# Lambda-Grid developments

## History - Present - Future

Cees de Laat

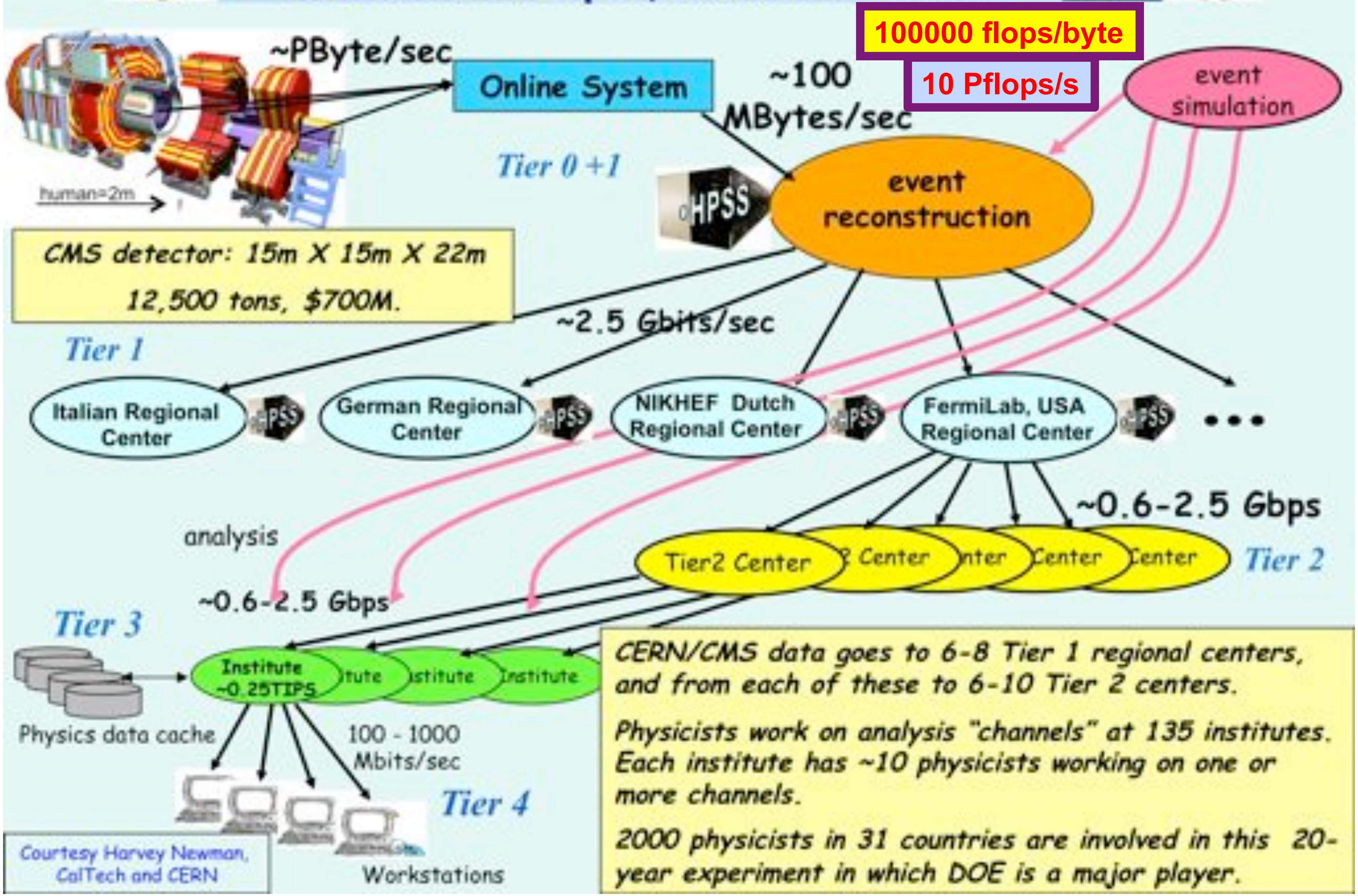
University of Amsterdam

# Contents

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

# LHC Data Grid Hierarchy

## CMS as example, Atlas is similar



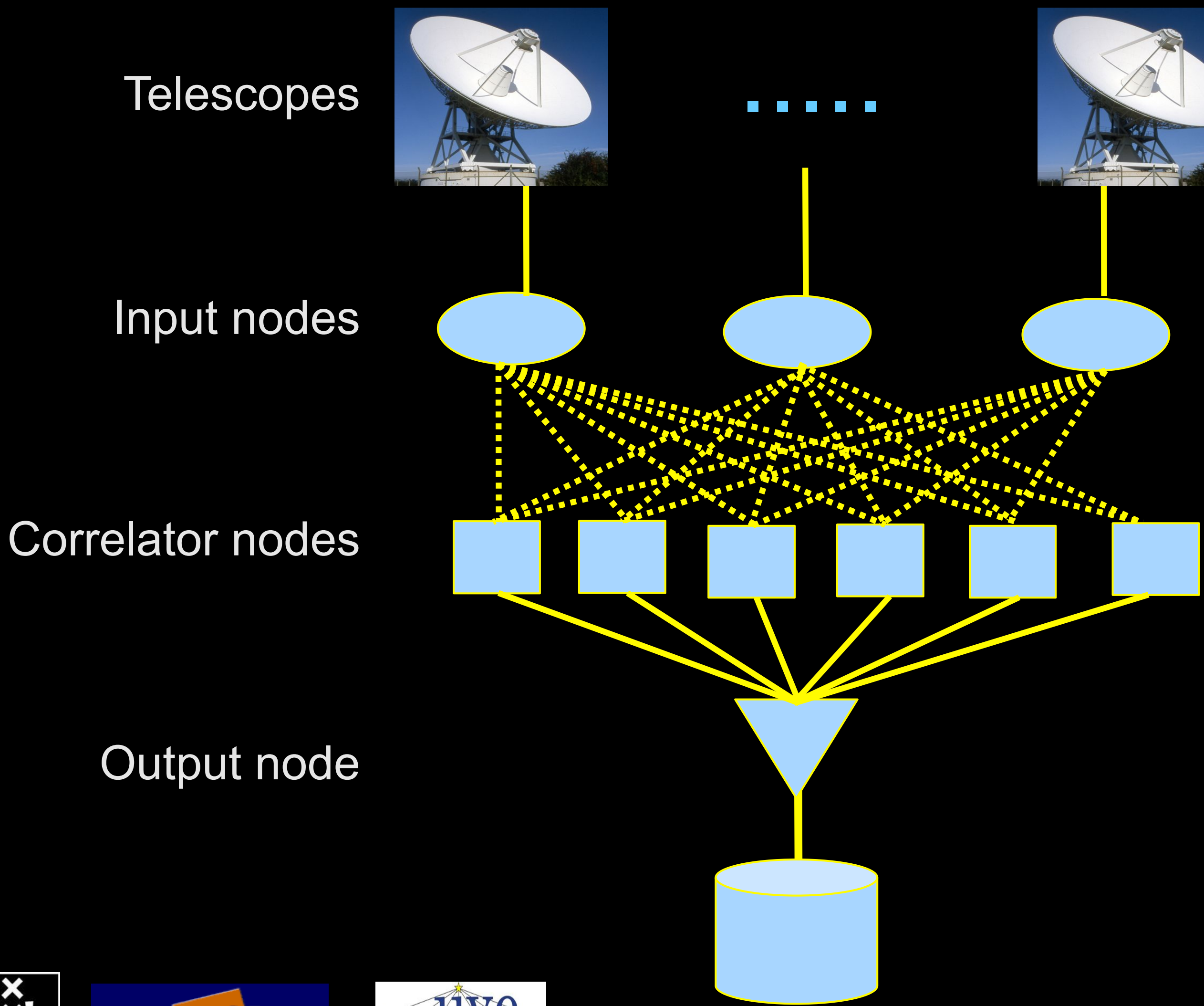

*CERN/CMS data goes to 6-8 Tier 1 regional centers, and from each of these to 6-10 Tier 2 centers.*

*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.*

# 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**

**0.1 Pflops/s**

THIS IS A DATA FLOW PROBLEM !!!

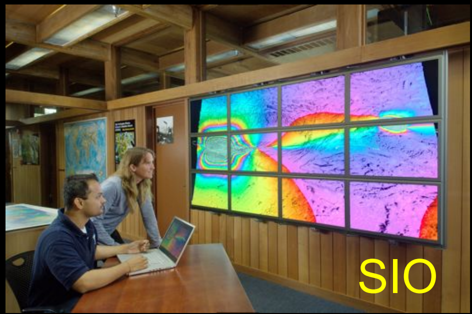
# LOFAR as a Sensor Network

20 flops/byte

2 Tflops/s

- LOFAR is a large distributed research infrastructure:
  - Astronomy:
    - >100 phased array stations
    - Combined in aperture synthesis array
    - 13,000 small “LF” antennas
    - 13,000 small “HF” tiles
  - Geophysics:
    - 18 vibration sensors per station
    - Infrasound detector per station
  - >20 Tbit/s generated digitally
  - >40 Tflop/s supercomputer
  - innovative software systems
    - new calibration approaches
    - full distributed control
    - VO and Grid integration
    - datamining and visualisation

# US and International OptIPortal Sites

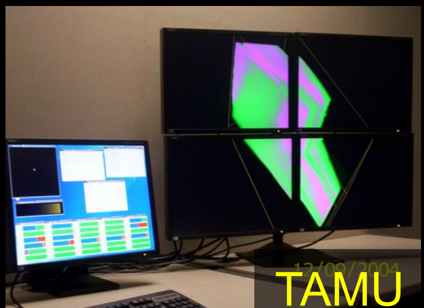
















**Real time, multiple 10 Gb/s**

# The "Dead Cat" demo

1 Mflops/byte

Real time issue

SC2004,
Pittsburgh,
Nov. 6 to 12, 2004
iGrid2005,
San Diego,
sept. 2005

Many thanks to:
AMC
SARA
GigaPort
UvA/AIR
Silicon Graphics, Inc.
Zoölogisch Museum

M. Scarpa, R.G. Belleman, P.M.A. Sloot and C.T.A.M. de Laat, "Highly Interactive Distributed Visualization", iGrid2005 special issue, Future Generation Computer Systems, volume 22 issue 8, pp. 896-900 (2006).

# IJKDIJK

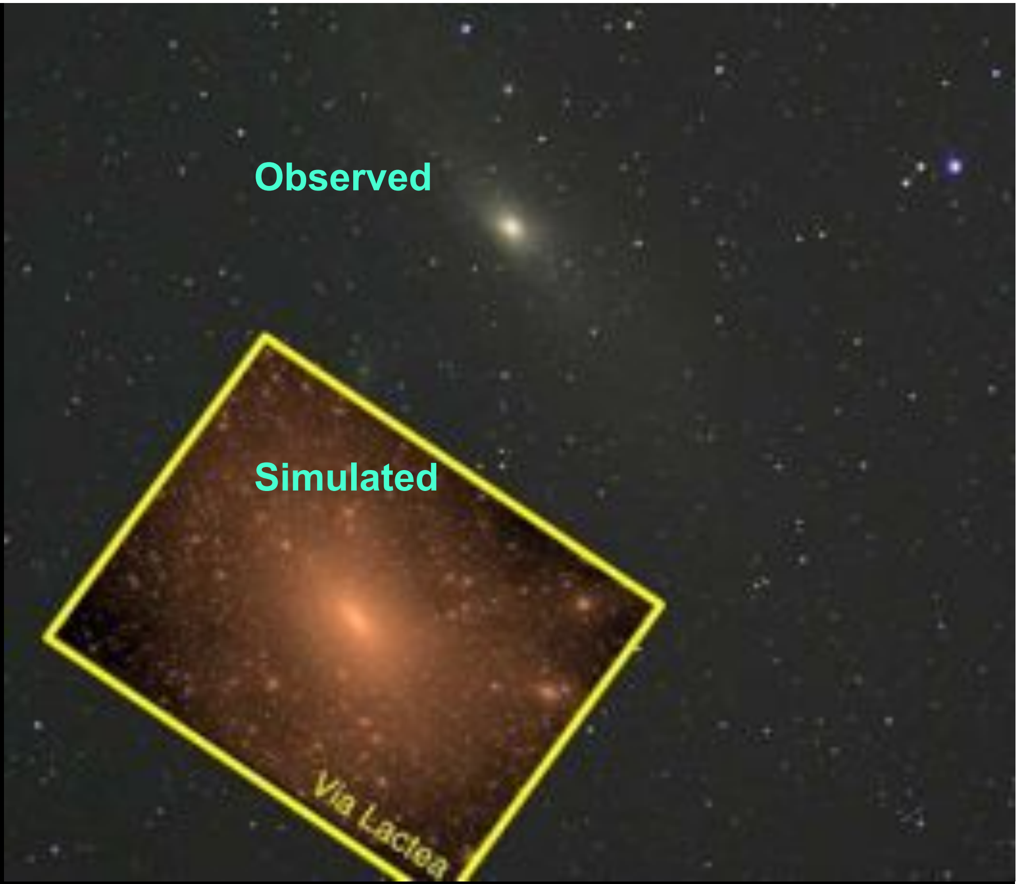
**300000 * 60 kb/s * 2 sensors (microphones) to cover all Dutch dikes**

# Sensor grid: instrument the dikes

## First controlled breach occurred on sept 27th ‘08:

# CosmoGrid

- Motivation:
  **previous simulations found >100 times more substructure than is observed!**




- Simulate large structure formation in the Universe
  - Dark Energy (cosmological constant)
  - Dark Matter (particles)
- Method: Cosmological *N*-body code
- Computation: Intercontinental SuperComputer Grid

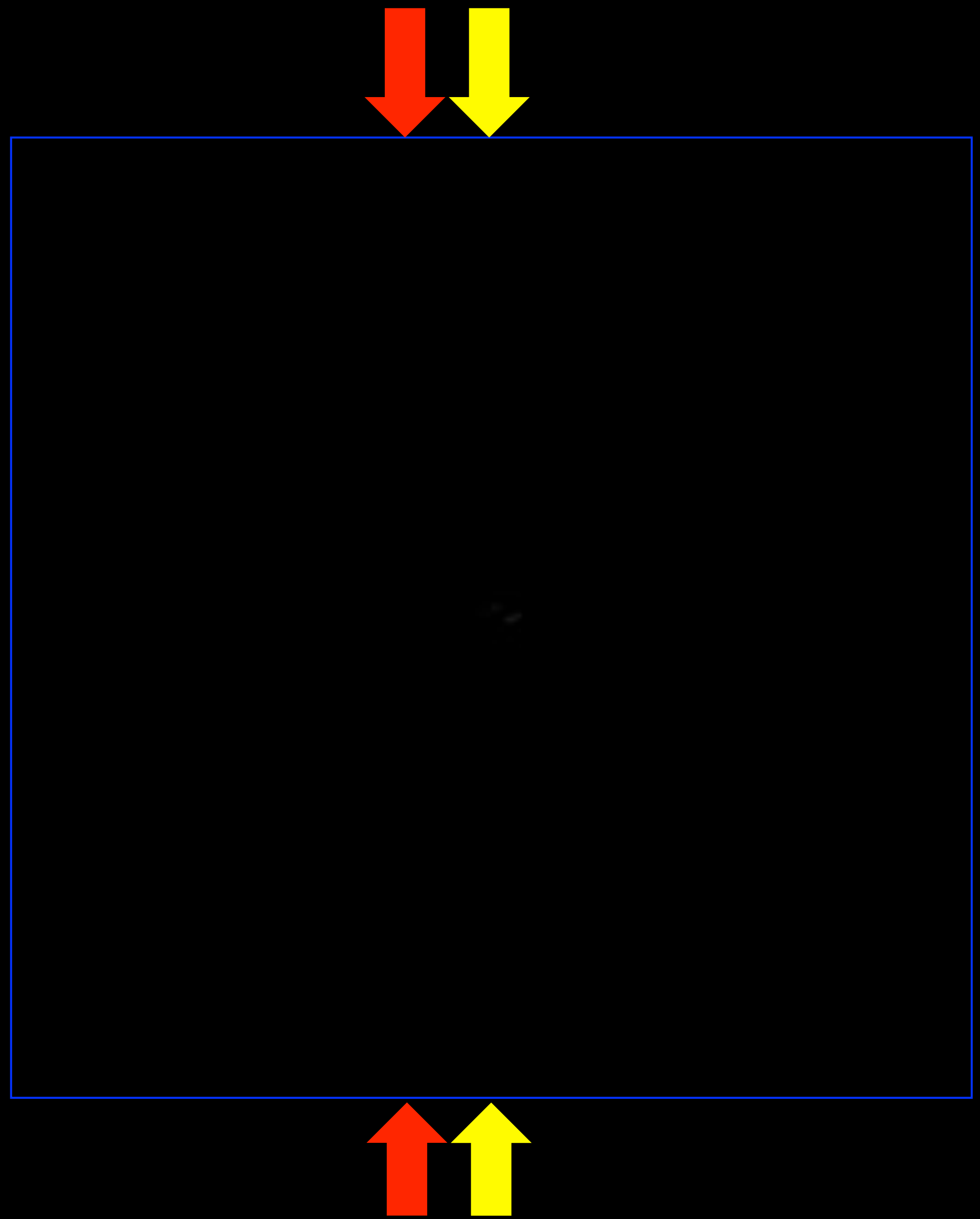
# The hardware setup

10 Mflops/byte

1 Eflops/s

- 2 supercomputers :
  - 1 in Amsterdam (60Tflops Power6 @ SARA)
  - 1 in Tokyo (30Tflops Cray XD0-4 @ CFCA)
- Both computers are connected via an intercontinental optical 10 Gbit/s network

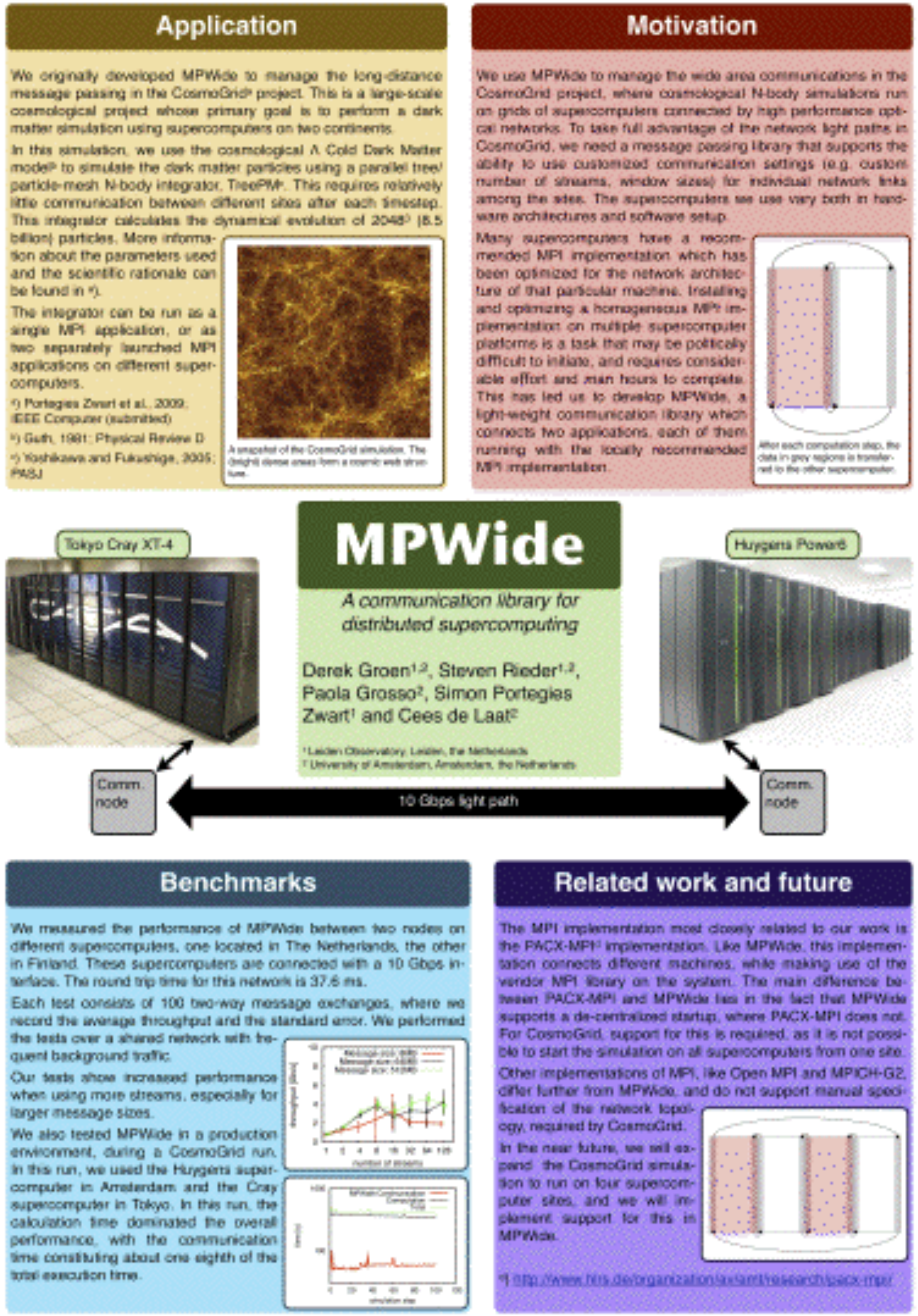
# Auto-balancing Supers



# CosmoGrid

# Supercomputing Grid across Continents and Oceans

# And yes, it works!



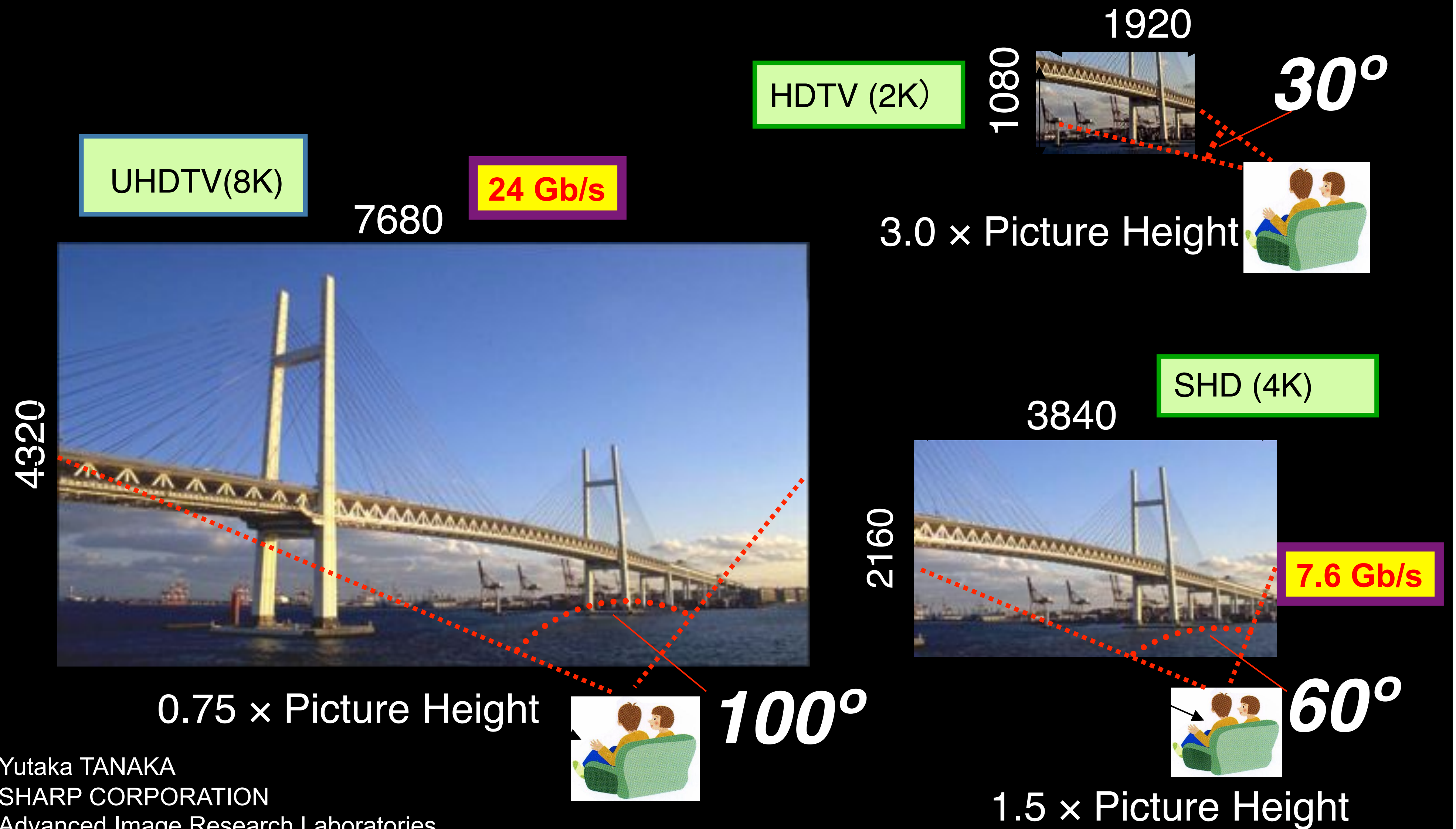
7.6 Gb/s

Real time issue

# CineGrid @ Holland Festival 2007

# Why is more resolution is better?

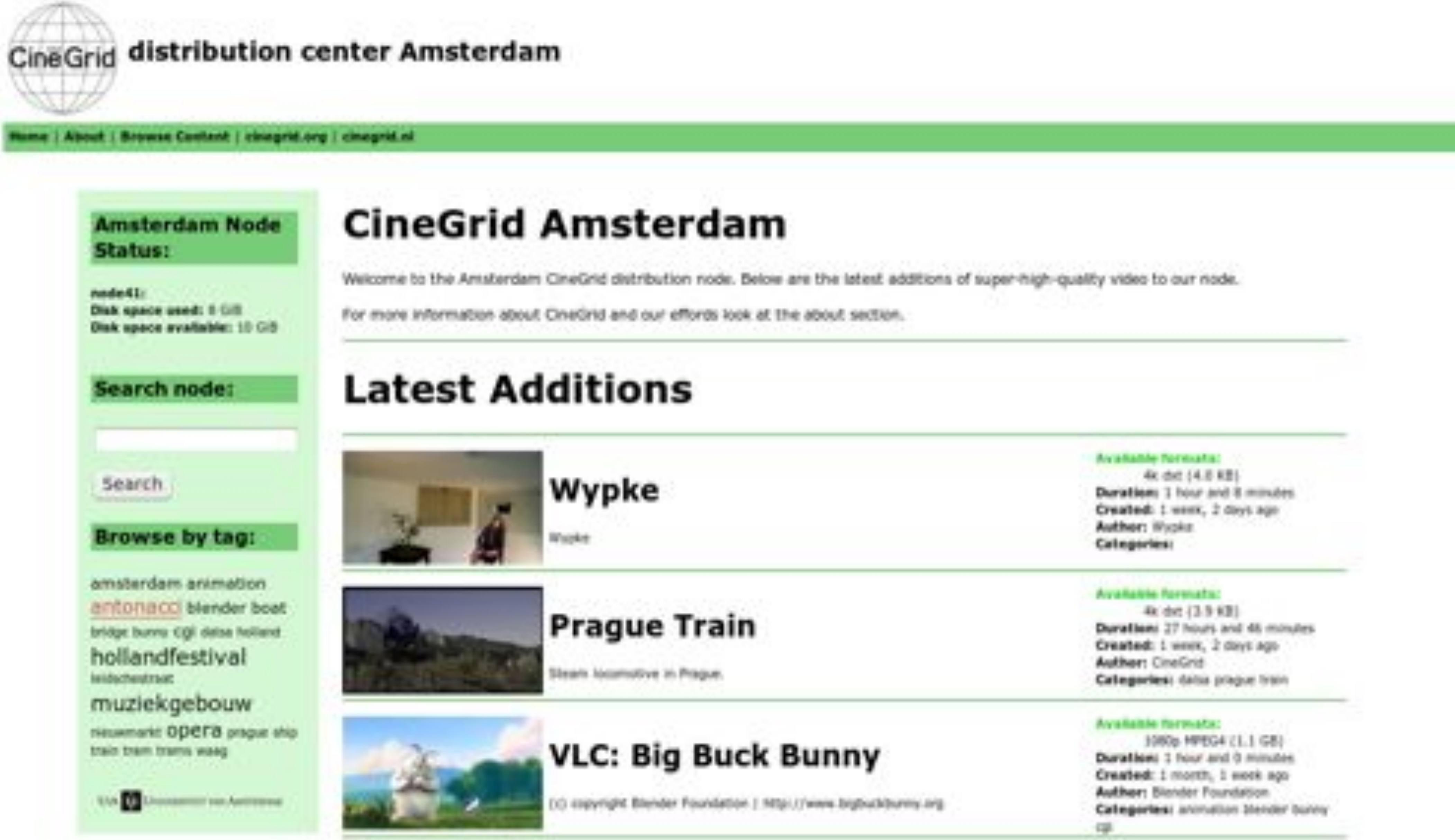
1. More Resolution Allows Closer Viewing of Larger Image
2. Closer Viewing of Larger Image Increases Viewing Angle
3. Increased Viewing Angle Produces Stronger Emotional Response



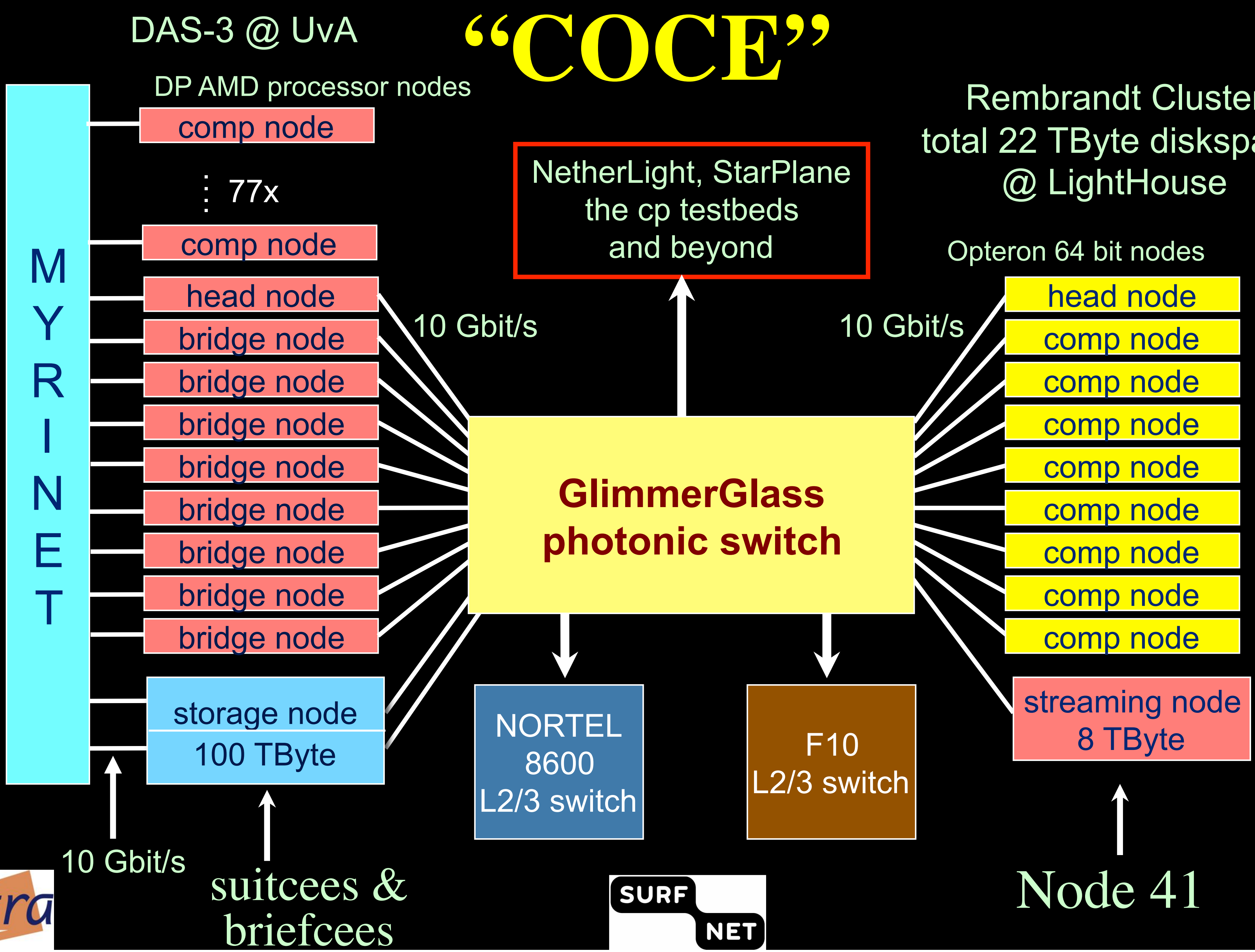


Yutaka TANAKA
SHARP CORPORATION
Advanced Image Research Laboratories

# CineGrid portal



# Amsterdam CineGrid S/F node



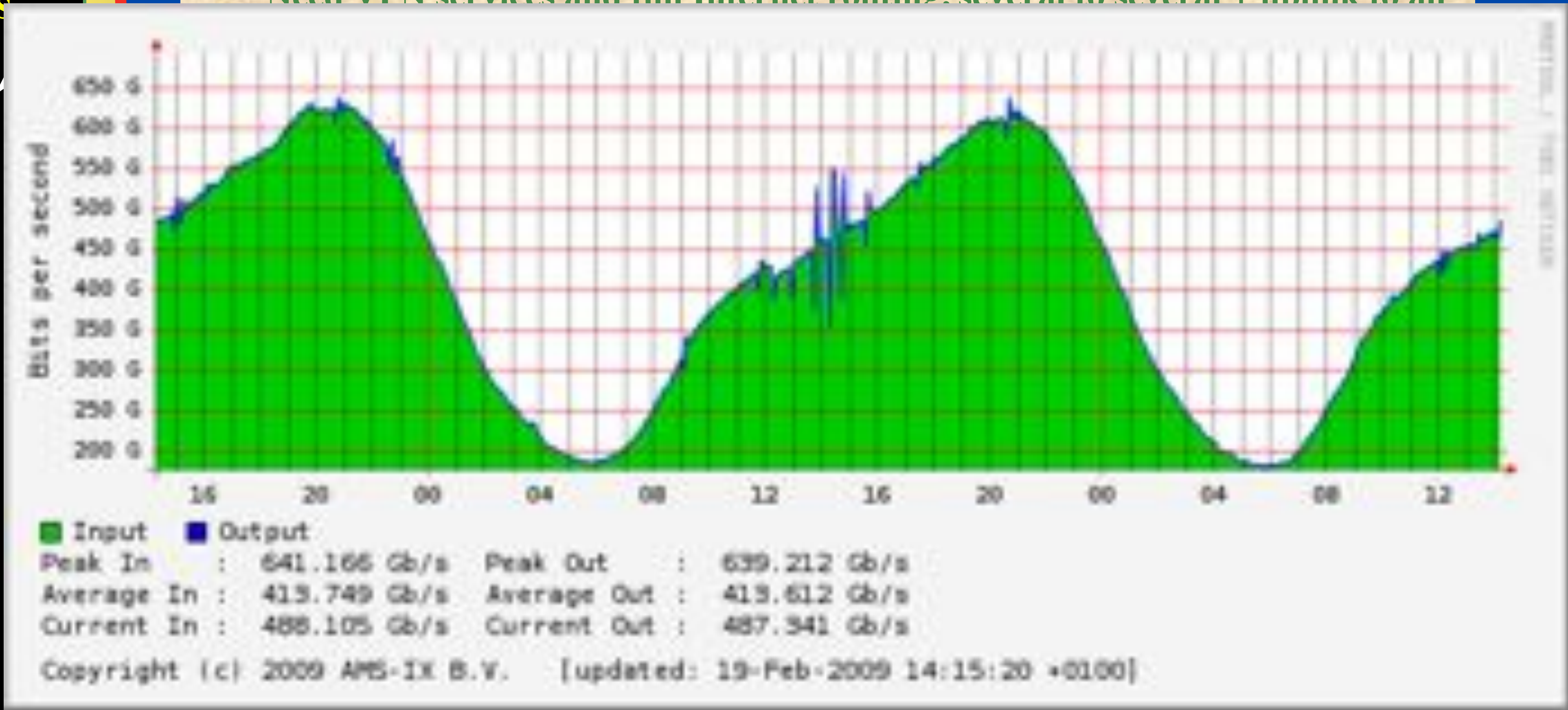
#users

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




B

C

ADSL (12 Mbit/s)

GigE

BW

Ref: Cees de Laat, Erik Radius, Steven Wallace, "The Rationale of the Current Optical Networking Initiatives"
iGrid2002 special issue, Future Generation Computer Systems, volume 19 issue 6 (2003)

# 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

# Hybrid computing

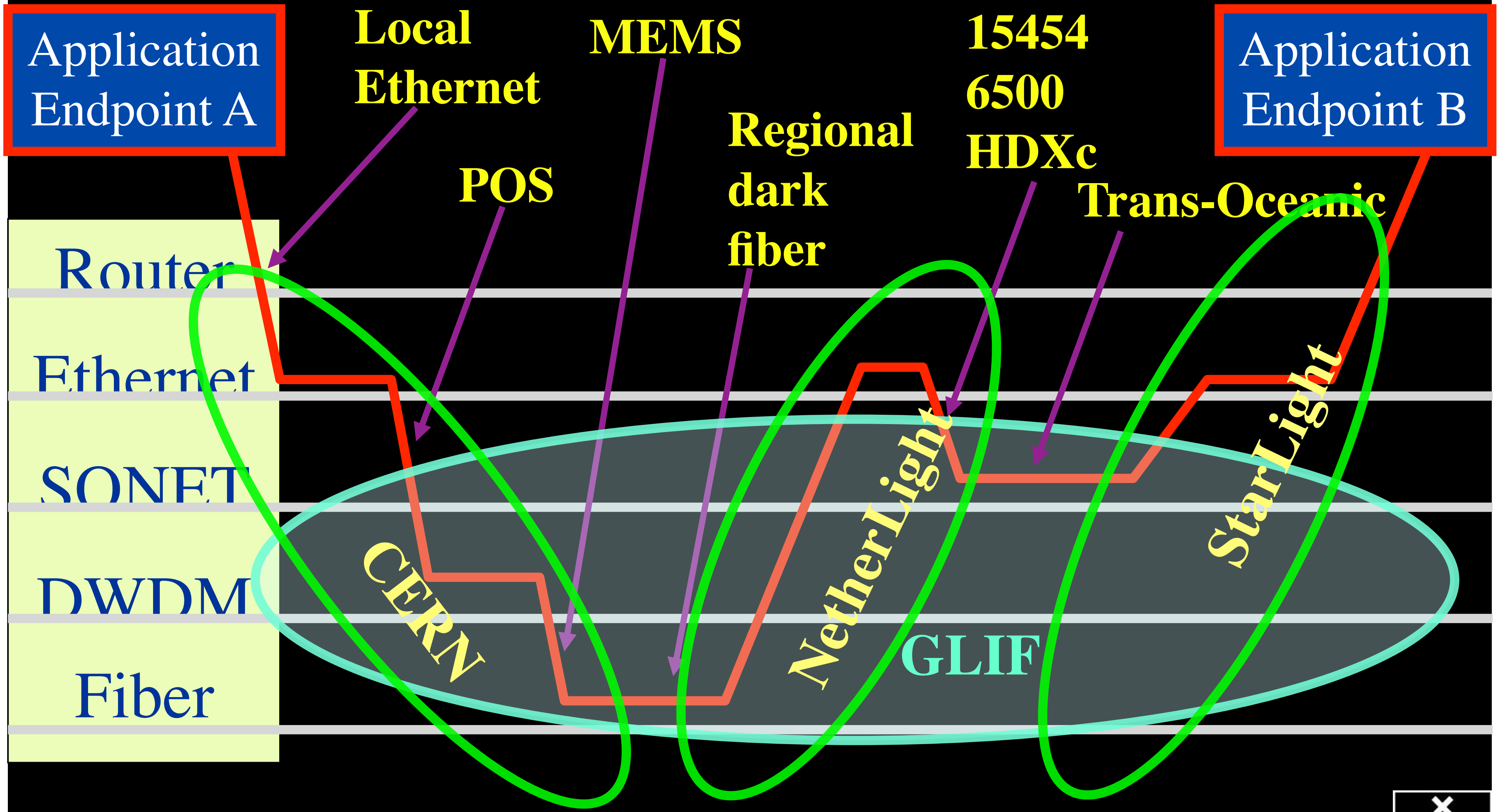
| | | |
|---|---|---|
| Routers | ←→ | Supercomputers |
| Ethernet switches | ←→ | Grid & Cloud |
| Photonic transport | ←→ | GPU's |

What matters:

Energy consumption/multiplication

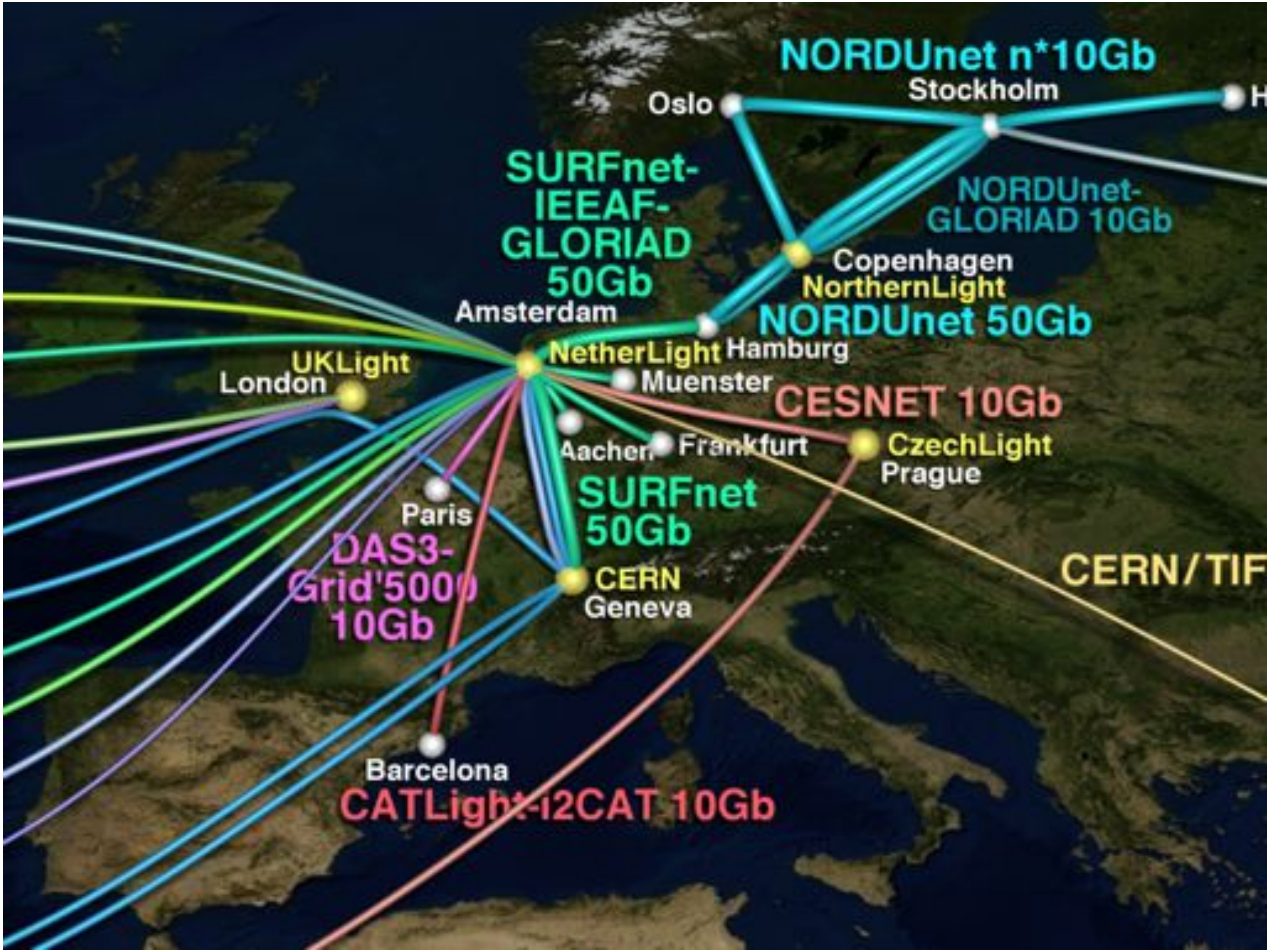
Energy consumption/bit transported

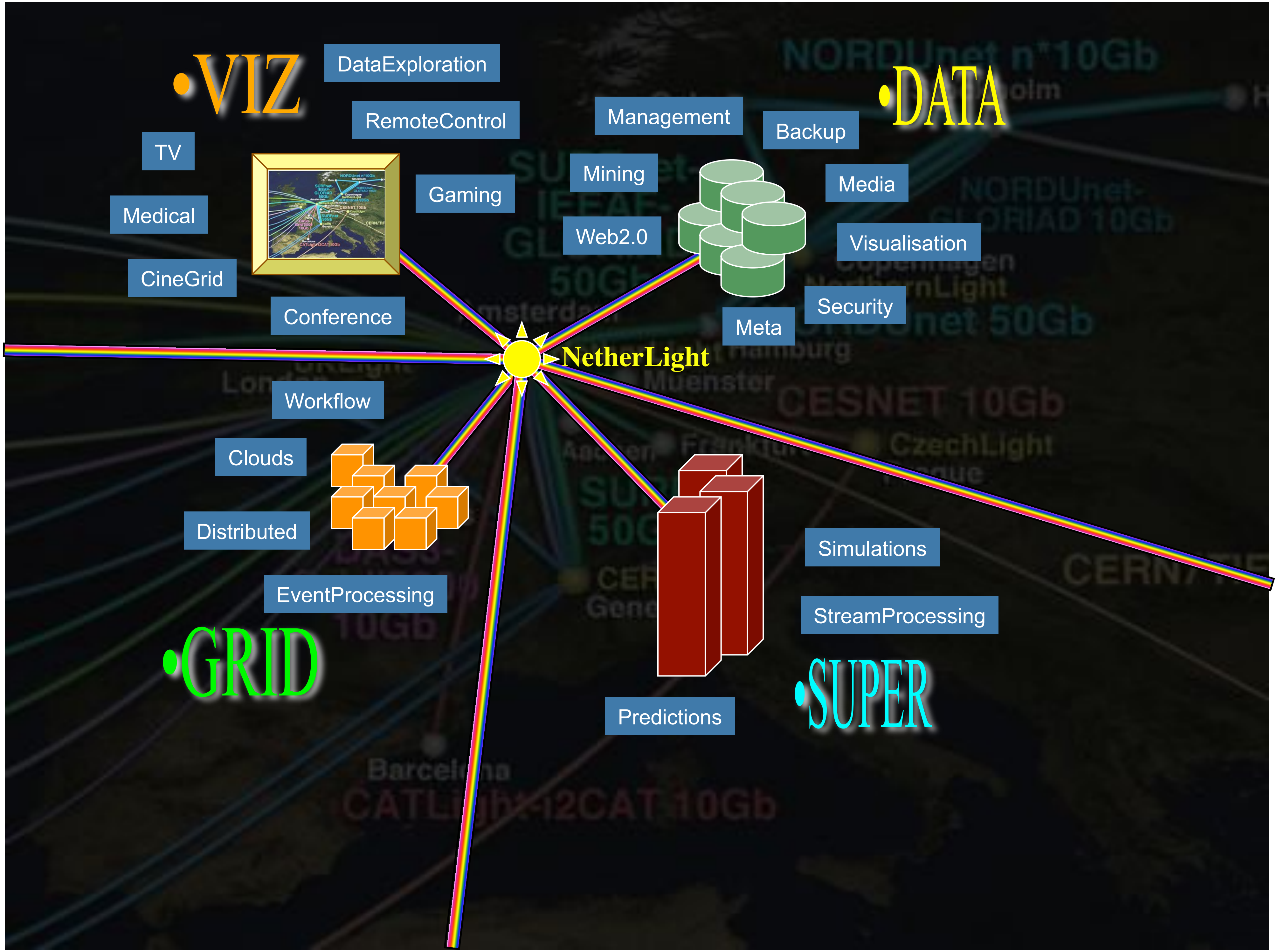
# How low can you go?



GLIF 2008

**Visualization courtesy of Bob Patterson, NCSA**
**Data collection by Maxine Brown.**


NORDUnet n*10Gb
Stockholm
Oslo
SURFnet-
IEEAF-
GLORIAD
50Gb
NORDUnet-
GLORIAD 10Gb
Copenhagen
NorthernLight
Amsterdam
NORDUnet 50Gb
NetherLight
Hamburg
UKLight
London
Muenster
CESNET 10Gb
Aachen
Frankfurt
CzechLight
Prague
SURFnet
50Gb
Paris
DAS3-
Grid'5000
10Gb
CERN
Geneva
CERN/TIF
Barcelona
CATLight-i2CAT 10Gb



•VIZ
DataExploration
RemoteControl
TV
Medical
CineGrid
Gaming
Conference
Management
Backup
•DATA
Mining
Web2.0
Media
Visualisation
Security
Meta
NetherLight
Workflow
Clouds
Distributed
EventProcessing
•GRID
Simulations
StreamProcessing
Predictions
•SUPER


In The Netherlands SURFnet connects between 180:

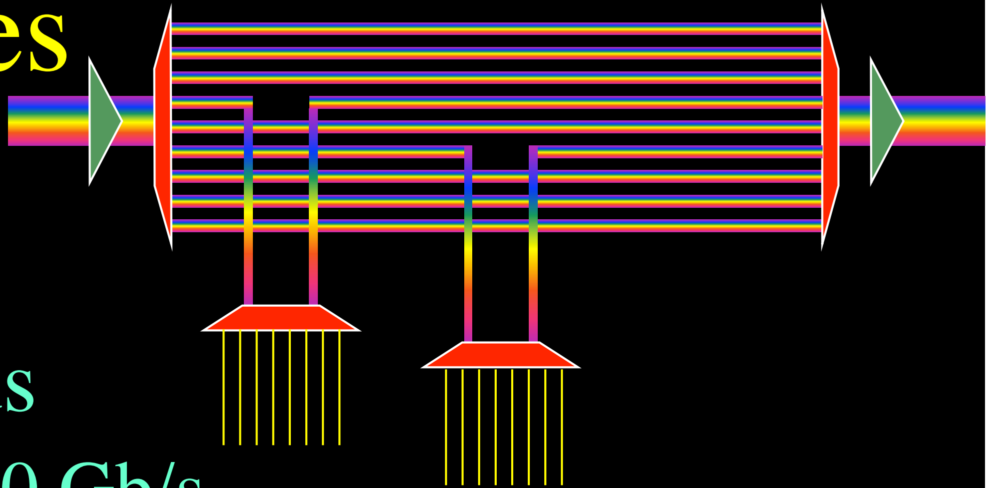
- universities;
- academic hospitals;
- most polytechnics;
- research centers.

with an indirect ~750K user base

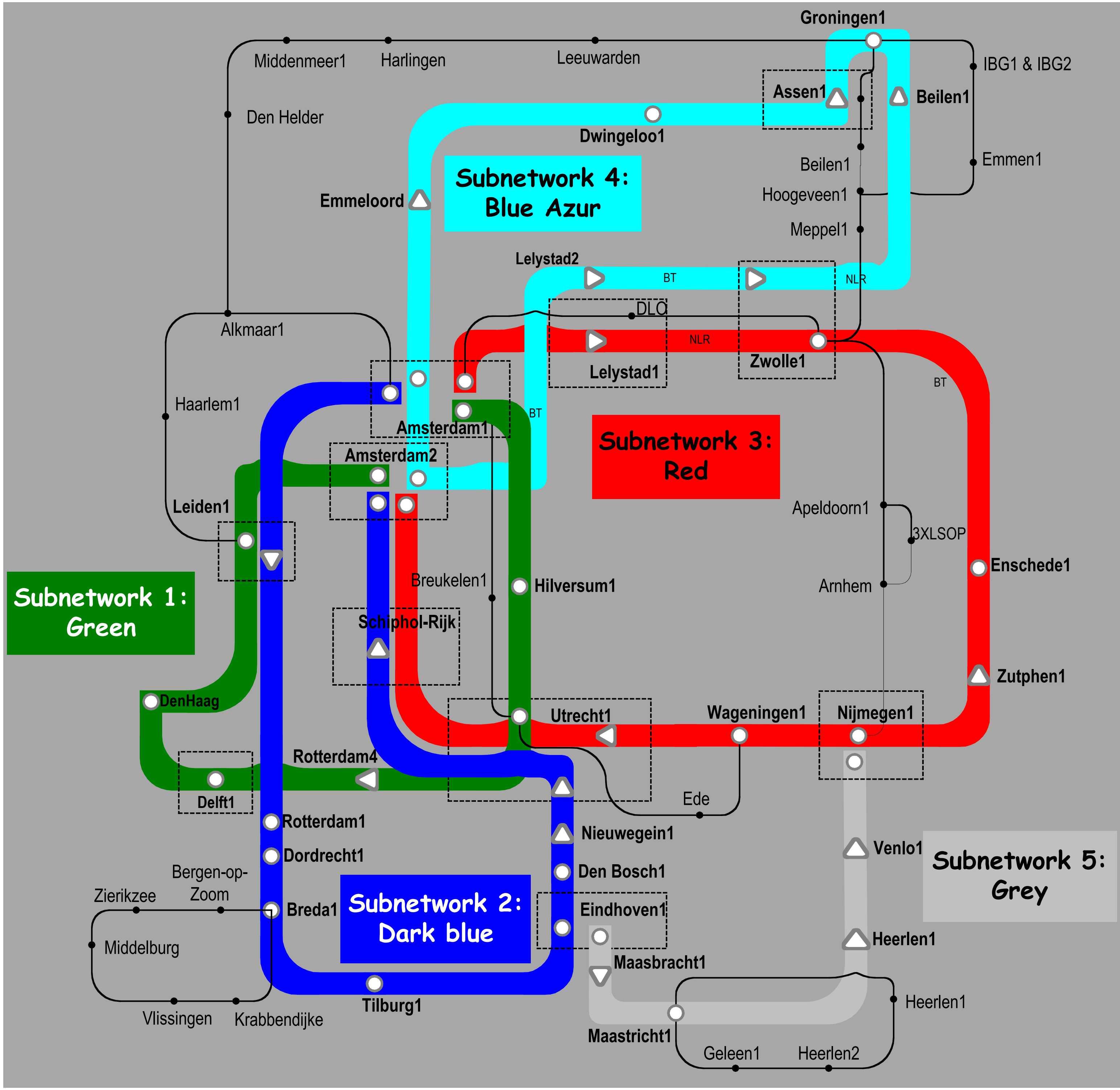
~ 8860 km

scale comparable to railway system

# SURFnet 6 principles



- Based on dark fiber
- 4 DWDM rings of 9 bands
  - Each capable of 10, 40, 100 Gb/s
  - each 4 (100 GHz spacing) or 8 (50 GHz spacing) colors
- Universities each have 1 band to connect their Routers +LightPaths
- Connect with 1 or 10 Gb/s Ethernet LanPhy
- Routing in Amsterdam in 2 core POP's!
- International connectivity in Amsterdam
- Lambda service between ring POP's and to NetherLight



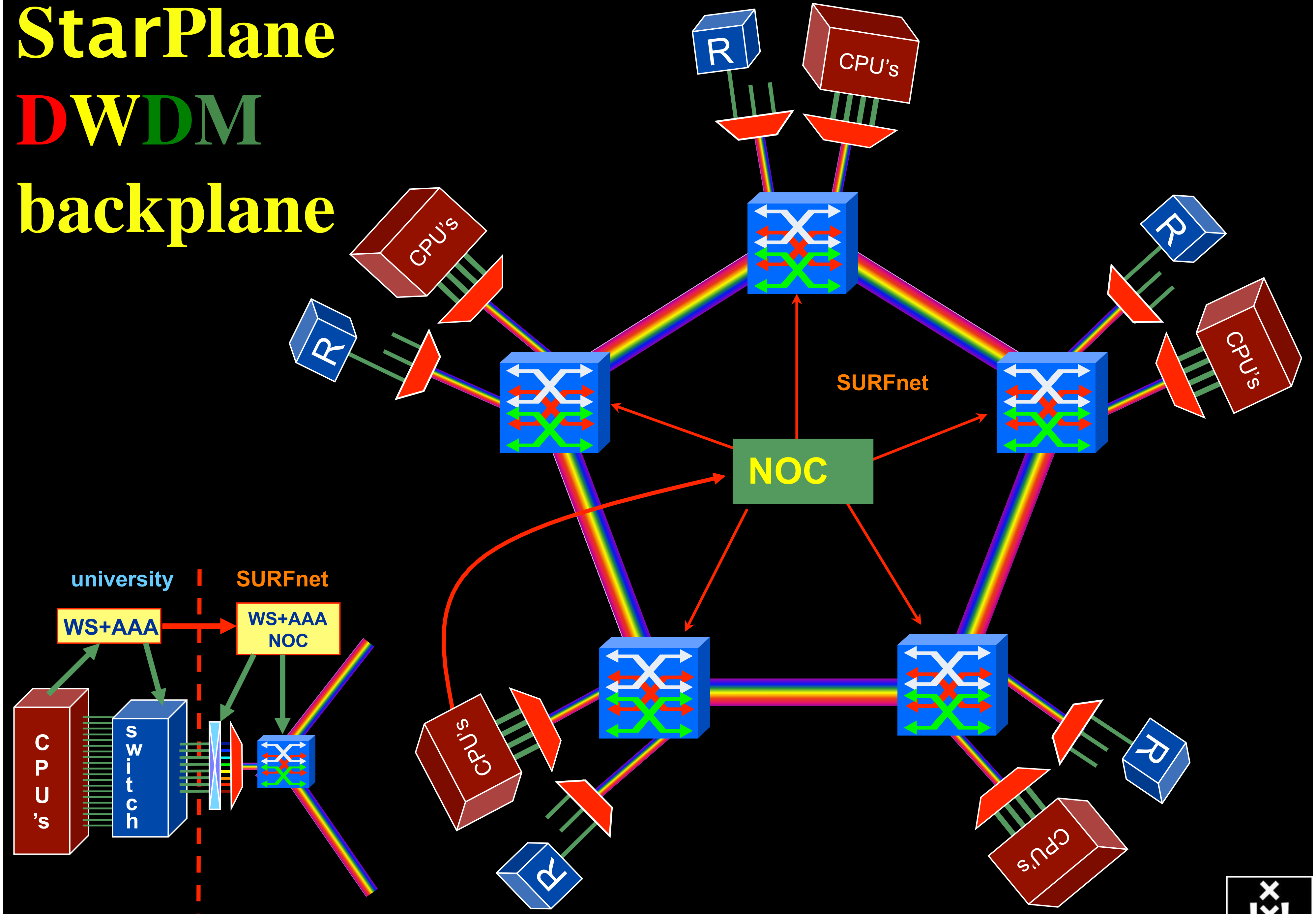

# Common Photonic Layer (CPL) in SURFnet6

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. Cross Domain Authorization using Tokens
4. RDF/Network Description Language
5. Tera-networking
6. Programmable networks

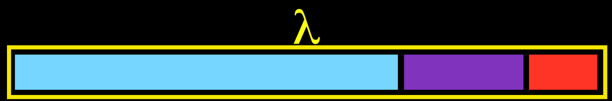
# StarPlane DWDM backplane



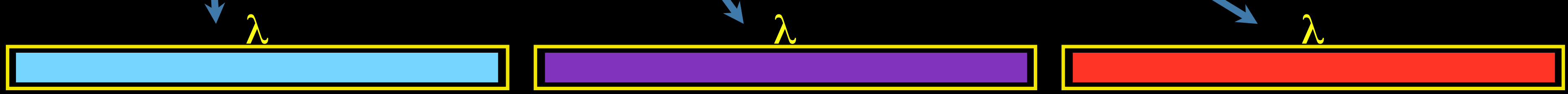

CdL

# QOS in a non destructive way!

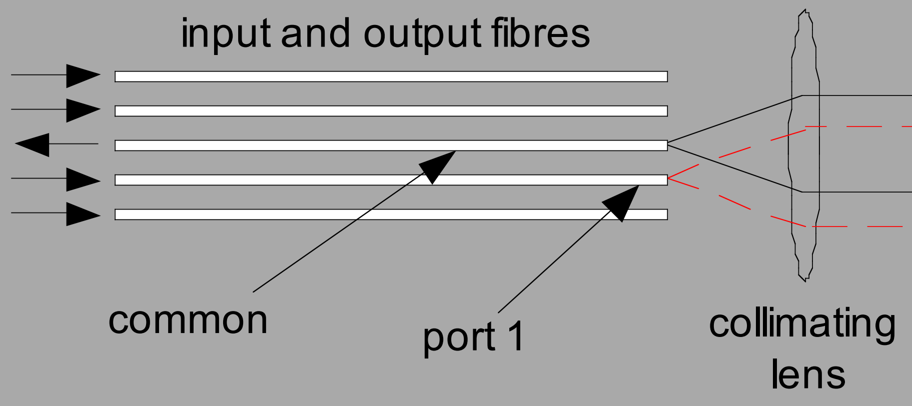
- Destructive QOS:
  - have a link or λ
  - set part of it aside for a lucky few under higher priority
  - rest gets less service

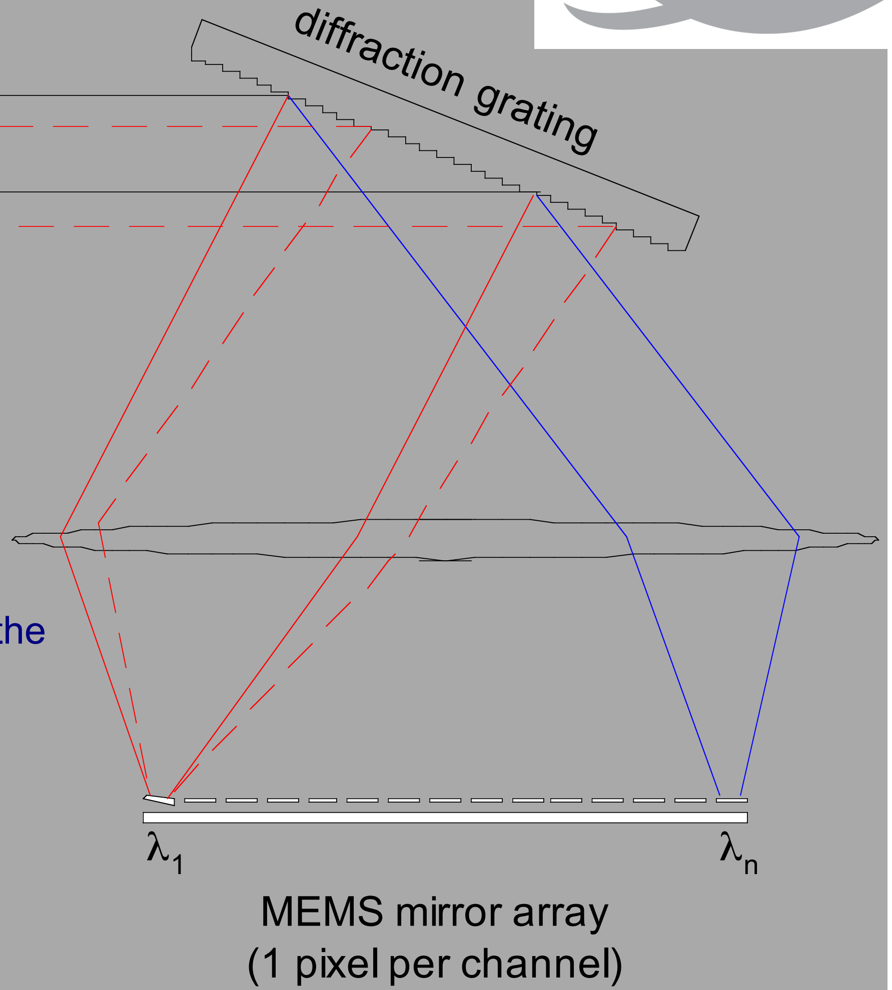



- Constructive QOS:
  - have a λ
  - add other λ's as needed on separate colors
  - move the lucky ones over there
  - rest gets also a bit happier!



# 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

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

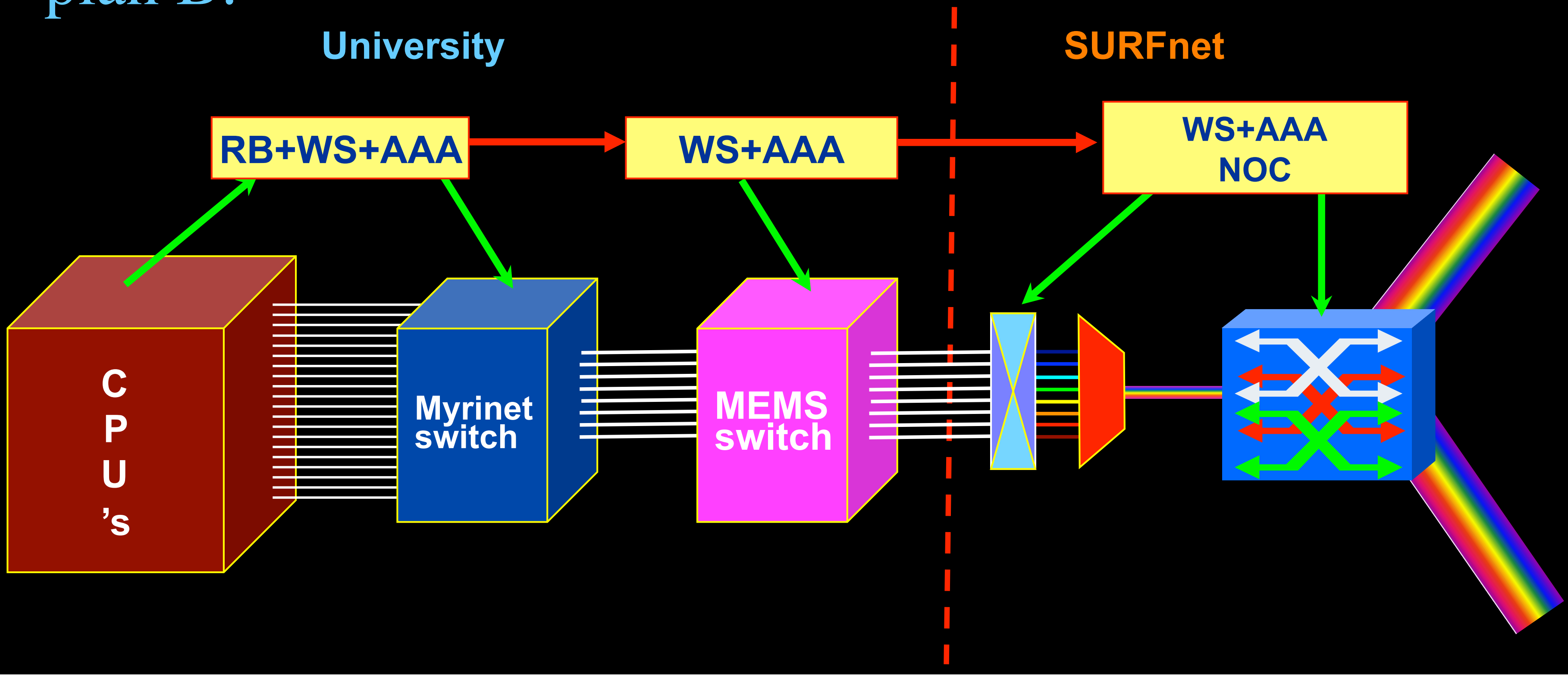
T(f)

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

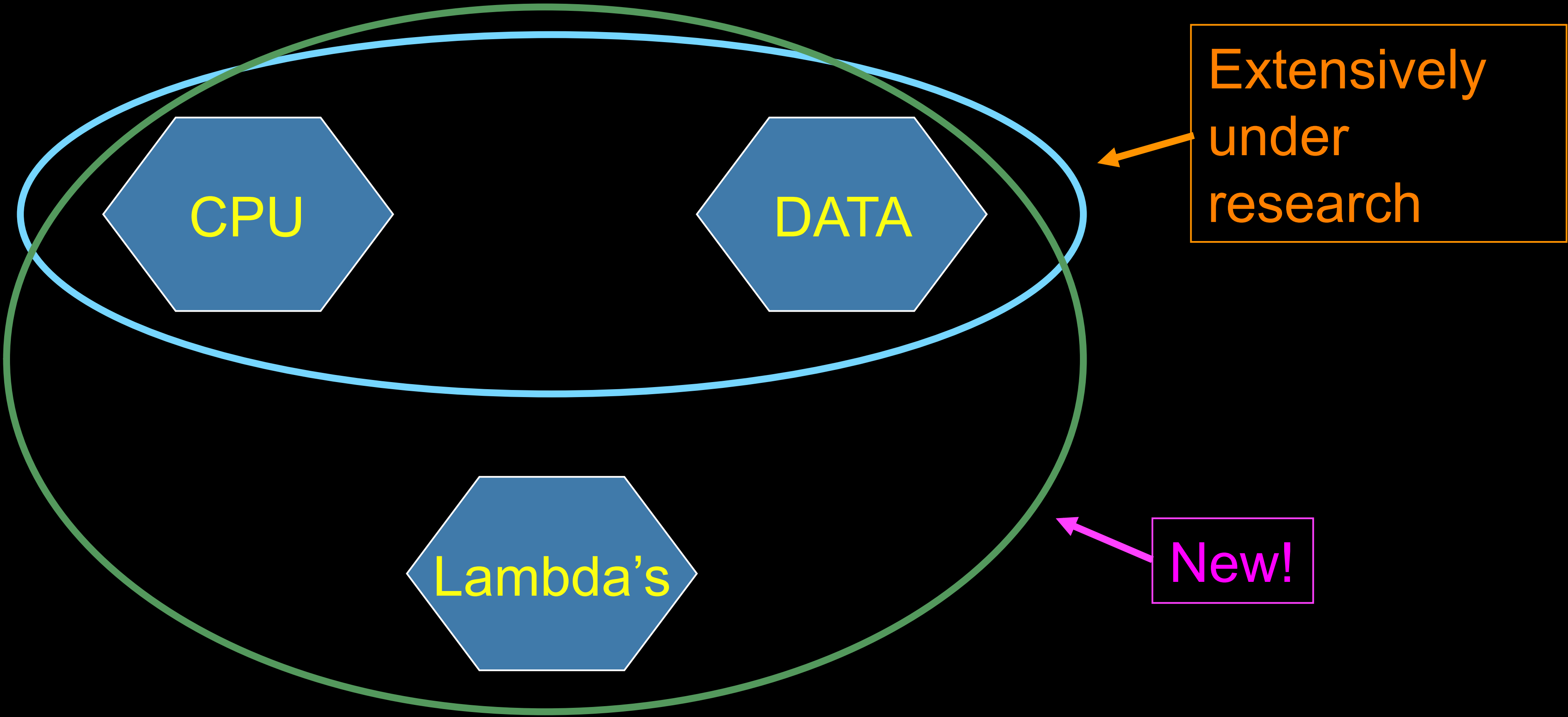
(ps. due to power ~ square E the signal to send **looks** like uncompensated received but is not)

# 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:



# 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.

Net Tests between DAS-3 Hosts

http://rembrandt0.uva.netherlight.nl/rtpl/das3/table/net_data.html

# MAY 31th 2007

## 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**

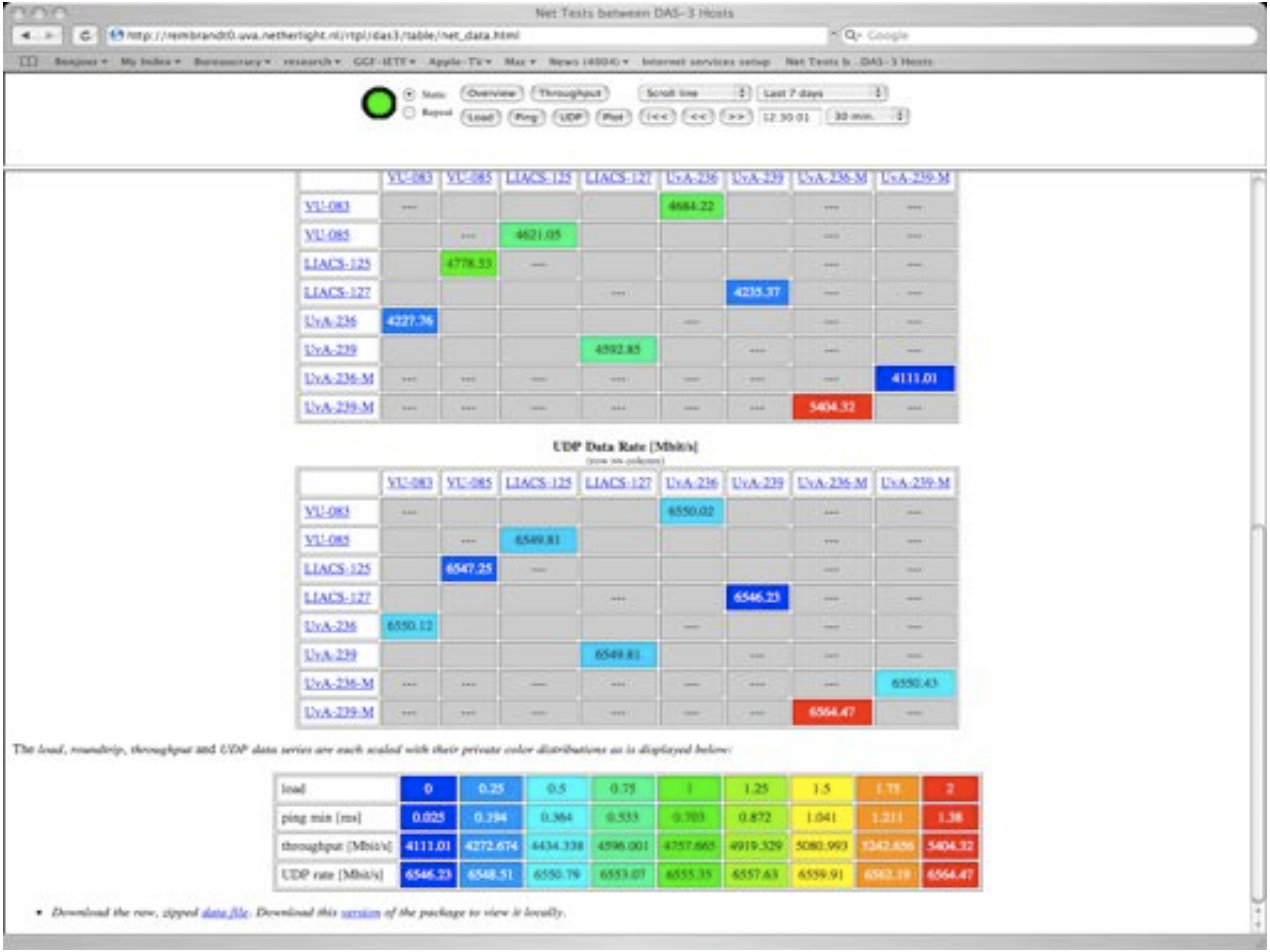
| VU-083 | VU-085 | LIACS-125 | LIACS-127 | UvA-236 | UvA-239 | UvA-236-M | UvA-239-M |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0.087 | 0 | 0.013 | 0.01 | 0.017 | 0.15 |

**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 | | --- | 4621.05 | | | | --- | --- |


Net Tests between DAS-3 Hosts

http://rembrandt0.uva.netherlight.nl/rtpl/das3/table/net_data.html

| | 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 | | | | --- | --- |
| LIACS-125 | | 4778.53 | --- | | | | --- | --- |
| LIACS-127 | | | | --- | | 4235.37 | --- | --- |
| UvA-236 | 4227.76 | | | | --- | | --- | --- |
| UvA-239 | | | | 4592.85 | | --- | --- | --- |
| UvA-236-M | --- | --- | --- | --- | --- | --- | --- | 4111.01 |
| UvA-239-M | --- | --- | --- | --- | --- | --- | 5404.32 | --- |

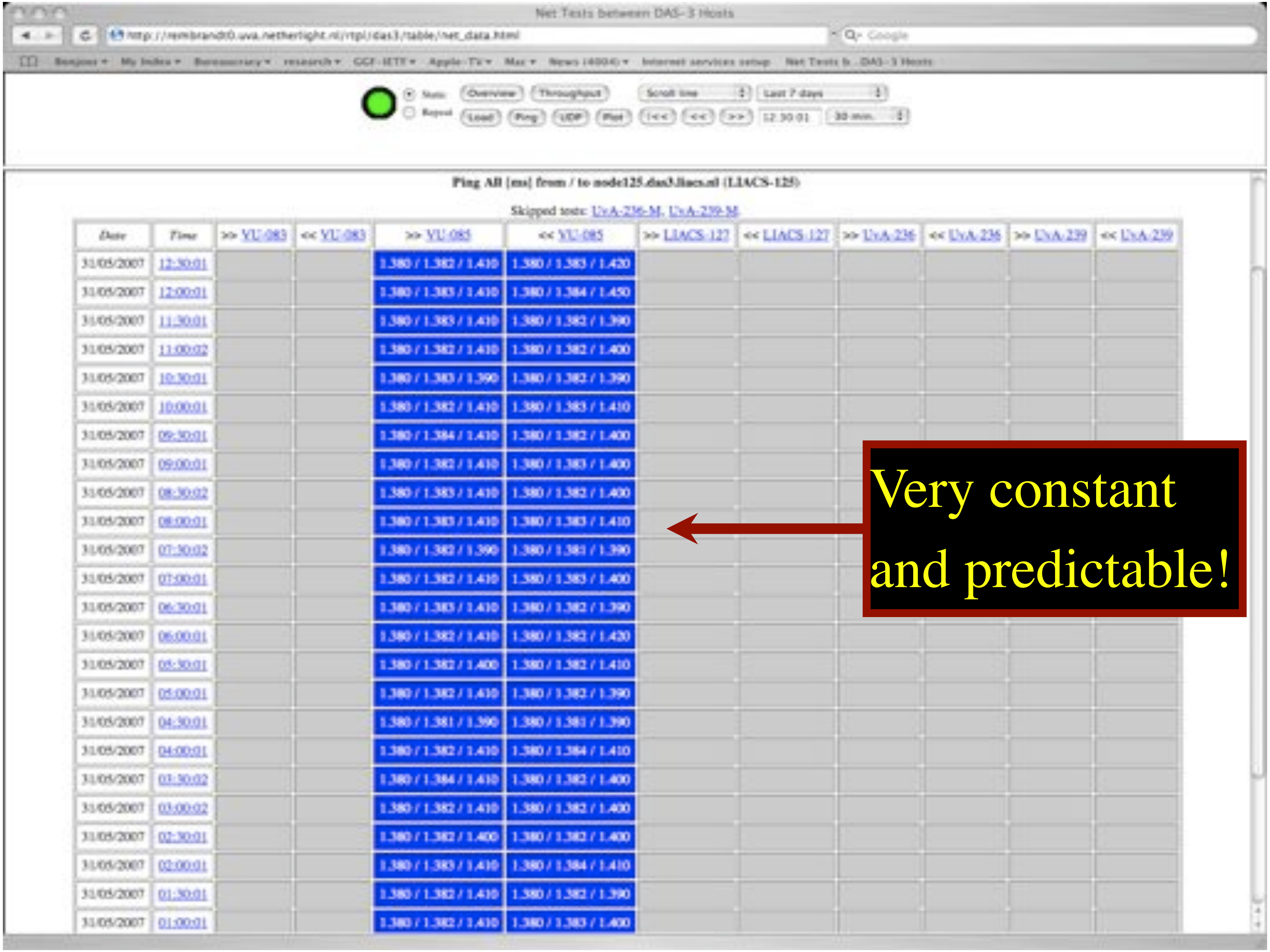
**UDP Data Rate [Mbit/s]**

| | VU-083 | VU-085 | LIACS-125 | LIACS-127 | UvA-236 | UvA-239 | UvA-236-M | UvA-239-M |
|---|---|---|---|---|---|---|---|---|
| VU-083 | --- | | | | 6550.02 | | --- | --- |
| VU-085 | | --- | 6549.81 | | | | --- | --- |
| LIACS-125 | | 6547.25 | --- | | | | --- | --- |
| LIACS-127 | | | | --- | | 6546.23 | --- | --- |
| UvA-236 | 6550.12 | | | | --- | | --- | --- |
| UvA-239 | | | | 6549.81 | | --- | --- | --- |
| UvA-236-M | --- | --- | --- | --- | --- | --- | --- | 6550.43 |
| UvA-239-M | --- | --- | --- | --- | --- | --- | 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 | 0.194 | 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.*



Net Tests between DAS-3 Hosts

http://rembrandt0.uva.netherlight.nl/rtpl/das3/table/net_data.html

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

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

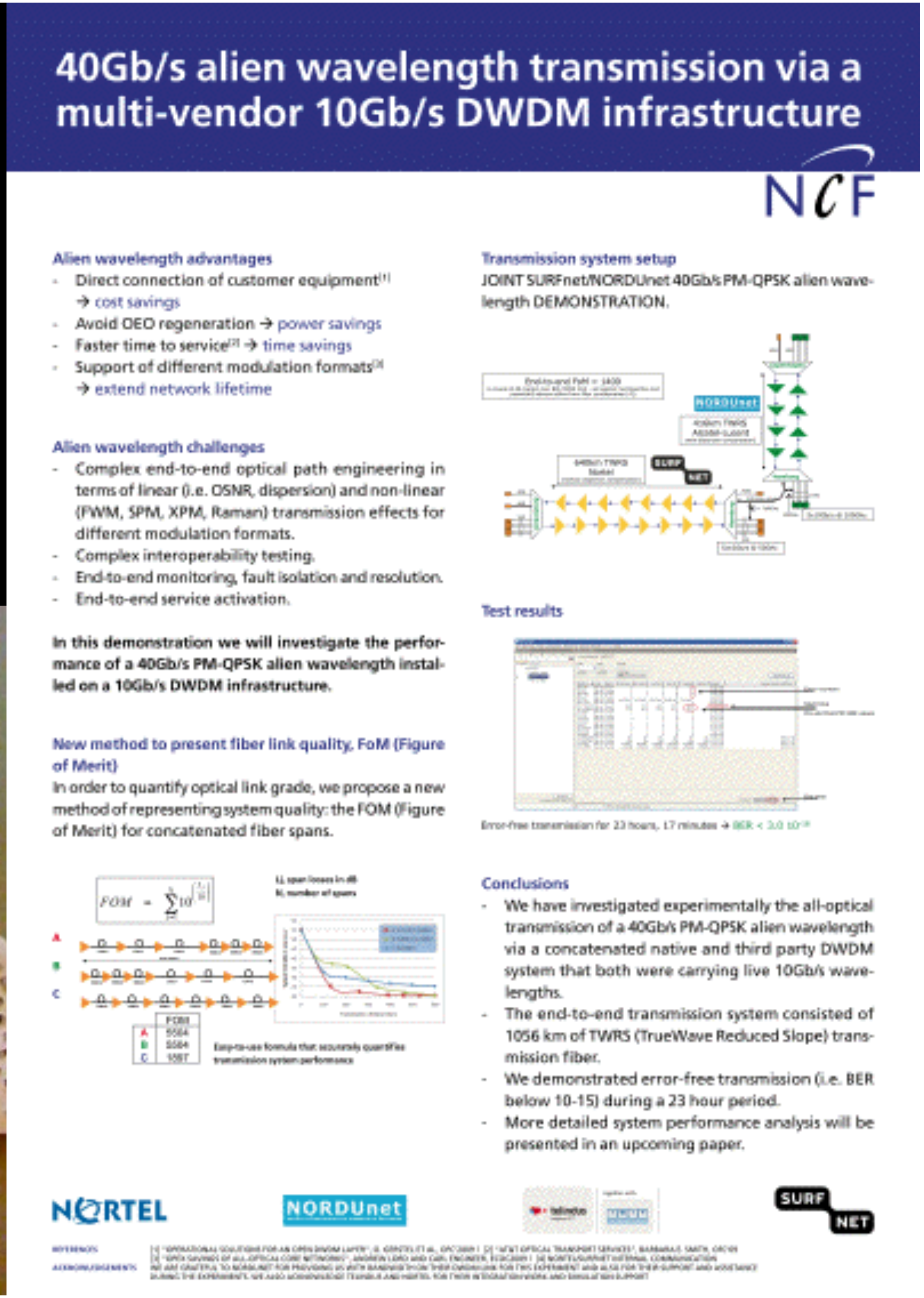
| *Date* | *Time* | >> VU-083 | << VU-083 | >> VU-085 | << VU-085 | >> LIACS-127 | << LIACS-127 | >> UvA-236 | << UvA-236 | >> UvA-239 | << UvA-239 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 31/05/2007 | 12:30:01 | | | 1.380 / 1.382 / 1.410 | 1.380 / 1.383 / 1.420 | | | | | | |
| 31/05/2007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

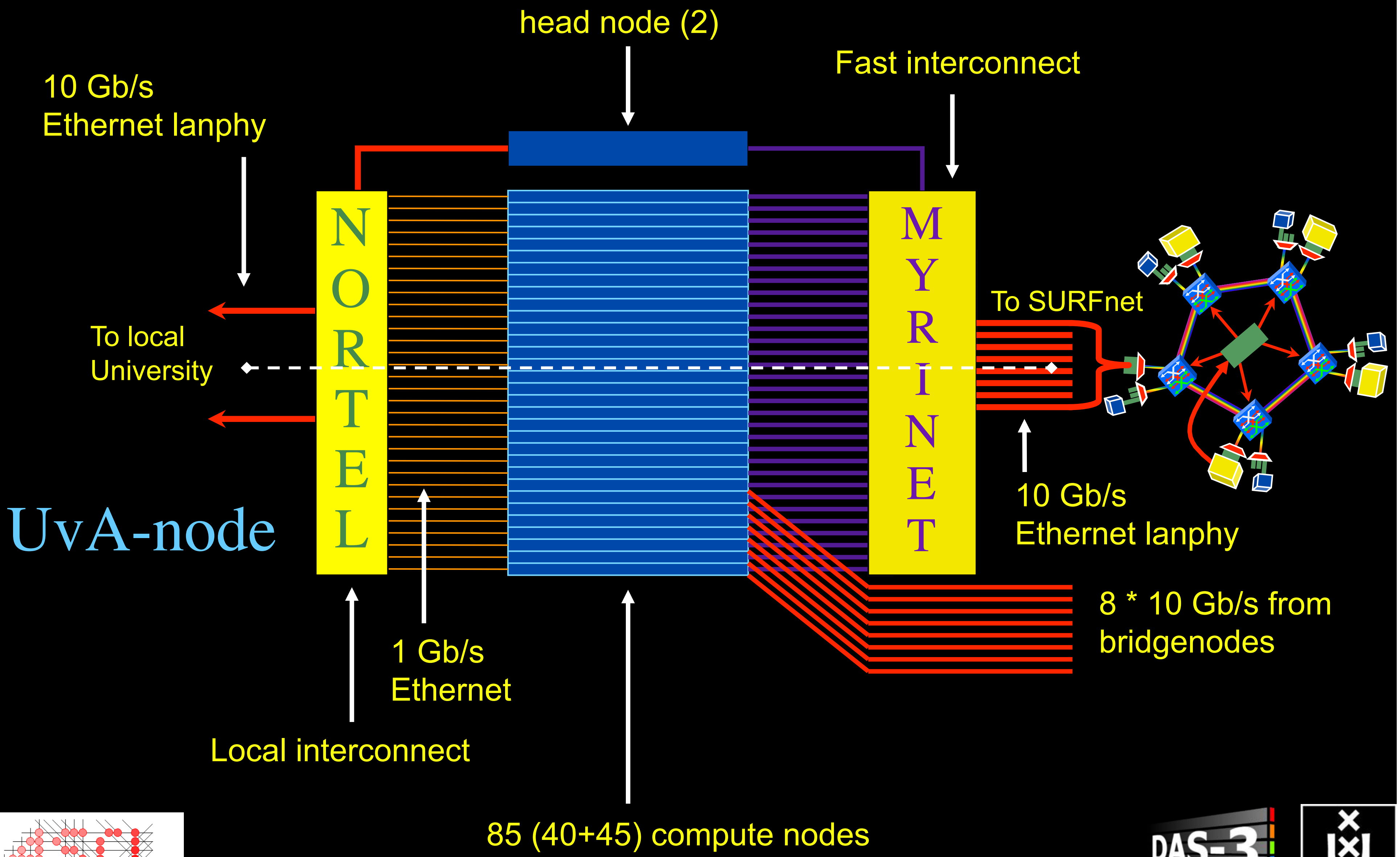
# What makes StarPlane fly?

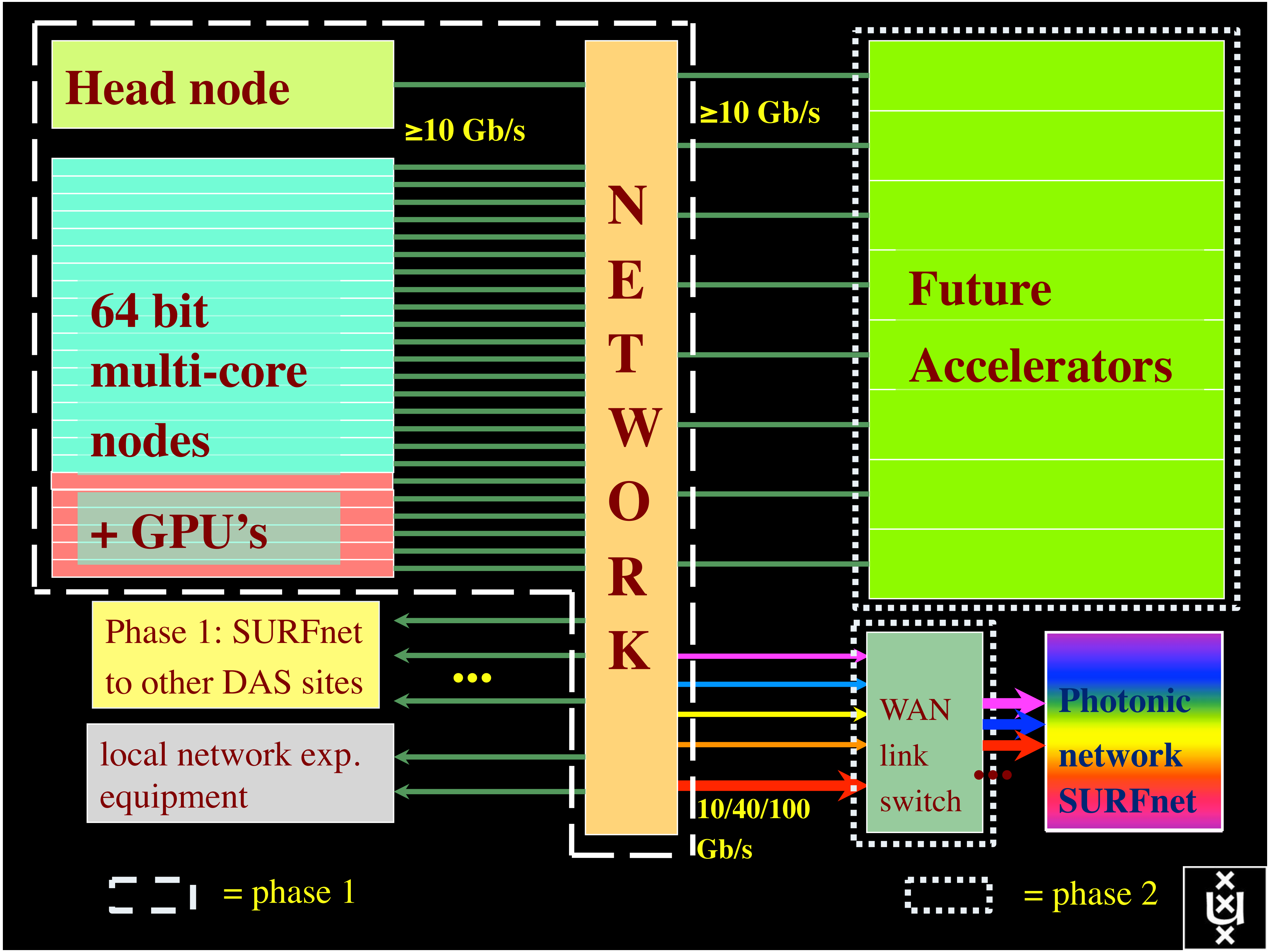
- Wavelength Selective Switches
  - for the “low cost” photonics
- Sandbox by confining StarPlane to one band
  - for experimenting on a production network
- Optimization of the controls to turn on/off a Lambda
  - direct access to part of the controls at the NOC
- electronic Dynamically Compensating 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 facilitate policy enabled cross domain signaling

# Alien light
# From idea to realisation!



# DAS-3 Cluster Architecture

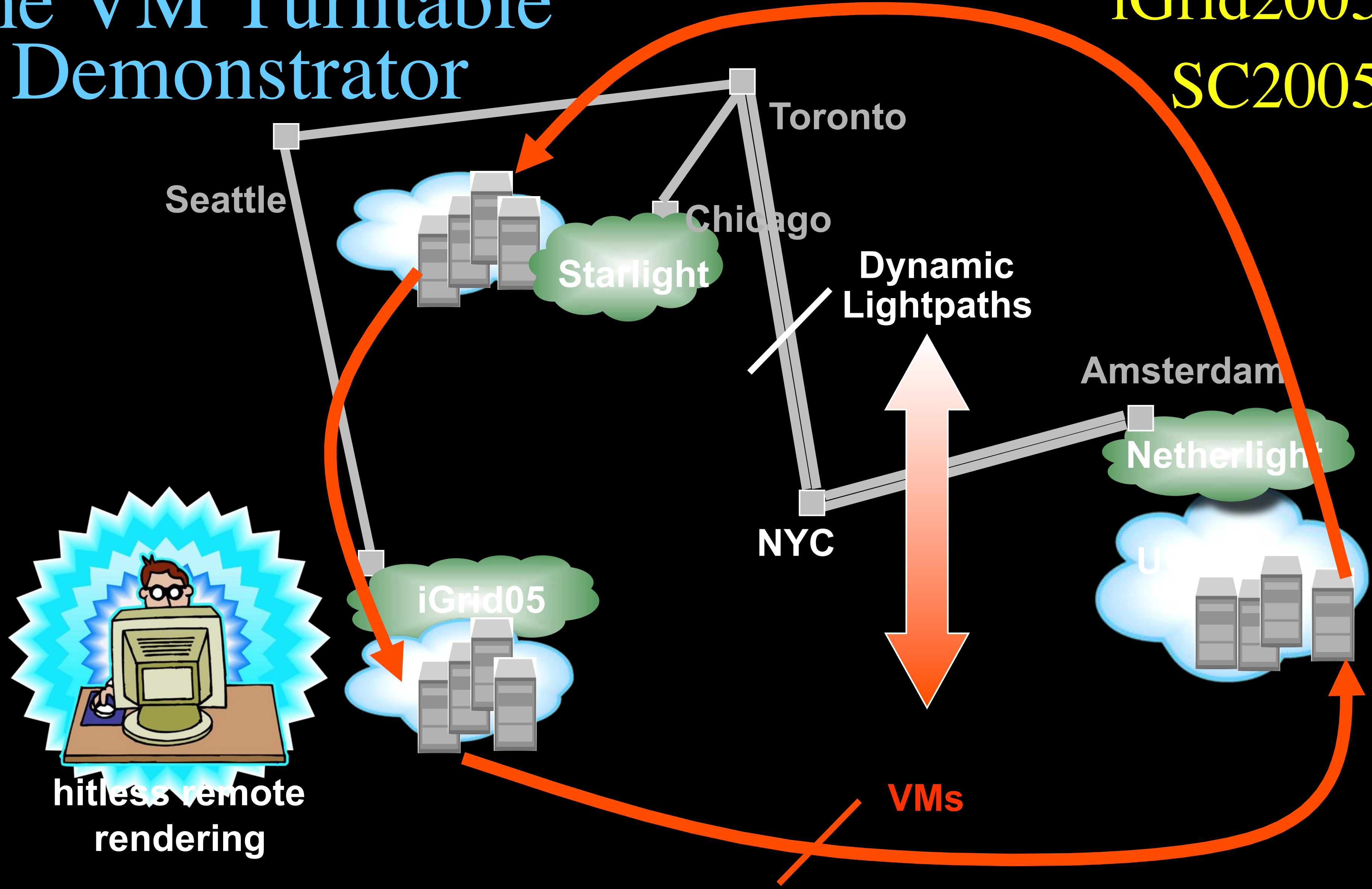



Head node
≥10 Gb/s
64 bit
multi-core
nodes
+ GPU's
NETWORK
≥10 Gb/s
Future
Accelerators
Phase 1: SURFnet
to other DAS sites
...
local network exp.
equipment
WAN
link
switch
10/40/100
Gb/s
Photonic
network
SURFnet
= phase 1
= phase 2

# The VM Turntable Demonstrator

iGrid2005

SC2005




**The VMs that are live-migrated run an iterative search-refine-search workflow against data stored in different databases at the various locations. A user in San Diego gets hitless rendering of search progress as VMs spin around**

# 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!

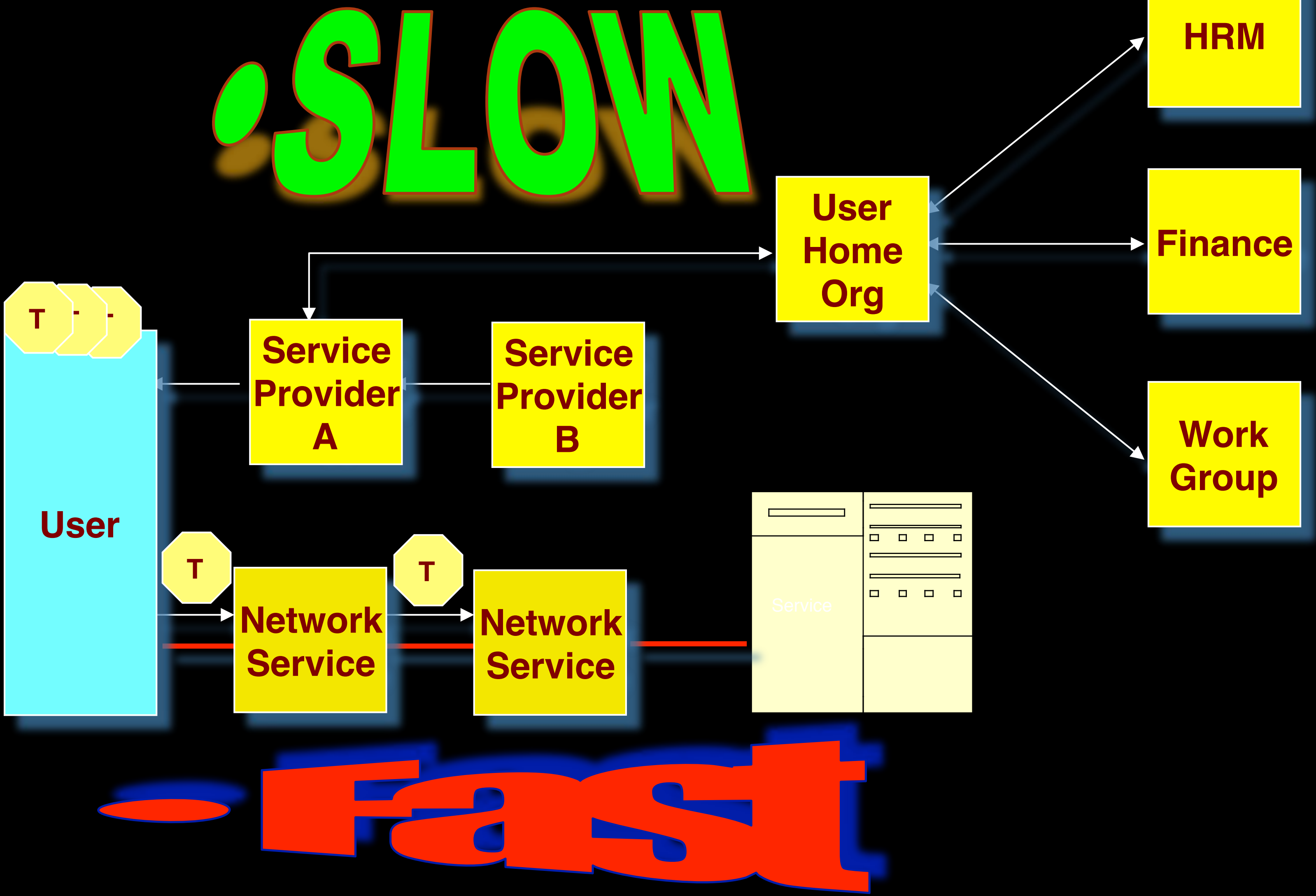
# Contents

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

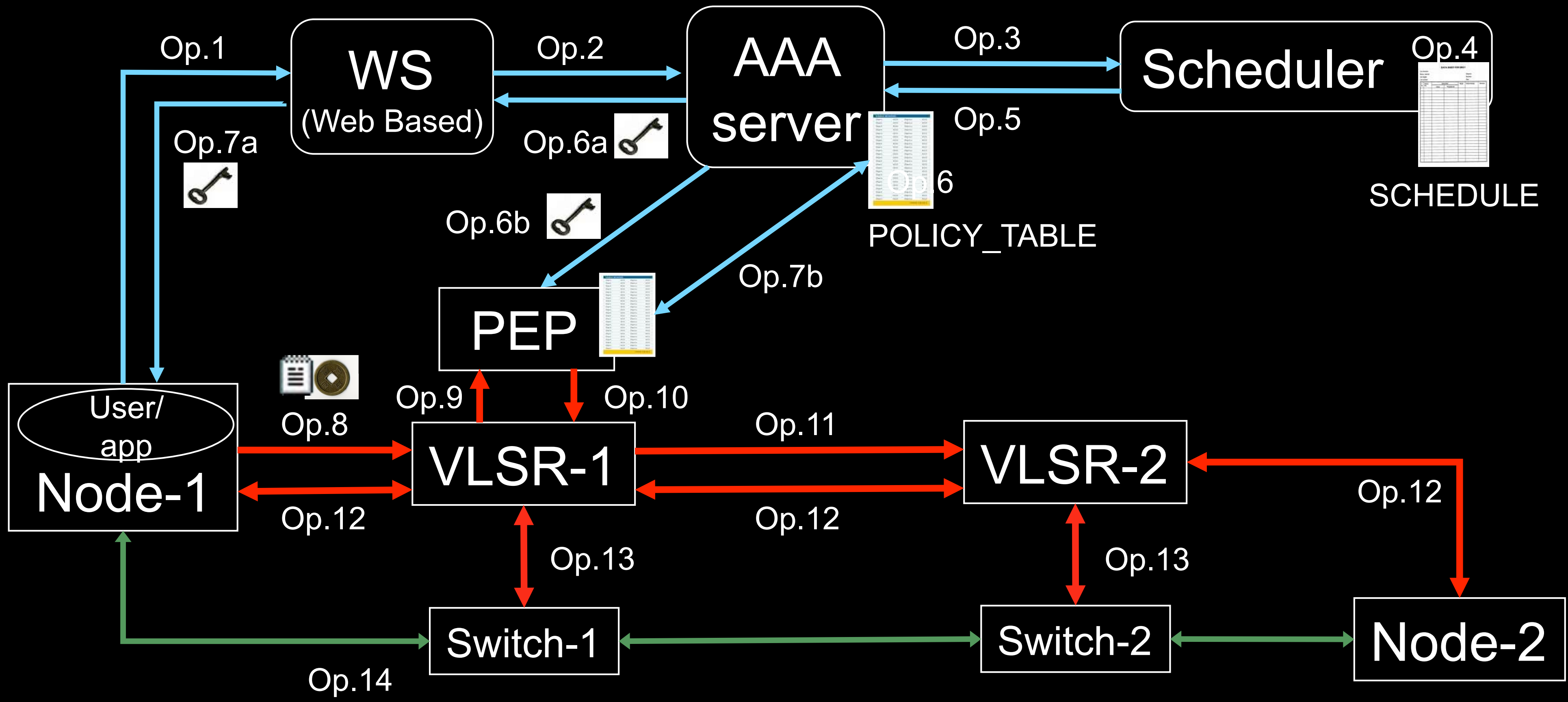
# Simple service access


Pitlochry, Scotland - Summer 2005

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



# DRAGON GMPLS & TBN Demo, SC06 Tampa




1. User (on Node1) requests a path via web to the WS.
2. WS sends the XML requests to the AAA server.
3. AAA server calculates a hashed index number and submits a request to the Scheduler.
4. Scheduler checks the SCHEDULE and add new entry.
5. Scheduler confirms the reservation to the AAA.
6. AAA server updates the POLICY_TABLE.

6a. AAA server issues an encrypted key to the WS.

6b. AAA server passes the same key to the PEP.

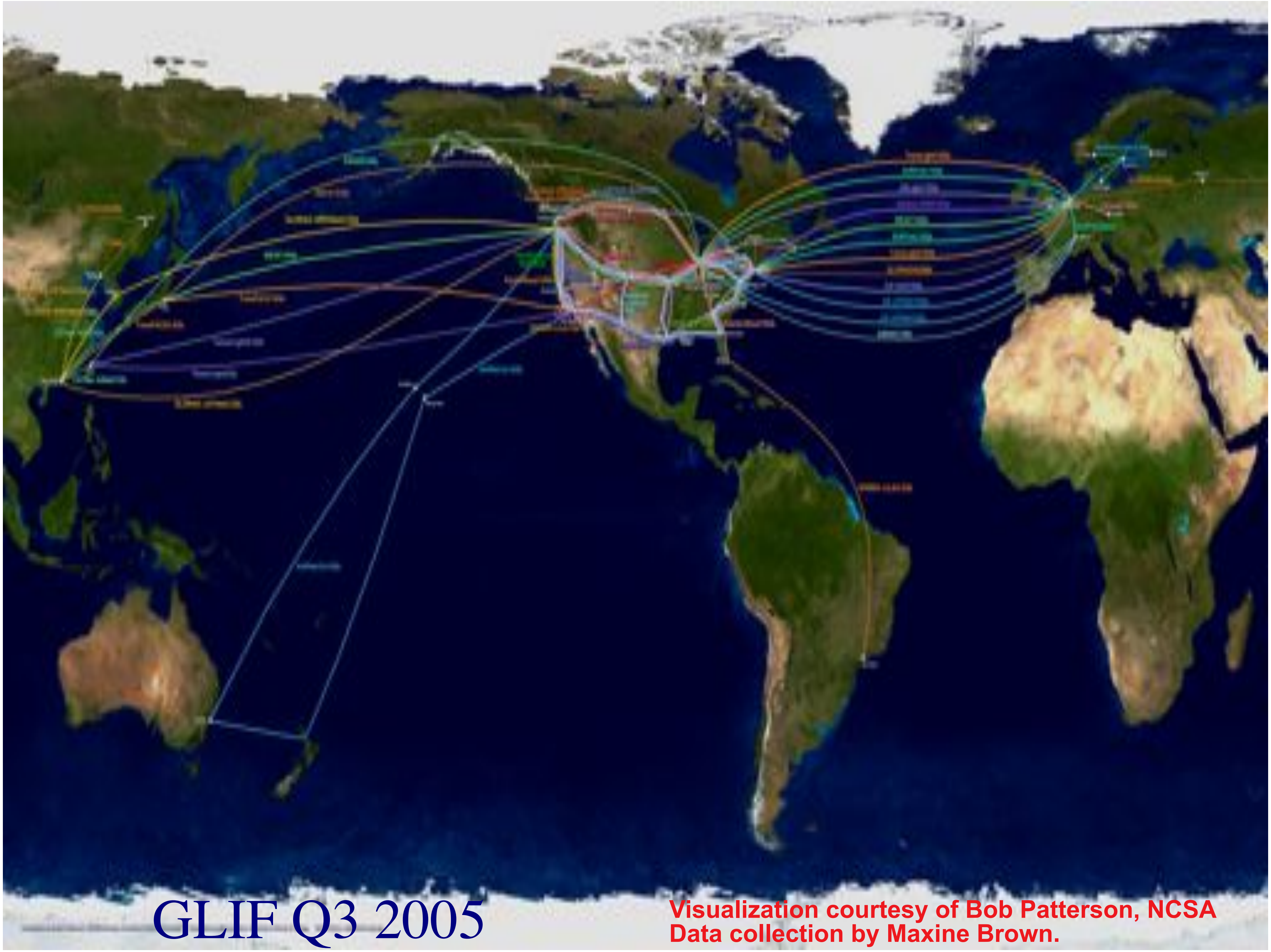
7a. WS passes the key to the user.

7b. AAA server interacts with PEP to update the local POLICY_TABLE on the PEP.

8. User constructs the RSVP message with extra Token data by using the key and sends to VLSR-1.
9. VLSR-1 queries PEP whether the Token in the RSVP message is valid.
10. PEP checks in the local POLICY_TABLE and return YES.
11. When VLSR-1 receives YES from PEP, it forwards the RSVP message.
12. All nodes process RSVP message(forwarding/response)
13. The Ethernet switches are configured
14. LSP is set up and traffic can flow

# Contents

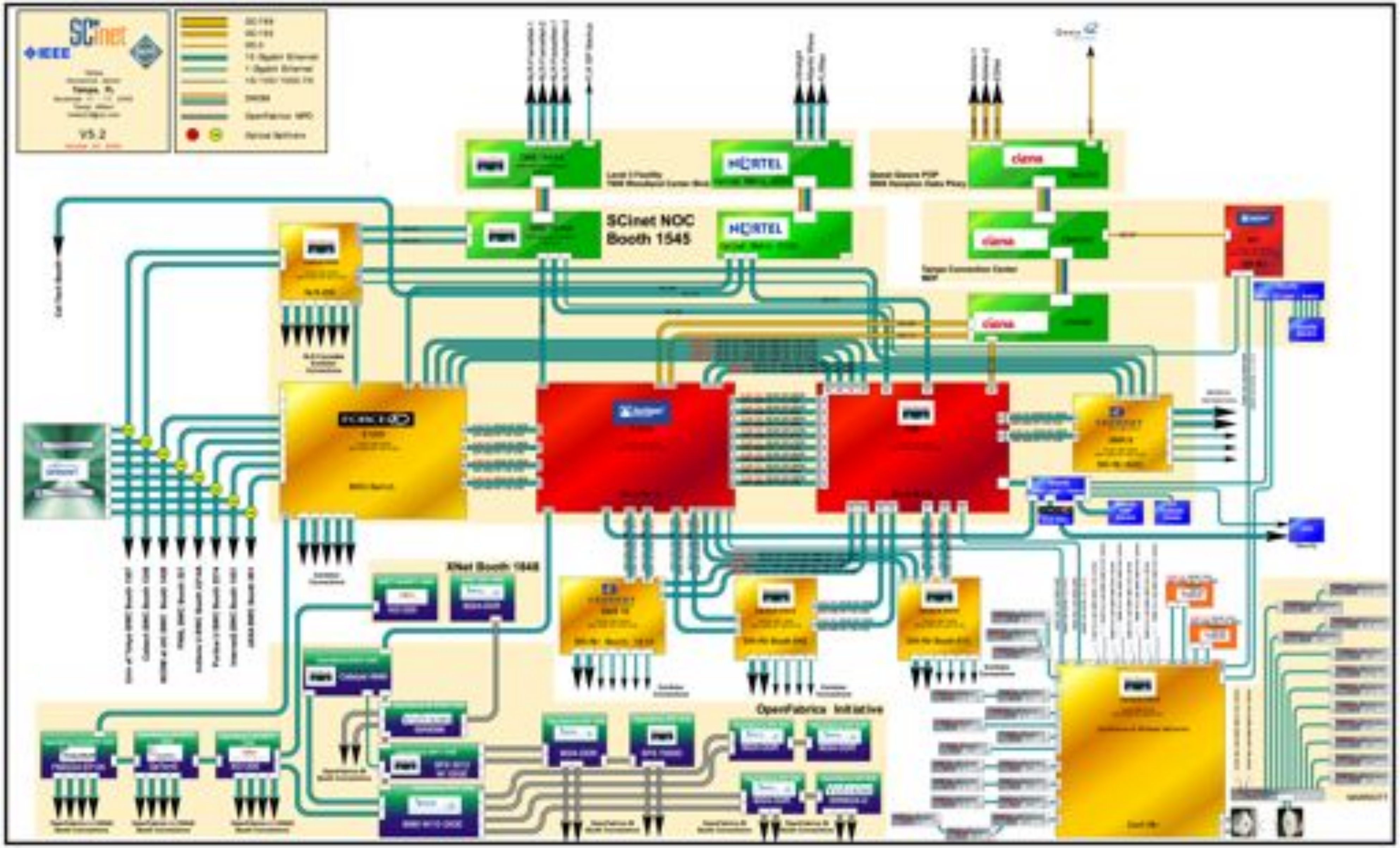
1. The need for hybrid networking
2. StarPlane; a grid controlled photonic network
3. Cross Domain Authorization using Tokens
4. RDF/Network Description Language
5. Tera-networking
6. Programmable networks



GLIF Q3 2005

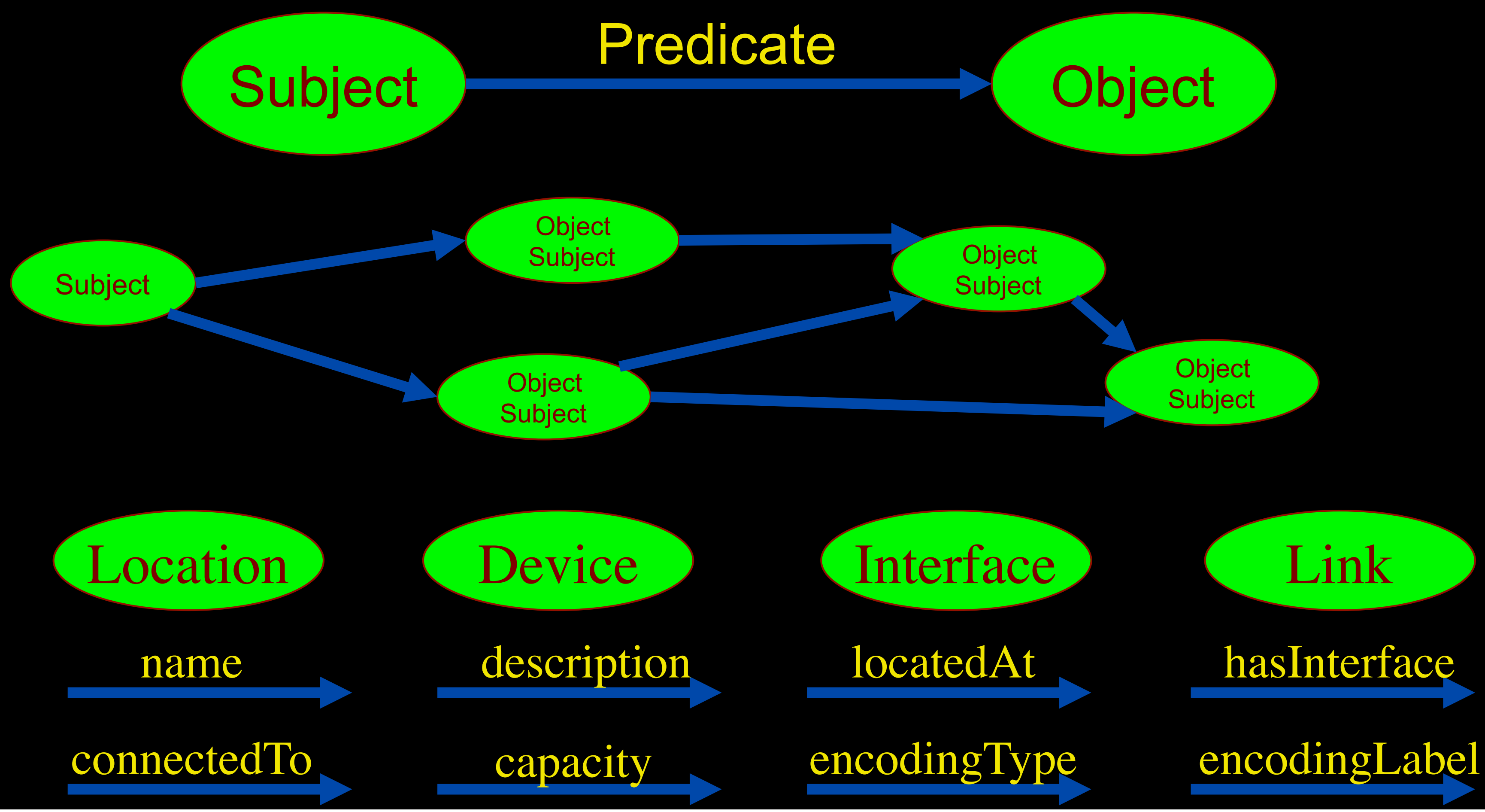
**Visualization courtesy of Bob Patterson, NCSA**
**Data collection by Maxine Brown.**

# Architecture SC06



# Network Description Language

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

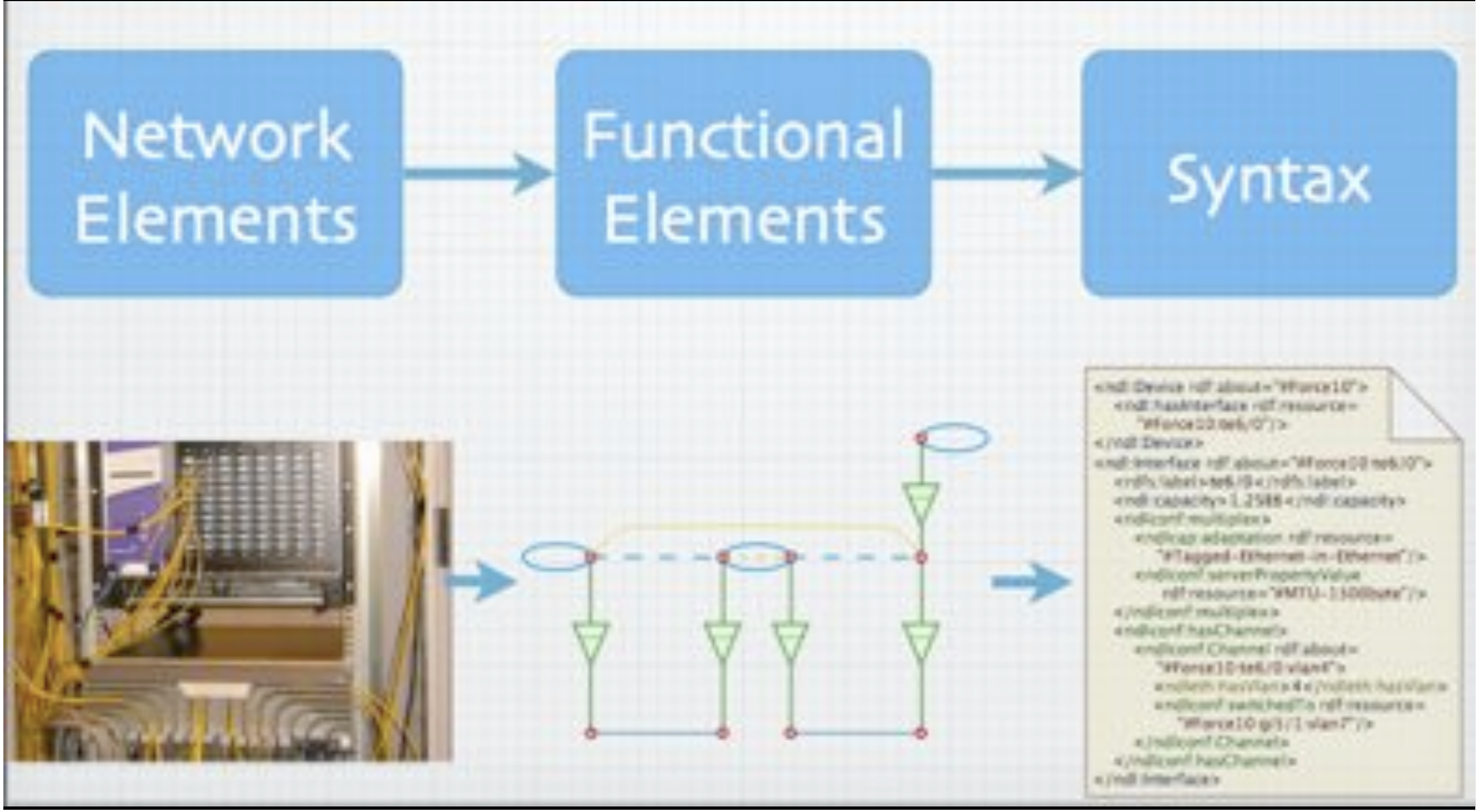


# Network Description Language

Choice of RDF instead of flat XML descriptions

Grounded modeling based on G0805 description:

Article: F. Dijkstra, B. Andree, K. Koymans, J. van der Ham, P. Grosso, C. de Laat, *"A Multi-Layer Network Model Based on ITU-T G.805"*

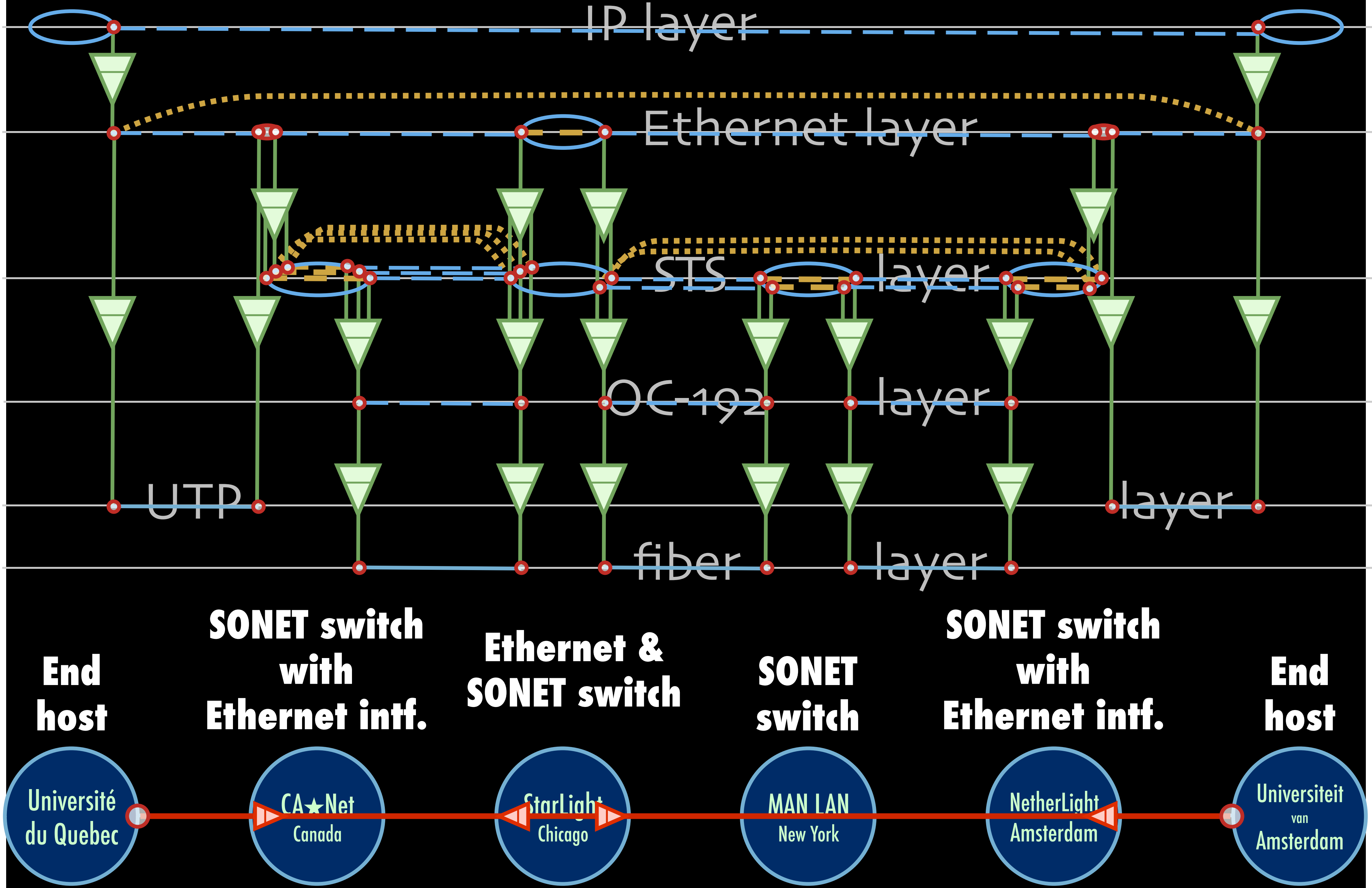


# NetherLight in RDF

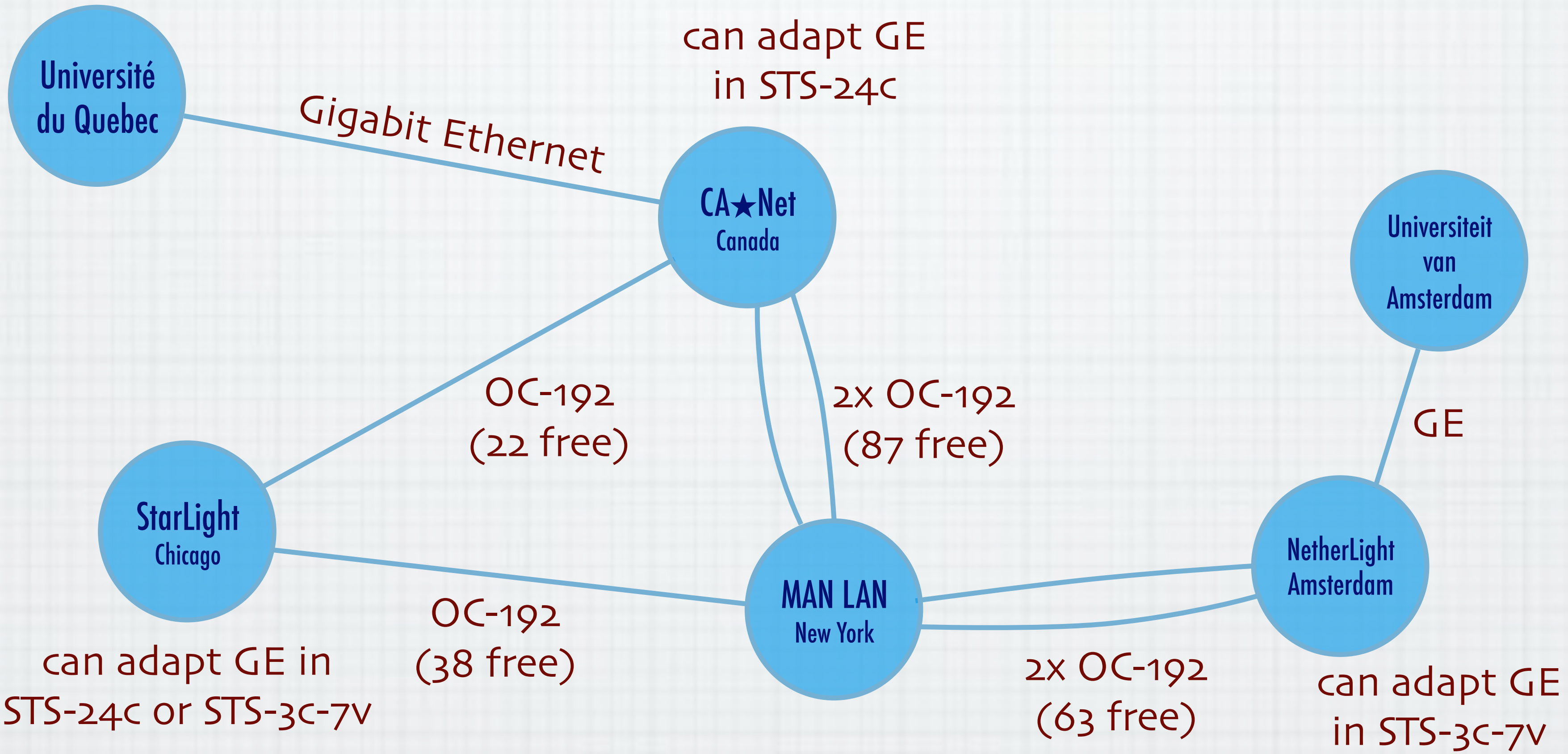
```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:ndl="http://www.science.uva.nl/research/air/ndl#">
<!-- Description of Netherlight -->
<ndl:Location rdf:about="#Netherlight">
      <ndl:name>Netherlight Optical Exchange</ndl:name>
</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:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
```

```
<!-- 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>
```

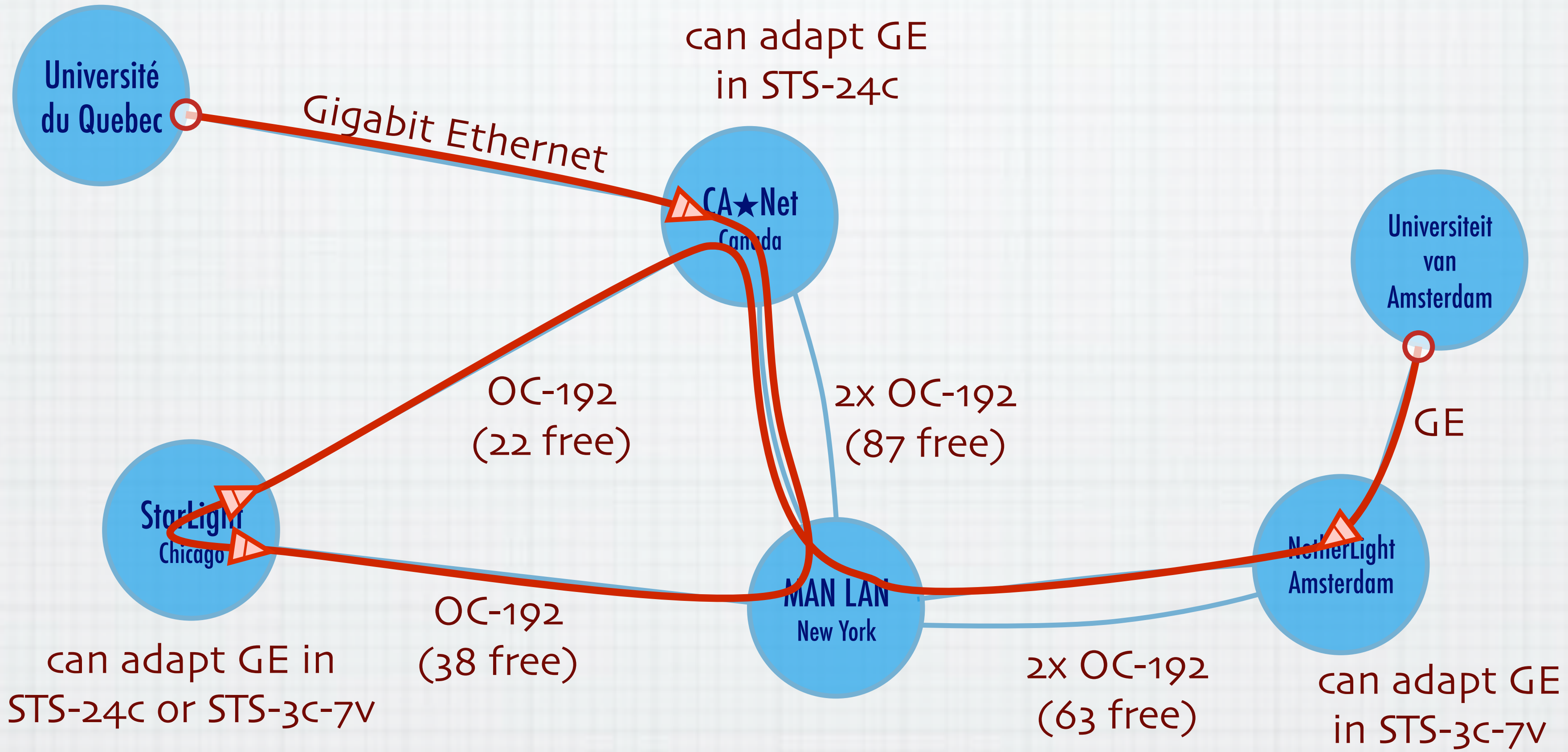
# Multi-layer descriptions in NDL



# A weird example

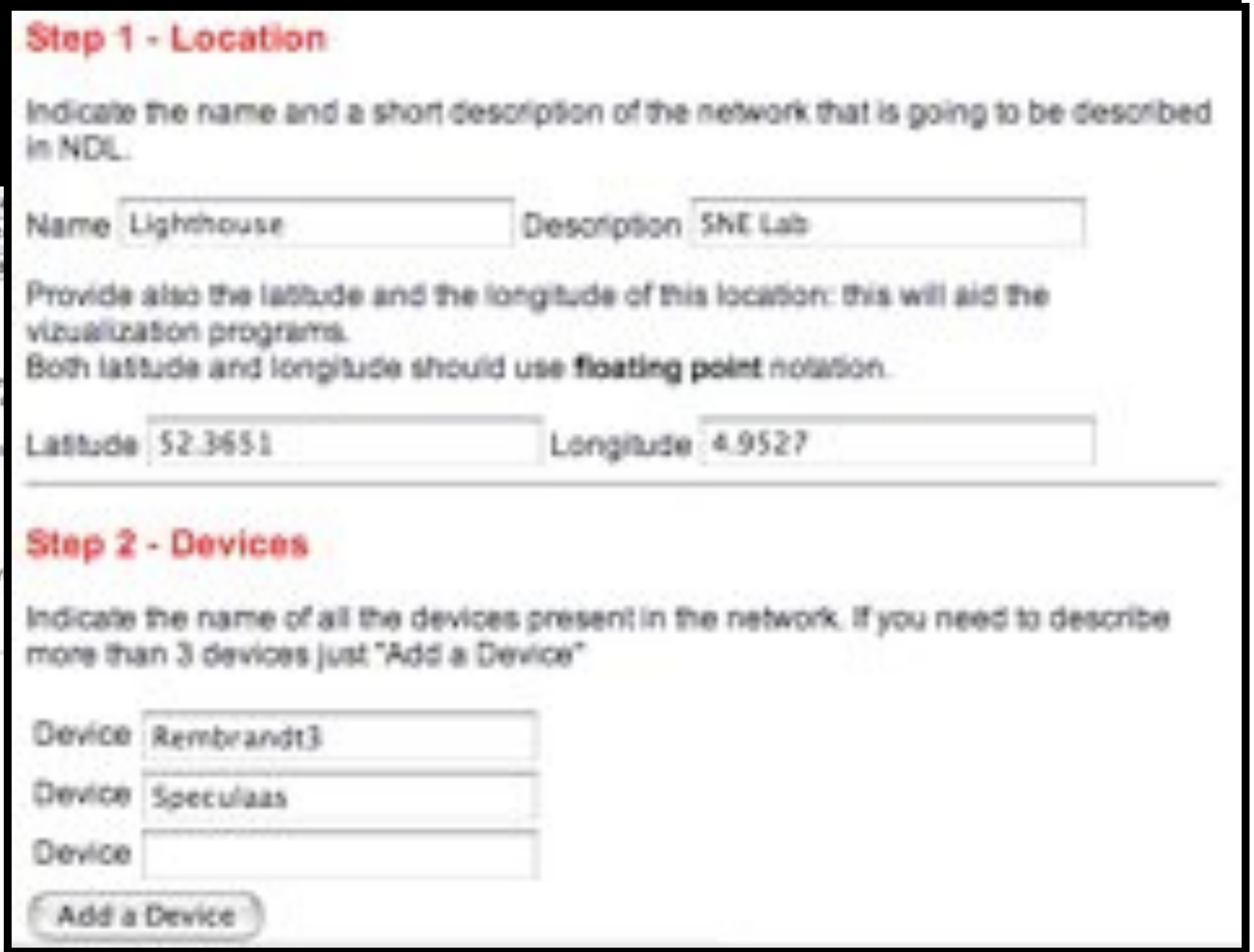


# The result :-)



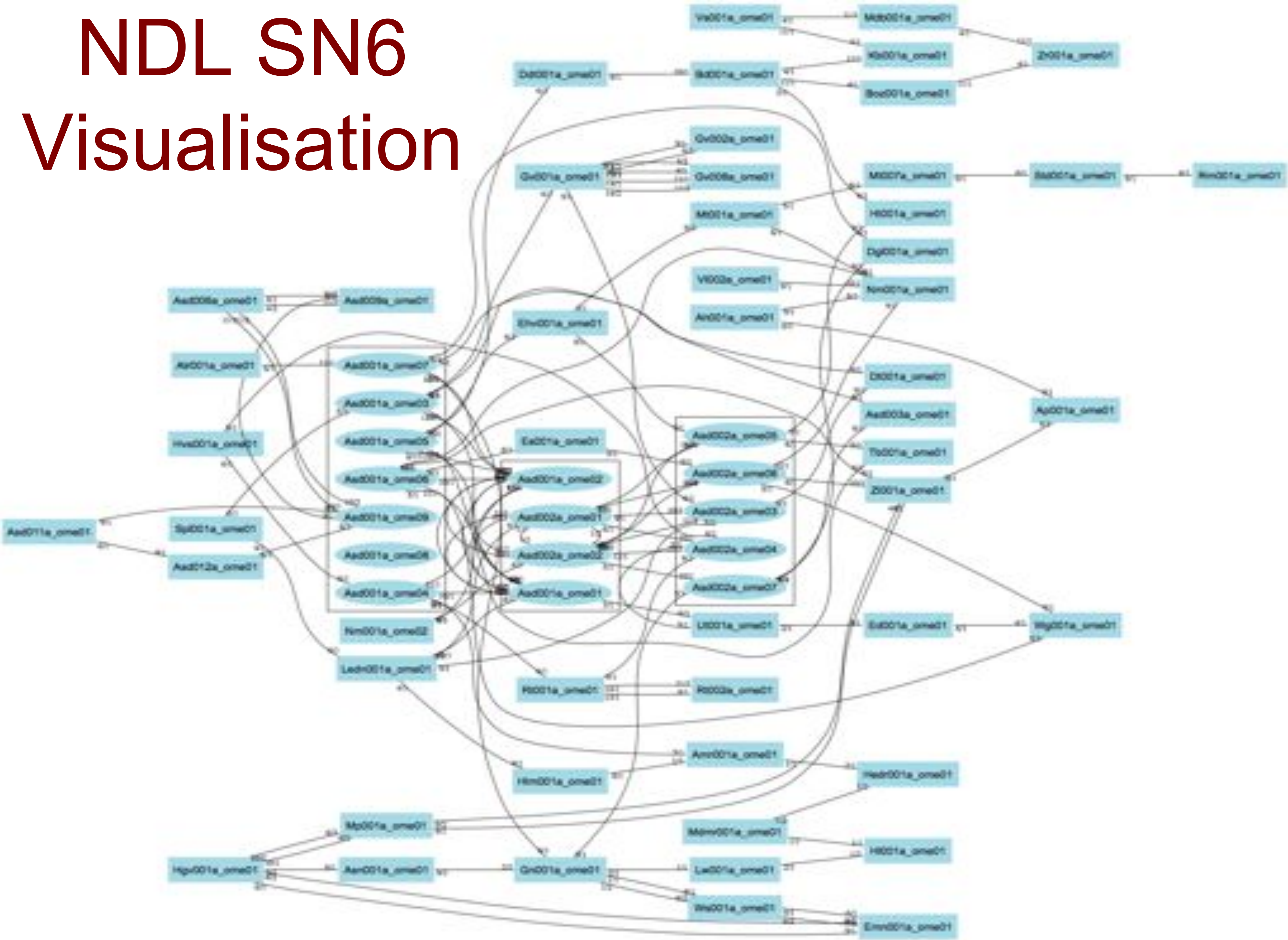

Thanks to Freek Dijkstra & team

# NDL Generator and Validator


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

# NDL SN6 Visualisation

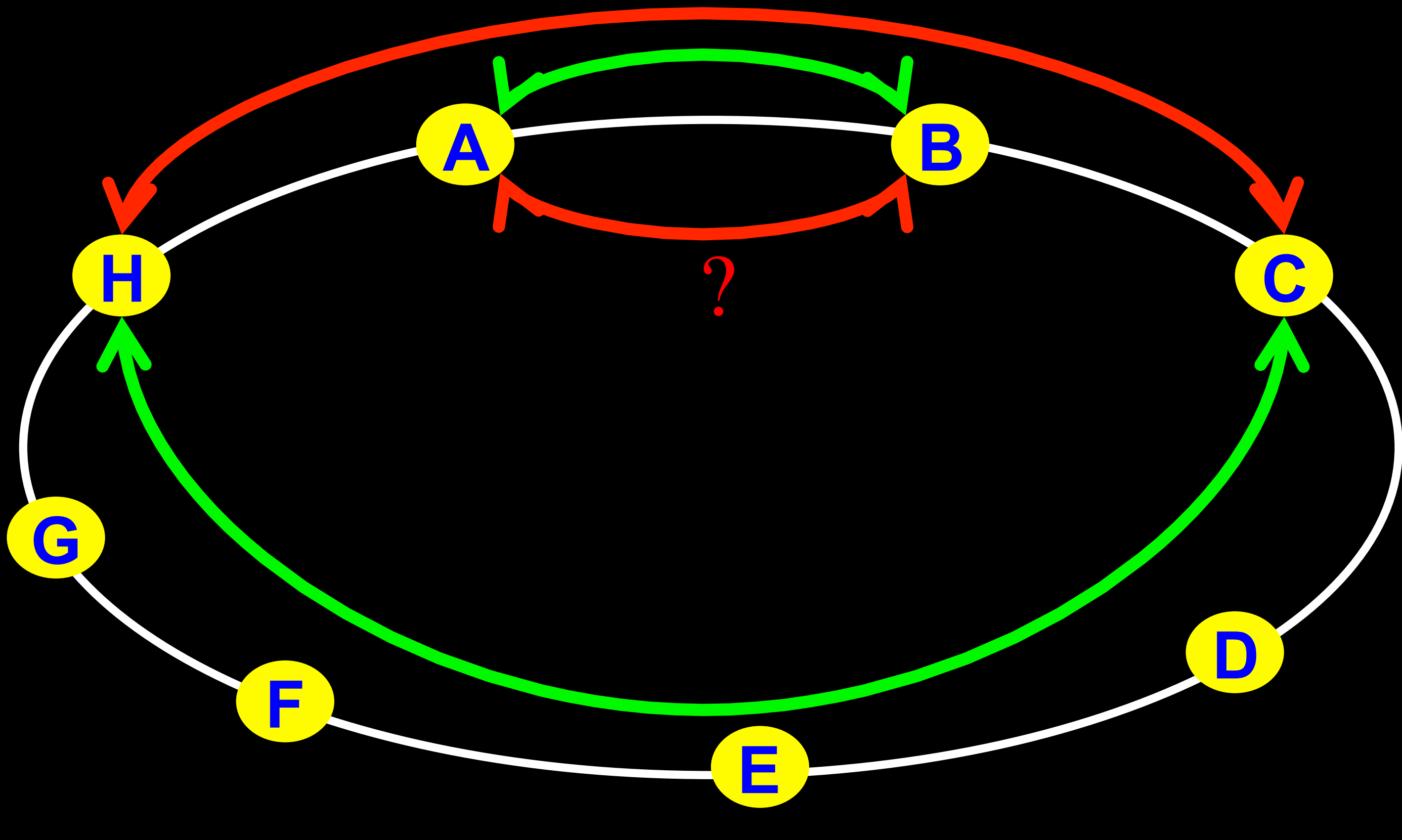


# The Problem

I want HC and AB

Success depends on the order

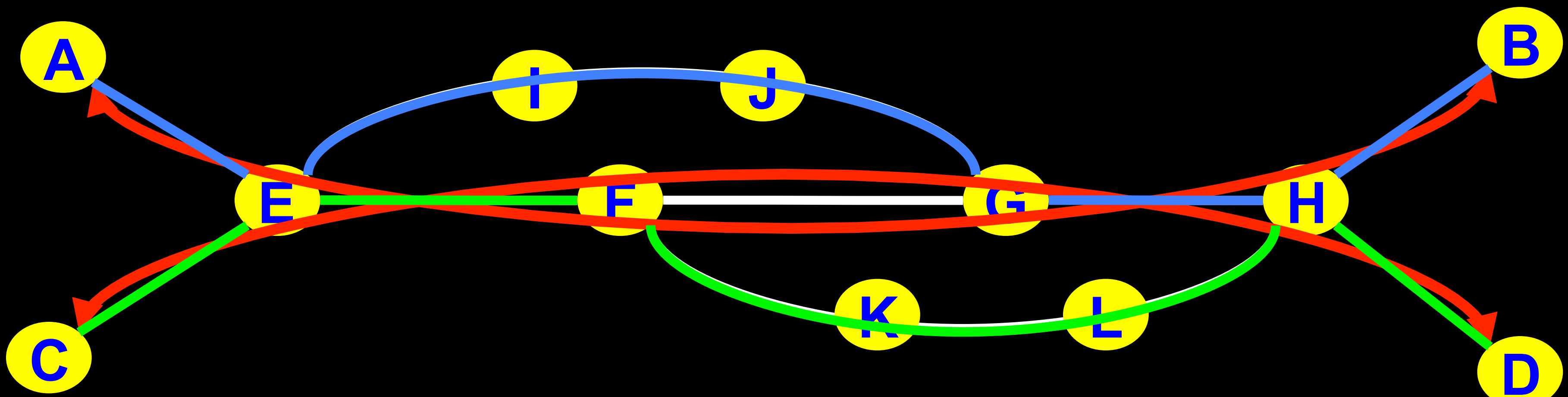
Wouldn't it be nice if I could request [HC, AB, ...]



# Another one ☺

## I want AB and CD

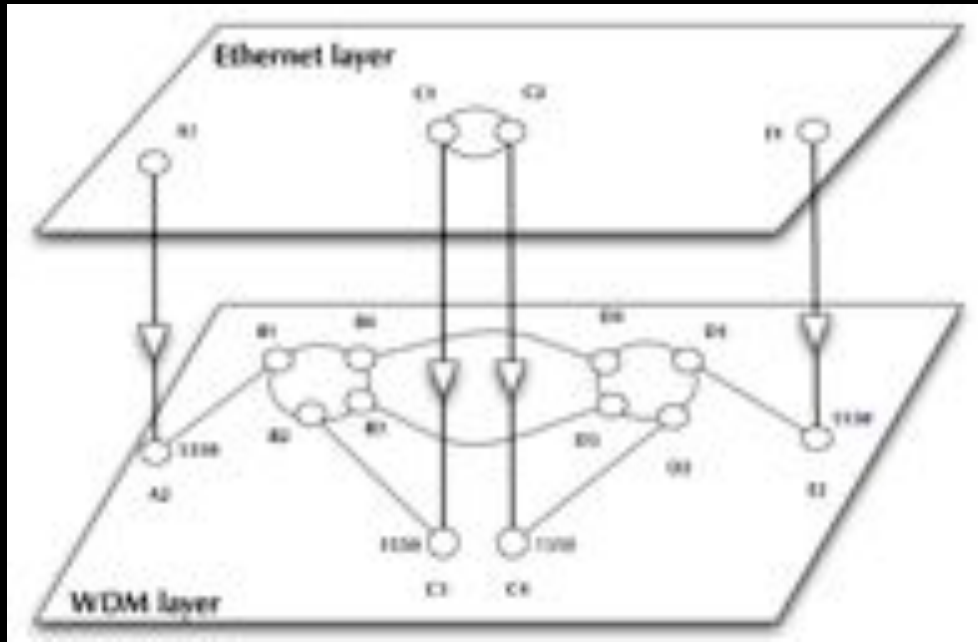
Success does not even depend on the order!!!

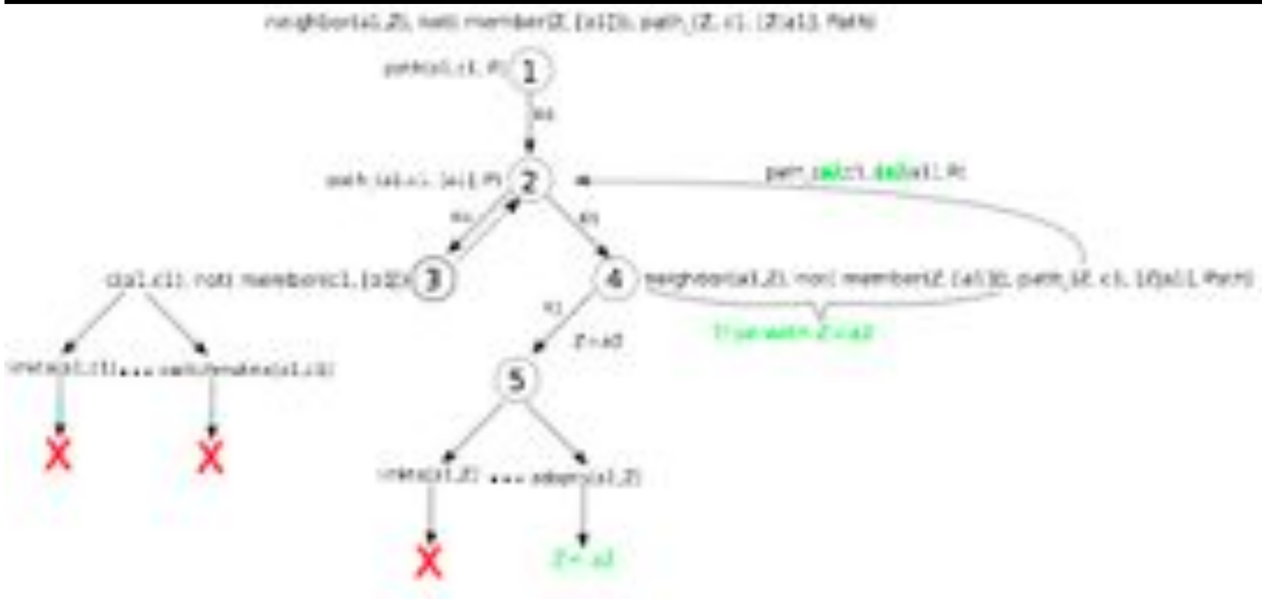


# NDL + PROLOG

Research Questions:

- order of requests
- complex requests
- usable leftovers



- Reason about graphs
- Find sub-graphs that comply with rules

# Multi layer multi domain networks

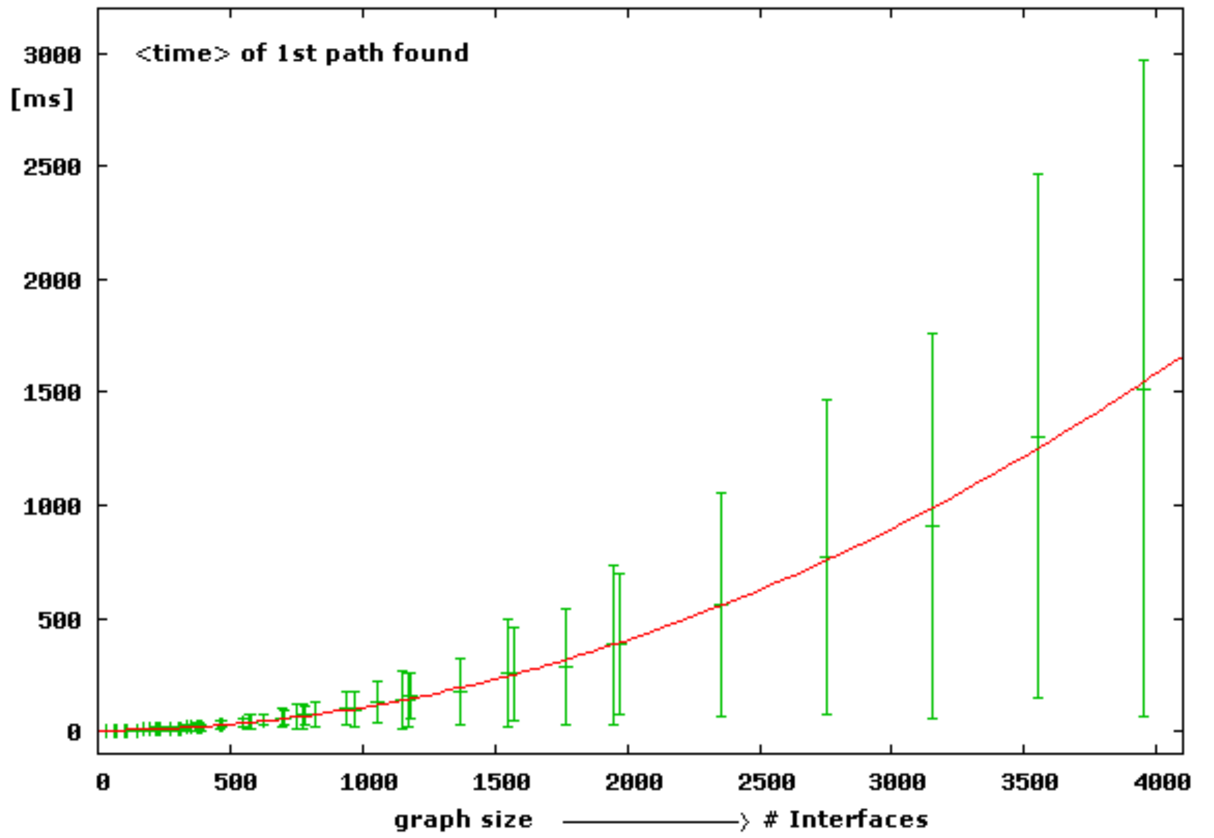
The networks for e-Science where applications use dedicated optical circuits.

*Is declarative programming more suitable to find paths in multi-domain multi-layer networks? Especially in presence of constraints and complex requests?*

Our approach:

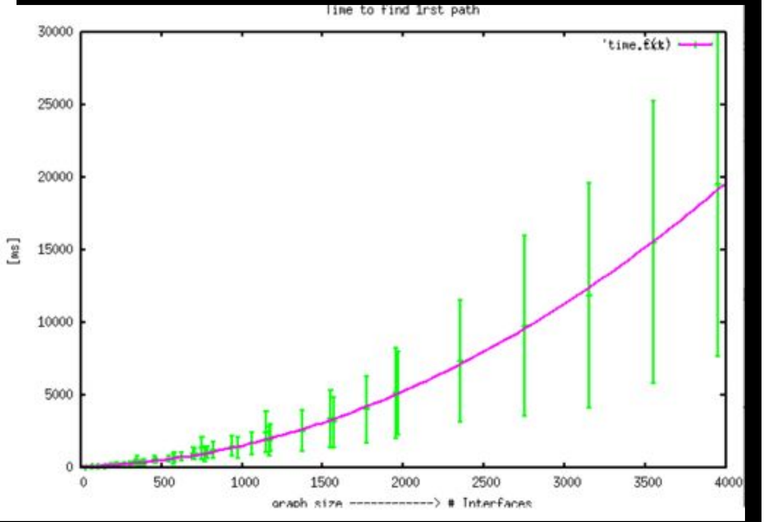
1. We generate BA network graphs with a varying number of domains and nodes. Barabasi-Albert scale free graphs are a good representation of these networks.
2. We represent the graphs in NDL – Network Description Language, the RDF schemas.
3. We load the RDF files in Prolog and Python programs
4. We perform a modified DFS –Depth First Search- algorithm to find paths.

# Single layer networks: results




- *Number of interfaces,*
- *given N nodes per domain D*
- *4*(D-2) + D*4*(N-2) for D > 2*

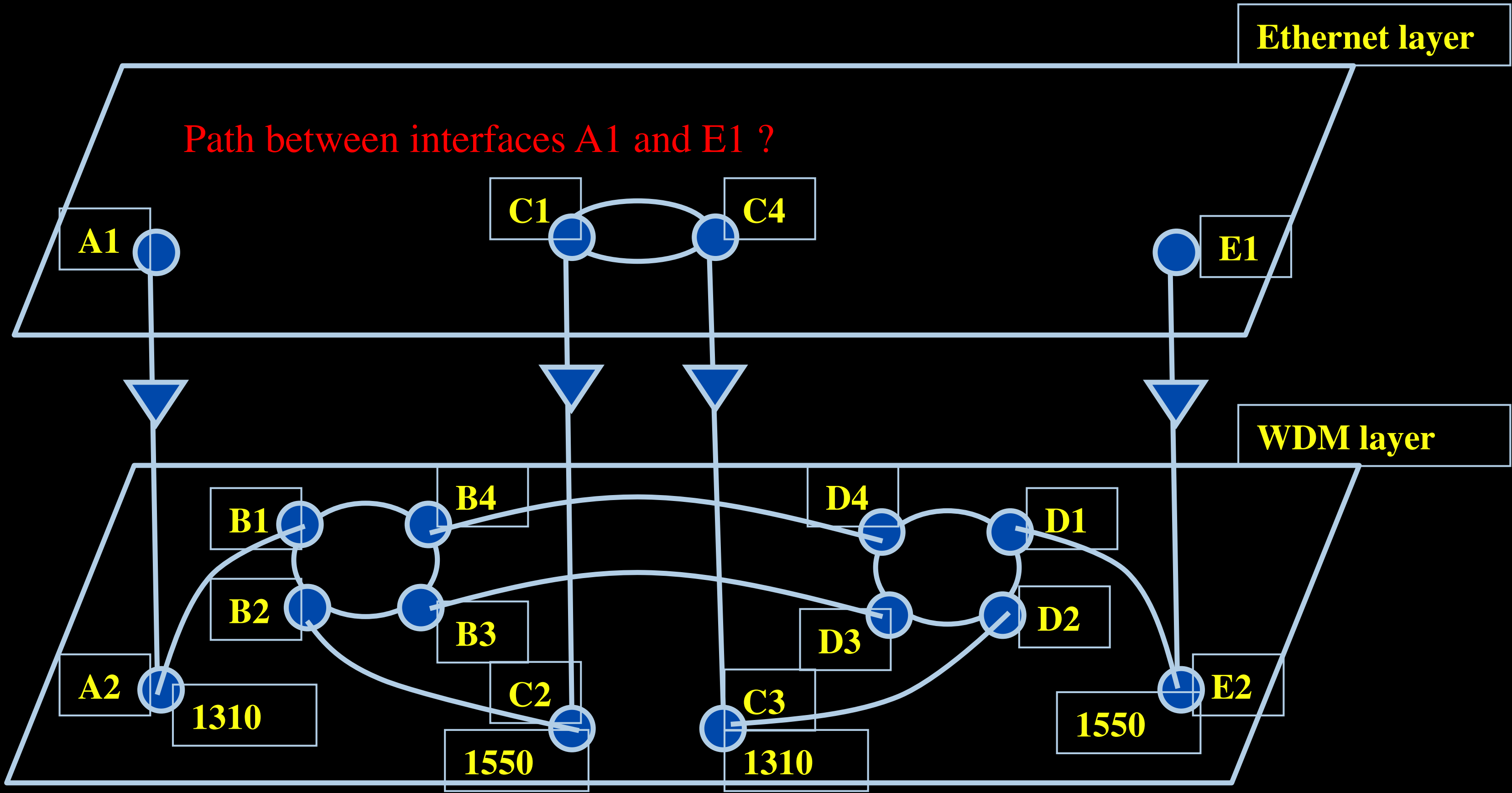
## Pynt-based DFS




## Prolog DFS

- Prolog time to find first path shorter than Python time.
- We observe a quadratic dependence.
- Length of paths found comparable.

# Multi-layer network




**Prolog rule:**

```
linkedto( Intf1, Intf2, CurrWav ):-                %-- is there a link between  Intf1 and Intf2 for wavelength CurrWav ?
   rdf_db:rdf( Intf1, ndl:'layer', Layer ),        %-- get layer of interface Intf1 → Layer
   Layer == 'wdm#LambdaNetworkElement',            %-- are we at the WDM-layer ?
   rdf_db:rdf( Intf1, ndl:'linkedTo', Intf2 ),     %-- is Intf1 linked to Intf2 in the RDF file?
   rdf_db:rdf( Intf2, wdm:'wavelength', W2 ),      %-- get wavelength of Intf2 → W2
   compatible_wavelengths( CurrWav, W2 ).          %-- is CurrWav compatible with W2 ?
```

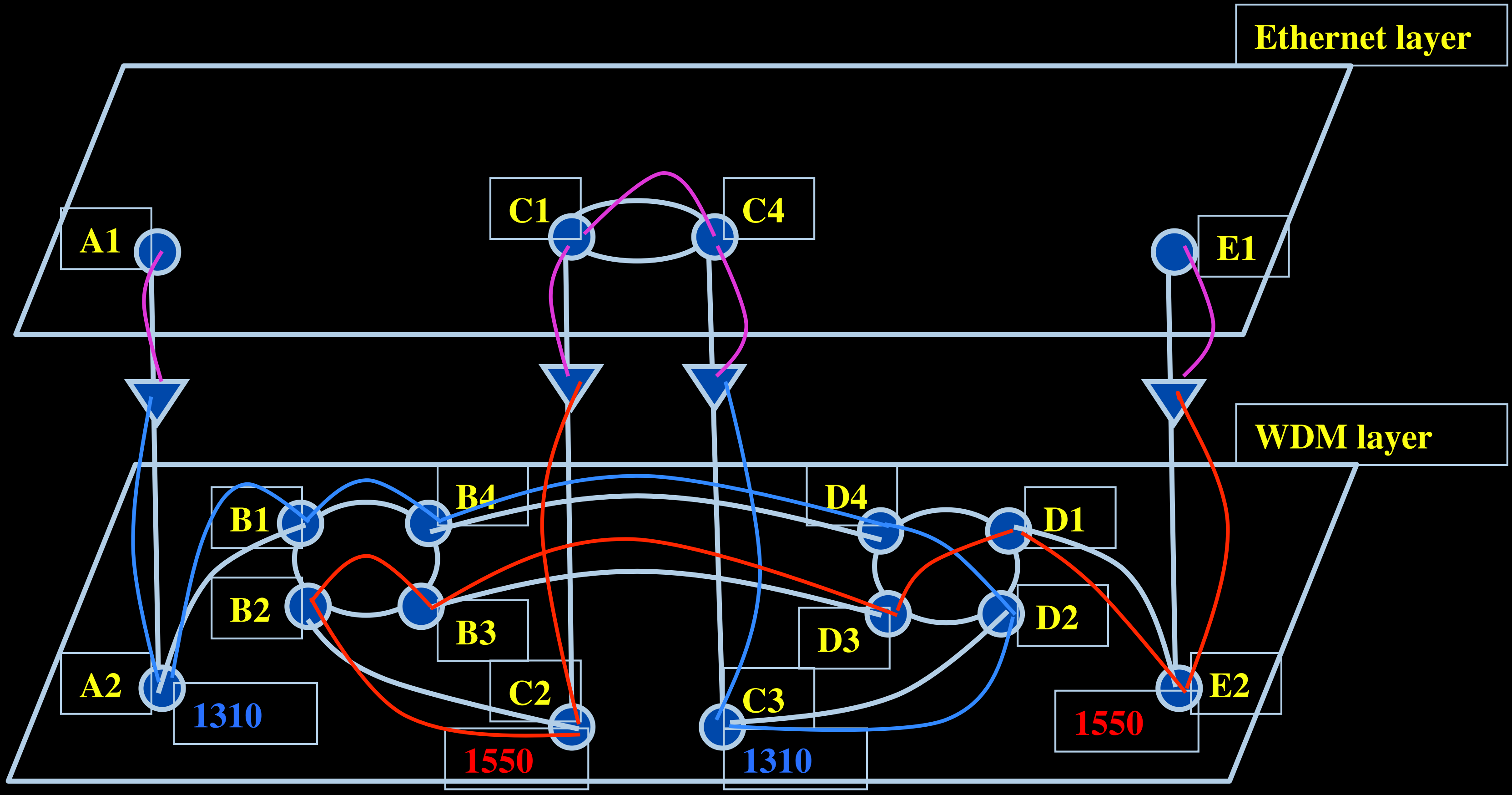
**linkedto( B4, D4, CurrWav ) is true for any value of CurrWav**

**linkedto( D2, C3, CurrWav) is true if CurrWav == 1310**

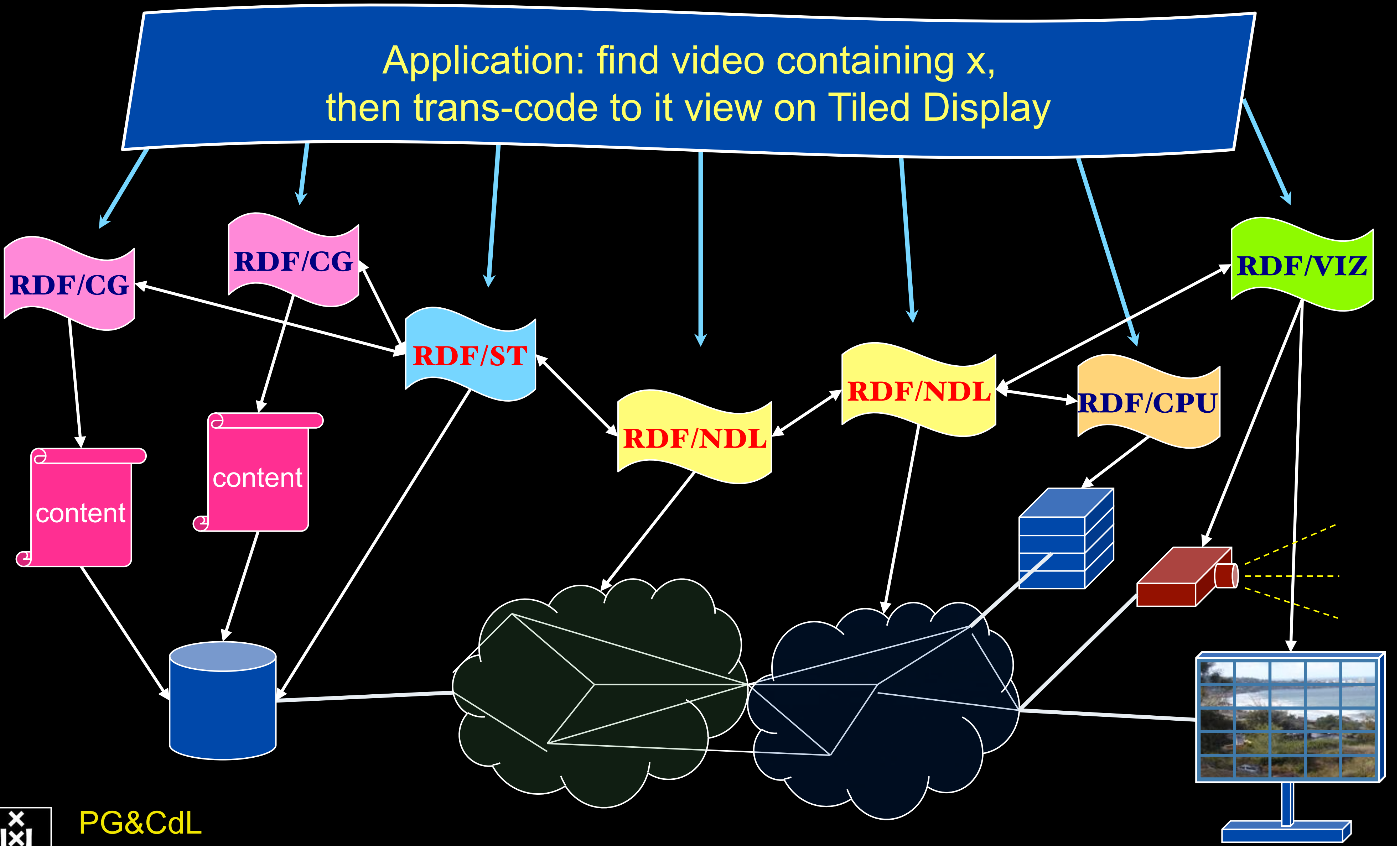
# Multi-layer

Path between interfaces A1 and E1:

A1-A2-B1-B4-D4-D2-C3-C4-C1-C2-B2-B3-D3-D1-E2-E1

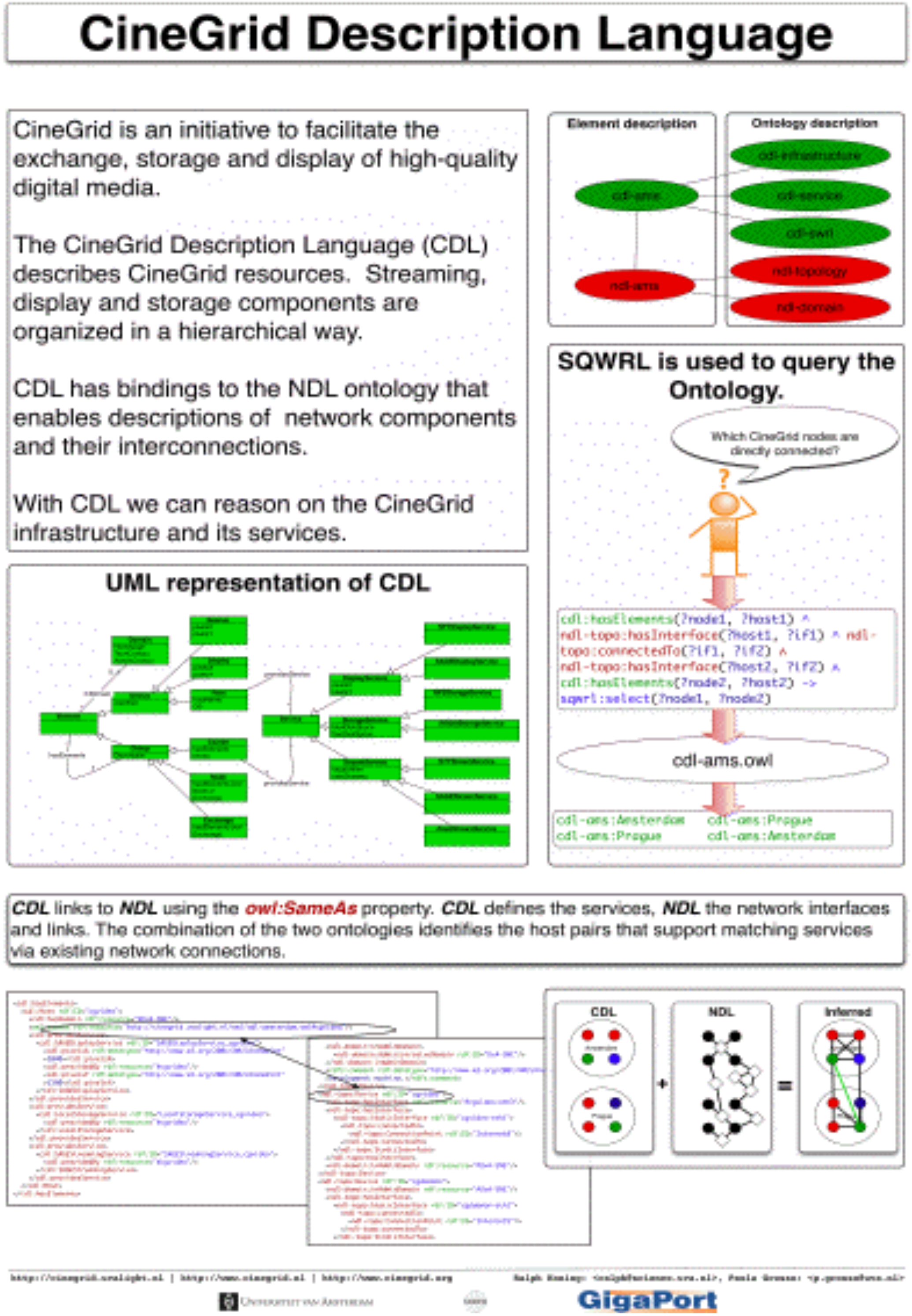


# RDF describing Infrastructure




PG&CdL

Applications and Networks become aware of each other!



# Contents

1. The need for hybrid networking
2. StarPlane; a grid controlled photonic network
3. Cross Domain Authorization using Tokens
4. RDF/Network Description Language
5. Tera-networking
6. 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

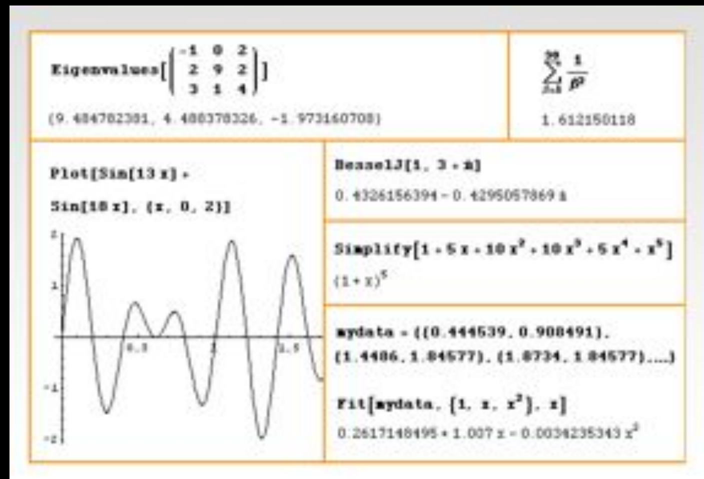
- it takes a core to receive 1 or 10 Gbit/s in a computer
- it takes one or two cores to deal with 10 Gbit/s storage
- same for Gigapixels
- same for 100's of Gflops
- Capacity of every part in a system seems of same scale
- look at 80 core Intel processor
  - cut it in two, left and right communicate 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!

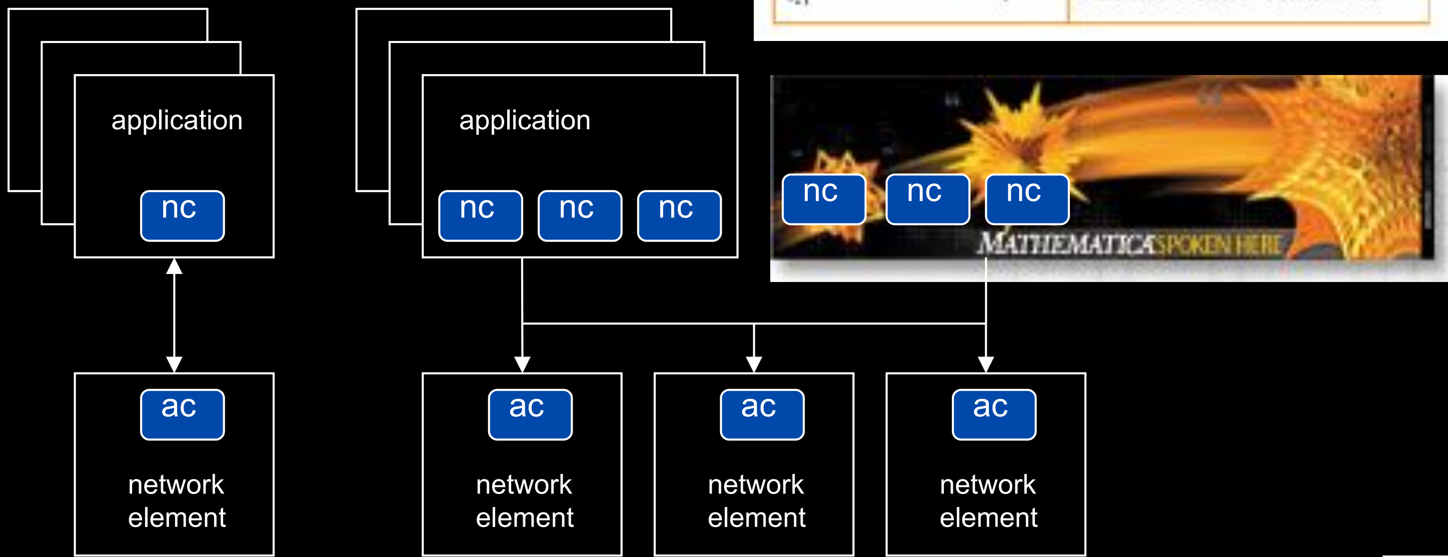
# Contents

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

# 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



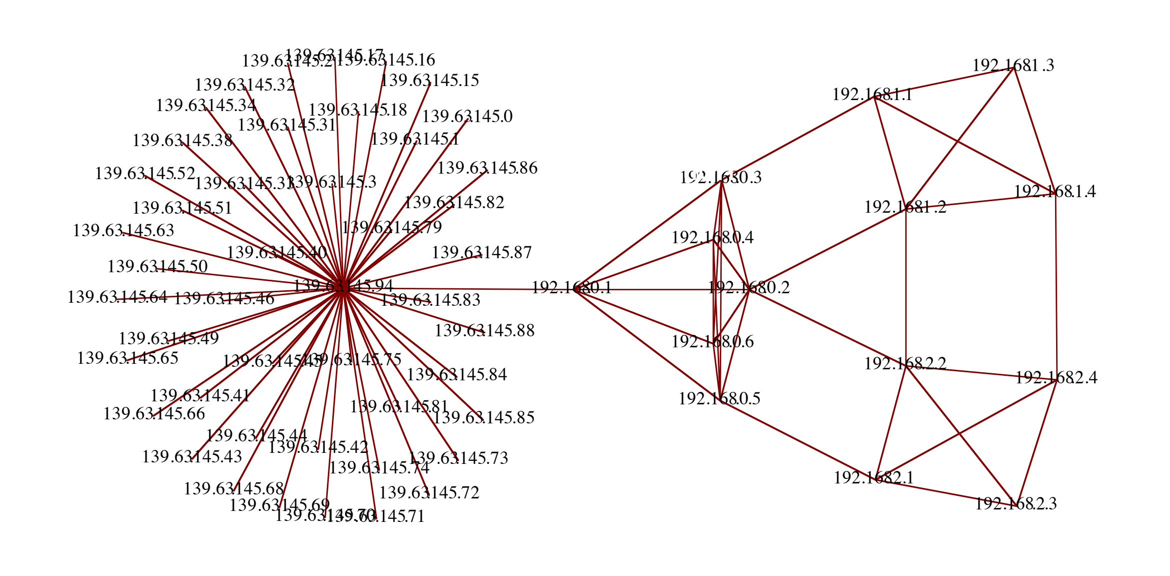


# Mathematica enables advanced graph queries, visualizations and real-time network manipulations on UPVNs

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

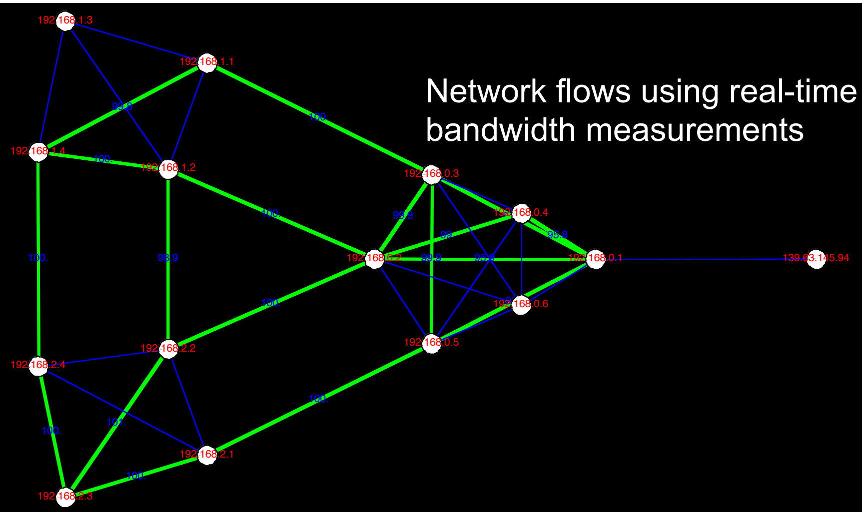
### Initialization and BFS discovery of NEs

```
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
```




### Transaction on shortest path with tokens

```
nodePath = ConvertIndicesToNodes[
Internal links: {192.168.2.3}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

# TouchTable Demonstration @ SC08

# Interactive programmable networks

*Questions ?*