

Building a Data-Intensive Supercomputer Architecture for the National Research Community

Shawn Strande, Project Manager, San Diego Supercomputer Center
Maria McLaughlin, Product Marketing Director, Cray

Cray Inc.

Table of Contents

Introduction.....	3
System Concept and Requirements.....	3
Gordon: SDSC's Data Intensive Supercomputer.....	4
Building a Data-Intensive Supercomputer.....	5
Data Oasis: Integrated High Performance Parallel File System	8
Tackling Big Data Challenges	8
Looking Ahead in HPC.....	10
Acknowledgements	11
References	11



Introduction

The Gordon supercomputer project began in 2009 as a proposal from the San Diego Supercomputer Center (SDSC) to the National Science Foundation (NSF) to build a data-intensive supercomputer. Based on the Cray CS300-AC™ cluster supercomputer, the proposed system was innovative in several respects: use of high performance solid-state drives (SSD), very large memory nodes, a very high performance parallel file system, and a dual-rail 3D torus interconnect. Taken together, these elements made the proposed system extremely well-suited to problems challenged by I/O and serial applications requiring very large memory. Additionally, the design represented a major departure from traditional high performance computing (HPC) clusters which are intended to excel at floating point operations.

SDSC proposed and won the five-year, \$20 million grant to build and operate “Gordon,” a data-intensive HPC system based on the Cray CS300-AC™ cluster supercomputer and one of the world’s first high performance computers to emphasize data movement rather than raw computational speed. This paper describes its development. It includes an overview and detailed specifications of the system and describes how Gordon has been addressing HPC big data challenges for SDSC and the national research community. Finally, it describes SDSC’s Data Oasis storage architecture which provides an integrated, high performance storage infrastructure for all of SDSC’s HPC systems.

System Concept and Requirements

In 2008 the NSF released a solicitation requesting systems in three distinct categories: 1) a data-intensive, high performance computing system; 2) an experimental high performance computing system; and 3) a high performance grid test-bed. With regard to the first category, the solicitation requested a proposal for data-intensive HPC systems to be optimized to support research that required or generated very large datasets or had very large I/O requirements. Additionally, the solicitation required the system to achieve a total peak computing capacity of at least 200 flops.

SDSC proposed a unique data intensive architecture, named Flash Gordon¹. As described in the proposal abstract, the project would support the acquisition, deployment and operation of a new system architecture based on the Cray CS300-AC cluster supercomputer. This new architecture was complemented with technologies from Intel and ScaleMP to address the latency gap between main memory and rotating disk storage in modern computing systems. The proposal consisted of very large shared virtual memory, cache-coherent “supernodes” to support a versatile set of programming paradigms and use flash memory to provide a level of dense, affordable, low-latency storage that could be configured as either extended swap space or a very fast file system. The combination of large addressable virtual memory, low-latency flash memory and a user-friendly programming environment would provide a step-up in capability for data-intensive applications that scale poorly on current large-scale architectures, providing a resource enabling transformative research in many domains.

Initially designed in 2009, the system had a planned deployment date of mid-2011 to take advantage of future technologies not available when the system was proposed, including new Intel processors, emerging interconnect products and topology, and SSDs. Gordon needed to meet the following key requirements:

- 2011 deployment
- Fixed cost
- High availability
- High flops
- High I/O capability between compute nodes and storage

¹ SDSC amended the name to “Gordon” just prior to its launch



- High bandwidth and IOPS capability to local storage
- High bandwidth, low latency interconnect
- Easy to manage

The Gordon project was rolled out in two phases. First, SDSC debuted “Dash” — a prototype system which served as a platform for testing the innovative features of the future Gordon production system. Deployed in September 2009, Dash provided the research community with access to the technology for application development and testing. The system became an allocated NSF resource in April 2010.

Based on the Cray GreenBlade™, building block platform for the Cray CS300 cluster supercomputer, Dash delivered 5.2 teraflops of performance. The 64-node system included four SSD flash-based I/O nodes, as well as ScaleMP’s vSMP Foundation™ software. Over its two-year production life, Dash was instrumental in reducing the risk in implementing Gordon’s more innovative features and in helping SDSC and the research community better understand how to bring such features to bear on data-intensive problems.

Gordon: SDSC’s Data Intensive Supercomputer

Gordon debuted at number 48 on the November 2011 Top500® list of the world’s fastest supercomputers and began production operations in early 2012 following extensive acceptance and reliability testing. Since its launch, Gordon has enabled researchers around the world to delve into some of the most challenging data-intensive projects.

As the first NSF open computing platform to include the Intel® Xeon® E5 processor, Gordon leapfrogged many systems on the Top500 list with twice as many cores thanks to the Intel® Advanced Vector Extensions support in the processor which provides twice the operations per clock cycle.

While this benchmark is important, Gordon’s I/O capabilities are more relevant to data-intensive computing. Gordon has over 300 terabytes of Intel® Solid-State Drive 710 series and achieves over 36 million IOPS (I/O operations per second). This is meaningful because many data-intensive applications are characterized by random data access patterns or I/O that is dominated by lots of small reads. With latencies an order of magnitude lower than spinning disk, Gordon provides a quantum leap in performance over traditional clusters. In addition to the SSDs, Gordon boasts a 4 petabyte (PB) Lustre file system rated at 100 gigabytes/s (GB). This number is staggering considering that Gordon is a modest 341 teraflops. Similar systems in the open science community have much lower bytes/flops ratio, which is another indication of Gordon’s balance for data-intensive applications (see Figure 1).

Gordon is part of the NSF’s Extreme Science and Engineering Discovery Environment (XSEDE), a nationwide partnership comprising 16 supercomputers and high-end visualization and data analysis resources. Gordon currently has over 60 distinct projects from a wide range of domains including computational cosmology, genomics, materials science, financial markets analysis, computational biology and more.

Gordon Network Architecture

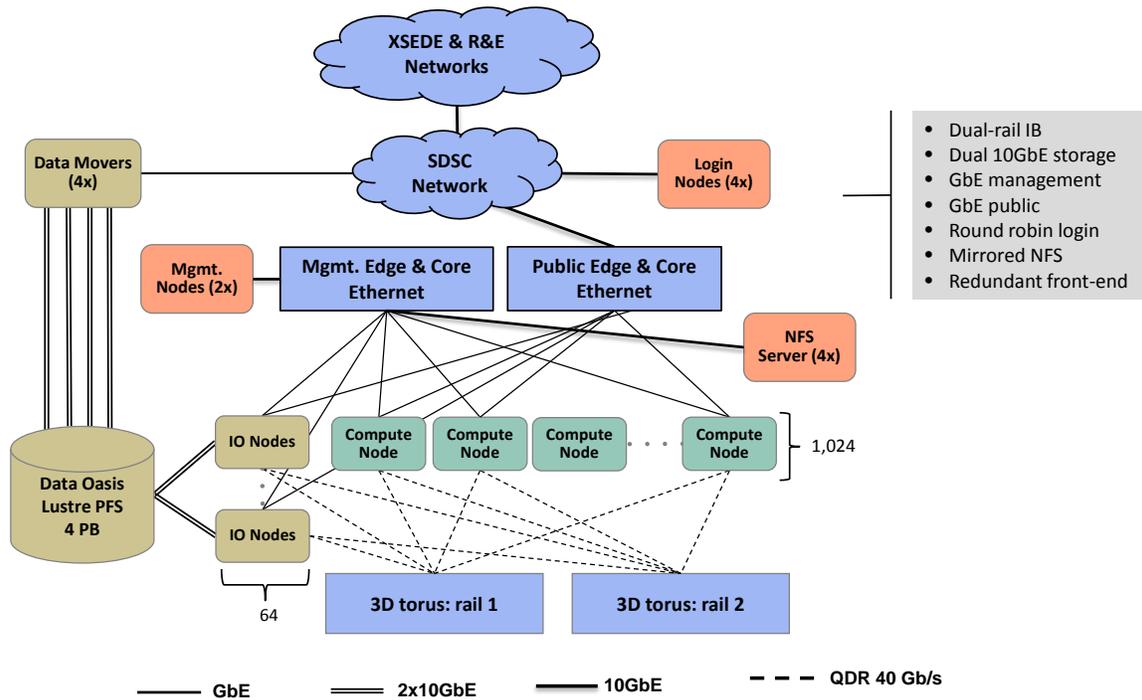


Figure 1: Gordon Architecture

Building a Data-Intensive Supercomputer

Gordon is composed of 1,024 compute nodes and 64 I/O nodes. Each of these dual-socket compute nodes has two 8-core 2.6 GHz Intel® Xeon® E5 processors and 64GB of DDR3-1333 memory. Each compute node also has an 80GB Intel SSD that is used as the system disk. Each I/O node has two 6-core 2.67 GHz Intel® Xeon® processor X5650, 48GB of DDR3-1333 memory and 16 300GB Intel® Solid-State Drives 710. The network topology is a dual rail 4x4x4 3D torus of switches with adjacent switches connected by three 4x QDR InfiniBand links (120Gb/s). Compute nodes (16 per switch) and I/O nodes (1 per switch) are connected to the switches by 4x QDR (40Gb/s). The theoretical peak performance of Gordon is 341 teraflops.

Each server platform in the Gordon system features two-socket CPUs configured with the latest Intel Xeon processor E5 integrated I/O, supporting PCI Express 3.0 specification and Intel® Data Direct I/O technologies which makes the processor intelligently and dynamically determine the optimal path for I/O traffic based on the overall system utilization while allowing system memory to remain in a low-power state. This feature reduces processor latency bottlenecks while improving the system performance and memory bandwidth. Gordon also takes advantage of Intel Advanced Vector Extensions to achieve 8 flops/clock cycles, which is twice the performance of any other processor with the same core count and frequency. This feature provides dramatic performance improvements for scientific applications that are dominated by flops.



Figure 2: Gordon consists of 21 racks in a compact, energy-efficient 48U form factor

Gordon is the first HPC system to employ massive amounts of SSDs. The SSDs are served via 64 I/O nodes and each I/O node is capable of more than 560,000 IOPS, or 35 million IOPS for the entire system. The Intel Solid-State Drive 710 series is based on High Endurance Technology (HET) which incorporates multi-level cell (MLC) memory offering the same high levels of performance as single-level cell (SLC) memory. It also delivers nearly the same endurance as SLC-based NAND SSDs, yet utilizes the higher capacity, more cost-effective MLC NAND.

Gordon also features large-memory “supernodes” capable of presenting more than 2 terabytes (TB) of cache-coherent memory via ScaleMP’s vSMP Foundation high performance software aggregation layer. Some applications are not parallel but can take advantage of the large memory and cores in a vSMP supernode. This feature is another key way Gordon supports data-intensive computing.

<u>Gordon Technical Summary</u>	
System Component	Configuration
<u>Intel EM64T Xeon E5 Compute Nodes</u>	
Sockets	2
Cores	16
Clock speed	2.6 GHz
Flop speed	333 Gflop/s
Memory capacity	64 GB
Memory bandwidth	85 GB/s
<u>Intel Xeon Processor X5650-based I/O Nodes</u>	
Sockets	2
Cores	12
Clock speed	2.67 GHz
Memory capacity	48 GB
Memory bandwidth	64 GB/s
SSD Flash memory	4.8 TB
<u>Full System</u>	
Total compute nodes	1024
Total compute cores	16,384
Peak performance	341 Tflop/s
Total memory	64 TB
Total memory bandwidth	87 TB/s
Total flash memory	300 TB
<u>QDR InfiniBand Interconnect</u>	
Topology	3D Torus
Link bandwidth	8 GB/s (bidirectional)
MPI latency	1.3 μ s
<u>Disk I/O Subsystem</u>	
File Systems	NFS, Lustre
Storage capacity (usable)	4 PB
I/O bandwidth	100 GB/s

<u>Gordon Software Environment</u>	
Software Function	Description
Cluster Management	Rocks
Operating System	CentOS
File Systems	NFS, Lustre
Scheduler and Resource Manager	Catalina, TORQUE
XSEDE Software	CTSS
User Environment	Modules
Compilers	Intel & PGI Fortran, C, C++
Message Passing	Intel MPI, MVAPICH, Open MPI
Debugger	DDT
Performance	IPM, mpiP, PAPI, TAU

Data Oasis: Integrated High Performance Parallel File System

All three of SDSC's HPC resources are connected to Data Oasis, SDSC's high-performance parallel file system. The Data Oasis backbone network architecture consists of a pair of Arista 7508 10GB/s Ethernet switches for dual-path reliability and performance. These form the extreme-performance network hub of SDSC's parallel network file system and cloud-based storage with more than 450 active 10GbE connections (capacity for 768 in total). A 10GbE backbone enables systems with three fundamentally different HPC interconnect architectures to interoperate and enable researchers to seamlessly move into these systems from national and campus research networks. The design of Data Oasis was driven by Gordon's performance requirement of 100 GB/s to the parallel file system. Part of the Gordon delivery included one half of this infrastructure, amounting to 2PB of high speed disk (see Figure 3).

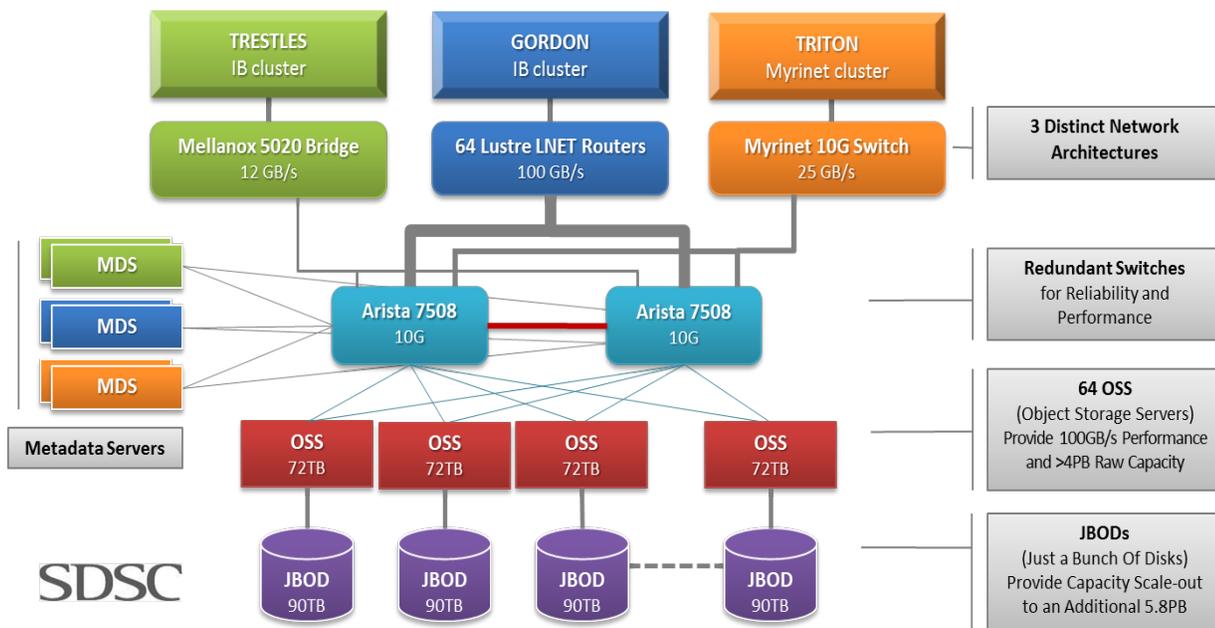


Figure 3: Data Oasis Architecture

Tackling Big Data Challenges

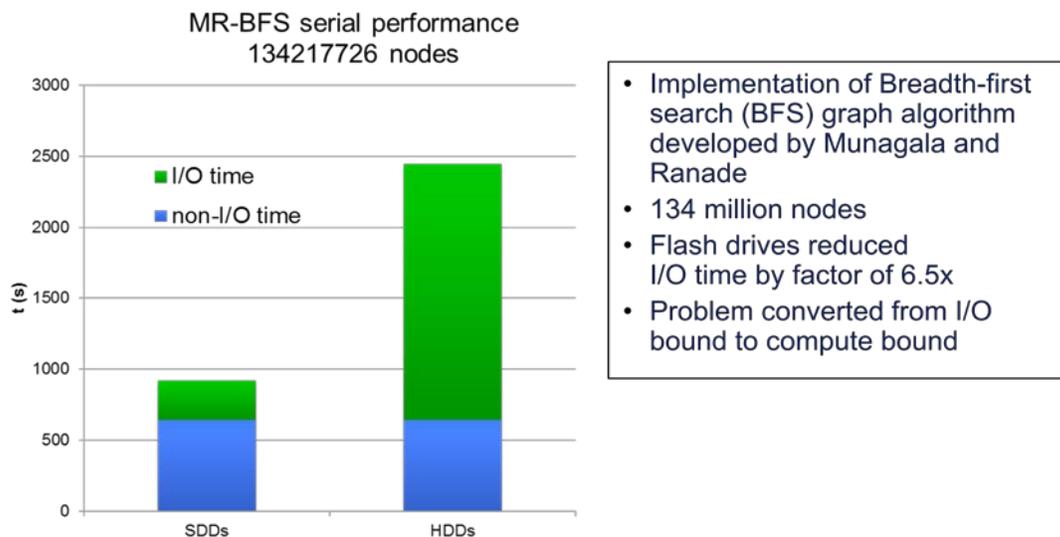
The NSF describes "Big Data" as "large, diverse, complex, longitudinal, and/or distributed datasets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future." Many of these datasets are so voluminous that most conventional computers and software cannot effectively process them.

Gordon is helping address these Big Data challenges by making it easier for researchers to move and analyze data. Its integrated data-intensive infrastructure and software tools allow researchers to effectively manipulate volumes of data at a much faster rate than other systems, which in turn helps extract knowledge and discovery from this data.

Since entering production in early 2012, Gordon has been instrumental in several projects. Among the earliest projects run on Gordon were two designed to predict climate simulations and study severe weather occurrences, including tornadoes, thunderstorms and tropical cyclones.

- Ming Xue, a professor at the School of Meteorology at the University of Oklahoma, leveraged Gordon to gain a better understanding of the impact of severe weather patterns and create accurate numerical simulations of these patterns.
- Atmospheric research scientist Craig Mattocks used Gordon to generate incredibly detailed climate change projections and simulations which will eventually be used to help guide water resource management decisions in South Florida.

Gordon's applications extend well beyond climate and weather patterns. The system is being used to study everything from large graph problems to quantum chemistry. Figure 4 shows how SSDs dramatically improve the I/O performance of network graph applications, reducing I/O time by a factor of 6.5 over traditional hard drives.

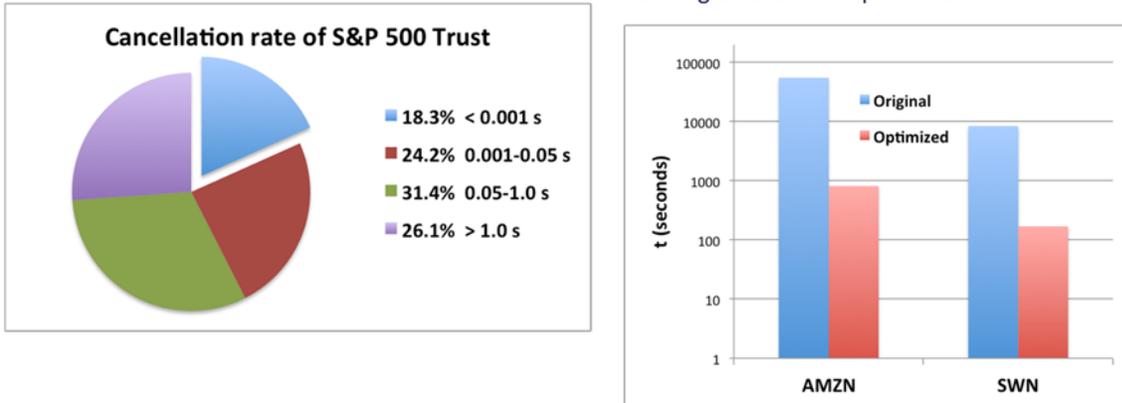


Source: Sandeep Gupta, San Diego Supercomputer Center. Used by permission. 2011

Figure 4: Breadth-first search comparison using SSD and HDD

Mao Ye of the University of Illinois at Urbana-Champaign is using Gordon to analyze massive amounts of NASDAQ market data to study stock market trends. Such research is already proving instrumental in predicting the types of occurrences that lead to market crashes and is expected to lead to revisions in Securities and Exchange Commission (SEC) guidelines and policies. Working closely with SDSC, Ye made dramatic improvements in application performance. Runs which were taking over 10 hours are now completing in 15 minutes (see Figure 5).

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



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

Figure 5: Impact of high-frequency trading on financial markets

Looking Ahead in HPC

The high performance computing industry is one of continual innovation. While groundbreaking technology is constantly being introduced, the next major project is always on the horizon with a new set of technological challenges.

One of the most significant future challenges is exascale computing. While still several years away, advancements such as those made with Gordon provide a platform for understanding the challenges of handling massive amounts of data which is critical enabler for exascale systems.

Acknowledgements

Cray congratulates the San Diego Supercomputer Center and its partners on their success with Gordon. We also extend our condolences to the loved ones of Allan Snavely, who passed away July 14, 2012. Allan made lasting contributions to the Gordon project and the entire supercomputing industry. With his passing, the supercomputing community has lost a true innovator and visionary.

References

<http://www.sdsc.edu/supercomputing/gordon/>

http://www.hpcwire.com/hpcwire/2012-06-05/san_diego_supercomputing_center_complete_data_oasis.html

<http://www.intel.com/content/www/us/en/solid-state-drives/ssd-710-series-brief.html>

<http://www.intel.com/content/www/us/en/processors/xeon/xeon-e5-solutions.html>

© 2013 Cray Inc. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of the copyright owners.

Cray is a registered trademark, and the Cray logo, Cray CS300-AC and Cray GreenBlade are trademarks of Cray Inc. Other product and service names mentioned herein are the trademarks of their respective owners.