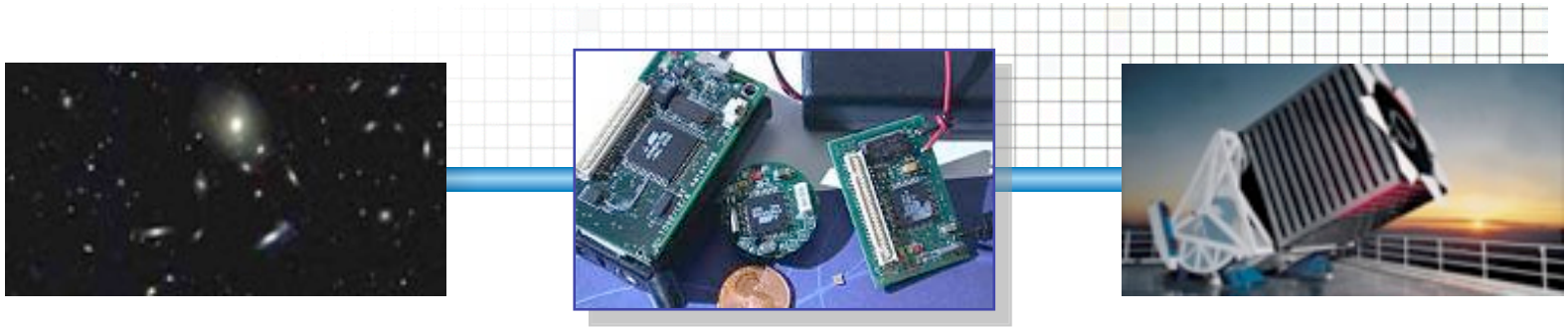


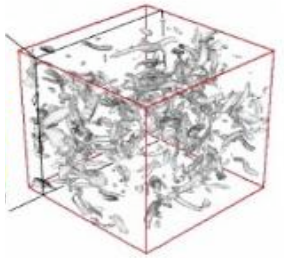
The Fourth Paradigm of Science

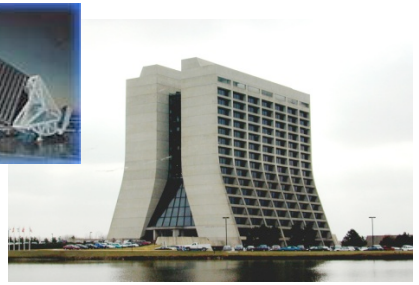
Alex Szalay
The Johns Hopkins University

Evolving Science

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  synthesizing theory, experiment and computation with advanced data management and statistics
  → new algorithms!

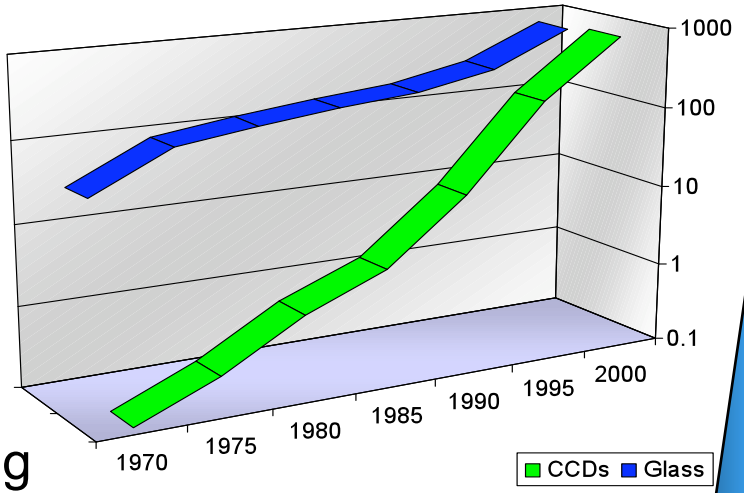
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$





Living in an Exponential World

- Scientific data doubles every year
  - *caused by successive generations of inexpensive sensors + exponentially faster computing*



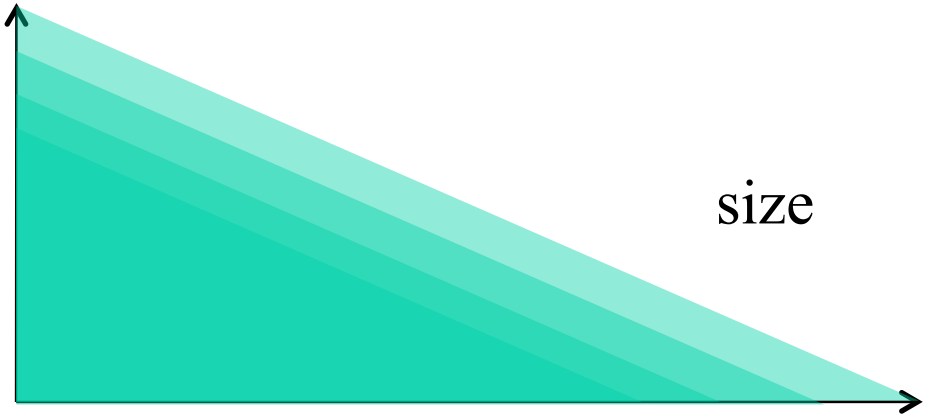
- Changes the nature of scientific computing
- Cuts across disciplines (eScience)
- It becomes increasingly harder to extract knowledge
- 20% of the world's servers go into huge data centers by the "Big 5"
  - *Google, Microsoft, Yahoo, Amazon, eBay*

Collecting Data

- Very extended distribution of data sets:

  data on all scales!

- Most datasets are small, and manually maintained (Excel spreadsheets)
- Total amount of data dominated by the other end (large multi-TB archive facilities)
- Most bytes today are collected via electronic sensors



Scientific Data Analysis

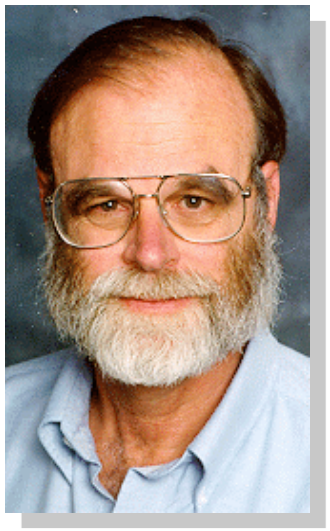
- Data is everywhere, never will be at a single location
- Architectures increasingly CPU-heavy, IO-poor
- Data-intensive scalable architectures needed
- Need randomized, incremental algorithms
  - *Best result in 1 min, 1 hour, 1 day, 1 week*
- Most scientific data analysis done on small to midsize BeoWulf clusters, from faculty startup
- Universities hitting the “power wall”
- **Not scalable, not maintainable…**

- Clouds on the horizon… but they look better from far away then close-up

Gray's Laws of Data Engineering

Jim Gray:

- Scientific computing is revolving around **data**
- Need **scale-out** solution for analysis
- Take the **analysis to the data**!
- Start with "**20 queries**"
- Go from "**working to working**"



DISC: Data Intensive Scalable Computing

Building Scientific Databases

- 10 years ago we set out to explore how to cope with the data explosion (with Jim Gray)
- Started in astronomy, with the Sloan Digital Sky Survey
- Expanded into other areas, while exploring what can be transferred
- During this time data sets grew from 100GB to 100TB
- Interactions with every step of the scientific process
  - *Data collection, data cleaning, data archiving, data organization, data publishing, mirroring, data distribution, data curation…*

Reference Applicatons

Some key projects at JHU

- **SDSS:** *10TB total, 3TB in DB, soon 10TB, in use for 6 years*
- **NVO :** *~5TB, many B rows, in use for 4 years*
- **PanStarrs:** *80TB by 2009, 300+ TB by 2012*
- **Immersive Turbulence***: 30TB now, 300TB next year, can change how we use HPC simulations worldwide*
- ***Sensor Networks:*** *200M measurements now, billions next year, forming complex relationships*

Key Questions:

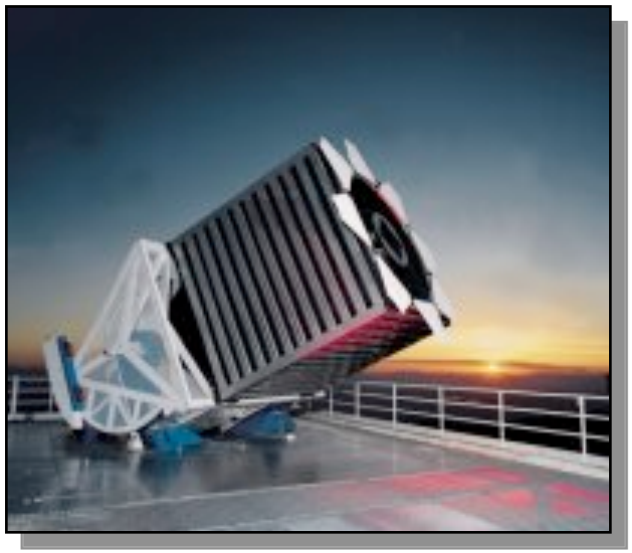
- How do we build a scalable architecture?
- How do we interact with petabytes of data?

Sloan Digital Sky Survey

- "**The Cosmic Genome Project**"
- Two surveys in one
  - Photometric survey in 5 bands
  - Spectroscopic redshift survey
- Data is public
  - 40 TB of raw data
  - 5 TB processed catalogs
  - 2.5 Terapixels of images
- Started in 1992, finishing in 2008
- Database and spectrograph built at JHU (SkyServer)

The University of Chicago
Princeton University
The Johns Hopkins University
The University of Washington
New Mexico State University
Fermi National Accelerator Laboratory
US Naval Observatory
The Japanese Participation Group
The Institute for Advanced Study
Max Planck Inst, Heidelberg

Sloan Foundation, NSF, DOE, NASA

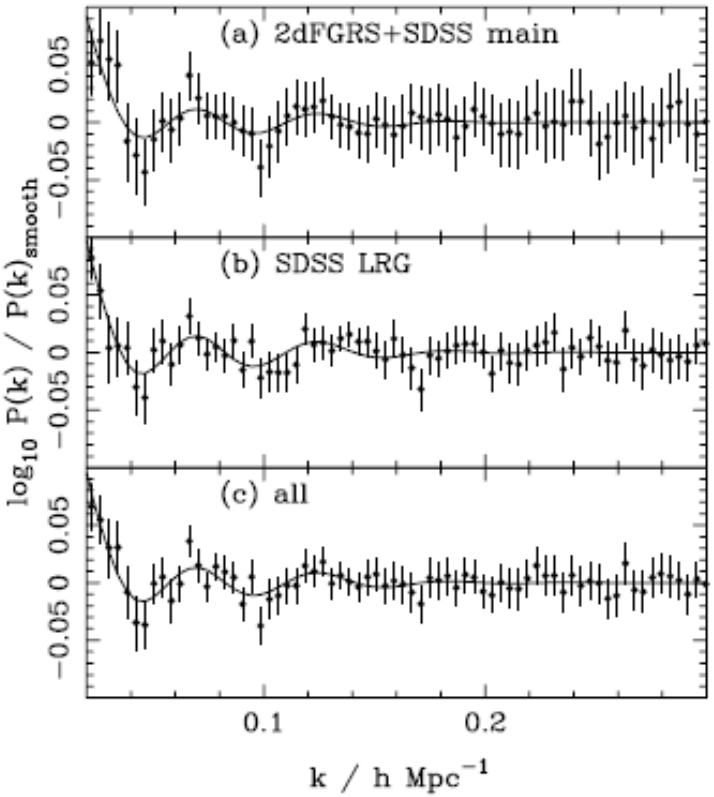


Primordial Sound Waves in SDSS

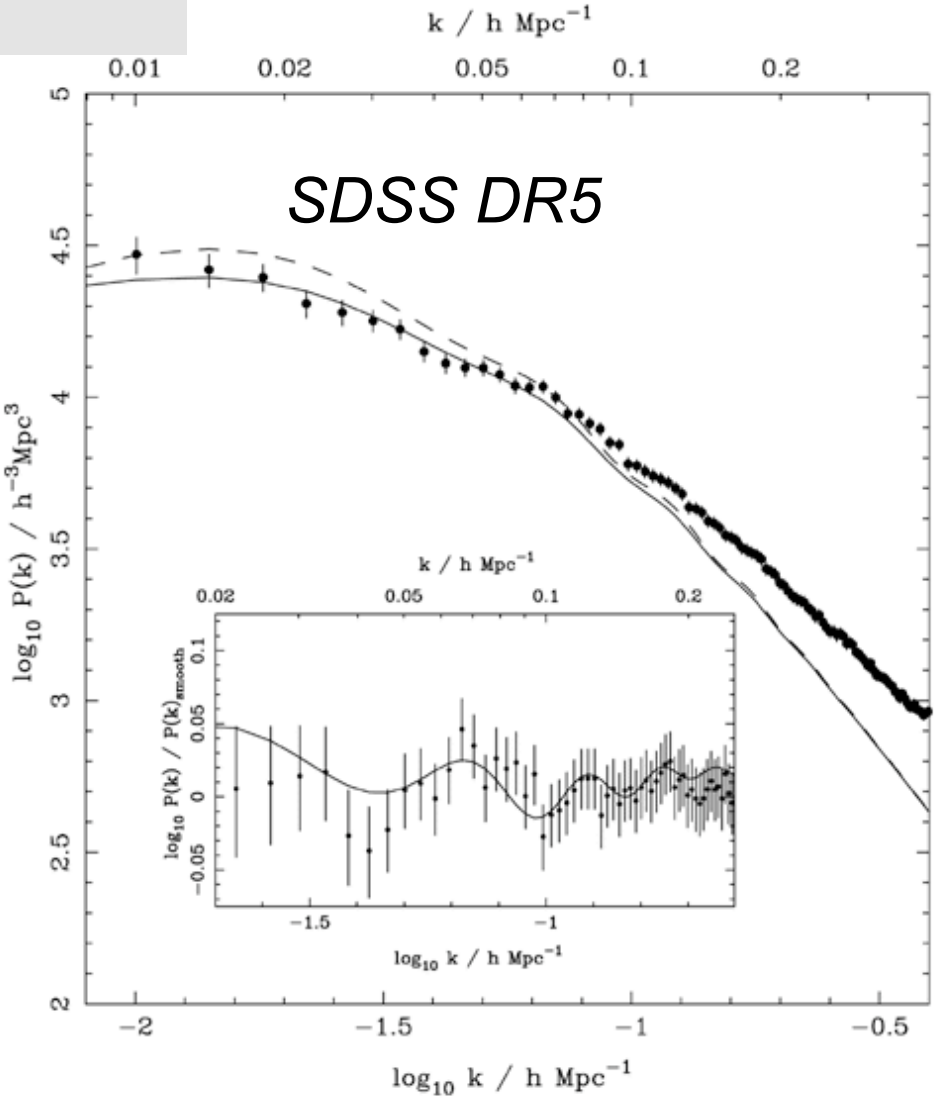
800K galaxies

Power Spectrum

(Percival et al 2006, 2007)



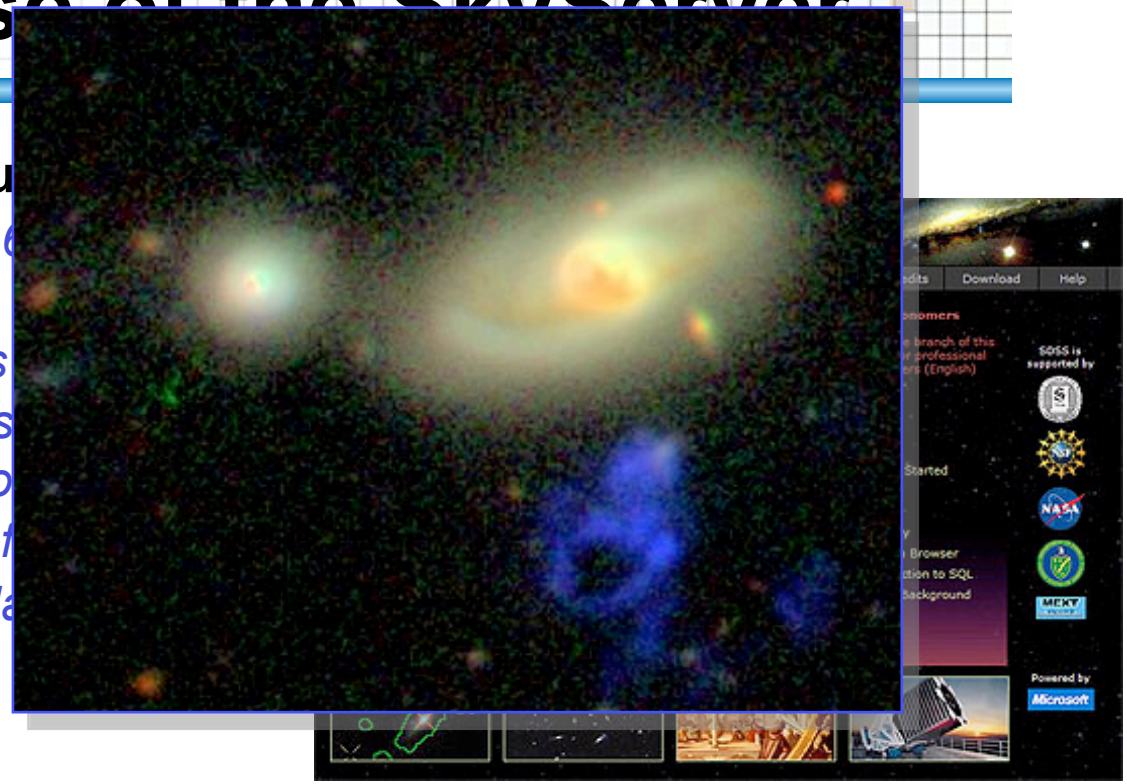
SDSS DR6+2dF



SDSS DR5

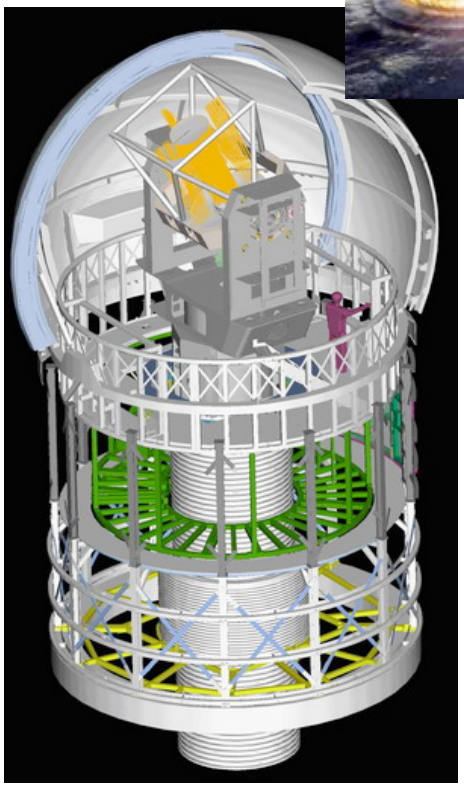
Public Use of the SkyServer

- **Prototype in 21st Centu**
  - *400 million web hits in 6*
  - *930,000 distinct users vs 10,000 astronomers*
  - *Delivered 50,000 hours of lectures to high scho*
  - *Delivered 100B rows of*
  - *Everything is a power la*



- **GalaxyZoo**
  - *40 million visual galaxy classifications by the public*
  - *Enormous publicity (CNN, Times, Washington Post, BBC)*
  - *100,000 people participating, blogs, poems, ….*
  - *Now truly amazing original discovery by a schoolteacher*

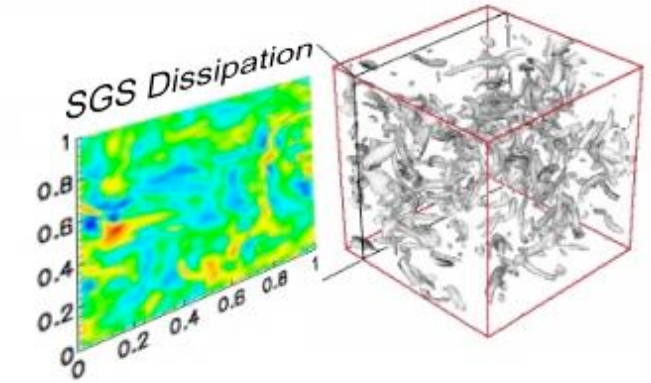
Pan-STARRS

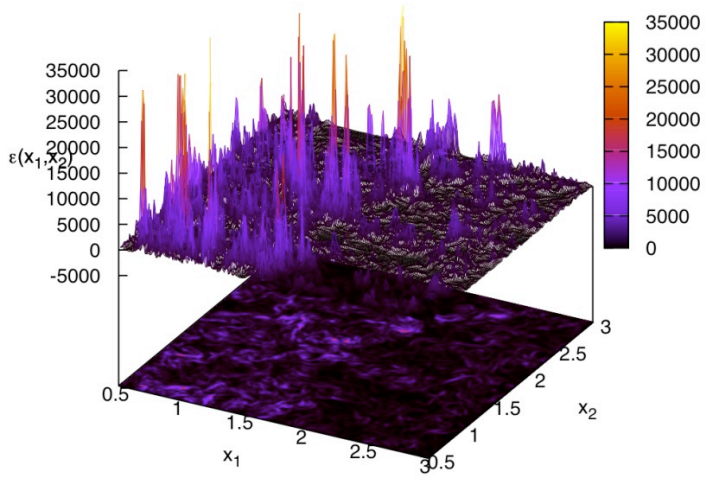


- **Detect 'killer asteroids'**
  - *PS1: starting in November 2008*
  - *Hawaii + JHU + Harvard/CfA + Edinburgh/Durham/Belfast + Max Planck Society*
- **Data Volume**
  - *>1 Petabytes/year raw data*
  - *Camera with 1.4Gigapixels*
  - *Over 5B celestial objects plus 250B detections in database*
  - *80TB SQLServer database built at JHU, **the largest astronomy DB in the world***
  - *3 copies for redundancy*
- **PS4:** 4 identical telescopes in 2012, 4PB/yr

Immersive Turbulence

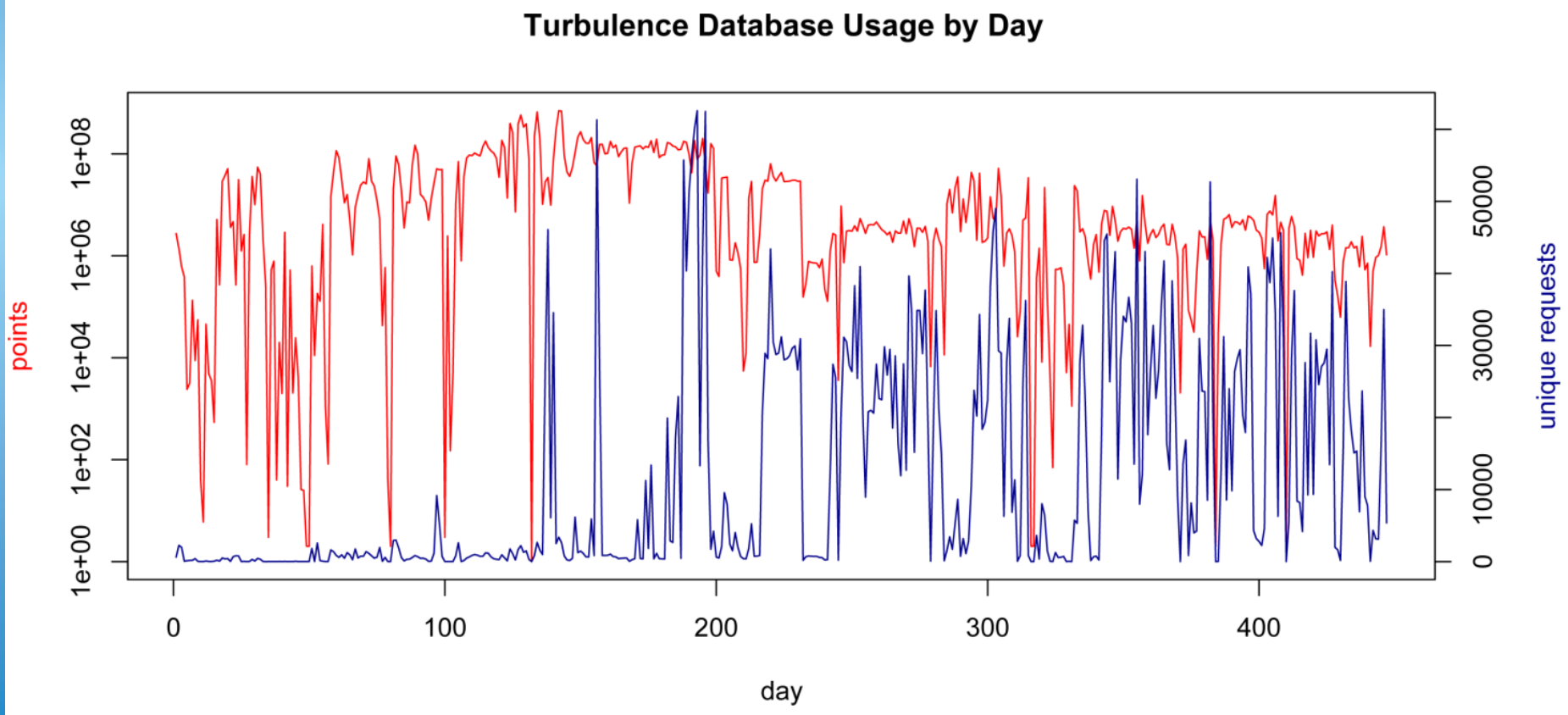
- **Understand the nature of turbulence**
  - *Consecutive snapshots of a $1{,}024^3$ simulation of turbulence: now 30 Terabytes*
  - *Treat it as an experiment, observe the database!*
  - *Throw test particles (sensors) in from your laptop, immerse into the simulation, like in the movie Twister*
- **New paradigm** for analyzing HPC simulations!





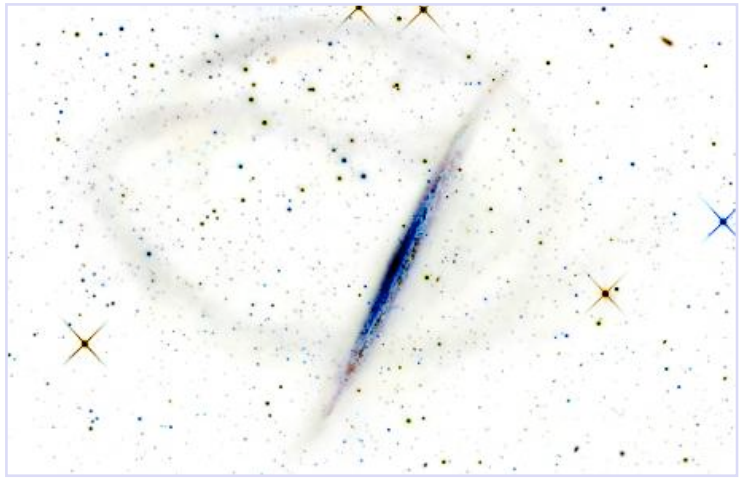
with C. Meneveau, S. Chen (Mech. E), G. Eyink (Applied Math), R. Burns (CS)

Daily Usage


Turbulence Database Usage by Day
points
1e+00
1e+02
1e+04
1e+06
1e+08
0
100
200
300
400
day
0
10000
30000
50000
unique requests

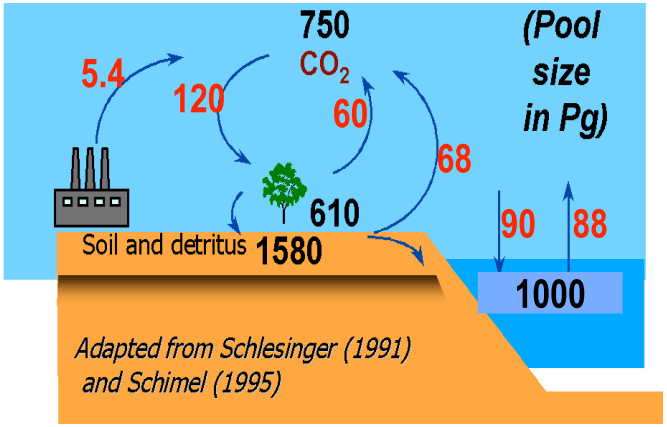
The Milky Way Laboratory

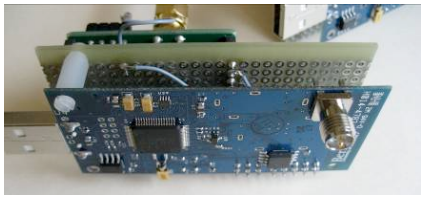
- Pending NSF Proposal to use cosmology simulations as an immersive laboratory for general users
- Use Via Lactea-II (20TB) as prototype, then Silver River (500TB+) as production (15M CPU hours)
- Output 10K+ hi-rez snapshots (200x of previous)
- Users insert test particles into system and follow trajectories in precomputed simulation
- Realistic “streams” from tidal disruption
- Users interact remotely with 0.5PB in ‘real time’



Life Under Your Feet

- **Role of the soil in Global Change**
  - *Soil CO_2 emission thought to be >**15 times** of anthropogenic*
  - *Using sensors we can measure it directly, in situ, over a large area*
- **Wireless sensor network**
  - *Use 100+ wireless computers (motes), with 10 sensors each, monitoring*
    - Air +soil temperature, soil moisture, …
    - Few sensors measure CO_2 concentration
  - *Long-term continuous data, 180K sensor days, 30M samples*
  - *Complex database of sensor data, built from the SkyServer*
  - *End-to-end data system, with inventory and calibration databases*



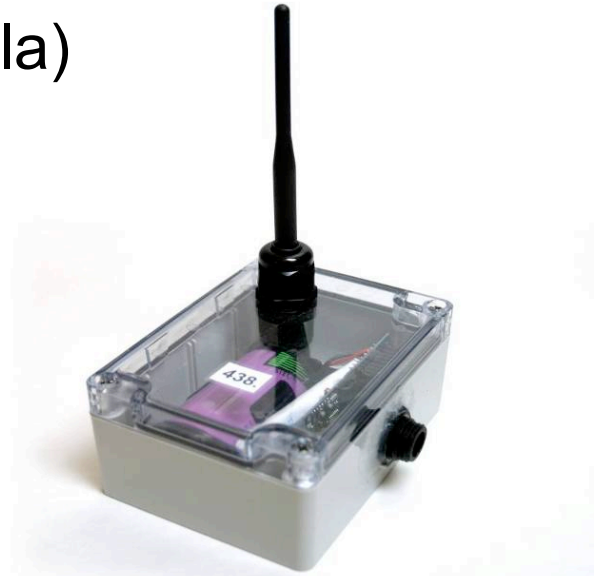


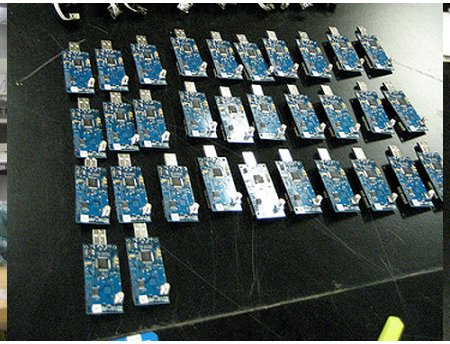
with K.Szlavecz (Earth and Planetary), A. Terzis (CS)

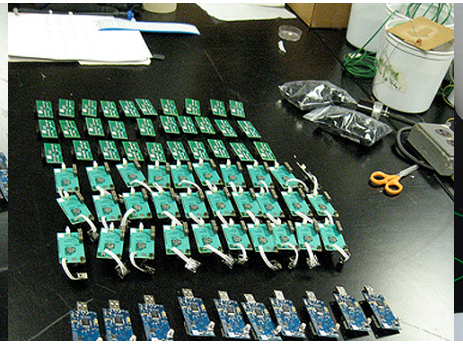
http://lifeunderyourfeet.org/

Current Status

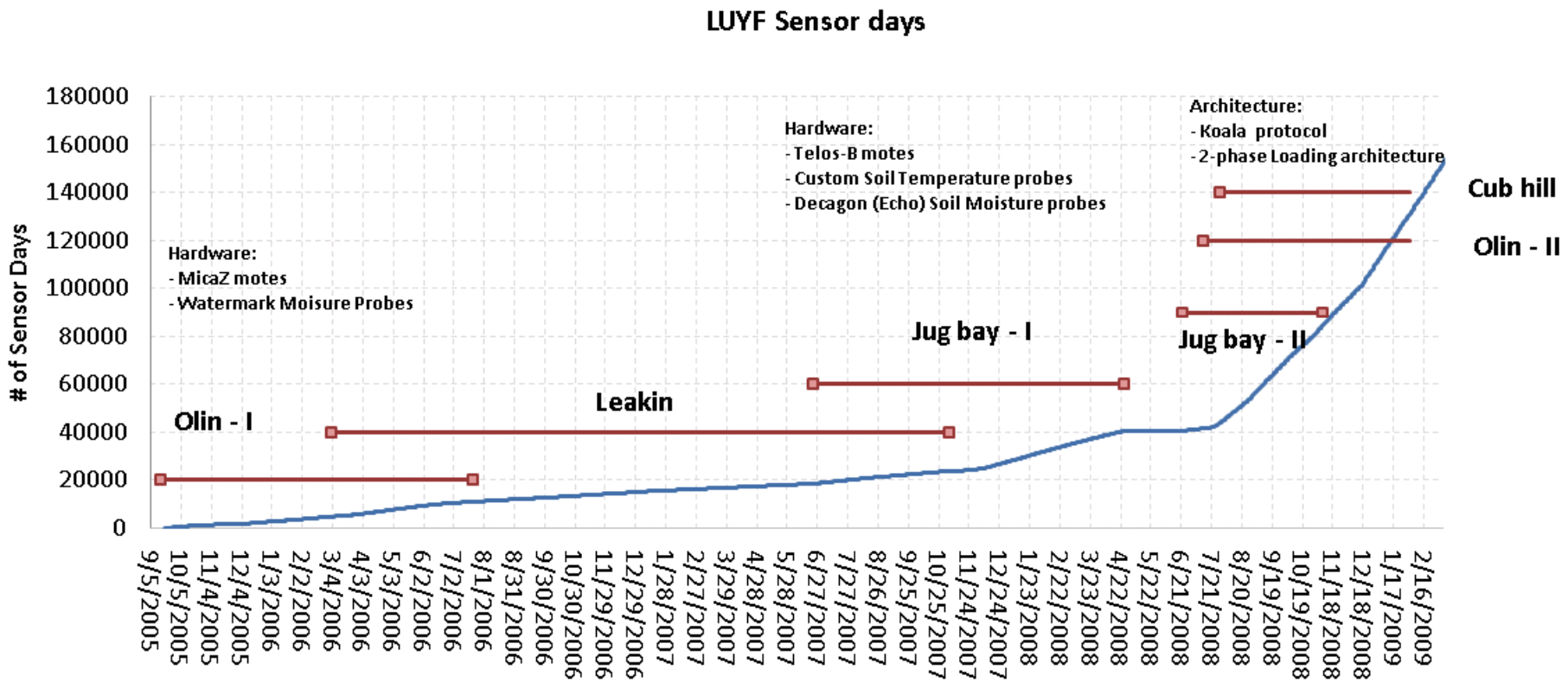
- Designed and built 2nd generation mote platform
  - *Telos SkyMote + own DAQ board*
- Hierarchical network architecture (Koala)
- Improved mote software
  - *Support for large-scale deployments*
  - *Over-the-air reprogramming*
  - *Daily log file written*
  - *Increased power efficiency (2 years on a single battery)*







Cumulative Sensor Days



Commonalities

- Huge amounts of data, aggregates needed
  - *But also need to keep raw data*
  - *Need for parallelism*
- Use patterns enormously benefit from indexing/DB
  - *Rapidly extract small subsets of large data sets*
  - *Geospatial everywhere*
  - *Compute aggregates*
  - *Fast sequential read performance is criticial!!!*
  - *But, in the end everything goes…. search for the unknown!!*
- Data will never be in one place
  - *Newest (and biggest) data are live, changing daily*
- Fits DB quite well, but no need for transactions
- Design pattern: class libraries wrapped in SQL UDF
  - *Take analysis to the data!!*

Continuing Growth

How long does the data growth continue?

- High end always linear
- Exponential comes from technology + economics
  - ⇔ rapidly changing generations
  - *like CCD's replacing plates, and become ever cheaper*
- How many new generations of instruments do we have left?
- Are there new growth areas emerging?
- **Software is becoming a new kind instrument**
  - *Value added federated data sets*
  - *Simulations*
  - *Hierarchical data replication*
- How do we build a scalable architecture?

Amdahl's Laws

Gene Amdahl (1965): Laws for a balanced system

i. Parallelism: max speedup is S/(S+P)

ii. One bit of IO/sec per instruction/sec (BW)

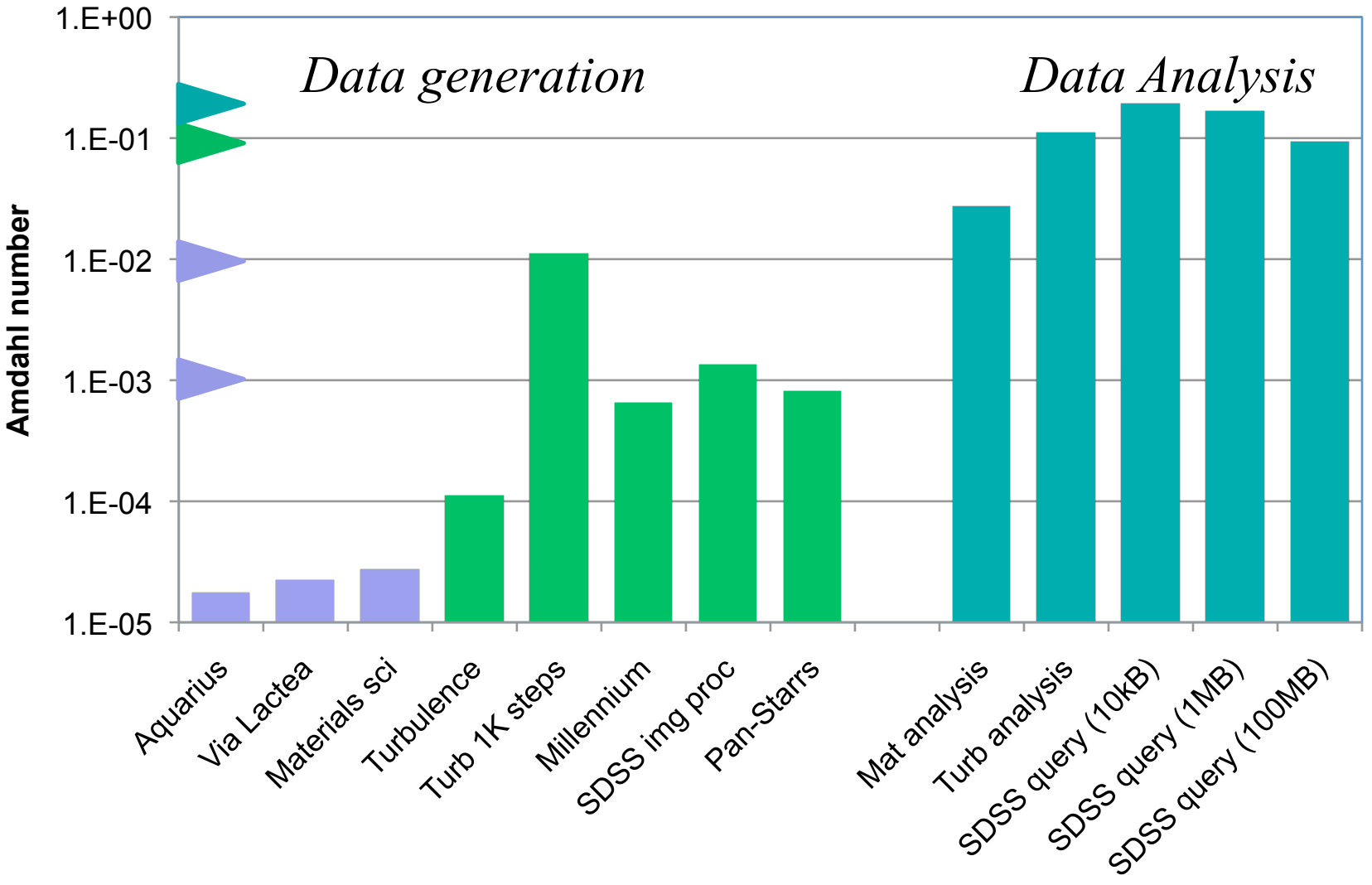
iii. One byte of memory per one instruction/sec (MEM)

Modern multi-core systems move farther away from Amdahl's Laws (Bell, Gray and Szalay 2006)

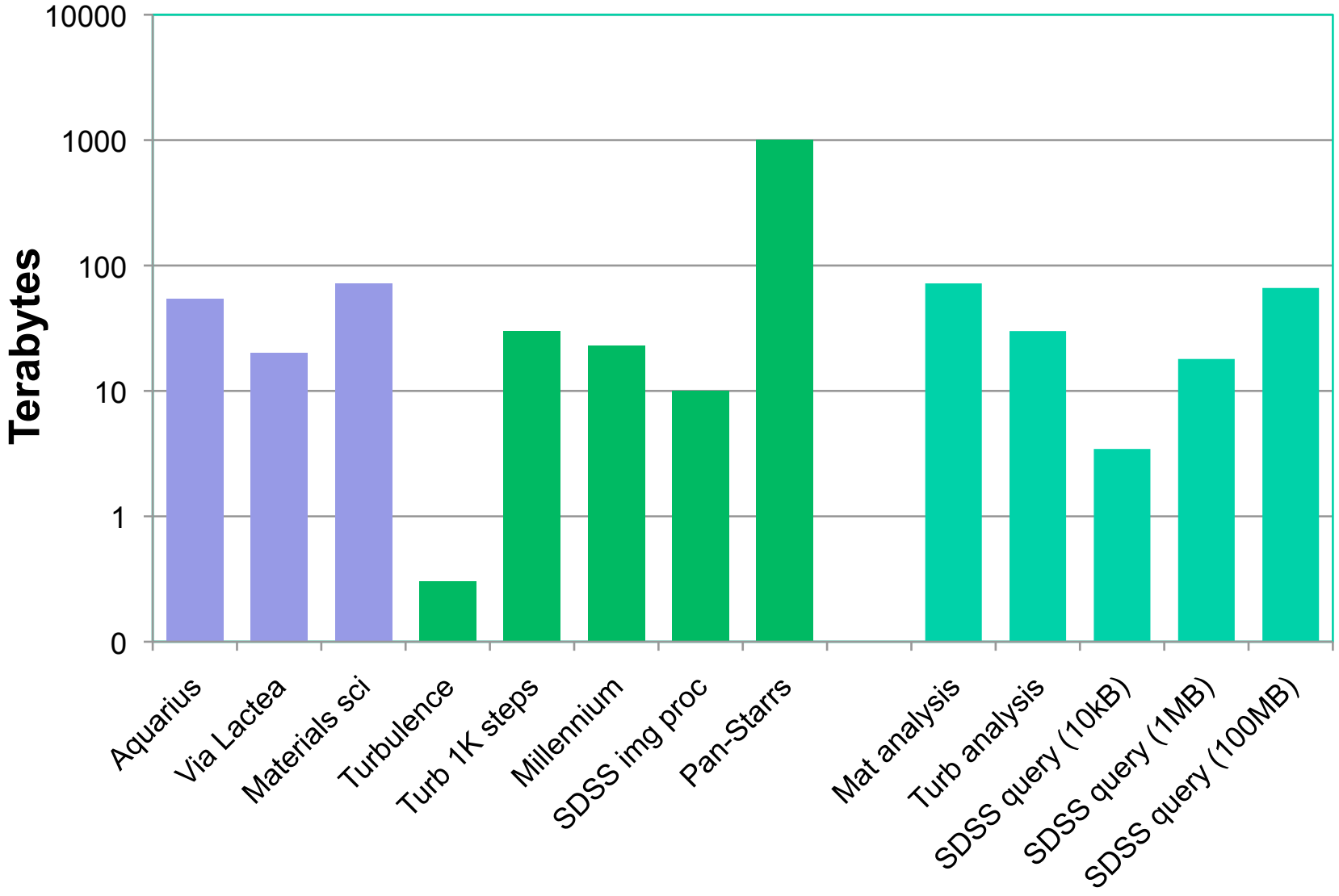
Typical Amdahl Numbers

System	CPU count	GIPS [GHz]	RAM [GB]	diskIO [MB/s]	Amdahl	
					RAM	IO
BeoWulf	100	300	200	3000	0.67	0.08
Desktop	2	6	4	150	0.67	0.2
Cloud VM	1	3	4	30	1.33	0.08
SC1	212992	150000	18600	16900	0.12	0.001
SC2	2090	5000	8260	4700	1.65	0.008
GrayWulf	416	1107	1152	70000	1.04	0.506

Amdahl Numbers for Data Sets



The Data Sizes Involved



Petascale Computing at JHU

- Distributed SQL Server cluster/cloud w.
- 50 servers, 1PB disk, 500 CPU
- Connected with 20 Gbit/sec Infiniband
- 10Gbit lambda uplink to UIC
- Funded by Moore Foundation, Microsoft and Pan-STARRS
- Dedicated to eScience, provide public access through services
- Linked to 1000 core compute cluster
- Room contains >100 of wireless temperature sensors

graywulf

GrayWulf Performance

- Demonstrated large scale scientific computations involving ~200TB of DB data
- DB speeds close to “speed of light” (72%)
- Scale-out over SQL Server cluster
- Aggregate I/O over 12 nodes
  - *17GB/s for raw IO, 12.5GB/s with SQL*
- **Scales to over 70GB/s for 46 nodes from $700K**
- Cost efficient: $10K/(GB/s)
- Excellent Amdahl number : 0.50

Emerging Trends for DISC

- Large data sets are here, solutions are not
- National Infrastructure does not match power law
- Even HPC projects choking on IO
- Scientists are “cheap”, also pushing to the limit
  - *We are still building your own…*
- Data archives become analysis facilities with smart data services
- Sociological trends:
  - *Data collection in ever larger collaborations (VO)*
  - *Analysis decoupled, off archived data by smaller groups*
- Data will be never co-located
  - *Streaming algorithms*
  - *“Data pipes” for distributed workflows*
  - *“Data diffusion”*

Cyberbricks/Amdahl Blades?

- **Scale down** the CPUs to the disks!
  - *Solid State Disks (SSDs)*
  - *1 low power motherboard per SSD*
- Current SSD parameters
  - *OCZ Vertex 120GB, 250MB/s read, 15,000 IOPS, $350*
  - *Intel X25-E 32GB, 250MB/s read, 35,000 IOPS, $450*
  - *Power consumption 0.2W idle, 1-2W under load*
- Typical low power motherboards
  - *Intel Atom Z530 + US15W chipset 5W at 1.6GHz*
- Combination is perfect Amdahl blade:
  - *200MB/s=1.6Gbits/s ⇔ 1.6GHz of Atom*

Amdahl Cluster at JHU

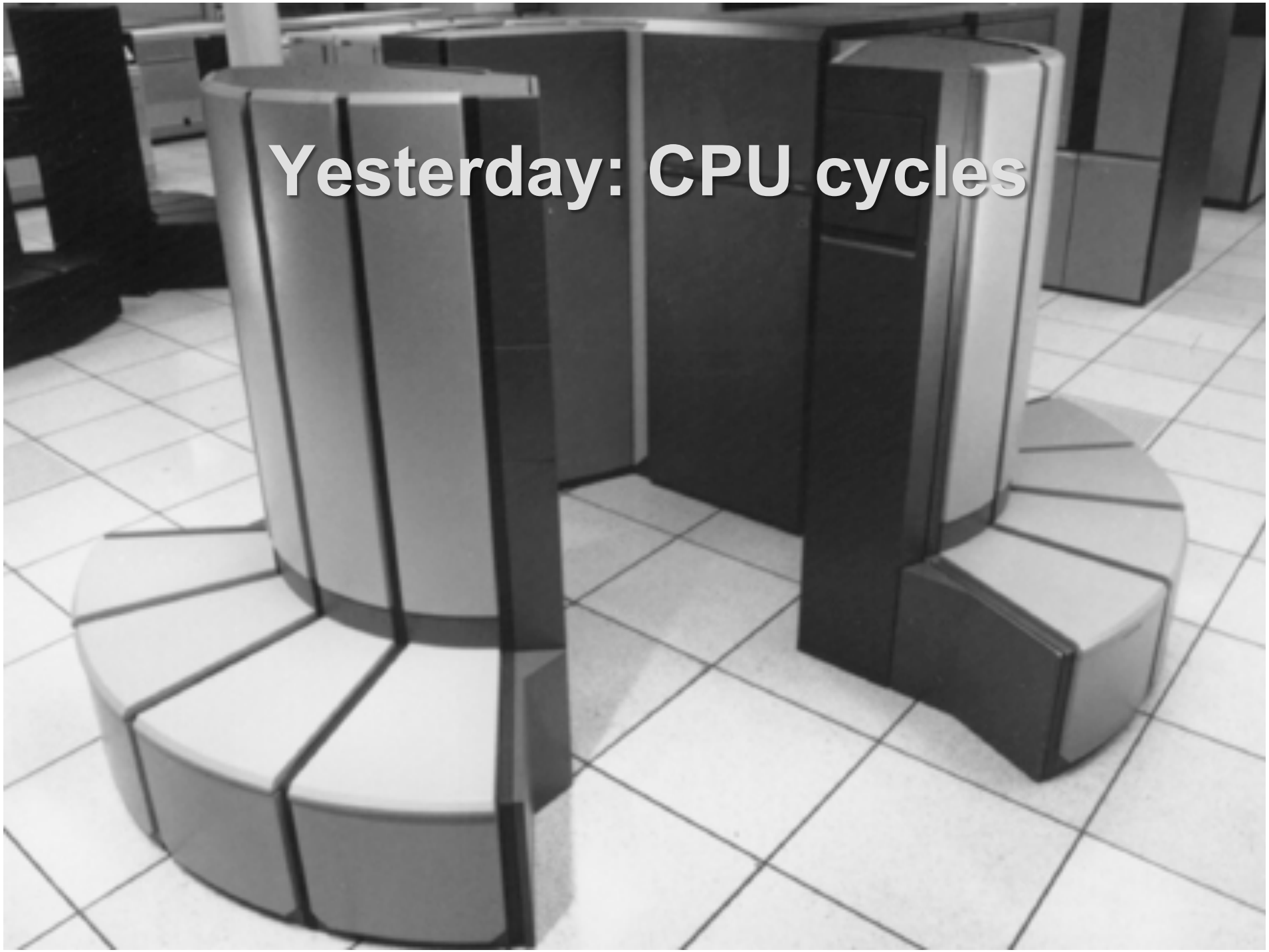
- 36-node cluster using 1200W (same as one GW!)
- Zotac Atom/ION motherboards
  - *4GB of memory, N330 dual core Atom, 16 GPU cores*
- Aggregate disk space 43.6TB
  - *63 x 120GB SSD = 7.7 TB*
  - *27x 1TB Samsung F1 = 27.0 TB*
  - *18x.5TB Samsung M1= 9.0 TB*
- Blazing I/O Performance: 18GB/s
- Cost is less than $30K
- Using the GPUs for data mining:
  - *6.4B multidimensional regressions in 5 minutes over 1.2TB*

How to Crawl Petabytes?

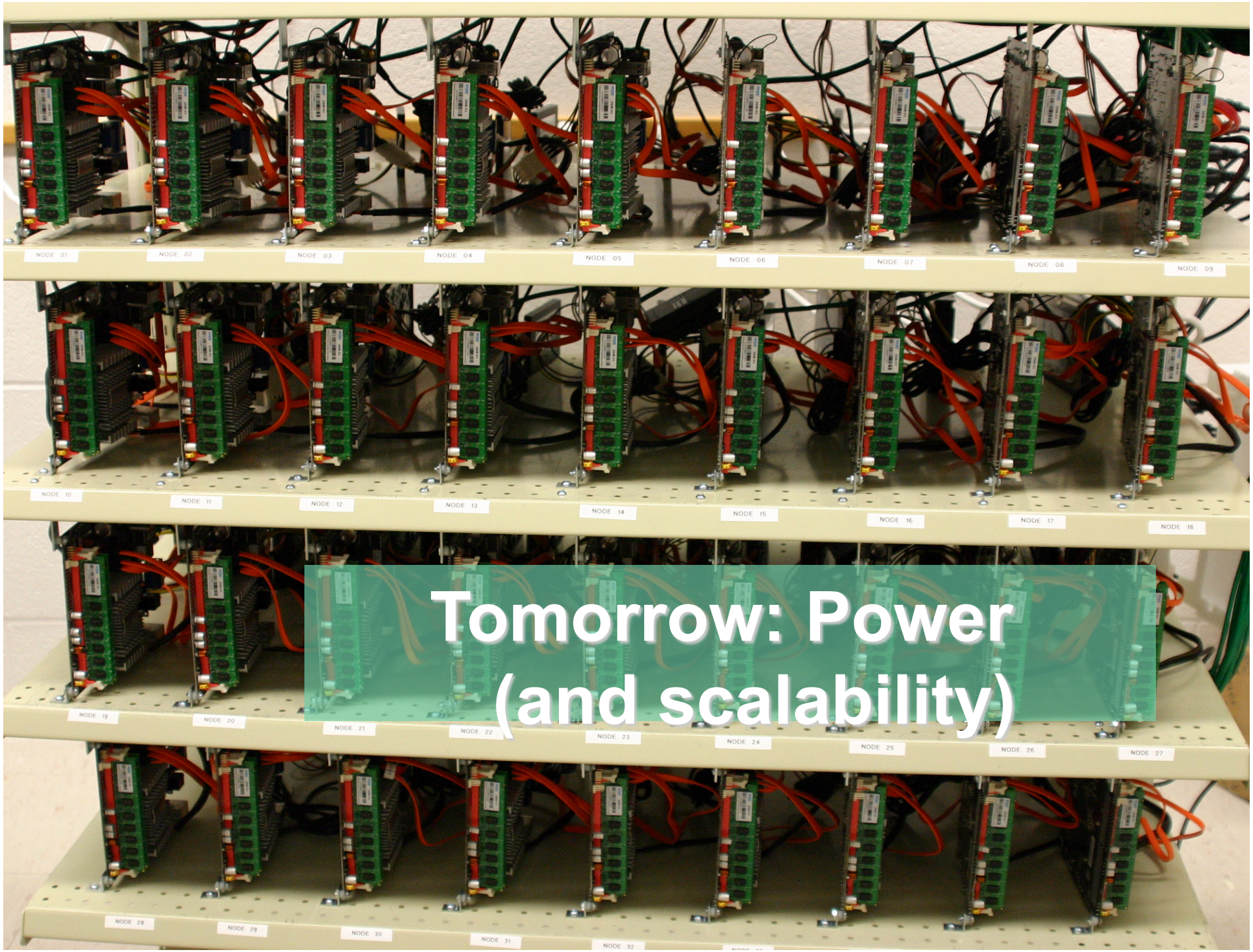
- Databases offer substantial performance advantages over MR (deWitt/Stonebraker)
- However: running a single SQL query over a monolithic dataset of petabytes is not meaningful
- Vision: scale-out partitioned systems, low level DB functionality, with high level intelligent crawling – relevance based priority queues (Nolan Li thesis@JHU)
- Automatic fault recovery, and self healing
- Stop, when answer is good enough….

Summary

- Science community starving for storage and IO
  - *Data-intensive computations as close to data as possible*
- Need an objective metrics for DISC systems
  - *Amdahl number appears to be good match to apps*
- Current architectures cannot scale much further
  - *Need to get off the curve leading to power wall*
  - *Multicores/GPGPUs + SSDs are a disruptive change!*
- Future in low-power, fault-tolerant architectures
  - *We propose scaled-out "Amdahl Data Clouds"*
  - *Smart "crawlers" emerging*
- Real reference applications for objective metrics
  - *Use large data sets for scalability studies (100TB+) e.g. SDSS, Pan-STARRS, Sensors, Turbulence*
- A new, Fourth Paradigm of science is emerging


Yesterday: CPU cycles

Today: Data Access


Tomorrow: Power
(and scalability)