Microsoft Research
FacultySummit

Scientific Data Analysis Today: From Terabytes to Petabytes

Alexander Szalay
The Johns Hopkins University

2031
FUTURE/WORLD
2011

The Science of Big Data

- Data growing exponentially, in all science
- Changes the nature of all science
- Non-incremental!
- Industry and government faces the same challenges
  - Microsoft, Google, Yahoo, DOD,….
- Convergence of physical and life sciences through Big Data (statistics and computing)
- A new scientific revolution

  => a rare and unique opportunity

Non-Incremental Changes

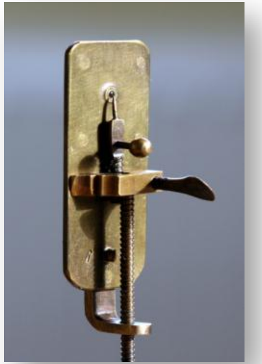
- Science is moving from hypothesis-driven to data-driven discoveries

Astronomy has always been data-driven....
now becoming more generally accepted

- Multifaceted challenges:
  - New data intensive scalable architectures
  - New randomized, incremental algorithms
  - New computational tools and strategies

  ... not just statistics, not just computer science, not just astronomy...

- Need a microscope of data



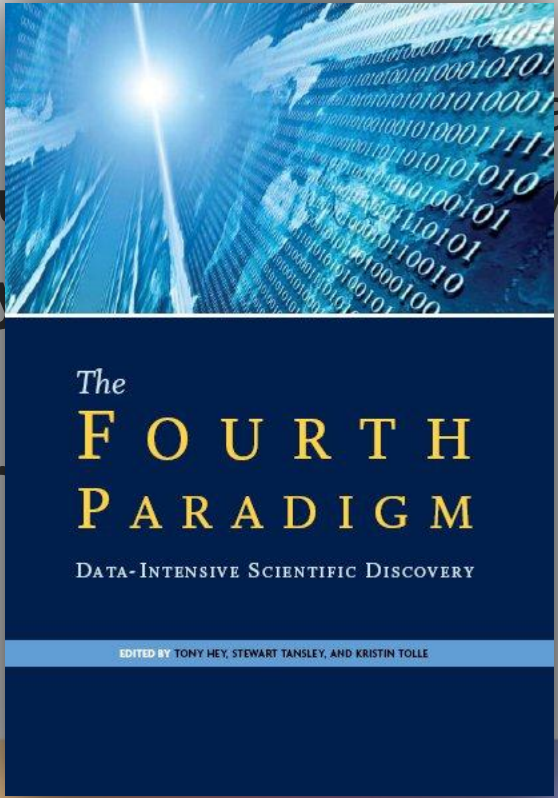
Scientific Data Analysis Today

- Scientific data is doubling every year, now reaching PBs
- Architectures increasingly CPU-heavy, IO-poor
  - New, more data-intensive scalable architectures are needed
- Databases are a good starting point, but scientists need special features (arrays, GPUs)
- Need new, incremental and randomized algorithms
- Most data analysis done on midsize BeoWulf clusters
- Universities hitting the “power wall”
- **Not scalable, not maintainable…**

Gray's Laws of Data Engineering

Jim Gray:

- Scientific com around **data**
- Need **scale-o** sis
- Take the **anal**
- Start with "**20**
- Go from "**wor**

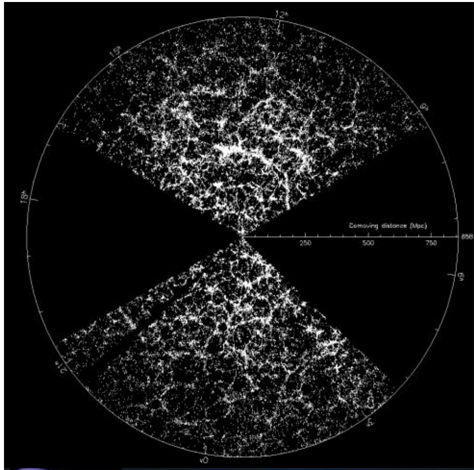


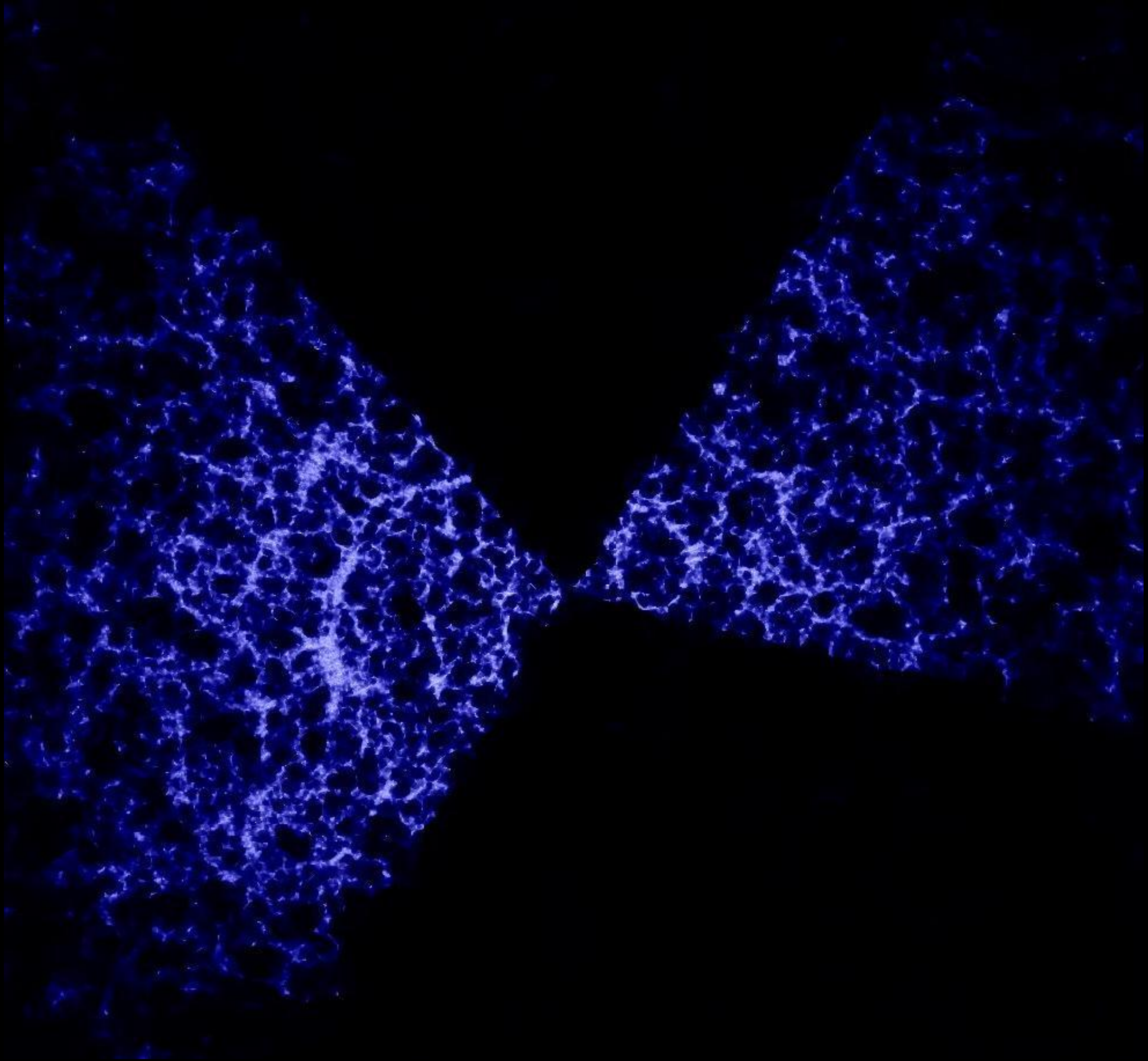
Building Scientific Databases

- 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
- Do the scientific computations inside the database!
- During this time data sets grew from 100GB to 1PB
- Interactions with every step of the scientific process

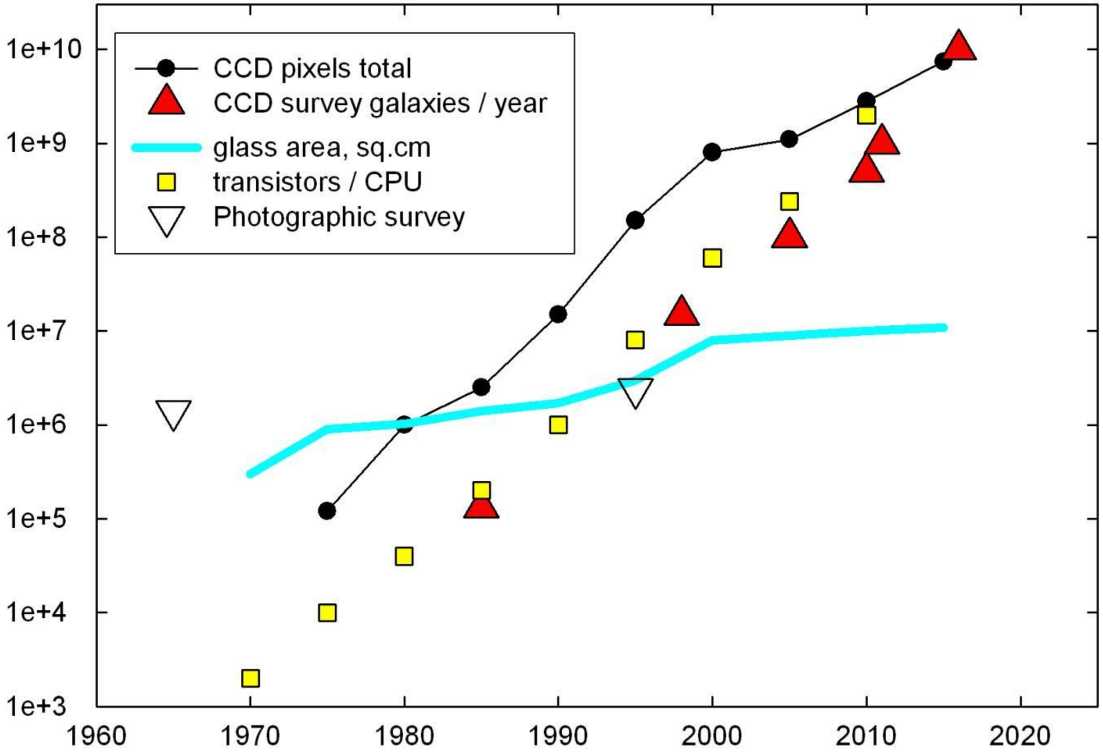
Sloan Digital Sky Survey

- **"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
- Extra data volume enabled by
  - Moore's Law, Kryder's Law





Survey Trends

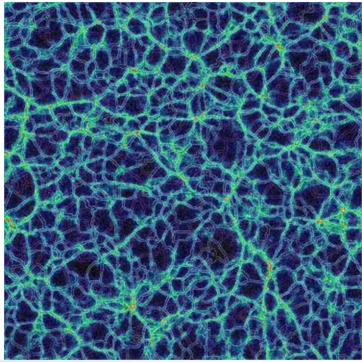


T.Tyson (2010)

Continuing Growth

How long does the data growth continue?

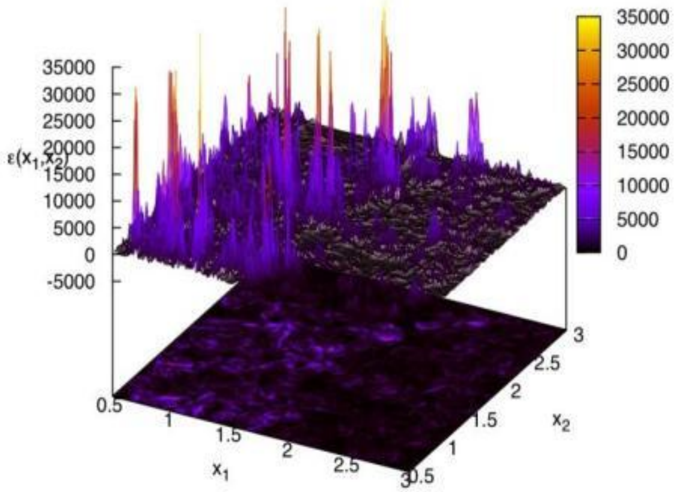
- High end always linear
- Exponential comes from technology + economics
  - rapidly changing generations
  - like CCD's replacing plates, and become ever cheaper
- How many generations of instruments are left?
- Are there new growth areas emerging?
- **Software is becoming a new kind of instrument**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



Immersive Turbulence

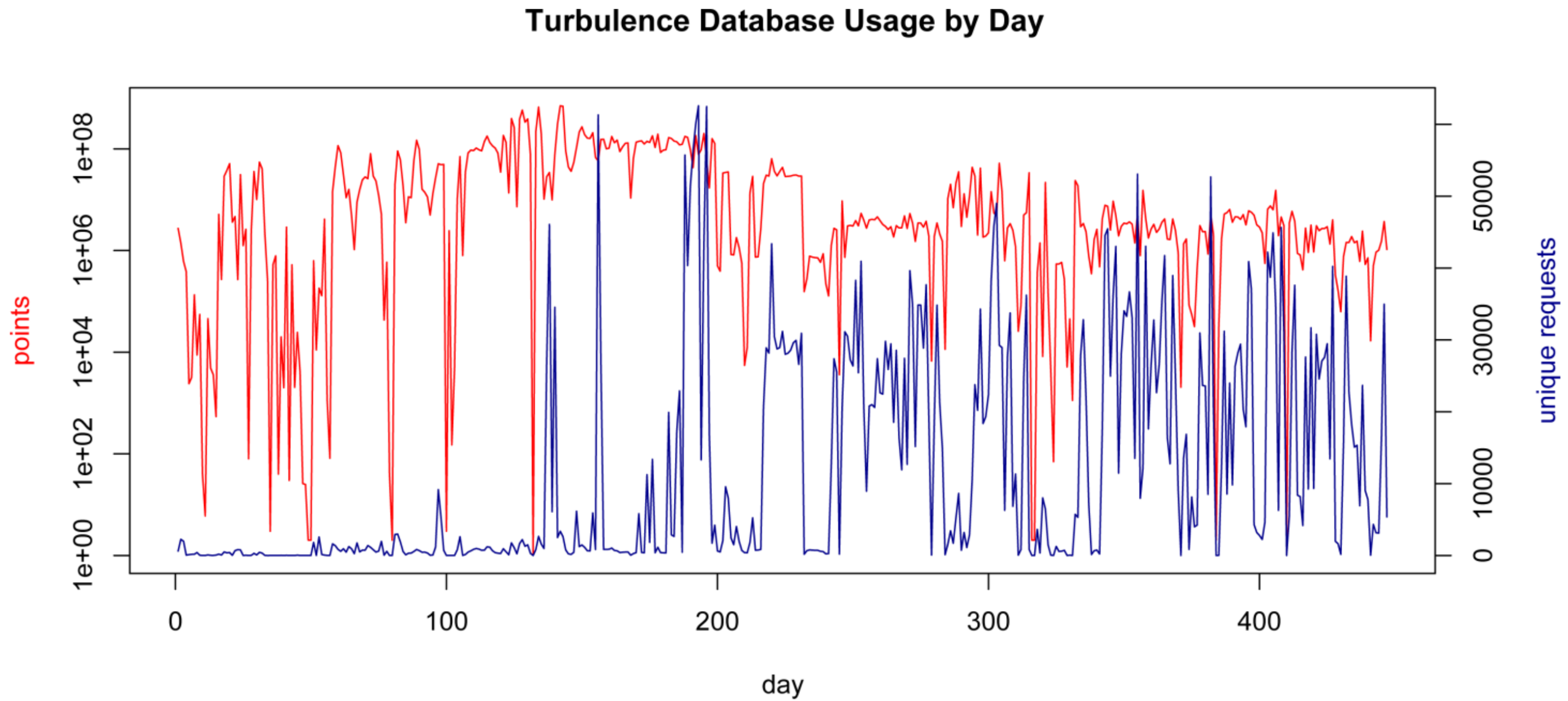
"… the last unsolved problem of classical physics…" Feynman

- **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)

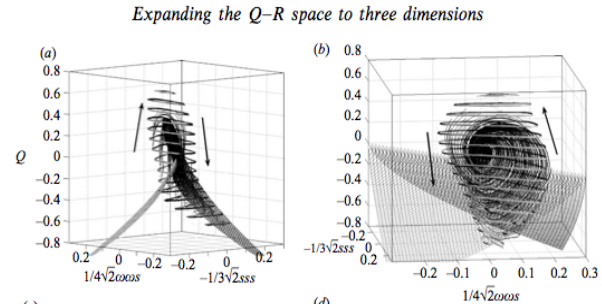
Daily Usage



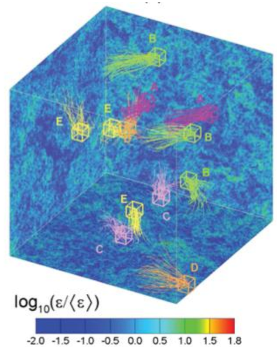
Turbulence Database Usage by Day
points
1e+00
1e+02
1e+04
1e+06
1e+08
unique requests
0
10000
30000
50000
0
100
200
300
400
day

Turbulence Research with the Database

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)

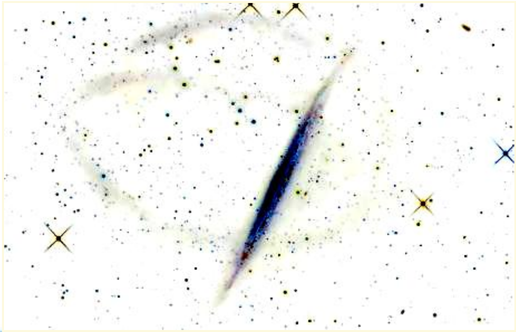


Lagrangian time correlation in turbulence
Yu & Meneveau,
Phys. Rev. Lett. 104, 084502 (2010)

The Milky Way Laboratory

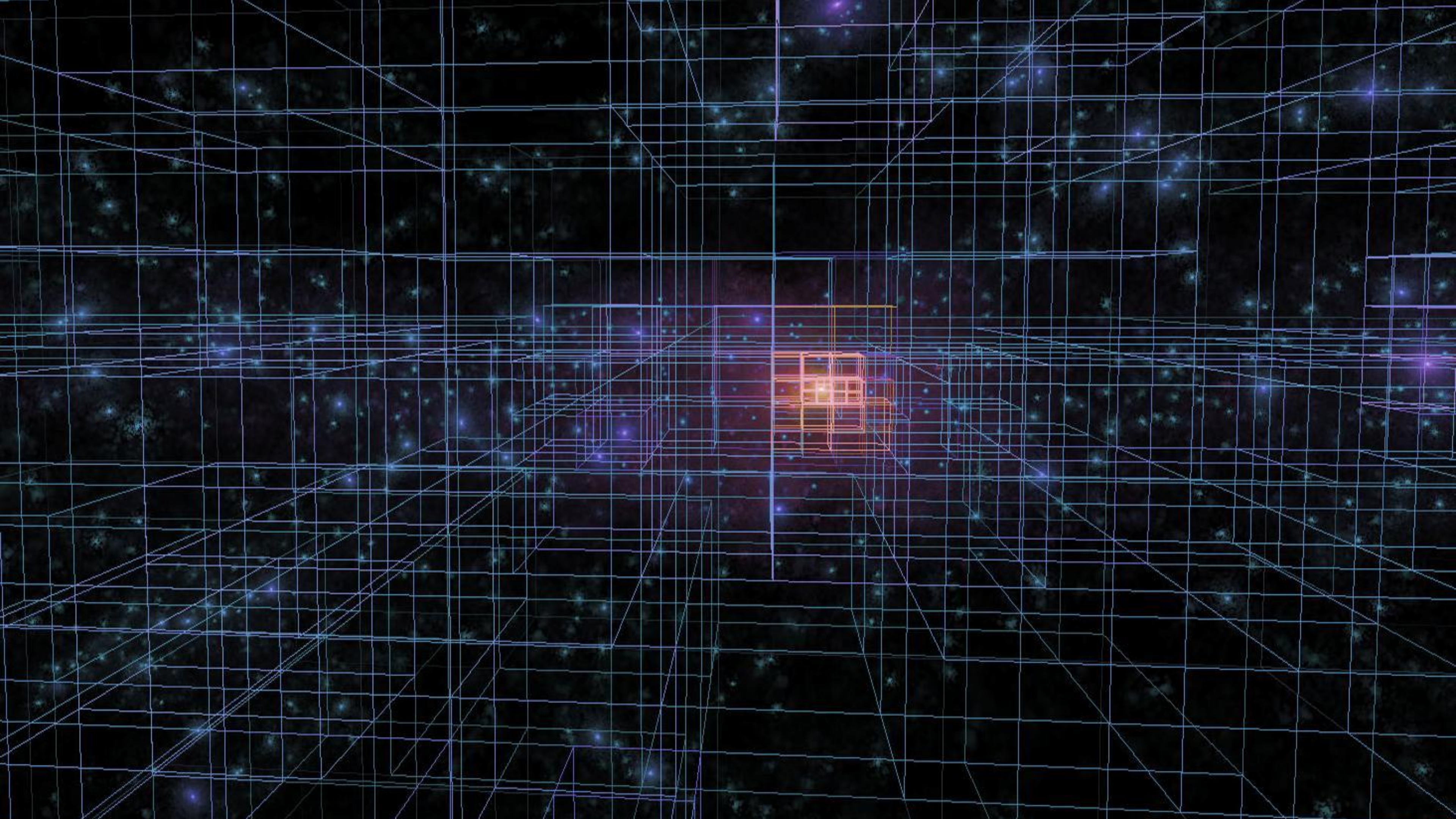
- Use cosmology simulations as immersive laboratory for general users
- Via Lactea-II (20TB) as prototype, then Silver River (50B particles) as production (15M CPU hours at the Oak Ridge Jaguar)
- 800+ hi-rez snapshots (2.6PB) => 800TB in DB
- Users can insert test particles (dwarf galaxies) into system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

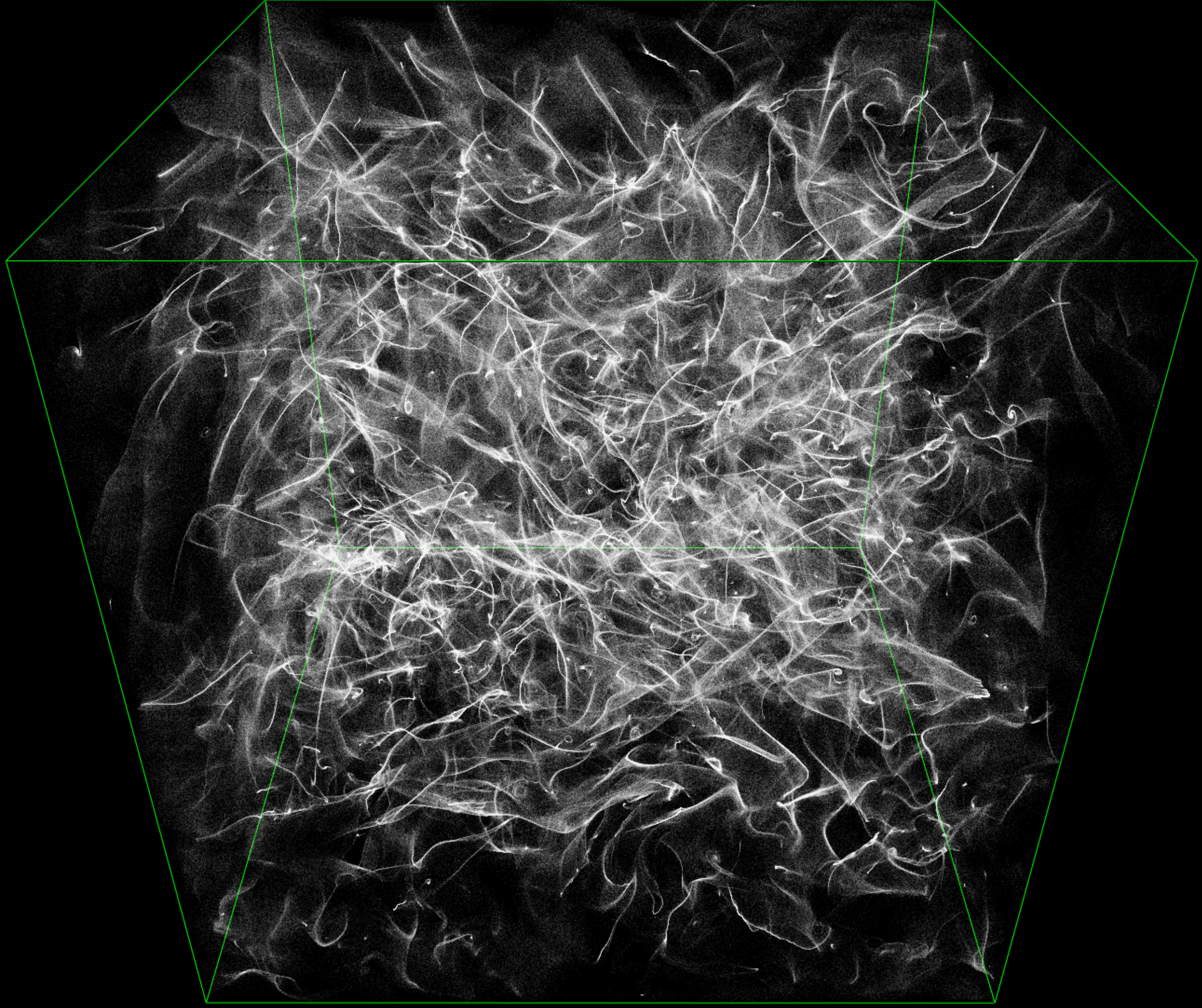
Madau, Rockosi, Szalay, Wyse, Silk, Lemson, Westermann, Blakeley, just funded by the NSF

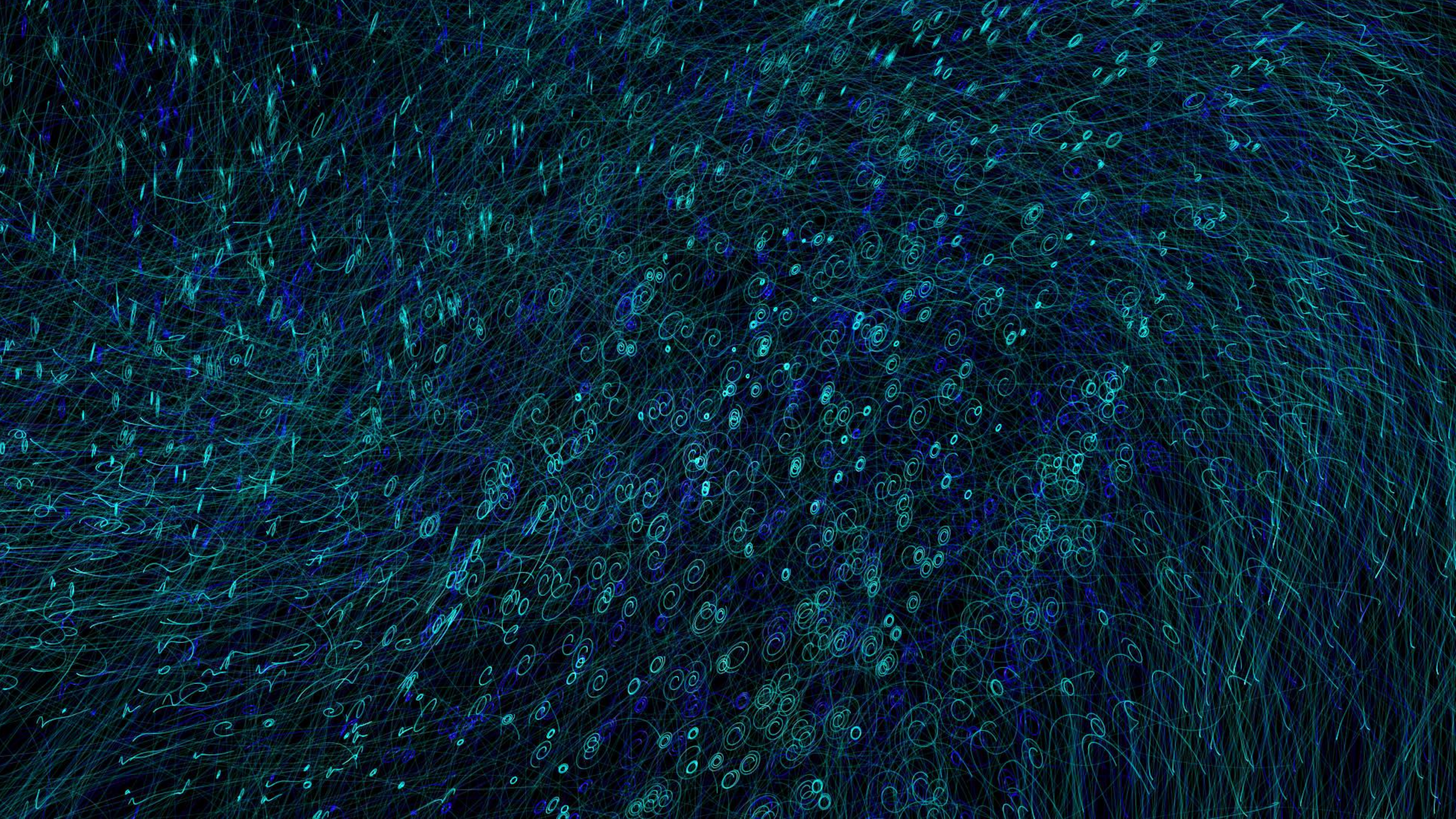


Visualizing Petabytes

- 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 SQL Server to GPU
- N-body simulations next







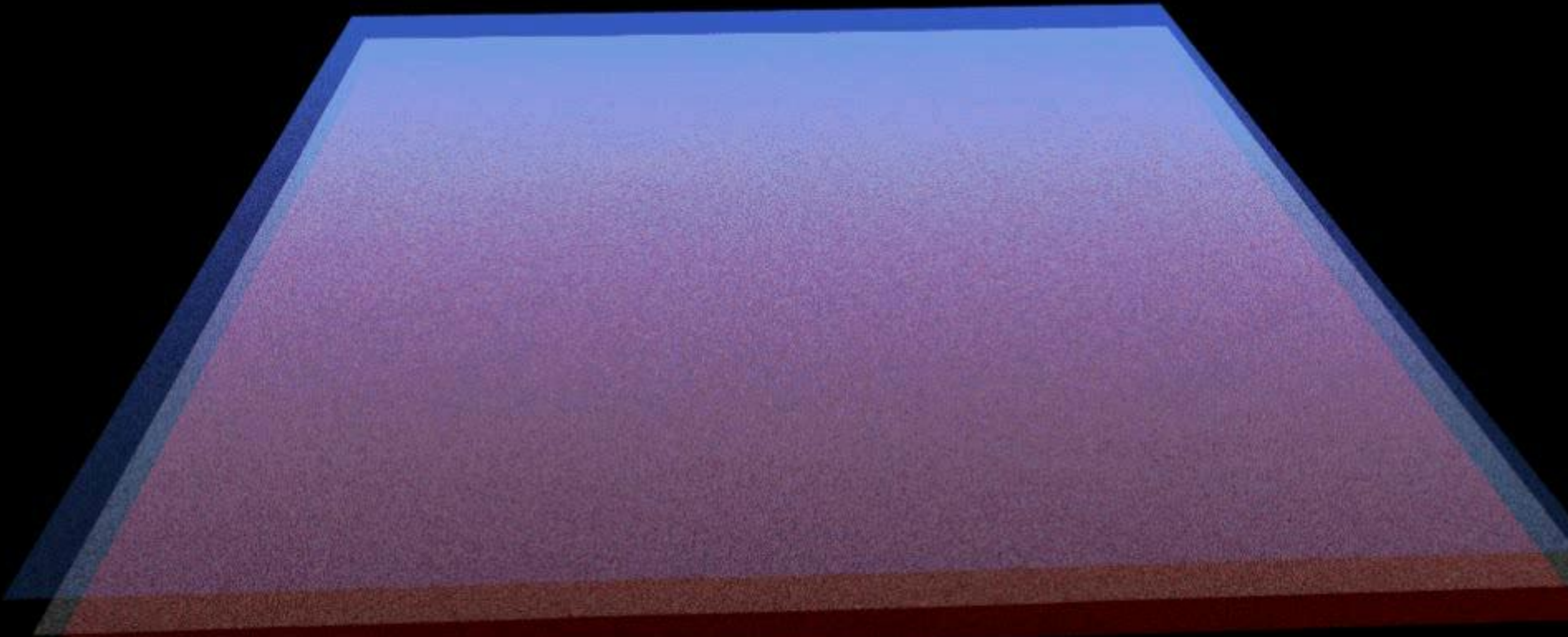
Microsoft Research
FacultySummit

VIDEO

3D Vorticity in a Turbulent Flow

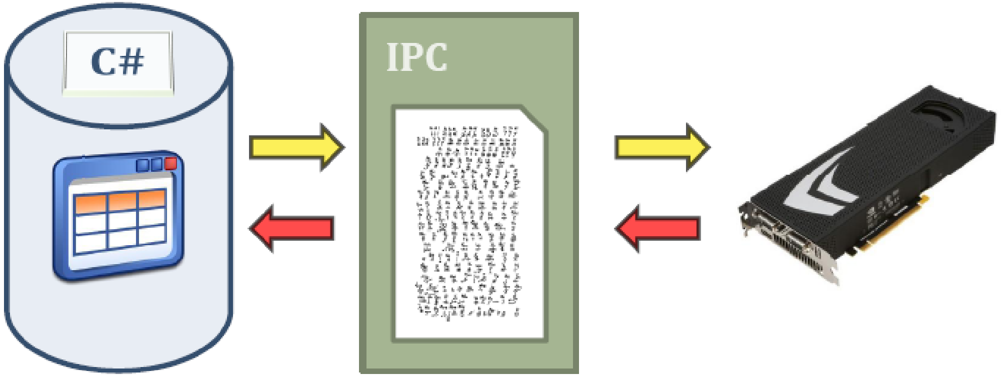
Kai Buerger and Alex Szalay
Technische Universitat Munich, and JHU

FUTURE WORLD 2011 2031



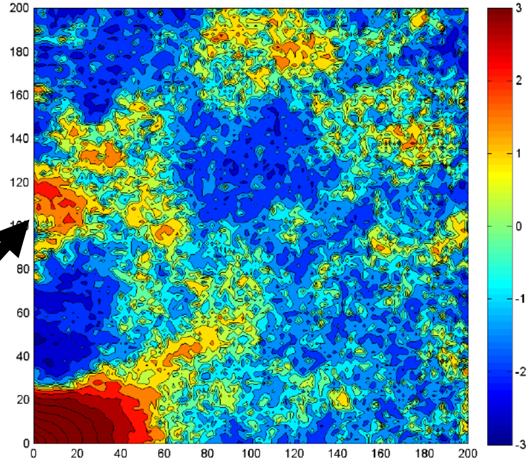
Extending Databases

- 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



Galaxy Correlations: Impact of GPUs

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



Acoustic Resonance Frequency of the Universe

Large Arrays in SQL Server

- Recent effort by Laszlo Dobos (w. J. Blakeley and D. Tomic)
- 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')
```

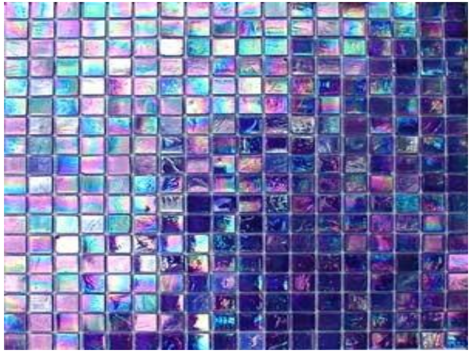
@a is an array of doubles with 3 indices
The first command averages the array over 4×4×4 blocks, returns indices and the value of the average into a table
Then we build a new (collapsed) array from its output

Querying Petabytes

- Add a layer to existing RDBMS that supports…
  - Statistical queries
  - Procedural queries
  - Fault tolerance for big queries
  - Scalable behavior
  - “Map/Reduce”-like crawler but with indexing
- Database already good…but not scalable enough
  - Break up data into small partitions (“tiles”)
  - Intercept and modify SQL
  - Run incremental query stream on tile set
  - Determine streaming order dynamically
  - Fast convergence for aggregate statistics

TileDB

- Distributed DB that adapts to query patterns
- No set physical schema
  - Represents data as tiles
  - Tiles replicate/migrate based on actual traffic
- Can automatically load from existing DB
  - Inherits schema (for querying only!)
- Fault tolerance
  - From one query, derive many
  - Each mini-query is a checkpoint
  - Can also estimate overall progress though ‘tiling’
- Execution order can be determined by sampling
  - Faster then sqrt(N) convergence



Nolan Li thesis
2011, JHU

Table

C1	C2	C3
A	1	-1
B	2	-2
C	3	-3
D	4	-4
E	5	-5
F	6	-6
G	7	-7

```
SELECT *
FROM TABLE
```

Tile Set

C1	C2	C3
A	1	-1
B	2	-2
C	3	-3
D	4	-4
E	5	-5
F	6	-6
G	7	-7

```
SELECT C1, C2
FROM TABLE
WHERE C3 <> -7
```

Tiles

C1	C2	C3
A	1	-1
B	2	-2
C	3	-3
D	4	-4
E	5	-5
F	6	-6
G	7	-7

```
SELECT C1, C2
FROM TABLE
WHERE C3 <> -7
   AND C1 >= 1 AND C2 < 3

SELECT C1, C2
FROM TABLE
WHERE C3 <> -7
   AND C1 >= 3 AND C2 < 5
```

Table -> Tiles

- Start with a table
- A *tile set* is some high-granularity partition of the table
- *Tiles* describe divisions of a tile set
    - Based on a covering partition of a tile set
    - Roughly equivalent in query cost
- Tile sets and tiles are fully described with SQL

Data Analysis Needs Today

- Disk space, disk space, disk space!!!!
- Current problems not on Exabyte scale yet:
  - 10-30TB easy, 100TB doable, 300TB really hard
  - For detailed analysis we need to park data for several months
- If not sequential access for a 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

"Extreme computing is about tradeoffs"

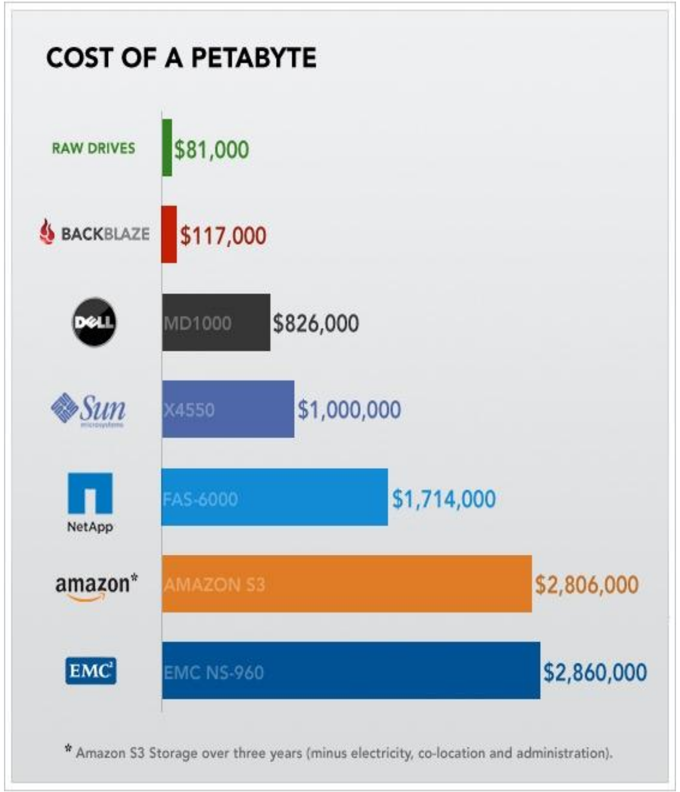
Stu Feldman (Google)

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 (-> slower CPUs, lots of disks/mobo)

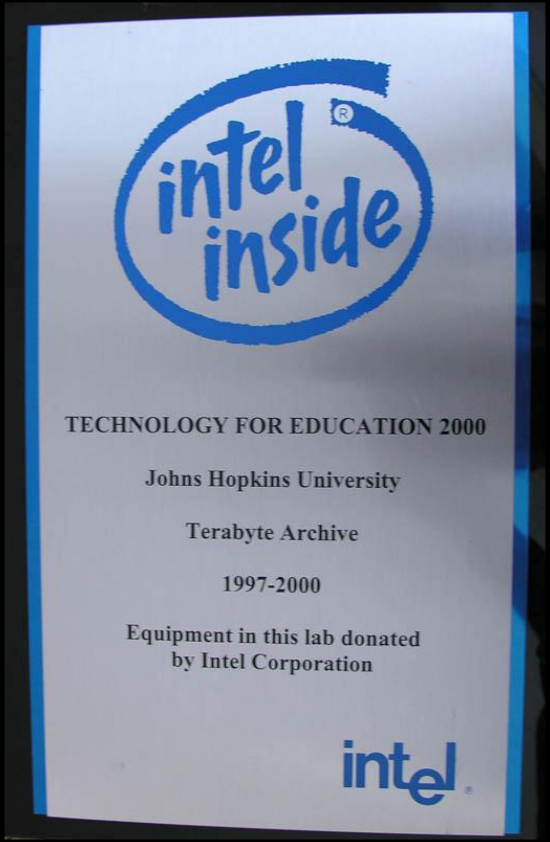
The order will be different in a few years...and scalability may appear as well

Cost of a Petabyte



From backblaze.com
Aug 2009

1TB in 2000



1PB: $\times 1000 = 2^{10}$

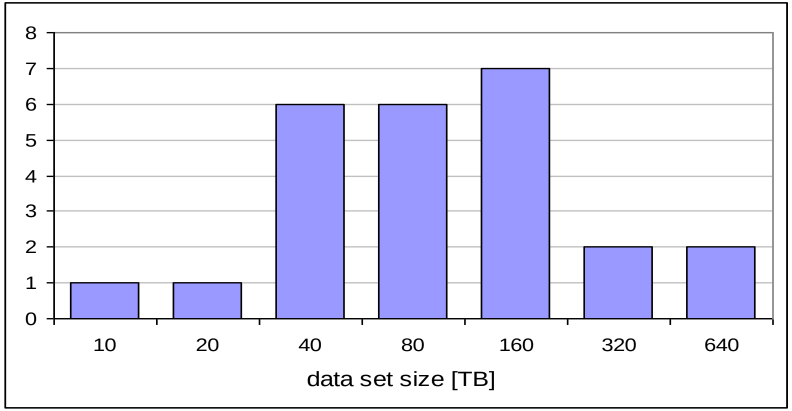
JHU Data-Scope

- 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 (450+GBps), but special purpose

	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

Increased Diversification

One shoe does not fit all!

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

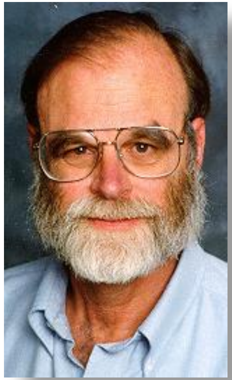
At the same time

- What remains in the middle?
    - Common denominator is Big Data
- Boutique systems dead, commodity rules
- We are still building our own…

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

Summary

- Science is increasingly driven by large data sets
- Large data sets are here, cheap, off-the-shelf solutions are not
  - 100TB is the current practical limit
- We need a new instrument: a "microscope" and "telescope" for data
- Increasing diversification over commodity hardware
- 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….

"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"

Microsoft® Research
FacultySummit

FUTURE WORLD 2011 2031

Microsoft® Research
FacultySummit

2031
FUTURE/WORLD
2011