Big Data 2015: Sponsor and Participants Research Event

Center for Large-scale Data Systems Research, CLDS

San Diego Supercomputer Center

UC San Diego

Agenda

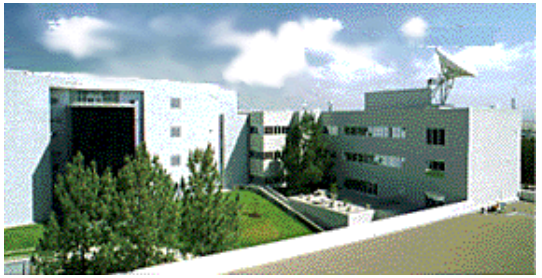
- **Welcome and introductions**
- **SDSC: Who we are**
- **CLDS: Our interests in big data and cloud**
  - Collaborative model with industry

Goals / Objectives of the meeting

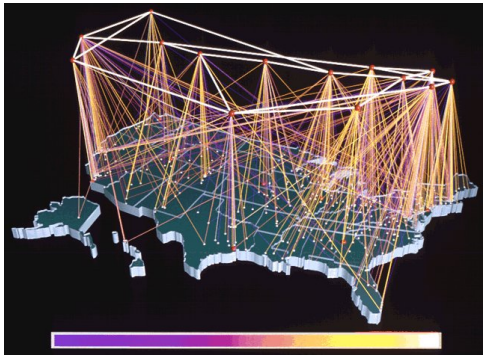
- **Get you interested in CLDS…**
- **…Enough to begin sponsoring the Center**
- **Work together on joint efforts**
  - Immediate projects as well as proposals to NSF et al
- **Some ideas:**
  - Benchmarking: broadly
  - Designing “appliances”
  - Architecture for mixing different types of data access (realtime, online, batch) in the same system
  - Design and build a commodity-based, big-data machine for the NSF community

SDSC: A Brief History

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



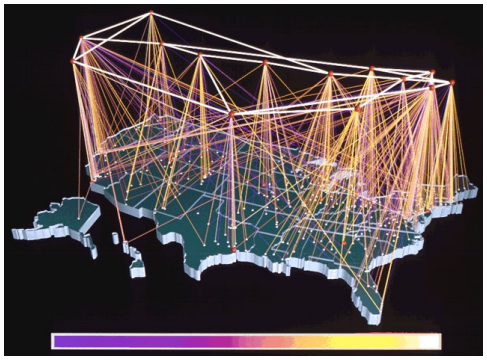
- **1997-2007: NSF PACI program leadership center; managed by UCSD**
  - PACI: Partnerships for Advanced Computational Infrastructure



- **2007-2009: transition to U. Calif. supported research computing resource**
  - still NSF national "resource provider"

- **2009-future: multi-constituency cyberinfrastructure (CI) center**
  - provide data-intensive CI resources, services, and expertise for campus, state, nation, industry

MISSION
Transforming Science and Society Through "Cyberinfrastructure"



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

D. Atkins, NSF Office of Cyberinfrastructure

CI Research & Development

Multi-Disciplinary Research Employing Computational Science…

caida

Next Generation Biology Workbench

GEON The Geosciences Network

PDB PROTEIN DATA BANK

OptIPuter

ORION ocean research interactive observatory networks

LCA Laboratory for Computational Astrophysics University of California, San Diego

SIOExplorer

Groups & Labs

- Advanced Cyberinfrastructure Development Group
- Advanced Query Processing Laboratory
- Applied Network Research
- Bourne Laboratory
- Complex Systems
- Cryoelectron Microscopy and 3D Image Reconstruction
- Cooperative Association for Internet Data Analysis
- Homeland Security
- Laboratory for Computational Astrophysics
- Laboratory for Environmental and Earth Science
- Molecular Interaction and Crystallography
- Next Generation Tools for Biology
- Oceanography and Biodiversity Research Group
- Pacific Rim Activities
- Performance Modeling and Characterization Laboratory
- Scientific Workflow Automation Technologies Laboratory
- SDSC Education Group
- Spatial Information Systems Laboratory

CIPRES

camera Marine Microbial Ecology

BIRN BIOMEDICAL INFORMATICS RESEARCH NETWORK

NEESgrid

SEEK

HPWREN

CHRONOPOLIS

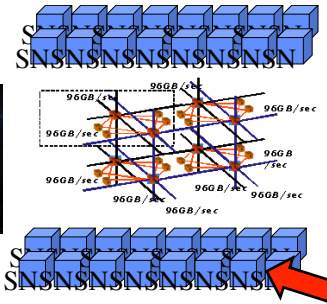
REGIONAL WORKBENCH CONSORTIUM Collaborative Research, Outreach and Education for Sustainable Development

CONVERSATIONS WITH HISTORY

ROADNet

CLDS Center for Large-scale Data Systems Resea

SDSC's "Research Cyberinfrastructure" Portfolio



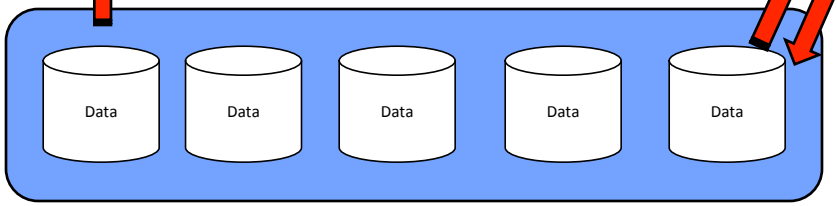
GORDON
- End 2011
- > 200 TF
- Unique open architecture for data-intensive computing
- Can allocate 10% to industry

INTERNET

High-speed network access—for free!

TRESTLES
- In production
- 100 TF, #111 on "Top500"
- Can allocate 10% to industry

SDSC Cloud Storage (Summer 2011)
- 5.5 PB raw
- Standard cloud protocols (OpenStack/S3)
- HIPAA/FISMA compliance (future)



DATA OASIS:
- Under construction
- 4,000 TB disk storage system
- Lustre parallel file system

HIGH THROUGHPUT INSTRUMENTS

TRITON RESOURCE
- 29 TF, in production
- Large memory nodes good for genomics computing
- Can allocate significant % to industry

SDSC Systems Portfolio

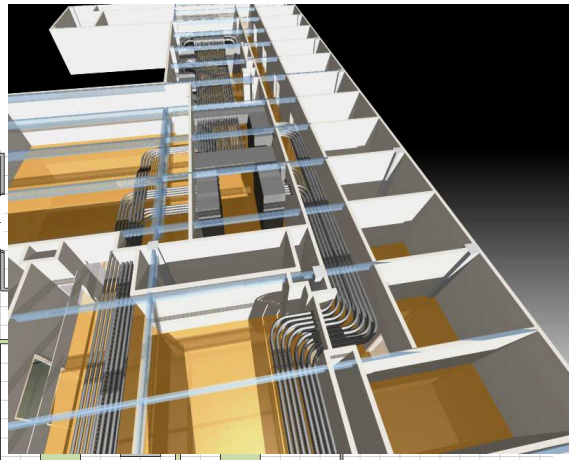
System	# Nodes / Processors / Memory per Node (GB)	Peak Performance ("TeraFLOPS"	Features / Use
Triton Compute Cluster (UC)	256 / 2,048 / 24	20	Medium-scale system for general purpose research computing
Petascale Analysis Facility (PDAF) (UC)	28 / 896 / 256 or 512	9	Unique hybrid SMP cluster for data-intensive research computing
"ShaRCS" (UC)	544 / 2176 / 24	21	Pilot project for UC-wide shared HPC. Currently fully reserved.
"Dash" (NSF/ TeraGrid))	64 / 512 / 48	5	Prototype/experimental system using Solid State Disk (SSD) and software-based virtual memory for data-intensive analyses.
"Trestles" (NSF/ TeraGrid)	324 / 10,368 / 64	100	New system for national researchers; deployed 12/2010. Employs SSD. #111 on "Top 500" list.
"Gordon" (NSF/ TeraGrid)	1024 compute servers with TBD processors / 64TB RAM and 256 TB flash per "supernode"	TBA	Large scale national system to be deployed in 2011. Experimental open-architecture system using SSD and software VM for data-intensive computing. Should rank around #30 on Top 500 list.

Datacenter Highlights – a Working Lab for Energy Efficiency

13,500 sq. ft. Legacy
4,500 sq. ft. New
= 18,000 sq. ft. Total Datacenter Space

13 MW Total Datacenter Power

Major Power Expansion



Computer Room Air Handling (CRAH) VFD Retrofits, Sensors to Throttle with Load

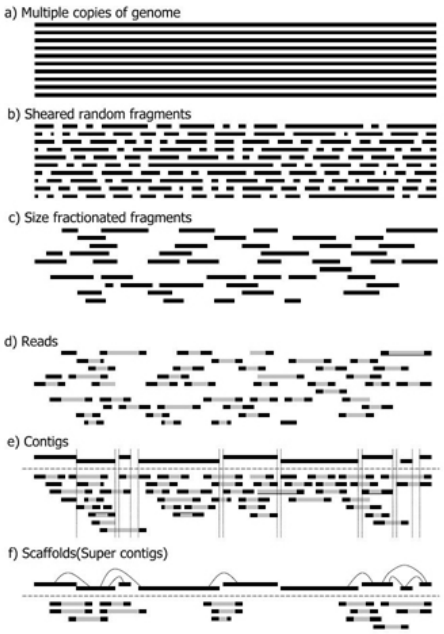
Thermal Containment…

Cold Aisle Containment (New Datacenter)

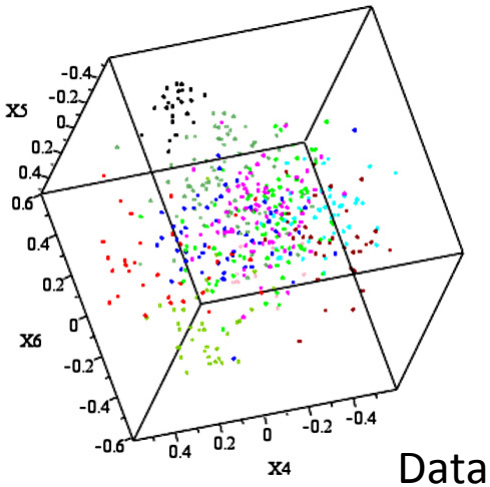
- **Expect 50-100% better cooling efficiency than traditional hot/cold aisle**
- **Knuerr Coolflex**
  - First of its kind deployment in the United States
- **Separates cold and hot air in the datacenter, eliminating the need for over-cooling, over-blowing**
  - Works best in a standardized datacenter
  - Server intakes from enclosed cold aisles (70-78F)
  - Server exhaust to room as a whole, which runs hot (85-100F)
  - Allows for high temperature differentials, maximizing efficiency of cooling equipment

SDSC's Specialty: Data-Intensive Computing

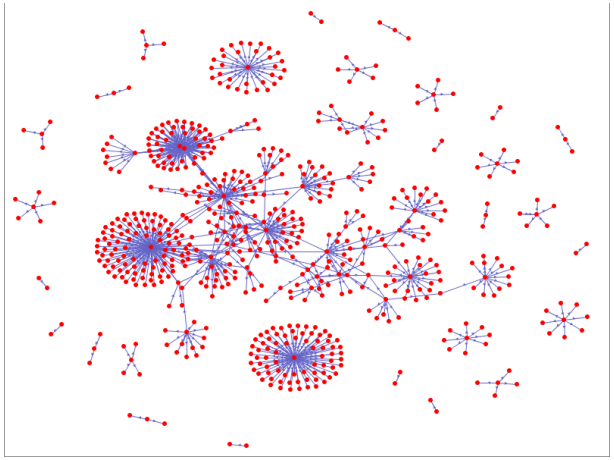
- The nature of high performance computing is changing
- The new problems are "data-intensive"
- SDSC is developing new computing architectures to address these problems



Genome Assembly



Data Mining



Social Network Analysis

Some of SDSC's interactions with industry in data

- **@backup – spinoff from General Atomics**
  - Backup from home computers to a remote archive facility (HPSS), via Web
- **Part of development team for DoE / IBM HPSS**
  - HPSS: tape-based archival storage
- **IBM Shared University Research (SUR) project**
  - Deploy IBM DB2 tablespace containers on HPSS
- **Nirvana Storage**
  - Spinoff from GA of the SDSC Storage Resource Broker software (www.nirvanastorage.com)

Context for CLDS

- **SDSC is a data-intensive supercomputer center**
- **SDSC has projects in CS and Computational Science research, development, and production**
- **Cloud computing is of interest even in the NSF world**
  - Example: projects like EarthCube. OOI is using Amazon AWS.
- **Cloud / big data have technical as well as management issues**

Ideas for CLDS projects, activities

- **Reference and probe benchmarking**
- **Appliance design**
- **Technology management and business implications**
- **Manageability**
- **How Much Information—Big Data?**
- **Executive education in big data and clouds**

Benchmarking

- **Users need guidance: on performance, price/performance**
  - Still lots of doubt out there
- **Big data is an end-to-end problem**
  - Need to be able to benchmark the end-to-end flow
- **We have experience with TPC**
- **"Probe" benchmarks: The Cloud Weather Service™**
  - Cloud performance can be variable
  - "Application-level" performance probes, a la Network Weather Service

- **In a fully digital world, big data is here to stay**
  - In NSF jargon "cyberinfrastructure" was a code word for modernizing as well as democratizing access to resources (computing, storage, software)
  - ➔ Big data: code word for making large-scale data processing routine task
- **NSF has expressed some interest in the benchmarking activity**

Appliance Design

- **Scalable storage infrastructure for clouds**
- **Environment for testing at scale (e.g. IBM DB2)**
- **Role of “fat” vs “skinny” nodes in a cluster**
- **Networking issues for big data (from CPU to data)**
- **Provide end-users decision-making framework for candidate workloads**
- **Serving different data processing requirements (e.g. realtime vs batch) within the same system**

Executive Education Program

- Series on Big Data and Clouds offered via the Rady School of Management, UCSD
- CLDS Sponsors receive 3 free seats for their employees and/or customers
- First course, Nov 15th

Rady | UC San Diego School of Management — Center for Executive Development

Competitive Advantage Through Cloud Computing

Understanding and exploiting one of the most significant computing trends of the 21st century

November 15, 2011
8:00 a.m. - 12:00 p.m.

www.rady.ucsd.edu/cloud-computing

"The desktop is dead. Welcome to the Internet cloud, where massive facilities across the globe will store all the data you'll ever use." – Wired Magazine

"Gartner Executive Programs Worldwide Survey of More Than 2,000 CIOs Identifies Cloud Computing as Top Technology Priority for CIOs in 2011" – Gartner Research

More and more CIO surveys report that top IT executives view cloud computing as a key resource necessary to maintain an effective and secure global IT infrastructure for 21st-century business. To address this need, UC San Diego's Rady School of Management and the Center for Large-scale Data Systems (CLDS) at the San Diego Supercomputer Center are jointly developing a new series of executive programs in cloud computing. The inaugural half day course, Competitive Advantage Through Cloud Computing, explains the key technologies and business imperatives behind cloud computing. The program will be taught by experts in computer science, IT systems and business strategy and will present technology concepts, the business imperatives, review current practice, and present the future opportunities for senior IT and business management, innovators and information entrepreneurs.

The rapid growth of the digital economy has resulted in major Internet and technology companies building massive datacenters distributed around the globe. The same companies are now providing advanced tools for individuals, businesses and entrepreneurs for developing applications and services that run in the "cloud." Are enterprise servers and dedicated data centers going the way of the dinosaur? Will cloud computing be the ultimate solution for outsourcing IT functions? Competitive Advantage Through Cloud Computing will answer these questions and more, and give you technical insights and the tools needed to make decisions about this critical trend in IT.

PROGRAM BENEFITS:

- Understand the technology, not the hype
- Explore cutting-edge cloud strategies to increase business competitiveness
- Understand the costs and benefits of cloud
- Learn how to assess which applications will migrate into the cloud and which ones will not
- Learn best practices and migration strategies for IT and business

WHO SHOULD ATTEND

- C-level and senior executives of any sized business looking for new ways to gain competitive business advantage
- CIO and Senior business executives with responsibility for IT strategy
- Executives from infrastructure vendors
- Entrepreneurs and innovators

REGISTER ONLINE

rady.ucsd.edu/cloud-computing
Space is limited.

PROGRAM FEE

$295 Participant Registration
(Includes tuition, course materials, campus parking and breakfast.)

PROGRAM FACULTY

Chaitan Baru, PhD: Director, Center for Large-Scale Data, San Diego SuperComputer Center

James Short, PhD: Lead Scientist, Center for Large-Scale Data, San Diego SuperComputer Center

PROGRAM CONSULTANT

Josh Pingel
858.822.0575
jpingel@ucsd.edu

www.rady.ucsd.edu/exec

Sponsorship Model - Tiers

	Sponsor Profile
CLDS Participating Sponsor Variable $15K – $35K annual	Partner with 2-4 other participating firms to define a joint research project and tailor project activities and outputs. For example: cloud performance benchmarking, cost-benefit analysis of transitioning to cloud. Commit time and data to the project. Receive quarterly updates and project compilations. Offer advice and feedback. Participate in Center events.
CLDS Program Sponsor 50K annual	Help define and participate in multiple program areas. For example: performance and capacity benchmarking, appliance design, measuring the benefits and costs of cloud computing, measuring information value and workload performance. Work directly with Center directors and the research team to direct major program areas. Receive project reports, quarterly research briefings, project data, attend events. Receive 1 free slot on each executive education module.
CLDS Affiliate 100K (multi-year)	Center Affiliate membership. Nominate 3 company officers to be CLDS SDSC Affiliates. Receive 3 free slots on each executive education module. Work directly with the Center Director and research team to guide the Center. Implement custom individual projects; have access to all CLDS resources including online social media.

CLDS Calendar

- **Launch benchmarking activity -- Now!**
  - With TPC and other companies
- **Launch Cloud Weather Service™ (CWS) – Now!**
  - Define the service; collect performance probe data; perform analytics
- **Exec Education course – Nov 15**
- **Join in proposal activities – Now!**
  - For Benchmarking
  - For building a national-scale Hadoop-style system for science and Computer Science research
    - Can be a vendor sandbox

- **Join CLDS to participate in all of the above**