ON*VECTOR

ON*VECTOR Working Group on Terabit Local Area Networks (T-LAN WG)

Thoughts on T-LAN

Cees de Laat
University of Amsterdam

Working Group Meeting

02/10/2010

Hosted by UCSD/Calit2

Sponsored by NTT Network Innovation Laboratory

T-LAN WG 2010 Question #1

- **What if T-LAN were to become widely available?**
  - How would processing, memory and/or storage architectures change?
    - Way more parallel
  - How would user interface environments like SAGE change?
    - For Jason
  - How would application architectures and organizational structures change? For example, scientific visualization, media production or distribution, CSCW, data mining, data preservation, grid computing, cloud computing, etc?
    - Separation of traffic at the source
  - How could T-LAN contribute to total energy savings for Green IT?
    - Not necessarily but it would require the rest to conserve even more

Progress

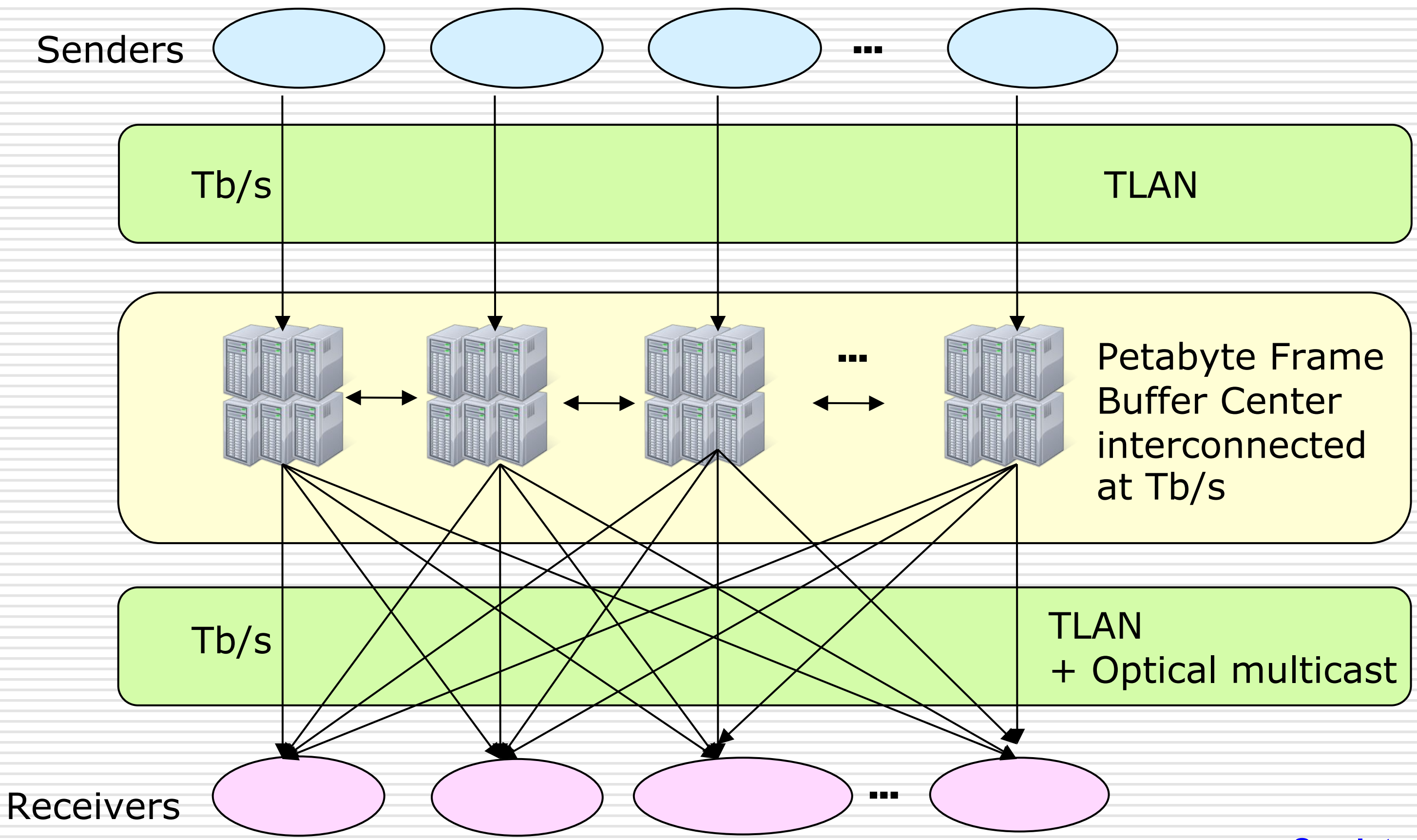
- Kilobit/s ← → keyboard
- Megabit/s ← → process ques / rpc's
- Gigabit/s ← → discs
- Terabit/s ← → GPU

T-LAN WG 2010 Question #2

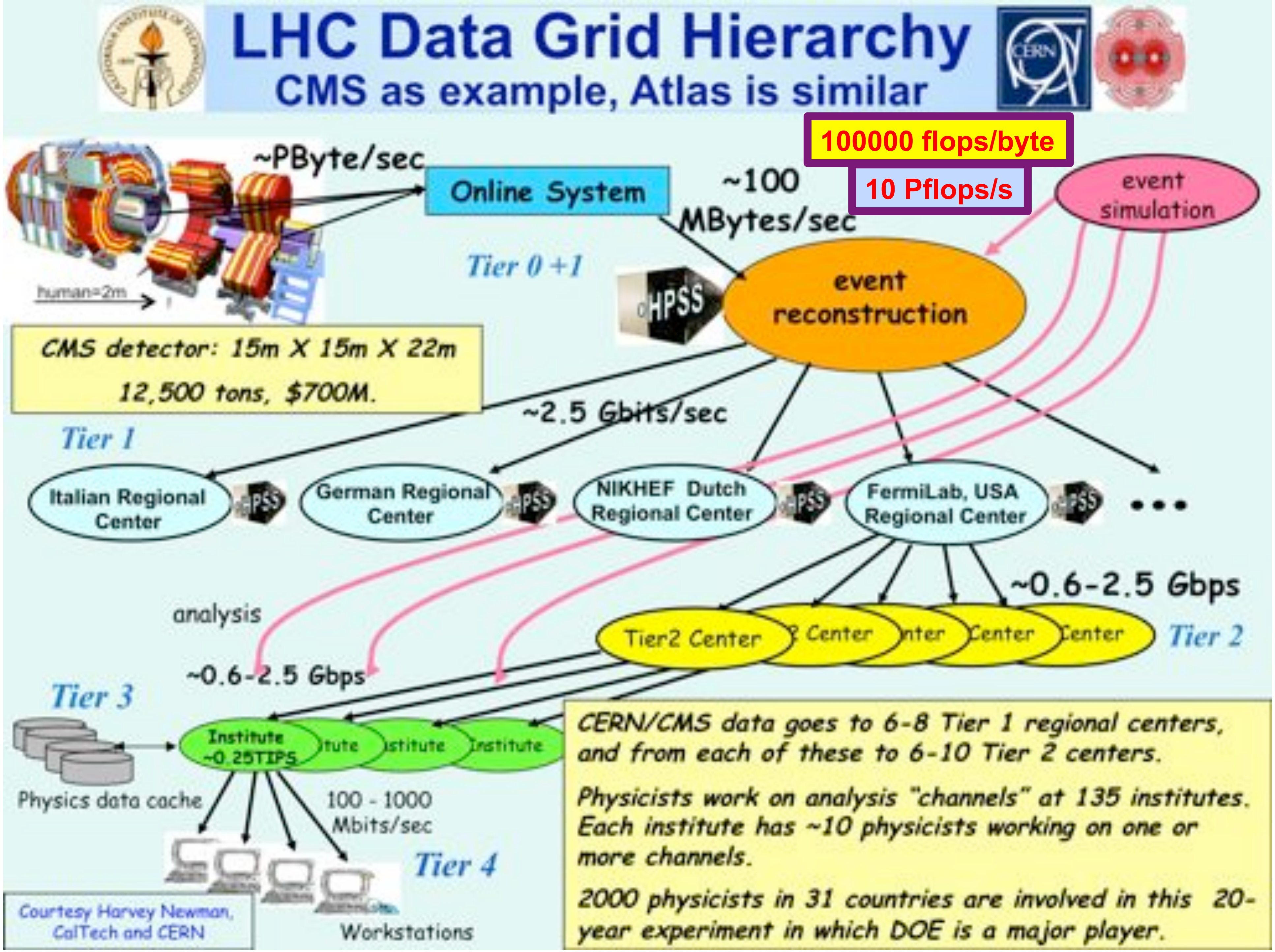
- **What are the technical challenges to terabit scalability?**
  - Energy consumption limit? Where?
    - Depends what it replaces and how much of the traffic it can push to the photonics
  - I/O bandwidth limit? Where?
    - At some point the system becomes very unbalanced
  - Distance limits? Can T-LAN be applied to T-WAN?
    - It will inevitably link with it so the T-WAN should carry the properties of T-LAN
  - Demand limits? Are synchronized high speed links really needed?
    - nope
  - Networking issues? Is point-to-point T-LAN the only option?
    - Nope, scalable hybrid services
  - Network node intelligence? What the best implementation?
  - QoS using burst traffic allowance or network shaping?
    - I would go for shaping if needed because of memory, **QOS at those speeds, forget it.**
  - Control plane limits? Can T-LAN and/or T-WAN networks be managed using current control plane concepts?
    - Possibly NSI, but it is a hard problem

T-LAN WG 2010 Question #3

- **Is this architectural concept a reasonable goal?**

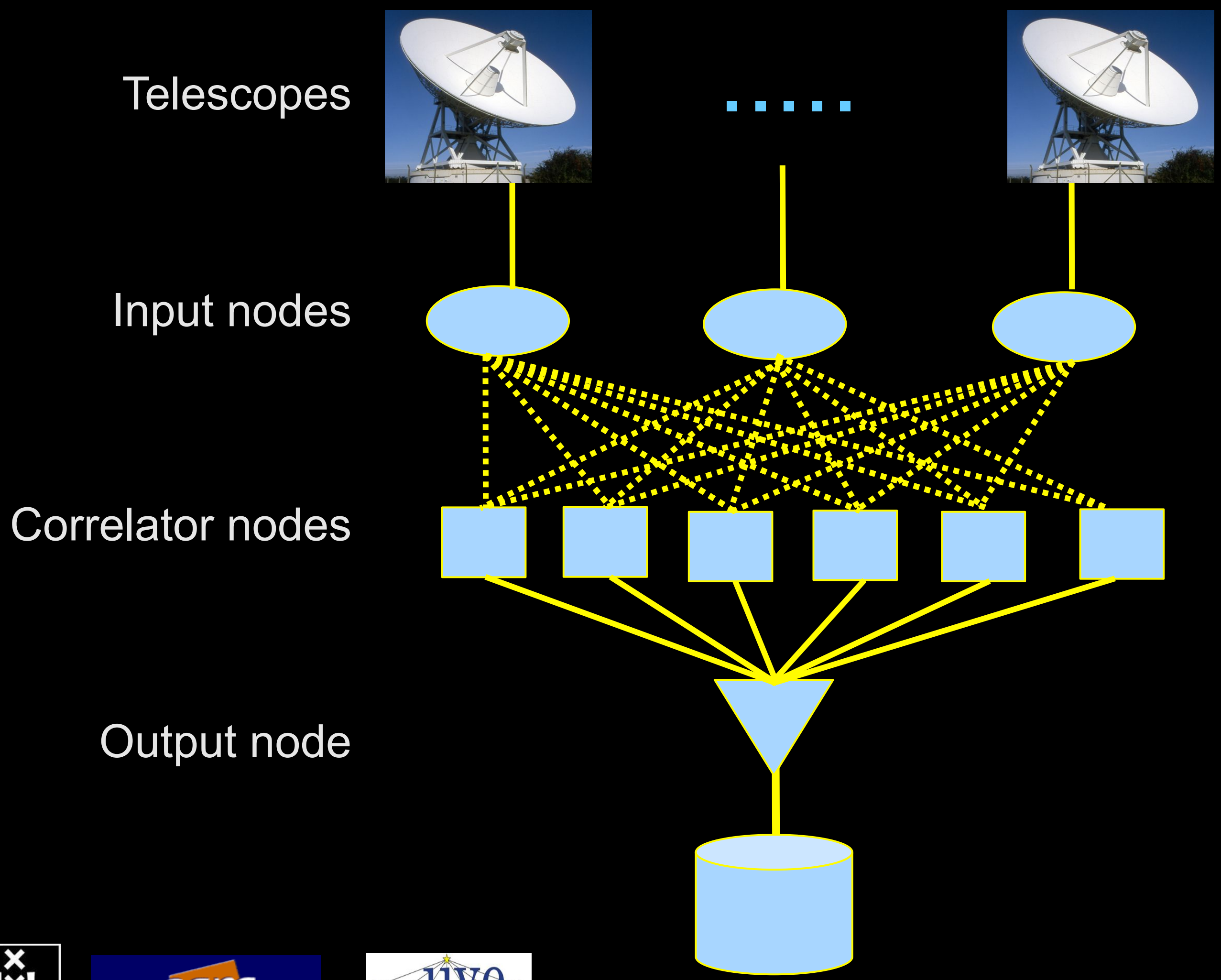


See later


LHC Data Grid Hierarchy
CMS as example, Atlas is similar
100000 flops/byte
10 Pflops/s
~PByte/sec
Online System
~100 MBytes/sec
event simulation
human=2m
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
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

0.1 Pflops/s

THIS IS A DATA FLOW PROBLEM !!!

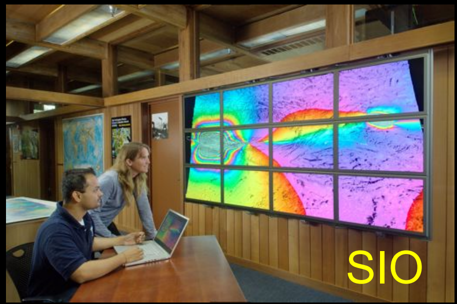
LOFAR as a Sensor Network

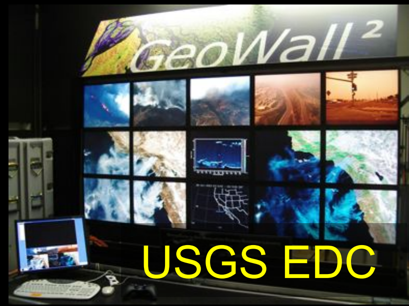
20 flops/byte

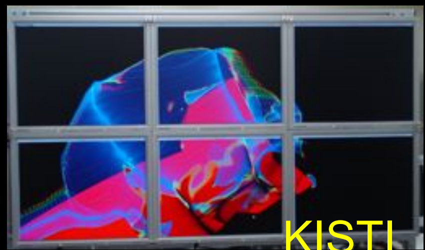
2 Tflops/s

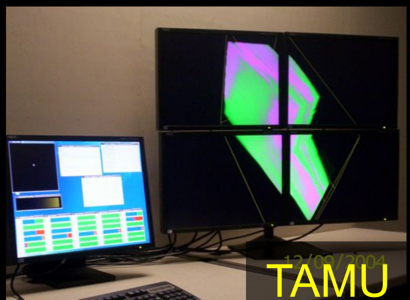
- LOFAR is a large distributed research infrastructure:
  - Astronomy:
    - >100 phased array stations
    - Combined in aperture synthesis array
    - 13,000 small “LF” antennas
    - 13,000 small “HF” tiles
  - Geophysics:
    - 18 vibration sensors per station
    - Infrasound detector per station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

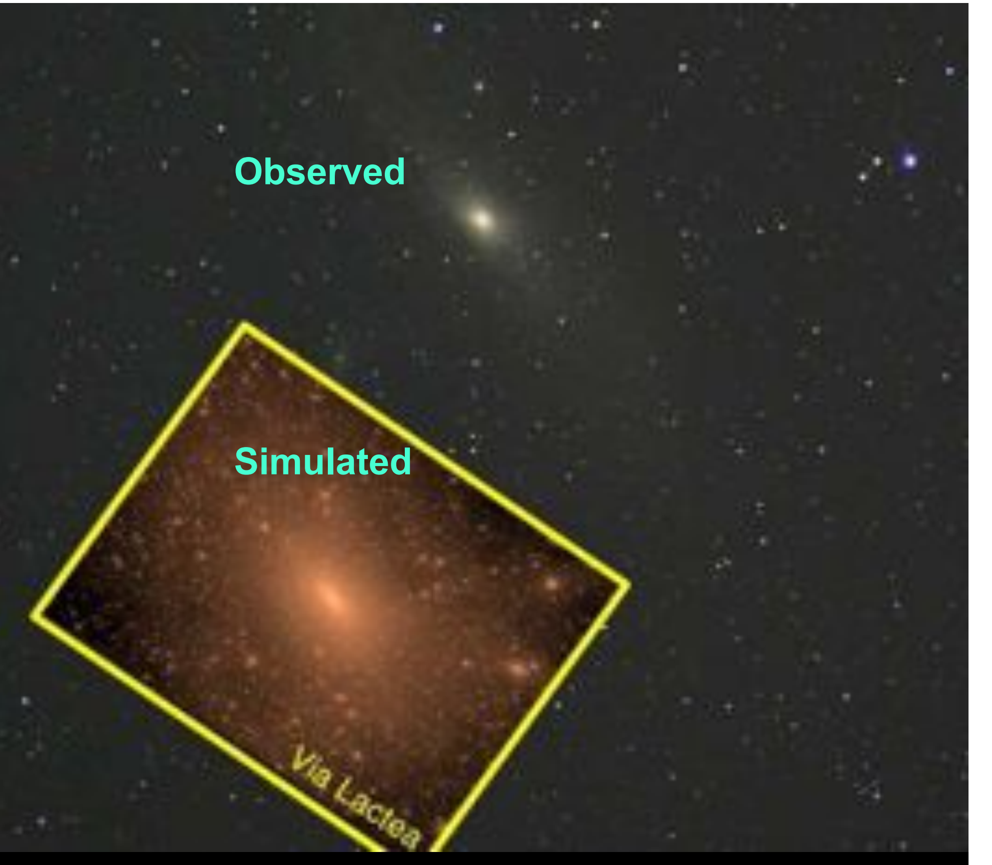
First controlled breach occurred on sept 27th '08:

Many Pflops/s

Many small flows -> 36 Gb/s

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

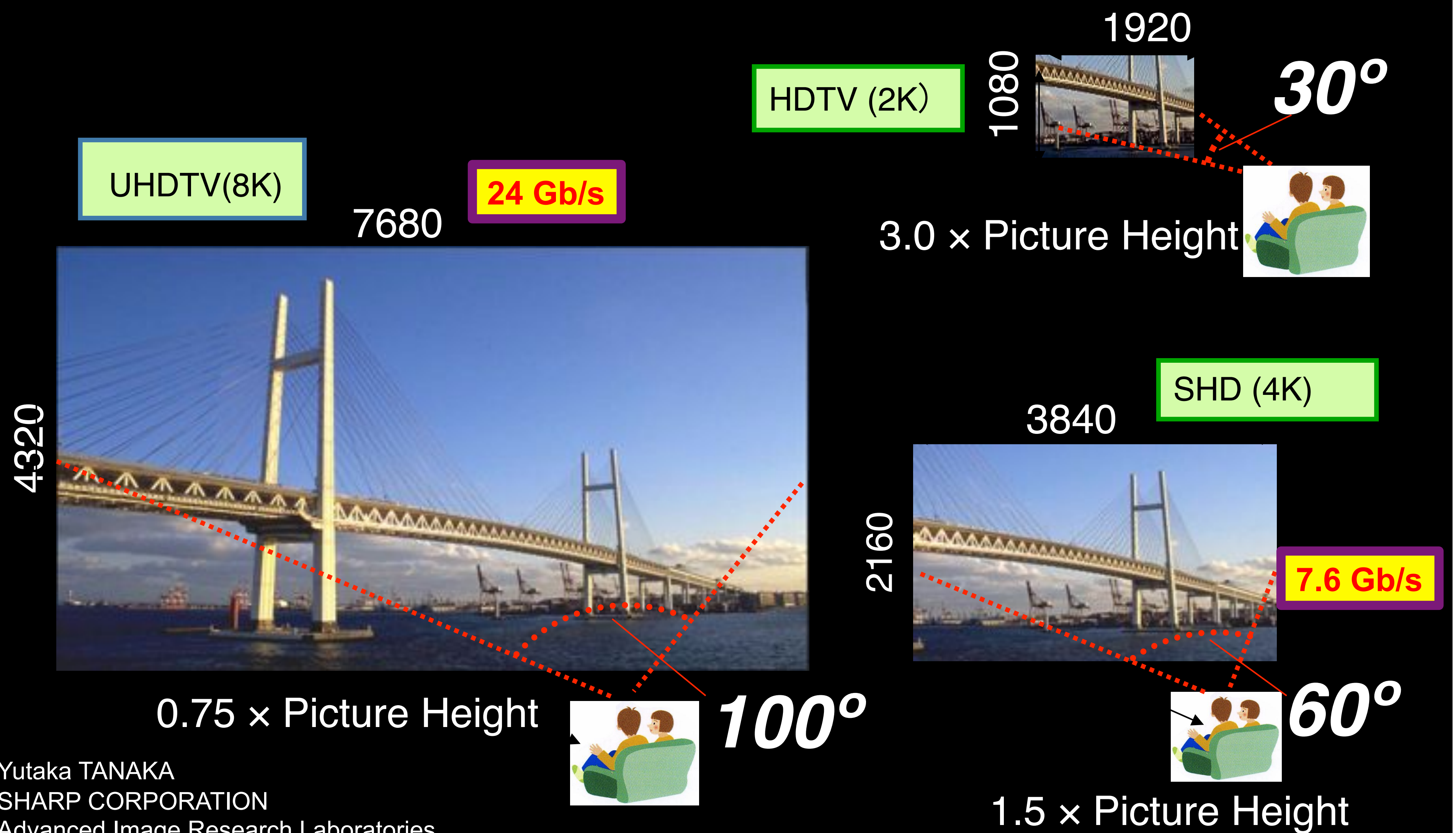
1 Eflops/s

- 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

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



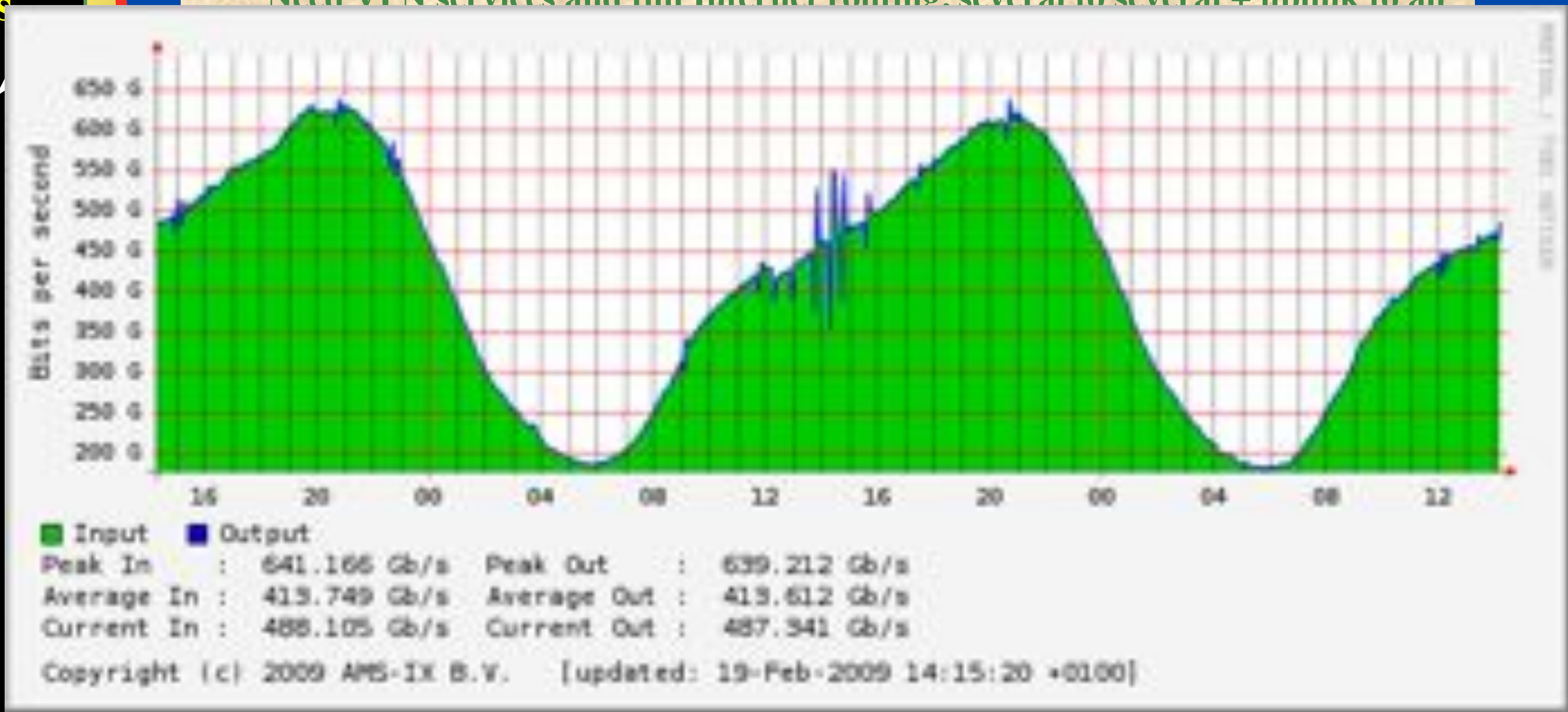
Yutaka TANAKA
SHARP CORPORATION
Advanced Image Research Laboratories

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



users

B

C

ADSL (12 Mbit/s)

GigE

BW

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!

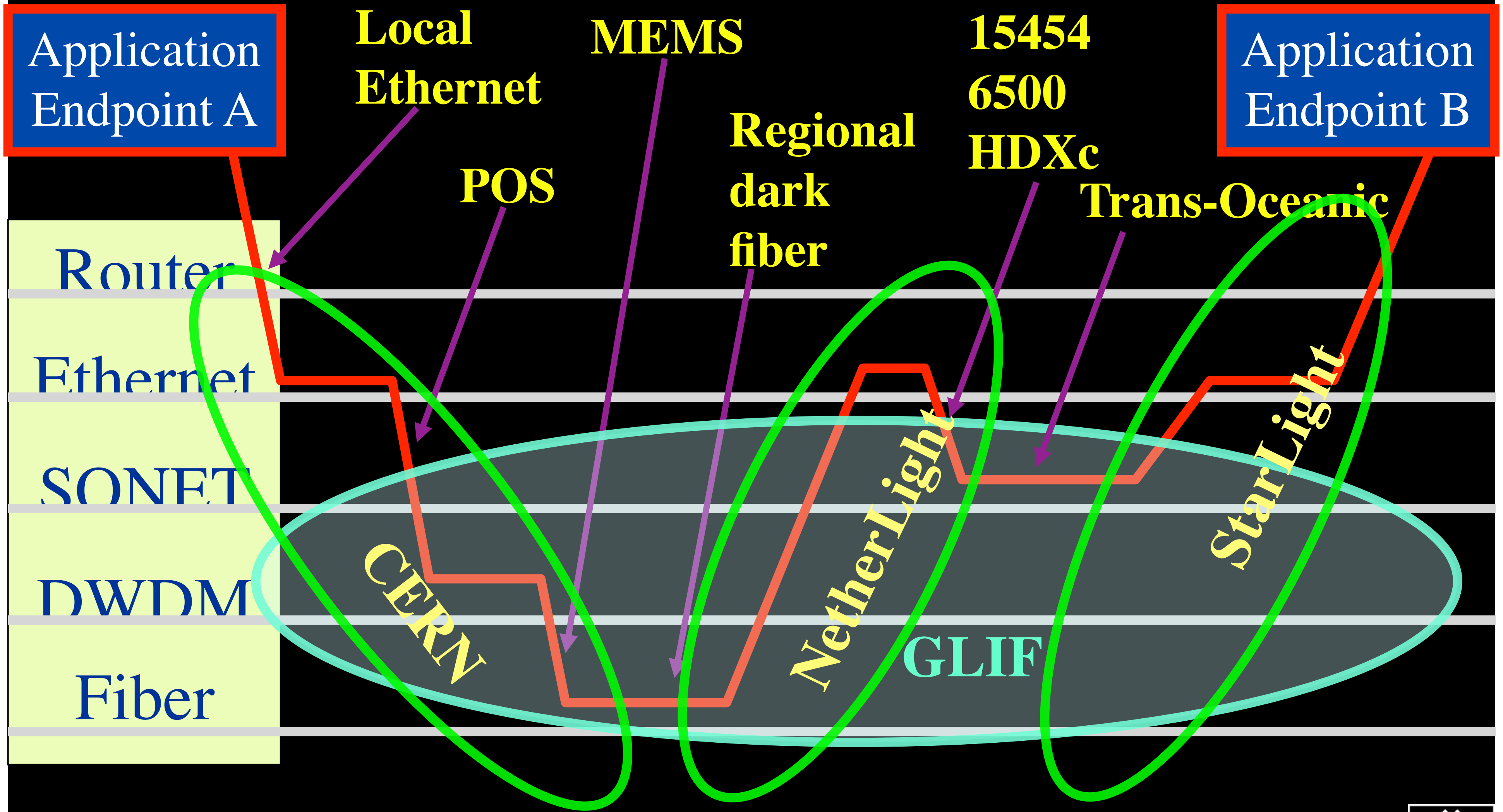
- Costs of photonic equipment 10% of switching 10 % of full routing
  - for same throughput!
  - Photonic vs Optical (optical used for SONET, etc, 10-50 k$/port)
  - DWDM lasers for long reach expensive, 10-50 k$
- Bottom line: look for a hybrid architecture which serves all classes in a cost effective 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

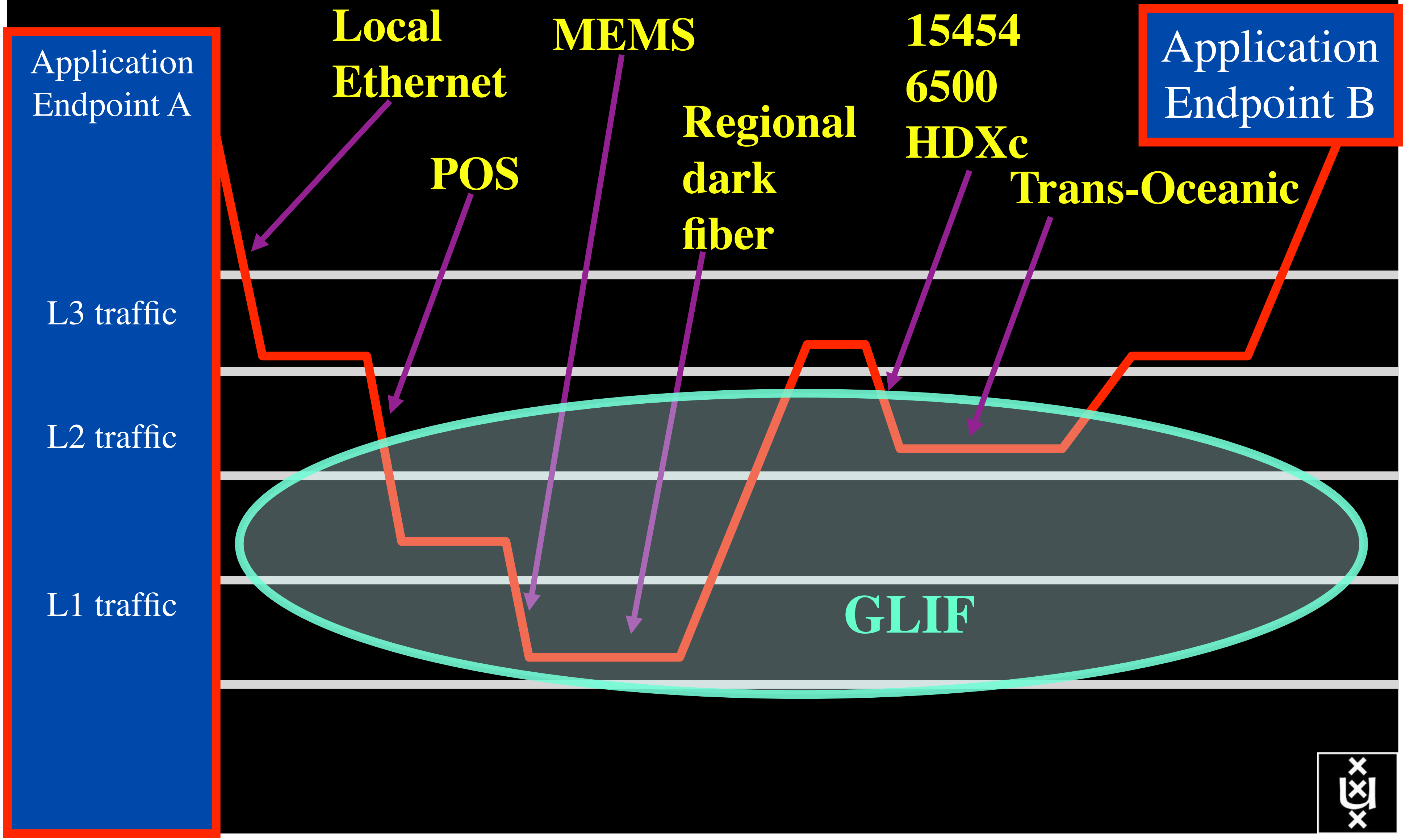
L2 ≈ 5-8 k$/port

L3 ≈ 75+ k$/port

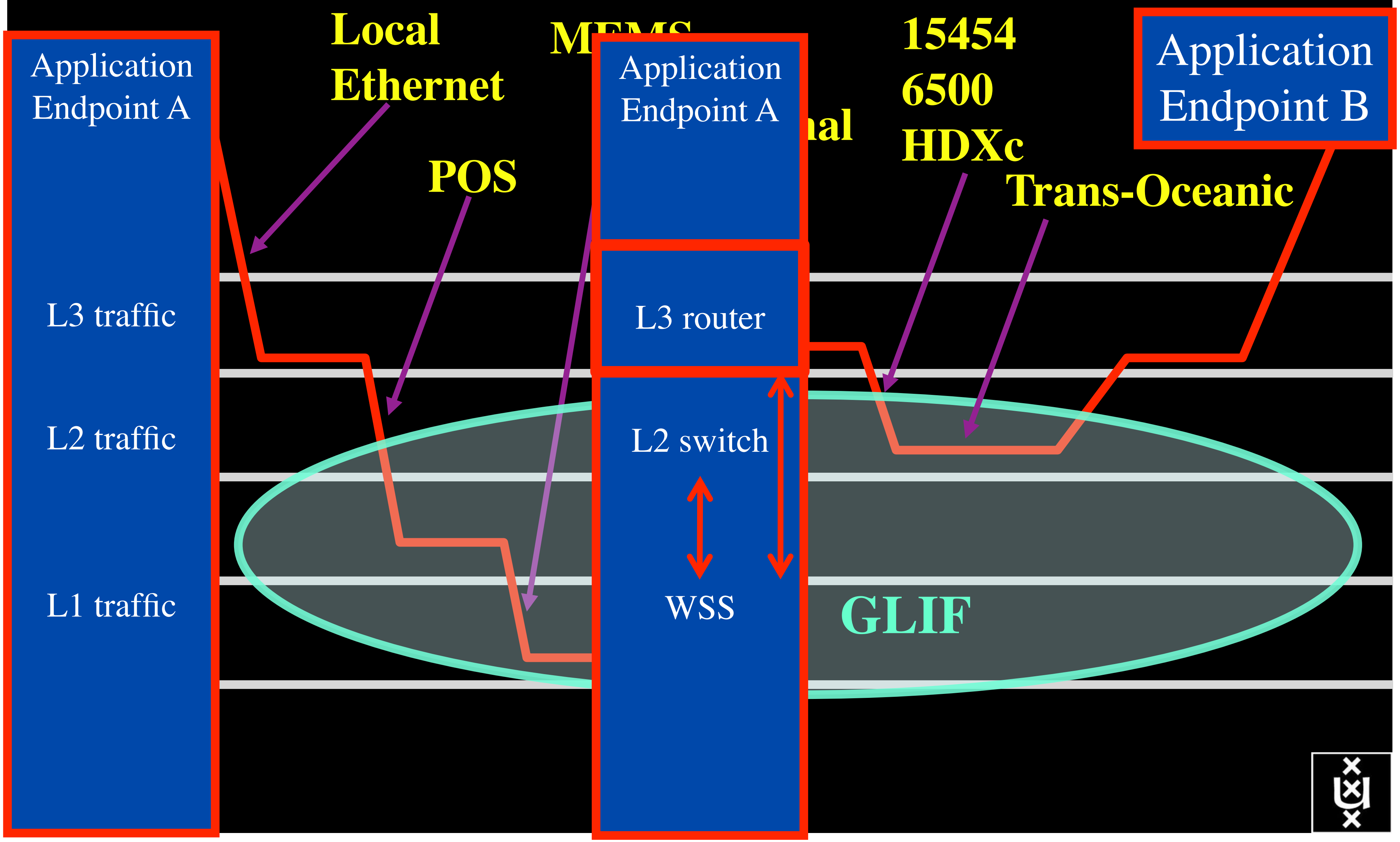
How low can you go?



Architecture



Architecture



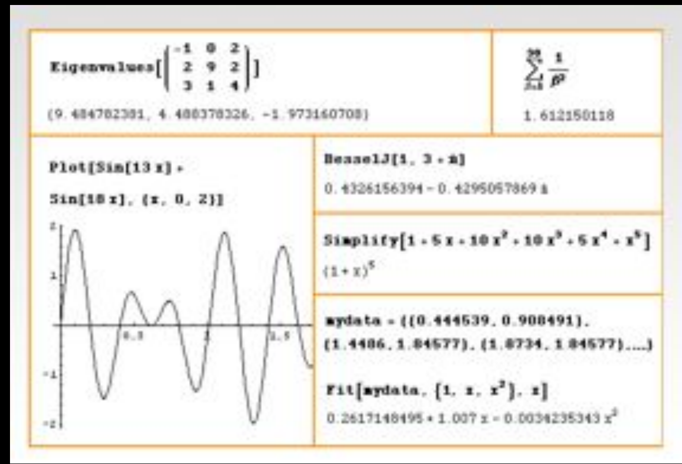
TeraThinking

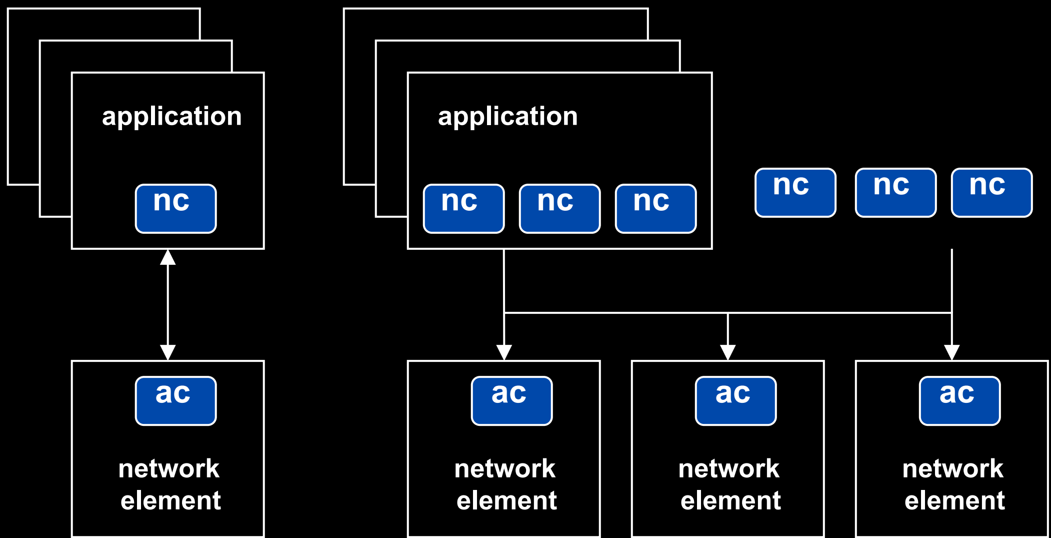
- What constitutes a Tb/s network?
- CALIT2 has 8000 Gigabit drops ?->? Terabit Lan?
- look at 80 core Intel processor
  - cut it in two, left and right communicate 8 TB/s
- think back to teraflop computing!
  - MPI turns a room full of pc's in a teraflop machine
- massive parallel channels in hosts,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



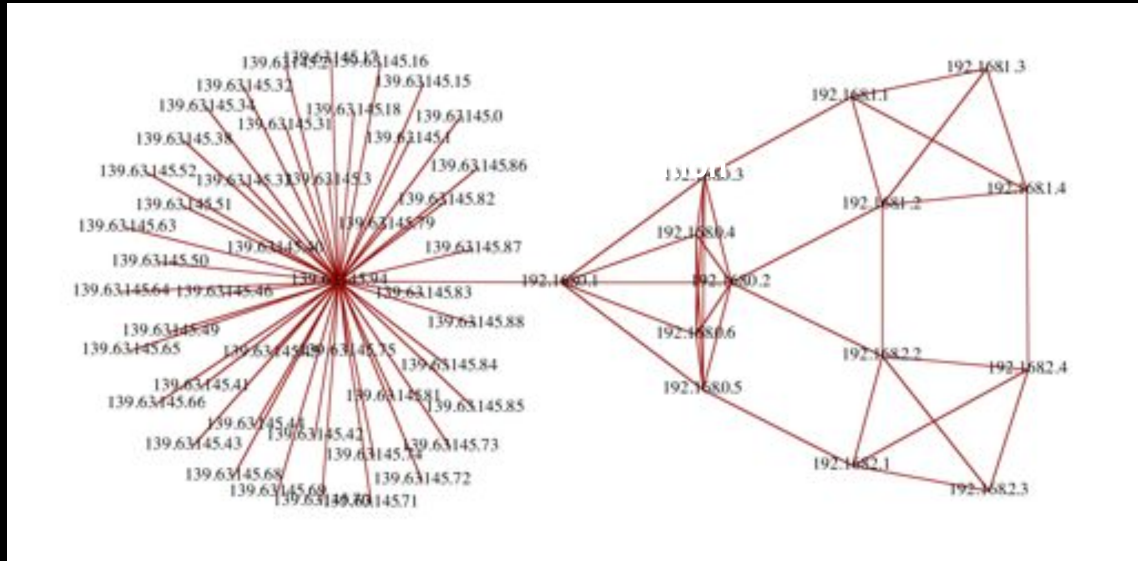


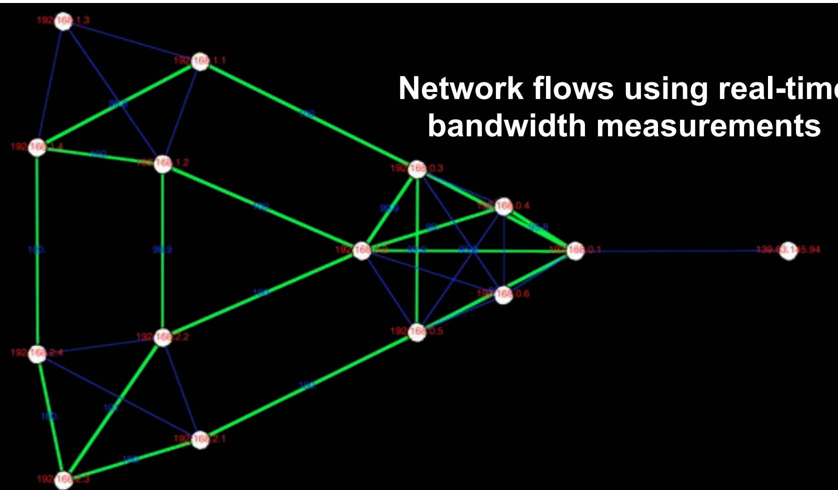
Mathematica enables advanced graph queries, visualizations and real-time network manipulations on UPVNs

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}
```





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.

Questions?