CineGrid  GRID & Networking

Cees de Laat

University of Amsterdam

With grid slides thanks to David Groep (NIKHEF)
CineGrid Mission

To build an interdisciplinary community that is focused on the research, development, and demonstration of networked collaborative tools to enable the production, use and exchange of very-high-quality digital media over photonic networks.

http://www.cinegrid.org/
The Grid

- Grid ‘coined’ in 1997 by
  - Carl Kesselman (ISI/USC) and
  - Ian Foster (ANL)

- builds on a tradition of distributed computing
  - 1969: Creaper & Reaper
  - 1978: RPC concept
  - 1985: Condor
  - 1991: CORBA
  - 1991: DCE/DFS
Exploding Data

Collected data in science (and industry) grows exponentially:

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Bible</td>
<td>5 MByte</td>
</tr>
<tr>
<td>X-ray image</td>
<td>5 MByte/day</td>
</tr>
<tr>
<td>Functional MRI</td>
<td>1 GByte/day</td>
</tr>
<tr>
<td>Bio-informatics databases</td>
<td>500 GByte each</td>
</tr>
<tr>
<td>Refereed journal papers</td>
<td>1 TByte/yr</td>
</tr>
<tr>
<td>Satellite world imagery</td>
<td>5 TByte/yr</td>
</tr>
<tr>
<td>US LoC contents</td>
<td>20 TByte</td>
</tr>
<tr>
<td>Internet Archive 1996-2002</td>
<td>100 TByte</td>
</tr>
<tr>
<td>Particle Physics today</td>
<td>1 PByte/yr</td>
</tr>
<tr>
<td>LHC era physics</td>
<td>20 PByte/yr</td>
</tr>
</tbody>
</table>
The Grid label

Many distributed computing middlewares are now called “grid”

- Oracle 10g
- BOINC (formerly SETI@home)
- Sun Grid Engine
- Unicore
- Globus Toolkit 4
- gLite
- ...

And then there is middleware to build grids that is not usually branded as such

- Condor
- ...

...
But they are not all that ‘griddy’

- **Oracle 10g** = database on a cluster with node function changes
- **BOINC (formerly SETI@home)** = single application client/server
- **Sun Grid Engine** = cluster batch system
- Unicore
- Globus Toolkit 4
- gLite
- ...

- Condor
- ...

...
Grid Middleware

- Software that enables Grid
  - term deliberately vague, like the term Grid itself
  - but, from experience, one needs at least these services
    - resource discovery
    - resource scheduling
    - uniform compute access
    - uniform data access (to both files and structured data)
    - asynchronous information sources
    - authentication, delegation and secure communications
    - identity management
    - system management and system access
  - and these services should have a standard, common, interface

- In general, ‘middleware’ is used to describe the layer between network and application
Grid Middleware and their position
Typical Grid Topology
Job Description Language

This is JDL that the user might send to the Resource Broker

Executable = "catfiles.sh";
StdOutput = "catted.out";
StdError = "std.err";
Arguments = "EssentialJobData.txt LogicalJobs.jdl /etc/motd";

InputSandbox = {
"/home/davidg/tmp/jobs/LogicalJobs.jdl",
"/home/davidg/tmp/jobs/catfiles.sh"
};

OutputSandBox = {
"catted.out", "std.err"
};

InputData = "LF:EssentialJobData.txt";
ReplicaCatalog = 
ldap://rls.edg.org/1c=WPSIX,dc=cnrs,dc=fr";
DataAccessProtocol = "gsiftp";

RetryCount = 2;
How to you see what’s in the Grid?

Broker matches the user’s request with the site
- ‘information supermarket’ matchmaking (using Condor Matchmaking)
- uses the information published by the site

Grid Information system
- ‘the only information a user ever gets about a site’
- So: should be reliable, consistent and complete
- Standard schema (GLUE) to describe sites, queues, storage (complex schema semantics)
- Currently presented as an LDAP directory

LDAP Browser Jarek Gawor: www.mcs.anl.gov/~gawor/ldap
Attributes set per Site

- **Site information**
  - SiteSysAdminContact: mailto: grid-admin@example.org
  - SiteSecurityContact: mailto: security@example.org

- **Cluster info**
  - **GlueSubClusterUniqueID=gridgate.cs.tcd.ie**
    - HostApplicationSoftwareRunTimeEnvironment: LCG-2_6_0
    - HostApplicationSoftwareRunTimeEnvironment: VO-atlas-release-10.0.4
    - HostBenchmarkSI00: 1300
    - GlueHostNetworkAdapterInboundIP: FALSE
    - GlueHostNetworkAdapterOutboundIP: TRUE
    - GlueHostOperatingSystemName: RHEL
    - GlueHostOperatingSystemRelease: 3.5
    - GlueHostOperatingSystemVersion: 3

    - GlueCEStateEstimatedResponseTime: 519
    - GlueCEStateRunningJobs: 175
    - GlueCEStateTotalJobs: 248

- **Storage: similar info** (paths, max number of files, quota, retention, ...)


Grid in operation

LCG Real Time Monitor, Gidon Moont, RAL GOC
## Format - Numbers - Bits (examples!)

<table>
<thead>
<tr>
<th>Format</th>
<th>X</th>
<th>Y</th>
<th>Rate</th>
<th>Color bits/pix</th>
<th>Frame pix</th>
<th>Frame MByte</th>
<th>Flow MByt/s</th>
<th>Stream Gbit/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>720p HD</td>
<td>1280</td>
<td>720</td>
<td>60</td>
<td>24</td>
<td>921600</td>
<td>2.8</td>
<td>170</td>
<td>1.3</td>
</tr>
<tr>
<td>1080p HD</td>
<td>1920</td>
<td>1080</td>
<td>30</td>
<td>24</td>
<td>2073600</td>
<td>6.2</td>
<td>190</td>
<td>1.5</td>
</tr>
<tr>
<td>2k</td>
<td>2048</td>
<td>1080</td>
<td>24</td>
<td>48</td>
<td>2211840</td>
<td>10</td>
<td>240</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>480</td>
<td>2.4</td>
</tr>
<tr>
<td>SHD</td>
<td>3840</td>
<td>2160</td>
<td>30</td>
<td>24</td>
<td>8294400</td>
<td>25</td>
<td>750</td>
<td>6.0</td>
</tr>
<tr>
<td>4k</td>
<td>4096</td>
<td>2160</td>
<td>24</td>
<td>36</td>
<td>8847360</td>
<td>40</td>
<td>960</td>
<td>7.6</td>
</tr>
</tbody>
</table>

Note: this is excluding sound!
Note: these are raw uncompressed data rates!
What is a LightPath

• A LightPath is a circuit like connection that connects end systems to each other. This uses usually the same infrastructure as the Internet, but a LightPath gets dedicated resources next to Internet.

• A LightPath can be a combination of:
  – A color in a fiber (Lambda)
  – Sonet/sdh circuit in a sonet infrastructure
  – Vlans and dedicated ports in an ethernet switch
  – Etc.

• Aim is to get predictable and knowable connection characteristics
Very constant and predictable!
Internet Transport Protocols

- **IP = Internet Protocol**
  - Connectionless packet transport service
  - Datagrams of max 64 kByte that can be fragmented down the way
  - Packets can get lost, duplicated or out of order!

- **TCP/IP = Transmission Control Protocol**
  - Reliable byte-stream over potentially unreliable packet service
  - Connection oriented, exactly once and in order, end to end duplex

- **UDP = User Datagram Protocol**
  - Packet service up to 64 kByte
  - Connectionless, unidirectional, L2 switches may start flooding
  - Unreliable delivery, can get out of order, duplicated, lost
Issues & protocols

• When using UDP watch for bottleneck!
• About 10 other non standard protocols
• FAST TCP
  – Modified receiver algorithms
• RBUDP
  – Runs on top of UDP, simple back-off and retransmission scheme
Windows and buffering for reliable protocols

• Round Trip Time (rtt) is time it takes to send a shortest message and get the answer back (unix tool ping).
• That is the shortest time the sender can know traffic arrived at the other end.
• Sender can only discard old data after receiving seq + 1 ack.
• Lightspeed in fiber = 200000 km/s.
• 100 km = 200 km round trip:
  – Amsterdam - Geneve: $3\text{ ms}$
  – Amsterdam - Chicago: $90\text{ ms}$
  – Amsterdam - San Diego: $160\text{ ms}$
  – Amsterdam - Tokyo: $267\text{ ms}$
  – Amsterdam - Sydney: $300\text{ ms}$

Therefore:
$NL = 6\text{ ms}^2$
## Buffer space

Window = RTT * BW

<table>
<thead>
<tr>
<th>RTT</th>
<th>100 Mbit/s</th>
<th>1 Gbit/s</th>
<th>10 Gbit/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.5 kB</td>
<td>125 kB</td>
<td>1.25 MB</td>
</tr>
<tr>
<td>2</td>
<td>25 kB</td>
<td>250 kB</td>
<td>2.5 MB</td>
</tr>
<tr>
<td>5</td>
<td>62.5 kB</td>
<td>615 kB</td>
<td>6.15 MB</td>
</tr>
<tr>
<td>10</td>
<td>125 kB</td>
<td>1.25 MB</td>
<td>12.5 MB</td>
</tr>
<tr>
<td>20</td>
<td>250 kB</td>
<td>2.5 MB</td>
<td>25 MB</td>
</tr>
<tr>
<td>50</td>
<td>625 kB</td>
<td>6.25 MB</td>
<td>62.5 MB</td>
</tr>
<tr>
<td>100</td>
<td>1.25 MB</td>
<td>12.5 MB</td>
<td>125 MB</td>
</tr>
<tr>
<td>200</td>
<td>2.5 MB</td>
<td>25 MB</td>
<td>250 MB</td>
</tr>
<tr>
<td>500</td>
<td>6.25 MB</td>
<td>62.5 MB</td>
<td>625 MB</td>
</tr>
<tr>
<td>1000</td>
<td>12.5 MB</td>
<td>125 MB</td>
<td>1250 MB</td>
</tr>
</tbody>
</table>
TCP Tuning (if not auto-tuning)

• 1 Gbit/s on 160 ms RTT (= Amsterdam - San Diego):
  – sysctl -w kern.ipc.maxsocketbuf=50000000
  – sysctl -w net.inet.tcp.sendspace=21000000
  – sysctl -w net.inet.tcp.recvspace=21000000
  – sysctl -w net.inet udp.maxdgram=57344
  – sysctl -w net.inet udp.recvspace=74848
  – sysctl -w net.local.stream.sendspace=32768
  – sysctl -w net.local.stream.recvspace=32768
  – sysctl -w kern.ipc.somaxconn=512
  – sysctl -w net.inet.tcp.mssdflt=1460
  – sysctl -w net.inet.tcp.delayed_ack=2
  – sysctl -w net.inet.tcp.rfc1323=1
  – sysctl -w net.inet.tcp.rfc1644=1
  – sysctl -w net.inet.tcp.newreno=1
End System Issues

• Ethernet card interface to computer bus system
  – PCI-X
    • 32/64 bit 66/133/266 MHZ -> about 8 Gbit/s max in 133 MHZ mode
  – PCI-Express
    • 2.5 Gbit/s per lane, 4, 8, 16 lanes

• Memory organization

• CPU cache
  – Effect when things go out of cache (small windows, etc.)

• CPU core
  – Takes 1 core to handle network (affinity may help)

• Disk raid subsystem
  – raid0 twice as fast as raid5
  – One disk does typically 40 MB/s write, 60 MB/s read
Amsterdam CineGrid S/F node

DAS-3 - 4U set @UvA

Rembrandt Cluster
total 22 TByte disk space @ LightHouse

NetherLight, StarPlane
the cp testbeds and beyond

10 Gbit/s
DP AMD processor nodes
head node (?)
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node

GlimmerGlass
photonic switch
10 Gbit/s

NORTEL
8600 L2/3 switch

F10 L2/3 switch

head node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node

Opteron 64 bit nodes

head node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node

streaming node 8 TByte

storage node 96 TByte

10 Gbit/s

10 Gbit/s

MYRINET

M Y R I N E T

32-77x
RDF describing Infrastructure

Application: find video containing x, then trans-code to it view on Tiled Display
Questions?

www.cinegrid.org
www.cinegrid.nl
www.supertube.org
www.science.uva.nl/~delaat