Lambda-Grid developments

www.science.uva.nl/~delaat

Cees de Laat

GigaPort EU

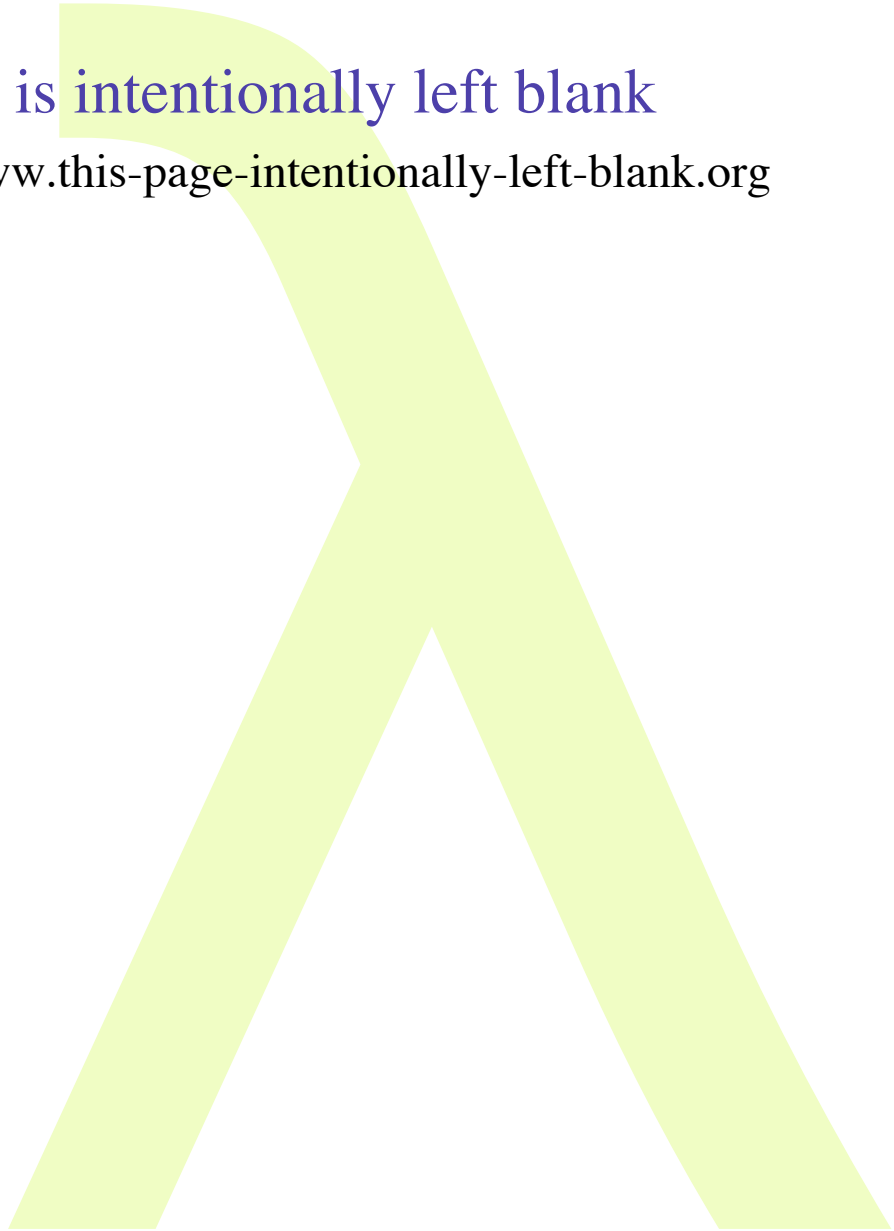
University of Amsterdam

SARA

NCF

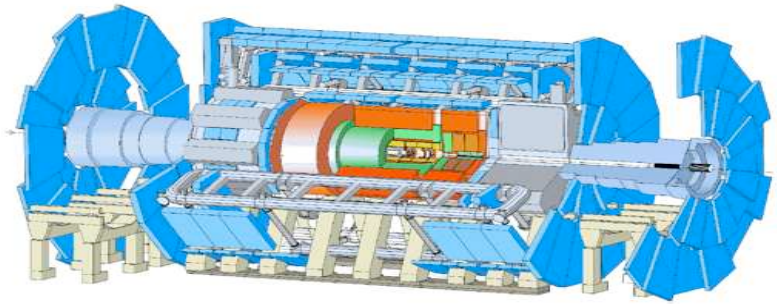
Contents of this talk

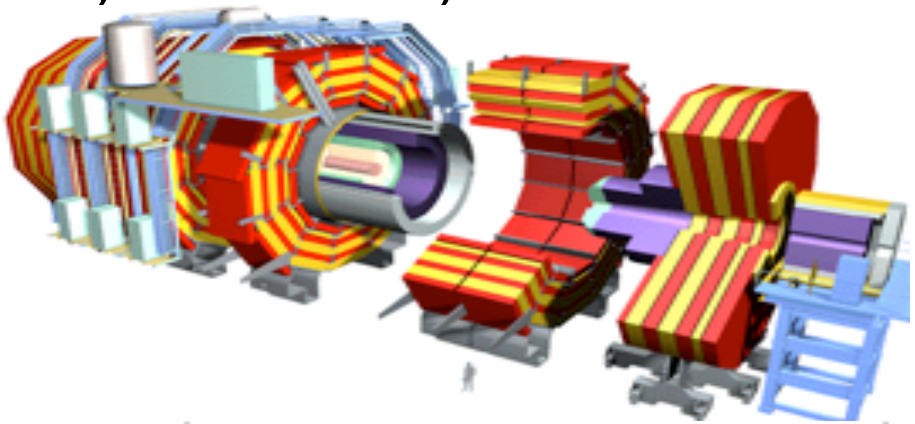
- This page is intentionally left blank
  - Ref: www.this-page-intentionally-left-blank.org

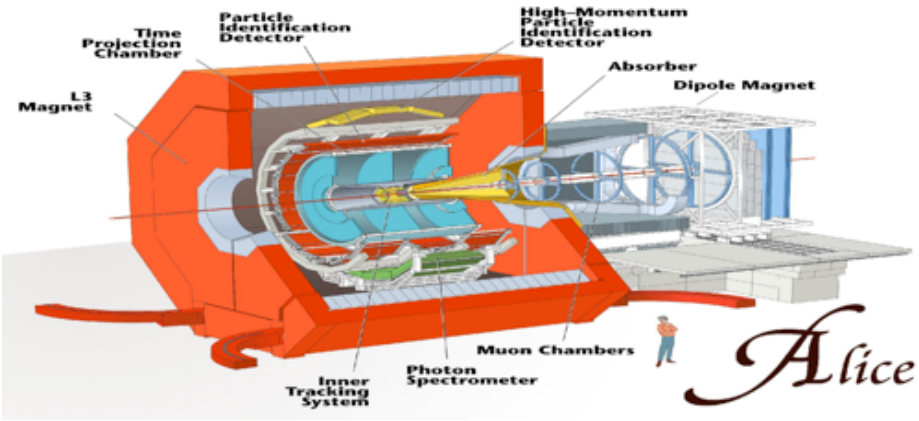


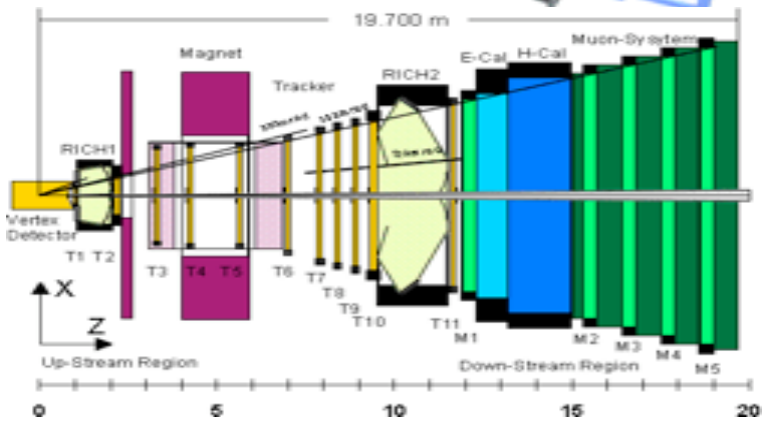
Four LHC Experiments: The Petabyte to Exabyte Challenge

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









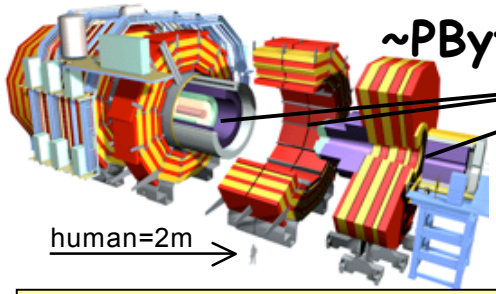
6000+ Physicists & Engineers; 60+ Countries; 250 Institutions

Tens of PB 2008; To 1 EB by ~2015

Hundreds of TFlops To PetaFlops

LHC Data Grid Hierarchy

CMS as example, Atlas is similar



~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.

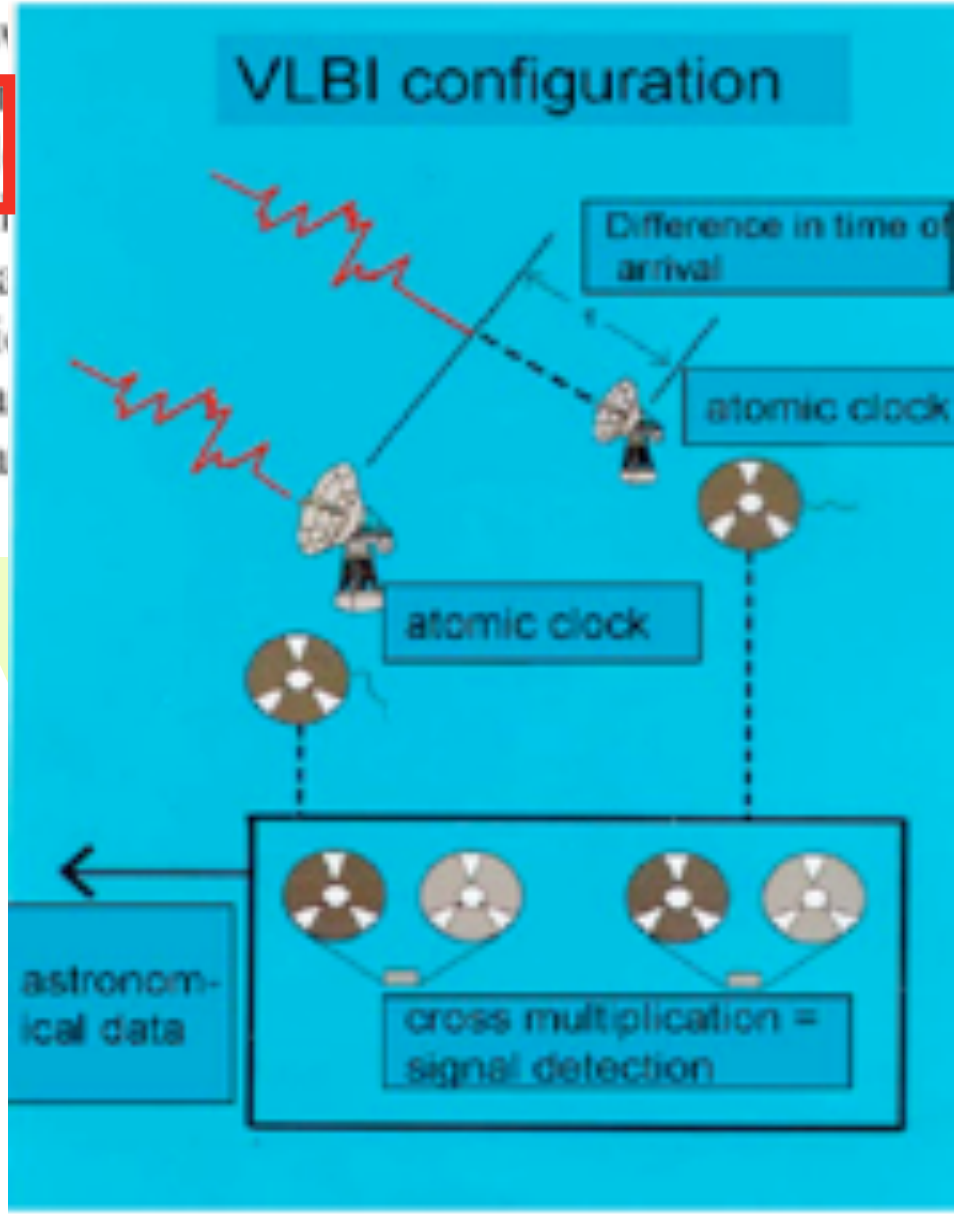
Physicists work on analysis "channels" at 135 institutes. Each institute has ~10 physicists working on one or more channels.

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

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
. Rates of 8Gb/s or more are entirely feasibl
der 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

GigaPort

LOFAR

www.lofar.org

1 - 45 Tbit/s,

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

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

SURFnet

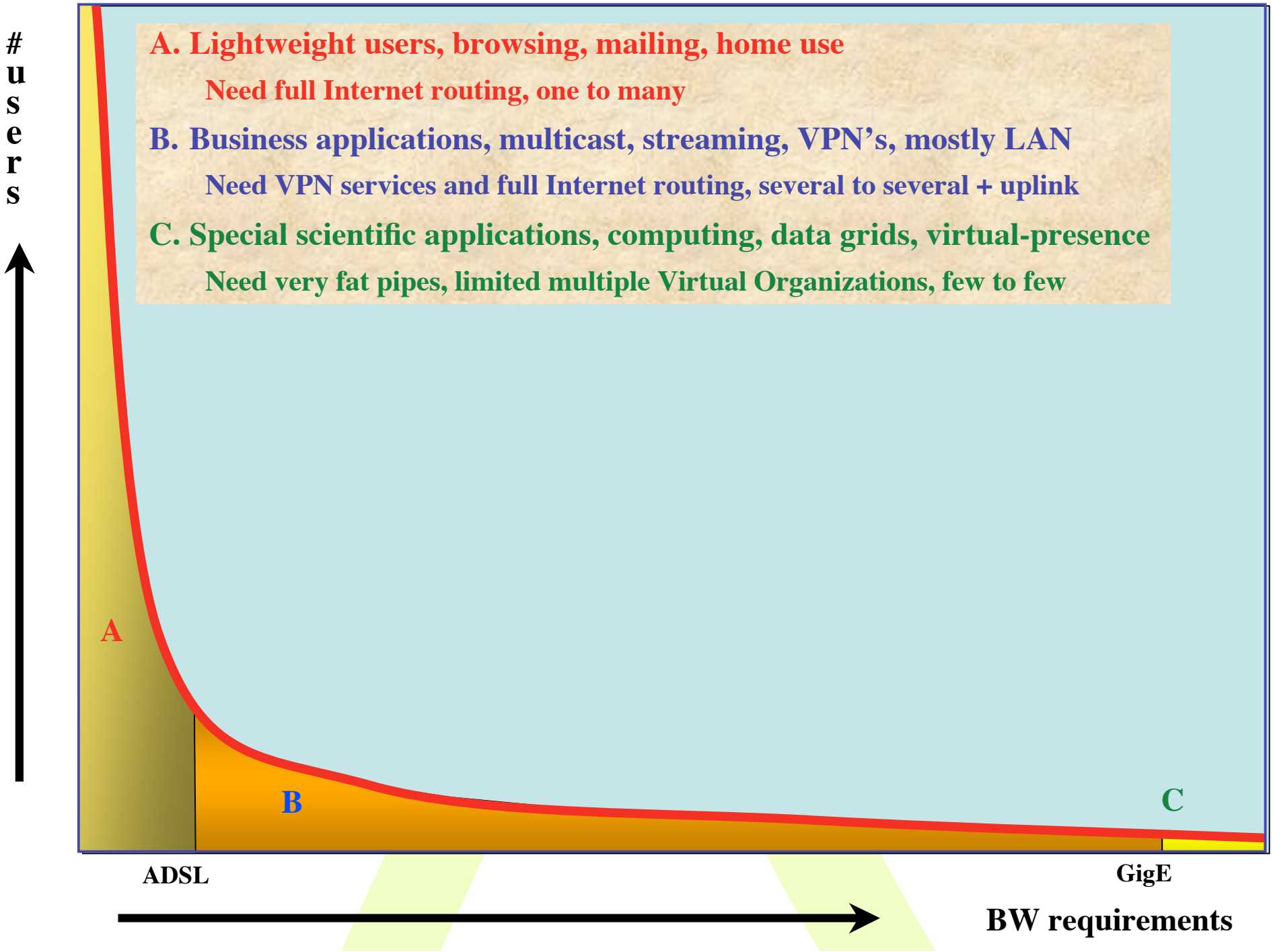
Grids

Showed you:

- Computational Grids
  - HEP and LOFAR analysis requires massive CPU capacity
- Data Grid
  - 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

UCSD
UCIrvine

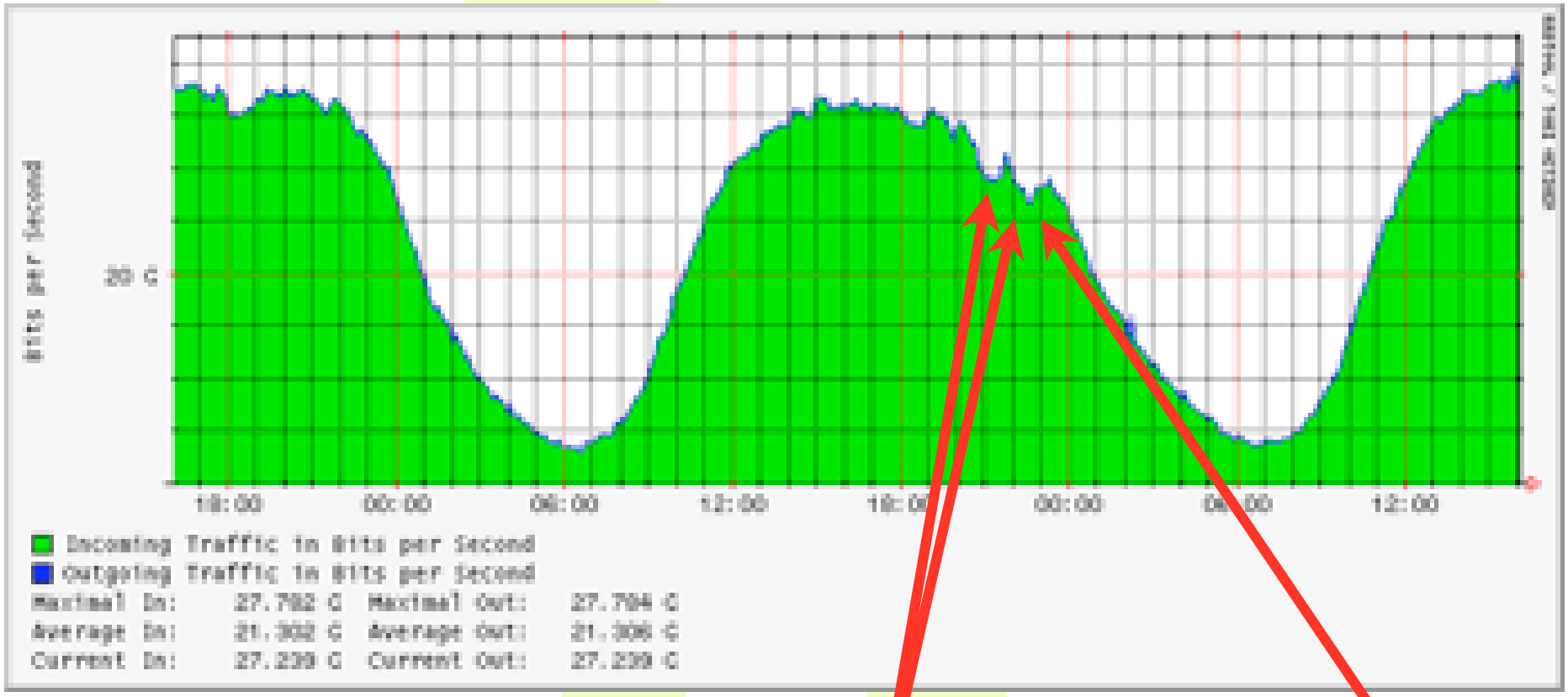
it²


#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. Special scientific applications, computing, data grids, virtual-presence
Need very fat pipes, limited multiple Virtual Organizations, few to few
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 :-(

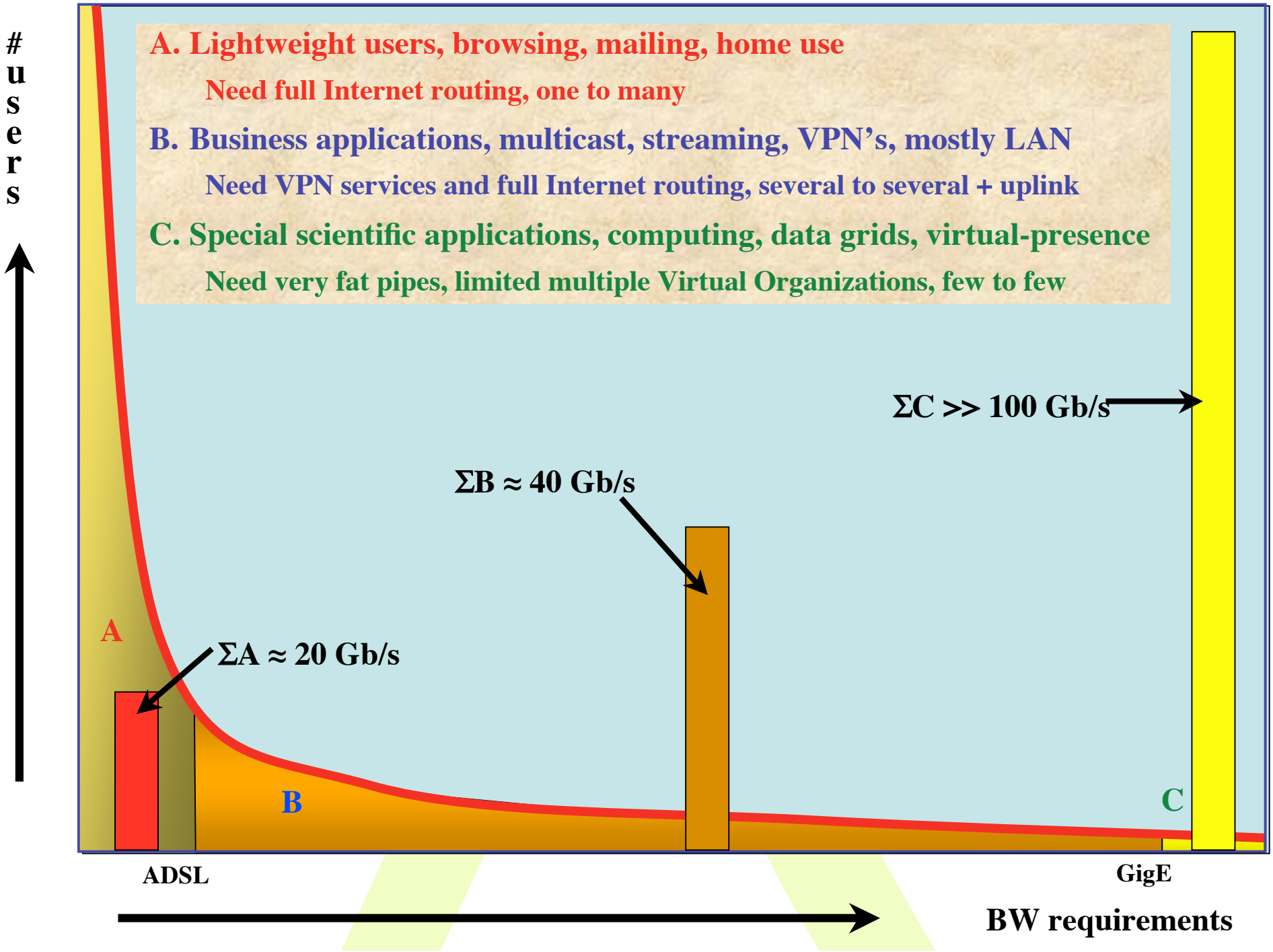
European championship football **Holland -- Czech Republic**

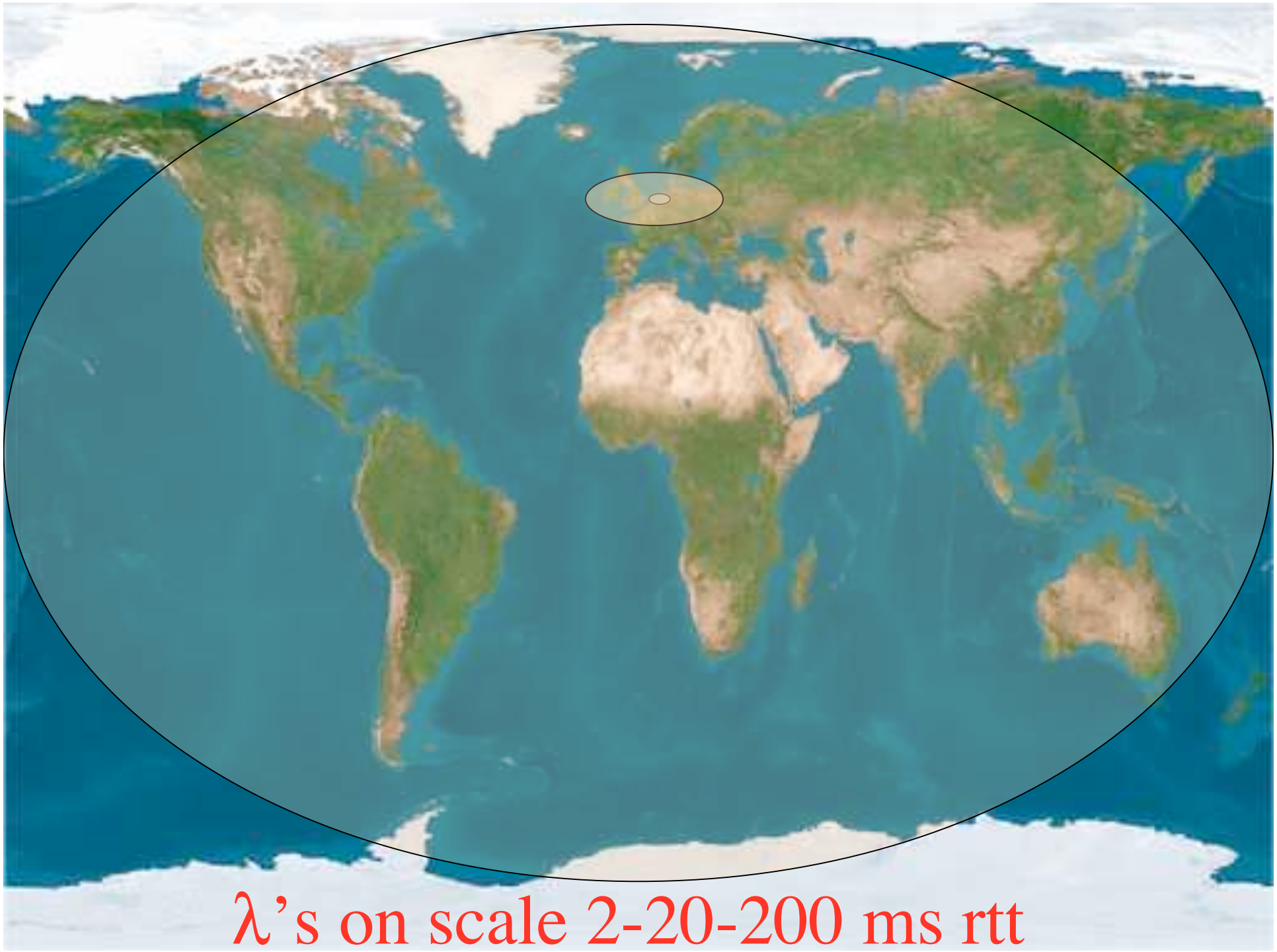
The Dutch Situation

- **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**
  - **SURFnet has 10 Gb/s to about 12 institutes and 0.1 to 1 Gb/s to 180 customers, estimate same for industry (overestimation) ==> 20-40 Gb/s**

The Dutch Situation

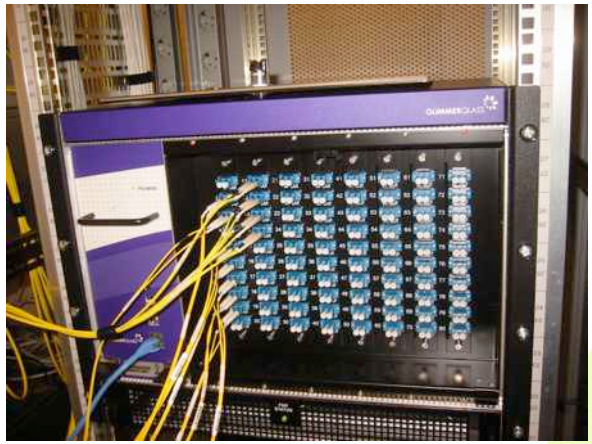
- **Estimate A**
  - **17 M people, 6.4 M households, 25 % penetration of 0.5-2.0 Mb/s ADSL, 40 times under-provisioning ==> 20 Gb/s**
- **Estimate B**
  - **SURFnet has 10 Gb/s to about 12 institutes and 0.1 to 1 Gb/s to 180 customers, estimate same for industry (overestimation) ==> 20-40 Gb/s**
- **Estimate C**
  - **Leading HEF and ASTRO + rest ==> 80-120 Gb/s**
  - **LOFAR ==> ≈ 26 Tbit/s**

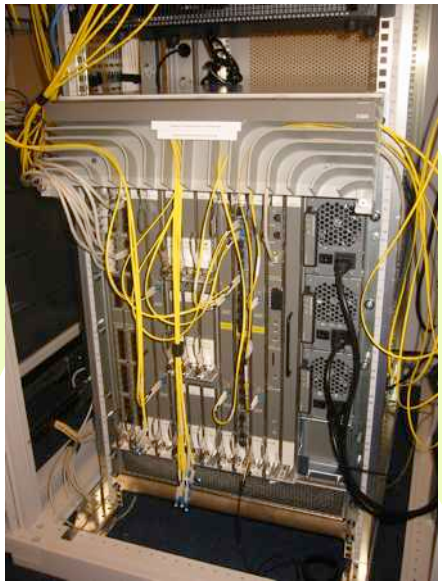

#users
A. Lightweight users, browsing, mailing, home use
Need full Internet routing, one to many
B. Business applications, multicast, streaming, VPN's, mostly LAN
Need VPN services and full Internet routing, several to several + uplink
C. Special scientific applications, computing, data grids, virtual-presence
Need very fat pipes, limited multiple Virtual Organizations, few to few
ΣC >> 100 Gb/s
ΣB ≈ 40 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 -> 100-300 k$, 10G switch port -> 10-20 k$, MEMS port -> 0.7 k$**
  - **DWDM lasers for long reach expensive, 10-50k$ (???)**
  - **64 Byte packet @ 10 Gbit/s -> 52 ns -> time to look up destination in 140 kEntries routing table (light speed from me to you (15 meter)!)**
- **Bottom line: look for a hybrid architecture which serves all classes in a cost effective way (A -> L3 , B -> L2 , C -> L1)**
- **Give each packet in the network the service it needs, but no more**



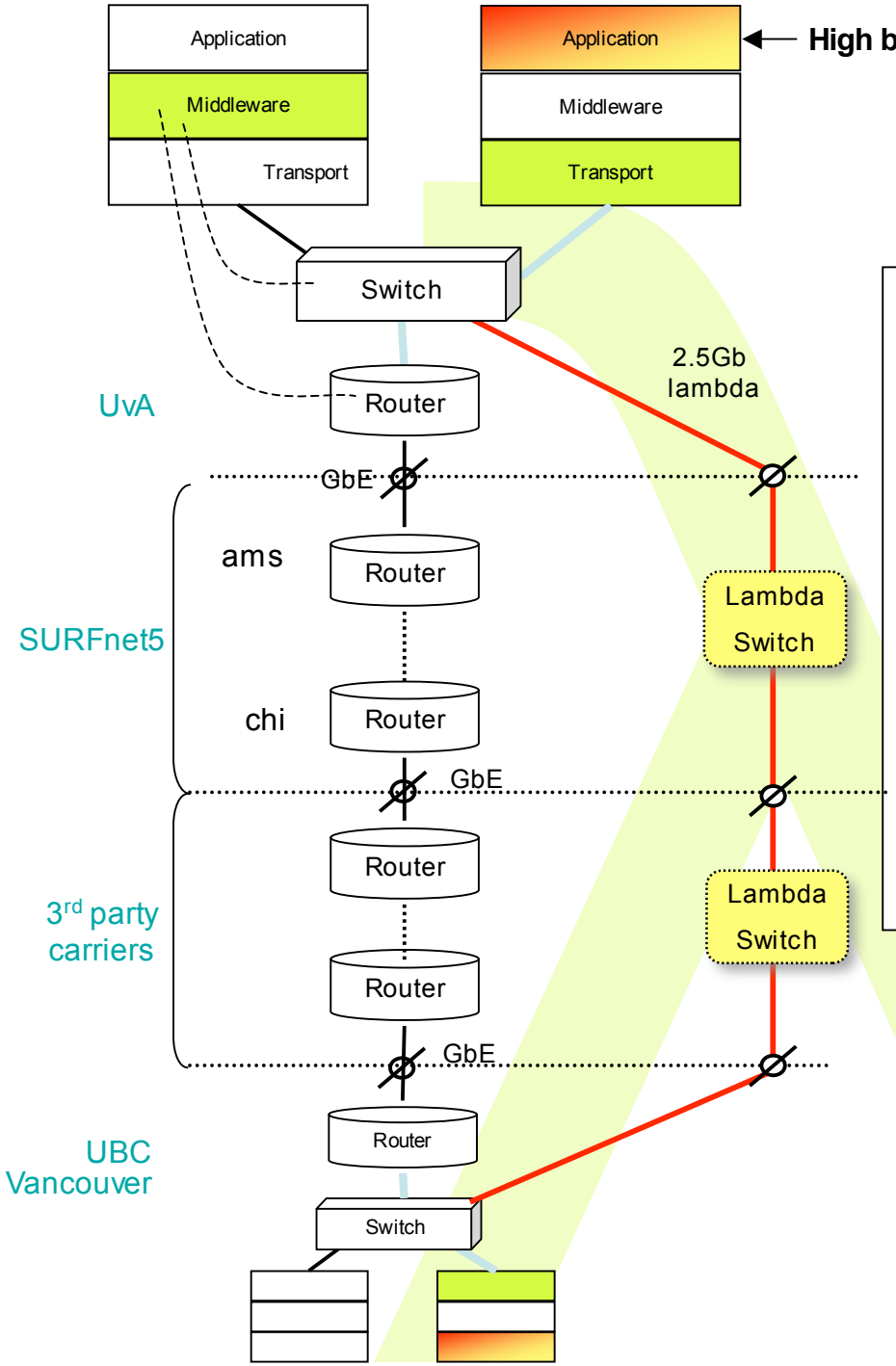


UVA/EVL's
64*64
Optical Switch
@ NetherLight
in SURFnet POP @
SARA

Costs 1/100th of a similar throughput router or 1/10th of an Ethernet switch but with specific services!

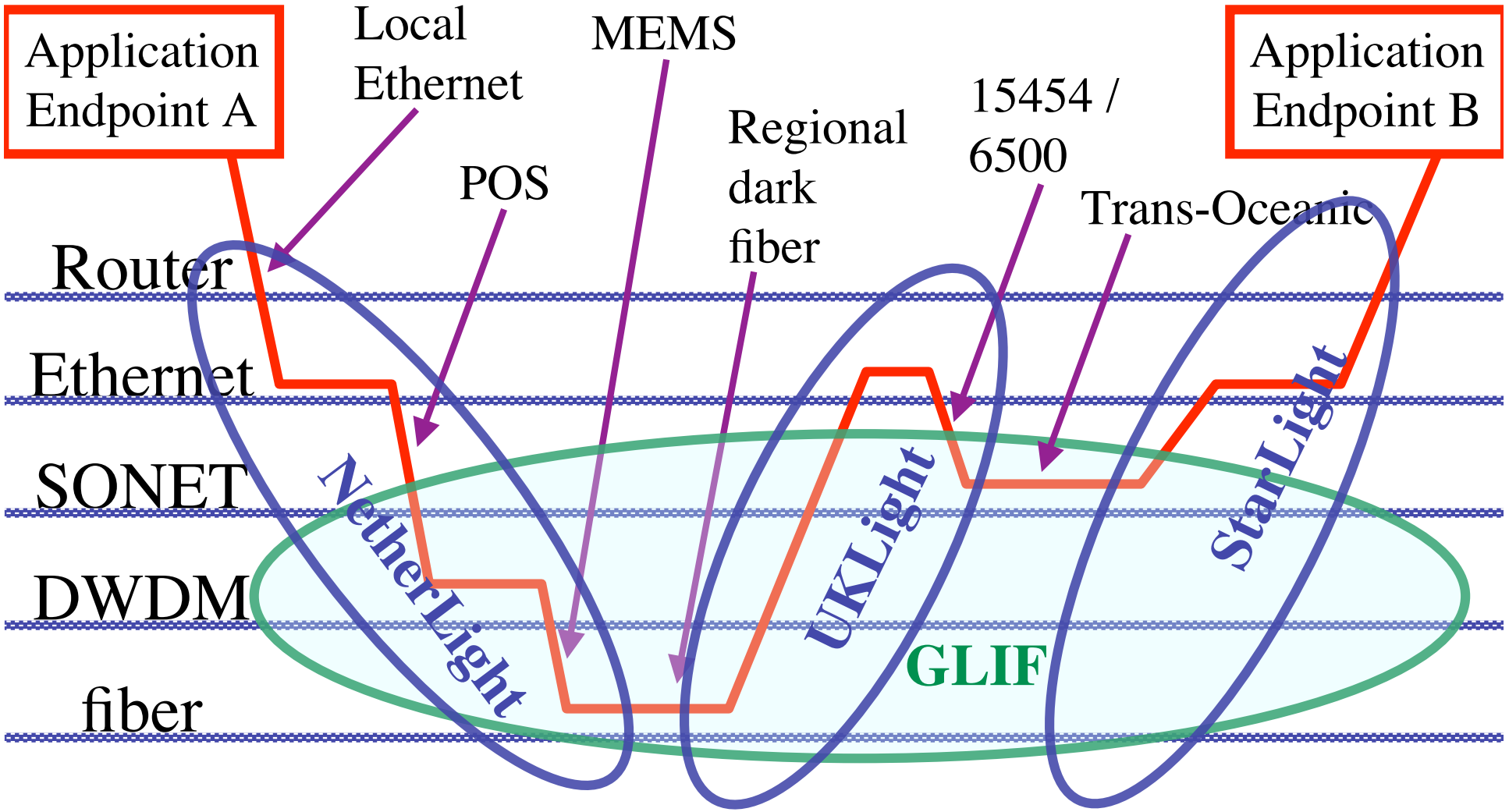
Services

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

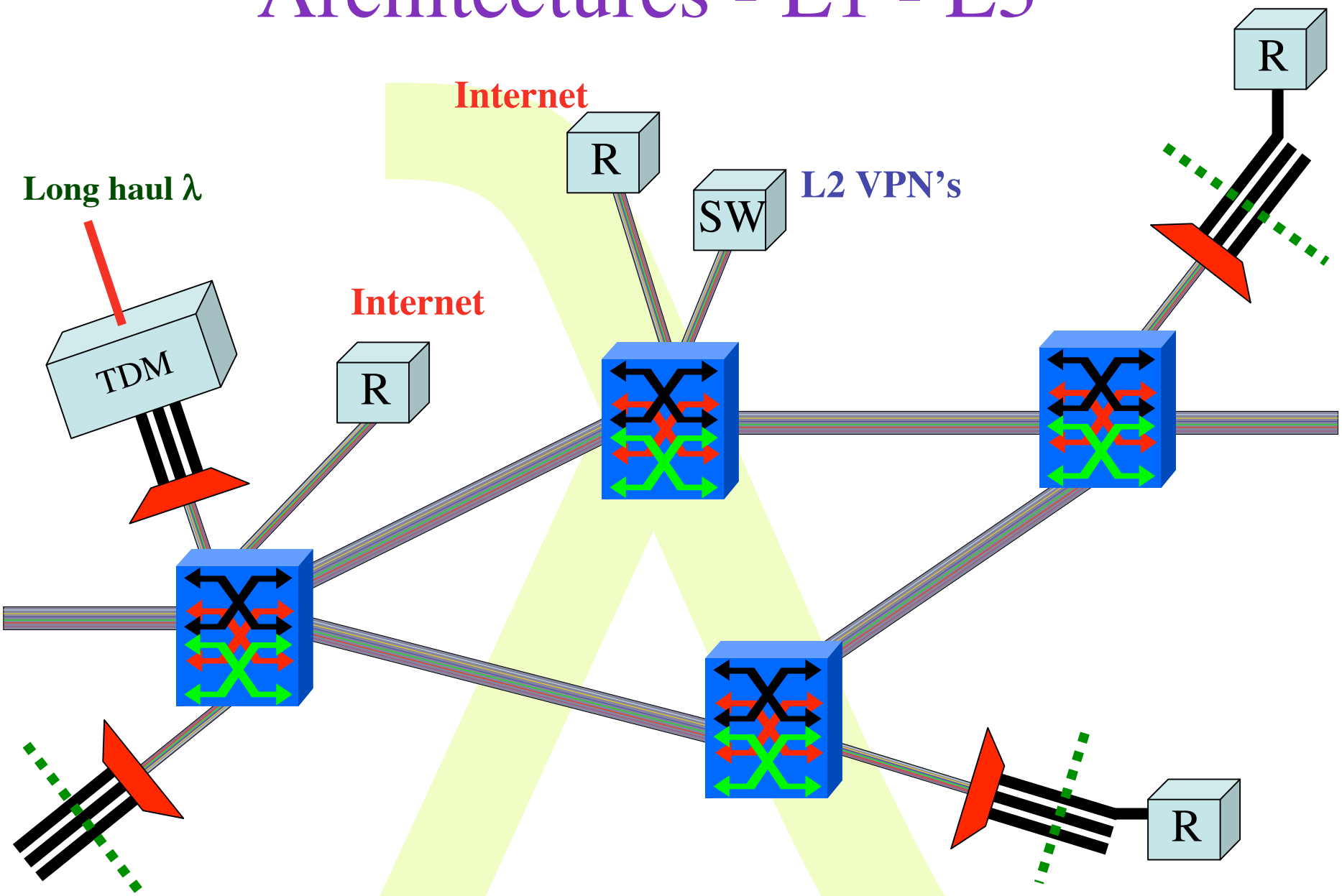


- lambda for high bandwidth applications
  - Bypass of production network
  - Middleware may request (optical) pipe
- RATIONALE:
  - Lower the cost of transport per packet
  - Use Internet as controlplane!

How low can you go?

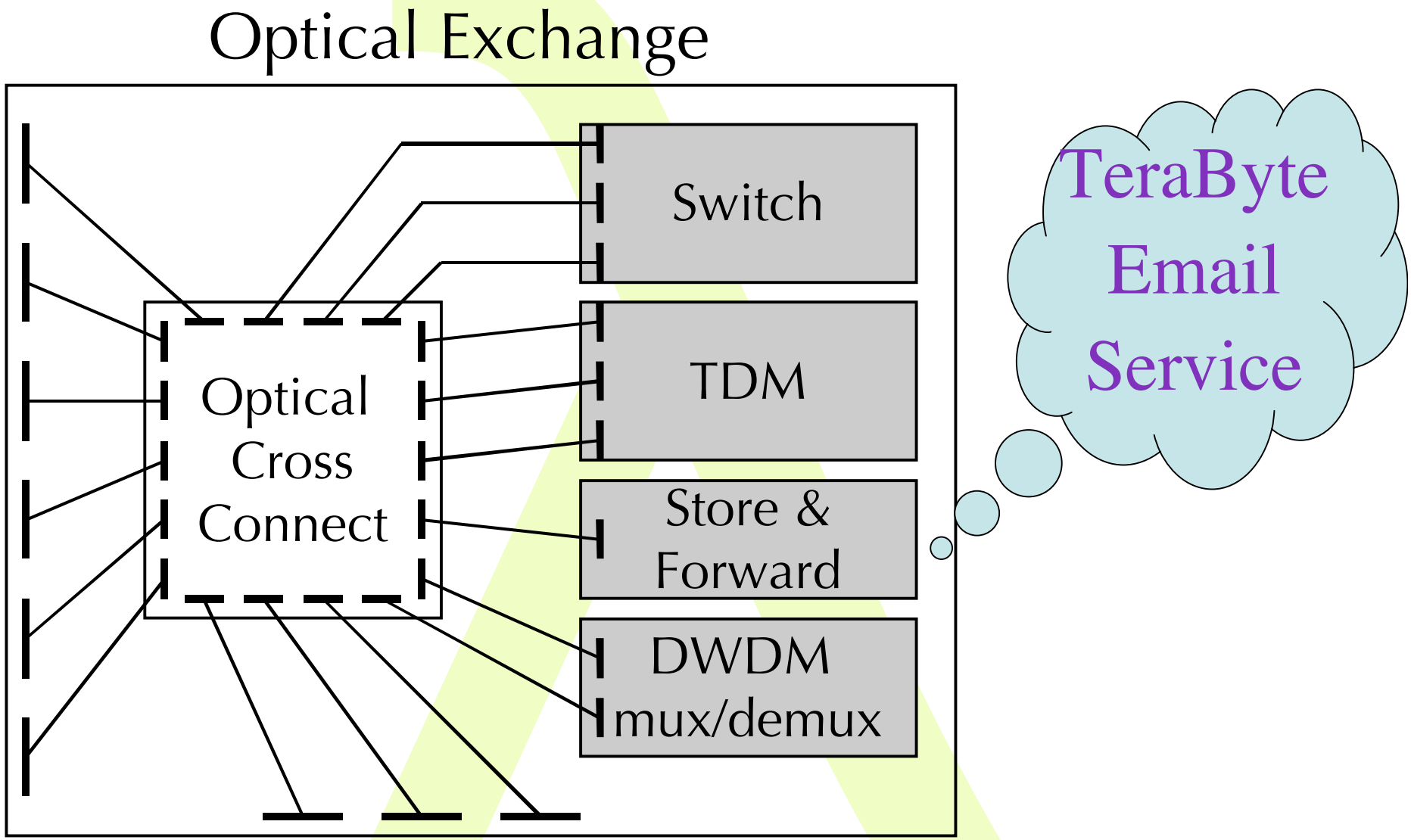


Architectures - L1 - L3



Bring plumbing to the users, not just create sinks in the middle of nowhere

Optical Exchange as Black Box

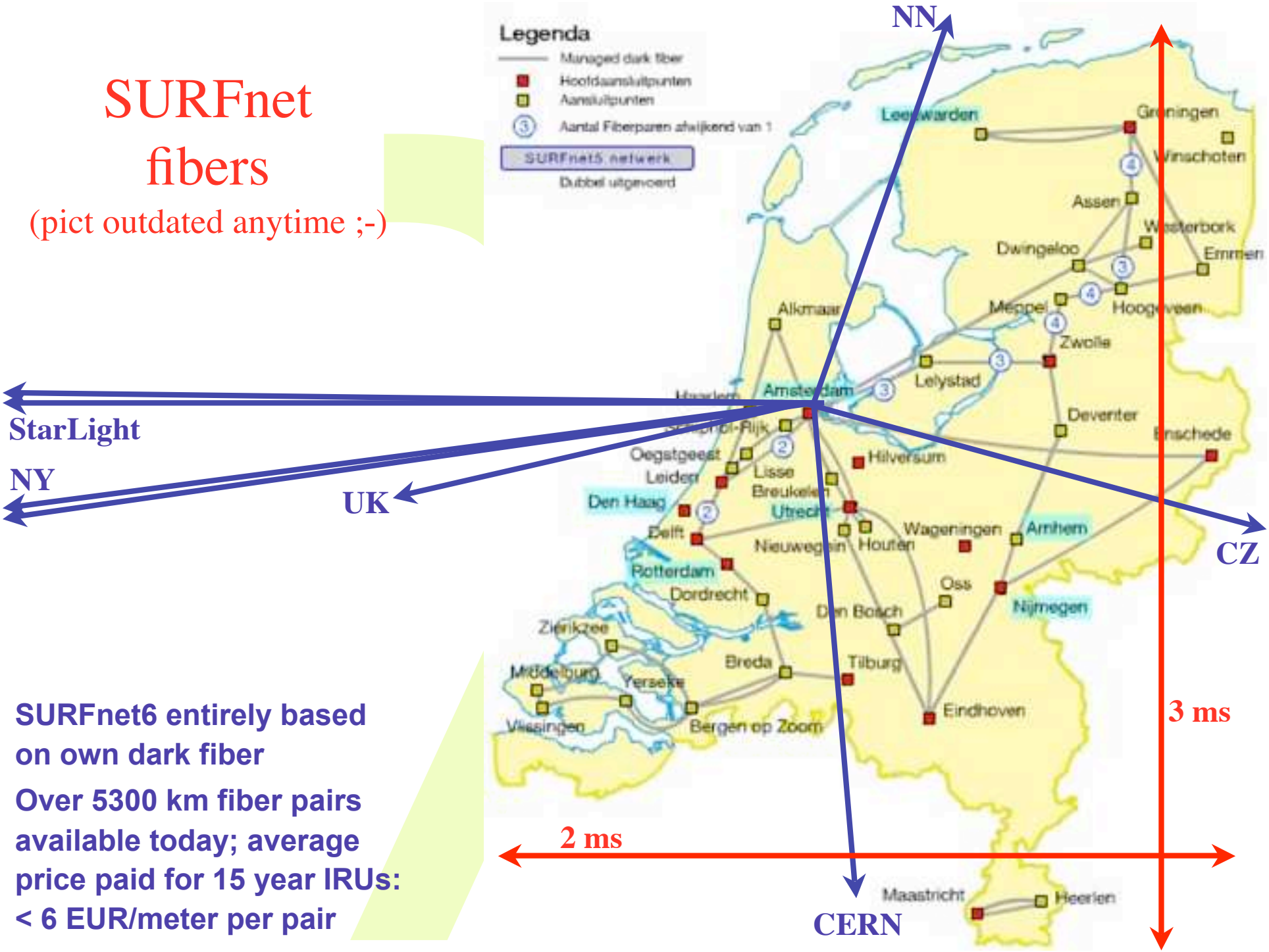


Service Matrix

From \ To	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 inspection in Science Park Amsterdam :-)

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: To build a new grid-computing paradigm, in which the central architectural element is optical networks, not computers, to support this decade's most demanding e-science applications.

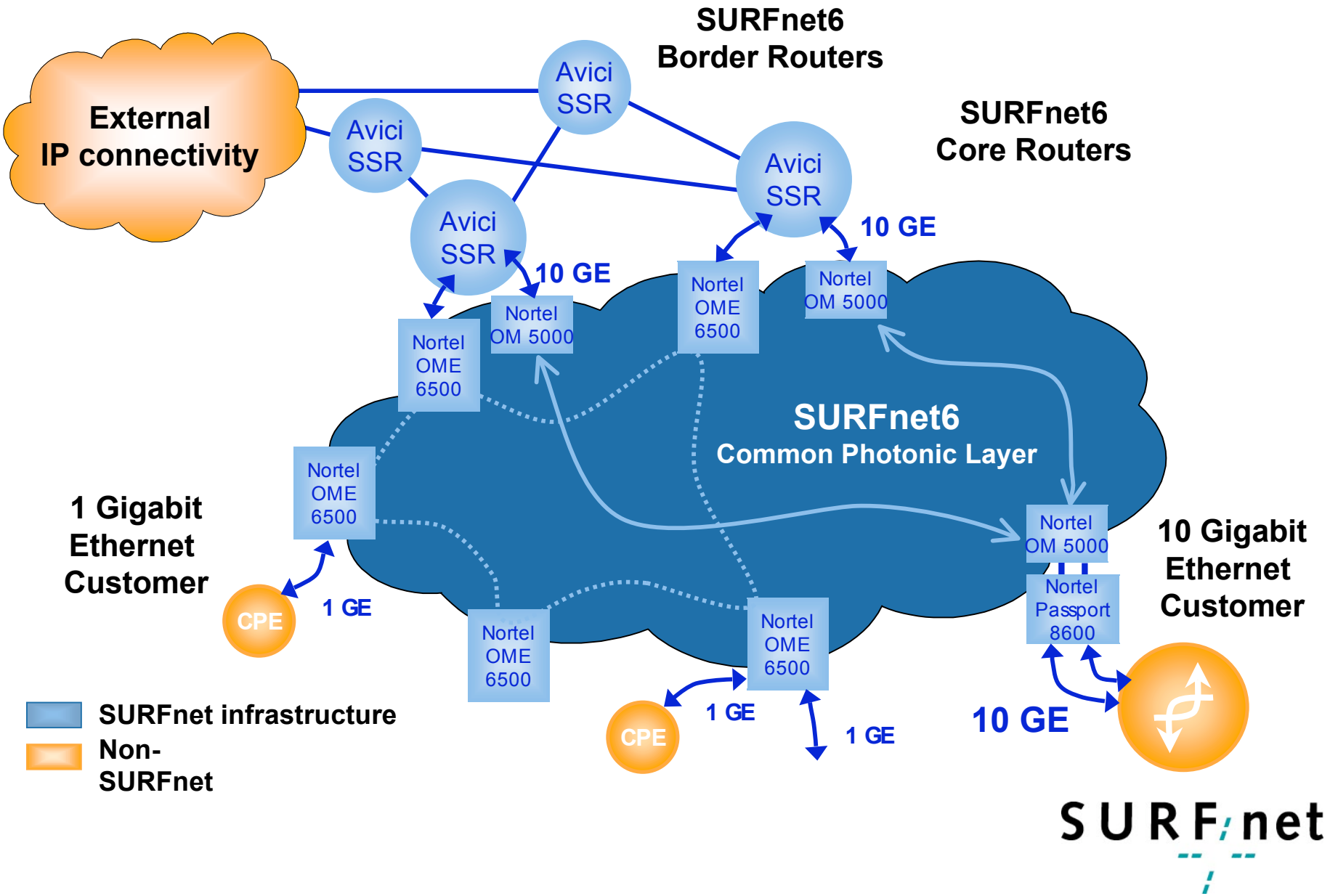
Coordinated by UvA, SURFnet and UIC

GLIF Q3 2004

Visualization courtesy of
Bob Patterson, NCSA.

IP network implementation

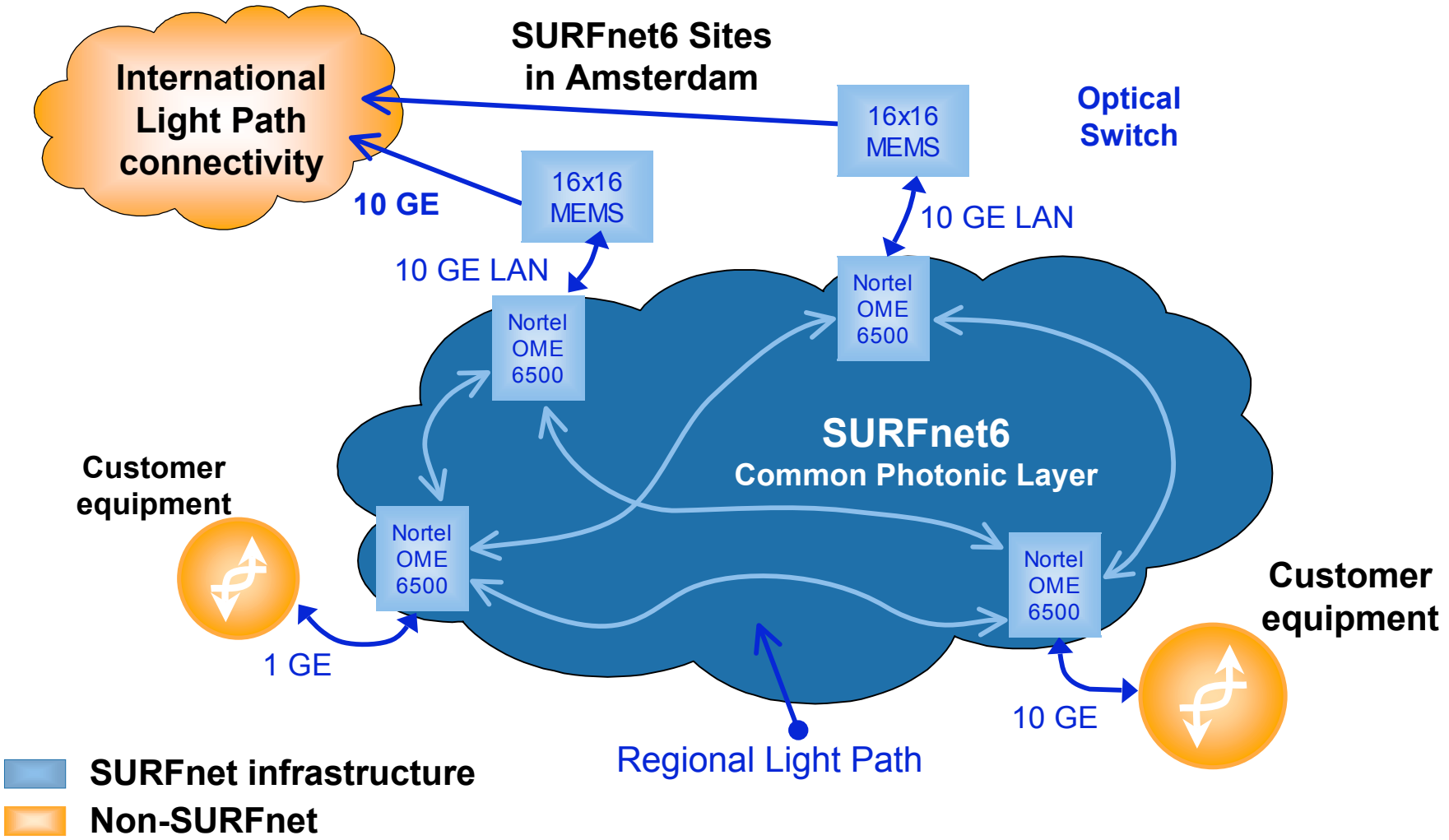
GigaPort



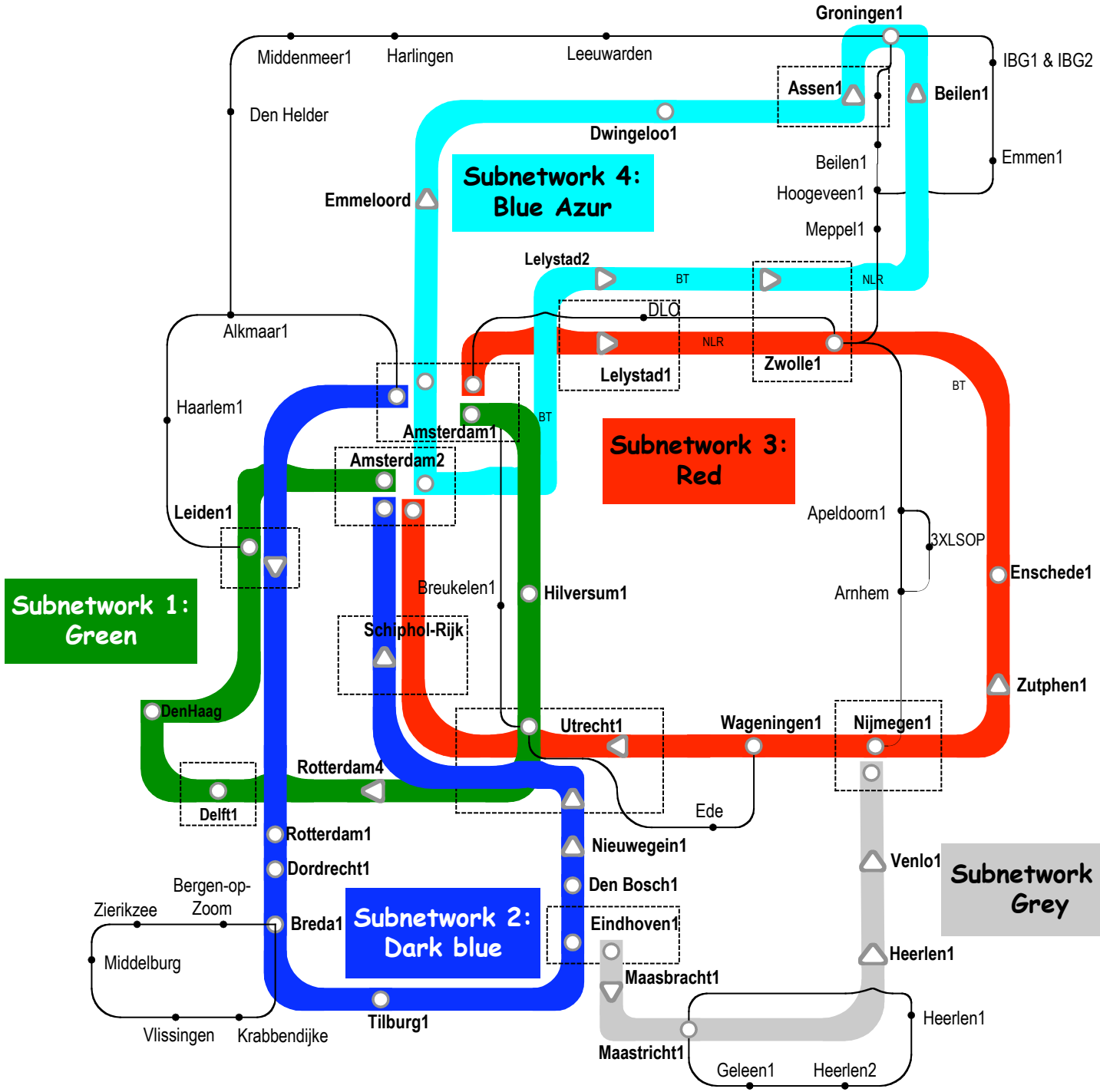
SURFnet

Light Paths provisioning implementation

GigaPort



SURFnet

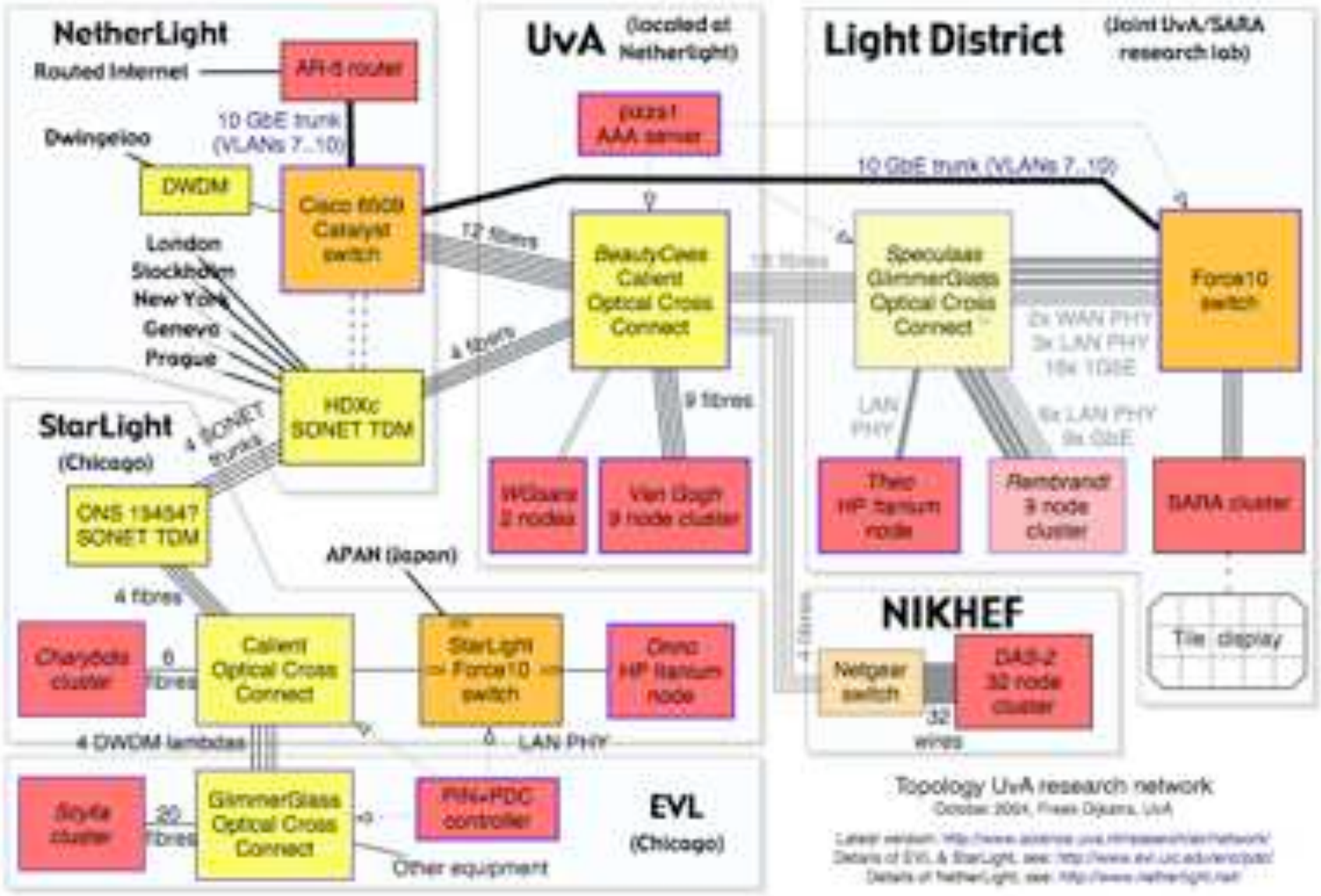


GigaPort

Common Photonic Layer (CPL) in SURFnet6

SURFnet

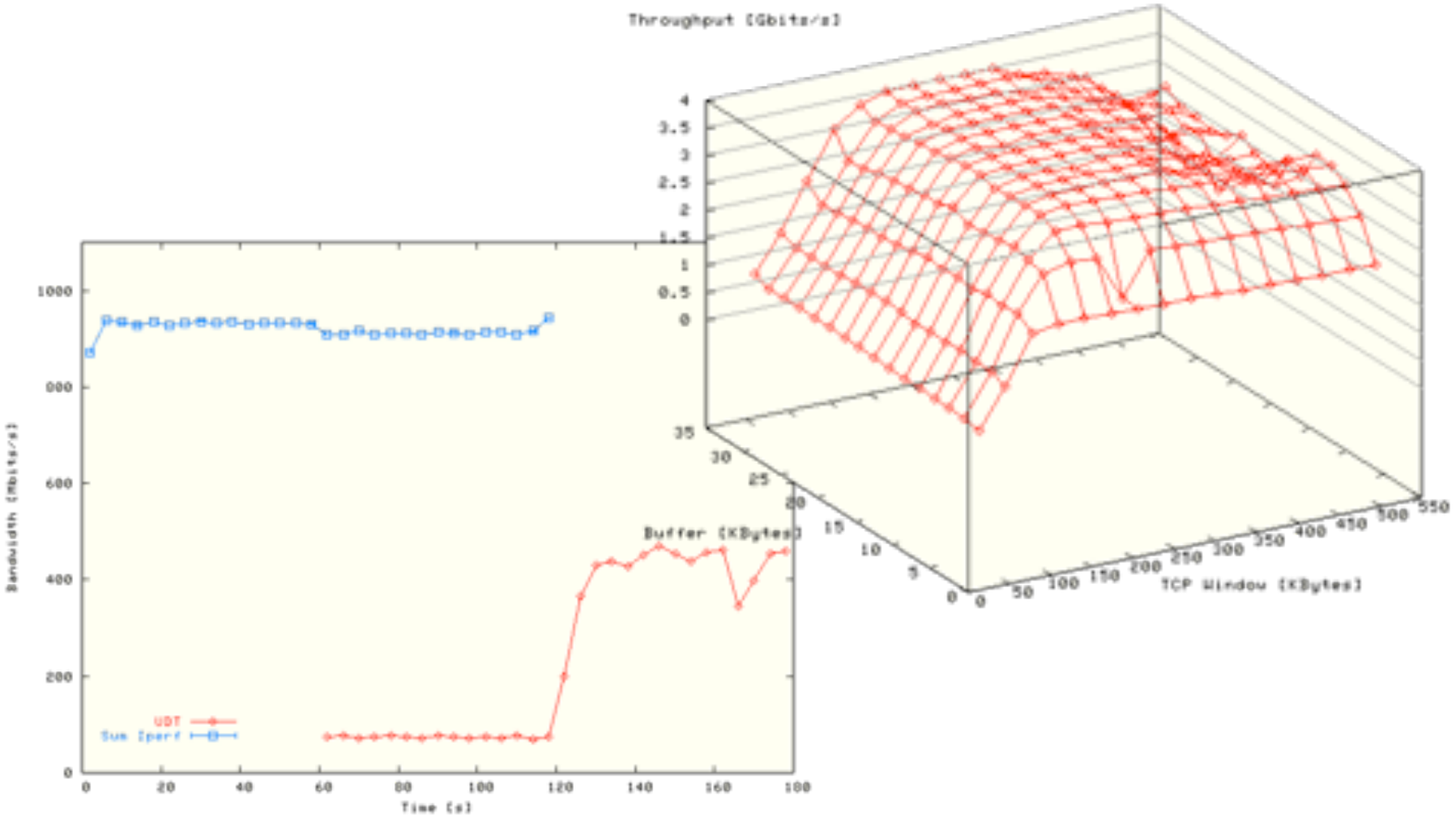
LightHouse


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
4 SONET trunks
StarLight (Chicago)
ONS 154547 SONET TDM
4 fibres
APAN (Japan)
Charybdis cluster
6 fibres
Calient Optical Cross Connect
StarLight Force10 switch
Cisco HP Itanium node
4 DWDM lambdas
LAN PHY
Scylla cluster
20 fibres
GlimmerGlass Optical Cross Connect
PIN-PDC controller
Other equipment
EVL
(Chicago)
UvA
(located at NetherLight)
uuns1 AAA server
12 fibres
4 fibres
BeautyCees Calient Optical Cross Connect
9 fibres
WGasmo 2 nodes
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
2x LAN PHY
16x 1GbE
LAN PHY
8x LAN PHY
8x GbE
Thor HP Itanium node
Rembrandt 8 node cluster
SARA cluster
NIKHEF
Netgear switch
32 wires
DAS-2 32 node cluster
Tile display
Topology UvA research network
October 2004, Pieter Dijsma, UvA

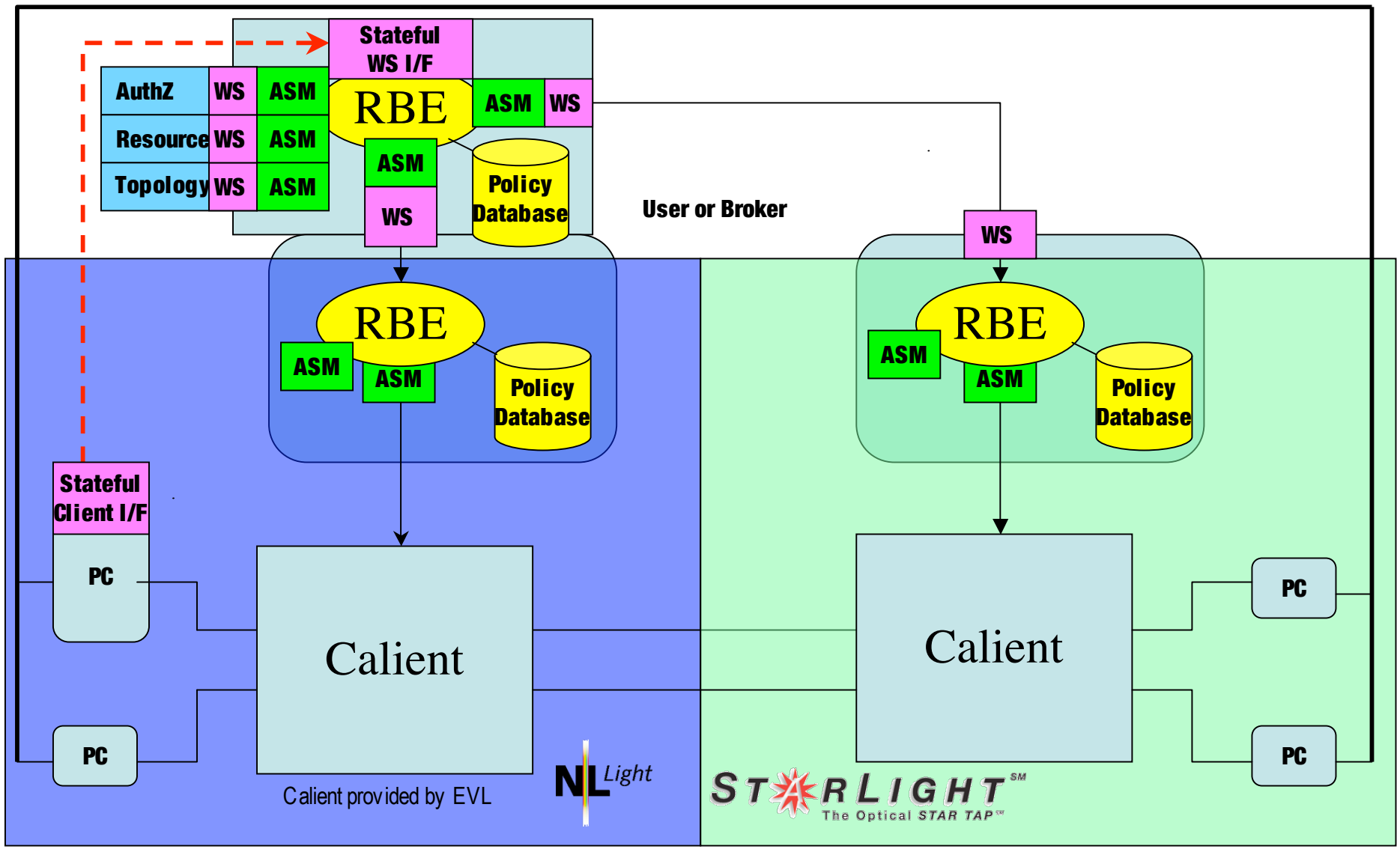
Research topics

- 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

Example Measurements



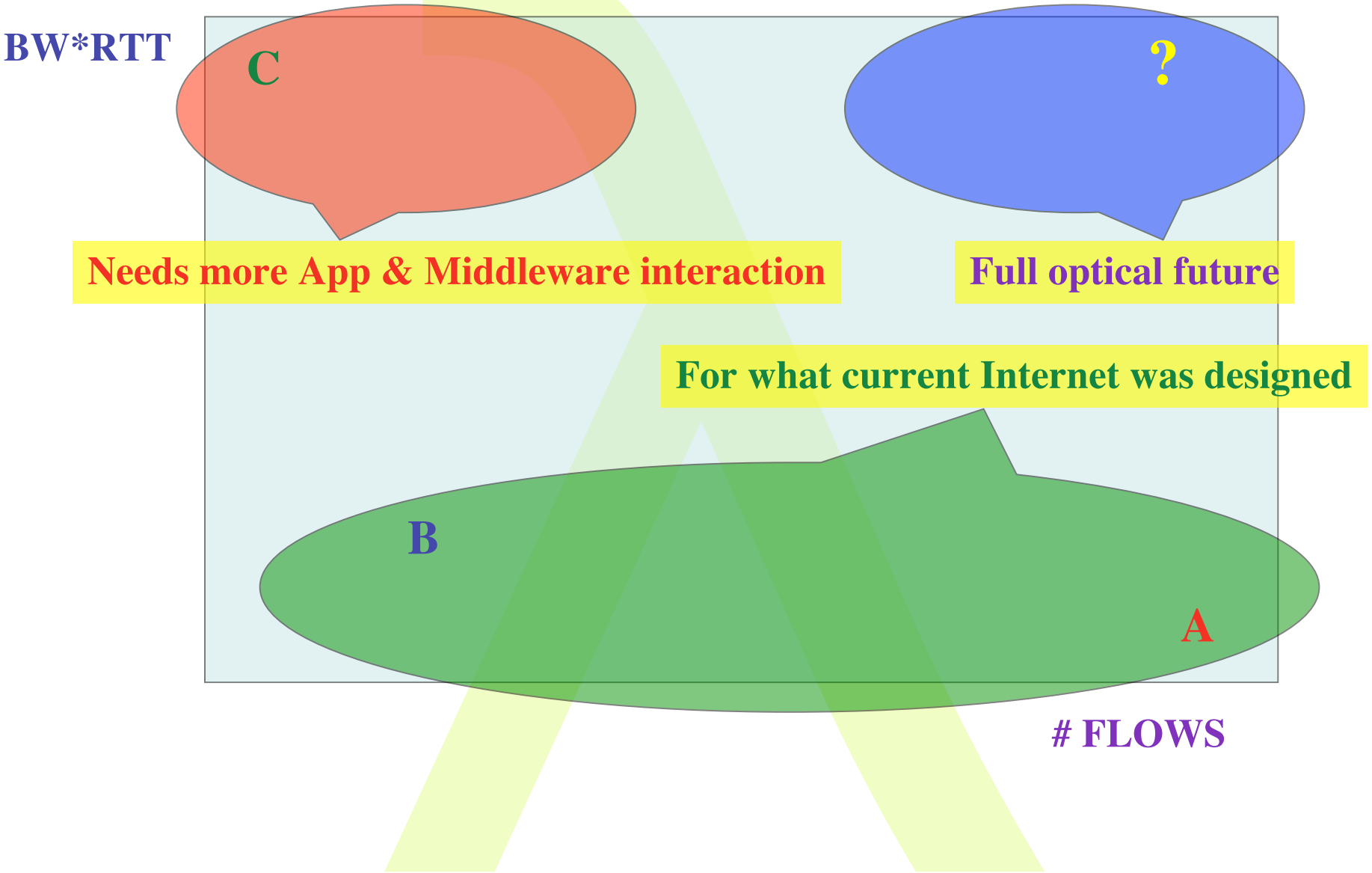
AAA based demo at SC2003



Conlusions

- Demanding applications
  - (Science) data repositories mirroring
  - Instrumentation grids
  - Visualisation and collaboration support
- Model of Lambda networking
  - Identify traffic types
  - Scales of infrastructure
  - Map efficiently to lower the cost/packet
- Current experiments
  - NetherLight
  - VLE/eScience Amsterdam
  - Networking research
    (control plane, transport protocols, optical net models)

Transport in the corners



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

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

[1957-2004]