ICT & e-Science

Cees de Laat

GLIF.is & CineGrid.org founding member

SURFnet

BSIK

NWO

EU

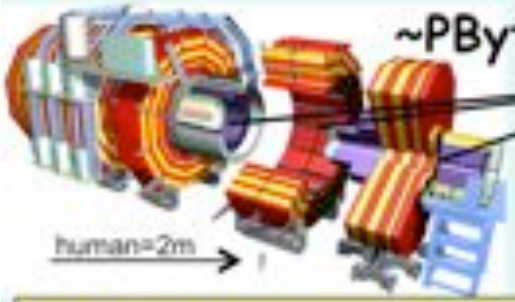
University of Amsterdam

TNO

NCF

LHC Data Grid Hierarchy

CMS as example, Atlas is similar



~PByte/sec

Online System

~100 MBytes/sec

100000 flops/byte

10 Pflops/s

event simulation

Tier 0 +1

HPSS

event reconstruction

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

Tier 2

~0.6-2.5 Gbps

Tier 3

Institute ~0.25TIPS

Physics data cache

100 - 1000 Mbits/sec

Tier 4

Workstations

Courtesy Harvey Newman, CalTech and CERN

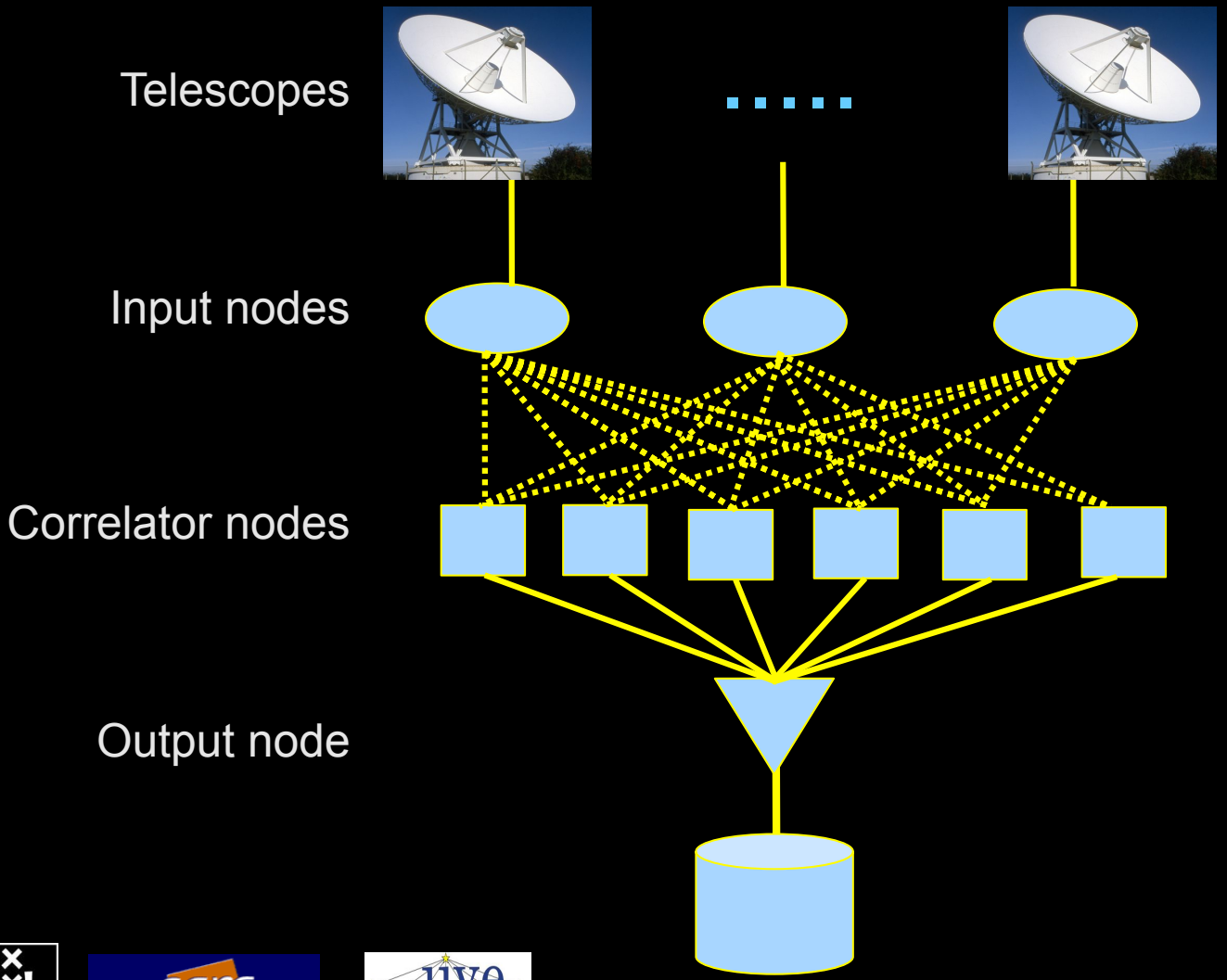
CERN/CMS data goes to 6-8 Tier 1 regional centers, and from each of these to 6-10 Tier 2 centers.

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

0.1 Pflops/s

THIS IS A DATA FLOW PROBLEM !!!

sara Computing & Networking Services

jive JOINT INSTITUTE FOR VLBI IN EUROPE

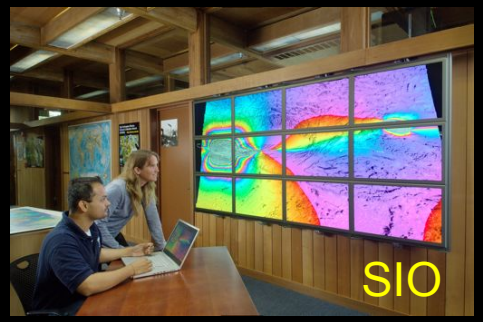
LOFAR as a Sensor Network

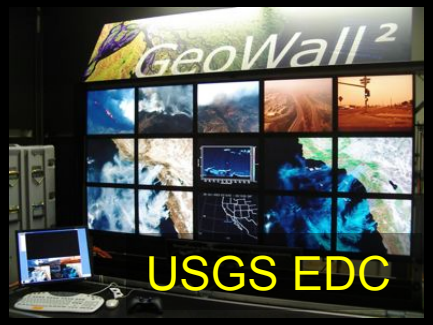
20 flops/byte

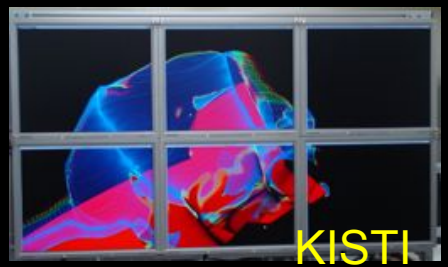
2 Tflops/s

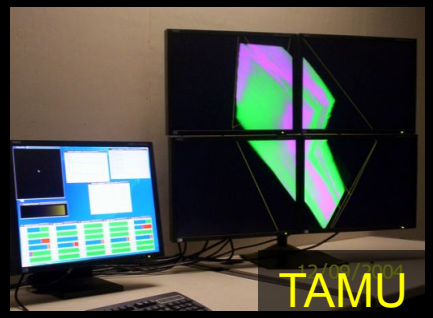
- LOFAR is a large distributed research infrastructure:
  - Astronomy:
    - >100 phased array stations
    - Combined in aperture synthesis array
    - 13,000 small “LF” antennas
    - 13,000 small “HF” tiles
  - Geophysics:
    - 18 vibration sensors per station
    - Infrasound detector per station
  - >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

1 Mflops/byte

Real time issue

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

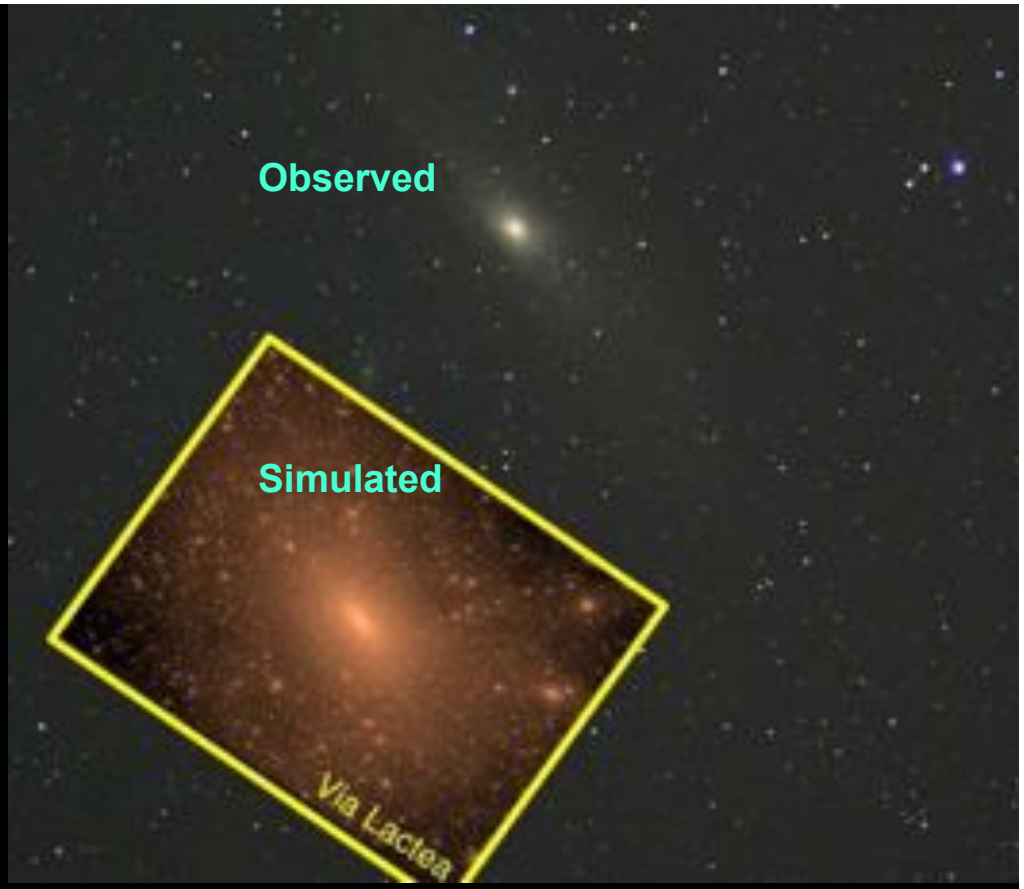
300000 * 60 kb/s * 2 sensors (microphones) to cover all Dutch dikes

Sensor grid: instrument the dikes

First controlled breach occurred on sept 27th ‘08:

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

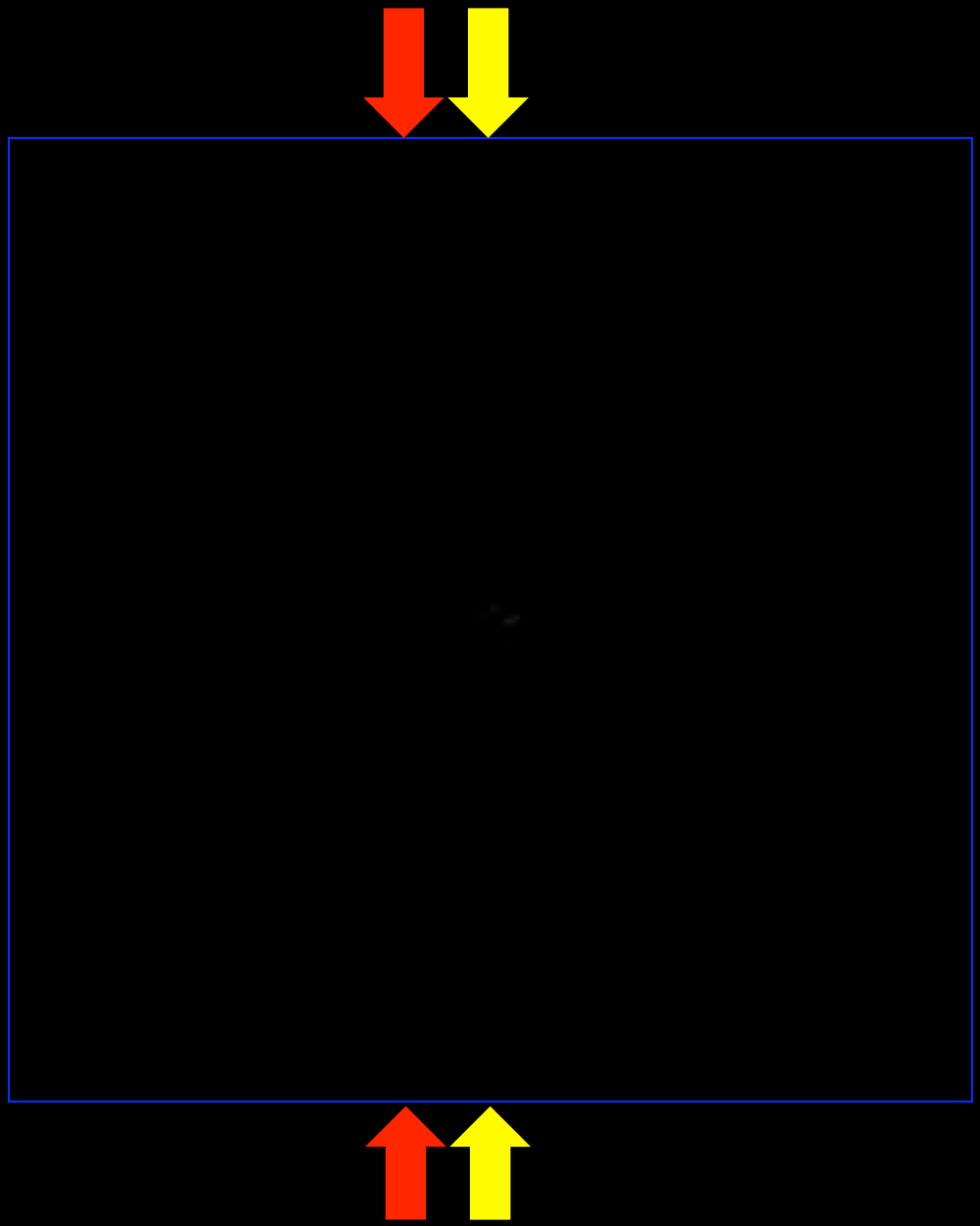
The hardware setup

10 Mflops/byte

1 Eflops/s

- 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

Auto-balancing Supers



CosmoGrid

Supercomputing Grid across Continents and Oceans

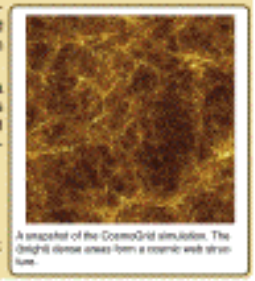
And yes, it works!

Application

We originally developed MPWide to manage the long-distance message passing in the CosmoGrid[a] project. This is a large-scale cosmological project whose primary goal is to perform a dark matter simulation using supercomputers on two continents.

In this simulation, we use the cosmological Λ Cold Dark Matter model[b] to simulate the dark matter particles using a parallel tree/particle-mesh N-body integrator, TreePM. This requires relatively little communication between different sites after each timestep. This integrator calculates the dynamical evolution of 2048³ (8.5 billion) particles. More information about the parameters used and the scientific rationale can be found in [a].

The integrator can be run as a single MPI application, or as two separately launched MPI applications on different supercomputers.



A snapshot of the CosmoGrid simulation. The (bright) dense areas form a cosmic web structure.

[a] Portegies Zwart et al., 2009, IEEE Computer (submitted)

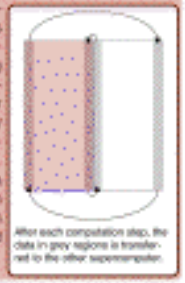
[b] Guth, 1981; Physical Review D

[c] Yoshikawa and Fukushige, 2005; PASJ

Motivation

We use MPWide to manage the wide area communications in the CosmoGrid project, where cosmological N-body simulations run on grids of supercomputers connected by high performance optical networks. To take full advantage of the network light paths in CosmoGrid, we need a message passing library that supports the ability to use customized communication settings (e.g. custom number of streams, window sizes) for individual network links among the sites. The supercomputers we use vary both in hardware architectures and software setup.

Many supercomputers have a recommended MPI implementation which has been optimized for the network architecture of that particular machine. Installing and optimizing a homogeneous MPI implementation on multiple supercomputer platforms is a task that may be politically difficult to initiate, and requires considerable effort and man hours to complete. This has led us to develop MPWide, a light-weight communication library which connects two applications, each of them running with the locally recommended MPI implementation.



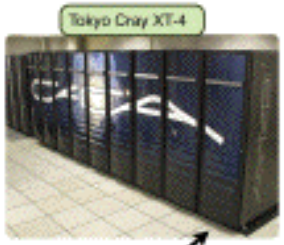
After each computation step, the data in grey regions is transferred to the other supercomputer.

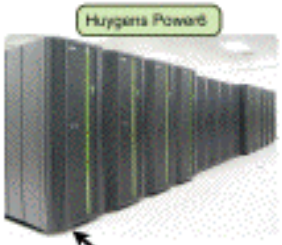
MPWide

A communication library for distributed supercomputing

Derek Groen[1,2], Steven Rieder[1,2], Paola Grosso[2], Simon Portegies Zwart[1] and Cees de Laat[2]

[1] Leiden Observatory, Leiden, the Netherlands
[2] University of Amsterdam, Amsterdam, the Netherlands





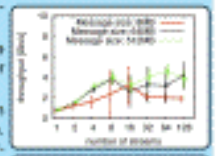
Benchmarks

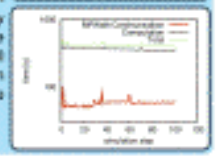
We measured the performance of MPWide between two nodes on different supercomputers, one located in The Netherlands, the other in Finland. These supercomputers are connected with a 10 Gbps interface. The round trip time for this network is 37.6 ms.

Each test consists of 100 two-way message exchanges, where we record the average throughput and the standard error. We performed the tests over a shared network with frequent background traffic.

Our tests show increased performance when using more streams, especially for larger message sizes.

We also tested MPWide in a production environment, during a CosmoGrid run. In this run, we used the Huygens supercomputer in Amsterdam and the Cray supercomputer in Tokyo. In this run, the calculation time dominated the overall performance, with the communication time constituting about one eighth of the total execution time.

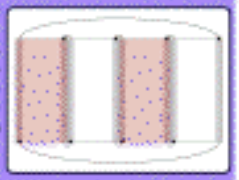




Related work and future

The MPI implementation most closely related to our work is the PACX-MPI[d] implementation. Like MPWide, this implementation connects different machines, while making use of the vendor MPI library on the system. The main difference between PACX-MPI and MPWide lies in the fact that MPWide supports a de-centralized startup, where PACX-MPI does not. For CosmoGrid, support for this is required, as it is not possible to start the simulation on all supercomputers from one site. Other implementations of MPI, like Open MPI and MPICH-G2, differ further from MPWide, and do not support manual specification of the network topology, required by CosmoGrid.

In the near future, we will expand the CosmoGrid simulation to run on four supercomputer sites, and we will implement support for this in MPWide.



[d] http://www.hlrs.de/organization/av/amt/research/pacx-mpi/


7.6 Gb/s
Real time issue
CineGrid
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

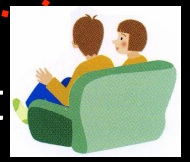
HDTV (2K)

1920

1080

30°

3.0 × Picture Height



UHDTV(8K)

24 Gb/s

7680

4320

0.75 × Picture Height

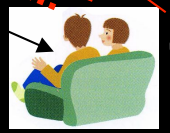
100°

SHD (4K)

3840

2160

7.6 Gb/s



60°

1.5 × Picture Height

Yutaka TANAKA
SHARP CORPORATION
Advanced Image Research Laboratories

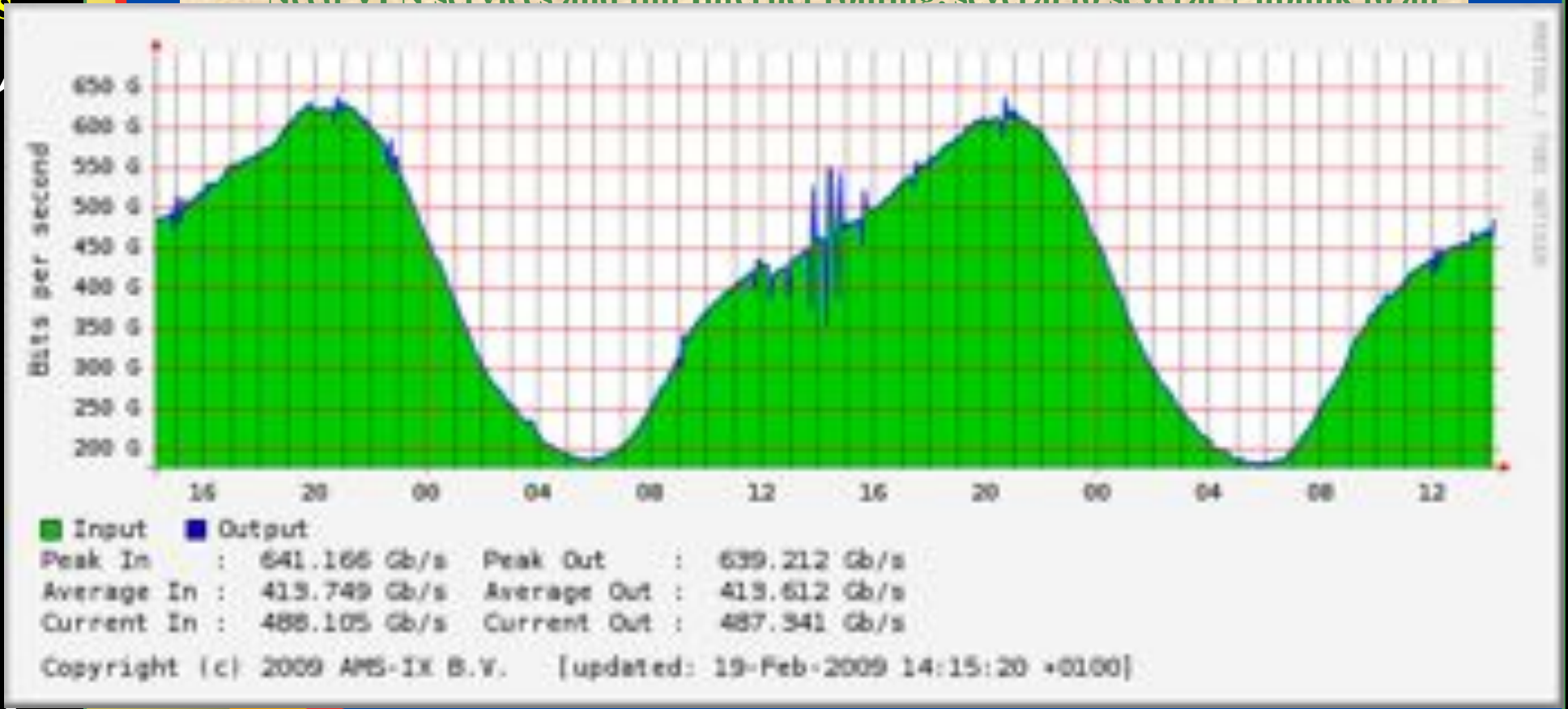
#users

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 (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 ≈ 2-3 k$/port

L2 ≈ 5-8 k$/port

L3 ≈ 75+ k$/port

Hybrid computing

Routers ⇐ ⇒ Supercomputers

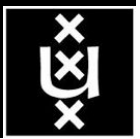
Ethernet switches ⇐ ⇒ Grid & Cloud

Photonic transport ⇐ ⇒ GPU's

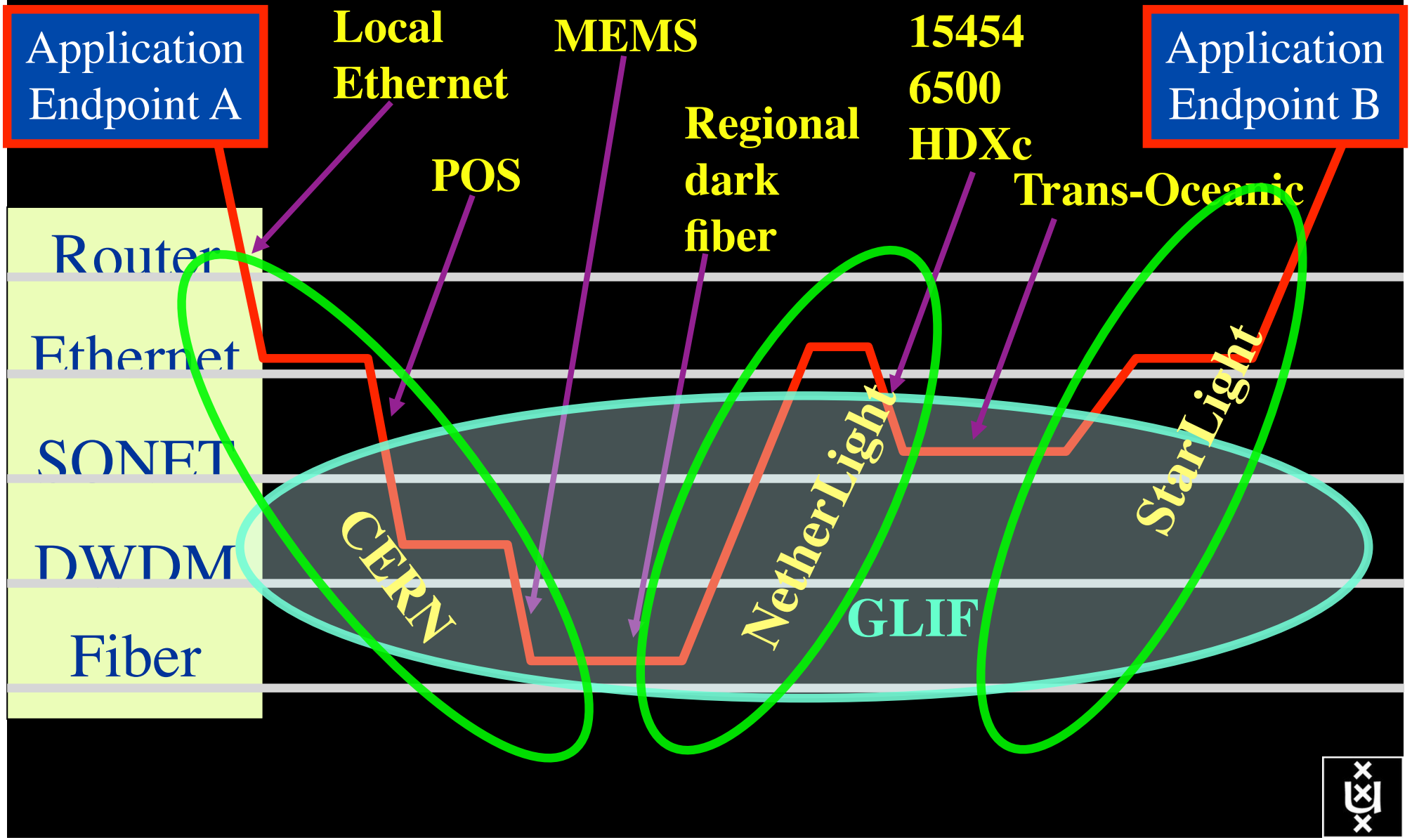
What matters:

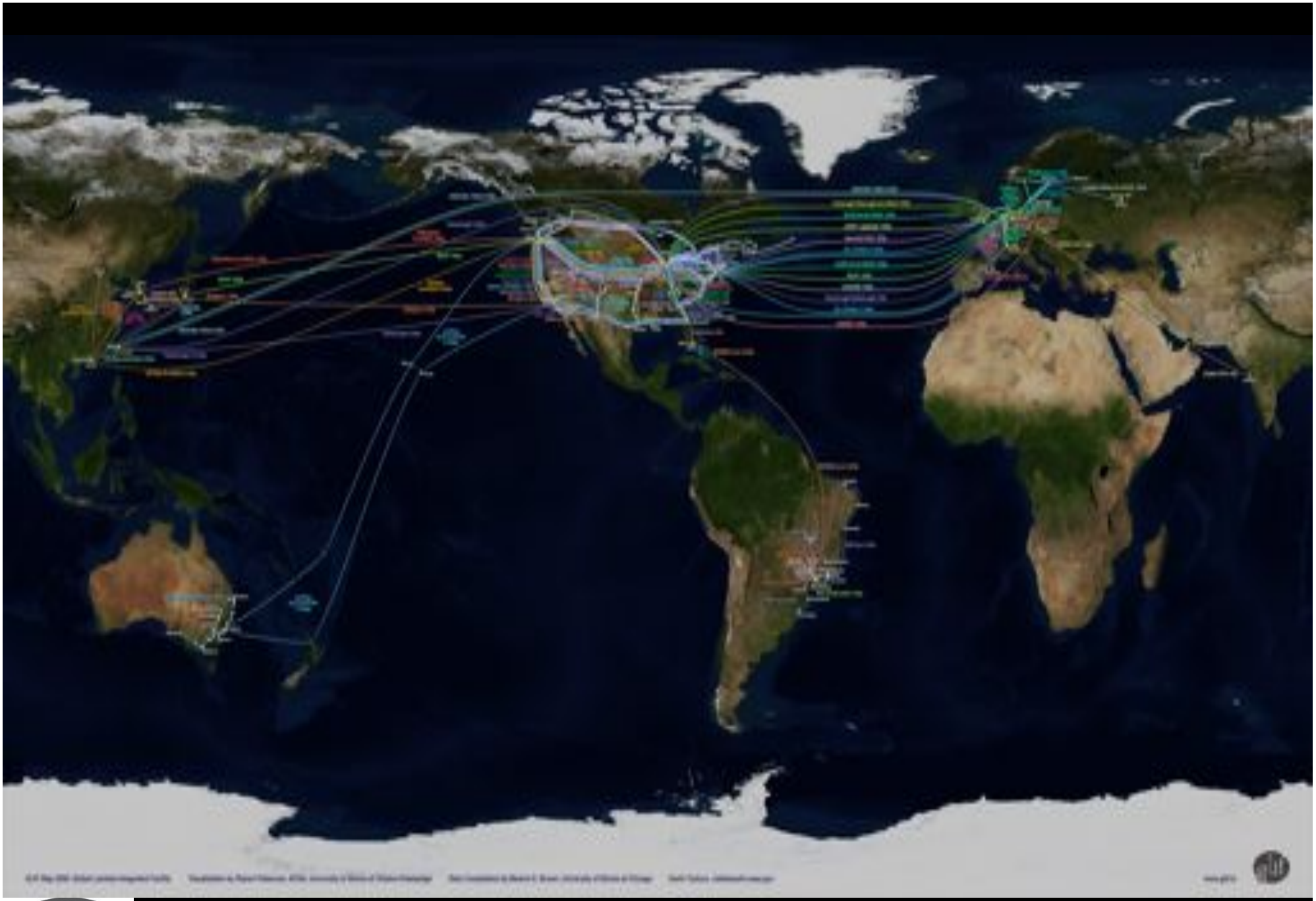
Energy consumption/multiplication

Energy consumption/bit transported



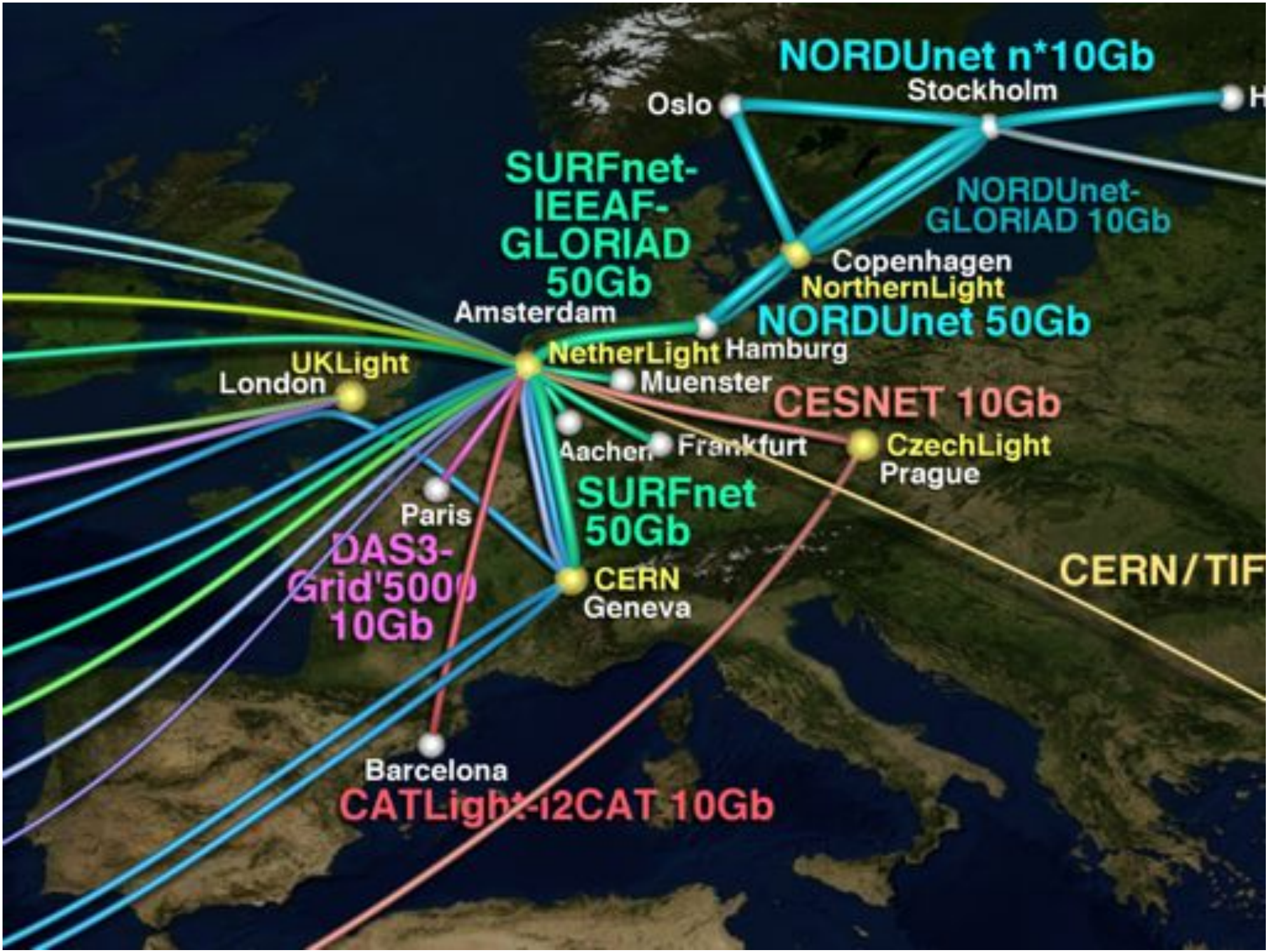
How low can you go?

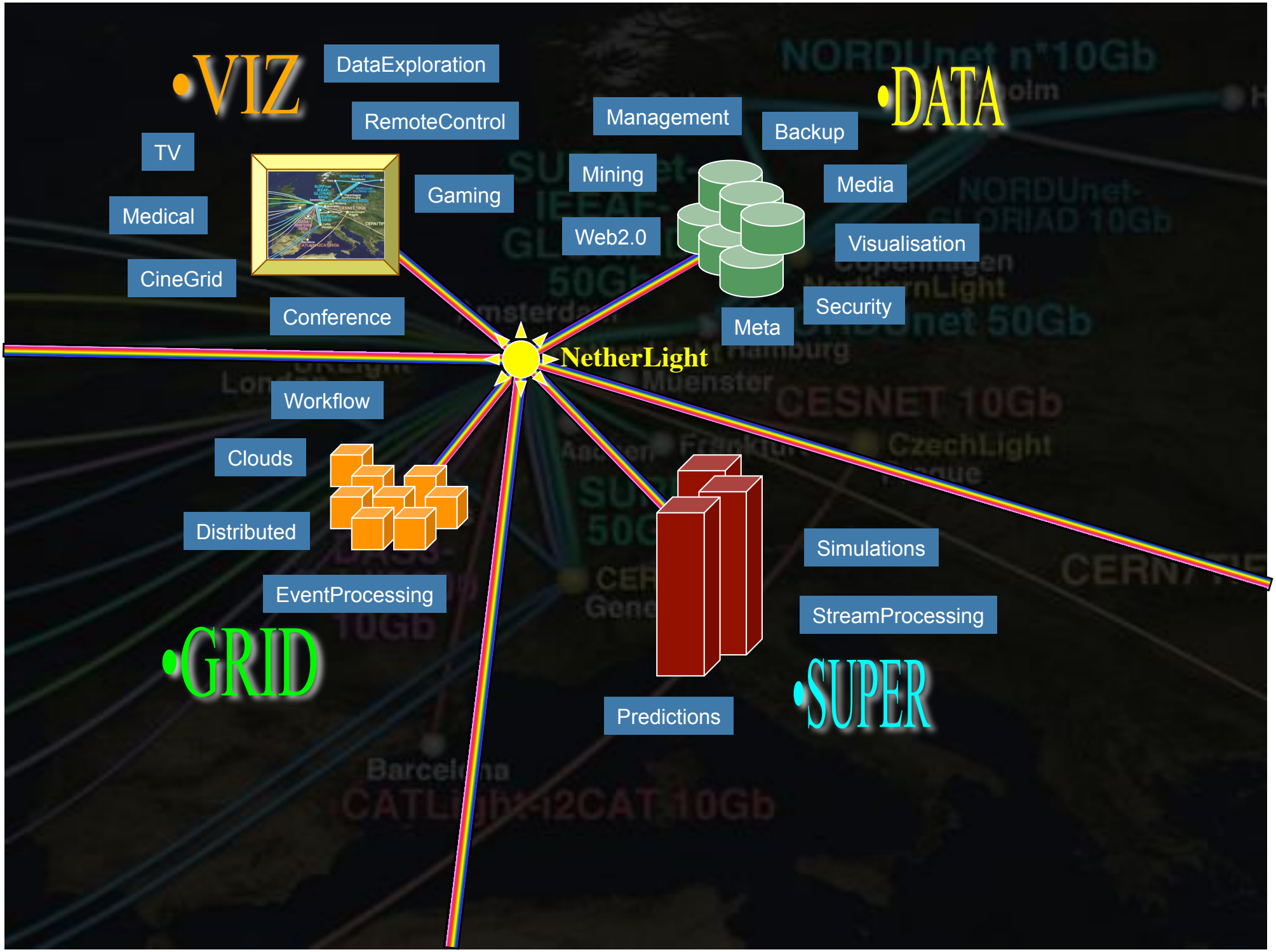




GLIF 2008

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


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

In The Netherlands SURFnet connects between 180:

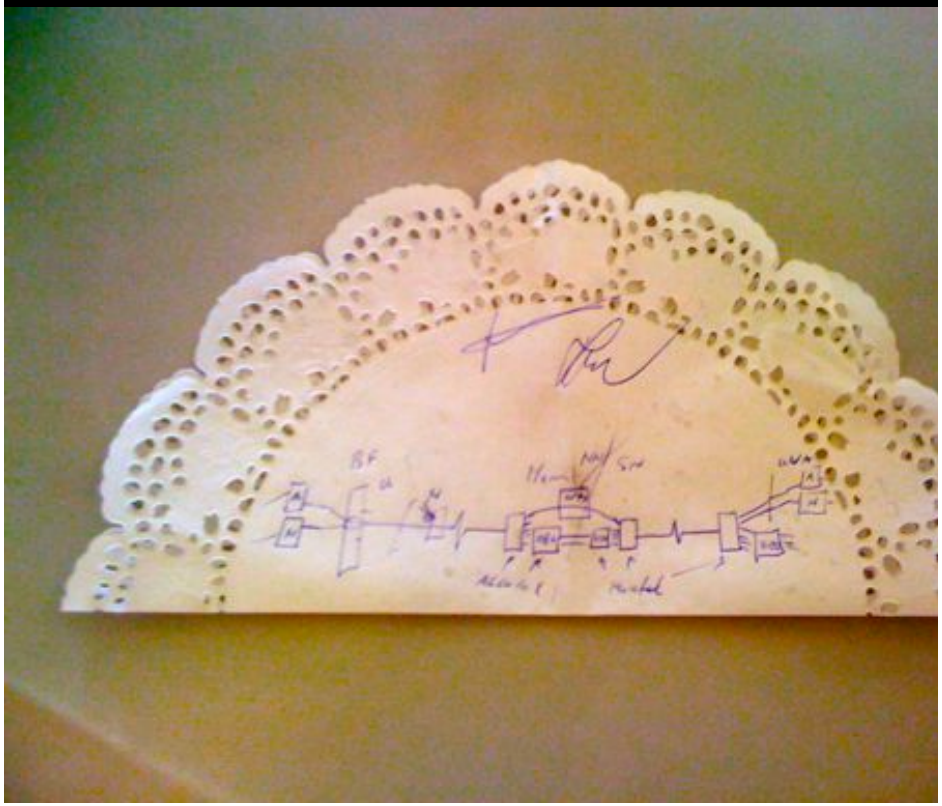
- universities;
- academic hospitals;
- most polytechnics;
- research centers.

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

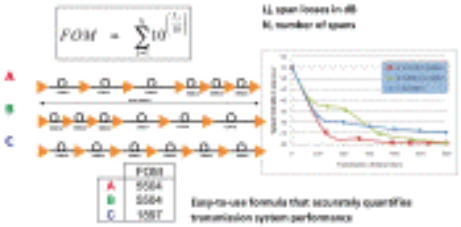
Alien wavelength challenges

- 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.

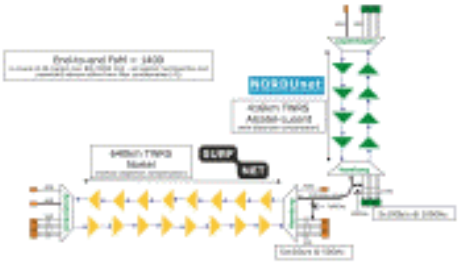
New method to present fiber link quality, FoM (Figure of Merit)

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 SURF NET

REFERENCES

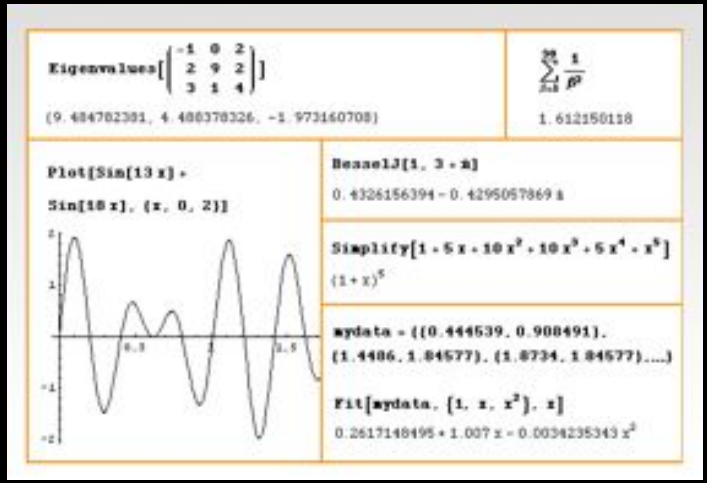
[1] "OPERATIONAL SOLUTIONS FOR AN OPEN DWDM LAYER", D. GRANLUND ET AL., OFC 2009 | [2] "AT&T OPTICAL TRANSPORT SERVICES", BARBARA E. SMITH, OFC/NFOEC 2008 | [3] "OPEN DWDMS OF ALL-OPTICAL CORE NETWORKS", ANDREW LORD AND GIRI ENGINEER, ECOC 2009 | [4] "SURFNET/SURFNET OPTIMAL COMMUNICATION"

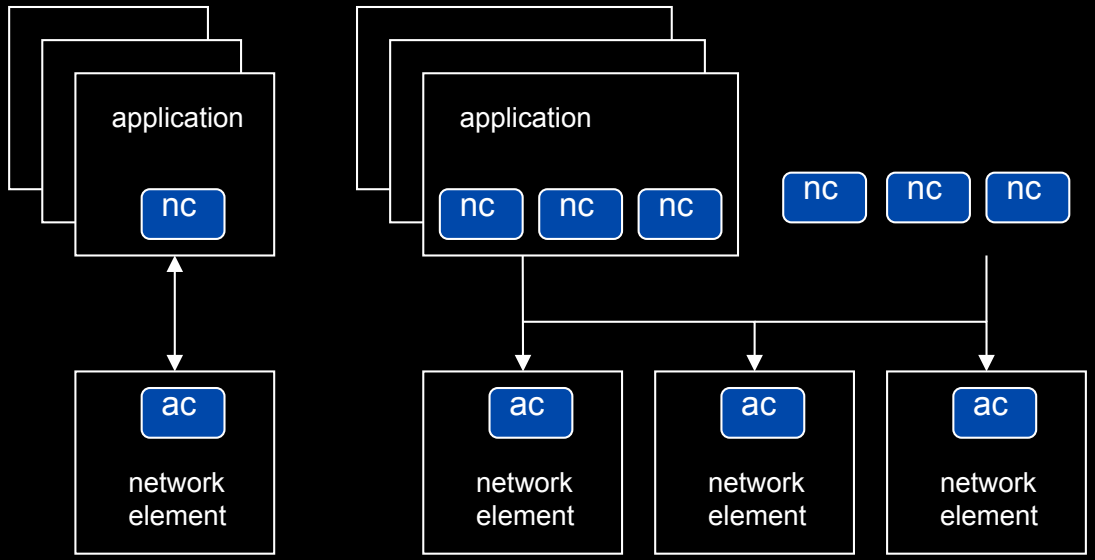
ACKNOWLEDGEMENTS

WE ARE GRATEFUL TO NORDUNET FOR PROVIDING US WITH BANDWIDTH ON THEIR OWN LINK FOR THIS EXPERIMENT AND ALSO FOR THEIR SUPPORT AND ASSISTANCE DURING THE EXPERIMENT. WE ALSO ACKNOWLEDGE TELINDUS AND NORTEL FOR THEIR INTEGRATION SERVICES AND SIMILAR SUPPORT.

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





TouchTable Demonstration @ SC08

Interactive programmable networks

Applications and Networks become aware of each other!

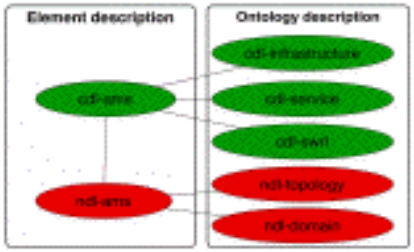
CineGrid Description Language

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.

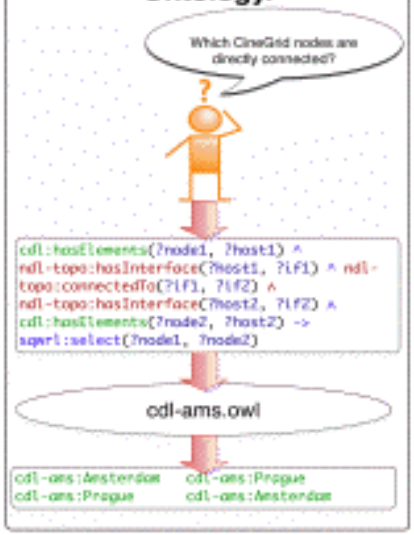
CDL has bindings to the NDL ontology that enables descriptions of network components and their interconnections.

With CDL we can reason on the CineGrid infrastructure and its services.



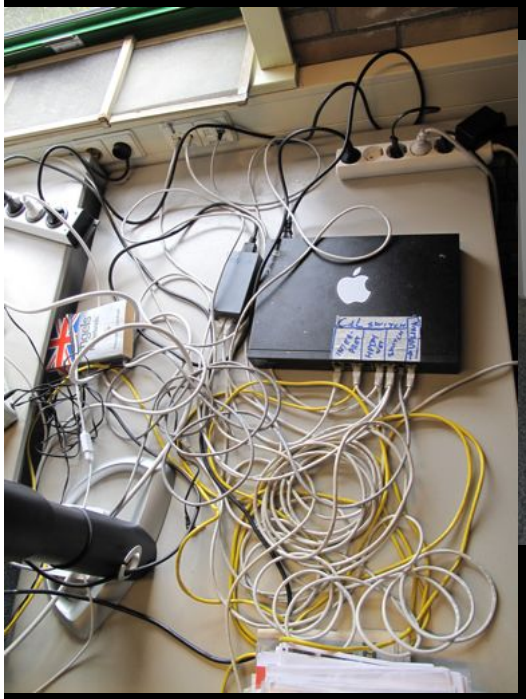
UML representation of CDL

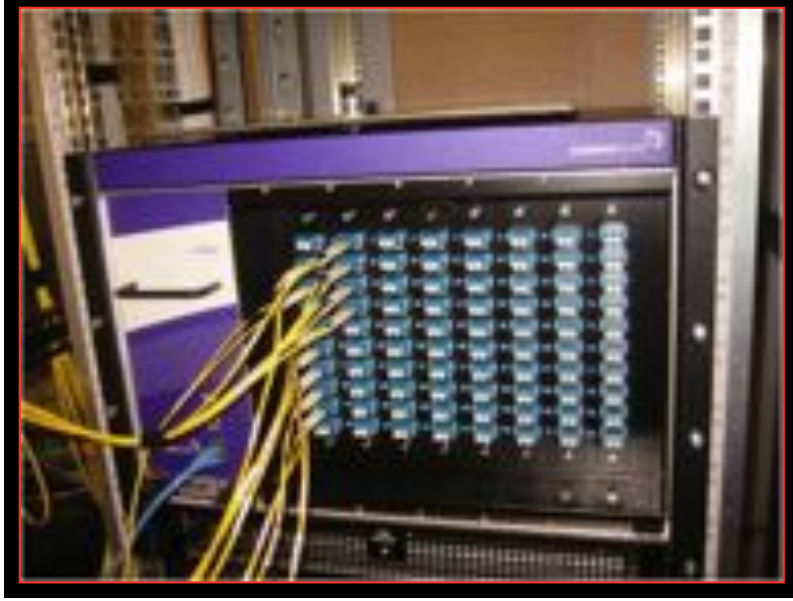
SQWRL is used to query the Ontology.



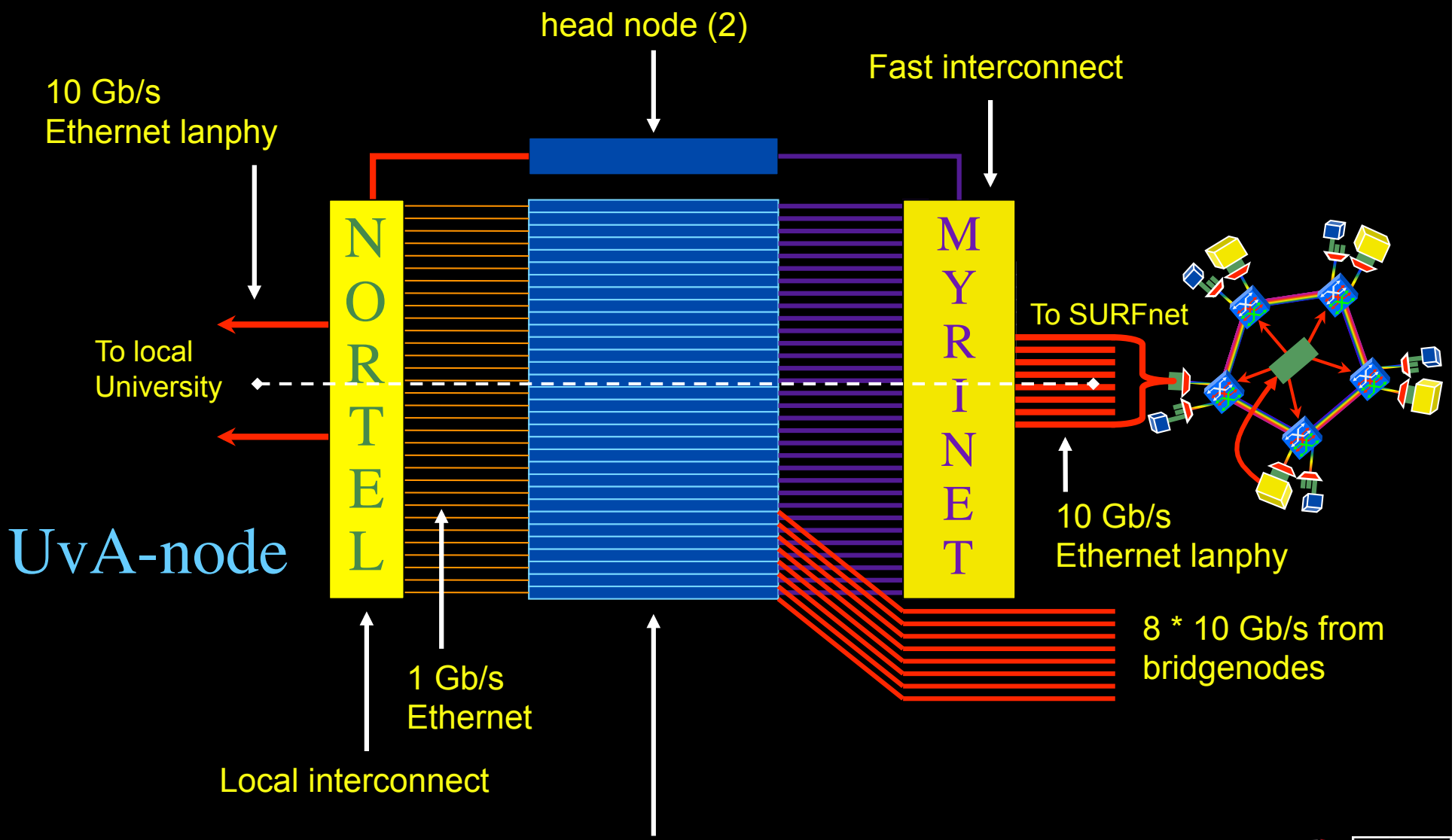
CDL links to ***NDL*** using the ***owl:SameAs*** property. ***CDL*** defines the services, ***NDL*** the network interfaces and links. The combination of the two ontologies identifies the host pairs that support matching services via existing network connections.

Our Christmas Trees ☺



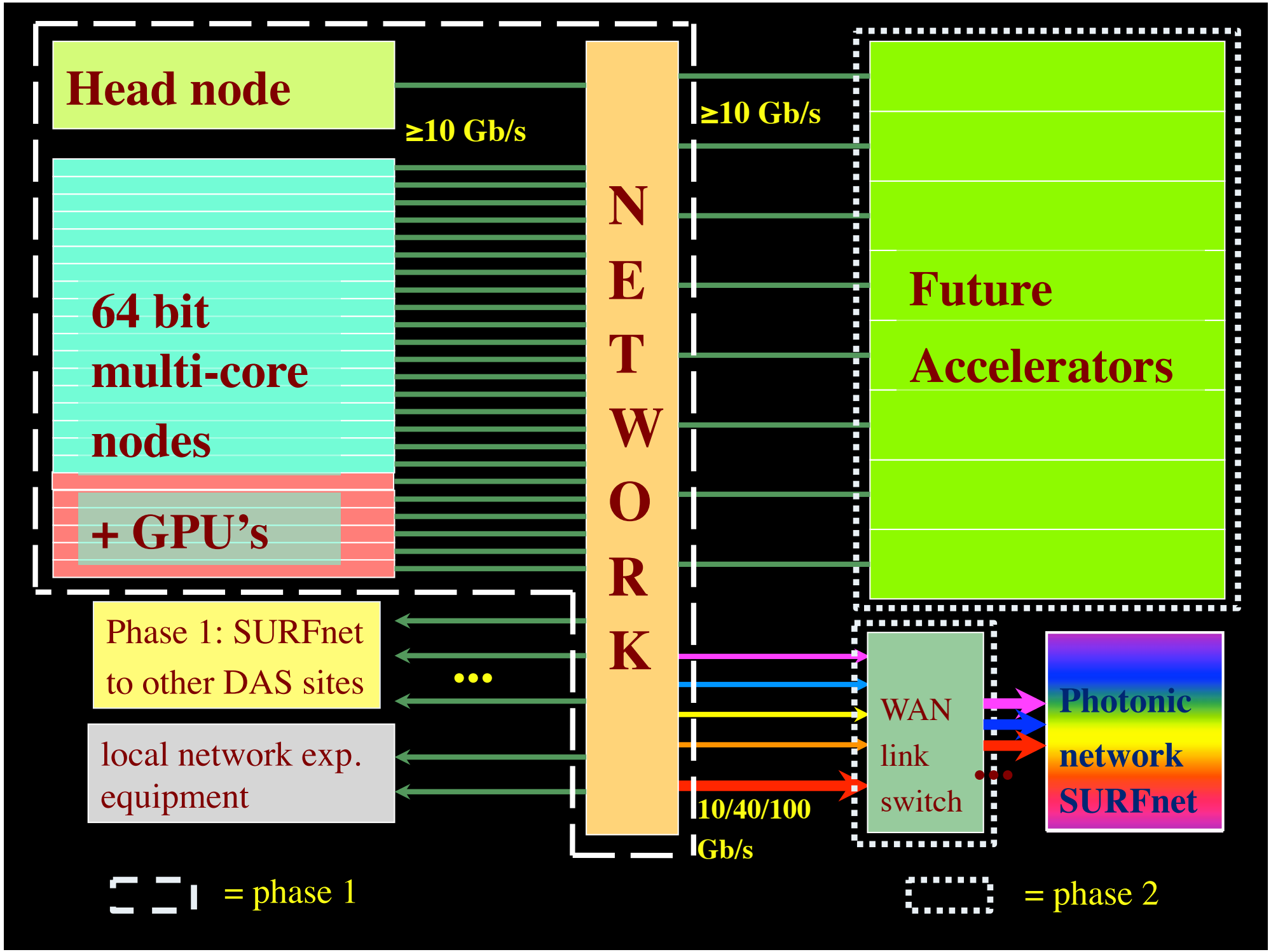


DAS-3 Cluster Architecture



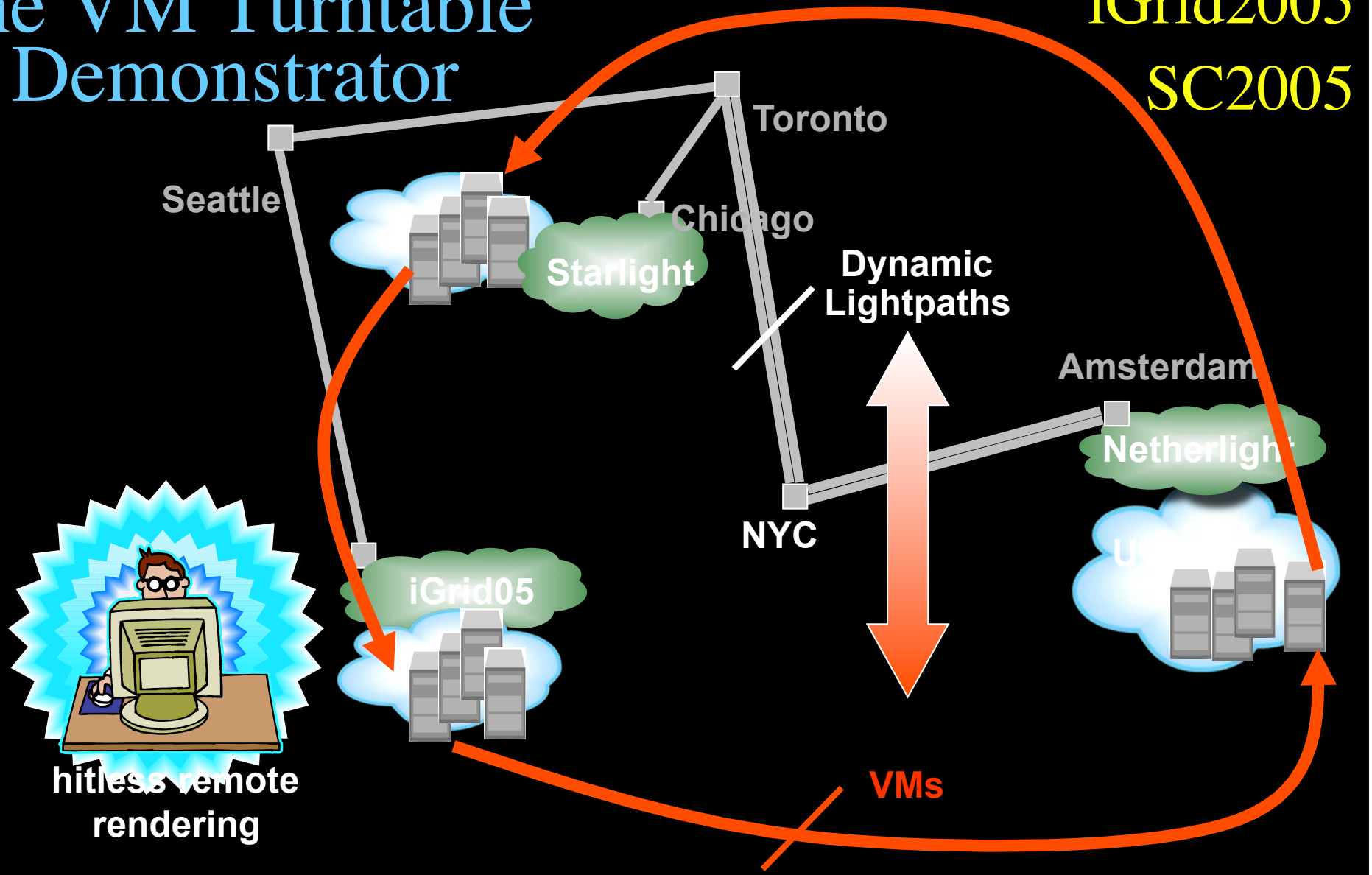
Advanced School for Computing and Imaging

DAS-3


Head node
≥10 Gb/s
64 bit
multi-core
nodes
+ GPU's
N
E
T
W
O
R
K
≥10 Gb/s
Future
Accelerators
Phase 1: SURFnet
to other DAS sites
local network exp.
equipment
WAN
link
switch
10/40/100
Gb/s
Photonic
network
SURFnet
= phase 1
= phase 2

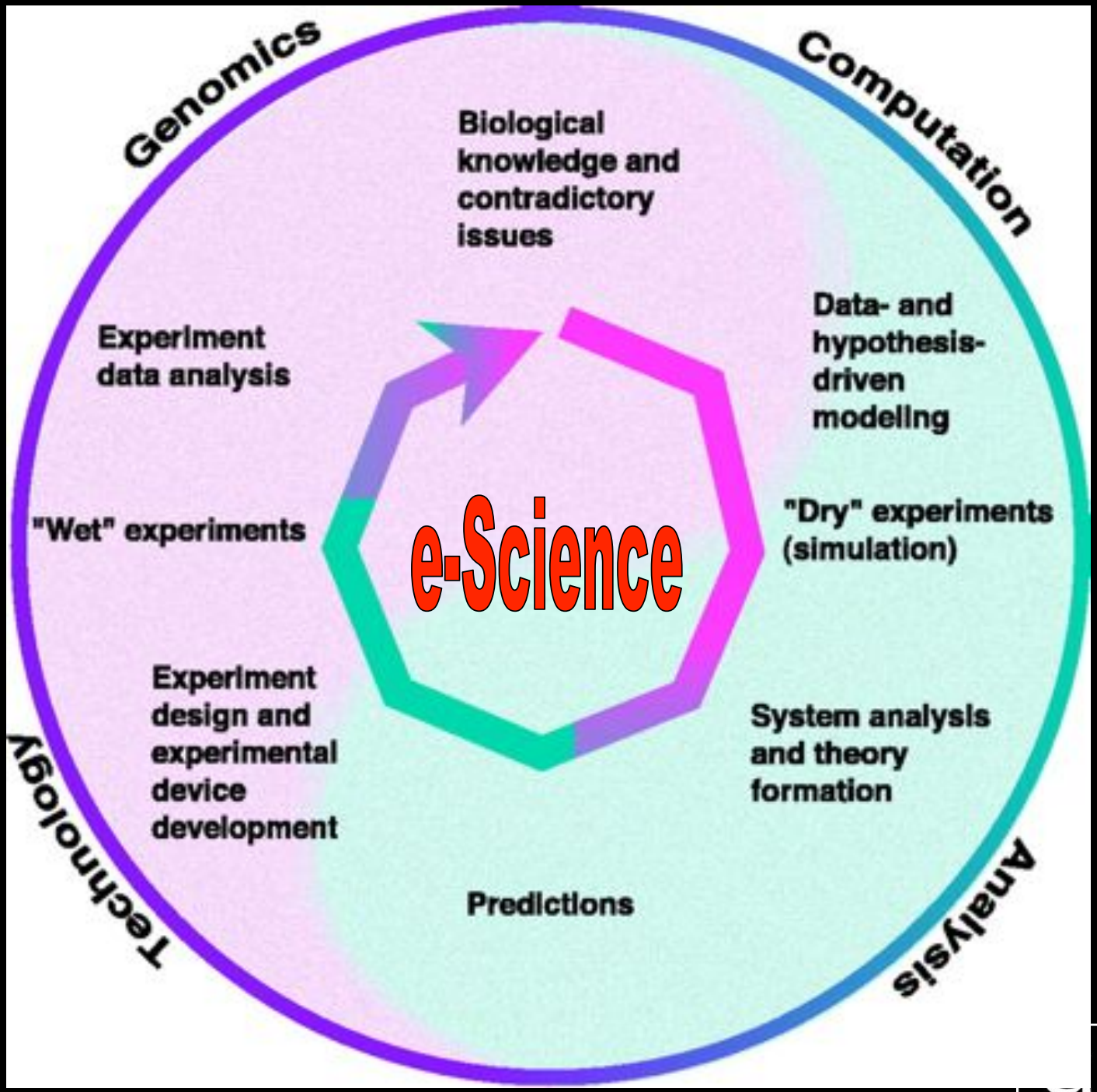
The VM Turntable Demonstrator

iGrid2005
SC2005



The VMs that are live-migrated run an iterative search-refine-search workflow against data stored in different databases at the various locations. A user in San Diego gets hitless rendering of search progress as VMs spin around

e-Science toegepast in biologie



Waarom e-Science?

- ICT probleem voor wetenschap, Industrie en maatschappij
  - Er wordt heel veel data verzameld
    - LOFAR, CERN, Life Sciences, Earth sciences, etc
  - Er wordt heel veel data gegenereerd
    - Klimaatmodellen, watermanagement, drugdesign, etc
    - Simulaties van ‘catastrofes’ die niet experimenteel getoetst kunnen worden i.v.m. Veiligheid, complexiteit, capaciteit en/of kosten
  - Nieuwe vormen van wetenschap
    - Interdisciplinair: Impact van High Troughput techniques (sequencing) in biologie, MRI in medische wetenschappen en farmacie
    - Science 2.0
- Nu Inefficiënt: ieder domein (biologen, chemici, etc.) vindt het wiel opnieuw uit.
  - Geen kennisoverdracht op gebied van techniek en methoden
  - Te beperkte interdisciplinair onderzoek
  - Te weinig professionele ICT support waardoor er kostbare ‘science’ tijd verloren gaat
  - Door te weinig gebruik van ICT in onderzoek kunnen we, op een aantal gebieden, de aansluiting bij de wereld gaan missen.

Problemen

- Coördinatie activiteiten
- Te weinig interdisciplinaire samenwerking
- Gekwalificeerd personeel
- Financiering
  - NWO etc zijn klassiek discipline georganiseerd
  - Nationale infrastructuur vereist nationale fondsen
  - Structureel geld en niet alleen impuls (core)
- Status: Science of Engineering?
- Mensen, met kennis van Science en ICT.
  - > Zijn belangrijker dan Computers!

Themes for next years

- 40 and 100 Gbit/s
- Network modeling and simulation
- **Cross domain Alien Light switching**
- **GreenLight - GreenSonar**
- **Network and infrastructure descriptions & WEB2.0**
- **Reasoning about services**
- Cloud Data - Computing
- Web Services based Authorization
- Network Services Interface (N-S and E-W)
- Fault tolerance, Fault isolation, **Monitoring**
- eScience integrated services
- Data and Media specific services

➔ **Smart e-Infrastructure**

Progress

- Kilobit/s ⇐ ⇒ keyboard
- Megabit/s ⇐ ⇒ process ques / rpc’s
- Gigabit/s ⇐ ⇒ discs, screens
- Terabit/s ⇐ ⇒ GPU

Onwards!

- We aim for extreme [comp,data,net,viz] experiments!
- Computer & Computational & e- Science needs open and unrestricted environments for experimentation with ICT!
- Our laboratory must be very well connected!
- Our laboratory must be easy accessible and nearby!
- Participation in master education.
- We need participation from IC to build & operate & utilize our laboratory!
- We need to be able to break things!

If we never broke something we did not try hard enough!

Networks

- Make sure we couple to the SURFnet hybrid services
  - Routed plus lightpaths implemented as NGE
  - Go to 100 Gb/s in next 5 years
- Create a scalable dwdm photonics layer
- Limit routers to the minimum (2)
- Use some form of Layer2 for traffic transport& aggregation
  - MPLS-TE or PBT, PLSB, etc.
- Make sure it is controllable & manageable
  - and ultimately partly by users & applications

Questions ?

Thanks: Paola Grosso & Henri Bal & Hans Dijkman & Bob Hertzberger
& Jeroen vd Ham & Freek Dijkstra & team for several of the slides.