Lambda-Grid developments
Global Lambda Integrated Facility

www.science.uva.nl/~delaat

Cees de Laat

SURFnet

EU

University of Amsterdam
Sensor Grids

\textbf{eVLBI}

Longer term VLBI is easily capable of generating more 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.

\textit{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 2

Tier 2 Center

Tier 3

Institute
100 - 1000 Mbits/sec

Tier 4

Workstations

~0.25 TIPS Workstations

~100 MBytes/sec

~0.6-2.5 Gbps

~2.5 Gbits/sec

~PByte/sec

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

Physics data cache

1000 - 1000 Mbytes/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.
Data intensive scientific computation through global networks

Nuclear experiments
Belle Experiments

Very High-speed Network

Data Reservoir

Distributed Shared files

Local Accesses

Data Reservoir

Digital Sky Survey

SUBARU Telescope

Grape6

Data analysis at University of Tokyo
Co-located interactive 3D visualization

The markers are tracked by infrared cameras. The positions are transmitted to the visualization system. 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. The new image is transmitted to the display.

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

SGI Onyx4 at SARA

Pittsburgh

Amsterdam
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
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 2005)

- **Estimate A**
  - 17 M people, 6.4 M households, 25 % penetration of 0.5 - 8 Mb/s ADSL, 40 times under-provisioning ==> ~ 40 Gb/s
AMS-IX

European championship football  Holland -- Czech Republic

June 19th 2004  May 2005  Lost :-(

Lost :-(
The Dutch Situation (in 2005)

• **Estimate A**
  - 17 M people, 6.4 M households, 25 % penetration of 0.5 - 8 Mb/s ADSL, 40 times under-provisioning ==> ~ 40 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

Month & year

Tbyte per month

1900 TByte/month ≈ 6 Gbits/second

March 2005
1.9 TByte

Slide courtesy Kees Neggers
The Dutch Situation (in 2005)

- **Estimate A**
  - 17 M people, 6.4 M households, 25% penetration of 0.5 - 8 Mb/s ADSL, 40 times under-provisioning ==> ~ 40 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

- $\Sigma A \approx 40 \text{ Gb/s}$
- $\Sigma B \approx 30 \text{ Gb/s}$
- $\Sigma C >> 100 \text{ Gb/s}$
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!

L1 \approx 1 \text{k$/port}

L2 \approx 7-15 \text{k$/port}

L3 \approx 100+ \text{k$/port}
## Services

<table>
<thead>
<tr>
<th>SCALE</th>
<th>CLASS</th>
<th>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$</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Switches + E-WANPHY VPN’s</td>
<td>Switches + E-WANPHY (G)MPLS</td>
<td>ROUTER$</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>dark fiber DWDM MEMS switch</td>
<td>DWDM, TDM / SONET Lambda switching</td>
<td></td>
<td>Lambdas, VLAN’s SONET Ethernet</td>
</tr>
</tbody>
</table>
How low can you go?

Application Endpoint A

Router

Application Endpoint B

Ethernet

SONET

POS

MEMS

DWDM

Regional dark fiber

NetherLight

Fiber

GLIF

StarLight

15454 / 6500

Trans-Oceanic

Local Ethernet

173

MEMS

UKLight
Optical Exchange as Black Box

Optical Exchange

Switch
TDM
Store & Forward
DWDM mux/demux

TeraByte Email Service
NetherLight Q1 2005
GLIF History

- Brainstorming in Antalya at Terena conf. 2001
- 1st 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
- 3rd 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
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 and research

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!)
Discipline Networks

Internet

HEP

ASTRO

Earth Science

Lambda

Fibers
SURFnet 6 principles

• Based on dark fiber
• 4 DWDM rings of 9 bands
  – each 4, later 8, colors
  – Each capable of 10, later 40 Gb/s
• Universities have POP’s on ring, each 1 band
• Connect with 1 or 10 Gb/s Ethernet
• Routing in Amsterdam in 2 core POP’s!
• International connectivity in Amsterdam
• Lambda service between ring POP’s and to NetherLight
6000 km fiber
Iru for 15 years
3900 km railways
SURFnet on Lambda inspection in Science Park Amsterdam :-}
StarPlane
DWDM
backplane
GRID-Colocation problem space

CPU

DATA

Extensively under research

Lambda’s

New!
Research @ AIR

- 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
The Lab
Protocol tests
TCP is bursty due to sliding window protocol and slow start algorithm.

Window = BandWidth * RTT & BW == slow

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

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

RFCC 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
IXP series Network Processor Units

Features:

- The IXP 2850 is able to perform packet functions at 10 gb/s
- 16 programmable Micro Engines to allow parallel dataplane processing.
- Two crypto units support bulk security algorithms (AES, DES, 3DES, SHA1)
- Designed for IPSec, however is general enough to do other things.
- Supports Cypher Block Chaining in combination with MAC.
The IXP2850 network processor implements the same store-and-forward design as the IXP2800, including 16 multi-threaded microengines in the dataplane and a high-performance Intel XScale core for control plane functions. The IXP2850 adds two cryptography blocks.
Fundamental places for token switches

1. At network ingress point (switch, access server etc.) admitting only valid tokens.
2. Inside a network.
4. At network egress point (switch, gateway device)
Example experiment agent model
Transport of flows

For what current Internet was designed

Needs more App & Middleware interaction

Full optical future

GLIF now

GLIF Future?

# FLOWS
Not Quite END

Thanks to:

Kees Neggers, Tom DeFanti, Joel Mambretti, Bill St. Arnaud,
Freek Dijkstra, Hans Byron, Leon Gommans, Bas van Oudenaarde,
Arle Taal, Peter de Beer, Bert Andree, Fred Wan,
Jeroen van der Ham, Karst Koymans, Paola Grosso,
Yuri Demchenko, Rob Meijer, Vic Papam.

Partially complete list:

Caas, Chase, Cess, Case, Cece, Case, Cas, Cass.