Lambda-Grid developments
Global Lambda Integrated Facility

www.science.uva.nl/~delaat

Cees de Laat

GigaPort
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 8 Gbit/s or more. The sensitivity of the VLBI array scales with bandwidth (=data-rate) and there is a strong push to modems. Rates of 8 Gbit/s or more are entirely feasible. It is expected that parallelized correlator will remain the most efficient approach to under development. It is expected that parallelized correlator will remain the most efficient approach for solving distributed multi-gigabit correlator and scaling factor.
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
Tier 1

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

Tier 2

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.

Tier 3

Tier 4

Physics data cache

Workstations

~0.6-2.5 Gbps

~0.6-2.5 TIPS

~2.5 Gbits/sec

~100 MBytes/sec

~PByte/sec

Online System

event reconstruction

event simulation

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

Nuclear experiments

Belle Experiments

Data Reservoir

Very High-speed Network

Data Reservoir

Data analysis at University of Tokyo

Local Accesses

Distributed Shared files

Digital Sky Survey

SUBARU Telescope

Grape6

Nobeyama Radio Observatory (VLBI)

X-ray astronomy Satellite ASUKA

Nobeyama Radio Observatory (VLBI)

Digital Sky Survey

SUBARU Telescope

Grape6

Data Reservoir

Data Reservoir

Data Reservoir

Data Reservoir
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.
Pittsburgh
Amsterdam
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
StarPlane
DWDM
backplane
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

• 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
European championship football  Holland -- Czech Republic

June 19th 2004

Lost :-(

AMS-IX
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

ΣA ≈ 20 Gb/s

ΣB ≈ 30 Gb/s

ΣC >> 100 Gb/s

BW requirements
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, 10-50k$

- 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 - 1 k$/port  
L2 - 7-15 k$/port  
L3 - 100-500 k$/port
## 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$</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Switches + E-WANPHY VPN’s</td>
<td>Switches + E-WANPHY (G)MPLS</td>
<td>ROUTER$</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>dark fiber DWDM MEMS switch</td>
<td>DWDM, TDM / SONET Lambda switching</td>
<td>Lambdas, VLAN’s SONET Ethernet</td>
<td></td>
</tr>
</tbody>
</table>
How low can you go?

Application Endpoint A

Router
Ethernet
SONET
DWDM
Fiber

Application Endpoint B

Local Ethernet
POS
MEMS
Regional dark fiber

15454 / 6500

Trans-Oceanic

UKLight
GLIF
StarLight
Optical Exchange as Black Box

Optical Exchange

- Switch
- TDM
- Store & Forward
- DWDM mux/demux

TeraByte Email Service
Common Photonic Layer (CPL) in SURFnet6
Laying of fiber near/at Science Park Amsterdam

Pictures by Yuri Demchenko
SURFnet on Lambda inspection in Science Park Amsterdam :-)

[Image of three people outdoors, one pointing and two others conversing]
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.
GLIF Q4 2004

Visualization courtesy of Bob Patterson, NCSA.
Discipline Networks

- Internet
- HEP
- ASTRO
- Earth Science
- Fibers
- Lambdas
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?
Advanced Internet Research Group @ UvA

- 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
  - Network & Grid integration and Applications
Protocol tests
TCP is bursty due to sliding window protocol and slow start algorithm.

Window = BandWidth * RTT & BW == slow

Memory-at-bottleneck = \frac{\text{fast} - \text{slow}}{\text{fast}} * \text{slow} * \text{RTT}

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

API

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

GLIF now

For what current Internet was designed

Full optical future

GLIF Future?

Needs more App & Middleware interaction

B

C

A

# FLOWS

BW

RTT
Open research questions

- LightPath into organizations like ours
- Sub-second true Lambda provisioning
- Massive data transport, storage and handout
- Scheduling of resources for workflows
- Complex authorization
- Brokering
- Security
iGrid 2005 was held at Science park Amsterdam

September 26-30, 2005
University of California, San Diego
California Institute for Telecommunications and Information Technology [Cal-(IT)^2]
United States