Thanks to NVIDIA, Microsoft External Research, NSF, Moore Foundation, OCZ Technology

Low-Power Amdahl-Balanced Blades for Data-Intensive Computing

Alex Szalay , Andreas Terzis, Alainna White, Howie Huang, Jan Vandenberg, Ani Thakar, Tamas Budavari, Sam Carliles, Alireza Murazavi, Gordon Bell, Jose Blakeley, David Luebke, Michael Schuette

Problem Statement

- Data sets reaching 100s of TB, soon PBs
- Large data sets are here, solutions are not
- National infrastructure does not match needs
  - *Facilities focused primarily on CPU cycles*
- Data does not fit into memory ➔ need sequential IO
- Even HPC projects choking on IO
- Scientists are “cheap”, also pushing to the limit
- A similar vacuum led to BeoWulf ten years ago

Amdahl's Laws

Gene Amdahl (1965): Laws for a balanced system

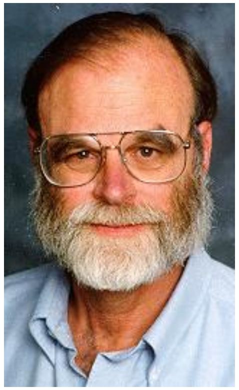
i. Parallelism: max speedup is S/(S+P)

ii. **One bit of IO/sec per instruction/sec (BW)**
Amdahl number

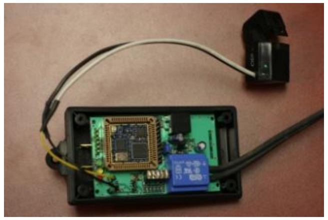
iii. One byte of memory per one instruction/sec (MEM)

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

graywulf



- Distributed SQLServer cluster/cloud
  - *50 servers, 1.1PB disk, 500 CPU*
  - *Connected with 20 Gbit/sec Infiniband*
  - *Linked to 1500 core compute cluster*
  - *Extremely high speed seq I/O*
  - *Each node 1.5GB/s, 1150W*
- Cost efficient: $10K/GBps
- Excellent Amdahl number : 0.56
- **But: hitting the “power wall”!!!!**



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, $360*
  - *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

Szalay, Bell, Huang, Terzis, White (HotPower09 paper):

Evaluation of many different motherboard + SSD combinations

Sweet Spot

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

The ION Advantage

Zotac ION/ITX motherboard

- *NVIDIA ION chipset for the Atom*
- *Supports 4GB of memory,*
- *PCI, 2xSATA channels (3 ports)*
- *16x GPU cores*
- *Needs about 28-30W total*
- *Unified memory for Atom+GPU*
- *CUDA 2.2: "Zero Copy Option"*

Our Hardware Configuration

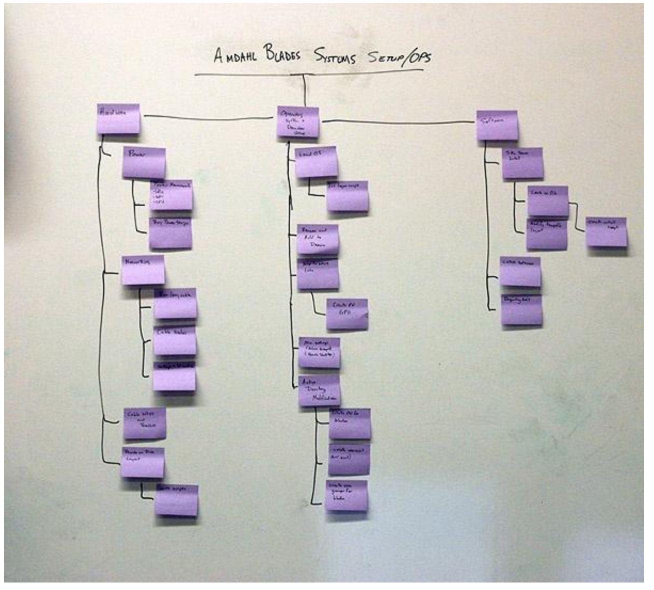
- 36-node cluster using 1200W (same as one GW!)
- Zotac Atom/ION motherboards
  - *4GB of memory, N330 dual core Atom, 16 GPU cores*
- Four rows of 9 nodes, different disk configurations

	C:	D:	E:
Row 1	SSD	SSD	SSD
Row 2	SSD	SSD	1TB
Row 3	SSD	1TB	1TB
Row 4	SSD	0.5TB	0.5TB

- 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*

Software Used

- Windows 7 Release Candidate
- SQL Server 2008 Enterprise RTM
- SQLIO test suite
- PerfMon + SQL Performance Counters
- Built in Monitoring Data Warehouse
- SQL batch scripts for testing
- C# application with CUDA, driven from SQL Server
- DPV for looking at results



Data Layout

- Data derived from the SDSS database
- 1.2TB database table 'sliced' 36 ways, each replicated over three nodes
- Further 12 partitions pre-computed over the primary key and info stored on the head node
- Variable number of partitions used for a given job:
  – *Dynamic load balancing*

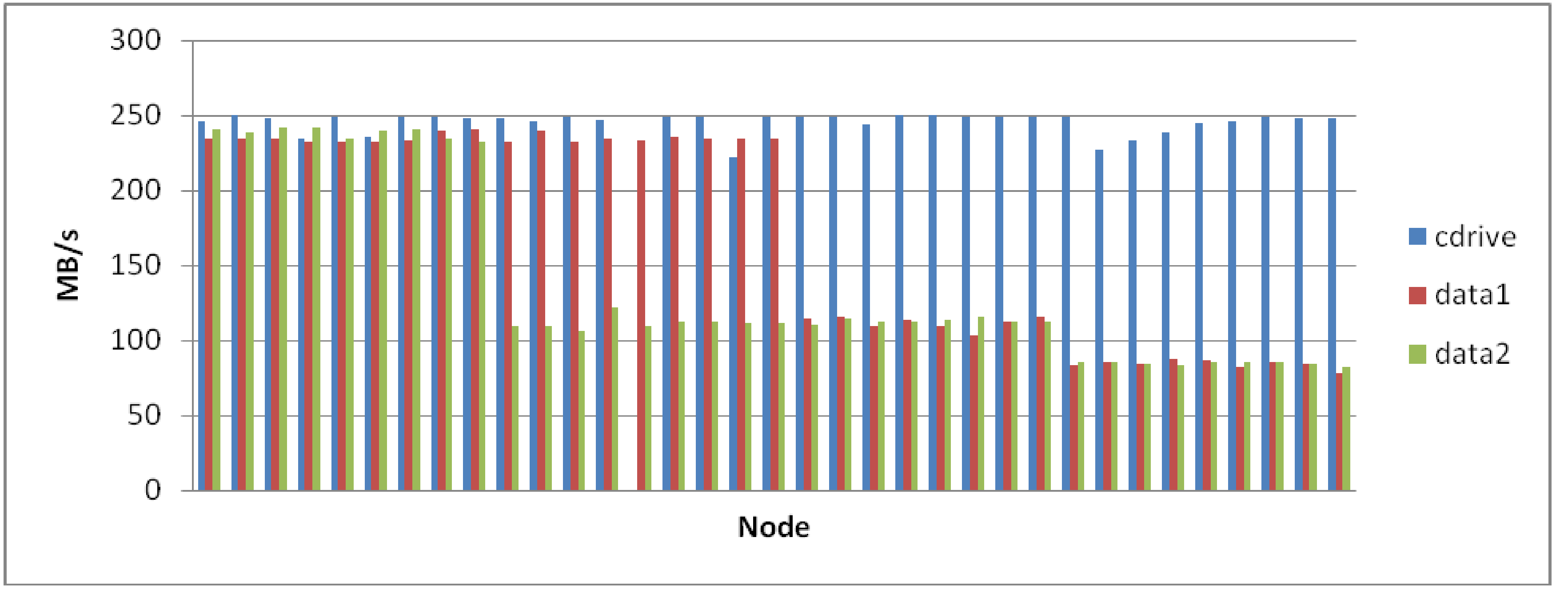
Node	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**			**31**	**32**	**33**	**34**	**35**	**36**
C Volume	A1	A2	A3	A4	A5	A6	A7	A8	. . .		A31	A32	A33	A34	A35	A36
Data 1	A2	A3	A4	A5	A6	A7	A8			A31	A32	A33	A34	A35	A36	A1
Data 2	A3	A4	A5	A6	A7	A8			A31	A32	A33	A34	A35	A36	A1	A2

Performance Tests

- Evaluate cluster performance
  - *Try both pure SSD and hybrid nodes*
  - *Samsung F1 drives 1TB, 128MB/s at 7.5W*
  - *Samsung M1 drives 500GB, 85MB/s at 2.5W*
- Run low level benchmarks
- Experiment with scalability, using SQL tests
- Compare real life apps, combining CUDA from within SQL Server
- “Photometric redshifts” for 300M galaxies from SDSS in 10 minutes, migrating an R module to CUDA and running it from SQL Server

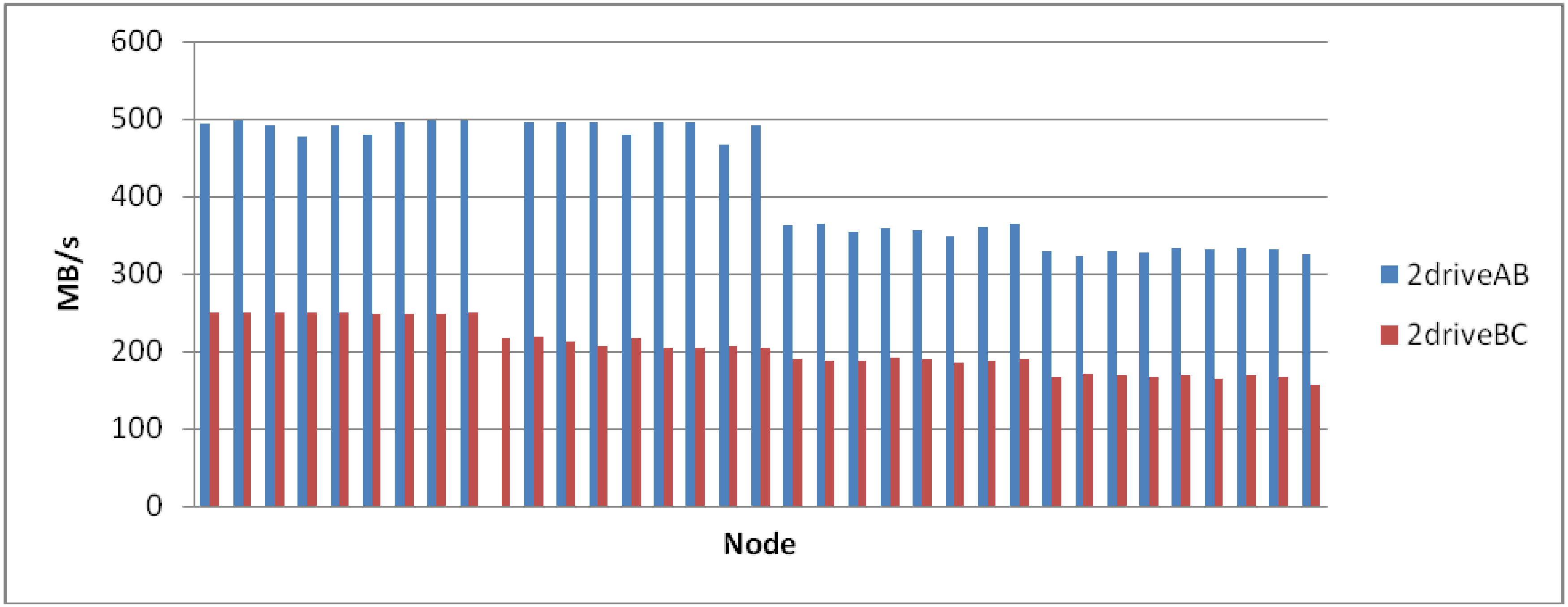
Single Drive IO Tests

- Performs as expected:
  - *OCZ: 250MB/s* (cdrive)
  - *Samsung 1TB: 118MB/s* (data1)
  - *Samsung 500GB: 78MB/s* (data2)



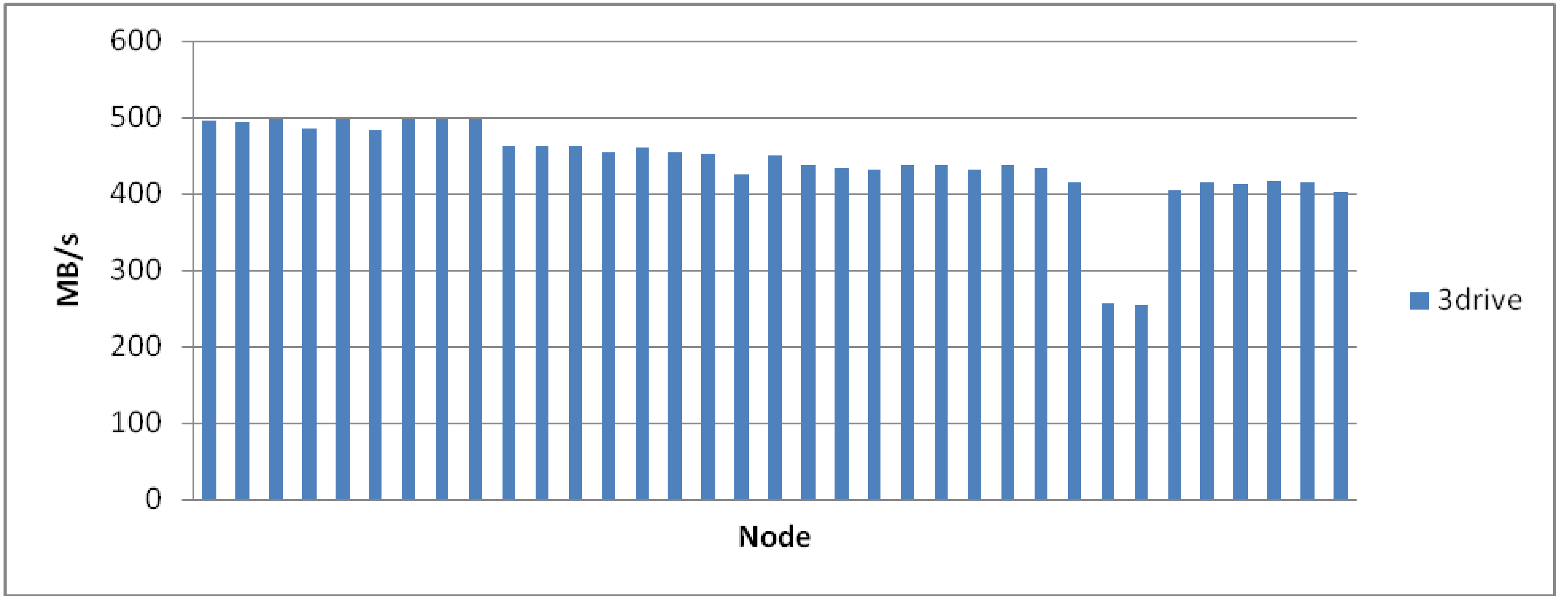
Two Drive IO Tests

- AB: Two SSDs max out at 500MB/s
- AB: SSD+1TB: 365MB/s, SSD+0.5TB: 330MB/s
- BC: 2SSD: 250MB/s, anything else: 160-200MB/s



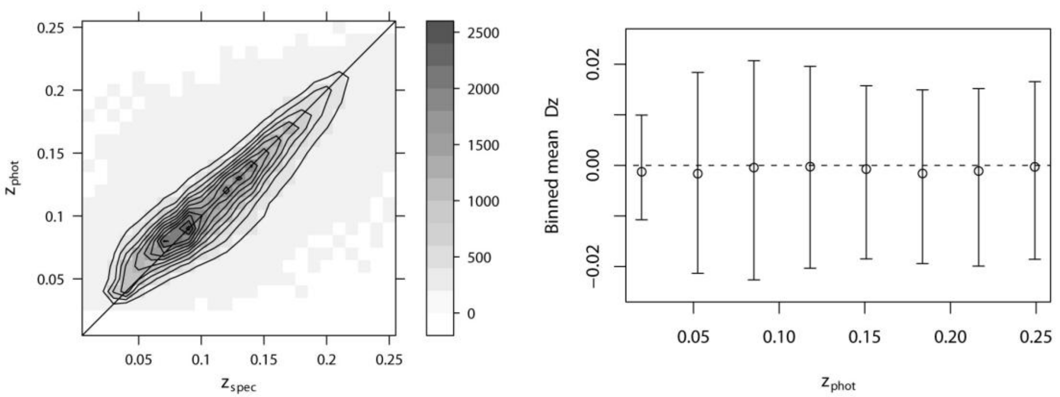
Three Drive IO Tests

- 3SSD: 500MB/s (no gain from 3rd drive)!
- SSD+2X: between 400 and 460MB/s!

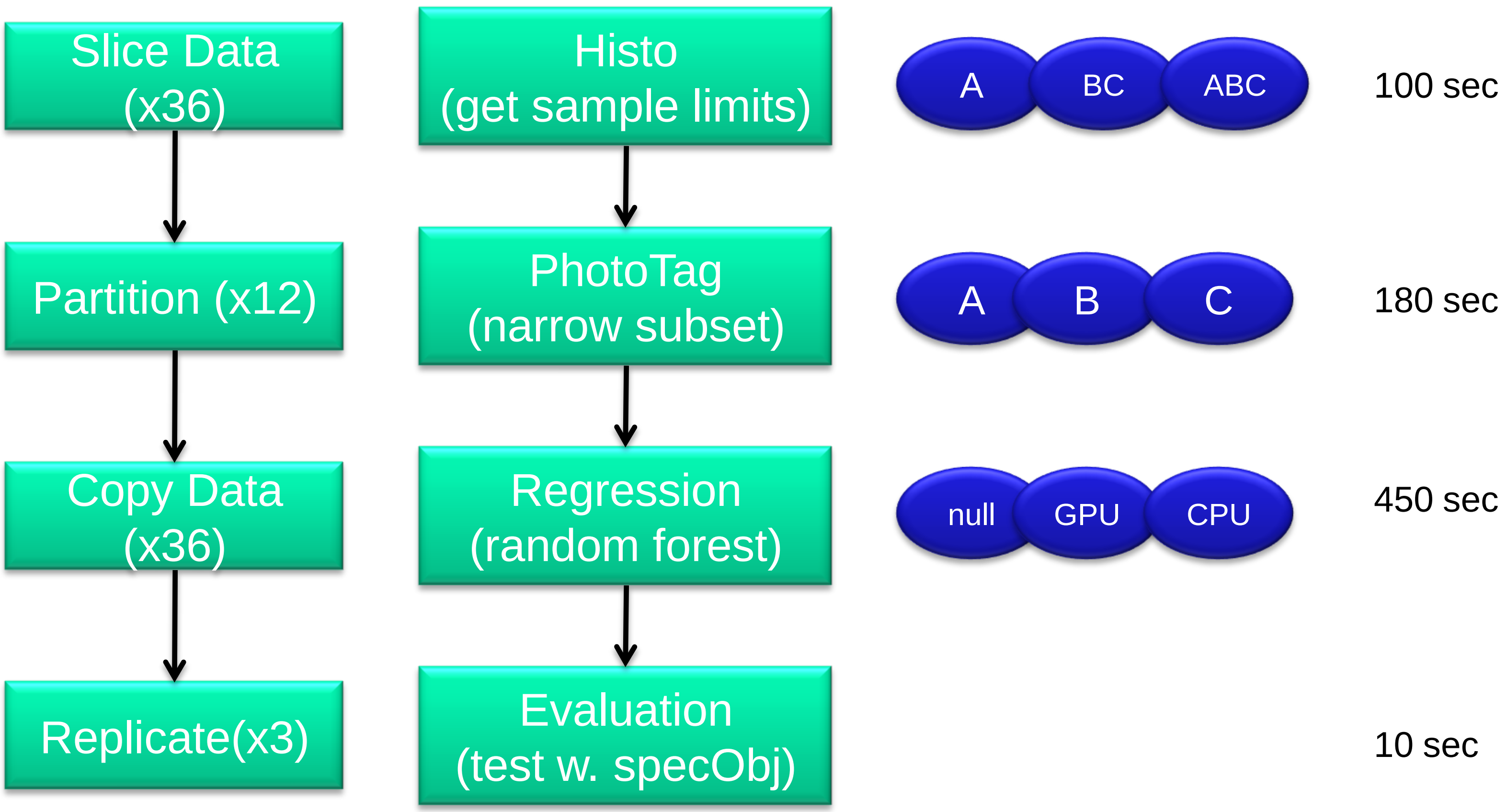


Photometric Redshifts

- Galaxies in SDSS have 5-band images
- Can be considered as a low resolution spectrum
  ➔ *redshift* ➔ *distance*
- From a training set create regression trees
- Random Forest algorithm (Berman):
  - *Use many tree-based regression estimators (20 here)*
  - *Average of the tree estimators is the forest estimator*
  - *Extremely robust*
- 128M galaxies bright enough
- Soon more than 1B



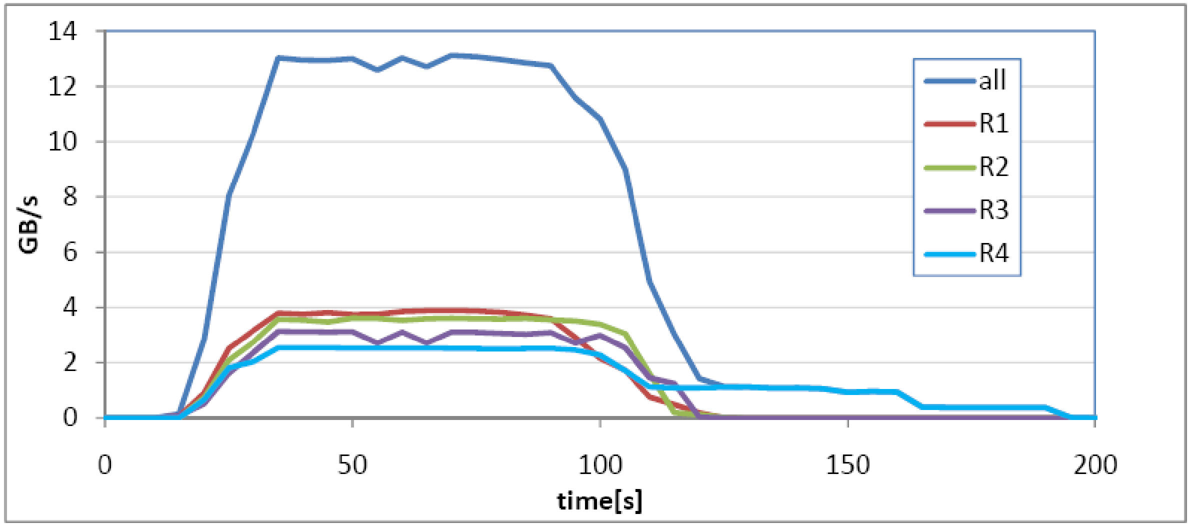
Astronomy Test Workflow



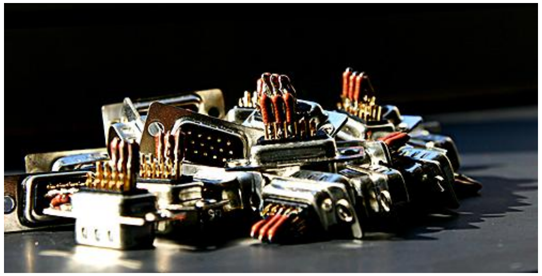
For a data set of 544M objects =>128M galaxies (1.2TB)

Histogram Test

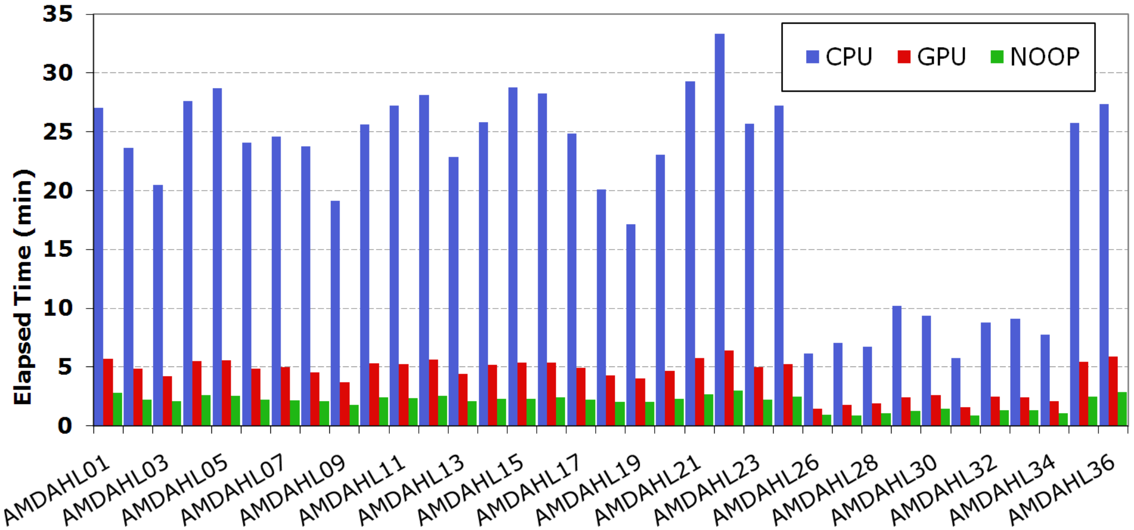
- Build histogram of 544 million objects from 1.2TB
- Reasonably complex selection criteria
- Distrubuted SQL query with dynamic load balancing
- Runs in 100 sec, average throughput: **13GB/sec**
- SSD at 421MB/s, SSD+1TB at 397MB/s & 379 MB/s



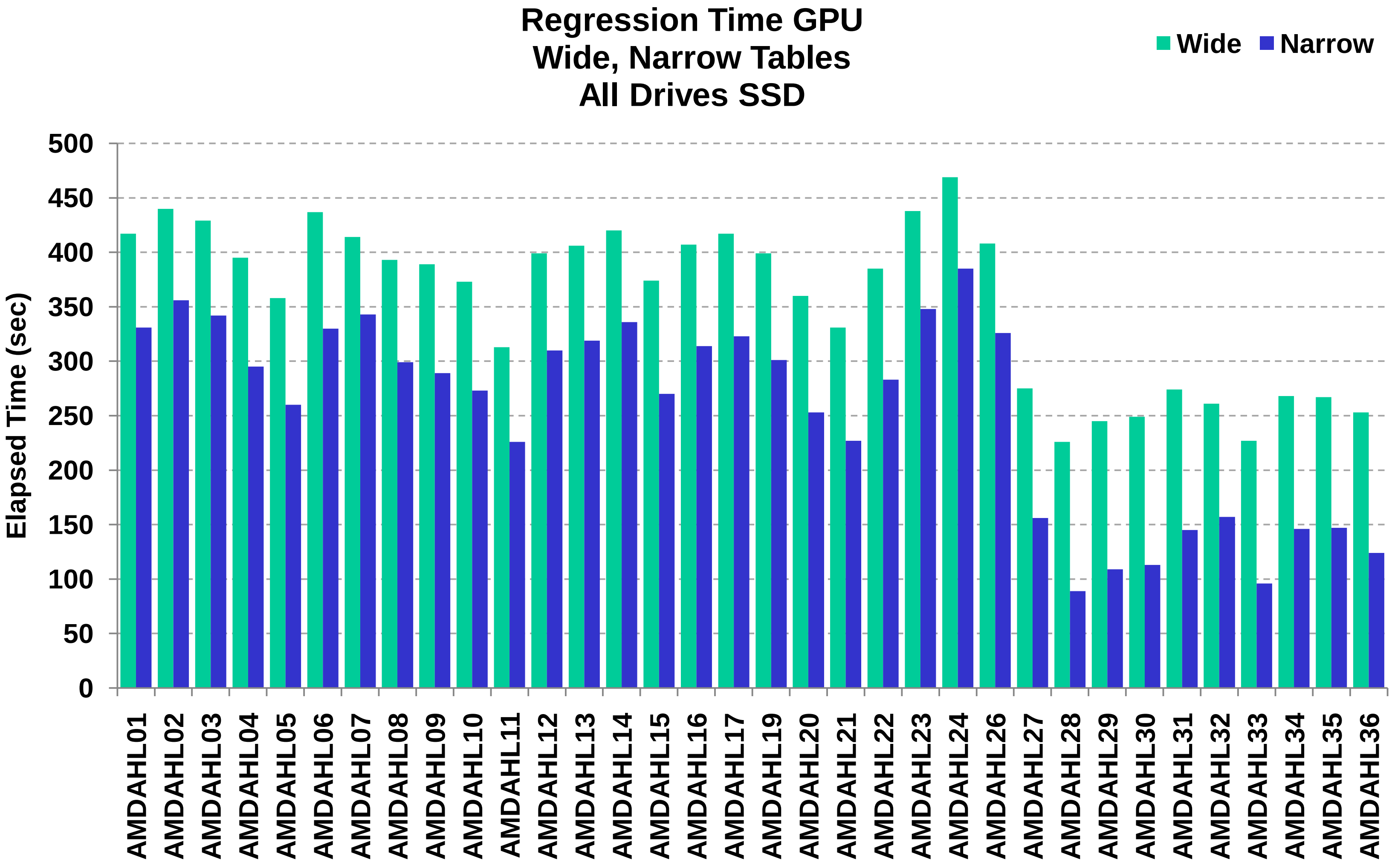
Random Forest



- Three tests: IO only (NOOP), GPU, CPU
- Code developed in C#, with a wrapper around CUDA
- Lots of little ‘traps’, like tricking Windows into having a real monitor (“dongle”), SQL+CUDA now friends…
- Query:
  - *544M objects*
  - *CPU: 1500s*
  - *GPU: 300s!!*
  - *IO: 150s*



Wide and Narrow Tests



Comparisons

- Test averages on the Amdahl cluster:

	NOOP	CPU	GPU	CPU/GPU	dC/dG
narrow	122	1252	259	4.67+-0.8	8.08+-0.8

- For a typical partition (A02)

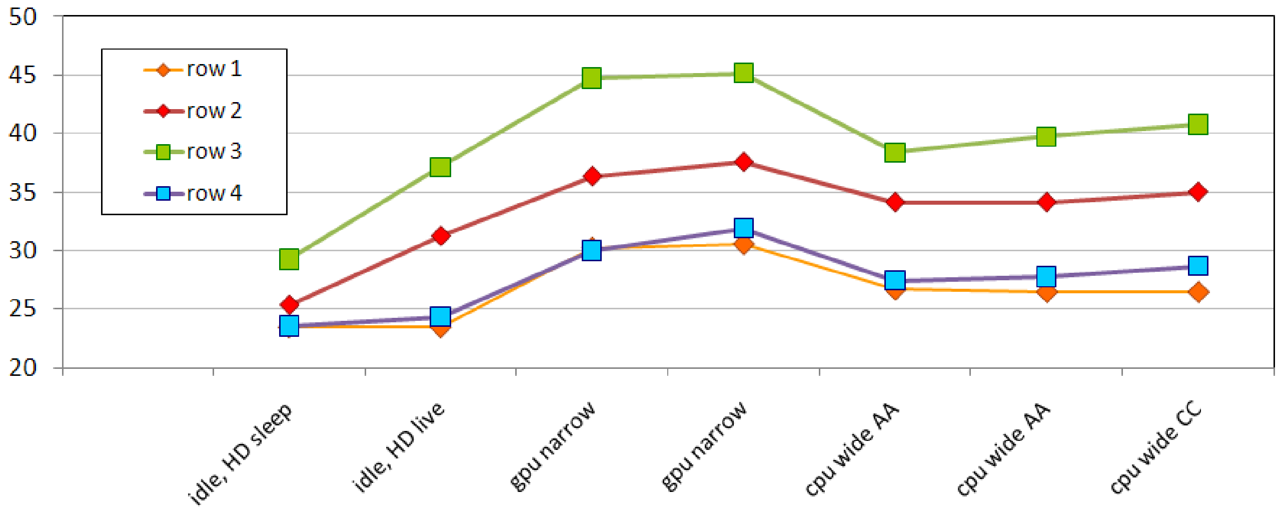
	NOOP	CPU	GPU	CPU/GPU	dC/dG
narrow	156	1530	331	4.62	7.85
wide	258	1570	417	3.76	8.25

- Comparisons to GrayWulf (on partition A01):

	Narrow				Wide			
	NOOP	CPU	GPU	delta	NOOP	CPU	GPU	delta
Amdahl01	156	1530	331	1374	258	1570	417	1312
GW	58	289		231	127	288		161
(G/A)	0.37	0.19	0.87	0.17	0.49	0.18	0.69	0.12

Power Consumption

- Total power between 885W and 1261W
- Lowest Row1 and Row4, highest Row3 (2x1TB)
- High GPU load contributes about 5W extra
- Little difference between lo and hi CPU loads



Conclusions

- 13 times more real sequential performance than a GrayWulf at a constant power!
- Hybrids give best tradeoff

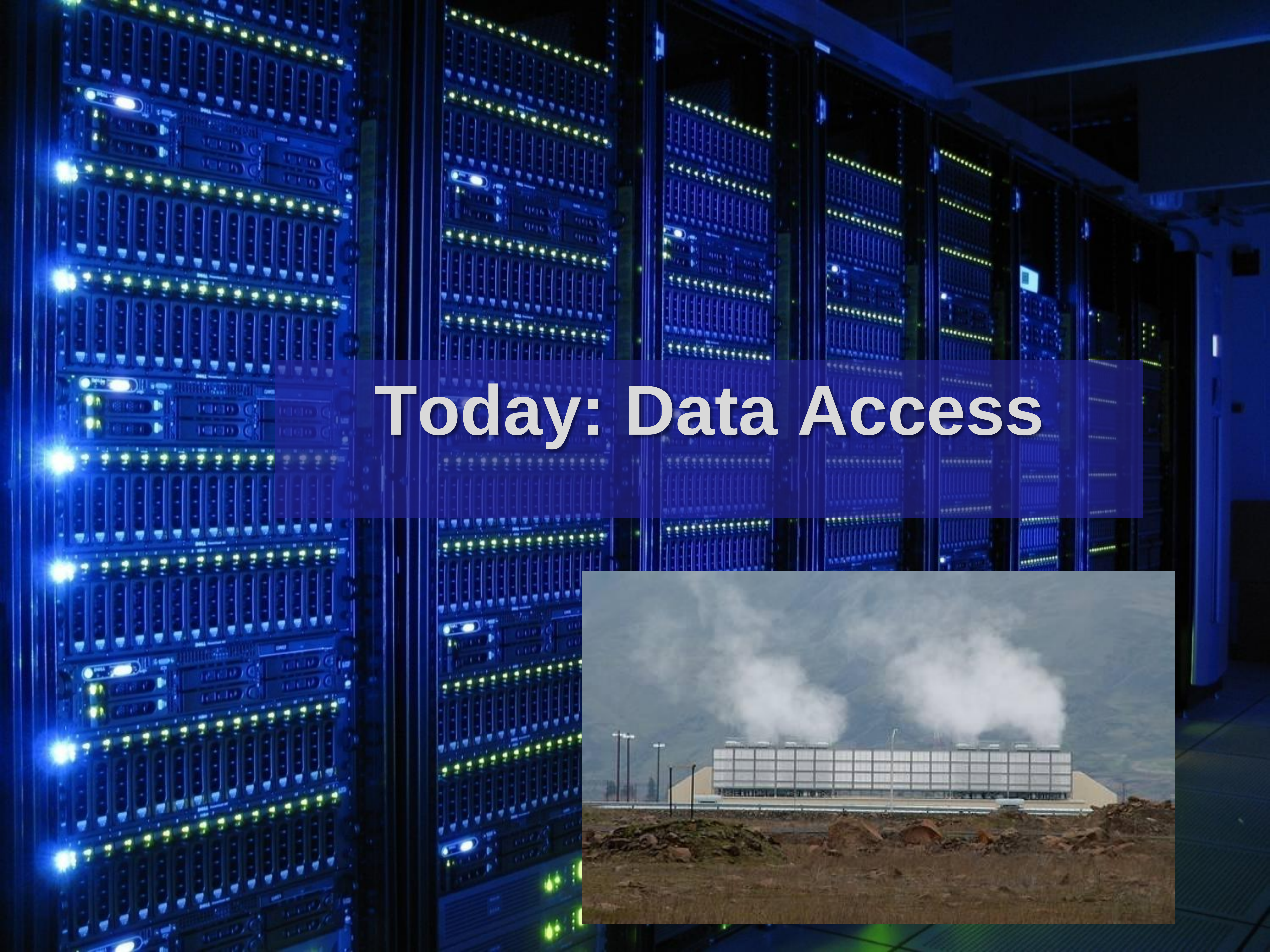
1SSD + 2HD:	450MB/s	2.12TB	40W
2SSD + 1HD:	500MB/s	1.24TB	34W

- Real life, compute/data-intensive astronomy analysis in <5 mins lapse time, processing 1.2TB and doing 6.4B tree traversals
- GPUs: a factor of 8 speedup in a complex task

"Down-and-Out and Low on Power"

Yesterday: CPU cycles

Today: Data Access



Tomorrow: Power