Balanced Scalable Architecture for Data-Intensive Supercomputing

Alex Szalay (JHU)
Howie Huang(GWU)
Andreas Terzis(JHU)

Scientific Data Analysis Today

- Scientific (and other) data double every year
- Trend driven by inexpensive sensors
- Data is everywhere, never will be at a single location
- More data-intensive scalable architectures needed

- Most scientific data analysis done on small to midsize BeoWulf clusters, from faculty startup
- Universities hitting the “power wall”
- **Not scalable, not maintainable…**

- *How to build a scalable, data-intensive architecture?*

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)

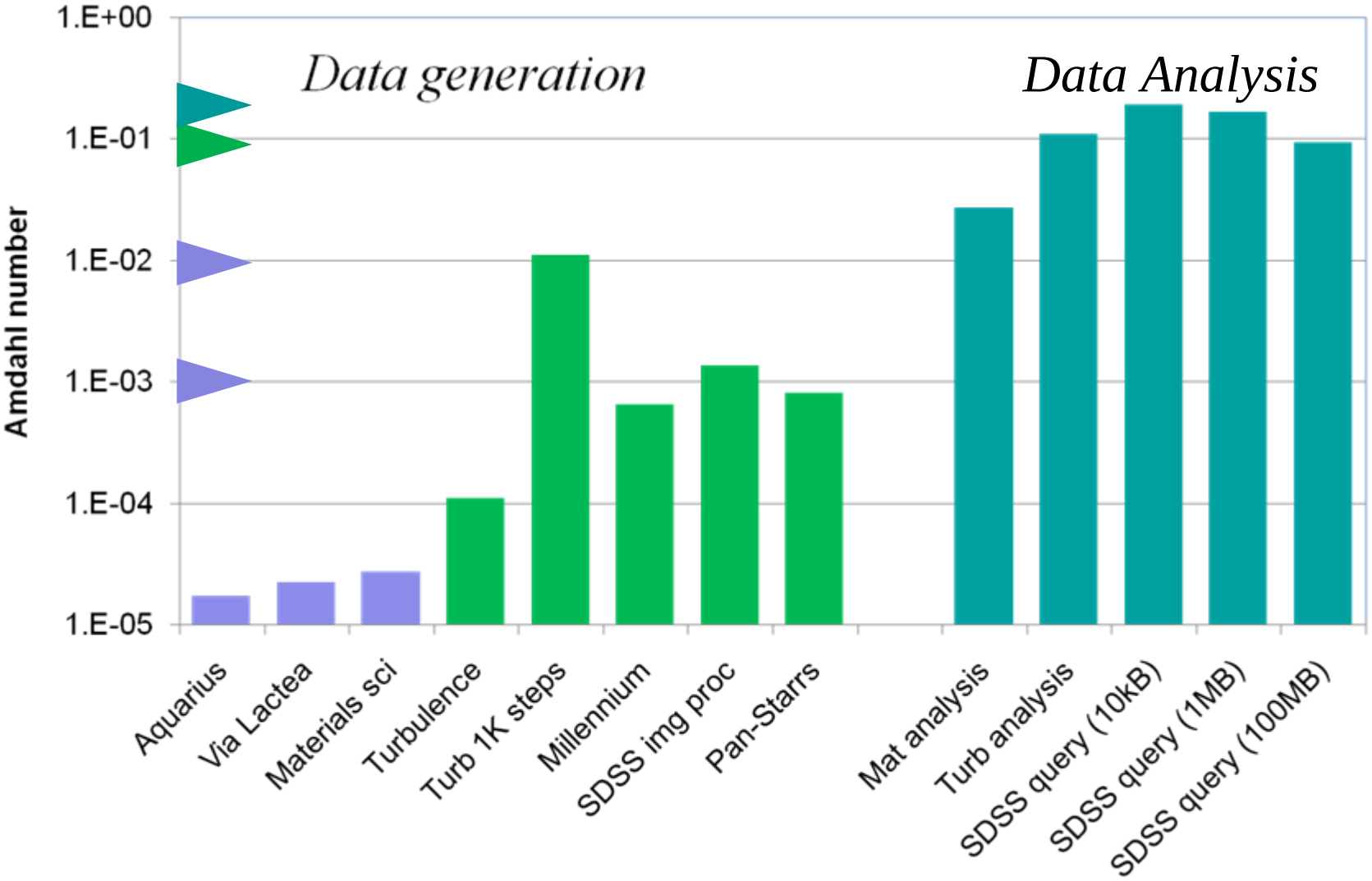
iii. One byte of memory per one instruction/sec (MEM)

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

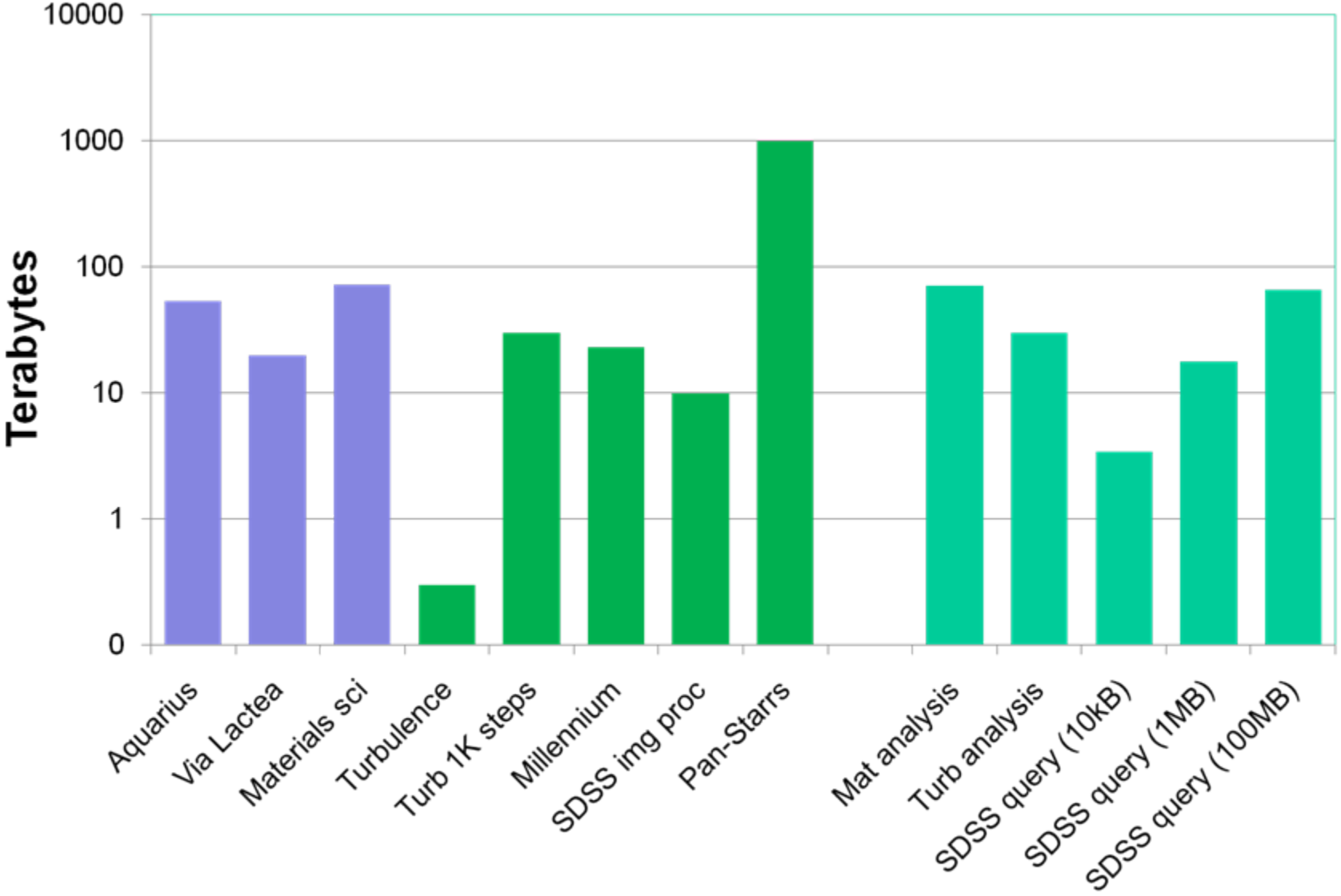
Typical Amdahl Numbers

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



Emerging Trends

- Large data sets are here, solutions are not
- National infrastructure does not match needs
- Facilities focused primarily on CPU cycles
- Even HPC projects choking on IO
- Sociology:
  - *Data collection in larger collaborations: Virtual Observatories*
  - *Analysis decoupled, off archived data by smaller groups*
  - *Data sets in 100s of TB, soon PBs*
- Scientists are “cheap”, also pushing to the limit
- A similar vacuum led to BeoWulf ten years ago

Prelude: GrayWulf

- Implement Jim Gray's vision of data-intensive, scale-out computing
- Build a machine with a high Amdahl number
- Modular commodity components (Dell)
- Create a scalable architecture
- 50 servers, 1PB disk, 500 CPU
- Connected with 20 Gbit/sec Infiniband
- 10Gbit lambda uplink to UIC
- Implement various reference projects

Moore Foundation, Microsoft and Pan-STARRS

graywulf

GrayWulf Performance

- Demonstrated large scale computations involving ~200TB of DB data (won SC08 Storage Challenge)
- DB speeds close to “speed of light” (72%)
- Scale-out over SQL Server cluster
- Aggregate I/O over 12 nodes
  - *17GBps for raw IO, 12.5GB/s with SQL*
- Scales to over 70GB/s for 46 nodes from <$700K
- Very cost efficient: $10K/GBps
- Excellent Amdahl number : 0.56
- **But: hitting the “power wall”!!!!**

20 Years - Unleashing the Power of HPC

SC08

Storage Challenge Winner

Alexander Szalay, Maria Nieto-Santisteban, Jan Vandenberg, Alainna Wonders, Randal Burns, Eric Perlman, Ani Thakar, Mike McCarty, Dean Zariello
Johns Hopkins University
Gordon Bell, Tony Hey, Roger Barga, Yogesh Simmhan, Michael Thomassy, Lubor Kollar,
Microsoft Corporation
Catherine Van Ingen, Microsoft Research
Robert Grossman, David Hanley, Yunhong Gu, Michael Sabala
University of Illinois at Chicago
Jim Heasley, University of Hawaii
Tim Carrol, Eric Barnes, Mike Rowland
Dell Inc.

GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

Patricia J. Teller — Conference Chair
Raymond L. Paden — Storage Challenge Chair

Austin Convention Center • November 15 - 21, 2008

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, $300*
  - *Power consumption 0.2W idle, 1-2W under load*
- Low power motherboards
  - *Intel dual Atom N330 + ION chipset 28W at 1.6GHz*
- Combination is 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

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	

Scaling: Sweet Spot Found

Scaled to a fixed sequential read rate

system	CPU[GHz]	seqIO[GB/s]	kIOPS	disk[TB]	power[W]	cost [$]	rel. power	nodes
GrayWulf	21.3	*1.5*	6	22.5	1150	19250	1.000	1
ASUS	19.4	*1.5*	56	3.0	230	9920	0.200	12
Intel	9.6	*1.5*	30	1.5	84	3530	0.073	3
Zotac	9.6	*1.5*	31	1.5	90	3570	0.078	3
AxiomTek	20.0	*1.5*	50	3.1	188	12430	0.163	12.5
Alix 3C2	30.0	*1.5*		0.5	240	19510	0.209	60

Cost contains 3 years of operation plus HW

Status

- Compared many low power motherboards, SSDs
- Building 50 node cluster for under $50K
- Zotac Atom/Ion motherboards received from NVIDIA
  - *N330 dual core CPU + 4GB memory*
  - *16 GPUs with integrated memory controller*
  - *3 SATA ports*
- Adding two OCZ Vertex drive we measured
  - *500MB/sec sequential read*
  - *400MB/sec sequential write*
  - *20,000 IOPS*
  - *28W power consumption*
- Just got HECURA grant!

Plans

- Evaluate 50 node cluster performance
  - *Try both pure SSD and hybrid nodes*
  - *Projections: 25GBps seq read, 1500W power*
- Run low level benchmarks (Win7, LINUX)
- Build generalized “data crawler”
- Experiment with scalability
- Compare various real life apps, using both DB and Hadoop based solutions

Summary

- Science community starving for storage and IO
  - *Data-intensive computations as close to data as possible*
- Need an objective metrics for DISC systems
  - *Amdahl number appears to be good match to apps*
- Current architectures cannot scale much further
  - *Need to get off the curve leading to power wall*
- Future in low-power, fault-tolerant architectures
  - *We propose scale-out "Amdahl Data Clouds"*
  - *On our way to a medium size testbed*
- Real reference applications for objective metrics
  - *Use large data sets for scalability studies (100TB+) e.g. SDSS, Pan-STARRS, Sensors, Turbulence*