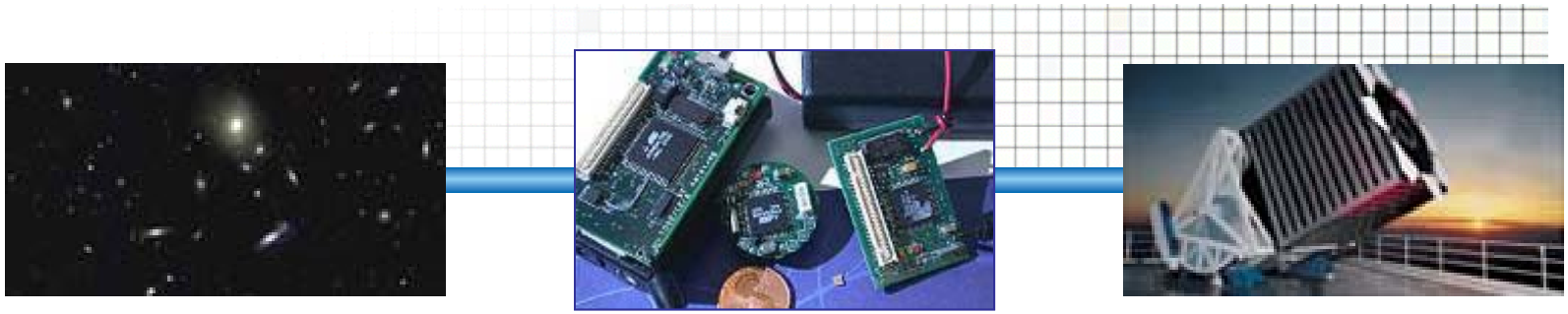
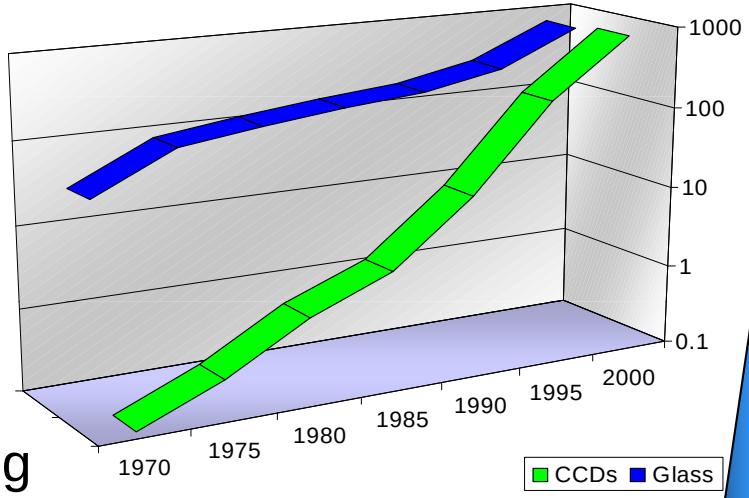


Amdahl's Laws and Extreme Data-Intensive Computing

Alex Szalay
The Johns Hopkins University

Living in an Exponential World

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



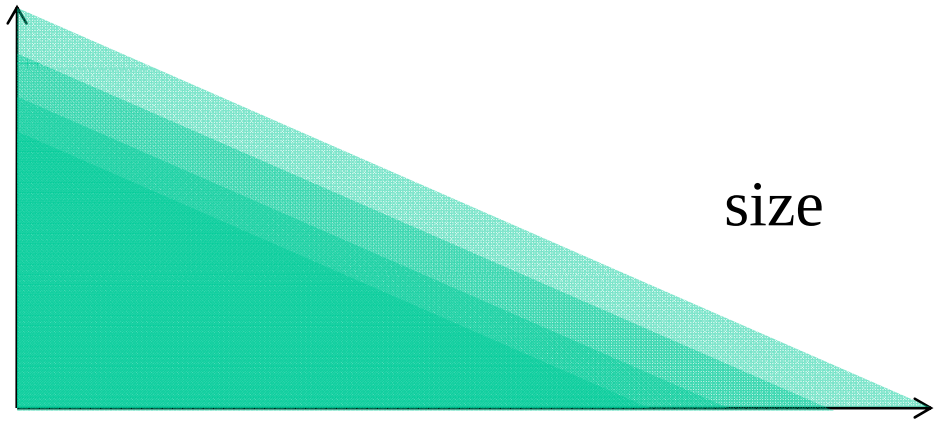
- Changes the nature of scientific computing
- Cuts across disciplines (eScience)
- It becomes increasingly harder to extract knowledge
- 20% of the world's servers go into huge data centers by the "Big 5"
  - *Google, Microsoft, Yahoo, Amazon, eBay*

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



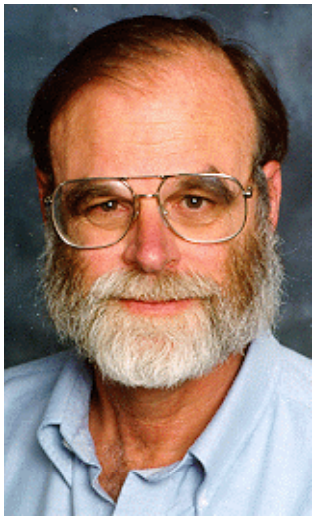
Scientific Data Analysis

- 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…**

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 Scientific 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:** *100TB total, 35TB in DB, in use for 8 years*
- **NVO :** *~5TB, few billion rows, in use for 4 years*
- **PanStarrs:** *80TB by 2011, 300+ TB by 2012*
- **Immersive Turbulence***: 30TB now, 100TB by Dec 2010*
- ***Sensor Networks:*** *200M measurements now, forming complex relationships*

Key Questions:

- What are the reasonable tradeoffs for DISC?
- 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

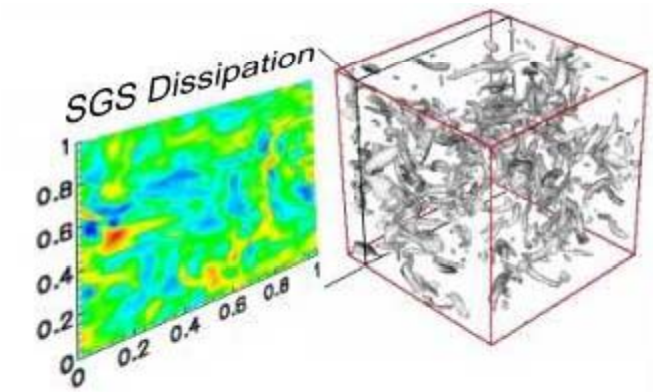
Public Use of the SkyServer

- **Prototype in 21st Century data access**
  - *774 million web hi*
  - *930,000 distinct us*
    vs 10,000 astrono
  - *Delivered 50,000 l*
    of lectures to high
  - *Delivered 100B ro*
  - *Everything is a po*

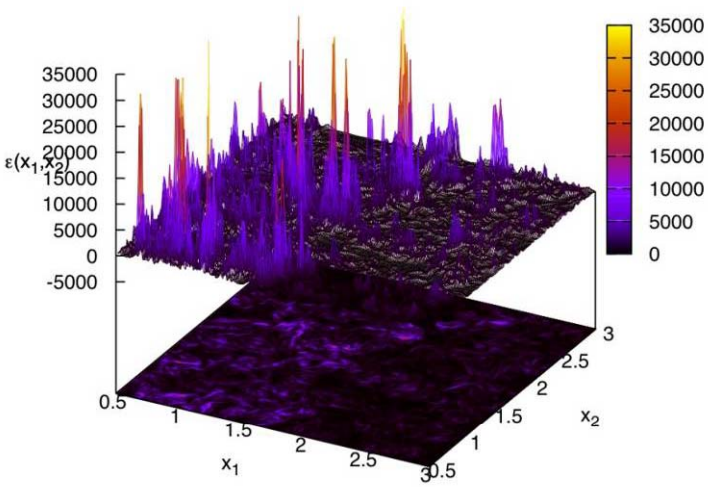
- **GalaxyZoo**
  - *40 million visual g*
  - *Enormous publicit*
  - *100,000 people participating, blogs, poems, ….*
  - *Now truly amazing original discovery by a schoolteacher from Holland*

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!



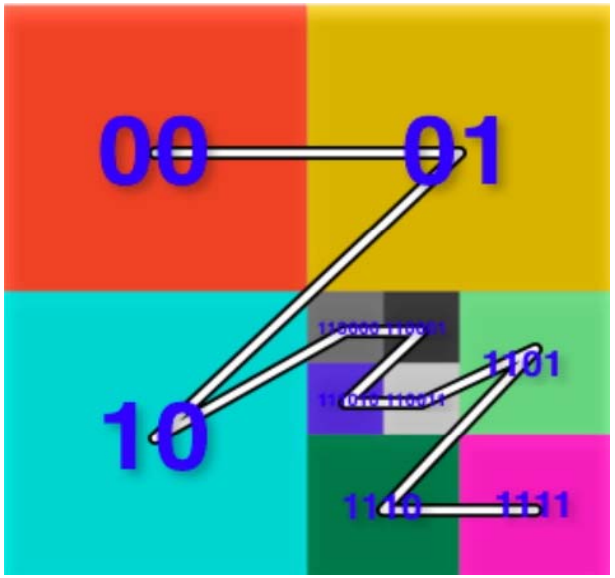
with C. Meneveau, S. Chen (Mech. E), G. Eyink (Applied Math), R. Burns (CS)

Database design philosophies

- "Move operations as close as possible to the data":
  Most elementary operations in analysis of CFD data (constrained on locality):
  - Differentiation (high-order finite-differencing)
  - Interpolation (Lagrange polynomial interpolation)

- "Storage schema must facilitate rapid searches"
  - Most basic search: given *x,y,z,t* position, find field variables (*u,v,w,p*).
  - Define elementary data-cube (optimize size relative to typical queries) and arrange along Z curve and indexing using oct-tree:

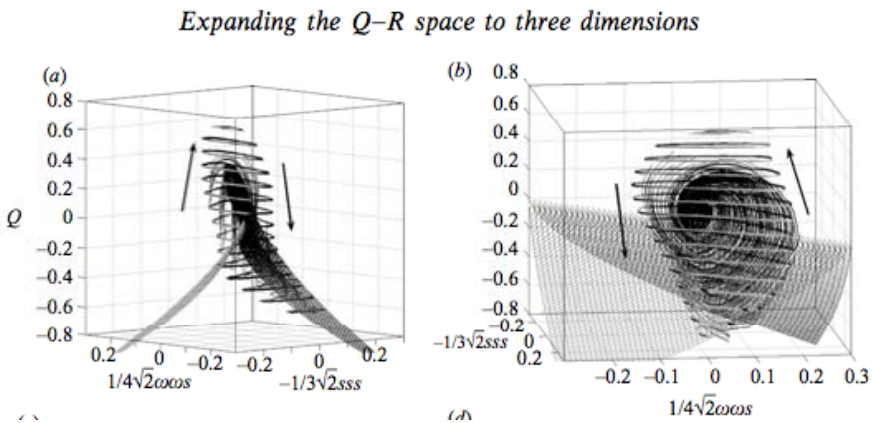
$j=$	0	1	2	3	4	5	6	7
$i=0$	0	1	4	5	16	17	20	21
$i=1$	2	3	6	7	18	19	22	23
$i=2$	8	9	12	13	24	25	28	29
$i=3$	10	11	14	15	26	27	30	31
$i=4$	32	33	36	37	48	49	52	53
$i=5$	34	35	38	39	50	51	54	55
$i=6$	40	41	44	45	56	57	60	61
$i=7$	42	43	46	47	58	59	62	63



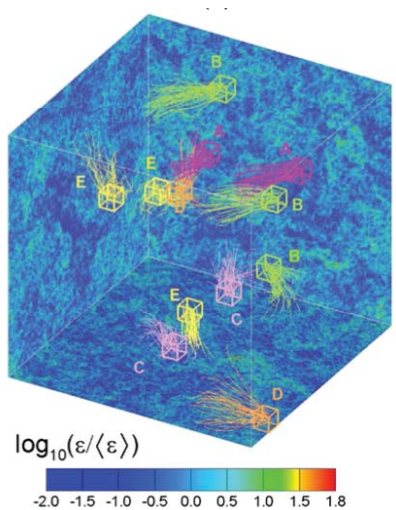
Sample Applications

Experimentalists testing PIV-based pressure-gradient measurement
(X. Liu & Katz, 61 APS-DFD meeting, November 2008)

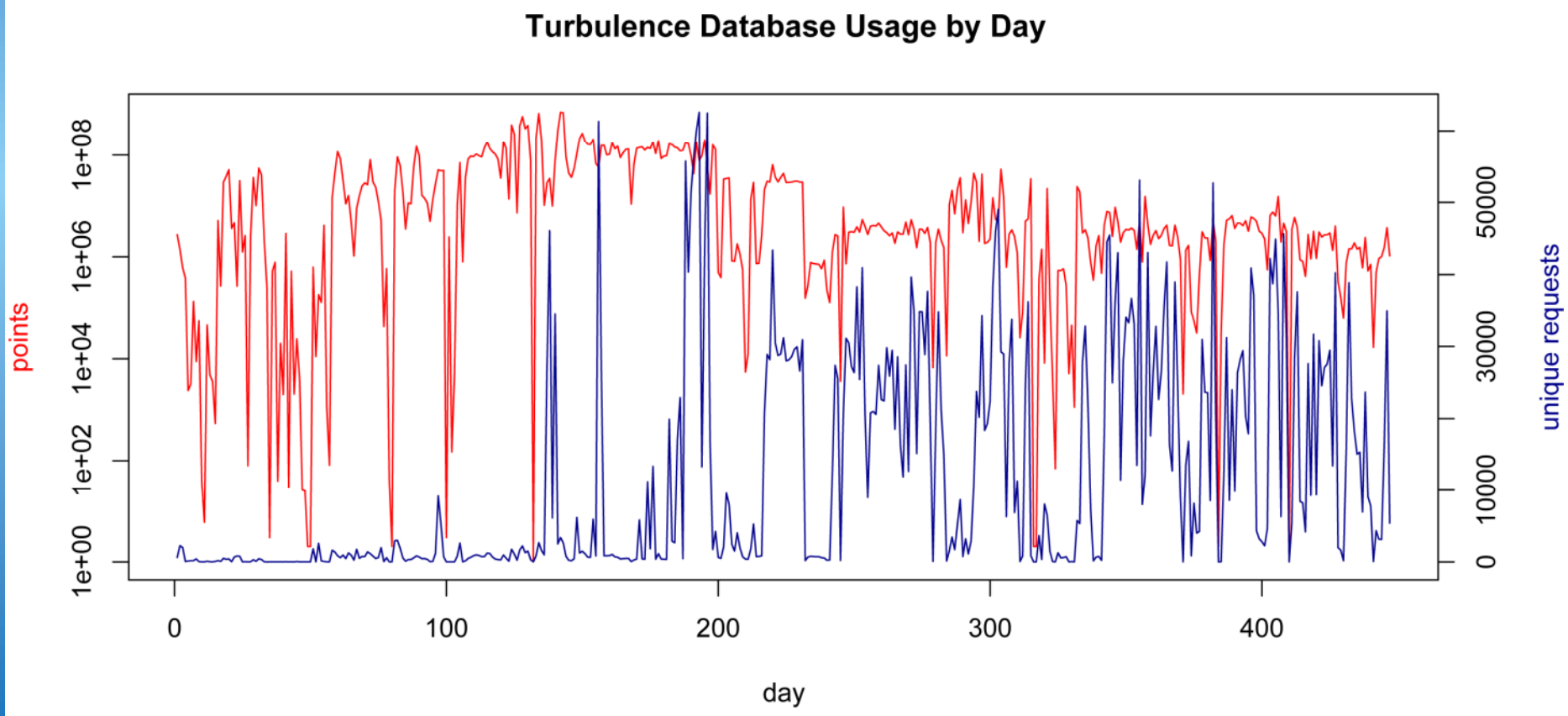
Measuring velocity gradient using a new set of 3 invariants
Luethi, Holzner & Tsinober,
J. Fluid Mechanics 641, pp. 497-507 (2010)



Lagrangian time correlation in turbulence
Yu & Meneveau
Phys. Rev. Lett. 104, 084502 (2010)

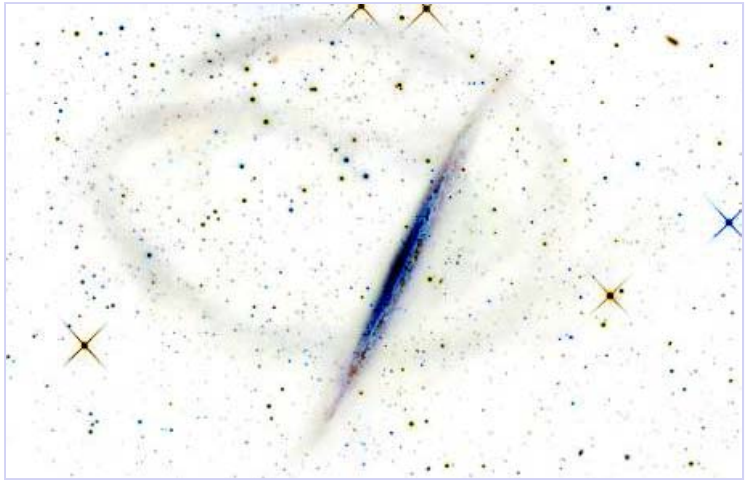


Daily Usage


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

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)*
- E.g. users insert test particles (dwarf galaxies) into system and follow trajectories in precomputed simulation
  - *Realistic "streams" from tidal disruption*
  - *Users interact remotely with 0.5PB in 'real time'*



500 Universes Project

- Gadget2/tree with 320^3 particles, 640^3 grid
  - *500 realizations of a 1(Gpc/h)^3 box*
  - *Cosmological parameters following Millennium XL (WMAP5)*
  - *In progress…*
- Outputs:
  - *100 snapshots of (x,v) and density grid*
  - *1000 snapshots of large scale Fourier modes (128^3)*
  - *Total storage ~100TB*
- Science
  - *Detailed topology estimators (walls, filaments, halos)*
  - *Conditional non-linear growth of large scale modes*
  - *NL evolution of baryon bumps in redshift space*

  … public facility …

Commonalities

- Huge amounts of data, aggregates needed
  - *But also need to keep raw data*
  - *Need for parallelism*
- Use patterns <u>enormously benefit</u> from indexing
  - *Rapidly extract small subsets of large data sets*
  - *Geospatial everywhere*
  - *Compute aggregates*
  - *Fast sequential read performance is critical!!!*
  - *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 <u>no need for transactions</u>
- Design pattern: class libraries wrapped in SQL UDF
  - *Take analysis to the data – run it all inside the DB!!*

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*

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)

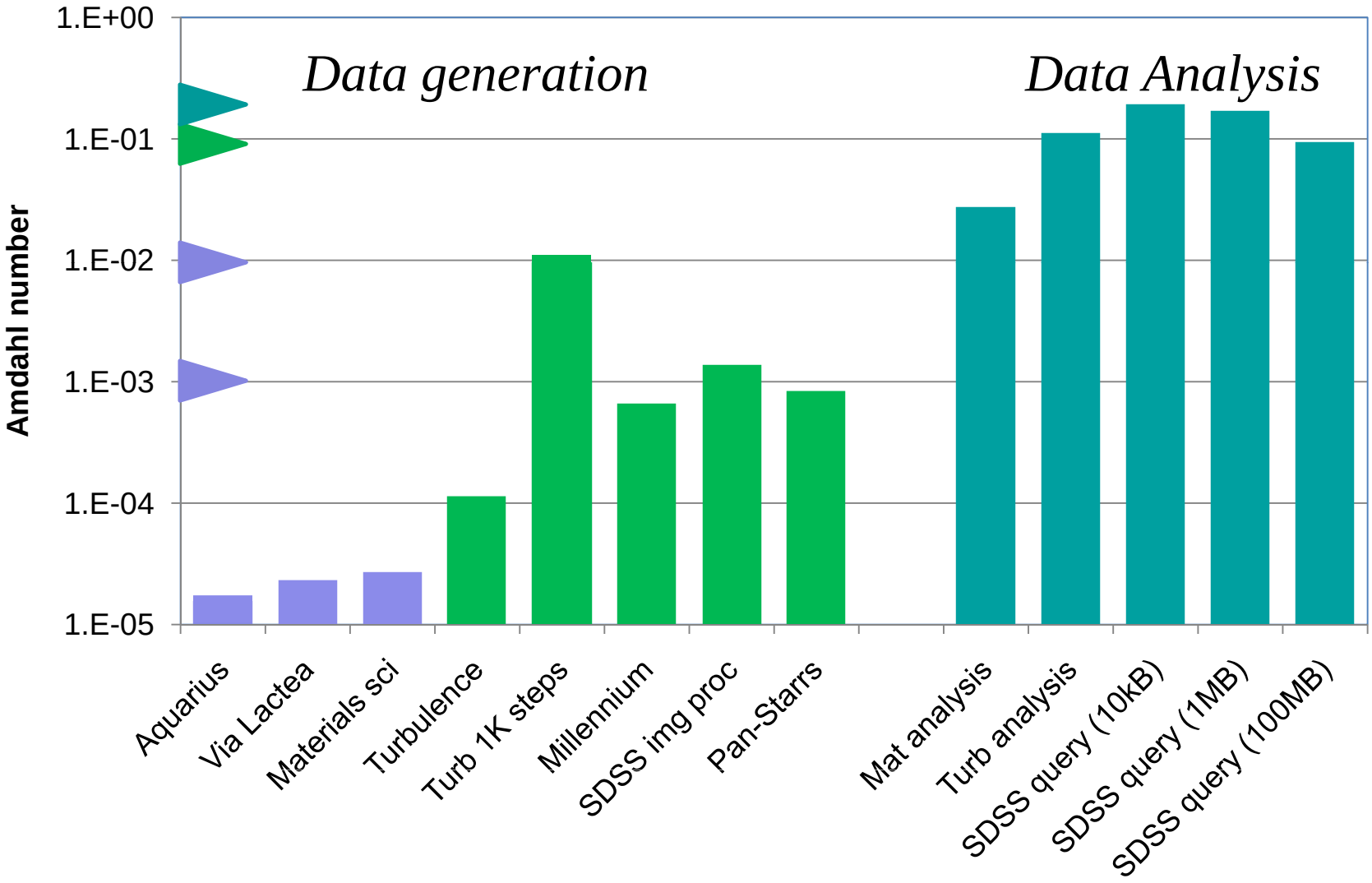
iii. One byte of memory per one instruction/sec (MEM)

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

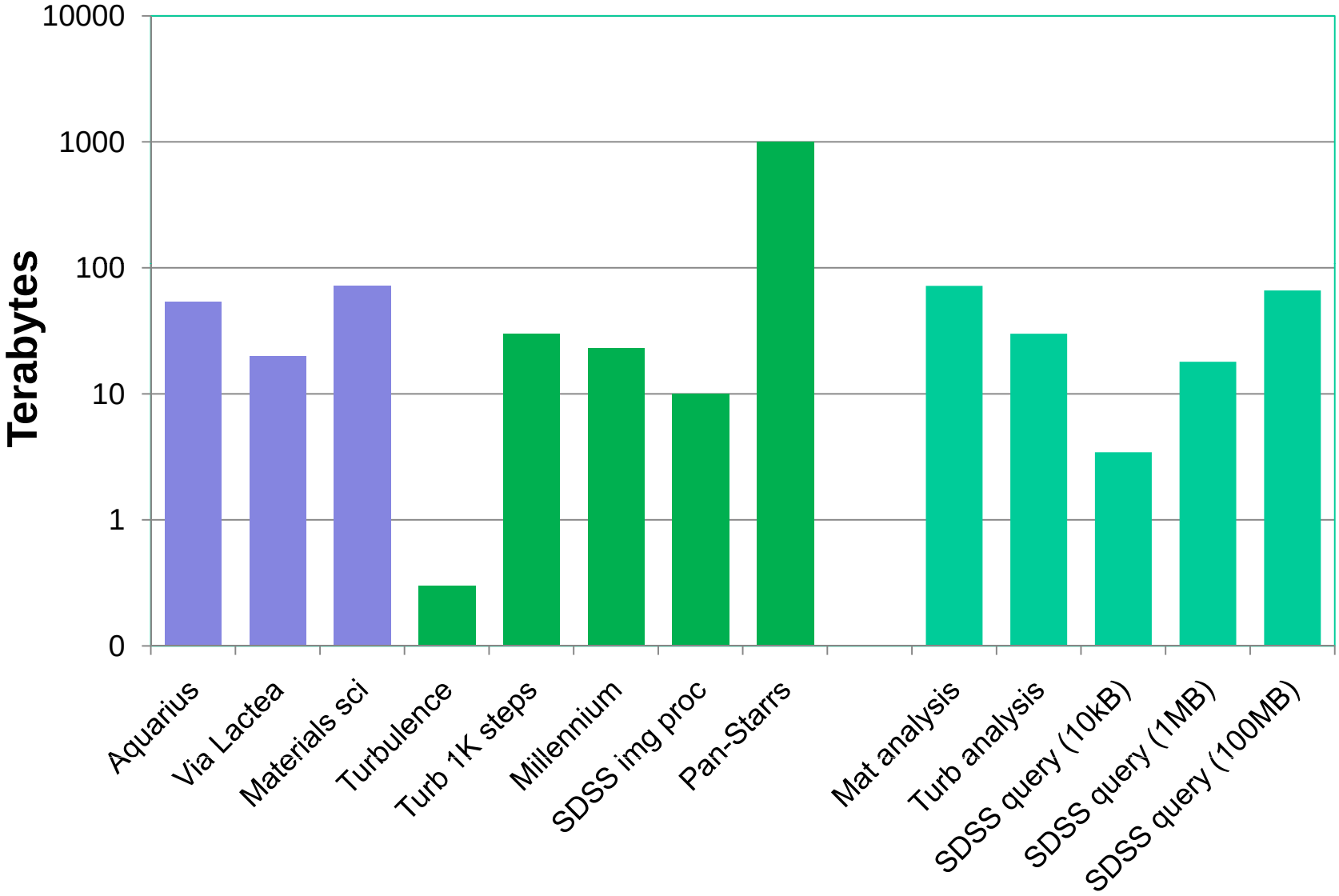
Typical Amdahl Numbers

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



DISC Needs Today

- Disk space, disk space, disk space!!!!
- Current problems not on Google scale yet:
  - *10-30TB easy, 100TB doable, 300TB really hard*
  - *For detailed analysis we need to park data for several mo*
- Sequential IO bandwidth
  - *If not sequential for large data set, we cannot do it*
- How do can move 100TB within a University?
  - *1Gbps 10 days*
  - *10 Gbps 1 day (but need to share backbone)*
  - *100 lbs box few hours*
- From outside?
  - *Dedicated 10Gbps or FedEx*

Tradeoffs Today

Stu Feldman (Google):

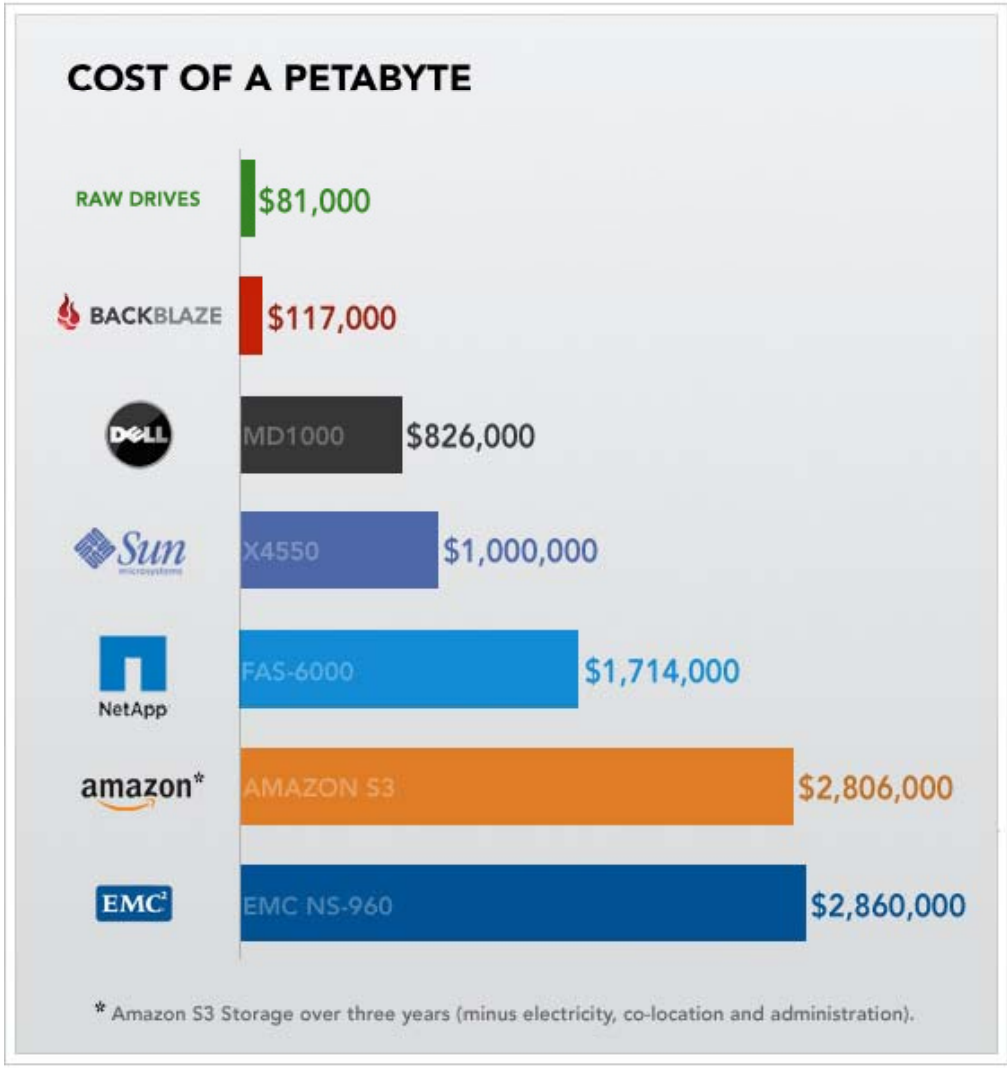
Extreme computing is about tradeoffs!

Ordered priorities for data-intensive science:

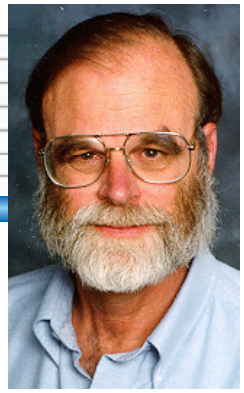
1. *Total storage (low redundancy)*
2. *Cost (total cost vs price of raw disks)*
3. *Sequential IO (total throughput vs. all disks, locally attached)*
4. *Fast stream processing (GPUs inside server, next to disks)*
5. *Low power (slow normal CPUs, lots of disks/mobo)*

The order will be different in a few years...and scalability will appear as well

Cost of a Petabyte



From backblaze.com



- Distributed SQLServer cluster/cloud
  - *50 servers, 1.1PB disk, 500 CPU*
  - *Connected with 20 Gbit/sec Infiniband*
  - *Linked to 1500 core compute cluster*
  - *Extremely high speed seq I/O (75GB/s)*
  - *Balanced: Amdahl number >0.5*

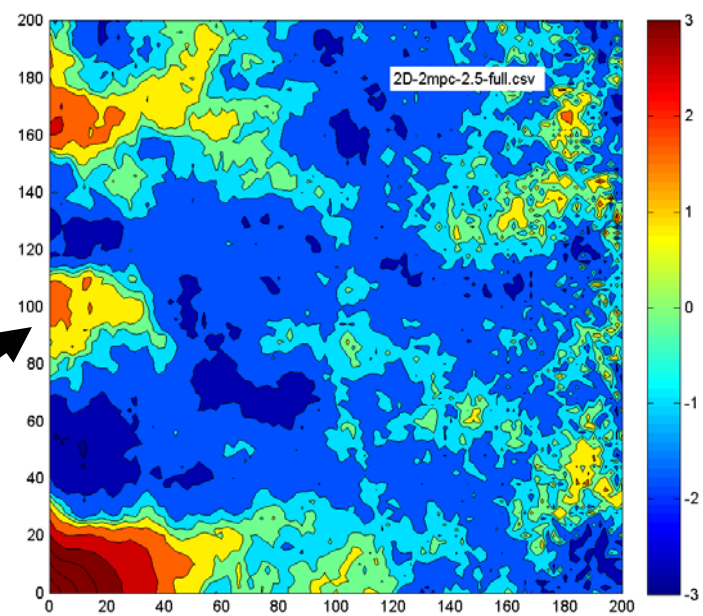
- Dedicated to eScience, provide public access through services
- Funded by Moore Foundation, Microsoft Research and Pan-STARRS
- Now similar machines in Potsdam and in Munich

Future: Cyberbricks?

- 36-node Amdahl cluster using 1200W total
- 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
- Amdahl number = 1!
- Cost is less than $30K
- Using the GPUs for data mining:
  - *6.4B multidimensional regressions in 5 minutes over 1.2TB*

The Impact of GPUs

- We need to reconsider the N logN only approach
- Once we can run 100K threads, maybe running SIMD N^2 on smaller partitions is also acceptable
- Potential impact on genomics huge
  - *Sequence matching using parallel brute force vs dynamic programming?*
- Recent JHU effort on integrating CUDA with SQL Server, using SQL UDF
- Galaxy correlation functions: 400 trillion galaxy pairs!



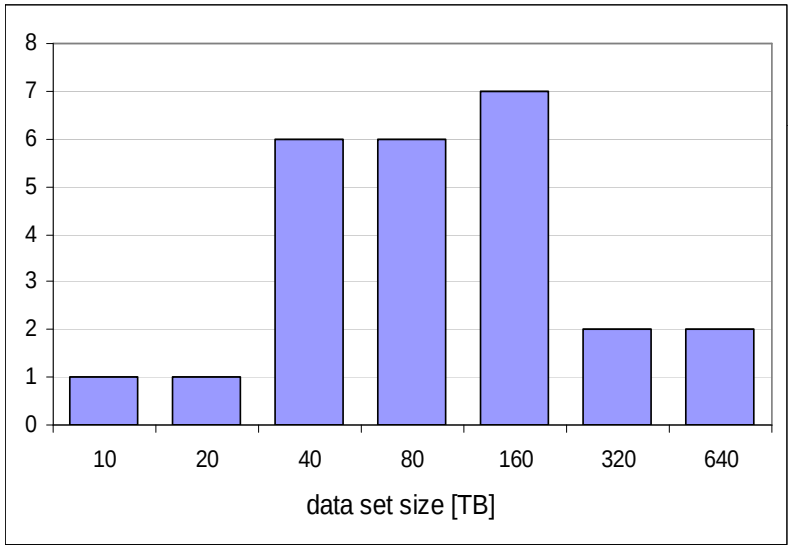
Data-Scope

- Proposal to NSF MRI to build a new ‘instrument’ to look at data
- 102 servers for $1M + about $200K switches+racks
- Two-tier: performance (P) and storage (S)
- Large (5PB) + cheap + fast (460GBps), but …
  . ..a special purpose instrument

	1P	*1S*	*90P*	*12S*	*Full*	
servers	1	1	90	12	102	
rack units	4	12	360	144	504	
capacity	24	252	2160	3024	5184	TB
price	8.5	22.8	766	274	1040	$K
power	1	1.9	94	23	116	kW
GPU	3	0	270	0	270	TF
seq IO	4.6	3.8	414	45	459	GBps
netwk bw	10	20	900	240	1140	Gbps

Proposed Projects at JHU

Discipline	data [TB]
Astrophysics	930
HEP/Material Sci.	394
CFD	425
BioInformatics	414
Environmental	660
Total	2823



19 projects total, data lifetimes between 3 mo and 3 yrs

Short Term Trends

- Large data sets are here, solutions are not
  - *100TB is the current practical limit*
- National Infrastructure does not match power law
- No real data-intensive computing facilities available
  - *Some are becoming a "little less CPU heavy"*
- Even HPC projects choking on IO
- Cloud hosting currently very expensive
- Cloud computing tradeoffs different from science needs
- Scientists are "cheap", also pushing the limit
  - *We are still building our own…*
  - *We will see campus level aggregation*
  - *May become the gateways to future cloud hubs*

5 Year Trend

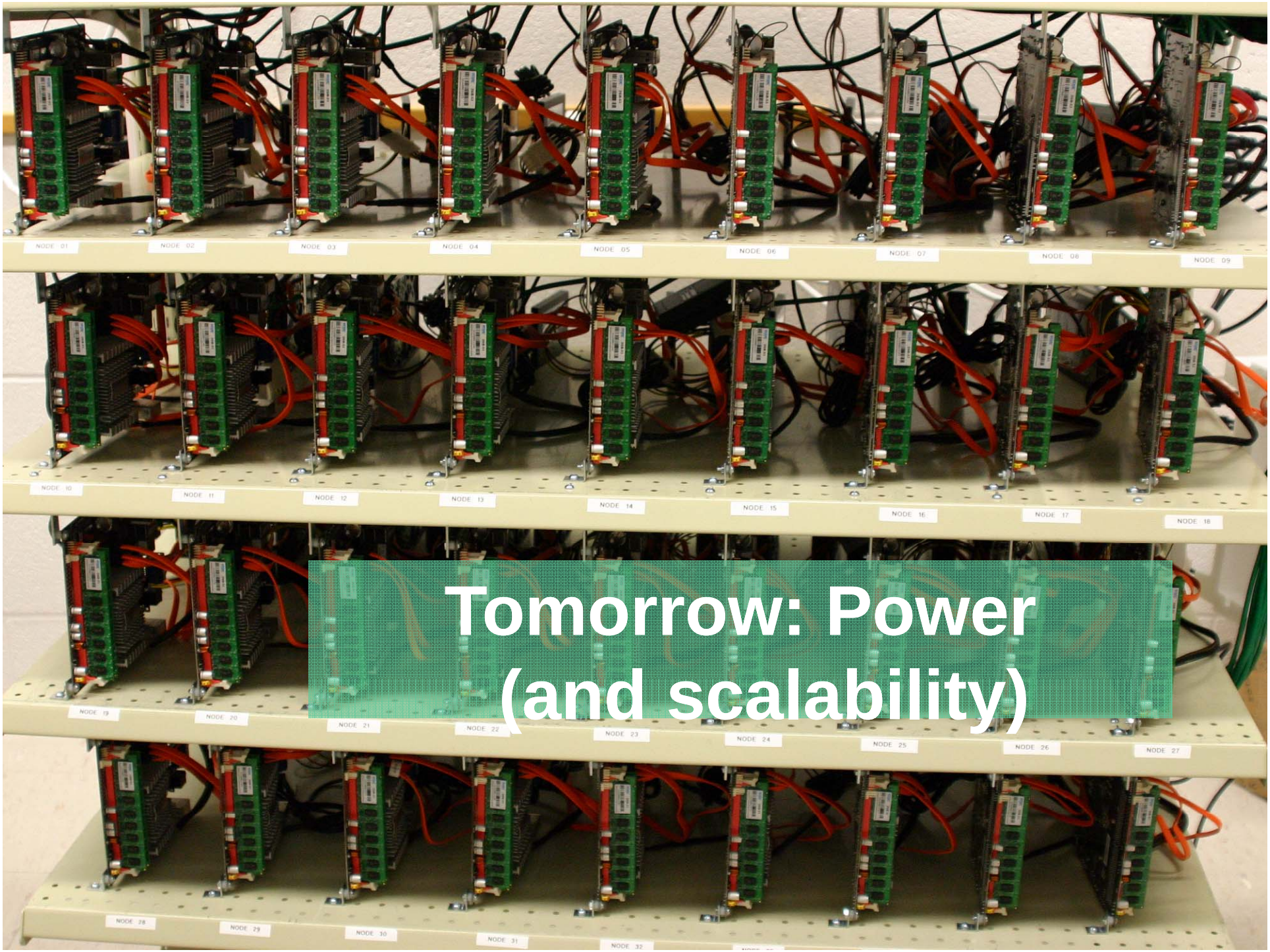
- Sociology:
  - *Data collection in ever larger collaborations (VO)*
  - *Analysis decoupled, off archived data by smaller groups*
  - *Data sets cross over to multi-PB*
- Some form of a scalable Cloud solution inevitable
  - *Who will operate it, what business model, what scale?*
  - *How does the on/off ramp work?*
  - *Science needs different tradeoffs than eCommerce*
- Scientific data will never be fully co-located
  - *Geographic origin tied to experimental facilities*
  - *Streaming algorithms, data pipes for distributed workflows*
  - *"Data diffusion"?*
  - *Containernet (Church, Hamilton, Greenberg 2010)*

Summary

- Science community starving for storage and IO
  - *Data-intensive computations as close to data as possible*
- Real multi-PB solutions are needed NOW!
  - *We have to build it ourselves*
- Current architectures cannot scale much further
  - *Need to get off the curve leading to power wall*
  - *Multicores/GPGPUs + SSDs are a disruptive change!*
- Need an objective metrics for DISC systems
  - *Amdahl number appears to be good match to apps*
- Future in low-power, fault-tolerant architectures
  - *We propose scaled-out "Amdahl Data Clouds"*
- A new, Fourth Paradigm of science is emerging
  - *Many common patterns across all scientific disciplines*

Yesterday: CPU cycles

Today: Data Access



Tomorrow: Power (and scalability)