PACE
Predictive Analytics Center of Excellence
@ San Diego Supercomputer Center, UCSD

Natasha Balac, Ph.D.
Brief History of SDSC

- **1985-1997:** NSF national supercomputer center; managed by General Atomics

- **1997-2007:** NSF PACI program leadership center; managed by UCSD
  - PACI: Partnerships for Advanced Computational Infrastructure

- **2007-2009:** Internal transition to support more diversified research computing
  - still NSF national “resource provider”

- **2009-future:** Multi-constituency cyberinfrastructure (CI) center
  - provide data-intensive CI resources, services, and expertise for campus, state, and nation
  - Approaching $1B in lifetime contract and grant activity!
Mission: Transforming Science and Society Through “Cyberinfrastructure”

“The comprehensive infrastructure needed to capitalize on dramatic advances in information technology has been termed cyberinfrastructure.”

D. Atkins, NSF Office of Cyberinfrastructure
Today’s Information Technologies Drive 21\textsuperscript{st} Century Solutions

- \textit{Means to an end:} Cyberinfrastructure is the foundation for modern research and education

- \textbf{Cyberinfrastructure components:}
  - Digital data
  - Computers
  - Wireless and wireline networks
  - Personal digital devices
  - Scientific instruments
  - Storage
  - Software
  - Sensors
  - People …
Simulation of SoCal Earthquake
The TeraGrid: Providing High-End Resources to Academic Researchers

- 11 Resource Providers (RPs)
- Grid Infrastructure Group (GIG)
  - Led by UC/ANL
- Annual Budget
  - RPs: ~$50M total
  - GIG: ~$15M
- Compute Power
  - ~20 compute systems
  - >2 PF aggregate
  - Most flops in TACC’s Ranger & NICS’ Kraken
  - New systems coming
- Additional HW Resources
  - >4 PB disk storage
  - Wide-area file systems
  - >60 PB tape storage
• A “data-intensive” supercomputer based on SSD flash memory and virtual shared memory
  • Emphasizes MEM and IO over FLOPS

• A custom-designed system based on COTS to accelerate access to massive databases in all fields of science, engineering, medicine, and social science

• The NSF’s most recent Track 2 award to SDSC
Gordon is designed specifically for data-intensive HPC applications

- Such applications involve “very large data-sets or very large input-output requirements”

- Two data-intensive application classes are important and growing

**Data Mining**

“the process of extracting hidden patterns from data... with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information.”

Wikipedia

**Data-Intensive Predictive Science**

solution of scientific problems via simulations that generate large amounts of data
Why Gordon?

EXPONENTIAL DATA GROWTH

- Growth of digital data is exponential
- Driven by advances in sensors, networking, and storage technologies

RED SHIFT

- While processor speeds have been increasing at Moore’s Law DRAM and disk speeds stay almost constant
- System utilization is cut in half every 18 months?
- It takes 2x time to read your disk every 18 months?
Modern HPC systems have a latency “gap”
Gordon architecture fills that gap

- Registers, $O(kB)$, 1 cycle
- Cache, $O(MB)$, 10 cycles
- Memory, $O(GB)$, 100 cycles
- Disk, $O(TB)$, 10,000 cycles

Flash $O(TB)$, 1,000 cycles
### High Performance Computing (HPC) vs High Performance Data (HPD)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>HPC</th>
<th>HPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key HW metric</td>
<td>Peak FLOPS</td>
<td>Peak IOPS</td>
</tr>
<tr>
<td>Architectural features</td>
<td>Many small-memory multicore nodes</td>
<td>Fewer large-memory SMP nodes</td>
</tr>
<tr>
<td>Typical application</td>
<td>Numerical simulation</td>
<td>Database query</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data mining</td>
</tr>
<tr>
<td>Concurrency</td>
<td>High concurrency</td>
<td>Low concurrency or serial</td>
</tr>
<tr>
<td>Data structures</td>
<td>Data easily partitioned e.g. grid</td>
<td>Data not easily partitioned e.g. graph</td>
</tr>
<tr>
<td>Typical disk I/O patterns</td>
<td>Large block sequential</td>
<td>Small block random</td>
</tr>
<tr>
<td>Typical usage mode</td>
<td>Batch process</td>
<td>Interactive</td>
</tr>
</tbody>
</table>
SDSC Wins SC09 Data Challenge With Dash Prototype System

- 5.2 Tflops prototype of Gordon with 4 TB of flash
- New Teragrid resource that has already sped up some scientific calculations by 2 orders-of-magnitude
Exponential Growth of Research Data
Protein Data Bank @ SDSC
SDSC’s Triton Resource
...computational and data resource for UC researchers

**Data Oasis**
- 2 - 4 Petabytes
- Large-scale Storage for research data sets
- Administered with the UCSD Libraries

**PDAF (Petascale Data Analysis Facility)**
- Designed for analysis of very large data sets

**Triton Compute Cluster**
- Condo-style cluster “starter” to create expandable UCSD computing resources.

**Repository for valuable research databases**

**UCSD/UC Research Labs and Sensor Nets**

**Unique computer for extreme-scale data analysis**

**Seed for UCSD’s “home grown” supercomputer**
Questions?

- [www.sdsc.edu](http://www.sdsc.edu)
- For further information, contact Natasha Balac (natashab@sdsc.edu)