High Speed Transport Protocols Evaluation in Grid5000

Pascale Vicat-Blanc Primet
Senior Researcher at INRIA
Leader of the RESO team
LIP Laboratory
UMR CNRS-INRIA-ENS-UCBL
Ecole Normale Supérieure de Lyon France
Pascale.primet@inria.fr

NRENs and Grids
TERENA workshop
7/12/06
Outline

Grid Internetworking Research

Grid5000 testbed

HS transport protocols evaluation

Conclusion & perspective
Grid Internetworking research

Data movements & bandwidth sharing

The GridNetwork
**EC-GIN: Grid Internetworking**

- **Original Internet technology**
- **Enriched with customised network mechanisms**

**Today’s Grid applications**
- Driving a racing car on a public road

**Traditional Internet applications** (web browser, ftp, ..)

**EC-GIN**

**EC-GIN enabled Grid applications**
- Applications with special network properties and requirements
- Real-time multimedia applications (VoIP, video conference, ..)

⇒ Faster Grid: network mechanisms based on Grid peculiarities
⇒ Economic Grid traffic management and security
EC-GIN : Research Challenges

• How to model Grid traffic?
  - Much is known about web traffic (e.g. self-similarity) - but the Grid is different!

• How to simulate a Grid-network?
  - Necessary for checking various environment conditions
  - May require traffic model (above)
  - Currently, Grid-Sim / Net-Sim are two separate worlds
    (different goals, assumptions, tools, people)

• How to specify network requirements?
  - Explicit or implicit, guaranteed or “elastic”, various possible levels of
    granularity (=> new or extended APIs?)

• How to align network and Grid economics?
  - Grid service model, charging model for grid services, and network model
    for such Grid services
  - Network Mgmt mechanisms in support of those three areas in an
    integrated fashion
Grid Internetwork

- The shared resources are interconnected by a complex internetwork
- Applications use Internet protocols: TCP/IP

Main Networking Issues:
1: Security
2: E2E performance prediction and control

Local area networks
Gigabit/10Gb/s
Eth, IB, Myri

Access Link
1, 10 Gb/s Ethernet

Core network
Internet
MPLS VPN
GMPLS
OBS...

⇒ Main Networking Issues:
⇒ 1: Security
⇒ 2: E2E performance prediction and control
E2E performance

Combination of many factors:

-> cross all layers and all elements of the E2E chain

Problems related to the network
  • if not overprovisionned or if no QoS support ...

Problems related to the TCP protocol
  • TCP designed first and foremost to be robust and when congestion is detected, TCP accommodates at the expense of reduced performance.

Problems related to the TCP configuration
  • small buffer space or SACK improperly negotiated

Problems related to the end system: hardware & OS
  • to the processor, bus speed, I/O devices
  • to the NIC with its associated driver;

Problems due to the applications
  • small messages or pauses in the data flow

-> quantify the contribution of the different layers and different elements
E2E performance

Objective function: MCT
- Minimum completion time

Speedup depends on \( C/T \)
- \( C \): computing time / image
- \( T \): transfer time / image

Speedup is very low in a congested network
Speedup is good in a controlled network
E2E performance
Flows interaction problem
The reality of TCP like congestion control algorithm at high speed
Outline

Grid Internetworking Research

Grid5000 testbed

HS transport protocols evaluation

Conclusion & perspective
GRID5000 initiative

A nation wide experimental platform for Grid researches

- 9 geographically distributed sites
- every site hosts a cluster (from 256 CPUs to 1K CPUs)
- All sites are connected by RENATER (10Gb/s DWDM VPN)
- A system/middleware environment for safe and repeatable experiments

Run Grid experiments in real life conditions

- Address critical issues of Grid system/middleware:
  - Programming, Scalability, Fault Tolerance, Scheduling
- Address critical issues of Grid Networking
  - High performance transport, QoS, measurement, distributed security
- Port and test applications
- Investigate innovative approaches
  - P2P resources discovery, Desktop Grids, active grids
Grid5000 network

RENATER-4

9 Clusters with 256 to 1K CPUS
=> about 2600 CPUs

10Gb/s
Dedicated lambdas

Grid5000 software:
Resource reservation

OAR
Automatic reconfiguration

KADEPLOY

INRIA
Special features

4 main features:

• A high security for Grid’5000 and the Internet, despite the deep reconfiguration feature
  --> Grid’5000 is confined: communications between sites are isolated from the Internet and Vice versa (level2 MPLS, Dedicated lambda).

• A software infrastructure allowing users to access Grid’5000 from any Grid’5000 site and have simple view of the system
  --> A user has a single account on Grid’5000, Grid’5000 is seen as a cluster of clusters, 9 (1 per site) unsynchronized home directories

• A reservation/scheduling tools allowing users to select nodes and schedule experiments
  a reservation engine + batch scheduler (1 per site) + OAR Grid (a co-reservation scheduling system)

• A user toolkit to reconfigure the nodes
  software image deployment and node reconfiguration tool
Reservation & Batch Scheduler
• Experiment: Geophysics: Seismic Ray Tracing in 3D mesh of the Earth

Building a seismic tomography model of the Earth using seismic wave propagation characteristics.

Seismic waves are modeled from events detected by sensors.

Ray tracing algorithm: waves are reconstructed from rays traced between the epicenter and one sensor.

A MPI parallel program composed of 3 steps:
1) Master-worker: ray tracing and mesh update by each process with blocks of rays successively fetched from the master process,
2) All-to-all communications to exchange submesh information between the processes,
3) Merging of cell information associated with each process.

Reference: 32 CPUs

IPGS: “Institut de Physique du Globe de Strasbourg”
Outline

Grid Internetworking Research

Grid5000 testbed

HS transport protocols evaluation

Conclusion & perspective
E2E performance problem

=> Wizard gap problem

Slide from Matt Mathis - PSC
Grid5000 network

Black fibers are rent by the network provider
RENATER is enlightened by RENATER

Source: Cees de Laat (UvA)

Next step? Lambdas on demand?

Sharing Grid5000 & DAS3
Is there a wizard gap problem in Grid5K?

Novice: 1Gb/s measurement, with default kernel images => goodput in Mb/s

<table>
<thead>
<tr>
<th></th>
<th>Bo</th>
<th>Gr</th>
<th>Li</th>
<th>Ly</th>
<th>Na</th>
<th>Or</th>
<th>Re</th>
<th>So</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bo</td>
<td></td>
<td></td>
<td>58.1</td>
<td>61.8</td>
<td>55.9</td>
<td>81.2</td>
<td>111</td>
<td>76.3</td>
<td>68.9</td>
</tr>
<tr>
<td>Gr</td>
<td>32.3</td>
<td></td>
<td></td>
<td>34.0</td>
<td>151</td>
<td>39.8</td>
<td>33.7</td>
<td>34.3</td>
<td>52.6</td>
</tr>
<tr>
<td>Li</td>
<td>53.3</td>
<td>70.0</td>
<td></td>
<td>53.6</td>
<td>112</td>
<td>199</td>
<td></td>
<td>55</td>
<td>44.3</td>
</tr>
<tr>
<td>Ly</td>
<td>61.5</td>
<td>230</td>
<td>71.2</td>
<td></td>
<td>97.6</td>
<td>106</td>
<td>49.8</td>
<td>100</td>
<td>72.0</td>
</tr>
<tr>
<td>Na</td>
<td>48.0</td>
<td>162</td>
<td>78.5</td>
<td>52.4</td>
<td></td>
<td></td>
<td>777</td>
<td>54.7</td>
<td>43.3</td>
</tr>
<tr>
<td>Or</td>
<td>67.8</td>
<td>54.1</td>
<td>150</td>
<td>58.8</td>
<td>936</td>
<td></td>
<td></td>
<td>68.7</td>
<td>36.2</td>
</tr>
<tr>
<td>Re</td>
<td>64.2</td>
<td>33.6</td>
<td>46.6</td>
<td>41.4</td>
<td>45.5</td>
<td>56.5</td>
<td></td>
<td>27.4</td>
<td>26.3</td>
</tr>
<tr>
<td>So</td>
<td>47</td>
<td>46.1</td>
<td>29.5</td>
<td>67.4</td>
<td>28.9</td>
<td>22.3</td>
<td>25.1</td>
<td></td>
<td>34.0</td>
</tr>
<tr>
<td>To</td>
<td>166</td>
<td>47.6</td>
<td>29.8</td>
<td>65.7</td>
<td>29.7</td>
<td>44.3</td>
<td>26.3</td>
<td>36.1</td>
<td></td>
</tr>
</tbody>
</table>

YES!

R <10%
Insufficient buffer size signature

newRENO; 100ms; skb=BDP

newRENO; 100ms; skb<BDP

• BDP : Bandwidth delay product, buffer size has to equal to BDP

• BDP mean in GRID5000 = 10e9 x 0,01 = 10e7 bits = 2,5MB

• Default buffer size = 170KB
  => max throughput = 128 x 8 x 100 = 102 400 Kb/s = 100Mb/s
Is there a wizard gap problem in G5K?

Expert: 1Gb/s measurement, with tuned kernel: goodput in Mb/s

<table>
<thead>
<tr>
<th></th>
<th>Bo</th>
<th>Gr</th>
<th>Li</th>
<th>Ly</th>
<th>Na</th>
<th>Or</th>
<th>Re</th>
<th>So</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bo</td>
<td></td>
<td>771</td>
<td>725</td>
<td>862</td>
<td>911</td>
<td>884</td>
<td>852</td>
<td>875</td>
<td>685</td>
</tr>
<tr>
<td>Gr</td>
<td>900</td>
<td></td>
<td>701</td>
<td>925</td>
<td>812</td>
<td>893</td>
<td>787</td>
<td>911</td>
<td>647</td>
</tr>
<tr>
<td>Li</td>
<td>738</td>
<td>838</td>
<td></td>
<td>120</td>
<td>922</td>
<td>848</td>
<td>916</td>
<td>598</td>
<td>579</td>
</tr>
<tr>
<td>Ly</td>
<td>425</td>
<td>912</td>
<td>786</td>
<td></td>
<td>904</td>
<td>740</td>
<td>864</td>
<td>926</td>
<td>730</td>
</tr>
<tr>
<td>Na</td>
<td>725</td>
<td>851</td>
<td>742</td>
<td>865</td>
<td></td>
<td>854</td>
<td>938</td>
<td>931</td>
<td>622</td>
</tr>
<tr>
<td>Or</td>
<td>799</td>
<td>866</td>
<td>777</td>
<td>869</td>
<td>936</td>
<td></td>
<td>849</td>
<td>878</td>
<td>523</td>
</tr>
<tr>
<td>Re</td>
<td>912</td>
<td>831</td>
<td>787</td>
<td>859</td>
<td>914</td>
<td>912</td>
<td></td>
<td>839</td>
<td>651</td>
</tr>
<tr>
<td>So</td>
<td>901</td>
<td>839</td>
<td>653</td>
<td>543</td>
<td>611</td>
<td>900</td>
<td>321</td>
<td></td>
<td>694</td>
</tr>
<tr>
<td>To</td>
<td>928</td>
<td>859</td>
<td>784</td>
<td>882</td>
<td>933</td>
<td>923</td>
<td>939</td>
<td>909</td>
<td></td>
</tr>
</tbody>
</table>

YES!
G > 80% of $C_T$
G < 9% of 10Gb/s
**High Speed-TCP approaches:**
Modify the Congestion control algorithm

High Perf TCP congestion control aim at minimizing this surface
(HS-TCP, S-TCP, H-TCP, BIC, CuBIC...)

High Perf transport protocols issues:
- Fairness, convergence, efficiency
- RTT fairness, Friendliness
- Reaction to available bandwidth dynamic
Example of testbed setup

13 x PCs

12 x 1 GbE

10 GbE

12x 1 GbE

13 x PCs

(futur) 1 x 10GbE

Grid5000
Or 10Gb/s WAN emulator

iperf

iperfd

13 x PCs

(futur) 1 x 10GbE
CUBIC in Grid5000 (11.5ms Rennes-Nancy)
HSTCP in Grid5000 (11.5ms Rennes-Nancy)
Parallel streams study

BIC TCP: 11 flows with 1, 2, 5 or 10 streams
Long distance MPI optimisation
Outline

Grid Internetworking Research

Grid5000 testbed

HS transport protocols evaluation

Conclusion & perspective
Conclusion & perspectives

• Grid5000 provides a unique testbed for high speed transport protocol benchmarking.
  • network controllable, end nodes redeployable
  • network instrumentation is necessary (on going work with Renater)
  • flow level monitoring at 10Gb/s is needed but very challenging

• We work at connecting Grid5000 with other international testbeds

• Many more studies are planned to better understand how end user can fully and systematically benefit from huge available capacity taking into account:
  • Hardware evolution
  • Networking technology evolution
  • New network & protocol architectures
GRID5000 networking collaborations

Interconnection of GRID5000 and DAS3 testbeds
- via RENATER- GEANT- SURFNET
- France - Netherland
- 10Gb/s dedicated lambda through europe

Interconnection of GRID5000 and Naregi testbeds
- via RENATER- GEANT- SUPERsinet
- France - Japan
- 1Gb/s dedicated channel through atlantic, usa, pacific

Interconnection between Lyon and Chicago (IN2P3/FNAL): ANR IGTMD
- via RENATER- GEANT- ESNET
- France - USA
- 2Gb/s dedicated channel through atlantic
Contacts

- Pascale.Primet@ens-lyon.fr

First international IEEE GRIDNETS 2007 conference in LYON (France)
17-19 october 2007
http://gridnets.eu

Looking for sponsors and contributors
Reserve