

Networking in the Research and Education World

Cees de Laat

EU

SURFnet

BSIK

NWO

University of Amsterdam



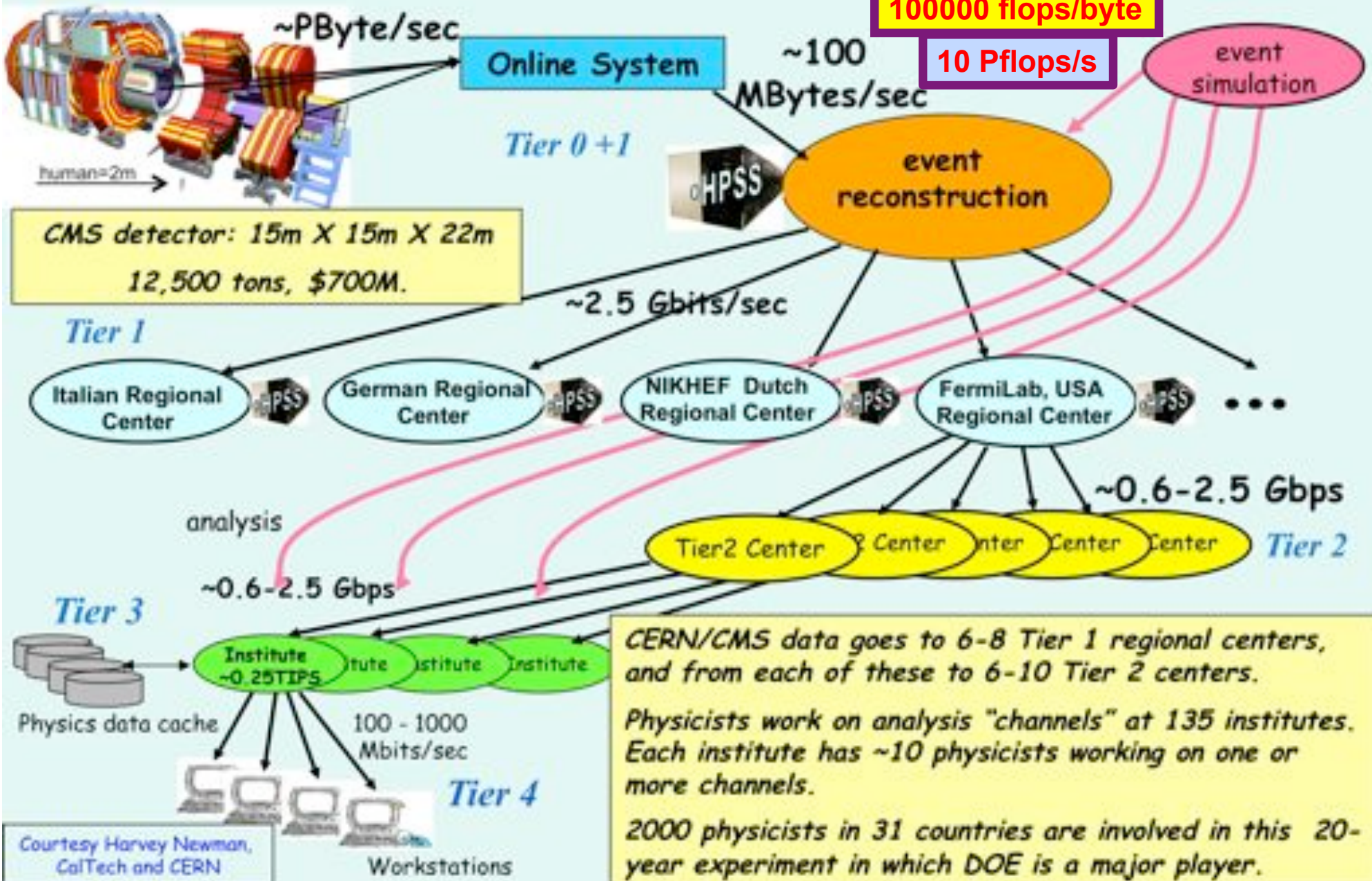
TNO
NCF





LHC Data Grid Hierarchy

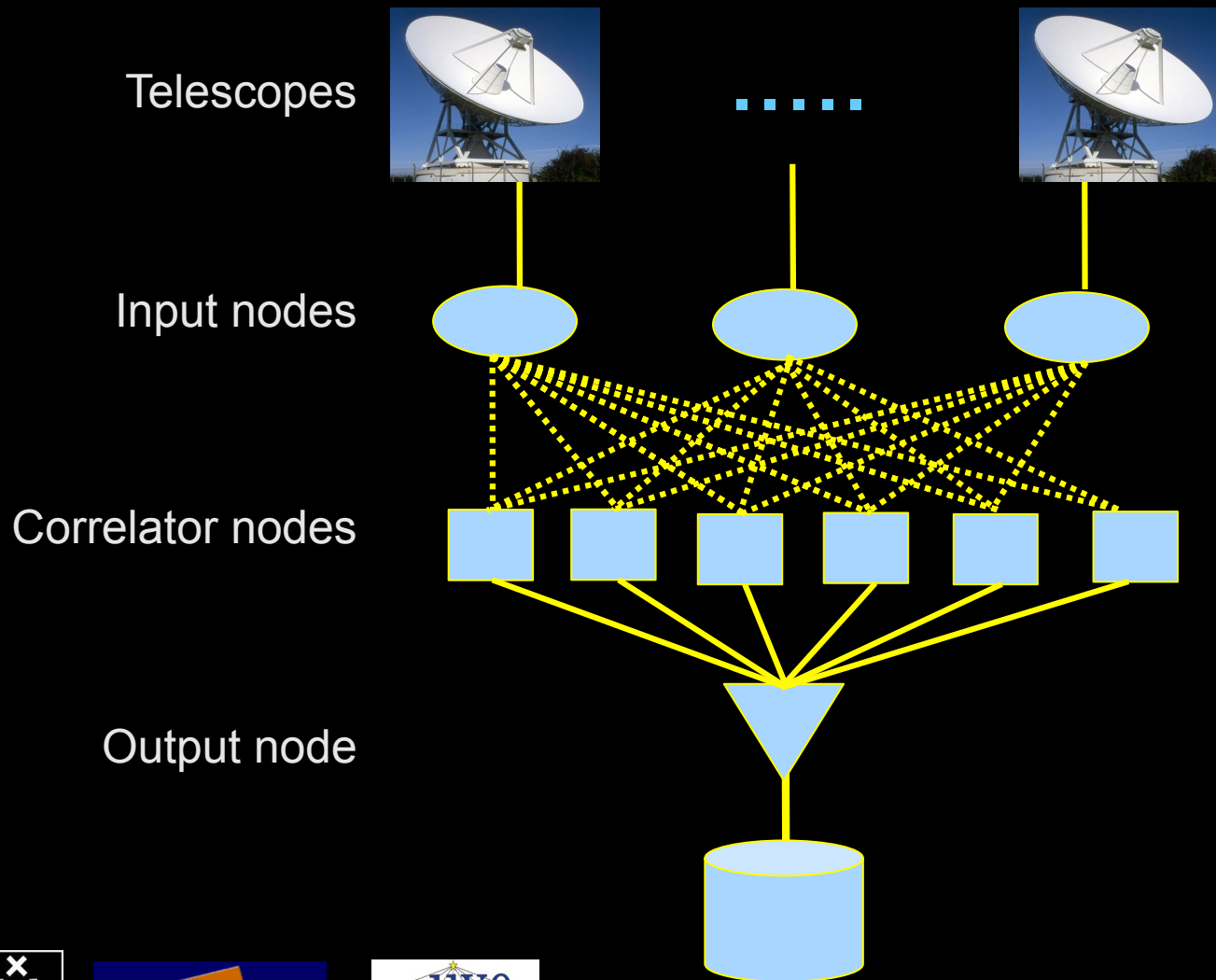
CMS as example, Atlas is similar



Courtesy Harvey Newman, CalTech and CERN

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

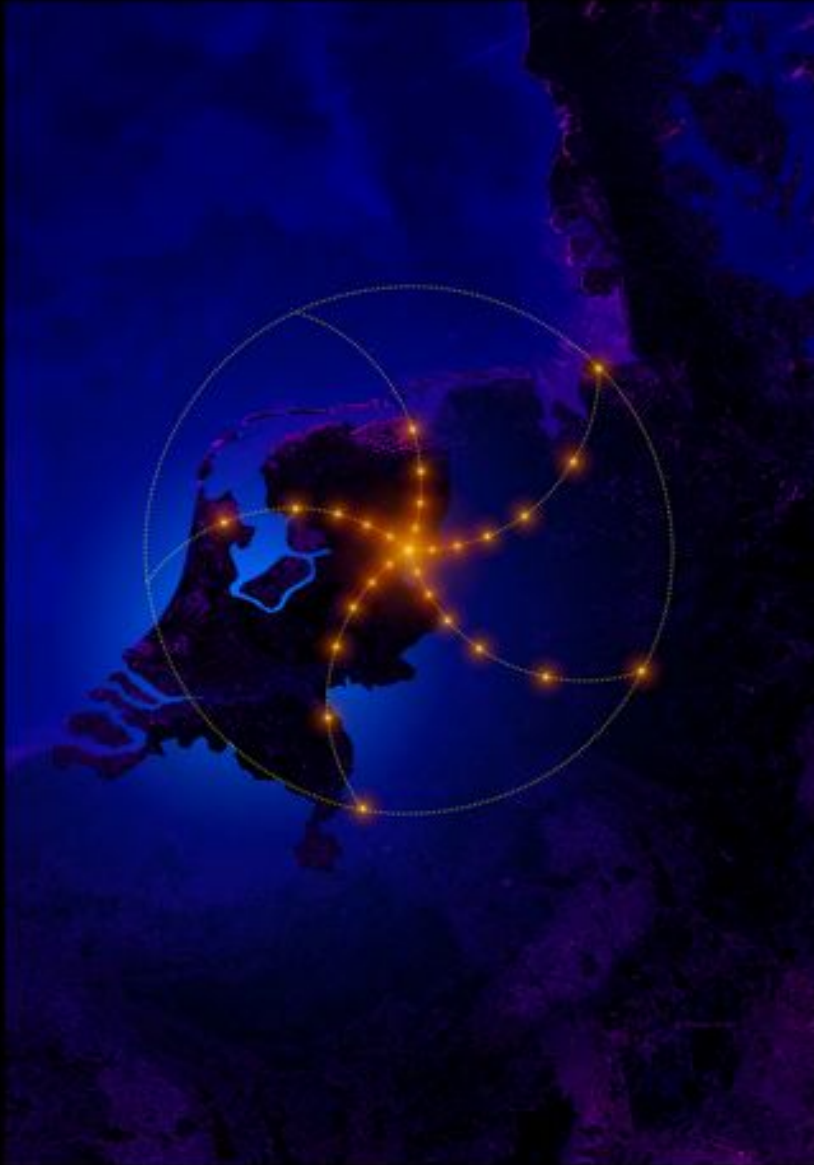
0.1 Pflops/s

**THIS IS A DATA FLOW
PROBLEM !!!**



LOFAR as a Sensor Network

20 flops/byte



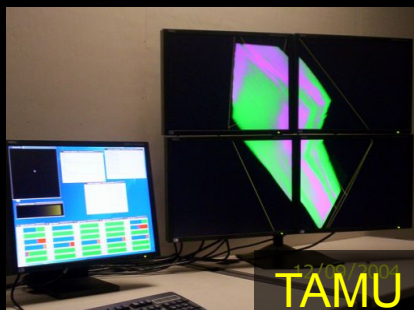
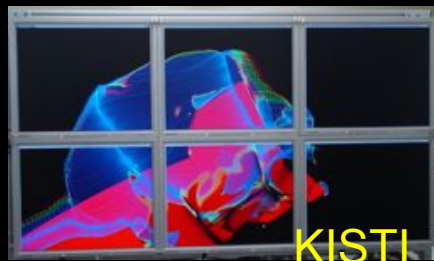
– LOFAR is a large distributed research infrastructure:

2 Tflops/s

- 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



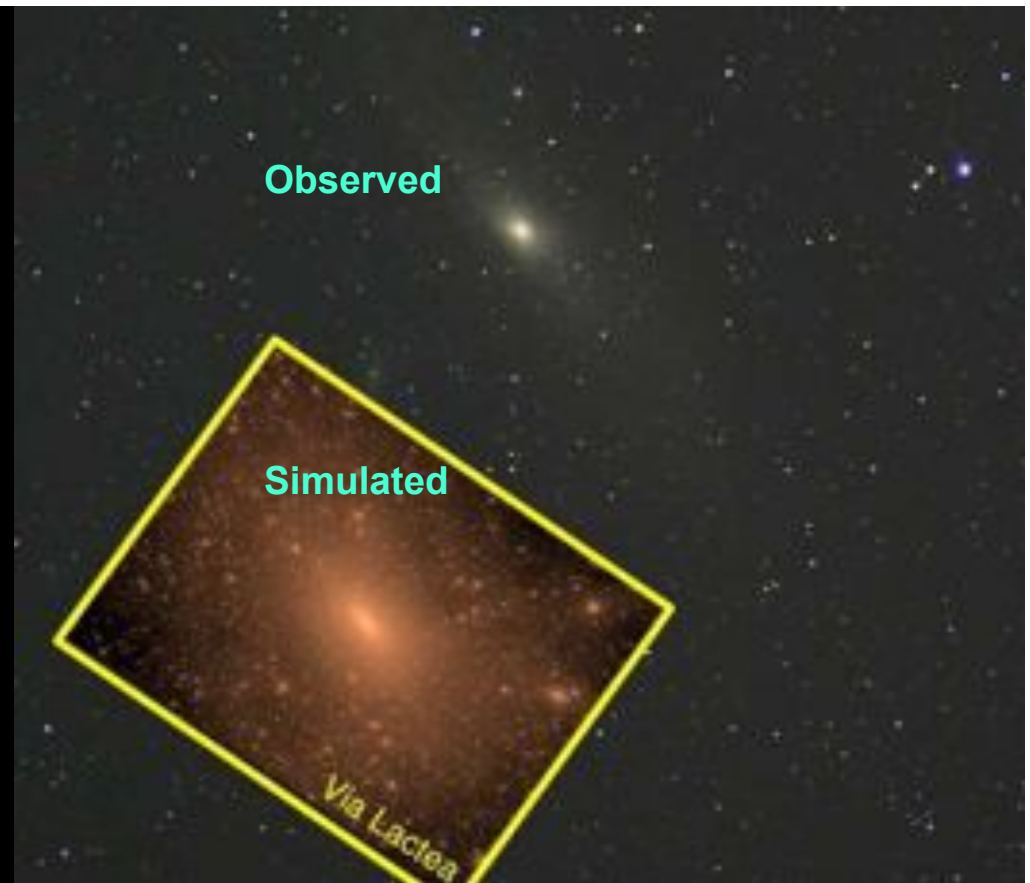
Sensor grid: instrument the dikes

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

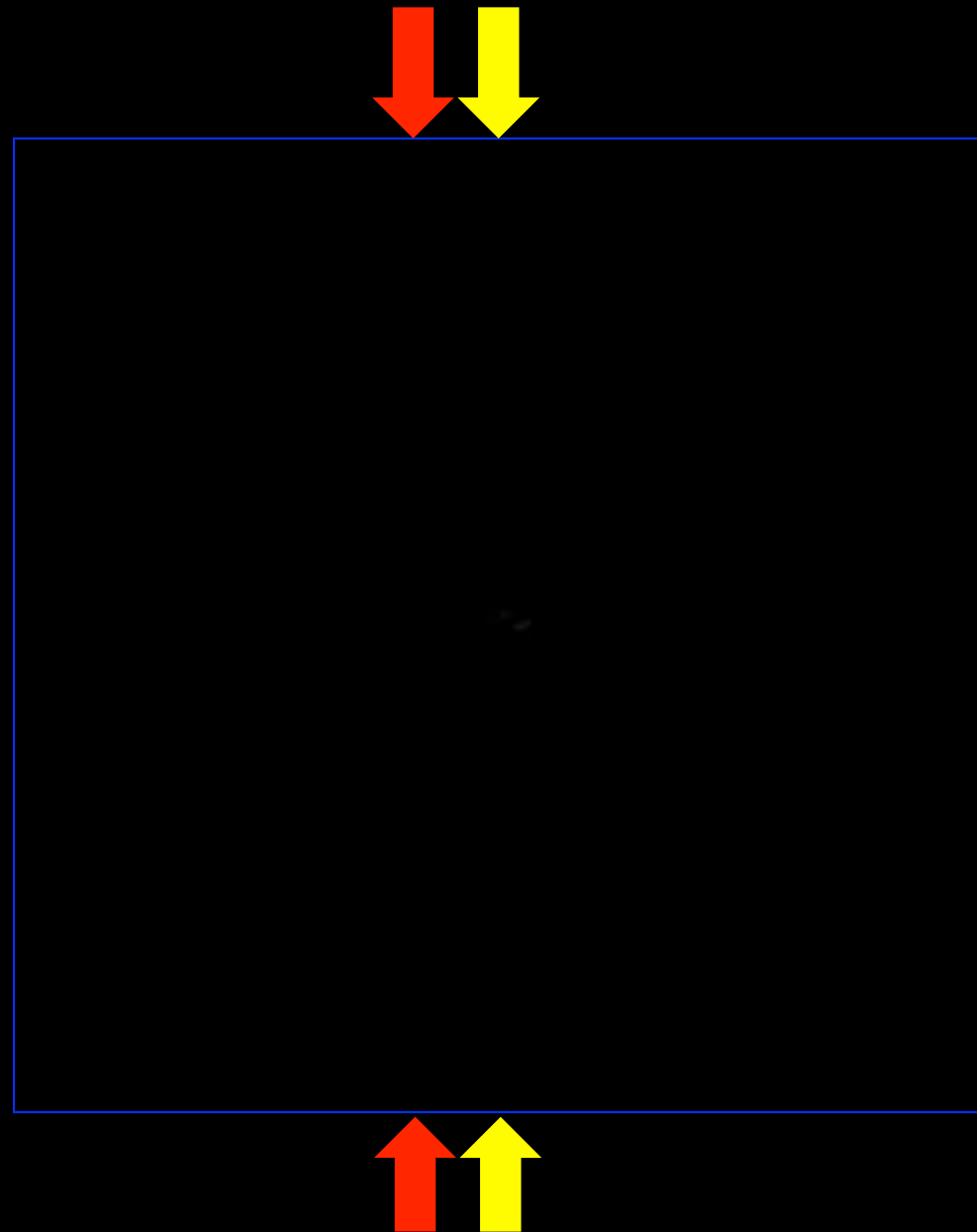
10 Mflops/byte

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

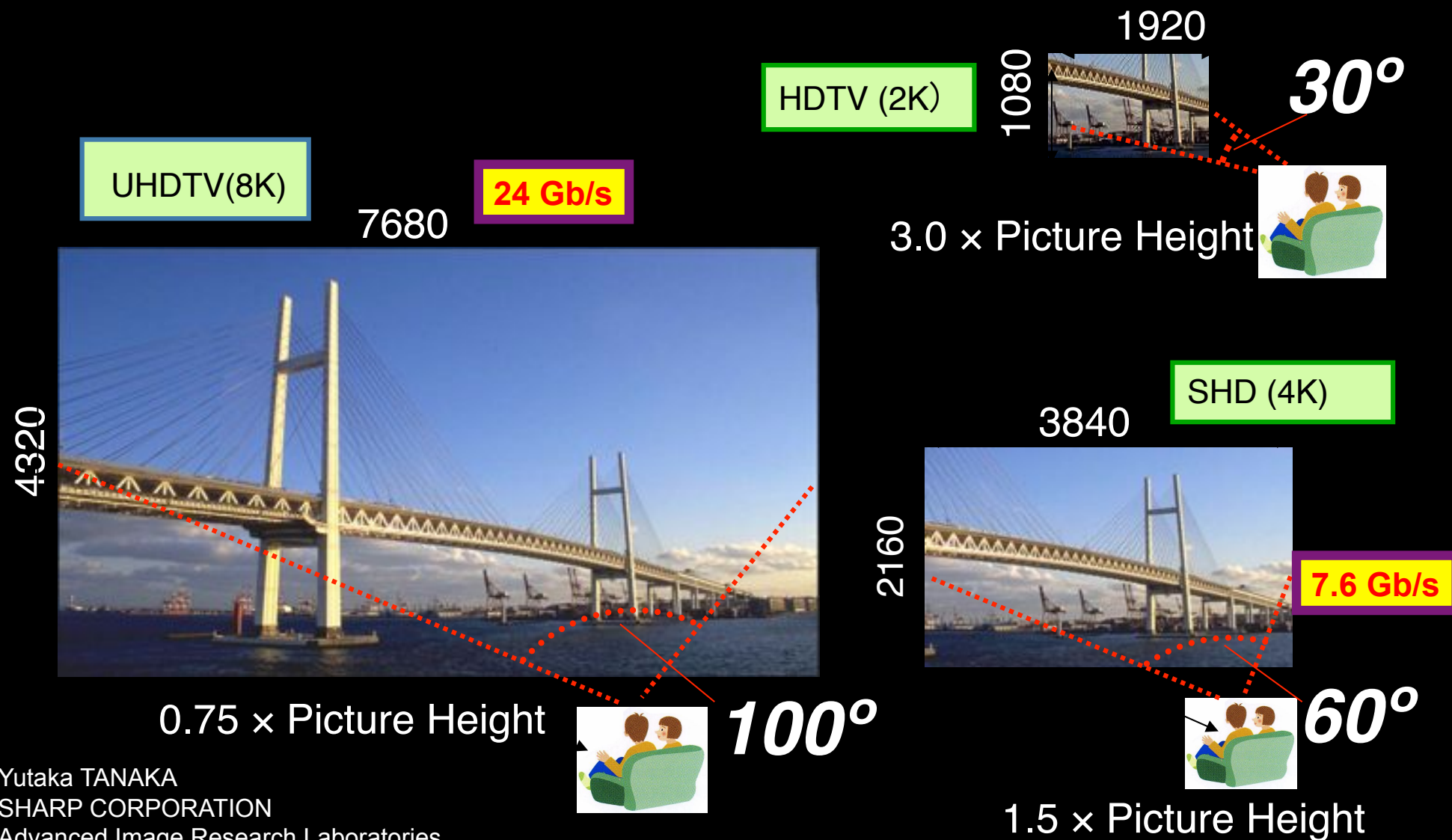


Auto-balancing Supers



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

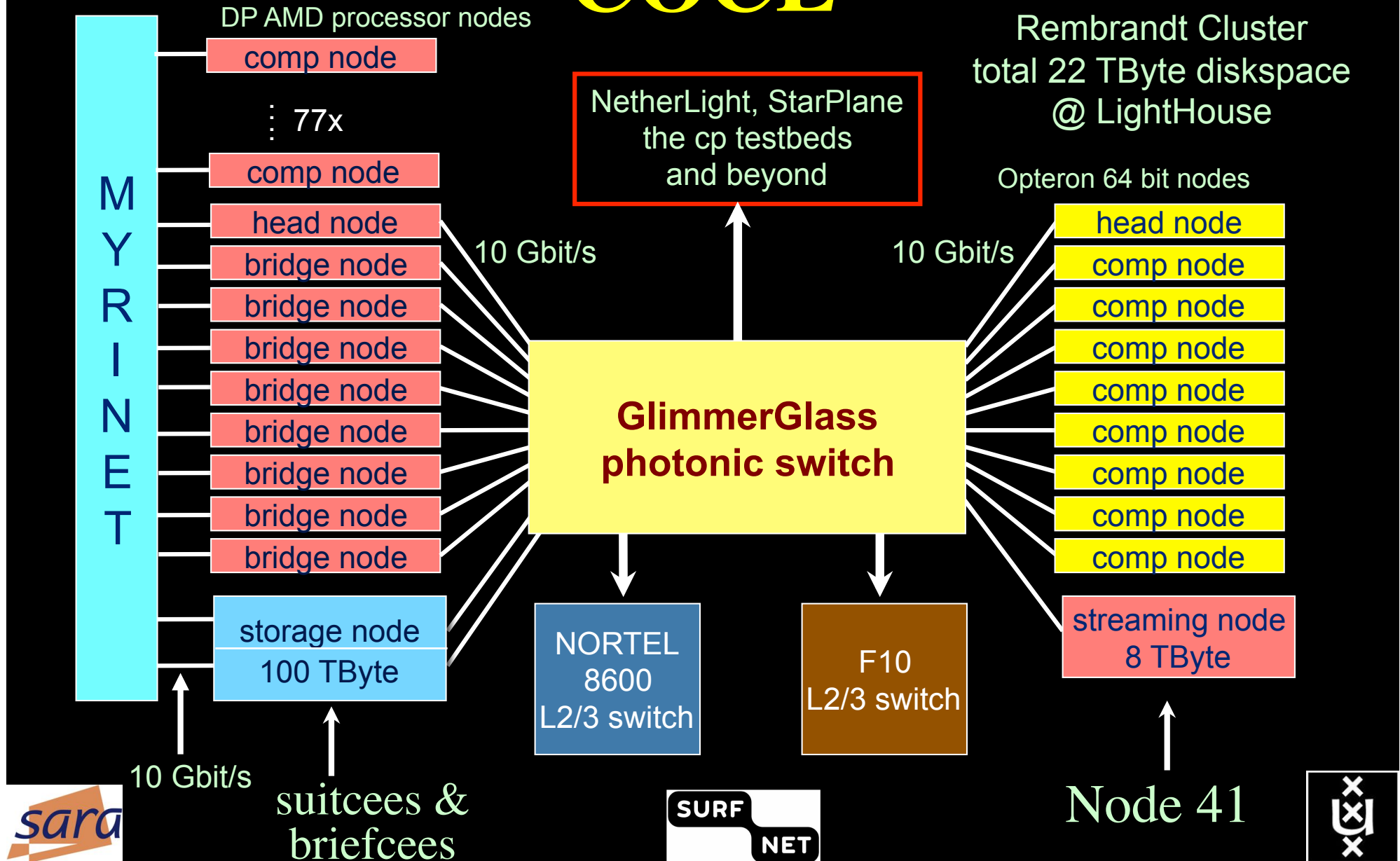


Amsterdam CineGrid S/F node

DAS-3 @ UvA

“COCE”

Rembrandt Cluster
total 22 TByte disk space
@ LightHouse



CineGrid portal



CineGrid distribution center Amsterdam

[Home](#) | [About](#) | [Browse Content](#) | [cinegrid.org](#) | [cinegrid.nl](#)

Amsterdam Node Status:

node41:
Disk space used: 8 GiB
Disk space available: 10 GiB

Search node:

Search

Browse by tag:

amsterdam animation
[antonacci](#) blender boat
bridge burny cgi delta holland
hollandfestival
leidsestraat
muziekgebouw
nieuwmarkt opera prague ship
train tram trams waag

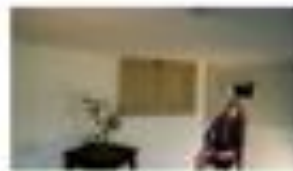
Via Distributed via Amsterdam

CineGrid Amsterdam

Welcome to the Amsterdam CineGrid distribution node. Below are the latest additions of super-high-quality video to our node.

For more information about CineGrid and our efforts look at the about section.

Latest Additions



Wypke

Wypke

Available formats:

4k dot (4.0 KB)

Duration: 1 hour and 8 minutes

Created: 1 week, 2 days ago

Author: Wypke

Categories:



Prague Train

Steam locomotive in Prague

Available formats:

4k dot (3.9 KB)

Duration: 27 hours and 46 minutes

Created: 1 week, 2 days ago

Author: CineGrid

Categories: delta prague train



VLC: Big Buck Bunny

(C) copyright Blender Foundation | <http://www.bigbuckbunny.org>

Available formats:

1080p MPEG4 (1.1 GB)

Duration: 1 hour and 0 minutes

Created: 1 month, 1 week ago

Author: Blender Foundation

Categories: animation blender bunny
cgi

u
s
e
r
s

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 2010

$$\Sigma A \approx 850 \text{ Gb/s}$$

$$\Sigma C \approx 350 \text{ Gb/s}$$

However:

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

A

B

C

ADSL (20 Mbit/s)

BW GigE

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 \approx 2-3 k\$/port



L2 \approx 5-8 k\$/port

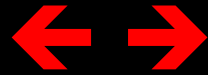


L3 \approx 75+ k\$/port



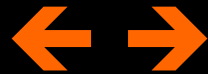
Hybrid computing

Routers



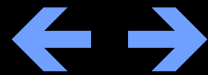
Supercomputers

Ethernet switches



Grid & Cloud

Photonic transport



GPU's

What matters:

Energy consumption/multiplication

Energy consumption/bit transported

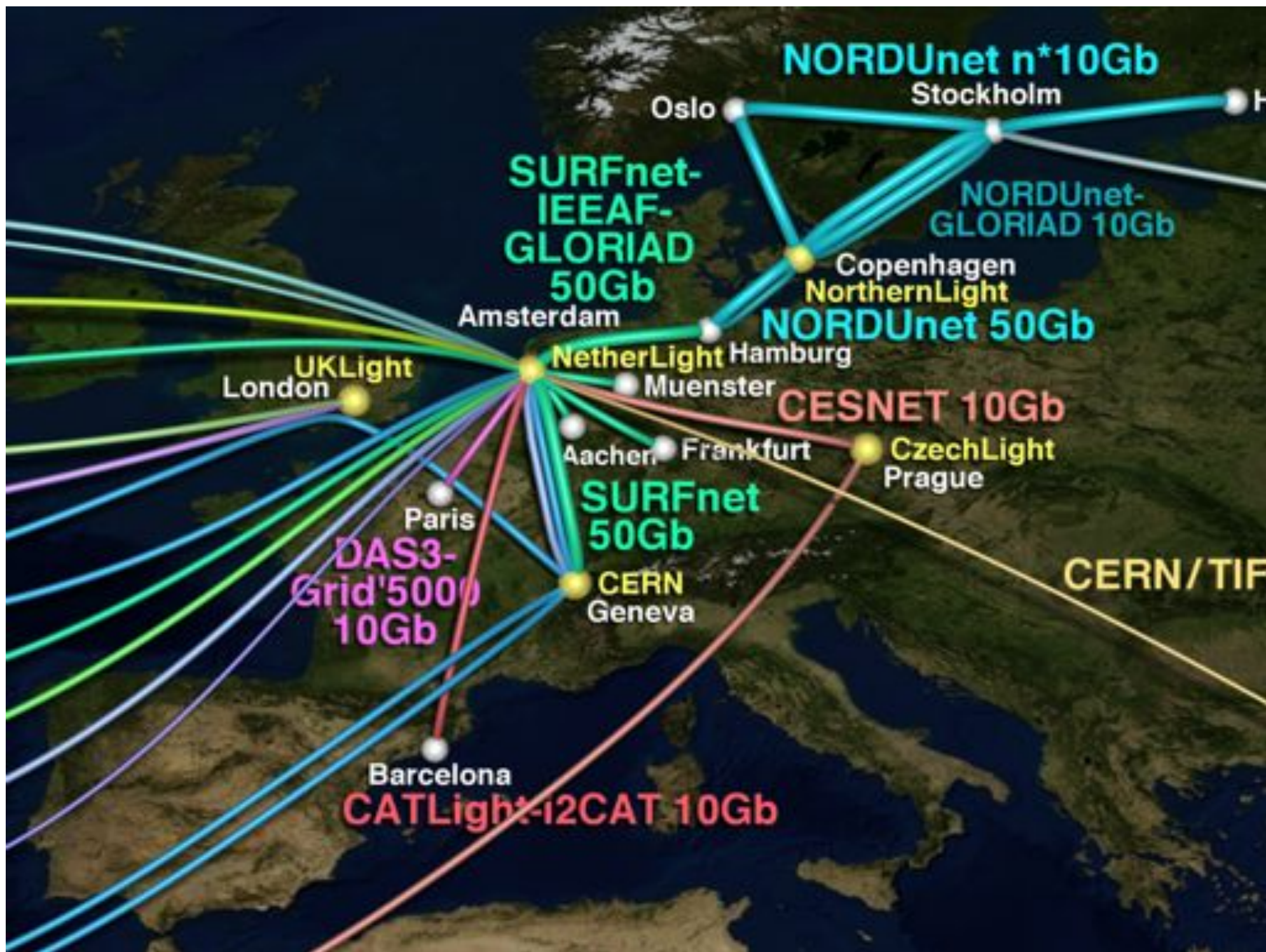




GLIF 2008

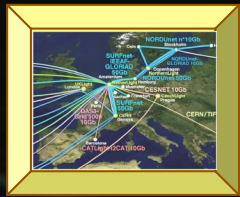
**Visualization courtesy of Bob Patterson, NCSA
Data collection by Maxine Brown.**





•VIZ

- DataExploration
- RemoteControl
- TV
- Medical
- CineGrid
- Gaming
- Conference



•DATA

- Management
- Backup
- Mining
- Web2.0
- Media
- Visualisation
- Security
- Meta



NetherLight

•GRID

- Workflow
- Clouds
- Distributed
- EventProcessing

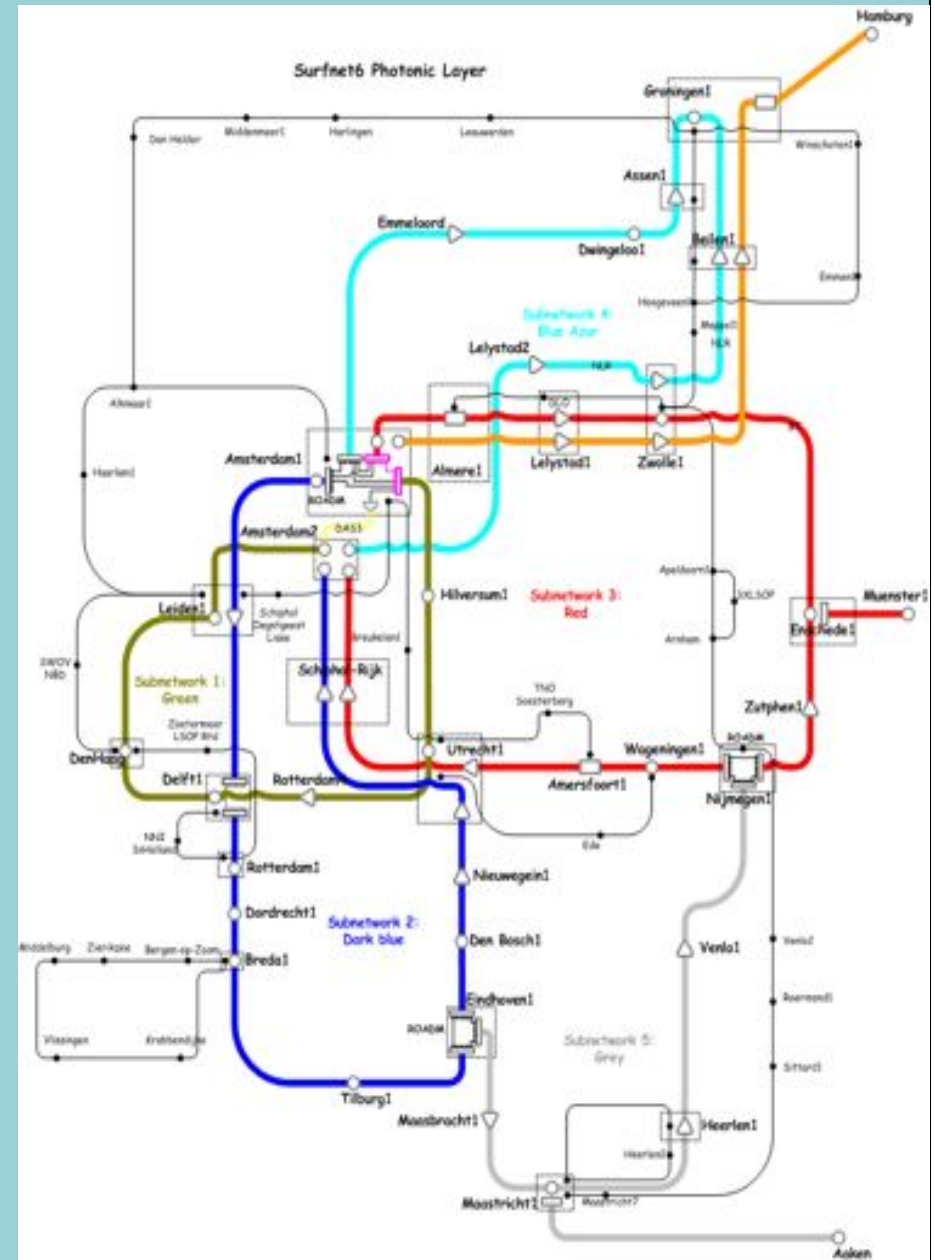


•SUPER

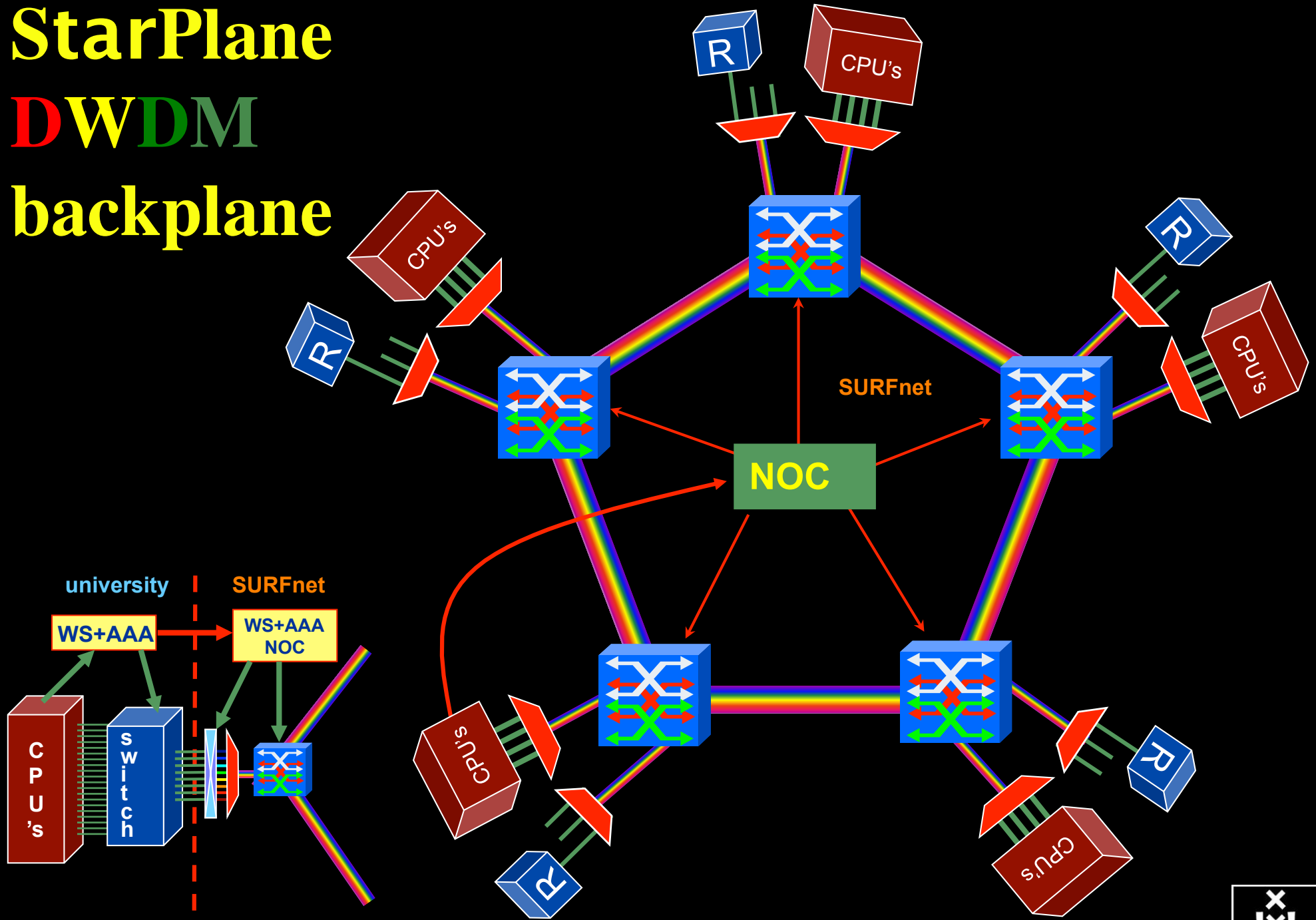
- Simulations
- StreamProcessing
- Predictions



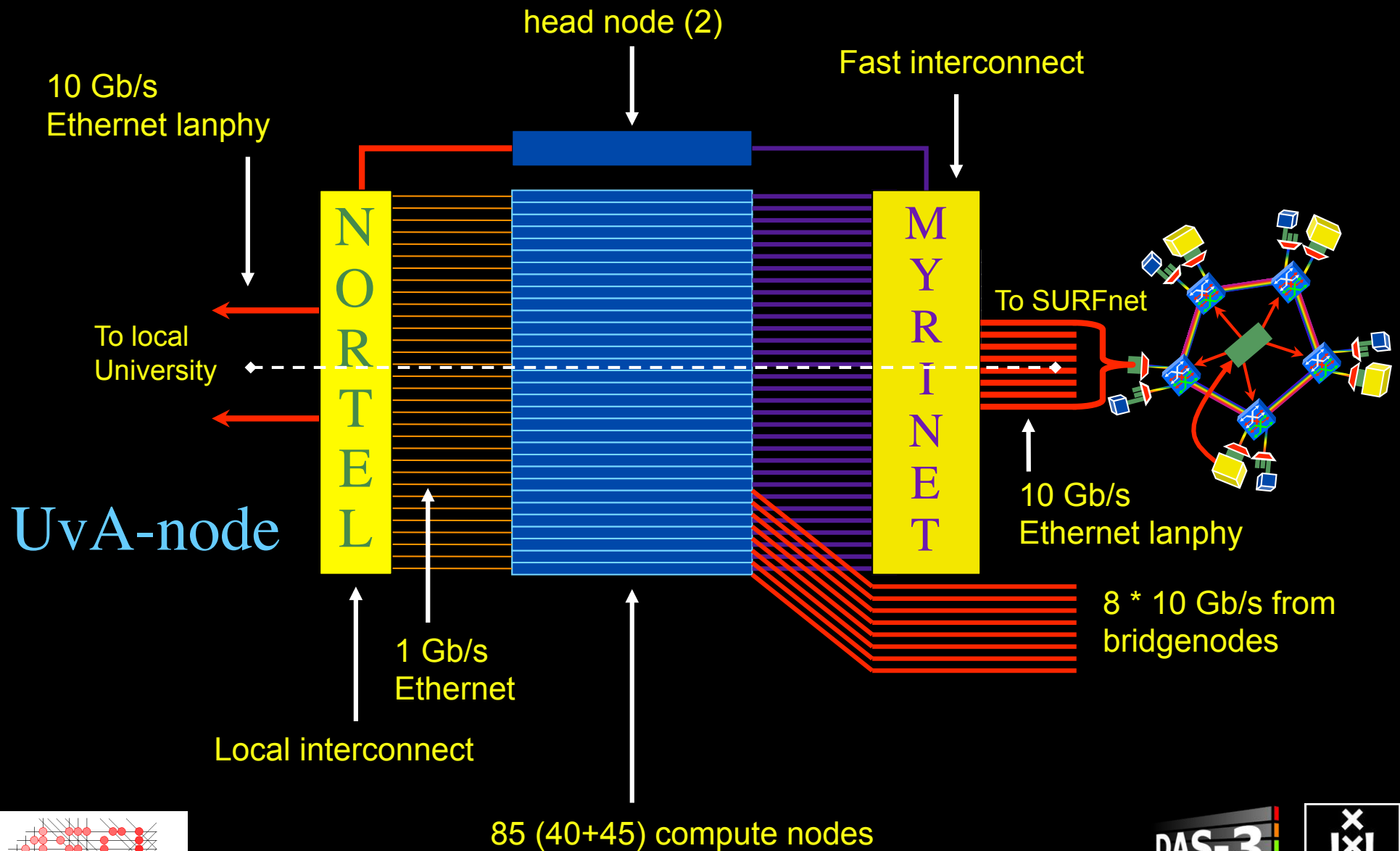
SURFnet6 on dark fiber



StarPlane DWDM backplane

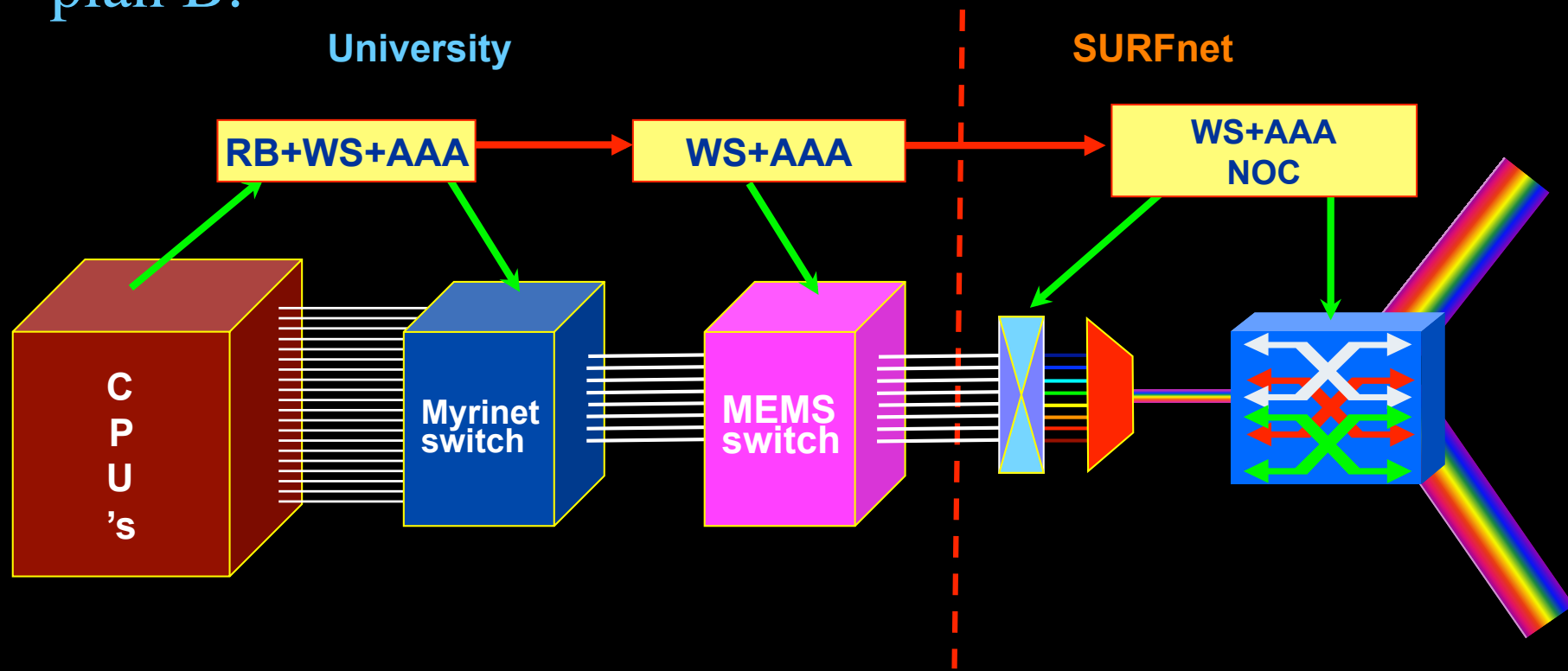


DAS-3 Cluster Architecture



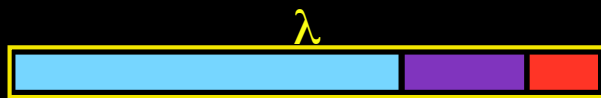
The challenge for sub-second switching

- bringing up/down a λ takes minutes
 - this was fast in the era of old time signaling (phone/fax)
 - λ 2λ influence (Amplifiers, non linear effects)
 - however minutes is historically grown, 5 nines, up for years
 - working with ~~Nortel~~ CIENA to get setup time significantly down
- plan B:

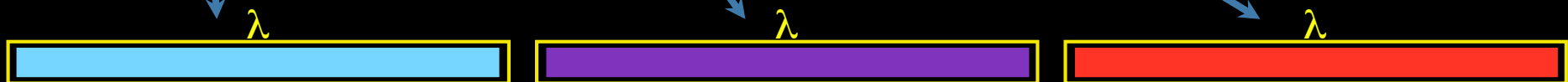


QOS in a non destructive way!

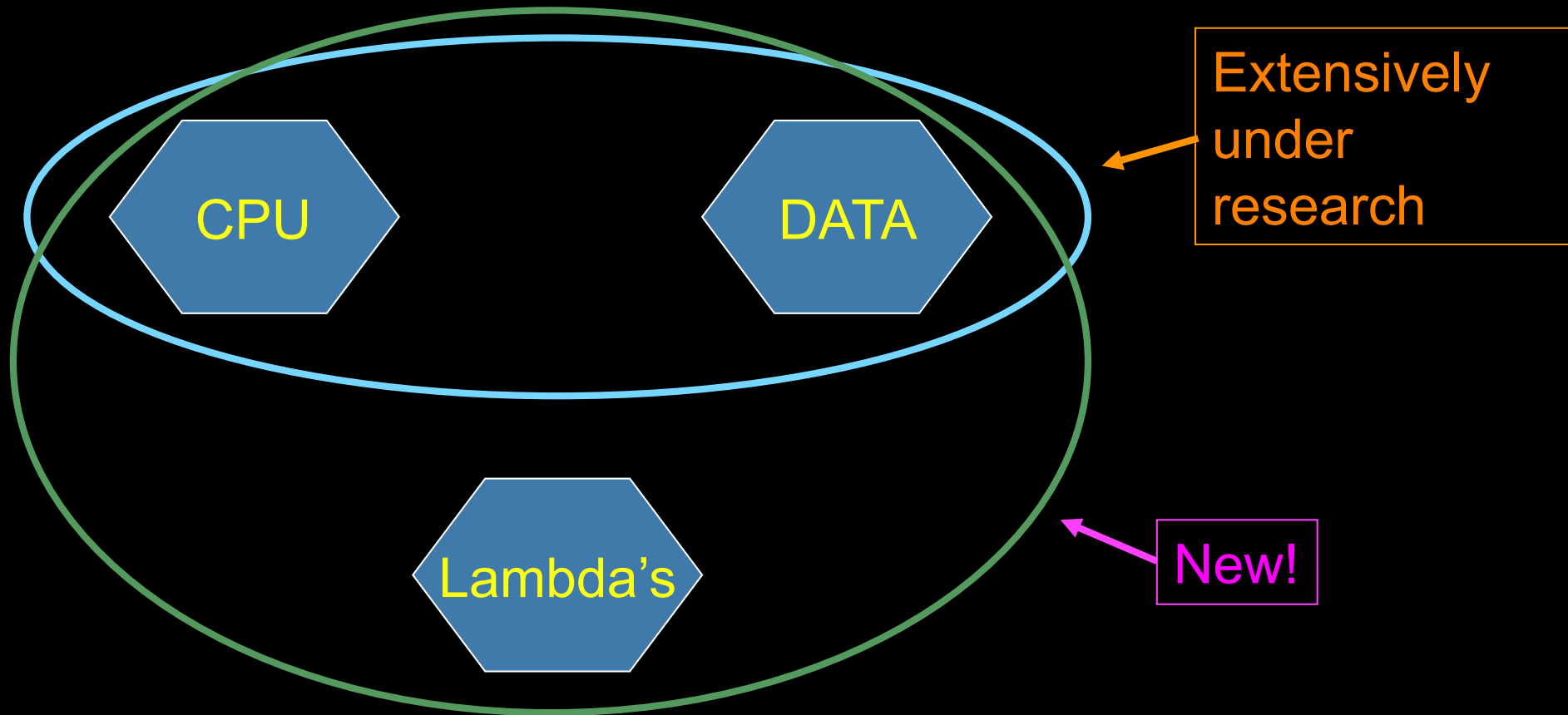
- Destructive QOS:
 - have a link or λ
 - set part of it aside for a lucky few under higher priority
 - rest gets less service



- 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.



Overview
 Throughput
 Load
 Ping
 UDP
 Plot

Scroll line: Last 7 days:

12:30:01 30 min.

Overview Net Tests between DAS-3 Hosts

- [Authenticate 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](#), [hist](#).

Select UDP value: [rate](#), [test](#).

MAY 31th 2007

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	0.097	0	0.013	0.01	0.017	0.15

Ping Min [ms]

(see 30 columns)

	VU-083	VU-085	LIACS-125	LIACS-127	UvA-236	UvA-239	UvA-236-M	UvA-239-M
VU-083	---				0.696		---	---
VU-085		---	1.390				---	---
LIACS-125		1.390	---				---	---
LIACS-127				---		1.230	---	---
UvA-236	0.696				---		---	---
UvA-239				1.230		---	---	---
UvA-236-M							---	0.025
UvA-239-M							0.025	---

Throughput [Mbit/s]

(see 30 columns)

	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				---	---

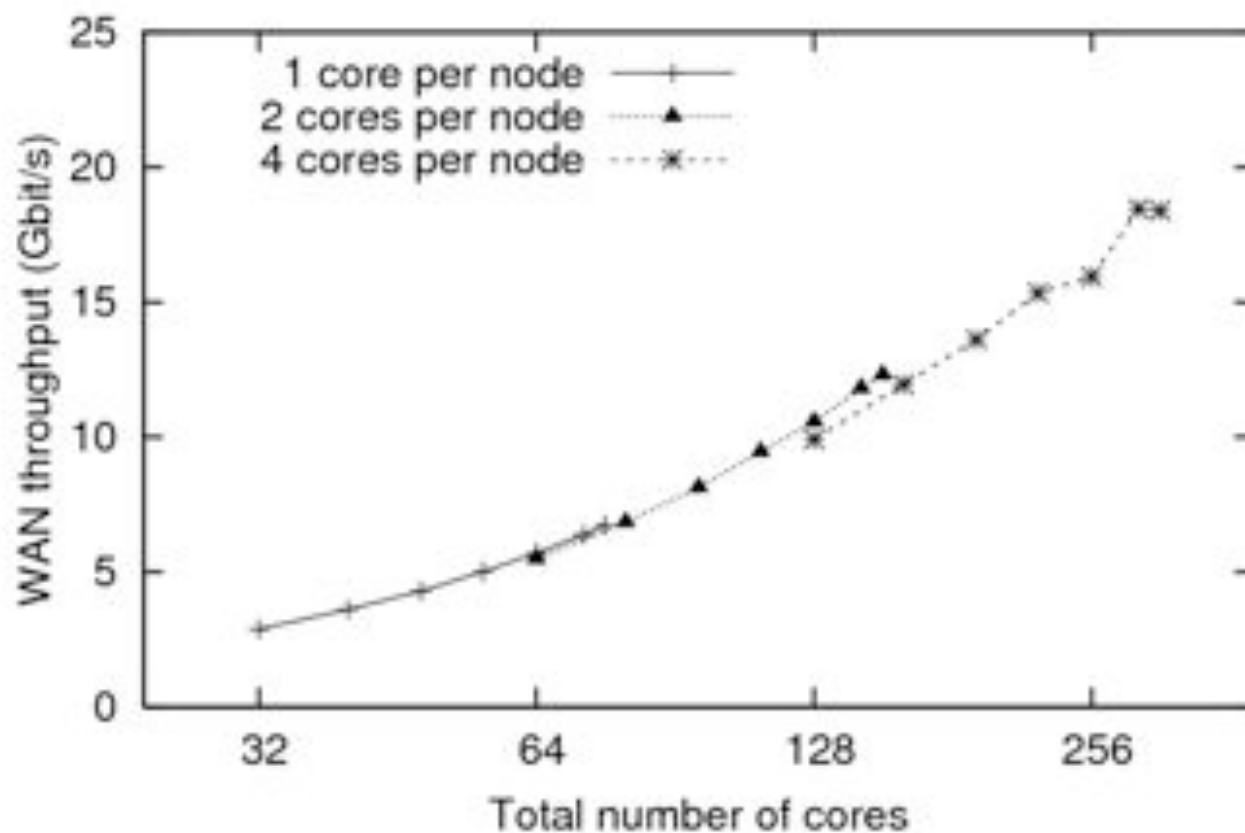
Games and Model Checking

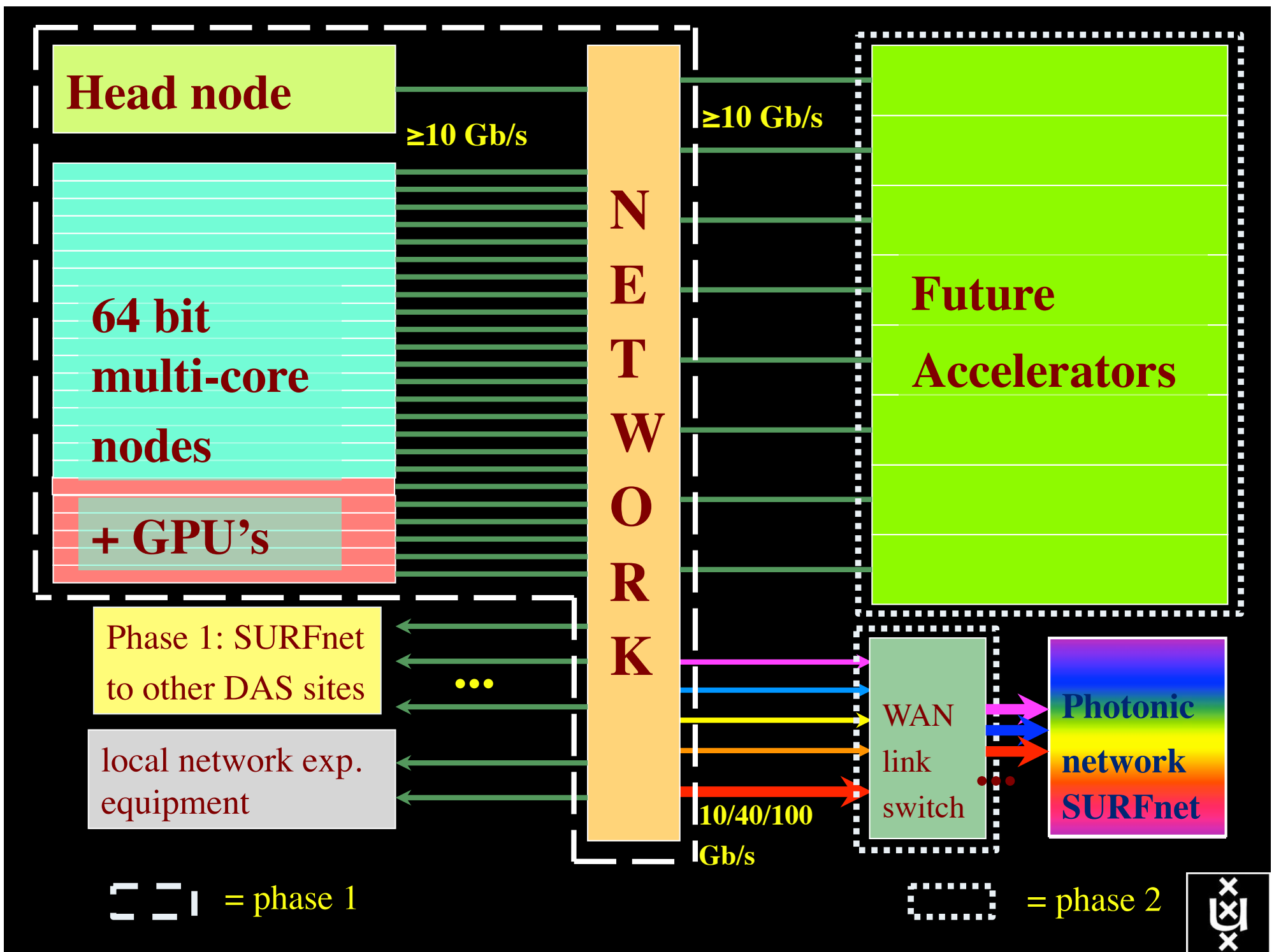
- Can solve entire Awari game on *wide-area* DAS-3 (889 B positions)
 - Needs 10G private optical network [CCGrid'08]
- Distributed model checking has very similar communication pattern
 - Search huge state spaces, random work distribution, bulk asynchronous transfers
- Can efficiently run DeVinE model checker on wide-area DAS-3, use up to 1 TB memory [IPDPS'09]





Required wide-area bandwidth





CBDF in Europe

Slide by Erik-Jan Bos





Video
Services
UK Scho

Alien light From idea to realisation!

40Gb/s alien wavelength transmission via a multi-vendor 10Gb/s DWDM infrastructure



Alien wavelength advantages

- Direct connection of customer equipment^[1]
→ cost savings
- Avoid OEO regeneration → power savings
- Faster time to service^[2] → time savings
- Support of different modulation formats^[3]
→ extend network lifetime

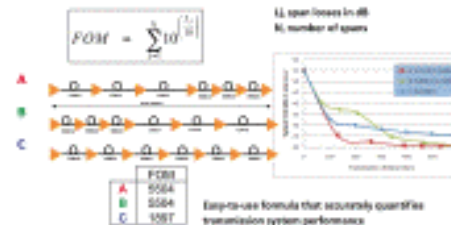
Alien wavelength challenges

- Complex end-to-end optical path engineering in terms of linear (i.e. OSNR, dispersion) and non-linear (FWM, SPM, XPM, Raman) transmission effects for different modulation formats.
- Complex interoperability testing.
- End-to-end monitoring, fault isolation and resolution.
- End-to-end service activation.

In this demonstration we will investigate the performance of a 40Gb/s PM-QPSK alien wavelength installed on a 10Gb/s DWDM infrastructure.

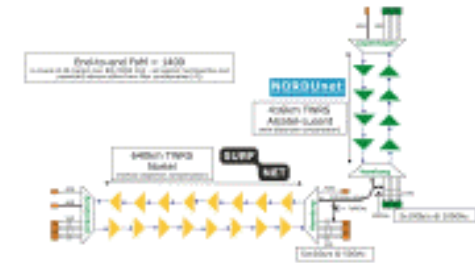
New method to present fiber link quality, FoM (Figure of Merit)

In order to quantify optical link grade, we propose a new method of representing system quality: the FOM (Figure of Merit) for concatenated fiber spans.

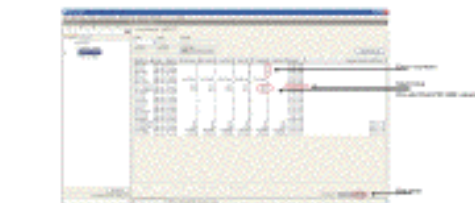


Transmission system setup

JOINT SURFnet/NORDUnet 40Gb/s PM-QPSK alien wavelength DEMONSTRATION.



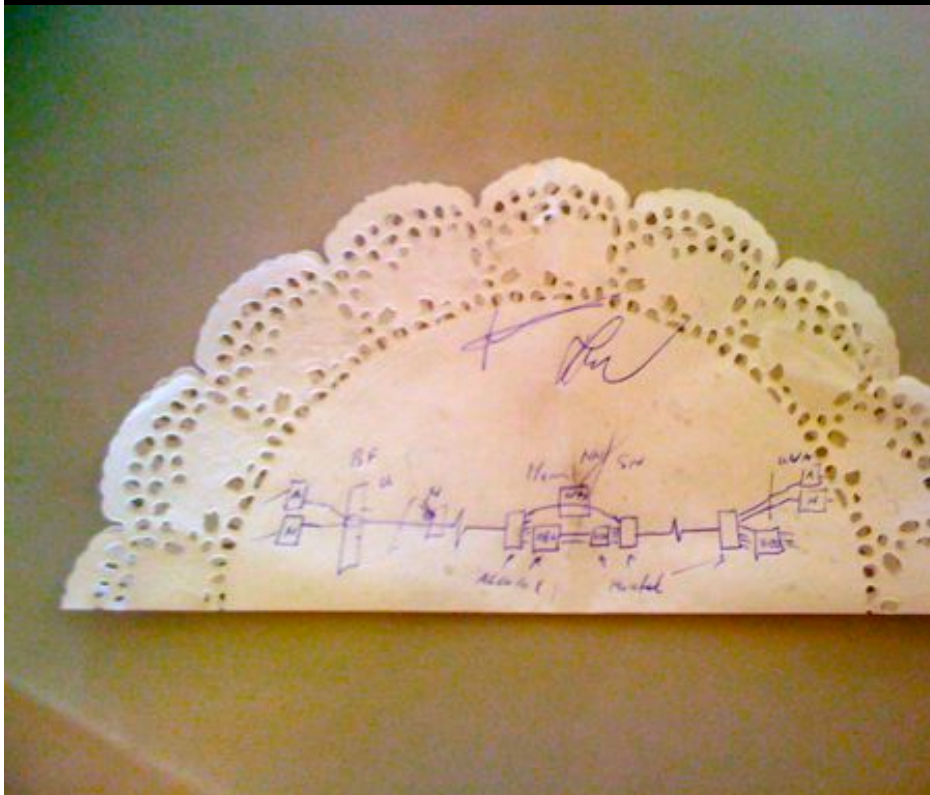
Test results



Error-free transmission for 23 hours, 17 minutes → BER < 3.0 x 10⁻¹⁶

Conclusions

- We have investigated experimentally the all-optical transmission of a 40Gb/s PM-QPSK alien wavelength via a concatenated native and third party DWDM system that both were carrying live 10Gb/s wavelengths.
- The end-to-end transmission system consisted of 1056 km of TWRS (TrueWave Reduced Slope) transmission fiber.
- We demonstrated error-free transmission (i.e. BER below 10⁻¹⁵) during a 23 hour period.
- More detailed system performance analysis will be presented in an upcoming paper.

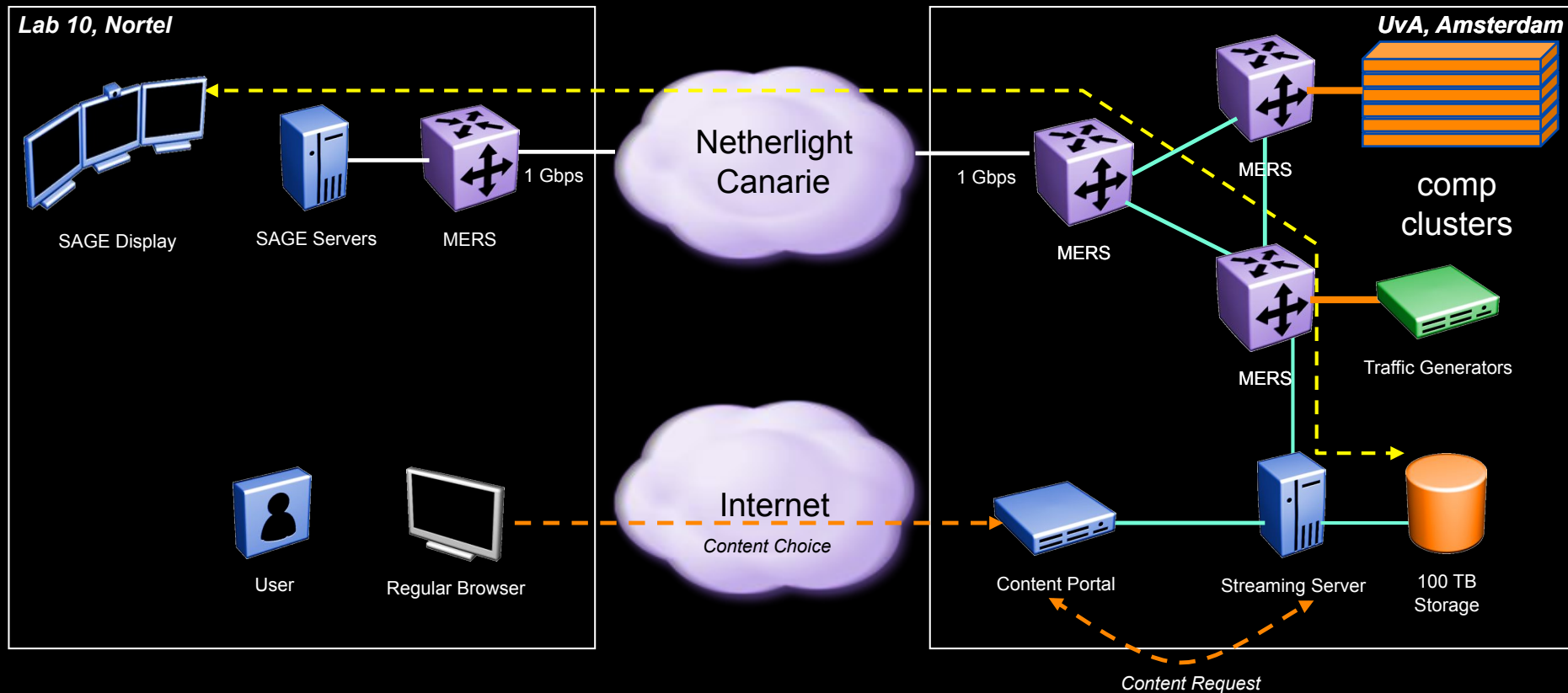


REFERENCES
[1] "OPERATIONAL SOLUTION FOR AN OPEN DWDM LAYER", B. GONTEL ET AL., OTC 2009. [2] "NET OPTICAL TRANSMIT SERVICES", MARRAS, SMITH, OTC09. [3] "SPIN SPINNING OF ALL-OPTICAL CORE NETWORKS", AMERIN, LORO AND CARL, ENLIGHTEN, ECOC2009. [4] "NON-LOCALITY IN FIBRE COMMUNICATIONS AND NEW CHALLENGES TO MANAGING THE MANAGED SERVICE", BARRAL, OTC09. [5] "NON-LOCALITY IN FIBRE COMMUNICATIONS DURING THE EXPERIMENT", WE ALSO ACKNOWLEDGE TELECOM AND NORTEL FOR THEIR IN-DOMAIN SERVICES AND TECHNICAL SUPPORT.

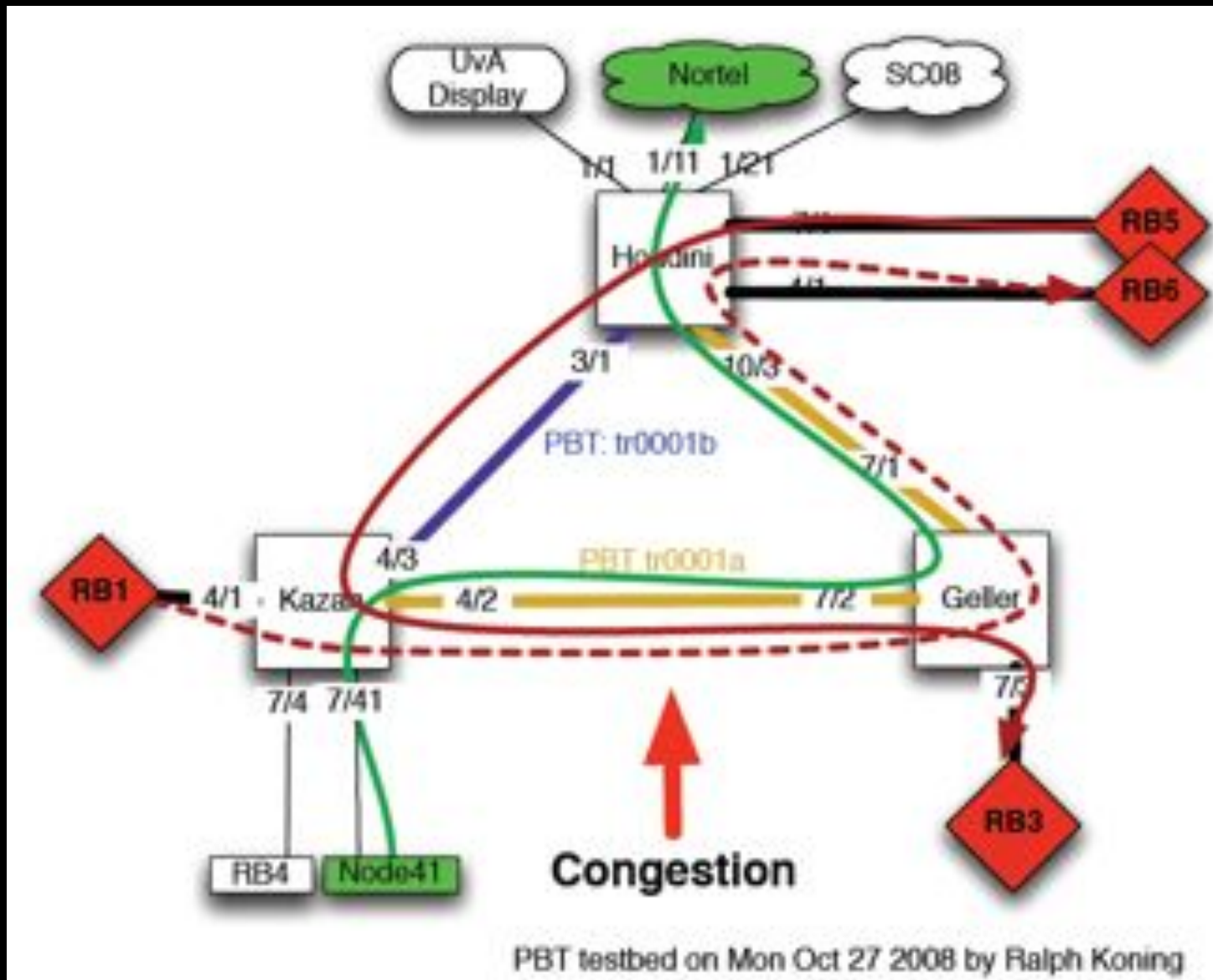
SC09

<file:///Users/delaat/Sites/sc09/index.html>

Diagram for SAGE video streaming to ATS



UvA Testbed



Congestion introduced in the network with multiple PBT paths carrying streamed SHD Content

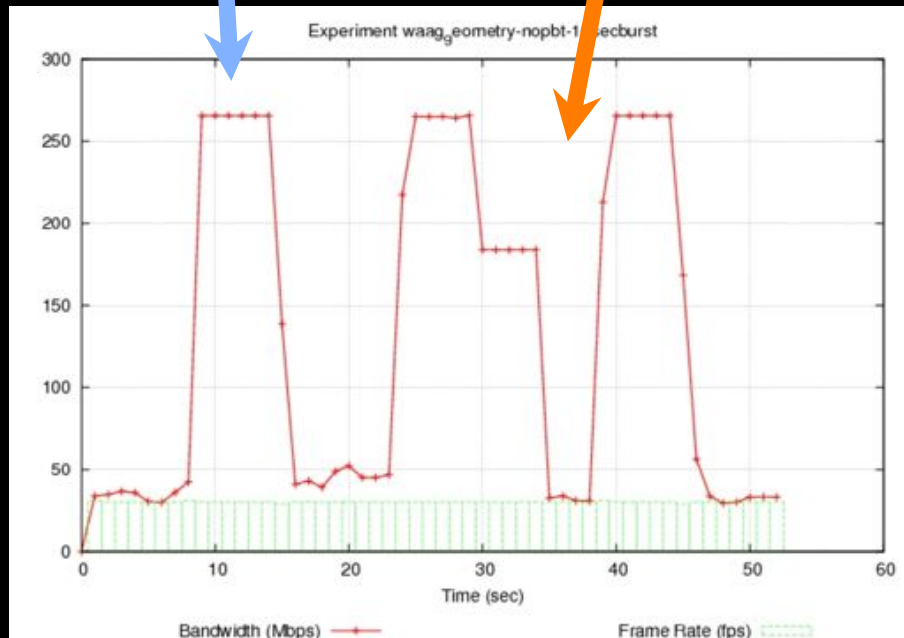


Experimental Data

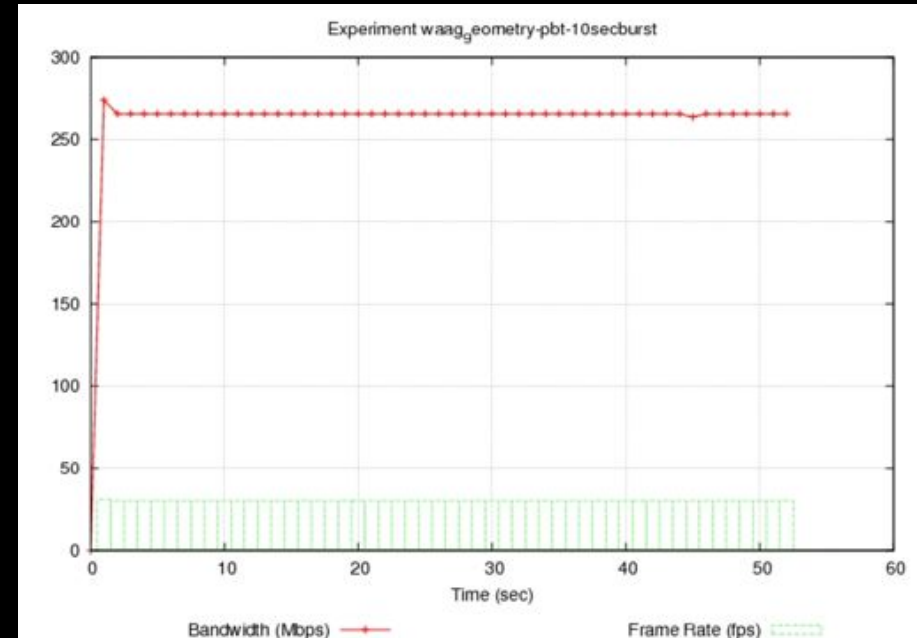


Sage without background traffic

Sage with background traffic



10 Second Traffic bursts with No PBT

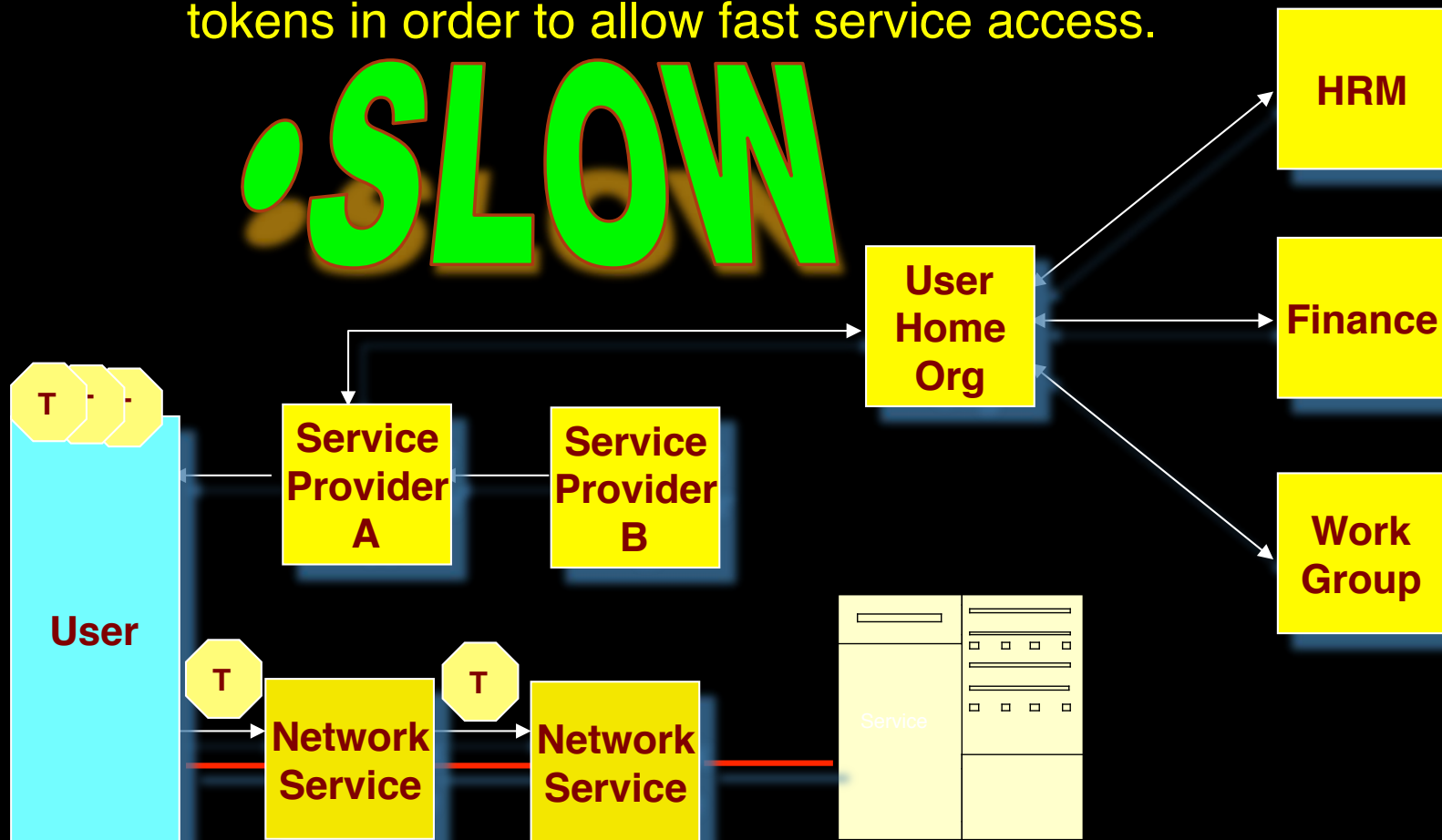


10 Second Traffic bursts with PBT

PBT is SIMPLE and EFFECTIVE technology to build a shared Media-Ready Network



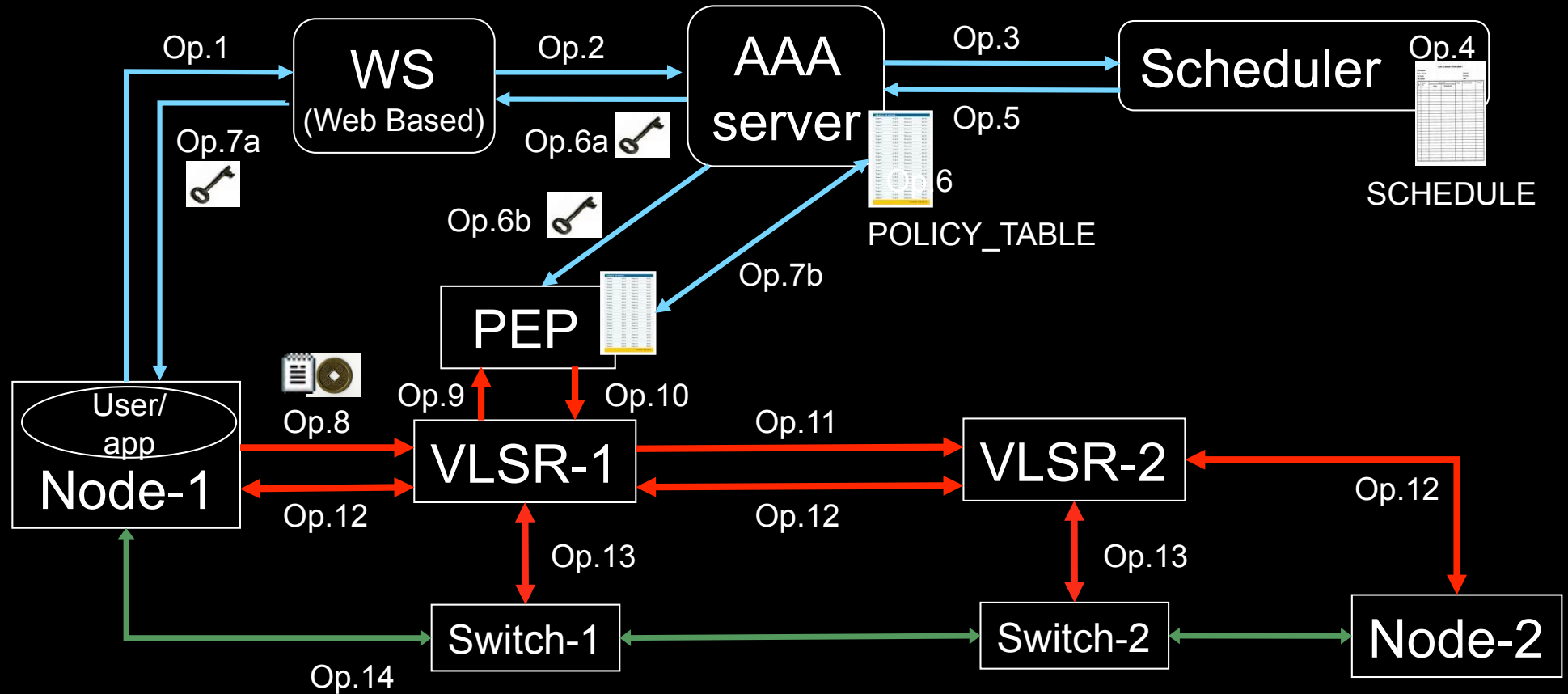
Use AAA concept to split (time consuming) service authorization process from service access using secure tokens in order to allow fast service access.



SLOW

Fast

Workflow



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





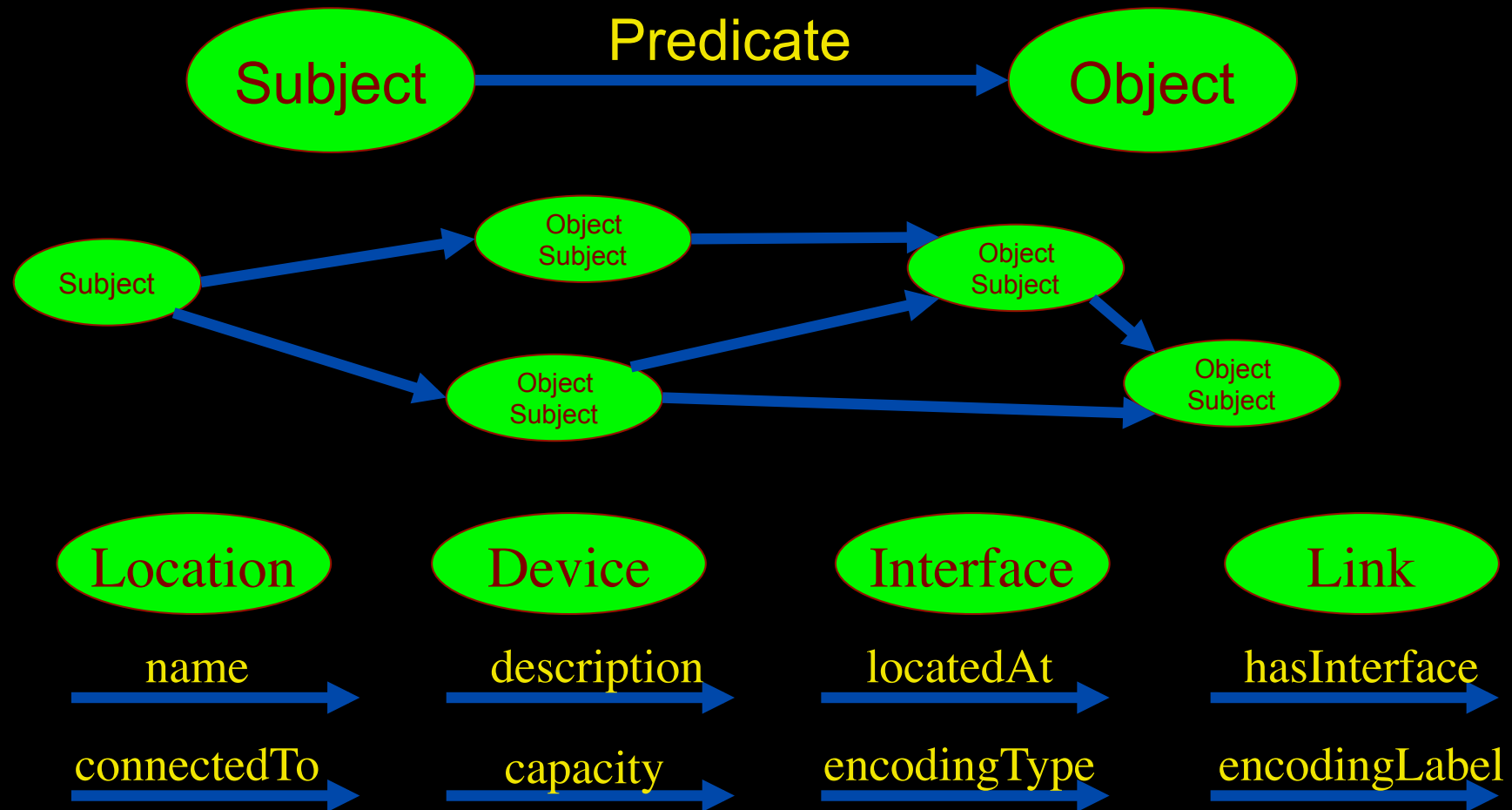
GLIF 2008

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



Network Description Language

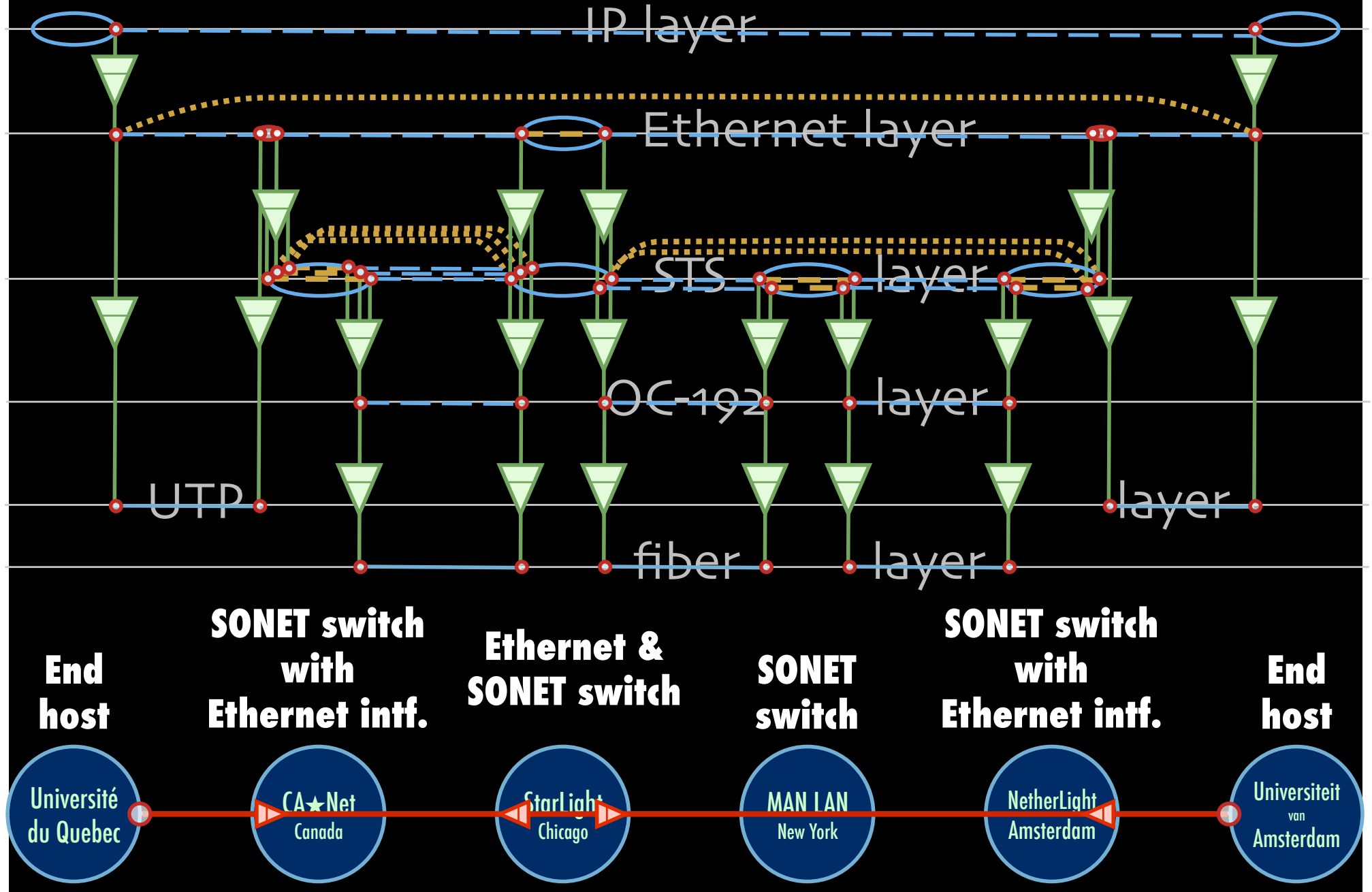
- From semantic Web / Resource Description Framework.
- The RDF uses XML as an interchange syntax.
- Data is described by triplets:



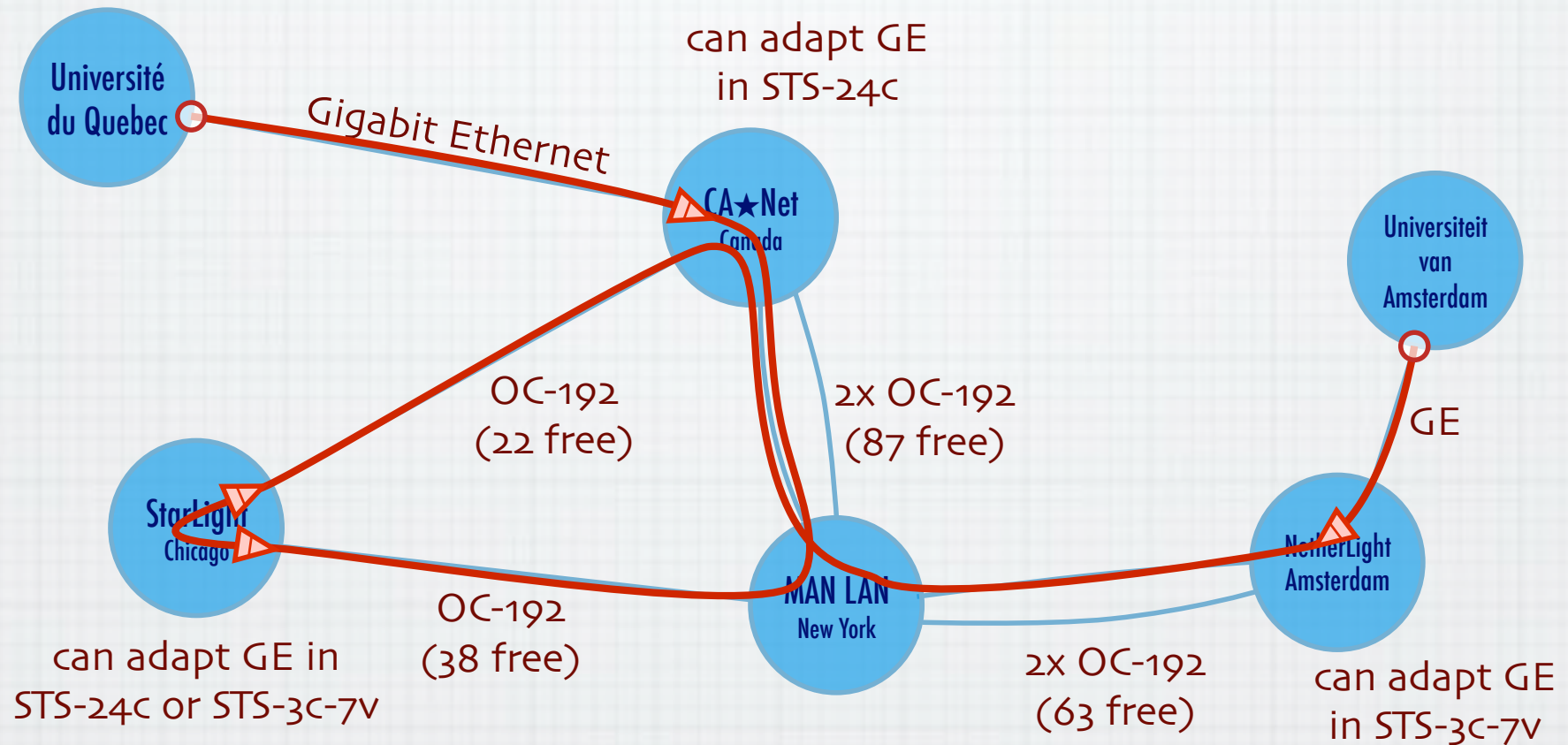
NetherLight in RDF

```
<?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:resource="#tdm3.amsterdam1.netherlight.net:503/2"/>
    <!-- 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

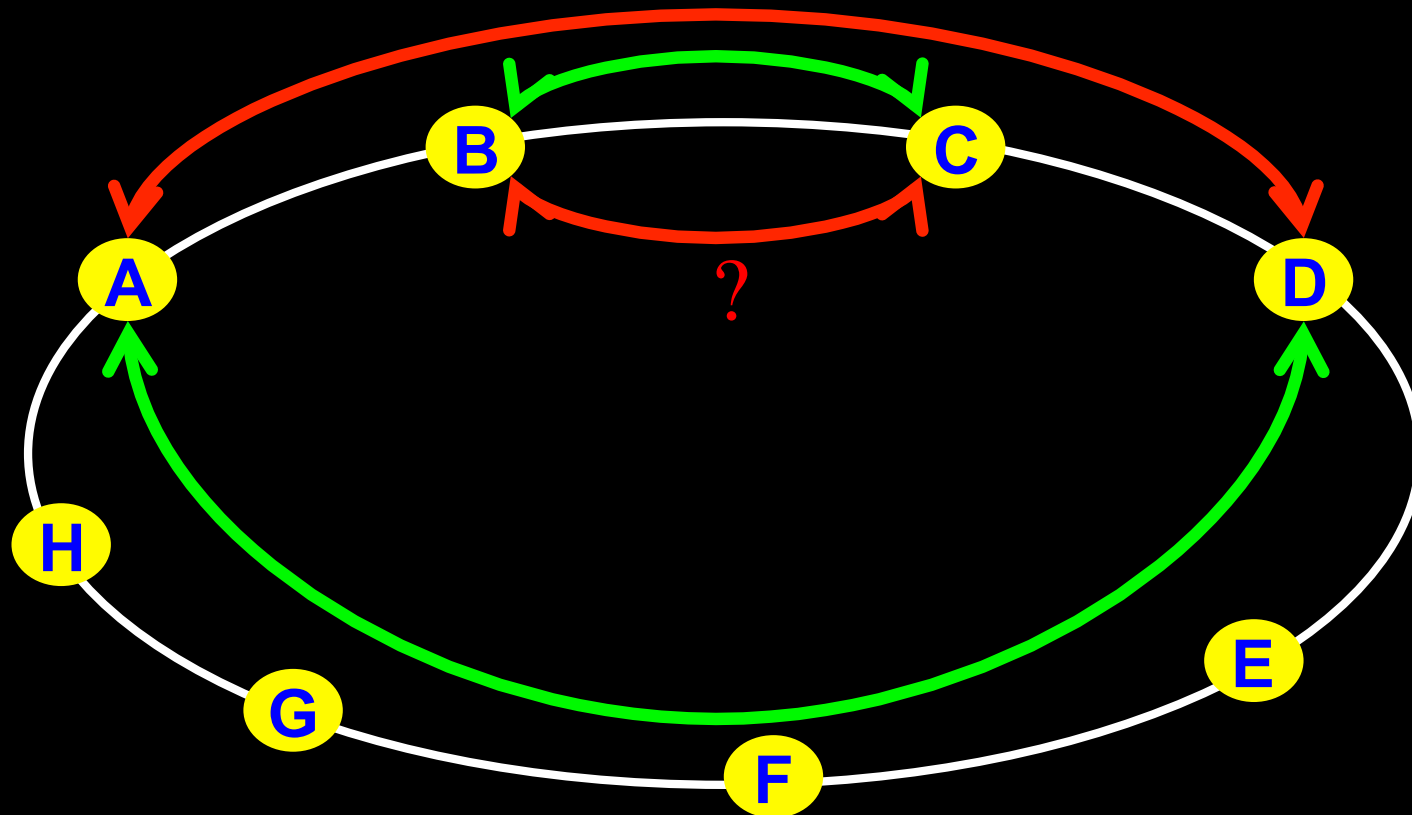


Thanks to Freek Dijkstra & team

The Problem

I want AD and BC

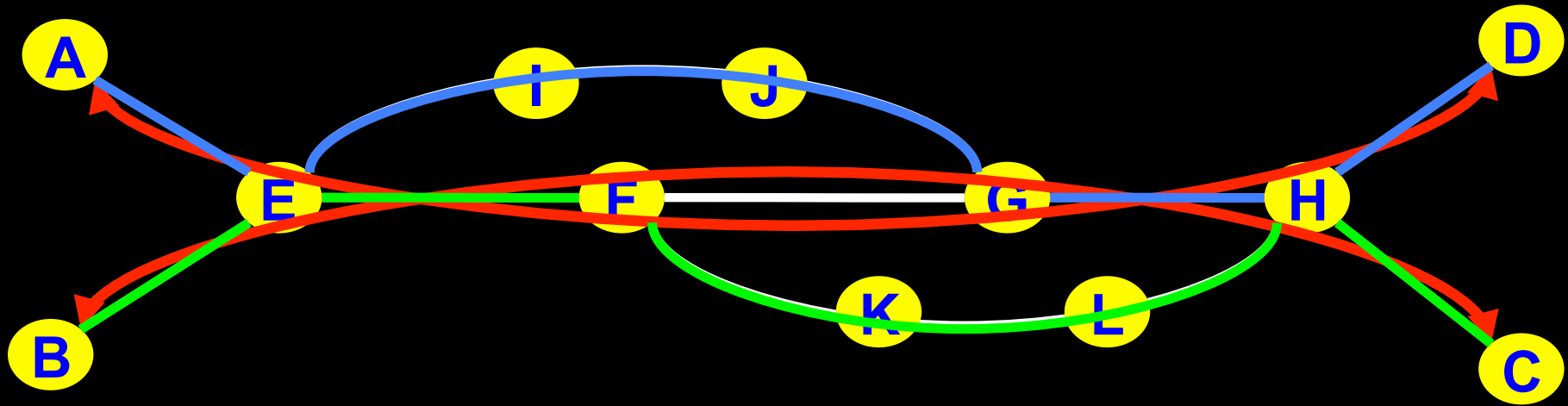
Success depends on the order



Another one 😊

I want AD and BC

Success does not even depend on the order!!!

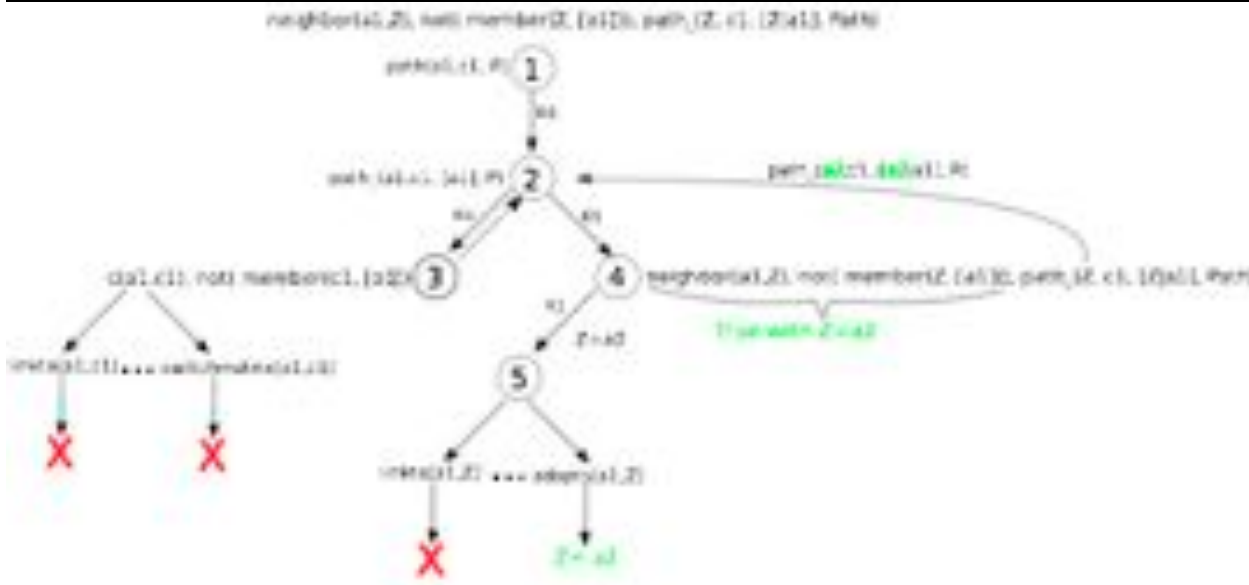
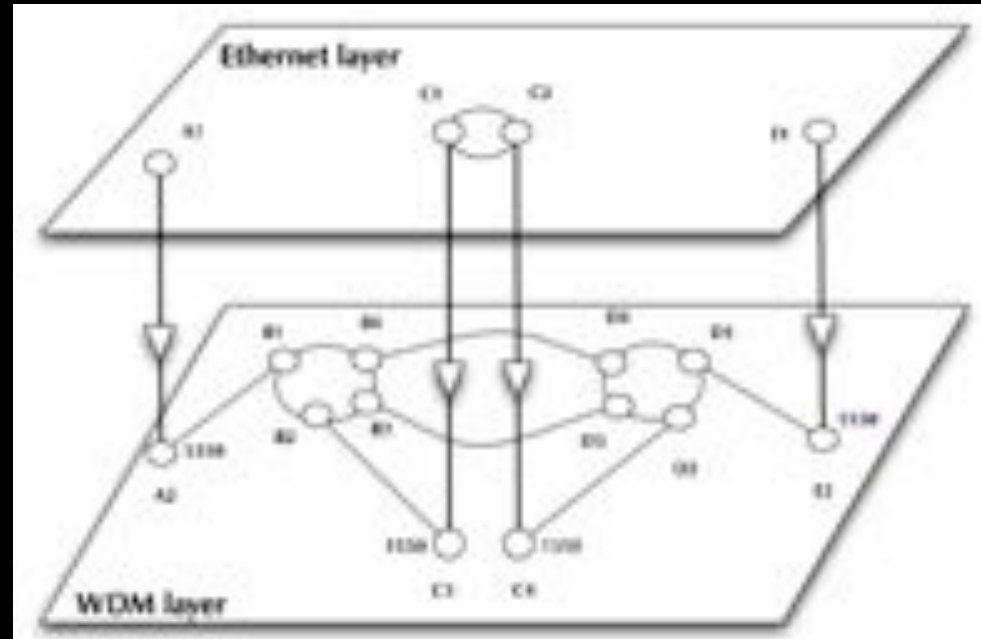


Wouldn't it be nice if I could request [AD, BC, ...]

NDL + PROLOG

Research Questions:

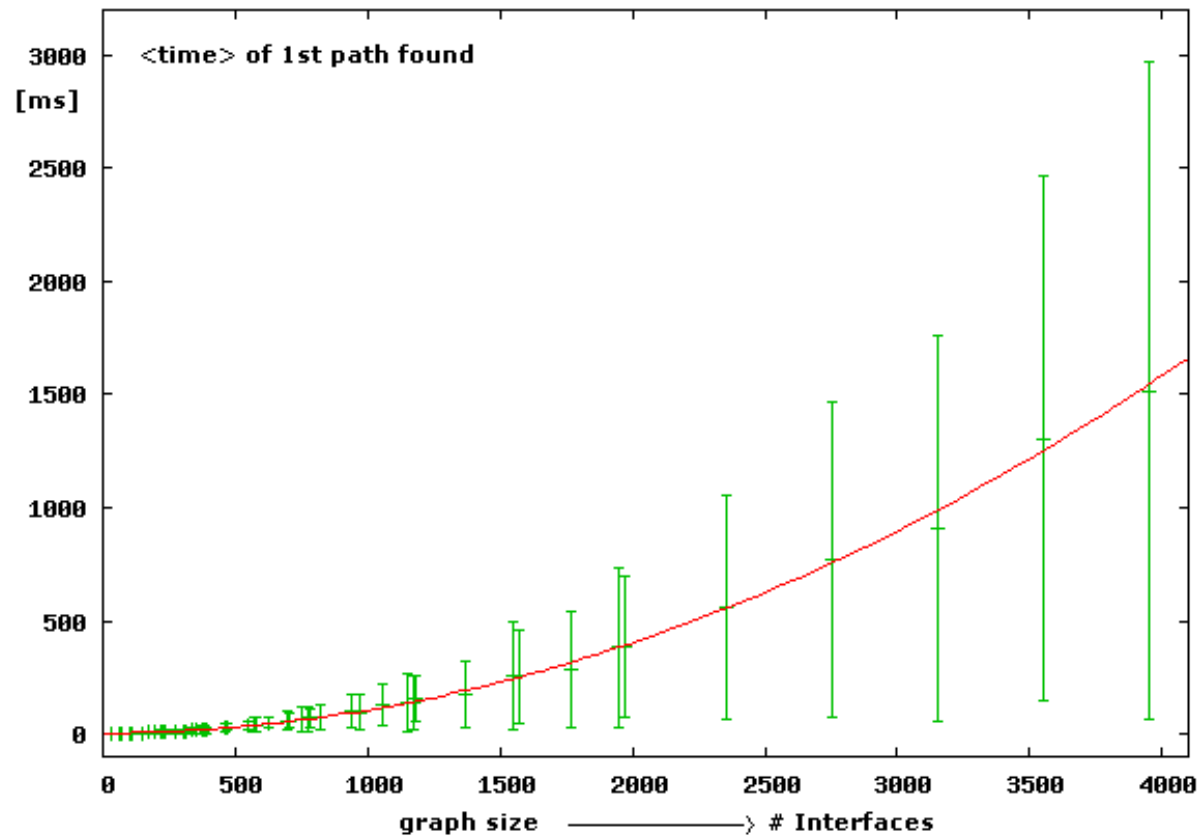
- order of requests
- complex requests
- usable leftovers



•Reason about graphs

•Find sub-graphs that comply with rules

Single layer networks: results

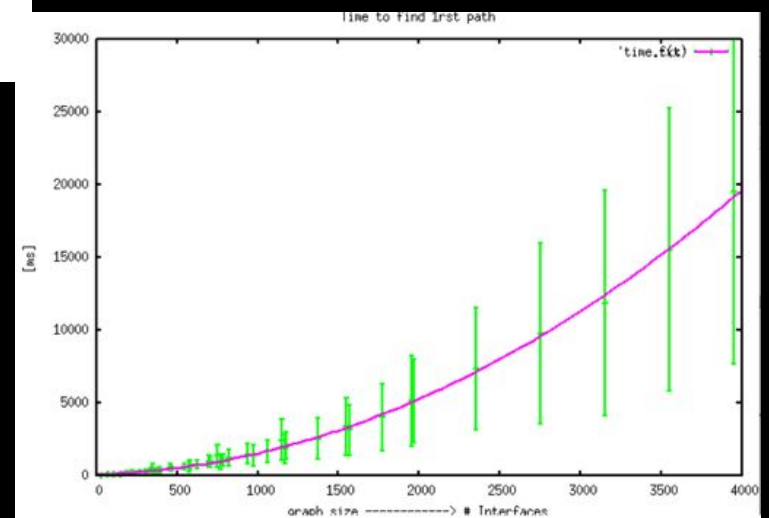


- Number of interfaces,
- given N nodes per domain D
- $4*(D-2) + D*4*(N-2)$ for $D > 2$

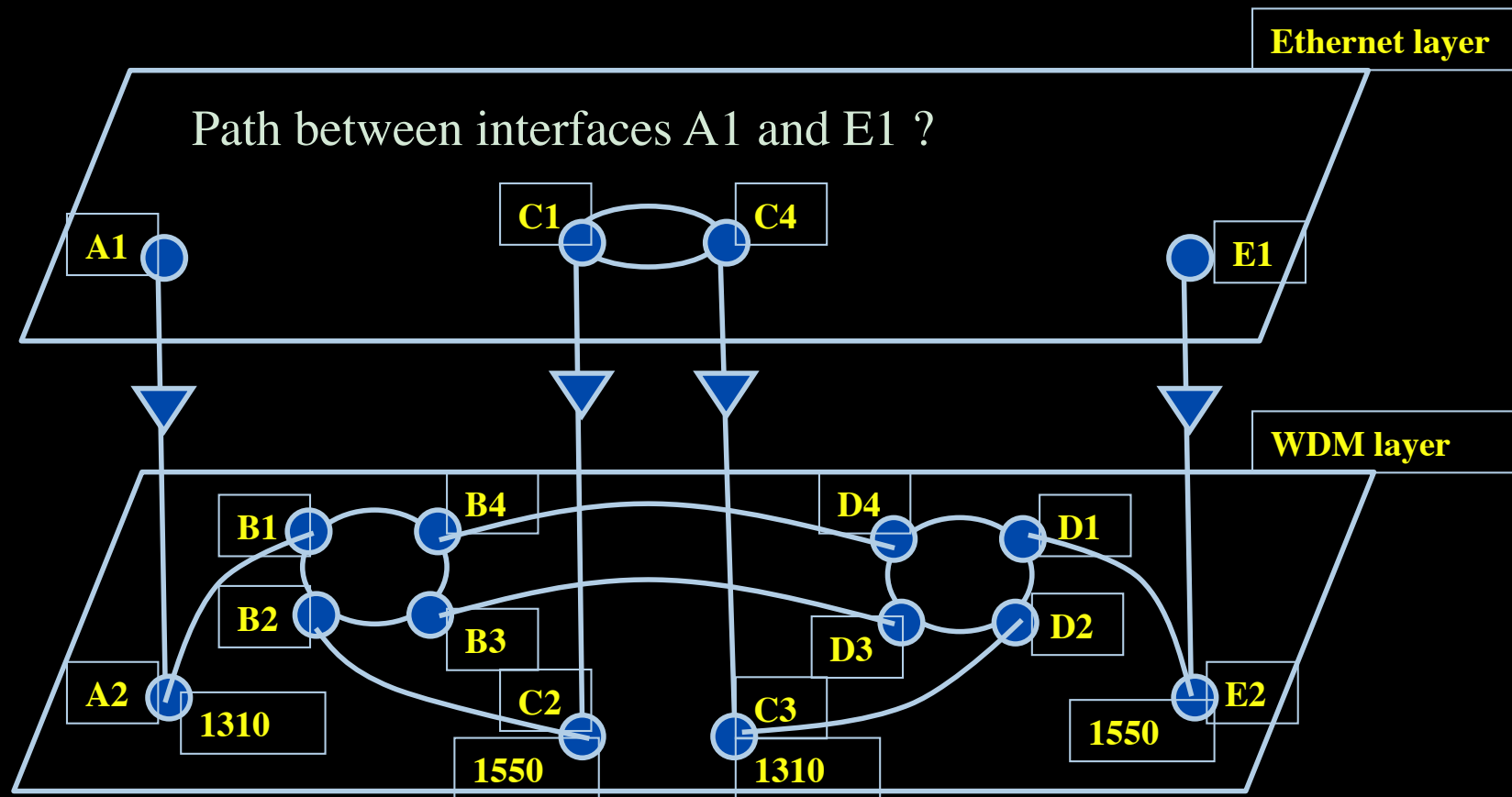
Pynt-based DFS

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 PathFinding



Prolog rule:

linkedto(Intf1, Intf2, CurrWav):-

```

rdf_db:rdf( Intf1, ndl:'layer', Layer ),
Layer == 'wdm#LambdaNetworkElement',
rdf_db:rdf( Intf1, ndl:'linkedTo', Intf2 ),
rdf_db:rdf( Intf2, wdm:'wavelength', W2 ),
compatible_wavelengths( CurrWav, W2 ).

```

%-- is there a link between Intf1 and Intf2 for wavelength CurrWav ?

%-- get layer of interface Intf1 → Layer

%-- are we at the WDM-layer ?

%-- is Intf1 linked to Intf2 in the RDF file?

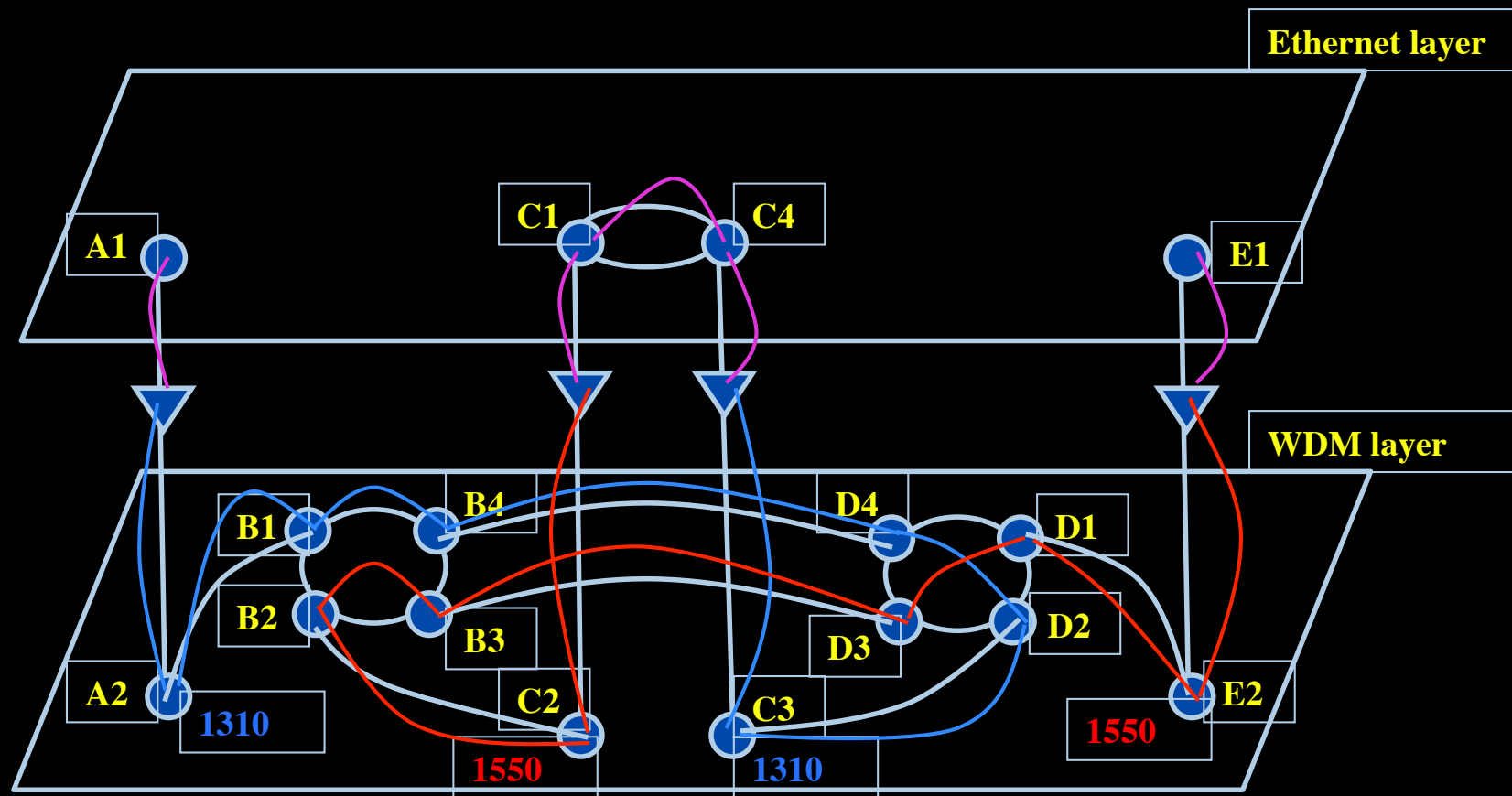
%-- get wavelength of Intf2 → W2

%-- 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 Network PathFinding

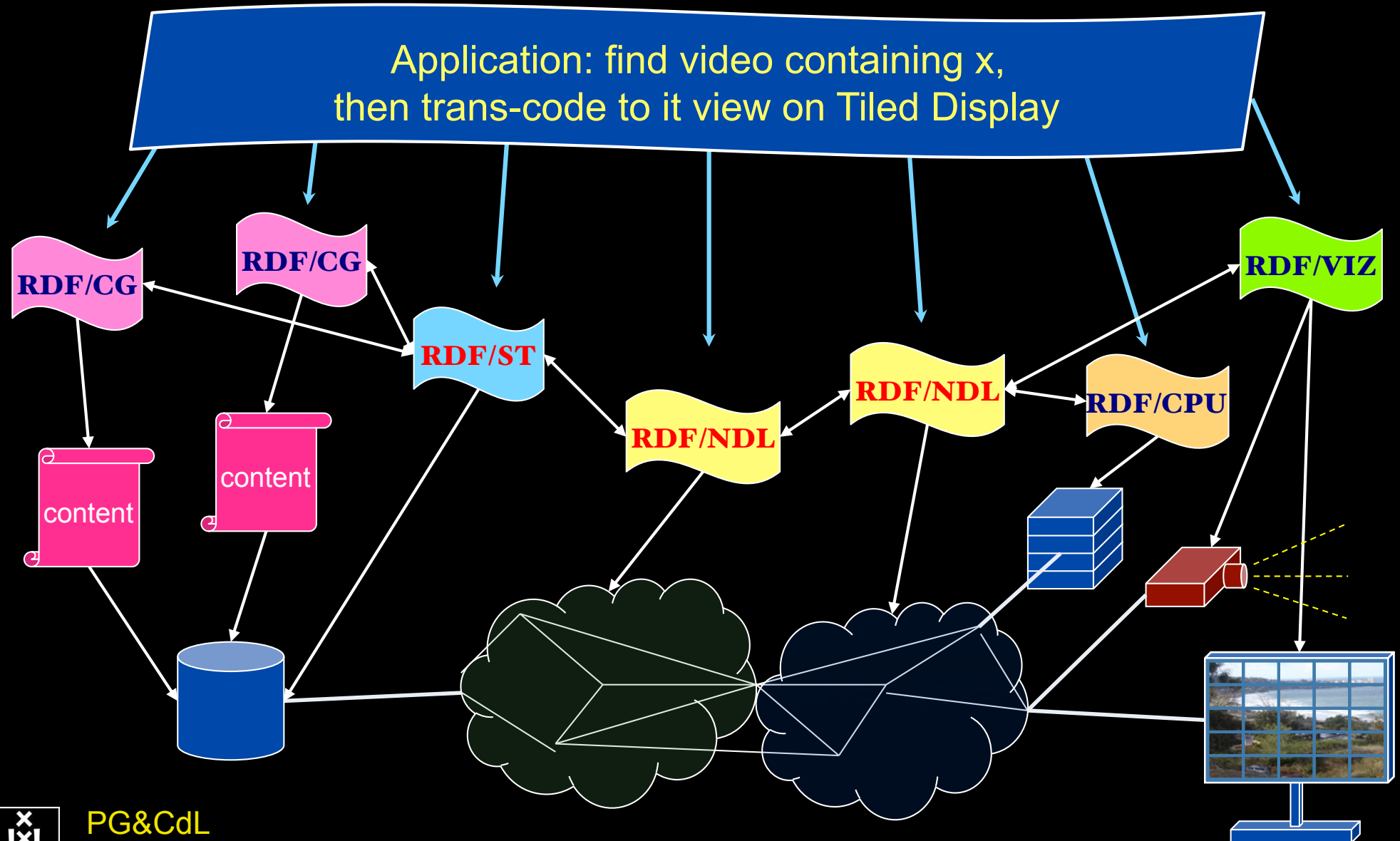


Path between interfaces A1 and E1:

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

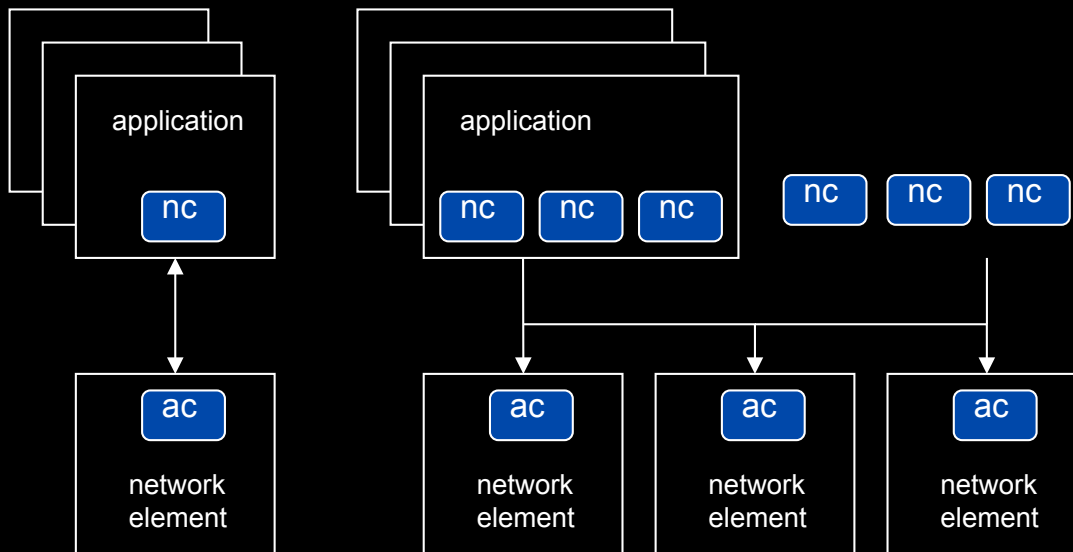
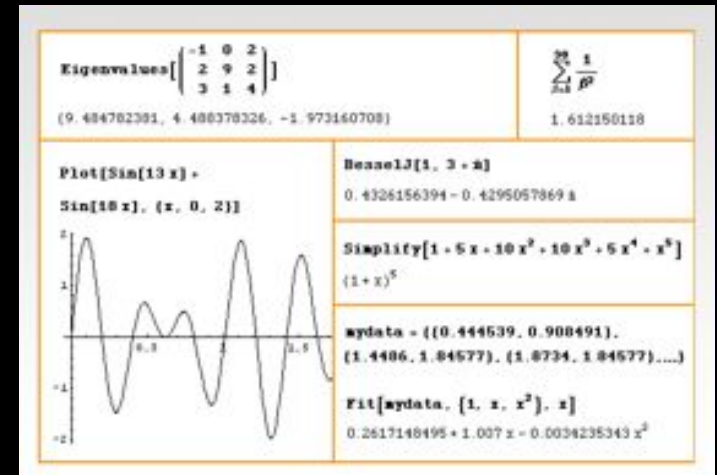
Scaling: Combinatorial problem

RDF describing Infrastructure



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, GetAllLinks, Remote,
NetworkTokenTransaction}
```

```
Global`upvnverbose = True;
```

```
AbsoluteTiming[nes = BFSDiscover["139.63.145.94"];][[1]]
```

```
AbsoluteTiming[result = BFSDiscoverLinks["139.63.145.94", nes];][[1]]
```

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

```
Getting neighbours of: 192.168.2.3
```

Transaction on shortest path with tokens

```
Internal links: {192.168.2.3}
```

```
nodePath = ConvertIndicesToNodes[
  ShortestPath[
    g,
    Node2Index[nids, "192.168.3.4"],
    Node2Index[nids, "139.63.77.49"]],
  nids];
```

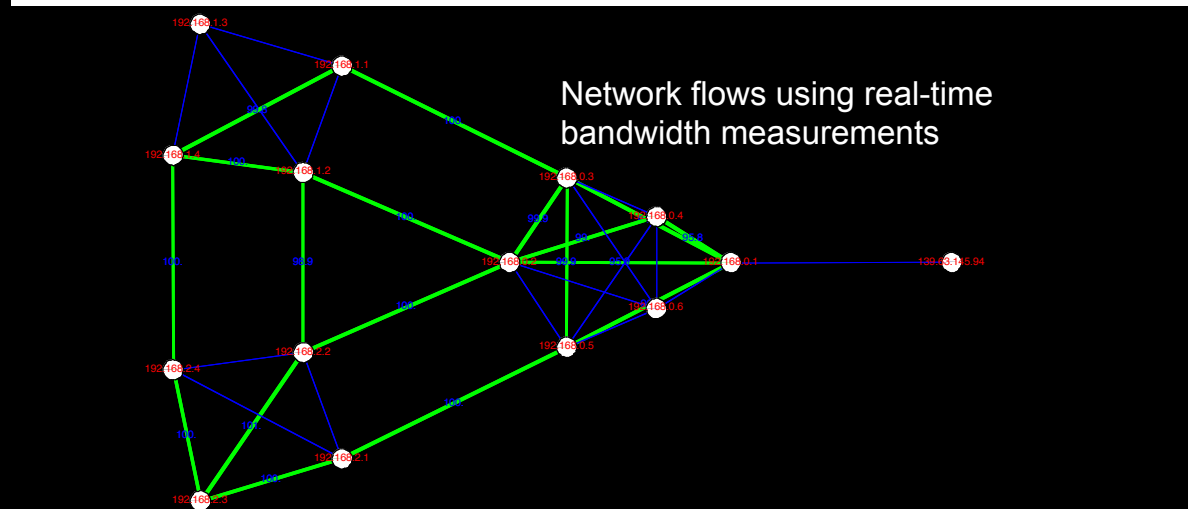
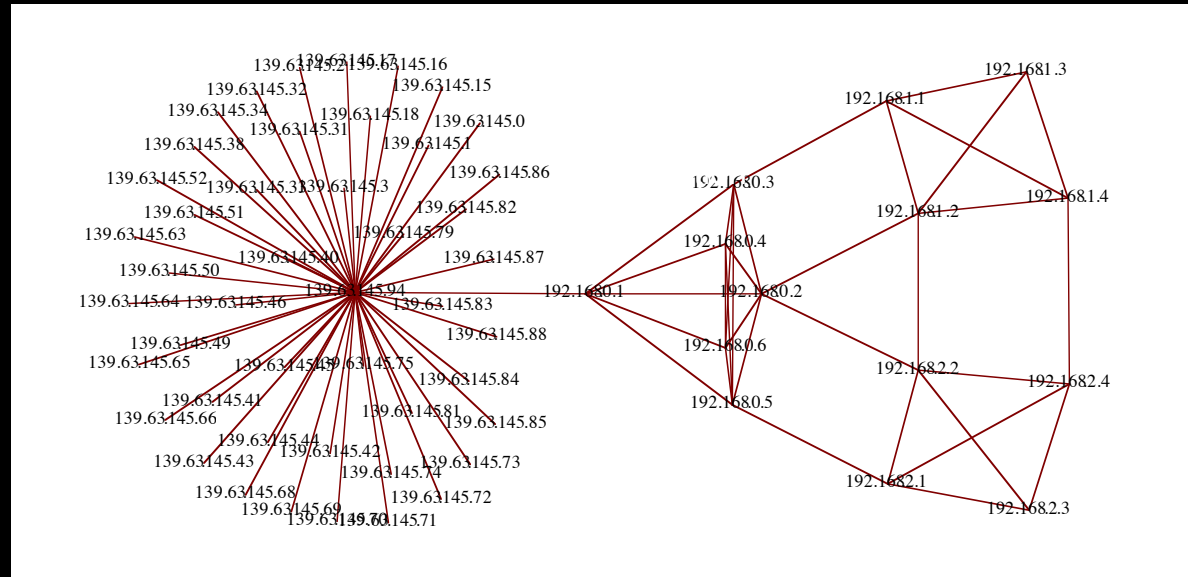
```
Print["Path: ", nodePath];
```

```
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 Virtualized Networks, accepted for publication to the IEEE e-Science 2006 conference Amsterdam.

StarPlane



TouchTable Demonstration @ SC08



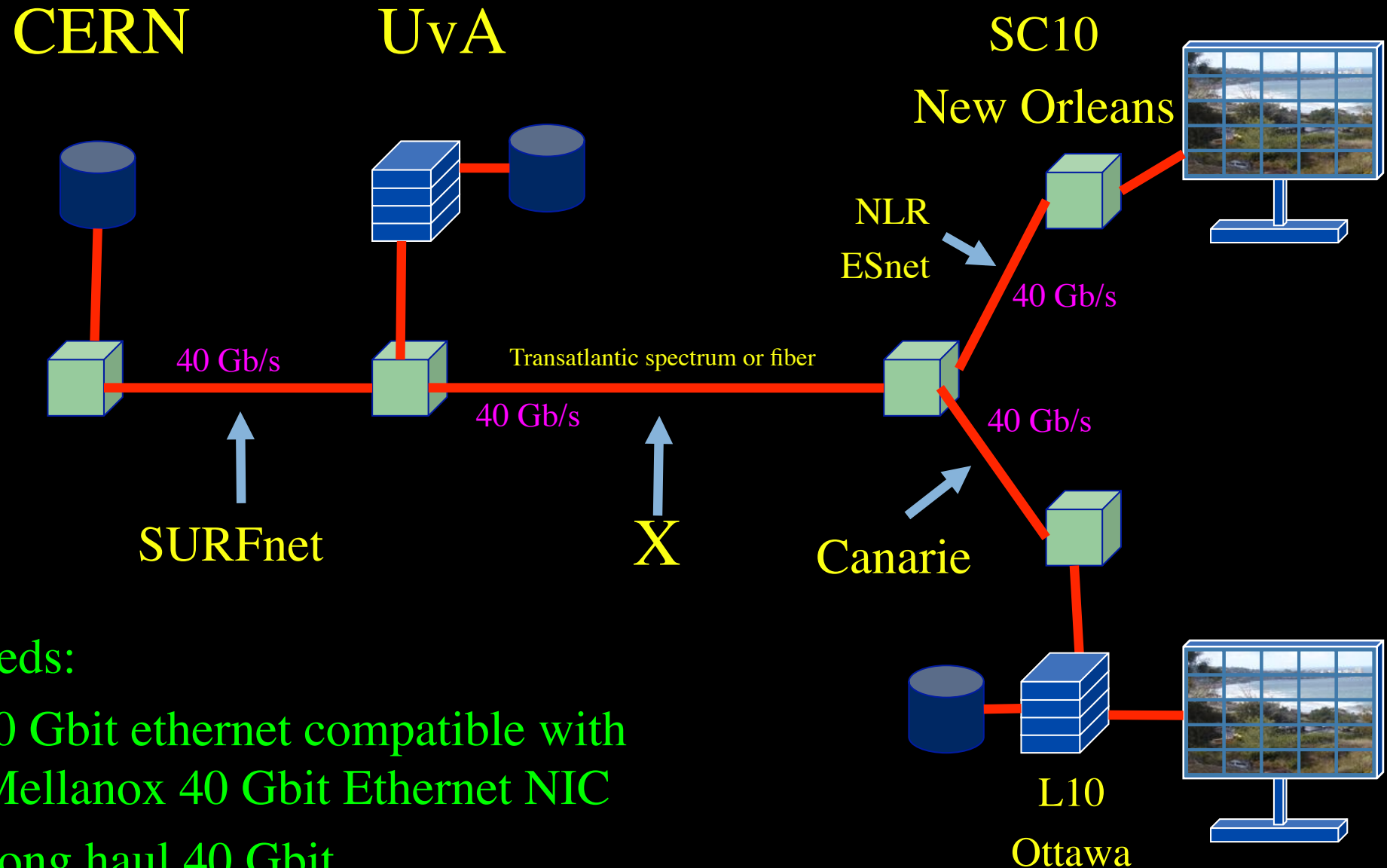
Themes for next years

- 40 and 100 Gbit/s
- Network modeling and simulation
- Cross domain Alien Light switching
- Green-Light
- Network and infrastructure descriptions & WEB2.0
- Reasoning about services
- Cloud Data – Computing - Virtualisation
- Web Services based Authorization
- Network Services Interface (N-S and E-W)
- e-Science integrated services
- Prototyping the Internet Exchange of the Future

Ideas for Collaboration

- Virtualized Network & Cloud node (NOVI)
 - EU project Federica
 - Lightpath based network slices (mac in mac)
 - VM's for app and protocol experimentation
 - Optionally virtualized routers for IP
- Multi layer Internet Exchange (GOLE)
 - Exchange traffic at layer 0, 1, 2
 - Control plane, policy, etc.
- PerfSonar → GreenSonar

Ideas for collaboration @ SC10



Needs:

- 40 Gbit ethernet compatible with Mellanox 40 Gbit Ethernet NIC
- Long haul 40 Gbit

CookReport
feb 2009 and feb-mar 2010

november '08
interview with
Kees Neggers (SURFnet),
Cees de Laat (UvA)

and furthermore
on november '09

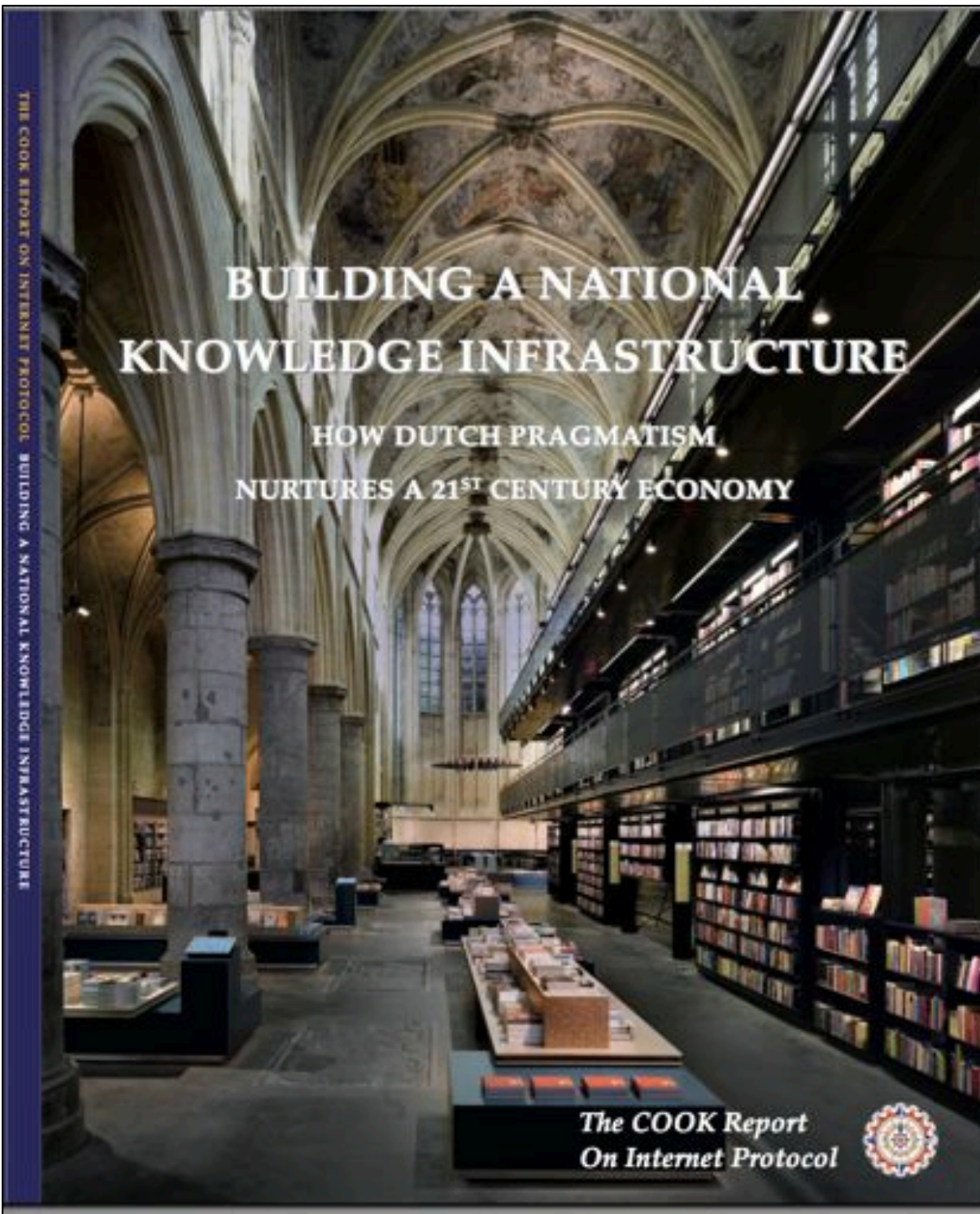
Wim Liebrandt (SURF),
Bob Hertzberger (UvA) and
Hans Dijkman (UvA)



BSIK projects
GigaPort &
VL-e / e-Science



cookreport.com



I did not talk about:

- Token Based Networking
- Privacy & Security
- Authorization, Policy and Trust
- Sensor networks
- Work Flow management

.....

Questions ?

