

Challenges for enabling Grid Computing over Optical Networks

Cees de Laat

SURFnet

EU

BSIK

NWO

University of Amsterdam

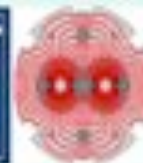
TNO
NCF



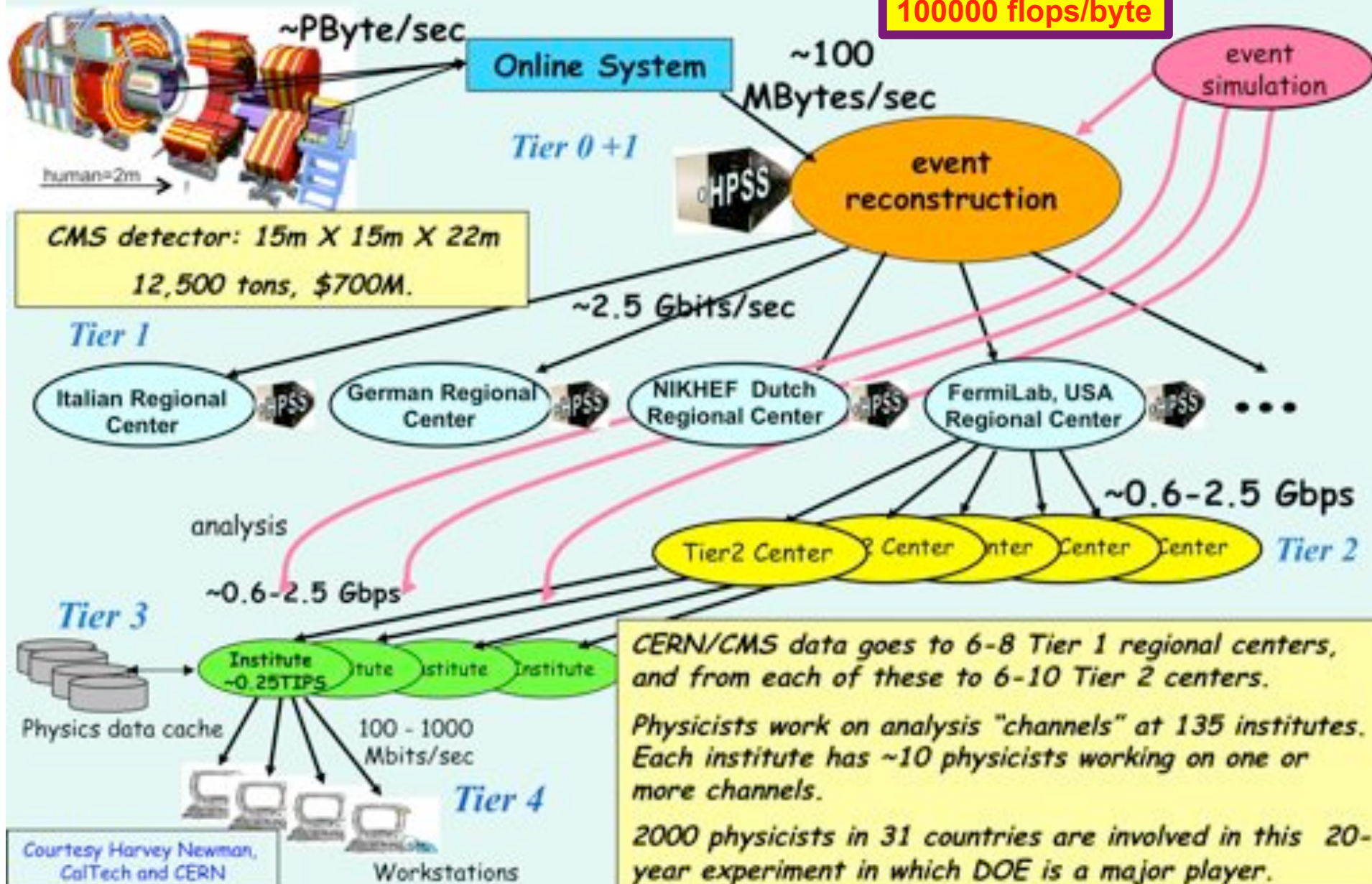


LHC Data Grid Hierarchy

CMS as example, Atlas is similar



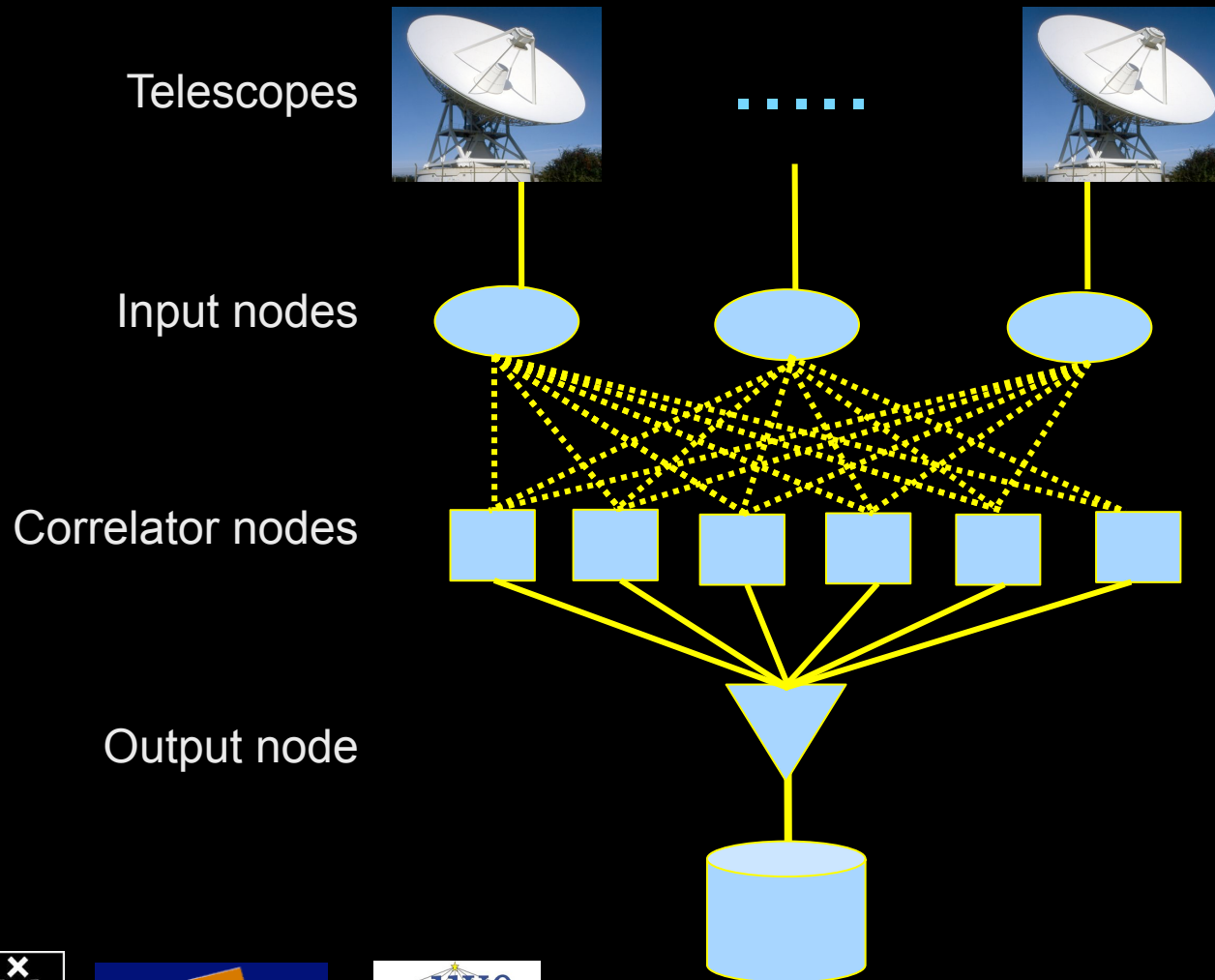
100000 flops/byte



Courtesy Harvey Newman, CalTech and CERN

The SCARIE project

SCARIE: a research project to create a Software Correlator for e-VLBI.
VLBI Correlation: signal processing technique to get high precision image from spatially distributed radio-telescope.



To equal the hardware correlator we need:

16 streams of 1Gbps

16 * 1Gbps of data

2 Tflops CPU power

2 TFlop / 16 Gbps =

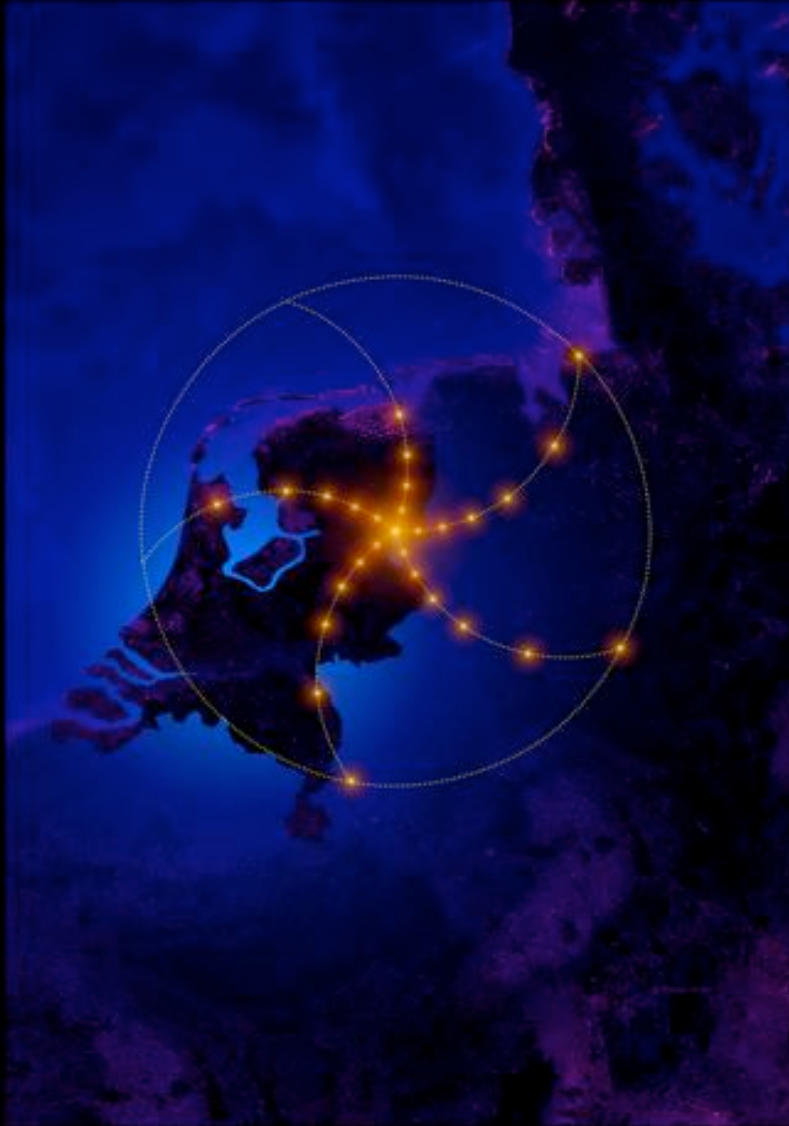
1000 flops/byte

THIS IS A DATA FLOW PROBLEM !!!



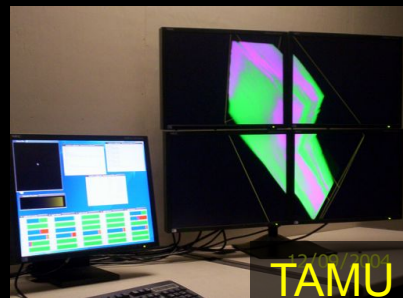
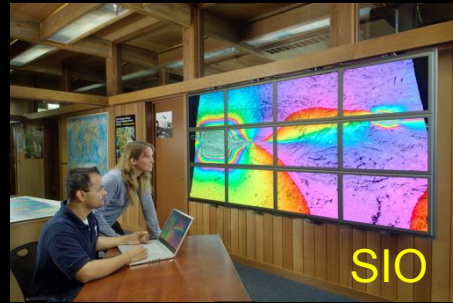
LOFAR as a Sensor Network

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

US and International OptIPortal Sites



Real time, multiple 10 Gb/s



The "Dead Cat" demo

1000000 flops/byte



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. Slood 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

Many small flows



Sensor grid: instrument the dikes

First controlled breach occurred on sept 27th '08:

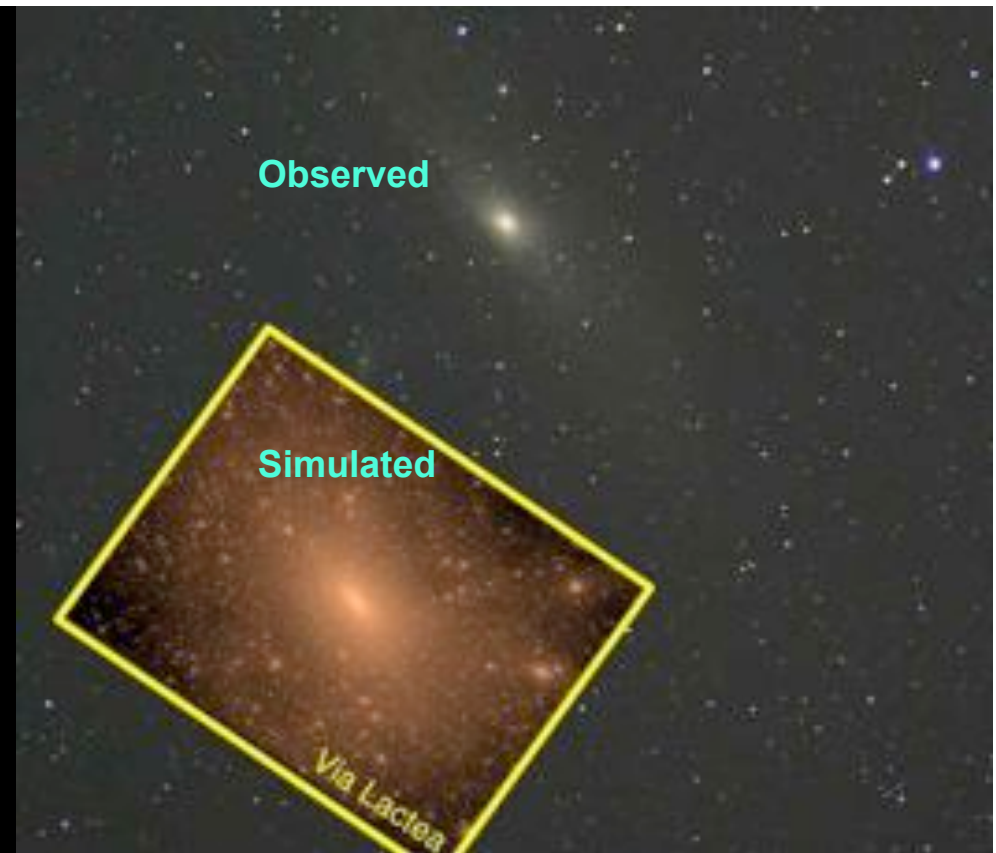


300000 * 60 kb/s * 2 sensors (microphones) to cover all Dutch dikes



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

10⁷

- 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



7.6 Gb/s

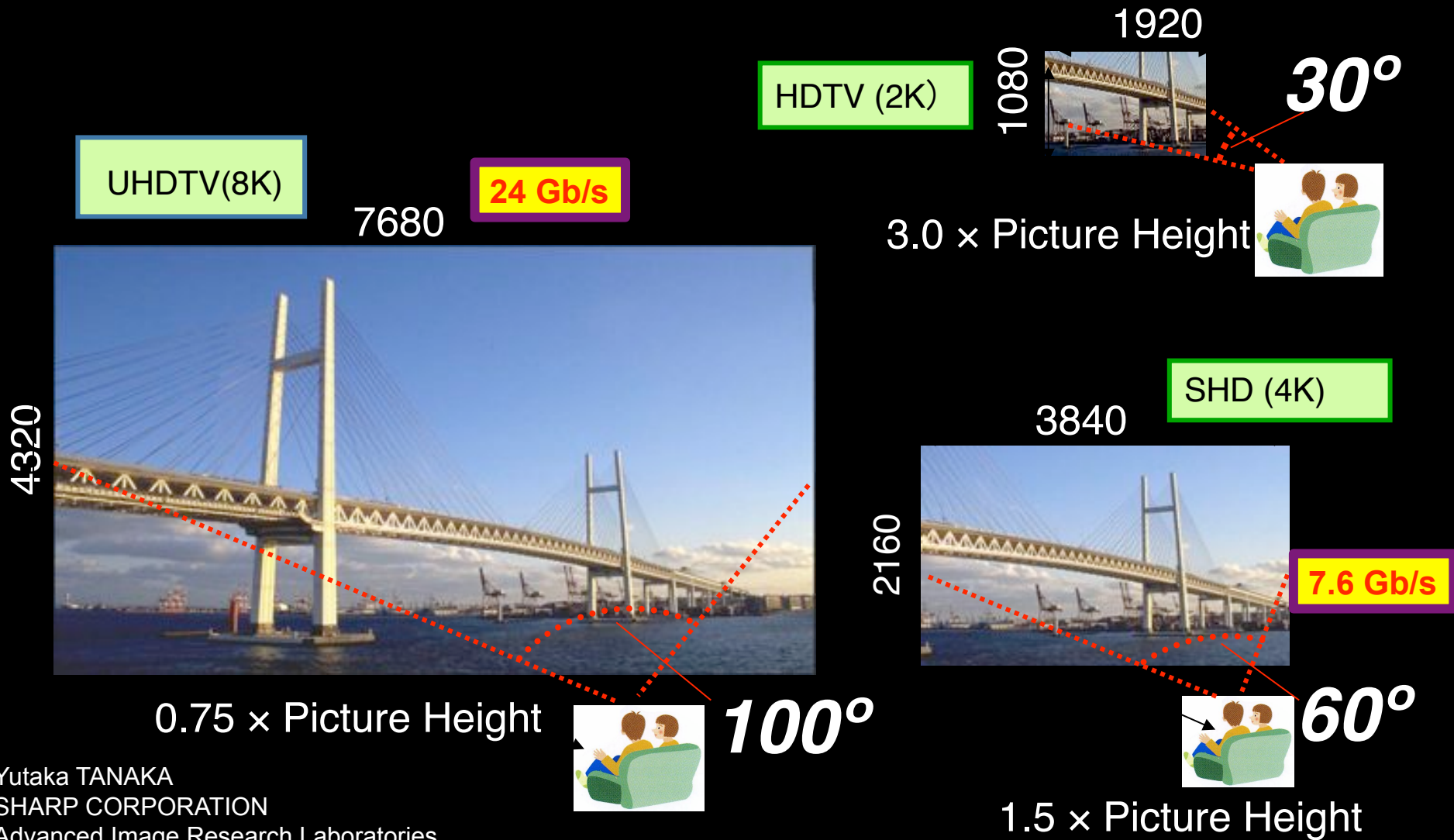


CineGrid @ Holland Festival 2007



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



CineGrid portal

100 Tbyte
Cache & Store & Forward



CineGrid distribution center Amsterdam

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

Amsterdam Node Status:

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

Search node:

Search

Browse by tag:

amsterdam animation
[antonacci](#) blender boat
bridge bunny cgi delta holland
hollandfestival
leidschestraat
muziekgebouw
nieuwmarkt opera prague ship
train tram trans waag

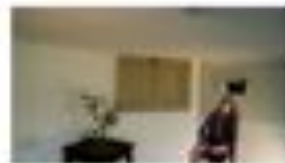
via licensed under Attribution-NonCommercial-ShareAlike

CineGrid Amsterdam

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

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

Latest Additions



Wypke

Wypke

Available formats:

4k drc (4.0 KB)
Duration: 1 hour and 8 minutes
Created: 1 week, 2 days ago
Author: Wypke
Categories:



Prague Train

Steam locomotive in Prague

Available formats:

4k drc (3.9 KB)
Duration: 27 hours and 46 minutes
Created: 1 week, 2 days ago
Author: CineGrid
Categories: delta prague train



VLC: Big Buck Bunny

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

Available formats:

1080p HPEG4 (1.1 GB)
Duration: 1 hour and 0 minutes
Created: 1 month, 1 week ago
Author: Blender Foundation
Categories: animation blender bunny
cgi

U
S
E
R
S

A. Lightweight users, browsing, mailing, home use

Need full Internet routing, one to all

B. Business/grid applications, multicast, streaming, VO's, mostly LAN

Need VPN services and full Internet routing, several to several + unlink to all



■ Input ■ Output
Peak In : 641.166 Gb/s Peak Out : 639.212 Gb/s
Average In : 419.749 Gb/s Average Out : 419.612 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)

BW

GigE

Ref: Cees de Laat, Erik Radius, Steven Wallace, "The Rationale of the Current Optical Networking Initiatives"
iGrid2002 special issue, Future Generation Computer Systems, volume 19 issue 6 (2003)



Towards Hybrid Networking!

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

L1 \approx 2-3 k\$/port



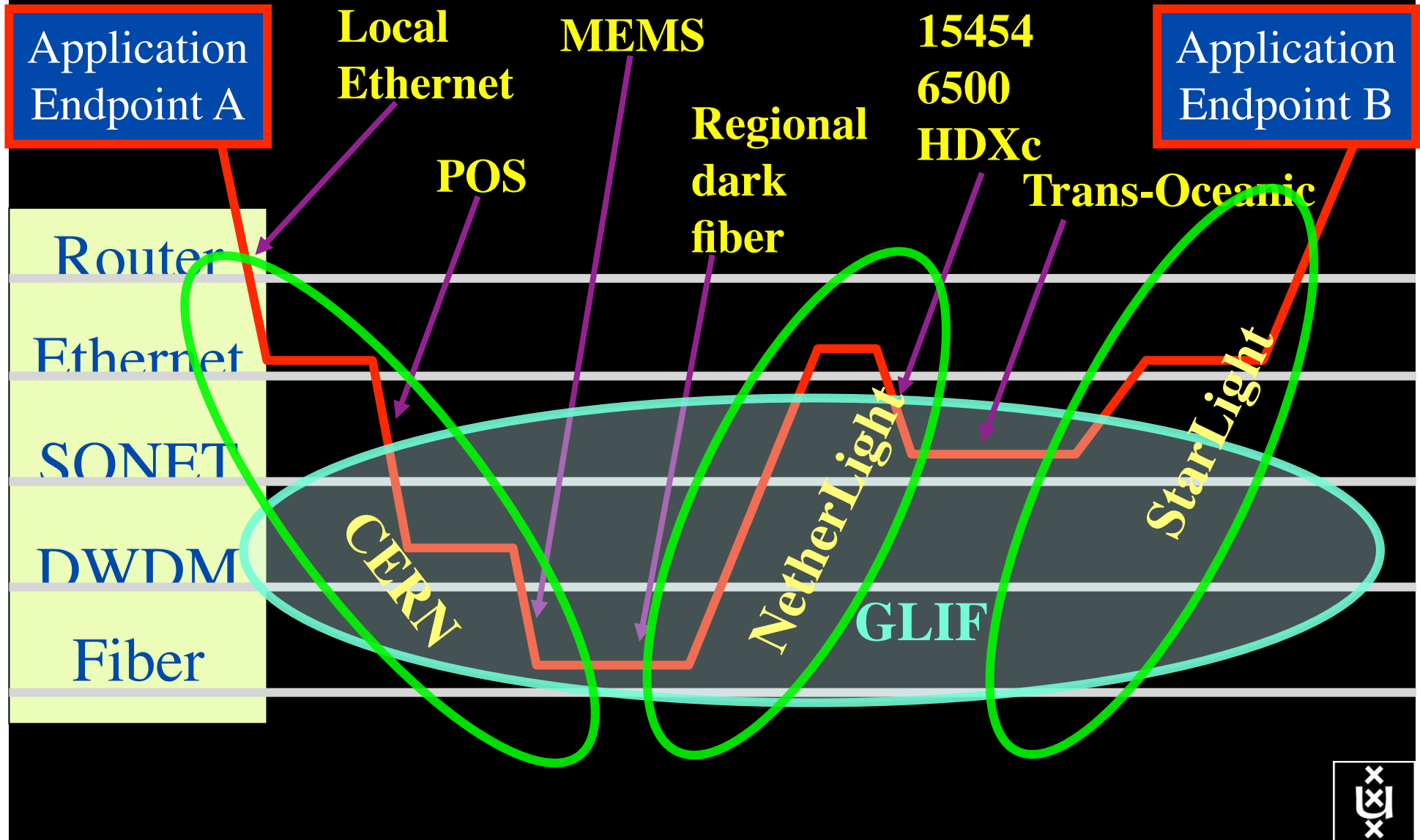
L2 \approx 5-8 k\$/port



L3 \approx 75+ k\$/port



How low can you go?

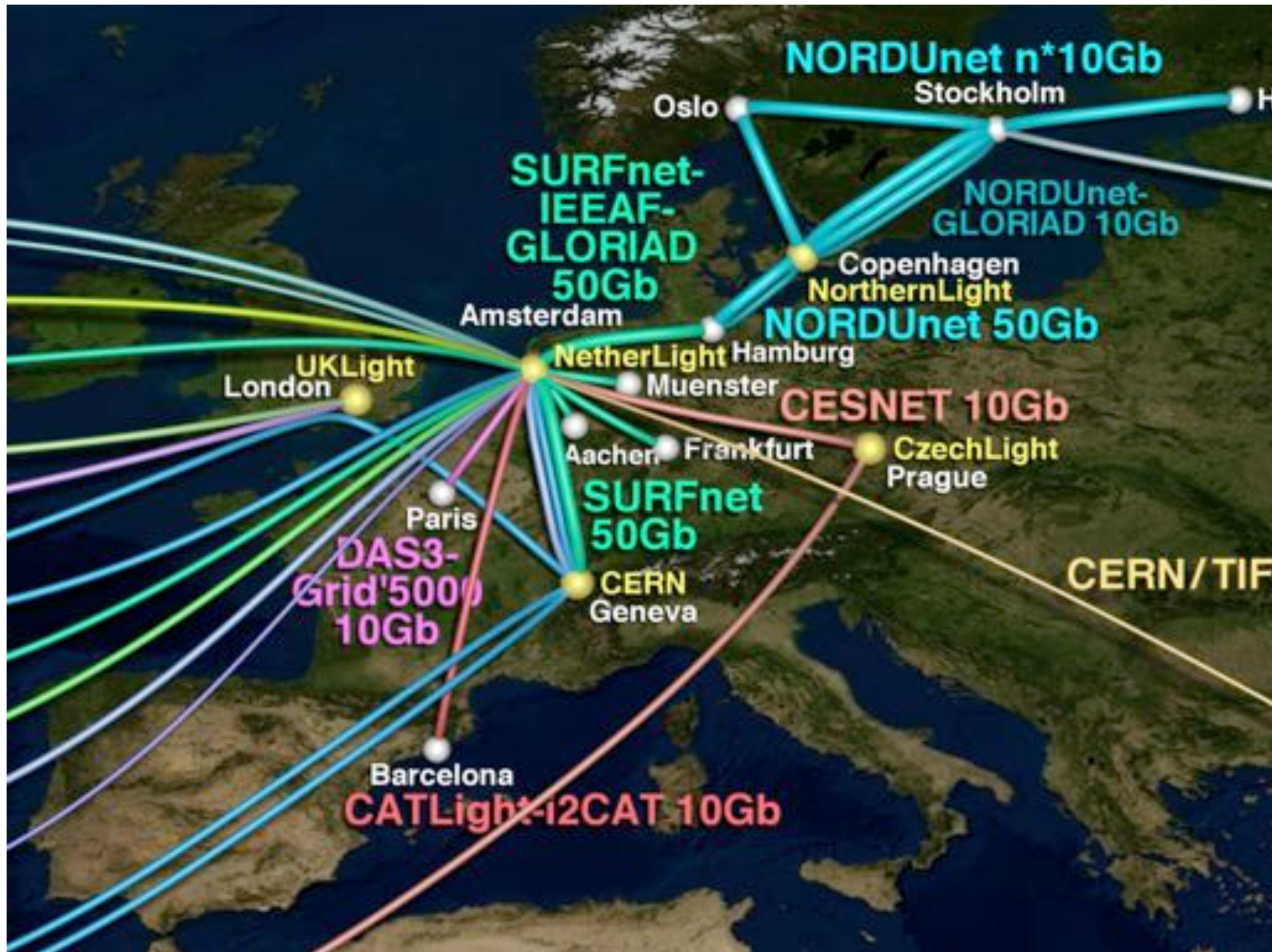




GLIF 2008

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





•VIZ

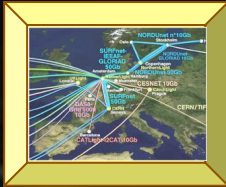
DataExploration

RemoteControl

TV

Medical

CineGrid



Gaming

Conference

Workflow

Clouds



Distributed

EventProcessing

•GRID

Management

Mining

Web2.0



Meta

•NetherLight

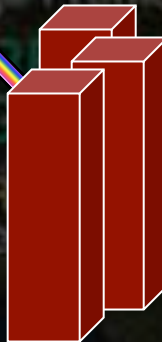
•DATA

Backup

Media

Visualisation

Security

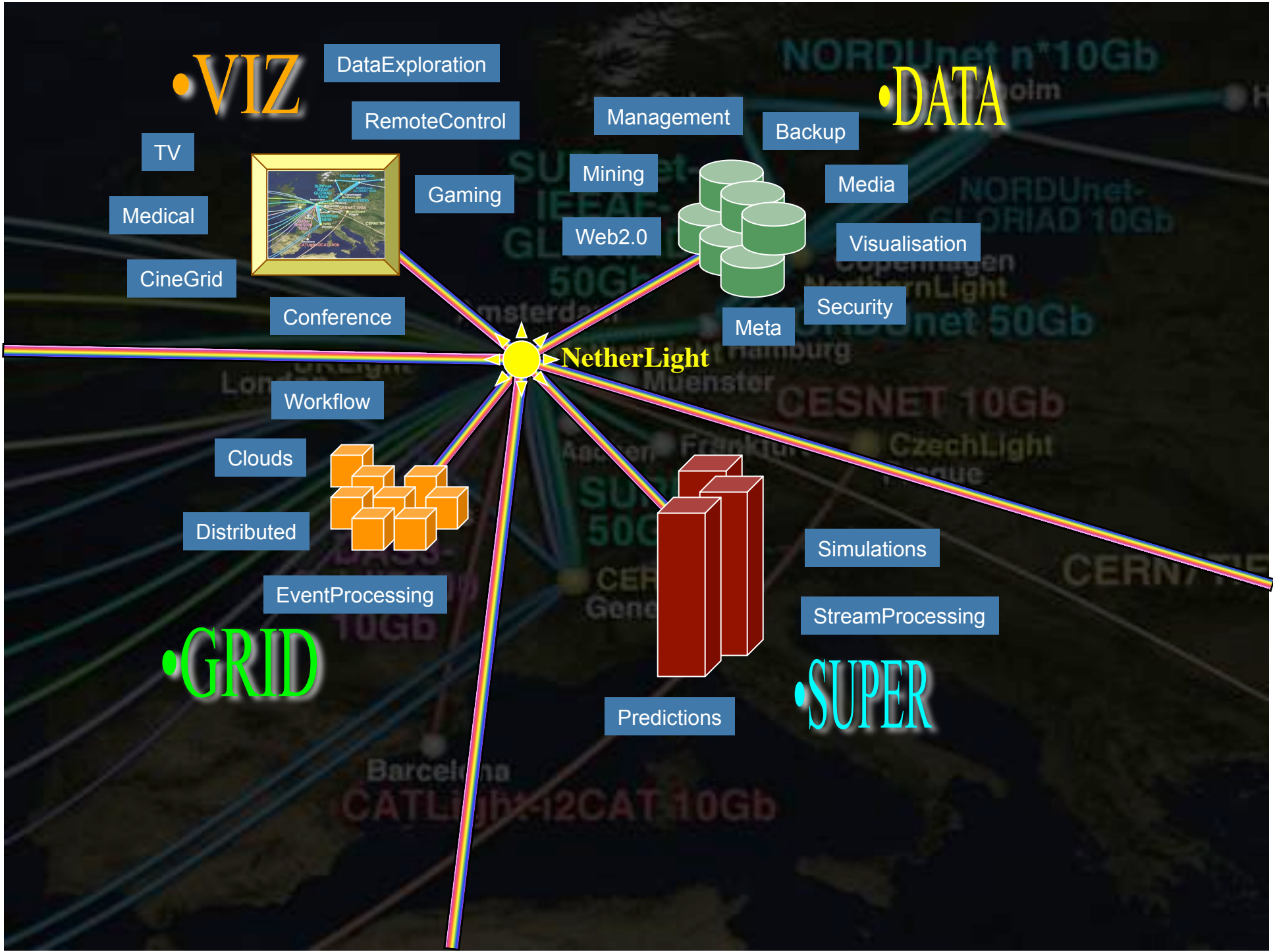


Simulations

StreamProcessing

Predictions

•SUPER



SURFnet

In The Netherlands SURFnet connects between 180:

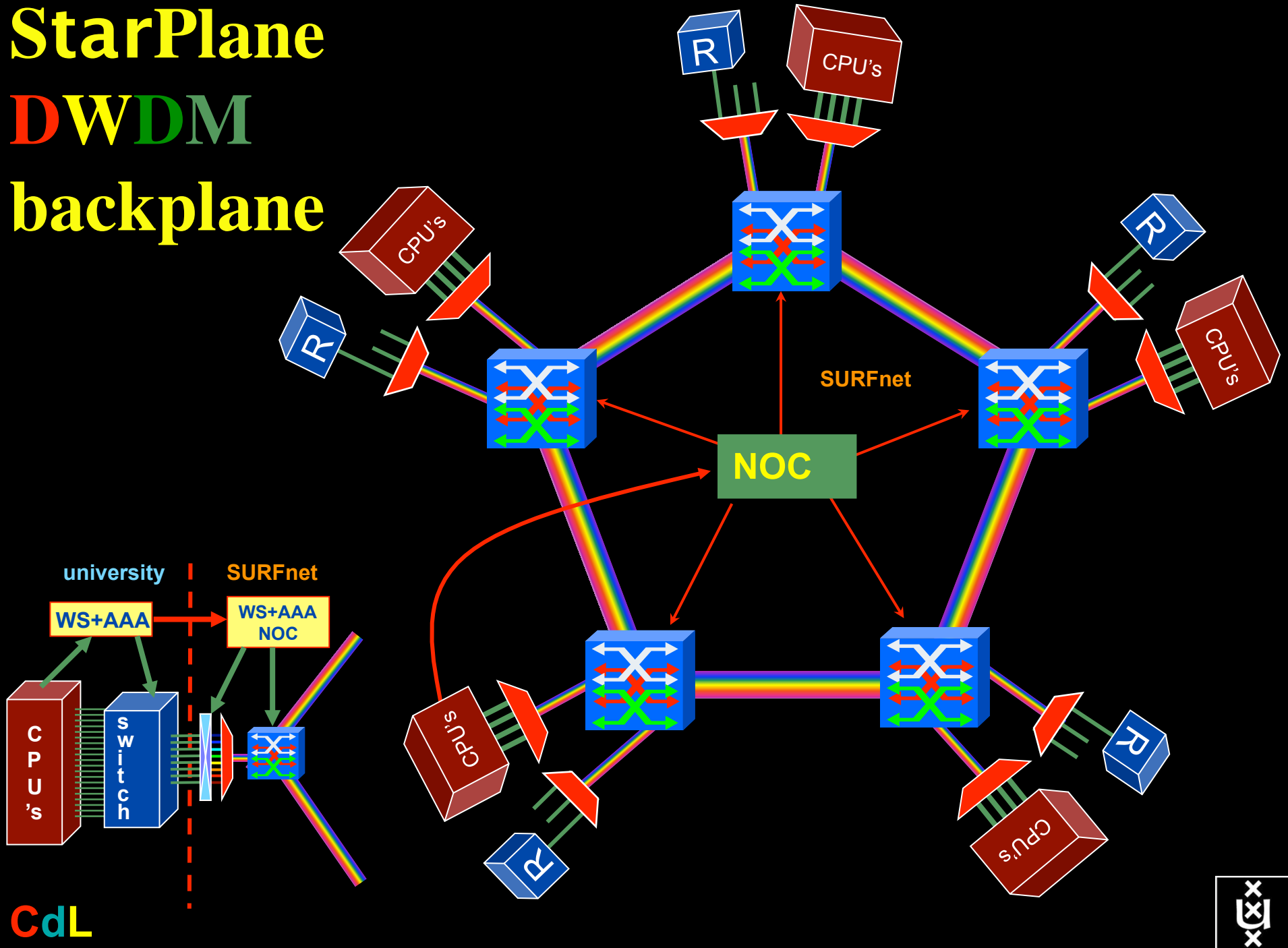
- universities;
- academic hospitals;
- most polytechnics;
- research centers.

with an indirect ~750K user base

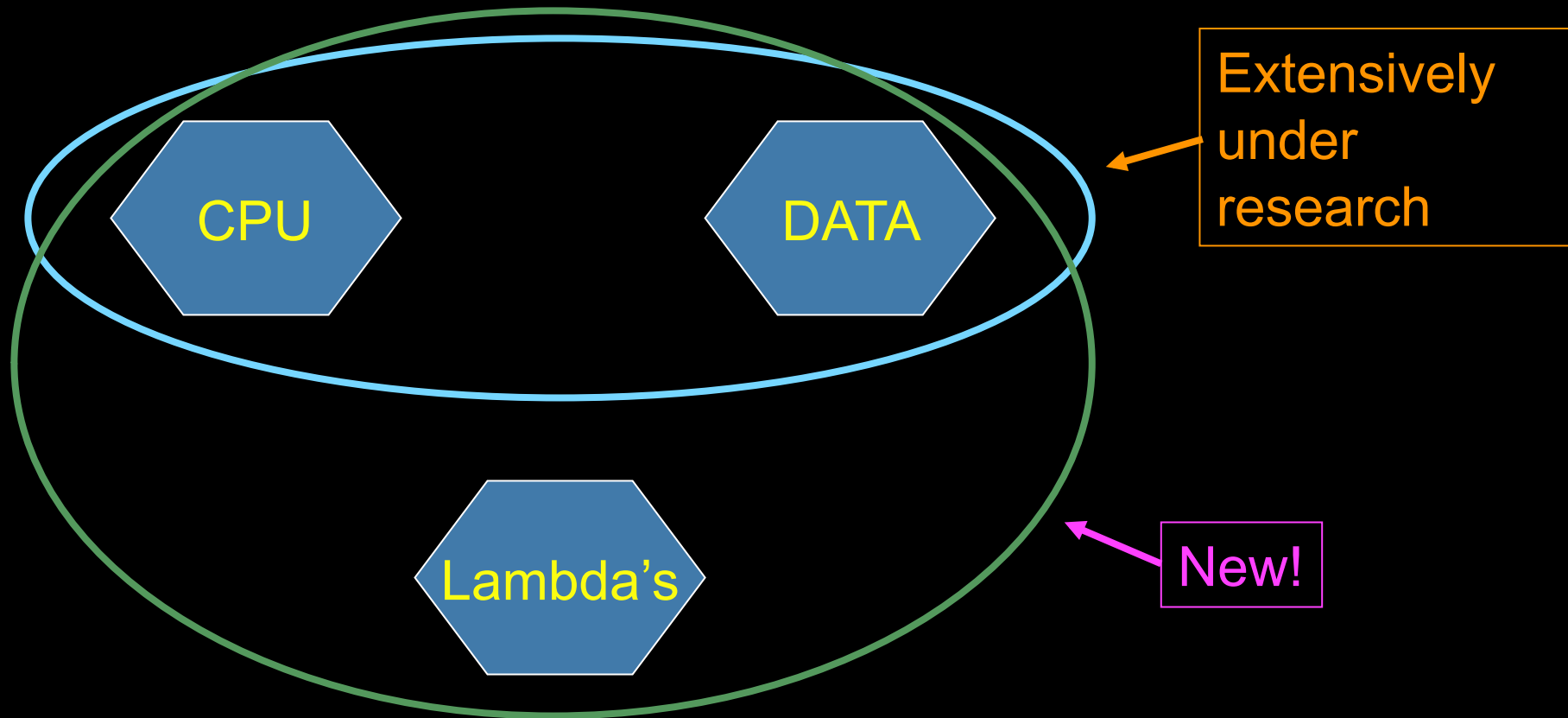
~ 8860 km
scale
comparable
to railway
system



StarPlane DWDM backplane



GRID Co-scheduling problem space

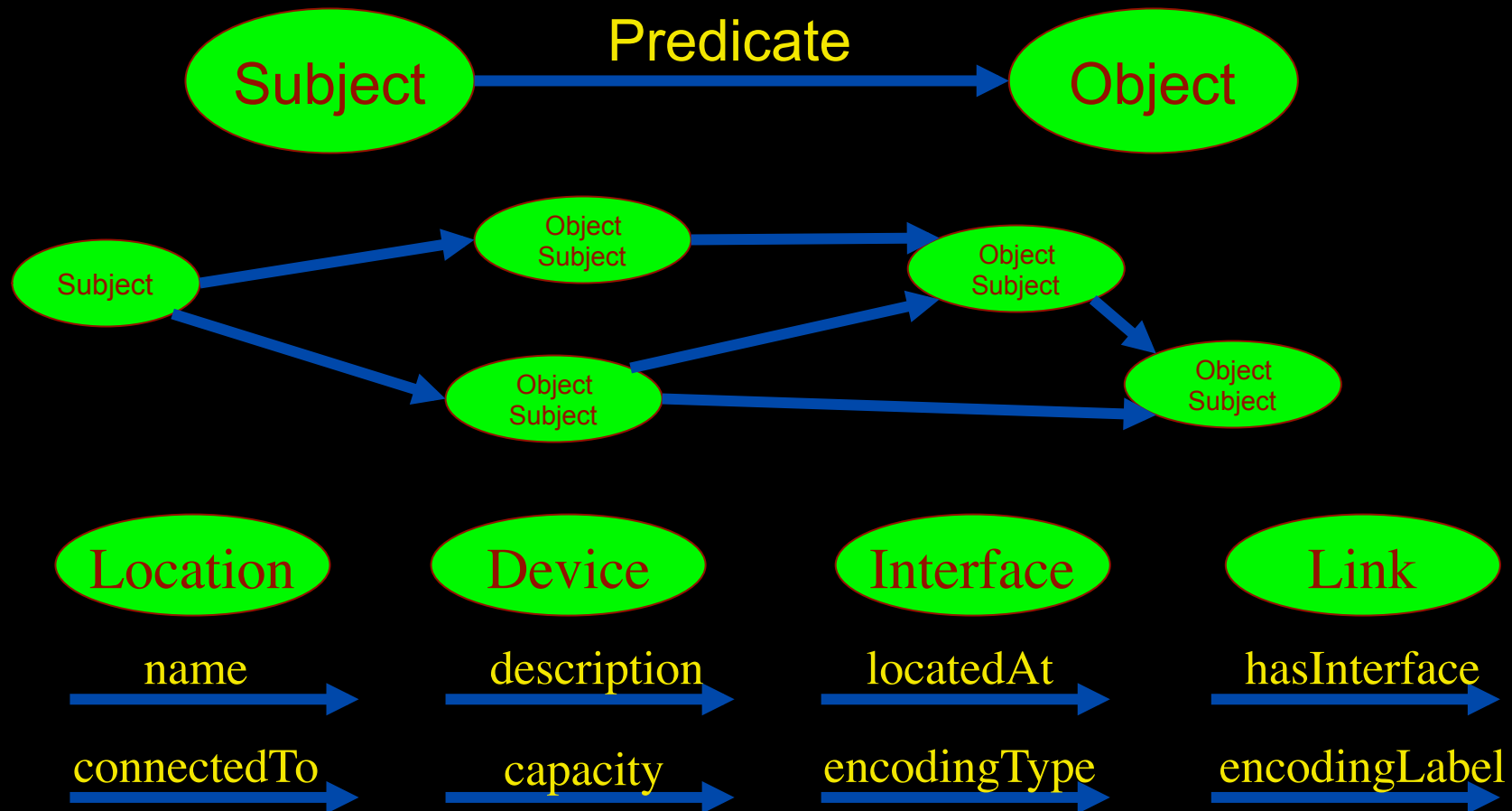


The StarPlane vision is to give flexibility directly to the applications by allowing them to choose the logical topology in real time, ultimately with sub-second lambda switching times on part of the SURFnet6 infrastructure.



Network Description Language

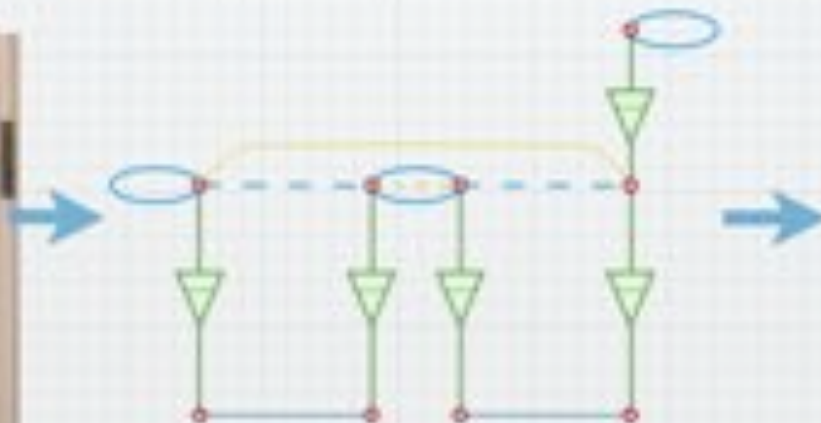
- From semantic Web / Resource Description Framework.
- The RDF uses XML as an interchange syntax.
- Data is described by triplets:



Network Description Language

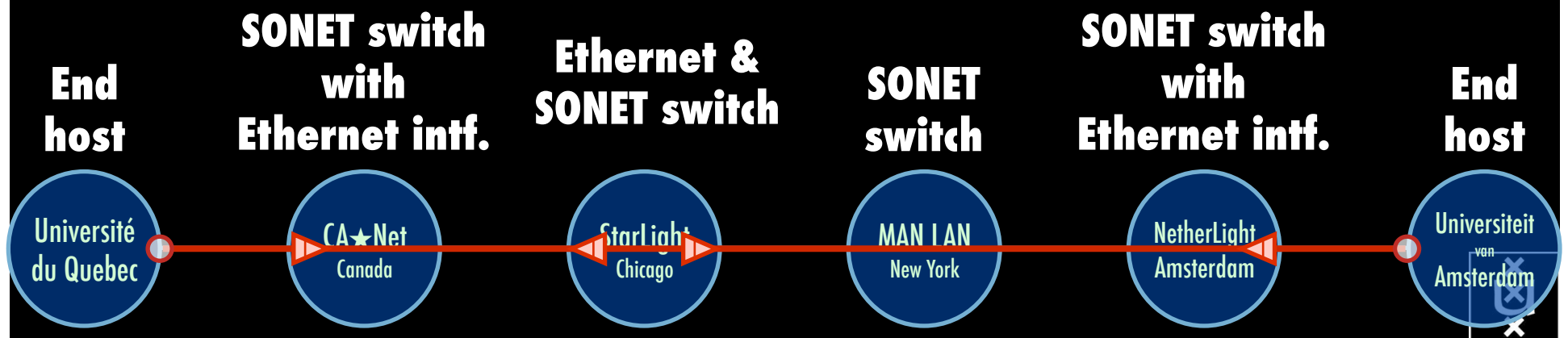
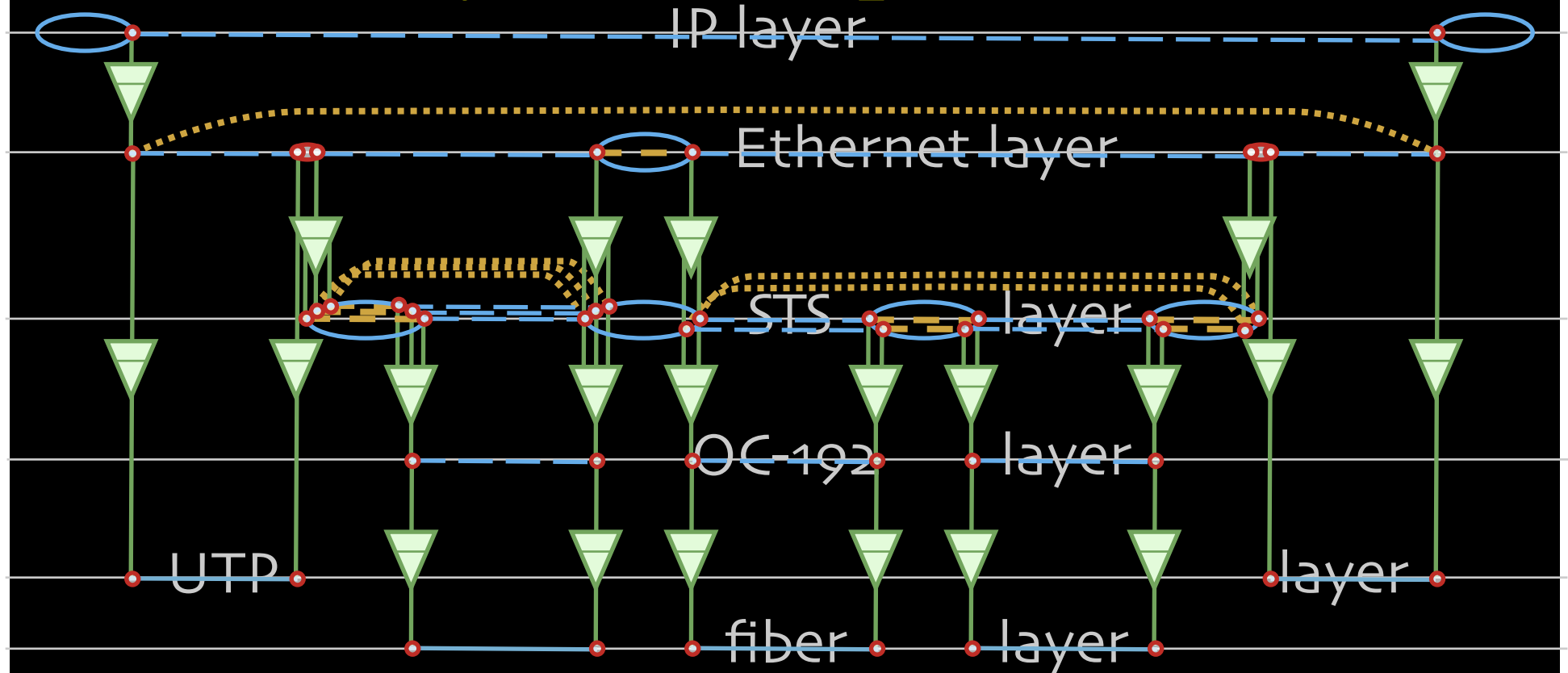
Choice of RDF instead of flat XML descriptions
Grounded modeling based on G0805 description:

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"



```
<nd:Device rdf:about="#Force10">
  <nd:hasInterface rdf:resource=
    "#Force10/eth/0"/>
</nd:Device>
<nd:Interface rdf:about="#Force10/eth/0">
  <nd:label="#eth/0">
  <nd:capacity=12588</nd:capacity>
  <nd:conf:multiplex>
  <nd:cap:adaptation rdf:resource=
    "#Tagged-Ethernet-in-Ethernet"/>
  <nd:conf:serverPropertyValue
    rdf:resource="#MTU-1500byte"/>
</nd:conf:multiplex>
  <nd:conf:hasChannels>
  <nd:conf:Channel rdf:about=
    "#Force10/eth/0/vlan1">
    <nd:eth:hasVlan=4</nd:eth:hasVlan>
    <nd:conf:switchedTo rdf:resource=
      "#Force10/g1/1/vlan7"/>
  </nd:conf:Channel>
</nd:conf:hasChannels>
</nd:Interface>
```


Multi-layer descriptions in NDL

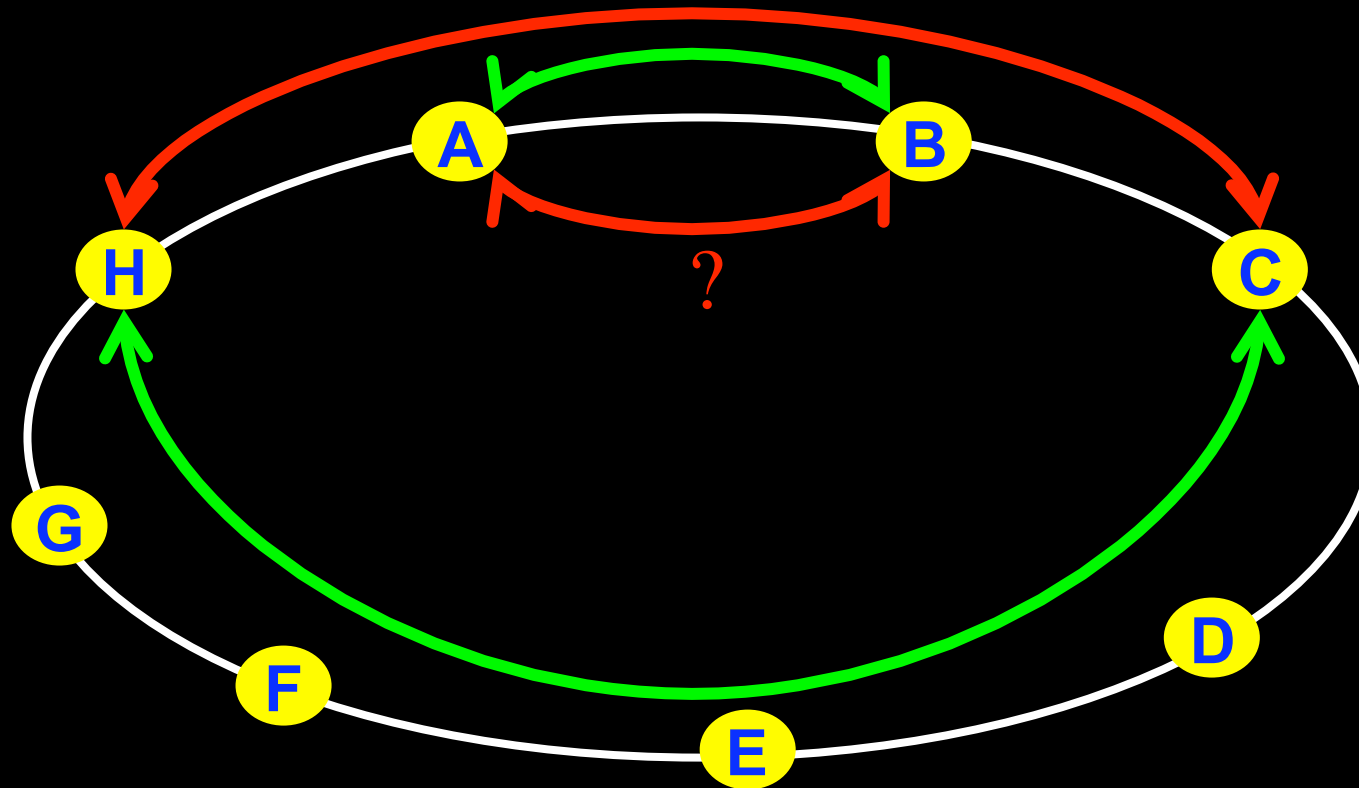


The Problem

I want HC and AB

Success depends on the order

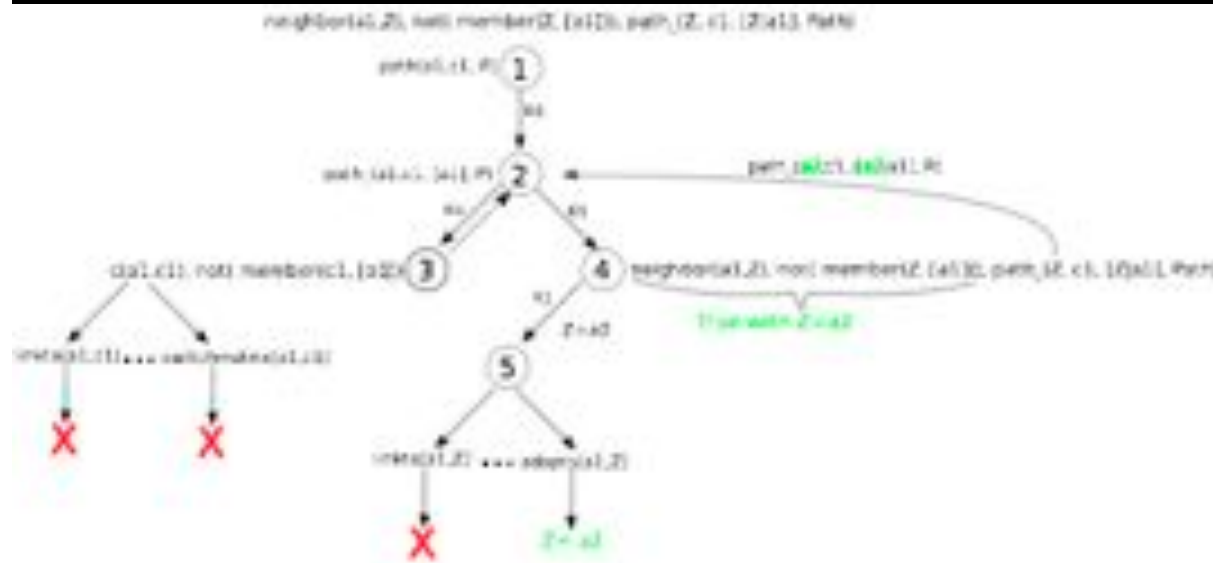
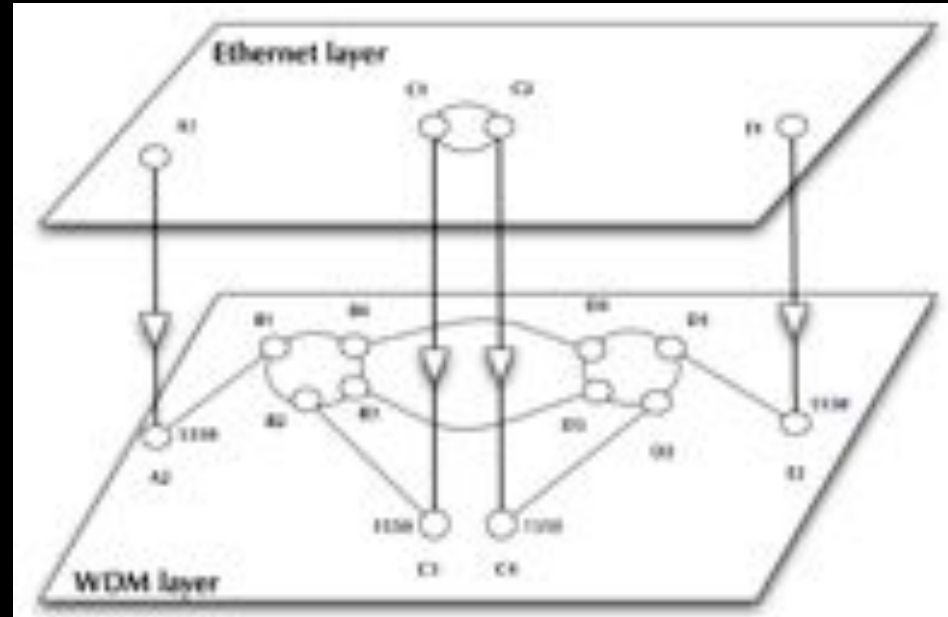
Wouldn't it be nice if I could request [HC, AB, ...]



NDL + PROLOG

Research Questions:

- order of requests
- complex requests
- usable leftovers



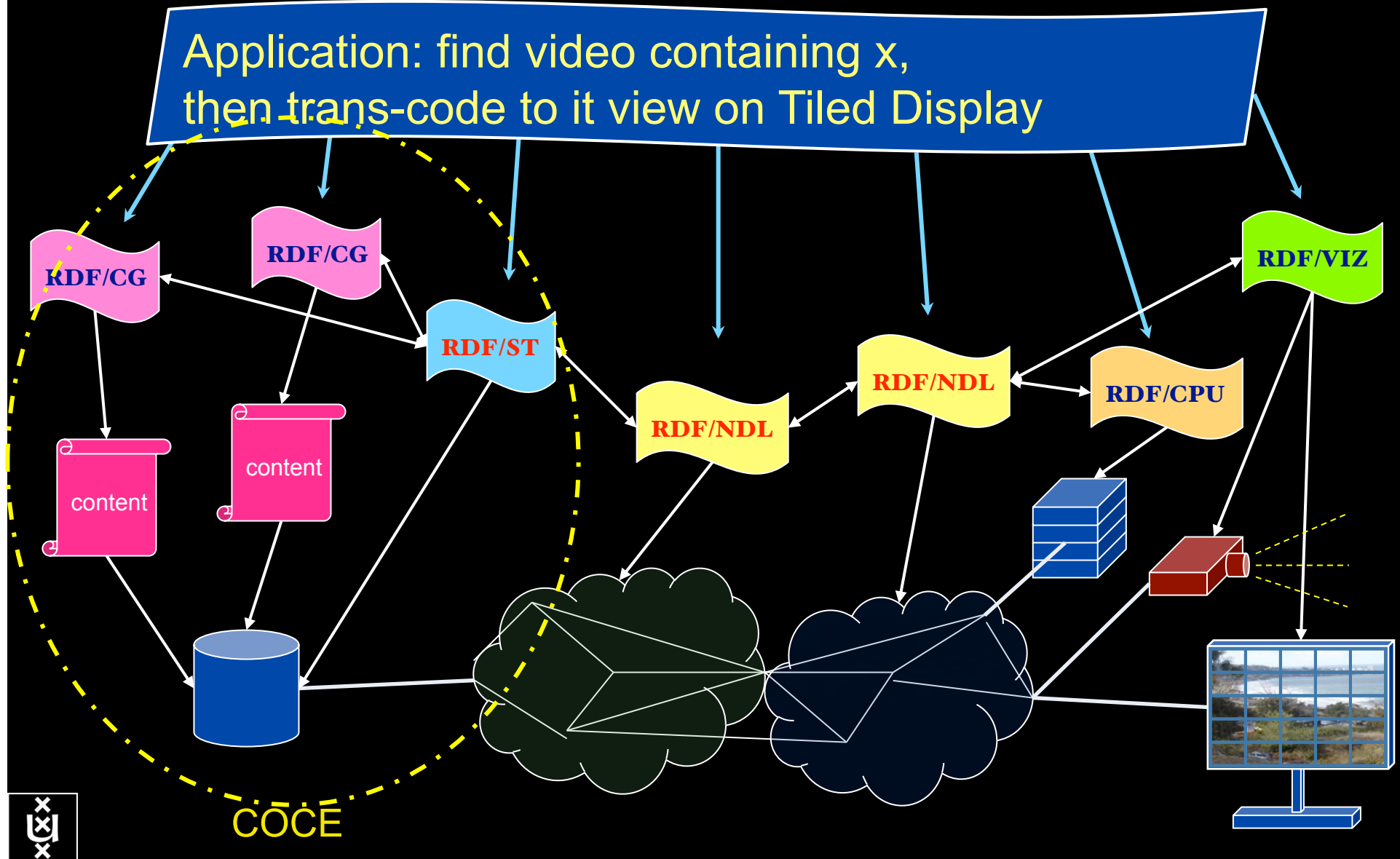
•Reason about graphs

•Find sub-graphs that comply with rules



RDF describing Infrastructure “I want”

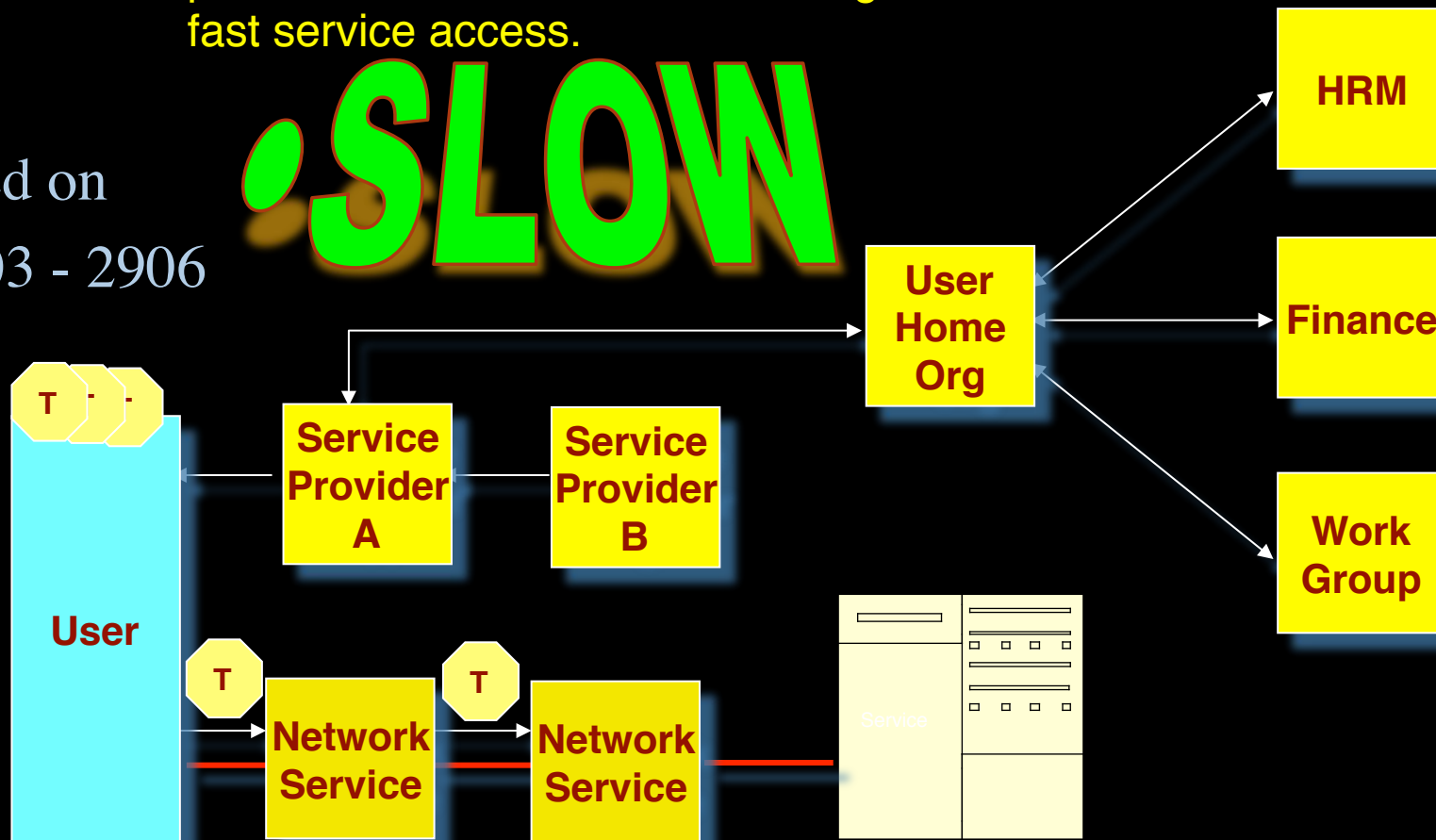
Application: find video containing x,
then trans-code to it view on Tiled Display



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

Based on
RFC 2903 - 2906

SLOW



Fast

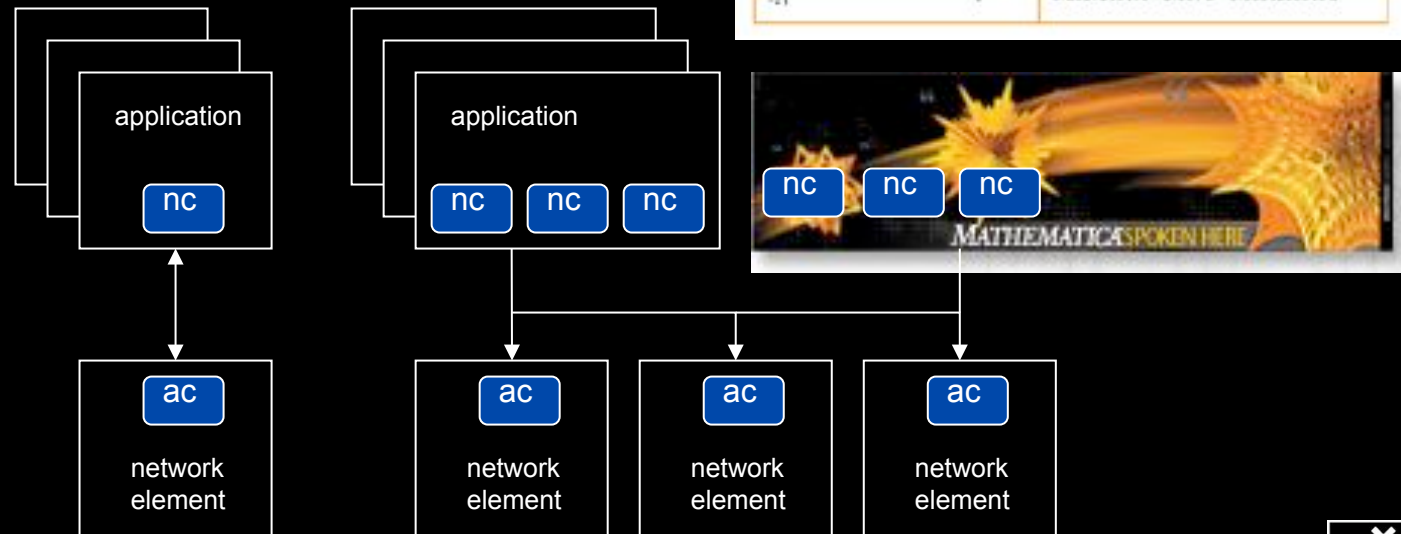
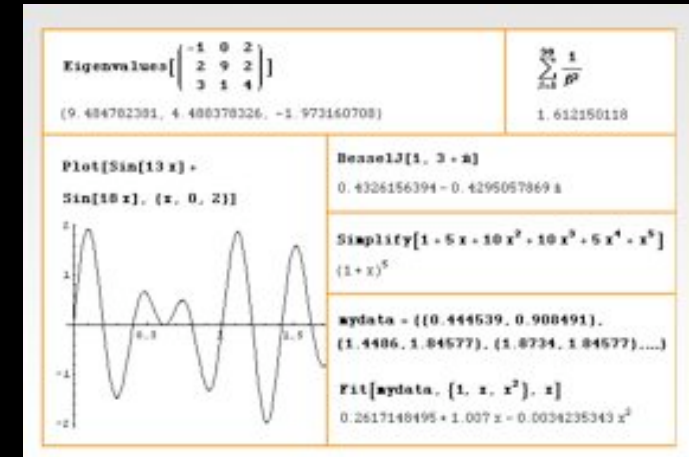
TeraThinking

- 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 turns a room full of pc's in 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 -> ?



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, GetAllpLinks, Remote,
NetworkTokenTransaction}
```

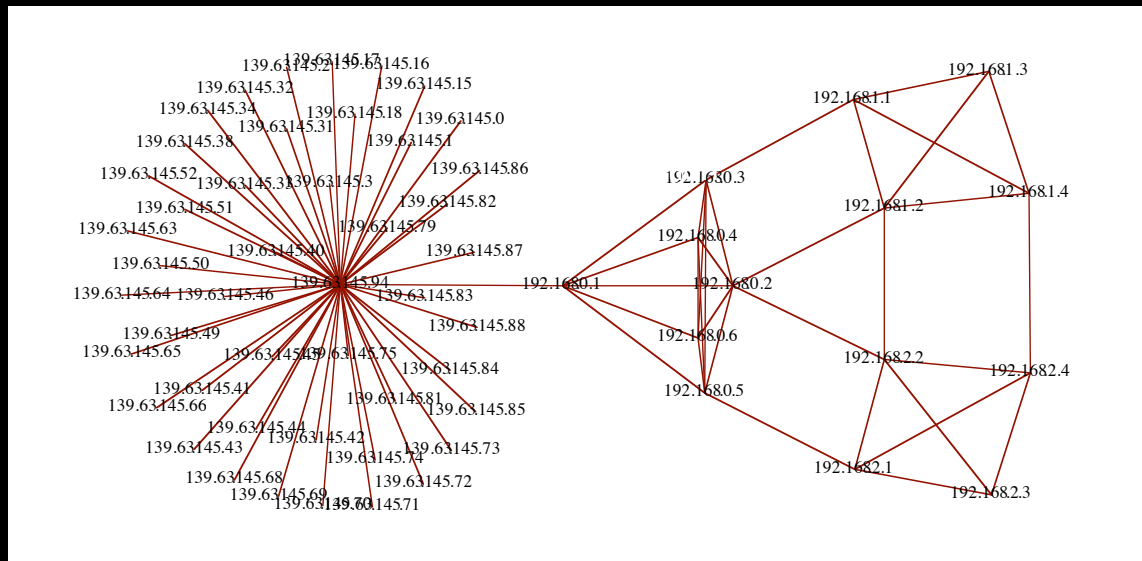
```
Global`upvnverbose = True;
```

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

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

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

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

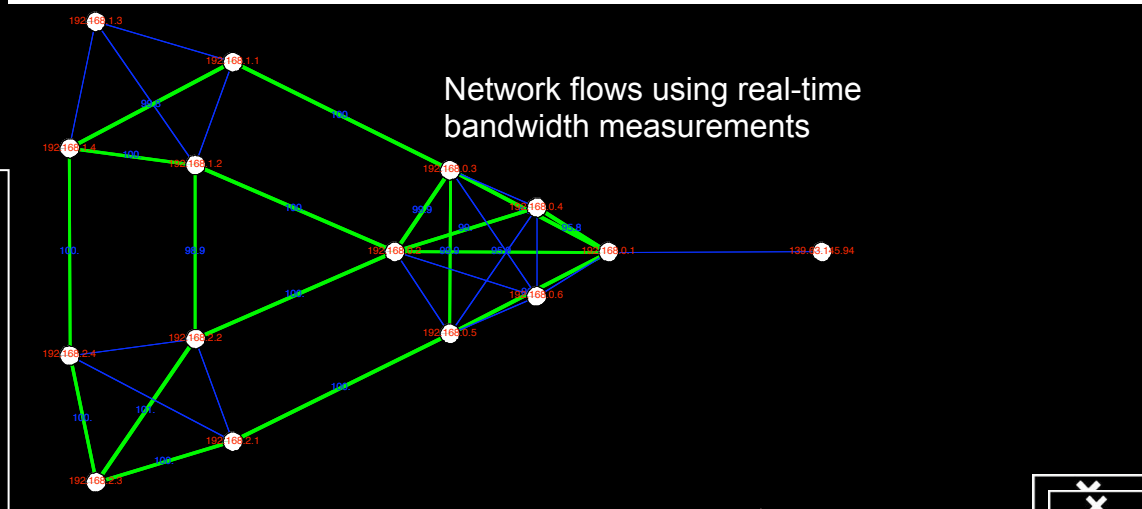


Transaction on shortest path with tokens

```
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

Interactive programmable networks



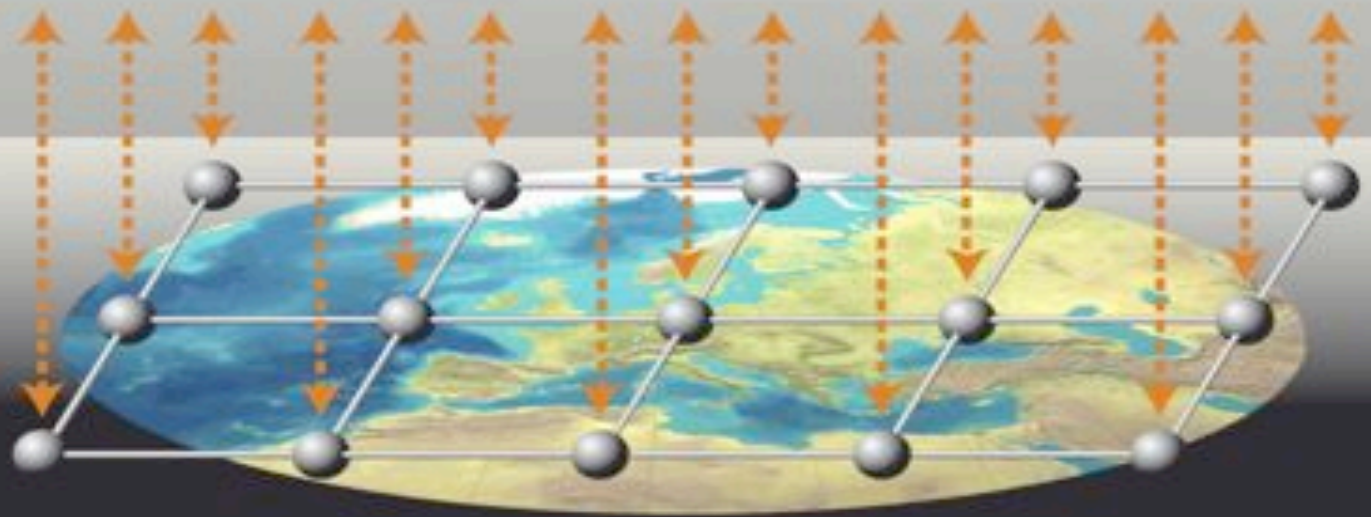
Themes

- eScience infrastructure virtualization
- Photonic networking -> Tb/s
- Capacity & Capability
- Data handling, integrity, security, privacy
- Cloud paradigm, green compute&store&net&viz
- **ENERGY dependency! (2009: 1Wy=1€)**

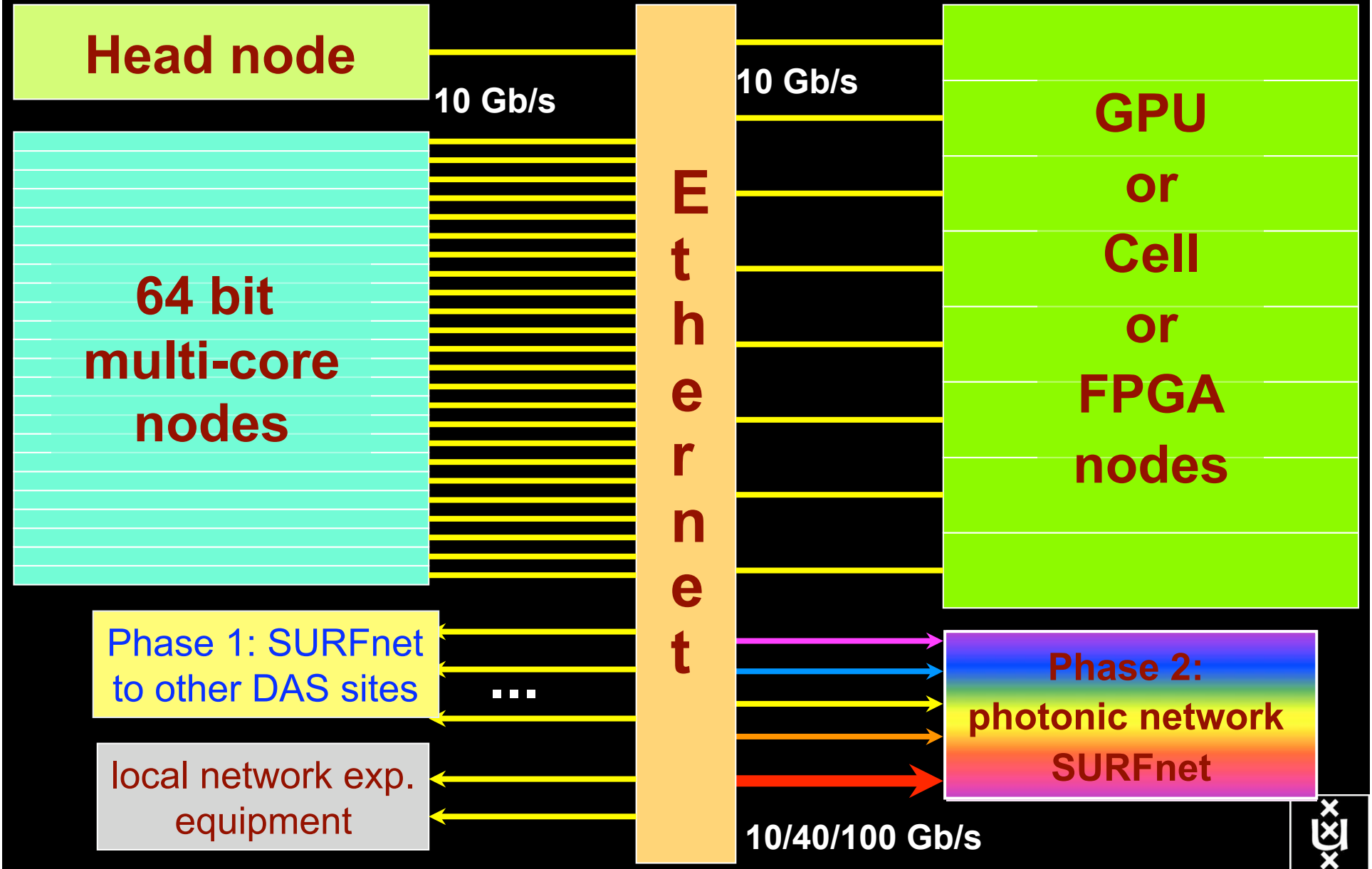


**Virtual Laboratory
generic e-Science services**

**High Performance & Distributed Computing
Web & Grid services**



DAS-4 Proposed Architecture



n.a.v. interview met Kees Neggers (SURFnet) & Cees de Laat (UvA)



- BSIK projects GigaPort &
- VL-e / e-Science



cookreport.com

ICT and E-Science as an Innovation Platform in The Netherlands

A National Research and Innovation Network

What Can the US Learn from Dutch Experience?

"The dogmas of the quiet past are inadequate to the stormy present. As our case is new, so we must think anew and act anew." Abraham Lincoln

By means of an examination of research networks in Holland, this issue presents some ideas for ways in which an American National Research, Education and Innovation Network could be structured.

possible are carried out by decentralized groups.

Volume XVII, No. 11
February 2009
ISSN 0975 - 4331

THE COOK REPORT ON INTERNET PROTOCOL

FEBRUARY 2009

The Basis for a Future Internet?

Optical Hybrid Networks and e-Science as Platforms for Innovation and Tech Transfer

Editor's Note: I continued the discussion begun on No-

slide shows our organization within the University and the

search department of KPN. He did a lot of virtualization

Questions ?

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

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

Thanks: Paola Grosso & Jeroen vd Ham & Freek

Dijkstra & team for several of the slides.

