Walking the Line

Cees de Laat

SURFnet

BSIK

EU

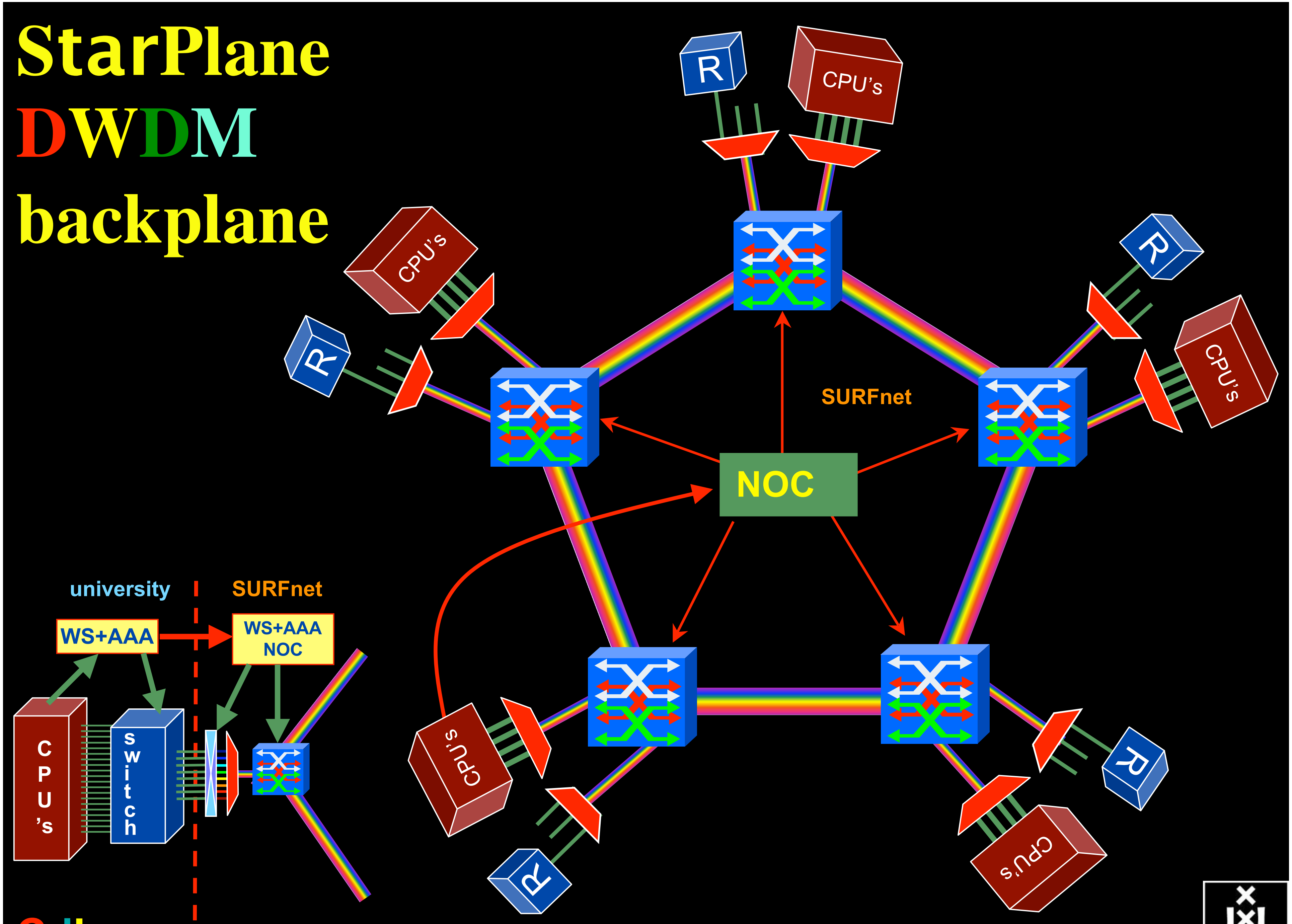
NWO

University of Amsterdam

TNO

NCF

StarPlane DWDM backplane



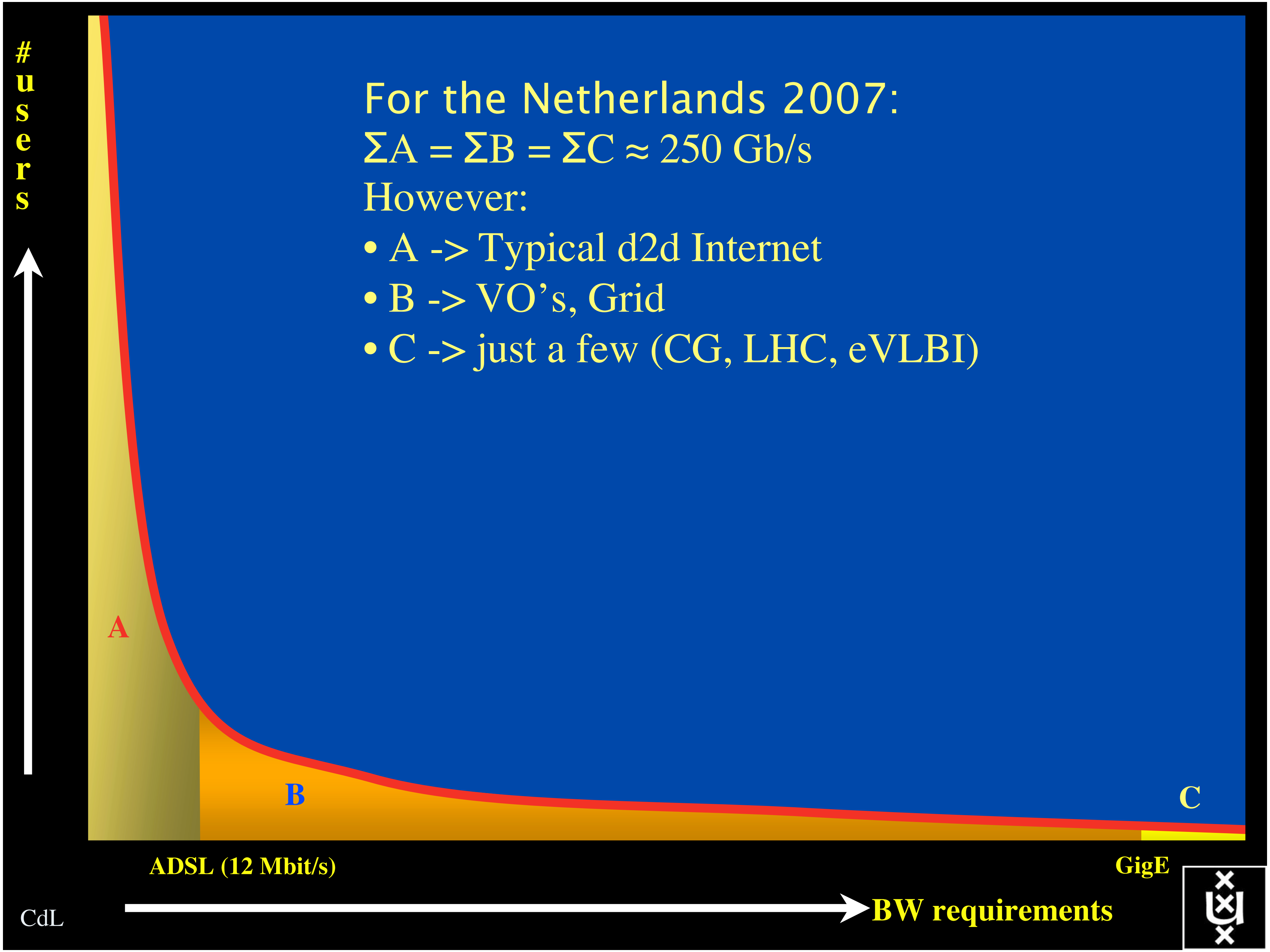
CdL

For the Netherlands 2007:

$\Sigma A = \Sigma B = \Sigma C \approx 250$ Gb/s

However:

- A -> Typical d2d Internet
- B -> VO’s, Grid
- C -> just a few (CG, LHC, eVLBI)



CdL

Towards Hybrid Networking!

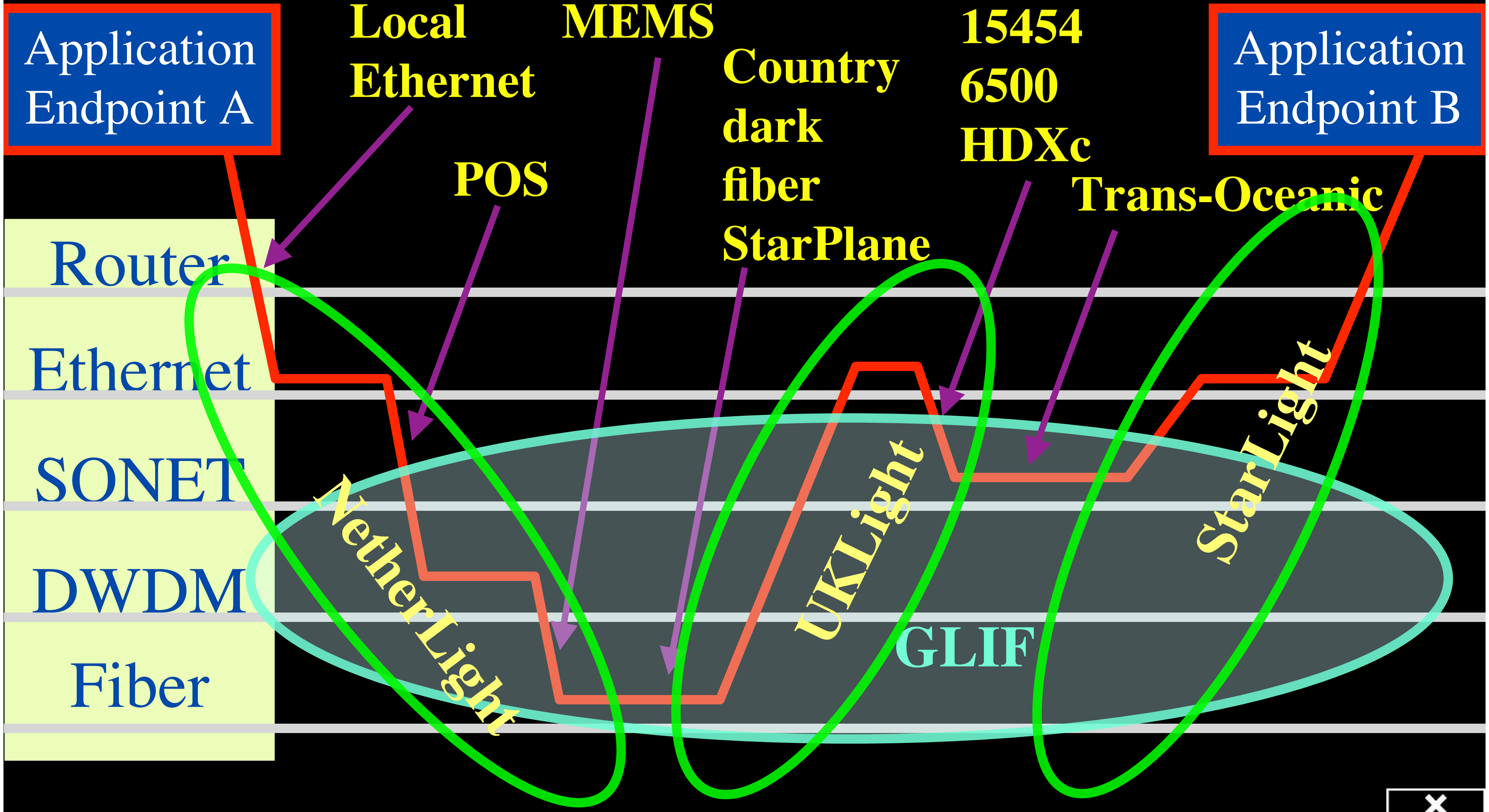
- Costs of photonic equipment 10% of switching 10 % of full routing
  - for same throughput!
  - Photonic vs Optical (optical used for SONET, etc, 10-50 k$/port)
  - DWDM lasers for long reach expensive, 10-50 k$
- Bottom line: look for a hybrid architecture which serves all classes in a cost effective 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

Application Endpoint B

Local Ethernet

MEMS

POS

Country dark fiber StarPlane

15454 6500 HDXc

Trans-Oceanic

Router

Ethernet

SONET

DWDM

Fiber

NetherLight

UKLight

StarLight

GLIF

Infrastructure Flexibility & Functionality

SCALE CLASS	Metro Country 2 ms RTT	Regional Continental 20 ms RTT	World Trans Ocean 200 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

Infrastructure Flexibility & Functionality

SCALE CLASS	Metro Country 2 ms RTT	Regional Continental 20 ms RTT	World Trans Ocean 200 ms RTT
A	Switching/ Routing	Routers	ROUTER$
B	Switches VPN's E-WANPHY	Routing Switches G-MPLS E-WANPHY	ROUTER$
C	dark fiber DWDM MEMS Photonic switch	DWDM TDM / SONET Lambda switching	VLAN's TDM SONET Ethernet

KUVN

PBT/PLSB

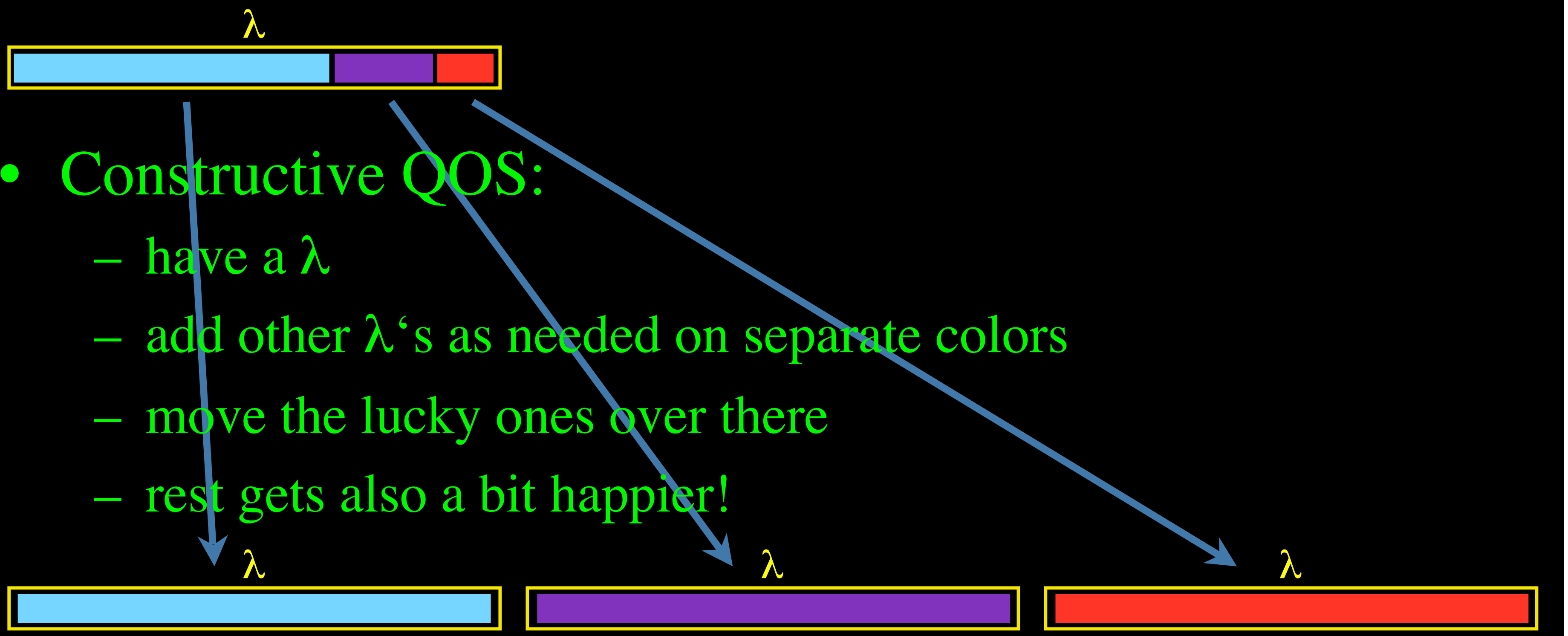
StarPlane

Phosphorus

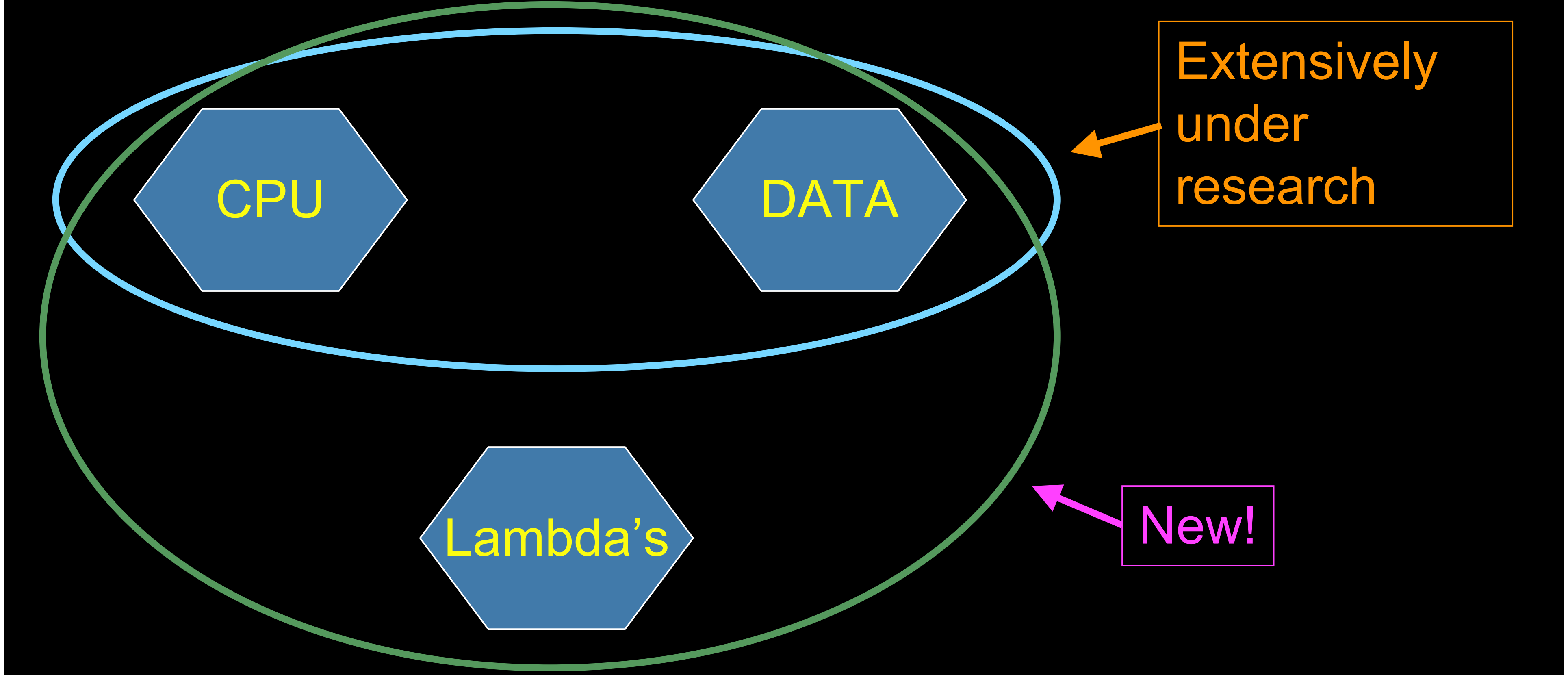
QOS in a non destructive way!

- 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!



GRID Co-scheduling problem space



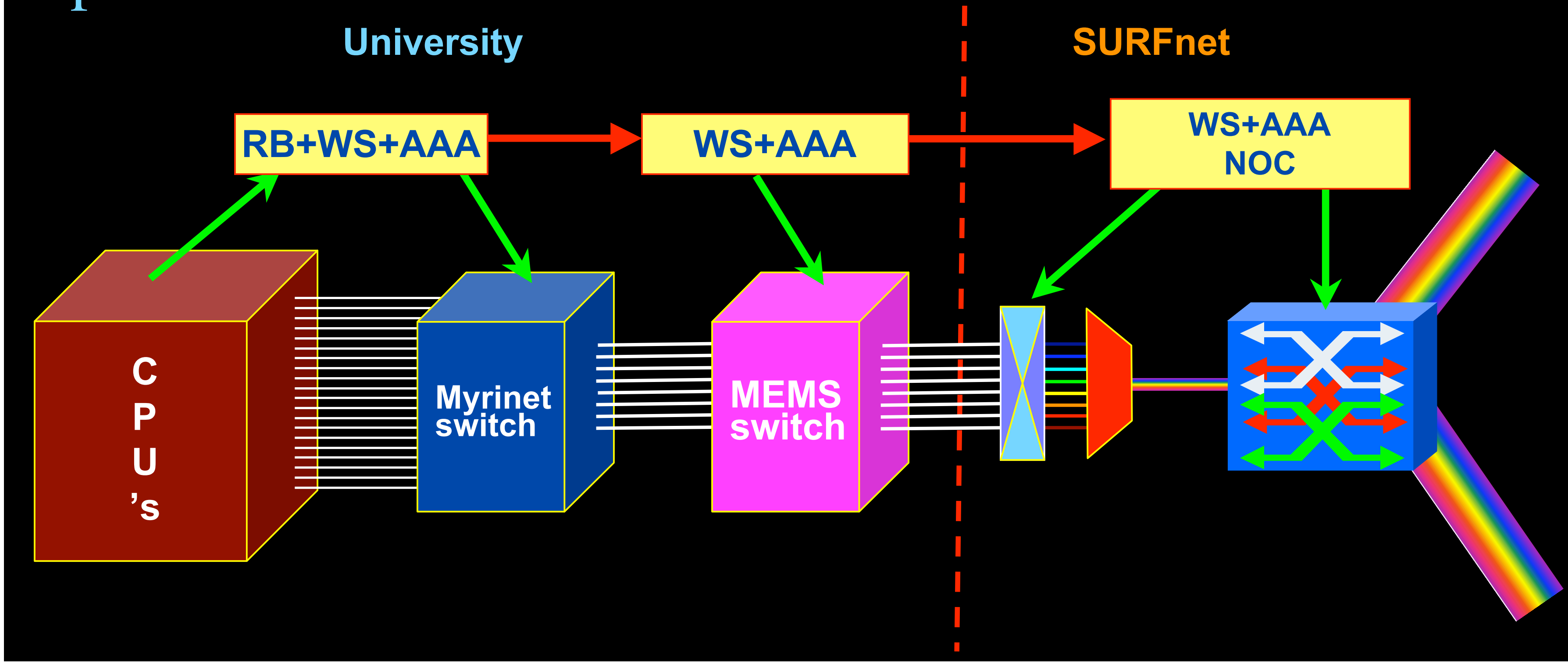
Extensively under research

New!

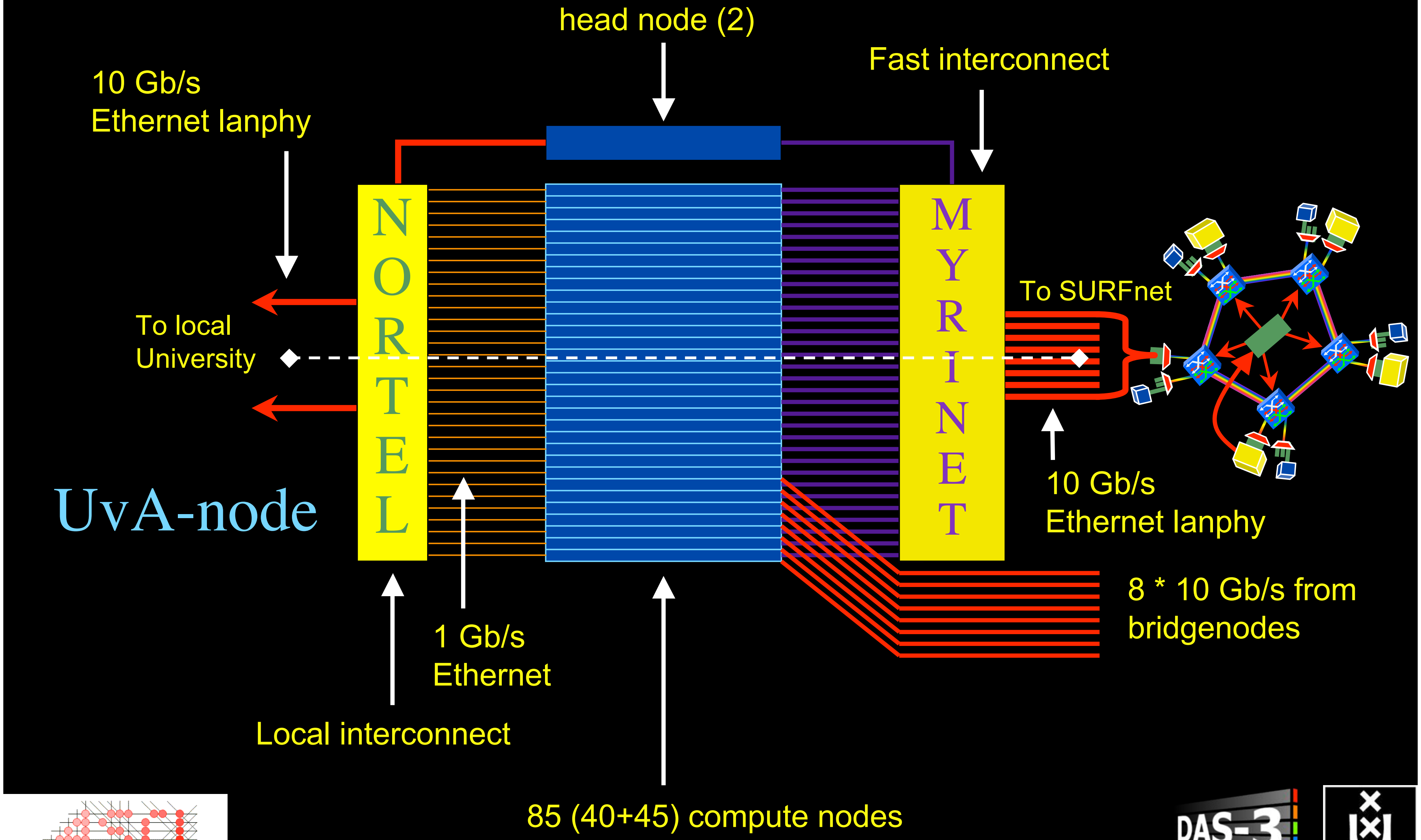
The StarPlane vision is to give flexibility directly to the applications by allowing them to choose the logical topology in real time, ultimately with sub-second lambda switching times on part of the SURFnet6 infrastructure.

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:



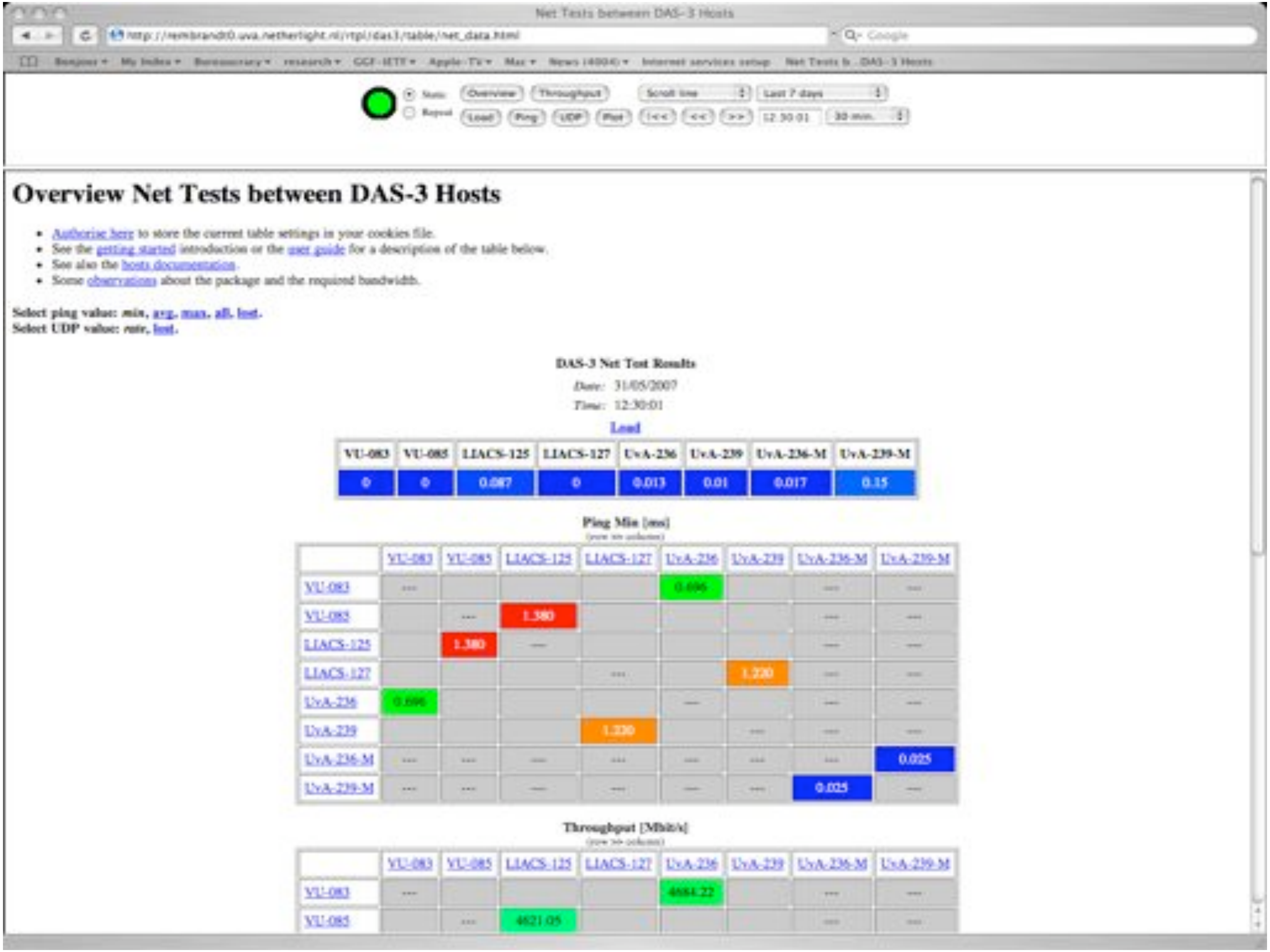
DAS-3 Cluster Architecture



Power is a big issue

- 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!



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

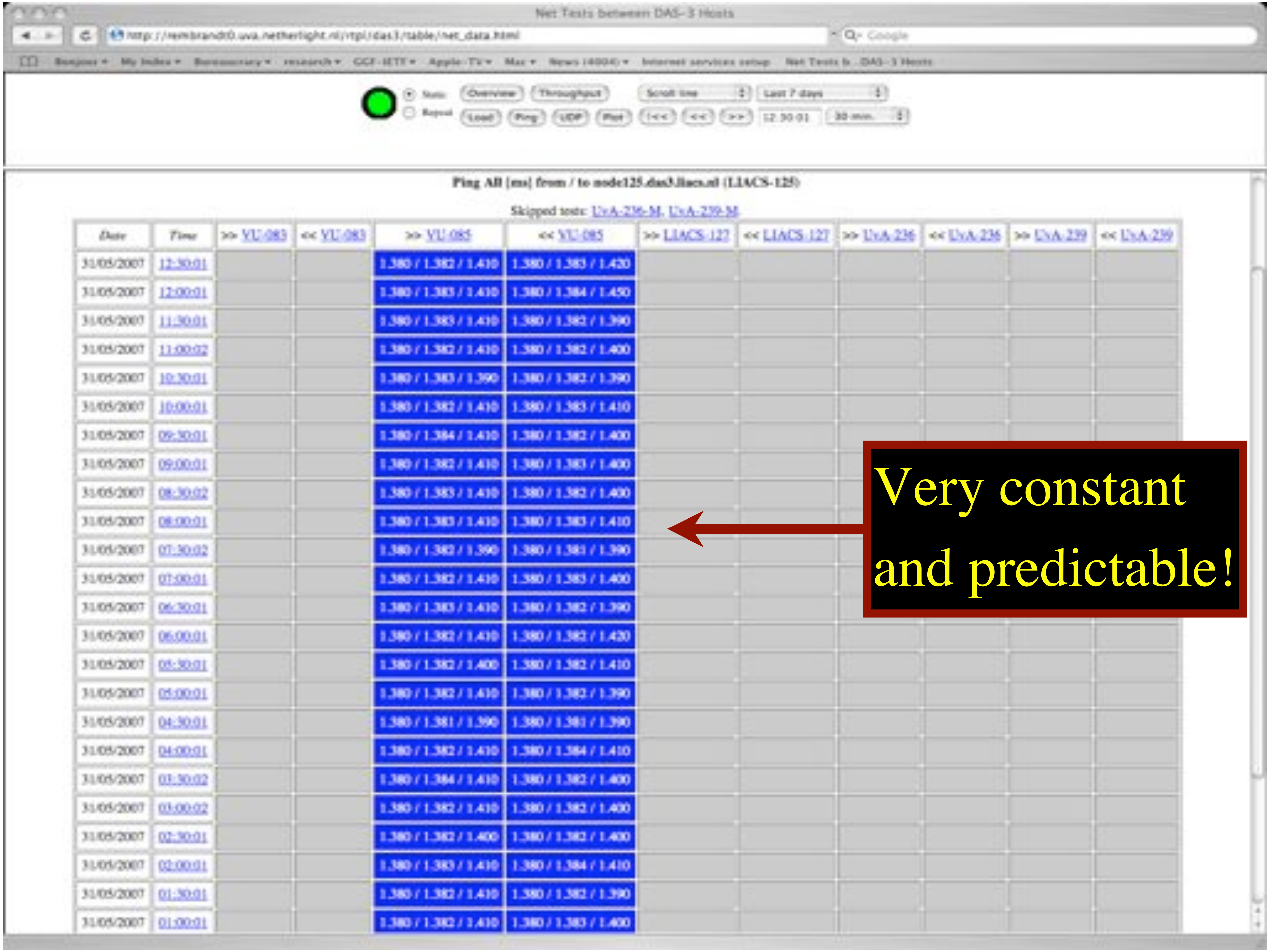
VU-083	VU-085	LIACS-125	LIACS-127	UvA-236	UvA-239	UvA-236-M	UvA-239-M
0	0	0.087	0	0.013	0.01	0.017	0.15

Ping Min [ms]
(row => column)

	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]
(row => column)

	VU-083	VU-085	LIACS-125	LIACS-127	UvA-236	UvA-239	UvA-236-M	UvA-239-M
VU-083	---				4684.22		---	---
VU-085		---	4621.05				---	---



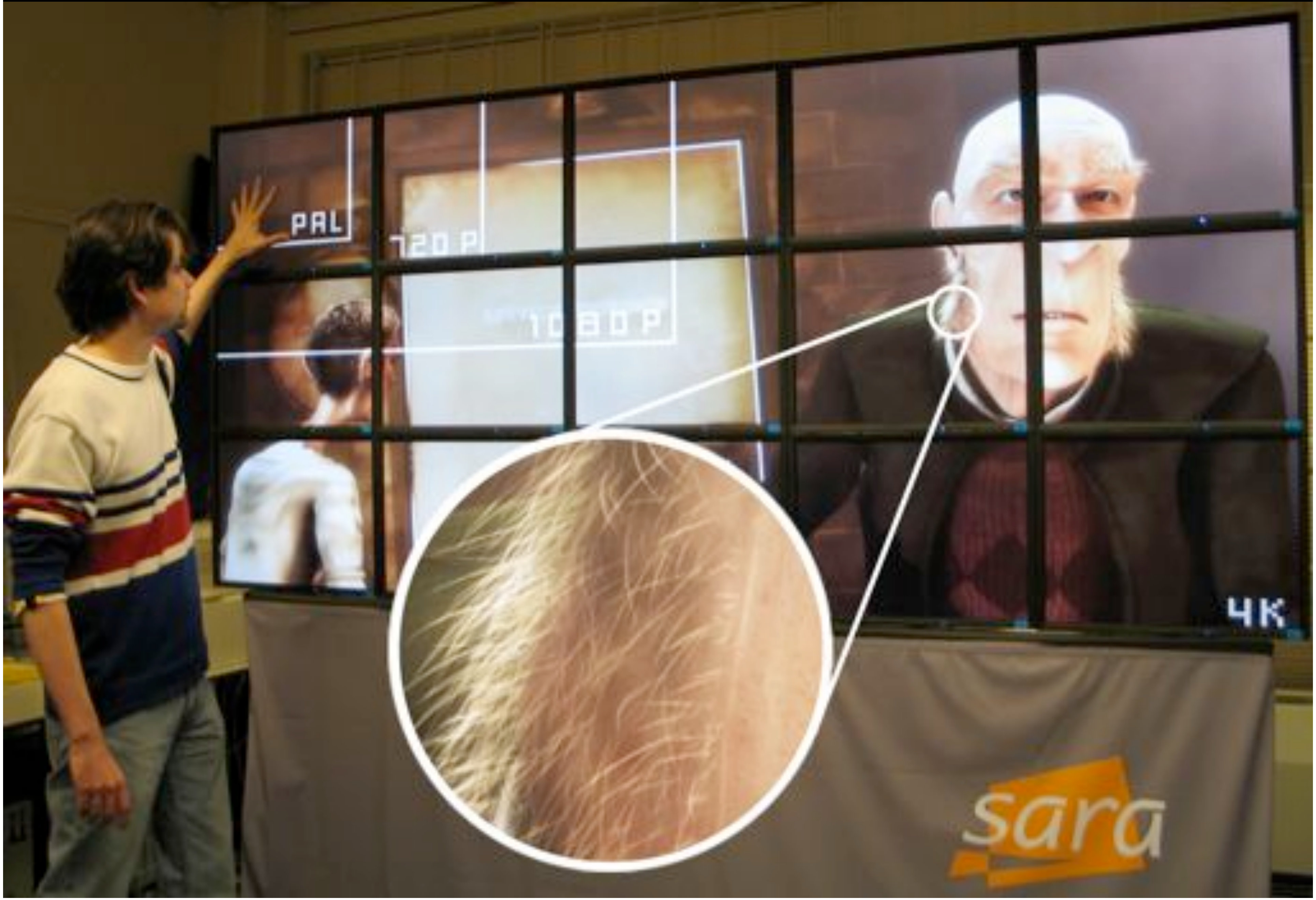
Ping All [ms] from / to node125.das3.liacs.nl (LIACS-125)

Skipped tests: UvA-236-M, UvA-239-M

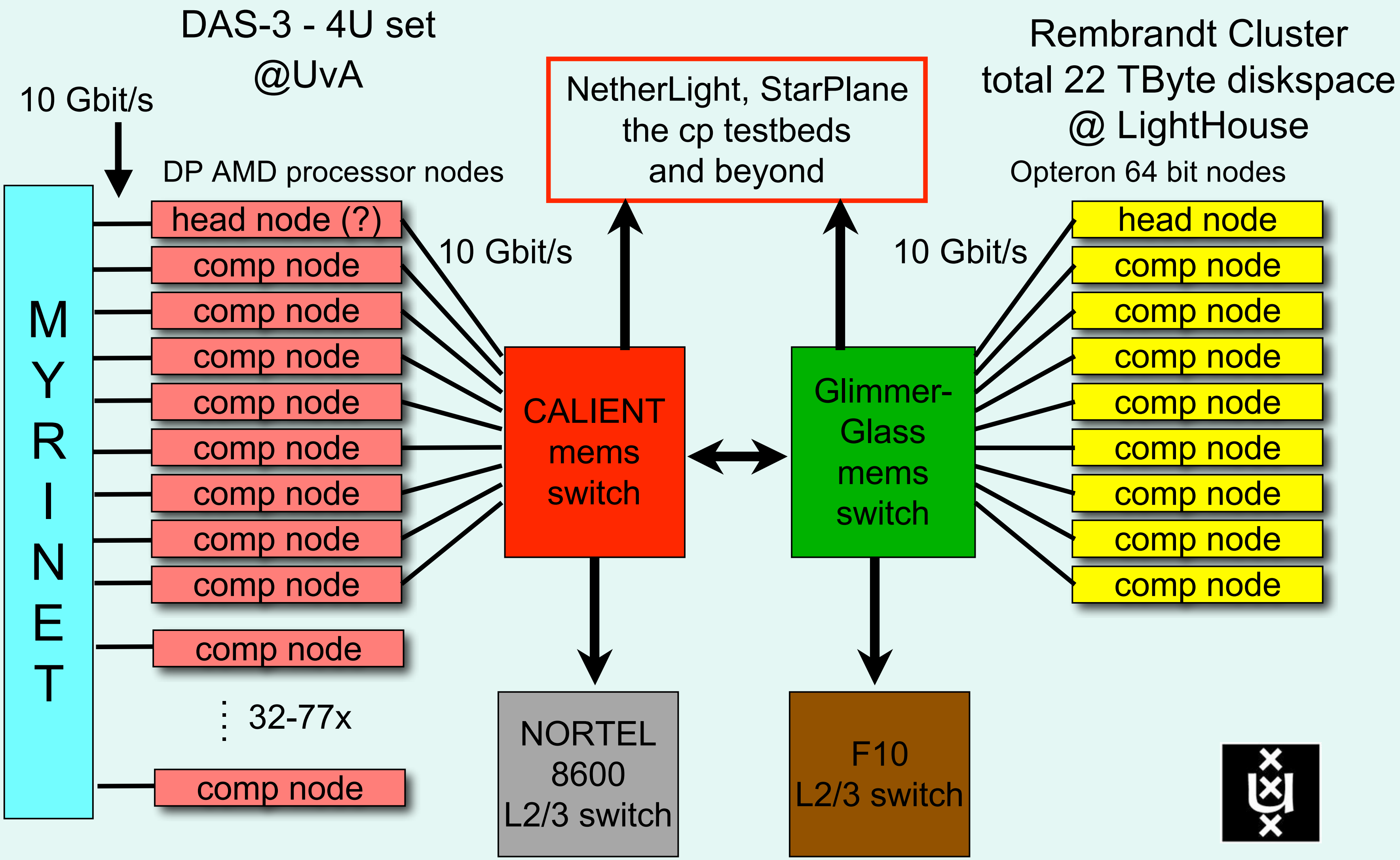
Date	*Time*	>> VU-083	<< VU-083	>> VU-085	<< VU-085	>> LIACS-127	<< LIACS-127	>> UvA-236	<< UvA-236	>> UvA-239	<< UvA-239
31/05/2007	12:30:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.420						
31/05/2007	12:00:01			1.380 / 1.383 / 1.410	1.380 / 1.384 / 1.450						
31/05/2007	11:30:01			1.380 / 1.383 / 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 / 1.381 / 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 / 1.381 / 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						

Very constant and predictable!

CineGrid@SARA



CineGrid



DAS-3 - 4U set
@UvA

10 Gbit/s

DP AMD processor nodes

MYRINET

head node (?)

comp node

comp 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

F10
L2/3 switch

Rembrandt Cluster
total 22 TByte diskspace
@ LightHouse

Opteron 64 bit nodes

10 Gbit/s

head node

comp node

comp node

comp node

comp node

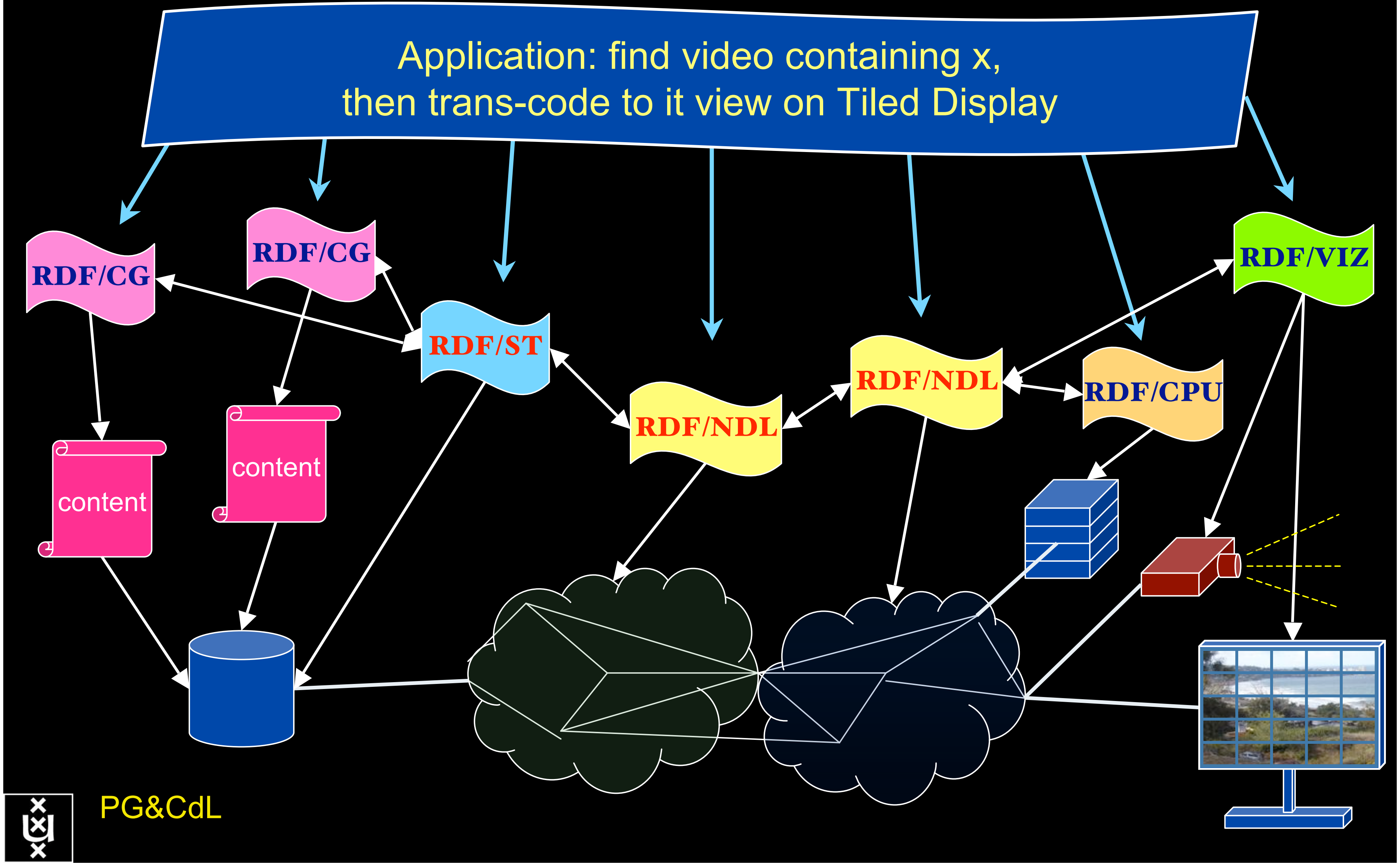
comp node

comp node

comp node

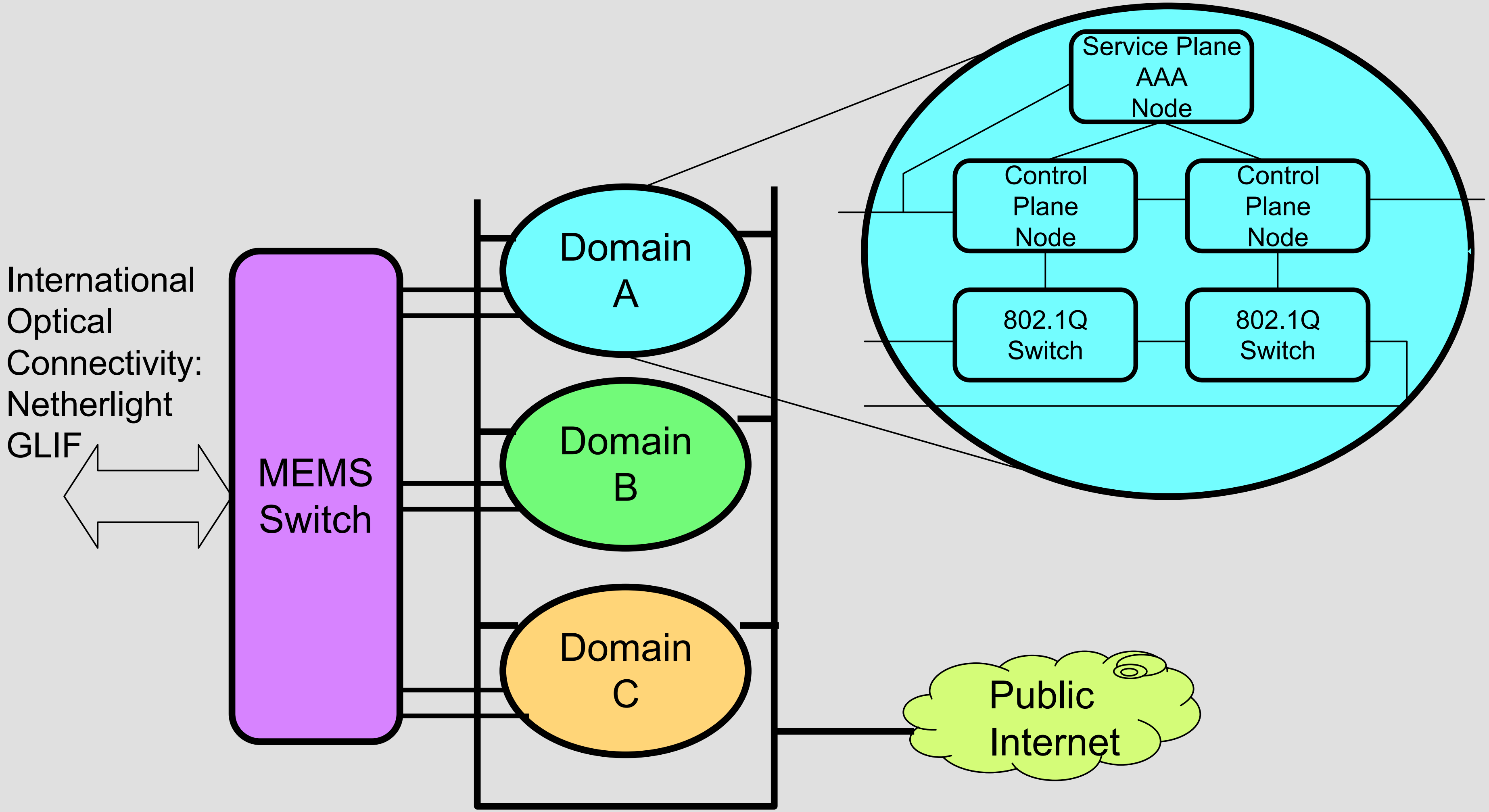
comp node

RDF describing Infrastructure


Application: find video containing x,
then trans-code to it view on Tiled Display
RDF/CG
RDF/CG
RDF/ST
RDF/NDL
RDF/NDL
RDF/CPU
RDF/VIZ
content
content

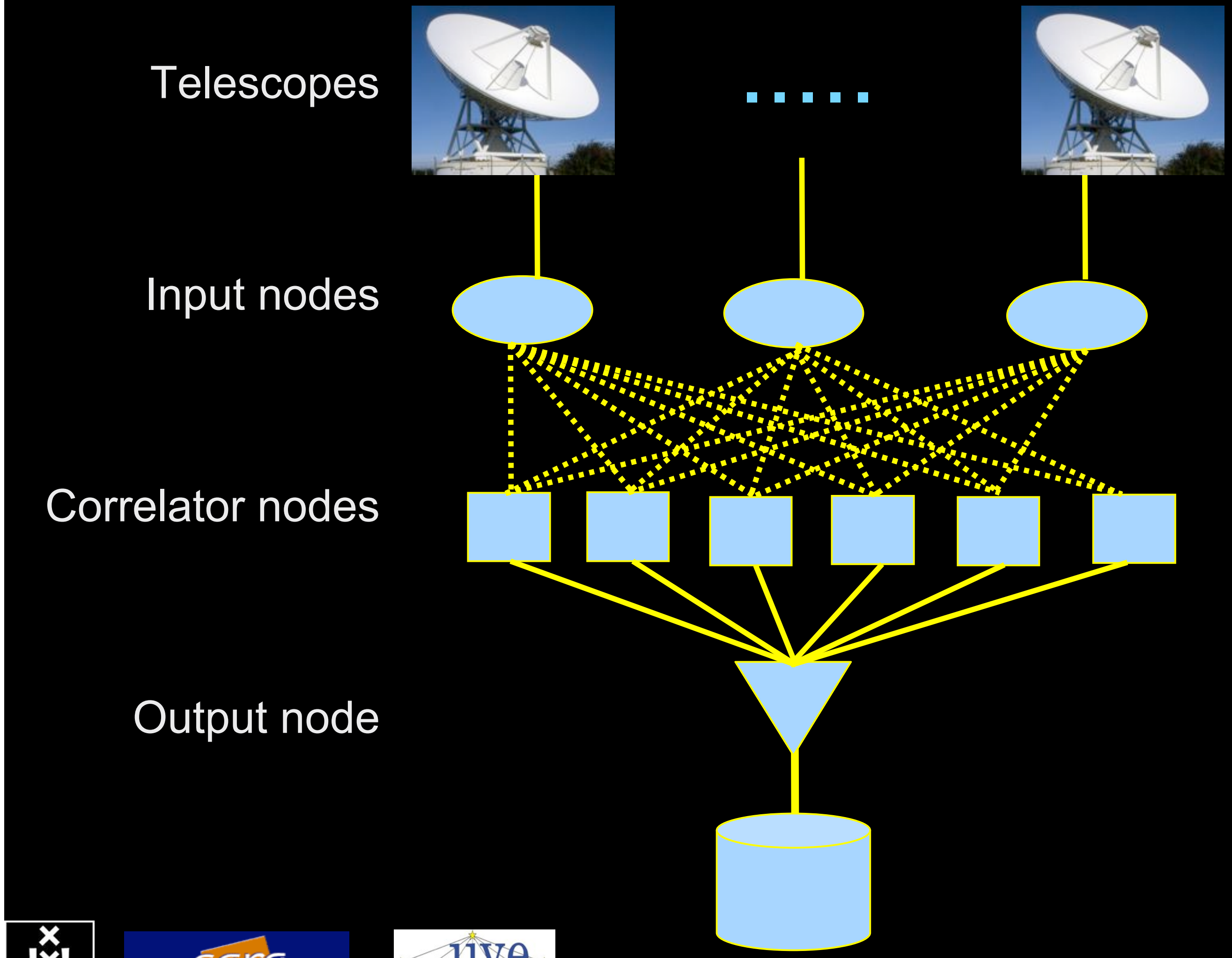
PG&CdL

Phosphorus AAA testbed



The SCARIe project

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.



Telescopes

Input nodes

Correlator nodes

Output node

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 !!!

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	->	?

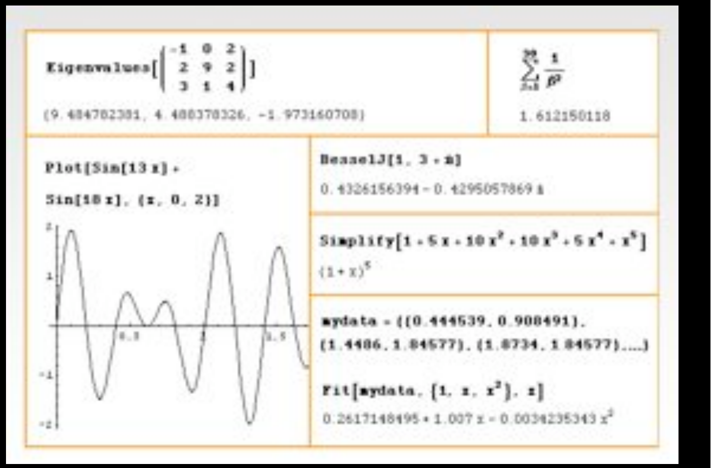
ref Larry Smarr & CdL

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
- 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



$$\text{Eigenvalues}\left[\begin{pmatrix} -1 & 0 & 2 \\ 2 & 9 & 2 \\ 3 & 1 & 4 \end{pmatrix}\right]$$

{9.484782381, 4.488378326, -1.973160708}

$$\sum_{\beta=1}^{30} \frac{1}{\beta^2}$$

1.612150118

Plot[Sin[13 x] + Sin[18 x], {x, 0, 2}]

BesselJ[1, 3 + ⅈ]

0.4326156394 - 0.4295057869 ⅈ

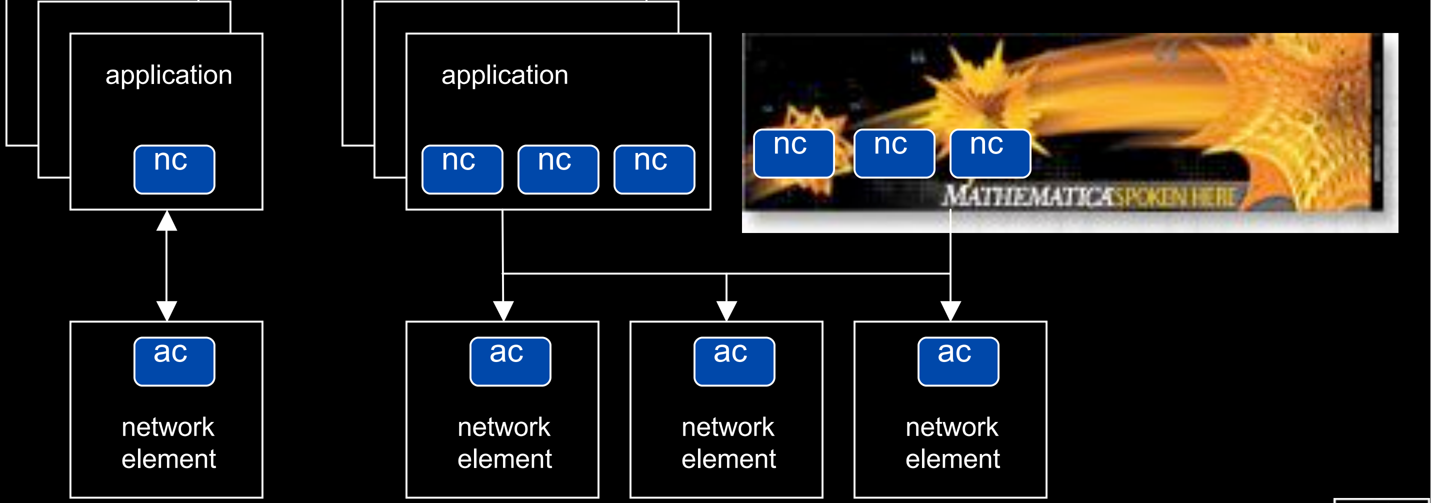
Simplify$[1 + 5 x + 10 x^2 + 10 x^3 + 5 x^4 + x^5]$

$(1 + x)^5$

mydata = {{0.444539, 0.908491}, {1.4486, 1.84577}, {1.8734, 1.84577},....}

Fit[mydata, $\{1, x, x^2\}$, x]

$0.2617148495 + 1.007 x - 0.0034235343 x^2$

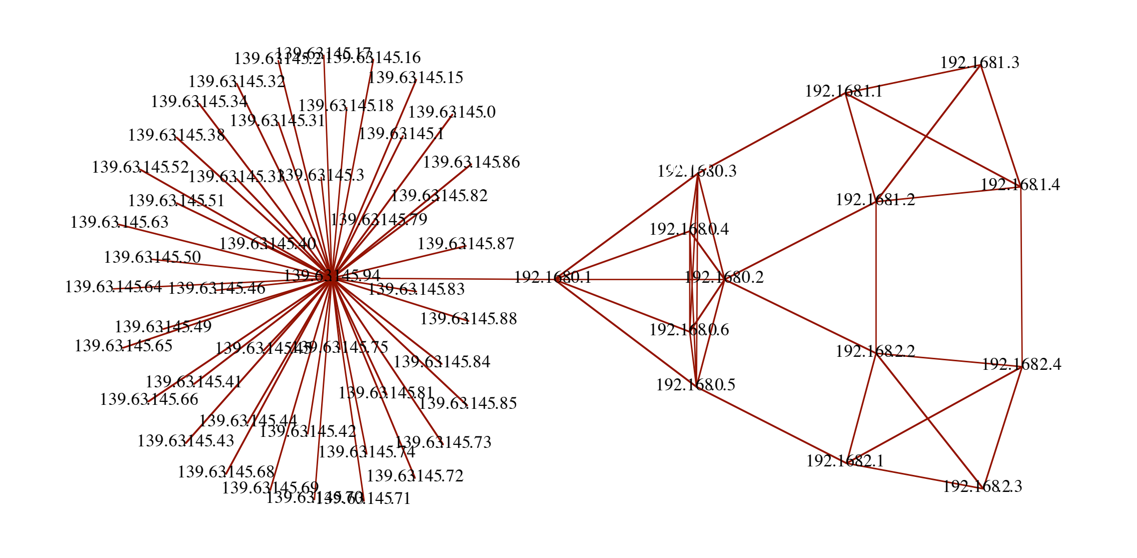


Mathematica enables advanced graph queries, visualizations and real-time network manipulations on UPVNs

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

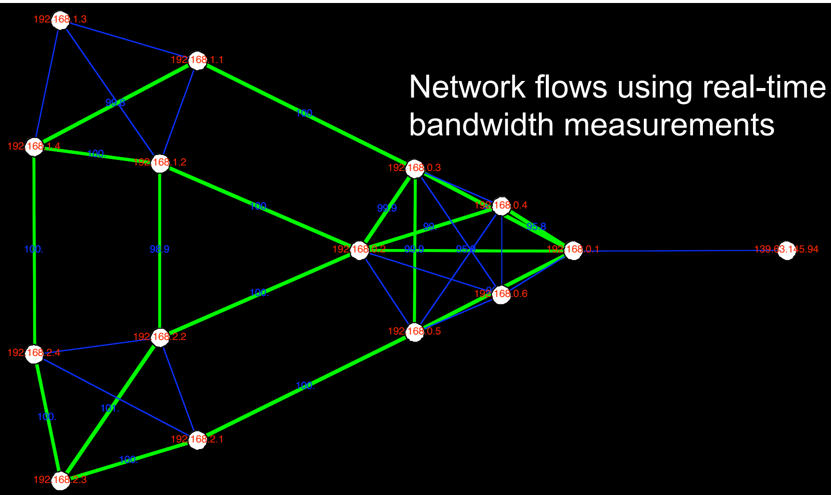
Initialization and BFS discovery of NEs

```
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

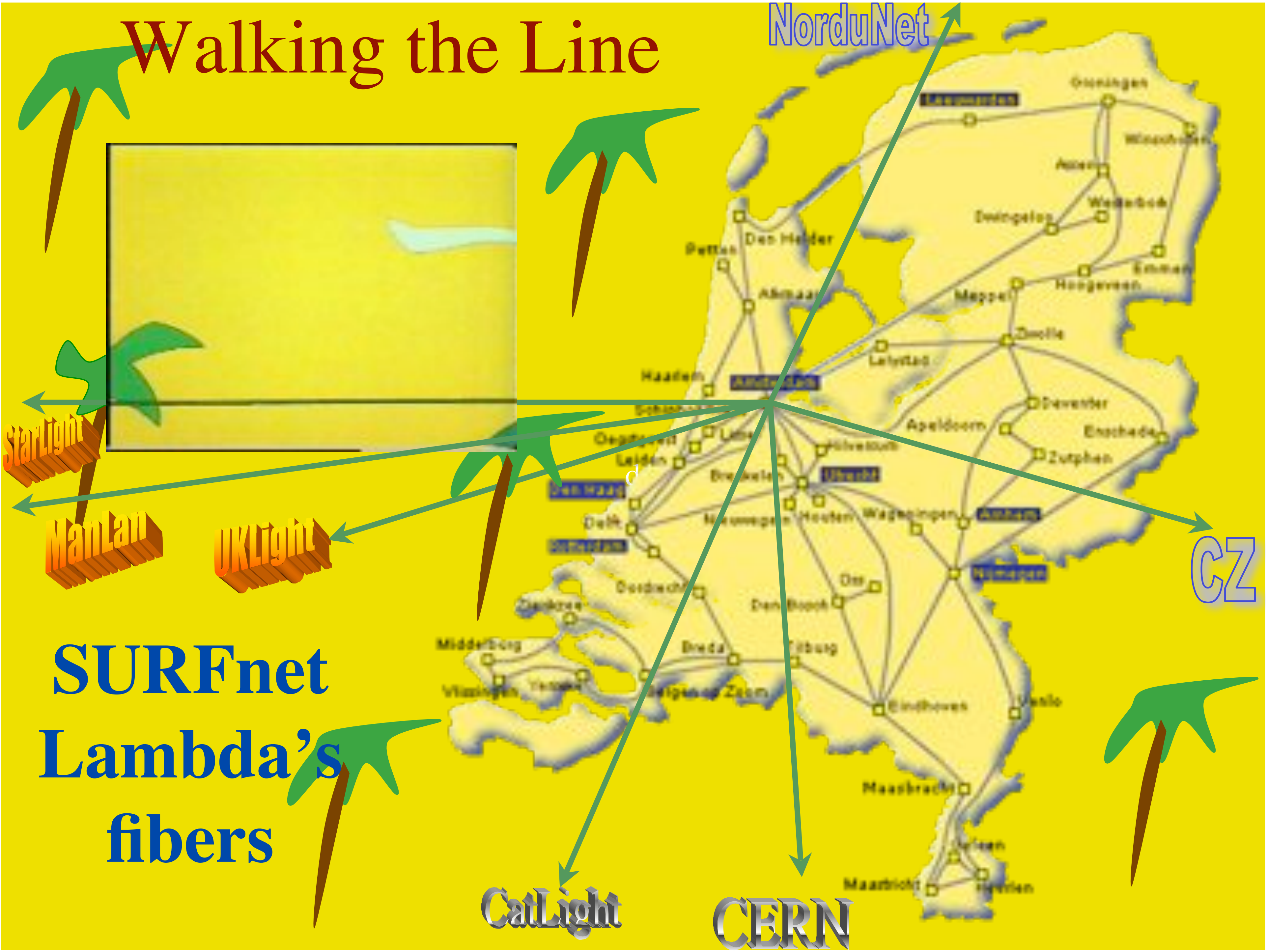
```
nodePath = ConvertIndicesToNodes[
                    ShortestPath[  g,
                                        Node2Index[nids,"192.168.3.4"],
                                        Node2Index[nids,"139.63.77.49"]],
                                        nids];
Print["Path: ", nodePath];
If[NetworkTokenTransaction[nodePath, "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.

Walking the Line



SURFnet Lambda's fibers

- *I did not talk about:*

AAA & TBN

Security

Grid, workflow

etc.etc.

Questions ?