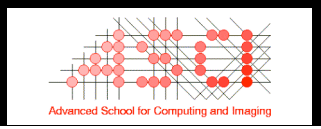
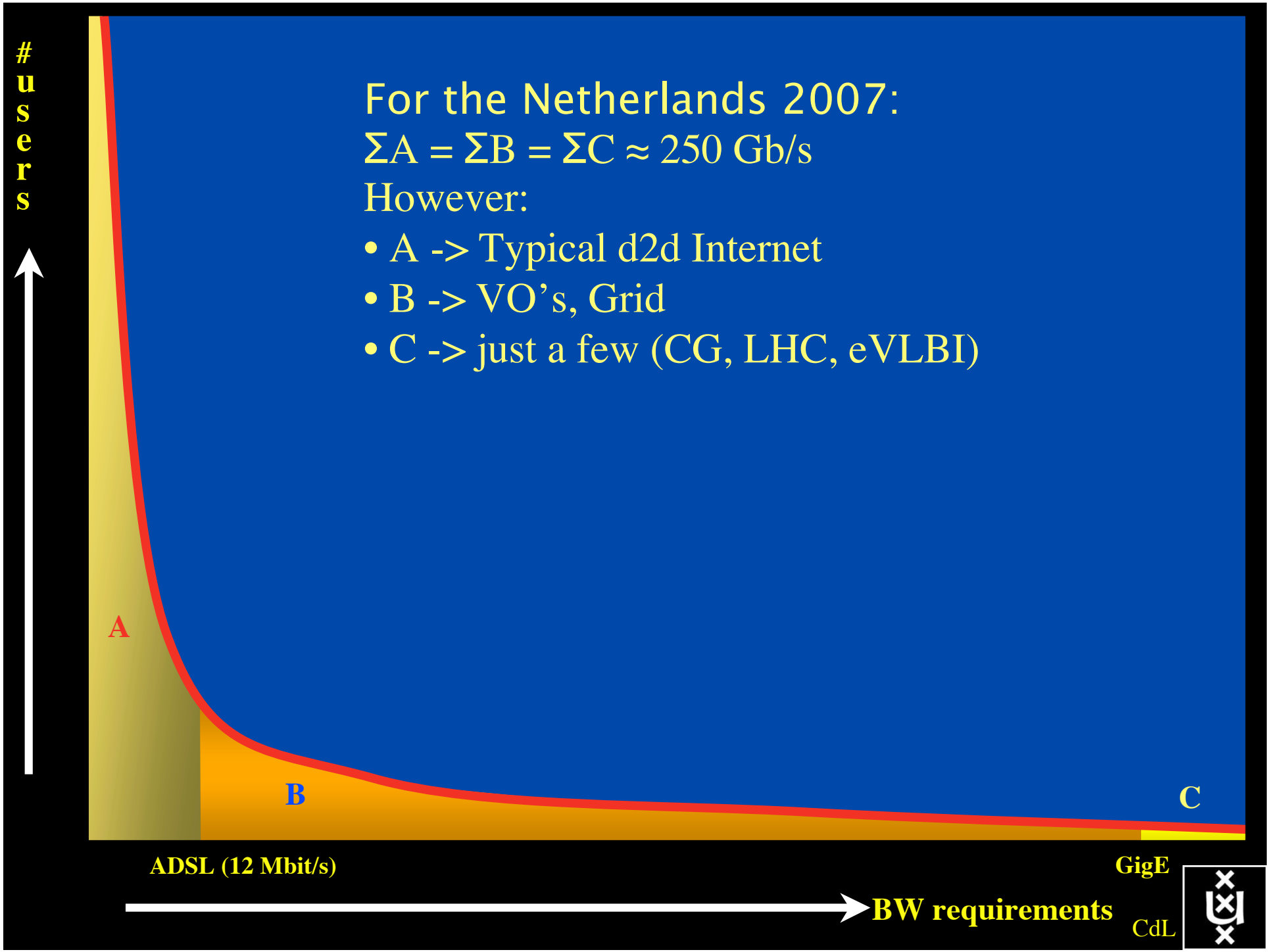
Lambda-Grid developments

StarPlane

Cees de Laat

University of Amsterdam




#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

Infrastructure

SCALE / CLASS	2 Metro	20 Regional	200 World
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

In The Netherlands SURFnet connects between 180:

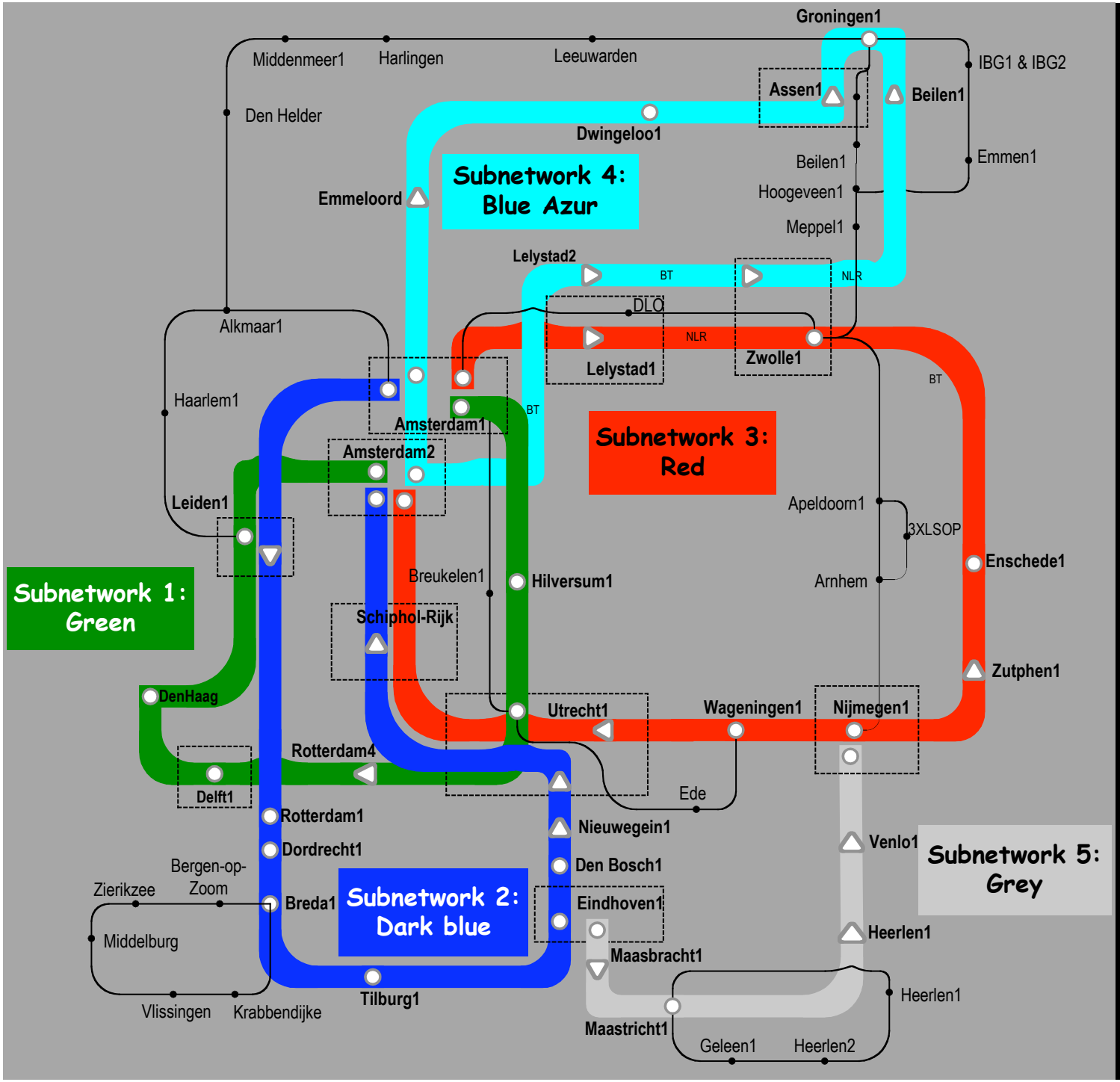
- universities;
- academic hospitals;
- most polytechnics;
- research centers.

with an indirect ~750K user base

~ 6000 km

scale comparable to railway system

SURF NET

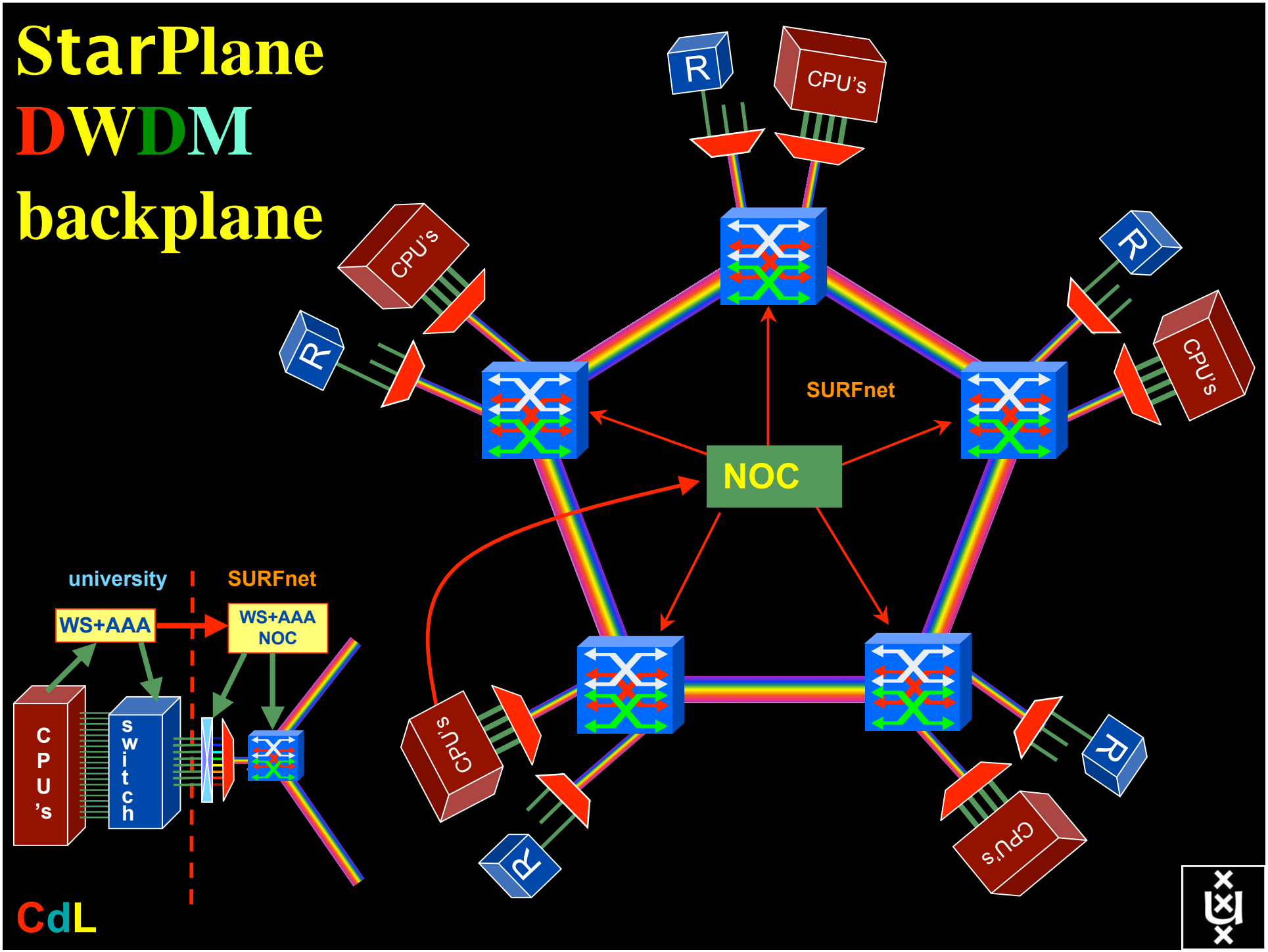


Common Photonic Layer (CPL) in SURFnet6

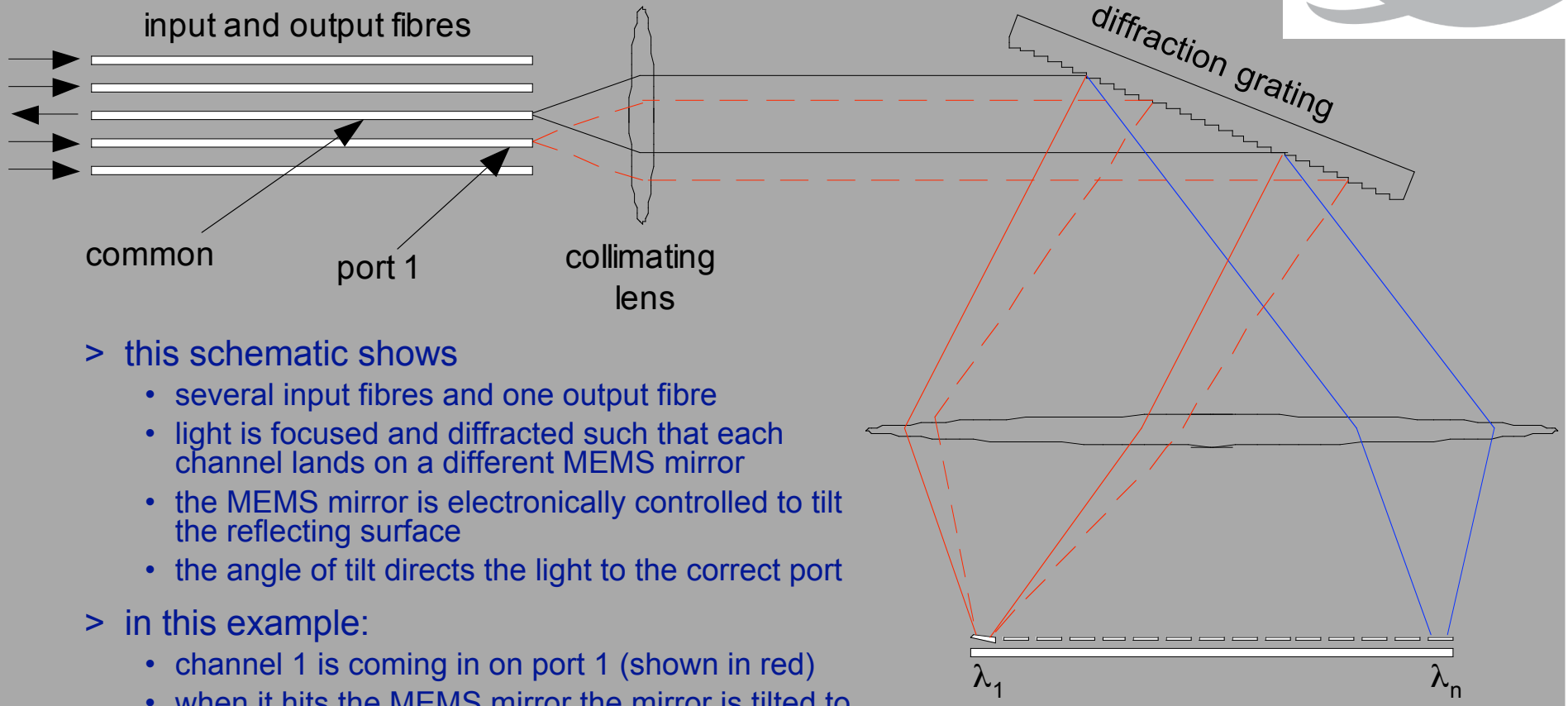
supports up to 72 Lambda's of 10 G each

future: 40/100 G.

StarPlane DWDM backplane



Module Operation



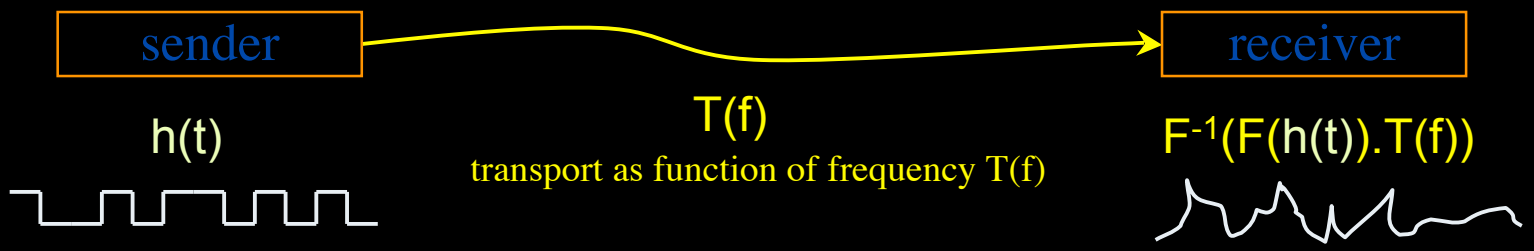
> this schematic shows
- several input fibres and one output fibre
- light is focused and diffracted such that each channel lands on a different MEMS mirror
- the MEMS mirror is electronically controlled to tilt the reflecting surface
- the angle of tilt directs the light to the correct port

> 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

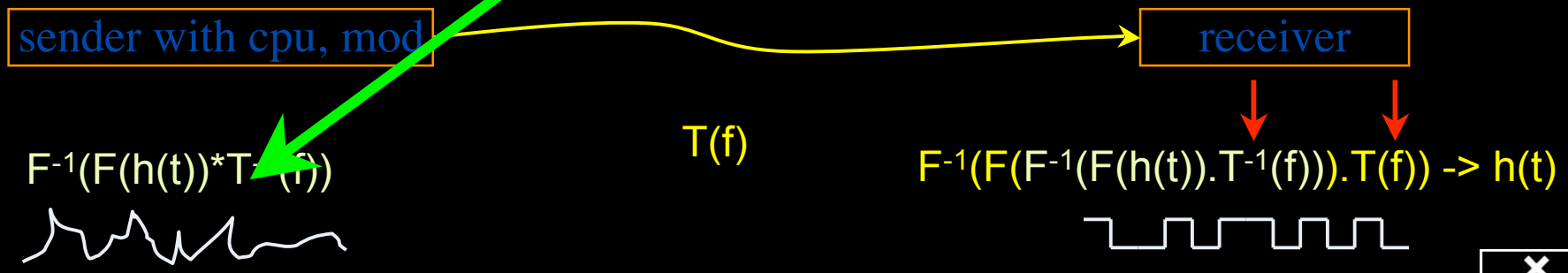
Dispersion compensating modem: eDCO from NORTEL

(Try to Google eDCO :-)



Solution in 5 easy steps for dummy's :

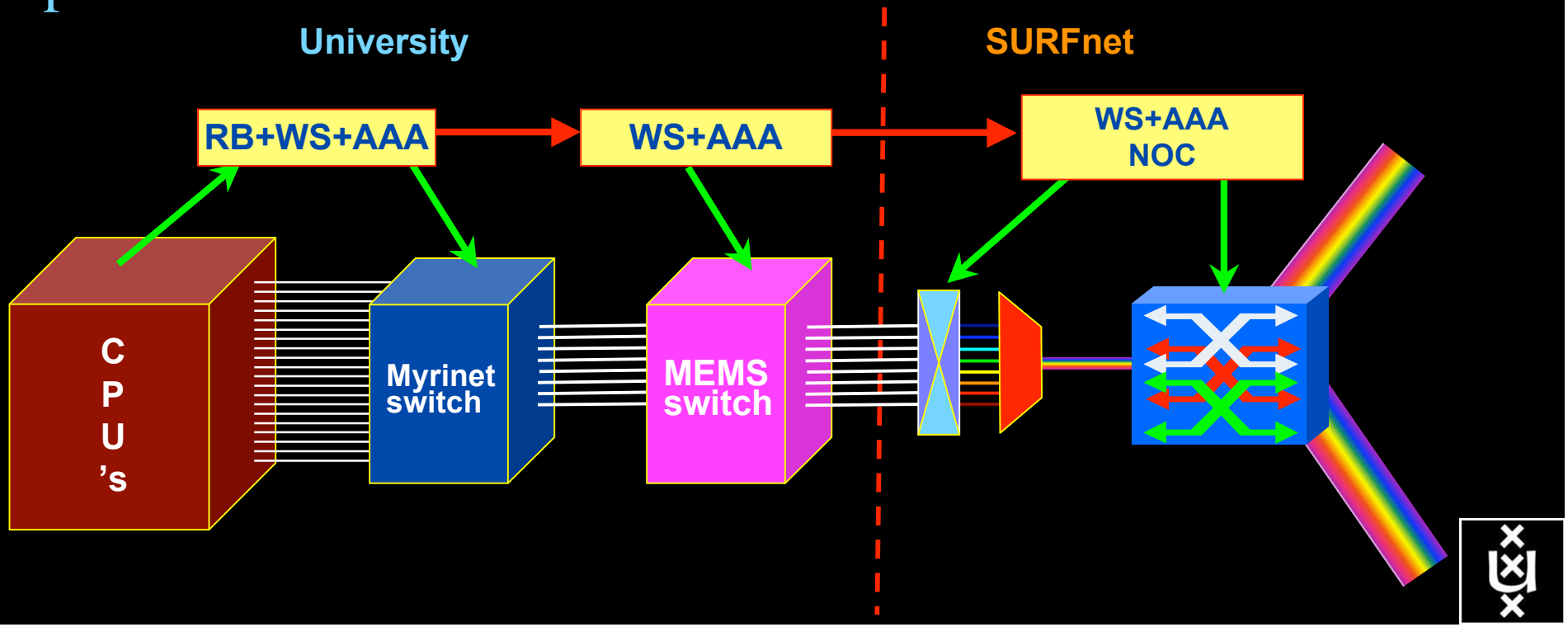
1. try to figure out T(f) by trial and error
2. invert T(f) -> $T^{-1}(f)$
3. computationally multiply $T^{-1}(f)$ with Fourier transform of bit pattern to send
4. inverse Fourier transform the result from frequency to time space
5. modulate laser with resulting $h'(t) = F^{-1}(F(h(t)).T^{-1}(f))$



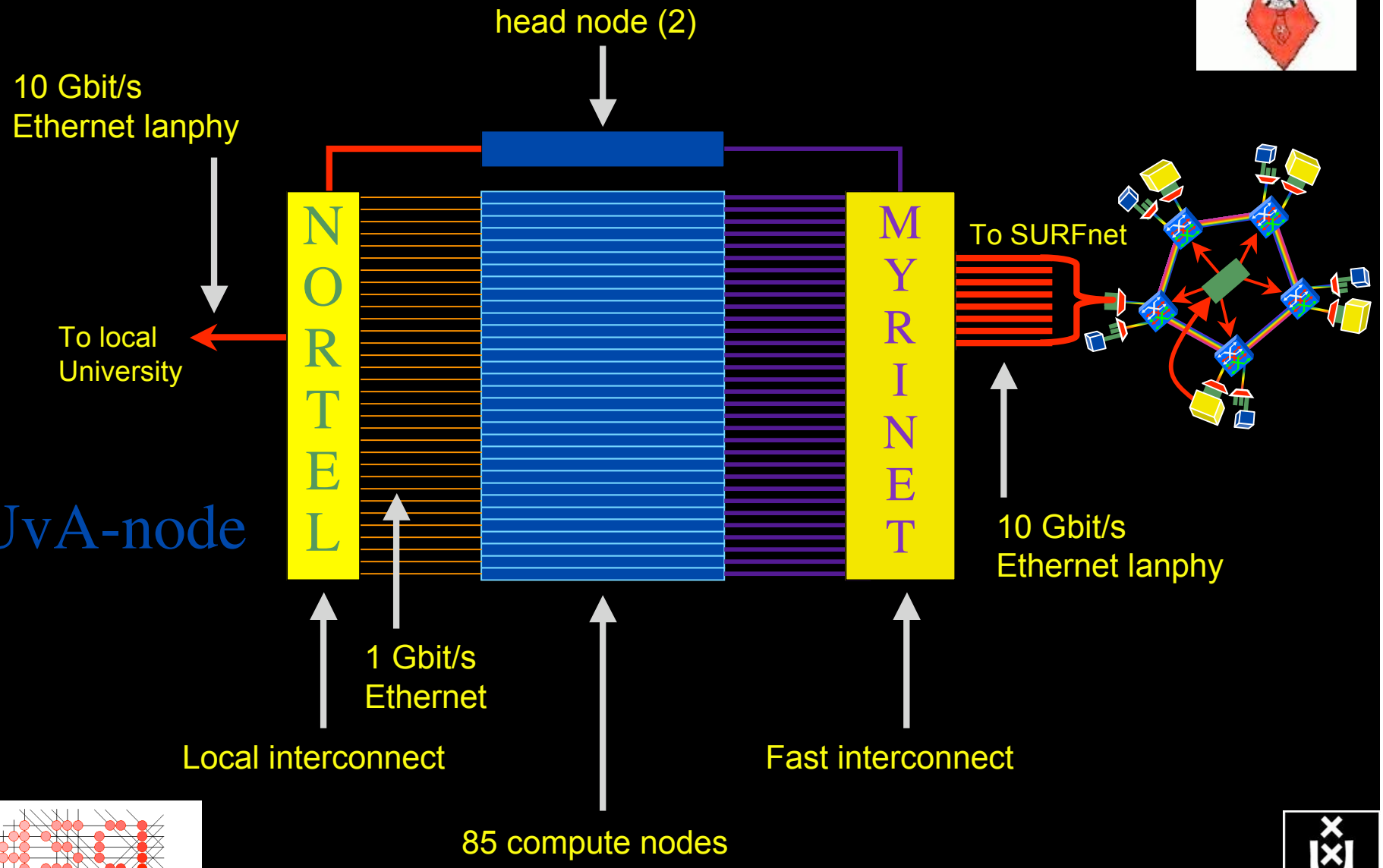
(ps. due to power ~ square E the signal to send **looks** like uncompensated received but is not)

The challenge for sub-second switching

- 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:



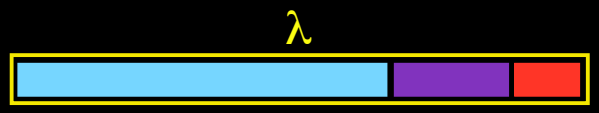
DAS-3 Cluster Architecture



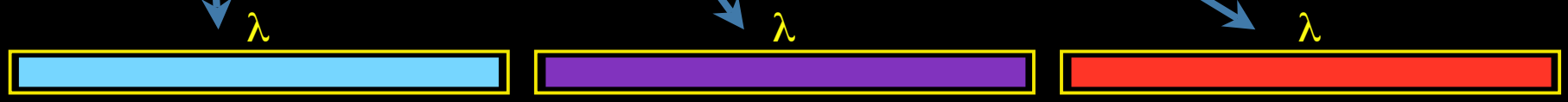
Advanced School for Computing and Imaging

QOS in a non destructive way!

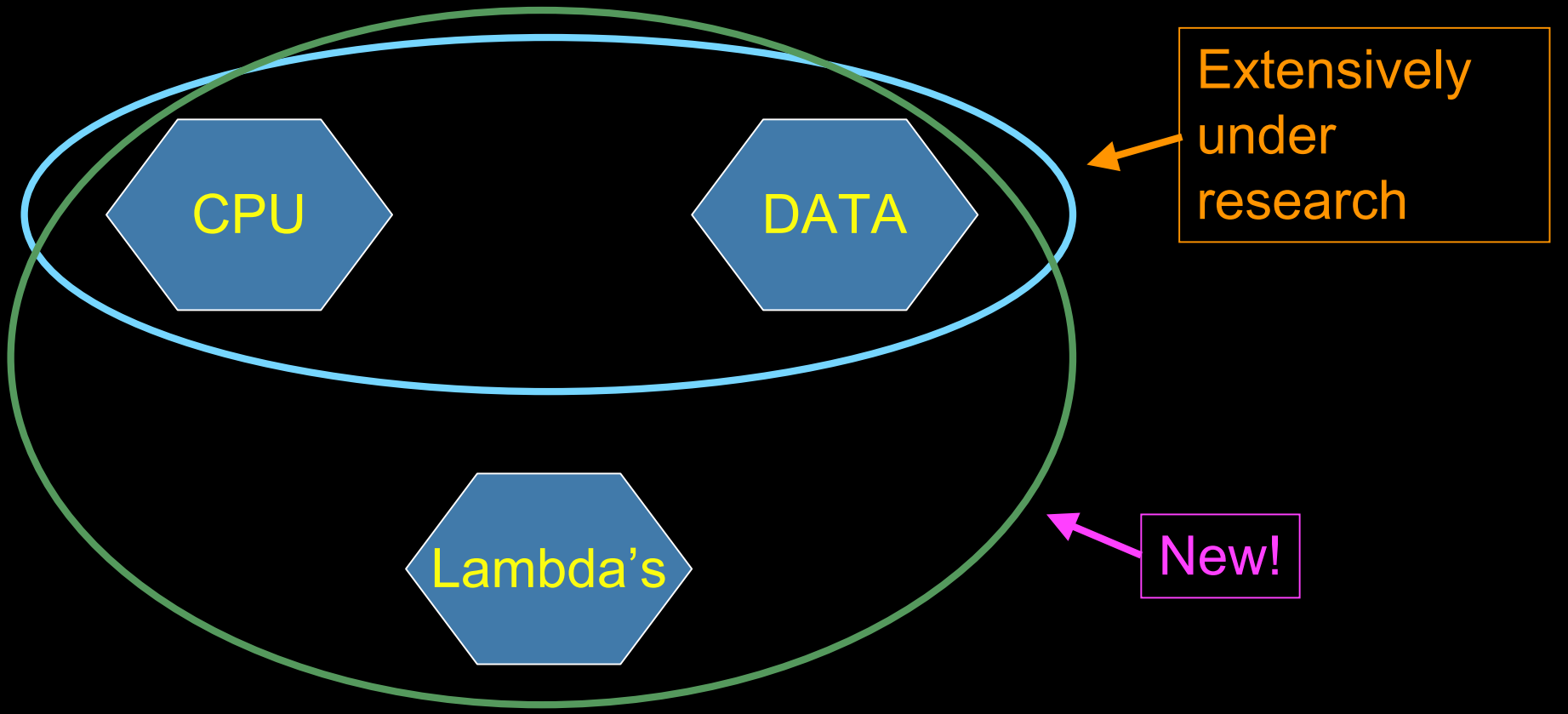
- Destructive QOS:
  - have a link or λ
  - set part of it aside for a lucky few under higher priority
  - rest gets less service



- 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!



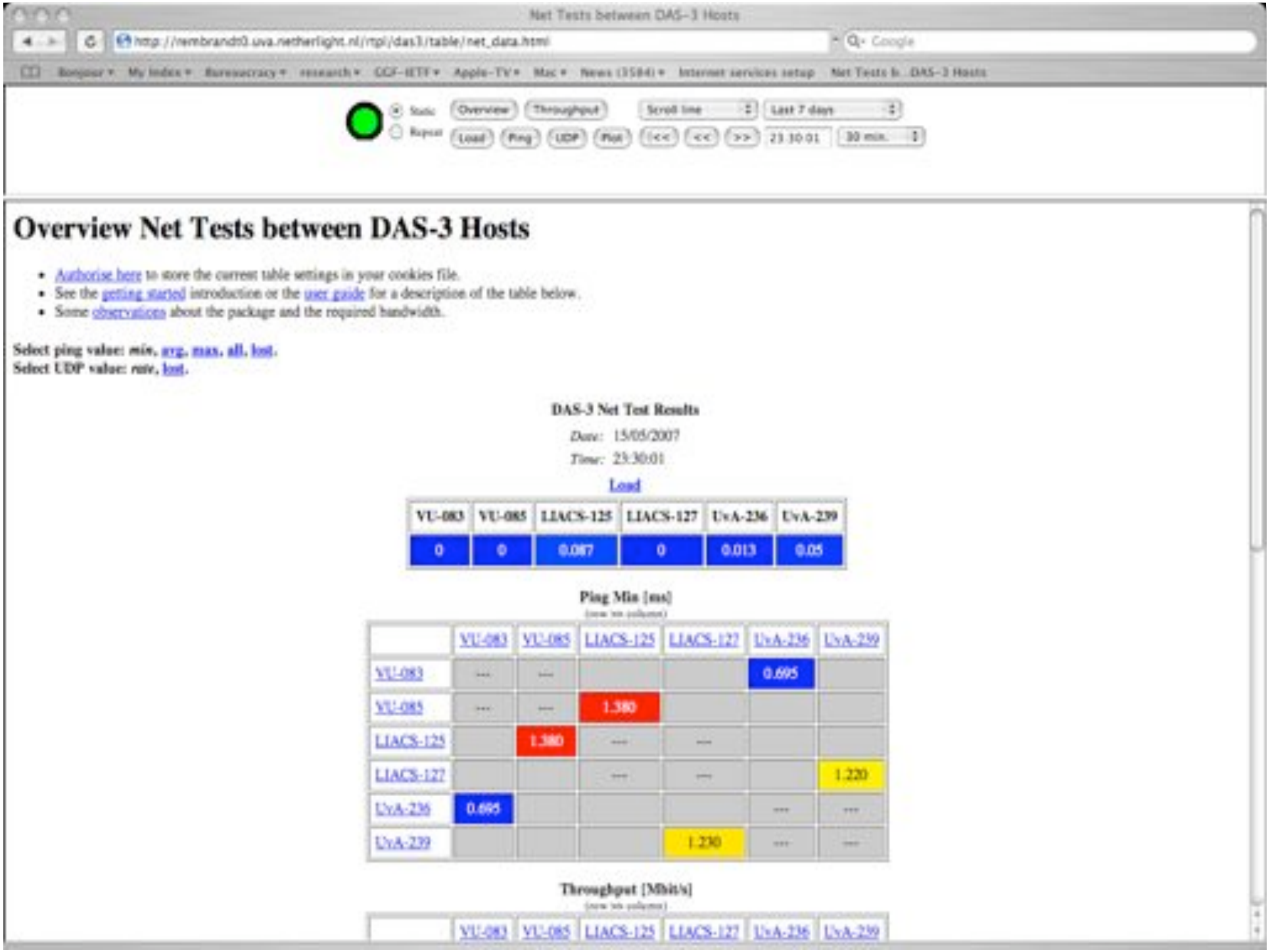
GRID Co-scheduling problem space

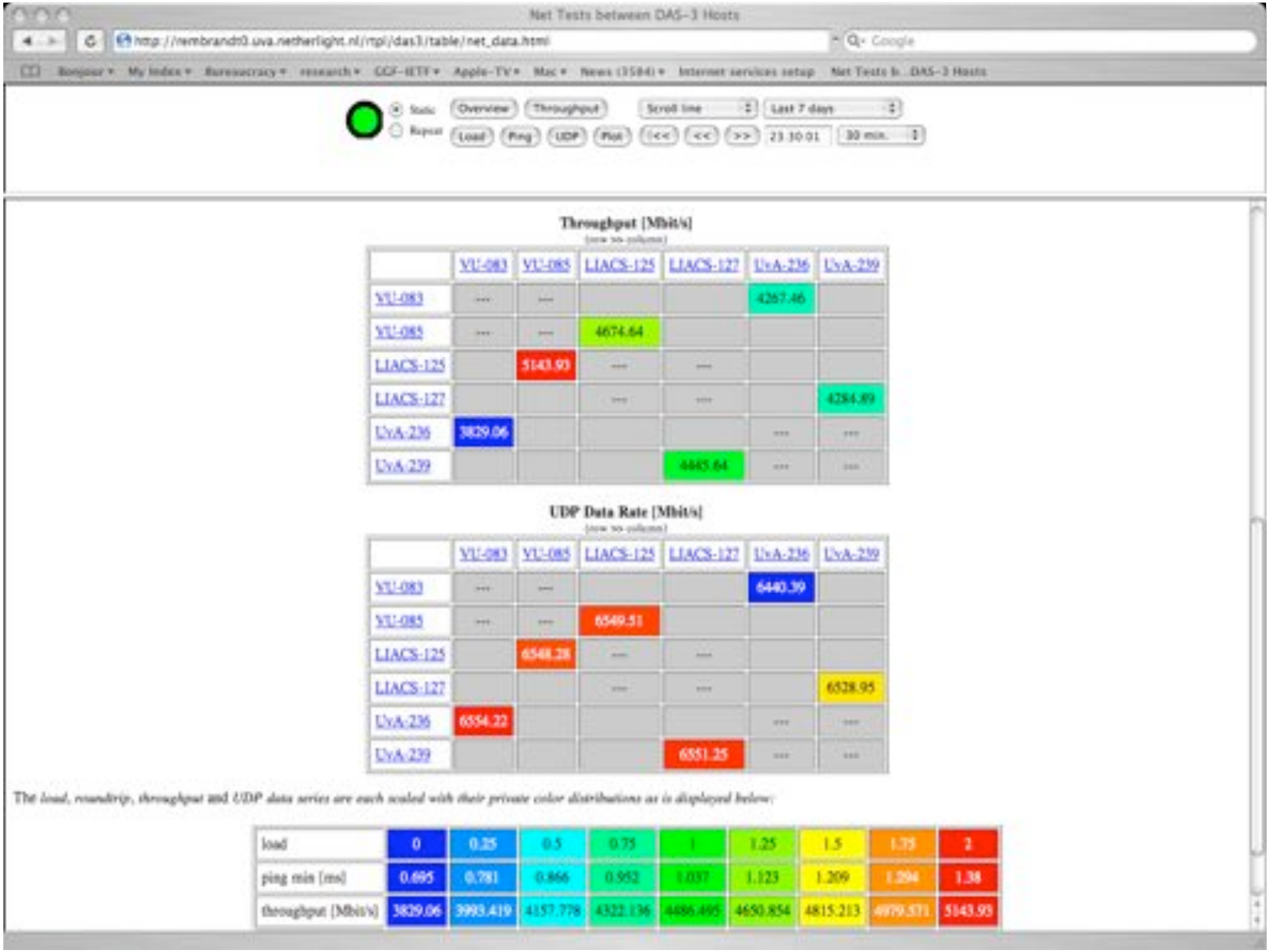


The StarPlane vision is to give flexibility directly to the applications by allowing them to choose the logical topology in real time, ultimately with sub-second lambda switching times on part of the SURFnet6 infrastructure.

What makes StarPlane fly?

- Wavelength Selective Switches
  - for the “low cost” photonics
- Sandbox by confining StarPlane to one band
  - for experimenting on a production network
- Optimization of the controls to turn on/off a Lambda
  - direct access to part of the controls at the NOC
- electronic Dynamically Compensating Optics (eDCO)
  - to compensate for changing lengths of the path
- traffic engineering
  - to create the OPN topologies needed by the applications
- Open Source GMPLS
  - to facillitate policy enabled cross domain signalling


Net Tests between DAS-3 Hosts
http://rembrandt0.uva.netherlight.nl/rtpl/das3/table/net_data.html
Overview
Throughput
Load
Ping
UDP
Plot
Last 7 days
30 min.
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.
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: 15/05/2007
Time: 23:30:01
Load
VU-083 VU-085 LIACS-125 LIACS-127 UvA-236 UvA-239
0 0 0.087 0 0.013 0.05
Ping Min [ms]
VU-083 VU-085 LIACS-125 LIACS-127 UvA-236 UvA-239
VU-083 --- --- 0.695
VU-085 --- --- 1.380
LIACS-125 1.380 --- ---
LIACS-127 --- --- 1.220
UvA-236 0.695 --- ---
UvA-239 1.230 --- ---
Throughput [Mbit/s]
VU-083 VU-085 LIACS-125 LIACS-127 UvA-236 UvA-239


Net Tests between DAS-3 Hosts
http://rembrandt0.uva.netherlight.nl/rpi/das3/table/net_data.html
Google
Bonjour
My Index
Bureaucracy
research
CGF-IETF
Apple-TV
Mac
News (3584)
Internet services setup
Net Tests b...DAS-3 Hosts
Static
Repeat
Overview
Throughput
Scroll line
Last 7 days
Load
Ping
UDP
Plot
<<
<<
>>
23:30:01
30 min.
Throughput [Mbit/s]
(row vs column)
VU-083
VU-085
LIACS-125
LIACS-127
UvA-236
UvA-239
4267.46
4674.64
5143.93
4284.89
3829.06
4445.64
UDP Data Rate [Mbit/s]
(row vs column)
VU-083
VU-085
LIACS-125
LIACS-127
UvA-236
UvA-239
6440.39
6549.51
6548.28
6528.95
6554.22
6551.25
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]
0.695
0.781
0.866
0.952
1.037
1.123
1.209
1.294
1.38
throughput [Mbit/s]
3829.06
3993.419
4157.778
4322.136
4486.495
4650.854
4815.213
4979.571
5143.93

Heterogeneous clusters

(# of unused ports)

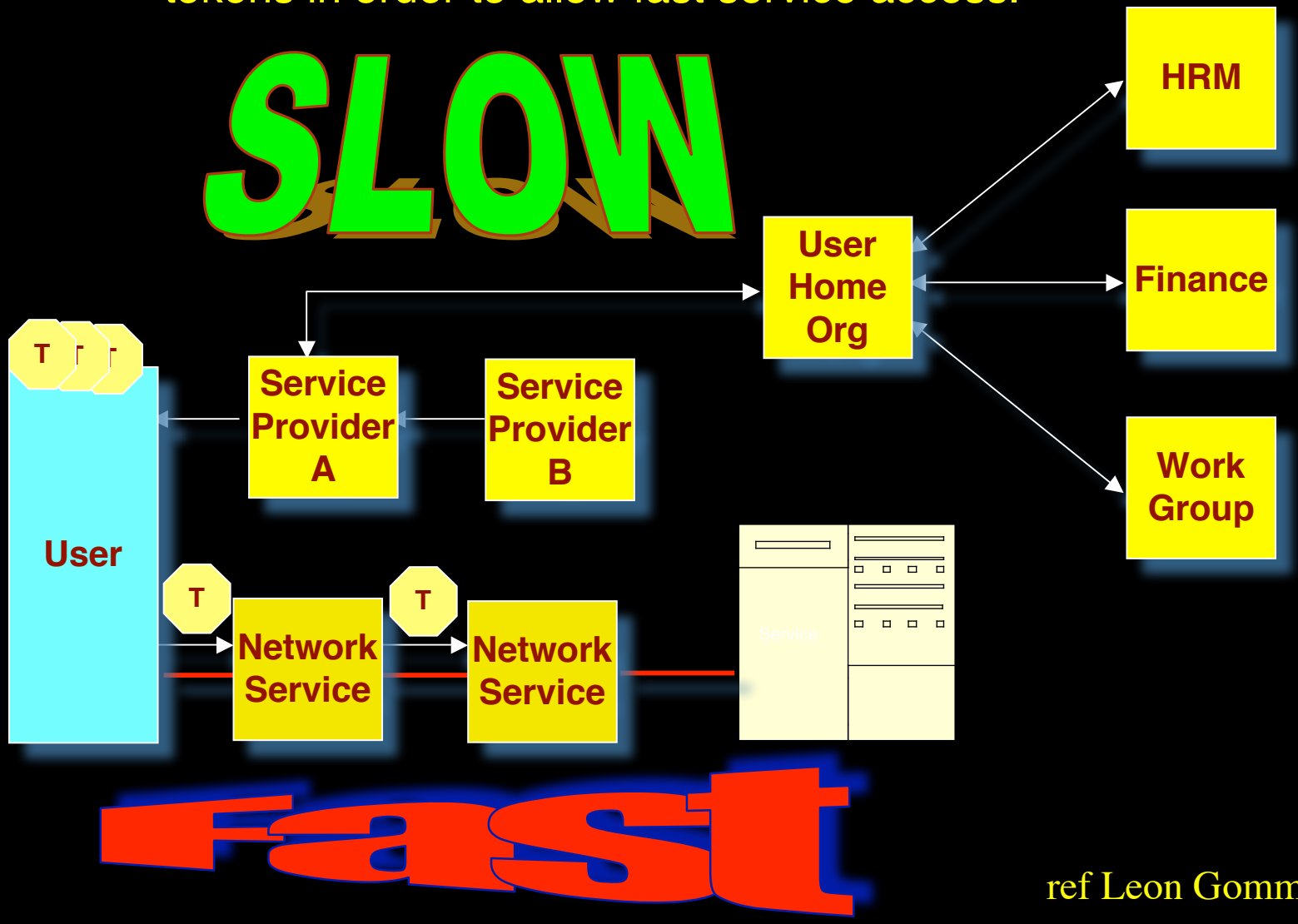
	LU	TUD	UvA-VLE	UvA-MN	VU	TOTALS
Head						
* storage	10TB	5TB	2TB	2TB	10TB	29TB
* CPU	2x2.4GHz DC	2x2.4GHz DC	2x2.2GHz DC	2x2.2GHz DC	2x2.4GHz DC	46.4 GHz
* memory	16GB	16GB	8GB	16GB	8GB	64GB
* Myri 10G	1		1	1	1	40 Gb/s
* 10GE	1	1	1	1	1	50 Gb/s
Compute	32	68	40 (+1)	46	85	271
* storage	400GB	250GB	250GB	2x250GB	250GB	84 TB
* CPU	2x2.6GHz	2x2.4GHz	2x2.2GHz DC	2x2.4GHz	2x2.4GHz DC	1.9 THz
* memory	4GB	4GB	4GB	4GB	4GB	1048 GB
* Myri 10G	1		1	1	1	2030 Gb/s
Myrinet						
* 10G ports	33 (7)		41	47	86 (2)	2070 Gb/s
* 10GE ports	8		8	8	8	320 Gb/s
Nortel						
* 1GE ports	32 (16)	136 (8)	40 (8)	46 (2)	85 (11)	339 Gb/s
* 10GE ports	1 (1)	9 (3)	2	2	1 (1)	

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!

Use AAA concept to split (time consuming) service authorization process from service access using secure tokens in order to allow fast service access.



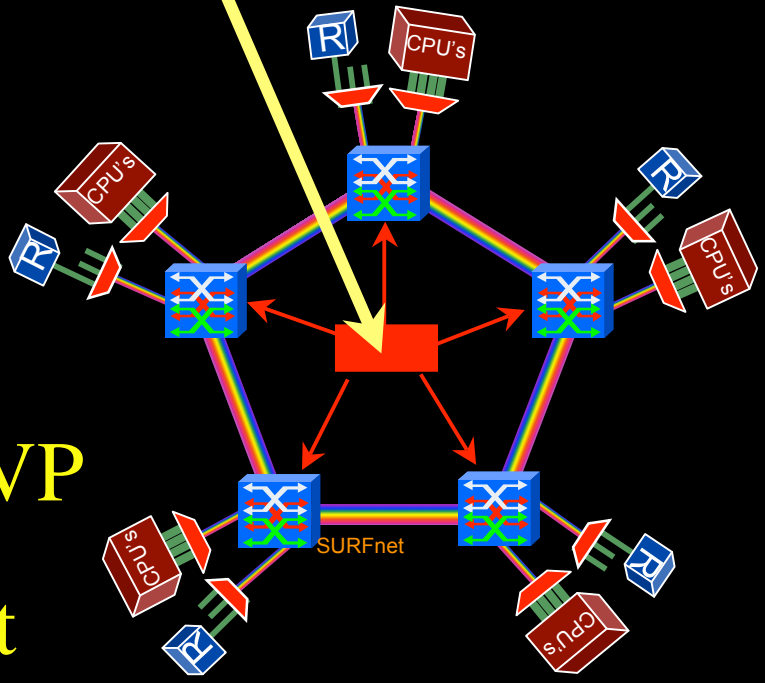
ref Leon Gommans

Control Plane

Application Scheduler

WS

DRAC

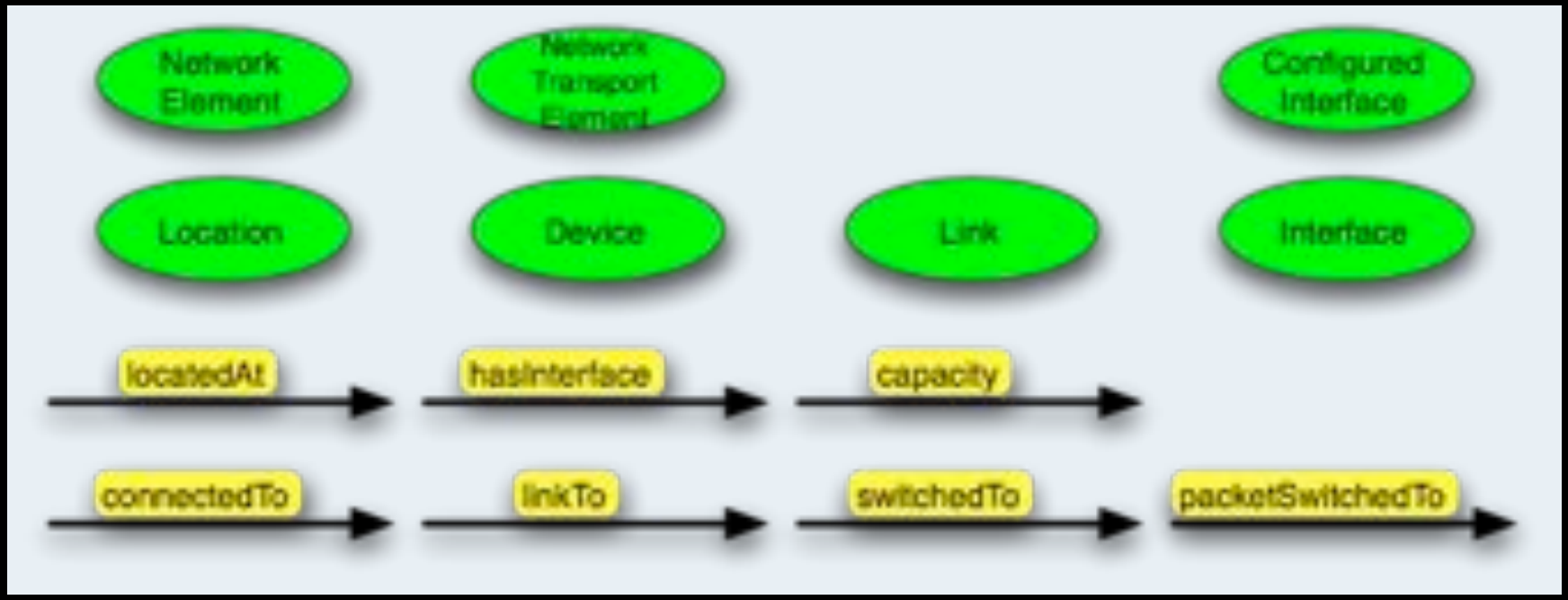


- currently implementing WS
- study interaction with TBN/RSVP
- also photonic switching at client

StarPlane and NDL

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

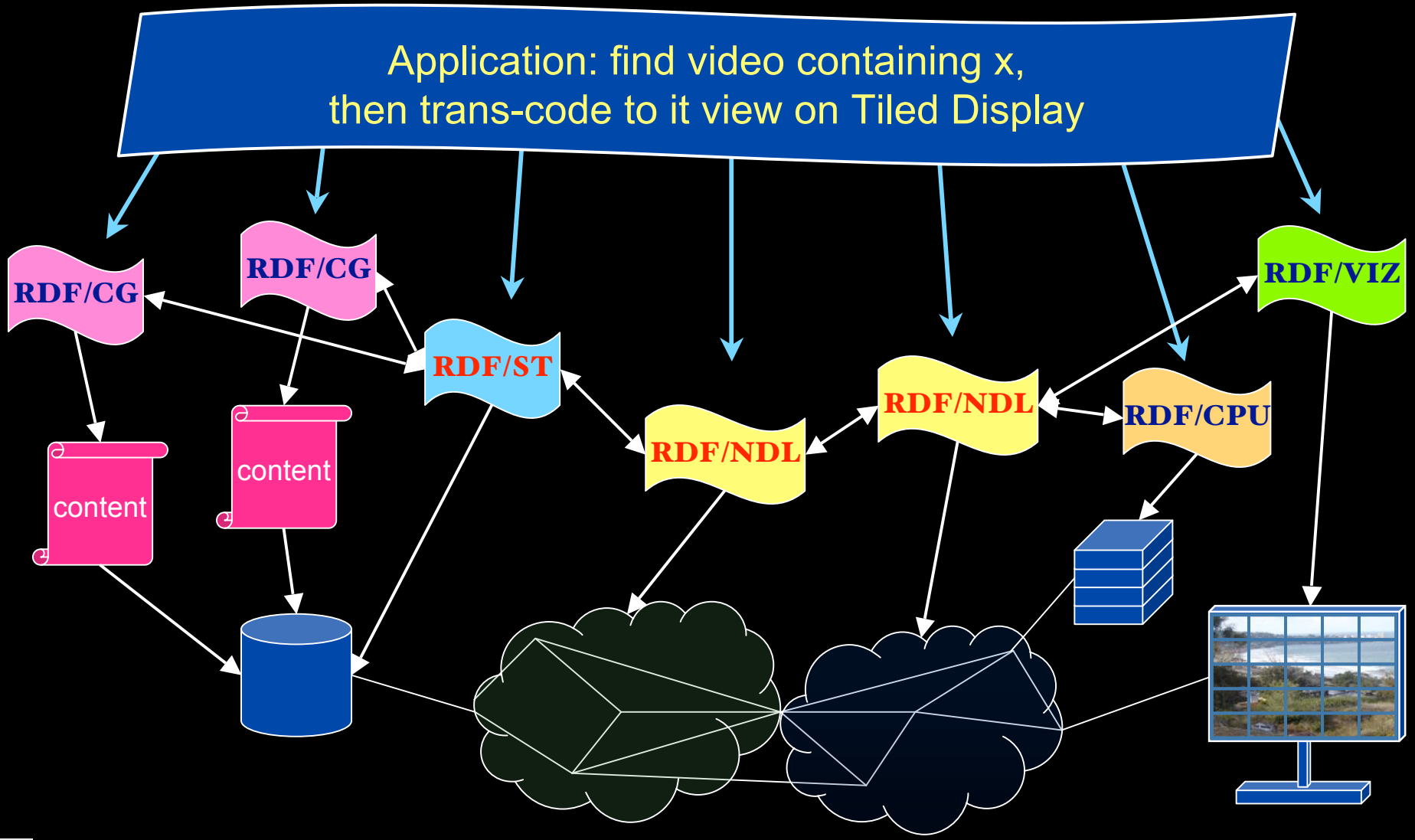
NDL is an RDF data model, based on idea of Semantic Web, for network topology descriptions.



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

ref: Talk from Paola Grosso on NDL/RDF

RDF describing Infrastructure

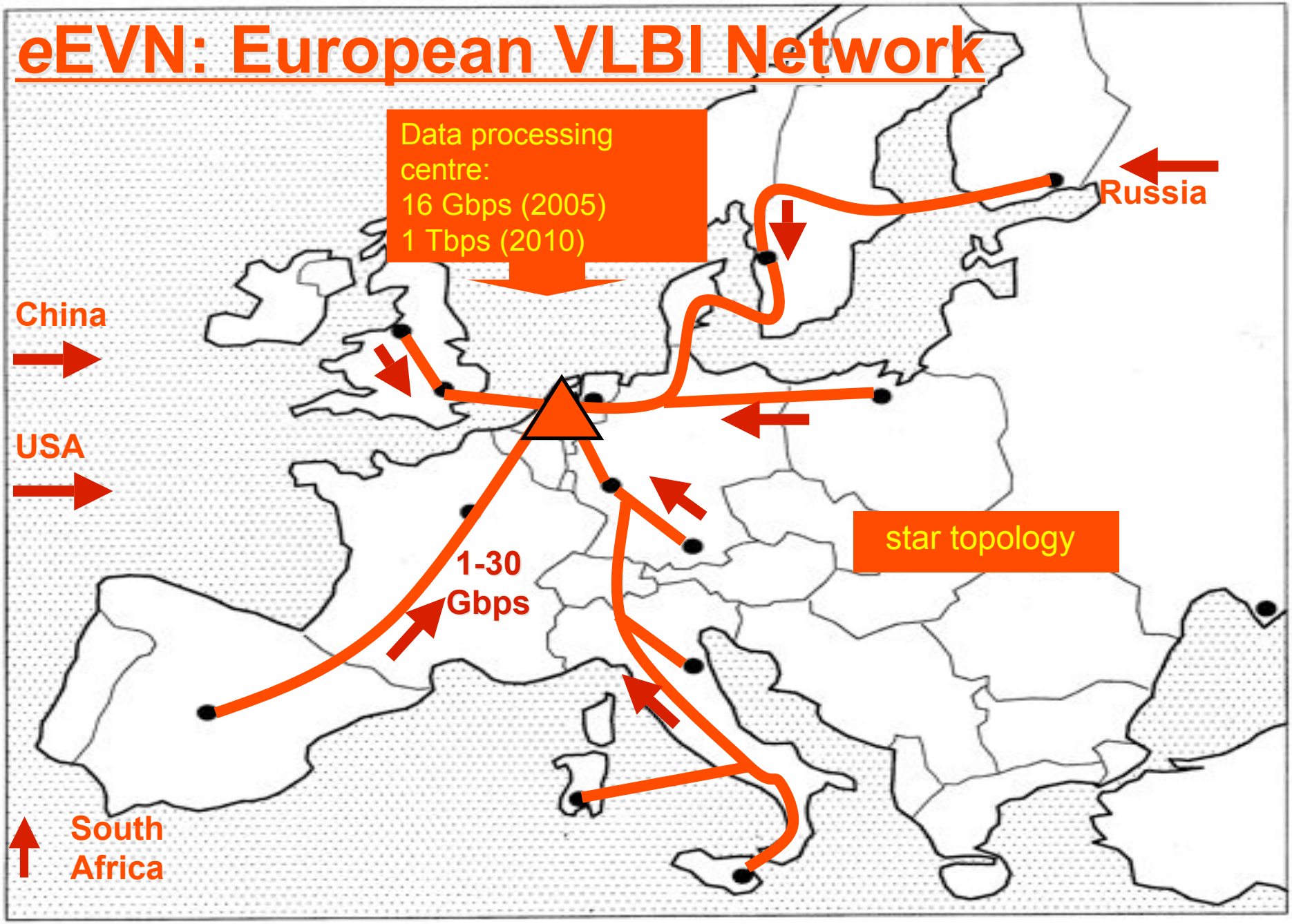


PG&CdL

StarPlane Applications

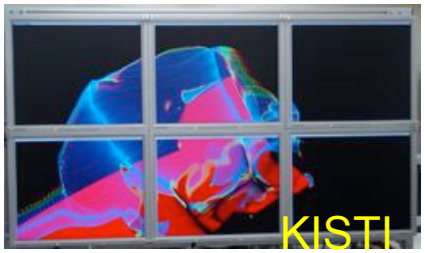
- Large ‘stand-alone’ file transfers
  - User-driven file transfers
  - Nightly backups
  - Transfer of medical data files (MRI)
- Large file (speedier) Stage-in/Stage-out
  - MEG modeling (Magneto encephalography)
  - Analysis of video data
- Application with static bandwidth requirements
  - Distributed game-tree search
  - Remote data access for analysis of video data
  - Remote visualization
- Applications with dynamic bandwidth requirements
  - Remote data access for MEG modeling
  - SCARI

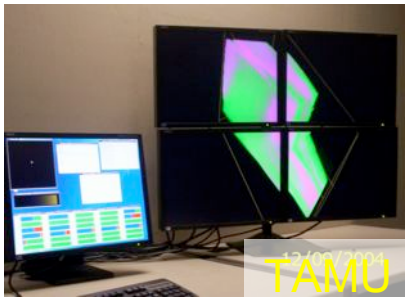
eEVN: European VLBI Network

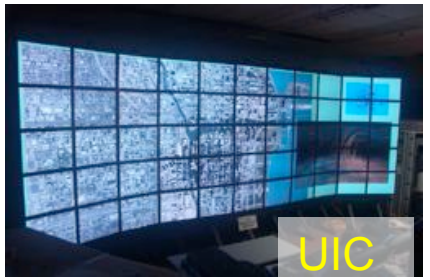


This slide courtesy of Richard Schilizzi <schilizzi@jive.nl>

US and International OptIPortal Sites







UCSD
UCIrvine

CineGrid@SARA

Tera-Thinking

- 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
- think back to teraflop computing!
  - MPI makes it a teraflop machine
- massive parallel channels in hosts, NIC's
- TeraApps programming model supported by
  - TFlops -> MPI / Globus
  - TBytes -> OGSA/DAIS
  - TPixels -> SAGE
  - TSensors -> LOFAR, LHC, LOOKING, CineGrid, ...
  - Tbit/s -> ?

ref Larry Smarr & CdL

Questions ?

Thanks to:

SURFnet, NWO (grant 643.000.504), NORTEL

Team: Li Xu, Jason Maasen,JP Velders, Leon Gommans, Paola Grosso, Herbert Bos, Henri Bal

Special thanks to Kees Neggers.