Data-Intensive Solutions at SDSC

Michael L. Norman
Interim Director, SDSC
Distinguished Professor of Physics
mlnorman@ucsd.edu

Michael L. Norman
Principal Investigator
Interim Director, SDSC

Allan Snavely
Co-Principal Investigator
Project Scientist

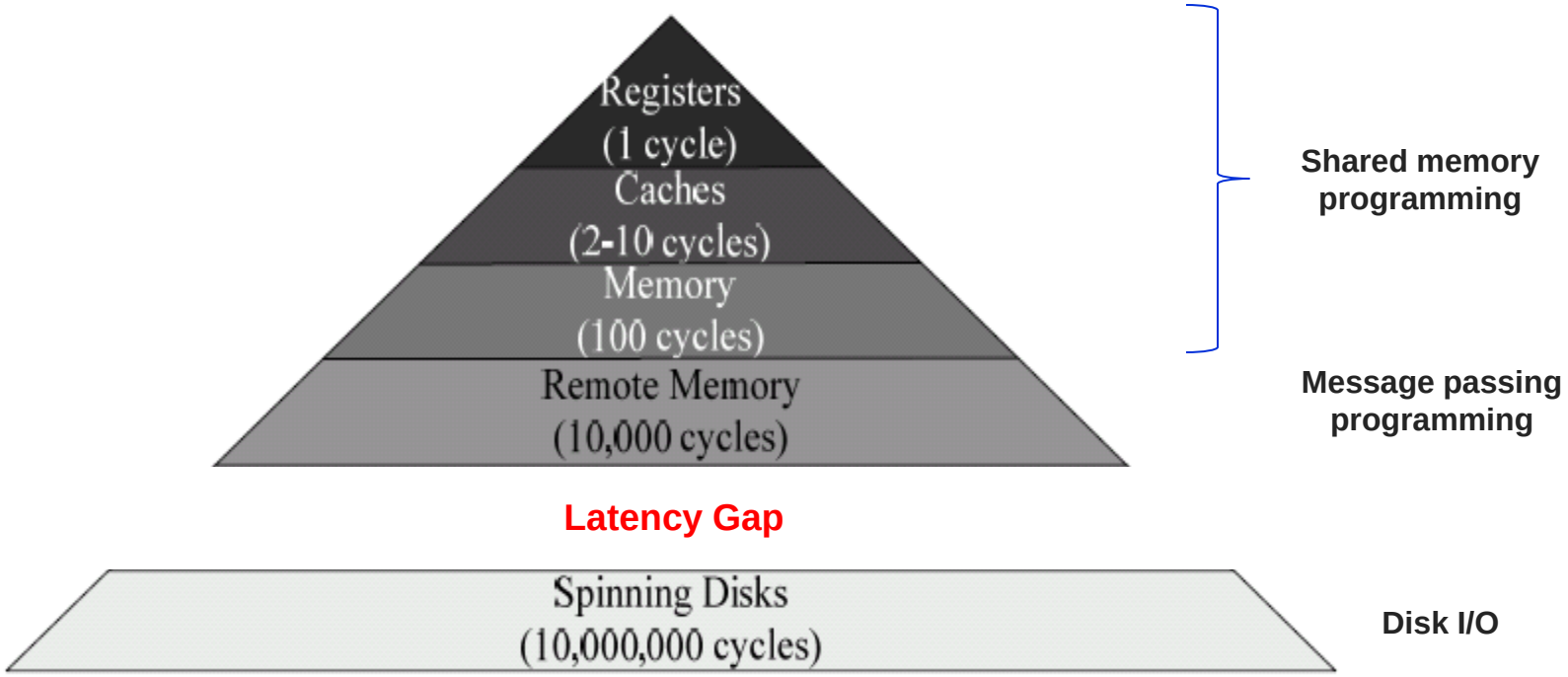
What is Gordon?

- **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 most recent Track 2 award to the San Diego Supercomputer Center (SDSC)**
- **Coming Summer 2011**

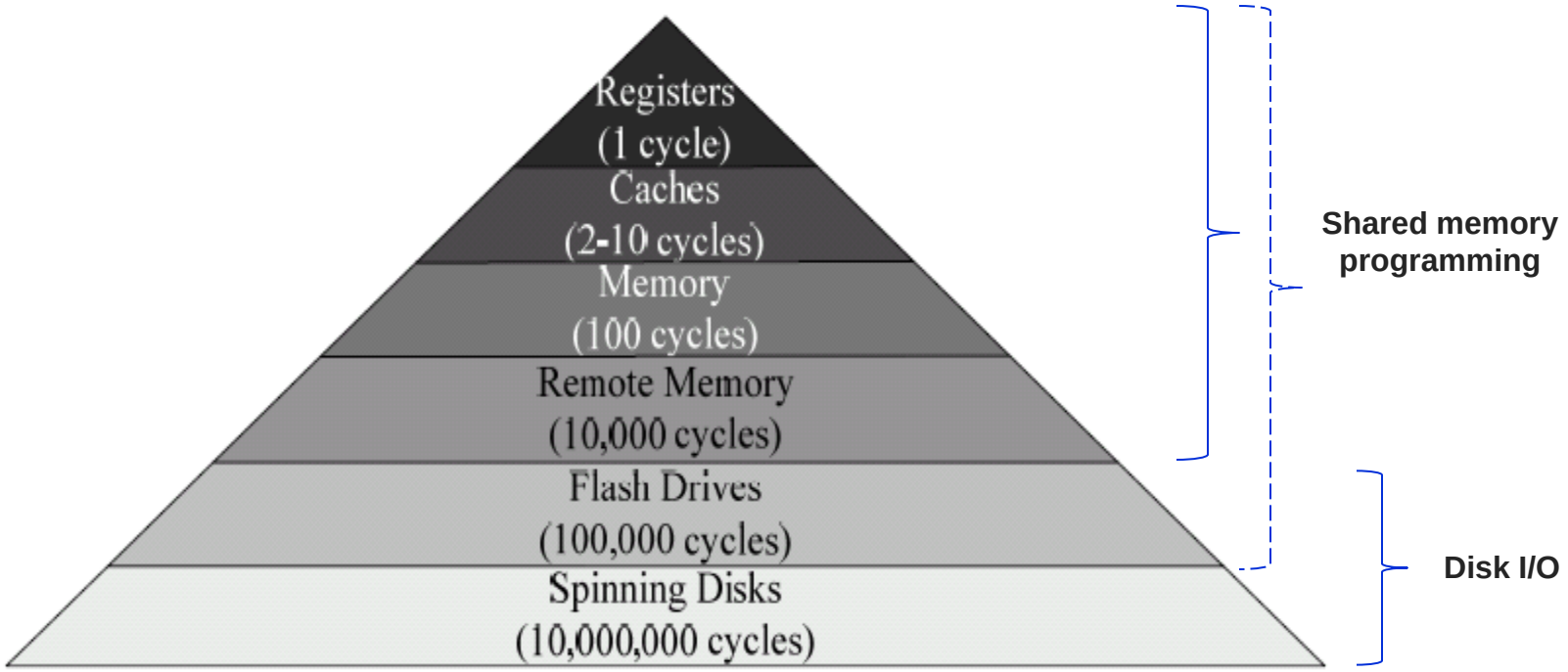
Why Gordon?

- **Growth of digital data is exponential**
  - "data tsunami"
- **Driven by advances in digital detectors, networking, and storage technologies**
- **Making sense of it all is the new imperative**
  - data analysis workflows
  - data mining
  - visual analytics
  - multiple-database queries
  - data-driven applications

The Memory Hierarchy of a Typical HPC Cluster

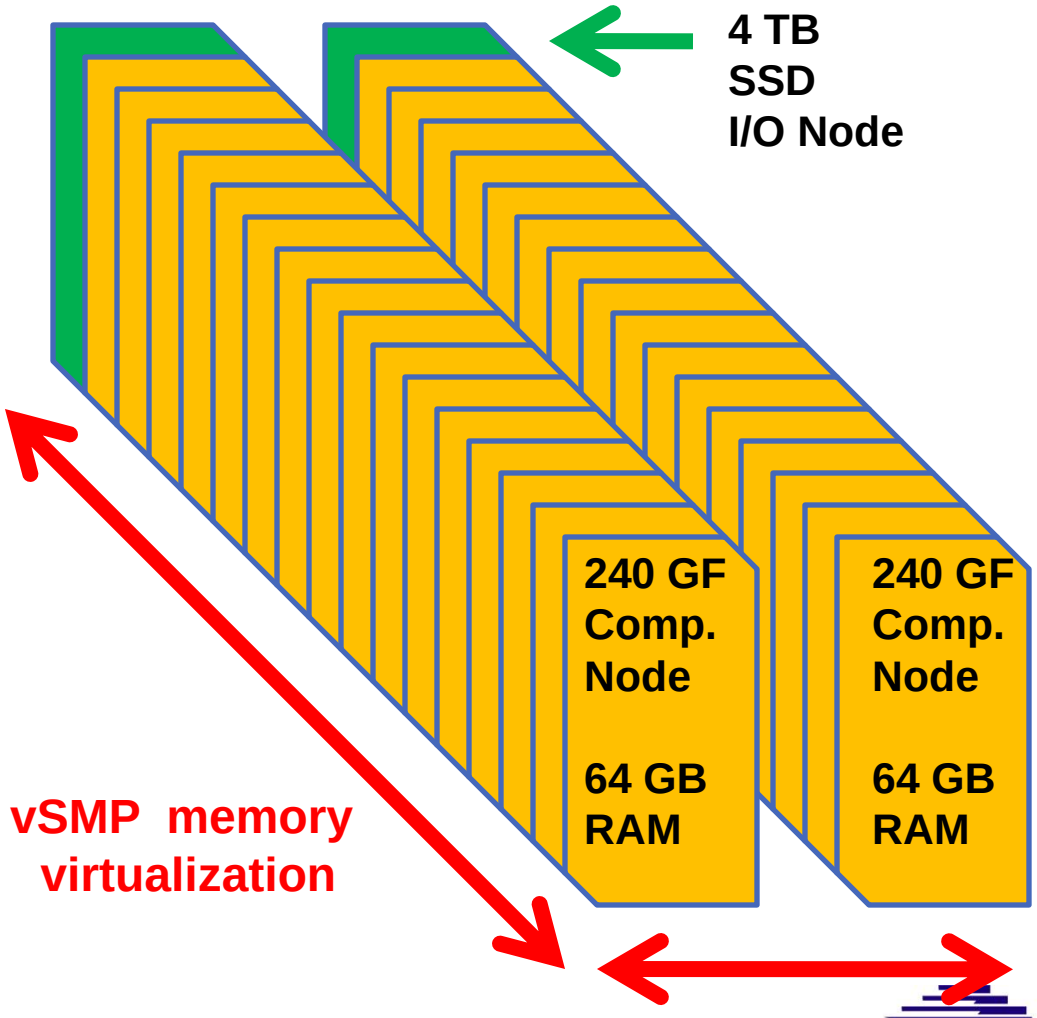


The Memory Hierarchy of Gordon



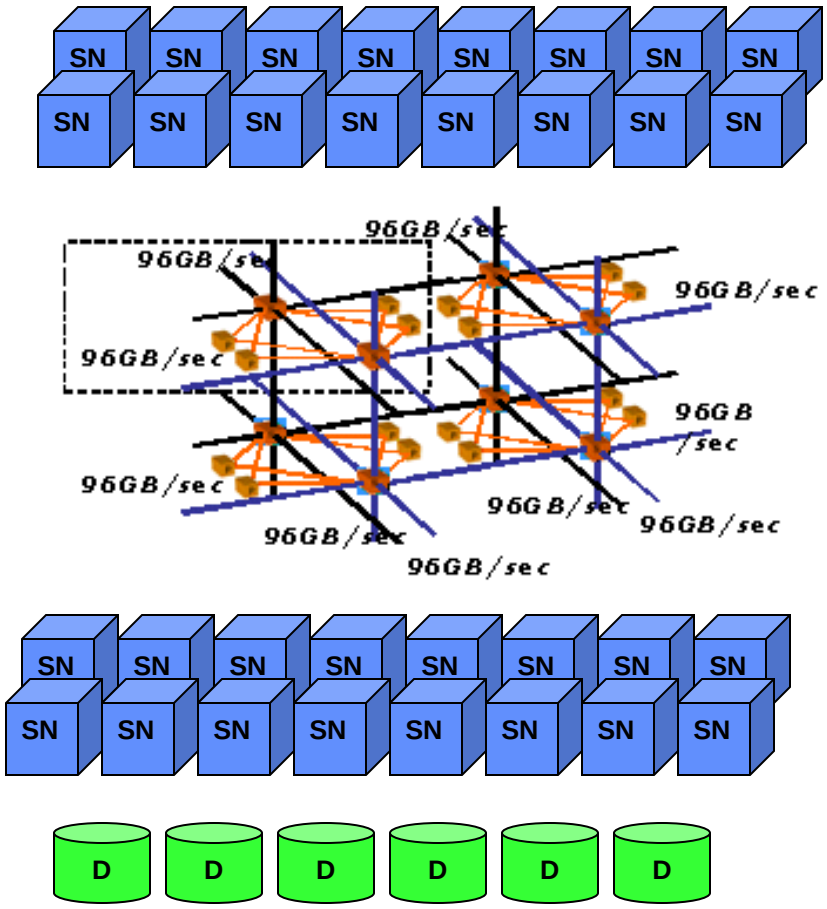
Gordon Architecture: "Supernode"

- **32 Appro Extreme-X compute nodes**
  - Dual processor Intel Sandy Bridge
    - 240 GFLOPS
    - 64 GB
- **2 Appro Extreme-X IO nodes**
  - Intel SSD drives
    - 4 TB ea.
    - 560,000 IOPS
- **ScaleMP vSMP virtual shared memory**
  - 2 TB RAM aggregate
  - 8 TB SSD aggregate



Gordon Architecture: Full Machine

- **32 supernodes = 1024 compute nodes**
- **Dual rail QDR Infiniband network**
  - 3D torus (4x4x4)
- **4 PB rotating disk parallel file system**
  - >100 GB/s



Gordon Aggregate Capabilities

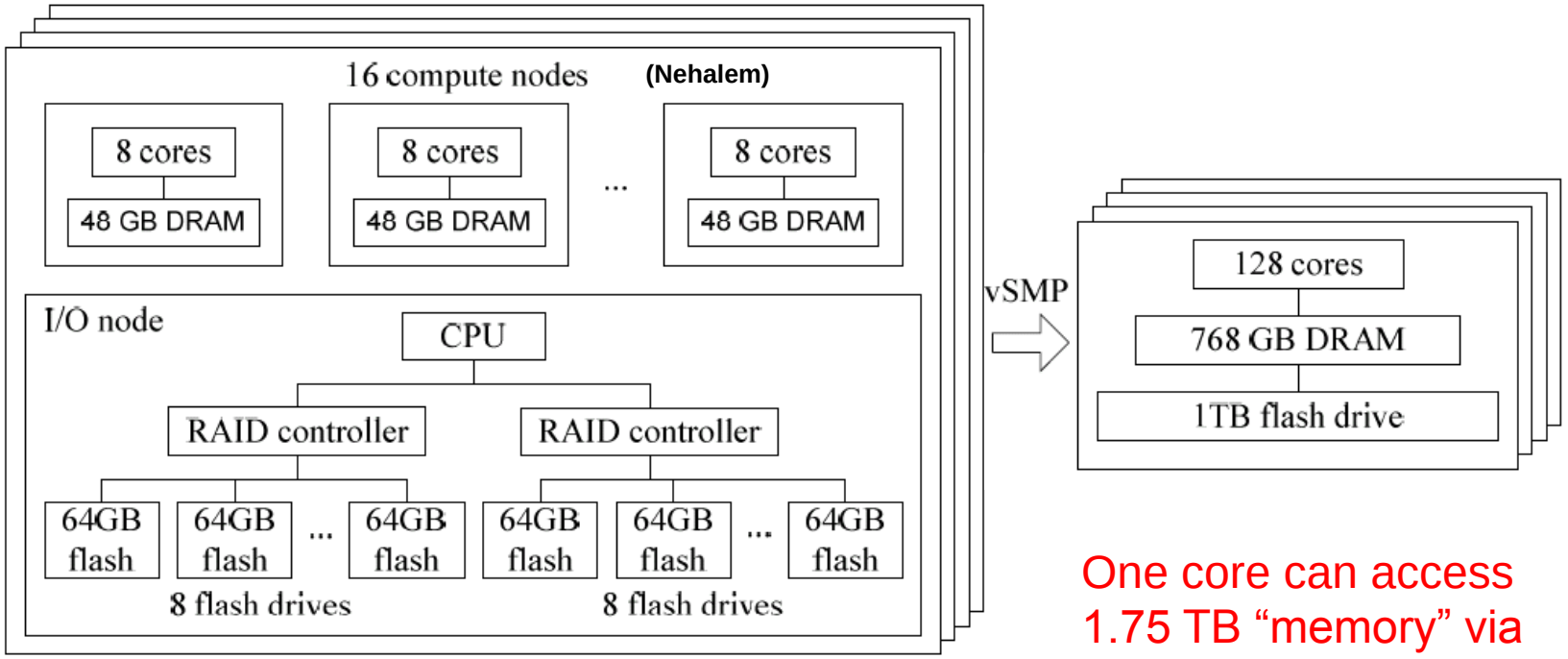
Speed	245 TFLOPS
Mem (RAM)	64 TB
Mem (SSD)	256 TB
Mem (RAM+SSD)	320 TB
Ratio (MEM/SPEED)	1.31 BYTES/FLOP
IO rate to SSDs	35 Million IOPS
Network bandwidth	16 GB/s bi-directional
Network latency	1 μsec.
Disk storage	4 PB
Disk IO Bandwidth	>100 GB/sec

Dash:

a working prototype of Gordon

Dash Architecture

4 supernodes = 64 physical nodes

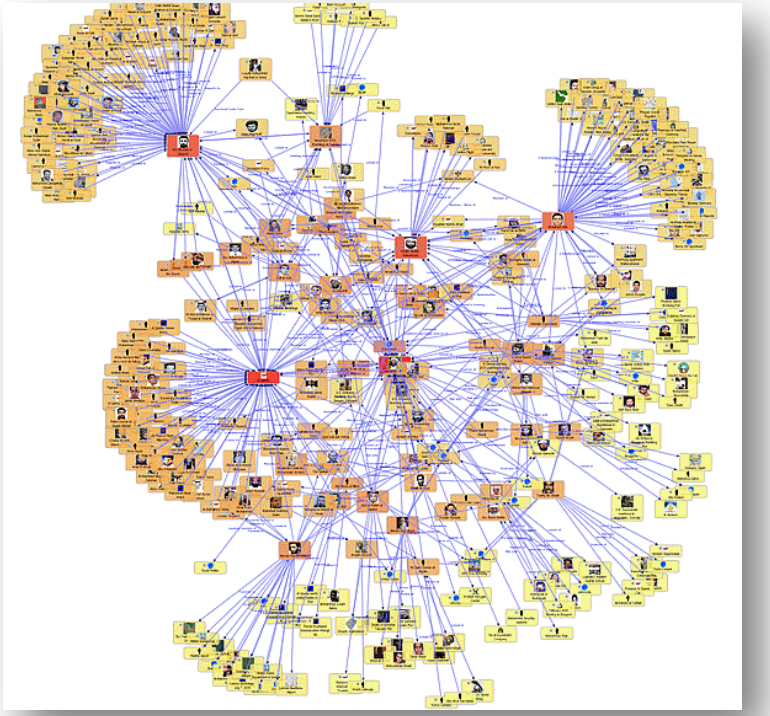


One core can access 1.75 TB “memory” via vSMP

NIH Biological Networks Pathway Analysis

- **Interaction networks or graphs occur in many disciplines, e.g. epidemiology, phylogenetics and systems biology**
- **Performance is limited by latency of a Database query and aggregate amount**

  of fast storage available

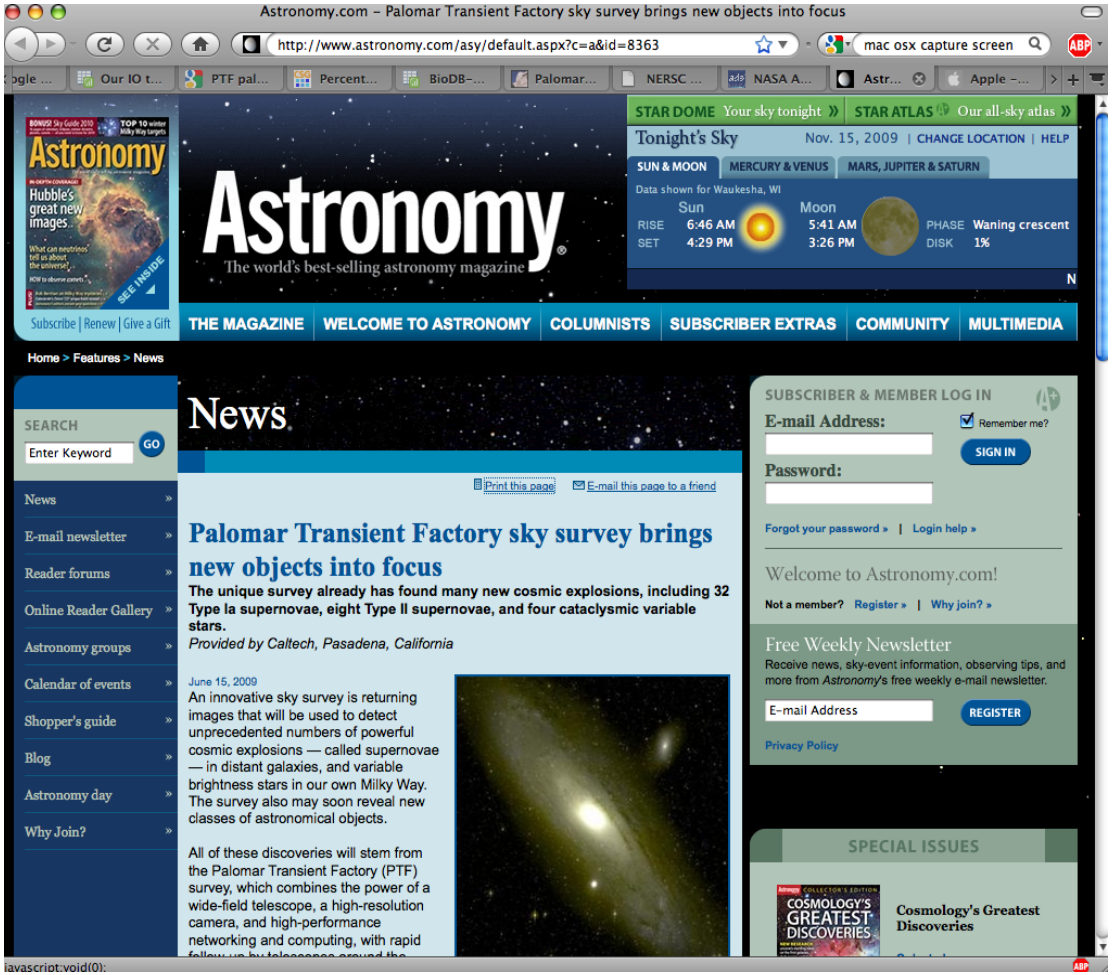


Biological Networks Query timing

TABLE 11: QUERY RESPONSE TIMES OF POPULAR QUERIES IN BIOLOGICAL NETWORKS ON DIFFERENT STORAGE MEDIA (HARD DISK, SSD AND MEMORY) AND THEIR SPEED-UP IN COMPARISON TO HARD DISK.

Query	Q2C	Q3D	Q5F	Q6G	Q7H
RAMFS (vSMP)	11338ms (1.42x)	62850ms (3.60x)	3ms (186x)	17957ms (1.54x)	211ms (5.64x)
SSD	11120ms (1.45x)	176873ms (1.28x)	11ms (50.73x)	24879ms (1.11x)	495ms (2.41s)
HDD	16090ms	226023ms	558ms	27661ms	1191ms

Palomar Transient Factory (PTF)



- Nightly wide-field surveys using Palomar Schmidt telescope
- Image data sent to LBL for archive/analysis
- 100 new transients every minute
- Large, random queries across multiple databases for IDs

PTF-DB Transient Search

	Forward Q1	Backward Q1
DASH-IO-SSD	11s (145x)	100s (24x)
Existing DB	1600s	2400s

Random Queries requesting very small chunks of data about the candidate observations

*Dash wins SC09 Storage Challenge at SC09**

***beating a team led by Alex Szalay and Gordon Bell**

Conclusions

- **Gordon architecture customized for data-intensive applications, but built entirely from commodity parts**
- **Basically a Linux cluster with**
  - Large RAM memory/core
  - Large amount of flash SSD
  - Virtual shared memory software
  - ➔10 TB of storage accessible from a single core
  - ➔shared memory parallelism for higher performance
- **Dash prototype accelerates real applications by 2-100x relative to disk depending on memory access patterns**
  - Random I/O accelerated the most
- **Dash prototype beats all commercial offerings in MB/s/$, IOPS/$, IOPS/GB, and IOPS/Watt**

Cost:Performance (Feb, 2010)

	HDD (SATA)	DASH SSD IO	DASH SUPER	FUSION IO SLC	SUN F5100
GB	2048	1024	768	160	480
$/GB	~0.15	19.43	112.63	41.06	90.62
MB/s/$	~0.4	0.16	0.49	0.12	0.07
IOPS/$	0.4-1.0	28	52	18	9
IOPS/GB	0.05-0.1	549	5853	725	828
5000 IOPS cost	$10000	176.96	96.22	283.17	547.22

Cost of DASH includes storage, n/w, processors.

Power Metrics

TABLE 2. COMPARISON OF POWER METRICS BETWEEN SSD AND HDD.

	DRAM 7x2 GB Dimms (14 GB)	Flash SSD 64GB	HDD 2TB
Active Power	70 W	2.4 W	11 W
Idle Power	35 W	0.1 W	7 W
IOPS per Watt	307	712	35