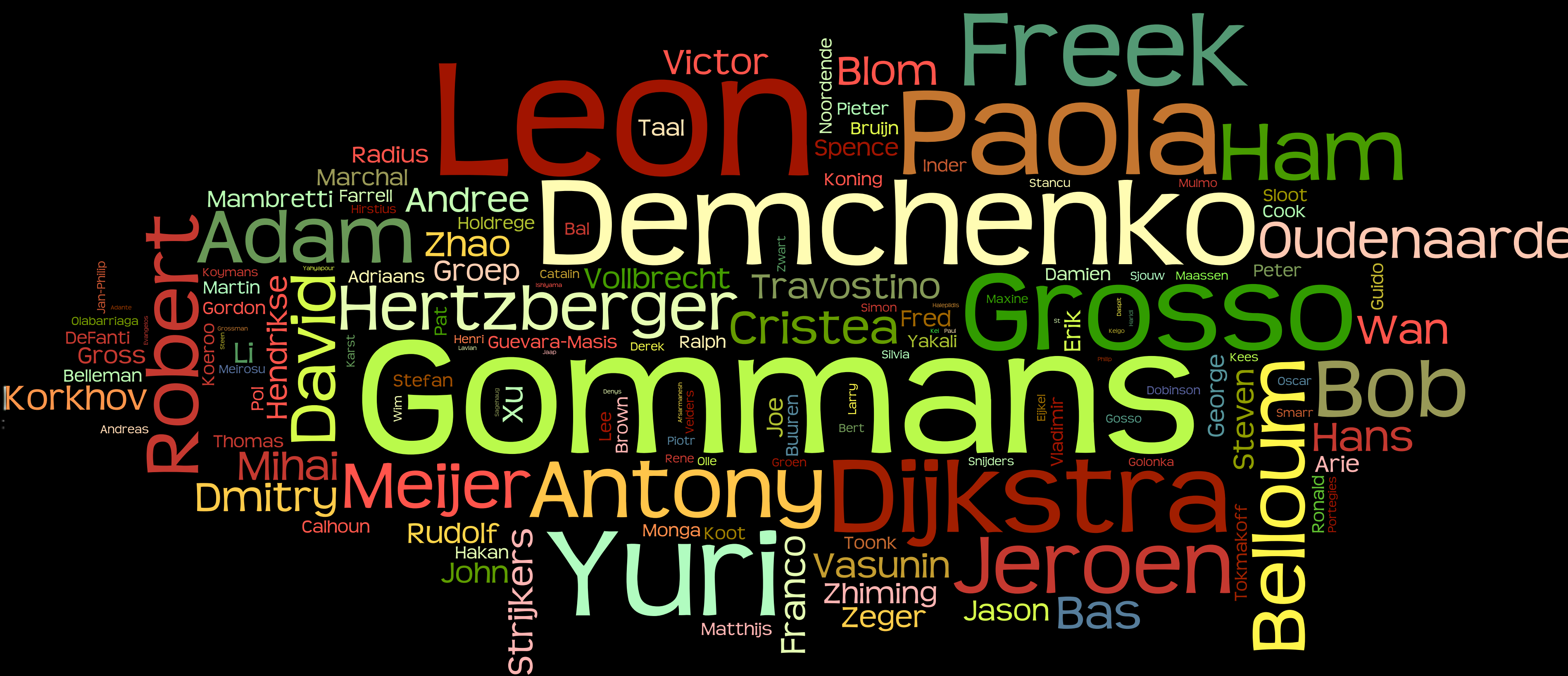
System and Network Engineering Research for Big Data Sciences

Cees de Laat

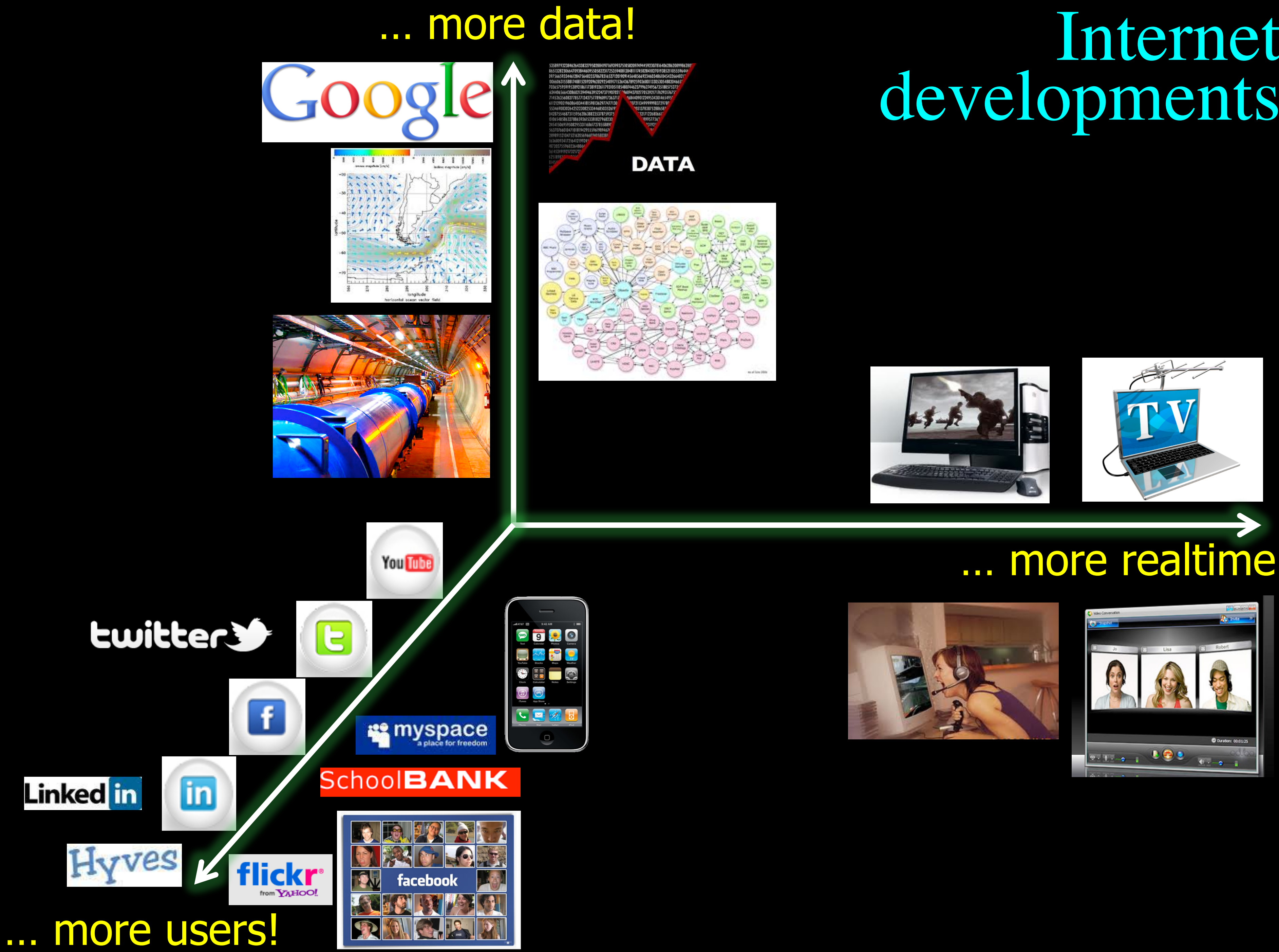


Mission SNE

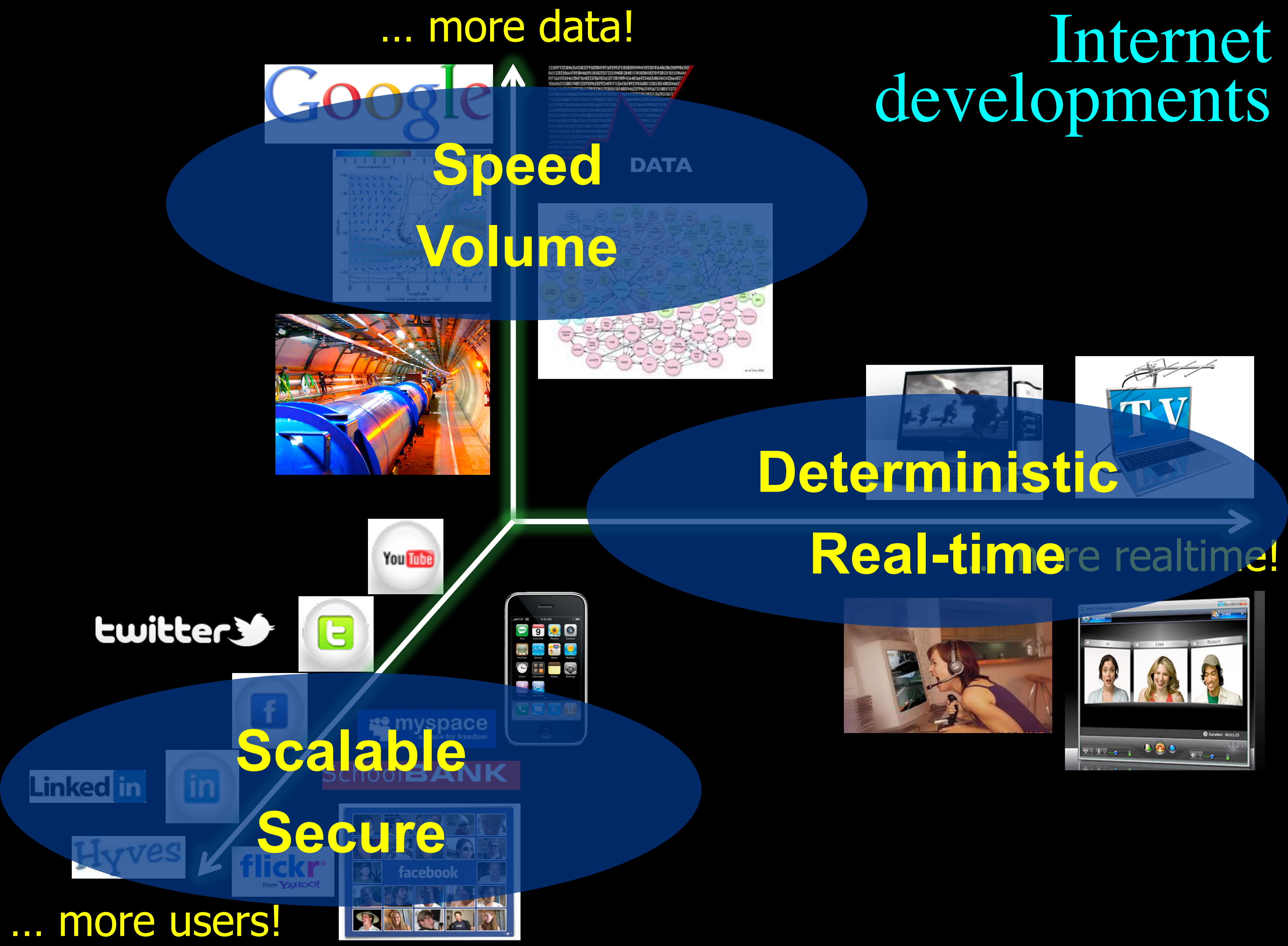
Can we create smart and safe data processing infrastructures that can be tailored to diverse application needs?

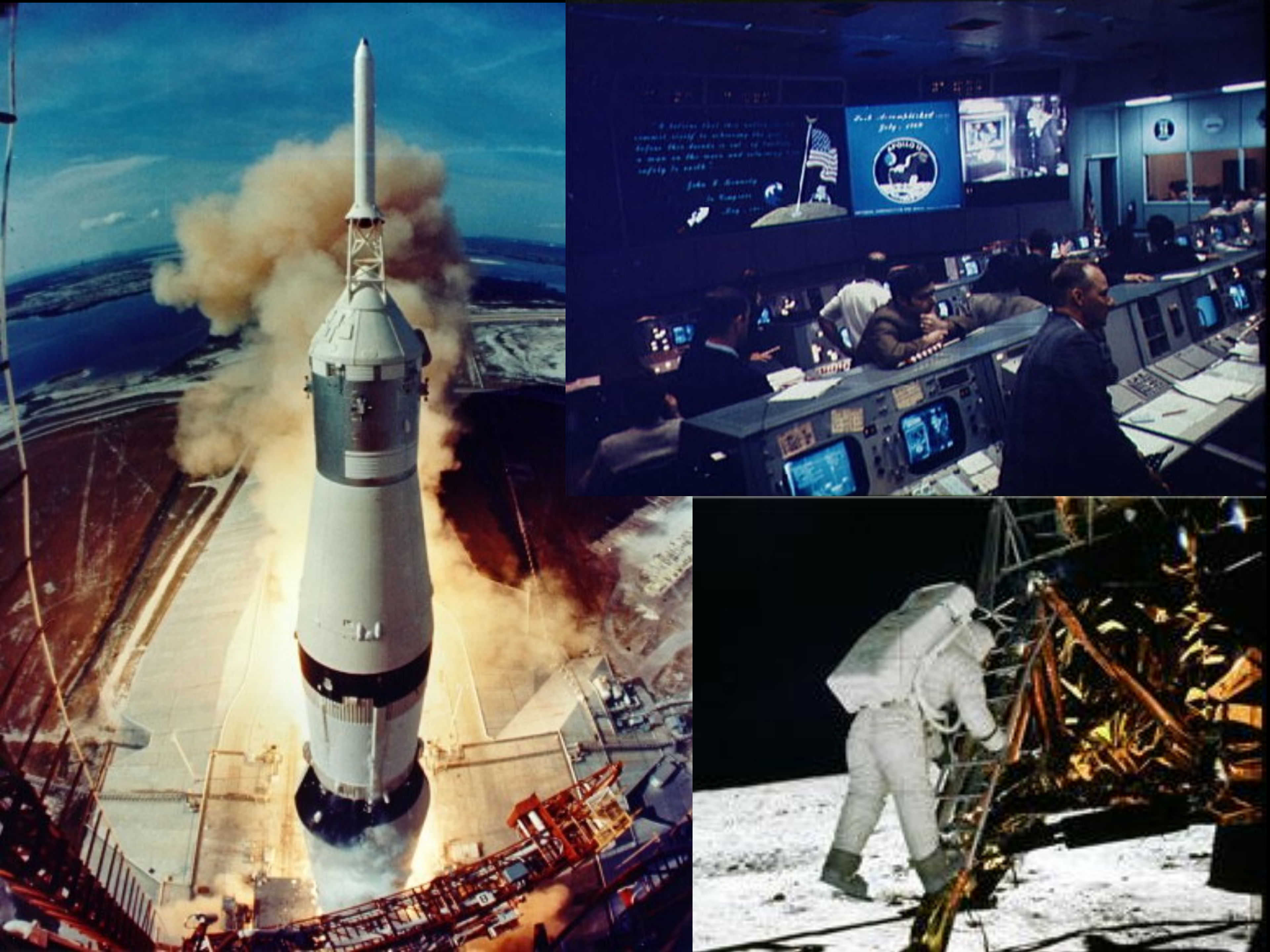
- *Capacity*
  - *Bandwidth on demand, QoS, architectures, photonics, performance*
- *Capability*
  - *Programmability, virtualization, complexity, semantics, workflows*
- *Security*
  - *Authorization, Anonymity, integrity of data in distributed data processing*
- *Sustainability*
  - *Greening infrastructure,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

Internet developments



Internet developments

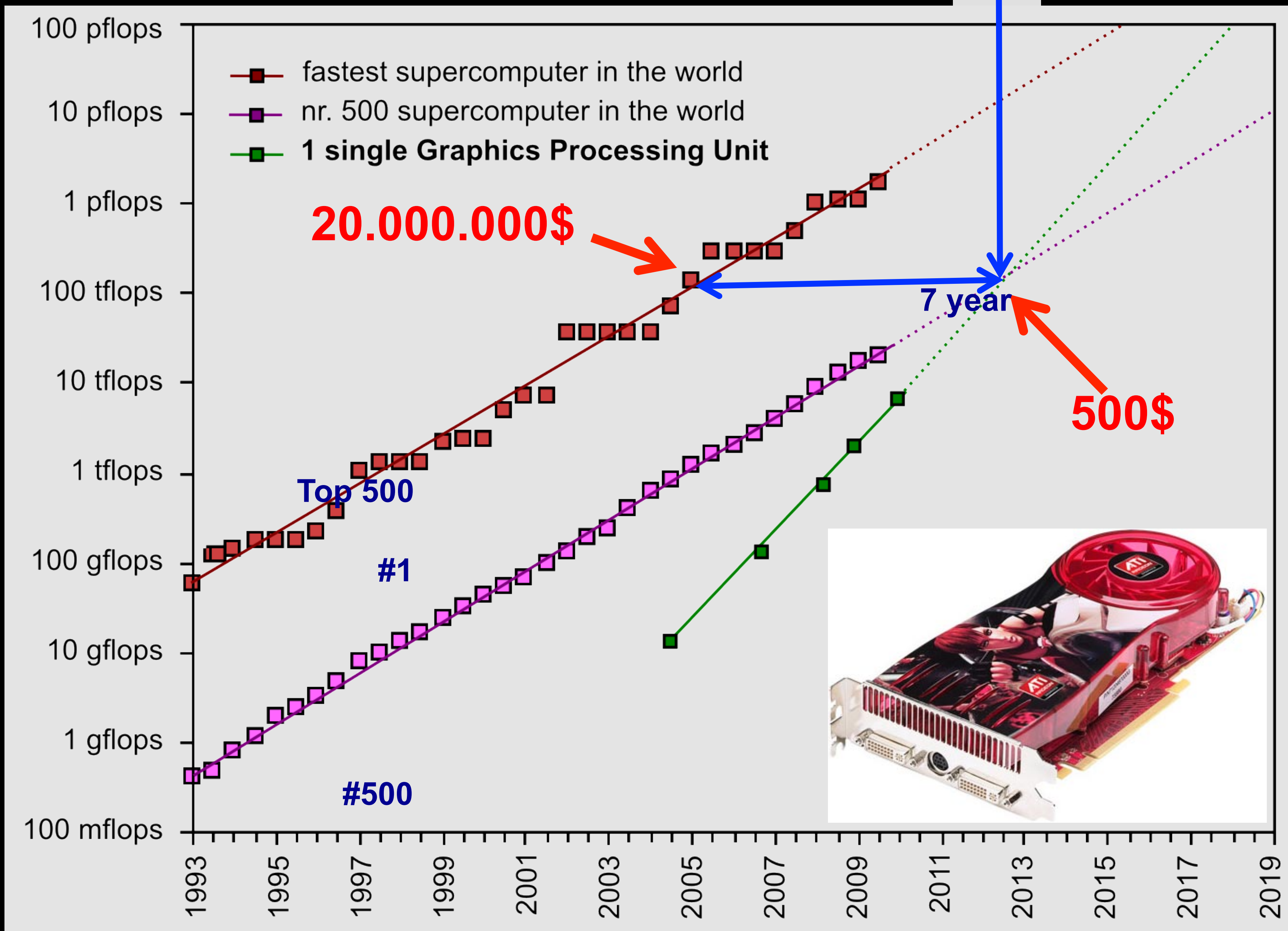




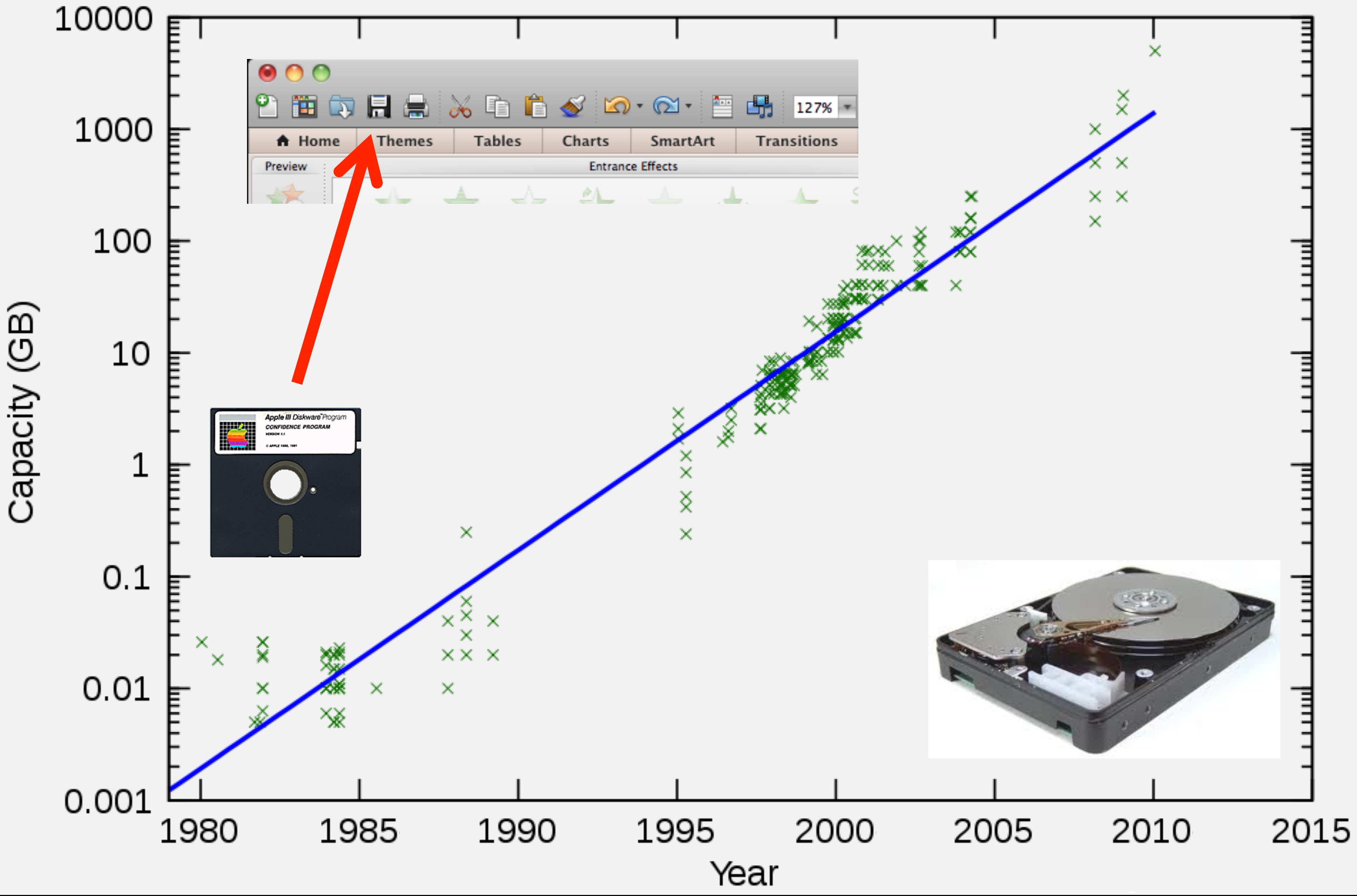

UPLINK ACTY
TEMP
AUTO
GIMBAL LOCK
PROG
RESTART
NO ATT
TRACKER
STBY
KEY REL
OPR ERR
COMP ACTY
VERB
NOUN
CLR
ENTR
RSET


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

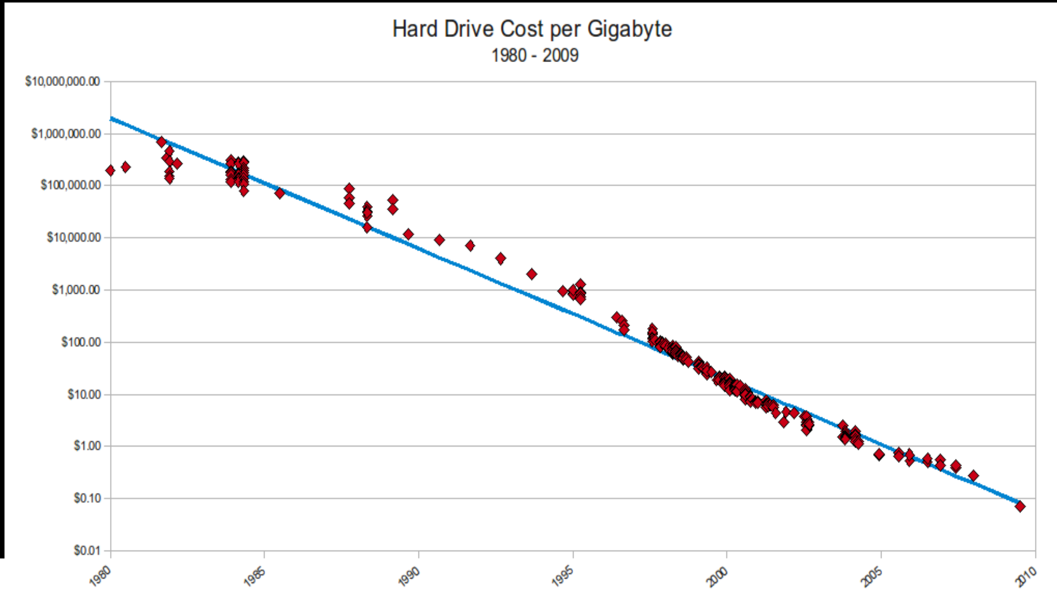
GPU cards are distruptive!



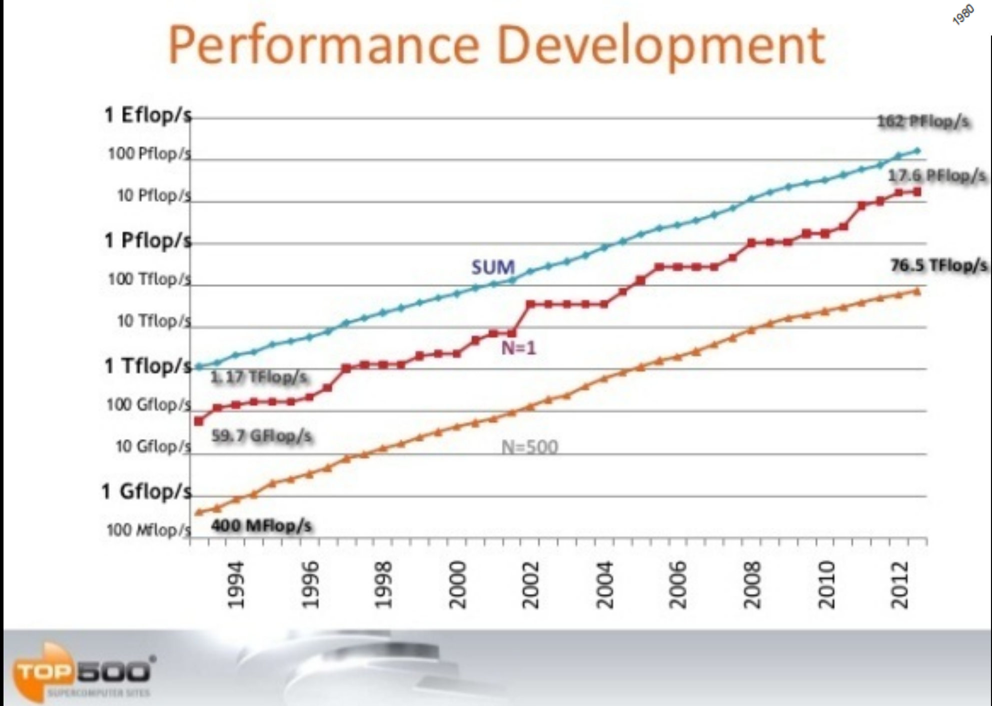
Data storage: doubling every 1.5 year!



Computing vs Data



Computing per unit cost has doubled roughly every 18 months.

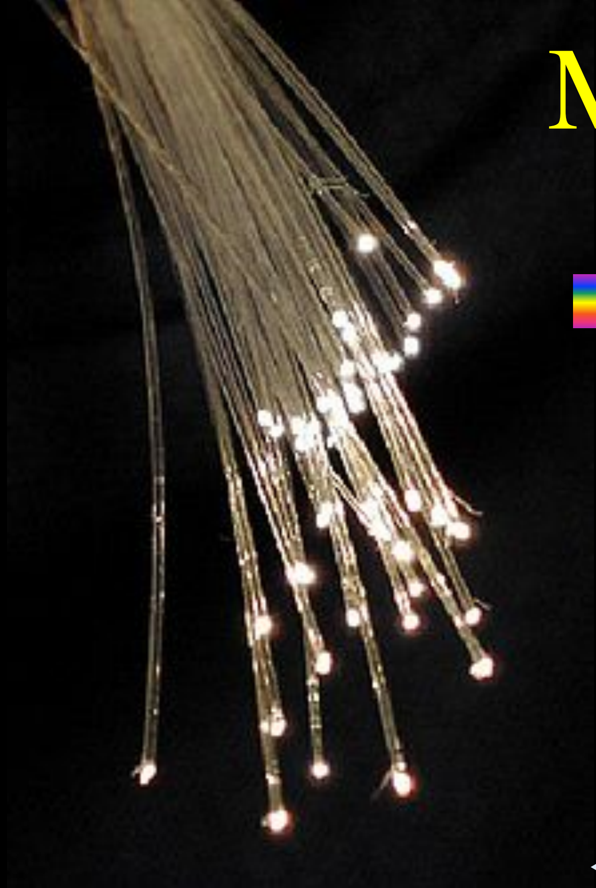


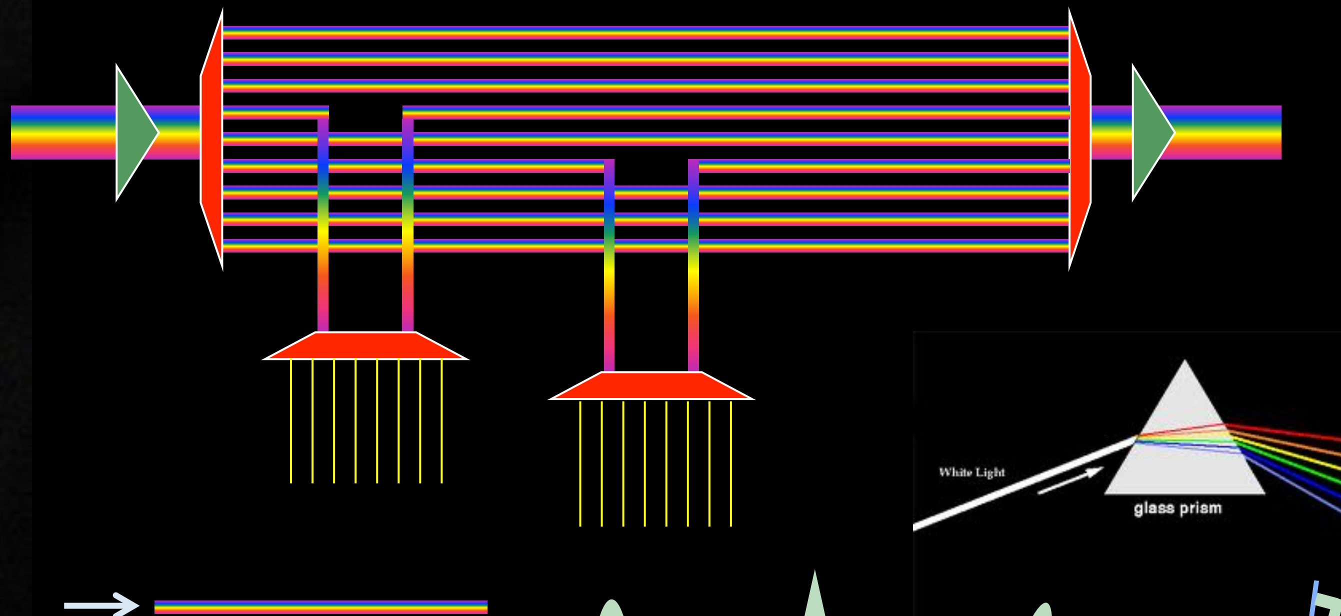
Space per unit cost has doubled roughly every 14 months.

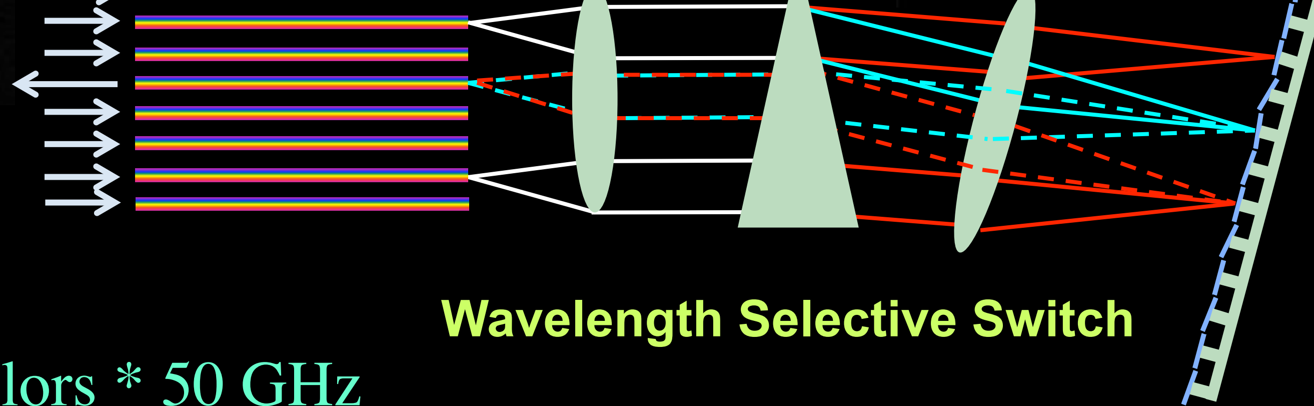
So: data becomes exponentially uncomputable.

http://www.mkomo.com/cost-per-gigabyte

Multiple colors / Fiber







Per fiber: ~ 80-100 colors * 50 GHz

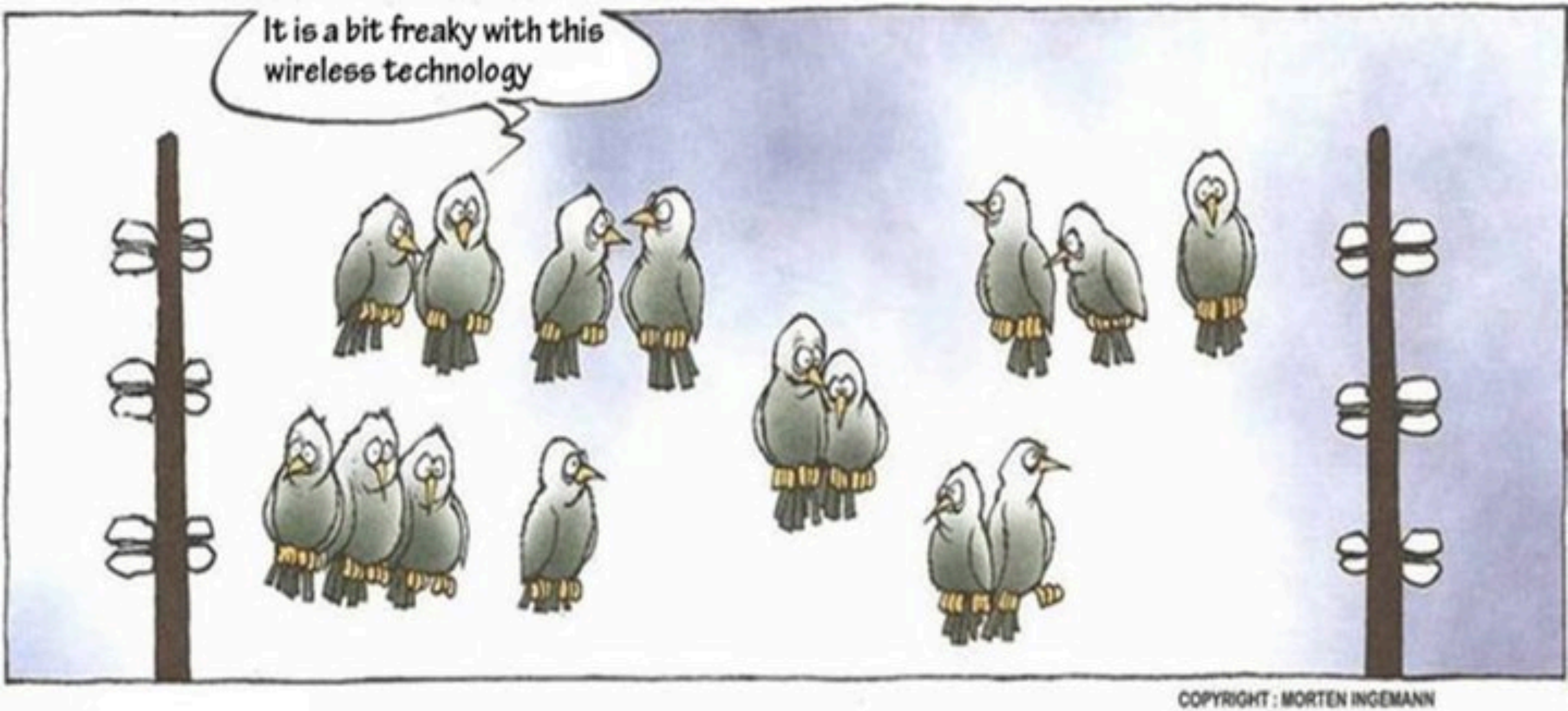
Per color: 10 – 40 – 100 Gbit/s

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

New: Hollow Fiber!

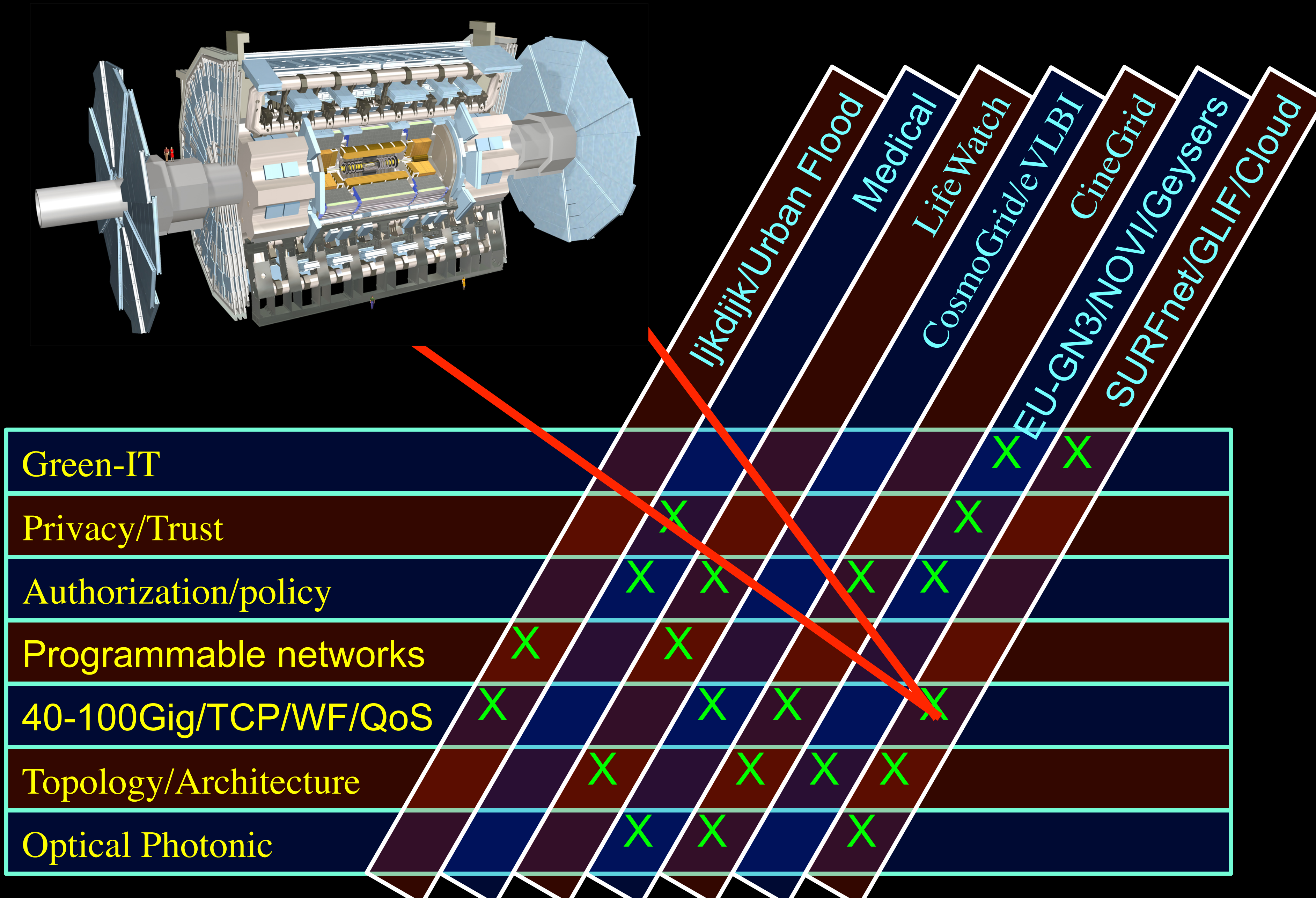
➔ less RTT!

Wireless Networks



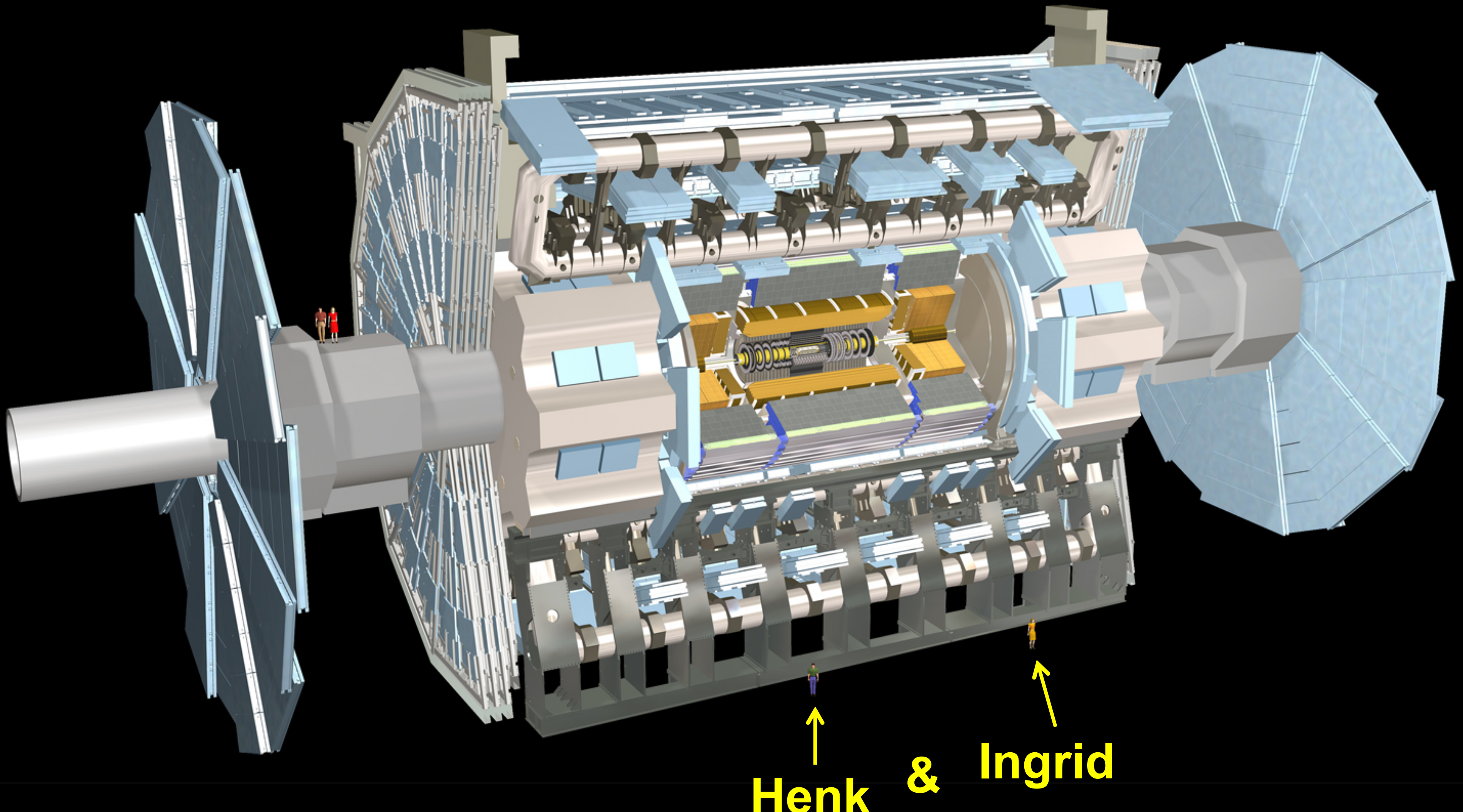
protocol LAN due to the easy comparison and convenience in the digital home. While consumer PC products has just started to migrate to a much higher bandwidth of 802.11n wireless LAN now working on next-generation standard definition is already in progress.

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

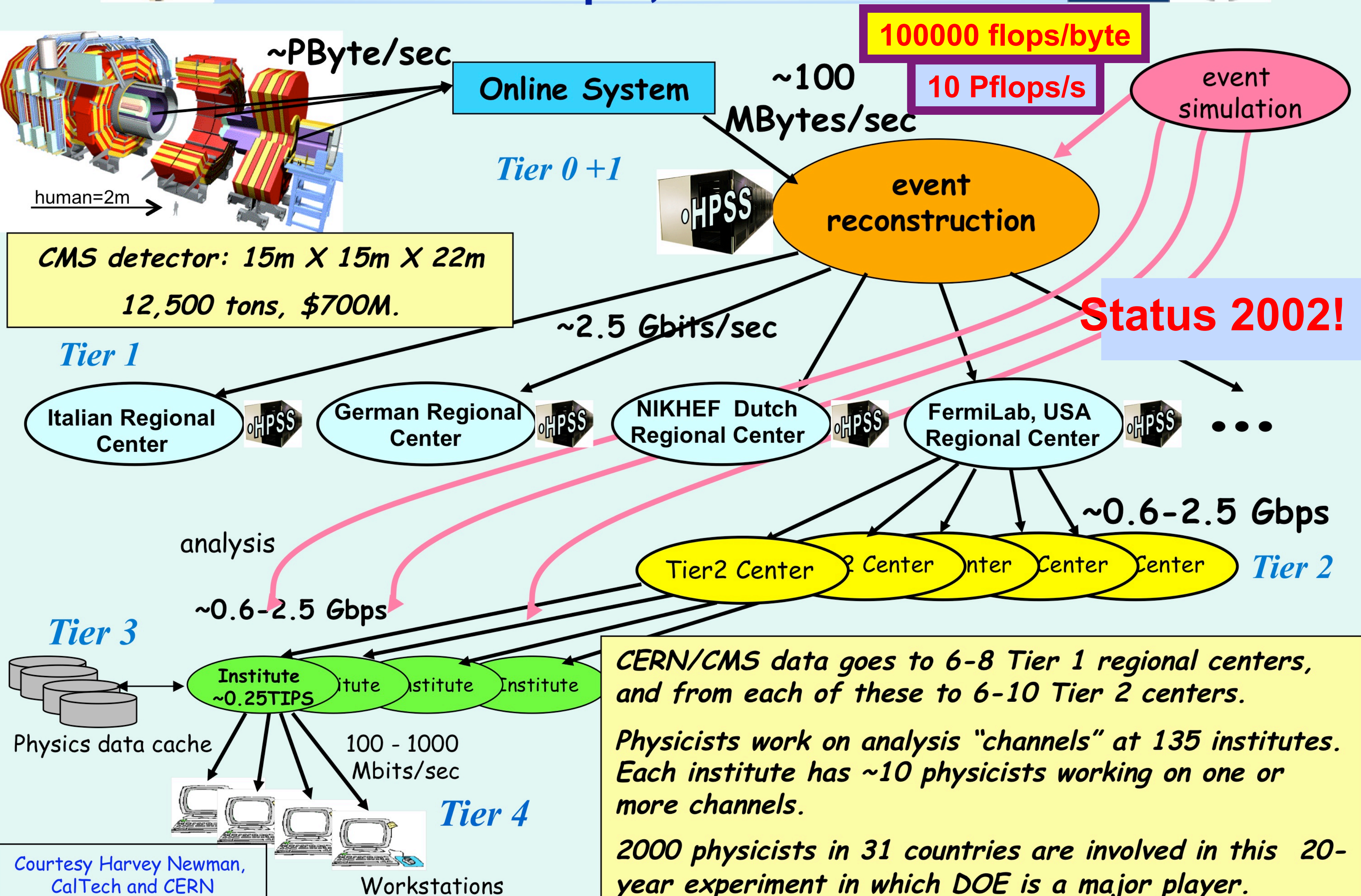
ATLAS detector @ CERN Geneve



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.

LOFAR as a Sensor Network

20 flops/byte

– LOFAR is a large distributed research infrastructure:

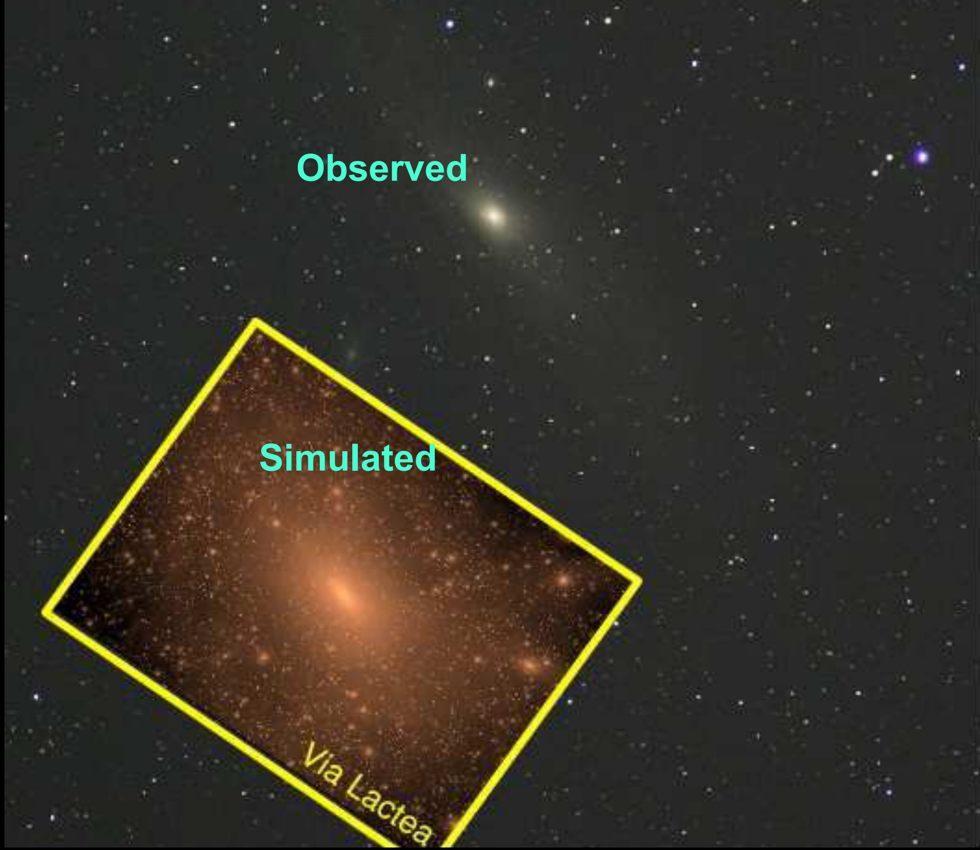
2 Tflops/s

- 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

CosmoGrid

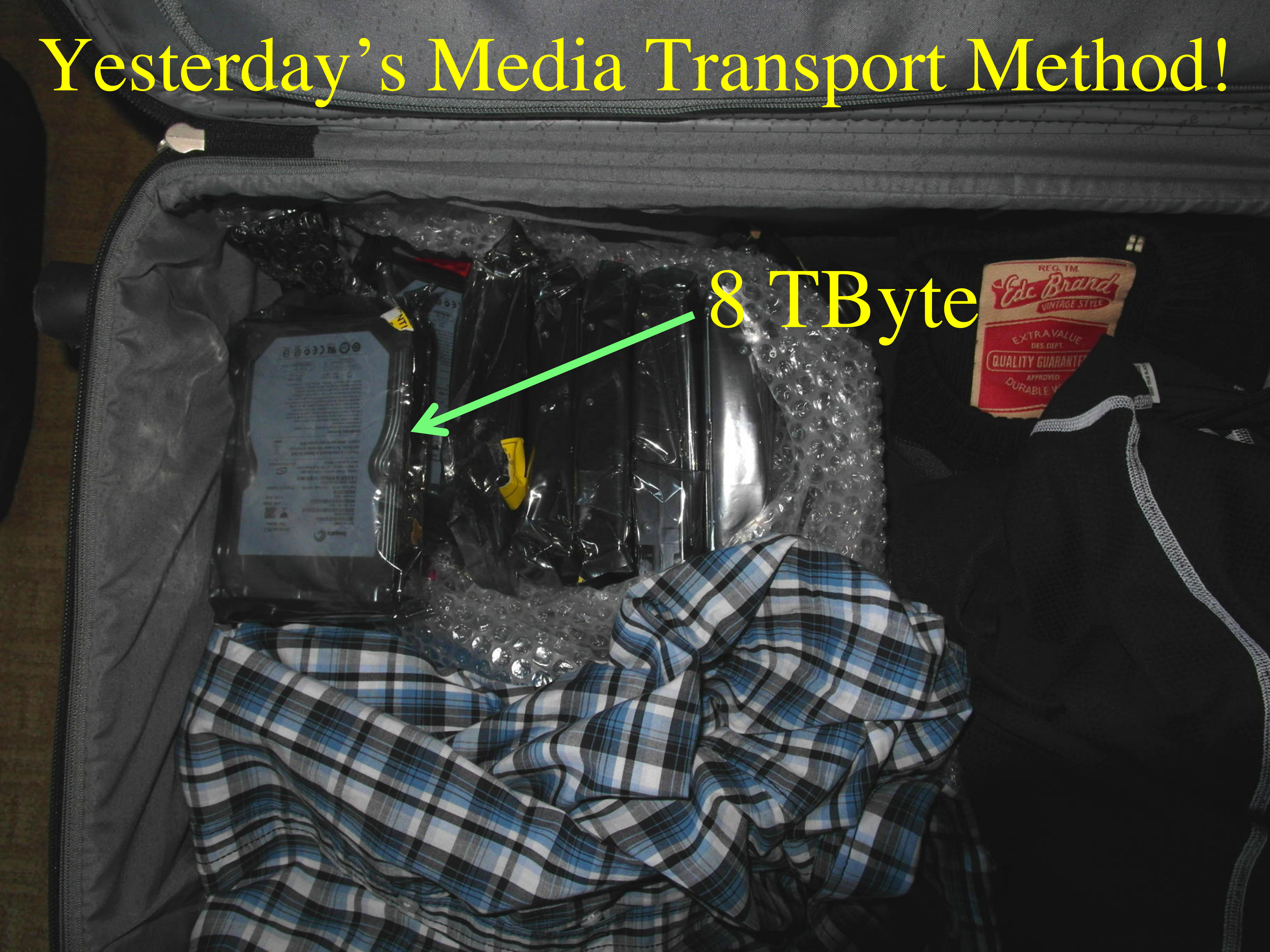
Simon Portegies Zwart et al.

- Motivation:
  - previous simulations found >100 times more substructure than is observed!
- Simulate large structure formation in the Universe
- Method: Cosmological *N*-body code
- Computation: Intercontinental SuperComputer Grid
- Current (2013) problem:
  - 2 PByte data in Oak Ridge!

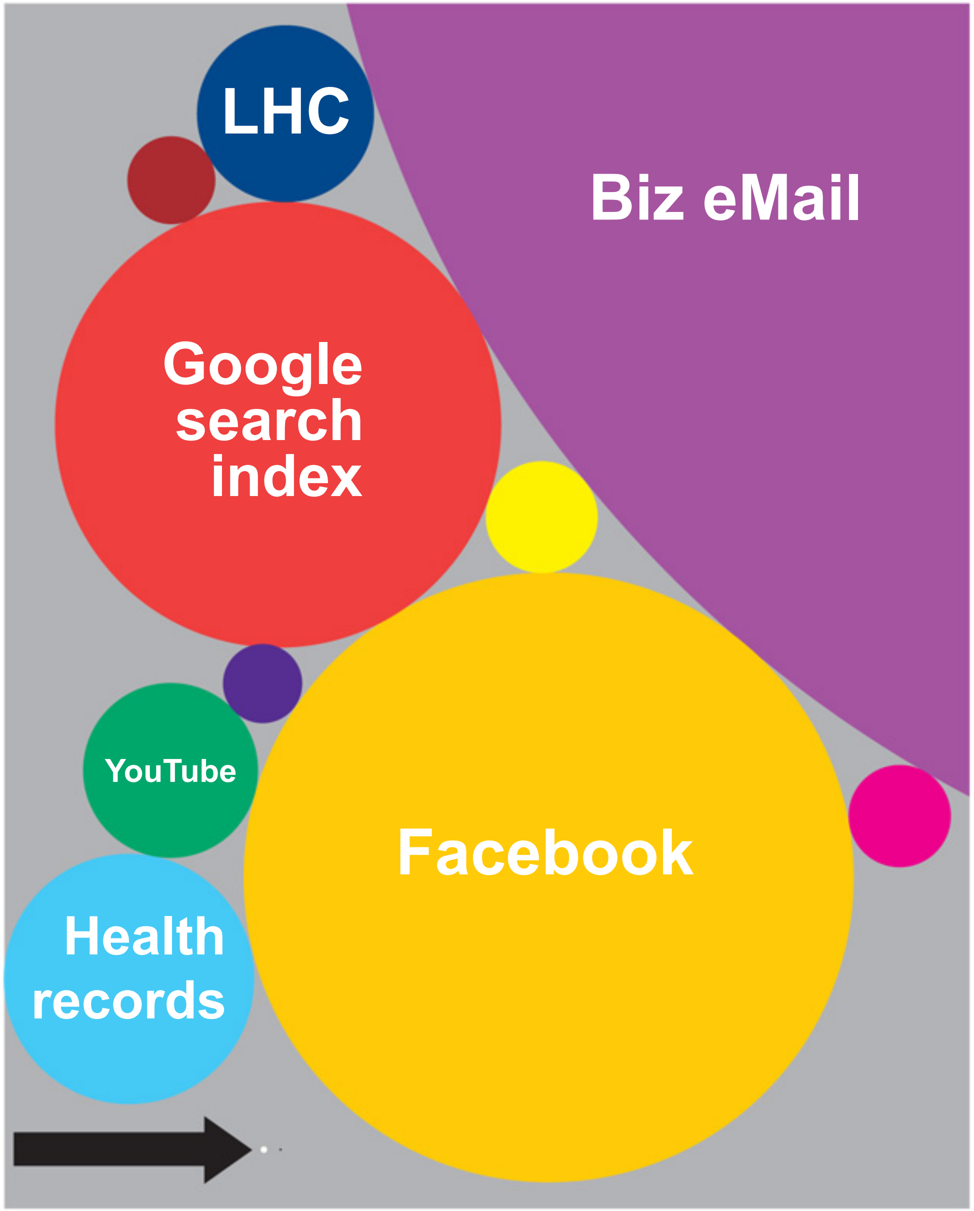


Moving Cinegrid Objects Globally

- **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!
8 TByte

What Happens in an Internet Minute?



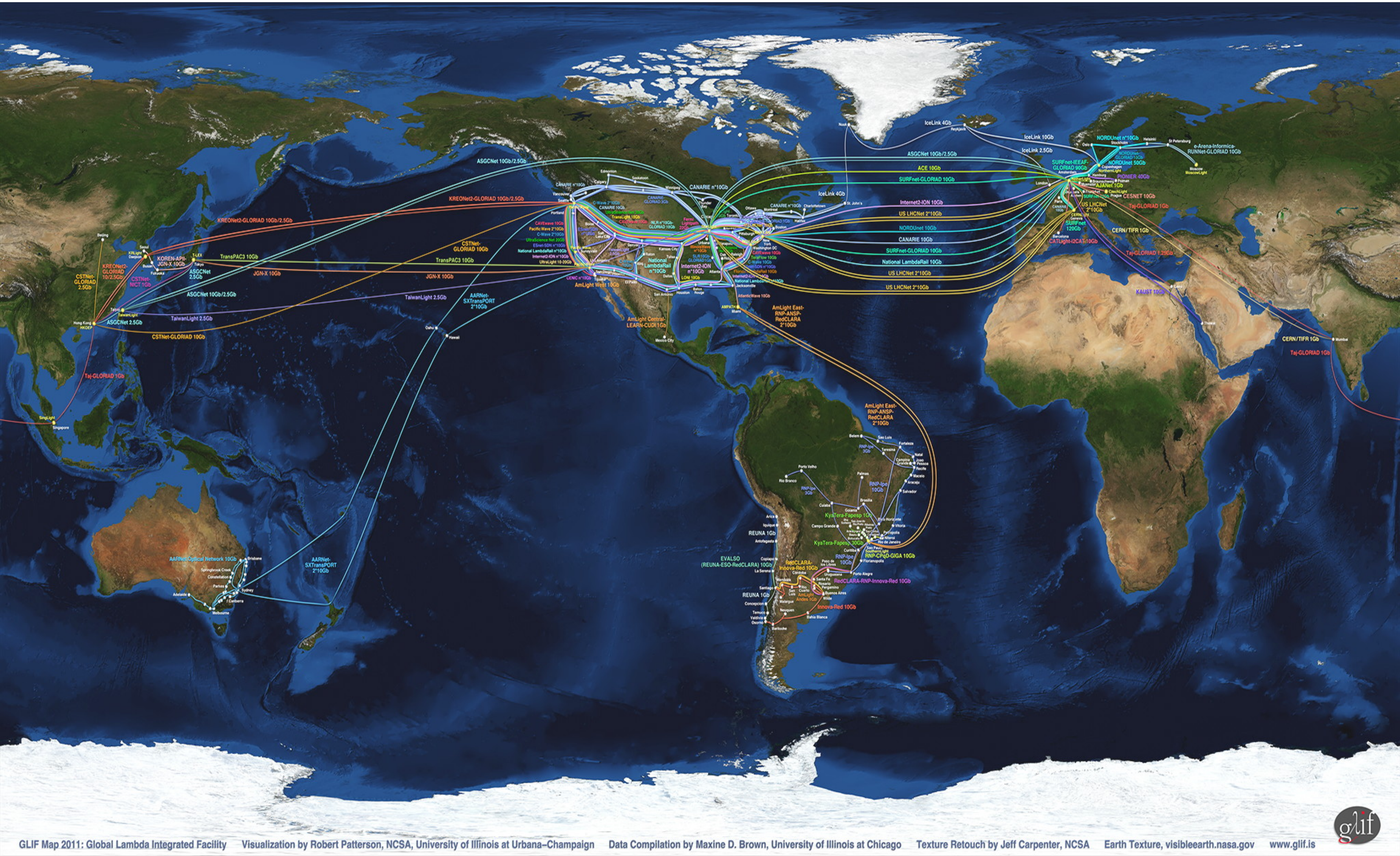
Size of data sets in terabytes

Business email sent per year	2,986,100
Content uploaded to Facebook each year	182,500
Google's search index	97,656
Kaiser Permanente's digital health records	30,720
Large Hadron Collider's annual data output	15,360
Videos uploaded to YouTube per year	15,000
National Climactic Data Center database	6,144
Library of Congress' digital collection	5,120
US Census Bureau data	3,789
Nasdaq stock market database	3,072
Tweets sent in 2012	19
Contents of every print issue of WIRED	1.26

There is always a bigger fish

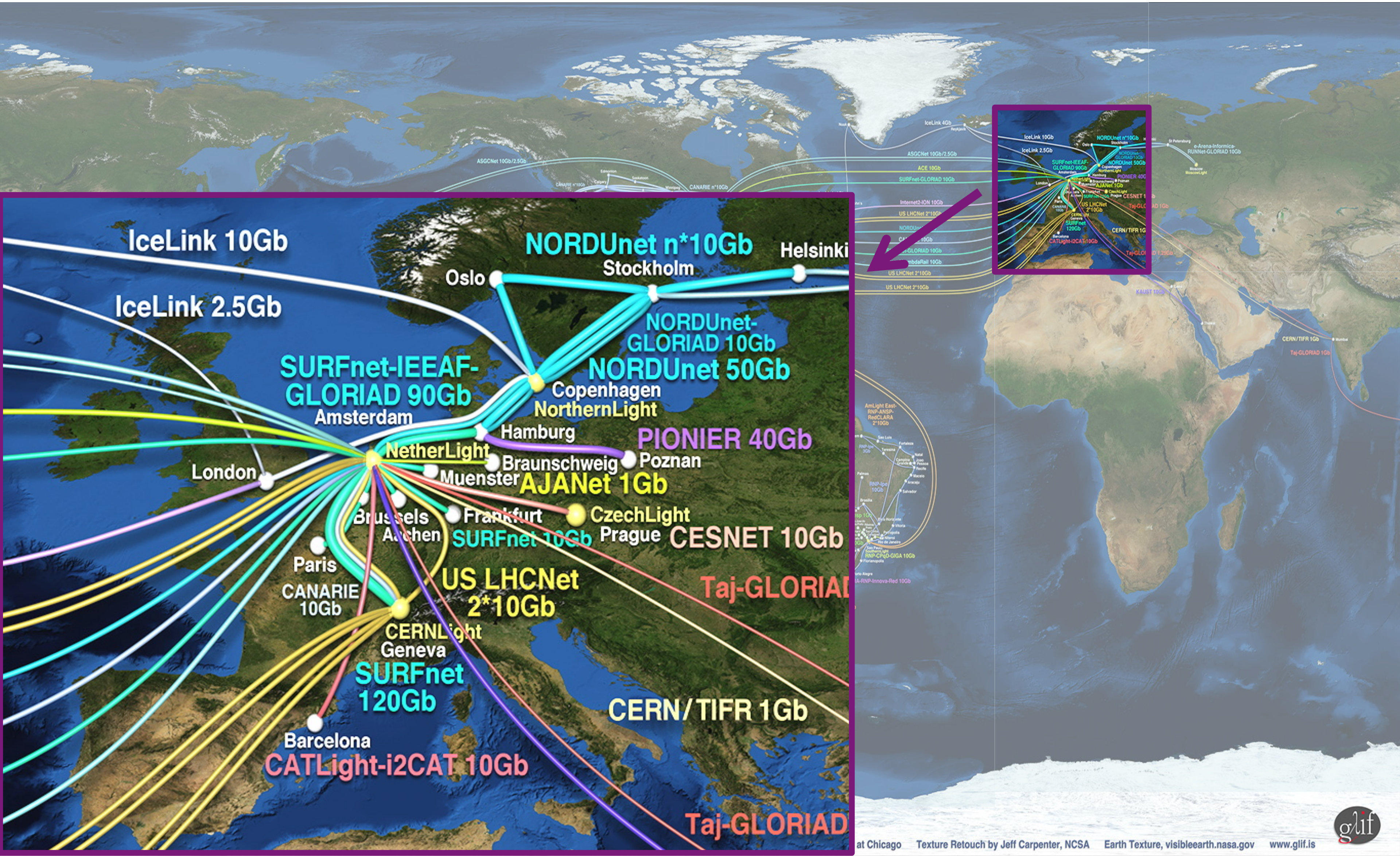
The GLIF – LightPaths around the World

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.



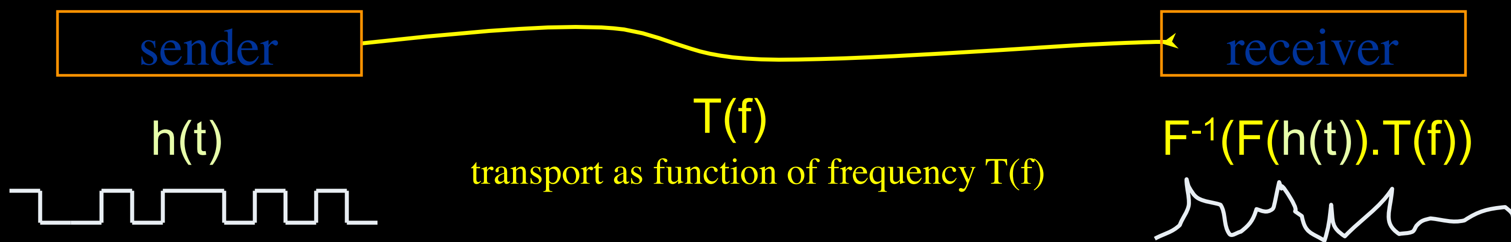
The GLIF – LightPaths around the World

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.



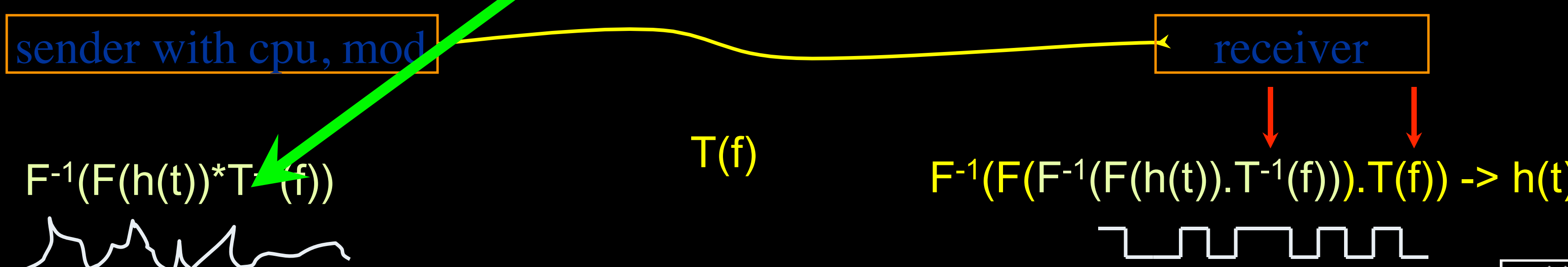
Dispersion compensating modem: eDCO from NORTEL

(Try to Google eDCO :-)



Solution in 5 easy steps for dummy's :

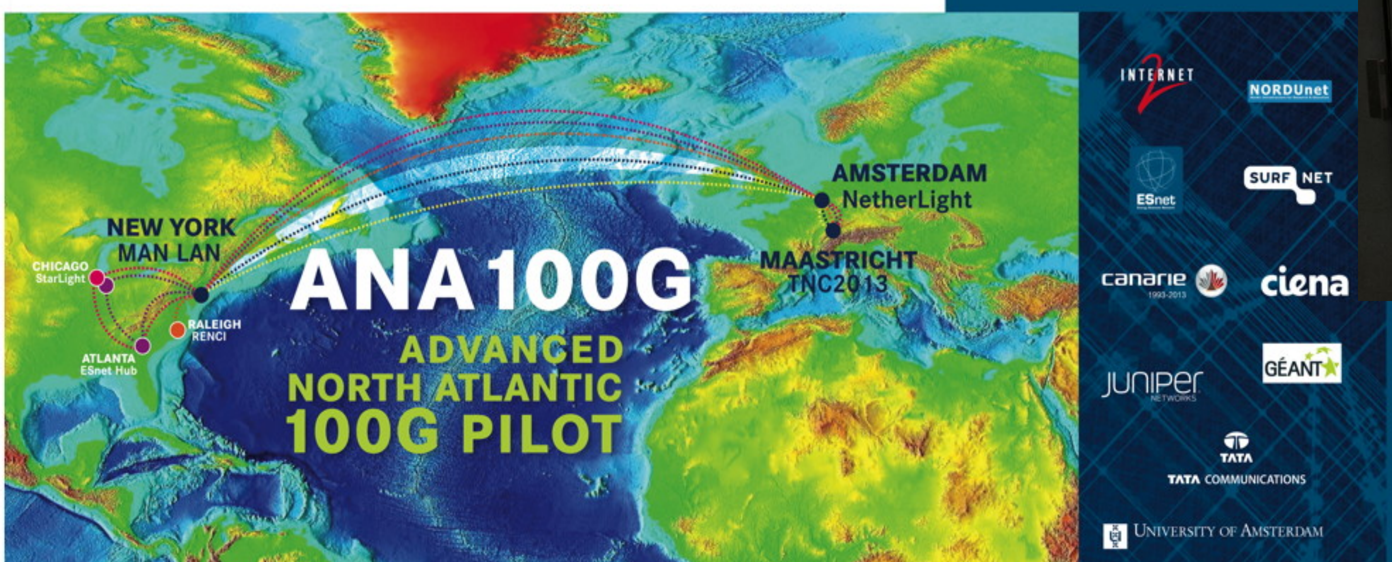
1. try to figure out T(f) by trial and error
2. invert T(f) -> $T^{-1}(f)$
3. computationally multiply $T^{-1}(f)$ with Fourier transform of bit pattern to send
4. inverse Fourier transform the result from frequency to time space
5. modulate laser with resulting h'(t) = $F^{-1}(F(h(t)).T^{-1}(f))$



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

ExoGeni @ UvA

Installed and up June 3th 2013



TNC2013 DEMOS JUNE, 2013

DEMO	TITLE	OWNER	AFFILIATION	E-MAIL	A-SIDE	Z-SIDE	PORTS(S) MAN LAN	PORTS(S) TNC2013	DETAILS
1	Big data transfers with multipathing, OpenFlow and MPTCP	Ronald van der Pol	SURFnet	ronald.vanderpol@surfnet.nl	TNC/MECC, Maastricht NL	Chicago, IL	Existing 100G link between internet2 and ESnet	2x40GE (Juniper)+ 2x10GE (OME6500)	In this demonstration we show how multipathing, OpenFlow and Multipath TCP (MPTCP) can help in large file transfers between data centres (Maastricht and Chicago). An OpenFlow application provisions multiple paths between the servers and MPTCP will be used on the servers to simultaneously send traffic across all those paths. This demo uses 2x40G on the transatlantic 100G link. ESnet provides 2x40G between MAN LAN and StarLight, ACE and USLHCnet provide additional 10GEs.
2	Visualize 100G traffic	Inder Monga	ESnet	imonga@es.net					Using an SNMP feed from the Juniper switch at TNC2013,and/or Brocade AL2S node in MANLAN, this demo would visualize the total traffic on the link, of all demos aggregated. The network diagram will show the transatlantic topology and some of the demo topologies.
3	How many modern servers can fill a 100Gbps Transatlantic Circuit?	Inder Monga	ESnet	imonga@es.net	Chicago, Ill	TNC showfloor	1x 100GE	8x 10GE	In this demonstration, we show that with the proper tuning and tool, only 2 hosts on each continent can generate almost 80Gbps of traffic. Each server has 4 10G NICS connected to a 40G virtual circuit, and has iperf3 running to generate traffic. ESnet's new "iperf3" throughput measurement tool, still in 'beta', combines the best features from other tools such as iperf, nuttcp, and netperf. See: https://my.es.net/demos/tnc2013/
4	First European ExoGENI at Work	Jeroen van der Ham	UvA	vdham@uva.nl	RENCI, NC	UvA, Amsterdam, NL	1x 10GE	1x 10GE	The ExoGENI racks at RENCI and UvA will be interconnected over a 10G pipe and be on continuously, showing GENI connectivity between Amsterdam and the rest of the GENI nodes in the USA.
5	Up and down North Atlantic @ 100G	Michael Enrico	DANTE	michael.enrico@dante.net	TNC showfloor	TNC showfloor	1x 100GE	1x 100GE	The DANTE 100GE test set will be placed at the TNC2013 showfloor and connected to the Juniper at 100G. When this demo is running a loop @ MAN LAN's Brocade switch will ensure that the traffic sent to MAN LAN returns to the showfloor. On display is the throughput and RTT (to show the traffic traveled the Atlantic twice)

Connected via the new 100 Gb/s transatlantic

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.

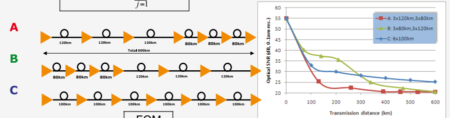
In this demonstration we will investigate the performance of a 40Gb/s PM-QPSK alien wavelength installed on a 10Gb/s DWDM infrastructure.

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.

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

Lj, span losses in dB
N, number of spans

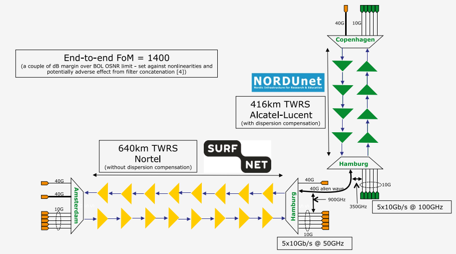


	FOM
A	5504
B	5504
C	1897

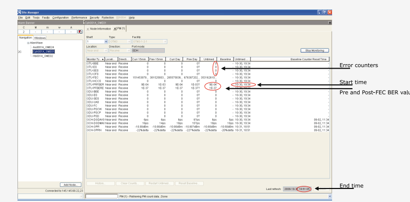
Easy-to-use formula that accurately quantifies transmission system performance

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

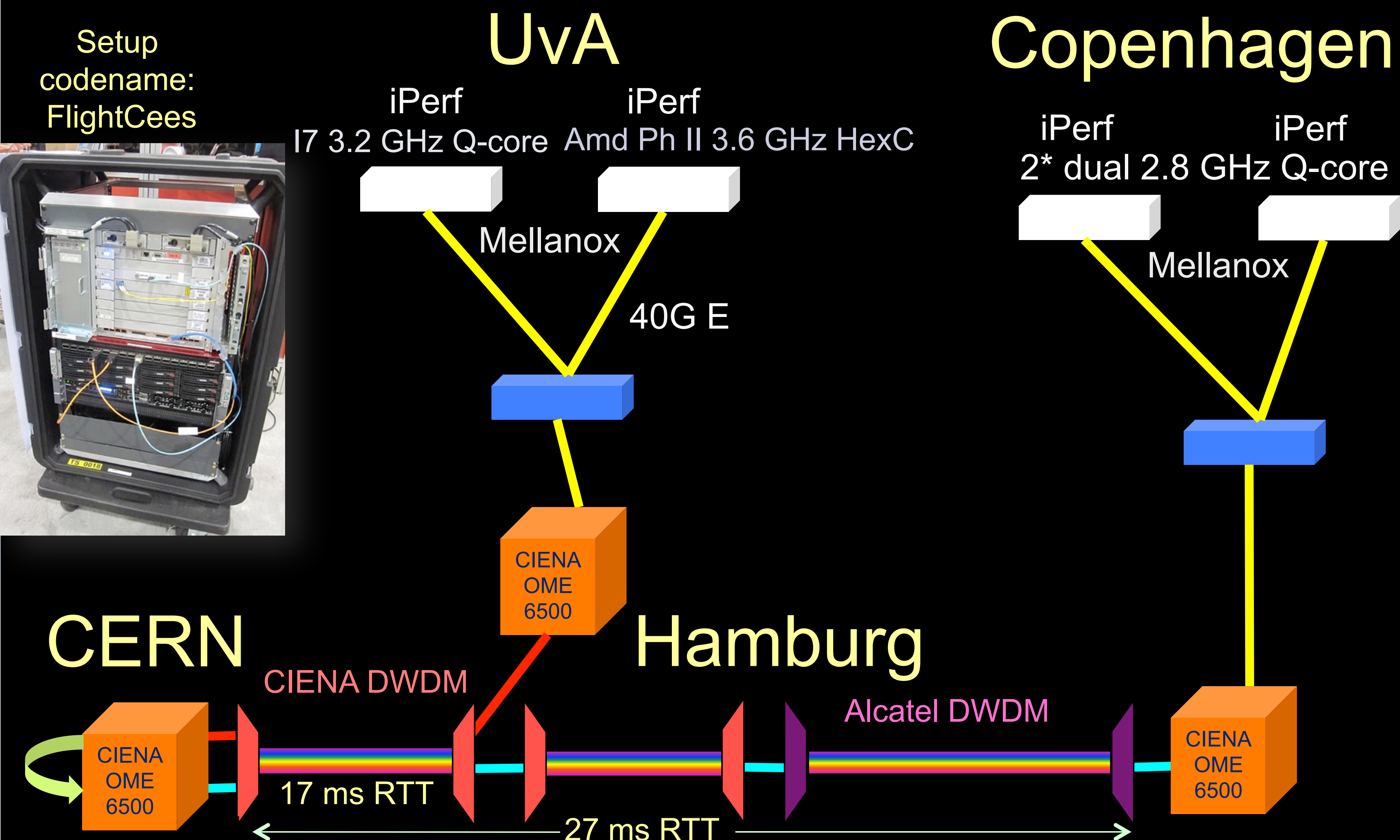
- We have investigated experimentally the all-optical transmission of a 40Gb/s PM-QPSK alien wavelength via a concatenated native and third party DWDM system that both were carrying live 10Gb/s wavelengths.
- The end-to-end transmission system consisted of 1056 km of TWRS (TrueWave Reduced Slope) transmission 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 SURF NET

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

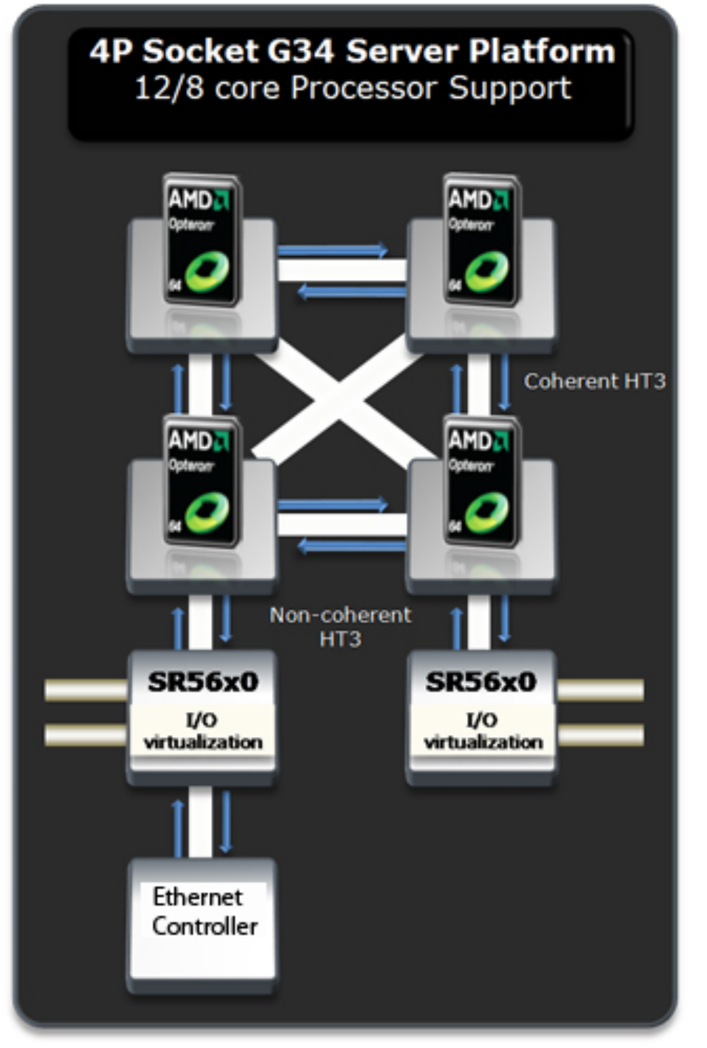


Amsterdam – Geneva (CERN) – Copenhagen – 4400 km (2700 km alien light)

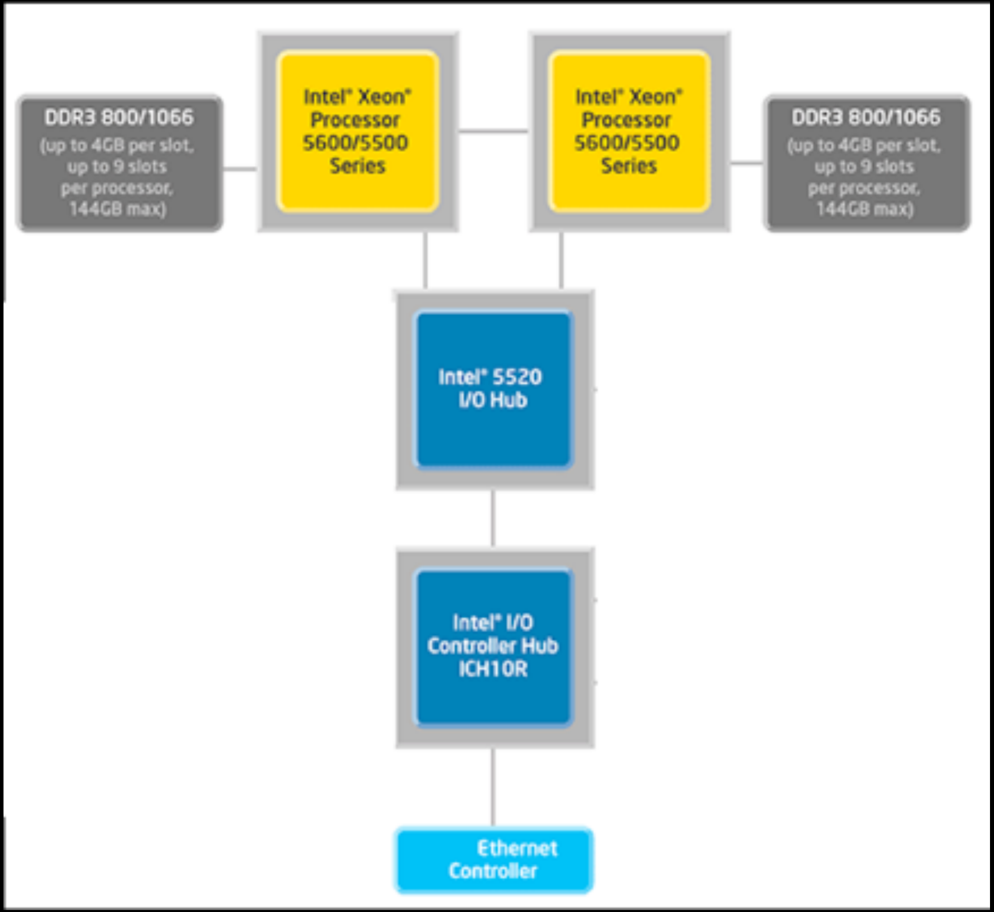
Results (rtt = 17 ms)

- 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

Server Architecture

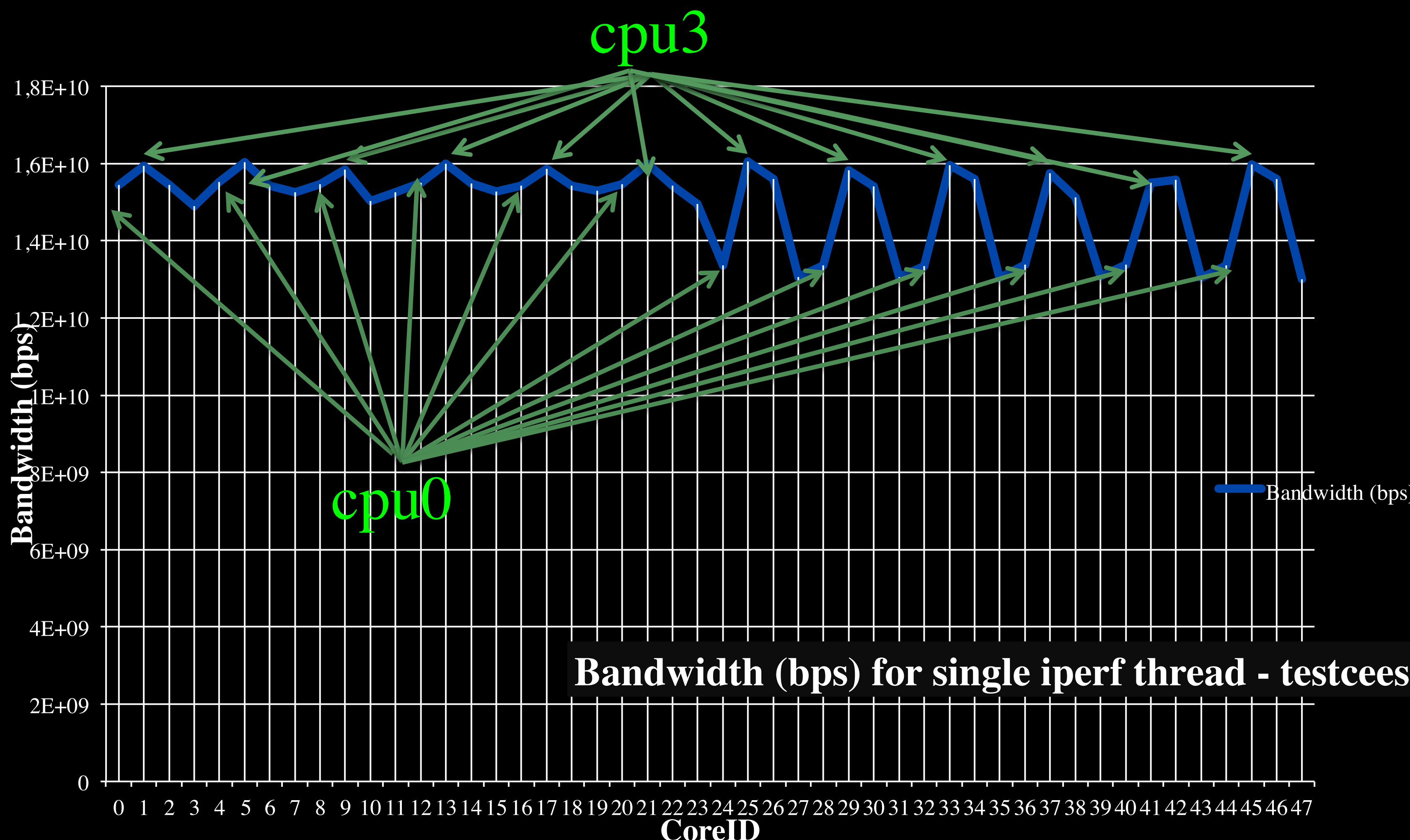


DELL R815
4 x AMD Opteron 6100



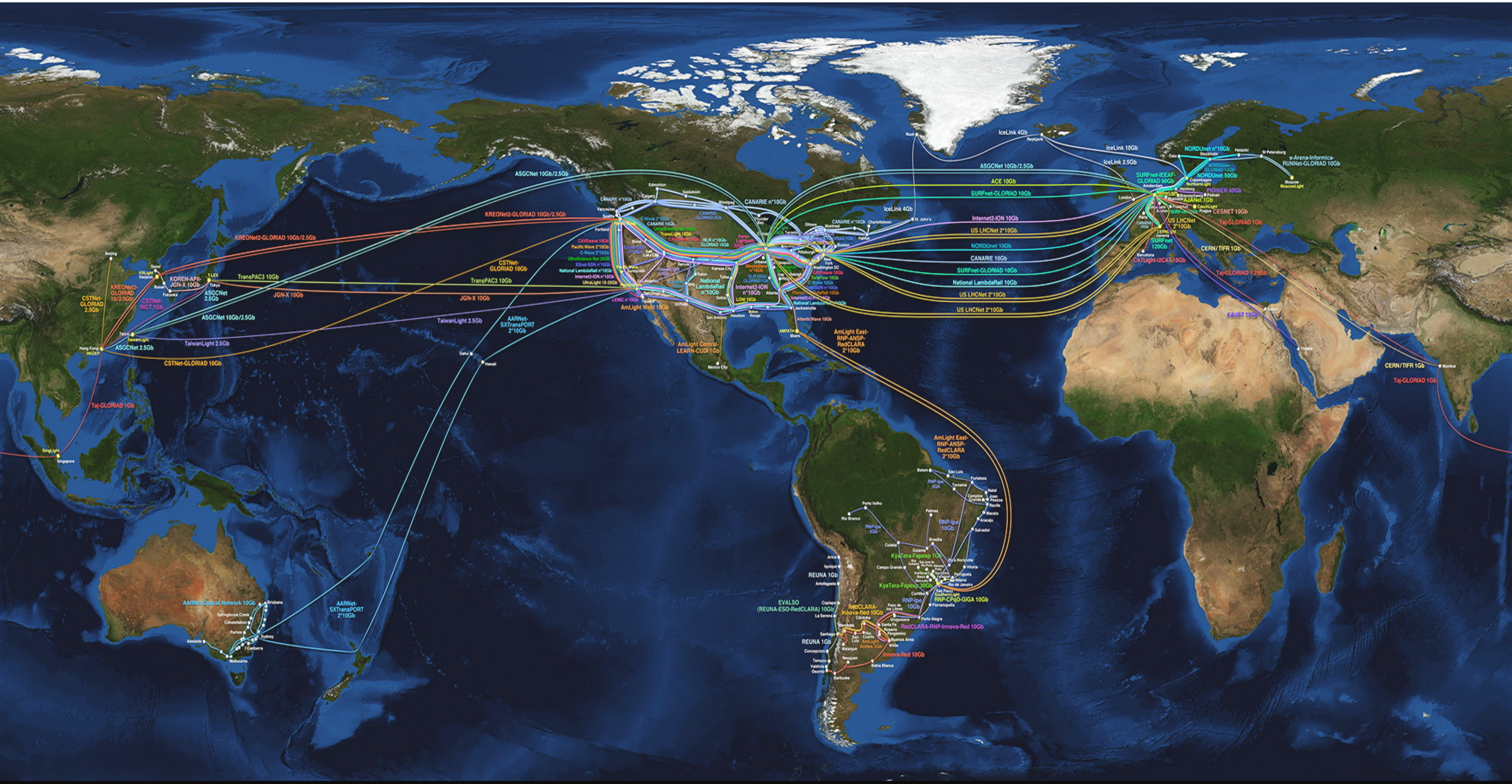
Supermicro X8DTT-HIBQF
2 x Intel Xeon

CPU Topology benchmark



We used numactl to bind iperf to cores

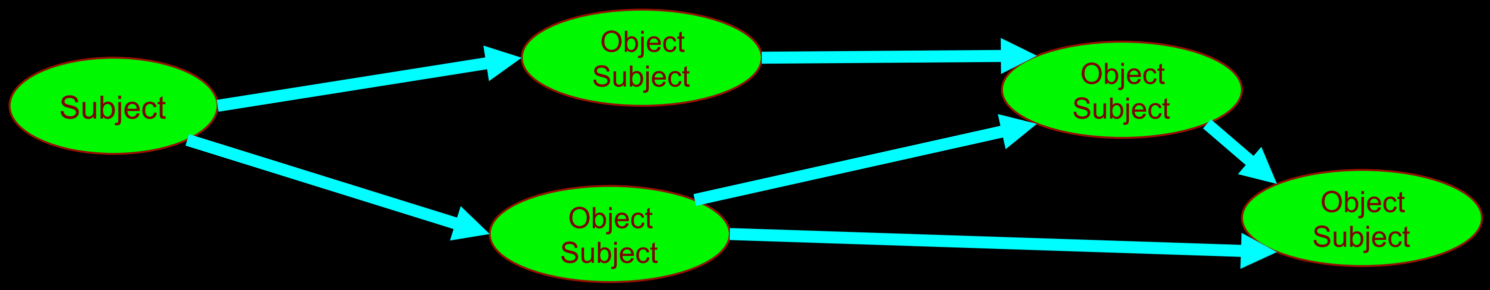
The GLIF – LightPaths around the World

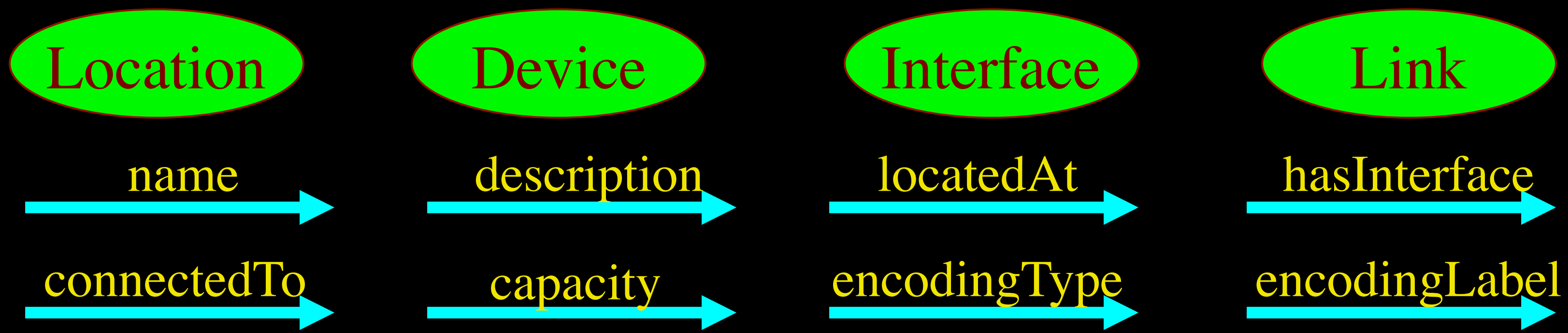


We investigate: TOMTOM for complex networks!

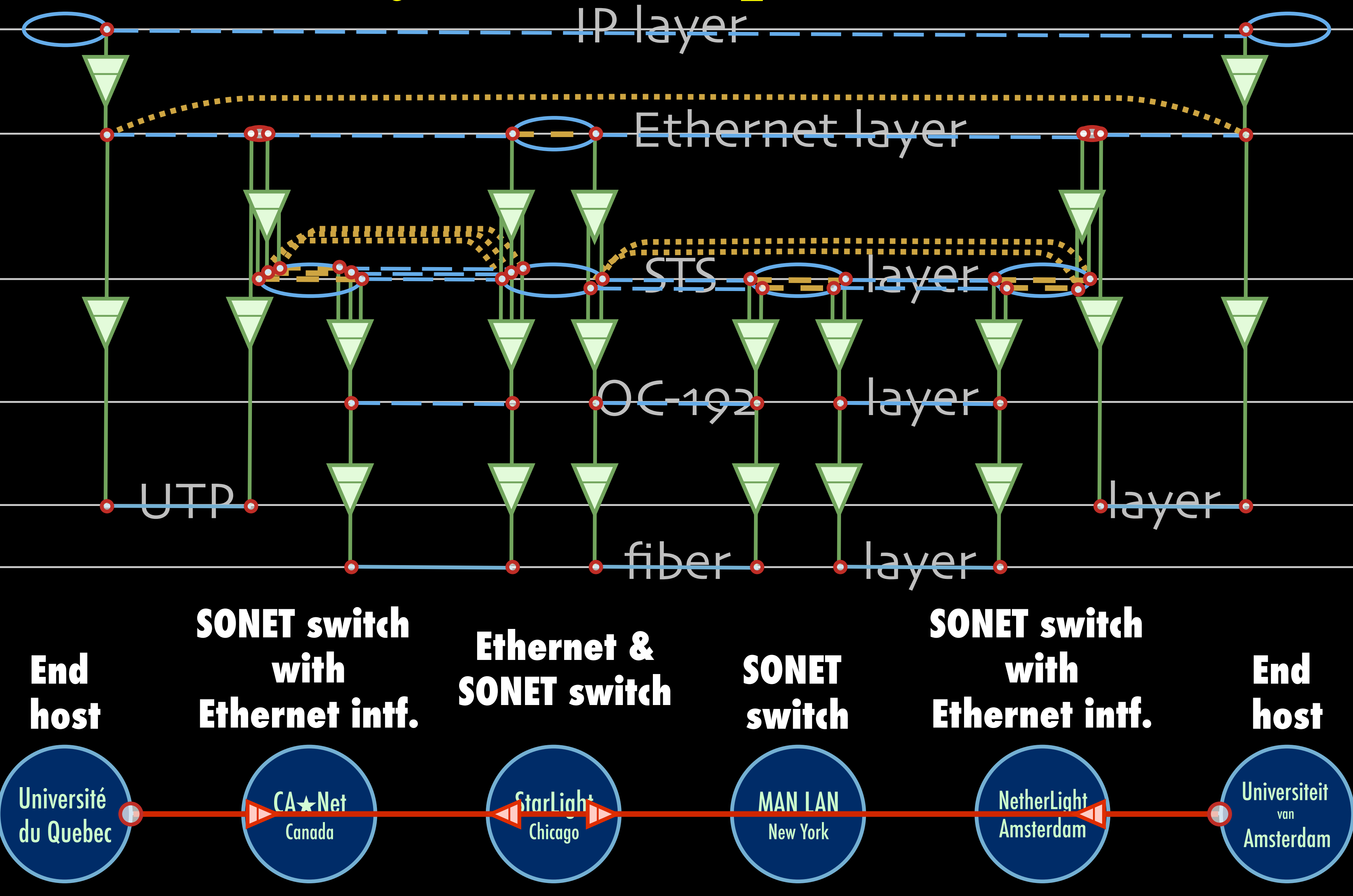
LinkedIN for Infrastructure

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

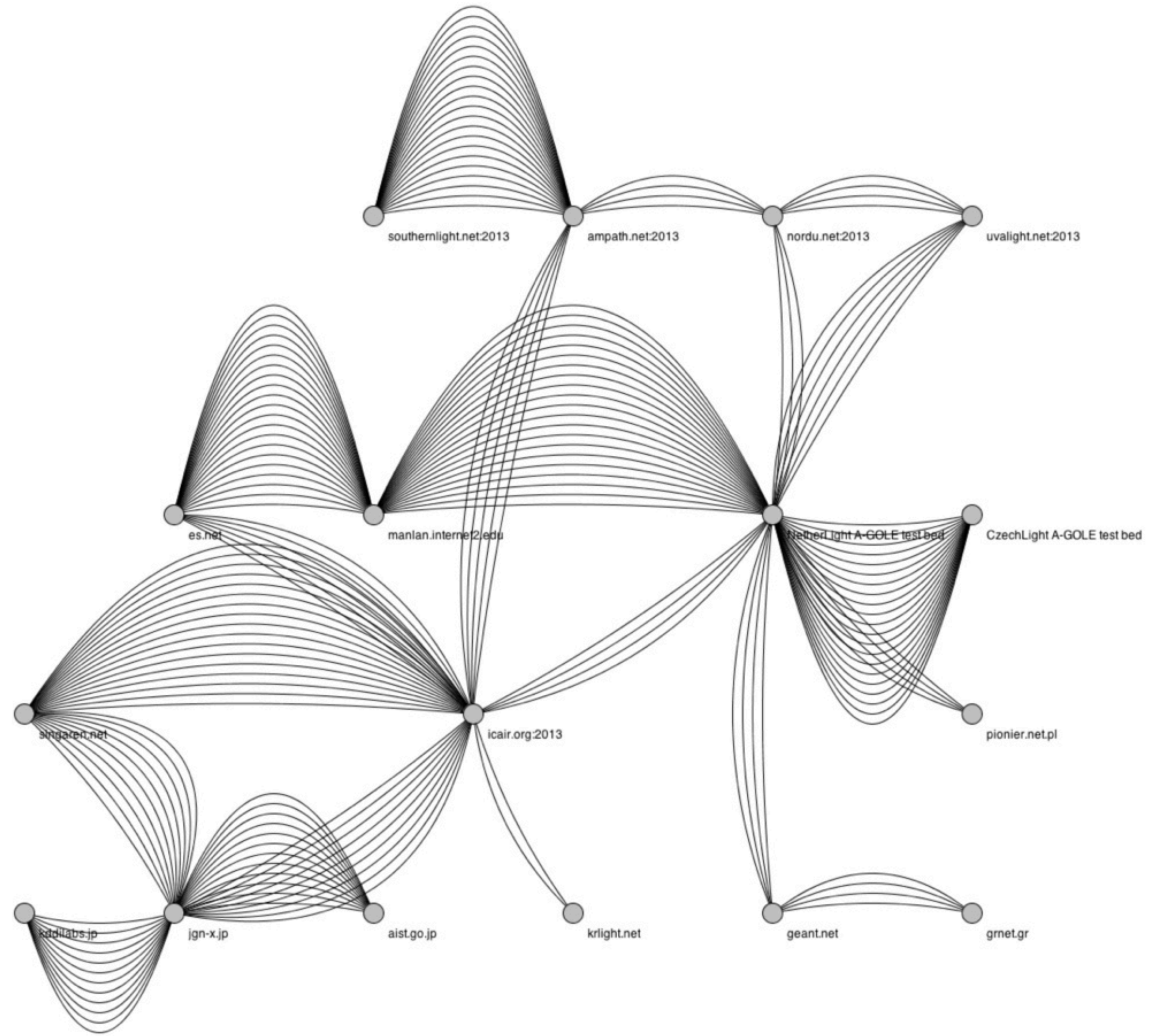




Multi-layer descriptions in NDL

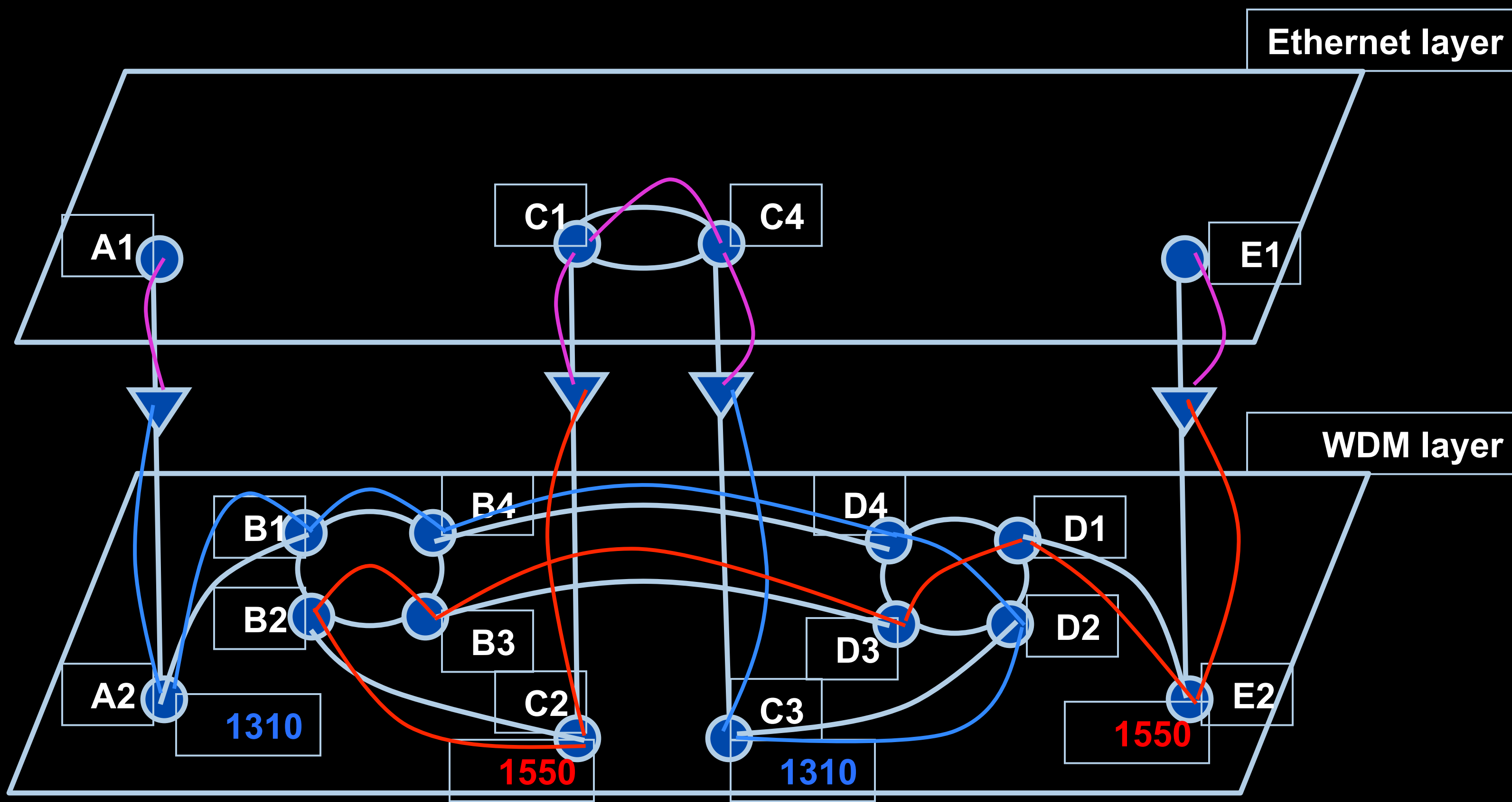


GLIF 2013



southernlight.net:2013
ampath.net:2013
nordu.net:2013
uvalight.net:2013
es.net
manlan.internet2.edu
NetherLight A-GOLE test bed
CzechLight A-GOLE test bed
singaren.net
icair.org:2013
pionier.net.pl
kddilabs.jp
jgn-x.jp
aist.go.jp
krlight.net
geant.net
grnet.gr

Multi-layer Network PathFinding



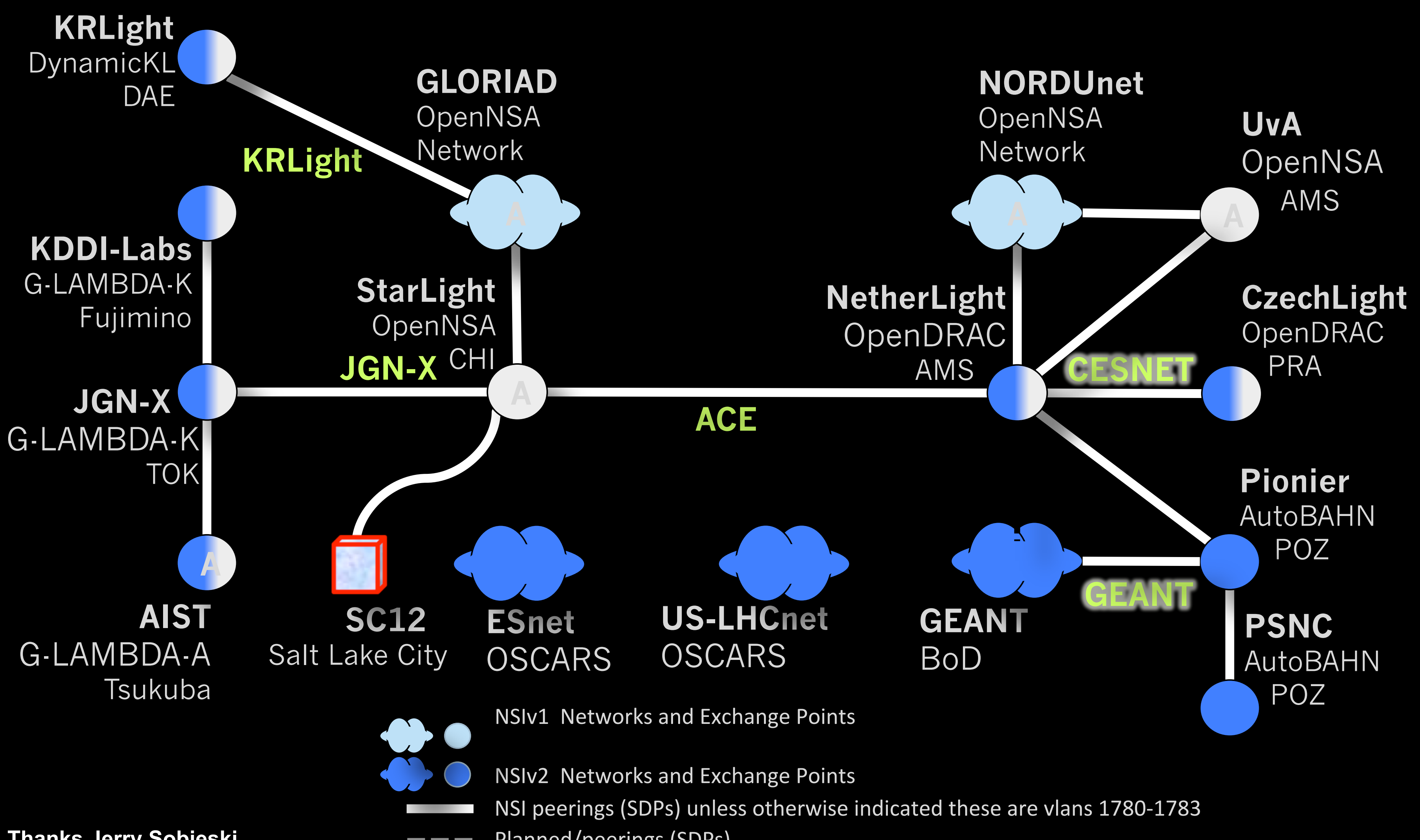
Path between interfaces A1 and E1:
A1-A2-B1-B4-D4-D2-C3-C4-C1-C2-B2-B3-D3-D1-E2-E1

Scaling: Combinatorial problem

Automated GOLE + NSI

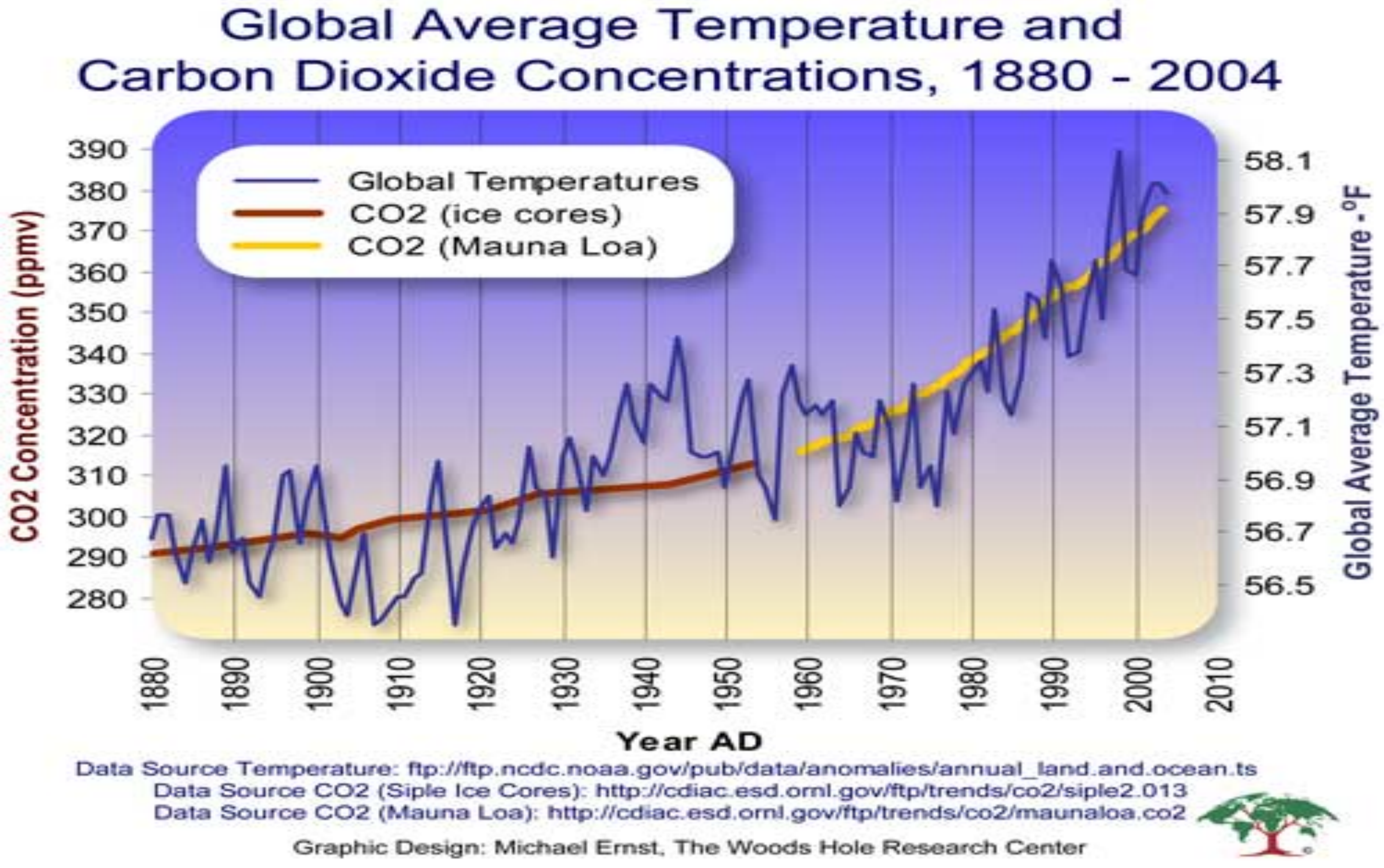
Joint NSI v1+v2 Beta Test Fabric Nov 2012

Ethernet Transport Service

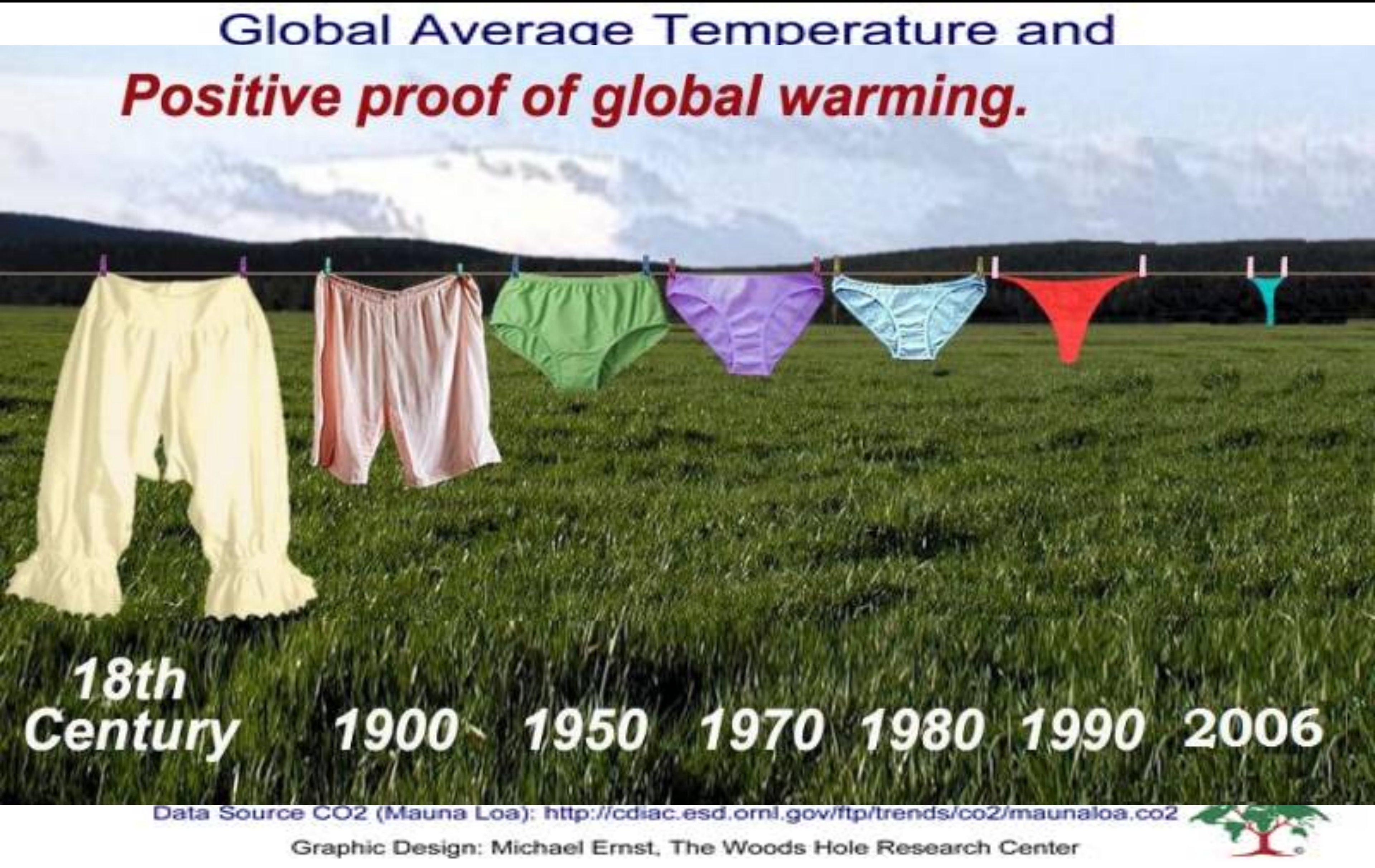


Thanks Jerry Sobieski

Need for GreenIT

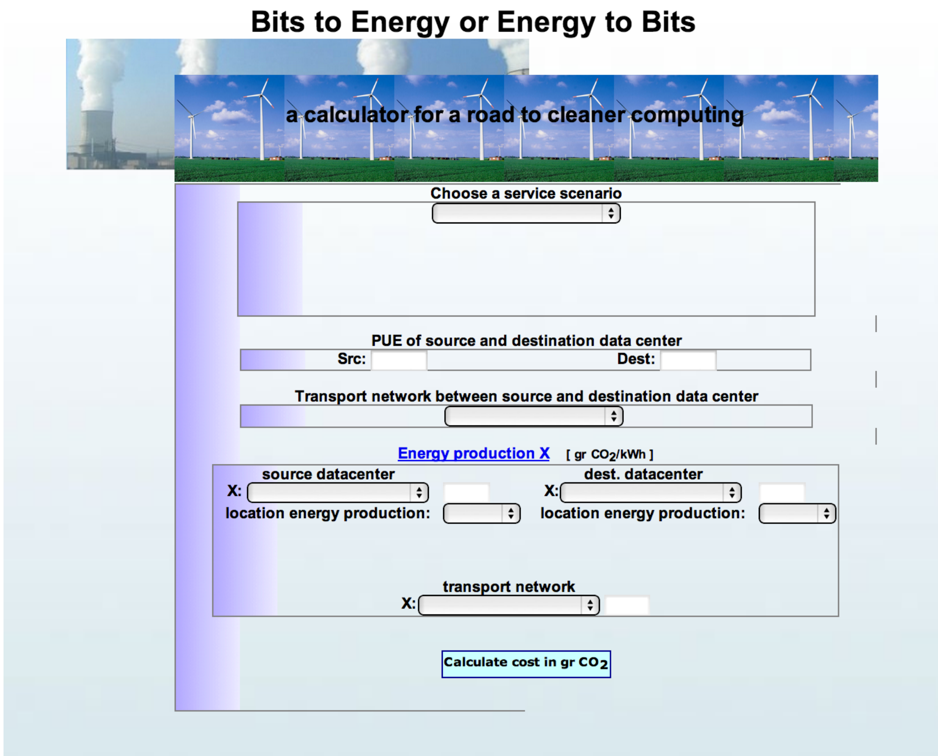


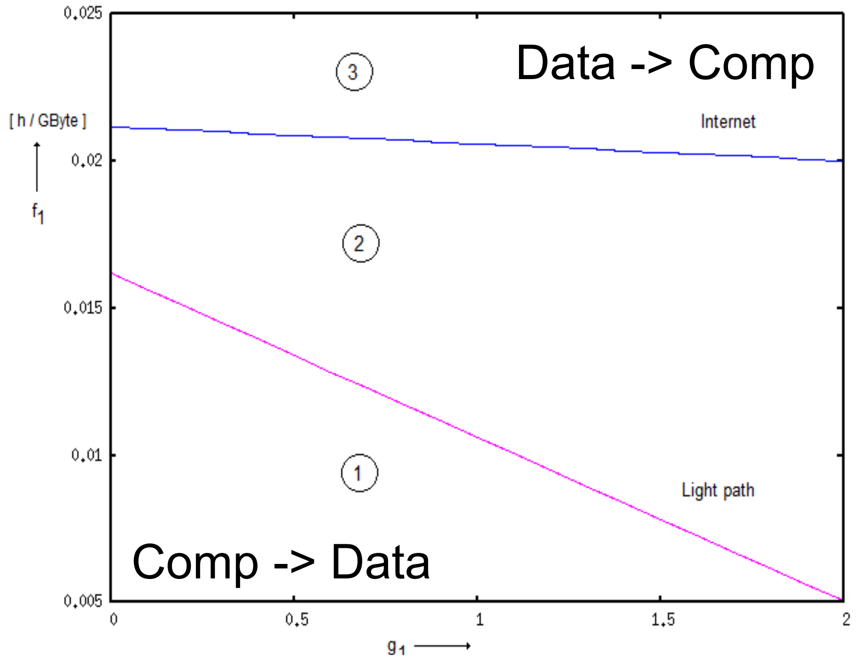
Need for GreenIT

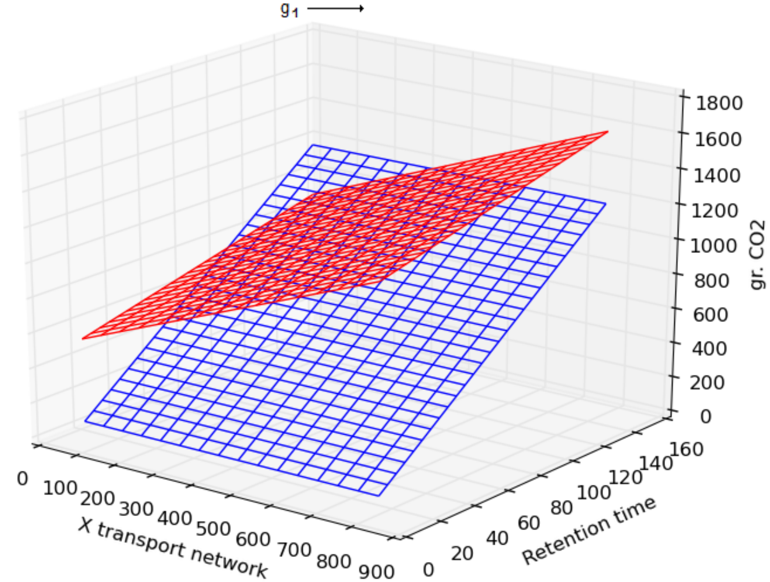


ECO-Scheduling

Bits-Nets-Energy







Storage to energy:

- When should you move hot or cold data to a green remote data centra for storage?
- Given different network paths what are the decision boundaries as function of the task complexity.

http://sne.science.uva.nl/bits2energy/ Taal & Grosso

Mission

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*
  - *Anonymity, integrity of data in distributed data processing*
- *Sustainability*
  - *Greening infrastructure, awareness*
- *Resilience*
  - *Systems under attack, failures, disasters*

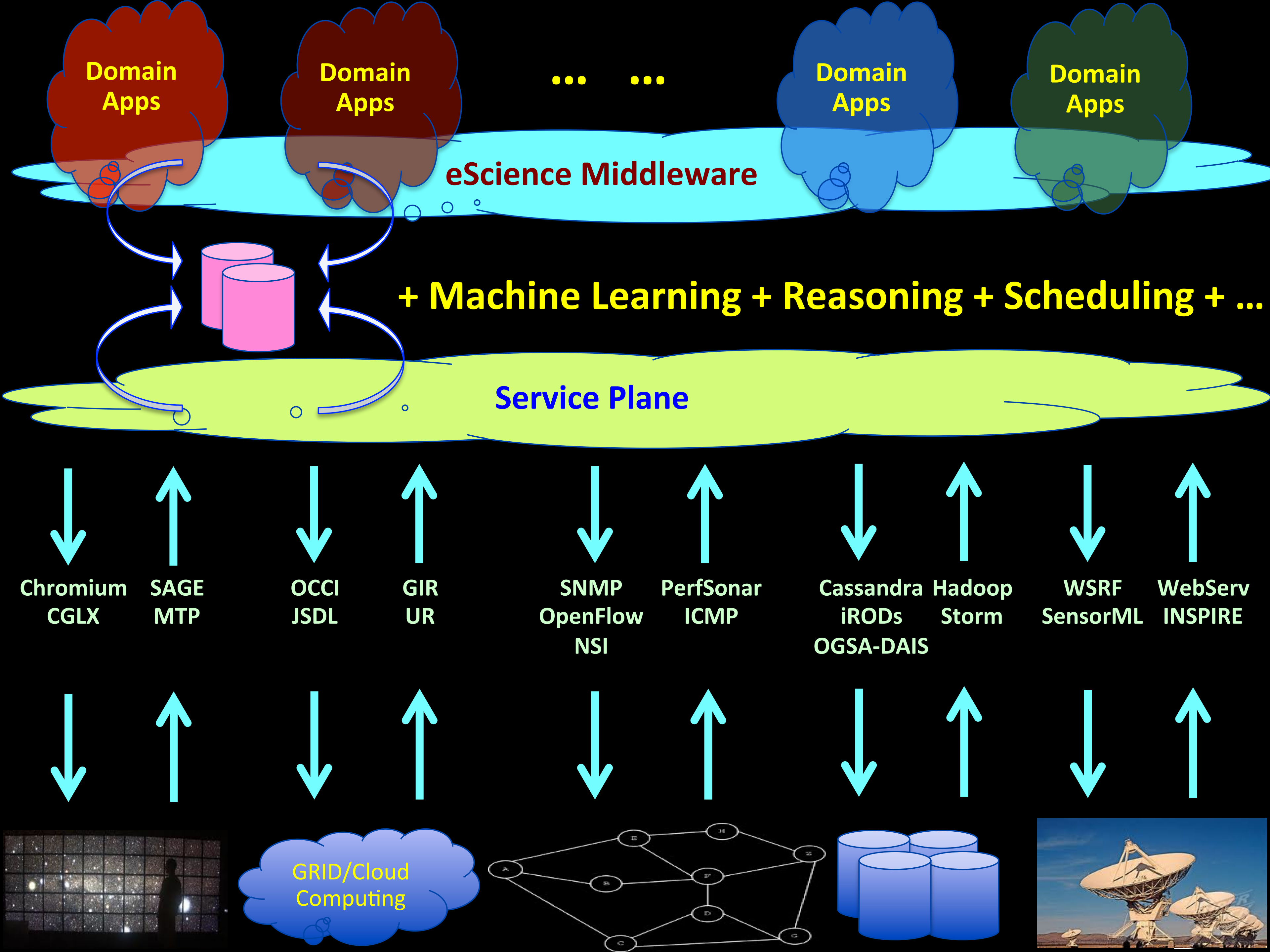
SMART

I want to

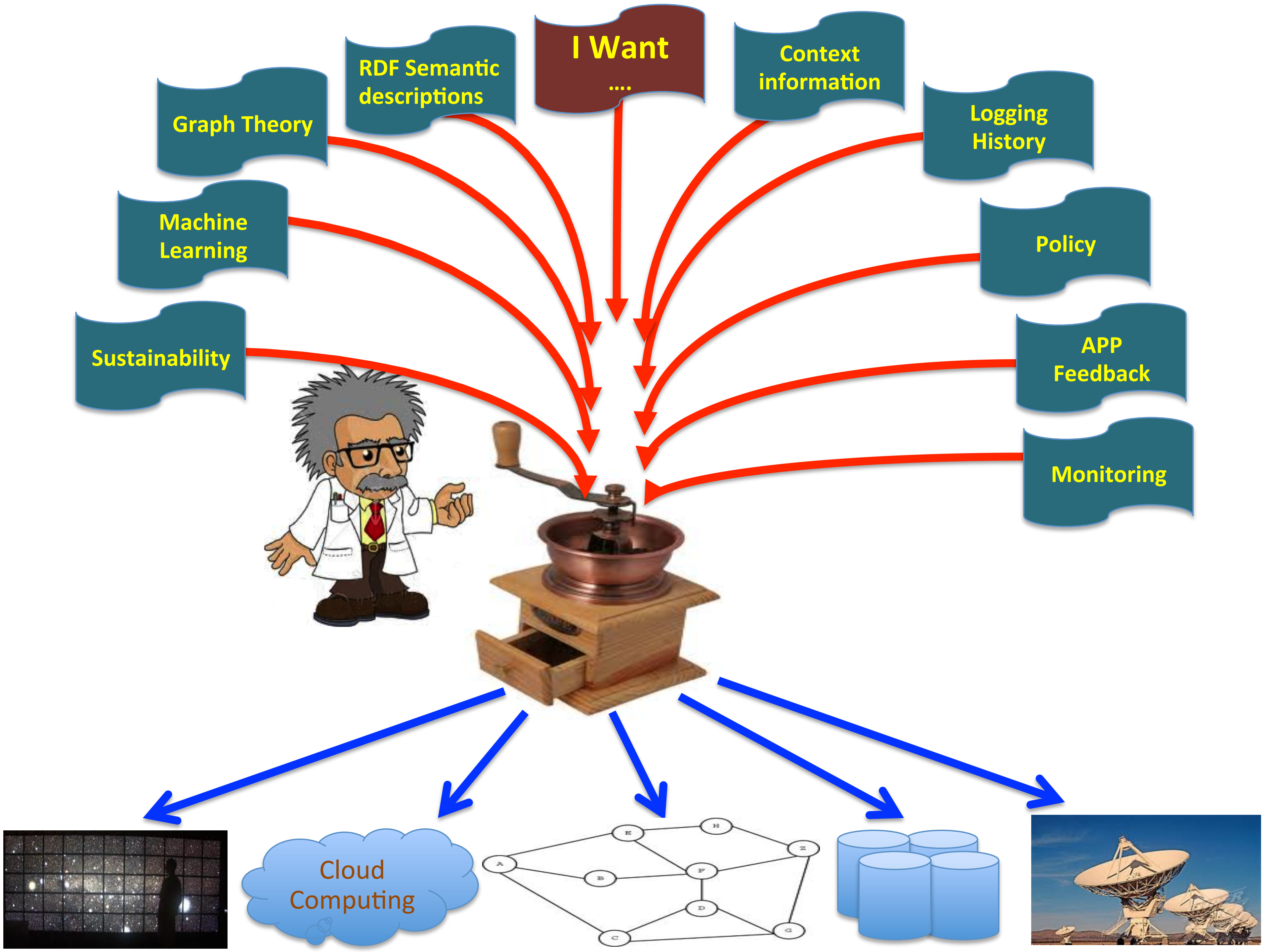
“Show Big Bug Bunny in 4K on my Tiled Display using green Infrastructure”

- 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!

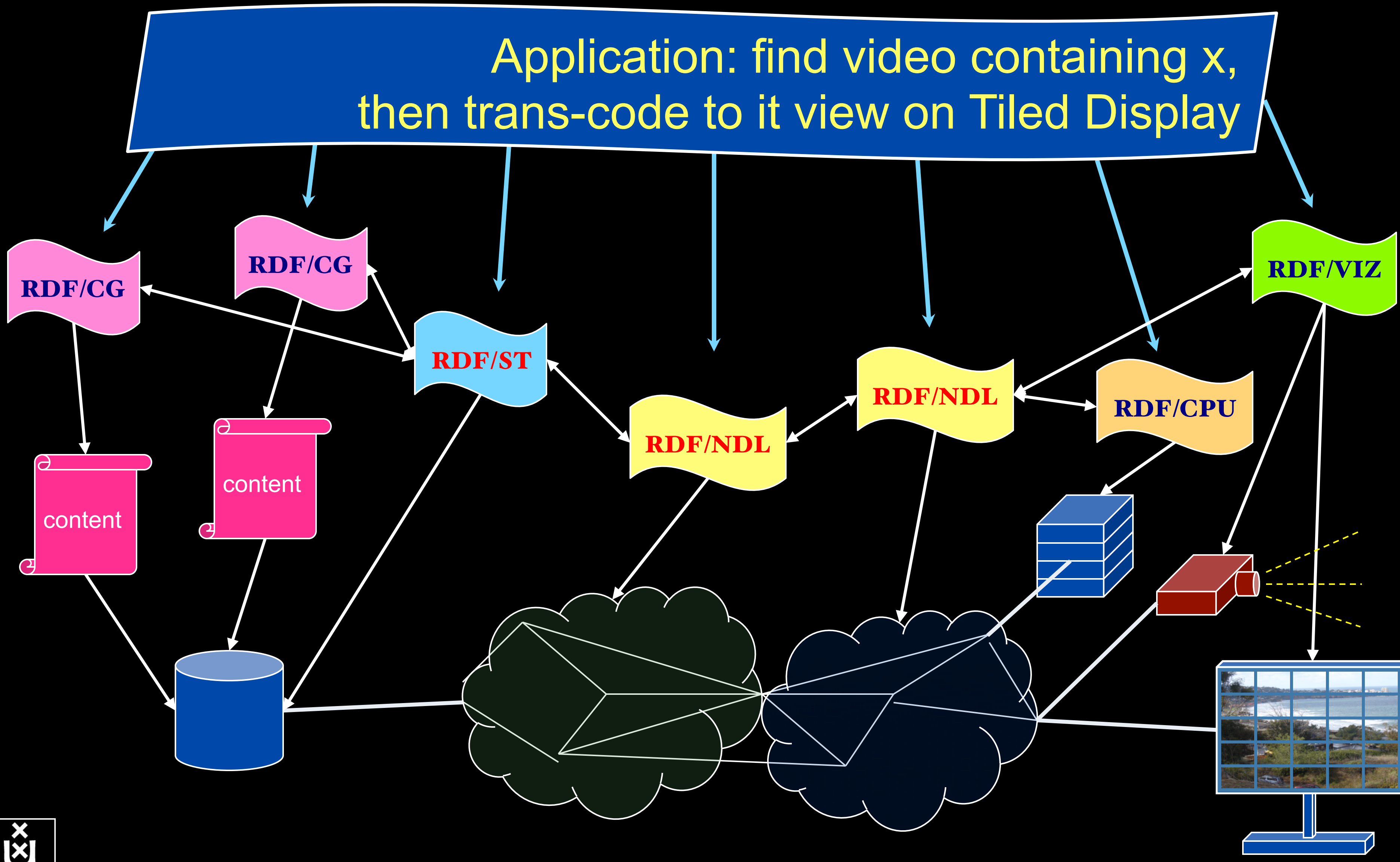


Domain Apps
Domain Apps
... ...
Domain Apps
Domain Apps
eScience Middleware
+ Machine Learning + Reasoning + Scheduling + ...
Service Plane
Chromium CGLX
SAGE MTP
OCCI JSDL
GIR UR
SNMP OpenFlow NSI
PerfSonar ICMP
Cassandra iRODs OGSA-DAIS
Hadoop Storm
WSRF SensorML
WebServ INSPIRE
GRID/Cloud Computing



I Want
....
RDF Semantic descriptions
Context information
Graph Theory
Logging History
Machine Learning
Policy
Sustainability
APP Feedback
Monitoring
Cloud Computing

RDF describing Infrastructure
“I want”



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.