e-Infrastructure aware Topology handling in the Global Lambda Integrated Facility

Cees de Laat

EU

SURFnet

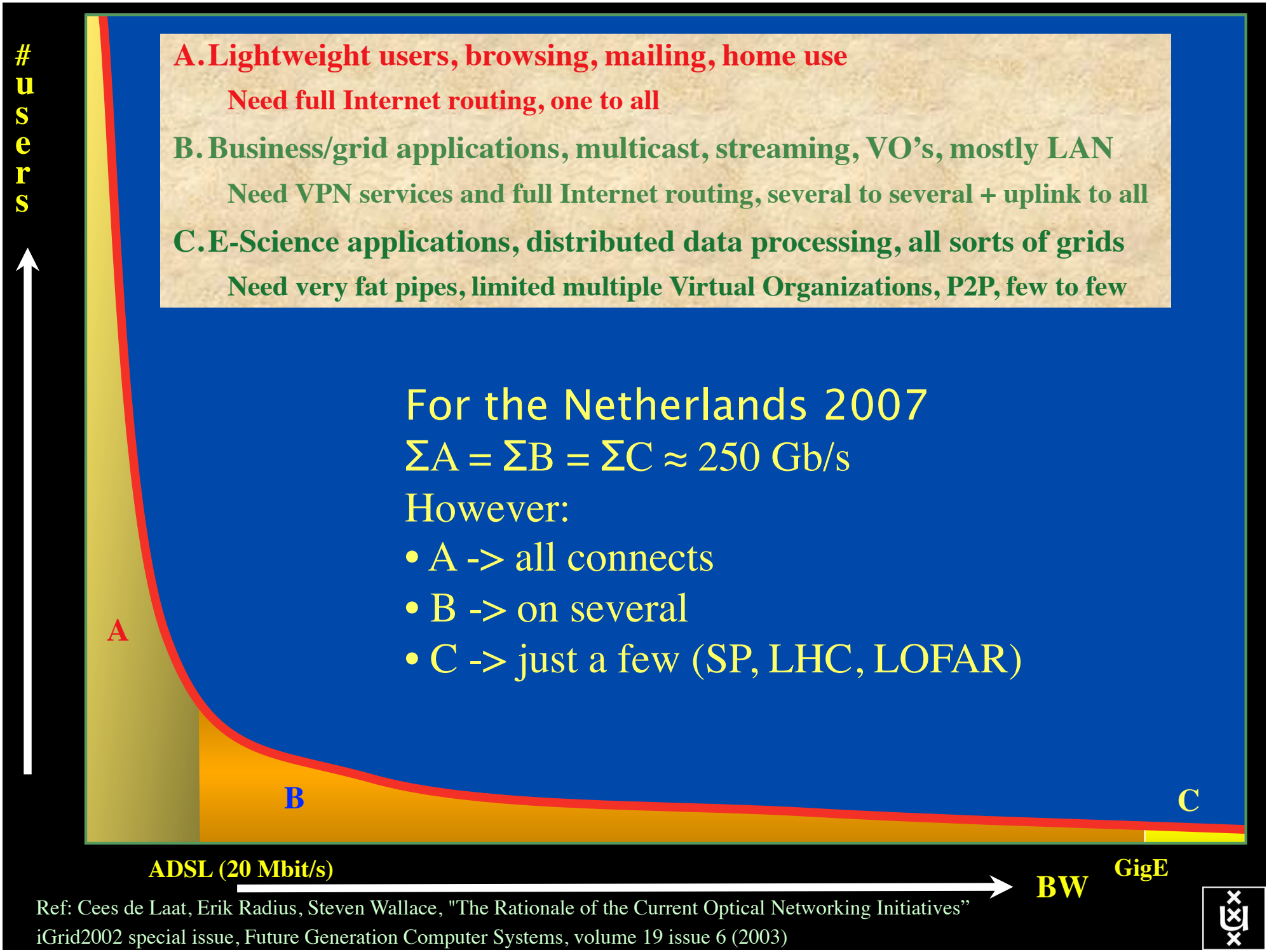
SURF-eScience

NWO

University of Amsterdam

TNO

NCF



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)

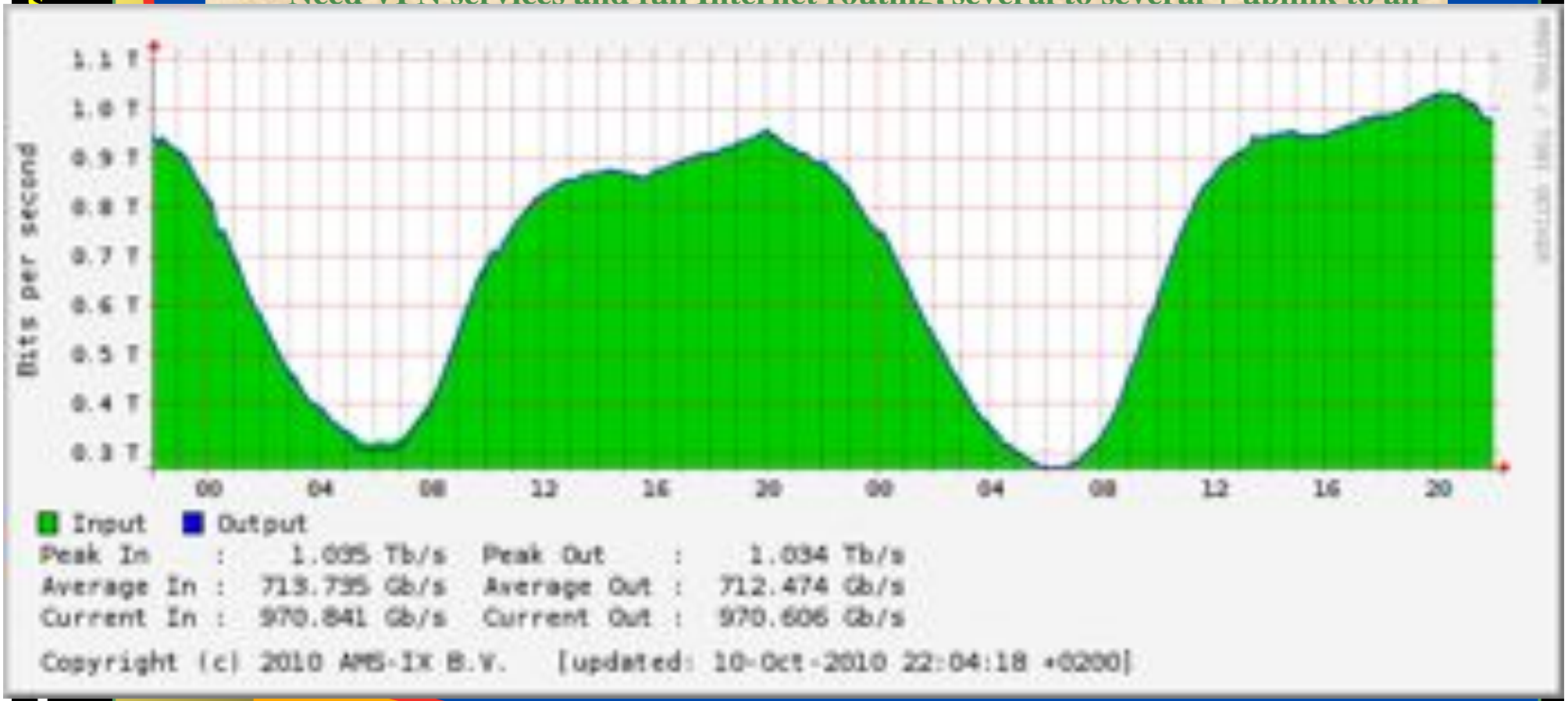
users

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

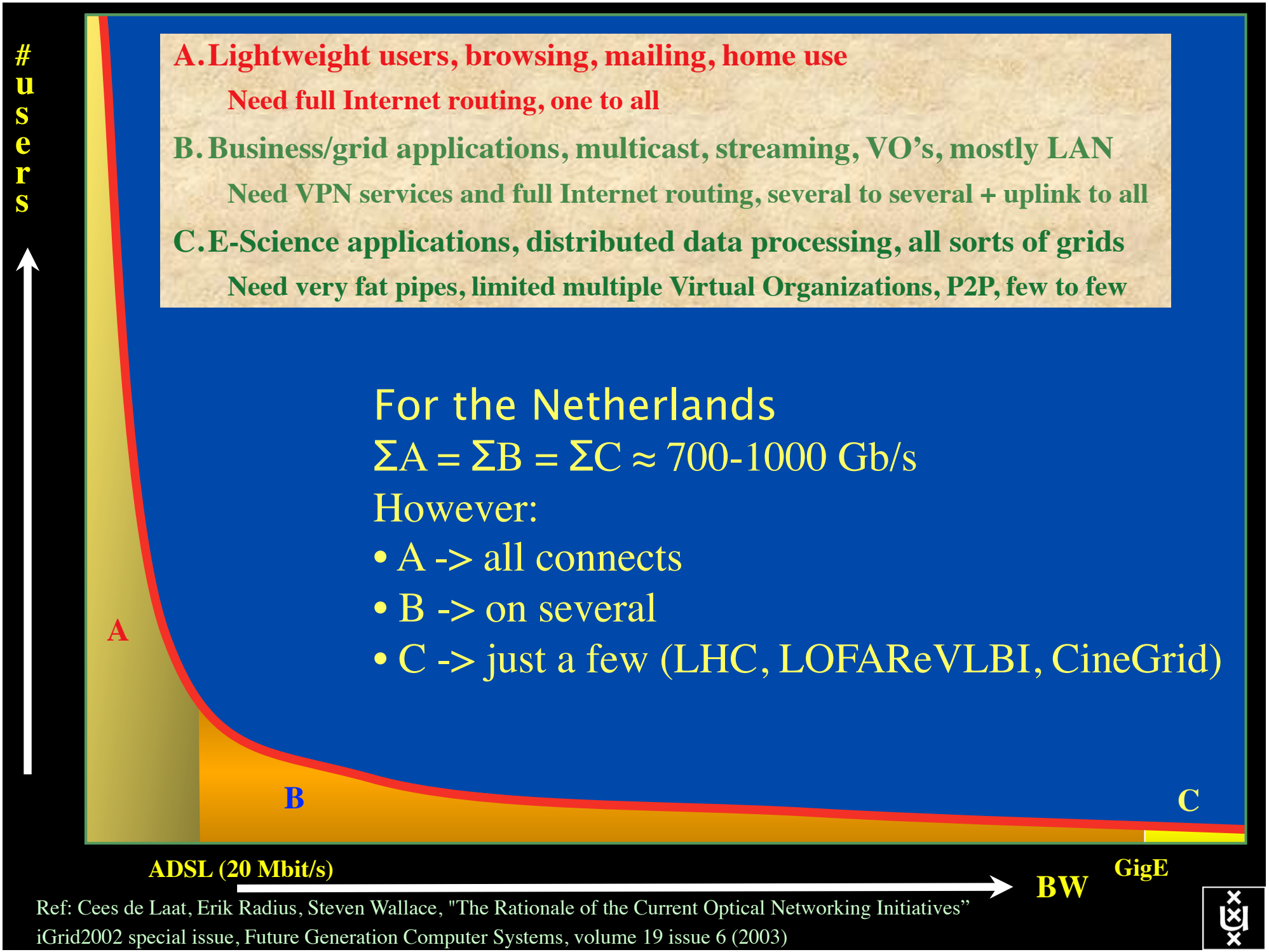


B

C

ADSL (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)



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!

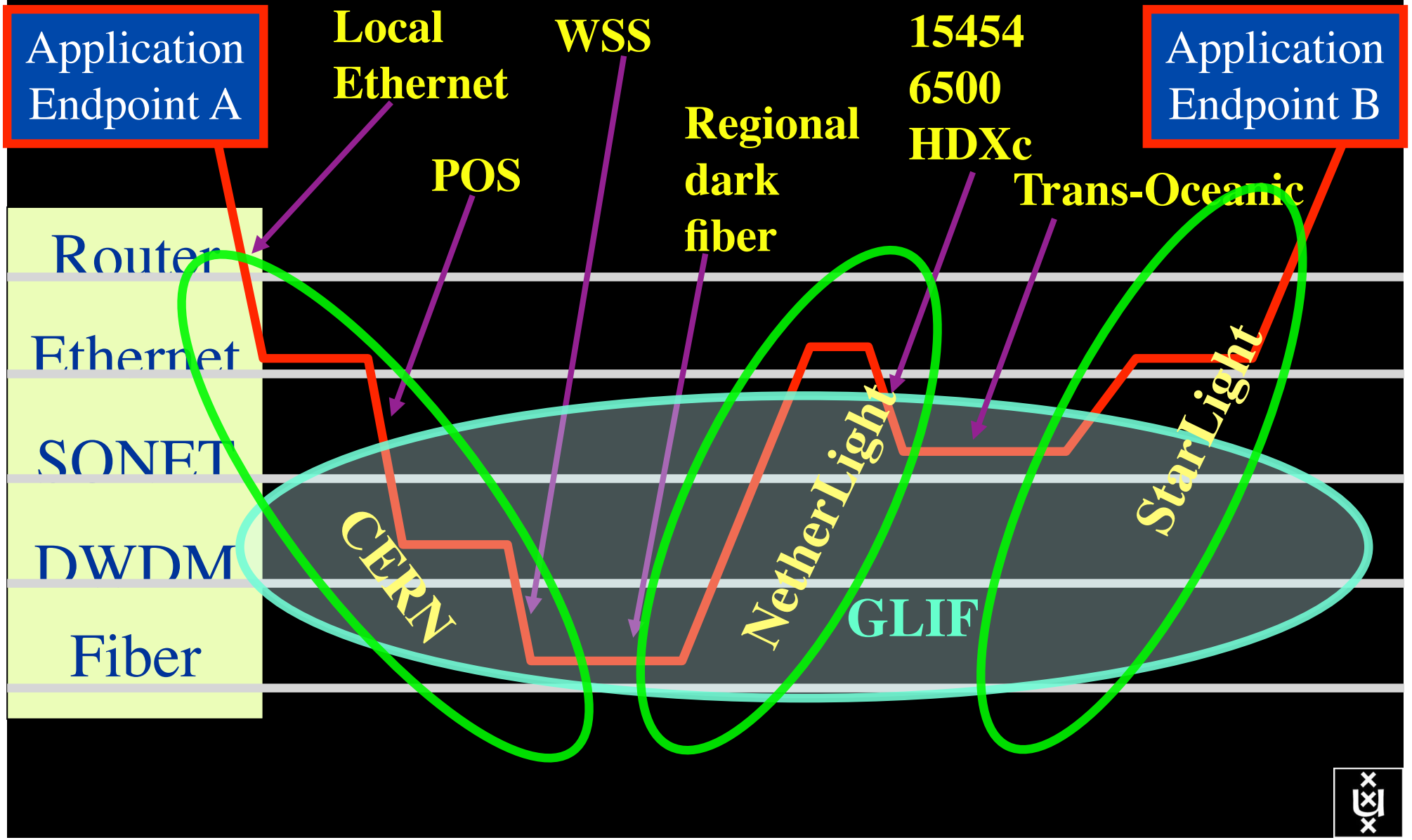
- Costs of photonic equipment 10% of switching 10 % of full routing
  - for same throughput!
  - Photonic vs Optical (optical used for SONET, etc, 10-50 k$/port)
  - DWDM lasers for long reach expensive, 10-50 k$
- Bottom line: look for a hybrid architecture which serves all classes in a cost effective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

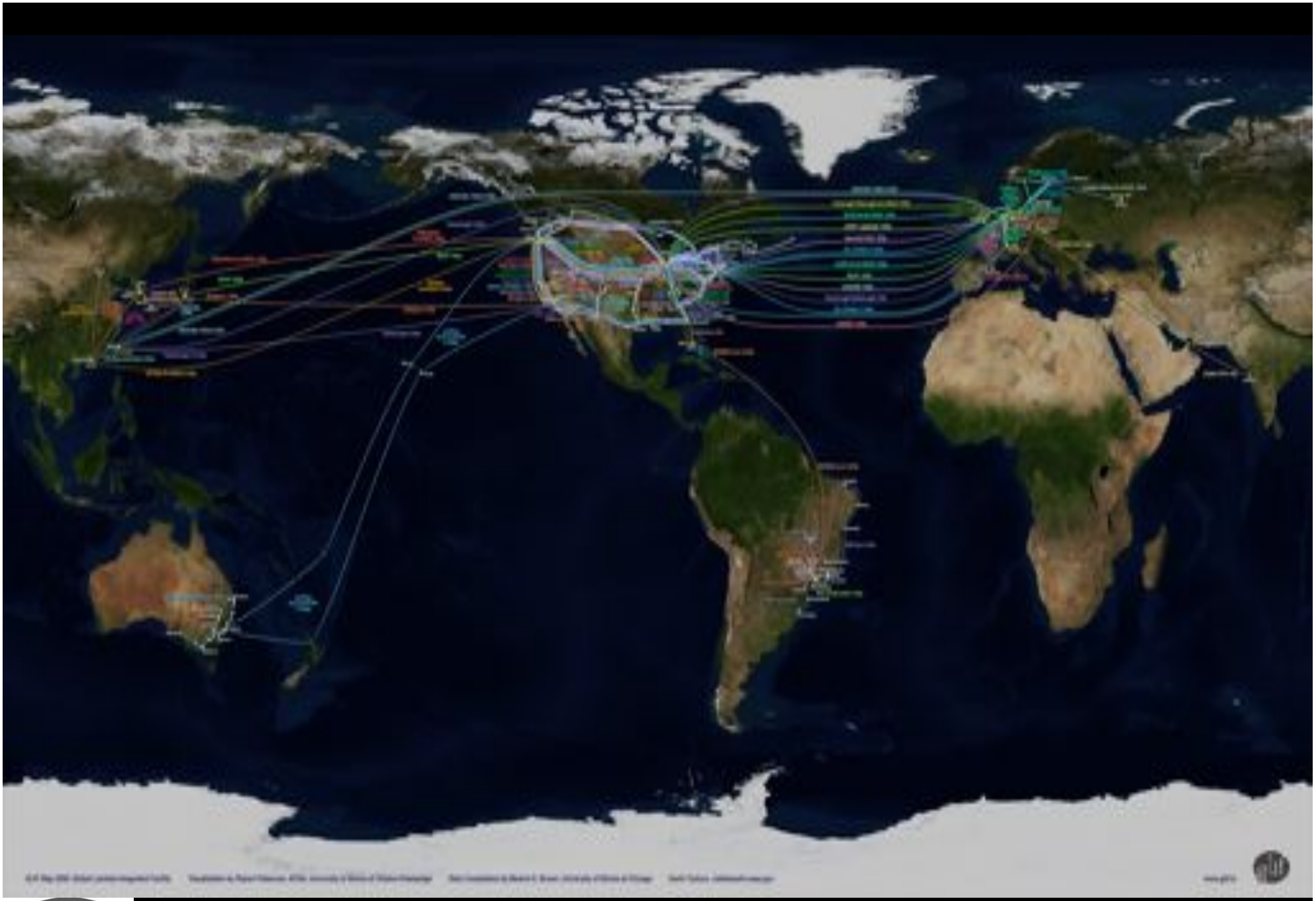
L1 ≈ 2-3 k$/port

L2 ≈ 2-5 k$/port

L3 ≈ 50+ k$/port

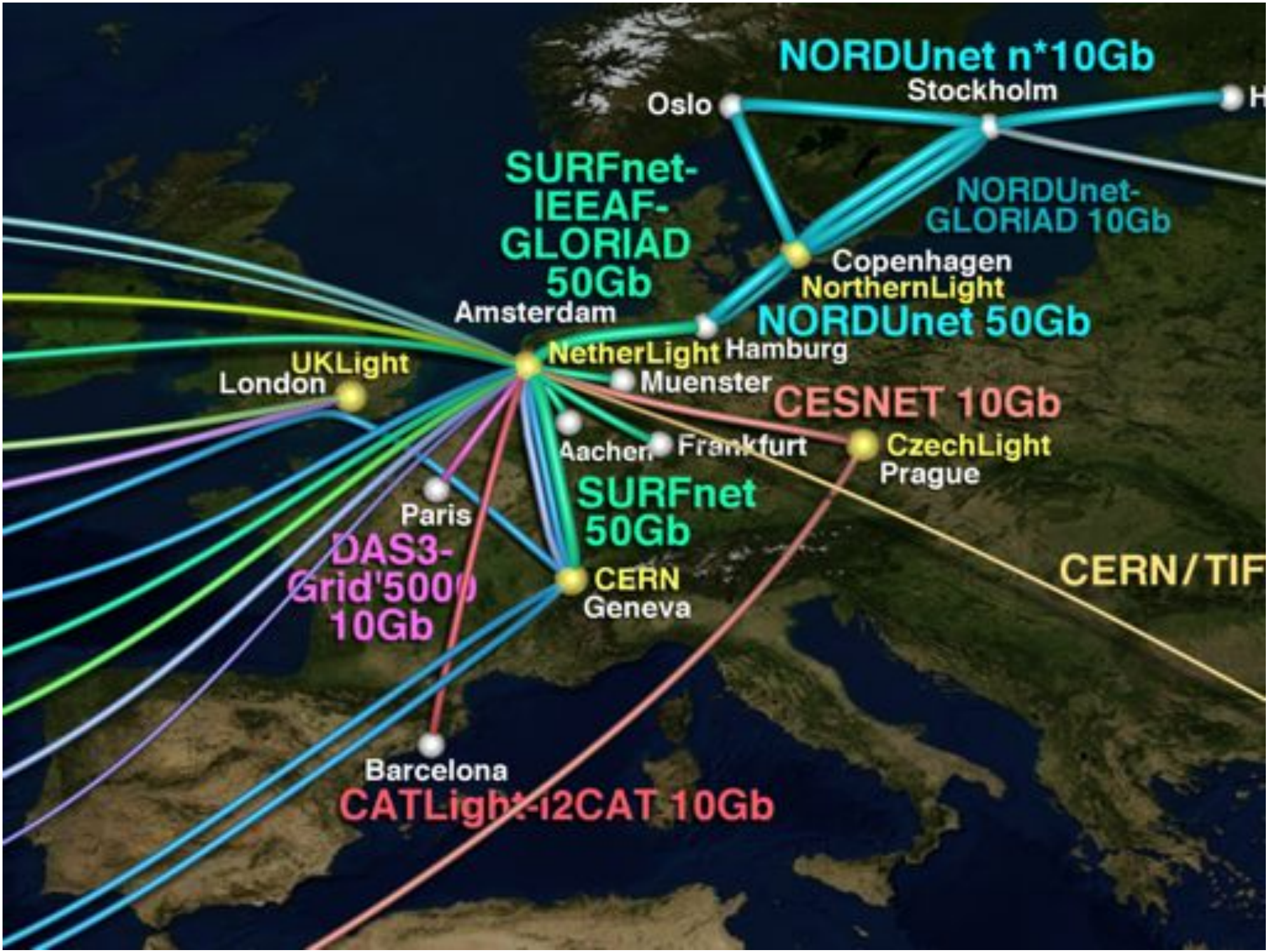
How low can you go?





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

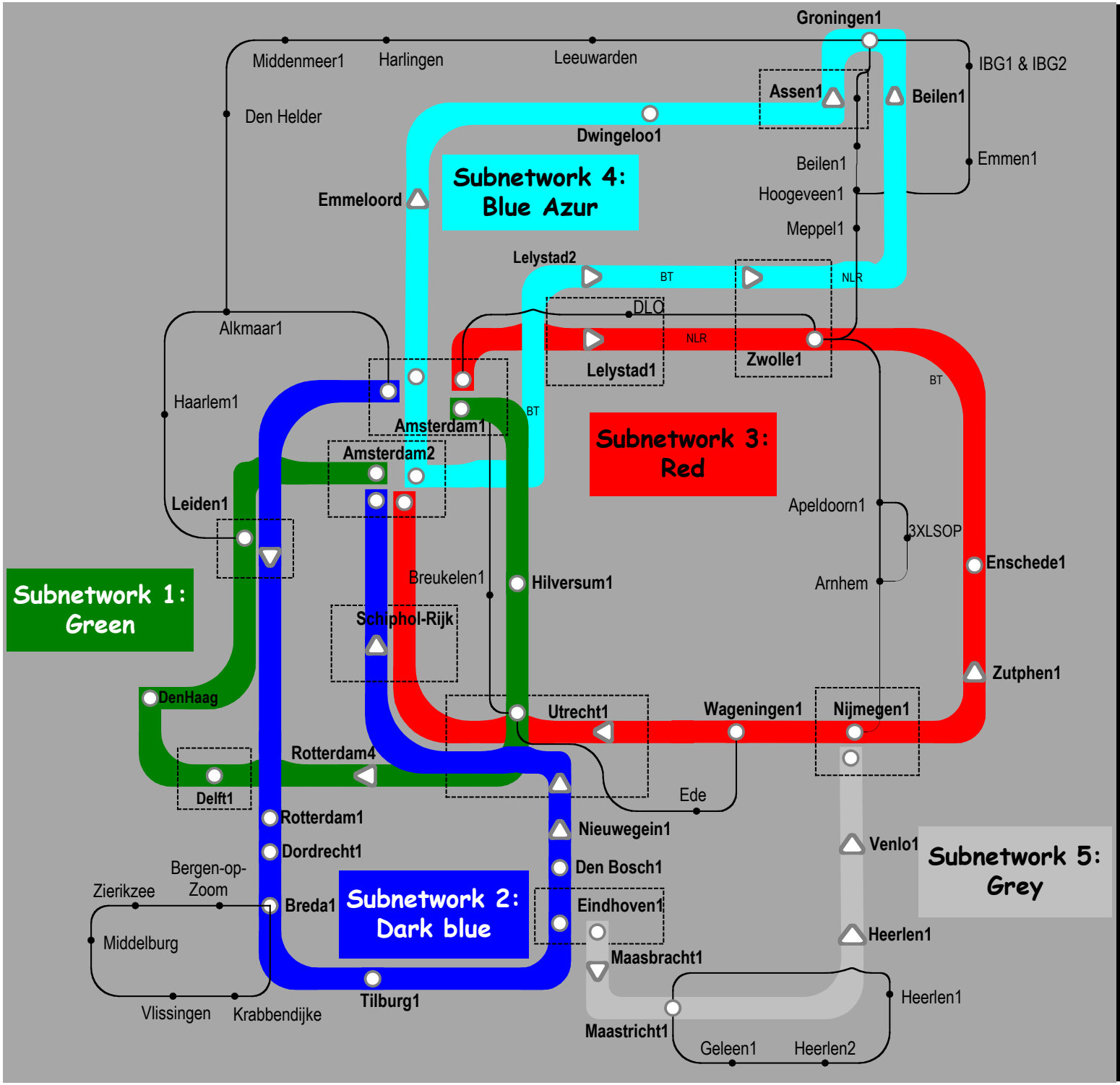
In The Netherlands SURFnet connects between 180:

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

with an indirect ~750K user base

~ 8860 km

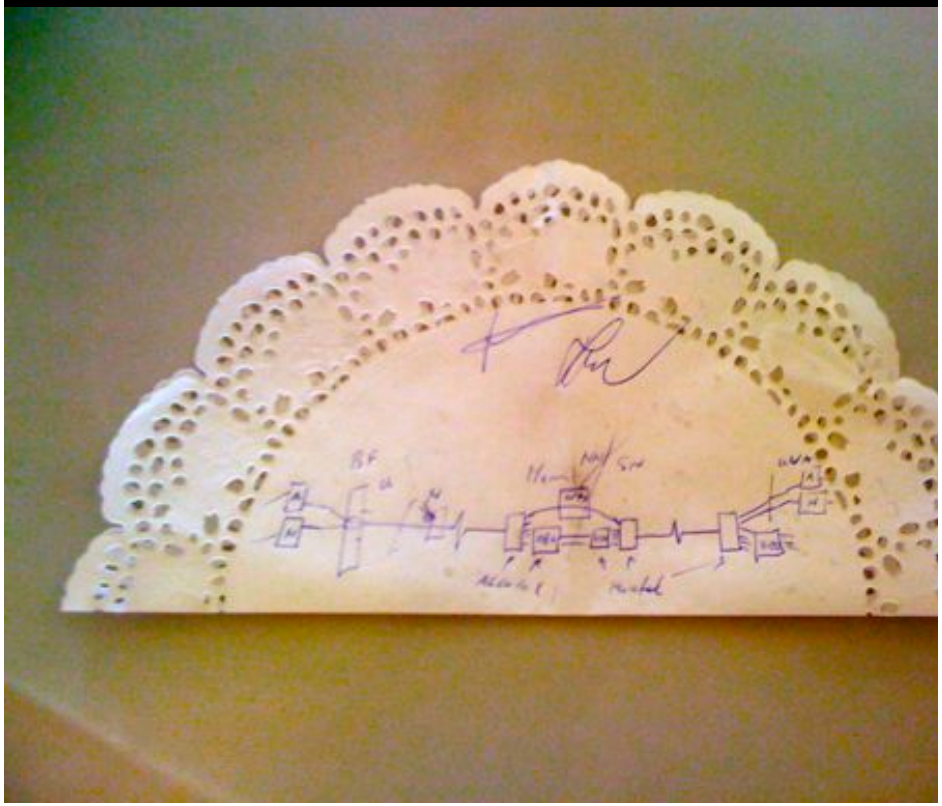
scale comparable to railway system



Common Photonic Layer (CPL) in SURFnet6

supports up to 72 Lambda's of 10 / 40 / 100 G

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

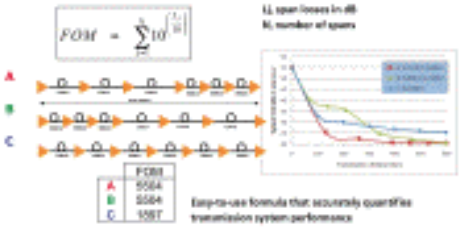
Alien wavelength challenges

- 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.

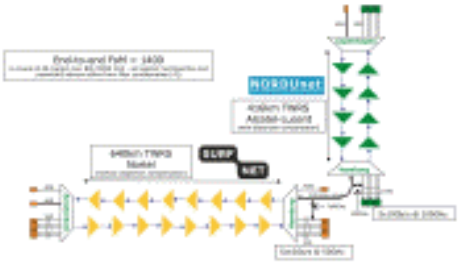
New method to present fiber link quality, FoM (Figure of Merit)

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

GLIF 2010

40 Gbps Lambda

Based on Ethernet

From UVA cluster

To CERN cluster

ClearStream

End-to-End Ultra Fast Transmission Over a Wide Area 40 Gbit/s Lambda

University of Amsterdam
Cosmin Dumitru
Cees de Laat
Ralph Koning
SURFnet
Erik-Jan Bos
Gerben van Malenstein
Ciena
David Yeung
Jan-Willem Ellion
Harry Peng
Kevin McKernan
Martin Bluethner
VU University Amsterdam
Kees Verstoep
Henri Bal
Mellanox
Erez Cohen
Bill Lee

Utilizing shared expertise in advanced photonic, leading edge hardware and high-performance computing, the team created a network application testbed using the 1650 km Cross Border Fiber between NetherLight and CERNLight, lit by SURFnet, connecting servers equipped with 40 Gigabit Ethernet network interface at the University of Amsterdam to remote servers with corresponding interfaces at GLIF 2010 in Geneva.

Network Setup

The Mellanox ConnectX-2 EN 40GbE is the first network interface that allows single stream ethernet transport far exceeding the common 10Gbps boundary limit.

The network infrastructure is based on Ciena's Optical Multiservice Edge (OME) 6500 equipped with 40 GbE interfaces, which enables data speeds to be seamlessly upgraded from 10 Gbps to 40 Gbps.

Application Setup

The DiVinE application is MPI based and in this setup uses TCP/IP as its network backend. DiVinE's runtime system is optimized to achieve good performance despite the very intensive traffic rate and high WAN latency over long distance.

We also use a server with basic UDP and TCP test tools to tune and measure capacities. Going beyond 10 Gbps leads to new challenges in applications, operating system tuning and system architecture design as new bottlenecks appear.

Special attention needs to be given to the setup of multi-core machines in order to have the best I/O performance and maximize the network throughput. During the demo the PCI-E x8 2.0 interface of the network card is saturated when using UDP traffic.

DiVinE

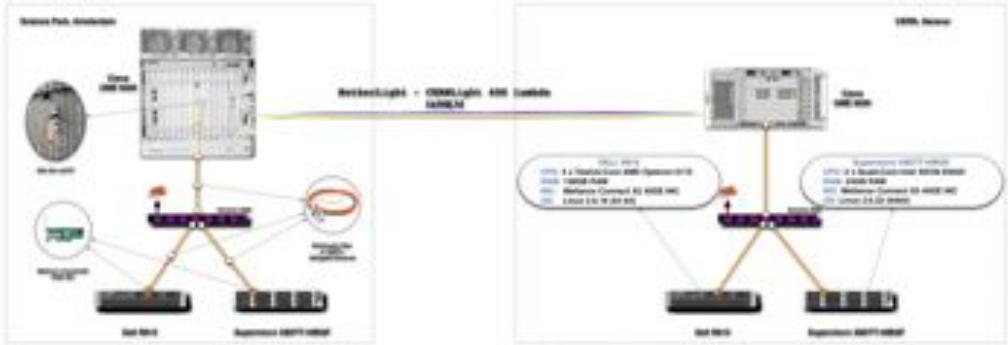
DiVinE is a tool for LTL model checking and reachability analysis of discrete distributed systems. The tool is able to efficiently exploit the aggregate computing power of multiple network-interconnected multi-cored workstations in order to deal with extremely large verification tasks.

Cluster-in-a-box

The Dell R815 is a 2U server powered by 48 AMD Opteron 6100 cores which make it as one of the densest x64 servers available on the market and is used to run the DiVinE application.

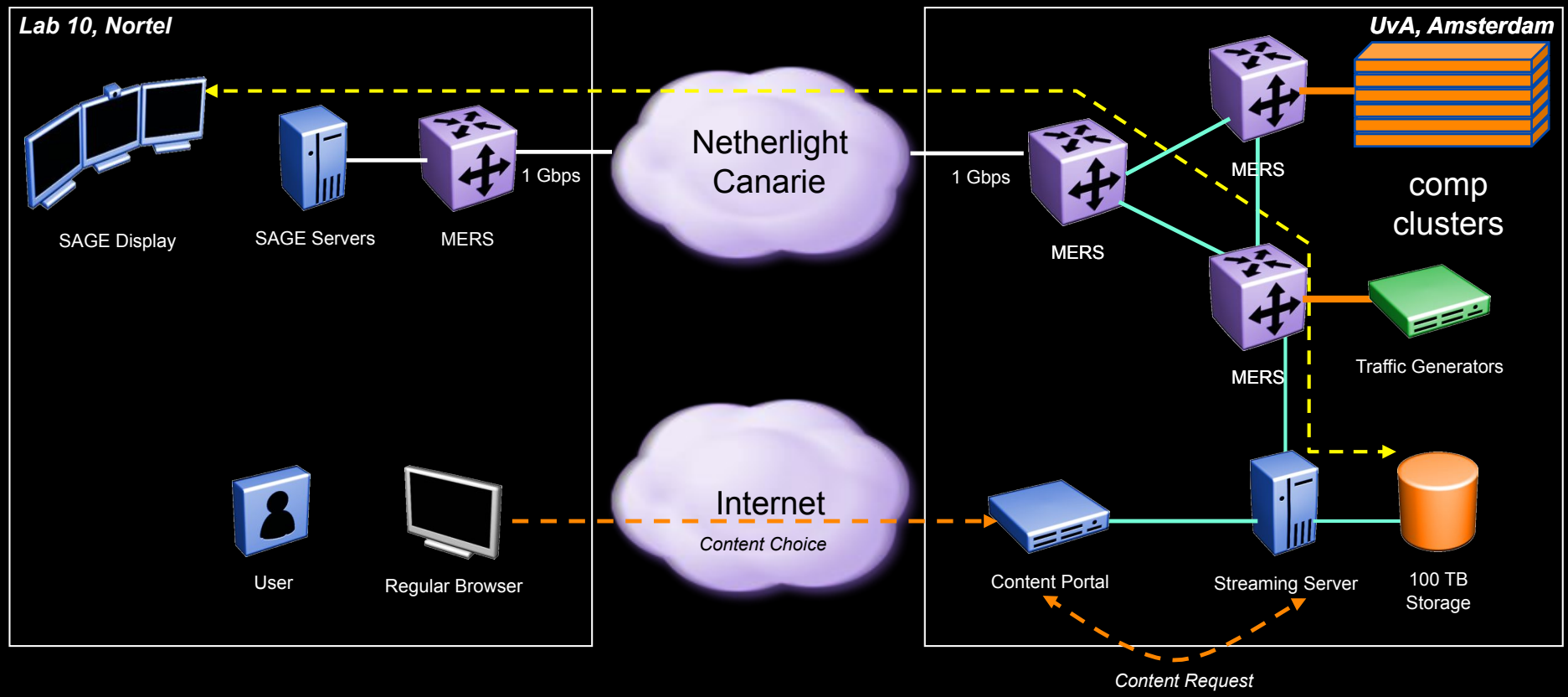
High Performance Node

Using a flexible I/O architecture, the Supermicro X8DTT with two quad-core Intel E5620 CPUs, allows extreme speeds of over 25 Gbps to be reached.

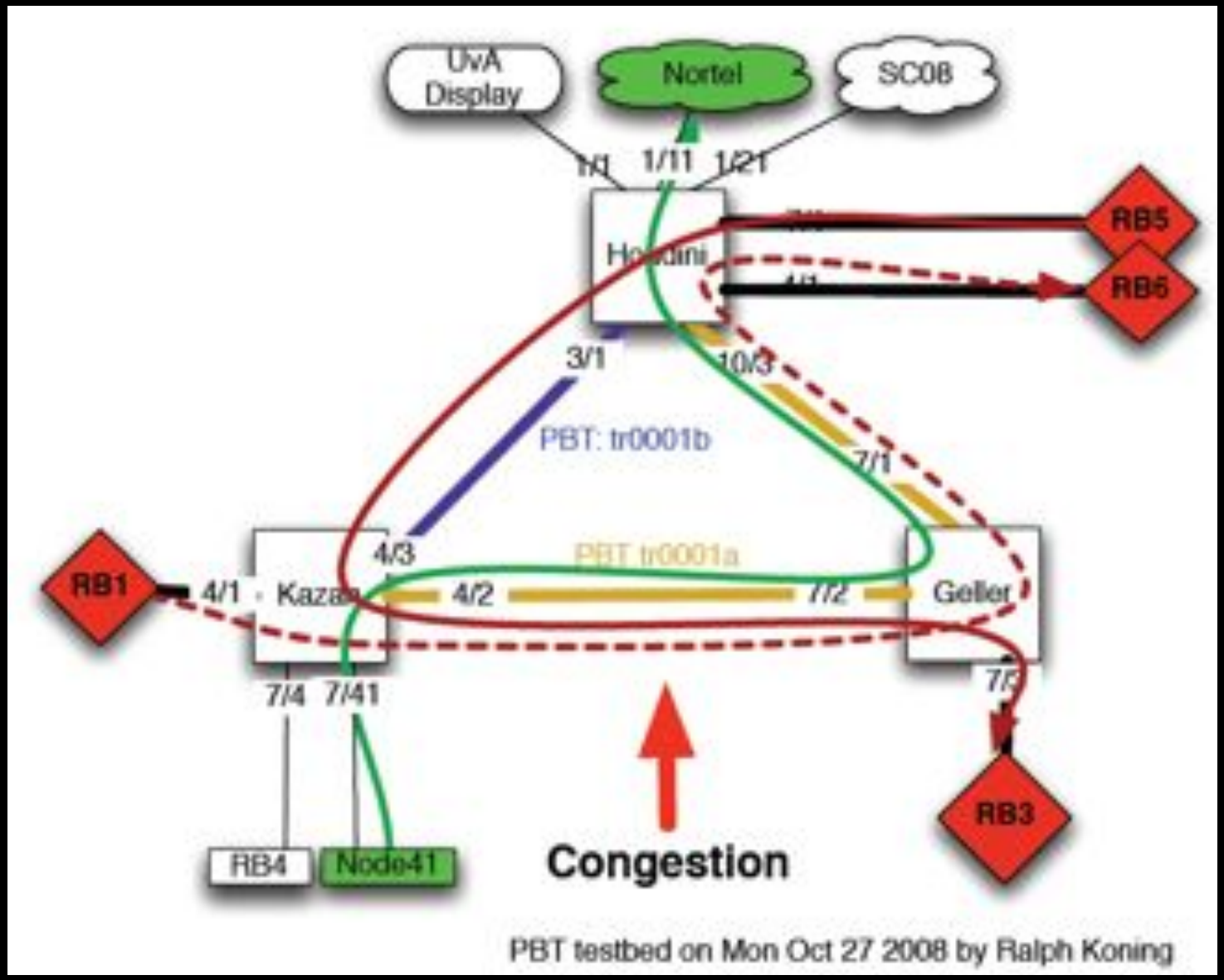


System and Network Engineering Research Group, Universiteit van Amsterdam
http://science.uva.nl/research/sne

Diagram for SAGE video streaming to ATS



UvA Testbed

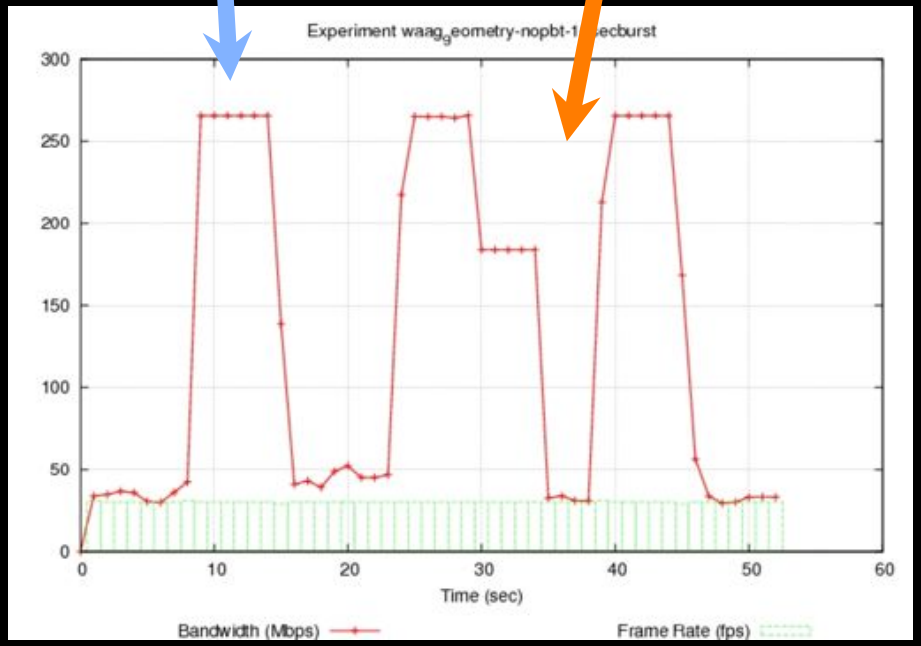


Congestion introduced in the network with multiple PBT paths carrying streamed SHD Content

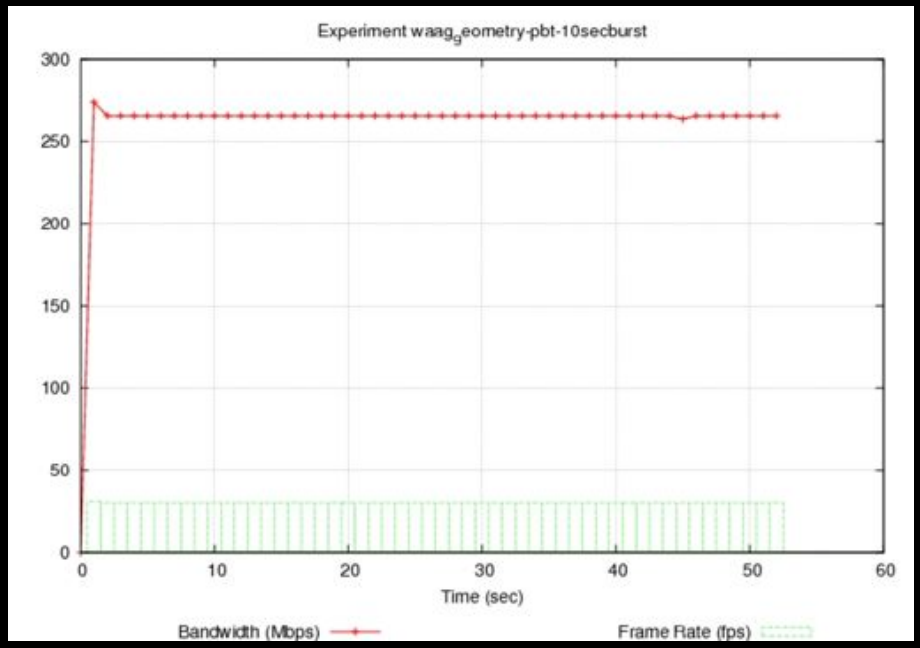
Experimental Data

Sage without background traffic

Sage with background traffic



10 Second Traffic bursts with No PBT

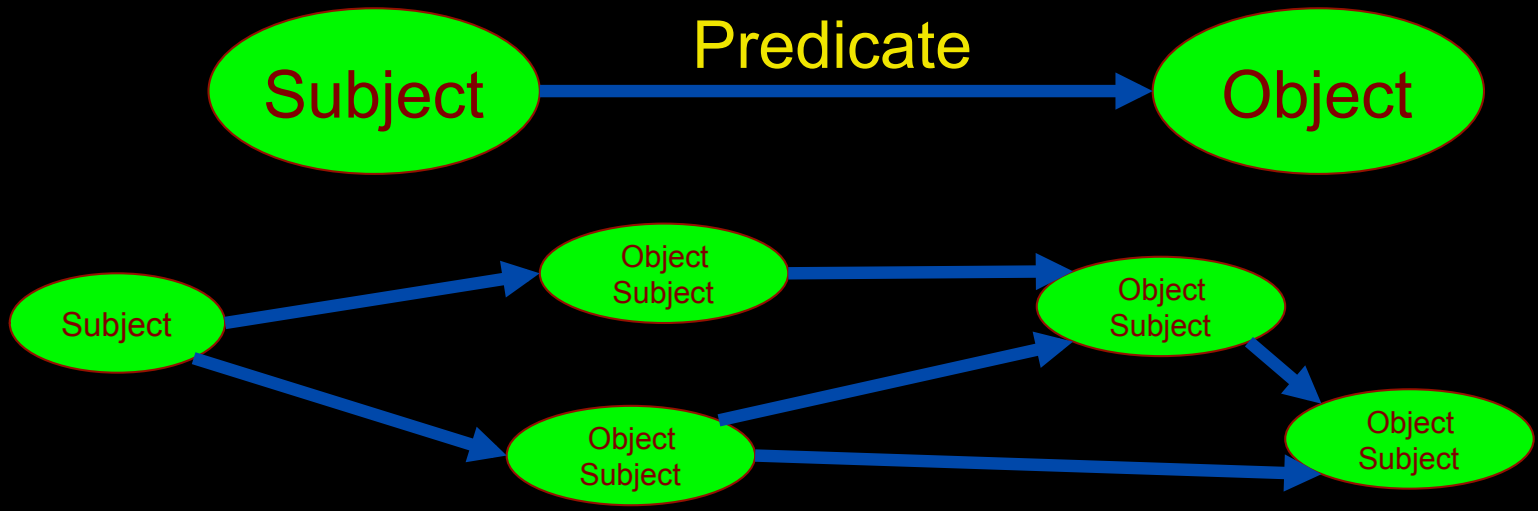


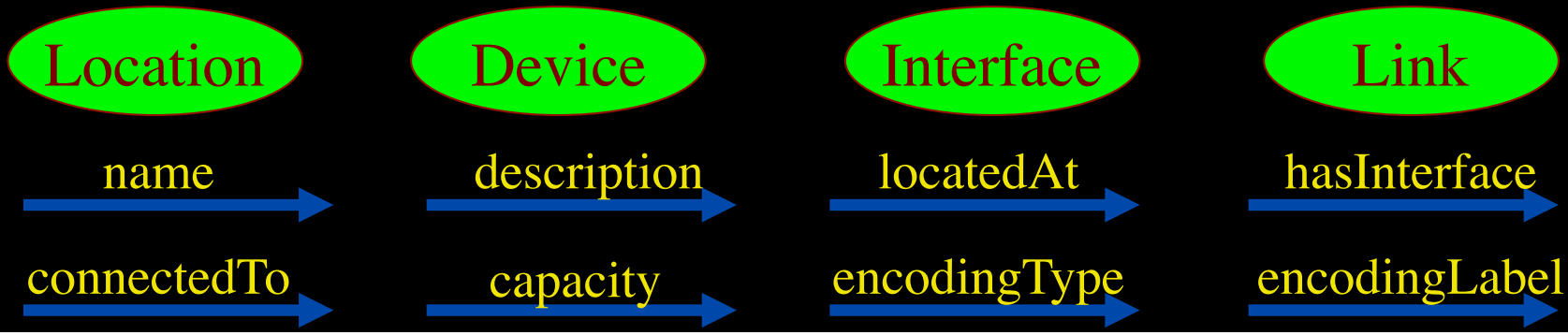
10 Second Traffic bursts with PBT

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

Network Description Language

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



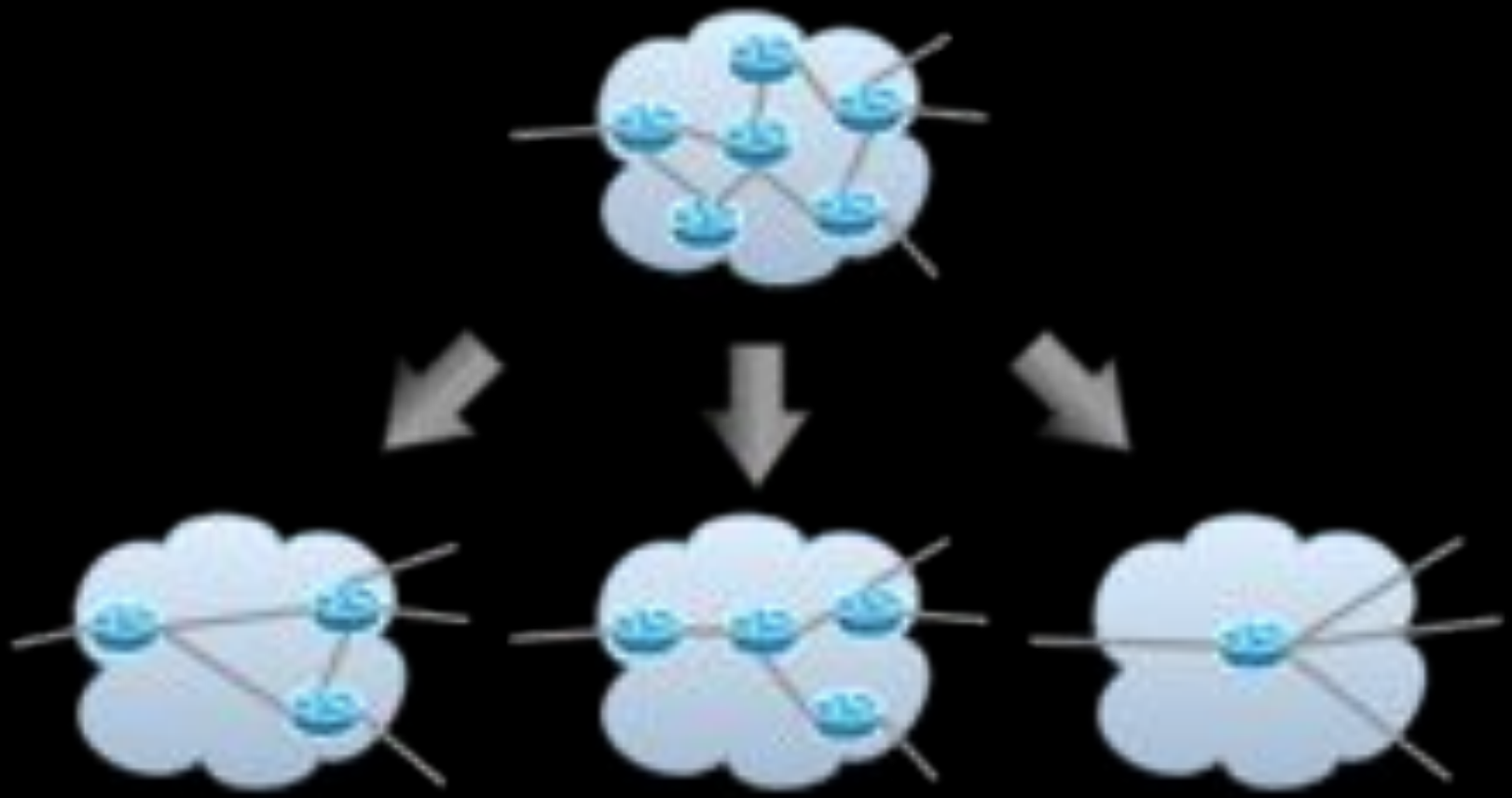


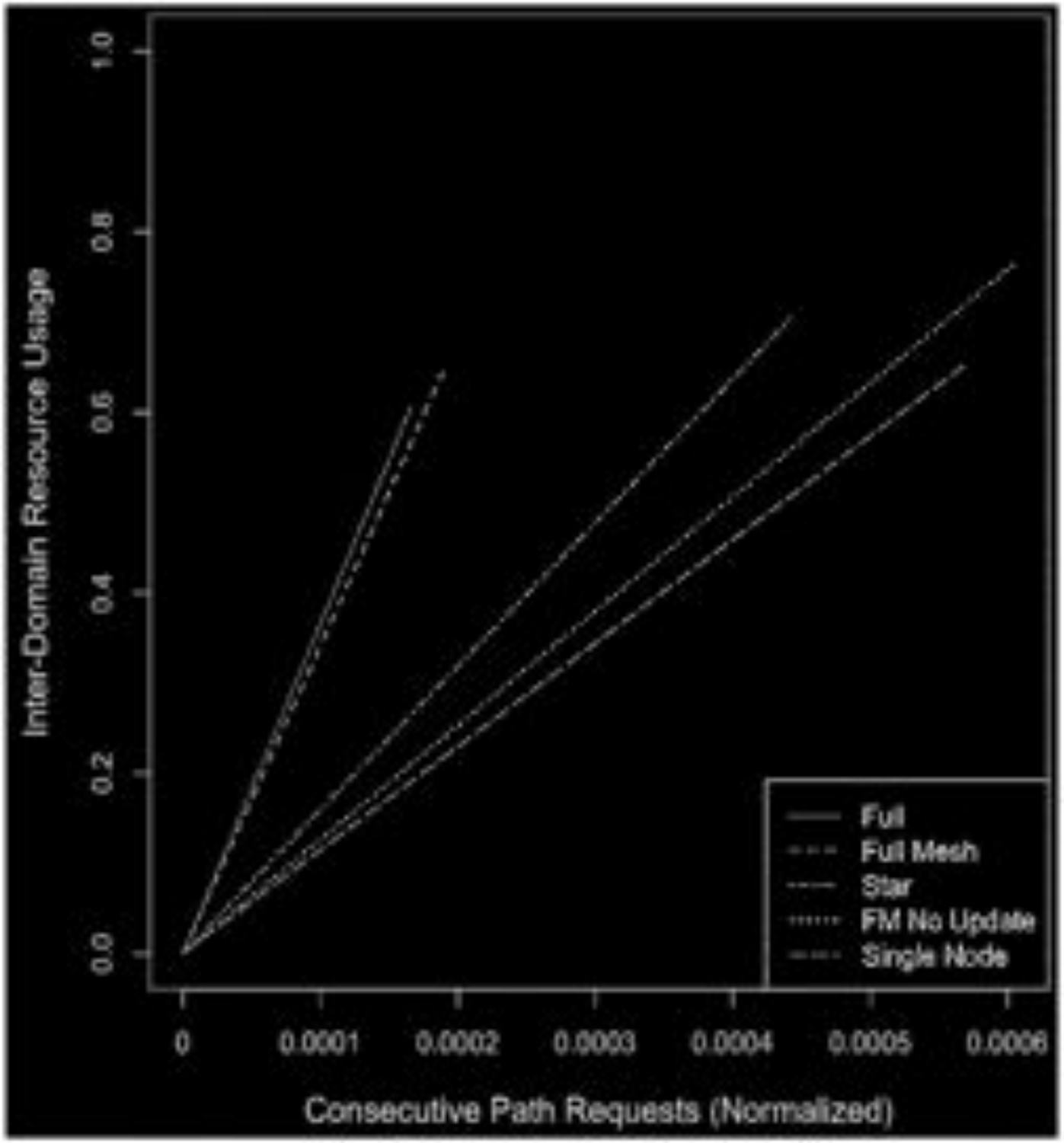
NetherLight in RDF

```
<?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:reso
      <ndl:hasInterface rdf:reso
      <ndl:hasInterface rdf:reso
      <ndl:hasInterface rdf:reso
      <ndl:hasInterface rdf:reso
      <ndl:hasInterface rdf:reso
      <ndl:hasInterface rdf:reso
```

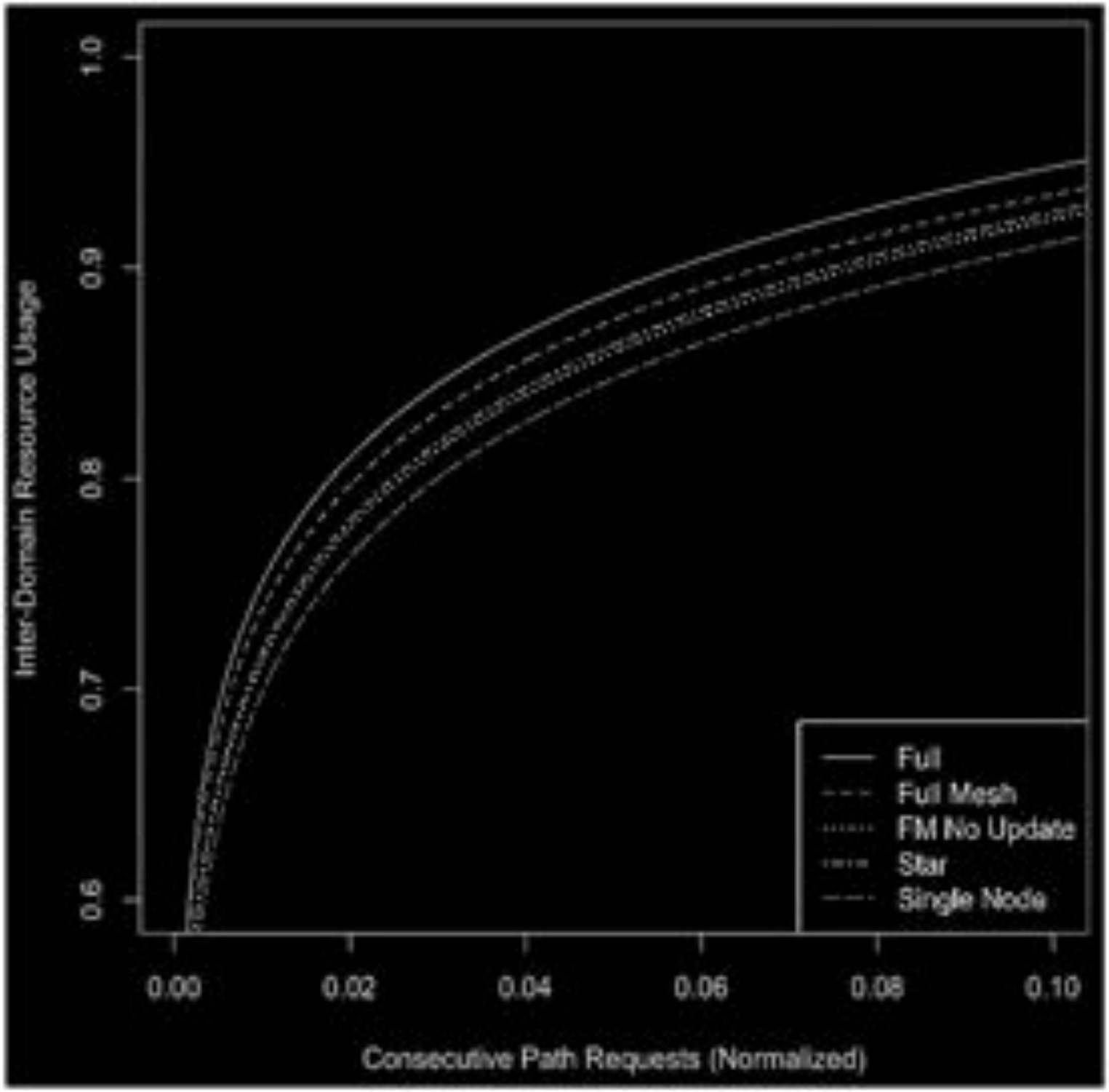
```
<!-- all the interfaces of TDM3.amsterdam1.netherlight.net -->

<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>
```

Topology Aggregation



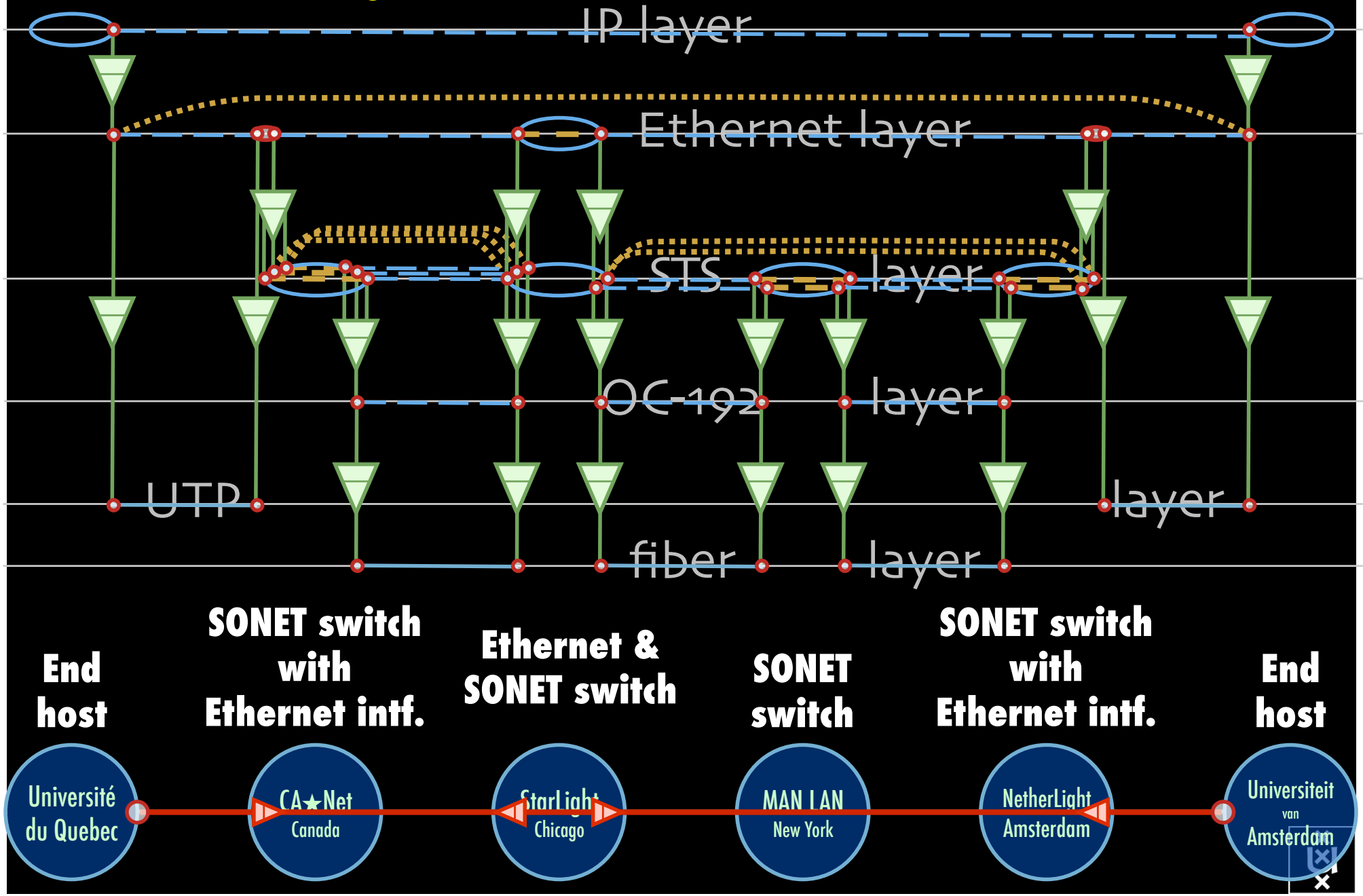


Topology Aggregation - Initial

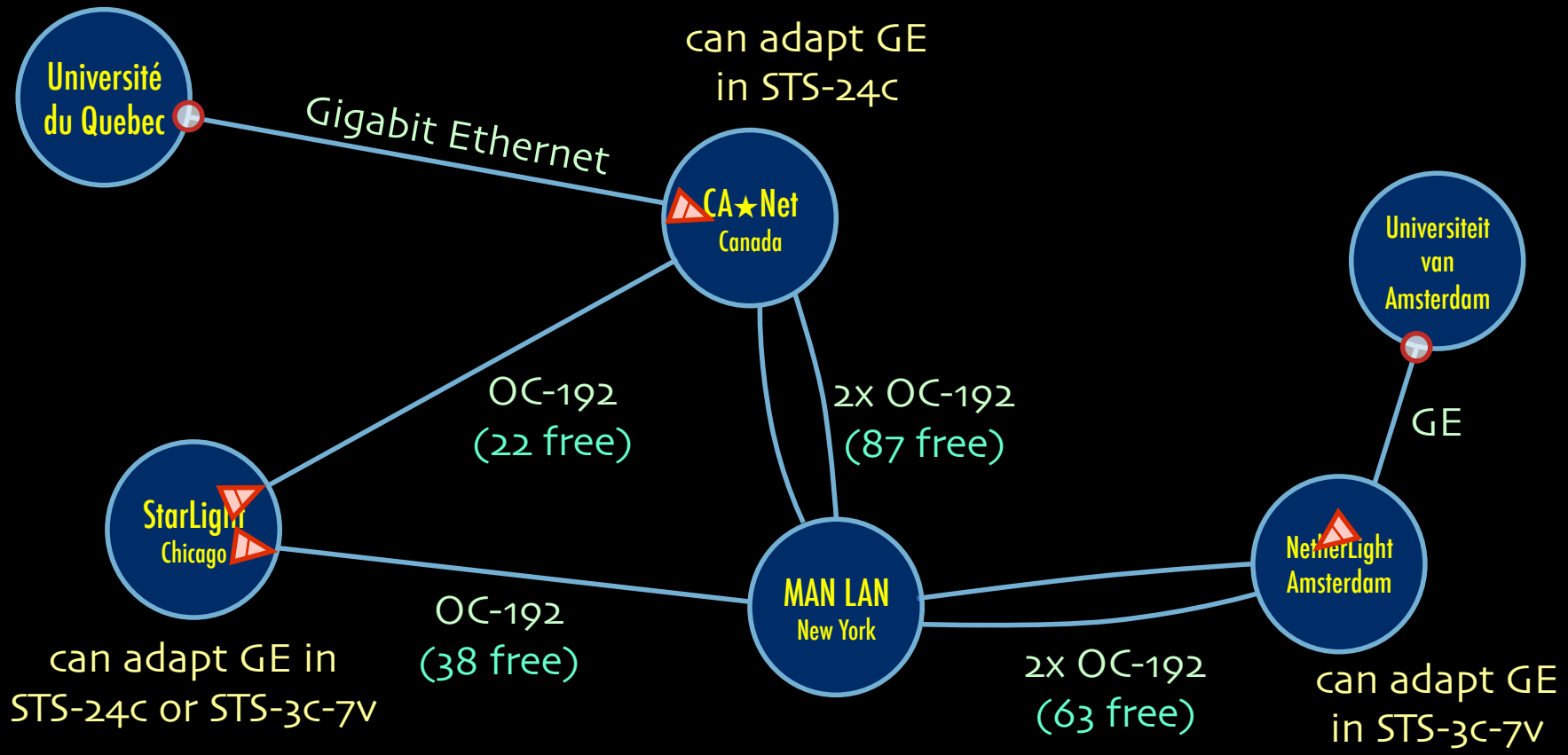


Topology Aggregation - Saturation

Multi-layer descriptions in NDL

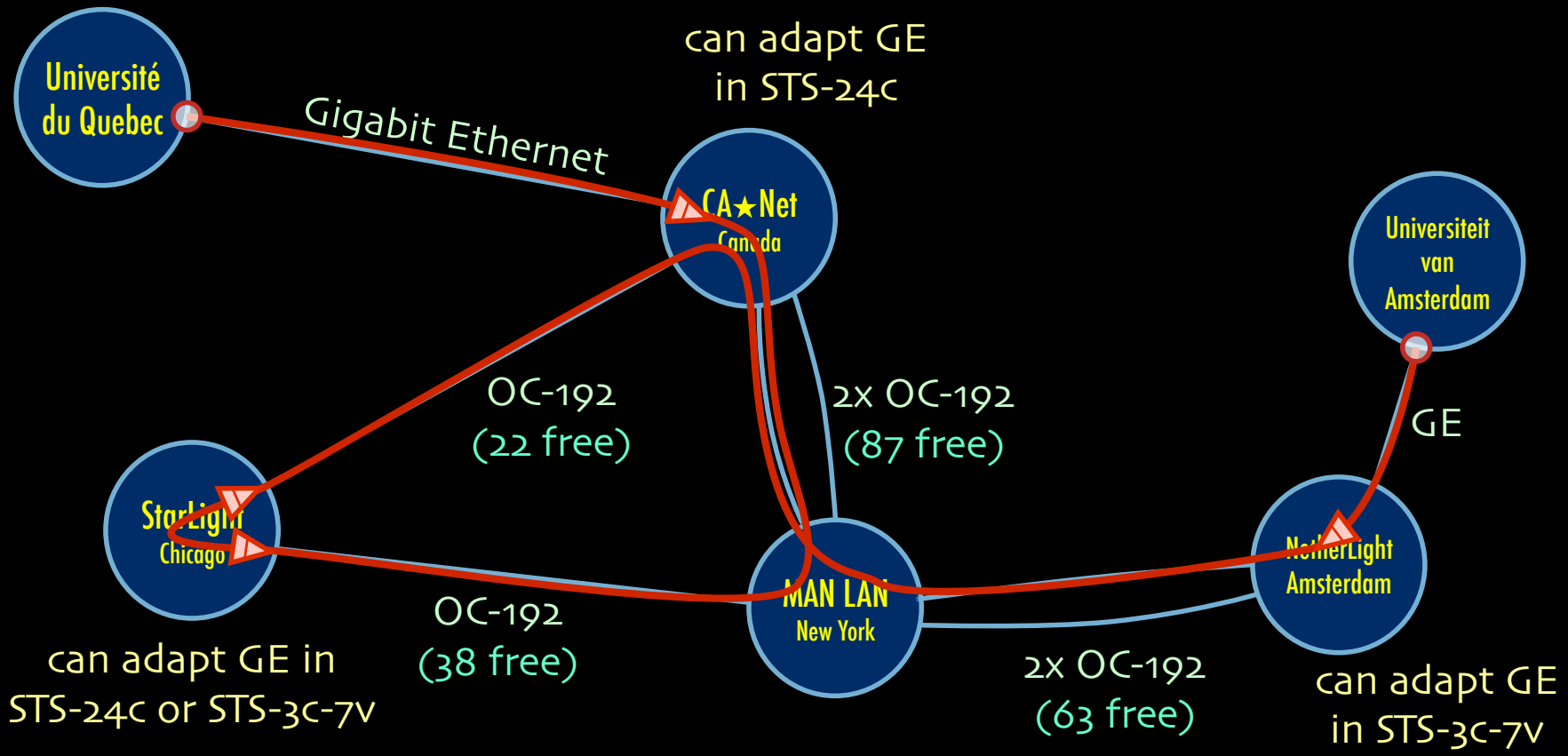


A weird example



Thanks to Freek Dijkstra & team

A weird example



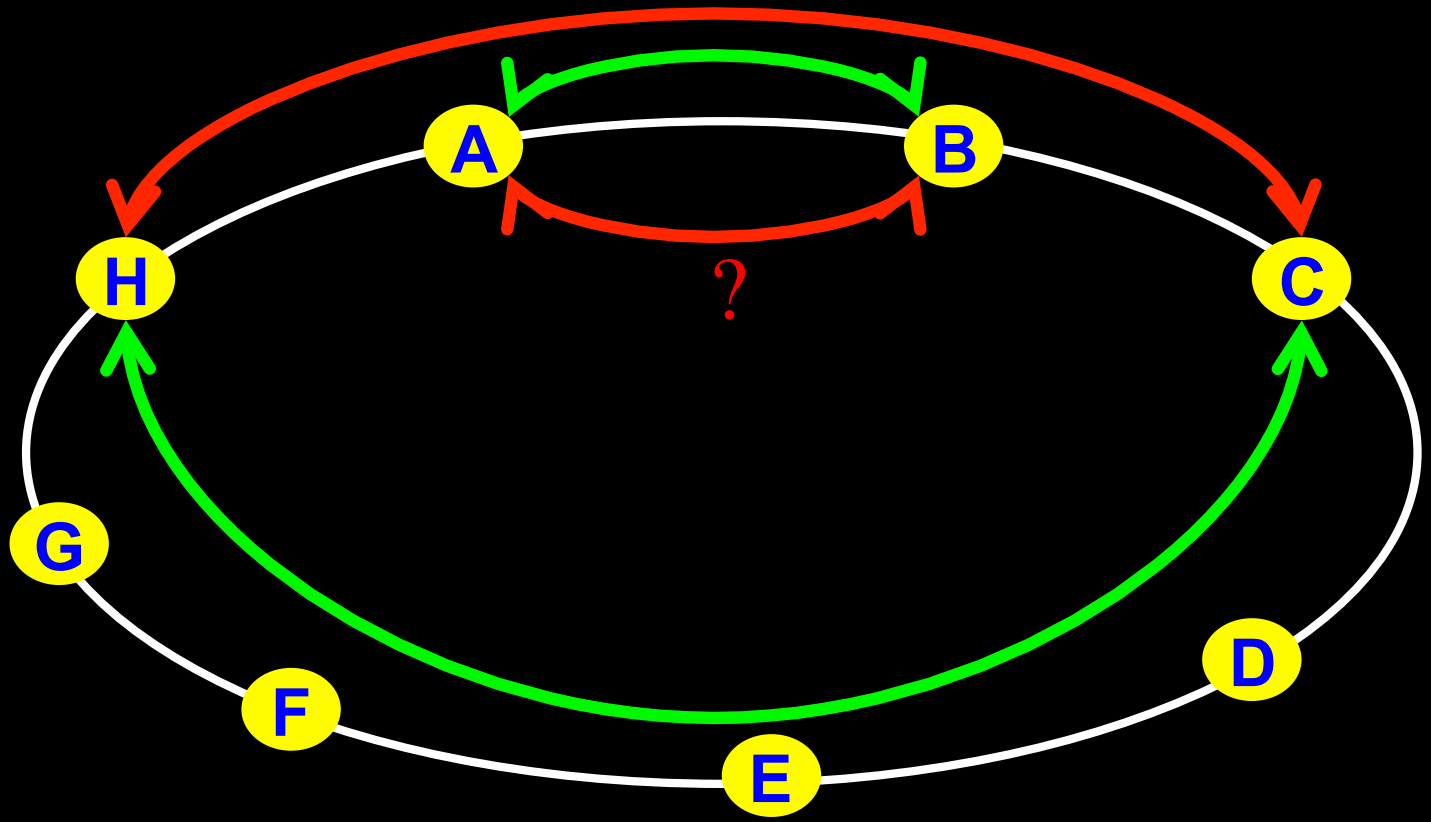
Thanks to Freek Dijkstra & team

The Problem

I want HC and AB

Success depends on the order

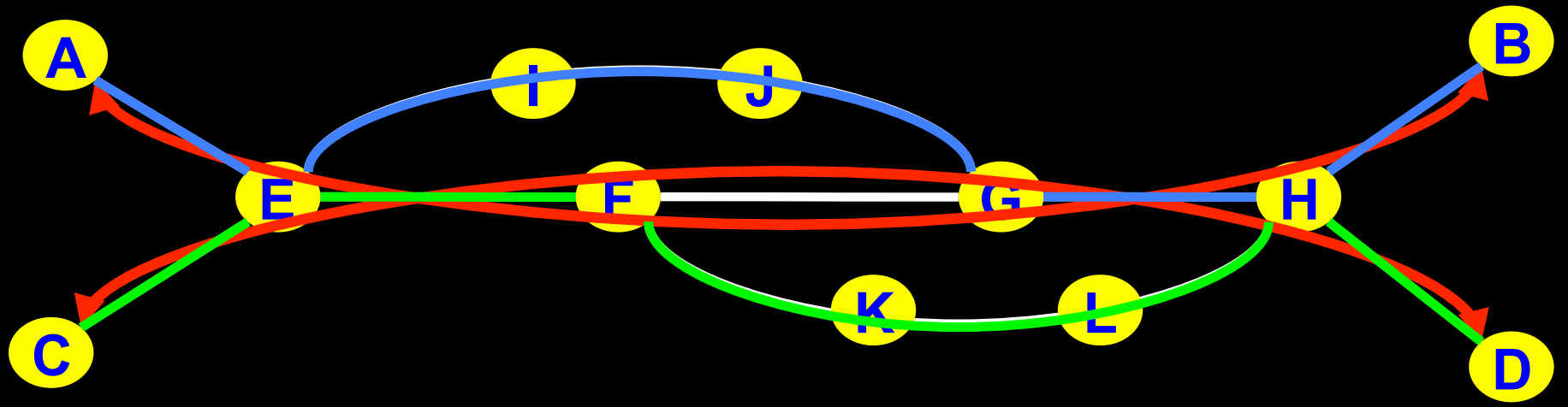
Wouldn't it be nice if I could request [HC, AB, ...]



Another one ☺

I want AB and CD

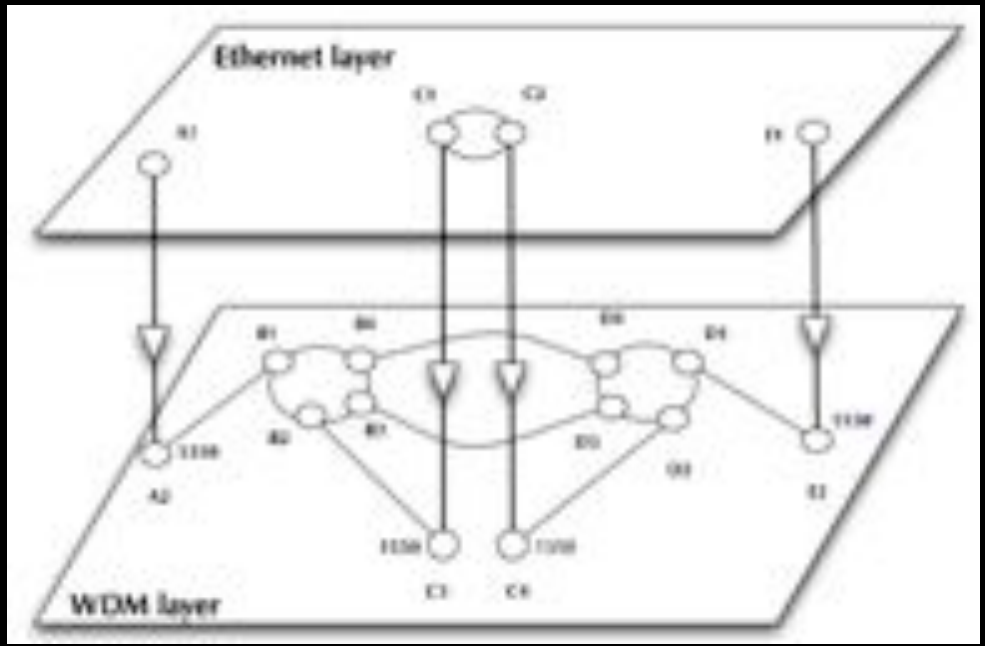
Success does not even depend on the order!!!

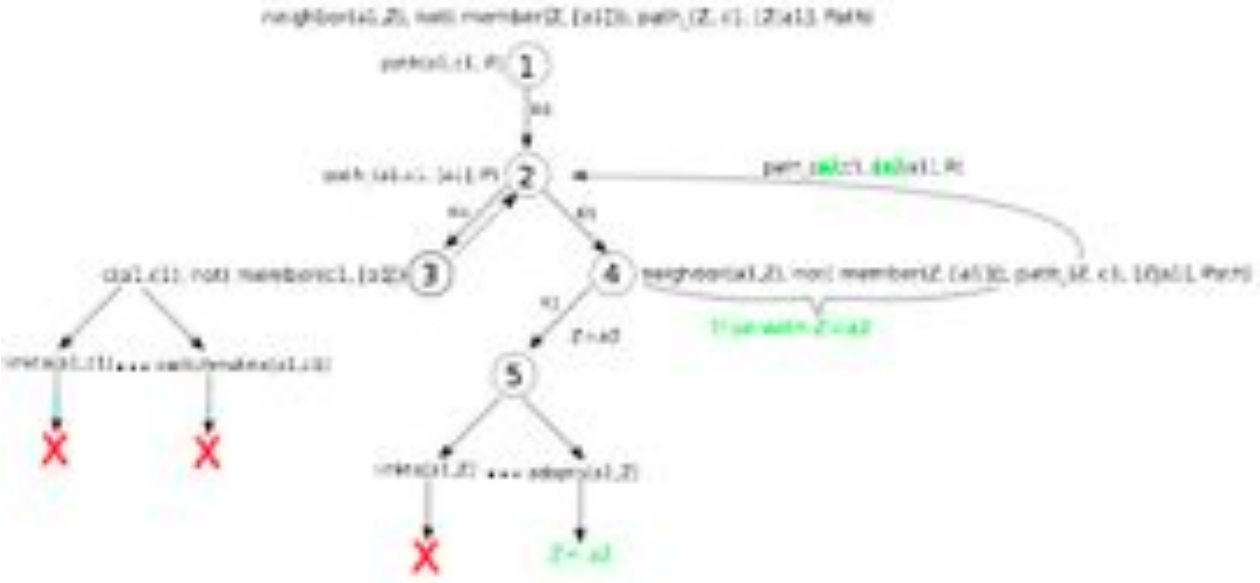


NDL + PROLOG

Research Questions:

- order of requests
- complex requests
- usable leftovers

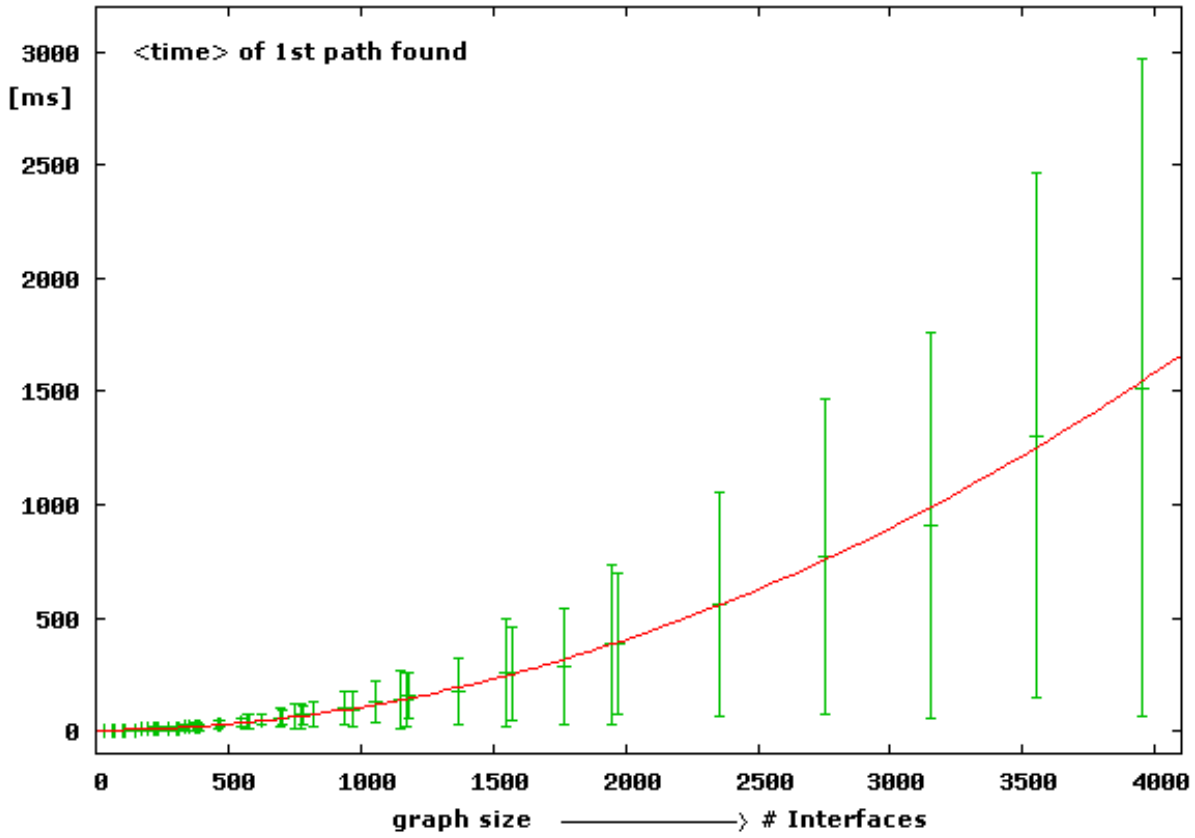




- Reason about graphs
- Find sub-graphs that comply with rules

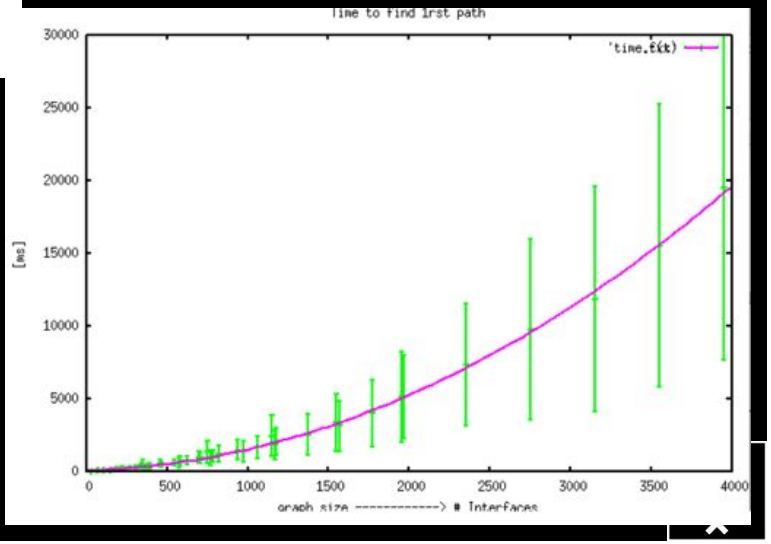
- Network descriptions are in NDL
- Use Prolog , a *logical programming* language:
  - clauses: facts and rules
  - goals: reached through backward chaining (goal-driven)
- Multi-layer pathfinding is a combinatorial bomb.
- Need features of networks to force Prolog to backtrack if it looks for an unnecessary long path.
- Introducing features (heuristics) speeds up the pathfinding but may lead to false negatives too.

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-domain 2-layer networks

How do multi-domain 2-layer networks look like?

Guess: Projection algorithm (2-layer: Ethernet /WDM)

Steps:

1. Generate a multi-domain graph by BA-algorithm
2. Generate a graph for each domain by BA-algorithm
3. For each domain graph project random nodes onto WDM layer
4. Connect the domains at each layer according to the multi-domain graph
5. Assign random wavelengths to the adaptation links

Advantage:

- Number of adaptations determined by the degree of the projected nodes
- Multi-domain Ethernet-layer as well as the multi-domain WDM-layer graph are not necessarily connected.

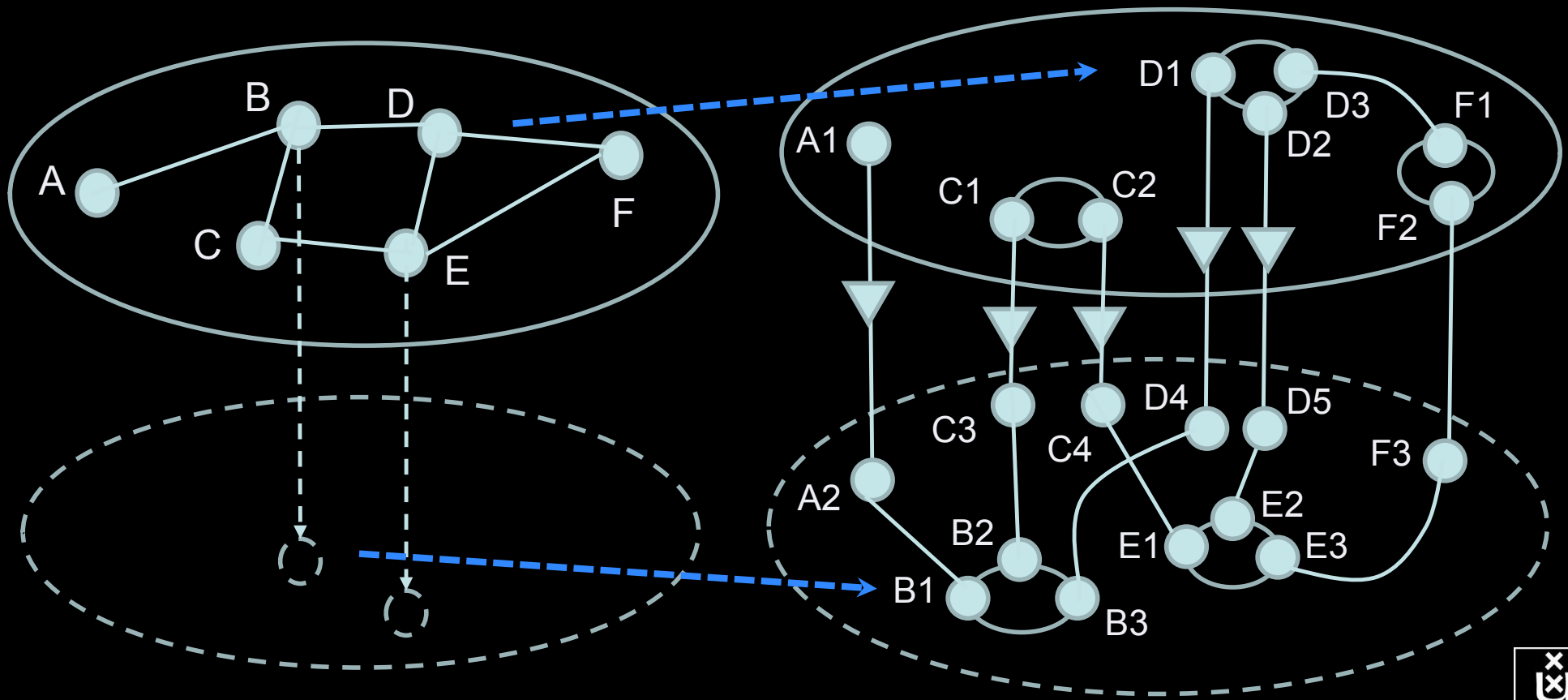
Input parameters:

- Number domains, number of nodes(devices) per domain
- Ratio of Ethernet-devices over WDM-devices per domain
- Distribution of wavelength

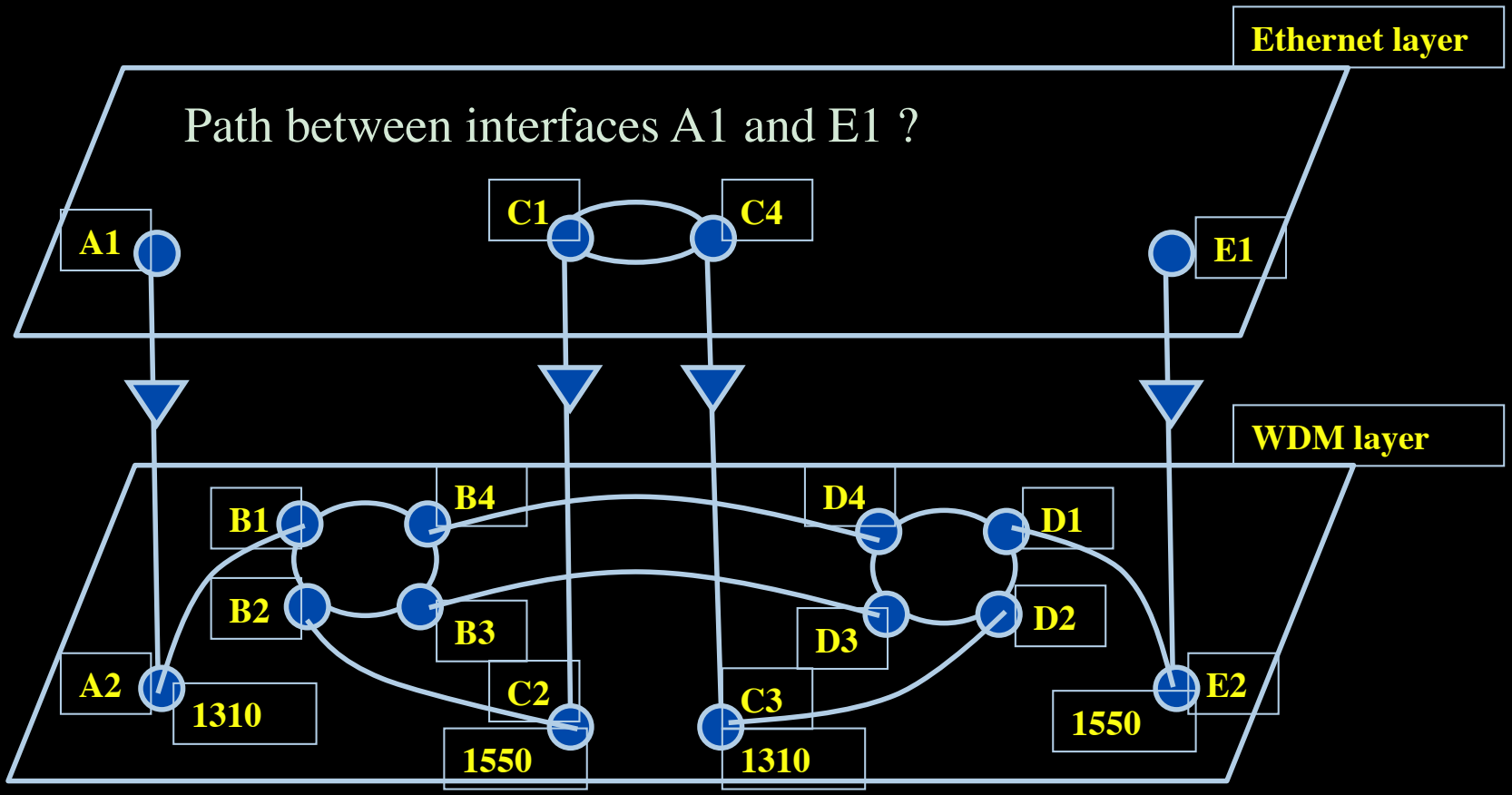
Multi-domain 2-layer networks

Projection algorithm

BA-algorithm to generate a graph for each domain
Project random nodes onto WDM layer



Multi-layer Network PathFinding



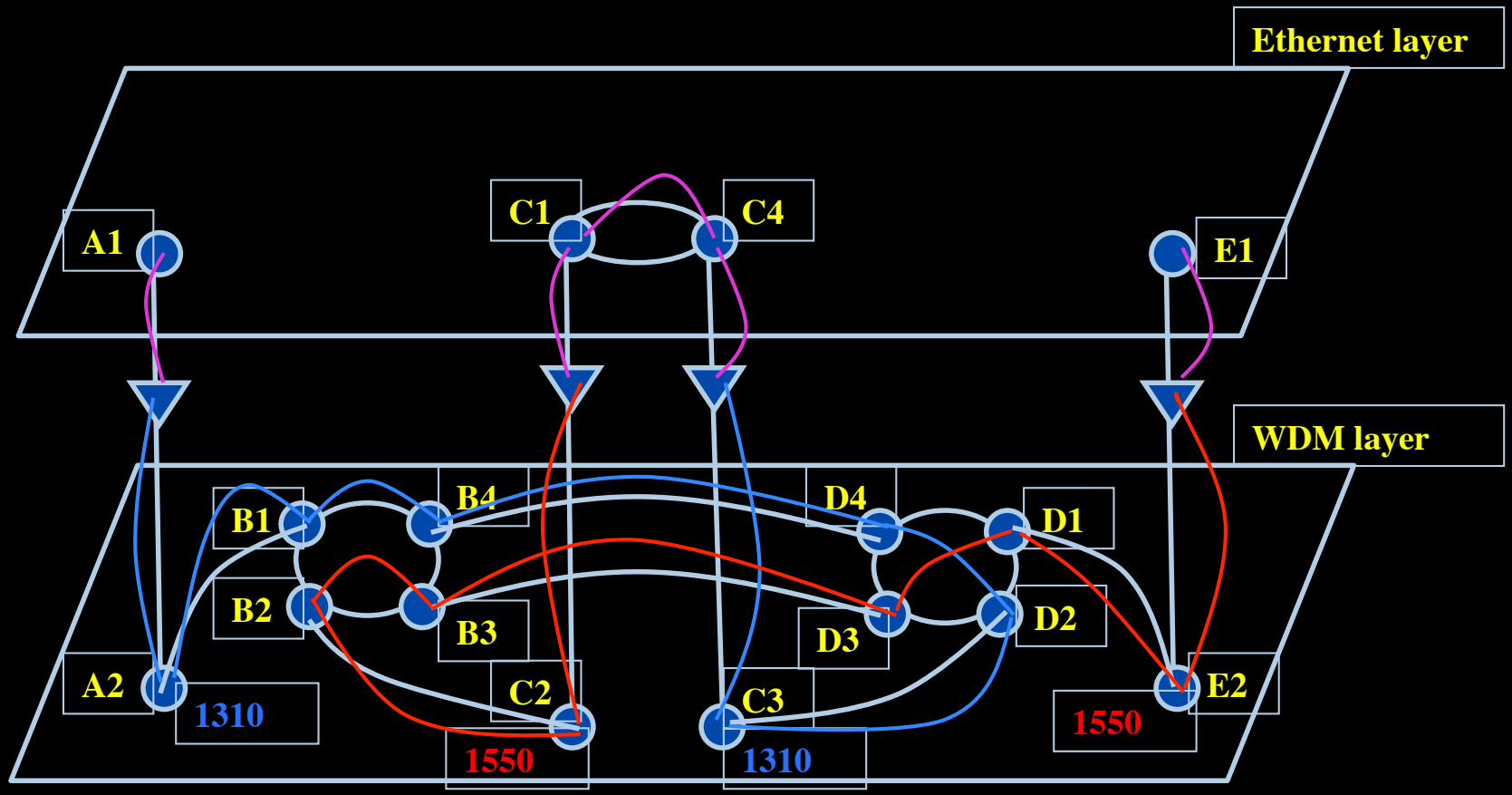
Prolog rule:

```
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

Prolog pathfinding results

DFS path constraints: Number of different wavelength

#Domains (#Ether:#WDM) (<#Intf>)(<#Adap>)	Prolog time [ms] μ(σ)	Timeouts	Success %	
3 (9:6)(55)(11)	20(4)	0	100	No max #wav
4 (48:32)(377)(73)	2620(8245)	74	92.6	
4 (96:64)(771)(147)	6592(11802)	207	79.3	
3 (9:6)(55)(11)	20(4)	0	100	#wav ≤3
4 (48:32)(377)(73)	1303(5052)	22	97.8	
4 (96:64)(771)(147)	3910(10045)	51	94.9	
3 (9:6)(55)(11)	20(4)	0	100	#wav ≤2
4 (48:32)(377)(73)	755(3210)	8	98.9	
4 (96:64)(771)(147)	3240(9052)	38	96.1	

Prolog pathfinding results

Parallel calls: A->B and B->A

Projection: A->B

#Domains (#Ether:#WDM) (<#Intf>)(<#Adap>)	Prolog time [ms] μ(σ)	Timeouts	Success %
3 (9:6)(55)(11)	20(4)	0	100
4 (48:32)(377)(73)	755(3210)	8	98.9*
4 (96:64)(771)(147)	3240(9052)	38	96.1*

#wav ≤2

Projection: first of A->B and B->A

3 (9:6)(55)(11)	19(1)	0	100
4 (48:32)(377)(73)	144(486)	0	100
4 (96:64)(771)(147)	601(2722)	2	99.6*

#wav ≤2

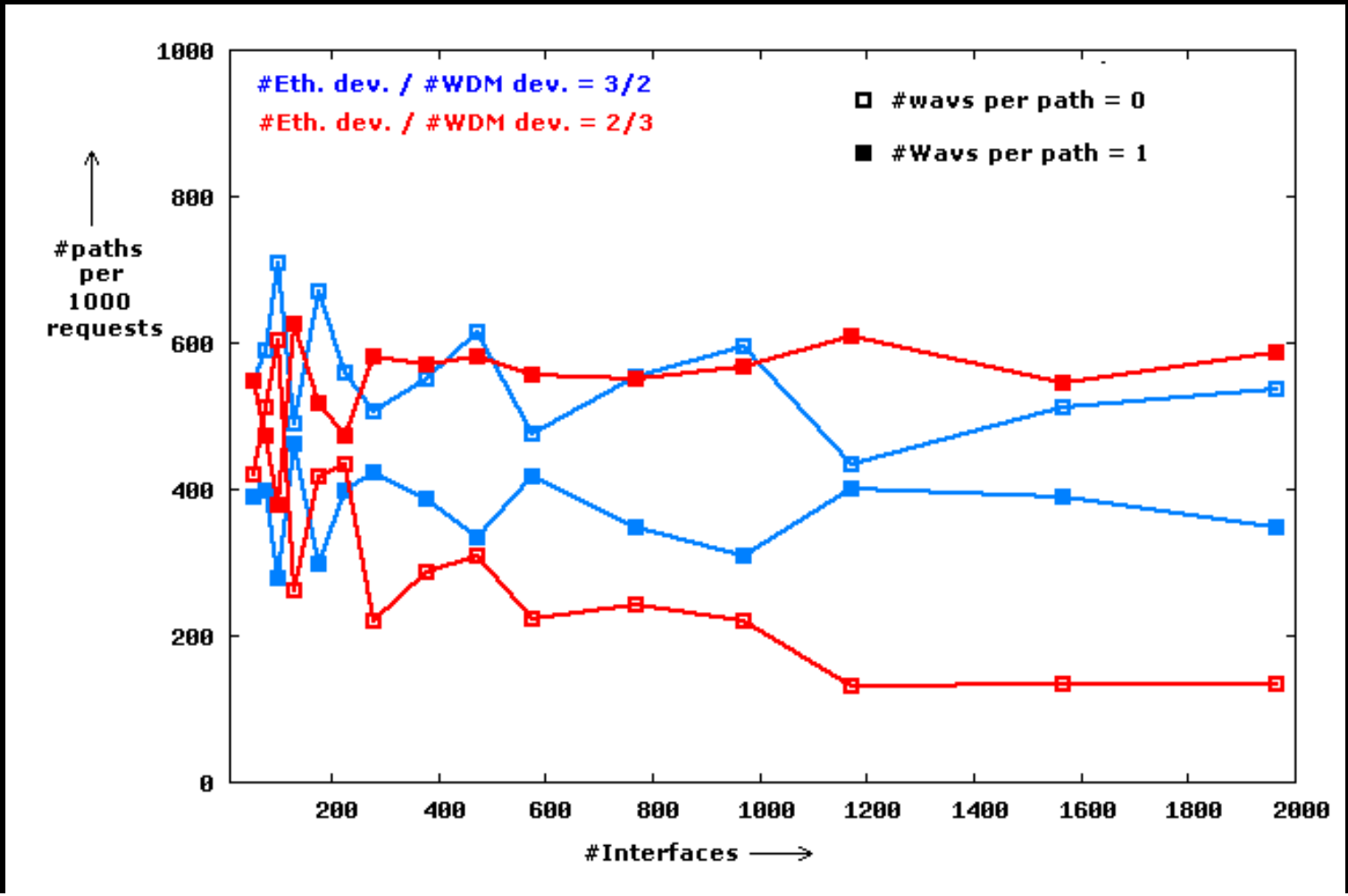
*false negatives also taken into account

Network features

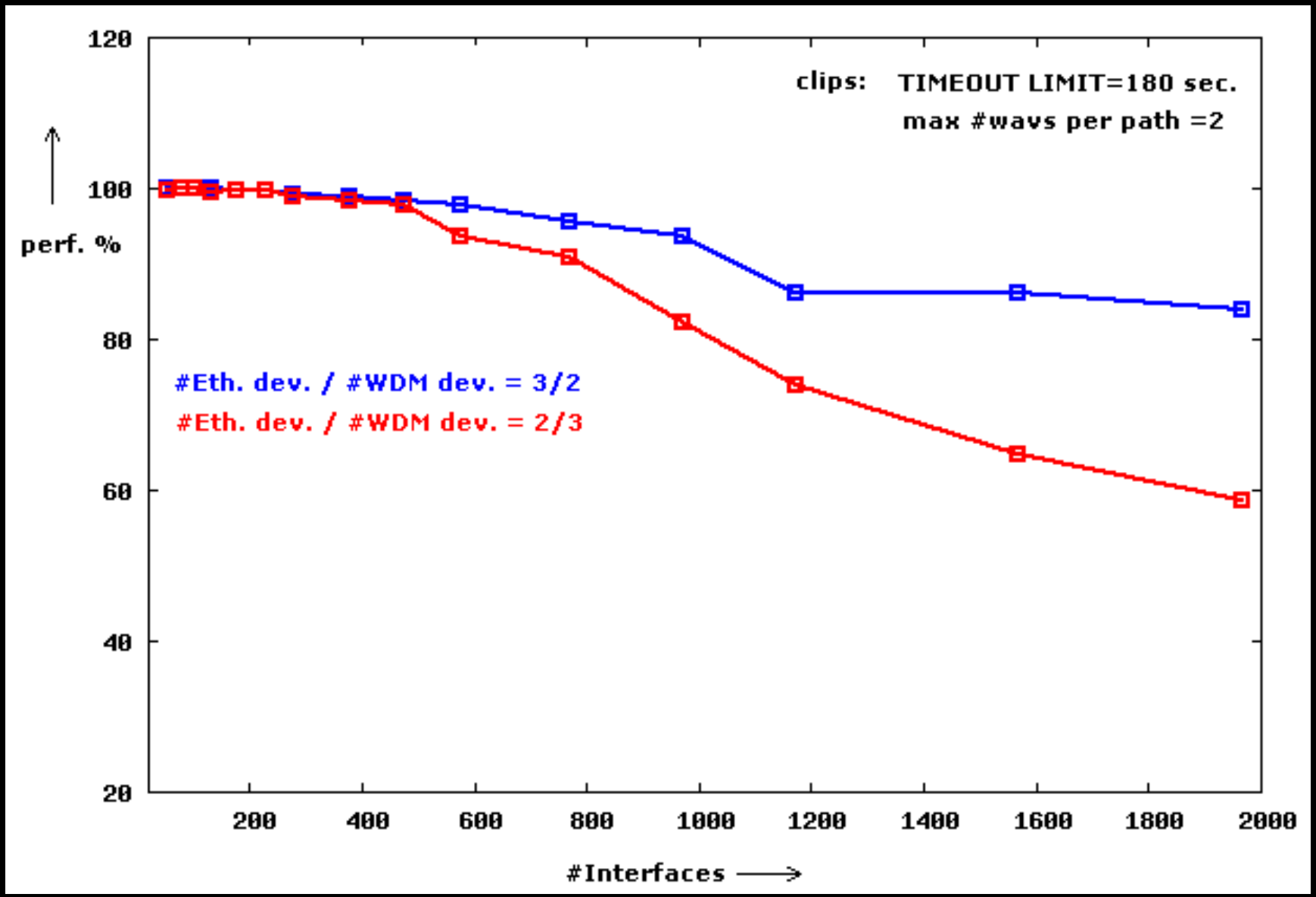
1000 random src-dst pairs per 50 topologies of a fixed size (#Interfaces):

#Eth. Dev / #WDM dev. = 3/2 : at least 90% of the paths with at most 1 wavelength

#Eth. Dev / #WDM dev. = 2/3 : #paths with 0 wavelengths decreases to 10% at least 50% of the paths with 1 wavelength.



Performance Prolog Depth-First Search

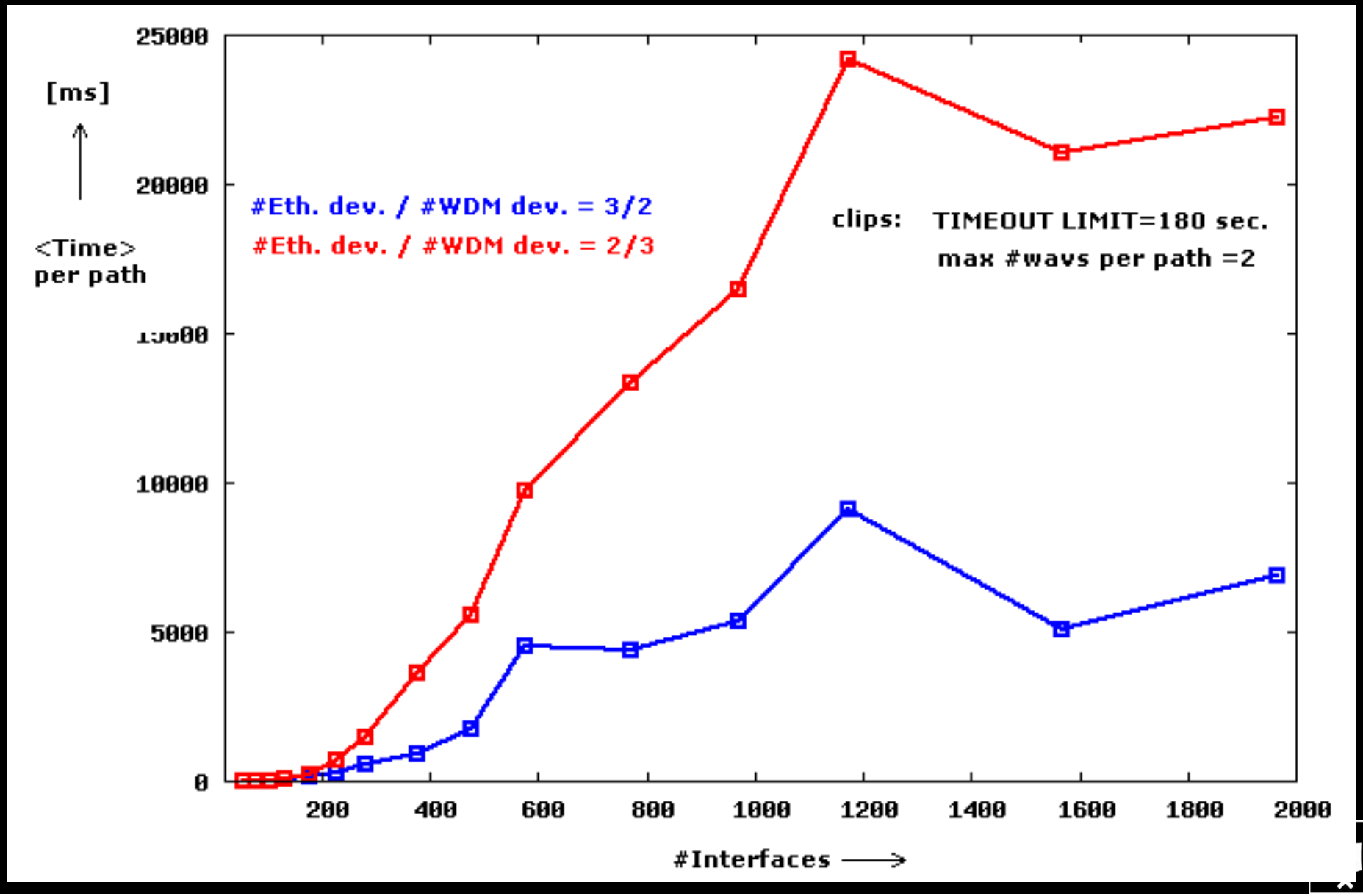


Performance drop mainly due to Timeout limit

False negatives due to max #wavelengths clip less than 1% of #paths

DAS3 cluste

Time Prolog Depth-First Search



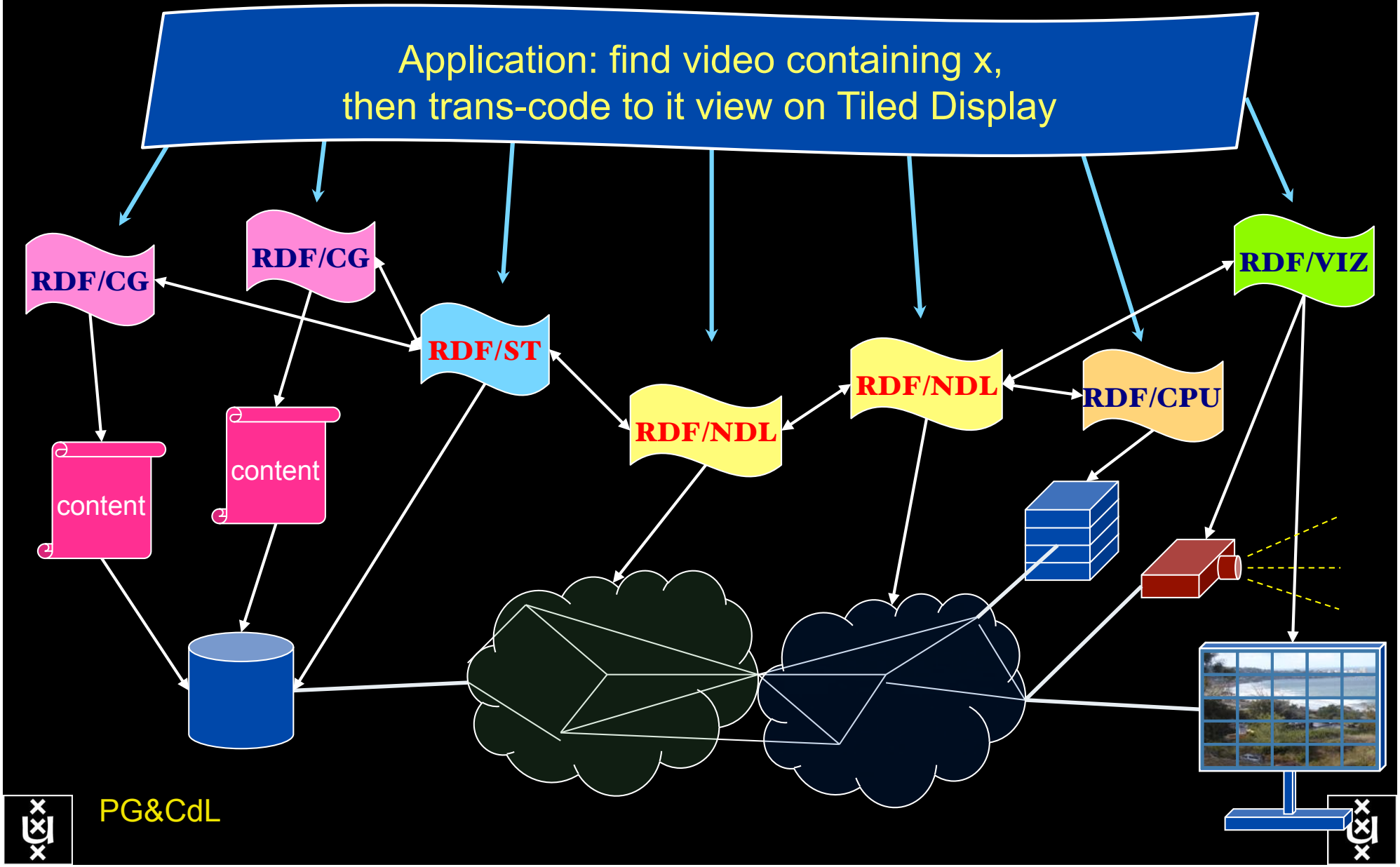
Activities

- RDF Infrastructure
- Integration of NDL and Fenius
- Cooperation with ESnet on OSCARS
- Standardization

Standardization

- NML is slowly progressing
  - Schema Document
- NSI is working frantically
  - Terminology Glossary
  - Architecture Document
  - NSI Protocol Document

RDF describing Infrastructure



Applications and Networks become aware of each other!

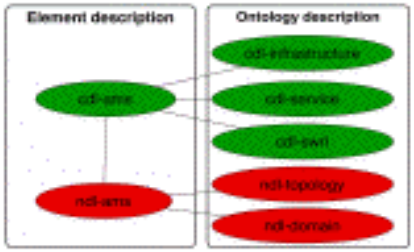
CineGrid Description Language

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

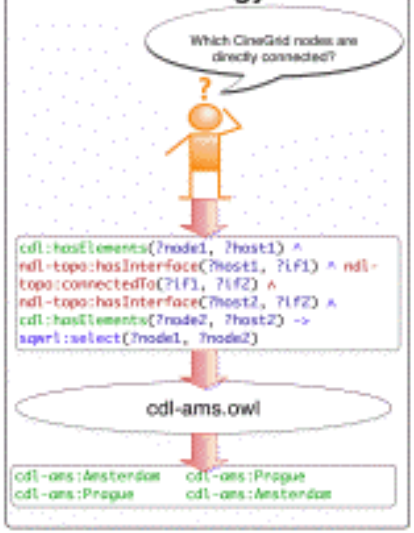
The CineGrid Description Language (CDL) describes CineGrid resources. Streaming, display and storage components are organized in a hierarchical way.

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.



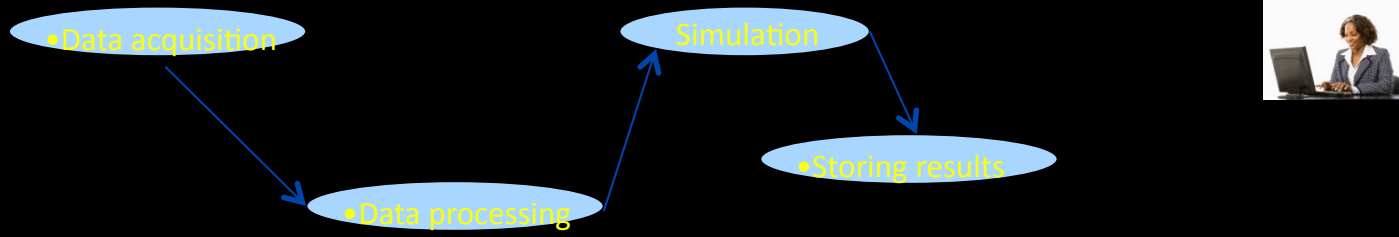
SQWRL is used to query the Ontology.



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.

Workflow execution: mapping between resources



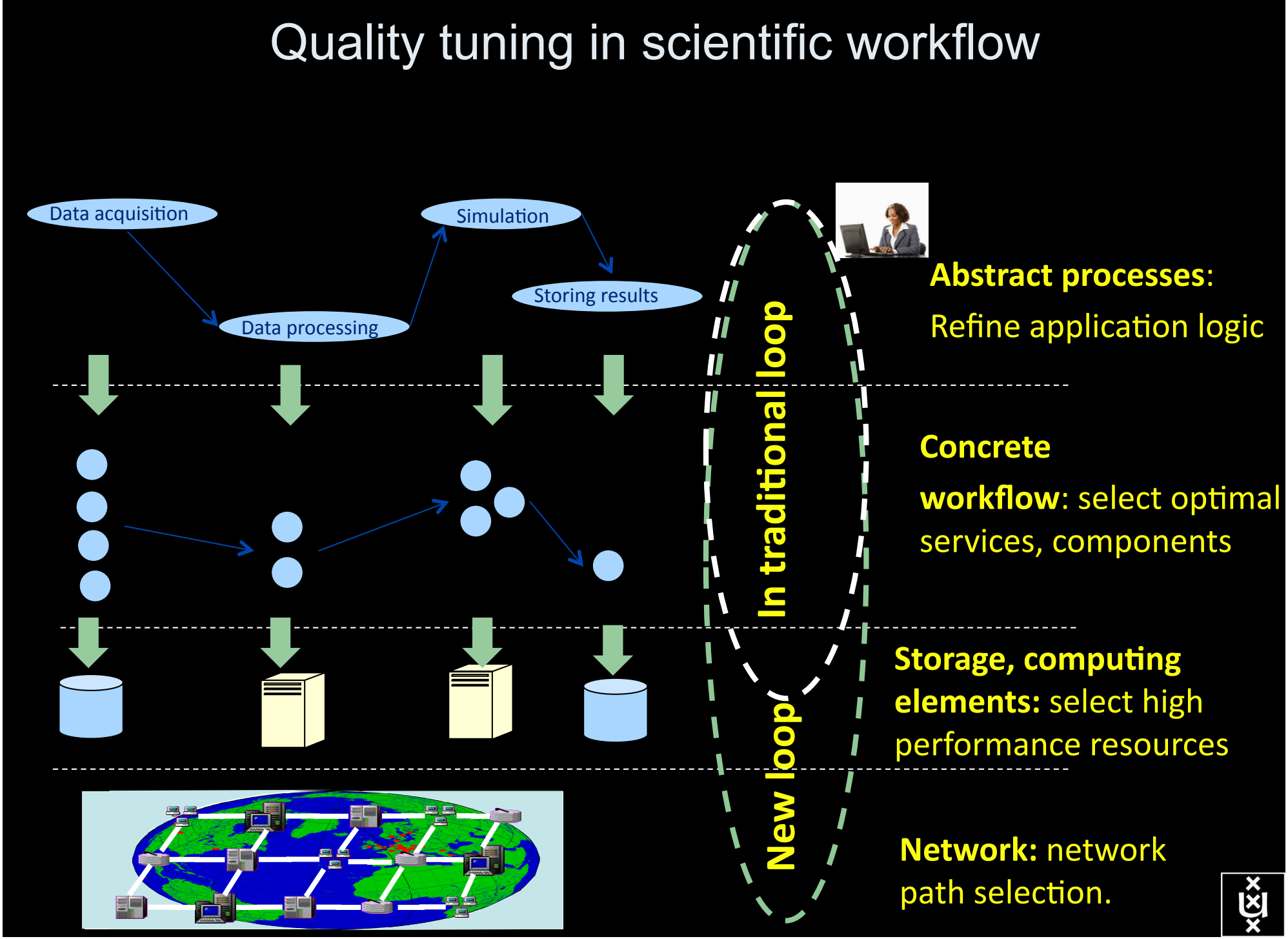
Abstract processes

Concrete workflow

Storage, computing elements

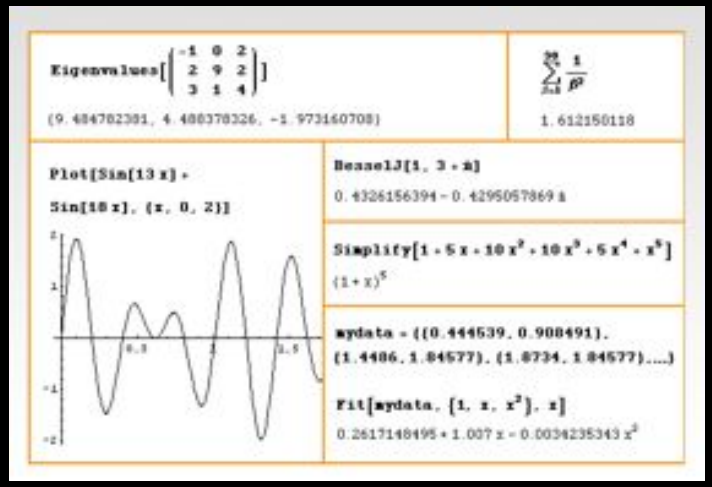
Network

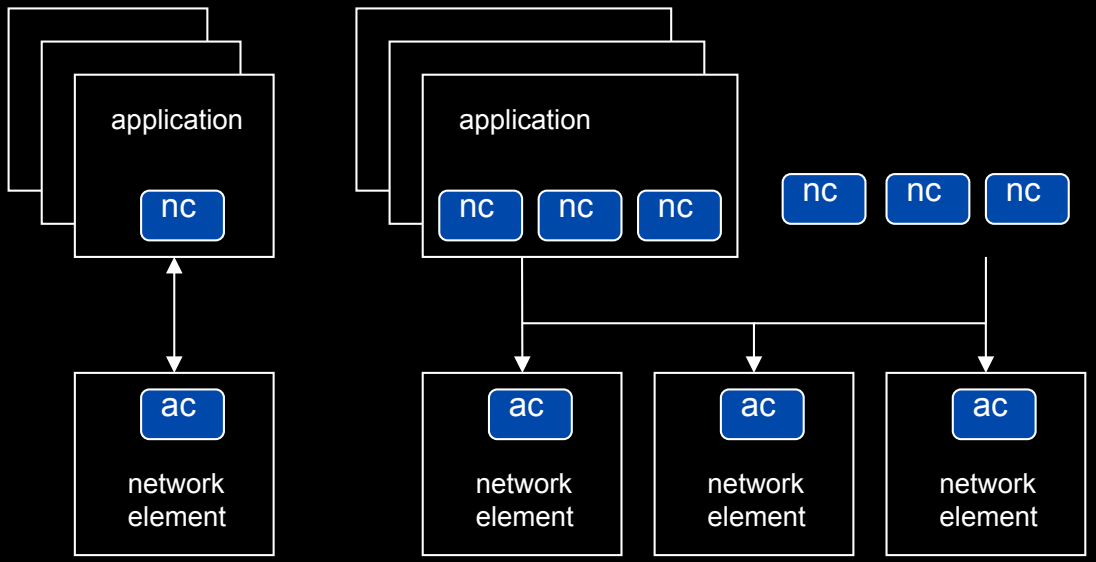
Quality tuning in scientific workflow



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

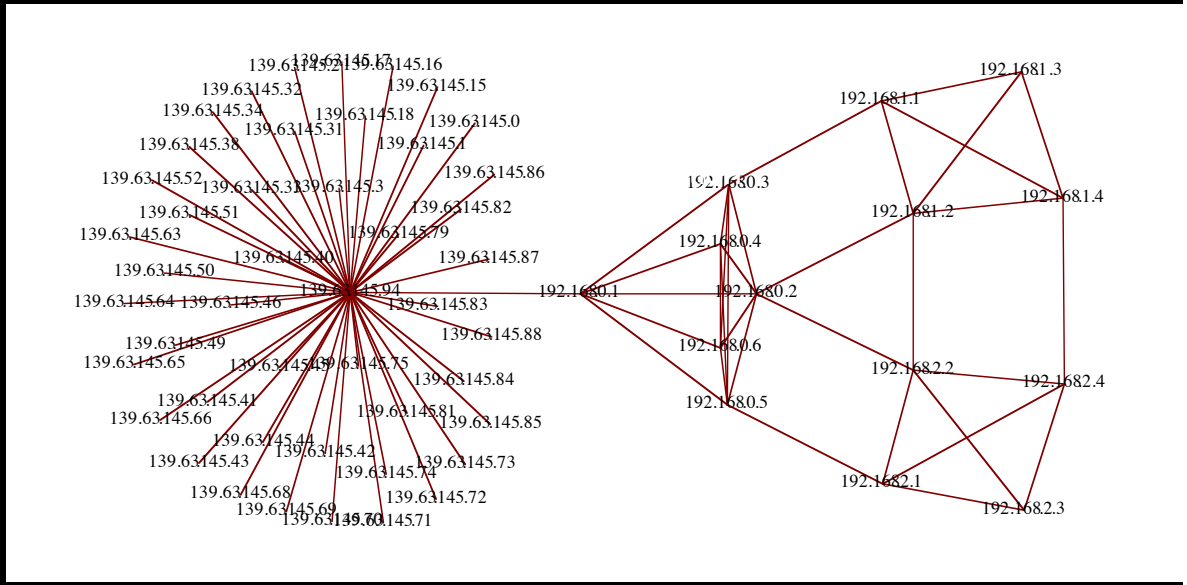
Initialization and BFS discovery of NEs

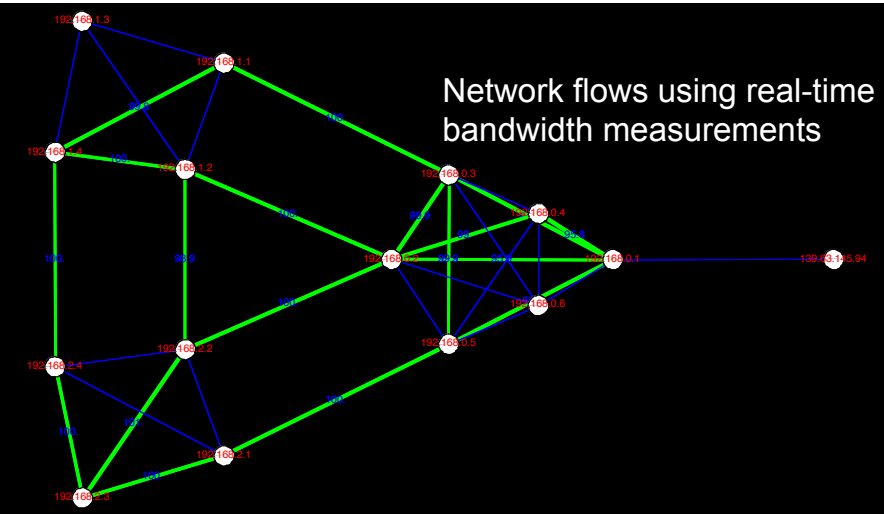
```
Needs["WebServices`"]
<<DiscreteMath`Combinatorica`
<<DiscreteMath`GraphPlot`
InitNetworkTopologyService["edge.ict.tno.nl"]

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

Hybrid computing

Routers ⇐ ⇒ Supercomputers

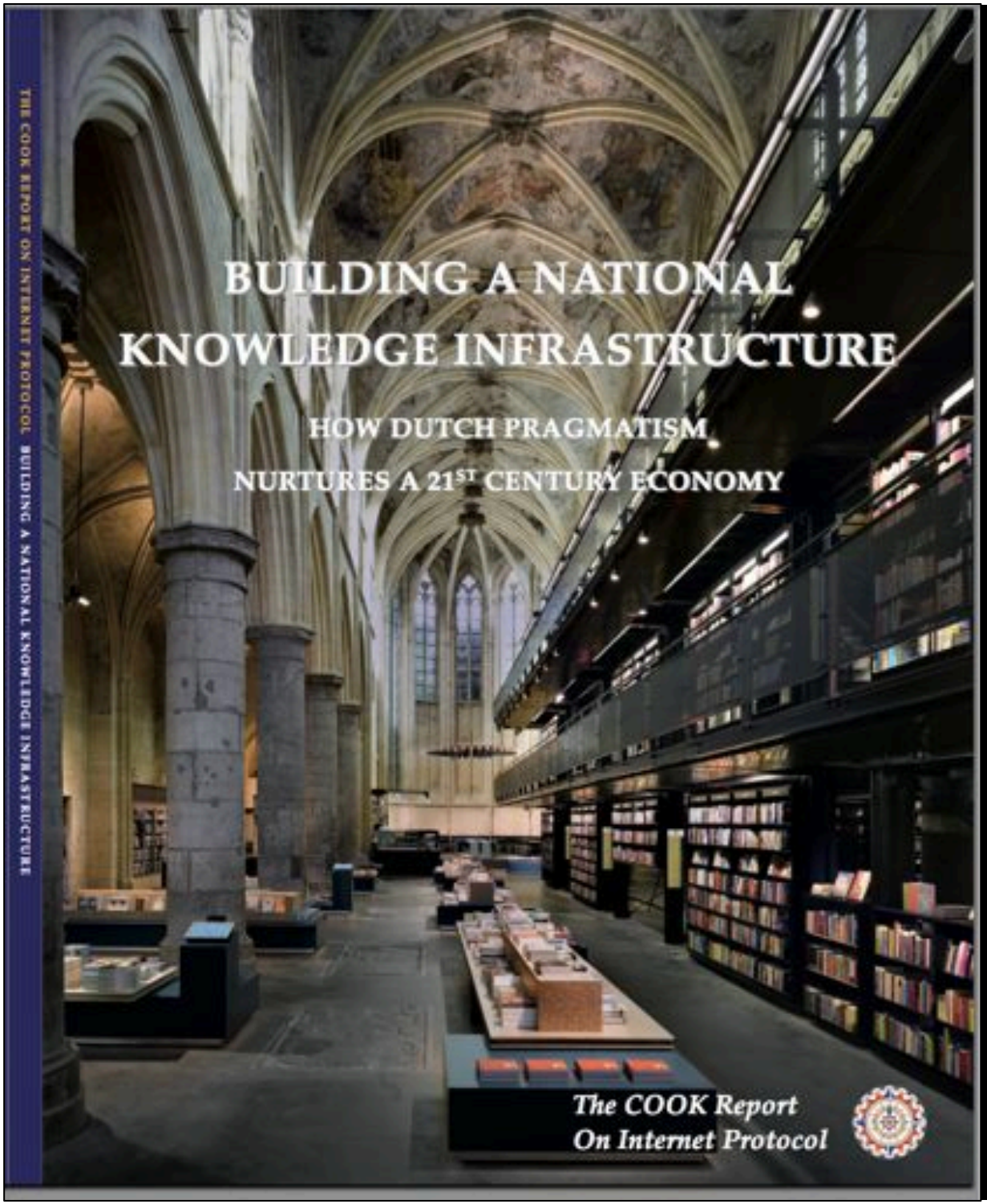
Ethernet switches ⇐ ⇒ Grid & Cloud

Photonic transport ⇐ ⇒ GPU's

What matters:

Energy consumption/multiplication

Energy consumption/bit transported



Questions ?

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

SURF NET

ext.delaat.net