Lambda-Grid developments

Global Lambda Integrated Facility

www.science.uva.nl/~delaat

Cees de Laat

GigaPort

EU

University of Amsterdam

SARA

NCF

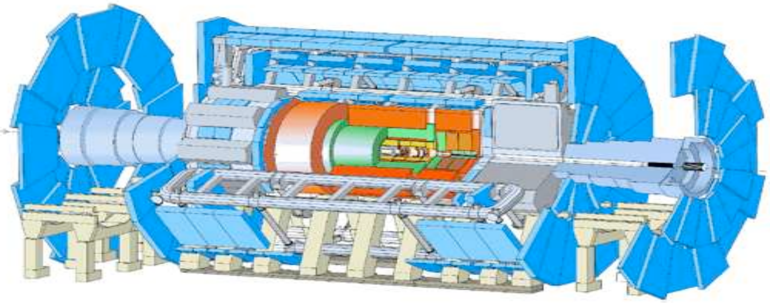
Contents

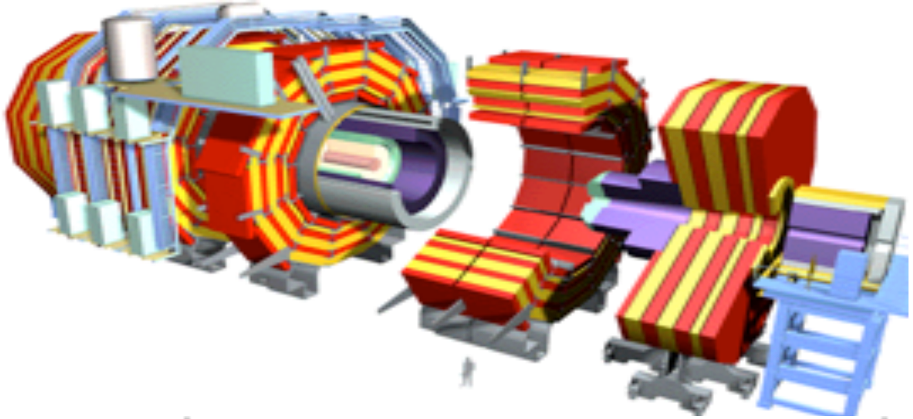
This page is intentionally left blank

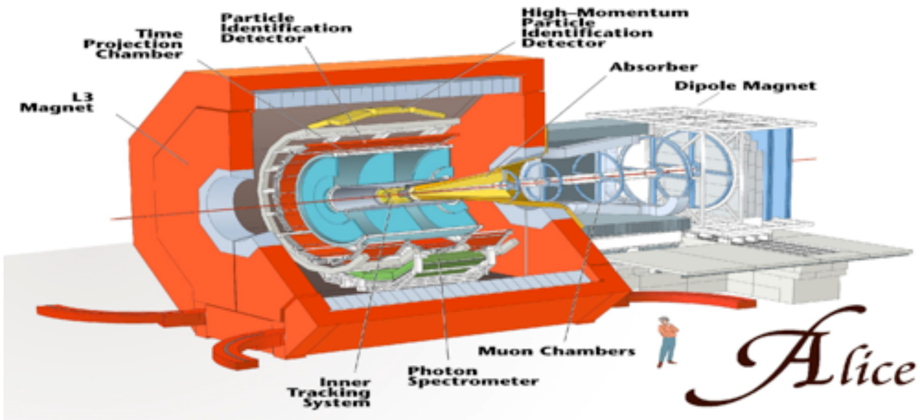
- Ref: www.this-page-intentionally-left-blank.org

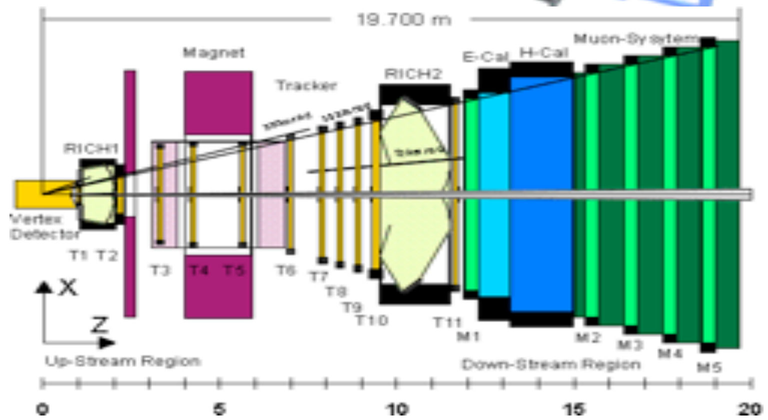
Four LHC Experiments: The Petabyte to Exabyte Challenge

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







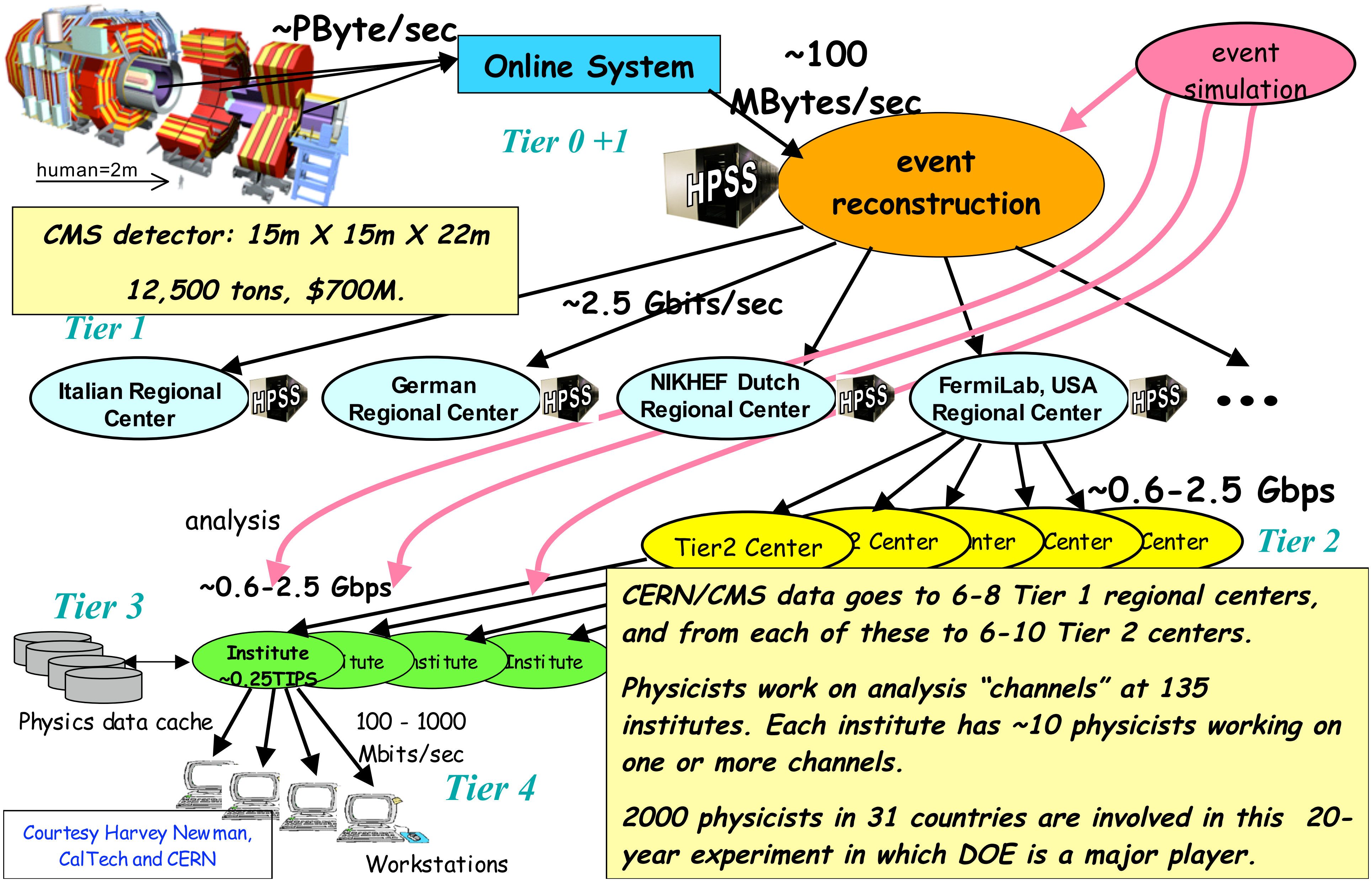


6000+ Physicists & Engineers; 60+ Countries; 250 Institutions

Tens of PB 2008; To 1 EB by ~2015

Hundreds of TFlops To PetaFlops

LHC Data Grid Hierarchy
CMS as example, Atlas is similar



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

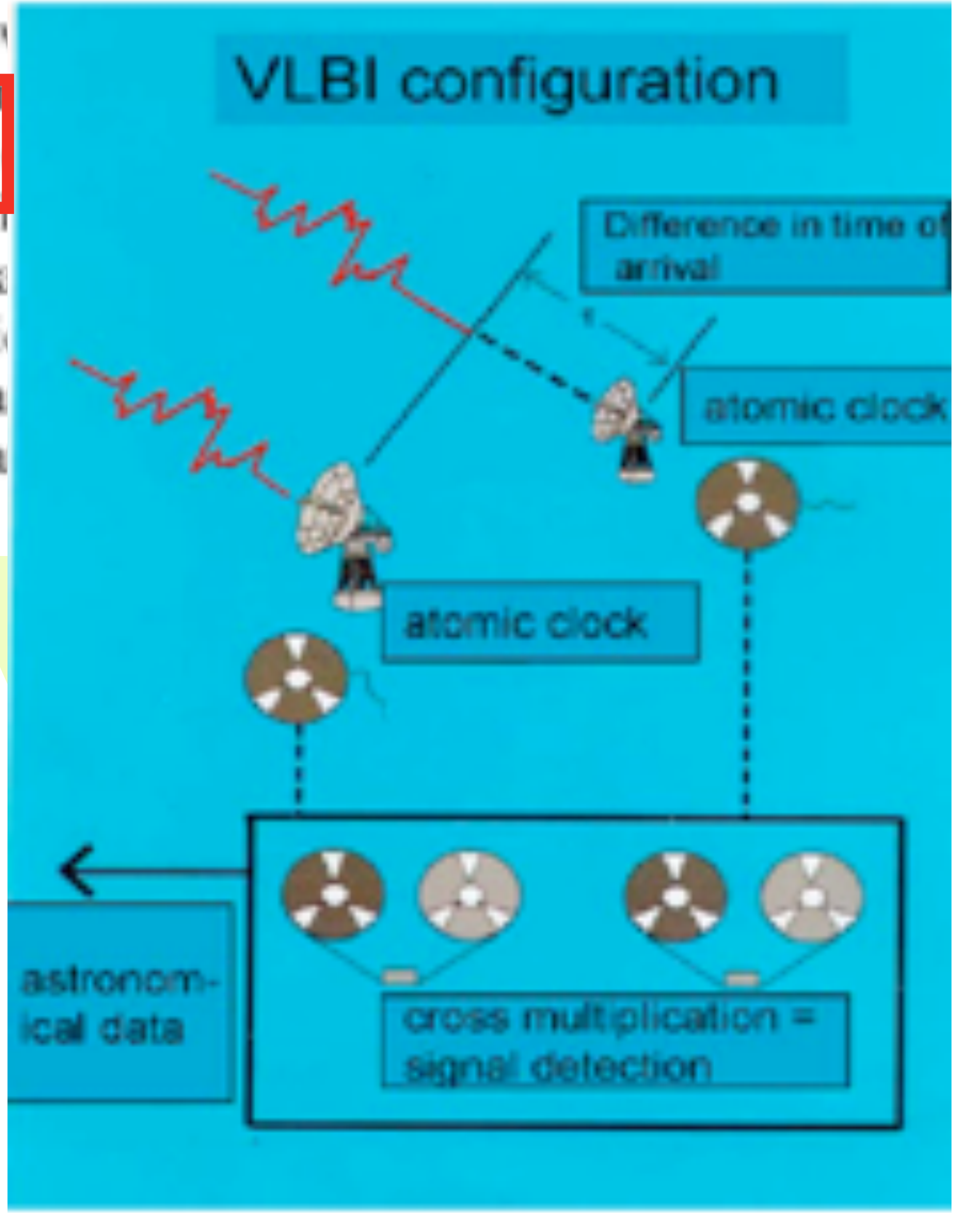
Physicists work on analysis "channels" at 135 institutes. Each institute has ~10 physicists working on one or more channels.

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

VLBI

ger term VLBI is easily capable of generating many Gb of data per
The sensitivity of the VLBI array scales w
(√data-rate) and there is a strong push to i
. Rates of 8Gb/s or more are entirely feasibl
nder development. It is expected that paral
orrelator will remain the most efficient approa
s distributed processing may have an appli
lti-gigabit data streams will aggregate into la
or and the capacity of the final link to the da
tor.

Westerbork Synthesis Radio Telescope - Netherlands



Lambdas as part of instruments

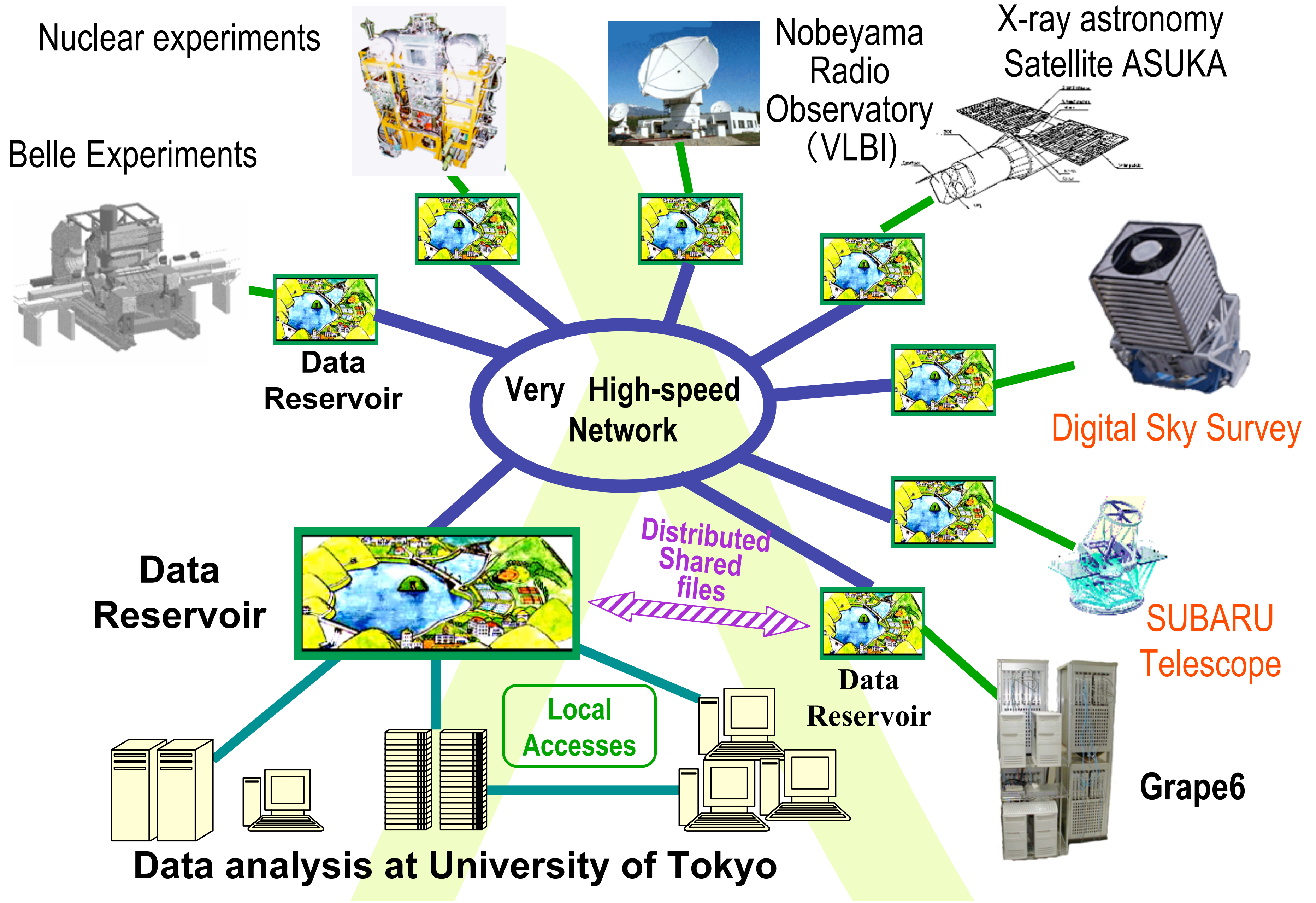
www.lofar.org

37 Tbit/s - 116 Tops/s

http://www.lofar.org/p/systems.htm

http://web.haystack.mit.edu/lofar/technical.html

Data intensive scientific computation through global networks



Co-located interactive 3D visualization

SC2004 “Dead Cat” demo

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

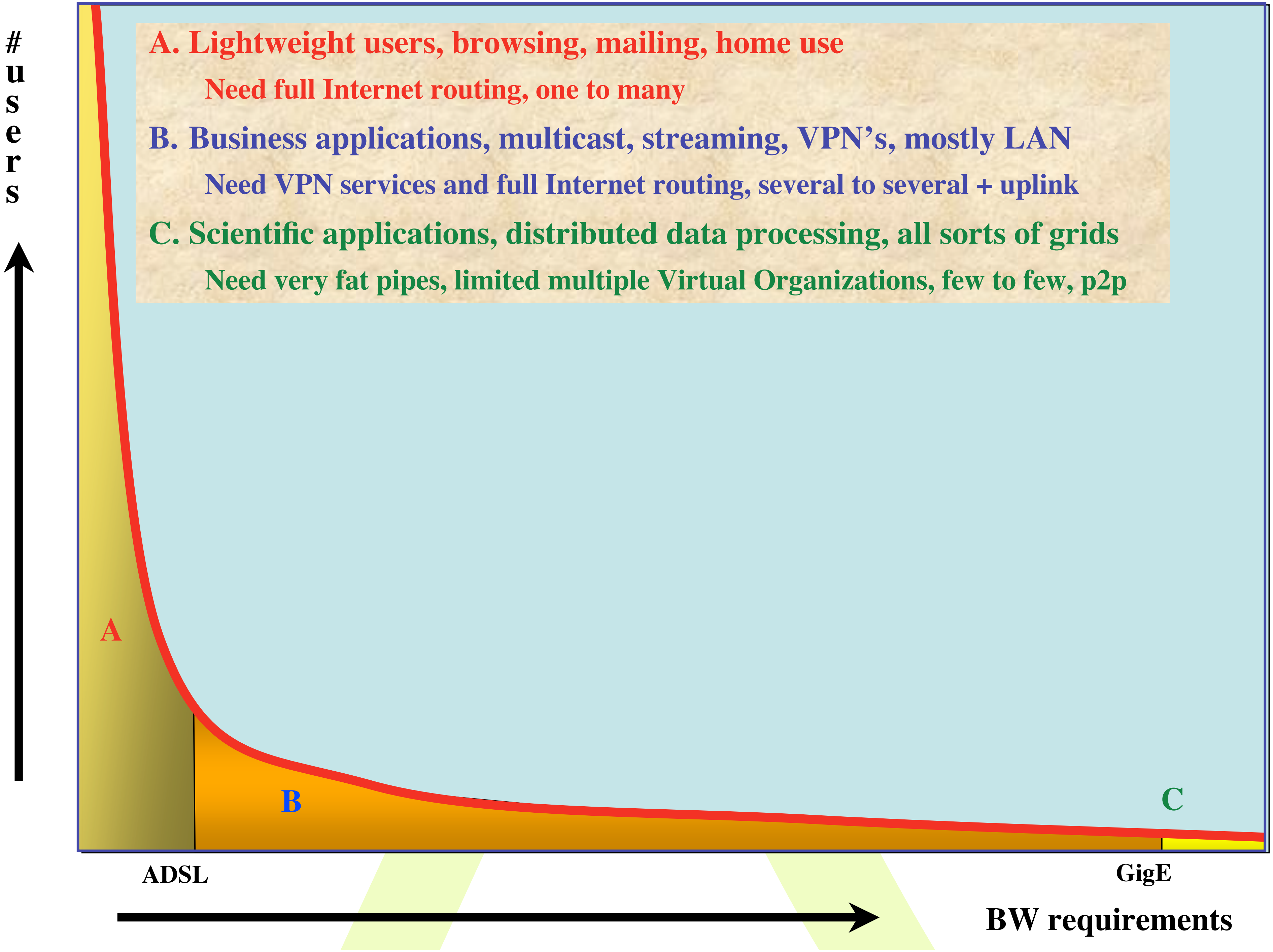
Produced by:
Michael Scarpa
Robert Belleman
Peter Sloot

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

Grids

Showed you:

- **Computational Grids**
  - **HEP and LOFAR analysis requires massive CPU capacity**
- **Data Grids**
  - **Storing and moving HEP, Bio and Health data sets is major challenge**
- **Instrumentation Grids**
  - **Several massive data sources are coming online**
- **Visualization Grids**
  - **Data object (TByte sized) inspection, anywhere, anytime**

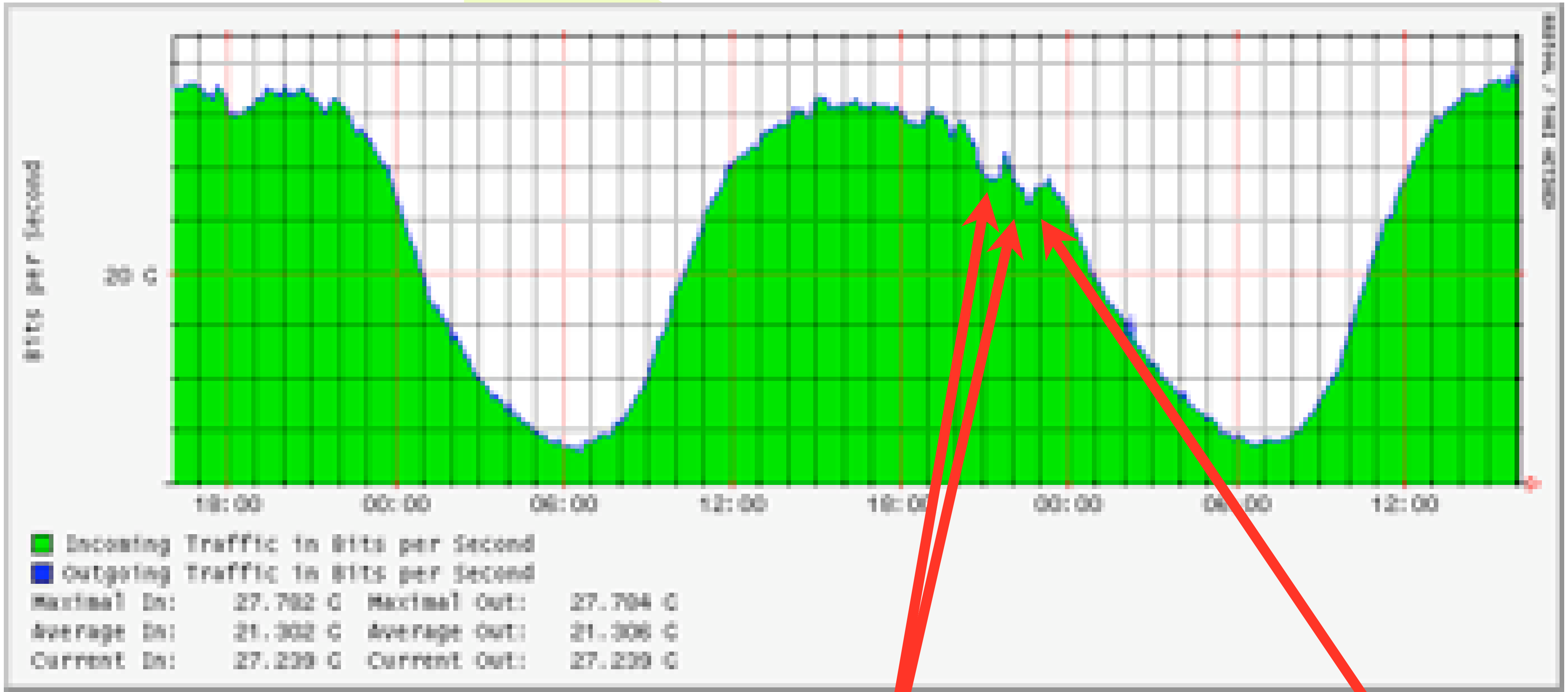


A. Lightweight users, browsing, mailing, home use
Need full Internet routing, one to many
B. Business applications, multicast, streaming, VPN's, mostly LAN
Need VPN services and full Internet routing, several to several + uplink
C. Scientific applications, distributed data processing, all sorts of grids
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



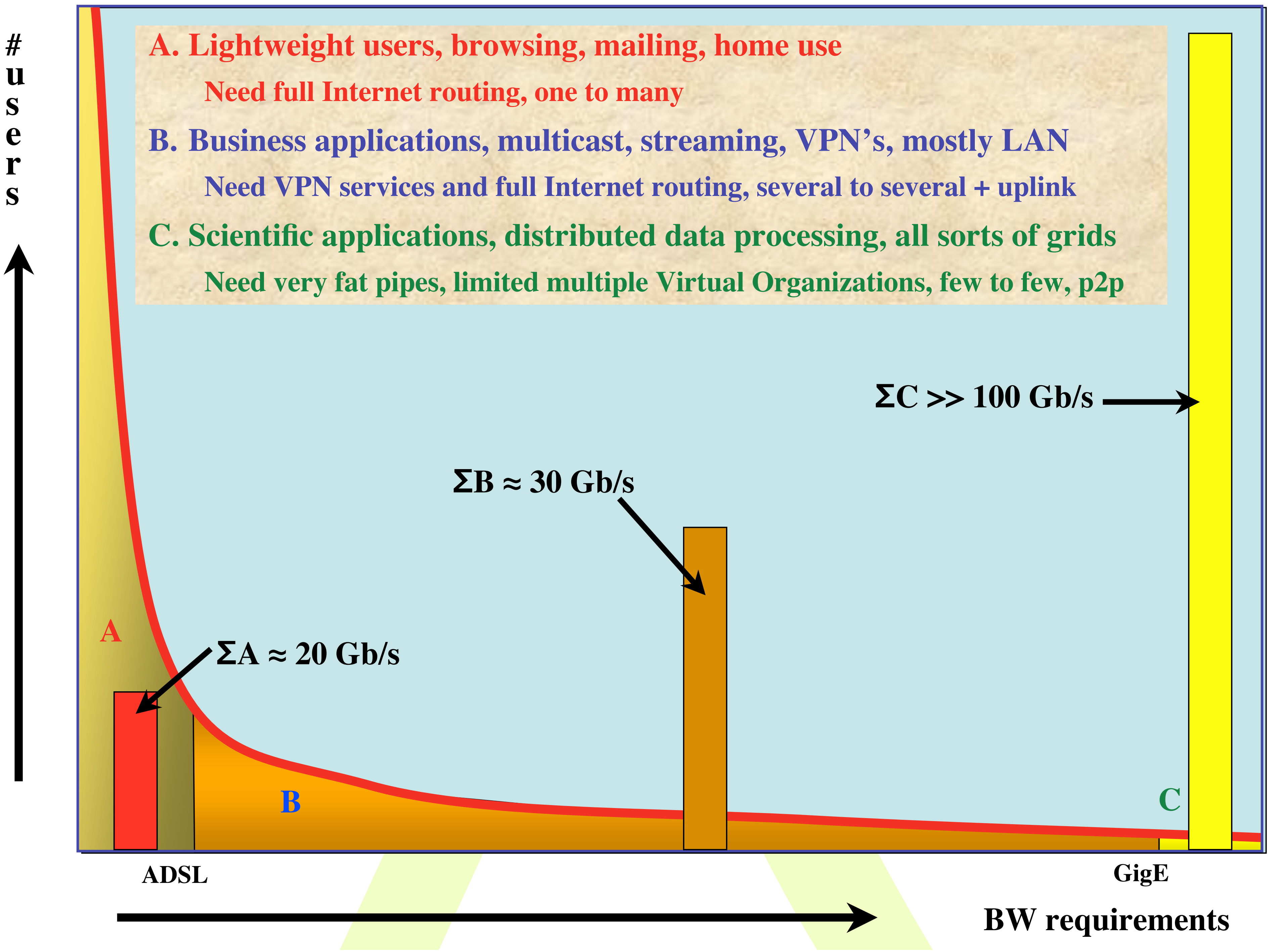
June 19th 2004

Lost :-(

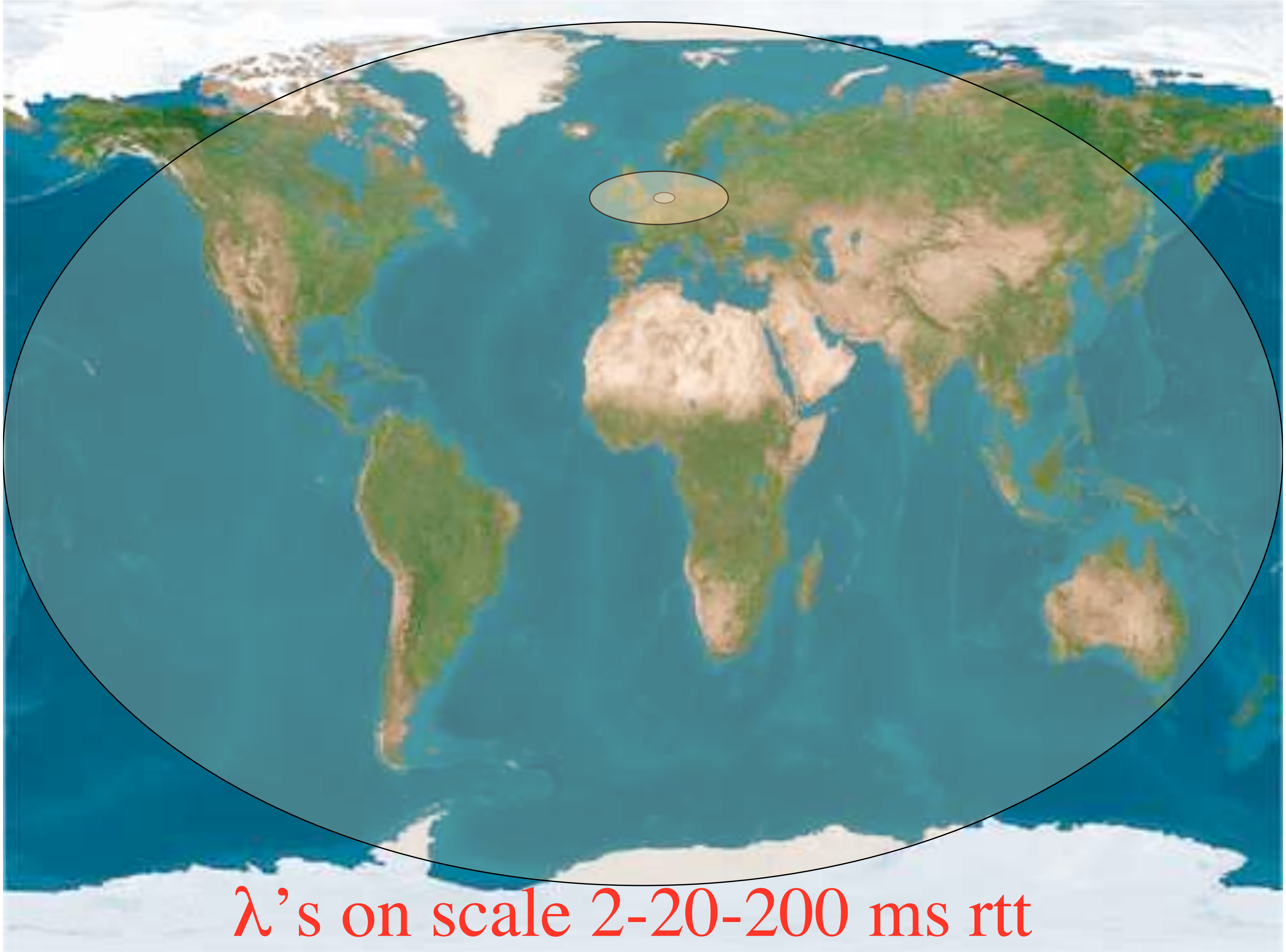
European championship football **Holland -- Czech Republic**

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**



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
ΣC >> 100 Gb/s
ΣB ≈ 30 Gb/s
ΣA ≈ 20 Gb/s
A
B
C
ADSL
GigE
BW requirements

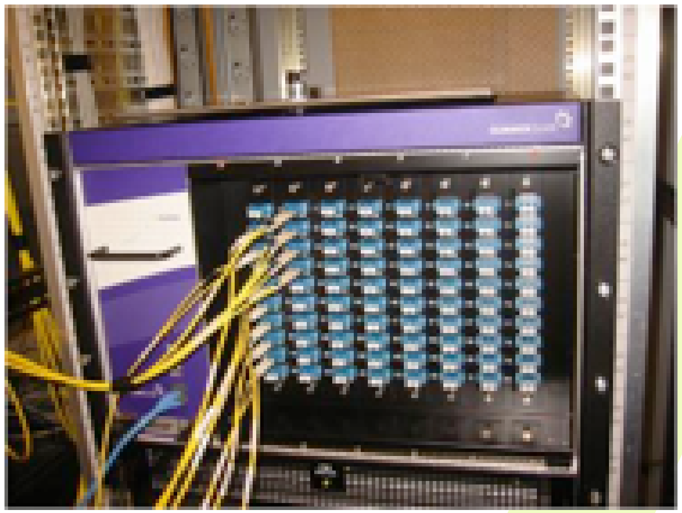


λ’s on scale 2-20-200 ms rtt

So what?

- 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



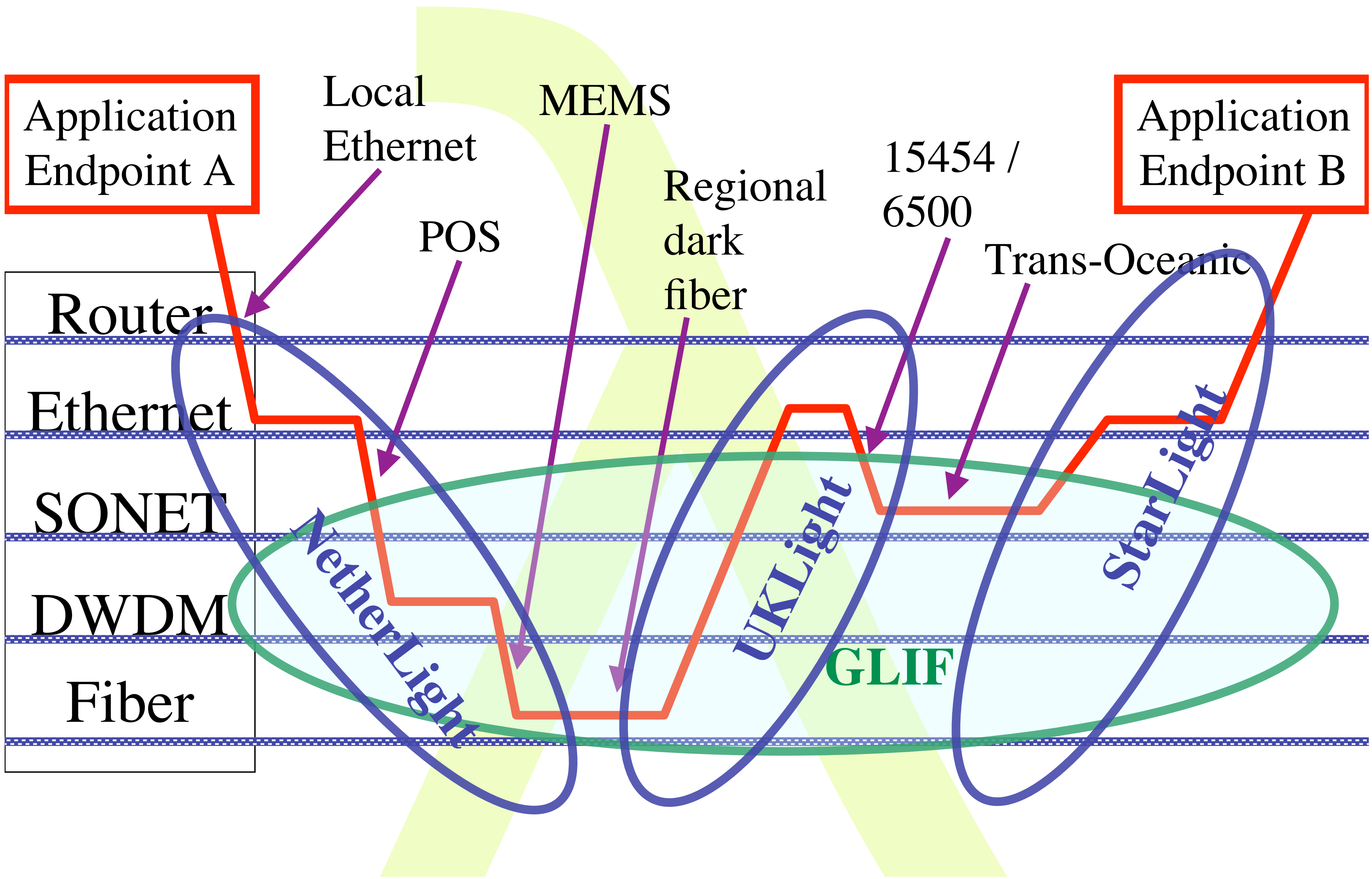
L2 - 10-20 k$/port

L3 - 100-500 k$/port

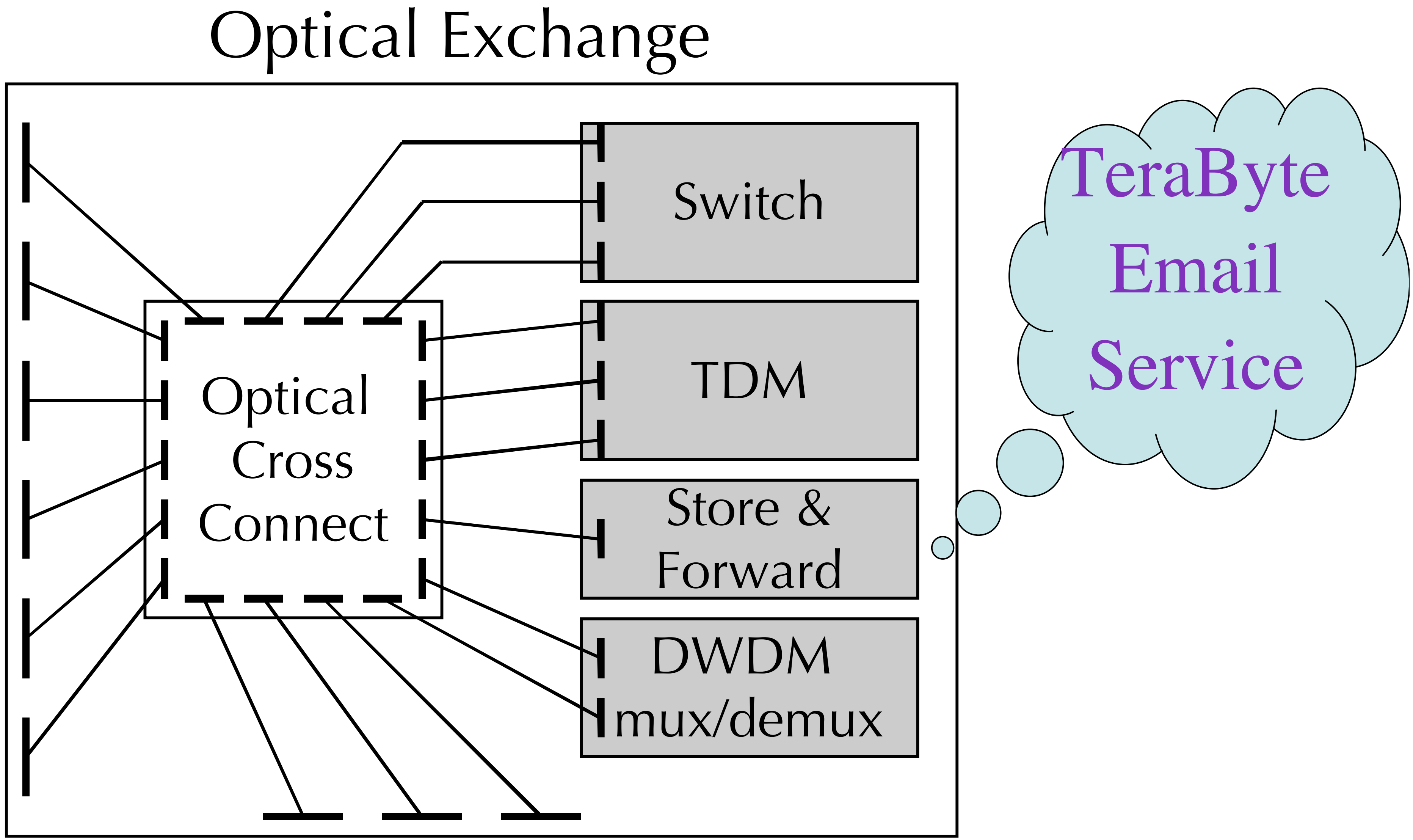
Services

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



Service Matrix

To From	WDM (multiple λ)	Single λ, any bitstream	SONET/ SDH	1 Gb/s Ethernet	LAN PHY Ethernet	WAN PHY Ethernet	VLAN tagged Ethernet	IP over Ethernet
WDM (multiple λ)	cross-connect multicast, regenerate, multicast	WDM demux	WDM demux*	WDM demux *	WDM demux *	WDM demux *	WDM demux *	WDM demux *
Single λ, any bitstream	WDM mux	cross-connect multicast, regenerate, multicast	N/A *	N/A *	N/A *	N/A *	N/A *	N/A *
SONET/SDH	WDM mux	N/A *	SONET switch, +	TDM demux *	TDM demux[6]	SONET switch	TDM demux *	TDM demux *
1 Gb/s Ethernet	WDM mux	N/A *	TDM mux	aggregate, Ethernet conversion +	aggregate, eth. convert	aggregate, Ethernet conversion	aggregate, VLAN encap	L3 entry *
LAN PHY Ethernet	WDM mux	N/A*	TDM mux[6]	aggregate, Ethernet conversion	aggregate, Ethernet conversion +	Ethernet conversion	aggregate, VLAN encap	L3 entry *
WAN PHY Ethernet	WDM mux	N/A *	SONET switch	aggregate, Ethernet conversion	Ethernet conversion	aggregate, Ethernet conversion +	aggregate, VLAN encap	L3 entry *
VLAN tagged Ethernet	WDM mux	N/A *	TDM mux	aggregate, VLAN decap	aggregate, VLAN decap	aggregate, VLAN decap	Aggregate, VLAN decap & encap +	N/A
IP over Ethernet	WDM mux	N/A *	TDM mux	L3 exit *	L3 exit *	L3 exit *	N/A	Store & forward, L3 entry/exit+

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

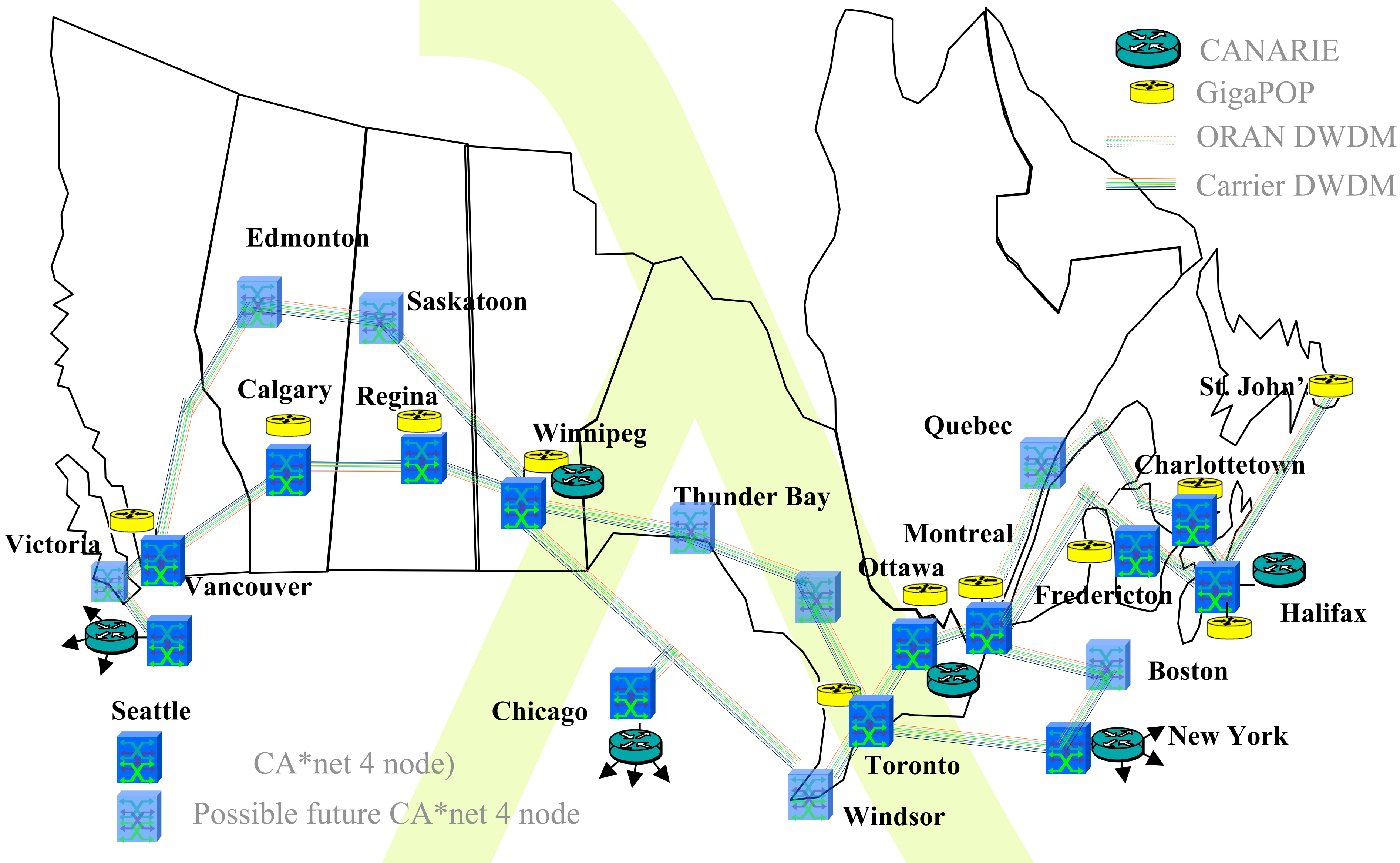

SURFnet on Lambda inspection in
Science Park Amsterdam :-)

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

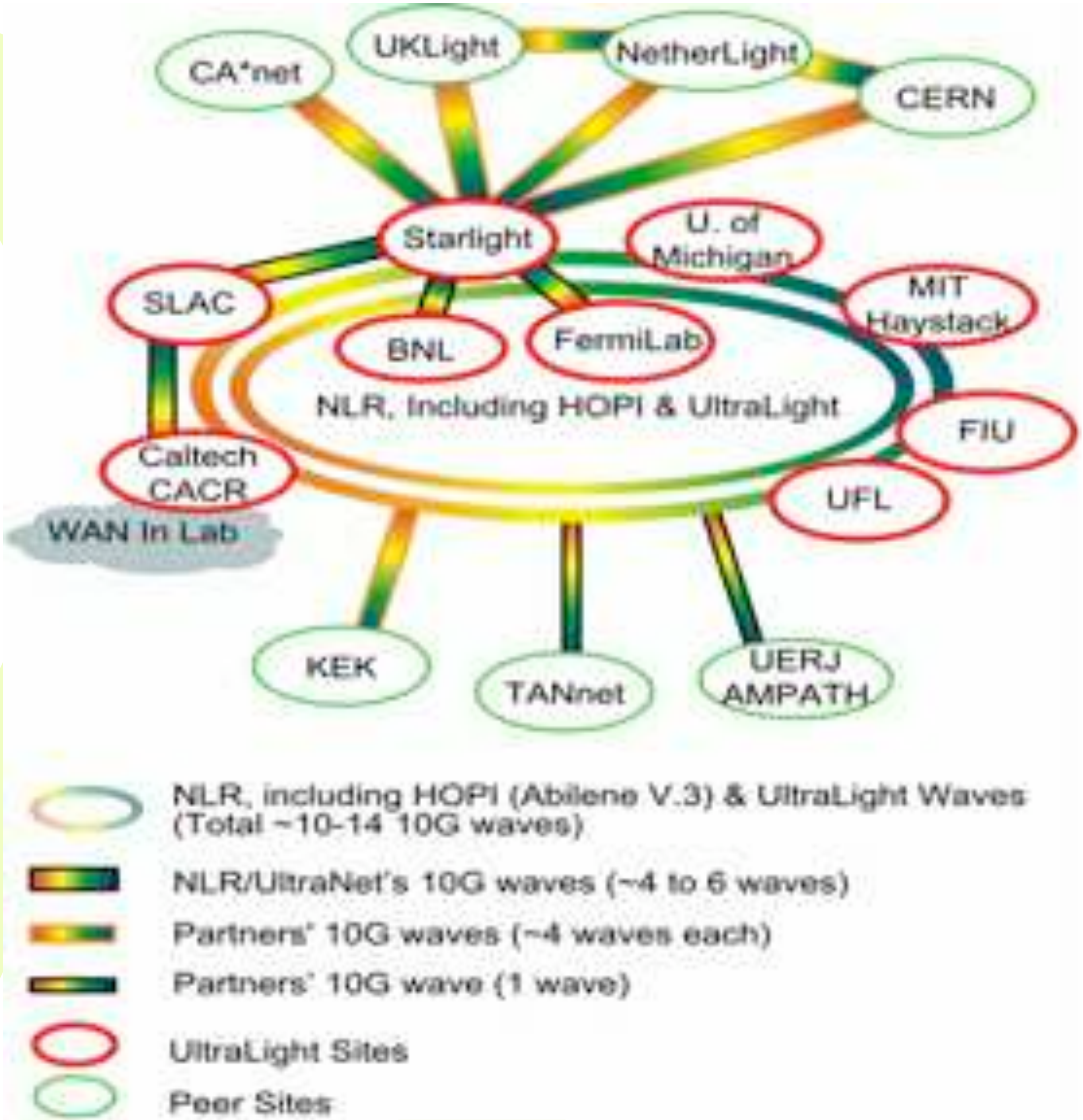
NLR Condominium lambda network

CA*net 4 Architecture

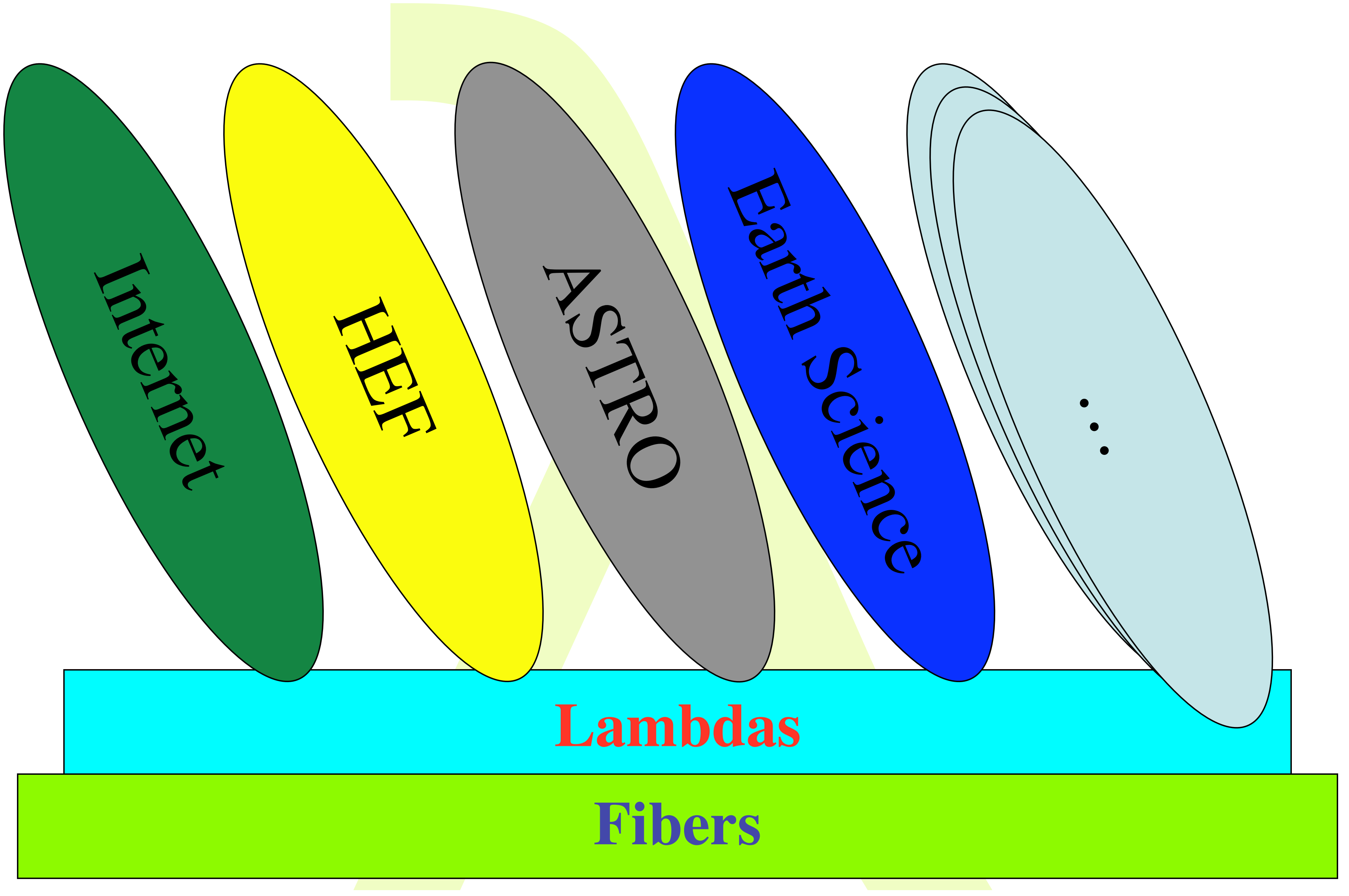


UltraLight Network: **PHASE III**

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



Discipline Networks



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

History of GLIF

- Brainstorming in Antalya at Terena conf. 2001
- 1th meeting at Terena offices 11-12 sep 2001
  - On invitation only (15) + public part
  - Thinking, SURFnet test lambda Starlight-Netherlight
- 2nd meeting appended to iGrid 2002 in Amsterdam
  - Public part in track, on invitation only day (22)
  - Core testbed brainstorming, idea checks, seeds for Translight
- 3th meeting Reykjavik, hosted by NORDUnet 2003
  - Grid/Lambda track in conference + this meeting (35!)
  - Brainstorm applications and showcases
  - Technology roadmap
  - GLIF established -> www.glif.is
- 4th at Nottingham 3 Sept 2004 hosted by UKERNA colocated UK-eScience
  - preparatory afternoon on 2 September
  - 60 participants
  - Attendance from China, Japan, Netherlands, Switzerland, US, UK, Taiwan, Australia, Tsjech, Korea, Canada, Ireland, Russia, Belgium, Denmark
  - Meeting of GOV, TEC and APP groups

GLIF Q3 2004

Visualization courtesy of
Bob Patterson, NCSA.

Research on Networks (CdL)

GigaPort

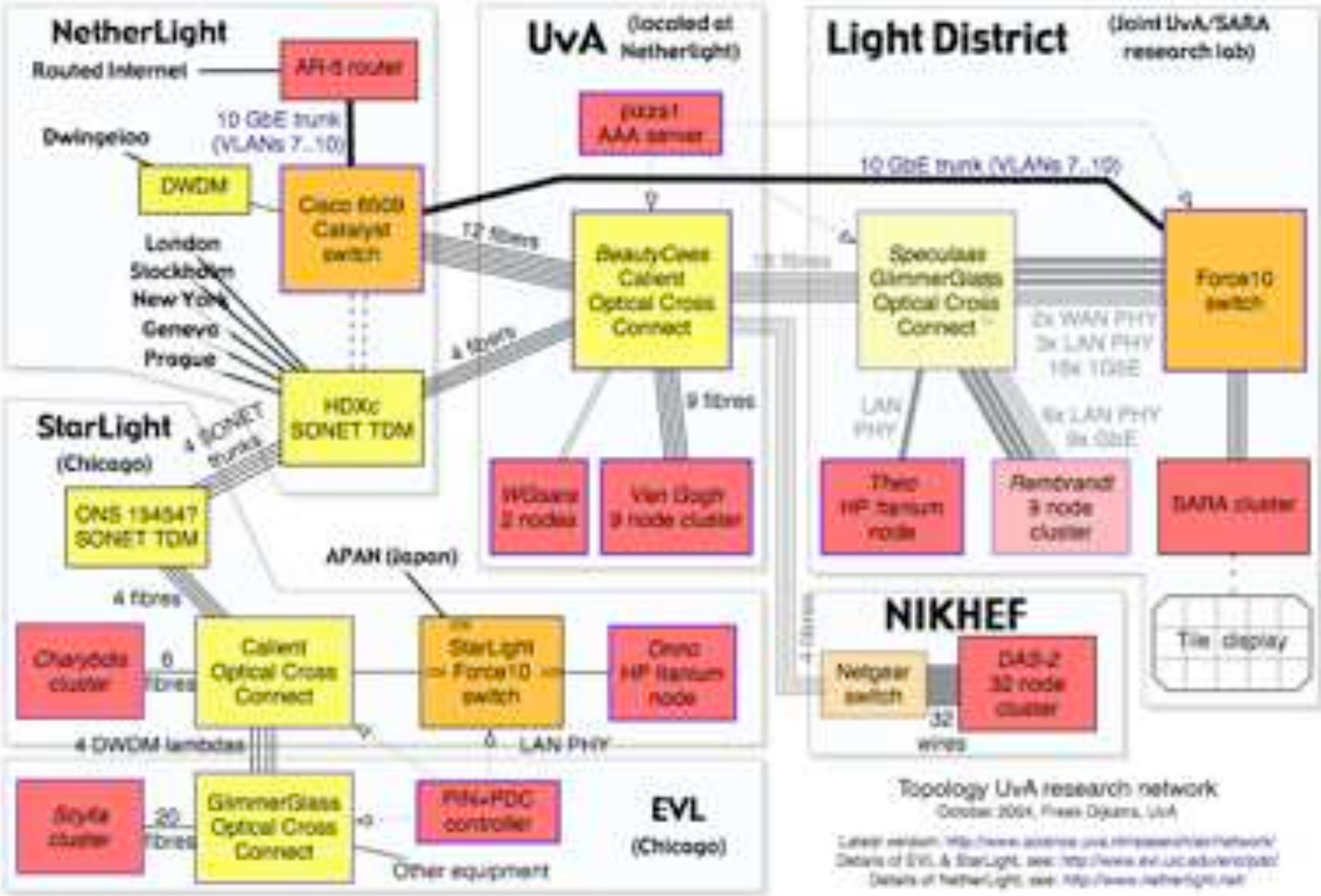
- **Optical Networking:**
  - What innovation in architectural models, components, control and light path provisioning are needed to integrate dynamically configurable optical transport networks and traditional IP networks to a generic data transport platform that provides end-to-end IP connectivity 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

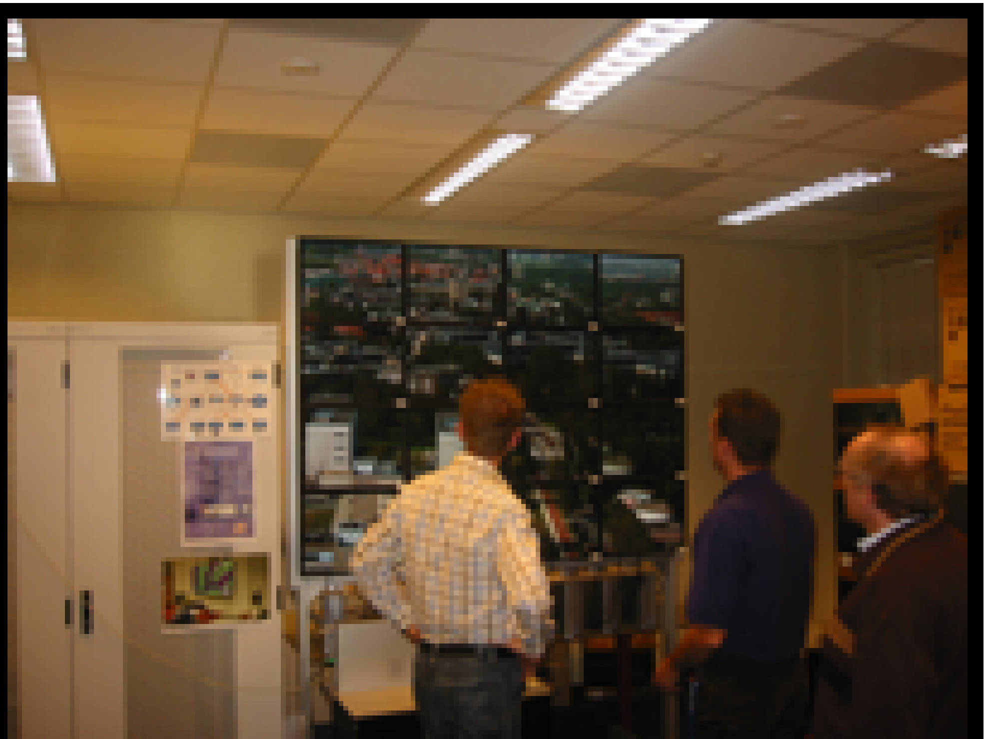
Research topics AIR @ UvA

- Optical networking architectures and models for usage
- Transport protocols for massive amounts of data
- Authorization of complex resources in multiple domains
- Embedding in Grid environments

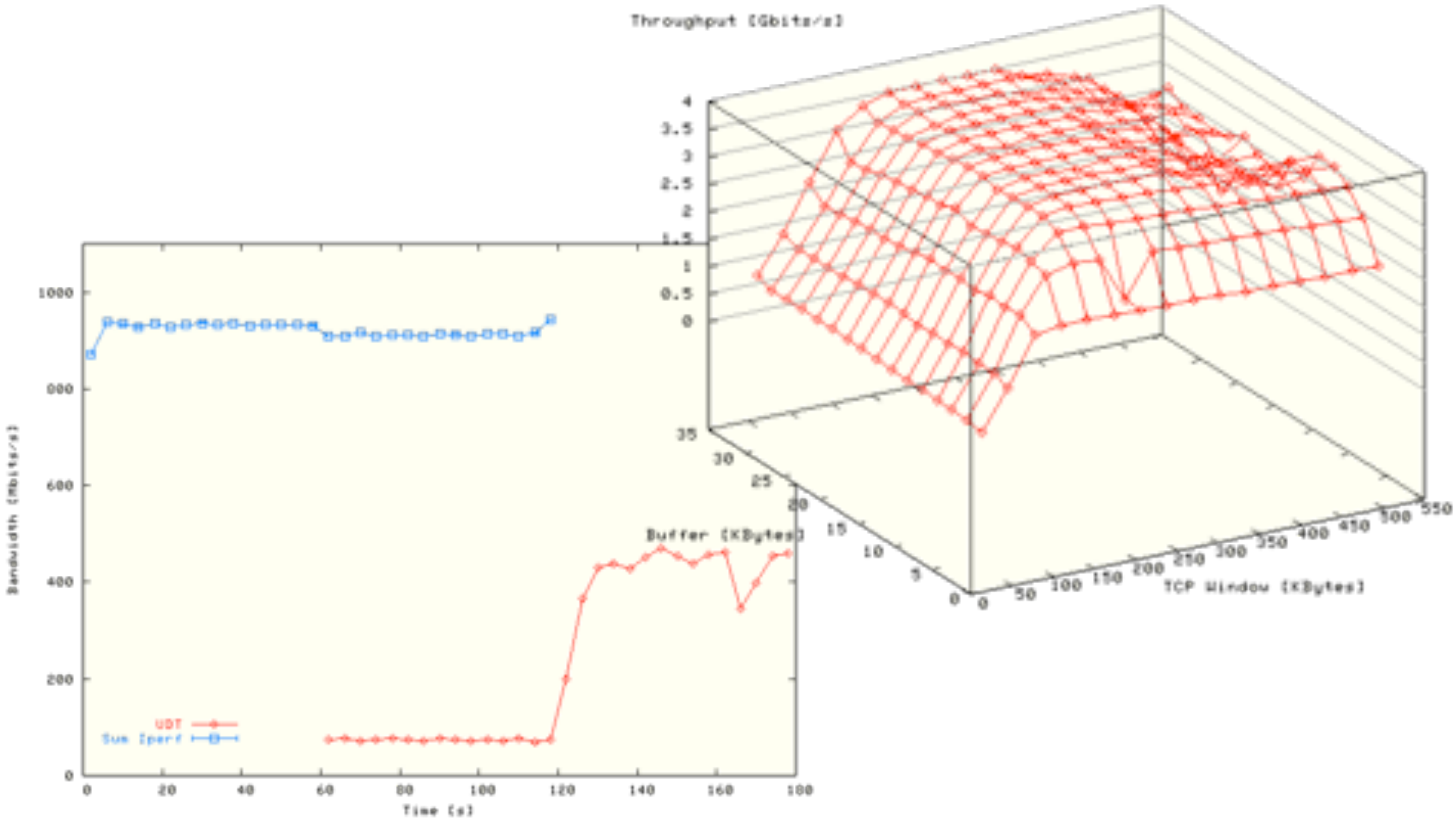
LightHouse



NetherLight
Routed Internet
Dwingeloo
DWDM
10 GbE trunk (VLANs 7..10)
Cisco 6509 Catalyst switch
London
Stockholm
New York
Geneva
Prague
HDXc SONET TDM
StarLight (Chicago)
4 SONET trunks
ONS 15454? SONET TDM
4 fibres
UvA (located at Netherlight)
pizza1 AAA server
12 fibres
4 fibres
9 fibres
BeautyCees Calient Optical Cross Connect
Van Gogh 9 node cluster
Light District (Joint UvA/SARA research lab)
10 GbE trunk (VLANs 7..10)
Speculaas GlimmerGlass Optical Cross Connect
Force10 switch
2x WAN PHY
3x LAN PHY
15x 1GbE
LAN PHY
6x LAN PHY
9x GbE
Rembrandt 9 node cluster
SARA cluster
APAN (Japan)
Calient Optical Cross Connect
StarLight Force10 switch
HP Itanium node
6 fibres
4 DWDM lambdas
LAN PHY
GlimmerGlass Optical Cross Connect
20 fibres
Other equipment
EVL (Chicago)
NIKHEF
Netgear switch
DAS-2 32 node cluster
32 wires
Tile display
Topology UvA research network

Example Measurements



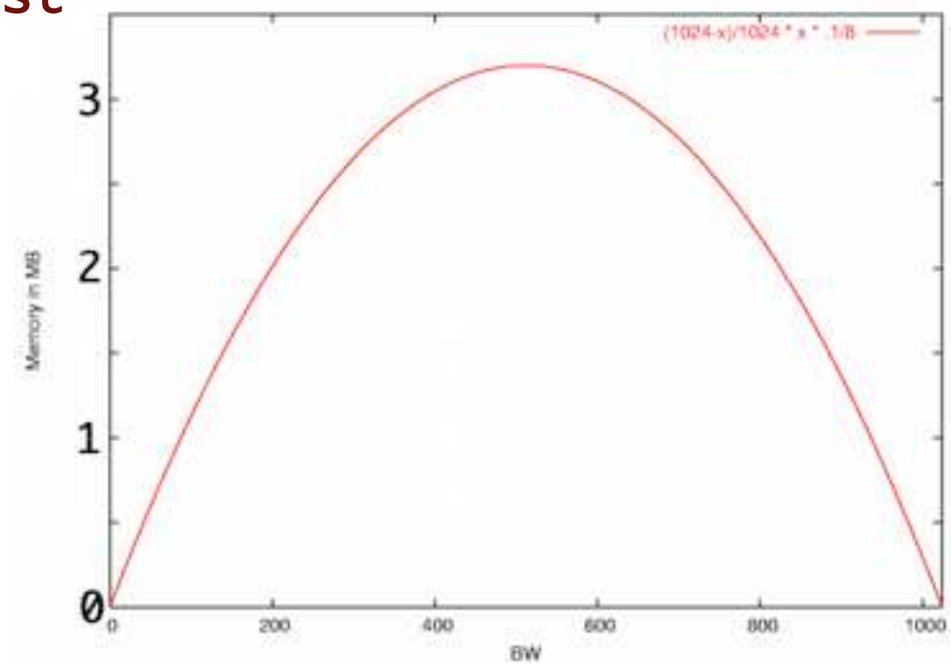
Layer - 2 requirements from 3/4

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

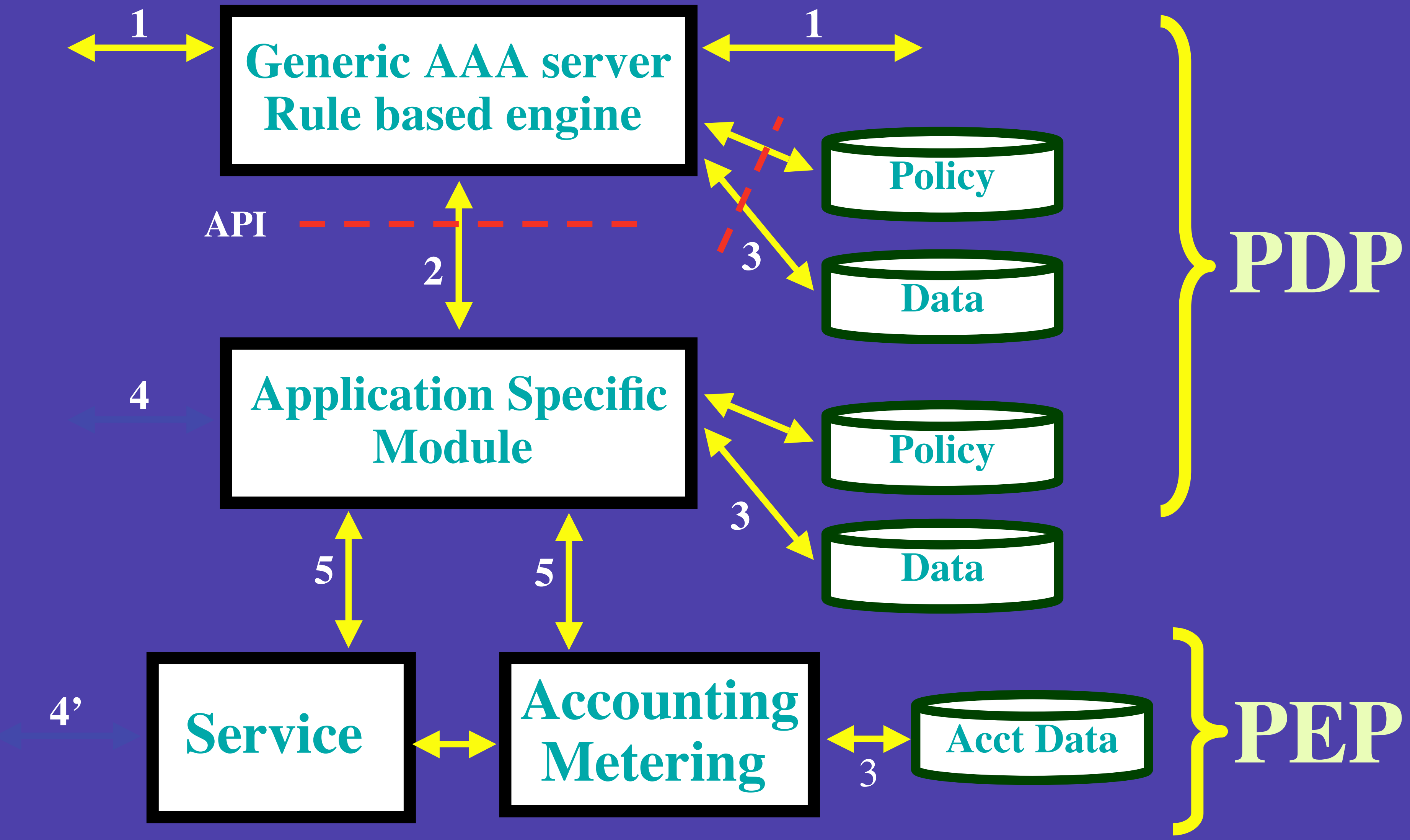
```
Window = BandWidth * RTT    &    BW == slow

                            fast - slow
Memory-at-bottleneck = ------------ * slow * RTT
                               fast
```

So pick from menu:

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

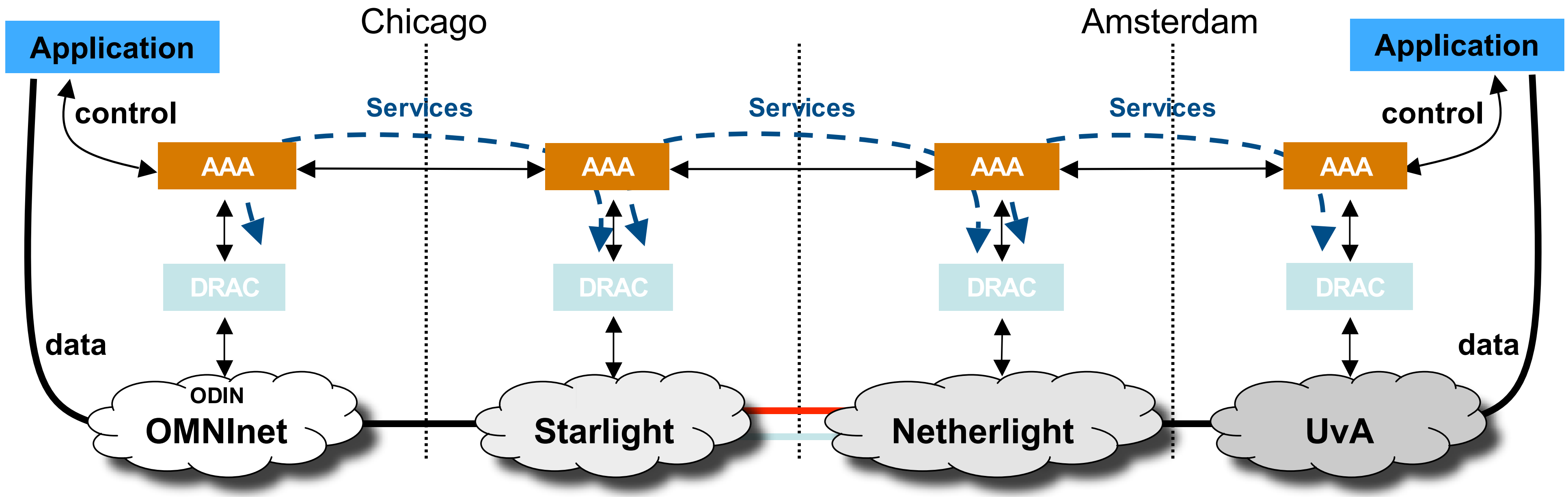


Starting point



RFC 2903 - 2906 , 3334 , policy draft

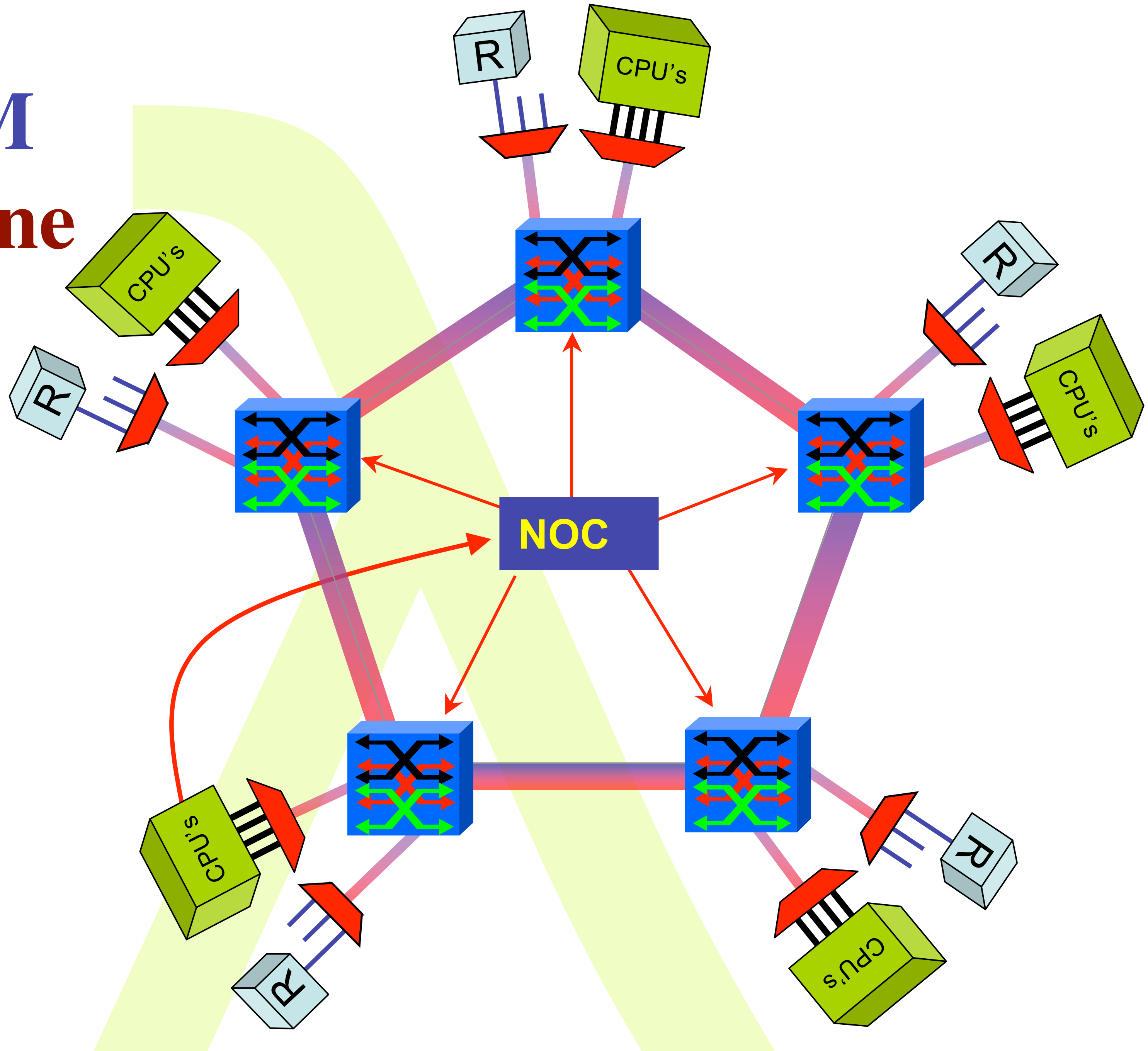
SC2004 CONTROL CHALLENGE



- 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

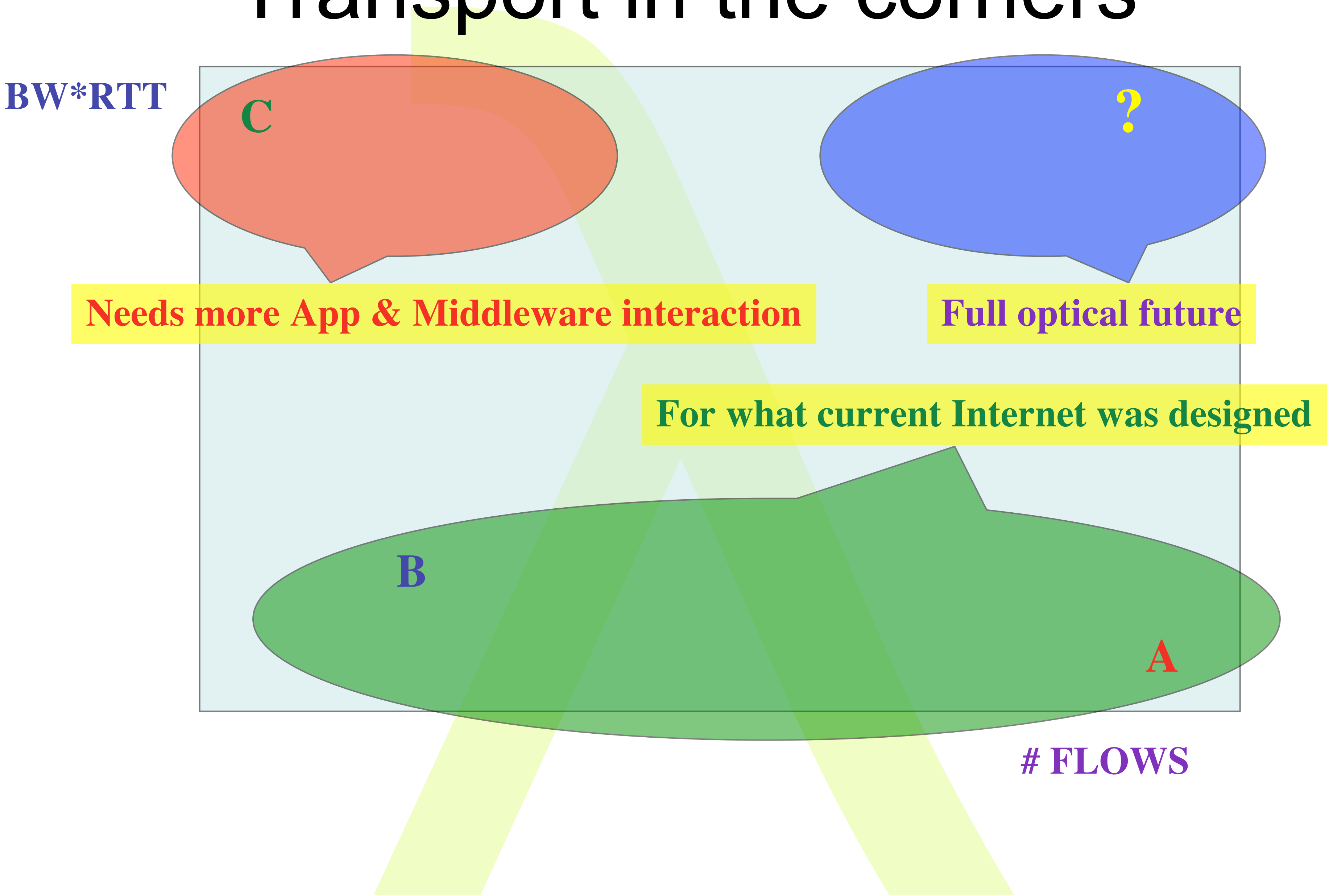

GOT
CONTROL
TOO?

DAS3 DWDM backplane



UvA-VLE
UvA-MM
VU
ULeiden
TUDelft

Transport in the corners



Revisiting the truck of tapes (one but last)

Consider one fiber

- **Current technology allows 320 λ in one of the frequency bands**
- **Each λ has a bandwidth of 40 Gbit/s**
- **Transport: 320 * $40*10^9$ / 8 = 1600 GByte/sec**

- **Take a 10 metric ton truck**
  - **One tape contains 50 Gbyte, weights 100 gr**
  - **Truck contains (10000 / 0.1) * 50 Gbyte = 5 PByte**

- **Truck / fiber = 5 PByte / 1600 GByte/sec = 3125 s ≈ one hour**
- **For distances further away than a truck drives in one hour (50 km) minus loading and handling 100000 tapes the fiber wins!!!**


Not quite THE END
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.
RESERVED
Cees
de Laat
Partially complete list:
Caas
Chase
Cess
Kess
Case