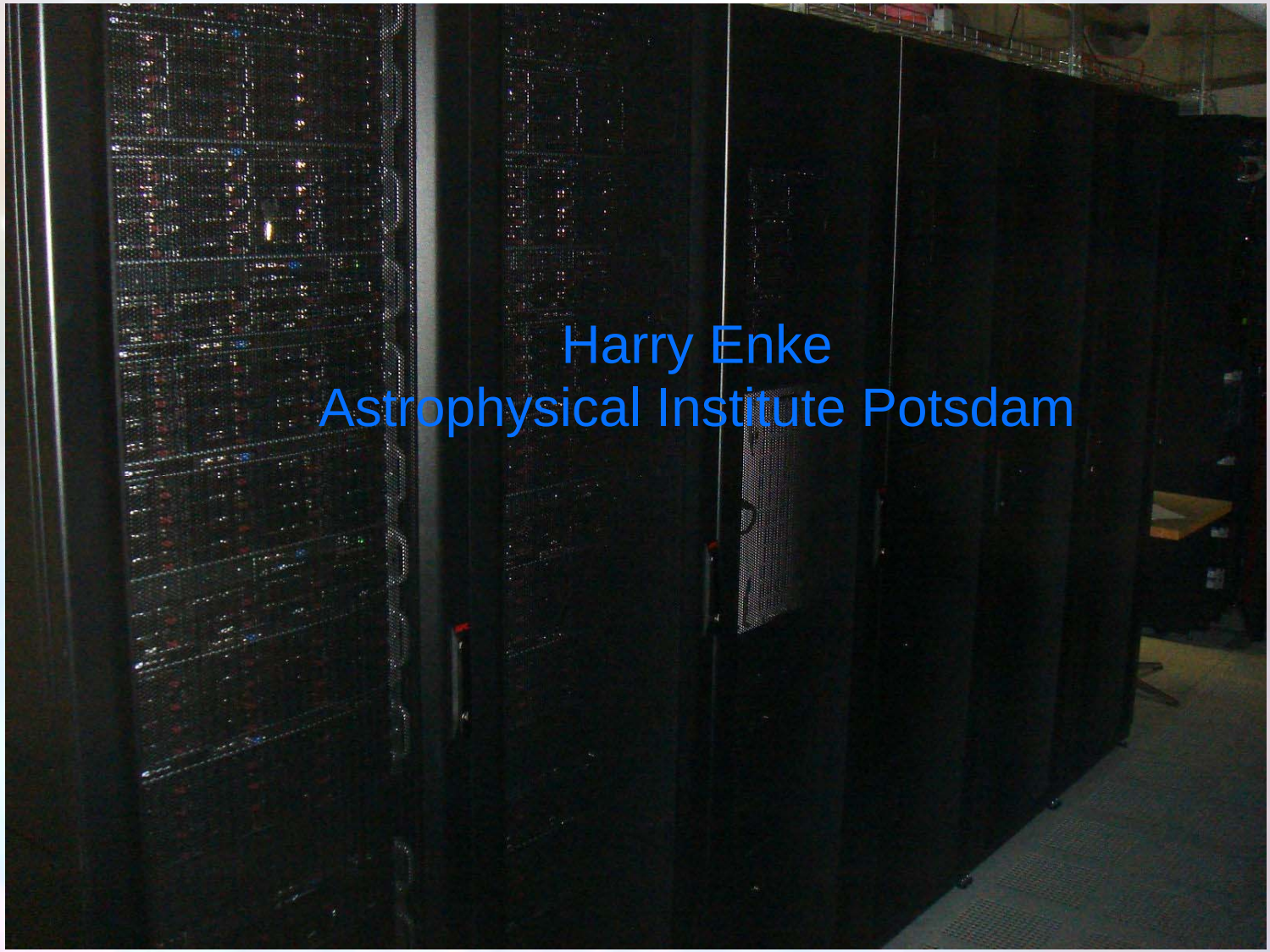
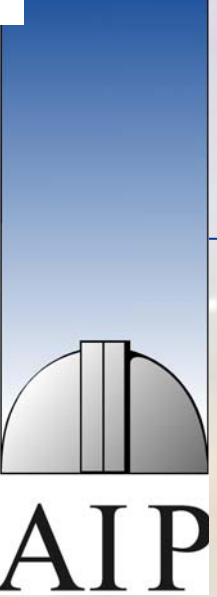
Almagest



Harry Enke
Astrophysical Institute Potsdam

Almagest - AIP Storage Cluster



AIP is building a new Storage cluster,

- dedicated to astronomical databases for
- current (SDSS) and
- upcoming PanSTARRS (and LSST) data

- Part of the Cluster is financed by BMBF
  - (Grid SI2008)
- The Storage Cluster is based on several concepts developped from the SDSS management group at JHU, summarized as the GrayWulf design
- The JHU cluster with this design won the SC‘08 Storage Challenge

Almagest: GrayWulf-Architecture*

Basic ideas, starting from two of Amdahls‘ laws:

A balanced system needs

1 Bit I/O for each CPU cycle (IO)

1 Byte of Memory for each CPU cycle (BW)

The table below shows numbers for some current systems:

System	CPU count	GIPS [GHz]	RAM [GB]	diskIO [MB/s]	Amdahl	
					RAM	IO
BeoWulf	100	300	200	3000	0.67	0.080
Desktop	2	6	4	150	0.67	0.200
Cloud VM	1	3	4	30	1.33	0.080
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

*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

Almagest: GrayWulf-Architecture*

I/O from Hard-Disk to Raid-Controller

- a harddisk has an I/O bandwidth of
  - 75-80 MB/s (seq. read)
  - 50-55 MB/s (write)
- for Raid 0 a jbod with 15 disks should scale linear to 15*75 MB/s = 1125 MB/s (seq. read)
- one SAS lane has a bandwidth of 3 Gb/s, the connection to the raidcontroller on the host has 4 lanes: avaliable bandwidth amounts to 1200 MB/s
- Result: the connection within the JBOD/Controller is sufficient to accommodate for the I/O bandwidth of ~1200MB/s

*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

Almagest: GrayWulf-Architecture*

I/O from Raid Controller to Infiniband

- The current raid controllers saturate at 800-900 MB/s, introducing a serious bottleneck
- PCIx8 slots show 1600MB/s bandwidth
- The PCI bus saturates at 2400 MB/s
- Result: the PCI bus can accomodate for 2-3 RAID controllers
- bandwidth for a copy operation between CPU and Memory is ~2400MB/s, write is ~4100MB/s and read is ~5700MB/s
- infiniband HCAs show 20Gb/s = 2500MB/s bandwidth

*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

Almagest: Hardware

GrayWulf architecture is based on commodity hardware

Principles:

- Each server has a separate raid controller (Areca) for each JBOD
- JBOD to raid controller has 4x3 SAS lanes
- SAS harddisks (Seagate, 1TB) avoid protocol overhead of ~20% compared to SAS/SATA
  - most current raid systems feature SAS interconnect and SATA harddisks

*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

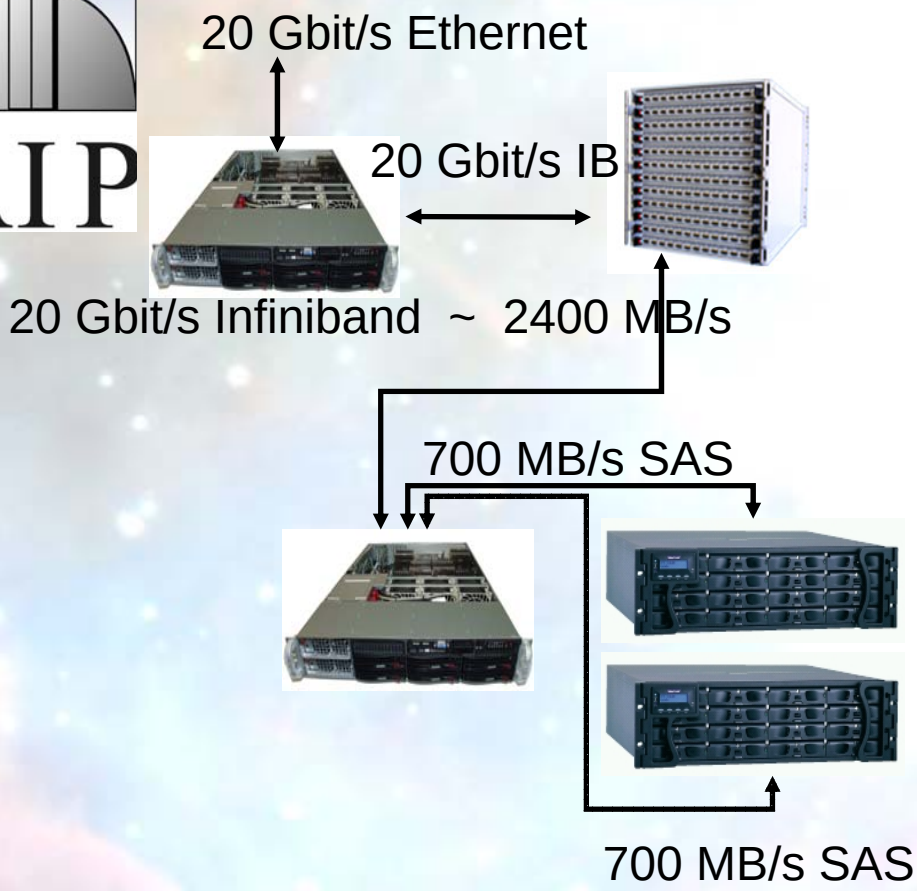
Almagest - AIP Storage Cluster

Hardware - Configuration

- 3 Tiers with different tasks
- Tier 3 shows 21 Servers
- 144-port Infiniband-Switch, aggregated bandwidth ~5.6 TBit/s on backplane
- 2 x 10GBit/sec uplink to Institute backbone via dedicated router

	CPU/Cores	RAM	RAID	Qty.
Tier 1	(4Px4C) 16	128	1 x 15	1
Tier 2	(4Px4C) 16	64	3 x 15	2
Tier 3	(2Px4C) 8	16	2 x 15	21
Total	(54P) 216C	592	735	24

Almagest: Connectivity



Disk-Server-Connection

Serial Attached SCSI (SAS)

- 1 to 3 Storageboxes (JBOD) per Server

- 1 Raid-Controller per Box

Server - Server Connection

Infiniband Socket Direct Protocol (SDP)

- 1 Inifiniband-HCA 20 GBit/s per Server

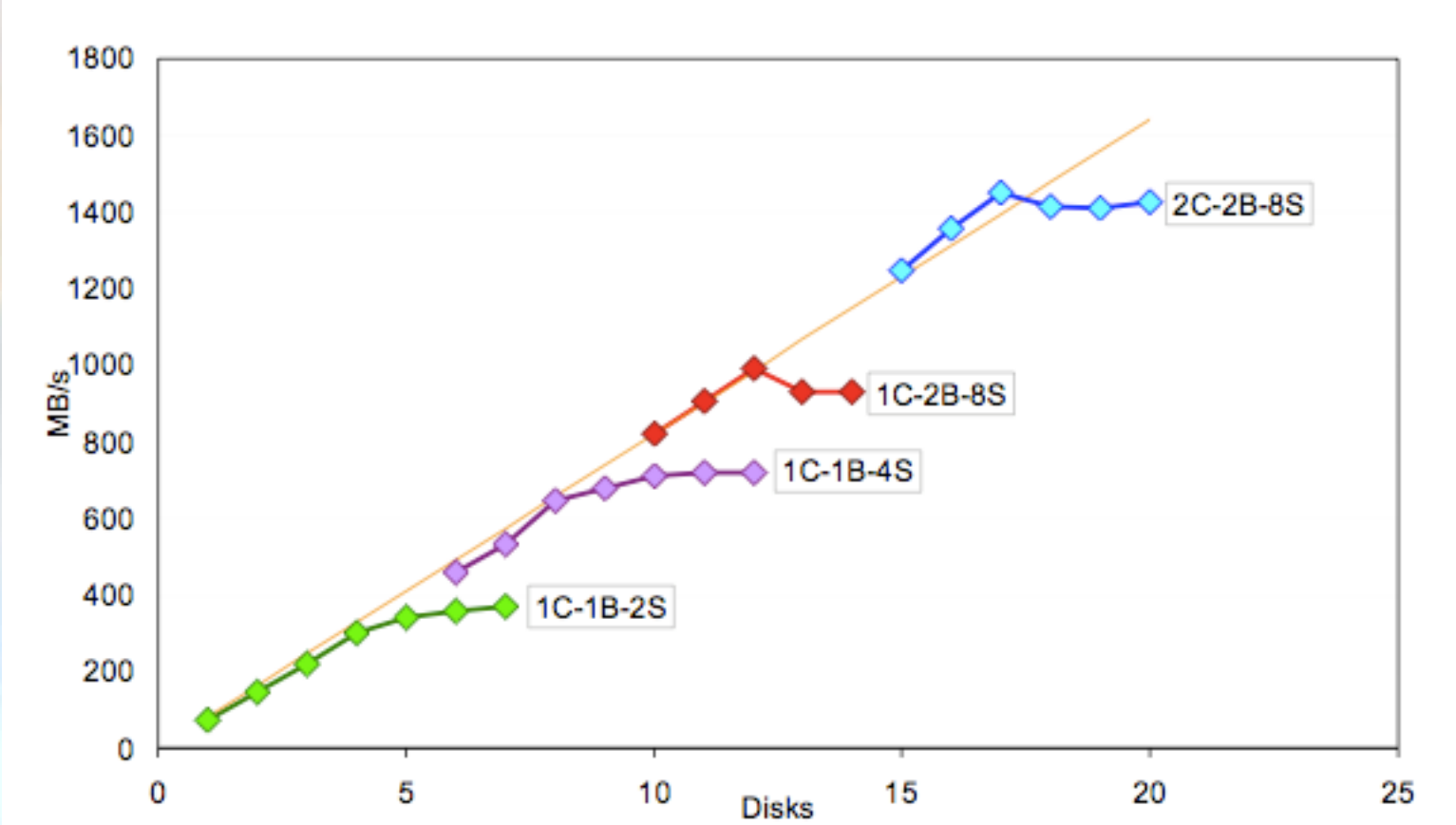
Internet - Infiniband

Ethernet to Infiniband via Router with

2x10 GBit/s Ethernet

2x 20 Gbit/s Inifiband

GrayWulf - Measurements*



Throughput measurements with different configurations
(S=SAS-Lane, B=JBOD, C=RAID-Controller)

*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

GrayWulf - Data Layout (SC'08)*

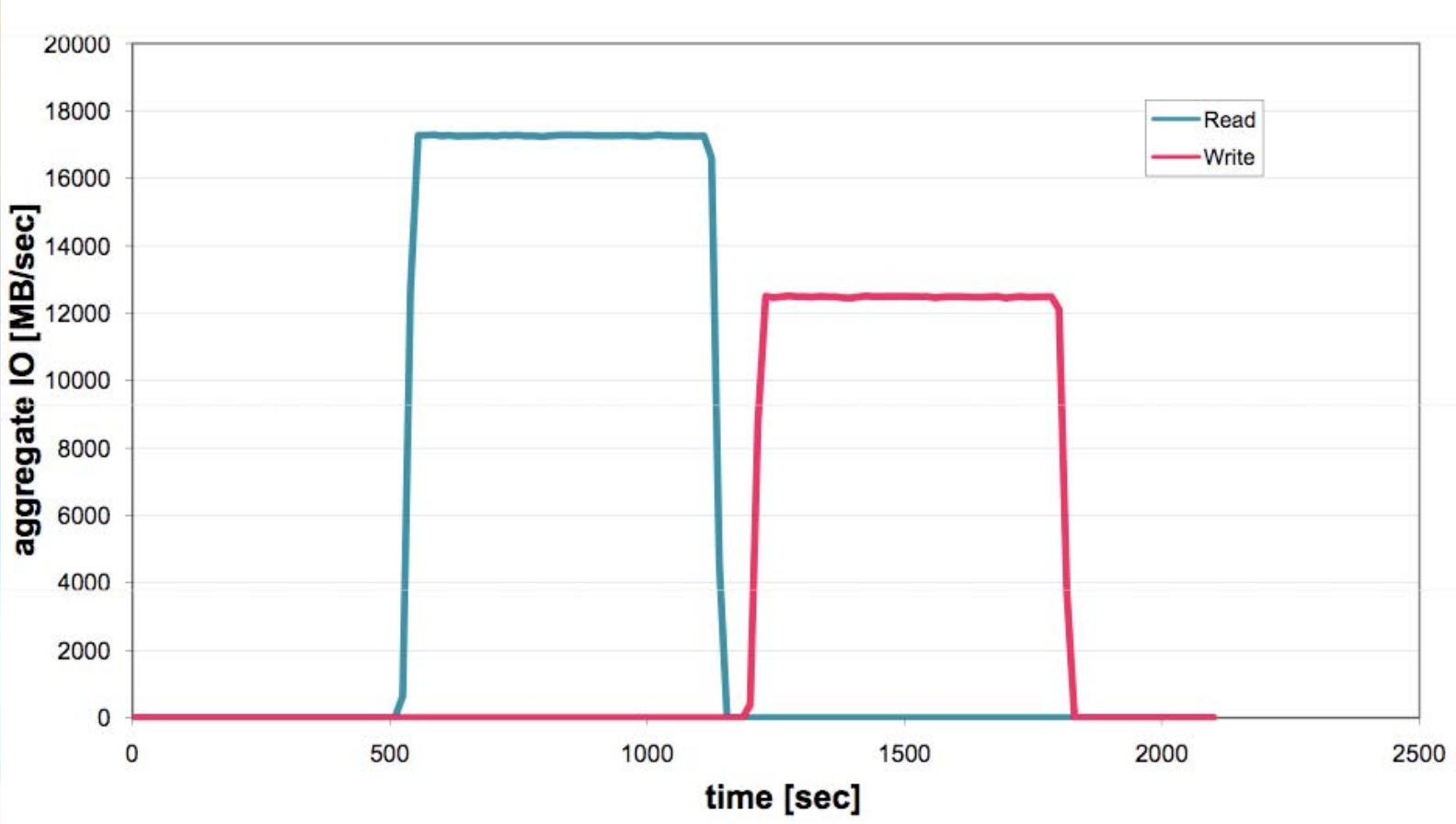
- 7.6TB database partitioned 4-ways
  - 4 data files (D1..D4), 4 log files (L1..L4)
- Data replicated twice to each server (12)
  - IB copy at 400MB/s over 4 threads
- Files interleaved across controllers
- Only one data file per volume
- All servers linked to head node
- All servers linked to head node
- Distributed Partitioned Views

		GW01	
ctrl	vol	82P	82Q
1	E	D1	L4
1	F	D2	L3
1	G	L1	D4
1	I	L2	D3
2	J	D4	L1
2	K	D3	L2
2	L	L3	D2
2	M	L4	D1

*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

GrayWulf - I/O Measurements

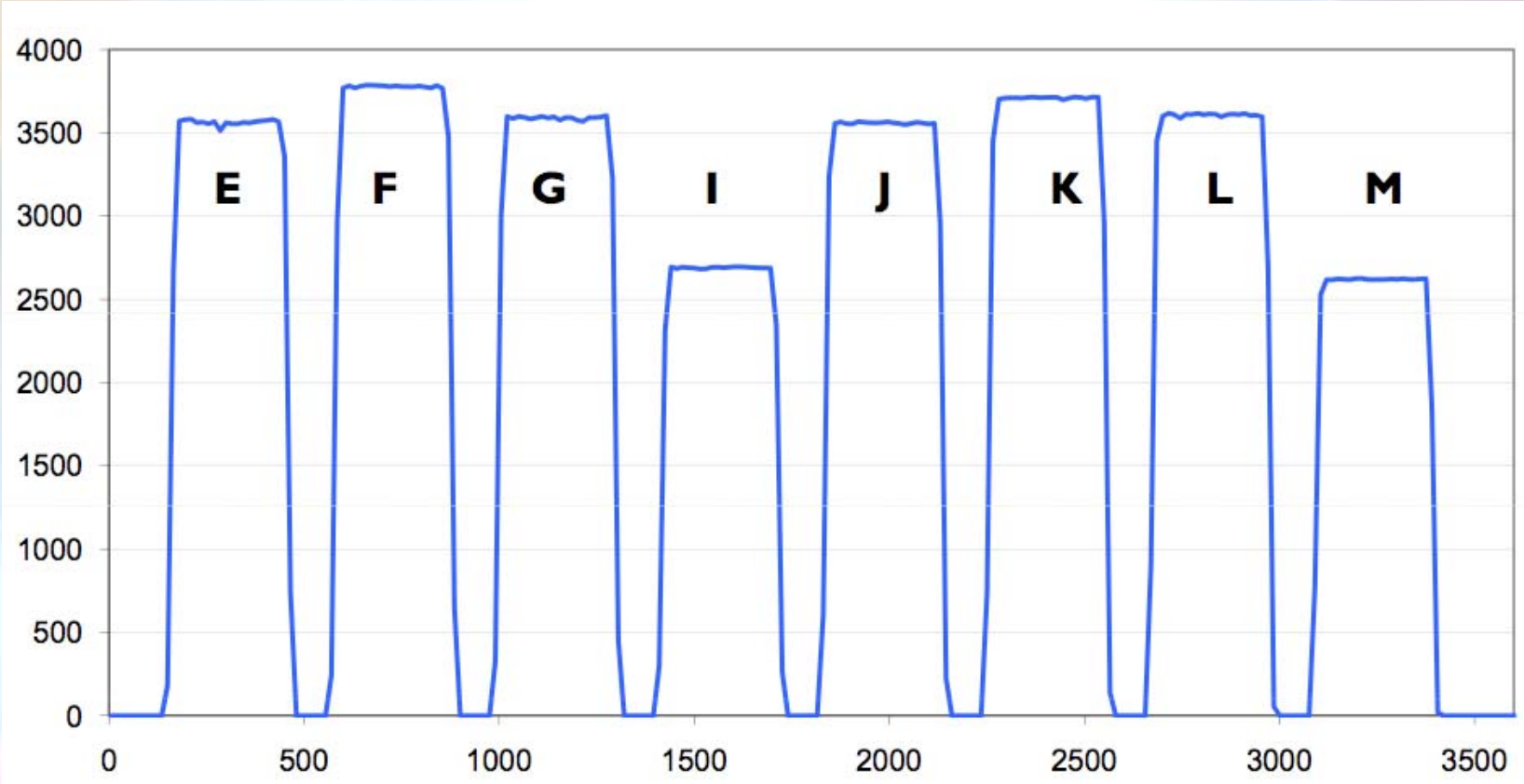
- SQLIO aggregate over 12 nodes



*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

GrayWulf - I/O Measurements

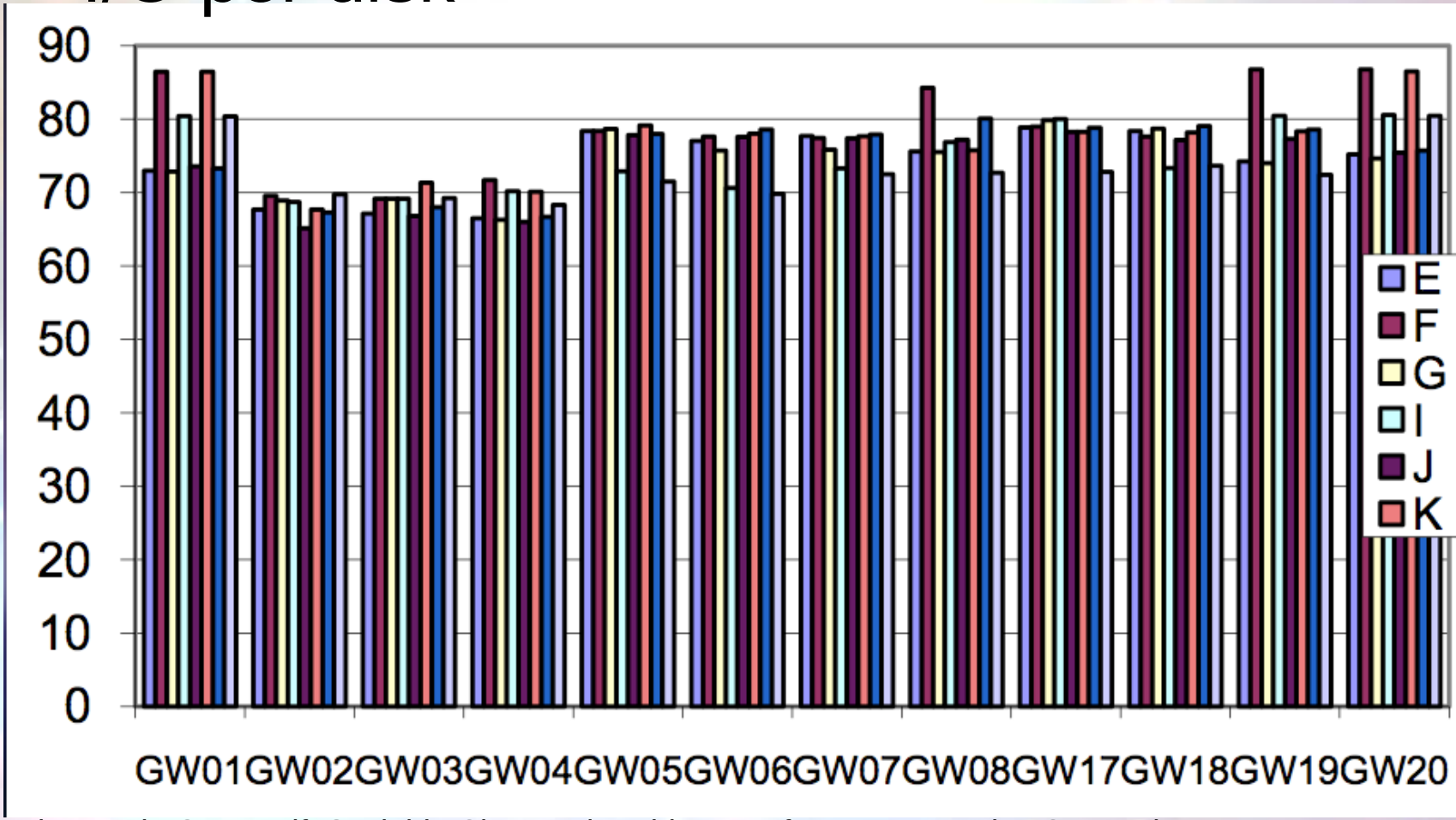
- Aggregate I/O per volume



*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

GrayWulf - I/O Measurements

- I/O per disk



*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

GrayWulf - Software

- Windows Server 2008 Enterprise Edition
- SQL Server 2008 Enterprise RTM
- SQLIO test suite
- PerfMon+ SQL Performance Counters
- Built in Monitoring Data Warehouse
- SQL batch scripts for testing
- DPV for looking at results

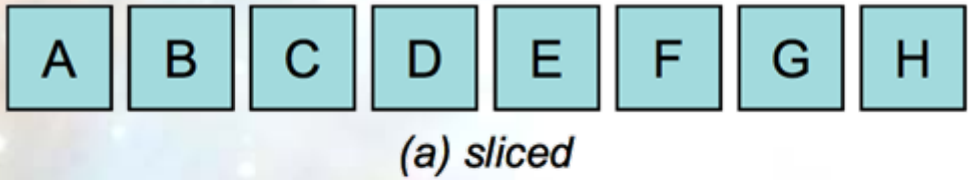
*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

Why DB based Storage Clusters

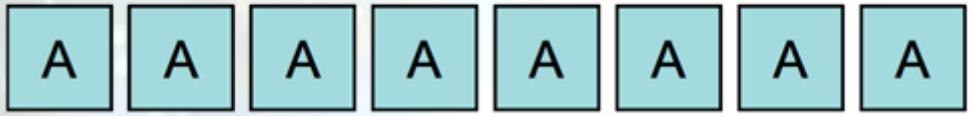
- LSST: 3200 Mpix camera,
  - 20000 sqdeg of the Sky,
  - 30 TB per Night
- PanSTARRS: 1200 Mpix camera
  - Scannig all Sky visible from Hawaii,
  - 6000sqdeg per Night
  - Expected: 40 Billion Objects, w. multiple images
- LOFAR: 150 Gb/s raw input data,
  - up to 20 TB per measurement

Why DB based Storage Clusters

Database Layout for SDSS and PanSTARRS*

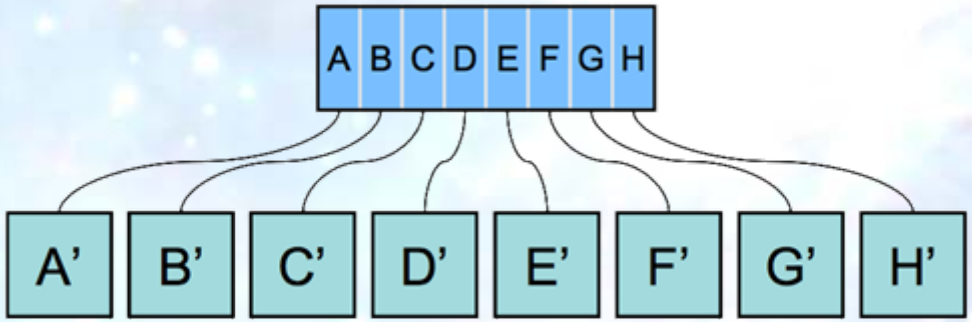


(a) sliced



(b) replicated

SDSS



(c) hierarchical

PanSTARRS

*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing