FZK Service Challenge Meeting
2/12/2004

Present:
Jos van Wezel(FZK), Jens Rehn (CERN/CMS), Cristopher Jung (FZK), Volker Buder (FZK), Kors Bos(NL), Doris Ressmann (FZK), Dennis Schieferdecker (Univ-KA), Ariel Garcia (Crossgrid/FZK), Lionel Schwarz (IN2P3), Laurent Caillat-Vallet (IN2P3), Thorsten Antoni (FZK), Andrew Sansum (RAL), Luca Dell'agnello (INFN), Stefano Zani (INFN), Les Robertson (CERN), Holger Marten(FZK), Klaus Peter-Mickel (FZK), Michael Ernst(DESY), Klaus Ullmann (DFN), Karin Schauerhammer (DFN), Bruno Hoeft(FZK)

Jos Van Wezel
Representing GridKa/FZK
Focus on interaction and discussion

Les Robertson – Phase II planning
Service Challenges are used to check the infrastructure and are build on the data and computing challenges. Planning group will look at important issues. One it tape storage. Access patterns to tape has not be looked at seriously.
Experiment Computing Challenges are used to check the experiment computing model
Data Challenges go on in the background on the existing grid infrastructure.

Experiments will have a copy of the raw data spread over the tier1s.
1.3GB/s aggregate data rate (from summer 2004 data) for Tier-0 to Tier-1
Have to be able to run at least twice this data rate due to problems at CERN or regional centre.

Series of Challenges to get us ready for full physics 6 months before startup
But we can do cosmics in early 2007.

Jos: What happens if aTier-1 goes down?

Les: **the tier-1 needs to plan for some extra capacity on tape – and the ability to write at twice the nominal rate through to tape.** Still not clear what data rates to tape that experiments have planned on. Ok for the streaming write – but what about ESD skimming reprocessing or Tier-2 support – this gives read rates which we don't know what they will be yet.

The challenges will help.

Klaus Ullmann: Is there redundancy planned?
Les: Don't know.
Klaus: Bandwidth is easy – how do we deal with long-term failures.
Thorsten Antoni: GridKa adds a second 10 Gb connect in 2008.
Bruno: Will cover it more in his talk

James Casey – CERN status
Jos: Shows need for separate setup – need to get this into the planning early.
Holger: anything know of the Fermi setup.
James: Fermi uses the CMS setup with 22 nodes
Jos: possibility to add a recommended list of hard and software to make tier 1's comparable
James: will look into it.

Bruno Hoeft – Progress at GridKa
Currently 10Gb via GEANT – testing
Going into production in 2005
In 2008 10g lambda direct to CERN. This would be for radiant traffic. Perhaps this will come a bit earlier.

Existing Geant 10g will be for Tier-2s and T1-T1 traffic

Management network overlayed- ganglia, nagios, cacti

10G testing – existing route via Italy – GEANT can't provide 10g via that route. They routed through France with a hardcoded LSP routing. Through France RTT of 19ms, via Italy 29 ms

Kors: Did INFN loose their 10Gb – why? since FZK go through France.
Luca: Not – they only have 1Gb- they will upgrade.

GEANT asks for FZK packets to be marked with LBEs – they are the first to be dropped!

Testbed
11 nodes – 8 x 1GB mem/ 1Gb nic/ IDE HDD, 3 x 2Gb, SCSI HDD
Local Storage – no RAID – LUNs exposed directly.

Iperf tests: 953Mbin to oplapro73, 884Mbit from oplapro73

Currently RENO at openlan, scalable TCP at FZK

Ramped up to 5Gbit iperf – GEANT noticed ☺ they seem the single stream TCP back-off , so go to many streams to smooth it out – not particularly efficient!

Test results : 2.5Gbit with gridftp over 11 nodes.

With the Tyan worker nodes, need to limit transfer. 4x 8 streams
On the IBM nodes 4x16

TCP Params - 16Max 8min

Future – GPFS SAN. Overcome the 25MB limit of single HD.

James: What should we do about alternate TCP stacks?
Jos: Do we need to sync on Kernel versions too?
James: We should provide recommendations.
Bruno: Good idea to provide a few nodes at the CERN end as pure test nodes- where we can change network params etc. on a higher freq.

Jules Wolfrat – SARA Status
Just got the network working yesterday.
Tested over GEANT line – poor results. Only got 100Mbit via iperf. Need to now look what we get on the dedicated links.

COFFEE BREAK

Killian Schwarz – gLite software for the Service Challenge
Summary of Den Haag meeting slides re: gLite middleware

Holger: What is the schedule and communication mechanism for the Alice DC3 to get the sites involved?
Killian: Currently just the Alice sites currently involved. They'll just install gLite on the Alien control nodes instead. In principle it starts now.

Doris Ressman - dCache implementation at FZK
Summary of dCache features. Implementation details at FZK
TSM used at FZK
15 POOL nodes added to dCache system at FZK for 10Gb tests.

Michael: Concerning dCache access. PRELOAD_LIBRARY to allow existing applications using dCache. Shown to work with CMS.
Concerning utilities: srmcp is not only client - lcg_utils too. GRIS provider for dCache.
DB Backend – also postgres impl as well as gdbm for PNFS backend.
Port range- no longer required to have an agreement for the port range. A java option exists that limits the port range. Used by SRM third-party copies.

Next dCache comes with the SRM that stores transfer state inside.

James: gLite DM have a cloudscape based version that removes the need for an external database.

James: SRM Copy provides the reliability layer – it will do the retries.

Michael: removing last single point of failure in dCache – PNFS database is he bottleneck. This will be removed by using a shared disk between two nodes with heartbeat and failover.
Special knowledge about cdap not needed. about dcap preload is available
Utilities: srmcp is not the only client. There is also vlgc utils. Additonally there is gris (in relation to the LCG SE)
Database based on gdbm but there exists a postgress implementation (FERMI)
A java client is available that allows you to set the ports
SRM database implementation is coming (transfer states are kept internal to the SRM implementation)
Failover of the writer node is in progress.

Andrew: What is reliability of underlying dCache infrastructure.
Doris: Pretty happy. Since it's a test setup, mostly the problems are with hardware – not production quality hardware yet.

LUNCH

James Casey – Demo
Jens: Monalisa was set up at FZK for CMS DC04 – it's not inactive, but can be set up again easily.
Jos: how easy to install client bundle for playing with?
James: Easy – but it stores passwords in the clear, and does direct SQL connections, so I prefer to limit it to a node at CERN right now – this will go away with the gLite clients.
There will be a monitoring link and access available (mrtg plots) wiki is available
Get cern account to join

Jens Rehn – Experiences with Data Movement via PHeDeX
CMS Model Data Rates:
Raw

- .6 +/- .3 MB per event (low lum)
- 1.5 +/- .5 MB per event (high lum)

RAW Prime

- Lossless compression

Reconstructed events (=DST)

Full Events

Phedex

- 100K file with 5 replicas each per month
- Reliable transfers
    - Checking file sizes and cksums
    - Multi-hop transfers with fallback routes
- Fulfil transfer needs
    - Push/pull and streaming models

    - Data subscriptions; metadata based on datasets
    - Web interface to requests and subscriptions
- Buffer space management
    - Cleaner to remove file
    - Stage pool management
- Monitoring
    - Status web page
    - Interface to monalisa
- Protocol matching
    - Multiple backends g-u-c, srmcp, dccp, lcg-rep
    - Automatic protocol matching

Data Movement GridKa import
Current
- Globus-url-copy
- Dccp to dCache interfaced MSS

Future
- Using srmcp
- Perhaps still with a dedicated buffer (so only reliably copied files make it into dCache)

Data Movement GridKa export
- Export from MSS via dCache
    - Auto buffer space management
    - No explicit cleaned needed
    - dCache interfaced with SRM
- Export from buffer disk
    - Only for intermediate transfers

Why SRM?
- Negotiates transfer protocol
- Checks available space
- Assume correct file transfers
- Initiates file staging (e.g. on dCache)

Transfers are sometimes unreliable. Check summing needed. Available in srmcp but Michael remarks the versions should agree on the algorithm to enable checksums.
Data rates = 2TB/day. Moved 40TB in total
See limitation at ~20MB/s per channel – both RAL and FZK – FZK has HTAR route? What is the limit due to this?

TMDB possible single point of failure – move to distributed TMDBs

Bruno: FZK has two lines – production (1 Gb) and test. Production moves to the 10Gb line soon.
Jos: How can they experiments use it?
James: We could set up some specific instances for experiments directly.
Jos: We also need to look at the levels of buffering inside a site in such a system – too many buffers happening.

Klaus Ullmann, X-Win and Geant2 – the next generation of research networks in Germany and Europe
Overview of networking both at the Germany NREN level (DFN) and the European level (Geant2).

Today, service provided by T-Systems. Finish end of 2005. 27 nodes around Germany

Future-
Technical Concept
- Inclusion of wavelength and dark fiber.
- Possibly more nodes

Protect investment in SDH-technology

Economic Concept

Re-define core network via packages of services
Make possible to establish more than one provider
Minimize risk for bidder

New model – split in four areas

Provisioning of dark fibre
Provisioning of wavelengths
Supervisor system
Support systems

Tendering

- Europe wide call in 2004
- No negotiation procedure
- Decision November 2004
- Start operation end 2005

Main results

- Most of the XWin core will be a fiber network. Rest with wavelengths
- Not what was expected !
- Fibre is cheap – in most cases (!) more economic than one wavelength – as opposed to DANTE where you need to run 3 or 4 lambdas before it is cheaper
- Future network now creates many new options – and is cheaper
- length of fibre is real length * 2

European vision

- Same vision true for Geant2
- Very likely be fibre from Geneva to Frankfurt
- This would appear ~2006

Site Messages

RAL – Andrew Samsun

Current site production Network is 2*1Gb.

Next TVN upgrade 2006 10Gbit/s

Test lightpath network (UKLIGHT)

Hardware on site
Connectivity in January
10Gbit possible if successful case made (finicial costs for UKERNA)

Internal Tier-1

Upgraded (3*stacks of Nortel 5510)
384ports/stack – 80 Git backplane
Upgrade to 10Gbit interconnect between stacks in 2005

Storage Infrastructure

Production deployment of dCache

1 head node
2 dcache pool nodes (gridftp portals)
NFS access to disk servers
CMS VO supported now, soon all VOs
Discussing back end to tape

Separate deployment of independent SRM interface to tape store

Hardware

200TB of disk
60 disk servers
120 external raid array (RAID5)
1500 spinning drives
STK powderhorn with 8 * 9940 drives
Running ADS
~2 PB available to 2007

Looking at tape to disc infrastructure right now

Performance
- Varies between hardware generations
- 60-120MB/s per array

INFN – Luca Dell'agnello

WAN
- Presently connected to GARR via 1Gbps links + 1 Gb (for tests)
- 10 GR GPoP collocated with Tier-1

LAN
- New core switch Extreme Black Diamond
- 4 x 10GE + 128 GE ports
- Additional summit 400 switch (2x 10GE +…)

Hardware
- 800 dual CPU proc (manlu Xeon)
- 200 TB disk (mainly SAN connected)
  - Access via NFS, rfio, gridftp
  - Started to deploy GPFS
- Currently testing other distributed file systems (PVFS, Lustre)
- STK L5500 library with 6 LTO2 drives
  - MSS access via CASTOR
- For the SC, tenders for 10 Gb in preparation

For SC, would look at SAN storage – probably running CASTOR gridftp/srm.

TODO Q; SL3 for opteron – experiences ???

IN2P3 – Lionel Schwarz

Network
- 1Gb link to REANTER

Hardware
- 2 nodes
- done ram to ram and disk to disk from Lyon to CERN: 70 MB/s

Software
- Globus gridftp
- SLC3
- 70MB/s disk to disk via gridftp

TODO
- ASAP
  - Increase # nodes and amount of disk
- Q105
  - SRM setup (dCache)/ HPSS
  - SRM to SRM transfers
- Mar05
  - SC Meeting Lyon
- Q305
  - 10 Gb link provisioned 2nd half 2005

Micheal: What is status of HPSS interface to dCache?
Lionel: It's working – need to do some production tests, since not enough nodes connected yet.
Holger: Why dCache and not your SRM to HPSS?
Lionel: Had problems with Berkeley SRM – wasn't easy to adapt to Lyon situation, due to different ways of accessing HPSS.

Next Meeting:
RAL 27th/28th
Agenda
- LHCC Computing model reviews

Network meeting
Milestone document

Discussion
Kors: Milestones document. When do we think we can do the real challenges – i.e. 5 sites to tape @ 300MB? July 05?

Les: It's important to try and get a period where we have resources from several sites to try and find bottlenecks- especially through to tape. Sites should estimate what percentage of the service they can provide.

Important thing is to keep it running day and night and see what problems occur.

Which 5 sites for March : FZK, CNAF, Fermi, SARA, RAL

Kors: Try and keep things going all the time at a lower data rate after the challenges.

Bruno: Starting slow and increasing bit by bit is easier. But also the high-speed challenges are useful. And we need "test test" infrastructure.

Jos: Try the Tier-1 to Tier-1 traffic soon.

James: Not sure what the Tier-1 to Tier-1 traffic is.

Kors: Also the Tier-1 to Tier-2 is interesting. We need to also think about the reprocessing at the Tier-1s as well. When can we test reprocessing at a Tier-1 while writing the raw to tape, and sending the ESD to tape and another Tier-1. This will be challenging.

Jos: Data is ok – we roughly know the issues. What about the cluster to the other tiers.

Kors: We do know – we know the rate of pre-processing. Model is known.

We need to create a model for re-processing and then test it in 2005.

Les: We need to get the draft out before next GDB. Discuss during January and try and decide.

Bruno: Go for short periods of high rate and longer low background.

Kors: Experiments will be doing transfers anyway – should just integrate them into a single service. Candidates are CMS and ATLAS.

Bruno: At CERN need 'service' for background traffic, setup to do service challenges on the side and some test nodes for tuning. And we need it early 2005.