Summary on the session :Hardware and Computing Fabrics

Takashi Sasaki,KEK

and

Jiri Chudoba, FZU

Track summary

- 41 papers were submitted in total
  - 17 papers are accepted as “oral”
  - The rest went to the poster
- Session 1 : Monday 16:00-
  - Very popular than LOC expectation
    - Audiences were almost twice of the capacity of the room(60), over 100 people
- Session 2 : Tuesday 14:00-
  - very popular
    - 80 audiences
- Session 3 : Tuesday 16:30-
  - Popular
    - 60 audiences

Paper category

- Benchmarking
  - 1
- operation experience
  - Data/computing center
    - 6
  - Experiments
    - 4
      - 2DAQ
- Data/computing center infrastructure
  - 1
- New technology
  - 5
  - Virtualization, SSD, new processor, file system

Authors

- North America
  - 5
- Europe
  - 12
- Asia
  - None

Session1:Monday, 22 March 2009 16:00-

- [19] A comparison of HEP code with SPEC benchmark on multicore worker nodes
  by Michele MICHELOTTO (INFN + Hepix)
- [435] Experience with low-power x86 processors (ATOM) for HEP usage
  by Mr. Sverre JARP (CERN)
- [387] Air Conditioning and Computer Centre Power Efficiency: the Reality
  by Tony CASS (CERN)
- [397] A High Performance Hierarchical Storage Management System For the Canadian Tier-1 Centre at TRIUMF
  by Mr. Simon LIU (TRIUMF)
- [431] Fair-share scheduling algorithm for a tertiary storage system
  by Mr. Pavel JAKL (Nuclear Physics Inst., Academy of Sciences, Praha)
- [216] Lustre File System Evaluation at FNAL
  by Stephen WOLBERS (FNAL)

HEPiX Benchmarking Group
Michele Michelotto at pd.infn.it

A comparison of HEP code with SPEC benchmark on multicore worker nodes

Why INT ?

- Since SPEC CPU 92 the HEP world decide to use INT as reference instead of FP (Floating Point)
- HEP programs of course make use of FP instructions but with minimal inpact on benchmarks
- I’ve never seen a clear proof of it

Results

- Very good correlation (>90%) for all experiments
- Both SI2006 and SFP2006 (multiple parallel) could be good substitute for SI2000
- Interesting talk from Andreas Hirstius from CERN-IT Openlab at HEPiX Spring 08 on “perfmon”

The choice

- SPECint2006 (12 applications)
  - Well established, published values available
  - HEP applications are mostly integer calculations
  - Correlations with experiment applications shown to be fine
- SPECfp2006 (17 applications)
  - Well established, published values available
  - Correlations with experiment applications shown to be fine
- SPECall_cpp2006 (7 applications)
  - Exactly as easy to run as is SPECint2006 or SPECfp2006
  - No published values (not necessarily a drawback)
  - Takes about 6 h (SPECint2006 or SPECfp2006 are about 24 h)
  - Best modeling of FP contribution to HEP applications
  - Important memory footprint
- Proposal to WLCG to adopt SPECall_cpp 2006, in parallel and to call it HEP SPEC06

Hep-Spec06

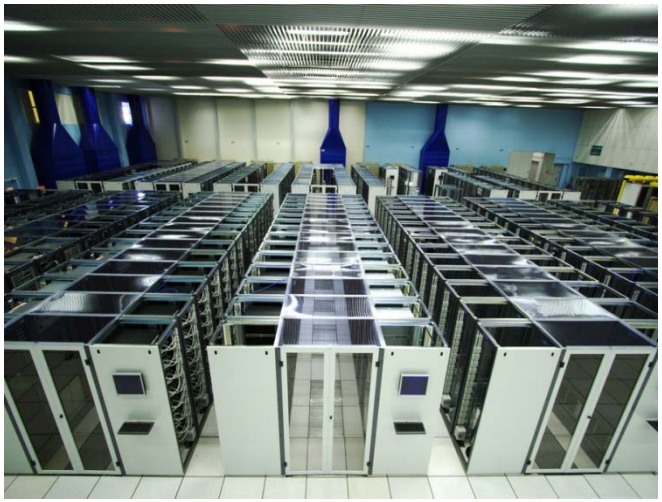
Machine	SPEC2000	SPEC2006 int 32	SPEC2006 fp 32	SPEC2006 CPP 32
lxbench01	1501	11.06	9.5	10.24
lxbench02	1495	10.09	7.7	9.63
lxbench03	4133	28.76	25.23	28.03
lxbench04	5675	36.77	27.85	35.28
lxbench05	6181	39.39	29.72	38.21
lxbench06	4569	31.44	27.82	31.67
lxbench07	9462	60.89	43.47	57.52
lxbench08	10556	64.78	46.48	60.76

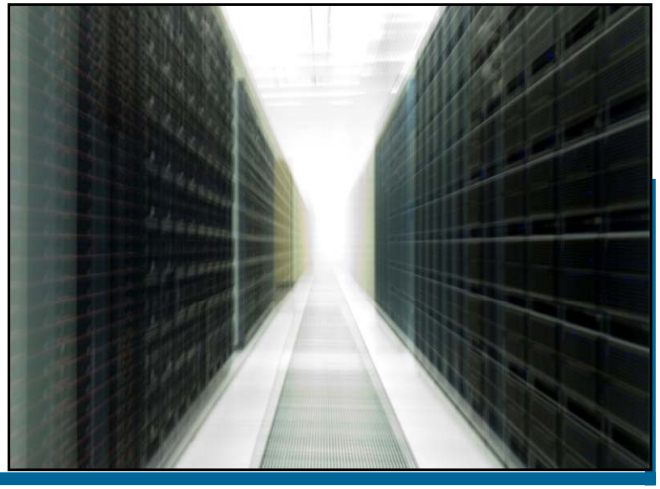
Is the Atom (N330) processor ready for High Energy Physics?

Gyorgy Balazs
Sverre Jarp
Andrzej Nowak

CERN openlab

CHEP09 – 23.3.2009





ATOM processor specifications

- ATOM N330 is the biggest member in the current family:

# cores	**2**
# hardware threads /core	**2**
Frequency	**1.6 Ghz**
Max (certified) memory config.	**2 GB**
L1 cache	**32KB+24KB**
L2 cache (per core)	**512KB**
Front-side bus frequency	**800 MHz**
64-bit enabled	**YES**
SIMD Extensions	**Incl. SSSE3**
In-order execution	**YES**

Price estimates (1)

- Taken “anonymously” from the Web (Oct. 08):

Motherboard+CPU	**110 CHF**
2GB DDR2 memory	**30 CHF**
Power supply, drives	**110 CHF**
Total	**250 CHF**

Atom

2x E5472 CPU	**3500 CHF**
1x4GB DDR2 memory	**300 CHF**
Other (board, PSU, drives)	**1400 CHF**
Total	**5200 CHF**

Harpertown

Of course, we can discuss “endlessly” whether the comparison is fair or not, so it is just meant as an indication!

Price estimates (2)

- **Memory adjustment (include 2GB/process)**
  - Taken "anonymously" from the Web (Oct. 08):

Motherboard+CPU	**110 CHF**
2*4GB DDR2 memory	**150 CHF**
Power supply, drives	**110 CHF**
Total	**370 CHF**

Atom

2x E5472 CPU	**3500 CHF**
4x4GB DDR2 memory	**1200 CHF**
Other (board, PSU, drives)	**1400 CHF**
Total	**6100 CHF**

Harpertown

Benchmark results

- **“test40” from Geant4 (in summary):**
  - Atom baseline: 1 process at 100% throughput at 47W
  - Atom peak: 4 processes at 302% throughput at 50W
  - Harpertown: 8 processes at 3891% throughput at 265W

	SETUP	USER TIME		ACTIVE	ADVANTAGE		
	#proc	Runtime AVG (us)	% of 1 proc	POWER (W)	Workload	Throughput	Throughput per Watt
ATOM 330 @ 1.6 GHz Fedora 9, GCC 4.3, 2GB RAM	1	156	100%	47 W	100%	100%	100%
	2	157	100%	48 W	200%	199%	195%
	3	192	123%	49 W	300%	244%	234%
	4	207	132%	50 W	400%	302%	287%
Harpertown @ 3.0 GHz SLC 4.7, GCC 4.3, 4GB RAM	1	32	21%	186 W	100%	488%	123%
	2	32	21%	202 W	200%	973%	227%
	4	32	21%	232 W	400%	1944%	394%
	8	32	21%	265 W	800%	3891%	690%

Benchmark results (cont'd)

- **"test40" from Geant4 (memory adjusted):**
  - Atom baseline: 1 process at 100% throughput at 53W
  - Atom peak: 4 processes at 302% throughput at 56W
  - Harpertown: 8 processes at 3891% throughput at 290W

	SETUP	USER TIME		ACTIVE	ADVANTAGE		
	#proc	Runtime AVG (us)	% of 1 proc	POWER (W)	Workload	Throughput	Throughput per Watt
Atom 330 @ 1.6 GHz Fedora 9, GCC 4.3, 2x4GB RAM	1	156	100%	53 W	100%	100%	100%
	2	157	100%	54 W	200%	199%	196%
	3	192	123%	55 W	300%	244%	235%
	4	207	132%	56 W	400%	302%	286%
Harpertown @ 3.0 GHz SLC 4.7, GCC 4.3, 4x4GB RAM	1	32	21%	210 W	100%	488%	123%
	2	32	21%	225 W	200%	973%	229%
	4	32	21%	255 W	400%	1944%	404%
	8	32	21%	290 W	800%	3891%	711%

Benchmark results (cont'd)

- **"test40" from Geant4 (in summary):**
  - Atom baseline: 1 process at 100% throughput at 53W
  - Atom peak: 4 processes at 302% throughput at 56W
  - Harpertown: 8 processes at 3891% throughput at 290W
- **In other words (Harpertown/Atom ratios):**
  - Cost ratio was: 16.5 (with adjusted memory)
  - 12.9x throughput advantage
  - 5.2x power increase
- **Atom N330 could be interesting in terms of performance/franc**
  - Currently uninteresting when looking at performance/watt

Air Conditioning and Computer Centre Power Efficiency The Reality

Christophe Martel

Tony Cass

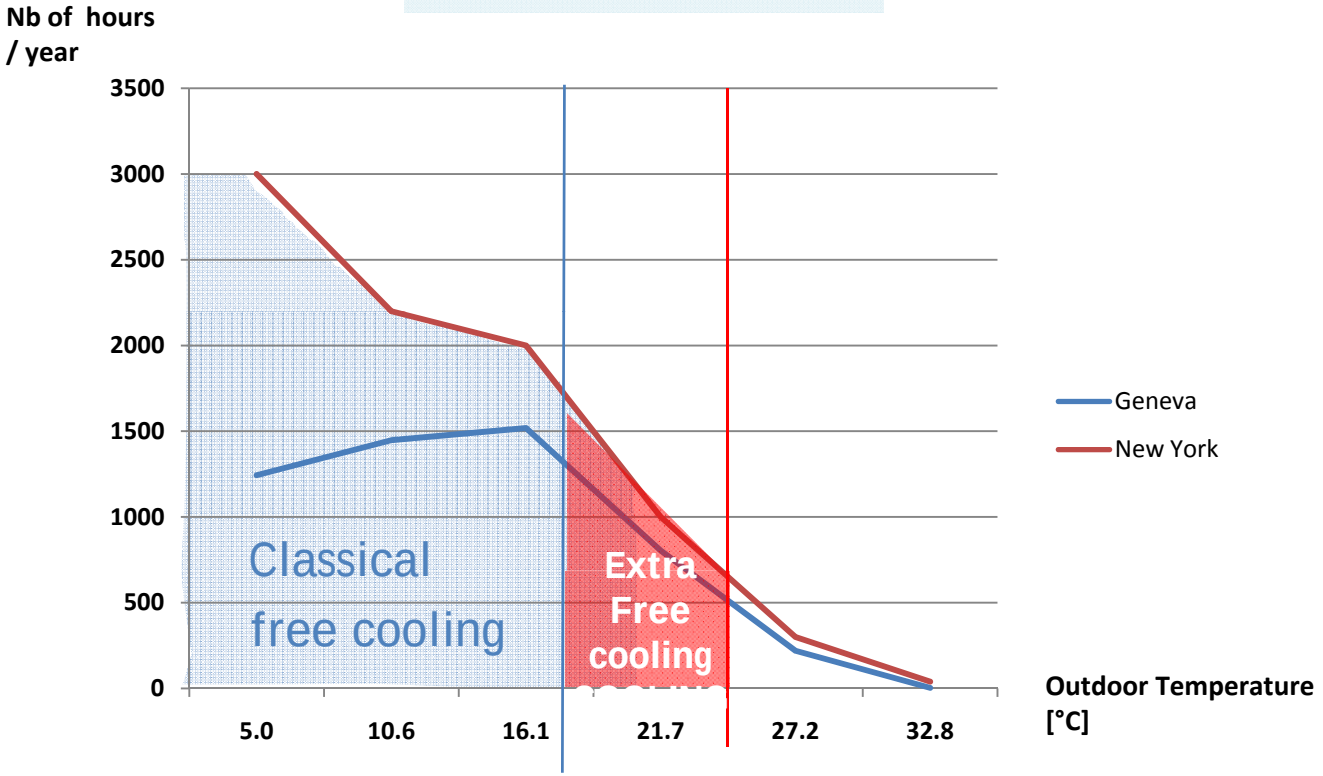
Data Centre Cooling Options

- Outside Air
- Cooled Air
  - Central distribution
  - Local distribu
  - Very local dist
- ~~Direct Water C~~

Basic Housekeeping is essential!

Basic Housekeeping is essential!

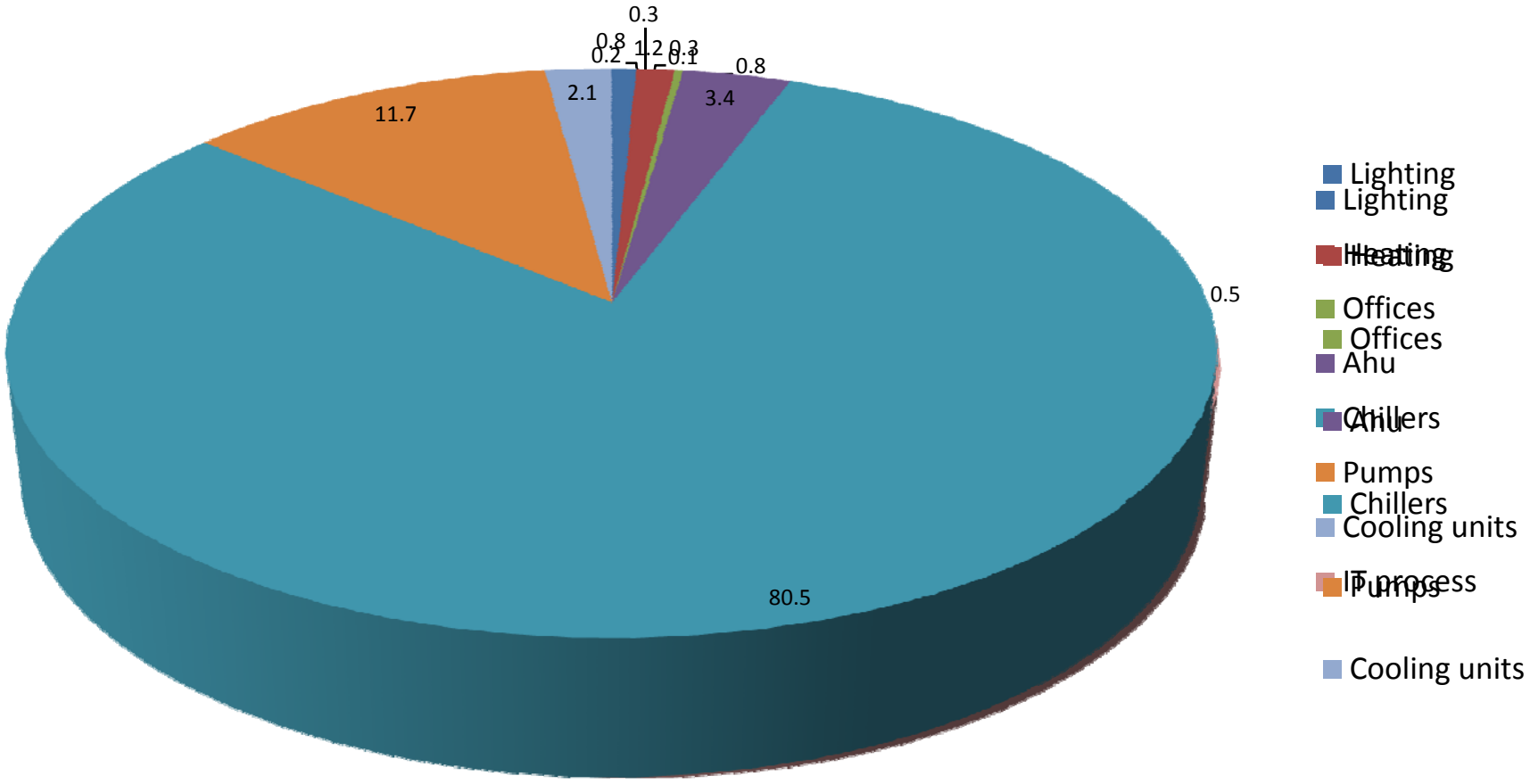
Hot is cool!



Indoor conditions from 20°C to 27°C

But not too hot: Vendors now accept that their systems will run at higher temperatures, but there are reports that server energy efficiency starts to drop above a certain temperature.

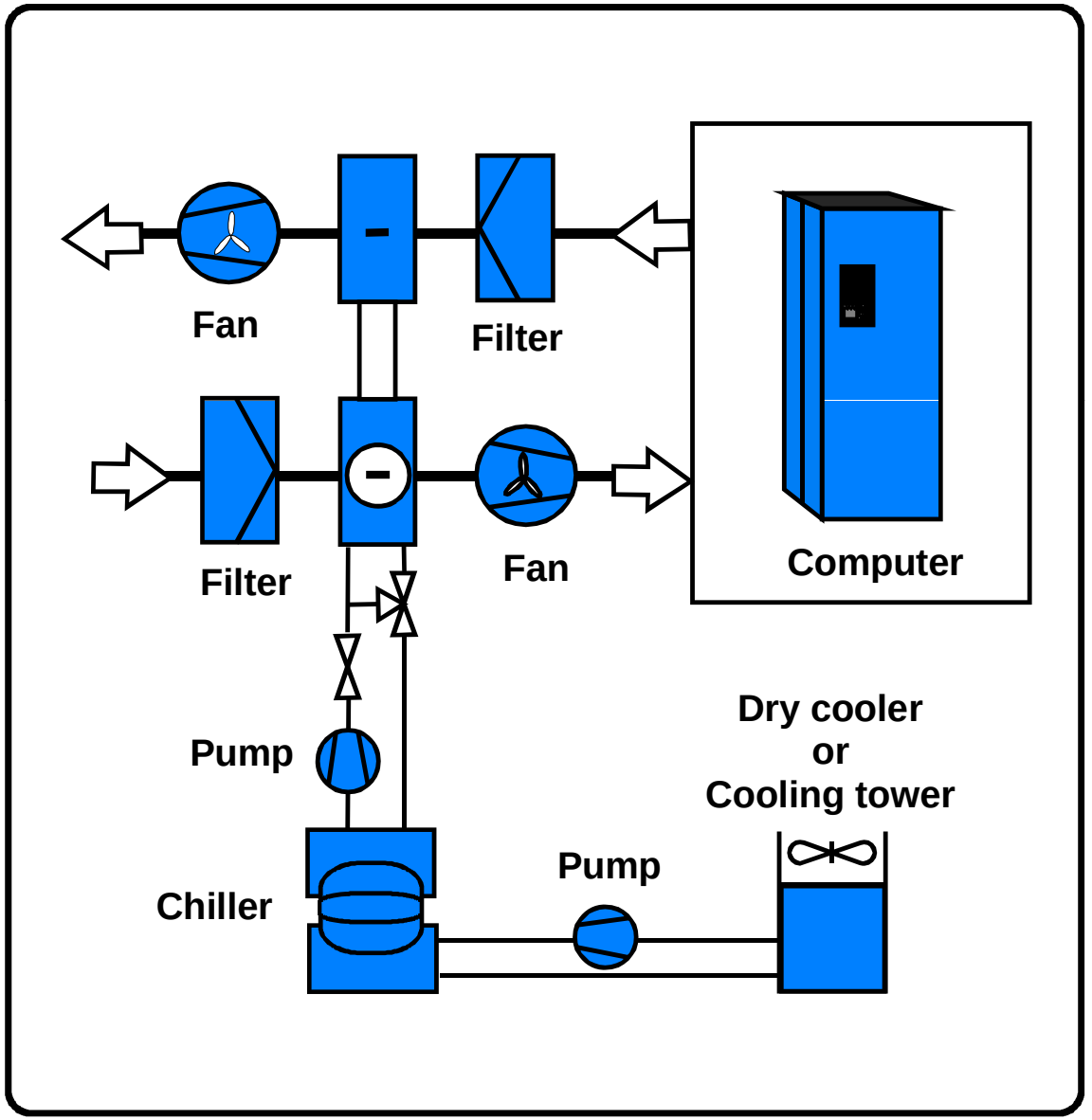
Annual Electricity Consumption



Efficient Data Centre — PUE=1.3

Conclusion

- **Optimise…**
- **… everywhere!**

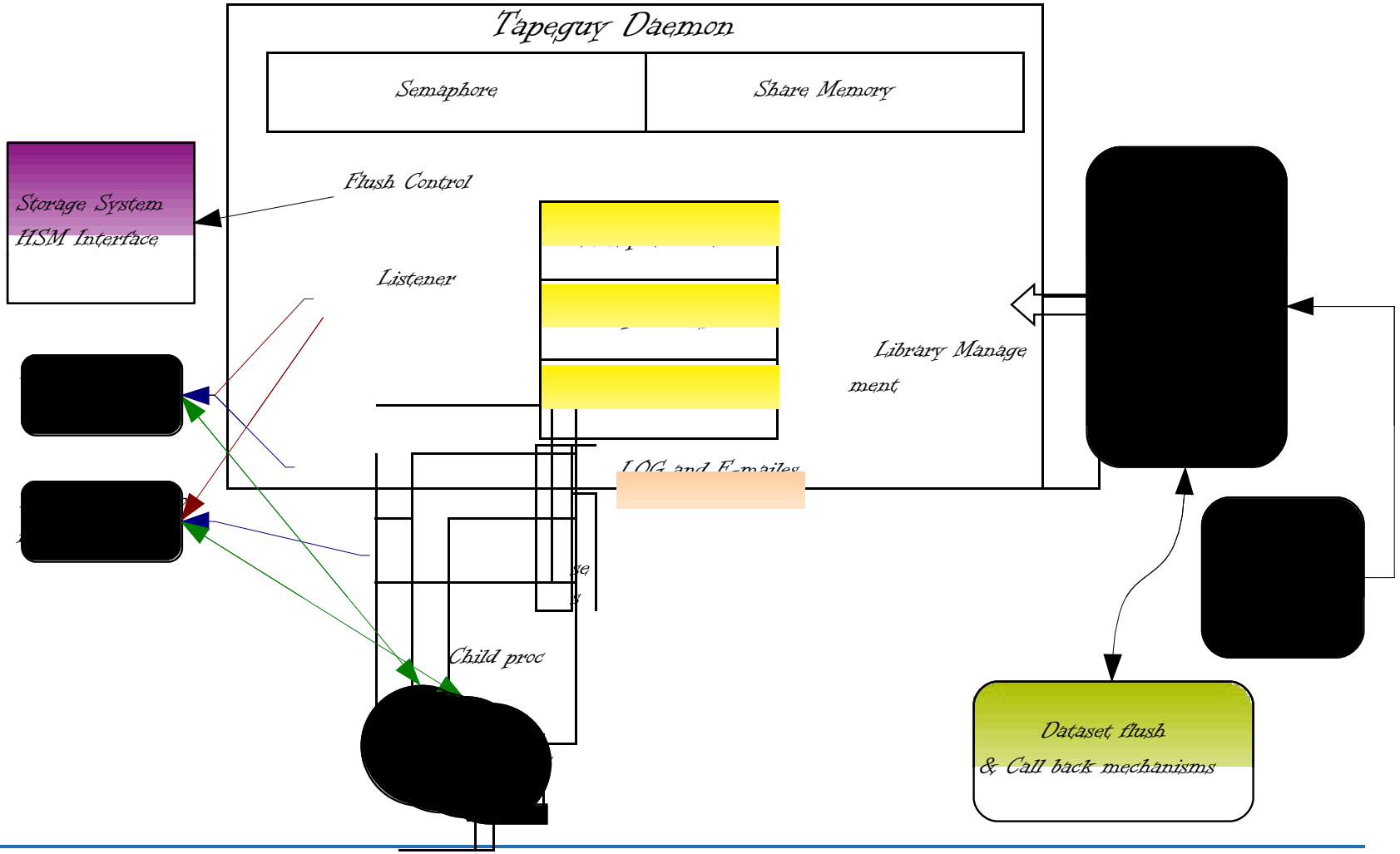


A High Performance Hierarchical Storage Management System For the Canadian Tier-1 Centre @ TRIUMF

Denice Deatrich, Simon Xinli Liu, Reda Tafirout

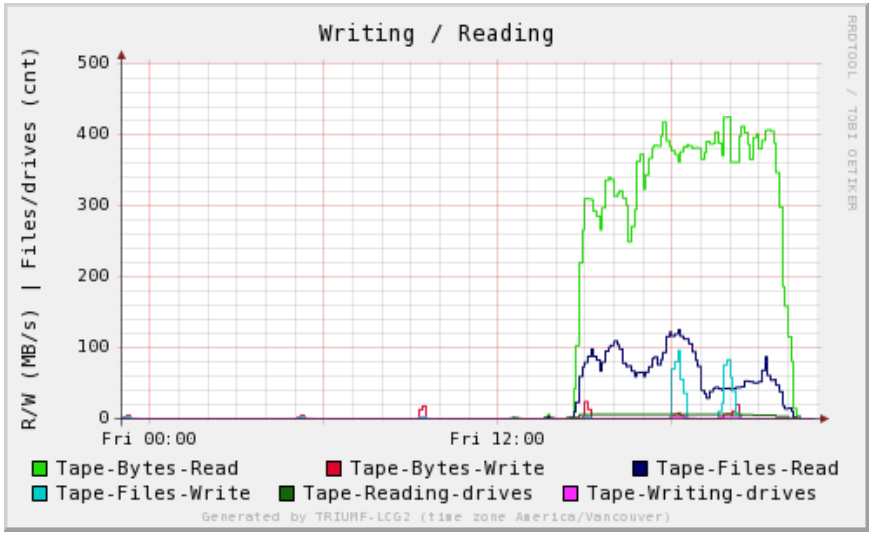
CHEP 09, Prague

Logical Architecture



Performance and Scalability(I)

- **Bulk pre-stage test**
  - **35 FDR datasets(3172 files)**
  - **9 TB data volume(13 tapes)**
  - **~8 hours to pre-stage (up to 6 drives)**
- **Currently can do: > 1 TB/hour**



Mass Storage Efficiency (MSS)

Date	R_Rate(MB/s)	W_Rate(MB/s)	Avg_File_R_Size(MB)	Avg_File_W_Size(MB)	R_Per_Mnt(MB)	W_Per_Mnt(MB)	R_Rep_Mnts	W_Rep_Mnts
2009Feb09	65.5	52.14	3001	4160	849740.4	37440	1.00(Total:11)	1.00(Total:0)

Performance and Scalability(II)

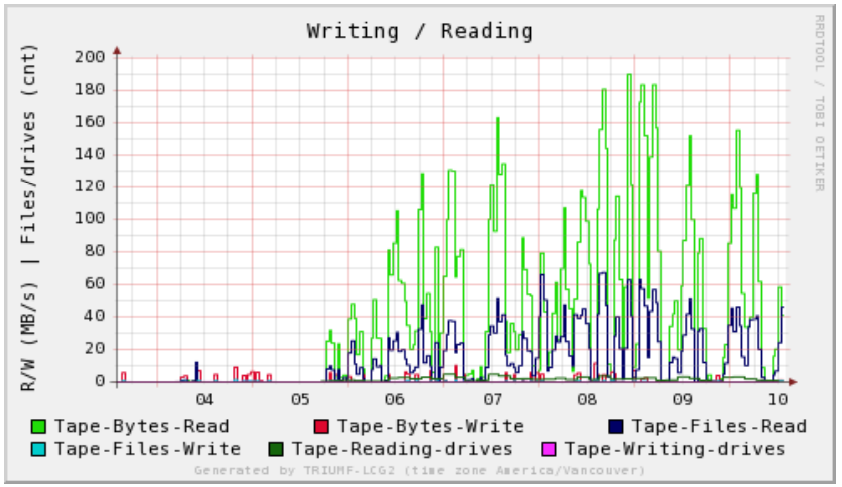
- **March-09 reprocessing (data to March 10)**

–

No file pre-stage in advance (not ideal sce nario, but reading still got benefit from dat aset level write grouping) –

105 datasets, 13987 files

–23 TB data volume (50 tapes involved)



Mass Storage Efficiency (MSS)

Date	R_Rate(MB/s)	W_Rate(MB/s)	Avg_File_R_Size(MB)	Avg_File_W_Size(MB)	R_Per_Mnt(MB)	W_Per_Mnt(MB)	R_Rep_Mnts	W_Rep_Mnts
2009Mar09	50.04	52.94	1831	3600	332270.36	43200	1.14(Total:16)	1.00(Total:0)
2009Mar08	40.61	59.82	1380	4373	240637.22	118080	1.50(Total:24)	1.00(Total:0)
2009Mar07	24.82	88.42	1820	3733	170268.62	100800	1.75(Total:28)	1.00(Total:0)
2009Mar06	36.45	79.73	1873	3960	149904.37	95040	1.41(Total:24)	1.00(Total:0)
2009Mar05	39.32	107.93	1808	4560	95840.5	54720	1.00(Total:3)	1.00(Total:0)

Conclusion

- **Tapeguy has been in production at the TRIUMF Tier-1 Centre since 2007 (a prototype version was developed in 2005 for Tier-1 service challenges)**
- **Provides greater control and flexibility than proprietary HSMs do**
- **Performance is good, and is expected to be scalable in order to match an increasing throughput demand in the coming years**

Fair-share scheduling algorithm for a tertiary storage system

Pavel Jakl[1] Jérôme Lauret[2]

[1]Nuclear Physics Institute, Academy of Science, Czech Republic

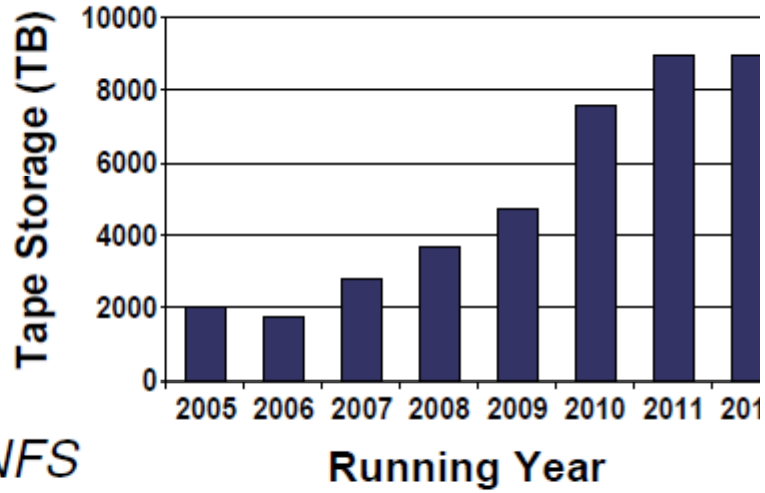
[2]Brookhaven National Laboratory, United States of America

CHEP 2009
Prague, Czech Republic

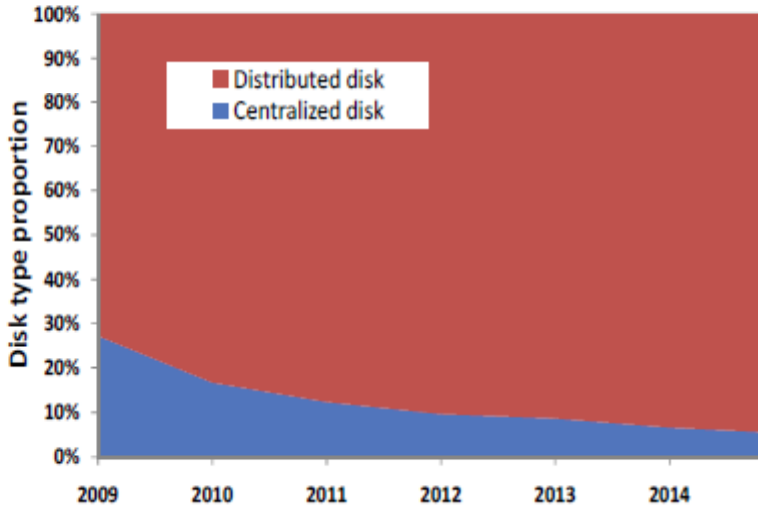
BROOKHAVEN NATIONAL LABORATORY

STAR ☆

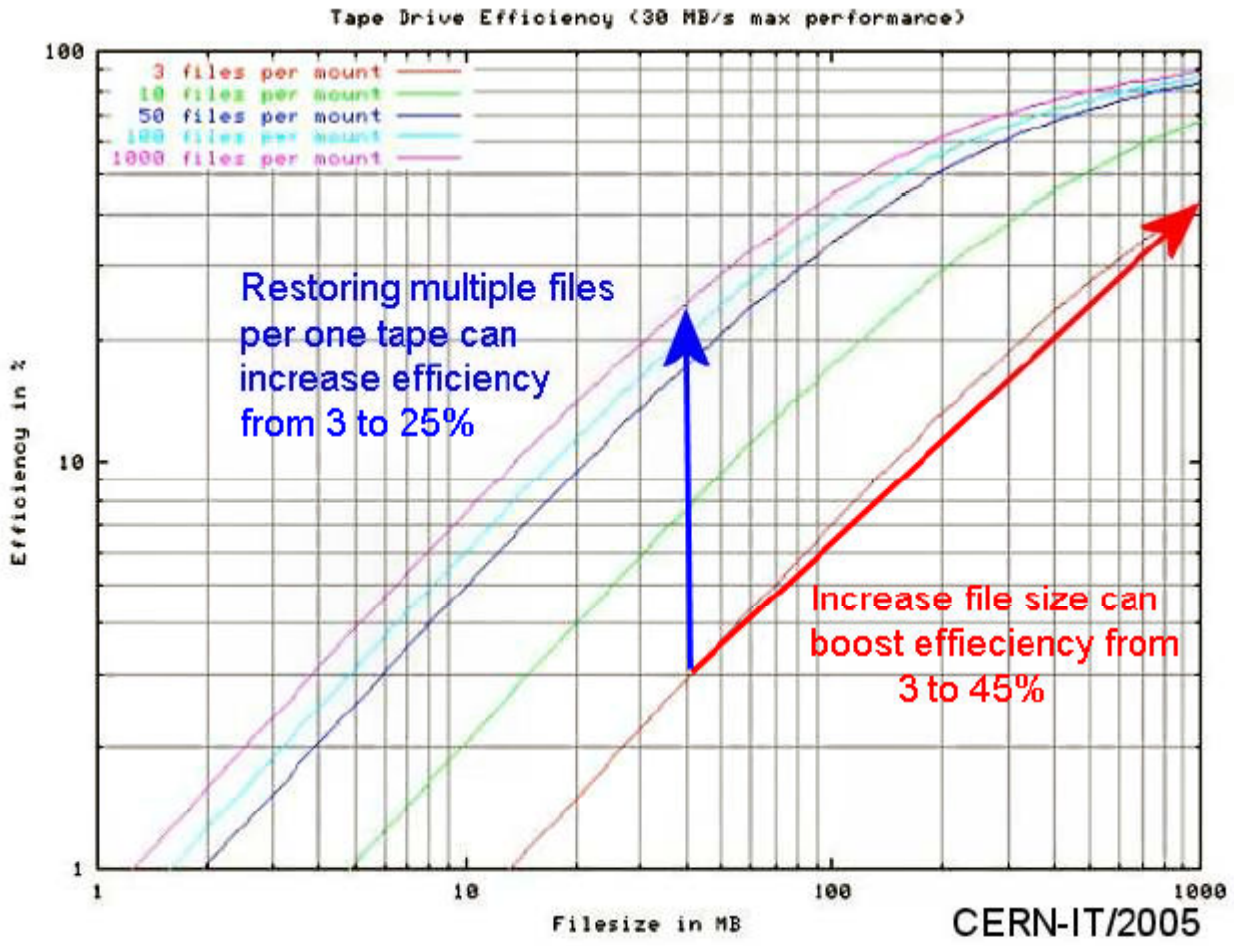
- over 1PB data per year at STAR
- *Permanent* location:
  - tape system (MSS): offers several **PBs**
- *Temporary* locations:
  - centralized disk space: **150 TB** *via NFS*
  - distributed disk space: **350 TB** *spread over 1000 nodes*



- **distributed** vs centralized disk:
  - \+ very low cost (factor of ∼10)
  - \+ less human resources to maintain
  - − worse manageability (one has to build aggregation)
  - − no native OS/system provides scalable/workable solution



Reasons for slow HPSS performance



Summary

- we have shown and demonstrated rational behind key performance parameters of the HPSS
  - future runs in STAR have optimal file size
  - pre-staging technique has to be used for the efficient tape optimizations
- scheduling algorithm should not only incorporate performance parameters but also has to be "enough" fair to the users
- simulation of the tape system distinguished efficient fair-share scheduling algorithm
  - we recommend Weighted Fair-share Grouping (WFSG) algorithm to achieve good throughput, maximal QoS and lowest delay
- Future work:
  - an implementation of the WFSG algorithm into the production system
  - an measurement of the algorithm's efficiency in the production

Lustre File System Evaluation at FNAL

Stephen Wolbers
for
Alex Kulyavtsev, Matt Crawford, Stu Fuess, Don Holmgren, Dmitry Litvintsev, Alexander Moibenko, Stan Naymola, Gene Oleynik,Timur Perelmutov, Don Petravick, Vladimir Podstavkov, Ron Rechenmacher, Nirmal Seenu, Jim Simone
Fermilab

•CHEP'09, Prague
March 23, 2009

Lustre Experience - HPC

- From our experience in production on Computational Cosmology Cluster (starting summer 2008) and limited pre-production on LQCD JPsi cluster (December 2008) the Lustre File system:
  - Lustre doesn't suffer the MPI deadlocks of dCache
  - direct access eliminates the staging of files to/from worker nodes that was needed with dCache (Posix IO)
  - improved IO rates compared to NFS and eliminated periodic NFS server "freezes"
  - reduced administration effort

Conclusions - HEP

- Lustre file system meets and exceeds our storage evaluation criteria in most areas, such as system capacity, scalability, IO performance, functionality, stability and high availability, accessibility, maintenance, and WAN access.
- Lustre has *much* faster metadata performance than our current storage system.
- At present Lustre can only be used for HEP applications not requiring large scale tape IO, such as LHC T2/T3 centers or scratch or volatile disk space at T1 centers.
- Lustre near term roadmap (about one year) for HSM in principle satisfies our HSM criteria. Some work will still be needed to integrate any existing tape system.

Session2:Tuesday, 24 March 2009 14:00-

- [38] The ALICE Online Data Storage System
  by Roberto DIVIà (CERN)
- [89] Integration of Virtualized Worker Nodes into Batch Systems.
  by Oliver OBERST (Karlsruhe Institute of Technology)
- [165] SL(C)5 for HEP - a status report
  by Ricardo SALGUEIRO DOMINGUES DA SILVA (CERN)
- [136] The NAF: National Analysis Facility at DESY
  by Andreas HAUPT (DESY); Yves KEMP (DESY)
- [224] Operational Experience with CMS Tier-2 Sites
  by Dr. Isidro GONZALEZ CABALLERO (Instituto de Fisica de Cantabria, Grupo de Altas Energias)
- [270] ScotGrid: Providing an Effective Distributed Tier-2 in the LHC Era
  by Dr. Graeme Andrew STEWART (University of Glasgow); Dr. Michael John KENYON (University of Glasgow); Dr. Samuel SKIPSEY (University of Glasgow)

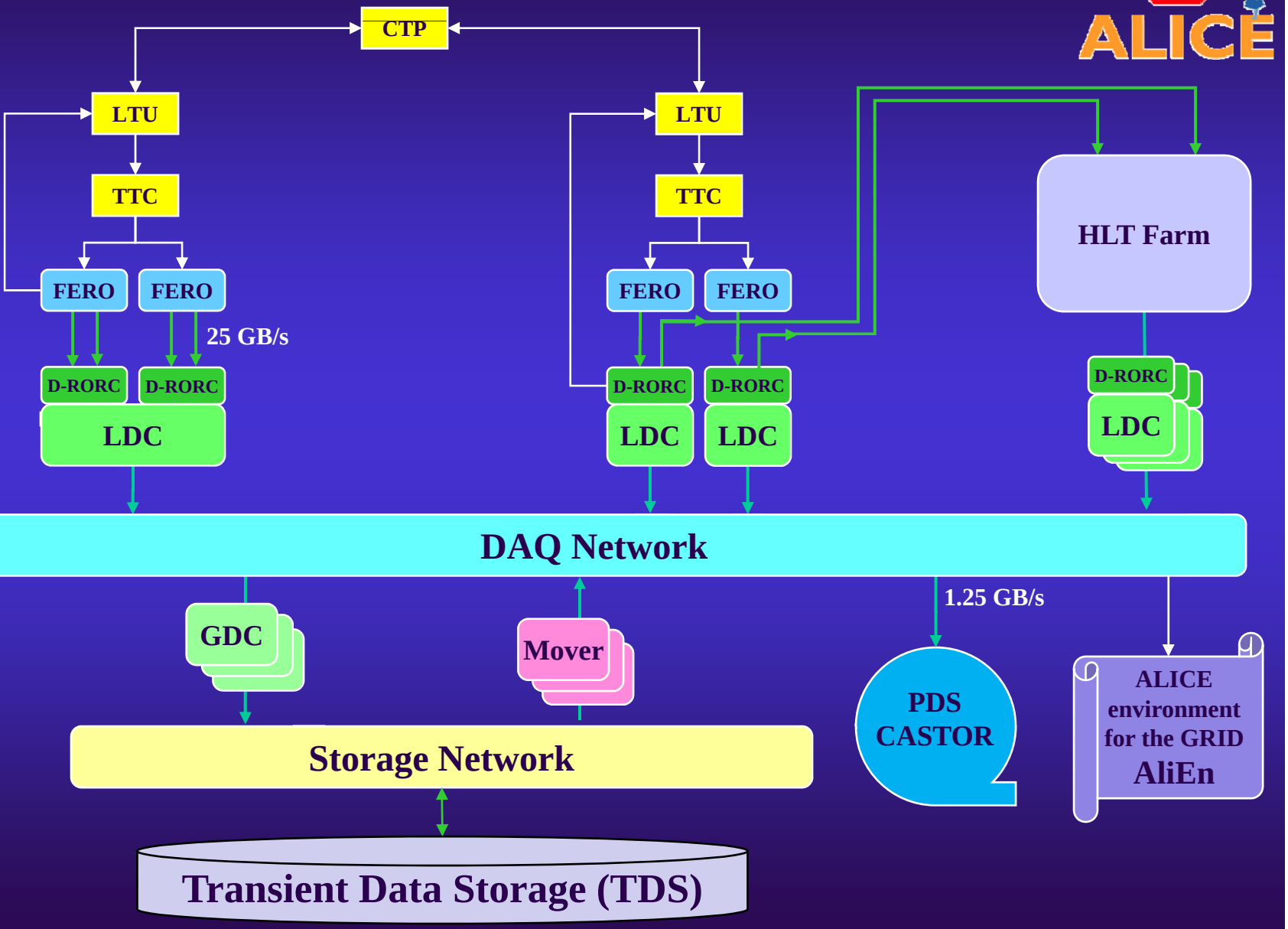
The ALICE Online Data Storage System

Roberto Divià (CERN), Ulrich Fuchs (CERN), Irina Makhlyueva (CERN), Pierre Vande Vyvre (CERN)

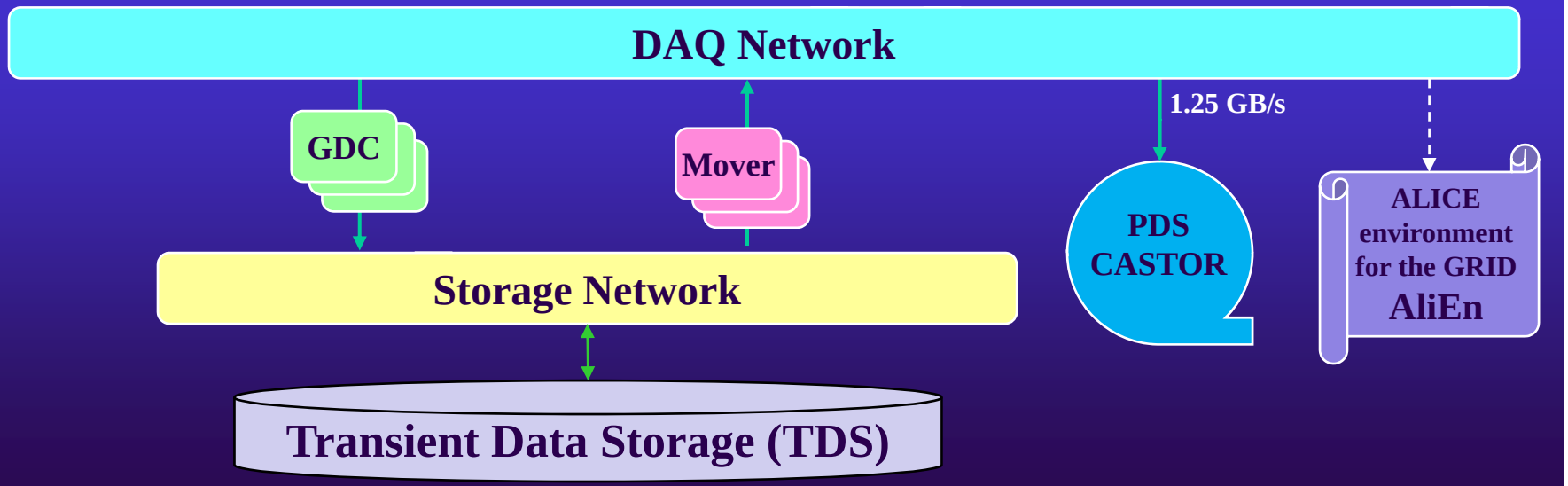
Valerio Altini (CERN), Franco Carena (CERN), Wisla Carena (CERN), Sylvain Chapeland (CERN), Vasco Chibante Barroso (CERN), Filippo Costa (CERN), Filimon Roukoutakis (CERN), Klaus Schossmaier (CERN), Csaba Soòs (CERN), Barthelemy Von Haller (CERN)

For the ALICE collaboration

ALICE trigger, DAQ & HLT



ALICE trigger, DAQ & HLT



Our objectives

Ensure steady and reliable data flow up to the design specs

Avoid stalling the detectors with data flow slowdowns

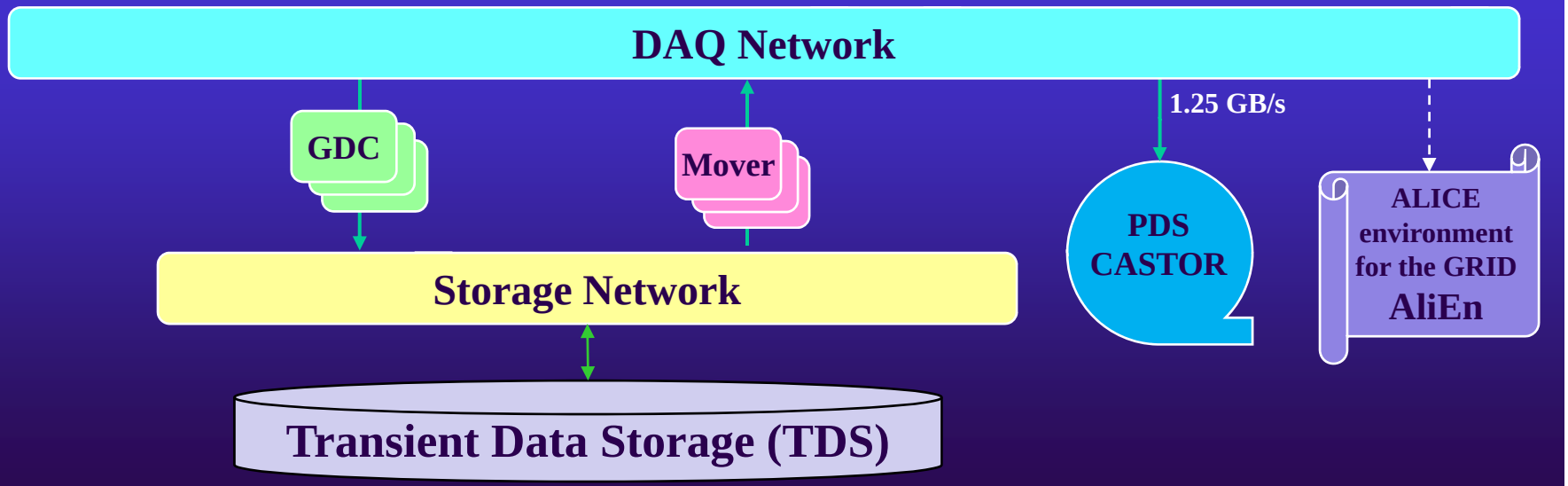
Give sufficient resources for online objectification in ROOT format via AliROOT

- very CPU-intensive procedure

Satisfy needs from ALICE parallel runs and from multiple detectors commissioning

Allow a staged deployment of the DAQ/TDS hardware

Provide sufficient storage for a complete LHC spill in case the transfer between the experiment and the CERN Computer Center does not progress



In conclusion…

- Continuous evaluation of HW & SW components proved the feasibility of the TDS/TDSM architecture
- All components validated and profiled
- ADCs gave highly valuable information for the R&D process
  - Additional ADCs added to the ALICE DAQ planning for 2009
- Detector commissioning went smoothly & all objectives were met
- No problems during cosmic and preparation runs
- Staged commissioning on its way
- Global tuning in progress

We are ready for LHC startup

FIO Fabric Infrastructure and Operations

CERN IT Department

SL(C) 5 Migration at CERN

CHEP 2009, Prague

Ulrich SCHWICKERATH

Ricardo SILVA

CERN, IT-FIO-FS

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Motivation – Context and lifecycle (1)

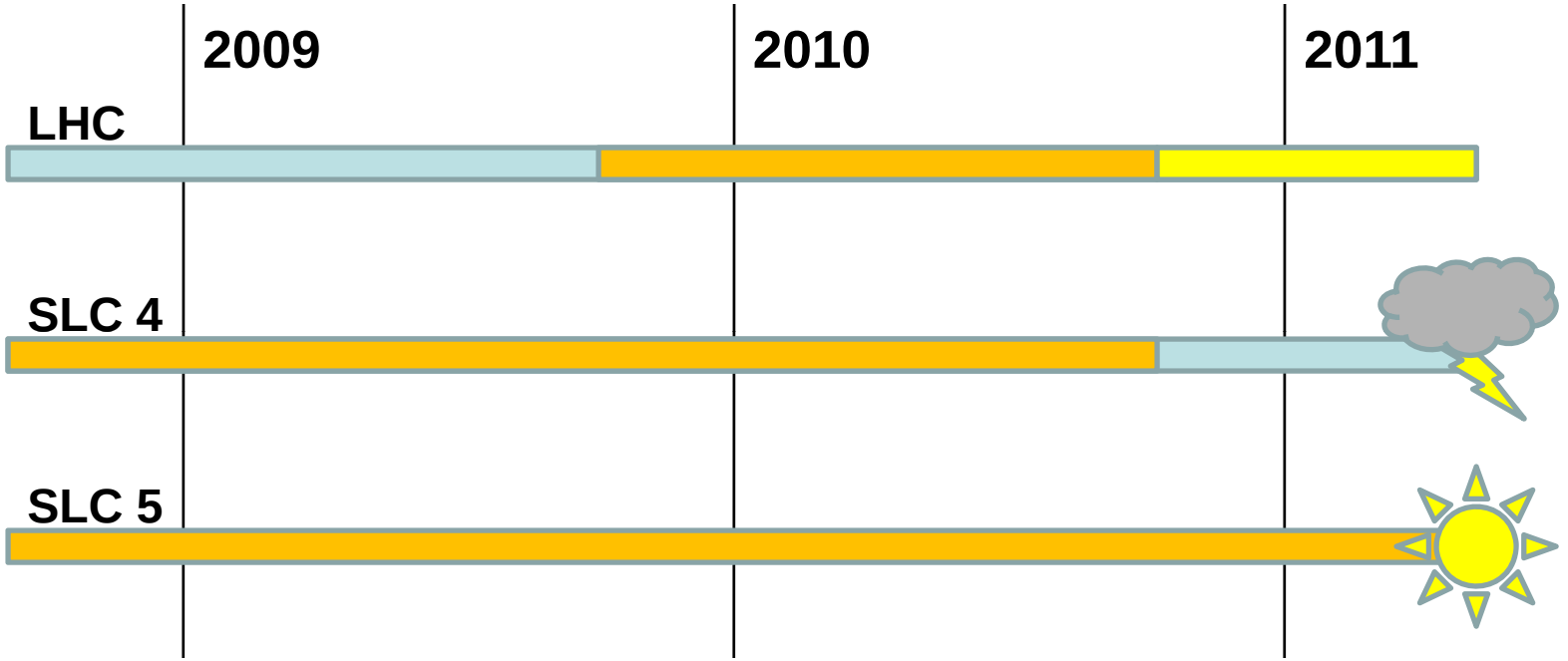
RHEL 4
(Feb 2005 - Feb 2009*)
▼
SL 4
(Apr 2005 - Oct 2010)
▼
SLC 4
SLF 4

RHEL 5
(Mar 2007 - Mar 2011*)
▼
SL 5
(May 2007 - 2011(?))
▼
SLC 5
SLF 5

* (End of Production 1 Phase)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Motivation – Context and lifecycle (2)



- We want stability during the LHC run period!

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Conclusions

- Move to SLC5 as main operating system well ahead of data taking
  - GDB: “Every experiment is interested on a transition to SLC5/64bit of the Grid resources as soon and as short as possible.”
- CERN has been providing SLC5 resources for several months
- Close collaboration with the experiments in the move to SLC5
- Extensively tested and production ready
- Confident on a quick and painless transition
- No known showstoppers for a large scale migration

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The NAF: National Analysis Facility at DESY.

The NAF concept and its place

- In the German HEP field
- In the Global Grid

The building blocks, with an emphasis on

- Usage of VOMS for Grid-type resources
- Interactive and batch cluster and integration with PROOF
- Usage of Grid-Proxies to access workgroup servers, AFS and dCache data
- The usage and operation of Lustre for fast data access.

Running experience

PHYSICS AT THE TERA SCALE
Helmholtz Alliance

HELMHOLTZ ASSOCIATION

Andreas Haupt, Yves Kemp (DESY)
Prague, CHEP 2009, 24.3.2009

DESY

PHYSICS AT THE TERASCALE
Helmholtz Alliance

HELMHOLTZ GEMEINSCHAFT

- Helmholtz Alliance: Physics at the Terascale
- Collaboration between ~20 german universities and research centers
  - People working on “Tera-eV physics”, e.g. LHC, ILC, and respective theorists
- Many different research fields
  - Physics Analysis, Detector Technologies, Accelerator Physics, Grid Computing
- http://www.terascale.de for more information

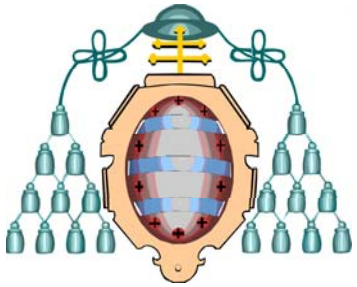
- NAF is part of the Grid Computing research topic
  - Give users of the German institutes of the LHC and ILC experiments additional resources for analysis
  - Atlas, CMS, LHCb and ILC
  - Size ~1.5 average LHC Tier2, with emphasis on data storage
  - Intended as distributed facility, starting at DESY (with its two sites Hamburg and Zeuthen)

Summary and Outlook

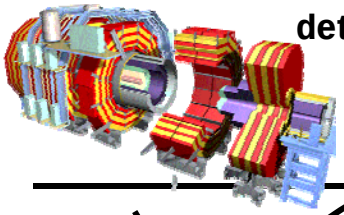
- The NAF is working: ~300 registered users
- Hardware resources already substantial, enlargement in 2009
- Generic approach:
  - All analysis workflows supported
  - All communities supported on one infrastructure
- More information, documentation and links to support:
  - http://naf.desy.de/
- **We all are waiting for our first great challenge: The first LHC colliding-beam data!**
- Questions? Comments? Welcome!
  - Yves.Kemp@desy.de / Andreas.Haupt@desy.de

Operational Experience with CMS Tier-2 Sites

I. González Caballero
(Universidad de Oviedo)
*for the **CMS Collaboration***



CMS Computing Model



detector

event filter (selection & reconstruction)

reconstruction

raw data

event summary data

event reprocessing

Tier-2

processed data

batch physics analysis

analysis

analysis objects (extracted by physics topic)

interactive physics analysis

event simulation

simulation

Tier-2

les.robertson@cern.ch

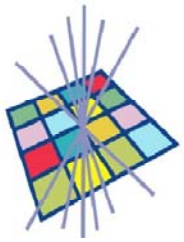
Future plans…

- The main goal in the near future is to completely integrate all the CMS Tier-2s into CMS computing operations
  - Using dedicated task forces to help sites meet the Site Readiness metrics
- Improve the availability and reliability of the sites to increase further the efficiency of both analysis and production activities
- Complete the data transfer mesh by commissioning the missing links
  - Specially Tier-2 → Tier-1 links
  - And continue checking the already commissioned links
- Improve the deployment of CMS Software loosening the requisites at the sites
- Install CRAB Servers at more sites:
  - CRAB Server takes care of some user routine interactions with the GRID improving the user experience
  - Improves the accounting and helps spotting problems and bugs in CMS software
  - A new powerful machine and special software needs to be installed by local operators
- CMS is building the tools to allow users to share their data with other users or groups
  - This will impact on the way data is handled at the sites

Conclusions

- Tier-2 sites play a very important role in the CMS Computing Model: They are expected to provide more than one third of the CMS computing resources
- CMS Tier-2 sites handle a mix of centrally controlled activity (MC production) and chaotic workflows (user analysis)
  - CPU needs to be appropriately set to ensure enough resources are given to each workflow
- CMS has built the tools to facilitate the day by day handling of data at the sites
  - The PhEDEx servers located at every site helps transferring data in an unattended way
  - A Data Manager appointed at every site links CMS central data operations with the local management
- CMS has established metrics to validate the availability and readiness of the Tier-2s to contribute efficiently to the collaboration computing needs
  - By verifying the ability to transfer and analyze data
  - A big number of tools have been developed by CMS and the GS group (CERN IT) to monitor every aspect of a Tier-2 in order to better identify and correct the problems that may appear
- CMS Tier-2s have proved to be already well prepared for massive data MC production, dynamic data transfer, and efficient data serving to local GRID clusters
- CMS Tier-2s have proved to be able to provide our physicists with the infrastructure and the computing power to perform their analysis efficiently

CMS Tier-2s have a crucial role to play in the coming years in the experiment, and are already well prepared for the LHC collisions and the CMS data taking



ScotGrid:

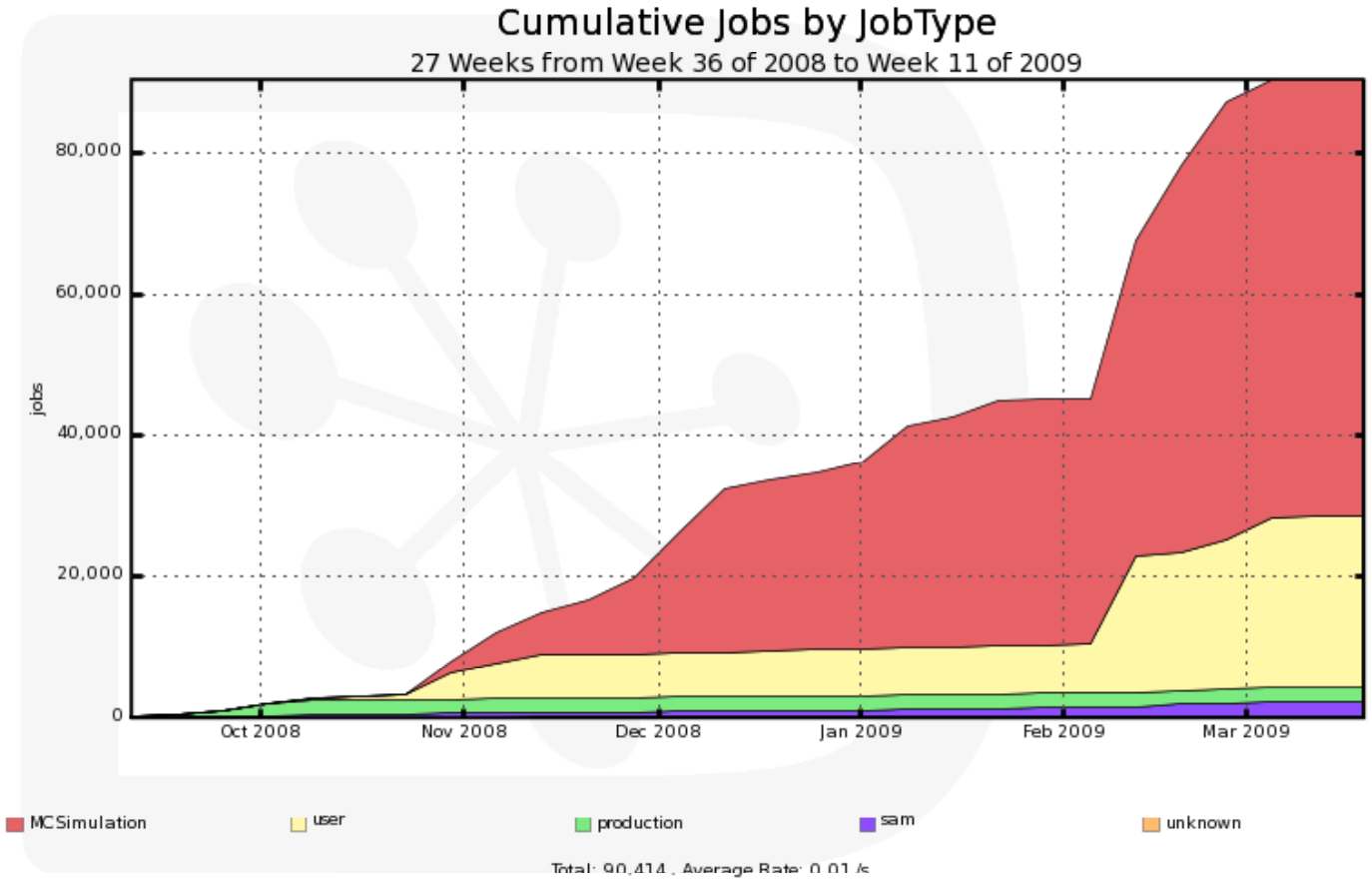
Providing an Effective Distributed Tier-2 in the LHC Era

Sam Skipsey

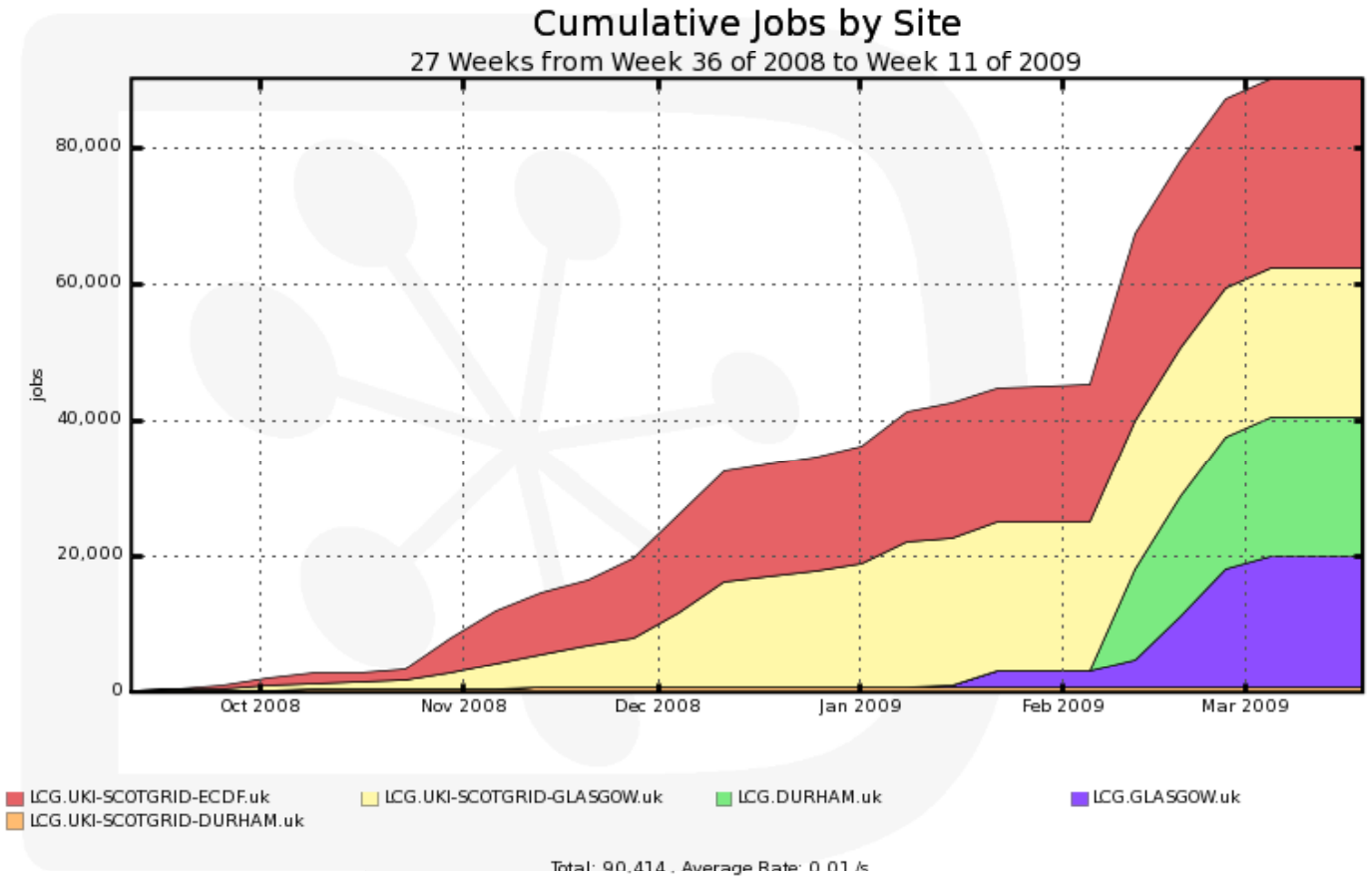
David Ambrose-Griffith, Greig Cowan, Mike Kenyon, Orlando Richards Phil Roffe, Graeme Stewart

Universities of Glasgow, Edinburgh and Durham

LHCb Jobtype changes



LHCb usage across sites



Conclusions

- Communication is essential!
- Be prepared to be flexible.
- Local copies of “central” services
  - Split load
  - But add overhead.

Session 3:**Tuesday, 24 March 2009 16:00-**

- [395] Study of Solid State Drives performance in PROOF distributed analysis system
  by Dr. Sergey PANITKIN (Department of Physics - Brookhaven National Laboratory (BNL))
- [282] Monitoring Individual Traffic Flows in the Atlas TDAQ Network
  by Mr. Rune SJOEN (Bergen University College)
- 17:00 [28] Oracle and storage IOs, explanations and experience at CERN
  by Mr. Eric GRANCHER (CERN)
- [229] A Service-Based SLA for the RACF at Brookhaven National Lab
  by Ms. Mizuki KARASAWA (Brookhaven National Laboratory); Dr. Jason SMITH (Brookhaven National Laboratory)
- [233] The Integration of Virtualization into the U.S. ATLAS Tier 1 Facility at Brookhaven
  by Mr. Christopher HOLLOWELL (Brookhaven National Laboratory); Mr. Robert PETKUS (Brookhaven National Laboratory)

SSD tests at PROOF farm at BNL

Michael Ernst, Sergey Panitkin, Robert Petkus, Ofer Rind, Torre Wenaus

BNL

March, 24
CHEP 2009
Prague, Czech Republic

395-SSD.pdf - Adobe Reader

ファイル(F) 編集(E) 表示(V) 文書(D) ツール(T) ウィンドウ(W) ヘルプ(H)

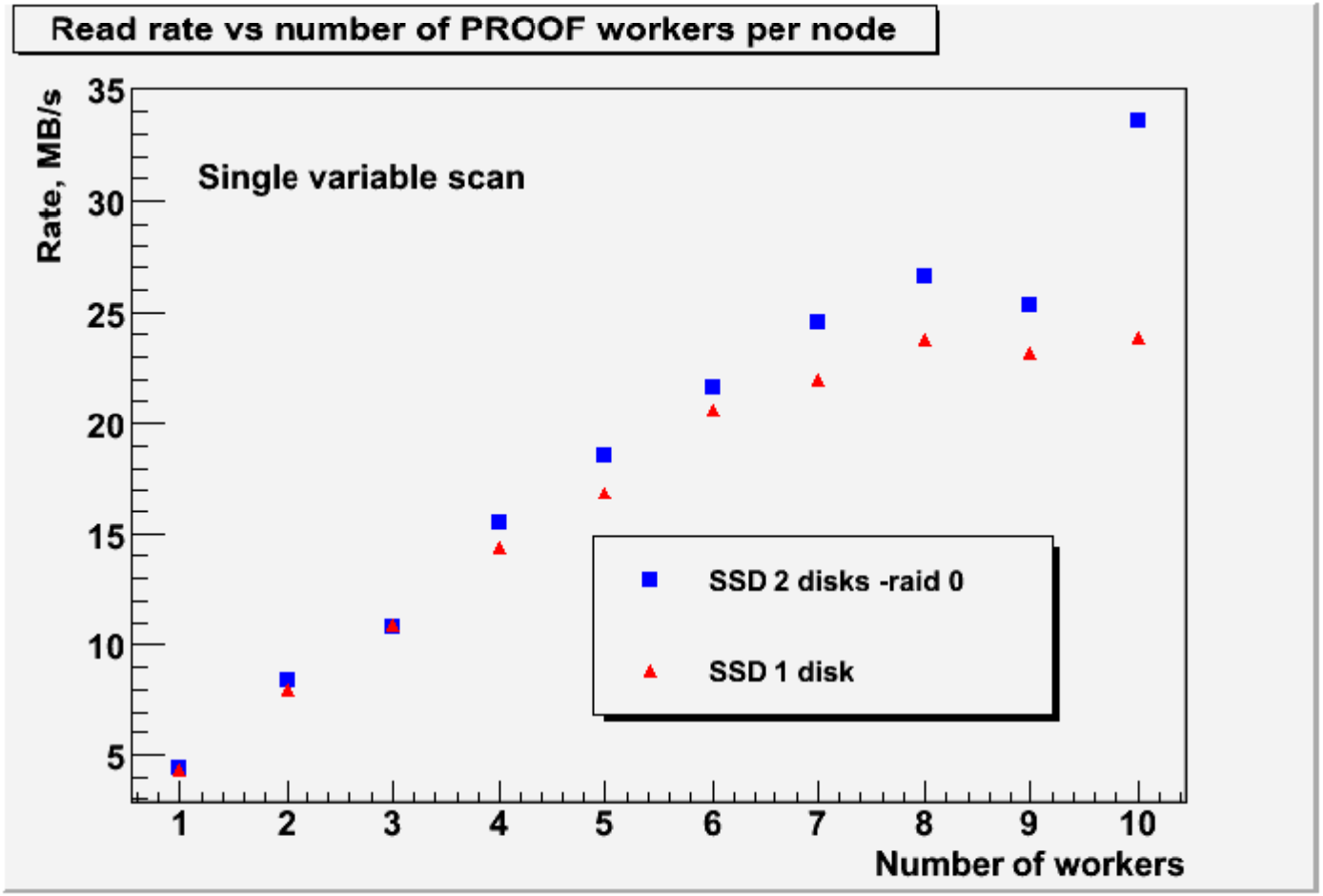
Solid State Disks Used for Tests

- Model: Mtron MSP-SATA7035064
- Capacity 64 GB
- Average access time ~0.1 ms (typical HD ~10ms)
- Sustained read ~120MB/s
- Sustained write ~80 MB/s
- IOPS (Sequential/ Random) 81,000/18,000
- Write endurance >140 years @ 50GB write per day
- MTBF 1,000,000 hours
- 7-bit Error Correction Code

395-SSD.pdf - Adobe Reader

ファイル(F) 編集(E) 表示(V) 文書(D) ツール(T) ウィンドウ(W) ヘルプ(H)

SSD: single disk vs RAID

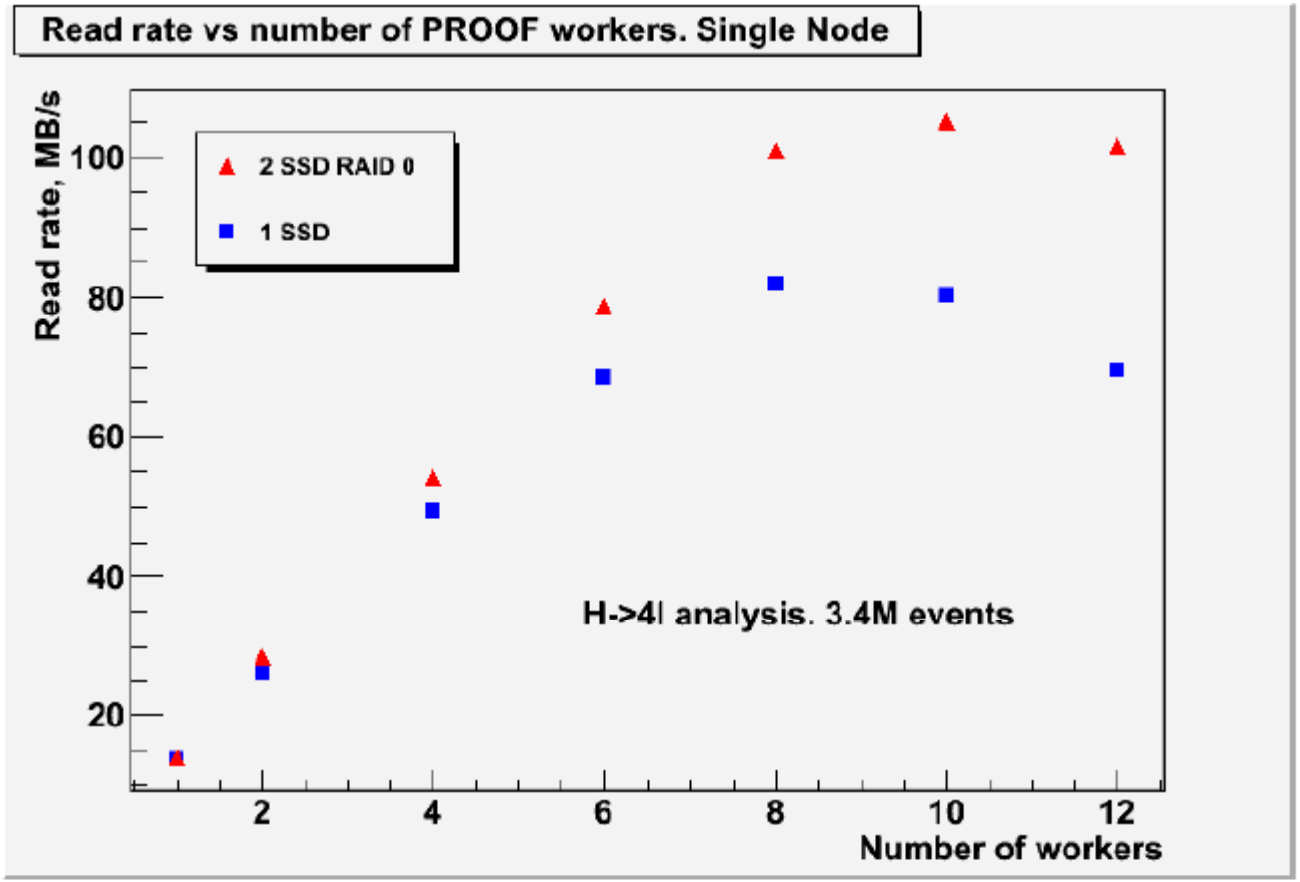


- SSD RAID has minimal impact until 8 simultaneously running jobs

395-SSD.pdf - Adobe Reader

ファイル(F) 編集(E) 表示(V) 文書(D) ツール(T) ウィンドウ(W) ヘルプ(H)

H->4l analysis. SSD RAID 0



SSD 2 disk RAID 0 shows little impact up to 4 worker load

395-SSD.pdf - Adobe Reader

ファイル(F) 編集(E) 表示(V) 文書(D) ツール(T) ウィンドウ(W) ヘルプ(H)

Summary and Discussion

- SSD technology offer significant performance advantage in concurrent analysis environment
- We observed~x10 better read performance than HDD in our test
- The main issue, in PROOF context, is matching of local I/O demand and supply
- Some observations from our tests
  - Single analysis worker in PROOF can generate ~10-15 MB/s read load
  - One SATA HDD can sustain ~2-3 PROOF workers
  - HDD RAID array can sustain ~ 3 to 6 workers
  - One Mtron SSD can sustain ~8 workers, almost at peak performance
  - SSD RAID is nice, but not really necessary with current hardware
- Currently the main issue with SSD is size (and cost) .
- Multi tiered local disk sub-system, with automatic pre-staging of data from HDD to SSD may be a promising solution which can provide both capacity and speed. Efficient data management is needed.

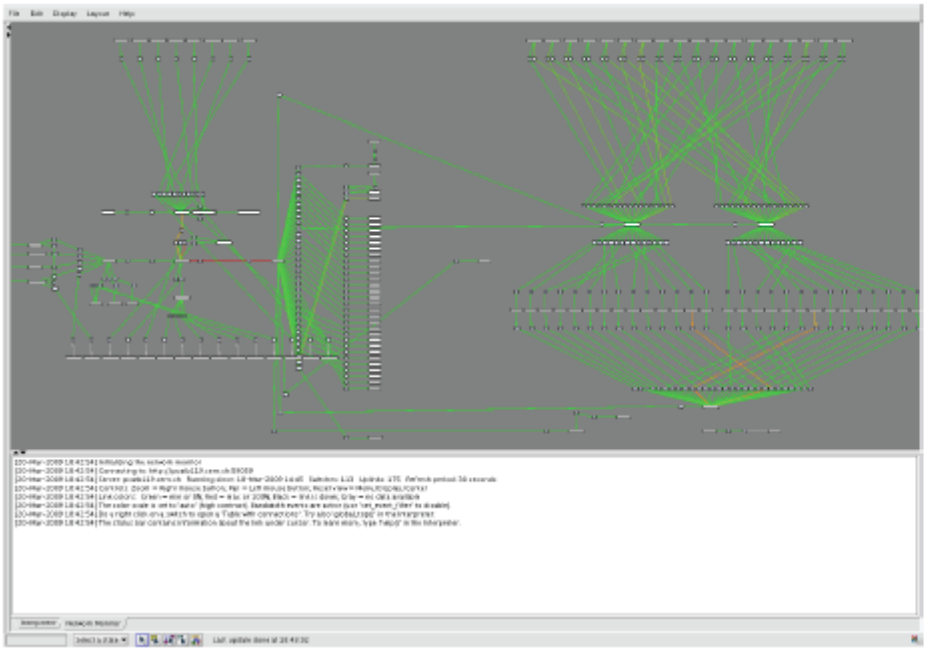
  We plan to investigate this option

Monitoring Individual Traffic Flows in the ATLAS TDAQ Network

R.Sjoen, S.Stancu, M.Ciobotaru, S.M.Batraneanu, L.Leahu, B.Martin, A.Al-Shabibi

March 21, 2009

Introduction



- 5 multi-blade chassis devices
- 200 edge switches
- 2000 processors
- Well known set of applications
- Classical SNMP-based monitoring to provide statistics on aggregate traffic
- Difficult to monitor, troubleshoot and quantify single traffic flows

Conclusion

- Taking it one step further compared to classical SNMP based monitoring
- By gaining a deeper knowledge about the traffic flows we have the ability to identify unknown traffic patterns
- When a problem is detected diagnostics can be performed immediately without having to reproduce the problem
- An intuitive interface allow non-expert users to easily obtain the desired information
- Historical analysis of events only limited by storage space

CERN IT Department

Oracle and storage IOs, explanations and experience at CERN
CHEP 2009 Prague [id. 28]

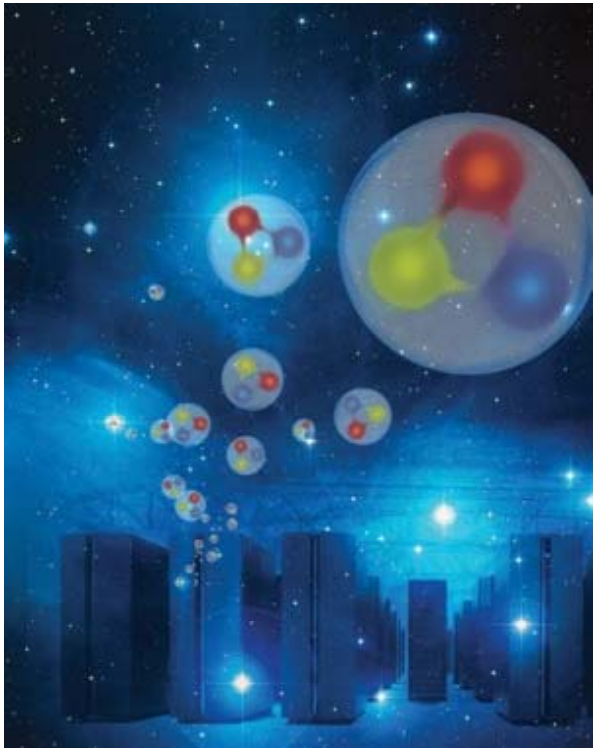


Image courtesy of Forschungszentrum Jülich / Seitenplan, with material from NASA, ESA and AURA/Caltech

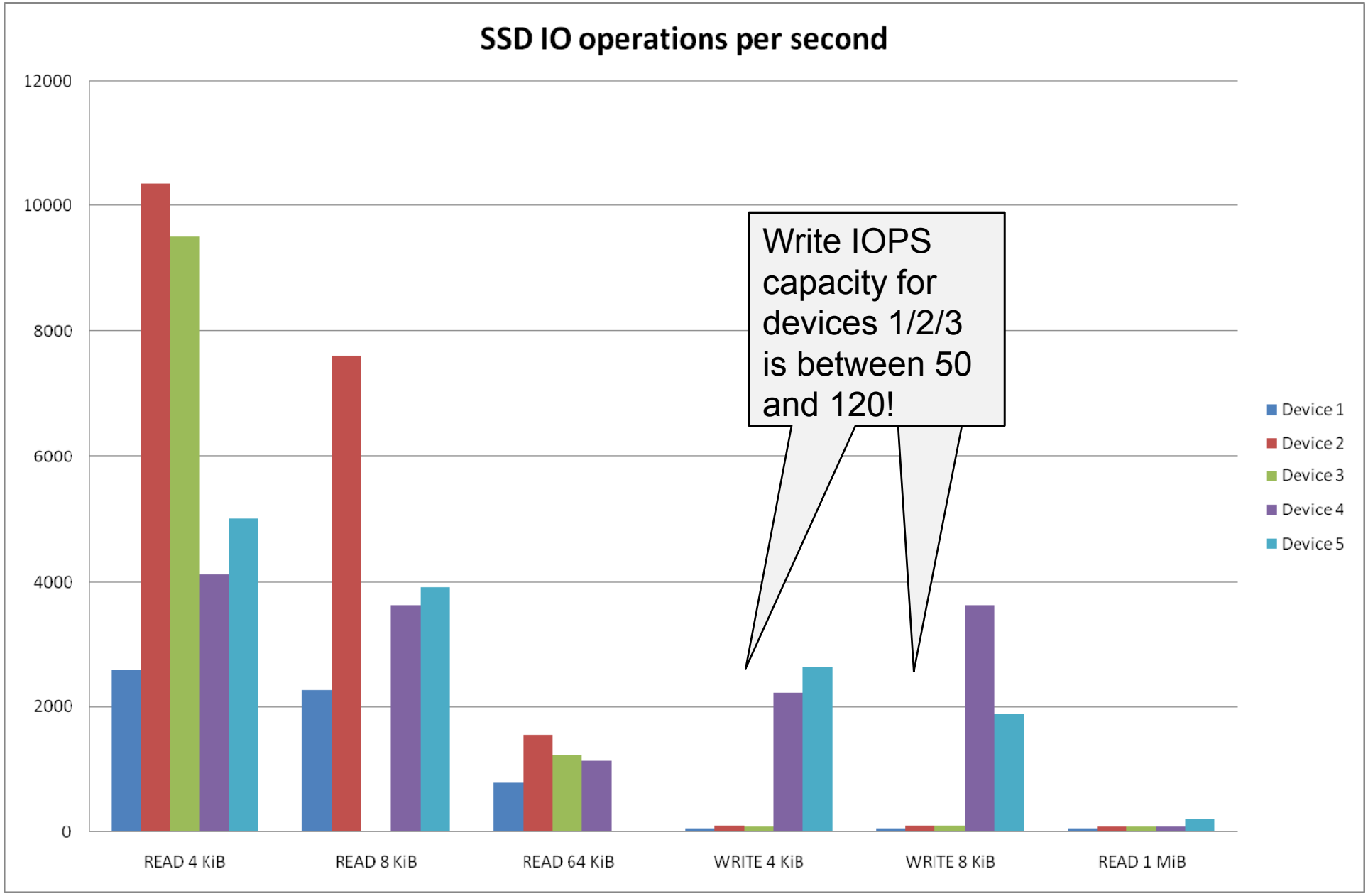
Eric Grancher
eric.grancher@cern.ch
CERN IT department

Conclusions

- New tools like ASH and DTrace change the way we can track IO operations
- Overload in IO and CPU can not be seen from Oracle IO views
- Exadata offloading operations can be interesting (and promising)
- Flash SSD are coming, a lot of differences between them. Writing is the issue (and is a driving price factor). Not applicable for everything. Not to be used for everything for now (as write cache? Oracle redo logs). They change the way IO operations are perceived.

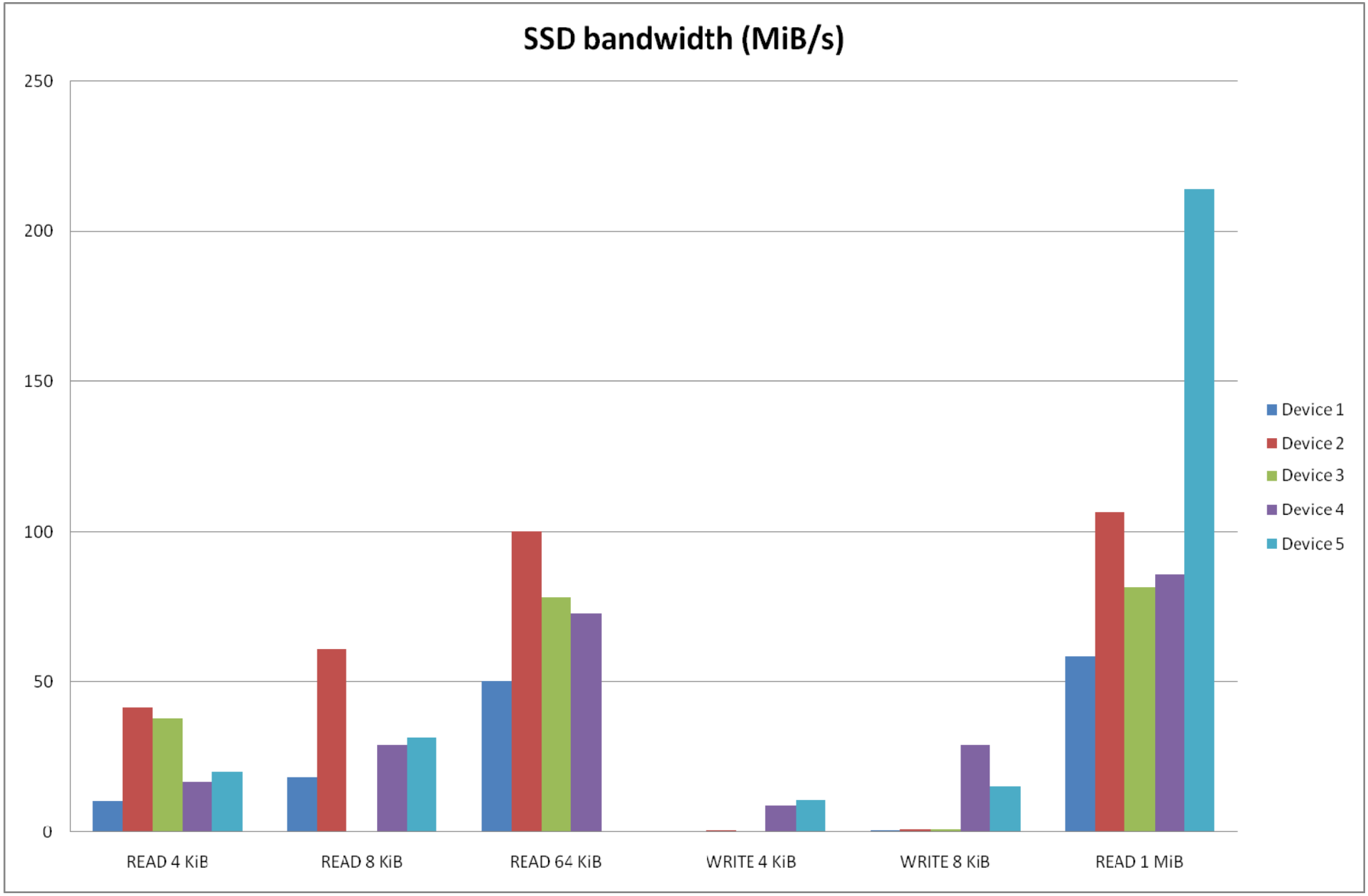
SSD (3/6)

CERN IT Department



SSD (4/6)

CERN IT Department



Service Level Agreement(SLA)
The intelligence layer

Tony Chan
Jason Smith
Mizuki Karasawa

March 24, 2009

Motivation

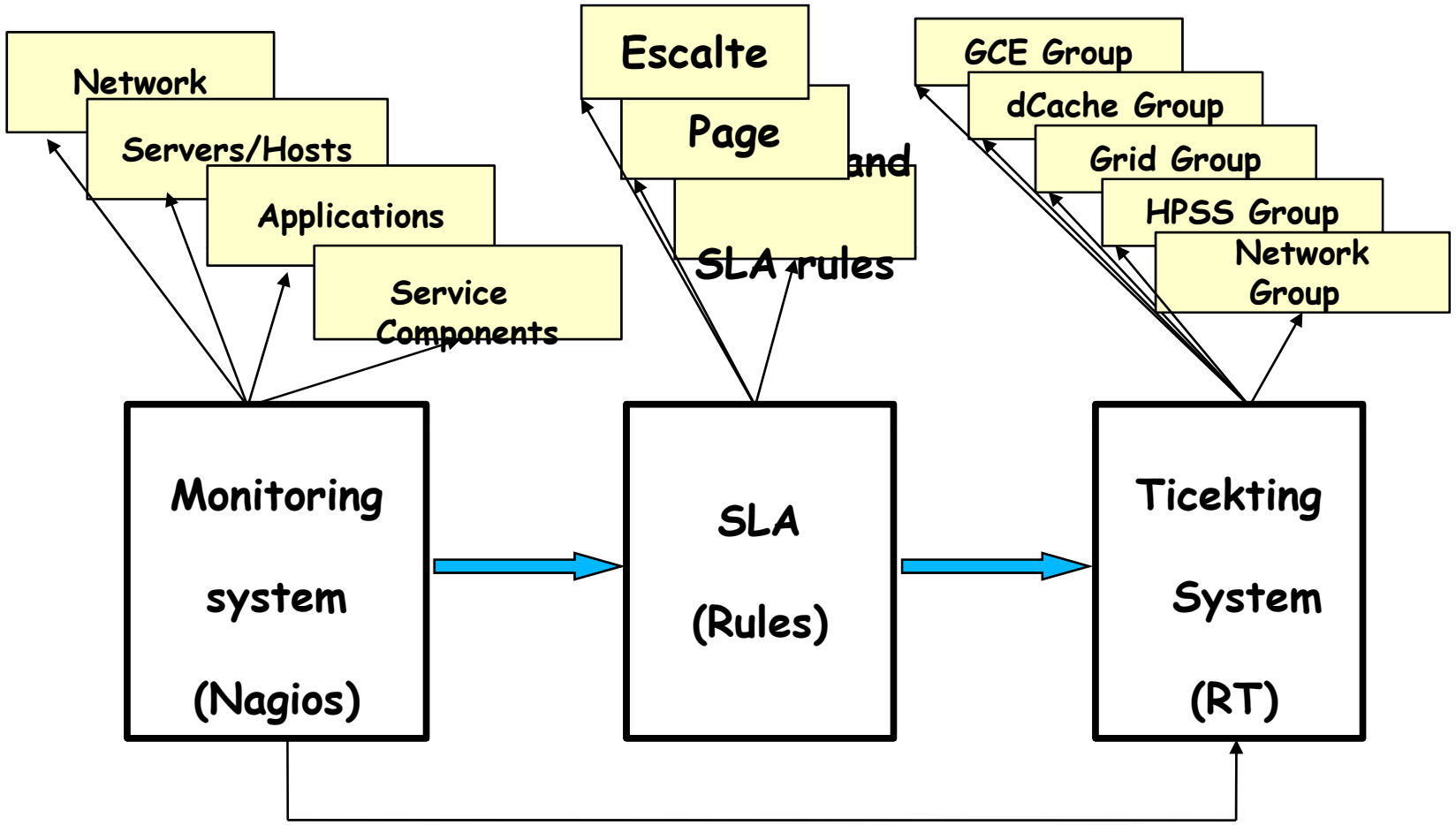
- The continue growth of the facility, the diverse needs of the scientific problem and increasingly prominent role of distributed computing requires RACF to change from a system-based to a service-based SLA with our user communities.

- SLA allows RACF to coordinate more efficiently the operation, maintenance and the development of the facility by creating a new, configurable alarm management that automates service alerts and notification of operations staff.

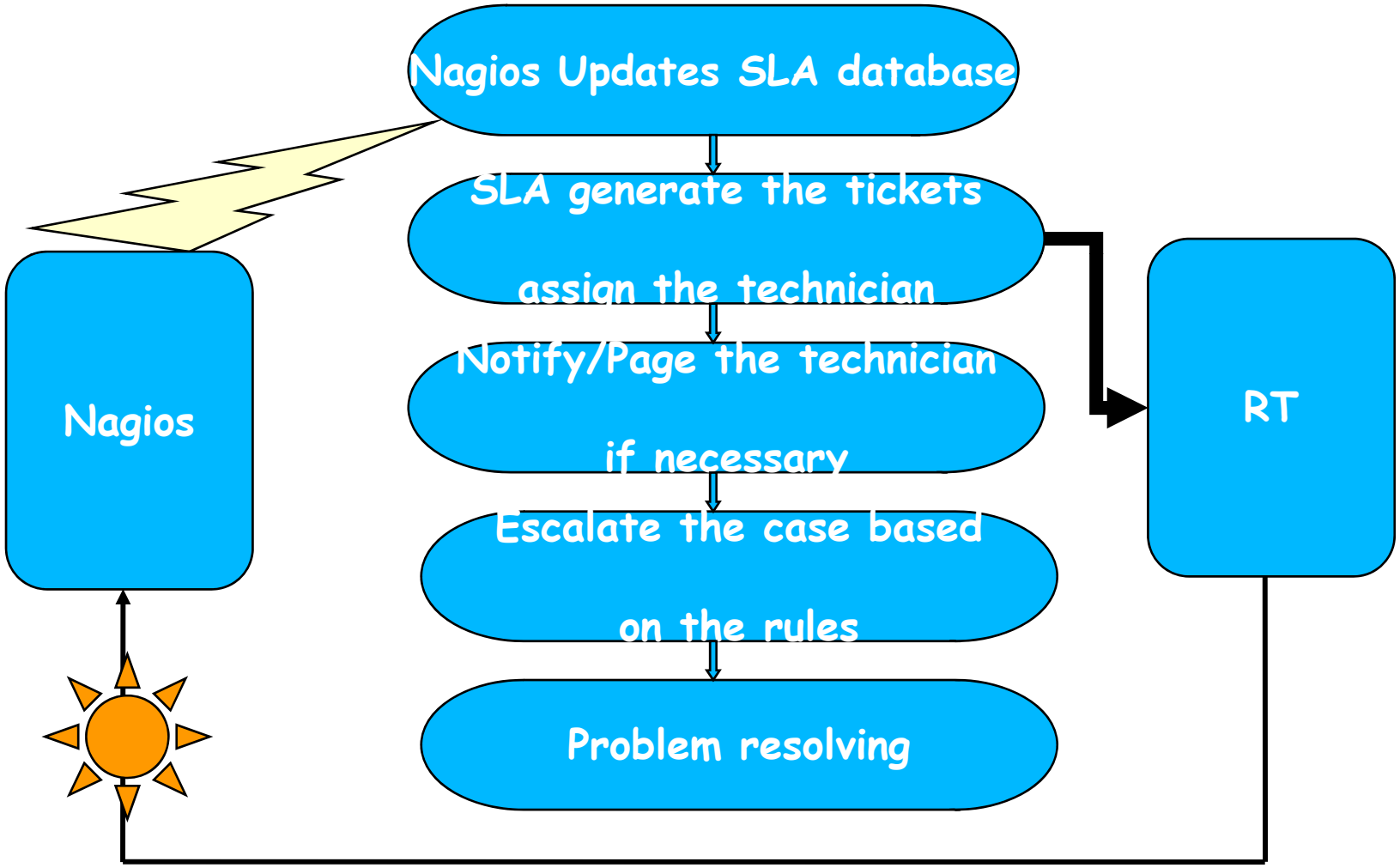
The SLA Concept

- The SLA records a common understanding about services, priorities, responsibilities, guarantees.
- Each area of service scope should have the 'level of service' define.
- The agreement relates to the service that users receives and how the service provider delivers that service.

New Architectures



Process Flow



Status update/Future plan

- Merge SLA to RT due to the close relationship between SLA & RT.
- Change the notification method from Nagios to SLA to avoid user misbehave. Reading directly from Nagios object cache to keep the consitancy and accuracy.
- Enhance the rule engine in order to deliver more efficient/informative alerts.
- Enhance the Web UI to give the visual outlook of the condition of the infrustrature.

233.pdf - Adobe Reader

ファイル(F) 編集(E) 表示(V) 文書(D) ツール(T) ウィンドウ(W) ヘルプ(H)

The Integration of Virtualization into the U.S. ATLAS Tier 1 Facility at Brookhaven

Christopher Hollowell <hollowec@bnl.gov>
RHIC/ATLAS Computing Facility (RACF)
Physics Department
Brookhaven National Laboratory

BROOKHAVEN

Virtualizaton at the RACF

- Running Xen 3.0.3, as shipped with RHEL5/SL5
- Used to split multicore hosts into individual virtual servers where OS segmentation is desirable or necessary
  - Allows for the most efficient use of increasingly prevalent multicore hardware
  - Specific operating system version requirements
  - Testbeds
  - Isolation of low and high security services
  - Reduction of resource contention (i.e. memory, disk space), and the impact of OS crashes

Virtualizaton at the RACF (Cont.)

- U.S. ATLAS Tier1 Processor Farm
  - 12 8-core physical machines paravirtualized into 40 servers: 2-3 guests + 1 Dom0 per host
  - Each physical system contains a single interactive virtual machine, and one or more batch/testbed host components
  - 32-bit SL5 Dom0 (control only), 32-bit SL4 DomUs
  - Physical CPUs pinned to guests
  - Networking via bridging, partitions for virtual disk devices
  - All interactive systems/submit hosts virtualized
    - Many interactive hosts desired for service redundancy
    - Current usage does not require more than 2 CPUs per host
    - Eliminates interactive vs. batch process contention for

Virtualizaton at the RACF (Cont.)

- Potential Future Use
  - Many of the RACF processor farm batch hosts are also dCache or xrootd/rootd storage pools. To increase reliability, it may be desirable to isolate the storage components from batch processing via virtual machines
  - Cloud computing?
  - Both would imply the deployment of Xen on the majority of our 1900-node processor farm systems. Challenges:
    - Number of available public IP addresses at BNL
    - Configuration/management
  - Live migration of VMs to minimize downtime for hardware maintenance/issues

NDGF Tier1 for WLCG: challenges and plans - Adobe Reader

ファイル(F) 編集(E) 表示(V) 文書(D) ツール(T) ウィンドウ(W) ヘルプ(H)

Analysis of internal NDGF Tier-1/2 network requirements

Josva Kleist, Ph.D.
Software Coordinator, NDGF

CHEP 2009

Prague, March 24th, 2009

NDGF
NORDIC DATAGRID FACILITY

NDGF Tier-1 Resource Centers

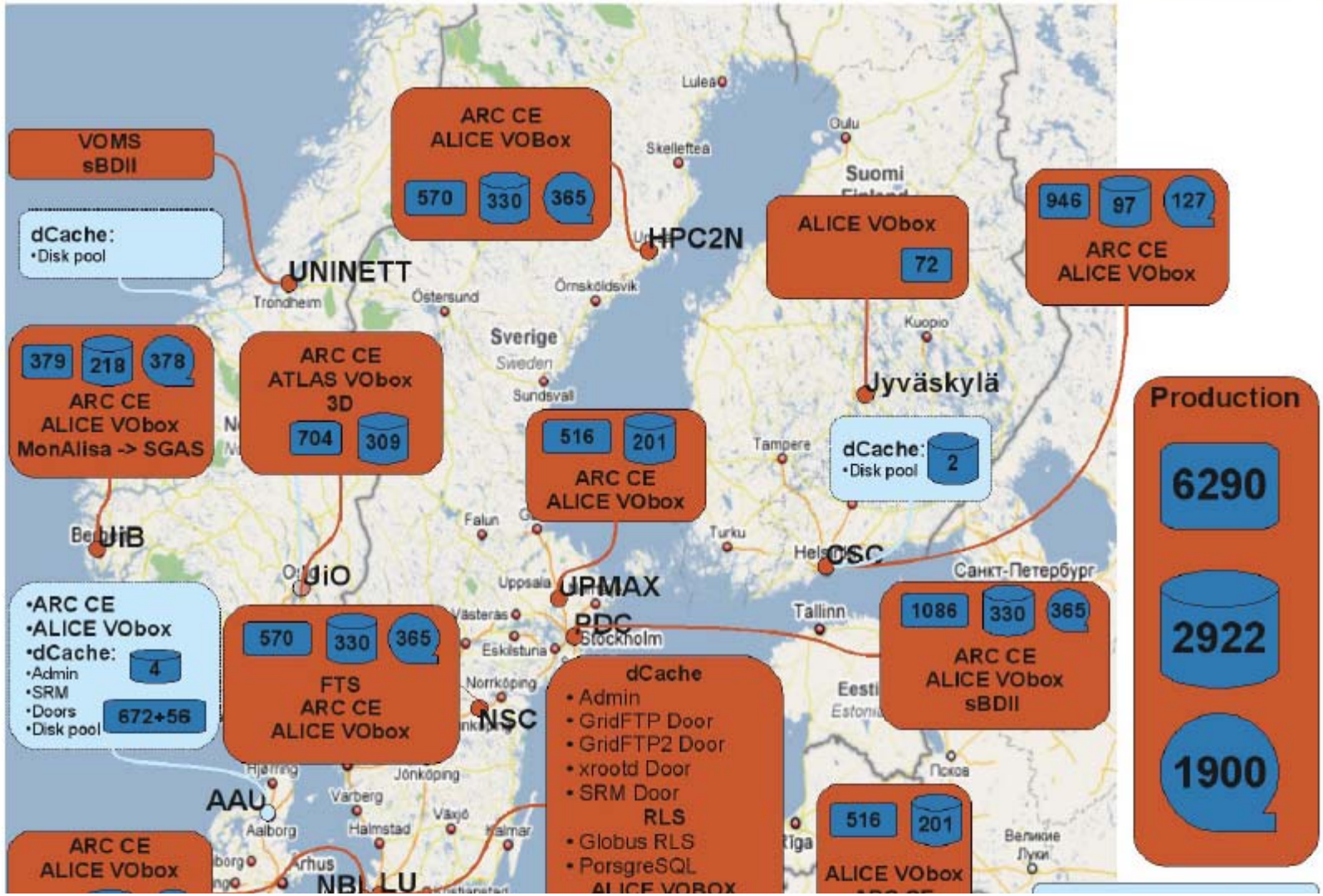
- The 7 biggest Nordic compute centers, dTier-1s, form the NDGF Tier-1
- Resources (Storage and Computing) are scattered
- Services can be centralized
- Advantages in redundancy
- Especially for 24x7 data taking

NDGF Tier1 for WLCG: challenges and plans - Adobe Reader

ファイル(F) 編集(E) 表示(V) 文書(D) ツール(T) ウィンドウ(W) ヘルプ(H)

NDGF
NORDIC DATAGRID FACILITY

Tier-1/2 overview



Summary and conclusion

- This track was very successful
  - Interesting papers
  - Many audiences
    - We needed larger rooms for the sessions
    - Less papers submitted is not necessary meant less audiences expected
- Thanks for speakers, contributors, chair persons and organizers

USB virus

- Somebody’s USB stick was influenced by USB virus
  - autorun.inf
- Scan you PC and USB sticks as soon as possible with the latest virus data if you have other person’s USB device in your PC this week