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 a 40 Tbit/s system. The sensitivity of the VLBI array scales with data rate, and there is a strong push to modems. Rates of 8Gb/s or more are entirely feasible.

Westerbork Synthesis Radio Telescope - Netherlands
Four LHC Experiments: The Petabyte to Exabyte Challenge

ATLAS, CMS, ALICE, LHCB

6000+ Physicists & Engineers; 60+ Countries; 250 Institutions

Tens of PB 2008; To 1 EB by ~2015

Hundreds of TFlops To PetaFlops
Tier 1

Italian Regional Center

German Regional Center

NIKHEF Dutch Regional Center

Fermilab, USA Regional Center

Tier 0 +1

event reconstruction

~0.6-2.5 Gbps

Tier 2

Tier 3

Tier 4

Physics data cache

Workstations

~0.25 TIPS

~100 MBytes/sec

~2.5 Gbits/sec

event simulation

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 4

Institute

~0.25 TIPS

~0.6-2.5 Gbps

analysis

human=2m

~PByte/sec

Online System

~100 MBytes/sec

Tier 1

Tier 0

CMS detector: 15m X 15m X 22m

12,500 tons, $700M.

~2.5 Gbits/sec

CMS as example, Atlas is similar

~0.6-2.5 Gbps

Tier 2

Tier 3

Tier 4

Institute

~0.25 TIPS

~100 - 1000 Mbits/sec

Workstations

Courtesy Harvey Newman, CalTech and CERN

1000 tarns, $700M.
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
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
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
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
AMS-IX

June 19th 2004

European championship football  Holland -- Czech Republic

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
Routed L3 traffic growth

SURFnet customer traffic: Monthly volume

1600 Tbyte/month ≈ 5 Gbits/second

Slide courtesy Kees Neggers
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
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!

\[
\begin{align*}
L1 & \approx 1 \text{k$/port} \\
L2 & \approx 7-15 \text{k$/port} \\
L3 & \approx 100+ \text{k$/port}
\end{align*}
\]
<table>
<thead>
<tr>
<th>SCALE</th>
<th>CLASS</th>
<th>Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>Metro</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Switching/routing</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>National/regional</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routing</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>World</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ROUTER$</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>Switches + E-WANPHY VPN’s</td>
</tr>
<tr>
<td></td>
<td>Switches + E-WANPHY (G)MPLS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ROUTER$</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>dark fiber</td>
</tr>
<tr>
<td></td>
<td>DWDM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MEMS switch</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DWDM, TDM / SONET</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lambda switching</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lambda, VLAN’s SONET Ethernet</td>
<td></td>
</tr>
</tbody>
</table>
How low can you go?

Application Endpoint A
Application Endpoint B

Router
Ethernet
SONET
DWDM
Fiber

Local Ethernet
MEMS
POS
Regional dark fiber

15454 / 6500

Trans-Oceanic
StarLight

NetherLight
UKLight
GLIF
Optical Exchange as Black Box

Optical Exchange

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

TeraByte Email Service
Laying of fiber near/at Science Park
Amsterdam

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

[Image of three individuals engaged in a discussion with one pointing out into the distance, surrounded by greenery.]
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

Lambdas

Fibers
Protocol tests
TCP is bursty due to sliding window protocol and slow start algorithm.

Window = BandWidth * RTT & BW == slow

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

So pick from menu:

- Flow control
- Traffic Shaping
- RED (Random Early Discard)
- Self clocking in TCP
- Deep memory
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

GLIF now

Full optical future

GLIF Future?

# 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.

Partially complete list:

Caas
Chase
Cess
Kess
Case