Lambda-Grid developments
Global Lambda Integrated Facility

www.science.uva.nl/~delaat

Cees de Laat

GigaPort
EU

University of Amsterdam
SARA
NCF
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
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

Tier 2

Tier 3
Institute

Tier 4

Tier 2 Center

Physics data cache
Workstations

~0.6-2.5 Gbps
~0.25TIPS
100 - 1000 Mbits/sec

~PByte/sec

~100 MB/sec

~2.5 Gbits/sec

~1000 Mbits/sec

~0.6-2.5 Gbps

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.

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

Tier 2

Tier 3
Institute

Tier 4

Tier 2 Center

Physics data cache
Workstations

~0.6-2.5 Gbps
~0.25TIPS
100 - 1000 Mbits/sec

~PByte/sec

~100 MB/sec

~2.5 Gbits/sec

~1000 Mbits/sec

~0.6-2.5 Gbps

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.

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

Tier 2

Tier 3
Institute

Tier 4

Tier 2 Center

Physics data cache
Workstations

~0.6-2.5 Gbps
~0.25TIPS
100 - 1000 Mbits/sec

~PByte/sec

~100 MB/sec

~2.5 Gbits/sec

~1000 Mbits/sec

~0.6-2.5 Gbps

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.
VLBI

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 and there is a strong push to increase the data rate. Rates of 8Gb/s or more are entirely feasible under development. It is expected that parallel multi-gigabit correlators will remain the most efficient approach. Distributed processing may have an application. Multi-gigabit data streams will aggregate into larger correlators and the capacity of the final link to the data collector.

Westerbork Synthesis Radio Telescope - Netherlands
Lambdas as part of instruments

www.lofar.org

37 Tbit/s - 116 Tops/s
http://www.lofar.org/p/systems.htm
http://web.haystack.mit.edu/lofar/technical.html

GigaPort
Data intensive scientific computation through global networks

Nuclear experiments
Belle Experiments

Data Reservoir

Very High-speed Network

Nobeyama Radio Observatory (VLBI)
X-ray astronomy Satellite ASUKA

Digital Sky Survey

SUBARU Telescope

Grape6

Data analysis at University of Tokyo

Data Reservoir

Local Accesses
Distributed Shared files
Co-located interactive 3D visualization

10 Gigabit/s path on the SURFnet and Abilene networks

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.

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
Grids

Showed you:

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

• **Data Grids**
  – 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. Scientific applications, distributed data processing, all sorts of grids
   Need very fat pipes, limited multiple Virtual Organizations, few to few, p2p
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 :-(
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**
  - 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

\[ \Sigma A \approx 20 \text{ Gb/s} \]
\[ \Sigma B \approx 30 \text{ Gb/s} \]
\[ \Sigma C >> 100 \text{ Gb/s} \]
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 -> 100-500 k$, 10G switch port -> 10-20 k$, MEMS port -> 0.7 k$
  – DWDM lasers for long reach expensive, 10-50k$

• 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 - 0.7 k$/port
L2 - 10-20 k$/port
L3 - 100-500 k$/port
<table>
<thead>
<tr>
<th>SCALE</th>
<th>CLASS</th>
<th>Services</th>
</tr>
</thead>
</table>
|       | A     | Switching/routing  
|       | B     | Switches + E-WANPHY VPN’s  
|       | C     | dark fiber DWDM MEMS switch  
|       |       | DWDM, TDM / SONET Lambda switching  
|       |       | Lambda, VLAN’s SONET Ethernet  
|       |       | 20 National/regional  
|       |       | 200 World  
|       |       | ROUTERS$  
|       |       | ROUTERS$  
|       |       | ROUTERS$  

**Services**

- **Lambda**, **VLAN’s**, **SONET**, **Ethernet**
- **DWDM**, **TDM**, **MEMS switch**
- **Routing**, **Switches + E-WANPHY**, **(G)MPLS**
- **Router**, **Switches + E-WANPHY**, **VPN’s**
- **200 World**, **20 National/regional**, **2 Metro**
- **SCALE**
- **CLASS**
How low can you go?
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</th>
<th>WDM (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 multicast, regenerate, multicast</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>
<td>WDM demux*</td>
</tr>
<tr>
<td>Single λ, any bitstream</td>
<td>WDM mux</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>
<td>N/A *</td>
<td>N/A *</td>
</tr>
<tr>
<td>SONET/SDH</td>
<td>WDM mux</td>
<td>N/A *</td>
<td>SONET switch, +</td>
<td>TDM demux*</td>
<td>TDM demux*</td>
<td>TDM demux*</td>
<td>TDM demux*</td>
<td>TDM demux*</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, Ethernet conversion +</td>
<td>aggregate, Ethernet conversion</td>
<td>aggregate, Ethernet conversion</td>
<td>aggregate, VLAN encap</td>
<td>L3 entry *</td>
<td></td>
</tr>
<tr>
<td>LAN PHY Ethernet</td>
<td>WDM mux</td>
<td>N/A *</td>
<td>TDM mux*</td>
<td>aggregate, Ethernet conversion</td>
<td>aggregate, Ethernet conversion +</td>
<td>Ethernet conversion</td>
<td>aggregate, VLAN encap</td>
<td>L3 entry *</td>
<td></td>
</tr>
<tr>
<td>WAN PHY Ethernet</td>
<td>WDM mux</td>
<td>N/A *</td>
<td>SONET switch</td>
<td>aggregate, Ethernet conversion</td>
<td>Ethernet conversion</td>
<td>aggregate, Ethernet conversion +</td>
<td>aggregate, VLAN encap</td>
<td>L3 entry *</td>
<td></td>
</tr>
<tr>
<td>VLAN tagged Ethernet</td>
<td>WDM mux</td>
<td>N/A *</td>
<td>TDM mux</td>
<td>aggregate, VLAN decap</td>
<td>aggregate, VLAN decap</td>
<td>aggregate, VLAN decap</td>
<td>Aggregate, VLAN decap &amp; encap +</td>
<td>N/A</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, L3 entry/exit+</td>
<td></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 Lambda inspection in Science Park Amsterdam :-}
UCLP intended for projects like National LambdaRail

CAVEwave partner acquires a separate wavelength between San Diego and Chicago and wants to manage it as part of its network including add/drop, routing, partition etc.
UltraLight Network: PHASE III

- Move into production
- Optical switching fully enabled amongst primary sites
- Integrated international infrastructure
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:

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!
History of GLIF

• **Brainstorming in Antalya at Terena conf. 2001**
  – On invitation only (15) + public part
  – Thinking, SURFnet test lambda Starlight-Netherlight

• **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 -> www.glif.is

• **4th at Nottingham 3 Sept 2004 hosted by UKERNA collocated UK-eScience**
  – preparatory afternoon on 2 September
  – 60 participants
  – Attendance from China, Japan, Netherlands, Switzerland, US, UK, Taiwan, Australia, Tsjech, Korea, Canada, Ireland, Russia, Belgium, Denmark
  – Meeting of GOV, TEC and APP groups
Research on Networks (CdL)

• **Optical Networking:**
  - What innovation in architectural models, components, control and light path provisioning are needed to integrate dynamically configurable optical transport networks and traditional IP networks to a generic data transport platform that provides end-to-end IP connectivity as well as light path (lambda and sub-lambda) services?

• **High performance routing and switching:**
  - What developments need to be made in the Internet Protocol Suite to support data intensive applications, and scale the routing and addressing capabilities to meet the demands of the research and higher education communities in the forthcoming 5 years?

• **Management and monitoring:**
  - What management and monitoring models on the dynamic hybrid network infrastructure are suited to provide the necessary high level information to support network planning, network security and network management?

• **Grids and access; reaching out to the user:**
  - What new models, interfaces and protocols are capable of empowering the (grid) user to access, and the provider to offer, the network and grid resources in a uniform manner as tools for scientific research?

• **Testing methodology:**
  - What are efficient and effective methods and setups to test the capabilities and performance of the new building blocks and their interworking, needed for a correct functioning of a next generation network?
Research topics AIR @ UvA

- **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
LightHouse

Diagram showing network configurations and equipment such as routers, switches, and optical cross-connects.
Example Measurements
Layer - 2 requirements from 3/4

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
Starting point

Generic AAA server
Rule based engine

Application Specific Module

Service
Accounting Metering

Policy
Data

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
DAS3
DWDM
backplane

UvA-VLE
UvA-MM
VU
ULeiden
TUDelft
Transport in the corners

BW*RTT

C

Needs more App & Middleware interaction

Full optical future

For what current Internet was designed

A

# FLOWS
Revisiting the truck of tapes

Consider one fiber

• Current technology allows $320\lambda$ in one of the frequency bands

• Each $\lambda$ has a bandwidth of 40 Gbit/s

• Transport: $320 \times 40 \times 10^9 / 8 = 1600$ GByte/sec

• Take a 10 metric ton truck

• One tape contains 50 Gbyte, weights 100 gr

• Truck contains $(10000 / 0.1) \times 50$ Gbyte = 5 PByte

• Truck / fiber = 5 PByte / 1600 GByte/sec = 3125 s ≈ one hour

• For distances further away than a truck drives in one hour (50 km) minus loading and handling 100000 tapes the fiber wins!!!
Not quite END

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

Partially complete list:
Caas
Chase
Cess
Kess
Case