Lambda-Grid developments

History - Present - Future

Cees de Laat

EU

SURFnet

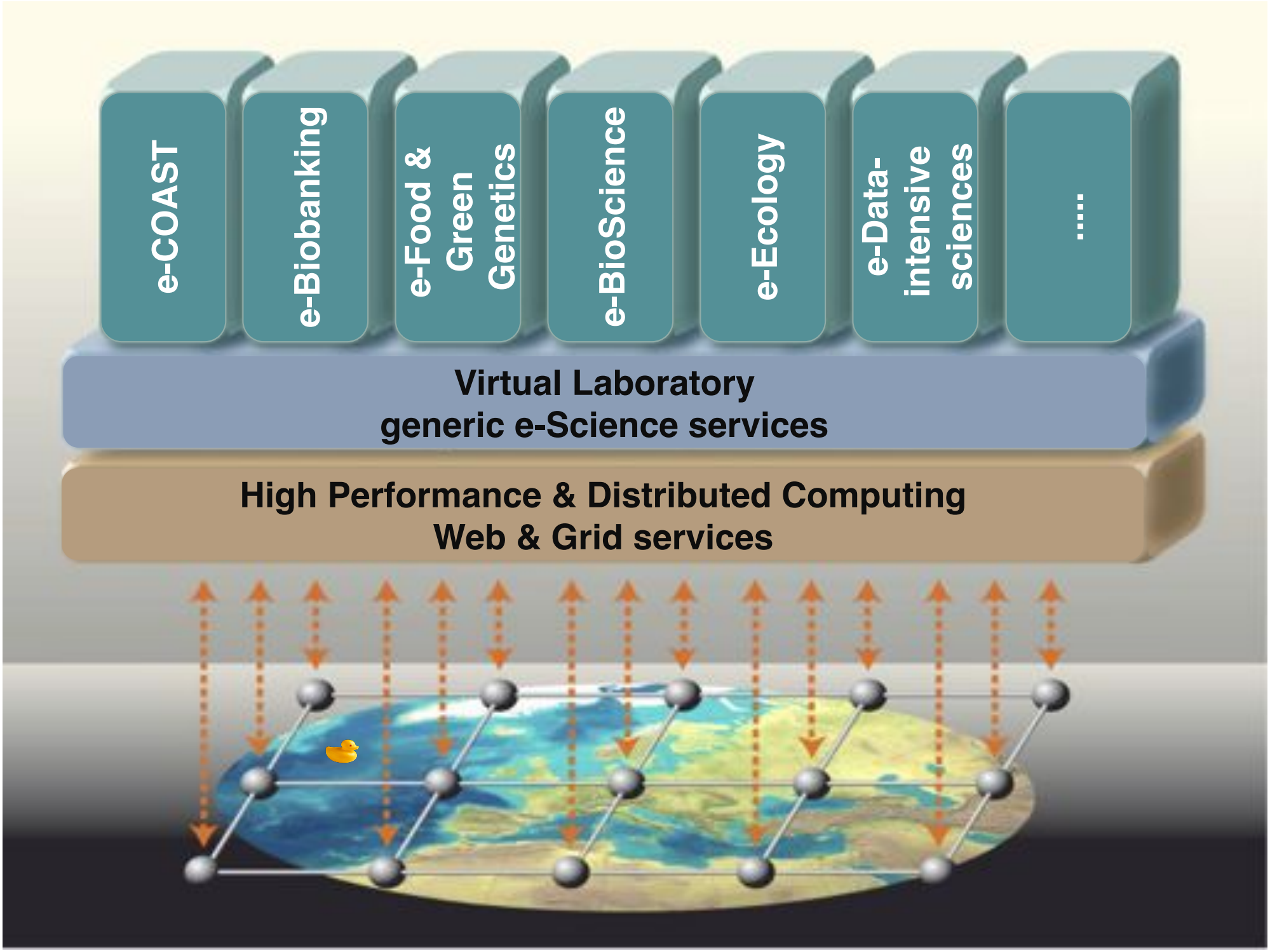
BSIK

NWO

University of Amsterdam

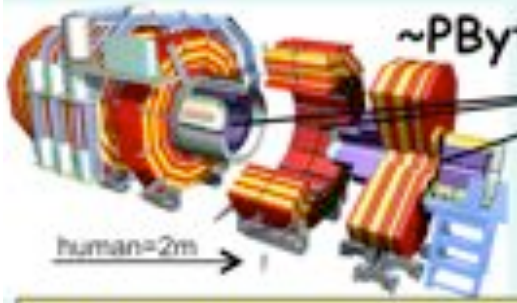
TNO

NCF


e-COAST
e-Biobanking
e-Food & Green Genetics
e-BioScience
e-Ecology
e-Data-intensive sciences
.....
Virtual Laboratory
generic e-Science services
High Performance & Distributed Computing
Web & Grid services

LHC Data Grid Hierarchy

CMS as example, Atlas is similar



~PByte/sec

Online System

~100 MBytes/sec

100000 flops/byte

10 Pflops/s

event simulation

Tier 0 +1

HPSS

event reconstruction

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

analysis

Tier2 Center

Tier 2

~0.6-2.5 Gbps

Tier 3

Institute ~0.25TIPS

Physics data cache

100 - 1000 Mbits/sec

Tier 4

Workstations

Courtesy Harvey Newman, CalTech and CERN

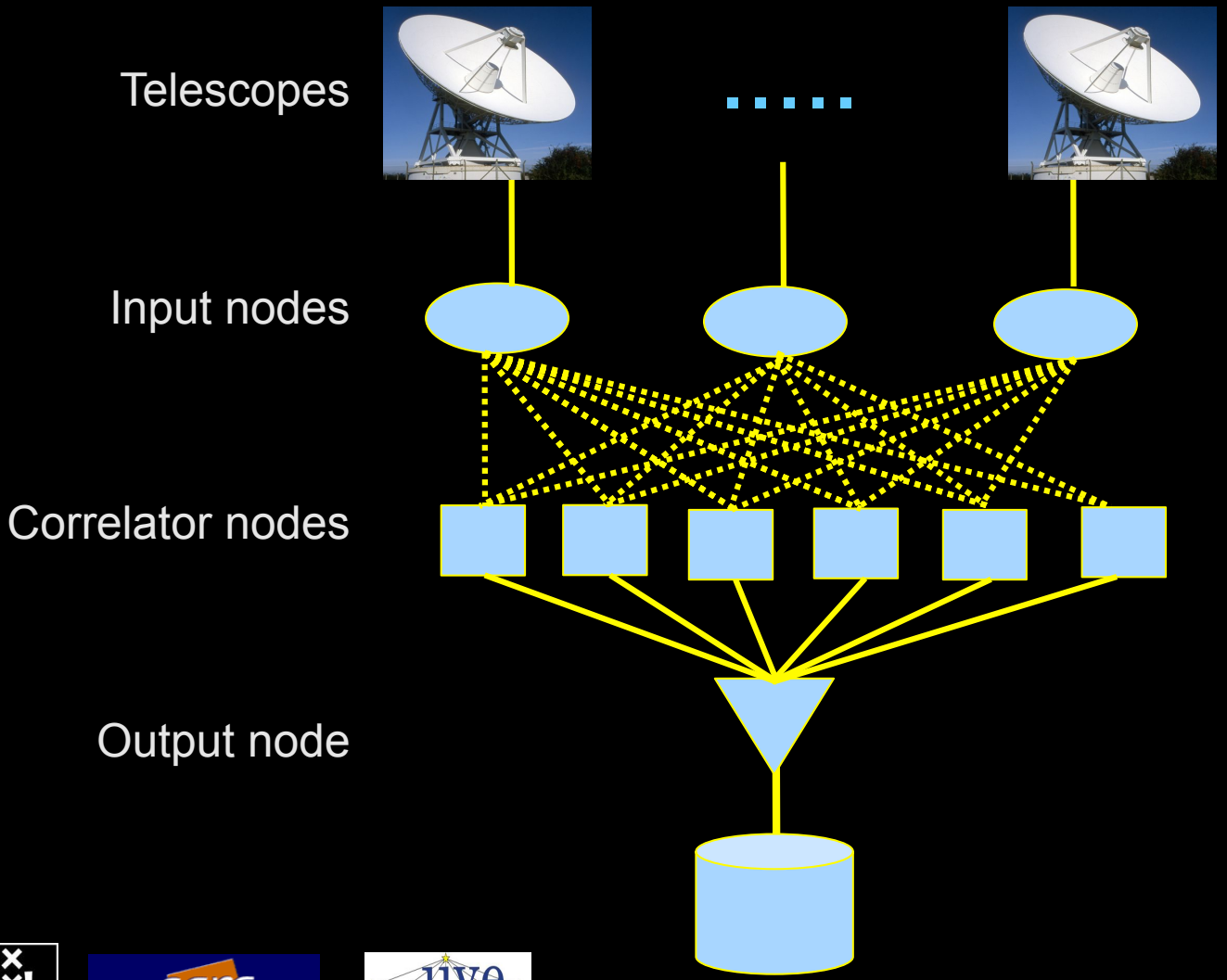
CERN/CMS data goes to 6-8 Tier 1 regional centers, and from each of these to 6-10 Tier 2 centers.

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.

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 !!!

sara Computing & Networking Services

jive JOINT INSTITUTE FOR VLBI IN EUROPE

LOFAR as a Sensor Network

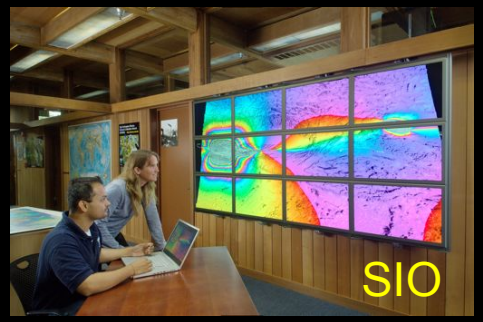
20 flops/byte

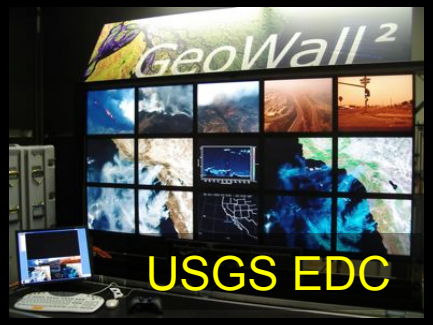
- LOFAR is a large distributed research infrastructure:

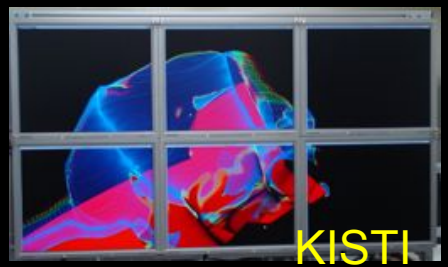
2 Tflops/s

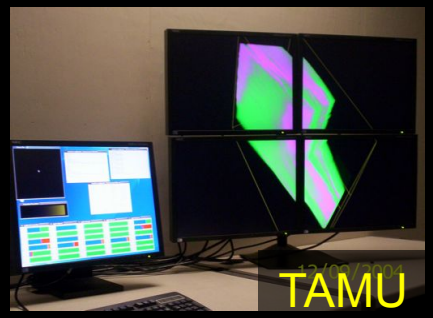
  - Astronomy:
    - >100 phased array stations
    - Combined in aperture synthesis array
    - 13,000 small “LF” antennas
    - 13,000 small “HF” tiles
  - Geophysics:
    - 18 vibration sensors per station
    - Infrasound detector per station
  - >20 Tbit/s generated digitally
  - >40 Tflop/s supercomputer
  - 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

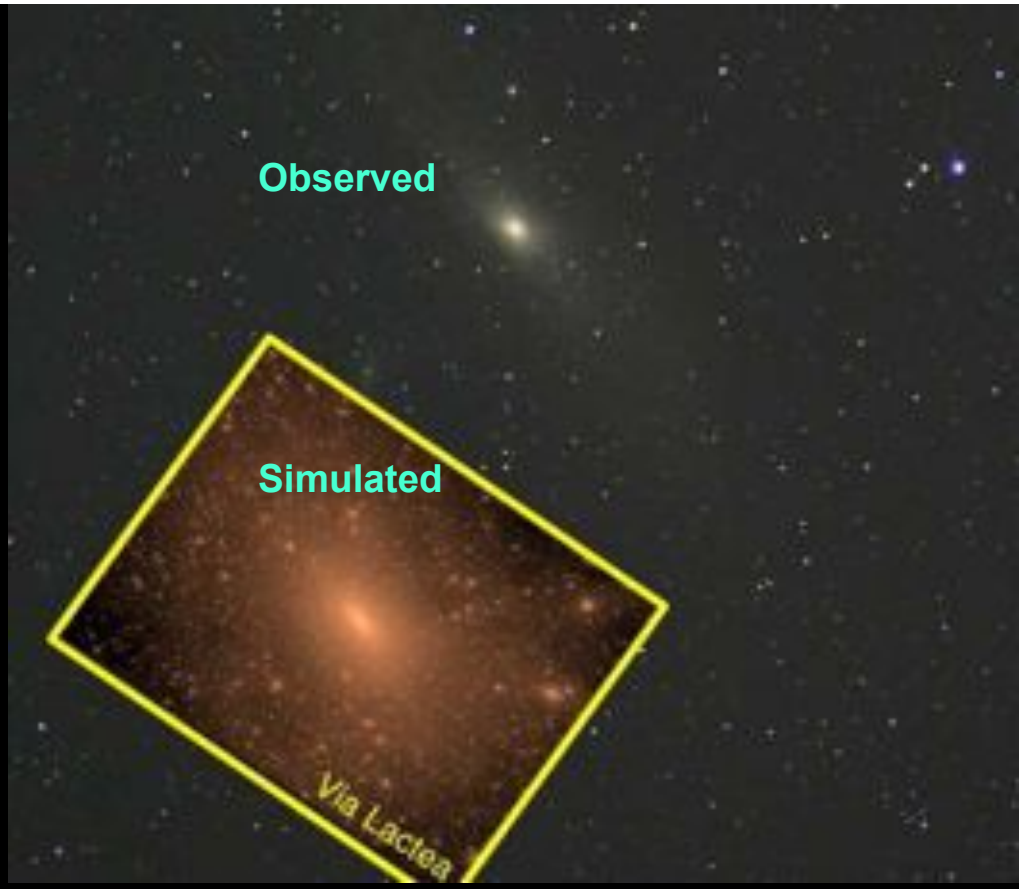
300000 * 60 kb/s * 2 sensors (microphones) to cover all Dutch dikes

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

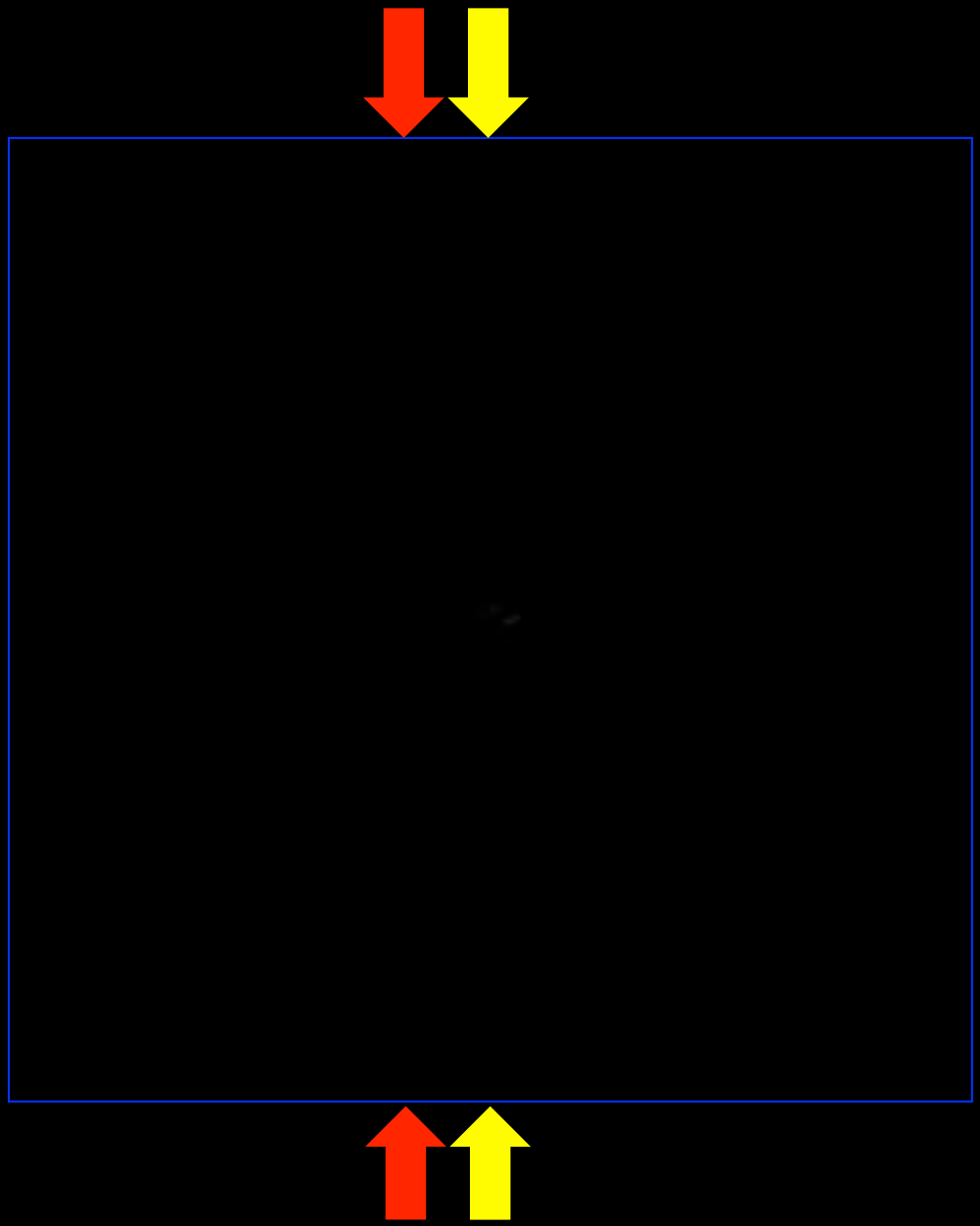
The hardware setup

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

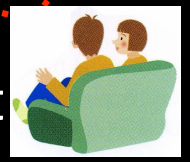
HDTV (2K)

1920

1080

30°

3.0 × Picture Height



UHDTV(8K)

24 Gb/s

7680

4320

0.75 × Picture Height

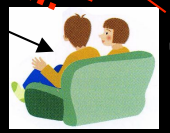
100°

SHD (4K)

3840

2160

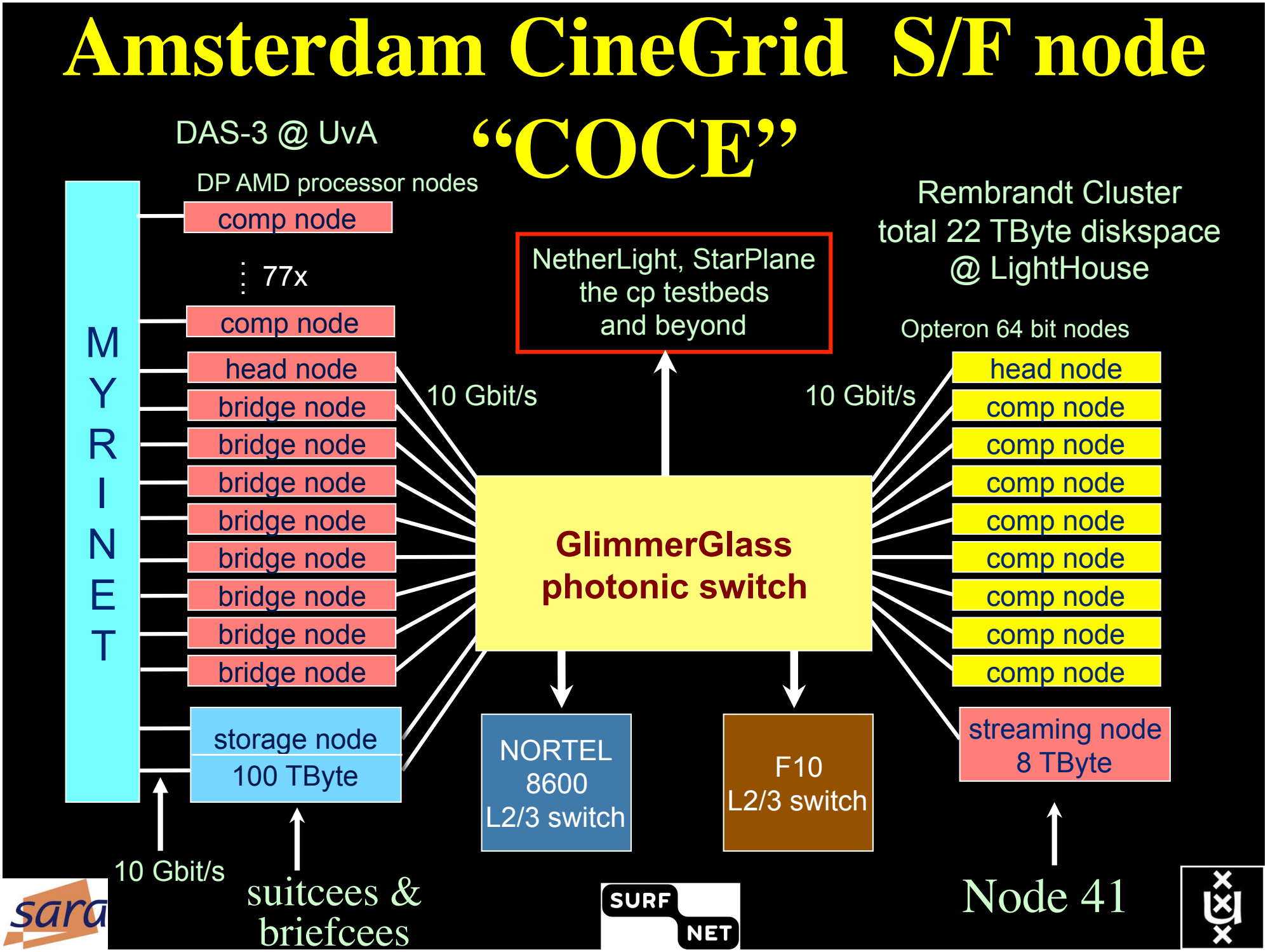
7.6 Gb/s



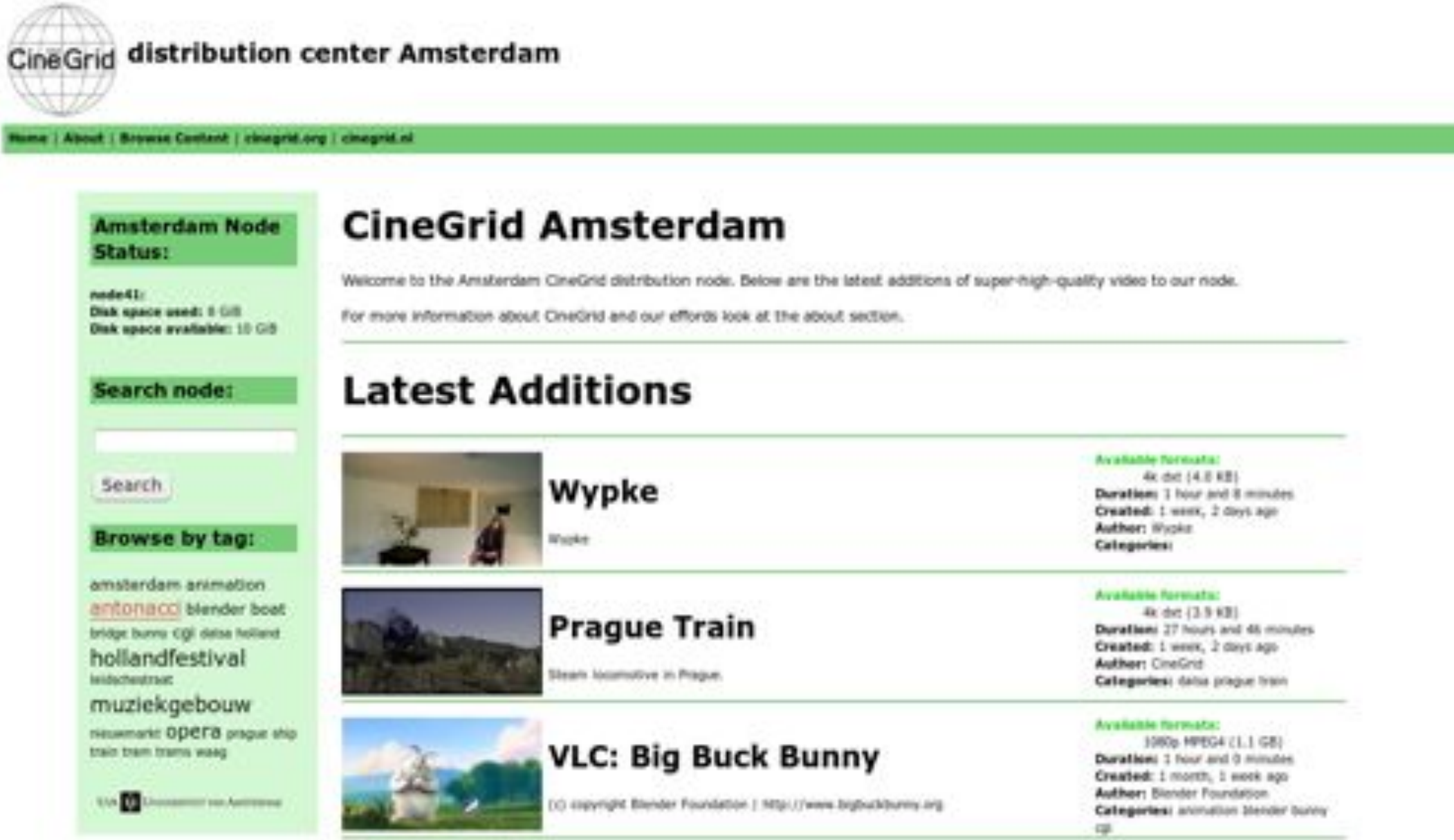
60°

1.5 × Picture Height

Yutaka TANAKA
SHARP CORPORATION
Advanced Image Research Laboratories


Amsterdam CineGrid S/F node
"COCE"
DAS-3 @ UvA
DP AMD processor nodes
comp node
77x
comp node
head node
bridge node
bridge node
bridge node
bridge node
bridge node
bridge node
bridge node
bridge node
storage node
100 TByte
MYRINET
10 Gbit/s
suitcees & briefcees
NetherLight, StarPlane
the cp testbeds
and beyond
10 Gbit/s
GlimmerGlass
photonic switch
NORTEL
8600
L2/3 switch
F10
L2/3 switch
Rembrandt Cluster
total 22 TByte diskspace
@ LightHouse
Opteron 64 bit nodes
10 Gbit/s
head node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
comp node
streaming node
8 TByte
Node 41
sara
SURF NET

CineGrid portal



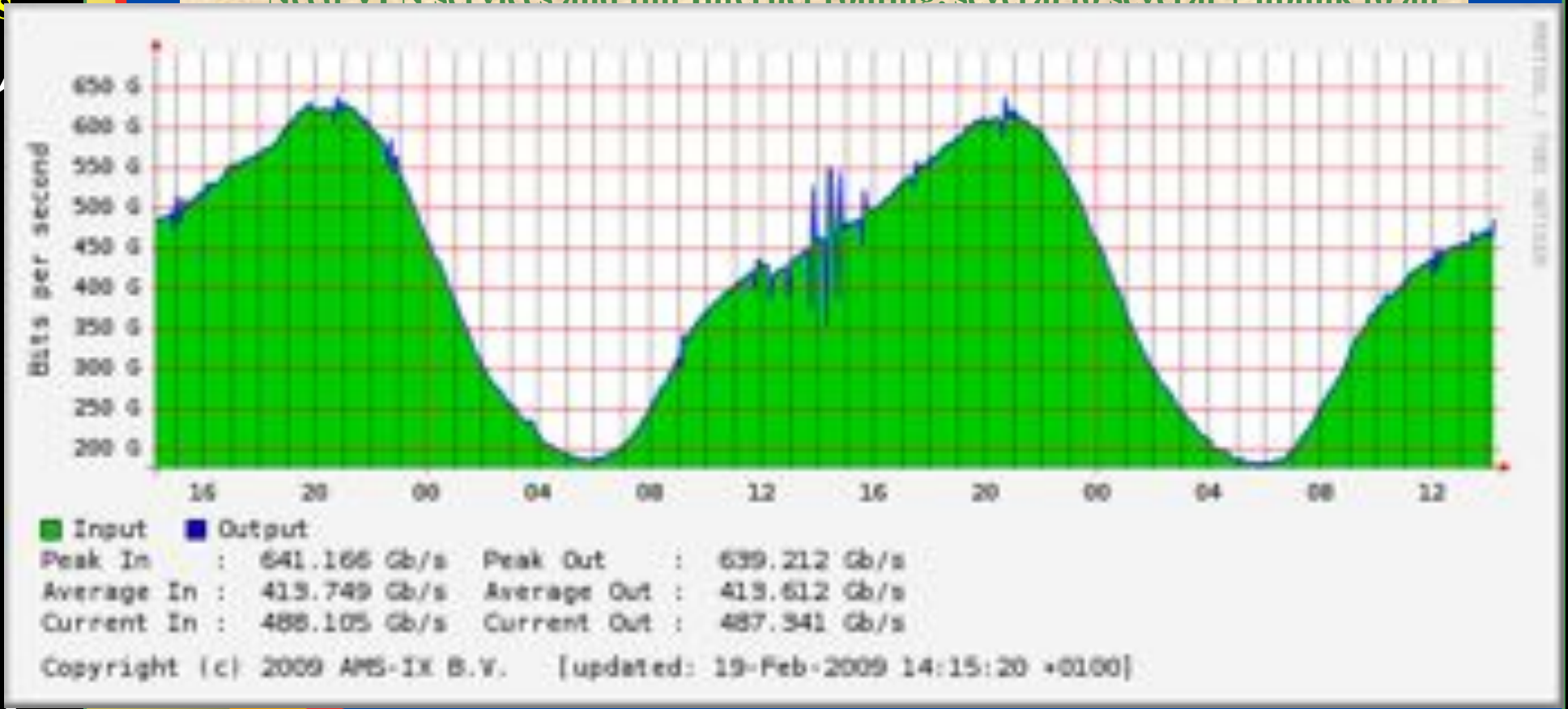
#users

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) → 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!

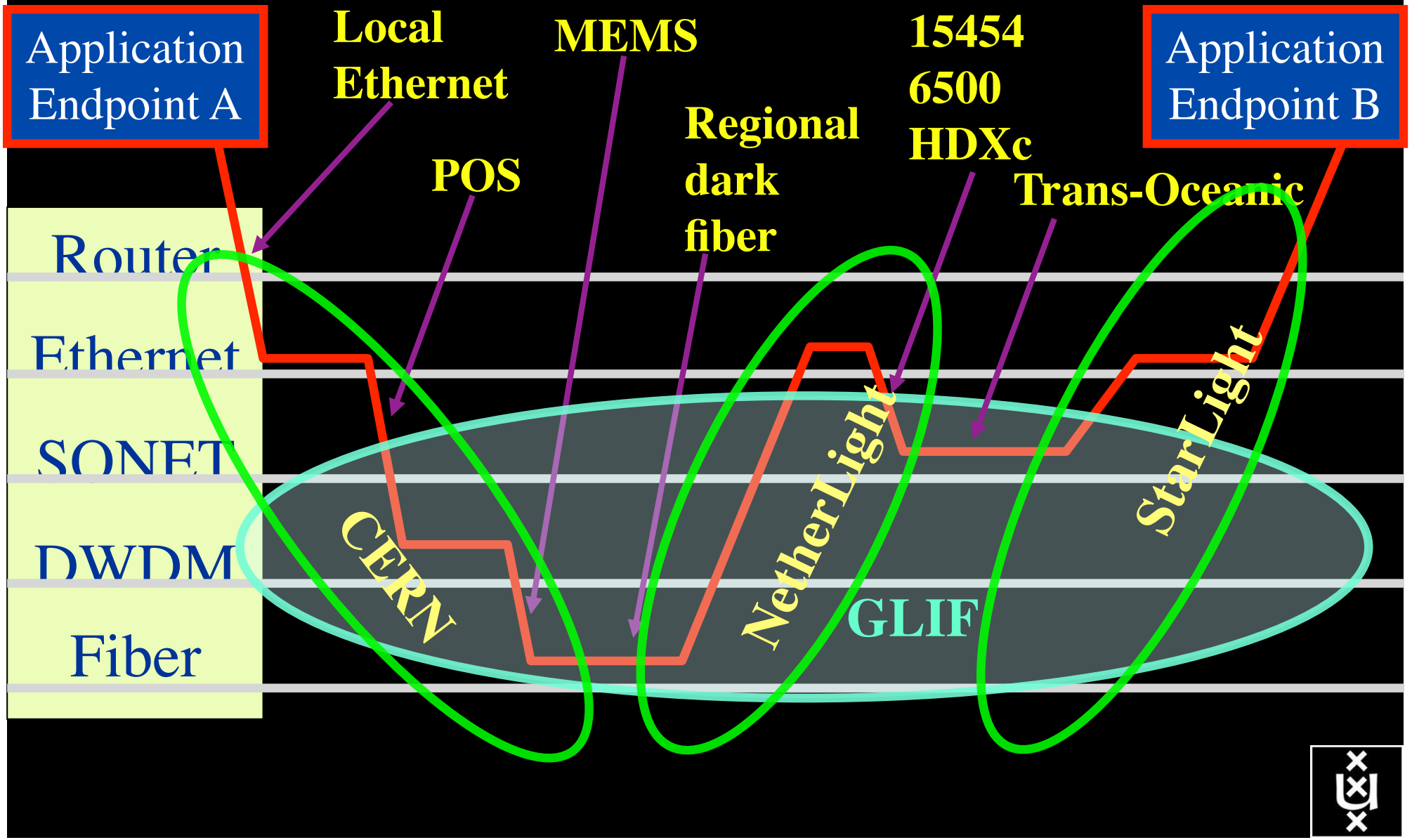
- Costs of photonic equipment 10% of switching 10 % of full routing
  - for same throughput!
  - Photonic vs Optical (optical used for SONET, etc, 10-50 k$/port)
  - DWDM lasers for long reach expensive, 10-50 k$
- Bottom line: look for a hybrid architecture which serves all classes in a cost effective way
  - map A -> L3 , B -> L2 , C -> L1 and L2
- Give each packet in the network the service it needs, but no more !

L1 ≈ 2-3 k$/port

L2 ≈ 5-8 k$/port

L3 ≈ 75+ k$/port

How low can you go?



Hybrid computing

Routers ⇐ ⇒ Supercomputers

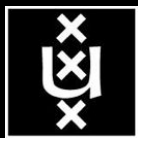
Ethernet switches ⇐ ⇒ Grid & Cloud

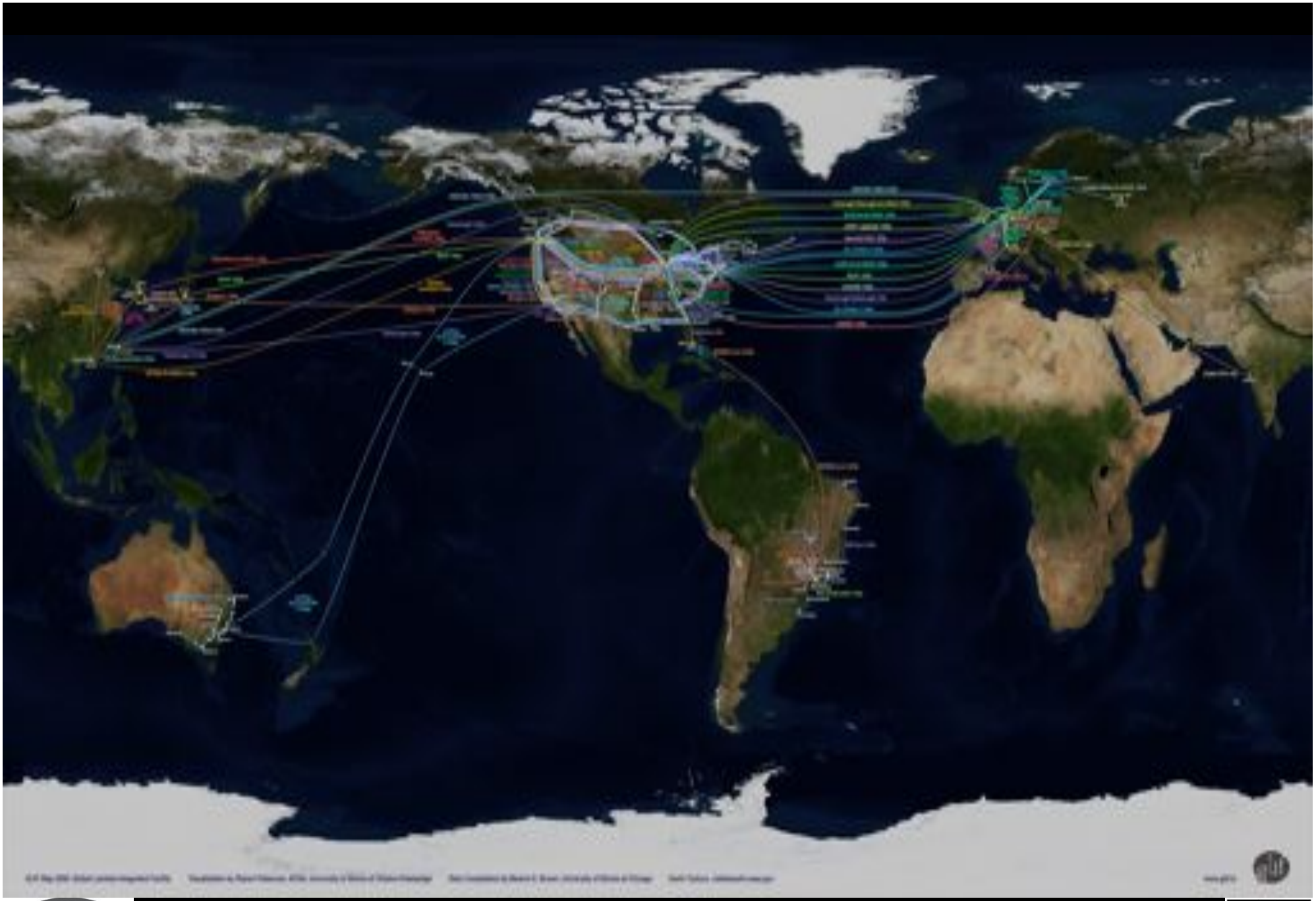
Photonic transport ⇐ ⇒ GPU's

What matters:

Energy consumption/multiplication

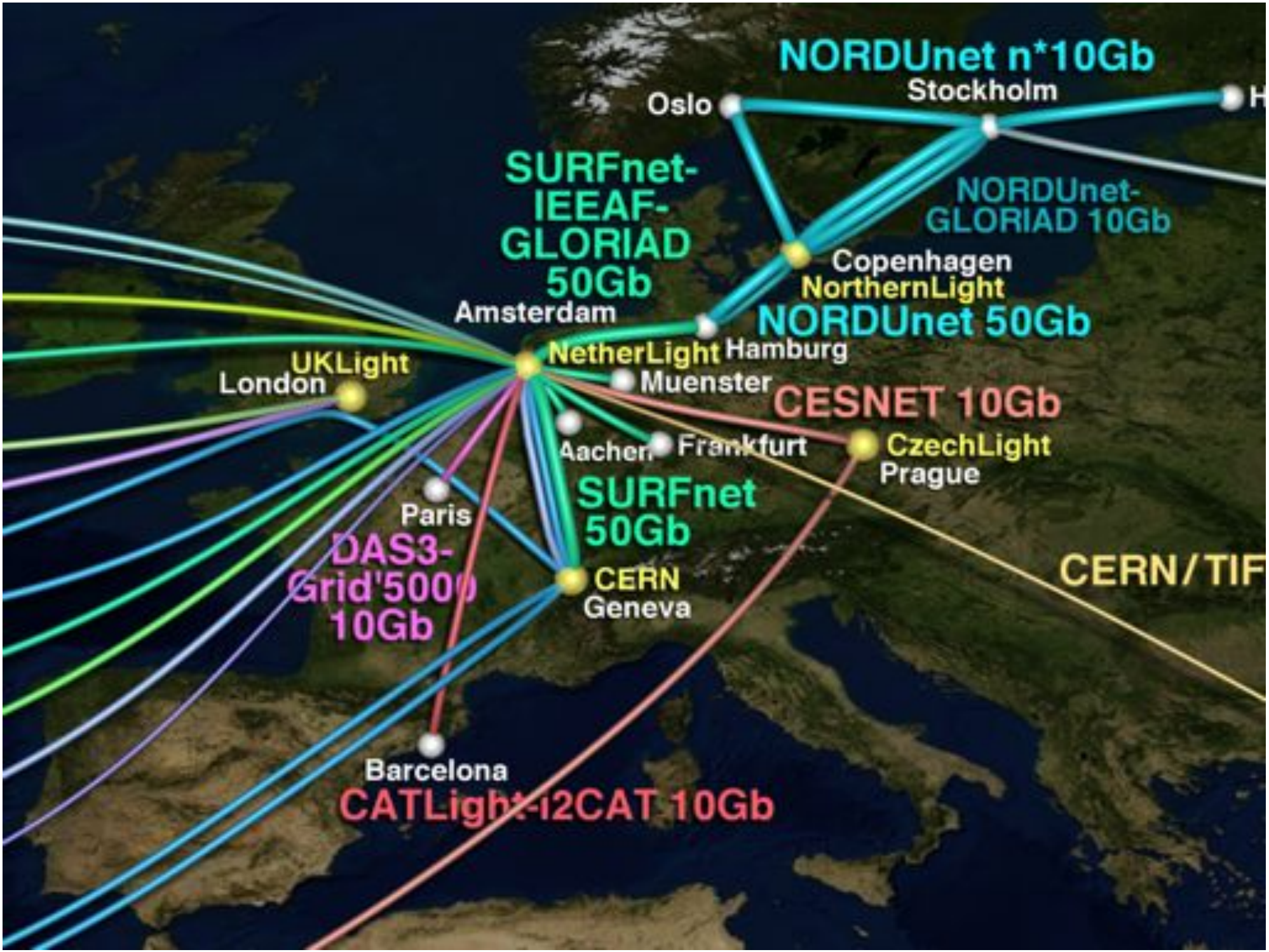
Energy consumption/bit transported

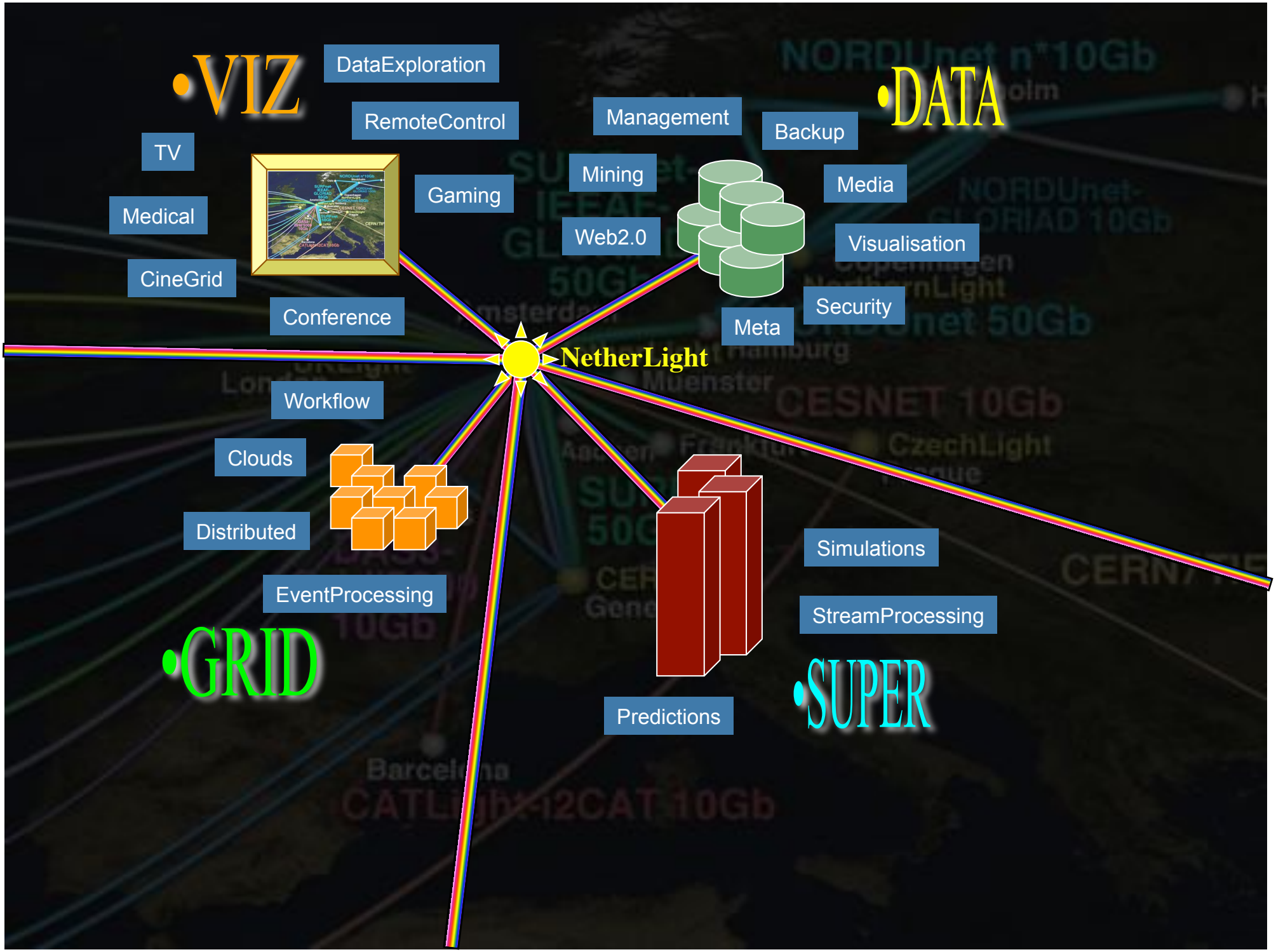




GLIF 2008

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


NORDUnet n*10Gb
Oslo
Stockholm
SURFnet-
IEEAF-
GLORIAD
50Gb
NORDUnet-
GLORIAD 10Gb
Copenhagen
NorthernLight
NORDUnet 50Gb
Amsterdam
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
•DATA
Management
Backup
Mining
Media
Web2.0
Visualisation
Security
Meta
NetherLight
Workflow
Clouds
Distributed
EventProcessing
•GRID
Simulations
StreamProcessing
Predictions
•SUPER

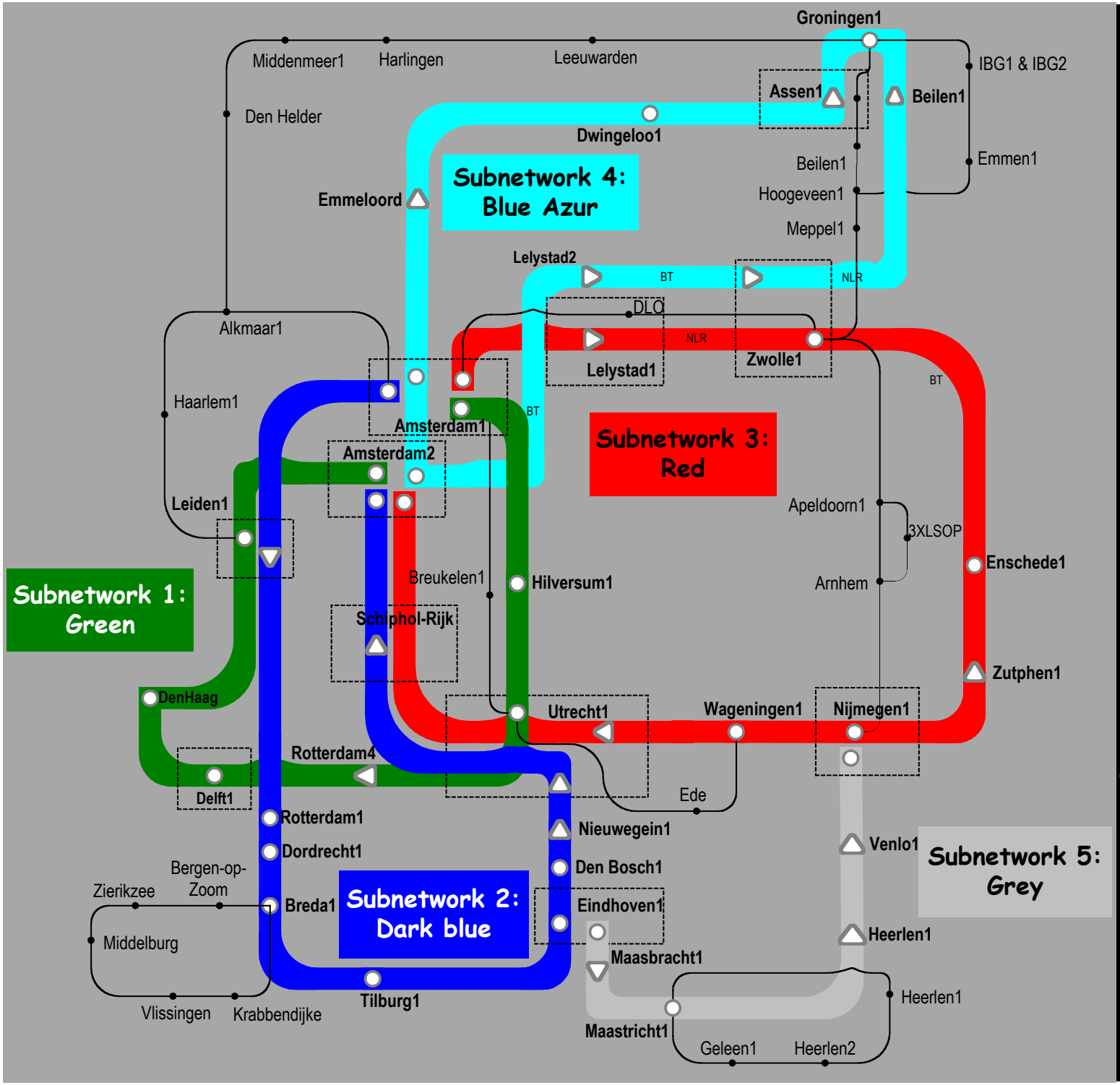
In The Netherlands SURFnet connects between 180:

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

with an indirect ~750K user base

~ 8860 km

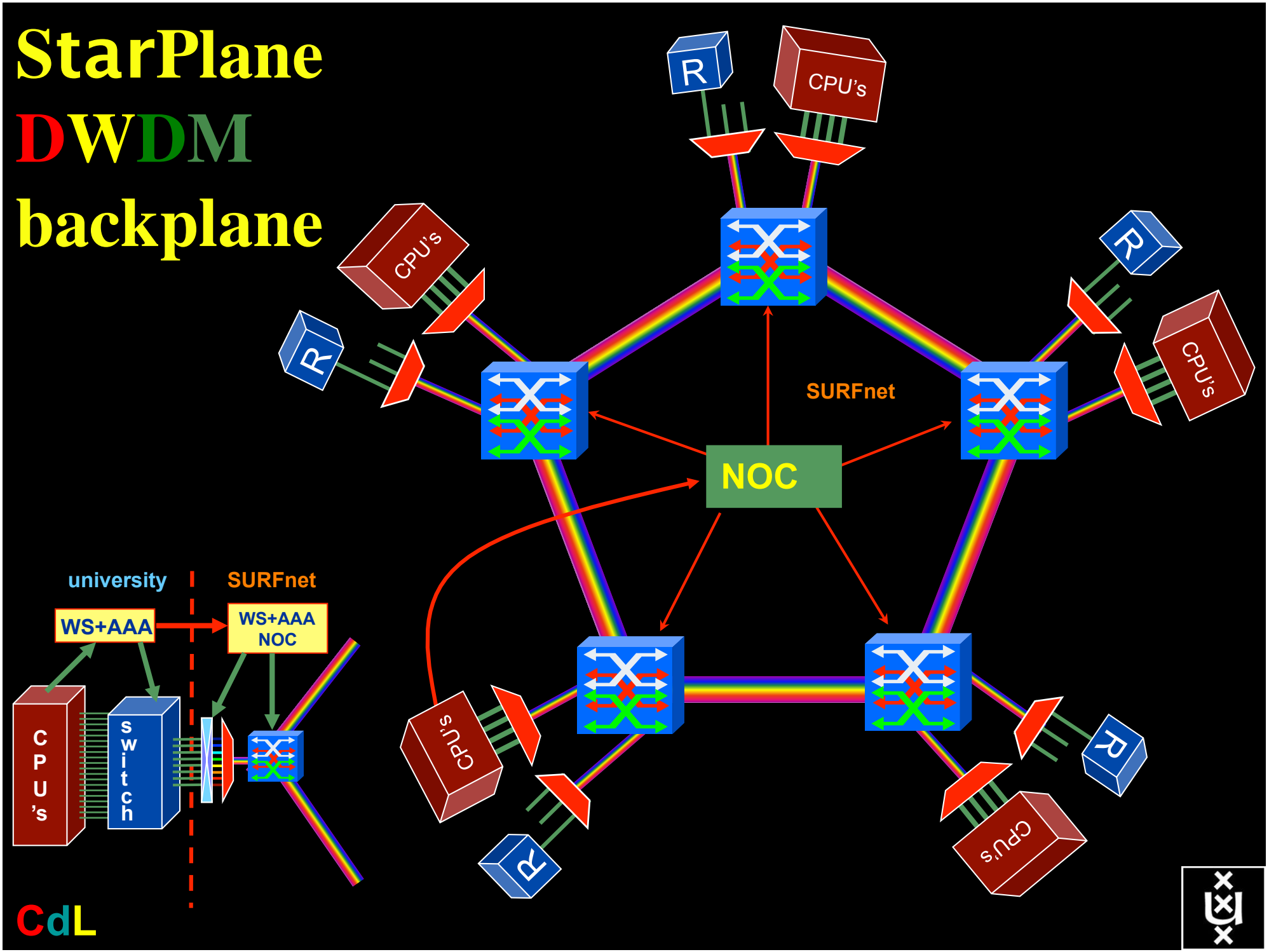
scale comparable to railway system



Common Photonic Layer (CPL) in SURFnet6

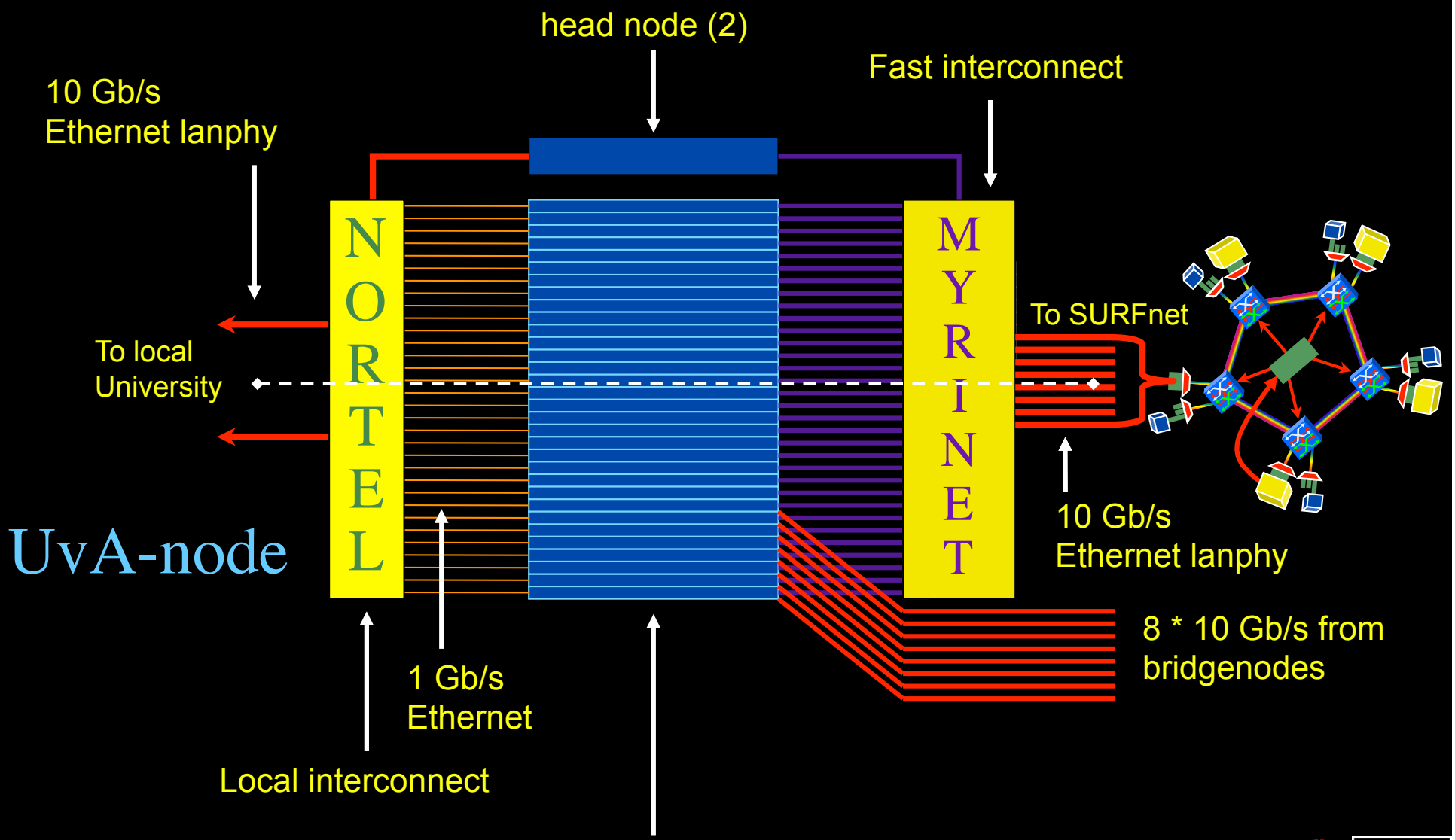
supports up to
72 Lambda's of
10 G each
40 G soon.

StarPlane
DWDM
backplane



CdL

DAS-3 Cluster Architecture

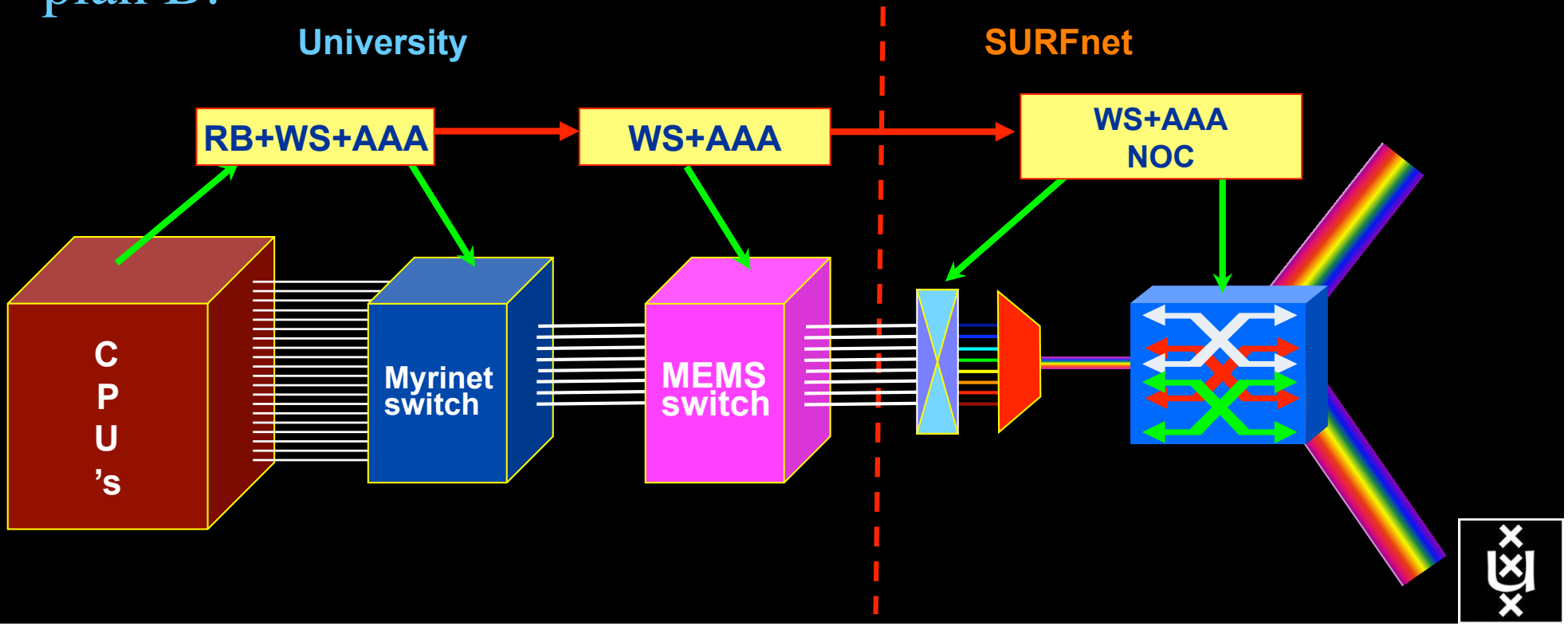


Advanced School for Computing and Imaging

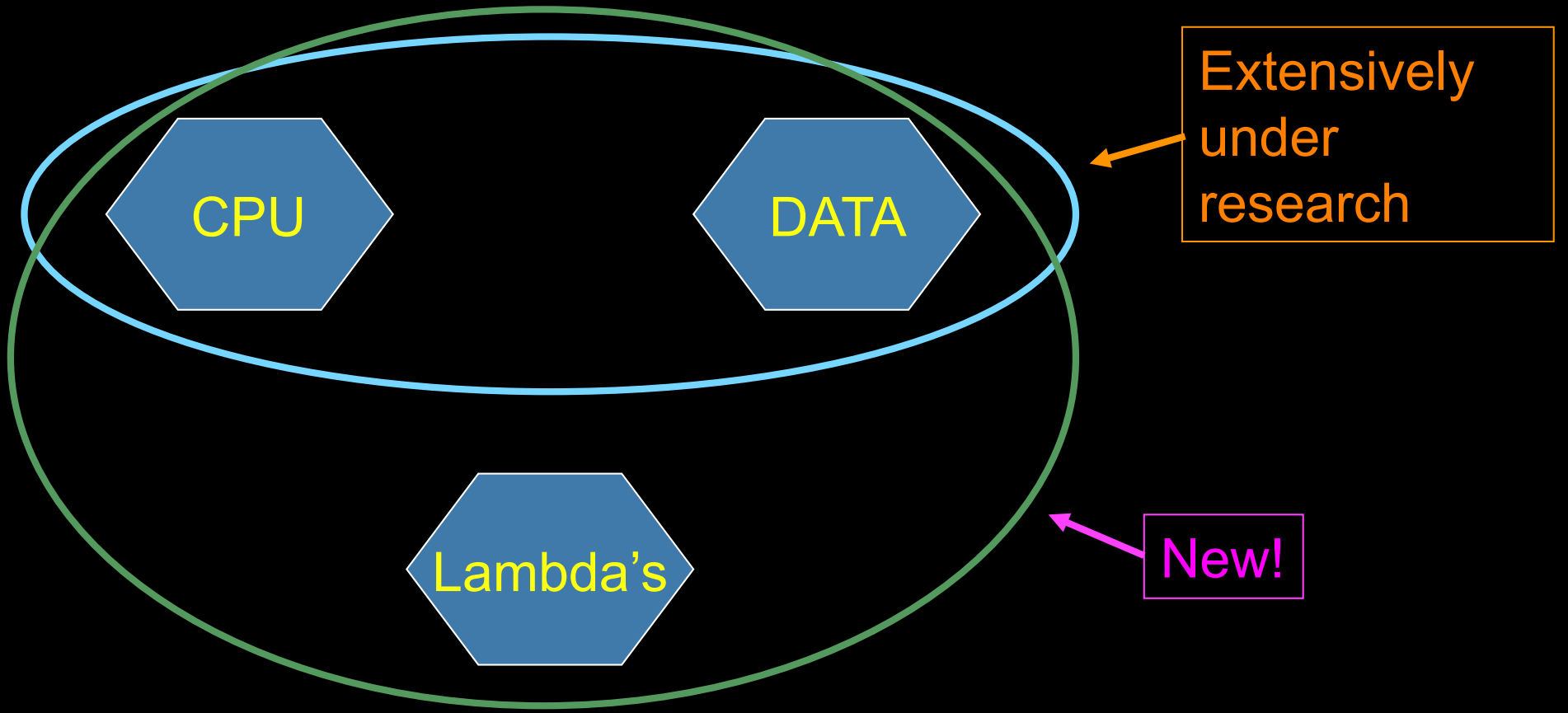
DAS-3

The challenge for sub-second switching

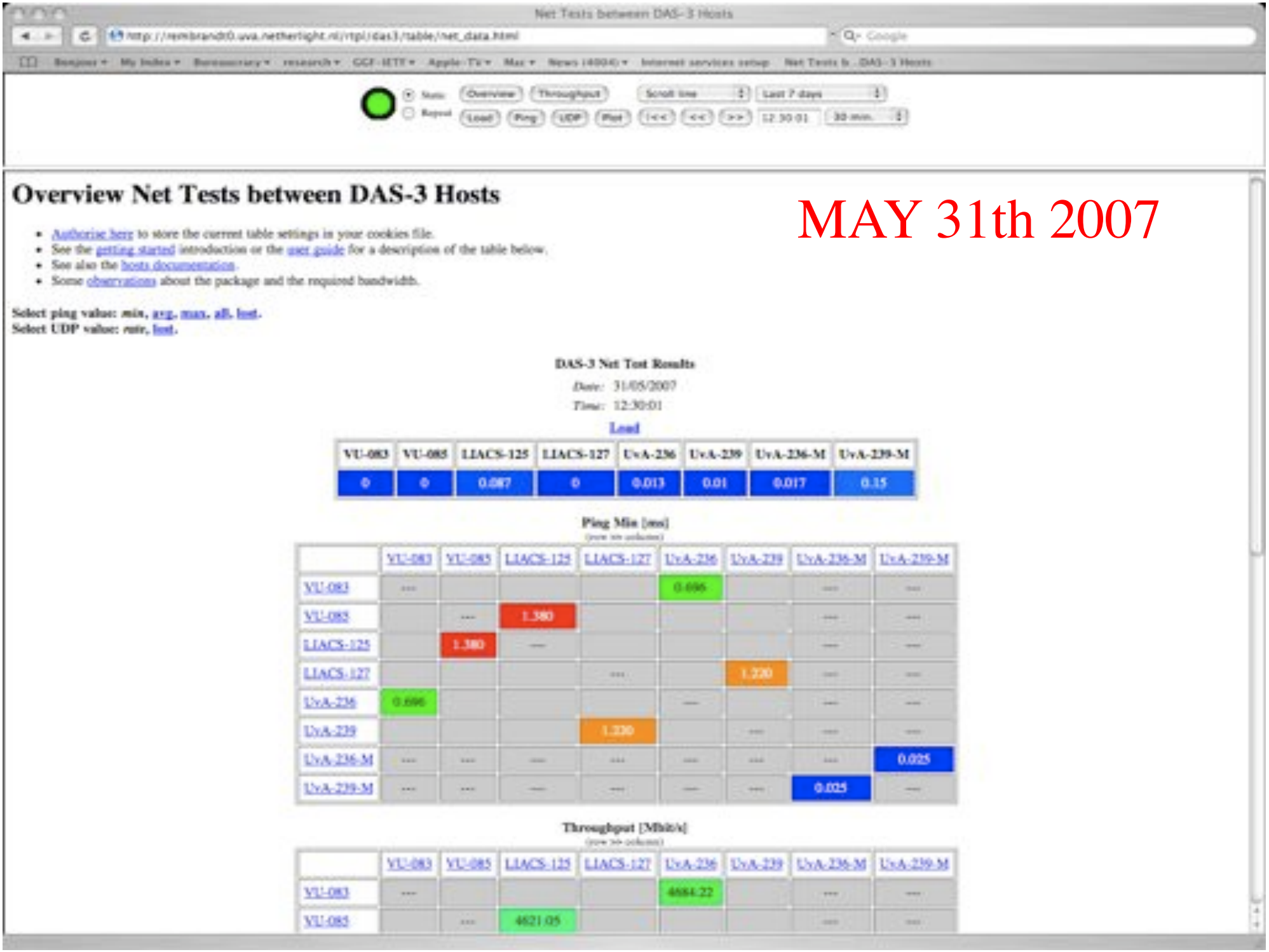
- bringing up/down a λ takes minutes
  - this was fast in the era of old time signaling (phone/fax)
  - λ 2 λ influence (Amplifiers, non linear effects)
  - however minutes is historically grown, 5 nines, up for years
  - working with ~~Nortel~~ CIENA to get setup time significantly down
- plan B:



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.

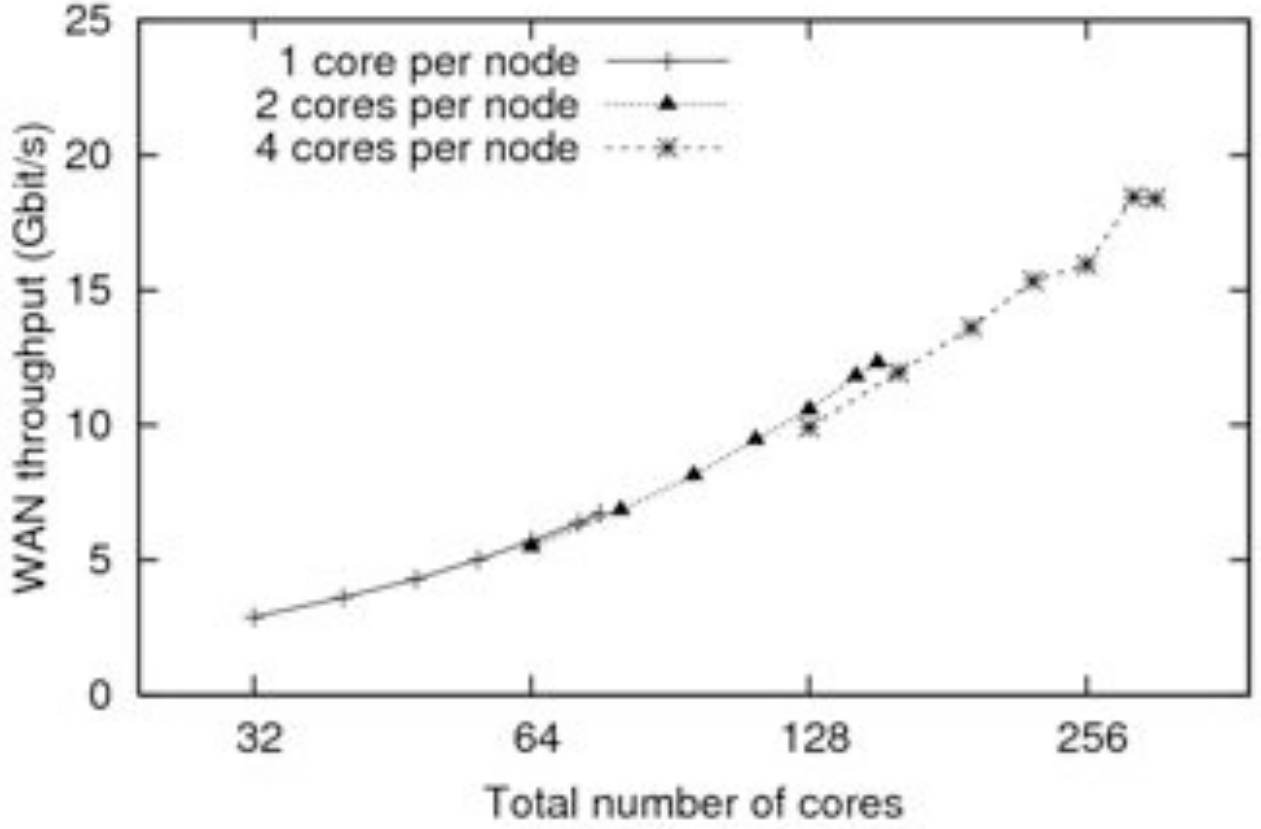

Net Tests between DAS-3 Hosts
Overview
Throughput
Load
Ping
UDP
Plot
Overview Net Tests between DAS-3 Hosts
MAY 31th 2007
Authorise here to store the current table settings in your cookies file.
See the getting started introduction or the user guide for a description of the table below.
See also the hosts documentation.
Some observations about the package and the required bandwidth.
Select ping value: min, avg, max, all, host.
Select UDP value: rate, host.
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

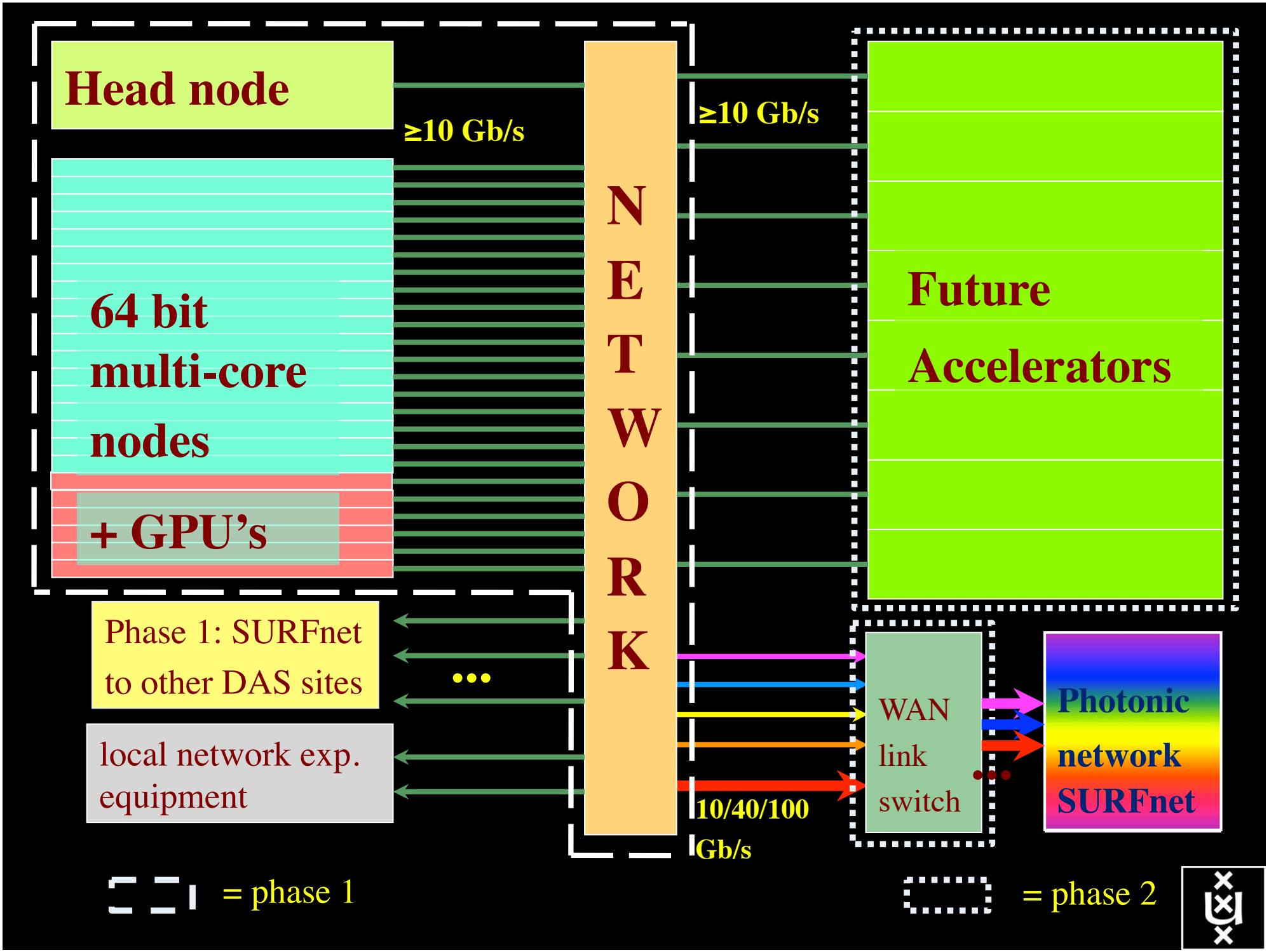
Games and Model Checking

- 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

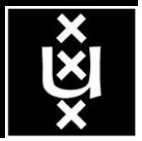



Head node
≥10 Gb/s
64 bit
multi-core
nodes
+ GPU's
N
E
T
W
O
R
K
≥10 Gb/s
Future
Accelerators
Phase 1: SURFnet
to other DAS sites
local network exp.
equipment
WAN
link
switch
Photonic
network
SURFnet
10/40/100
Gb/s
= phase 1
= phase 2

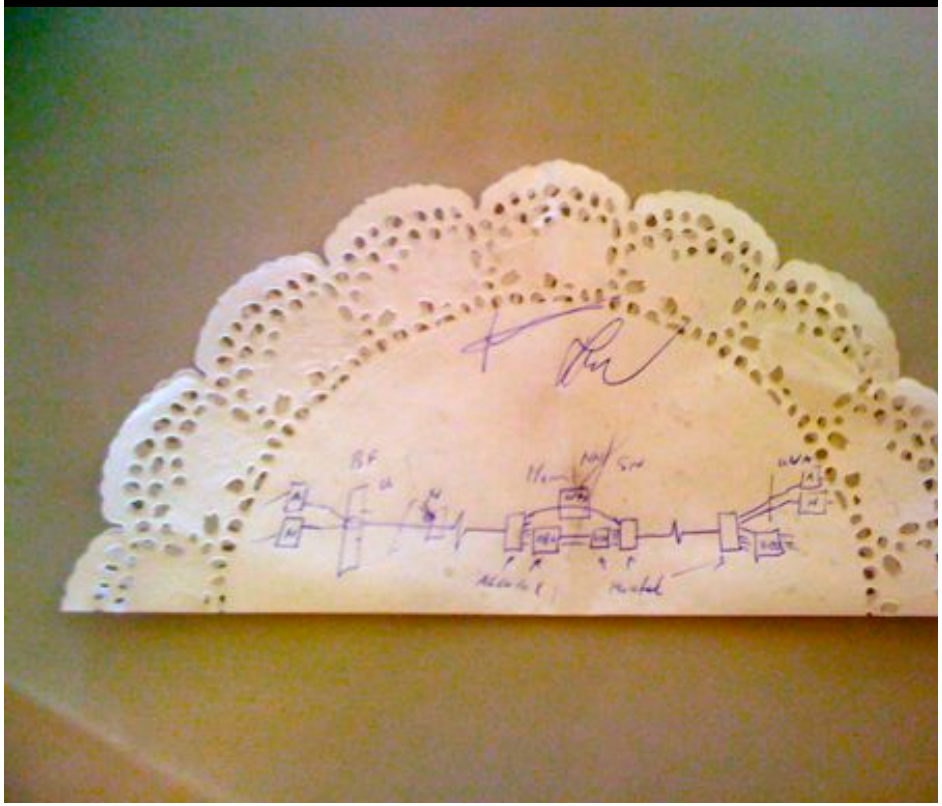
Power is a big issue

- 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!



Alien light From idea to realisation!



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

NCF

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

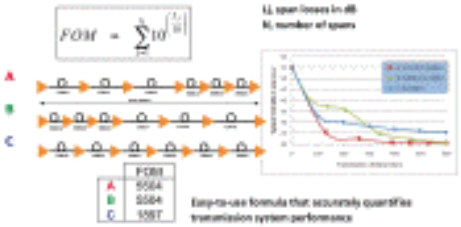
Alien wavelength challenges

- 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.

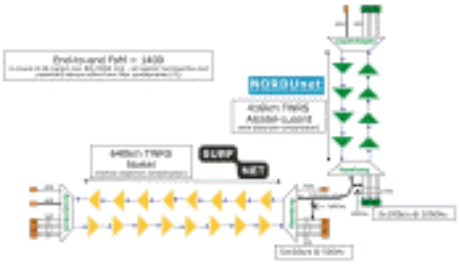
New method to present fiber link quality, FoM (Figure of Merit)

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 10^-15

Conclusions

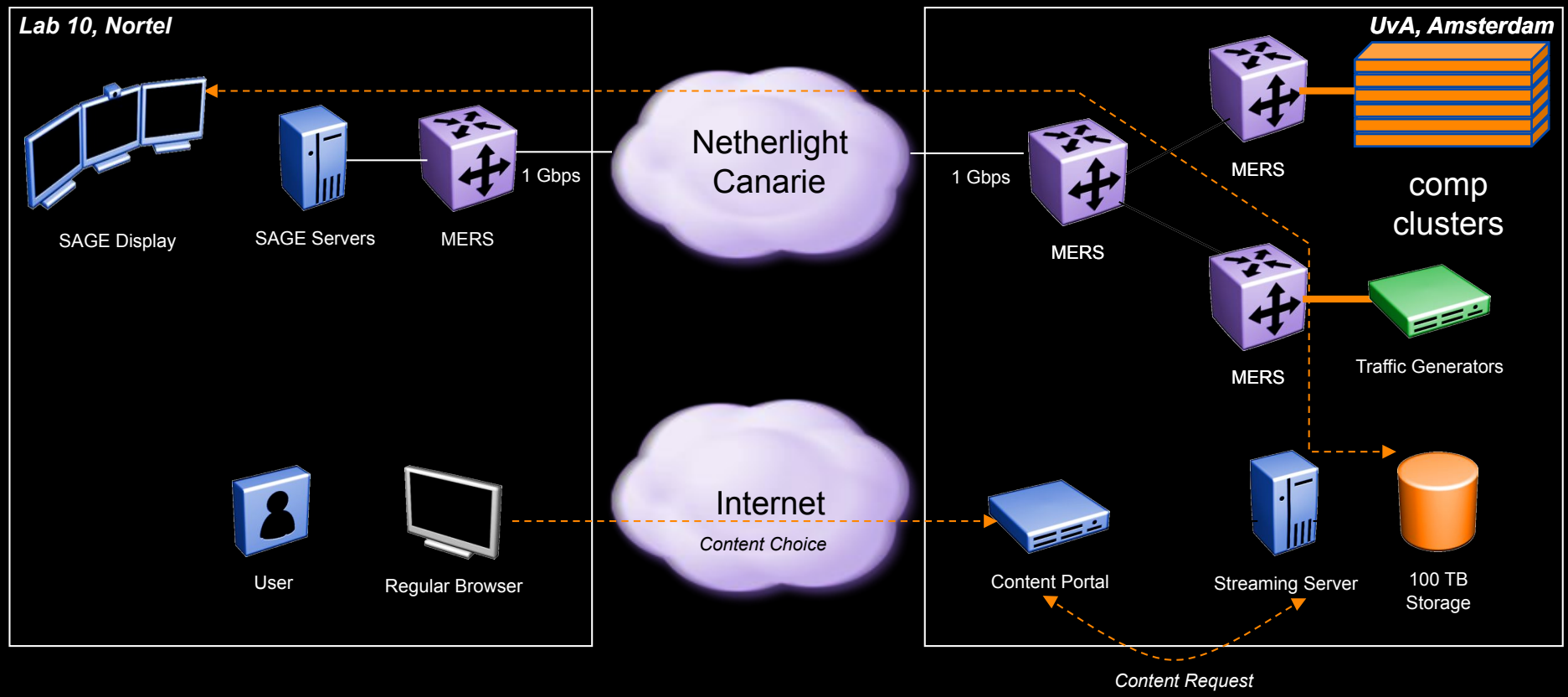
- We have investigated experimentally the all-optical transmission of a 40Gb/s PM-QPSK alien wavelength via a concatenated native and third party DWDM system that both were carrying live 10Gb/s wavelengths.
- The end-to-end transmission system consisted of 1056 km of TWRS (TrueWave Reduced Slope) transmission 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.

NORTEL NORDUnet SURF NET

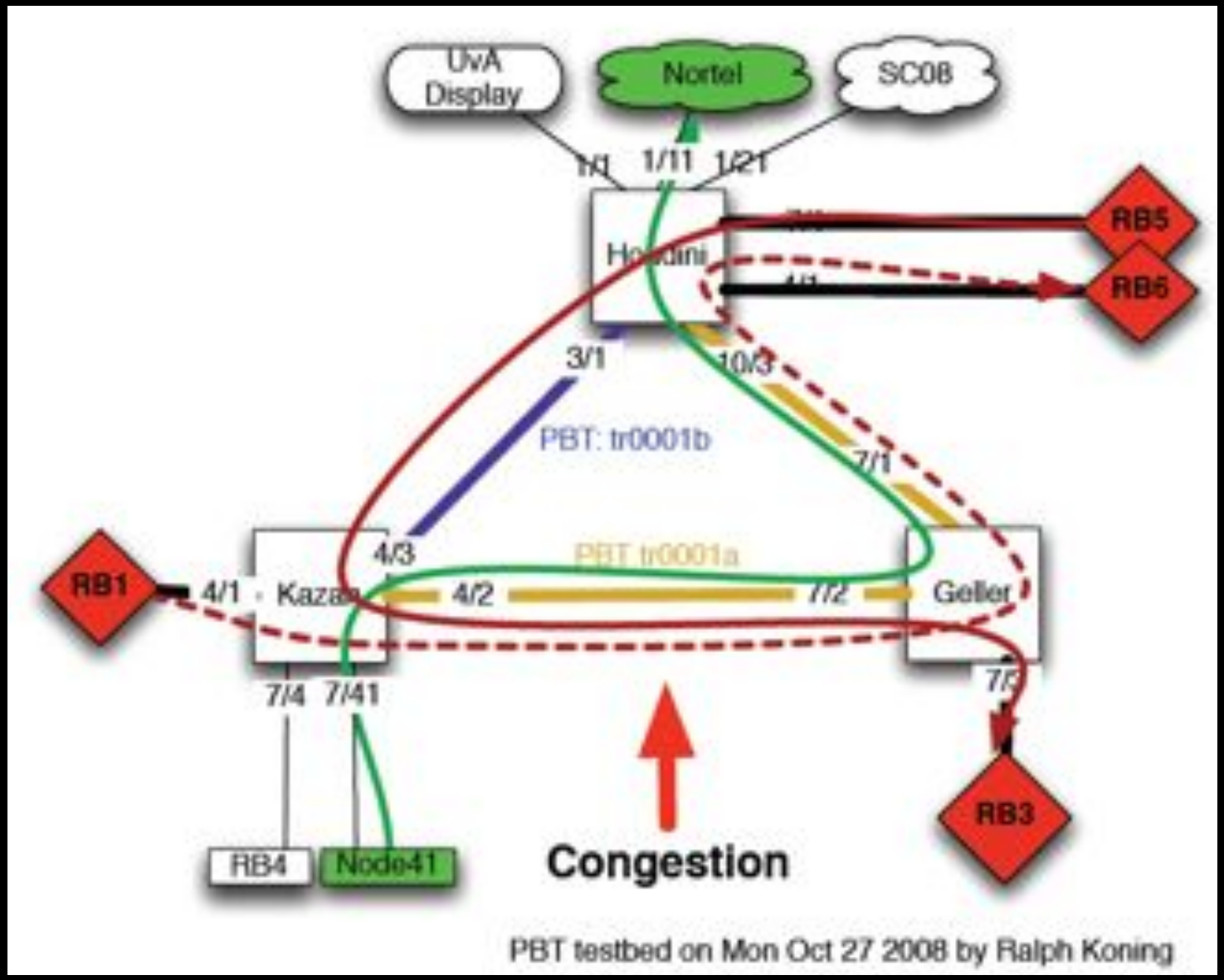
REFERENCES

ACKNOWLEDGEMENTS

Diagram for SAGE video streaming to ATS



UvA Testbed

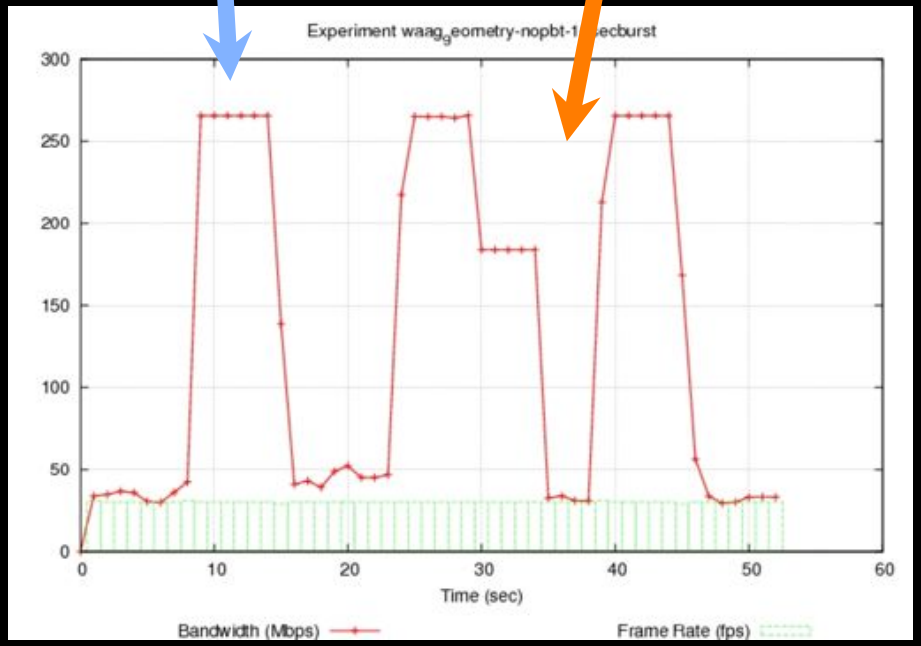


Congestion introduced in the network with multiple PBT paths carrying streamed SHD Content

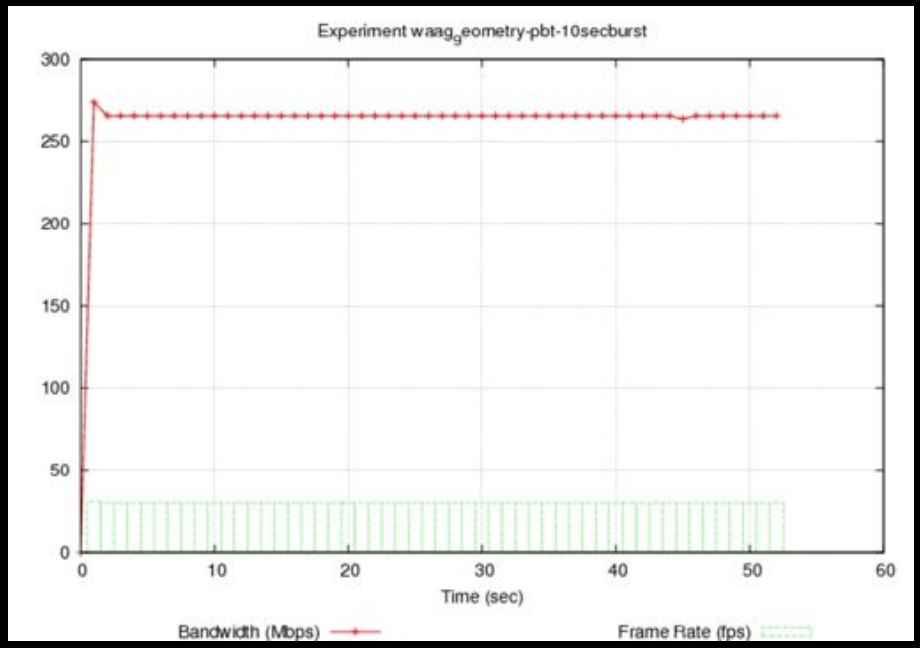
Experimental Data

Sage without background traffic

Sage with background traffic



10 Second Traffic bursts with No PBT

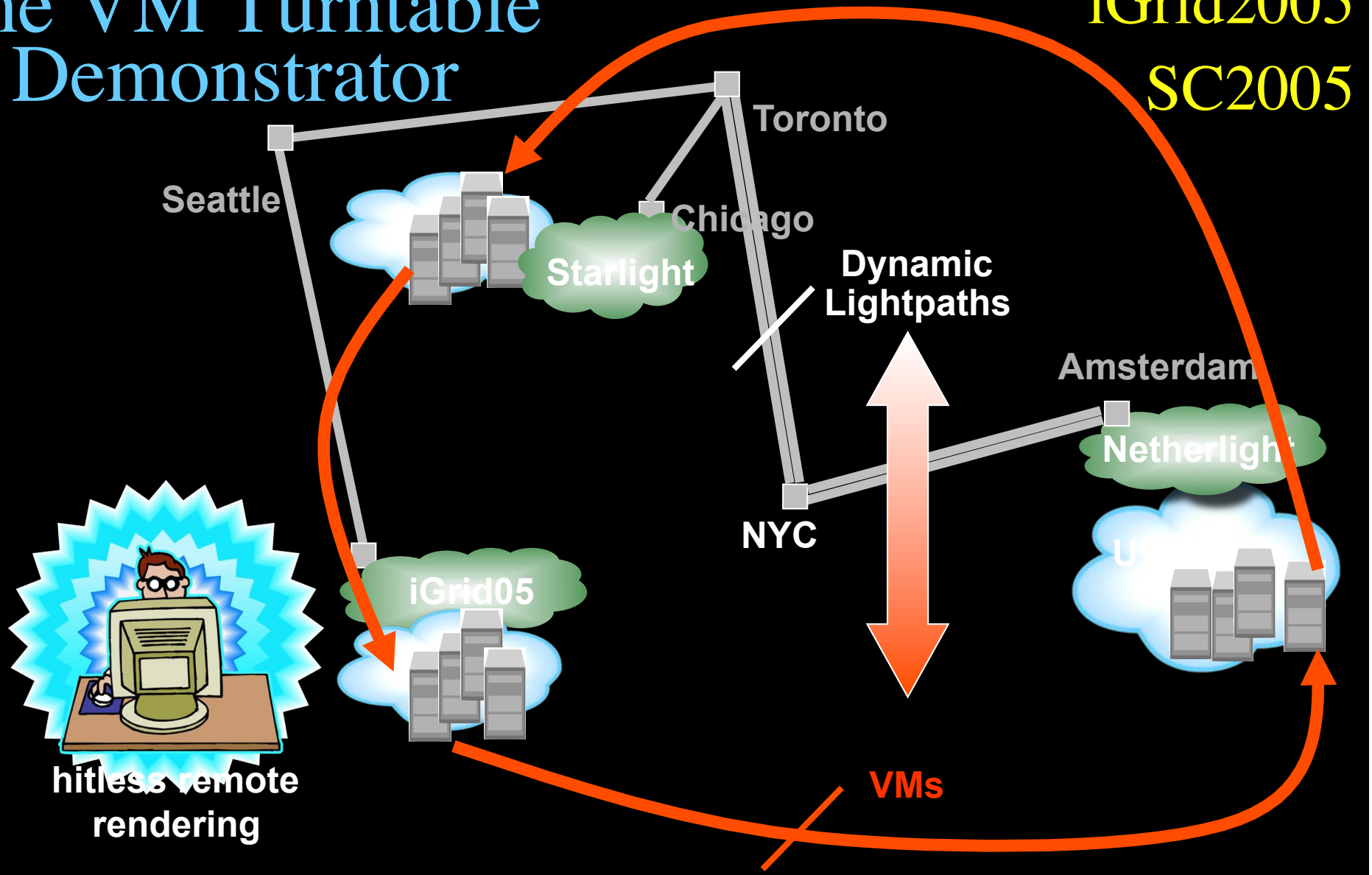


10 Second Traffic bursts with PBT

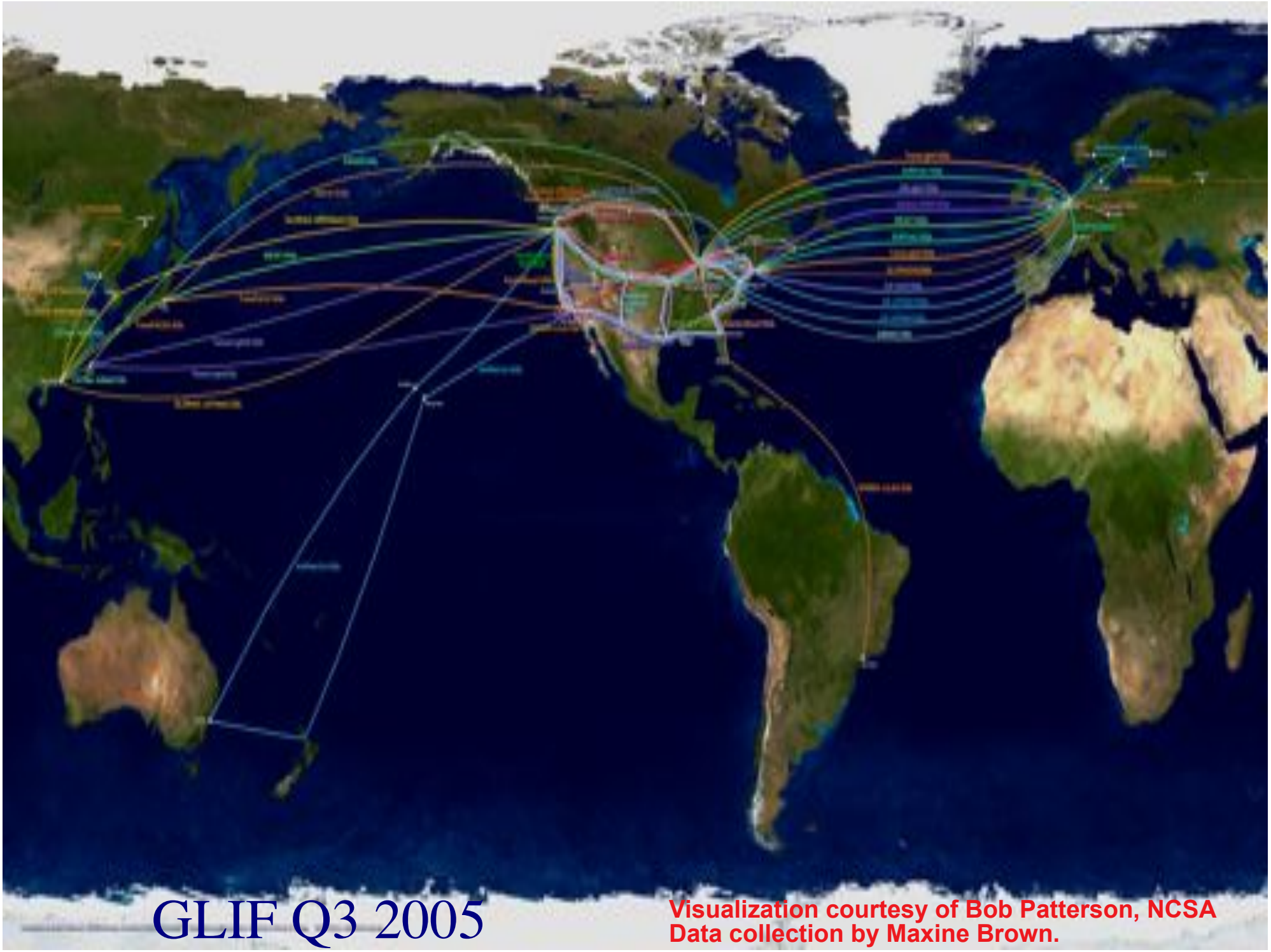
PBT is *SIMPLE* and *EFFECTIVE* technology to build a shared Media-Ready Network

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

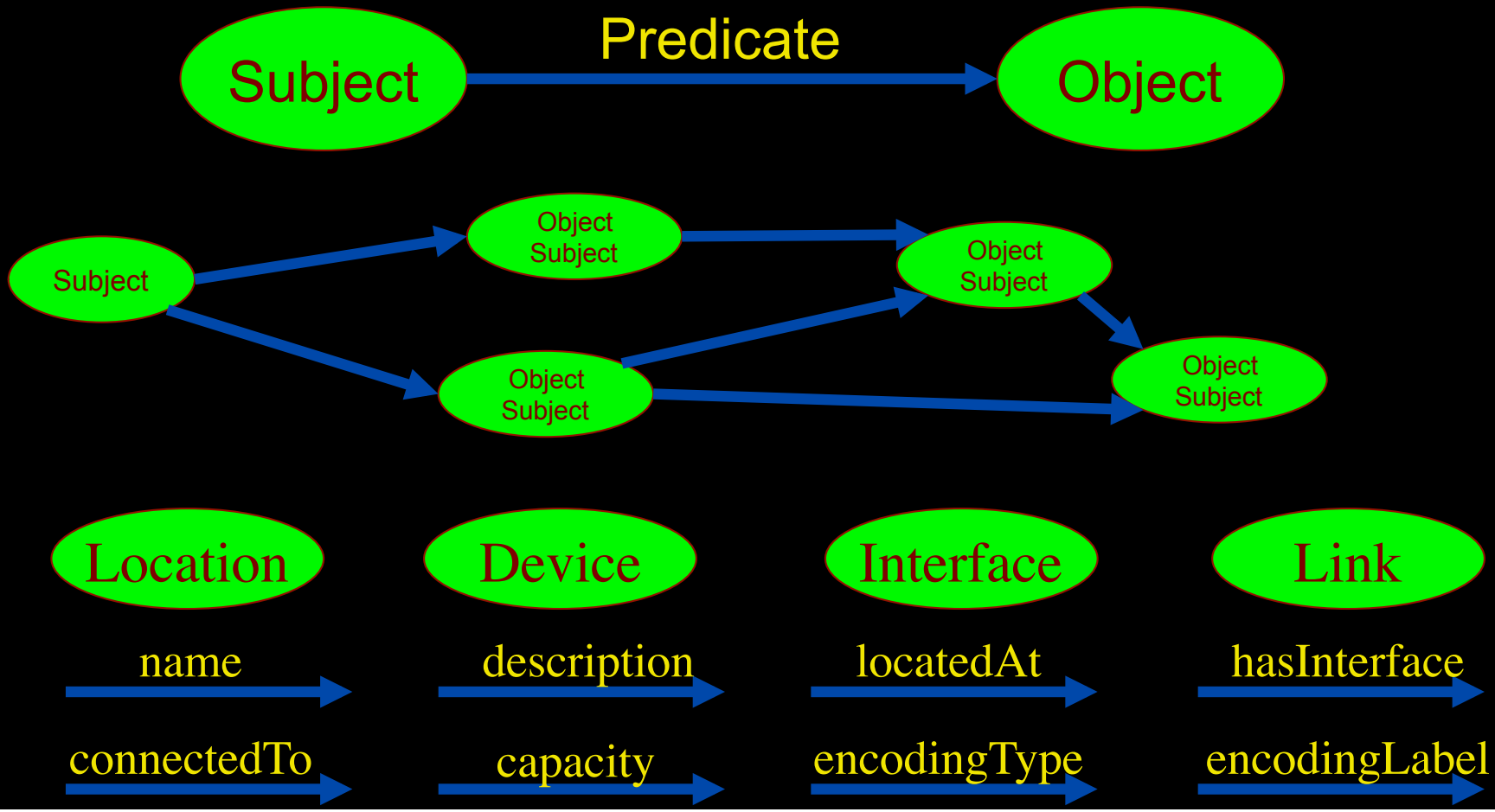


GLIF Q3 2005

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:

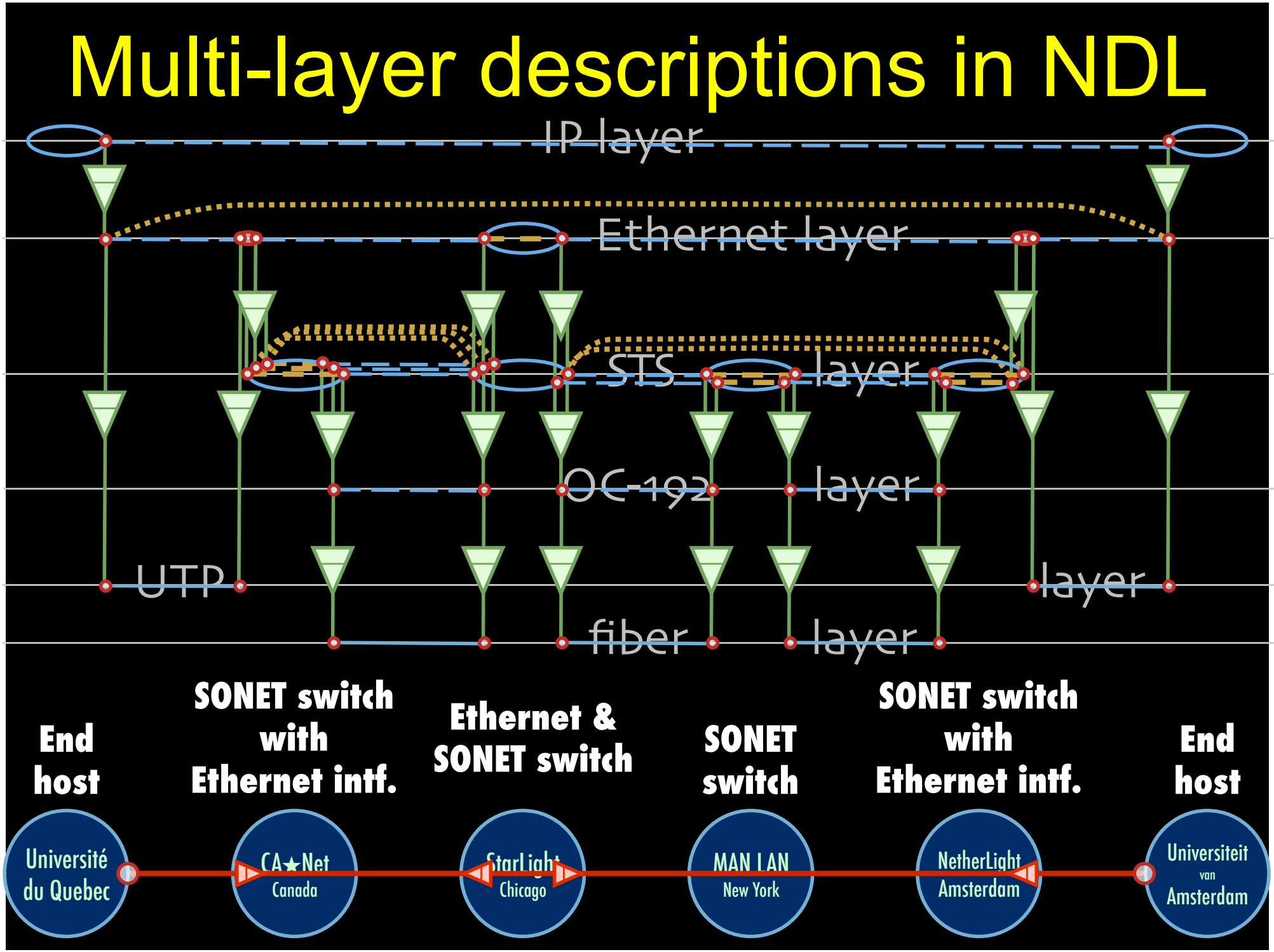


NetherLight in RDF

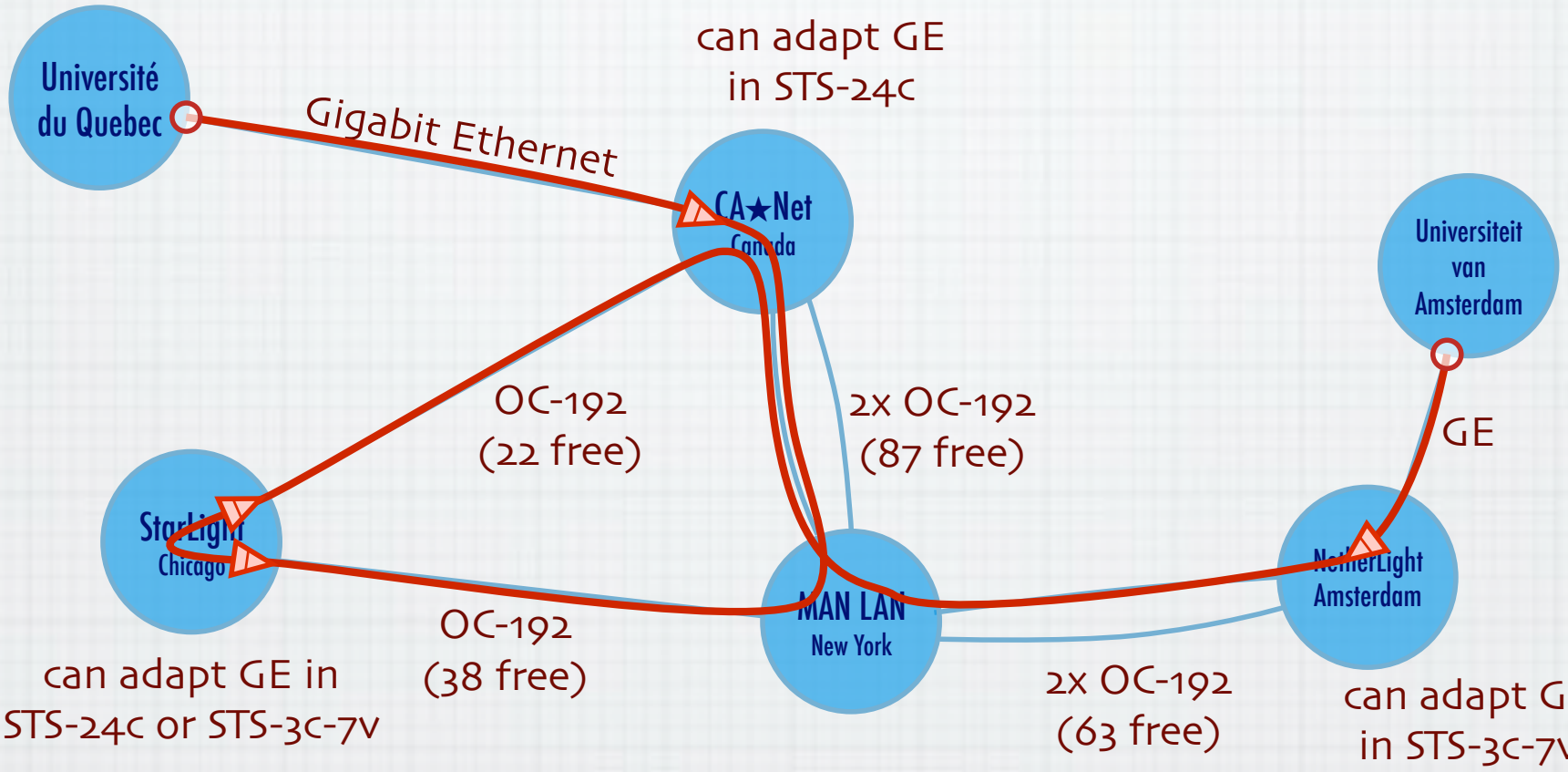
```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:ndl="http://www.science.uva.nl/research/air/ndl#">
<!-- Description of Netherlight -->
<ndl:Location rdf:about="#Netherlight">
      <ndl:name>Netherlight Optical Exchange</ndl:name>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



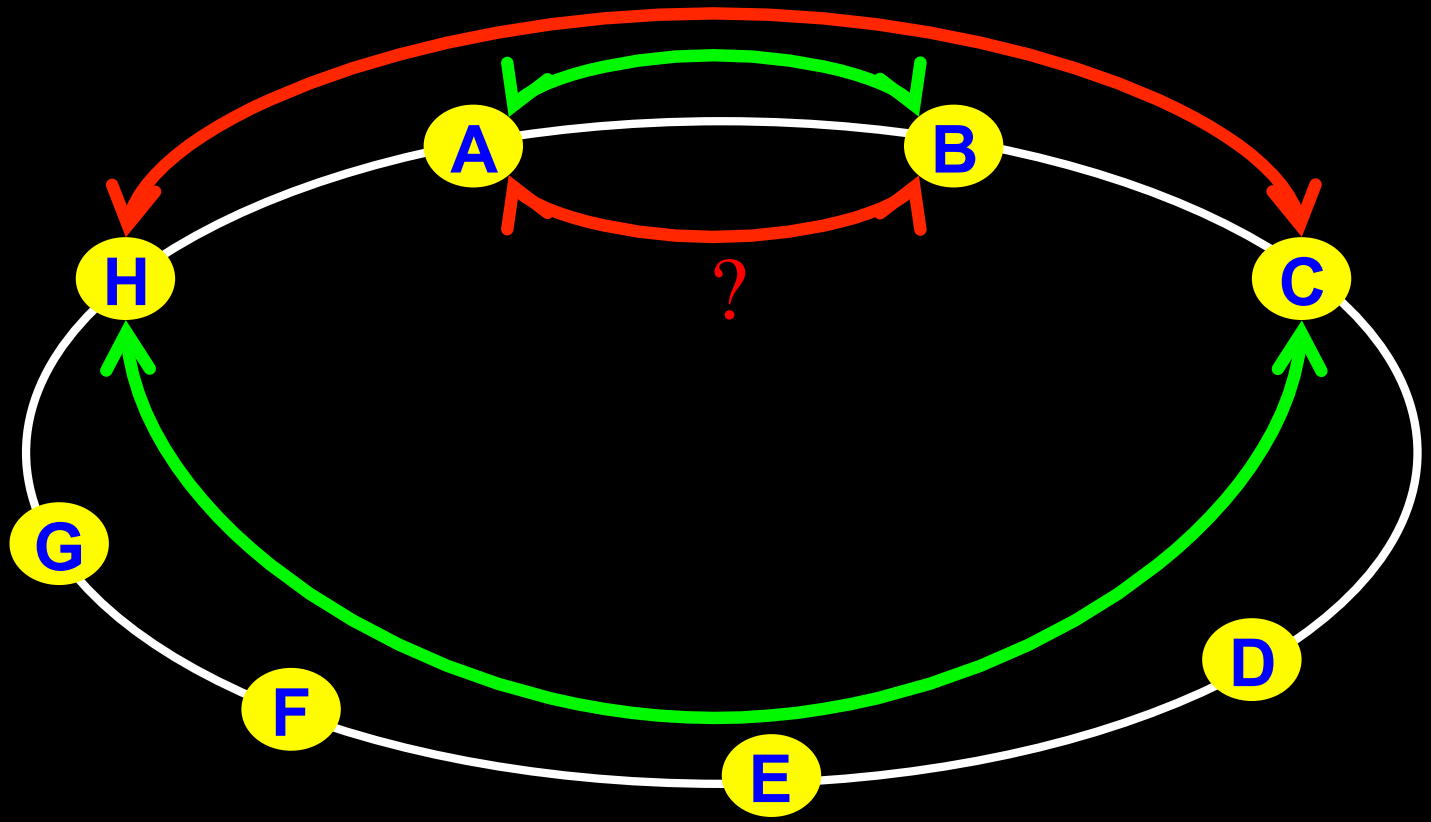
Thanks to Freek Dijkstra & team

The Problem

I want HC and AB

Success depends on the order

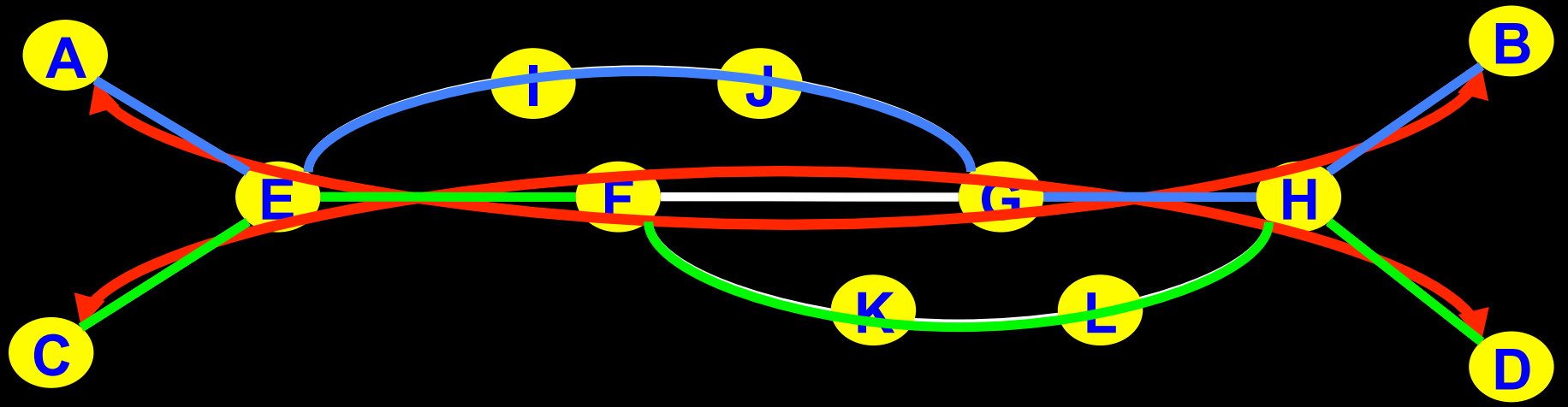
Wouldn't it be nice if I could request [HC, AB, ...]



Another one ☺

I want AB and CD

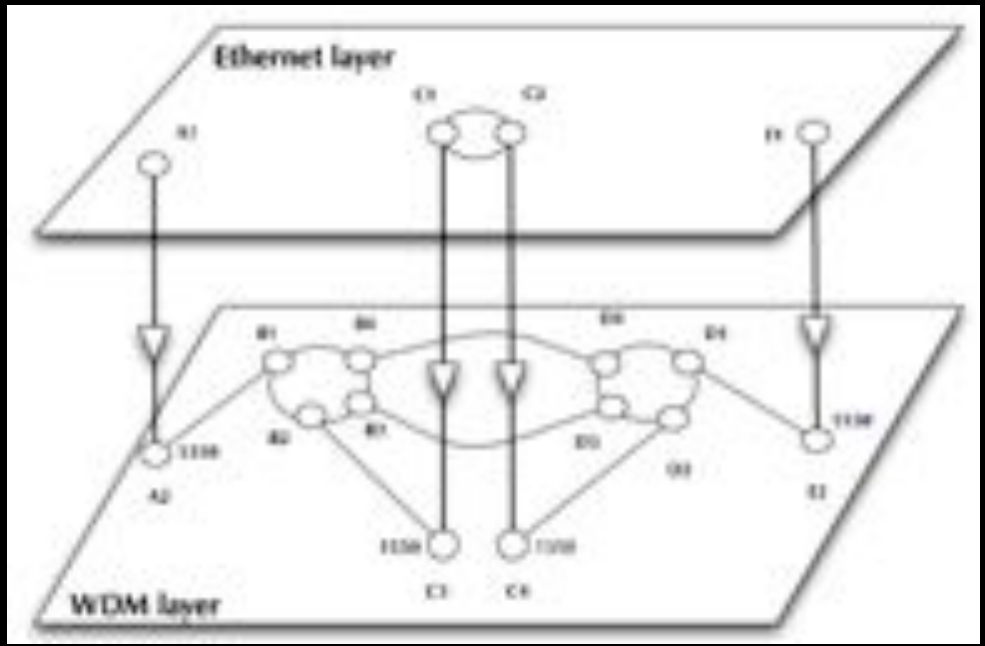
Success does not even depend on the order!!!

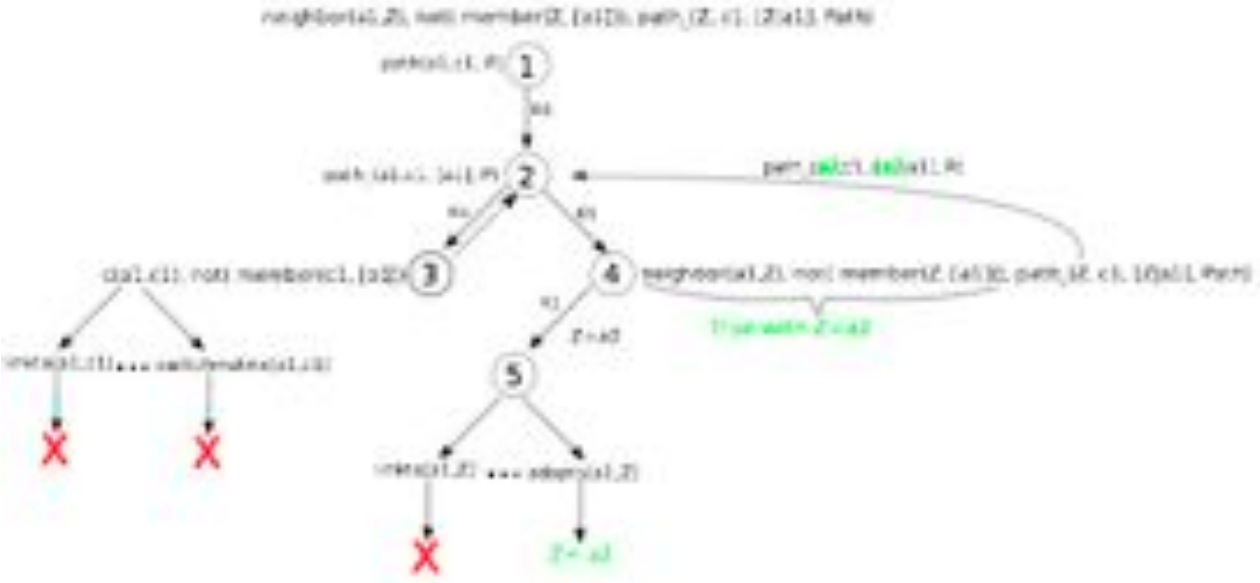


NDL + PROLOG

Research Questions:

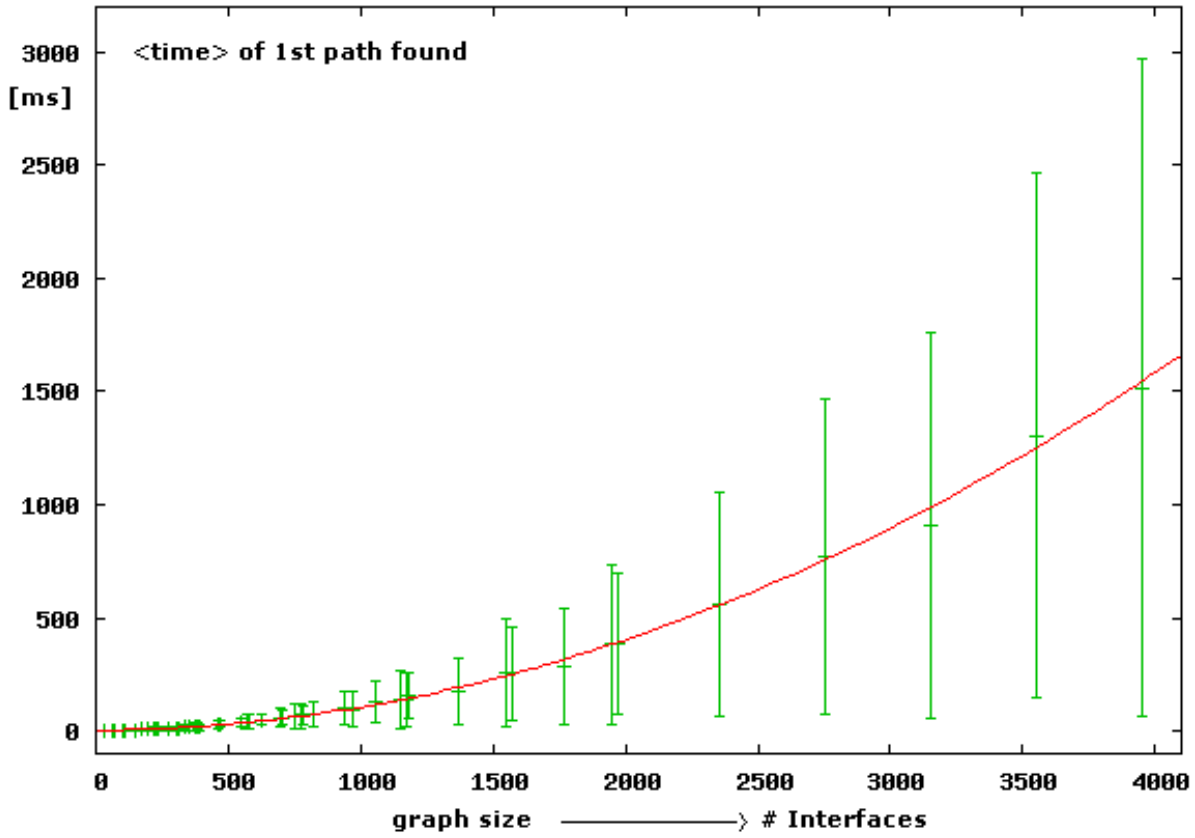
- order of requests
- complex requests
- usable leftovers





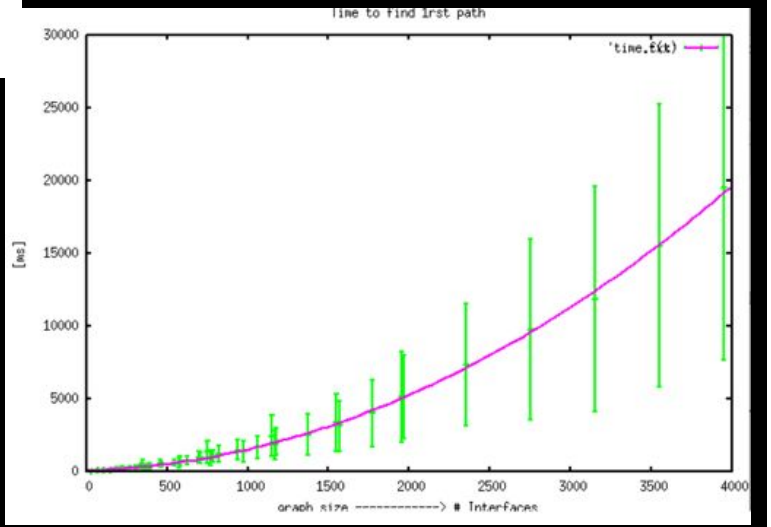
- Reason about graphs
- Find sub-graphs that comply with rules

Single layer networks: results



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

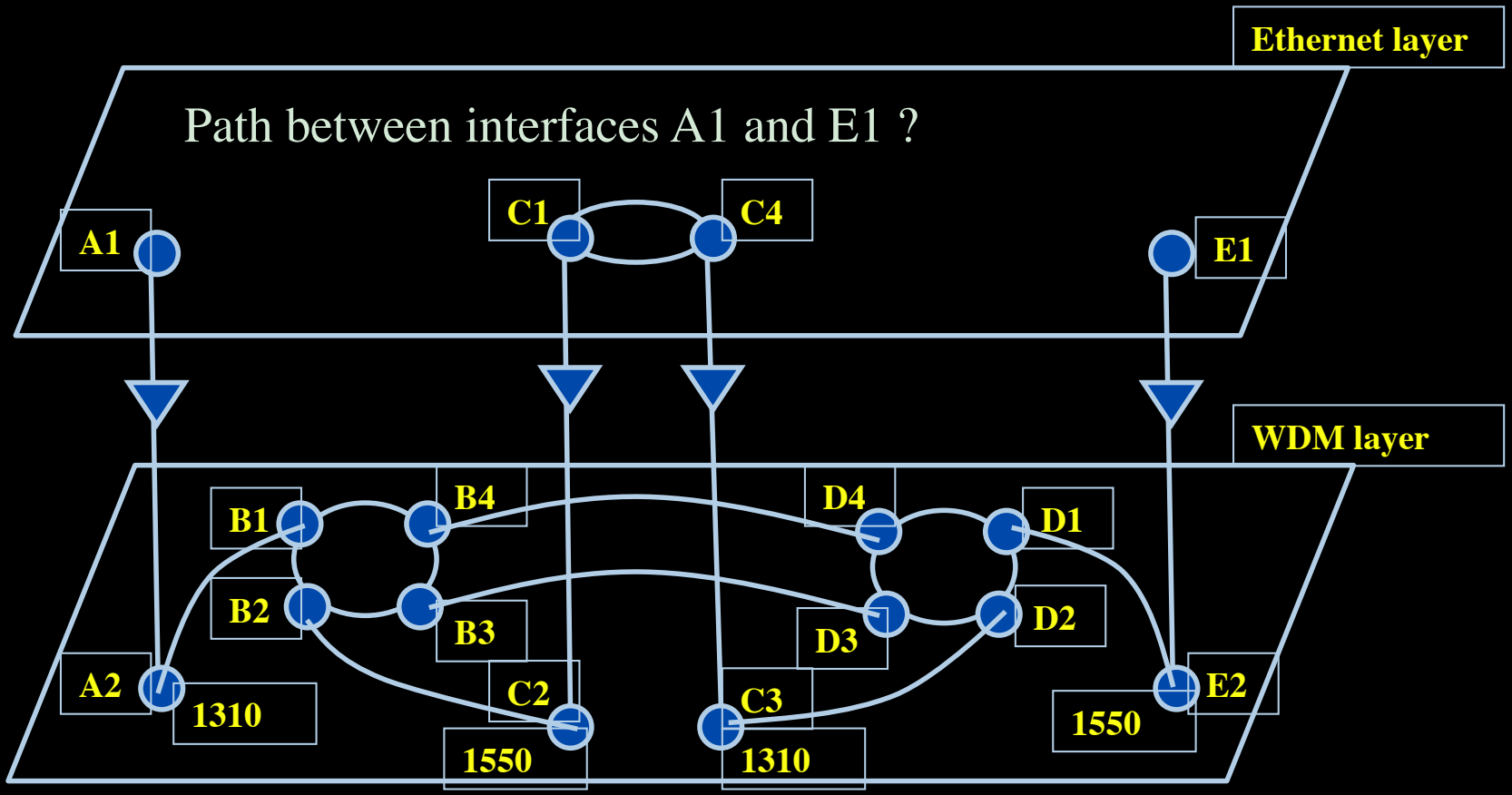
Pynt-based DFS



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



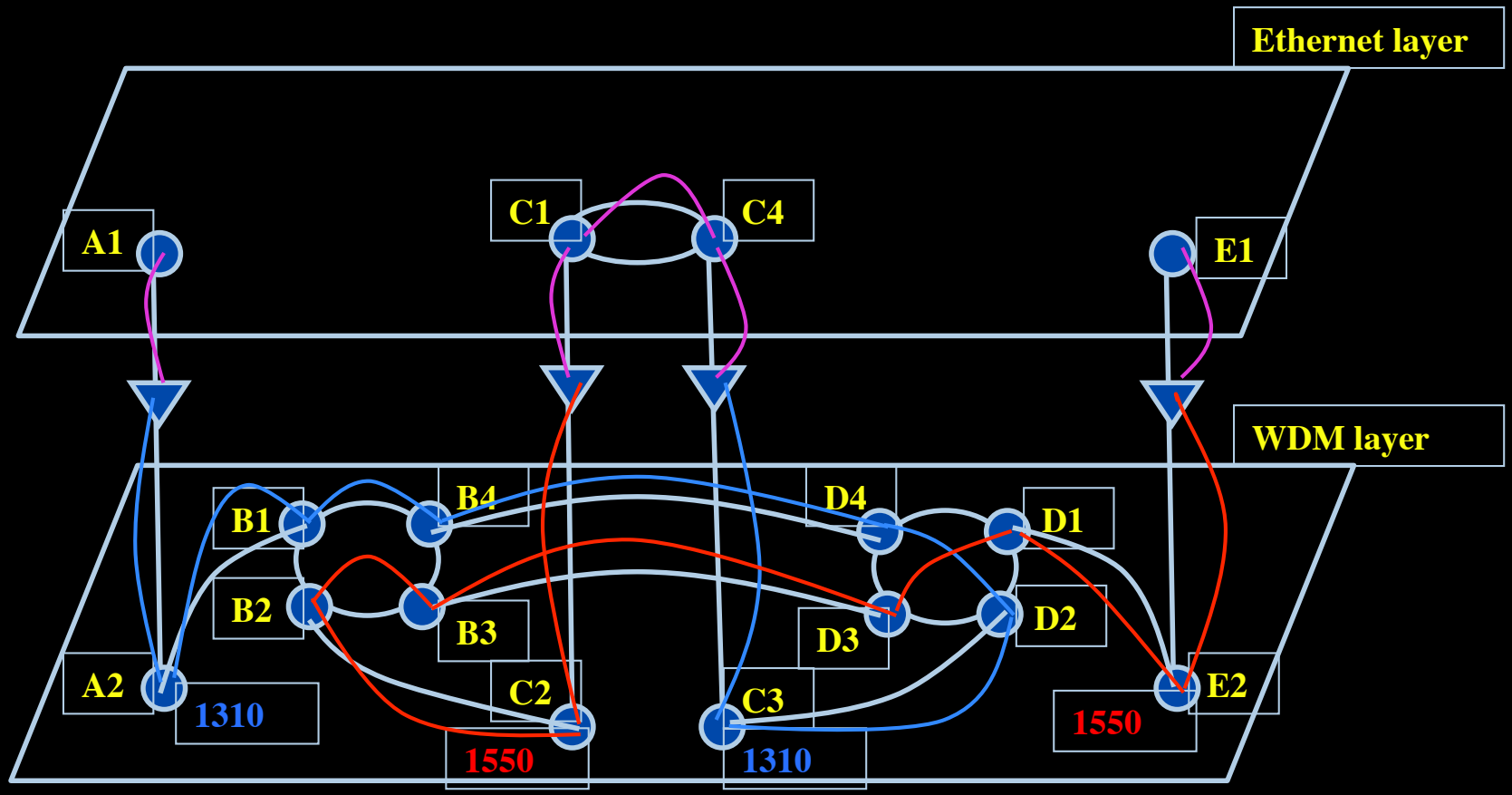
Prolog rule:

```
linkedto( Intf1, Intf2, CurrWav ):-                    %-- 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 Network PathFinding

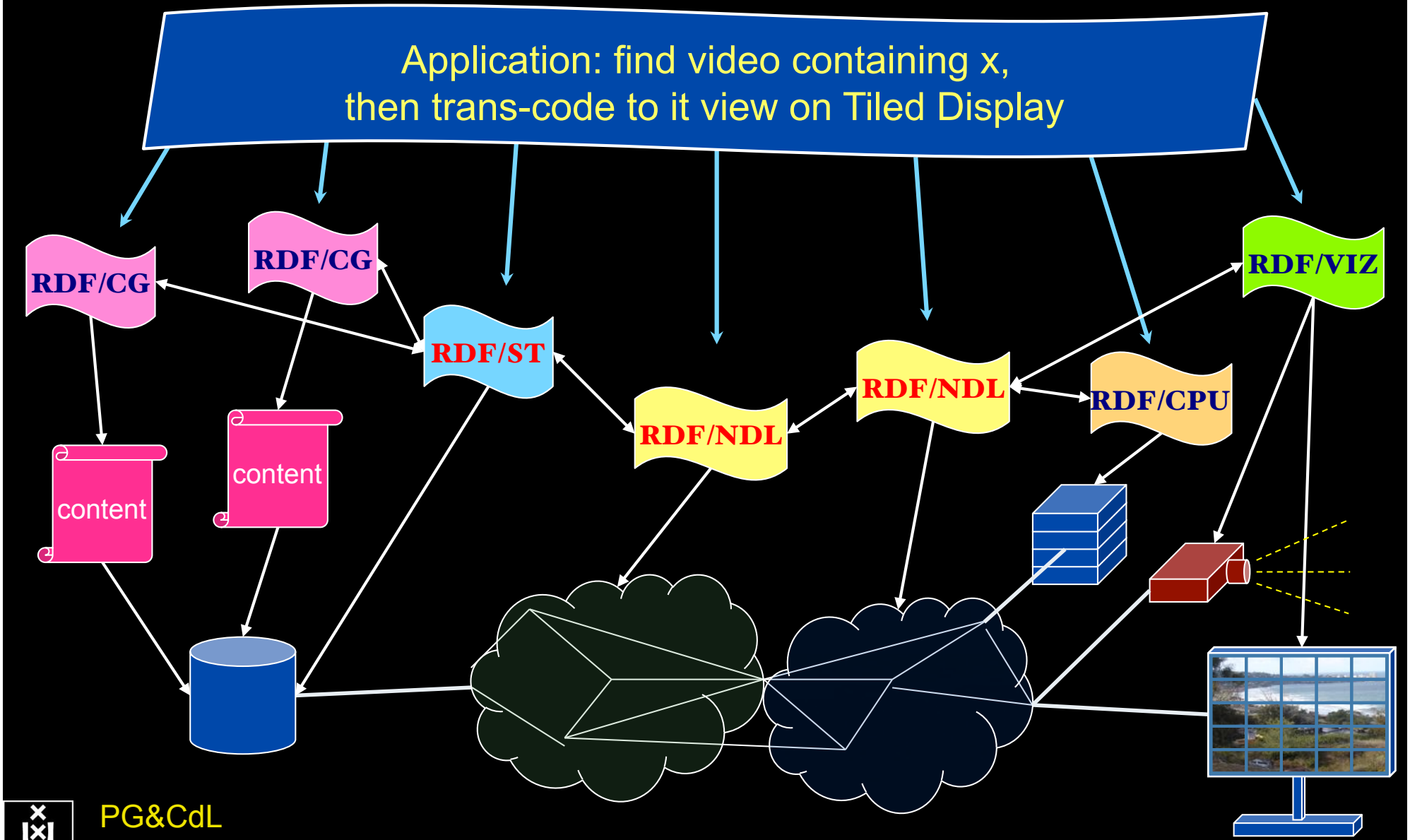


Path between interfaces A1 and E1:

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

Scaling: Combinatorial problem

RDF describing Infrastructure



PG&CdL

Applications and Networks become aware of each other!

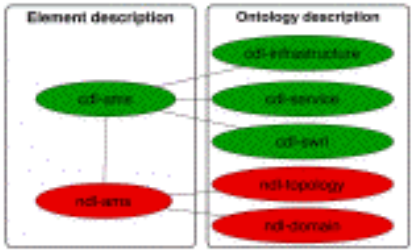
CineGrid Description Language

CineGrid is an initiative to facilitate the exchange, storage and display of high-quality digital media.

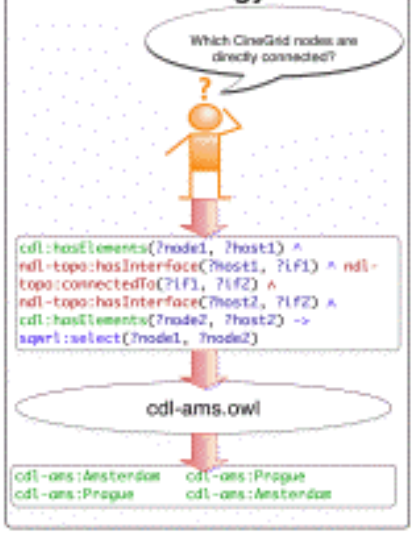
The CineGrid Description Language (CDL) describes CineGrid resources. Streaming, display and storage components are organized in a hierarchical way.

CDL has bindings to the NDL ontology that enables descriptions of network components and their interconnections.

With CDL we can reason on the CineGrid infrastructure and its services.



SQWRL is used to query the Ontology.

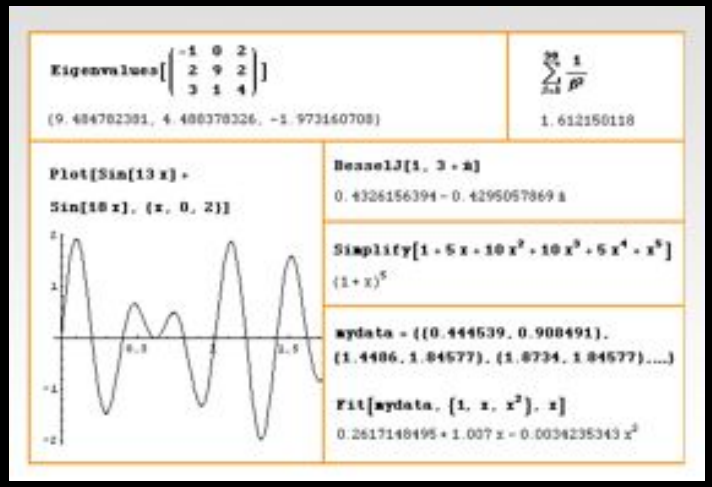


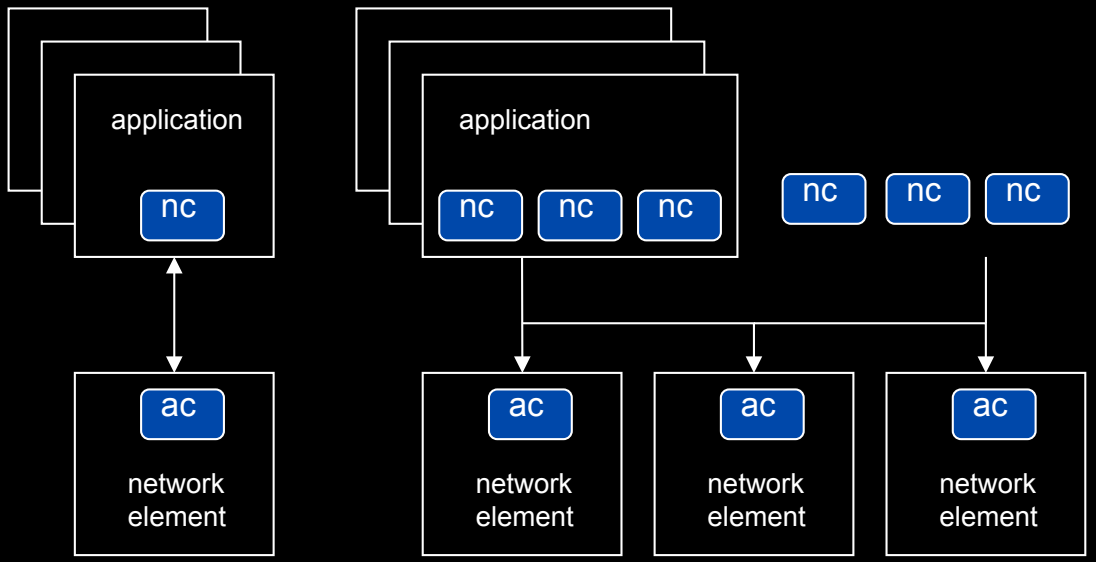
UML representation of CDL

CDL links to ***NDL*** using the ***owl:SameAs*** property. ***CDL*** defines the services, ***NDL*** the network interfaces and links. The combination of the two ontologies identifies the host pairs that support matching services via existing network connections.

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





StarPlane

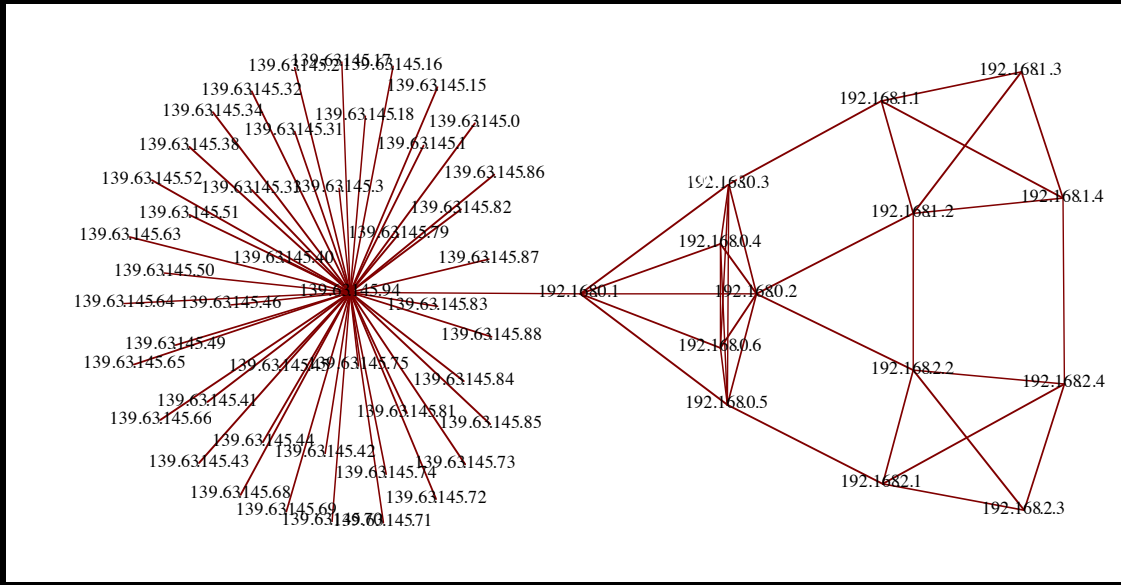
Mathematica enables advanced graph queries, visualizations and real-time network manipulations on UPVNs

Topology matters can be dealt with algorithmically

Results can be persisted using a transaction service built in UPVN

Initialization and BFS discovery of NEs

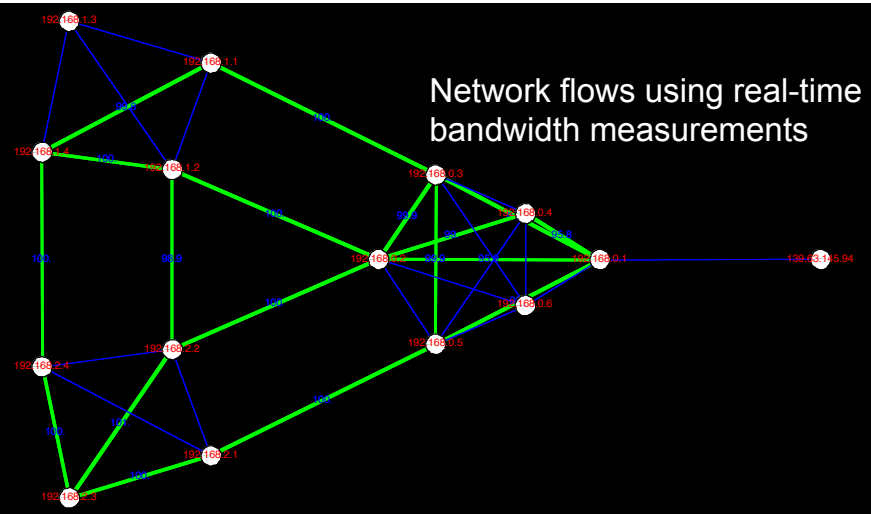
```
Needs["WebServices`"]
<<DiscreteMath`Combinatorica`
<<DiscreteMath`GraphPlot`
InitNetworkTopologyService["edge.ict.tno.nl"]

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
```



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
```



Network flows using real-time bandwidth measurements

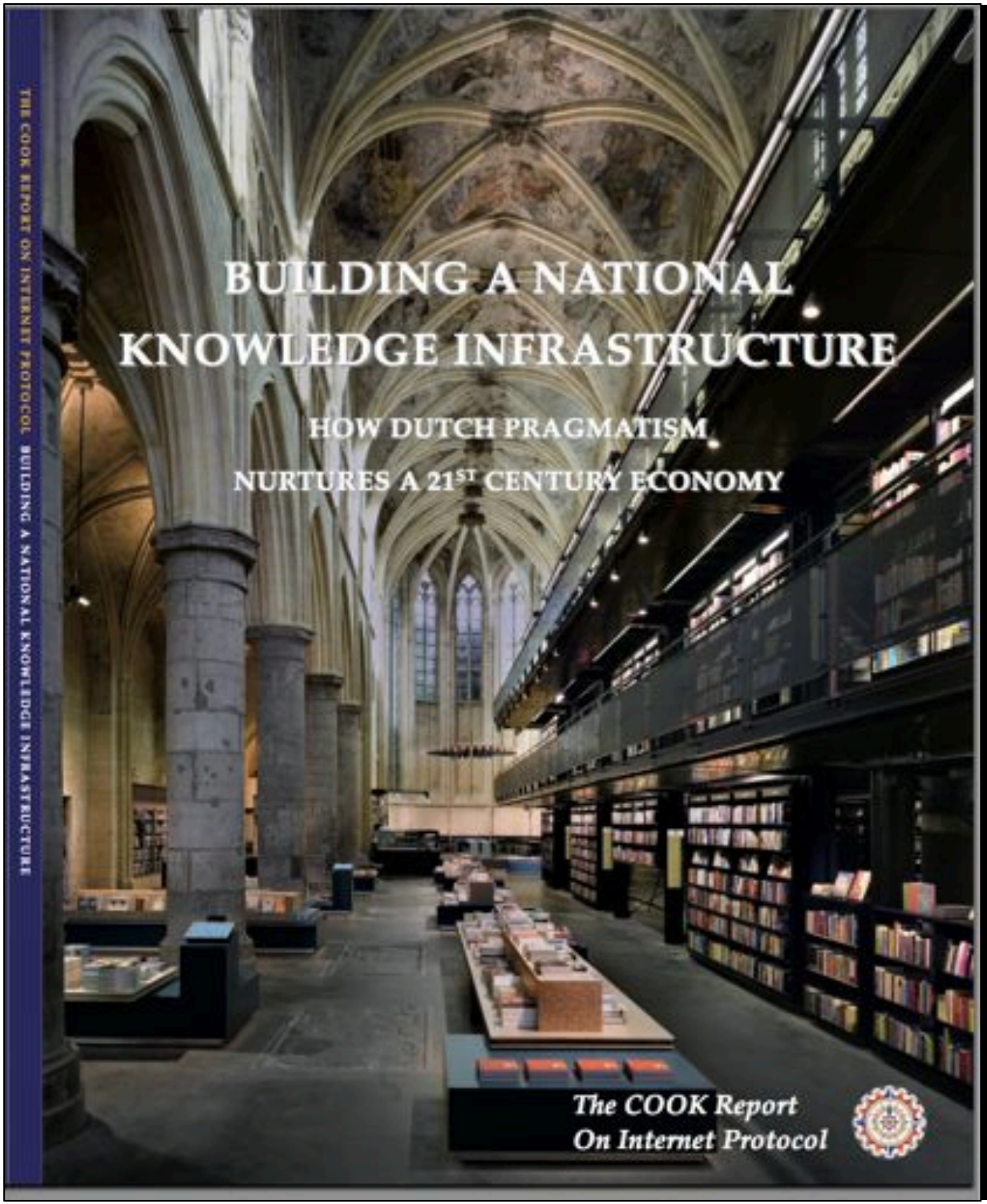
ref: Robert J. Meijer, Rudolf J. Strijkers, Leon Gommans, Cees de Laat, User Programmable Virtualiized Networks, accepted for publication to the IEEE e-Science 2006 conference Amsterdam.

StarPlane

TouchTable Demonstration @ SC08

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



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 ?

SURF NET