Lambda-Grid developments

www.science.uva.nl/~delaat

Cees de Laat

GigaPort

EU

University of Amsterdam

SARA
Contents of this talk

- This page is intentionally left blank
  - Ref: www.this-page-intentionally-left-blank.org
Four LHC Experiments: The Petabyte to Exabyte Challenge

ATLAS, CMS, ALICE, LHCB

- Tens of PB 2008; To 1 EB by ~2015
- Hundreds of TFlops To PetaFlops

6000+ Physicists & Engineers; 60+ Countries; 250 Institutions
CMS detector: 15m X 15m X 22m
12,500 tons, $700M.

Physics data cache

 tier: 1
 Italian Regional Center
 German Regional Center
 NIKHEF Dutch Regional Center
 FermiLab, USA Regional Center
 Tier 2
 Tier 3
 Institute Institute Institute
 ~0.25TIPS
 100 - 1000 Mbits/sec
 Workstations

Tier 4

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.
In the longer term, VLBI is easily capable of generating many Gb of data per hour. The sensitivity of the VLBI array scales with the data rate (data-rate) and there is a strong push to increase data rates. Rates of 8Gb/s or more are entirely feasible under development. It is expected that parallel correlator will remain the most efficient approach. Distributed processing may have an application. Multi-gigabit data streams will aggregate into labor and the capacity of the final link to the data center.

Westerbork Synthesis Radio Telescope - Netherlands
Lambdas as part of instruments

www.lofar.org

1 - 45 Tbit/s,
http://www.lofar.org/p/systems.htm
http://web.haystack.mit.edu/lofar/technical.html
Showed you:

• **Computational Grids**
  – HEP and LOFAR analysis requires massive CPU capacity

• **Data Grid**
  – Storing and moving HEP, Bio and Health data sets is major challenge

• **Instrumentation Grids**
  – Several massive data sources are coming online

• **Visualization Grids**
  – Data object (TByte sized) inspection, anywhere, anytime
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
The Dutch Situation

- **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

European championship football  Holland -- Czech Republic

June 19th 2004

Lost :-(

Bits per second

Maximal In: 27.792 G  Maximal Out: 27.794 G
Current In: 27.239 G  Current Out: 27.239 G
The Dutch Situation

- **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**
  - SURFnet has 10 Gb/s to about 12 institutes and 0.1 to 1 Gb/s to 180 customers, estimate same for industry (overestimation) ==> 20-40 Gb/s
The Dutch Situation

- **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**
  
  - SURFnet has 10 Gb/s to about 12 institutes and 0.1 to 1 Gb/s to 180 customers, estimate same for industry (overestimation) ==> 20-40 Gb/s

- **Estimate C**
  
  - Leading HEF and ASTRO + rest ==> 80-120 Gb/s
  
  - LOFAR ==> ≈ 26 Tbit/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. Special scientific applications, computing, data grids, virtual-presence
   Need very fat pipes, limited multiple Virtual Organizations, few to few

\[ \Sigma A \approx 20 \text{ Gb/s} \]

\[ \Sigma B \approx 40 \text{ Gb/s} \]

\[ \Sigma C >> 100 \text{ Gb/s} \]
λ’s on scale 2-20-200 ms rtt
So what?

- Costs of optical equipment 10% of switching 10% of full routing equipment for same throughput
  - 10G routerblade -> 100-300 k$, 10G switch port -> 10-20 k$, MEMS port -> 0.7 k$
  - DWDM lasers for long reach expensive, 10-50k$ (???)
  - 64 Byte packet @ 10 Gbit/s -> 52 ns -> time to look up destination in 140 kEntries routing table (light speed from me to you (15 meter)!)  
- Bottom line: look for a hybrid architecture which serves all classes in a cost effective way ( A -> L3 , B -> L2 , C -> L1)
- Give each packet in the network the service it needs, but no more
UVA/EVL’s 64*64 Optical Switch @ NetherLight in SURFnet POP @ SARA

Costs 1/100th of a similar throughput router or 1/10th of an Ethernet switch but with specific services!
<table>
<thead>
<tr>
<th>CLASS</th>
<th>SCALE</th>
<th>2 Metro</th>
<th>20 National/regional</th>
<th>200 World</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>AMSIX</td>
<td>Switching/routing</td>
<td>Routing</td>
<td>ROUTER$</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>Switches + ETH-WANPHY VPN’s</td>
<td>Switches + ETH-WANPHY (G)MPLS</td>
<td>ROUTER$</td>
</tr>
<tr>
<td>C</td>
<td>NetherLight</td>
<td>dark fiber DWDM MEMS switch</td>
<td>DWDM, TDM / SONET Lambda switching</td>
<td>Lambdas, VLAN’s SONET Ethernet</td>
</tr>
</tbody>
</table>
- **lambda for high bandwidth applications**
  - Bypass of production network
  - Middleware may request (optical) pipe
- **RATIONALE:**
  - Lower the cost of transport per packet
  - Use Internet as controlplane!
How low can you go?
Architectures - L1 - L3

Bring plumbing to the users, not just create sinks in the middle of nowhere
Optical Exchange as Black Box

Optical Exchange

Switch

TDM

Store & Forward

DWDM mux/demux

TeraByte Email Service
### Service Matrix

<table>
<thead>
<tr>
<th>From</th>
<th>To (multiple λ)</th>
<th>Single λ, any bitstream</th>
<th>SONET/SDH</th>
<th>1 Gb/s Ethernet</th>
<th>LAN PHY Ethernet</th>
<th>WAN PHY Ethernet</th>
<th>VLAN tagged Ethernet</th>
<th>IP over Ethernet</th>
</tr>
</thead>
<tbody>
<tr>
<td>WDM (multiple λ)</td>
<td>cross-connect</td>
<td>WDM demux</td>
<td>WDM demux*</td>
<td>WDM demux</td>
<td>WDM demux*</td>
<td>WDM demux*</td>
<td>WDM demux*</td>
<td>WDM demux*</td>
</tr>
<tr>
<td></td>
<td>multicast,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>regenerate,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>multicast</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single λ, any bitstream</td>
<td>WDM mux</td>
<td>cross-connect</td>
<td>N/A*</td>
<td>N/A*</td>
<td>N/A*</td>
<td>N/A*</td>
<td>N/A*</td>
<td>N/A*</td>
</tr>
<tr>
<td></td>
<td>multicast,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>regenerate,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>multicast</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SONET/SDH</td>
<td>WDM mux</td>
<td>N/A*</td>
<td>SONET switch, +</td>
<td>TDM demux*</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>TDM demux*</td>
<td>TDM demux*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ethernet</td>
<td>Ethernet</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>conversion +</td>
<td>conversion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Gb/s Ethernet</td>
<td>WDM mux</td>
<td>N/A*</td>
<td>TDM mux</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>L3 entry*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ethernet</td>
<td>Ethernet</td>
<td>VLAN decap</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>conversion +</td>
<td>conversion</td>
<td>encap</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAN PHY Ethernet</td>
<td>WDM mux</td>
<td>N/A*</td>
<td>TDM mux6</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>L3 entry*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ethernet</td>
<td>Ethernet</td>
<td>VLAN decap</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>conversion +</td>
<td>conversion</td>
<td>encap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WAN PHY Ethernet</td>
<td>WDM mux</td>
<td>N/A*</td>
<td>SONET switch</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>L3 entry*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ethernet</td>
<td>Ethernet</td>
<td>VLAN decap</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>conversion +</td>
<td>conversion</td>
<td>encap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VLAN tagged Ethernet</td>
<td>WDM mux</td>
<td>N/A*</td>
<td>TDM mux</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>aggregate,</td>
<td>Aggregate,</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VLAN decap</td>
<td>VLAN decap</td>
<td>VLAN decap</td>
<td>VLAN decap &amp; encap +</td>
<td></td>
</tr>
<tr>
<td>IP over Ethernet</td>
<td>WDM mux</td>
<td>N/A*</td>
<td>TDM mux</td>
<td>L3 exit*</td>
<td>L3 exit*</td>
<td>L3 exit*</td>
<td>N/A</td>
<td>Store &amp; forward,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>L3 entry/exit+</td>
</tr>
</tbody>
</table>
SURFnet fibers
(pict outdated anytime ;-)

SURFnet6 entirely based on own dark fiber
Over 5300 km fiber pairs available today; average price paid for 15 year IRUs:
< 6 EUR/meter per pair
SURFnet on inspection in Science Park Amsterdam :-)
GLIF: Global Lambda Integrated Facility

- Established at the 3rd Lambda Grid Workshop, August 2003 in Reykjavik, Iceland
- Collaborative initiative among worldwide NRENs, institutions and their users
- A world-scale Lambda-based Laboratory for application and middleware development

GLIF vision: To build a new grid-computing paradigm, in which the central architectural element is optical networks, not computers, to support this decade’s most demanding e-science applications.

Coordinated by UvA, SURFnet and UIC
GLIF Q3 2004

Visualization courtesy of Bob Patterson, NCSA.
Light Paths provisioning implementation

International Light Path connectivity

10 GE

10 GE LAN

10 GE LAN

10 GE

SURFnet6 Sites in Amsterdam

16x16 MEMS

16x16 MEMS

Nortel OME 6500

Nortel OME 6500

Nortel OME 6500

GigaPort

Optical Switch

SURFnet6 Common Photonic Layer

Regional Light Path

Customer equipment

SURFnet infrastructure

Non-SURFnet

Nortel OME 6500

16x16 MEMS

Customer equipment
Research topics

- **Optical** networking architectures and models for usage
- Transport protocols for massive amounts of data
- Authorization of complex resources in multiple domains
- Embedding in Grid environments
Example Measurements
AAA based demo at SC2003

Calient provided by EVL
Conclusions

• Demanding applications
  • (Science) data repositories mirroring
  • Instrumentation grids
  • Visualisation and collaboration support

• Model of Lambda networking
  • Identify traffic types
  • Scales of infrastructure
  • Map efficiently to lower the cost/packet

• Current experiments
  • NetherLight
  • VLE/eScience Amsterdam
  • Networking research
    (control plane, transport protocols, optical net models)
Transport in the corners

BW*RTT

For what current Internet was designed

Needs more App & Middleware interaction

Full optical future

# FLOWS
Not quite The END

Thanks to

SURFnet: Kees Neggers, UIC&iCAIR: Tom DeFanti, Joel Mambretti, CANARIE: Bill St. Arnaud
Freek Dijkstra, Hans Blom, Leon Gommans, Bas van oudenaarde, Arie Taal, Pieter de Boer, Bert Andree, Martijn de Munnik, Antony Antony, Rob Meijer, VL-team.

Partially complete list:

Caas
Chase
Cess
Kess
Case
The END

Thanks to
SURFnet: Kees Neggers, UIC&iCAIR: Tom DeFanti, Joel Mambretti, CANARIE: Bill St. Arnaud
Freek Dijkstra, Hans Blom, Leon Gommans, Bas van oudenaarde, Arie Taal, Pieter de Boer, Bert Andree, Martijn de Munnik, Antony Antony, Rob Meijer, VL-team.

Partially complete list:
Caas
Chase
Cess
Kess
Case
[1957-2004]