10Gbit between GridKa and openlab and their obstacles

Forschungszentrum Karlsruhe GmbH
Institute for Scientific Computing
P.O. Box 3640
D-76021 Karlsruhe, Germany

http://www.gridka.de

Bruno Hoeft
Outline

• LAN of GridKa
  – Structure
  – Prevision of installation in 2008

• WAN
  – History (1G/2003)
  – Current 10G
  – Testbed
  – Challenges crossing multi NREN
  – Quality and quantity evaluation of network between GridKa and openlap
  – File transfer
Projects at GridKa

1,5 Mio. Jobs and 4,2 Mio. hours calculation in 2004

calculating jobs with “real” data

LHC Experimente

non-LHC Experimente

Bruno Hoeft

# gradual extension of GridKa resources

<table>
<thead>
<tr>
<th></th>
<th>Apr 2004</th>
<th>Okt 2004</th>
<th>Apr 2005</th>
<th>% of 2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processors</td>
<td>680</td>
<td>1070</td>
<td>1.280</td>
<td>30 %</td>
</tr>
<tr>
<td>Computing power / kSI2k</td>
<td>580</td>
<td>920</td>
<td>1290</td>
<td>12 %</td>
</tr>
<tr>
<td>Disk [TB]</td>
<td>160</td>
<td>220</td>
<td>270</td>
<td>18 %</td>
</tr>
<tr>
<td>Tape [TB]</td>
<td>280</td>
<td>375</td>
<td>475</td>
<td>12 %</td>
</tr>
<tr>
<td>Internet [Gb/s]</td>
<td>2</td>
<td>10*</td>
<td>10</td>
<td>50 %</td>
</tr>
</tbody>
</table>

* Internet connection for sc

**April 2005:**
- biggest Linux-Cluster within the German Science Society
- largest Online-Storage at a single installation in Germany
- strongest Internet connection in Germany
- available at the Grid with over 100 installations in Europe
Network installation

DFN

router

1 Gbit/s
GridKa

PIX

Storage
direct
internet access

router

private network

switch

.....

ComputeNode
ComputeNode
ComputeNode

.....

FS
NAS

.....

FS
NAS

.....

FS

.....

FS

.....

Ethernet 320 Mbit
 Ethernet 100 Mbit
 Ethernet 1 Gbit
 FiberChannel 2Gbit
 Ethernet 10 Gbit
Network installation

incl. Management Network

- Ganglia
- Nagios
- Cacti

Ethernet 320 Mbit
Ethernet 100 Mbit
Ethernet 1 Gbit
FiberChannel 2Gbit
Ethernet 10 Gbit
Projection of installation in 2008

- **2005**
  - Completed end 2005
  - Block A: CN - Rack, FS

- **2006**
  - Block B: CN - Rack, FS

- **2007**
  - Block C: CN - Rack, FS

- **2008**
  - Block D: CN - Rack, FS

- 10 Gbit Internet
- 10 Gbit
- 10 Gbit
- 10 Gbit
- Backbone Router

- WorkerNodes, FileServer, DiskStorage

Bruno Hoeff

GridKa – WAN connectivity planning

- 10G internet connection since Oct 2004
- Mid to end of 2005 → production
- X-Win partner of DFN
- “direct“ light path between CERN and GridKa via X-Win and Geant2
WAN connection and Gigabit test with CERN

Bruno Hoeft

Géant

DFN

Frankfurt

Karlsruhe

Geneva

10 Gbps

2.4 Gbps

2x 1 Gbps

GridFTP tested over 1 Gbps

GridFTP tested over 1 Gbps

GridFTP server
WAN connection
10Gigabit test with CERN

DFN
10 Gbps
Frankfurt
Karlsruhe

Géant
10 Gbps
Geneva

CERN GridFTP server

10 Gbps
swiCE3-G4-3.switch.ch
swiCE2-P6-1.switch.ch
it.ch1.ch.geant.net
de.it1.it.geant.net
dfn.de1.de.geant.net
cr-karlsruhe1-po12-0.g-win.dfn.de
ar-karlsruhe1-ge5-2-700.g-win.dfn.de
r-internet.fzk.de

GridFTP server
WAN connection
10Gigabit test with CERN

DFN
10 Gbps
Frankfurt

Géant
10 Gbps
Geneva

Karlsruhe
10 Gbps

LSP Routing
swiCE3-G4-3.switch.ch
fr.ch1.ch.geant.net
de.fr1.fr.geant.net
dfn.de1.de.geant.net
cr-karlsruhe1-po12-0.g-win.dfn.de
ar-karlsruhe1-ge5-2-700.g-win.dfn.de
r-internet.fzk.de

GridFTP server
CERN

GridFTP server
WAN connection
10Gigabit test with CERN

- Bandwidth evaluation (tcp/udp)
- MPLS via France
- LBE
- GridFTP server pool HD to HD
  Storage to Storage
- SRM
Hardware

- Various Xeon dual 2.8 and 3.0 GHz IBM x-series (Intel and Broadcom NIC)
- Recently added 3.0 GHz EM64T (800 FSB)
- Cisco 6509 with 4 10 Gb ports and lots of 1 Gb
- Storage: Datadirect 2A8500 with 16 TB
- Linux RH ES 3.0 (U2 and U3), GPFS
- 10 GE Link to GEANT via DFN (least best effort 😞)
  TCP/IP stack
  - 4 MB buffer
  - 2 MB window size
Quality Evaluation

scalable

LAN

W A N

953 Mbit; jitter: 0.018 ms; ooo: 215 – 0.008%

884 Mbit; jitter: 0.015 ms; ooo: 4239 – 0.018%

W A N

WAN

[root@10gtk111 root]# iperf -s -u -f m -i 1
[root@10gtk113 root]# iperf -c 192.108.46.111 -u -b 1000M -f m -i 1 -t 10
[ ID] Interval Transfer Bandwidth Jitter Lost/Total Datagrams
[ 3 ] 0.0-10.0 sec 1141 MBytes 957 Mbits/sec 0.028 ms 0/813885 (0%)

[ oplapro73] ~ > iperf -s -u -f m -i 1
[root@10gtk111 root]# iperf -c 192.108.46.113 -u -b 1000M -f m -i 1 -t 30
[ ID] Interval Transfer Bandwidth Jitter Lost/Total Datagrams
[ 3 ] 0.0-30.0 sec 3422 MBytes 884 Mbits/sec 0.021 ms 0/2440780 (0%)

[lan]scalable

WAN

[ oplapro73] ~ > iperf -s -u -f m -i 1
[root@10gtk113 root]# iperf -c 192.108.46.113 -u -b 1000M -f m -i 1 -t 30
[ ID] Interval Transfer Bandwidth Jitter Lost/Total Datagrams
[ 3 ] 0.0-30.0 sec 3408 MBytes 953 Mbits/sec 0.019 ms 7299/2438606 (0.3%)
[ 3 ] 0.0-30.0 sec 215 datagrams received out-of-order

LAN

WAN

[lan]scalable

[ oplapro73] ~ > iperf -s -u -f m -i 1
[root@10gtk113 root]# iperf -c 192.108.46.113 -u -b 1000M -f m -i 1 -t 30
[ ID] Interval Transfer Bandwidth Jitter Lost/Total Datagrams
[ 3 ] 0.0-30.0 sec 3162 MBytes 884 Mbits/sec 0.015 ms 130855/2386375 (5.5%)
[ 3 ] 0.0-30.0 sec 4239 datagrams received out-of-order

ooo - out of order
Bandwidth evaluation

Reno

W A N 348Mbit/sec – single stream W A N

iperf

Reno

Bruno Hoeft
Bandwidth evaluation

Reno

W A N 5 nodes a 112MByte/sec – 24 parallel streams

iperf

Oplapro7[1-5]

Reno

Bruno Hoeft
Gridftp sc1 throughput

Sc1 – 500MByte sustained

19 Nodes
- 15 WorkerNodes
- 1 FileServer
- 3 FileServer

* 20MByte IDE/ATA HD
* 50MByte SCSI HD
* 50MByte SAN
Evaluation of max throughput

9 nodes each site
-8 * 845 Mbit
-1 * 540 Mbit
higher speed at the one stream is resulting in a packet loss
Future Work

- Gpfs for data dest.
- Digging into HW details to discover bottlenecks (packet drop due to bad PCI timing)
- Stabilise the transport
- Installation of SRM
- Installation of dcache

Long Term

- Lightpath via X-Win (DFN) / Geant2 to CERN
thanks for your attention

Questions