Lambda-Grid developments

History - Present - Future

Cees de Laat

SURFnet

EU

BSIK

NWO

University of Amsterdam

TNO

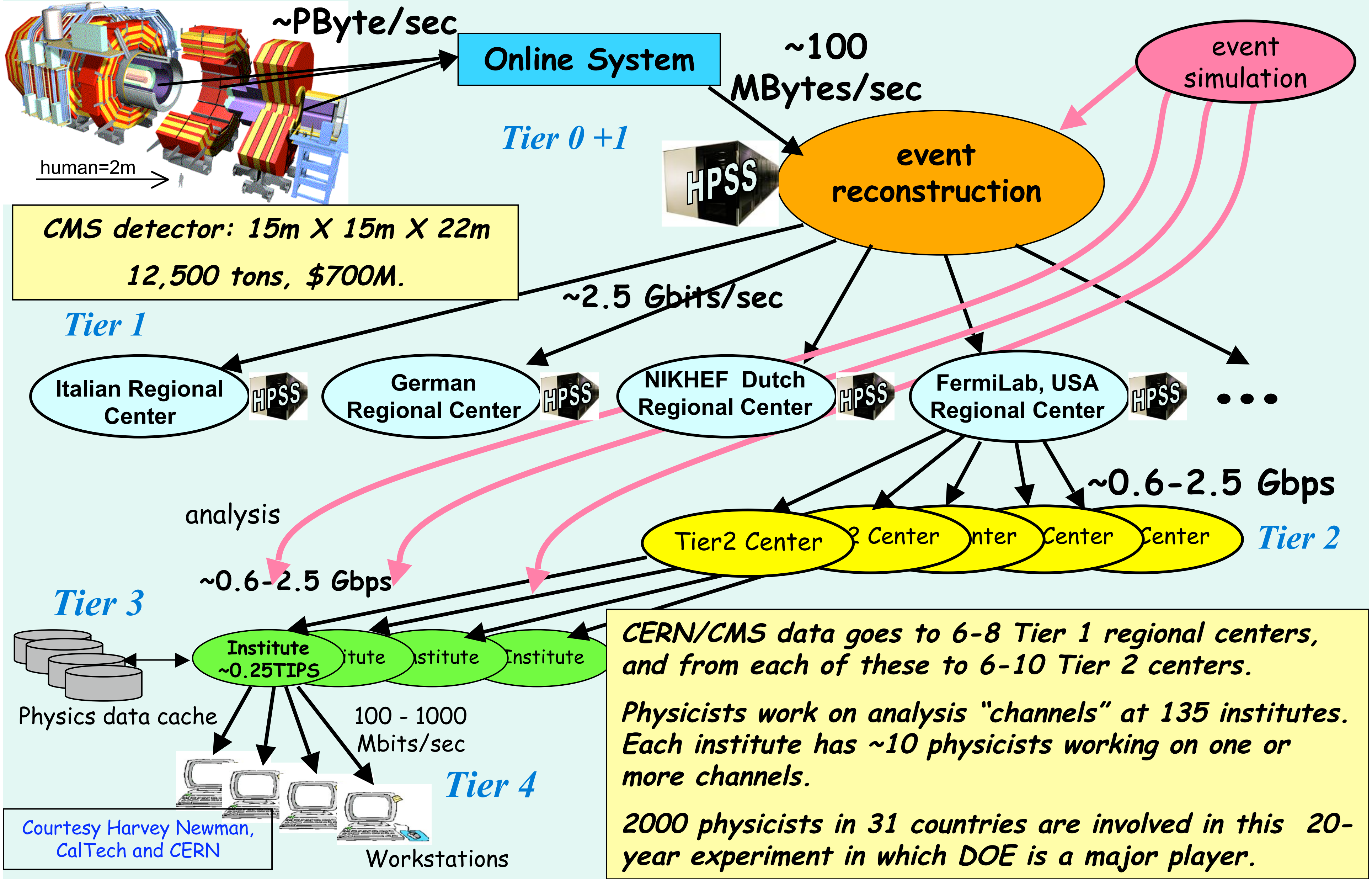
NCF

Contents

1. The need for hybrid networking
2. StarPlane; a grid controlled photonic network
3. RDF/Network Description Language
4. Tera-networking
5. Programmable networks

LHC Data Grid Hierarchy

CMS as example, Atlas is similar



~PByte/sec

Online System

~100 MBytes/sec

human=2m

Tier 0 +1

HPSS

event reconstruction

event simulation

CMS detector: 15m X 15m X 22m

12,500 tons, $700M.

~2.5 Gbits/sec

Tier 1

Italian Regional Center

German Regional Center

NIKHEF Dutch Regional Center

FermiLab, USA Regional Center

• • •

~0.6-2.5 Gbps

Tier2 Center

Tier 2

analysis

~0.6-2.5 Gbps

Tier 3

Institute ~0.25TIPS

Institute

Physics data cache

100 - 1000 Mbits/sec

Tier 4

Workstations

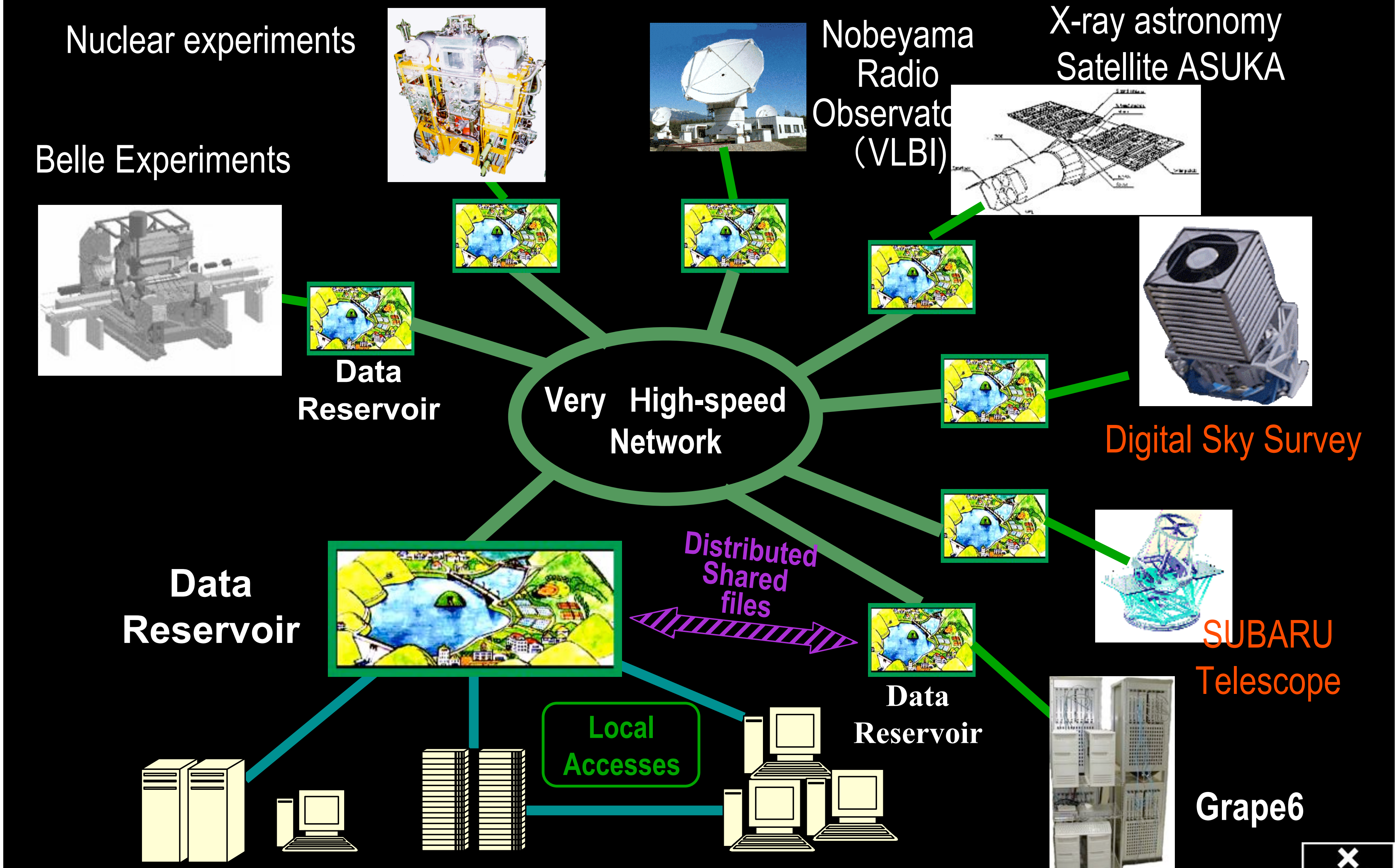
Courtesy Harvey Newman, CalTech and CERN

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.

Data intensive scientific computation through global networks



Nuclear experiments

Nobeyama Radio Observato (VLBI)

X-ray astronomy Satellite ASUKA

Belle Experiments

Data Reservoir

Very High-speed Network

Digital Sky Survey

Distributed Shared files

Data Reservoir

SUBARU Telescope

Local Accesses

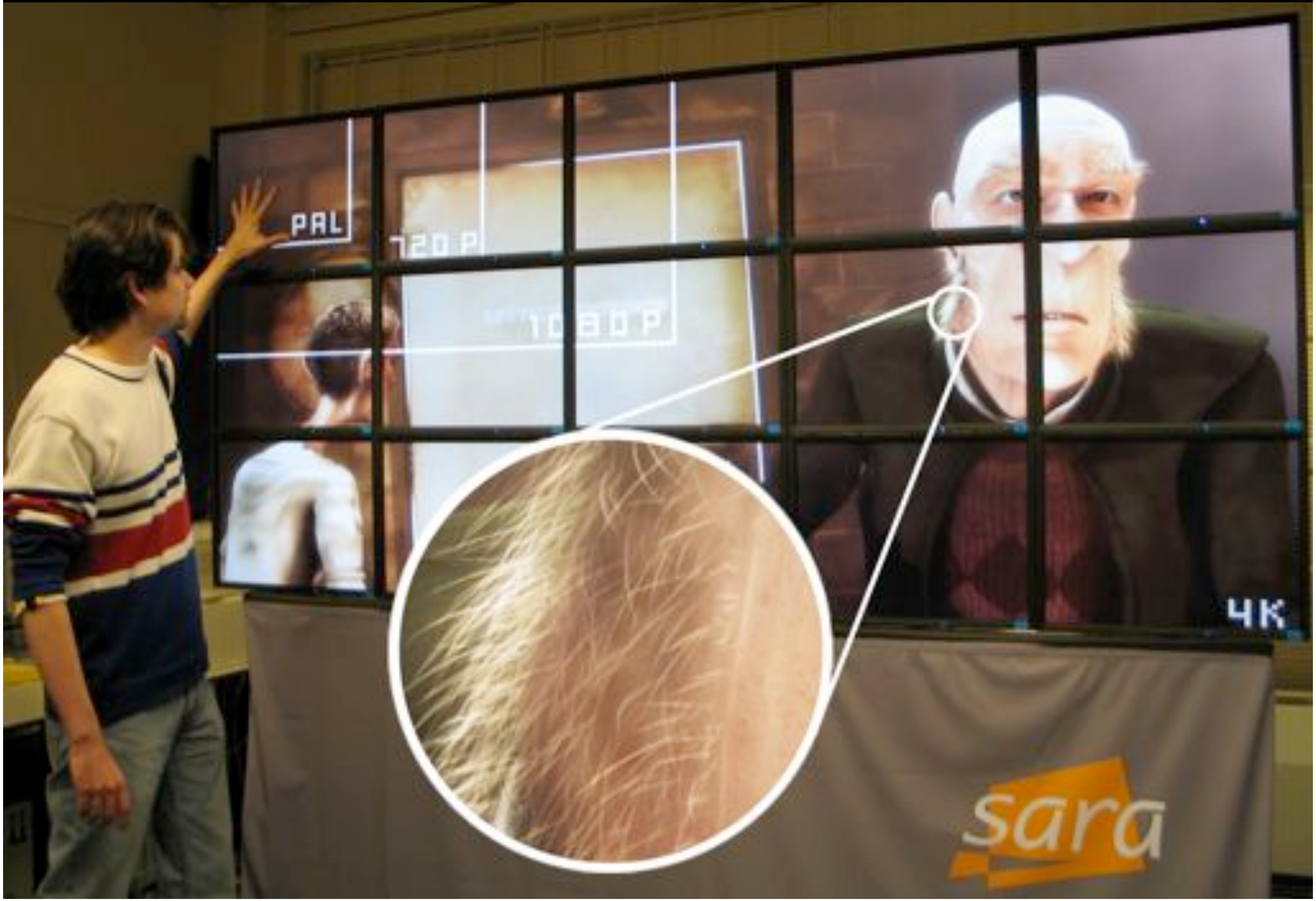
Data Reservoir

Grape6

Data analysis at University of Tokyo

CineGrid@SARA

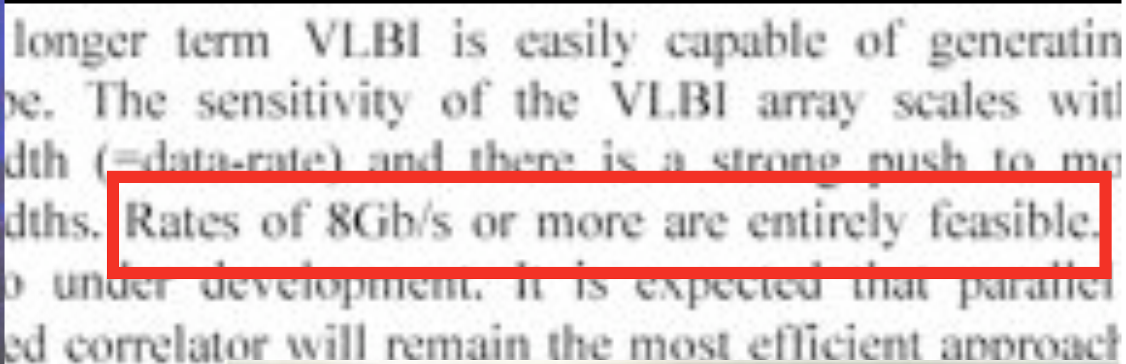
2b of 5



Sensor Grids

~ 40 Tbit/s
www.lofar.org

eVLBI

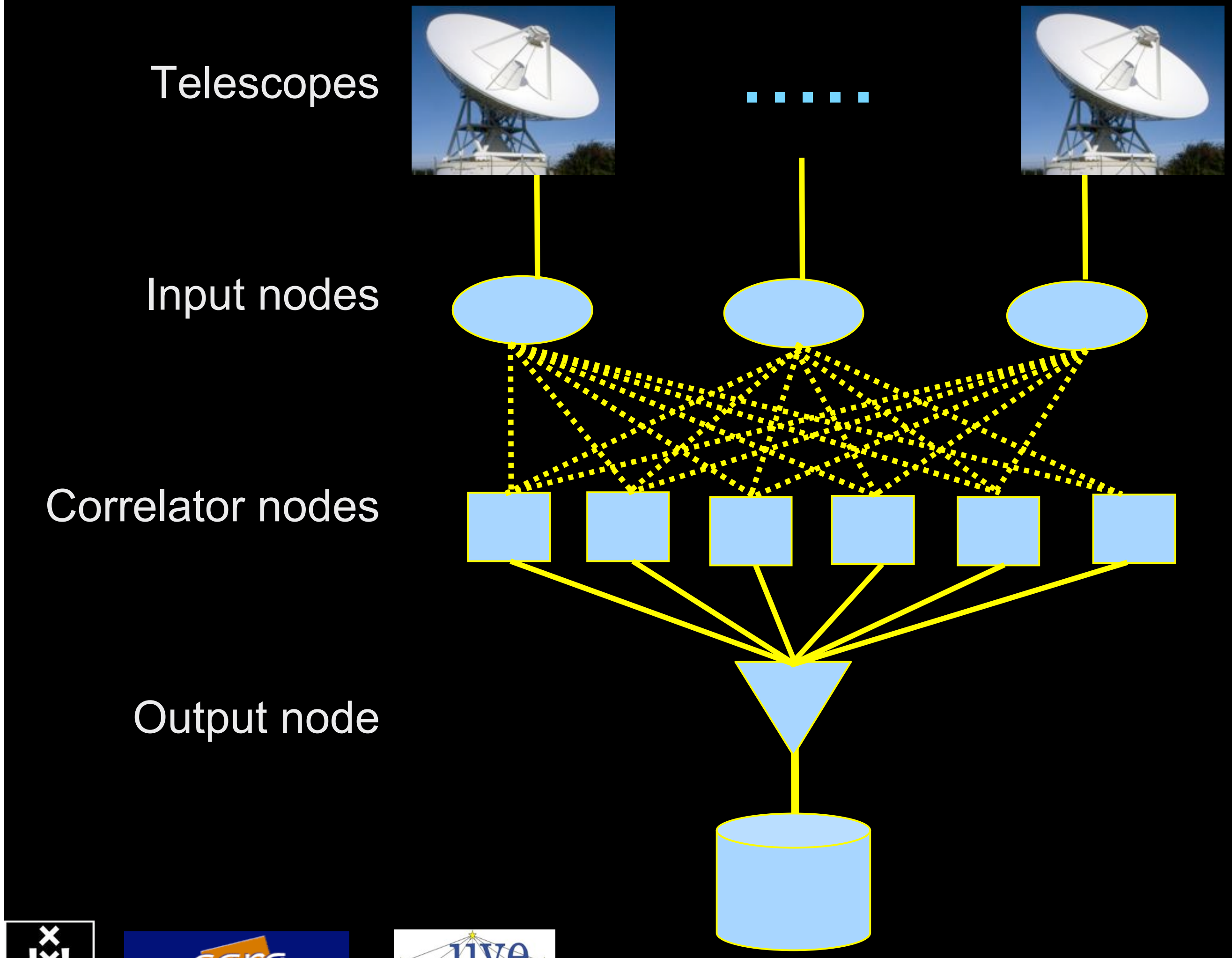


longer term VLBI is easily capable of generatin
pe. The sensitivity of the VLBI array scales wit
dth (=data-rate) and there is a strong push to mo
dths. Rates of 8Gb/s or more are entirely feasible.
o under development. It is expected that parallel
ed correlator will remain the most efficient approach
olves dist
, multi-gig
relator and
factor.

Westerbork Synthesis Radio Telescope - Netherlands

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.



To equal the hardware correlator we need:

16 streams of 1Gbps

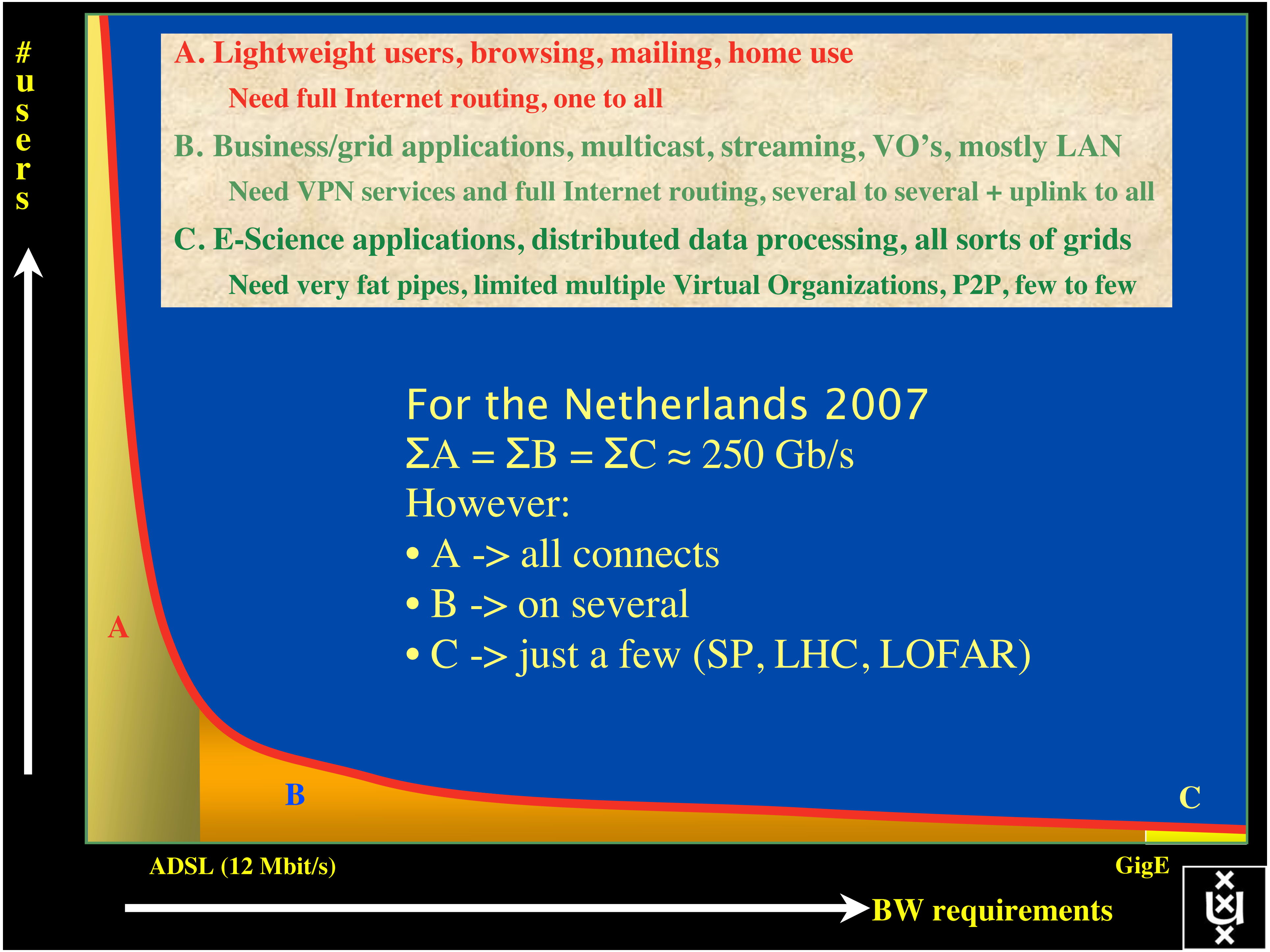
16 * 1Gbps of data

2 Tflops CPU power

2 TFlop / 16 Gbps =

1000 flops/byte

THIS IS A DATA FLOW PROBLEM !!!



A. Lightweight users, browsing, mailing, home use

Need full Internet routing, one to all

B. Business/grid applications, multicast, streaming, VO's, mostly LAN

Need VPN services and full Internet routing, several to several + uplink to all

C. E-Science applications, distributed data processing, all sorts of grids

Need very fat pipes, limited multiple Virtual Organizations, P2P, few to few

For the Netherlands 2007

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

However:

- A -> all connects
- B -> on several
- C -> just a few (SP, LHC, LOFAR)

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 and L2
- Give each packet in the network the service it needs, but no more !

L1 ≈ 2-3 k$/port

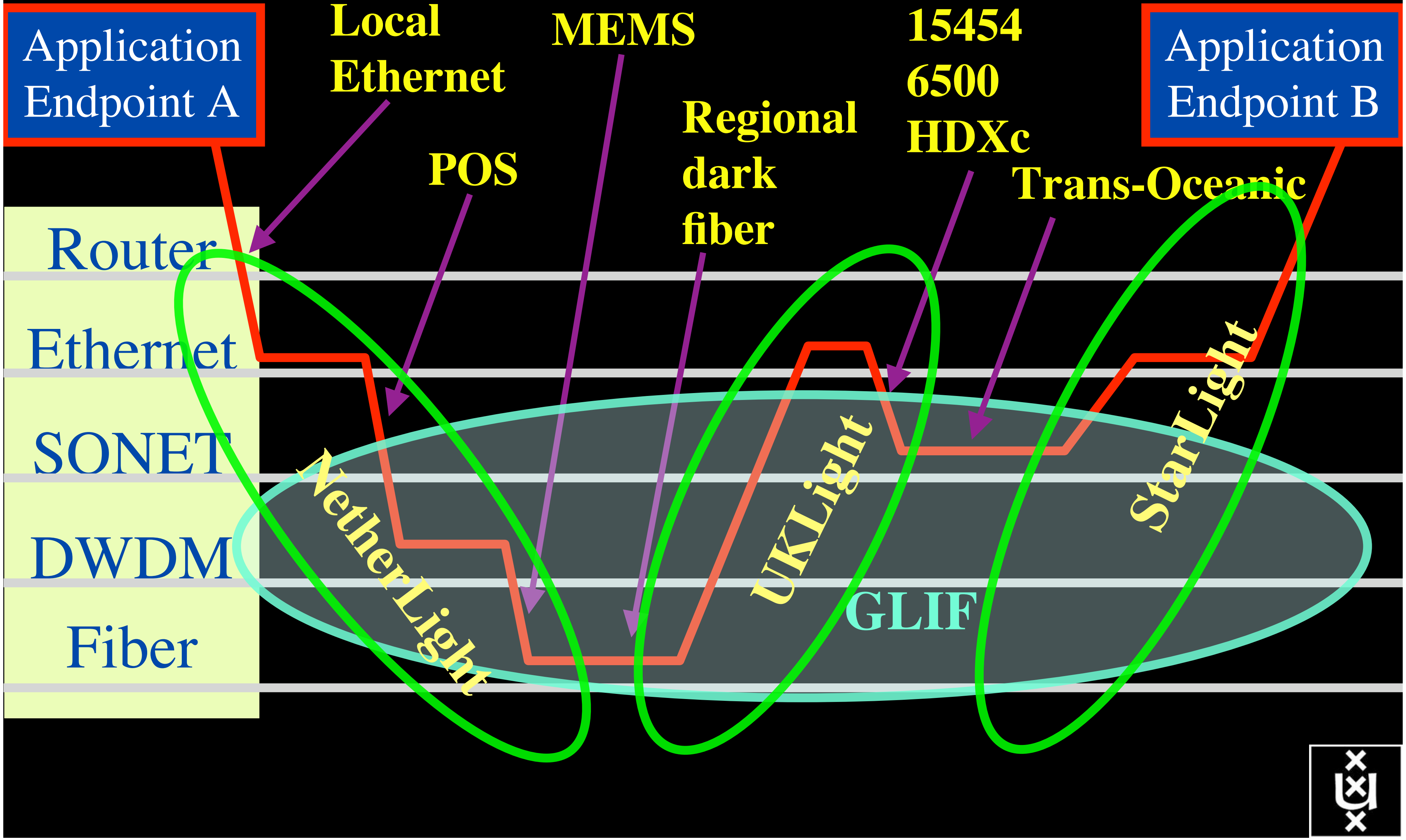
L2 ≈ 5-8 k$/port

L3 ≈ 75+ k$/port

Trends

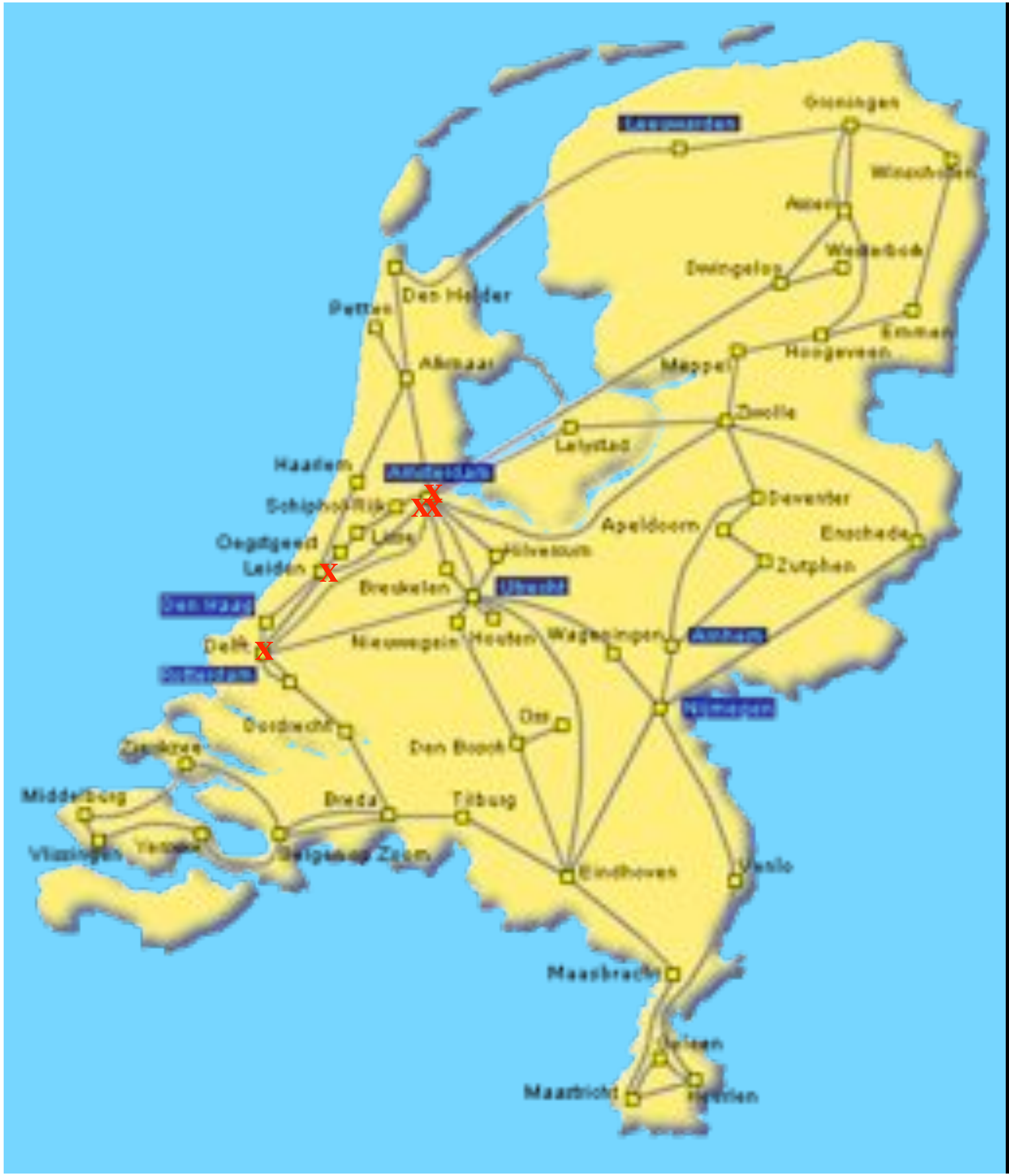
- We have made baby-steps on the path to optical networking
  - Still many mails and phone calls
- See several trends:
  - lambda's get fatter and cheaper
  - photonic technology cheap per bandwidth
  - embedded computation capacity increasing
  - latency and high bandwidth congestion avoidance conflict
  - ethernet is getting circuit properties (PBT)
  - applications need more and more predictable behaviour

How low can you go?



c5 of 6

The playfield => GLIF



In The Netherlands SURFnet connects between 180:

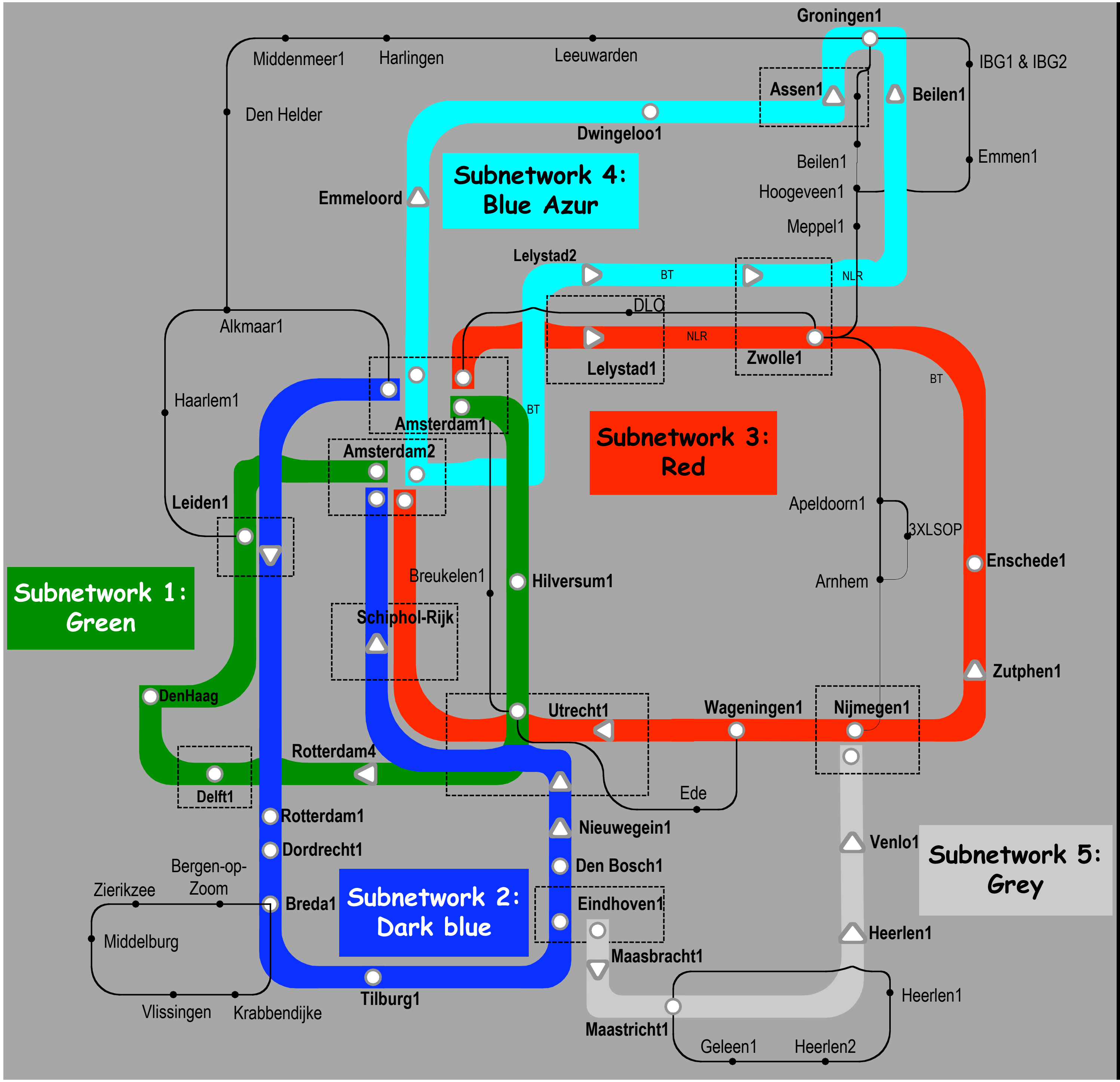
- universities;
- academic hospitals;
- most polytechnics;
- research centers.

with an indirect ~750K user base

Red crosses = StarPlane

~ 6000 km

scale comparable to railway system



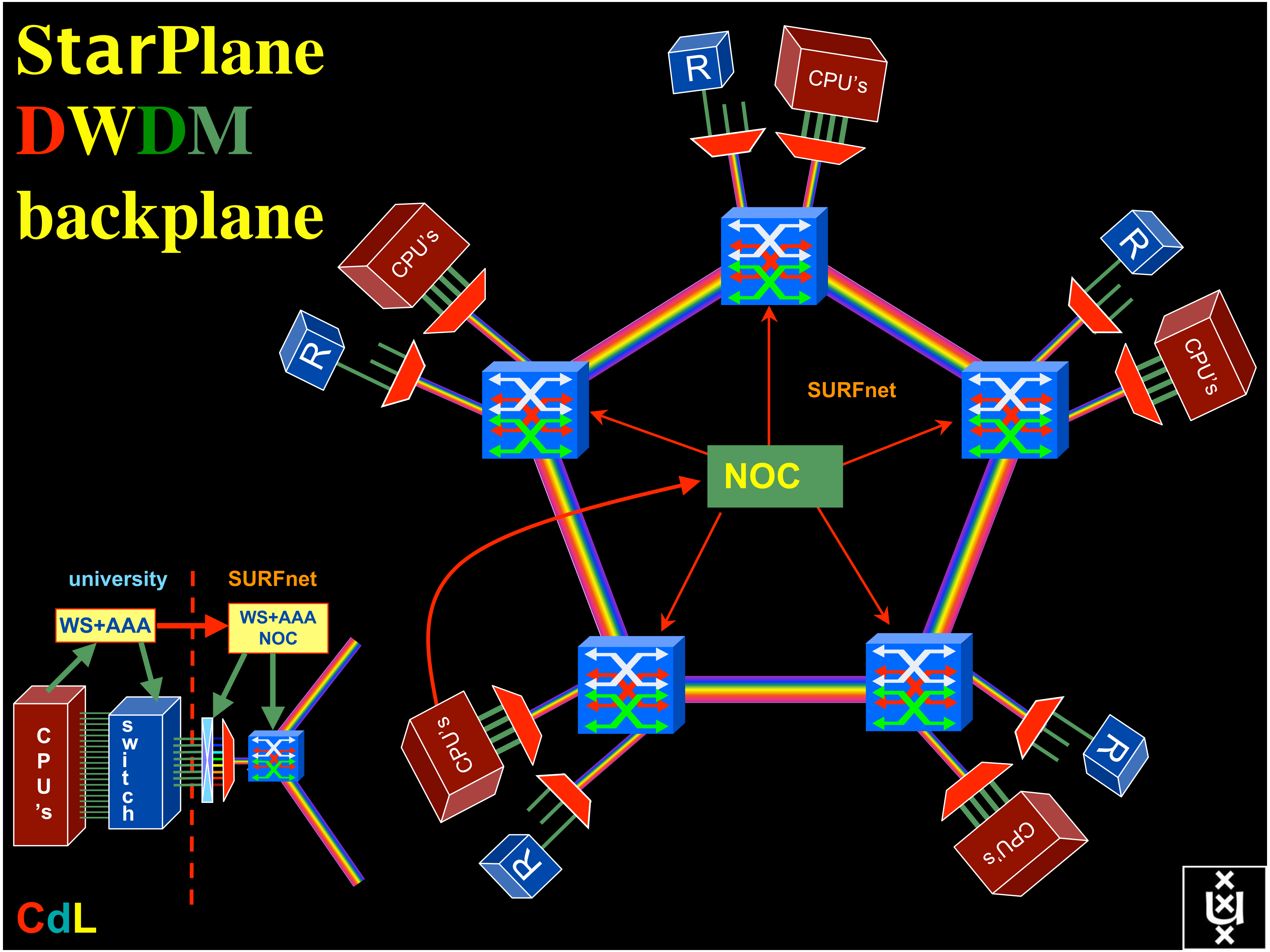
Common Photonic Layer (CPL) in SURFnet6

supports up to
72 Lambda's of
10 G each
40 G soon.

Contents

1. The need for hybrid networking
2. StarPlane; a grid controlled photonic network
3. RDF/Network Description Language
4. Tera-networking
5. Programmable networks

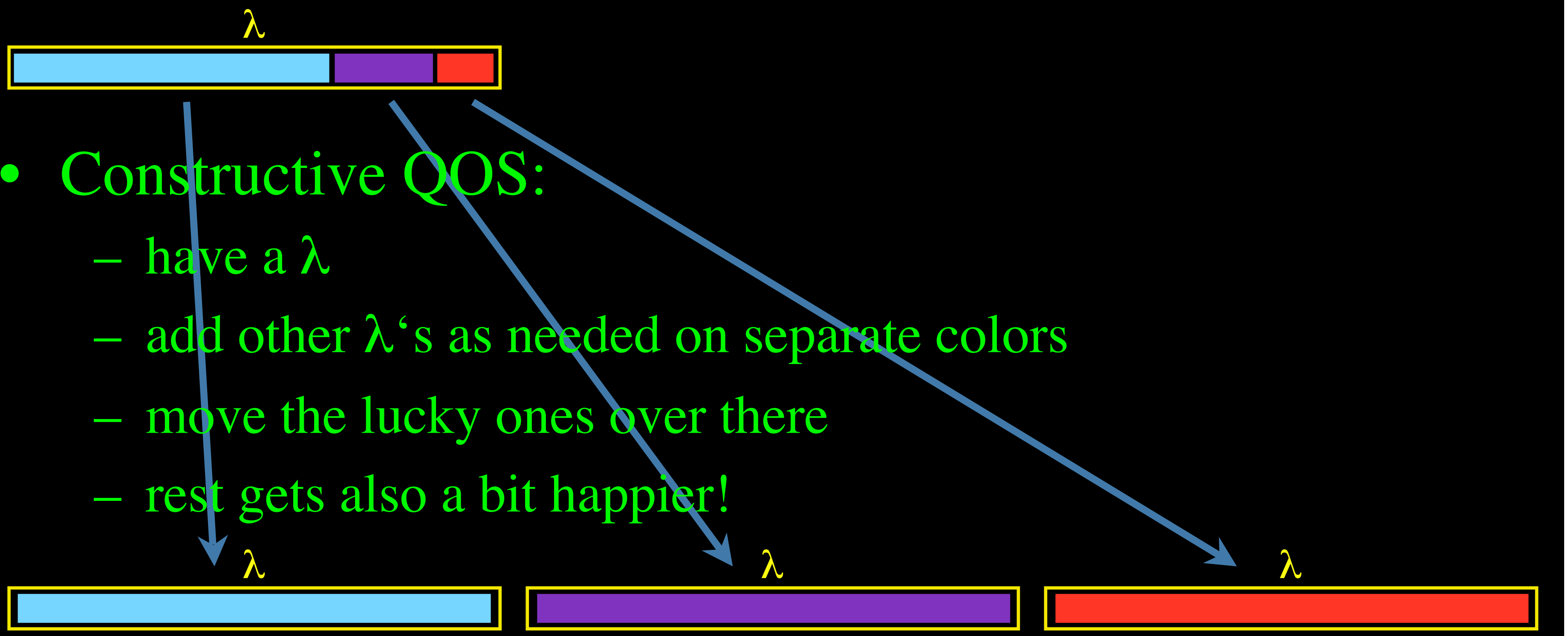
StarPlane DWDM backplane



CdL

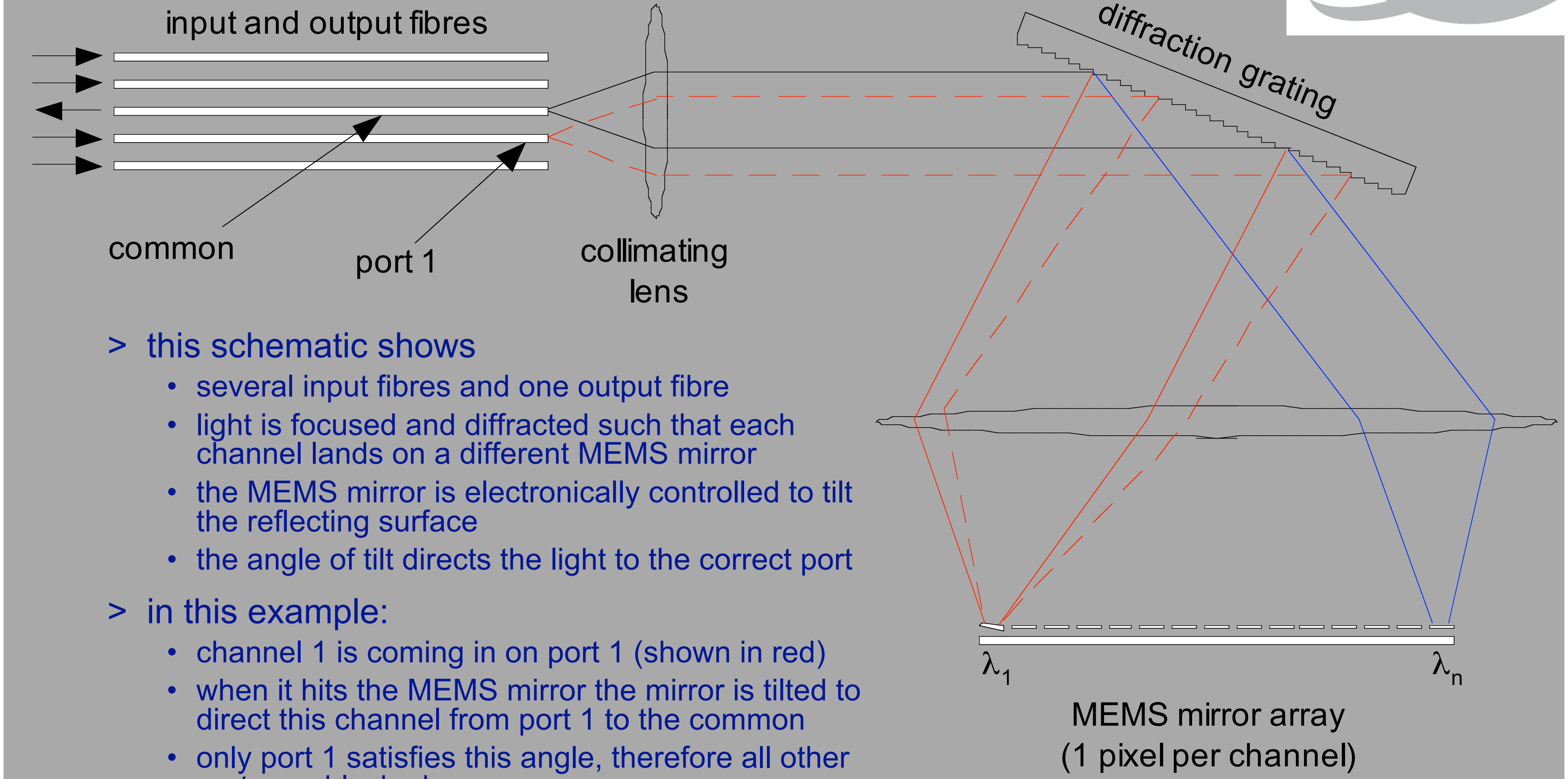
QOS in a non destructive way!

- 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

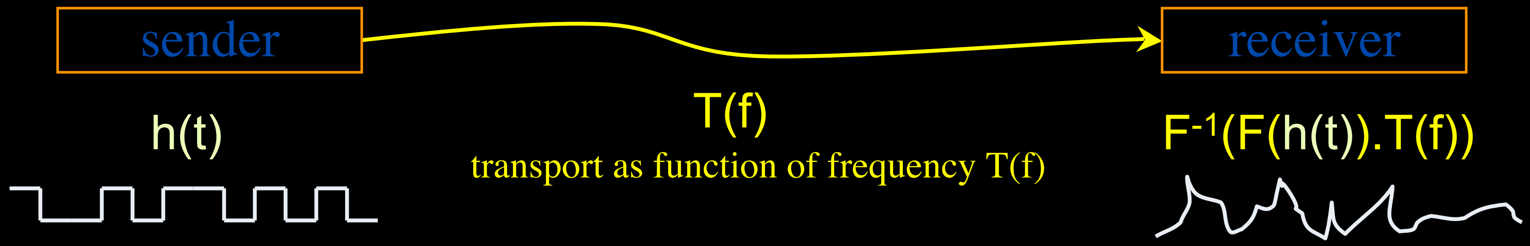
> in this example:

- 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

(Try to Google eDCO :-)



sender

receiver

$h(t)$

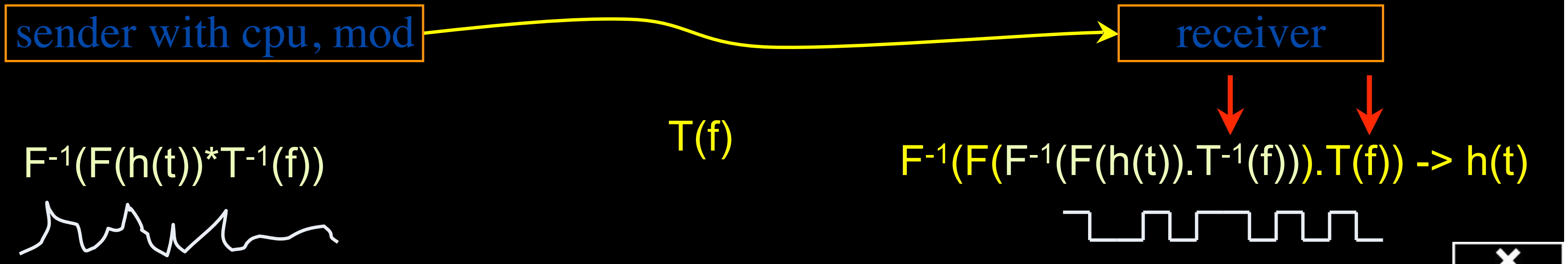
$T(f)$

transport as function of frequency T(f)

$F^{-1}(F(h(t)).T(f))$

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) = $F^{-1}(F(h(t)).T^{-1}(f))$



sender with cpu, mod

receiver

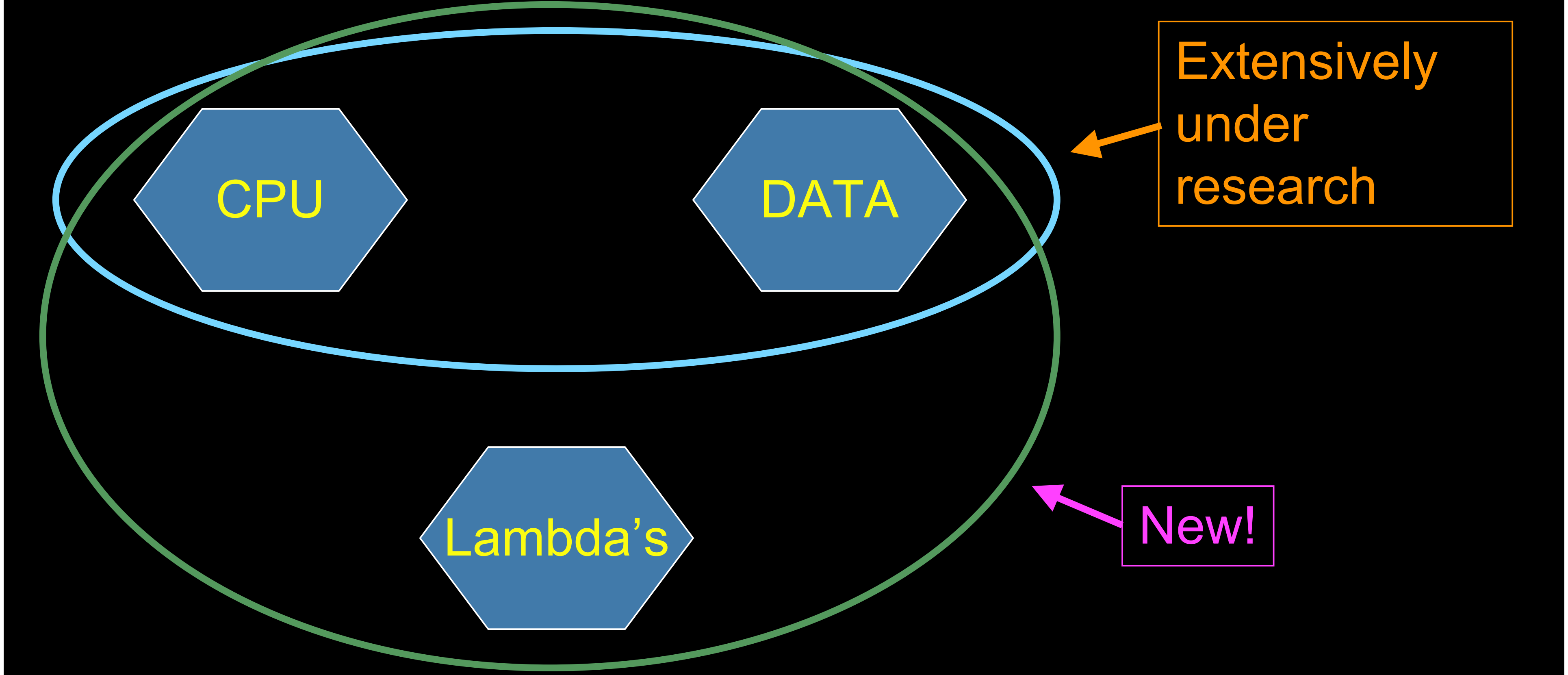
$T(f)$

$F^{-1}(F(h(t))*T^{-1}(f))$

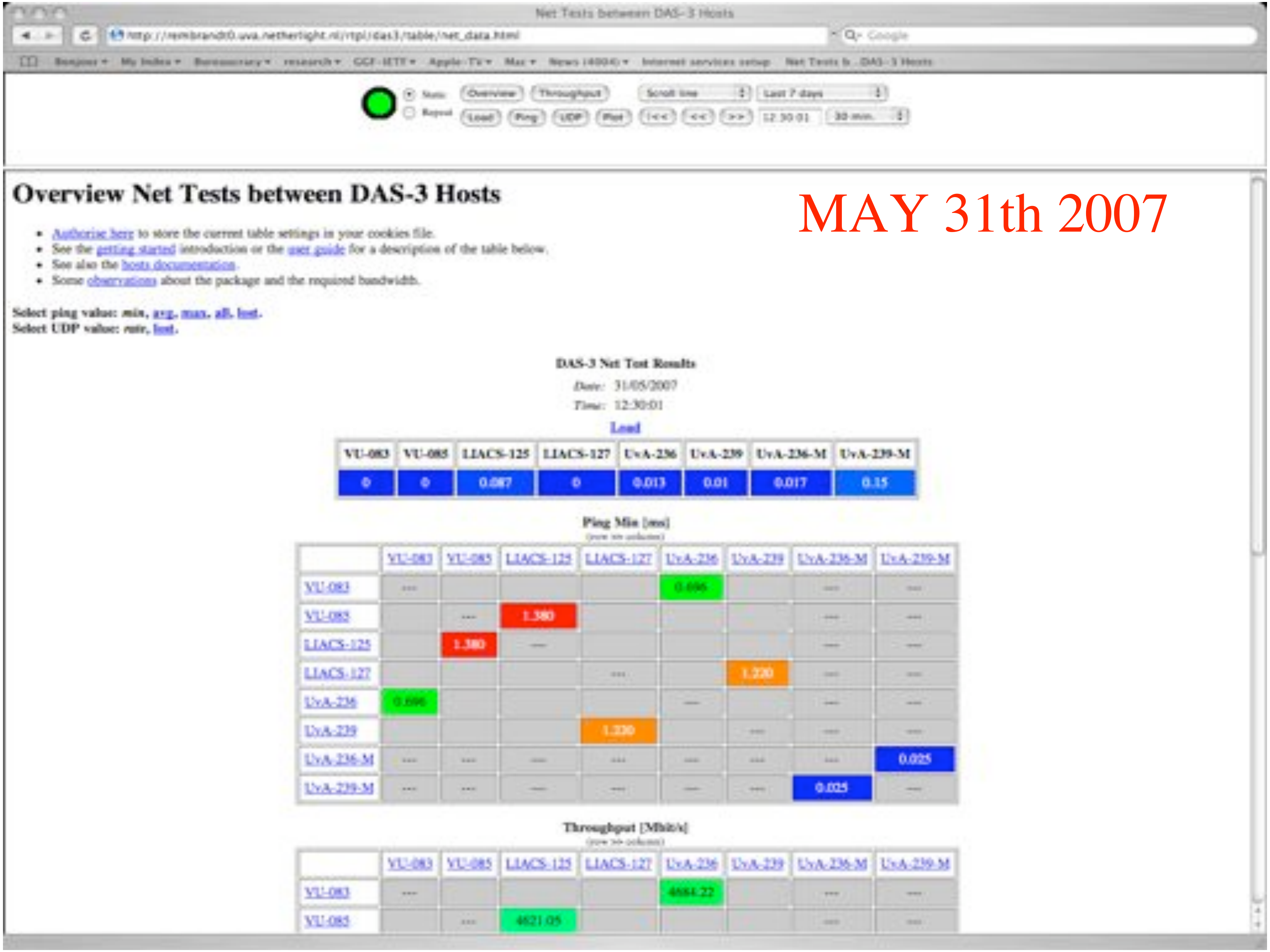
$F^{-1}(F(F^{-1}(F(h(t)).T^{-1}(f))).T(f))$ -> $h(t)$

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

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.



Overview Net Tests between DAS-3 Hosts

MAY 31th 2007

- 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

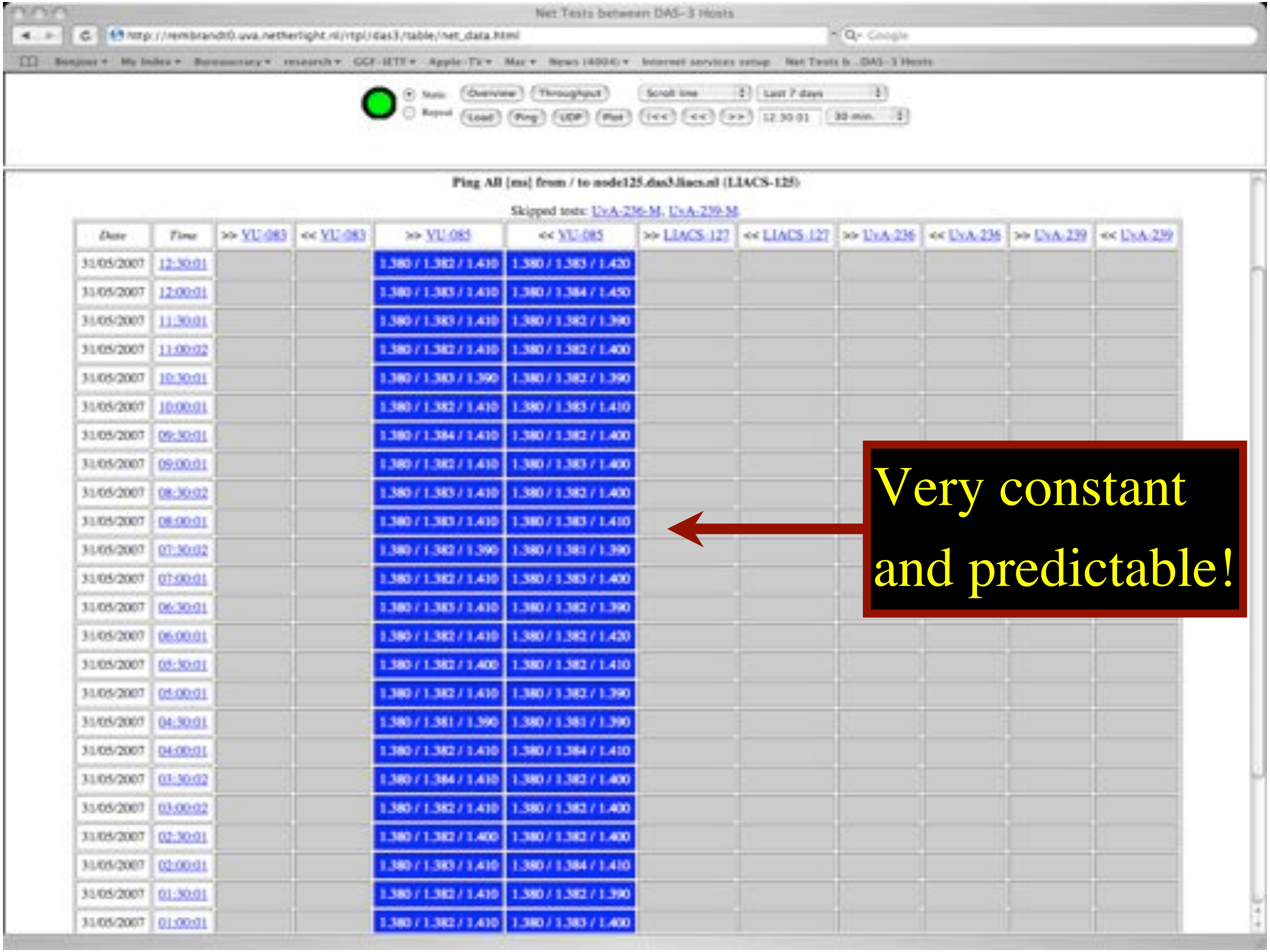
VU-083	VU-085	LIACS-125	LIACS-127	UvA-236	UvA-239	UvA-236-M	UvA-239-M
0	0	0.087	0	0.013	0.01	0.017	0.15

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				---	---



Net Tests between DAS-3 Hosts
http://rembrandt0.uva.netherlight.nl/rtpl/das3/table/net_data.html
Google
Bonjour ▾ My Index ▾ Bureaucracy ▾ research ▾ GGF-IETF ▾ Apple-TV ▾ Mac ▾ News (4004) ▾ Internet services setup Net Tests b...DAS-3 Hosts
● Static
○ Repeat
Overview
Throughput
Load
Ping
UDP
Plot
Scroll time
Last 7 days
|<<
<<
>>
12:30:01
30 min.

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

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

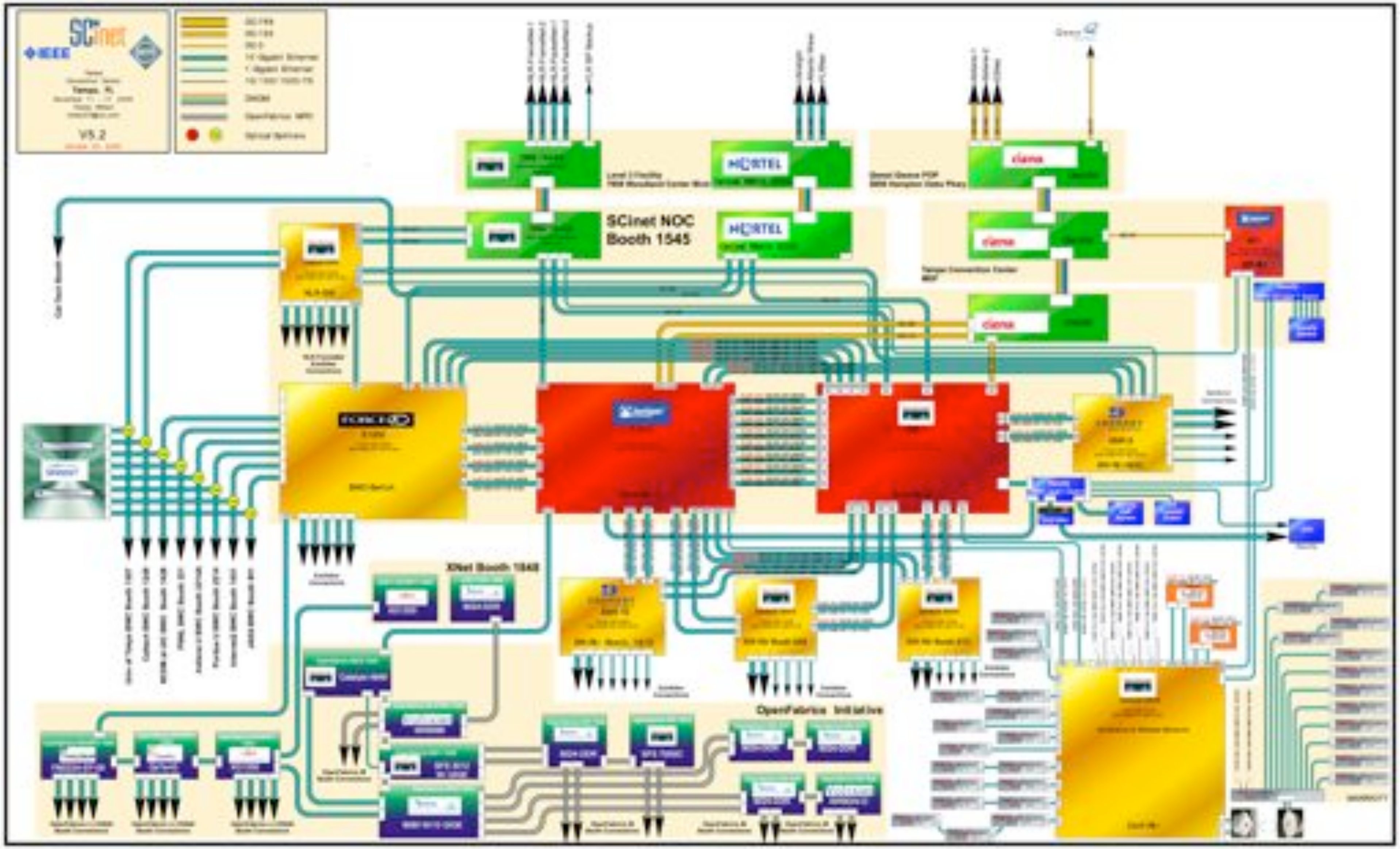
Date	Time	>> VU-083	<< VU-083	>> VU-085	<< VU-085	>> LIACS-127	<< LIACS-127	>> UvA-236	<< UvA-236	>> UvA-239	<< UvA-239
31/05/2007	12:30:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.420						
31/05/2007	12:00:01			1.380 / 1.383 / 1.410	1.380 / 1.384 / 1.450						
31/05/2007	11:30:01			1.380 / 1.383 /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

Contents

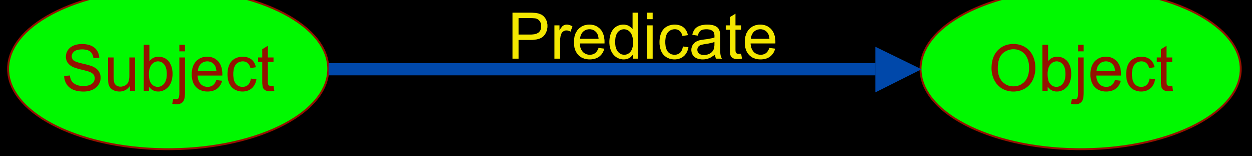
1. The need for hybrid networking
2. StarPlane; a grid controlled photonic network
3. RDF/Network Description Language
4. Tera-networking
5. Programmable networks

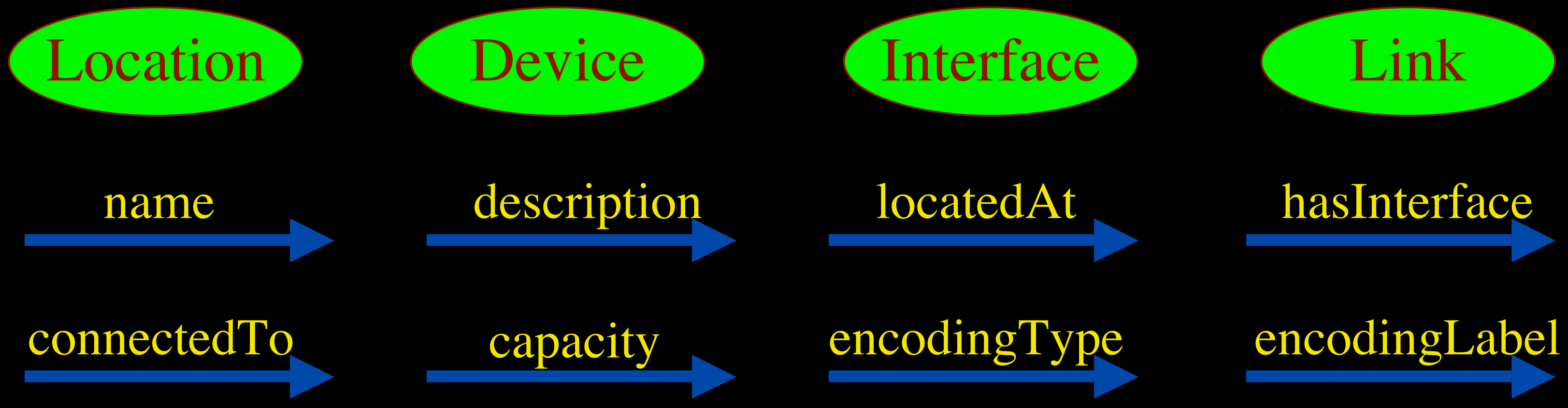
Architecture SC06



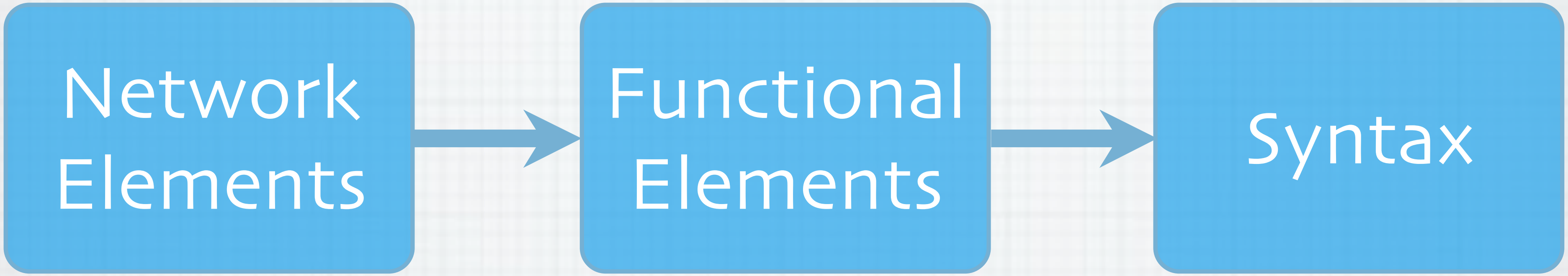
Network Description Language

- From semantic Web / Resource Description Framework.
- The RDF uses XML as an interchange syntax.
- Data is described by triplets:





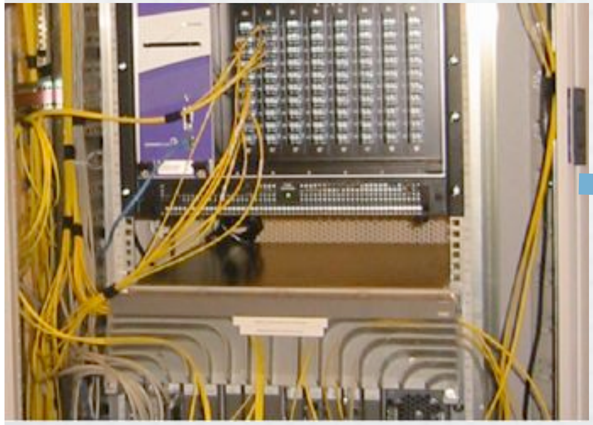
The Modelling Process

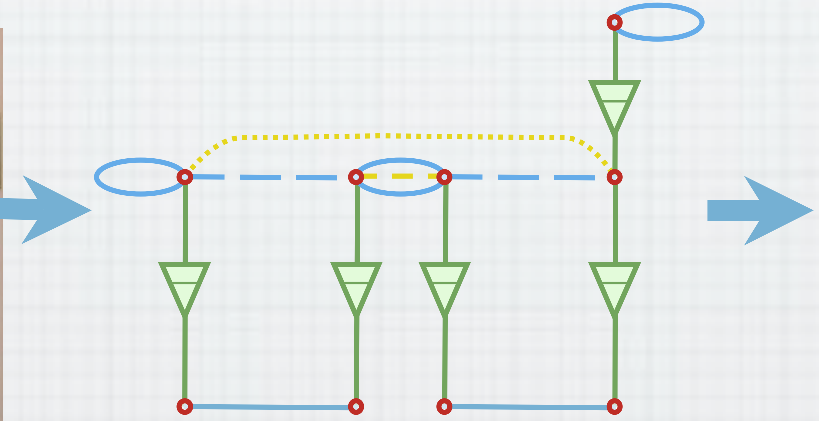


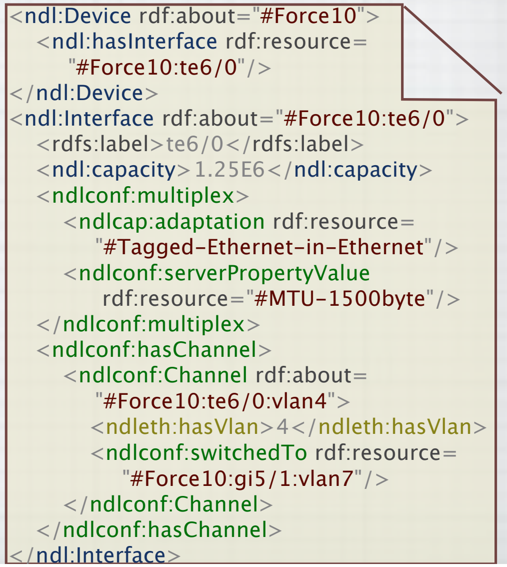
Network Elements

Functional Elements

Syntax







```
<ndl:Device rdf:about="#Force10">
   <ndl:hasInterface rdf:resource=
       "#Force10:te6/0"/>
</ndl:Device>
<ndl:Interface rdf:about="#Force10:te6/0">
   <rdfs:label>te6/0</rdfs:label>
   <ndl:capacity>1.25E6</ndl:capacity>
   <ndlconf:multiplex>
      <ndlcap:adaptation rdf:resource=
          "#Tagged-Ethernet-in-Ethernet"/>
      <ndlconf:serverPropertyValue
          rdf:resource="#MTU-1500byte"/>
   </ndlconf:multiplex>
   <ndlconf:hasChannel>
      <ndlconf:Channel rdf:about=
          "#Force10:te6/0:vlan4">
         <ndleth:hasVlan>4</ndleth:hasVlan>
         <ndlconf:switchedTo rdf:resource=
             "#Force10:gi5/1:vlan7"/>
      </ndlconf:Channel>
   </ndlconf:hasChannel>
</ndl:Interface>
```

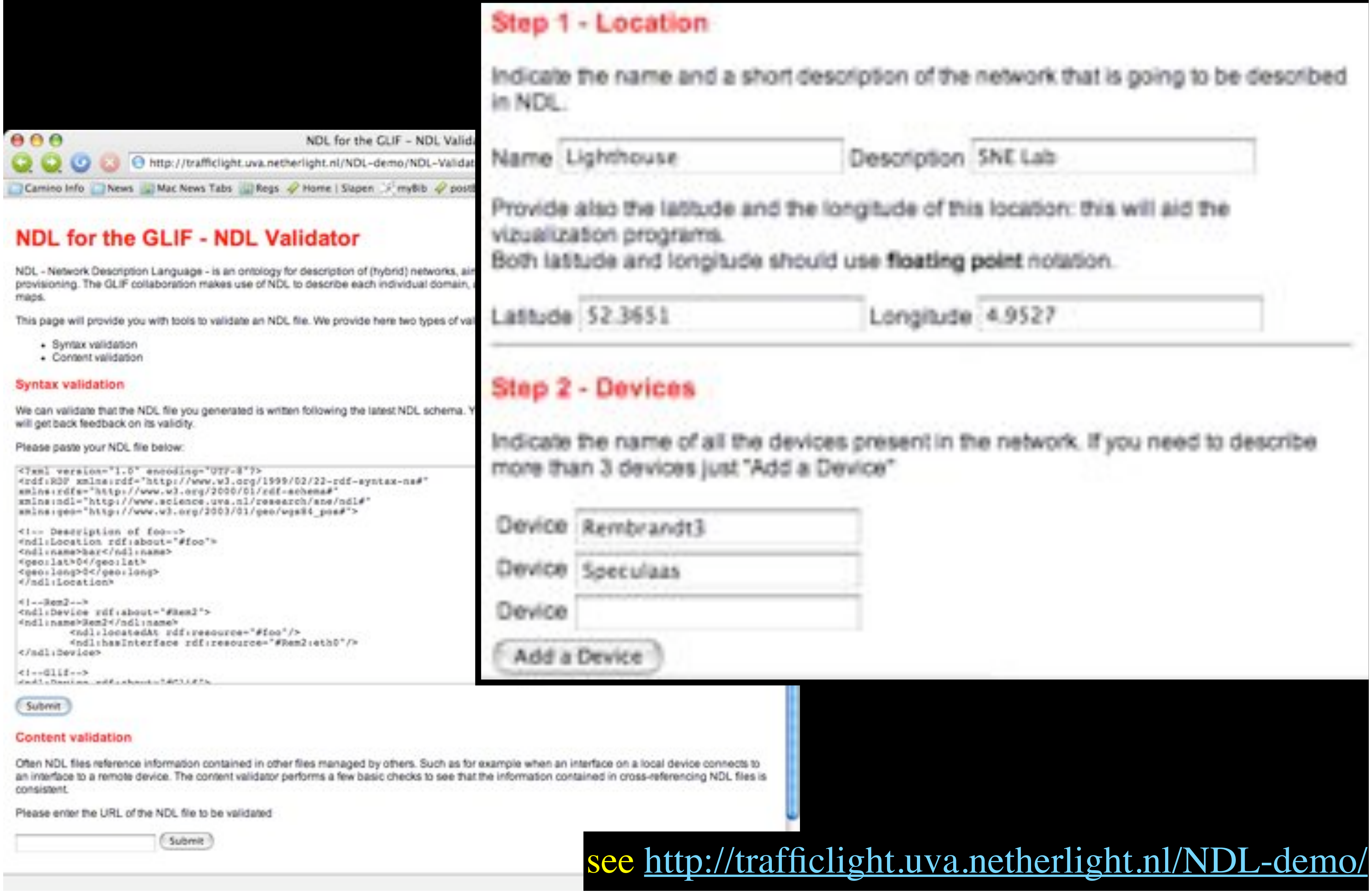
NetherLight in RDF

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:ndl="http://www.science.uva.nl/research/air/ndl#">
<!-- Description of Netherlight -->
<ndl:Location rdf:about="#Netherlight">
      <ndl:name>Netherlight Optical Exchange</ndl:name>
</ndl:Location>
<!-- TDM3.amsterdam1.netherlight.net -->
<ndl:Device rdf:about="#tdm3.amsterdam1.netherlight.net">
      <ndl:name>tdm3.amsterdam1.netherlight.net</ndl:name>
      <ndl:locatedAt rdf:resource="#amsterdam1.netherlight.net"/>
      <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:501/1"/>
      <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:501/3"/>
      <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:501/4"/>
      <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:503/1"/>
      <ndl:hasInterface rdf:resour
      <ndl:hasInterface rdf:resour
      <ndl:hasInterface rdf:resour
      <ndl:hasInterface rdf:resour
      <ndl:hasInterface rdf:resour
      <ndl:hasInterface rdf:resour
      <ndl:hasInterface rdf:resour
      <ndl:hasInterface rdf:resour
```

```
<!-- all the interfaces of TDM3.amsterdam1.netherlight.net -->

<ndl:Interface rdf:about="#tdm3.amsterdam1.netherlight.net:501/1">
            <ndl:name>tdm3.amsterdam1.netherlight.net:POS501/1</ndl:name>
            <ndl:connectedTo rdf:resource="#tdm4.amsterdam1.netherlight.net:5/1"/>
 </ndl:Interface>
<ndl:Interface rdf:about="#tdm3.amsterdam1.netherlight.net:501/2">
            <ndl:name>tdm3.amsterdam1.netherlight.net:POS501/2</ndl:name>
            <ndl:connectedTo rdf:resource="#tdm1.amsterdam1.netherlight.net:12/1"/>
 </ndl:Interface>
```

NDL Generator and Validator



NDL for the GLIF – NDL Valida
http://trafficlight.uva.netherlight.nl/NDL-demo/NDL-Validat
Camino Info
News
Mac News Tabs
Regs
Home | Slapen
myBib
post

Step 1 - Location

Indicate the name and a short description of the network that is going to be described in NDL.

Name	Lighthouse	Description	SNE Lab

Provide also the latitude and the longitude of this location: this will aid the vizualization programs.
Both latitude and longitude should use **floating point** notation.

Latitude	52.3651	Longitude	4.9527

Step 2 - Devices

Indicate the name of all the devices present in the network. If you need to describe more than 3 devices just "Add a Device"

Device	Rembrandt3
Device	Speculaas
Device	

NDL for the GLIF - NDL Validator

NDL - Network Description Language - is an ontology for description of (hybrid) networks, ai
provisioning. The GLIF collaboration makes use of NDL to describe each individual domain,
maps.

This page will provide you with tools to validate an NDL file. We provide here two types of va

- Syntax validation
- Content validation

Syntax validation

We can validate that the NDL file you generated is written following the latest NDL schema. Y
will get back feedback on its validity.

Please paste your NDL file below:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:ndl="http://www.science.uva.nl/research/sne/ndl#"
xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#">

<!-- Description of foo-->
<ndl:Location rdf:about="#foo">
<ndl:name>bar</ndl:name>
<geo:lat>0</geo:lat>
<geo:long>0</geo:long>
</ndl:Location>

<!--Rem2-->
<ndl:Device rdf:about="#Rem2">
<ndl:name>Rem2</ndl:name>
        <ndl:locatedAt rdf:resource="#foo"/>
        <ndl:hasInterface rdf:resource="#Rem2:eth0"/>
</ndl:Device>

<!--Glif-->
<ndl:Device rdf:about="#Glif">
```

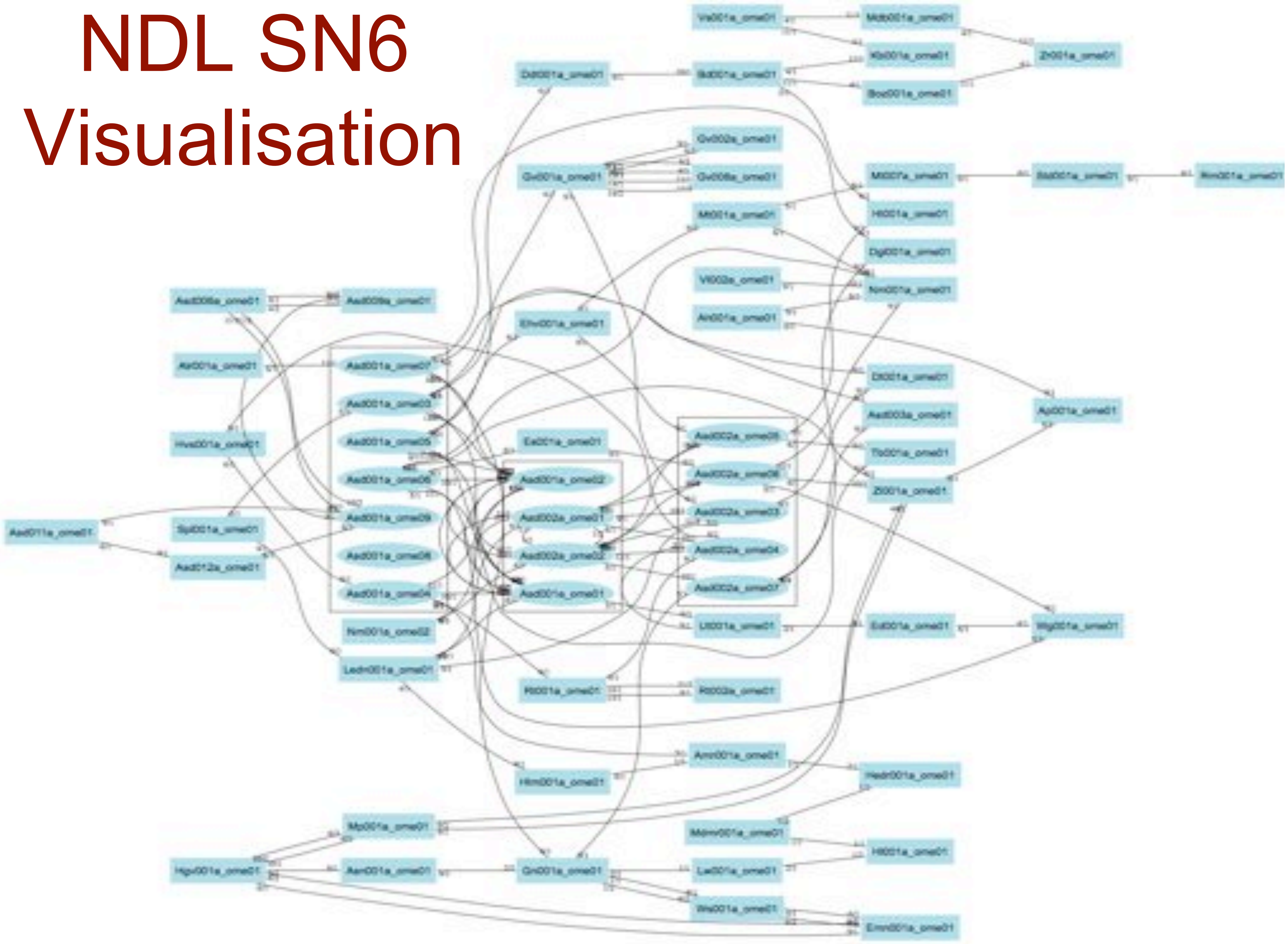
Content validation

Often NDL files reference information contained in other files managed by others. Such as for example when an interface on a local device connects to an interface to a remote device. The content validator performs a few basic checks to see that the information contained in cross-referencing NDL files is consistent.

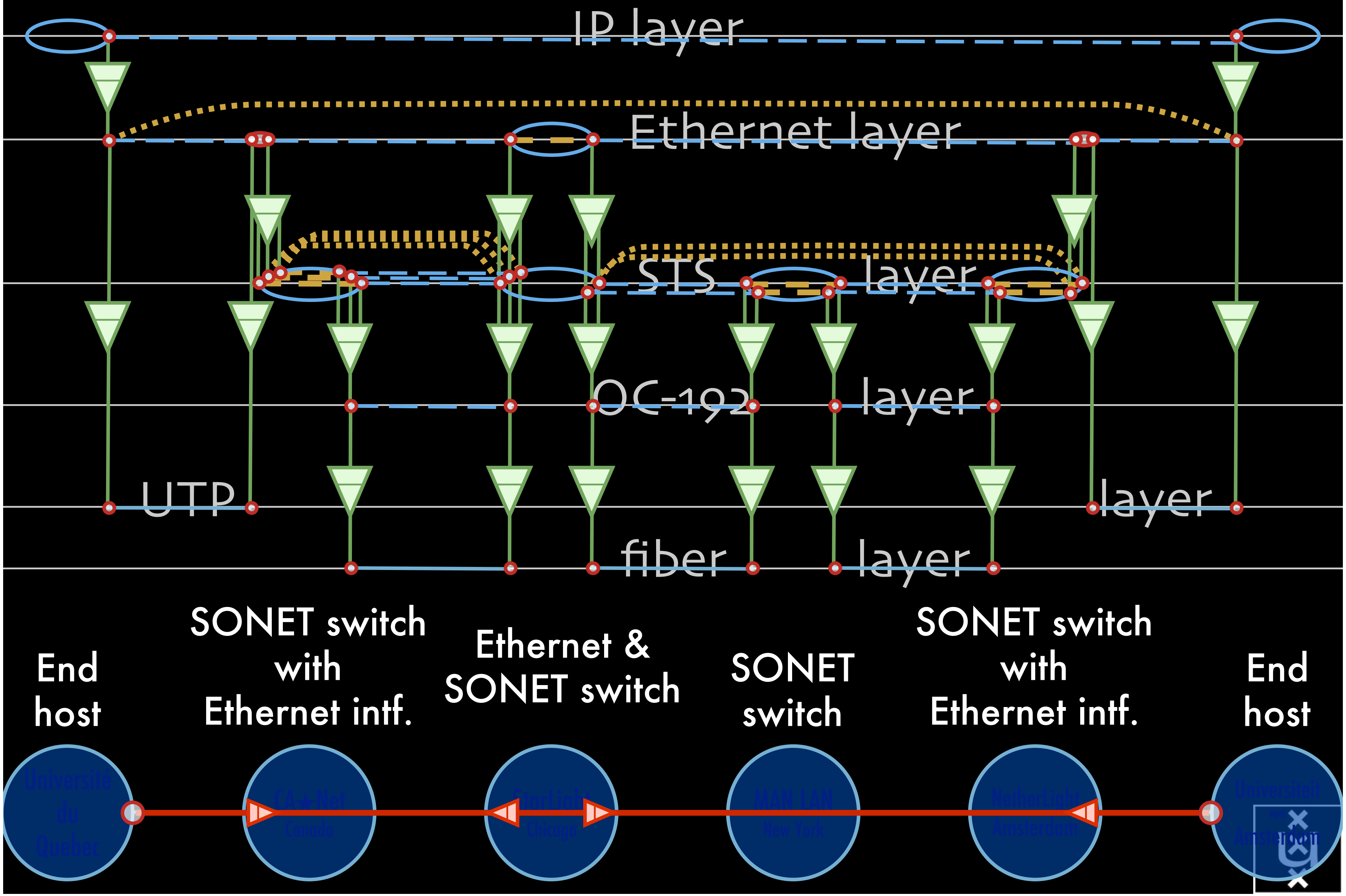
Please enter the URL of the NDL file to be validated

see http://trafficlight.uva.netherlight.nl/NDL-demo/

NDL SN6 Visualisation



Multi-layer extensions to NDL



IP layer

Ethernet layer

STS layer

OC-192 layer

UTP layer

fiber layer

End host

SONET switch with Ethernet intf.

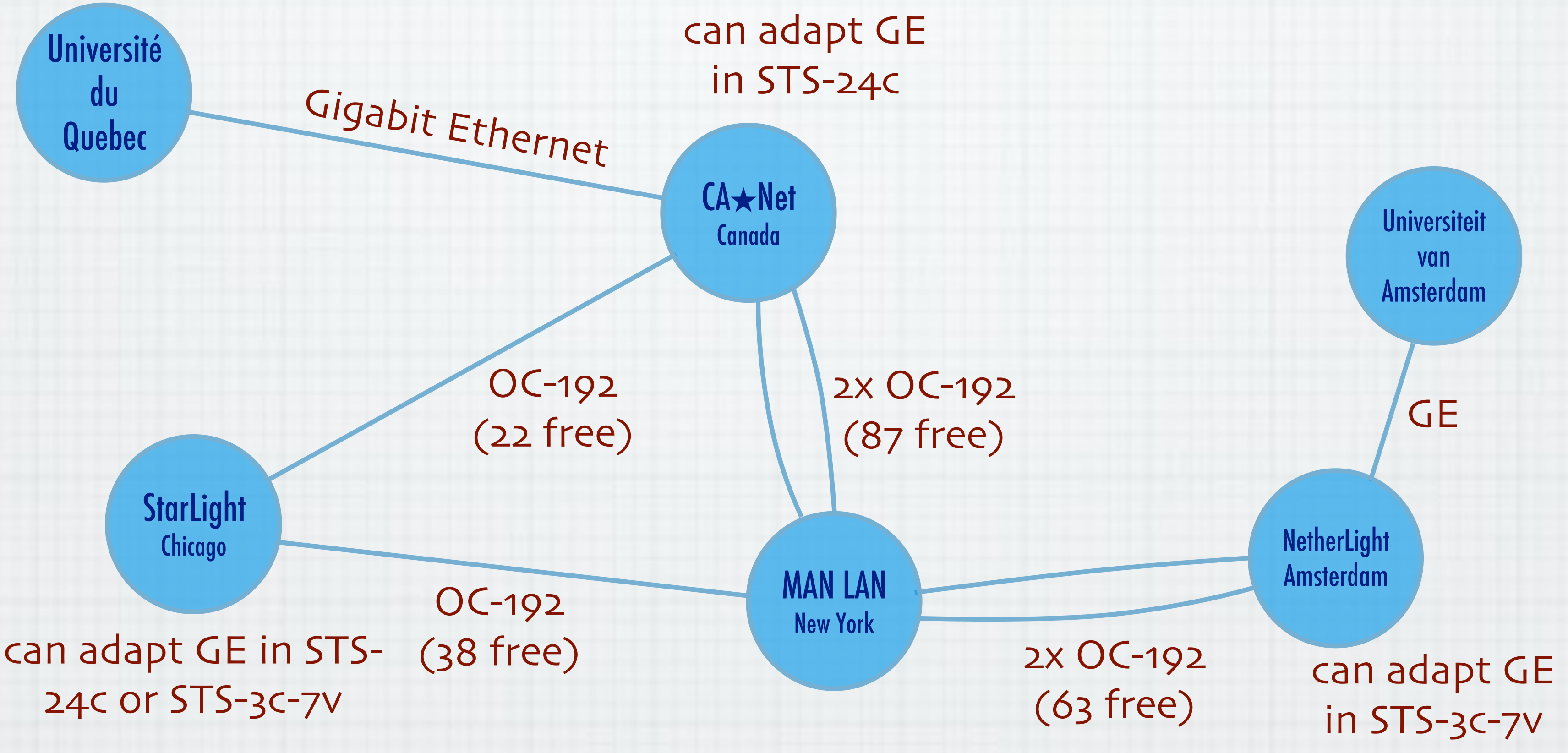
Ethernet & SONET switch

SONET switch

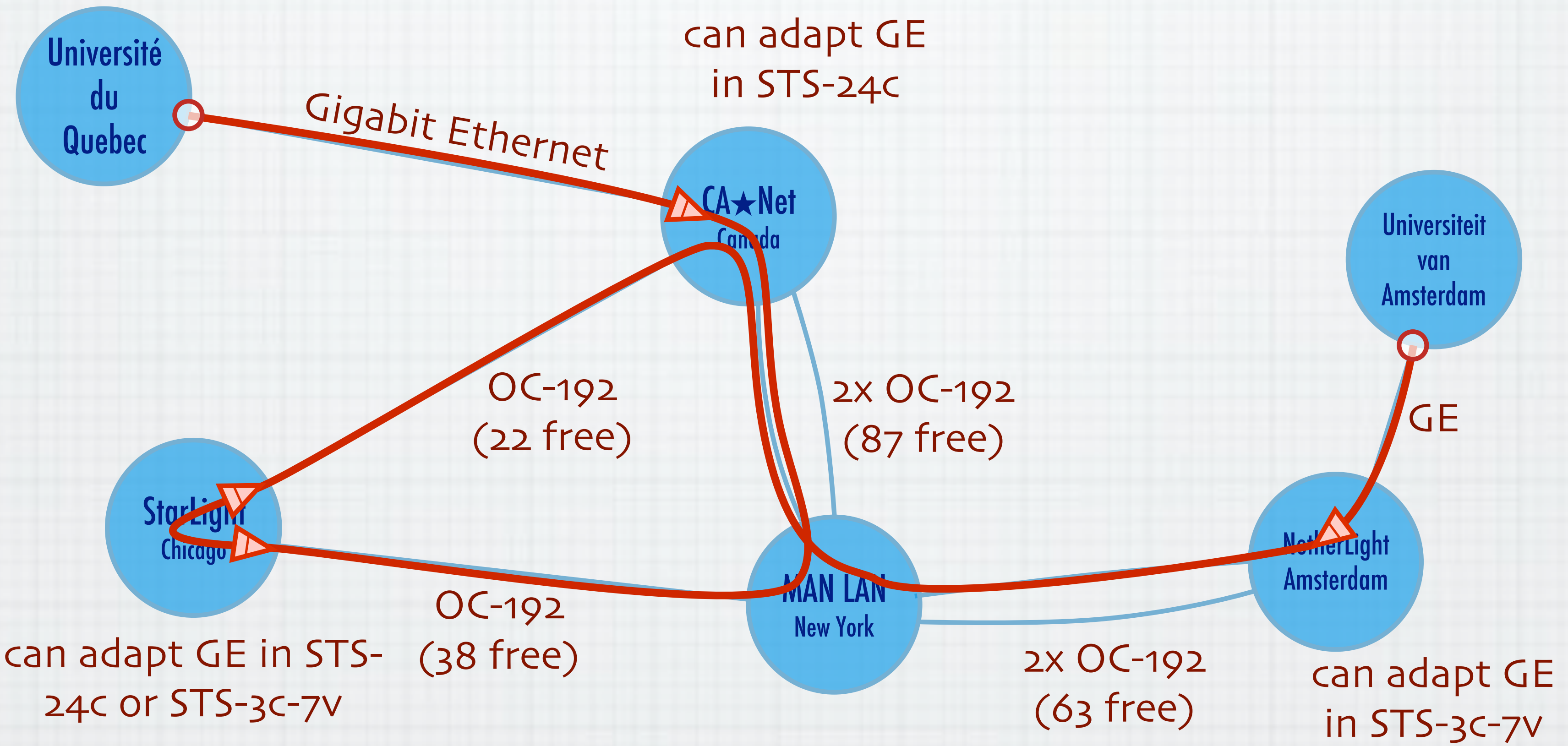
SONET switch with Ethernet intf.

End host

A weird example

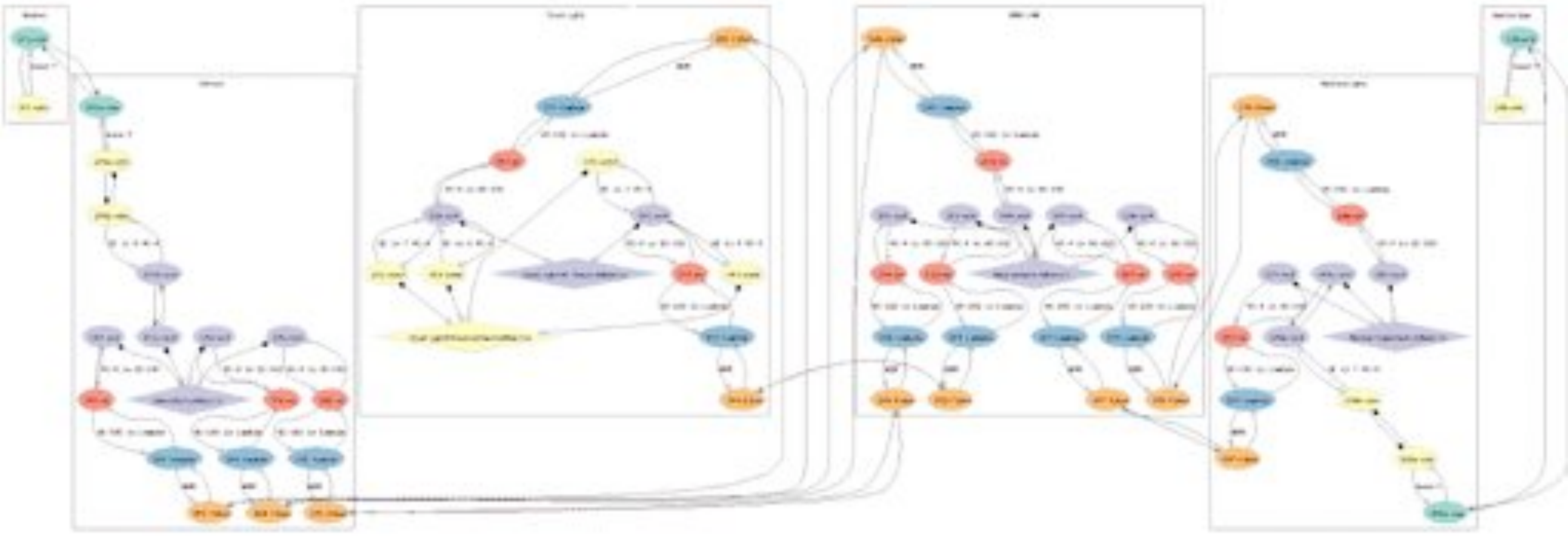


The result :-)

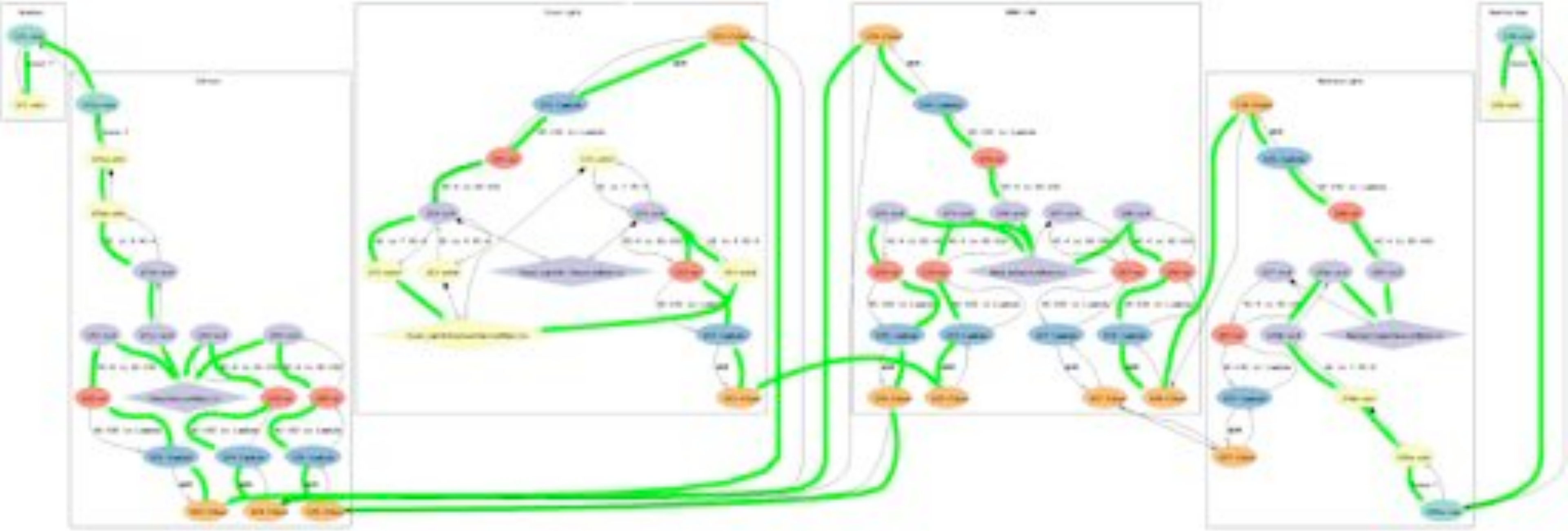


Thanks to Freek Dijkstra & team

MultiDomain MultiLayer pathfinding in action



MultiDomain MultiLayer pathfinding in action



OGF NML-WG

Open Grid Forum - Network Markup Language workgroup

Chairs:

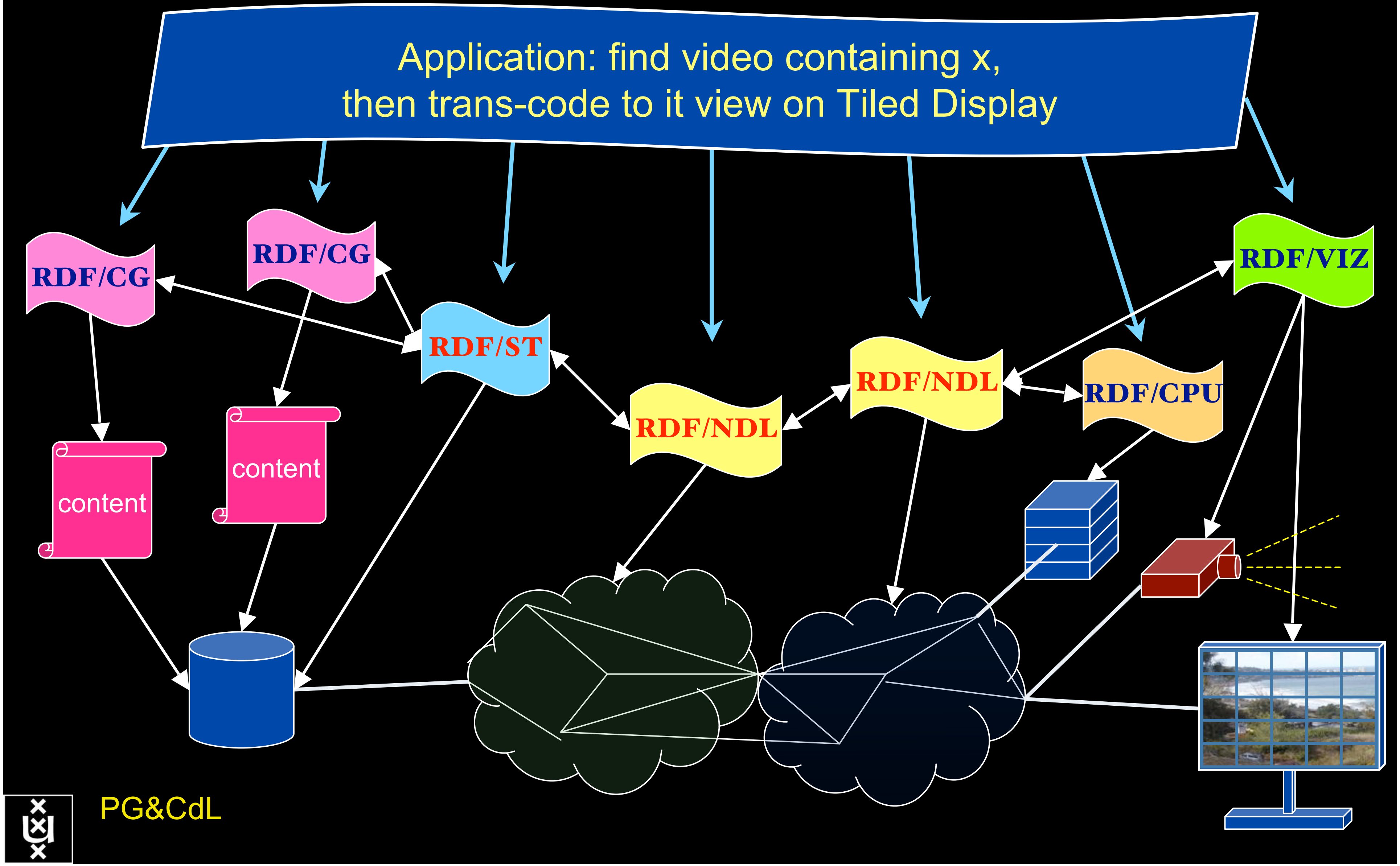
- Paola Grosso – Universiteit van Amsterdam
- Martin Swany – University of Delaware

Purpose:

To describe network topologies, so that the outcome is a standardized network description ontology and schema, facilitating interoperability between different projects.

https://forge.gridforum.org/sf/projects/nml-wg

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

Contents

1. The need for hybrid networking
2. StarPlane; a grid controlled photonic network
3. RDF/Network Description Language
4. Tera-networking
5. Programmable networks

TeraThinking

- 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 -> ?

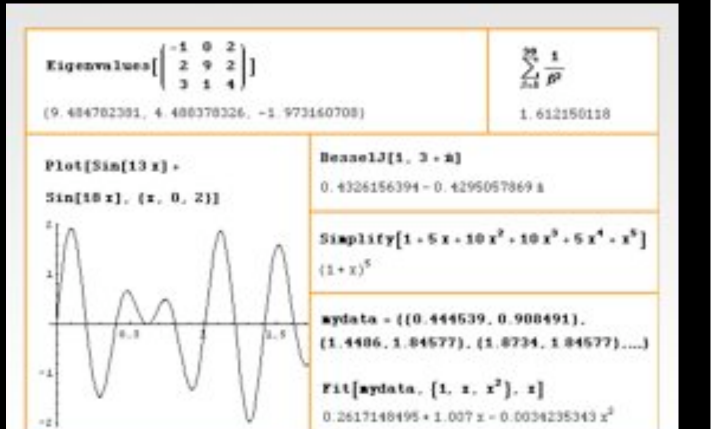
ref Larry Smarr & CdL

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



```
Eigenvalues[( -1 0 2 )
            (  2 9 2 )
            (  3 1 4 )]
{9.484782381, 4.488378326, -1.973160708}
```

$$\sum_{j=1}^{\infty} \frac{1}{j^3}$$

1.612150118

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

```
BesselJ[1, 3 + i]
0.4326156394 - 0.4295057869 i
```

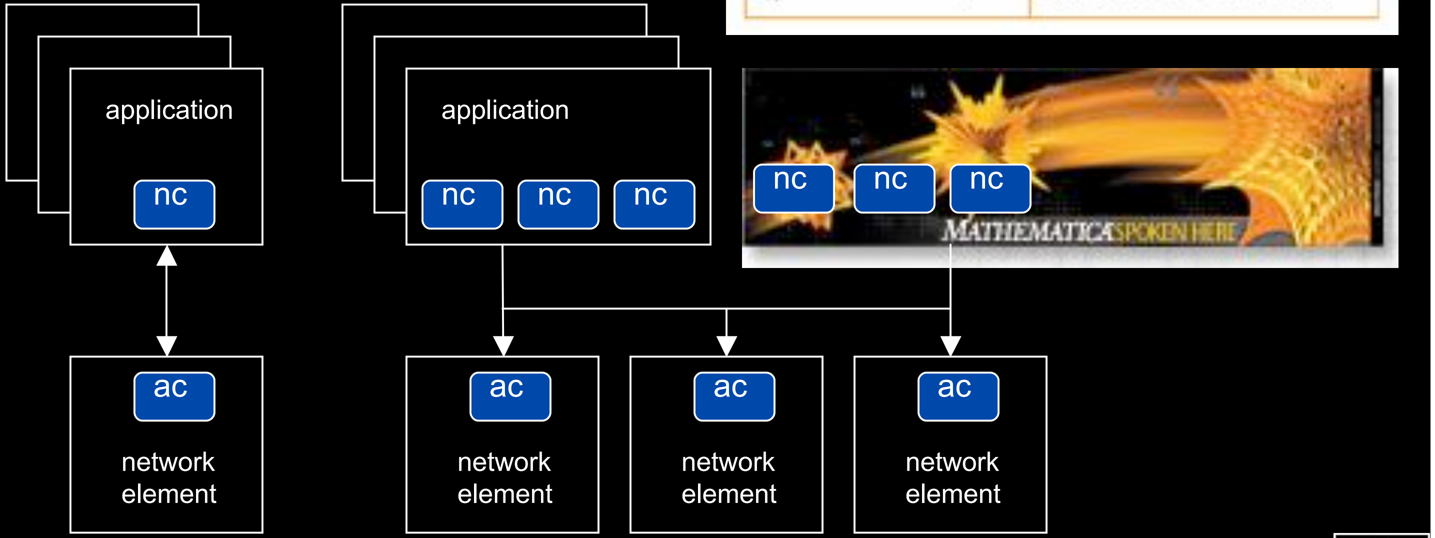
$$\text{Simplify}[1 + 5x + 10x^2 + 10x^3 + 5x^4 + x^5]$$

$$(1 + x)^5$$

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

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

$$0.2617148495 + 1.007 x - 0.0034235343 x^2$$

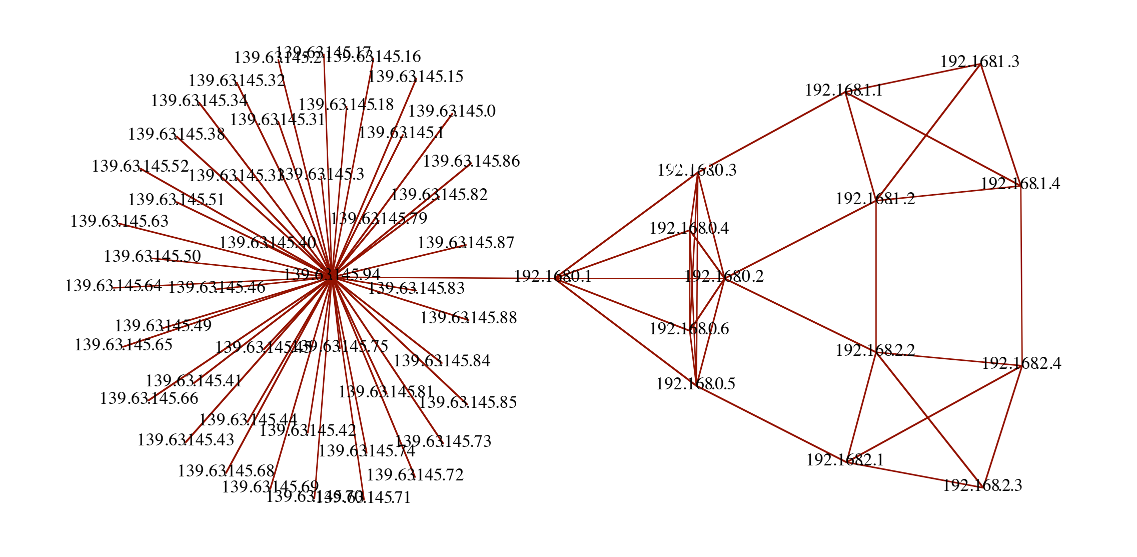


StarPlane

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



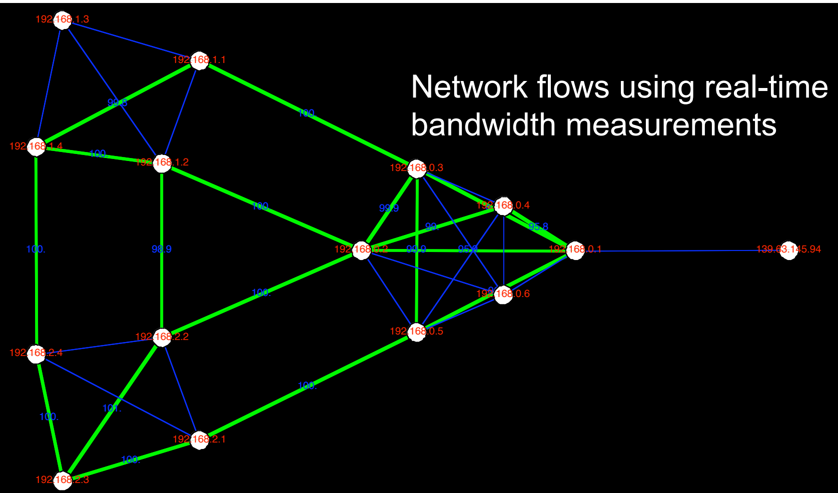
Initialization and BFS discovery of NEs

```
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}
```



Network flows using real-time bandwidth measurements

Transaction on shortest path with tokens

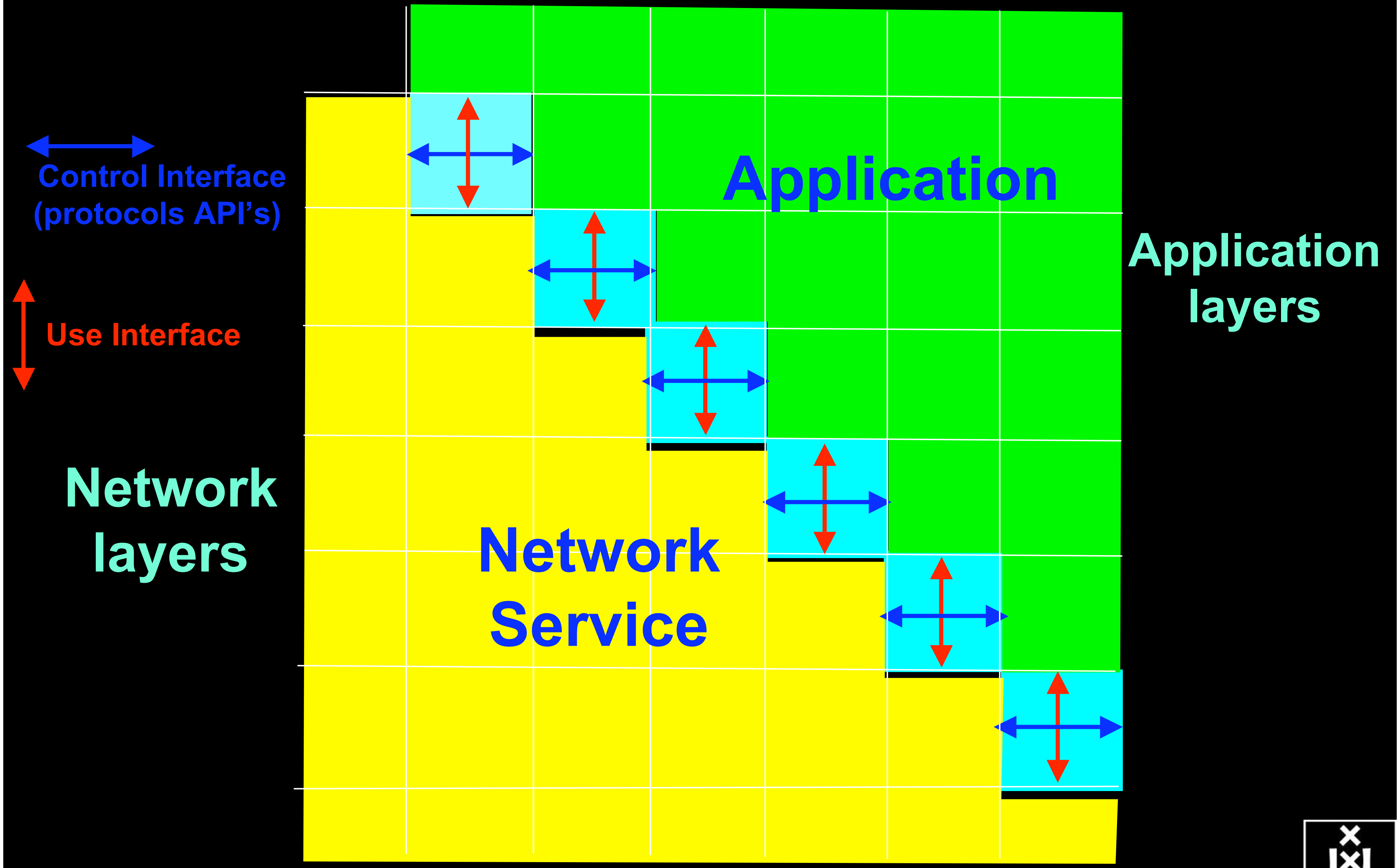
```
nodePath = ConvertIndicesToNodes[
                    ShortestPath[  g,
                                   Node2Index[nids,"192.168.3.4"],
                                   Node2Index[nids,"139.63.77.49"]],
                                   nids];
Print["Path: ", nodePath];
If[NetworkTokenTransaction[nodePath, "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 Virtualiized Networks, accepted for publication to the IEEE e-Science 2006 conference Amsterdam.

Functional building blocks



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!

Questions ?

I did not talk about StarPlane

AAA

…