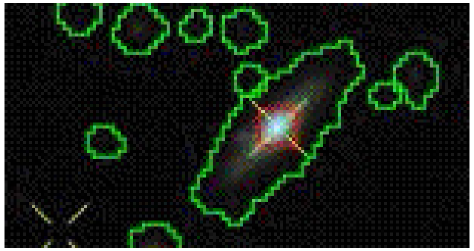
New Challenges in Petascale Scientific Databases

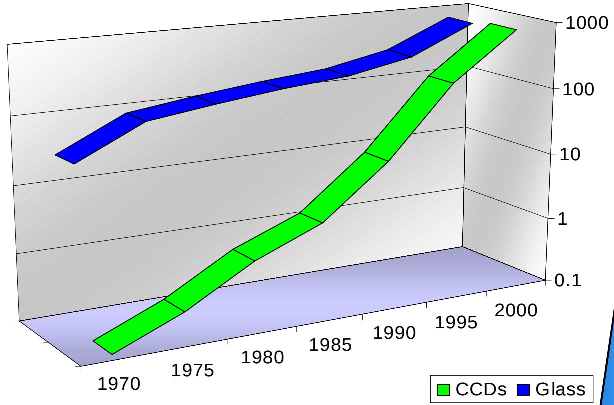
Alex Szalay

The Johns Hopkins University



Living in an Exponential World

- Scientific data doubles every year
  - *caused by successive generations of inexpensive sensors + exponentially faster computing*



- Changes the nature of scientific computing
- Cuts across disciplines (eScience)
- It becomes increasingly harder to extract knowledge
- 20% of the world's servers go into huge data centers by the "Big 5"
  - *Google, Microsoft, Yahoo, Amazon, eBay*
- So it is not only the scientific data!

Astronomy Trends

CMB Surveys (pixels)

- 1990 COBE 1000
- 2000 Boomerang 10,000
- 2002 CBI 50,000
- 2003 WMAP 1 Million
- 2008 Planck 10 Million

Angular Galaxy Surveys (obj)

- 1970 Lick 1M
- 1990 APM 2M
- 2005 SDSS 200M
- 2009 PANSTARRS 1200M
- 2015 LSST 3000M

Time Domain

- QUEST
- SDSS Extension survey
- Dark Energy Camera
- PanStarrs
- SNAP…
- LSST…

Galaxy Redshift Surveys (obj)

- 1986 CfA 3500
- 1996 LCRS 23000
- 2003 2dF 250000
- 2005 SDSS 750000

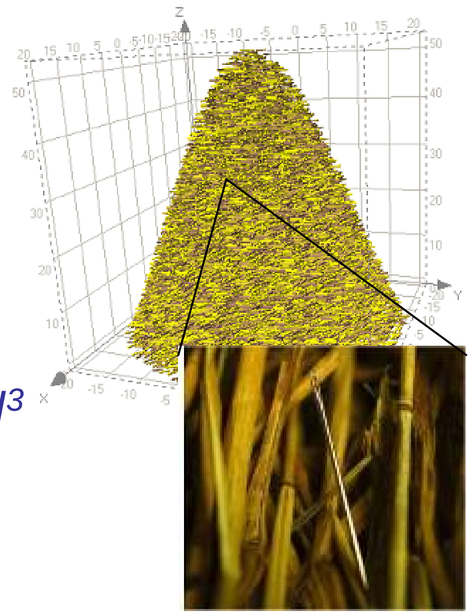
Petabytes/year by the end of the decade…

Collecting Data

- Very extended distribution of data sets:

  data on all scales!
- Most datasets are small, and manually maintained (Excel spreadsheets)
- Total amount of data dominated by the other end (large multi-TB archive facilities)
- Most bytes today are collected via electronic sensors

Next-Generation Data Analysis



- Looking for
  - *Needles in haystacks – the Higgs particle*
  - *Haystacks: Dark matter, Dark energy*
- Needles are easier than haystacks
- 'Optimal' statistics have poor scaling
  - *Correlation functions are N^2, likelihood techniques N^3*
  - *For large data sets main errors are not statistical*
- As data and computers grow with Moore's Law, we can only keep up with $N \log N$
- A way out: sufficient statistics?
  - *Discard notion of optimal (data is fuzzy, answers are approximate)*
  - *Don't assume infinite computational resources or memory*
- Requires combination of statistics & computer science
  - *Clever data structures, new, randomized algorithms*

Data Intensive Scalable Computing

- The nature of scientific computing is changing
- It is about the data
- Adding more CPUs makes the IO lag further behind
- Getting even worse with multi-core
- We need more balanced architectures

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)

iv. One IO per 50,000 instructions (IO)

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

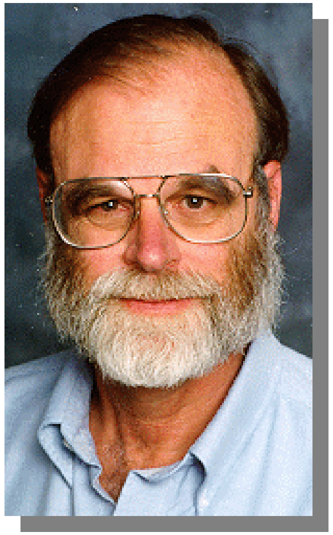
For a Blue Gene the BW=0.013, MEM=0.471.

For the JHU cluster BW=0.664, MEM=1.099

Gray’s Laws of Data Engineering

Jim Gray:

- Scientific computing is revolving around **data**
- Need **scale-out** solution for analysis
- Take the **analysis to the data**!
- Start with “**20 queries**”
- Go from “**working to working**”



Reference Applicatons

Several key projects at JHU

- **SDSS:** *10TB total, 3TB in DB, soon 10TB, in use for 6 years*
- **NVO Apps:** *~5TB, many B rows, in use for 4 years*
- **PanStarrs:** *80TB by 2009, 300+ TB by 2012*
- **Immersive Turbulence***: 30TB now, 300TB next year, can change how we use HPC simulations worldwide*
- **SkyQuery***: perform fast spatial joins on the largest astronomy catalogs / replicate multi-TB datasets 20 times for much faster query performance (1Bx1B in 3 mins)*
- **OncoSpace***: 350TB of radiation oncology images today, 1PB in two years, to be analyzed on the fly*
- ***Sensor Networks:*** *200M measurements now, billions next year, forming complex relationships*

Sloan Digital Sky Survey

Goal

Create the most detailed map of the Northern sky

"The Cosmic Genome Project"

Two surveys in one

Photometric survey in 5 bands

Spectroscopic redshift survey

Automated data reduction

150 man-years of development

High data volume

40 TB of raw data

5 TB processed catalogs

Data is public

2.5 Terapixels of images

Now officially FINISHED

The University of Chicago
Princeton University
The Johns Hopkins University
The University of Washington
New Mexico State University
Fermi National Accelerator Laboratory
US Naval Observatory
The Japanese Participation Group
The Institute for Advanced Study
Max Planck Inst, Heidelberg

Sloan Foundation, NSF, DOE, NASA

SDSS Now Finished!

- As of May 15, 2008 SDSS is officially complete
- Final data release (DR7.2) later this year
- Final archiving of the data in progress
  - *Paper archive at U. Chicago Library*
  - *Digital Archive at JHU Library*
- Archive will contain >100TB
  - *All raw data*
  - *All processed/calibrated data*
  - *All version of the database*
  - *Full email archive and technical drawings*
  - *Full software code repository*

Database Challenges

- Loading (and scrubbing) the Data
- Organizing the Data (20 queries, self-documenting)
- Accessing the Data (small and large queries, visual)
- Delivering the Data (workbench)
- Analyzing the Data (spatial, scaling…)

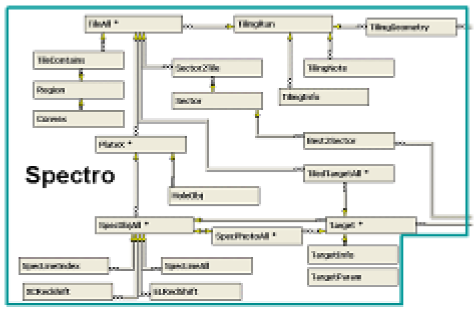
MyDB: Workbench

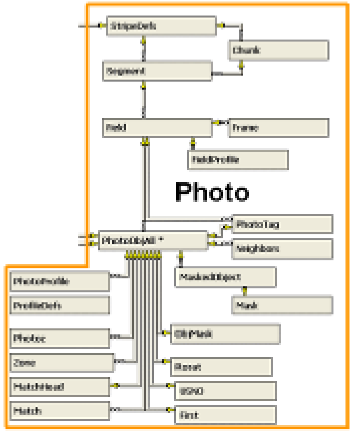
- Need to register ‘power users’, with their own DB
- Query output goes to ‘MyDB’
- Can be joined with source database
- Results are materialized from MyDB upon request
- Users can do:
  - *Insert, Drop, Create, Select Into, Functions, Procedures*
  - *Publish their tables to a group area*
- Data delivery via the CASJobs (C# WS)

=> Sending analysis to the data!

User Level Services

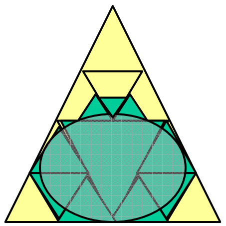
- Three different applications on top of the same core
  - *Finding Chart (arbitrary size)*
  - *Navigate (fixed size, clickable navigation)*
  - *Image List (display many postage stamps on same page)*
- Linked to
  - *One another*
  - *Image Explorer (link to complex schema)*
  - *On-line documentation*





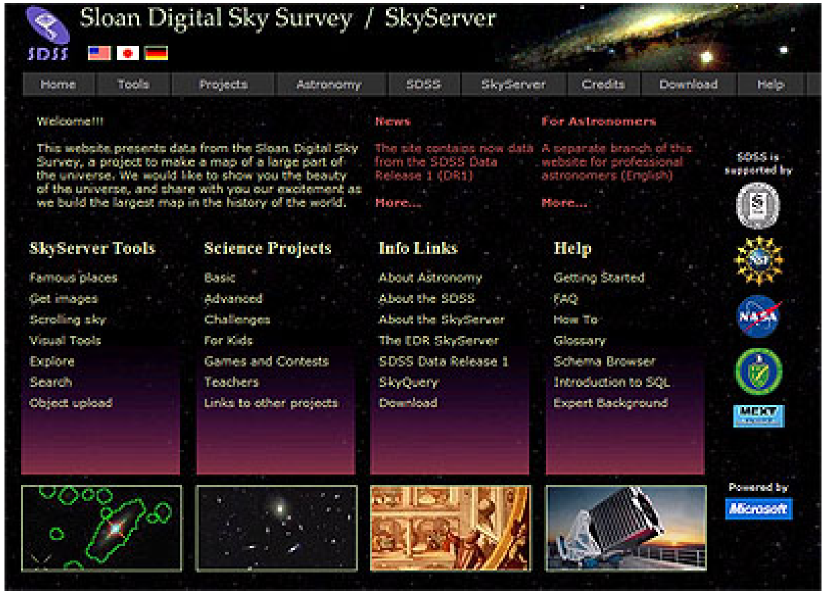
Geometries

- SDSS has lots of complex boundaries
  - *60,000+ regions*
  - *6M masks, represented as spherical polygons*
- A GIS-like library built in C++ and SQL
- Now converted to C# for direct plugin into SQLServer 2005 (17 times faster than C++)
- Precompute arcs and store in database for rendering
- Functions for point in polygon, intersecting polygons, polygons covering points, all points in polygon
- Using spherical quadtrees (HTM)



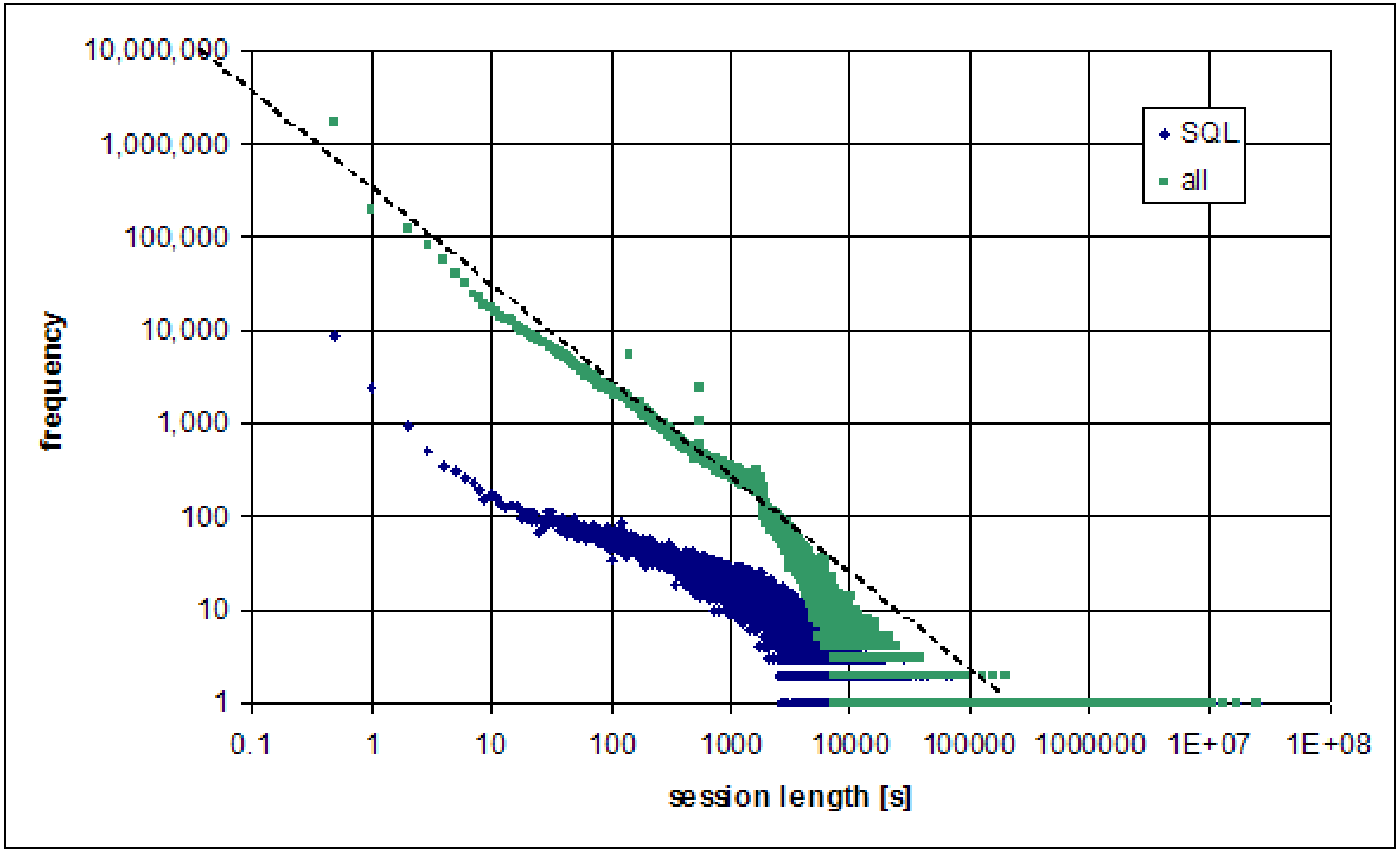
Public Use of the SkyServer

- Prototype in data publishing
  - *450 million web hits in 6 years*
  - *930,000 distinct users vs 15,000 astronomers*
  - *Delivered 50,000 hours of lectures to high schools*
  - *Delivered >100B rows of data*
  - *Everything is a power law*



- Interactive workbench
  - *Casjobs/MyDB*
  - *Power users get their own database, no time limits*
  - *They can store their data server-side, link to main data*
  - *They can share results with each other*
  - *Simple analysis tools (plots, etc)*
  - *Over 1,600 'power users'*

Skyserver Sessions



Vic Singh et al (Stanford/ MSR)

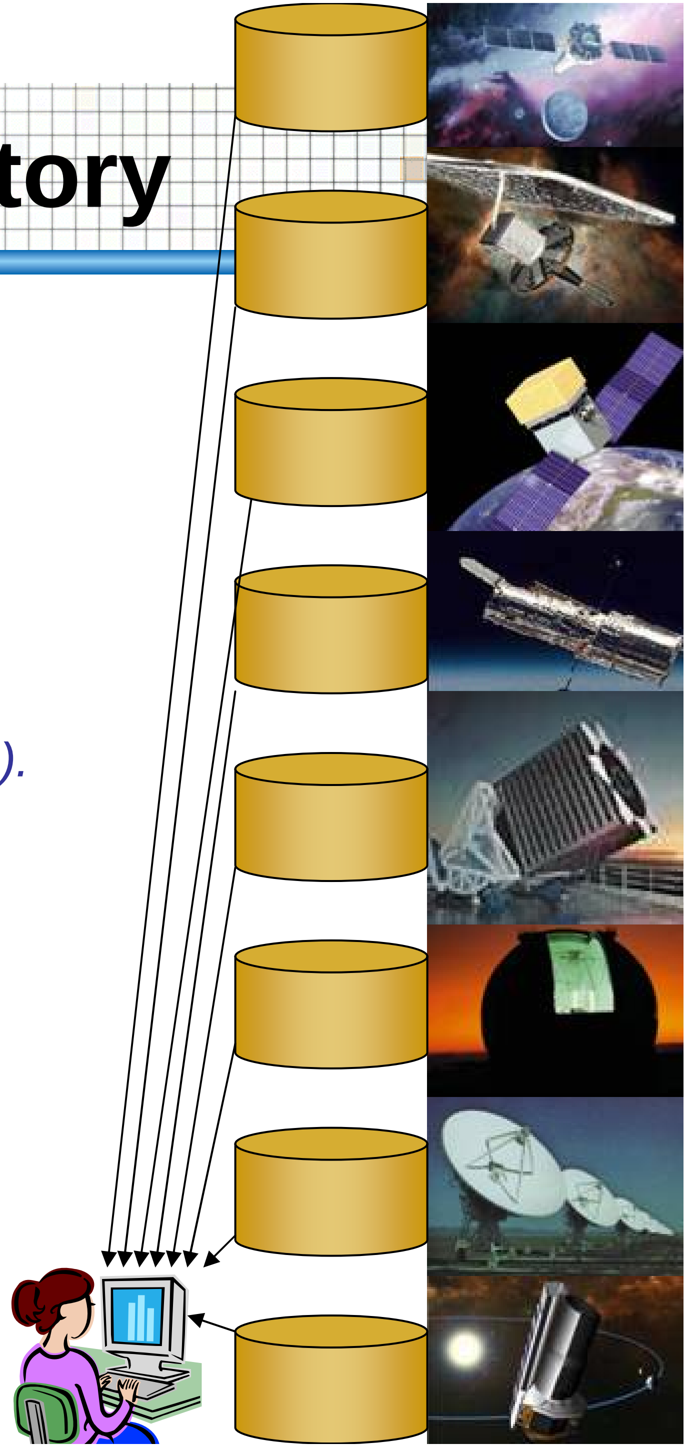
Why Is Astronomy Special?

- Especially attractive for the wide public
- Community is not very large
- **It has no commercial value**
  - *No privacy concerns, freely share results with others*
  - *Great for experimenting with algorithms*
- It is real and well documented
  - *High-dimensional (with confidenc*
  - *Spatial, temporal*
- Diverse and distributed
  - *Many different instruments from m places and times*
- The questions are interesting
- There is a lot of it (soon petaby

WORTHLESS!

The Virtual Observatory

- Premise: most data is (or could be online)
- The Internet is the world's best telescope:
  - *It has data on every part of the sky*
  - *In every measured spectral band: optical, x-ray, radio..*
  - *As deep as the best instruments (2 years ago).*
  - *It is up when you are up*
  - *The "seeing" is always great*
  - *It's a smart telescope: links objects and data to literature on them*
- Software became the capital expense
  - *Share, standardize, reuse..*



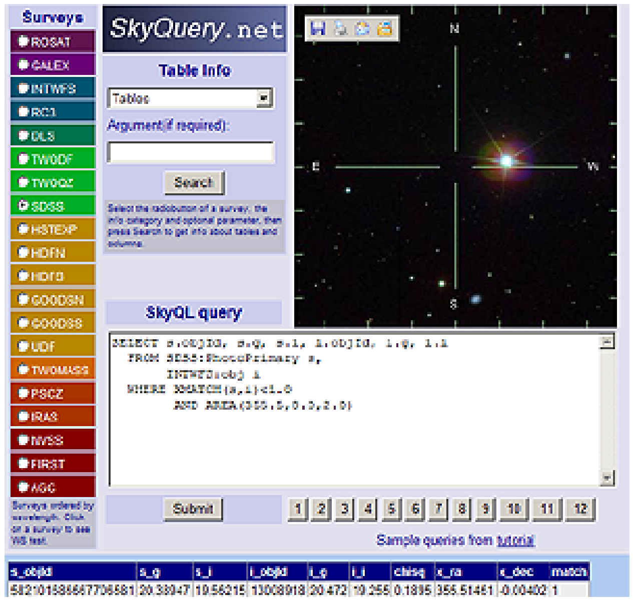
National Virtual Observatory

- NSF ITR project, “Building the Framework for the National Virtual Observatory” is a collaboration of 17 funded and 3 unfunded organizations
  - *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

SkyQuery

- Distributed Query tool using a set of web services
- Many astronomy archives from Pasadena, Chicago, Baltimore, Cambridge (England).
- Implemented in C# and .NET
- After 6 months users wanted to perform joins between catalogs of ~1B cardinality
- Current time for such queries is 1.2h
- We need a parallel engine
- With 20 servers we can deliver 5 min turnaround for these joins



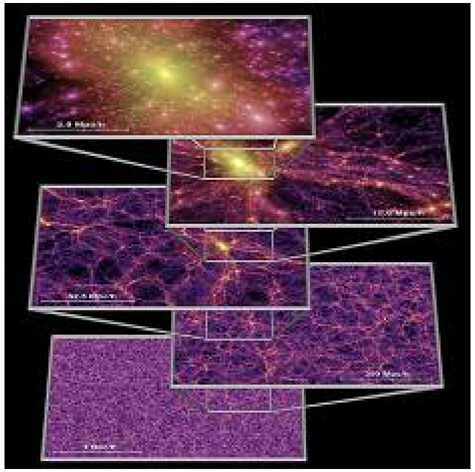
SkyQuery: Interesting Patterns

- Sequential crossmatch of large data sets
  - *Fuzzy spatial join of 1B x 1B*
- Several sequential algorithms, require sorting
- Can be easily parallelized
- Current performance
  - *1.2 hours for 1B x 1B on a single server over whole sky*
  - *Expect 20-fold improvement on SQL cluster*
- How to deal with “success”?
  - *Many users, more and more random access*
- Ferris Wheel
  - *Circular “scan machine”, you get on any time, off after one circle*
  - *Uses only sequential reads*
  - *Can be distributed through synchronizing (w. Grossman)*

Simulations

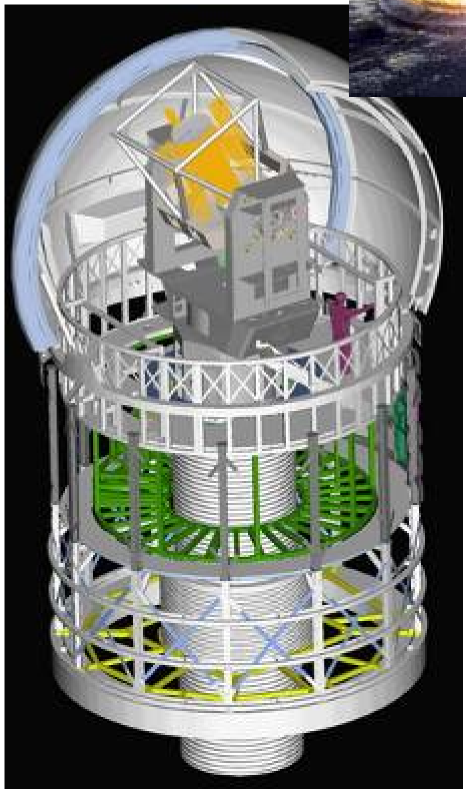
Cosmological simulations have 10^9 particles and produce over 30TB of data (Millennium)

- 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



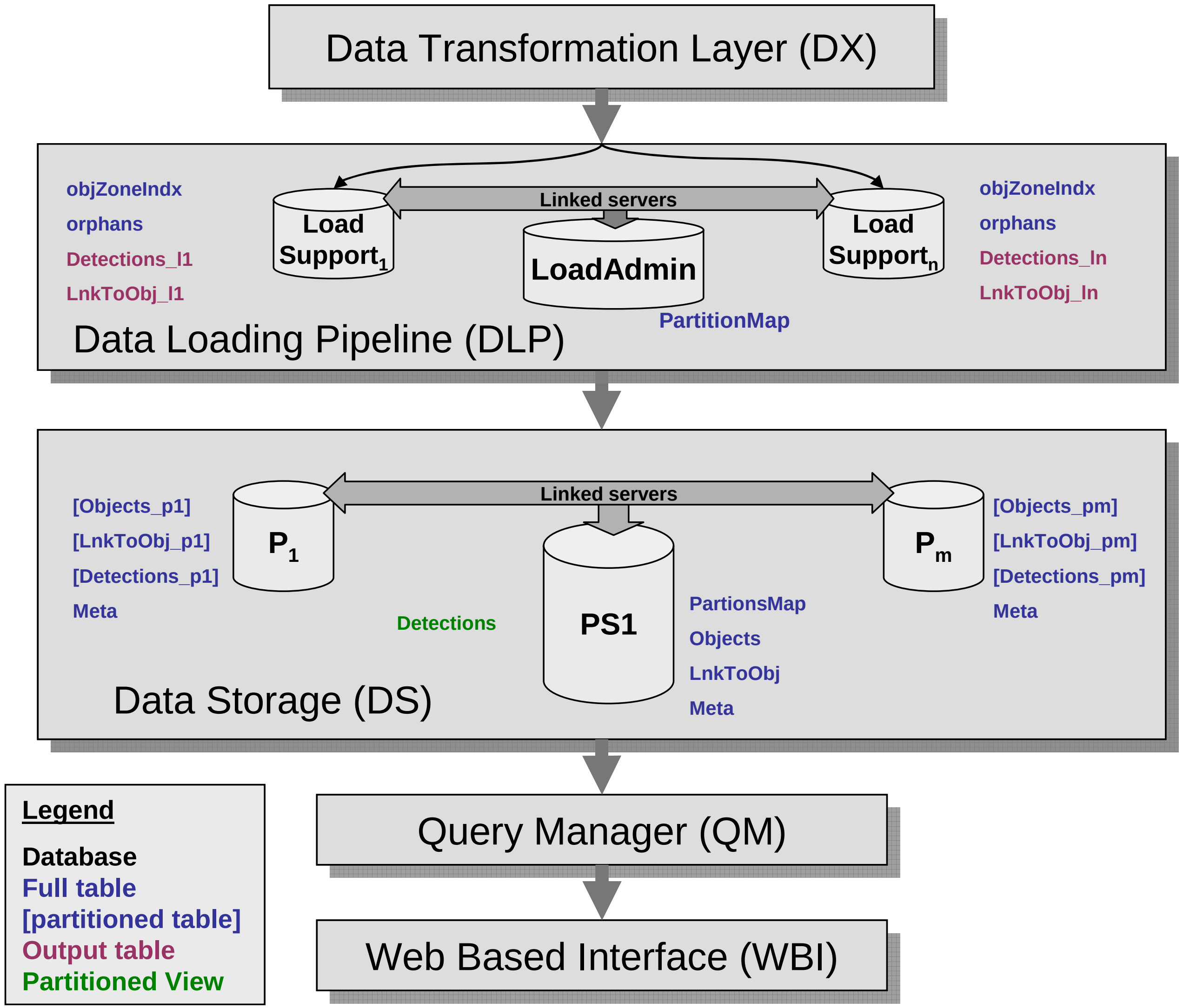
- Too few realizations (now 50)
- Hard to analyze the data afterwards -> need DB
- What is the best way to compare to real data?

Pan-STARRS



- **Detect 'killer asteroids'**
  - *PS1: starting in November 2008*
  - *Hawaii + JHU + Harvard/CfA + Edinburgh/Durham/Belfast + Max Planck Society*
- **Data Volume**
  - *>1 Petabytes/year raw data*
  - *Over 5B celestial objects plus 250B detections in database*
  - *80TB SQLServer database built at JHU,* ***the largest astronomy DB in the world***
  - *3 copies for redundancy*
- **PS4**
  - *4 identical telescopes in 2012, generating 4PB/yr*

PS1 ODM High-Level Organization



PS1 Table Sizes - Monolithic

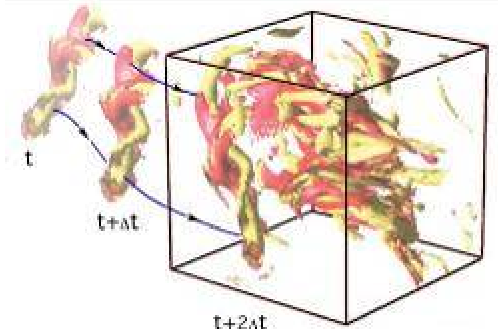
Table	Year 1	Year 2	Year 3	Year 3.5
Objects	2.03	2.03	2.03	2.03
StackDetection	6.78	13.56	20.34	23.73
StackApFlx	0.62	1.24	1.86	2.17
StackModelFits	1.22	2.44	3.66	4.27
P2Detection	8.02	16.03	24.05	28.06
StackHighSigDelta	1.76	3.51	5.27	6.15
Other Tables	*1.78*	*2.07*	*2.37*	*2.52*
Indexes (+20%)	*4.44*	*8.18*	*11.20*	*13.78*
Total	**26.65**	**49.07**	**71.50**	**82.71**

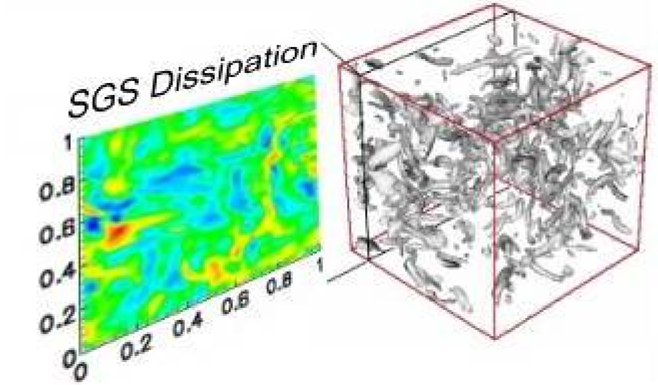
Sizes are in TB

Immersive Turbulence

- **Understand the nature of turbulence**
  - *Consecutive snapshots of a 1,024^3 simulation of turbulence: now 30 Terabytes*
  - *Soon 6K^3 and 300 Terabytes (IBM)*
  - *Treat it as an experiment, observe the database!*
  - *Throw test particles in from your laptop, immerse yourself into the simulation, like in the movie Twister*

- **New paradigm** for analyzing HPC simulations!





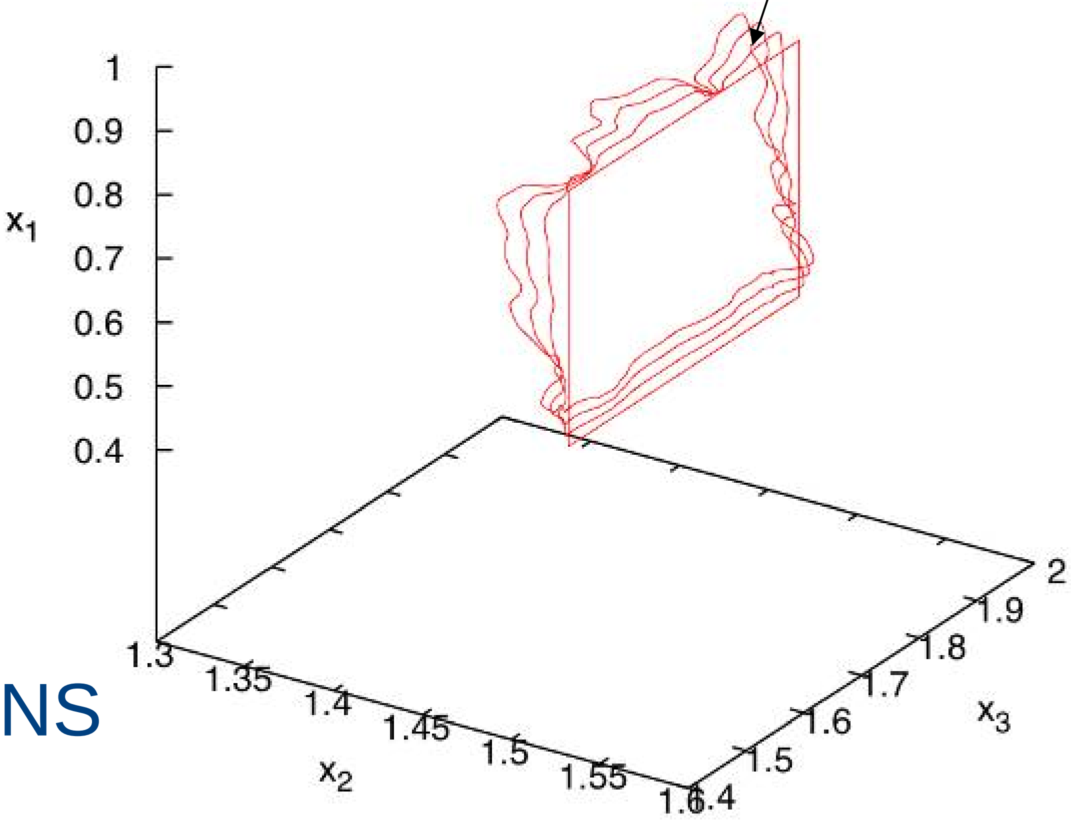
with C. Meneveau, S. Chen (Mech. E), G. Eyink (Applied Math), R. Burns (CS)

Sample code (gfortran 90!) running on a laptop

```
!
 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

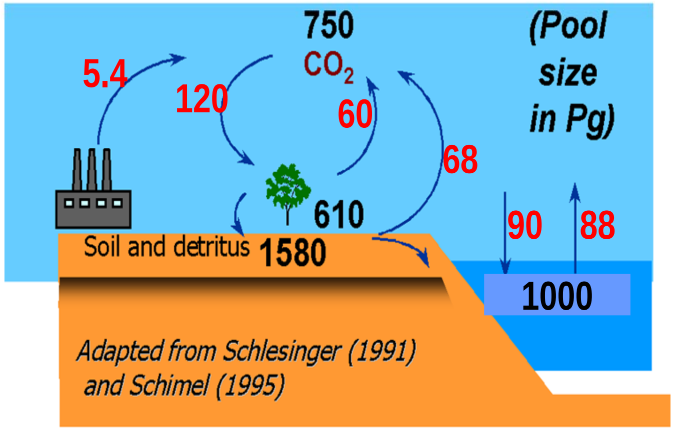
advect backwards in time !



Not possible during DNS

Life Under Your Feet

- **Role of the soil in Global Change**
  - *Soil CO_2 emission thought to be >**15 times** of anthropogenic*
  - *Using sensors we can measure it directly, in situ, over a large area*



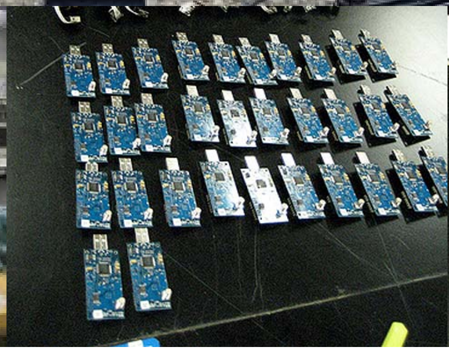
- **Wireless sensor network**
  - *Use 200 wireless (Intel) computers, with 10 sensors each, monitoring*
    - Air +soil temperature, moisture, …
    - Few sensors measure CO_2 concentration
  - *Long-term continuous data, >200M measurements/year*
  - *Complex database of sensor data, built from the SkyServer*

with K.Szlavecz (Earth and Planetary), A. Terzis (CS)

http://lifeunderyourfeet.org/

Next deployment

- Integration with Baltimore Ecosystem Study LTER
  - *End of July 08*
  - *Deploy 200 2nd gen motes*
  - *Goal: Improve understanding of coupled water and carbon cycle in the soil*
  - *Use better sensors*



Ongoing BES Data Collection

precipitation depth (point gages) NEXRAD precipitation precipitation chemistry air temperature humidity atmospheric pressure wind speed wind direction total solar irradiance photosynthetically active radiation net radiation CO_2 flux CO_2 profile leaf wetness profile dewpoint temperature	soil permanent plots temperature moisture content trace gas fluxes (N_2O, CO_2, CH_4) nutrients (total N, NO_3^-, NH_4^+, PO_4^-) matric potential heat flux fauna toxics soil survey data surficial geology well logs rock core records groundwater levels stream flow stage velocity,discharge	stream chemistry water temperature air temperature pH dissolved oxygen specific conductance anions (NO_3^-, Cl^-, SO_4^{2-}, PO_4^-) cations (Na^+, K^+, Mg^{2+}, Ca^{2+}) total N and P total suspended solids turbidity fecal coliform toxics pathogens biota	storm sewer flows storm sewer toxics storm sewer nutrients vegetation inventory vegetation remote sensing digital elevation model LIDAR elevation data sediment surveys channel profiles channel stability data land use/land cover demographic data economic data

Welty and McGuire 2006

Commonalities

- Huge amounts of data, aggregates needed
  - *But also need to keep raw data*
  - *Need for parallelism*
- Requests enormously benefit from indexing
- Very few predefined query patterns
  - *Everything goes….*
  - *Rapidly extract small subsets of large data sets*
  - *Geospatial everywhere*
  - *Buckets and crawlers….*
- Data will never be in one place
  - *Remote joins will not go away*
- Not much need for transactions
- Data scrubbing is crucial

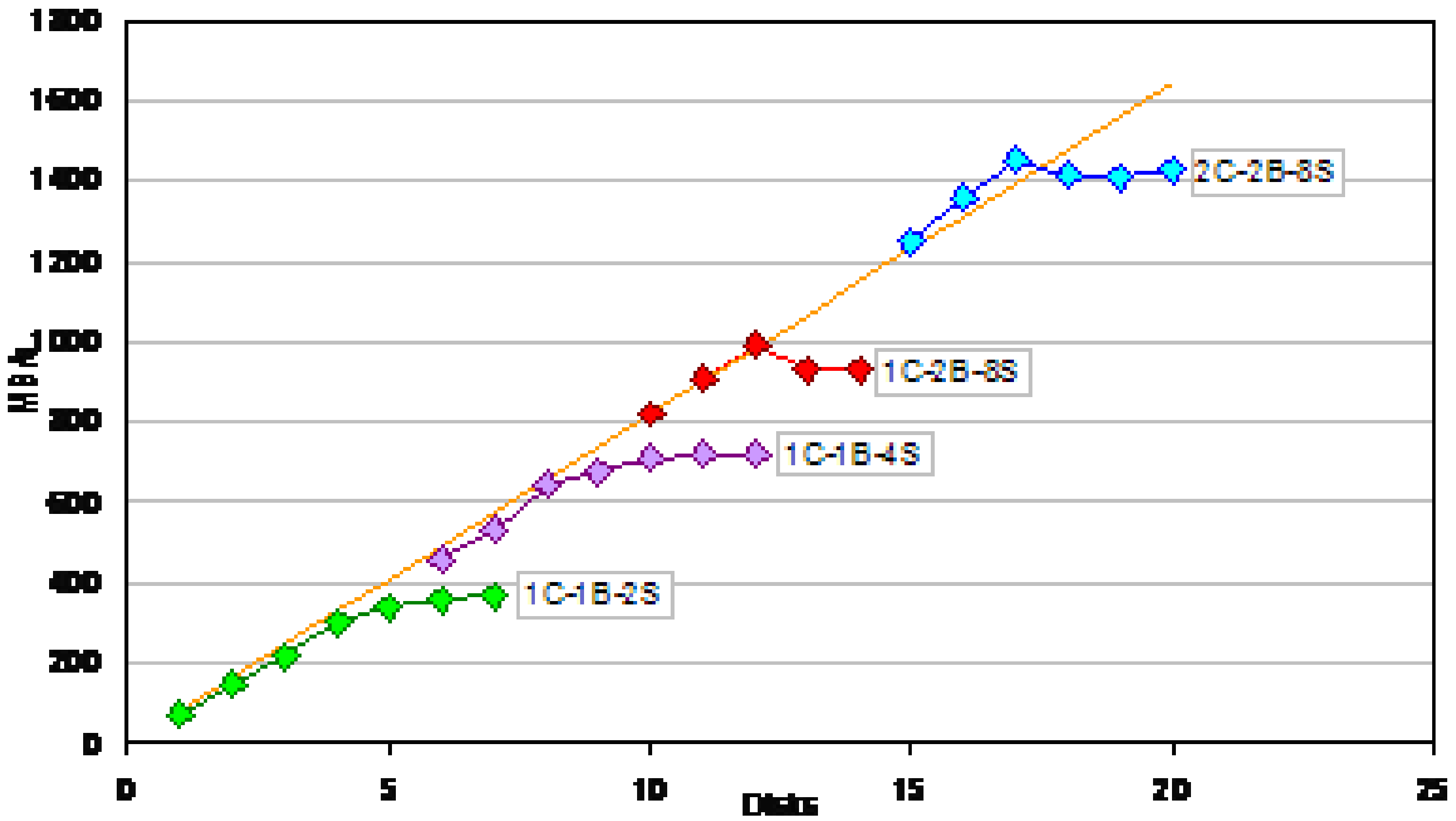
Emerging Trends for DISC

- Large data sets are here, solutions are not
- Scientists are "cheap"
  - *Giving them SW is not enough*
  - *Need recipe for solutions*
- Emerging sociological trends:
  - *Data collection in ever larger collaborations (VO)*
  - *Analysis decoupled, off archived data by smaller groups*
- Even HPC projects choking on IO
- Exponential data growth
  - *> data will be never co-located*
- "Data cleaning" is much harder than data loading

Petascale Computing at JHU

- We are building a distributed SQL Server cluster exceeding 1 Petabyte
- Just becoming operational
- 40x8-core servers with 22TB each,
  6x16-core servers with 33TB each,
  connected with 20 Gbit/sec Infiniband
- 10Gbit lambda uplink to StarTap
- Funded by Moore Foundation, Microsoft and the Pan-STARRS project
- Dedicated to eScience,
  will provide public access

IO Measurements on JHU System

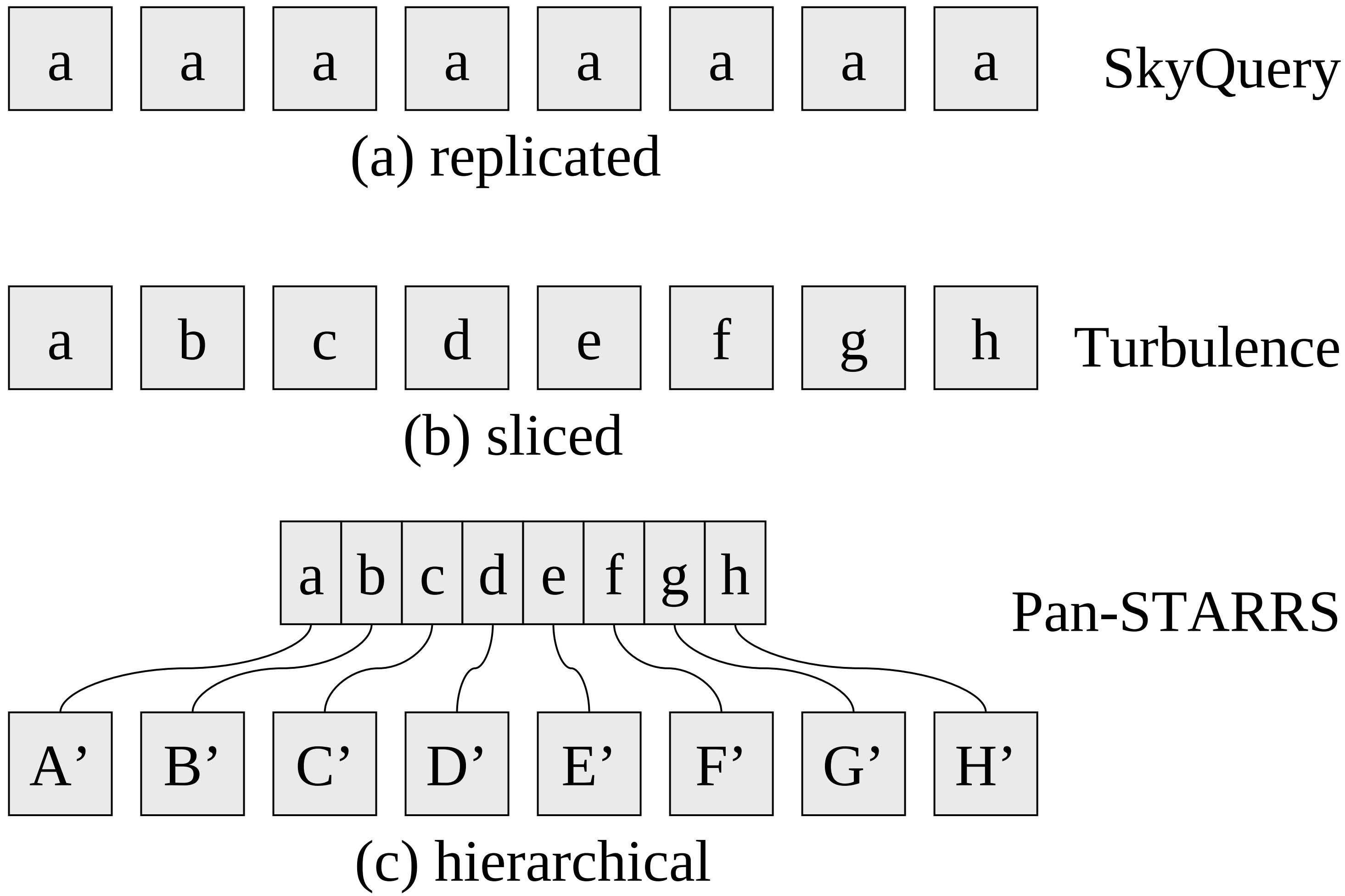


1 server: 1.4 Gbytes/sec, 22.5TB, $12K

Aggregate Performance

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

Data Layouts



Components

- Data must be heavily partitioned
- It must be simple to manage

- Distributed SQL Server cluster
- Management tools
- Configuration tools
- Workflow environment for loading/system jobs
- Workflow environment for user requests
- Provide crawler framework
  - *Both SQL and procedural languages*
- User workspace environment (MyDB)

The Road Ahead

- Build Pan-Starrs (be pragmatic)
- Generalize to GrayWulf prototype
- Fill with interesting datasets
- Create publicly usable dataspace
- Add procedural language support for user crawlers
- Adopt Amazon-lookalike service interfaces
  - *S4 -> Simple Storage Services for Science (Budavari)*
- Distributed workflows across geographic boundaries
  - *“Ferris-wheel”/streaming algorithms (w. B. Grossman)*
  - *“Data pipes” for distributed workflows (w. B. Bauer)*
  - *“Data diffusion” (w I. Foster and I. Raicu)*

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 new generations of instruments do we have left?
- Are there new growth areas emerging?
- **Software is also an instrument**
  - *hierarchical data replication*
  - *virtual data*
  - *data cloning*

Technology+Sociology+Economics

- Neither of them is enough
- **Technology** changing rapidly
  - *Sensors, Moore's Law*
  - *Trend driven by changing generations of technologies*
- **Sociology** is changing in unpredictable ways
  - *YouTube, tagging,…*
  - *Best presentation interface may come from left field*
  - *In general, people will use a new technology if it is*
    - Offers something entirely new
    - Or substantially cheaper
    - Or substantially simpler
- **Economics**: funding is not changing

Grand Challenges (a'la TPC)?

- Benchmark characteristics
  - *Define simple filter (<100 cycles per byte)*
  - *Set selectivity threshold to a few percent (1%-2%-5%)*
  - *Apply it to a PB of data, extract the result*

1. Fastest post-processing of a petabyte data set
2. The cheapest Terabit/sec for a PB data set
   - *Apply similar pattern as above, measure continuous IO*
3. Do a Terabit/sec join between separated PB data

   (a) Data sets reside on separate machines

   (b) Data sets reside at different geographic locations
4. Load/Analyze a PB
   - *Same as 1, but also need to load the data as well*

Summary

- Data growing exponentially
- Petabytes/year by 2010
  - *Need scalable solutions*
  - *Move analysis to the data*
  - *Spatial and temporal features essential*
- Explosion is coming from inexpensive sensors
- Same thing happening in all sciences
  - *High energy physics, genomics, cancer research, medical imaging, oceanography, remote sensing, …*
- Science with so much data requires a new paradigm
  - *Computational methods, algorithmic thinking will come just as naturally as mathematics today*
- We need to come up with new HPC architectures
- ***eScience***: an emerging new branch of science