Lambda-Grid developments

www.science.uva.nl/~delaat

Cees de Laat

GigaPort EU

University of Amsterdam

SARA

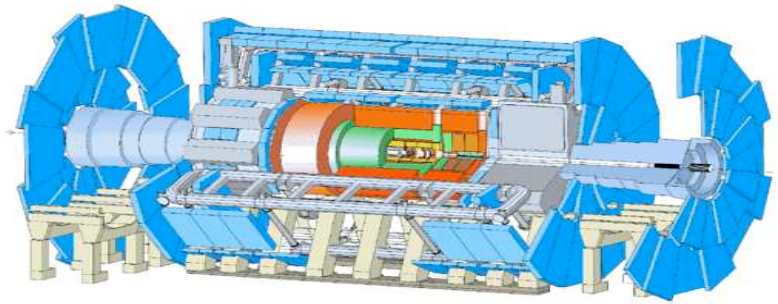
NCF

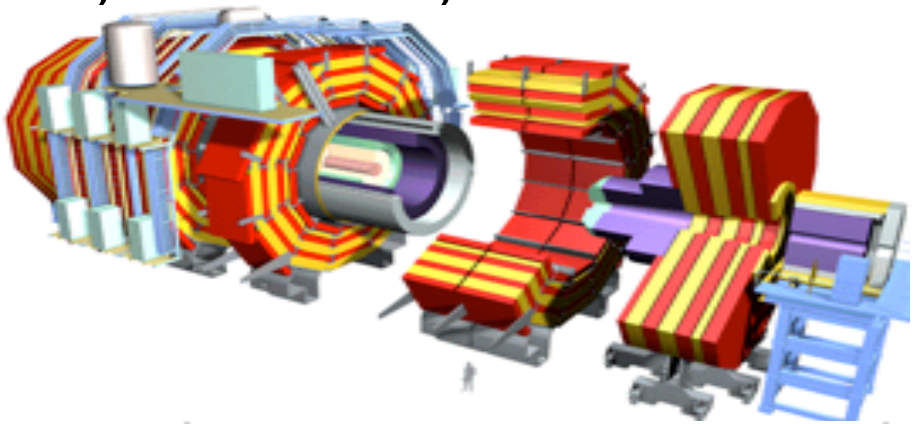
Contents of this talk

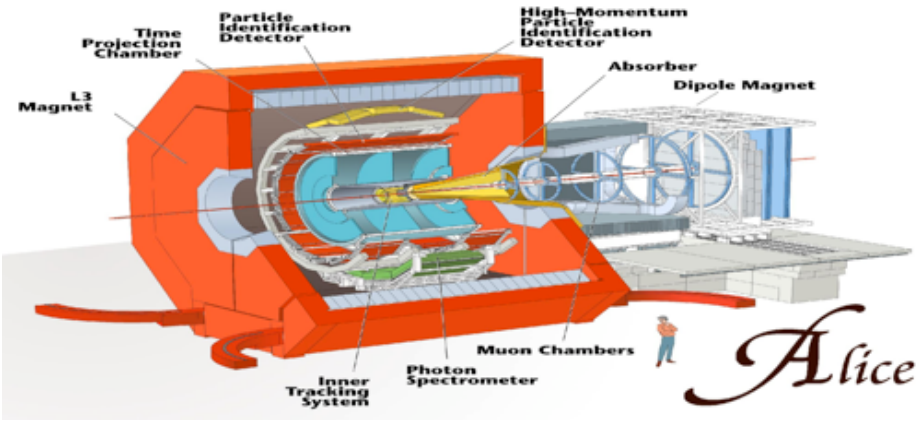
- Demanding applications
- Model of Lambda networking
- Current experiments

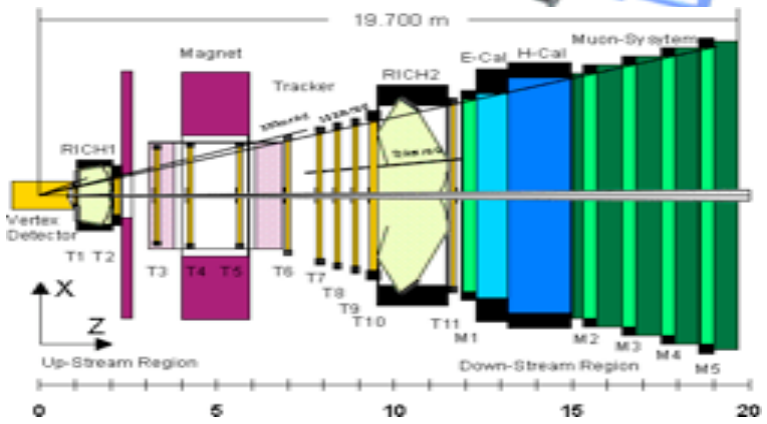
Four LHC Experiments: The Petabyte to Exabyte Challenge

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









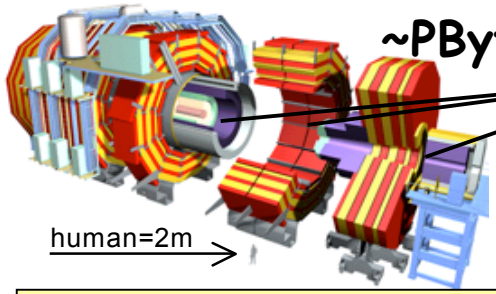
6000+ Physicists & Engineers; 60+ Countries; 250 Institutions

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.

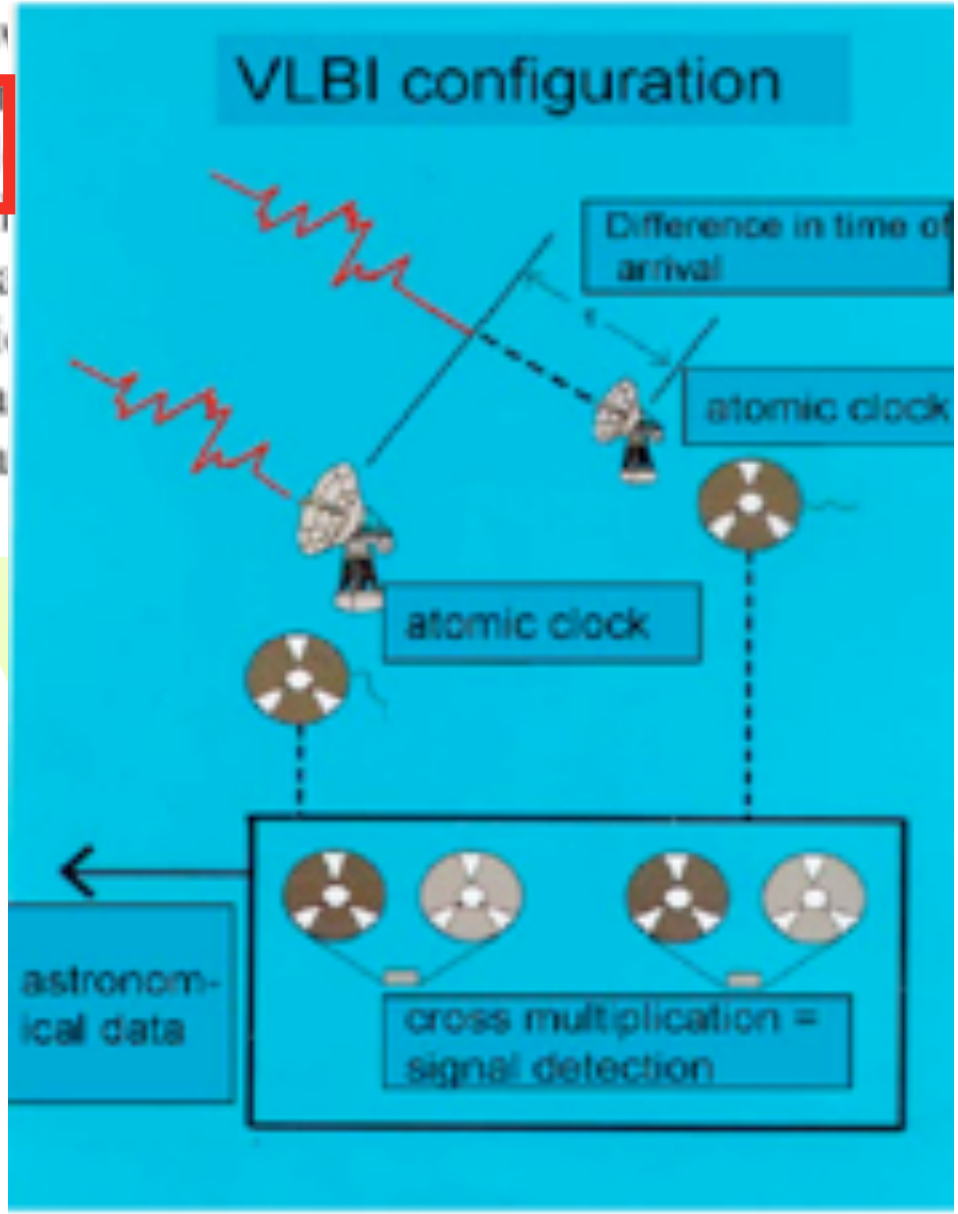
Physicists work on analysis "channels" at 135 institutes. Each institute has ~10 physicists working on one or more channels.

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

SURFnet

OptIPuter Project Goal: Scaling to 100 Million Pixels

- **JuxtaView (UIC EVL) for PerspecTile LCD Wall**
  - Digital Montage Viewer
  - 8000x3600 Pixel Resolution~30M Pixels
- **Display Is Powered By**
  - 16 PCs with Graphics Cards
  - 2 Gigabit Networking per PC

Source: Jason Leigh, EVL, UIC; USGS EROS

UCSD UCIrvine

it²

Grids

Showed you:

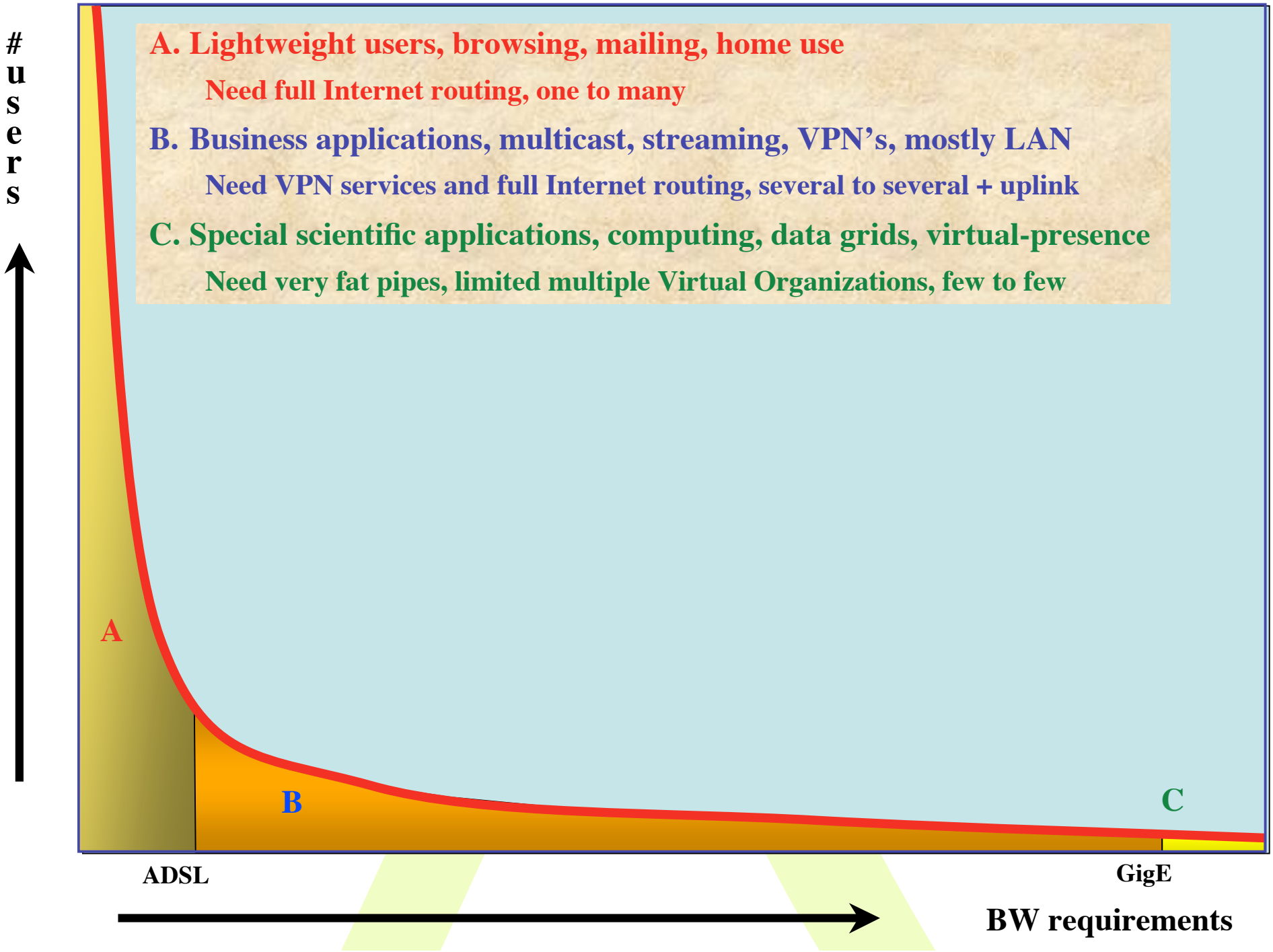
- Computational Grids
  - HEP and LOFAR analysis requires massive CPU capacity
- Data Grid
  - Storing and moving HEP, Bio and Health data sets is major challenge
- Instrumentation Grids
  - Several massive data sources are coming online
- Visualization Grids
  - Data object (TByte sized) inspection,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

UCSD
UCIrvine

it²

Contents of this talk

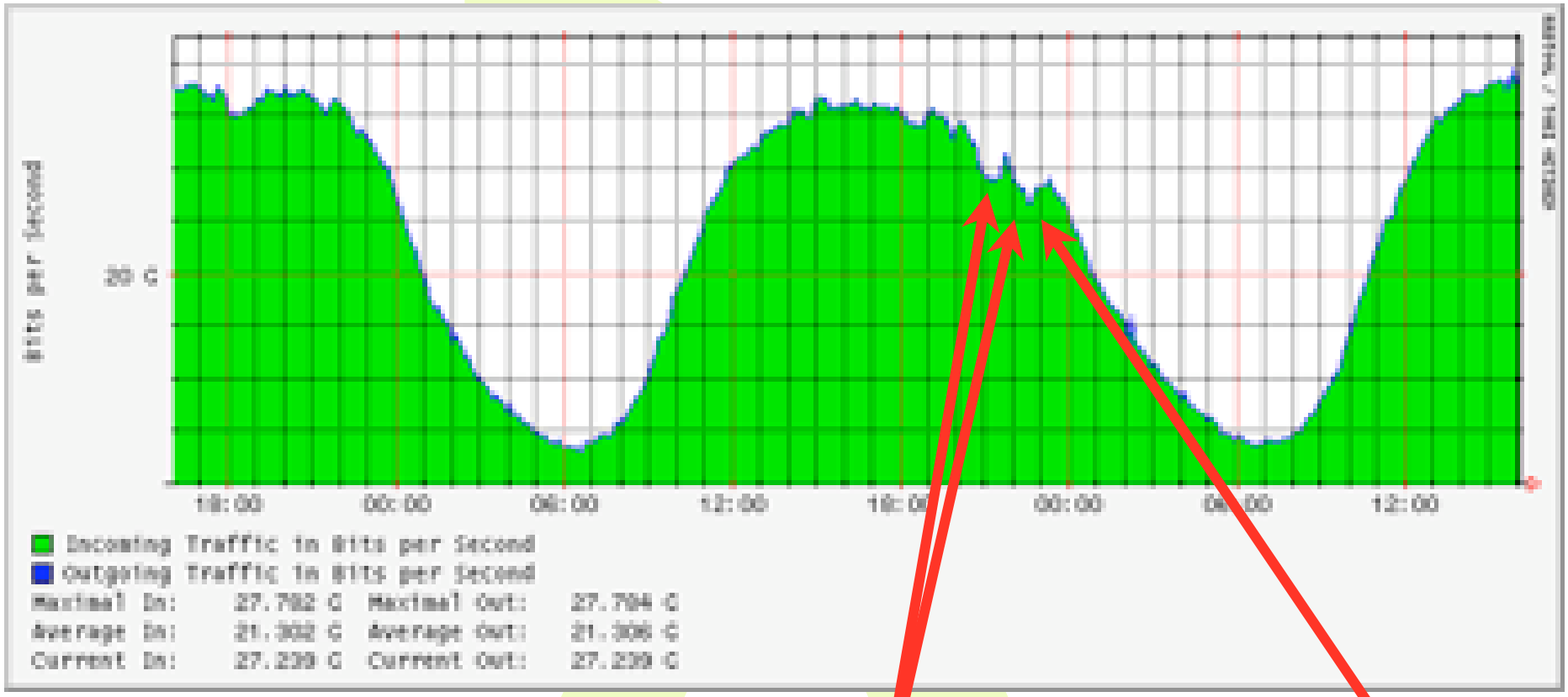
- Demanding applications
- Model of Lambda networking
- Current experiments


#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 :-(

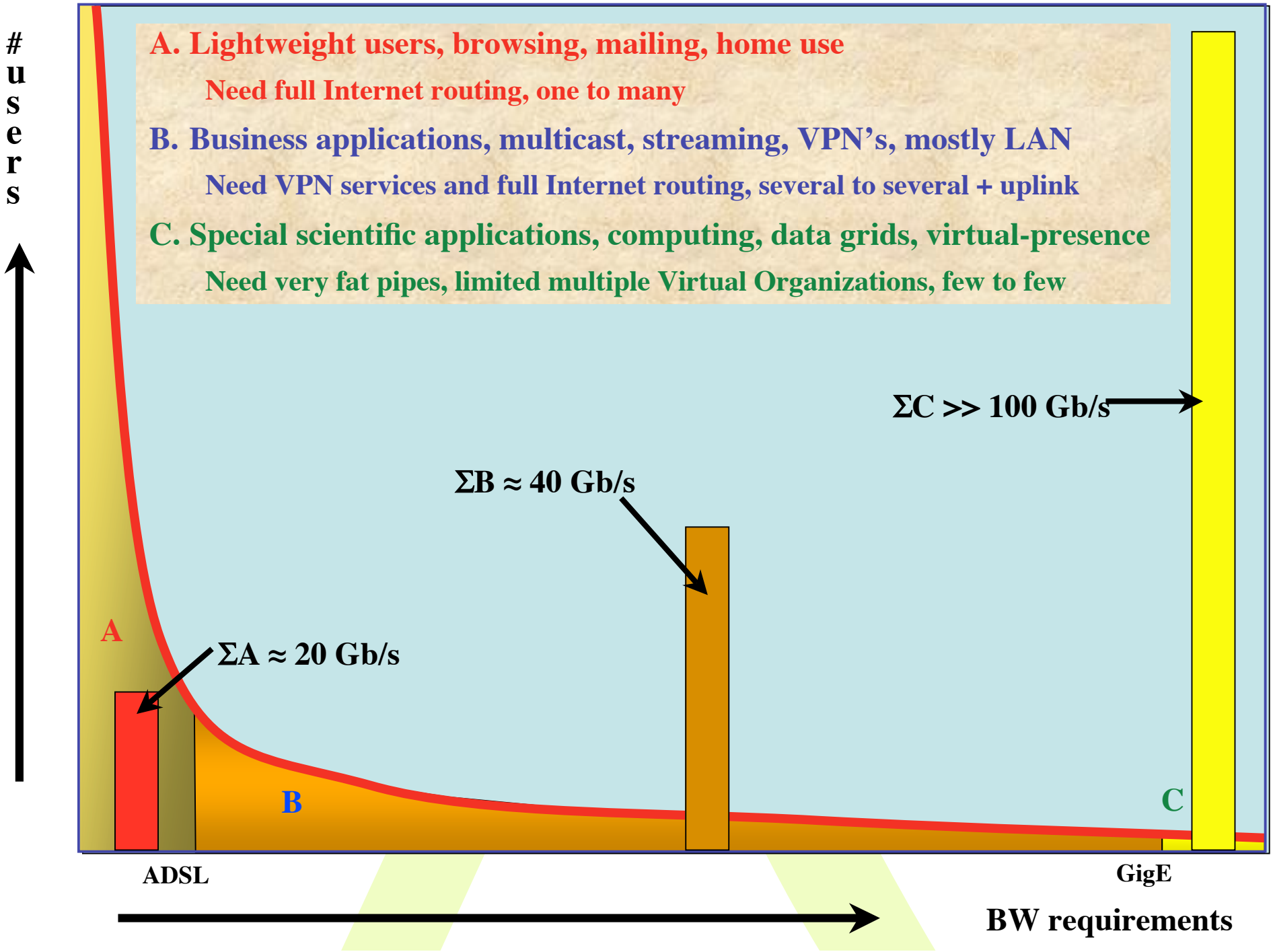
European championship football **Holland -- Czech Republic**

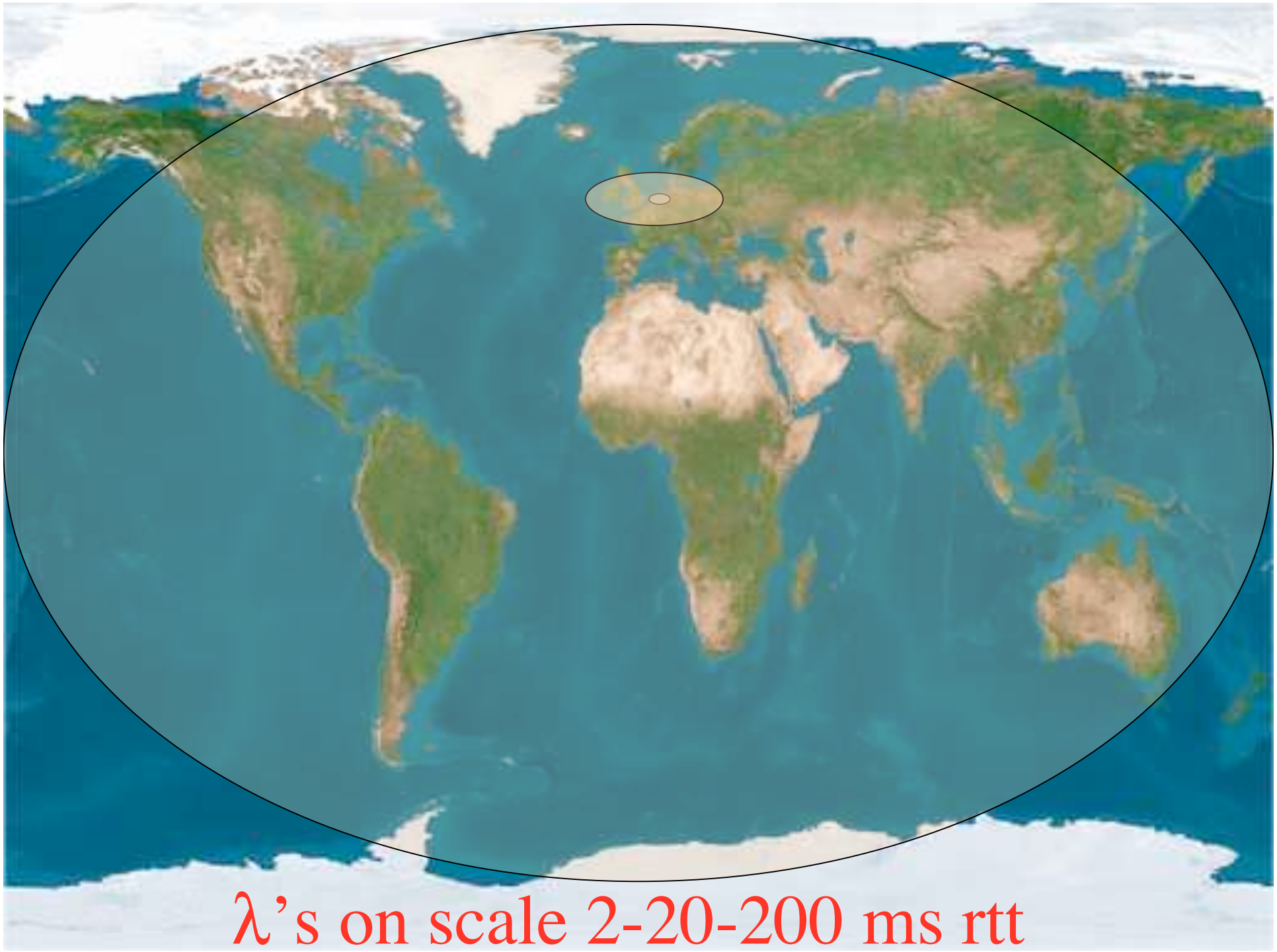
The Dutch Situation

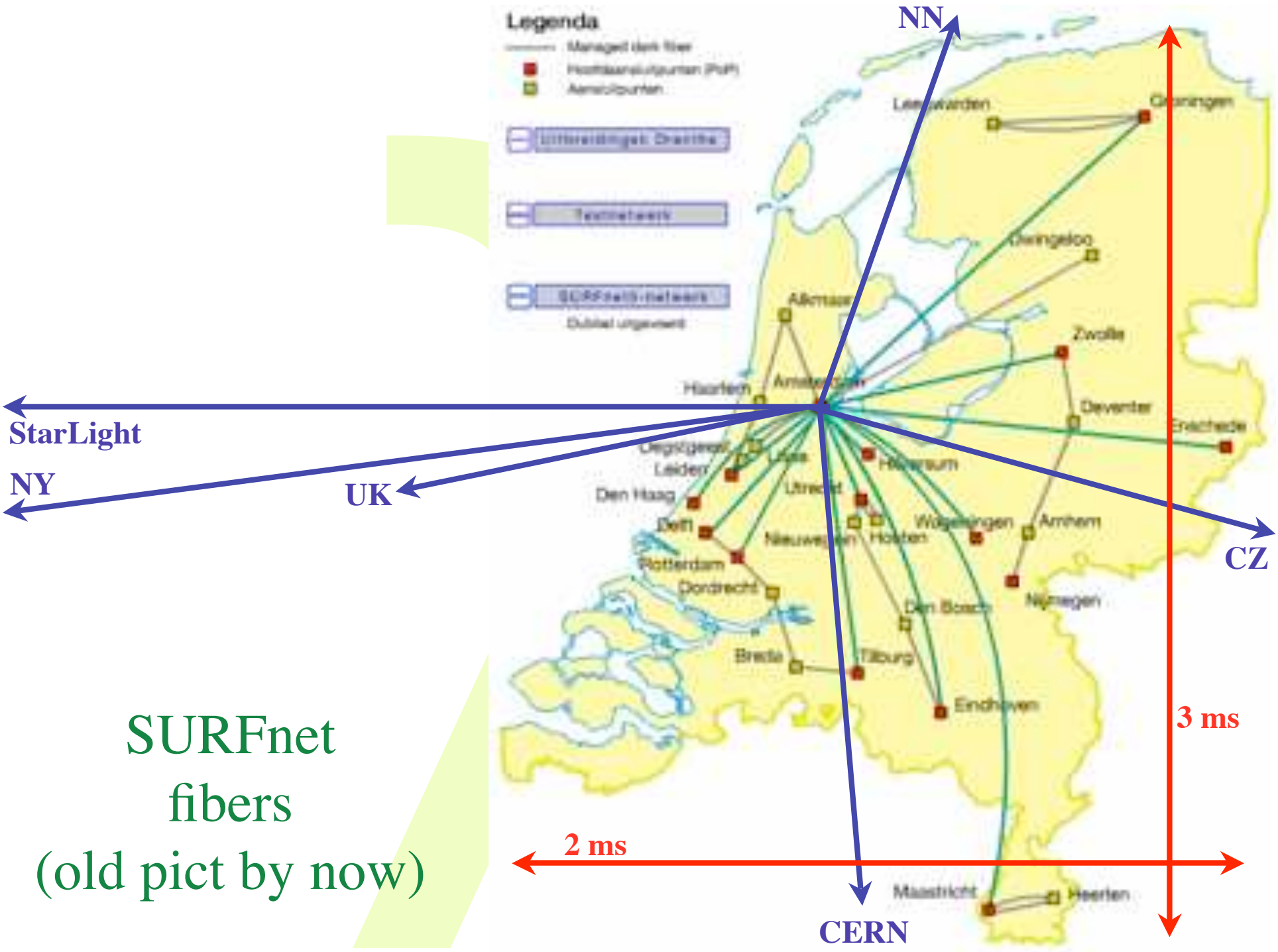
- **Estimate A**
  - **17 M people, 6.4 M households, 25 % penetration of 0.5-2.0 Mb/s ADSL, 40 times under-provisioning ==> 20 Gb/s**
- **Estimate B**
  - **SURFnet has 10 Gb/s to about 12 institutes and 0.1 to 1 Gb/s to 180 customers, estimate same for industry (overestimation) ==> 20-40 Gb/s**

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



SURFnet
fibers
(old pict by now)

The only formula

$$\#\lambda(rtt,t) \approx \frac{200 * e^{(t-2002)}}{rtt}$$

Compares very well with SURFnet's resources and Lambda's @ NetherLight

- **1 Transatlantic Lambda in 2002, now ~10 from EU+US**
- **5300 km dark fiber in Holland ≈ railway net**

So what?

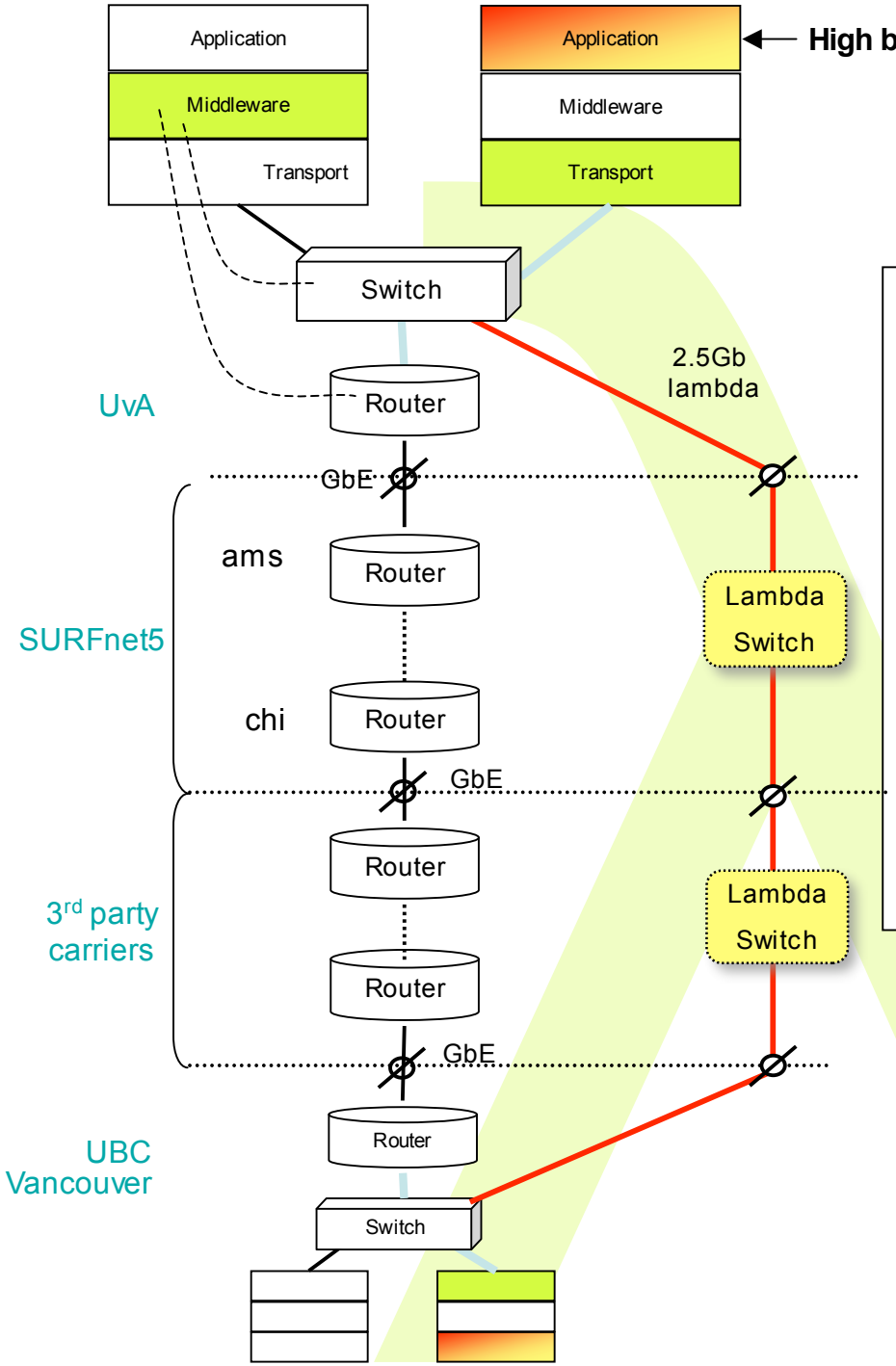
- **Costs of optical equipment 10% of switching 10 % of full routing equipment for same throughput**
  - **10G routerblade -> 100-300 k$, 10G switch port -> 10-20 k$, MEMS port -> 0.7 k$**
  - **DWDM lasers for long reach expensive, 10-50k$ (???)**
  - **64 Byte packet @ 10 Gbit/s -> 52 ns -> time to 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**
- **Look at worldwide ethernet infrastructure:**
  - **Tested 10 Gbit/s Ethernet WANPHY Amsterdam-CERN**
  - **http://www.surfnet.nl/en/publications/pressreleases/021003.html**
- **Look at worldwide lambda structure:**
  - **http://www.glif.is/**

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!

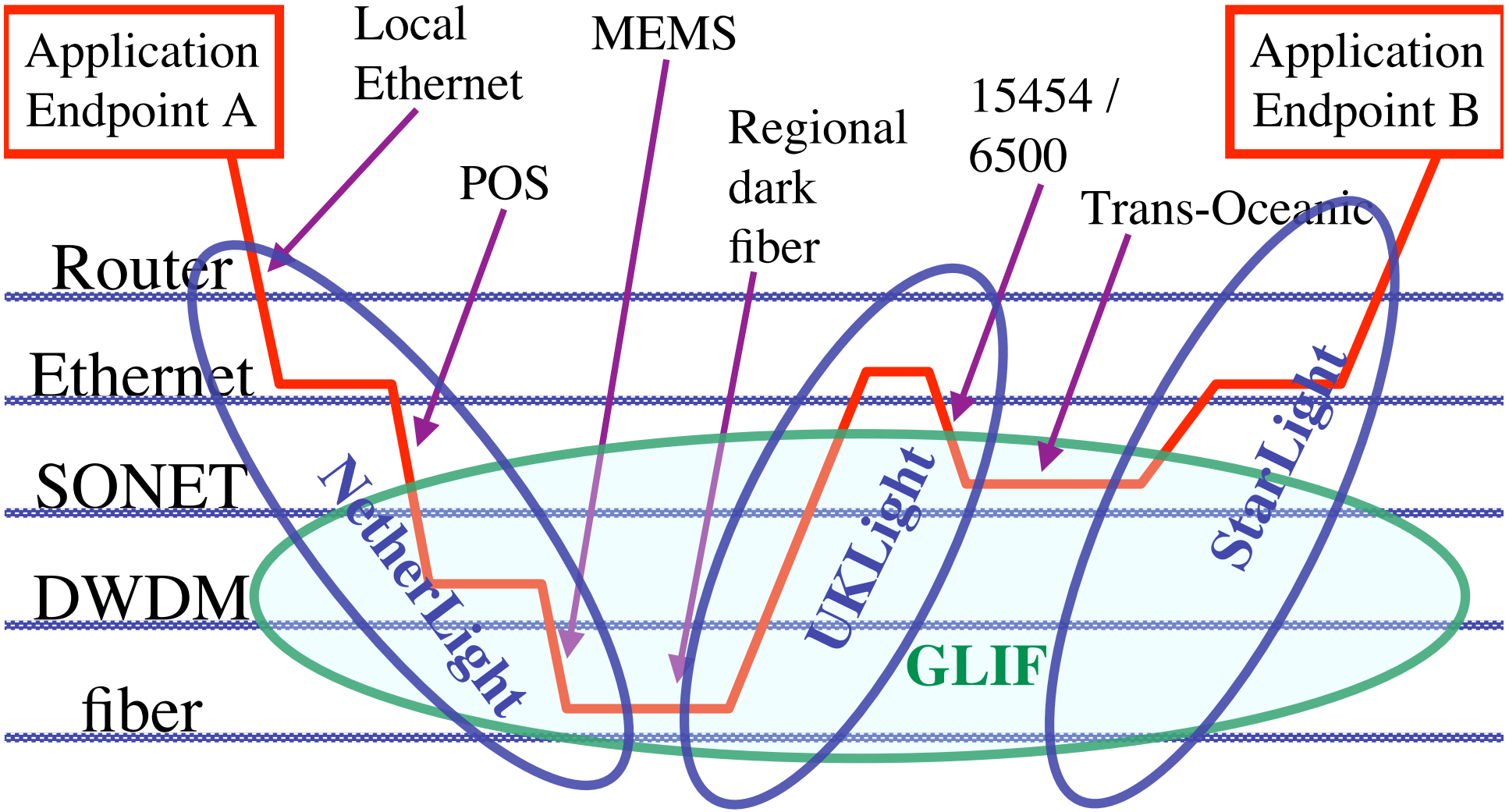
Services

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

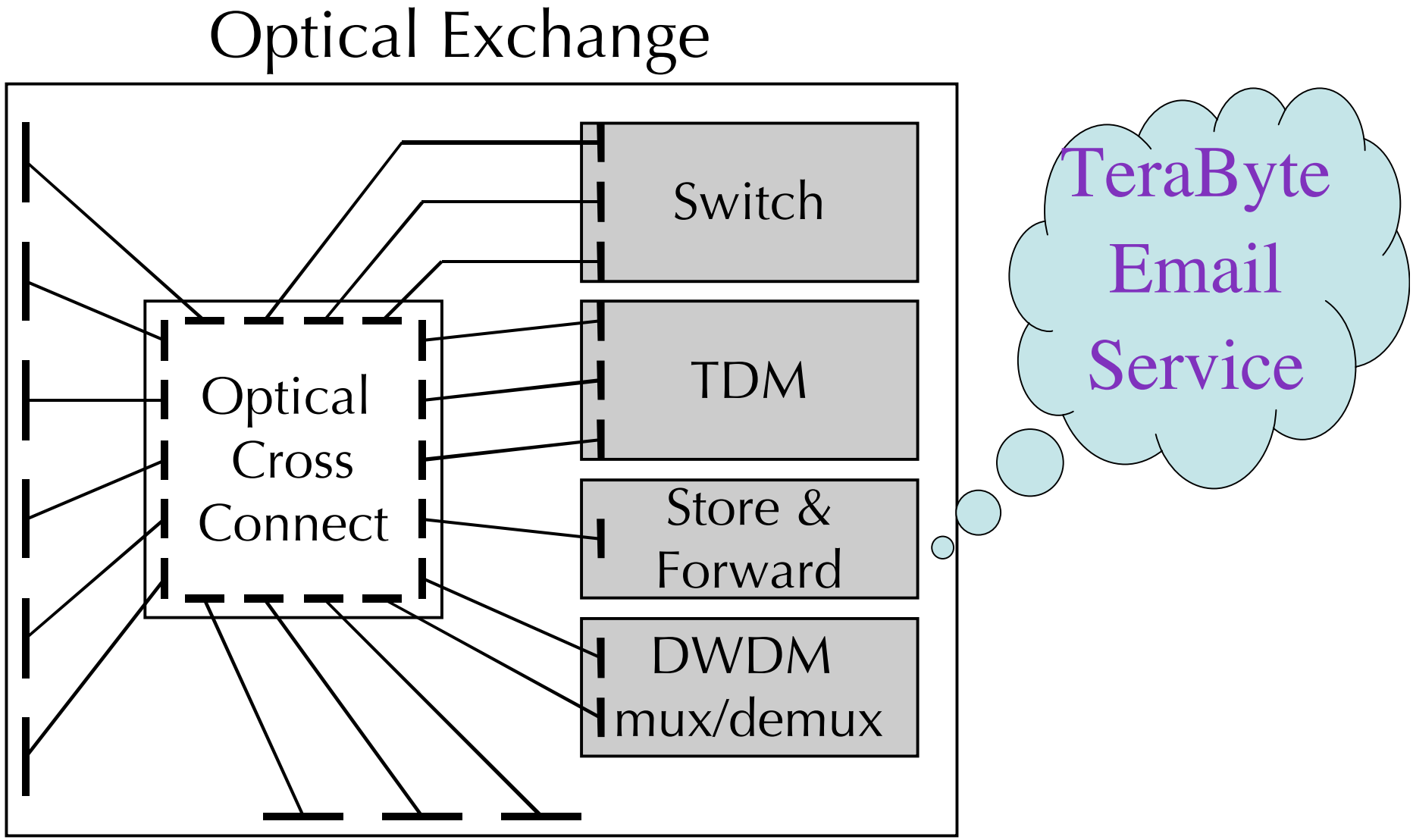


- lambda for high bandwidth applications
  - Bypass of production network
  - Middleware may request (optical) pipe
- RATIONALE:
  - Lower the cost of transport per packet
  - Use Internet as controlplane!

How low can you go?



Optical Exchange as Black Box



Contents of this talk

- Demanding applications
- Model of Lambda networking
- Current experiments

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

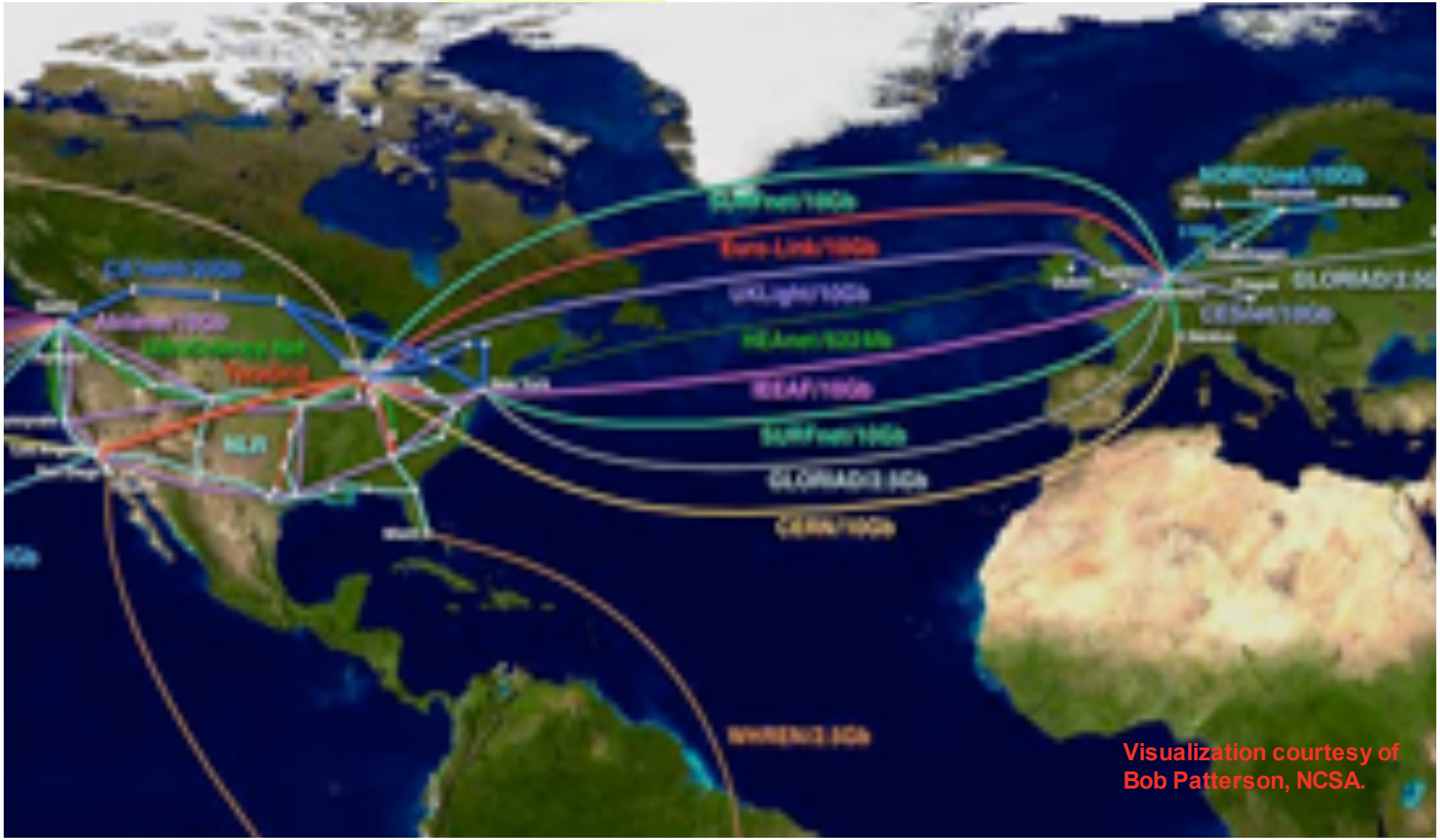
GLIF vision:

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.

GLIF Q3 2004

Visualization courtesy of
Bob Patterson, NCSA.

GLIF Q3 2004



Little GLORIAD

http://www.nsf.gov/od/lpa/news/03/pr03151.htm

SURFnet6 on dark fiber

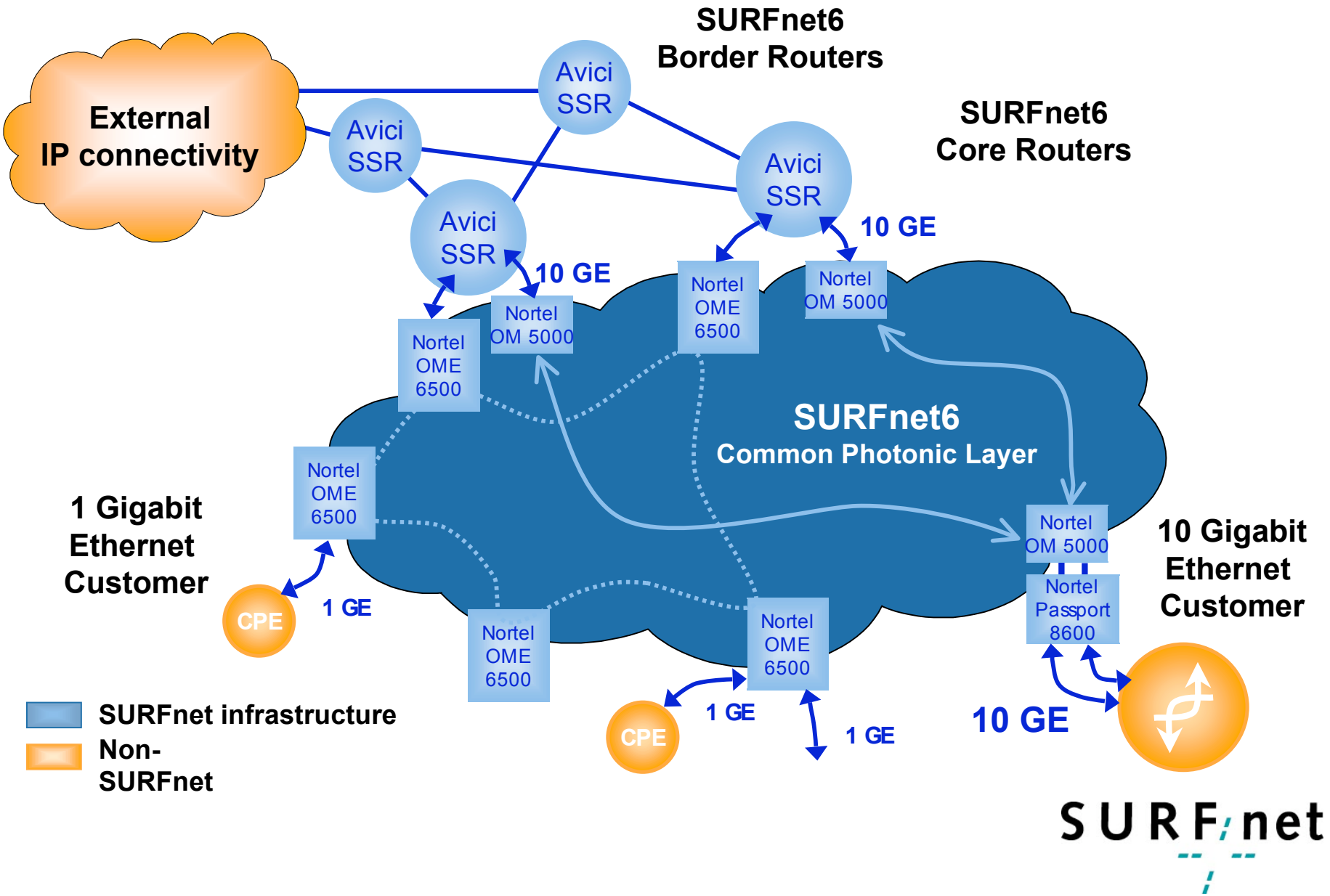
GigaPort

- SURFnet6 will be 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
- Managed dark fiber infrastructure will be extended with new routes, to be ready for SURFnet6

SURFnet

IP network implementation

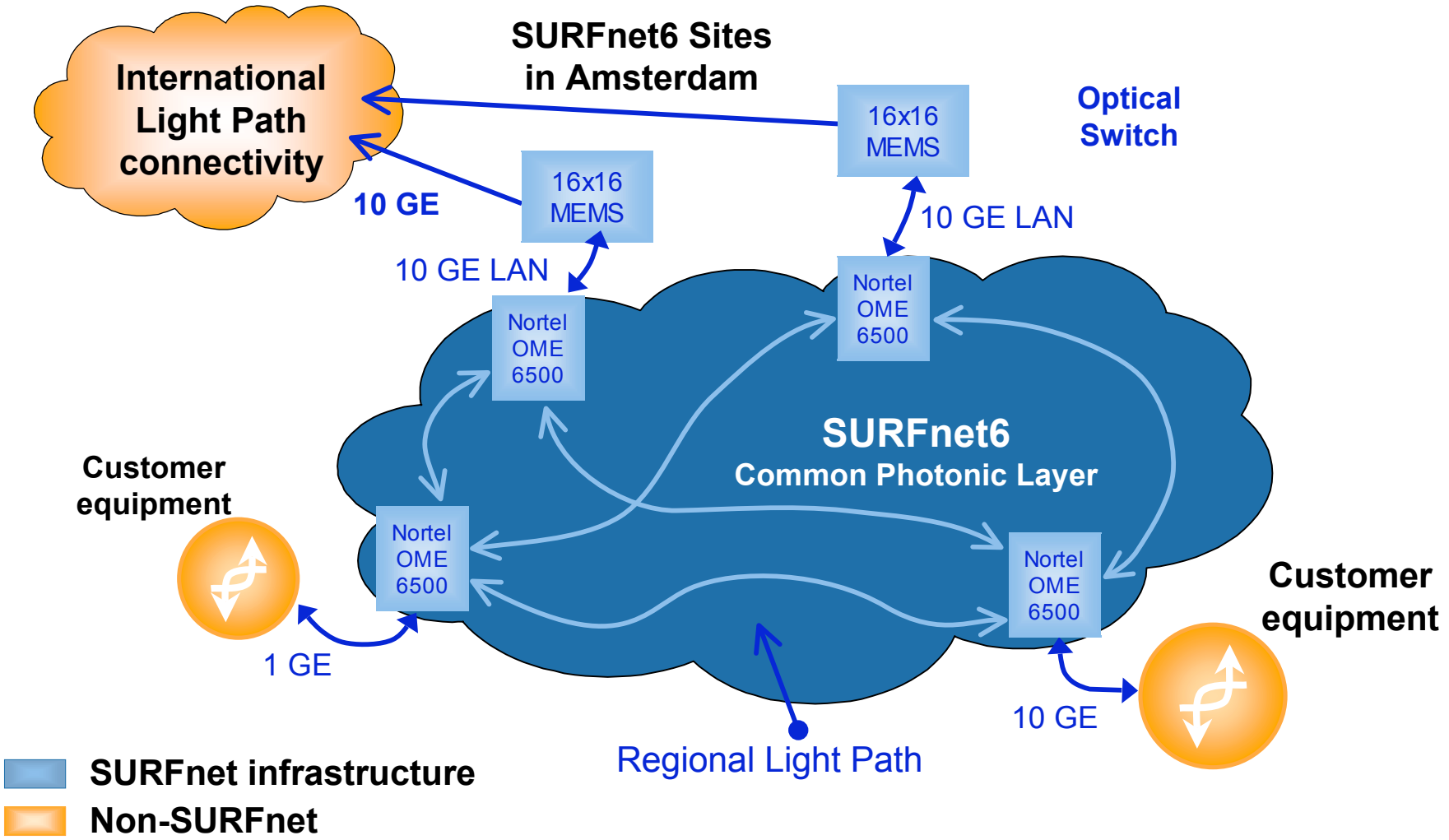
GigaPort



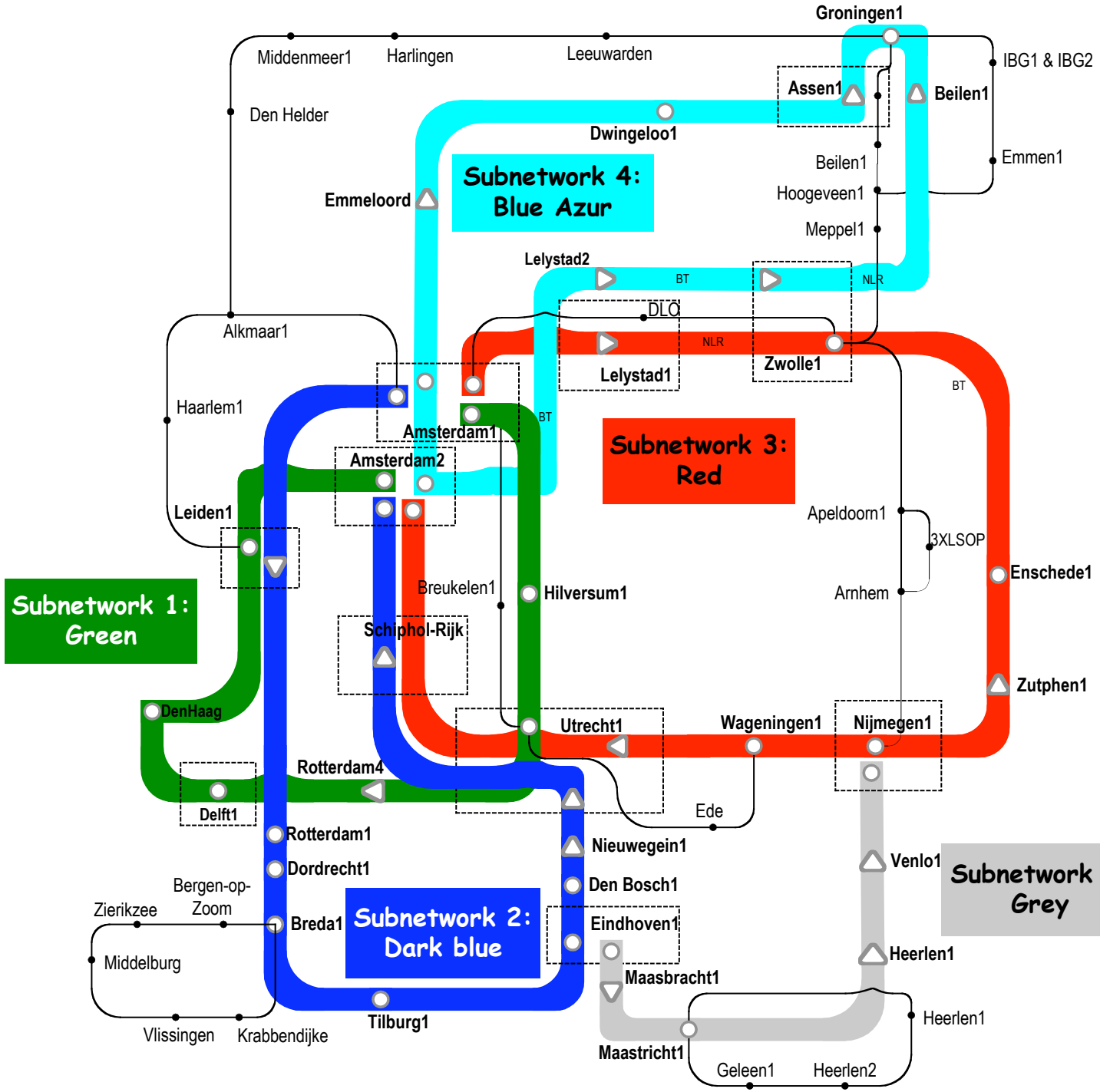
SURFnet

Light Paths provisioning implementation

GigaPort



SURFnet

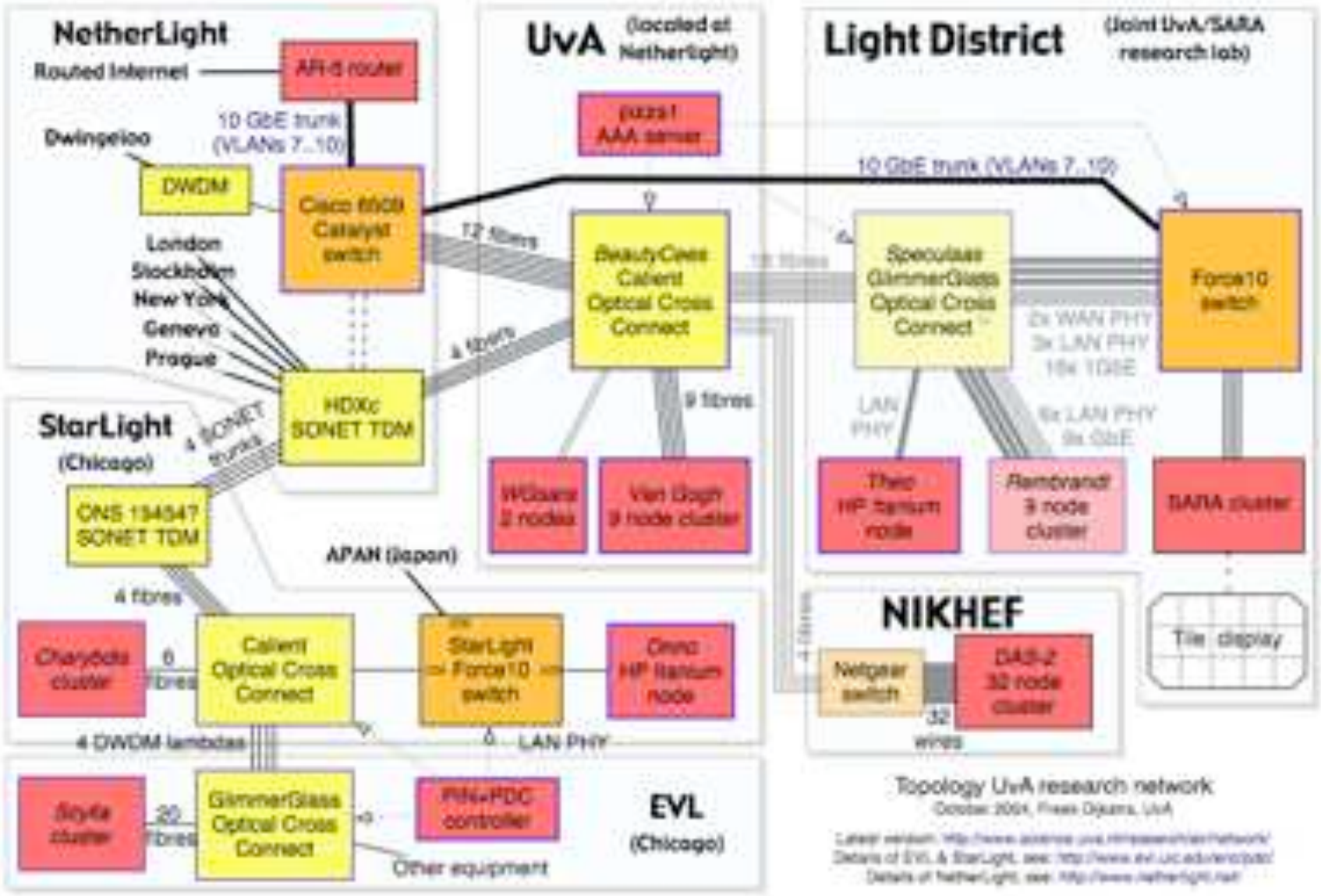


GigaPort

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

SURFnet is looking after your Lambda's in Science Park Amsterdam

Research on Networks (CdL)

GigaPort

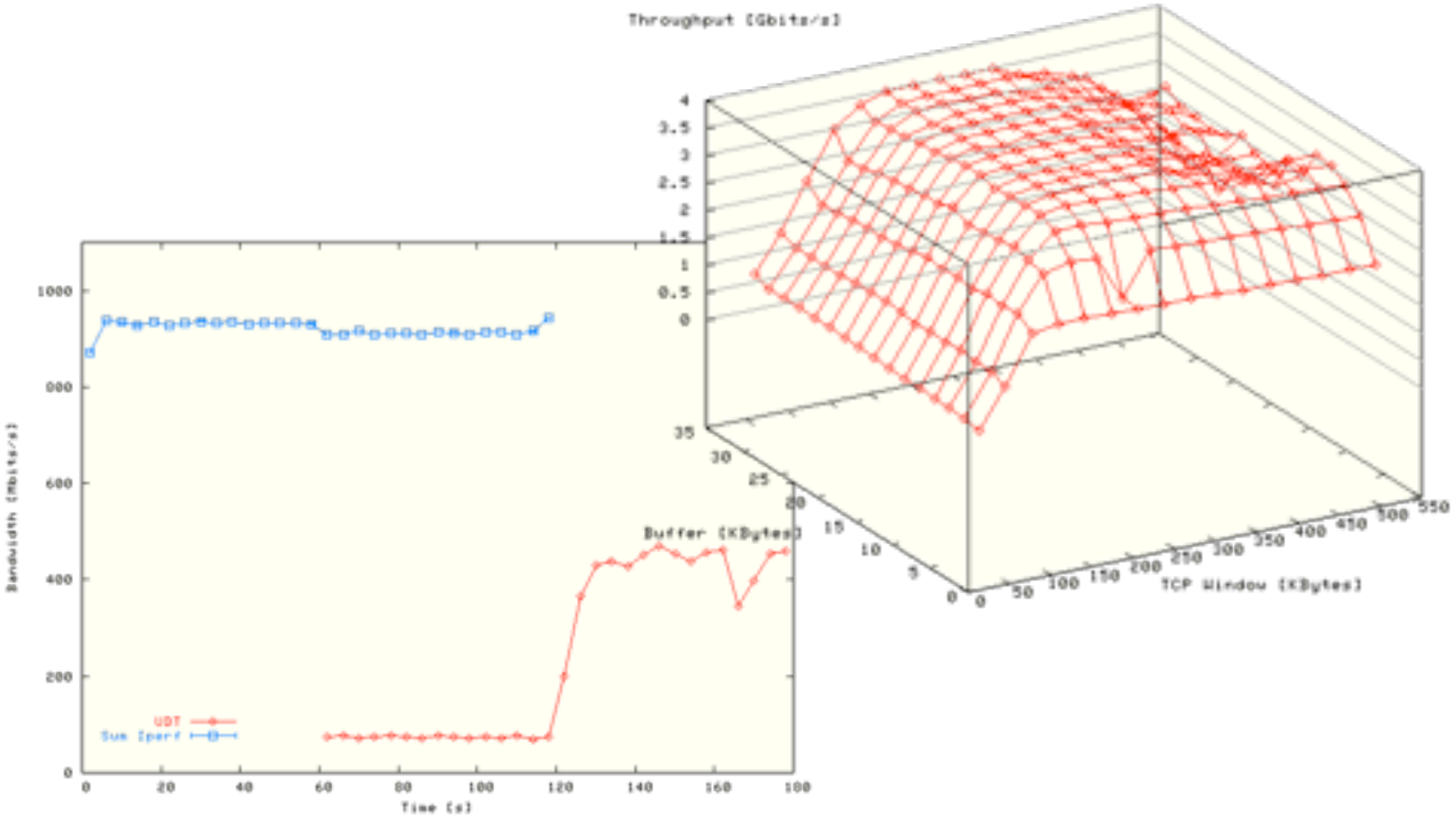
- **Optical Networking:** What innovation in architectural models, components, control and light path provisioning are needed to integrate dynamically configurable optical transport networks and traditional IP networks to a generic data transport platform that provides end-to-end IP connectivity as well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

- 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



Layer - 2 requirements from 3/4

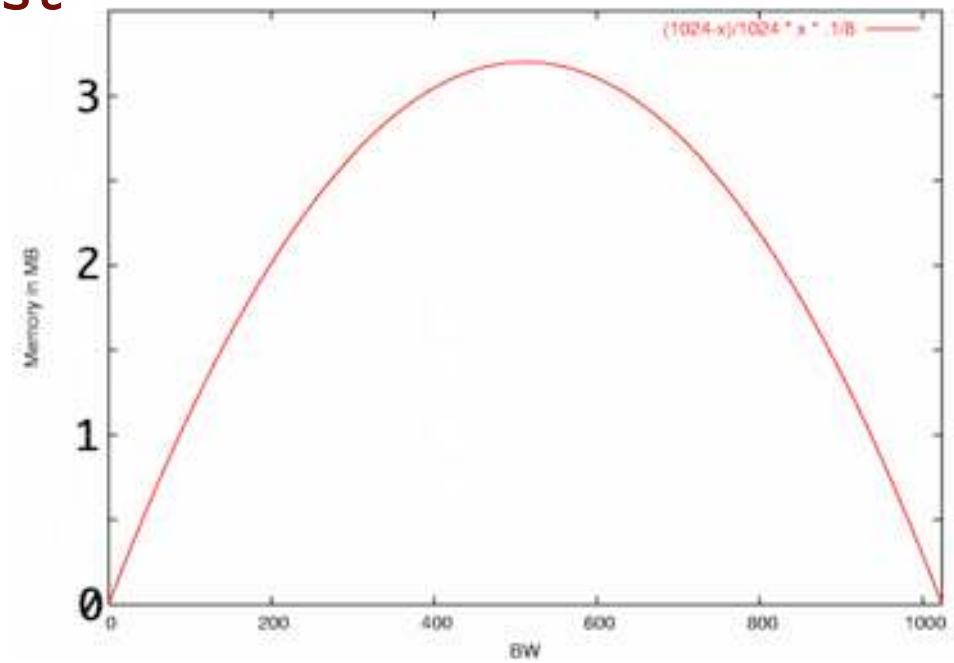
TCP is bursty due to sliding window protocol and slow start algorithm.

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

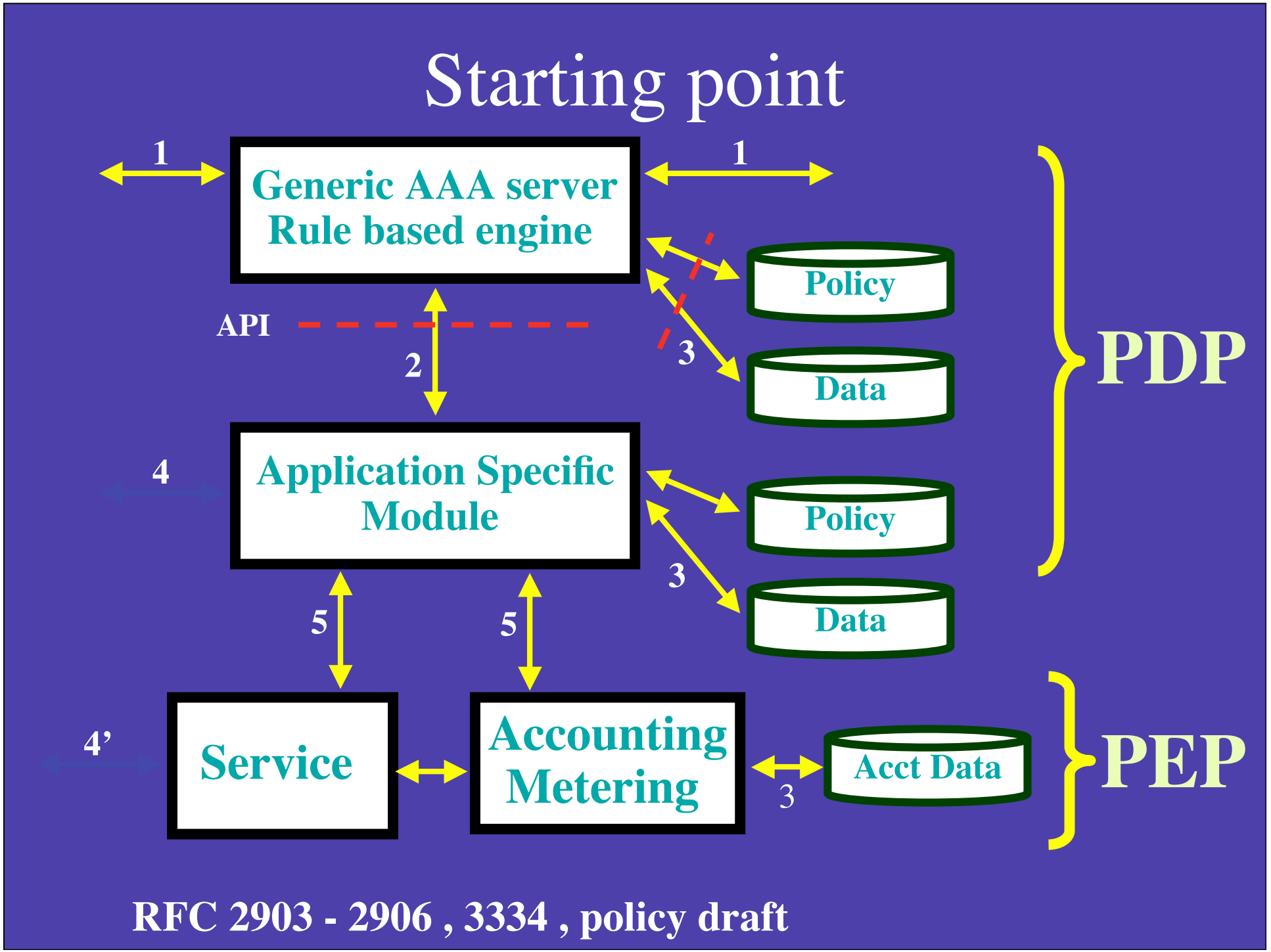
$$\text{Memory-at-bottleneck} = \frac{\text{fast} - \text{slow}}{\text{fast}} * \text{slow} * \text{RTT}$$

So pick from menu:

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

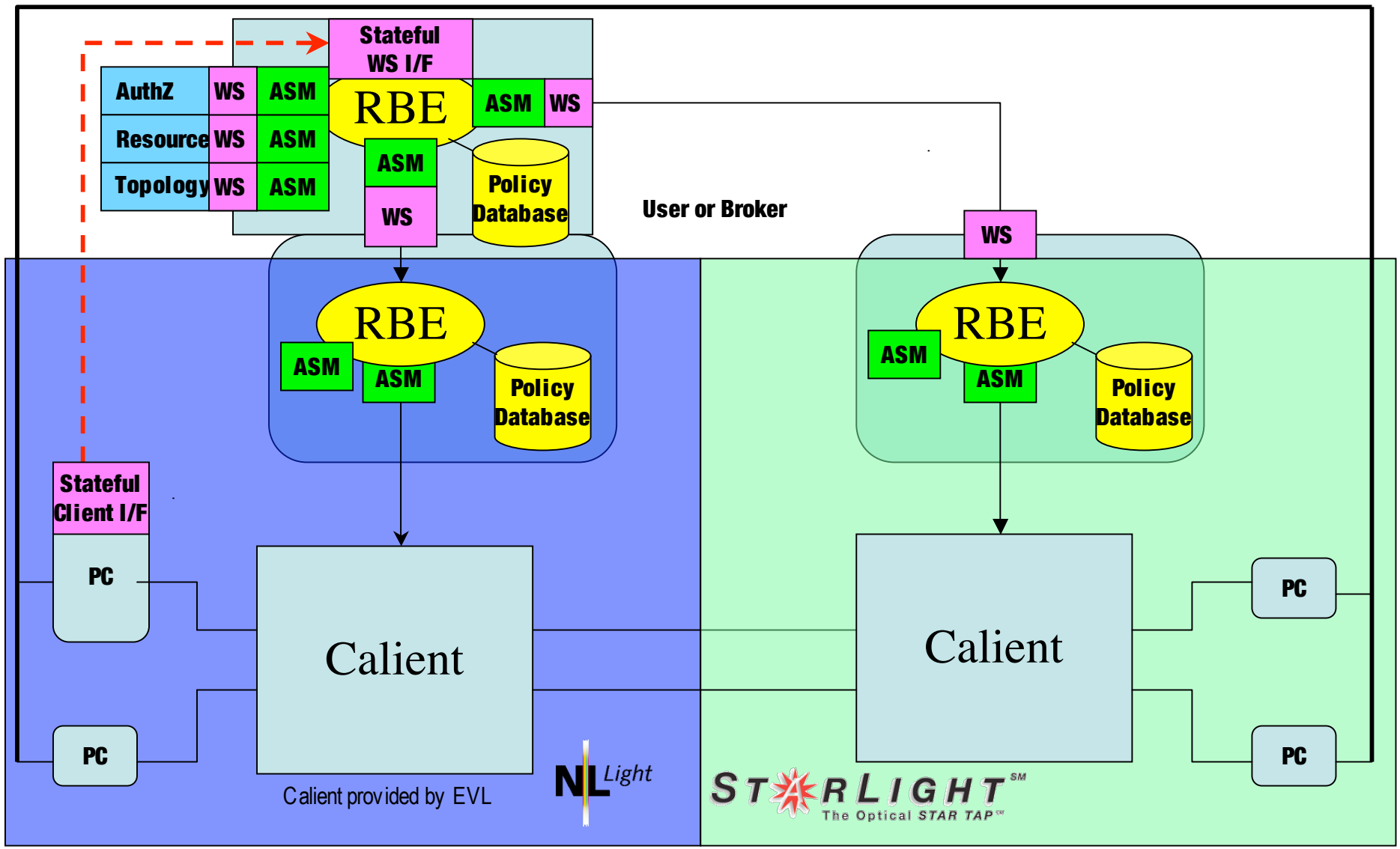


Starting point



RFC 2903 - 2906 , 3334 , policy draft

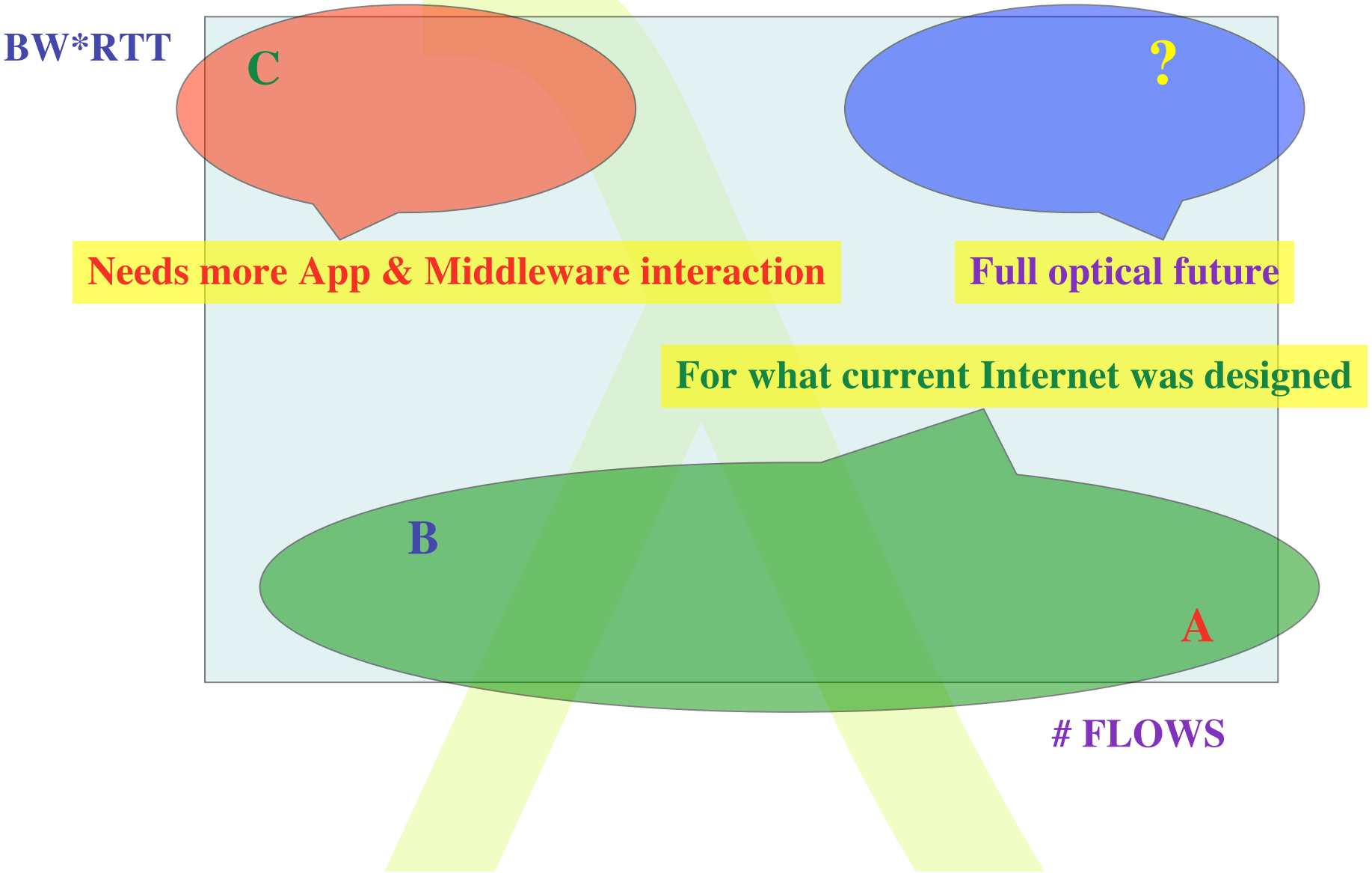
AAA based demo at SC2003



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

SURFnet

sara
Computing & Networking Services