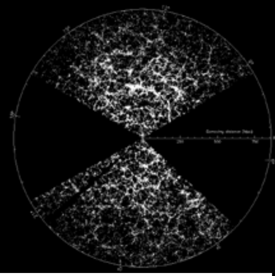
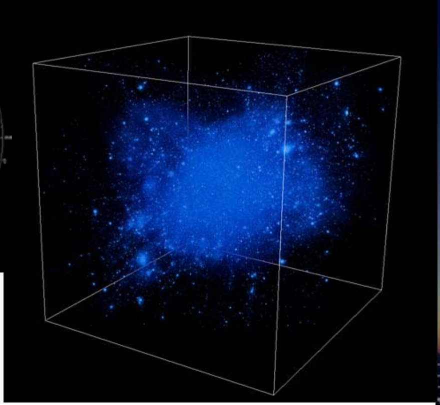
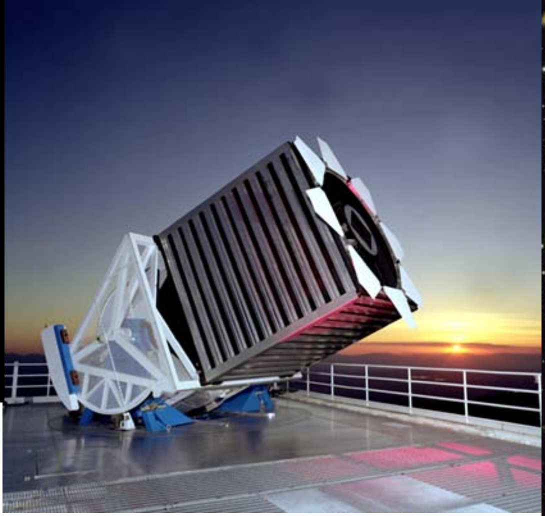




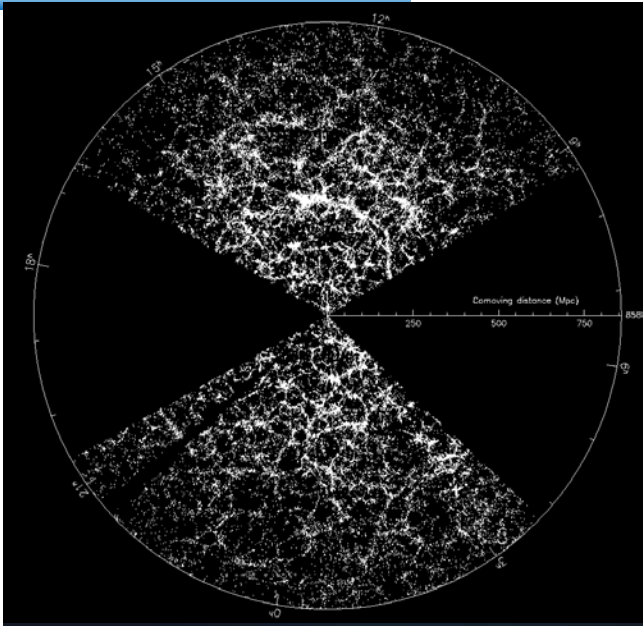


Extreme Data-Intensive Computing in Astrophysics

Alex Szalay
The Johns Hopkins University

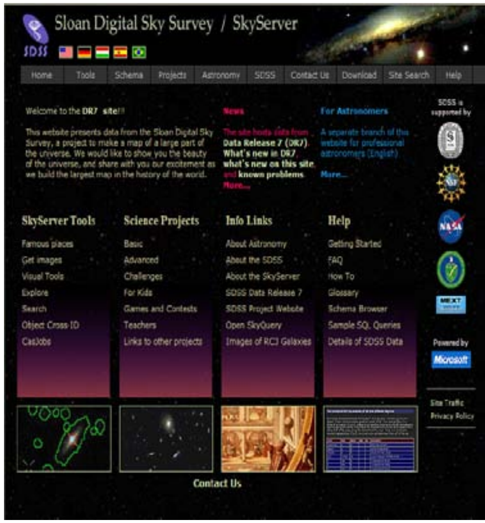
Sloan Digital Sky Survey

- **“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 et al

- Prototype in 21st Century data access
  - – *847 million web hits in 10 years*
  - – *The world's most used astronomy facility today*
  - – *1,000,000 distinct users vs. 10,000 astronomers*
  - – *The emergence of the "Internet scientist"*



Sloan Digital Sky Survey / SkyServer

Welcome to the **DR7 site**!!

This website presents data from the Sloan Digital Sky Survey, a project to make a map of a large part of the universe. We would like to show you the beauty of the universe, and share with you our excitement as we build the largest map in the history of the world.

News

The site hosts data from **Data Release 7 (DR7)**. **What's new in DR7**, **what's new on this site**, and **known problems**. More...

For Astronomers

A separate branch of this website for professional astronomers (English)

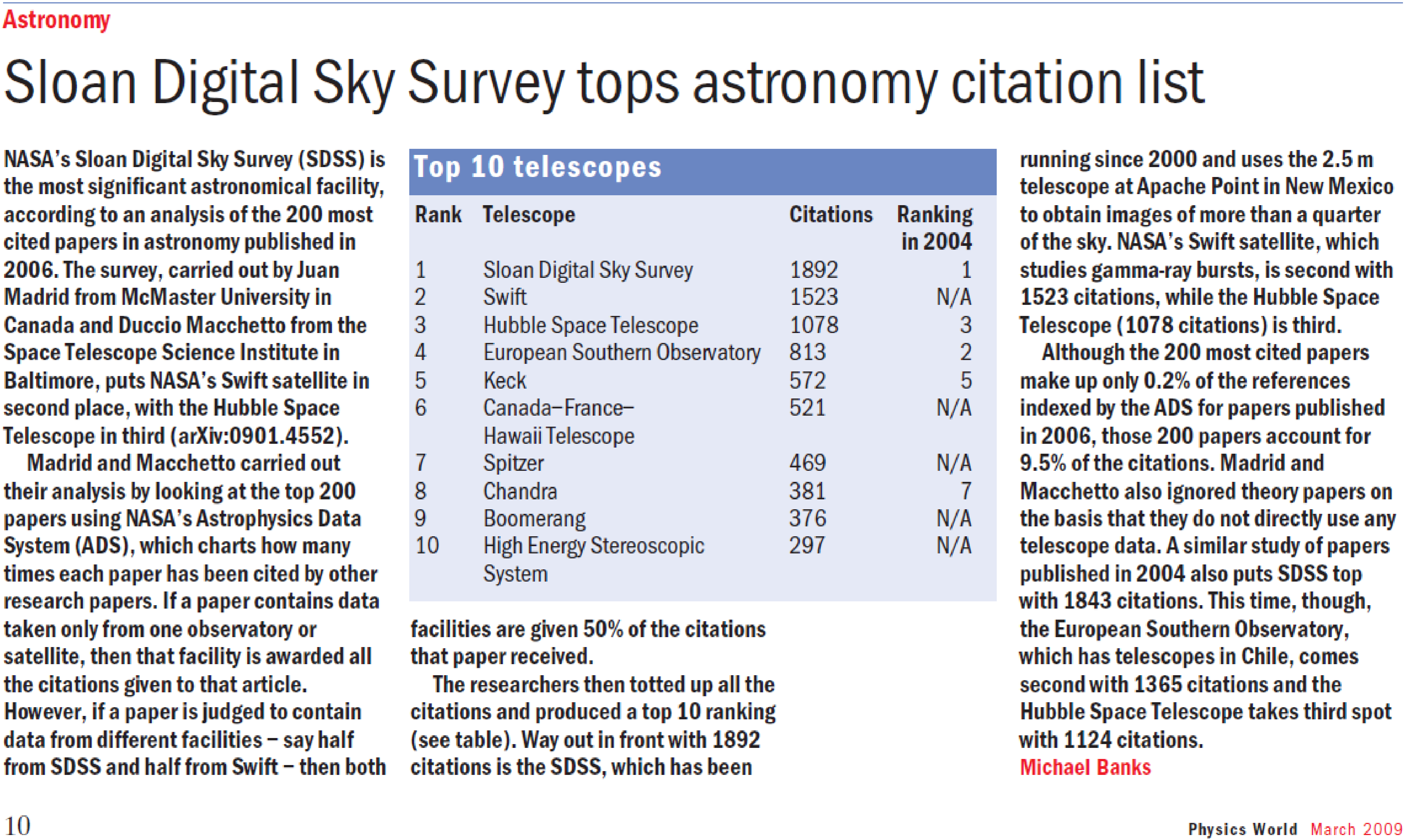
More...

SDSS is supported by

Powered by Microsoft

- 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…*
  - – *Amazing original discoveries (Voorwerp, Green Peas)*

Impact of Sky Surveys



Astronomy

Sloan Digital Sky Survey tops astronomy citation list

NASA's Sloan Digital Sky Survey (SDSS) is the most significant astronomical facility, according to an analysis of the 200 most cited papers in astronomy published in 2006. The survey, carried out by Juan Madrid from McMaster University in Canada and Duccio Macchetto from the Space Telescope Science Institute in Baltimore, puts NASA's Swift satellite in second place, with the Hubble Space Telescope in third (arXiv:0901.4552).

Madrid and Macchetto carried out their analysis by looking at the top 200 papers using NASA's Astrophysics Data System (ADS), which charts how many times each paper has been cited by other research papers. If a paper contains data taken only from one observatory or satellite, then that facility is awarded all the citations given to that article. However, if a paper is judged to contain data from different facilities – say half from SDSS and half from Swift – then both facilities are given 50% of the citations that paper received.

Top 10 telescopes

Rank	Telescope	Citations	Ranking in 2004
1	Sloan Digital Sky Survey	1892	1
2	Swift	1523	N/A
3	Hubble Space Telescope	1078	3
4	European Southern Observatory	813	2
5	Keck	572	5
6	Canada–France–Hawaii Telescope	521	N/A
7	Spitzer	469	N/A
8	Chandra	381	7
9	Boomerang	376	N/A
10	High Energy Stereoscopic System	297	N/A

The researchers then totted up all the citations and produced a top 10 ranking (see table). Way out in front with 1892 citations is the SDSS, which has been running since 2000 and uses the 2.5 m telescope at Apache Point in New Mexico to obtain images of more than a quarter of the sky. NASA's Swift satellite, which studies gamma-ray bursts, is second with 1523 citations, while the Hubble Space Telescope (1078 citations) is third.

Although the 200 most cited papers make up only 0.2% of the references indexed by the ADS for papers published in 2006, those 200 papers account for 9.5% of the citations. Madrid and Macchetto also ignored theory papers on the basis that they do not directly use any telescope data. A similar study of papers published in 2004 also puts SDSS top with 1843 citations. This time, though, the European Southern Observatory, which has telescopes in Chile, comes second with 1365 citations and the Hubble Space Telescope takes third spot with 1124 citations.

Michael Banks

Virtual Observatory

- Started with NSF ITR project, “Building the Framework for the National Virtual Observatory”, collaboration of 20 groups
  - *Astronomy data centers*
  - *National observatories*
  - *Supercomputer centers*
  - *University departments*
  - *Computer science/information technology specialists*
- Similar projects now in 15 countries world-wide

⇒ International Virtual Observatory Alliance

NSF+NASA=> VAO

VO Challenges

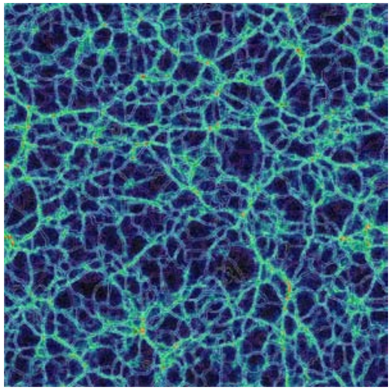
- Most challenges are sociological, not technical
- Trust: scientists want trustworthy, calibrated data with occasional access to low-level raw data
- Career rewards for young people still not there
- Threshold for publishing data is still too high
- Robust applications are hard to build (factor of 3…)
- Archives (and data) on all scales, all over the world

- Astronomy has successfully passed the first hurdles!

Continuing Growth

How long does the data growth continue?

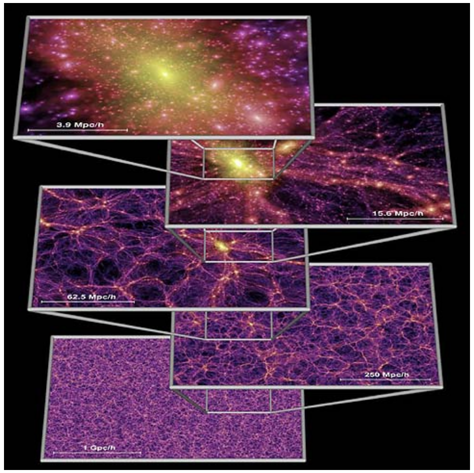
- High end always linear
- Exponential comes from technology + economics
  - *rapidly changing generations*
  - *like CCD's replacing plates, and become ever cheaper*
- How many generations of instruments are left?
- Are there new growth areas emerging?
- **Software is becoming a new kind of instrument**
  - *Value added data*
  - *Hierarchical data replication*
  - ***Large and complex simulations***



Cosmological Simulations

In 2000 cosmological simulations had 10^9 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



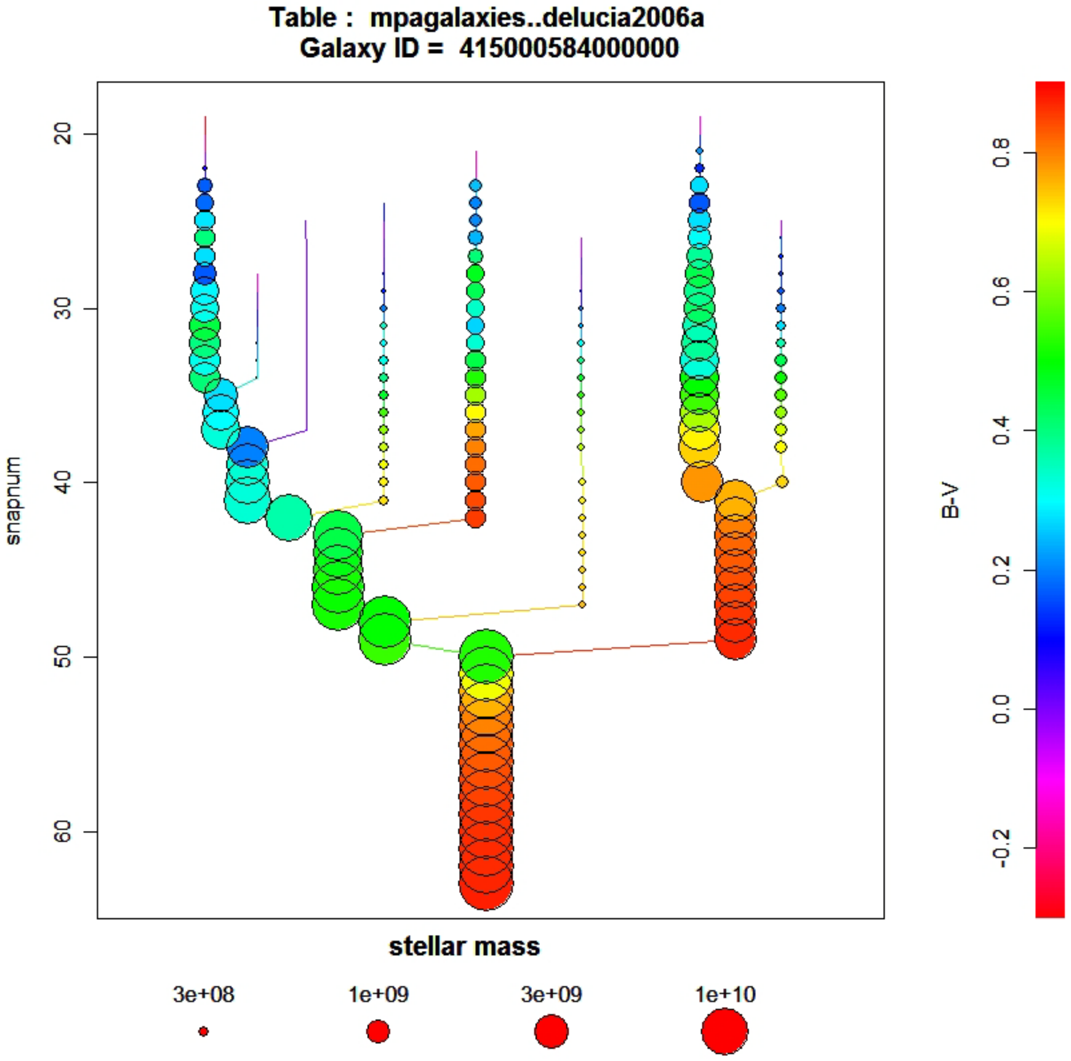
- Today: simulations with 10^{12} particles and PB of output are under way (MillenniumXXL, Exascale-Sky, etc)
- Hard to analyze the data afterwards -> need DB
- What is the best way to compare to real data?

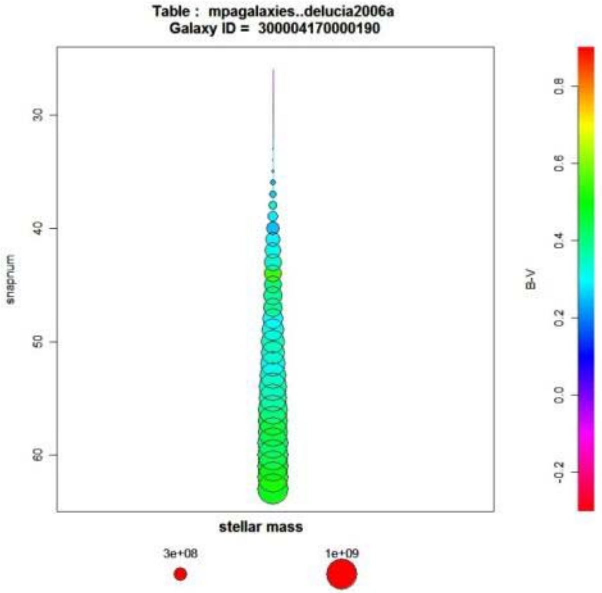
Millennium Database

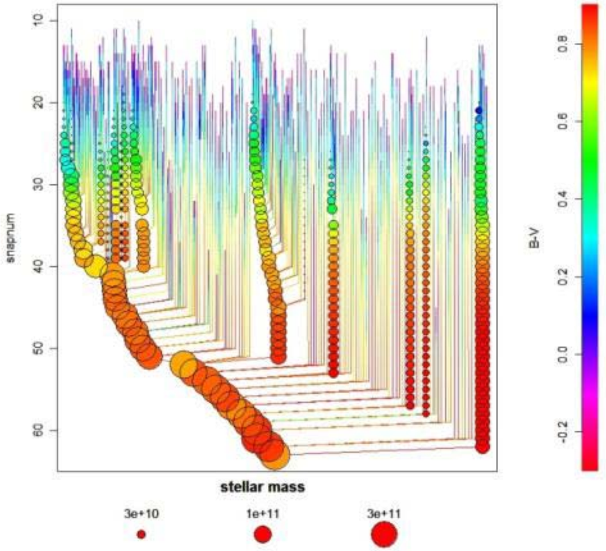
- **Density field on 256^3 mesh**
  - *CIC*
  - *Gaussian smoothed: 1.25,2.5,5,10 Mpc/h*
- Friends-of-Friends (FOF) groups
- SUBFIND Subhalos
- Galaxies from 2 semi-analytical models (SAMs)
  - *MPA (L-Galaxies, DeLucia & Blaizot, 2006)*
  - *Durham (GalForm, Bower et al, 2006)*
- Subhalo and galaxy formation histories: merger trees
- Mock catalogues on light-cone
  - *Pencil beams (Kitzbichler & White, 2006)*
  - *All-sky (depth of SDSS spectral sample)*

Gerard Lemson 2006

Time evolution: merger trees





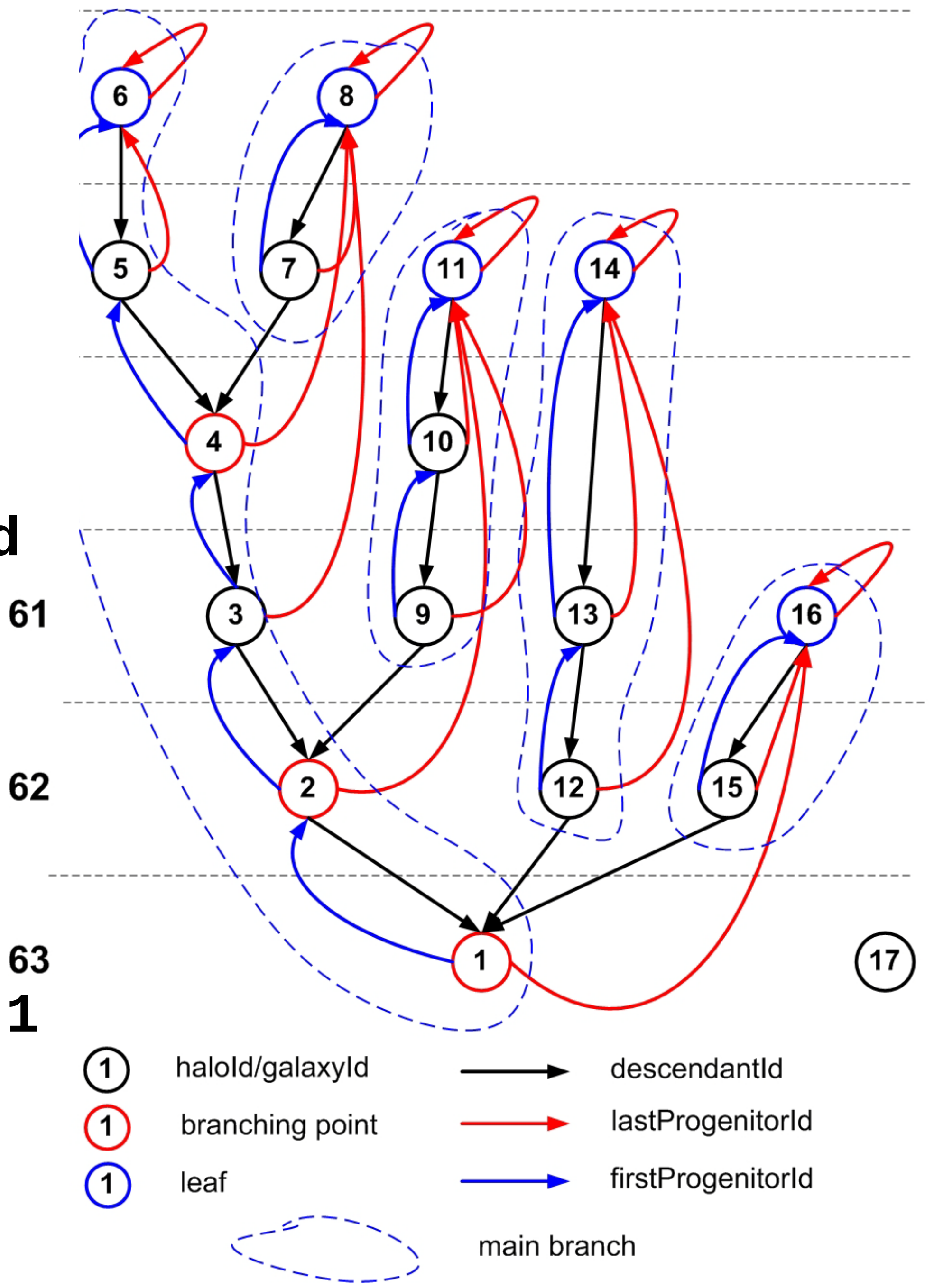


Merger trees :

```
select prog.*
  from galaxies d
  ,    galaxies p
where d.galaxyId = @id
  and p.galaxyId
  between d.galaxyId
  and d.lastProgenitorId
```

Branching points :

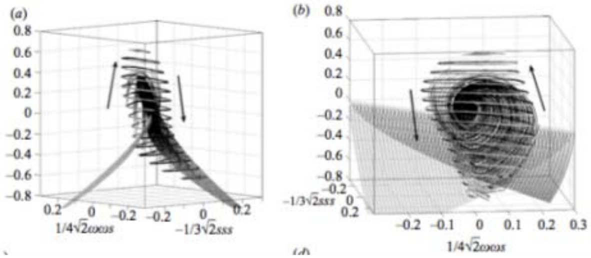
```
select descendantId
  from galaxies d
 where descendantId != -1
 group by descendantId
  having count(*) > 1
```

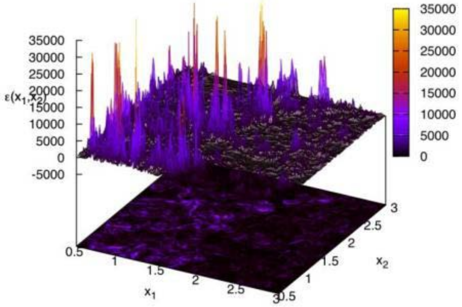


Immersive Turbulence

"… 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*



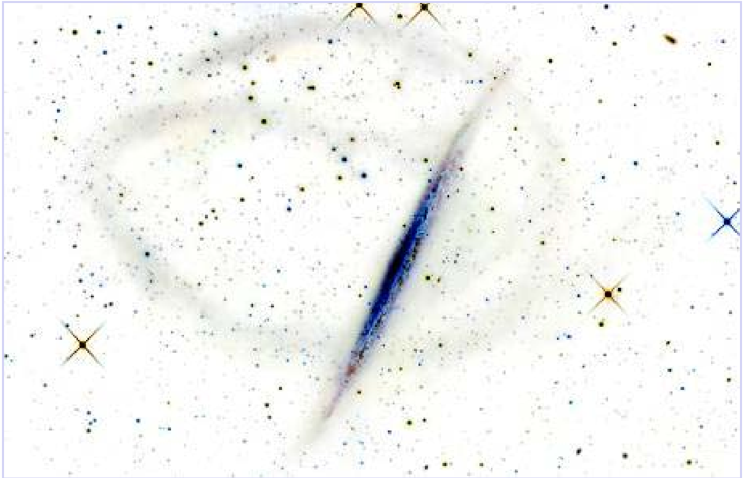


- **New paradigm** for analyzing simulations!

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

The Milky Way Laboratory

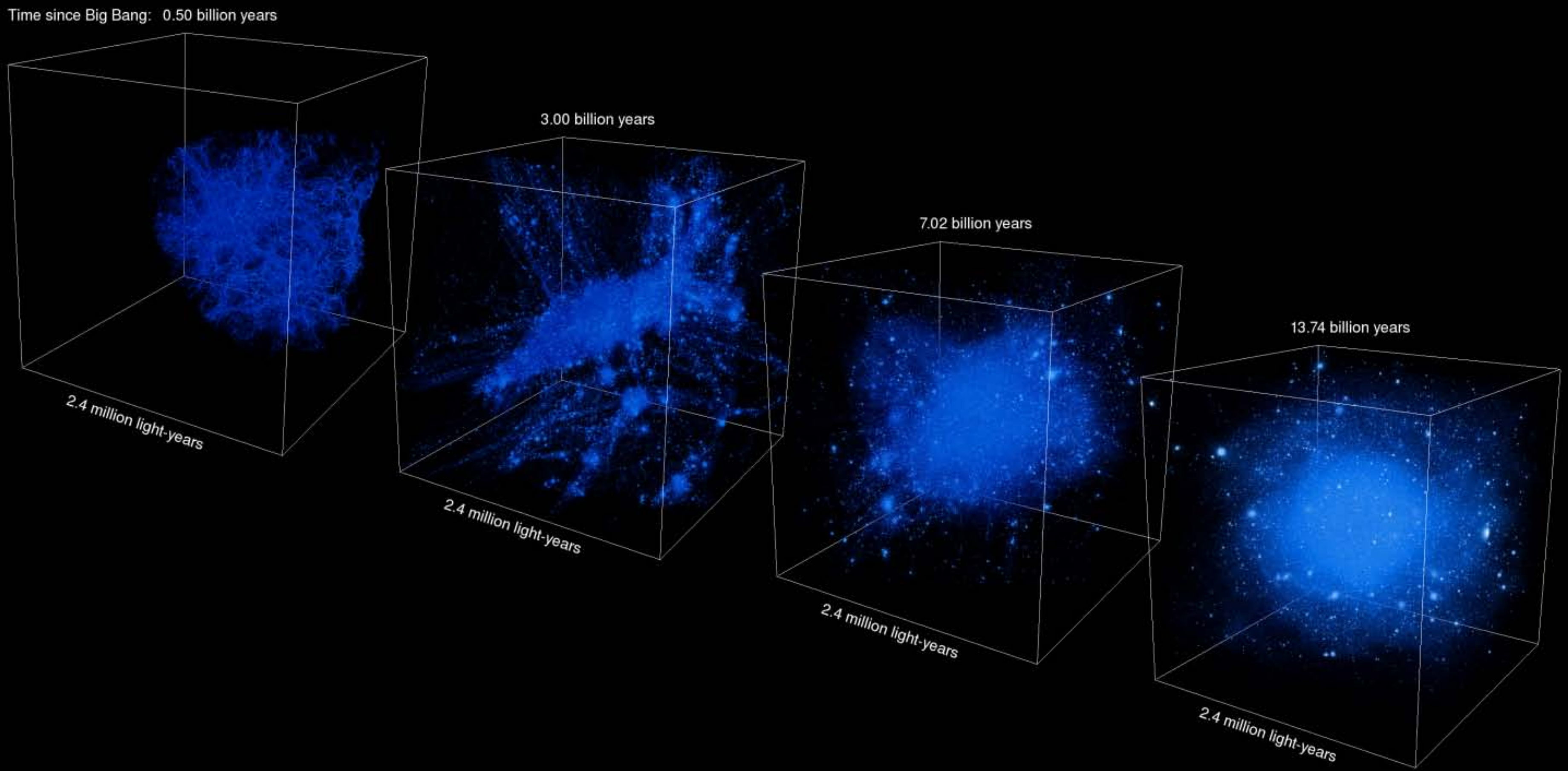
- Use cosmology simulations as an immersive laboratory for general users
- Via Lactea-II (20TB) as prototype, then Silver River (50B particles) as production (15M CPU hours)
- 800+ hi-rez snapshots (2.6PB) => 800TB in DB
- Users can insert test particles (dwarf galaxies) into system and follow trajectories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

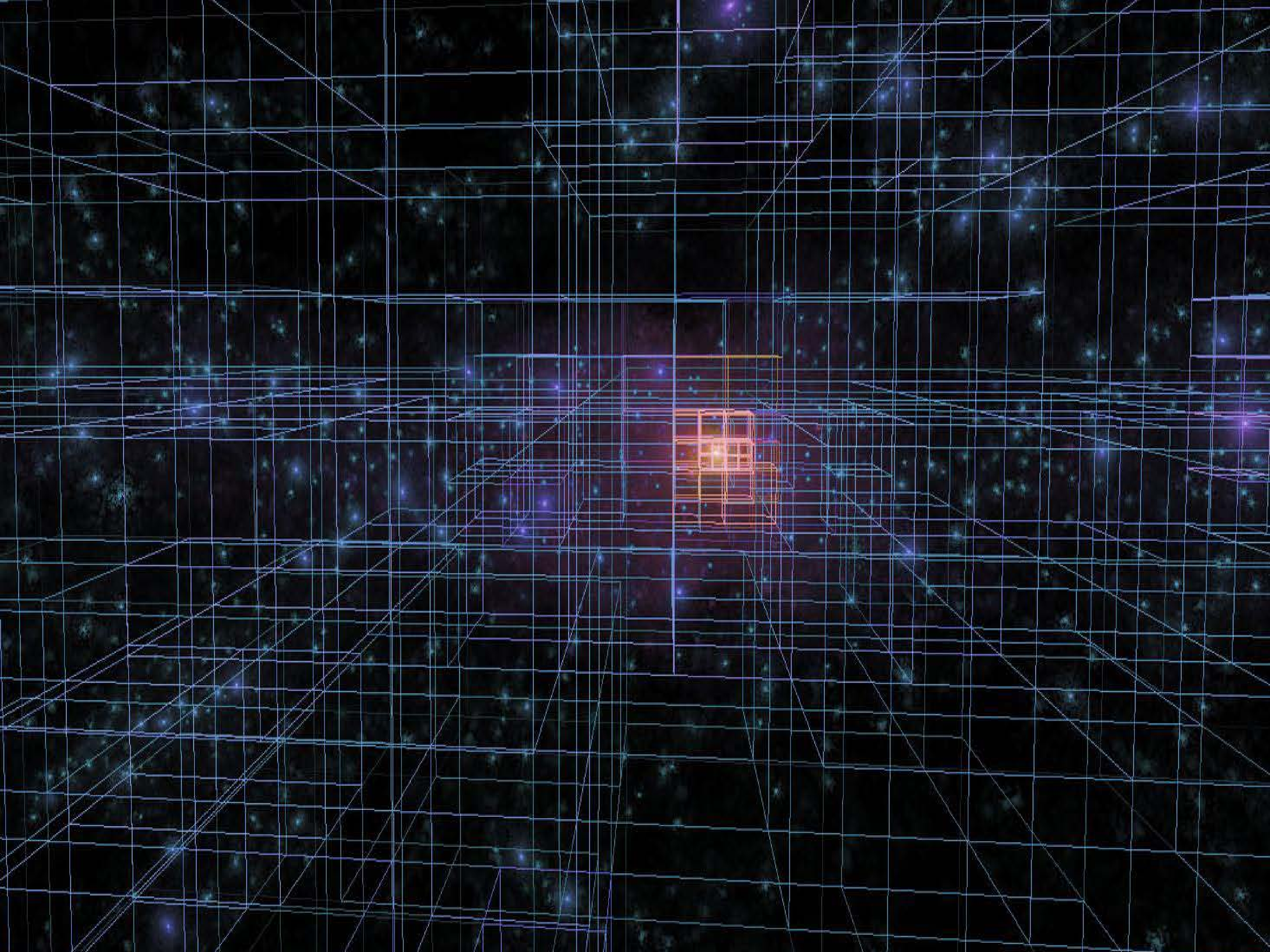


Stadel, Moore, Madau, Kuehlen
Szalay, Wyse, Silk, Lemson,
Westermann, Blakeley

Visualizing PB Simulations

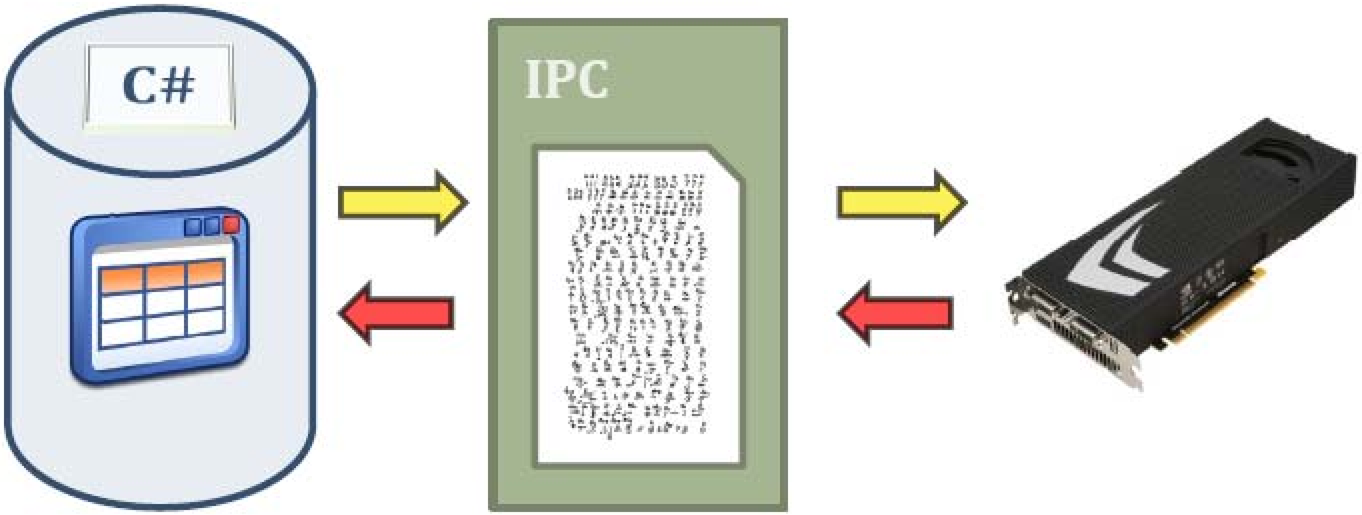
- Needs to be done where the data is…
- It is easier to send a HD 3D video stream to the user than all the data
- Interactive visualizations driven remotely
- Visualizations are becoming IO limited: precompute octree and prefetch to SSDs
- It is possible to build individual servers with extreme data rates (5GBps per server… see Data-Scope)
- Prototype on turbulence simulation already works: data streaming directly from DB to GPU
- N-body simulations next


Time since Big Bang: 0.50 billion years
2.4 million light-years
3.00 billion years
2.4 million light-years
7.02 billion years
2.4 million light-years
13.74 billion years
2.4 million light-years



Extending SQL Server

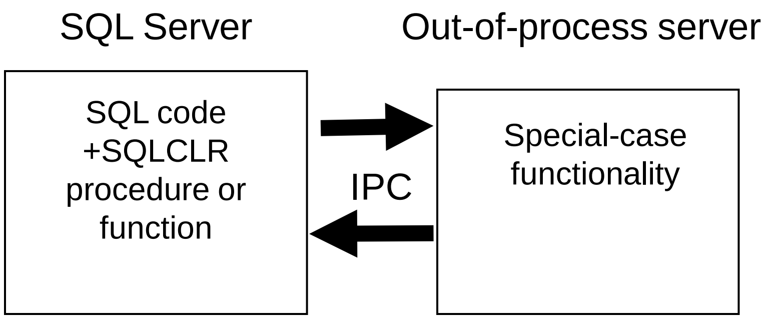
- User Defined Functions in DB execute inside CUDA
  - *100x gains in floating point heavy computations*
- Dedicated service for direct access
  - *Shared memory IPC w/ on-the-fly data transform*



Richard Wilton and Tamas Budavari (JHU)

SQLCLR Out-of Process Server

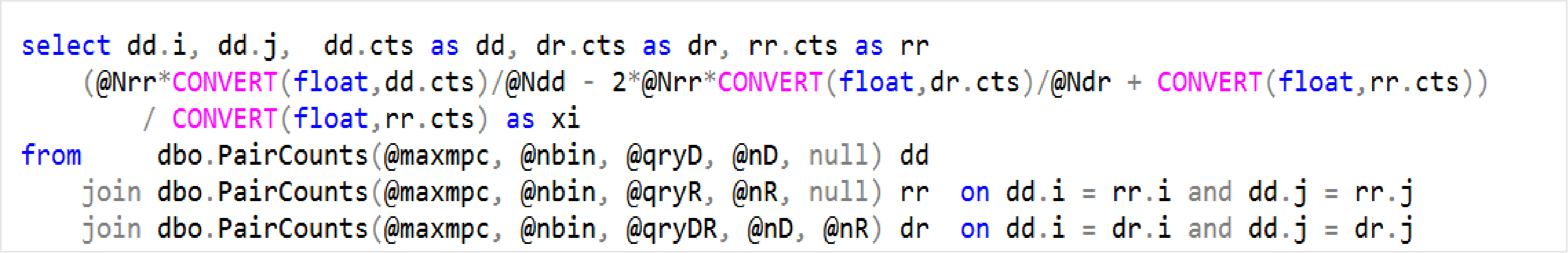
- The basic concept
  - *Implement computational functionality in a separate process from the SQL Server process*
  - *Access that functionality using IPC*
- Why
  - *Avoid memory, threading, and permissions restrictions on SQLCLR implementations*
  - *Load dynamic-link libraries*
  - *Invoke native-code methods*
  - *Exploit lower-level APIs (e.g. SqlClient, bulk insert) to move data between SQL Server and CUDA*


SQL Server
Out-of-process server
SQL code +SQLCLR procedure or function
IPC
Special-case functionality

```
declare @sql nvarchar(max)
set @sql = N'exec SqA1.dbo.SWGPerf @qOffset=311, @chrNum=7, @dimTile=128'
exec dbo.ExecSWG @sqlCmd=@sql, @targetTable='##tmpX08'
```

Demo: Galaxy Correlations

- Generated 16M random points with correct radial and angular selection for SDSS-N
- Done on an NVIDIA GeForce 295 card
- Brute force massively parallel N^2 code is much faster than tree-code for hi-res correlation function
- All done inside the JHU SDSS SQL Server database
- Correlation function is now SQL UDF:

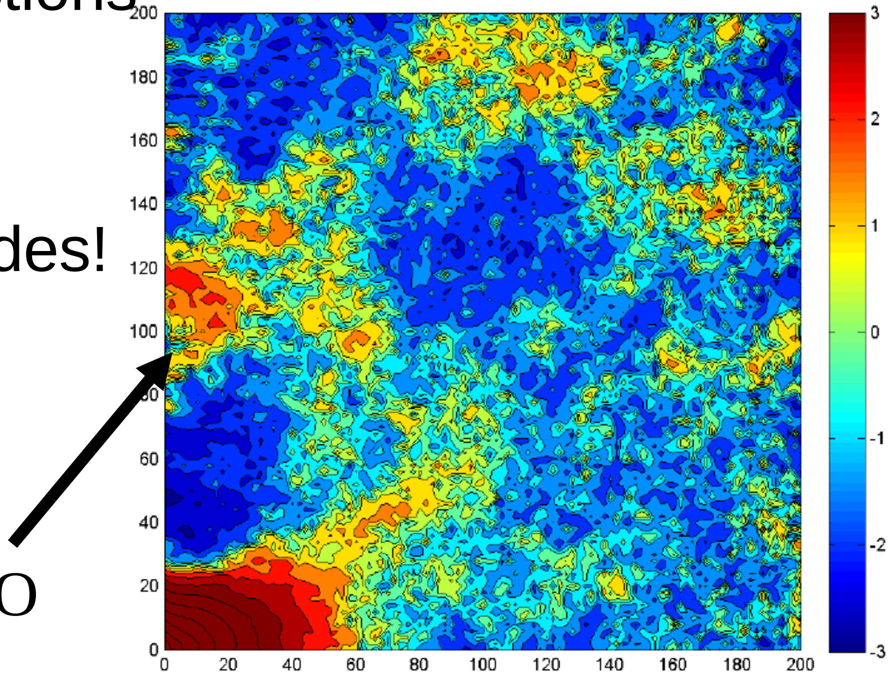


```sql
select dd.i, dd.j,  dd.cts as dd, dr.cts as dr, rr.cts as rr
    (@Nrr*CONVERT(float,dd.cts)/@Ndd - 2*@Nrr*CONVERT(float,dr.cts)/@Ndr + CONVERT(float,rr.cts))
        / CONVERT(float,rr.cts) as xi
from     dbo.PairCounts(@maxmpc, @nbin, @qryD, @nD, null) dd
    join dbo.PairCounts(@maxmpc, @nbin, @qryR, @nR, null) rr  on dd.i = rr.i and dd.j = rr.j
    join dbo.PairCounts(@maxmpc, @nbin, @qryDR, @nD, @nR) dr  on dd.i = dr.i and dd.j = dr.j
```

Correlations: Impact of GPUs

- Reconsider the N logN only approach
- Once we can run 100K threads, maybe running SIMD N^2 on smaller partitions is also acceptable
- Inside the DB: integrating CUDA with SQL Server, with SQL User Defined Functions
- Galaxy spatial correlations: **600 trillion galaxy pairs**
- Much faster than the tree codes!

Tian, Budavari, Neyrinck, Szalay
ApJ 2011



BAO

Arrays in SQL Server

- Recent effort by Laszlo Dobos
- Written in C++
- Arrays packed into varbinary(8000) or varbinary(max)
- 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

Amdahl's Laws

Gene Amdahl (1965): **Laws for a balanced system**

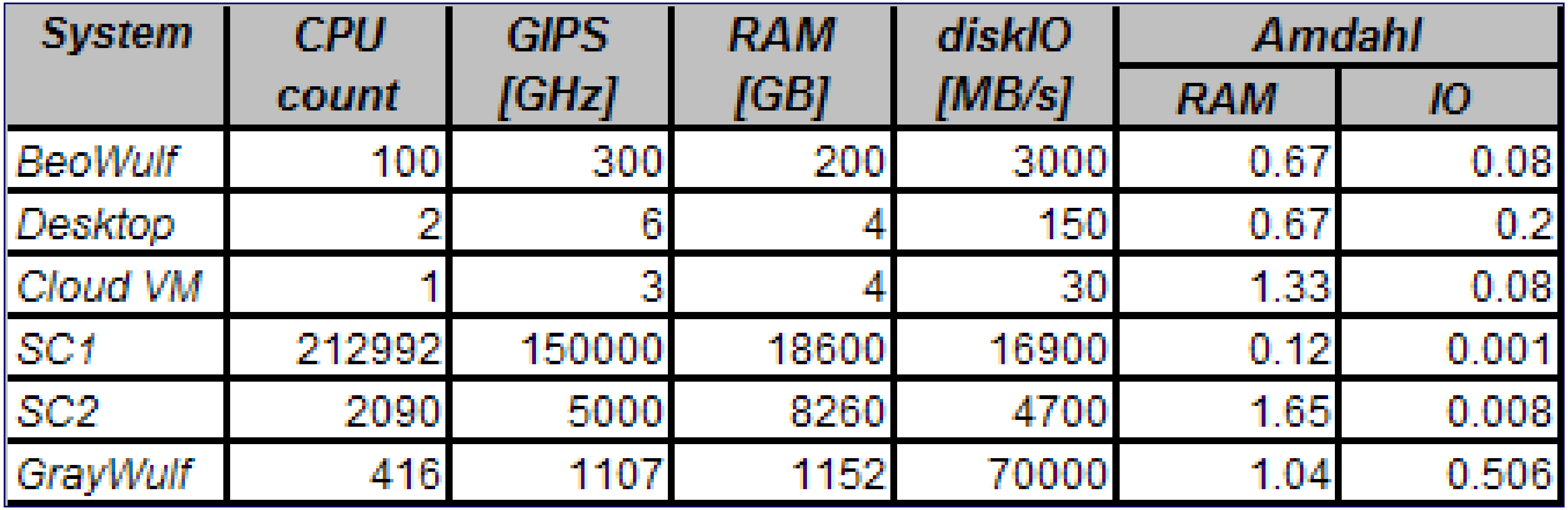
i. Parallelism: max speedup is S/(S+P)

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

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

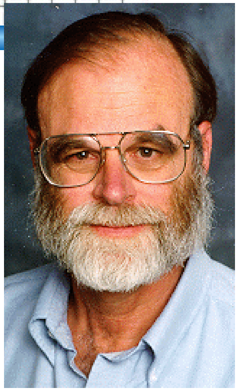
Modern multi-core systems move farther
away from Amdahl's Laws
(Bell, Gray and Szalay 2006)

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

Petascale Computing at JHU



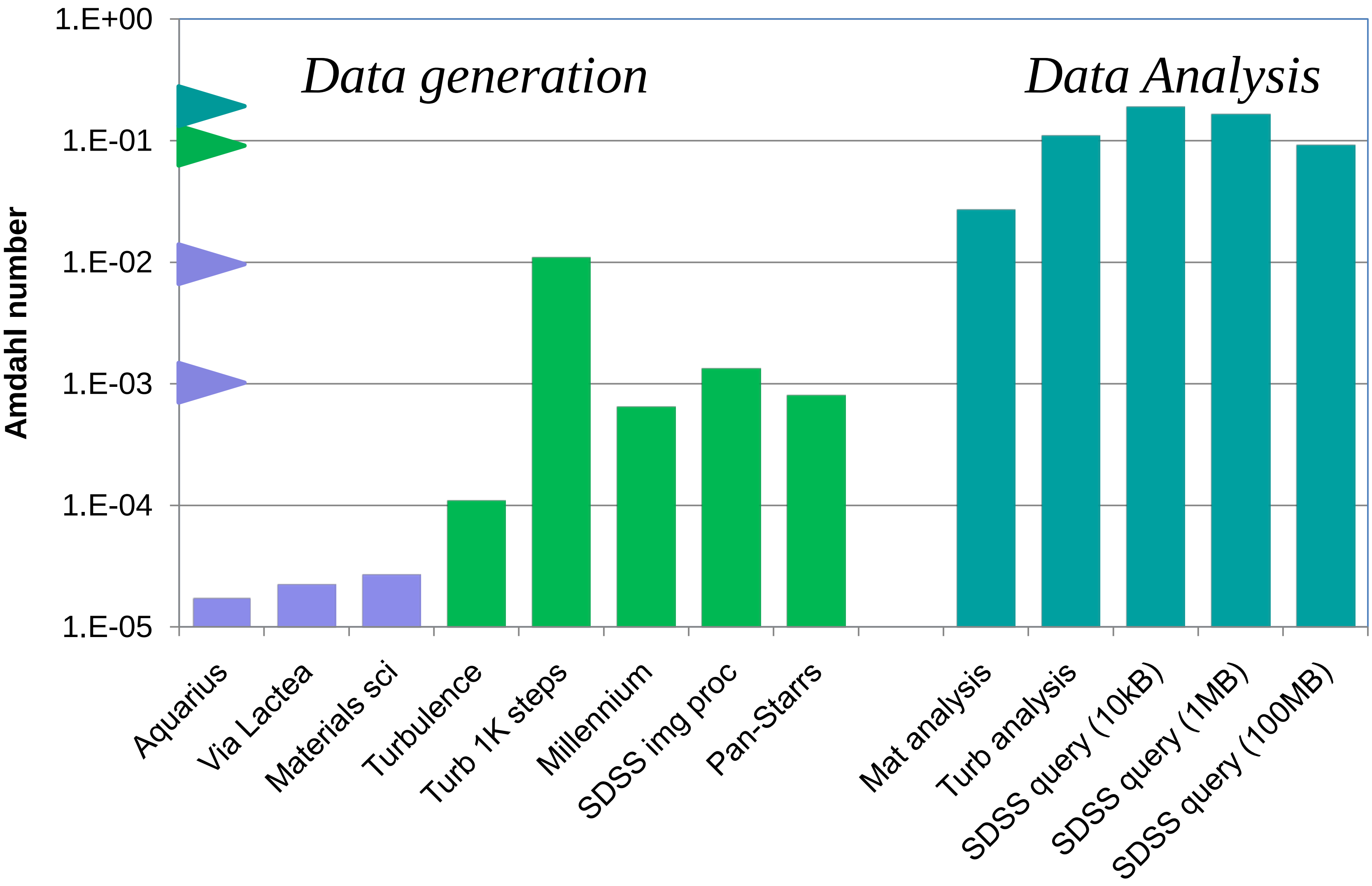
- Distributed SQL Server cluster/cloud w.
- 50 Dell servers, 1PB disk, 500 CPU
- Connected with 20 Gbit/sec Infiniband
- 10Gbit lambda uplink to UIC
- Funded by Moore Foundation, Microsoft and Pan-STARRS
- Dedicated to eScience, provide public access through services
- Linked to 1000 core compute cluster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

graywulf

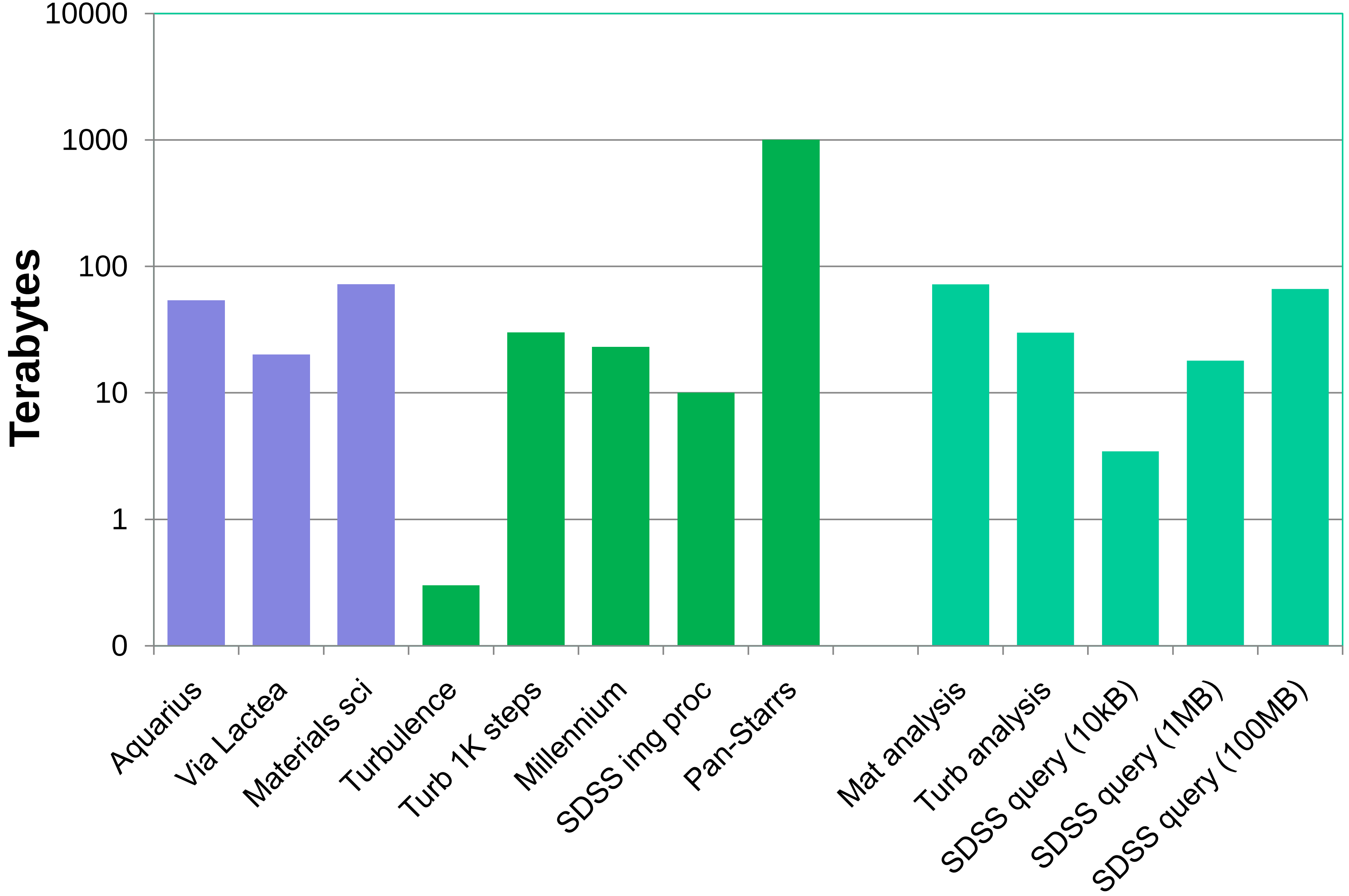
Photo-Z on Cyberbricks

- 36-node Amdahl cluster using 1200W total
- Zotac Atom/ION motherboards
  - *4GB of memory, N330 dual core Atom, 16 GPU cores*
- Aggregate disk space 76TB
  - *36 x 120GB SSD = 4.3 TB*
  - *72x 1TB Samsung F1 = 72.0 TB*
- Blazing I/O Performance: 18GB/s
- Amdahl number = 1 for under $30K
- Using the GPUs for data mining:
  - *6.4B multidimensional regressions (photo-z) over 1.2TB of data (128M x 50 estimators) in 5 minutes*
  - *Random Forest algorithm inside the DB on GPUs*

Amdahl Numbers for Data Sets



The Data Sizes Involved

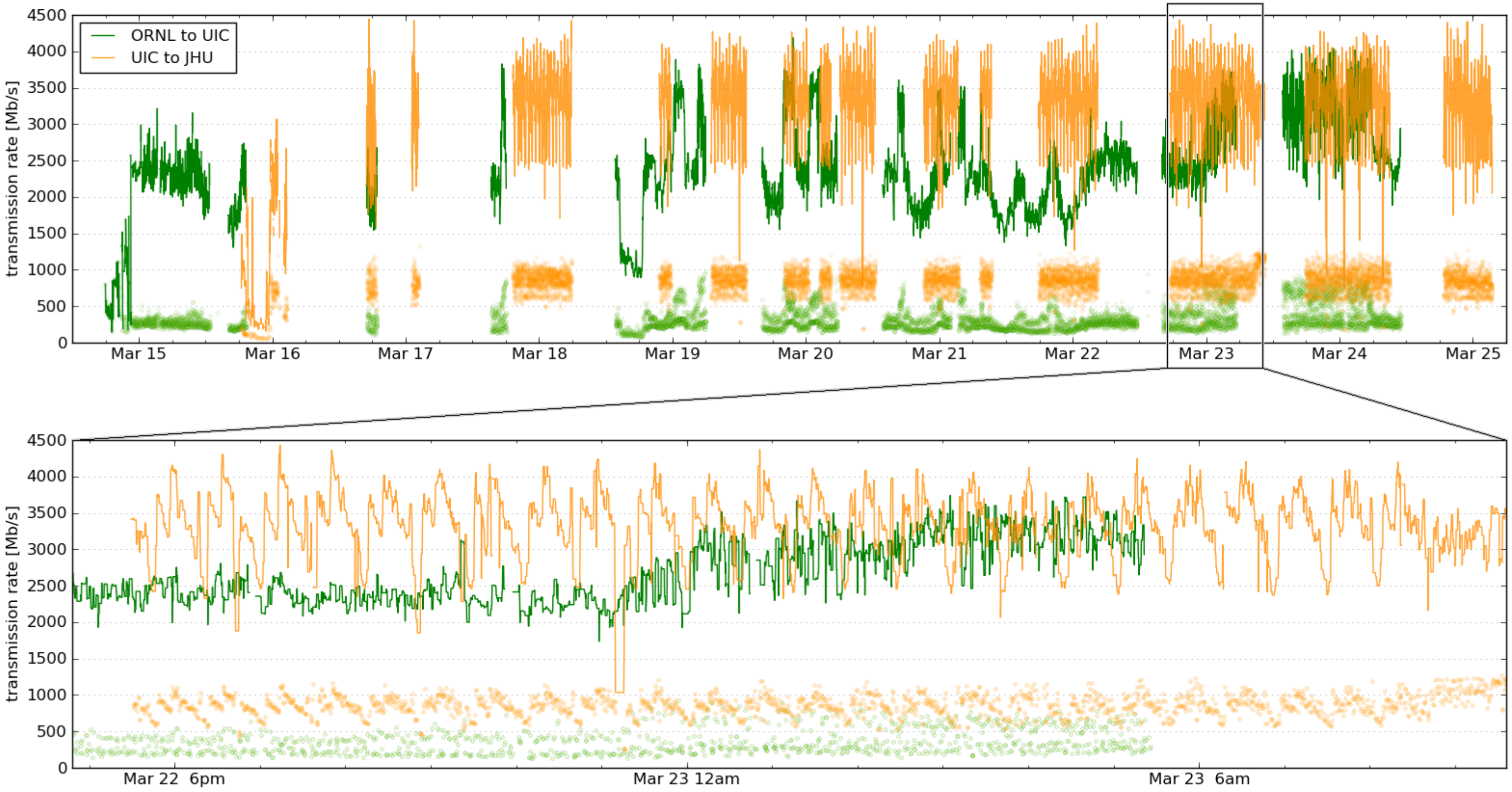


DISC Needs Today

- 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*

Silver River Transfer

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



Tradeoffs Today

“Extreme computing is about tradeoffs”

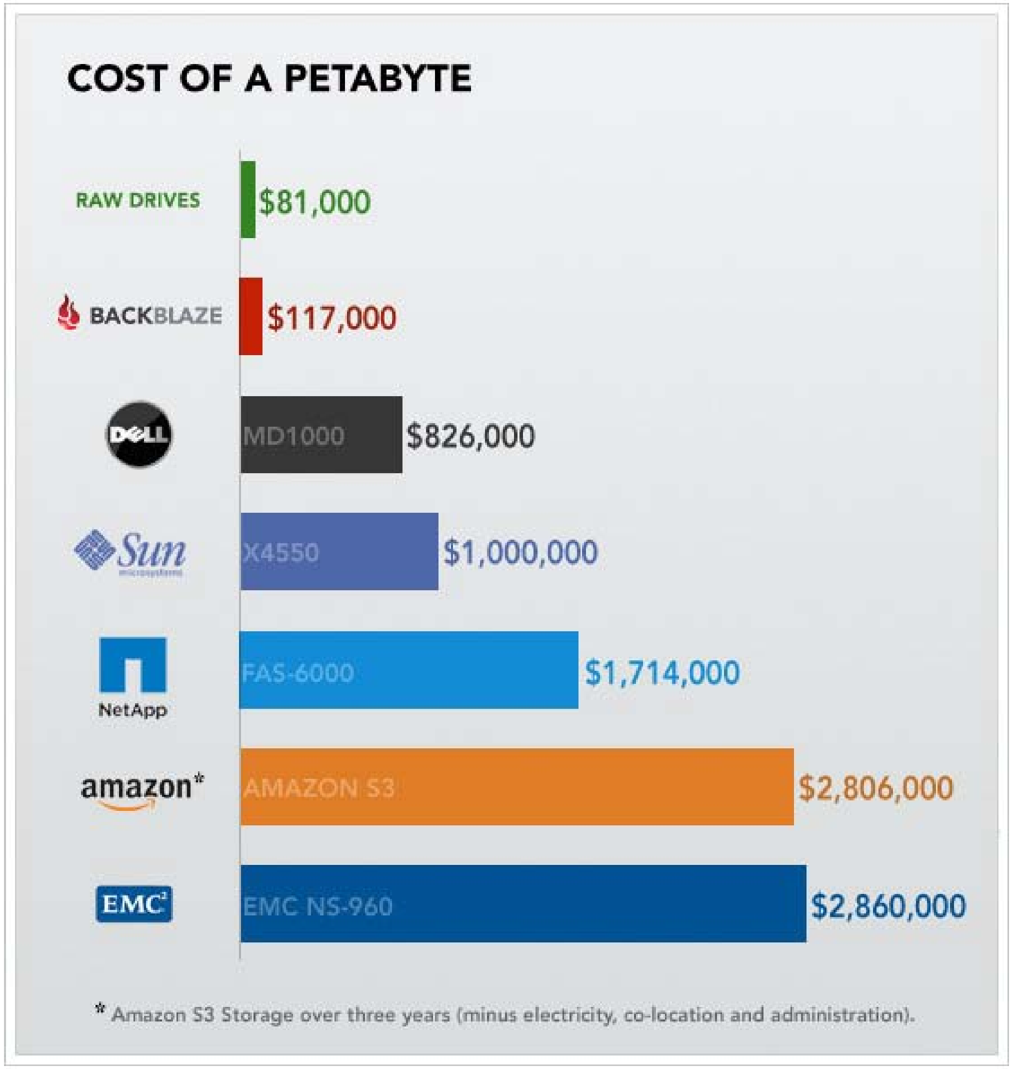
Stu Feldman (Google)

Ordered priorities for data-intensive scientific computing

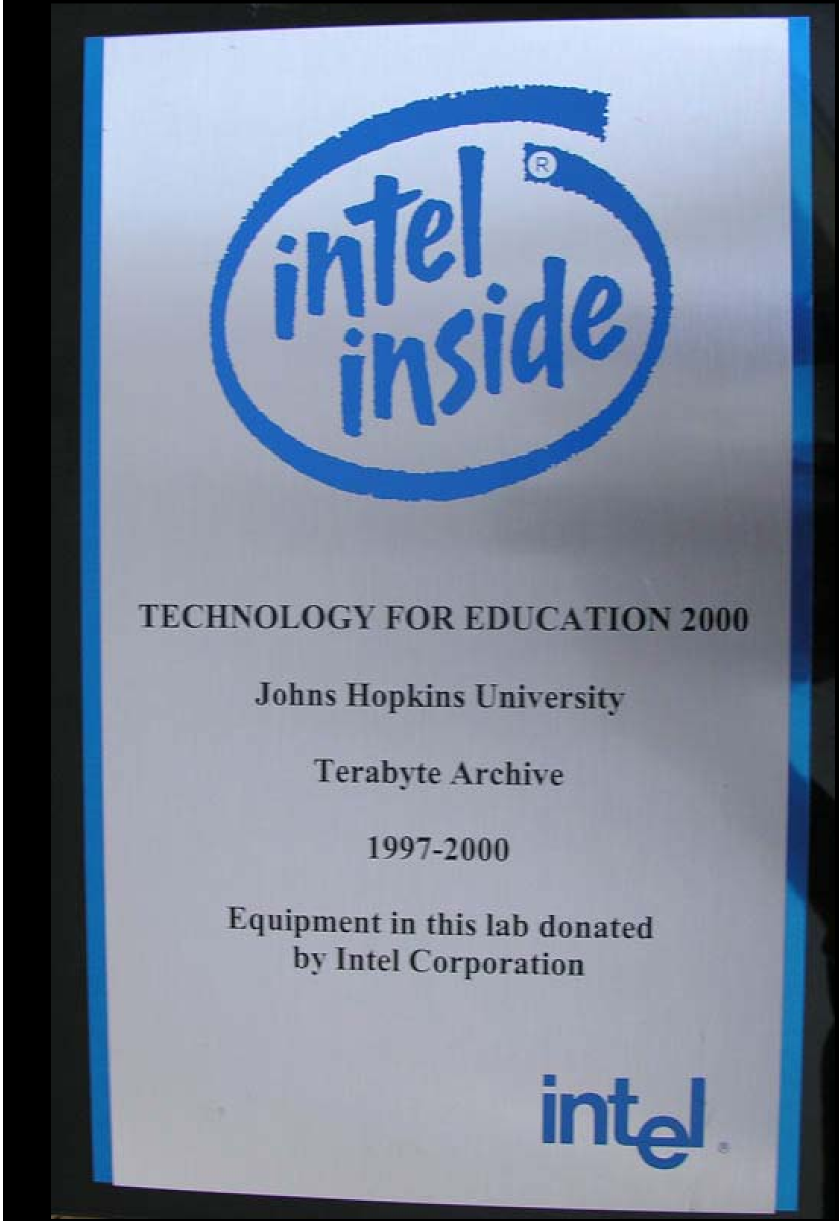
1. *Total storage (-> low redundancy)*
2. *Cost (-> total cost vs price of raw disks)*
3. *Sequential IO (-> locally attached disks, fast ctrl)*
4. *Fast stream processing (->GPUs inside server)*
5. *Low power (-> slow normal CPUs, lots of disks/mobo)*

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

Cost of a Petabyte



From backblaze.com
Aug 2009



intel inside

TECHNOLOGY FOR EDUCATION 2000

Johns Hopkins University

Terabyte Archive

1997-2000

Equipment in this lab donated
by Intel Corporation

intel

graywulf

JHU Data-Scope

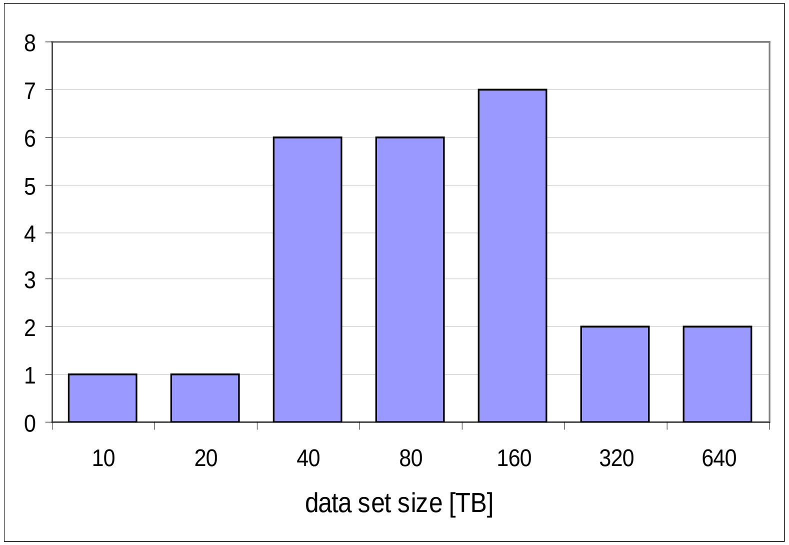
- Funded by NSF MRI to build a new ‘instrument’ to look at data
- Goal: 102 servers for $1M + about $200K switches+racks
- Two-tier: performance (P) and storage (S)
- Large (5PB) + cheap + fast (400+GBps), but …

  . ..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 proposed for the Data-Scope, more coming, data lifetimes between 3 mo and 3 yrs

Fractal Vision

- The Data-Scope created a lot of excitement but also a lot of fear at JHU…
  - *Pro: Solve problems that exceed group scale, collaborate*
  - *Con: Are we back to centralized research computing?*
- Clear impedance mismatch between monolithic large systems and individual users
- e-Science needs different tradeoffs from eCommerce
- Larger systems are more efficient
- Smaller systems have more agility
- How to make it all play nicely together?

Increased Diversification

One shoe does not fit all!

- Diversity grows naturally, no matter what
- Evolutionary pressures help
  - *Large floating point calculations move to GPUs*
  - *Large data moves into the cloud*
  - *RandomIO moves to Solid State Disks*
  - *Stream processing emerging (SKA…)*
  - *noSQL vs databases vs column store vs SciDB …*
- Individual groups want subtle specializations

At the same time

- What remains in the middle (common denominator)?
- Boutique systems dead, commodity rules
- We are still building our own…

DISC Sociology

- What happens to a discipline after the world's largest instrument is built?
  - *We should not take for granted that there will be a next*
- Broad sociological changes
  - *Data collection in ever larger collaborations (VO)*
  - *Analysis decoupled, on archived data by smaller groups*
- The impact of power laws
  - *we need to look at problems in octaves*
  - *Pareto rule (90% of the people only look at 10% of data)*
  - *the scientists may only be the tail of our users*
  - *there is never a discrete end or a sharp edge (except for our funding)*

DISC Economics

- What is the price of **software**?
  - *30% from SDSS, more for LSST*
  - *Repurpose for other disciplines, do not reinvent the wheel*
- What is the price of **hardware**?
  - *Moore's Law comes to the rescue…*
    we could build the LSST HW today, no problem in 10 years
  - *Extreme computing is about extreme tradeoffs….*
- What is the price (value) of **data**?
  - *$100,000 /paper (Ray Norris)*
- The cost of total ownership and business model contrasted with level budgets

DISC Technology

- **Storage**: by next year we will have Petabytes
- **Networking**: we need to move them (LSST)
- **Stream processing**: cannot even save them (SKA)!
- **Information Science**: need systematic data curation (Data Conservancy)
- **Computations**: archives == computational services

- **Tradeoffs**: What is the right balance among economy of scale, diversification and agility?

Summary

- Science is increasingly driven by large data sets
- Large data sets are here, COTS solutions are not
  - *100TB is the current practical limit*
- We need a new instrument: a “microscope” and “telescope” for data=> a **Data-Scope**!
- Increasing diversification over commodity HW
- Changing sociology:
  - *Data collection in large collaborations (VO)*
  - *Analysis done on the archived data, possible (and attractive) for individuals*
- A new, Fourth Paradigm of Science is emerging…

but it is not incremental….

[redacted]

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

Henry Ford

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