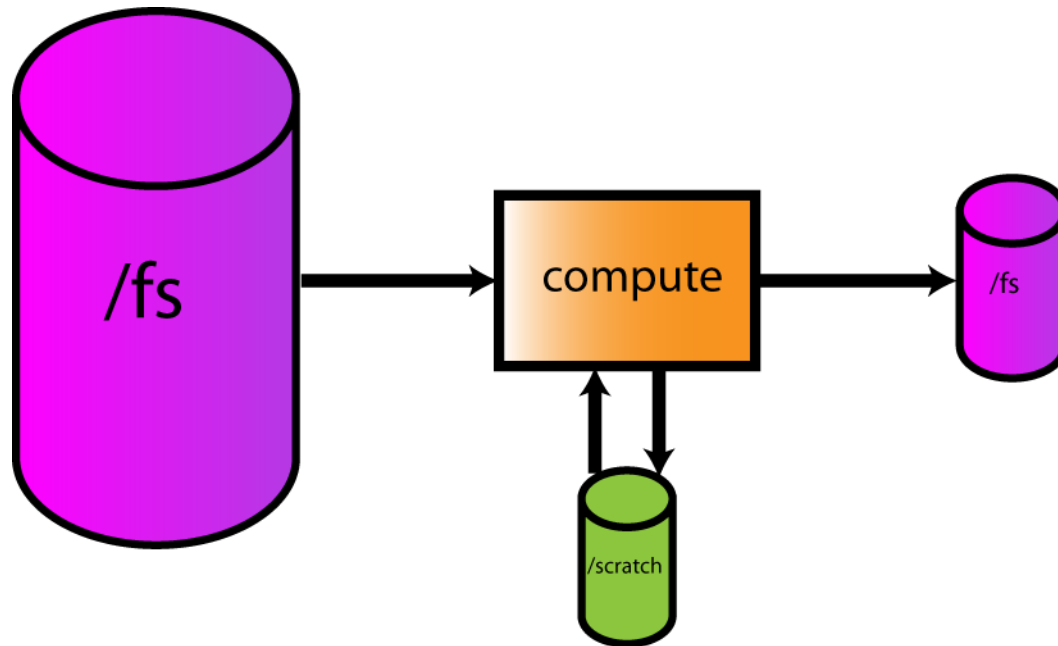
# Challenge for academic HPC

SDSC (and TACC, PSC, NCSA, NICS, etc.) have very limited control over their user base.

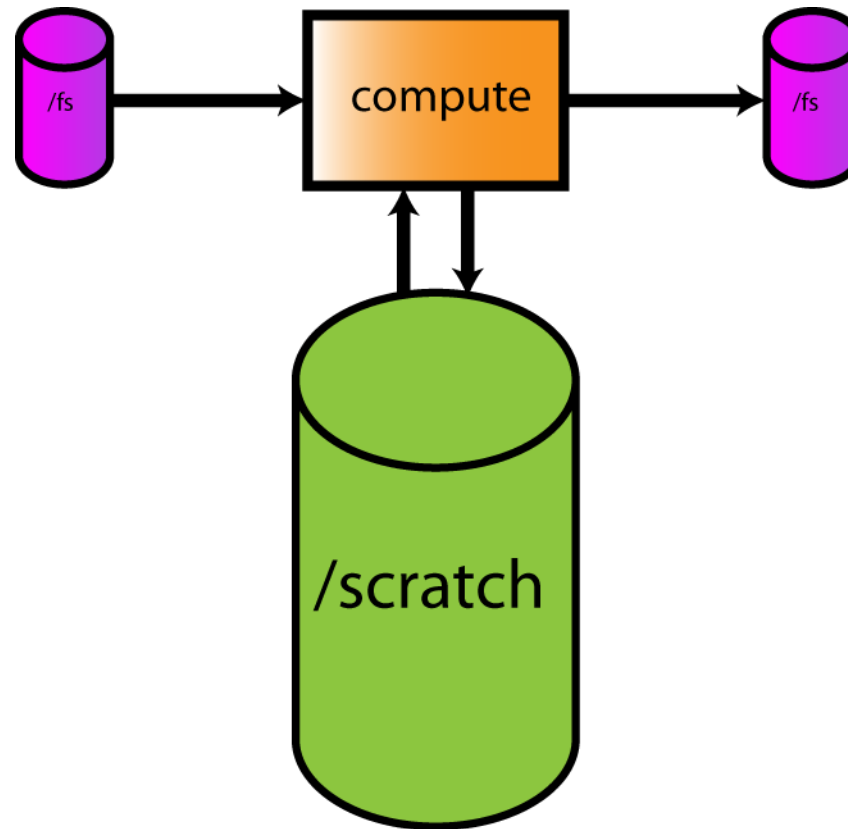
Awards on NSF supported machines are mostly decided by an external allocation committee. At most we can advise the committee on the proper use of the machine.

As a result, even a system that was designed for particular types of problems must still be a general purpose machine that is suitable for a wide range of traditional and novel users

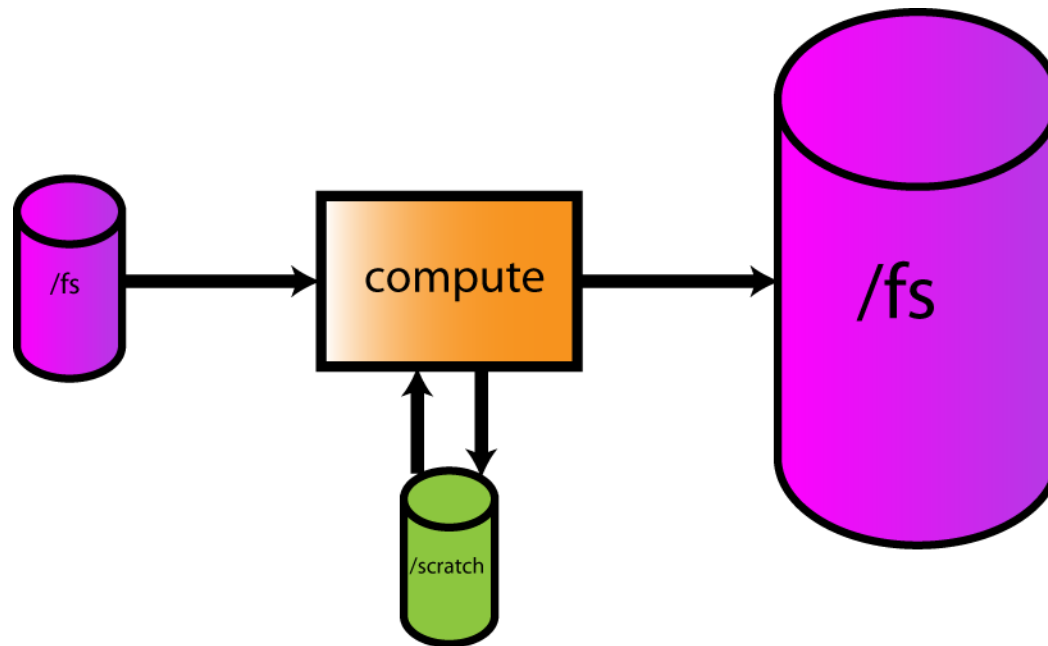
Data analytics and certain types of visualization applications often require processing large amounts of raw data, but can end up producing fairly small amounts of output. In some cases, the result can be single number



Many problems in domains such as graph algorithms, de novo sequence assembly, and quantum chemistry require intermediate files that are disproportionately large relative to the size of the input/output files



Simulations involving integration of ODEs (e.g. molecular dynamics) or PDEs (e.g. CFD, structural mechanics, weather and climate modeling) may involve modest amounts of input data, but end up generating large amounts of output – 4D data sets proportional to problem size x number of time steps



---

# Gordon Hardware overview

- **1024 dual-socket compute nodes**
  - 2 x Intel EM64T Xeon E5 (Sandy Bridge) 2.6 GHz processors
  - 64 GB DDR3-1333 memory
  - 80 GB local Flash memory
  - *64 TB total DRAM, 341 TFlop peak performance*
- **64 dual-socket I/O nodes**
  - 2 x Intel Westmere 2.66 GHz processors
  - 48 GB DDR3-1333 memory
  - 16 x 300 GB Intel 710 (Lyndonville) Solid State Drives (SSD)
  - *300 TB total flash memory*
- **Dual-rail 3D torus InfiniBand QDR (40 Gbit/s) network**
- **4 PB Lustre-based parallel file system**
  - *Capable of delivering up to 100 GB/s to Gordon*

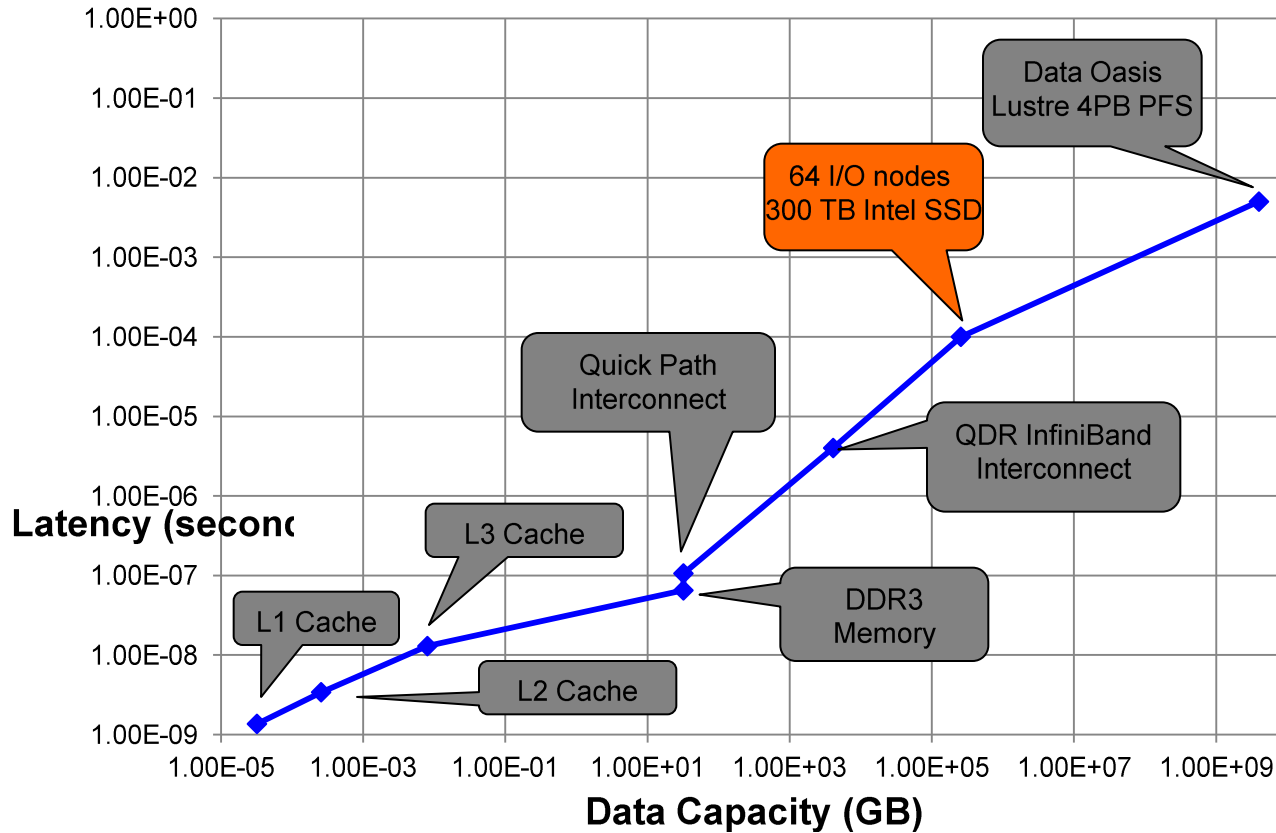


Gordon placed #48 on 11/2011 Top500 list  
(based on 788 nodes)

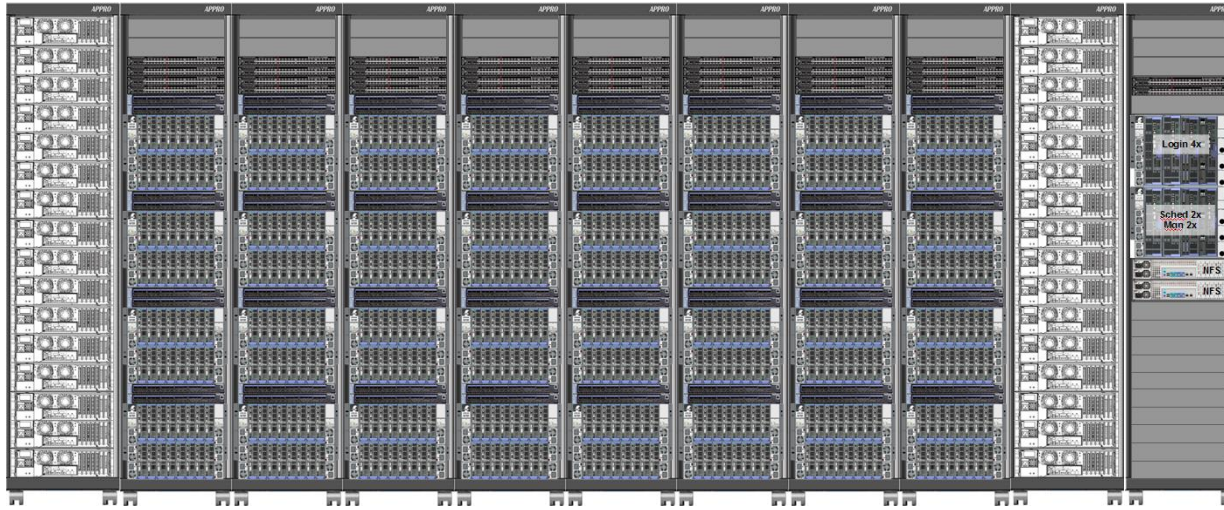
Linpack benchmark give ~ 85% theoretical peak  
285 TF @ 1010 nodes

Rank	Site	Computer/Year Vendor	Cores	R <sub>max</sub>	R <sub>peak</sub>	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	705024	10510.00	11280.38	12659.9
2	National Supercomputing Center in Tianjin China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT	186368	2566.00	4701.00	4040.0
3	DOE/SC/Oak Ridge National Laboratory United States	Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.0
47	Universitaet Aachen/RWTH Germany	Bullx B500 Cluster, Xeon X56xx 3.06Ghz, QDR Infiniband / 2011 Bull	25448	219.84	270.54	
48	UCSD/San Diego Supercomputer Center United States	Xtreme-X GreenBlade GB512X, Xeon E5 (Sandy Bridge - EP) 8C 2.60GHz, Infiniband QDR / 2011 Appro	12608	218.10	262.25	252.2
49	INPE (National Institute for Space Research) Brazil	Cray XT6 12-core 2.1 GHz / 2010 Cray Inc.	30720	205.10	258.05	

# Access to Data Comes with a Latency Penalty



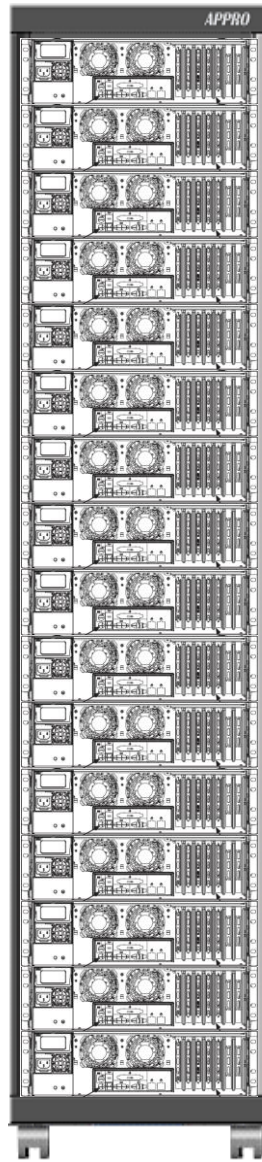
# Gordon Rack Layout



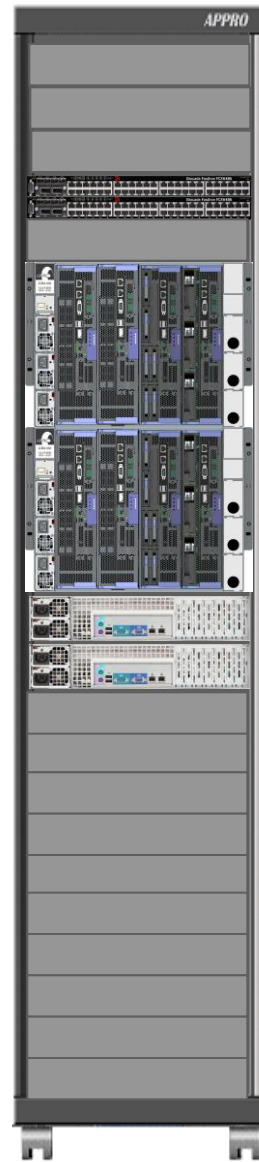
16 compute node racks  
4 I/O node racks  
1 service rack



CN Rack



ION Rack



Service Nodes Rack

Compute node racks:  
4 Appro subracks  
64 blades

ION racks:  
16 Gordon I/O nodes

Service rack:  
4 login nodes  
2 NFS servers  
2 Scheduler nodes  
2 management nodes

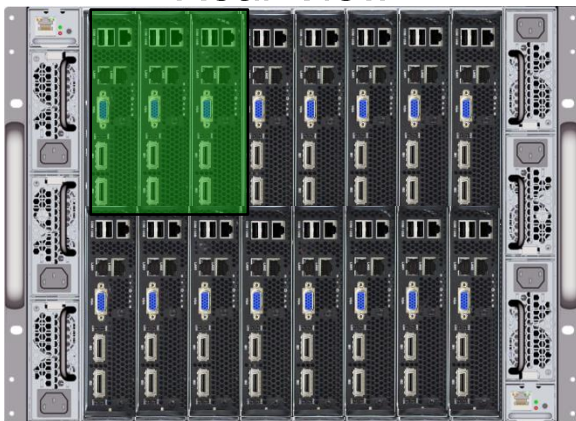
# Based on Appro GreenBlade™ 8000 Series designed for improved reliability and energy efficiency

Front View

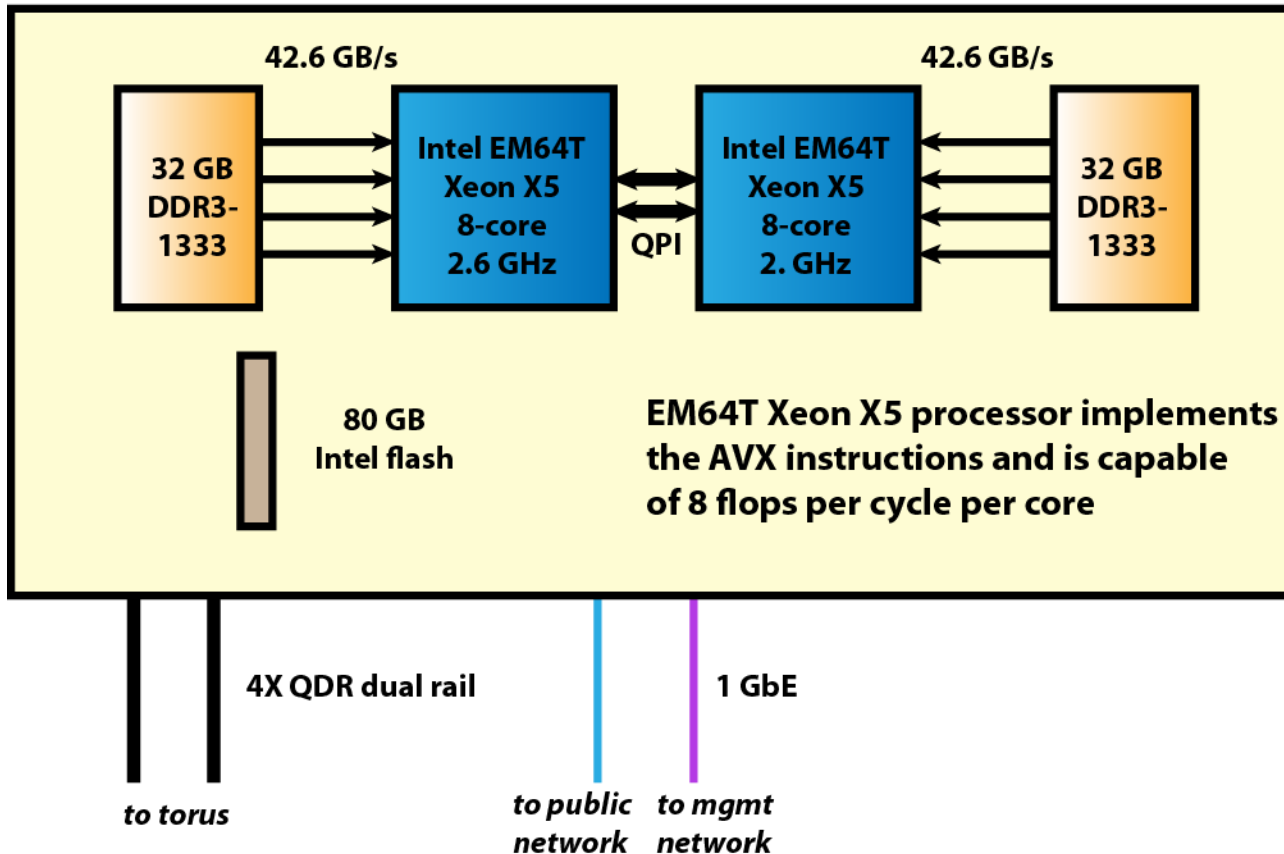


- 5 RU Subrack
- Supports 16 x 2P Intel Sandy-Bridge Blades
- Support for up to six high-efficiency 1625W hot-swappable PS in N+1 configuration
- Support for dual-redundant platform management modules
- Supports six hot-swappable, redundant fan modules
- Shared reduces power consumption by up to 20W per blade over previous design

Rear View



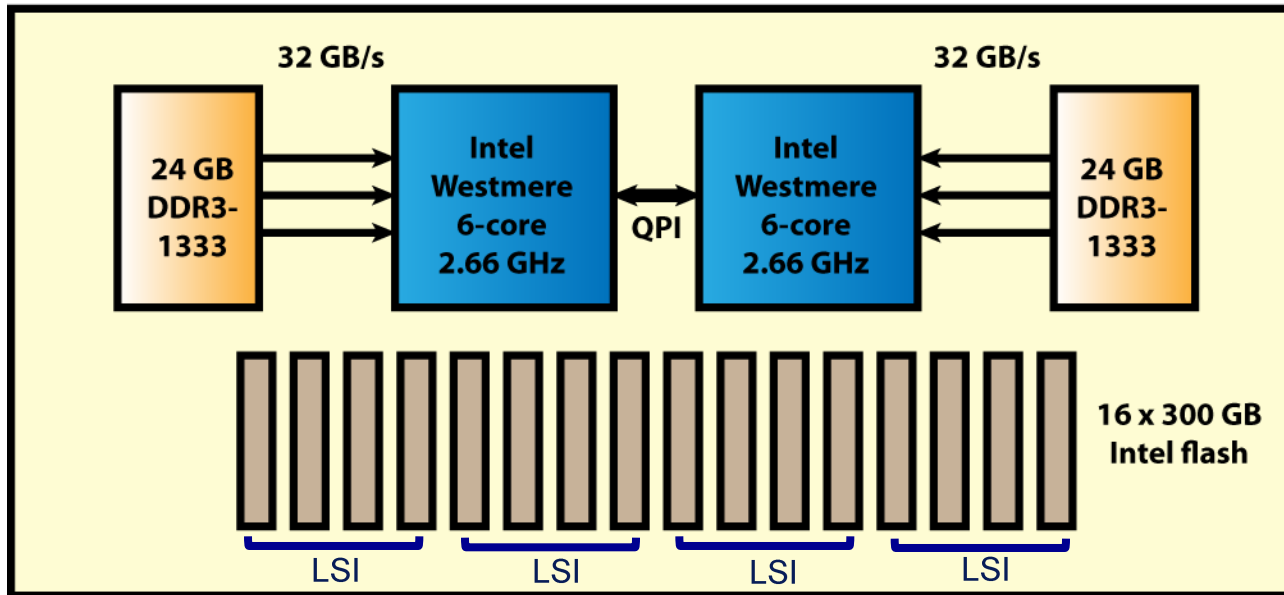
# Gordon compute node



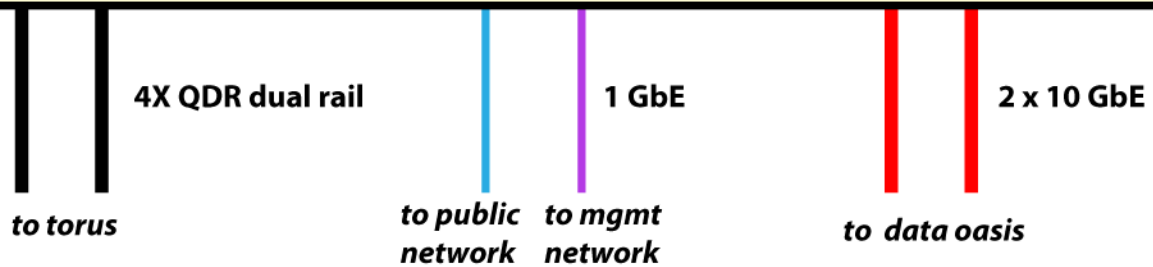
- summary**
- 64 GB DRAM
  - 16 cores
  - 2.6 GHz
  - 80 GB flash

For more information on AVX, see <http://software.intel.com/en-us/avx/>

# Gordon I/O node

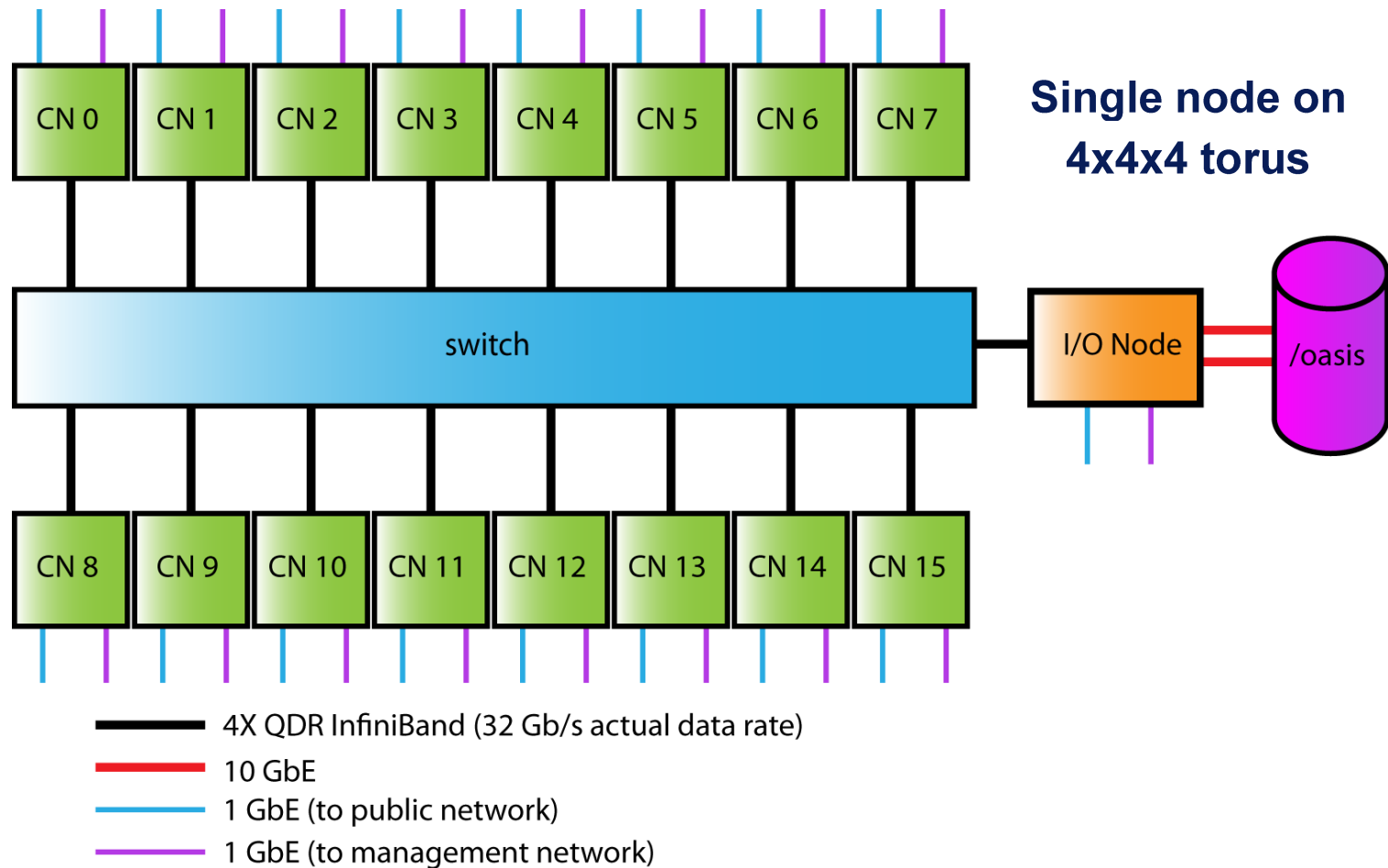


- summary**
- 48 GB DRAM
- 12 cores
- 2.66 GHz
- 4.8 TB flash



Bonded into single channel  
~ 1.6 GB/s bandwidth

Simplified single rail view of Gordon connectivity showing routing between compute nodes on same switch, I/O node, and data oasis.



---

# Exporting flash to compute nodes

Flash will be made available to compute nodes using iSER (iSCSI Extensions for RDMA). Various deployments possible, but will most likely use one of two models

- Export one flash drive to each compute node, with separate XFS file system on each drive
- Configure 16 SSDs on I/O node as RAID0 device, with single shared file system (e.g. OCFS) available to all compute nodes on switch

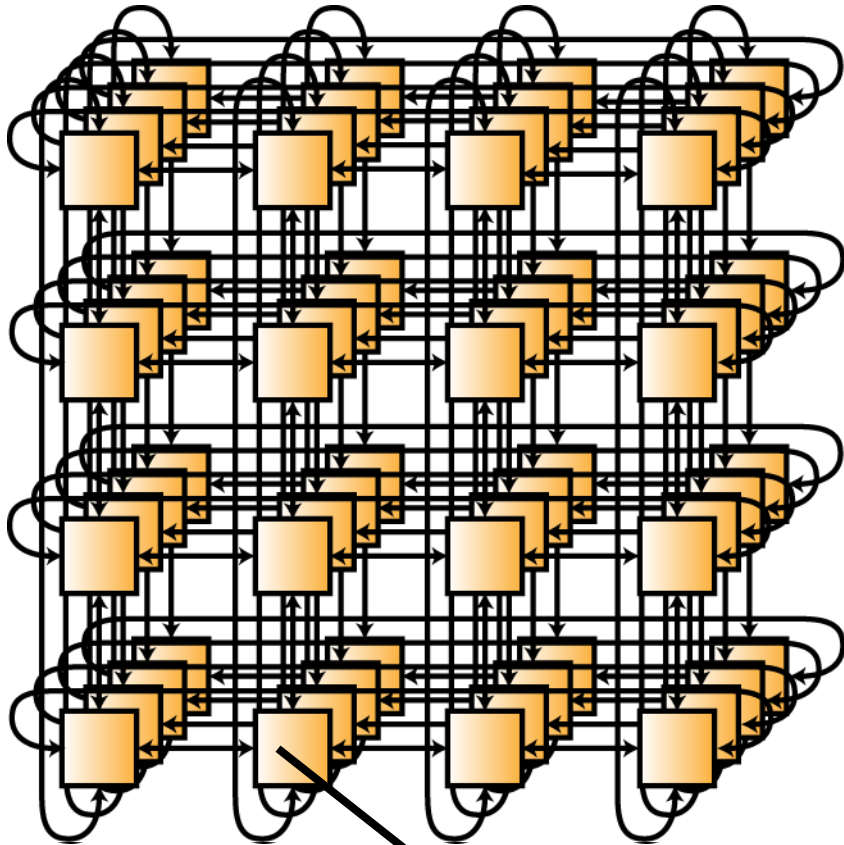
# 3D Torus Interconnect

Gordon switches connected in dual rail 4x4x4 3D torus

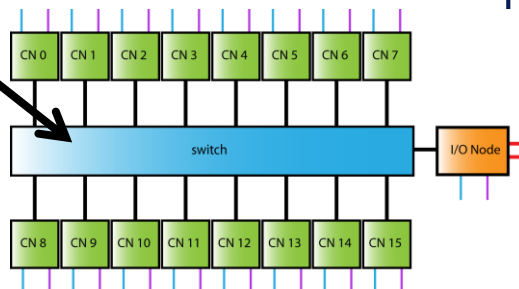
Maximum of six hops to get from one node to furthest node in cluster

Fault tolerant, requires up to 40% fewer switches and 25-50% fewer cables than other topologies

Scheduler will be aware of torus geometry and assign nodes to jobs accordingly



*Note – three 40 Gbit/s connections between neighboring switches, only 1 shown*

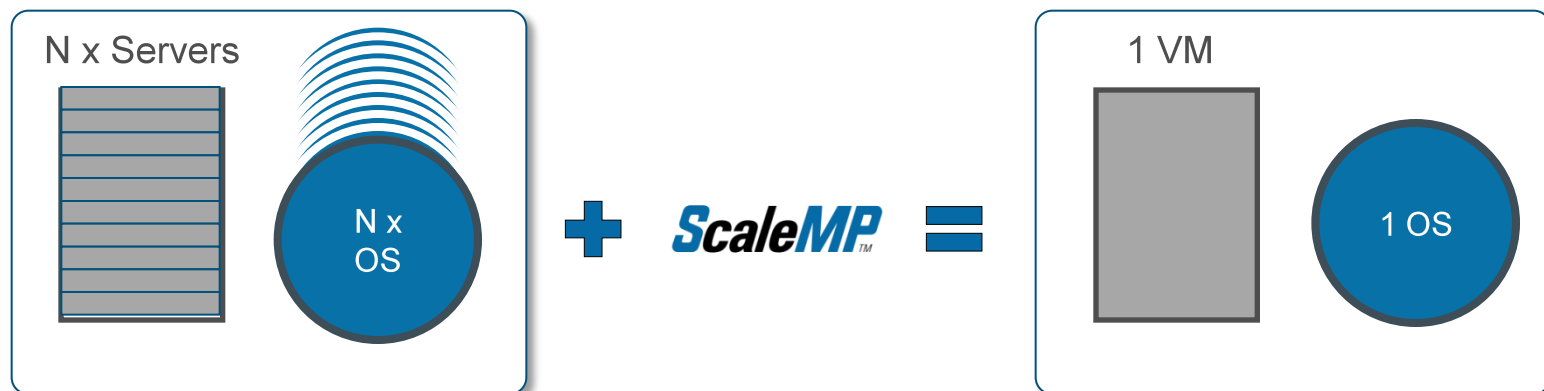


---

## Gordon is a Green System

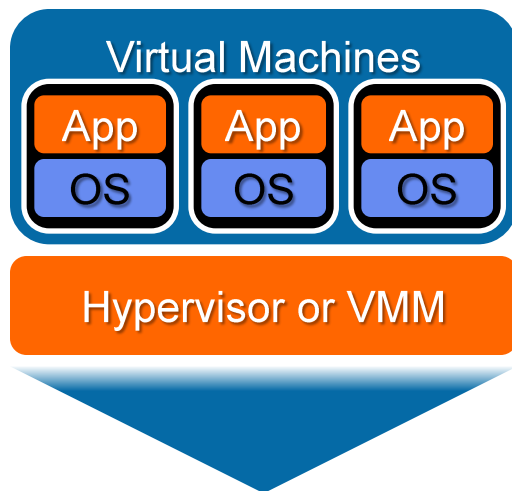
- Compute nodes and login nodes use 80 GB SSDs instead of hard disks as system drives
  - SSD ~ 3-5 W active / 700 mW idle
  - HDD ~ 7-9 W active / slightly less than idle
- 3D Torus requires ~ 40% fewer switches than fat tree
- Appro GreenBlade system packs eight dual-socket Intel Sandy Bridge (E5) nodes into 5U subrack. Efficient power delivery and cooling
- Linpack peak / max power dissipation = 252 TF / 291 kW  
864 MFlop/W → #30 on Top Green 500 (11/2011)

# Introduction to vSMP

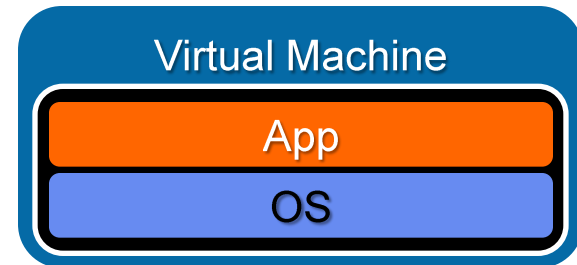


Virtualization **software** for **aggregating** multiple **off-the-shelf** systems into a single virtual machine, providing improved usability and higher performance

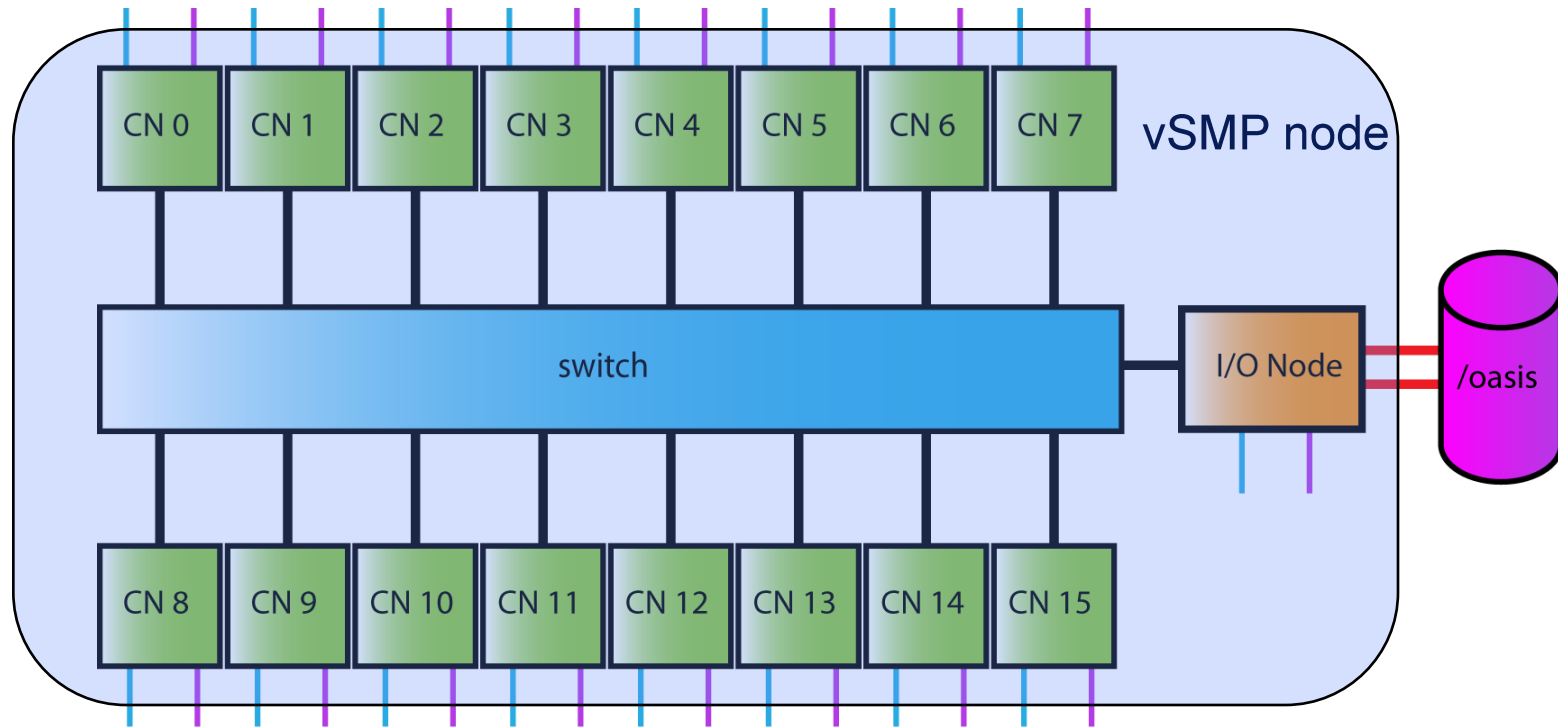
# PARTITIONING



# AGGREGATION

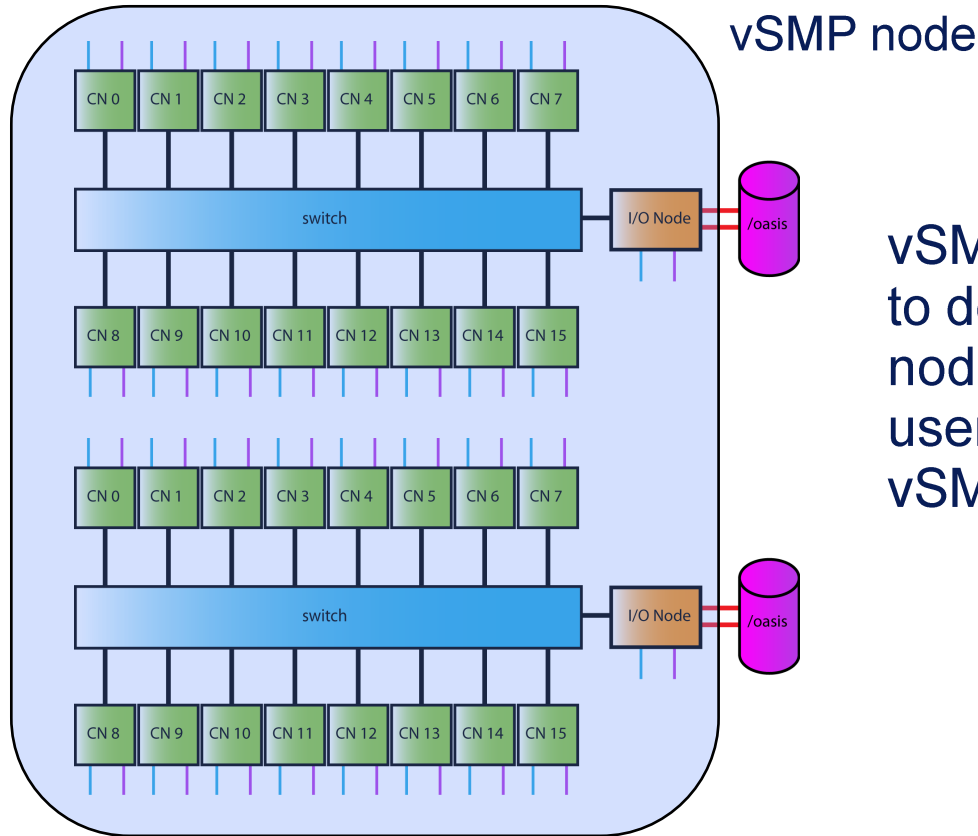


vSMP node configured from 16 compute nodes and one I/O node



To user, logically appears as a single, large SMP node with ~1 TB memory (16 x 64 GB) and 256 compute cores (16 x 16)

## vSMP node configured from 32 compute nodes and two I/O node



vSMP software provides flexibility to deploy logical shared memory nodes in sizes demanded by users. Can potentially configure vSMP nodes on the fly

To user, logically appears as a single, large SMP node with ~2 TB memory (32 x 64 GB) and 512 compute cores (32 x 16)

# Access to Gordon - Academic researchers

To apply, the principal investigator (PI) must be a researcher or educator at a U.S.-based institution, including federal research labs or commercial organizations, though additional information may be needed from researchers not affiliated with academic or non-profit research institutions

- Time awarded on a competitive basis
- Proposals submitted through the Extreme Science and Engineering Discovery Environment (XSEDE)  
<https://www.xsede.org/allocations>
- Startup proposals: 100,000 SU limit / reviewed continuously
- Research proposal: no limit / reviewed quarterly
- Provide strong justification for why you need Gordon (parallelism, memory footprint, vSMP, flash)

# Access to Gordon – Industrial Users



## Research Cyberinfrastructure (RCI)

Making connections is fundamental to scientific research.

RCI offers UC San Diego researchers the computing, network, and human infrastructure needed to create, manage, and share data. Principal investigators are encouraged to use the campus's RCI in addressing federal sponsors' existing and new data management requirements. SDSC's favorable pricing can help researchers justify costs associated with data generation and management—costs that are increasingly included in proposal budgets.

<http://rci.ucsd.edu>

Ron Hawkins, SDSC Director of Industry Relations [rhawkins@sdsc.edu](mailto:rhawkins@sdsc.edu)

---

# A novel allocations model for data-intensive project

Allow projects dedicated use of an I/O node plus 1-4 compute nodes for up to one year.

Users get 4.8 TB of persistent flash storage to provide fast random access to frequently used data sets

Also benefit from direct network connection to I/O node, compute power for data processing, 2 x 10GbE connections to Lustre based parallel file system, and administrative support from SDSC staff

---

# Who's using Gordon

Traditional users (just another machine):

CFD, high energy physics, climate & weather

Plus some that use Gordon's novel features:

Automated annotation of multimedia data

Analysis of Protein Data Bank

Graph problems

Computational chemistry

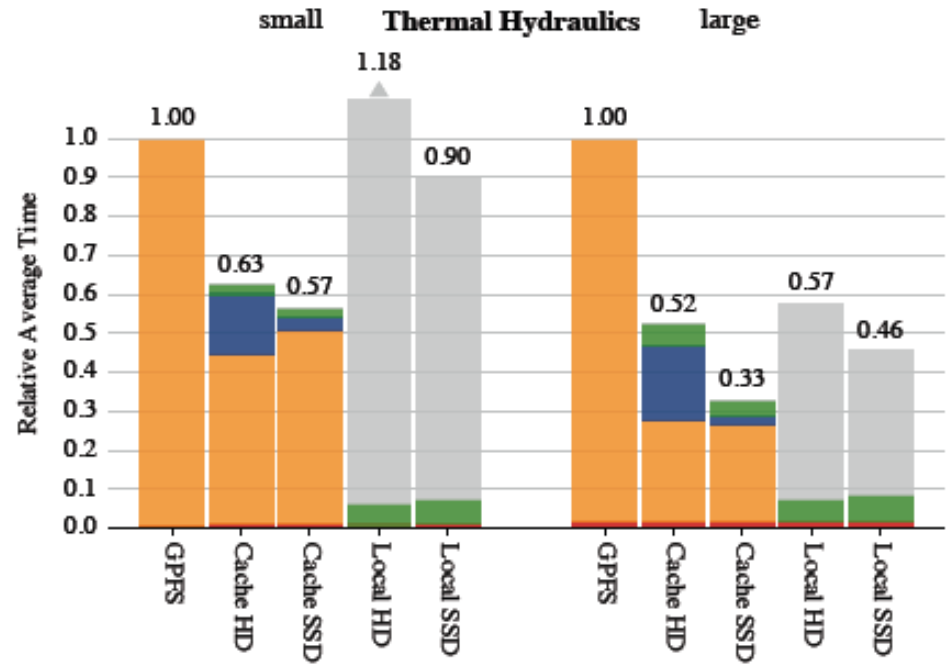
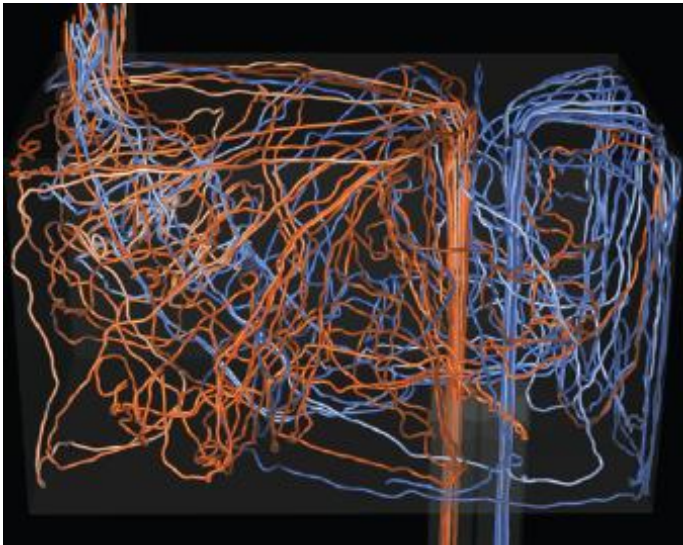
LSST Moving Object Pipeline System

Genome assembly – Velvet, SOAPdenovo

Finance / Impact of high-speed trading

Astroinformatics (workshop at SDSC this summer)

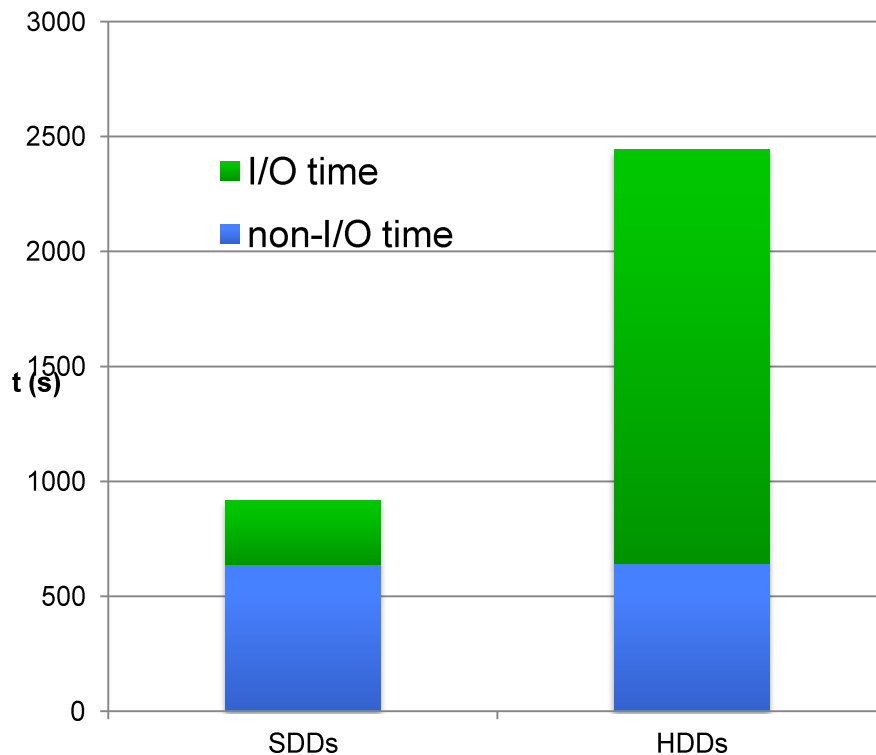
# Flash case study – Parallel Streamline Visualization



Caching data to drives results in better performance than reading directly from GPFS or preloading into local disk. SSDs perform better than HDDs

# Flash case study – Breadth First Search

MR-BFS serial performance  
134217726 nodes



Implementation of Breadth-first search (BFS) graph algorithm developed by Munagala and Ranade

Benchmark problem: BFS on graph containing 134 million nodes

Use of flash drives reduced I/O time by factor of 6.5x. As expected, no measurable impact on non-I/O operations

Problem converted from I/O bound to compute bound

# Gordon Team

## **SDSC**

Mike Norman – PI  
Allan Snaveley – co-PI  
Shawn Strande – Project Manager  
Bob Sinkovits – Applications Lead  
Mahidhar Tatineni – User support / applications  
Jerry Greenberg – Applications (chem, MATLAB)  
Pietro Cicotti – Applications & benchmarking  
Wayne Pfeiffer – Applications (genomics)  
Jeffrey Bennett – Storage Engineer  
Eva Hocks – Systems Administration  
William Young - Systems  
Chaitan Baru – Database applications  
Kenneth Yoshimoto – Scheduling (Catalina)  
Susan Rathbun – Project Coordinator  
Diane Baxter - EOT  
Jim Ballew – acceptance testing and design  
Amit Majumdar – ASTA  
Nancy Wilkins – Science Portals

## **UCSD**

Steve Swanson  
Adrian Caulfield  
Jiahua He (now at Amazon)  
Meenakshi Bhaskaran

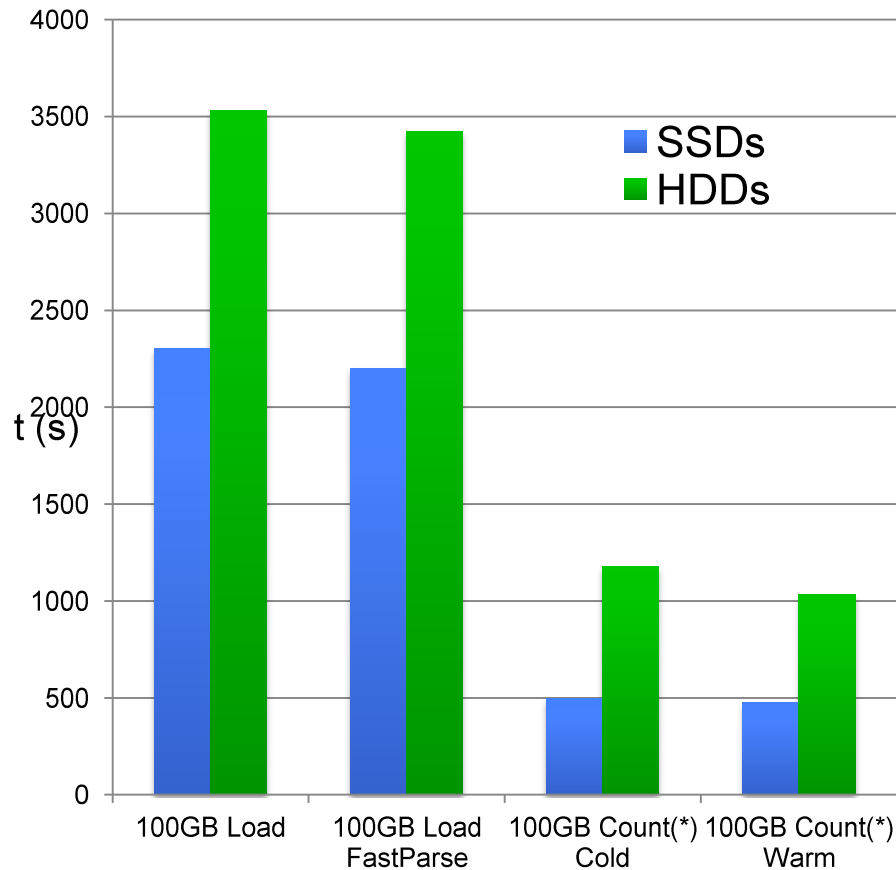
## **ScaleMP**

Nir Paikowsky  
(and many others)

## **Appro**

Steve Lyness  
Greg Faussette  
Adrian Wu  
Roland Wong

# Flash case study – LIDAR

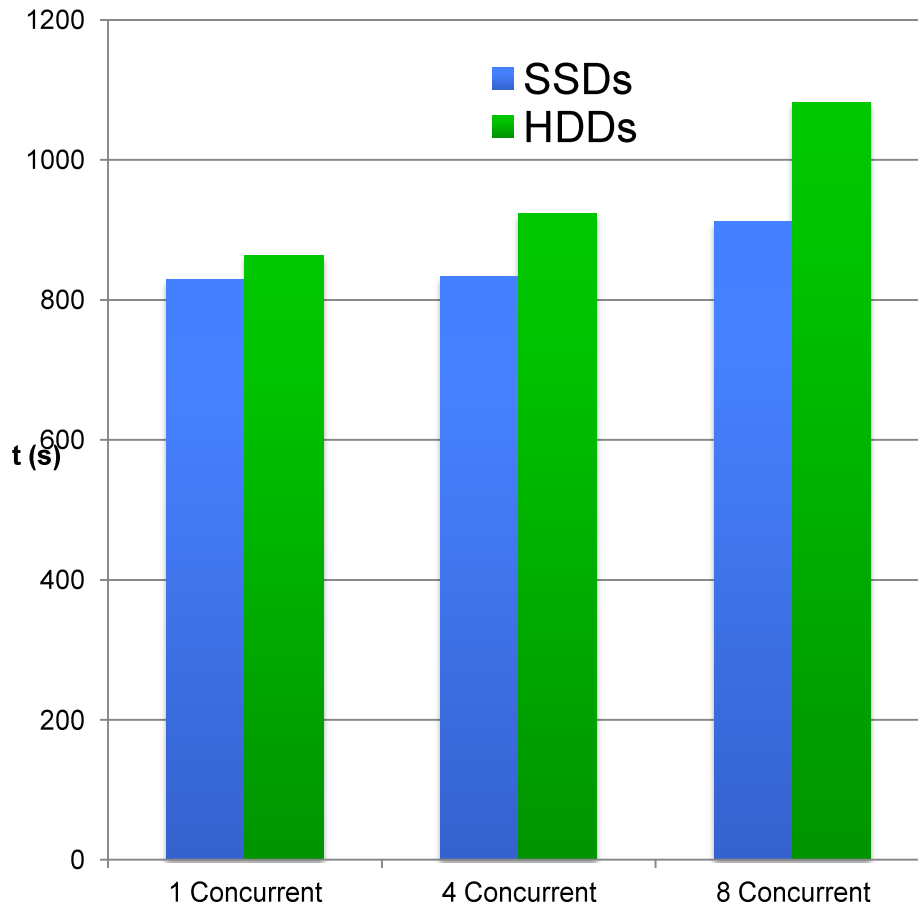


Remote sensing technology used to map geographic features with high resolution

Benchmark problem: Load 100 GB data into single table, then count rows. DB2 database instance

Flash drives 1.5x (load) to 2.4x (count) faster than hard disks

## Flash case study – LIDAR

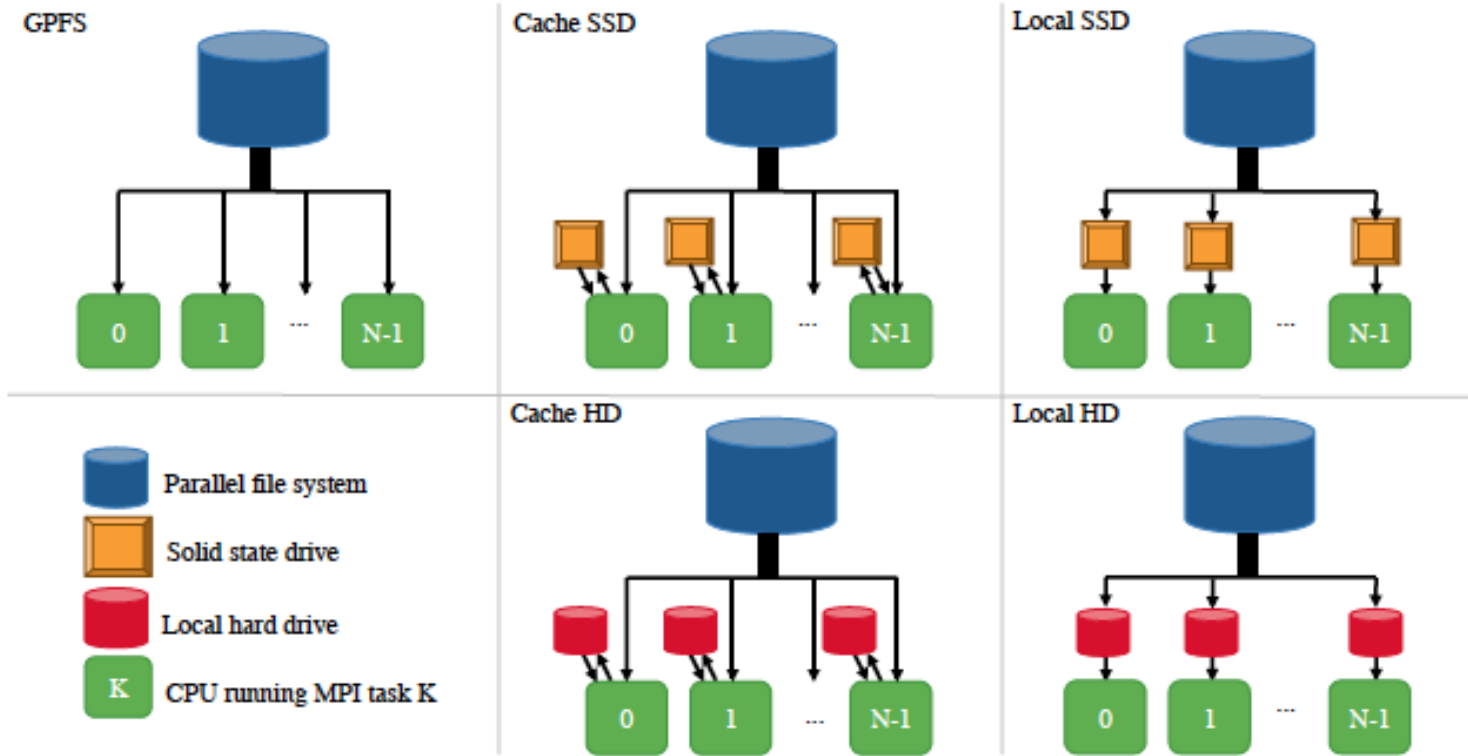


Remote sensing technology used to map geographic features with high resolution

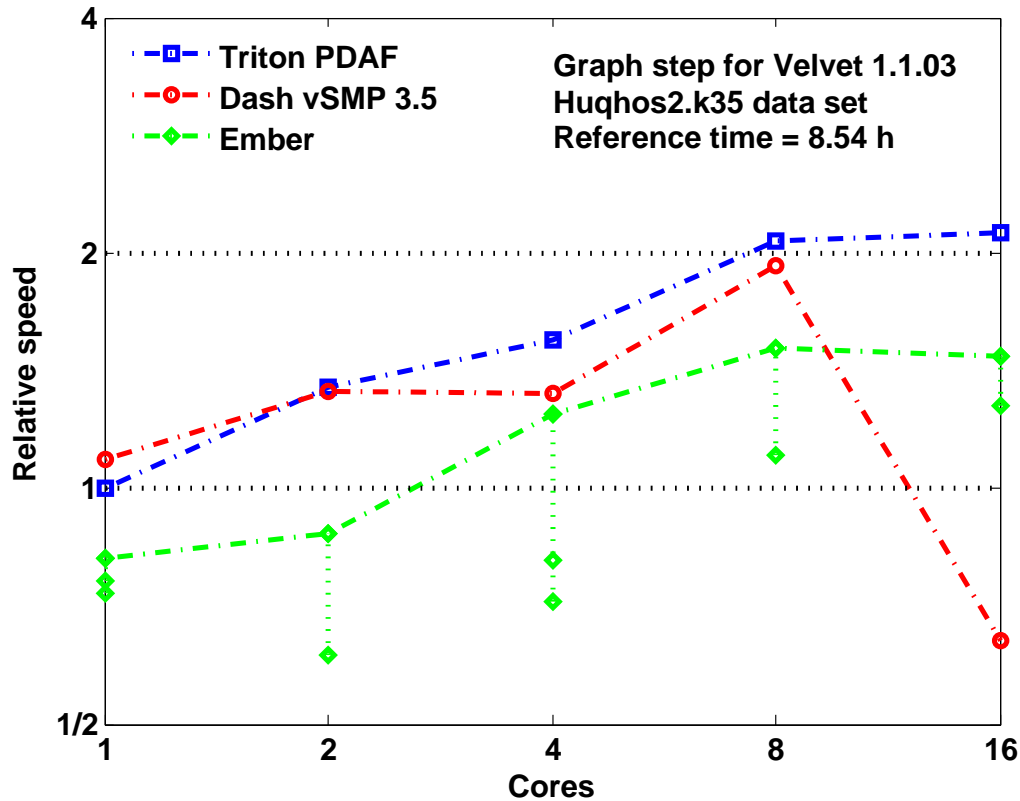
Comparison of runtimes for concurrent LIDAR queries obtained with flash drives (SSD) and hard drives (HDD) using the Alaska Denali-Totschunda data collection.

Impact of SSDs was modest, but significant when executing multiple simultaneous queries

# Flash case study – Parallel Streamline Visualization



# vSMP case study – Velvet (genome assembly)

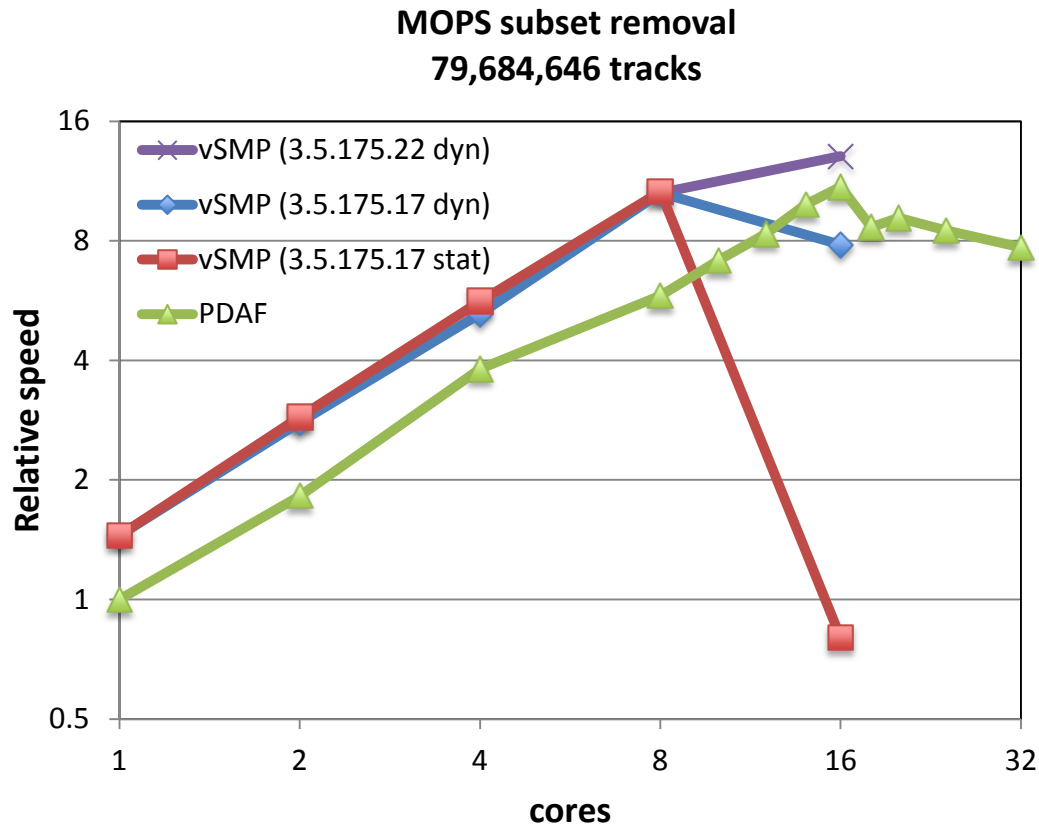


Total memory usage ~ 116 GB (3 boards)

De novo assembly of short DNA reads using the de Bruijn graph algorithm. Code parallelized using OpenMP directives.

Benchmark problem: Daphnia genome assembly from 44-bp and 75-bp reads using 35-mer

# vSMP case study – MOPS (subset removal)



Total memory usage ~ 100 GB (3 boards)

Sets of detections collected using the Large Synoptic Survey Telescope are grouped into tracks representing potential asteroid orbits

Subset removal algorithm used to identify and eliminate those tracks that are wholly contained within other tracks

7.3x speedup on 8 cores is better than that obtained on large shared memory node. Dynamic thread scheduling mitigates impact of using CPUs off board.

Although preloading entire data set into flash typically takes longer than just reading from GPFS, still worth doing if multiple visualizations will be performed while data is in flash

