High Performance Networking for Grid Applications

www.science.uva.nl/~delaat

Cees de Laat

Faculty of Science
High Performance Networking for Grid Applications

www.science.uva.nl/~delaat

Cees de Laat

EU

SURFnet

University of Amsterdam
Contents of this talk

This slide is intentionally left blank
eVLBI
The sensitivity of the VLBI array scales with the square root of the (data-rate) and there is a strong push to increase this. Rates of 8Gb/s or more are entirely feasible under development. It is expected that parallel correlators will remain the most efficient approach; distributed processing may have an application. Multi-gigabit data streams will aggregate into larger networks and the capacity of the final link to the data correlator.
• 28 demonstrations from 16 countries: Australia, Canada, CERN, France, Finland, Germany, Greece, Italy, Japan, The Netherlands, Singapore, Spain, Sweden, Taiwan, United Kingdom, United States
• Applications demonstrated: art, bioinformatics, chemistry, cosmology, cultural heritage, education, high-definition media streaming, manufacturing, medicine, neuroscience, physics, tele-science
• Grid technologies demonstrated: Major emphasis on grid middleware, data management grids, data replication grids, visualization grids, data/visualization grids, computational grids, access grids, grid portals
• 25Gb transatlantic bandwidth (100Mb/attendee, 250x iGrid2000!)

www.igrid2002.org
Experimental Networks

- High-performance trials of new technologies that support *application-dictated* development of software toolkits, middleware, computing and networking.
- Provide *known and knowable characteristics* with deterministic and repeatable behavior on a persistent basis, while encouraging experimentation with innovative concepts.
- Experimental Networks are seen as the *missing link* between Research and Production Networks.

http://www.evl.uic.edu/activity/NSF/index.html
http://www.calit2.net/events/2002/nsf/index.html
What is a LambdaGrid?

- A **grid** is a set of networked, middleware-enabled computing resources.
- A **LambdaGrid** is a grid in which the lambda networks themselves are resources that can be scheduled, like all other computing resources. The ability to schedule and provision lambdas provides **deterministic** end-to-end network performance for real-time or time-critical applications, which cannot be achieved on today’s grids.
A. Lightweight users, browsing, mailing, home use
   Need full Internet routing, one to many

B. Business applications, multicast, streaming, VPN’s, mostly LAN
   Need VPN services and full Internet routing, several to several + uplink

C. Special scientific applications, computing, data grids, virtual-presence
   Need very fat pipes, limited multiple Virtual Organizations, few to few
A. Lightweight users, browsing, mailing, home use
   Need full Internet routing, one to many

B. Business applications, multicast, streaming, VPN’s, mostly LAN
   Need VPN services and full Internet routing, several to several + uplink

C. Special scientific applications, computing, data grids, virtual-presence
   Need very fat pipes, limited multiple Virtual Organizations, few to few
Scale 2-20-200
The only formula’s

\[ \# \lambda(\text{rtt}) \approx \frac{200 \times e^{(t-2002)}}{\text{rtt}} \]

Now, having been a High Energy Physicist we set
\[ c = 1 \]
\[ e = 1 \]
\[ \bar{h} = 1 \]
and the formula reduces to:
\[ \# \lambda \approx \frac{200 \times e^{(t-2002)}}{\text{rtt}} \]
SURFnet Lambda’s fibers (old already)
# Services

<table>
<thead>
<tr>
<th>SCALE</th>
<th>CLASS</th>
<th>2 Metro</th>
<th>20 National/regional</th>
<th>200 World</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Switching/routing</td>
<td>Routing</td>
<td>ROUTER$</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>VPN’s, (G)MPLS</td>
<td>VPN’s Routing</td>
<td>Routing</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>dark fiber Optical switching</td>
<td>Lambda switching</td>
<td>Sub-lambdas, ethernet-sdh</td>
<td></td>
</tr>
</tbody>
</table>

\[
\# \lambda \approx \frac{200 \times e^{(t-2002)}}{rtt}
\]
Current technology + (re)definition

- Current (to me) available technology consists of SONET/SDH switches, 10 gig ethernet and dark fiber environments
- Optical switch installed (this week)!
- DWDM+switching included
- Starlight/NetherLight deploy VLAN’s on Ethernet switches to connect [exactly two] ports (but also routing)
- We want to understand routerless limited environments
- So redefine a λ as:

  “a λ is a pipe where you can inspect packets as they enter and when they exit, but principally not when in transit. In transit one only deals with the parameters of the pipe: number, color, bandwidth”
MEMS optical switch (CALIENT)
So what are the facts

- Costs of fat pipes (fibers) are one/third of cost of equipment to light them up
  - Is what Lambda salesmen tell me

- Costs of optical equipment 10% of switching 10% of full routing equipment for same throughput
  - 100 Byte packet @ 40 Gb/s -> 20 ns to look up in 140 kEntries routing table (light speed from me to you!)

- Big sciences need fat pipes

- Bottom line: look for a hybrid architecture which serves all users in a cost effective way
Architectures - L1 - L3

Long haul $\lambda$

Internet

TDM

R

Internet

R

SW

L2 VPN’s

(10b of 13)
- **lambda** for high bandwidth applications
  - Bypass of production network
  - Middleware may request (optical) pipe
- **RATIONALE:**
  - Lower the cost of transport per packet
How low can you go?

Application Endpoint A

Local Ethernet

MEMS

Regional dark fiber

ONS 15454

Trans-Atlantic

Application Endpoint B

Router

Ethernet

SONET

DWDM

fiber

NetherLight

TransLight

POS
Virtual Organization on L2
NetherLight Network: 2003
Emerging international lambda grid

- Stockholm Northern Light
- Amsterdam NetherLight
- Chicago StarLight
- London UKLight
- Geneva CERN
- Prague CzechLight
- Dwingeloo ASTRON/JIVE

Network connections:
- Tyco/IEEAF 10 Gbit/s
- SURFnet 10 Gbit/s
- CESNET 2.5 Gbit/s
- NSF 10 Gbit/s

Operational 1H03:
- New York City
- Chicago StarLight
- London UKLight
- Stockholm Northern Light

Expected 2H03:
- Prague CzechLight
SURFnet backbone

Lambda’s to
- Chicago,
- Geneve,
- Praha,
- NYC
- London

Dark fiber To Dwingeloo

NetherLight

15454

Fat pc

1 Gbps

4 HP servers

DAS: 32*2cpu’s IBM Myrinet 100Mbs

1 Gbps

10 Gbps

UvA/NikHEF/SARA
Early Lambda/LightPath TDM experiences
5000 1 kByte UDP packets
TCP is bursty due to sliding window protocol and slow start algorithm.

\[ \text{Window} = \text{BandWidth} \times \text{RTT} \quad \& \quad \text{BW} = \text{slow} \]

\[ \text{Memory-at-bottleneck} = \frac{\text{fast} \times \text{slow} \times \text{RTT}}{} \]

So pick from menu:
- Flow control
- Traffic Shaping
- RED (Random Early Discard)
- Self clocking in TCP
- Deep memory
Self-clocking of TCP

- **WS**
- **L2** (fast) \(\rightarrow\) (slow)
- **high RTT**
- **L2** (slow) \(\rightarrow\) (fast)

- 14 µsec
- 20 µsec
- 20 µsec
Layer - 2 requirements from 3/4

Window = BandWidth * RTT & BW == slow

Memory-at-bottleneck = __________ * slow * RTT
                         fast

Given M and f, solve for slow ===>

0 = s² - f * s + ________
    RTT

s₁, s₂ = \frac{f}{2} (1 +/- sqrt(1 - 4 \frac{M}{f \cdot RTT}))
Forbidden area, solutions for $s$ when $f = 1 \text{ Gb/s}$, $M = 0.5 \text{ Mbyte}$ AND NOT USING FLOWCONTROL

158 ms = RTT Amsterdam - Vancouver
Problem Solving Environment

Applications and Supporting Tools

Application Development Support

Collective Grid Services
- Brokering
- Global Queuing
- Co-Scheduling
- Data Cataloguing
- Auditing
- Authorization
- Monitoring
- Fault Management

Common Grid Services
- Grid Information Service
- Uniform Resource Access
- Global Event Services
- Uniform Data Access
- Data Replication
- Communication Services

Grid Security Infrastructure (authentication, proxy, secure transport)

Communication

Grid access (proxy authentication, authorization, initiation)

Fabric
- Grid task initiation

Local Resources
- Resource Manager
  - CPUs
- Resource Manager
  - Monitors
- Resource Manager
  - On-Line Storage
- Resource Manager
  - Scientific Instruments
- Resource Manager
  - Tertiary Storage
- Resource Manager
  - Highspeed Data Transport

layers of increasing abstraction taxonomy

High performance computing and Processor memory co-allocation
Security and Generic AAA
Optical Networking

Researched in other programlines
Imported from the Globus toolkit
Starting point

Generic AAA server
Rule based engine

API

Application Specific Module

Service
Accounting Metering

Policy
Data

PDP

PEP

RFC 2903 - 2906, 3334, policy draft

(19e of 20)
(Future) Projects

• National:
  • NCF Grid project
  • VLE
  • GigaPort-NG
  • LOFAR

• European
  • DataGrid
  • DataTAG

• International
  • NetherLight
  • StarLight
  • AnyLight, LowLight, BackLight
  • Optiputer

Research:
Models of Lambda networking
Transport
AAA
Transport in the corners

- BW*RTT
- Needs more App & Middleware interaction
- For what current Internet was designed
- Full optical future
- # FLOWS

(19h of 20)
Thanks to

SURFnet: Kees Neggers, UIC & iCAIR: Tom DeFanti, Joel Mambretti, CANARIE: Bill St. Arnaud
This work is supported by: SURFnet, EU-IST project DATATAG