StarPlane DWDM backplane
For the Netherlands 2007:
\[ \Sigma A = \Sigma B = \Sigma C \approx 250 \text{ Gb/s} \]
However:
- A -> Typical d2d Internet
- B -> VO’s, Grid
- C -> just a few (CG, LHC, eVLBI)
Towards Hybrid Networking!

• Costs of photonic equipment 10% of switching 10% of full routing
  – for same throughput!
  – Photonic vs Optical (optical used for SONET, etc, 10-50 k$/port)
  – DWDM lasers for long reach expensive, 10-50 k$

• 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 ≈ 0.5-1.5 k$/port
L2 ≈ 5-8 k$/port
L3 ≈ 75+ k$/port
How low can you go?

Application Endpoint A

Router

Ethernet

SONET

DWDM

Fiber

Local Ethernet

POS

MEMS

Country dark fiber StarPlane

15454 6500 HDXc

Trans-Oceanic

NetherLight

UKLight

GLIF

Application Endpoint B

StarPlane

NetherLight

UKLight

GLIF
<table>
<thead>
<tr>
<th>SCALE</th>
<th>Metro Country (2 ms RTT)</th>
<th>Regional Continental (20 ms RTT)</th>
<th>World Trans Ocean (200 ms RTT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASS</td>
<td>Switching/Routing</td>
<td>Routers</td>
<td>ROUTER$</td>
</tr>
<tr>
<td>A</td>
<td>Switches VPN’s E-WANPHY</td>
<td>Routing Switches (G)MPLS E-WANPHY</td>
<td>ROUTER$</td>
</tr>
<tr>
<td>B</td>
<td>dark fiber, DWDM, WSS, Photonic switch</td>
<td>DWDM, TDM / SONET Lambda switching</td>
<td>VLAN’s TDM SONET Ethernet</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCALE</td>
<td>CLASS</td>
<td>Metro Country</td>
<td>Regional Continental</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>---------------</td>
<td>----------------------</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>Switching/ Routing</td>
<td>Routers</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>Switching/ E-WANPHY</td>
<td>Lambda switching</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>Photonic switch</td>
<td>DWDM, TDM / SONET</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RTT</th>
<th>SCALE</th>
<th>CLASS</th>
<th>Metro Country</th>
<th>Regional Continental</th>
<th>World Trans Ocean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 ms RTT</td>
<td>A</td>
<td></td>
<td>Switching/ Routing</td>
<td>Routers</td>
<td>ROUTER$</td>
</tr>
<tr>
<td>20 ms RTT</td>
<td>B</td>
<td></td>
<td>Switching/ E-WANPHY</td>
<td>Lambda switching</td>
<td>ROUTER$</td>
</tr>
<tr>
<td>200 ms RTT</td>
<td>C</td>
<td></td>
<td>Photonic switch</td>
<td>DWDM, TDM / SONET</td>
<td>VLAN’s Ethernet</td>
</tr>
</tbody>
</table>
QOS in a non destructive way!

• Destructive QOS:
  – have a link or $\lambda$
  – set part of it aside for a lucky few under higher priority
  – rest gets less service

• Constructive QOS:
  – have a $\lambda$
  – add other $\lambda$‘s as needed on separate colors
  – move the lucky ones over there
  – rest gets also a bit happier!
The StarPlane vision is to give flexibility directly to the applications by allowing them to choose the logical topology in real time, ultimately with sub-second lambda switching times on part of the SURFnet6 infrastructure.
The challenge for sub-second switching

- bringing up/down a λ takes minutes
  - this was fast in the era of old time signaling (phone/fax)
  - λ 2 λ influence (Amplifiers, non linear effects)
  - however minutes is historically grown, 5 nines, up for years
  - working with Nortel to get setup time significantly down

- plan B:

  University

  RB+WS+AAA → Myrinet switch → WS+AAA → MEMS switch → WS+AAA

  SURFnet

  WS+AAA NOC
DAS-3 Cluster Architecture

- **85 (40+45) compute nodes**
- **1 Gb/s Ethernet**
- **Local interconnect**
- **10 Gb/s Ethernet lanphy**
- **Fast interconnect**
- **To local University**
- **To SURFnet**
- **8 * 10 Gb/s from bridgenodes**
Power is a big issue

- UvA cluster uses (max) 30 kWh
- 1 kWh ~ 0.1 €

- per year \(\rightarrow\) 26 k€/y
- add cooling 50\% \(\rightarrow\) 39 k€/y
- Emergency power system \(\rightarrow\) 50 k€/y
- per rack 10 kWh is now normal

**YOU BURN ABOUT HALF THE CLUSTER OVER ITS LIFETIME!**

- Terminating a 10 Gb/s wave costs about 200 W
- Entire loaded fiber \(\rightarrow\) 16 kW
- Wavelength Selective Switch : few W!
Overview Net Tests between DAS-3 Hosts

- Authorize here to store the current table settings in your cookies file.
- See the getting started introduction or the user guide for a description of the table below.
- See also the hosts documentation.
- Some observations about the package and the required bandwidth.

Select ping value: min, avg, max, all, last.
Select UDP value: rtt, lost.

DAS-3 Net Test Results
Date: 31/05/2007
Time: 12:30:01

Load

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>VU-083</td>
<td>0</td>
<td>0</td>
<td>0.087</td>
<td>0</td>
<td>0.01</td>
<td>0.017</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>YU-085</td>
<td>0</td>
<td>0</td>
<td>0.087</td>
<td>0.013</td>
<td>0.01</td>
<td>0.017</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>LIACS-125</td>
<td>1.380</td>
<td>---</td>
<td>1.380</td>
<td>---</td>
<td>1.230</td>
<td>1.230</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>LIACS-127</td>
<td>---</td>
<td>1.380</td>
<td>---</td>
<td>1.230</td>
<td>---</td>
<td>1.230</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>UvA-236</td>
<td>---</td>
<td>---</td>
<td>1.380</td>
<td>---</td>
<td>1.230</td>
<td>1.230</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>UvA-239</td>
<td>---</td>
<td>---</td>
<td>1.380</td>
<td>---</td>
<td>1.230</td>
<td>1.230</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>UvA-236-M</td>
<td>---</td>
<td>---</td>
<td>1.380</td>
<td>---</td>
<td>1.230</td>
<td>1.230</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>UvA-239-M</td>
<td>---</td>
<td>---</td>
<td>1.380</td>
<td>---</td>
<td>1.230</td>
<td>1.230</td>
<td>0.025</td>
<td></td>
</tr>
</tbody>
</table>

Ping Min [ms]

Throughput [Mbit/s]
Very constant and predictable!
CineGrid

DAS-3 - 4U set @UvA

Rembrandt Cluster
total 22 TByte diskspace @ LightHouse
Opteron 64 bit nodes

DP AMD processor nodes

10 Gbit/s

head node (?)
comp node
comp node
comp node
comp node
comp node
comp node
comp node

32-77x

comp node

10 Gbit/s

NetherLight, StarPlane
the cp testbeds
and beyond

CALIENT
mems switch

Glimmer-Glass
mems switch

NORTEL
8600
L2/3 switch

head node
comp node
comp node
comp node
comp node
comp node
comp node
comp node

F10
L2/3 switch
RDF describing Infrastructure

Application: find video containing x, then trans-code to it view on Tiled Display

- RDF/CG
- RDF/CG
- RDF/ST
- RDF/ST
- RDF/NL
- RDF/NL
- RDF/NDL
- RDF/NDL
- RDF/VIZ
- RDF/VIZ
- RDF/CPU
- RDF/CPU

PG&CdL

StarPlane
Phosphorus AAA testbed

International Optical Connectivity: Netherlight GLIF

MEMS Switch

Domain A

Domain B

Domain C

Service Plane AAA Node

Control Plane Node

802.1Q Switch

Public Internet

802.1Q Switch

Control Plane Node
The SCARle project

SCARle: a research project to create a Software Correlator for e-VLBI.
VLBI Correlation: signal processing technique to get high precision image from spatially distributed radio-telescope.

To equal the hardware correlator we need:

- 16 streams of 1Gbps
- 16 * 1Gbps of data
- 2 Tflops CPU power

\[
\frac{2 \text{ TFlop}}{16 \text{ Gbps}} = 1000 \text{ flops/byte}
\]

THIS IS A DATA FLOW PROBLEM !!!
Tera-Thinking

• What constitutes a Tb/s network?
• 128 times 10 Gbit/s between renderer and tiled display?
• CALIT2 has 8000 Gigabit drops ?->? Terabit Lan?
• think back to teraflop computing!
  – MPI makes it a teraflop machine

• TeraApps programming model supported by
  – TFlops  ->  MPI / Globus
  – TBytes  ->  OGSA/DAIS
  – TPixels ->  SAGE
  – TSensors ->  LOFAR, LHC, LOOKING, CineGrid, ...
  – Tbit/s  ->  ?

ref Larry Smarr & CdL
Need for discrete parallelism

- it takes a core to receive 1 or 10 Gbit/s in a computer
- it takes one or two cores to deal with 10 Gbit/s storage
- same for Gigapixels
- same for 100’s of Gflops
- Capacity of every part in a system seems of same scale
  - look at 80 core Intel processor
    - cut it in two, left and right communicate 8 TB/s
- massive parallel channels in hosts, NIC’s
- Therefore we need to go massively parallel allocating complete parts for the problem at hand!
User Programmable Virtualized Networks allows the results of decades of computer science to handle the complexities of application specific networking.

- The network is virtualized as a collection of resources
- UPVN's enable network resources to be programmed as part of the application
- Mathematica, a powerful mathematical software system, can interact with real networks using UPVN's
Mathematica enables advanced graph queries, visualizations and real-time network manipulations on UPVN.

Topologies matters can be dealt with algorithmically. Results can be persisted using a transaction service built in UPVN.

Initialization and BFS discovery of NEs

```
Needs["WebServices"]
<< DiscreteMath`Combinatorica`
<< DiscreteMath`GraphPlot
InitNetworkTopologyService["edge.ict.tno.nl"]

Available methods:
{DiscoverNetworkElements, GetLinkBandwidth, GetAllIpLinks, Remote, NetworkTokenTransaction}
```

Global`upvnverbose = True;
AbsoluteTiming[nes = BFSDiscover["139.63.145.94"];];[[1]]
AbsoluteTiming[result = BFSDiscoverLinks["139.63.145.94", nes];];[[1]]

Getting neighbours of: 139.63.145.94
Internal links: {192.168.0.1, 139.63.145.94}
(...)
Getting neighbours of: 192.168.2.3
Internal links: {192.168.2.3}

Transaction on shortest path with tokens

```
nodePath = ConvertIndicesToNodes[ShortestPath[g, Node2Index[nids,"192.168.3.4"], Node2Index[nids,"139.63.77.49"]], nids];

Print["Path: ", nodePath];
If[NetworkTokenTransaction[nodePath, "green"]==True,
Print["Committed"], Print["Transaction failed"];]
```

Path: 
{192.168.3.4,192.168.3.1,139.63.77.30,139.63.77.49}

Committed

ref: Robert J. Meijer, Rudolf J. Strijkers, Leon Gommans, Cees de Laat, User Programmable Virtualized Networks, accepted for publication to the IEEE e-Science 2006 conference Amsterdam.
Walking the Line

SURFnet

Lambda’s fibers

NorduNet
I did not talk about:
AAA & TBN
Security
Grid, workflow
etc.etc.

Questions?