Gordon: A novel high performance computing system for data and memory intensive applications

Robert Sinkovits
Gordon Applications Lead
San Diego Supercomputer Center

OREGON
Klamath Lake
Goose Lake
Shasta Lake
Eagle Lake
Sacramento
Pyramid Lake
Reno
Lake Tahoe
CARSON CITY
Rosa
SACRAMENTO
NEVADA
Oakland
ncisco
Stockton
San Jose
Mono Lake
Monterey Bay
Salinas
Fresno
Owens Lake
CALIFORNIA
Lake Mead
Bakersfield
Las Veg
Point Arguello
Los Angeles
Riverside
Channel Islands
Long Beach
Salton Sea
Colorado
San Diego
Tijuana
MEX.
MEXICALI

SDSC's mission is to transform science and society at UC San Diego and across the nation through world-leading cyberinfrastructure innovation, development and expertise

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

Trend in U.S. academic computing toward large node counts, with smaller memory per node

machine	peak (TF)	nodes	mem (TB)	mem/node (GB)
Kraken	1174	9408	147	16
Ranger	579	3936	123	32
Lonestar4	302	1888	44	24
Gordon	341	1024	64	64 (1024, 2048, …)
Athena	166	4152	16	4
Trestles	100	324	20	64
Steele	66	893	28	16-32
Condor Pool	60	1750	27	0.5-32
Lincoln	48	192	3	16
Blacklight	37	256	32	128 (16384)

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

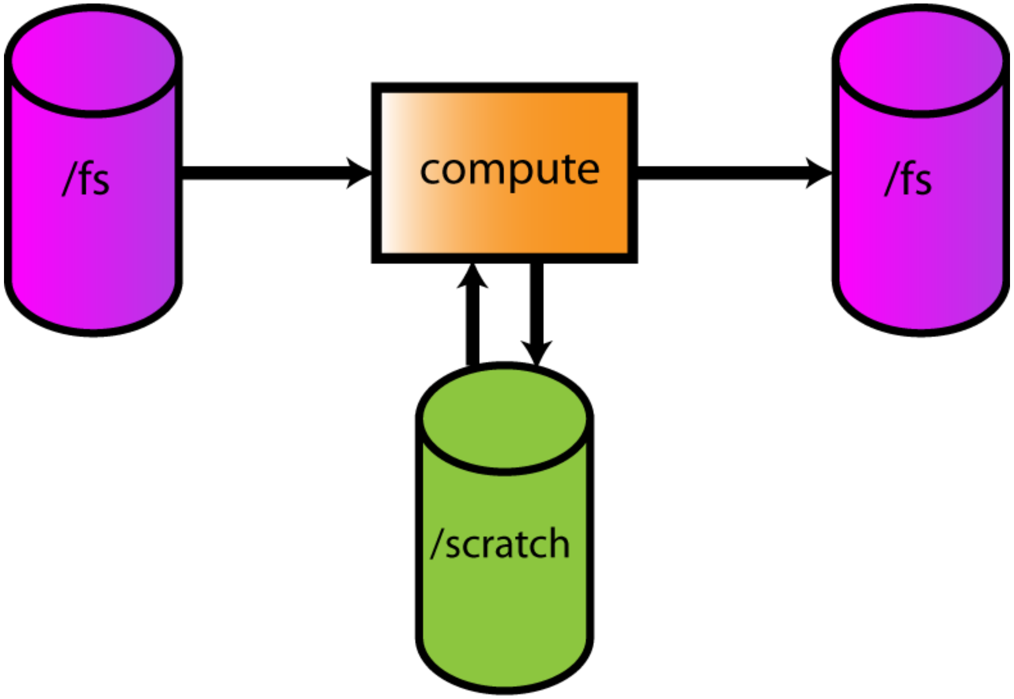
Available to all U.S. academic researchers on a competitive basis and on a limited basis for-fee to most non-academic users (+ foreign users in collaboration with U.S. PIs)

UCSD SDSC	Design Deployment Support	intel	Processors Motherboards Flash drives
APPRO HPC Cluster Solutions	Integrator	ScaleMP™	vSMP Foundation
NSF	Funding OCI #0910847	Mellanox TECHNOLOGIES	3D Torus

Gordon Hardware overview

- **1024 dual-socket compute nodes**
  - 2 x Intel EM64T Xeon E5 (Sandy Bridge) 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 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***

Data intensive computing

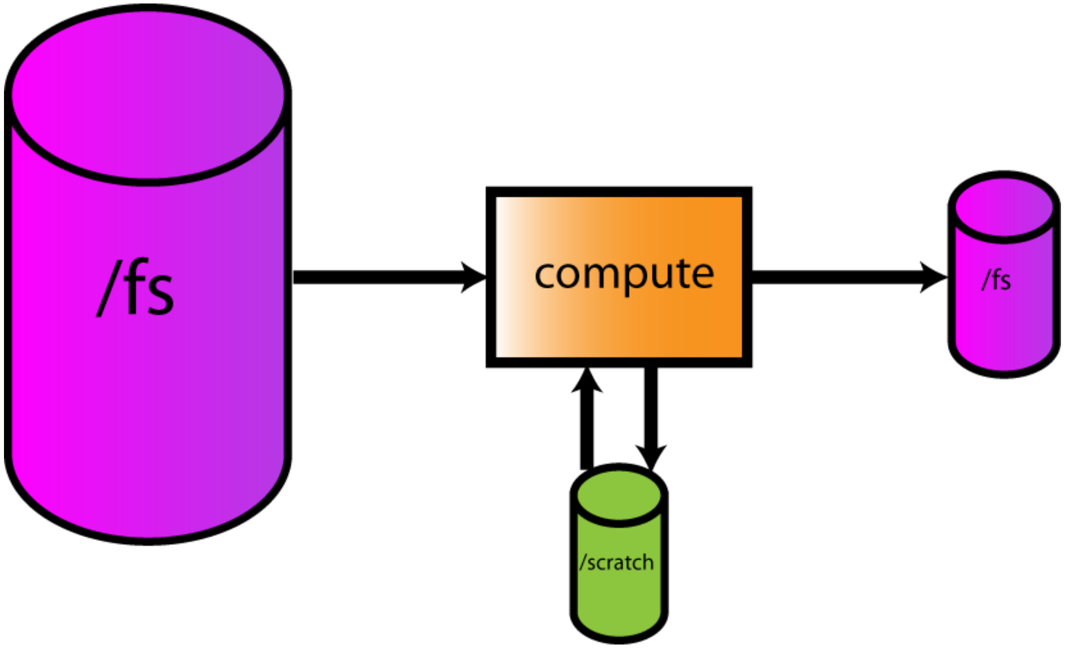


Data intensive problems can be characterized by the sizes of the input, output, and intermediate data sets

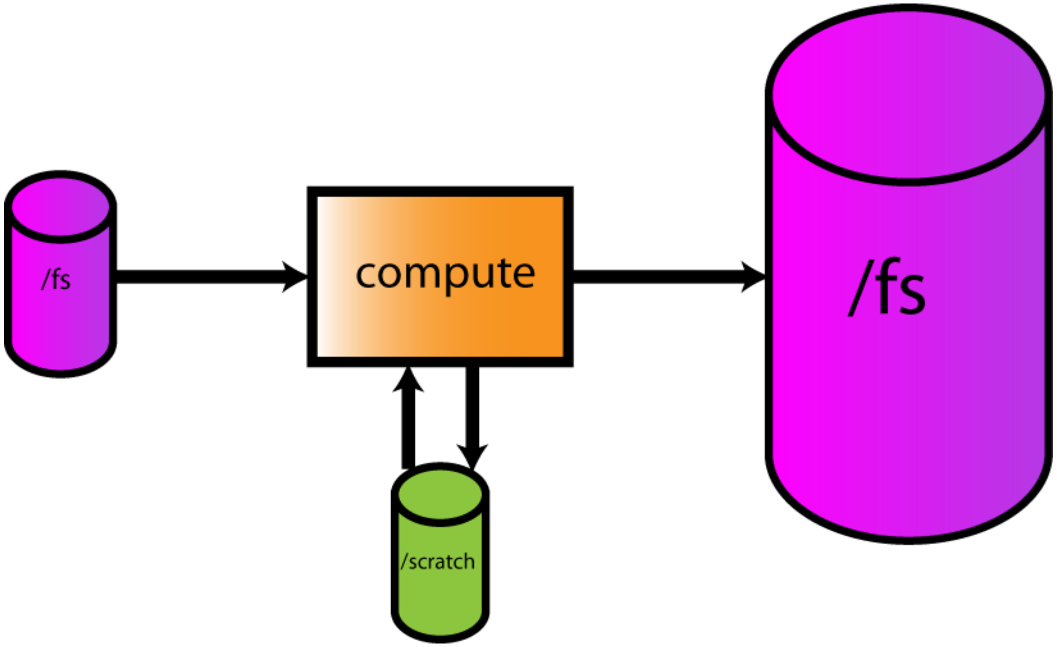
Can also be classified according to patterns of data access (e.g. random vs. sequential, small vs. large reads/writes)

Performance can be improved through changes to hardware, systems software, file systems, or user application

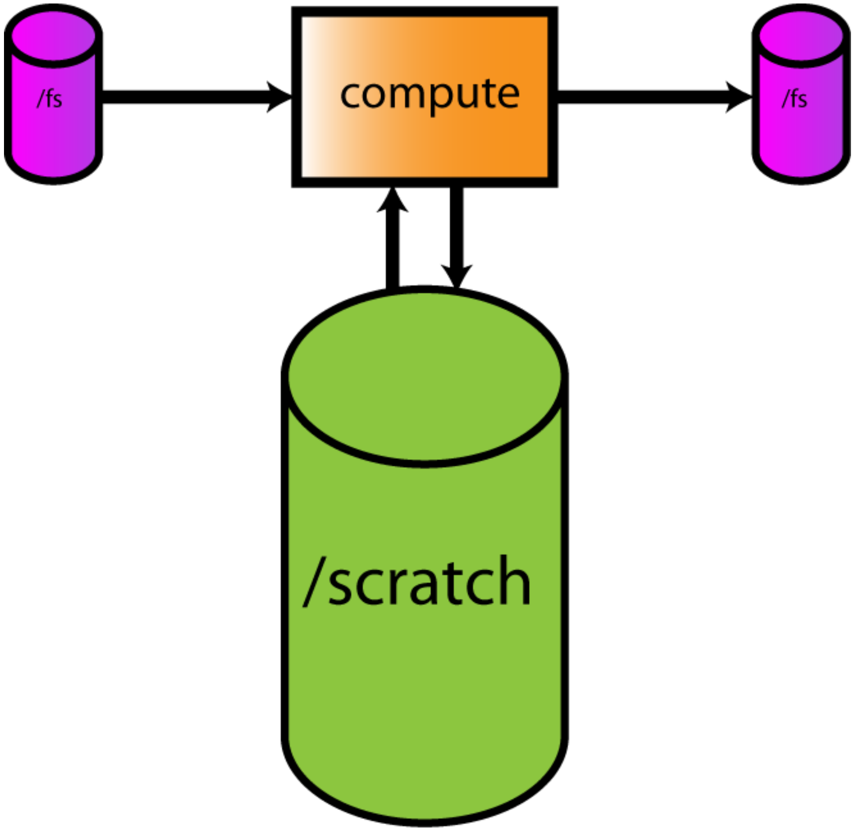
Data mining and certain types of visualization applications often require processing large amounts of raw data, but can end up producing fairly small amounts of output. In some cases, the result can be single number



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



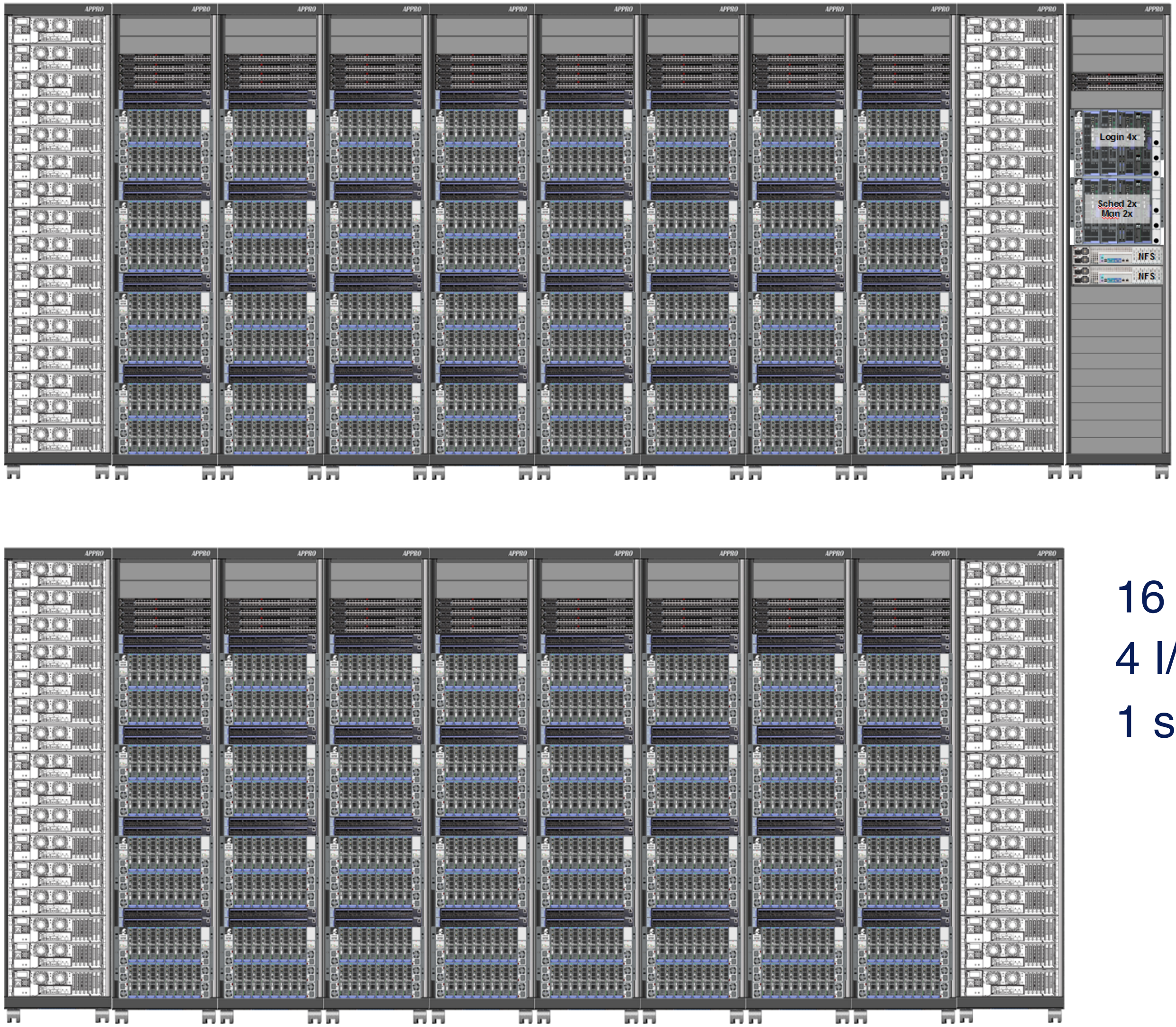
Many problems in domains such as graph algorithms, de novo sequence assembly, and quantum chemistry require intermediate files that are disproportionately large relative to the size of the input/output files



A great Gordon application will …

- Make use of the flash storage for scratch/staging
  64 I/O nodes each w/ 16 x 300 GB SSDs (4.4 TB usable)
- Need the large, logical shared memory
  ~ 1 TB DRAM & 256 cores, w/ larger configurations possible
- Be a threaded app that scales to very large number of cores
- Require large physical memory per node
  64 GB/node, 4 GB/core
- Be able to use the AVX instructions (8 flops/cycle/core)
- Need a high-bandwidth, low-latency inter-processor network

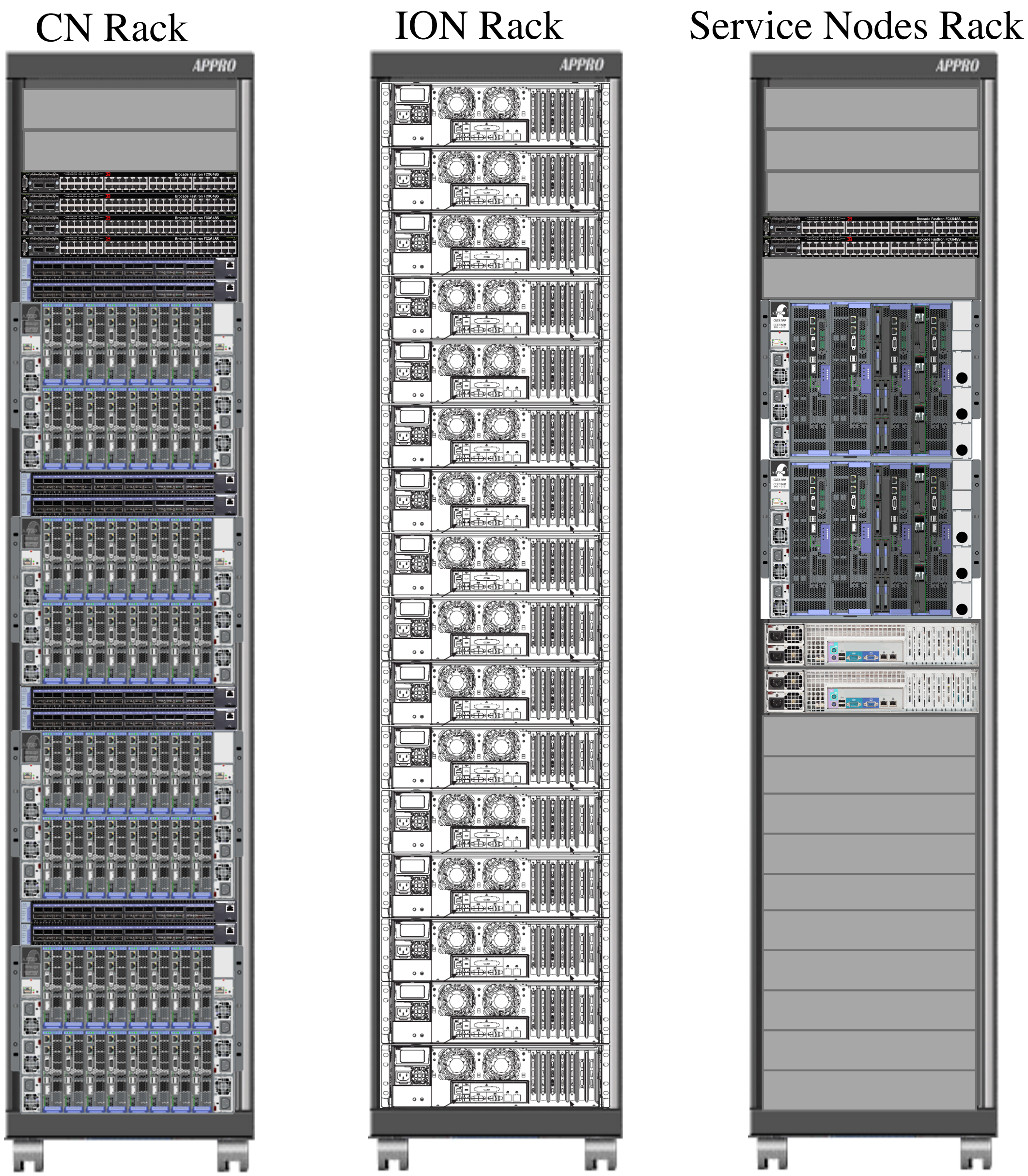
Gordon Rack Layout



16 compute node racks

4 I/O node racks

1 service rack



Compute node racks:
4 Appro subracks
64 blades

ION racks:
16 Gordon I/O nodes

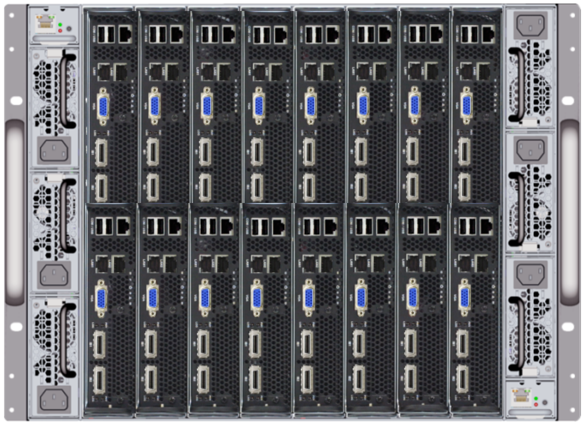
Service rack:
4 login nodes
2 NFS servers
2 Scheduler nodes
2 management nodes

Based on Appro GreenBlade™ 8000 Series

designed for improved reliability and energy efficiency

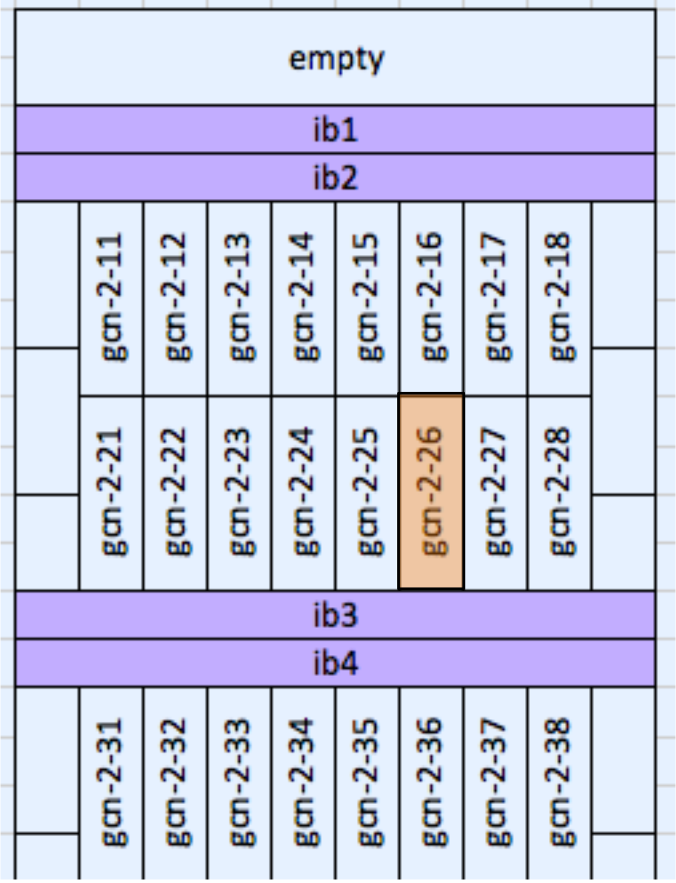
Front View

Rear View



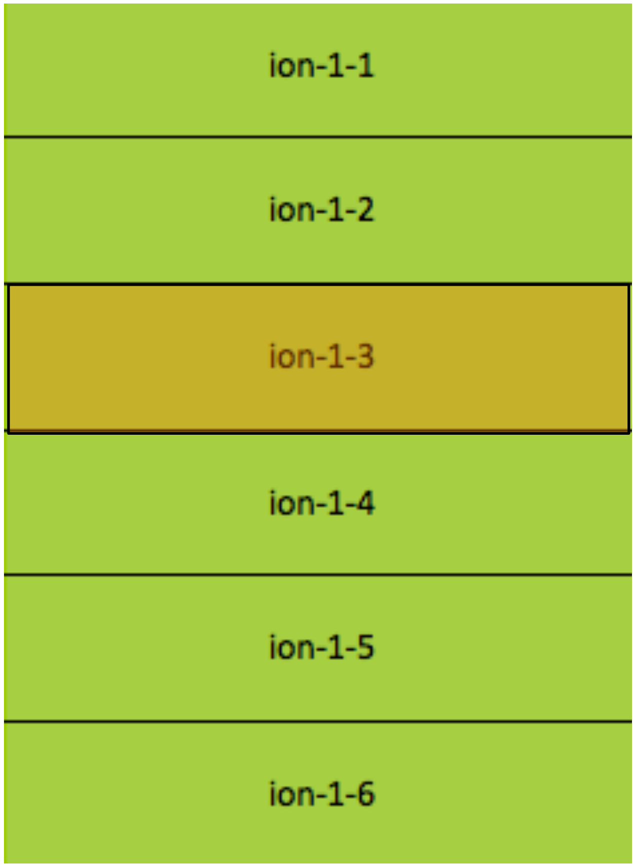
- 8RU Subrack
- Supports 16x 2P Intel Sandy-Bridge Blades
- Support for up to six high-efficiency 1625W hot-swappable PS in N+1 configuration
- Support for dual-redundant platform management modules
- Supports six hot-swappable, redundant fan modules
- Shared reduces power consumption by up to 20W per blade over previous design

Gordon naming conventions



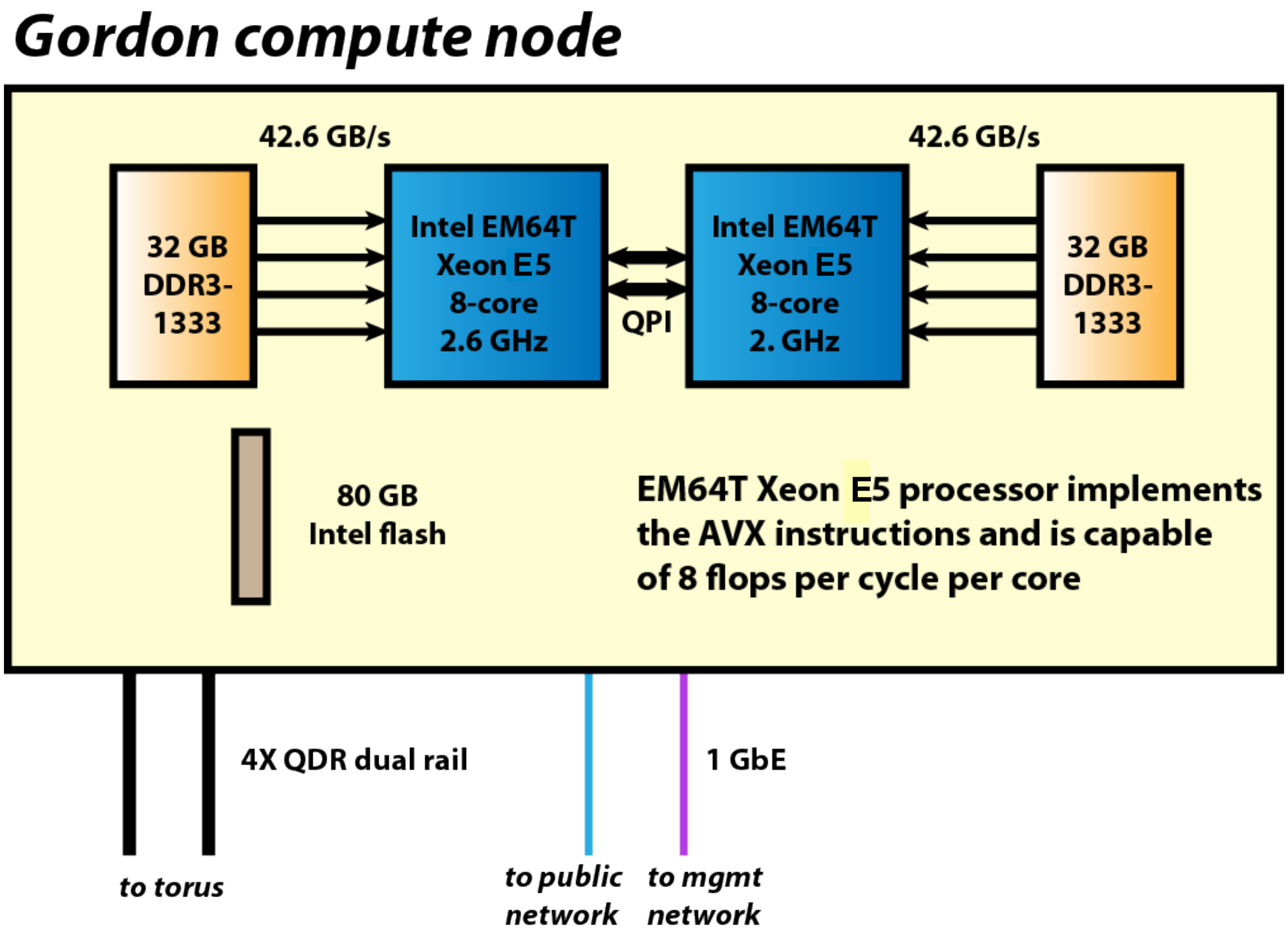
empty							
ib1							
ib2							
gcn-2-11	gcn-2-12	gcn-2-13	gcn-2-14	gcn-2-15	gcn-2-16	gcn-2-17	gcn-2-18
gcn-2-21	gcn-2-22	gcn-2-23	gcn-2-24	gcn-2-25	gcn-2-26	gcn-2-27	gcn-2-28
ib3							
ib4							
gcn-2-31	gcn-2-32	gcn-2-33	gcn-2-34	gcn-2-35	gcn-2-36	gcn-2-37	gcn-2-38

Gordon Compute Node
rack 2, row 2, node 6



ion-1-1
ion-1-2
ion-1-3
ion-1-4
ion-1-5
ion-1-6

I/O Node
rack 1, node 3


Gordon compute node
42.6 GB/s
42.6 GB/s
32 GB DDR3-1333
Intel EM64T Xeon E5 8-core 2.6 GHz
QPI
Intel EM64T Xeon E5 8-core 2. GHz
32 GB DDR3-1333
80 GB Intel flash
EM64T Xeon E5 processor implements the AVX instructions and is capable of 8 flops per cycle per core
4X QDR dual rail
1 GbE
to torus
to public network
to mgmt network

summary

64 GB DRAM

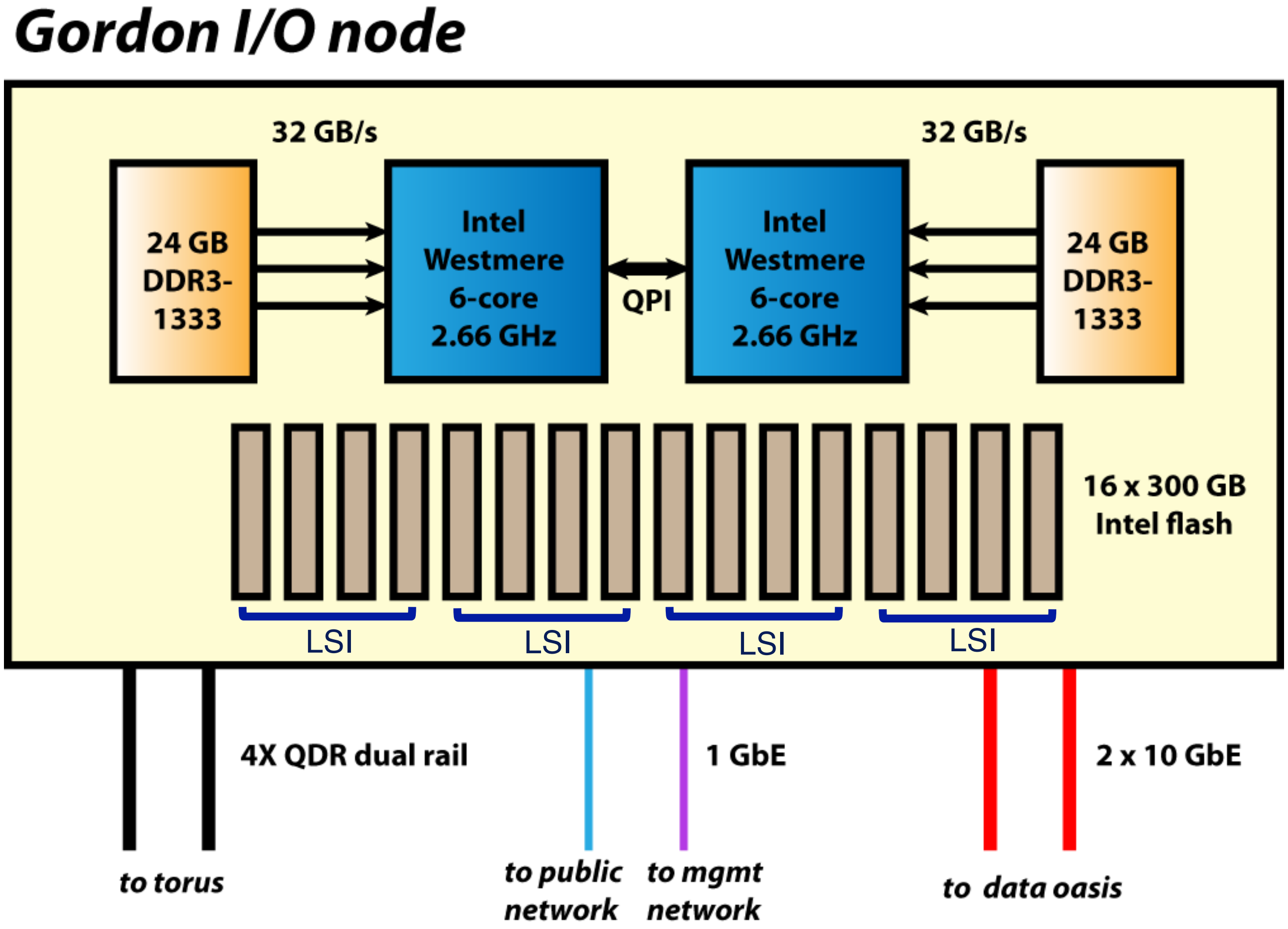
16 cores

2.6 GHz

80 GB flash

333 GFlops

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



summary

48 GB DRAM

12 cores

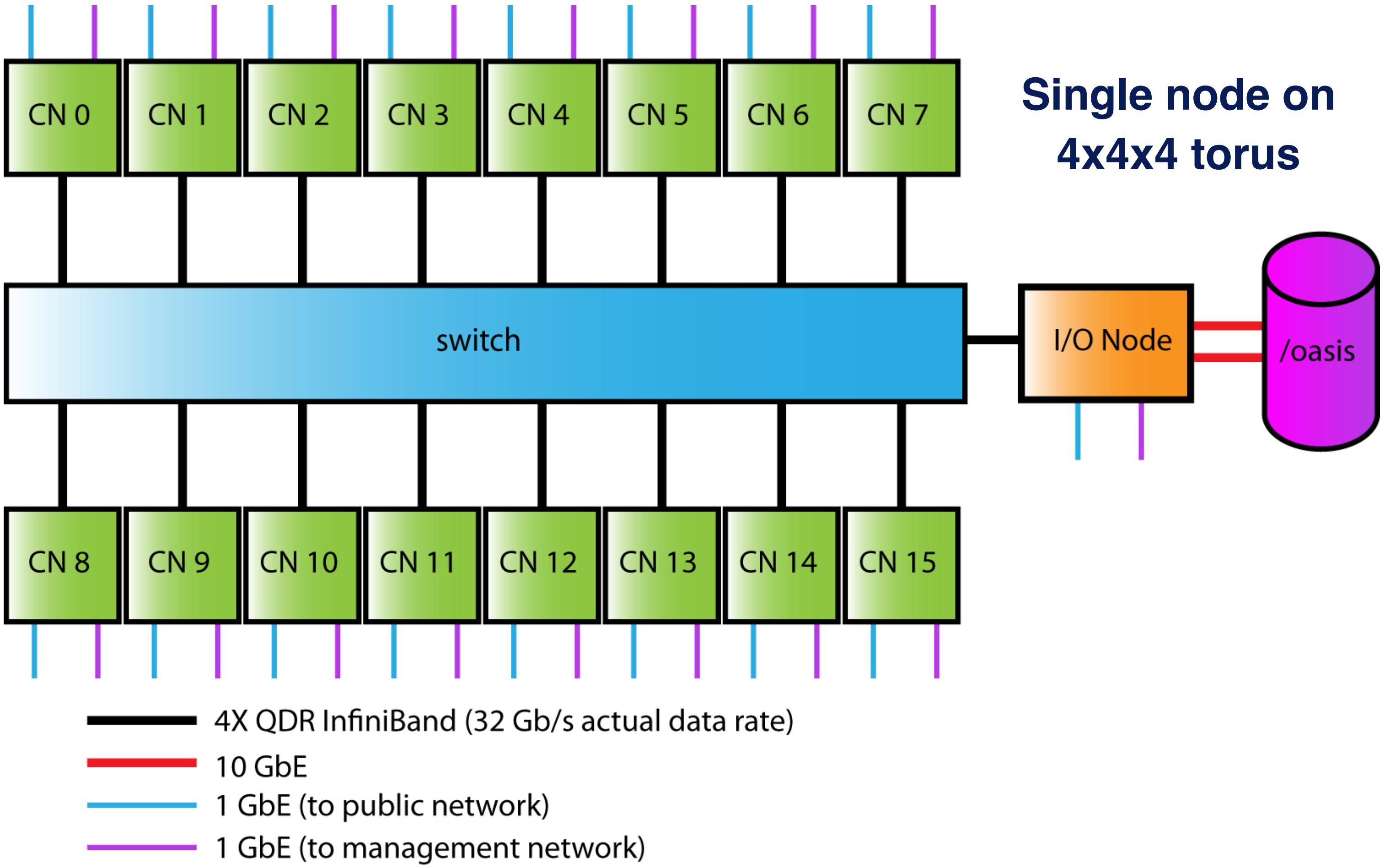
2.66 GHz

4.8 TB flash

Bonded into single channel

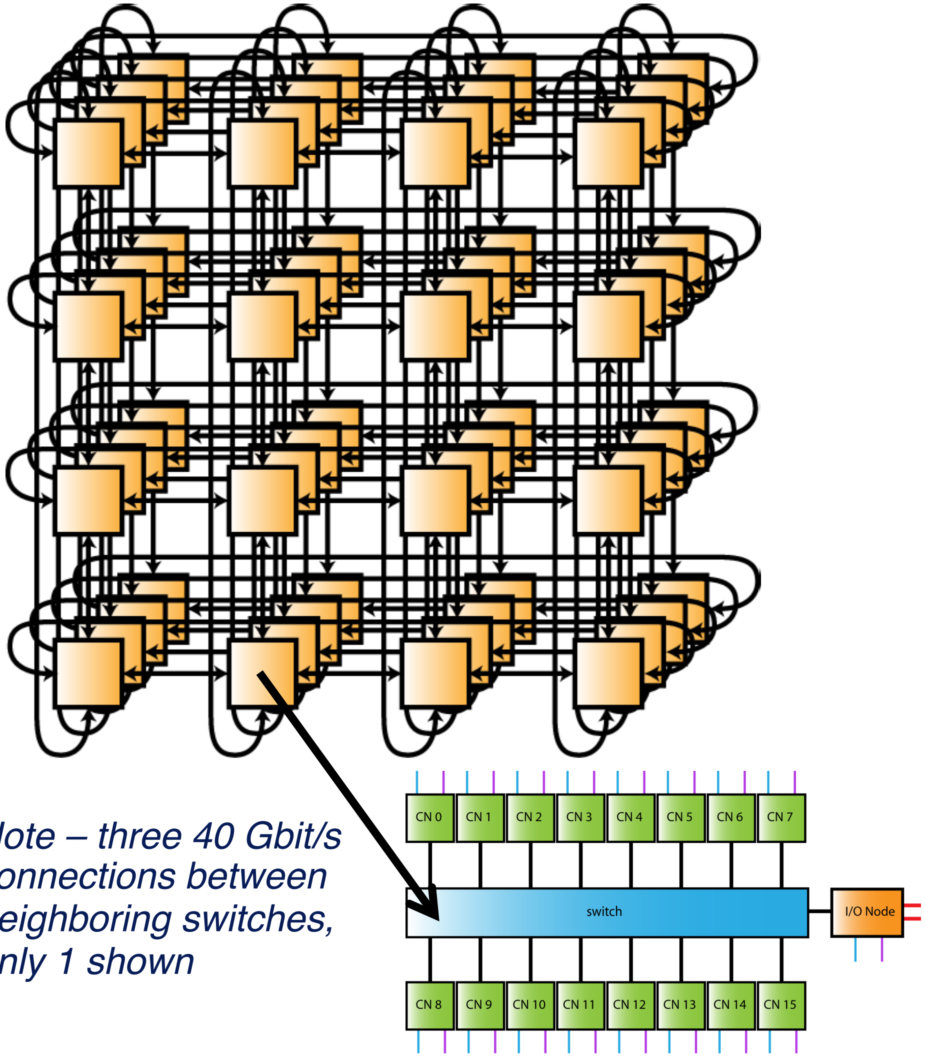
~ 1.6 GB/s bandwidth

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


CN 0
CN 1
CN 2
CN 3
CN 4
CN 5
CN 6
CN 7
switch
I/O Node
/oasis
CN 8
CN 9
CN 10
CN 11
CN 12
CN 13
CN 14
CN 15
4X QDR InfiniBand (32 Gb/s actual data rate)
10 GbE
1 GbE (to public network)
1 GbE (to management network)

Single node on 4x4x4 torus

3D Torus Interconnect



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

Gordon switches connected in dual rail 4x4x4 3D torus

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

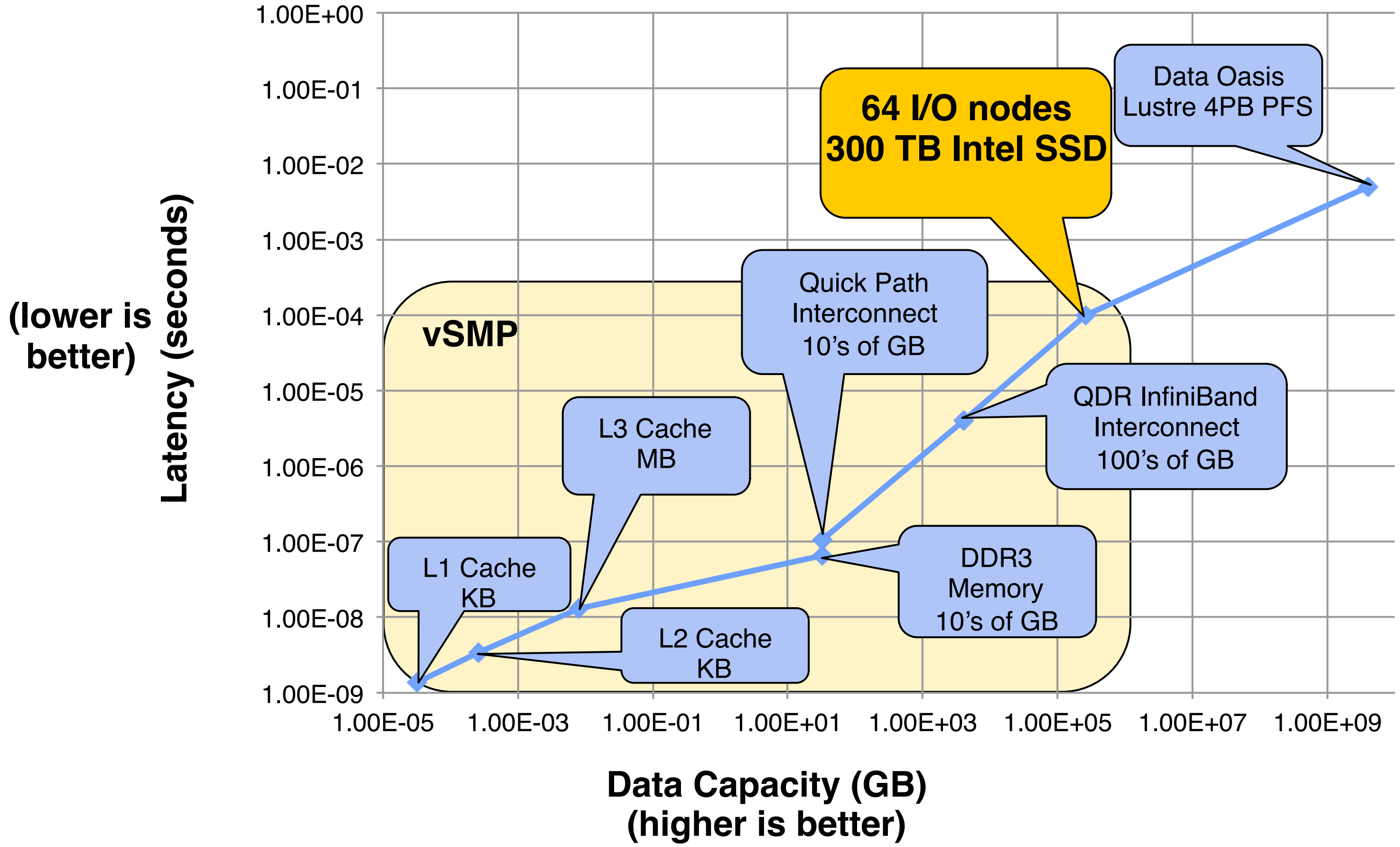
Fault tolerant, requires up to 40% fewer switches and 25-50% fewer cables than other topologies

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

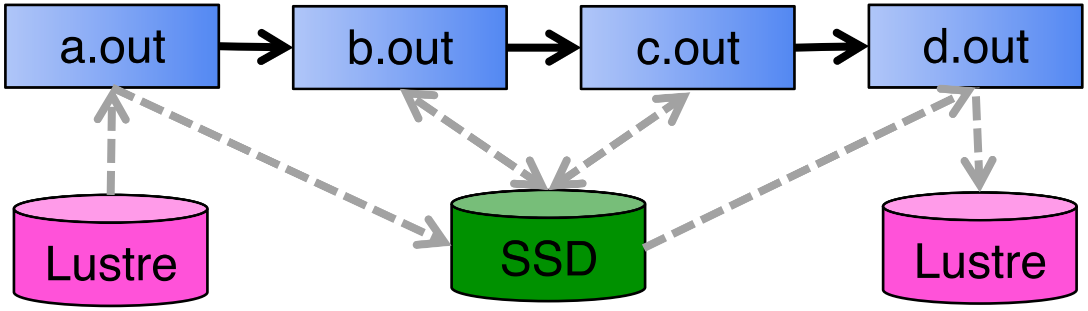
SSDs are a good fit for data-intensive computing

	Flash Drive (Intel 710 Series)	Typical HDD	Good for Data Intensive Apps
Latency (r/w)	75/85 µs r/w	10 ms	✔
Bandwidth (r/w)	270 / 210 MB/s	100-150 MB/s	✔
IOPS (r/w)	38,500 / 2700	100	✔
Power consumption	3.7 W (0.7 W idle)	6-10 W	✔
Price/GB	$3/GB	$.50/GB	-
Endurance	1.1 PB	N/A	✔
Total Cost of Ownership	Jury is still out.		

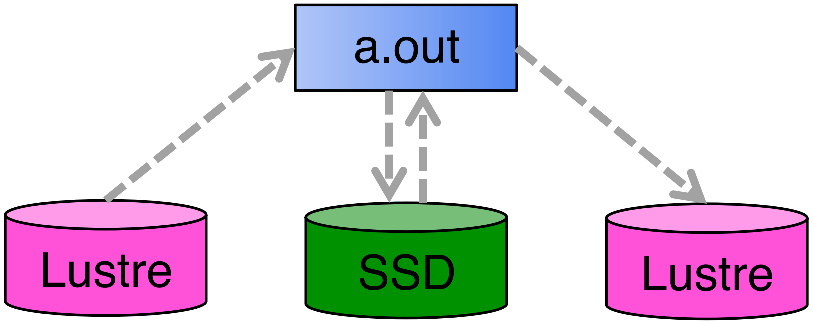
vSMP and Flash Bridge the Latency & Capacity Gap



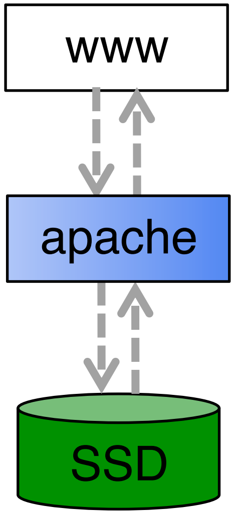
Flash use scenarios



Intermediate results in processing pipeline written(read) to(from) flash

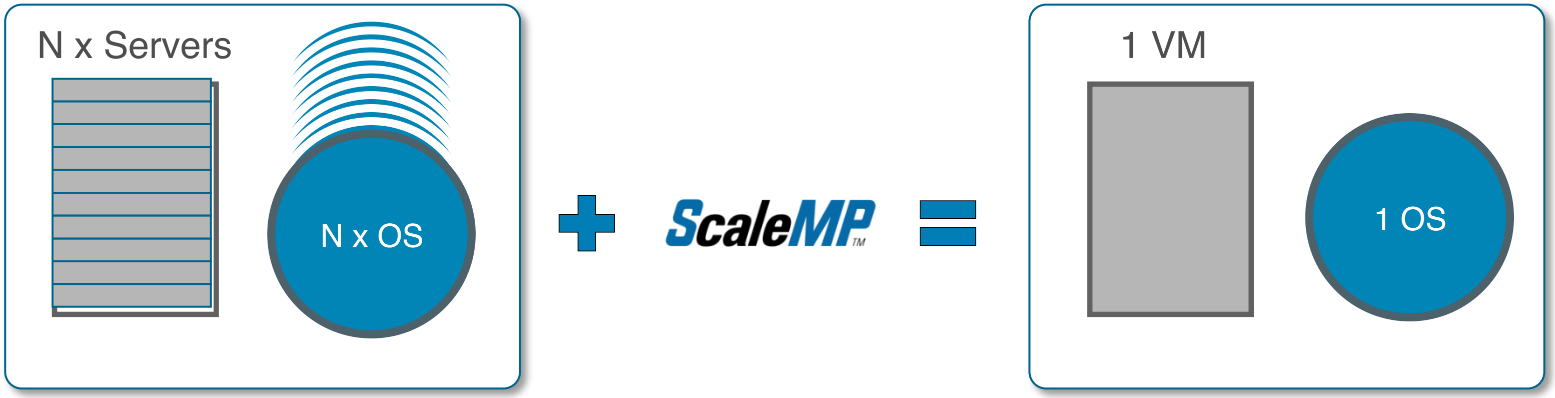


Use flash as fast scratch space



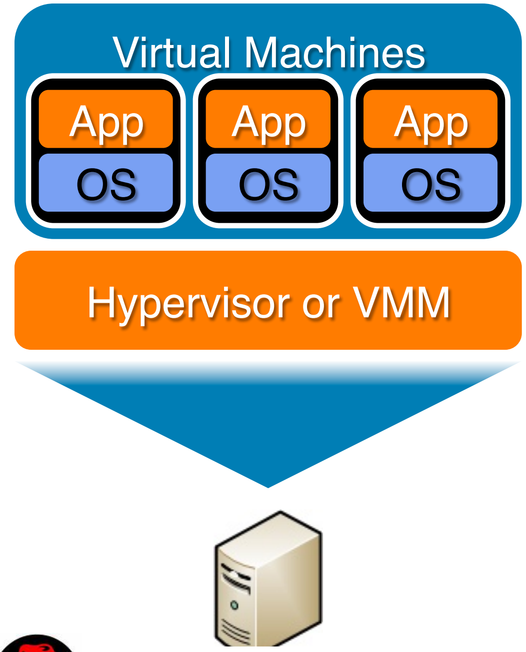
House data base in flash and serve to web

Introduction to vSMP



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

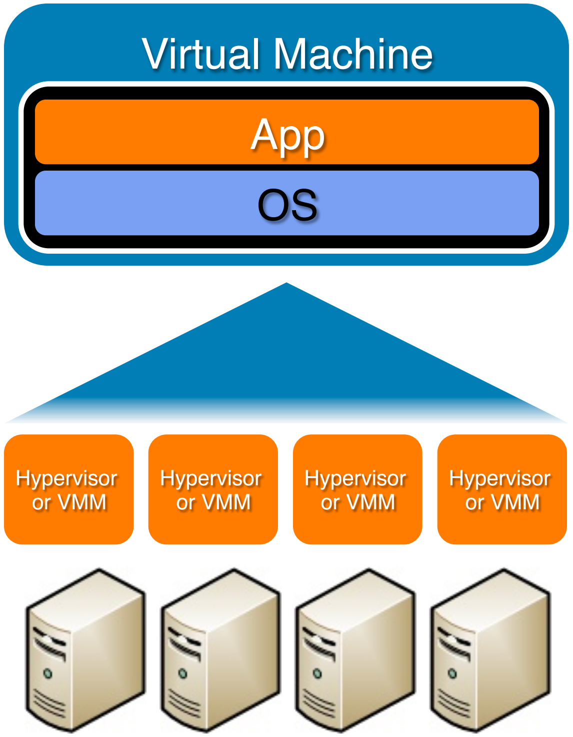
PARTITIONING



redhat vmware CITRIX

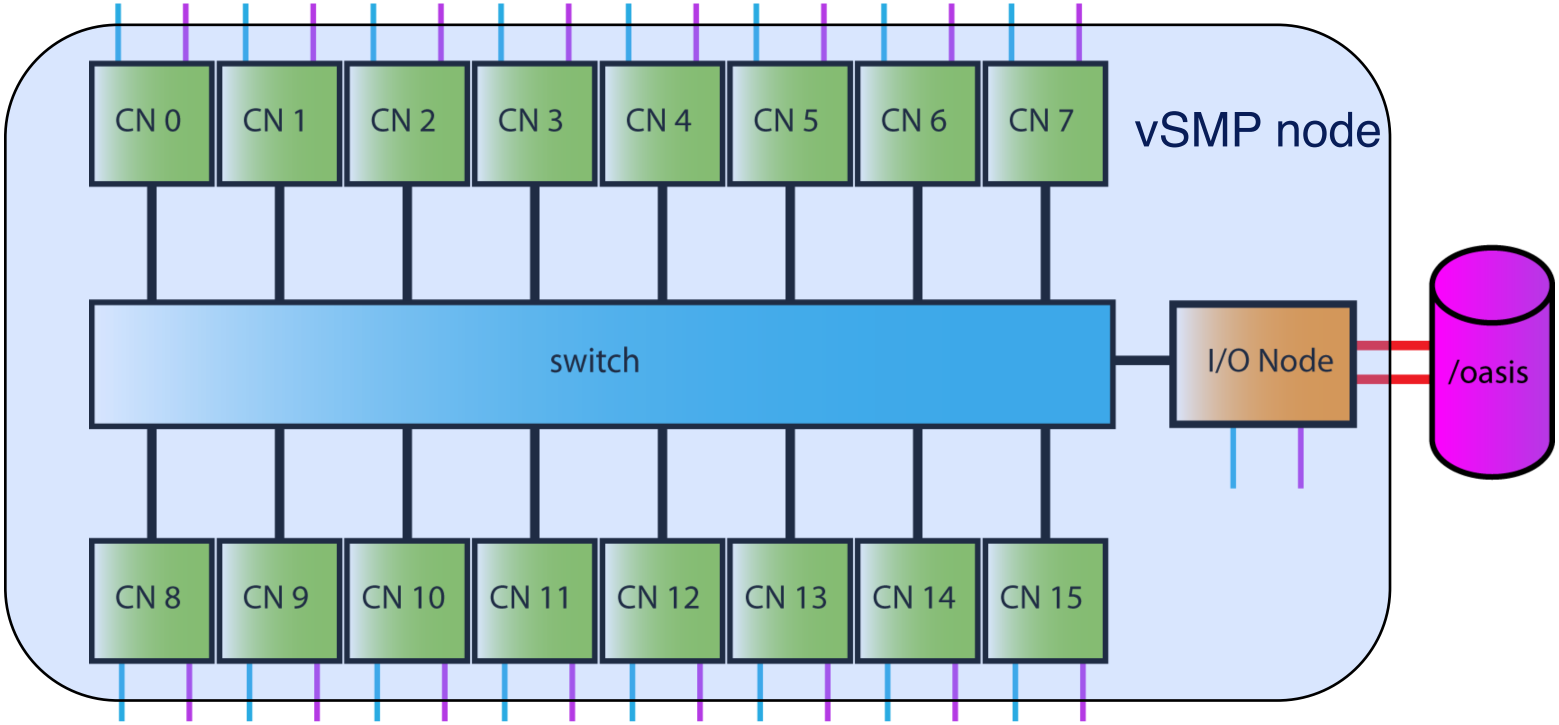
QUMRANET Microsoft Xen Source

AGGREGATION



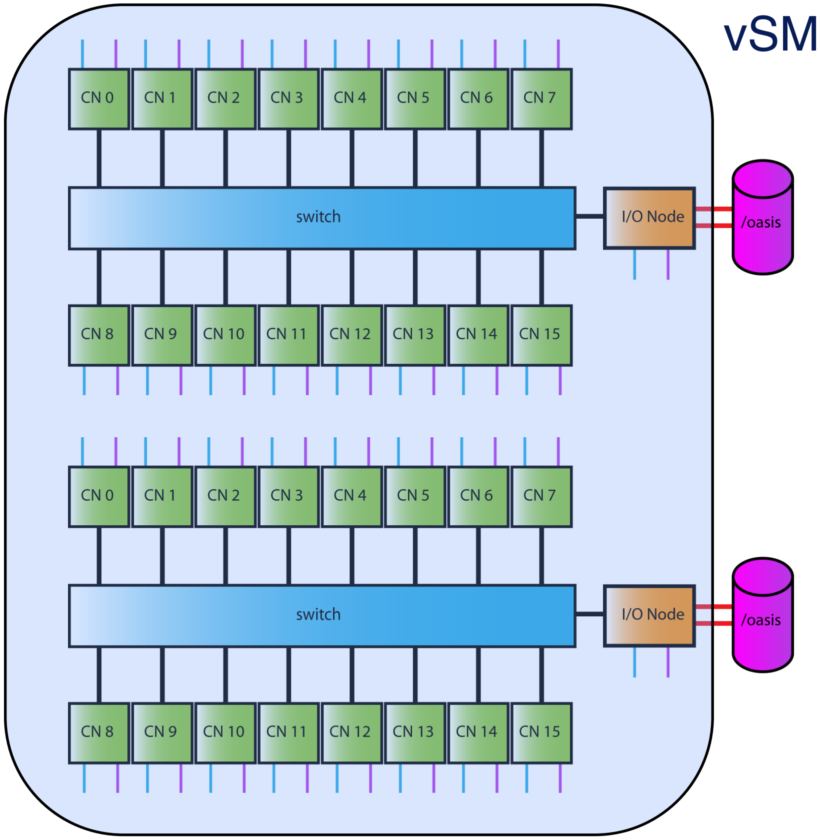
ScaleMP™

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



To user, logically appears as a single, large SMP node

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



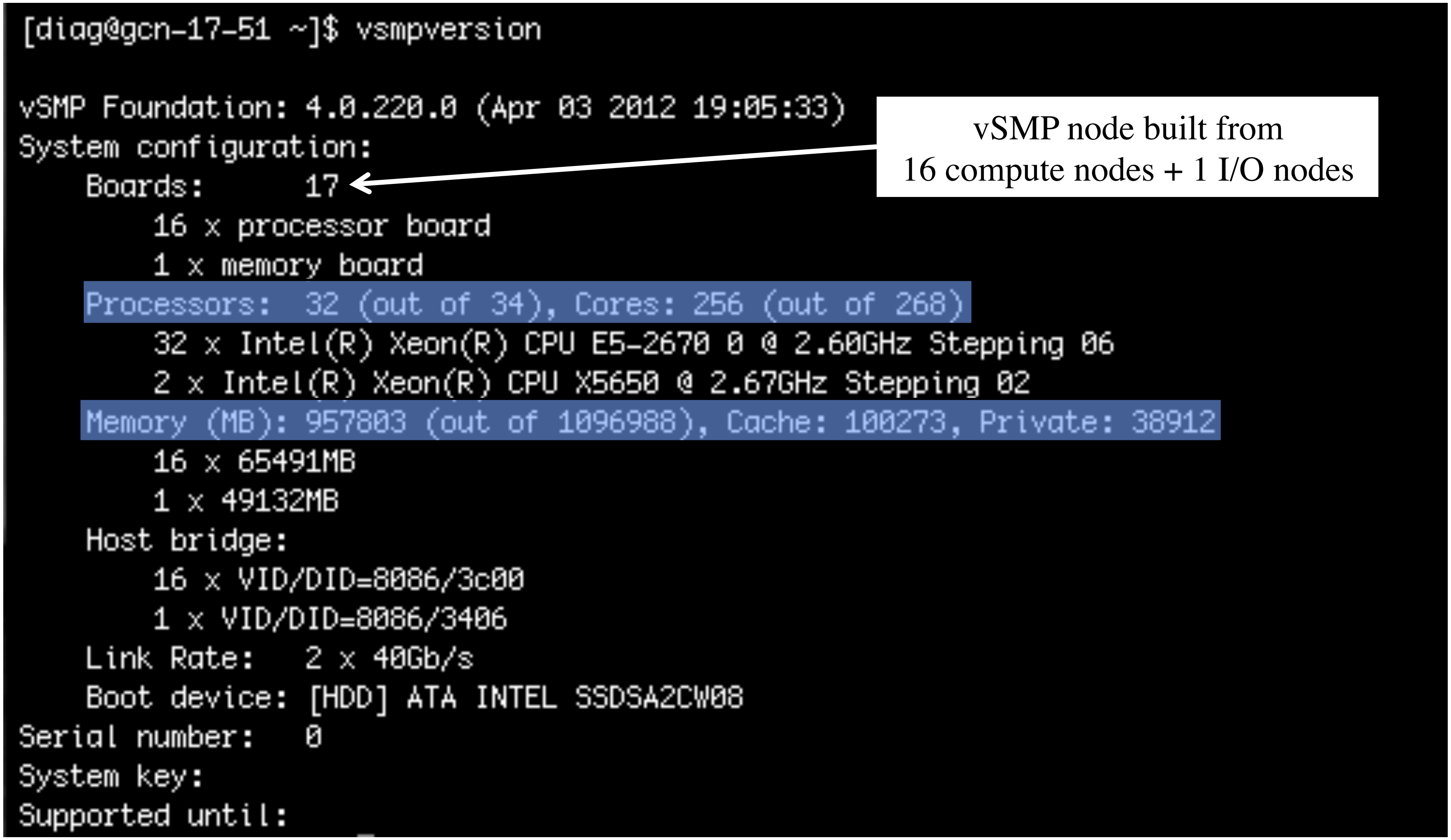
vSMP node

vSMP software provides flexibility to deploy logical shared memory nodes in sizes demanded by users.

Can potentially configure vSMP nodes on the fly (but not quite there yet)

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)

Overview of a vSMP node

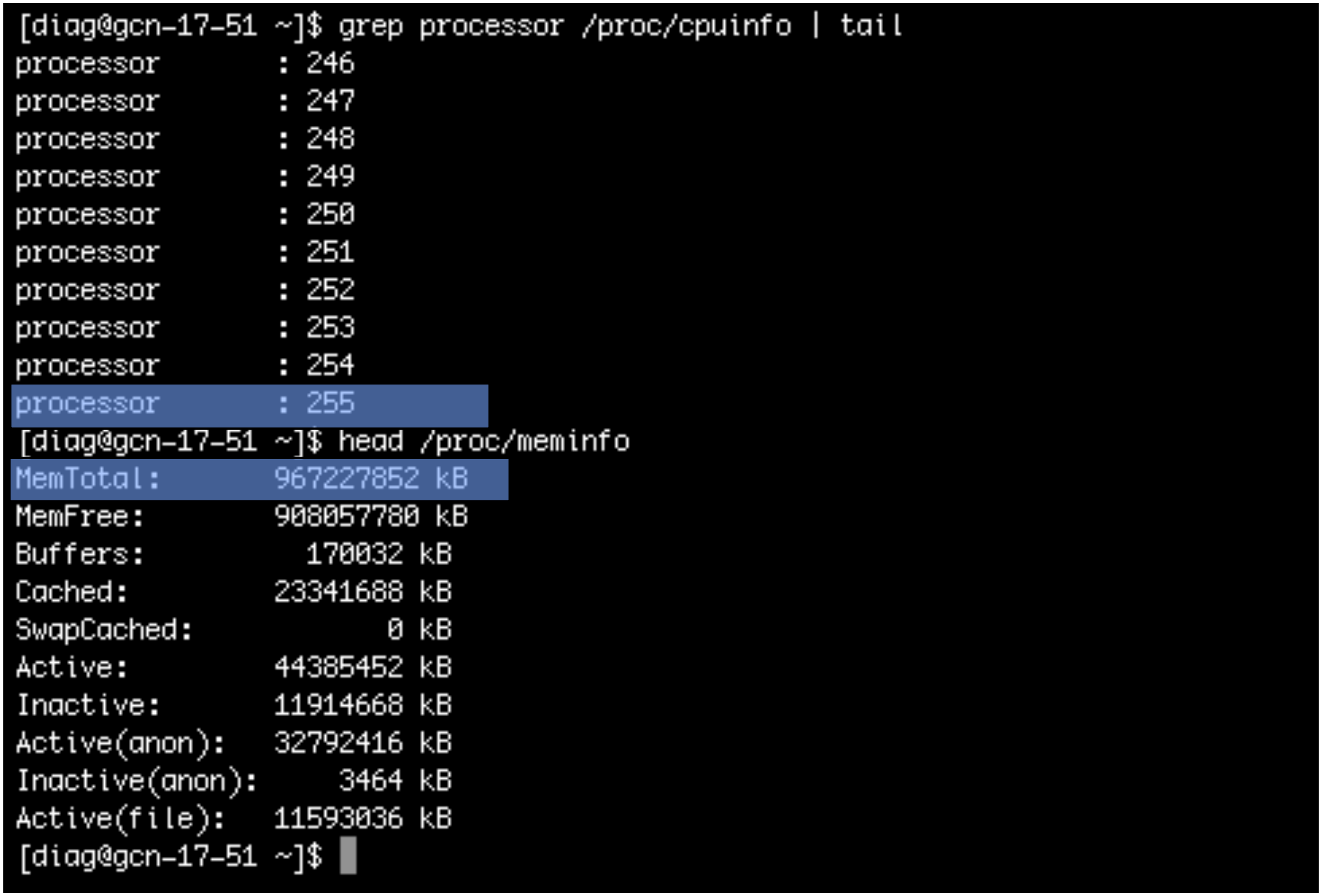


```
[diag@gcn-17-51 ~]$ vsmpversion

vSMP Foundation: 4.0.220.0 (Apr 03 2012 19:05:33)
System configuration:
    Boards:      17
        16 x processor board
        1 x memory board
    Processors:  32 (out of 34), Cores: 256 (out of 268)
        32 x Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz Stepping 06
        2 x Intel(R) Xeon(R) CPU X5650 @ 2.67GHz Stepping 02
    Memory (MB): 957803 (out of 1096988), Cache: 100273, Private: 38912
        16 x 65491MB
        1 x 49132MB
    Host bridge:
        16 x VID/DID=8086/3c00
        1 x VID/DID=8086/3406
    Link Rate:   2 x 40Gb/s
    Boot device: [HDD] ATA INTEL SSDSA2CW08
Serial number:   0
System key:
Supported until:
```

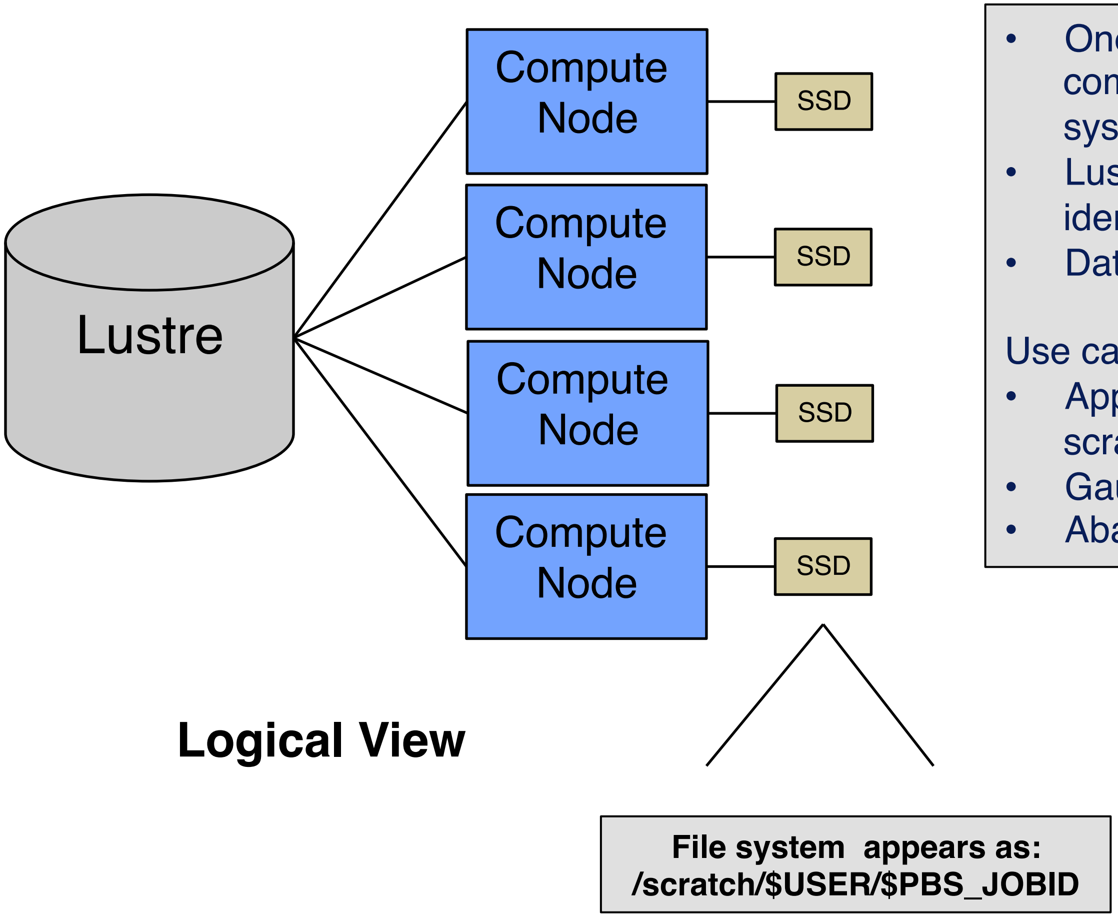
vSMP node built from
16 compute nodes + 1 I/O nodes

Overview of a vSMP node



```
[diag@gcn-17-51 ~]$ grep processor /proc/cpuinfo | tail
processor       : 246
processor       : 247
processor       : 248
processor       : 249
processor       : 250
processor       : 251
processor       : 252
processor       : 253
processor       : 254
processor       : 255
[diag@gcn-17-51 ~]$ head /proc/meminfo
MemTotal:       967227852 kB
MemFree:        908057780 kB
Buffers:          170032 kB
Cached:         23341688 kB
SwapCached:            0 kB
Active:         44385452 kB
Inactive:       11914668 kB
Active(anon):   32792416 kB
Inactive(anon):     3464 kB
Active(file):   11593036 kB
[diag@gcn-17-51 ~]$ 
```

Exporting Flash
Model A: One SSD per Compute Node



Logical View

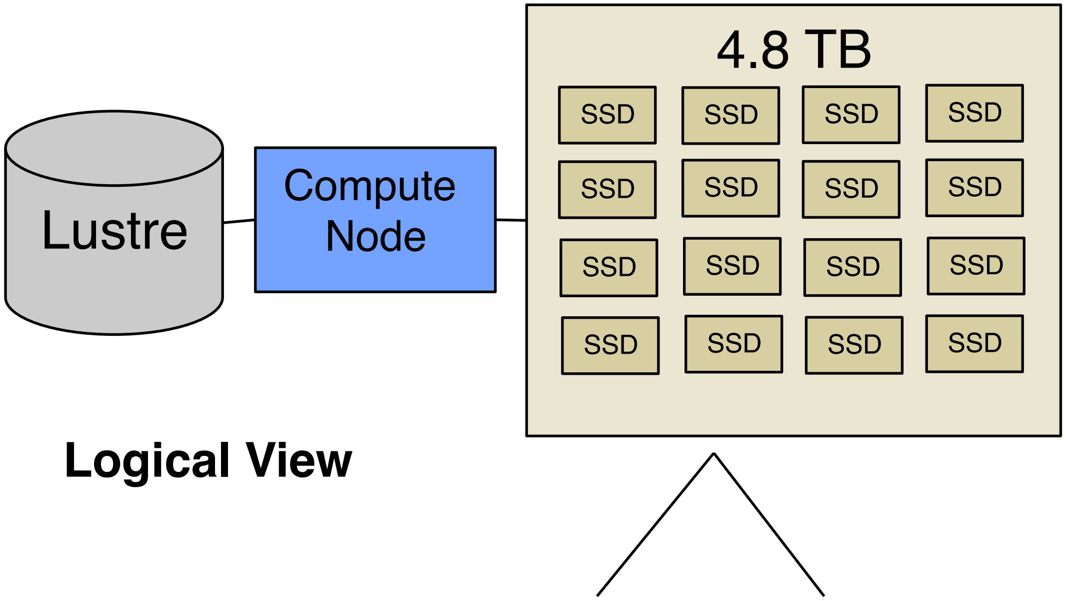
File system appears as: /scratch/$USER/$PBS_JOBID

- One 300 GB flash drive exported to each compute node appears as a local file system
- Lustre parallel file system is mounted identically on all nodes.
- Data is purged at the end of the run

Use cases:

- Applications that need local, temporary scratch
- Gaussian
- Abaqus

Exporting Flash
Model B: 16 SSD's for 1 Compute Node



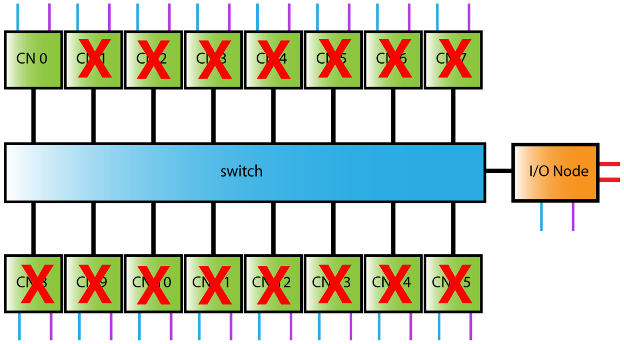
Logical View

File system appears as:
/scratch/$USER/$PBS_JOBID

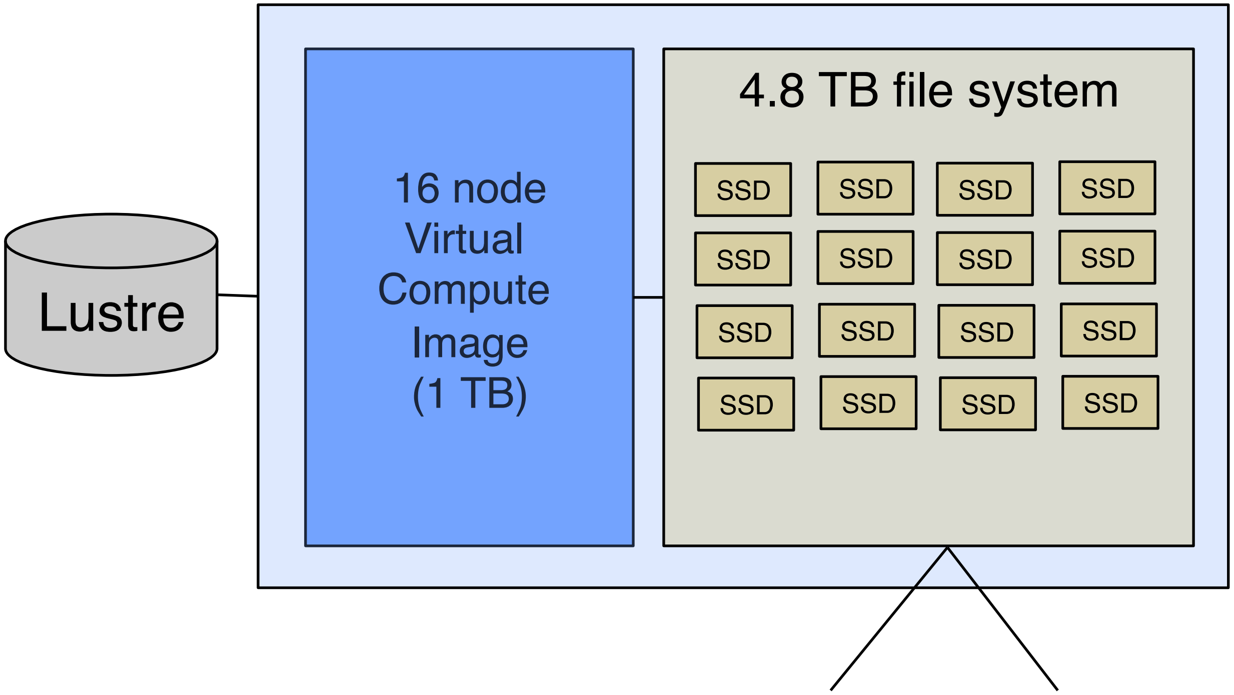
- 16 SSD's in a RAID0 appear as a single 4.8 TB file system to the compute node.
- Flash I/O and Lustre traffic uses Rail 1 of the torus.

Use cases:

- Database
- Data mining
- Gaussian
- Abaqus



Exporting Flash
Model C: 16 SSD’s within a vSMP Supernode



File system appears as:
/scratch1/$USER/$PBS_JOBID
(/scratch2 available if using a
32-node supernode)

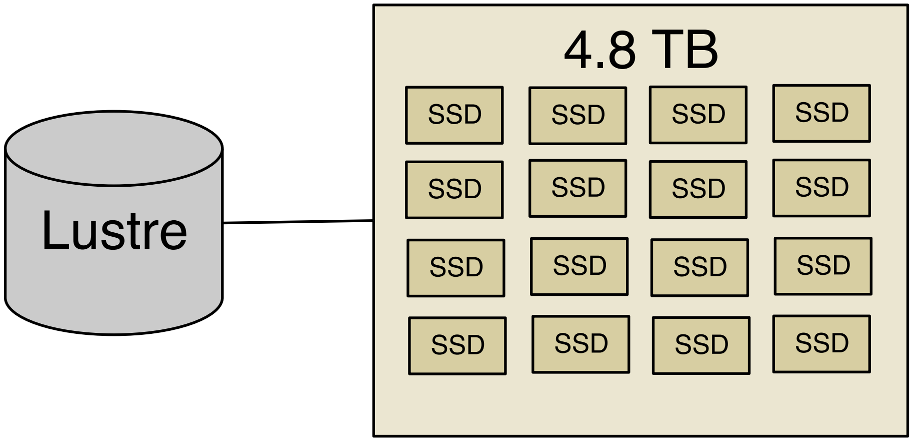
Logical View

- 4.8 TB flash as a single XFS file system
- Flash I/O uses both rail 0 and rail 1
- Data purged at the end of a run

Use cases:

- Serial and threaded applications that need large memory and local disk
- Abaqus
- Genomics (Velvet, Allpaths, etc)

Exporting Flash
Model D: Dedicated I/O node



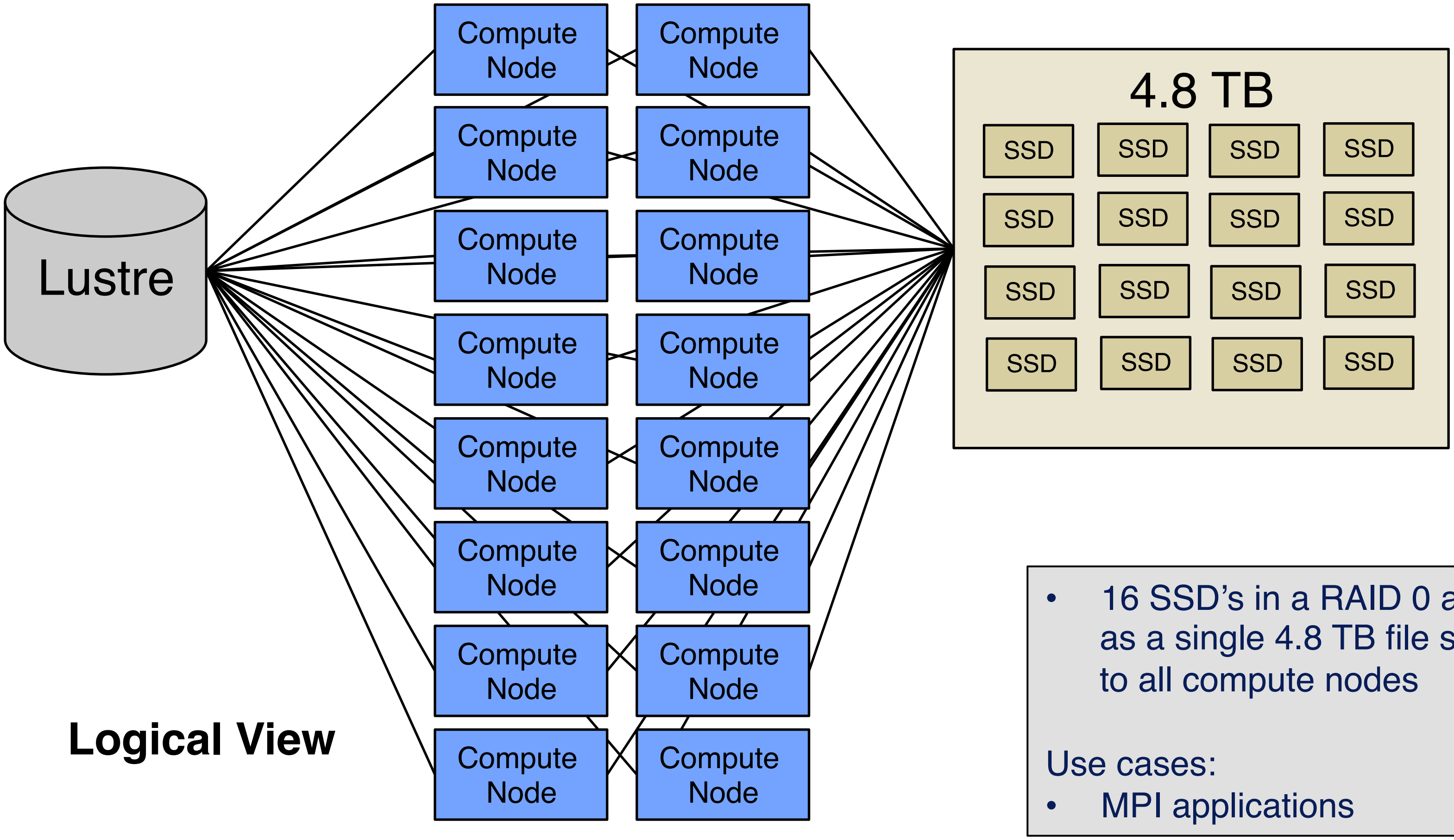
Logical View

- I/O node is allocated for up to one year
- Users run applications directly on the I/O node
- Users may request dedicated compute nodes or access them through the scheduler
- Data is persistent

Use cases:

- Database
- Data mining
- Community applications (XSEDE Science Gateways)

Exporting Flash
Model E: 16 SSD's/ 16 compute node – Flash mounted as parallel file system (Under development)



Logical View

- 16 SSD's in a RAID 0 appear as a single 4.8 TB file system to all compute nodes

Use cases:

- MPI applications

Computational Style Code
Answering the question: Why Gordon?

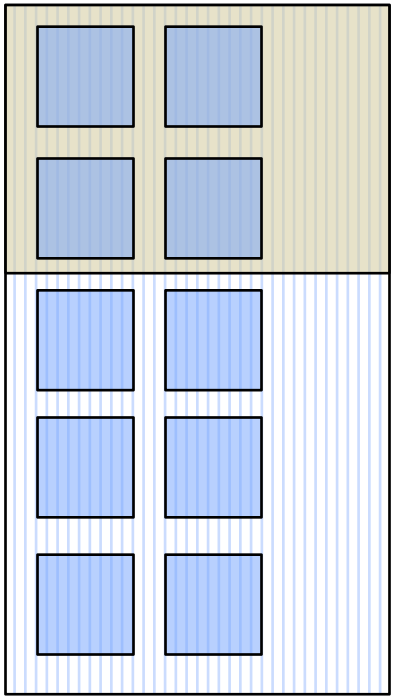
V	M	F
C	T	L

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

Foxglove Calculation using Gaussian 09 with vSMP - MP2 Energy Gradient Calculation

The Foxglove plant (Digitalis) is studied for its medicinal uses. *Digoxin*, an extract of the Foxglove, is used to treat a variety of conditions including diseases of the heart. There is some recent research that suggests it may also be a beneficial cancer treatment.

🟦 = 1 Compute node
(16 cores/node)
64 GB/node)



Processor footprint - 4 nodes
64 threads

Time to solution:

43,000s

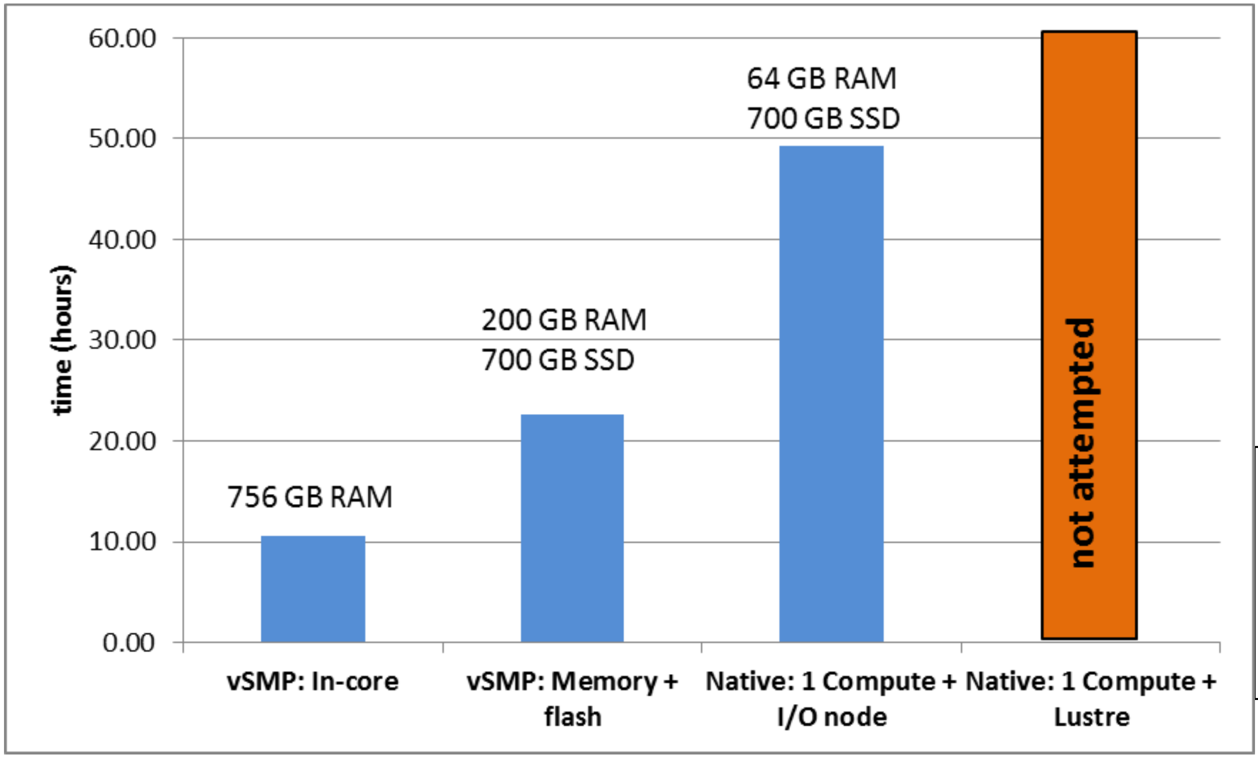
Memory footprint – 10 nodes
700 GB

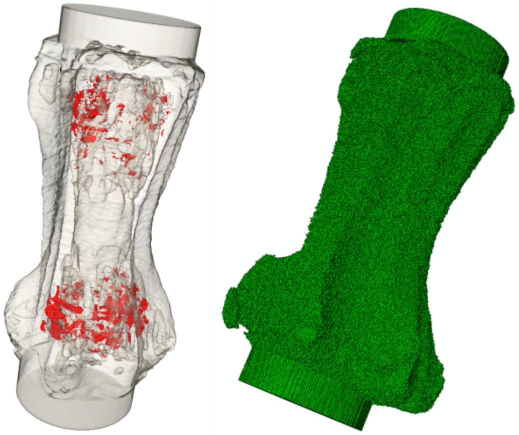
V	M	F
C	T	L

Source: *Jerry Greenberg, San Diego Supercomputer Center. January, 2012.*

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

V	M	F
C	T	L

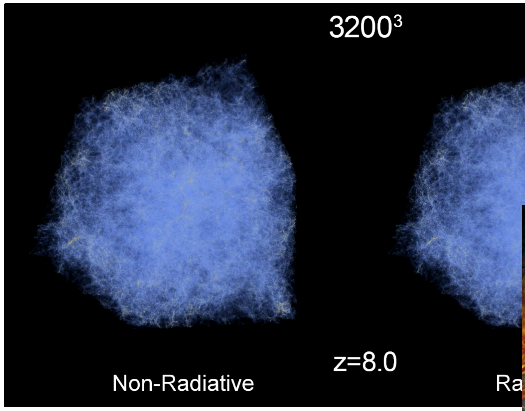
Source: *Matthew Goff, Chris Hernandez. Cornell University. Used by permission. 2012*

Cosmology simulation - matter power spectrum measurement using vSMP

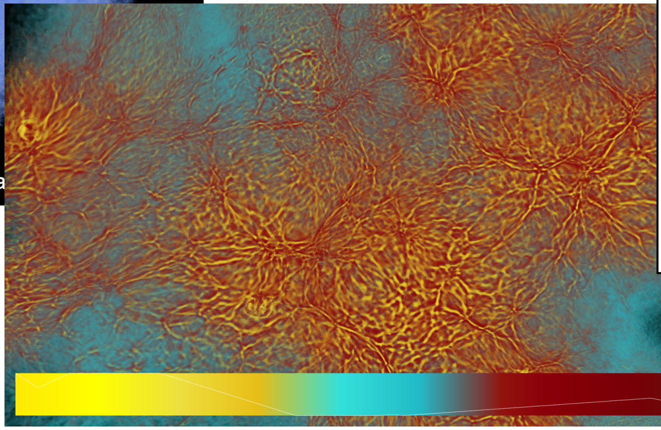
Goal is to measure the effect of the light from the first stars on the evolution of the universe. To quantitatively compare the matter distribution of each simulation, we use radially binned 3D power spectra.

- 2 simulations
- 3200^3 uniform 3D grids
- 15k+ files each

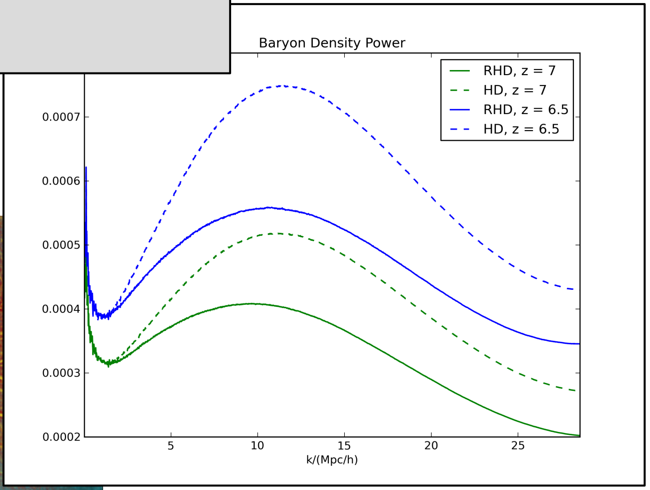
- Existing OpenMP code
- ~256GB memory used
- ~5 ½ hours per field
- 0 development effort



Individual simulations



Difference



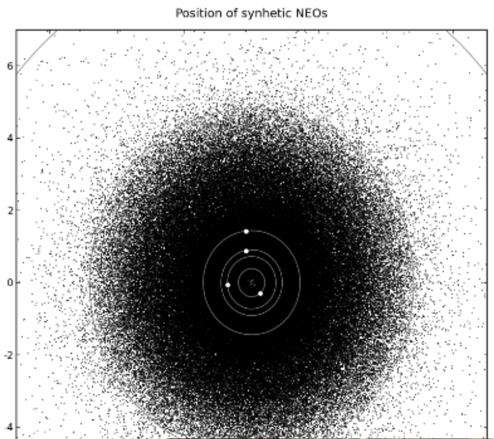
Power spectra

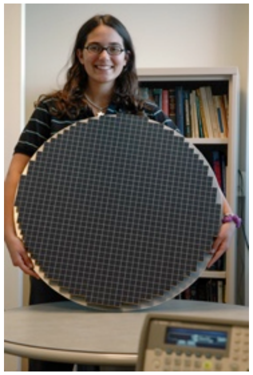
V	M	F
C	T	L

Source: *Rick Wagner, Michael L. Norman. SDSC.*

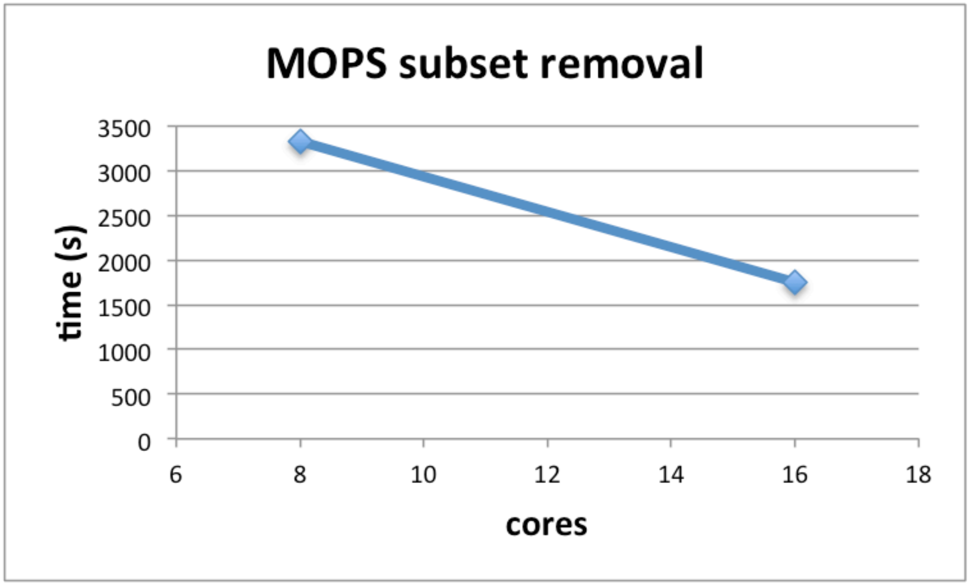
LSST – Moving Object Pipeline System

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



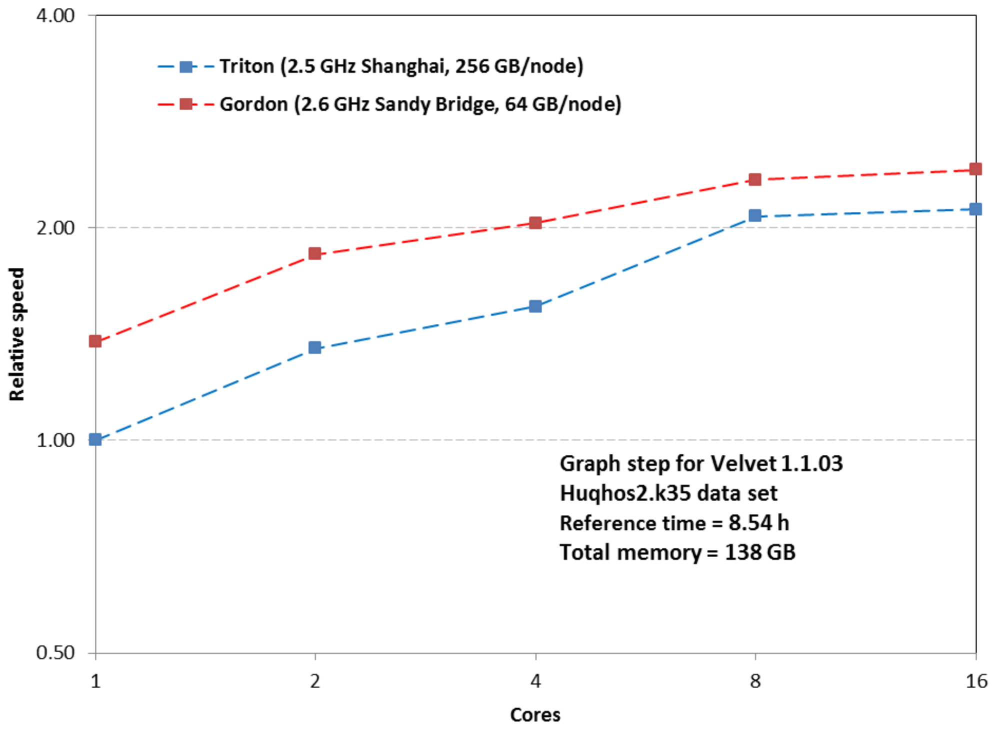
V	M	F
C	T	L

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

Daphnia Genome Assembly using Velvet and vSMP

Daphnia (a.k.a. water flea), is a model species used for understanding mechanisms of inheritance and evolution, and as a surrogate species for studying human health in responses to environmental changes.

Photo: Dr. Jan Michels, Christian-Albrechts-University, Kiel



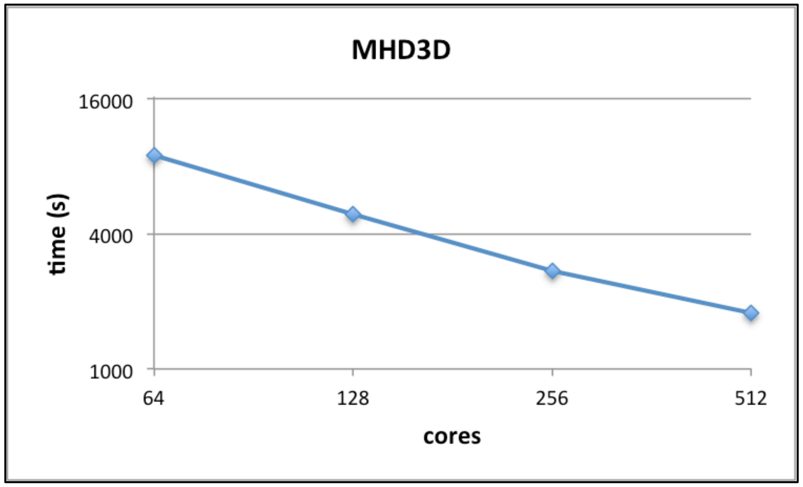
Graph step for Velvet 1.1.03
Huqhos2.k35 data set
Reference time = 8.54 h
Total memory = 138 GB

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

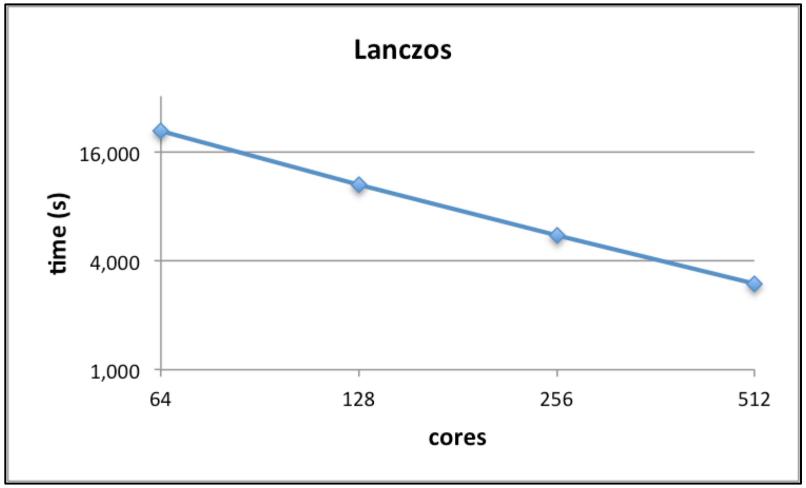
V	M	F
C	T	L

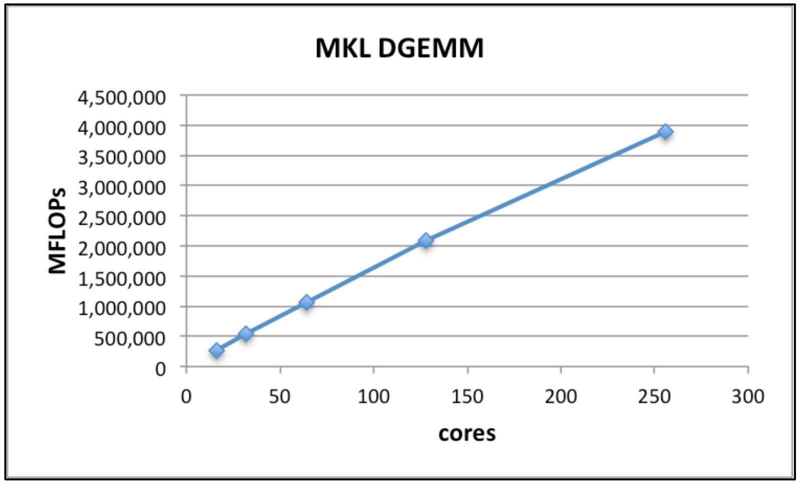
***Source:** Wayne Pfeiffer, San Diego Supercomputer Center. Used by permission.*



Although the main benefit of vSMP is access to large memory, can also be useful for shared memory applications that scale to large numbers of cores

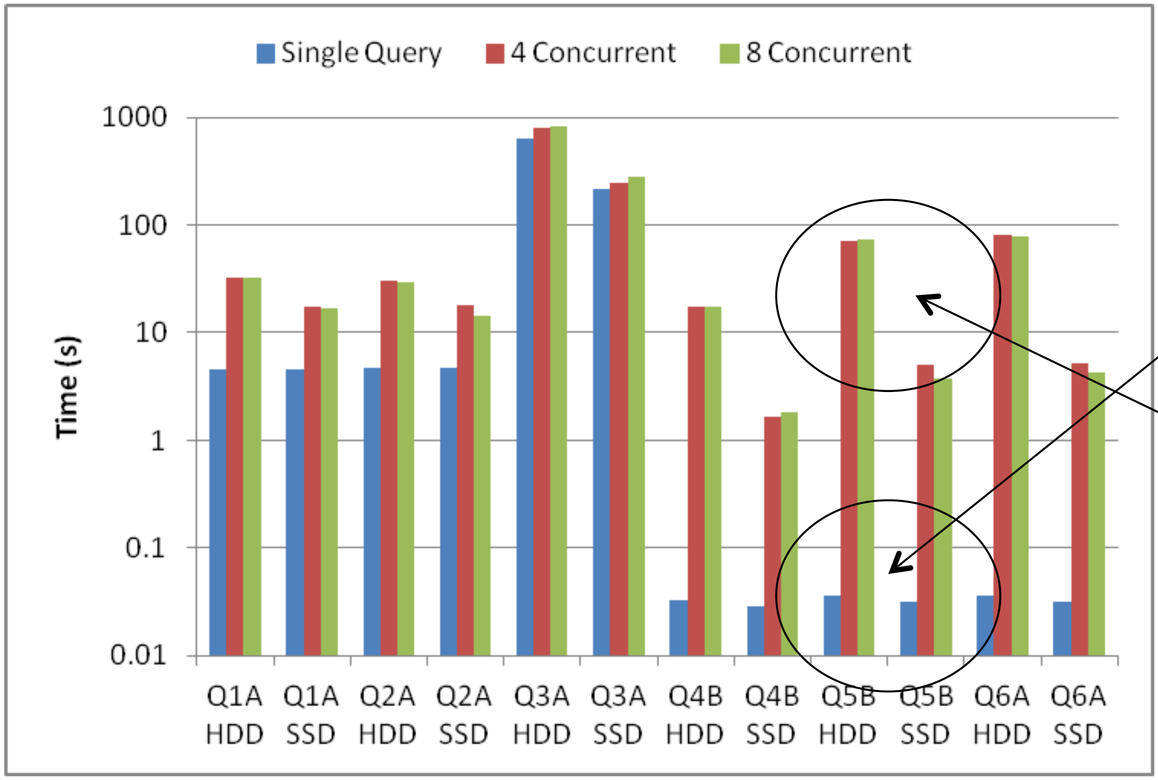
The MHD3D, Lanczos, and MKL DGEMM tests all fit within memory of a single node (< 64 GB)





PDB Query Comparisons, with DB2 Database on two Gordon I/O Nodes: One with HDD's, One with SSD's

The Protein Data Bank (PDB): Is the single worldwide repository of information about the 3D structures of large biological molecules. These are the molecules of life that are found in all organisms. Understanding the shape of a molecule helps to understand how it works.



- For single queries, HDD and SSD perform about the same.
- For concurrent queries, SSD's achieve big speedup.
- Q5B is > 10x, and performance varies by type of query

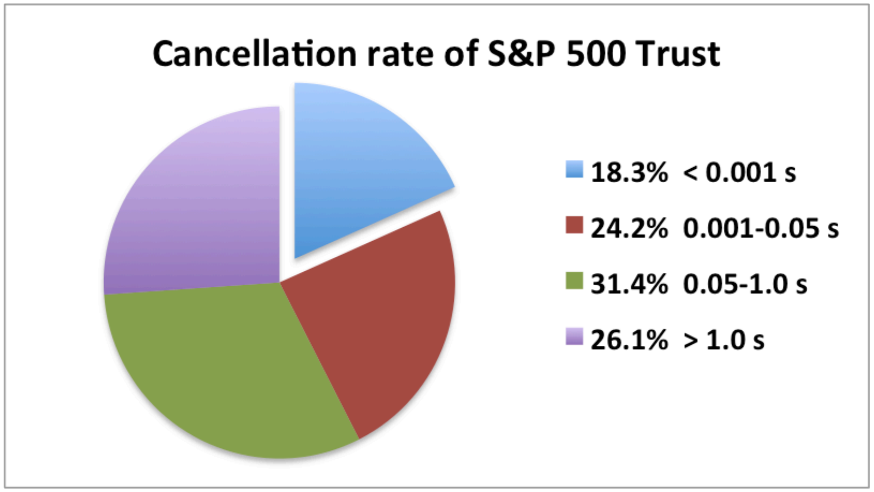
V	M	F
C	T	L

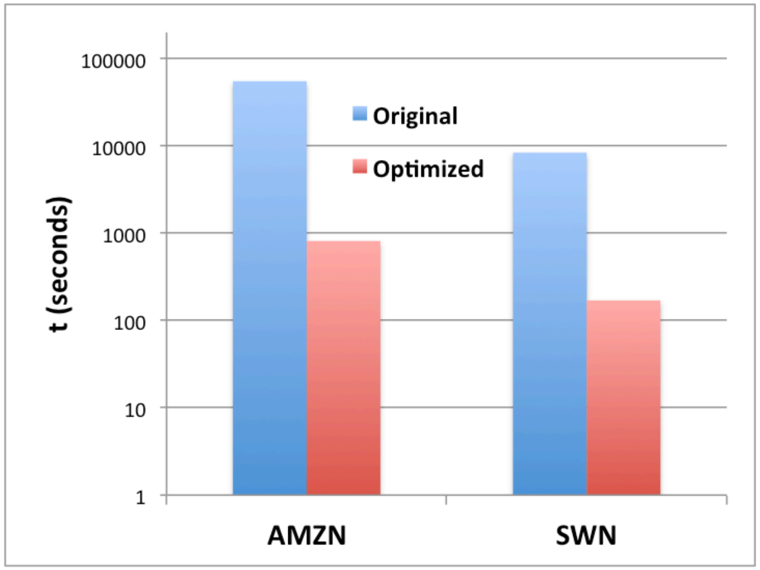
Source: *Vishwinath Nandigam, San Diego Supercomputer Center. 2011*

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


Cancellation rate of S&P 500 Trust

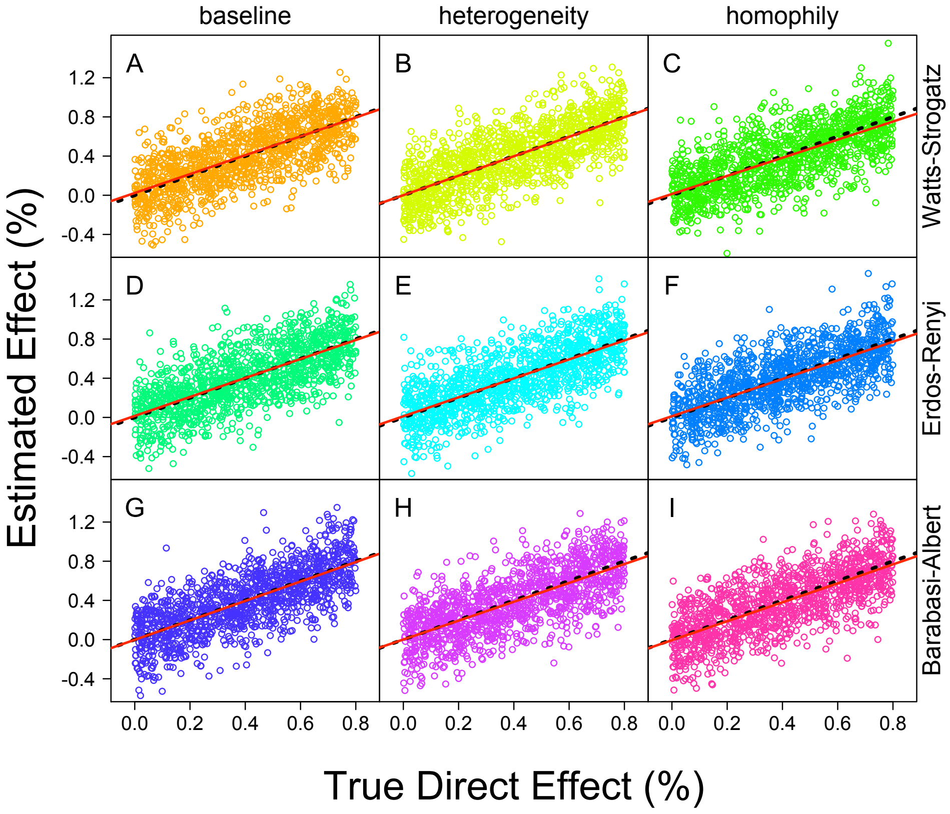


V	M	F
C	T	L

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

Monte Carlo network permutation simulations using R

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



INSIGHT
Gut microbes and health

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

VOTE

0 1 2 2 6 0 3 0 0
People on Facebook Voted

CAMPAIGN TRIAL

Close friendships beat weaker Facebook links in test of election-day influence PAGES 212 & 295

MICROWAVE WARFARE
UNEXPLODED MYTH?
The Pentagon's quest for the e-bomb
PAGE 198

RESEARCH CAREERS
PICKING WINNERS
A formula to predict your future h*-index*
PAGE 201

LITERATURE
FORGOTTEN FUTURES
When science fiction came of age
PAGE 204

NATURE.COM/NATURE
13 September 2012 £10
Vol. 489, No. 7415

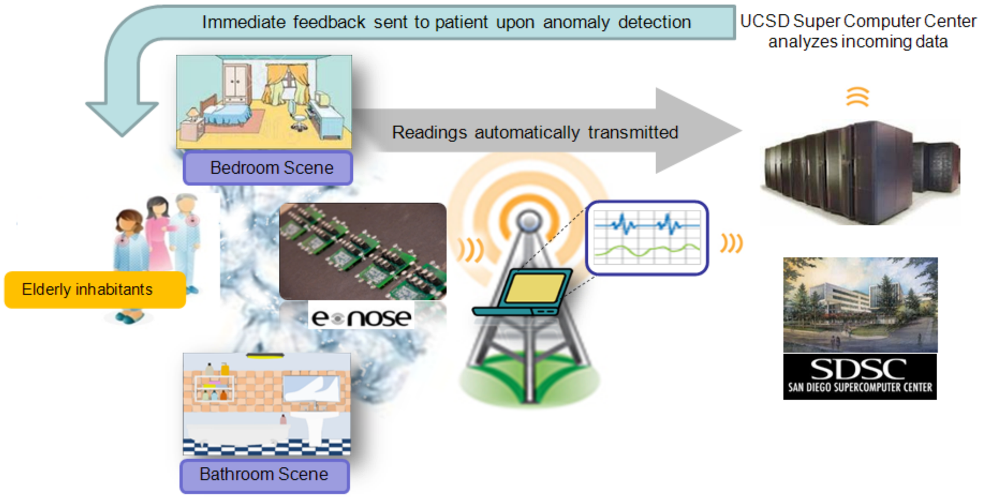
***Source:** Chris Fariss, UCSD Dept. of Political Science Used by permission. 6/1/2012*

V	M	F
C	T	L

SDSC SAN DIEGO SUPERCOMPUTER CENTER

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.



Immediate feedback sent to patient upon anomaly detection

UCSD Super Computer Center analyzes incoming data

Readings automatically transmitted

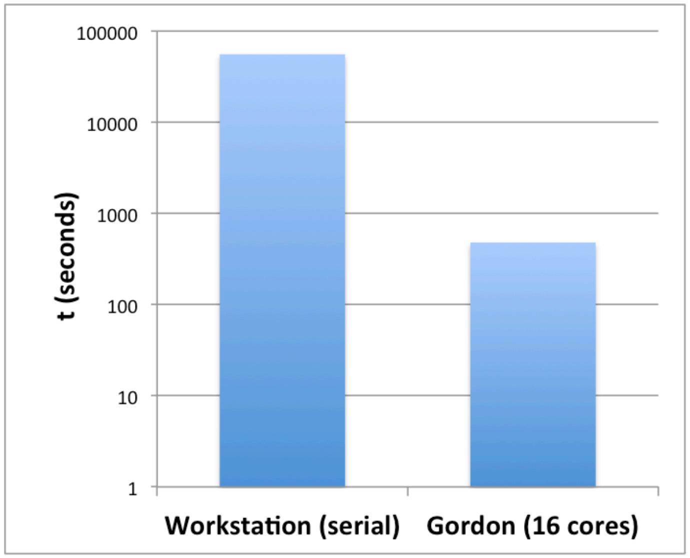
Bedroom Scene

Elderly inhabitants

e-nose

Bathroom Scene

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

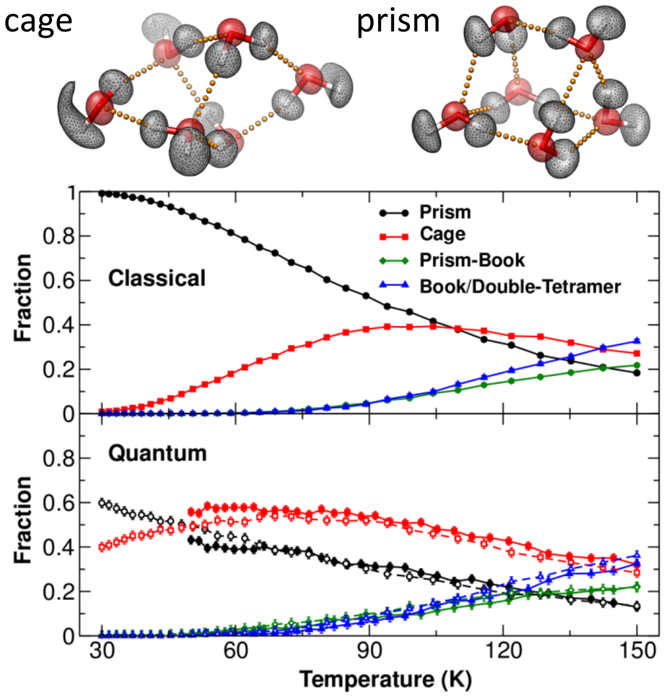


V	M	F
C	T	L

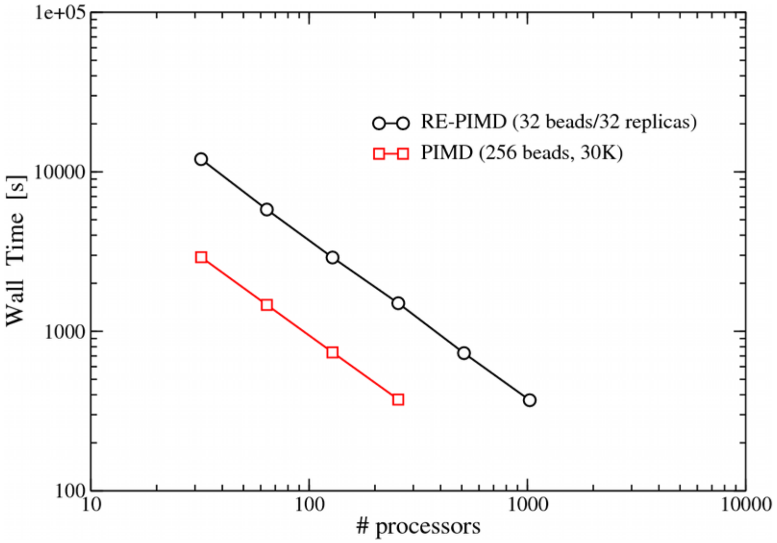
Source: *Ramon Huerta, UCSD Bio Circuits Institute Used by permission 6/1/2012*

Quantum simulations of water clusters

The water hexamer is the smallest water cluster having a fully 3D structure. Conflicting predictions for the relative stability of different hexamer isomers underscores the sensitivity of this prototypical H-bonded system to the description of the underlying molecular interactions. Quantum simulations on the full-dimensional *ab initio* WHBB potential energy surface made it possible to determine for the first time relative populations of different isomers as a function of temperature.



Scaling behavior of quantum simulations on Gordon



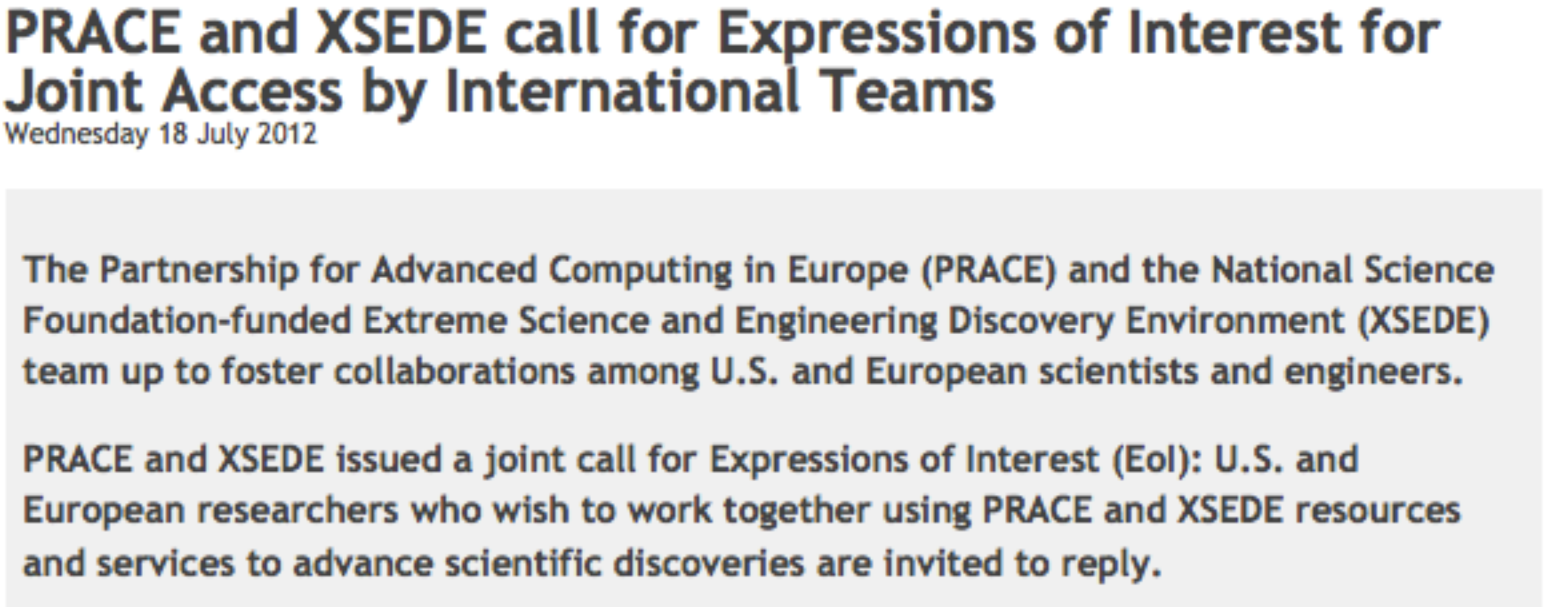
V	M	F
C	T	L

Source: *Francesco Paesani, UCSD Dept. of Chemistry Used by permission. 6/1/2012*

An invitation to collaborate

Actively looking for data/memory intensive applications to run on Gordon. Focus has been on either traditional user base (sci/eng simulations) or new user communities (e.g. social sciences, finance) that are starting to scale up. Want to expand to predictive analytics and data mining.

SDSC can provide a number of flash and large memory environments: remotely mounted SSDs, locally mounted SSD and HDD (new), early access to hardware, vSMP nodes of various sizes (N x 16 compute+1I/O)



PRACE and XSEDE call for Expressions of Interest for Joint Access by International Teams

Wednesday 18 July 2012

The Partnership for Advanced Computing in Europe (PRACE) and the National Science Foundation-funded Extreme Science and Engineering Discovery Environment (XSEDE) team up to foster collaborations among U.S. and European scientists and engineers.

PRACE and XSEDE issued a joint call for Expressions of Interest (EoI): U.S. and European researchers who wish to work together using PRACE and XSEDE resources and services to advance scientific discoveries are invited to reply.

I I-shaped people

П Pi-shaped people

Ш Sha-shaped people

Gordon Team

SDSC: Mike Norman, Allan Snavely, Shawn Strande, Bob Sinkovits, Mahidhar Tatineni Jerry Greenberg, Pietro Cicotti, Wayne Pfeiffer, Jeffrey Bennett, Eva Hocks, William Young (now at Roche), Chaitan Baru, Kai Lin, Kenneth Yoshimoto, Susan Rathbun, Diane Baxter, Jim Ballew, Amit Majumdar, Nancy Wilkins, Christopher Irving, Rick Wagner, Natasha Balac, Paul Rodriguez, Nicole Wolter

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

ScaleMP: Nir Paikowsky, Shai Fultheim and many others

Appro: Steve Lyness, Greg Faussette (now at Aeon Computing), Adrian Wu, Roland Wong