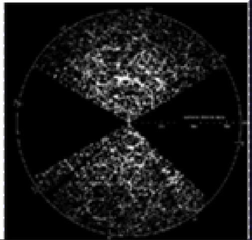
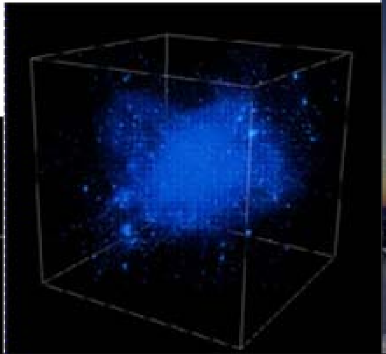
Institute for Data Intensive Engineering and Science

idies

Extreme Data-Intensive Scientific Computing

Alexander Szalay
JHU





Big Data in Science

- Data growing exponentially, in all science
- All science is becoming data-driven
- This is happening very rapidly
- Data becoming increasingly open/public
- Non-incremental!
- Convergence of physical and life sciences through Big Data (statistics and computing)
- The "long tail" is important
- A scientific revolution in how discovery takes place

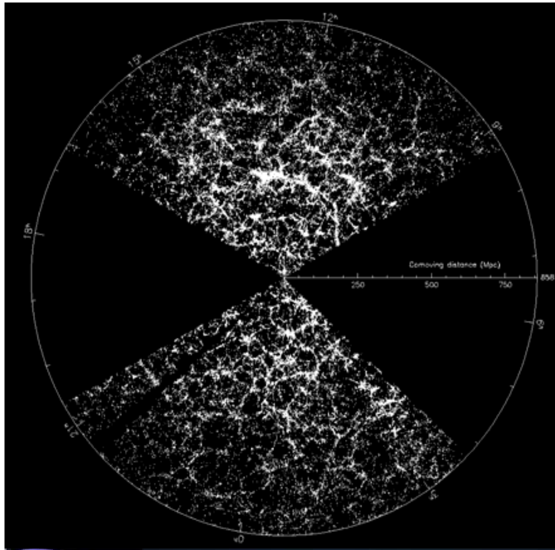
=> a rare and unique opportunity

Scalable Data-Intensive Analysis

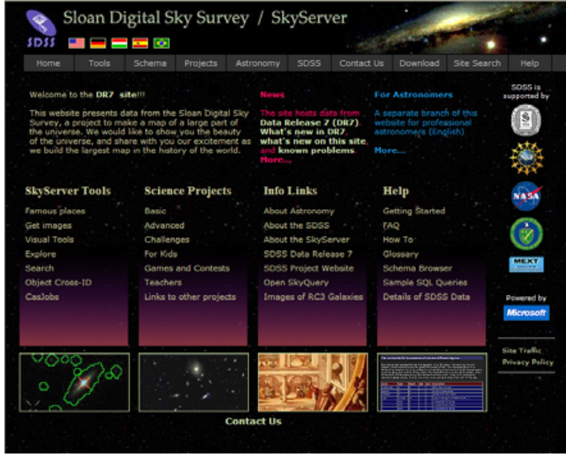
- Large data sets => data resides on hard disks
- Analysis has to move to the data
- Hard disks are becoming sequential devices
  - For a PB data set you cannot use a random access pattern
- Analyses and visualization become streaming problems
- Same thing is true with searches
  - Massively parallel sequential crawlers (MR, Hadoop, etc)
- Indexing needs to be maximally sequential
  - Space filling curves (Peano-Hilbert, Morton,…)
- Need streaming versions of our algorithms

Sloan Digital Sky Survey

- "The Cosmic Genome Project"
- Two surveys in one
  - Photometric survey in 5 bands
  - Spectroscopic redshift survey
- Data is public
  - 2.5 Terapixels of images => 5 Tpx
  - 10 TB of raw data => 120TB processed
  - 0.5 TB catalogs => 35TB in the end
- Started in 1992, finished in 2008
- Data volume enabled by Moore's Law, Kryder's Law



SkyServer

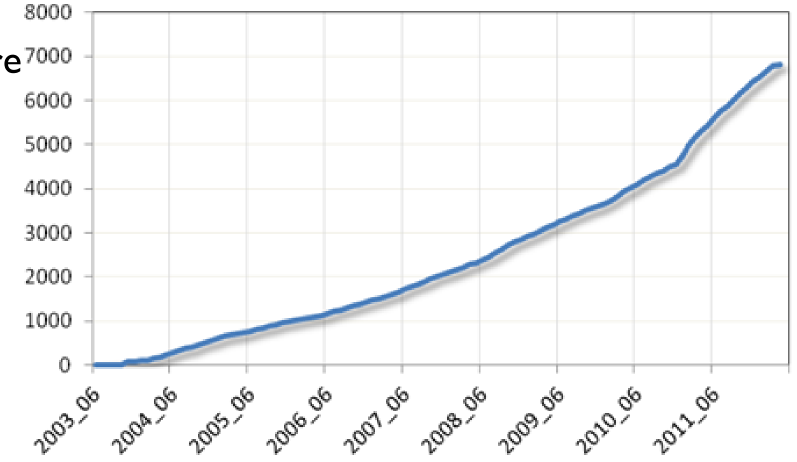


- Prototype in 21st Century data access
  - One billion web hits in 10 years
  - 4,000,000 distinct users vs. 15,000 astronomers
  - The emergence of the "Internet scientist"
  - The world's most used astronomy facility today
  - Collaborative server-side analysis done

MyDB: Workbench

- Registered 'power users', with their own server-side DB (Nolan Li)
- Query output goes to 'MyDB'
- Can be joined with source database (contexts) or with other tables
- Results are materialized from MyDB upon request
- Users can collaborate!
  - Insert, Drop, Create, Select Into, Functions, Procedure
  - **Publish/share** their tables to a group area
  - Flexibility "at the edge"/ Read-only big DB
- Data delivery via Web Services

=> Sending analysis to the data!



What is Different about Data-Intensive?

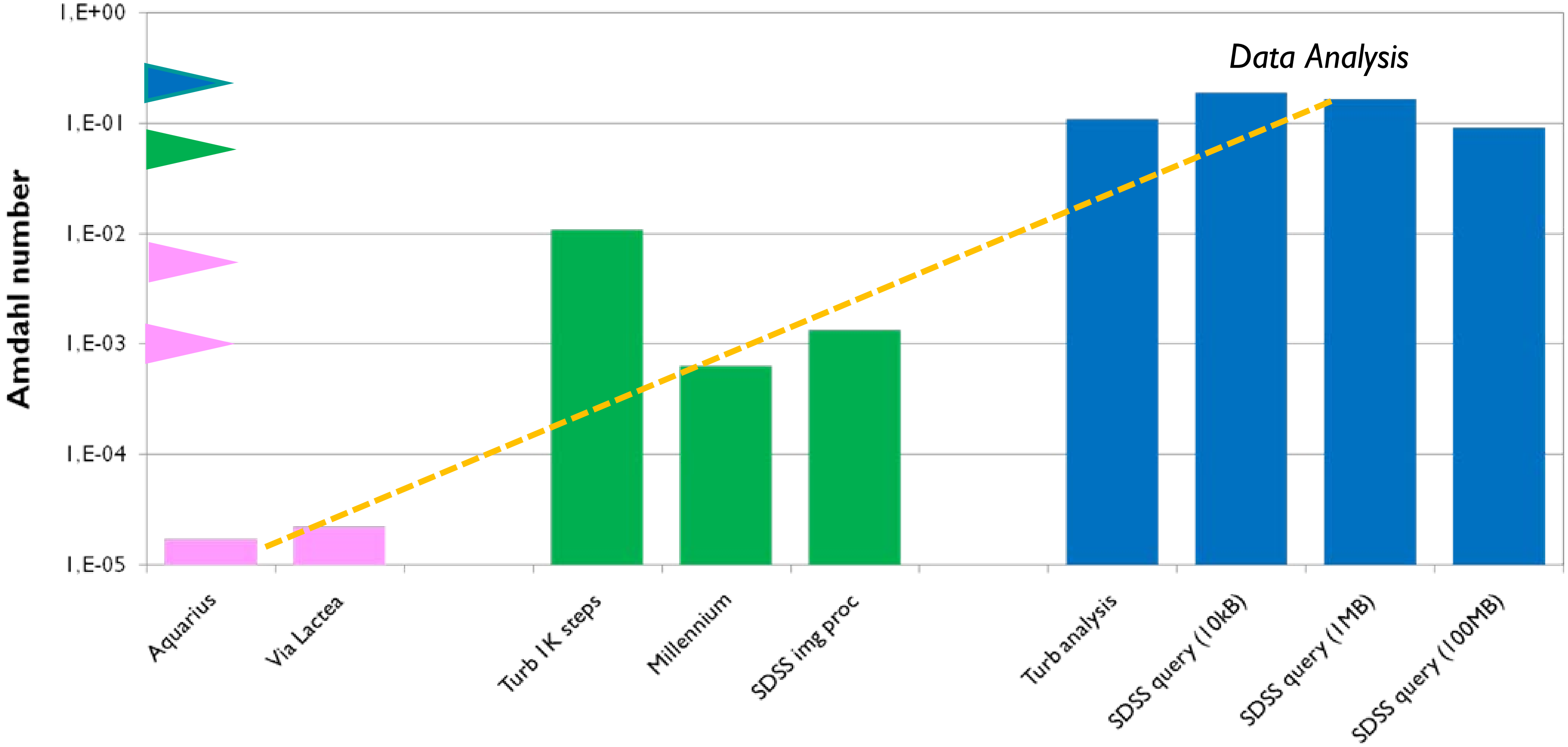
- Data is hard (and costly) to move
- Data locality is the key!
- Typical data analysis scenarios are hierachical
- At least the first stage requires data filtering/censoring/extraction
  - Usually very low cycles/byte of data
- **Amdahl (1965): Laws for a balanced system**
  1. Parallelism: max speedup is S/(S+P)
  2. One bit of IO/sec per instruction/sec (BW)
  3. One byte of memory per one instruction/sec (MEM)

Typical Amdahl Numbers

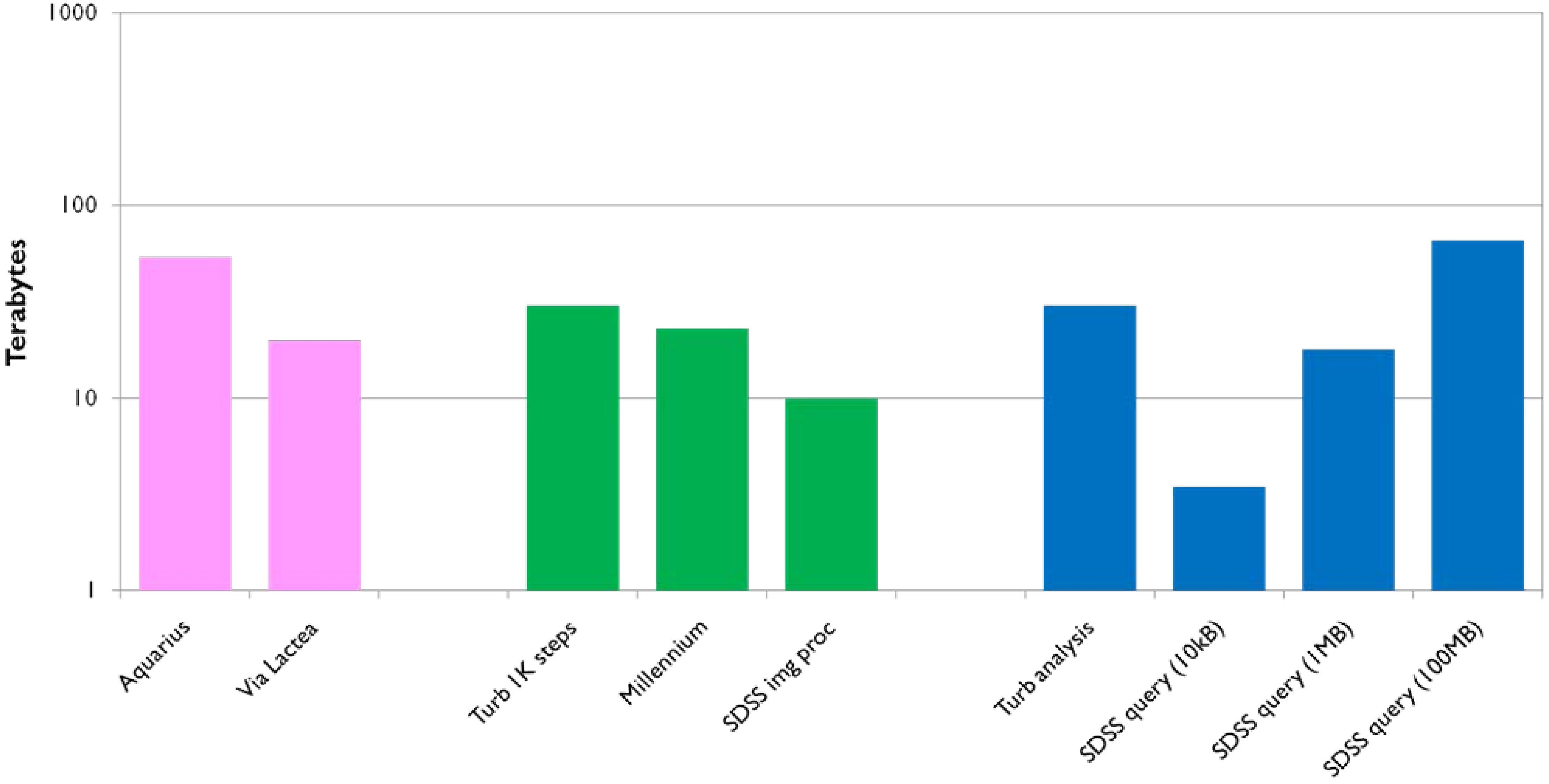
System	CPU count	GIPS [GHz]	RAM [GB]	diskIO [MB/s]	Amdahl	
					RAM	IO
BeoWulf	100	300	200	3000	0.67	0.08
Desktop	2	6	4	150	0.67	0.2
Cloud VM	1	3	4	30	1.33	0.08
SC1	212992	150000	18600	16900	0.12	0.001
SC2	2090	5000	8260	4700	1.65	0.008
GrayWulf	416	1107	1152	70000	1.04	0.506

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

Amdahl Numbers for Data Sets



Data Sizes Involved



Data in HPC Simulations

- HPC is an instrument in its own right
- Largest simulations approach petabytes
  - from supernovae to turbulence, biology and brain modeling
- Need public access to the best and latest sims through interactive numerical laboratories
- Creates new challenges in how to:
  - Move the petabytes of data (high speed networking)
  - Interface (virtual sensors, immersive analysis)
  - Look at it (render on top of the data, drive remotely)
  - Analyze (algorithms, scalable analytics)
  - Support and archive (long term strategies)

Usage Scenarios for Simulation Outputs

- On-the fly analysis (immediate)
- Private reuse (short/mid term)
- Public reuse (mid term)
- Public service portal (mid/long term)
- Archival and curation (long term)

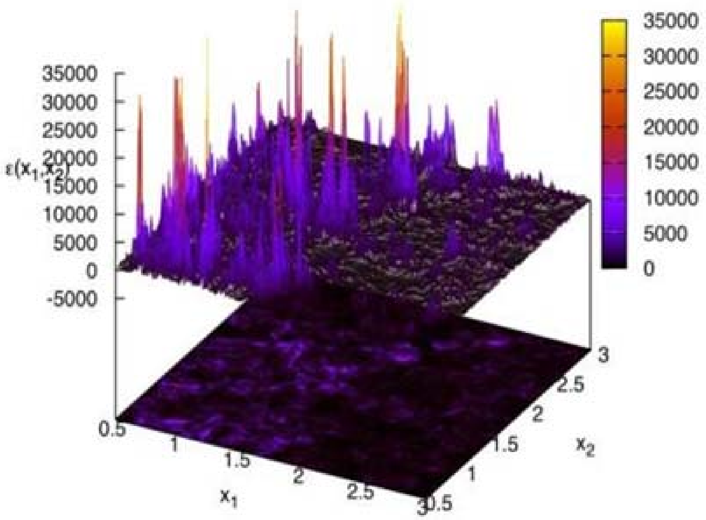
Silver River Network Transfer

- Simulation run on Jaguar
- 150TB in less than 10 days from Oak Ridge to JHU using a dedicated 10G connection

Immersive Turbulence

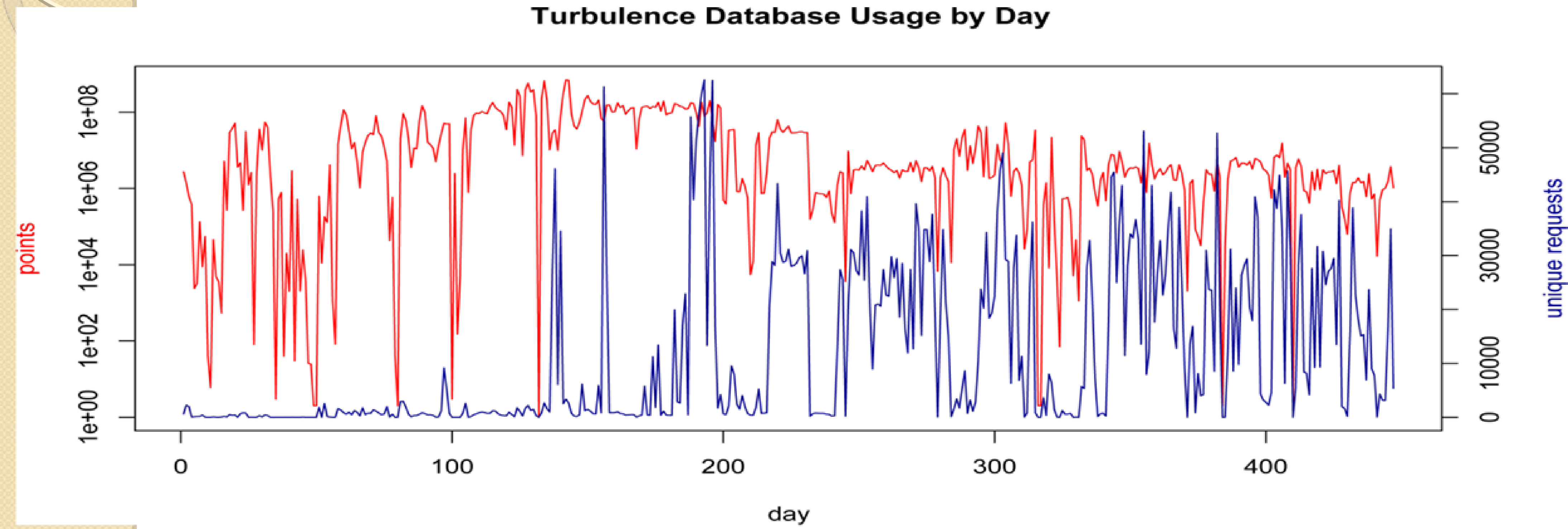
"… the last unsolved problem of classical physics…"
Feynman

- Understand the nature of turbulence
  - Consecutive snapshots of a large simulation of turbulence: now 30 Terabytes
  - Treat it as an experiment, **play** with the database!
  - **Shoot test particles** (sensors) from your laptop into the simulation, like in the movie Twister
  - Next: 70TB MHD simulation
- New paradigm for analyzing simulations!



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

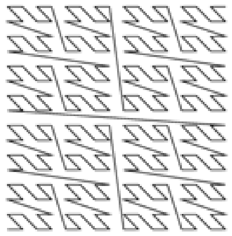
Typical Daily Usage

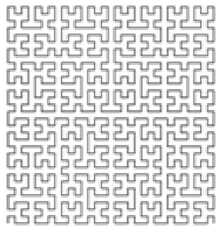


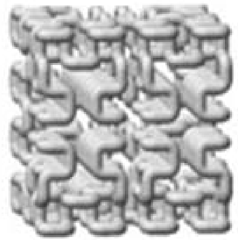
2011: exceeded 100B points publicly delivered

Spatial queries, random samples

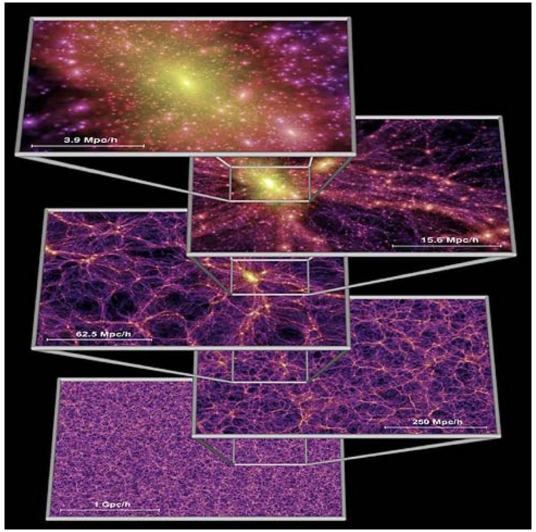
- Spatial queries require multi-dimensional indexes.
- (x,y,z) does not work: need discretisation
  - index on (ix,iy,iz) with ix=floor(x/8) etc
- More sophisticated: space filling curves
  - bit-interleaving/octree/Z-Index
  - Peano-Hilbert curve
  - Need custom functions for range and volume queries
  - Plug in modular space filling library (Budavari)







Cosmological Simulations



In 2000 cosmological simulations had 10^{10} particles and produced over 30TB of data (Millennium)

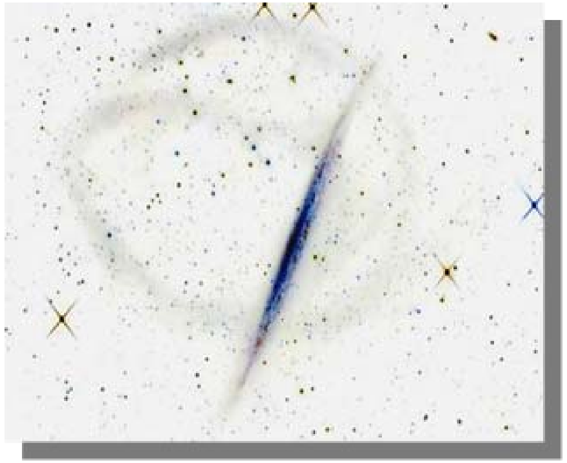
- Build up dark matter halos
- Track merging history of halos
- Use it to assign star formation history
- Combination with spectral synthesis
- Realistic distribution of galaxy types
- More than 1,000 CASJobs/MyDB users

Today: simulations with 10^{12} particles and PB of output are under way (MillenniumXXL, Silver River, Exascale Sky)
but there is not enough disk space to store the output!

The Milky Way Laboratory

- Cosmology simulations as immersive laboratory for general users
- Via Lactea-II (20TB) as prototype, then Silver River (50B particles) as production (50M CPU hours)
- 800+ hi-rez snapshots (2.6PB) => 1PB in DB
- Users can insert test particles (dwarf galaxies) into the system and follow trajectories in pre-computed simulation
- Compute dark matter annihilation maps interactively
- Users will interact remotely with a PB in 'real time'

Madau, Rockosi, Szalay, Wyse, Silk, Stadel,
Kuhlen, Lemson, Westermann, Blakeley

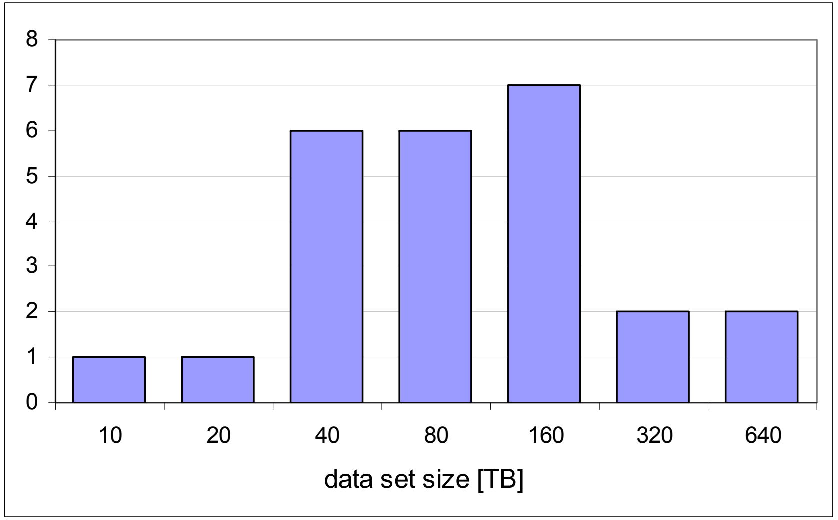


Visualizing Petabytes

- Send the rendering to the data …
- It is easier to send a HD 3D video stream to the user than all the data
- Interactive visualizations driven remotely
- Visualizations are becoming IO limited
- It is possible to build individual servers with extreme data rates
- Prototype on turbulence simulation already works: data streaming directly from DB to GPU
- N-body simulations next

Current Data-Intensive Projects at JHU

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



19 projects total proposed, more coming,
data lifetimes between 3 mo and 3 yrs

Tradeoffs for Data Analysis

Today, we have no good and cheap architecture for large scale data analysis

Extreme computing is about tradeoffs
--- Stu Feldman

Ordered priorities for data-intensive scientific computing

1. Total storage (-> low redundancy)
2. Cost (-> total cost vs price of raw disks)
3. Sequential IO (-> locally attached disks, fast ctrl)
4. Fast streams (->GPUs inside server)
5. Low power (-> slow normal CPUs, lots of disks/mobo)

Idea

- Data analysis: we need a fast scanning engine!
  - Users can park hundreds of TBs of data for months (but not permanent)
  - Two tiered architecture for split functionalities (analysis vs checkpointing)
  - Overall, minimize costs as possible, only use free SW
  - Use as fast an interconnect as possible
  - Build a BeoWulf-like template that can be replicated at other institutions
- Performance/analysis tier
  - Sacrifice distributed file system for locally attached storage with share nothing
  - Maximize data streaming from disk to GPU (5GBytes/sec nodes)
- Storage tier
  - Provide truly inexpensive large storage for checkpointing($70K/PB)
  - Maximize recovery from network

=> JHU Data-Scope

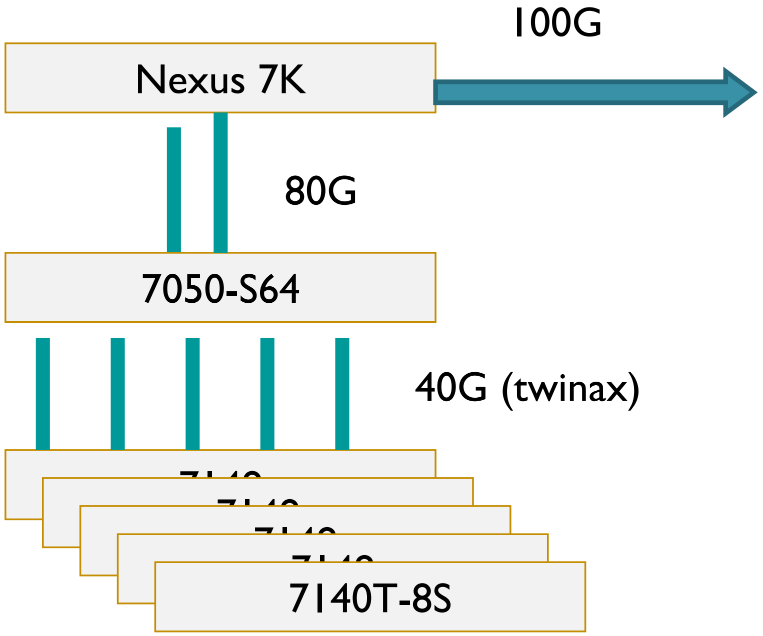
Data-Scope Specs

	Revised					
	1P	*1S*	*All P*	*All S*	*Full*	
servers	1	1	90	6	102	
rack units	4	34	360	204	564	
capacity	24	720	2160	4320	6480	TB
price	8.8	57	8.8	57	792	$K
power	1.4	10	126	60	186	kW
GPU*	1.35	0	121.5	0	122	TF
seq IO	5.3	3.8	477	23	500	GBps
IOPS	240	54	21600	324	21924	kIOPS
netwk bw	10	20	900	240	1140	Gbps

* Without the GPU costs (it is about $1,600/ card)

Network Architecture

- Arista Networks switches (10G, copper and SFP+)
- 5x 7140T-8S for the Top of the Rack (TOR) switches
  - 40 CAT6, 8 SFP+
- 7050-S64 for the core
  - 64x SFP+, 4x QSFP+ (40G)
- Fat-tree architecture
- Uplink to Cisco Nexus 7K
  - 2x100G card
  - 6x40G card



Increased Diversification

One shoe does not fit all!

- Diversity grows naturally, no matter what
- Evolutionary pressures help
- Individual groups want specializations

- Large floating point calculations move to GPUs
- Big data moves into a cloud (private or public)
- RandomIO moves to Solid State Disks
- High-Speed stream processing emerging
- noSQL vs databases vs column store vs SciDB …

At the same time

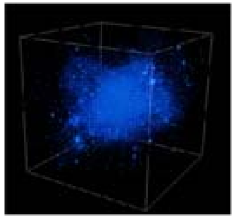
- What remains in the middle?
  - Common denominator is Big Data
- Data management
  - Everybody needs it, nobody enjoys doing it
- We are still building our own…

over and over…

The Long Tail

- The "Long Tail" of a huge number of small data sets
  - The integral of the "long tail" is big!
- Facebook: bring many small, seemingly unrelated data to a single cloud and new value emerges
  - What is the science equivalent?
- The DropBox lesson
  - Simple interfaces are much more powerful than complex ones
  - API public

Sociology



- Broad sociological changes
  - Convergence of Physical and Life Sciences
  - Data collection in ever larger collaborations
  - Virtual Observatories: CERN, VAO, NCBI, NEON, OOI,…
  - Analysis decoupled, off archived data by smaller groups
  - Emergence of the citizen/internet scientist
- Need to start training the next generations
  - Π-shaped vs I-shaped people
  - Early involvement in "Computational thinking"

Summary

- Science is increasingly driven by data (large and small)
- Large data sets are here, COTS solutions are not
- Analyzing large data requires a different approach
- We need new instruments: “microscopes & telescopes” for data
- Changing sociology
- From hypothesis-driven to data-driven science
- Same problems present in HPC data
- A new, Fourth Paradigm of Science is emerging…