StarPlane

StarPlane - Lambda Network under Control of Grid Applications

Cees de Laat

SURFnet

BSIK

EU

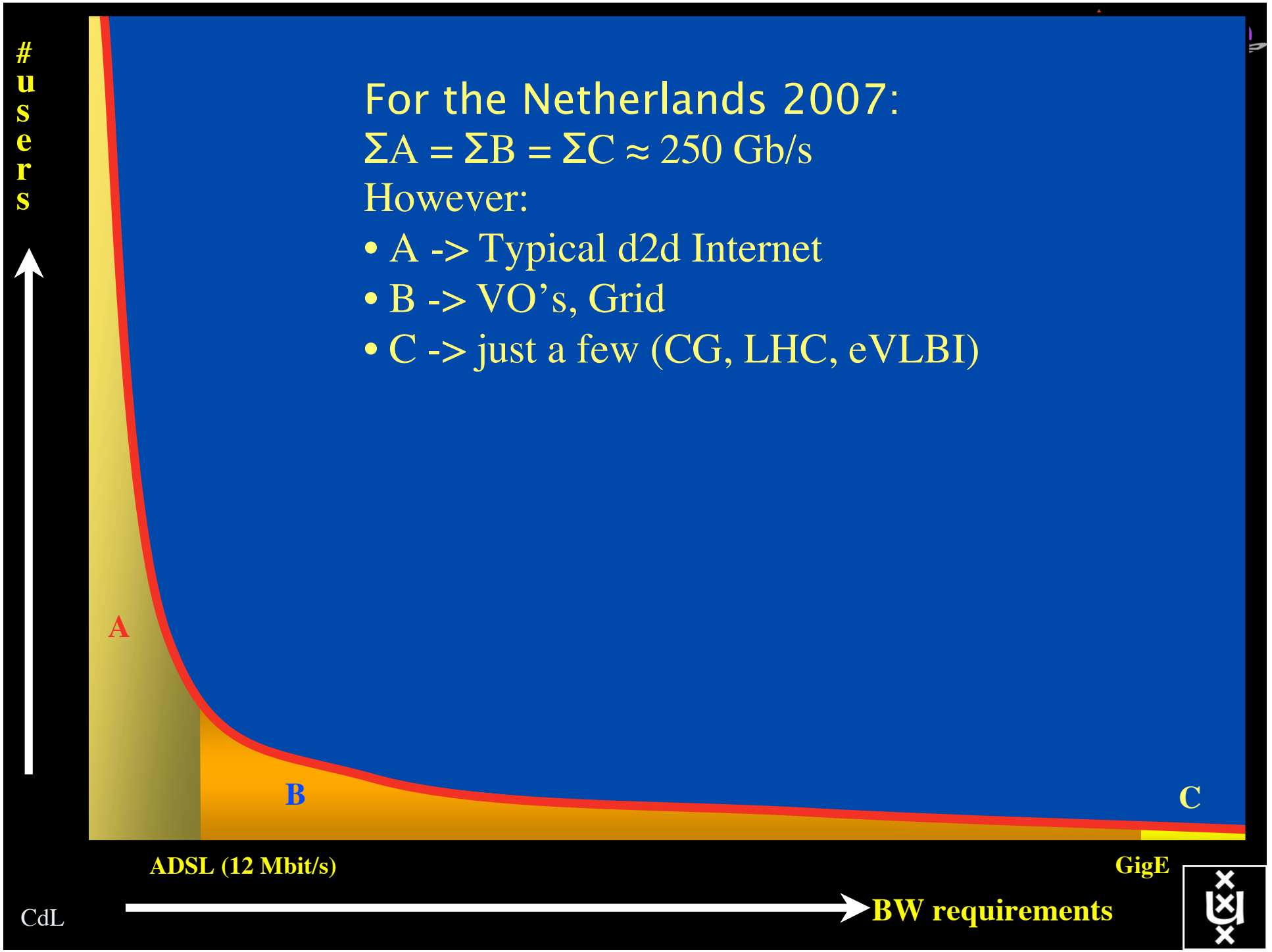
NWO

University of Amsterdam

TNO
NCF

SURF NET

DAS-3


#users
For the Netherlands 2007:
ΣA = ΣB = ΣC ≈ 250 Gb/s
However:
• A -> Typical d2d Internet
• B -> VO's, Grid
• C -> just a few (CG, LHC, eVLBI)
A
B
C
ADSL (12 Mbit/s)
GigE
BW requirements
CdL

Towards Hybrid Networking!

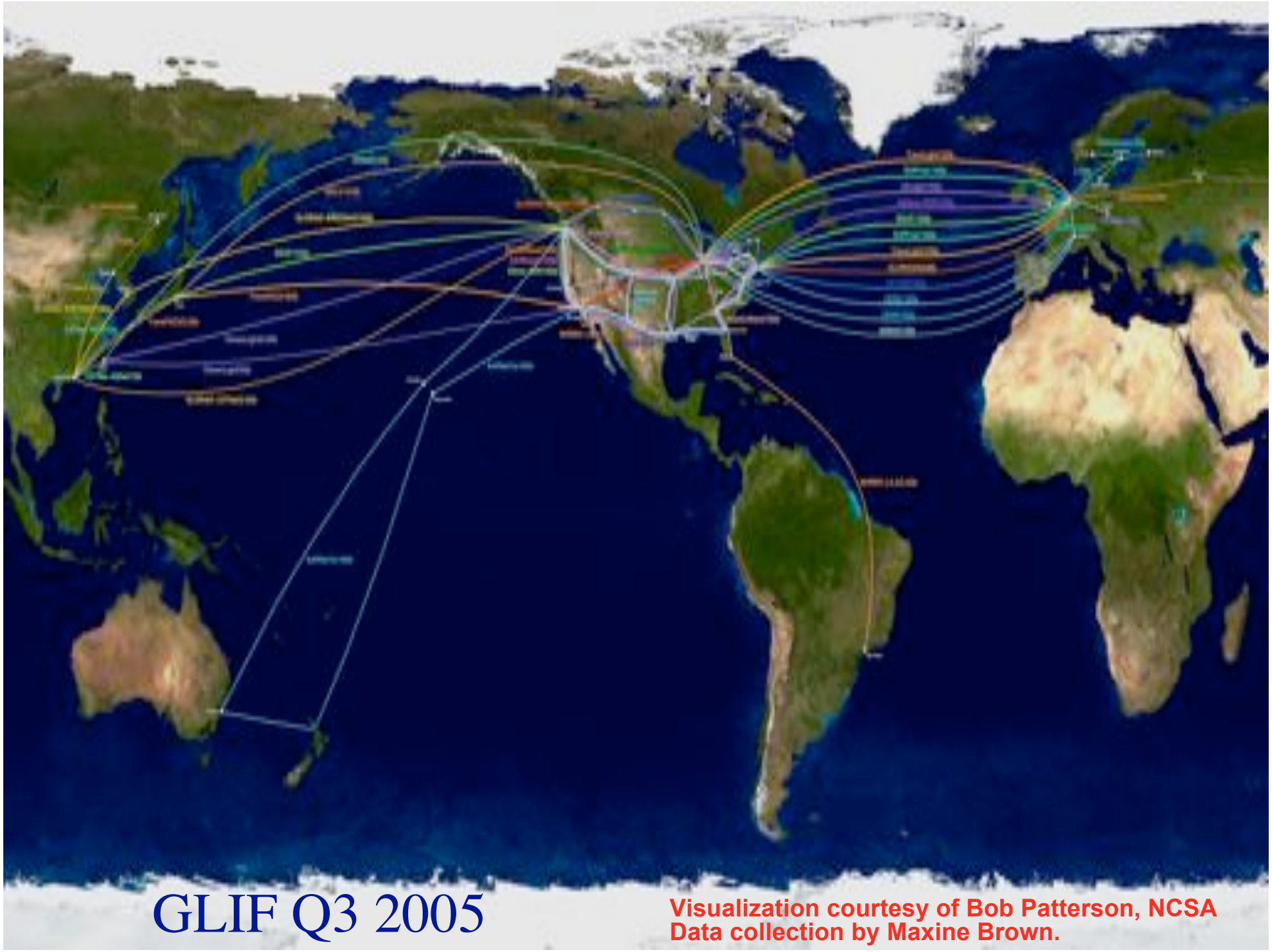
StarPlane

- 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


GLIF Q3 2005
Visualization courtesy of Bob Patterson, NCSA
Data collection by Maxine Brown.

Infrastructure Flexibility & Functionality

StarPlane

SCALE / CLASS	Metro Country 2 ms RTT	Regional Continental 2 ms RTT	World Trans Ocean 2 ms RTT
A	Switching/ Routing	Routers	ROUTER$
B	Switches VPN's E-WANPHY	Routing Switches (G)MPLS E-WANPHY	ROUTER$
C	dark fiber DWDM WSS Photonic switch	DWDM, TDM / SONET Lambda switching	VLAN's TDM SONET Ethernet

StarPlane

In The Netherlands SURFnet connects between 180:

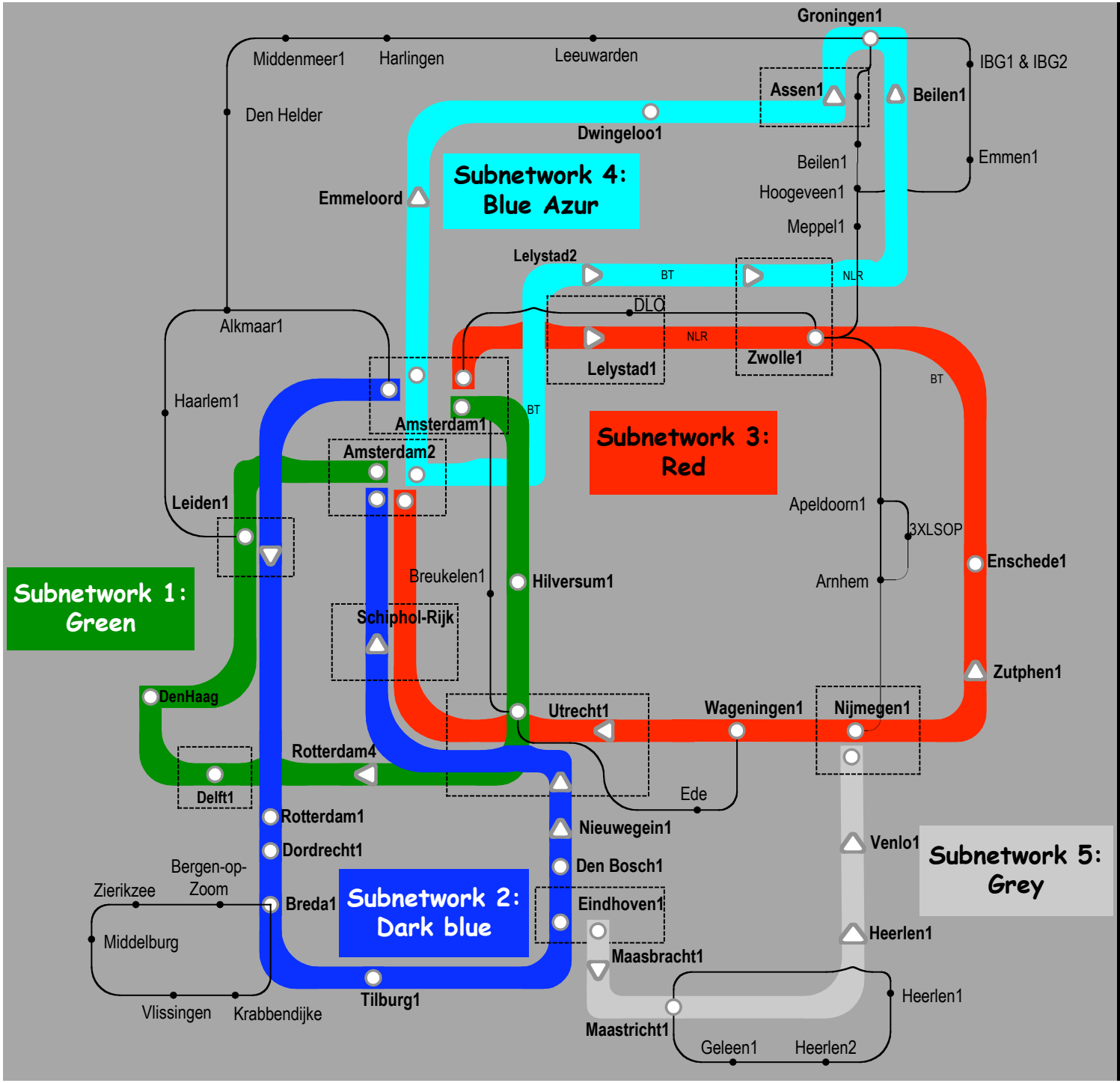
- universities;
- academic hospitals;
- most polytechnics;
- research centers.

with an indirect ~750K user base

~ 6000 km

scale comparable to railway system

SURFNET



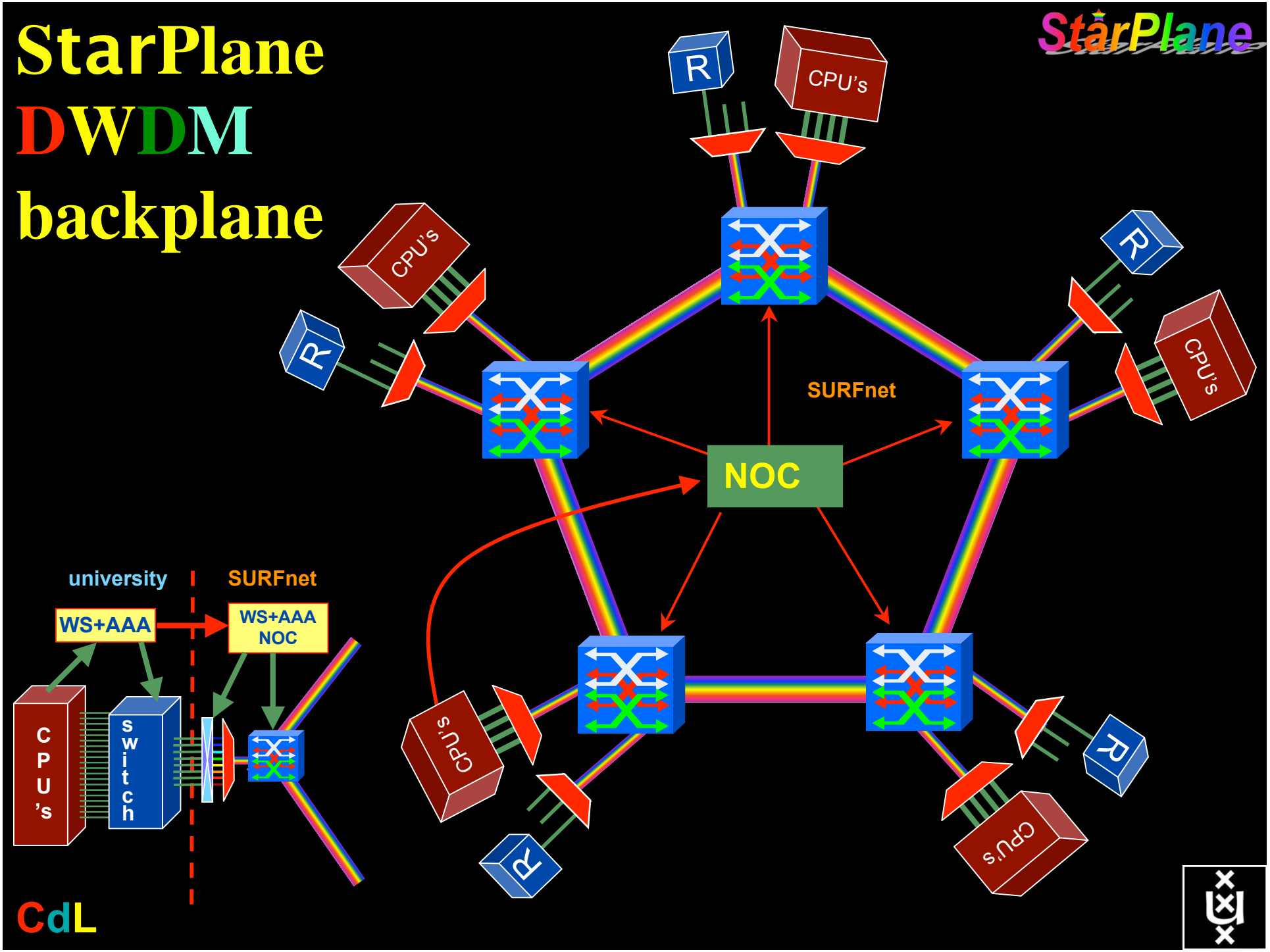
StarPlane

Common Photonic Layer (CPL) in SURFnet6

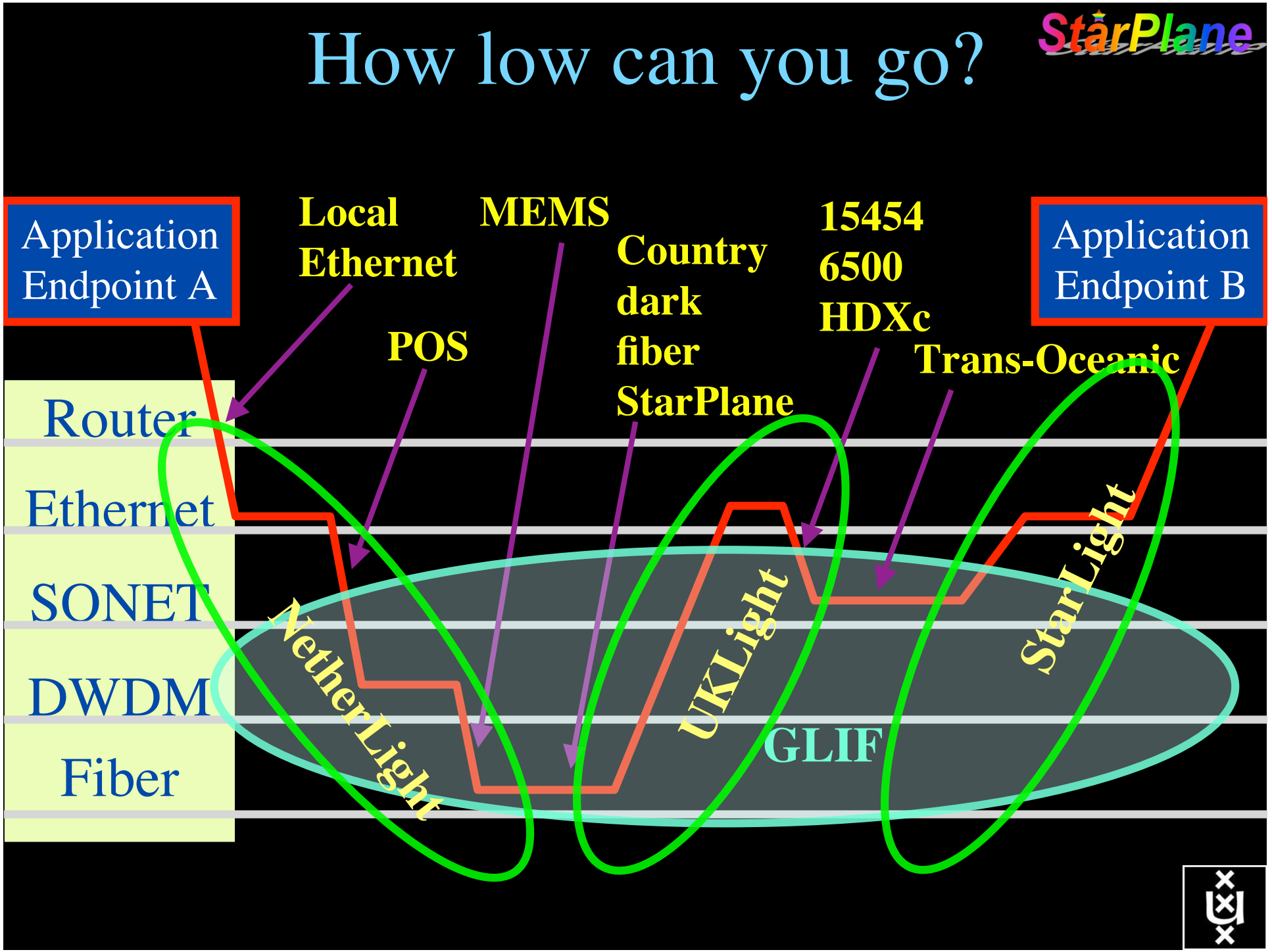
Supports up to 72 Lambda's of 10 Gb/s each future: 40/100 Gb/s

SURFNET

StarPlane DWDM backplane

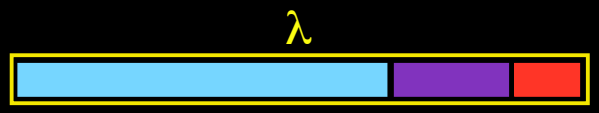


How low can you go?

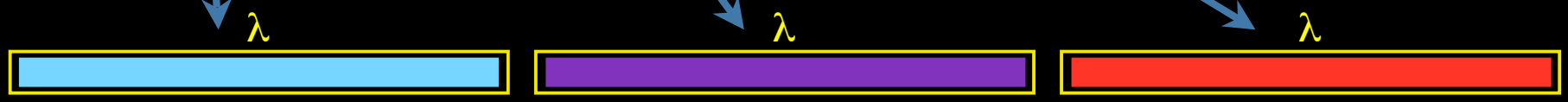


QOS in a non destructive way!

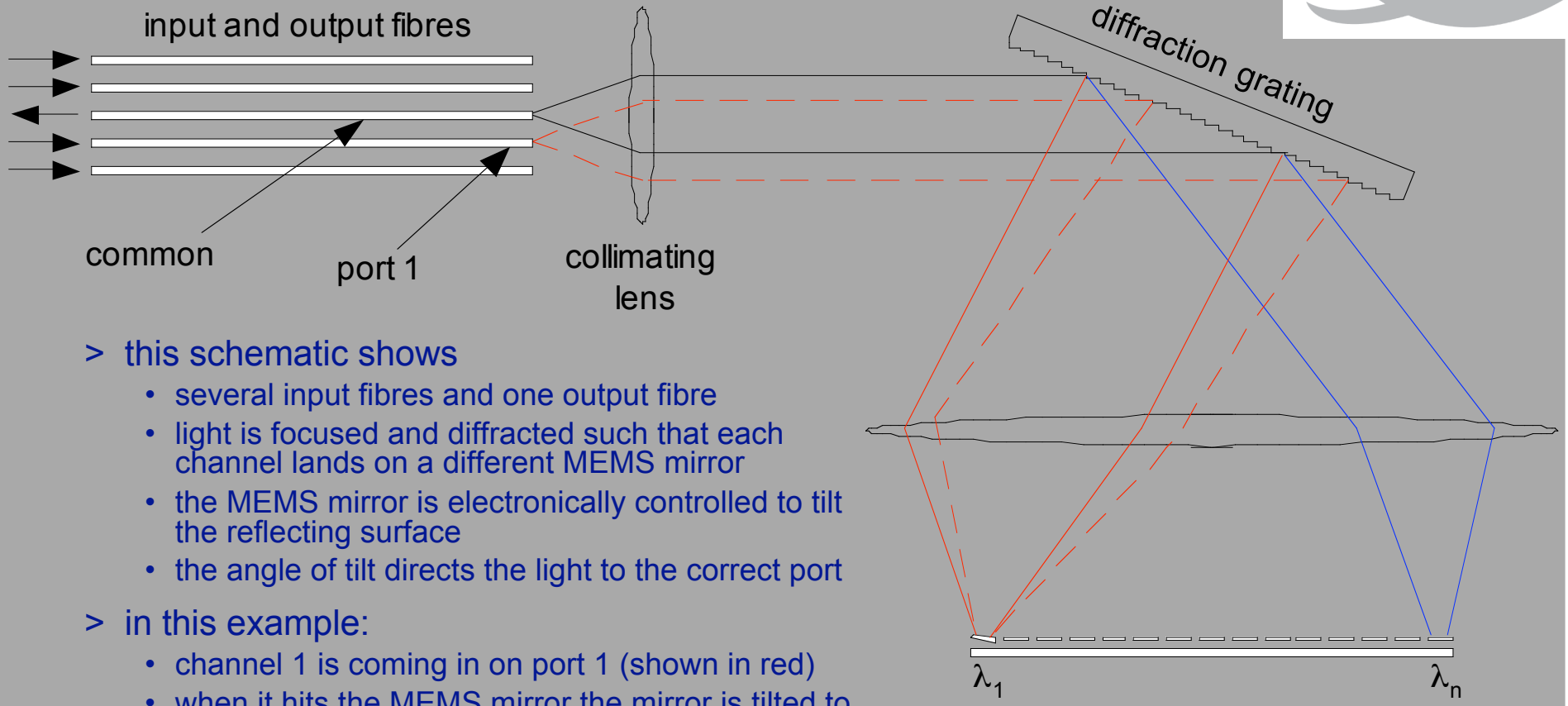
- Destructive QOS:
  - have a link or λ
  - set part of it aside for a lucky few under higher priority
  - rest gets less service



- Constructive QOS:
  - have a λ
  - add other λ's as needed on separate colors
  - move the lucky ones over there
  - rest gets also a bit happier!



Module Operation



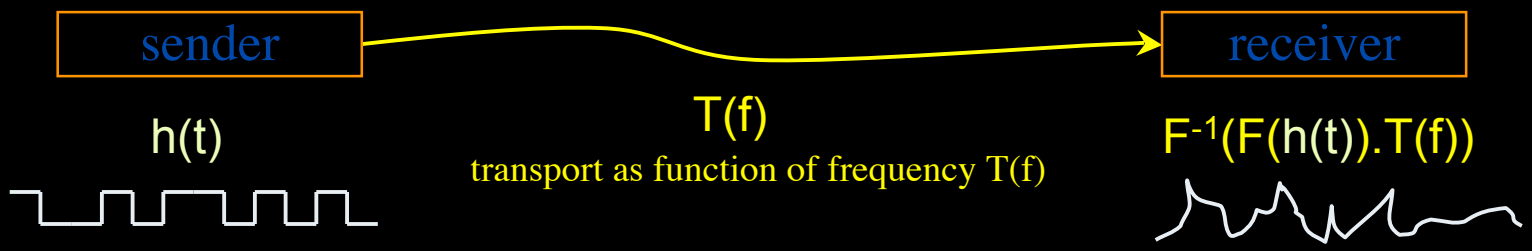
> this schematic shows
  - several input fibres and one output fibre
  - light is focused and diffracted such that each channel lands on a different MEMS mirror
  - the MEMS mirror is electronically controlled to tilt the reflecting surface
  - the angle of tilt directs the light to the correct port

> in this example:
  - channel 1 is coming in on port 1 (shown in red)
  - when it hits the MEMS mirror the mirror is tilted to direct this channel from port 1 to the common
  - only port 1 satisfies this angle, therefore all other ports are blocked

ref Eric Bernier, NORTEL

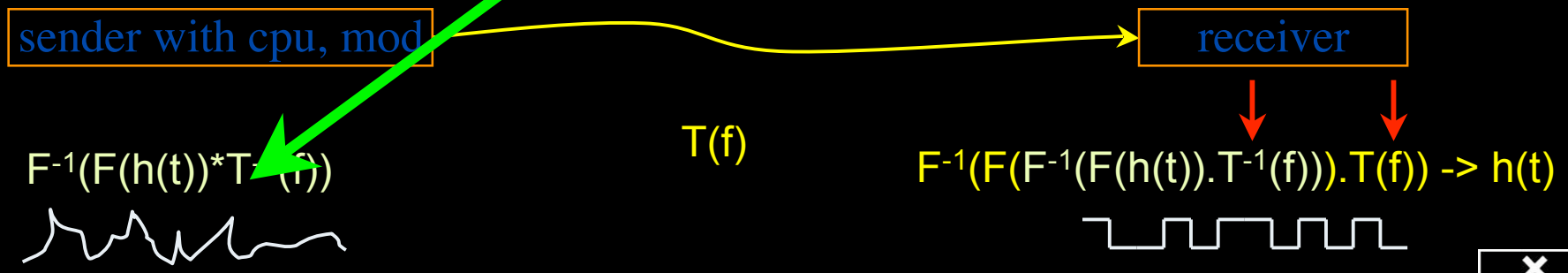
Dispersion compensating modem: eDCO from NORTEL

StarPlane



Solution in 5 easy steps for dummy's :

1. try to figure out T(f) by trial and error
2. invert T(f) -> $T^{-1}(f)$
3. computationally multiply $T^{-1}(f)$ with Fourier transform of bit pattern to send
4. inverse Fourier transform the result from frequency to time space
5. modulate laser with resulting $h'(t) = F^{-1}(F(h(t)).T^{-1}(f))$

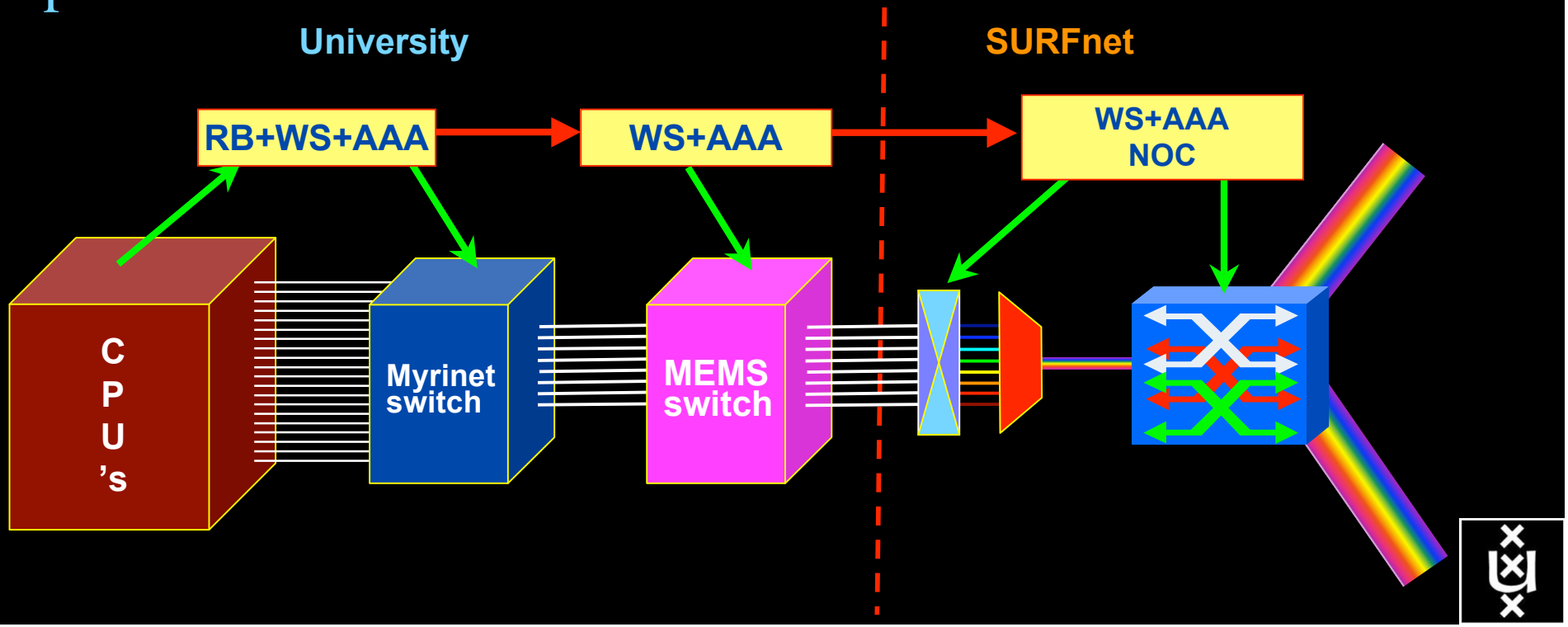


(ps. due to power ~ square E the signal to send **looks** like uncompensated received but is not)

The challenge for sub-second switching

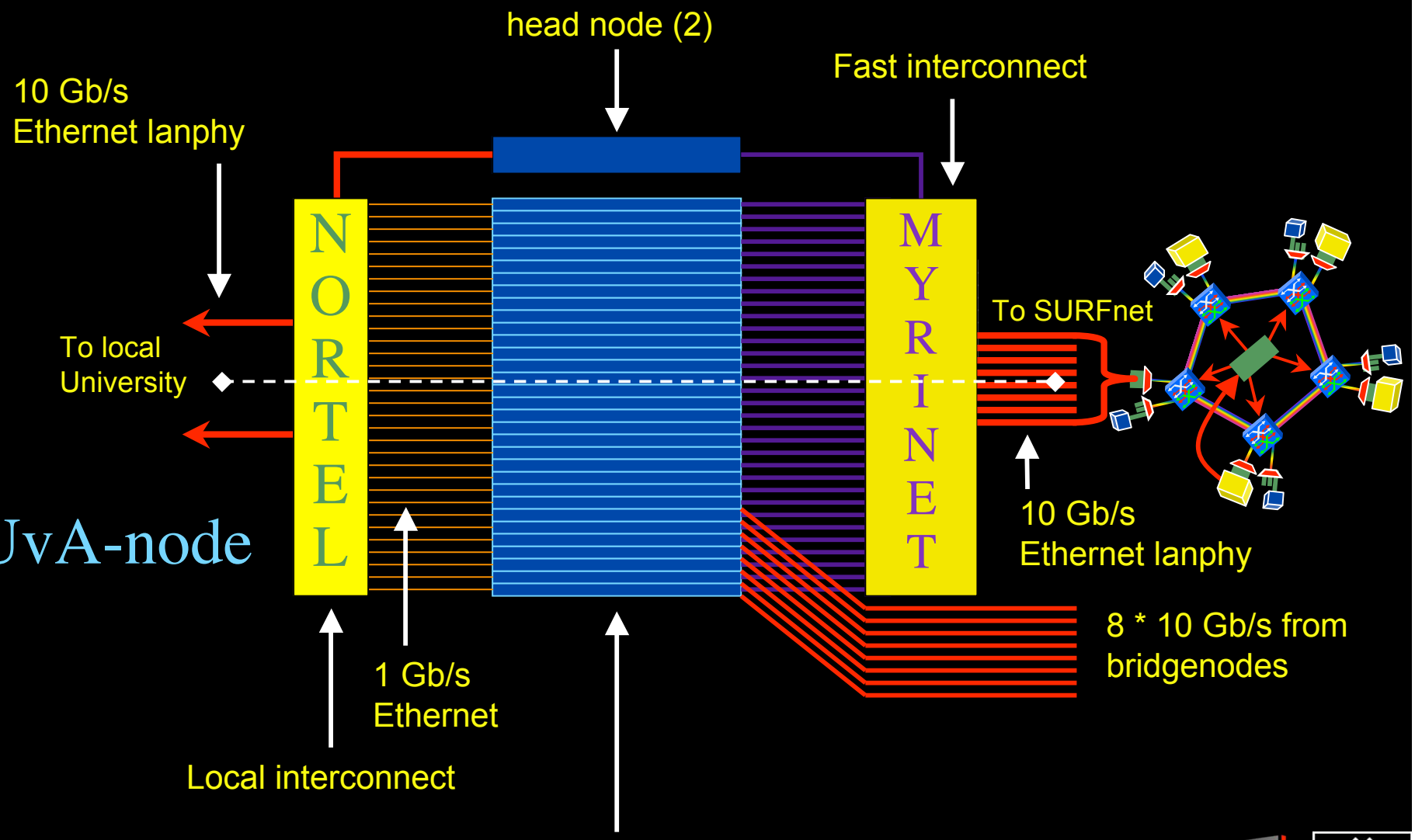
StarPlane

- 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:



DAS-3 Cluster Architecture

StarPlane



Advanced School for Computing and Imaging

DAS-3

Power is a big issue

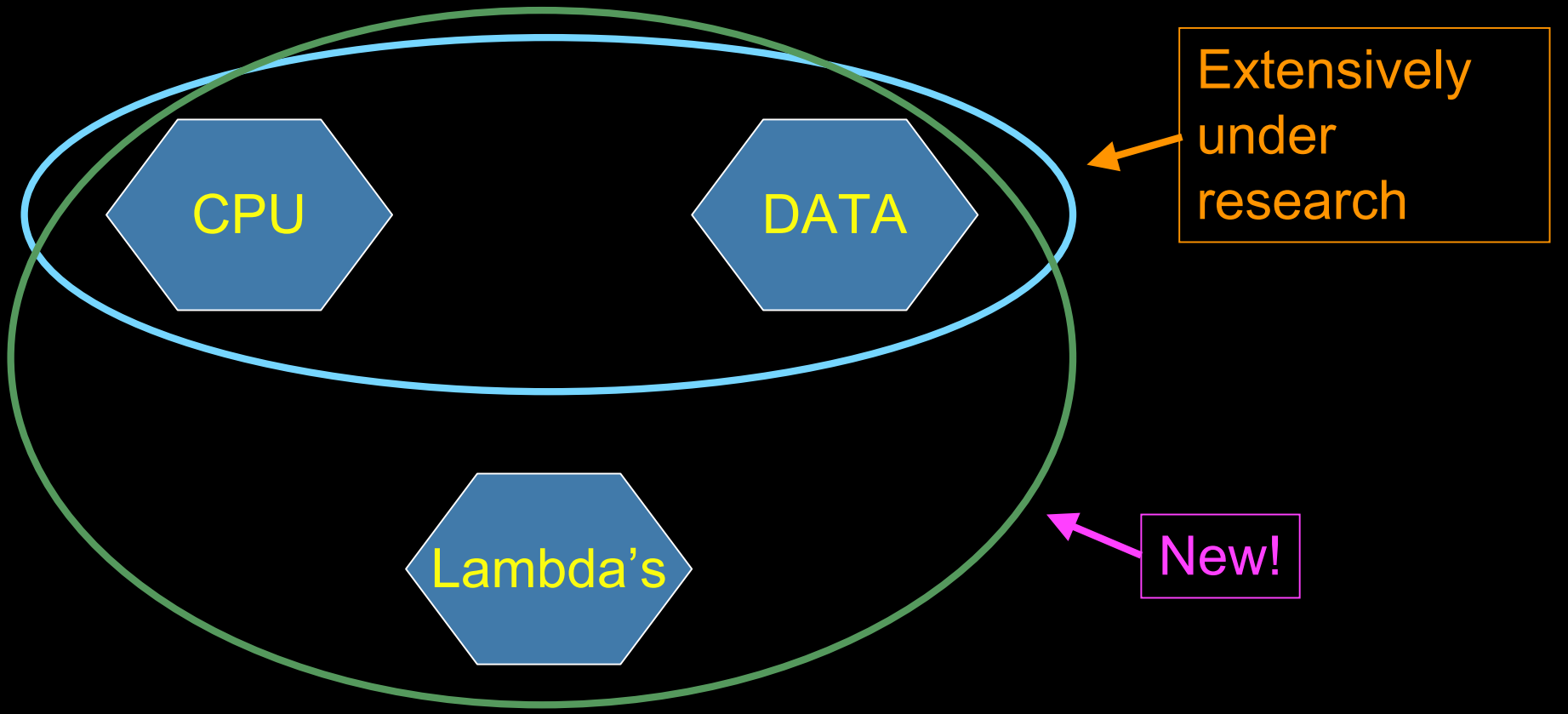
StarPlane

- UvA cluster uses (max) 30 kWh
- 1 kWh ~ 0.1 €
- per year -> 26 k€/y
- add cooling 50% -> 39 k€/y
- Emergency power system -> 50 k€/y
- per rack 10 kWh is now normal
- **YOU BURN ABOUT HALF THE CLUSTER OVER ITS LIFETIME!**

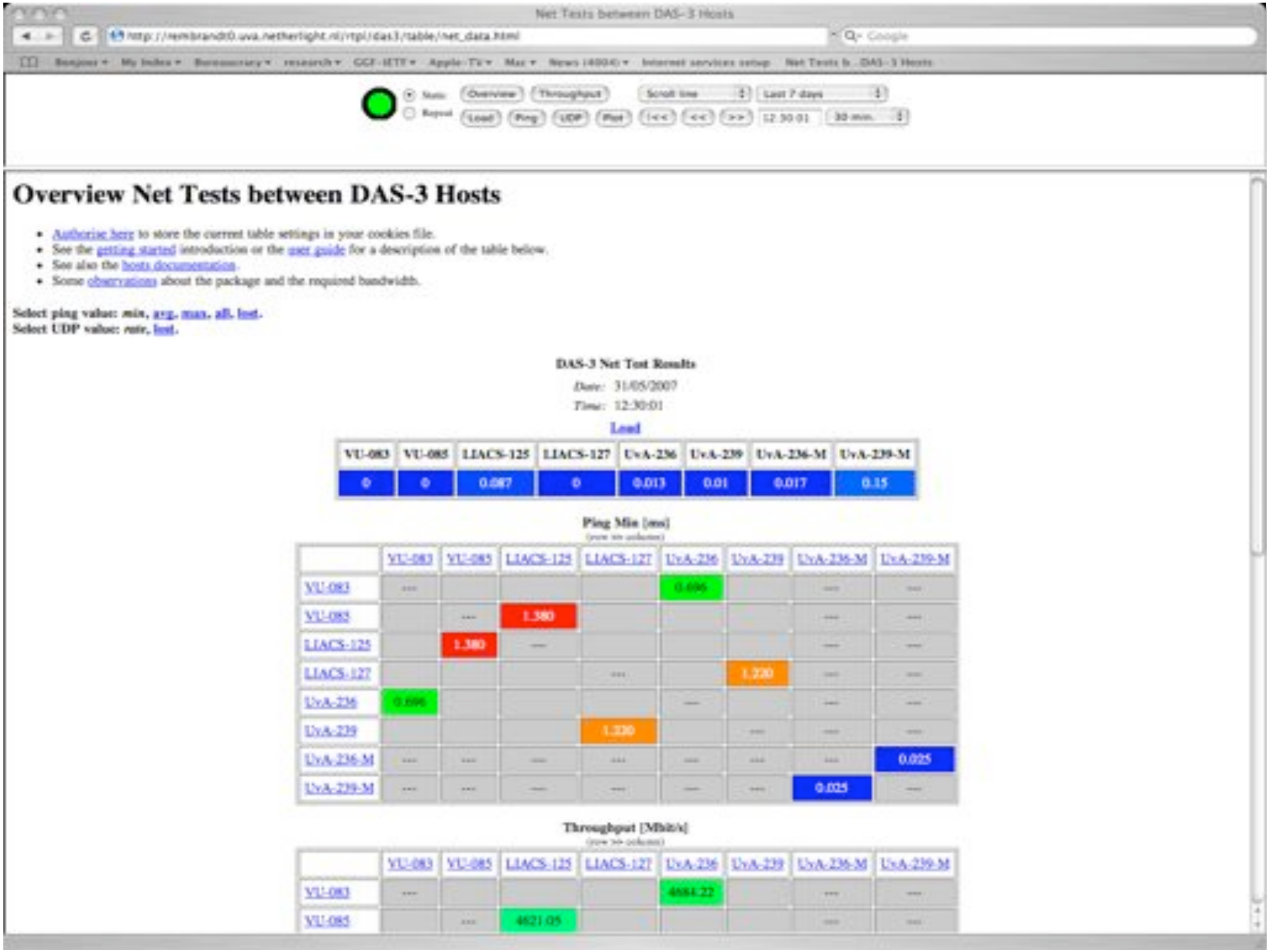
- Terminating a 10 Gb/s wave costs about 200 W
- Entire loaded fiber -> 16 kW
- Wavelength Selective Switch : few W!

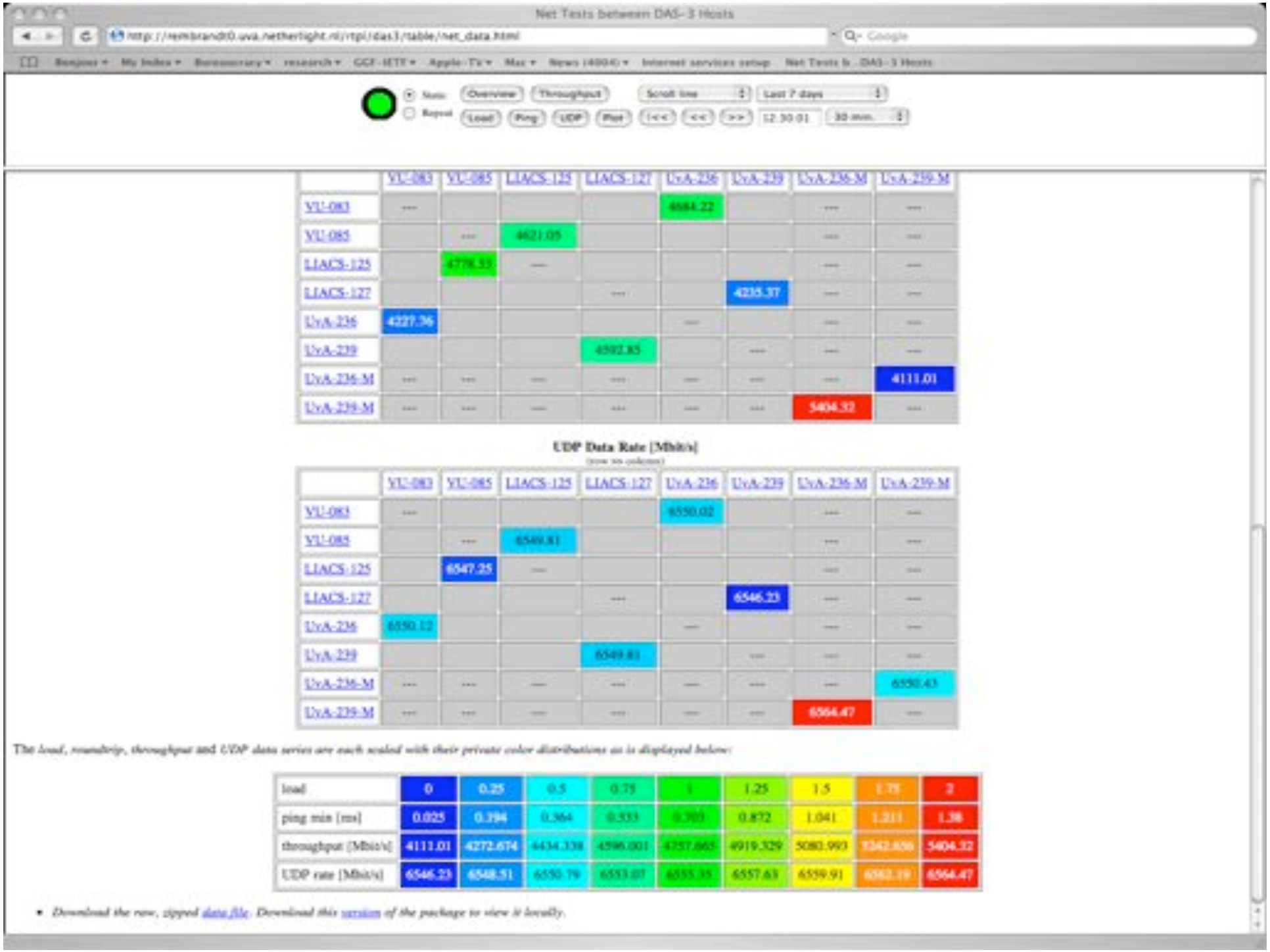
GRID Co-scheduling problem space

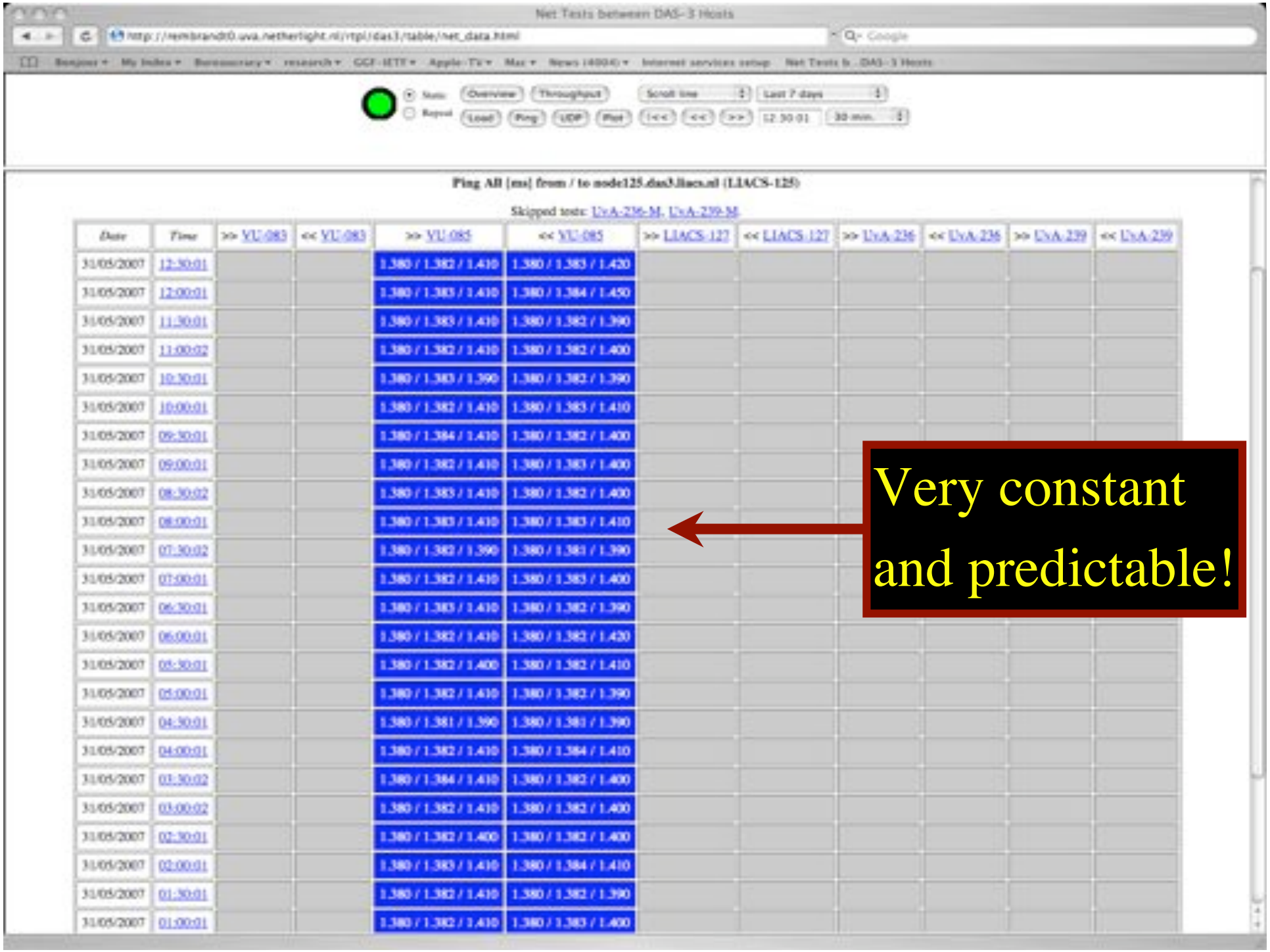
StarPlane



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.


Net Tests between DAS-3 Hosts
http://rembrandt0.uva.netherlight.nl/rtpl/das3/table/net_data.html
Overview
Throughput
Load
Ping
UDP
Plot
Overview Net Tests between DAS-3 Hosts
Authorise 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, lost.
Select UDP value: rate, lost.
DAS-3 Net Test Results
Date: 31/05/2007
Time: 12:30:01
Load
VU-083 VU-085 LIACS-125 LIACS-127 UvA-236 UvA-239 UvA-236-M UvA-239-M
0 0 0.087 0 0.013 0.01 0.017 0.15
Ping Min [ms]
VU-083 VU-085 LIACS-125 LIACS-127 UvA-236 UvA-239 UvA-236-M UvA-239-M
VU-083 --- 0.696
VU-085 --- 1.380
LIACS-125 1.380 ---
LIACS-127 --- 1.220
UvA-236 0.696 ---
UvA-239 1.220 ---
UvA-236-M --- --- --- --- --- --- --- 0.025
UvA-239-M --- --- --- --- --- --- 0.025 ---
Throughput [Mbit/s]
VU-083 VU-085 LIACS-125 LIACS-127 UvA-236 UvA-239 UvA-236-M UvA-239-M
VU-083 --- 4684.22
VU-085 --- 4821.05


Net Tests between DAS-3 Hosts
http://rembrandt0.uva.netherlight.nl/rtpl/das3/table/net_data.html
Google
Overview
Throughput
Scroll time
Last 7 days
Load
Ping
UDP
Plot
12:30:01
30 min.
VU-083
VU-085
LIACS-125
LIACS-127
UvA-236
UvA-239
UvA-236-M
UvA-239-M
4664.22
4621.05
4778.53
4235.37
4227.36
4592.85
4111.01
5404.32
UDP Data Rate [Mbit/s]
6550.02
6549.81
6547.25
6546.23
6550.12
6549.81
6550.43
6564.47
The load, roundtrip, throughput and UDP data series are each scaled with their private color distributions as is displayed below:
load
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
ping min [ms]
0.025
0.394
0.364
0.533
0.703
0.872
1.041
1.211
1.38
throughput [Mbit/s]
4111.01
4272.674
4434.338
4596.001
4757.665
4919.329
5080.993
5242.656
5404.32
UDP rate [Mbit/s]
6546.23
6548.51
6550.79
6553.07
6555.35
6557.63
6559.91
6562.19
6564.47
Download the raw, zipped data file. Download this version of the package to view it locally.



Date	Time	>> VU-083	<< VU-083	>> VU-085	<< VU-085	>> LIACS-127	<< LIACS-127	>> UvA-236	<< UvA-236	>> UvA-239	<< UvA-239
31/05/2007	12:30:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.420						
31/05/2007	12:00:01			1.380 / 1.383 / 1.410	1.380 / 1.384 / 1.450						
31/05/2007	11:30:01			1.380 / 1.383 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	11:00:02			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	10:30:01			1.380 / 1.383 / 1.390	1.380 / 1.382 / 1.390						
31/05/2007	10:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.410						
31/05/2007	09:30:01			1.380 / 1.384 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	09:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.400						
31/05/2007	08:30:02			1.380 / 1.383 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	08:00:01			1.380 / 1.383 / 1.410	1.380 / 1.383 / 1.410						
31/05/2007	07:30:02			1.380 / 1.382 / 1.390	1.380 / 1.383 / 1.390						
31/05/2007	07:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.400						
31/05/2007	06:30:01			1.380 / 1.383 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	06:00:01			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.420						
31/05/2007	05:30:01			1.380 / 1.382 / 1.400	1.380 / 1.382 / 1.410						
31/05/2007	05:00:01			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	04:30:01			1.380 / 1.381 / 1.390	1.380 / 1.383 / 1.390						
31/05/2007	04:00:01			1.380 / 1.382 / 1.410	1.380 / 1.384 / 1.410						
31/05/2007	03:30:02			1.380 / 1.384 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	03:00:02			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	02:30:01			1.380 / 1.382 / 1.400	1.380 / 1.382 / 1.400						
31/05/2007	02:00:01			1.380 / 1.383 / 1.410	1.380 / 1.384 / 1.410						
31/05/2007	01:30:01			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	01:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.400						

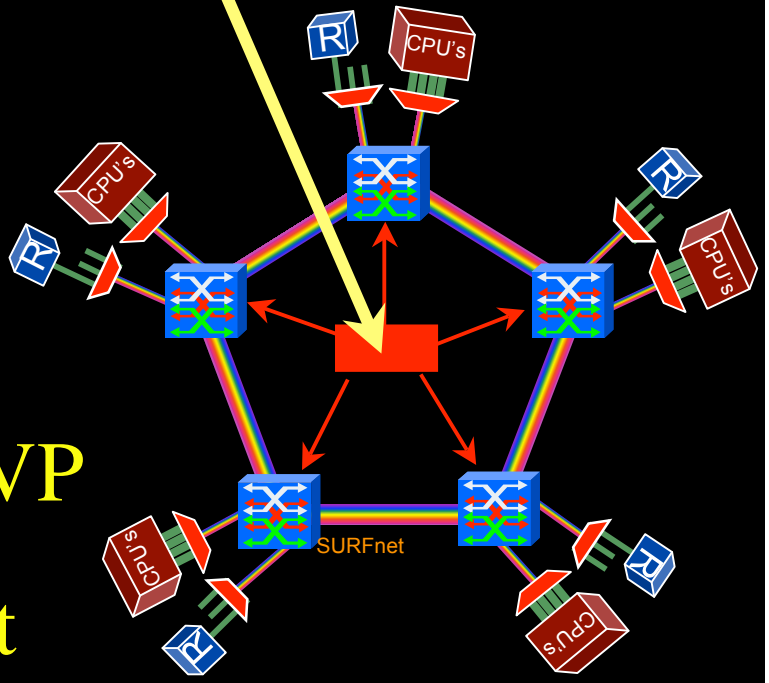
Control Plane

StarPlane

Application Scheduler

WS

DRAC

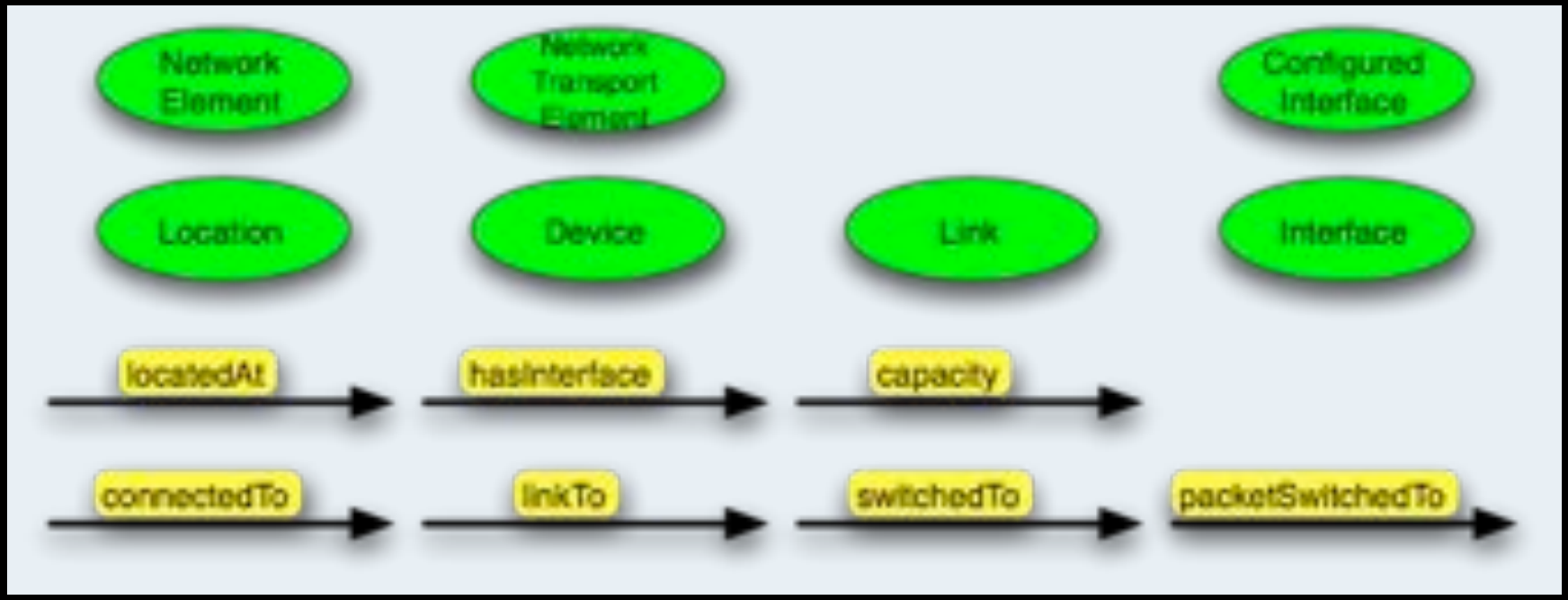


- currently implementing WS
- study interaction with TBN/RSVP
- also photonic switching at client

StarPlane and NDL

While on topologies. SNE group is working on NDL - Network Description Language.

NDL is an RDF data model, based on idea of Semantic Web, for network topology descriptions.

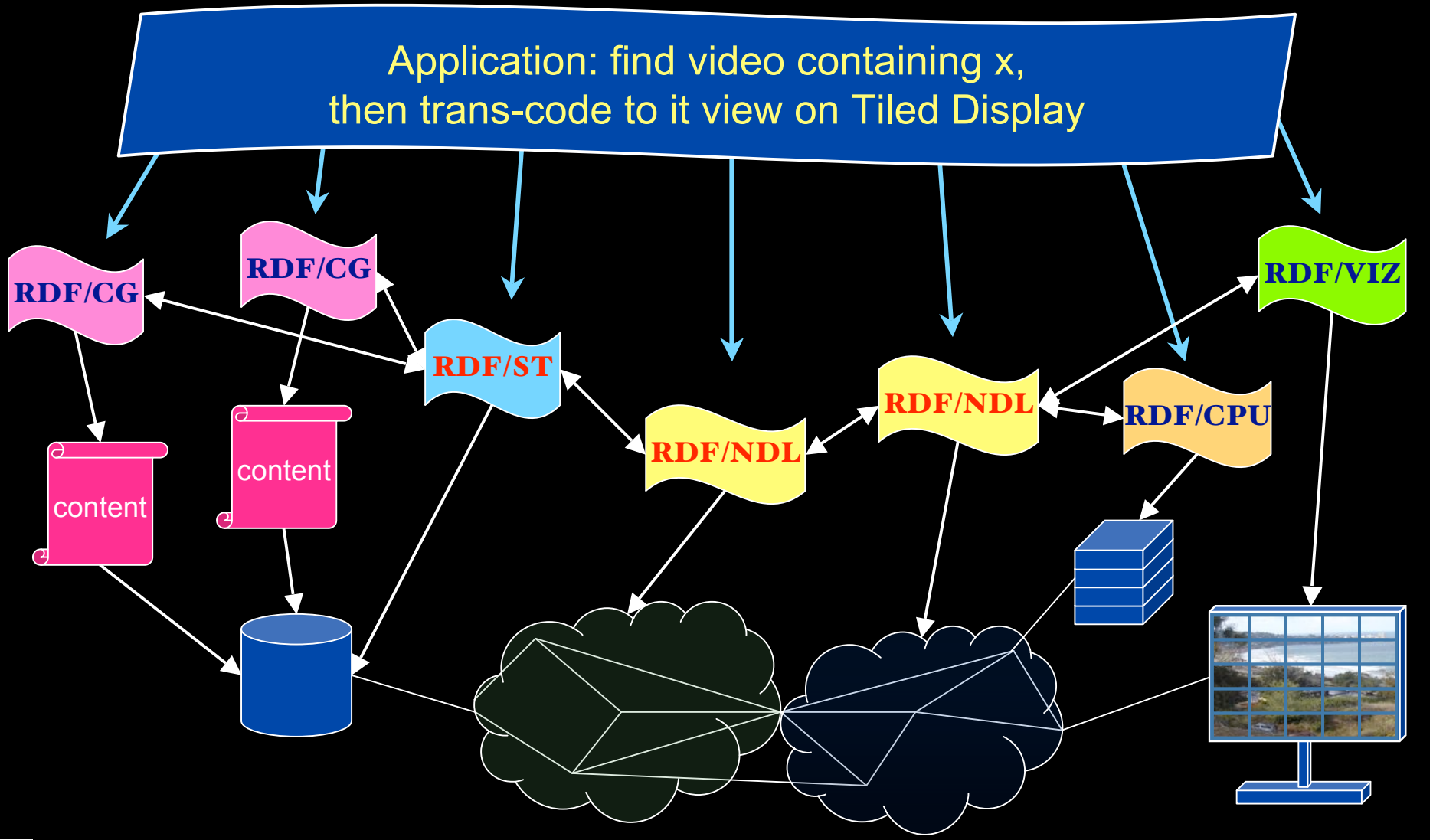


In StarPlane we are researching use of NDL for topology exchange and topology requests from clients.

ref: Talk from Paola Grosso on NDL/RDF at TNC2007

RDF describing Infrastructure

StarPlane

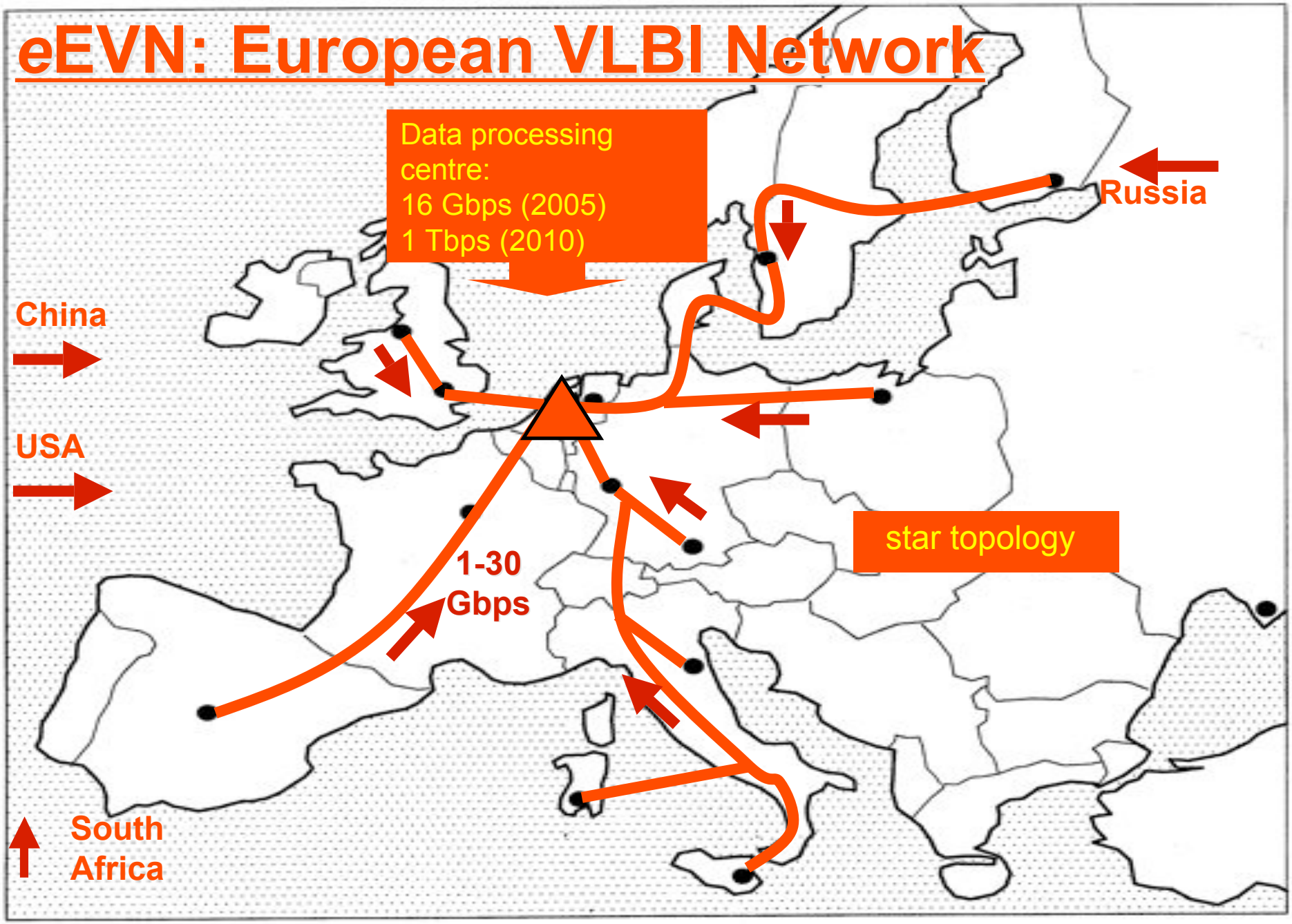


PG&CdL

CineGrid@SARA

StarPlane

eEVN: European VLBI Network

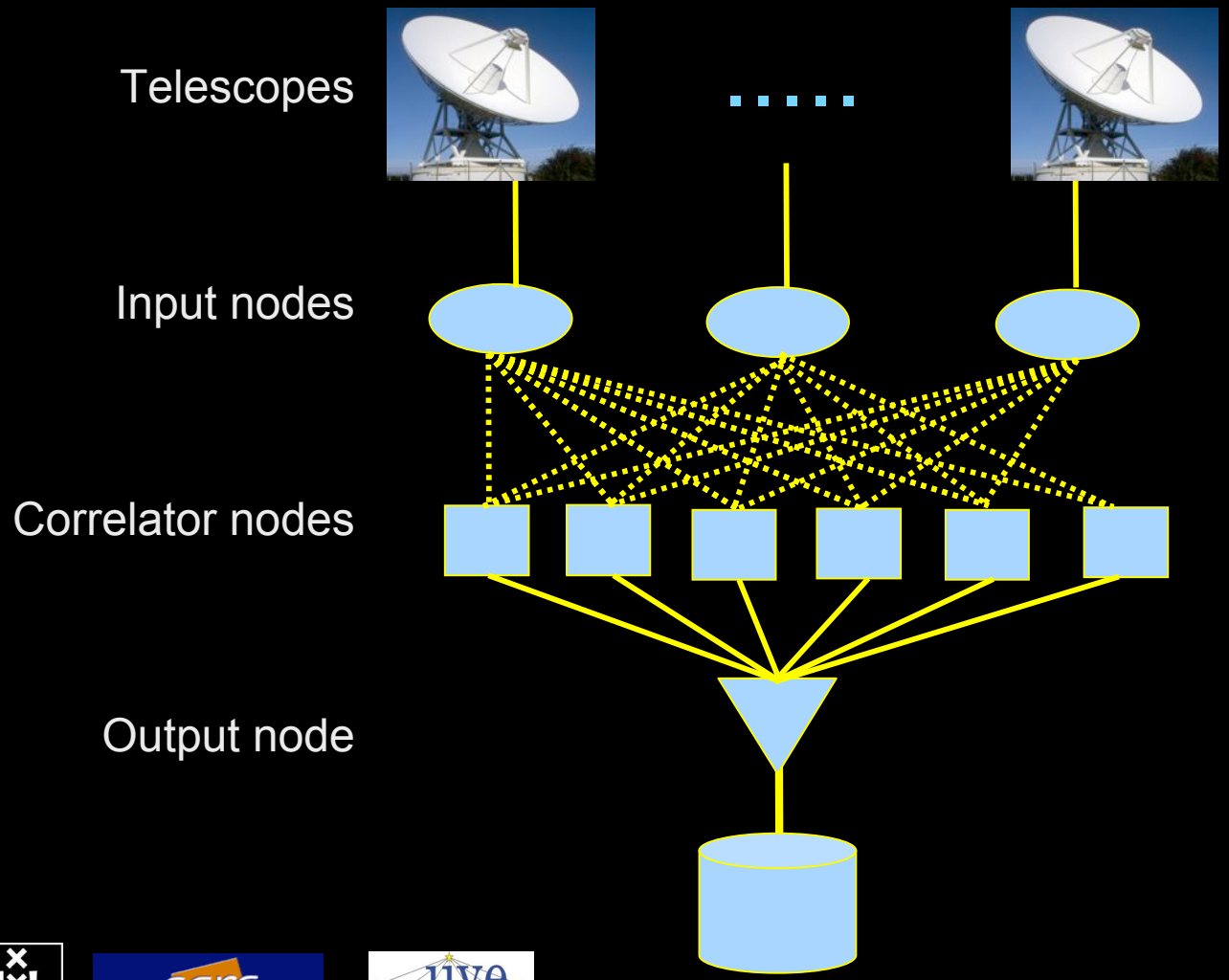


This slide courtesy of Richard Schilizzi <schilizzi@jive.nl>

The SCARIe project

StarPlane

SCARIe: 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

2 TFlop / 16 Gbps =

1000 flops/byte

THIS IS A DATA FLOW PROBLEM !!!

sara Computing & Networking Services

jive JOINT INSTITUTE FOR VLBI IN EUROPE

Tera-Thinking

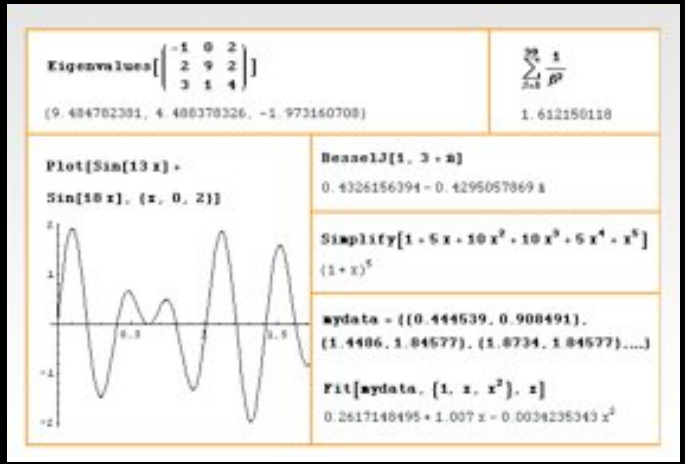
StarPlane

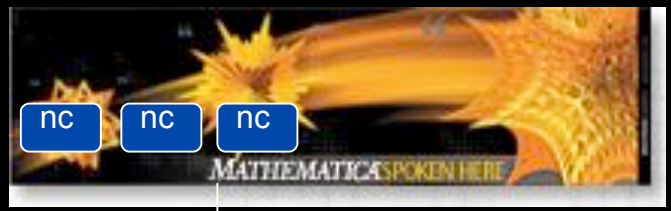
- What constitutes a Tb/s network?
- CALIT2 has 8000 Gigabit drops ?->? Terabit Lan?
- look at 80 core Intel processor
  - cut it in two, left and right communicate 8 TB/s
- massive parallel channels in hosts, NIC's
- think back to teraflop computing!
  - MPI makes it a teraflop machine
- TeraApps programming model supported by
  - TFlops -> MPI / Globus
  - TBytes -> OGSA/DAIS
  - TPixels -> SAGE
  - TSensors -> LOFAR, LHC, LOOKING, CineGrid, ...
  - Tbit/s -> ?

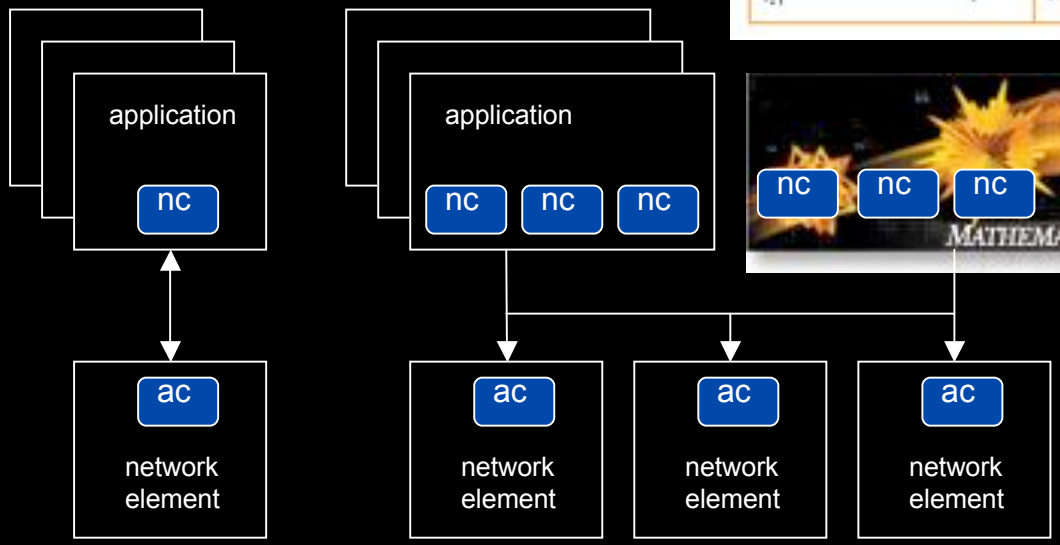
ref Larry Smarr & CdL

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
- UPVNs enable network resources to be programmed as part of the application
- Mathematica, a powerful mathematical software system, can interact with real networks using UPVNs







StarPlane

Mathematica enables advanced graph queries, visualizations and real-time network manipulations on UPVNs

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

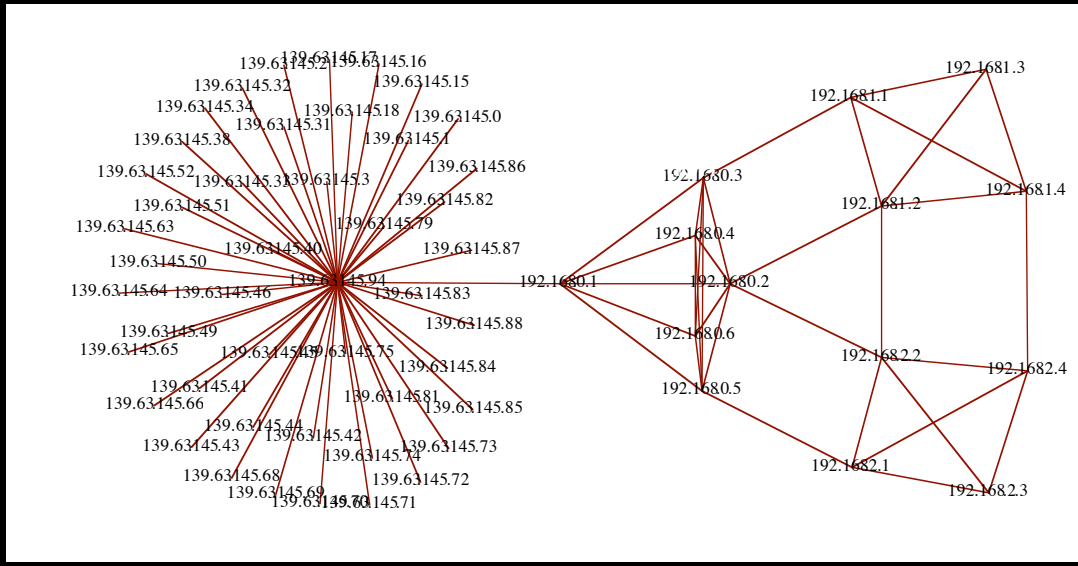
Initialization and BFS discovery of NEs

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

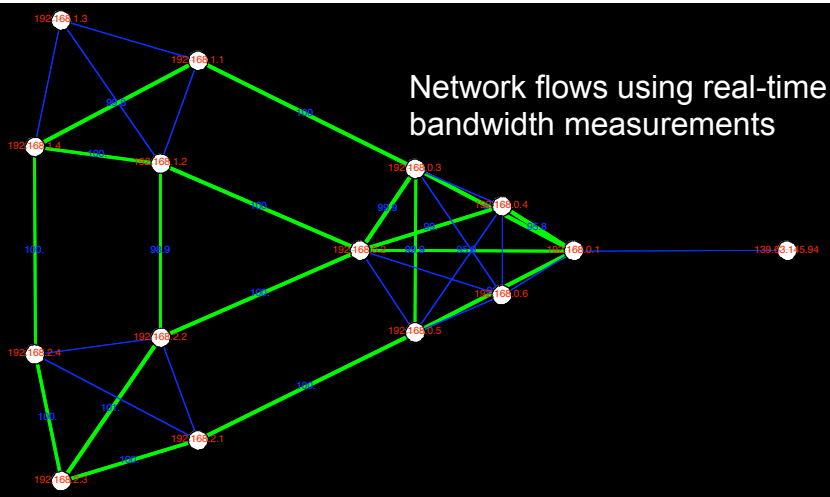


Transaction on shortest path with tokens

```
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
```



Network flows using real-time bandwidth measurements

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

StarPlane

StarPlane

Questions ?

Thanks to:

SURFnet, BSIK (GigaPort grant), NWO (grant 643.000.504), NORTEL

StarPlane team: Li Xu, Jason Maasen, JP Velders, Paola Grosso, Herbert Bos, Henri Bal

DAS-3 admins

Special thanks to Kees Neggers and his team.

Advanced School for Computing and Imaging

DAS-3

SURF NET