Lambda-Grid developments
Global Lambda Integrated Facility
www.science.uva.nl/~delaat
Cees de Laat
SURFnet
EU
University of Amsterdam
Contents

This page is intentionally left blank

- Ref: www.this-page-intentionally-left-blank.org
Sensor Grids

~ 40 Tbit/s
www.lofar.org

eVLBI

longer term VLBI is easily capable of generating such data. The sensitivity of the VLBI array scales with bandwidth (data-rate) and there is a strong push to move to higher bandwidths. Rates of 8Gb/s or more are entirely feasible, and are under development. It is expected that parallel-processed correlator will remain the most efficient approach to process distributed correlators with multi-gigabit data rate. The synchrotron factor.

Westerbork Synthesis Radio Telescope - Netherlands
LHC Data Grid Hierarchy
CMS as example, Atlas is similar

Tier 1
- Italian Regional Center
- German Regional Center
- NIKHEF Dutch Regional Center
- FermiLab, USA Regional Center

Tier 0 +1
- Online System

Tier 2
- CERN/CMS data goes to 6-8 Tier 1 regional centers, and from each of these to 6-10 Tier 2 centers.
- ~0.6-2.5 Gbps

Tier 3
- Tier 2 Center
- ~0.6-2.5 Gbps

Tier 4
- Institute
- Physics data cache
- ~0.25 TIPS
- Workstations
- 100 - 1000 Mbits/sec

Analysis

Event reconstruction

~2.5 Gbits/sec

~100 MBytes/sec

~PByte/sec

Physicists work on analysis “channels” at 135 institutes. Each institute has ~10 physicists working on one or more channels.

2000 physicists in 31 countries are involved in this 20-year experiment in which DOE is a major player.

Tier 0

Tier 1

Tier 2

Tier 3

Tier 4

CMS detector: 15m X 15m X 22m
12,500 tons, $700M.

~0.25 TIPS Workstations

~100 MBytes/sec

~100 Mbits/sec

~2.5 Gbits/sec

~0.6-2.5 Gbps

~0.6-2.5 Gbps

~2.5 Gbits/sec

CERN/CMS data goes to 6-8 Tier 1 regional centers, and from each of these to 6-10 Tier 2 centers.

Physicists work on analysis “channels” at 135 institutes. Each institute has ~10 physicists working on one or more channels.

2000 physicists in 31 countries are involved in this 20-year experiment in which DOE is a major player.

Courtesy Harvey Newman, CalTech and CERN
Data intensive scientific computation through global networks

Nuclear experiments

Belle Experiments

Very high-speed Network

Data Reservoir

Distributed Shared files

Local Accesses

Data analysis at University of Tokyo

Grape6

SUBARU Telescope

Digital Sky Survey

X-ray astronomy Satellite ASUKA

Nobeyama Radio Observatory (VLBI)
Showed you 5 types of Grids

- **Sensor Grids**
  - Several massive data sources are coming online

- **Computational Grids**
  - HEP and LOFAR analysis needs massive CPU capacity
  - Research: dynamic nation wide optical backplane control

- **Data (Store) Grids**
  - Moving and storing HEP, Bio and Health data sets is major challenge

- **Visualization Grids**
  - Data object (TByte sized) inspection, anywhere, anytime

- **Lambda Grids**
  - Hybrid networks
Co-located interactive 3D visualization

The markers are tracked by infrared cameras.

The positions are transmitted to the visualization system.

The new image is transmitted to the display.

The visualization system uses the reported positions to render a new image of the visualized data.

The volumetric data resides locally on the visualization system.

10 Gigabit/s path on the SURFnet and Abilene networks.

SGI Onyx4 at SARA.

Amsterdam

Pittsburgh
SC2004 “Dead Cat” demo

SuperComputing 2004, Pittsburgh, Nov. 6 to 12, 2004

Produced by:
Michael Scarpa
Robert Belleman
Peter Sloot

Many thanks to:
AMC
SARA
GigaPort
UvA/AIR
Silicon Graphics, Inc.
Zoölogisch Museum
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. Scientific applications, distributed data processing, all sorts of grids
   Need very fat pipes, limited multiple Virtual Organizations, few to few, p2p
The Dutch Situation (in 2004)

- **Estimate A**
  - 17 M people, 6.4 M households, 25 % penetration of 0.5-2.0 Mb/s ADSL, 40 times under-provisioning ==> 20 Gb/s
AMS-IX

June 19th 2004

European championship football  Holland -- Czech Republic

Lost :-(

Holland -- Czech Republic

Maximal In: 27.782 G Maximal Out: 27.794 G
Current In: 27.239 G Current Out: 27.239 G

Bits per second

10:00  00:00  06:00  12:00  18:00
The Dutch Situation (in 2004)

• Estimate A
  – 17 M people, 6.4 M households, 25 % penetration of 0.5-2.0 Mb/s ADSL, 40 times under-provisioning ==> 20 Gb/s

• Estimate B
  – SURFnet5 has 2*10 Gb/s to about 15 institutes and 0.1 to 1 Gb/s to 170 customers, estimate same for industry (overestimation) ==> 10-30 Gb/s
Routed L3 traffic growth

SURFnet customer traffic: Monthly volume

1600 Tbyte/month ≈ 5 Gbits/second

Slide courtesy Kees Neggers
The Dutch Situation (in 2004)

- **Estimate A**
  - 17 M people, 6.4 M households, 25% penetration of 0.5-2.0 Mb/s ADSL, 40 times under-provisioning ==> 20 Gb/s

- **Estimate B**
  - SURFnet5 has 2*10 Gb/s to about 15 institutes and 0.1 to 1 Gb/s to 170 customers, estimate same for industry (overestimation) ==> 10-30 Gb/s

- **Estimate C**
  - Leading HEF and ASTRO + rest ==> 80-120 Gb/s
  - LOFAR ==> ≈ 37 Tbit/s ==> ≈ n x 10 Gb/s
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. Scientific applications, distributed data processing, all sorts of grids  
   Need very fat pipes, limited multiple Virtual Organizations, few to few, p2p

\[ \Sigma A \approx 20 \text{ Gb/s} \]
\[ \Sigma B \approx 30 \text{ Gb/s} \]
\[ \Sigma C \gg 100 \text{ Gb/s} \]
λ’s on scale 2-20-200 ms rtt
Towards Hybrid Networking!

- Costs of optical equipment 10% of switching 10% of full routing equipment for same throughput
  - 10G routerblade -> 100-500 k$, 10G switch port -> 7-15 k$, MEMS port -> 1 k$
  - DWDM lasers for long reach expensive, 10-50 k$

- Bottom line: look for a hybrid architecture which serves all classes in a cost effective way (map A -> L3, B -> L2, C -> L1)

- Give each packet in the network the service it needs, but no more!

\[
\begin{align*}
L1 & \approx 1 \text{ k$/port} \\
L2 & \approx 7-15 \text{ k$/port} \\
L3 & \approx 100+ \text{ k$/port}
\end{align*}
\]
## Services

<table>
<thead>
<tr>
<th>SCALE</th>
<th>CLASS</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Switching/routing</td>
<td>Switches + E-WANPHY VPN’s</td>
<td>dark fiber DWDM MEMS switch</td>
</tr>
<tr>
<td></td>
<td>Metro</td>
<td>Routing</td>
<td>Switches + E-WANPHY (G)MPLS</td>
<td>DWDM, TDM / SONET Lambda switching</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>20 National/regional</td>
<td>200 World</td>
<td>200 World</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ROUTER$</td>
<td>ROUTER$</td>
<td>ROUTER$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lambda, VLAN’s SONET Ethernet</td>
<td>Lambda, VLAN’s SONET Ethernet</td>
<td>Lambda, VLAN’s SONET Ethernet</td>
</tr>
</tbody>
</table>
How low can you go?

Application Endpoint A

Router

Ethernet

SONET

DWDM

Fiber

Local Ethernet

POS

MEMS

Regional dark fiber

Application Endpoint B

15454 / 6500

Trans-Oceanic

UKLight

StarLight

GLIF

NetherLight

Application
Optical Exchange as Black Box

Optical Exchange

Switch
TDM
Store & Forward
DWDM mux/demux

TeraByte Email Service
Subnetwork 1: Green

Subnetwork 2: Dark blue

Subnetwork 3: Red

Subnetwork 4: Blue Azur

Subnetwork 5: Grey

Common Photonic Layer (CPL) in SURFnet6
StarPlane
DWDM
backplane
GRID-Colocation problem space

Extensively under research

New!
Laying of fiber near/at Science Park
Amsterdam

Pictures by Yuri Demchenko
GLIF Q4 2004

Visualization courtesy of Bob Patterson, NCSA.
Discipline Networks

Internet
HEP
ASTRO
Earth Science

Fibers

Lambdas
GLIF History

• Brainstorming in Antalya at Terena conf. 2001
  • 1th meeting at Terena offices 11-12 sep 2001
    – On invitation only (15) + public part
    – Thinking, SURFnet test lambda Starlight-Netherlight
• 2nd meeting appended to iGrid 2002 in Amsterdam
  – Public part in track, on invitation only day (22)
  – Core testbed brainstorming, idea checks, seeds for Translight
• 3th meeting Reykjavik, hosted by NORDUnet 2003
  – Grid/Lambda track in conference + this meeting (35!)
  – Brainstorm applications and showcases
  – Technology roadmap
  – GLIF established --> glif.is
• 4th meeting Nottingham (UK), hosted by UKERNA, 2-3 September 2004
  – 60 participants
  – Attendance from China, Japan, Netherlands, Switzerland, US, UK, Taiwan, Australia, Tsjech, Korea, Canada, Ireland, Russia, Belgium, Denmark
  – Truly Worldwide!
GLIF Mission Statement

• GLIF is a world-scale Lambda-based Laboratory for application and middleware development on emerging LambdaGrids, where applications rely on dynamically configured networks based on optical wavelengths

• GLIF is an environment (networking infrastructure, network engineering, system integration, middleware, applications) to accomplish real work
Working groups

**GLIF Governance and policy**

Our small-scale Lambda Workshop is now turning into a global activity. TransLight and similar projects contribute to the infrastructure part of GLIF. A good and well understood governance structure is key to the manageability and success of GLIF. Our prime goal is to decide upon and agree to the GLIF governance and infrastructure usage policy.

**GLIF Lambda infrastructure and Lambda exchange implementations**

A major function for previous Lambda Workshops was to get the network engineers together to discuss and agree on the topology, connectivity and interfaces of the Lambda facility. Technology developments need to be folded into the architecture and the expected outcome of this meeting is an agreed view on the interfaces and services of Lambda exchanges and a connectivity map of Lambdas for the next year, with a focus on iGrid 2005 and the emerging applications.

**Persistent Applications**

Key to the success of the GLIF effort is to connect the major applications to the Facility. We, therefore, need a list of prime applications to focus on and a roadmap to work with those applications to get them up to speed. The demonstrations at SC2004 and iGrid 2005 can be determined in this meeting.

**Control Plane and Grid Integration**

The GLIF can only function if we agree on the interfaces and protocols that talk to each other in the control plane on the contributed Lambda resources. The main players in this field are already meeting, almost on a bi-monthly schedule. Although not essential, this GLIF meeting could also host a breakout session on control plane middleware.
GLIF - 5 meeting

- Collocated with iGrid2005 San Diego
- CAL-(IT)$^2$
- Thursday 29 sept 2005
  - Presentations track
- Friday 30 sept 2005
  - Work group meetings
- NOT on invitation only anymore!
  - Open meeting for participants
  - Industry rep’s only on workgroup chairs invitation (no marketing!)
• Optical networking architectures and models
  – Optical Internet Exchange architecture
  – Lambda routing and assignment

• IP transport protocols, performances monitoring and measurements
  – With respect to performance
  – Monitoring and reporting
  – Traffic generation with grid infrastructure

• Authorization, Authentication and Accounting
  – Concepts
  – Proof of concepts
  – Application
Protocol tests
TCP is bursty due to sliding window protocol and slow start algorithm.

$$\text{Window} = \text{BandWidth} \times \text{RTT} \quad \& \quad \text{BW} \Rightarrow \text{slow}$$

$$\frac{\text{fast} - \text{slow}}{\text{fast}} \times \text{slow} \times \text{RTT}$$

So pick from menu:
- Flow control
- Traffic Shaping
- RED (Random Early Discard)
- Self clocking in TCP
- Deep memory
Generic AAA server
Rule based engine

Application Specific Module

Service
Accounting Metering

Policy Data

API

Starting point

1

2

3

4

5

4'

RFC 2903 - 2906, 3334, policy draft
• finesse the control of bandwidth across multiple domains
• while exploiting scalability and intra-, inter-domain fault recovery
• thru layering of a novel SOA upon legacy control planes and NEs
Transport of flows

For what current Internet was designed

Needs more App & Middleware interaction

C

Full optical future

GLIF Future?

GLIF now

B

# FLOWS
Thanks to
SURFnet: Kees Neggers, UIC
KiCAIR: Tom DeFanti, Joel Mambretti, CANARIE: Bill St. Arnaud
Freek Hijestra, Hans Blom, Leon Gommans, Bas van oudenaarde, Arie Taal, Pieter de Beer, Bert Andree, Fred Wan,
Jeroen van der Ham, Karst Koymans, Paola Grosso, Yuri Demchenko, Rob Meijer, VL-team.

Not Quite END

Partially complete list:
Caas
Chase
Cess
Kess
Case