Institute for Data Intensive Engineering and Science

idies

Extreme Data-Intensive Scientific Computing

Alex 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

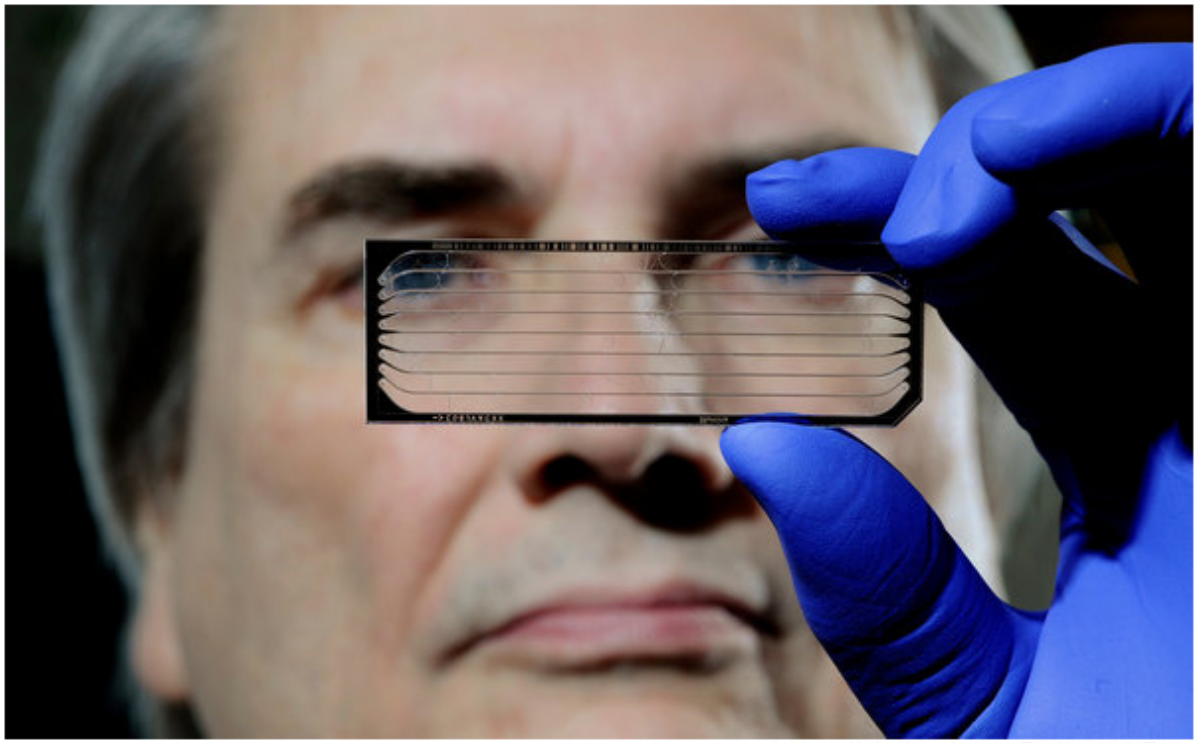
Scientific Data Analysis Today

- Scientific data is doubling every year, reaching PBs
  - *CERN is at 22PB today, 10K genomes ~5PB*
- Data will never will be at a single location
- Architectures increasingly CPU-heavy, IO-poor
- Scientists need special features (arrays, GPUs)
- Most data analysis done on midsize BeoWulf clusters
- Universities hitting the “power wall”
- Soon we cannot even store the incoming data stream
- **Not scalable, not maintainable…**

The New York Times

Business Day

DNA Sequencing Caught in Deluge of Data



Kathy Kmonicek for The New York Times

W. Richard McCombie, a professor of human genetics at the Cold Spring Harbor Laboratory, examining DNA samples.

By ANDREW POLLACK

Published: November 30, 2011

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

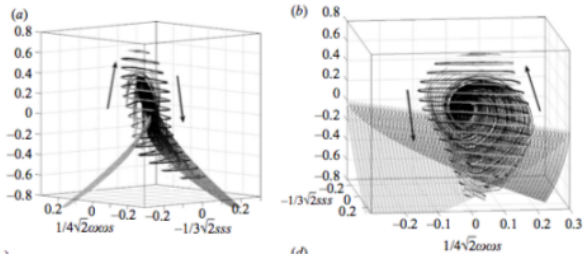
Data in HPC Simulations

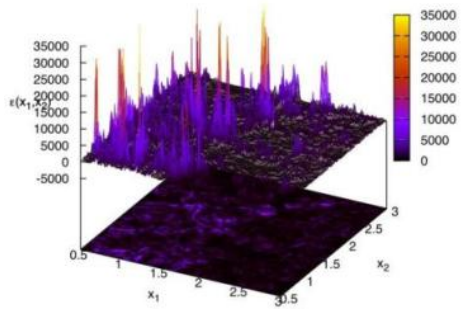
- HPC is an instrument in its own right
- Largest simulations approach petabytes
  - *from supernovae to turbulence, biology and brain modeling*
- Need public access to the best and latest through interactive numerical laboratories
- Creates new challenges in
  - *how to move the petabytes of data (high speed networking)*
  - *How to look at it (render on top of the data, drive remotely)*
  - *How to interface (virtual sensors, immersive analysis)*
  - *How to analyze (algorithms, scalable analytics)*

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)

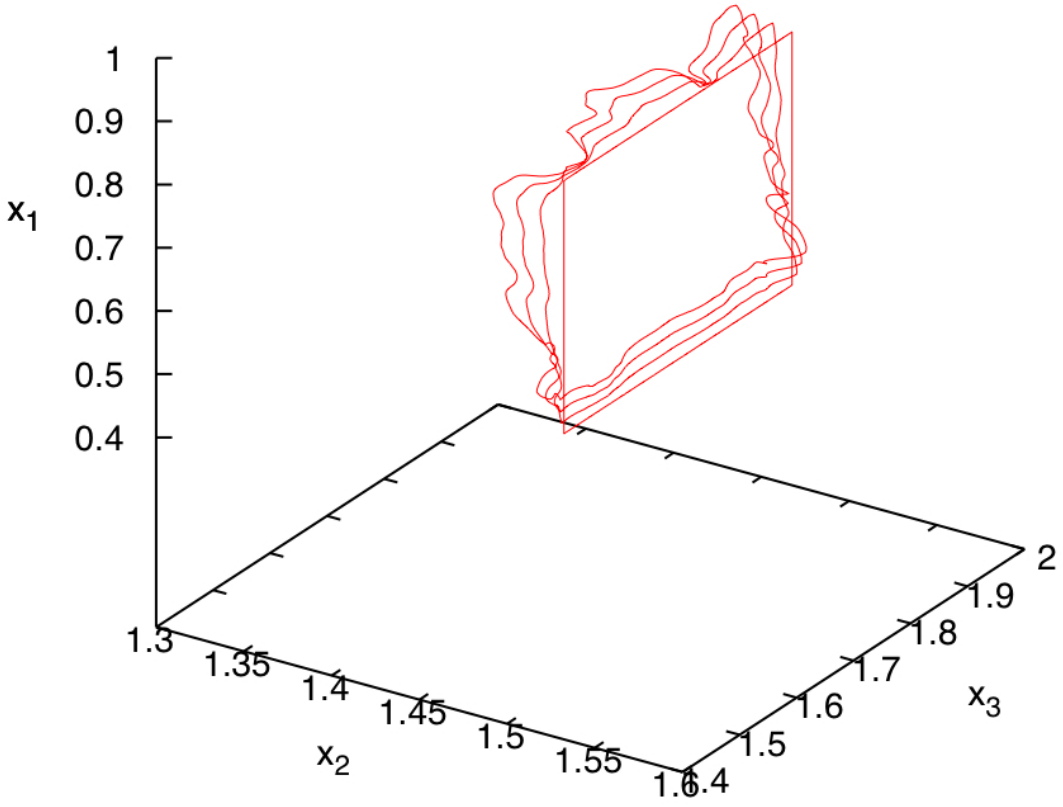
Sample code (fortran 90)

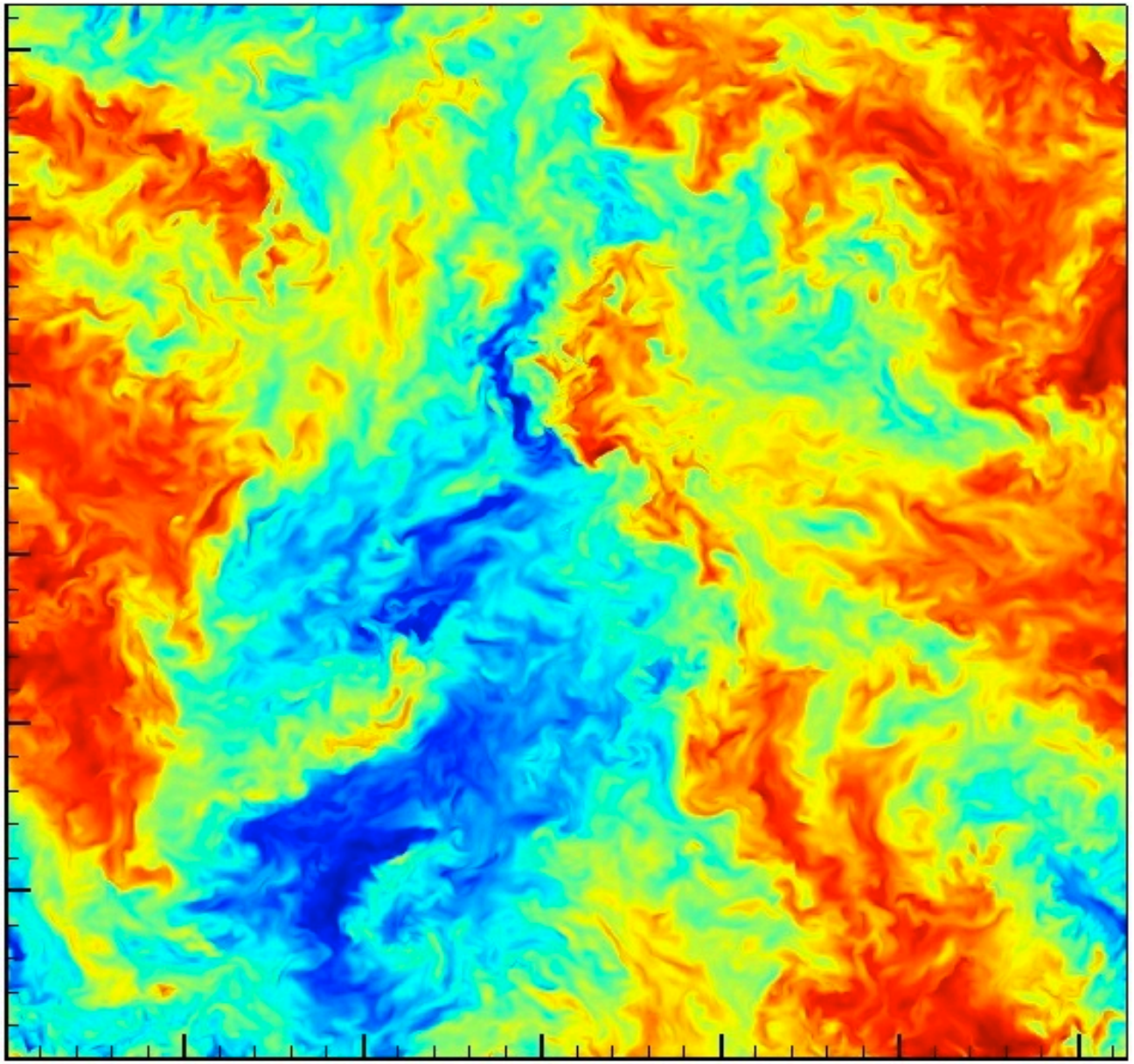
```
!
  do it = 1,15,1
!
    print *,'time = ',time
    time=time-deltat
!
    CALL getvelocity(authkey, dataset1, time, Lagrangian6thOrder , PCHIPInterpolation, 4*n, points, dataout)
 !
   do i=1,4*n
   do k=1,3
      points(k,i)=points(k,i)+dataout(k,i)*deltat
   end do
   end do
!
   if (it.eq.5.or.it.eq.10.or.it.eq.15) then
    do i=1,4*n
      write(10,*) points(1,i),points(2,i),points(3,i)
    end do
      write(10,*) points(1,1),points(2,1),points(3,1)
   endif
   write(10,*) '  '
   end do
!
   endif
```

minus

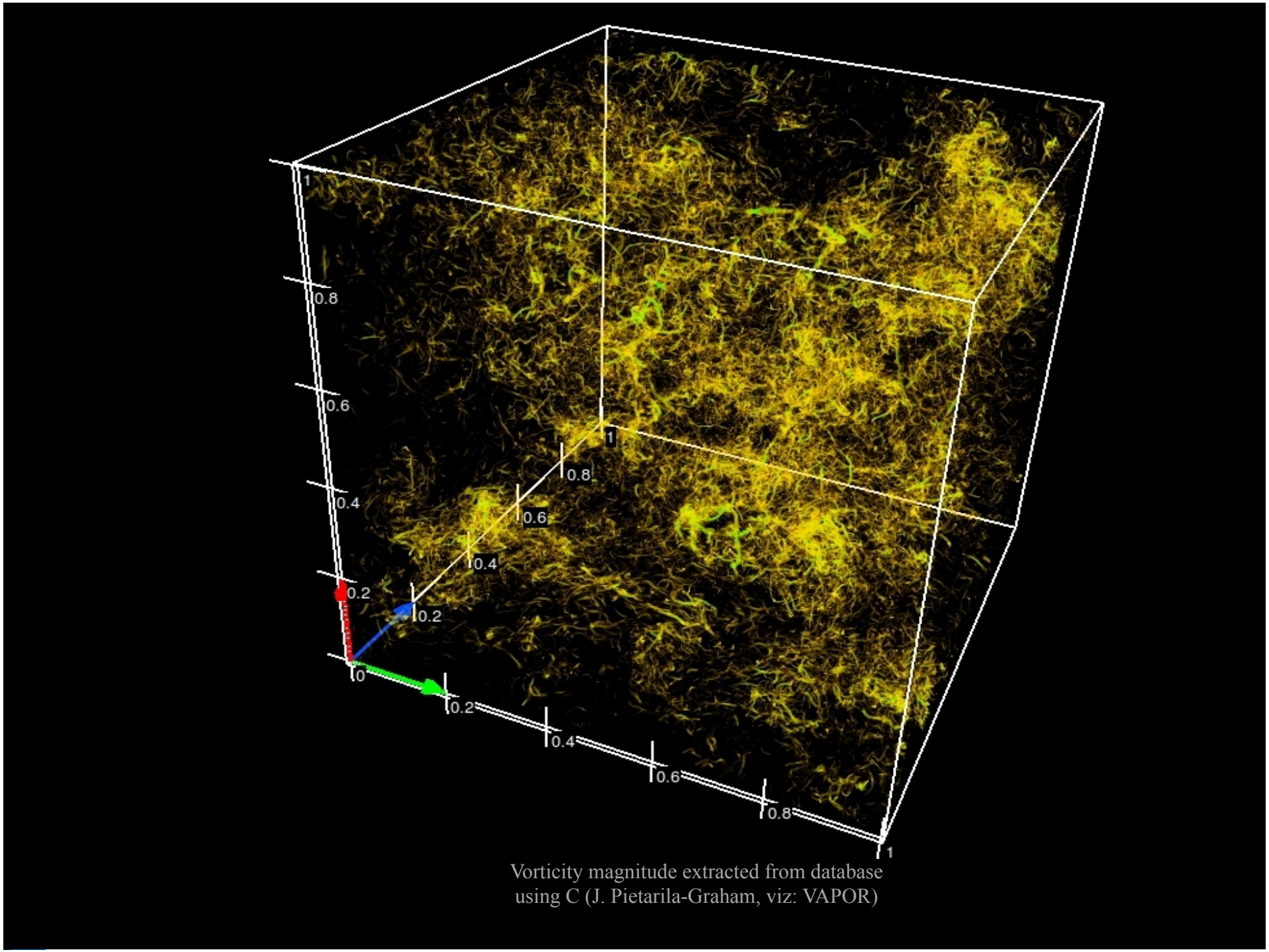
advect backwards in time !

Not possible during DNS



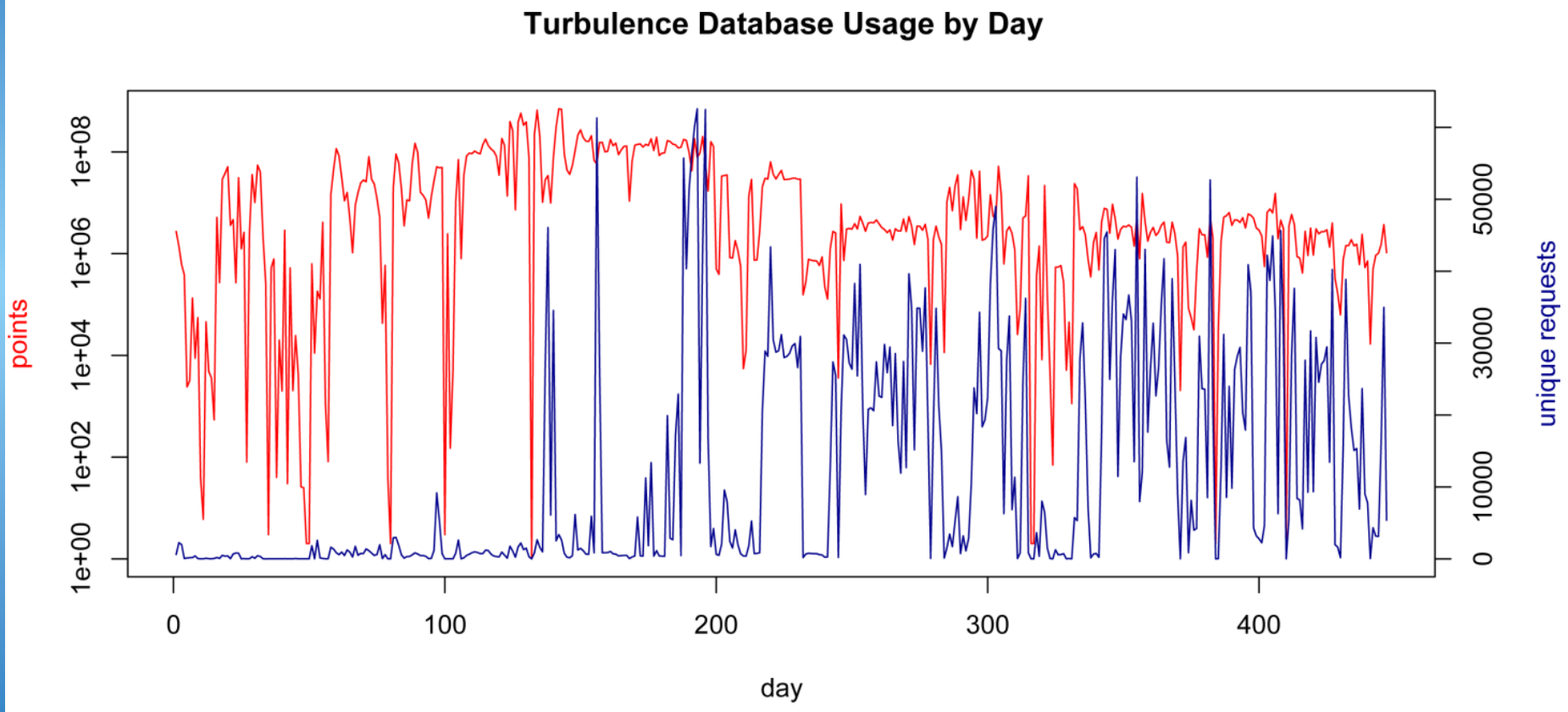


$u(x,y,z_0,t_0)$ extracted from database using Matlab (C. Verhulst)



Vorticity magnitude extracted from database
using C (J. Pietarila-Graham, viz: VAPOR)

Daily Usage

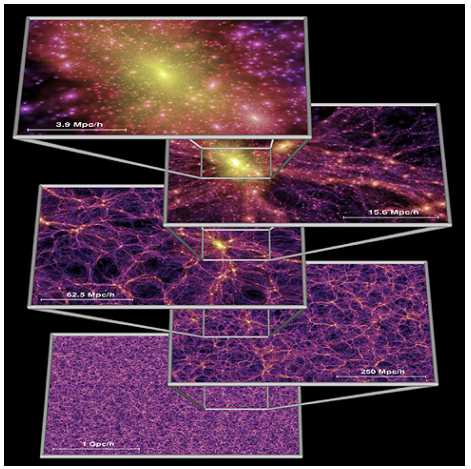


2011: exceeded 100B points, delivered publicly

Cosmological Simulations

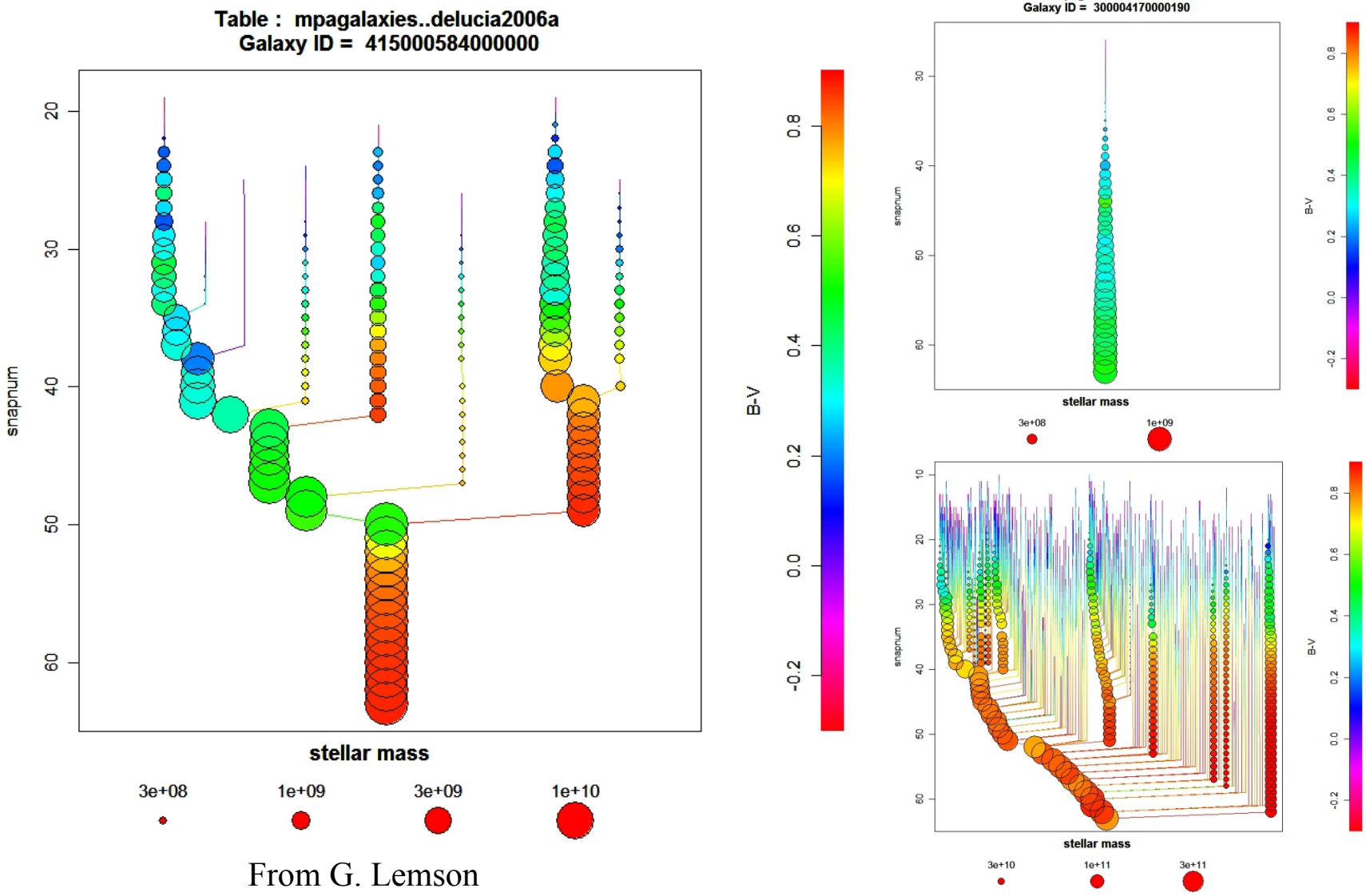
In 2005 cosmological simulations had 10^{10} particles and produced over 30TB of data (Millennium)

- 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



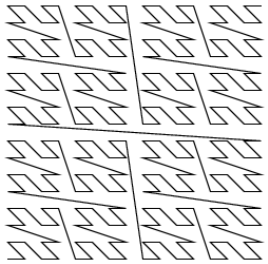
- Today: simulations with 10^{12} particles and PB of output are under way (MillenniumXXL, Silver River, etc)
- Hard to analyze the data afterwards
- What is the best way to compare to real data?

Time evolution: merger trees

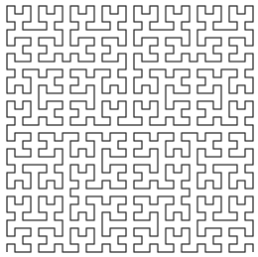


From G. Lemson

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)*

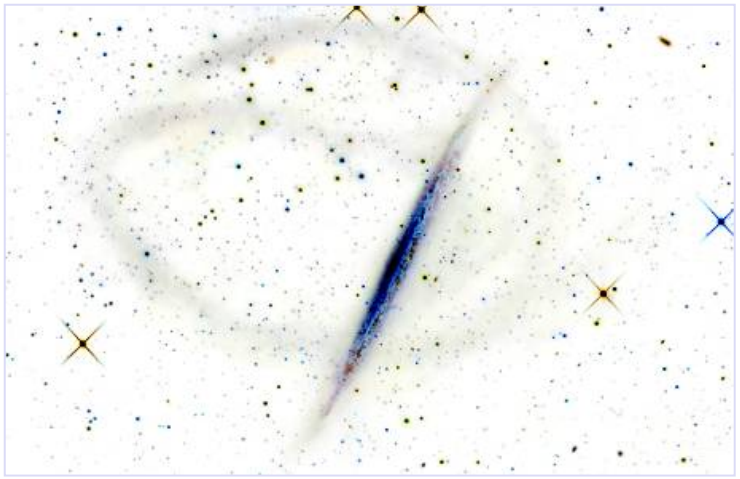


- Random sampling using a RANDOM column
  - *RANDOM from [0,1000000]*

The Milky Way Laboratory

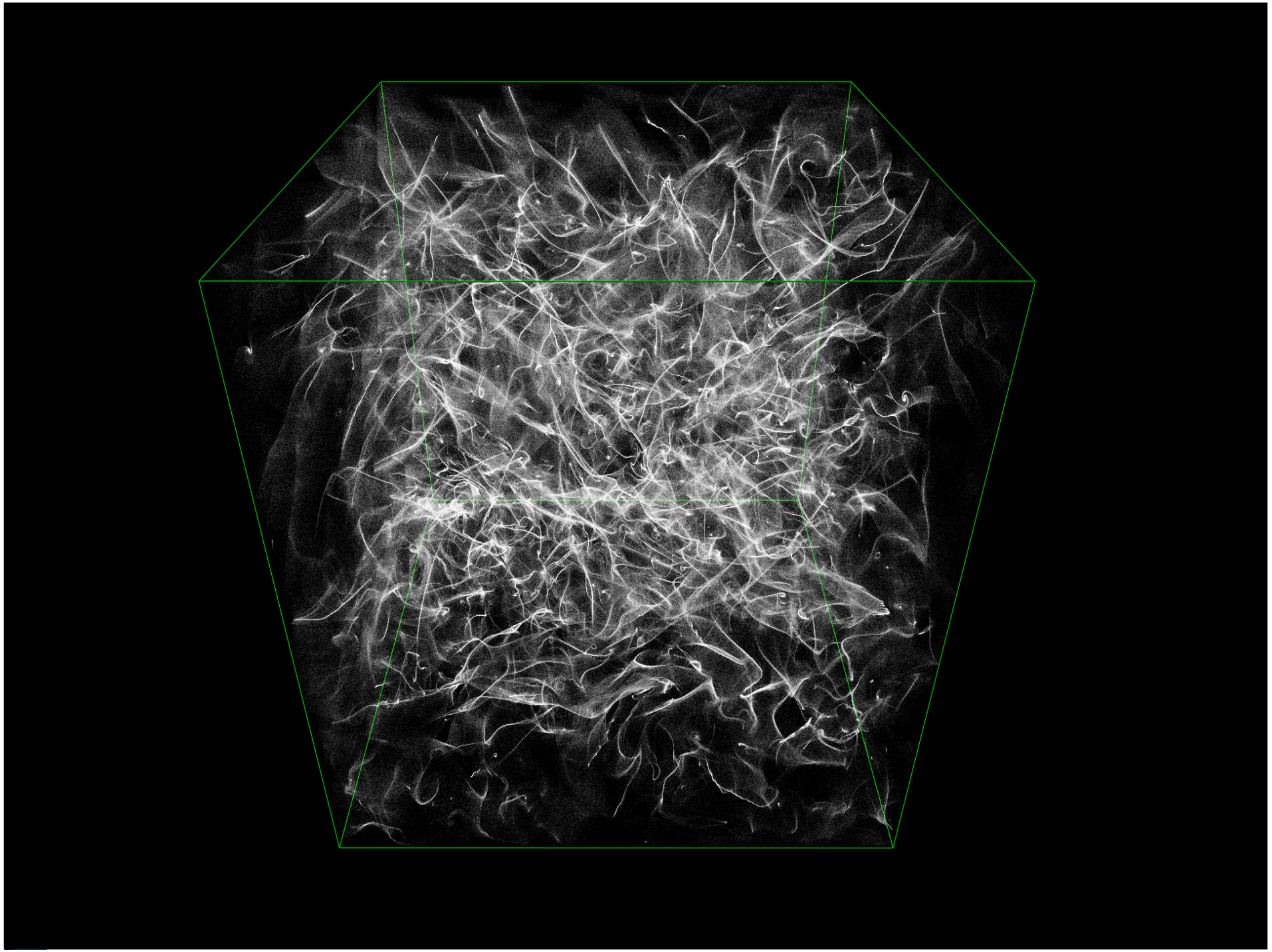
- Use cosmology simulations as an immersive laboratory for general users
- Via Lactea-II (20TB) as prototype, then Silver River (50B particles) as production (15M CPU hours)
- 800+ hi-rez snapshots (2.6PB) => 800TB in DB
- Users can insert test particles (dwarf galaxies) into system and follow trajectories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

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

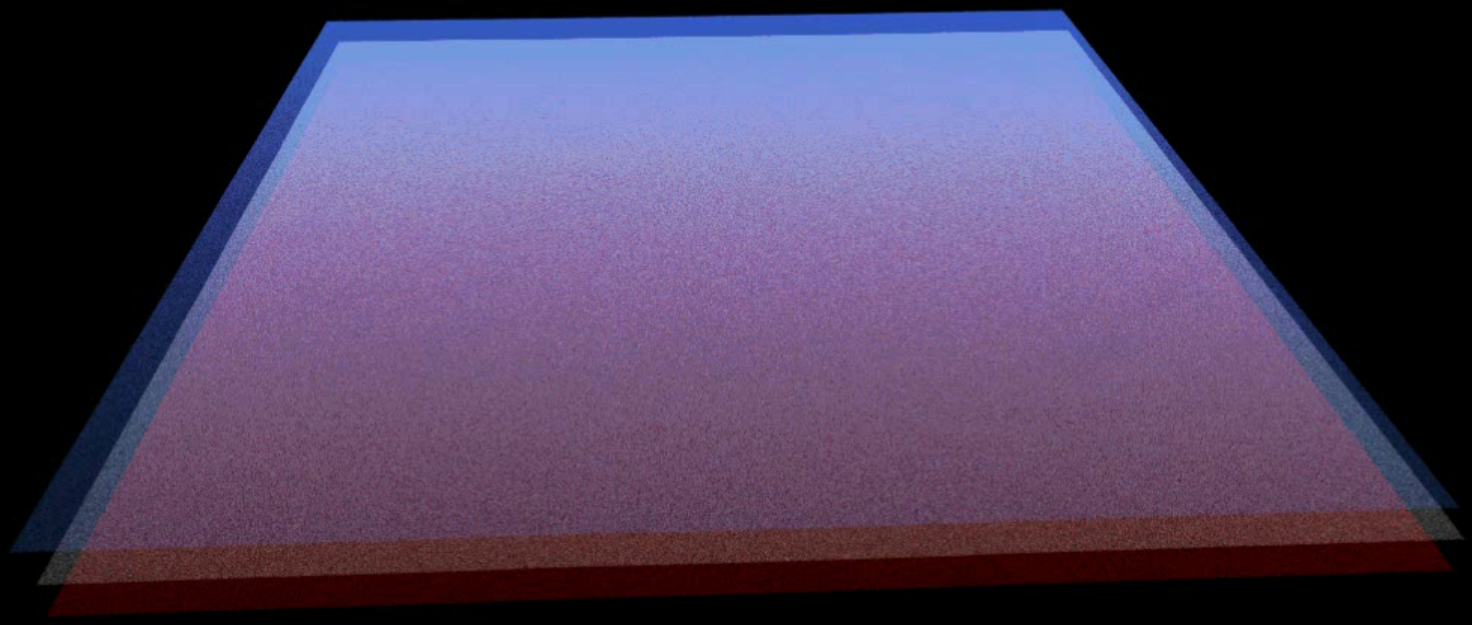


Visualizing Petabytes

- 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

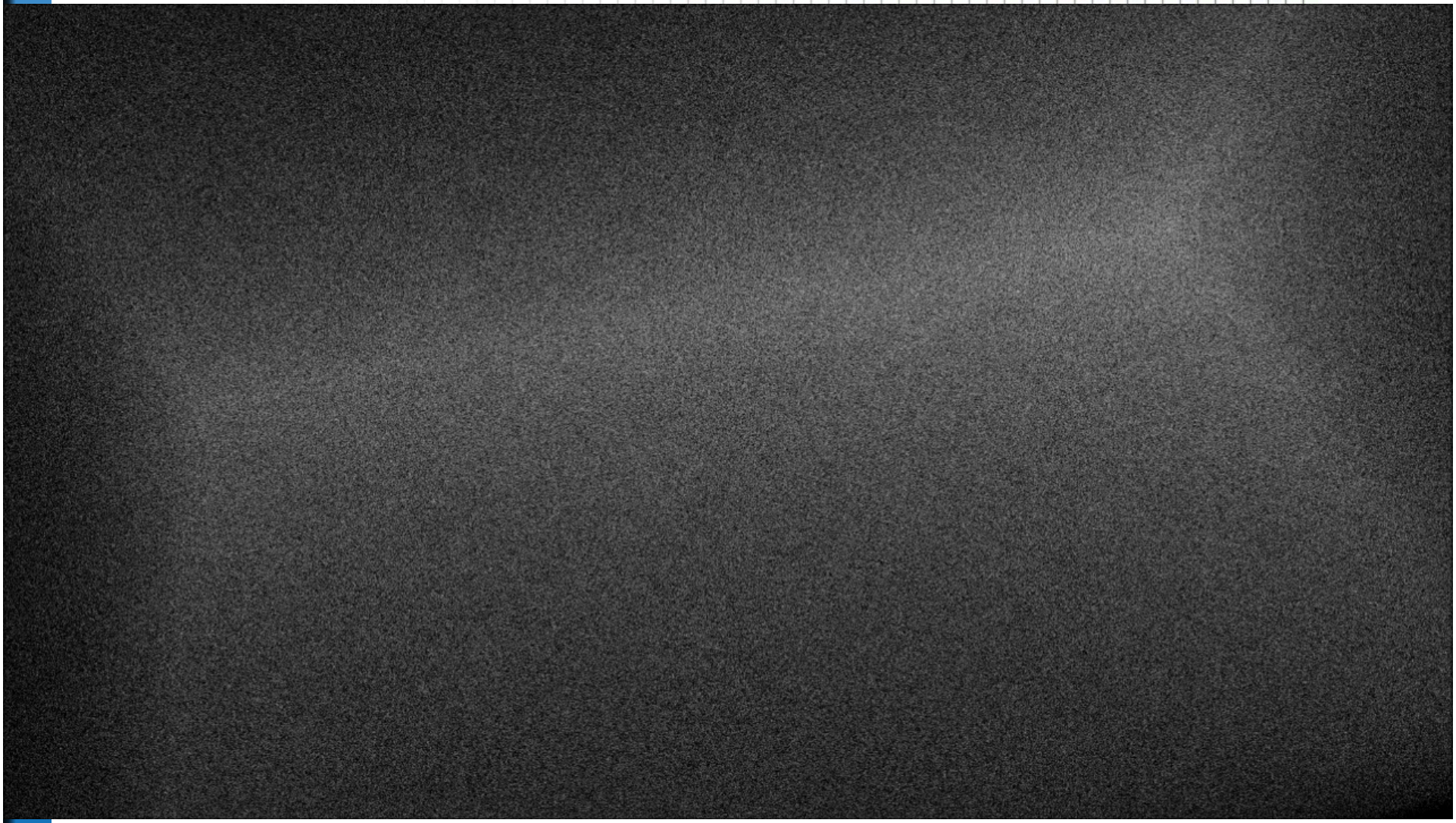


Streaming Visualization of Turbulence

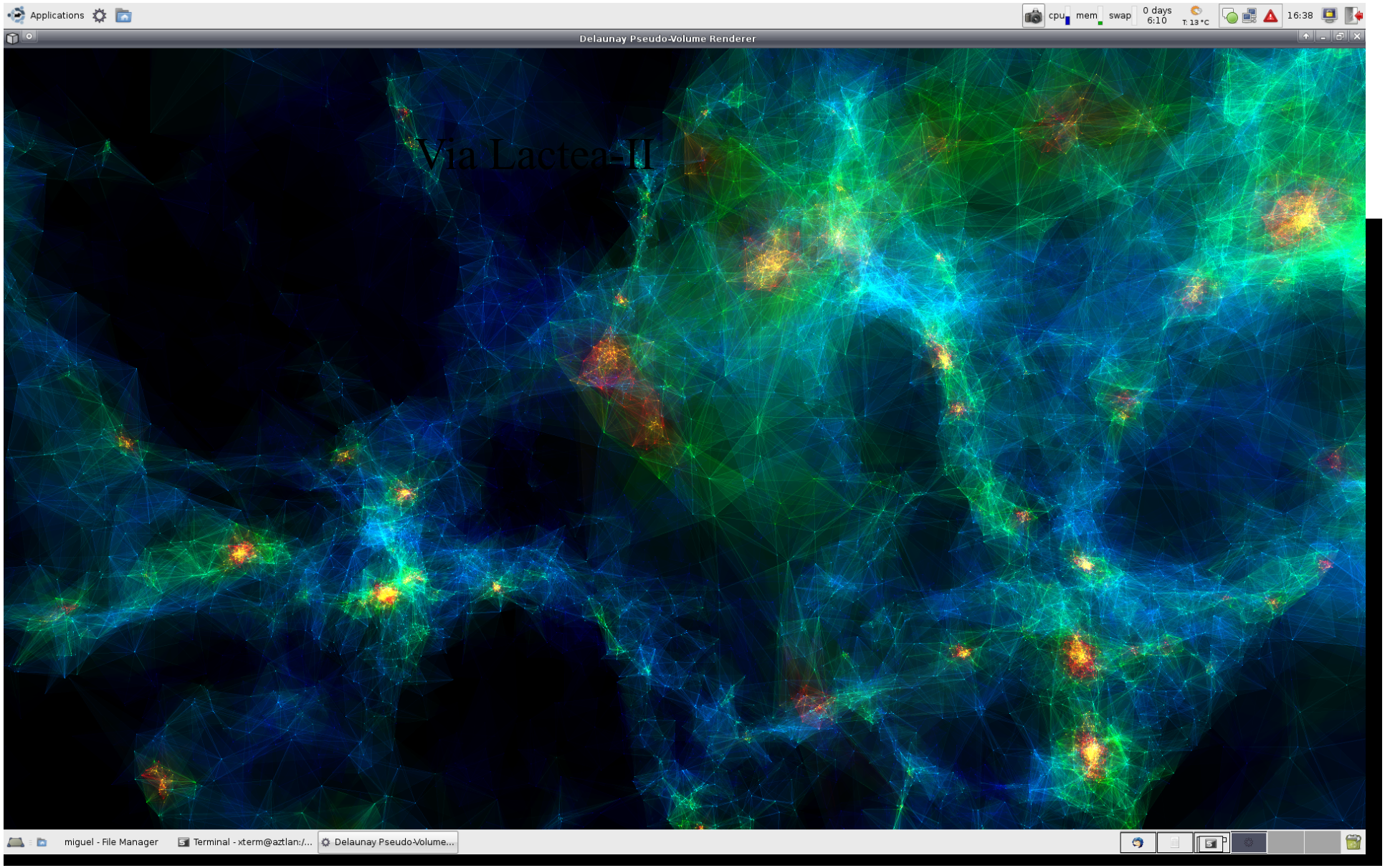


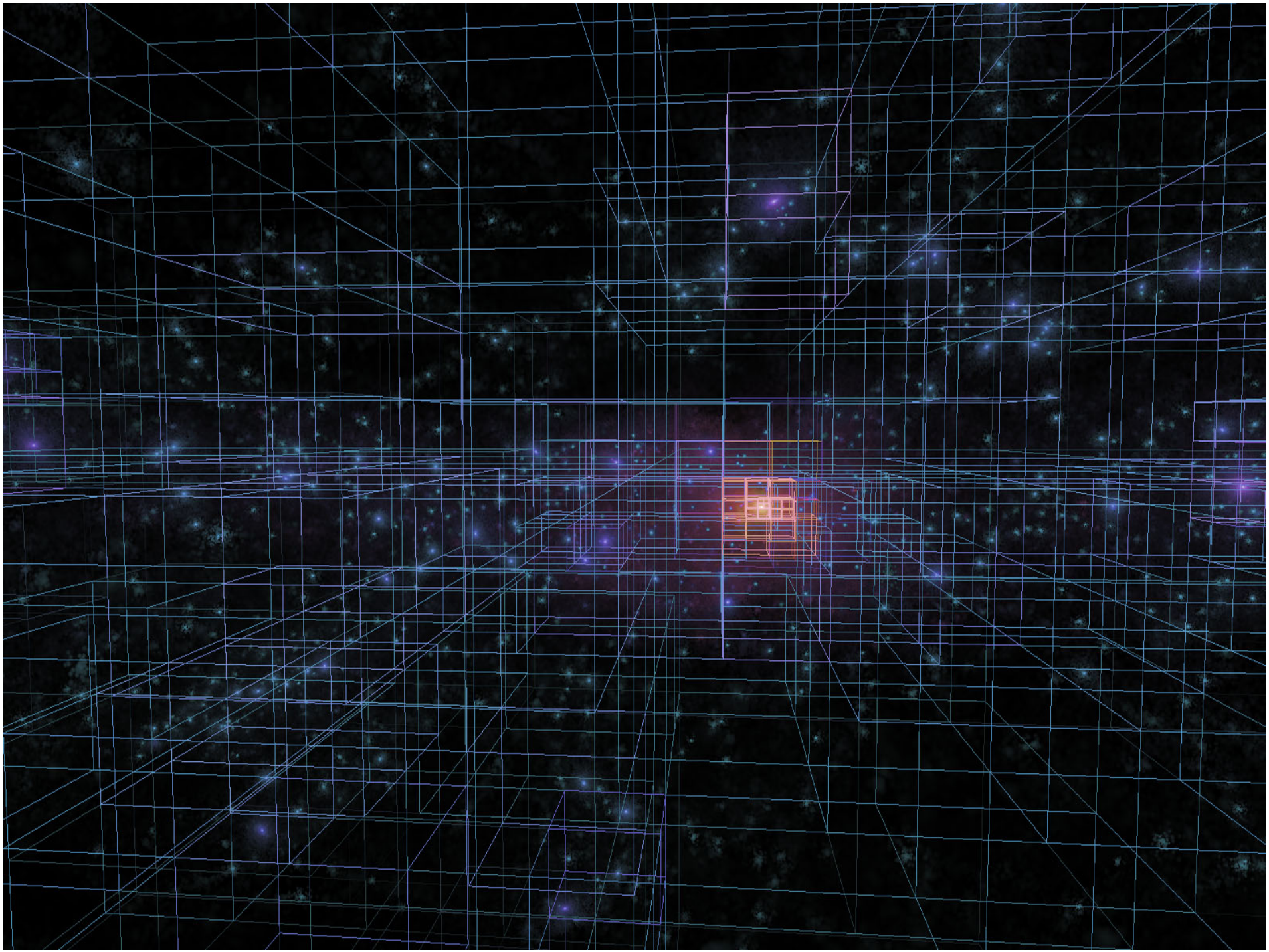
Kai Buerger, Technische Universitat Munich, 24 million particles

Visualization of the Vorticity



Kai Buerger, Technische Universitat Munich


Applications
cpu
mem
swap
0 days
6:10
T: 13 °C
16:38
Delaunay Pseudo-Volume Renderer
Via Lactea-II
miguel - File Manager
Terminal - xterm@aztlan:/...
Delaunay Pseudo-Volume...



DISC Challenges

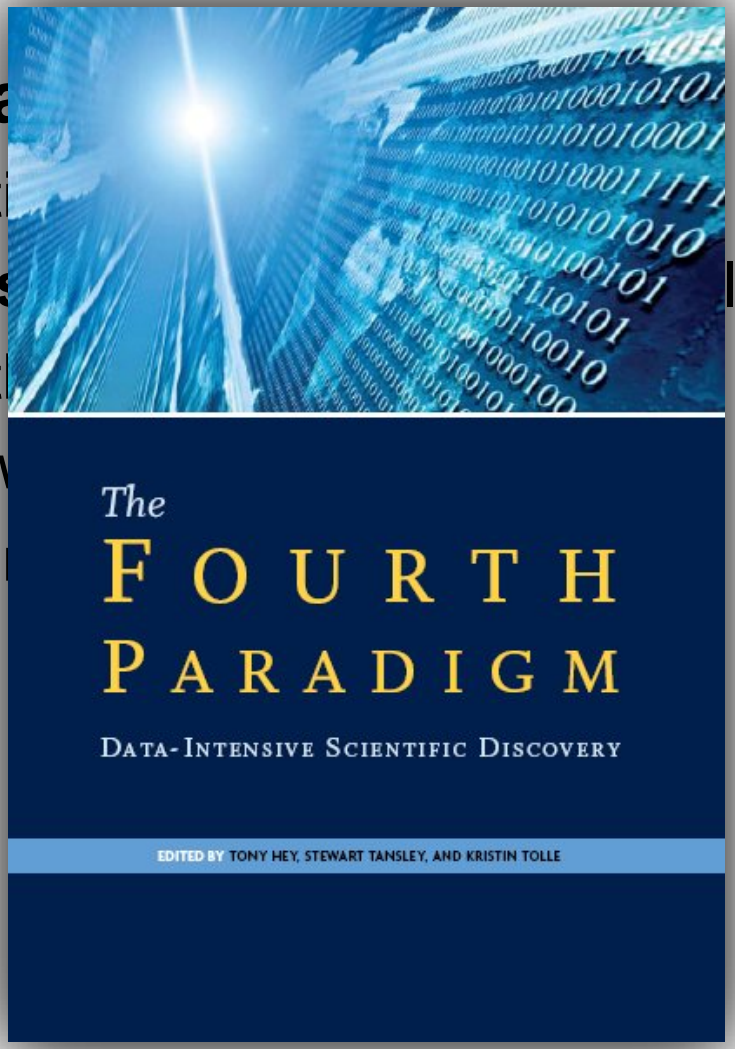
DISC: Data Intensive Scalable Computing

- Where are the systems challenges today?
  - *Storage size*
  - *System balance*
  - *Data mobility*
  - *Statistical algorithms*
  - *Scalability/power*
- What is being done to soften it?
  - *Scale up or scale out…*
  - *New SW platforms emerging*
  - *Testing disruptive technologies*
  - *New streaming algorithms*

Gray's Laws of Data Engineering

Jim Gra

- Scienti around **data**
- Need s lysis
- Take t
- Start w
- Go fro



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)

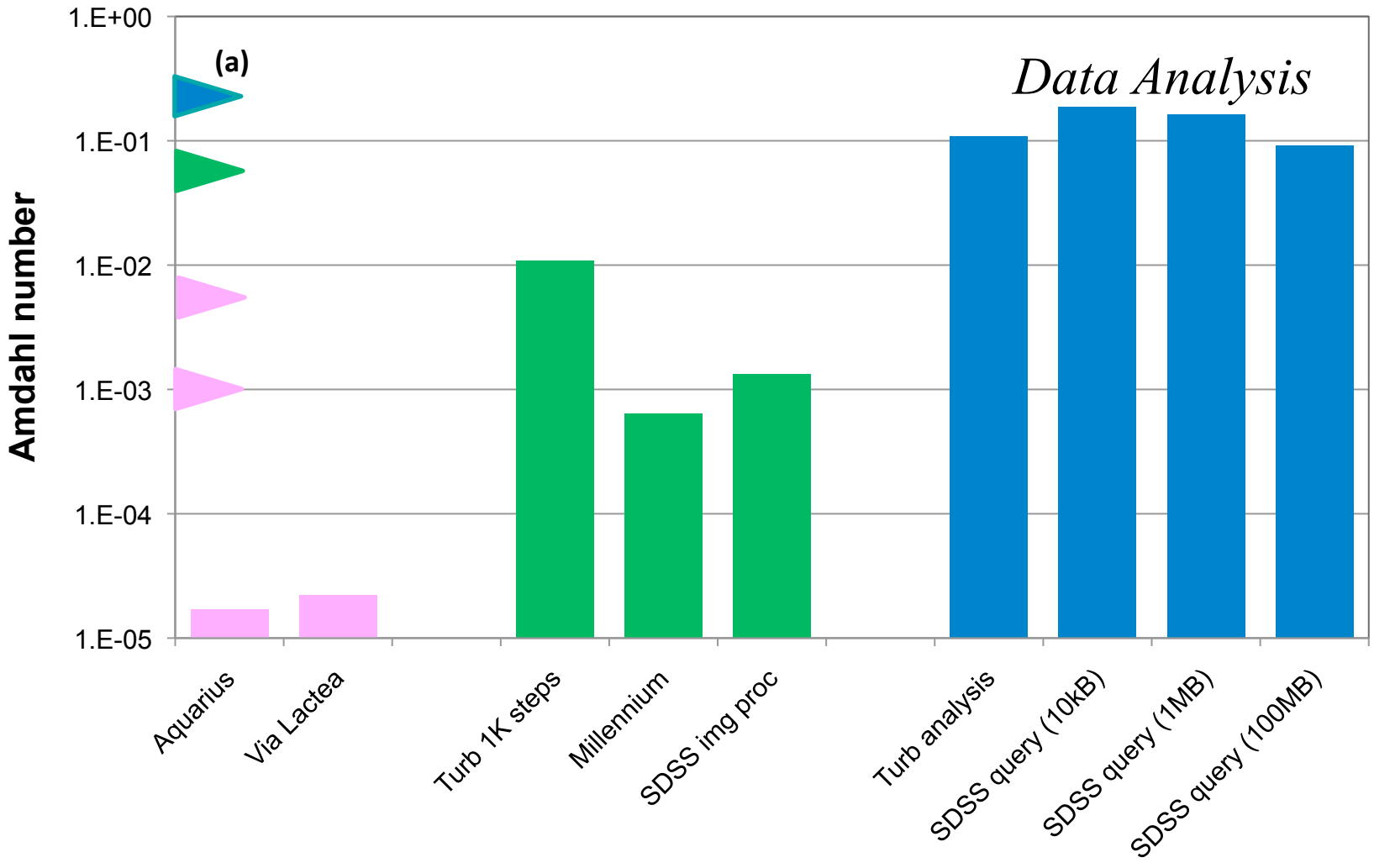
iii. One byte of memory per one instruction/sec (MEM)

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

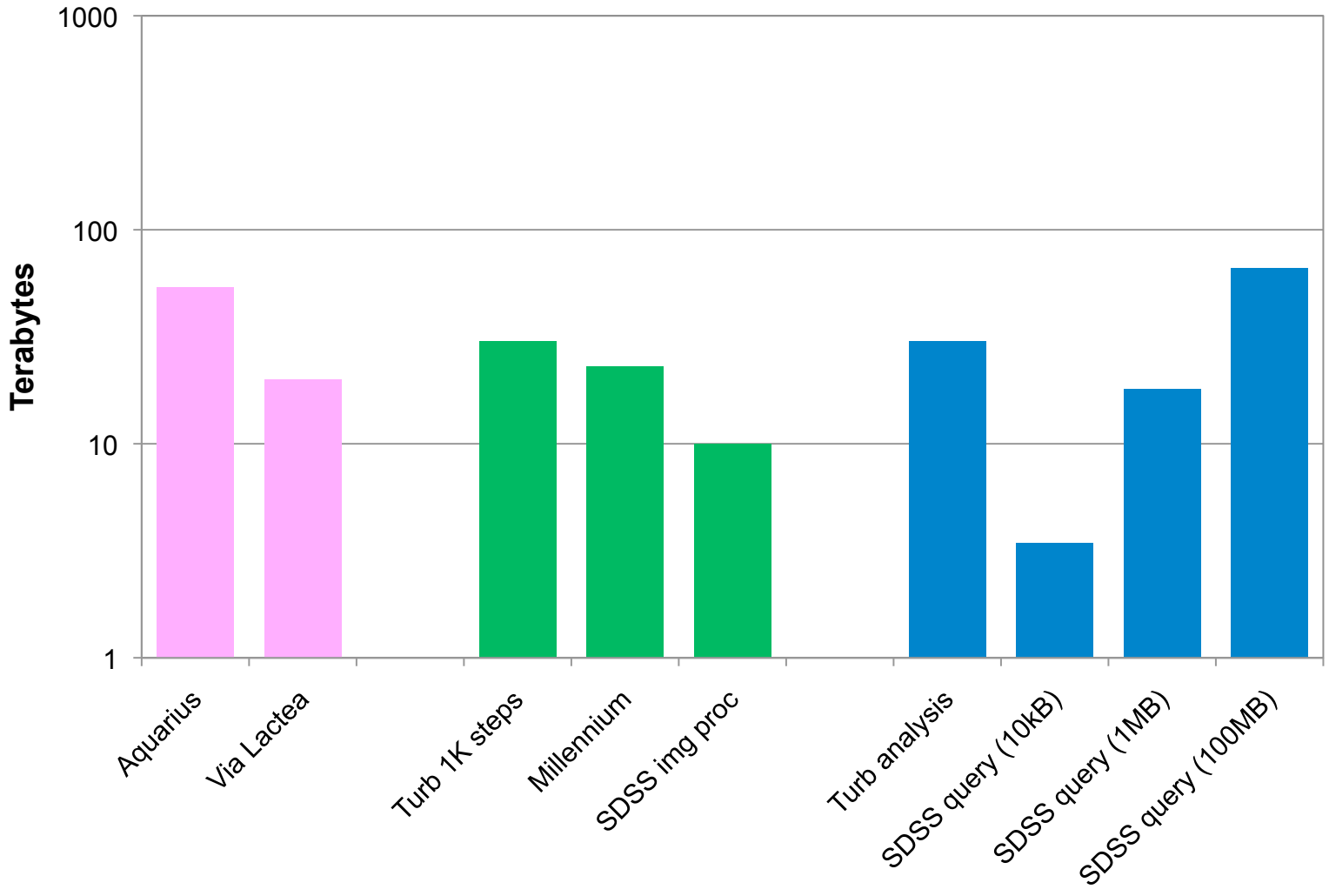
Typical Amdahl Numbers

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

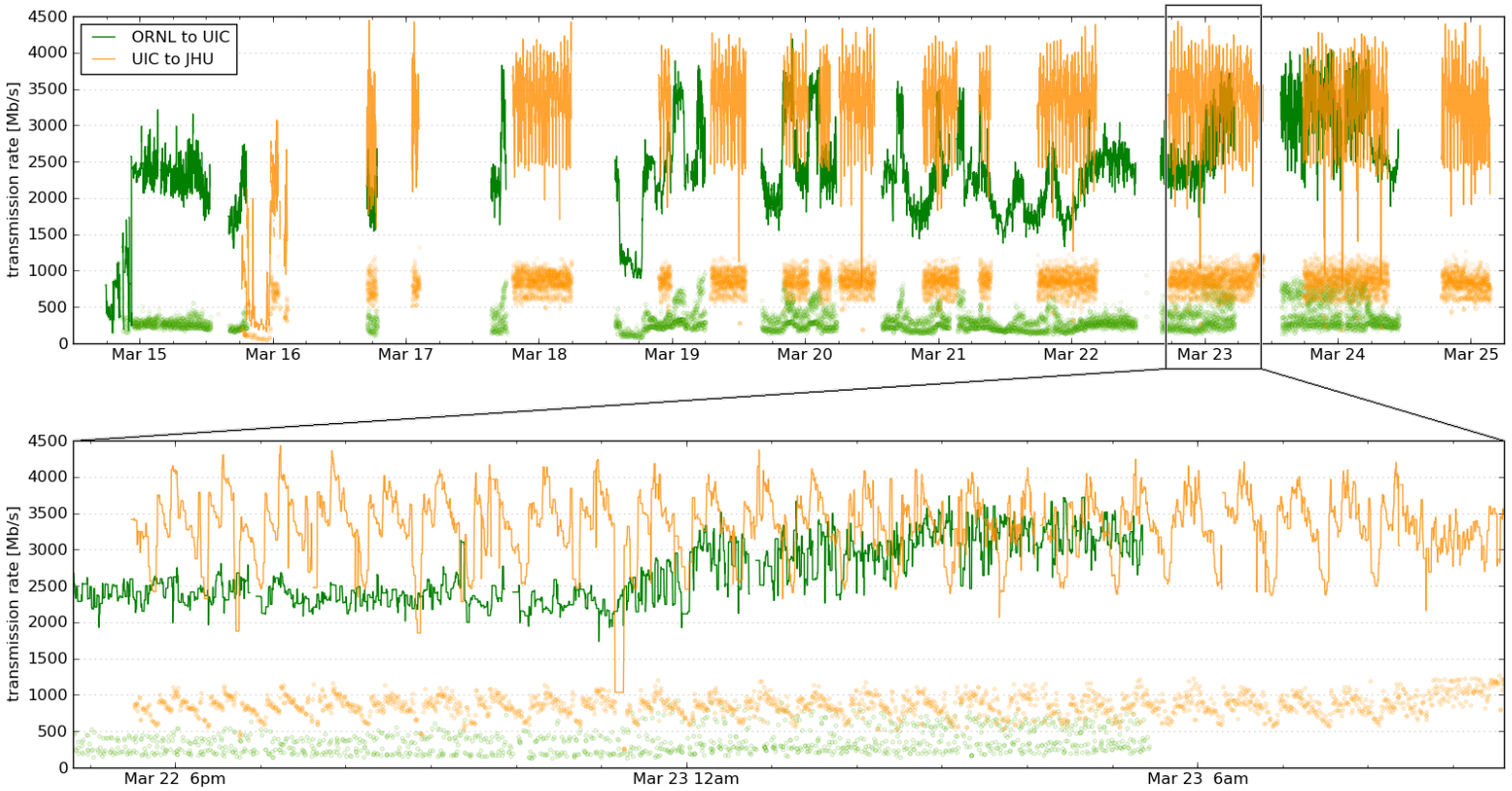


DISC Needs Today

- 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 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

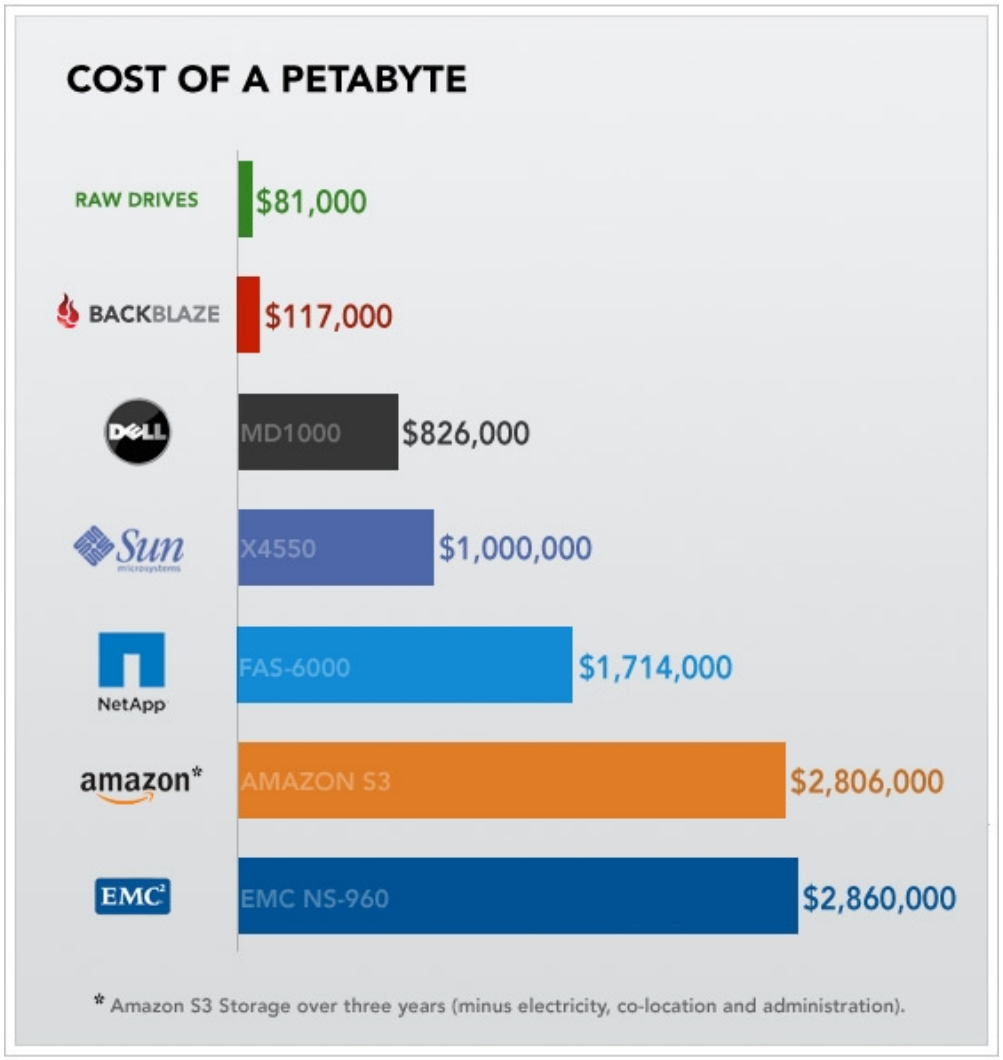
Stu Feldman: Extreme computing is about tradeoffs

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…

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: 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

	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

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 the cloud (private or public)
- RandomIO moves to Solid State Disks
- 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 to do it
- We are still building our own… over and over

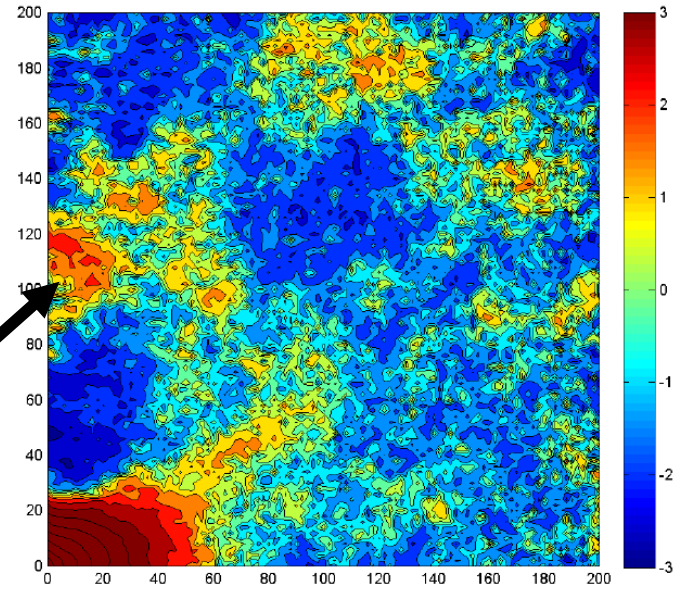
Cyberbricks?

- 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 for under $30K
- Using the GPUs for data mining:
  - *6.4B multidimensional regressions in 5 minutes over 1.2TB*
  - *Ported RF module from R in C#/CUDA*

Szalay, Bell, Huang, Terzis, White (Hotpower-09)

Correlation Function on GPUs

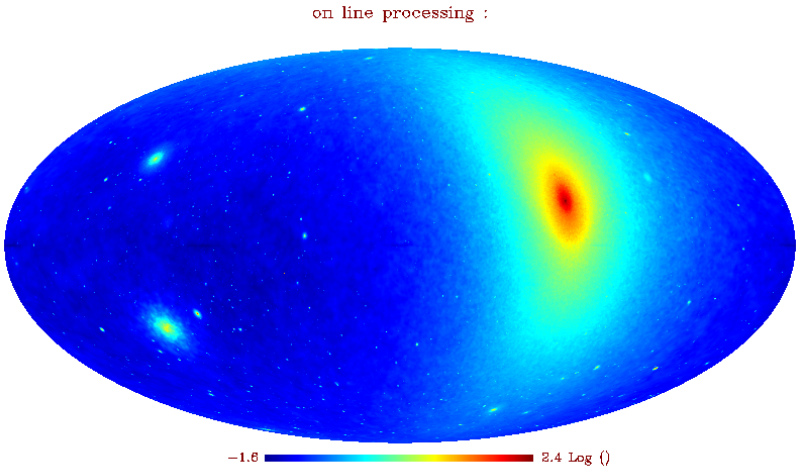
- We need to reconsider the N logN only approach
- Once we can run 100K threads, maybe running SIMD N^2 on smaller partitions is also acceptable
- Recent JHU effort on integrating CUDA with SQL Server, using SQL UDF
- Galaxy spatial correlations: 600 trillion real and random galaxy pairs using brute force N^2
- Much faster than the tree codes!
  - *This is because high resolution was needed…*

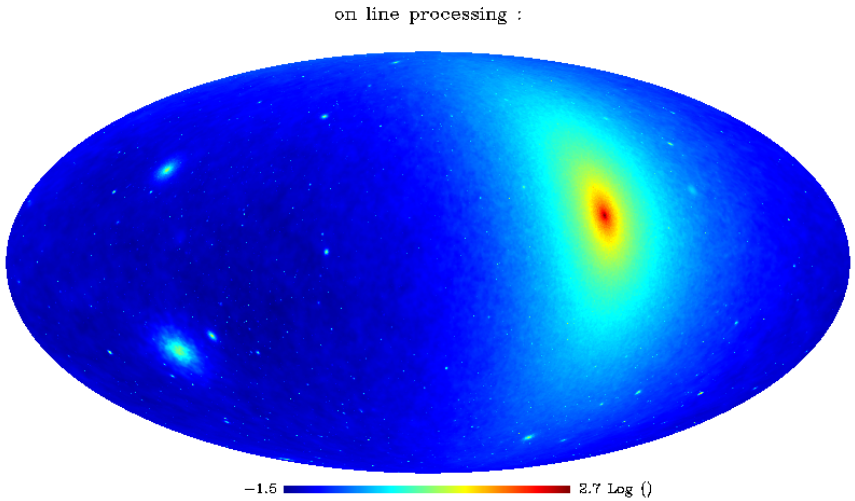


Tian, Budavari, Neyrinck, Szalay 2010

Via Lactea-II

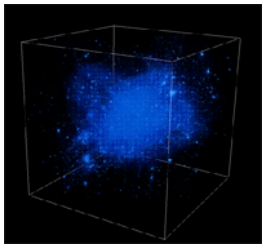
- Computing the gamma-ray annihilation map
- Goal: build an interactive service
- Original calculation:
  8 hrs/image
- GPU + Open GL:
  50 sec





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*
  - *Impact of demographic changes in science*
- 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
- Changing sociology
- From hypothesis-driven to data-driven science
- We need new instruments: “microscopes” and “telescopes” for data
- There is also a problem on the “long tail”
- Same problems present in business and society
- Data changes not only science, but society
- A new, Fourth Paradigm of Science is emerging…

A convergence of statistics, computer science, physical and life sciences…..