Computer Facilities, Production Grids and Networking

Track Summary

Maria Girone
Daniele Bonacorsi
Andreas Heiss

Computer Facilities, Production Grids and Networking

Track Summary

Usual disclaimers apply!
(Sorry, no poster coverage, biased, incomplete, my personal view etc.)

Maria Girone
Daniele Bonacorsi
Andreas Heiss

Statistics

98 contributions

- 19 accepted as oral
- 79 posters

4 sessions

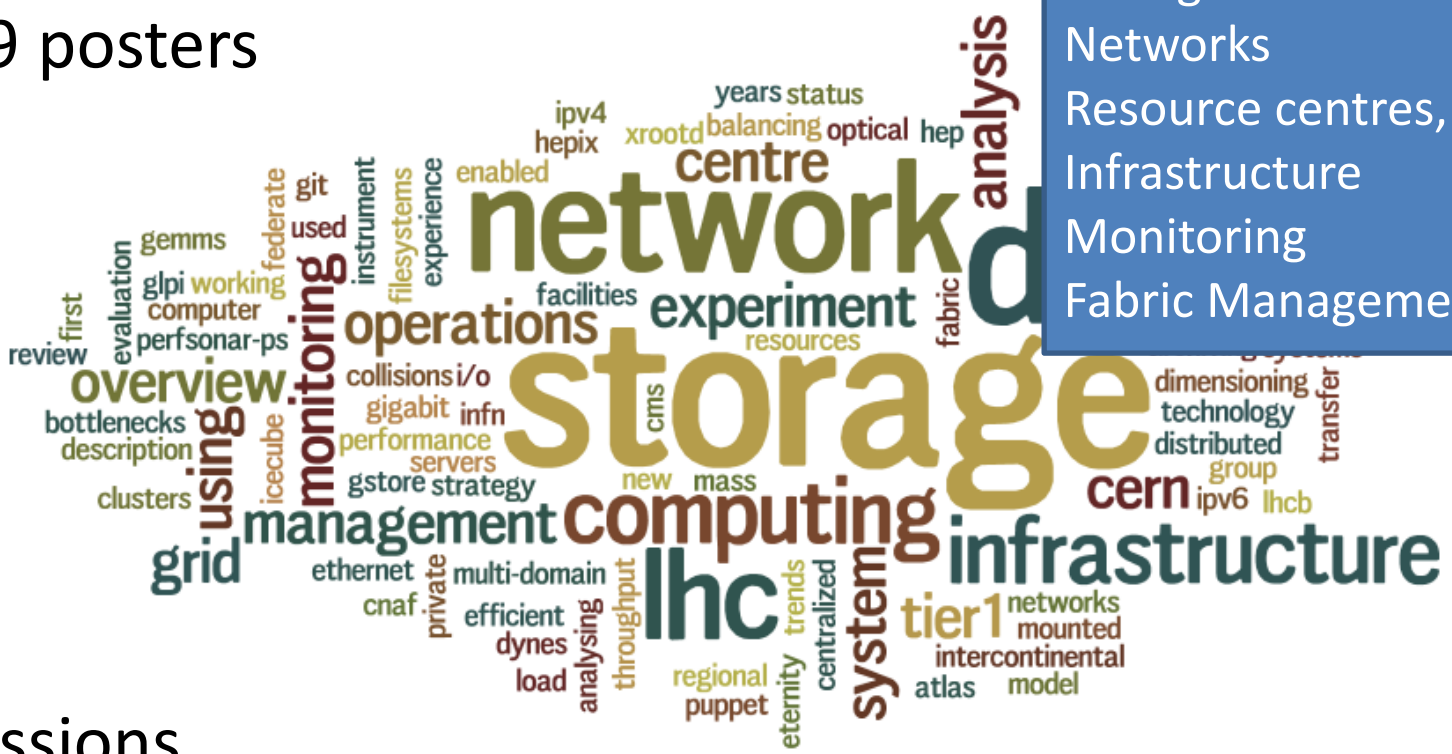
Audience:

~ 70 – 80 people

Statistics

98 contributions

- 19 accepted as oral
- 79 posters



Relative # of occurrences	
Storage	24%
Networks	17%
Resource centres, Infrastructure	16%
Monitoring	10%
Fabric Management	9%

4 sessions

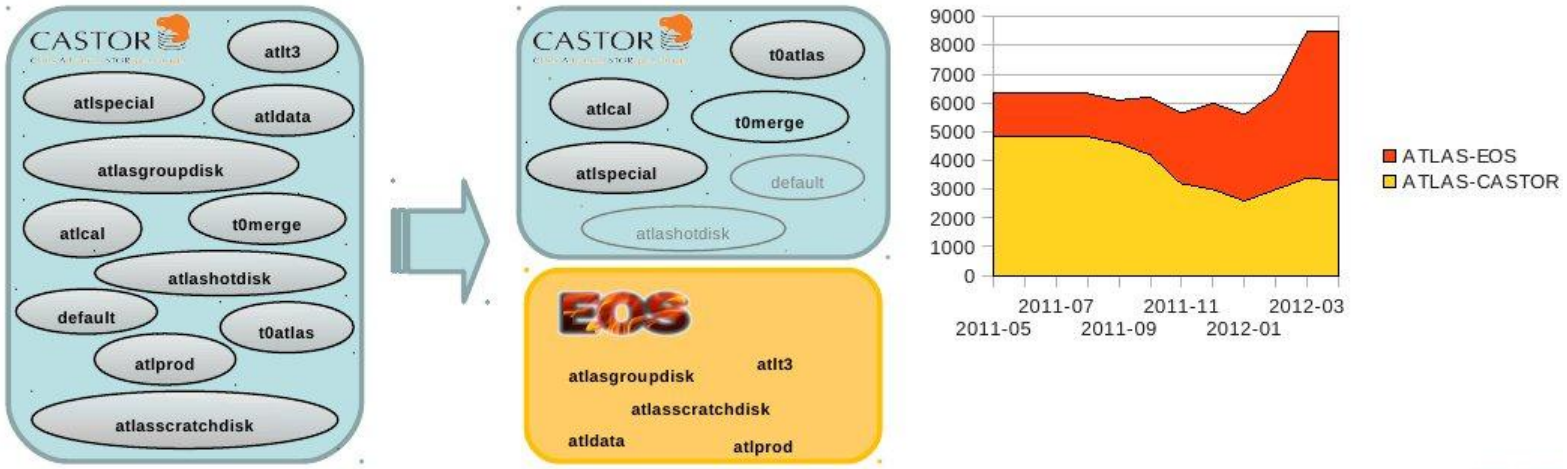
Audience:

~ 70 – 80 people

Storage Systems

Jan Iven: Overview of Storage Operations at CERN

- Castor HSM is optimized for Tier-0 flow, not for “random“ user analysis.
- Strong increase in such analysis suggested to introduce second type of storage System. -> EOS = xrootd + in-memory namespace “plugin“.



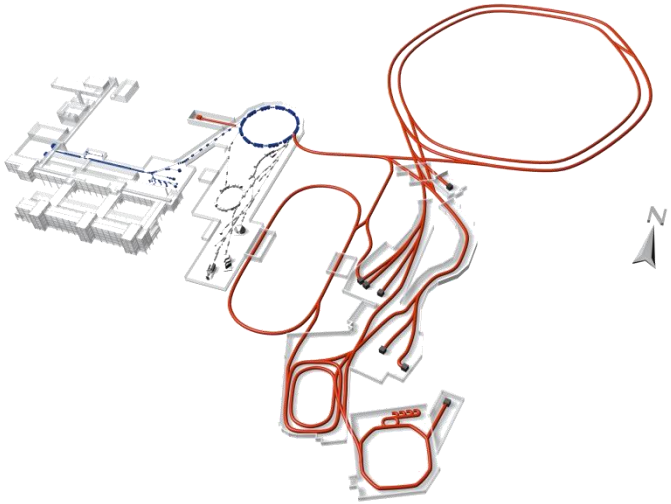
(credit: L.Mascetti)

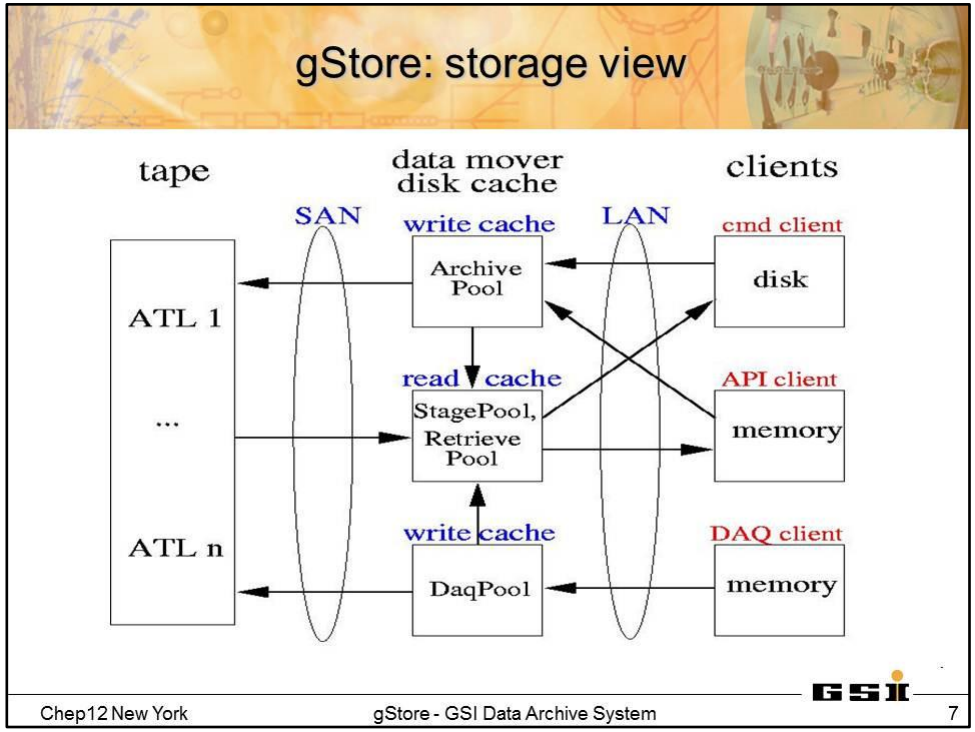
Experience of >1 year of EOS:
- easy setup and updating (no DB components)
- easier server draining
- less support requests

Storage Systems

Horst Göringer, GSI: High Performance Experiment Data Archiving with gStore

- Scaleable, local HSM based on TSM
- Easy scaleable data moving to Lustre for analysis.
- FAIR 2018: 33PB per year

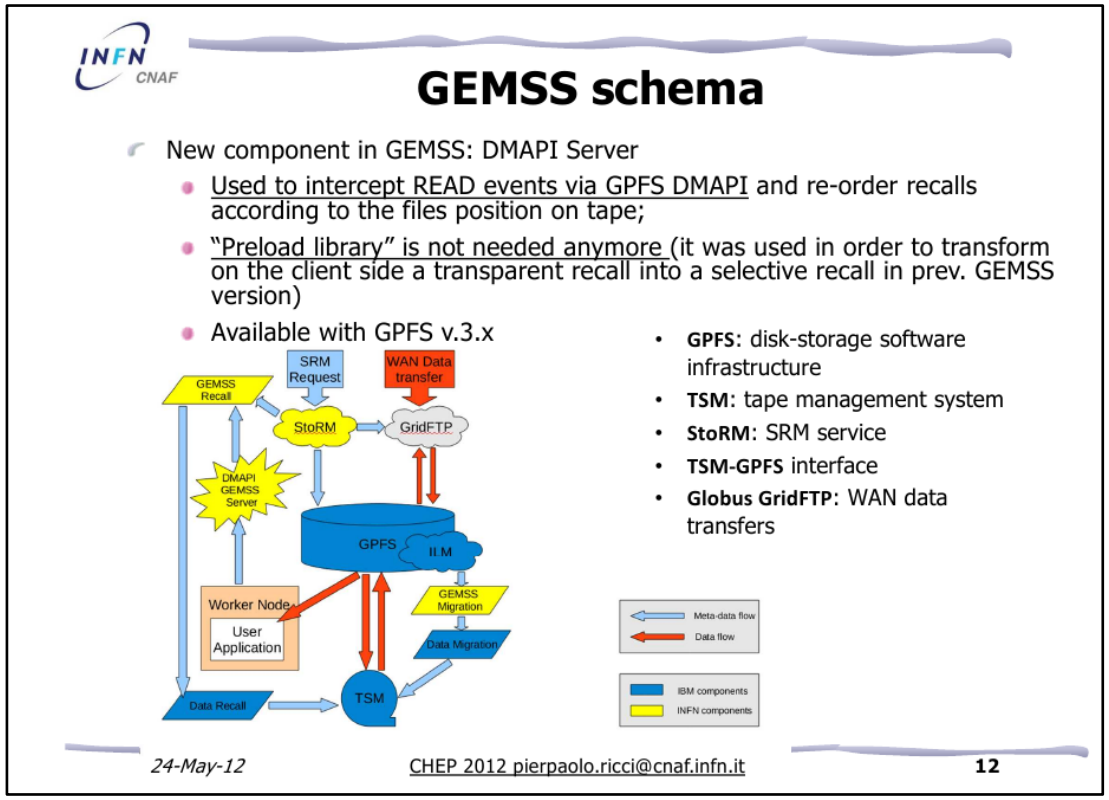




Storage Systems

Pier Paolo Ricci, INFN: The Grid Enabled Mass Storage System (GEMMS): the Storage and Data management system used at the INFN Tier1 at CNAF.

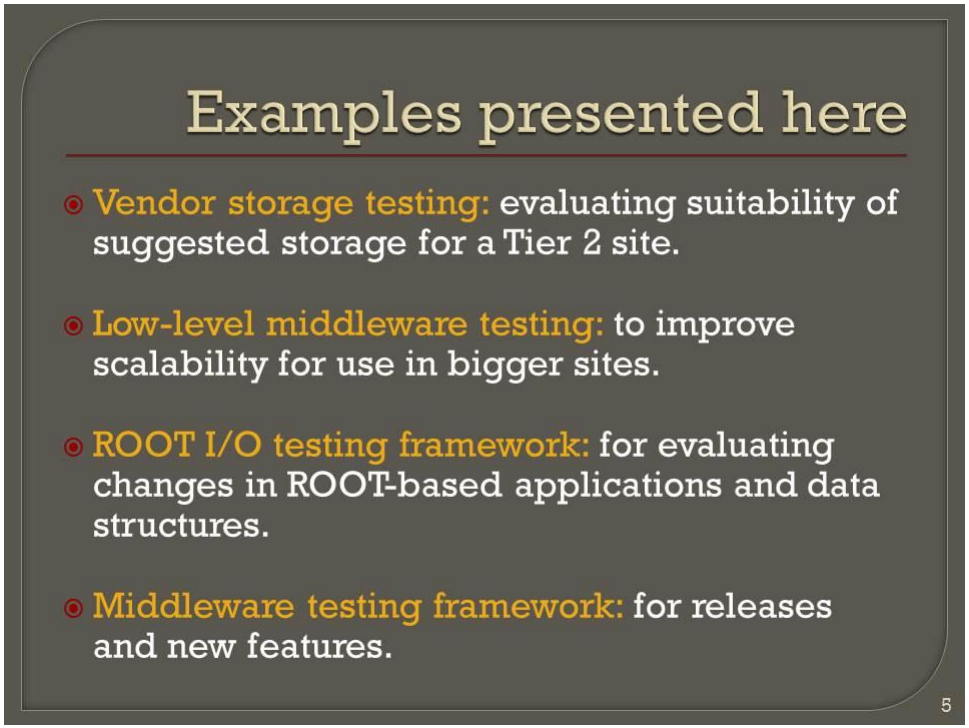
- Full HSM integration of GPFS, TSM and StoRM
- Read events triggered by Storm or by intercepted read requests from WNs.
- Clear improvement of stability compared to the formerly used CASTOR system.

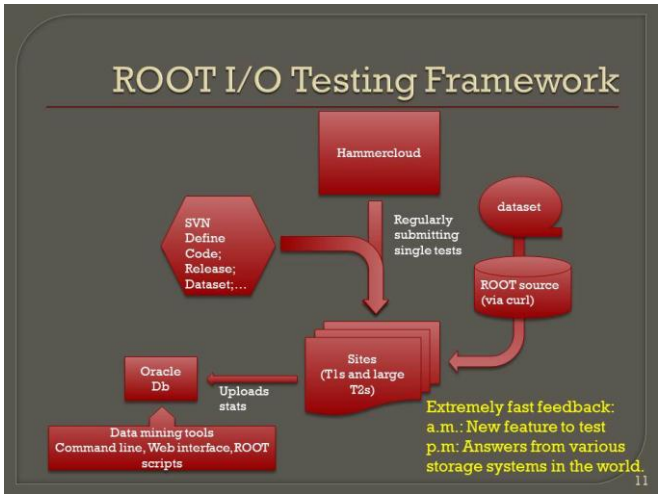


Storage Systems

Wahid Bhimji, University of Edinburgh: Analysing I/O bottlenecks in LHC data analysis on grid storage resources

Set of manual and automated tools to systematically test the I/O performance of different components: HW, middleware (e.g. DPM), application level (ROOT)



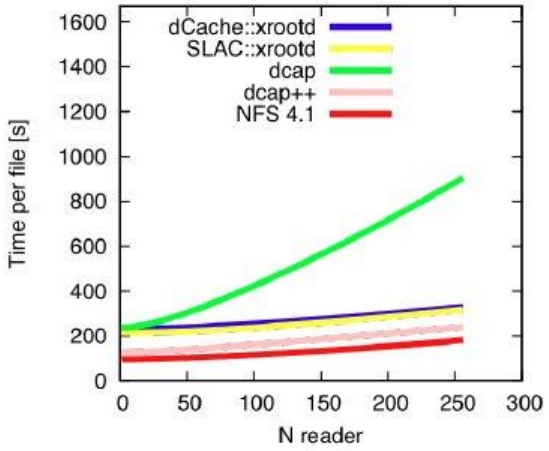


Hammercloud to pull code to a site and execute it.

Ives Kemp: Experience with HEP analysis on mounted filesystems

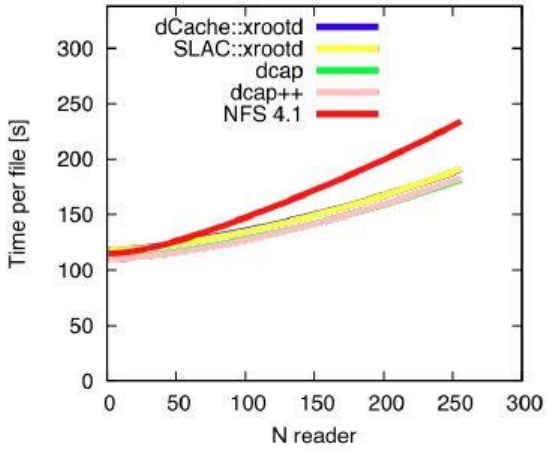
Results of protocol comparisons

- No clear winner: Depends on the read scenario
- NFS generally one of the fastest in this test setup



Optimized file, no TTreeCache, reading all branches

- VFS cache enhances analysis speed

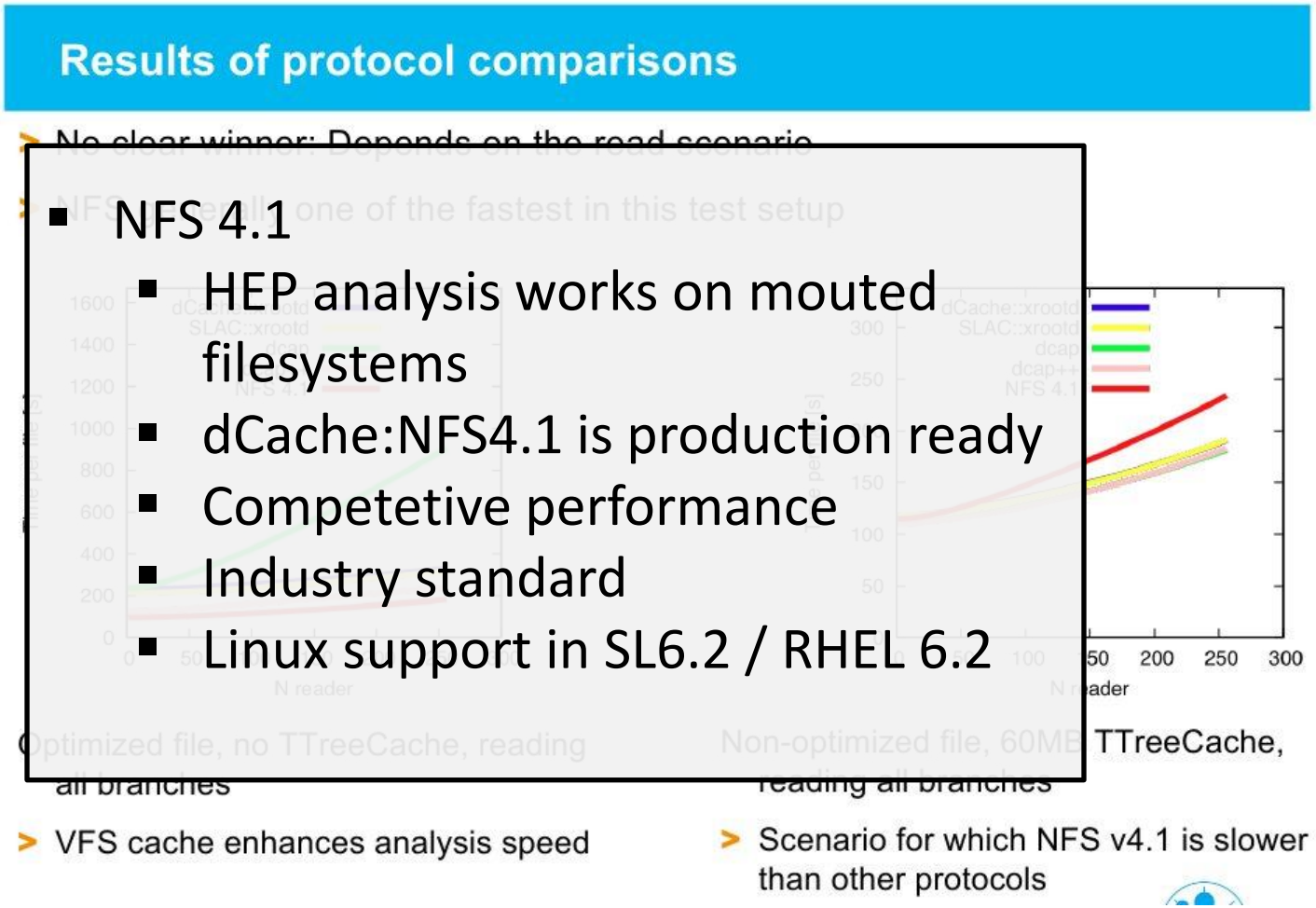


Non-optimized file, 60MB TTreeCache, reading all branches

- Scenario for which NFS v4.1 is slower than other protocols

Storage Systems

Ives Kemp: Experience with HEP analysis on mounted filesystems



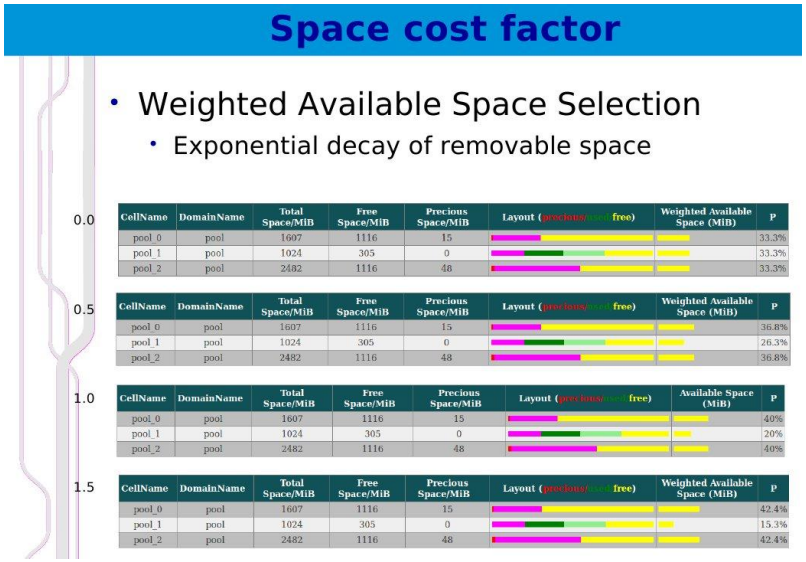
- NFS 4.1
  - HEP analysis works on mouted filesystems
  - dCache:NFS4.1 is production ready
  - Competetive performance
  - Industry standard
  - Linux support in SL6.2 / RHEL 6.2

Storage Systems

Mattias Wadenstein: A strategy for load balancing in distributed storage systems

-> Sometimes difficult to achieve even write pool selection in dCache.

-> New, dynamic pool selection of write pools in dCache to overcome issues like write clumping.



CellName	DomainName	Total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/used/free)	Weighted Available Space (MiB)	P
pool_0	pool	1607	1116	15			33.3%
pool_1	pool	1024	305	0			33.3%
pool_2	pool	2482	1116	48			33.3%

CellName	DomainName	Total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/used/free)	Weighted Available Space (MiB)	P
pool_0	pool	1607	1116	15			36.8%
pool_1	pool	1024	305	0			26.3%
pool_2	pool	2482	1116	48			36.8%

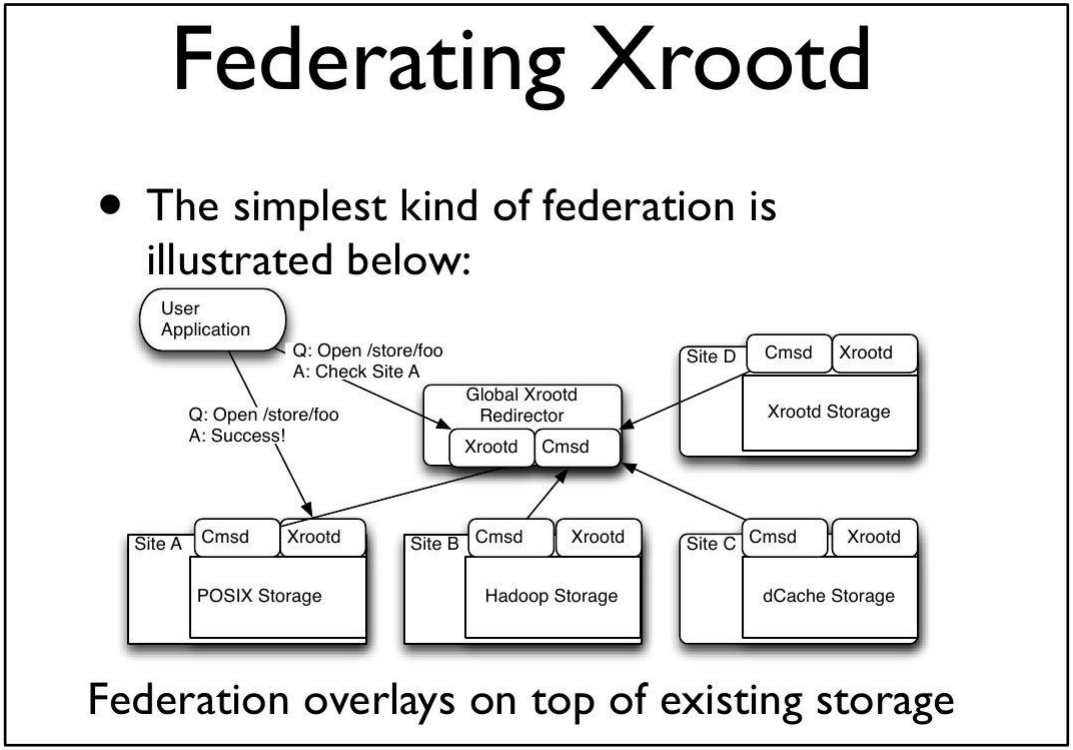
CellName	DomainName	Total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/used/free)	Available Space (MiB)	P
pool_0	pool	1607	1116	15			40%
pool_1	pool	1024	305	0			20%
pool_2	pool	2482	1116	48			40%

CellName	DomainName	Total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/used/free)	Weighted Available Space (MiB)	P
pool_0	pool	1607	1116	15			42.4%
pool_1	pool	1024	305	0			15.3%
pool_2	pool	2482	1116	48			42.4%

New method as default in dCache 2.2

Storage Systems

Brian Bockelman: Using Xrootd to federate regional storage



- Any storage system with posix of C interface can be integrated using a specific plugin, e.g. HDFS, Lustre
- Xrootd and dCache SEs can be integrated directly
- Global namespace: plugin to export global filename

Wide-area direct Xrootd access works but clients need to be aware of the sometimes higher network latency.

Storage Systems

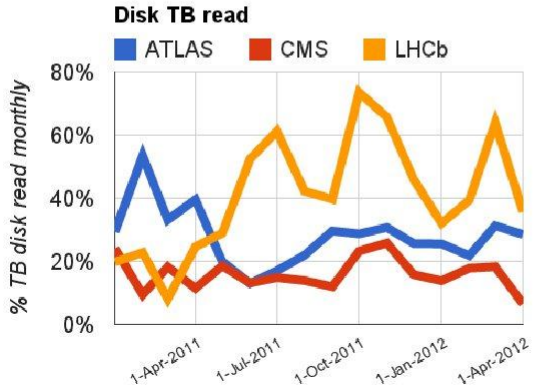
Gonzalo Merino: Dimensioning storage and computing clusters for efficient high throughput computing

- PIC T1 infrastructure
- Dimensioning of network backbone and storage systems
- Study of usage of data on disks by analyzing dCache billing logs.
  Only 20 to 25% of data on disk is touched at least once per month.

Data on disk usage

PIC
port d'informació
científica

Looking at the fraction of the data on disk (different files) which is read every month, one sees that it is often quite low.



Given that LHCb represents ~10% of the resources

=> every month, **75 - 80%** of all the bytes on disk at PIC are not read.

Network monitoring

- Modified computing and data distribution models of LHC experiments
- New network infrastructures -> LHCONE

$\Rightarrow$ Demand for better network performance monitoring

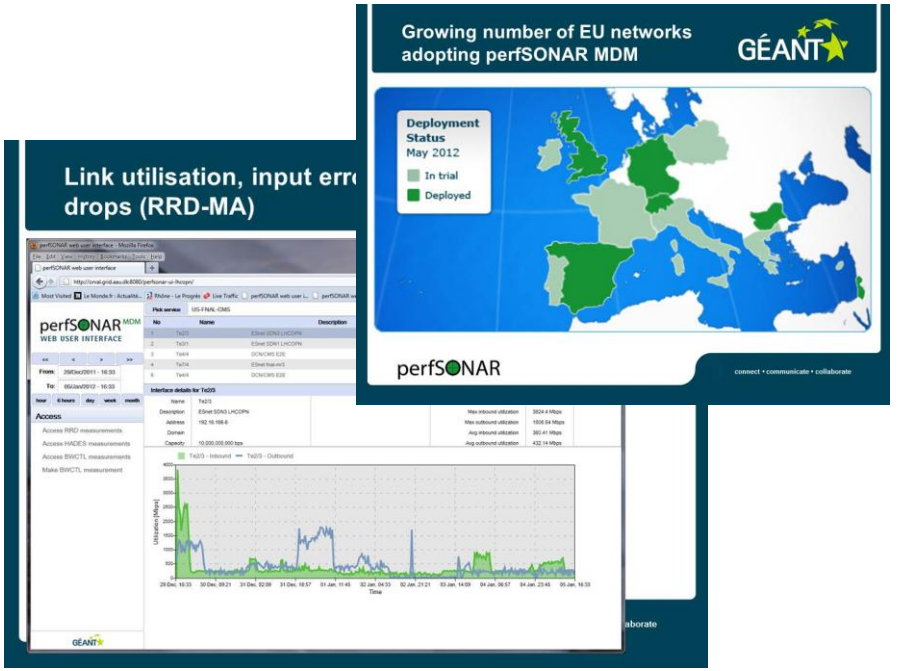
Two presentations of the perfSONAR tool in this track.

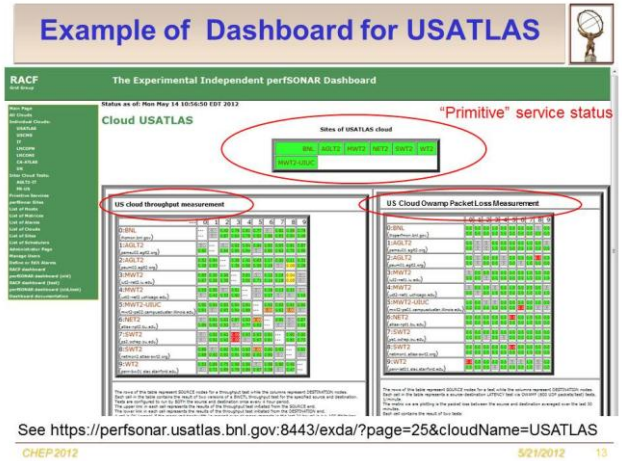
- Measuring standard network metrics
  - Bandwidth, latency, packet loss, …
- Two interoperable implementations

Networking

Domenico Vicinanza, GEANT:

- Differences of perfSONAR-MDM and perfSONAR-PS
- Deployment status in Europe
- Monitoring LHCOPN und LHCONE





Shawn McKee, Univ. of Michigan:

- perfSONAR-PS based monitoring of the US ATLAS network
- Future developments
  - Deployment at US CMS sites
  - Integration in dashboards

Networking

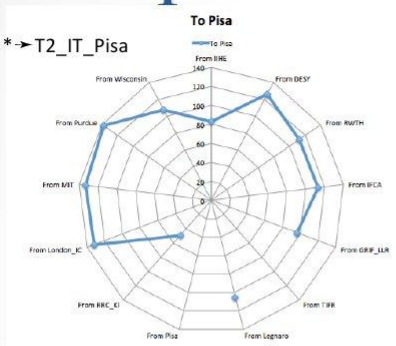
Rapolas Kaselis: CMS data transfer operations after the first year of LHC collisions

- PhEDEx data transfer monitoring
- Troubleshooting
- Test of the LHCONE

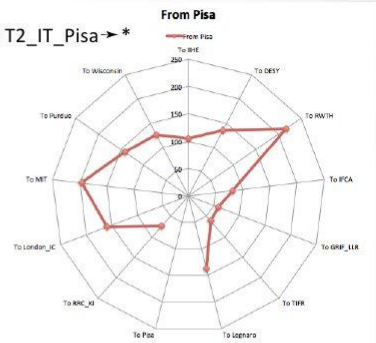
⇒ Smooth operations, but monitoring and debugging transfers takes lots of manpower.

A snapshot of test results

E.g. rates *→T2_IT_Pisa



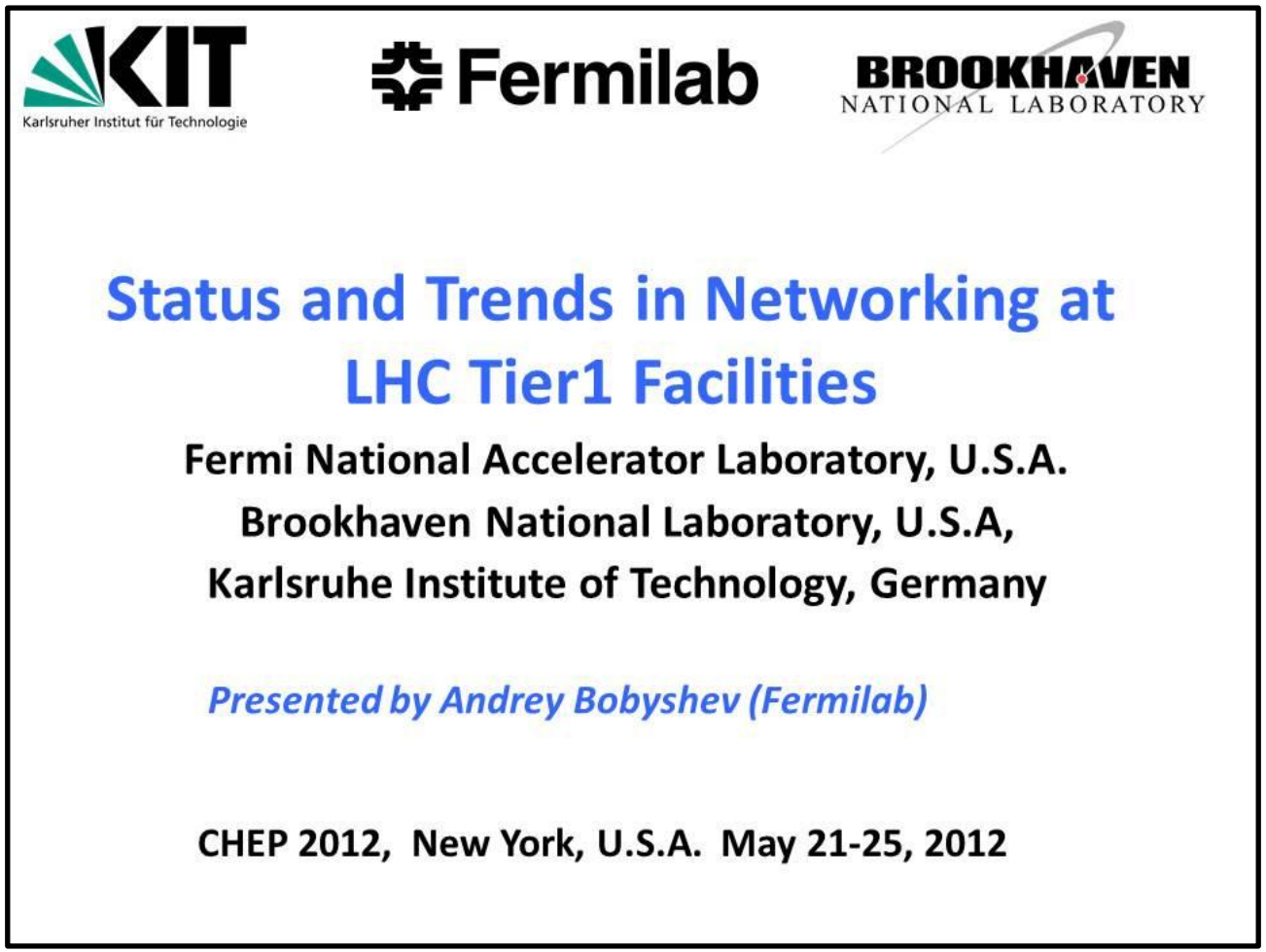
E.g. rates T2_IT_Pisa→*



Max rate in 1 hr [MiB/s] to / from	BE IIHE	DE DESY	DE RWTH	ES IFCA	FR GRIF LLR	IN TIFR	IT Legnaro	IT Pisa	RU RRC KI	UK London IC	US MIT	US Purdue	US Wisconsin
BE IIHE		85	105	49	50	60	96	83	38	79	75	80	76
DE DESY	105		518	85	62	61	148	126	65	182	260	109	256
DE RWTH	97	144		86	108	88	164	112	74	229	255	103	183
ES IFCA	77	85	97		61	76	87	113	66	102	122	72	136
FR GRIF_LLR	107	132	449	76		77	145	96	57	320	279	133	368
IN TIFR													
IT Legnaro	87	109	196	64	57	47		105	62	79	171	114	180
IT Pisa	105	135	217	81	59	61	137		74	160	197	141	126
RU RRC_KI	42	68	119	28	n/a	42	51	49		117	99	77	97
UK London_IC	64	110	414	93	88	87	139	132	63		305	116	287
US MIT	108	89	422	84	68	65	133	133	59	39		72	428
US Purdue	101	55	314	55	48	75	75	138	48	427	320		408
US Wisconsin	102	105	365	81	43	86	139	108	62	100	330	85	

Networking

Andrey Bobyshey, FNAL: "Tier-1 LAN party"



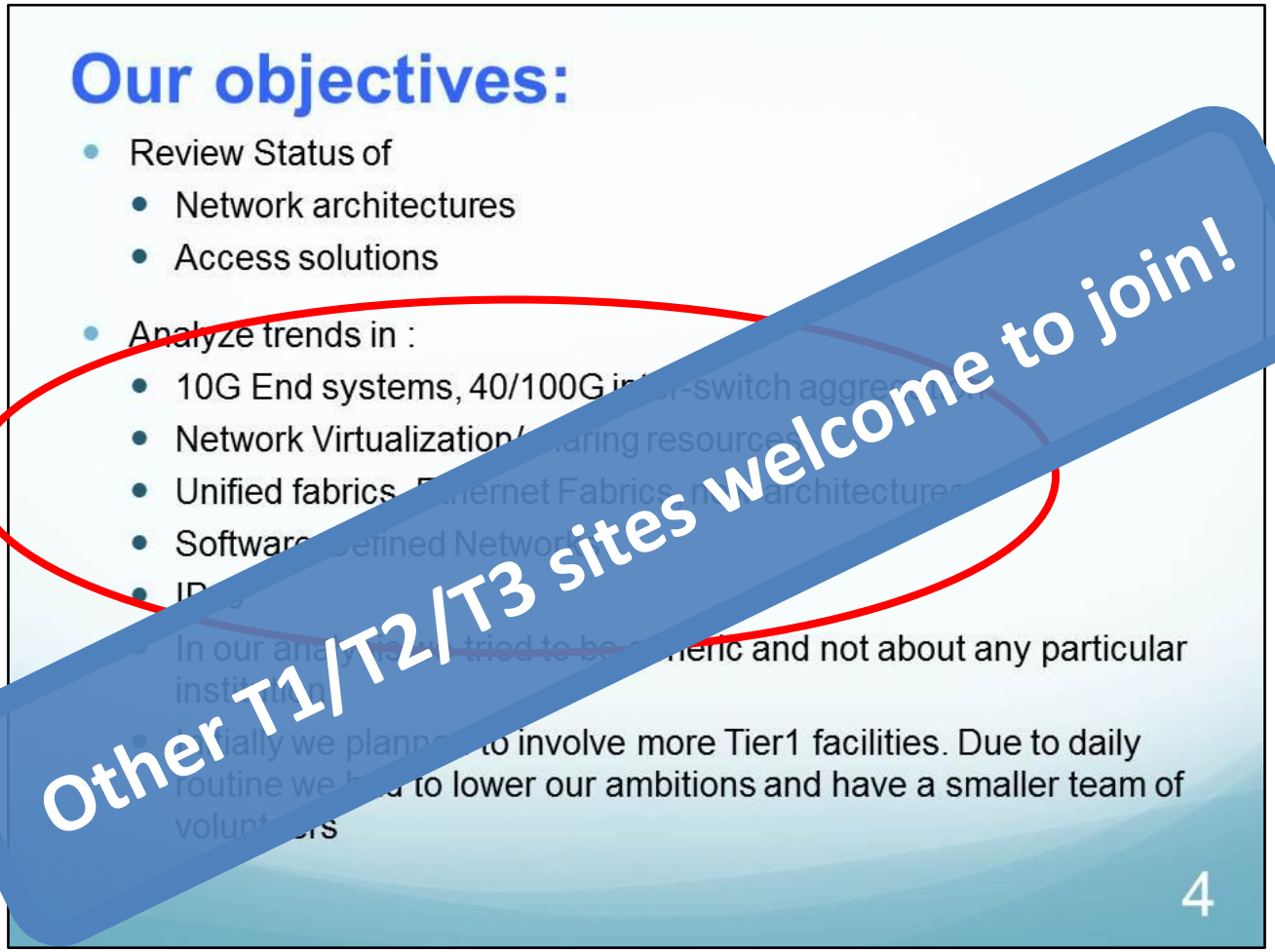
Networking

Our objectives:

- Review Status of
  - Network architectures
  - Access solutions
- Analyze trends in :
  - 10G End systems, 40/100G inter-switch aggregation
  - Network Virtualization/sharing resources
  - Unified fabrics, Ethernet Fabrics, new architectures
  - Software-Defined Networks
  - IPv6
  - In our analysis we tried to be generic and not about any particular institution
  - Initially we planned to involve more Tier1 facilities. Due to daily routine we had to lower our ambitions and have a smaller team of volunteers

4

Networking


Our objectives:
Review Status of
Network architectures
Access solutions
Analyze trends in :
10G End systems, 40/100G
Network Virtualization
Unified fabrics
Software
In our
and not about any particular
to involve more Tier1 facilities. Due to daily
to lower our ambitions and have a smaller team of
4
Other T1/T2/T3 sites welcome to join!

Networking

Eduardo Martelli: “From IPv4 to eternity”: the HEPiX IPv6 working group

Motivation:

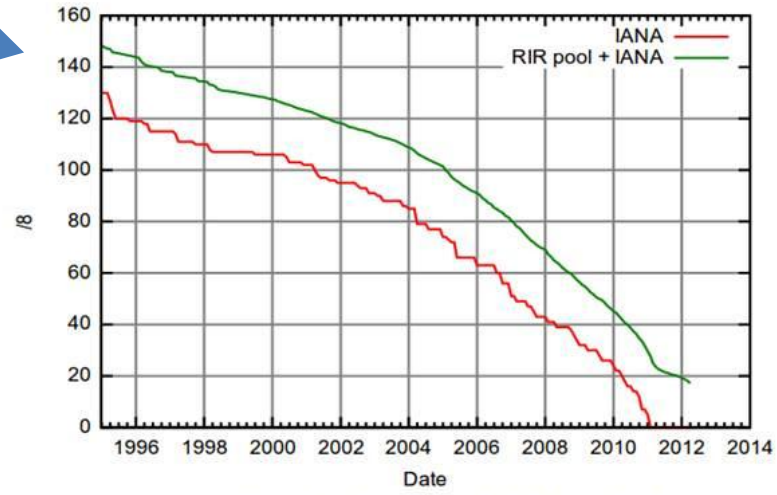
1)

2) VM and (commercial) cloud services will require IPv6

HEPiX

IPv4 address space depletion

Remaining IPv4 Free Addresses (/8 blocks):



Source: http://en.wikipedia.org/wiki/File:Ipv4-exhaust.svg

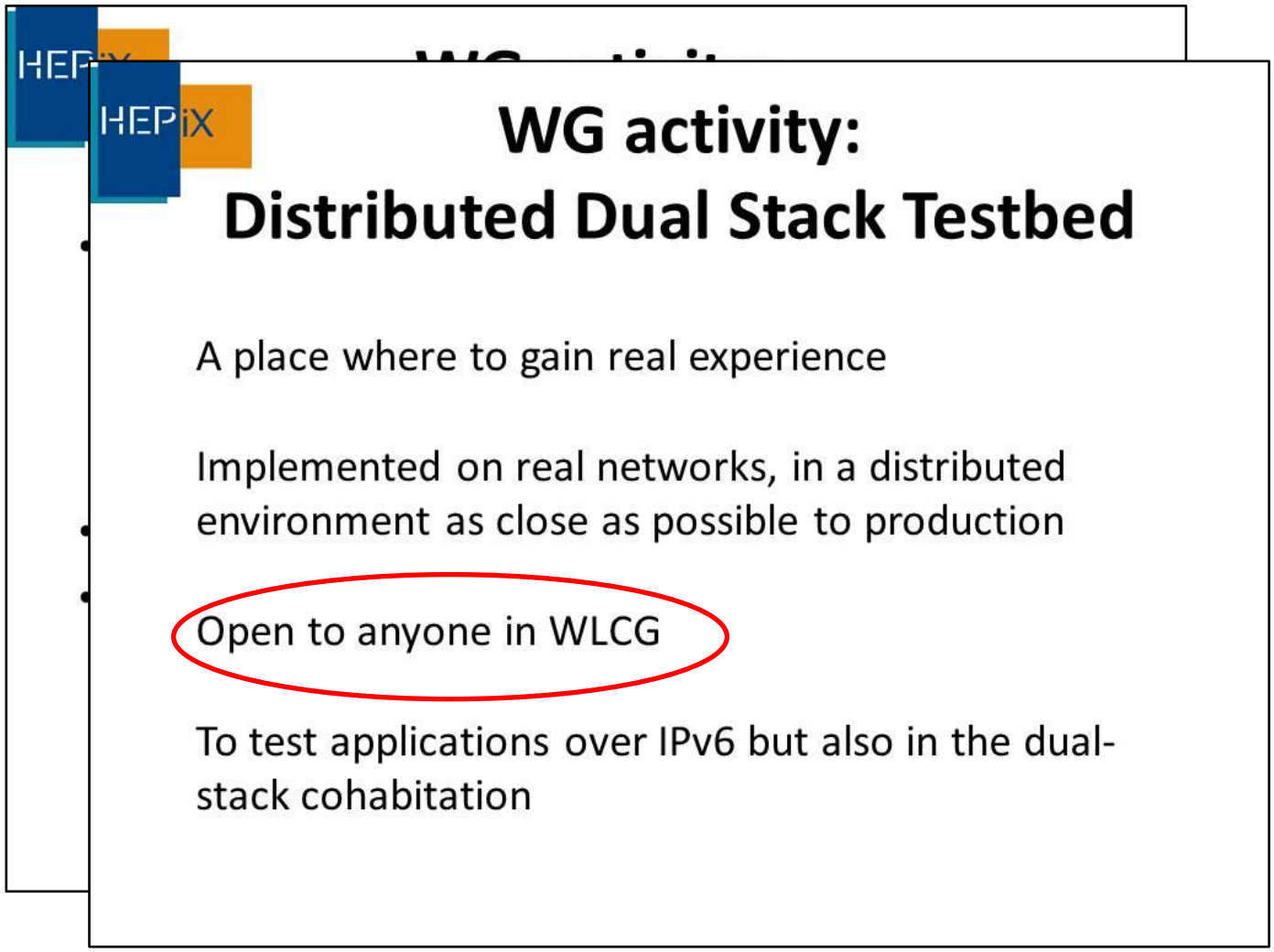
5

WG activity: Software & Tools IPv6 Survey

- An "Asset" survey is now underway
  - A spreadsheet to be completed by all sites and the LHC experiments
  - Includes **all** applications, middleware and tools
  - Tickets to be entered for all problems found
- If IPv6-readiness is known, can be recorded
- Otherwise we will need to investigate further
  - Ask developer and/or supplier
  - Scan source code or look for network calls while running
  - Test the running application under dual stack conditions

15

Networking


HEPiX
WG activity:
Distributed Dual Stack Testbed
A place where to gain real experience
Implemented on real networks, in a distributed environment as close as possible to production
Open to anyone in WLCG
To test applications over IPv6 but also in the dual-stack cohabitation

Networking

WG activity:

First results:

- Gridftp file transfers working
- Many tools and applications *not ready*

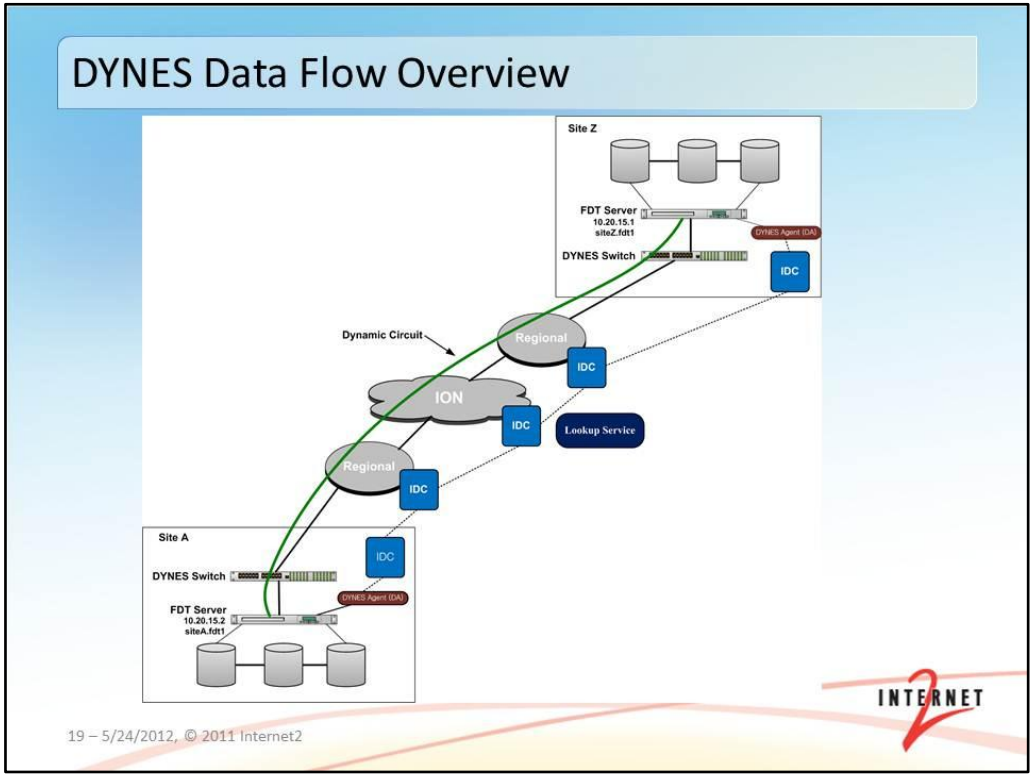
Future plans include:

- Expand testbed, consider merging EGI and HEPIX testbeds, include all WLCG services
- “IPv6 days“: turn on dual-stack on production infrastructure

Networking

Jason Zurawski: The DYNES Instrument: A Description and Overview

- A “Cyber-instrument“ extending Internet2‘s ION (on-demand network) service to regional and campus networks.
- Support large, long-distance science data flows
- ~40 US universities and 12 Internet2 connectors



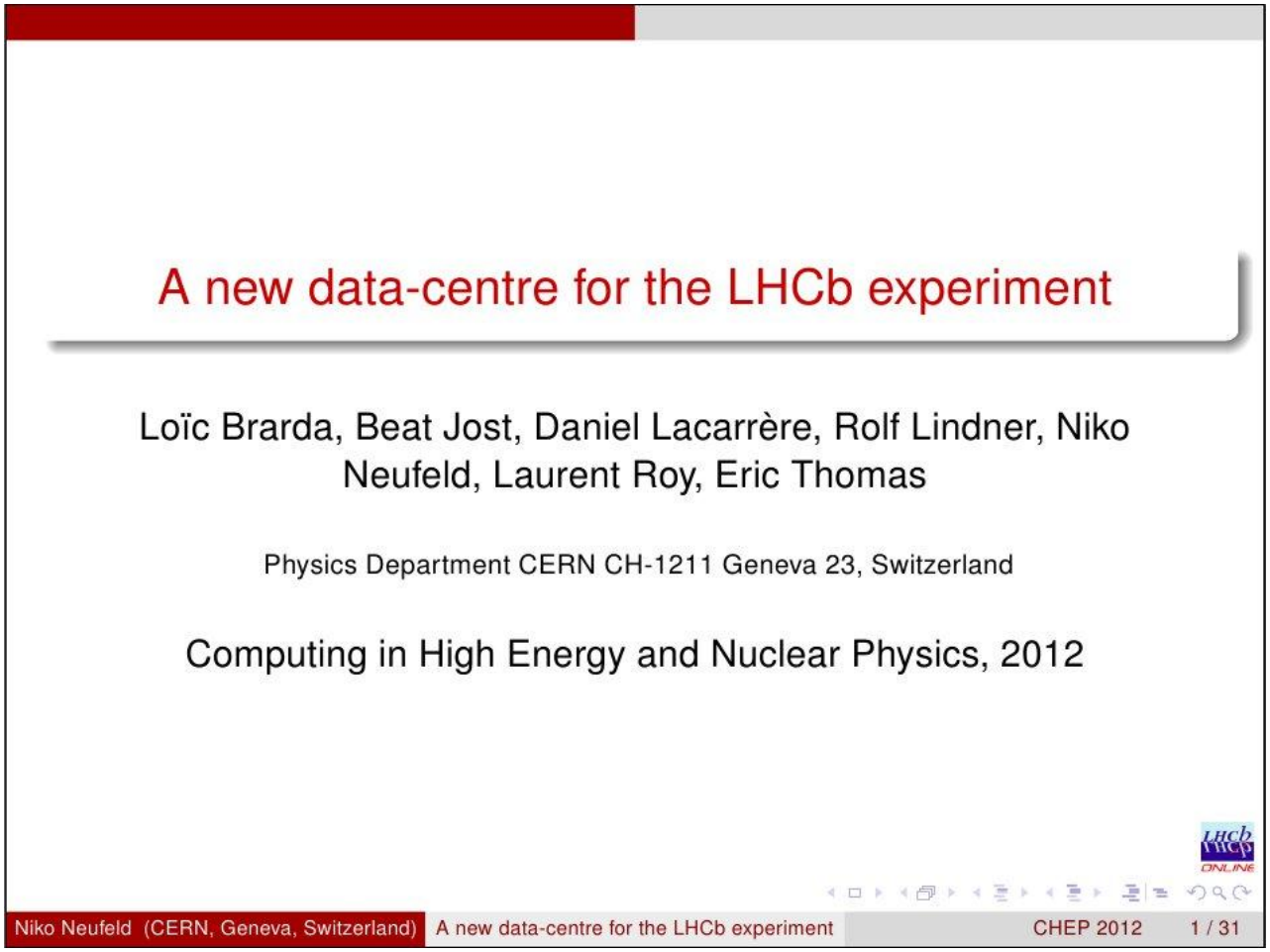
Out-of-the-box solution. Each site needs:

- Inter-domain (IDC) controller
- Dynes switch
- FDT server
- Storage

Resource Centres

New event filter farm and data centre for LHCb for the time after LS2

- Upgraded DAQ capable of reading out the entire detector @ 40 MHz



A new data-centre for the LHCb experiment

Loïc Brarda, Beat Jost, Daniel Lacarrère, Rolf Lindner, Niko Neufeld, Laurent Roy, Eric Thomas

Physics Department CERN CH-1211 Geneva 23, Switzerland

Computing in High Energy and Nuclear Physics, 2012

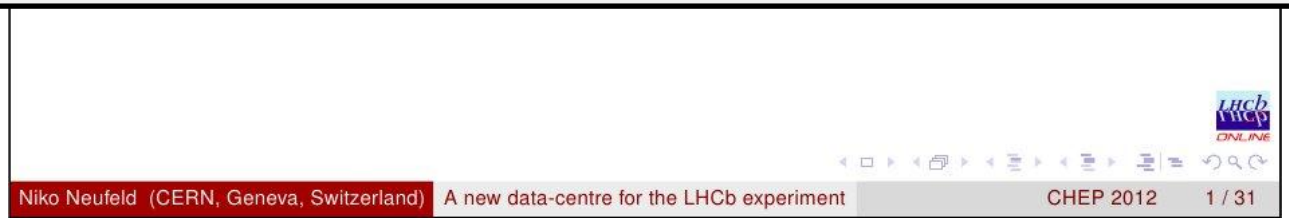
Niko Neufeld (CERN, Geneva, Switzerland) A new data-centre for the LHCb experiment CHEP 2012 1 / 31

Resource Centres

New event filter farm and data centre for LHCb for the time after LS2

- Upgraded DAQ capable of reading out the entire detector @ 40 MHz

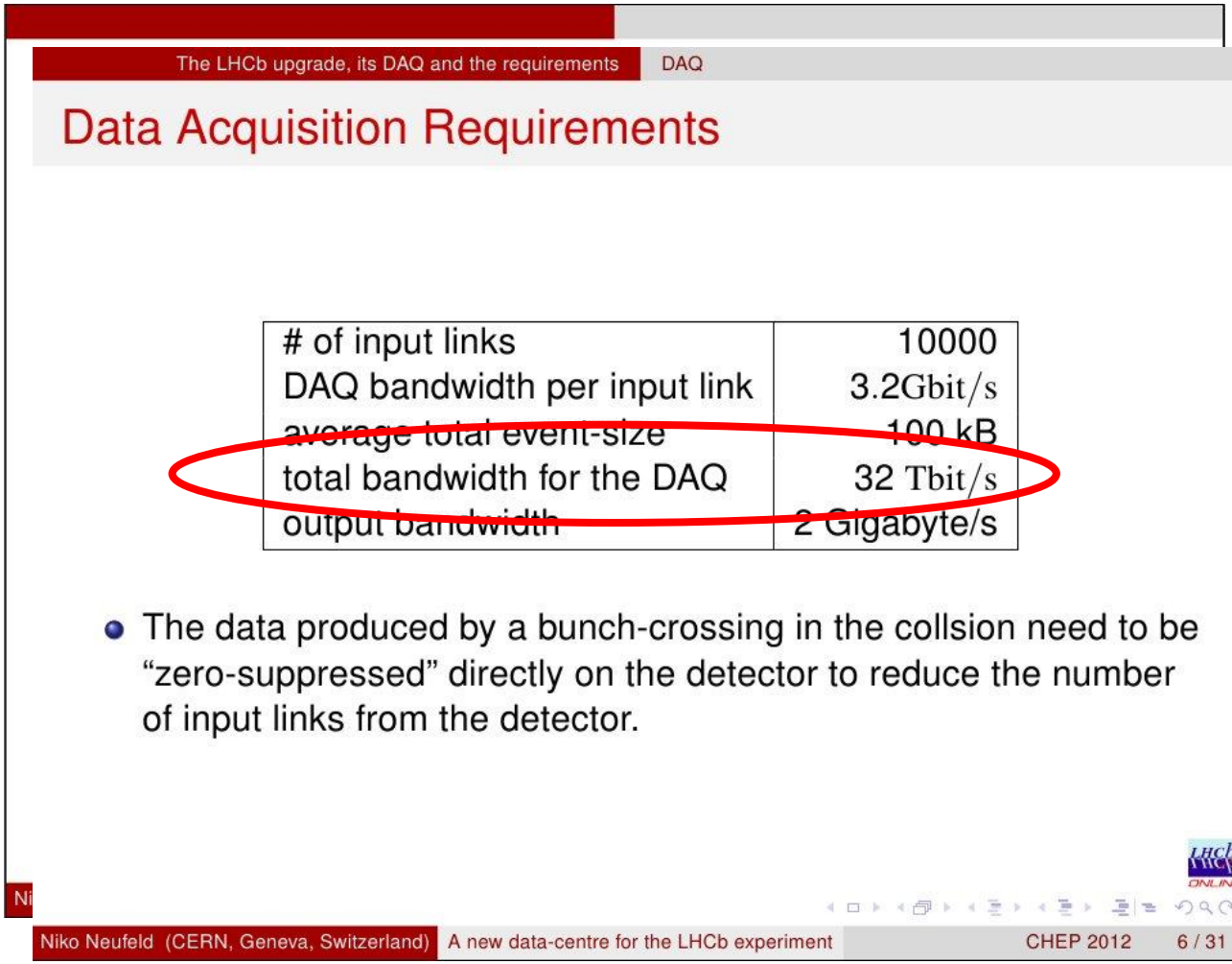
- Limited lifetime of the LHCb experiment suggests to consider remote hosting in existing building.
- New on-site data centre beyond budget
- Study of different options
  - Remote hosting off the CERN site
  - Remote hosting on the CERN site in existing centre



Resource Centres

New event filter farm and data centre for LHCb for the time after LS2

- Upgraded DAQ capable of reading out the entire detector @ 40 MHz

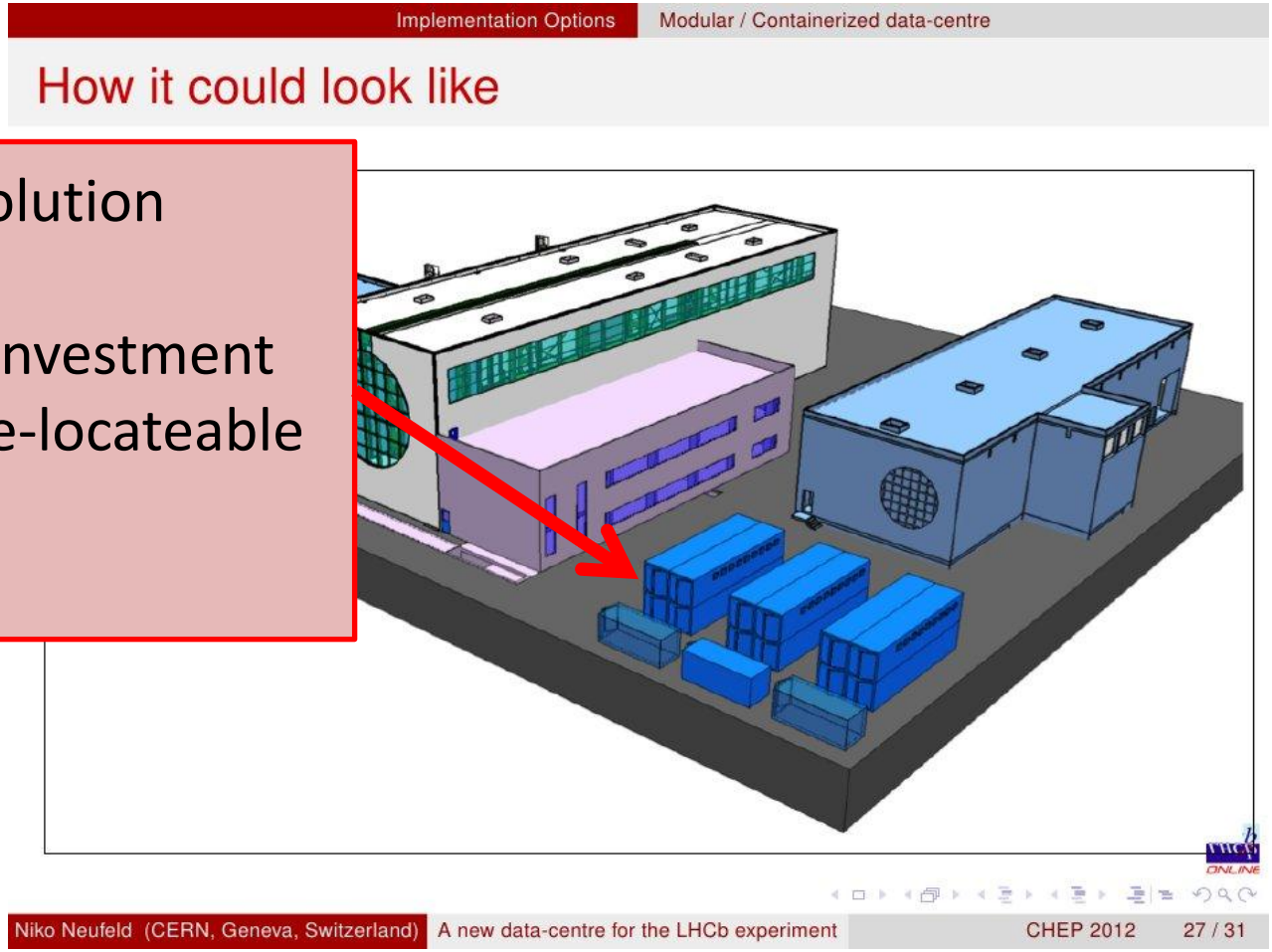


Resource Centres

New event filter farm and data centre for LHCb for the time after LS2

- Upgraded DAQ capable of reading out the entire detector @ 40 MHz

- Containerized solution advantages:
  - Low capital investment
  - Re-usable, re-locateable
  - Low PUE
  - On-site



Resource Centres

Data Centre Selection

CERN IT Department

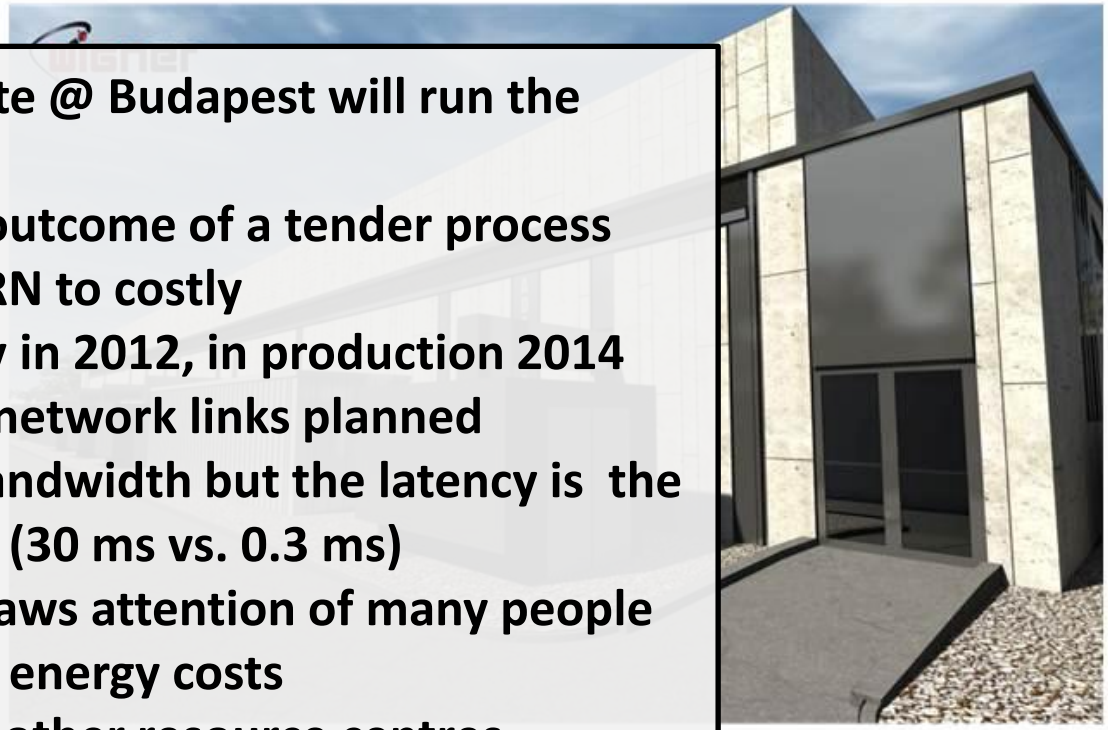
- Wigner Institute in Budapest, Hungary

Resource Centres

Data Centre Selection

CERN IT Department

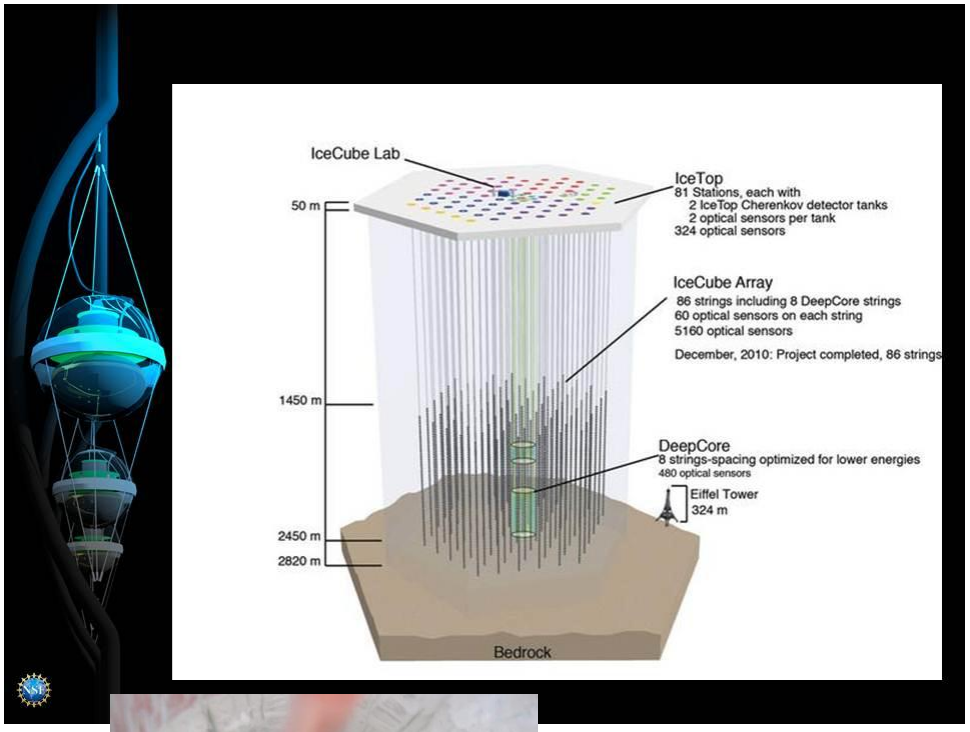
- Wigner Institute @ Budapest will run the remote-Tier0
- (unexpected) outcome of a tender process
- Building at CERN to costly
- Testing already in 2012, in production 2014
- Two 100 Gb/s network links planned
  - Not the bandwidth but the latency is the challenge. (30 ms vs. 0.3 ms)
- This project draws attention of many people
  - Increasing energy costs
  - Model for other resource centres, including HPC?



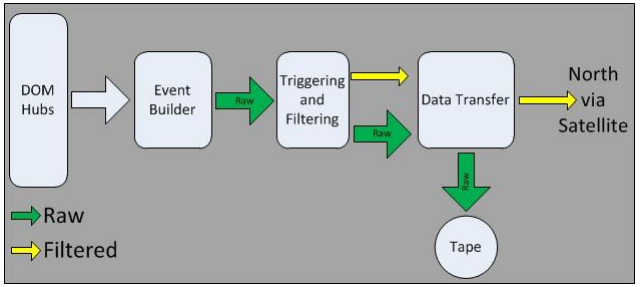
Wigner Institute in Budapest, Hungary

Non-LHC projects

Steve Barnet, University of Wisconsin: The Ice Cube Computing Model



- 1 TB of RAW data per day
- Satellite capacity: 100 GB per day



Tier-0: Madison, Wisconsin

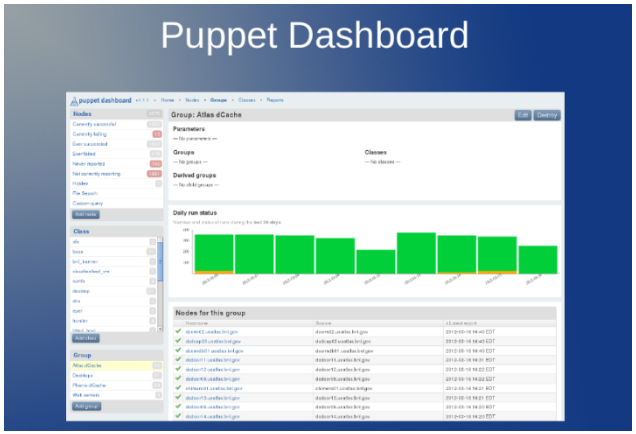
- RAW data
- Event selection

Tier-1: Zeuthen, Germany

- Copy of reconstructed data
- MC production and store

Other Topics

Fabric Management: The winner is …. Puppet (+ GIT + ….)



Jason Smith, BNL: Centralized Fabric Management Using Puppet, Git, and GLPI

Jason Smith, BNL:

- Powerful language
- Large user base and developer community
- Dashboard

Tim Bell, CERN:

- Maintenance costs for own tool too high
- CERN compute centre size not longer leading edge
- Meanwhile many open source solutions available
- Puppet: Large user an support community
- Better chances on the job market!

Tim Bell: Review of CERN Computer Centre Infrastructure

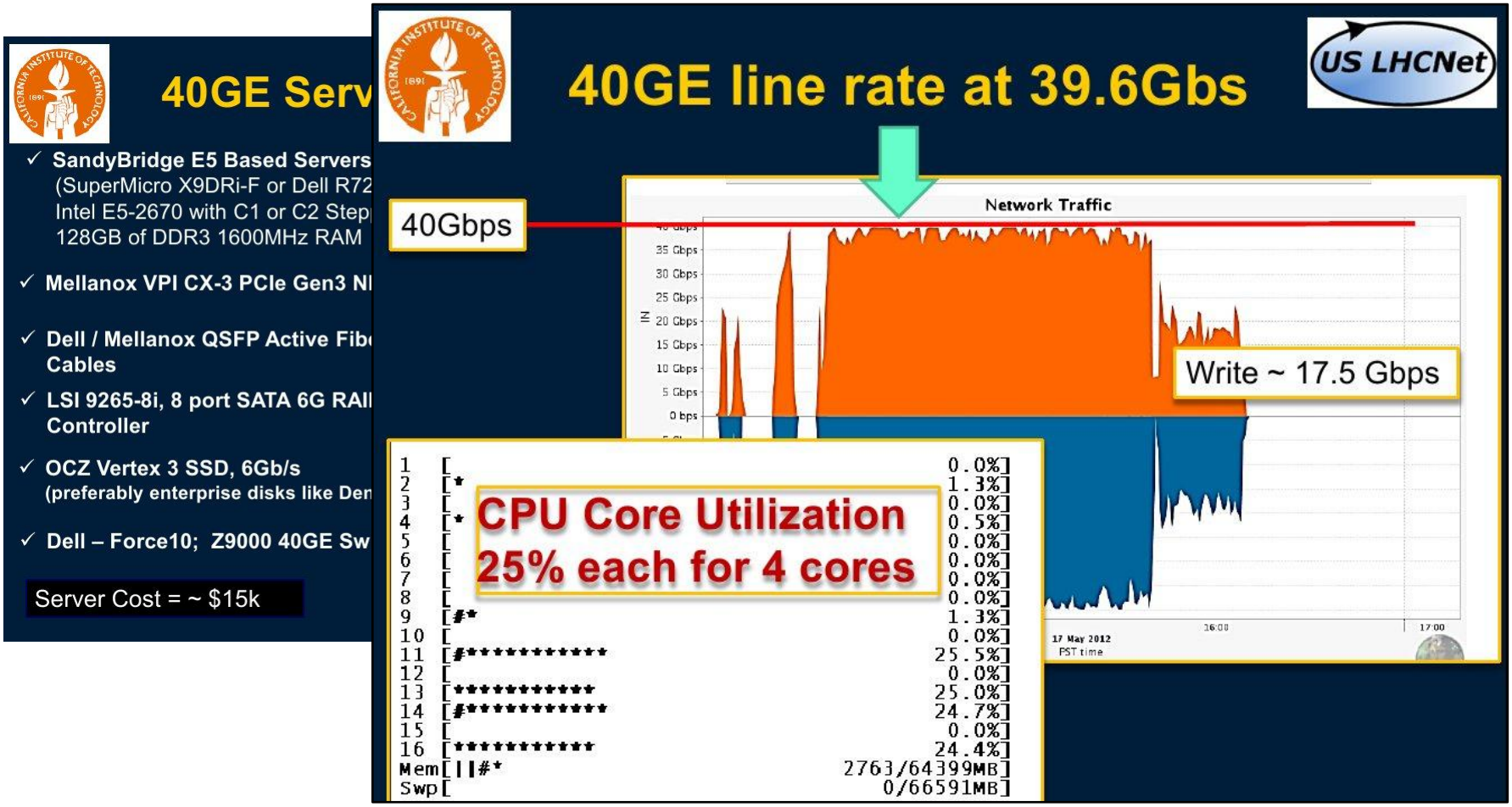
Other topics

Azher Mughal, Caltech: Evaluation of 40 Gigabit Ethernet technology for data servers

- Hardware and software tuning is necessary:
  - Latest firmwares
  - Enable PCIe Gen3
  - Bind NIC driver to CPU where PCIe lane is connected
  - Move raid driver IRQ to second CPU
  - Bind file transfer application (FDT) to second CPU.
  - Change kernel parameters
  - …

Other topics

Azher Mughal, Caltech: Evaluation of 40 Gigabit Ethernet technology for data servers



Summary of the Summary

- CHEP 2010: many presentations about ‘practical experiences‘ with real data
- CHEP 2012: again more visionary presentations
  - Experiences from data taking still used to optimize systems, storage, networks, …
  - But also a lot of planning for the future
    - High luminosity running, improved detectors => more data
    - Scaling capacity, e.g.
      - CERN remote T0
      - LHCb data centre
      - Exploitation of next generation networks begins (40, 100 Gbit/s, bandwidth on demand).
    - Going mainstream, less custom tools, save (wo)manpower
  - Storage systems are still a hot topic!

Thanks

- to speakers and poster presenters for many interesting presentations
- to the audience of the parallel sessions for their interest and the vital discussions
- to the conference organizers for a great CHEP 2012!