Research on Networks

Cees de Laat
University of Amsterdam
Contents

• StarPlane & Tests
• Fault Isolation
• Malan dark fiber project
• Power considerations
• SC06 activities
Contents

• StarPlane & Tests (HB, LX, JV, CdL, PG)
• Fault Isolation
• Malan dark fiber project
• Power considerations
• SC06 activities
StarPlane
DWDM
backplane
QOS in a non destructive way!

• Old QOS:
  – have a link or λ
  – set part of it aside for a lucky few under higher priority
  – rest gets less service

• New QOS:
  – have a λ
  – add other λ‘s as needed on separate colors
  – move the lucky ones over there
  – rest gets also a bit happier!
DAS-3 Cluster Tender
http://www.clustervision.com/pr_das3_uk.html

85 compute nodes

Fast interconnect

Local interconnect

To SURFnet

10 Gbit/s Ethernet lanphy

To local University

UvA-node

10 Gbit/s Ethernet lanphy

1 Gbit/s Ethernet
### Heterogeneous clusters

<table>
<thead>
<tr>
<th></th>
<th>LU</th>
<th>TUD</th>
<th>UvA-VLE</th>
<th>UvA-MN</th>
<th>VU</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Head</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>* storage</td>
<td>10TB</td>
<td>5TB</td>
<td>2TB</td>
<td>2TB</td>
<td>10TB</td>
<td>29TB</td>
</tr>
<tr>
<td>* CPU</td>
<td>2x2.4GHz DC</td>
<td>2x2.4GHz DC</td>
<td>2x2.2GHz DC</td>
<td>2x2.2GHz DC</td>
<td>2x2.4GHz DC</td>
<td>46.4 GHz</td>
</tr>
<tr>
<td>* memory</td>
<td>16GB</td>
<td>16GB</td>
<td>8GB</td>
<td>16GB</td>
<td>8GB</td>
<td>64GB</td>
</tr>
<tr>
<td>* Myri 10G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>40 Gb/s</td>
</tr>
<tr>
<td>* 10GE</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>50 Gb/s</td>
</tr>
<tr>
<td><strong>Compute</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>* storage</td>
<td>400GB</td>
<td>250GB</td>
<td>40 (+1)</td>
<td>46</td>
<td>85</td>
<td>271</td>
</tr>
<tr>
<td>* CPU</td>
<td>2x2.6GHz DC</td>
<td>2x2.4GHz DC</td>
<td>2x2.2GHz DC</td>
<td>2x2.4GHz DC</td>
<td>2x2.4GHz DC</td>
<td>1.9 THz</td>
</tr>
<tr>
<td>* memory</td>
<td>4GB</td>
<td>4GB</td>
<td>4GB</td>
<td>4GB</td>
<td>4GB</td>
<td>1048 GB</td>
</tr>
<tr>
<td>* Myri 10G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2030 Gb/s</td>
</tr>
<tr>
<td><strong>Myrinet</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>* 10G ports</td>
<td>33 (7)</td>
<td></td>
<td>41</td>
<td>47</td>
<td>86 (2)</td>
<td>2070 Gb/s</td>
</tr>
<tr>
<td>* 10GE ports</td>
<td>8</td>
<td></td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>320 Gb/s</td>
</tr>
<tr>
<td><strong>Nortel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>* 1GE ports</td>
<td>32 (16)</td>
<td>136 (8)</td>
<td>40 (8)</td>
<td>46 (2)</td>
<td>85 (11)</td>
<td>339 Gb/s</td>
</tr>
<tr>
<td>* 10GE ports</td>
<td>1 (1)</td>
<td>9 (3)</td>
<td>2</td>
<td>2</td>
<td>1 (1)</td>
<td></td>
</tr>
</tbody>
</table>
Status DAS-3

• clusters delivered end of august
  – housed @ UvA IvI building
  – 48 fiber pair cable to SURFnet6 rack @ SARA
  – in acceptance testing
  – Myrinet 10 G Ethernet cards expected in few weeks
  – 14 * 10 G NIC’s delivered for initial testing/bridging
Dispersion compensating modem: eDCO from NORTEL
(Try to Google eDCO :-)

Solution in 5 easy steps for dummy’s:
1. try to figure out $T(f)$ by trial and error
2. invert $T(f) \rightarrow T^{-1}(f)$
3. computationally multiply $T^{-1}(f)$ with Fourier transform of bit pattern to send
4. inverse Fourier transform the result from frequency to time space
5. modulate laser with resulting $h'(t) = F^{-1}(F(h(t)).T^{-1}(f))$

(ps. due to power \~ square $E$ the signal to send looks like uncompensated received but is not)
Preliminary DAS-3 test

- De metingen zijn in dit geval gedaan Iperf met drie node-paren met een applicatie-buffer size van 16 KB.
- Langs de X-as is de totale TCP window size / (# node paren) uitgezet en langs de Y-as het aantal parallele flows per node paar.
Other tests

• SURFnet6 tests
  – http://trafficlight.uva.netherlight.nl/SURFnet6Tests/
  – tests with CWI:
    • http://trafficlight.uva.netherlight.nl/SURFnet6Tests/cwi_direct_link/toc/index.html
  – Autotune TCP stack
    • http://trafficlight.uva.netherlight.nl/SURFnet6Tests/cwi_direct_link/scen_01/tcp/index.html#CPU-Aff-Auto-Tune-Sect
  – related UDP tests
    • http://trafficlight.uva.netherlight.nl/SURFnet6Tests/cwi_direct_link/scen_01/udp/perf_results/cpu_aff/index.html
  – reorder tests on Rembrandt
    • http://trafficlight.uva.netherlight.nl/SURFnet6Tests/cwi_direct_link/scen_01/udp/conclusions/index.html
Contents

• StarPlane & Tests
• Fault Isolation (NR, FD, CdL, PG)
• Malan dark fiber project
• Power considerations
• SC06 activities
Expert System: Recap

- Gather data from devices along a connection, and try to pinpoint the fault. Passive monitoring
- Two ways to think about it:
  - A connection does not work, what is wrong?
  - Predict with the available data if a link works or not.
- Current lack of tools to retrieve monitoring information in a uniform format. Takes an effort to write software for each device.
- Unclear if it is possible to cope with missing data.
Expert System Design

- Minor changes to the design:
  - Now uses NDL
  - Added client-server interaction
Expert System: Progress

- Now uses NDL
  - Easier to extend with new properties (multilayer information, device information)
  - Created NDL extension to describe device configuration (see talk Bert)
- On demand monitoring
  - Added support for Glimmerglass; Force10, Nortel switch and Calient in progress
  - No constant polling, but only when needed
- Added client-server interaction
  - Still in plain text; may be in XML in the future (webservice)
Expert System: Planning

- Detection and Isolation
  - Use top-down isolation: if a connection is correct at layer N, then it is not needed to check all details of layer N-1 (maybe only verify the path)

- Planning
  - First demo showing the expert system
  - Extension to NDL for device specific information (brand, model, so that the monitor knows how to gather data)
  - Extension to multilayer problem, using the top-down isolation (for now it just gathers all data)
Contents

• StarPlane & Tests
• Fault Isolation
• Malan dark fiber project (KK, JV, JvG, CdL, PG)
• Power considerations
• SC06 activities
MALAN

- Metropolitan Area LAN
- Between
  - SNE Lab (HvA / Amstel Station)
  - Lighthouse (SARA / Science Park)
- Using Eurofiber dark fiber
Example student projects

- Netbooting (Mac OS X; Linux)
- BW-intensive applications
- Distributed filesystem
- iSCSI
- IDS’ses
- University – ISP interface
- extend hybrid networking in the enterprise
Contents

• StarPlane & Tests
• Fault Isolation
• Malan dark fiber project
• Power considerations (CdL)
• SC06 activities
Power is a big issue

• UvA cluster uses (max) 30 kWh
• 1 kWh ~ 0.1 €
• per year -> 26 k€/y
• add cooling 50% -> 39 k€/y
• Emergency power system -> 50 k€/y
• per rack 10 kWh is now normal
• YOU BURN ABOUT HALF THE CLUSTER OVER ITS LIFETIME!
Power outages are a big problem

• on average about one outage per year
  – once the generator not starting/taking over
    • -> batteries
  – this summer weekend explosion of cable
    • -> generator fine!
• battery power for 5 minutes, generator to take over
• priorities for emergency power/cooling
• asked to shutdown compute nodes if temp rises
VM opportunity

- Head node
- Memory rich VM holder
- CPU nodes
- Switchable power strip
- Network
Contents

- StarPlane & Tests
- Fault Isolation
- Malan dark fiber project
- Power considerations
- SC06 activities (LX, JvdH, FW, MH, JP, LG, PG, CdL)
Activities @ SC06

• Accepted poster on NDL by JvdH et al.
• Booth poster + slideshow on StarPlane by CdL
• Booth poster + demo, slides on NDL by Jeroen et al.
• Booth poster + demo, slides on TBN by Leon et al.
• Booth web page plus WebCam’s by CdL
• SciNet routerheads team JP
• support demo NORTEL VM on Gloriad
• support demo “optical” multicast Joe Mambretti
• participation demo data mining IDS logs of Bob Grossman
The Dutch Booth #1805 at SC 2006, nov 13 - 16 2006 live! (made by C.T. de Lant)

This page was live during sc06, now archived, see us at sc07 in Reno!

Click on one of the windows to enlarge that view!

SC2006 demonstrators in the Dutch Booth

visit the TRIN expert homepage
visit the SARA TOPS project page
visit NIX for the GLIF page
visit System & Engineering @ UvA
visit StarPlane.org
visit Personal Space Station demo

Token-based GMPLS
Path Authorization and Resource Management by extending RSVP-TE with Tokens

Network Description Language
Semantics for Hybrid Networks

NLX

NORTEL

GigaPort

SURF.net

Universiteit van Amsterdam

Myriyce

StarPlane

NWO

NCF

SC2006 UvA Posters @ Dutch booth (click on poster to download pdf)
• FGCS special issue on sc2005

• # Jeroen van der Ham, Paola Grosso, Ronald van der Pol, Andree Toonk, Cees de Laat, "Using the Network Description Language in Optical Networks", Tenth IFIP/IEEE International Symposium on Integrated Management (IM 2007), 21-25 May 2007, Munich, Germany. Accepted paper -> still to be published

• # Yuri Demchenko, Leon Gommans, Cees de Laat, Rene van Buuren. "Domain Based Access Control Model for Distributed Collaborative Applications", Accepted paper, The 2nd IEEE International Conference on e-Science and Grid Computing, December 4-6, 2006, Amsterdam. -> Accepted paper -> still to be published

Questions ?