PHYSICAL STORAGE AND INTERCONNECTS

ALEXANDRU GRIGORE

DAQ WORKSHOP, 13.03.2013

AGENDA

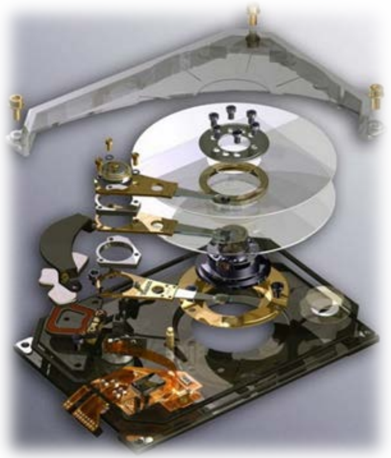
Storage media

- Hard Disk Drives
  present facts, tuning, near future
- Solid State Storage
  advantages, considerations
- Solid Class Memory

Media connectivity

- Trends

+ Experiments information



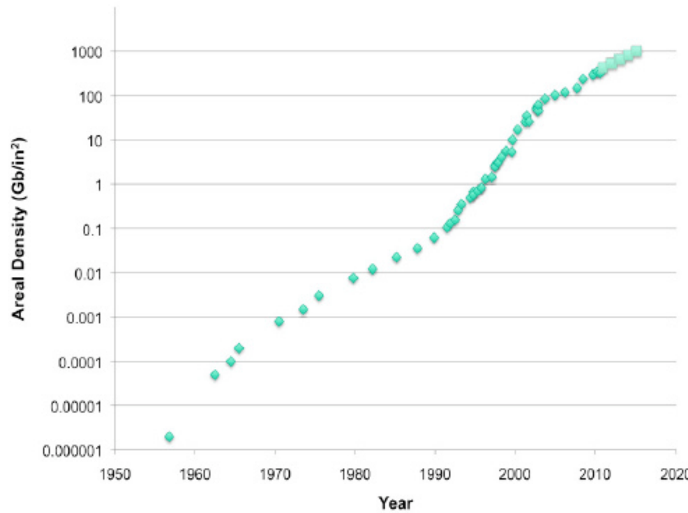
HDD STORAGE FACTS

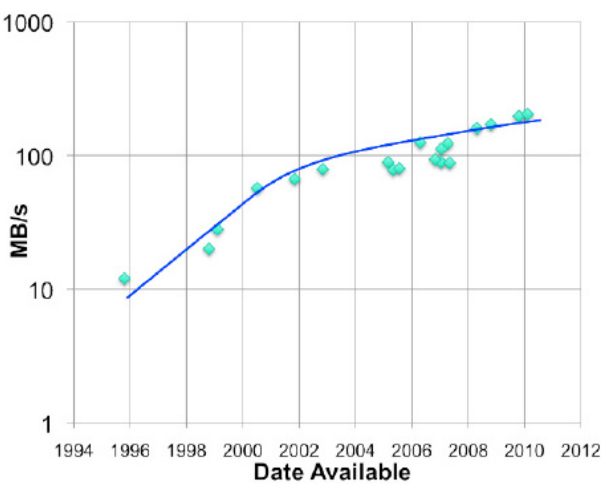
Areal density skyrocketed to 1TB/in^2

Latency has settled down to 2, 3 and 4.1 milliseconds

A high performance HDD would have a max. sustained bandwidth of ~171MB/s

Performance, not capacity, is the issue with HDD storage.





Source: IBM, 2011 “GPFS Scans 10 Billion Files in 43 Minutes”

HDD STORAGE SPARING MECHANICS

RAID

- Stripe data across multiple disks

Short stroking

- Place data only on outer section of disk platters to reduce the seek time

Thin Provisioning

- Use virtual volumes to constrain data to small areas of a physical disk

Parallel file systems

- Link multiple individual disk spindles to increase system throughput

Virtualization

- Use software to spread volumes across multiple drives; a more sophisticated extension of RAID 0

Faster drives

- 15K RPM drives over 10K RPM drives

HDD STORAGE NEAR FUTURE

Writing

- Heat Assisted Magnetic Recording

Mechanics

- Magnetic fluid bearings

Cooling

- Disk based cooling with helium

Capacity

- 60TB hard drives on their way
  ++ challenges in RAID
  [Uli's presentation]

RPMs & Hybrids

- A slower drive might appear in the enterprise space
- If this happens it will have been caused by the combination of solidstate storage and slower disk drives into storage systems that provide the same or better performance to cost ratio.

SOLID STATE STORAGE
FLASH

	SLC	MLC-2	MLC-3	MLC-4
Bits per cell	1	2	3	4
Performance	Fastest	⟷	⟷	Slowest
Endurance	Longest	⟷	⟷	Shortest
Capacity	Smallest	⟷	⟷	Largest
Error Prob.	Lowest	⟷	⟷	Highest
Price per GB	Highest	⟷	⟷	Lowest
Applications	Enterprise	Mostly Consumer	Consumer	Consumer

	$/GB	$/IOPS	IOPS/watt
SSD (SLC)	$5 – $40	$0.005 – $0.15	1000 – 15000
SSD (MLC)	$0.63 – $4	$0.004 – $0.05	1000 – 15000
HDD (enterprise)	$0.50 – $1	$1 – $3	10 – 30
HDD (desktop)	$0.05 – $0.37	$1 – $4	10 – 40

Source: Demartek 2012

FLASH MEMORY ARRAYS

RAMSAN 820
TEXAS MEMORY INSTRUMENTS

450.000 IOPS for reads or writes

sustained, ra

1RU

12 or 24TB us

Up to four 8 G

four 40 Gb/s Q

Integrated we

No single poi

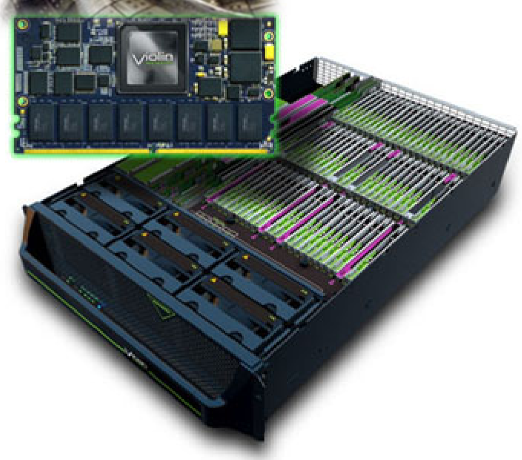
VIOLIN MEMORY
6000 SERIES FLASH

1 Million IOPS

40Gb/s Infiniband

[read]

algorithm(r)



SOLID STATE STORAGE TODAY'S FACTS

Price

- more expensive than hdd [varies a lot f() of the technology].
- steep decrease in price - NAND Flash, 2010 3.2$/GB, 2015 0.5$/GB [Source: IDC, #235155, June 2012]

Endurance

- Single Level Cell 10^5 writes/cell
- Multi Level Cell 10^4 writes/cell
- Triple Level Cell ~300 writes/cell

Reliability

- write amplification
  [wear leveling, data deduplication]

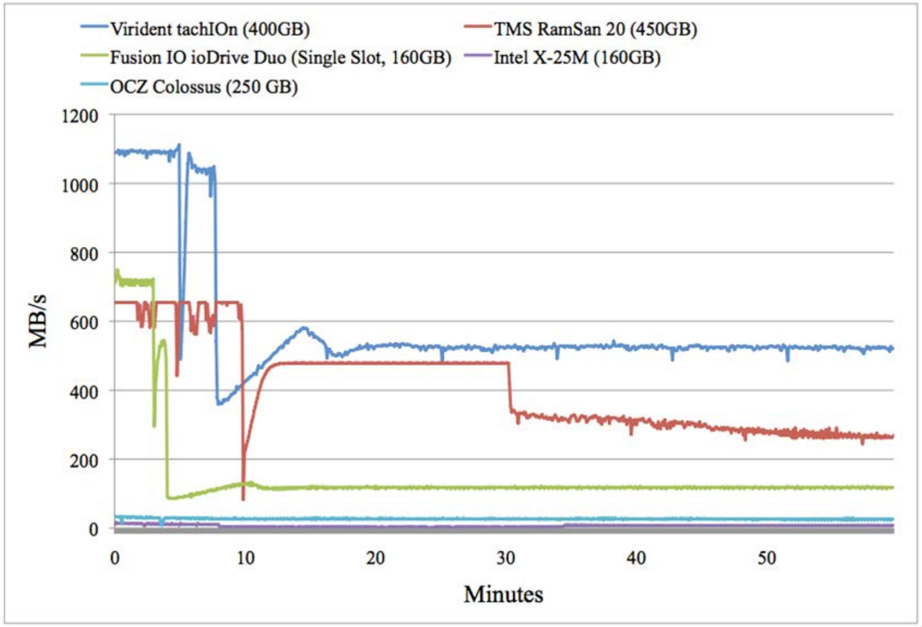


Image source: www.vmem.com

STORAGE CLASS MEMORY THE NEXT STORAGE SYSTEM

Projected characteristics of storage class memory devices

Capacity	1 TB
Read or write access time	100 ns
Data rate	>1 GB/s
Sustained I/O rate [1/(0.1 µs + 4 KB/1 GB/s) = 1/4.1 µs]	238,000 SIO/s
Sustained bandwidth (4 KB/4.1 µs = 975 MB/s)	975 MB/s
Write endurance	10^{12} writes

CPU cycles	*Device*	*Comment*
10^7–10^8	Disk	Nonvolatile, slow, and inexpensive
	—Gap in access time—	
10^3	SCM	Nonvolatile, fast, and inexpensive
10^2	DRAM	Volatile, fast, and expensive
10–100	L2 and L3 cache	Volatile, fast, and expensive
1	L1 cache	Volatile, fast, and expensive

PCM – Phase Change Memory

- The key concept involves the use of certain chalcogenide alloys (typically based on germanium, antimony, and tellurium) as the bistable storage material.
- A PCM-based SCM is expected to have roughly the specifications shown above by 2020.

Source: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5388608

CASE STUDY
ALICE TDS - HDD TO SSD

300GB, SATA, R6

IO transfer size: 16K

Application IO: 50/50

HDD size: 3.5”

RPM: 10k

# of drives: 40*16

% consumed: 10% || 22% ||40%

Maintenance: no

# of instances: 1

128GB SLC, R6

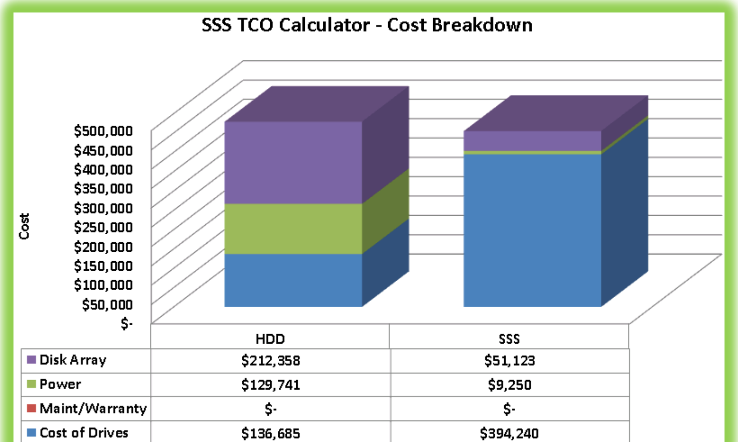
I/O improvement: 171% || 497% || 985%

IOPS gain: 60,541 || 175,611 ||348,216

Form factor: 2.5”

Total SSS: 150 || 330 || 600

Power reduction: 96.5% || 92.4% ||86,1%



TCO f() used storage space:

10% - TCO impact $258,684 (gain)

22% - TCO impact $8,583 (gain)

40% - TCO impact $401,617 (invest)

Source: snia.org/forums/sssi/programs/TCOcalc

EXPERIMENTS STORAGE

ALICE, ATLAS, CMS, LHCb all* use HDDs for data storage.

ALICE: SAN

ATLAS: DAS

CMS: SAN

LHCb: SAN [monolithic]

***LHCb uses an SSD disk pool for metadata storage.**

EXPERIMENTS STORAGE ALICE

75 Disk Arrays x 2FC 4G ports x 3 volumes, R6

SATA II

WD 320GB, ST 2TB HDDs

225 volumes, 610 TB

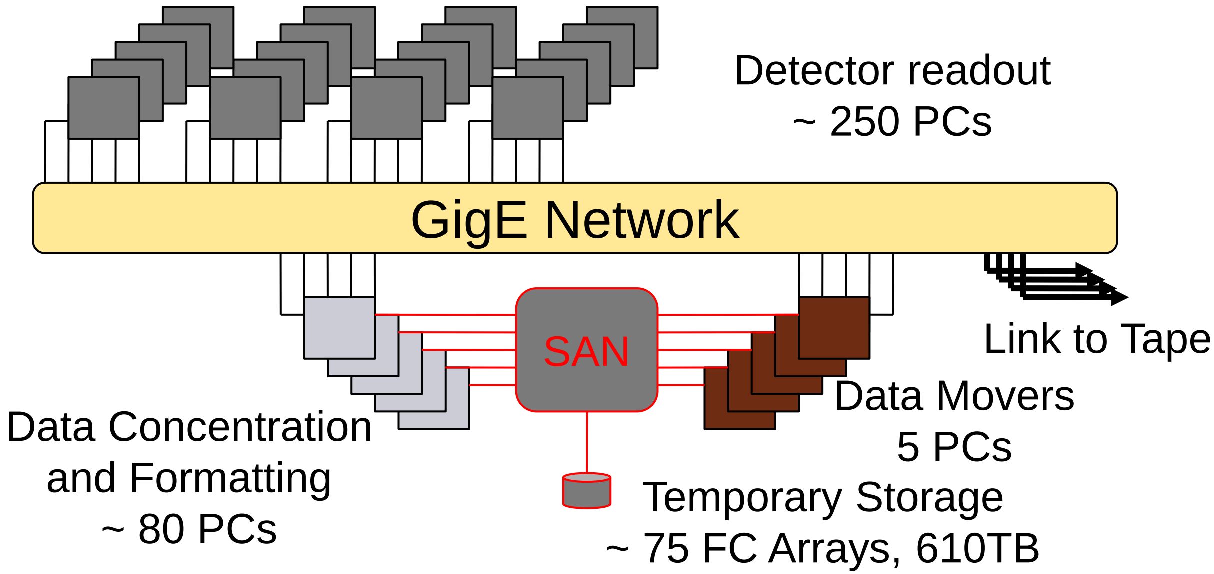
Stornext (affinity)

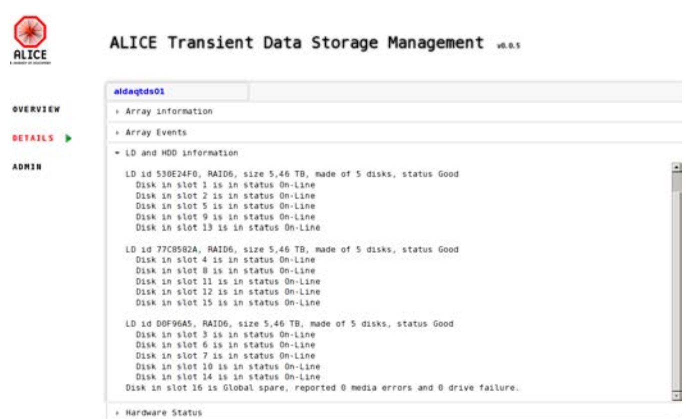
FC4Gb/s and 8Gb/s

Experienced the “RAID write hole” [-> Uli’s presentation]

LS1: same architecture, vertical scaling ?

LS2: in progress





EXPERIMENTS STORAGE ATLAS

4RU

6+3 SubFarmOutputs x

24x1TB HDDs, R5 x

3 RAID controllers

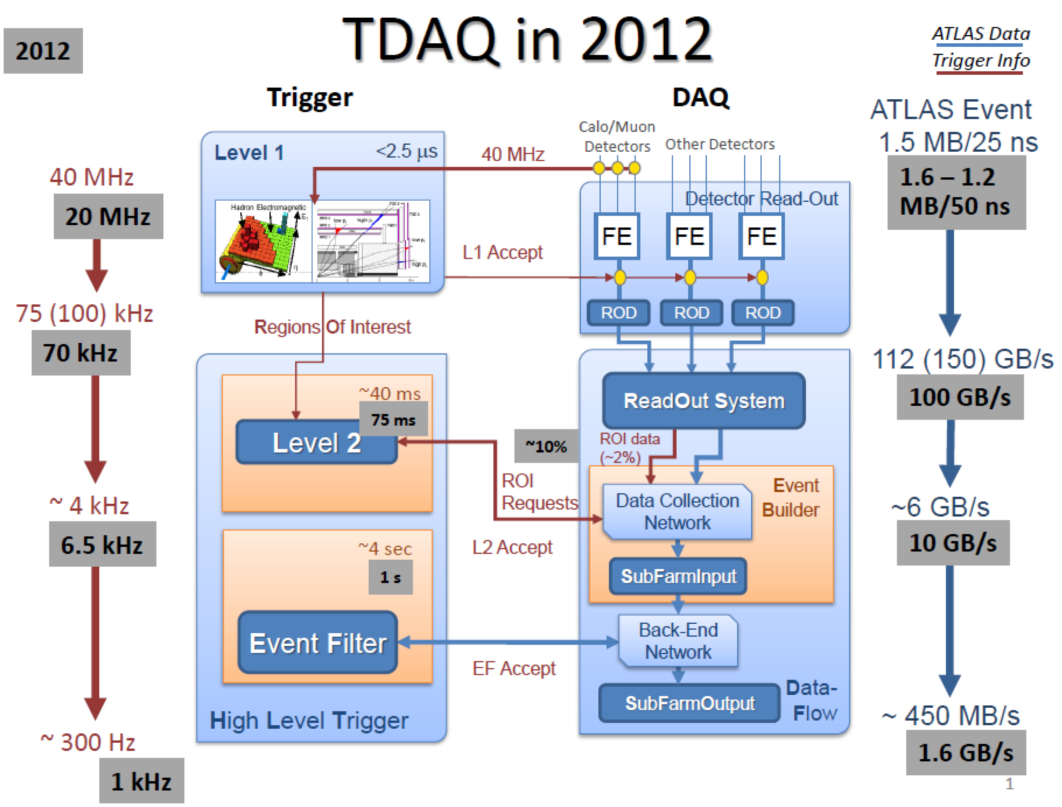
SAS / SATA

Highly scalable and cost effective, add one SFO, get an extra 300MB/s

LS1 – same architecture

Will fine tune horizontal and vertical scaling

LS2 – work in progress



EXPERIMENTS STORAGE CMS

SMs: 16x Dell 2950s

Each 2950 owns 2 SataBeast arrays

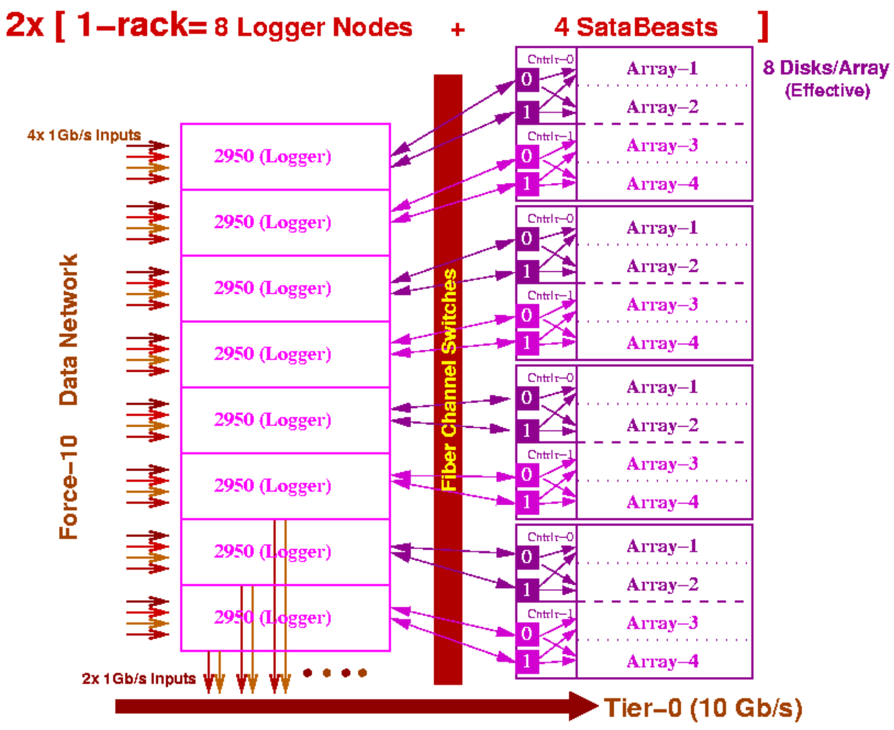
8x Nexsan SataBeast disk enclosures

RAID-6

256TB usable capacity

LS1: Change everything, probably go for NetApp

LS2: Work in progress



EXPERIMENTS STORAGE LHCb

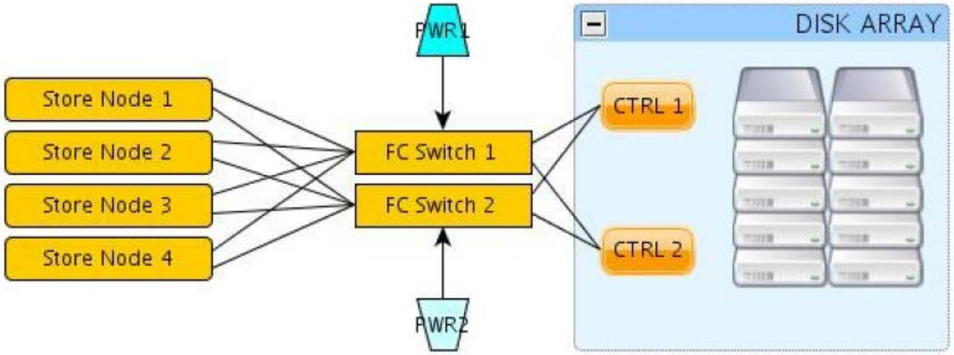
DDN9900, 150HDD [oo. 600], 1TB and 2TB

Storage is used for data, user’s home, programs

Published via NFS and SAMBA

Quantum Stornext

8Gb/s FC



LS1 upgrade

- more HDDs
- change machines

LS2

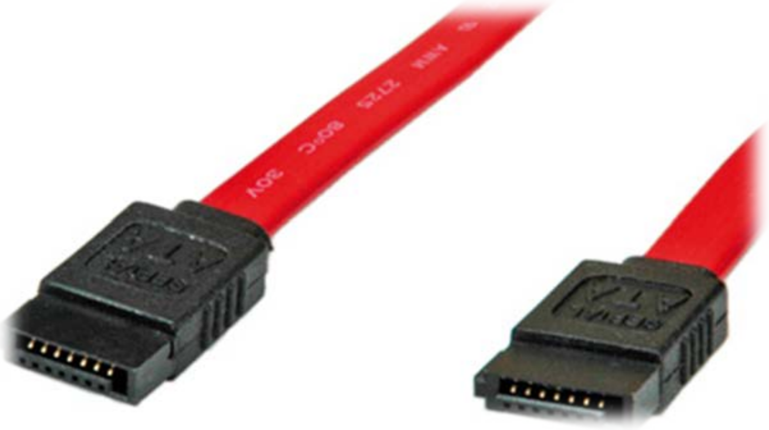
- Probably go to DAS architecture [(c) ATLAS]

Netapp machine with SSD for virtual machines

INTERCONNECTS

Interconnects trends

Experiments info



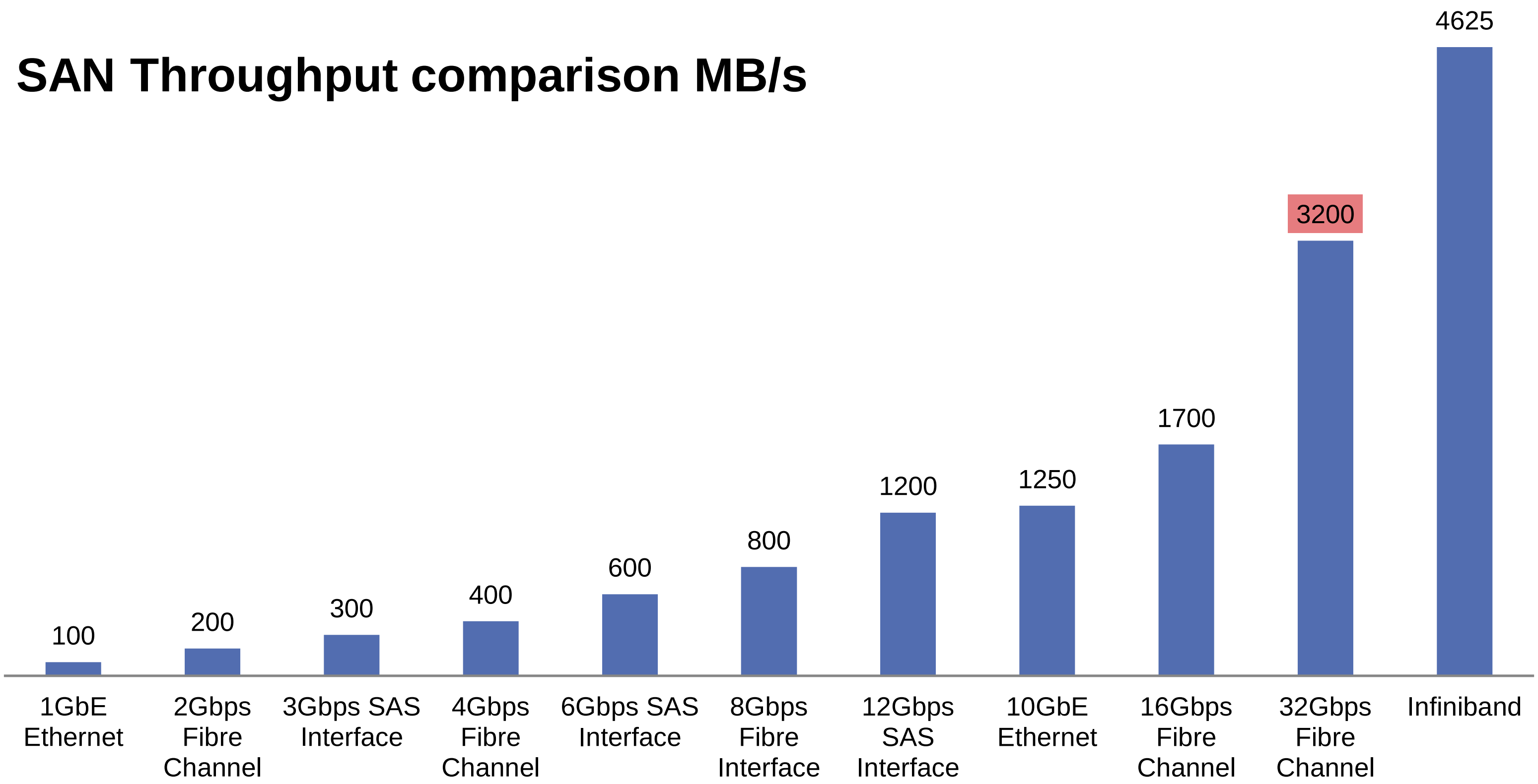
INTERCONNECTS TRENDS

(R)evolutions to notice:

- FC
- PCIe v4
- SATA Express
- SAS [24Gb/s in 2016]

	2005	2007	2008	2009	2010	2011	2012	2013	2014
Storage Networking									
FC	4Gb/s		8Gb/s			16Gb/s			32Gb/s
FCoE				10Gb/s					
IB	20Gb/s		40Gb/s			56Gb/s		100Gb/s	
iSCSI		10Gb/s							
Disk Drive Connectors									
SAS	3Gb/s			6Gb/s			12Gb/s		
SATA	3Gb/s				6Gb/s				
SATA uSSD						6Gb/s			
SATA Express								being ratified	
Host bus									
PCIe		4Gb/s			8Gb/s	v4 approved			

SAN INTERCONNECTS COMPARISON



EXPERIMENTS INTERCONNECTS

NOW [LS1]

HDD

SATA II, SAS, NL-SAS

Storage

FC: 4Gb/s, 8Gb/s

iSCSI

Networking

1GbE, 10GbE

THEN [LS2]

BUZZWORDS

Interconnects

- Intel buys QLogic’s Infiniband segment [Jan 2012]
- Melanox bought Voltaire [Nov 2010] - moving into Ethernet
- Oracle buys 10.2% of Melanox shares [Oct 2010]

SSS

- Violin enters PCIe SSD market [March 2013]
- IBM bought Texas Memory Instruments [Aug 2012]
- Apple bought Anobit [Jan 2012]
- Oracle bought Pillar [Jan 2011]

10GbE, 30GbE

- move from SAN to NAS?

Converged Networking

Unified Storage

Massive Array of Idle Disks – go green

THANK YOU

ALICE DAQ Team

Costin Grigoras [ALICE Offline]

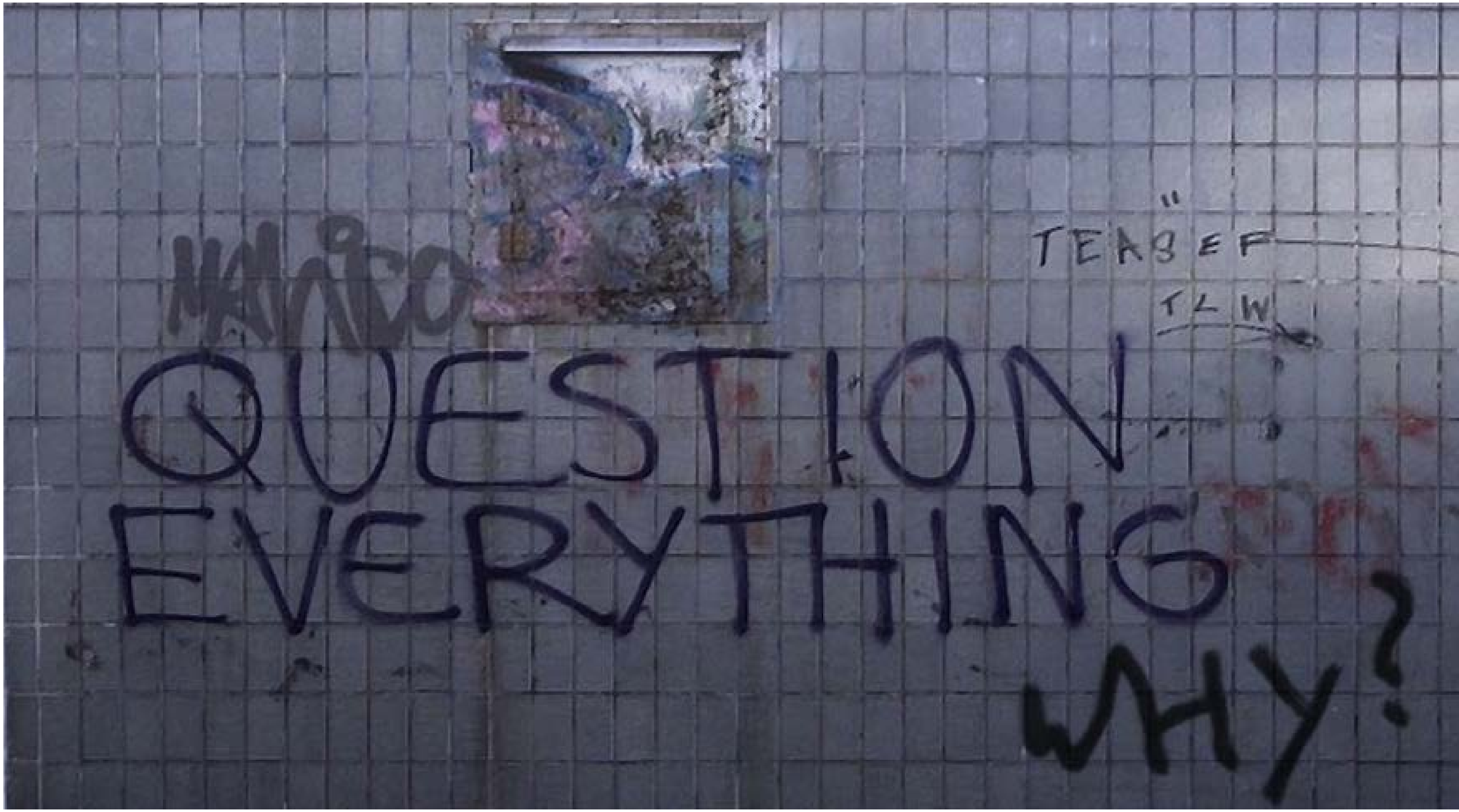
Wainer Vandelli [ATLAS]

Georgiana Lavinia Darlea [CMS]

Olivier Raginel [CMS]

Gerry Bauer [CMS]

Rainer Schwemmer [LHCb]


QUESTION
EVERYTHING
WHY?