Lambda-Grid developments

Global Lambda Integrated Facility

www.science.uva.nl/~delaat

Cees de Laat

GigaPort

EU

University of Amsterdam

SARA

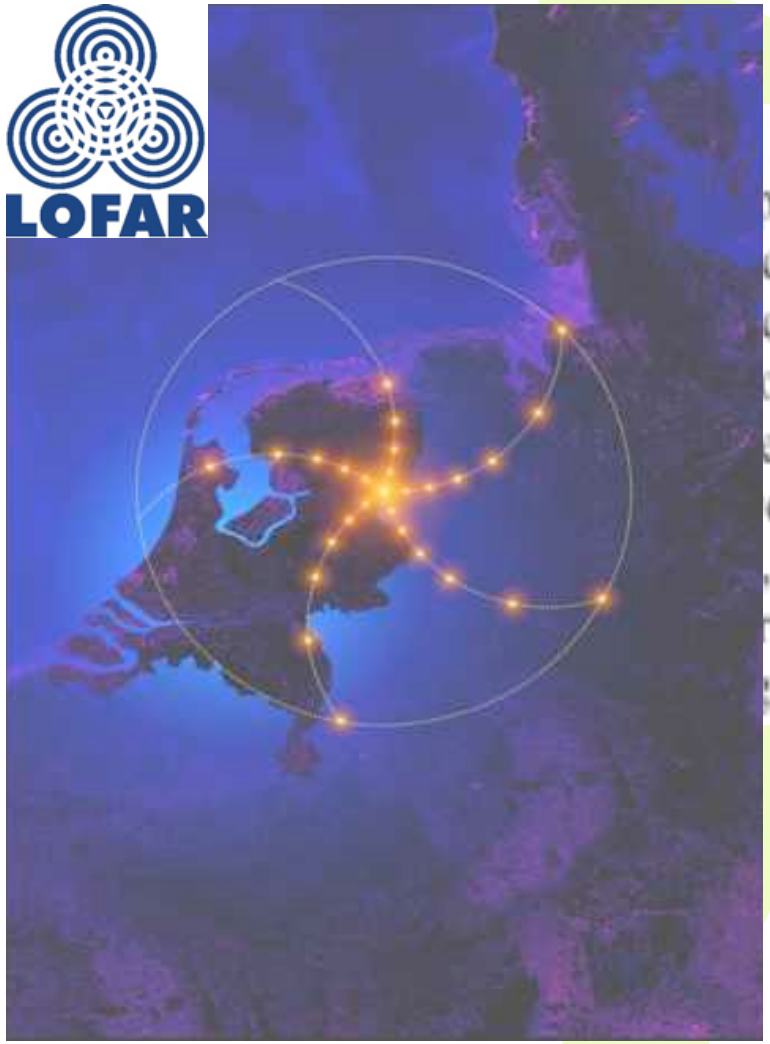
NCF

Contents

This page is intentionally left blank

- Ref: www.this-page-intentionally-left-blank.org

Sensor Grids



~ 40 Tbit/s
www.lofar.org

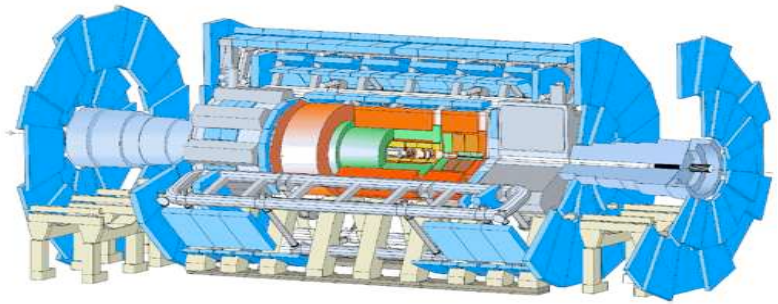
eVLBI

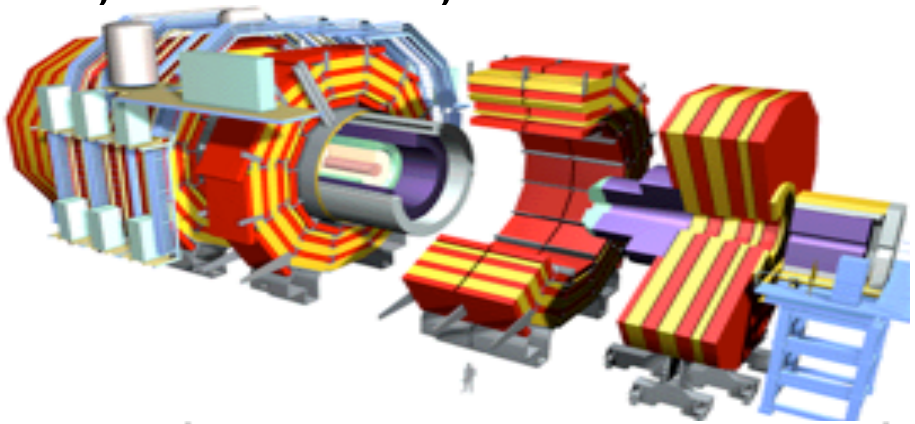
longer term VLBI is easily capable of generatin
e. The sensitivity of the VLBI array scales wit
dth (=data-rate) and there is a strong push to mo
dths. Rates of 8Gb/s or more are entirely feasible.
under development. It is expected that parallel
ed correlator will remain the most efficient approach
olves dist
multi-gig
elator and
factor.

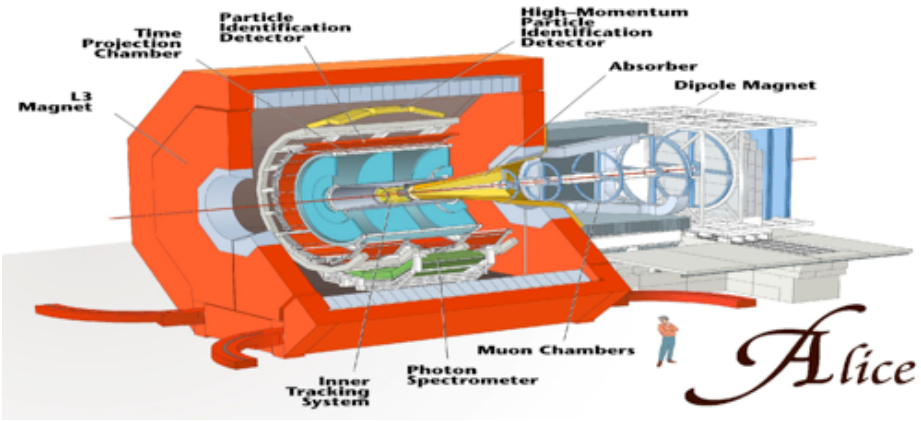
Westerbork Synthesis Radio Telescope - Netherlands

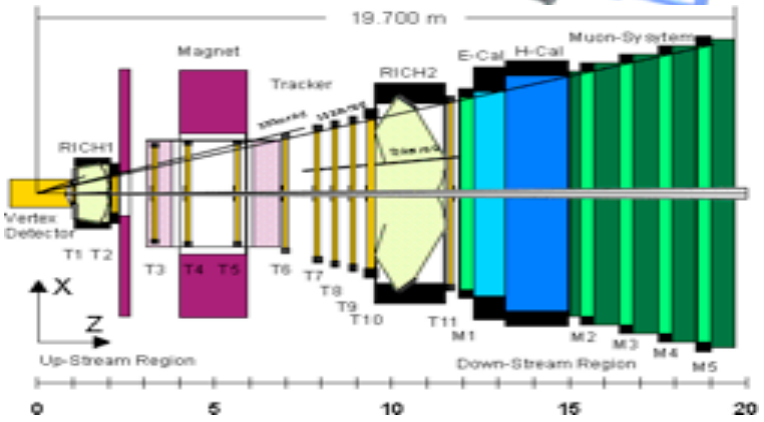
Four LHC Experiments: The Petabyte to Exabyte Challenge

- **ATLAS, CMS, ALICE, LHCB**









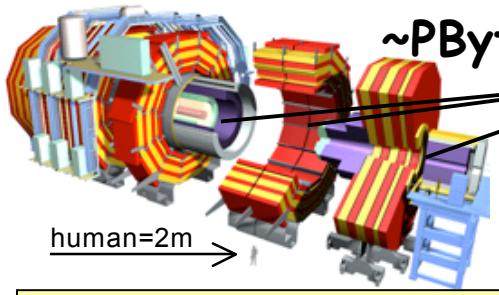
6000+ Physicists & Engineers; 60+ Countries; 250 Institutions

Tens of PB 2008; To 1 EB by ~2015

Hundreds of TFlops To PetaFlops

LHC Data Grid Hierarchy

CMS as example, Atlas is similar



~PByte/sec

Online System

~100 MBytes/sec

Tier 0 +1

HPSS

event reconstruction

event simulation

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

Tier2 Center

Tier 2

analysis

~0.6-2.5 Gbps

Tier 3

Institute ~0.25TIPS

Physics data cache

100 - 1000 Mbits/sec

Tier 4

Workstations

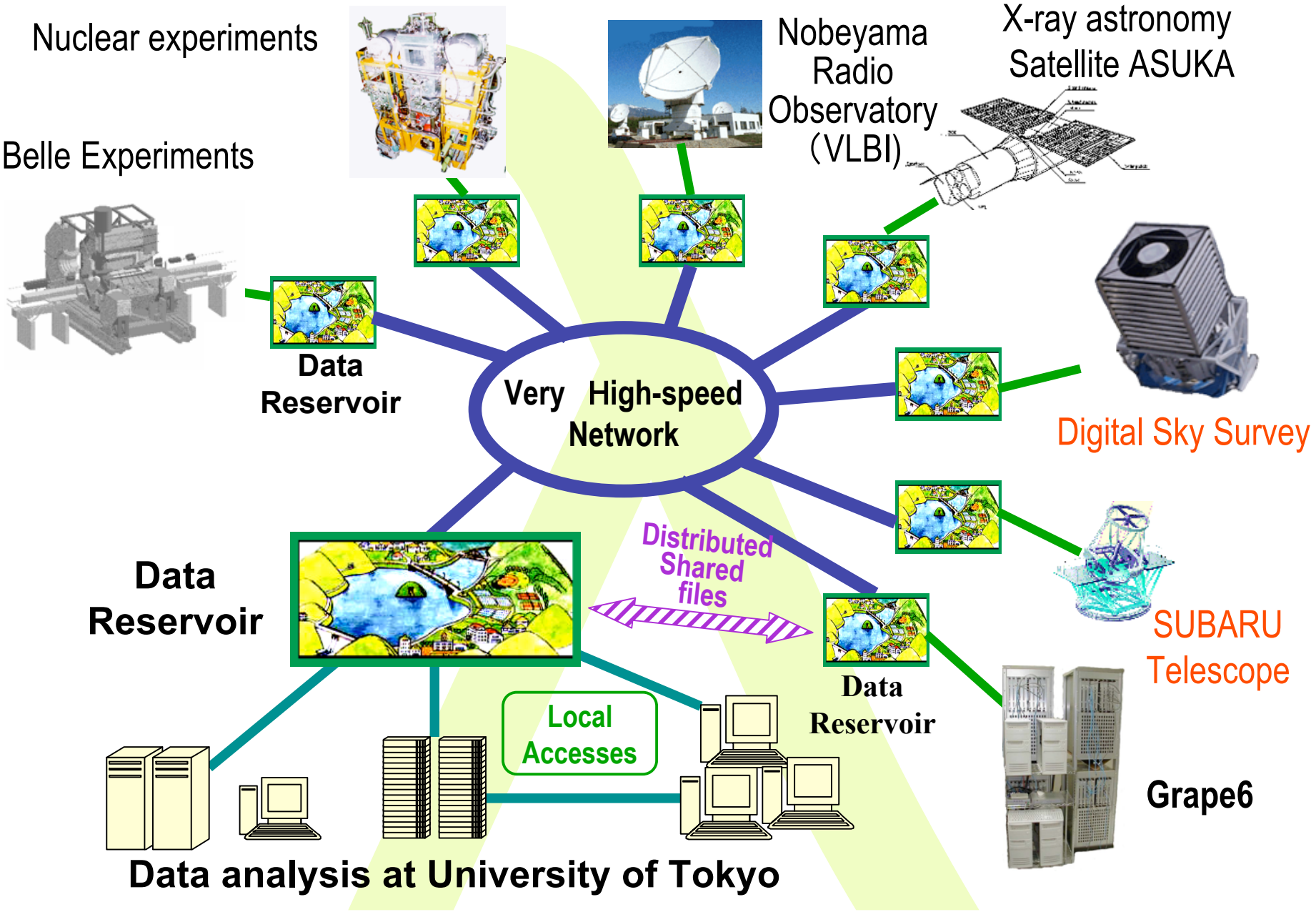
Courtesy Harvey Newman, CalTech and CERN

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

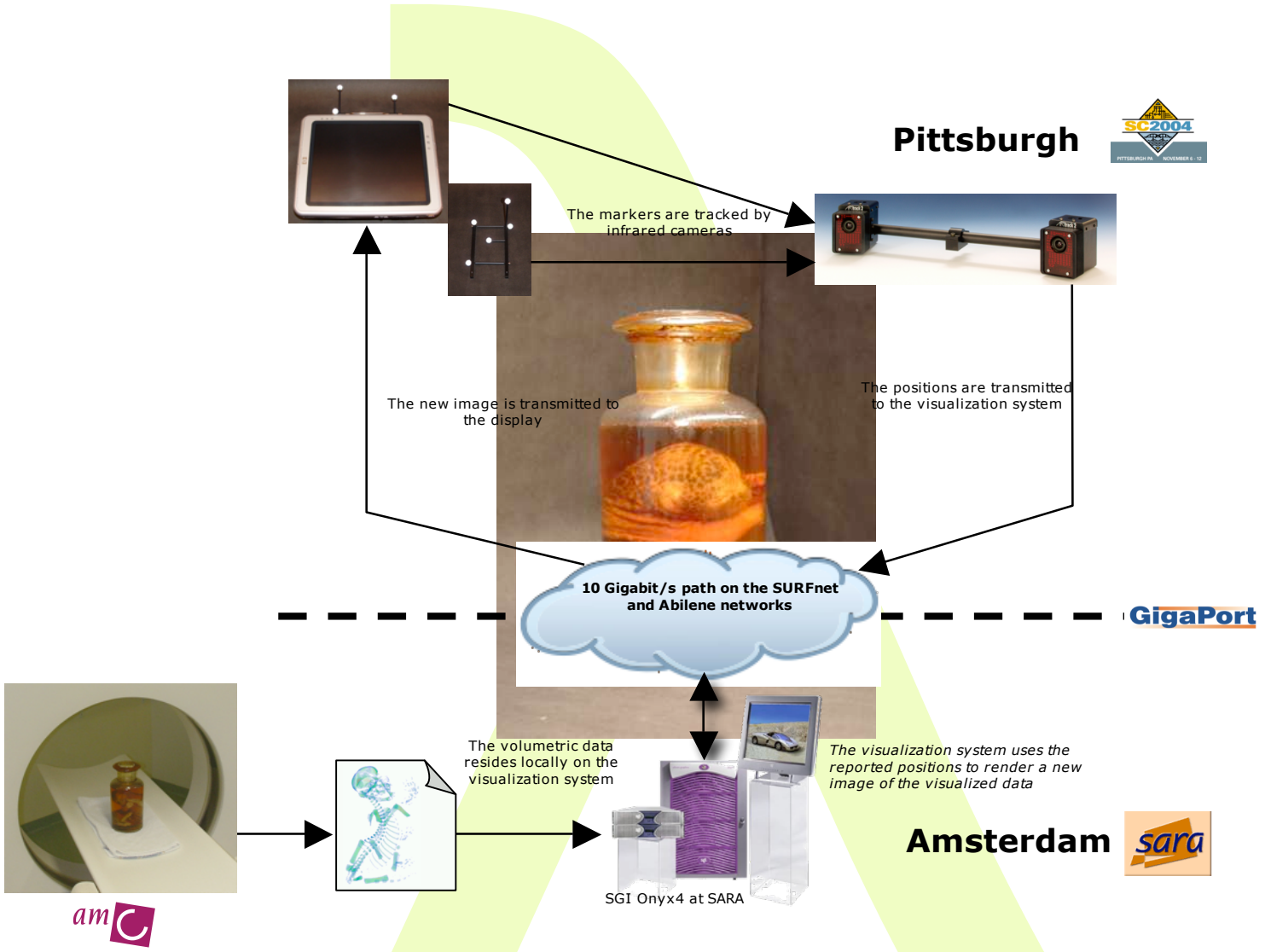
Physicists work on analysis "channels" at 135 institutes. Each institute has ~10 physicists working on one or more channels.

2000 physicists in 31 countries are involved in this 20-year experiment in which DOE is a major player.

Data intensive scientific computation through global networks



Co-located interactive 3D visualization



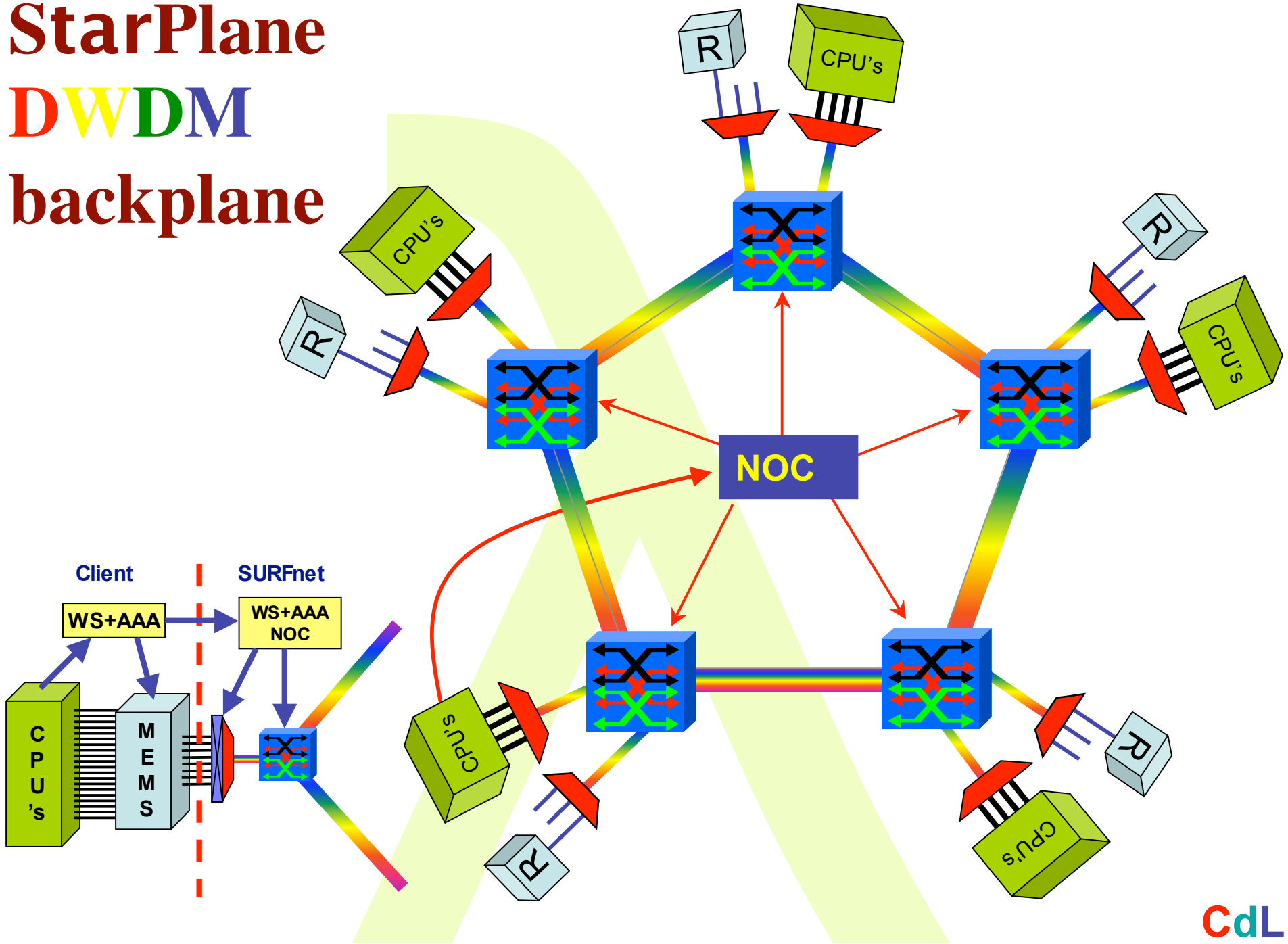
SC2004 “Dead Cat” demo

SuperComputing 2004,
Pittsburgh,
Nov. 6 to 12, 2004

Produced by:
Michael Scarpa
Robert Belleman
Peter Sloot

Many thanks to:
AMC
SARA
GigaPort
UvA/AIR
Silicon Graphics, Inc.
Zoölogisch Museum

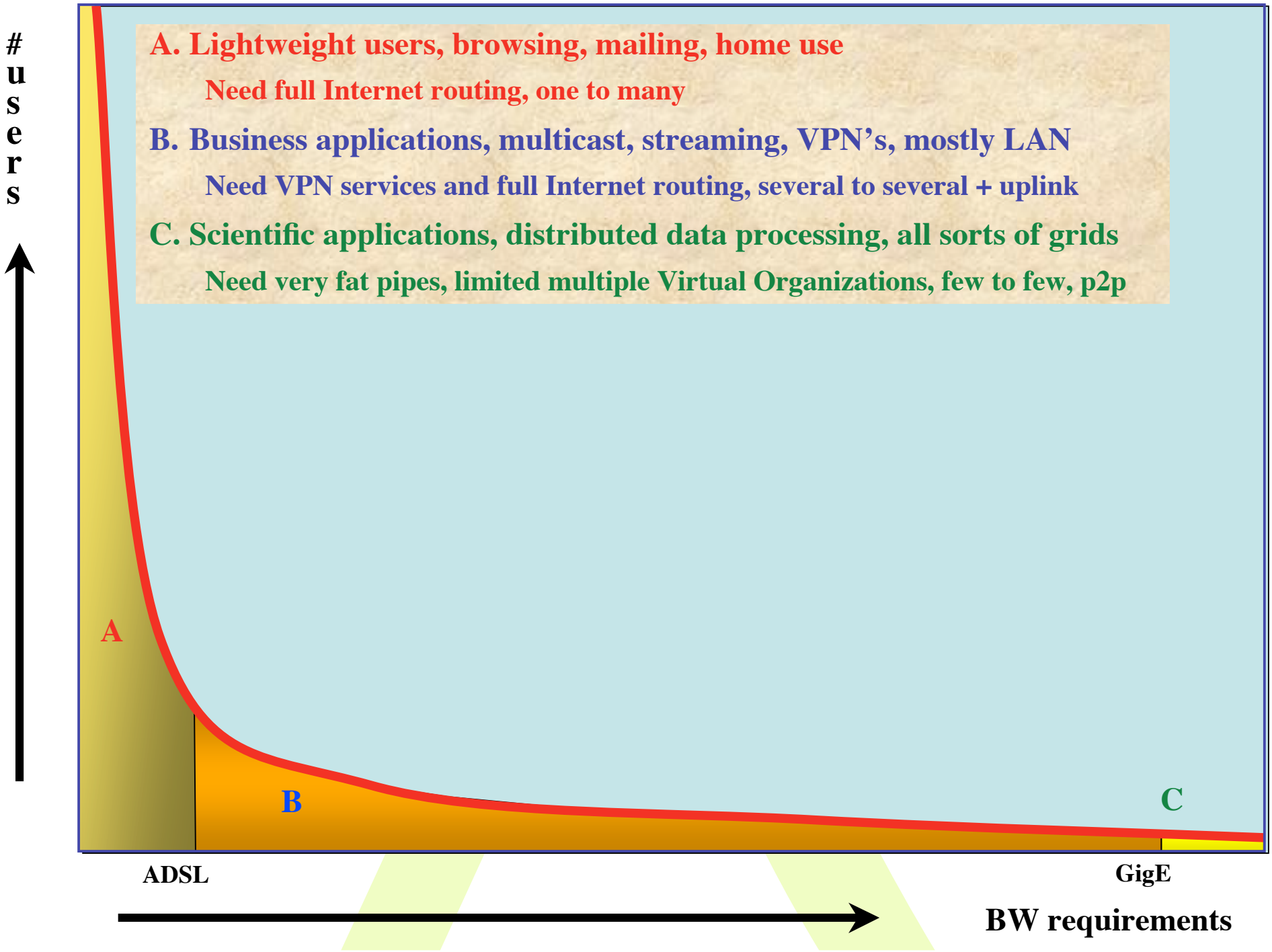
StarPlane DWDM backplane



CdL

Showed you 5 types of Grids

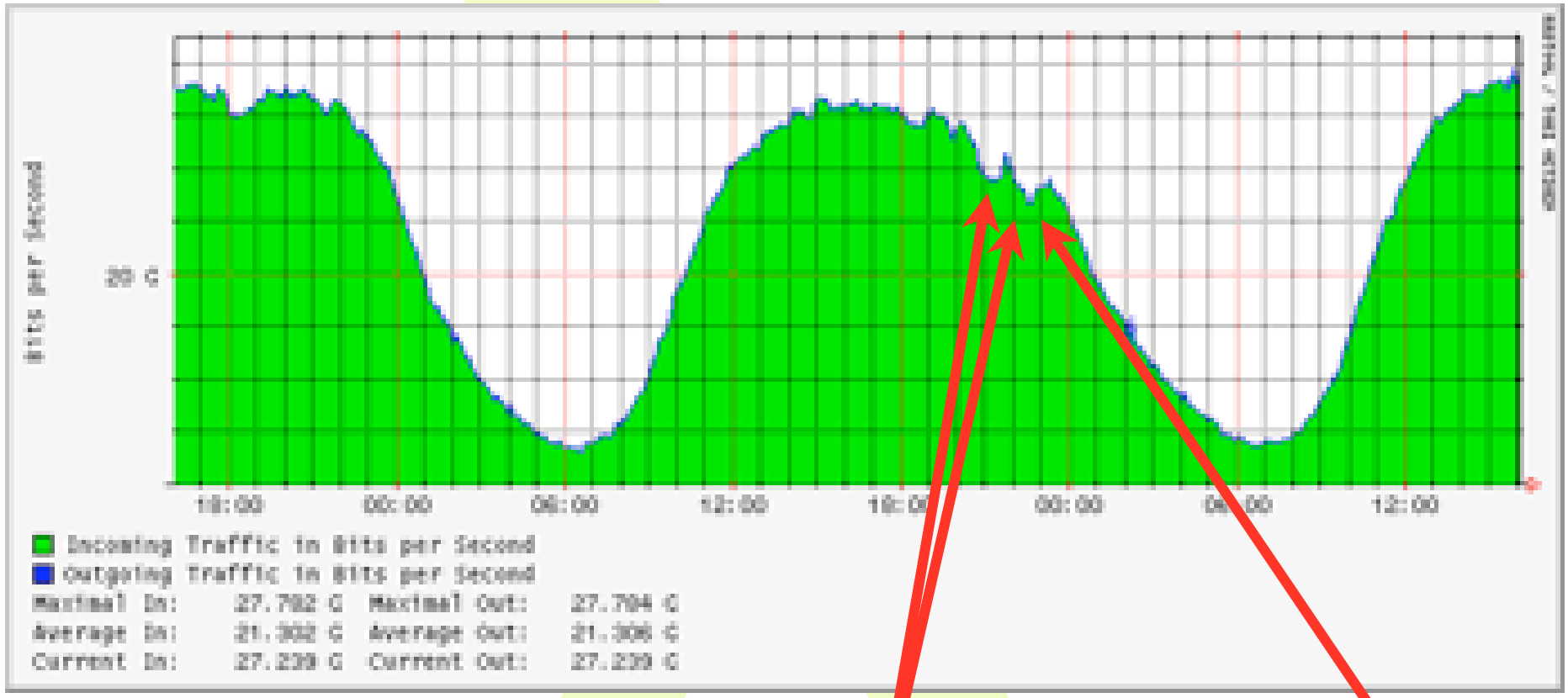
- Sensor Grids
  - Several massive data sources are coming online
- Computational Grids
  - HEP and LOFAR analysis needs massive CPU capacity
  - Research: dynamic nation wide optical backplane control
- Data (Store) Grids
  - Moving and storing HEP, Bio and Health data sets is major challenge
- Visualization Grids
  - Data object (TByte sized) inspection, anywhere, anytime
- Lambda Grids
  - Hybrid networks


A. Lightweight users, browsing, mailing, home use
Need full Internet routing, one to many
B. Business applications, multicast, streaming, VPN's, mostly LAN
Need VPN services and full Internet routing, several to several + uplink
C. Scientific applications, distributed data processing, all sorts of grids
Need very fat pipes, limited multiple Virtual Organizations, few to few, p2p
users
A
B
C
ADSL
GigE
BW requirements

The Dutch Situation

- **Estimate A**
  - **17 M people, 6.4 M households, 25 % penetration of 0.5-2.0 Mb/s ADSL, 40 times under-provisioning ==> 20 Gb/s**

AMS-IX



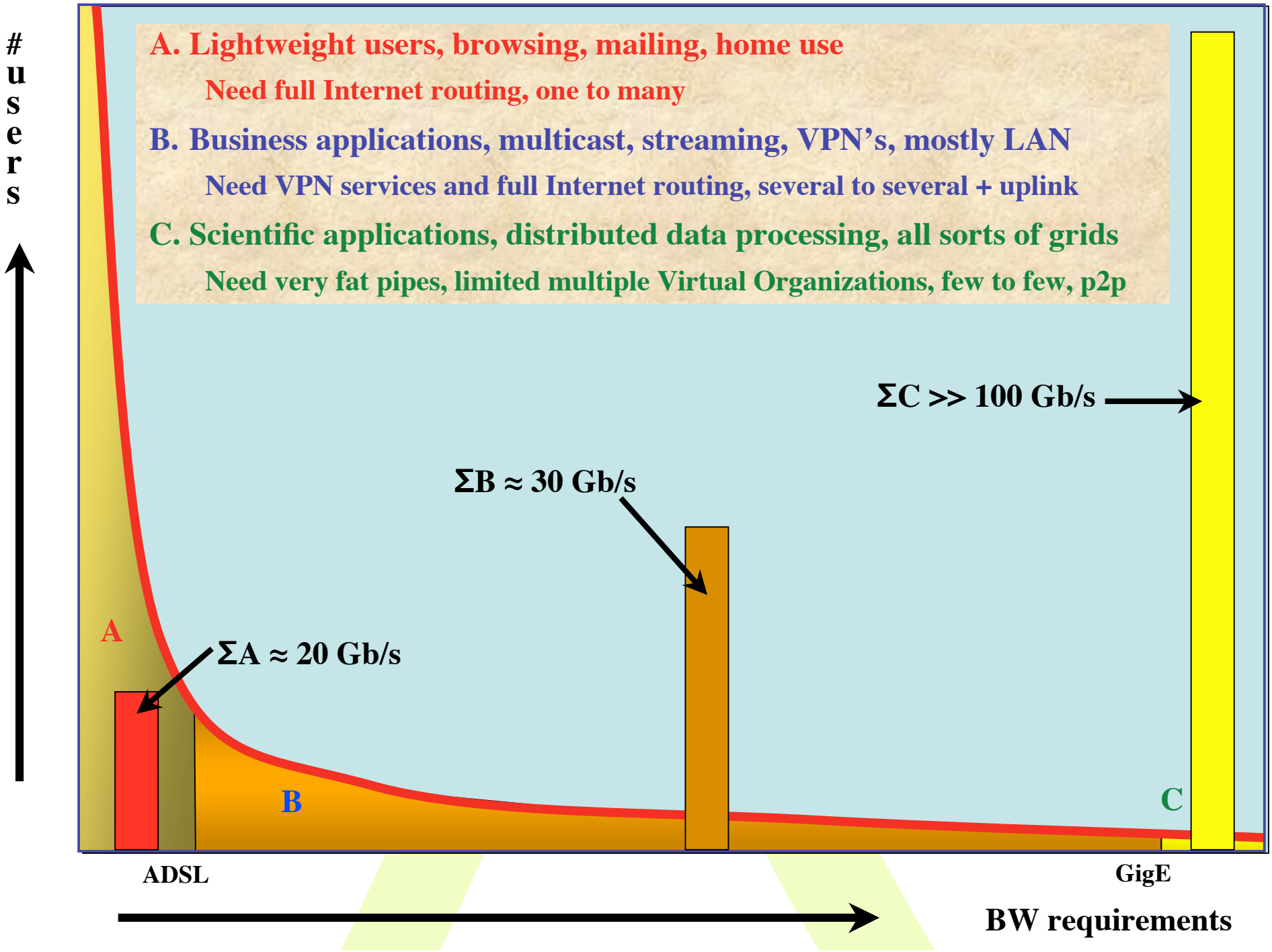
June 19th 2004

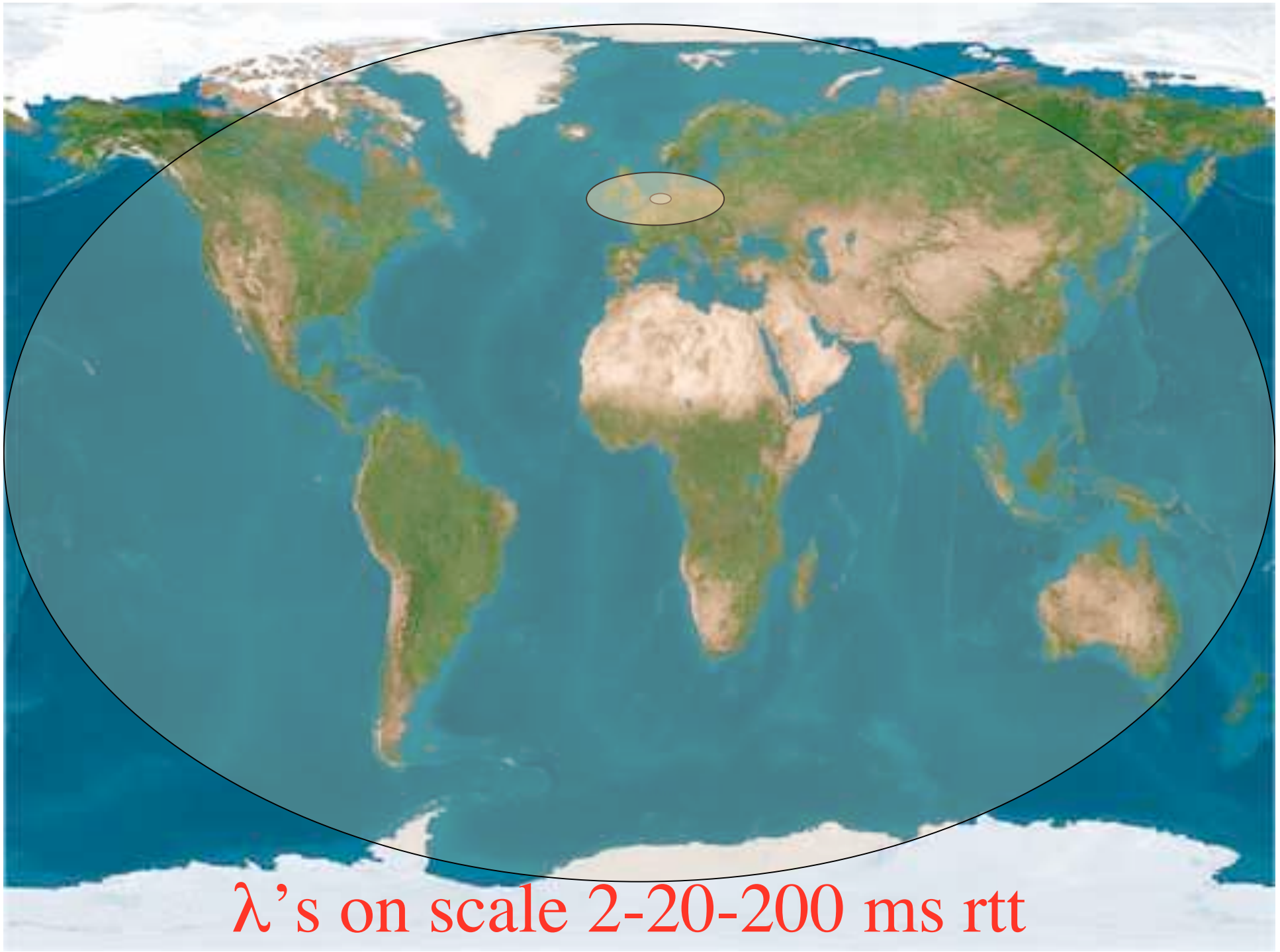
Lost :-(

European championship football Holland -- Czech Republic

The Dutch Situation

- **Estimate A**
  - **17 M people, 6.4 M households, 25 % penetration of 0.5-2.0 Mb/s ADSL, 40 times under-provisioning ==> 20 Gb/s**
- **Estimate B**
  - **SURFnet5 has 2*10 Gb/s to about 15 institutes and 0.1 to 1 Gb/s to 170 customers, estimate same for industry (overestimation) ==> 10-30 Gb/s**
- **Estimate C**
  - **Leading HEF and ASTRO + rest ==> 80-120 Gb/s**
  - **LOFAR ==> ≈ 37 Tbit/s ==> ≈ n x 10 Gb/s**

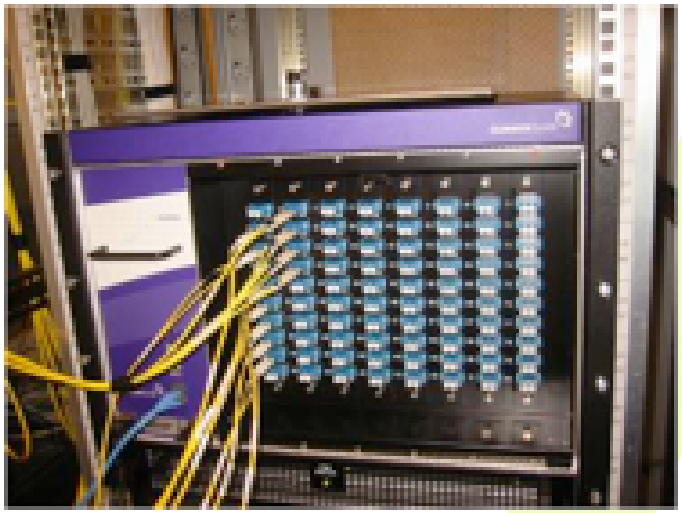

#users
A. Lightweight users, browsing, mailing, home use
Need full Internet routing, one to many
B. Business applications, multicast, streaming, VPN's, mostly LAN
Need VPN services and full Internet routing, several to several + uplink
C. Scientific applications, distributed data processing, all sorts of grids
Need very fat pipes, limited multiple Virtual Organizations, few to few, p2p
ΣC >> 100 Gb/s
ΣB ≈ 30 Gb/s
ΣA ≈ 20 Gb/s
A
B
C
ADSL
GigE
BW requirements


λ's on scale 2-20-200 ms rtt

Towards Hybrid Networking!

- **Costs of optical equipment 10% of switching 10 % of full routing equipment for same throughput**
  - **10G routerblade -> 100-500 k$, 10G switch port -> 10-20 k$, MEMS port -> 0.7 k$**
  - **DWDM lasers for long reach expensive, 10-50k$**
- **Bottom line: look for a hybrid architecture which serves all classes in a cost effective way (map A -> L3 , B -> L2 , C -> L1)**
- **Give each packet in the network the service it needs, but no more !**

L1 - 0.7 k$/port



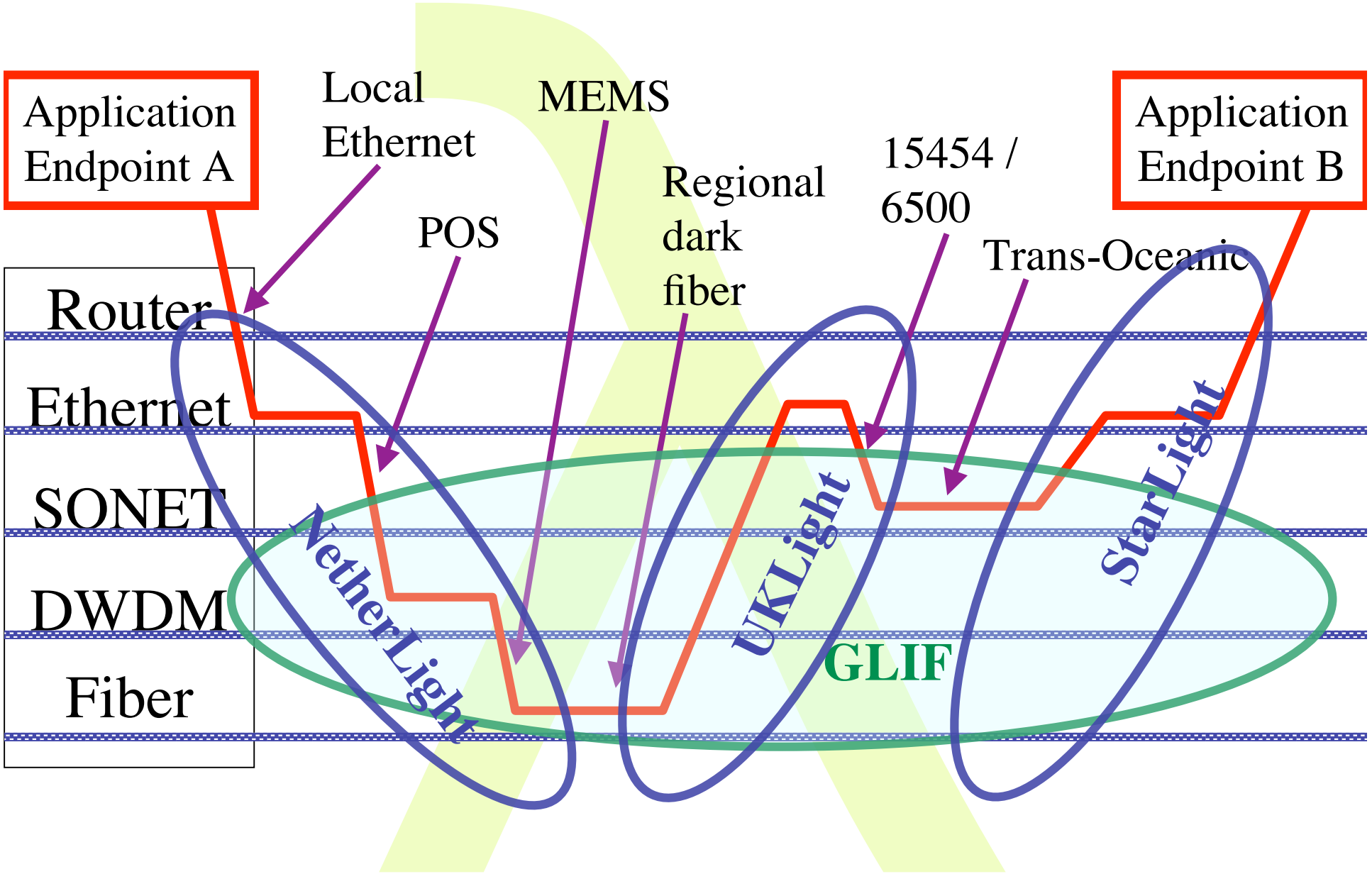
L2 - 10-20 k$/port

L3 - 100-500 k$/port

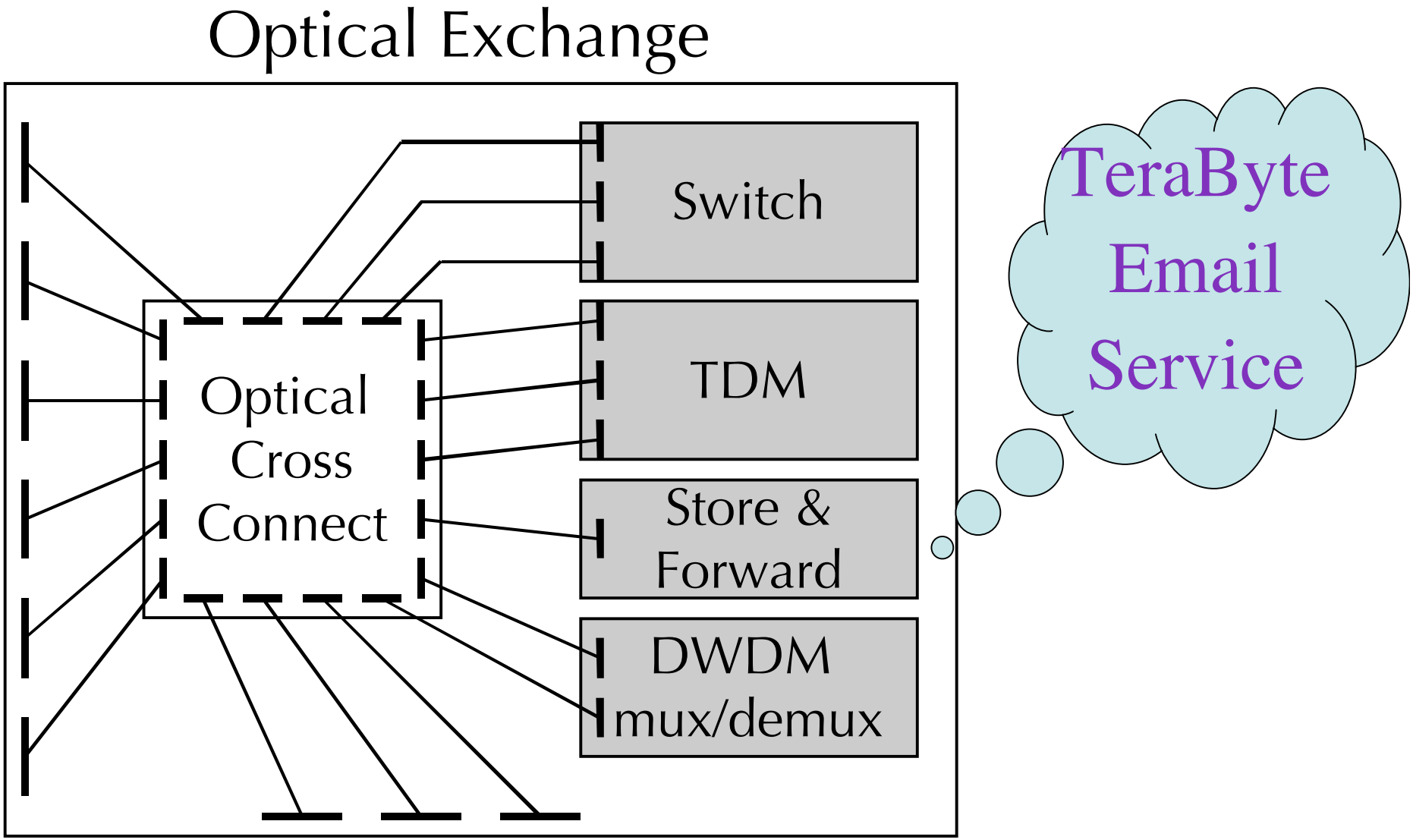
Services

SCALE / CLASS	2 Metro	20 National/ regional	200 World
A	Switching/ routing	Routing	ROUTER$
B	Switches + E-WANPHY VPN's	Switches + E-WANPHY (G)MPLS	ROUTER$
C	dark fiber DWDM MEMS switch	DWDM, TDM / SONET Lambda switching	Lambdas, VLAN's SONET Ethernet

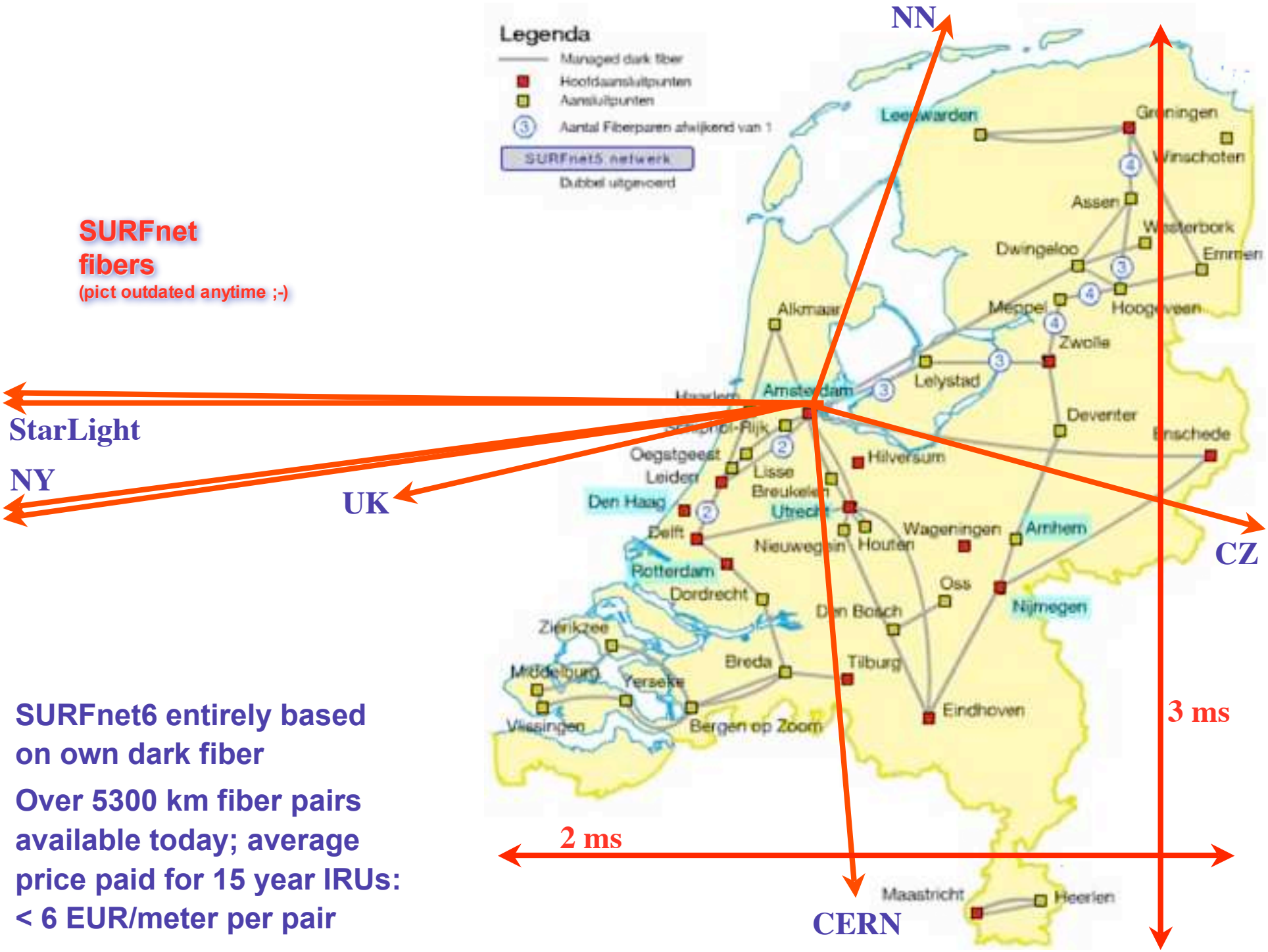
How low can you go?



Optical Exchange as Black Box

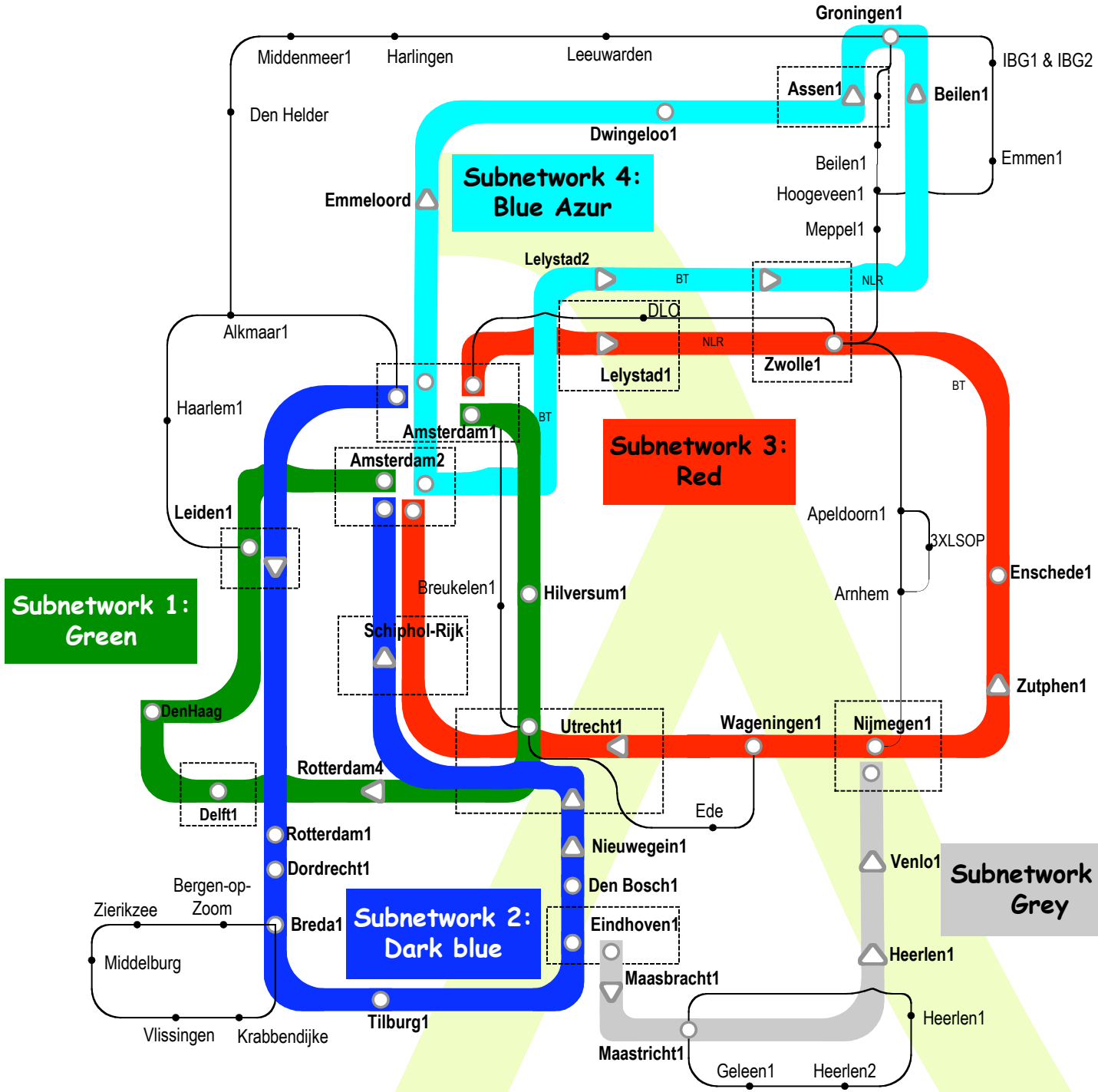


SURFnet fibers
(pict outdated anytime ;-)



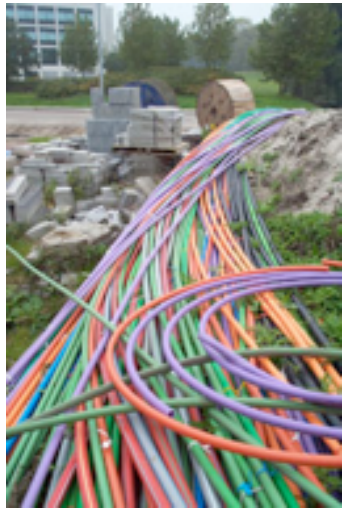
SURFnet6 entirely based on own dark fiber

Over 5300 km fiber pairs available today; average price paid for 15 year IRUs: < 6 EUR/meter per pair



Common Photonic Layer (CPL) in SURFnet6

Laying of fiber near/at Science Park Amsterdam



Pictures by Yuri Demchenko

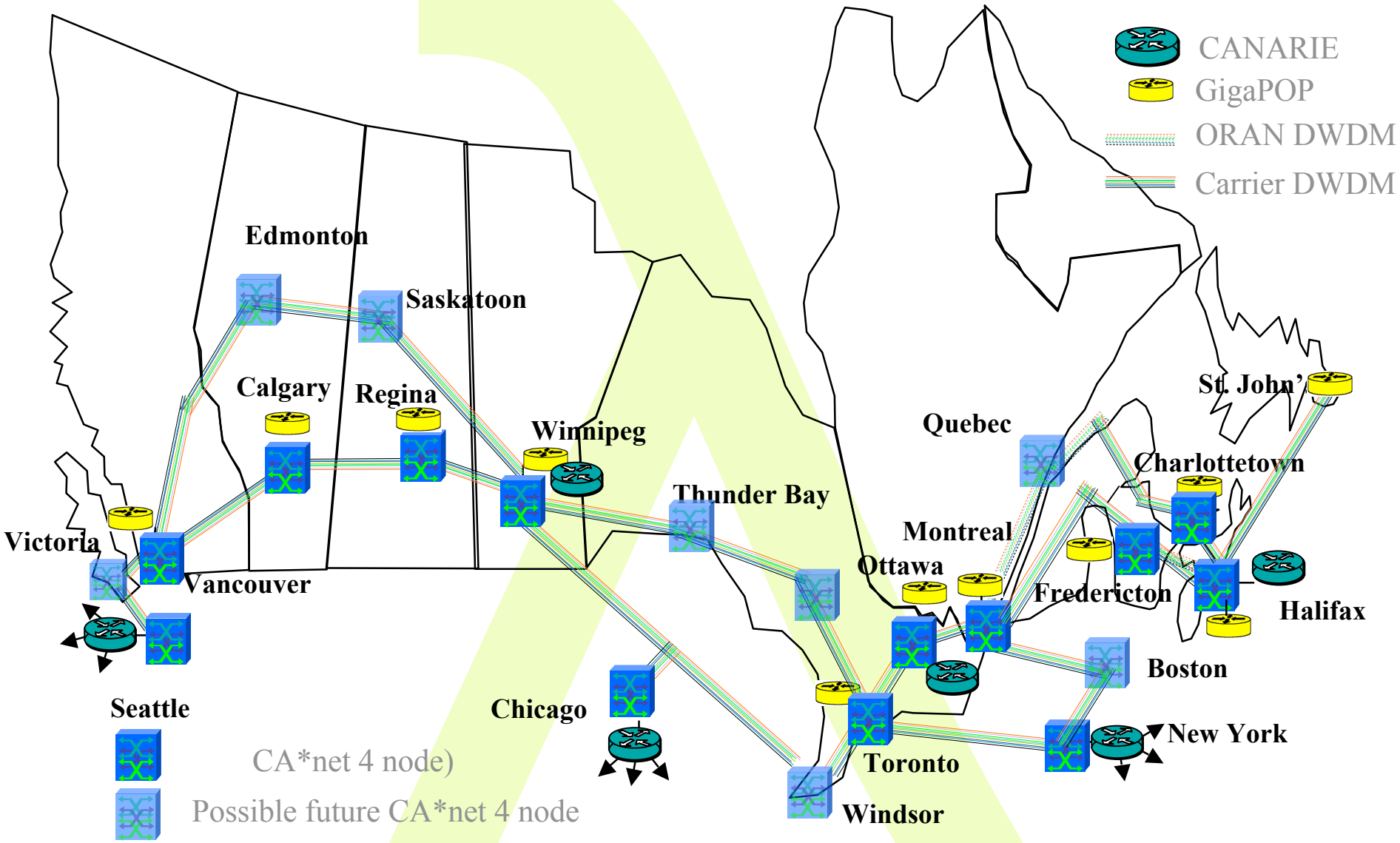
SURFnet on Lambda inspection in Science Park Amsterdam :-)

UCLP intended for projects like National LambdaRail

CAVEwave partner acquires a separate wavelength between San Diego and Chicago and wants to manage it as part of its network including add/drop, routing, partition etc

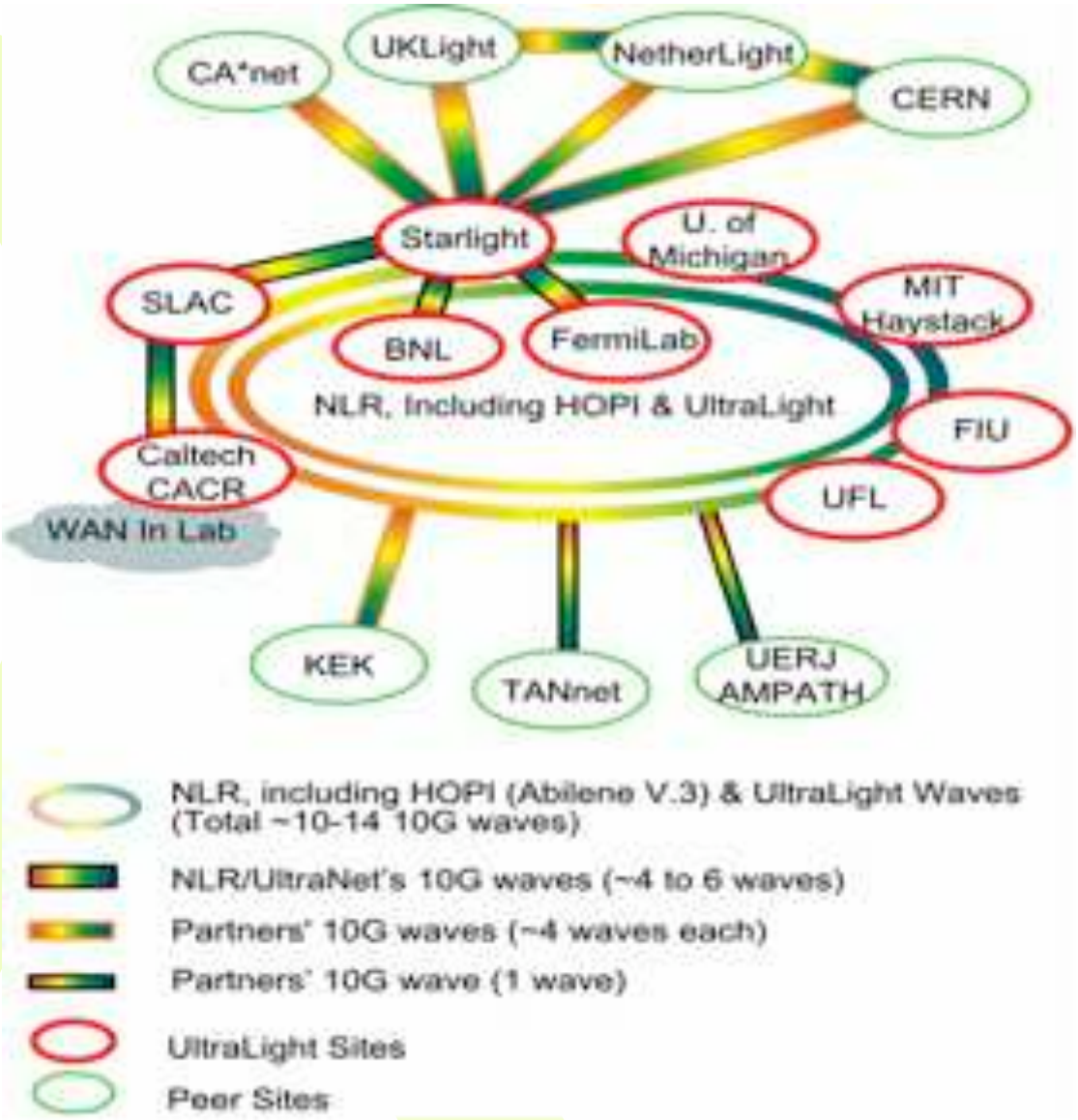
NLR Condominium lambda network

CA*net 4 Architecture



UltraLight Network: **PHASE III**

- Move into production
- Optical switching fully enabled amongst primary sites
- Integrated international infrastructure



GLIF: Global Lambda Integrated Facility

- Established at the 3rd Lambda Grid Workshop, August 2003 in Reykjavik, Iceland
- Collaborative initiative among worldwide NRENs, institutions and their users
- A world-scale Lambda-based Laboratory for application and middleware development

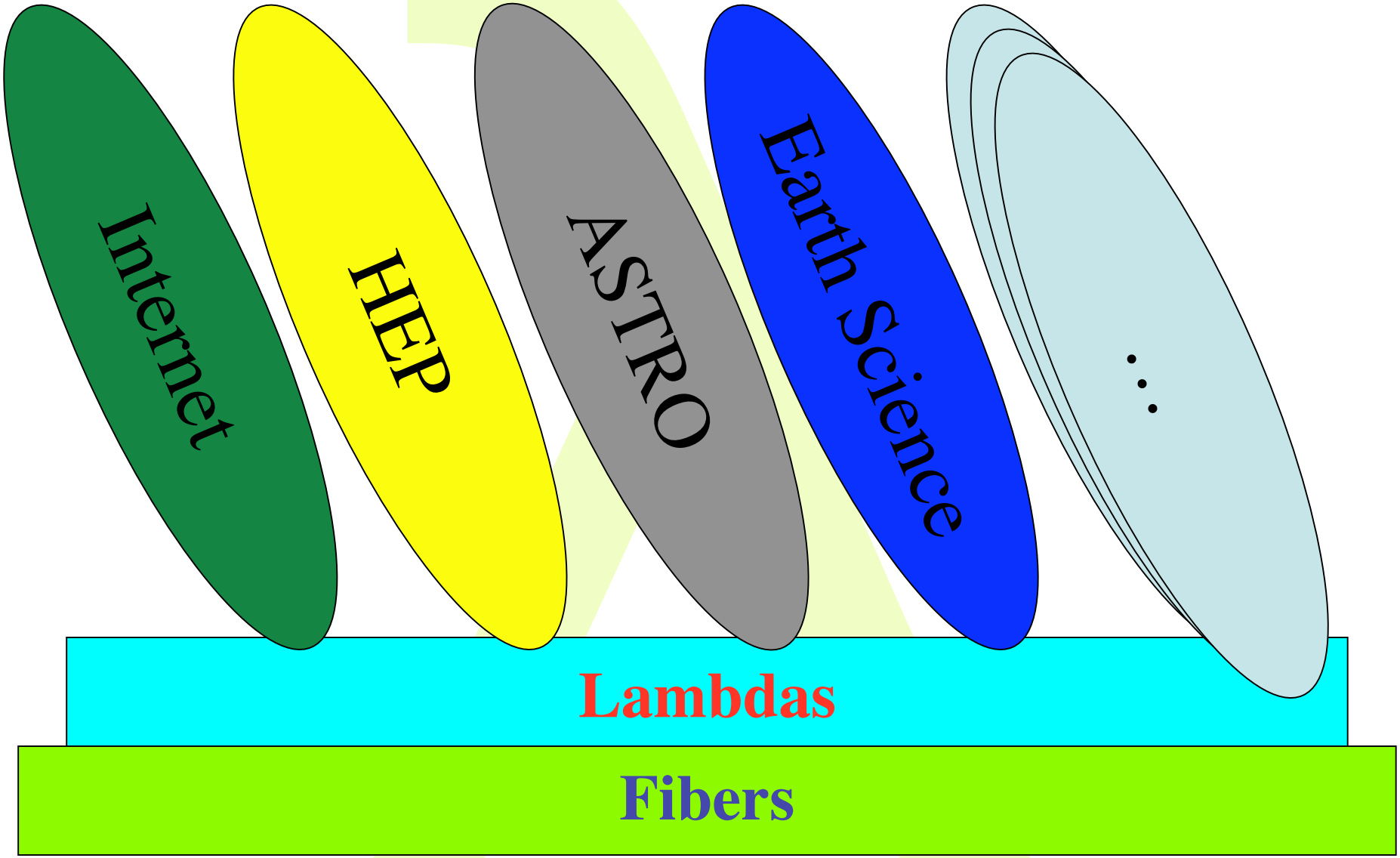
GLIF vision:

GLIF is a world-scale Lambda-based Laboratory for application and middleware development on emerging LambdaGrids, where applications rely on dynamically configured networks based on optical wavelengths!

GLIF Q4 2004

Visualization courtesy of
Bob Patterson, NCSA.

Discipline Networks



Research on Networks (CdL)

GigaPort

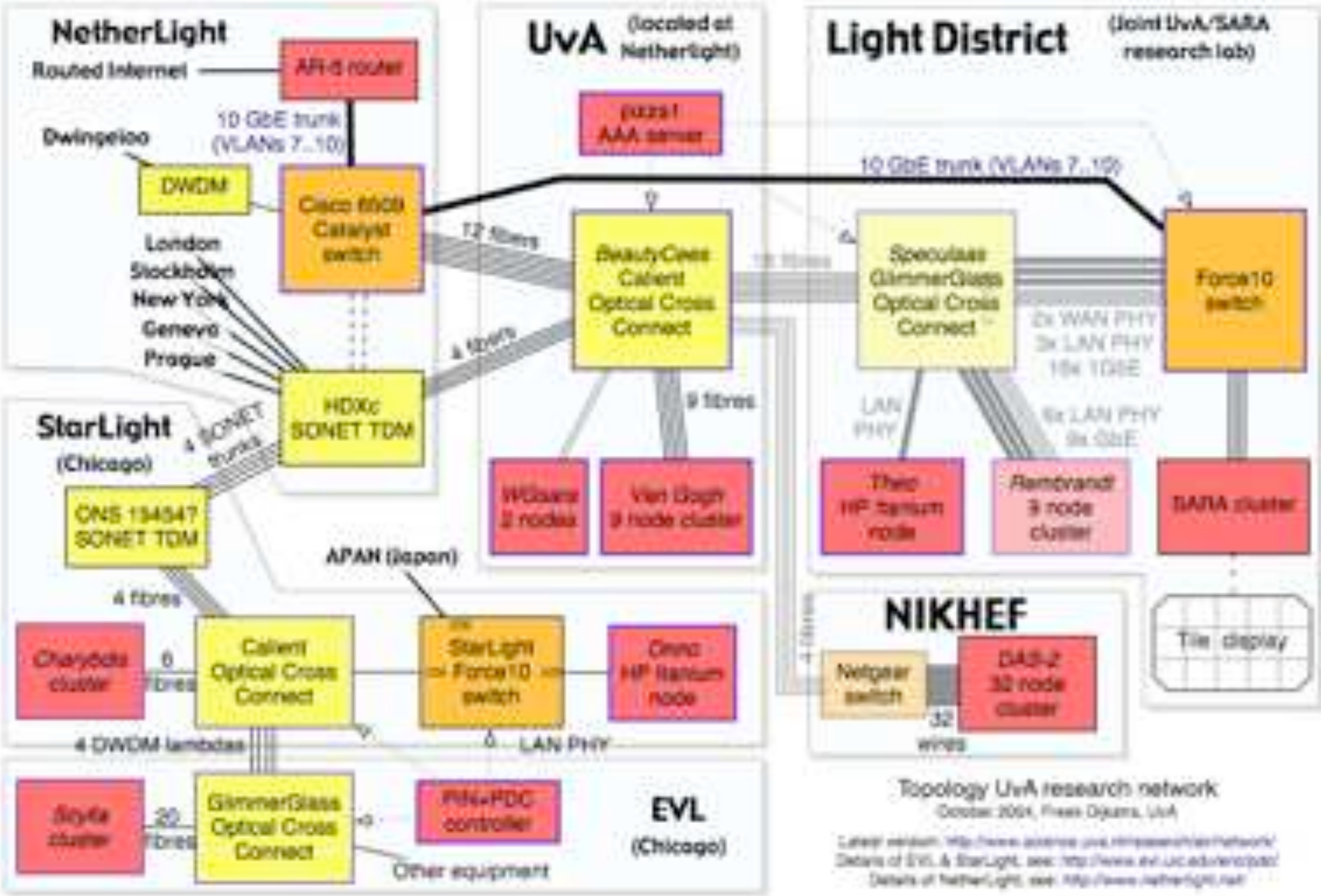
- **Optical Networking:**
  - What innovation in architectural models, components, control and light path provisioning are needed to integrate dynamically configurable optical transport networks and traditional IP networks to a generic data transport platform that provides end-to-end IP connectivity as well as light path (lambda and sub-lambda) services?
- **High performance routing and switching:**
  - What developments need to be made in the Internet Protocol Suite to support data intensive applications, and scale the routing and addressing capabilities to meet the demands of the research and higher education communities in the forthcoming 5 years?
- **Management and monitoring:**
  - What management and monitoring models on the dynamic hybrid network infrastructure are suited to provide the necessary high level information to support network planning, network security and network management?
- **Grids and access; reaching out to the user:**
  - What new models, interfaces and protocols are capable of empowering the (grid) user to access, and the provider to offer, the network and grid resources in a uniform manner as tools for scientific research?
- **Testing methodology:**
  - What are efficient and effective methods and setups to test the capabilities and performance of the new building blocks and their interworking, needed for a correct functioning of a next generation network?

SURFnet

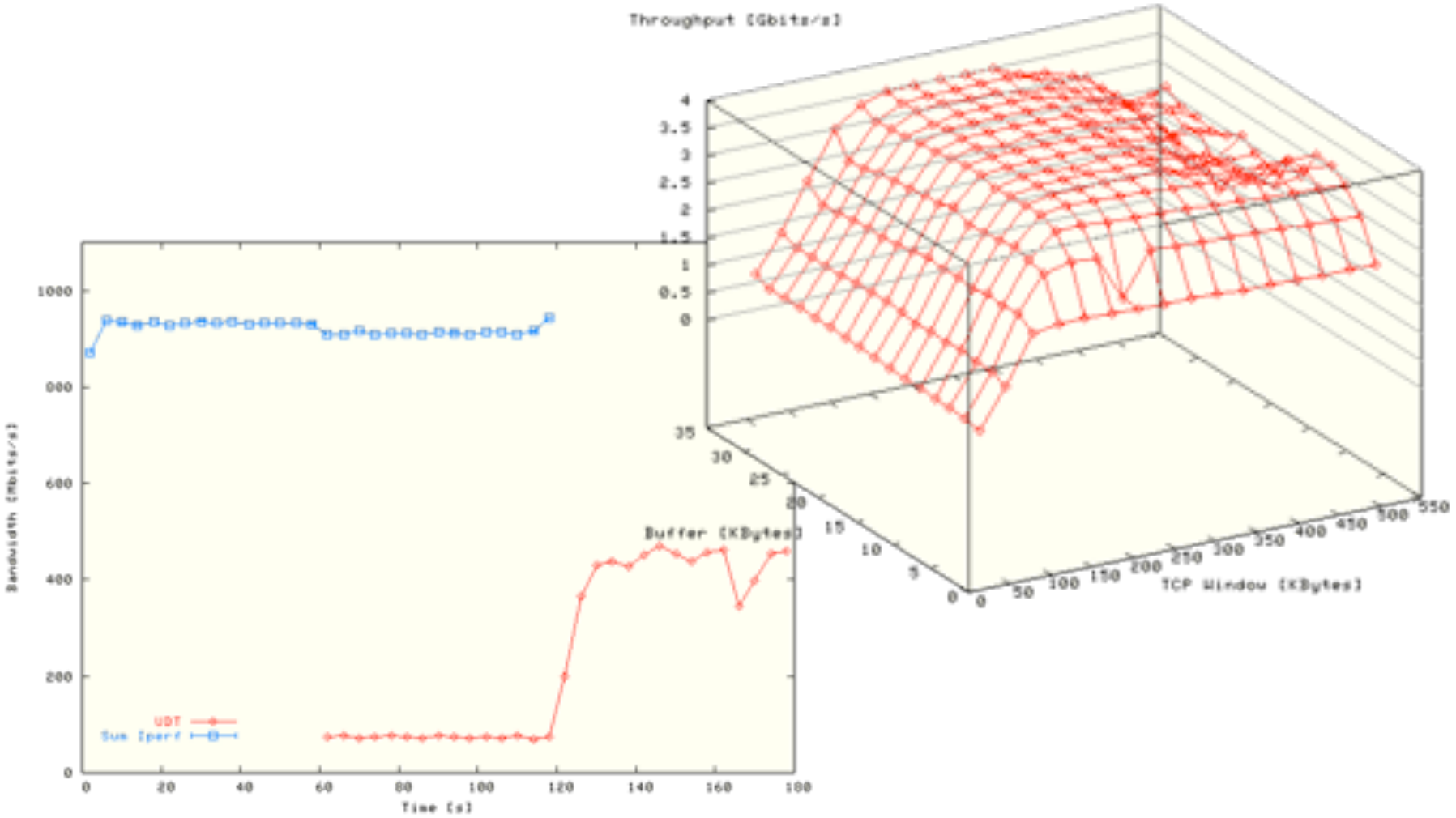
Advanced Internet Research Group @ UvA

- **Optical networking architectures and models**
  - **Optical Internet Exchange architecture**
  - **Lambda routing and assignment**
- **IP transport protocols, performances monitoring and measurements**
  - **With respect to performance**
  - **Monitoring and reporting**
  - **Traffic generation with grid infrastructure**
- **Authorization, Authentication and Accounting**
  - **Concepts**
  - **Proof of concepts**
  - **Network & Grid integration and Applications**

LightHouse


NetherLight
Routed Internet
AR-5 router
10 GbE trunk
(VLANs 7..10)
Dwingeloo
DWDM
Cisco 6509
Catalyst
switch
London
Stockholm
New York
Geneva
Prague
HDXc
SONET TDM
StarLight
(Chicago)
4 SONET trunks
ONS 15454?
SONET TDM
UvA
(located at
Netherlight)
pizza1
AAA server
10 GbE trunk (VLANs 7..10)
12 fibres
4 fibres
BeautyGlass
Calient
Optical Cross
Connect
9 fibres
WGlass
2 nodes
Van Gogh
9 node cluster
Light District
(Joint UvA/SARA
research lab)
Speculaas
GlimmerGlass
Optical Cross
Connect
Force10
switch
2x WAN PHY
3x LAN PHY
16x 1GbE
LAN
PHY
6x LAN PHY
9x GbE
Theo
HP Itanium
node
Rembrandt
9 node
cluster
SARA cluster
Tile display
NIKHEF
Netgear
switch
DAS-2
32 node
cluster
32
wires
APAN (Japan)
4 fibres
Charybdis
cluster
6
fibres
Calient
Optical Cross
Connect
StarLight
Force10
switch
Onno
HP Itanium
node
4 DWDM lambdas
LAN PHY
Scylla
cluster
20
fibres
GlimmerGlass
Optical Cross
Connect
PIN+PDC
controller
EVL
(Chicago)
Other equipment
Topology UvA research network

Example Measurements



Layer - 2 requirements from 3/4

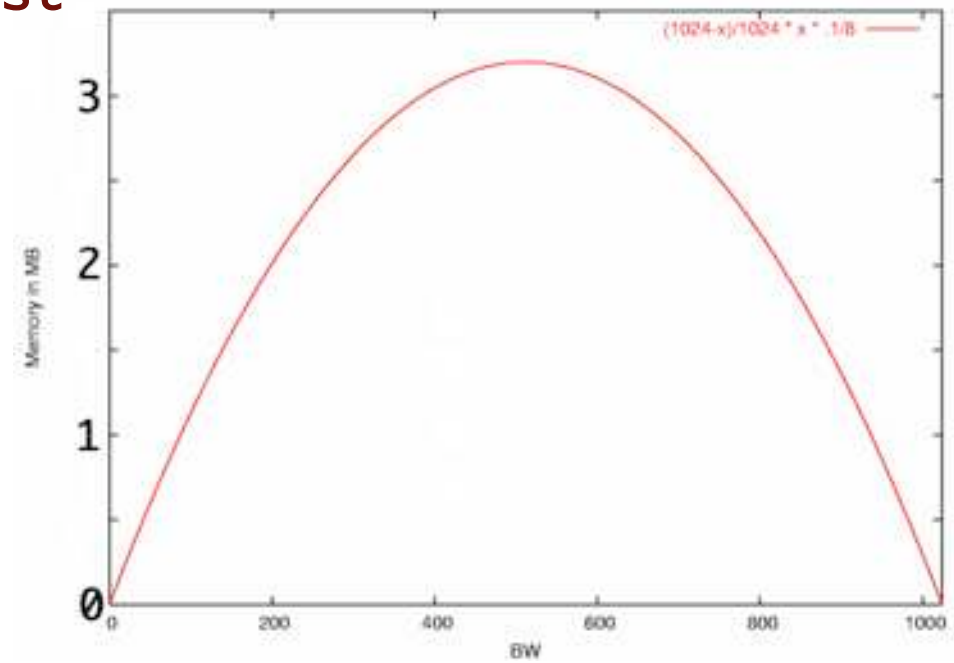
TCP is bursty due to sliding window protocol and slow start algorithm.

```
Window = BandWidth * RTT     &    BW == slow
```

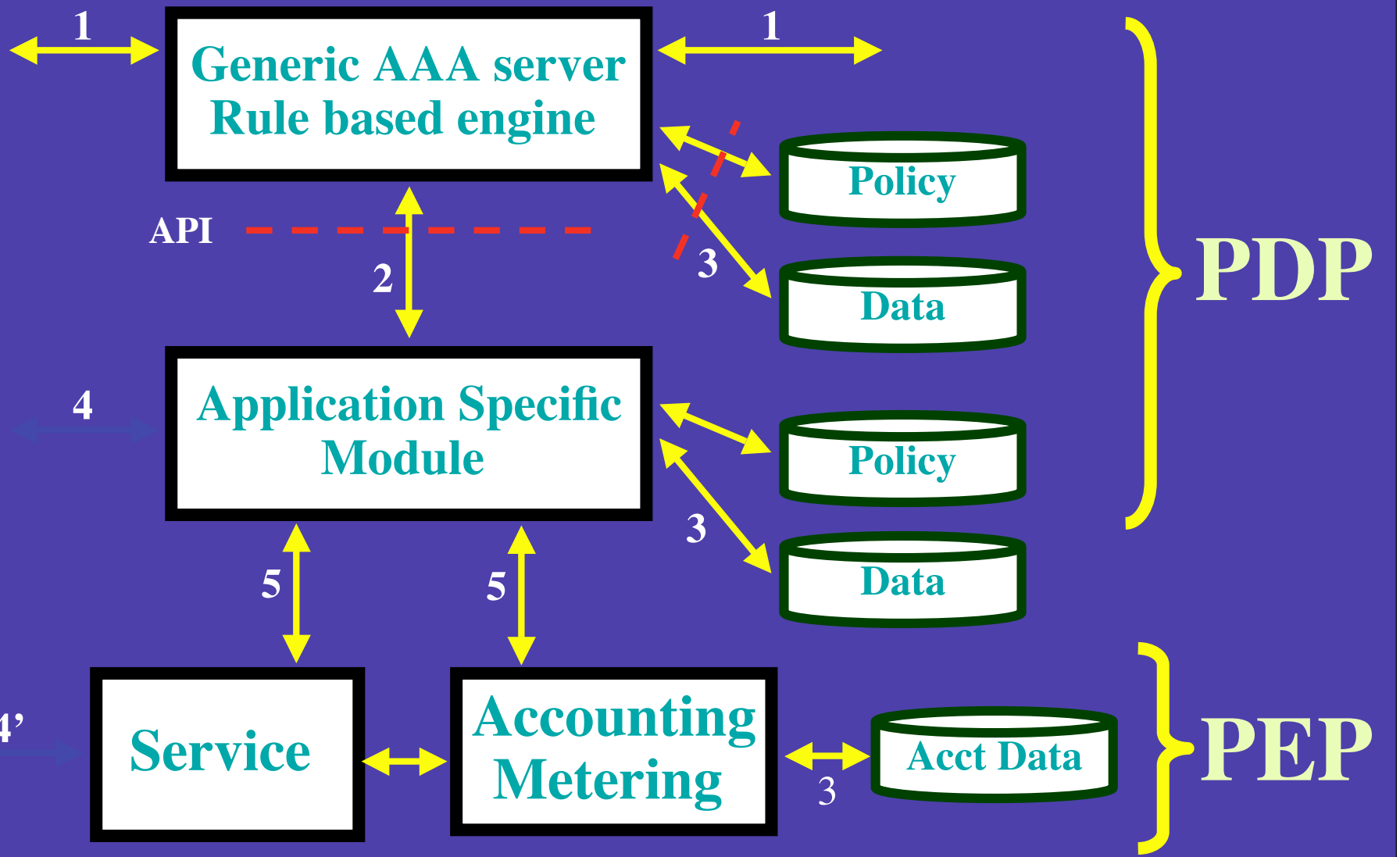
$$\text{Memory-at-bottleneck} = \frac{\text{fast} - \text{slow}}{\text{fast}} * \text{slow} * \text{RTT}$$

So pick from menu:

- *Flow control*
- *Traffic Shaping*
- *RED (Random Early Discard)*
- *Self clocking in TCP*
- *Deep memory*

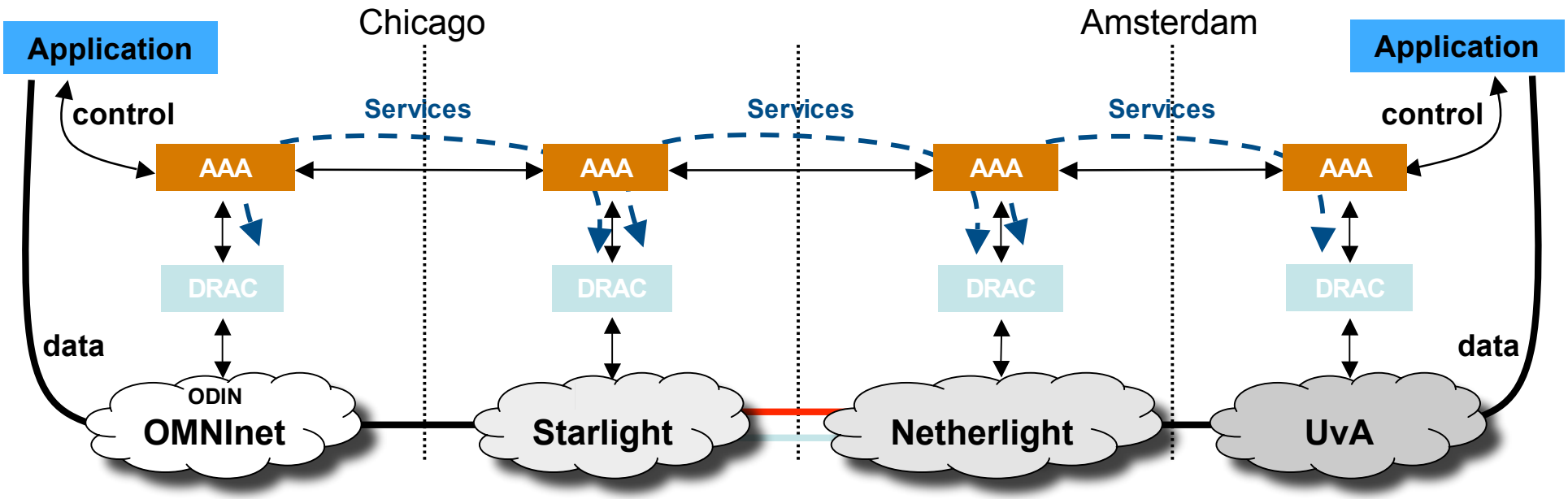


Starting point



RFC 2903 - 2906 , 3334 , policy draft

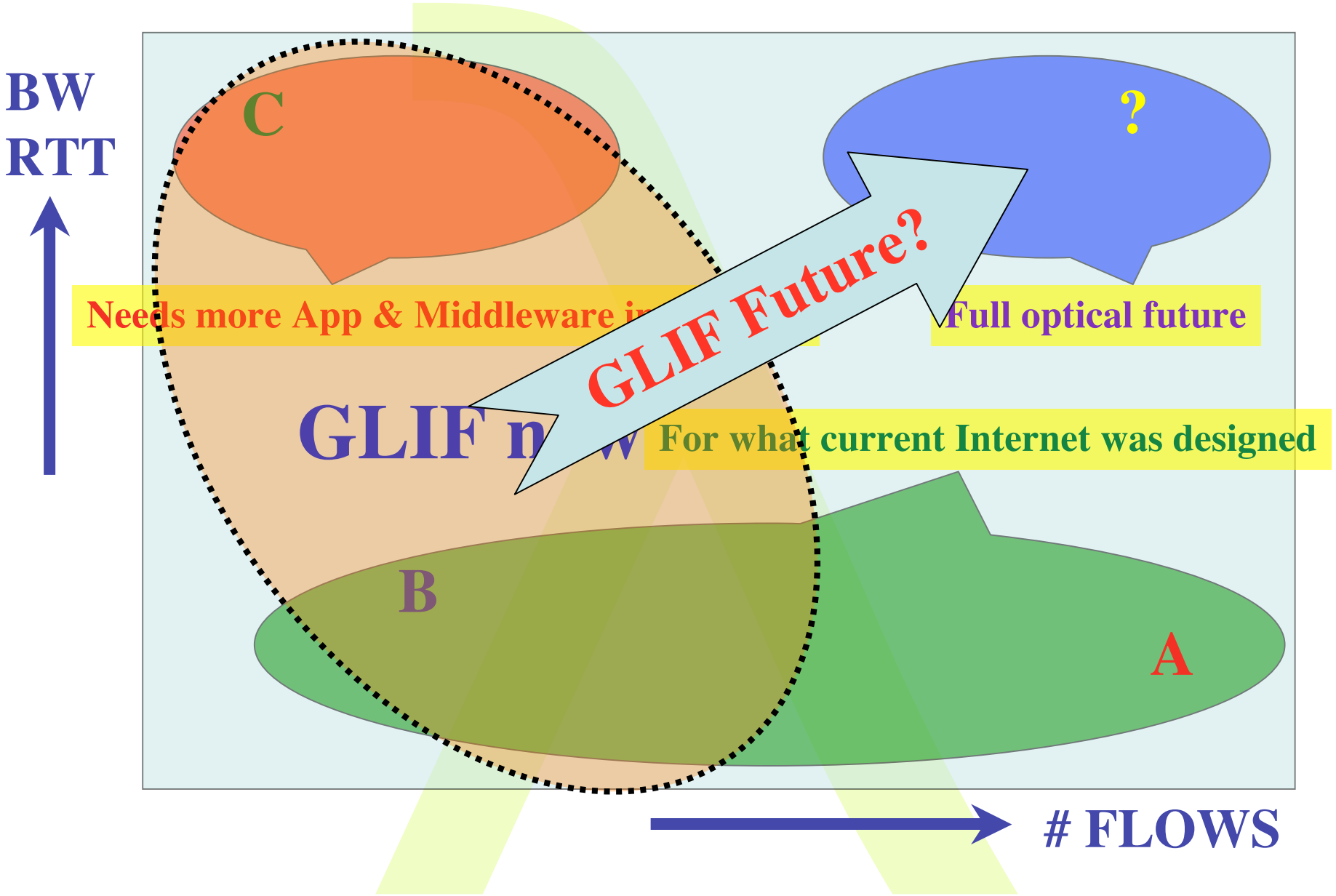
SC2004 CONTROL CHALLENGE



- finesse the control of bandwidth across multiple domains
- while exploiting scalability and intra- , inter-domain fault recovery
- thru layering of a novel SOA upon legacy control planes and NEs


GOT
CONTROL
TOO?

Transport of flows



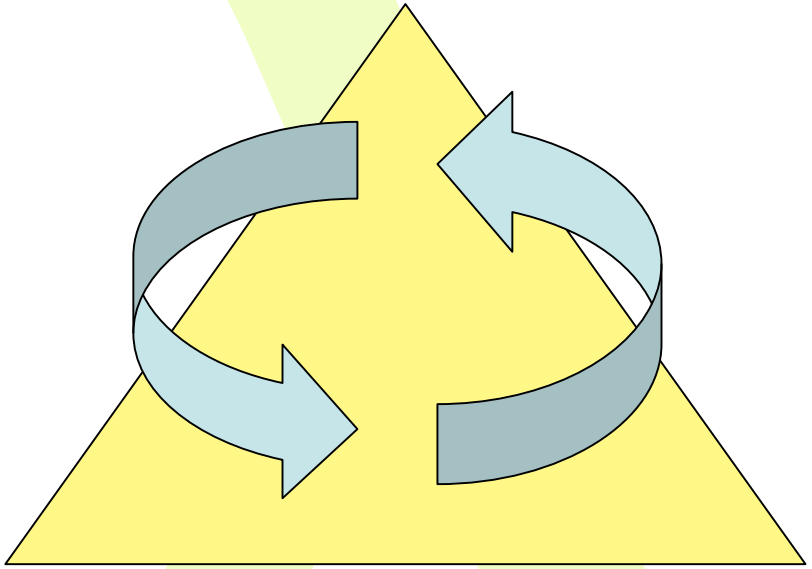
Open research questions

- LightPath into organizations like ours
- Sub-second true Lambda provisioning
- Massive data transport, storage and handout
- Scheduling of resources for workflows
- Complex authorization
- Brokering
- Security

Science park

Creators

University,
Institutes



Producers

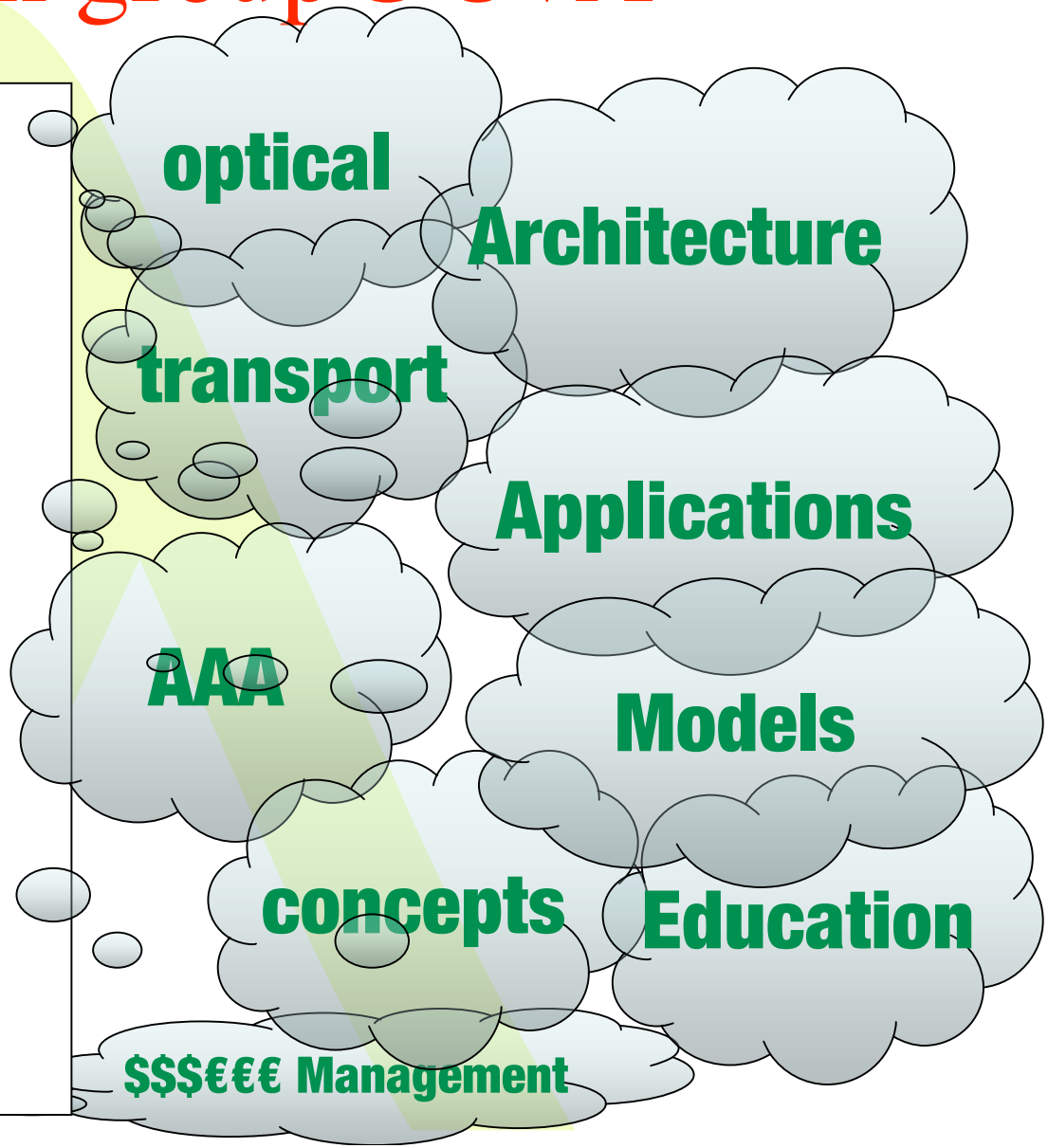
Supercomputer center,
Network exchanges,
Storage providers
Visualization facilities

Consumers

E-Sciences,
HEP, Bio, Solid State Phys.
Industry

OSI model Advanced Internet Research group@UvA

- Freek Dijkstra
- Hans Blom
- Bert Andree
- Paola Grosso
- Jeroen van der Ham
- Martijn Steenbakkers
- Bas van Oudenaarde
- Fred Wan
- Arie Taal
- Yuri Demchenko
- Leon Gommans
- Rob Meijer
- Karst Koymans
- Cees de Laat



Not quite THE END

Thanks to

SURFnet: Kees Neggers,UIC&iCAIR: Tom DeFanti, Joel Mambretti, CANARIE: Bill St, Arnaud

Freek Dijkstra, Hans Blom, Leon Gommans, Bas van oudenaarde, Arie Taal, Pieter de Boer, Bert Andree, Martijn de Munnik, Antony Antony, Rob Meijer, VL-team.

Partially complete list:

- Caas
- Chase
- Cess
- Kess
- Case

World of Tomorrow - 2005

September 26-30, 2005
University of California, San Diego
California Institute for Telecommunications and Information Technology [Cal-(IT)^2]
United States

iGrid 2002 was held at Science park Amsterdam