Application experiences with Gordon – a flash-based HPC system

Michael L. Norman
Director, SDSC
PI, Gordon Project

Funded by the NSF Office of Cyberinfrastructure: OCI-0910847
What is Gordon?

- A “data-intensive” supercomputer based on SSD flash memory and virtual shared memory SW
  - Emphasizes MEM and IOPS over FLOPS
- A system designed to accelerate access to massive data bases being generated in all fields of science, engineering, medicine, and social science
- The NSF’s Track 2 award to the San Diego Supercomputer Center (SDSC)
- In production since Feb. 2012
Outline

• Data-intensive computing

• Gordon architecture
  • The vision
  • The realization

• Application success stories

• Conclusions
What is Data-Intensive Computing?

- NSF solicitation: “very large data-sets or very large input-output requirements”

- Two data-intensive application classes are important and growing

Data Mining & Data Analytics

Data-Intensive Simulations
What is Data-Intensive Computing?

• Examples Applications:
  • Genome sequence assembly
  • Climate simulation analysis
  • Social network analysis
The Memory Hierarchy of a Typical HPC Cluster

- Registers (1 cycle)
- Caches (2-10 cycles)
- Memory (100 cycles)
- Remote Memory (10,000 cycles)

- Shared memory programming
- Message passing programming
- Disk I/O

Latency Gap

- Spinning Disks (10,000,000 cycles)
- GB
- TB
- PB
The Memory Hierarchy of Gordon

- Registers (1 cycle)
- Caches (2-10 cycles)
- Memory (100 cycles)
- Remote Memory (10,000 cycles)
- Flash Drives (100,000 cycles)
- Spinning Disks (10,000,000 cycles)

Shared memory programming

Disk I/O

GB
TB
100s TB
PB
Gordon’s 2 Big Ideas

• Massive amounts of flash SSD for Big Data analysis
• Gordon employs 300 TB of flash
  • Largest thumb drive in the world

• Enough to store Netflix entire 100,000 title collection 3 times over
• ..or 100,000 entire human genomes
Gordon’s 2 Big Ideas

• Make supercomputers simple to program via virtual shared memory

• ScaleMP SW creates “supernodes” that can run standard analysis software on very large data
Gordon 32-way Supernode

vSMP aggregation SW

ION

Dual WM IOP

4.8 TB flash SSD

ION

Dual WM IOP

4.8 TB flash SSD
Gordon 32-way Supernode

vSMP aggregation SW

8 TF compute

2 TB DRAM

9.6 TB SSD, >1 Million IOPS
The Gordon System - Supernode

- Basic building block is the Supernode
  - 8 TF compute (64 Sandy Bridge 8-core CPUs)
  - 2 TB virtual shared memory (ScaleMP)
  - 9.6 TB flash SSD
  - 3.2 GB/s IO BW
  - > 1 million IOPS
The Gordon System – Full Machine

- 32 Supernodes =
  - 300 TF compute
  - 64 TB RAM
  - 300 TB flash SSD
  - 100 GB/s
  - 36 million IOPS
Why is flash SSD better than disk?

• Read latency for random IO is up to 100x faster than HDD (read head seek time)

• This speeds up database accesses enormously
Why haven't we seen flash on HPC systems before now?

- **Two reasons:**
  - Cost
  - Durability
- **SLC flash SSD (enterprise)**
  - Durable but costly
- **MLC flash SSD (consumer)**
  - Affordable but not durable
- **Intel eMLC flash SSD**
  - Affordable and durable
  - Uses overprovisioning and wear-leveling to increase durability
Allocations Growth Rate by Parent Science
(since Feb. 2012)
Service Units Allocated by Parent Science (since inception)

- **Chemistry**: 108,052,070
- **Molecular Biosciences**: 88,232,093
- **Physics**: 54,916,942
- **Atmospheric Sciences**: 39,632,861
- **Astronomical Sciences**: 28,923,010
- **Materials Research**: 24,681,305
- **Earth Sciences**: 20,379,558
- **Environmental Biology**: 16,763,325
- **Training**: 8,647,131
- **Social and Economic Science**: 7,676,339
- **Avg of 16 others**: 1,353,154

Data Source: XSEDE. Powered by XDMoD/Highcharts.
Applications Success Stories
The Gordon Project enables science breakthroughs in a number of ways

- **Solid state drives** (SSDs) provide fast scratch space for applications that read/write large amounts of temporary data.

- **Dedicated I/O nodes** are vital to projects requiring fast, persistent access to multi-terabyte data sets.

- **vSMP nodes** provide large, logical shared memory and large core counts. Useful for big memory and highly scalable threaded apps.

- **64 GB DRAM** per compute node supports applications requiring both significant compute resources and relatively large memory.

- **Expert support** helps users effectively use Gordon’s novel features and make the transition from serial to parallel or workstation to HPC.
IO NODES

ION

Dual WM IOP

4.8 TB flash SSD
**Dedicated I/O node (Gordon ION) allocations**

Gordon ION awards are intended for projects that can specifically benefit from persistent, dedicated access to the flash storage. These include database / data mining applications and community projects.

- Award is for up to one year
- Can request dedicated use of up to 16 compute nodes on same switch

Gordon ION projects have included:

- IntegromeDB
- OpenTopography
- Protein Data Bank
- Large Scale Video Analytics
- Graduate level distributed computing class
- Analysis of LHC CMS experiment data
High Energy Physics Data Calibration (dedicated I/O node)

SDSC's Gordon Supercomputer Assists in Crunching Large Hadron Collider Data
Published: April 4, 2013. by University of California - San Diego

Gordon, the unique supercomputer launched last year by the San Diego Supercomputer Center (SDSC) at the University of California, San Diego, recently completed its most data-intensive task so far: rapidly processing raw data from almost one billion particle collisions as part of a project to help define the future research agenda for the Large Hadron Collider (LHC).

SDSC logo

World OSG/LHC resources w/o Gordon

World OSG/LHC resources with Gordon

SDSC logo

SAN DIEGO SUPERCOMPUTER CENTER

at the UNIVERSITY OF CALIFORNIA; SAN DIEGO
Gordon integrated with CMS-OSG workflow engine and FNAL archive

SSDs store detector calibration files
Contemporary culture is awash in moving images: there is more video uploaded to YouTube in a single day than a single person can ever view in a lifetime. The Large Scale Video Analytics (LSVA) project is addressing the implications of this storm of images on issues of identity, memory, history and politics.

The LSVA project is customizing the Medici content management system to apply various algorithms for image recognition and visualization into workflows that will allow real-time analysis of video.

To enable this work, the project was provided with a dedicated Gordon I/O node and four compute nodes. Persistent access to the flash storage in the I/O nodes has been critical for minimizing data access times and allowing interactive analysis.

Source: Virginia Kuhn (USC) Used by permission. 2013
**Protein Data Bank (dedicated I/O node)**

The RCSB Protein Data Bank (PDB) is the leading primary database that provides access to the experimentally determined structures of proteins, nucleic acids and complex assemblies. In order to allow users to quickly identify more distant 3D relationships, the PDB provides a pre-calculated set of all possible pairwise 3D protein structure alignments.

Although the pairwise structure comparisons are computationally intensive, the bottleneck is the centralized server that is responsible for assigning work, collecting results and updating the MySQL database.

Using a dedicated Gordon I/O node and the associated 16 compute nodes, work could be accomplished 4-6x faster than using the OSG.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Time for 15M alignments</th>
<th>speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference (OSG)</td>
<td>24 hours</td>
<td>1</td>
</tr>
<tr>
<td>Lyndonville</td>
<td>6.3 hours</td>
<td>3.8</td>
</tr>
<tr>
<td>Taylorsville</td>
<td>4.1 hours</td>
<td>5.8</td>
</tr>
</tbody>
</table>
OpenTopography Facility (dedicated I/O node)

The NSF funded OpenTopography Facility provides online access to Earth science-oriented high-resolution LIDAR topography data along with online processing tools and derivative products. Point cloud data are processed to produce digital elevation models (DEMs) - 3D representations of the landscape.

High-resolution bare earth DEM of San Andreas fault south of San Francisco, generated using OpenTopography LIDAR processing tools

Source: C. Crosby, UNAVCO

Illustration of local binning geometry. Dots are LIDAR shots ‘+’ indicate locations of DEM nodes at which elevation is estimated based

Local binning algorithm utilizes the elevation information from only the points inside of a circular search area with user specified radius. An out-of-core (memory) version of the local binning algorithm exploits secondary storage for saving intermediate results when the size of a grid exceeds that of memory.

Using a dedicated Gordon I/O node with the fast SSD drives reduces run times of massive concurrent out-of-core processing jobs by a factor of 20x

<table>
<thead>
<tr>
<th>Dataset and processing configuration</th>
<th># concurrent jobs</th>
<th>OT Servers</th>
<th>Gordon ION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lake Tahoe 208 Million LIDAR returns 0.2-m grid res and 0.2 m rad.</td>
<td>1</td>
<td>3297 sec</td>
<td>1102 sec</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>29607 sec</td>
<td>1449 sec</td>
</tr>
</tbody>
</table>
Predictive analytics graduate class (dedicated I/O node)

For operations on large data sets, the amount of time spent moving data between levels of the storage and memory hierarchy often dwarfs the computation time. In these cases, it can be more efficient to move the software to the data rather than the traditional approach of moving data to the software. Distributed computing frameworks such as Hadoop take advantage of this new paradigm.

Yoav Freund was awarded a dedicated I/O node to build a Hadoop cluster for use by a graduate level class in distributed computing. The HDFS resides on the SSDs to enable rapid data access.

Student projects included:

- Analysis of temperature and airflow sensors on UCSD campus
- Detection of failures in the Internet
- Prediction of medical costs in car accidents
- Registration of brain images
- Deciphering RNA regulatory code
- Analysis of election day Tweets

Source: Yuncong Chen, Velu Ganapath and Yoav Freund (UCSD)
The IntegromeDB is a large-scale data integration system and biomedical search engine. IntegromeDB collects and organizes heterogeneous data from over a thousand databases covered by the *Nucleic Acid* and millions of public biomedical, biochemical, drug and disease-related resources.

IntegromeDB is a distributed system stored in a PostgreSQL database containing over 5,000 tables, 500 billion rows and 50TB of data. New content is acquired using a modified version of the SmartCrawler web crawler and pages are indexed using Apache Lucene.

Project was awarded two Gordon I/O nodes, the accompanying compute nodes and 50 TB of space on Data Oasis. The compute nodes are used primarily for post-processing of raw data. Using the I/O nodes dramatically increased the speed of read/write file operations (10x) and I/O database operations (50x).

**Source:** Michael Baitaluk (UCSD)

*Used by permission 2013*
SOLID STATE DISK
Gordon is HPC system of choice for many quantum chemists

- Flash SSD accelerates access to large out-of-core electron integral files
- Fast processors
- Large DRAM / node

structure of water hexamers
(F. Paesani, UCSD)

Source: Lillian Chong (U Pittsburgh) Used by permission. 2013
The National Center for Microscopy and Imaging Research (NCMIR) develops technologies to bridge understanding of biological systems between the gross anatomical and molecular scales and to make these technologies broadly available to biomedical researchers. NCMIR provides expertise, infrastructure, technological development, and an environment in which new information about the 3D ultrastructure of tissues, cells, and macromolecular complexes may be accurately and easily obtained and analyzed.

Using 1.8 TB of storage on one of Gordon’s “big flash” nodes, NCMIR researchers were able to speedup image format conversion and generation of aligned image stacks by factors of 12x and 3x, respectively, relative to local resources.

Source: Mark Ellisman and Shava Smallen (UCSD)
Do not redistribute before publication
VIRTUAL SHARED MEMORY
Mathematical anthropology (vSMP)

The identification of cohesive subgroups in large networks is of key importance to problems spanning the social and biological sciences. A k-cohesive subgroup has the property that it is resistant to disruption by disconnection by removal of at least k of its nodes. This has been shown to be equivalent to a set of vertices where all members are joined by k independent vertex-independent paths (Menger’s theorem).

Doug White (UCI) and his collaborators are using software developed using R and the igraph package to study social networks.

The software was parallelized using the R multicore package and ported to Gordon’s vSMP nodes. Analyses for large problems (2400 node Watts-Strogatz model) are achieving estimated speedups of 243x on 256 compute cores. Work is underway to identify cohesive subgroups in large co-authorship networks.

The comparison of large scale cosmology simulations run with and without radiation transport can provide insight into the effect of light from the first stars on the evolution of the universe. Radially binned 3D matter power spectra are calculated to quantitatively compare the matter distributions of the two simulations.

Original ENZO calculation run on 3200^3 uniform grid. Existing OpenMP code for power spectrum calculation ported to vSMP with zero development effort. Memory footprint was ~256 GB and all 256 cores used.

Source: Rick Wagner, Michael L. Norman. SDSC. Used by permission. 2012
Climate simulation data analysis (vSMP)

Large Eddy Simulations (LES) model the lowest levels of turbulent flow rather than resolving them directly, thereby making real problems in engineering and atmospheric science tractable. Giga-LES is applied here to deep tropical convection in near equilibrium for 24 hours over an area comparable to that of a typical horizontal grid cell in a global climate model.

**Simulation Details**
- GigaLES Model Run Dataset (partial)
- 40 time-steps (24 hour simulation)
- 256 vertical layers
- 2048 x 2048 meters
- 100 m horizontal resolution

**R Analysis**
- 160 GB data set (40 netCDF files @ 4 GB each)
- 340 GB memory footprint
- ~ 3 ½ hours for data input and analysis

The visualization and analysis of the Giga-LES data required significantly more memory than is available on standard nodes. Porting to Gordon vSMP node enabled this work to be carried out in a reasonable amount of time.

*Multi-scale Modeling Framework: M. Kharoutdinov, SUNY Stonybrook*
*Visualization: J. Helly, A. Chourasia*
*Analysis: J. Helly, S. Strande*

The Center for Multi-scale Modeling of Atmospheric Processes (CMMAP) is an NSF Science and Technology Center focused on improving the representation of cloud processes in climate models.
The goal of the simulations is to analyze how small variances in boundary conditions affect 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 25³ micro CT images into voxel-based finite element models

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

Run times for LOB construction of heavily traded NASDAQ securities (June 4, 2010)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>wall time (s) orig. code</th>
<th>wall time (s) opt. code</th>
<th>speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWN</td>
<td>8400</td>
<td>128</td>
<td>66x</td>
</tr>
<tr>
<td>AMZN</td>
<td>55200</td>
<td>437</td>
<td>126x</td>
</tr>
<tr>
<td>AAPL</td>
<td>129914</td>
<td>1145</td>
<td>113x</td>
</tr>
</tbody>
</table>

Project benefited greatly by direct involvement of SDSC applications staff. Optimizations to the original thread-parallel code resulted in greater than 100x speedups. It is now possible to analyze entire day of NASDAQ activity in a few hours using 16 Gordon nodes. With new capabilities, beginning to consider analysis of options data with 100x greater memory requirements.

Source: Mao Ye, Dept. of Finance, U. Illinois. Used by permission. 6/1/2012
MEMORY
Heterogeneity in virus structure (memory)

Cryo-electron microscopy and three-dimensional image reconstruction are the techniques of choice for determining the structure of macromolecular complexes that resist crystallization. Using a novel maximum likelihood estimator (MLE) approach, it is also possible to determine heterogeneity of the structure and gain insight into the viral maturation process.

Cross-sections of the HK97 reconstructions for particles with (left) and without protease (right). Variance is indicated by color and image-specific color maps are used (blue is low and red is high).

The MLE approach is both computationally and memory intensive. Early work using this algorithm was hampered by access to local resources that only have 16 GB of memory per node. The software is written in MATLAB and parallelized using the Matlab Distributed Computing Server (MDCS).

After porting to software to Gordon, MLE algorithms are being applied to first novel biological calculation (Hong Kong 97 virus).

Source: Peter Doerschuk (Cornell), Jack Johnson (TSRI) Used by permission. 2013
Lattice QCD simulations at light quark masses are notoriously difficult, especially for protons and neutrons, due to the exponentially decreased signal-to-noise ratio as the quark masses become lighter. An improved numerical technique called All-Mode-Averaging (AMA), based on the lowest eigenvectors of very large sparse matrices is being used to speed up these calculations.

The AMA calculations are floating point intensive and most production runs use 512 or 1024 compute cores. Each eigenvector is 12 GB in size and large physical memory is required. Substantial output is generated and at times the project has used more than 60 TB of disk space.

Due to the large memory requirement, Gordon is the only large scale XSEDE resource appropriate for these calculations.

Source: Meifeng Lin (Boston U)
Used by permission. 2013
**Biochemistry of influenza virus (memory)**

Avian influenza type A is a major pandemic threat. Two membrane glycoproteins, hemagglutinin (HA) and neuraminidase (NA), play important roles in mediating critical interactions with host-cell surface receptors that contain terminal sialic-acid moieties. Modeling NA and HA as crowded membrane-bound proteins embedded in the viral envelope provides new insights into the biology of influenza.

Atomistic model was built of the entire influenza virus lipid envelope (~15 million atoms) with NA and HA glycoproteins distributed over the viral surface in a physiologically relevant manner.

The electrostatic potential of this model was subsequently calculated using the parallel Adaptive Poisson Boltzmann Solver (APBS) software.

Even after segmenting into 448 regions, each calculation still required the full 64 GB available on the Gordon nodes.

**Source:** Rommie Amaro (UCSD)  
Used by permission. 2013
The PH (pleckstrin homology) domains are one of the most common membrane targeting motifs in cellular signaling proteins. A combination of experimental and numerical techniques is used to explore the molecular mechanism of membrane binding for the PH domain of the general receptor for phosphoinositides-1 (GRP1) protein.

**Simulation of membrane proteins (processors/network)**

NAMD simulations of membrane-protein systems are up to four times faster than on Kraken and have led to previously unseen levels of molecular detail. The emphasis on small job sizes on Gordon also enables higher throughput (typical node counts are sixteen for the NAMD simulations).

*Source:* Greg Voth (U. Chicago)  
*Used by permission. 2013*
## Gordon supports new computing modalities and communities

<table>
<thead>
<tr>
<th>New Communities</th>
<th>New Modalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Political Science</td>
<td>Persistent databases</td>
</tr>
<tr>
<td>Anthropology</td>
<td>• RCBS Protein Data Bank</td>
</tr>
<tr>
<td>Cinema</td>
<td>• Hadoop processing</td>
</tr>
<tr>
<td>Psychology</td>
<td>• Yoav Freund (UCSD) big data analytics class</td>
</tr>
<tr>
<td>Finance</td>
<td>• vSMP for large graph processing</td>
</tr>
<tr>
<td>Predictive Analytics</td>
<td>• Doug White (UCI) social network cohesion analysis</td>
</tr>
<tr>
<td></td>
<td>• Streaming sensor data</td>
</tr>
<tr>
<td></td>
<td>• Smart energy grid, chemical sensors</td>
</tr>
</tbody>
</table>
Novel applications are in the pipeline

- Hadoop student projects (Yoav Freund, UCSD)
- dark matter/SUSY search (LHC/CMS collaboration)
- analysis of full market options trading data (Mao Ye, UIUC)
- vSMP for visualization of massive astrophysical simulations (Dylan Nelson, Harvard)
- Genomic analysis algorithms (David Haussler, UCSC)

**Student projects included**
- Analysis of temperature and airflow sensors on UCSD campus
  - Detection of failures in the Internet
- Prediction of medical costs in car accidents
  - Registration of brain images
  - Deciphering RNA regulatory code
  - Analysis of election day Tweets
In Summary: 
The Gordon project is like a Vector

• The direction is correct

• The magnitude is increasing
RESERVE SLIDES
Data Oasis* Heterogeneous Architecture
Lustre-based Parallel File System
(Only Gordon scratch file system shown)

*2012 HPCwire Editor’s Choice for Best HPC Storage Product
Data Intensive Computing - Hadoop

- UCSD Predictive Analytics Graduate Class among early adopters.
- Hadoop cluster was set up on dedicated I/O node and associated compute nodes (16).
- User services staff assisted in setting up several Hadoop based tools and infrastructure elements
  - Hbase – Scalable distributed database
  - PIG – High-level data-flow and execution framework
  - Revolution-R with Rhadoop - R packages that allow users to manage and analyze data with Hadoop.
- Monitoring tools (based on ganglia) installed.