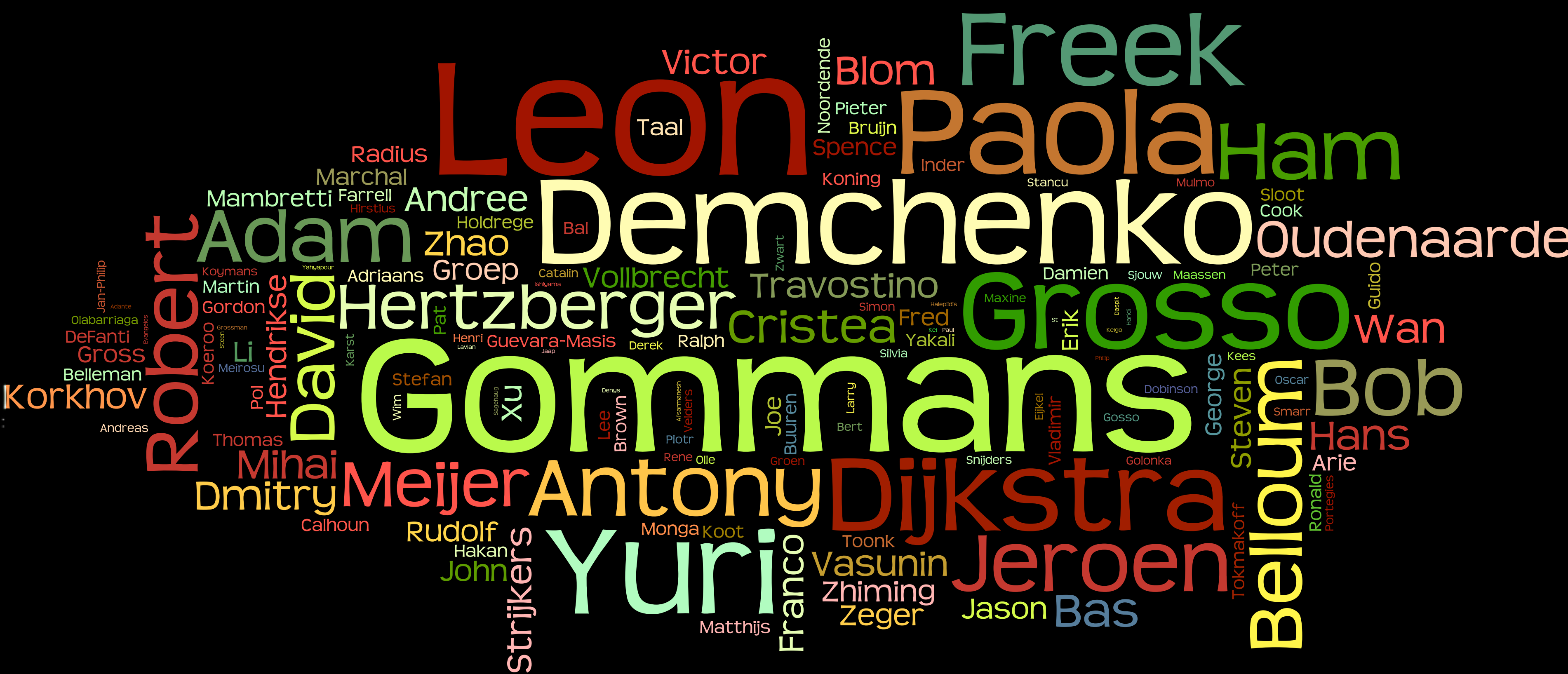
System and Network Engineering Research for Big Data Sciences

Cees de Laat



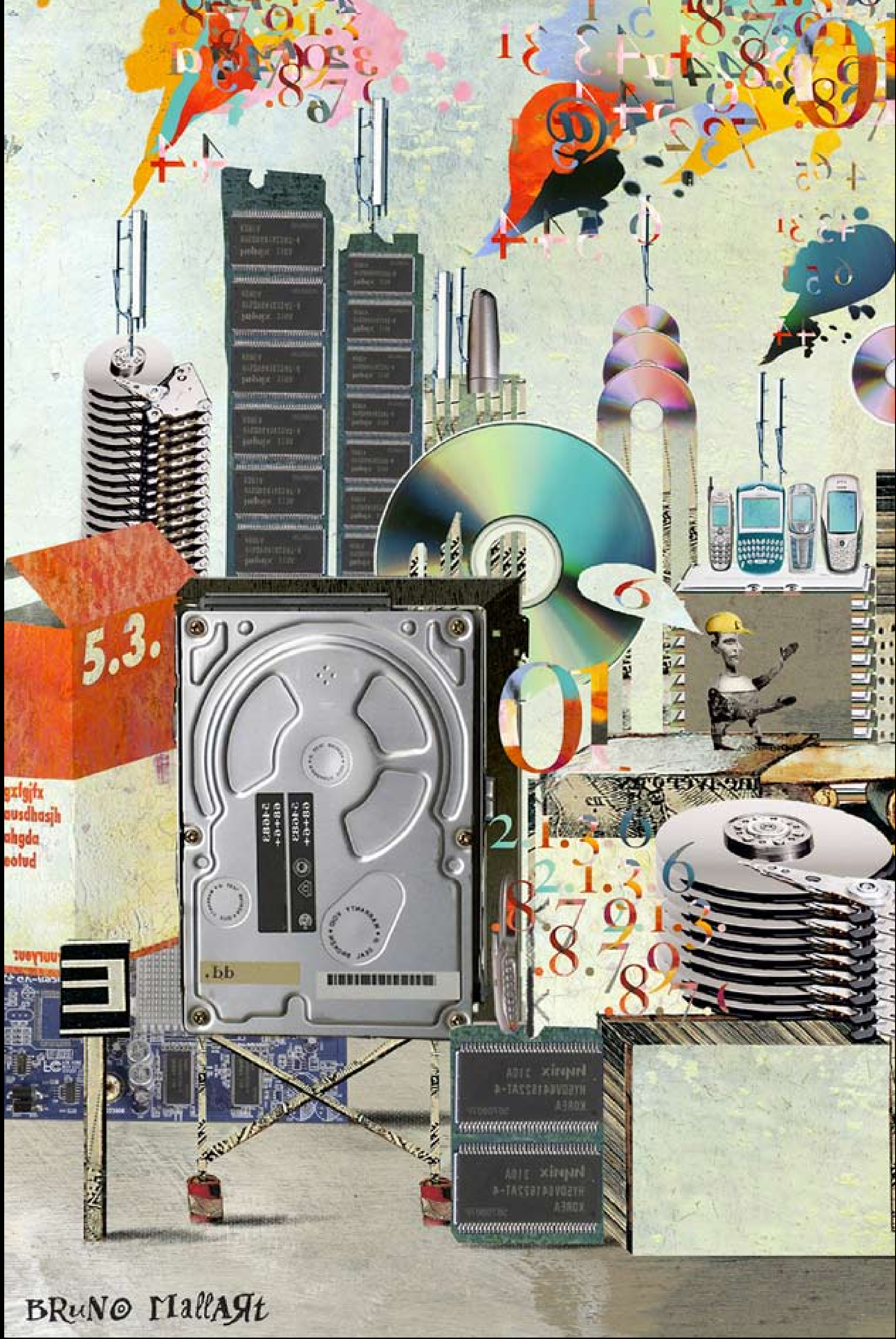
From King's Dutch Academy of Sciences

The Dutch Research Agenda

"Information technology (IT) now permeates all aspects of public, commercial, social, and personal life. bank cards, satnav, and weather radar... IT has become completely indispensable."

"But to guarantee the reliability and quality of constantly bigger and more complicated IT, we will need to find answers to some fundamental questions!"

http://www.knaw.nl/Content/Internet_KNAW/publicaties/pdf/20111029.pdf



Reduction of Complexity by Integration

By combining services such as telephony, television, data, and computing capacity within a single network, we can cut down on complexity, energy consumption and maintenance.

- How can we describe and analyze complex information systems effectively?
- How can we specify and measure the quality and reliability of a system?
- How can we combine various different systems?
- How can we design systems in which separate processors can co-operate efficiently via mutual network connections within a much larger whole?
- Can we design information systems that can diagnose their own malfunctions and perhaps even repair them?
- How can we specify, predict, and measure system performance as effectively as possible?

SNE addresses a.o. the **highlighted** questions!

http://www.knaw.nl/Content/Internet_KNAW/publicaties/pdf/20111029.pdf

Internet

From a network experiment that never ended (Vint Cerf)

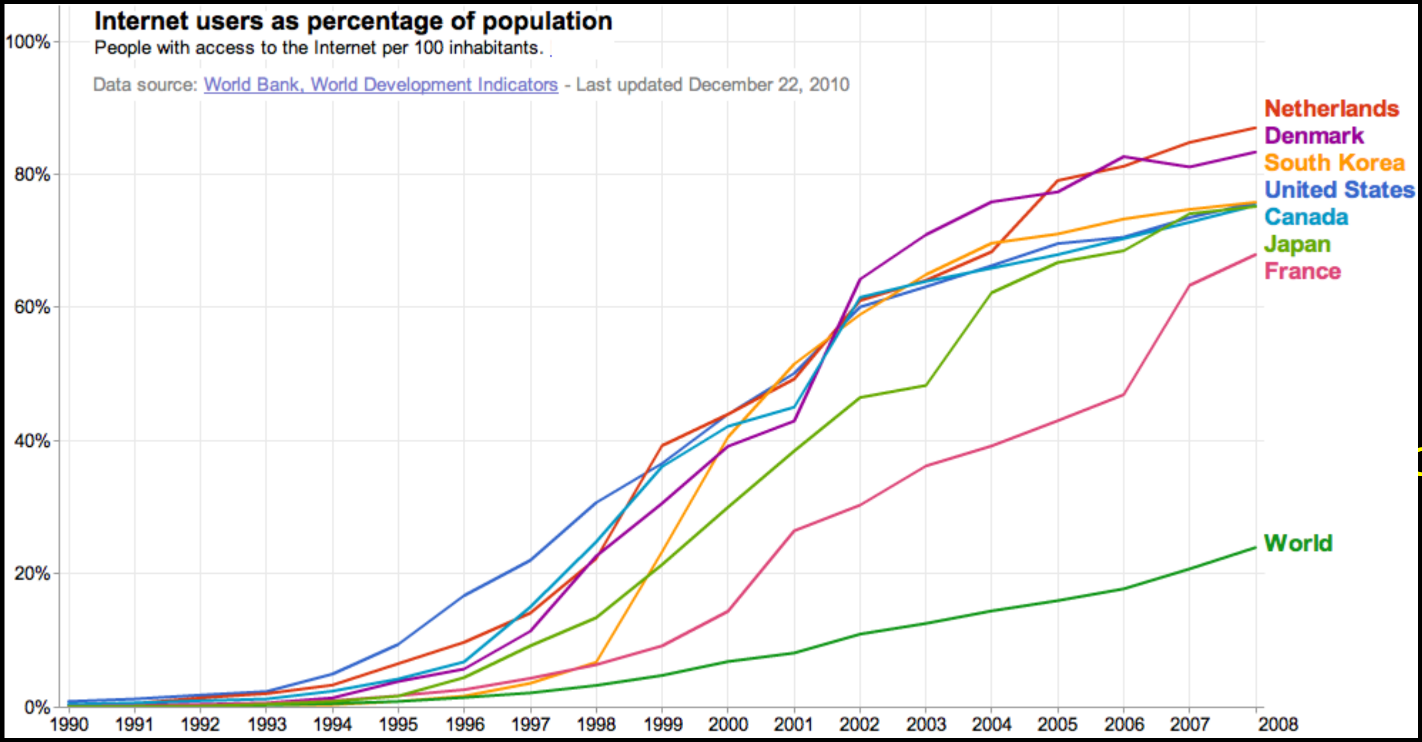
1974: for the first time the word **internet** (*RFC 675 - Specification of Internet Transmission Control Program) [note -> Open process!]*

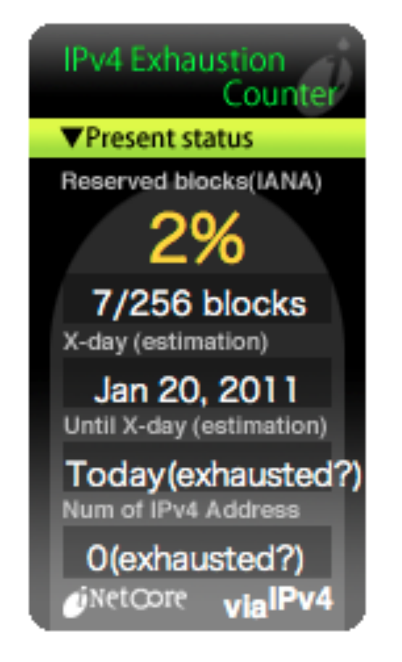
1981: the **TCP/IP** standard was ready to be adopted (*RFC 791,792,793*)

To a network for society

1989: WWW was born

2010



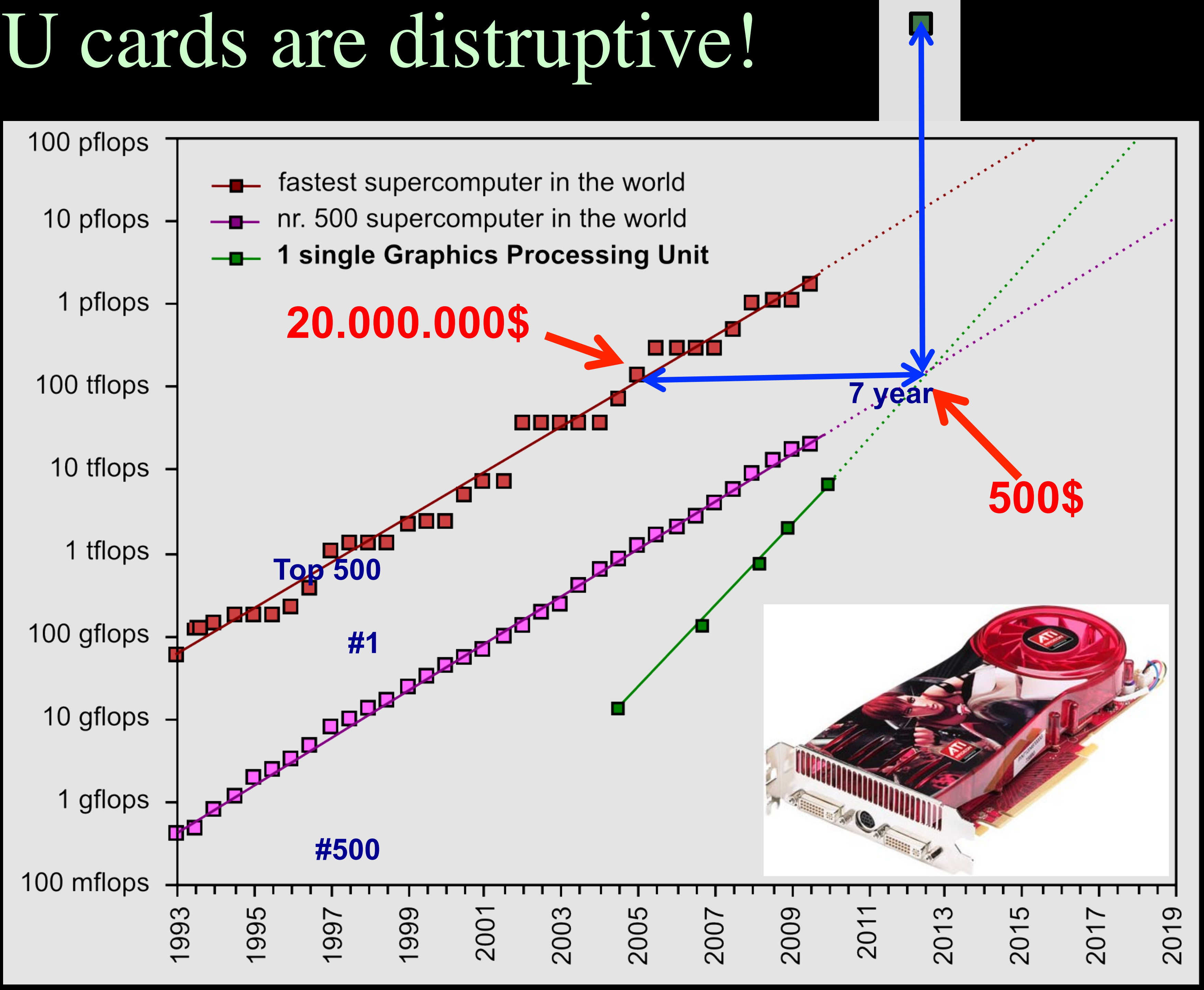


Ipv6day.nl

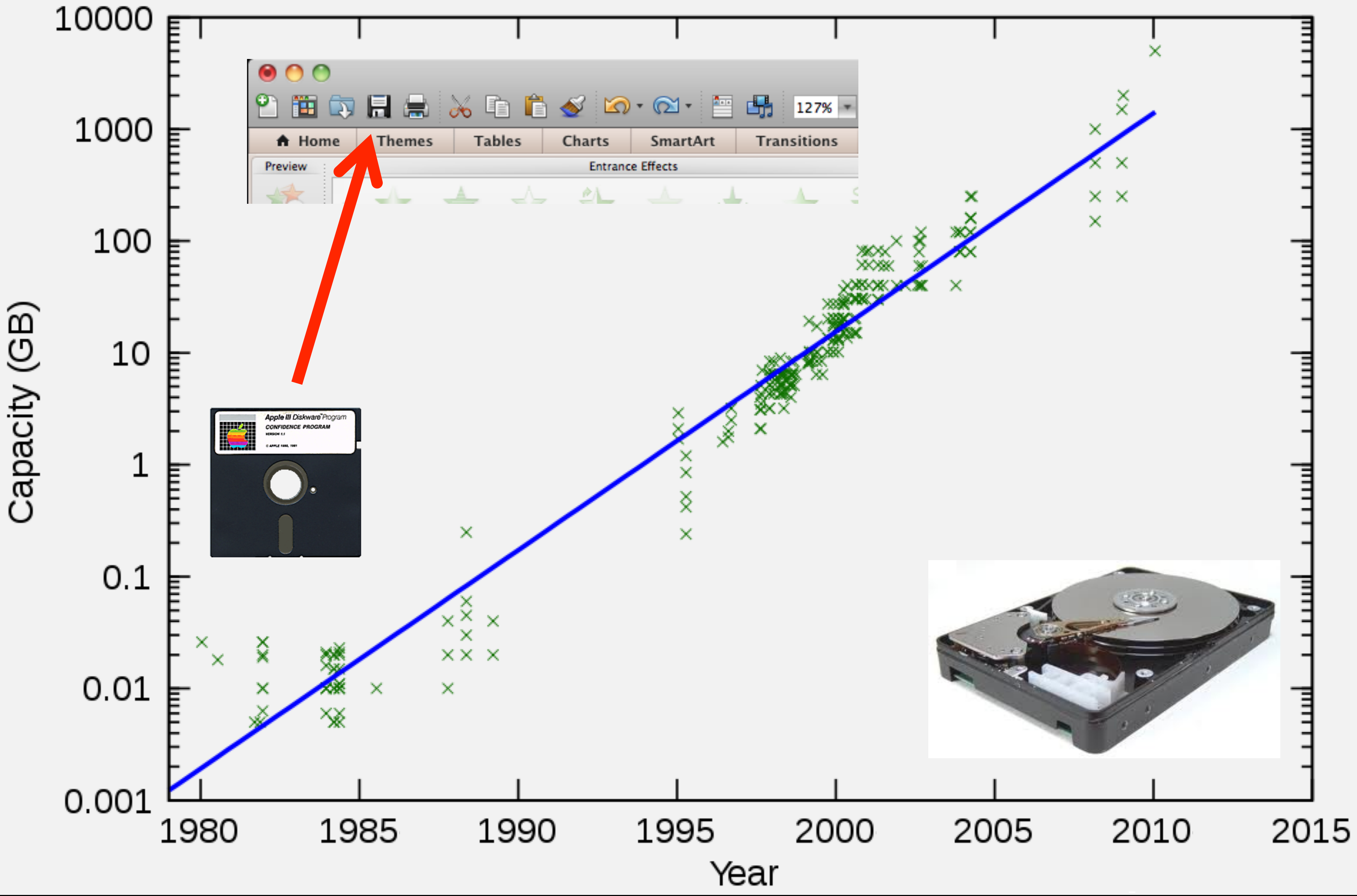

AT&T 3G
9:42 AM
Tuesday
9
SMS
Text
Calendar
Photos
Camera
280
73°
YouTube
Stocks
Maps
Weather
Clock
Calculator
Notes
Settings
iTunes
App Store
Phone
Mail
Safari
iPod


TEMP
RESTART
TRACKER
VERB
NOUN
CLR
ENTR
RSET
+
−
7
8
9
4
5
6
1
2
3

GPU cards are distruptive!



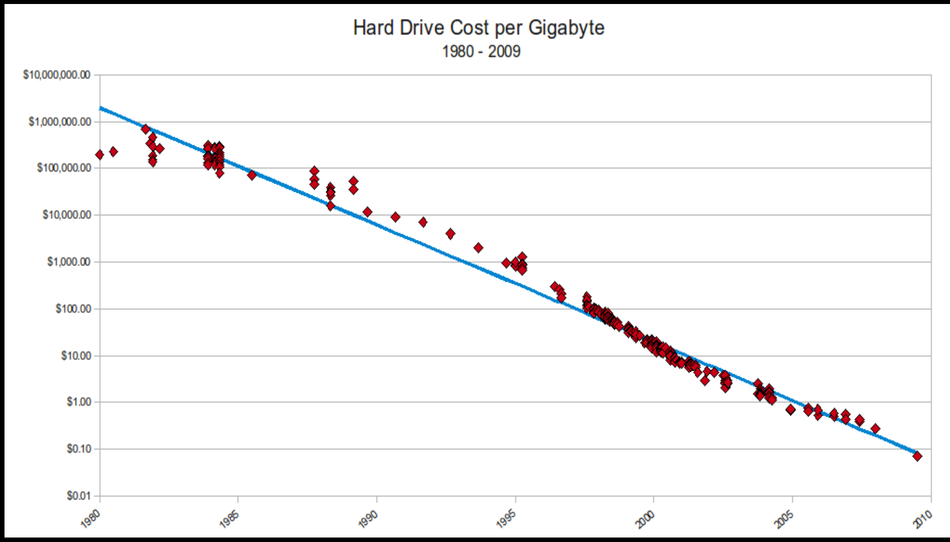
Data storage: doubling every 1.5 year!



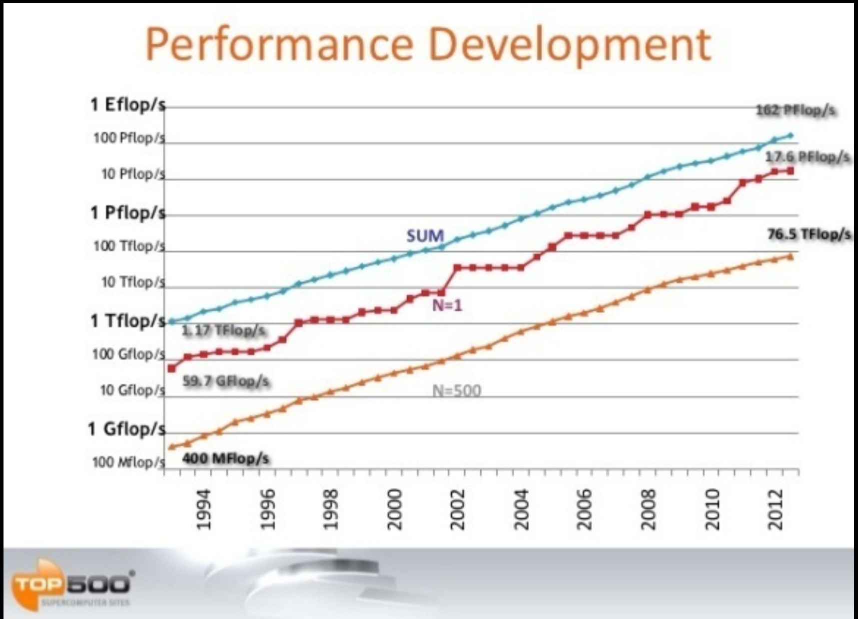
Reliable and Safe!

This omnipresence of IT makes us not only strong but also vulnerable.

- A virus, a hacker, or a system failure can instantly send digital shockwaves around the world.



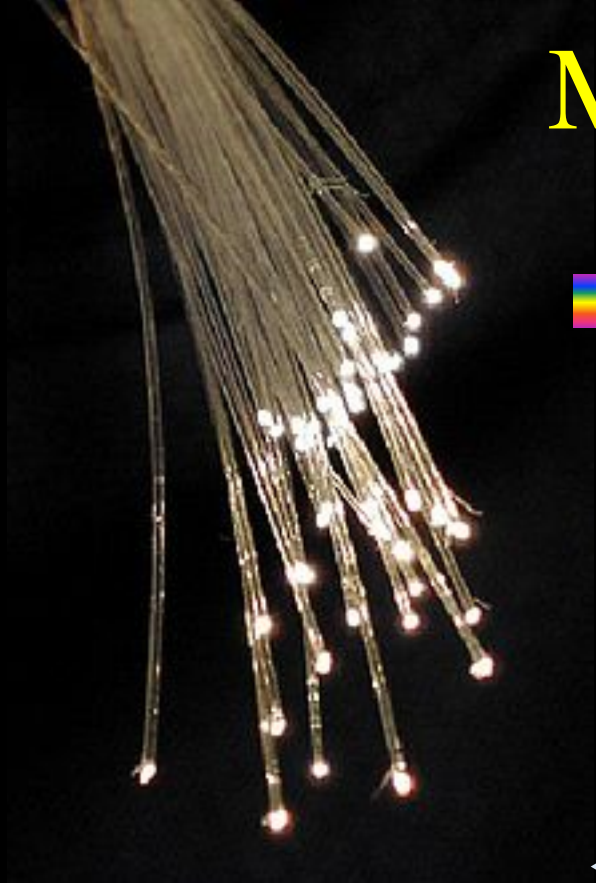
The hardware and software that allow all our systems to operate is becoming bigger and more complex all the time, and the capacity of networks and data storage is increasing by leaps and bounds.

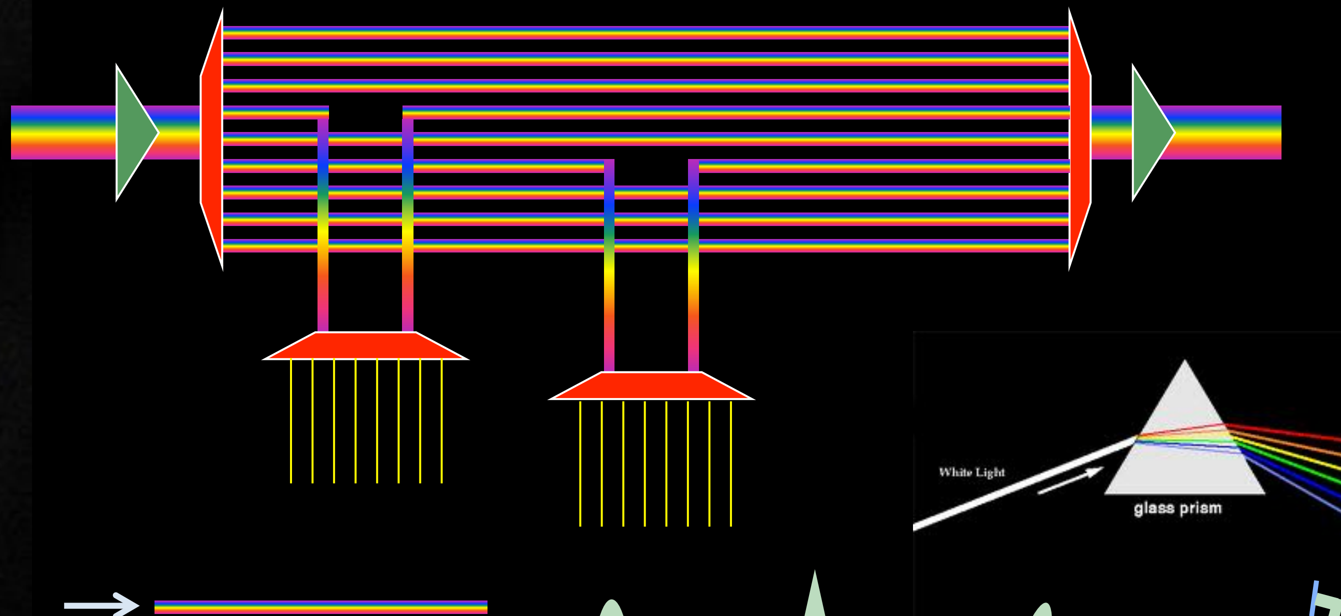


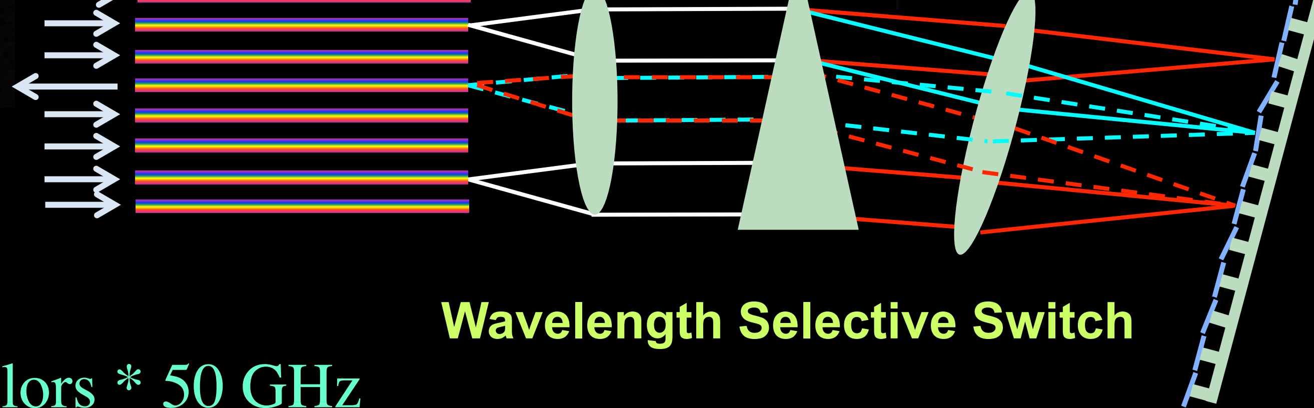
We will soon reach the limits of what is currently feasible and controllable.

http://www.knaw.nl/Content/Internet_KNAW/publicaties/pdf/20111029.pdf

Multiple colors / Fiber







Per fiber: ~ 80-100 colors * 50 GHz

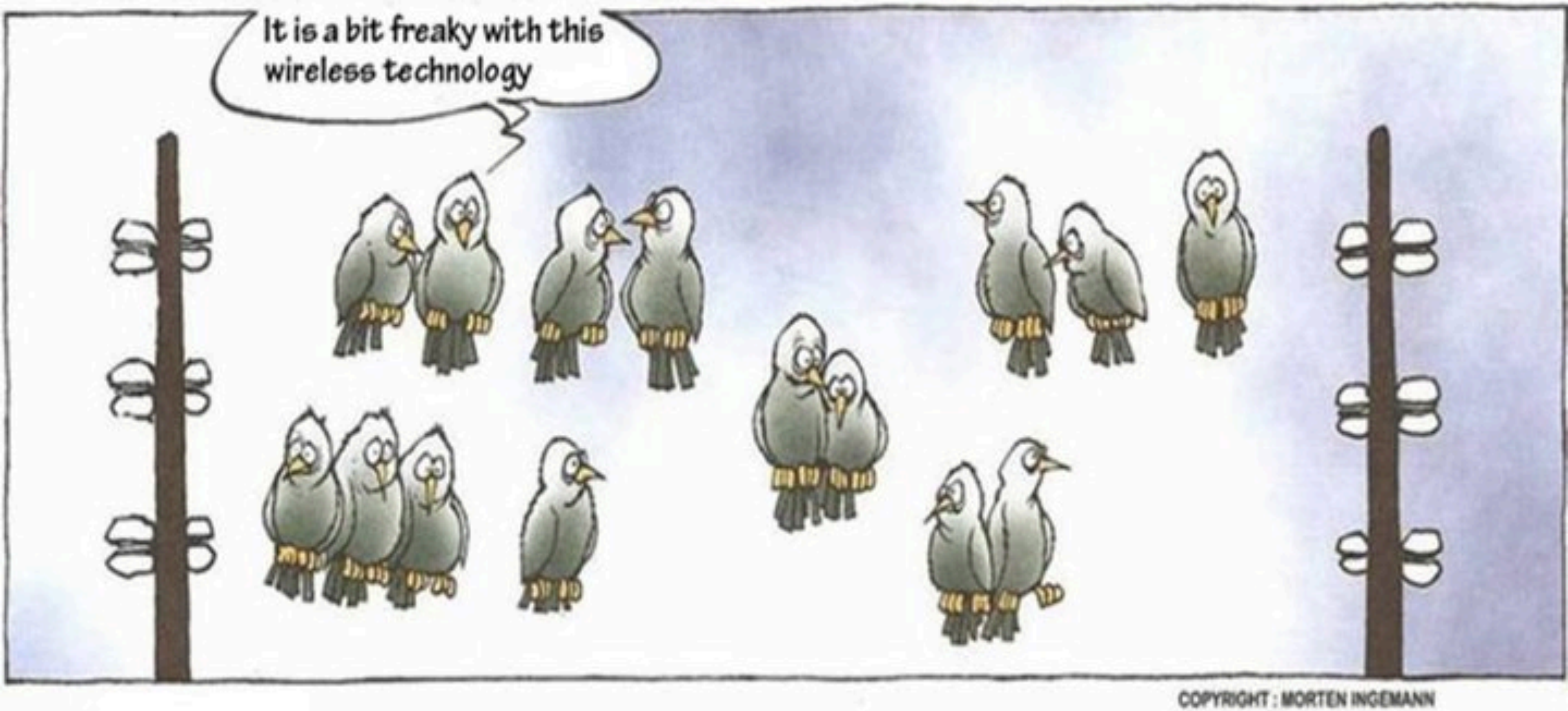
Per color: 10 – 40 – 100 Gbit/s

BW * Distance ~ $2*10^{17}$ bm/s

New: Hollow Fiber!

➔ less RTT!

Wireless Networks



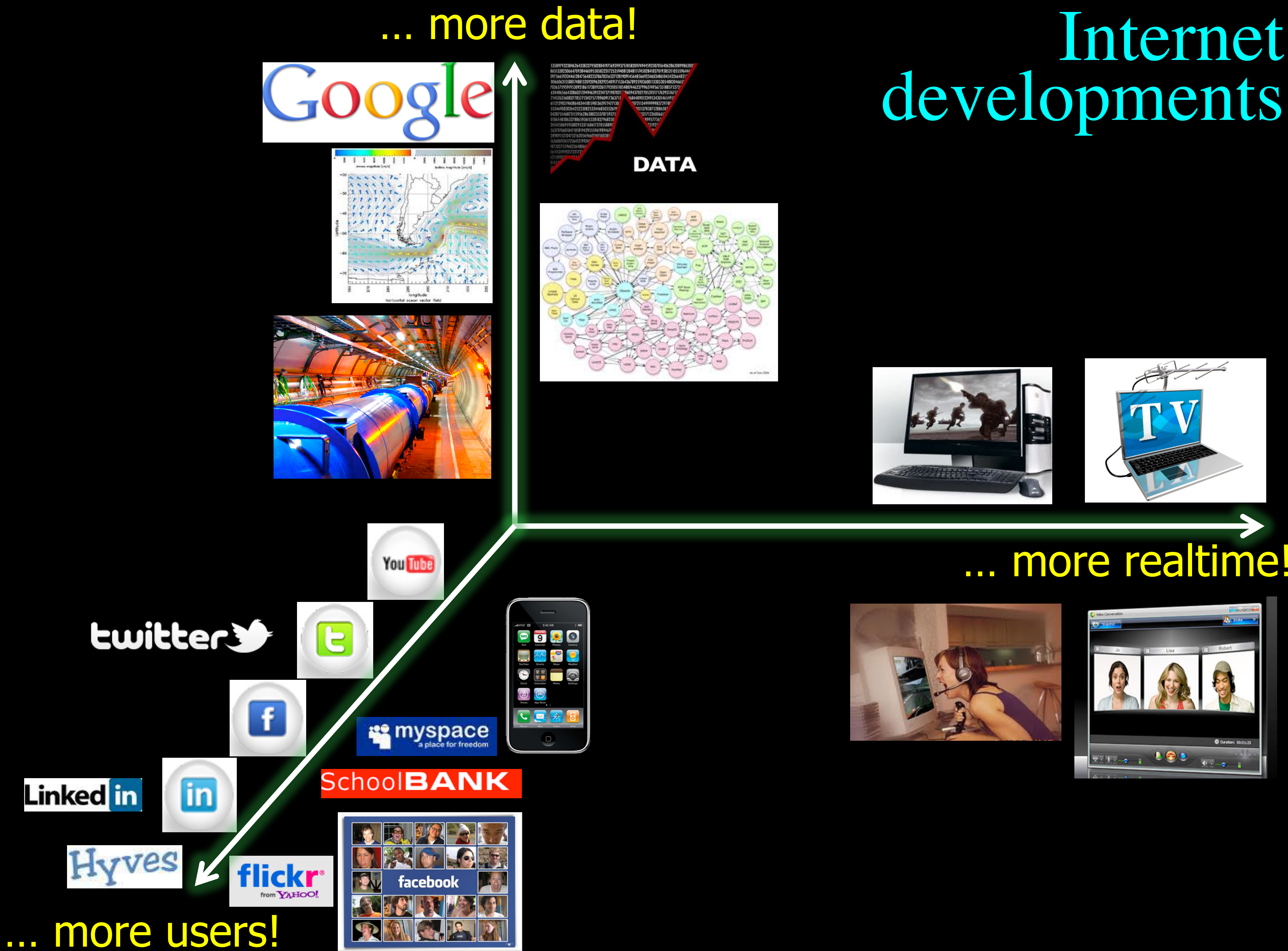
protocol LAN due to the easy comparison and convenience in the digital home. While consumer PC products has just started to migrate to a much higher bandwidth of 802.11n wireless LAN now working on next-generation standard definition is already in progress.

Mission SNE

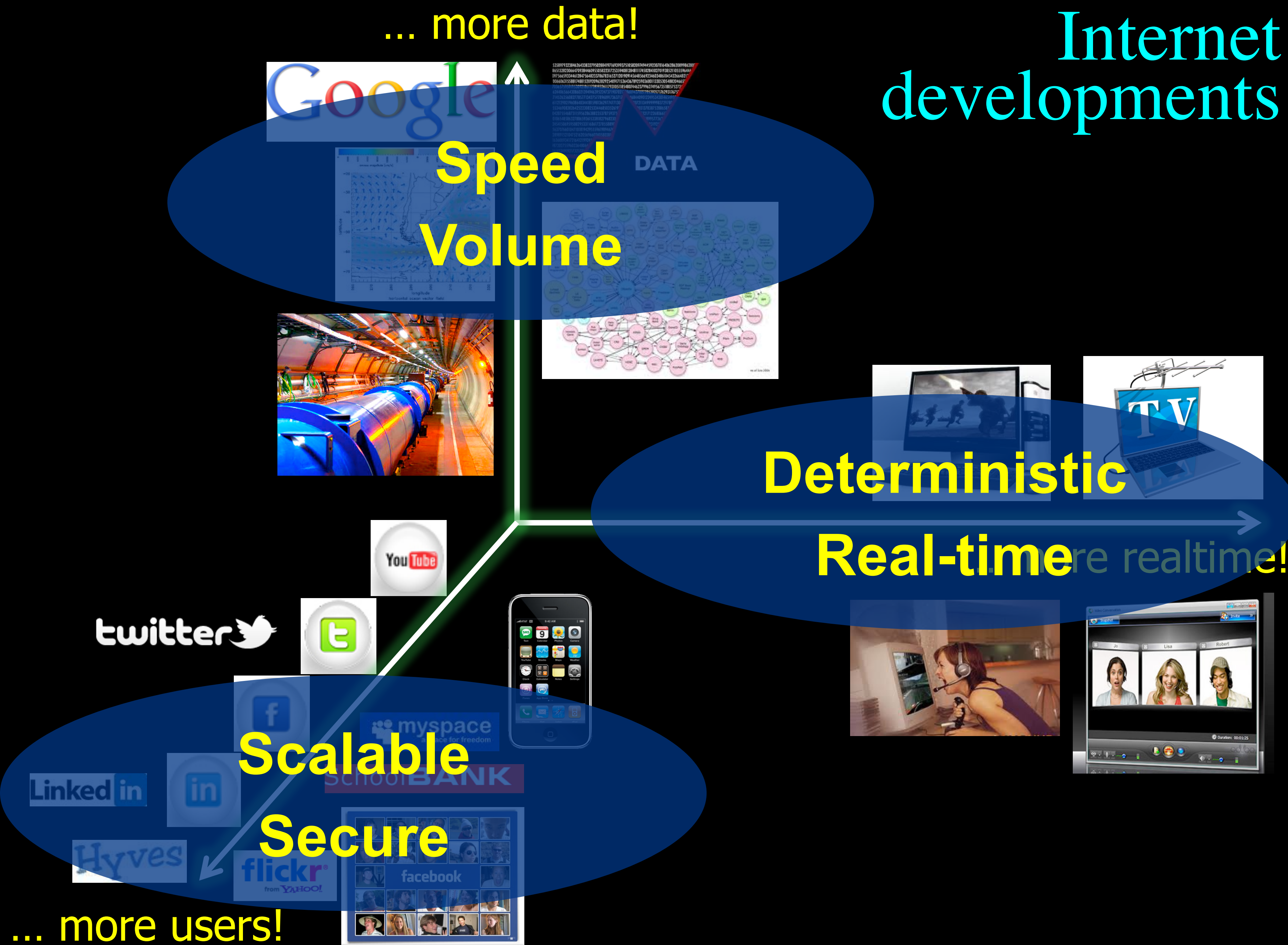
Can we create smart and safe data processing infrastructures that can be tailored to diverse application needs?

- *Capacity*
  - *Bandwidth on demand, QoS, architectures, photonics, performance*
- *Capability*
  - *Programmability, virtualization, complexity, semantics, workflows*
- *Security*
  - *Authorization, Anonymity, integrity of data in distributed data processing*
- *Sustainability*
  - *Greening infrastructure, awareness*
- *Resilience*
  - *Systems under attack, failures, disasters*

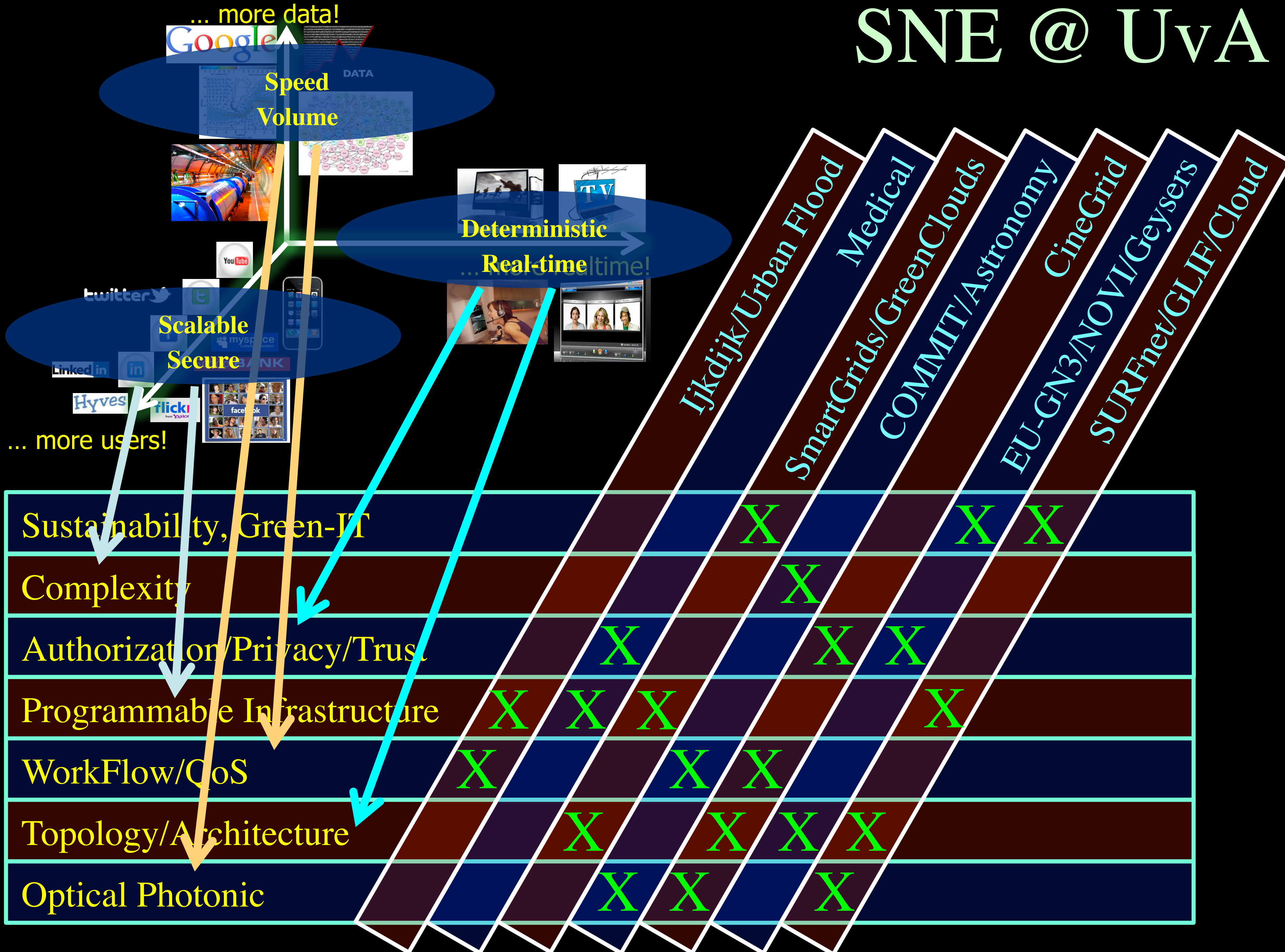
Internet developments



Internet developments

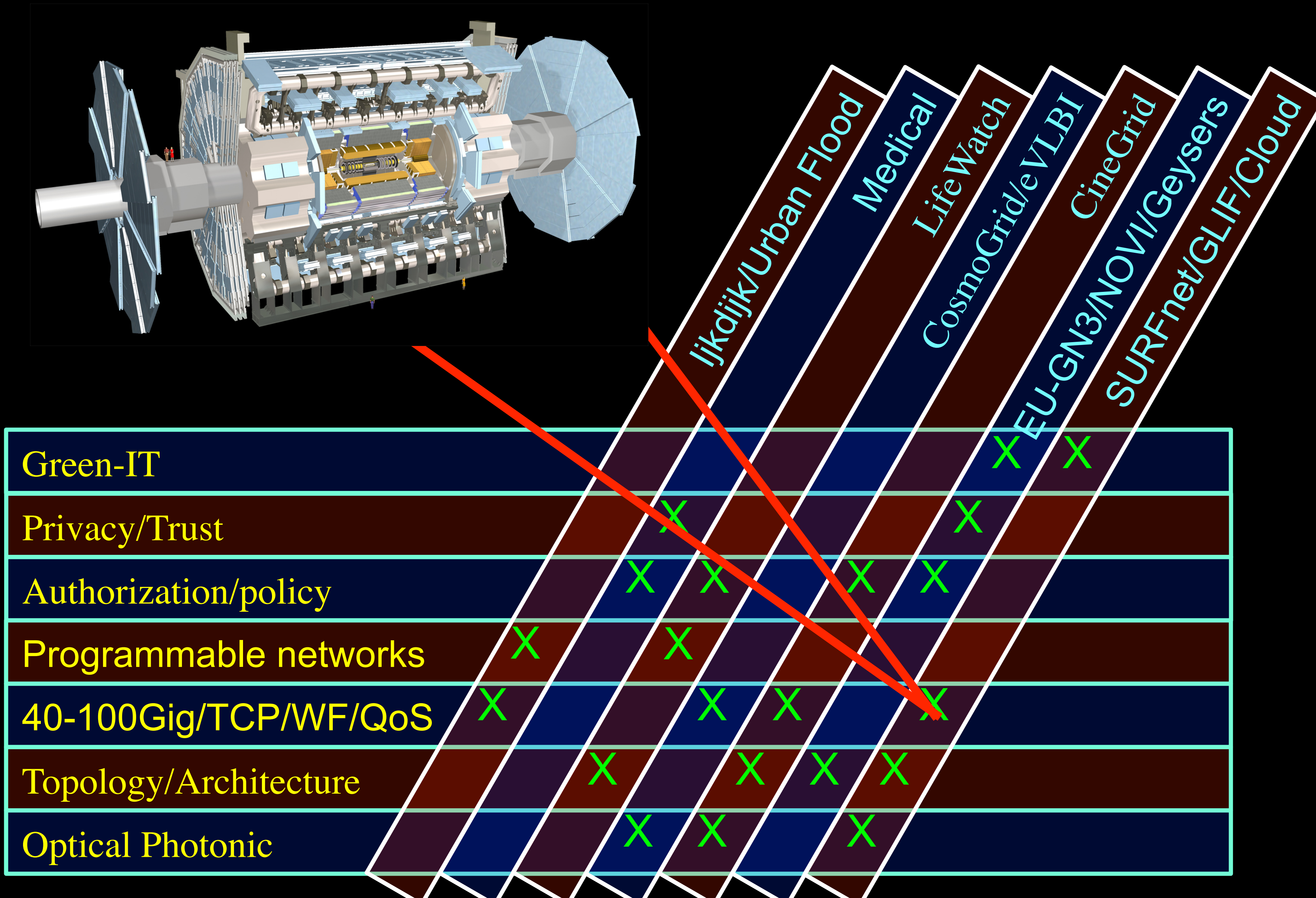


SNE @ UvA



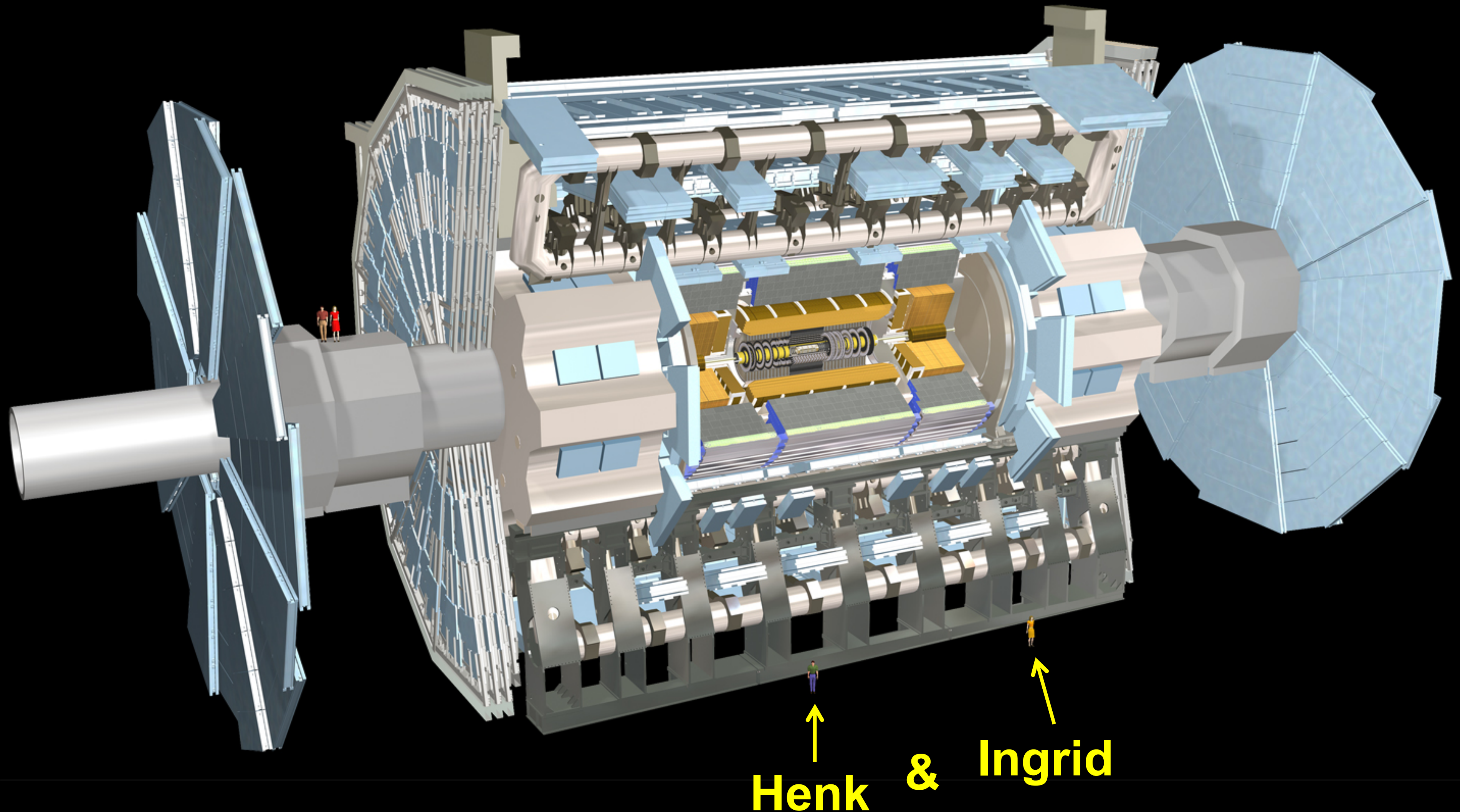
... more data!
Speed
Volume
Deterministic
Real-time
... more realtime!
Scalable
Secure
... more users!
Ijkdijk/Urban Flood
Medical
SmartGrids/GreenClouds
COMMIT/Astronomy
CineGrid
EU-GN3/NOVI/Geysers
SURFnet/GLIF/Cloud
Sustainability, Green-IT
Complexity
Authorization/Privacy/Trust
Programmable Infrastructure
WorkFlow/QoS
Topology/Architecture
Optical Photonic

SNE @ UvA



Ijkdijk/Urban Flood
Medical
LifeWatch
CosmoGrid/eVLBI
CineGrid
EU-GN3/NOVI/Geysers
SURFnet/GLIF/Cloud
Green-IT
Privacy/Trust
Authorization/policy
Programmable networks
40-100Gig/TCP/WF/QoS
Topology/Architecture
Optical Photonic

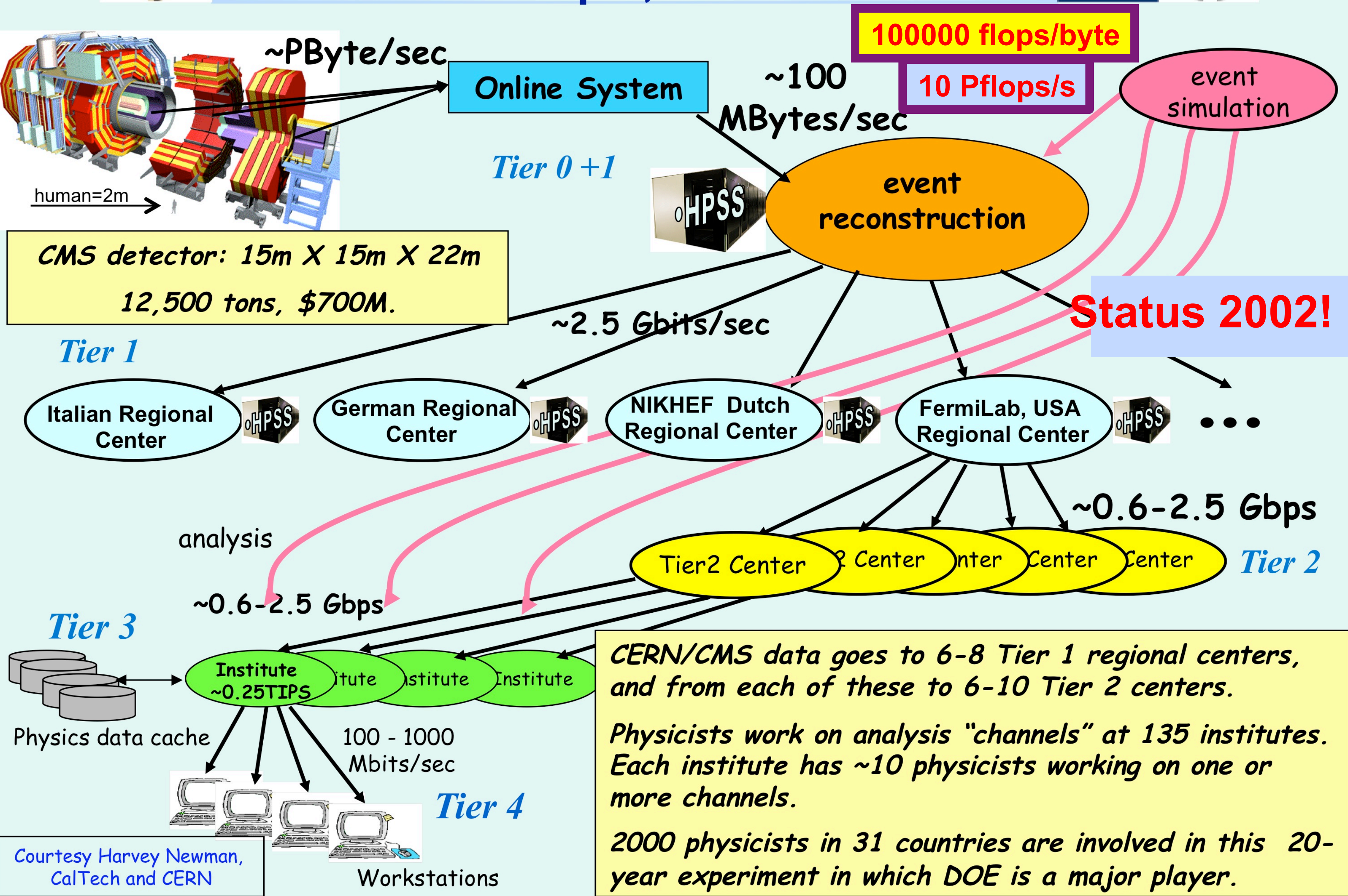
ATLAS detector @ CERN Geneve



ATLAS detector @ CERN Geneve

LHC Data Grid Hierarchy

CMS as example, Atlas is similar



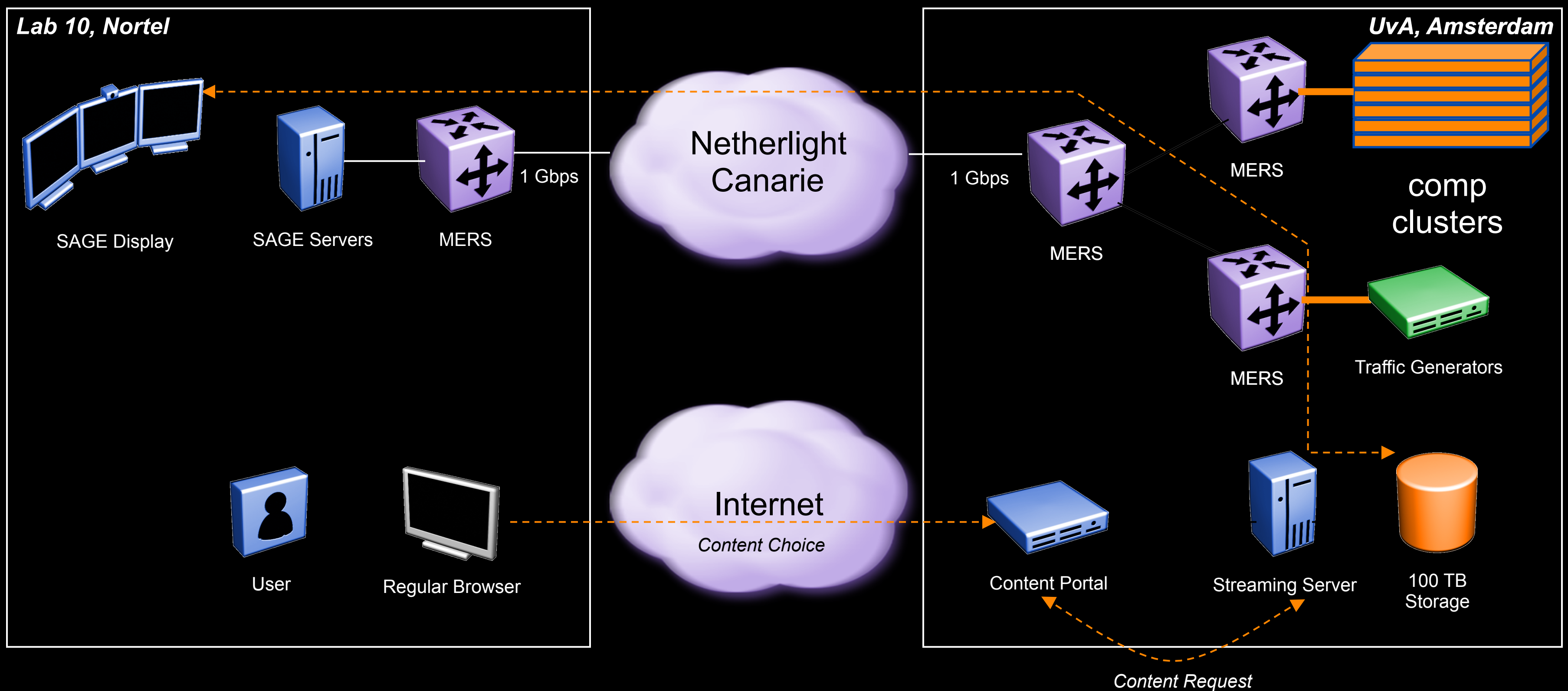
CERN/CMS data goes to 6-8 Tier 1 regional centers, and from each of these to 6-10 Tier 2 centers.

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.

Big and small flows don’t go well together on the same wire! ☹

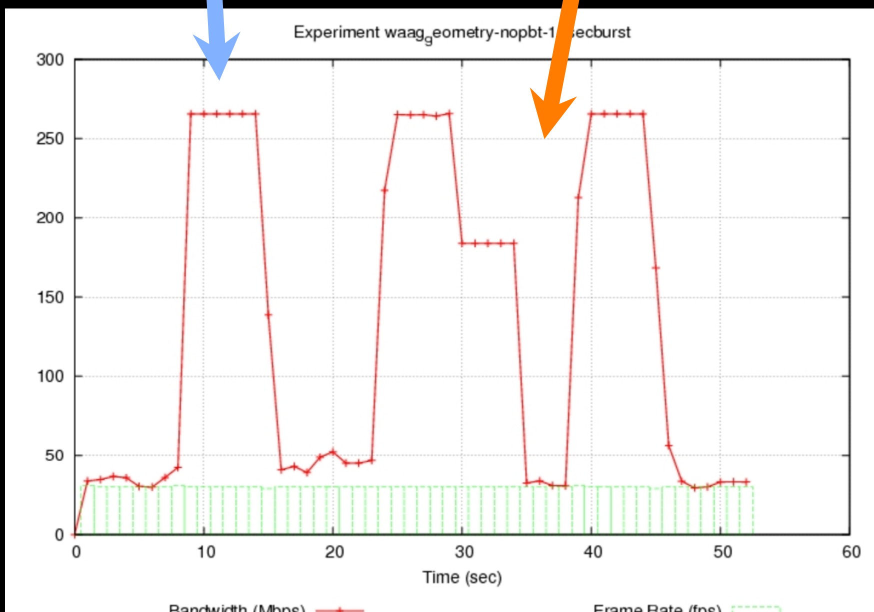
Diagram for SAGE video streaming to ATS



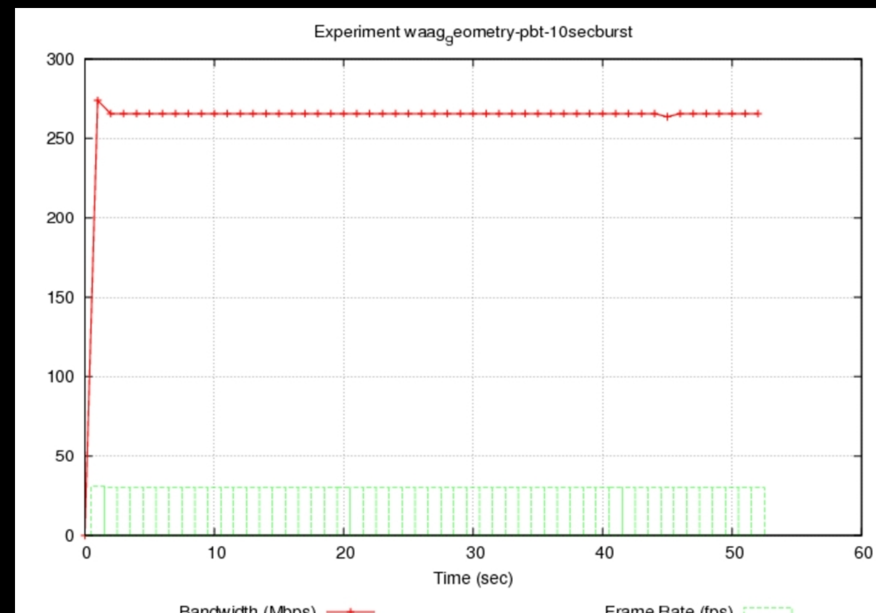
Experimental Data

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

Alien light From idea to realisation!

40Gb/s alien wavelength transmission via a multi-vendor 10Gb/s DWDM infrastructure

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.

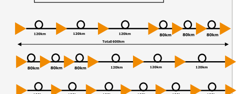
In this demonstration we will investigate the performance of a 40Gb/s PM-QPSK alien wavelength installed on a 10Gb/s DWDM infrastructure.

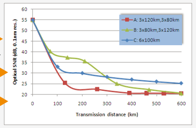
New method to present fiber link quality, FoM (Figure of Merit)

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.

$$FOM = \sum_{j=1}^{N} 10^{\left(\frac{L_j}{10}\right)}$$

Lj, span losses in dB
N, number of spans



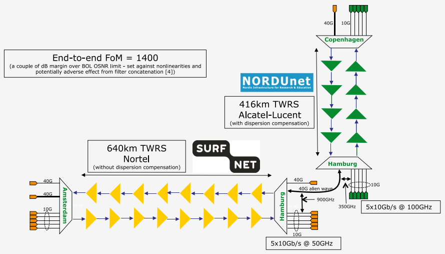


	FOM
A	5504
B	5504
C	1897

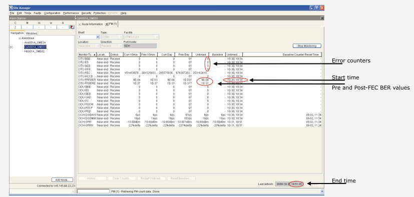
Easy-to-use formula that accurately quantifies transmission system performance

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^{-16}

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.

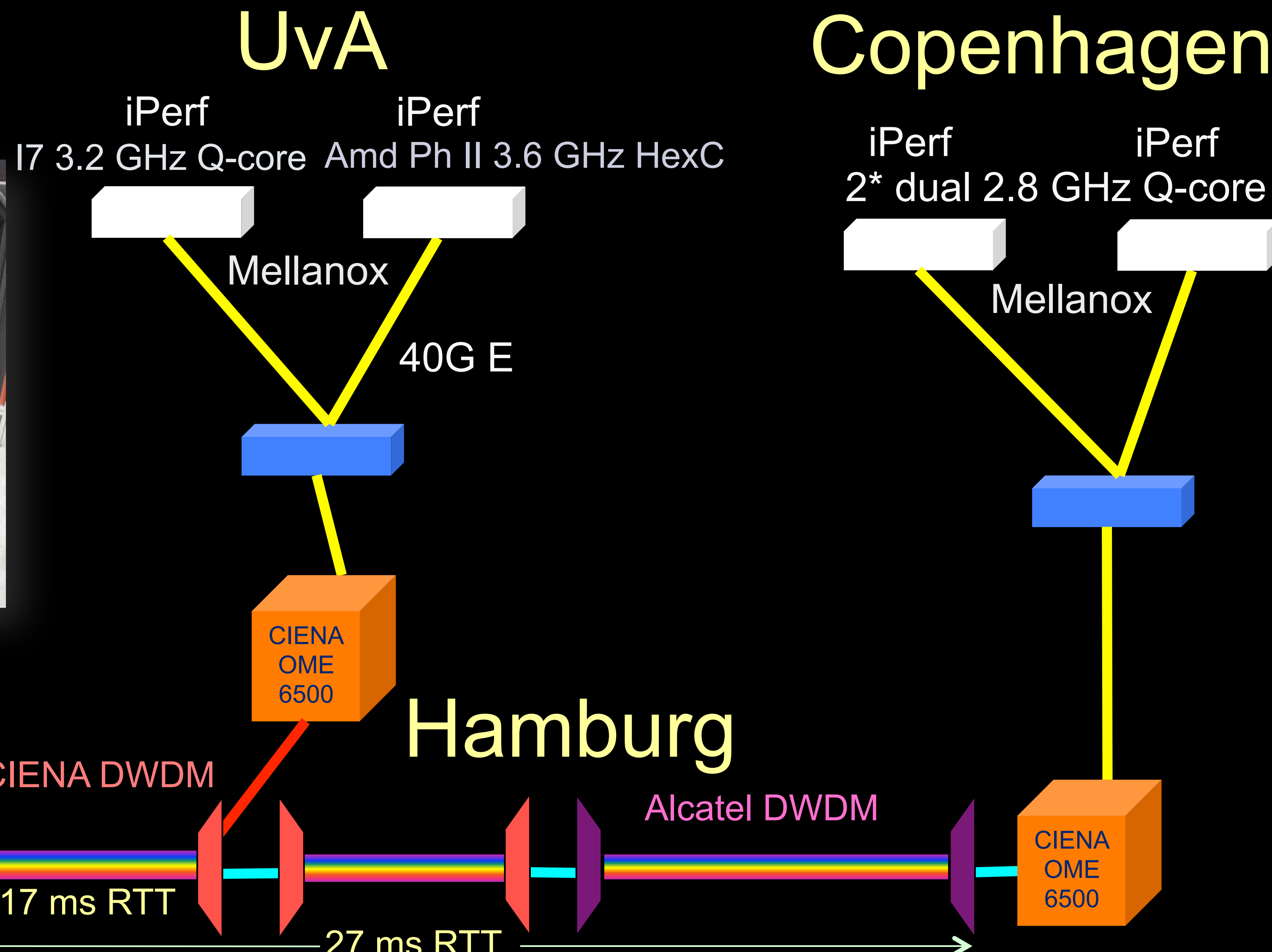
NORTEL NORDUnet telindus together with ISIT SURFNET

REFERENCES [1] "OPERATIONAL SOLUTIONS FOR AN OPEN DWDM LAYER", O. GERSTEL ET AL, OFC'2009 | [2] "AT&T OPTICAL TRANSPORT SERVICES", BARBARA E. SMITH, OFC'09 [3] "OPEX SAVINGS OF ALL-OPTICAL CORE NETWORKS", ANDREW LORD AND CARL ENGINEER, ECOC2009 | [4] NORTEL/SURFNET INTERNAL COMMUNICATION

ACKNOWLEDGEMENTS WE ARE GRATEFUL TO NORDUNET FOR PROVIDING US WITH BANDWIDTH ON THEIR DWDM LINK FOR THIS EXPERIMENT AND ALSO FOR THEIR SUPPORT AND ASSISTANCE DURING THE EXPERIMENTS. WE ALSO ACKNOWLEDGE TELINDUS AND NORTEL FOR THEIR INTEGRATION WORK AND SIMULATION SUPPORT

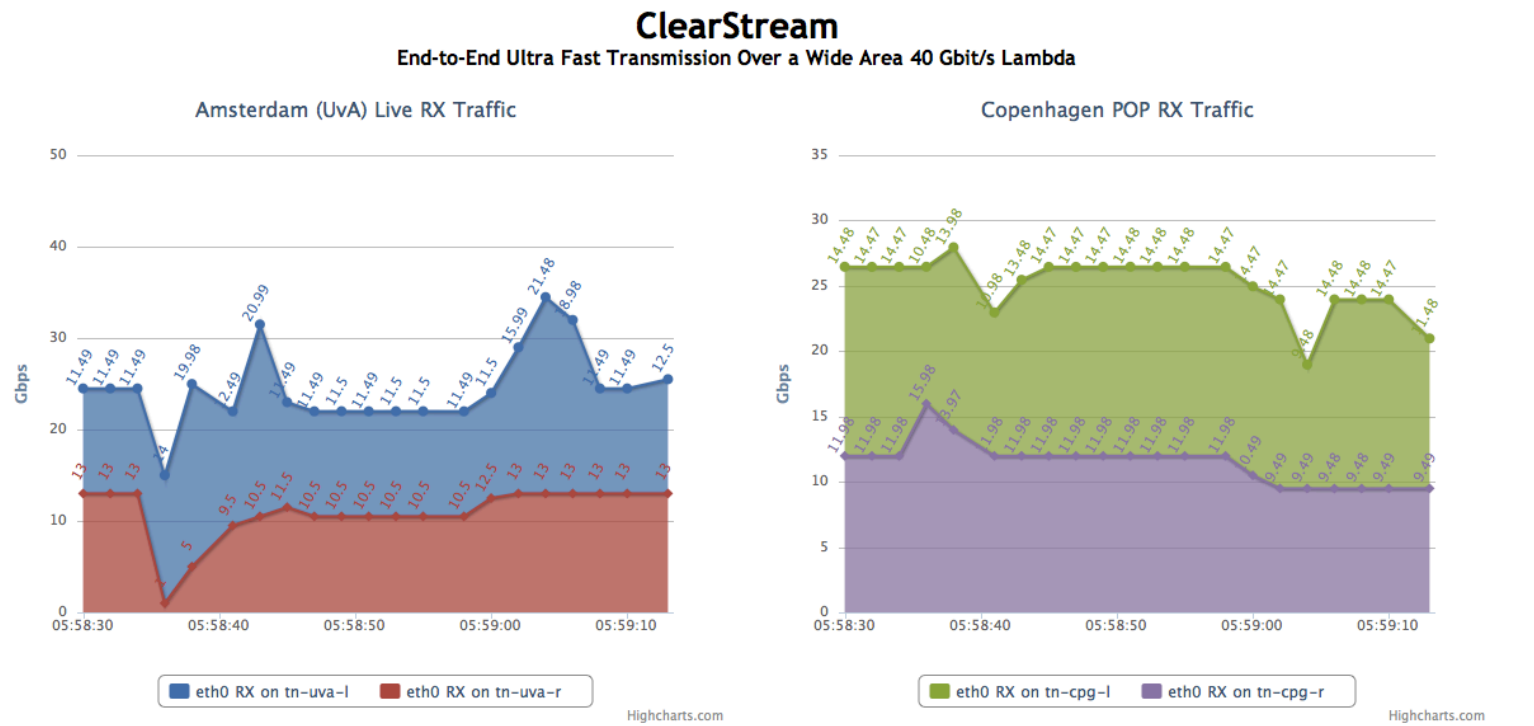
ClearStream @ TNC2011

Setup codename: FlightCees



Amsterdam – Geneva (CERN) – Copenhagen – 4400 km (2700 km alien light)

Visit CIENA Booth
surf to http://tnc11.delaat.net

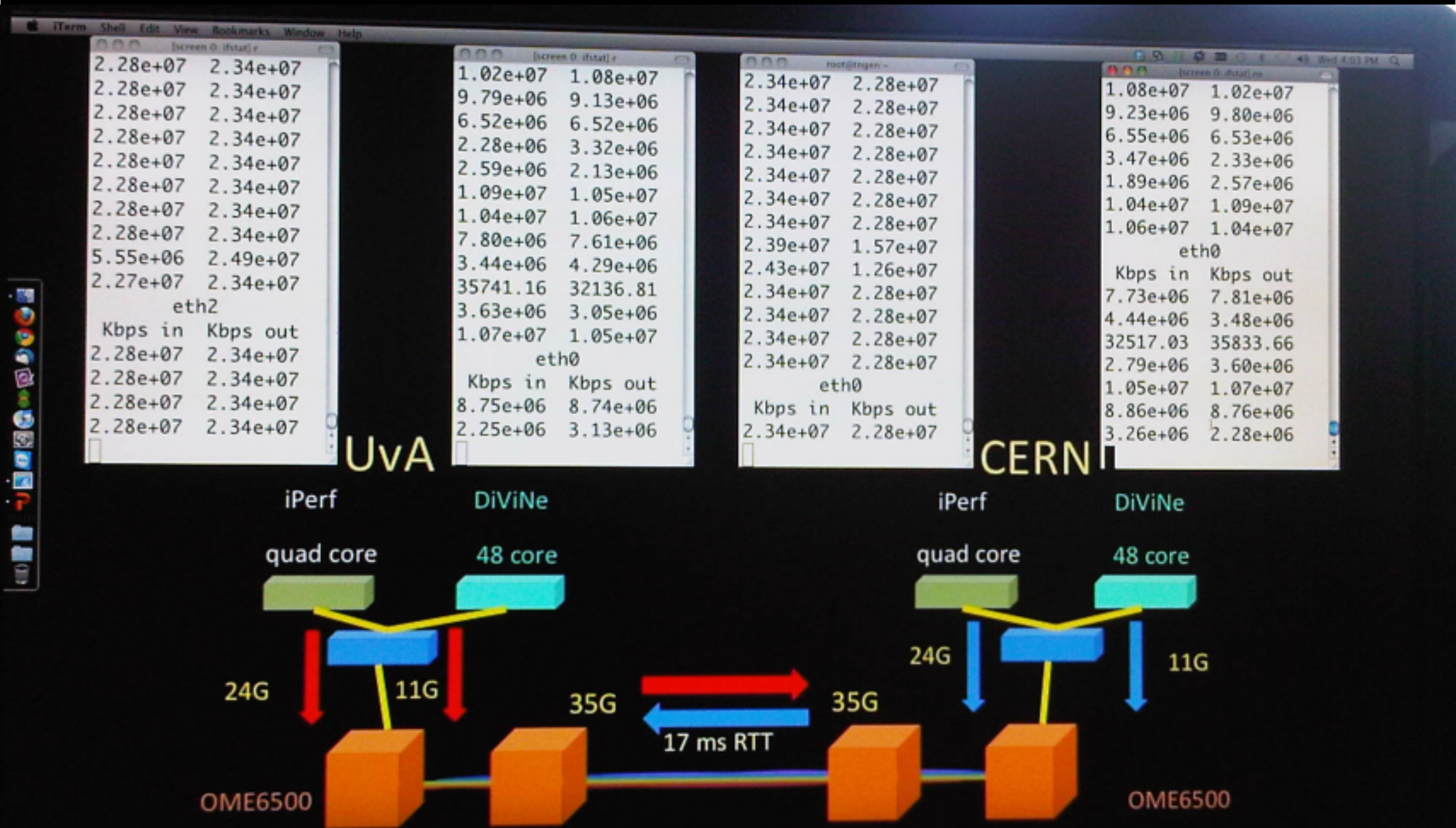


Incoming Amsterdam 25.5 Gbps

Incoming Copenhagen 20.97 Gbps

Total Throughput 46.47 Gbps RTT 44.032 ms

From GLIF October 2010 @ CERN



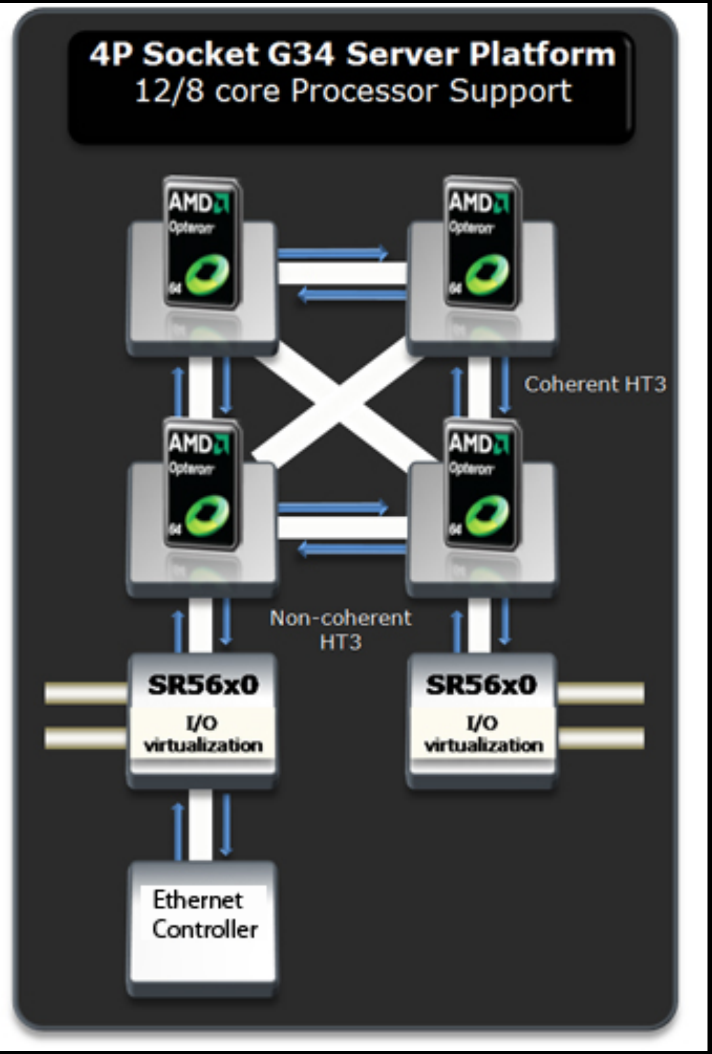
Results (rtt = 17 ms)

- Single flow iPerf 1 core -> 21 Gbps
- Single flow iPerf 1 core <> -> 15+15 Gbps
- Multi flow iPerf 2 cores -> 25 Gbps
- Multi flow iPerf 2 cores <> -> 23+23 Gbps
- DiViNe <> -> 11 Gbps
- Multi flow iPerf + DiVine -> 35 Gbps
- Multi flow iPerf + DiVine <> -> 35 + 35 Gbps

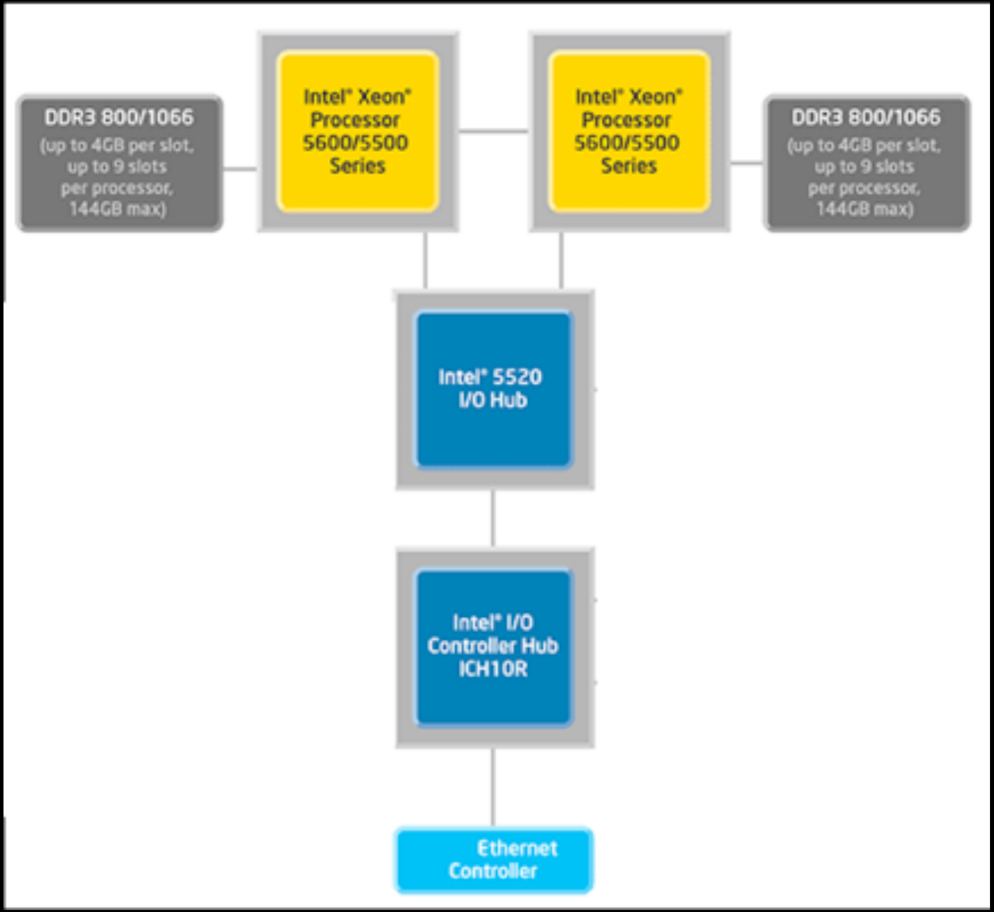
Performance Explained

- Mellanox 40GE card is PCI-E 2.0 8x (5GT/s)
- 40Gbit/s raw throughput but ….
- PCI-E is a network-like protocol
  - 8/10 bit encoding -> 25% overhead -> 32Gbit/s maximum data throughput
  - Routing information
- Extra overhead from IP/Ethernet framing
- Server architecture matters!
  - 4P system performed worse in multithreaded iperf

Server Architecture

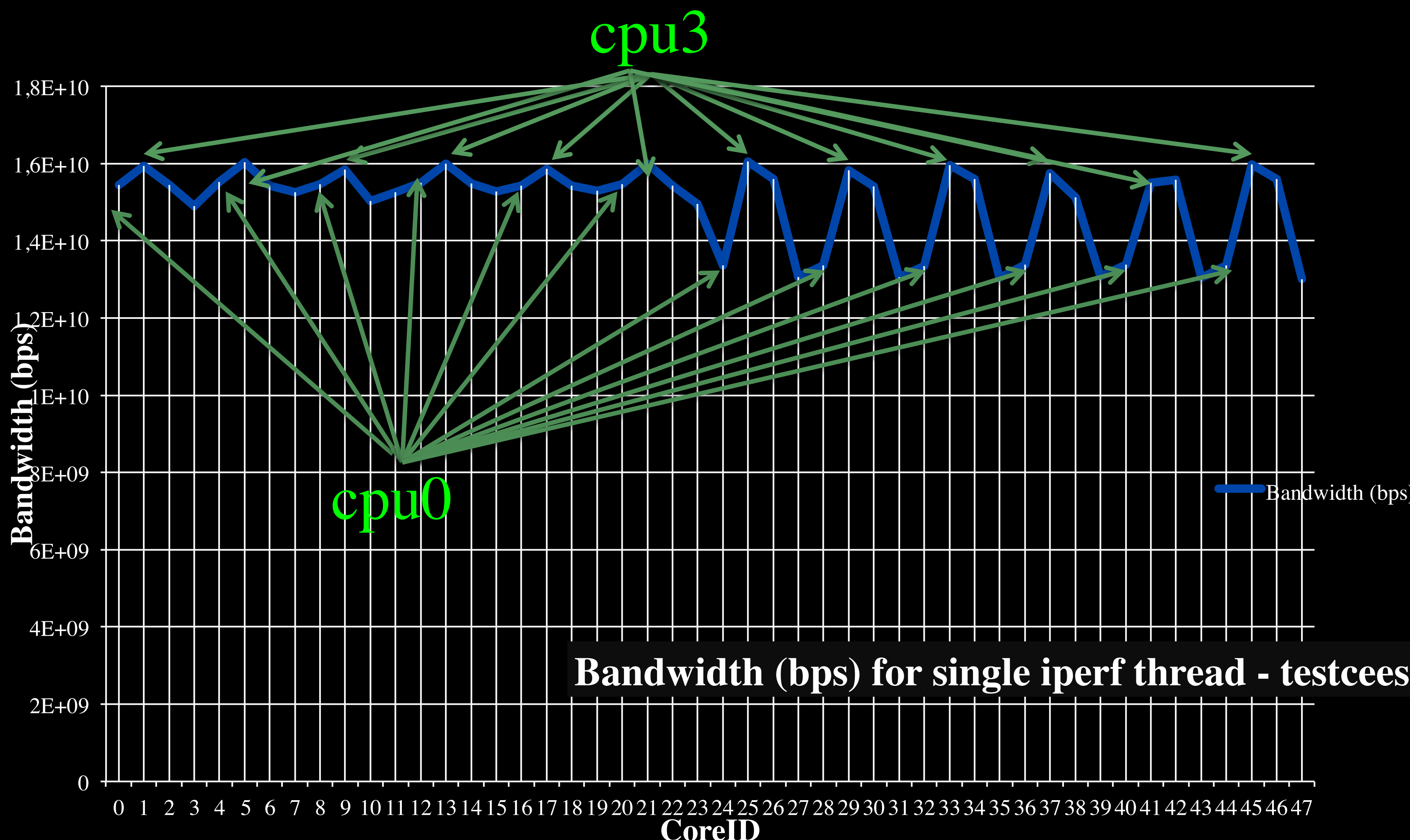


DELL R815
4 x AMD Opteron 6100



Supermicro X8DTT-HIBQF
2 x Intel Xeon

CPU Topology benchmark

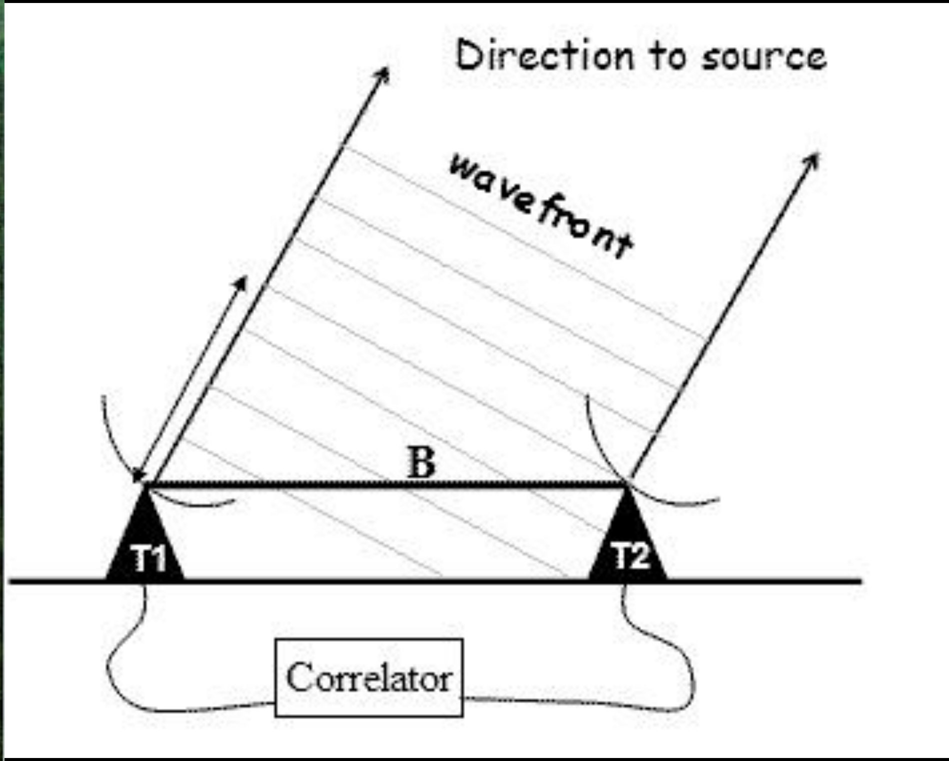


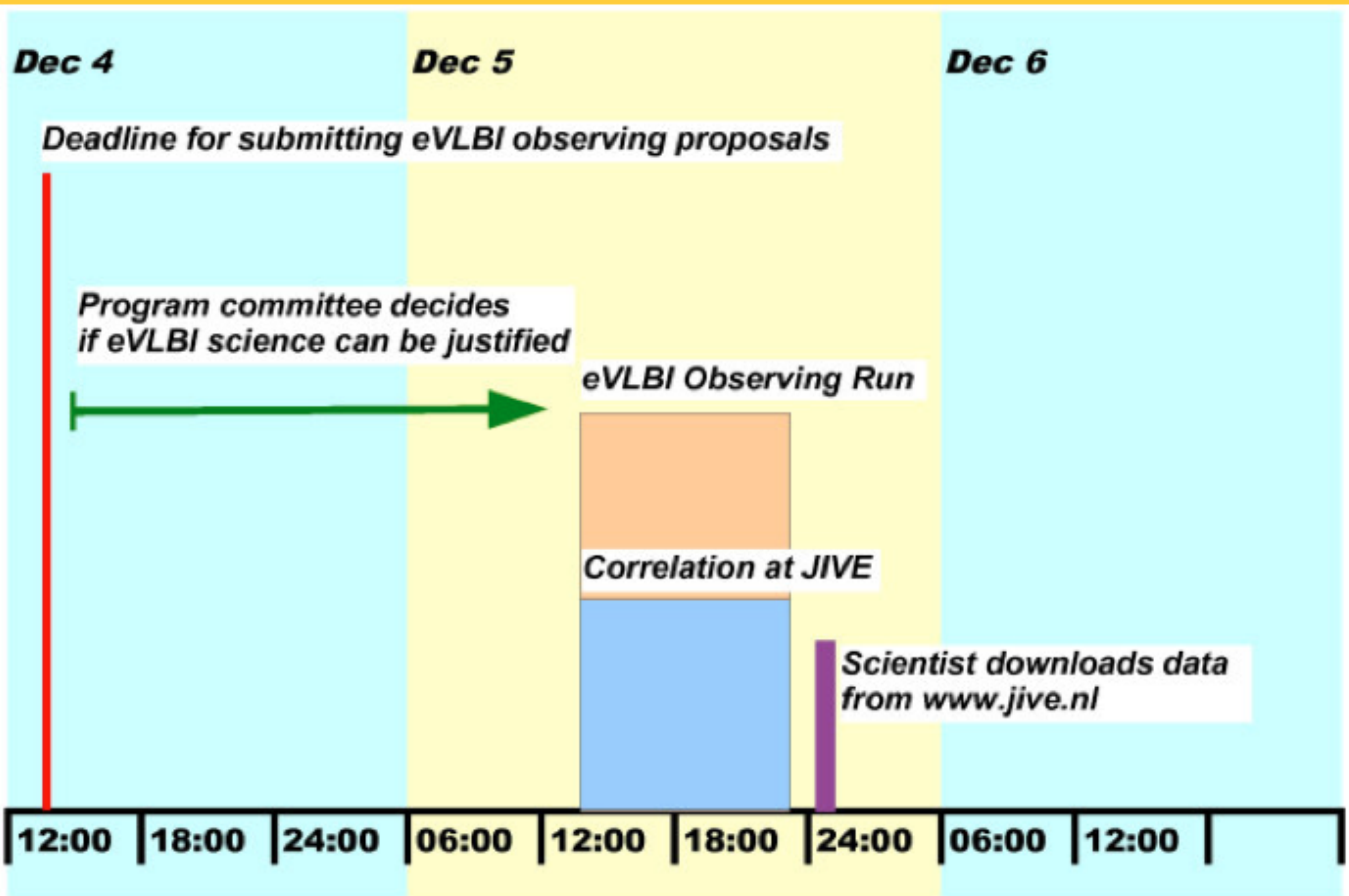
We used numactl to bind iperf to cores

SNE @ UvA

	Ijkdijk/Urban Flood	Medical	LifeWatch	CosmoGrid/eVLBI	CineGrid	EU-GN3/NOVI/Geysers	SURFnet/GLIF/Cloud
Green-IT						X	X
Privacy/Trust		X				X	
Authorization/policy		X	X		X	X	
Programmable networks	X		X				
40-100Gig/TCP/WF/QoS	X			X	X		X
Topology/Architecture			X		X	X	X
Optical Photonic				X	X		X

e -Very Large Base Interferometer



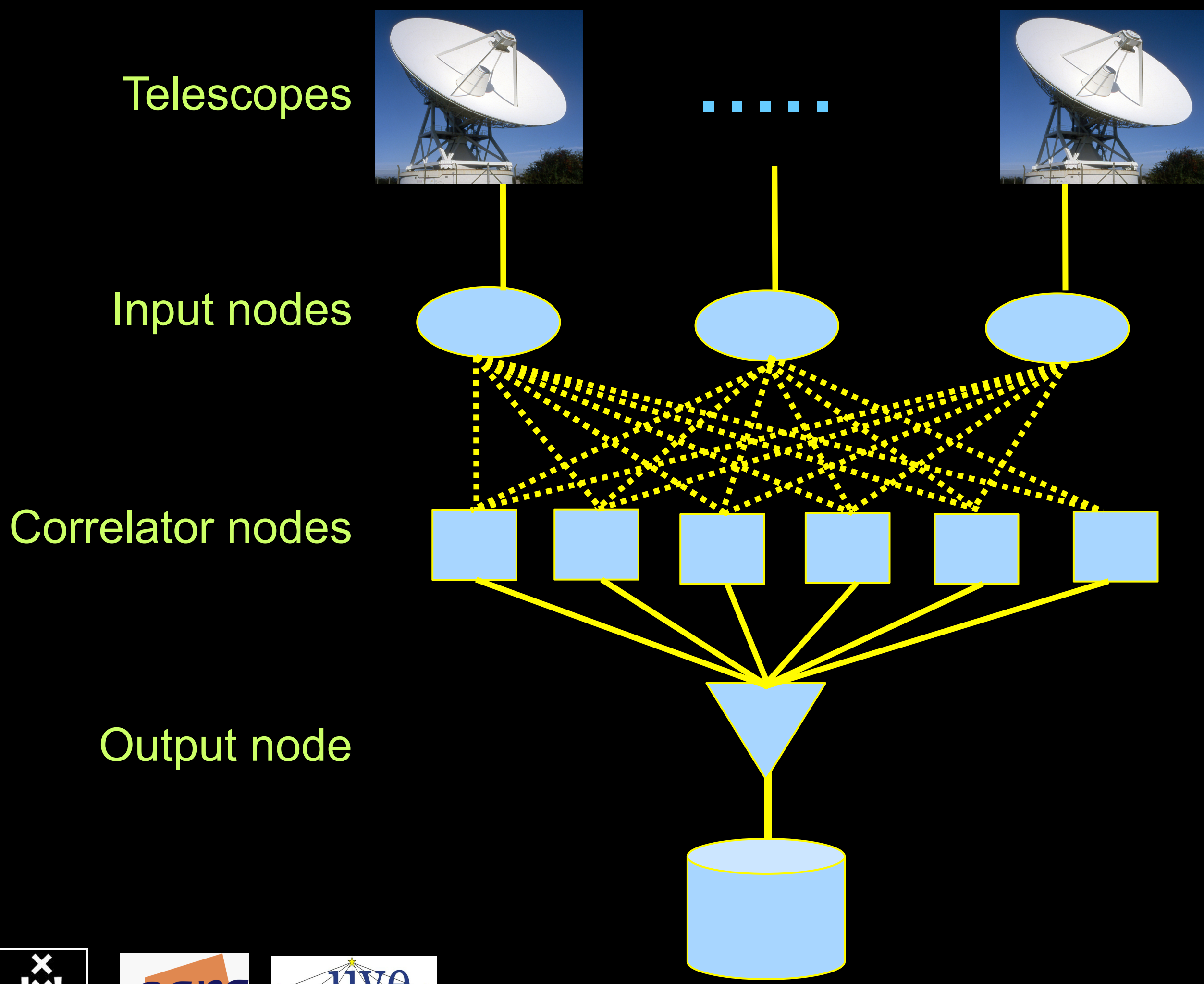


Slide courtesy of Richard Schilizzi <schilizzi@jive.nl>

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.

16 Gbit/s - 2 Tflop ➔
THIS IS A DATA FLOW PROBLEM !!!



Research:

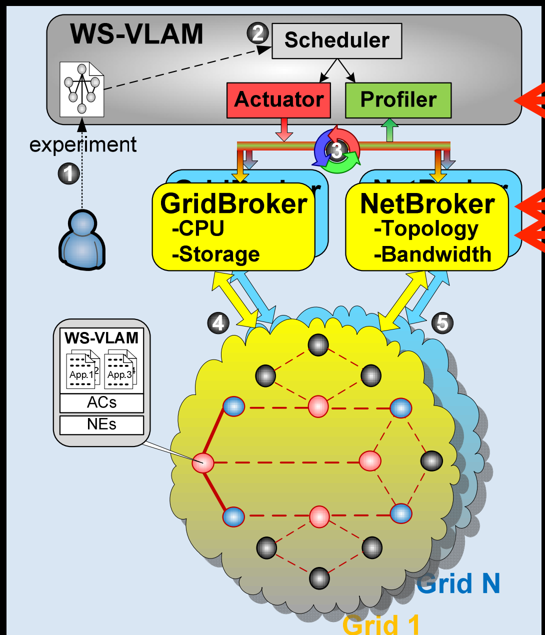


Figure 2. Grid architecture that includes programmable network services.

LOFAR as a Sensor Network

20 flops/byte

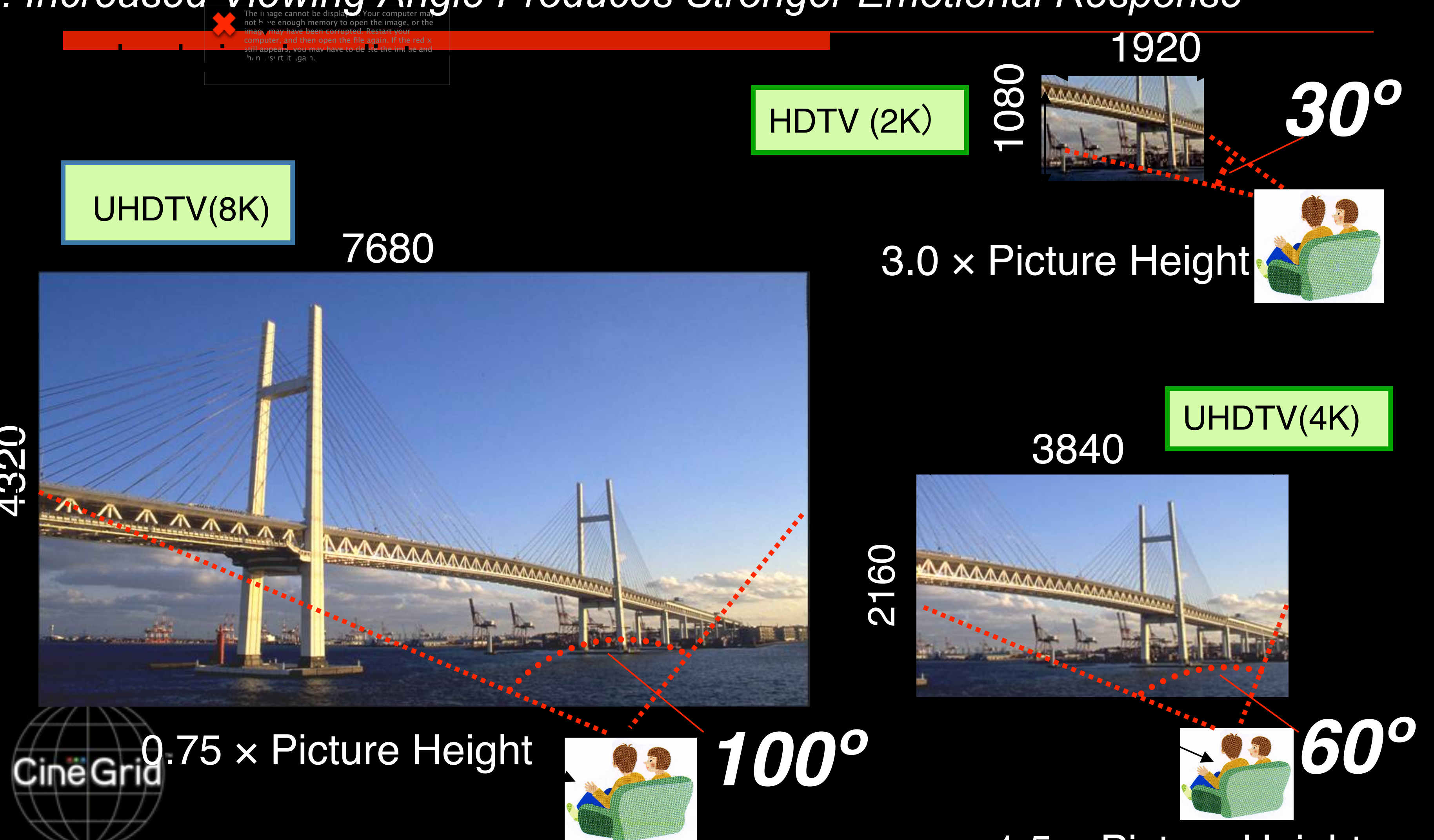
- LOFAR is a large distributed research infrastructure:

2 Tflops/s

  - 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

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

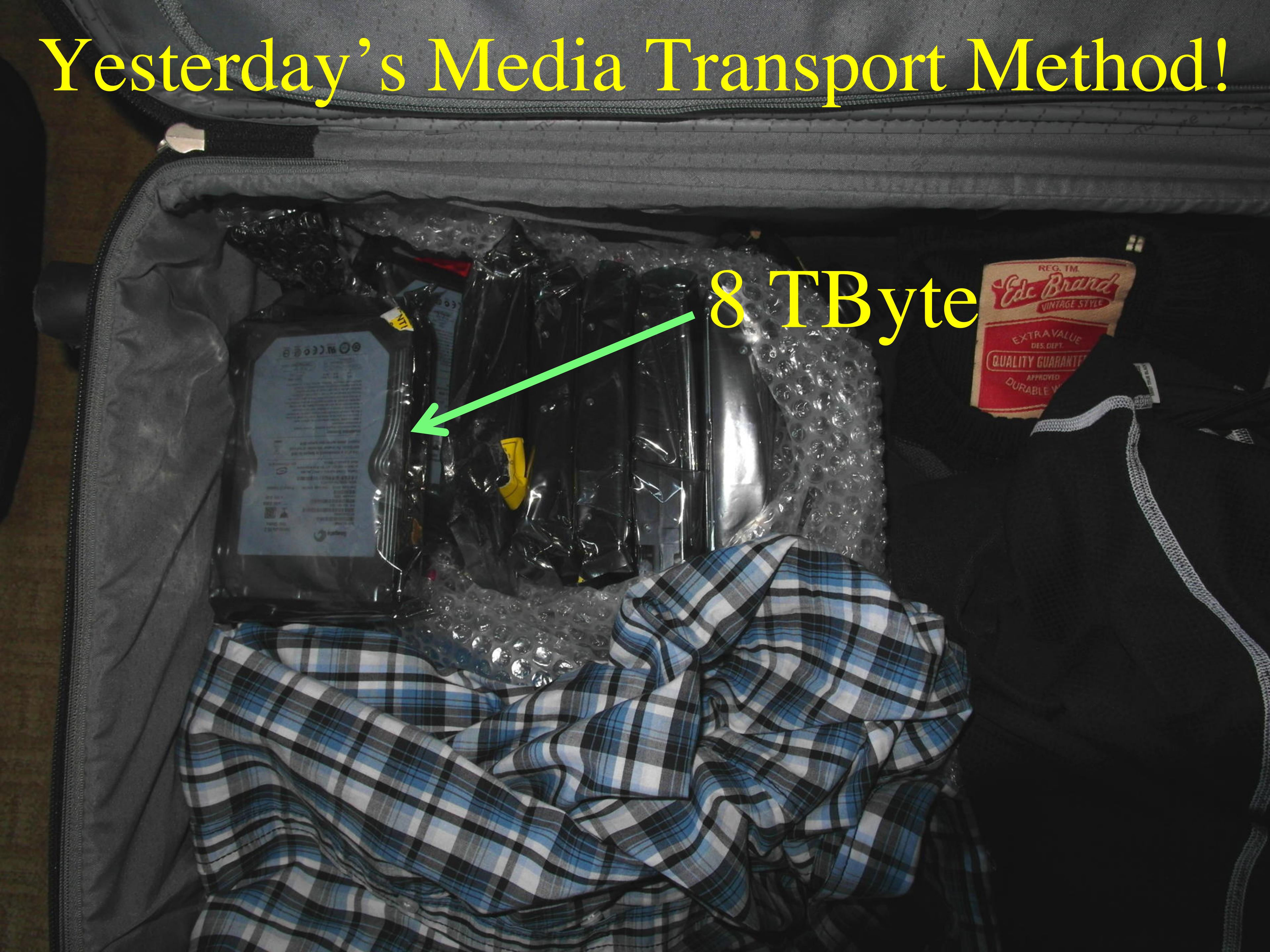


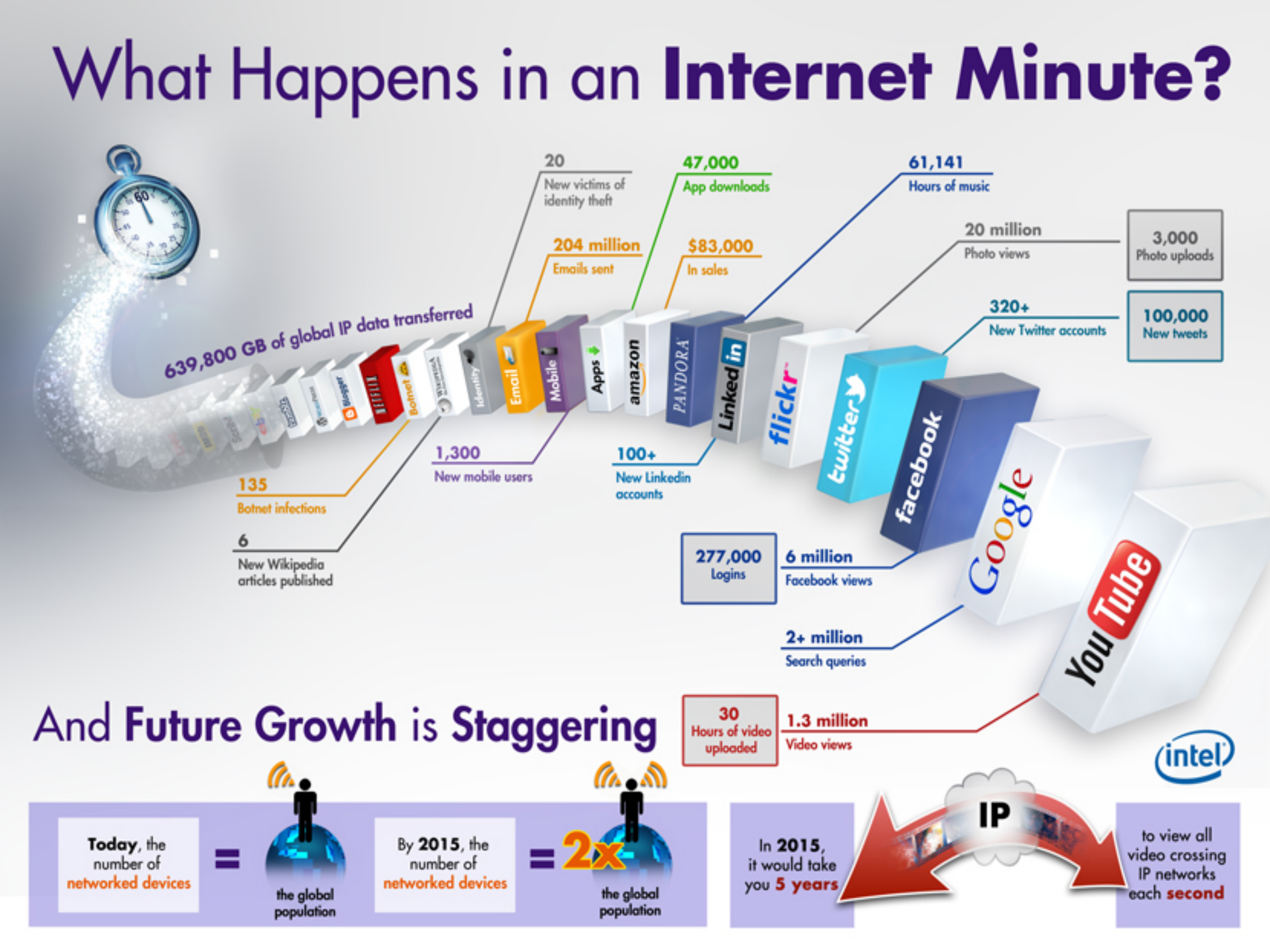
Yutaka TANAKA
SHARP CORPORATION
Advanced Image Research Laboratories

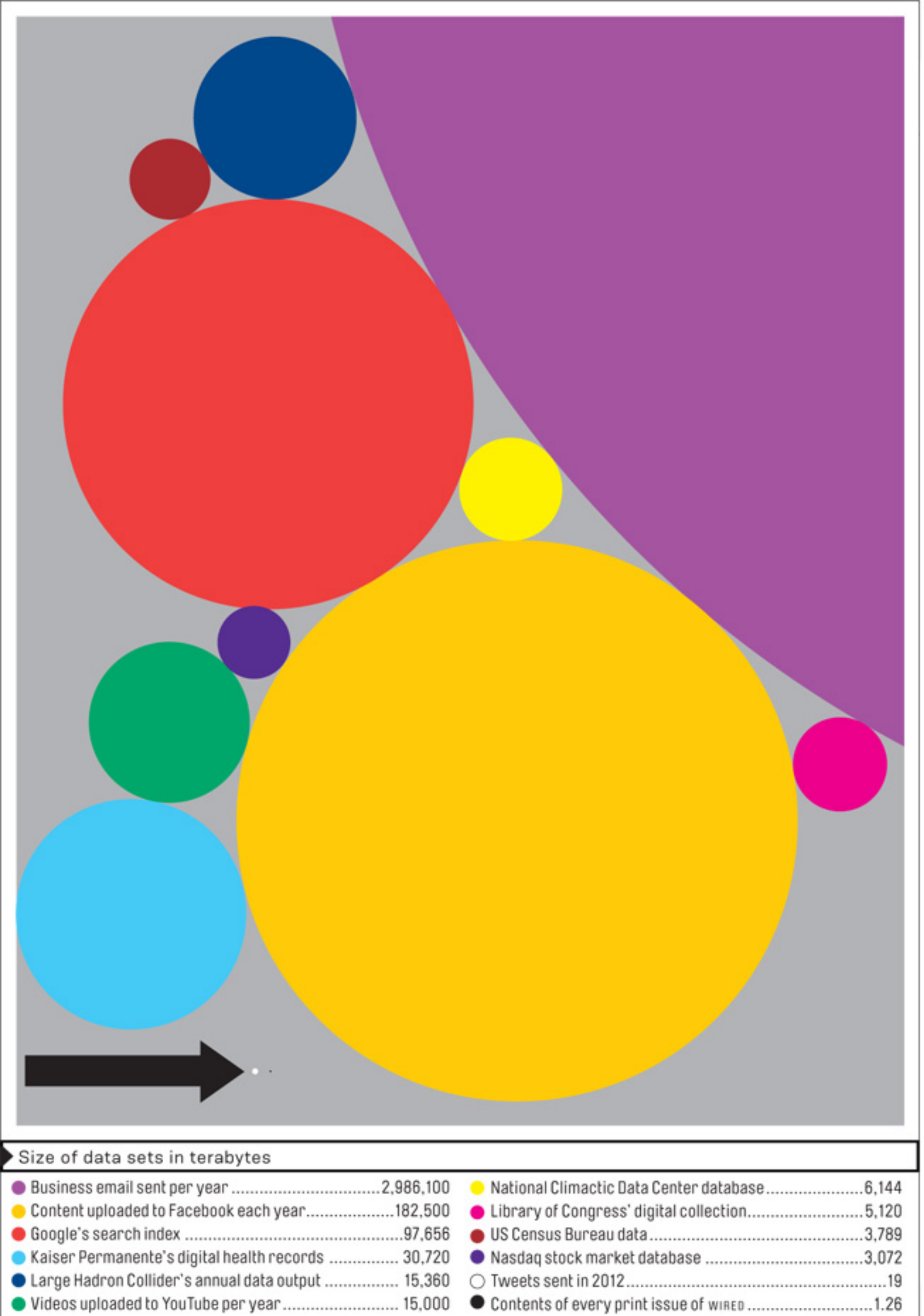
Moving Big Data Objects Globally

- **Digital Motion Picture for Audio Post-Production**
  - 1 TV Episode Dubbing Reference ~ 1 GB
  - 1 Theatrical 5.1 Final Mix ~ 8 GB
  - 1 Theatrical Feature Dubbing reference ~ 30 GB
- **Digital Motion Picture Acquisition**
  - 4K RGB x 24 FPS x 10bit/color: ~ 48MB/Frame uncompressed *(ideal)*
  - 6:1 ~ 20:1 shooting ratios => 48TB ~ 160TB digital camera originals
- **Digital Dailies**
  - HD compressed MPEG-2 @ 25 ~ 50 Mb/s
- **Digital Post-production and Visual Effects**
  - Gigabytes - Terabytes to Select Sites Depending on Project
- **Digital Motion Picture Distribution**
  - Film Printing in Regions
    - Features ~ 8TB
    - Trailers ~ 200GB
  - Digital Cinema Package to Theatres
    - Features ~ 100 - 300GB per DCP
    - Trailers ~ 2 - 4GB per DCP

Yesterday's Media Transport Method!



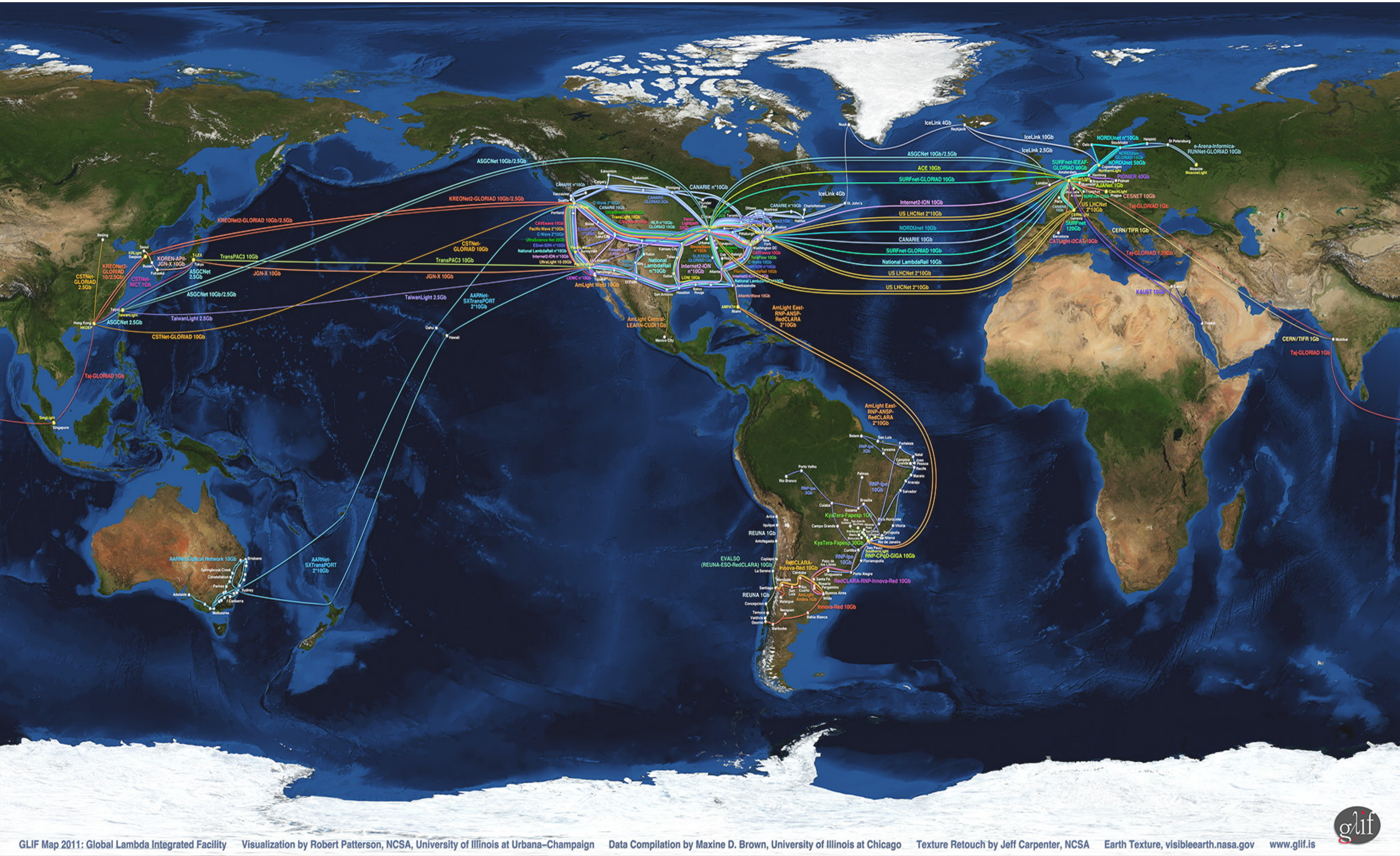

What Happens in an Internet Minute?
639,800 GB of global IP data transferred
20
New victims of identity theft
204 million
Emails sent
47,000
App downloads
$83,000
In sales
61,141
Hours of music
20 million
Photo views
3,000
Photo uploads
320+
New Twitter accounts
100,000
New tweets
135
Botnet infections
6
New Wikipedia articles published
1,300
New mobile users
100+
New Linkedin accounts
277,000
Logins
6 million
Facebook views
2+ million
Search queries
30
Hours of video uploaded
1.3 million
Video views
Botnet
Identity
Email
Mobile
Apps
amazon
PANDORA
Linked in
flickr
twitter
facebook
Google
YouTube
intel
And Future Growth is Staggering
Today, the number of networked devices
=
the global population
By 2015, the number of networked devices
=
2x
the global population
In 2015, it would take you 5 years
IP
to view all video crossing IP networks each second



There
is
always
a
bigger
fish

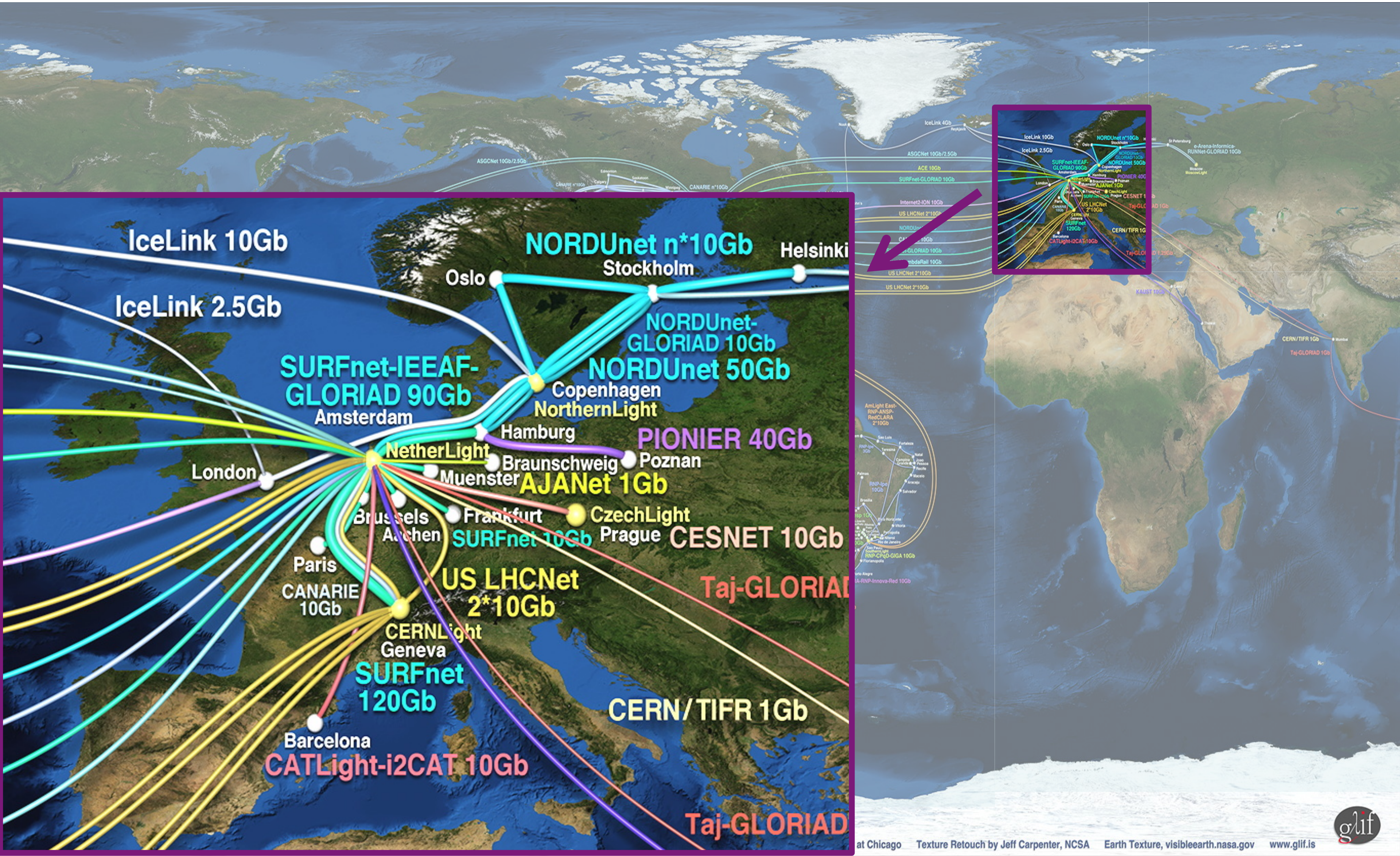
The GLIF – LightPaths around the World

F Dijkstra, J van der Ham, P Grosso, C de Laat, “A path finding implementation for multi-layer networks”, Future Generation Computer Systems 25 (2), 142-146.

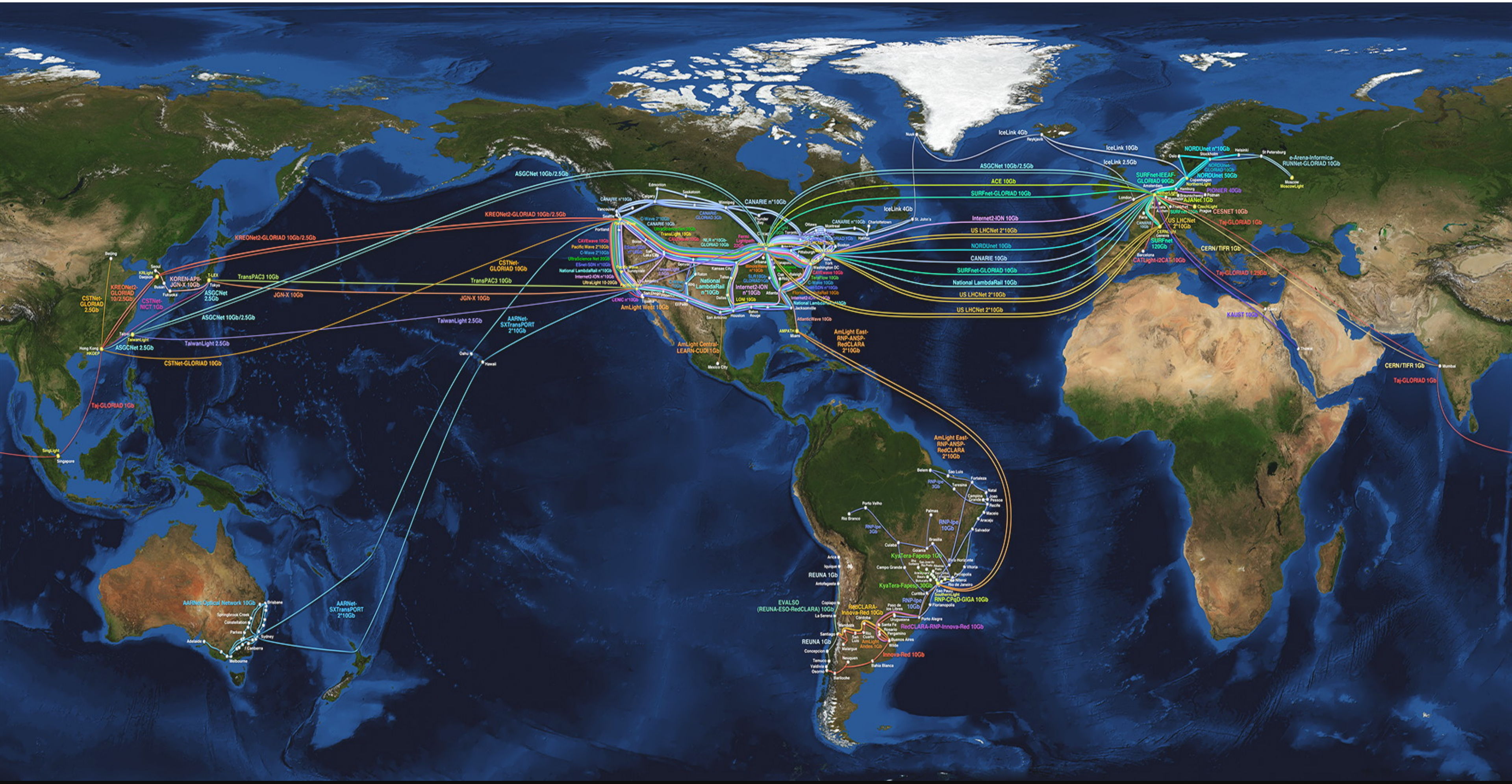


The GLIF – LightPaths around the World

F Dijkstra, J van der Ham, P Grosso, C de Laat, “A path finding implementation for multi-layer networks”, Future Generation Computer Systems 25 (2), 142-146.



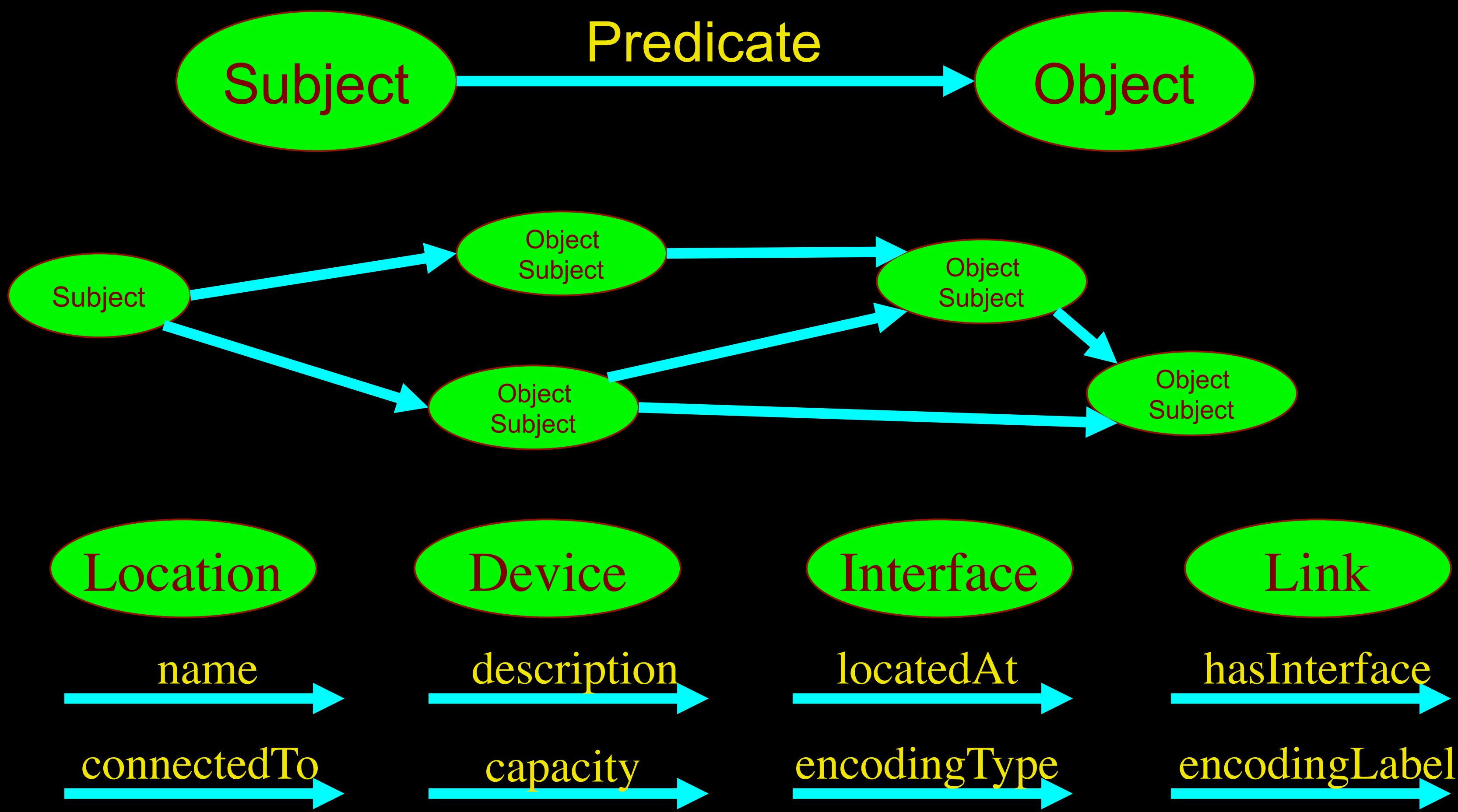
The GLIF – LightPaths around the World



We investigate: TOMTOM for complex networks!

LinkedIN for Infrastructure

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



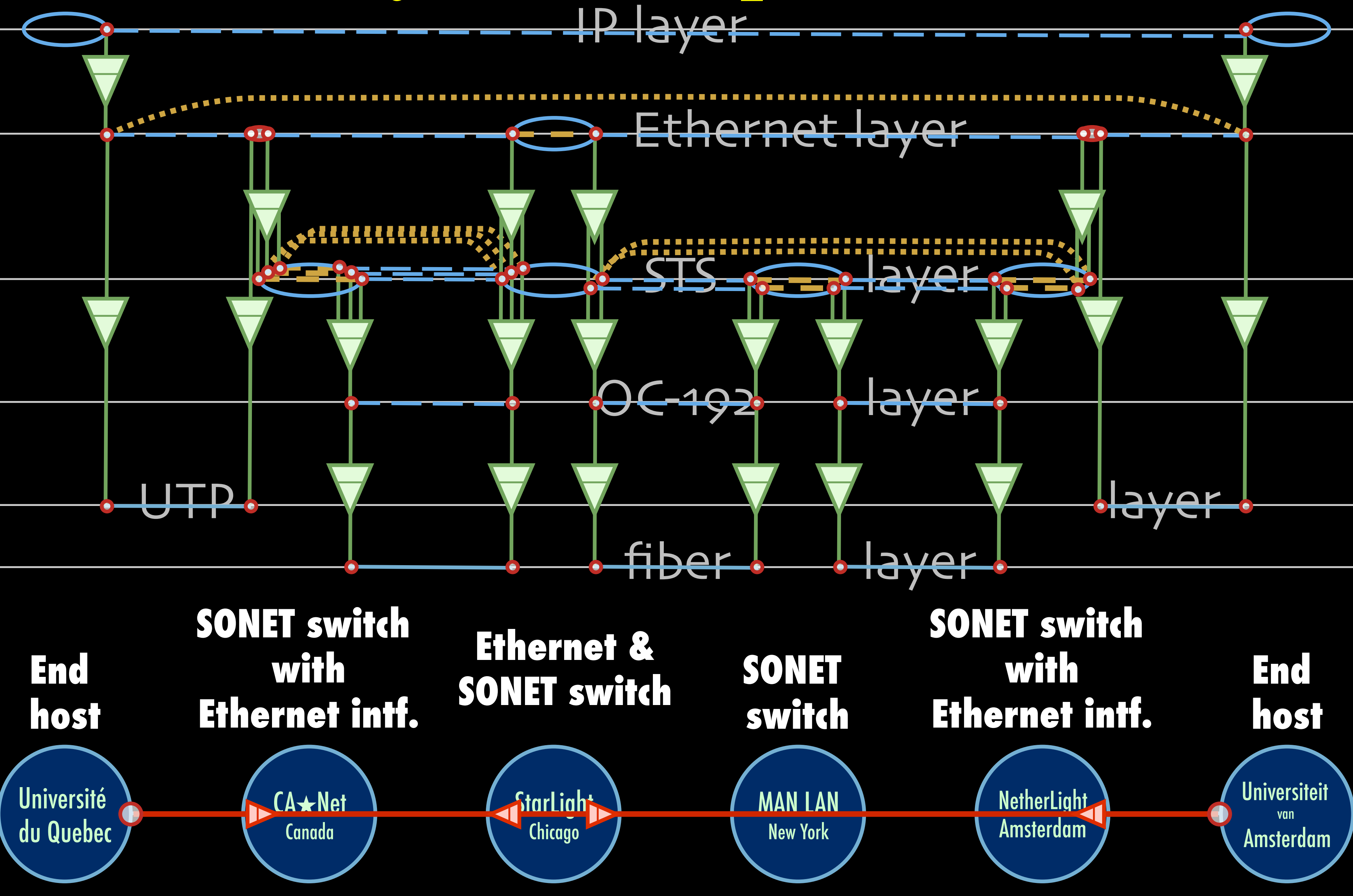
NetherLight in RDF

```
<?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:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
      <ndl:hasInterface rdf:resourc
```

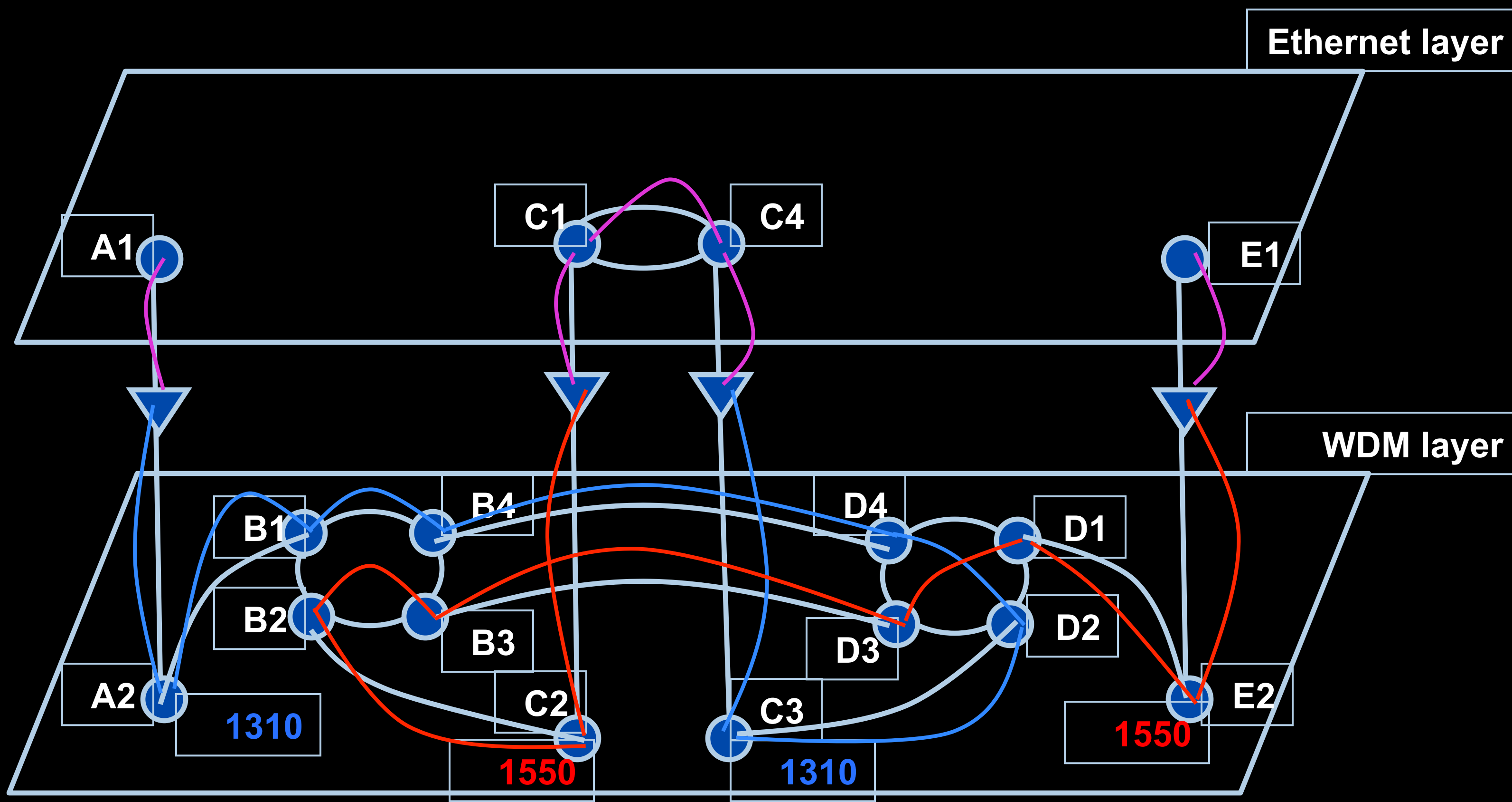
```
<!-- all the interfaces of TDM3.amsterdam1.netherlight.net -->

<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>
```

Multi-layer descriptions in NDL



Multi-layer Network PathFinding



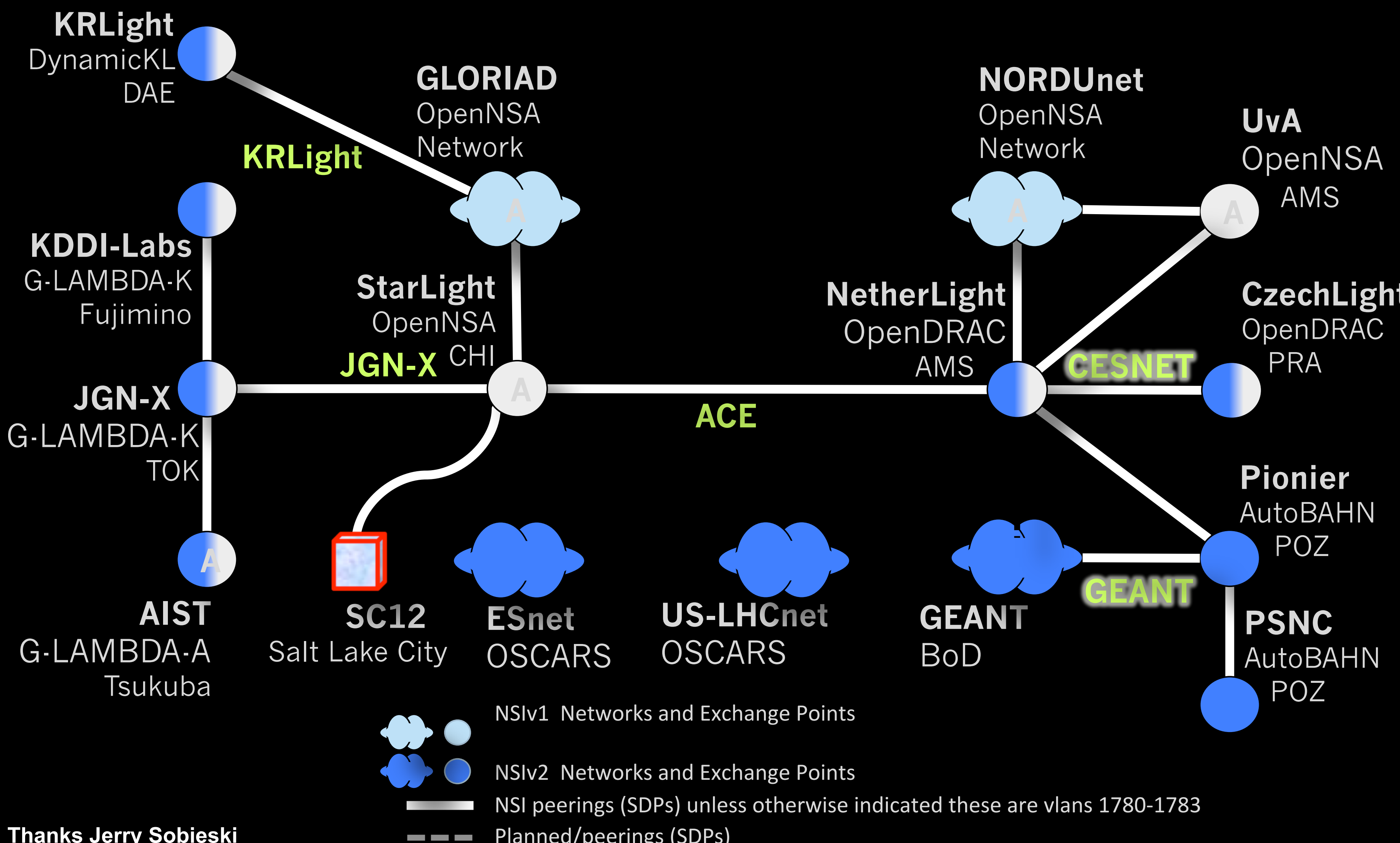
Path between interfaces A1 and E1:
A1-A2-B1-B4-D4-D2-C3-C4-C1-C2-B2-B3-D3-D1-E2-E1

Scaling: Combinatorial problem

Automated GOLE + NSI

Joint NSI v1+v2 Beta Test Fabric   Nov 2012

Ethernet Transport Service

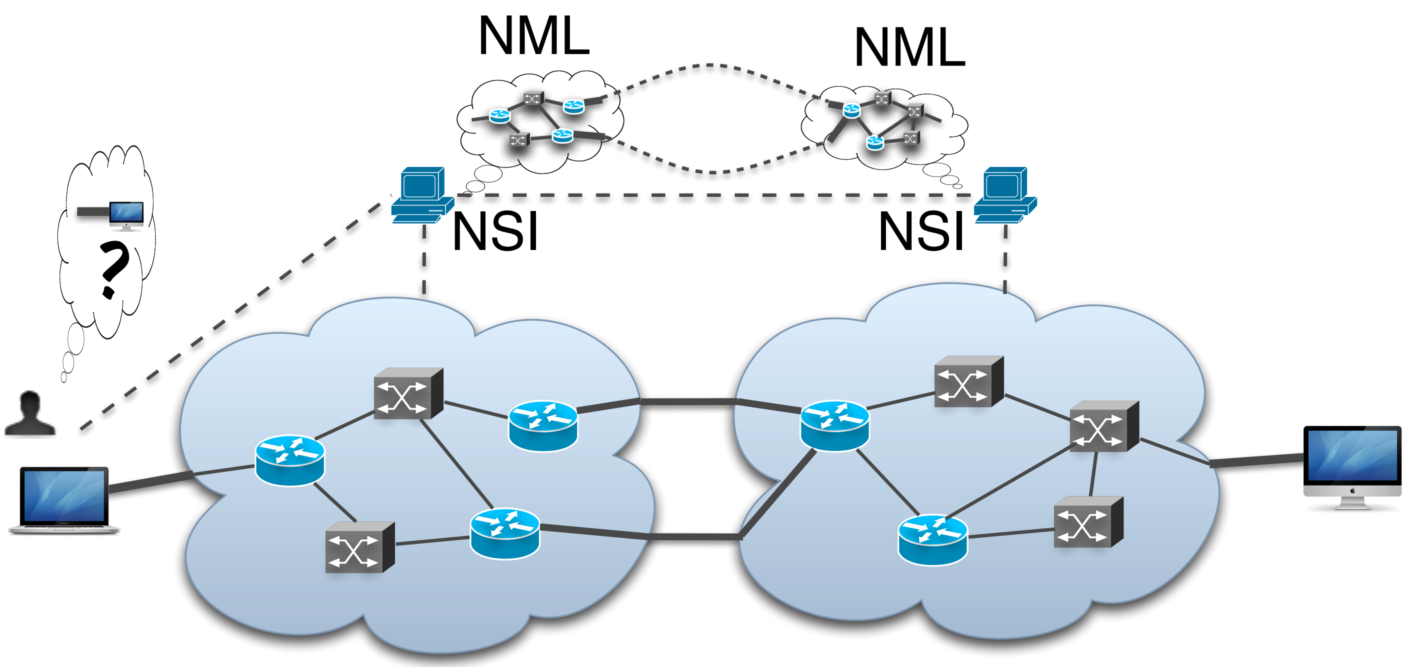


Thanks Jerry Sobieski

Network Topology Description

Network topology research supporting automatic network provisioning

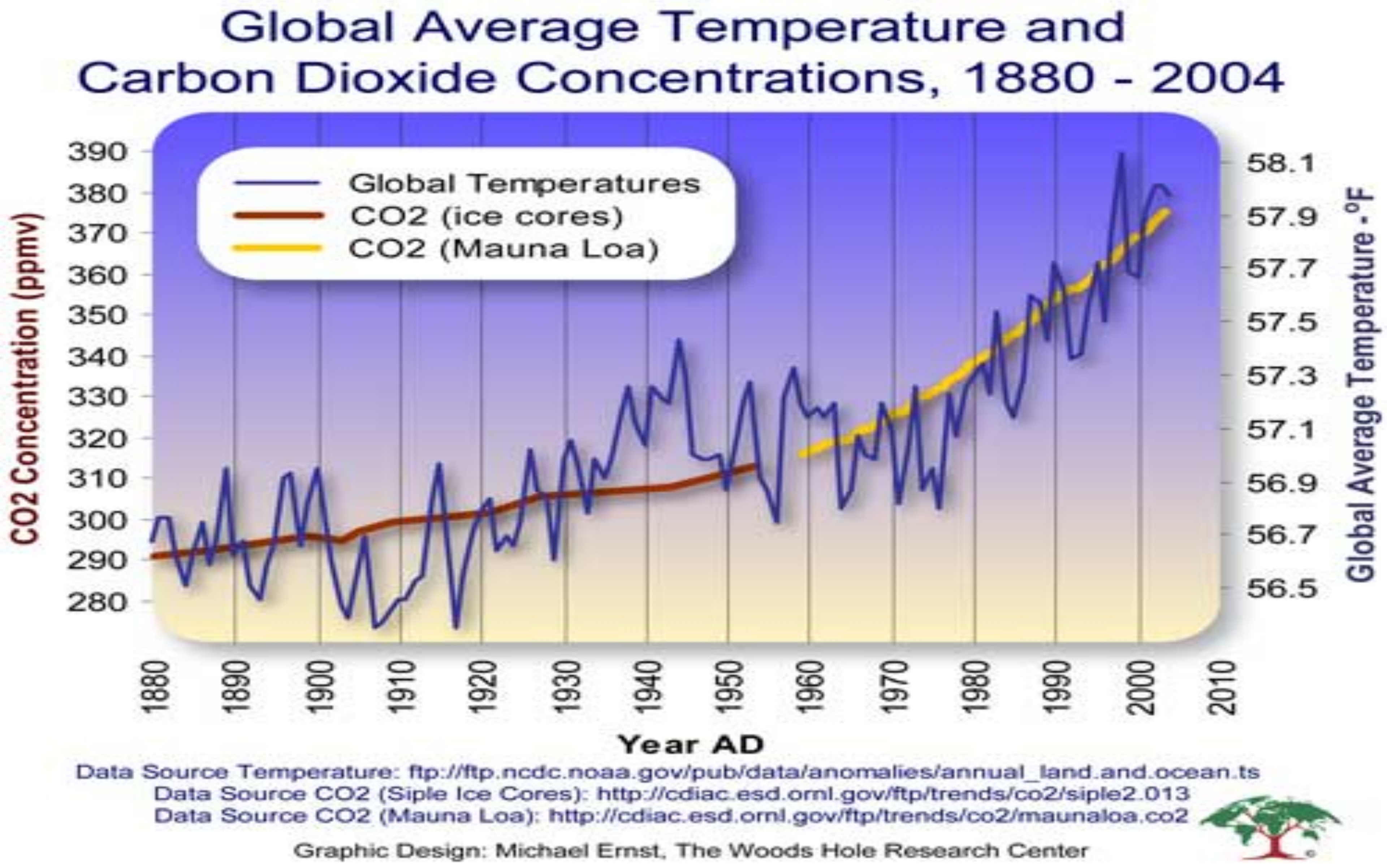
- Inter-domain networks
- Multiple technologies
- Based on incomplete information
- Possibly linked to other resources



http://redmine.ogf.org/projects/nml-wg
http://redmine.ogf.org/projects/nsi-wg

http://sne.science.uva.nl/ndl

Need for GreenIT



Greening the Processing System

ENERGY STAR

Turn Green Tech into Greenbacks

IT Certifications for Jobs That Make a Difference

Uptime Institute Accredited Tier Designer

The Uptime Institute has long been a proponent for green data center design and implementation. Its certification course on data center design embeds green principles into the curriculum.

CERTIFIED GREEN COMPUTING USER SPECIALIST

GCI™ Green Computing Initiative

SUSTAINABILITY

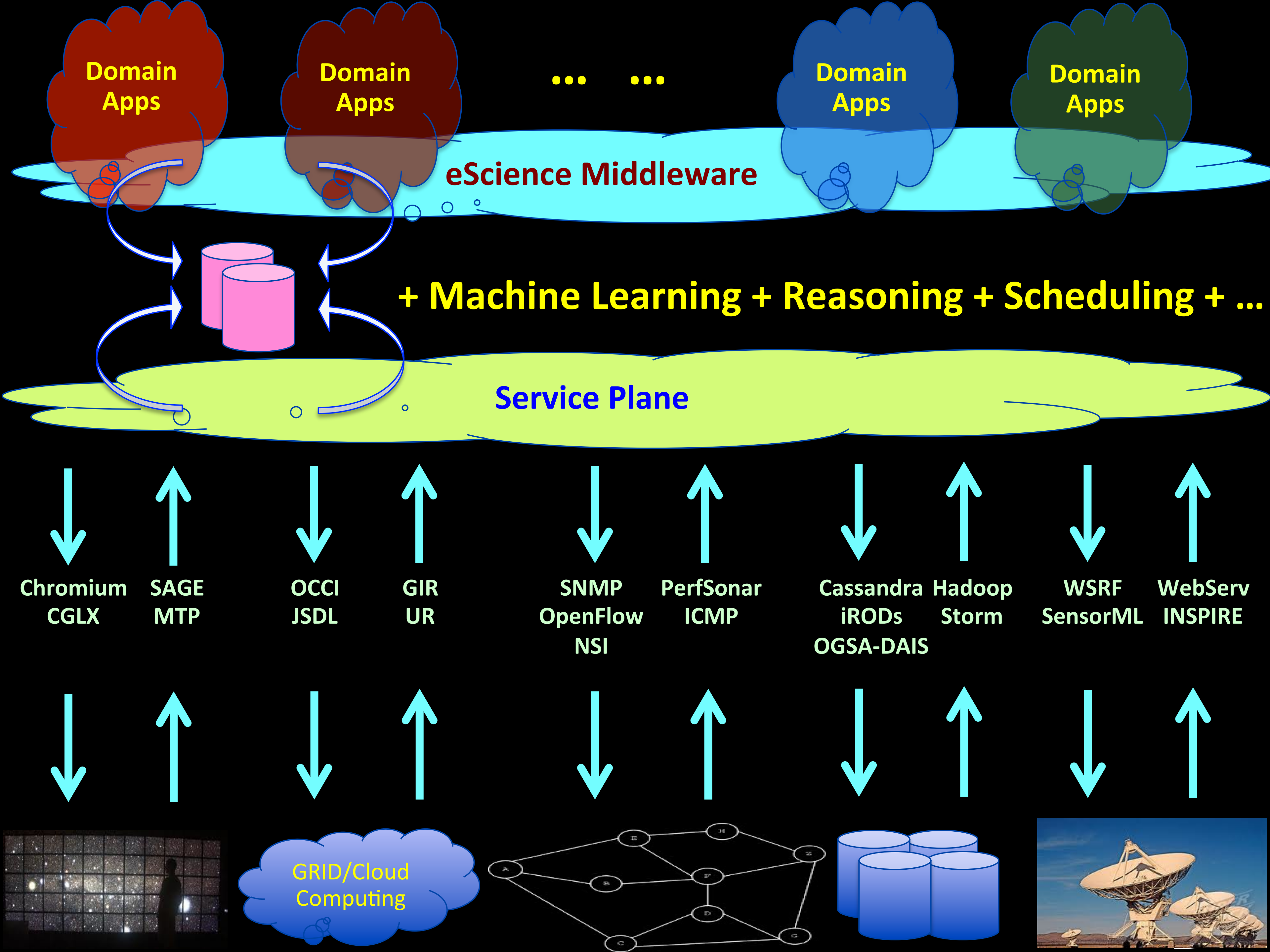
Your Career

ECO-Scheduling

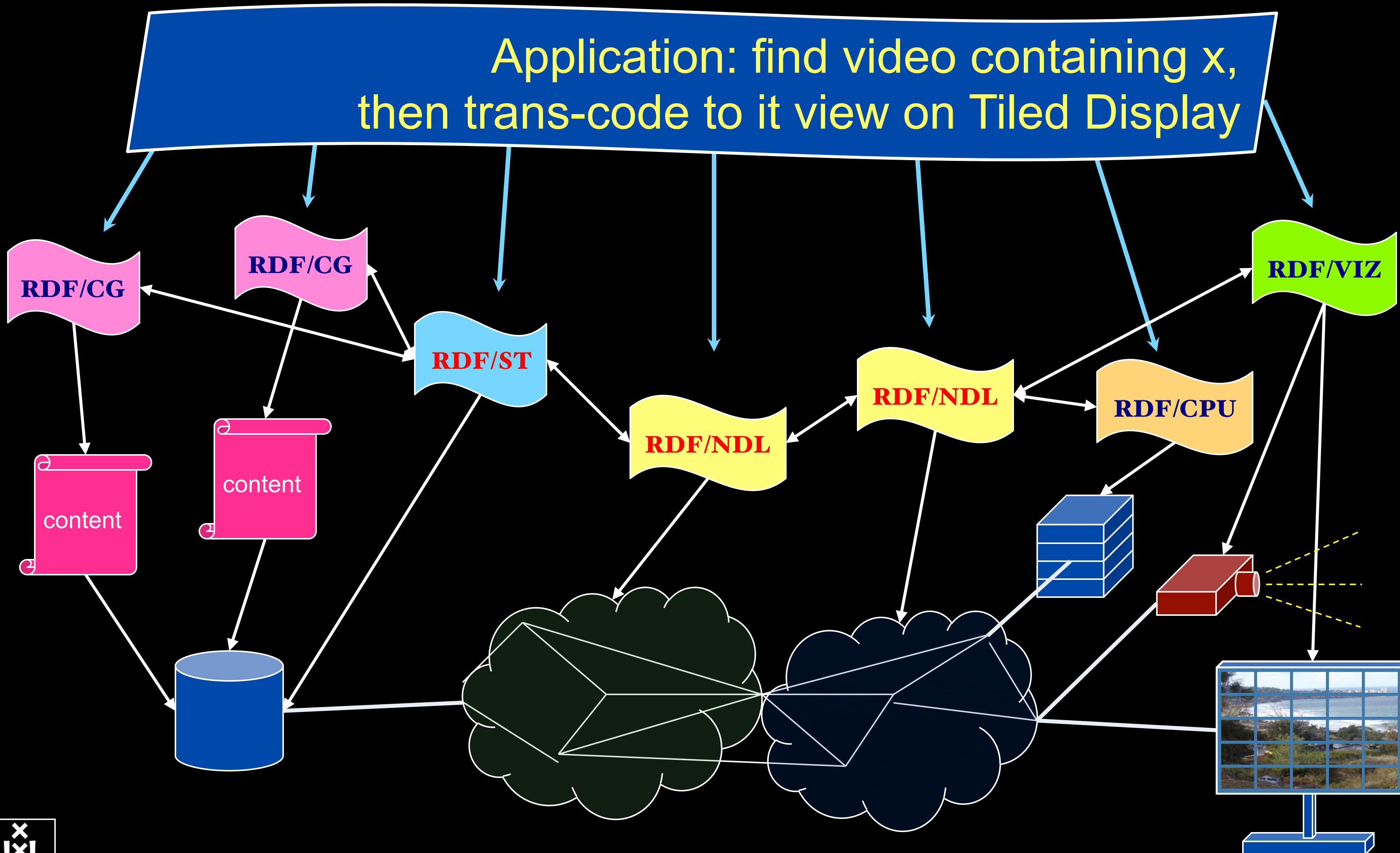
I want to

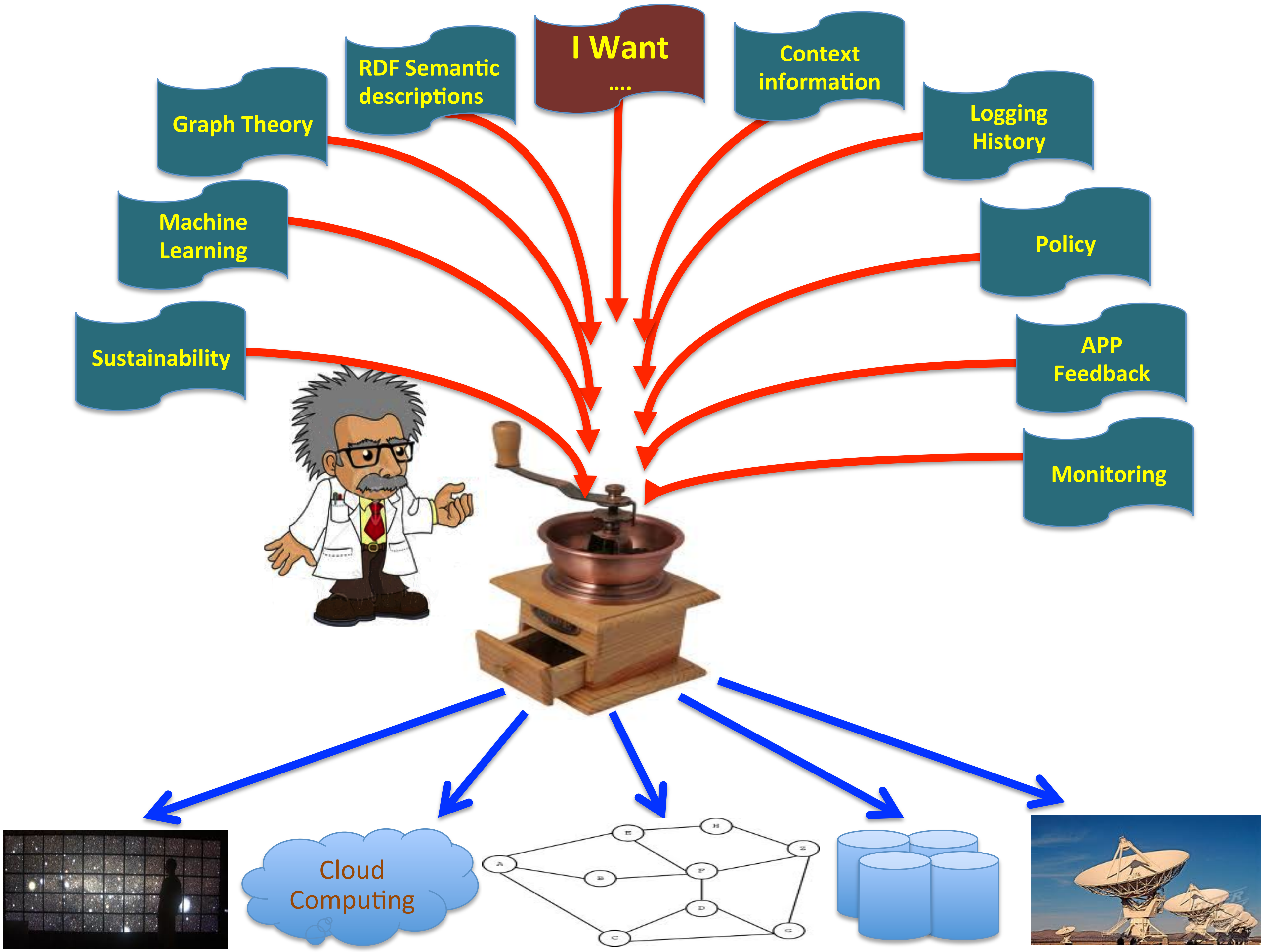
"Show Big Bug Bunny in 4K on my Tiled Display using green Infrastructure"

- Big Bugs Bunny can be on multiple servers on the Internet.
- Movie may need processing / recoding to get to 4K for Tiled Display.
- Needs deterministic Green infrastructure for Quality of Experience.
- Consumer / Scientist does not want to know the underlying details.
  ➔ His refrigerator also just works.


Domain Apps
Domain Apps
… …
Domain Apps
Domain Apps
eScience Middleware
+ Machine Learning + Reasoning + Scheduling + …
Service Plane
Chromium CGLX
SAGE MTP
OCCI JSDL
GIR UR
SNMP OpenFlow NSI
PerfSonar ICMP
Cassandra iRODs OGSA-DAIS
Hadoop Storm
WSRF SensorML
WebServ INSPIRE
GRID/Cloud Computing

RDF describing Infrastructure
"I want"




I Want
....
RDF Semantic descriptions
Context information
Graph Theory
Logging History
Machine Learning
Policy
Sustainability
APP Feedback
Monitoring
Cloud Computing
A
B
C
D
E
F
G
H
Z

Why?

Because we can!

Paper #1 + Q's

TRANSLIGHT

A GLOBAL-SCALE LAMBDAGRID FOR E-SCIENCE

This global experiment wants to see if high-end applications needing transport capacities of multiple Gbps for up to hours at a time can be handled through an optical bypass network.

Tom DeFanti, Cees de Laat, Joe Mambretti, Kees Neggers, Bill St. Arnaud.

Communications of the ACM, Volume 46, Issue 11 (November 2003), Pages: 34 – 41.

http://delaat.net/pubs/2003-j-6.pdf

Paper #1 + Q's

- Q1: This article is now 10 years old. Back then Twitter did not exist. What do you think will be the drivers for network capacity demand in Science and Society 10 years from now?

- Q2: List arguments why one would use photonic networks directly in science applications and arguments why not tu use photonics directly but use current Internet.

- Q3: This question is not directly from this paper but fun to figure out via search on the web: Fiber cable systems under the ocean are very expensive and cost 100's of millions to put in place. How many fibers do they put in one cable and why that amount?

Paper #2 + Q's

A distributed topology information system for optical networks based on the semantic web.

Jeroen van der Ham, Freek Dijkstra, Paola Grosso, Ronald van der Pol, Andree Toonk, Cees de Laat

Elsevier Journal on Optical Switching and Networking, Volume 5, Issues 2-3, June 2008, pp 85-93.

http://delaat.net/pubs/2008-j-4.pdf

Paper #2 + Q's

- Q1: Suppose this method of describing networks is a total worldwide success and allows to find superfast networking paths through the CI (CyberInfrastructure). The question becomes: Does it scale? Can you find reasons why and/or why not it could scale up to the size of the internet?

- Q2: Are the described methods and framework fault tolerant? If not, then list the issues in your view. What do you see best ways to do something about it.

- Q3: List advantages of NDL, or more generically, using semantic web methods for describing cyber infrastructure?

The constant factor in our field is Change!

The 50 years it took Physicists to find one particle, the Higgs, we came from:

“Fortran goto”, Unix, c, SmallTalk, DECnet, TCP/IP, c++, Internet, WWW, Semantic Web, Photonic networks, Google, grid, cloud, Data^3, App

to:

DDOS attacks destroying Banks and Bitcoins.

Conclusion:

Need for Safe, Smart, Resilient Sustainable Infrastructure.