Challenges for enabling eScience over Optical Networks

Cees de Laat

SURFnet

EU

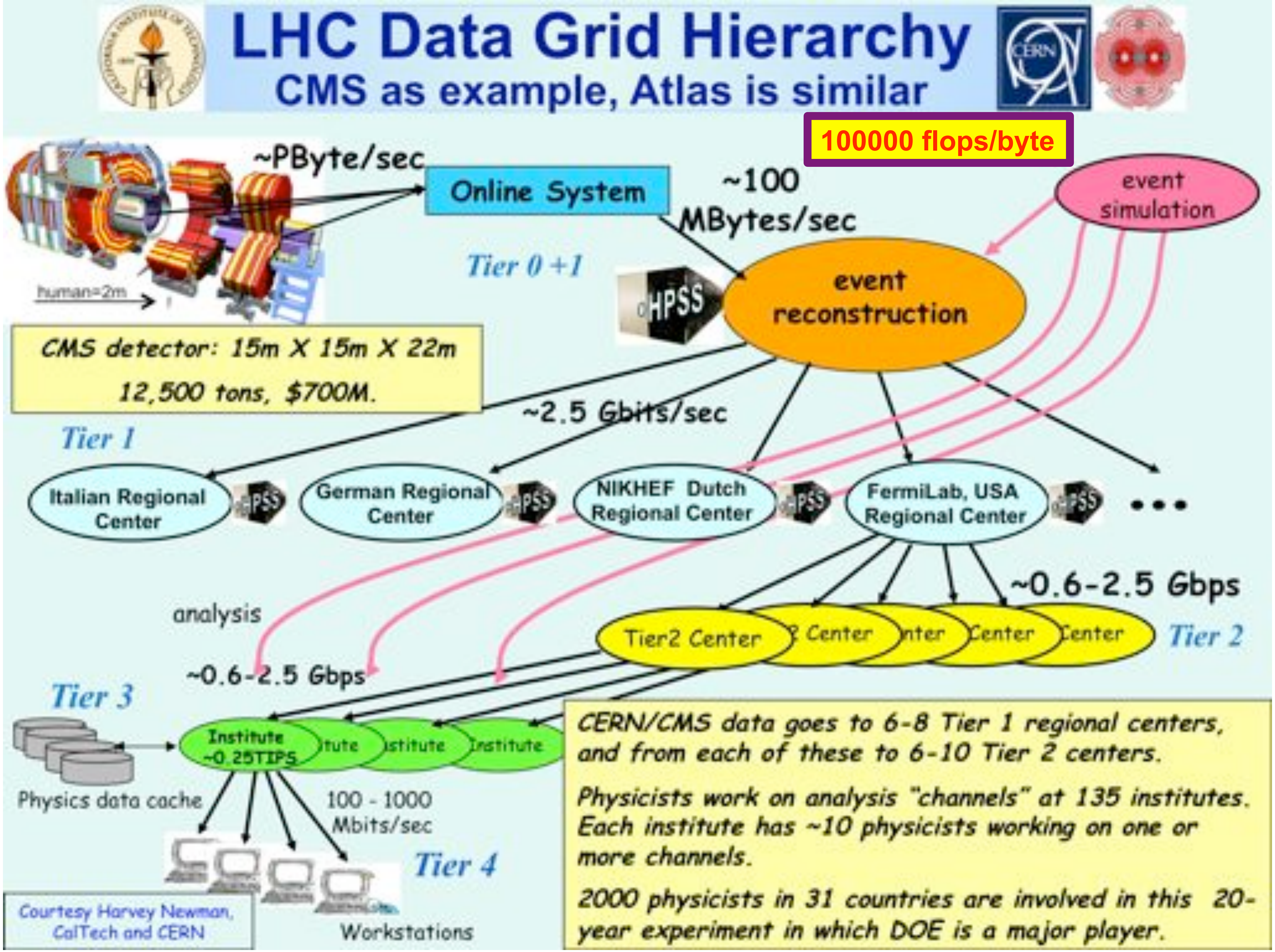
BSIK

NWO

University of Amsterdam

TNO

NCF



LHC Data Grid Hierarchy

CMS as example, Atlas is similar

100000 flops/byte

~PByte/sec

Online System

~100 MBytes/sec

event simulation

Tier 0 +1

HPSS

event reconstruction

human=2m

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

2 Center

nter

Center

Center

Tier 2

~0.6-2.5 Gbps

Tier 3

Institute ~0.25TIPS

itute

stitute

Institute

Physics data cache

100 - 1000 Mbits/sec

Tier 4

Workstations

Courtesy Harvey Newman, CalTech and CERN

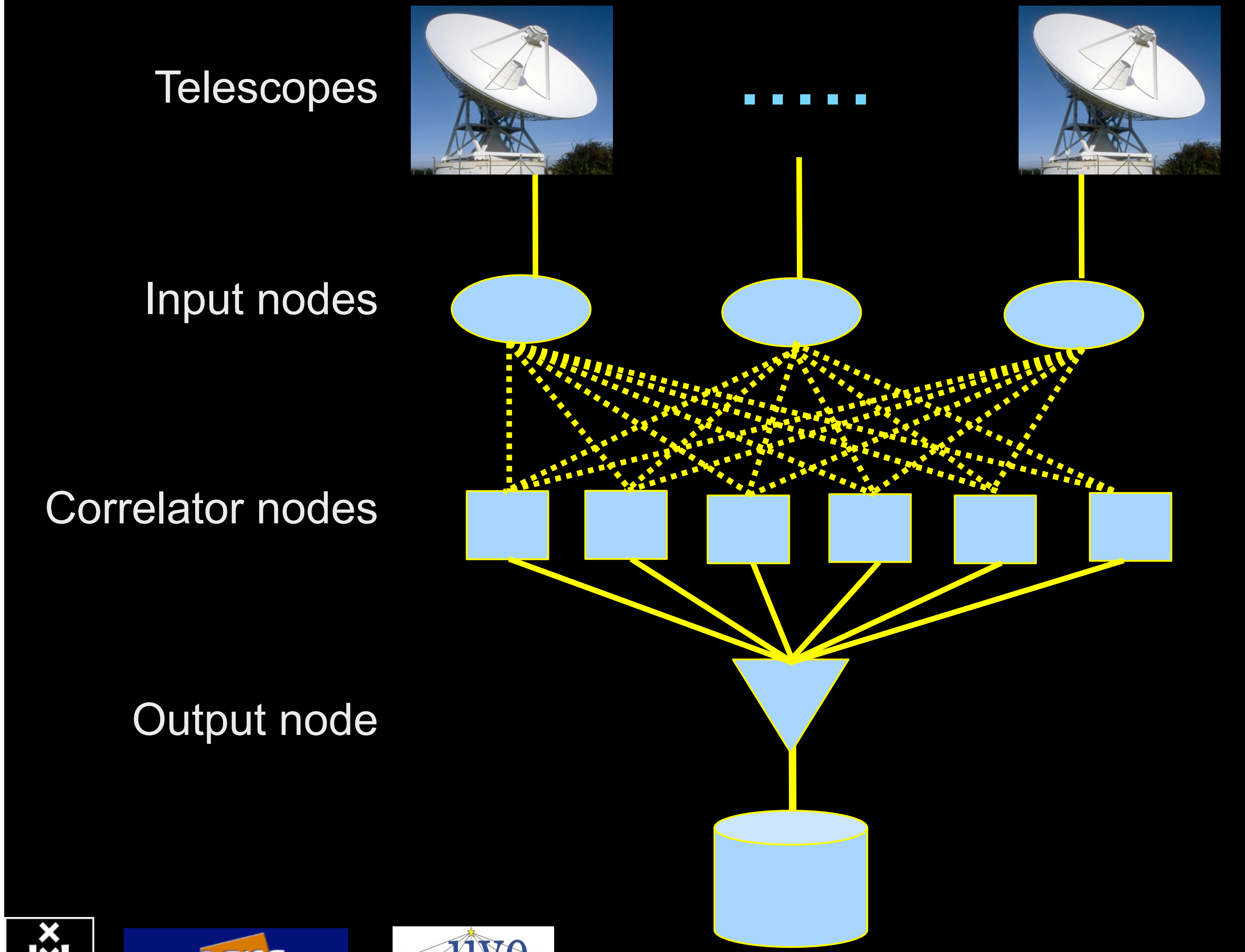
CERN/CMS data goes to 6-8 Tier 1 regional centers, and from each of these to 6-10 Tier 2 centers.

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

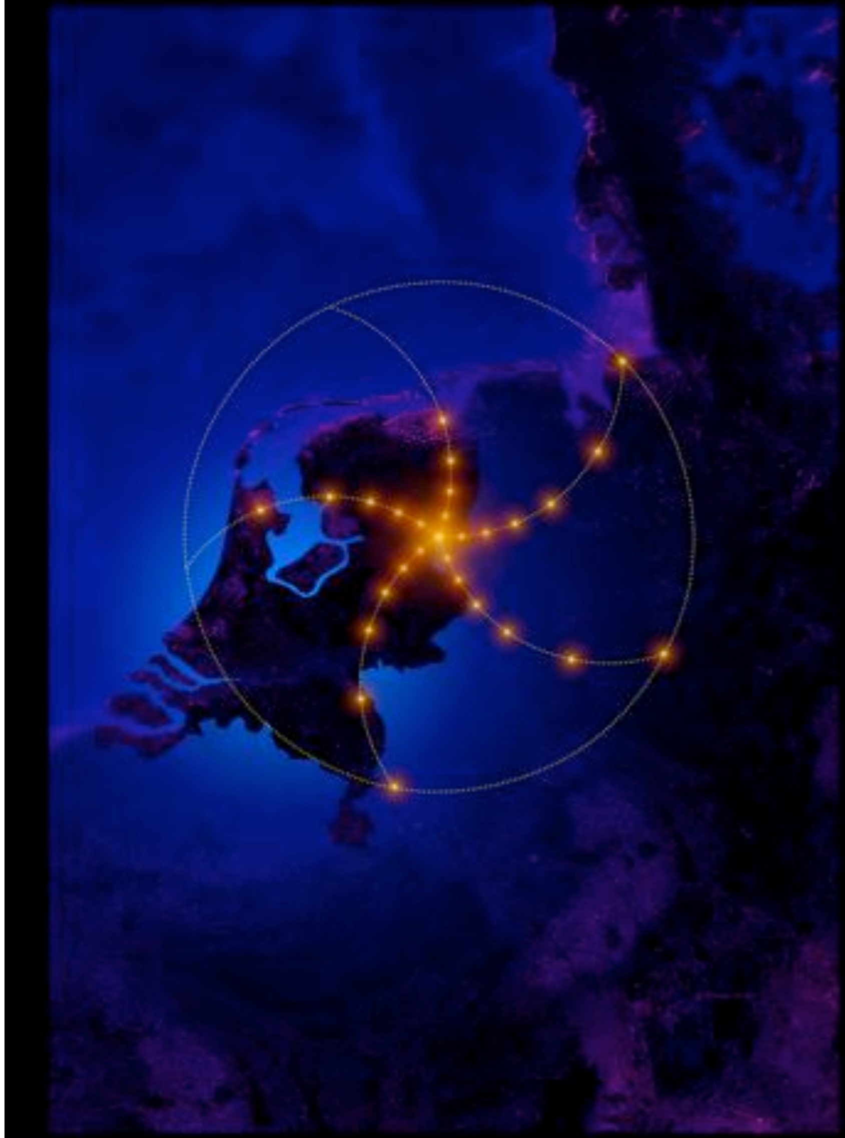
2 TFlop / 16 Gbps =

1000 flops/byte

THIS IS A DATA FLOW PROBLEM !!!

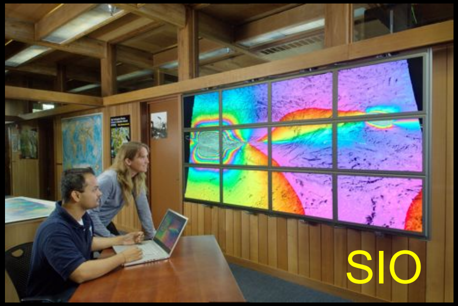
LOFAR as a Sensor Network

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


SIO

NCMIR

USGS EDC

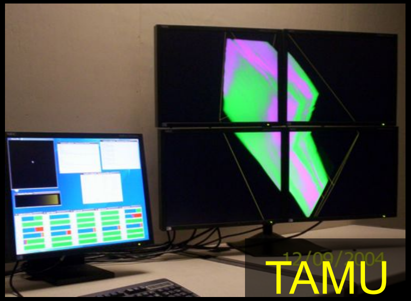
NCSA & TRECC

SARA

KISTI

AIST

RINCON & Nortel


TAMU

UCI

UIC

CALIT2

Real time, multiple 10 Gb/s

The “Dead Cat” demo

1 Mflops/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. 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:

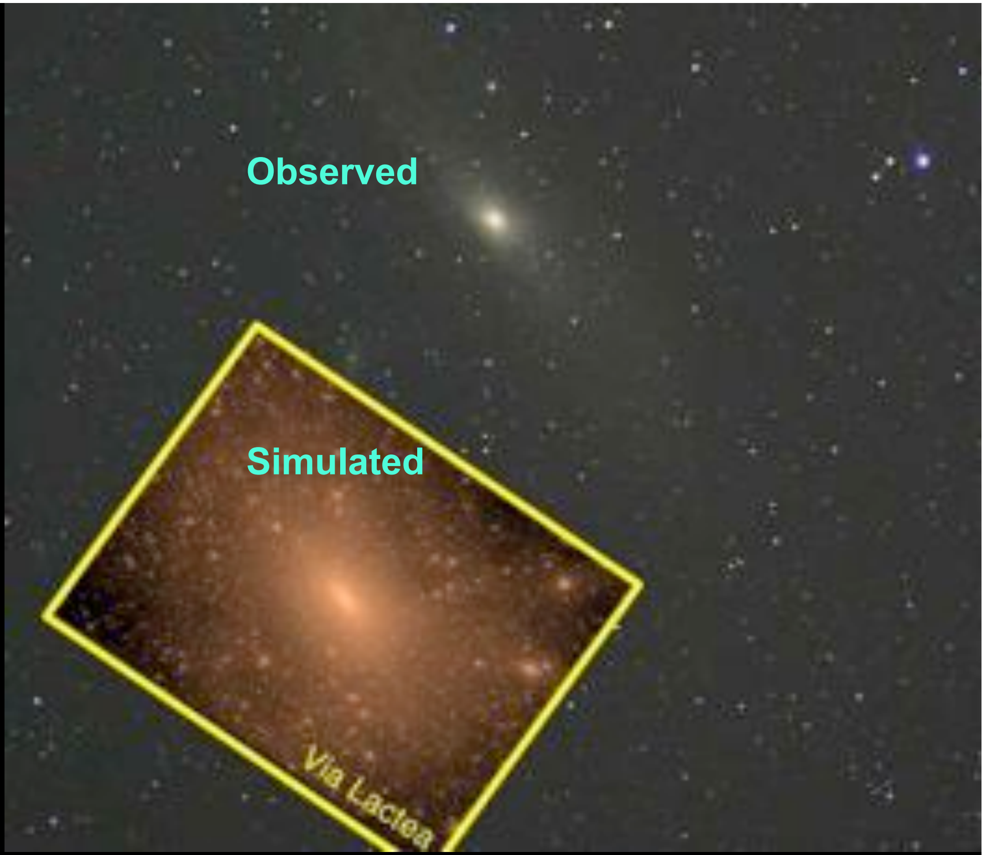
Many small flows -> 36 Gb/s

Urban_flow !

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

10 Mflops/byte

The hardware setup

- 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

CHINA
TOKYO
10 Gb/s dedicated network
GERMANY
BELGIUM

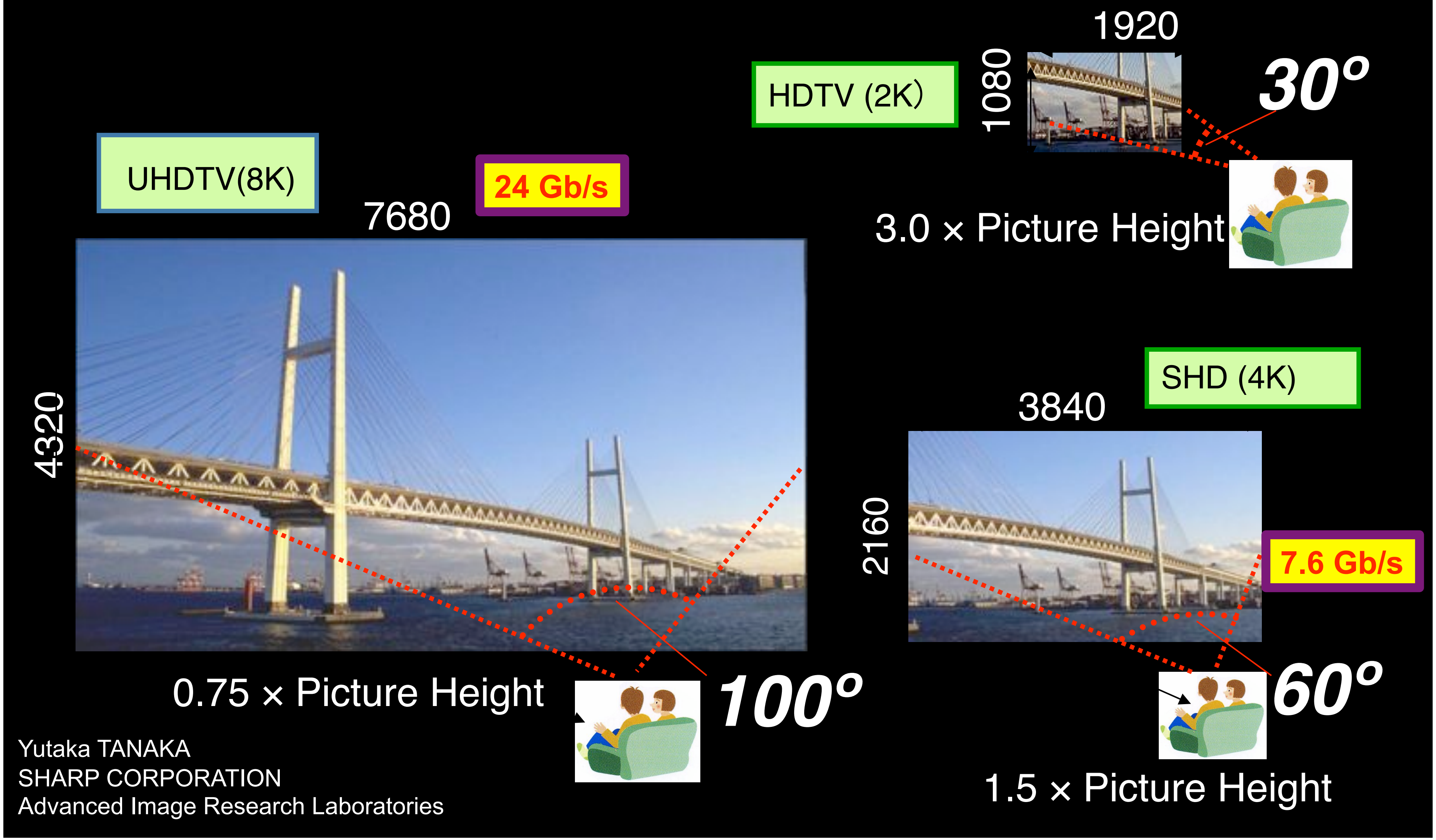
270 ms RTT

7.6 Gb/s

CineGrid @ Holland Festival 2007

CineGrid: 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

30º

3.0 × Picture Height

UHDTV(8K)

24 Gb/s

7680

4320

0.75 × Picture Height

100º

SHD (4K)

3840

2160

7.6 Gb/s

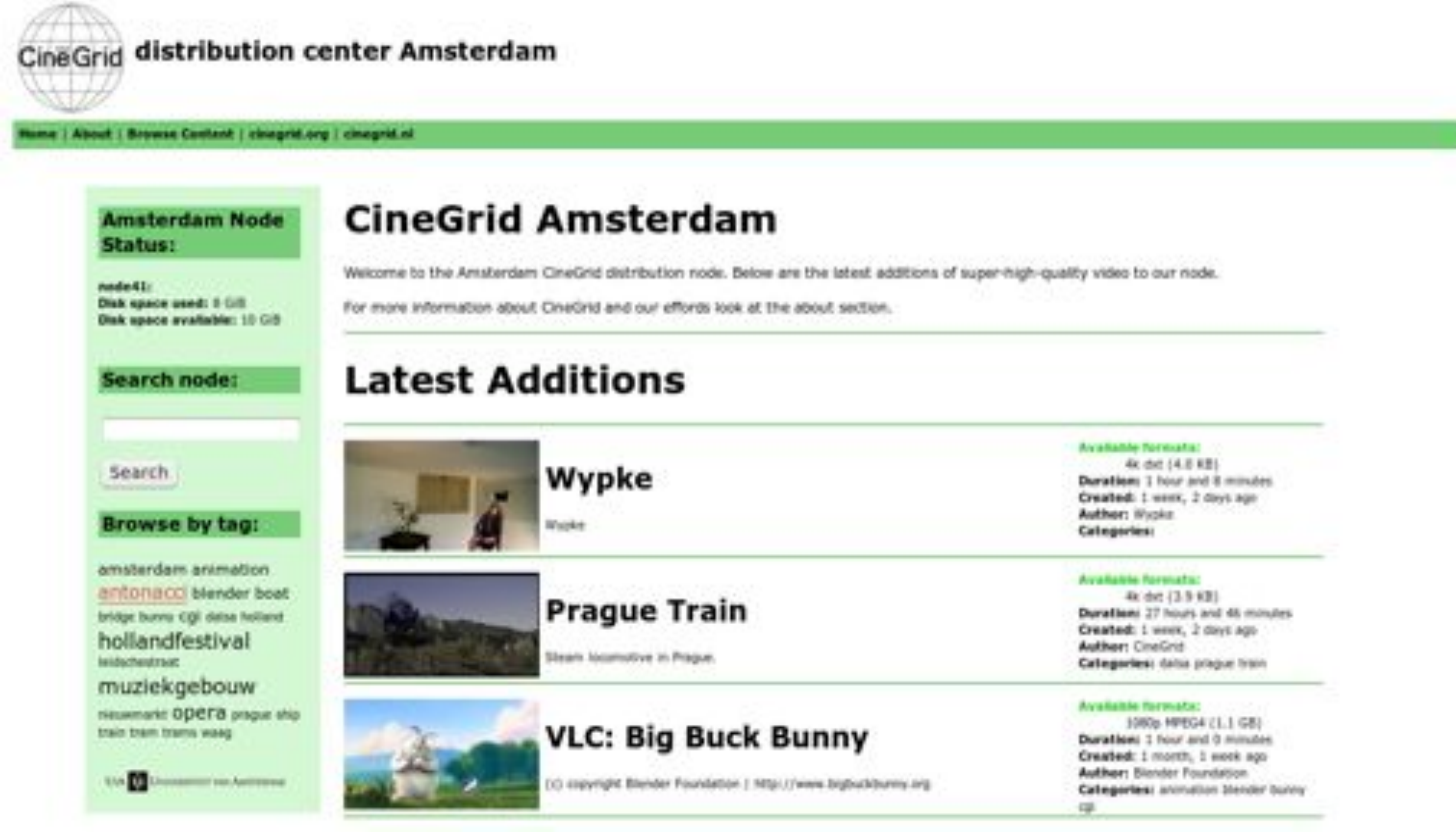
60º

1.5 × Picture Height

Yutaka TANAKA
SHARP CORPORATION
Advanced Image Research Laboratories

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 dalsa holland hollandfestival leidschestraat muziekgebouw nieuwmarkt opera prague ship train tram trams waag

UvA Universiteit van Amsterdam

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 effords look at the about section.

Latest Additions

Wypke

Wypke

Available formats:
4k dxt (4.8 KB)
Duration: 1 hour and 8 minutes
Created: 1 week, 2 days ago
Author: Wypke
Categories:

Prague Train

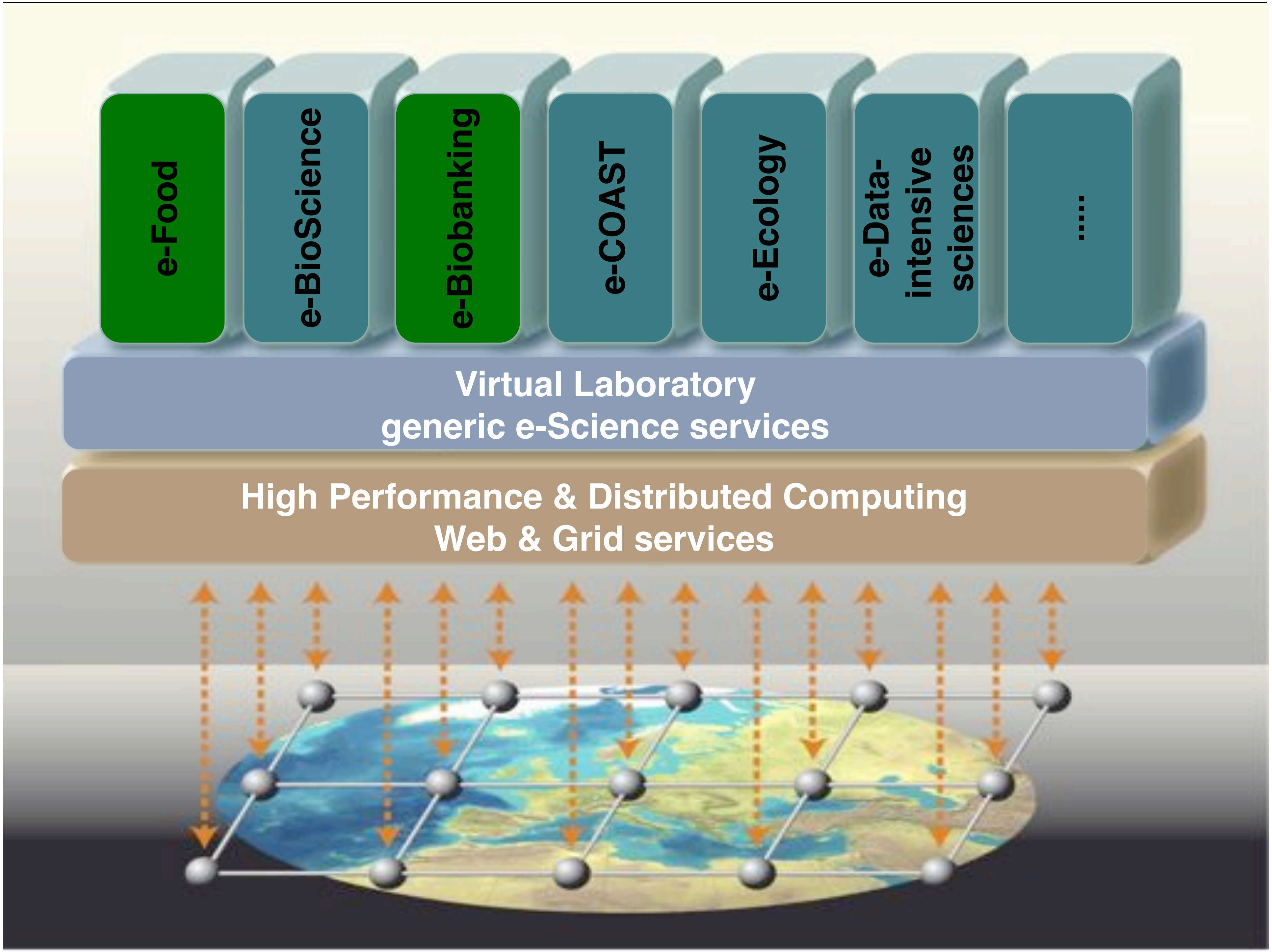
Steam locomotive in Prague.

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

VLC: Big Buck Bunny

(c) copyright Blender Foundation | http://www.bigbuckbunny.org

Available formats:
1080p MPEG4 (1.1 GB)
Duration: 1 hour and 0 minutes
Created: 1 month, 1 week ago
Author: Blender Foundation
Categories: animation blender bunny cgi



e-Food

e-BioScience

e-Biobanking

e-COAST

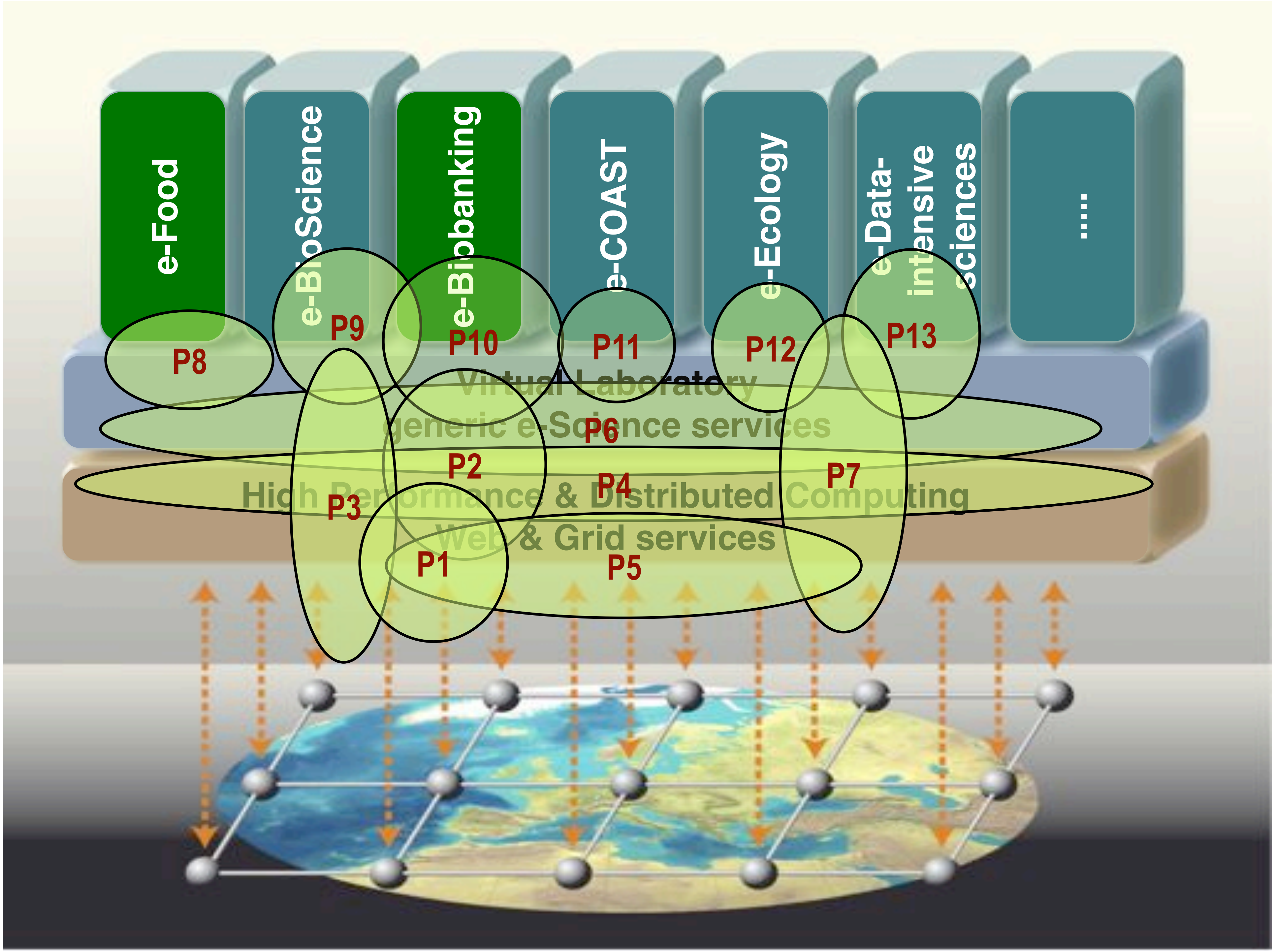
e-Ecology

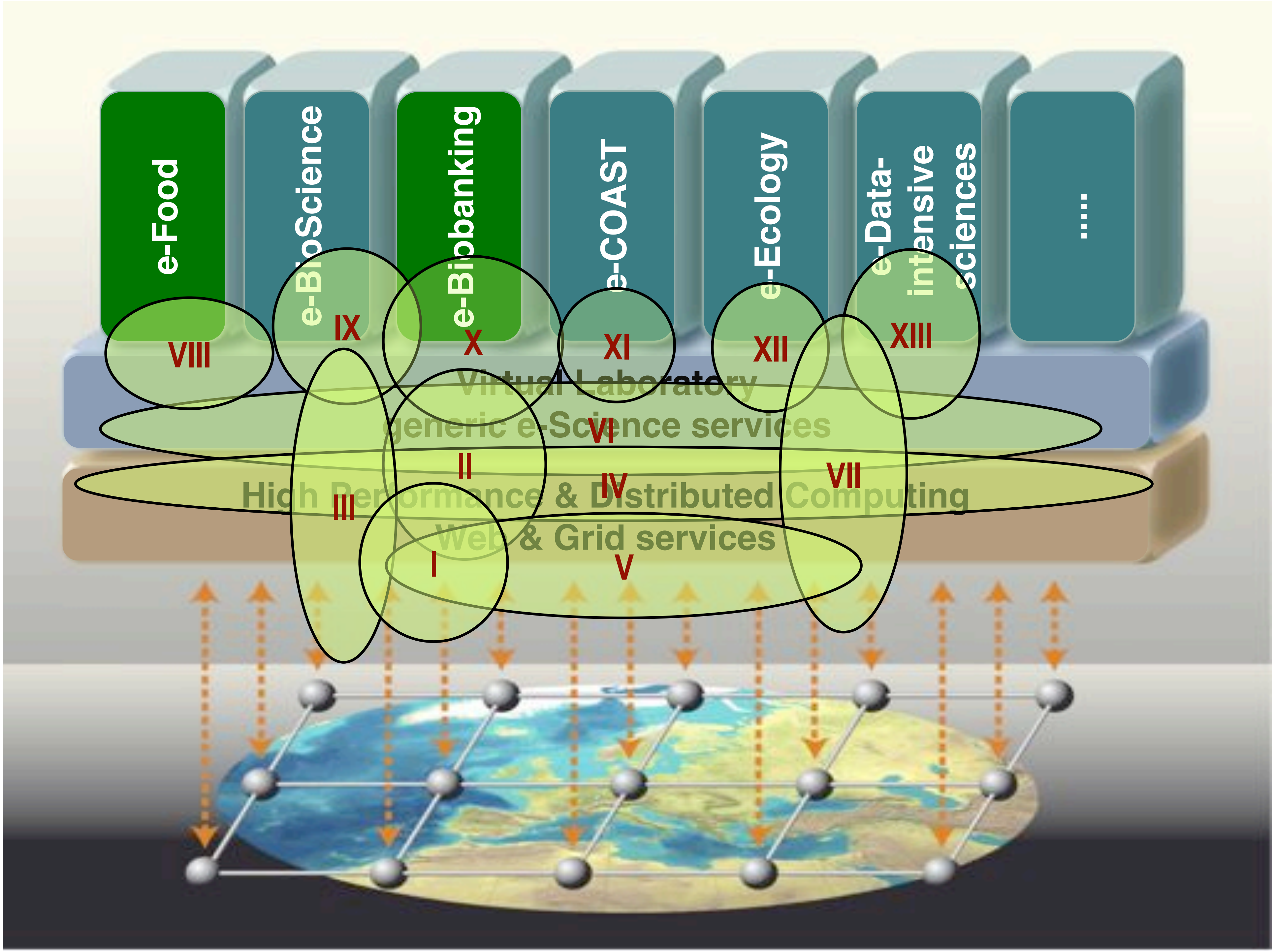
e-Data-intensive sciences

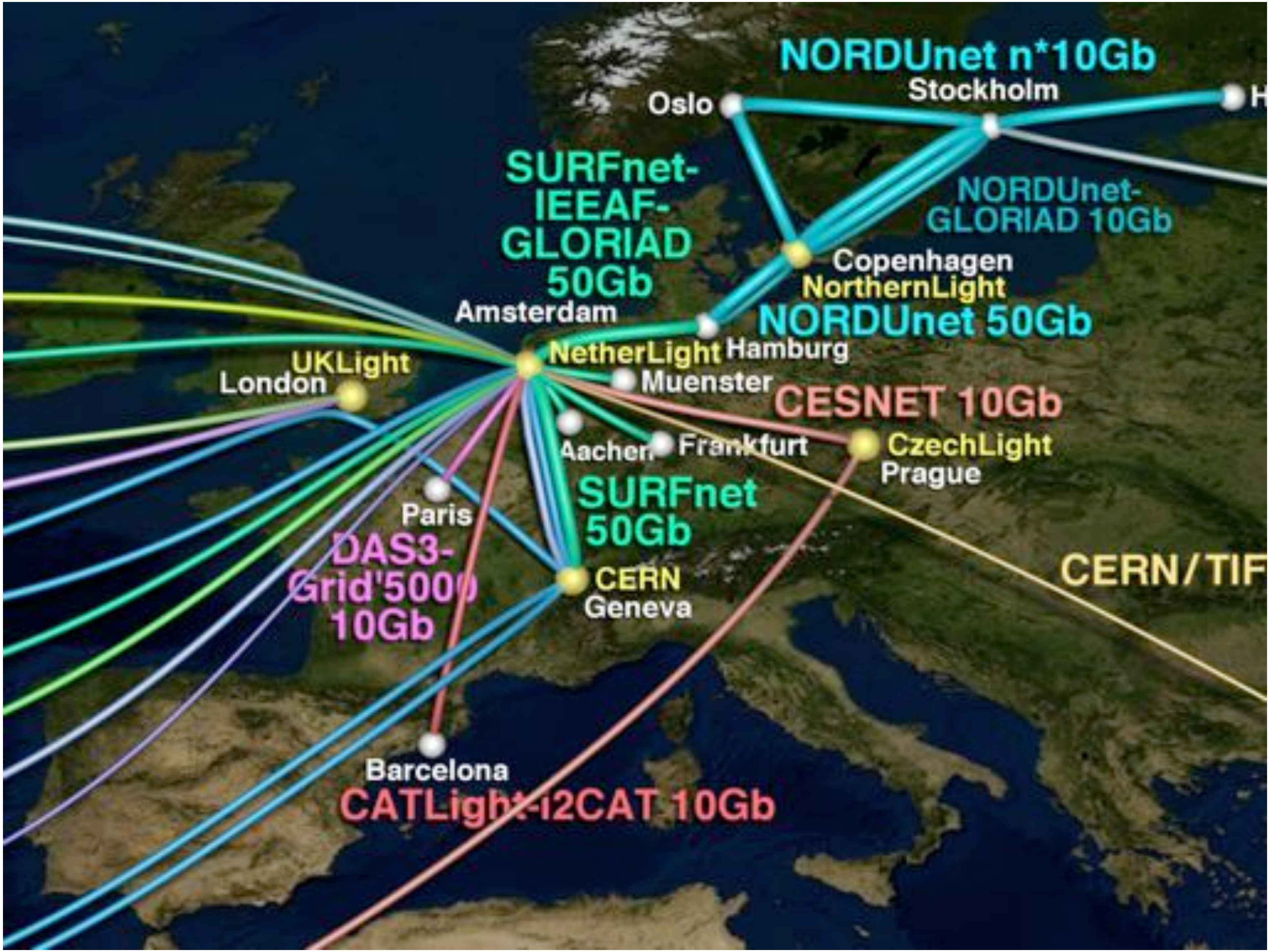
.....

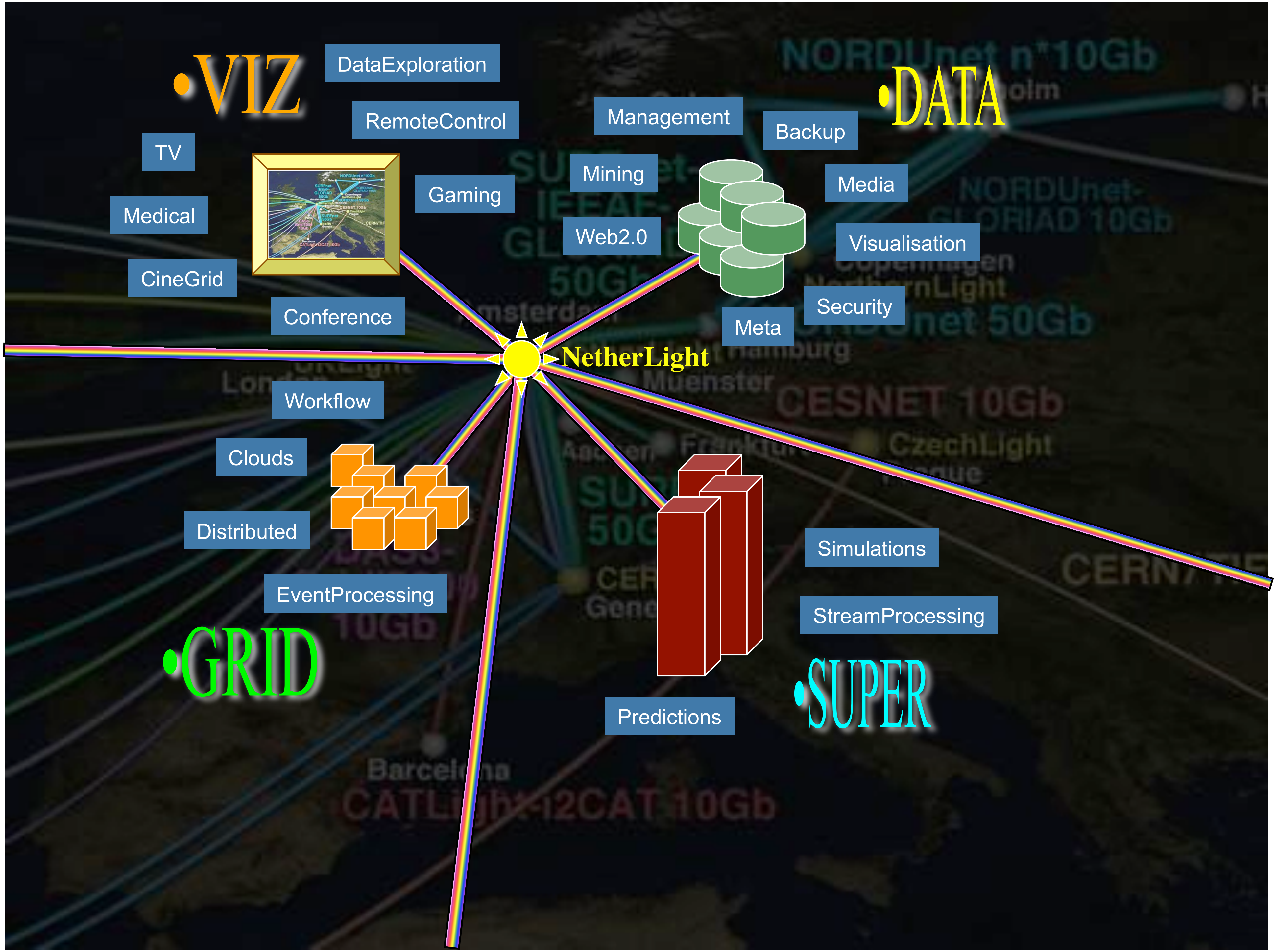
Virtual Laboratory
generic e-Science services

High Performance & Distributed Computing
Web & Grid services


e-Food
e-BioScience
e-Biobanking
e-COAST
e-Ecology
e-Data-intensive sciences
.....
Virtual Laboratory
generic e-Science services
High Performance & Distributed Computing
Web & Grid services
P8
P9
P10
P11
P12
P13
P6
P2
P4
P7
P3
P1
P5


e-Food
e-BioScience
e-Biobanking
e-COAST
e-Ecology
e-Data-intensive sciences
.....
Virtual Laboratory
generic e-Science services
High Performance & Distributed Computing
Web & Grid services
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII


NORDUnet n*10Gb
Stockholm
Oslo
H
SURFnet-
IEEAF-
GLORIAD
50Gb
NORDUnet-
GLORIAD 10Gb
Copenhagen
NorthernLight
NORDUnet 50Gb
Amsterdam
NetherLight
Hamburg
UKLight
London
Muenster
CESNET 10Gb
Aachen
Frankfurt
CzechLight
Prague
Paris
SURFnet
50Gb
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

•GRID

Workflow

Clouds

Distributed

EventProcessing

•SUPER

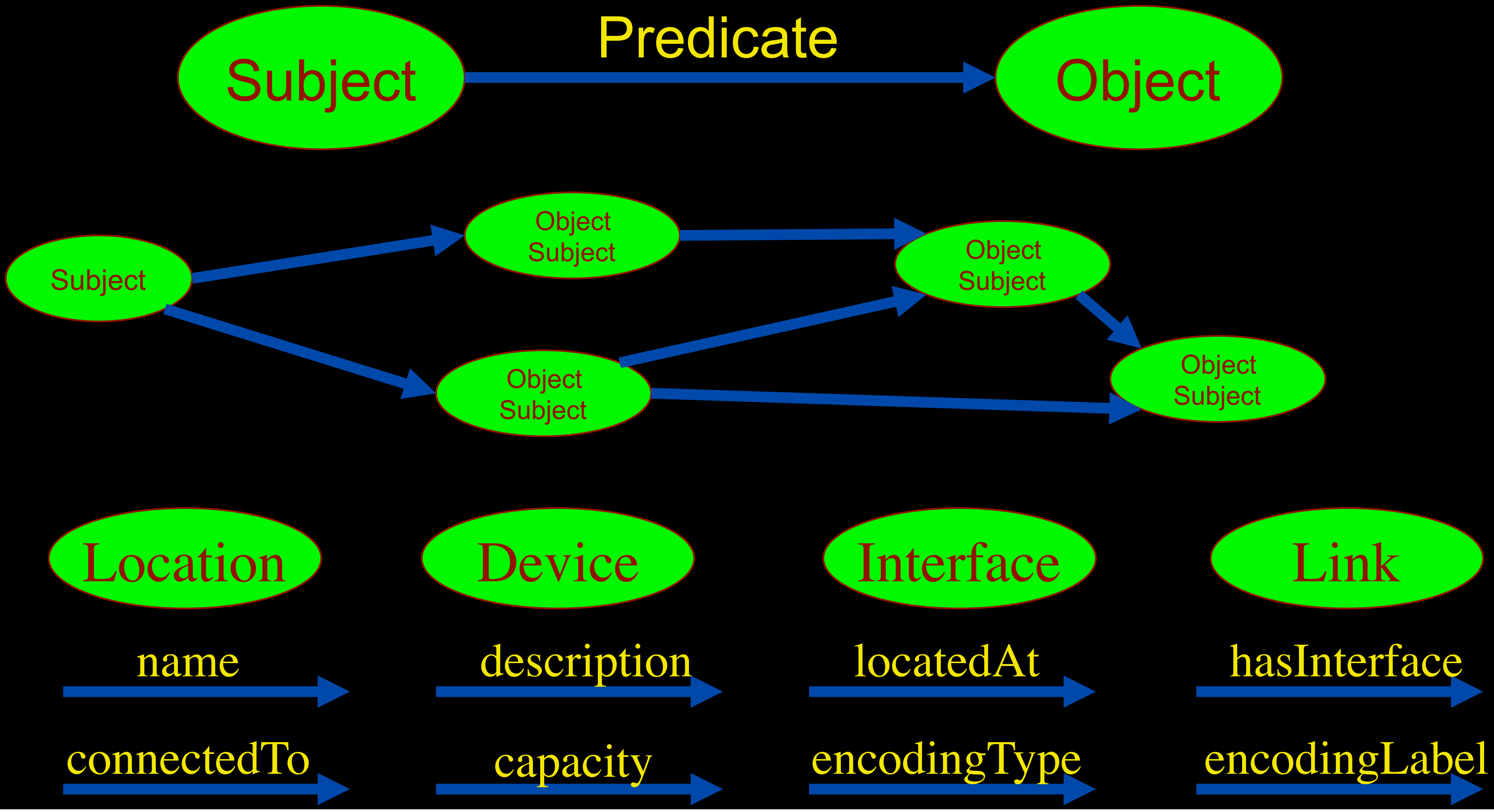
Simulations

StreamProcessing

Predictions

Network Description Language

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



Subject

Predicate

Object

Subject

Object
Subject

Object
Subject

Object
Subject

Object
Subject

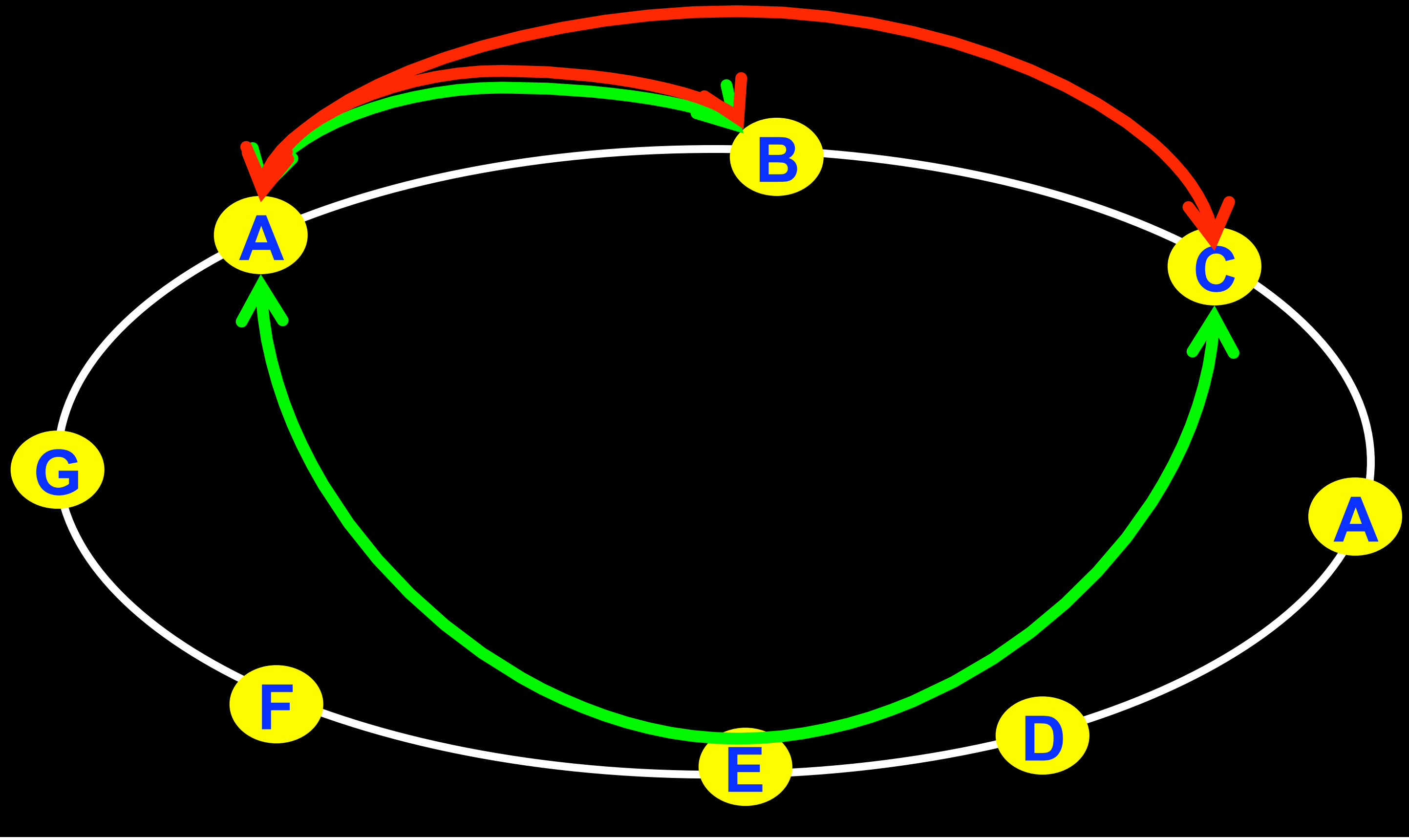
Location	Device	Interface	Link
name	description	locatedAt	hasInterface
connectedTo	capacity	encodingType	encodingLabel

The Problem

I want AC and AB

Success depends on the order of requests

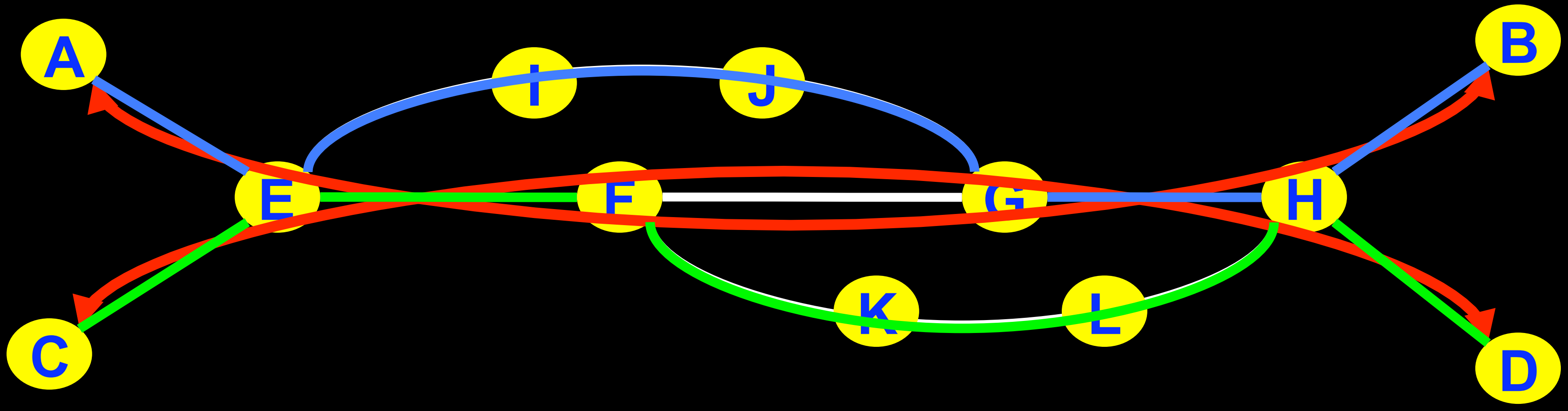
Wouldn’t it be nice if I could request [AB, AC, ...]



Another one ☺

I want AB and CD

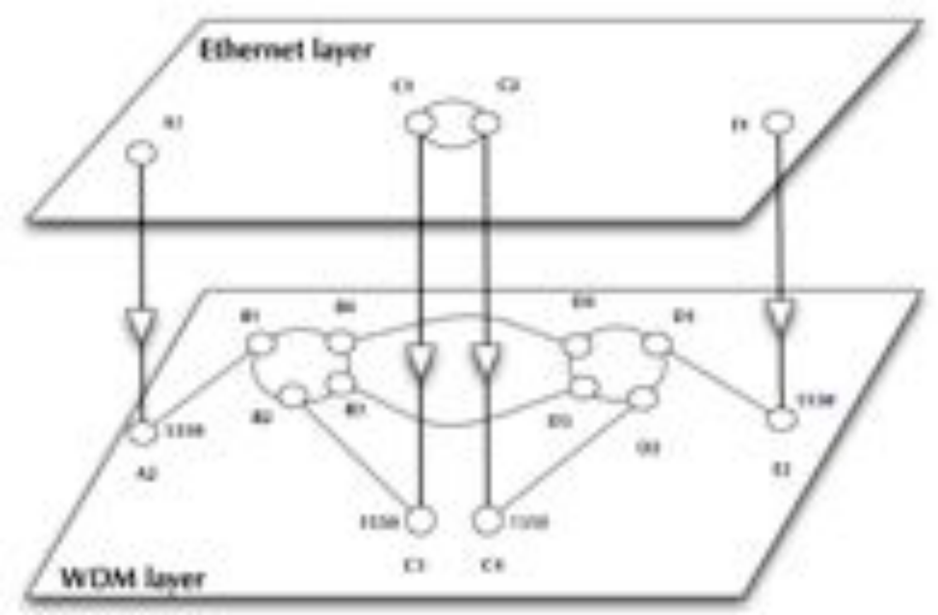
Success does not even depend on the order!!!

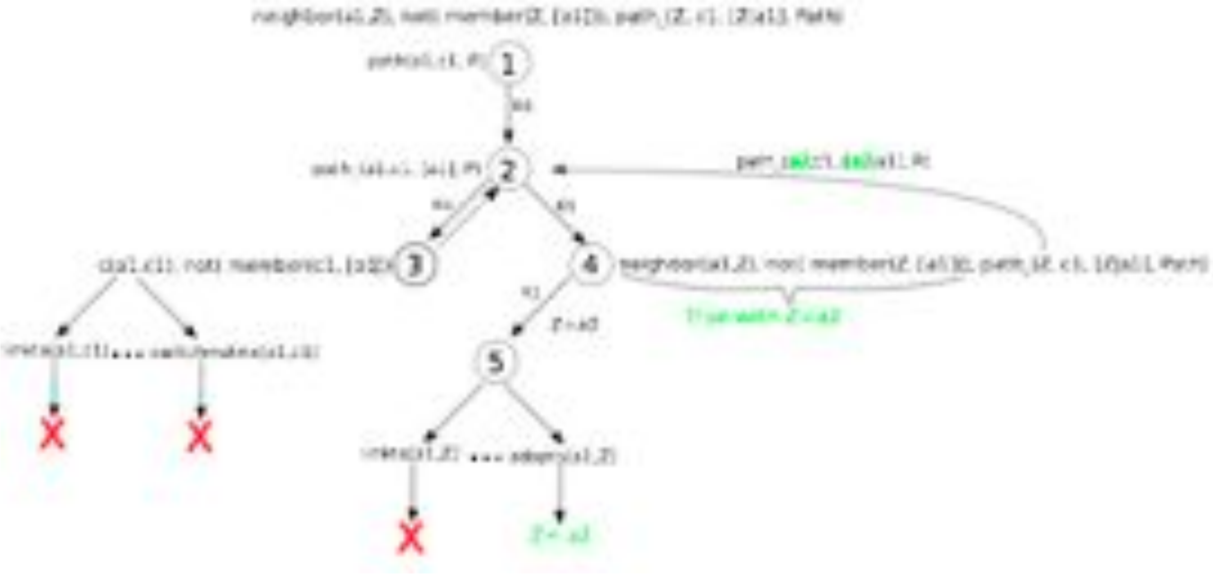


NDL + PROLOG

Research Questions:

- order of requests
- complex requests
- usable leftovers

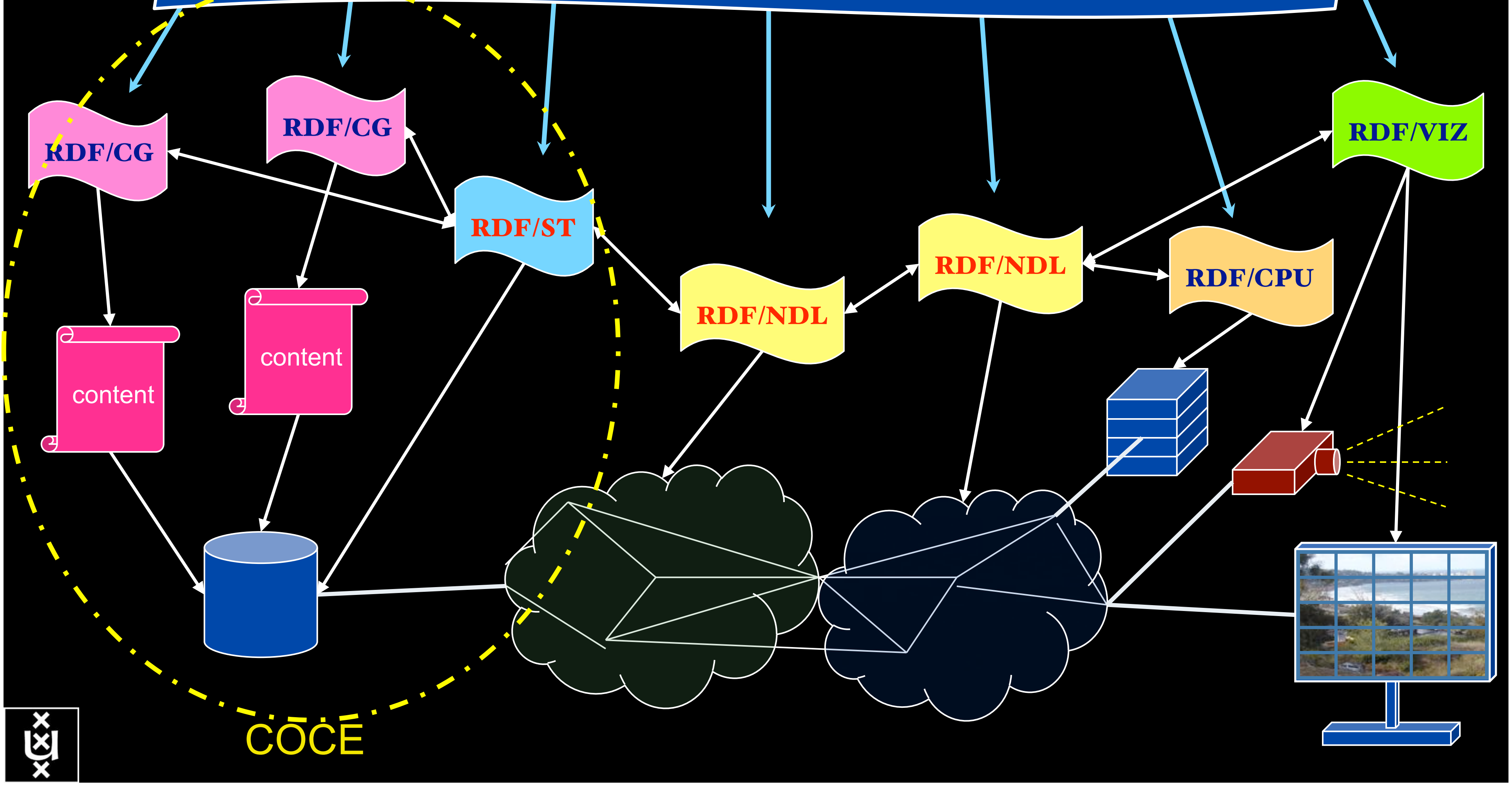




- Reason about graphs
- Find sub-graphs that comply with rules
- It finds solutions to previous slides!

RDF describing Infrastructure
“I want”

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



TeraThinking

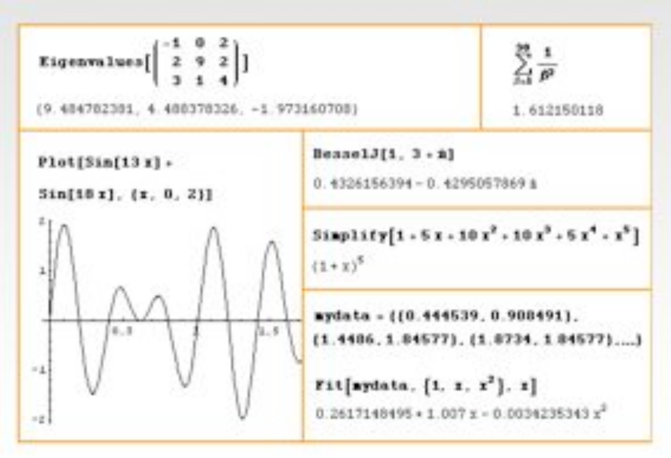
- What constitutes a Tb/s network?
- CALIT2 has 8000 Gigabit drops ?->? Terabit Lan?
- look at 80 core Intel processor
  - cut it in two, left and right communicate 8 TB/s
- think back to teraflop computing!
  - MPI turns a room full of pc's in a teraflop machine
- massive parallel channels in hosts, NIC's
- TeraApps programming model supported by

– TFlops	->	MPI / Globus
– TBytes	->	OGSA/DAIS
– TPixels	->	SAGE
– TSensors	->	LOFAR, LHC, LOOKING, CineGrid, ...
– Tbit/s	->	?

ref Larry Smarr & CdL

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



$$\text{Eigenvalues}\left[\begin{pmatrix} -1 & 0 & 2 \\ 2 & 9 & 2 \\ 3 & 1 & 4 \end{pmatrix}\right]$$

{9.484782381, 4.488378326, -1.973160708}

$$\sum_{j=1}^{30} \frac{1}{j^2}$$

1.612150118

Plot[Sin[13 x] + Sin[18 x], {x, 0, 2}]

BesselJ[1, 3 + i]

0.4326156394 - 0.4295057869 i

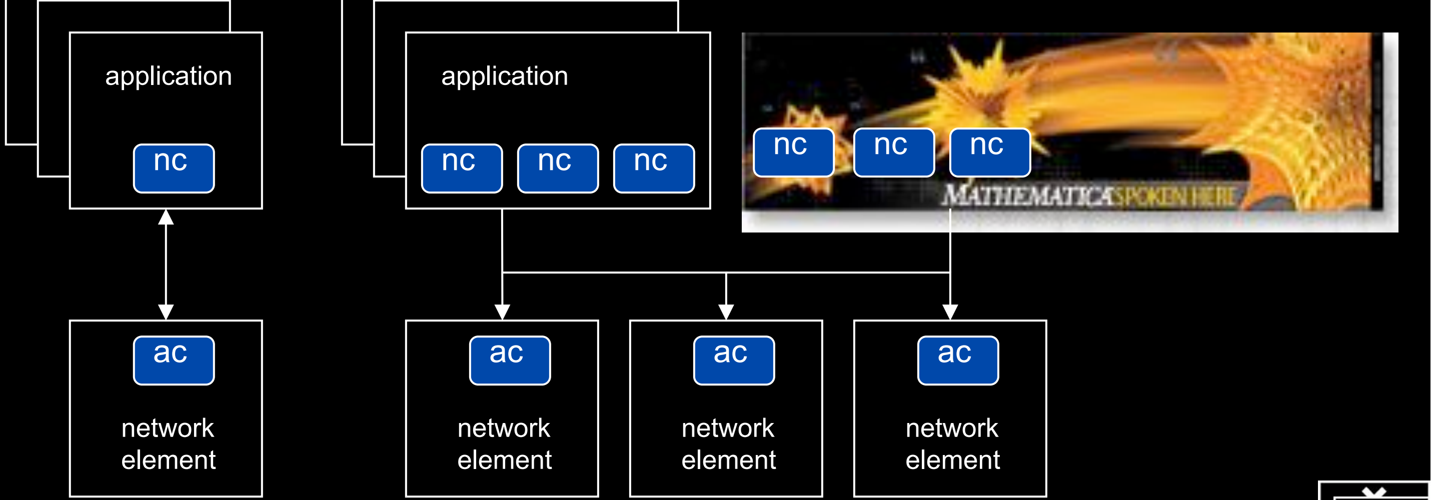
Simplify[1 + 5 x + 10 x^2 + 10 x^3 + 5 x^4 + x^5]

$(1 + x)^5$

mydata = {{0.444539, 0.908491}, {1.4486, 1.84577}, {1.8734, 1.84577},....}

Fit[mydata, {1, x, x^2}, x]

$0.2617148495 + 1.007\,x - 0.0034235343\,x^2$

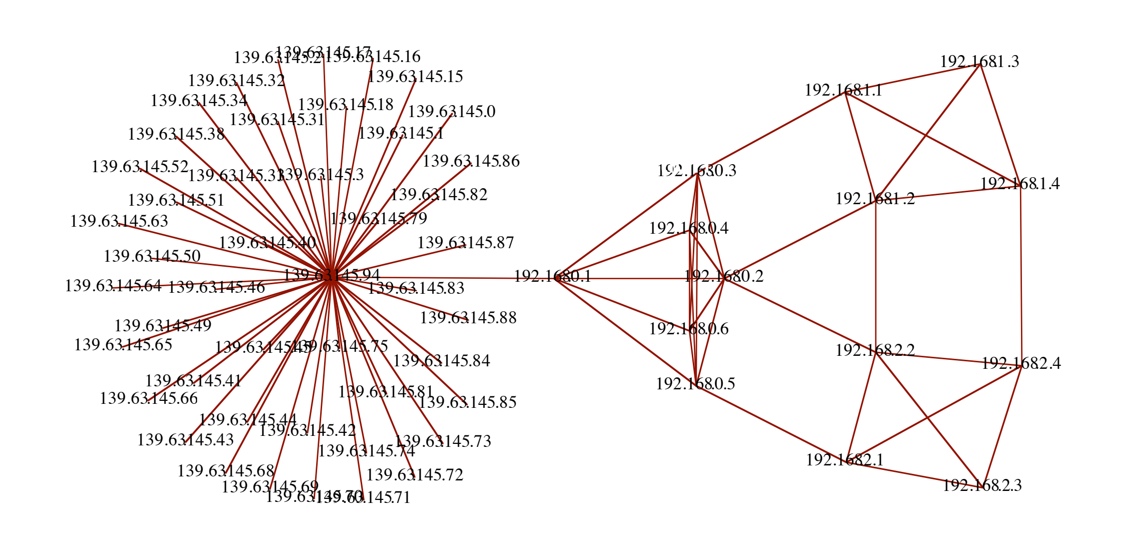


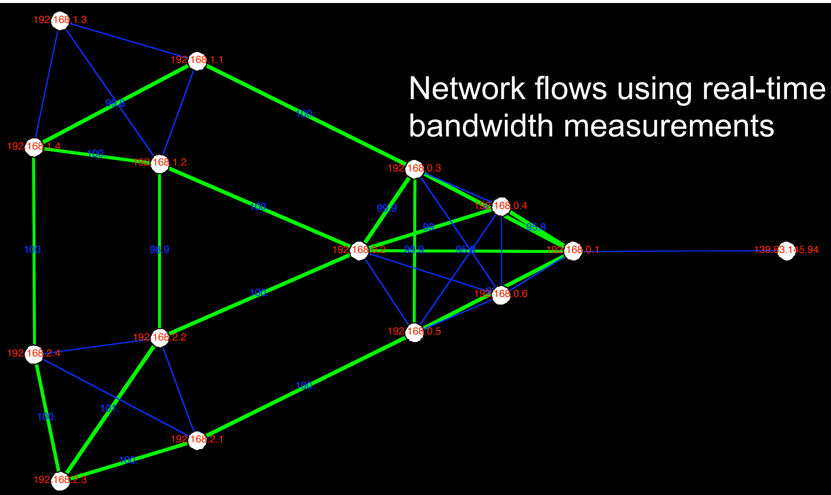
Mathematica enables advanced graph queries, visualizations and real-time network manipulations on UPVNs

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}
```





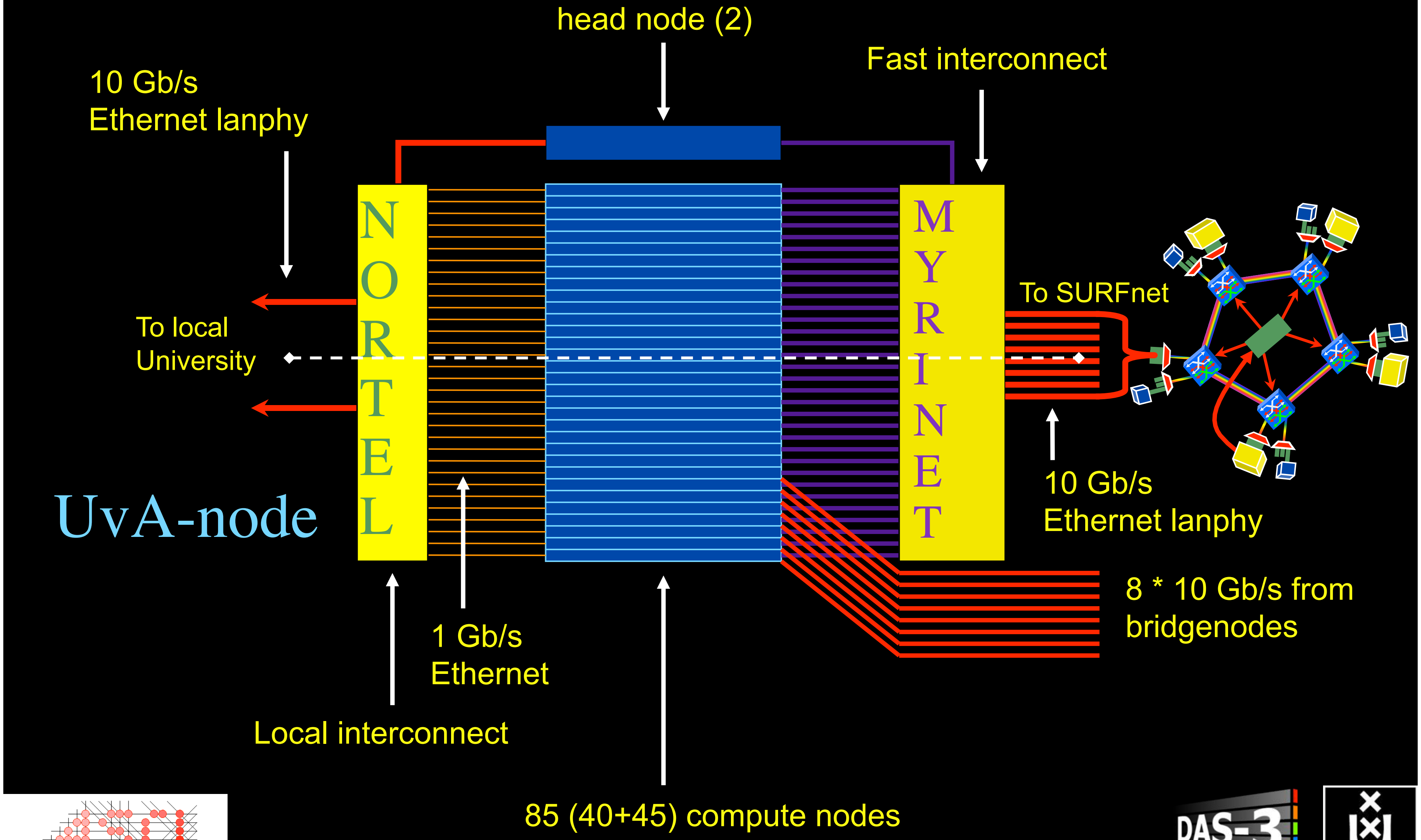
Network flows using real-time bandwidth measurements

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
```

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.

DAS-3 Cluster Architecture



head node (2)

Fast interconnect

10 Gb/s
Ethernet lanphy

To local
University

UvA-node

NORTEL

MYRINET

To SURFnet

10 Gb/s
Ethernet lanphy

8 * 10 Gb/s from
bridgenodes

1 Gb/s
Ethernet

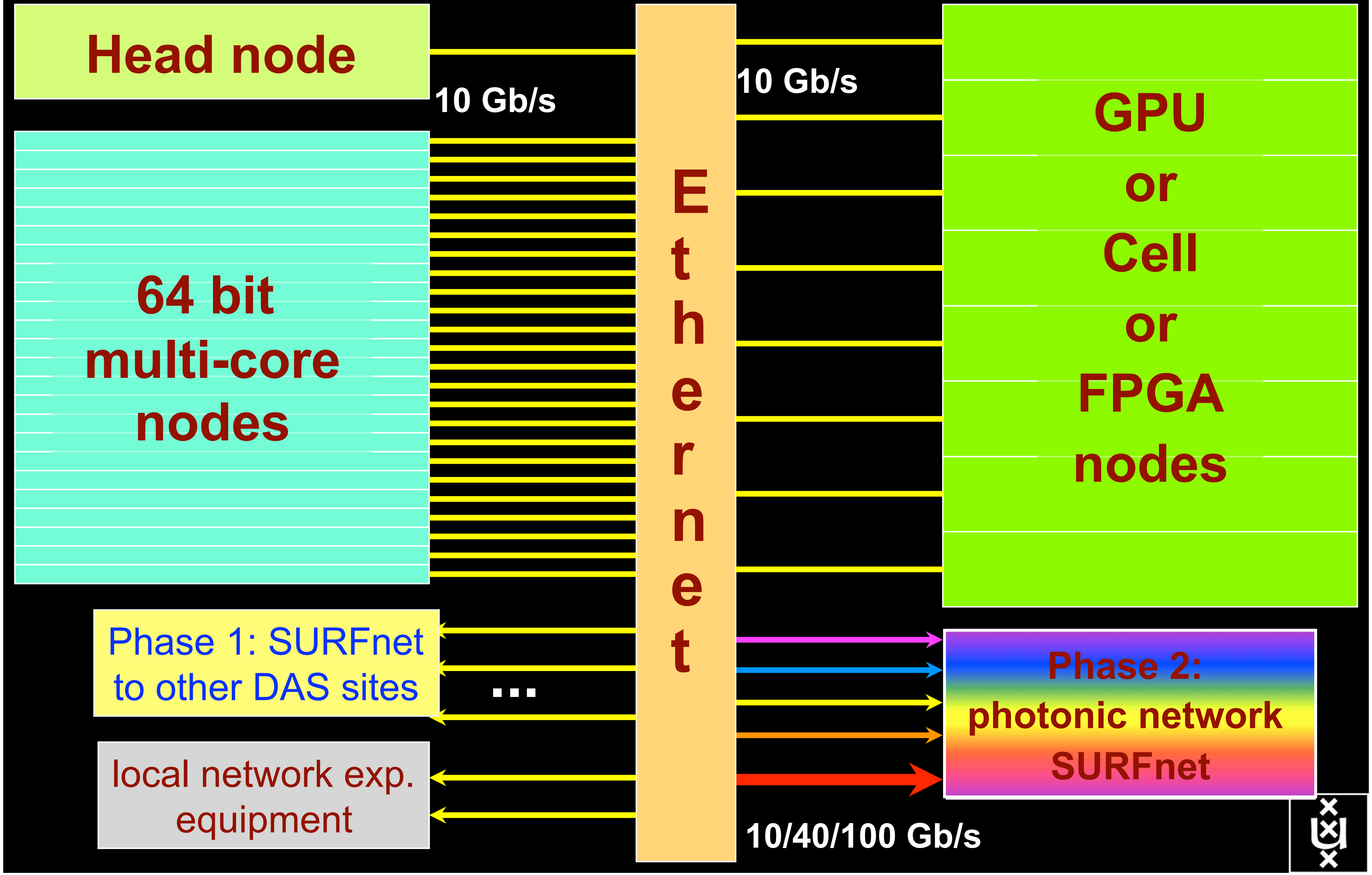
Local interconnect

85 (40+45) compute nodes

Advanced School for Computing and Imaging

DAS-3

DAS-4 Proposed Architecture



Head node

10 Gb/s

64 bit multi-core nodes

Ethernet

10 Gb/s

GPU or Cell or FPGA nodes

Phase 1: SURFnet to other DAS sites

...

local network exp. equipment

Phase 2: photonic network SURFnet

10/40/100 Gb/s

Themes for next years

- Network modeling and simulation
- Cross domain Alien Light switching
- eScience infrastructure virtualization (NSI)
- Photonic networking -> Tb/s
- Capacity & Capability
- Data handling, integrity, security, privacy
- Reasoning about services
- Fault tolerance, Fault isolation, monitoring
- eScience Data and Media specific services
- Cloud paradigm, green compute&store&net&viz
- ENERGY dependency! (2009: 1Wy=1€)

Quotes from OnVector 2008

prof. Ken-Ichi Sato:

- It is very difficult to predict future services, however, video is expected to be the king media used for bit rate demanding services. High-quality video technologies are rapidly advancing.
- TCP/IP bottleneck is becoming more and more tangible. It will limit the future envisaged network expansion -the energy bottleneck and throughput bottleneck need to be resolved.
- Fast optical circuit/path switching will play the key role to create cost effective and bandwidth abundant future networks.
- Hierarchical optical path network and the node technologies are very important, and hence they need to be fully developed soon.

Quotes from OnVector 2008

- dr. Kazuo Hagimoto:
- NTT is developing a system that automatically generates metadata such as title, summary, and key words that are extracted from voice or subtitles.

dr. Shimizu:

- Applications for Tbit networks:
  - High Resolution Simulation
  - Weather Forecast
  - Earthquake Forecast
  - City Planning
  - Digital Engineering
  - Nano Device Engineering
  - Protein Structural Analysis

Quotes from OnVector 2008

prof. Larry Smarr:

- Interconnecting Regional Optical Networks
  Is Driving Campus Optical Infrastructure Deployment

prof. Ed Seidel:

- Petascale computing will not only provide huge data, but will demand new computing modalities
- Will place new demands on networking, data management, visualization, resource co- allocation
- Applications need to be configurable for the new type of infrastructure, need to be aware of environment
- If we don’t solve these problems, people will use machines anyway, but science will suffer!

Bill s’Arnaud:

- “Optical networks (as opposed to electronic routed networks) have much smaller carbon footprint”

Interactive programmable networks

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.