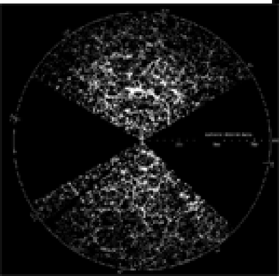
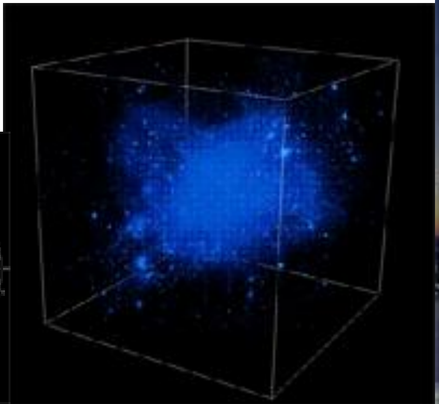
Institute for Data Intensive Engineering and Science

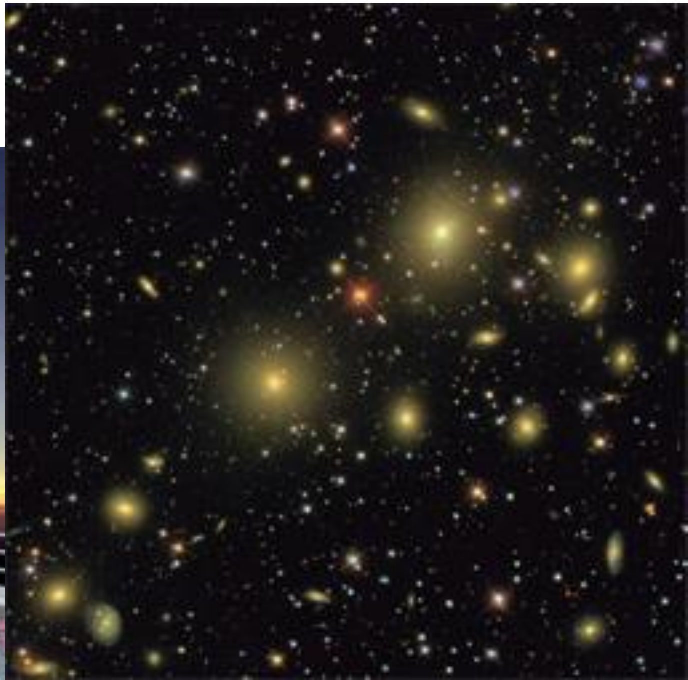
idies

Data-Intensive Computing in Astrophysics

Alex Szalay
JHU

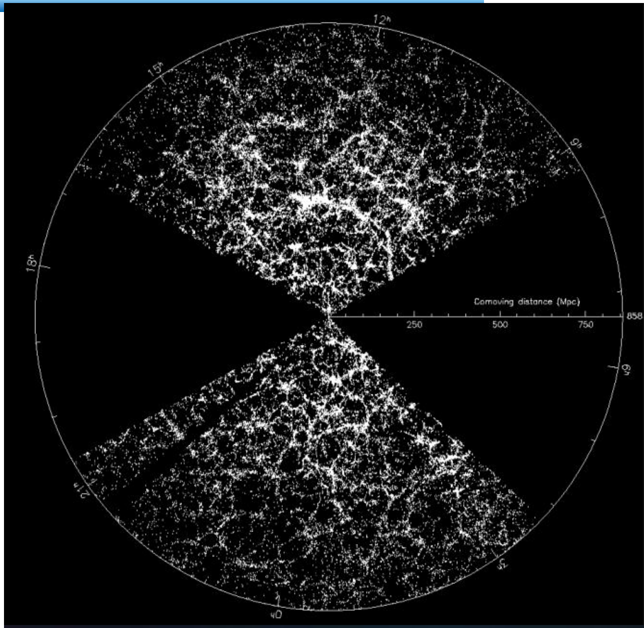






Sloan Digital Sky Survey

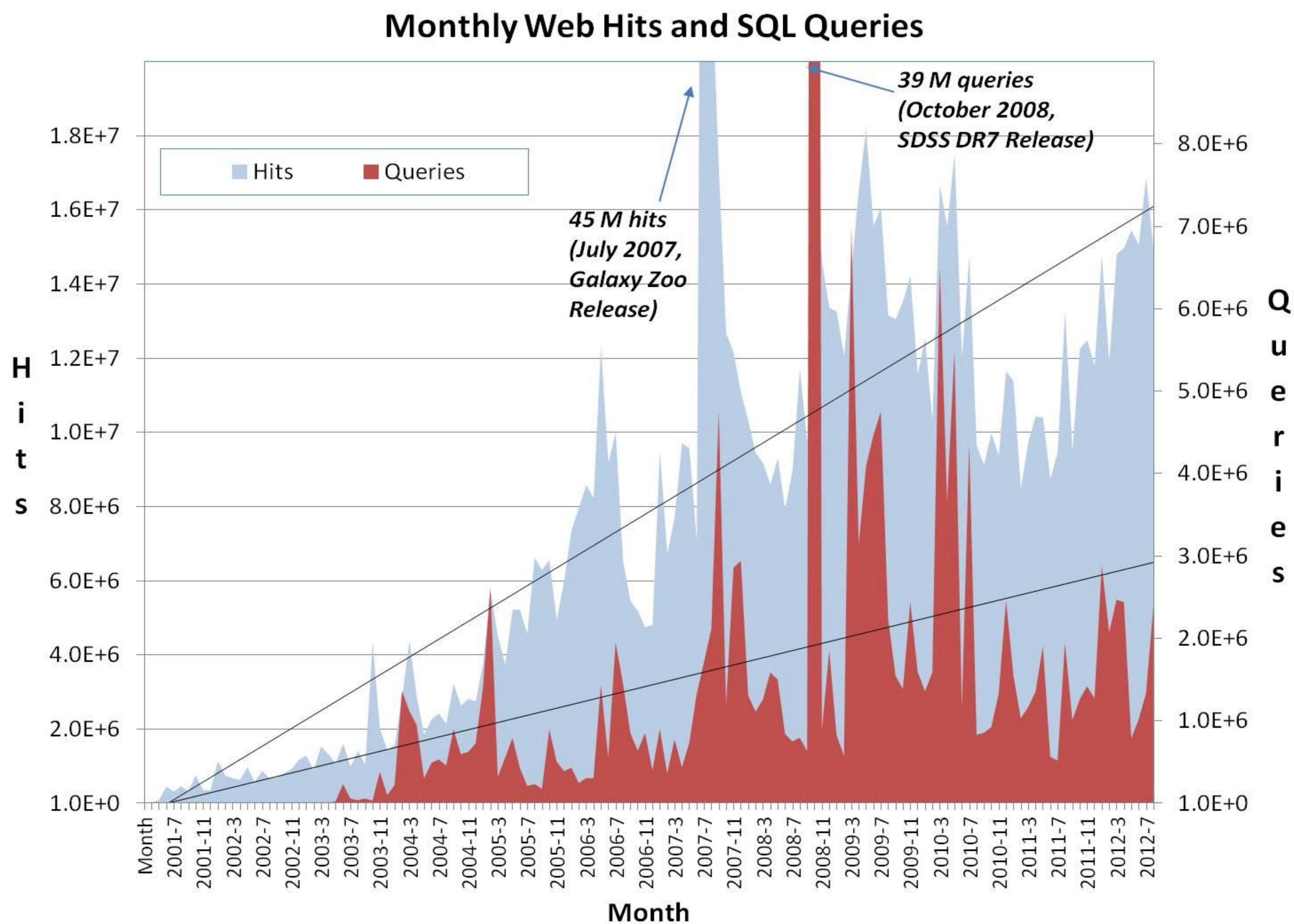
- **“The Cosmic Genome Project”**
- Two surveys in one
  - Photometric survey in 5 bands
  - Spectroscopic redshift survey
- Data is public
  - 2.5 Terapixels of images => 5 Tpx
  - 10 TB of raw data => 120TB processed
  - 0.5 TB catalogs => 35TB in the end
- Started in 1992, finished in 2008
- Database and spectrograph
  built at JHU (SkyServer)



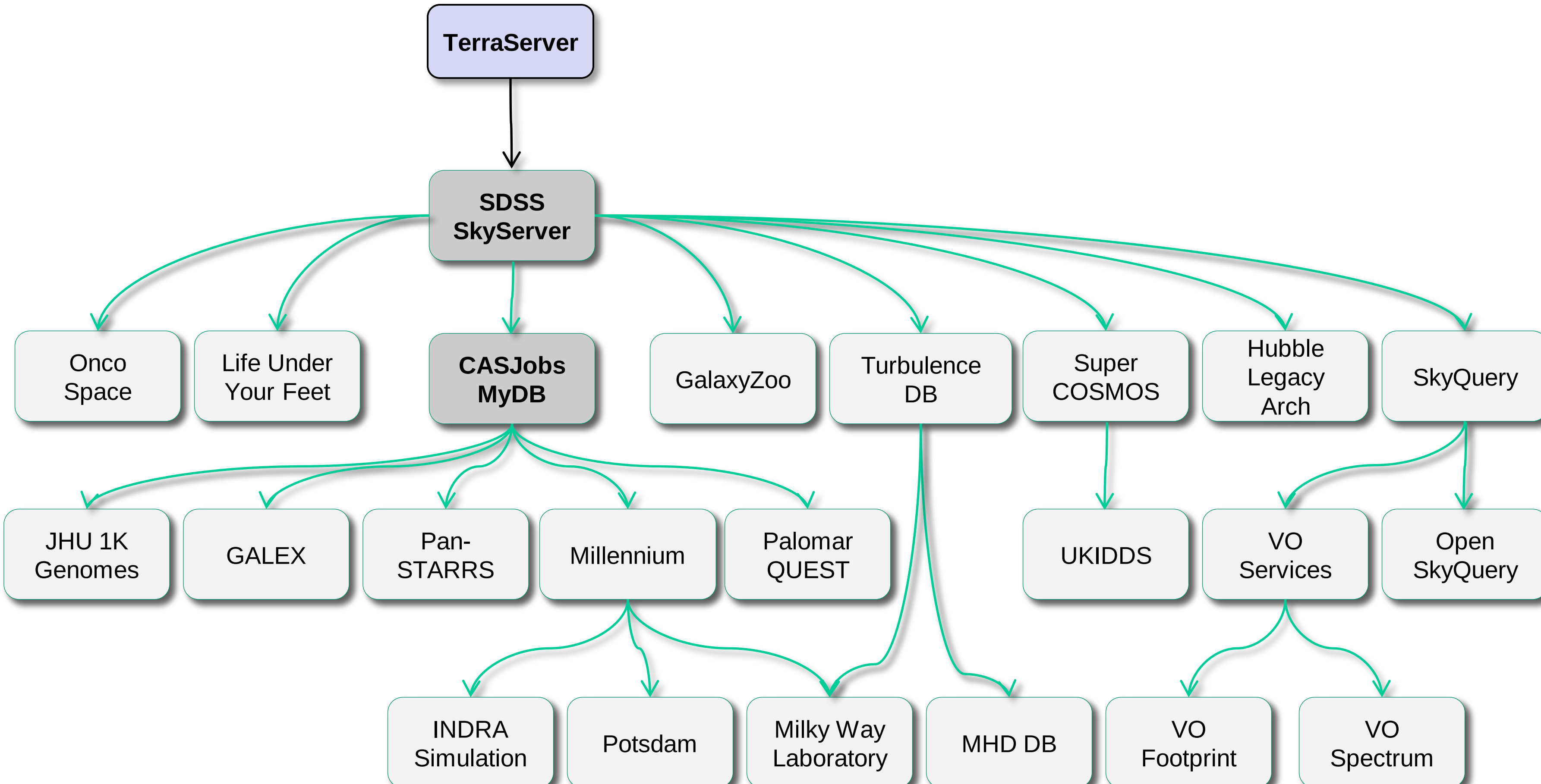
Skyserver

- Prototype in 21st Century data access
  - *1 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 by 5K astronomers (30%)*
- GalaxyZoo (Lintott et al)
  - *40 million visual galaxy classifications by the public*
  - *Enormous publicity (CNN, Times, Washington Post, BBC)*
  - *300,000 people participating, blogs, poems…*
  - *Original discoveries by the public (Voorwerp, Green Peas)*

Monthly Web Hits and SQL Queries



The SDSS Genealogy

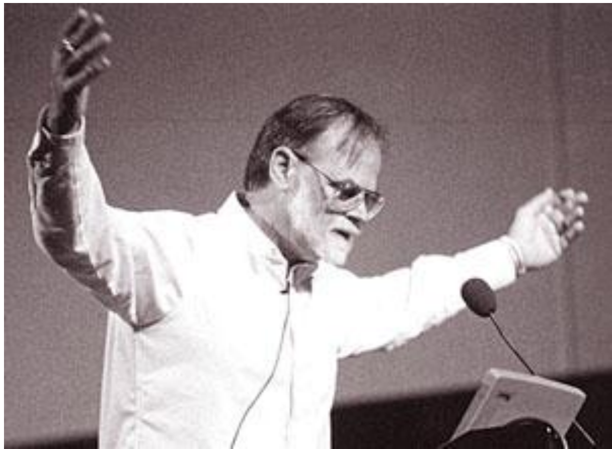


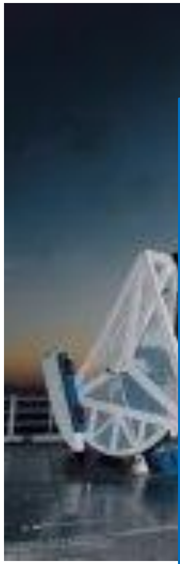
Why Is Astronomy Interesting?

- Approach inherently and traditionally data-driven
  - *Cannot do experiments…*
- Important spatio-temporal features
- Very large density contrasts in populations
- Real errors and covariances
- Many signals very subtle, buried in systematics
- Data sets large, pushing scalability
  - *LSST will be 100PB*

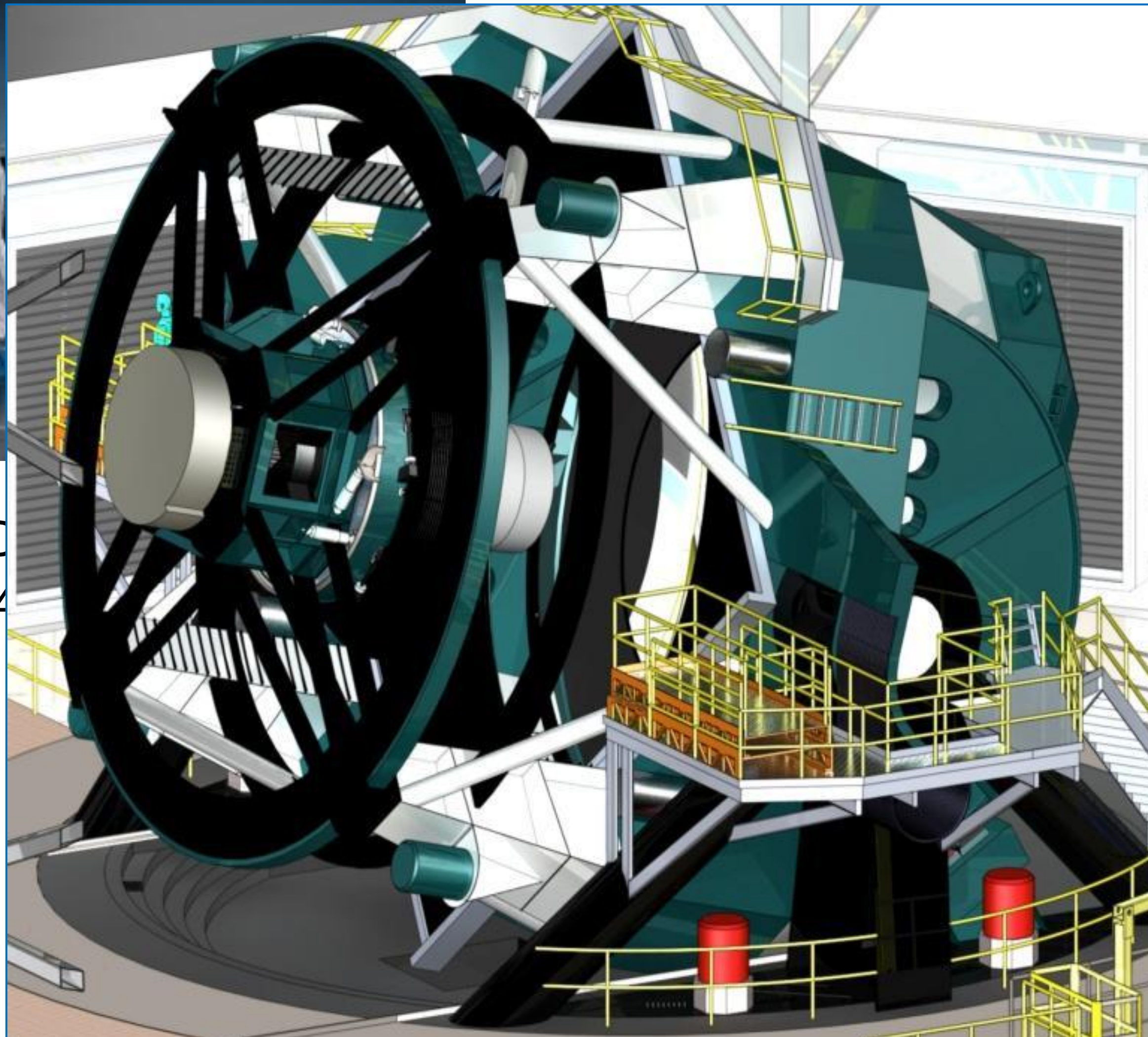
"Exciting, since it is ***worthless****!"*

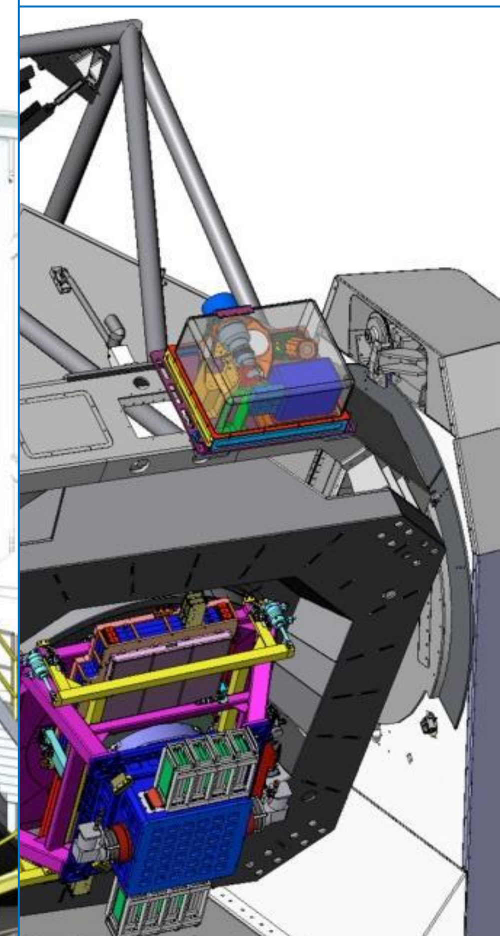
— Jim Gray





SD
2.4

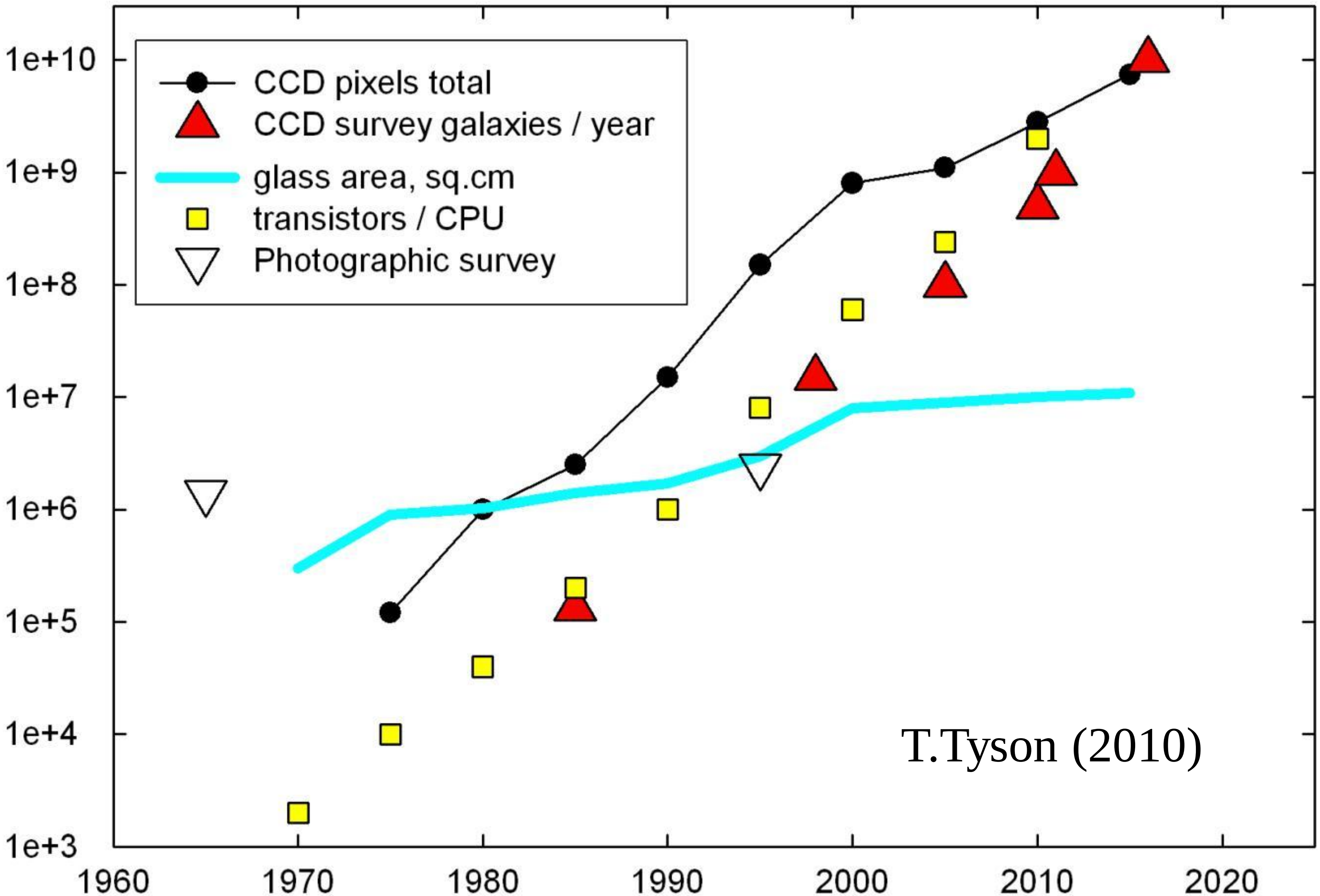




LSST
8.4m 3.2Gpixel

PanSTARRS
1.8m 1.4Gpixel

Survey Trends



Virtual Observatory

- Federate all astronomy data in the world into one system
- Most challenges are sociological, not technical
- Trust: scientists want trustworthy, calibrated data with occasional access to low-level raw data
- Career rewards for young people still not there
- Threshold for publishing data is still too high
- Robust applications are hard to build (factor of 3…)
- Archives (and data) on all scales, all over the world
- *Astronomy has successfully passed the first hurdles… but it is a long journey… no instant gratification*

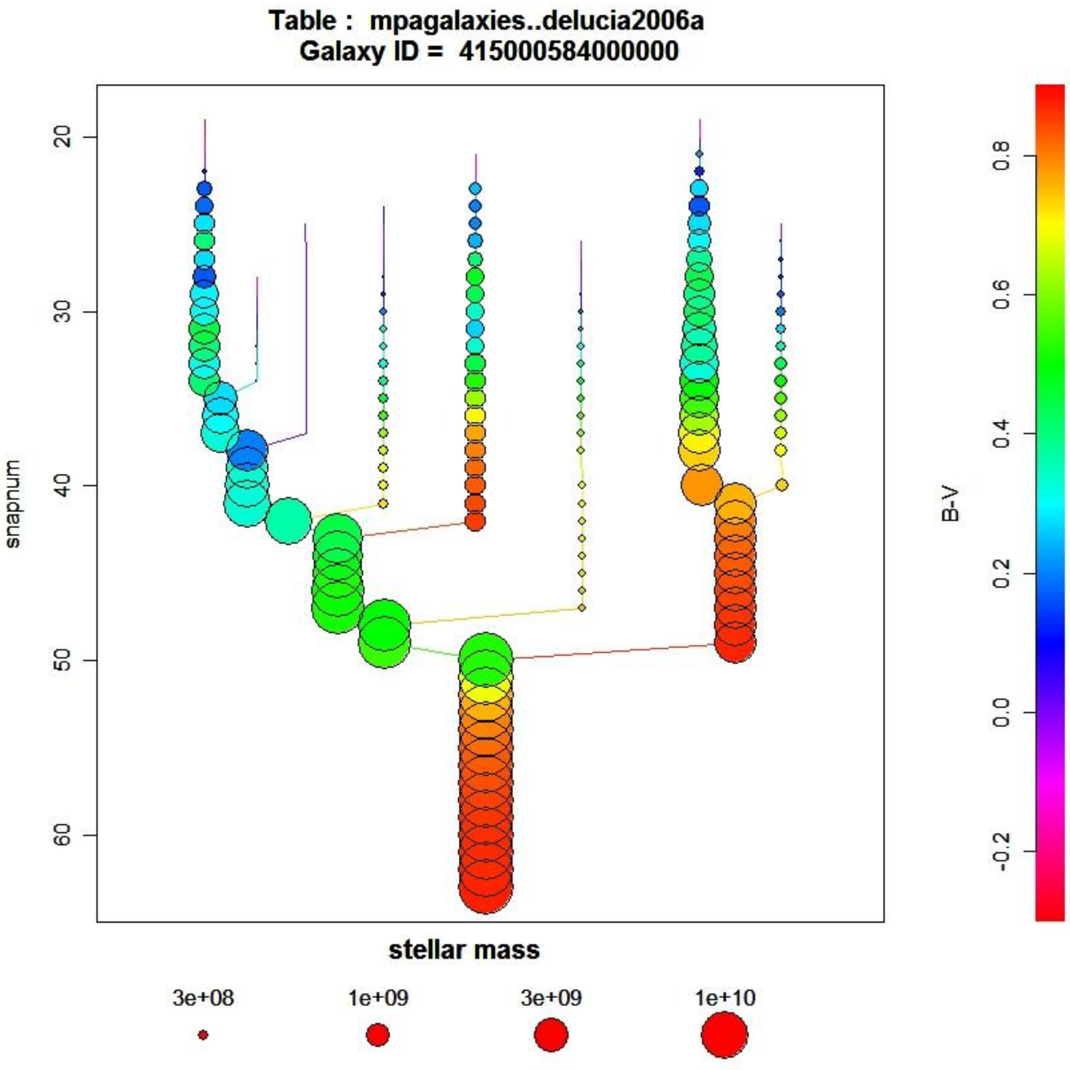
Data in HPC Simulations

- HPC is an instrument in its own right
- Largest simulations approach petabytes
  - *from supernovae to turbulence, biology and brain modeling*
- Pressure for public access to the best and latest through interactive numerical laboratories
- Creates new challenges in
  - *How to move the petabytes of data (high speed networking)*
  - *How to look at it (render on top of the data, drive remotely)*
  - *How to interface (smart sensors, immersive analysis)*
  - *How to analyze (value added services, analytics, …)*
  - *Architectures (supercomputers, DB servers, ??)*

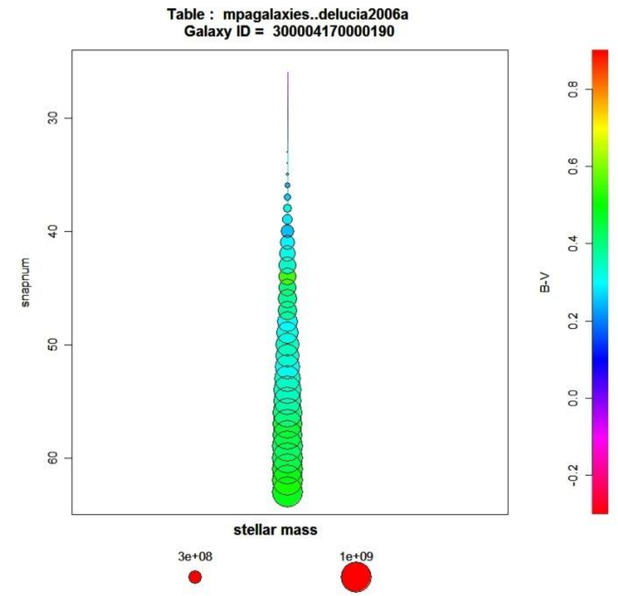
Cosmology Simulations

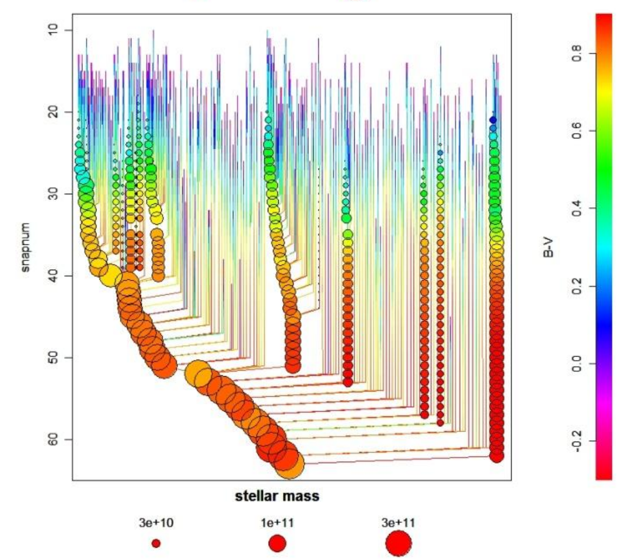
- Data size and scalability
  - *PB, trillion particles, dark matter*
  - *Where is the data located, how does it get there*
- Value added services
  - *Localized (SED, SAM, star formation history, resimulations)*
  - *Rendering (viz, lensing, DM annihilation, light cones)*
  - *Global analytics (FFT, correlations of subsets, covariances)*
- Data representations
  - *Particles vs hydro*
  - *Particle tracking in DM data*
  - *Aggregates, summary of uncertainty quantification (UQ)*

Time evolution: merger trees

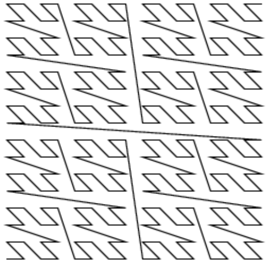


From G. Lemson

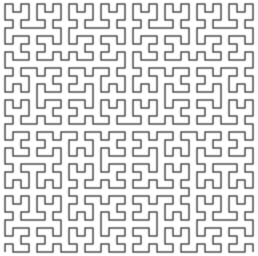




Spatial queries, random samples



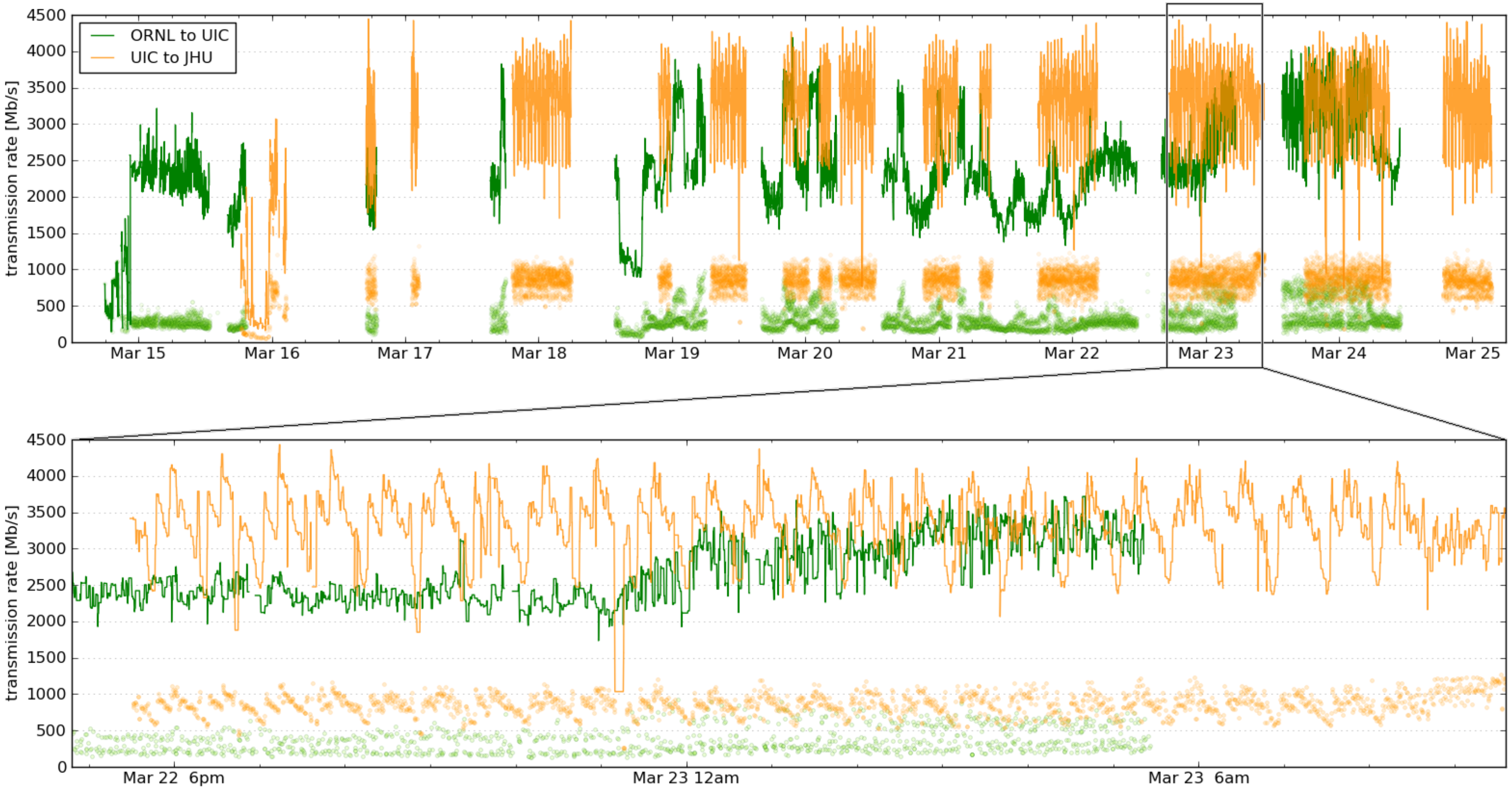
- Spatial queries require multi-dimensional indexes.
- (x,y,z) does not work: need discretisation
  - *index on (ix,iy,iz) withix=floor(x/10) etc*
- More sophisticated: space fillilng curves
  - *bit-interleaving/octtree/Z-Index*
  - *Peano-Hilbert curve*
  - *Need custom functions for range queries*
  - *Plug in modular space filling library (Budavari)*



- Random sampling using a RANDOM column
  - *RANDOM from [0,1000000]*

Silver River Transfer

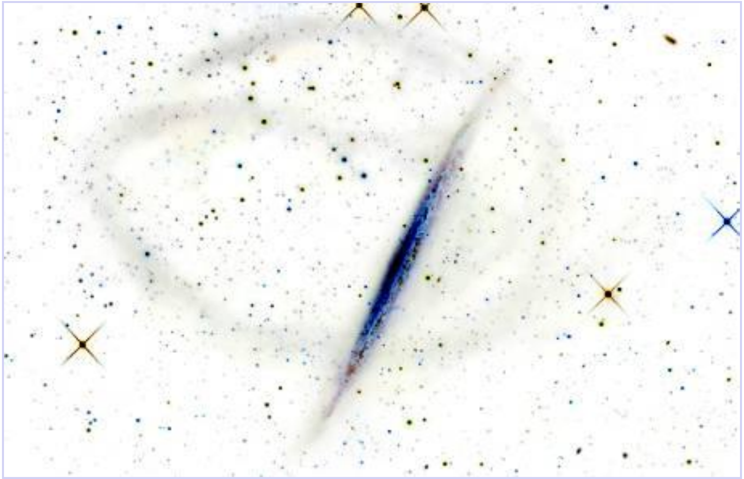
- 150TB in less than 10 days from Oak Ridge to JHU using a dedicated 10G connection



The Milky Way Laboratory

- Use cosmology simulations as an immersive laboratory for general users
- Via Lactea-II (20TB) as prototype, then Silver River (50B particles) as production (15M CPU hours)
- 800+ hi-rez snapshots (2.6PB) => 800TB in DB
- Users can insert test particles (dwarf galaxies) into system and follow trajectories in pre-computed simulation
- Users interact remotely with a PB in ‘real time’

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



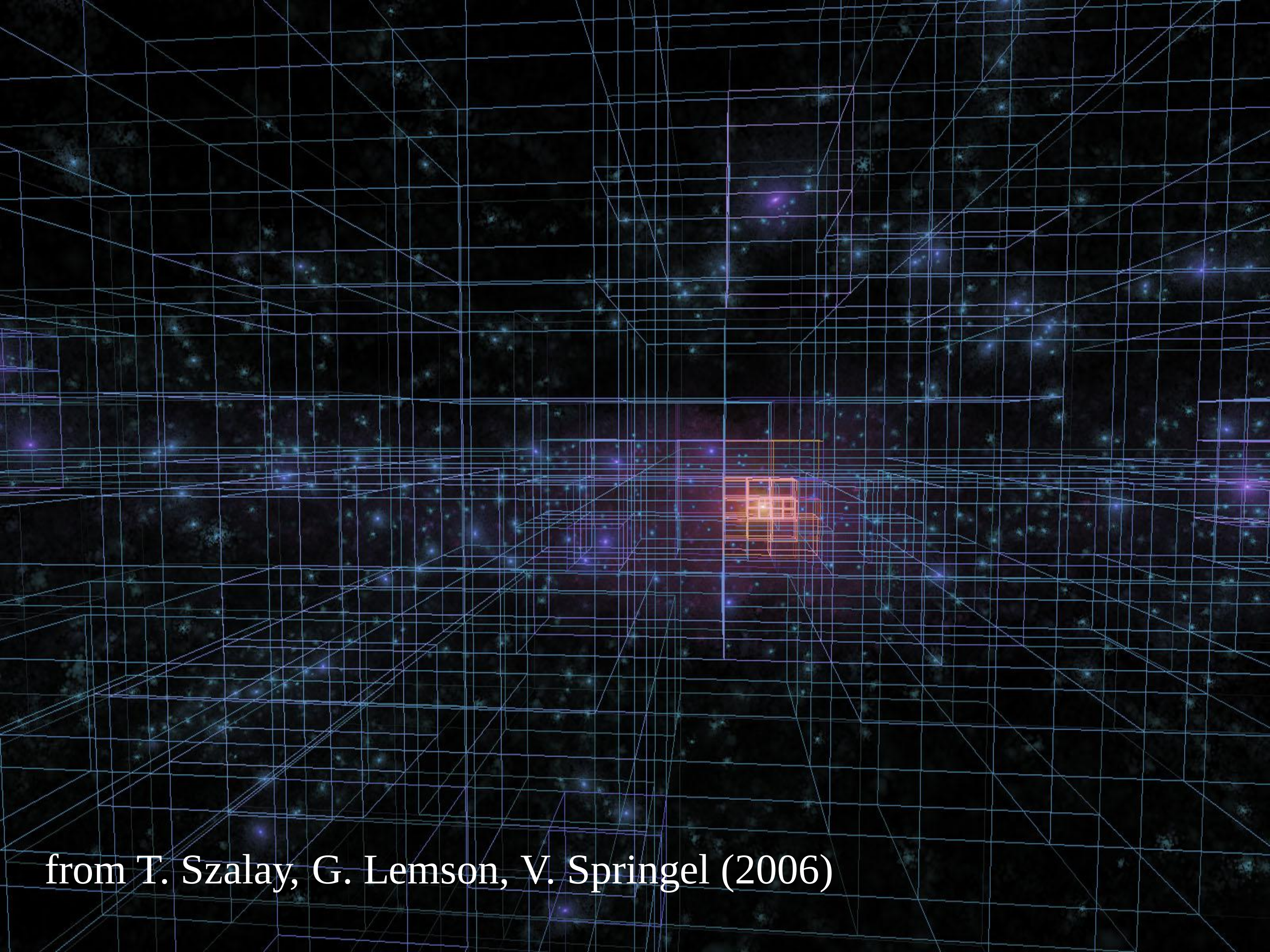
Visualizing Petabytes

- Needs to be done where the data is…
- 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: precompute octree and prefetch to SSDs
- It is possible to build individual servers with extreme data rates (5GBps per server… see Data-Scope)
- Prototype on turbulence simulation already works: data streaming directly from DB to GPU
- N-body simulations next

Real Time Interactions with TB

- Aquarius simulation (V.Springel, Heidelberg)
- 150M particles, 128 timesteps
- 20B total points, 1.4TB total
- Real-time, interactive on a single GeForce 9800
- Hierarchical merging of particles over an octree
- Trajectories computed from 3 subsequent snapshots
- Tag particles of interest interactively
- Limiting factor: disk streaming speed
- Done by an undergraduate over two months (Tamas Szalay) with Volker Springel and G. Lemson

http://arxiv.org/abs/0811.2055

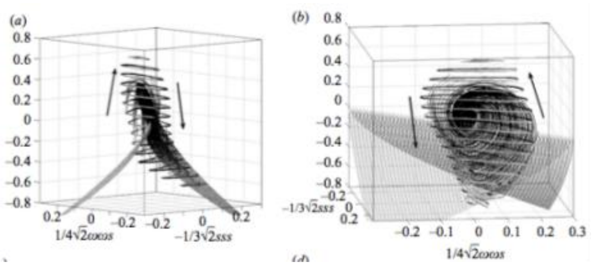


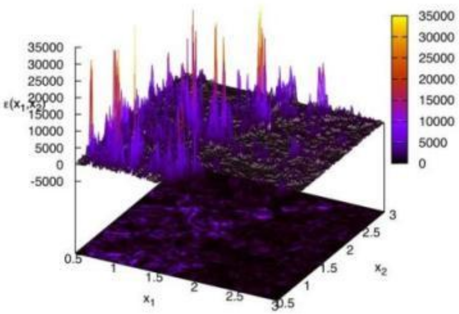
from T. Szalay, G. Lemson, V. Springel (2006)

Immersive Turbulence

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

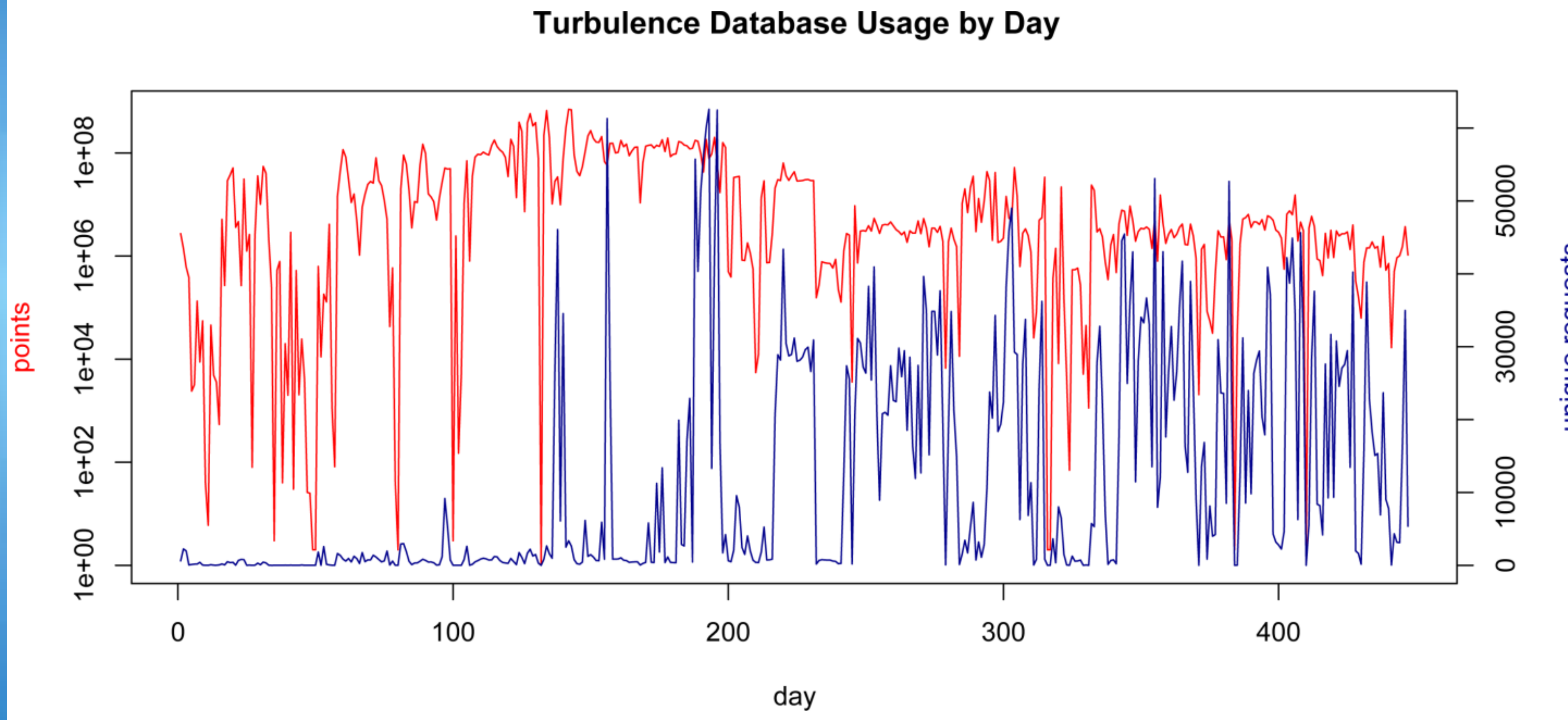
- **Understand the nature of turbulence**
  - *Consecutive snapshots of a large simulation of turbulence: now 30 Terabytes*
  - *Treat it as an experiment, **play** with the database!*
  - ***Shoot test particles** (sensors) from your laptop into the simulation, like in the movie Twister*
  - *Next: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





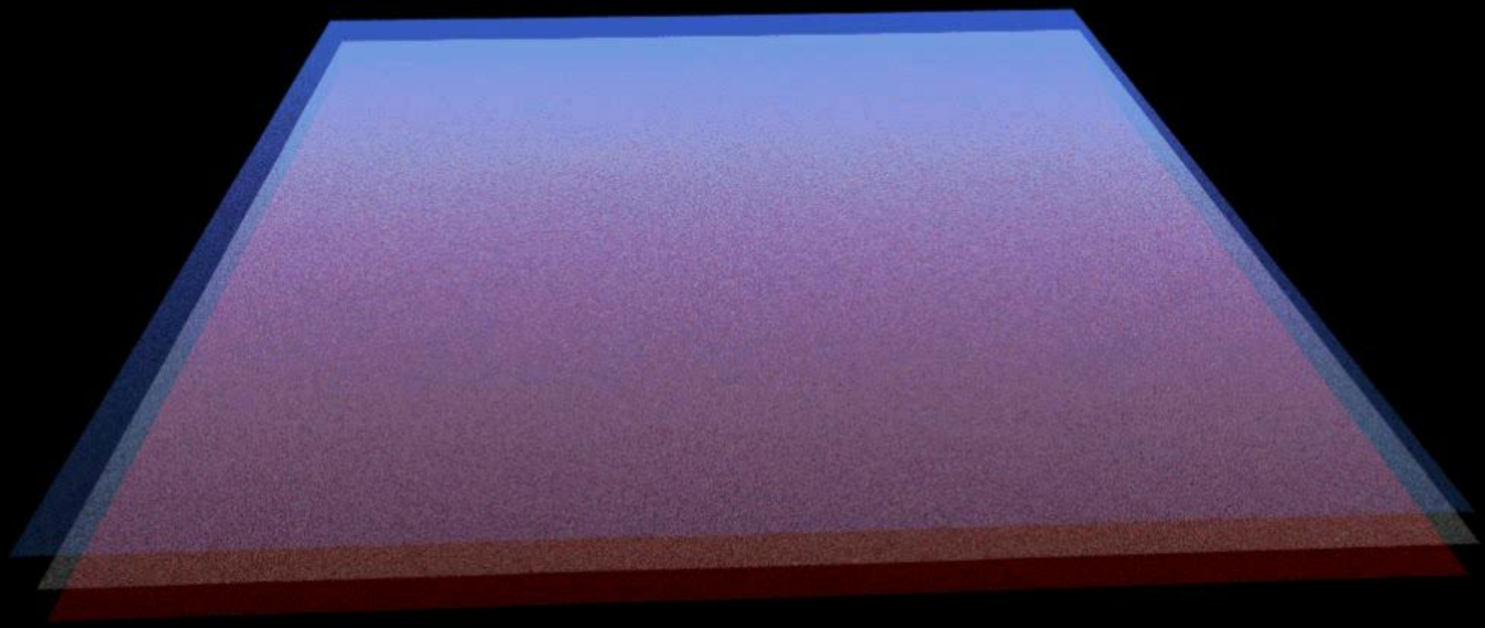
with C. Meneveau, S. Chen (Mech. E), G. Eyink (Applied Math), R. Burns (CS)

Daily Usage



2011: exceeded 100B points, delivered publicly

Streaming Visualization of Turbulence



Kai Buerger, Technische Universitat Munich, 24 million particles

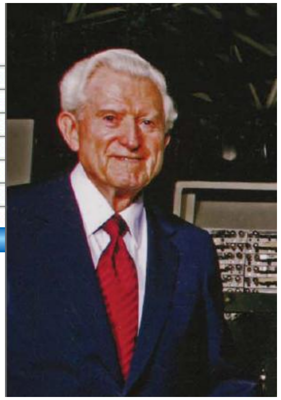
Visualization of the Vorticity

Kai Buerger, Technische Universitat Munich

Architectual Challenges

- How to build a system good for the analysis?
- Where should data be stored
  - *Not at the supercomputers (too expensive storage)*
  - *Computations and visualizations must be on top of the data*
  - *Need high bandwidth to source of data*
- Databases are a good model, but are they scalable?
  - *Google (Dremel, Tenzing, Spanner: exascale SQL)*
  - *Need to be augmented with value-added services*
- Makes no sense to build master servers, scale out
  - *Cosmology simulations are not hard to partition*
  - *Use fast, cheap storage, GPUs for some of the compute*
  - *Consider a layer of large memory systems*

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)

Table 1. Amdahl's laws applied to various system powers.

Operations per second	RAM	Disk I/O bytes/s	Disks for that bandwidth at 100 Mbytes/s/disk	Disk byte capacity (100x RAM)	Disks for that capacity at 1 Tbyte/disk
10^9	Gigabyte	10^8	1	10^{11}	1
10^{12}	Terabyte	10^{11}	1,000	10^{14}	100
10^{15}	Petabyte	10^{14}	1,000,000	10^{17}	100,000
10^{18}	Exabyte	10^{17}	1,000,000,000	10^{20}	100,000,000

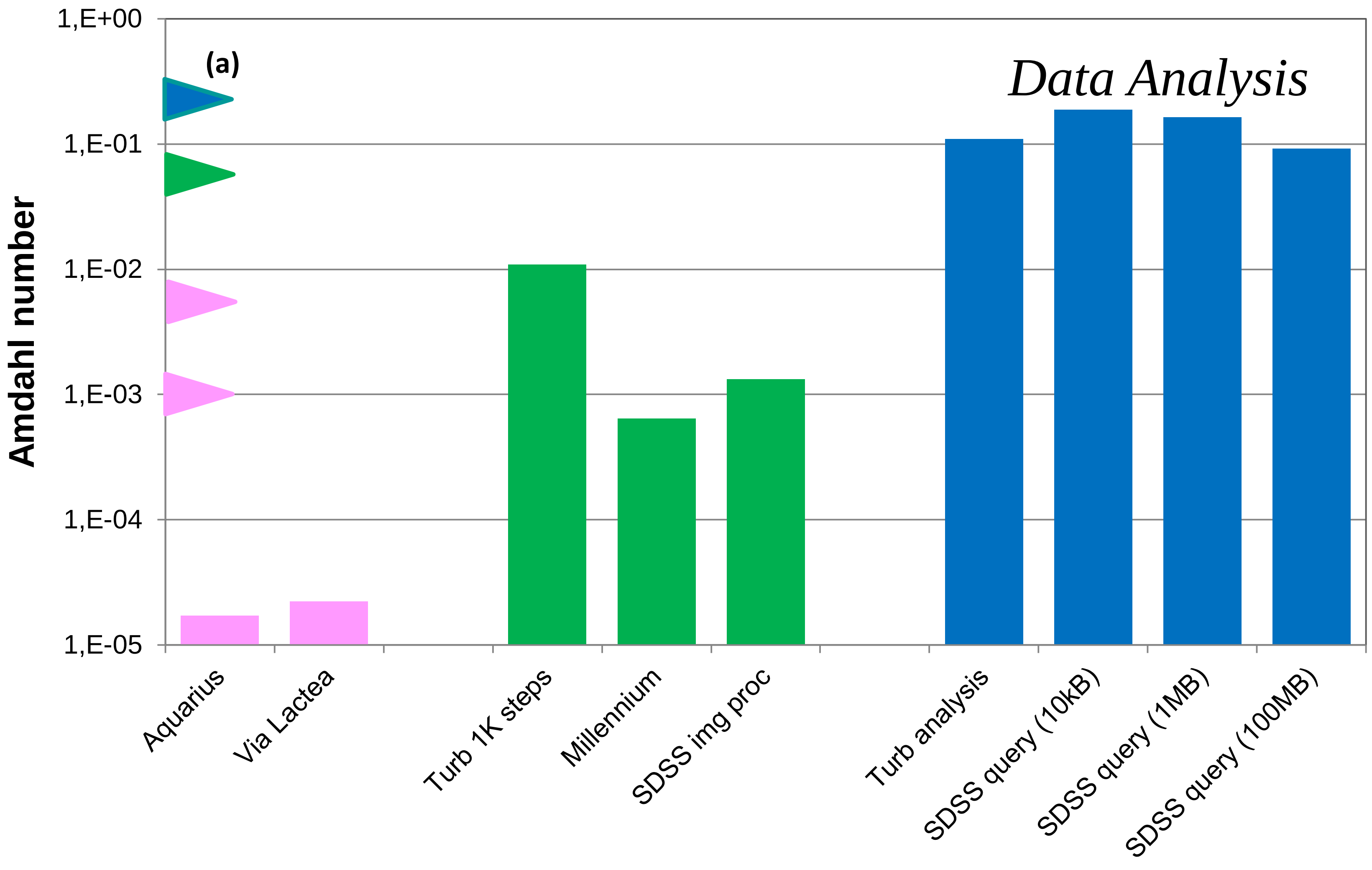
Modern multi-core systems move farther away from Amdahl's Laws (Bell, Gray and Szalay 2006)

Typical Amdahl Numbers

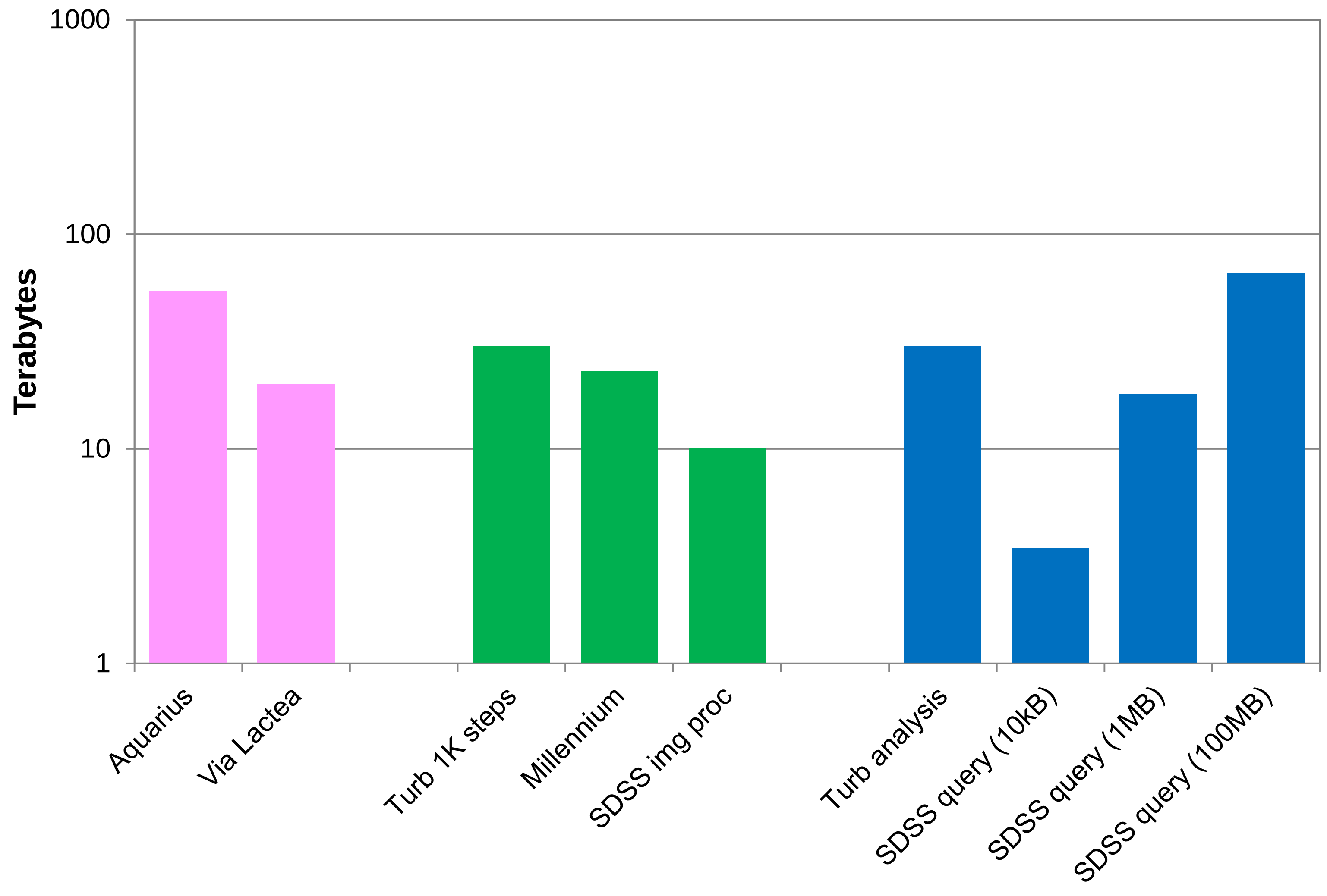
System	CPU count	GIPS [GHz]	RAM [GB]	diskIO [MB/s]	Amdahl	
					RAM	IO
BeoWulf	100	300	200	3000	0.67	0.08
Desktop	2	6	4	150	0.67	0.2
Cloud VM	1	3	4	30	1.33	0.08
SC1	212992	150000	18600	16900	0.12	0.001
SC2	2090	5000	8260	4700	1.65	0.008
GrayWulf	416	1107	1152	70000	1.04	0.506

BGAS	**17**	**1.6**	**16**	**2000**	**0.59**	**0.590**

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 months*
- 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: Extreme computing is about tradeoffs

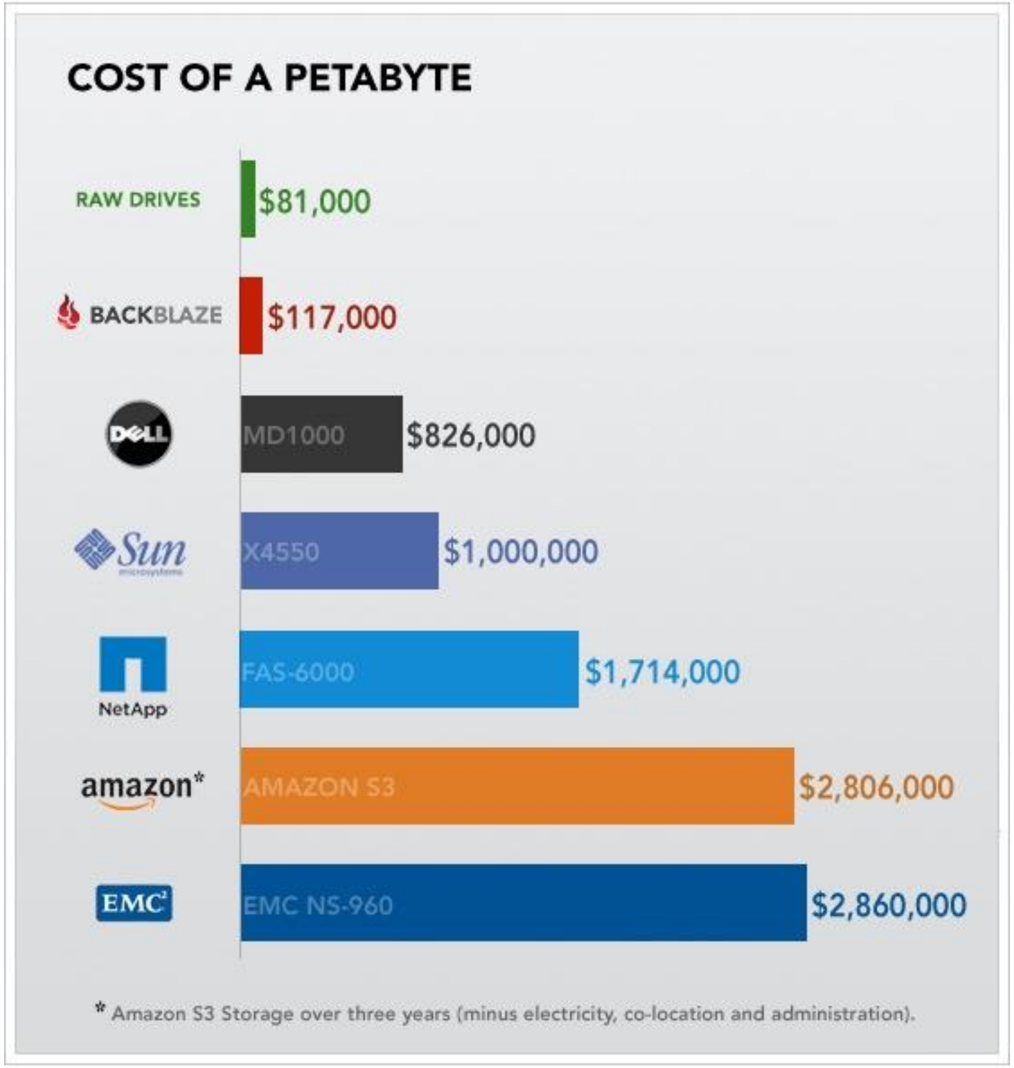
Ordered priorities for data-intensive scientific computing

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)*

The order will be different every year…

Challenges today: disk space, then memory

Cost of a Petabyte



From backblaze.com
Aug 2009

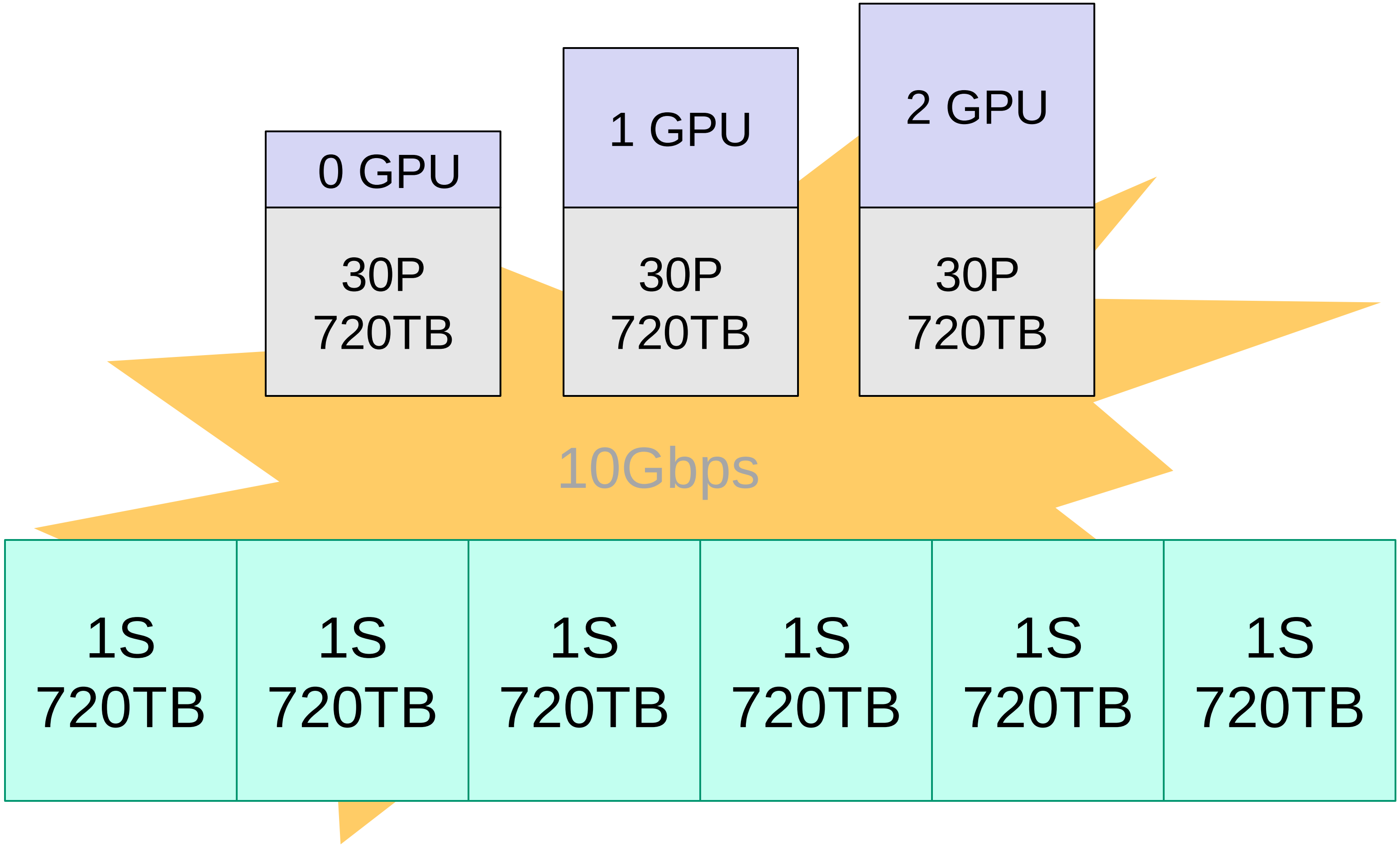
JHU Data-Scope

- Funded by NSF MRI to build a new ‘instrument’ to look at data
- Goal: ~100 servers for $1M + about $200K switches+racks
- Two-tier: performance (P) and storage (S)
- Mix of regular HDD and SSDs
- Large (6.5PB) + cheap + fast (500 GBps), but …
. ..a special purpose instrument

Amdahl Number
1.38

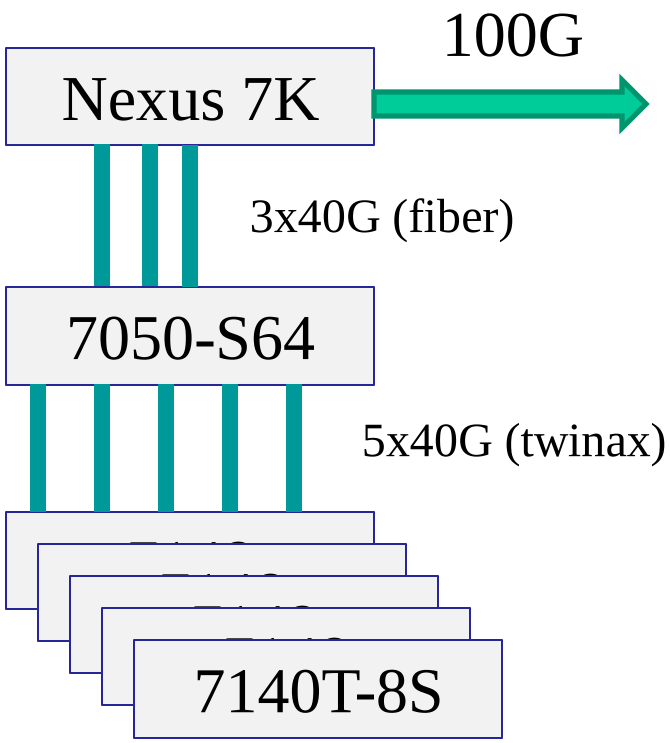
	Final configuration					
	1P	*1S*	*All P*	*All S*	*Full*	
servers	1	1	90	6	102	
rack units	4	34	360	204	564	
capacity	24	720	2160	4320	6480	TB
price	8.8	57	8.8	57	792	$K
power	1.4	10	126	60	186	kW
GPU*	1.35	0	121.5	0	122	TF
seq IO	5.3	3.8	477	23	500	GBps
IOPS	240	54	21600	324	21924	kIOPS
netwk bw	10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

Cluster Layout



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*



How to Use the Data-Scope?

- Write short proposal (<1 page)
- Bring your own data to JHU (10TB-1PB)
- Get a dedicated machines for a few months
- Pick your platform (Sci Linux or Windows+SQL)
- Pick special SW environment (MATLAB, etc)
- Partition the data
- Start crunching
- Take away the result or buy disks for cold storage

Crossing the PB Boundary

- Via Lactea-II (20TB) as prototype, then Silver River (50B particles) as production (15M CPU hours)
- 800+ hi-rez snapshots (2.6PB) => 800TB in DB
- Users can insert test particles (dwarf galaxies) into system and follow trajectories in pre-computed simulation
- Users interact remotely with a PB in ‘real time’

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

- INDRA (512 1Gpc box with 1G particles, 1.1PB)
  - *Bridget Falck talk*

Large Arrays in SQL Server

- Recent effort by Laszlo Dobos (w. J. Blakeley and D. Tomic)
- User defined data-type written in C++
- Arrays packed into varbinary(8000) or varbinary(max)
- Also direct translation to Matlab arrays
- Various subsets, aggregates, extractions and conversions in T-SQL (see regrid example:)

```
SELECT s.ix, DoubleArray.Avg(s.a)
INTO ##temptable
FROM DoubleArray.Split(@a,Int16Array.Vector_3(4,4,4)) s
SELECT @subsample = DoubleArray.Concat_N('##temptable')
```

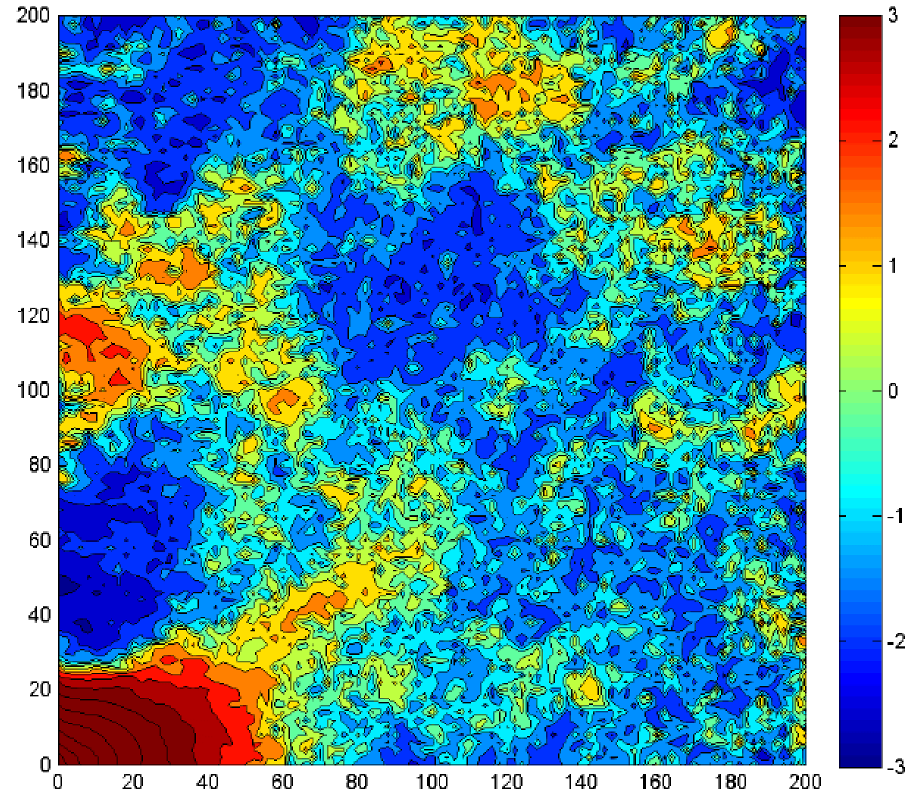
@a is an array of doubles with 3 indices
The first command averages the array over 4×4×4 blocks, returns indices and the value of the average into a table
Then we build a new (collapsed) array from its output

Using GPUs

- Massive multicore platform emerging
  - *GPGPU, multicores (Intel MIC/PHI)*
  - *Becoming mainstream (Titan, Blue Waters, Tsunami, Stampede)*
- Advantages
  - *100K+ data parallel threads*
  - *An order of magnitude lower Watts/GF*
  - *Integrated with the database (execute UDF on GPU)*
- Disadvantages
  - *Only 6GB of memory per board*
  - *PCIe limited*
  - *Special coding required*

Galaxy Correlations

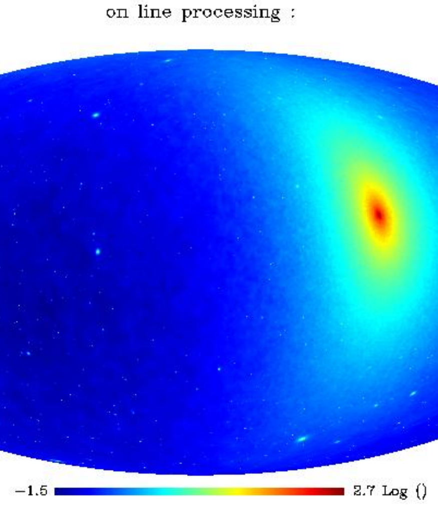
- Generated 16M random points with correct radial and angular selection for SDSS-N
- Use 500K galaxies with 3D distances from SDSS
- Originally done on an NVII
- **600 trillion** galaxy/random
- Brute force massively para faster than tree-code for hi




```
select dd.i, dd.j,  dd.cts as dd, dr.cts as dr, rr.
    (@Nrr*CONVERT(float,dd.cts)/@Ndd - 2*@Nrr*CONVE
        / CONVERT(float,rr.cts) as xi
from     dbo.PairCounts(@maxmpc, @nbin, @qryD, @nD,
    join dbo.PairCounts(@maxmpc, @nbin, @qryR, @nR,
    join dbo.PairCounts(@maxmpc, @nbin, @qryDR, @nI
```

Dark Matter Annihilation

- Data from the Via Lactea II Simulation (400M particles)
- Computing the dark matter annihilation
- Original code by M. Kuhlen runs in 8 hours for a single image
- New GPU based code runs in 24 sec, Open GL shader lang. (Lin Yang, 2nd year grad student at JHU)
- Soon: interactive service (design your own cross-section)
- Would apply very well to lensing and image generation



Summary

- Amazing progress in 7 years
- Millennium is prime showcase of how to use simulations
- Community is now using the DB as an instrument
- New challenges emerging:
  - *Petabytes of data, trillions of particles*
  - *Increasingly sophisticated value added services*
  - *Need a coherent strategy to go to the next level*
- It is not just about storage, but how to integrate access and computation
- Bridging the gap between data server and supercomputer
- Dramatically increase the use of HPC simulations

“If I had asked people what they wanted, they would have said faster horses…”

–Henry Ford

From a recent book by Eric Haseltine:
“Long Fuse and Big Bang”