Lambda-Grid developments

Global Lambda Integrated Facility

www.science.uva.nl/~delaat

Cees de Laat

SURFnet

EU

University of Amsterdam

SARA

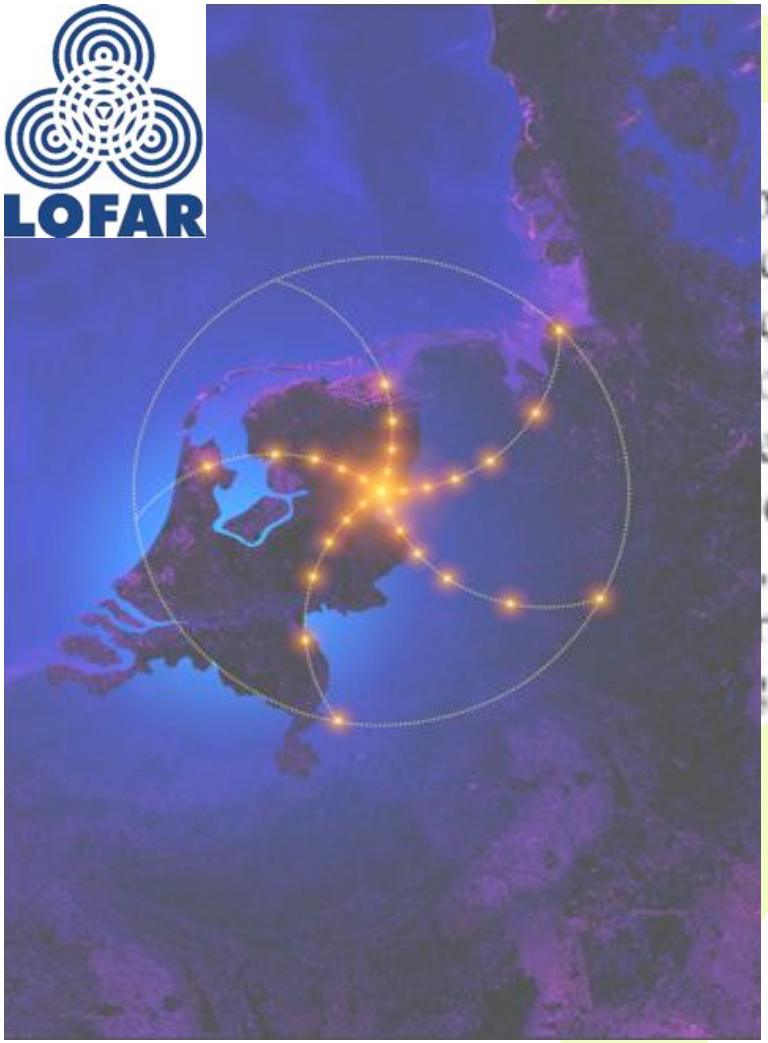
NCF

Contents

This page is intentionally left blank

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

Sensor Grids



~ 40 Tbit/s
www.lofar.org

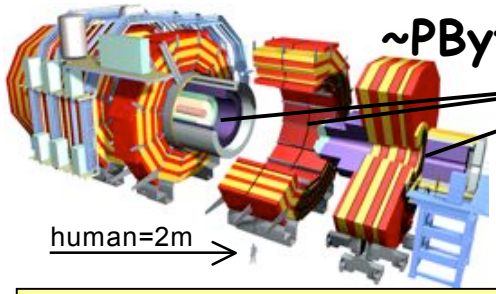
eVLBI

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 paralle
ed correlator will remain the most efficient approach
olves dist
multi-gig
elator and
factor.

Westerbork Synthesis Radio Telescope - Netherlands

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

analysis

Tier2 Center

Tier 2

~0.6-2.5 Gbps

Tier 3

Institute ~0.25TIPS

Physics data cache

100 - 1000 Mbits/sec

Tier 4

Workstations

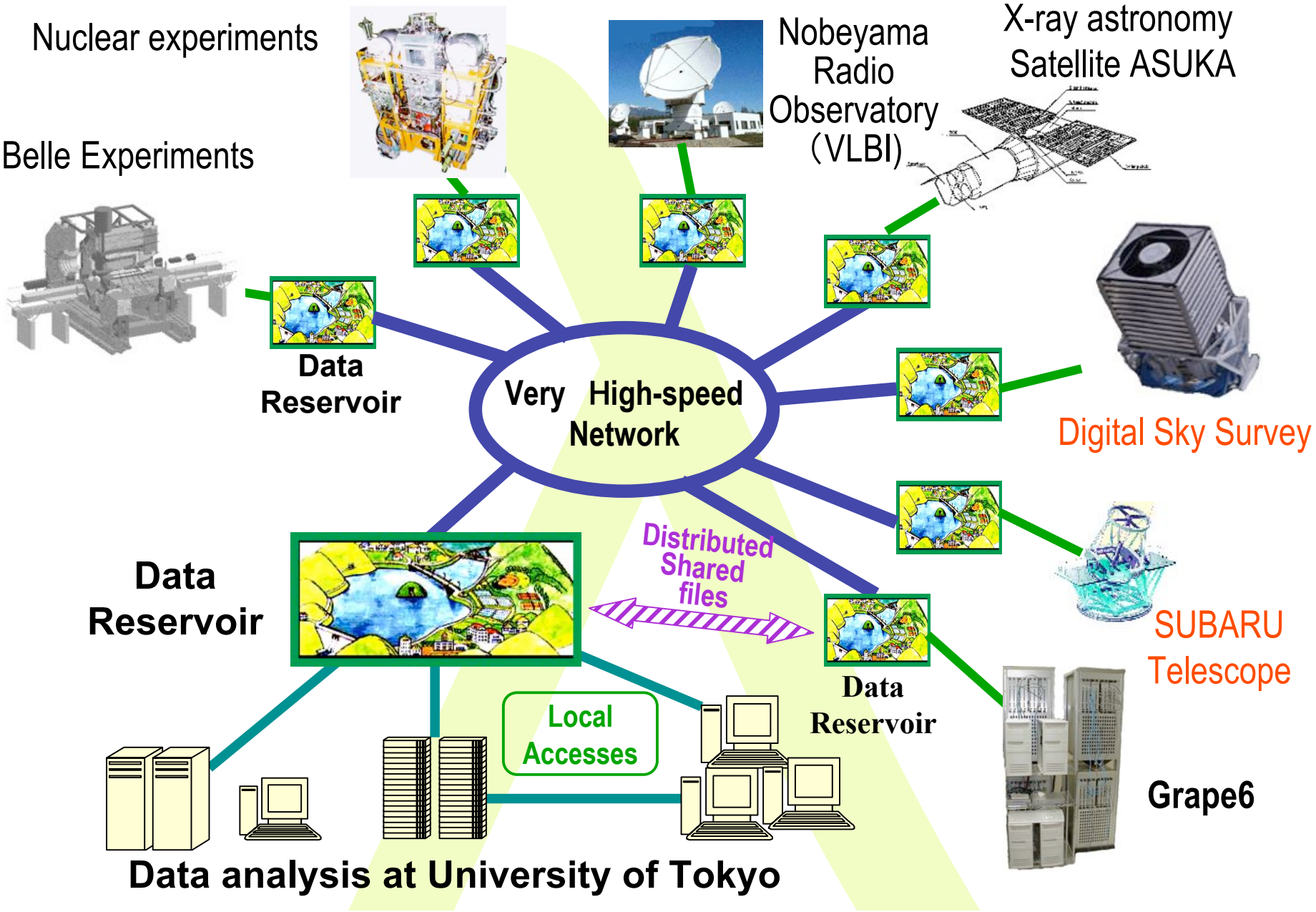
Courtesy Harvey Newman, CalTech and CERN

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

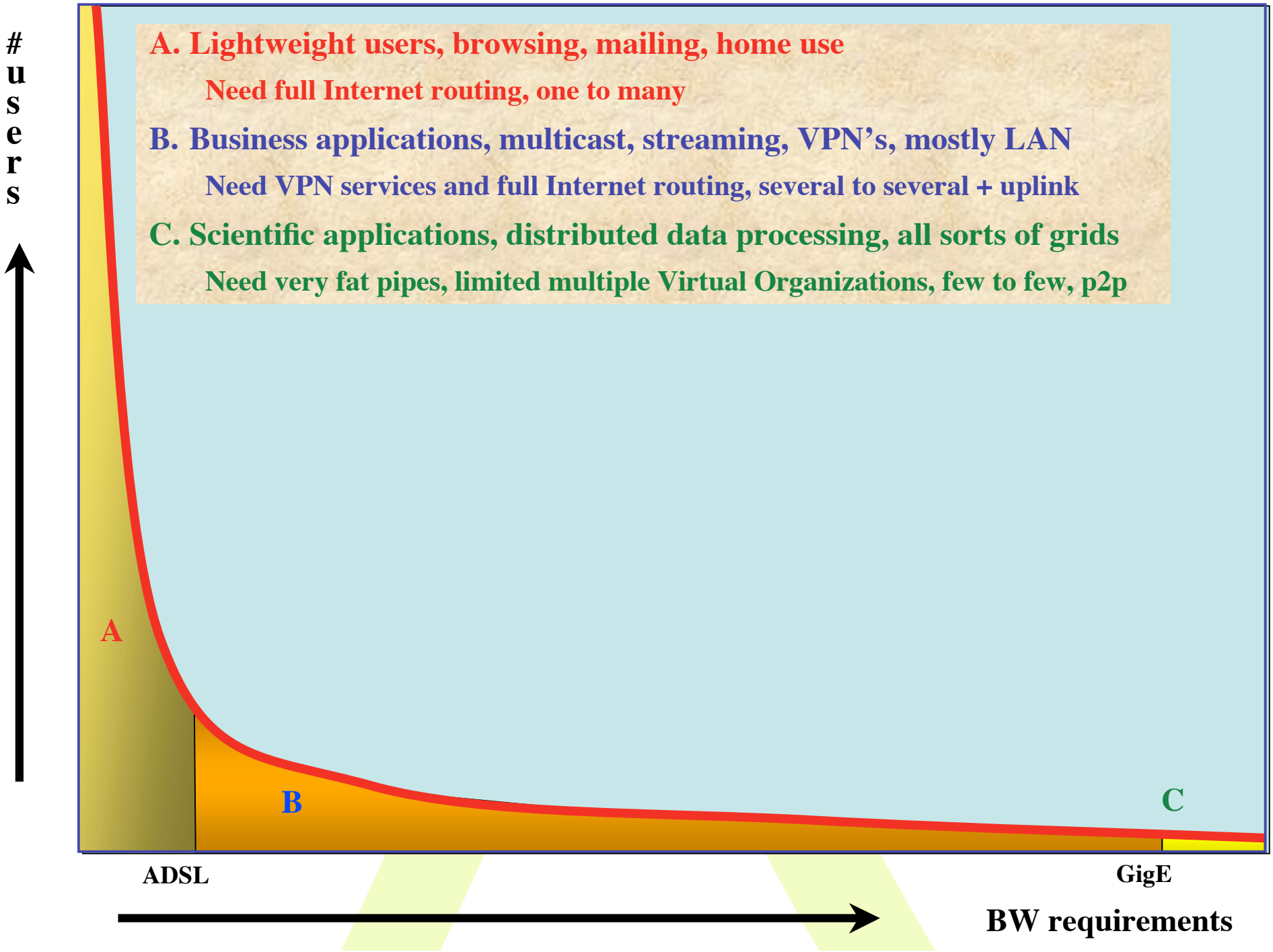


Showed you 4 types of Grids

- 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

Add another one:

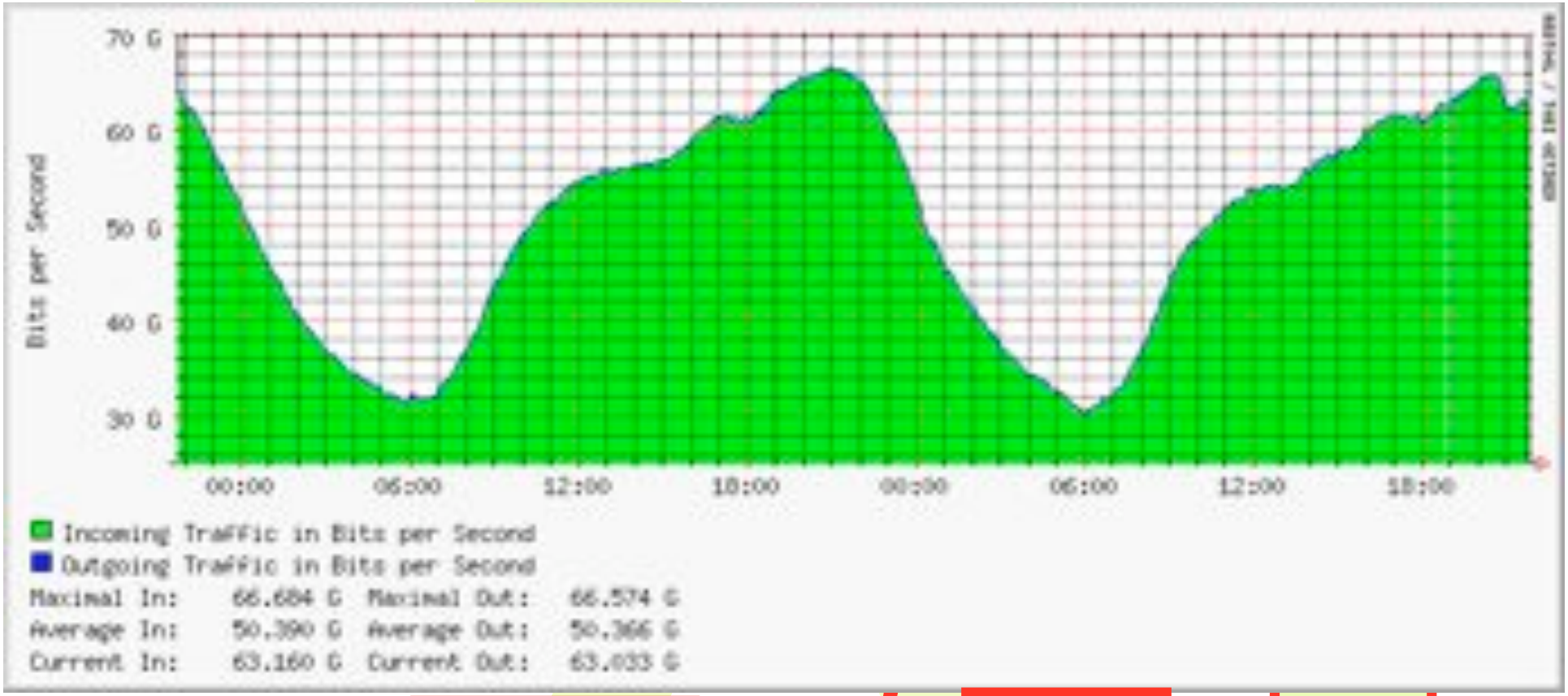
- Lambda Grids
  - Hybrid networks


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 (in 2005)

- **Estimate A**
  - **17 M people, 6.4 M households, 25 % penetration of 0.5 - 8 Mb/s ADSL, 40 times under-provisioning ==> ~ 40 Gb/s**

AMS-IX



June 19th 2004

May 2005

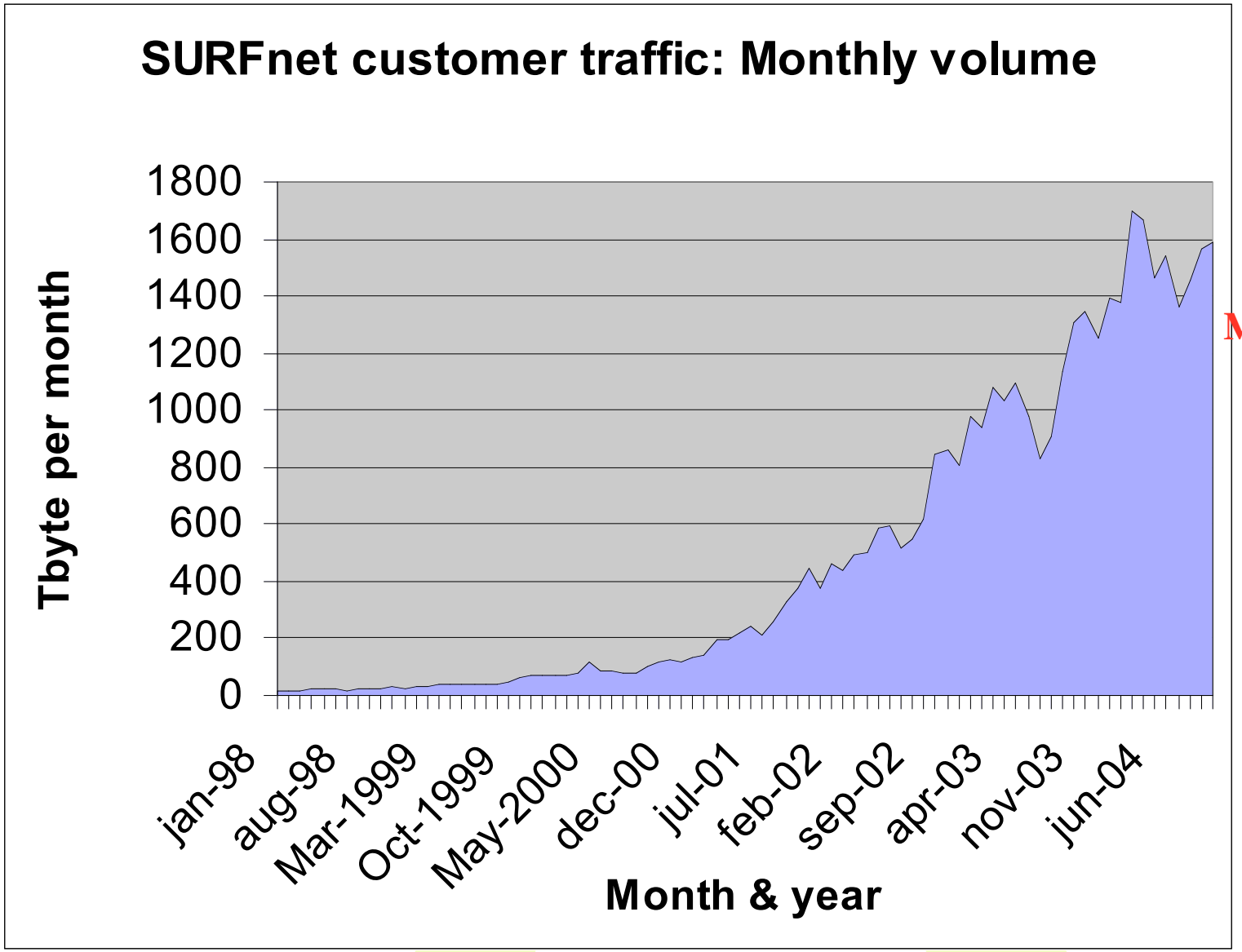
Lost :-(

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

The Dutch Situation (in 2005)

- **Estimate A**
  - **17 M people, 6.4 M households, 25 % penetration of 0.5 - 8 Mb/s ADSL, 40 times under-provisioning ==> ~ 40 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**

Routed L3 traffic growth



March 2005

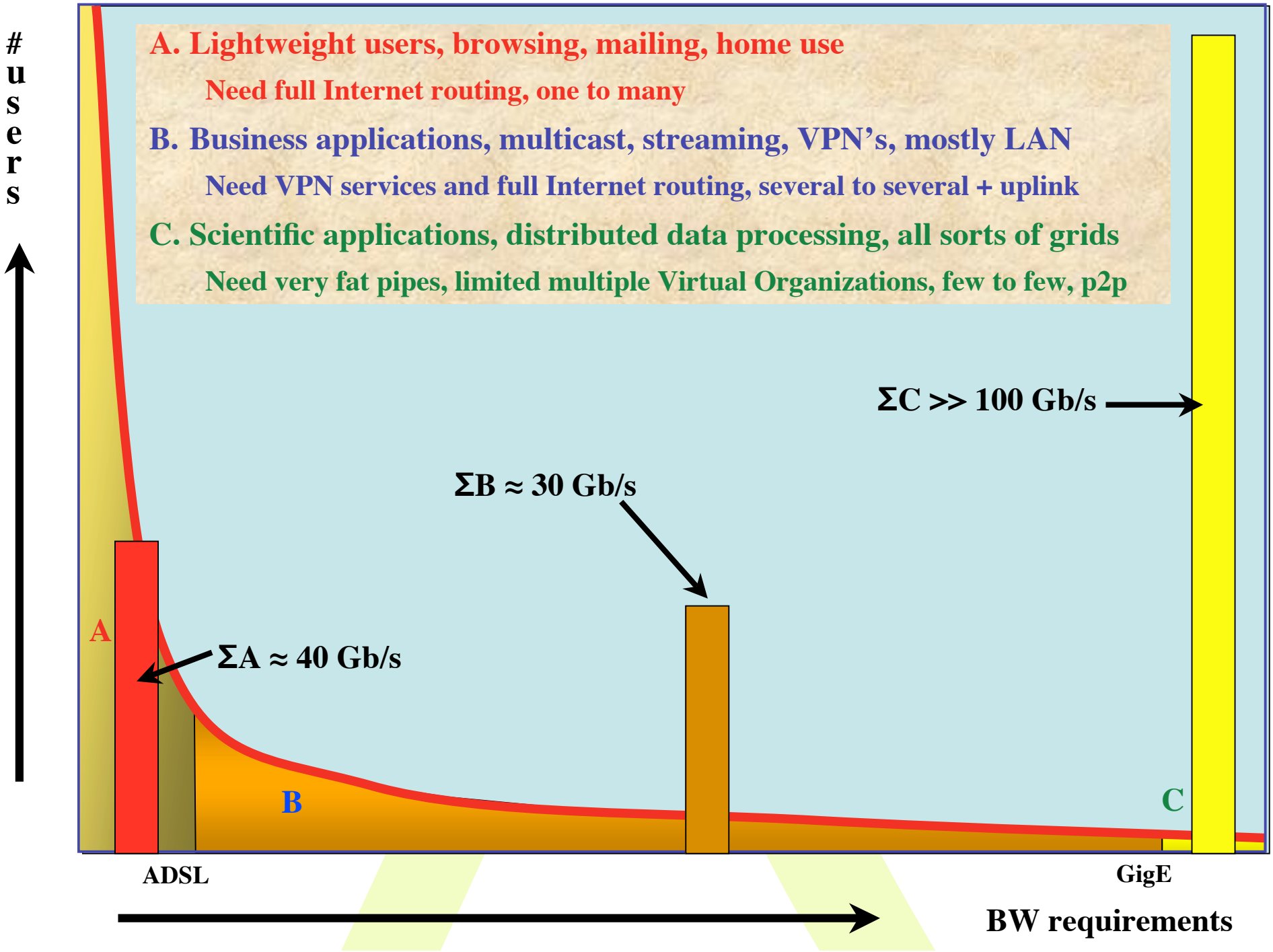
1.9 TByte

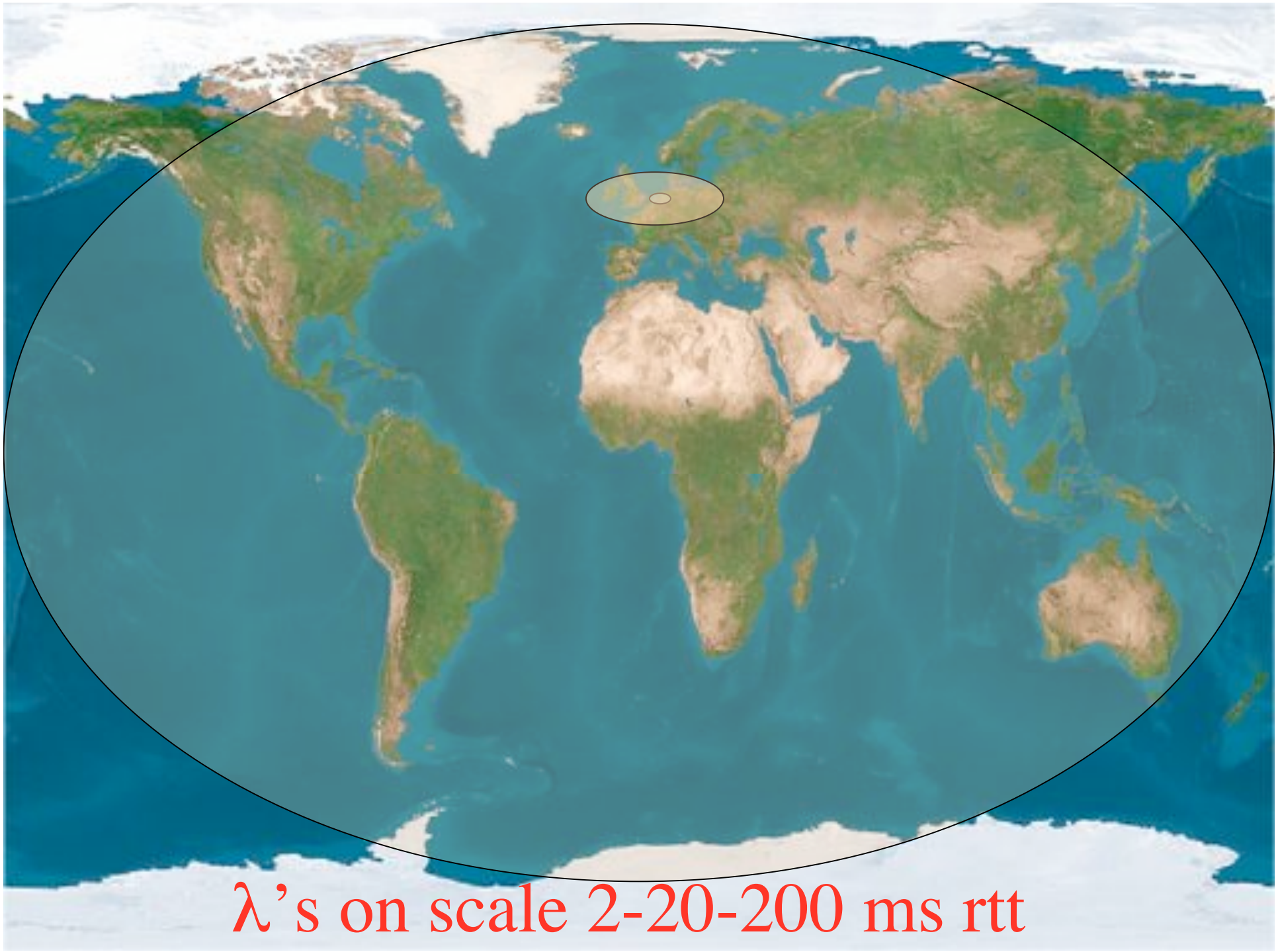
1900 TByte/month ≈ 6 Gbits/second

Slide courtesy Kees Neggers

The Dutch Situation (in 2005)

- **Estimate A**
  - **17 M people, 6.4 M households, 25 % penetration of 0.5 - 8 Mb/s ADSL, 40 times under-provisioning ==> ~ 40 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 ≈ 40 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 -> 75-300 k$, 10G switch port -> 7-15 k$, MEMS port -> 1 k$**
  - **DWDM lasers for long reach expensive, 10-50 k$**
- **Bottom line: look for a hybrid architecture which serves all classes in a cost effective way ==> map A -> L3 , B -> L2 , C -> L1**
- **Give each packet in the network the service it needs, but no more !**

L1 ≈ 1 k$/port

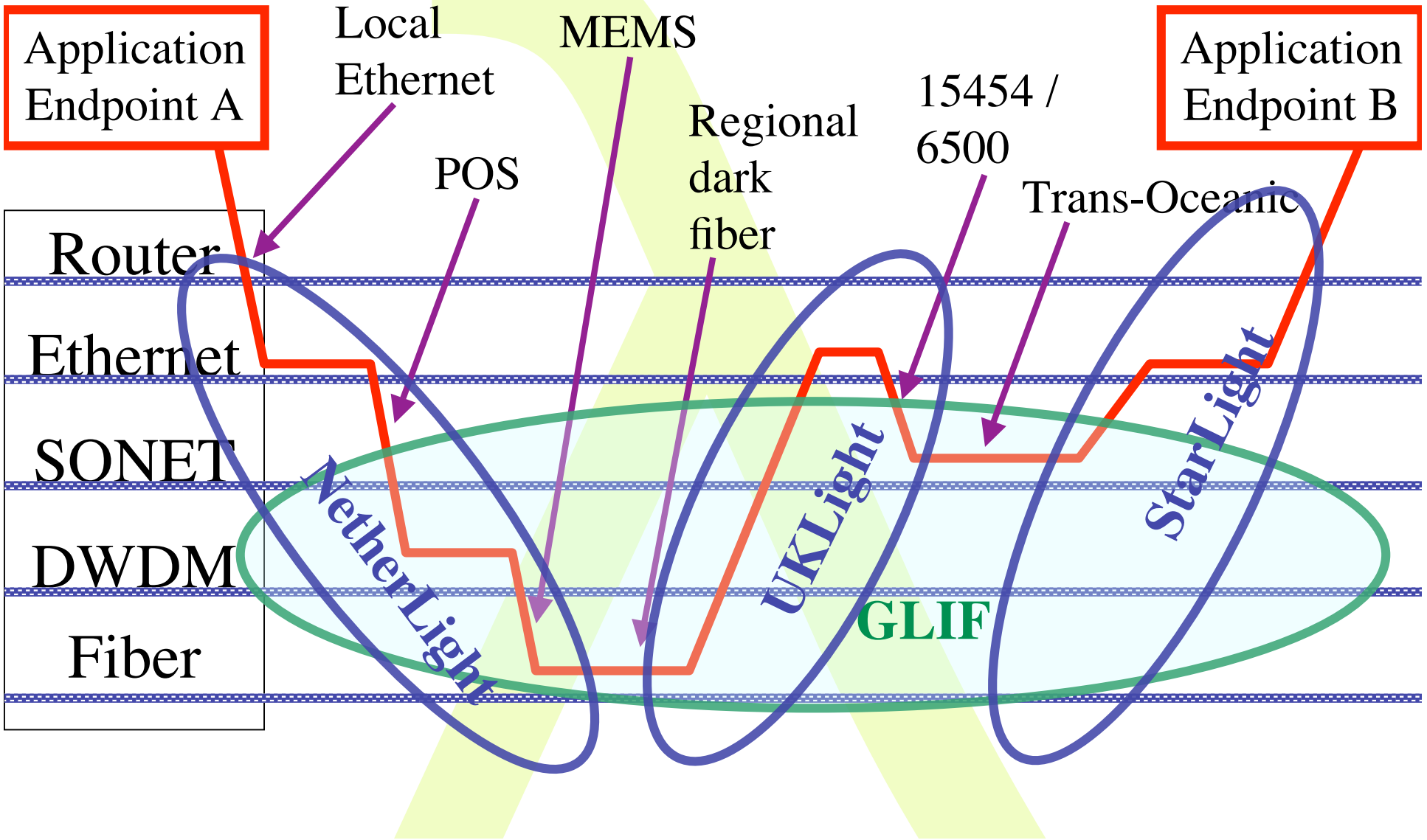
L2 ≈ 5-10 k$/port

L3 ≈ 75+ k$/port

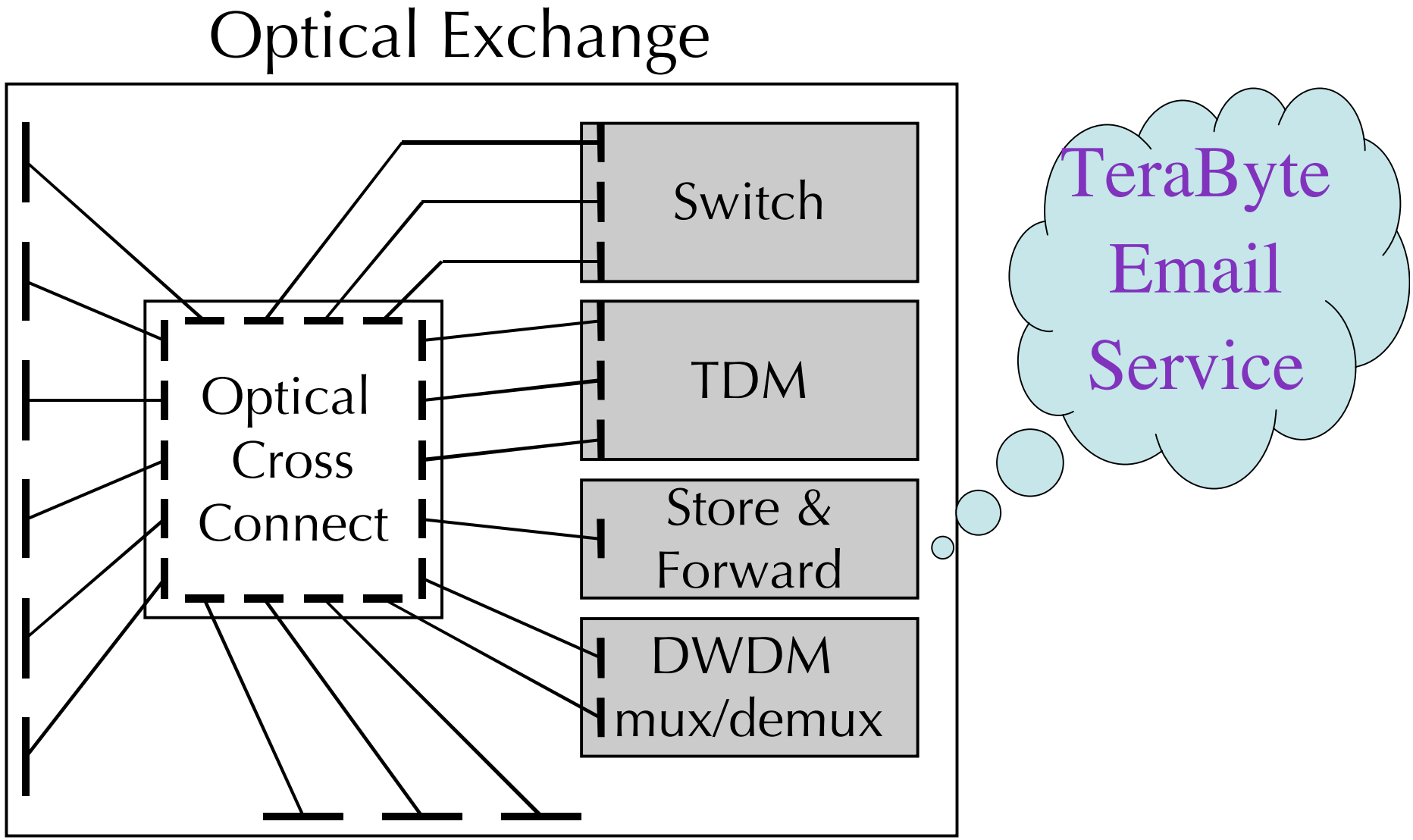
Services

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

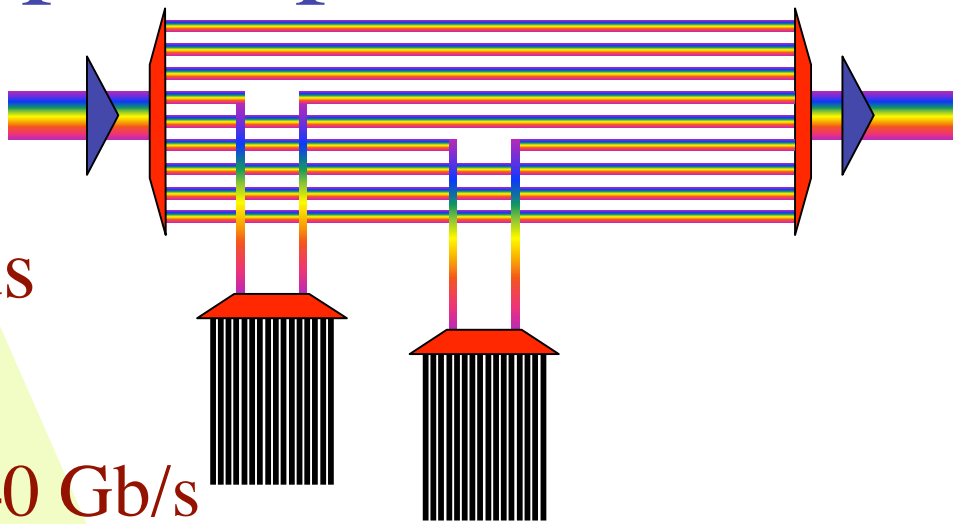
How low can you go?



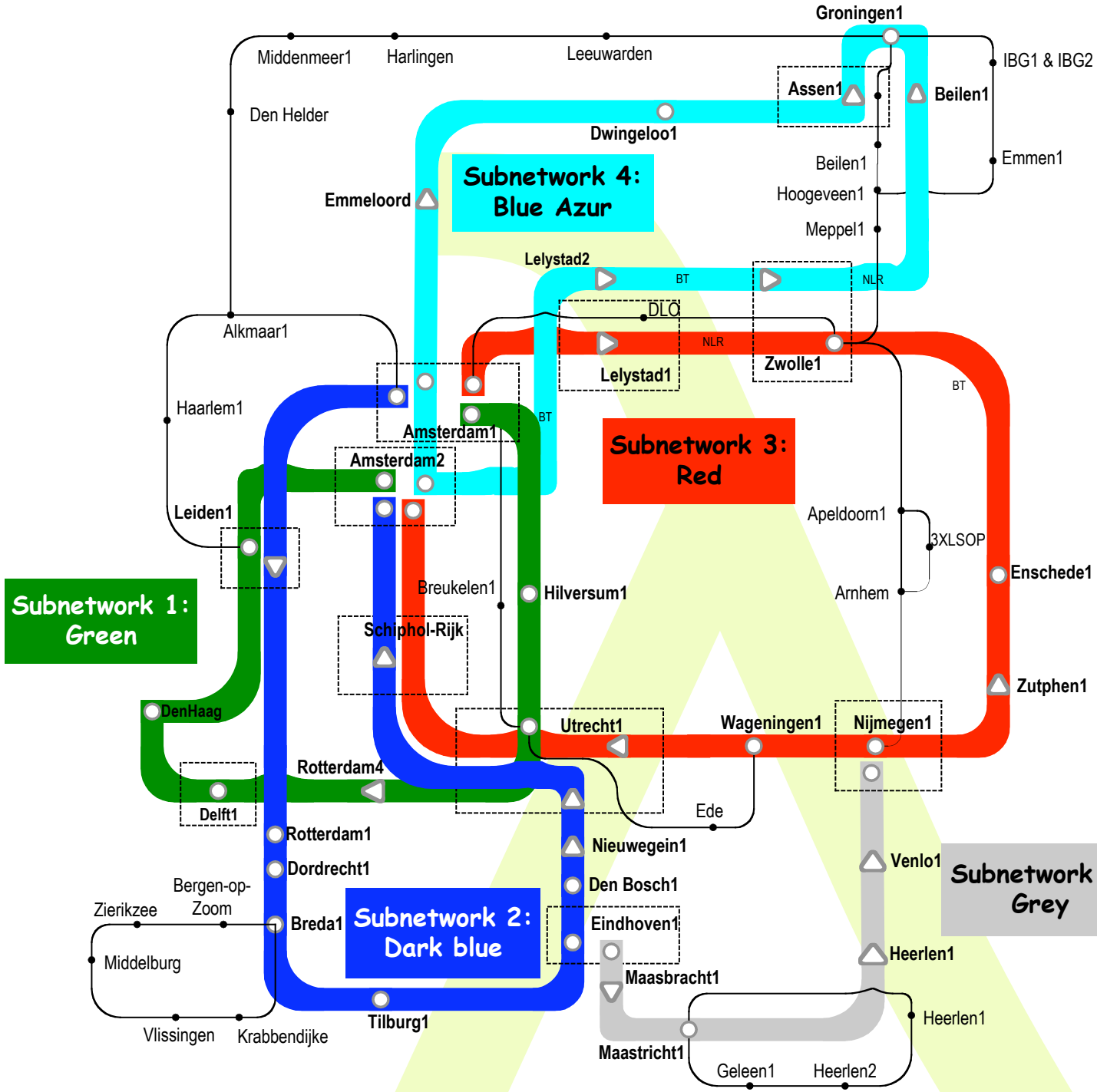
Optical Exchange as Black Box



SURFnet 6 principles

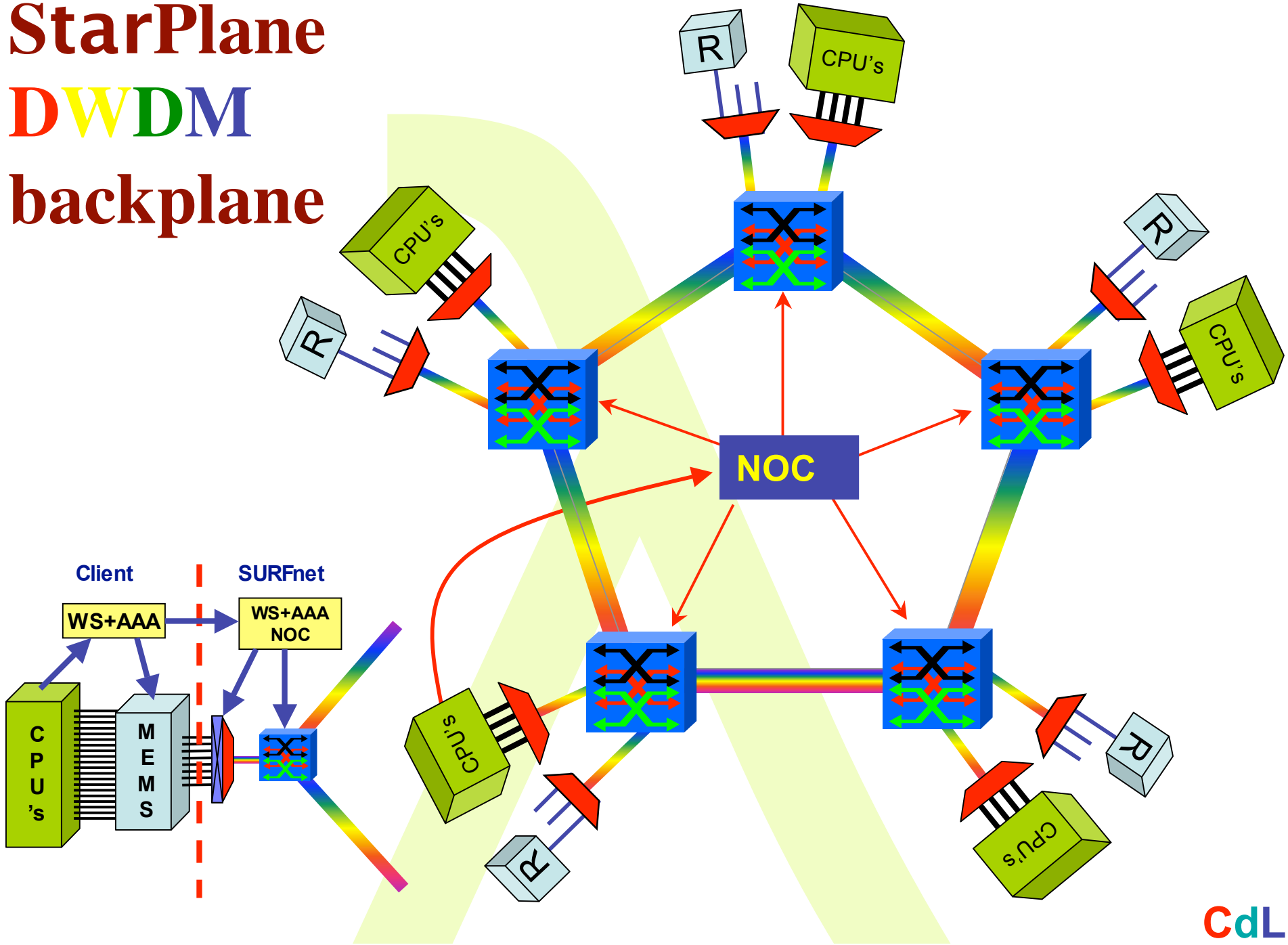


- Based on dark fiber
- 4 DWDM rings of 9 bands
  - each 4, later 8, colors
  - Each capable of 10, later 40 Gb/s
- Universities have POP's on ring, each 1 band
- Connect with 1 or 10 Gb/s Ethernet
- Routing in Amsterdam in 2 core POP's!
- International connectivity in Amsterdam
- Lambda service between ring POP's and to NetherLight



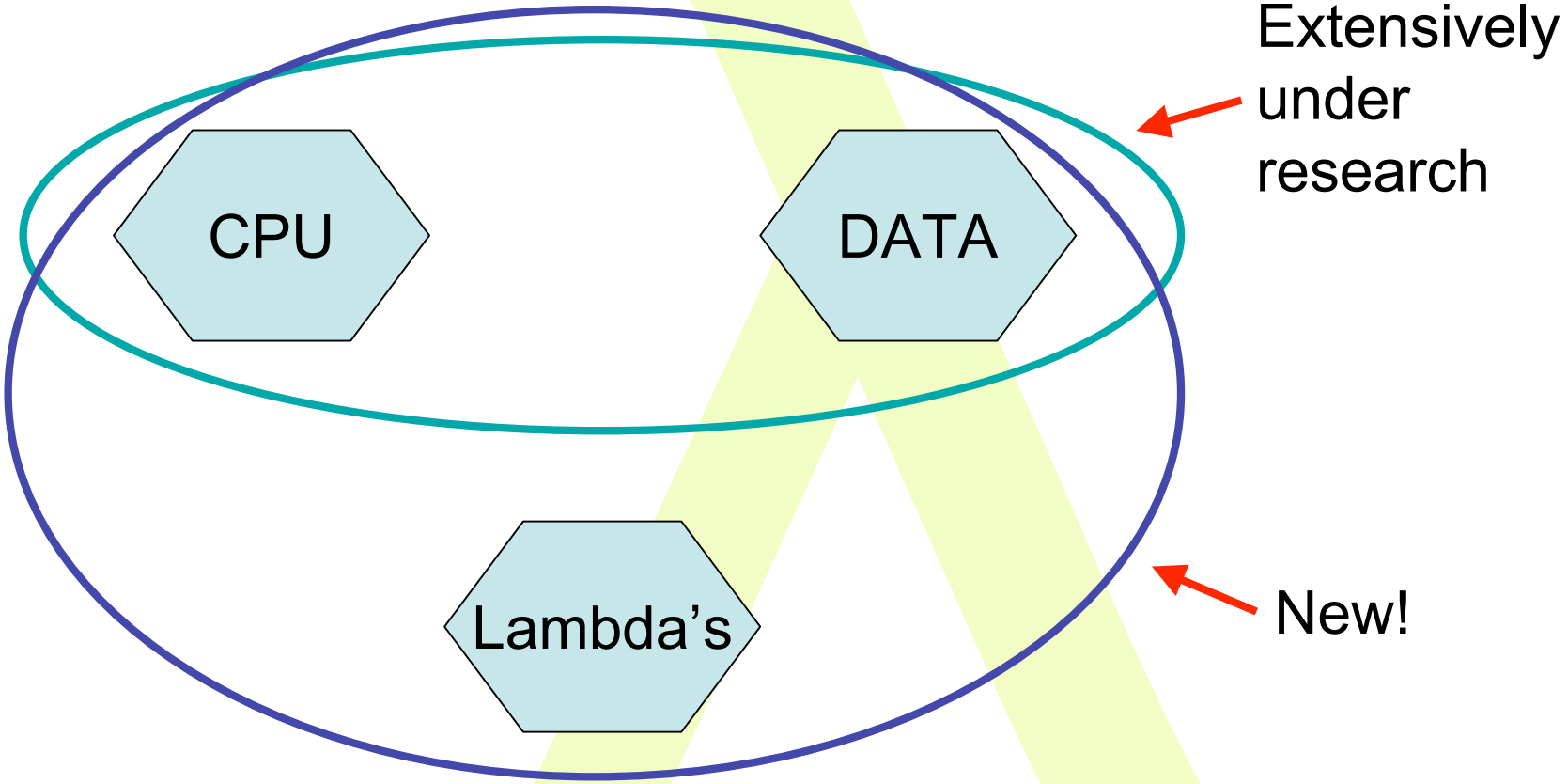
Common Photonic Layer (CPL) in SURFnet6

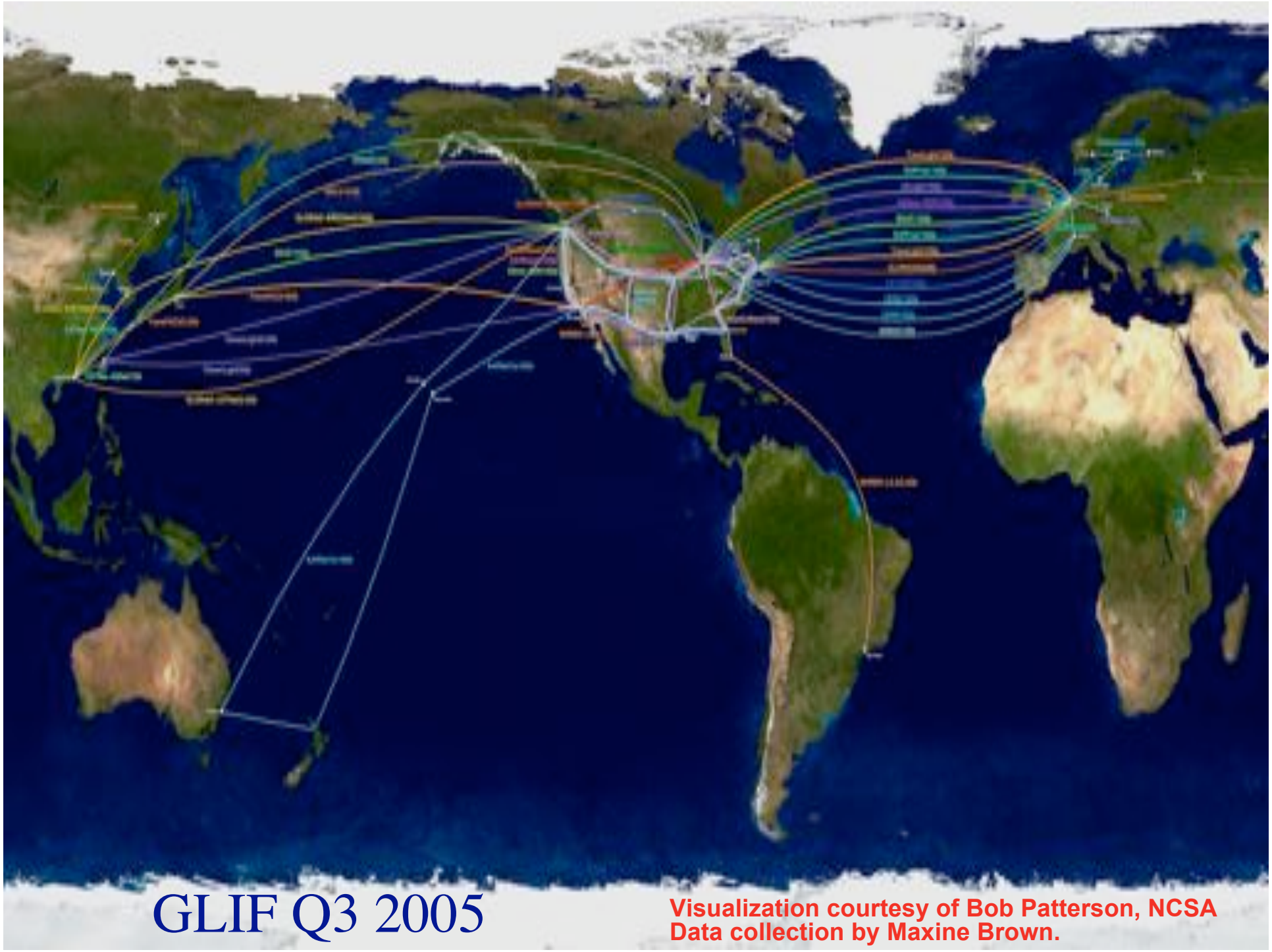
StarPlane
DWDM
backplane



CdL

GRID-Colocation problem space





GLIF Q3 2005

Visualization courtesy of Bob Patterson, NCSA
Data collection by Maxine Brown.

GLIF Mission Statement

- **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 is an environment (networking infrastructure, network engineering, system integration, middleware, applications) to accomplish real work**

Working groups

GLIF Governance and policy

Our small-scale Lambda Workshop is now turning into a global activity. TransLight and similar projects contribute to the infrastructure part of GLIF. A good and well understood governance structure is key to the manageability and success of GLIF. Our prime goal is to decide upon and agree to the GLIF governance and infrastructure usage policy.

GLIF Lambda infrastructure and Lambda exchange implementations

A major function for previous Lambda Workshops was to get the network engineers together to discuss and agree on the topology, connectivity and interfaces of the Lambda facility. Technology developments need to be folded into the architecture and the expected outcome of this meeting is an agreed view on the interfaces and services of Lambda exchanges and a connectivity map of Lambdas for the next year, with a focus on iGrid 2005 and the emerging applications.

Persistent Applications

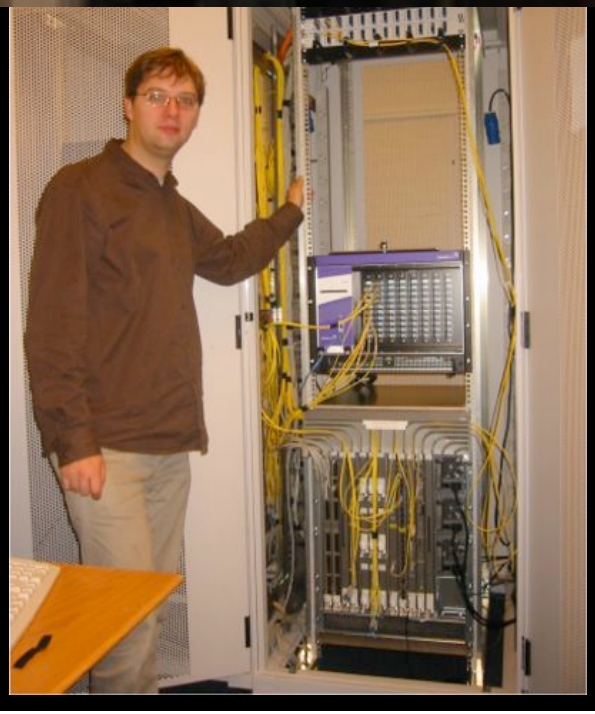
Key to the success of the GLIF effort is to connect the major applications to the Facility. We, therefore, need a list of prime applications to focus on and a roadmap to work with those applications to get them up to speed. The demonstrations at SC2004 and iGrid 2005 can be determined in this meeting.

Control Plane and Grid Integration

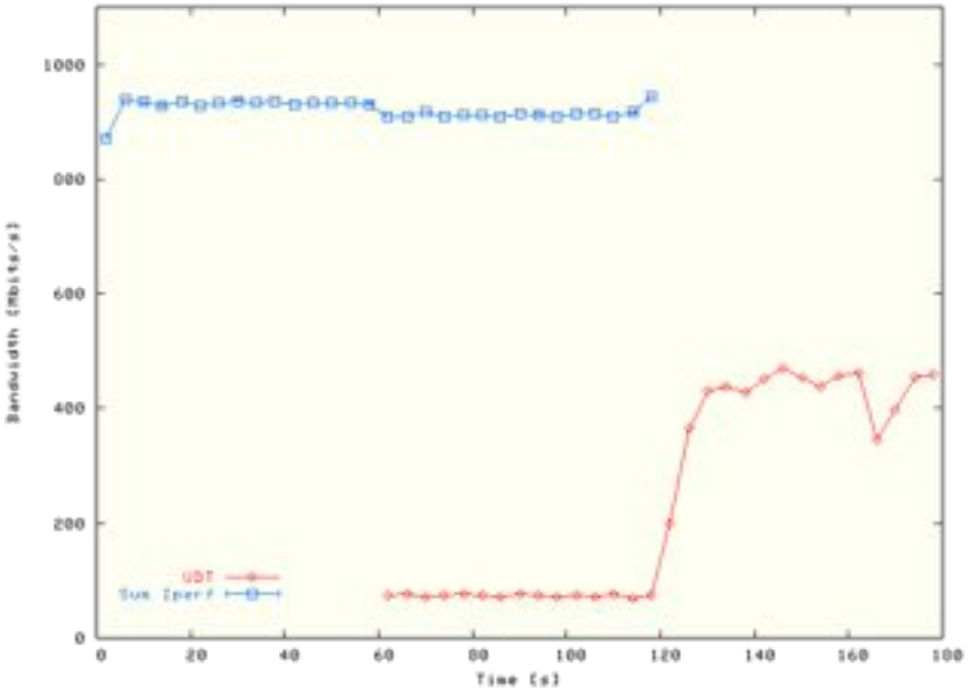
The GLIF can only function if we agree on the interfaces and protocols that talk to each other in the control plane on the contributed Lambda resources. The main players in this field are already meeting, almost on a bi-monthly schedule. Although not essential, this GLIF meeting could also host a breakout session on control plane middleware.

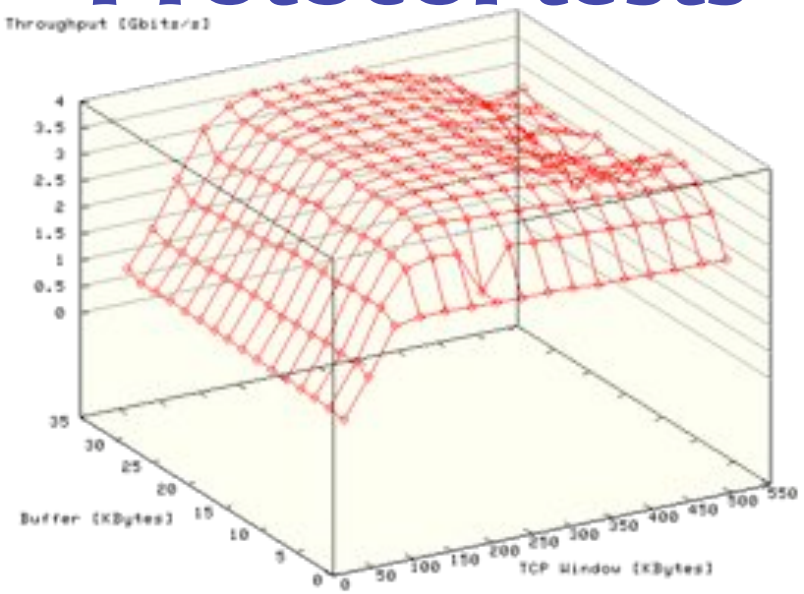
- Optical networking architectures and models
  - Optical Internet Exchange architecture
  - Lambda routing and assignment
- IP transport protocols, performances monitoring and measurements
  - With respect to performance
  - Monitoring and reporting
  - Traffic generation with grid infrastructure
-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

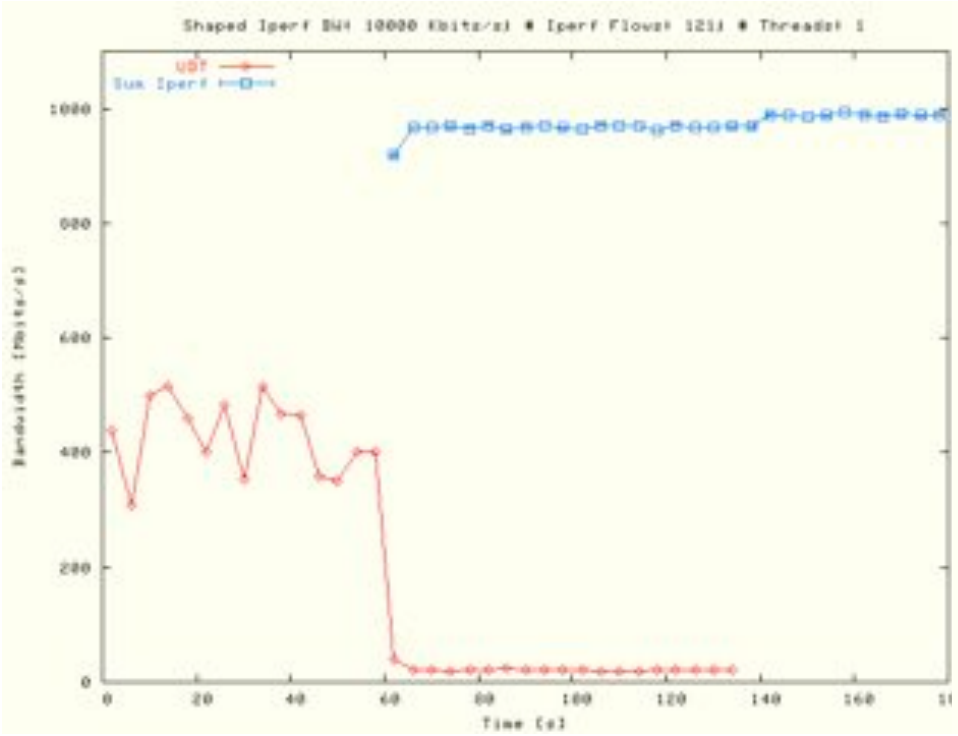
LightHouse

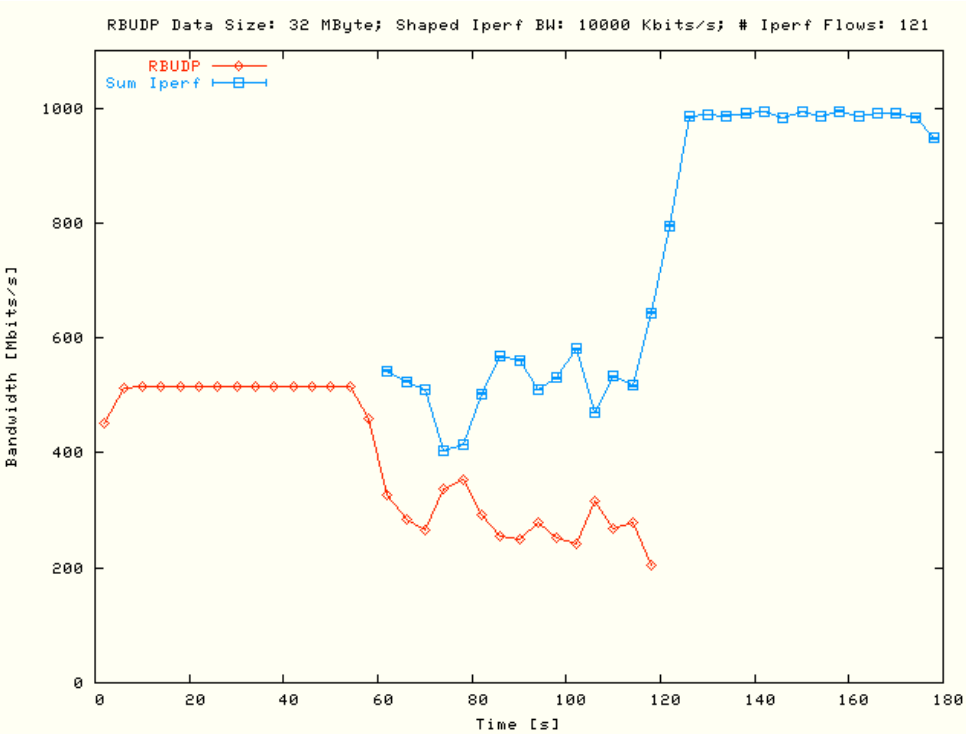


Protocol tests









Layer - 2 requirements from 3/4

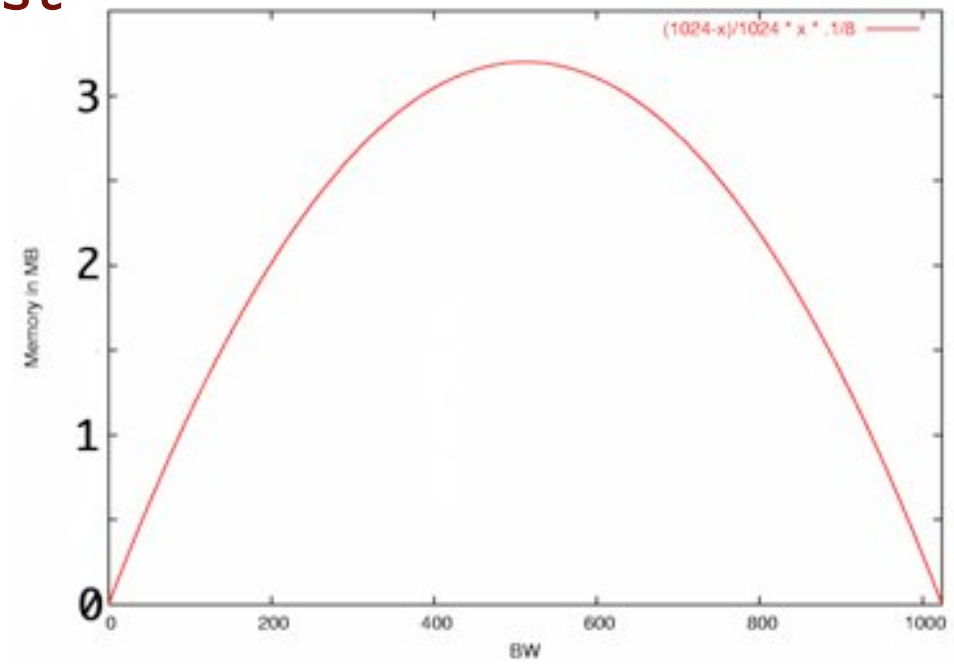
TCP is bursty due to sliding window protocol and slow start algorithm.

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

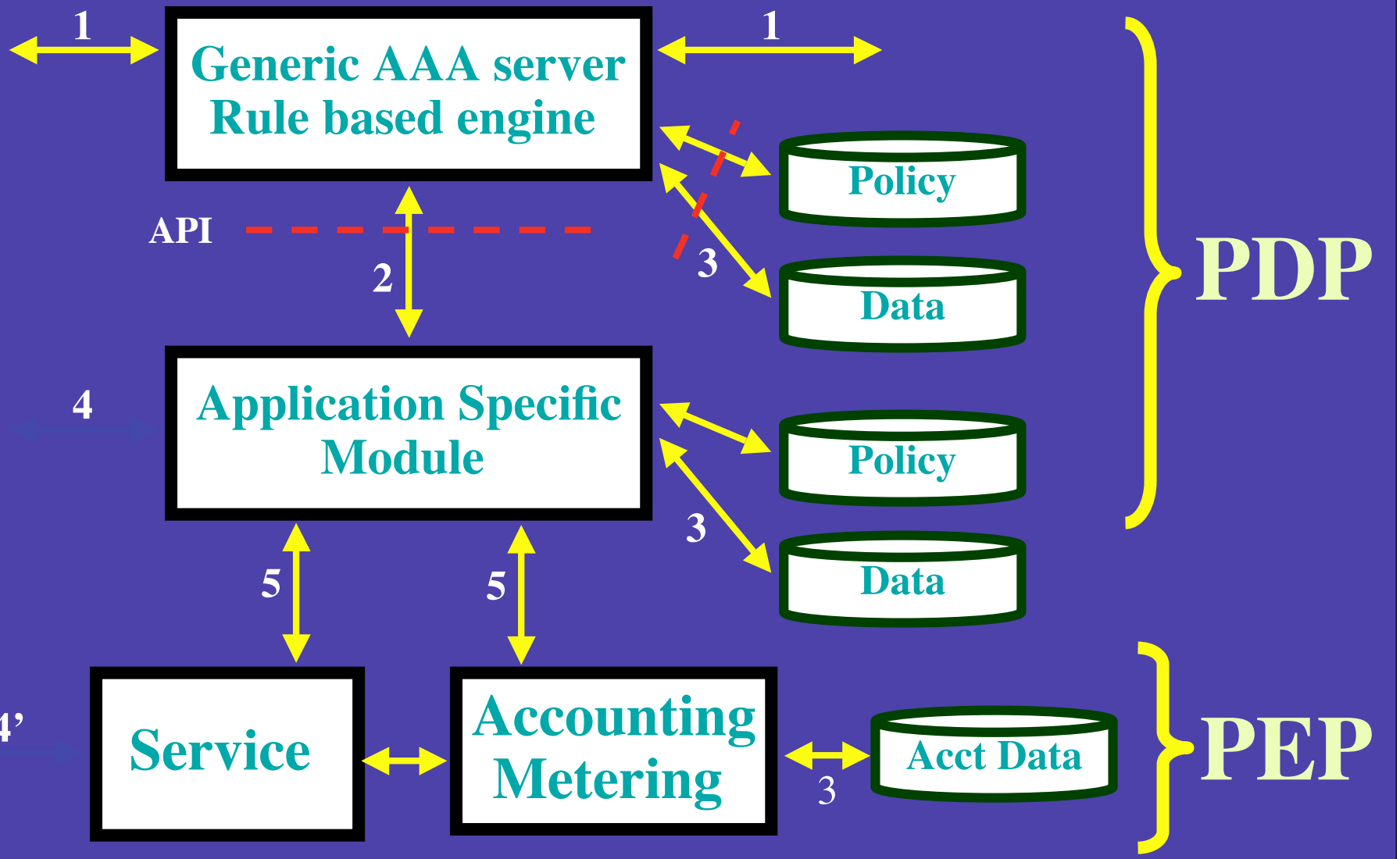
$$\text{Memory-at-bottleneck} = \frac{\text{fast} - \text{slow}}{\text{fast}} * \text{slow} * \text{RTT}$$

So pick from menu:

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

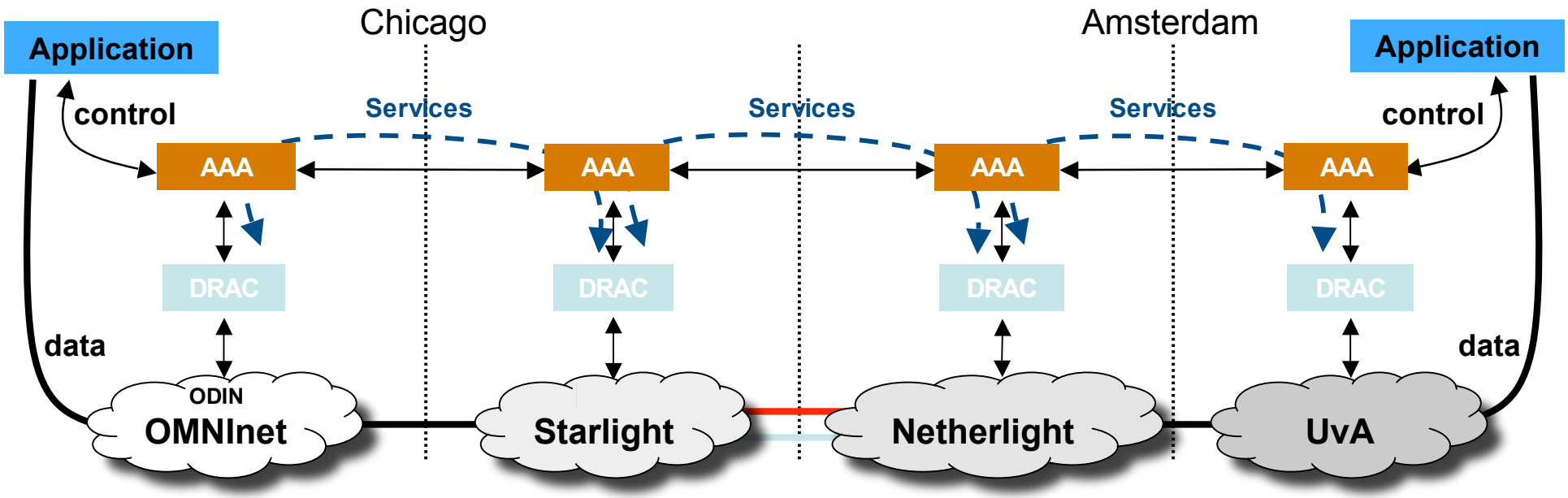


Starting point



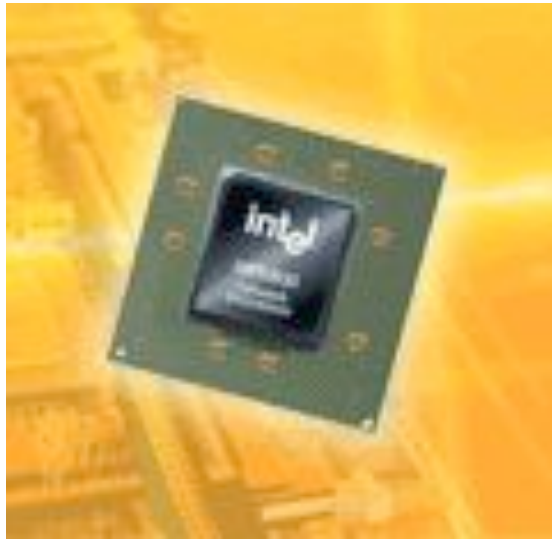
RFC 2903 - 2906 , 3334 , policy draft

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

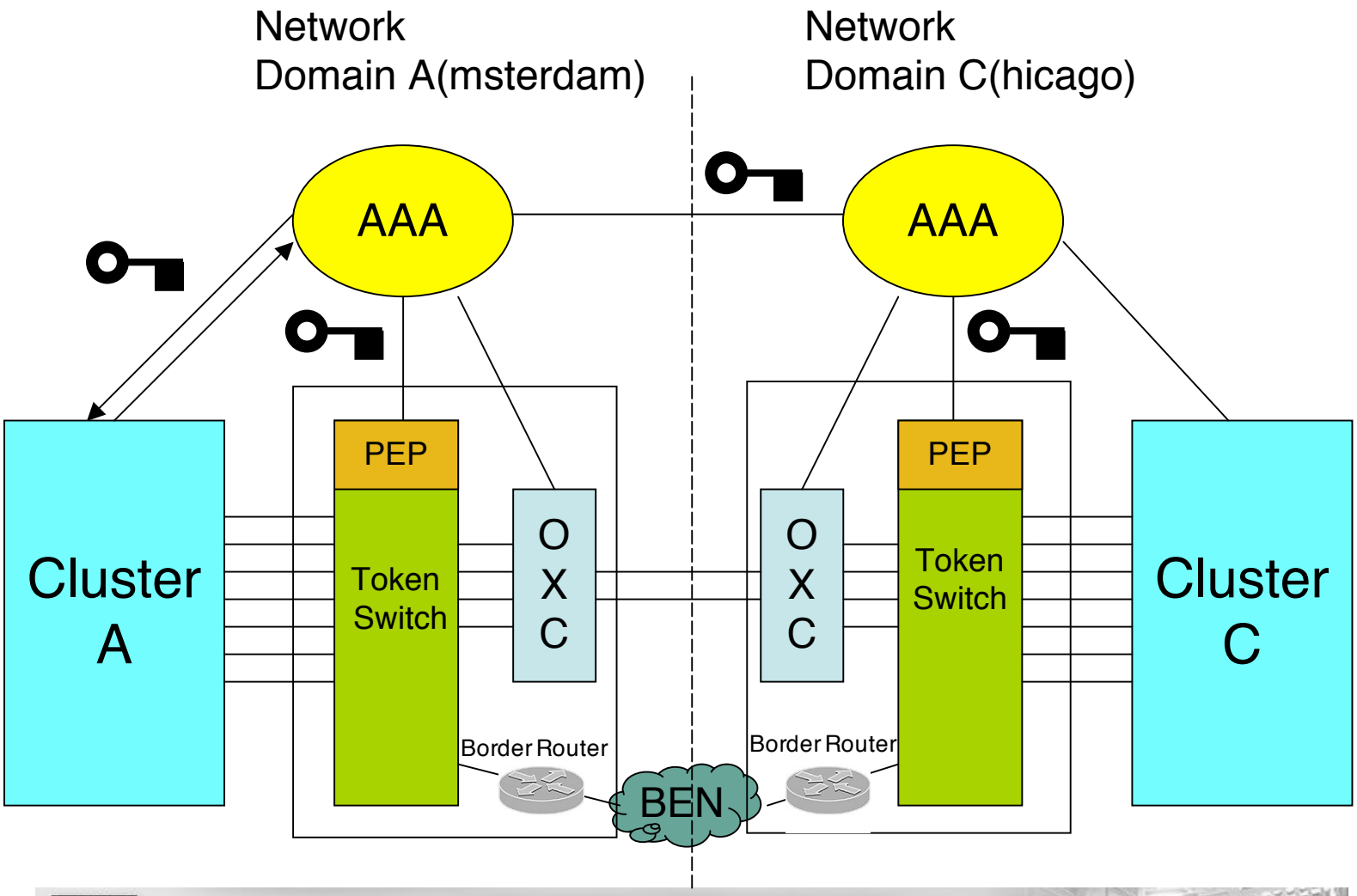
IXP series Network Processor Units



Features:

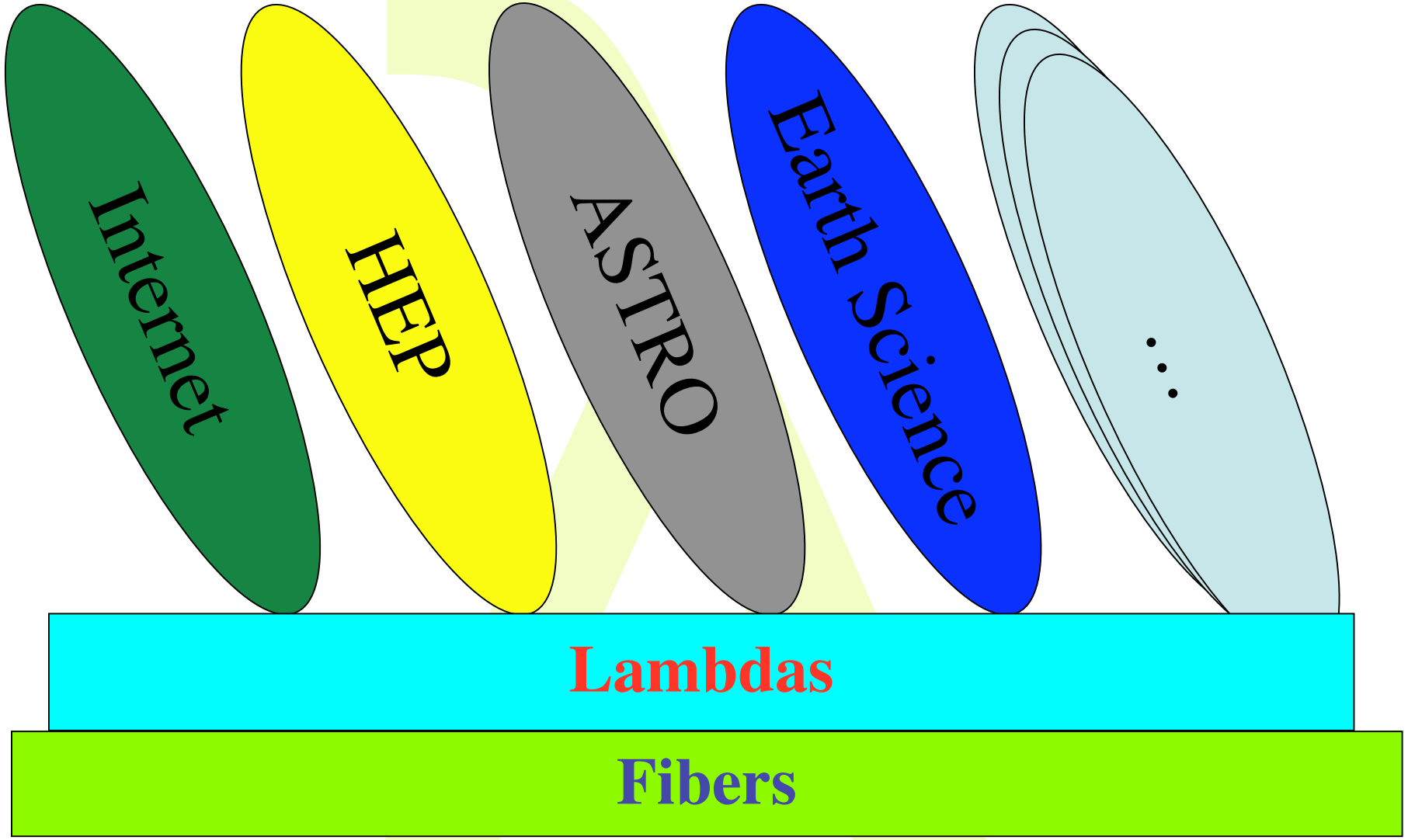
- The IXP 2850 is able to perform packet functions at 10 gb/s
- 16 programmable Micro Engines to allow parallel dataplane processing.
- Two crypto units support bulk security algorithms (AES, DES, 3DES, SHA1)
- Designed for IPSec, however is general enough to do other things.
- Supports Cypher Block Chaining in combination with MAC.

Example experiment agent model

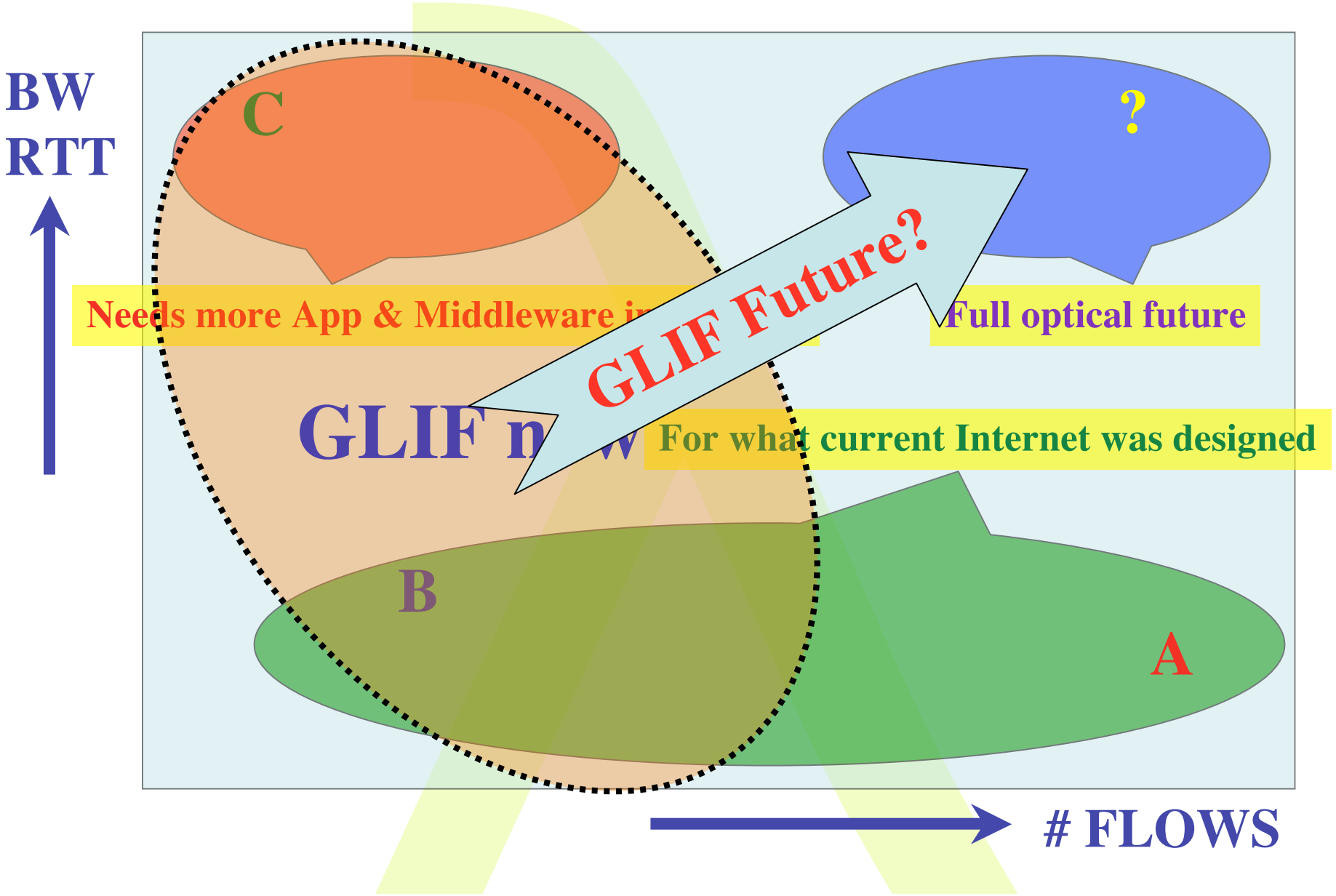


Co-located interactive 3D visualization

Discipline Networks



Transport of flows



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, Fred Wan, Jeroen van der Ham, Karst Koymans, Paola Grosso, Yuri Demchenko, Rob Meijer, VL-team.

Partially complete list:

- Caas
- Chase
- Cess
- Kess
- Case