HPEC
September 16, 2010

Dr. Mike Norman, PI
Dr. Allan Snavely, Co-PI

Project Goals and Overview

Dr. Michael Norman
Gordon PI
Director, SDSC
Professor, Physics, UCSD

Dr. Allan Snavely
Gordon Co-PI
Director, PMaC Lab, SDSC
Adjunct Professor, CSE, UCSD

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**

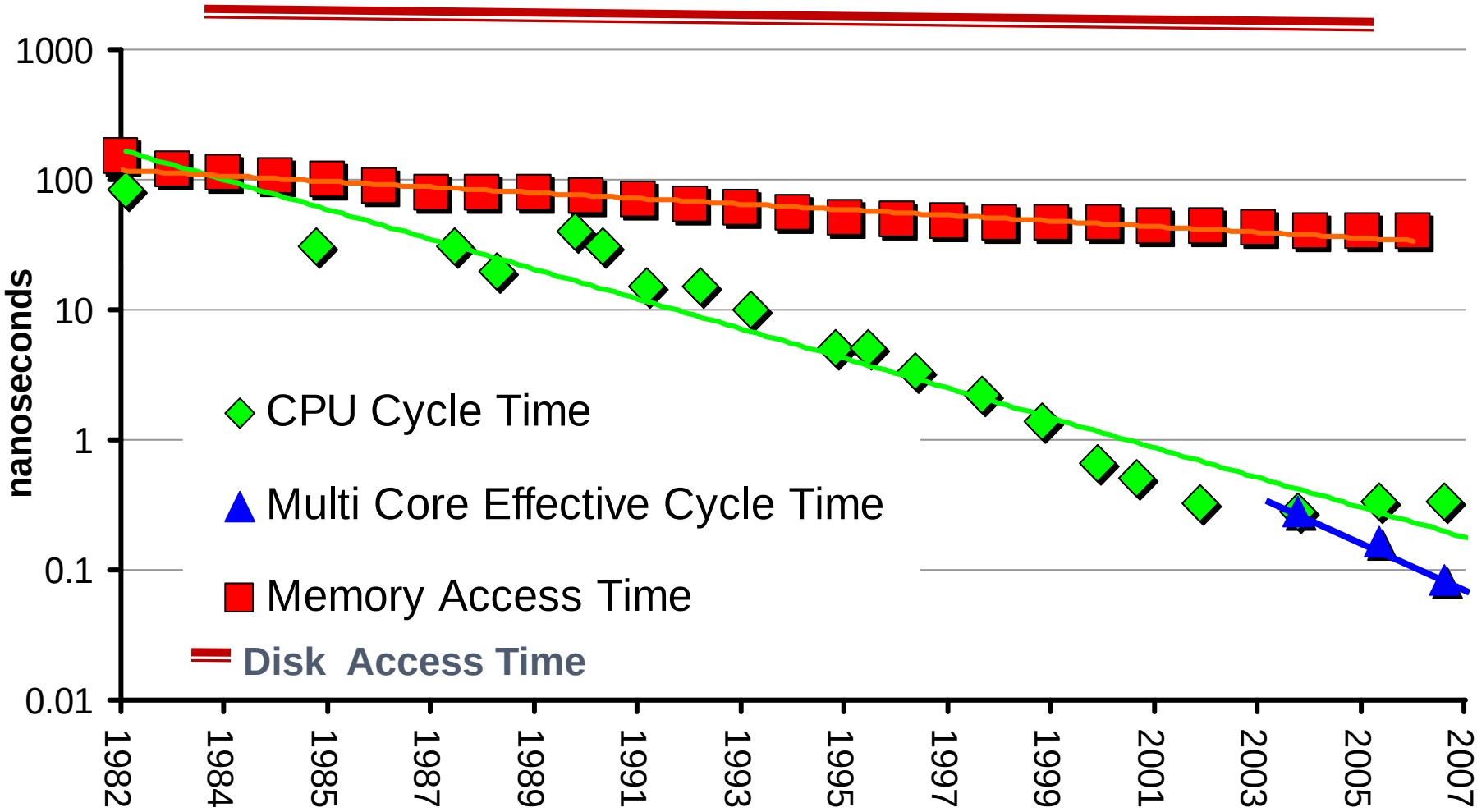
Why Gordon?

- **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

Cosmological Dark Energy Surveys

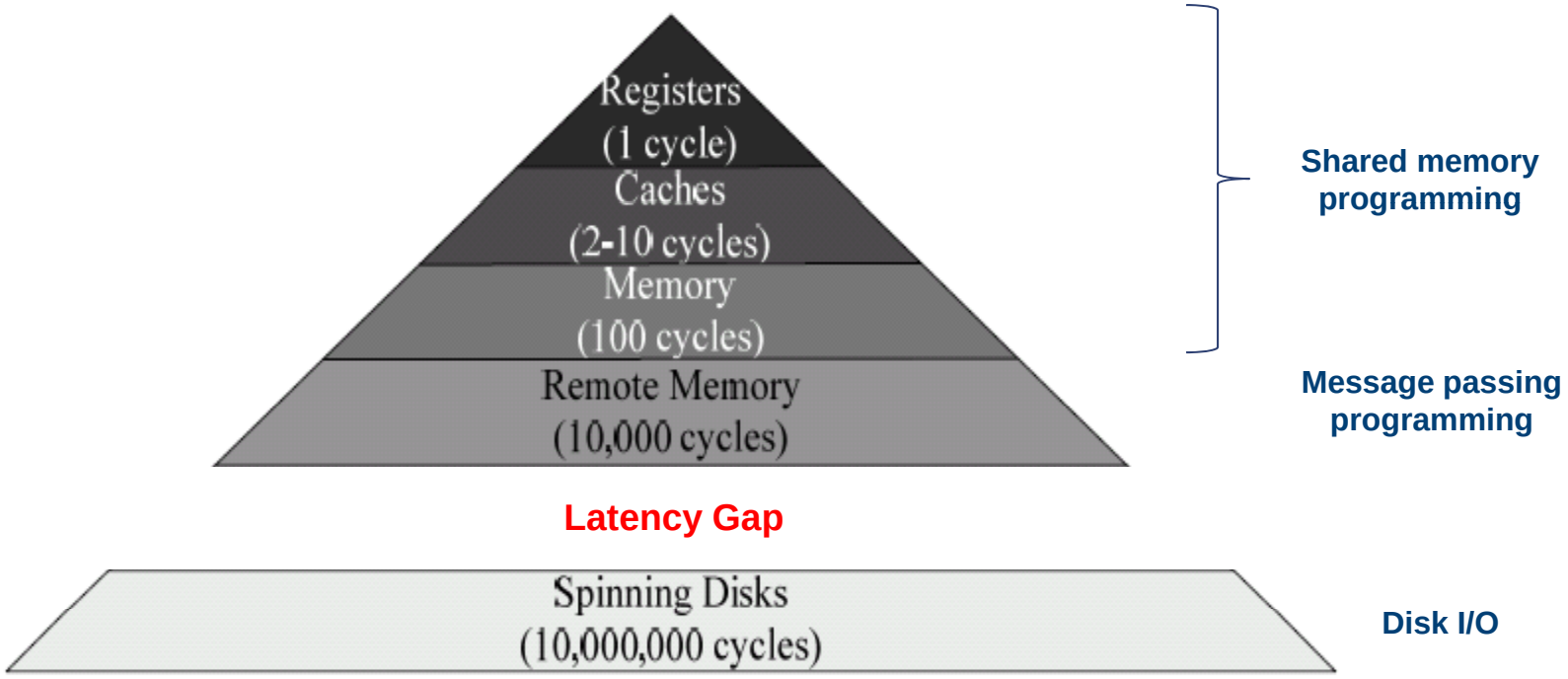
Survey	Area (sq. deg.)	Start date	Image Data (PB)	Object Catalog (PB)
Pan-STARRS-1	30,000	2009	1.5	0.1
Dark Energy Survey	5,000	2011	2.4	0.1
Pan-STARRS-4	30,000	2012	20	0.2
Large Synoptic Survey Telescope	20,000	~2015	60	30
Joint Dark Energy Mission	28,000	~2015	~60	~30

Red Shift: Data keeps moving further away from the CPU with every turn of Moore's Law

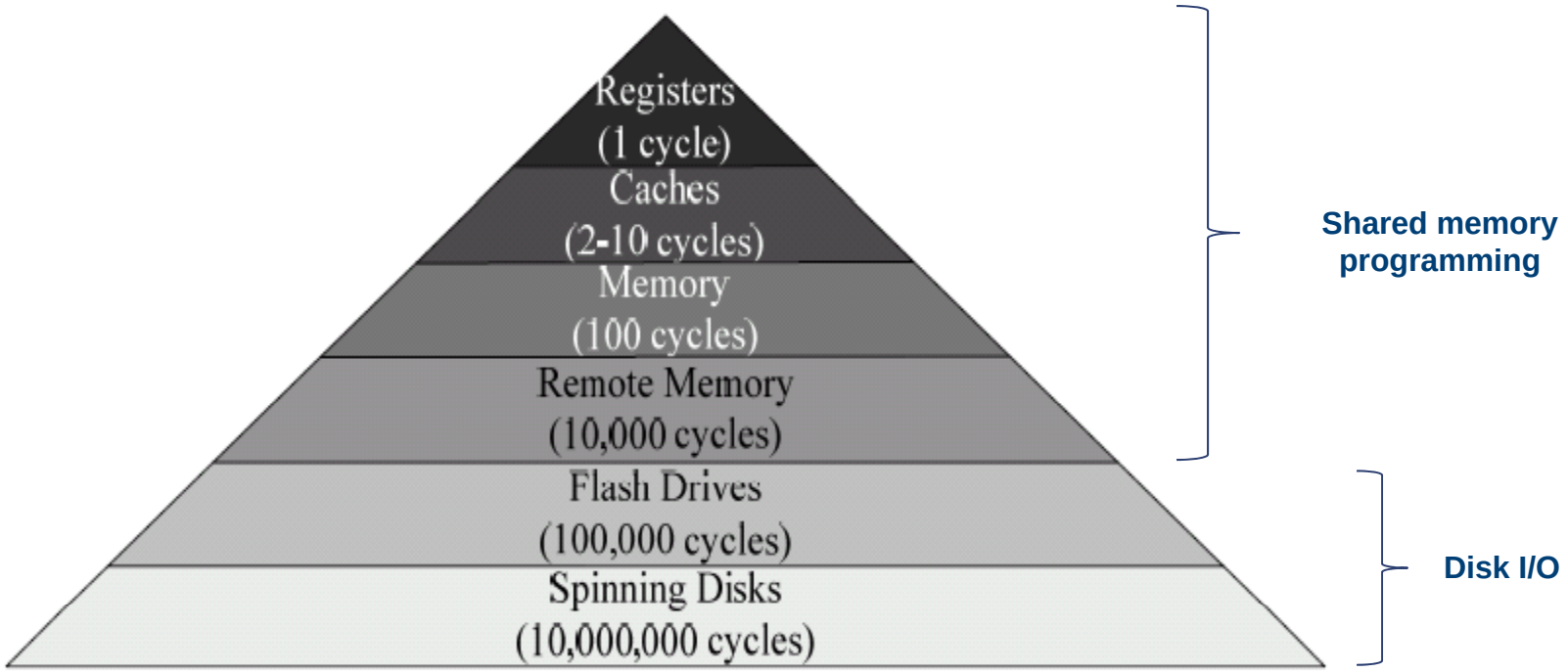


data due to Dean Klein of Micron

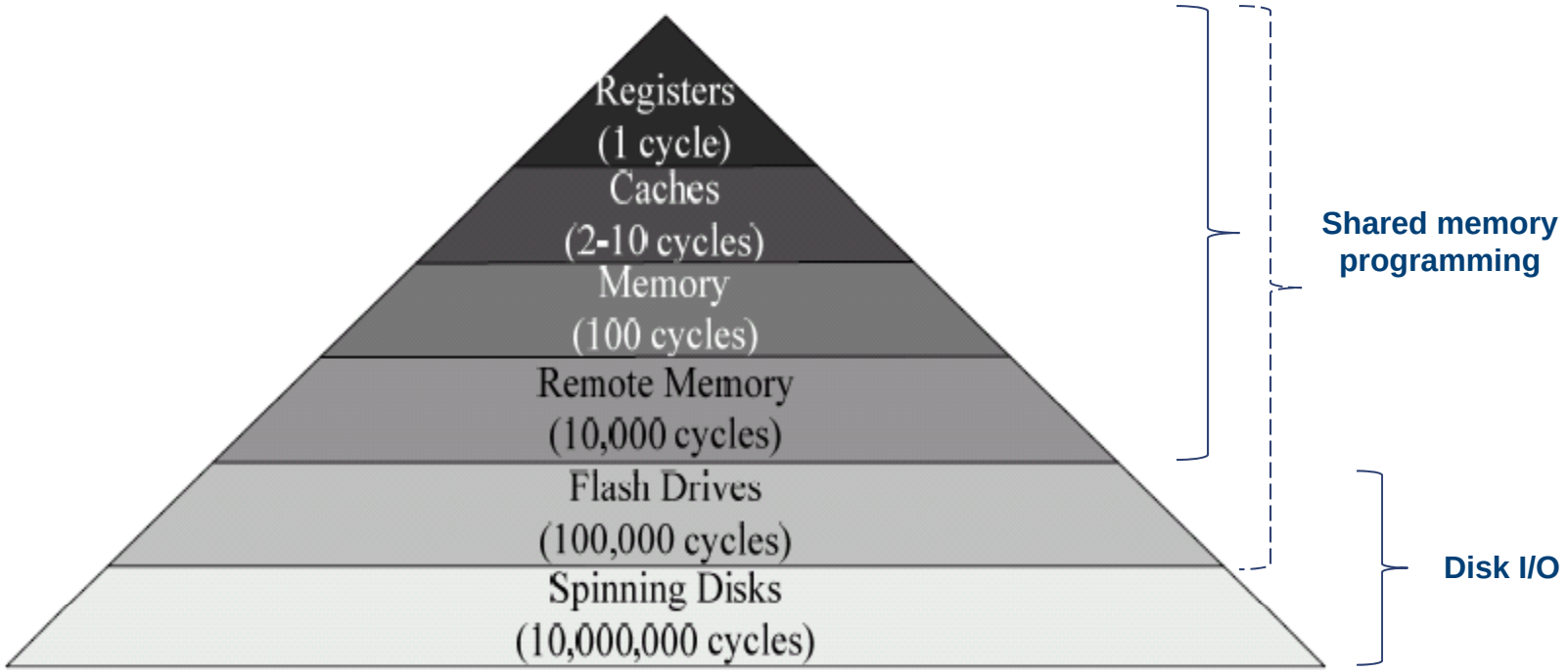
The Memory Hierarchy of a Typical HPC Cluster



The Memory Hierarchy of Gordon



The Memory Hierarchy of Gordon



Gordon is designed specifically for data-intensive HPC applications

- **Such applications involve "*very large data-sets or very large input-output requirements*" *(NSF Track 2D RFP)***
- **Two data-intensive application classes are important and growing**

Data Mining

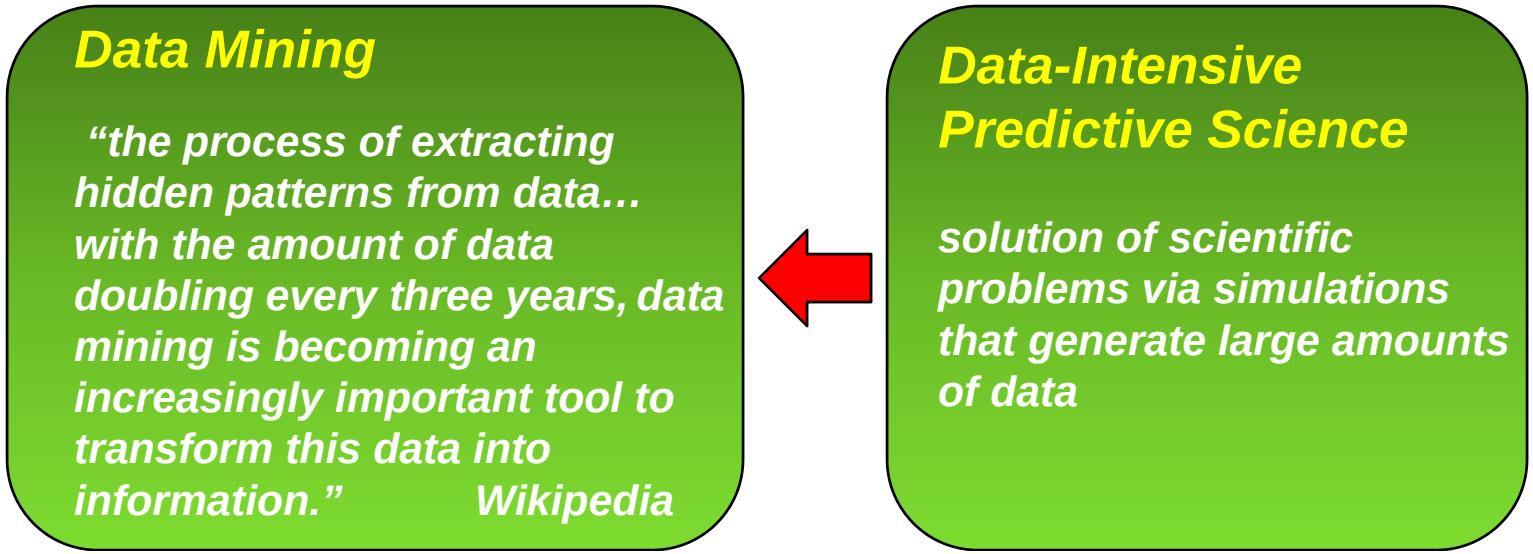
"the process of extracting hidden patterns from data… with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information." Wikipedia

Data-Intensive Predictive Science

solution of scientific problems via simulations that generate large amounts of data

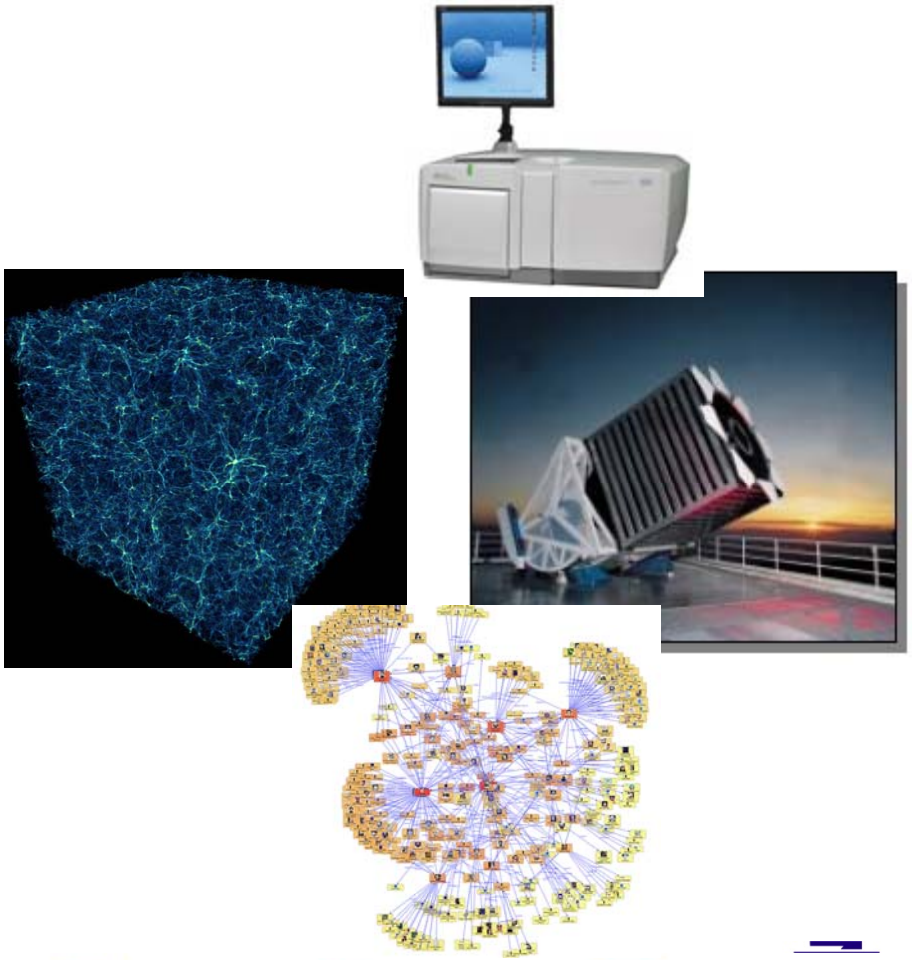
Gordon is designed specifically for data-intensive HPC applications

- **Such applications involve "*very large data-sets or very large input-output requirements*" *(NSF Track 2D RFP)***
- **Two data-intensive application classes are important and growing**



Data mining applications will benefit from Gordon

- ***De novo* genome assembly from sequencer reads & analysis of galaxies from cosmological simulations and observations**
  - *Will benefit from large shared memory*
- **Federations of databases and interaction network analysis for drug discovery, social science, biology, epidemiology, etc.**
  - *Will benefit from low latency I/O from flash*



De novo *genome assembly (1)*

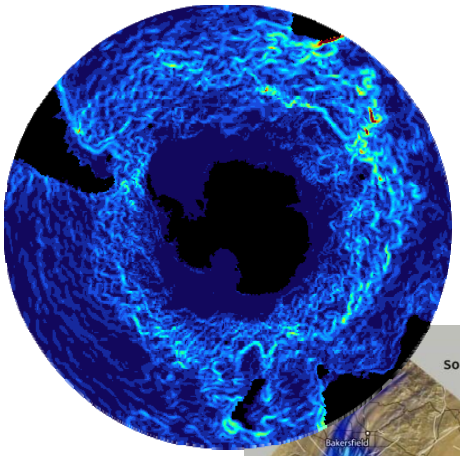
- **Problem**
  - Assemble genome sequence from millions of fragments (reads) generated by high-throughput sequencers *without an assumed template*
- **State-of-the-art codes include**
  - EULER-SR (UCSD)
  - Velvet (EMBL)
  - Edena (Geneva)

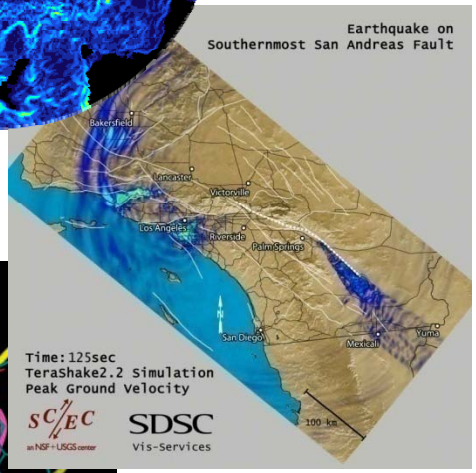
De novo *genome assembly (2)*

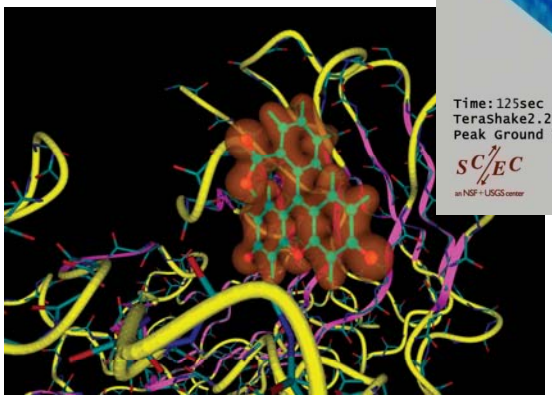
- **Method**
  - Correct reads for errors
  - Construct graph in memory from reads & modify graph to obtain final sequence
- **Performance issues**
  - Interesting problems require large amounts of memory
  - Graph construction hard to parallelize
  - Approach being prototyped with EULER-SR is to parallelize with OpenMP
  - Large shared memory will be essential (~ 500B per base pair)
- **New science capability enabled by Gordon**
  - Assembly of whole human genome within a super-node (~2 TB)
  - Could do 32 simultaneously on Gordon

Data-intensive predictive science will benefit from Gordon

- **Solution of inverse problems in oceanography, atmospheric science, & seismology**
  - *Will benefit from a balanced system, especially large RAM per core & fast I/O*
- **Modestly scalable codes in quantum chemistry & structural engineering**
  - *Will benefit from large shared memory*

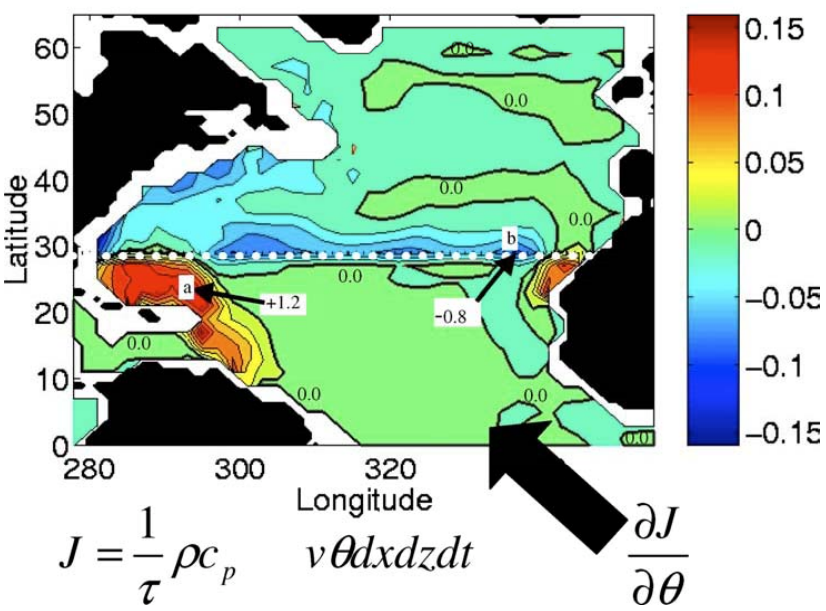






Solution of inverse problems in geoscience (1)

- **Problem**
  - Reconstruct 3D (or 3D+t) fields from measured data supplemented by sophisticated models
- **Examples include**
  - Ocean state estimation (MITgcm 4DVar)
  - Atmospheric data assim-ilation (WRF or ARPS: 3DVar, 4DVar, & EnKF)
  - Full 3D seismic tomography



Sensitivity of heat flux to sea-surface temperature: P.Heimbach, C. Hill, & R, Giering, *Future Generation Computer Systems,* 21, 1356-1371 (2005)

Solution of inverse problems in geoscience (2)

- **Method (in MITgcm)**
  - Solve nonlinear minimization problem by iteratively sweeping through forward and adjoint GCM equations
- **Three-level checkpointing (in MITgcm) requires**
  - 3 forward sweeps
  - 1 adjoint sweep (at ~2.5x the time of forward sweep)

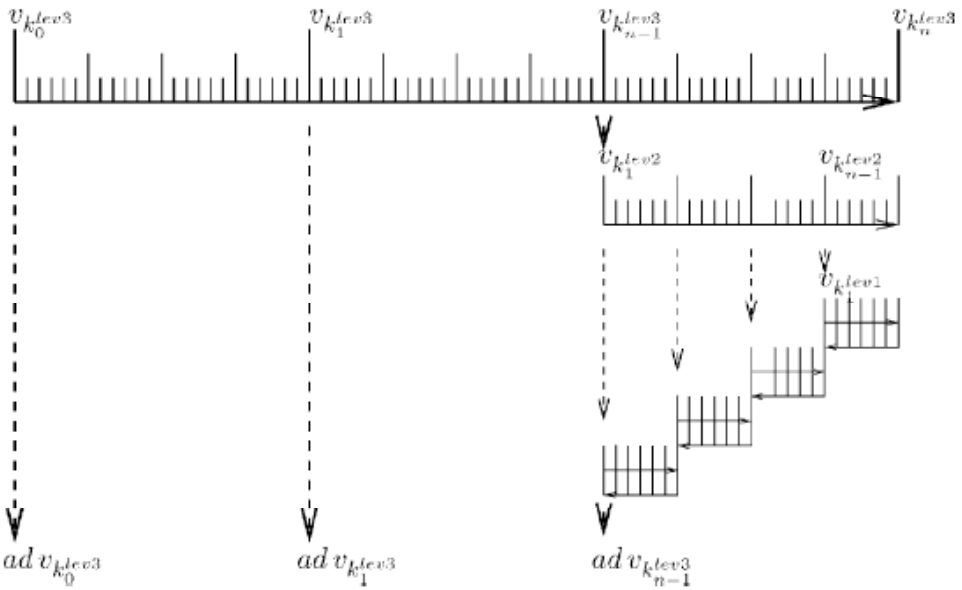


Fig. 3. Schematic view of intermediate dump and restart for 3-level checkpointing.

P.Heimbach, C. Hill, & R, Giering, *Future Generation Computer Systems*, 21, 1356-1371 (2005)

SDSC UCSD APPRO HPC Cluster Solutions intel ScaleMP

Solution of inverse problems in geoscience (3)

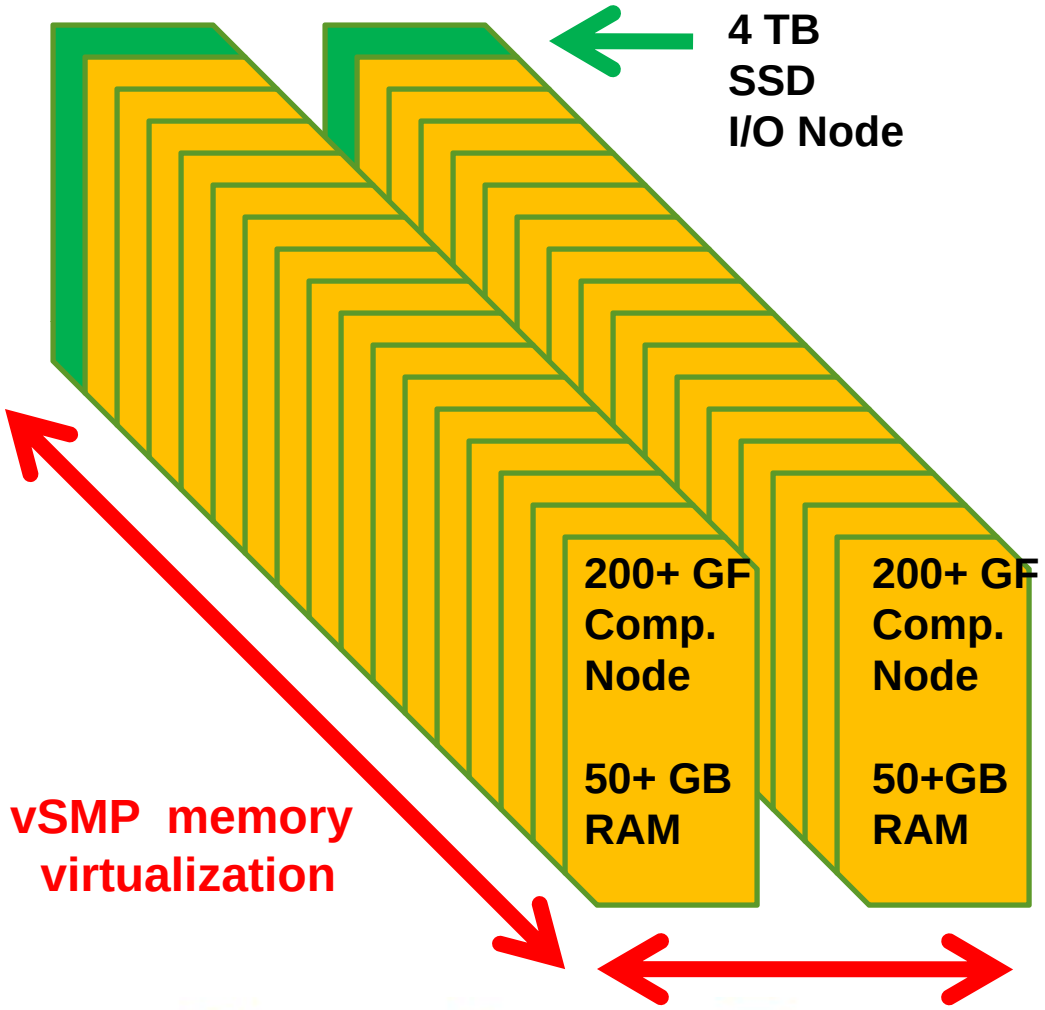
- **Performance issues (in MITgcm)**
  - Adjoint sweep requires state data at every time step
  - Multi-level checkpointing is required that
    - stores state data on disk (or flash) at checkpoints &
    - recomputes state data between checkpoints
  - Performance will benefit from
    - large RAM per core (fewer checkpoints)
    - high I/O bandwidth from flash and/or disk (faster reads)
- **New science capability enabled by Gordon**
  - Higher throughput (x4-5) for current models
  - higher resolution

High Performance Computing (HPC) vs High Performance Data (HPD)

Attribute	HPC	HPD
Key HW metric	Peak FLOPS	Peak IOPS
Architectural features	Many small-memory multicore nodes	Fewer large-memory vSMP nodes
Typical application	Numerical simulation	Database query Data mining
Concurrency	High concurrency	Low concurrency or serial
Data structures	Data easily partitioned e.g. grid	Data not easily partitioned e.g. graph
Typical disk I/O patterns	Large block sequential	Small block random
Typical usage mode	Batch process	Interactive

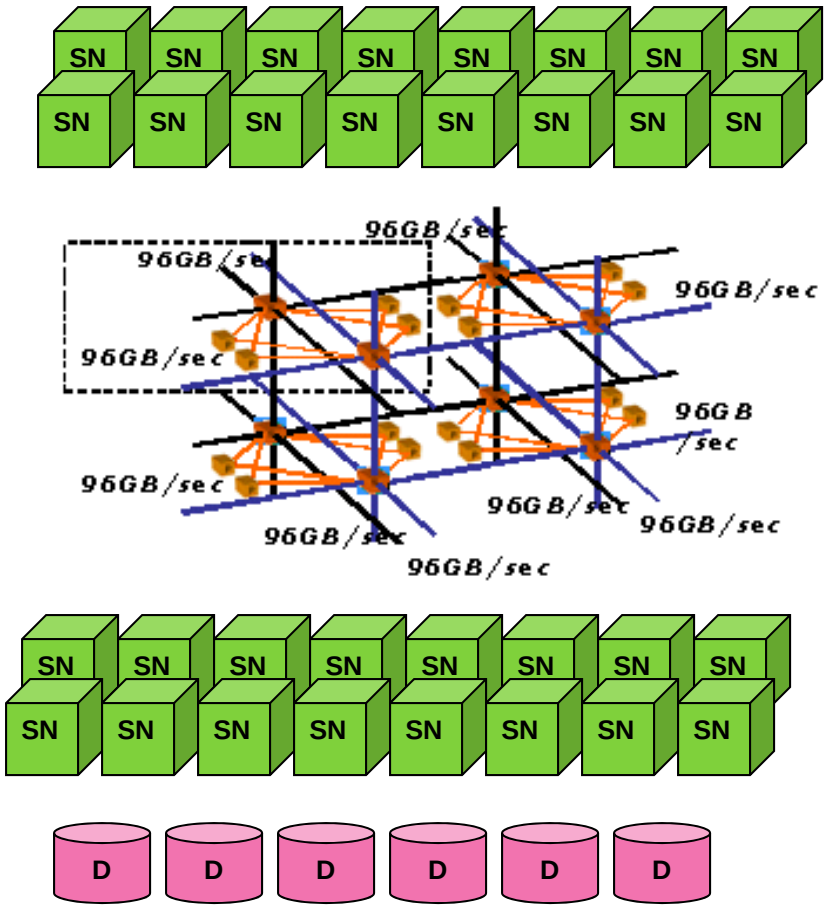
Gordon Architecture: "Supernode"

- **32 Appro Extreme-X compute nodes**
  - Dual processor Intel Sandy Bridge
    - 200+ GFLOPS
    - 50+ 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

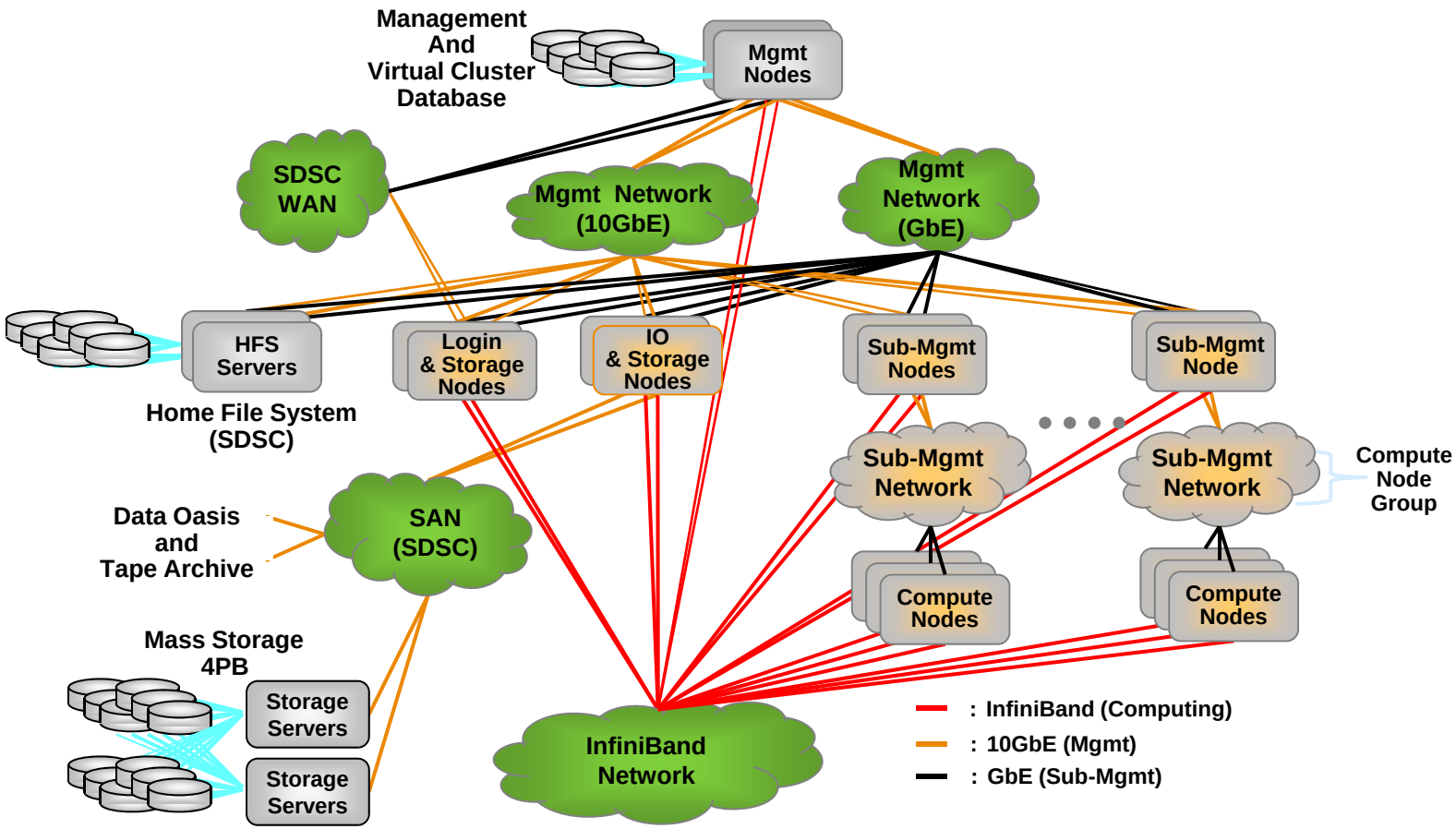


Gordon Architecture: Full Machine

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



Gordon Architecture



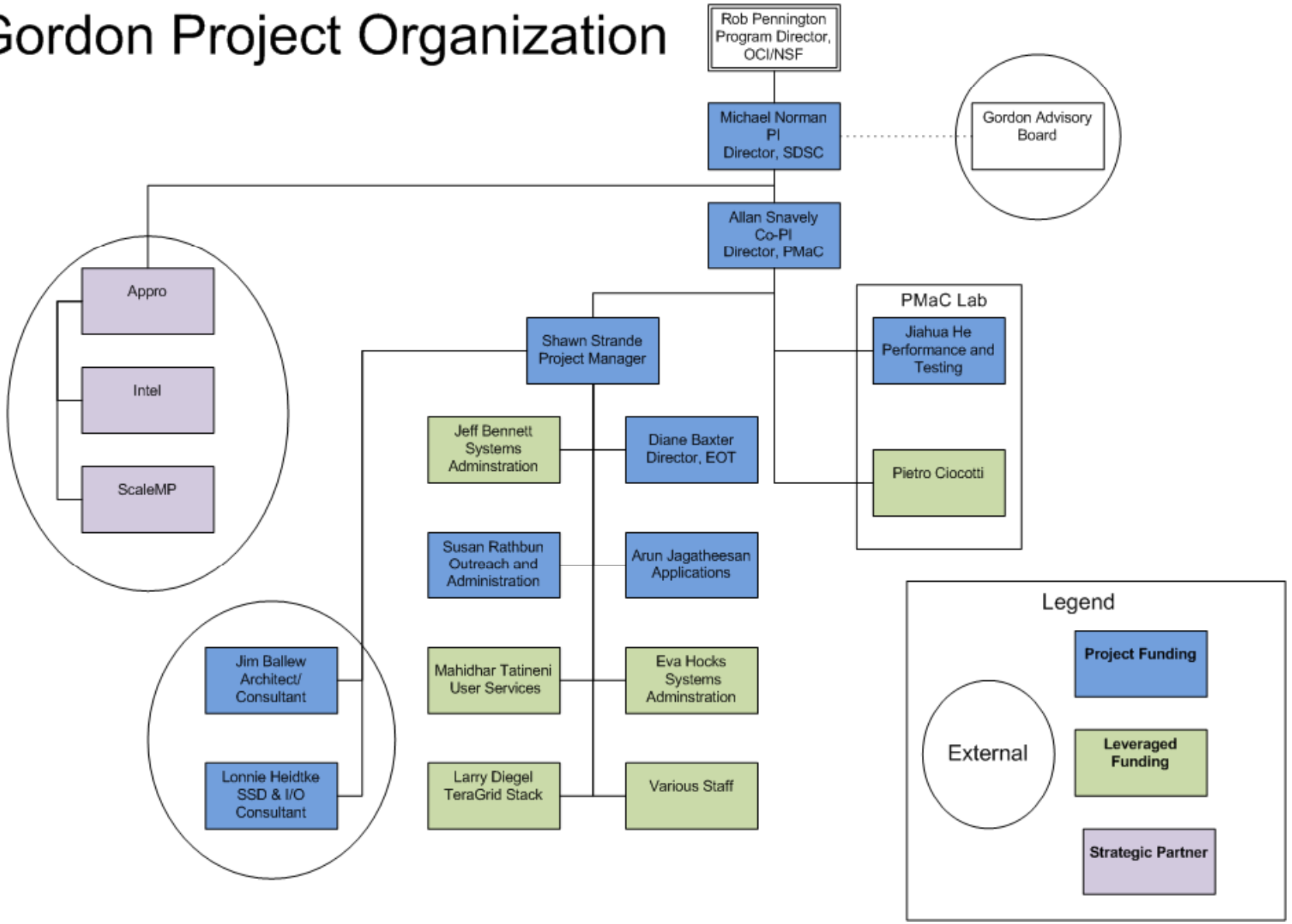
Gordon Technical Description

Speed	200+ TFLOPS
Memory (RAM)	50+ TB
Memory (SSD)	256+ TB
Memory (RAM+SSD)	300+ TB
Ratio (MEM/SPEED)	1.31 BYTES/FLOP
IO rate to SSDs	35 Million IOPS
Network bandwidth	16 GB/s full-duplex
Network latency	1 μsec.
Disk storage (external/PFS)	4 PB
IO Bandwidth to PFS	>100 GB/sec

Project Progress and Milestones

- ✓ Completed the 16, and 32-way vSMP Acceptance Tests
- ✓ Added Dash as a TeraGrid Resource in April 2010
- ✓ Allocated Users on Dash
- ✓ TeraGrid 2010 Presentations, Tutorials and BOFs
- ✓ SC ‘10 Papers
- ✓ SSD Testing and Optimization
- ✓ Critical Design Review
- ✓ Planning for October Deployment of 16 Gordon I/O Nodes
- ✓ Data Center Preparations Underway
- ✓ Data Intensive Workshop to be Held at SDSC in October 2010

Gordon Project Organization



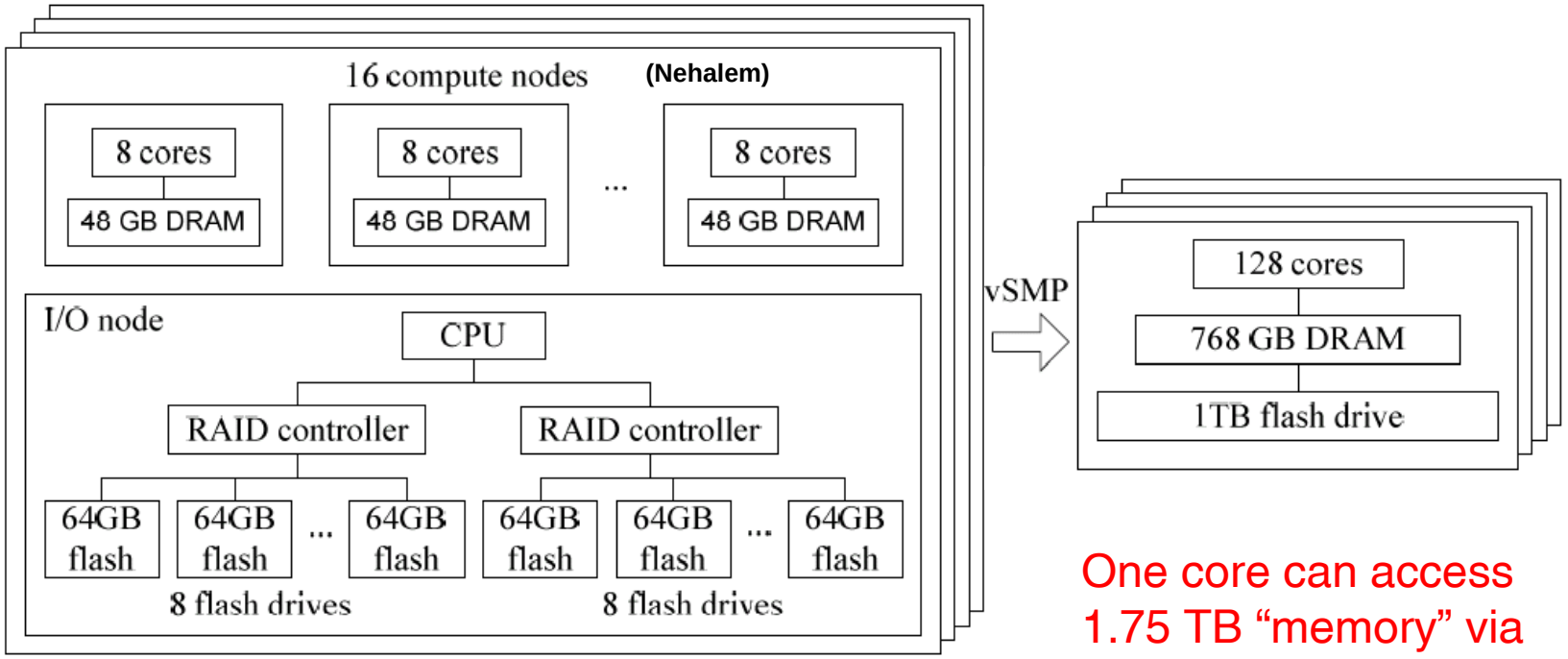
Dash: a working prototype of Gordon

Before Gordon There is Dash

- Dash has been deployed as a risk mitigator for Gordon
- Dash is an Appro cluster that embodies the core architectural features of Gordon and provides a platform for testing, evaluation, and porting/optimizing applications
  - 64 node, dual-socket, 4 core, Nehalem
  - 48GB memory per node
  - 4TB of Intel SLC Flash (X25E)
  - InfiniBand Interconnect
  - vSMP Foundation supernodes
- Using Dash for:
  - SSD Testing (vendors, controllers, RAID, file systems)
  - 16, and 32-Way vSMP Acceptance Testing
  - Early User Testing
  - Development of processes and procedures for systems administration, security, operations, and networking

Dash Architecture

4 supernodes = 64 physical nodes



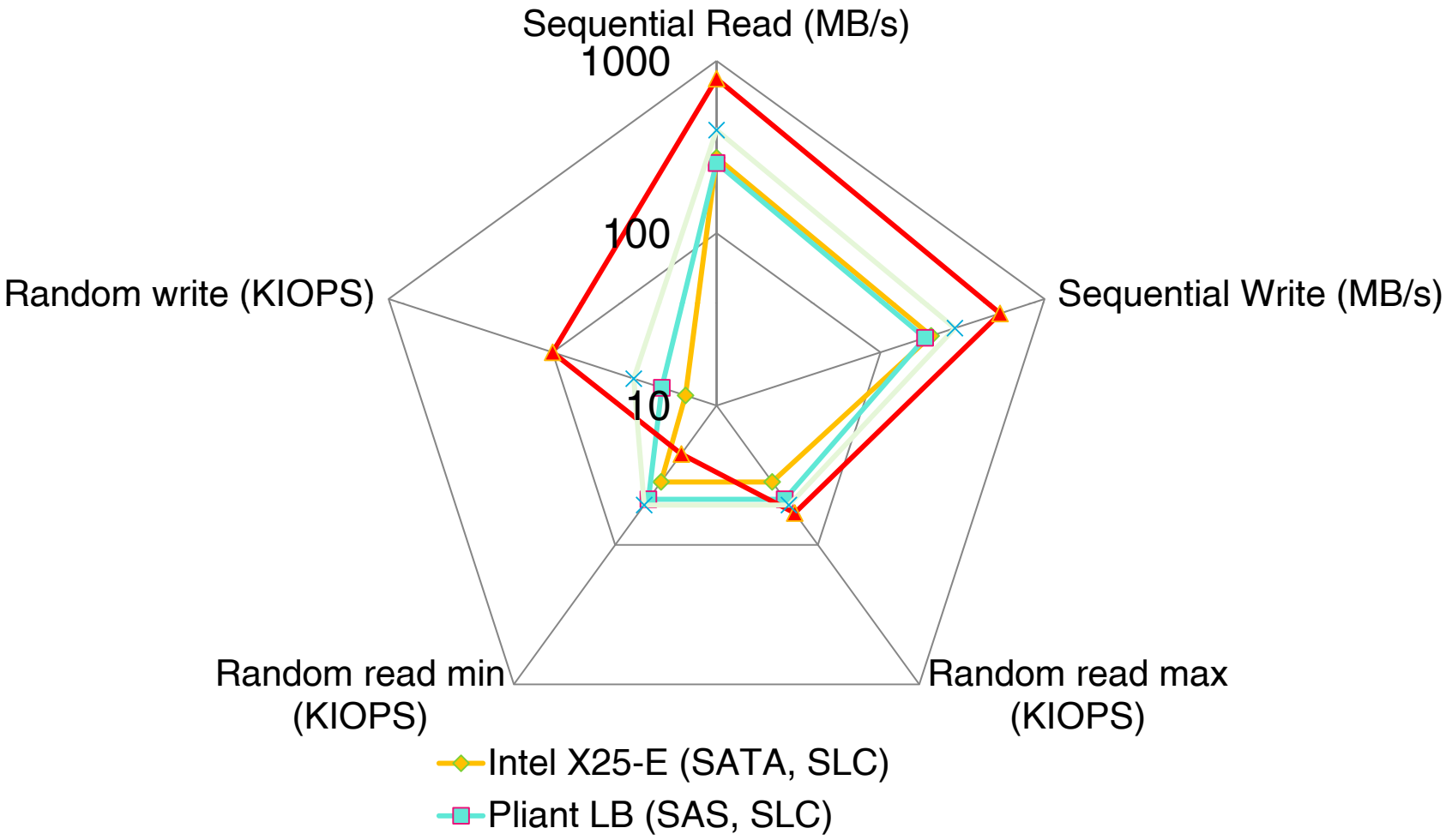
One core can access 1.75 TB “memory” via vSMP

Dash is a Platform for Technology Testing and Performance Evaluation

- Example: Using Dash to assess current SSD technology and performance:
  - MLC vs. SLC
  - SATA vs. SAS drive
  - Direct PCIe connected
  - Over-provisioning
  - Weal leveling
  - ECC
  - Developing a framework for assessing future developments in SSD's

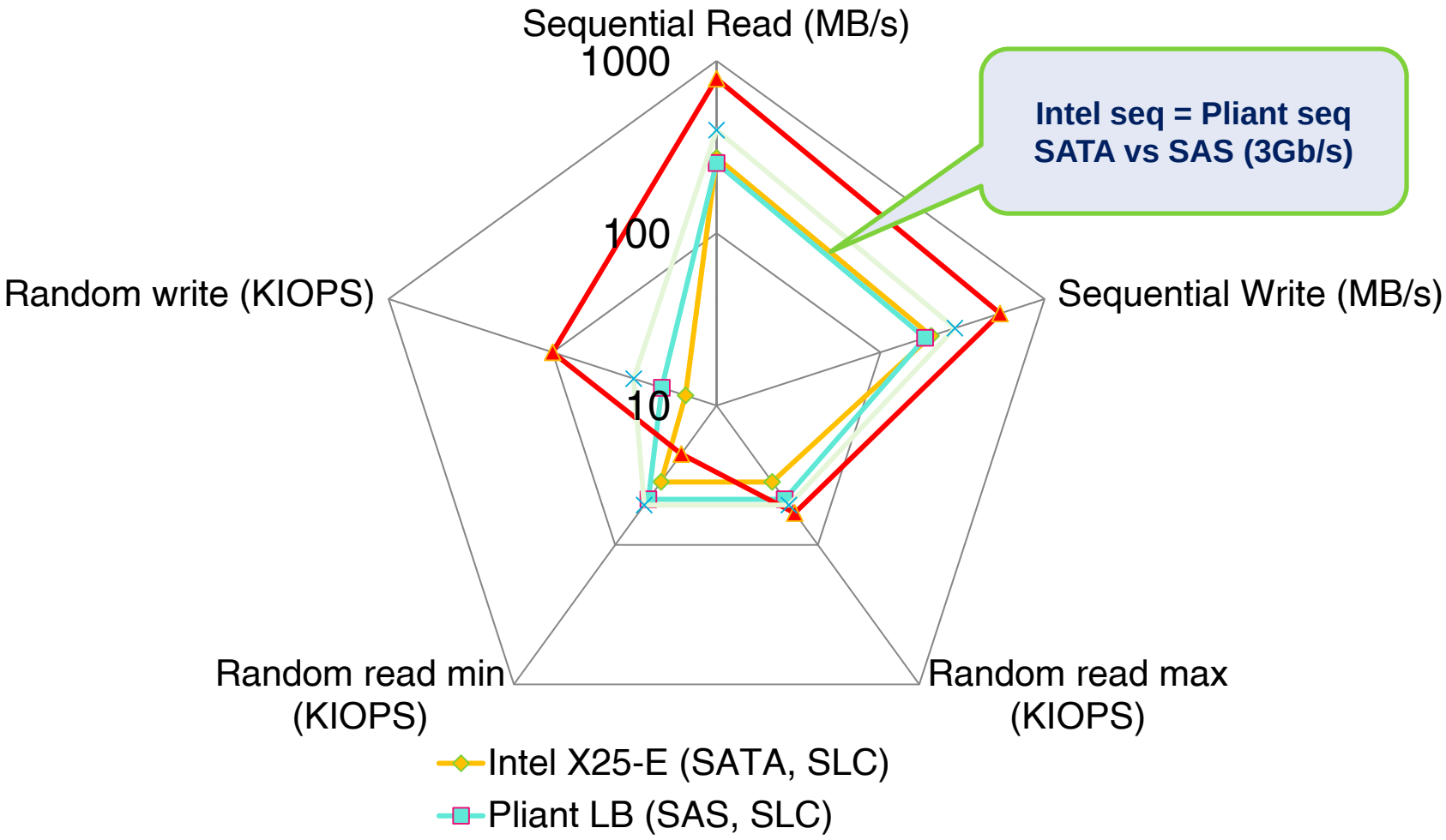
SSD Testing - Representative

(Note: log scale axes, larger is better)



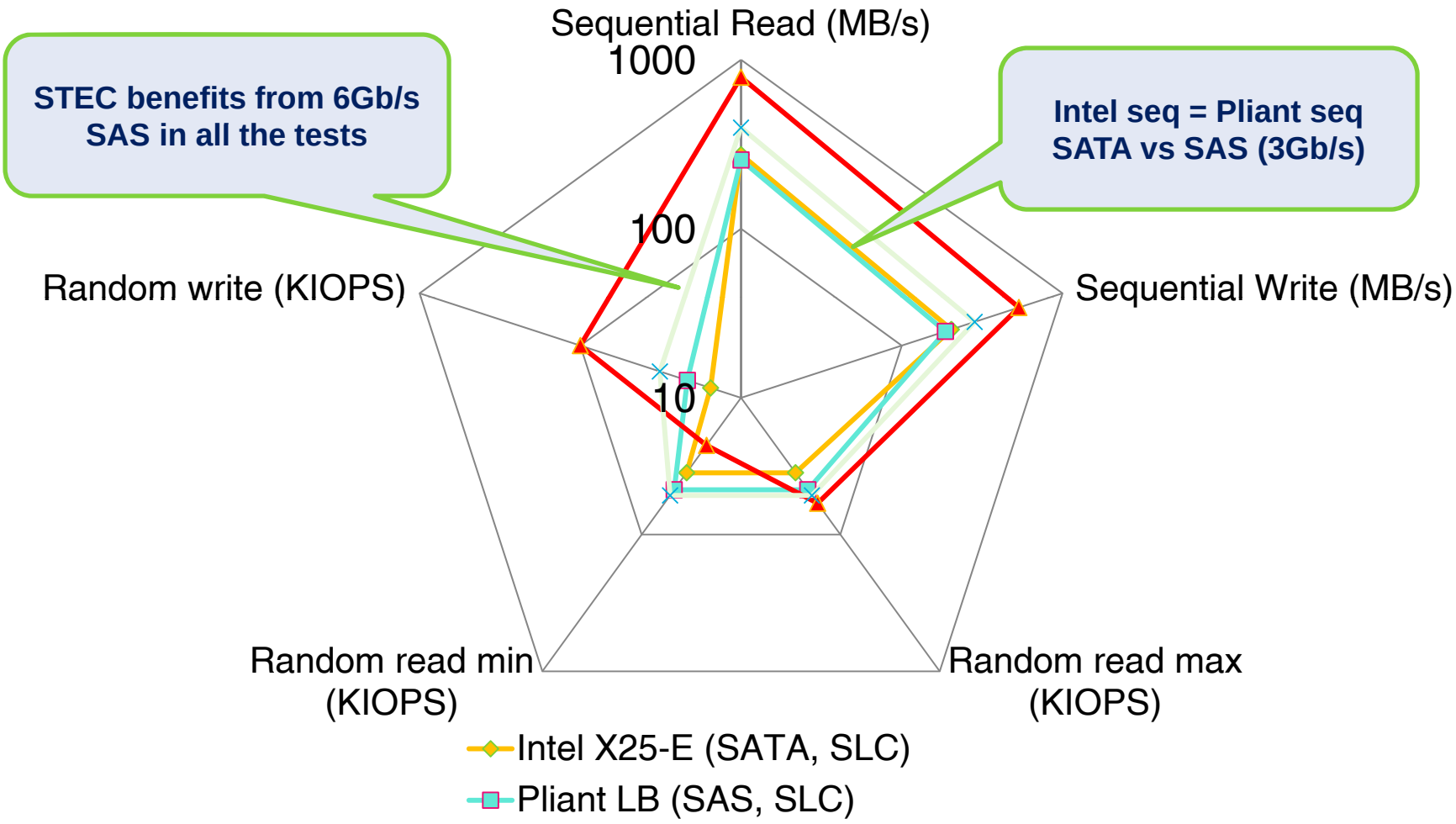
SSD Testing - Representative

(Note: log scale axes, larger is better)



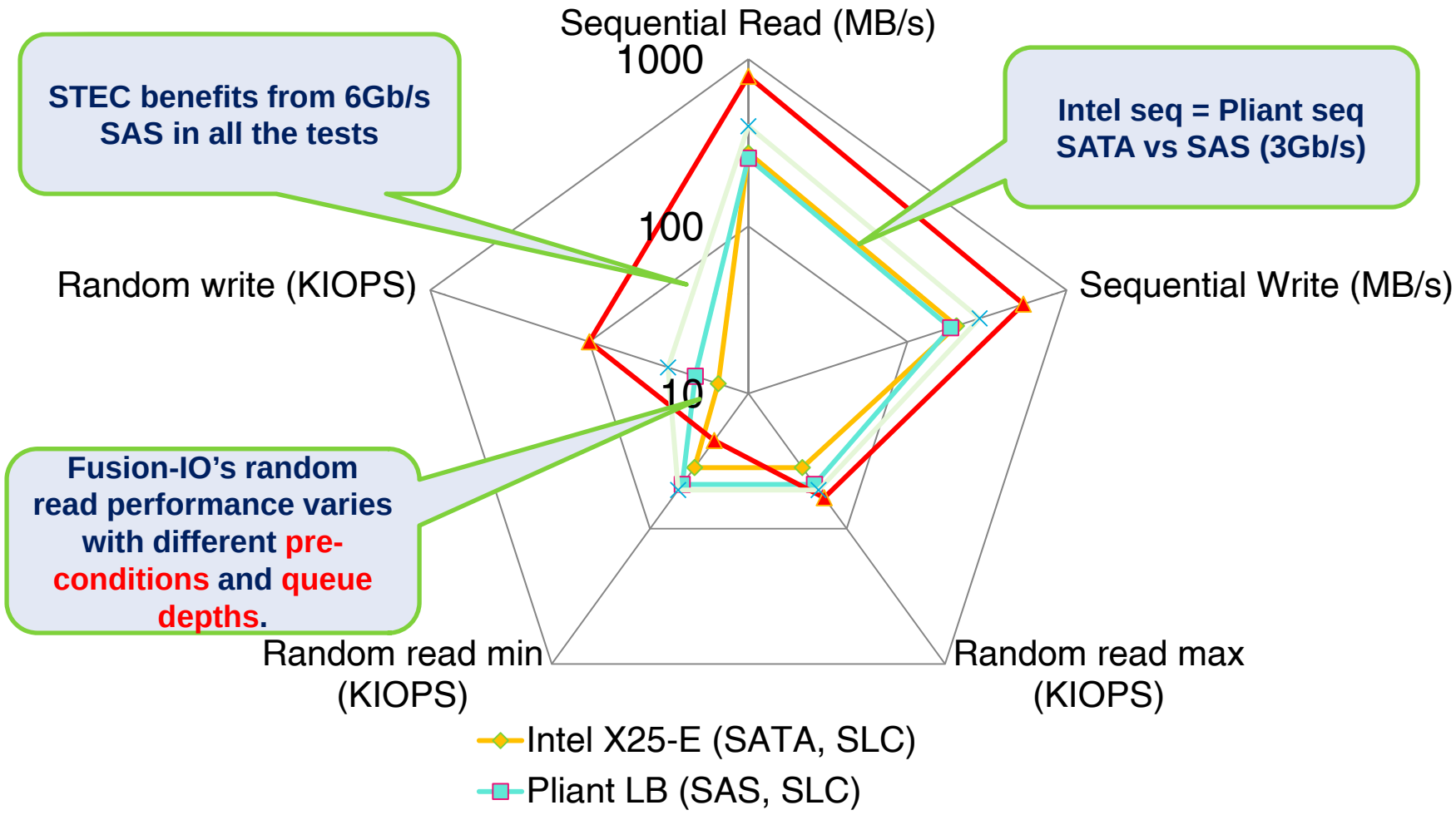
SSD Testing - Representative

(Note: log scale axes, larger is better)



SSD Testing - Representative

(Note: log scale axes, larger is better)



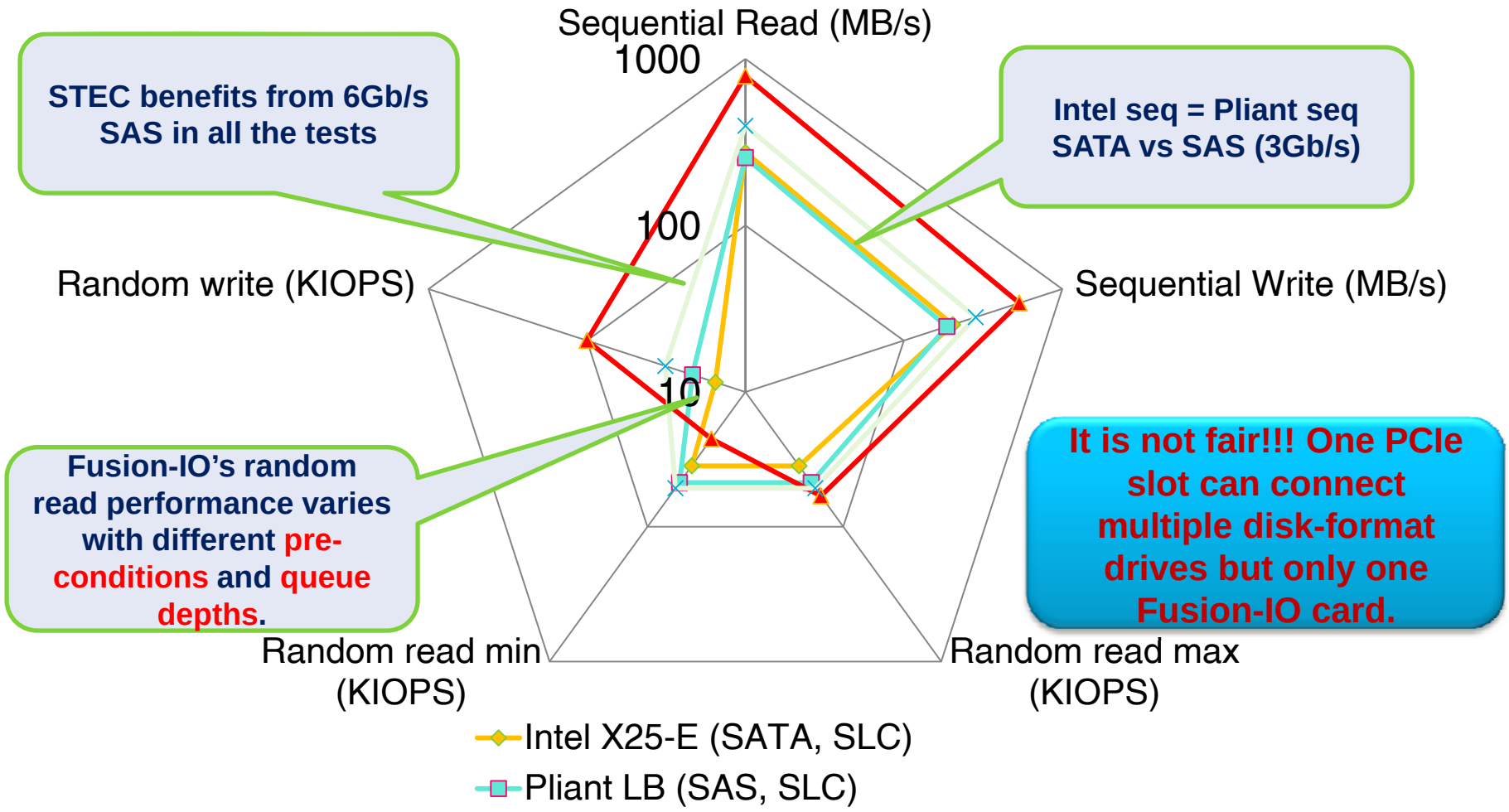
STEC benefits from 6Gb/s SAS in all the tests

Intel seq = Pliant seq SATA vs SAS (3Gb/s)

Fusion-IO's random read performance varies with different pre-conditions and queue depths.

SSD Testing - Representative

(Note: log scale axes, larger is better)



STEC benefits from 6Gb/s SAS in all the tests

Intel seq = Pliant seq SATA vs SAS (3Gb/s)

Fusion-IO's random read performance varies with different pre-conditions and queue depths.

It is not fair!!! One PCIe slot can connect multiple disk-format drives but only one Fusion-IO card.

Dash Became a TeraGrid Resource in April 2010

- Dash TeraGrid configuration
  - One 16-way vSMP supernode with 960 GB of SSD
  - One 16-node cluster with 1 TB of SSD
  - Additional nodes to be added pending 32-way vSMP acceptance
- Allocated users now on the system and supported via TeraGrid Users Services Staff
- Resource leverage with Dash as a TeraGrid Resource
  - Systems and Network Administration
  - User Support, Training, and Documentation
  - Identifying technical hurdles well before Gordon goes live
- Tutorial, Papers, and BOF at TeraGrid 2010, and SC '10

Dash Allocations/Early User Outreach

Project (or) user	Institution(s)	Science Community	Scope of Work	Status and Outcomes
Tim Axelrod	University of Arizona, LSST.org	Astronomical Sciences	Determine if and how the new paradigm of IO-oriented data-intensive supercomputing used by DASH/Flash Gordon can be used by LSST.	LSST has requested a startup account on DASH that has been approved by TG (30000 SUs)
Mark Miller	UCSD	Molecular Biosciences	Improve BFAST code performance using Dash features. Specifically SSDs will be used to accelerate file I/O operations.	Startup request approved on Dash (30000 SUs)
Sameer Shende	University of Oregon	Performance Evaluation and Benchmarking	Performance Evaluation Using the TAU Performance System (R). vSMP node will be used to analyze, visualize performance data.	Startup request approved on Dash (20000 SUs)
John Helly	University of California-San Diego	Atmospheric Sciences	Data Transposition Development for Exa-scale Data in Memory	Startup request approved on Dash (30000 SUs)
John Dennis	NCAR	Atmospheric Sciences		

Dash Early User Success Stories

- NIH Biological Networks Pathway Analysis
  - Queries on graphical data producing a lot of random IO, requiring significant IOPS. DASH vSMP speedup: **186 %**
- Protein Data Bank - Alignment Database
  - Predictive Science with queries on pair-wise comparisons and alignments of protein structures: **69%** DASH speedup
- Supercomputing Conference (SC)
  - 2009 HPC Storage Challenge Winner,
  - 2010 – Two Publications accepted. Finalist for Best paper and Best Student paper.

Seismic Exploration

-
  - A data intensive seismic imaging technique that is memory and I/O bound and therefore a good candidate for Dash and Gordon architectures.
- Goals of the project:
  - Profile the application to understand the combination of processor speed, memory, IO, and storage performance that will result in optimum performance.
  - Develop a general and accurate performance model for flash memory-based architectures and tune and evaluate performance on the Dash system.
- Early Results from this effort
  - Usage of flash results in ~3X performance increase over PFS and ~2x performance increase over node-local spinning disk as measured by decrease in wallclock time.
  - Performance appears to be directly proportional to amount of time code spends in I/O.
  - Using UCSD's Triton cluster as control for Dash benchmarking. Results show similar performance for NFS / PFS disk and node-local disk.

Dash/ Gordon Documentation

- Dash User Guide (SDSC site -->User Support --> Resources --> Dash)
- http://www.sdsc.edu/us/resources/dash/

- TeraGrid Resource Catalog (TeraGrid site --> User Support --> Resources --> Compute & Viz Resources): https://www.teragrid.org/web/user-support/compute_resources

- Gordon is mentioned under Dash's listing in the TG Resource Catalog as a future resource. It will have its own entry as the production date nears

- TeraGrid Knowledge Base, two articles
  (TeraGrid site --> Help & Support --> KB --> Search on "Dash" or "Gordon"):
- https://www.teragrid.org/web/user-support/kb
  - On the TeraGrid, what is Dash?
  - On the TeraGrid, what is Gordon?

Invited Speaker: M.L. Norman: *Accelerating Data-Intensive Science with Gordon and Dash*

Presentation: *DASH-IO: an Empirical Study of Flash-based IO for HPC;*
Jiahua He, Jeffrey Bennett, and Allan Snavely, SDSC

Birds of a Feather (2): *NSF's Track 2D and RVDAS Resources*; Richard Moore, Chair

Tutorial: Using vSMP and Flash Technologies for Data Intensive Applications

Presenters: Mahidhar Tatineni, Jerry Greenberg, and Arun Jagatheesan, San Diego Supercomputer Center (SDSC)
University of California, San Diego (UCSD)

Abstract:

Virtual shared-memory (vSMP) and flash memory technologies have the potential to improve the performance of data-intensive applications. Dash is a new TeraGrid resource at SDSC that showcases both of these technologies. This tutorial will be a basic introduction to using vSMP and flash technologies and how to access Dash via the TeraGrid. Hands-on material will be used to demonstrate the use and performance benefits.

Agenda for this half-day tutorial includes:

Dash Architecture, the Dash user environment ; hands-on examples on use of a vSMP node; hands-on examples illustrating flash memory use; and a Q&A session including hands on preliminary work with attendee codes on vSMP nodes and flash IO nodes.

SDSC SAN DIEGO SUPERCOMPUTER CENTER UCSD APPR HPC Cluster Solutions intel ScaleMP

Gordon/Dash Education, Outreach and Training Activities Supercomputing Conference 2010, New Orleans, LA November 13-19, 2010

"Understanding the Impact of Emerging Non-volatile Memories on High-performance, IO-Intensive Computing " ***Nominated*** *for a* ***best paper*** *as well as* ***best student paper****.*

Presenter: Adrian Caulfield

Authors: Adrian Caulfield, J. Coburn, T. Mollov, A. De, A. Akel, J. He, A. Jagatheesan, R. Gupta, A. Snavely, S. Swanson

"DASH: a Recipe for a Flash-based Data Intensive Supercomputer" focuses on the use of commodity hardware to achieve a significant cost/performance ratio for data-intensive supercomputing.

Presenter: Jiahua He

Authors: Jiahua He, A. Jagatheesan, S. Gupta, J. Bennett, A. Snavely

Grand Challenges in Data-Intensive Sciences

October 26-28, 2010

San Diego Supercomputer Center , UC San Diego

Confirmed conference *topics* and speakers :

- ***Needs and Opportunities in Observational Astronomy*** - **Alex Szalay, JHU**
- ***Transient Sky Surveys*** – **Peter Nugent, LBNL**
- ***Large Data-Intensive Graph Problems*** – **John Gilbert, UCSB**
- ***Algorithms for Massive Data Sets*** – **Michael Mahoney, Stanford U.**
- ***Needs and Opportunities in Seismic Modeling and Earthquake Preparedness*** - **Tom Jordan, USC**
- ***Needs and Opportunities in Fluid Dynamics Modeling and Flow Field Data Analysis*** – **Parviz Moin, Stanford U.**
- ***Needs and Emerging Opportunities in Neuroscience*** – **Mark Ellisman, UCSD**
- ***Data-Driven Science in the Globally Networked World*** – **Larry Smarr, UCSD**

SDSC SAN DIEGO SUPERCOMPUTER CENTER UCSD APPR HPC Cluster Solutions intel ScaleMP