# Hybrid Networking for eScience


## Cees de Laat

## EU

## SURFnet

## SURF-eScience

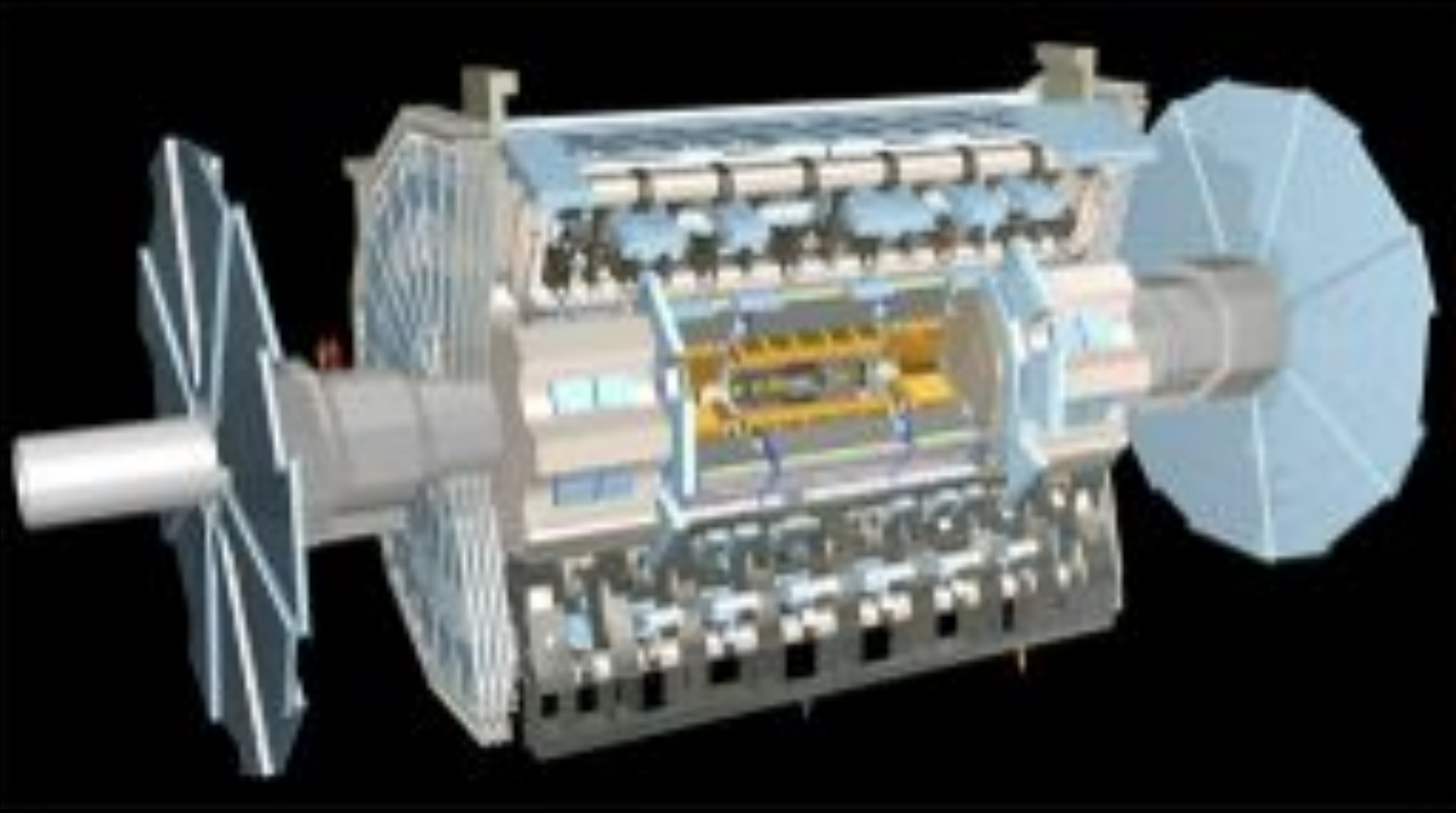
## NWO

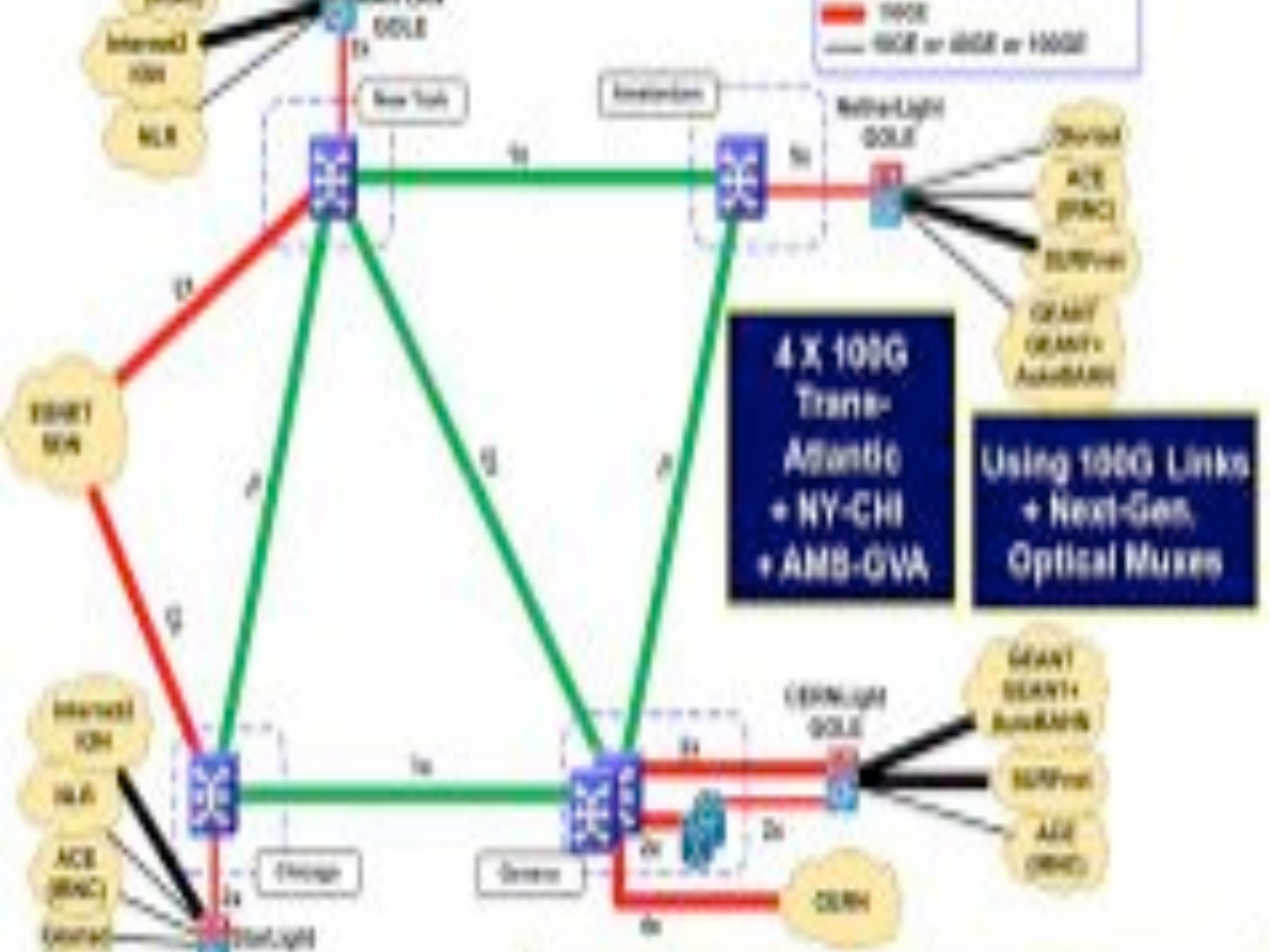
## University of Amsterdam

TNO

NCF

# ATLAS detector @ CERN Geneve



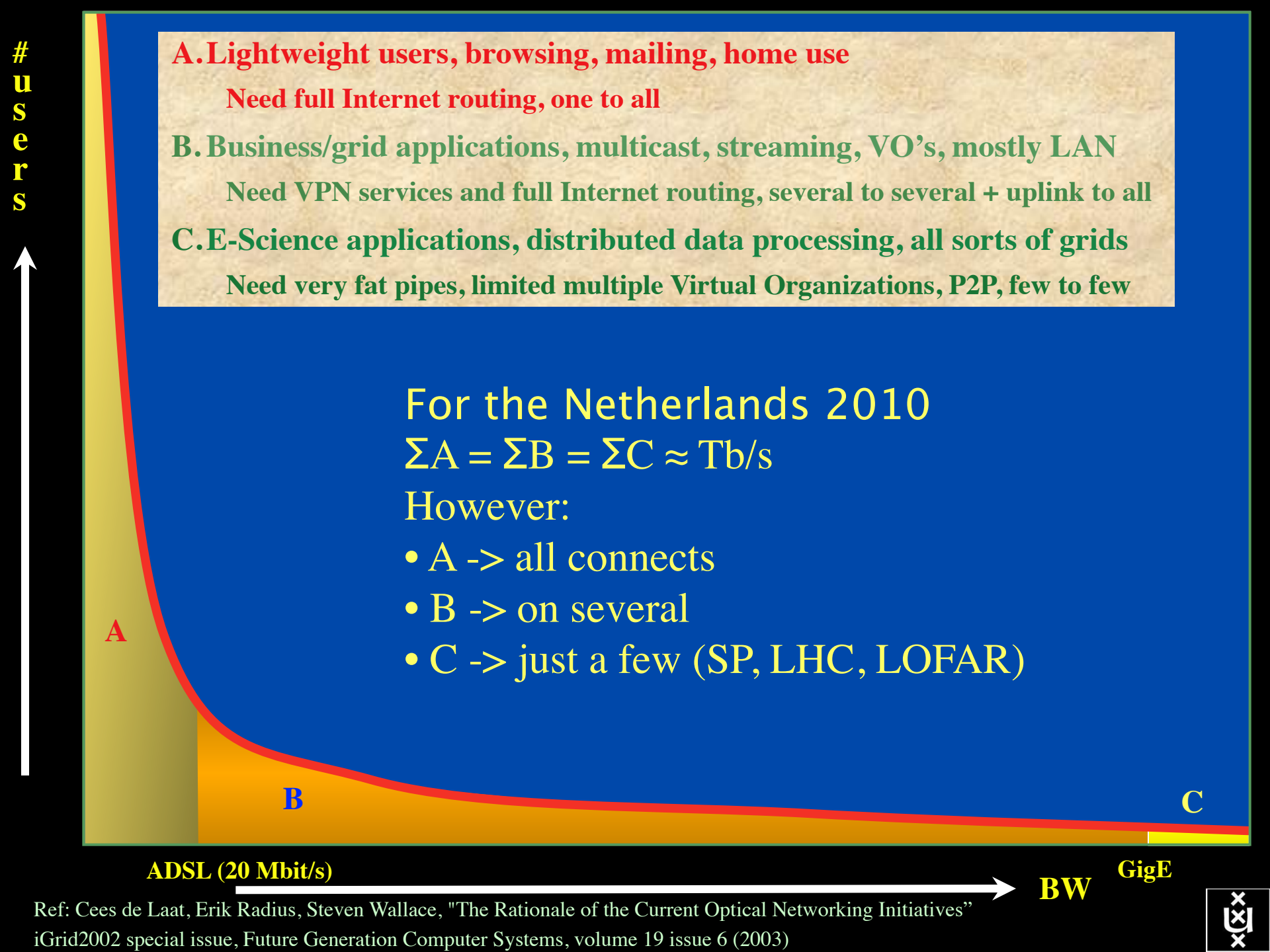

4 X 100G
Trans-
Atlantic
+ NY-CHI
+ AMS-GVA
Using 100G Links
+ Next-Gen
Optical Muxes

# LOFAR as a Sensor Network

20 flops/byte

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

2 Tflops/s




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!

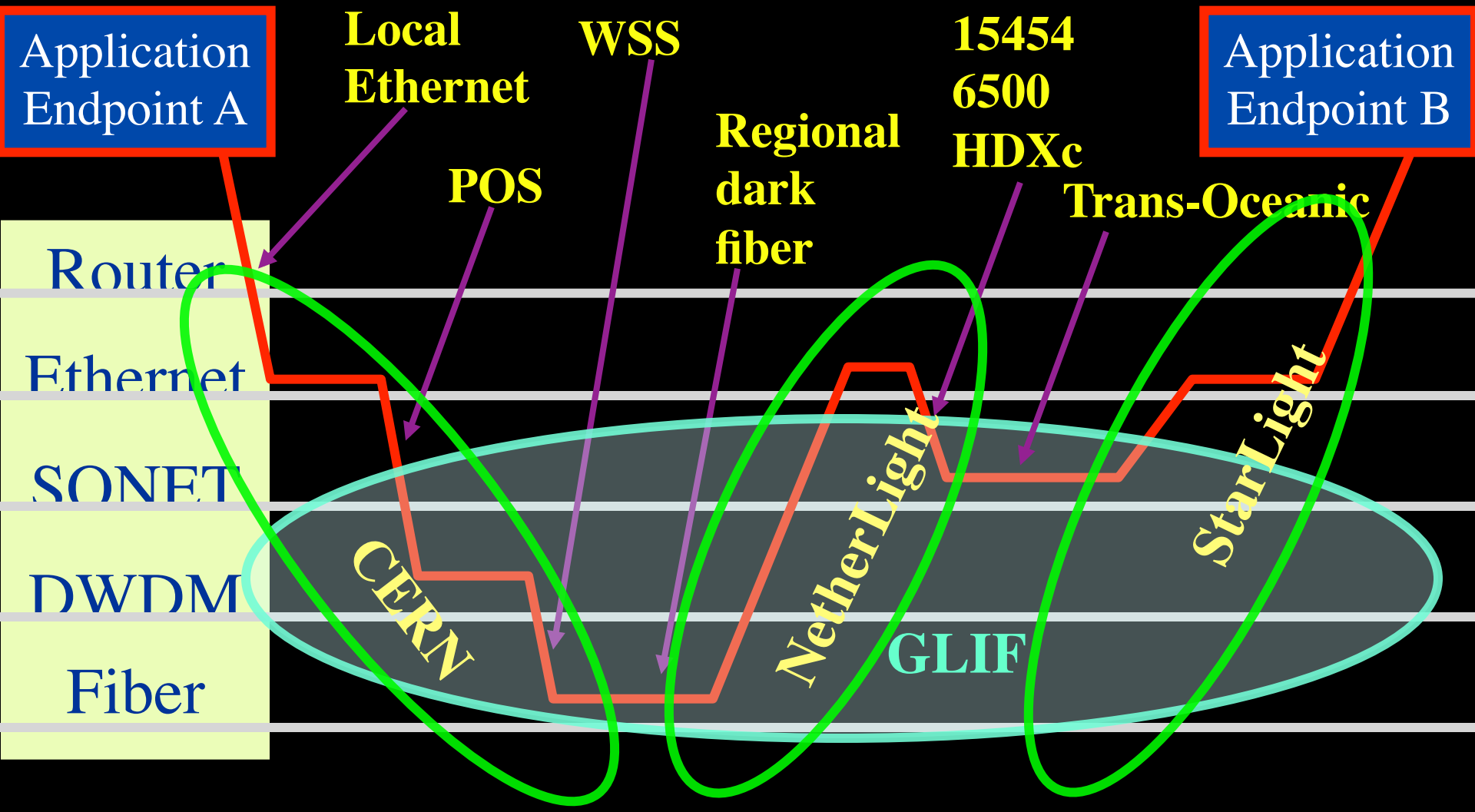
- Costs of photonic equipment 10% of switching 10 % of full routing
  - for same throughput!
  - Photonic vs Optical (optical used for SONET, etc, 10-50 k$/port)
  - DWDM lasers for long reach expensive, 10-50 k$
- Bottom line: look for a hybrid architecture which serves all classes in a cost effective 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
Stockholm
Oslo
SURFnet-
IEEAF-
GLORIAD
50Gb
NORDUnet-
GLORIAD 10Gb
Copenhagen
NorthernLight
NORDUnet 50Gb
Amsterdam
NetherLight Hamburg
UKLight
London
Munster
CESNET 10Gb
CzechLight
Prague
SURFnet
50Gb
Paris
DAS3-
Grid'500
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

# 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

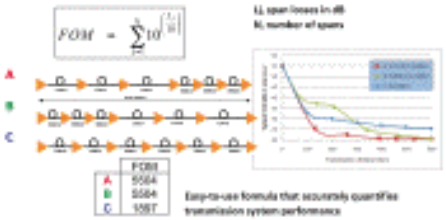
## Alien wavelength challenges

- 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.**

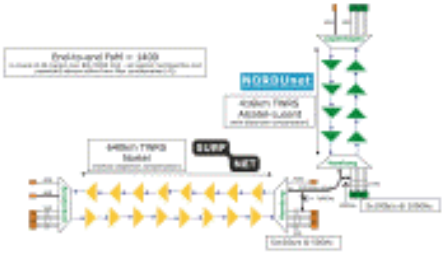
## New method to present fiber link quality, FoM (Figure of Merit)

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

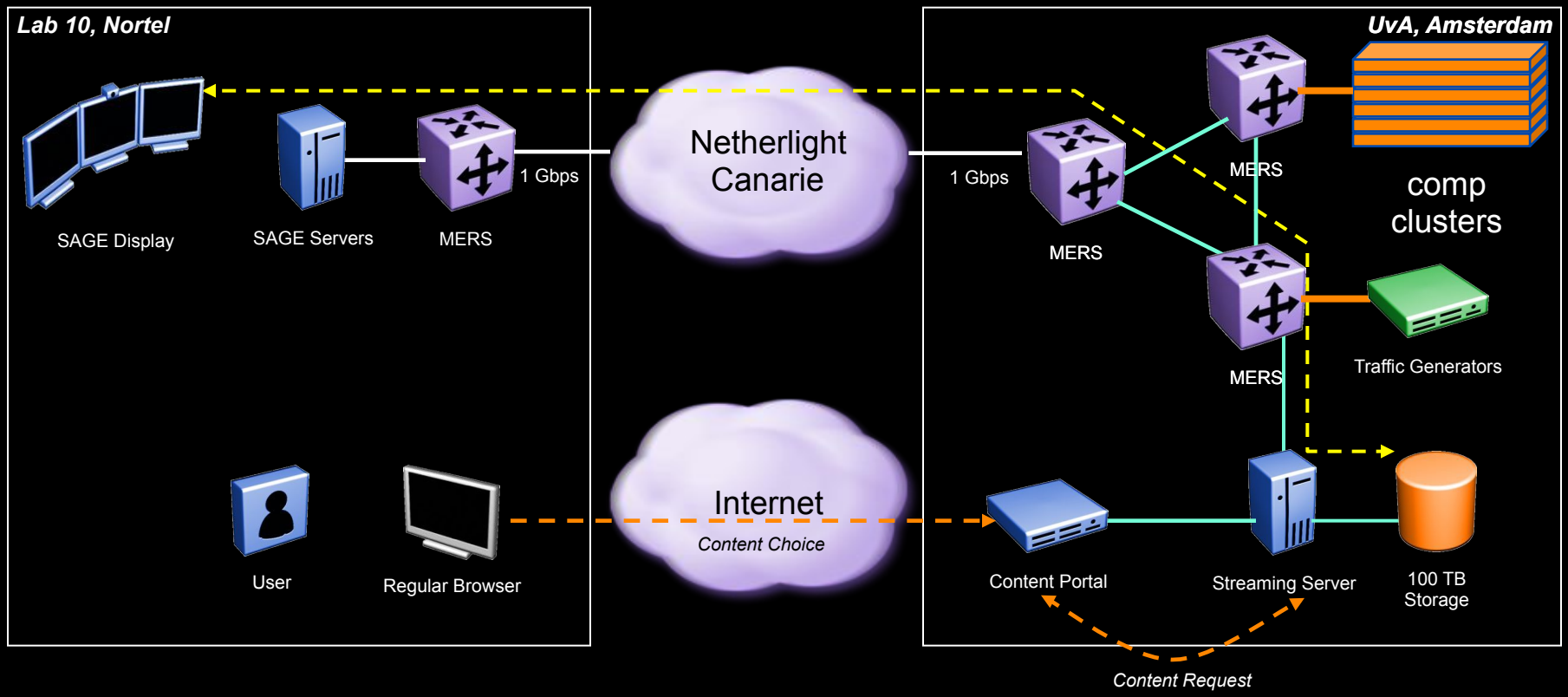
- We have investigated experimentally the all-optical transmission of a 40Gb/s PM-QPSK alien wavelength via a concatenated native and third party DWDM system that both were carrying live 10Gb/s wavelengths.
- The end-to-end transmission system consisted of 1056 km of TWRS (TrueWave Reduced Slope) transmission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

REFERENCES

ACKNOWLEDGEMENTS

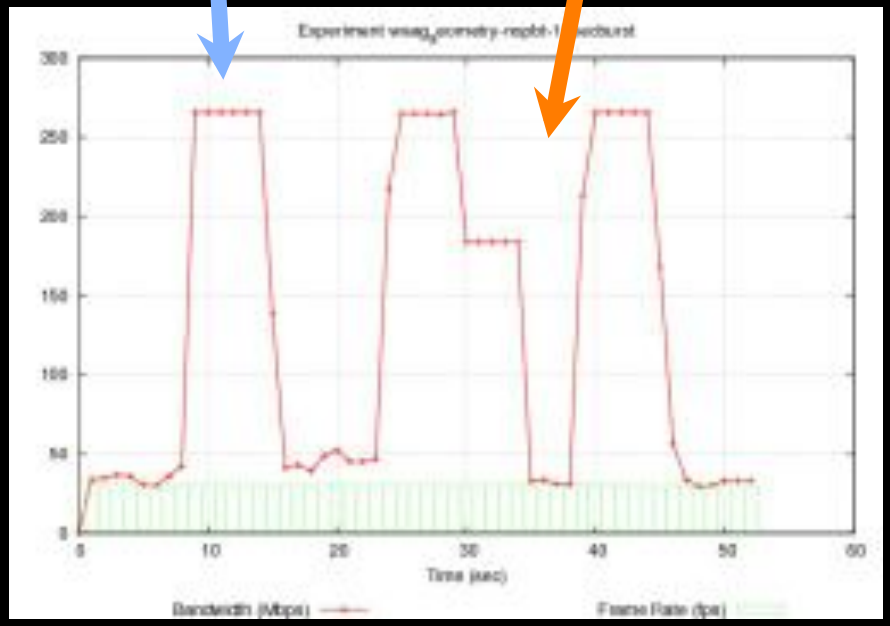
# Diagram for SAGE video streaming to ATS



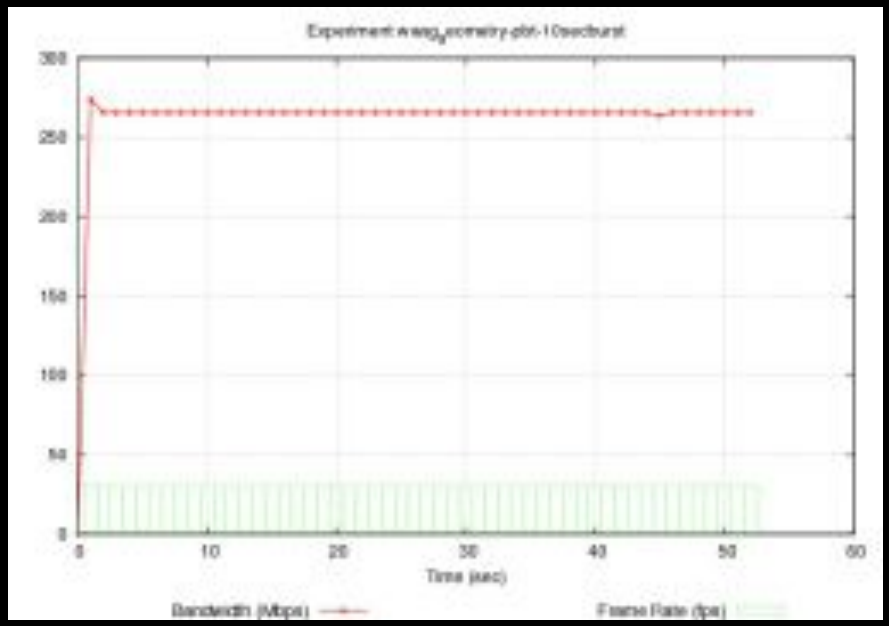
# Experimental Data

Sage without background traffic

Sage with background traffic

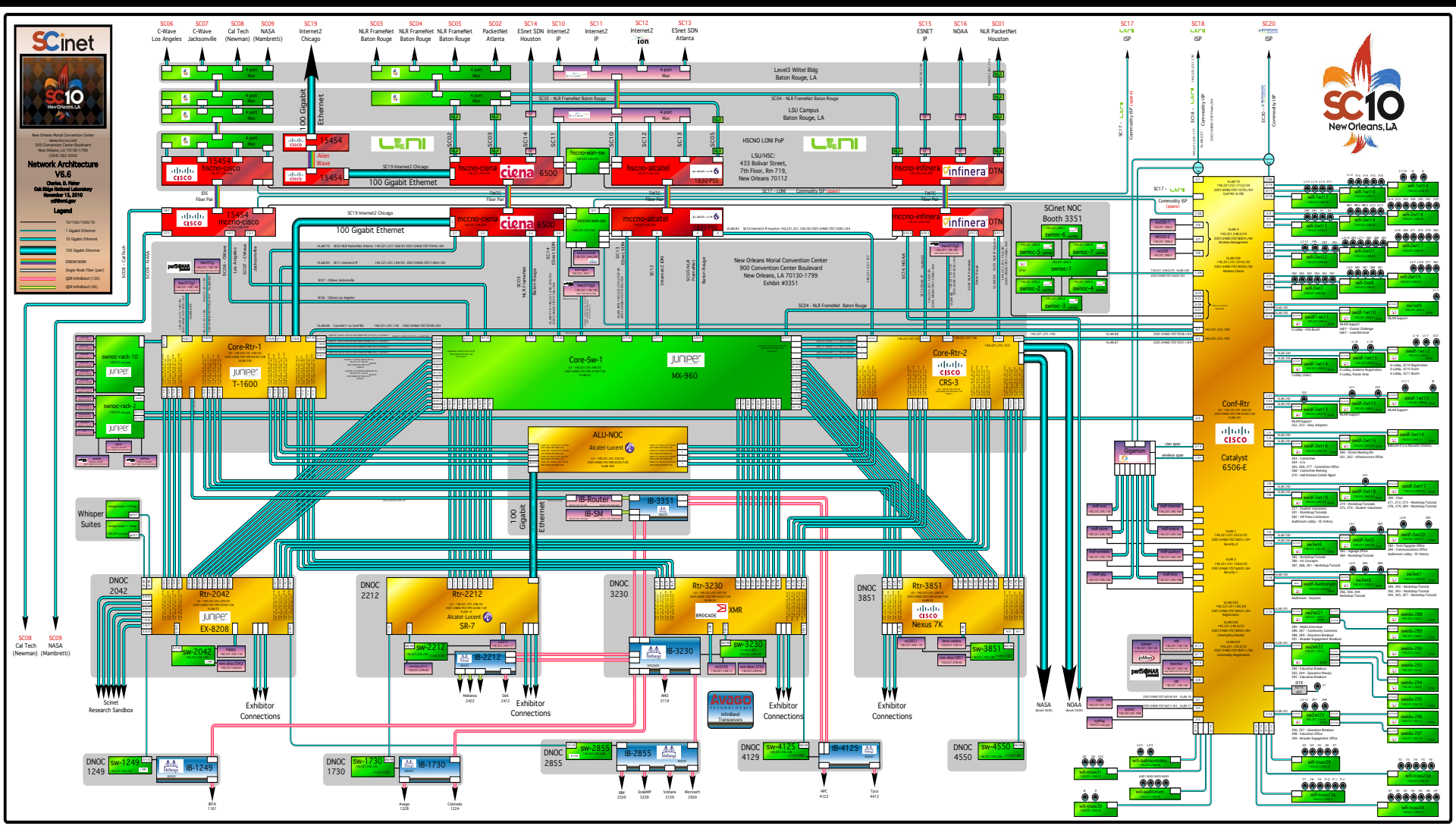



10 Second Traffic bursts with No PBT




10 Second Traffic bursts with PBT

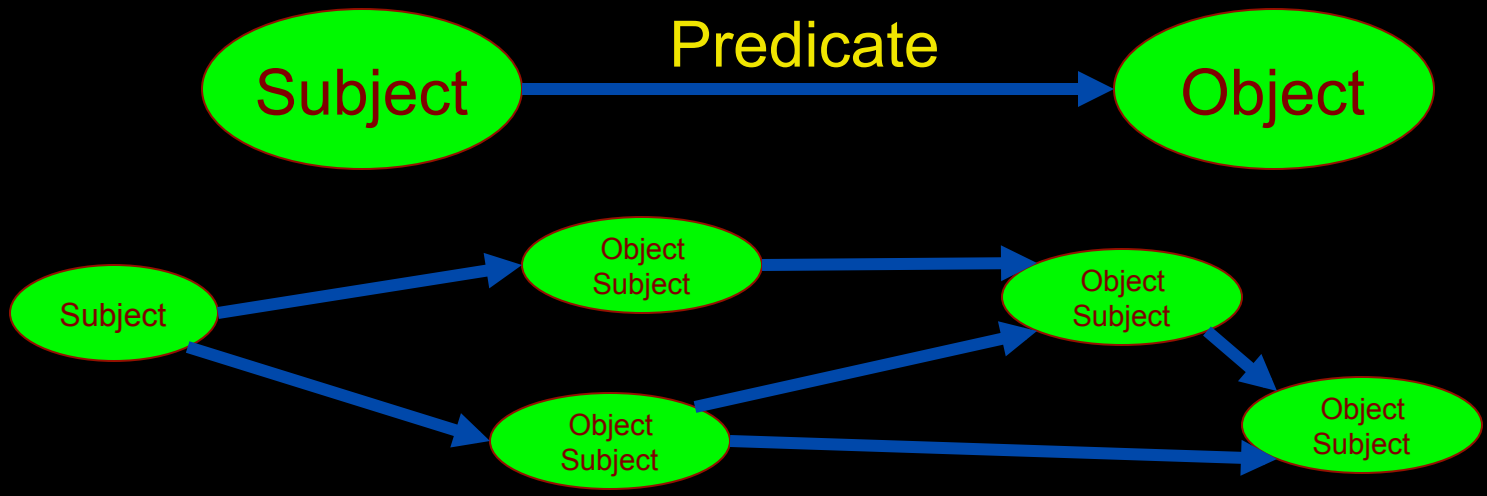
PBT is *SIMPLE* and *EFFECTIVE* technology to build a shared Media-Ready Network

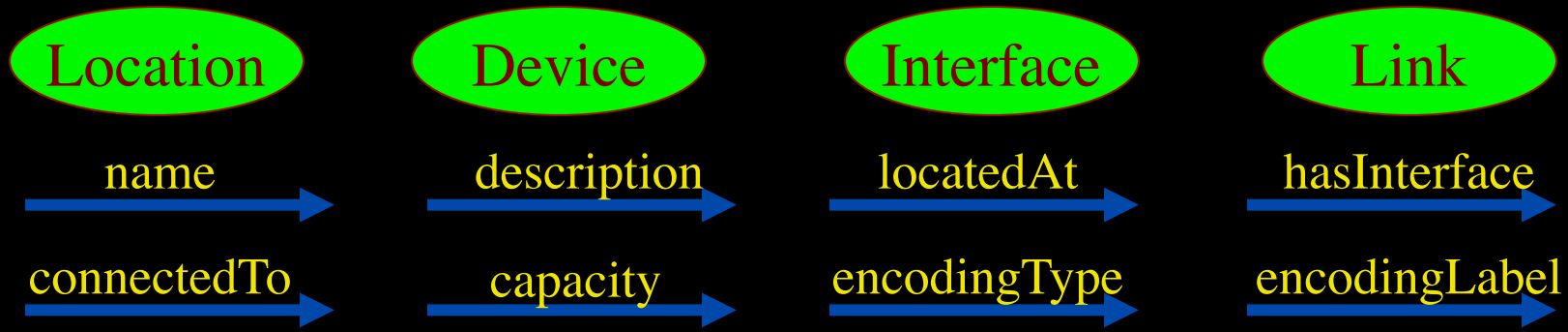

SCinet
SC10
New Orleans, LA
Network Architecture
V6.6
Legend
Core-Rtr-1
Core-Sw-1
Core-Rtr-2
Conf-Rtr
ALU-NOC
Rtr-2042
Rtr-2212
Rtr-3230
Rtr-3851
New Orleans Morial Convention Center
900 Convention Center Boulevard
New Orleans, LA 70130-1799
Exhibit #3351
SCinet NOC
Booth 3351
Level3 Wiltel Bldg
Baton Rouge, LA
LSU Campus
Baton Rouge, LA
HSCNO LONI PoP
LSU/HSC:
433 Bolivar Street,
7th Floor, Rm 719,
New Orleans 70112
Scinet
Research Sandbox
Exhibitor
Connections
NASA
NOAA

# Complexity

# Network Description Language

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



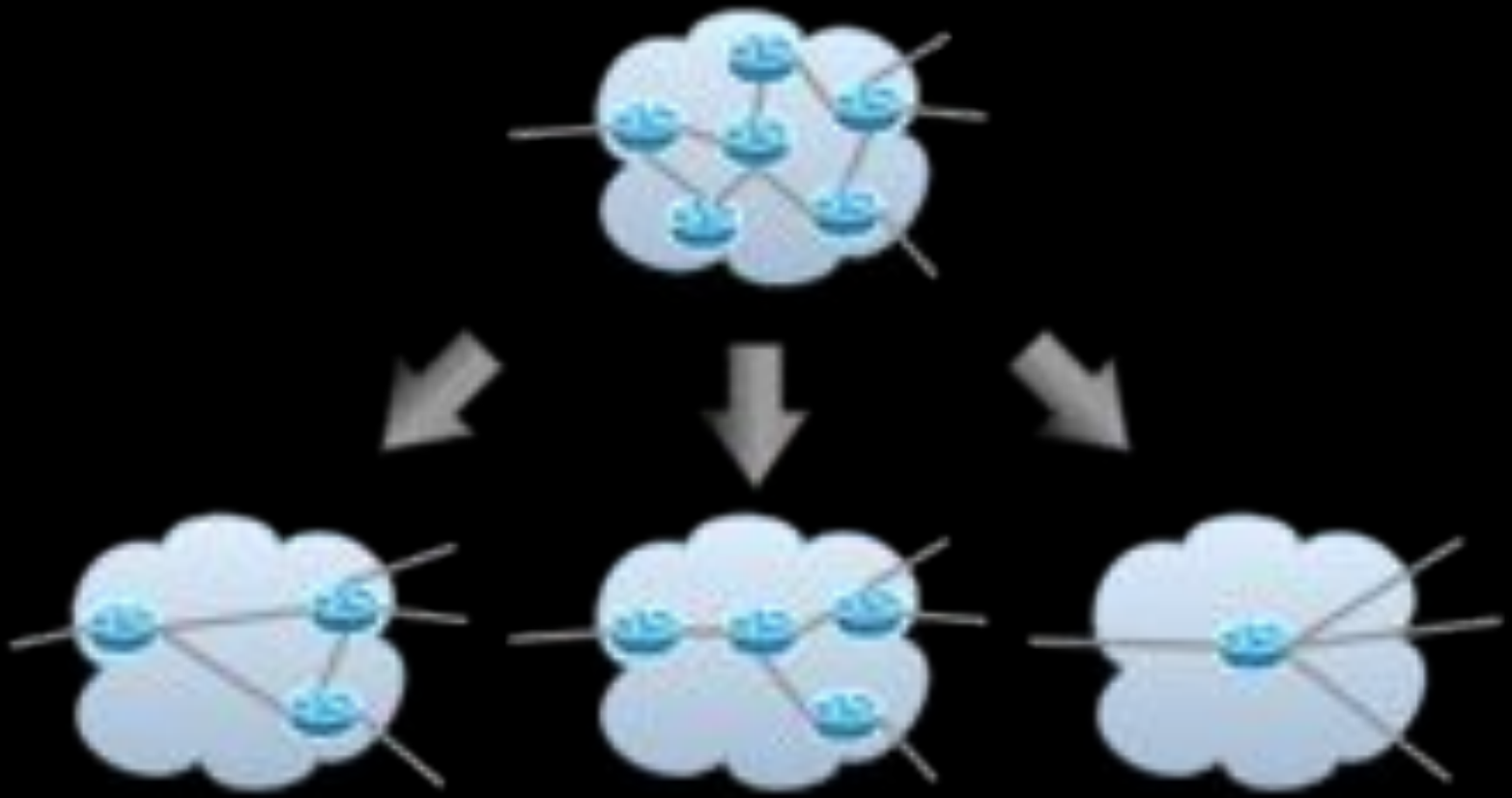
# NetherLight in RDF

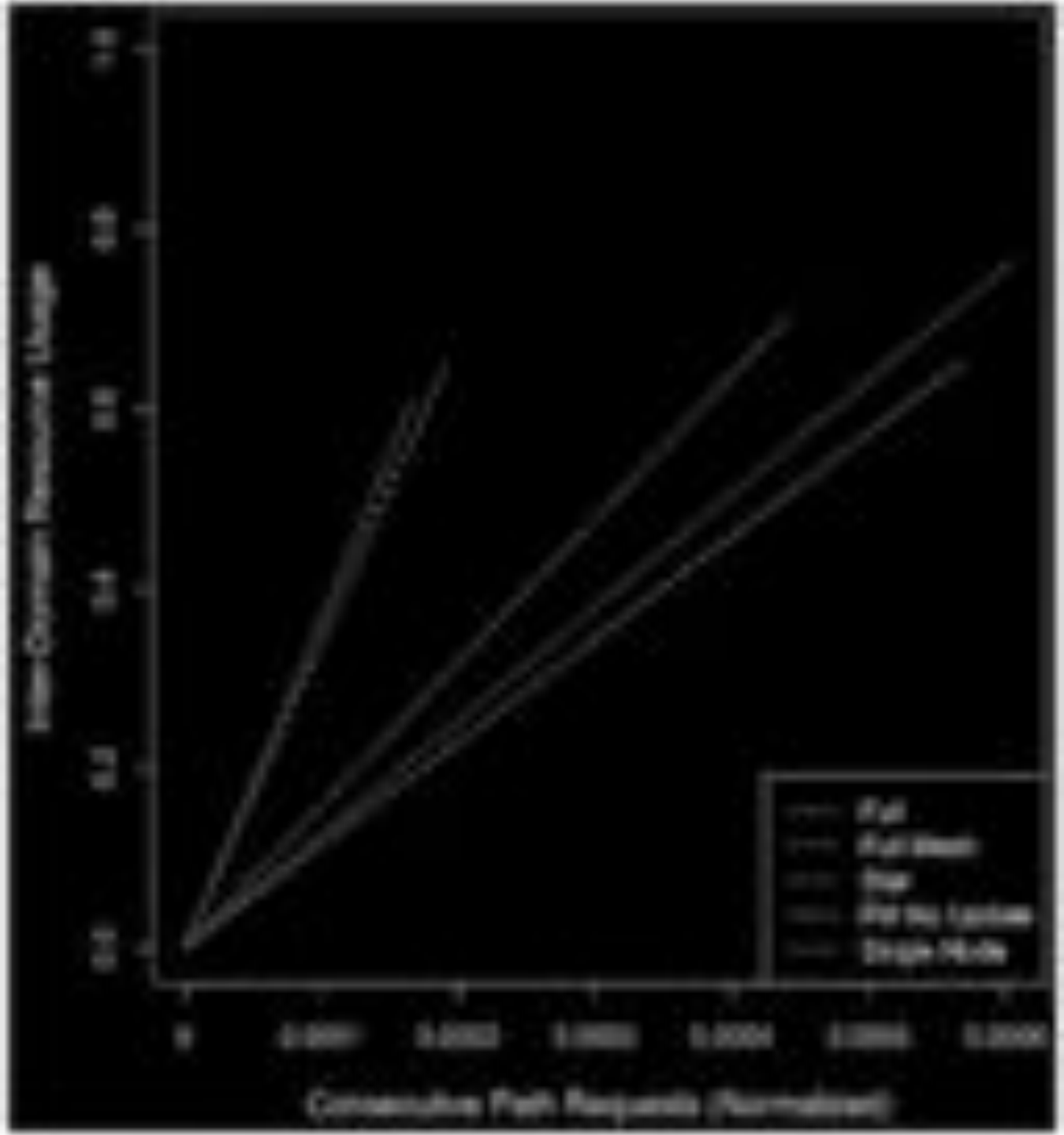
```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:ndl="http://www.science.uva.nl/research/air/ndl#">
<!-- Description of Netherlight -->
<ndl:Location rdf:about="#Netherlight">
      <ndl:name>Netherlight Optical Exchange</ndl:name>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

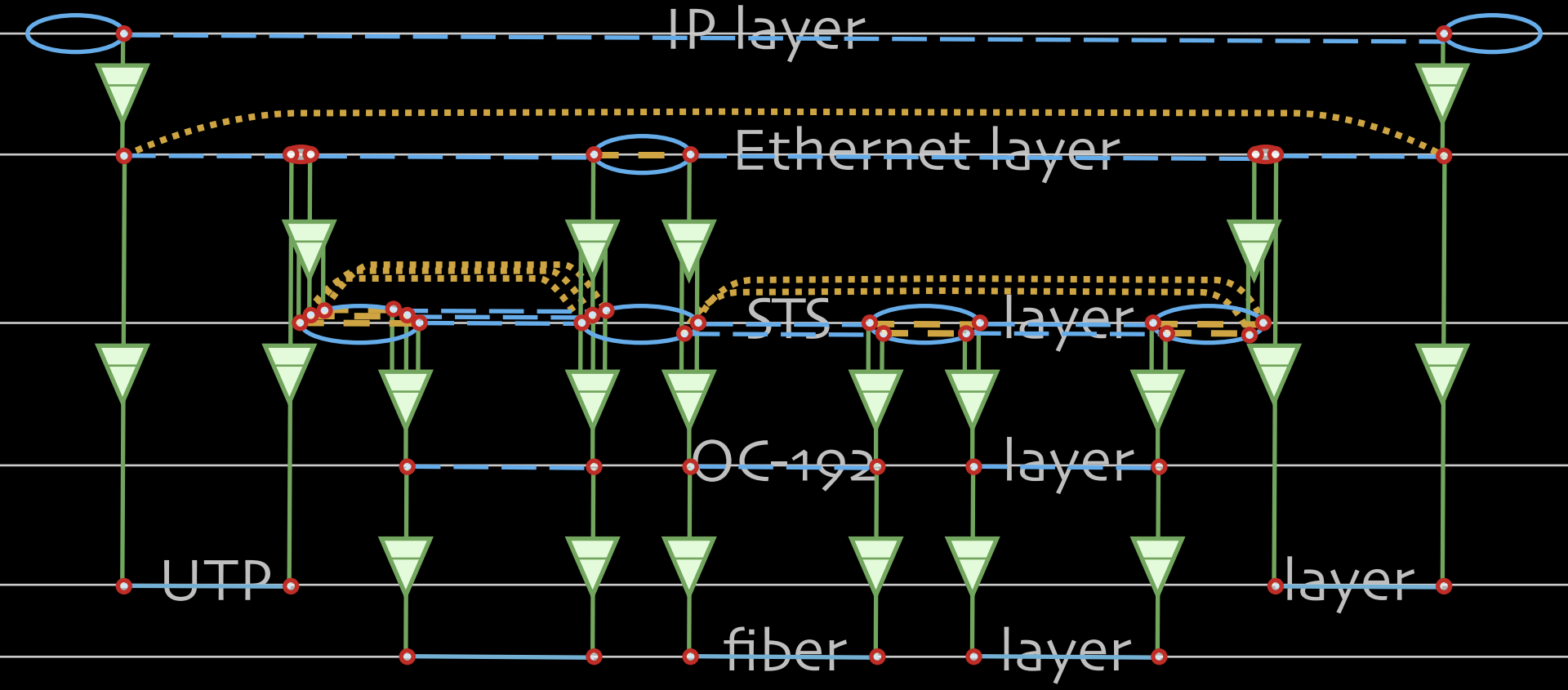
# Topology Aggregation

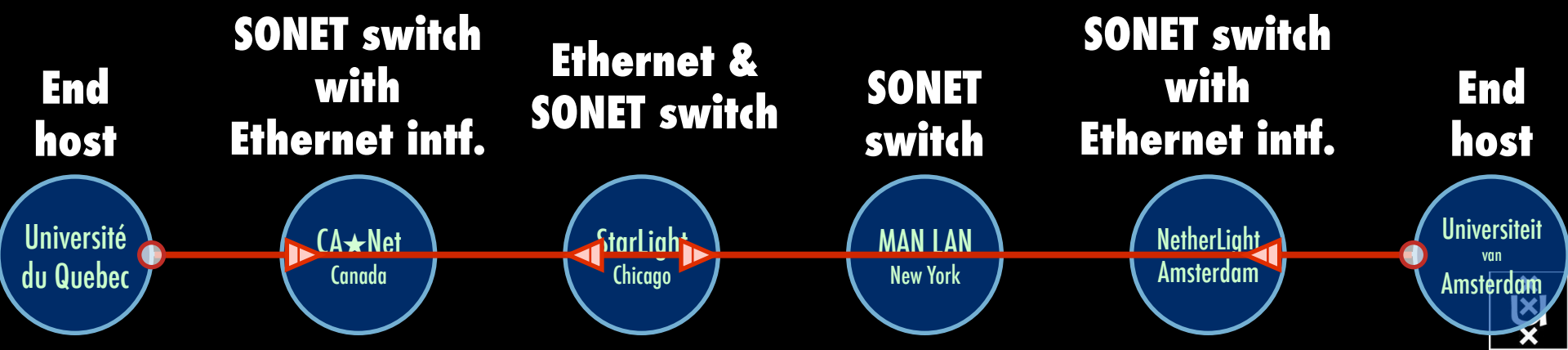




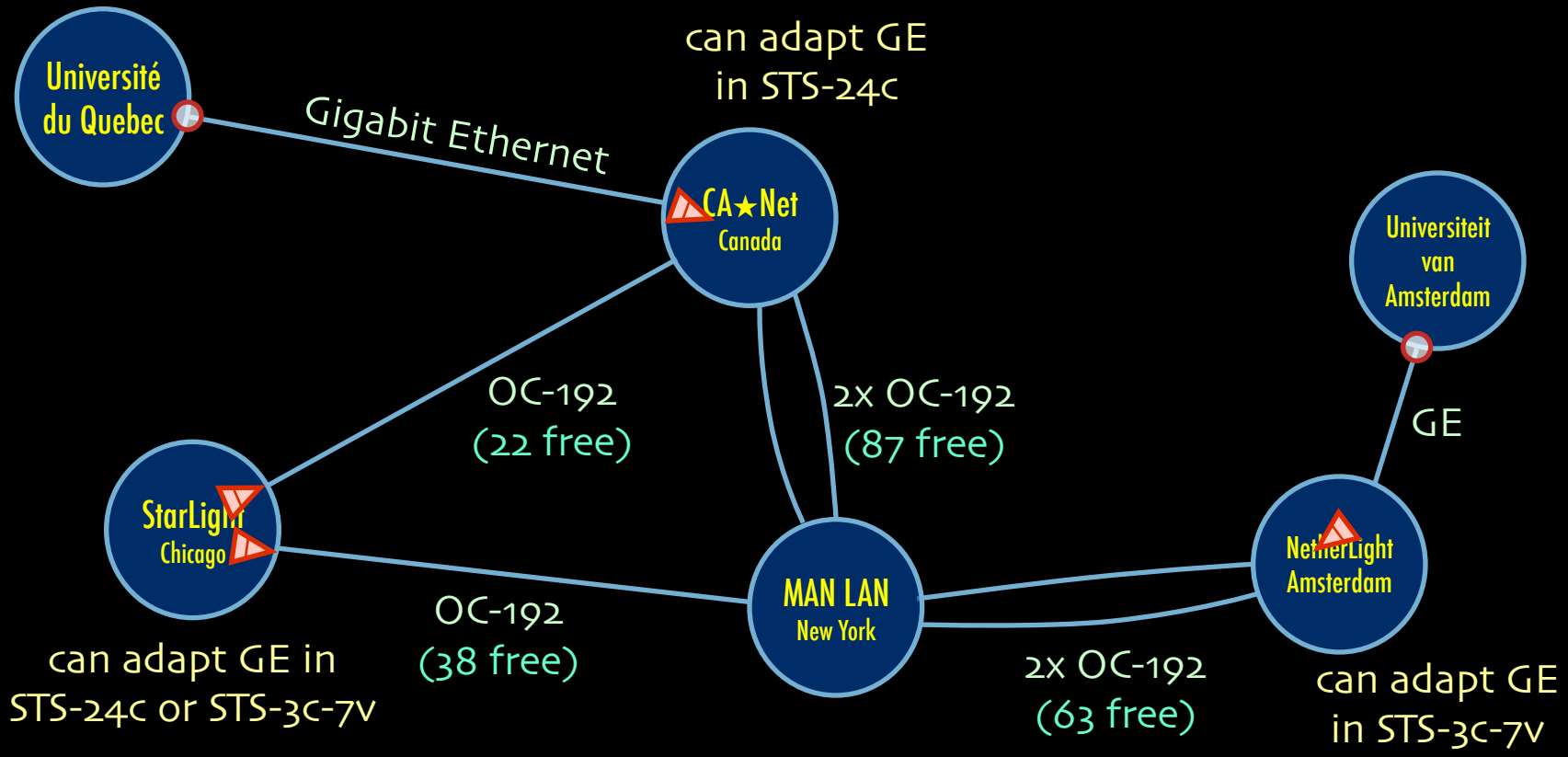
# Topology Aggregation - Initial

# Multi-layer descriptions in NDL



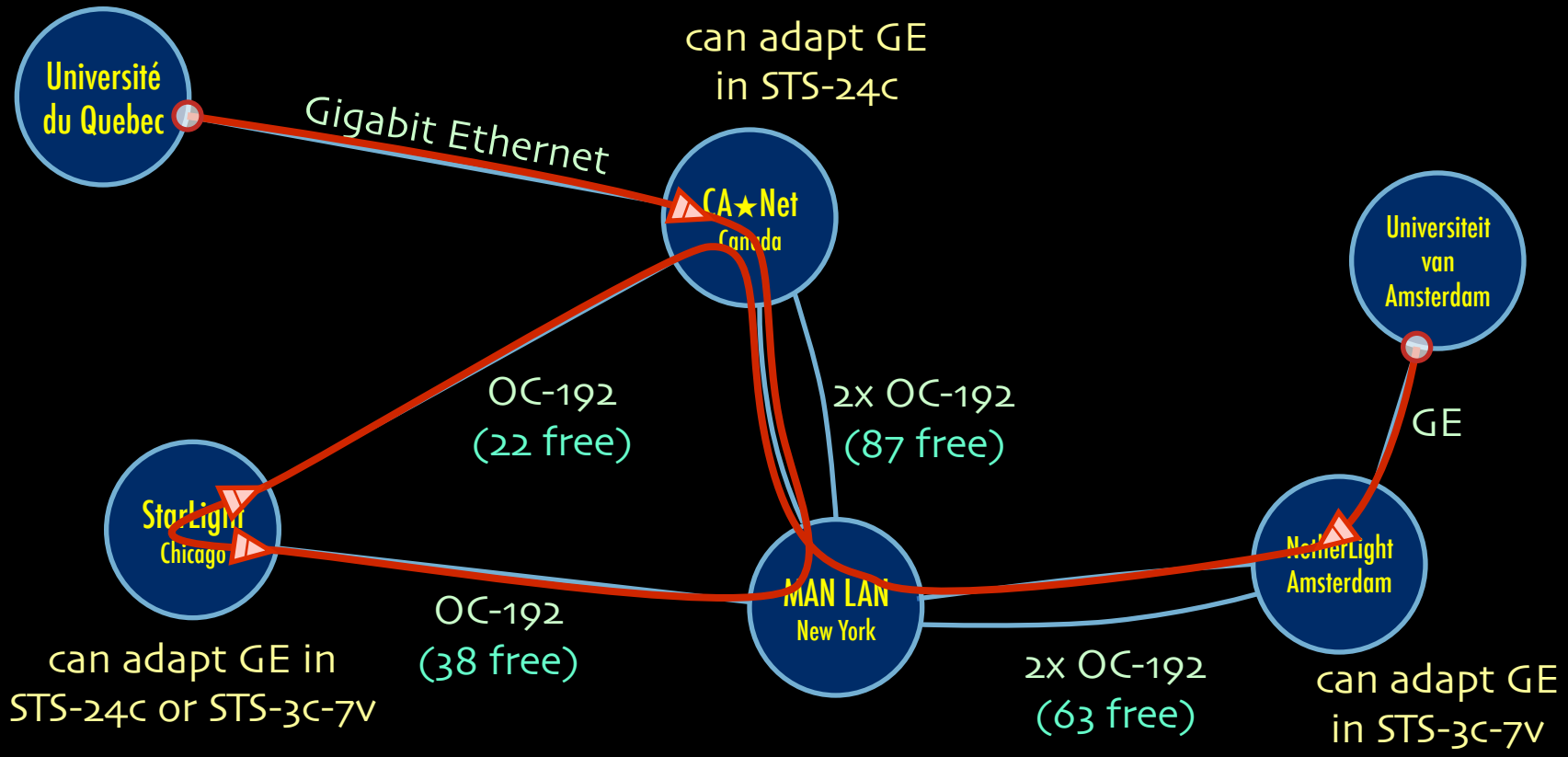



# A weird example




Thanks to Freek Dijkstra & team

# A weird example



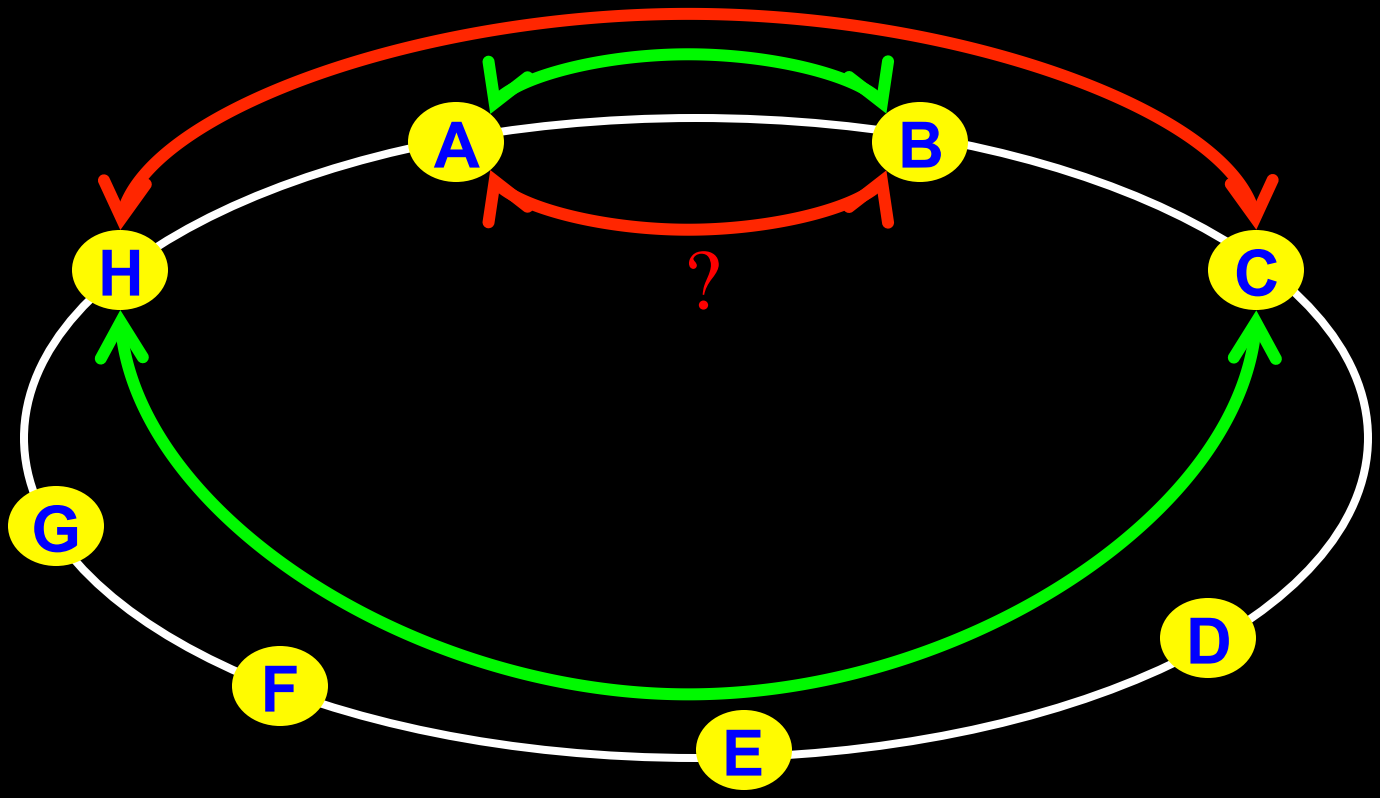

Thanks to Freek Dijkstra & team

# The Problem

I want HC and AB

Success depends on the order

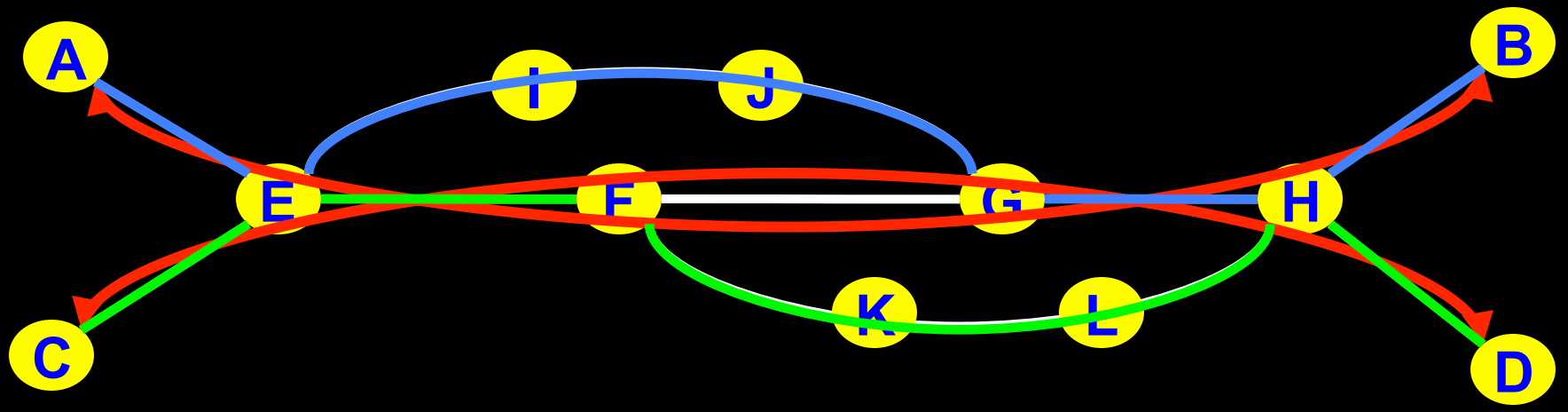
Wouldn't it be nice if I could request [HC, AB, ...]



# Another one ☺

I want AB and CD

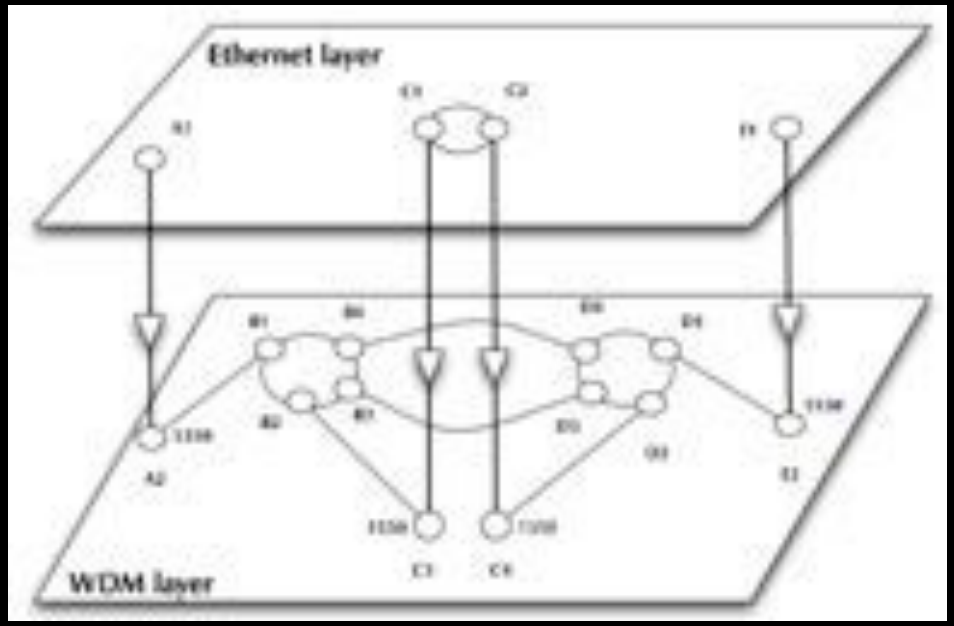
Success does not even depend on the order!!!

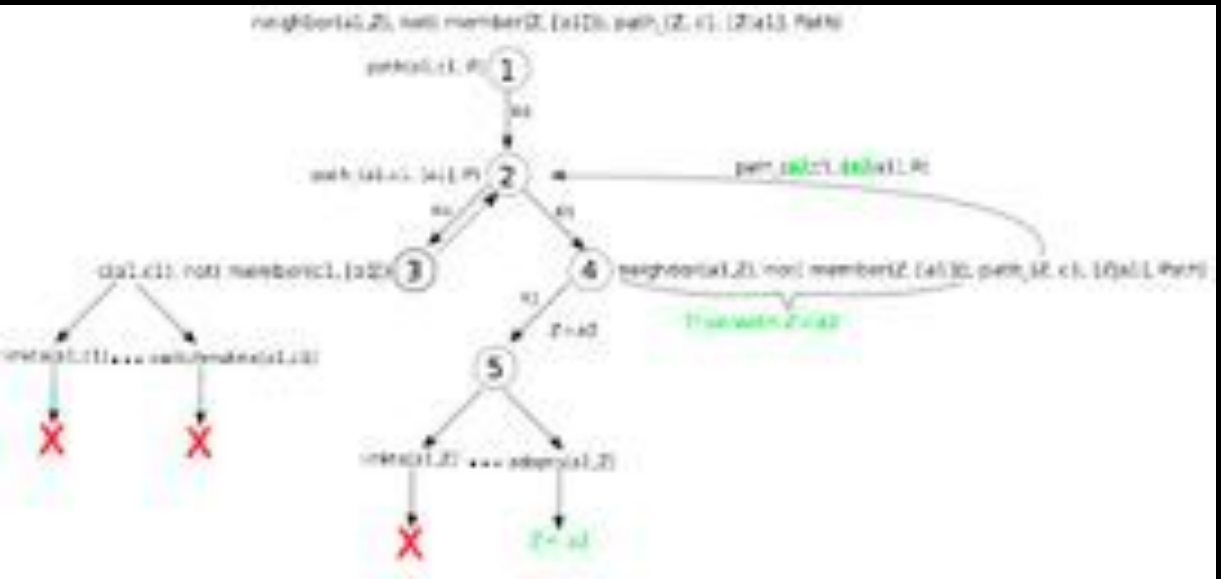


# NDL + PROLOG

Research Questions:

- order of requests
- complex requests
- usable leftovers



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

# 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.

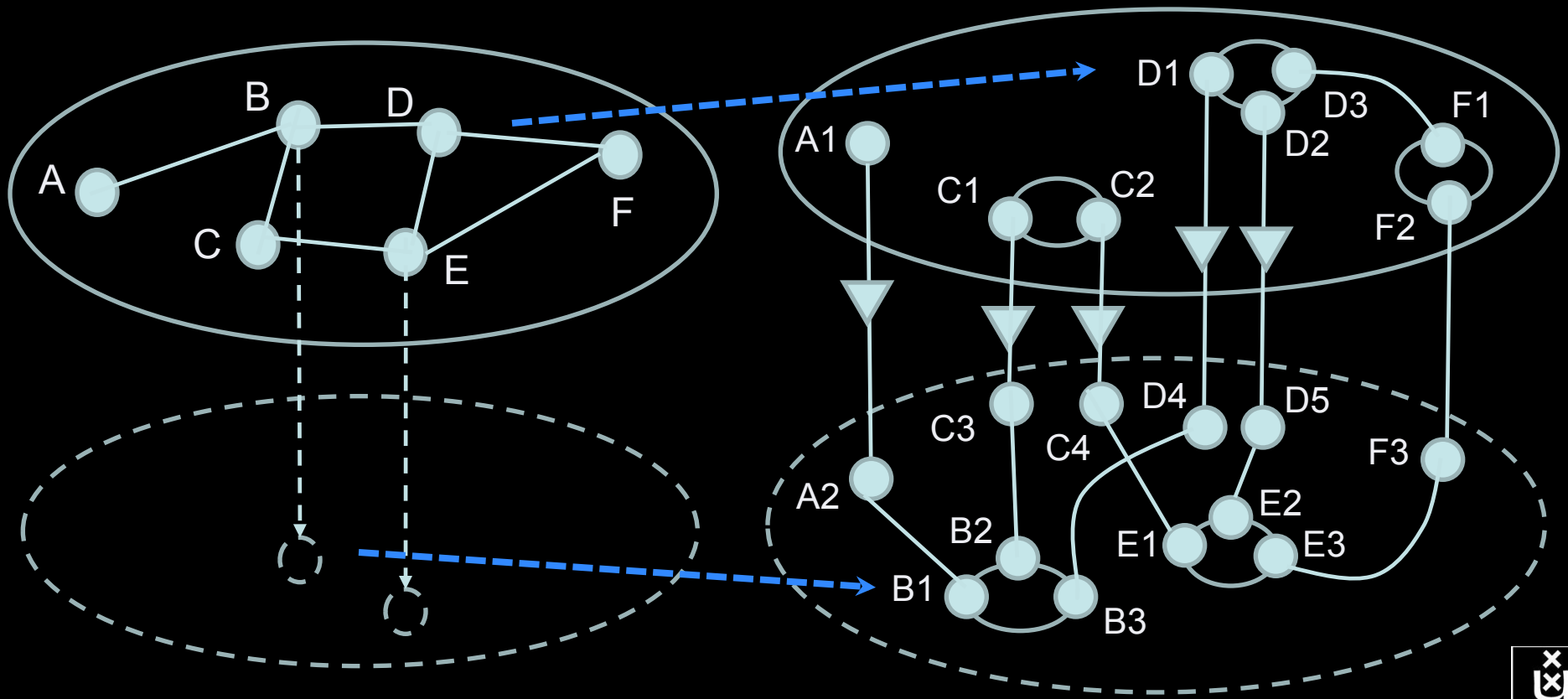
**Input parameters:**

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

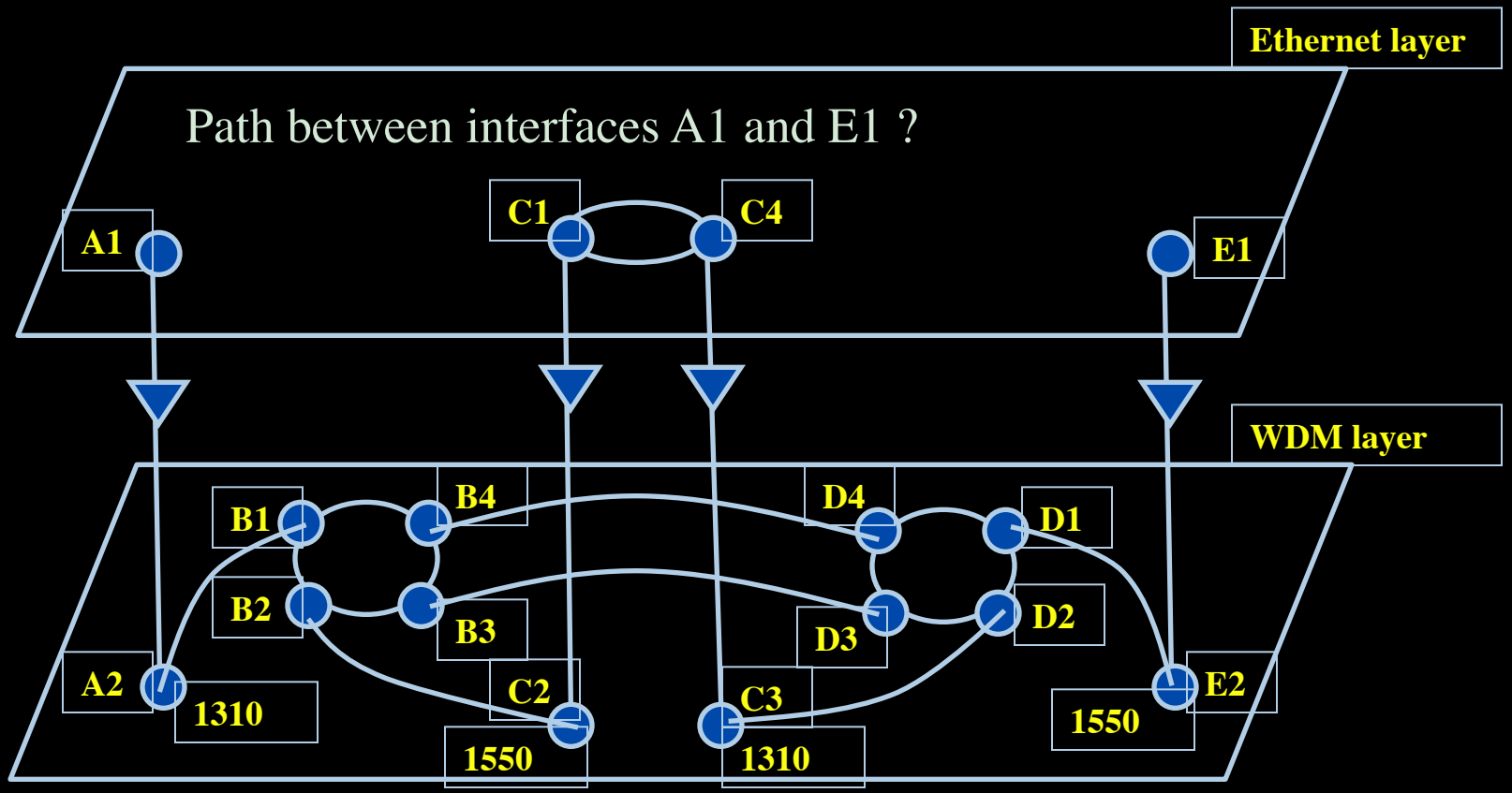
# Multi-domain 2-layer networks

**Projection algorithm**

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



# Multi-layer Network PathFinding



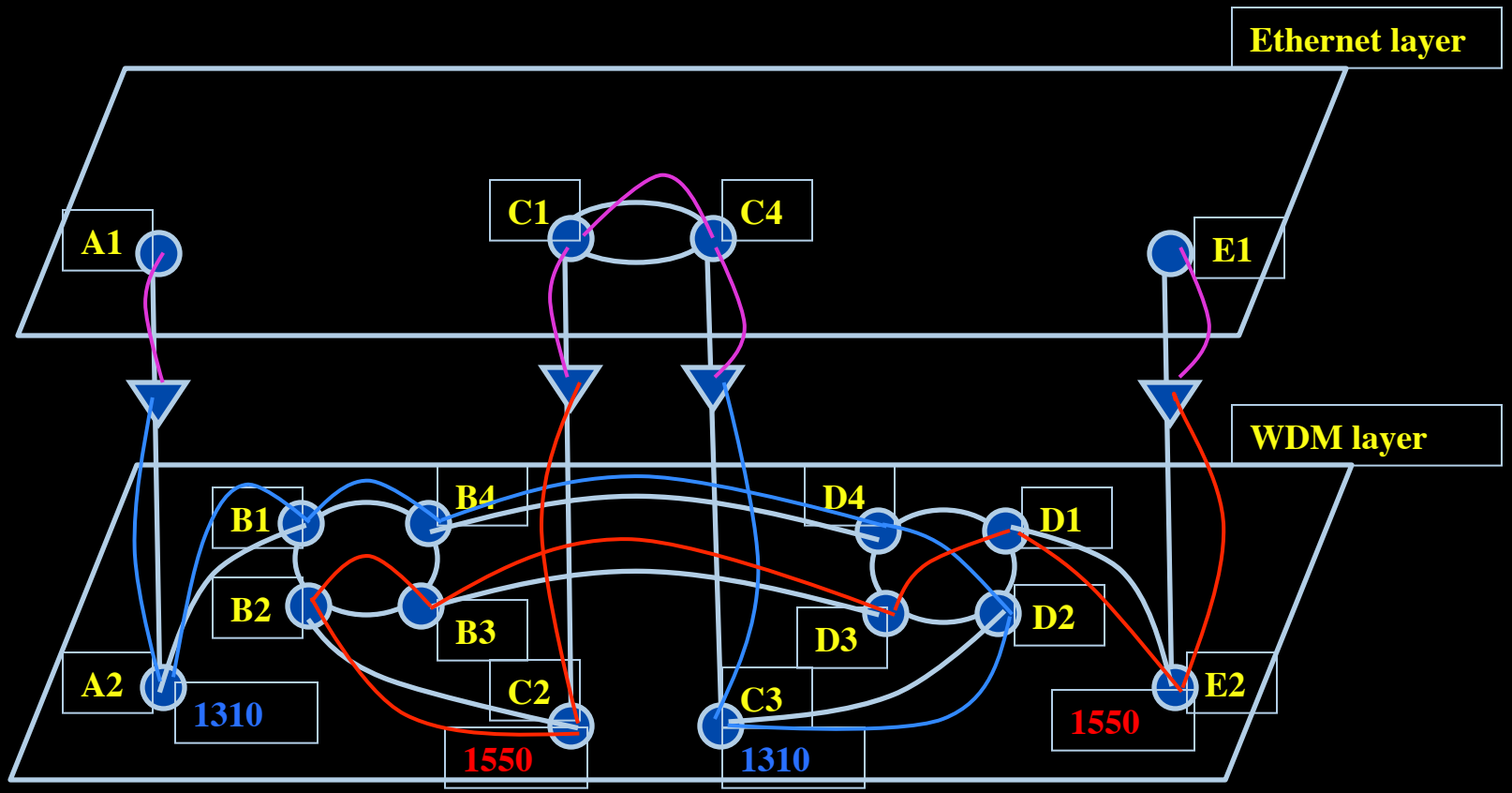

**Prolog rule:**

```
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 | |
| 4 (48:32)(377)(73) | 2620(8245) | 74 | 92.6 | No max #wav |
| 4 (96:64)(771)(147) | 6592(11802) | 207 | 79.3 | |
| | | | | |
| 3 (9:6)(55)(11) | 20(4) | 0 | 100 | |
| 4 (48:32)(377)(73) | 1303(5052) | 22 | 97.8 | #wav ≤3 |
| 4 (96:64)(771)(147) | 3910(10045) | 51 | 94.9 | |
| | | | | |
| 3 (9:6)(55)(11) | 20(4) | 0 | 100 | |
| 4 (48:32)(377)(73) | 755(3210) | 8 | 98.9 | #wav ≤2 |
| 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

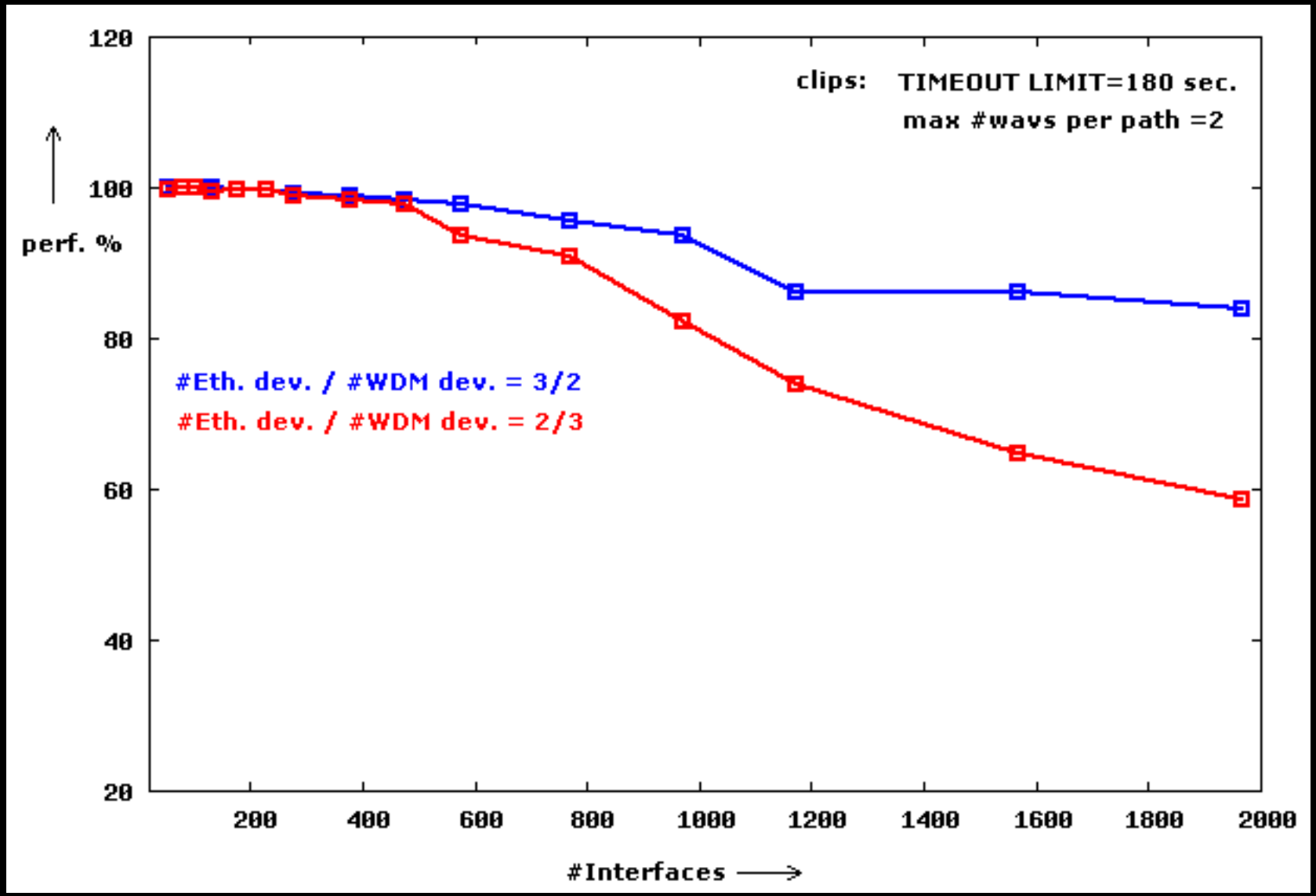
Projection: first of A->B and B->A

| | | | |
|---|---|---|---|
| 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

# Performance Prolog Depth-First Search

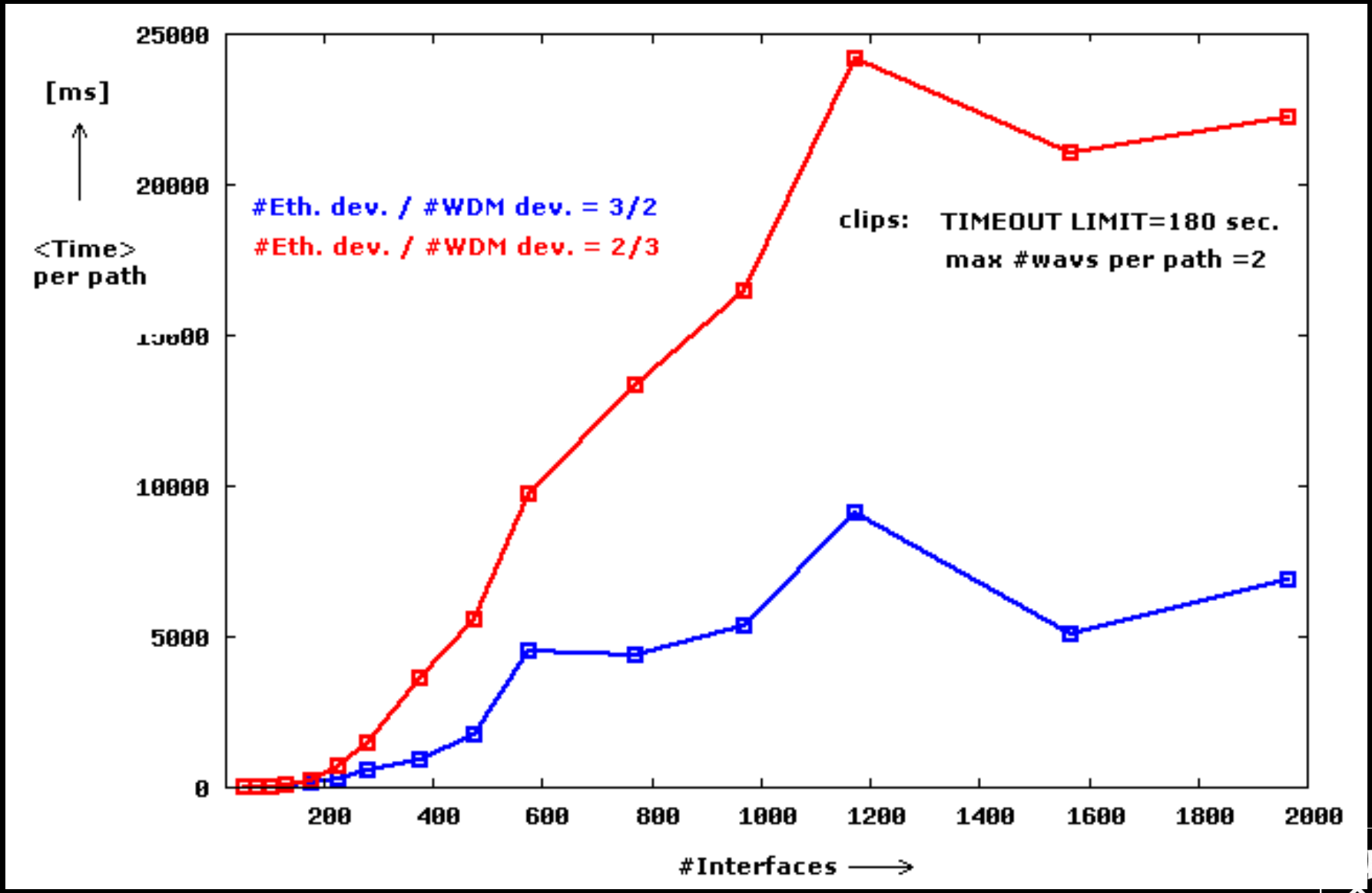



Performance drop mainly due to Timeout limit

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

# DAS3 cluster

## Time Prolog Depth-First Search

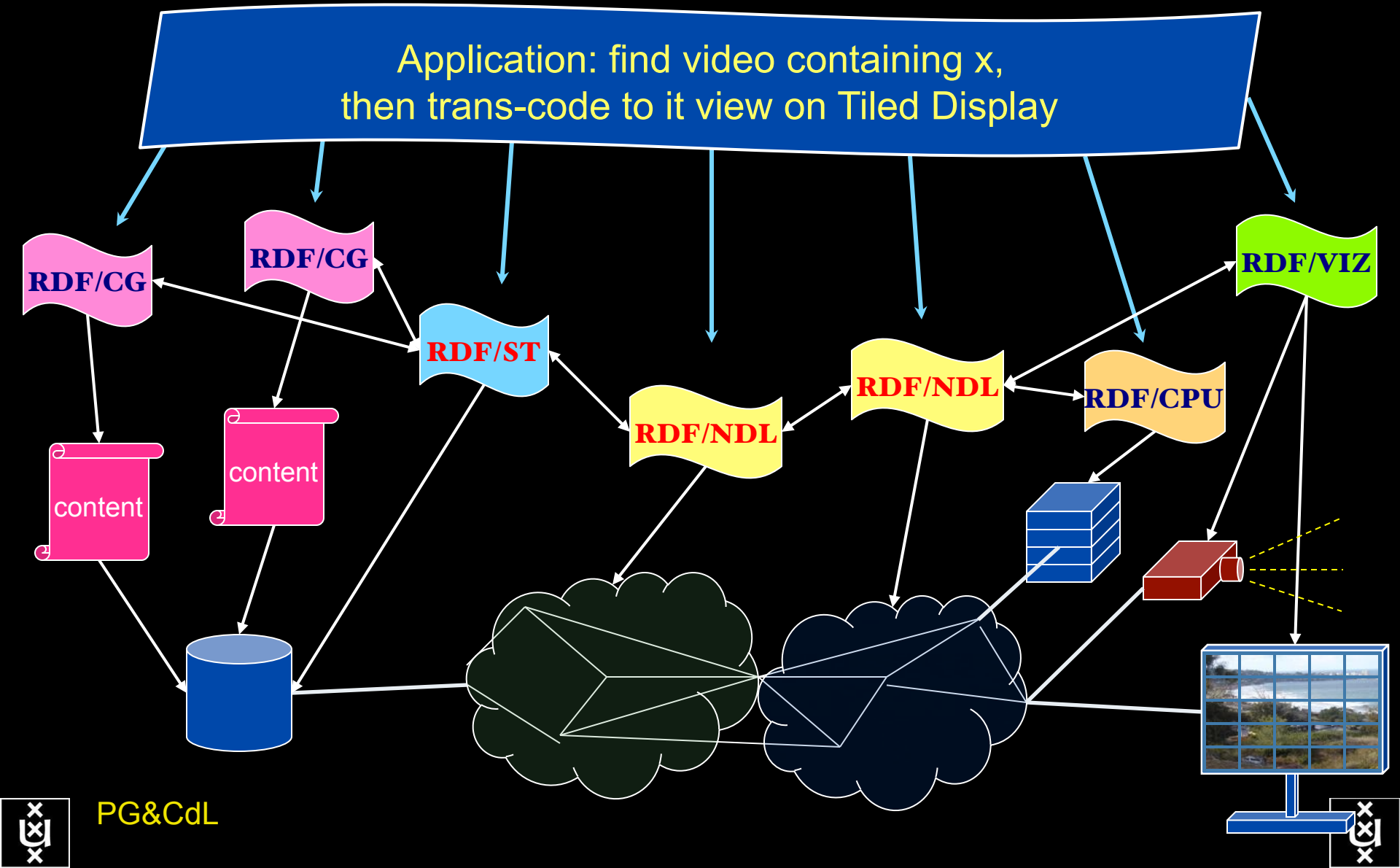


# Standardization

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

- 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.
- Constructed large set of multi-domain 2-layer networks of different sizes with the Barabási-Albert algorithm.
- Studied shortest paths between randomly chosen src-dst pairs by means of an memory unfriendly algorithm.

# RDF describing Infrastructure



Applications and Networks become aware of each other!

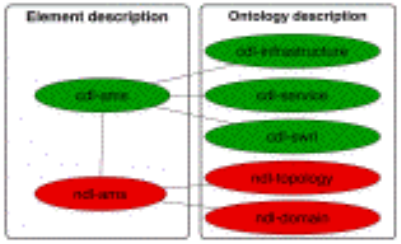
# CineGrid Description Language

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

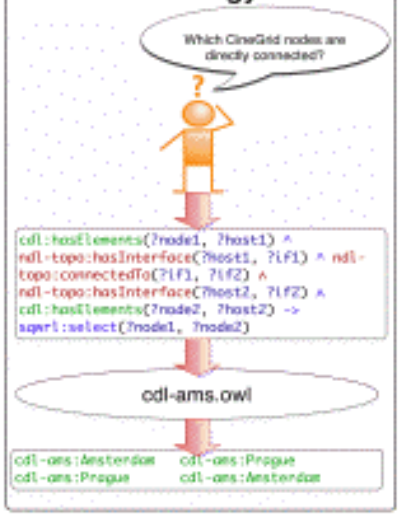
The CineGrid Description Language (CDL) describes CineGrid resources. Streaming, display and storage components are organized in a hierarchical way.

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.

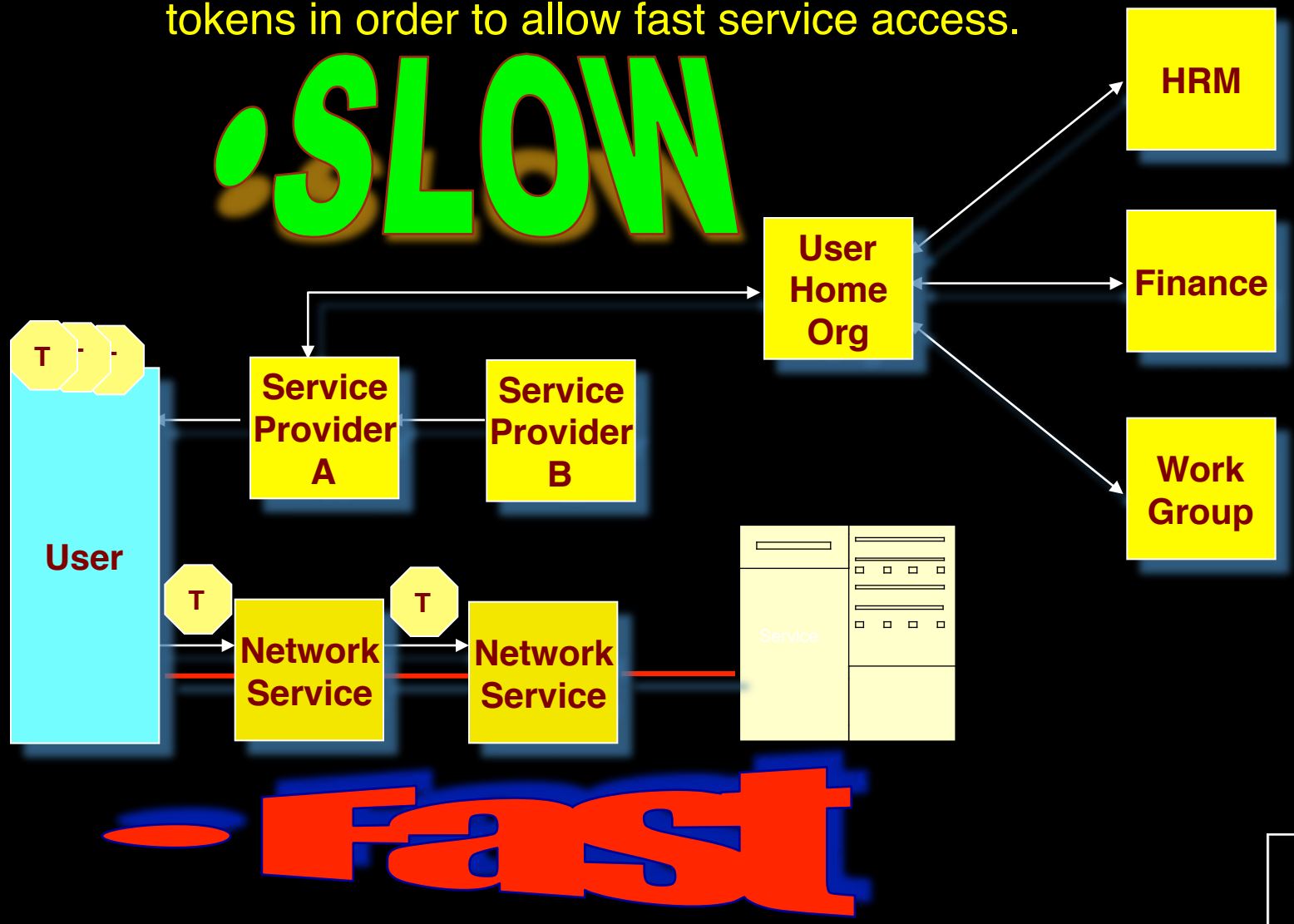



## UML representation of CDL

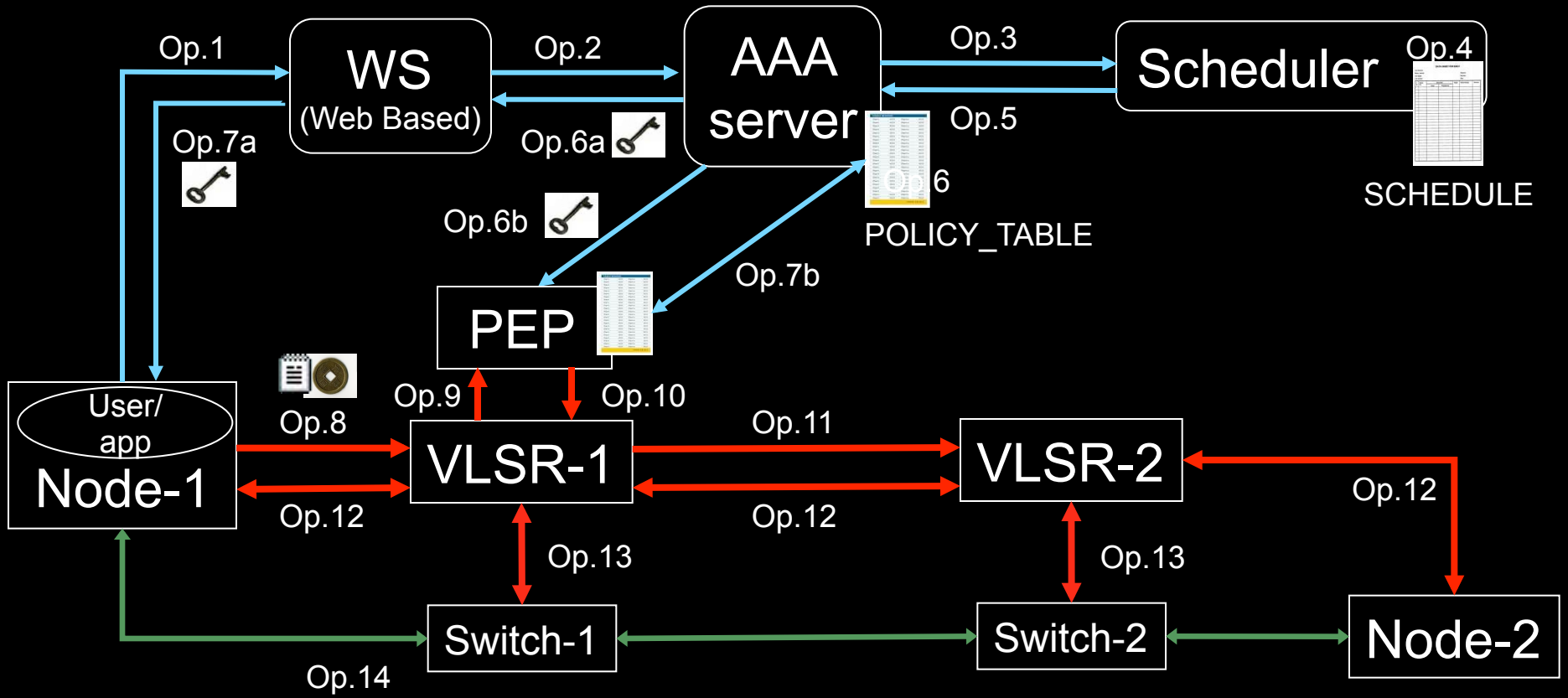
***CDL*** links to ***NDL*** using the ***owl:SameAs*** property. ***CDL*** defines the services, ***NDL*** the network interfaces and links. The combination of the two ontologies identifies the host pairs that support matching services via existing network connections.


GigaPort

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



# 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

# Hybrid computing

Routers ⇐ ⇒ Supercomputers

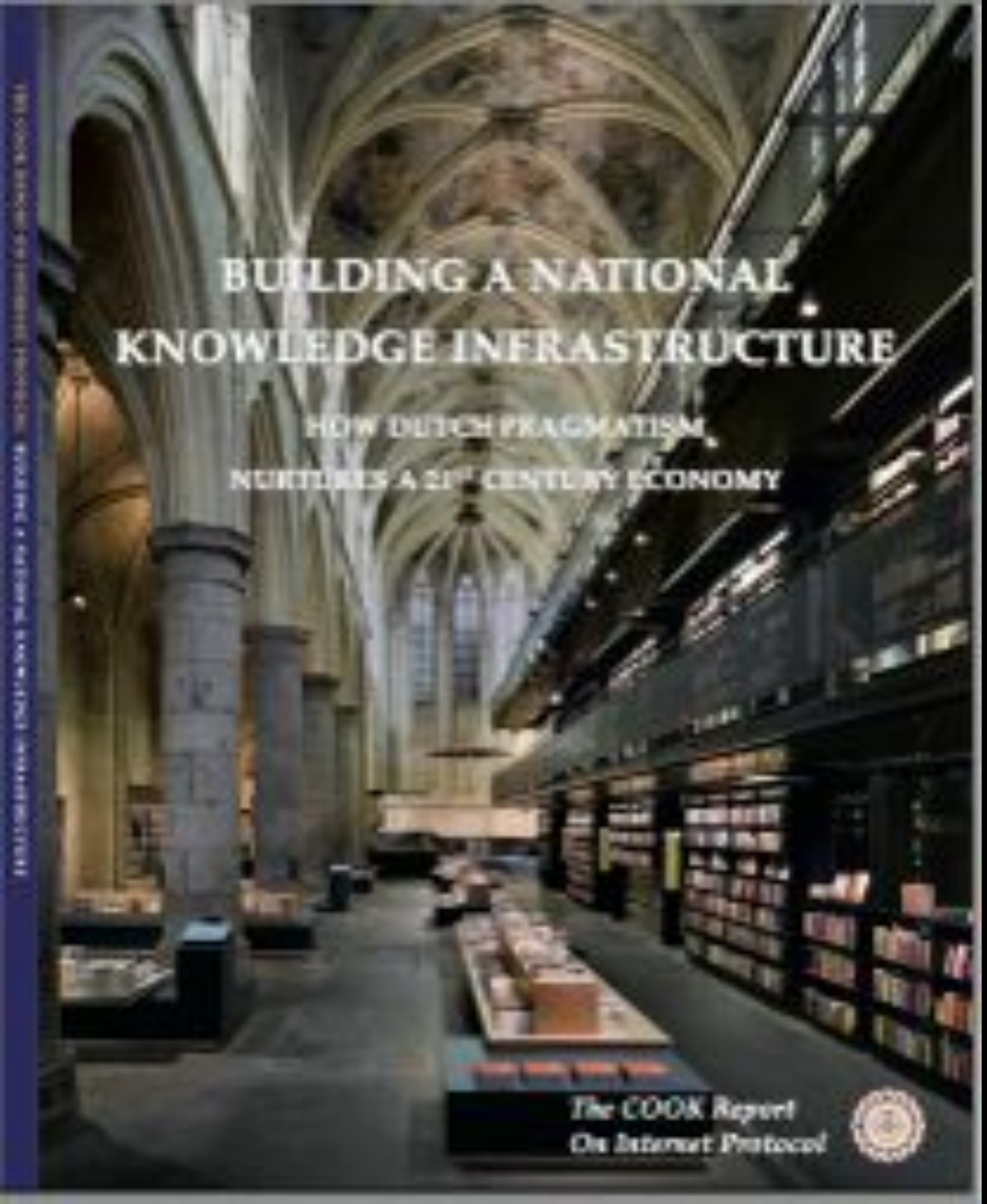
Ethernet switches ⇐ ⇒ Grid & Cloud

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