GNARP 2009:
Optical Networks for e-Science

Cees de Laat

GLIF.is founding member

SURFnet

EU

BSIK

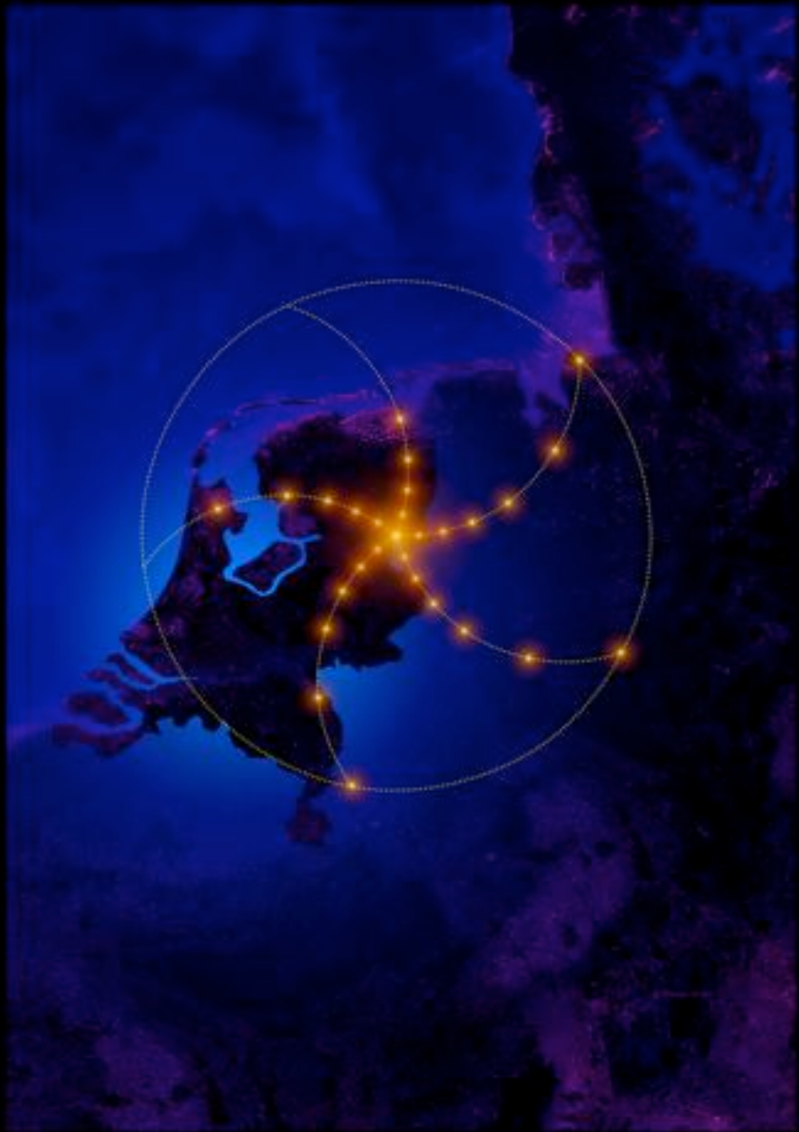
NWO

University of Amsterdam

TNO

NCF

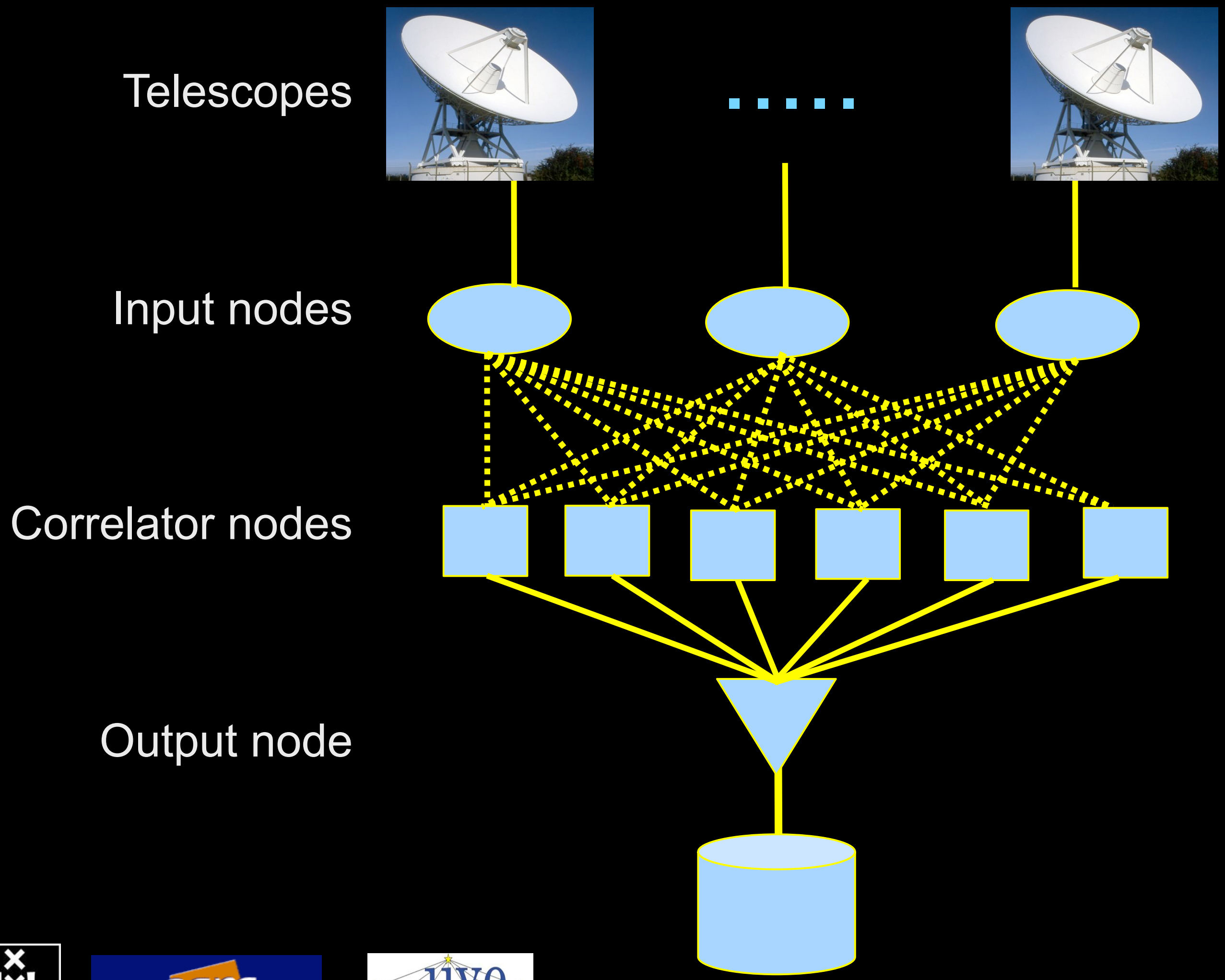
LOFAR as a Sensor Network



- 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

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

THIS IS A DATA FLOW PROBLEM !!!

The "Dead Cat" demo

SC2004 & iGrid2005

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

Produced by:
Michael Scarpa
Robert Belleman
Peter Sloot

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

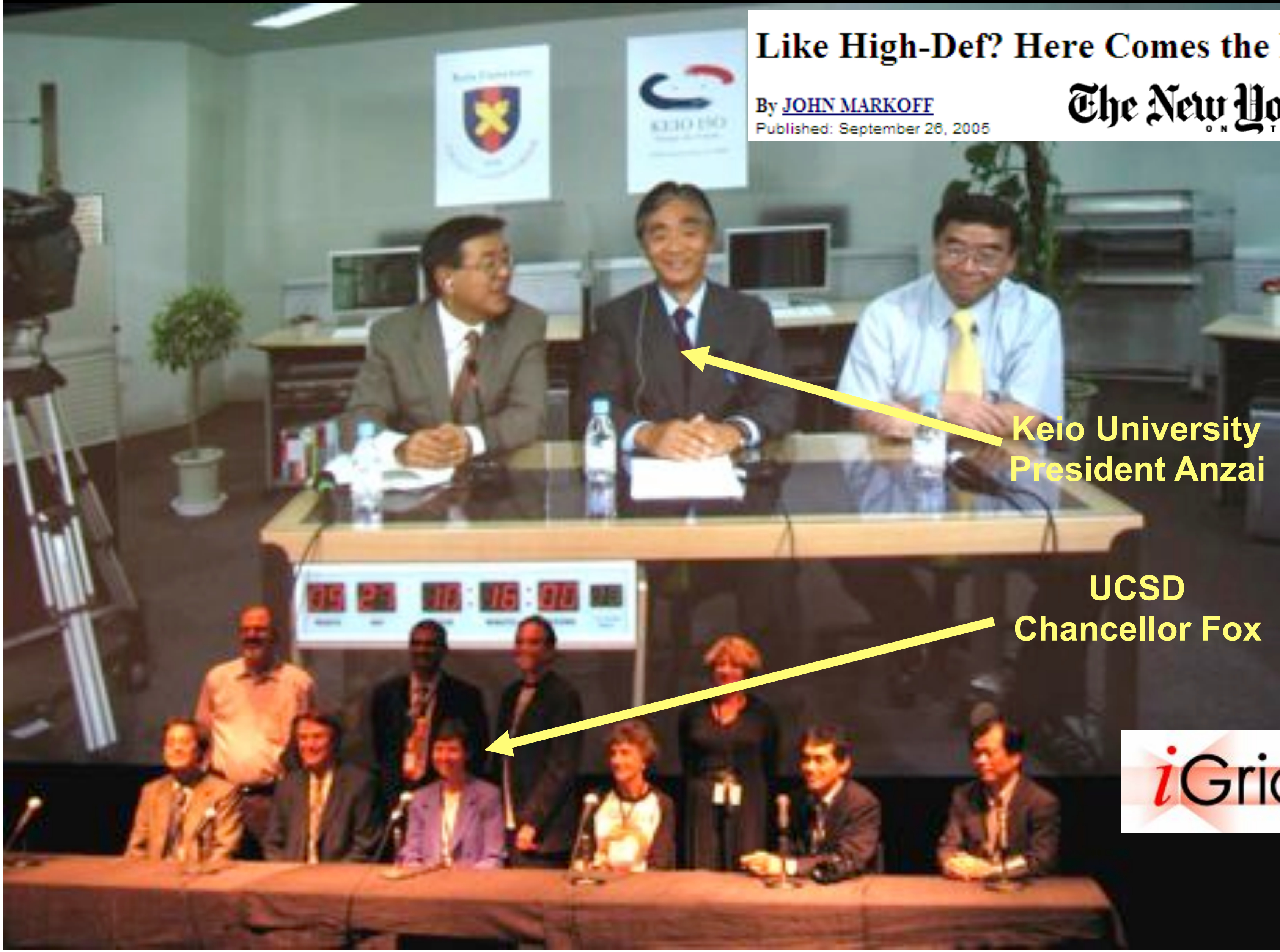
Keio/Calit2 Collaboration: Trans-Pacific 4K Teleconference

Like High-Def? Here Comes the Next Level

By JOHN MARKOFF

Published: September 26, 2005

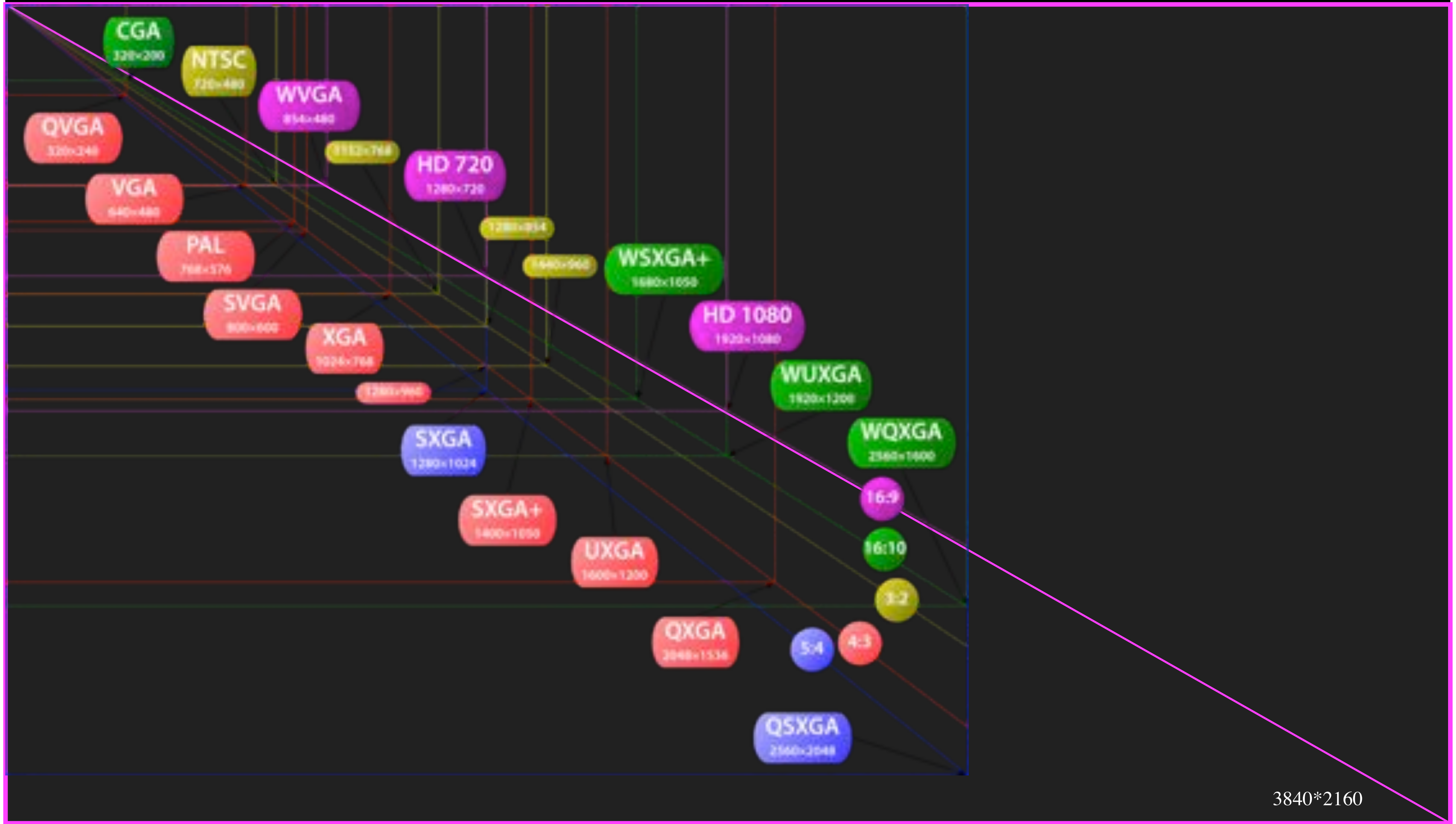
The New York Times ON THE WEB



Used 1Gbps Dedicated

Sony
NTT
SGI

iGrid 2005



Formats - Numbers - Bits

Format - Numbers - Bits (examples!)

Format	X	Y	Rate /s	Color bits/pix	Frame pix	Frame MByte	Flow MByt/s	Stream Gbit/s
720p HD	1280	720	60	24	921600	2.8	170	1.3
1080p HD	1920	1080	30	24	2073600	6.2	190	1.5
2k	2048	1080	24 48	36	2211840	10	240 480	1.2 2.4
SHD	3840	2160	30	24	8294400	25	750	6.0
4k	4096	2160	24	36	8847360	40	960	7.6

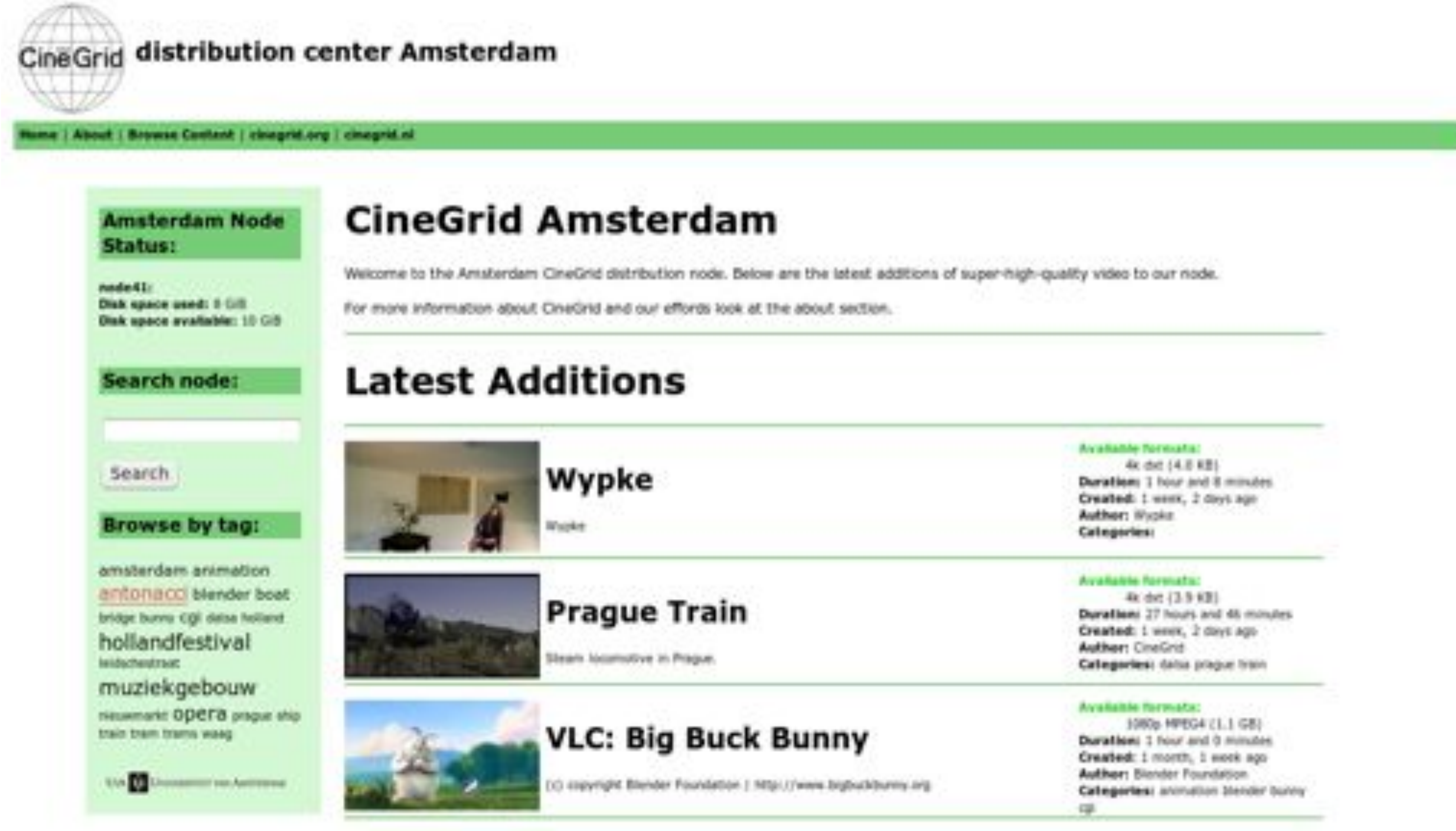
Note: this is excluding sound!
Note: these are raw uncompressed data rates ex overhead!

Buffer space

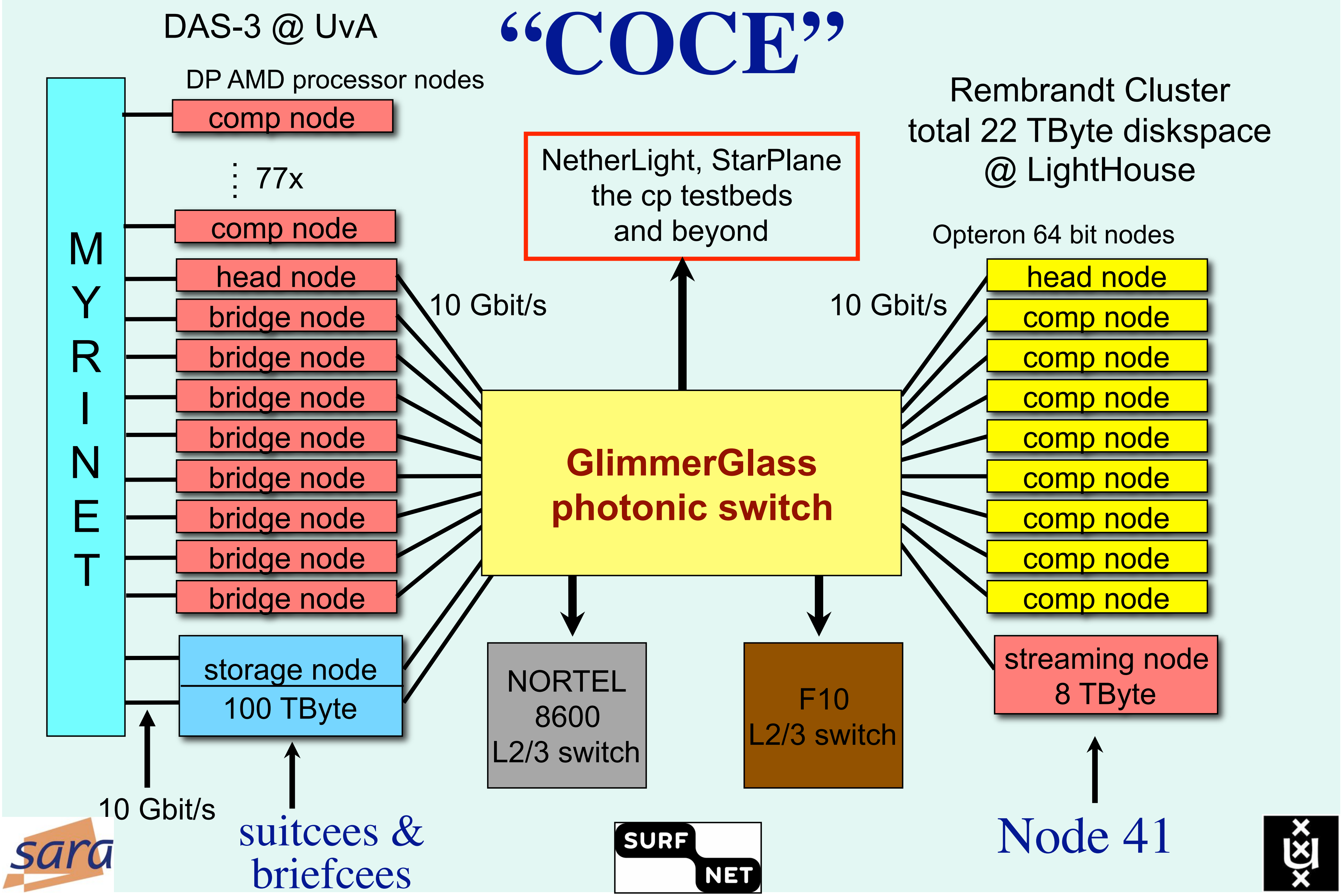
Window = RTT * BW

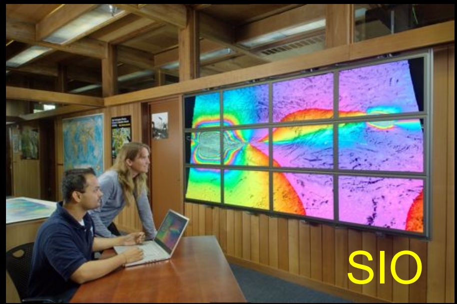
RTT	100 Mbit/s	1 Gbit/s	10 Gbit/s
1	12.5 kB	125 kB	1.25 MB
2	25 kB	250 kB	2.5 MB
5	62.5 kB	615 kB	6.15 MB
10	125 kB	1.25 MB	12.5 MB
20	250 kB	2.5 MB	25 MB
50	625 kB	6.25 MB	62.5 MB
100	1.25 MB	12.5 MB	125 MB
200	2.5 MB	25 MB	250 MB
500	6.25 MB	62.5 MB	625 MB
1000	12.5 MB	125 MB	1250 MB

CineGrid portal



Amsterdam CineGrid S/F node




SIO


NCMIR


USGS EDC

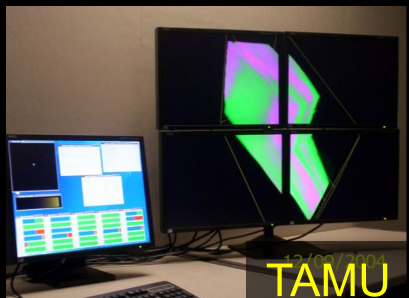

NCSA &
TRECC


SARA


KISTI


AIST


RINCON & Nortel


TAMU


UCI


UIC

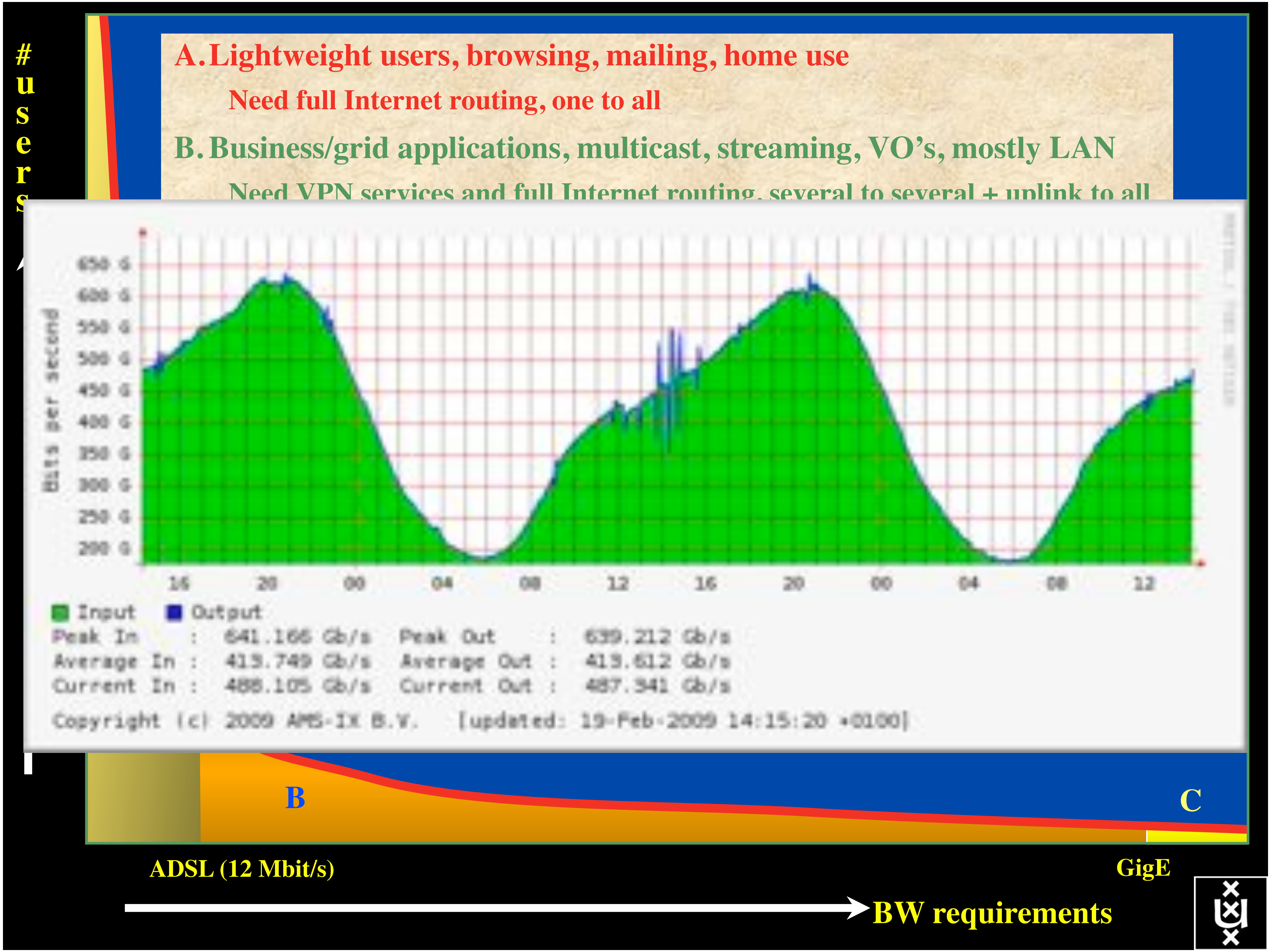

CALIT2

IJKDIJK

Sensor grid: instrument the dikes

First controlled breach occurred on sept 27th '08:

30000 sensors (microphones) to cover all Dutch dikes



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
users
Bits per second
650 G
600 G
550 G
500 G
450 G
400 G
350 G
300 G
250 G
200 G
16 20 00 04 08 12 16 20 00 04 08 12
Input
Output
Peak In : 641.166 Gb/s Peak Out : 639.212 Gb/s
Average In : 413.749 Gb/s Average Out : 413.612 Gb/s
Current In : 488.105 Gb/s Current Out : 487.341 Gb/s
Copyright (c) 2009 AMS-IX B.V. [updated: 19-Feb-2009 14:15:20 +0100]
B
C
ADSL (12 Mbit/s)
GigE
BW requirements

Towards Hybrid Networking!

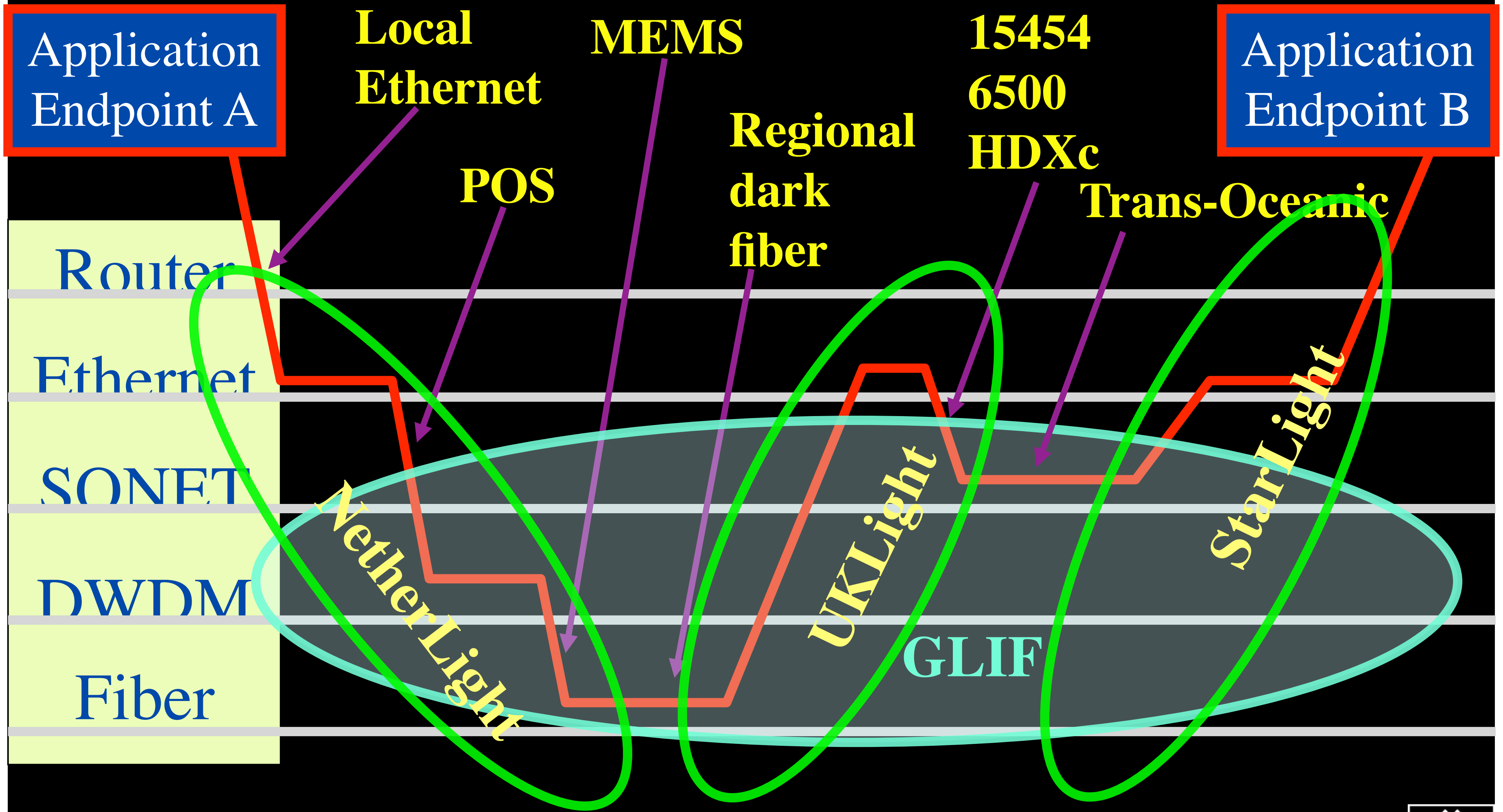
- Costs of photonic equipment 10% of switching 10 % of full routing
  – for same throughput!
  – Photonic vs Optical (optical used for SONET, etc, 10-50 k$/port)
  – DWDM lasers for long reach expensive, 10-50 k$
- Bottom line: look for a hybrid architecture which serves all classes in a cost effective 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
0.5 W/port

L2 ≈ 5-8 k$/port
10-15 W/port

L3 ≈ 75+ k$/port
250 W/port

How low can you go?



In The Netherlands SURFnet connects between 180:

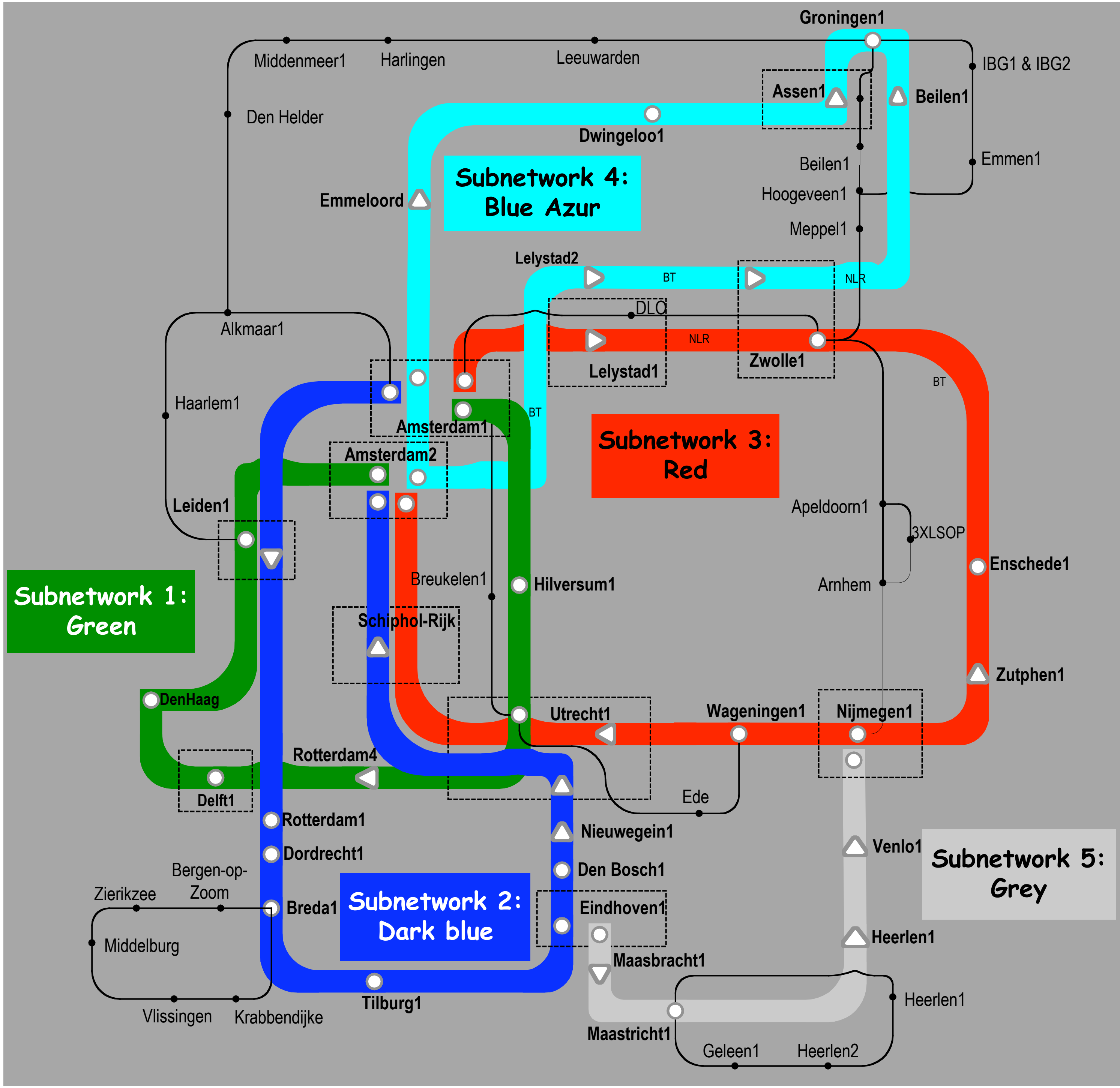
- universities;
- academic hospitals;
- most polytechnics;
- research centers.

with an indirect ~750K user base

~ 8860 km

scale comparable to railway system

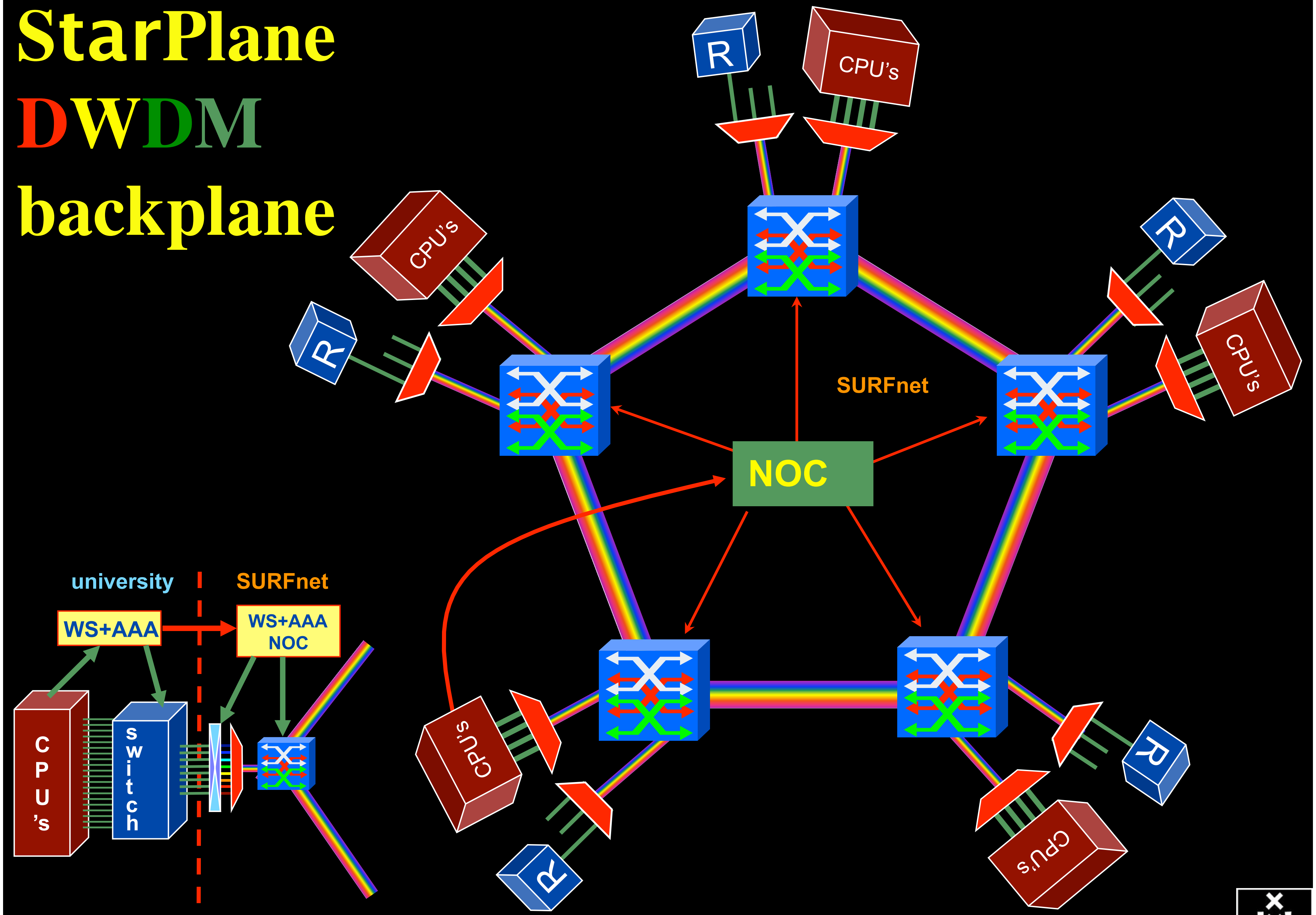
Common Photonic Layer (CPL) in SURFnet6



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

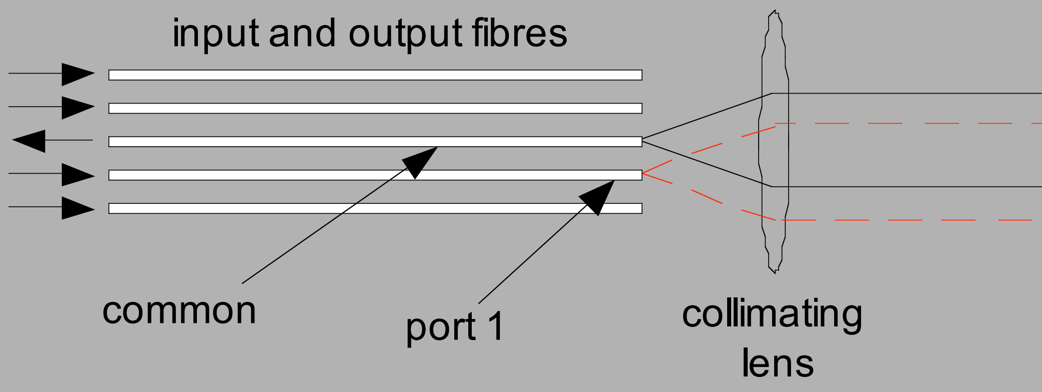
SURFnet

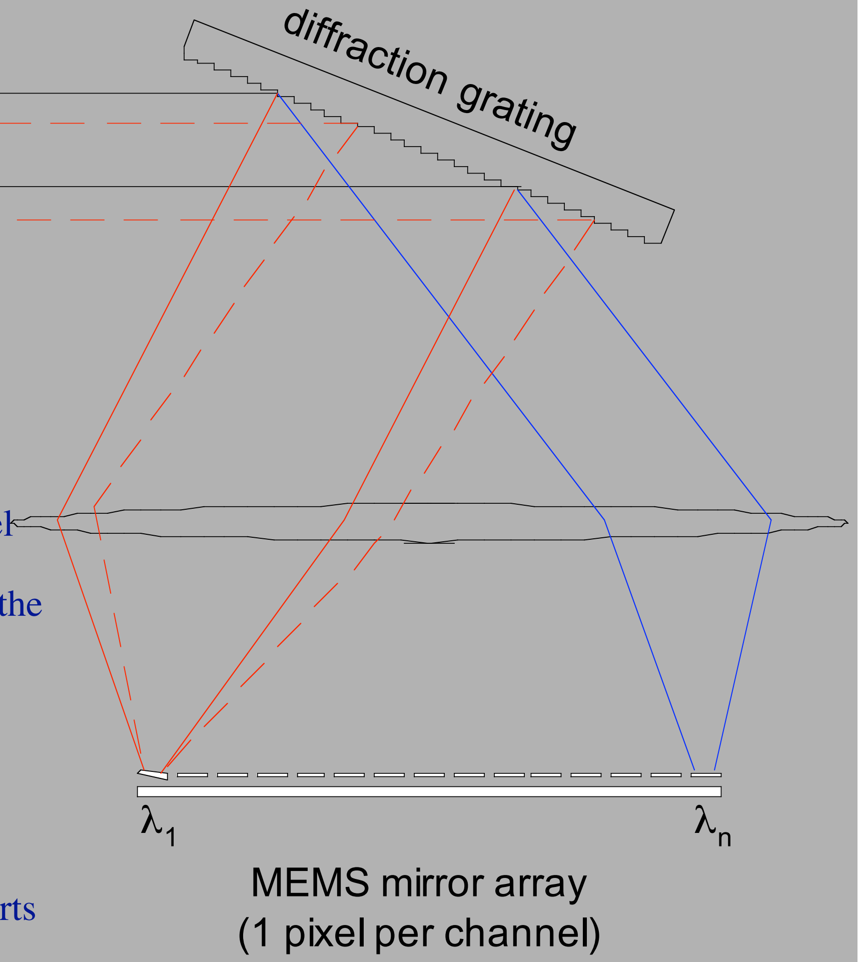
StarPlane DWDM backplane



CdL

Module Operation

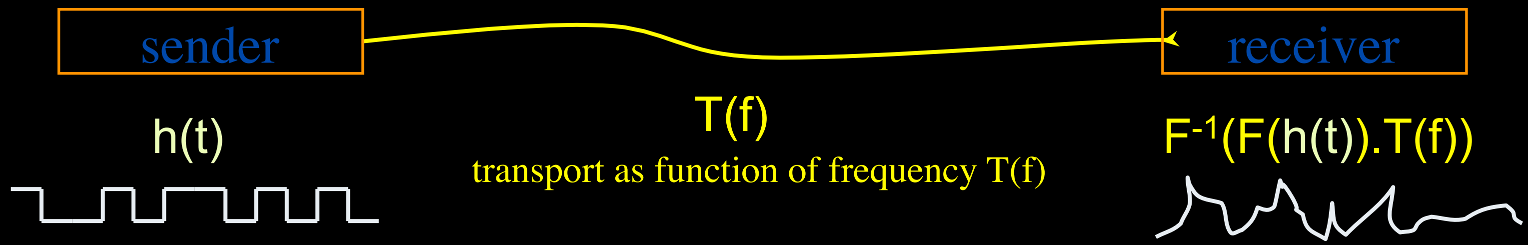




- this schematic shows
  - several input fibres and one output fibre
  - light is focused and diffracted such that each channel lands on a different MEMS mirror
  - the MEMS mirror is electronically controlled to tilt the reflecting surface
  - the angle of tilt directs the light to the correct port
- in this example:
  - channel 1 is coming in on port 1 (shown in red)
  - when it hits the MEMS mirror the mirror is tilted to direct this channel from port 1 to the common
  - only port 1 satisfies this angle, therefore all other ports are blocked

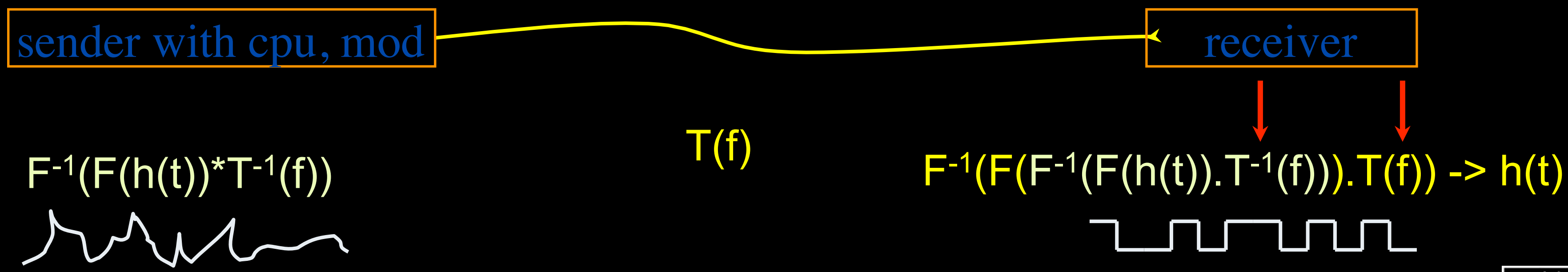
Dispersion compensating modem: eDCO from NORTEL

(Try to Google eDCO :-)



Solution in 5 easy steps for dummy's :

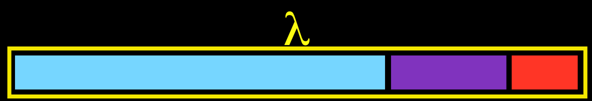
1. try to figure out T(f) by trial and error
2. invert T(f) -> $T^{-1}(f)$
3. computationally multiply $T^{-1}(f)$ with Fourier transform of bit pattern to send
4. inverse Fourier transform the result from frequency to time space
5. modulate laser with resulting $h'(t) = F^{-1}(F(h(t)).T^{-1}(f))$



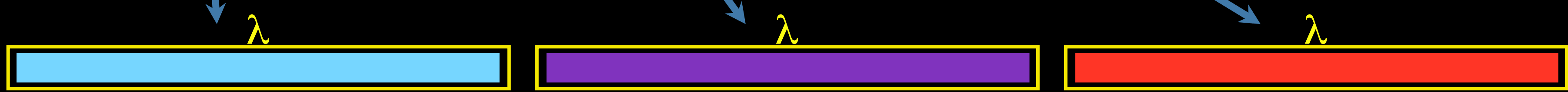
(ps. due to power ~ square E the signal to send **looks** like uncompensated received but is not)

QOS in a non destructive way!

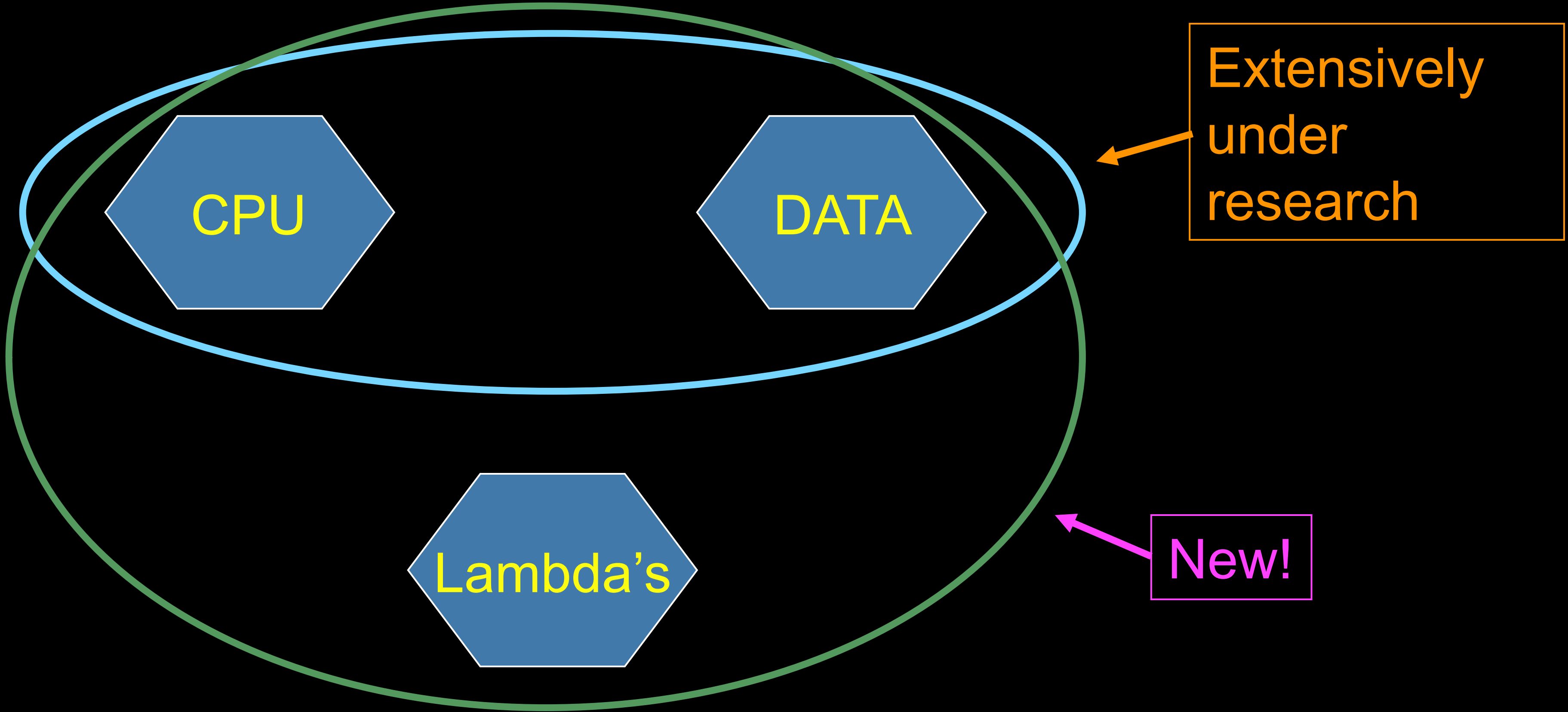
- Destructive QOS:
  - have a link or λ
  - set part of it aside for a lucky few under higher priority
  - rest gets less service



- 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!



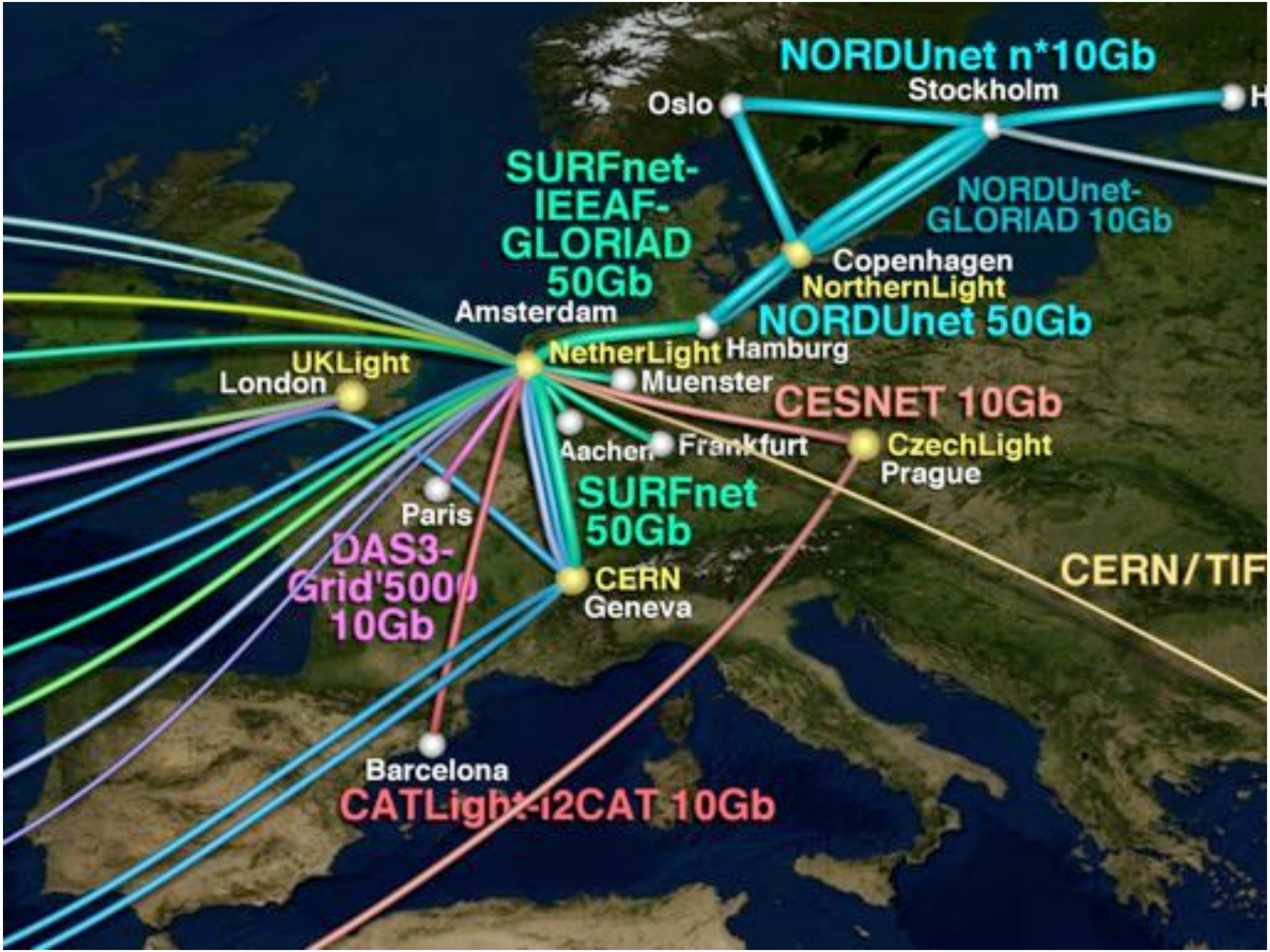
GRID Co-scheduling problem space

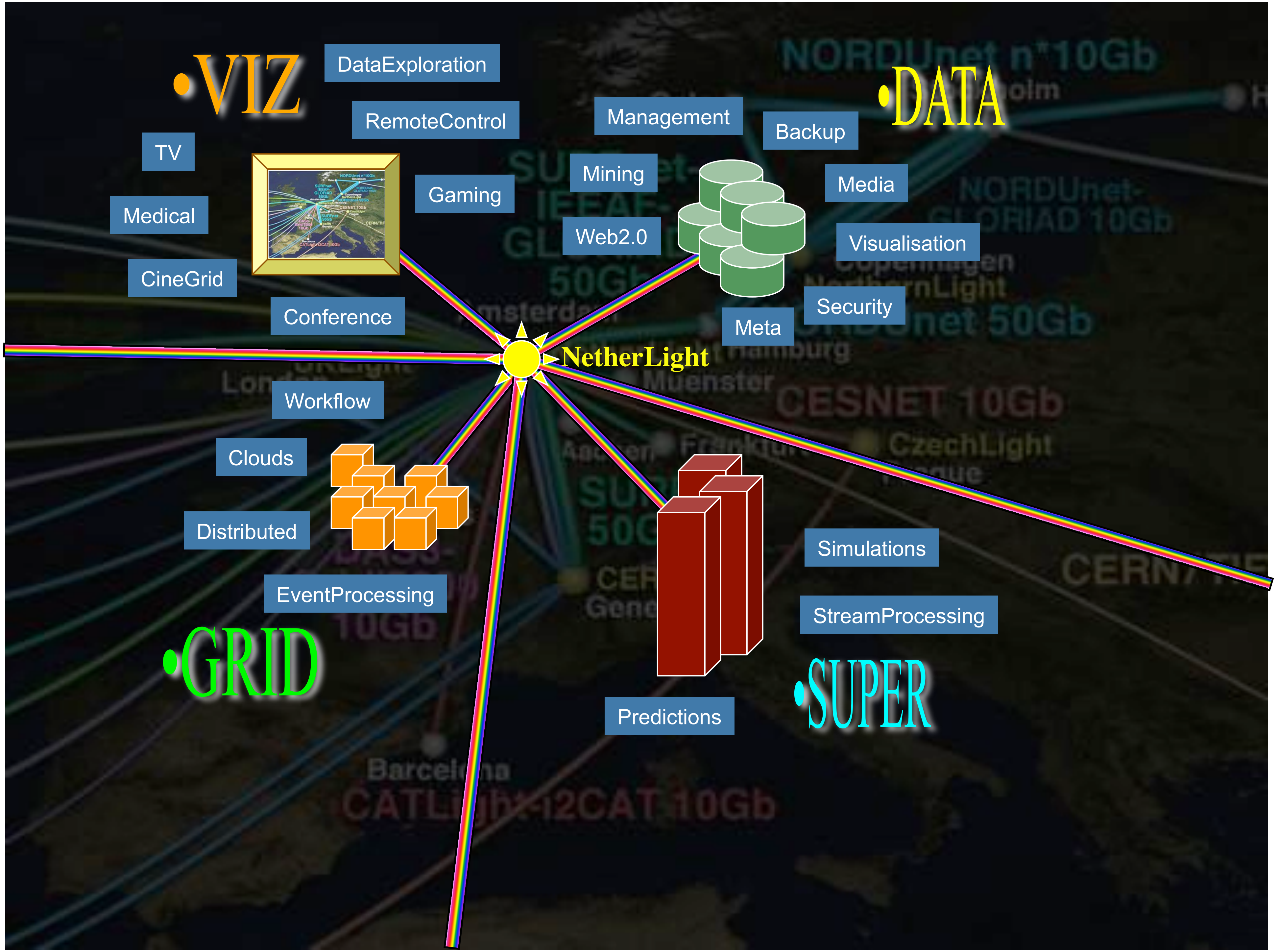


The StarPlane vision is to give flexibility directly to the applications by allowing them to choose the logical topology in real time, ultimately with sub-second lambda switching times on part of the SURFnet6 infrastructure.

GLIF 2008

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

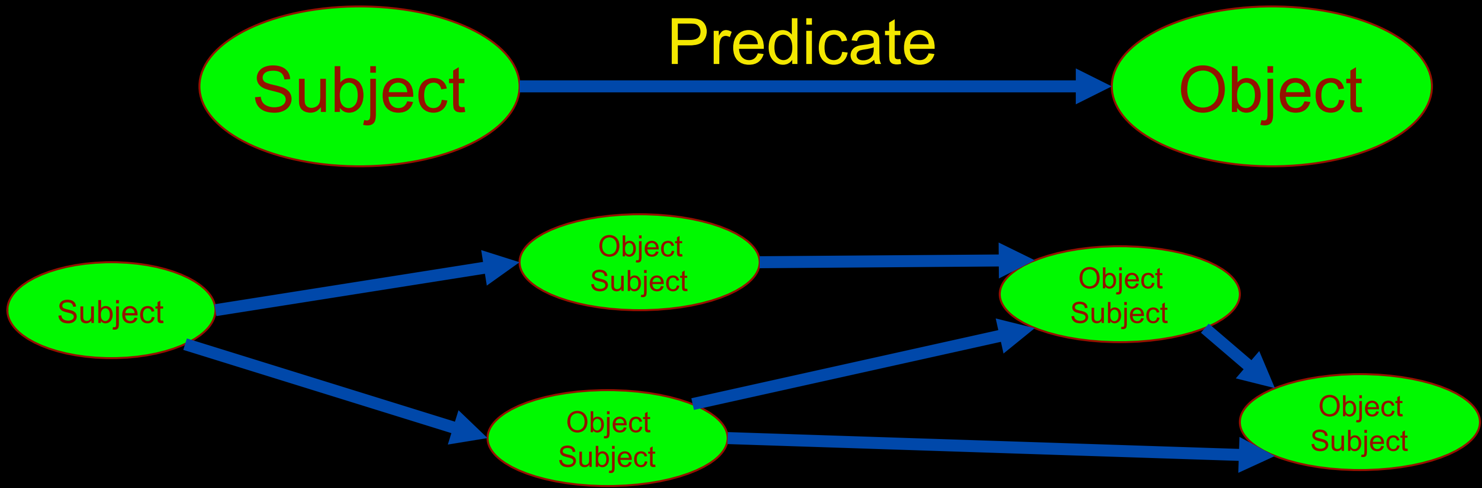

NORDUnet n*10Gb
Stockholm
Oslo
SURFnet-
IEEAF-
GLORIAD
50Gb
NORDUnet-
GLORIAD 10Gb
Copenhagen
NorthernLight
NORDUnet 50Gb
Amsterdam
NetherLight
Hamburg
UKLight
London
Muenster
CESNET 10Gb
Aachen
Frankfurt
CzechLight
Prague
SURFnet
50Gb
Paris
DAS3-
Grid'5000
10Gb
CERN
Geneva
CERN/TIF
Barcelona
CATLight-i2CAT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

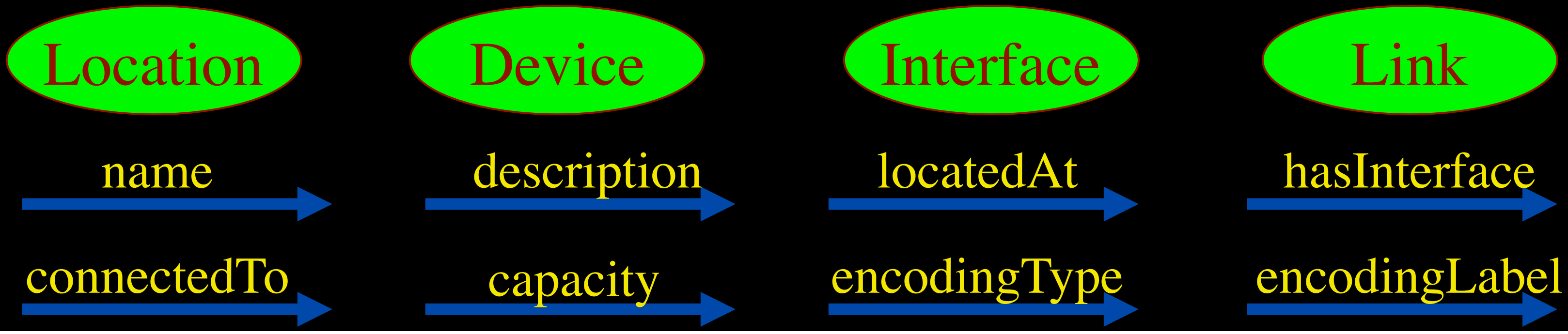


•VIZ
DataExploration
RemoteControl
TV
Medical
Gaming
CineGrid
Conference
Management
Backup
•DATA
Mining
Media
Web2.0
Visualisation
Security
Meta
NetherLight
Workflow
Clouds
Distributed
EventProcessing
•GRID
Simulations
StreamProcessing
•SUPER
Predictions

Network Description Language

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



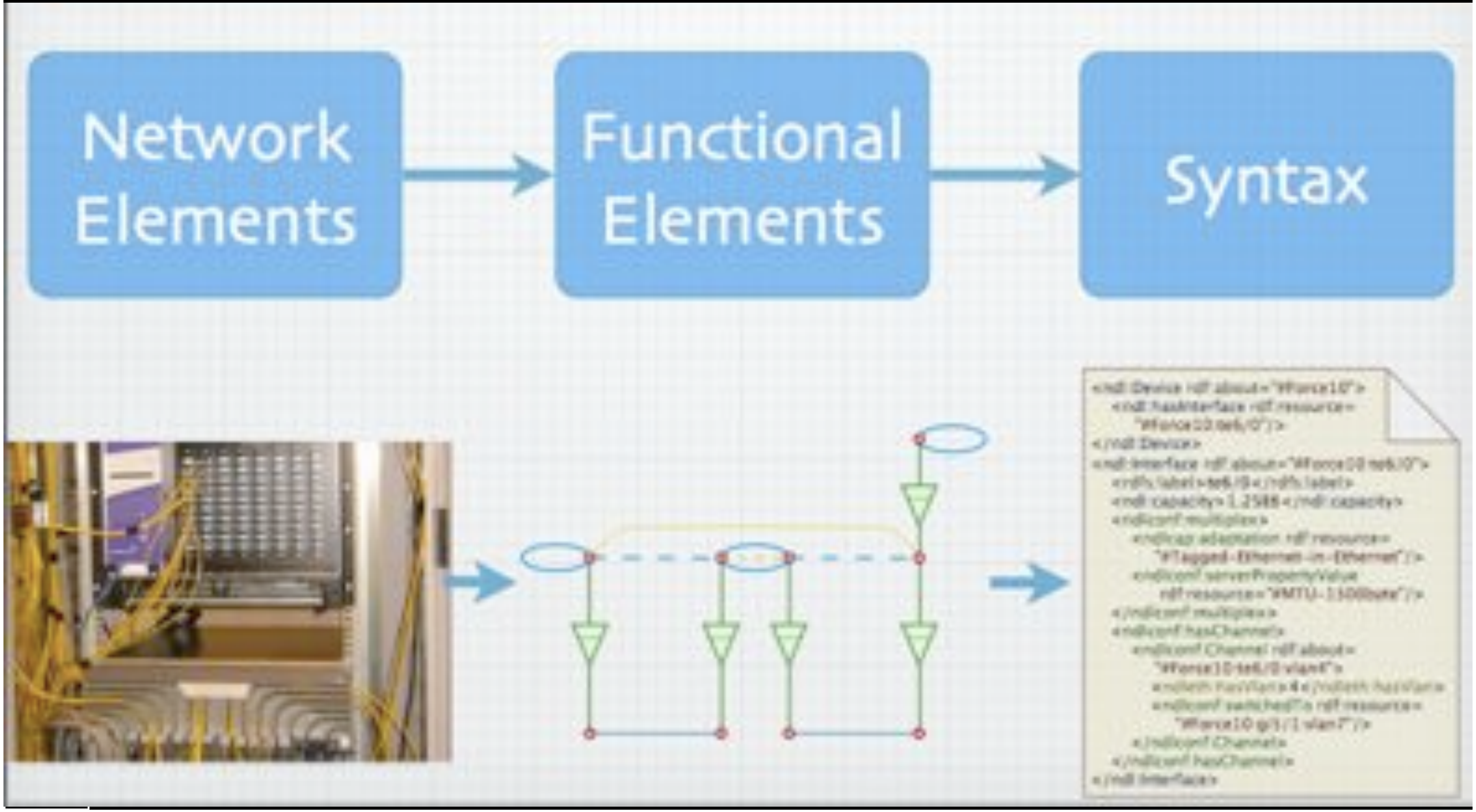


Network Description Language

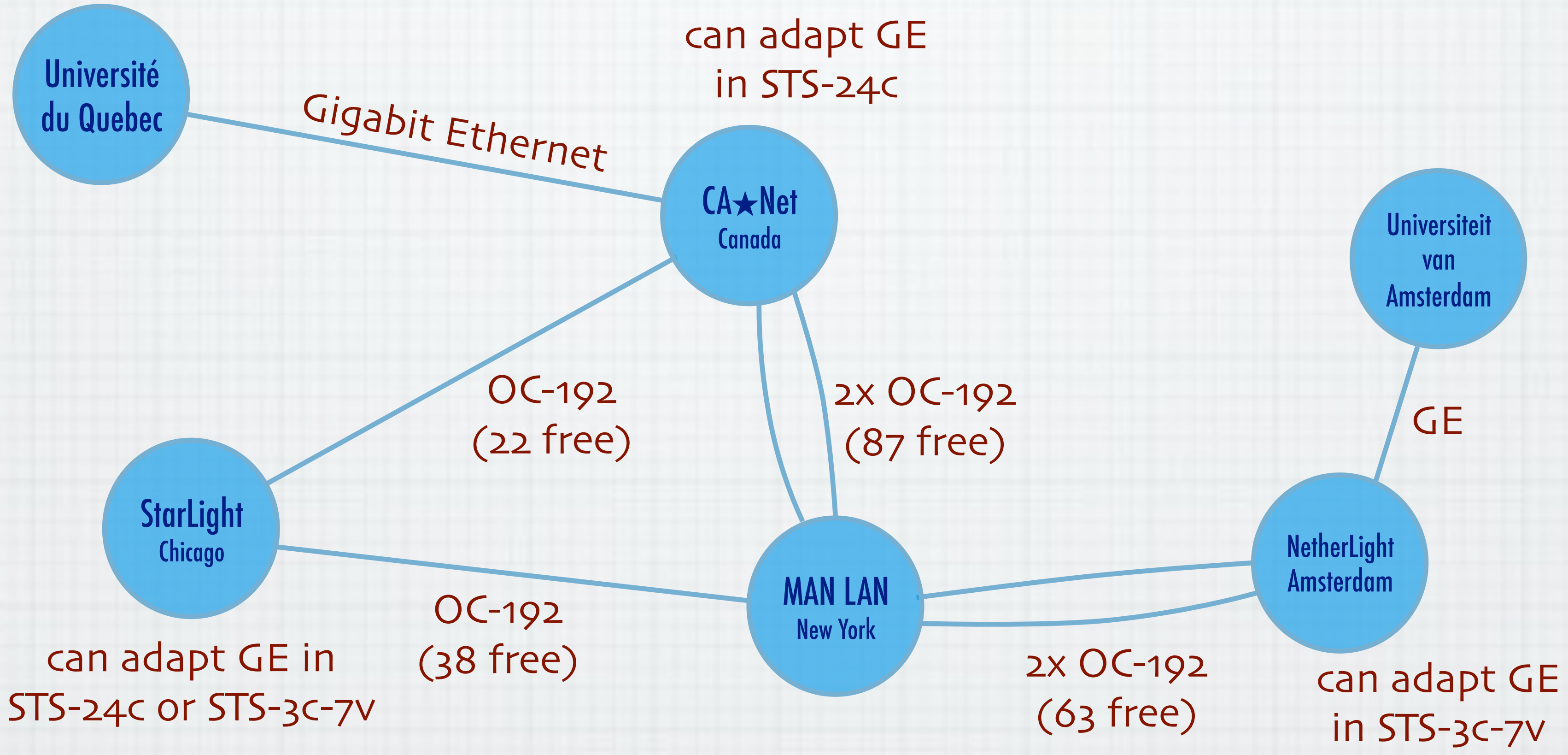
Choice of RDF instead of XML syntax

Grounded modeling based on G0805 description:

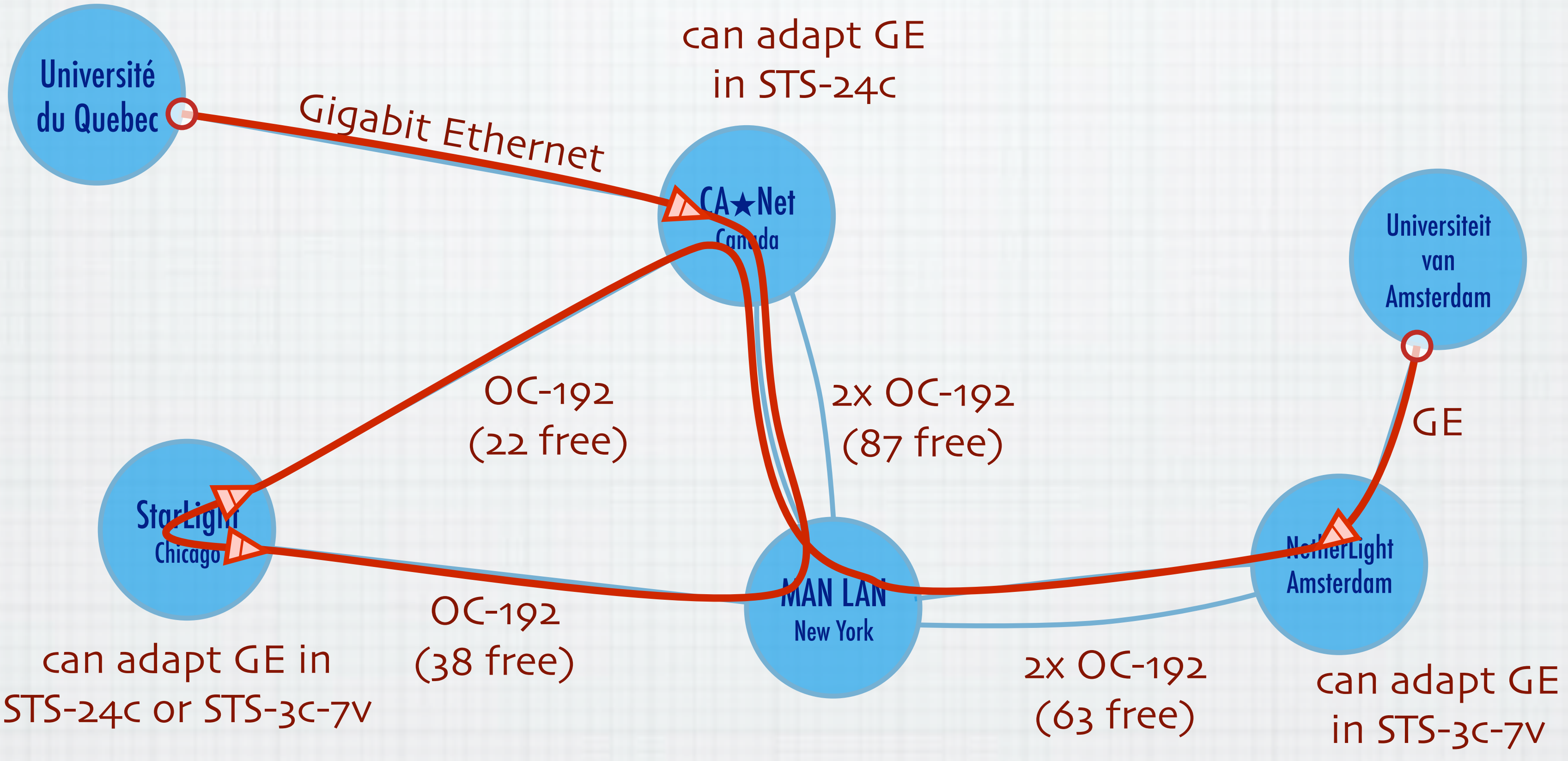
Article: F. Dijkstra, B. Andree, K. Koymans, J. van der Ham, P. Grosso, C. de Laat, *"A Multi-Layer Network Model Based on ITU-T G.805"*



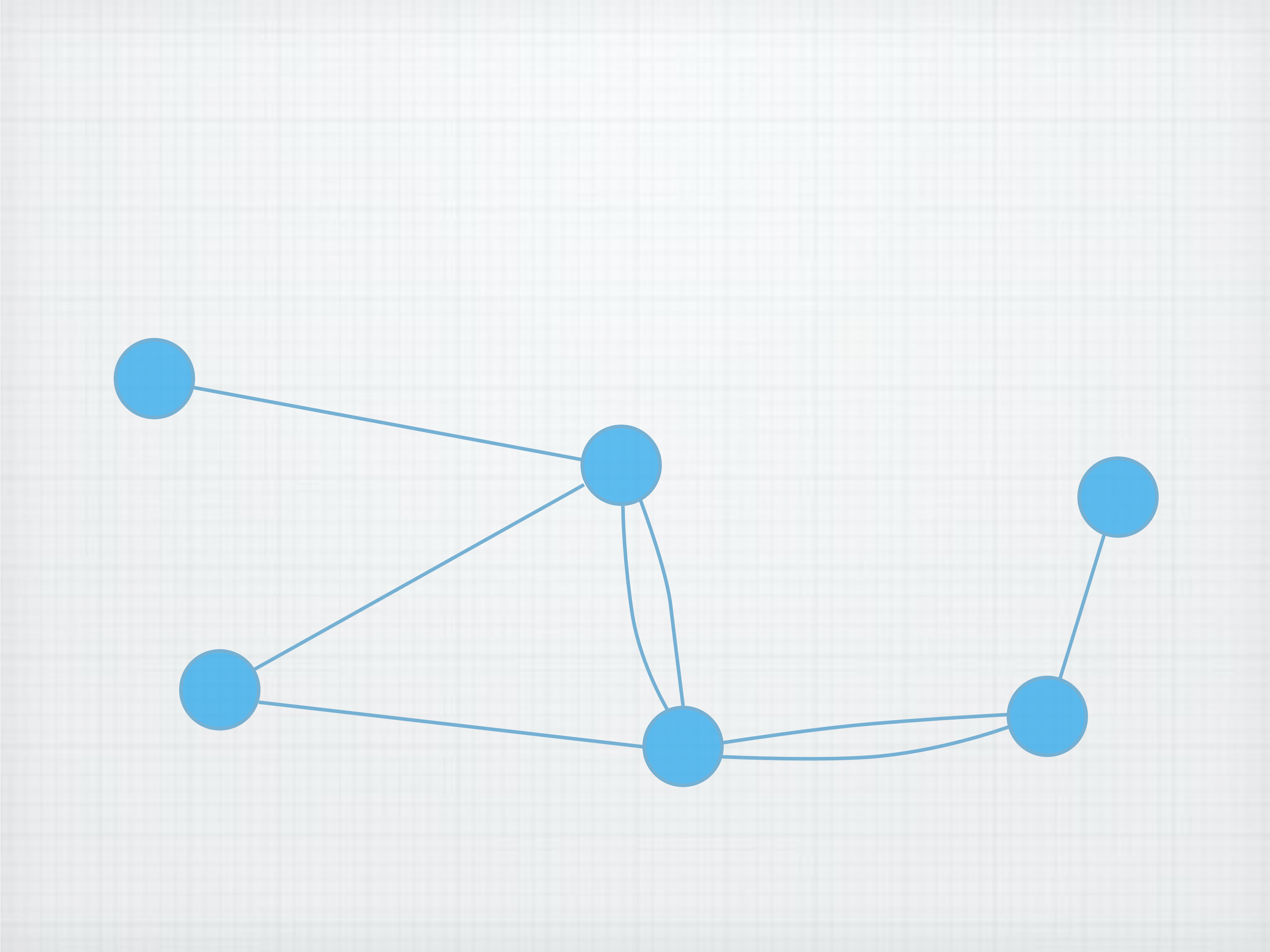
A weird example

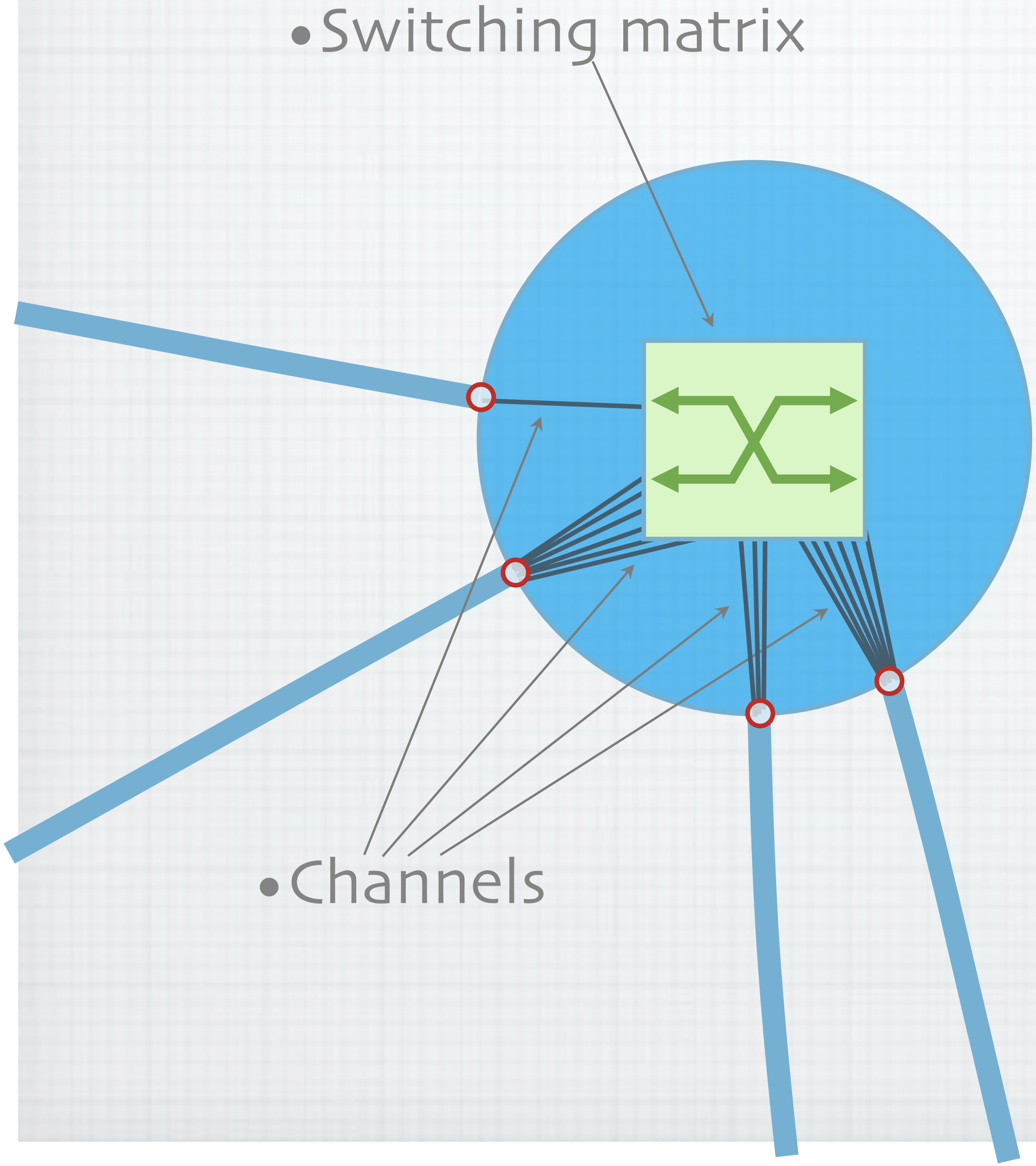


The result :-)



Thanks to Freek Dijkstra & team

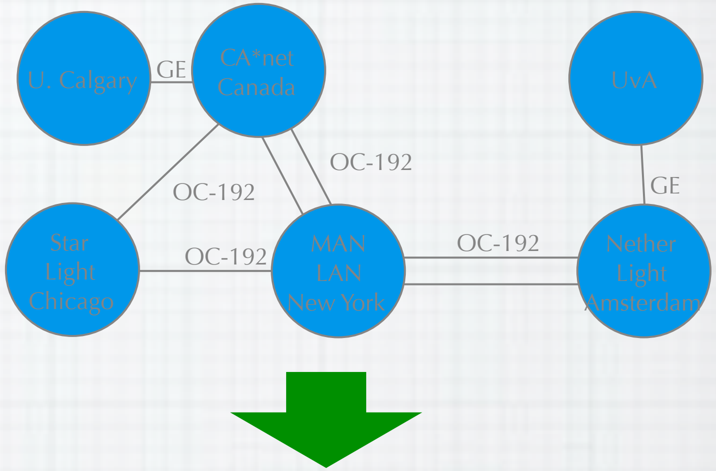


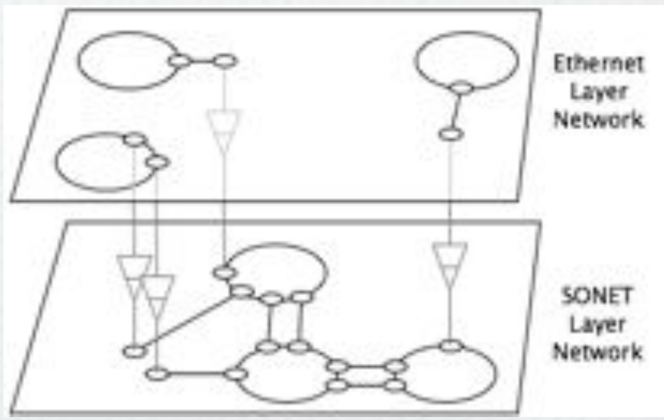


- A devices switches data based on:
  - The source interface
  - One or more labels
- Example label types:
  - Ethernet VLAN
  - SONET STS Channel
  - Wavelength (λ)
- For example, all data from channel 31 of interface 2 is forwarded to channel 28 of interface 4.

NDL Multilayer Extension

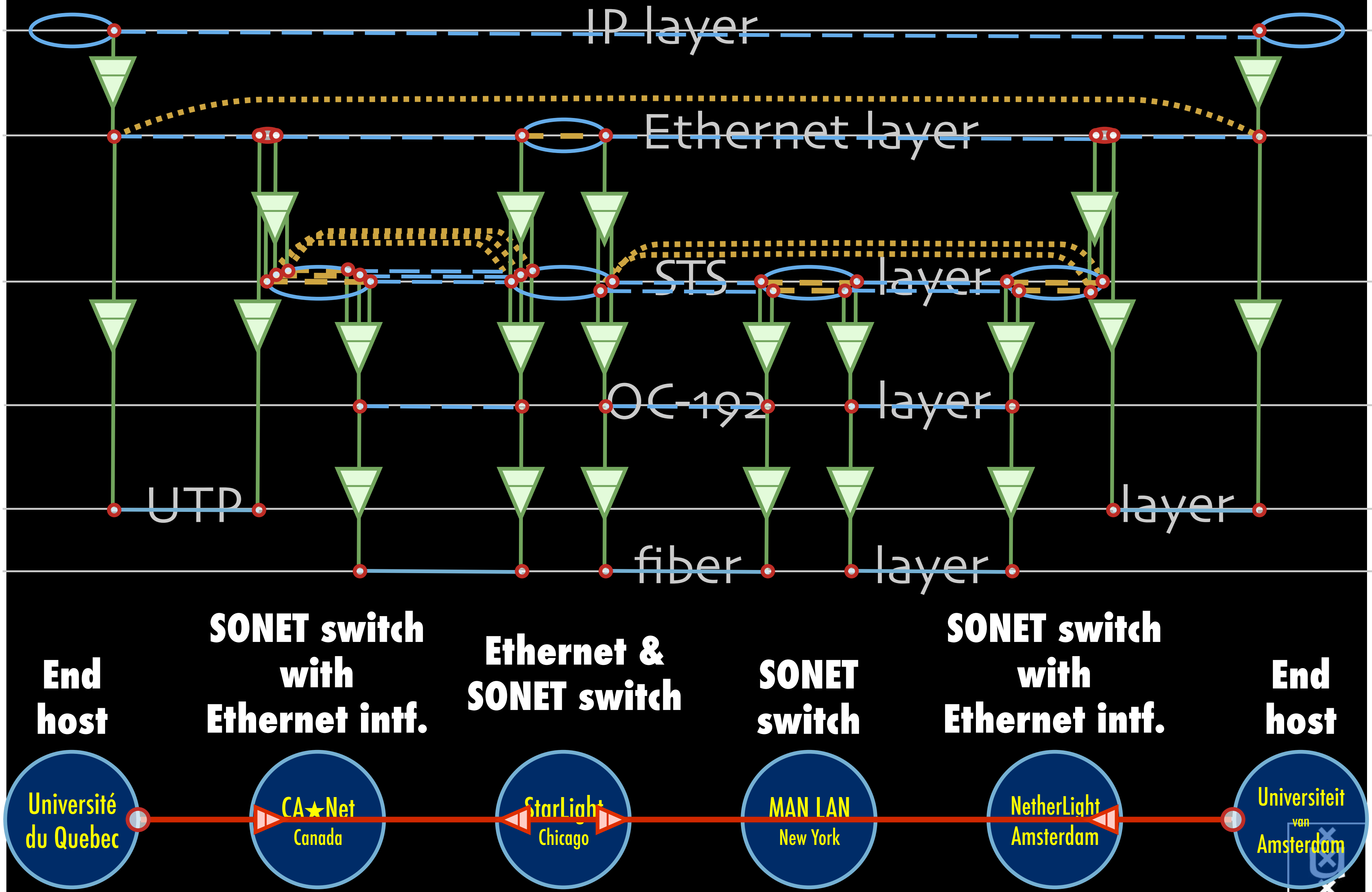
- ITU-T G.805 describes functional elements (e.g. adaptation, termination functions, link connections, etc.) to describe **network connections**.
- We extended these function elements (e.g. with potential adaptation functions) to describes **networks**.
- We created a model to map actual network elements to functional elements.
- Defined a simple algebra to define correctness of a connection
- We created a NDL extension to describe these functional elements.





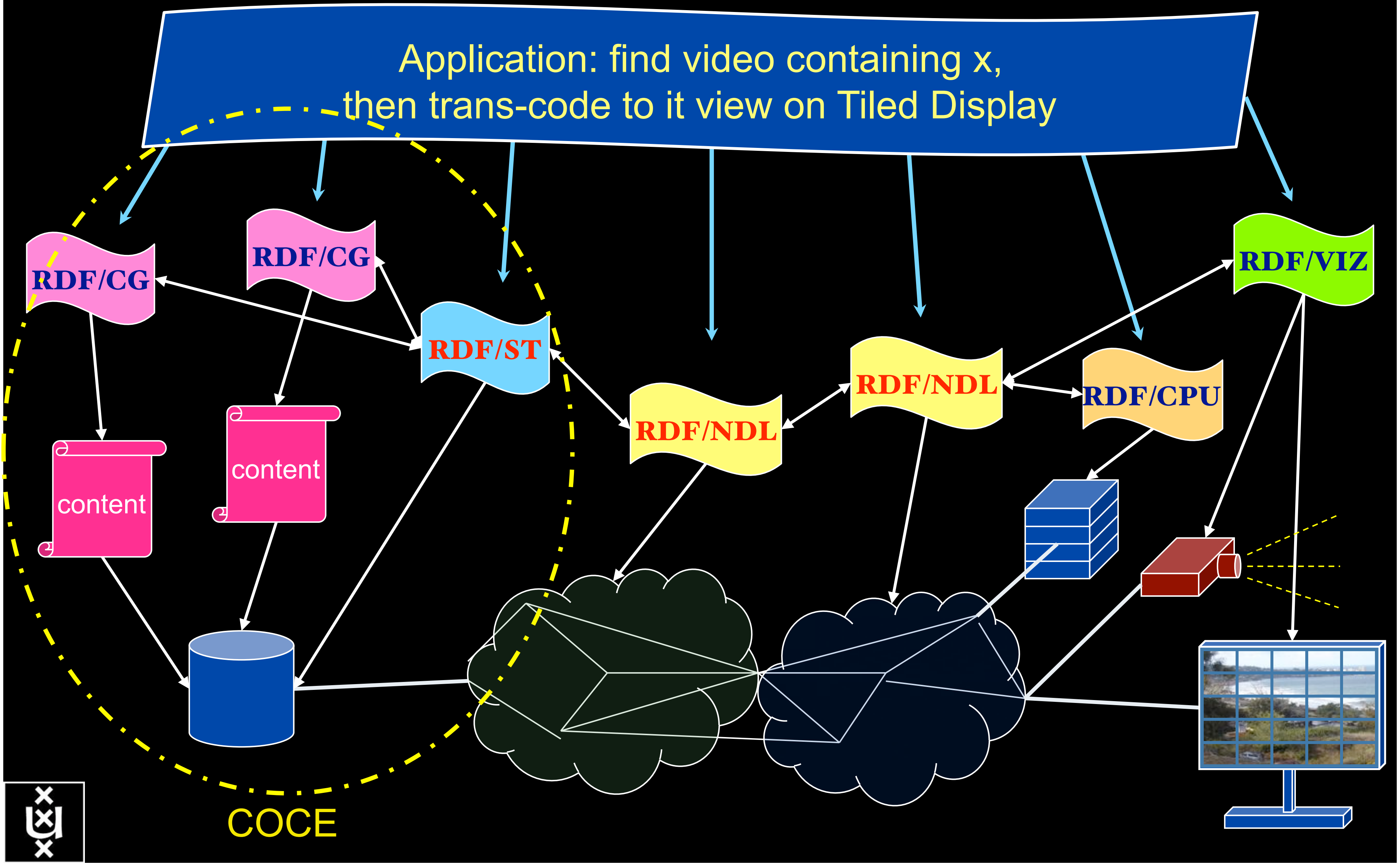
Simplified model to map network elements to functional elements

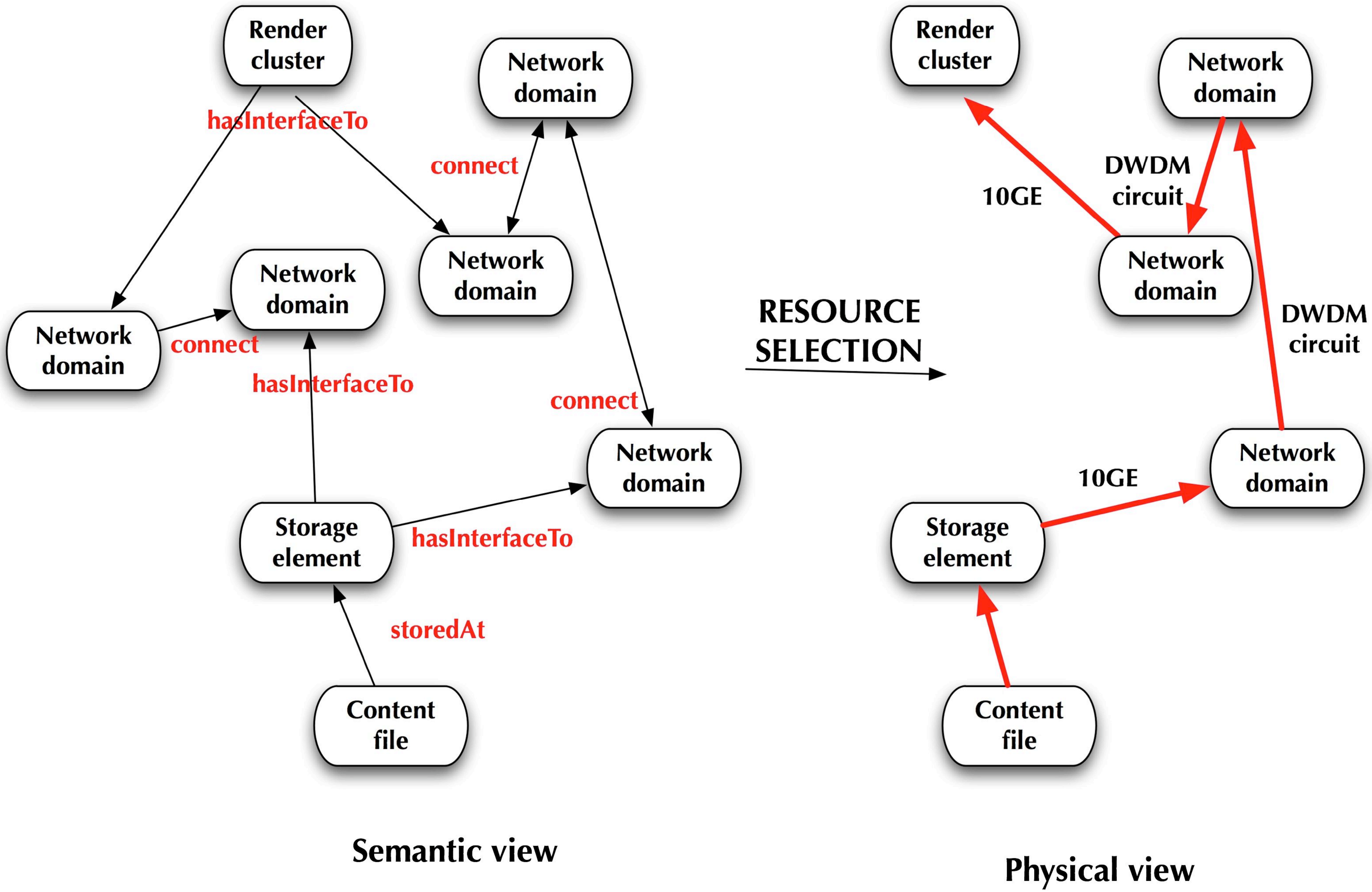
Multi-layer extensions to NDL




Map
Satellite
Hybrid
Greenland
Iceland
Canada
United States
North Pacific Ocean
North Atlantic Ocean
Japan
Mexico
Cuba
Venezuela
Colombia
Brasil
Peru
Bolivia
Chile
Argentina
Australia
South Pacific Ocean
South Atlantic Ocean
Norge Norway
España Spain
Italia Italy
Algeria
Libya
Mauritania
Mali
Niger
Chad
Nigeria
Angola
Namibia
Indonesia
Papua New Guinea
Google

RDF describing Infrastructure
"I want"





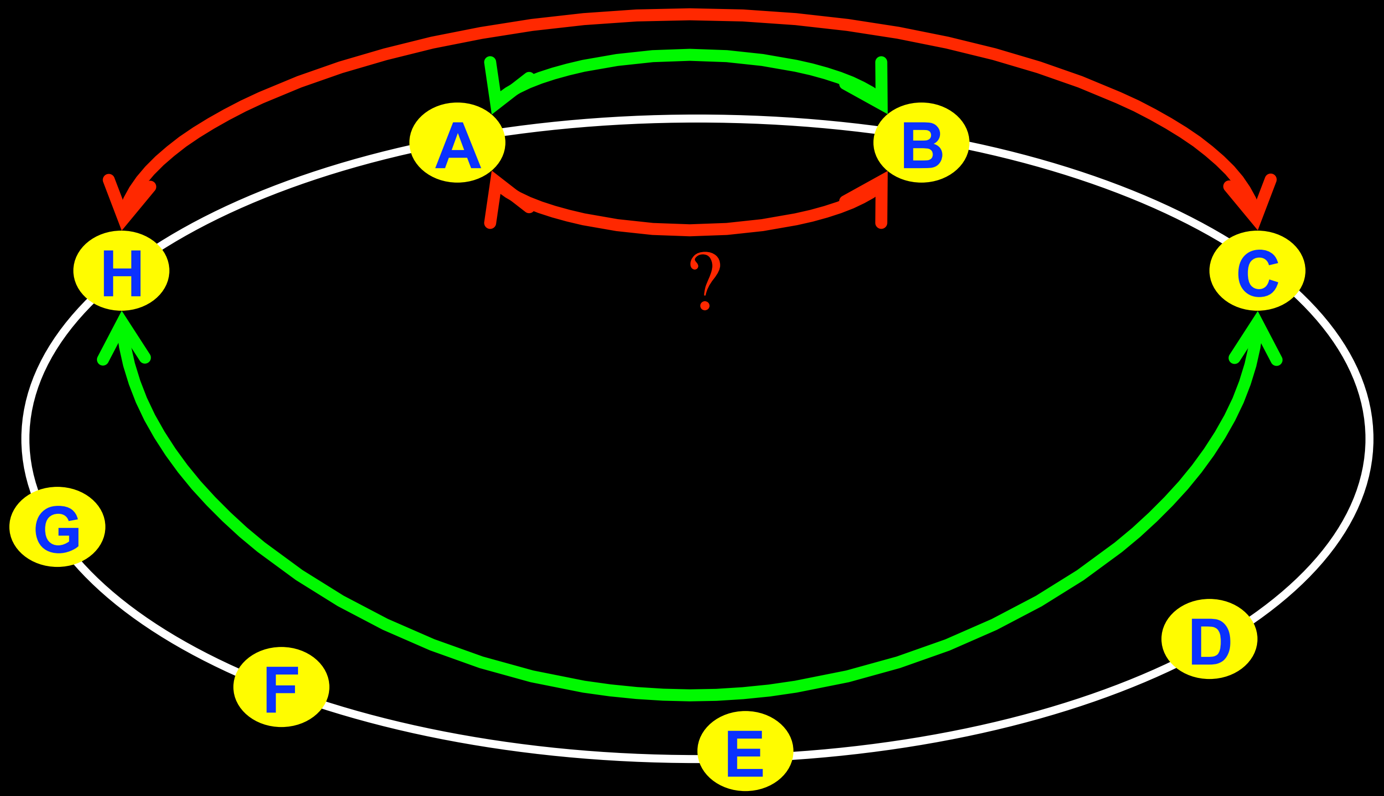
Semantic Reasoning

The Problem

I want HC and AB

Success depends on the order

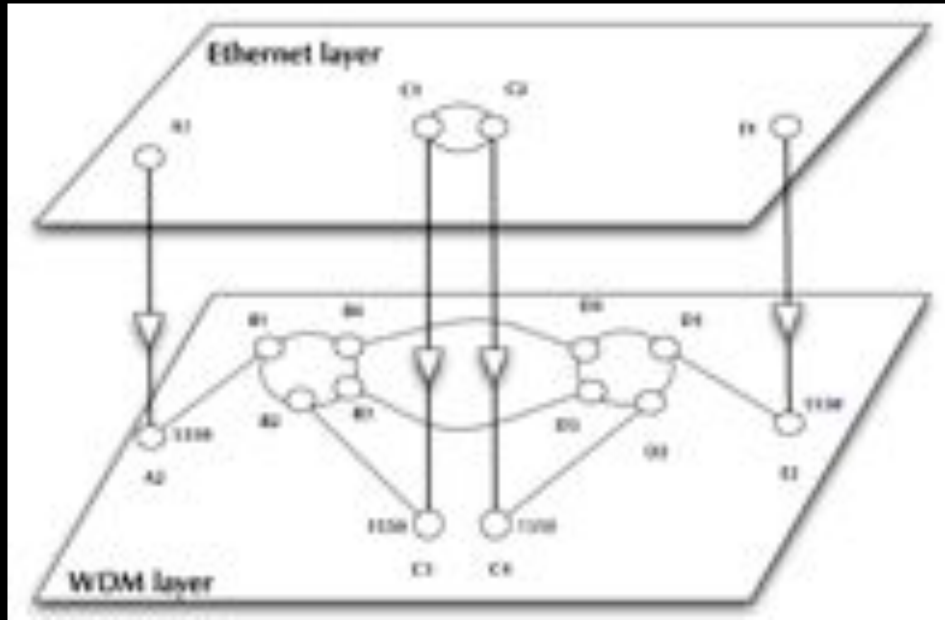
Wouldn't it be nice if I could request [HC, AB, ...]

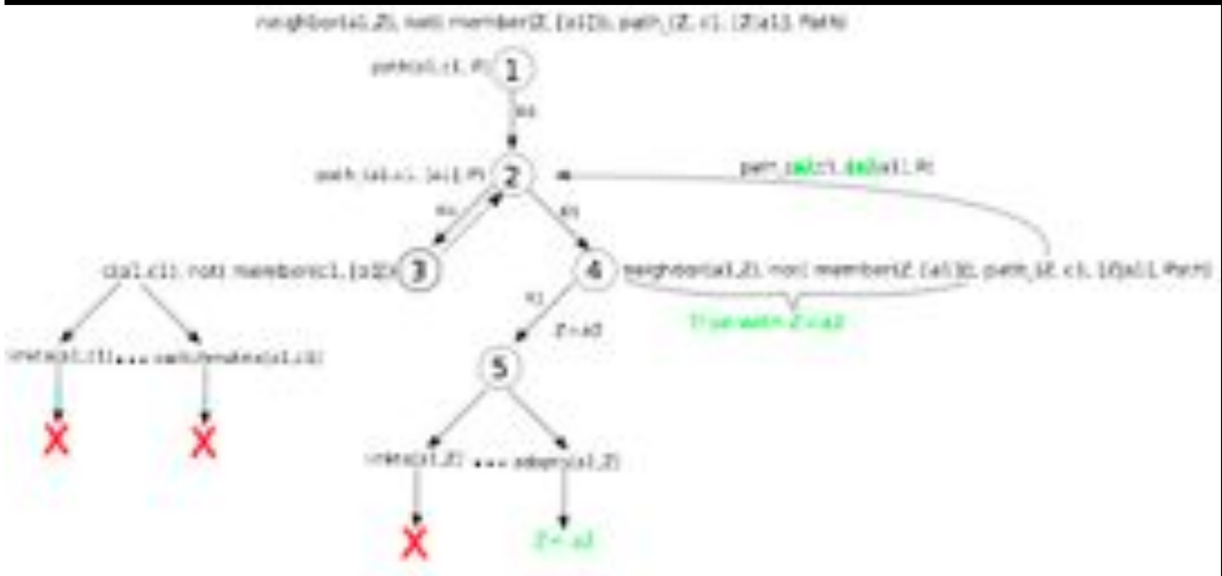


NDL + PROLOG

Research Questions:

- order of requests
- complex requests
- Usable leftovers

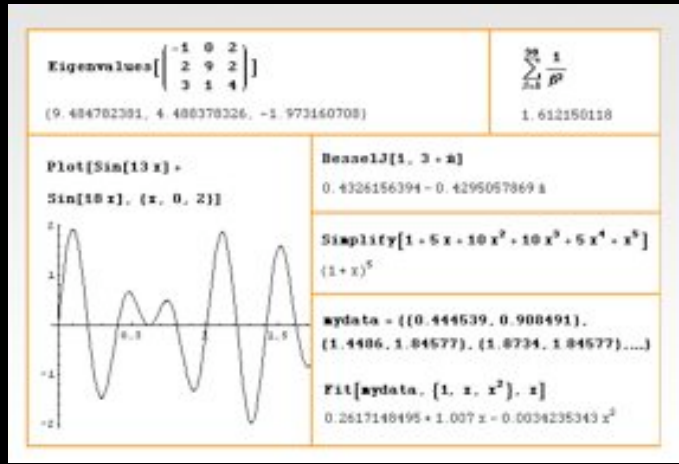


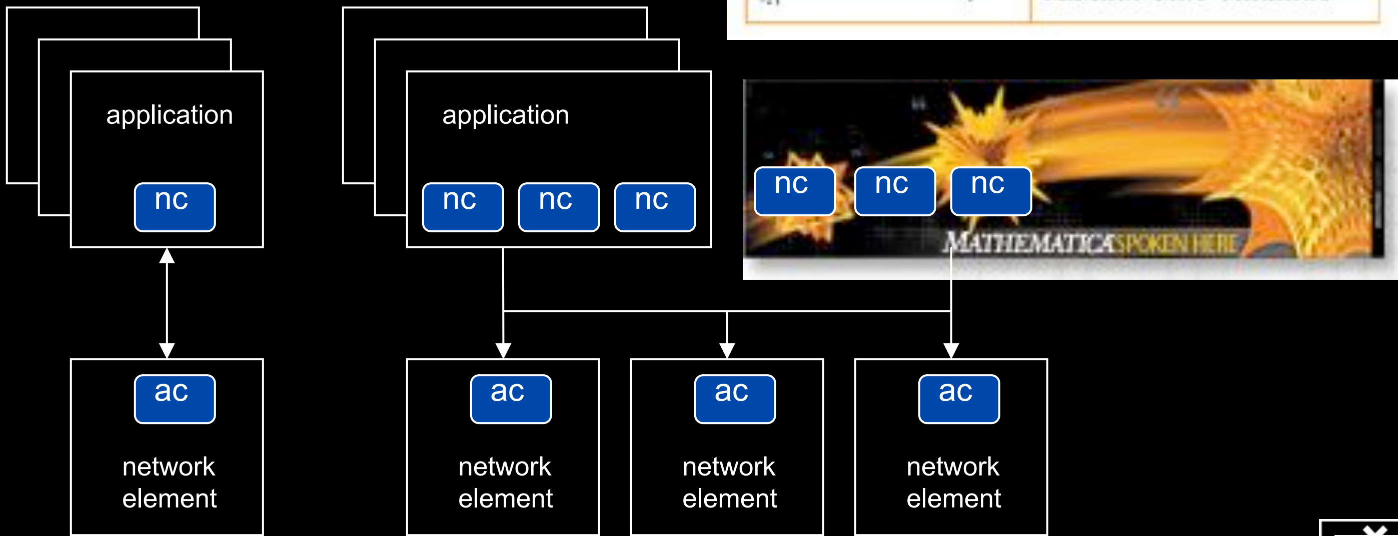


- Reason about graphs
- Find sub-graphs that comply with rules

User Programmable Virtualized Networks allows the results of decades of computer science to handle the complexities of application specific networking.

- The network is virtualized as a collection of resources
- UPVNs enable network resources to be programmed as part of the application
- Mathematica, a powerful mathematical software system, can interact with real networks using UPVNs

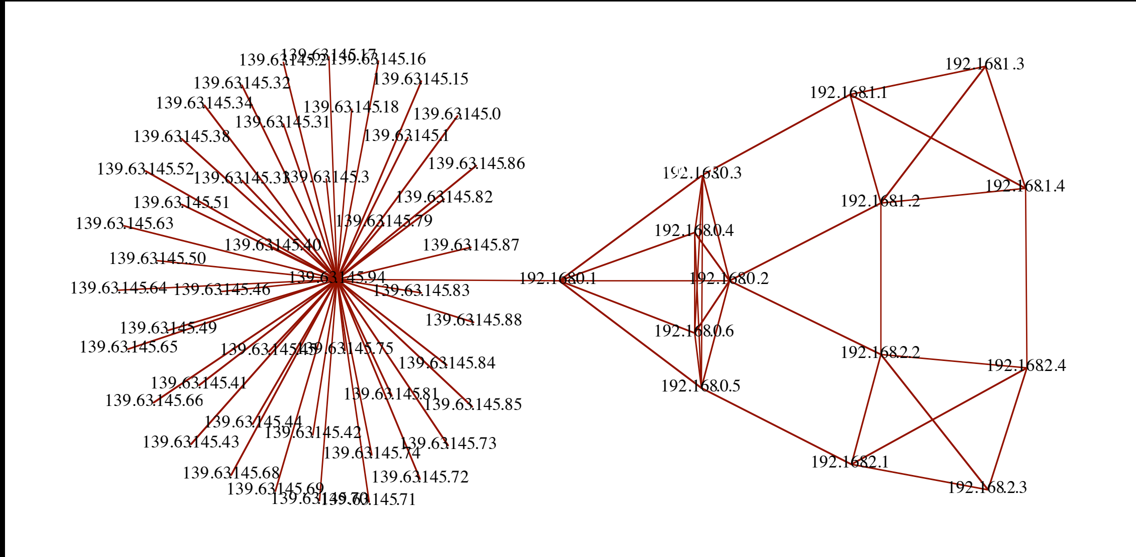




Mathematica enables advanced graph queries, visualizations and real-time network manipulations on UPVNs

Topology matters can be dealt with algorithmically

Results can be persisted using a transaction service built in UPVN

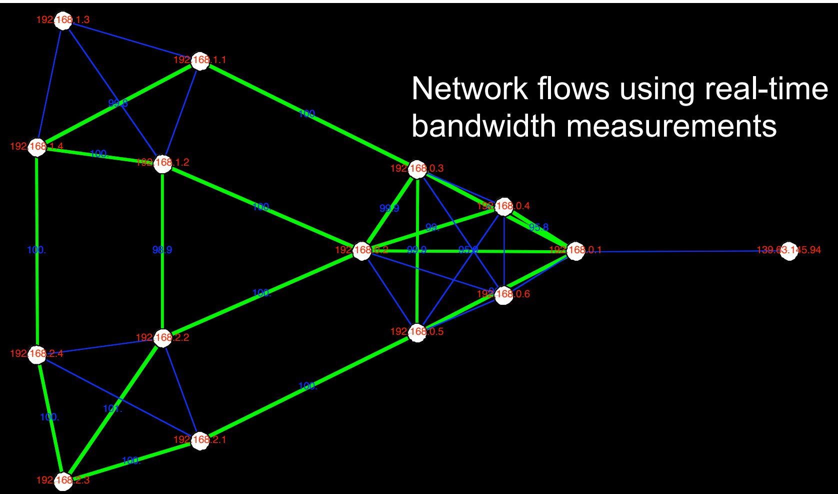


Initialization and BFS discovery of NEs

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

Available methods:

{DiscoverNetworkElements,GetLinkBandwidth,GetAllIpLinks,Remote,
NetworkTokenTransaction}

Global`upvnverbose = True;

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

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

Getting neigbours of: 139.63.145.94
Internal links: {192.168.0.1, 139.63.145.94}
(...)
Getting neigbours of:192.168.2.3
Internal links: {192.168.2.3}
```

Transaction on shortest path with tokens

```
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

TeraThinking

- What constitutes a Tb/s network?
- CALIT2 has 8000 Gigabit drops ?->? Terabit Lan?
- look at 80 core Intel processor
  - cut it in two, left and right communicate 8 TB/s
- think back to teraflop computing!
  - MPI makes it a teraflop machine
- massive parallel channels in hosts, NIC’s
- TeraApps programming model supported by
  - TFlops -> MPI / Globus
  - TBytes -> OGSA/DAIS
  - TPixels -> SAGE
  - TSensors -> LOFAR, LHC, LOOKING, CineGrid, ...
  - Tbit/s -> ?

ref Larry Smarr & CdL

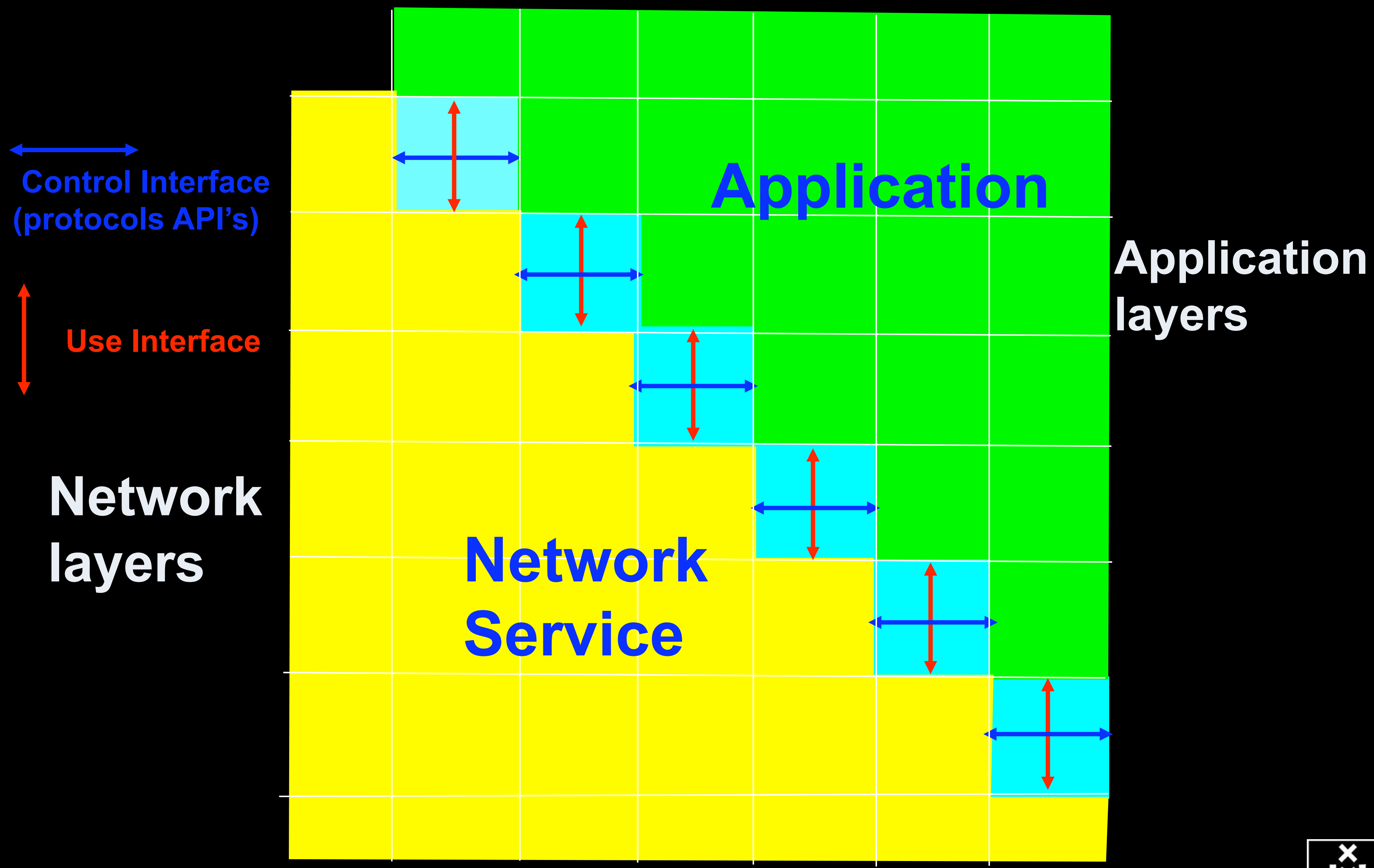
TouchTable Demonstration @ SC08

Interactive programmable networks

Need for discrete parallelism

- it takes a core to receive 1 or 10 Gbit/s in a computer
- it takes one or two cores to deal with 10 Gbit/s storage
- same for Gigapixels
- same for 100’s of Gflops
- Capacity of every part in a system seems of same scale
- look at 80 core Intel processor
  - cut it in two, left and right communicate 8 TB/s
- massive parallel channels in hosts, NIC’s
- Therefore we need to go massively parallel allocating complete parts for the problem at hand!

Multi Layer Service Architecture



Questions ?

Accepted paper: A Declarative Approach to Multi-Layer Path Finding Based on Semantic Network Descriptions.

Not on the memory stick, so:

http://delaat.net:/~delaat/papers/declarative_path_finding.pdf

Thanks: Paola Grosso & Jeroen vd Ham & Freek Dijkstra & team for several of the slides.