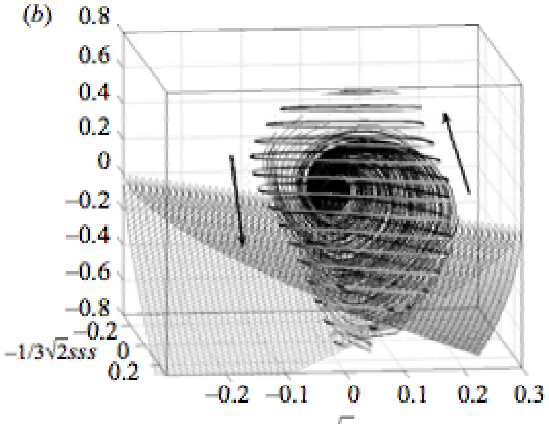


Amdahl's Laws and Extreme Data-Intensive Scientific Computing



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…**

Continuing Growth of Data

How long does the growth continue?

- High end always linear
- Exponential comes from technology + economics
  - *rapidly changing generations of sensors and processing*
  - *like CCD's replacing photographic plates*
- 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*
  - *Large and complex simulations*
  - *Hierarchical data replication*

Challenges

- Where are the challenges in DISC 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 algorithms*

Increased Diversification

One shoe does not fit all!

- Diversity grows naturally, no matter what
- Evolutionary pressures help
  - *Large floating point calculations move to GPUs*
  - *Fast IO moves to high Amdahl number systems*
  - *Stream processing emerging*
  - *noSQL vs databases vs column store etc*
  - *Large data moves into the cloud (but how?)*
- Individual groups want subtle specializations

At the same time

- What remains in the middle/mainstream?
- Boutique systems dead, commodity rules
- Large graph problems still hard to do (XMT or Pregel)

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)

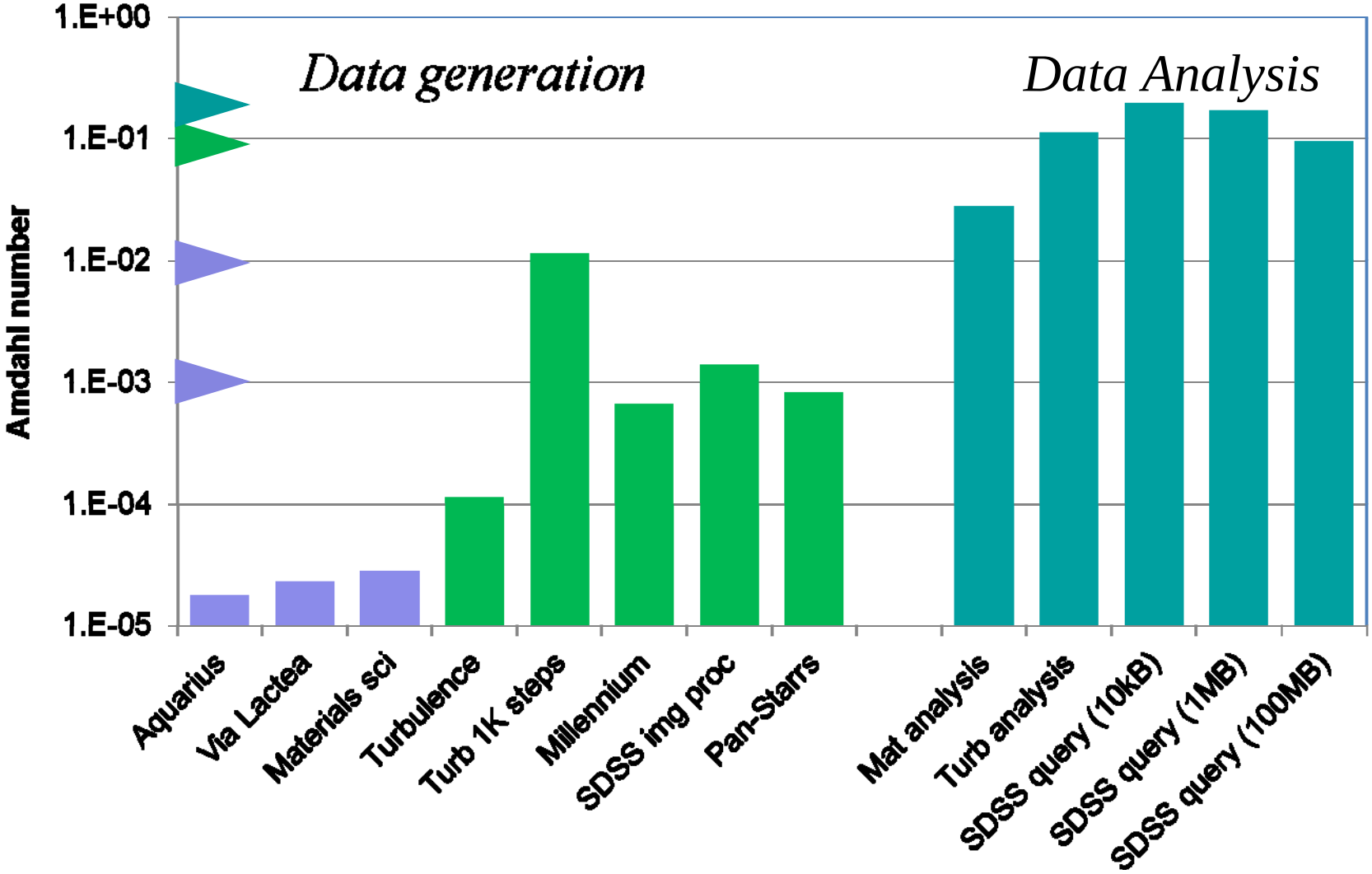
iii. One byte of memory per one instruction/sec (MEM)

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

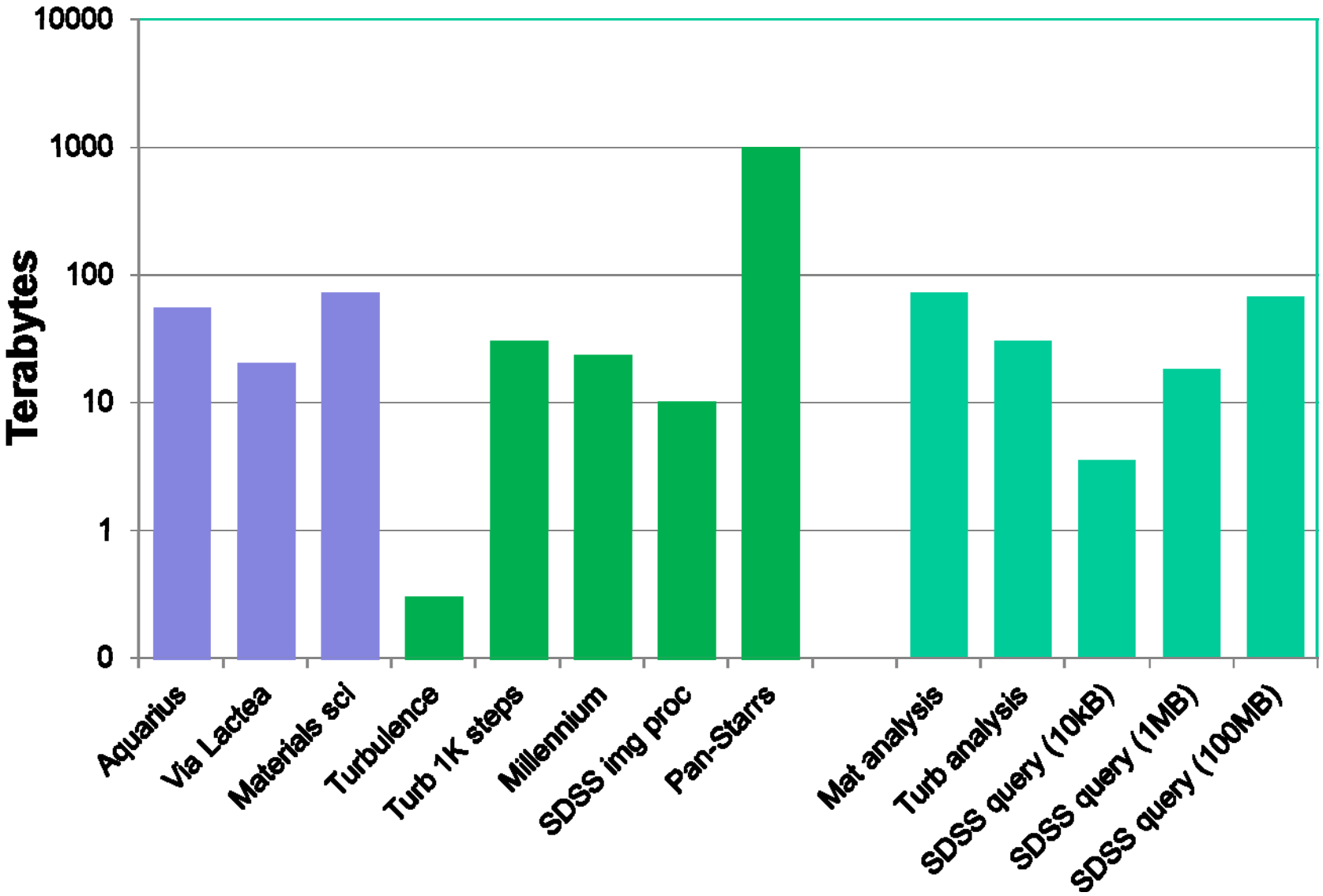
Typical Amdahl Numbers

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*
- Stream processing
  - *A continuum of Amdahl numbers*
- 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

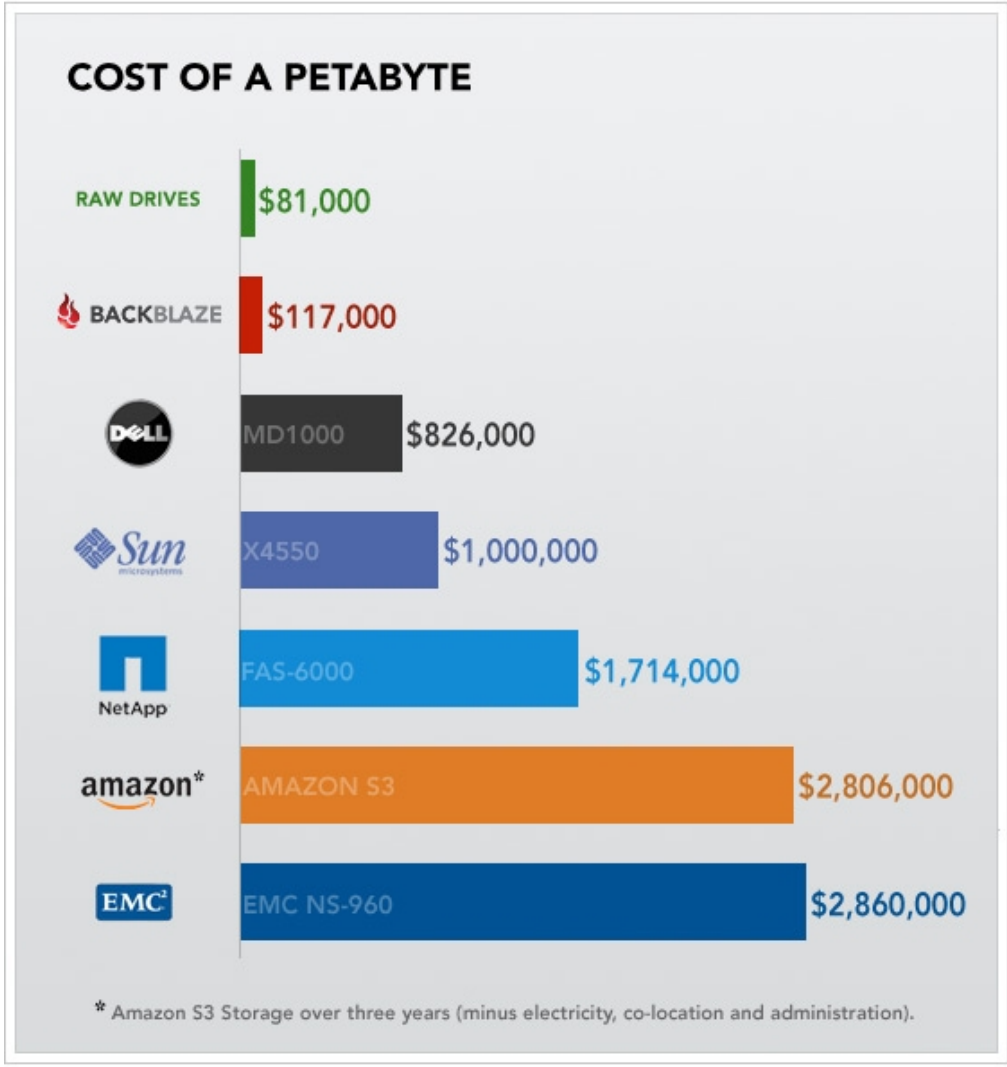
Stu Feldman: Extreme computing is about tradeoffs

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



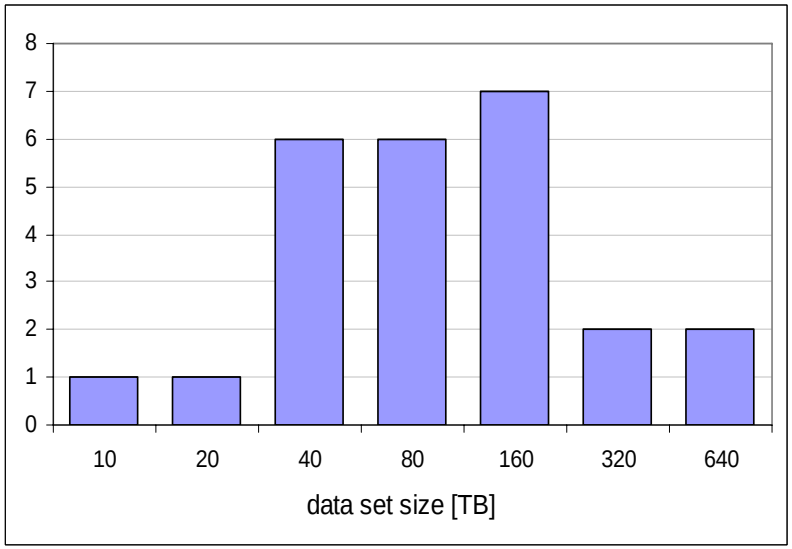
From backblaze.com
Aug 2009

JHU Data-Scope

- Funded by NSF MRI to build a new ‘instrument’ to look at data
- Goal: 100 servers for $1.2M, including 10G switches
- Two-tier: performance (P) and storage (S)
- Large (5PB) + cheap + fast (400+GBps), but …
  . ..a special purpose instrument

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

Solid State Disks (SSDs)

- Much higher throughput with lower power consumption
  - *250MB/s sequential read, 200MB/s sequential write*
  - *1-2W peak power*
- Incorporating SSDs to server designs
  - *Scale-up*
    - Install SSDs into existing high end servers
    - Quickly running out of PCI BW
  - *Scale down*
    - Amdahl Blade: one SSD per core
    - Ideal Amdahl number and IOPS ratio

Cyberbricks/Amdahl Blades

- **Scale down** the CPUs to the disks!
  - *Solid State Disks (SSDs)*
  - *1 low power CPU per SSD*
- Current SSD parameters
  - *OCZ Vertex 120GB, 250MB/s read, 10,000+ IOPS, $350*
  - *Power consumption 0.2W idle, 1-2W under load*
- Low power motherboards
  - *Intel dual Atom N330 + NVIDIA ION chipset 28W at 1.6GHz*
- Combination achieves perfect Amdahl blade
  - *200MB/s=1.6Gbits/s ⇔ 1.6GHz of Atom*

Building a Low Power Cluster

AS, G.Bell, H.Huang, A.Terzis, A.White (HotPower09):

Evaluation of many different motherboard + SSD combinations

System	Model	CPU	Chipset
ASUS	EeeBox	N270	945GSE
Intel	D945GCLF2	N330	945GC
Zotac	Ion	N330	ION
AxiomTek	Pico 820	Z530	US15W
Alix	3C2	LX800	AMD

Sweet spot: dual core Atom!

Basic Properties

Table 2: Basic properties of a single node for various systems considered

	CPU [GHz]	mem [GB]	seqIO [GB/s]	randIO [kIOPS]	disk [TB]	power [W]	cost [$]	relative power	Amdahl numbers seq	mem	rand
GrayWulf	21.3	24	1.500	6.0	22.5	1,150	19,253	1.000	0.56	1.13	0.014
ASUS	1.6	2	0.124	4.6	0.25	19	820	0.017	0.62	1.25	0.144
Intel	3.2	2	0.500	10.4	0.50	28	1,177	0.024	1.25	0.63	0.156
Zotac	3.2	4	0.500	10.4	0.50	30	1,189	0.026	1.25	1.25	0.163
AxiomTek	1.6	2	0.120	4.0	0.25	15	995	0.013	0.60	1.25	0.125
Alix 3C2	0.5	0.5	0.025		0.008	4	325	0.003	0.40	1.00	

Cost of ownership includes $0.15/kWh x 1.6 for AC overhead
Price assumes 250GB SSD for $400

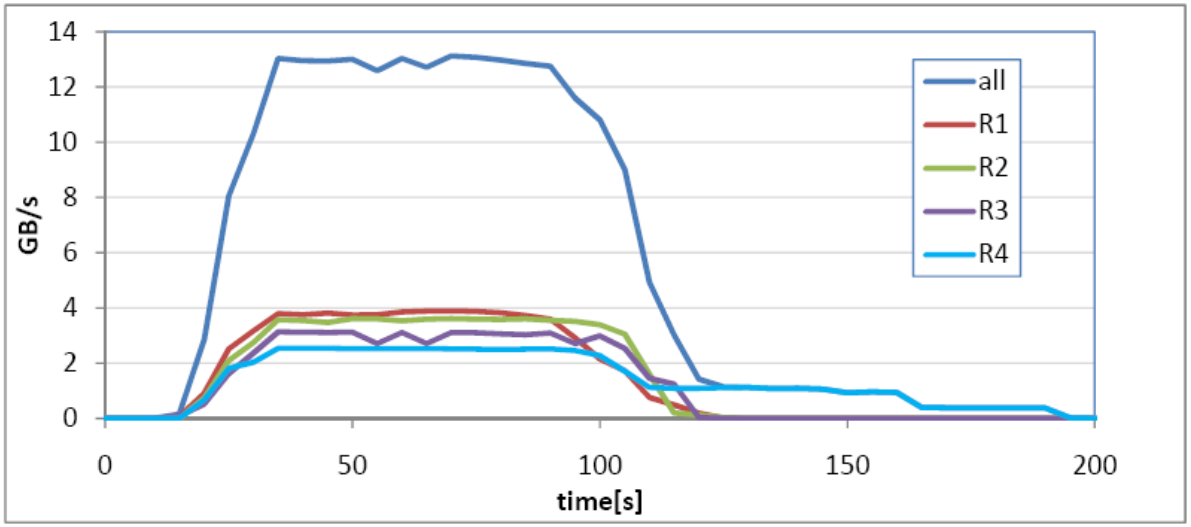
1GB/sec sequential IO for $2400

1kW Amdahl Cluster

- 36-node Amdahl cluster using 1000W 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*
- I/O Performance: 18GB/s
- Amdahl number = 1 for under $40K
- Using the GPUs for data mining:
  - *6.4B multidimensional regressions in 5 minutes over 1.2TB*
  - *Ported RF module from R in C#/CUDA*

Histogram Test

- Build histogram of 544 million objects from 1.2TB
- Reasonably complex selection criteria
- Distrubuted SQL query with dynamic load balancing
- Runs in 100 sec, average throughput: **13GB/sec**
- SSD at 421MB/s, SSD+1TB at 397MB/s & 379 MB/s



Scaling I

Sweet spot

constant price

	CPU	seq IO	randIO	disk	power	cost	rel power	price	elect	
	[GHz]	[GB/s]	[kIOPS/s]	[TB]	[W]	[$]		[$]	[$/3yr]	nodes
Gr[illegible]lf	21.3	1.5	6	22.5	1150	19253	1.0000	12000.0	7253.3	1.0
ASUS	37.6	2.9	108	5.9	446	19253	0.3880	16439.0	2814.3	23.5
Intel	52.4	8.2	164	8.2	458	19253	0.3984	16363.5	2889.8	16.4
Zotac	51.8	8.1	168	8.1	486	19253	0.4223	16189.9	3063.4	16.2
Pico820	31.0	2.3	77	4.8	290	19253	0.2525	17421.9	1831.4	19.4
Alix 3C2	42.7	2.1		0.7	342	19253	0.2973	17096.6	2156.6	85.5
hybrid	56.8	8.9		40.0	799	19253	0.6951	14211.4	5041.9	17.8

constant seq IO

	CPU	seq IO	randIO	disk	power	cost	rel power	price	elect	
	[GHz]	[GB/s]	[kIOPS/s]	[TB]	[W]	[$]		[$]	[$/3yr]	nodes
GrayWulf	21.3	1.5	6	22.5	1150	19253	1.0000	12000.0	7253.3	1.0
ASUS	19.4	1.5	56	3.0	230	9917	0.1999	8467.7	1449.6	12.1
Intel	9.6	1.5	30	1.5	84	3530	0.0730	3000.0	529.8	3.0
Zotac	9.6	1.5	31	1.5	90	3568	0.0783	3000.0	567.6	3.0
Pico820	20.0	1.5	50	3.1	188	12433	0.1630	11250.0	1182.6	12.5
Alix 3C2	30.0	1.5	0	0.5	240	13514	0.2087	12000.0	1513.7	60.0
hybrid	9.6	1.5	18	6.8	135	3251	0.1174	2400.0	851.5	3.0

Scaling II

constant power

	CPU	seq IO	randIO	disk	power	cost	rel power	price	elect	
	[GHz]	[GB/s]	[kIOPS/s]	[TB]	[W]	[$]		[$]	[$/3yr]	nodes
GrayWulf	21.3	1.5	6	22.5	1150	19253	1.0000	12000.0	7253.3	1.0
ASUS	96.8	7.5	278	15.1	1150	49622	1.0000	42368.4	7253.3	60.5
Intel									253.3	41.1
Zota									253.3	38.3
Pico									253.3	76.7
Alix									253.3	287.5
hybri									253.3	25.6

As soon as we have a 750GB SSD for $400, it will be cheaper to build PB-scale storage systems from Amdahl blades!!!

constant disk size

	CPU	seq IO	randIO	disk	power	cost	rel power	price	elect	
	[GHz]	[GB/s]	[kIOPS/s]	[TB]	[W]	[$]		[$]	[$/3yr]	nodes
GrayWulf	21.3	1.5	6	22.5	1150	19253	1.0000	12000.0	7253.3	1.0
ASUS	144.0	11.2	414	22.5	1710	73785	1.4870	63000.0	10785.3	90.0
Intel	144.0	22.5	450	22.5	1260	52947	1.0957	45000.0	7947.1	45.0
Zotac	144.0	22.5	468	22.5	1350	53515	1.1739	45000.0	8514.7	45.0
Pico820	144.0	10.8	360	22.5	1350	89515	1.1739	81000.0	8514.7	90.0
Alix 3C2	1406.3	70.3	0	22.5	11250	633456	9.7826	562500.0	70956.0	2812.5
hybrid	32.0	5.0	60	22.5	450	10838	0.3913	8000.0	2838.2	10.0

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

GPUs

- Hundreds of cores – 100K+ parallel threads!
- CPU is free, RAM is slow
- Amdahl's first Law applies
- GPU has >50GB/s bandwidth inside
- Still difficult to keep the cores busy
  - *IO bandwidth limiting many computations*

- How to integrate it with data intensive computations?
- How to integrate it with SQL?

See Budavari talk

Short Term Trends in DISC

- Large data sets are here, solutions are not
  - *100TB is the current practical limit in science projects*
  - *Much wider need for DISC than expected*
  - *Next year the pain threshold will be about 500TB*
- No real data-intensive computing facilities available
- Even HPC projects choking on IO, and lack storage
- Cloud computing tradeoffs different from science needs
- Increased diversification using commodity components
  - *SSDs, GPUs (Gordon, Amdahl blades,…)*
- Scientists are “frugal”, also pushing the limit
  - *Exploring disruptive technologies…*
  - *Still building our own…on the cheap*
  - *Finally, campus level aggregation emerging …*