



Amdahl's Laws and Extreme Data-Intensive Scientific Computing

Alex Szalay
The Johns Hopkins University

Scientific Data Analysis Today

- Scientific data is doubling every year, reaching PBs
- Data is everywhere, never will be at a single location
- Need randomized, incremental algorithms
 - *Best result in 1 min, 1 hour, 1 day, 1 week*
- Architectures increasingly CPU-heavy, IO-poor
- Data-intensive scalable architectures needed
- Most scientific data analysis done on small to midsize BeoWulf clusters, from faculty startup
- Universities hitting the “power wall”
- Soon we cannot even store the incoming data stream
- **Not scalable, not maintainable...**

Why Is Astronomy Special?

- Especially attractive for the wide public
- Community is not very large
- It is real and well documented
 - *High-dimensional (with confidence intervals)*
 - *Spatial, temporal*
- Diverse and distributed
 - *Many different instruments from many different places and times*
- The questions are interesting
- There is a lot of it (soon petabytes)

It has no commercial value

No privacy concerns, freely share results with others
Great for experimenting with algorithms

WORTHLESS!



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
 - 40 TB of raw data => 120TB processed
 - 5 TB catalogs => 35TB in the end
- Started in 1992, finished in 2008
- Long-term archiving of the data in progress

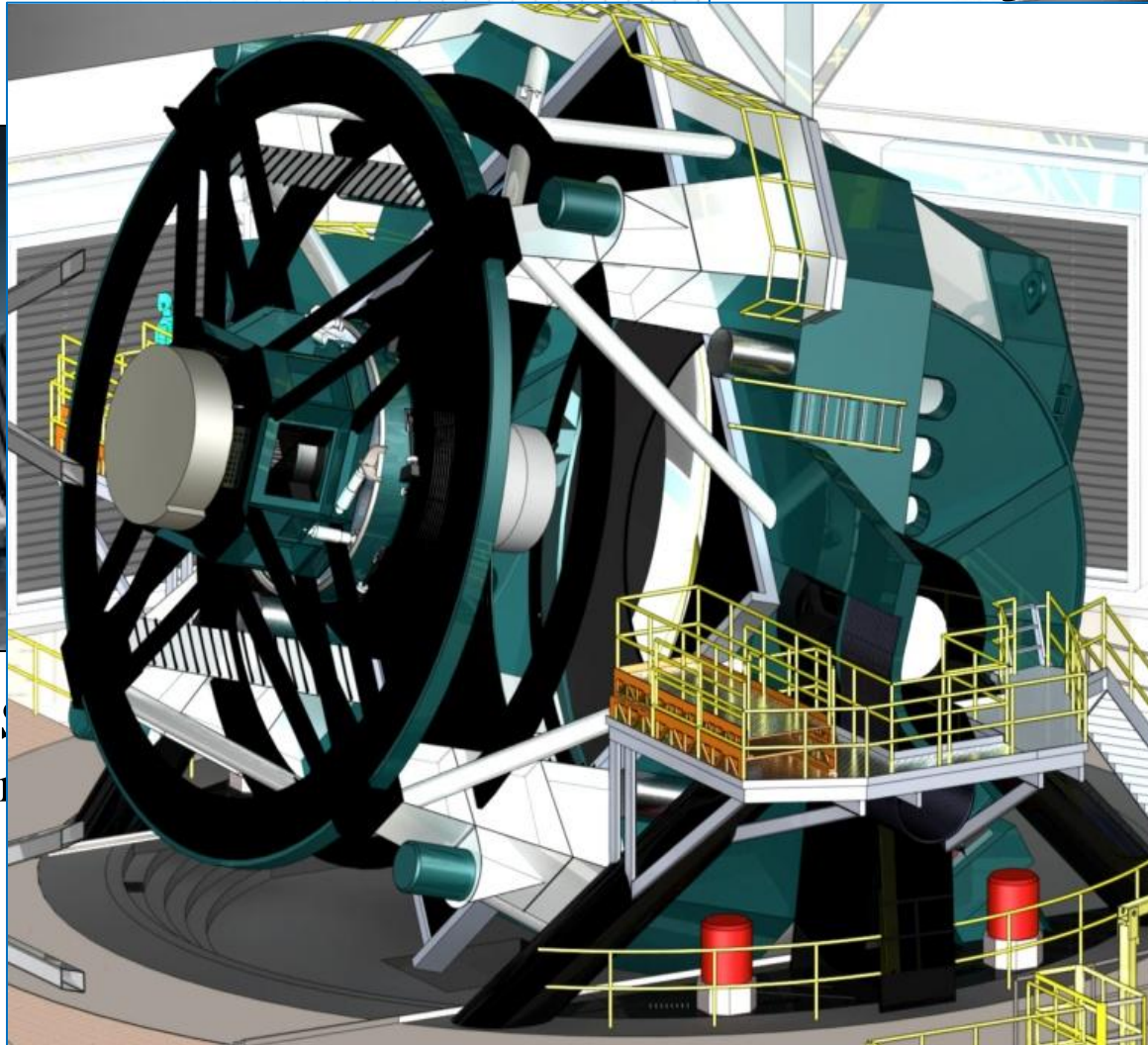
*The University of Chicago
Princeton University
The Johns Hopkins University
The University of Washington
New Mexico State University
Fermi National Accelerator Laboratory
US Naval Observatory
The Japanese Participation Group
The Institute for Advanced Study
Max Planck Inst, Heidelberg
Sloan Foundation,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



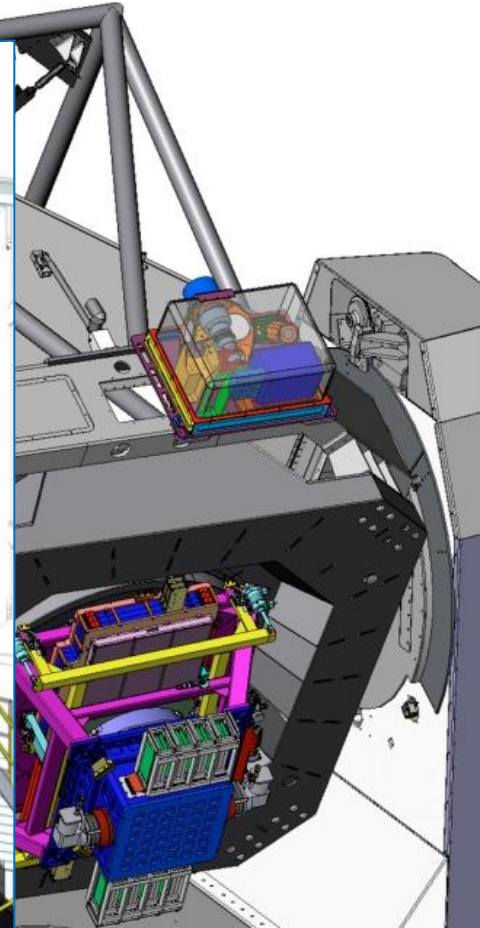
The Era of Surveys



SDSS
2.4m

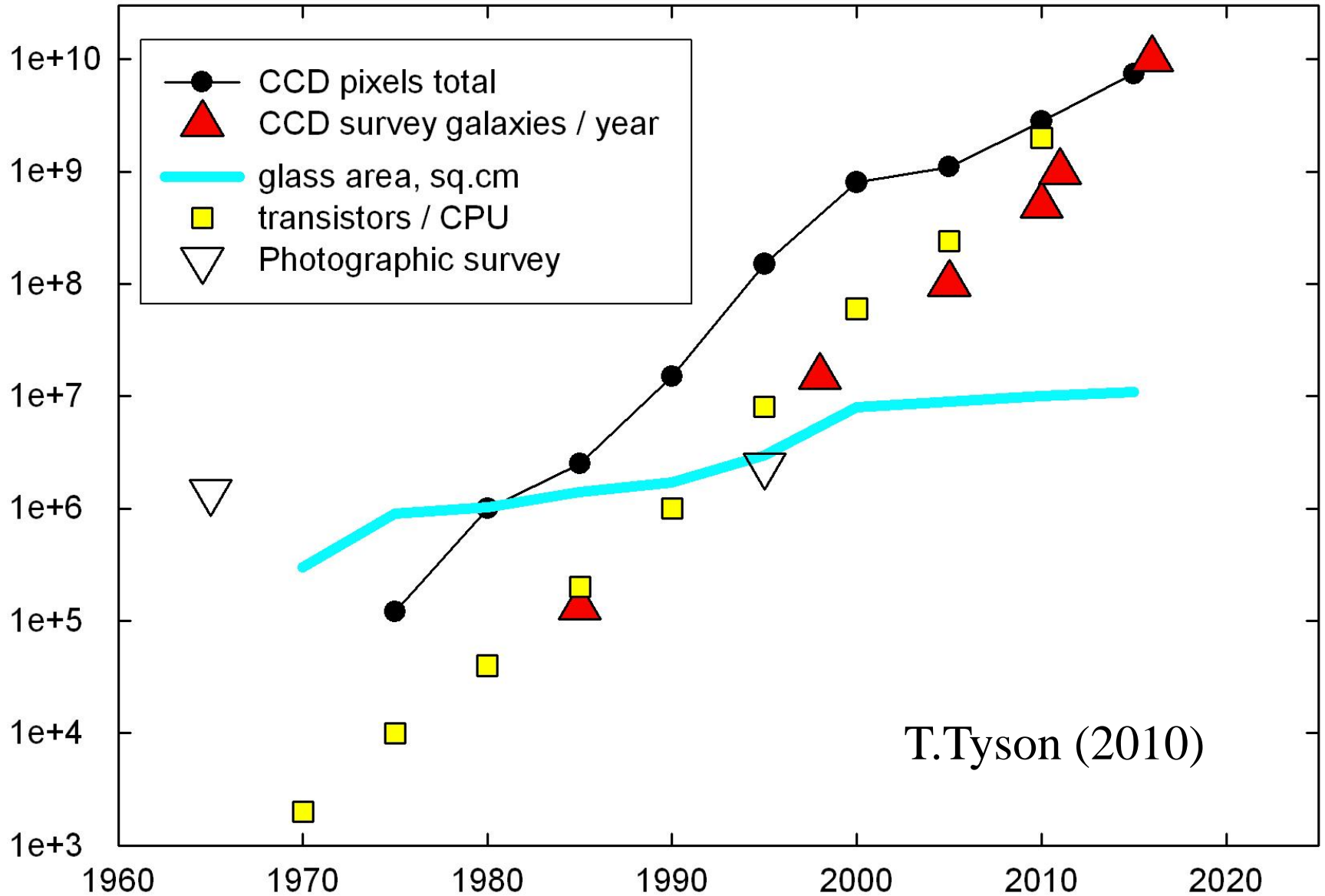


LSST
8.4m 3.2Gpixel



STARRS
1.8m 1.4Gpixel

Survey Trends



T.Tyson (2010)

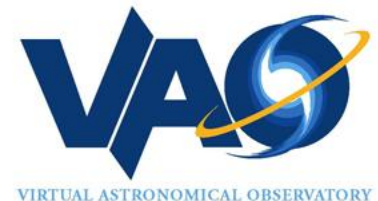
Virtual Observatory

- NSF ITR project, “Building the Framework for the National Virtual Observatory” collaboration of 20 groups
- Similar projects now in 15 countries
 - *International Virtual Observatory Alliance*
- Virtual Astronomical Observatory



Common VO Challenges

- Most challenges are sociological, not technical !!!!!!!!!!!!!
- Hard to find data (yellow pages/repository)
- Threshold for publishing data is currently too high
- Scientists want calibrated data with occasional access to low-level raw data
- Sophisticated data models take a long time...
- Robust applications are hard to build (factor of 3...)
- Geospatial everywhere, but GIS is not good enough
- Archives on all scales, all over the world
- Need distributed user repository



Continuing Growth

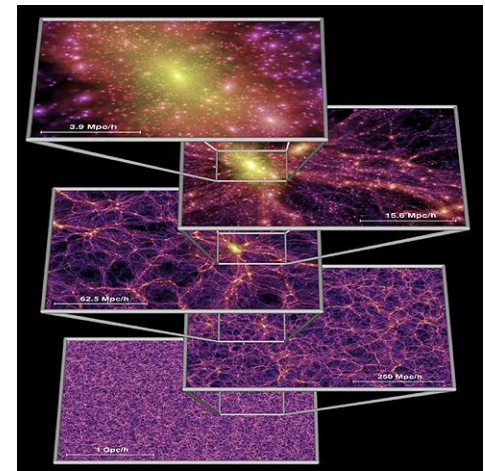
How long does the data growth continue?

- 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 federated data sets*
 - *Hierarchical data replication*
 - *Large and complex simulations*

Cosmological Simulations

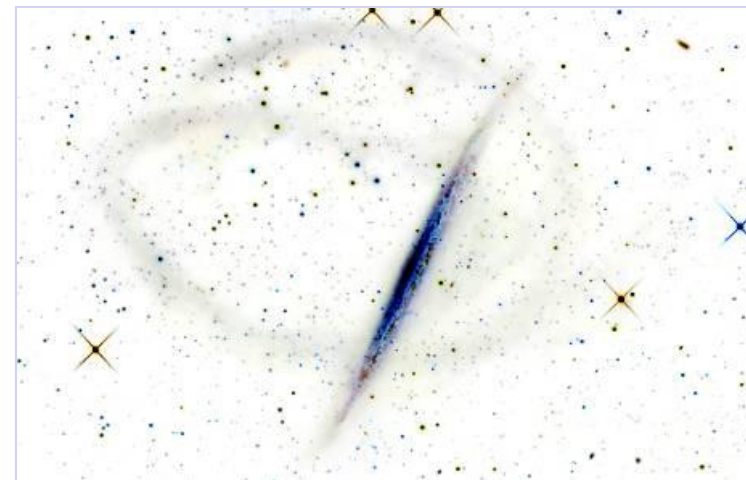
Cosmological simulations have 10^9 particles and produce over 30TB of data (Millennium)

- 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
-
- Hard to analyze the data afterwards -> need DB
 - What is the best way to compare to real data?
 - Next generation of simulations with 10^{12} particles and 500TB of output are under way (Exascale-Sky)



The Milky Way Laboratory

- Idea: use cosmology simulations as an immersive laboratory for general users
- Use Via Lactea-II (20TB) as prototype, then Silver River (500TB+) as production (15M CPU hours)
- Output 10K+ hi-rez snapshots (200x of previous)
- Users insert test particles (dwarf galaxies) into system and follow trajectories in precomputed simulation
- Users interact remotely with 500TB in 'real time'
- Madau, Rockosi, Wyse, Silk, Szalay, Westermann, Blakeley



Commonalities

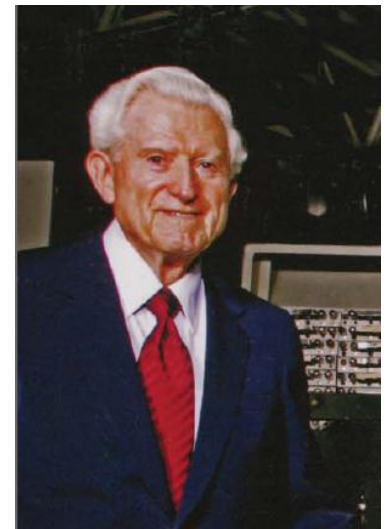
- Huge amounts of data, aggregates needed
 - *But also need to keep raw data*
 - *Need for parallelism*
- Use patterns enormously benefit from indexing
 - *Rapidly extract small subsets of large data sets*
 - *Geospatial everywhere*
 - *Compute aggregates*
 - *Fast sequential read performance is critical!!!*
 - *But, in the end everything goes.... search for the unknown!!*
- Data will never be in one place
 - *Newest (and biggest) data are live, changing daily*
- Fits DB quite well, but no need for transactions
- Design pattern: class libraries wrapped in SQL UDF
 - *Take analysis to the data!!*

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)

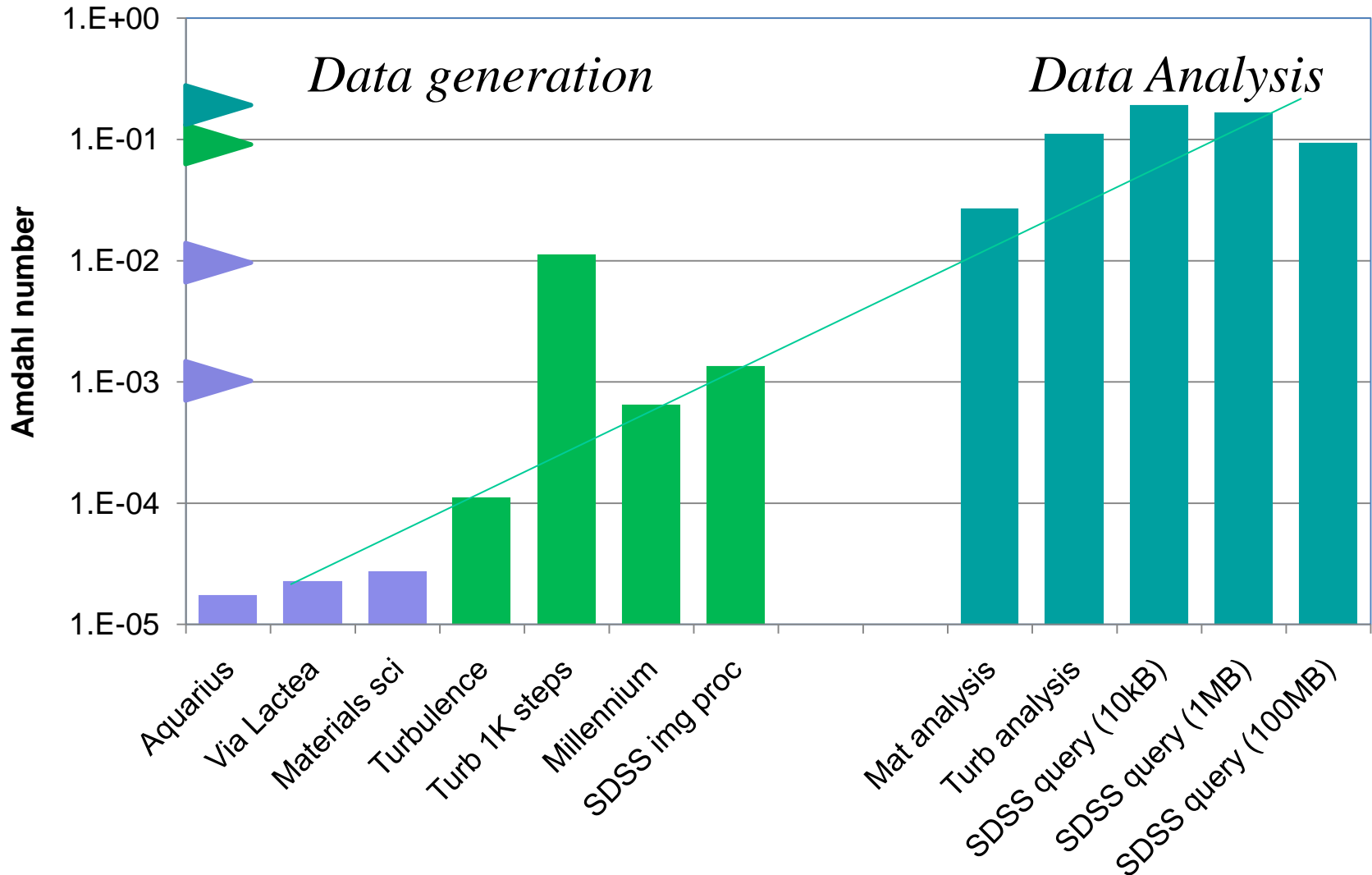
Modern multi-core systems move farther
away from Amdahl's Laws
(Bell, Gray and Szalay 2006)



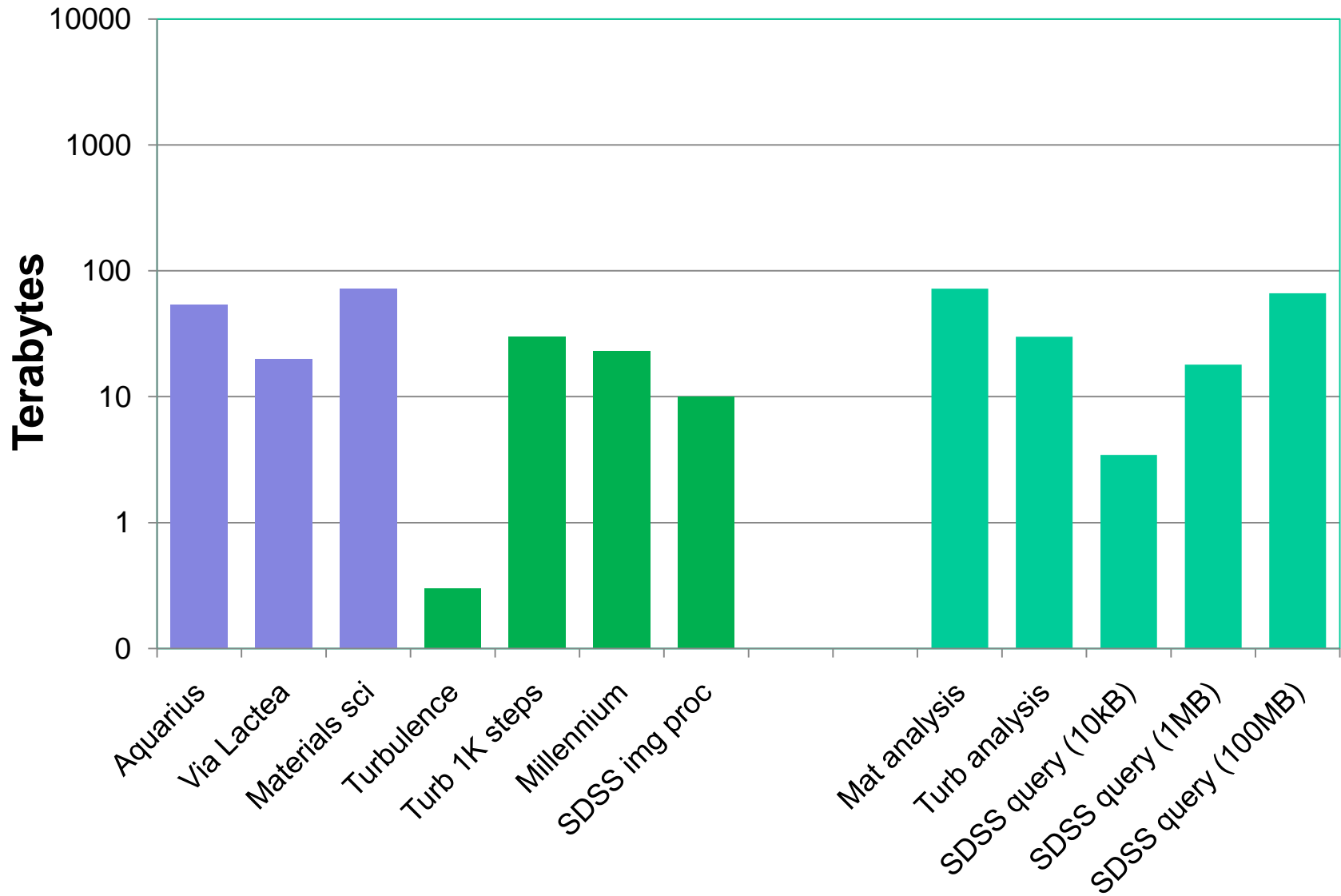
Typical Amdahl Numbers

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

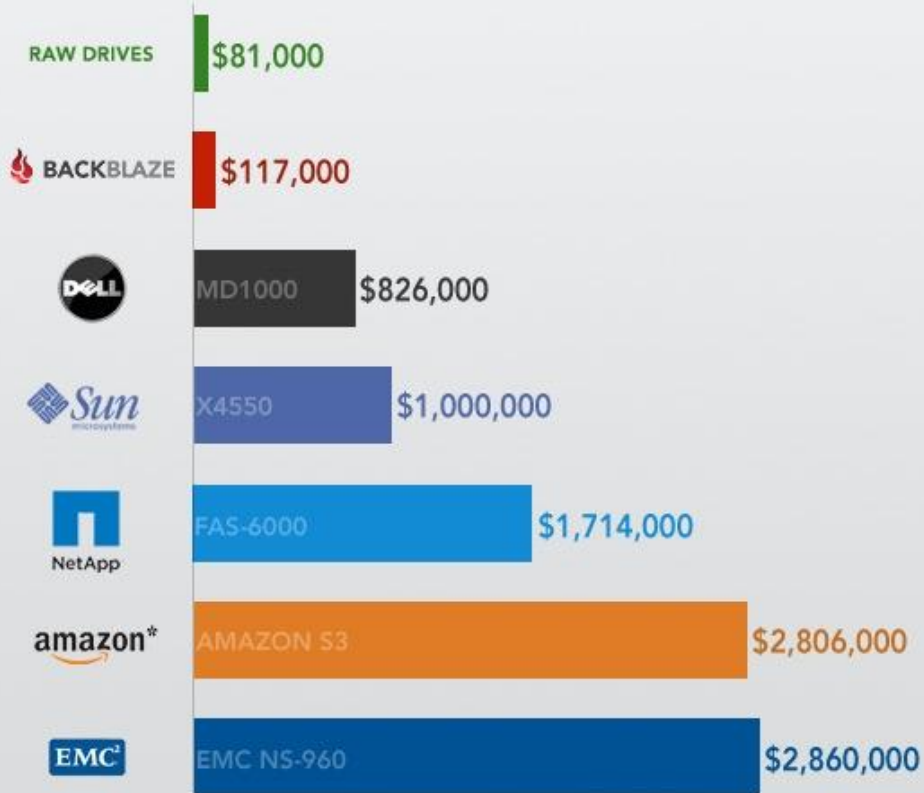
Ordered priorities for data-intensive scientific computing

1. *Total storage* (-> *low redundancy*)
2. *Cost* (-> *total cost vs price of raw disks*)
3. *Sequential IO* (-> *locally attached disks, fast ctrl*)
4. *Fast 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

COST OF A PETABYTE



* Amazon S3 Storage over three years (minus electricity, co-location and administration).

From backblaze.com
Aug 2009



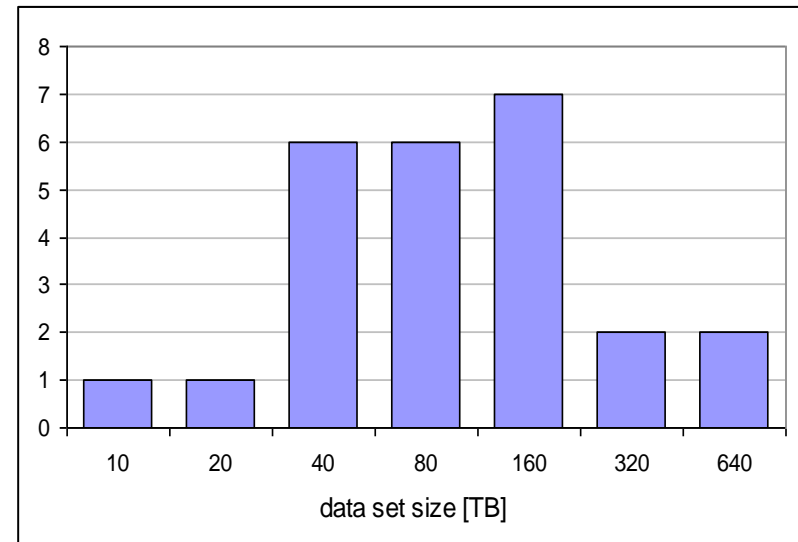
JHU Data-Scope

- 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

	<i>1P</i>	<i>1S</i>	<i>90P</i>	<i>12S</i>	<i>Full</i>	
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

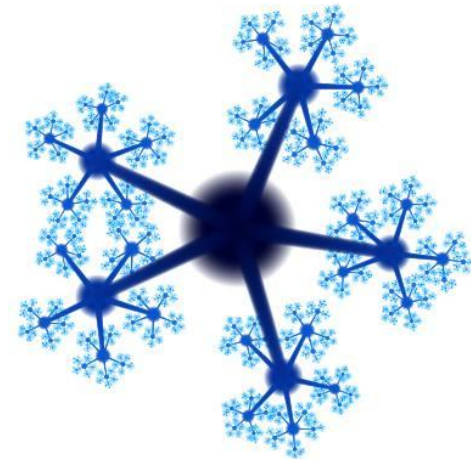
Discipline	data [TB]
Astrophysics	930
HEP/Material Sci.	394
CFD	425
Bioinformatics	414
Environmental	660
Total	2823



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



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 large data centers*
 - *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

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 (photo-z) in 5 minutes over 1.2TB*
 - *Ported RF module from R in C#/CUDA*

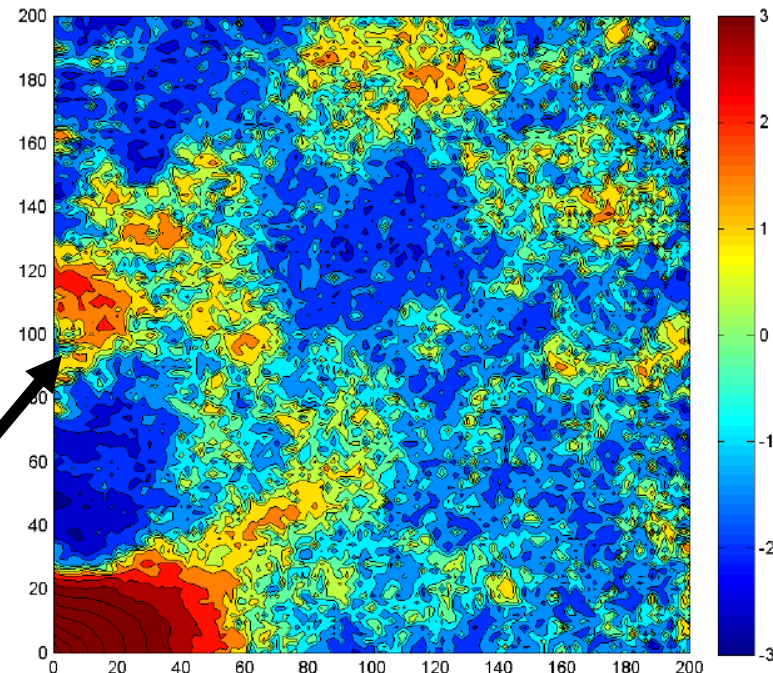


The Impact of GPUs

- Reconsider the $N \log N$ 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 galaxy pairs using brute force N^2 algorithm
- Faster than the tree codes!

Tian, Neyrinck,
Budavari, Szalay 2010

BAO



Summary

- Large data sets are here, solutions are not
 - *100TB is the current practical limit*
- Science community starving for storage and IO
- No real data-intensive computing facilities available
 - *Changing with Dash, Gordon, Data-Scope, GrayWulf...*
- Even HPC projects choking on IO
- Real multi-PB solutions are needed NOW!
- Cloud hosting currently very expensive
- Cloud computing tradeoffs different from science needs
- Scientists are “frugal”, also pushing the limit
- Current architectures cannot scale much further
- Astronomy representative for science data challenges

*“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”