SNR-2006: Very Dynamic LightPath Applications in DAS3 & StarPlane.

Cees de Laat

SURFnet

BSIK

EU

University of Amsterdam
History - 1

DAS = Distributed ASCI Supercomputer

• Project DAS-1 started in 1997 by Andrew Tanenbaum
• To prove distributed clusters were as effective as super...
• 4-5 clusters connected via high speed links
  – DAS-1 -> 6 Mbit/s full mesh ATM
  – DAS-2 -> Gbit/s L3
  – DAS-3 -> StarPlane
• DAS-1 ran BSD, changed to Linux (Andrew... :-)  
• DAS-1 and 2 uniform architecture, not so in DAS-3
• Over 200 users, 25 Ph.D. theses
• http://www.cs.vu.nl/das/
Examples cluster computing

- Communication protocols for Myrinet
- Parallel languages (Orca, Spar)
- Parallel applications
  - PILE: Parallel image processing
  - HIRLAM: Weather forecasting
  - Solving Awari (3500-year old game)
- GRAPE: N-body simulation hardware
Distributed supercomputing on DAS

- Parallel processing on multiple clusters
- Study non-trivially parallel applications
- Exploit hierarchical structure for locality optimizations
  - latency hiding, message combining, etc.
- Successful for many applications
  - E.g. Jem3D in ProActive [F. Huet, SC’04]
Example projects

- **Albatross**
  - Optimize algorithms for wide area execution

- **MagPle:**
  - MPI collective communication for WANs

- **Manta: distributed supercomputing in Java**

- **Dynamite: MPI checkpointing & migration**

- **ProActive (INRIA)**

- **Co-allocation/scheduling in multi-clusters**

- **Ensflow**
  - Stochastic ocean flow model
Grid & P2P computing: using DAS-2 as part of larger heterogeneous grids

- Ibis: Java-centric grid computing
- Satin: divide-and-conquer on grids
- Zorilla: P2P distributed supercomputing
- KOALA: co-allocation of grid resources
- Globule: P2P system with adaptive replication
- CrossGrid: interactive simulation and visualization of a biomedical system
DAS 1 - 2 Cluster

- Head node
- Fast interconnect
- Local interconnect
- Ethernet
- 32 compute nodes

To local University and wide area interconnect
SURFnet6 Architecture discussions 2001-2002

- photonic backbone
- (L2 and) L3 services
- NORTEL
- Static
- Summer 2004 K&C
- NWO-GLANCE
- StarPlane
- PHD-PD-SP
- Start 1-feb-06, Li Xu, Jan Philip Velders, Jason Maasen
  - Henri Bal, Paola Grosso, Herbert Bos, CdL, SN-folks.

ref: cdl-2002-01-18-UCL-opt.ppt
StarPlane Approach

- StarPlane is a NWO funded project with major contributions from SURFnet and NORTEL.

- The vision is to allow part of the photonic network infrastructure of SURFnet6 to be manipulated by Grid applications to optimize the performance of specific e-Science applications.

- StarPlane will use the physical infrastructure provided by SURFnet6 and the distributed supercomputer DAS-3.

- The novelty: to give flexibility directly to the applications by allowing them to choose the logical topology in real time, ultimately with subsecond lambda switching times.
GRID-Colocation problem space

- Extensively under research
- New!
In The Netherlands SURFnet connects between 180:
- universities;
- academic hospitals;
- most polytechnics;
- research centers.
with a user base of ~750K users

> 6000 km comparable to railway system
Common Photonic Layer (CPL) in SURFnet6 supports up to 72 Lambda’s of 10 G each 40 G soon.
DAS-3 Cluster Tender
http://www.clustervision.com/pr_das3_uk.html

32 compute nodes

Fast interconnect

10 Gbit/s Ethernet lanphy

To SURFnet

1 Gbit/s Ethernet

Local interconnect

To local University

head node
# Heterogeneous clusters

<table>
<thead>
<tr>
<th></th>
<th>LU</th>
<th>TUD</th>
<th>UvA</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></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></td>
</tr>
<tr>
<td>* 10GE</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></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>250GB</td>
<td>2x250GB</td>
<td>250GB</td>
<td>84 TB</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></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>41</td>
<td>47</td>
<td>86 (2)</td>
<td></td>
<td>320 Gb/s</td>
</tr>
<tr>
<td>* 10GE ports</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td></td>
<td></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>
Photonics
Module Operation

> this schematic shows
  • several input fibres and one output fibre
  • light is focused and diffracted such that each channel lands on a different MEMS mirror
  • the MEMS mirror is electronically controlled to tilt the reflecting surface
  • the angle of tilt directs the light to the correct port

> in this example:
  • channel 1 is coming in on port 1 (shown in red)
  • when it hits the MEMS mirror the mirror is tilted to direct this channel from port 1 to the common
  • only port 1 satisfies this angle, therefore all other ports are blocked

ref Eric Bernier, NORTEL
What makes StarPlane possible

- Wavelength Selective Switches
- Sandbox by confining StarPlane to a band
- Optimization of the controls to turn on/off a Lambda
- electronic Dynamically Compensating Optics (eDCO)
- traffic engineering
Traffic engineering
What do we need

- vlan’s
- trunking
- spanning tree modified?
- mac in mac?
- source routing modified
- Policy interfaces
- AAA interaction (EduRoam, Shibboleth)
StarPlane applications

- Large ‘stand-alone’ file transfers
  - User-driven file transfers
  - Nightly backups
  - Transfer of medical data files (MRI)

- Large file (speedier) Stage-in/Stage-out
  - MEG modeling (Magnetoencephalography)
  - Analysis of video data

- Application with static bandwidth requirements
  - Distributed game-tree search
  - Remote data access for analysis of video data
  - Remote visualization

- Applications with dynamic bandwidth requirements
  - Remote data access for MEG modeling
  - SCARI
Application - Network interaction

App 1

App 2

App 3
Workflow based App.

App 1

App 2

App 3
Risks

What have we today
What to avoid
Three Easy Steps:

Click the START button

Insert money...

$0.25 per minute...
Example:
$1 = 4 minutes
$5 = 20 minutes
No change is provided!

Surf the web!
Check your email here!
Conclusions

• We try to go for fast (subsecond) Lambda setup and teardown, that is different from most other initiatives

• We need to work on GMPLS, SOA, webservice, RDF, supporting tools to make this happen

• We need to stress the current control loops and procedures to get there

• Workflow systems and/or applications need to become network aware.
Questions?

Credits: some slides from Paola Grosso or Henri Bal