

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

eVLBI



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

Rates of 8Gb/s or more are entirely feasible.

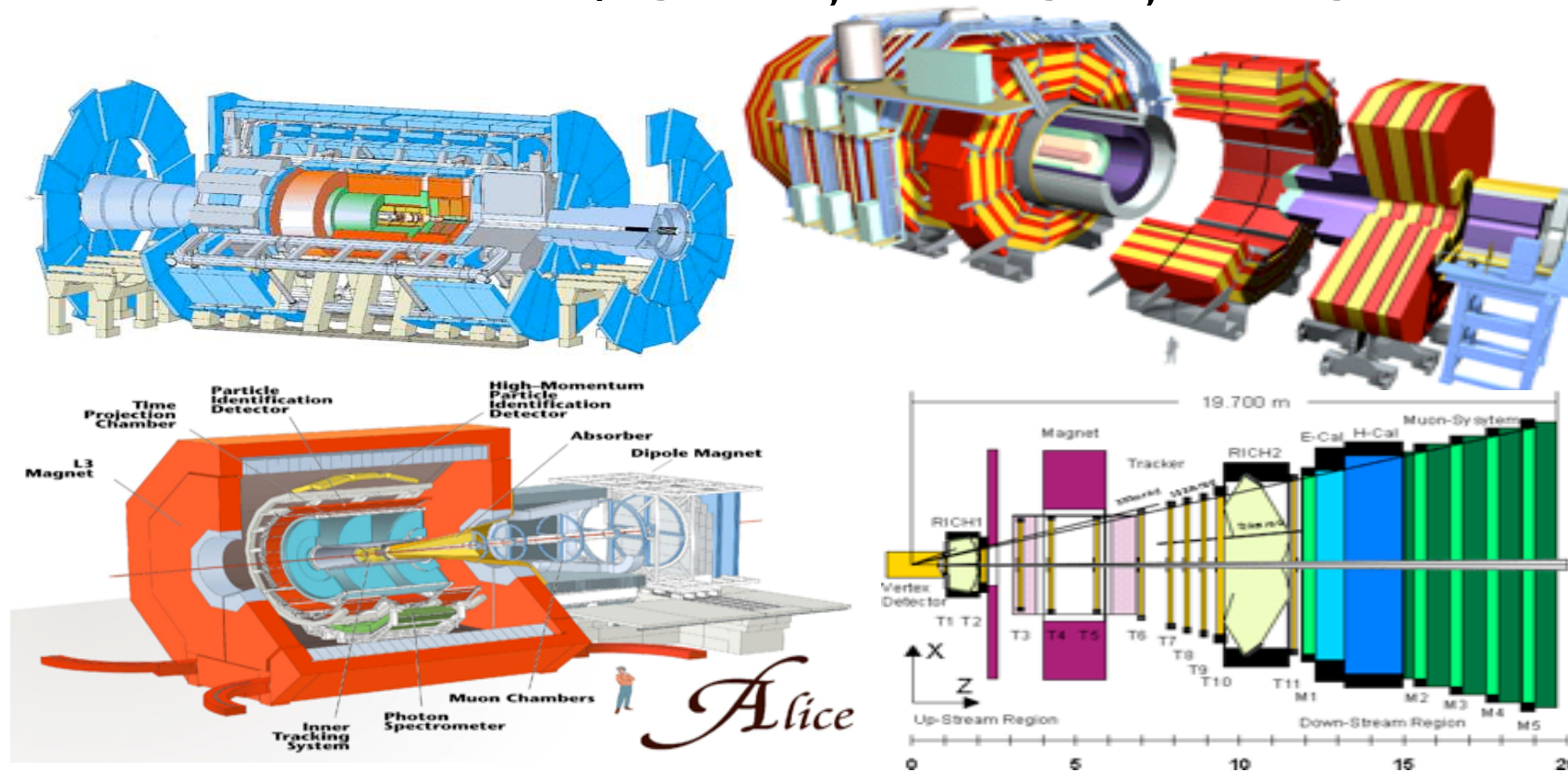


Westerbork Synthesis Radio Telescope - Netherlands

~ 40 Tbit/s
www.lofar.org

Four LHC Experiments: The Petabyte to Exabyte Challenge

- **ATLAS, CMS, ALICE, LHCb**

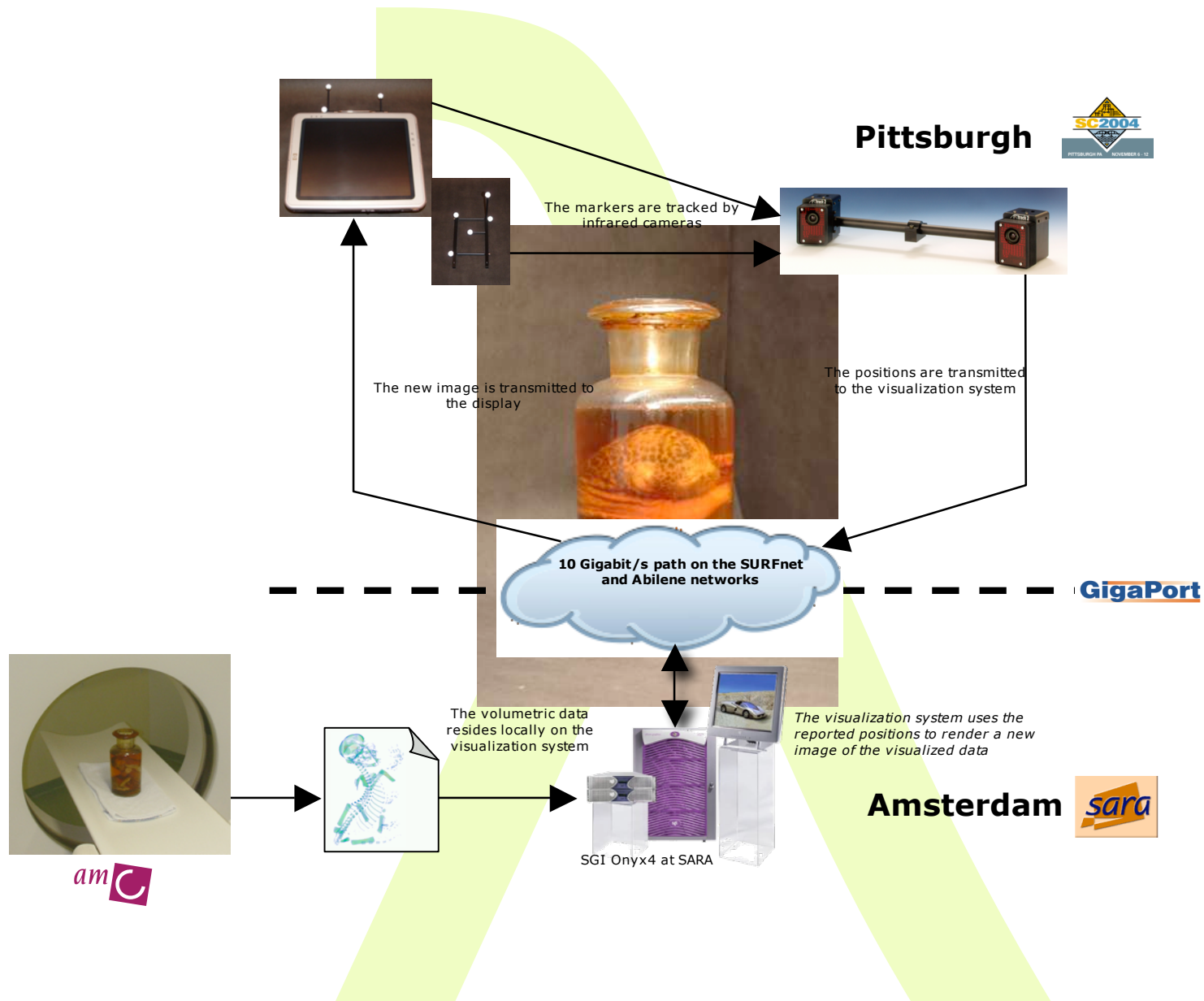


6000+ Physicists & Engineers; 60+ Countries; 250 Institutions

Tens of PB 2008; To 1 EB by ~2015
Hundreds of TFlops To PetaFlops



Co-located interactive 3D visualization



SC2004 “Dead Cat” demo

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

Produced by:

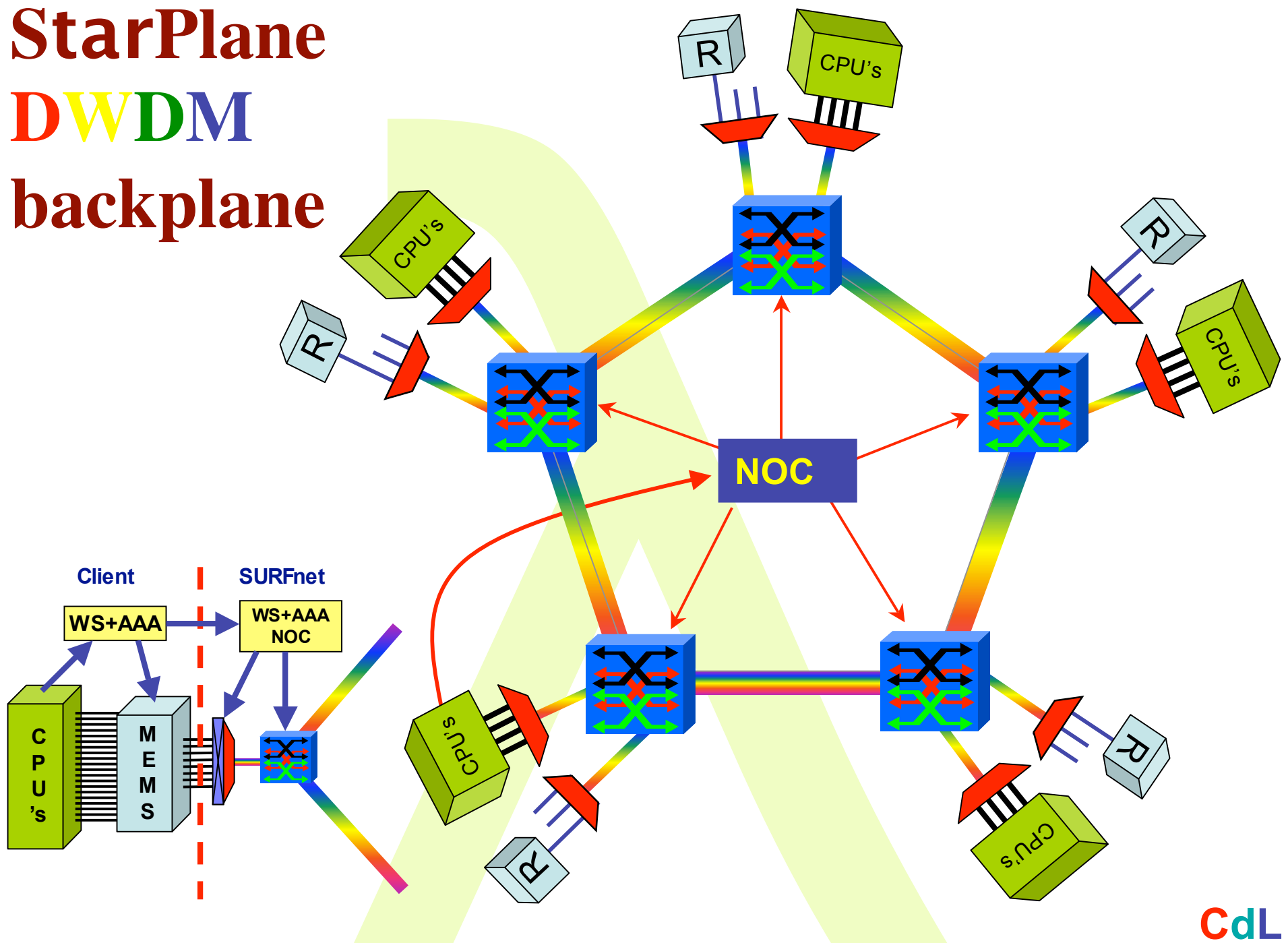
Michael Scarpa
Robert Belleman
Peter Slood

Many thanks to:

AMC
SARA
GigaPort
UvA/AIR
Silicon Graphics, Inc.
Zoölogisch Museum



StarPlane DWDM backplane

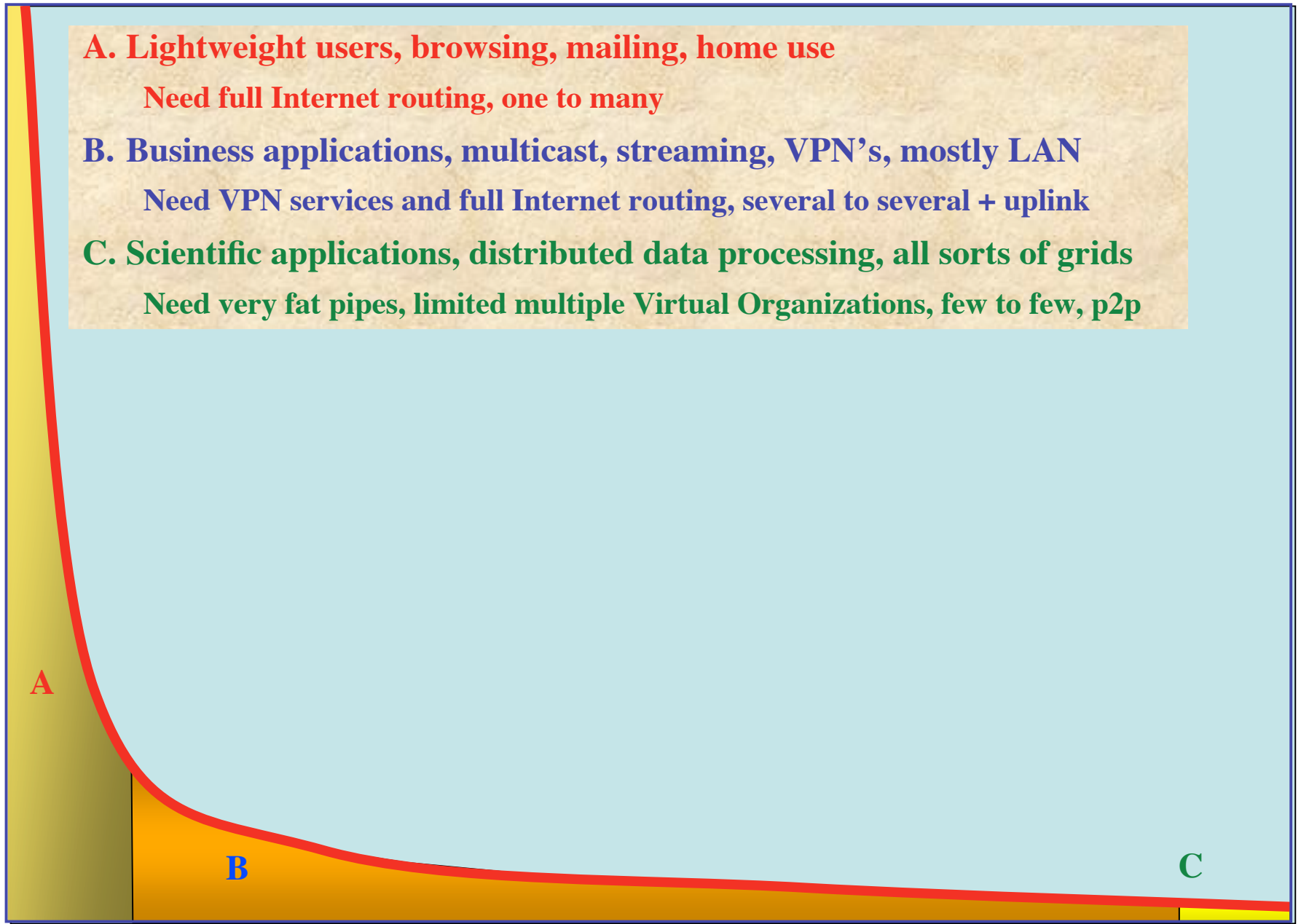


Showed you 5 types of Grids

- 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

U
S
E
R
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



ADSL

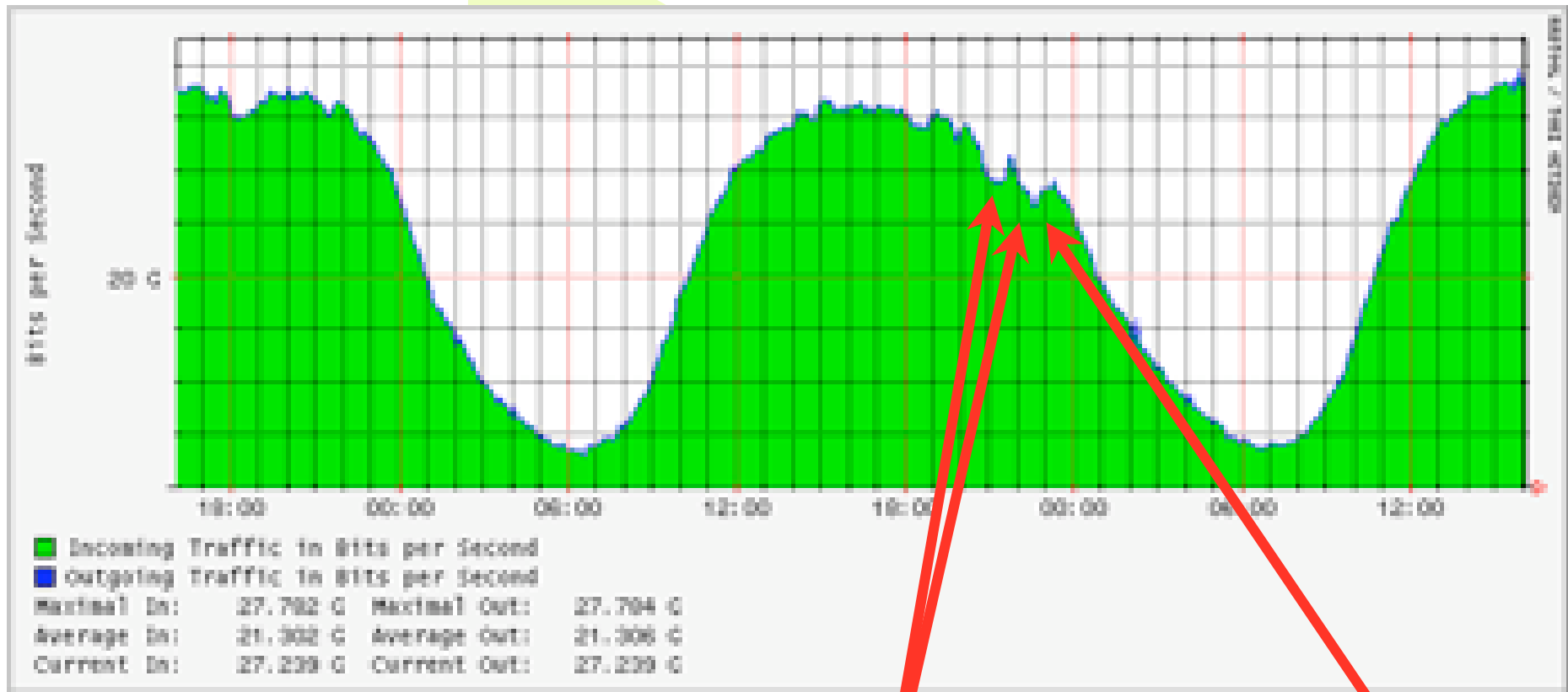
GigE

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 \implies 20 Gb/s

- **Estimate B**

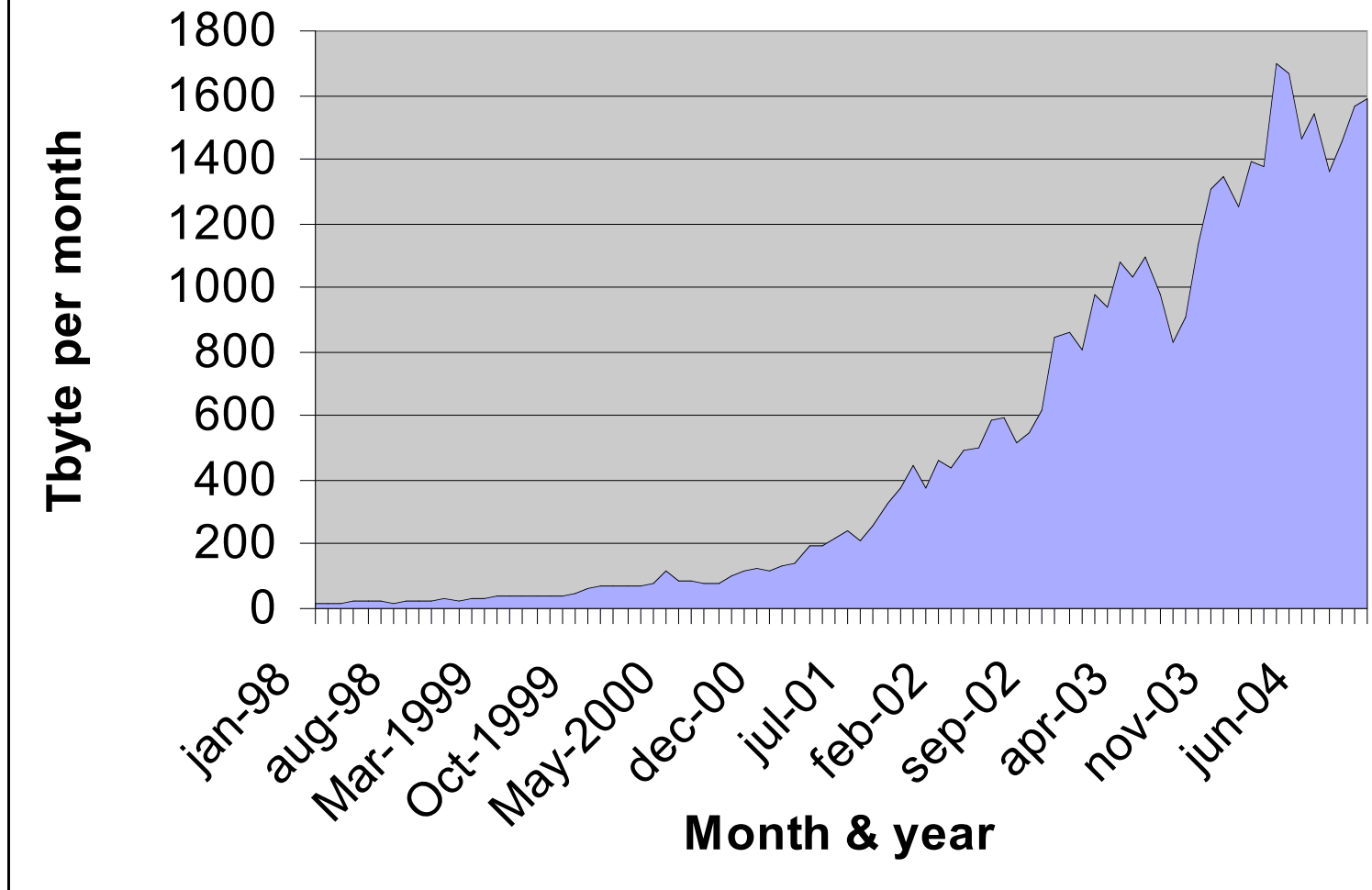
- SURFnet5 has 2*10 Gb/s to about 15 institutes and 0.1 to 1 Gb/s to 170 customers, estimate same for industry (overestimation) \implies 10-30 Gb/s

- **Estimate C**

- Leading HEF and ASTRO + rest \implies 80-120 Gb/s
- LOFAR $\implies \approx 37$ Tbit/s $\implies \approx n \times 10$ Gb/s

Routed L3 traffic growth

SURFnet customer traffic: Monthly volume



1600 Tbyte/month \approx 5 Gbits/second

Slide courtesy Kees Neggers

u
s
e
r
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

$\Sigma C \gg 100 \text{ Gb/s}$ →

$\Sigma B \approx 30 \text{ Gb/s}$

$\Sigma A \approx 20 \text{ Gb/s}$

A

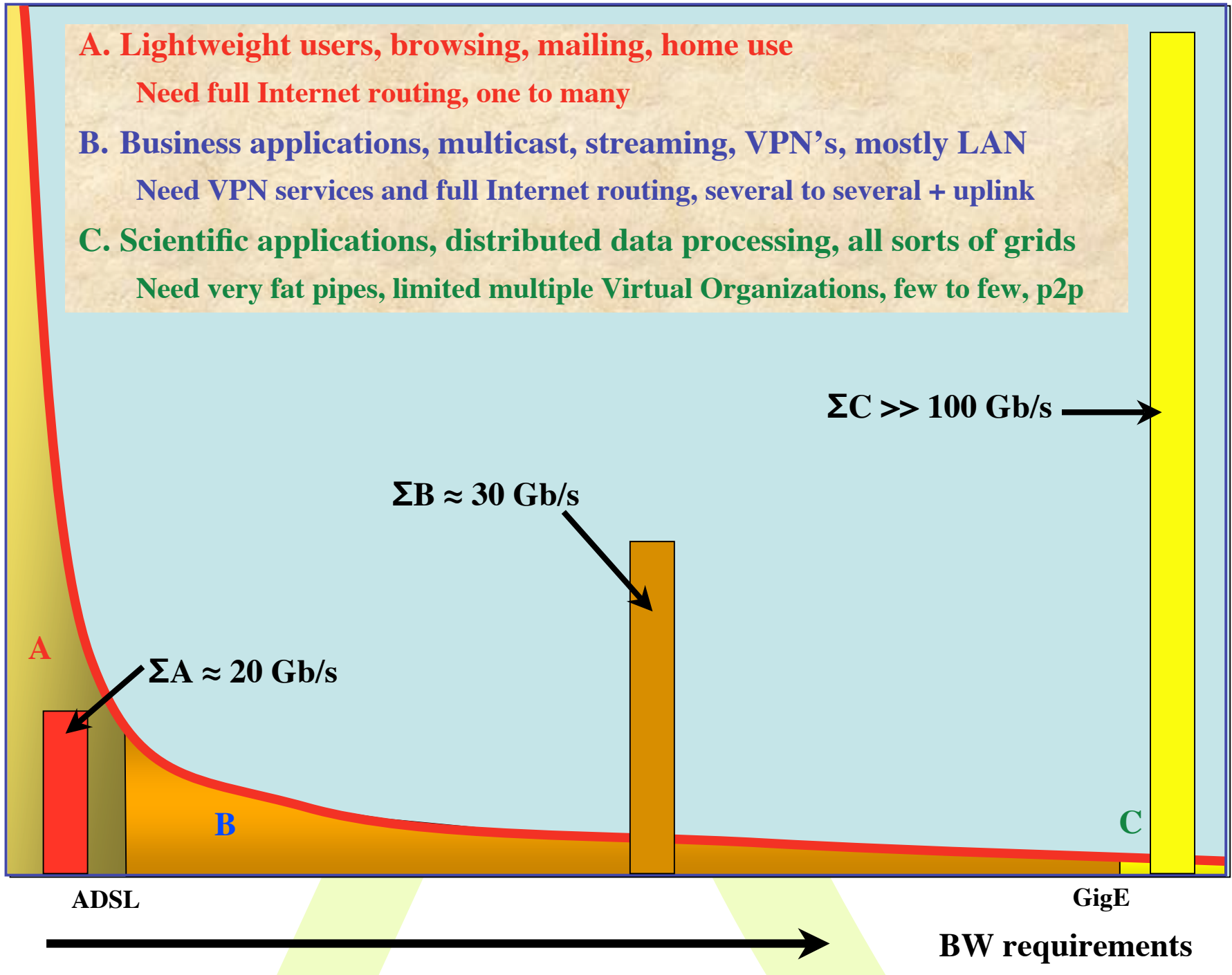
B

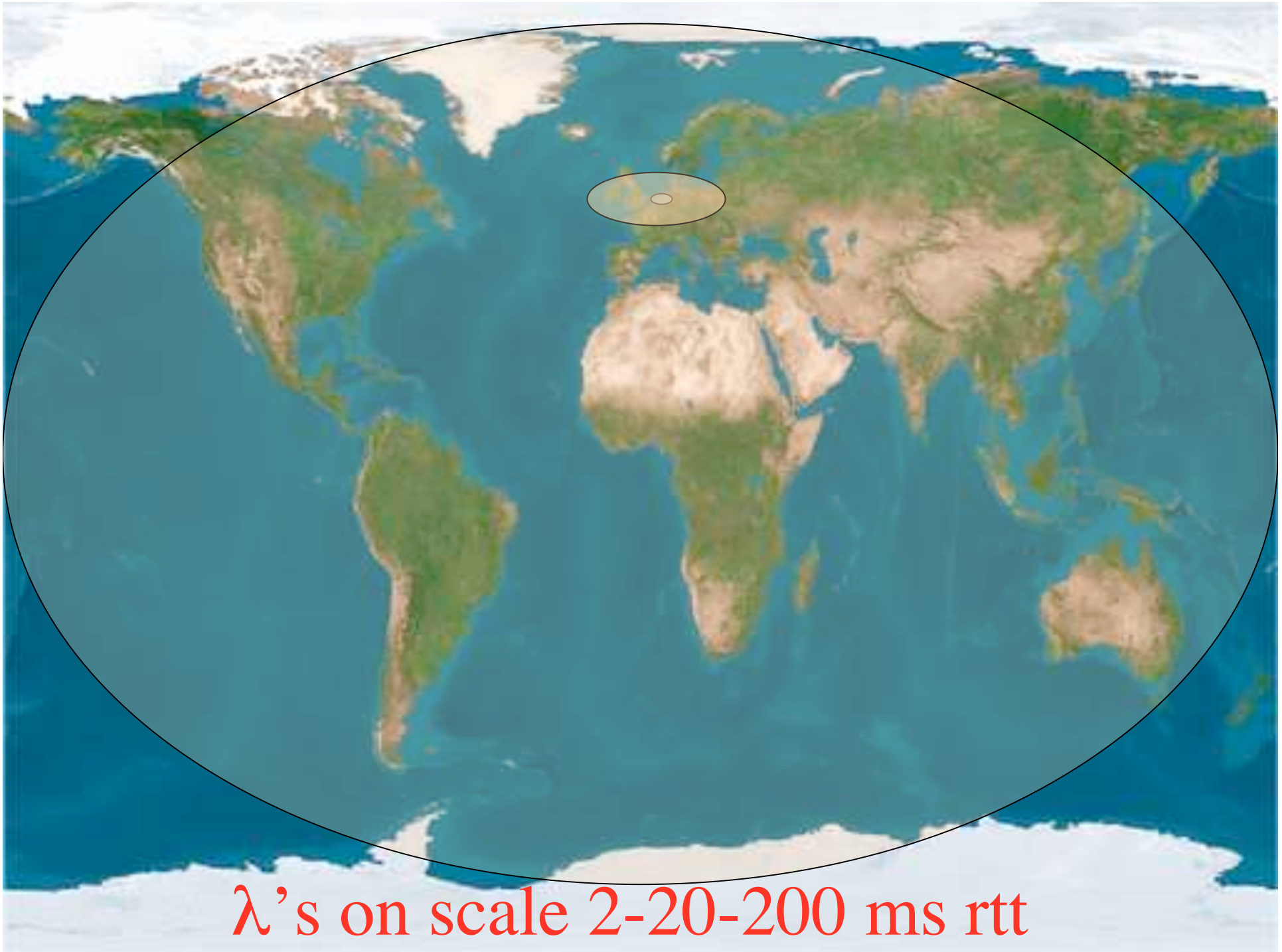
C

ADSL

GigE

→
BW requirements



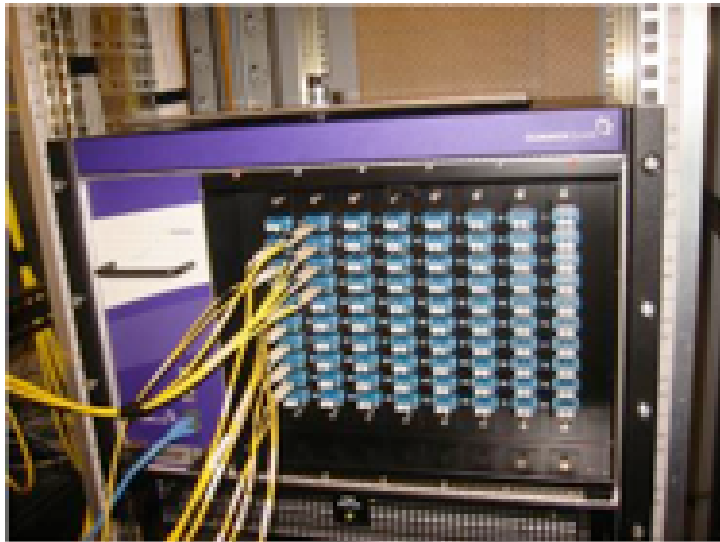


λ 's on scale 2-20-200 ms rtt

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 -> 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 \approx 1 k\$/port



L2 \approx 7-15 k\$/port



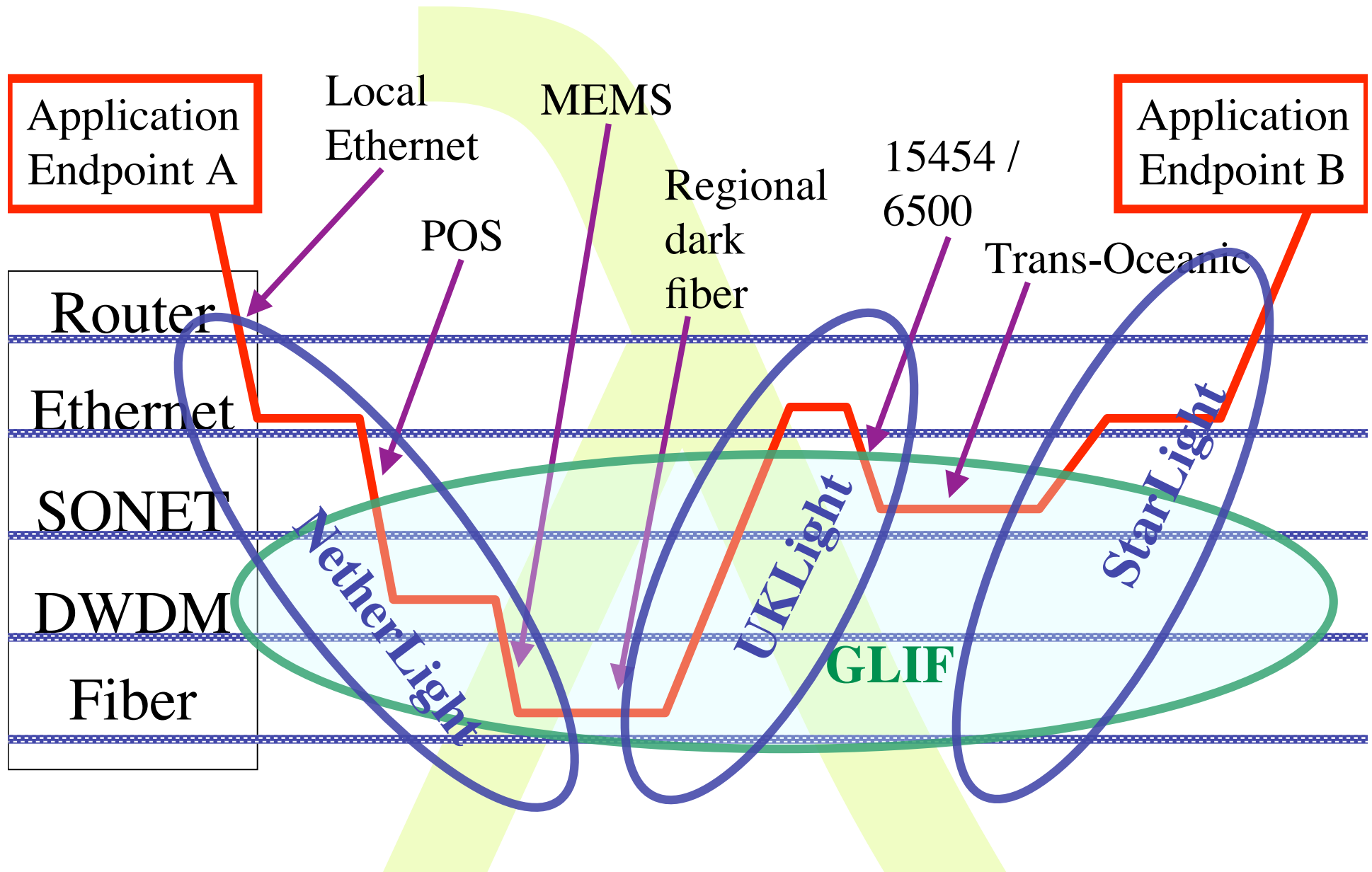
L3 \approx 100+ k\$/port



Services

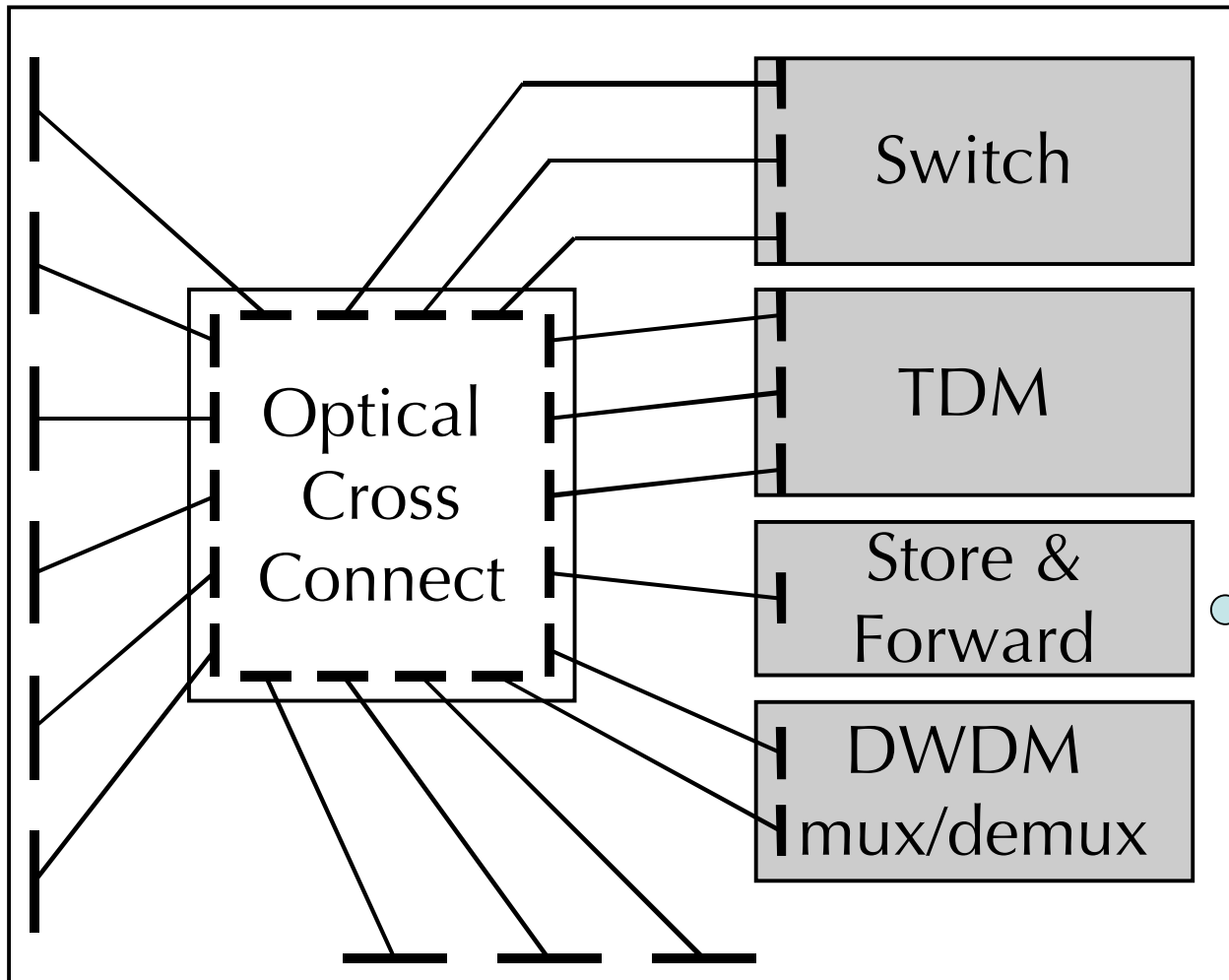
<div style="text-align: right;">SCALE</div> <div style="text-align: left;">CLASS</div>	2 Metro	20 National/ regional	200 World
A	Switching/ routing	Routing	ROUTER\$
B	Switches + E-WANPHY VPN's	Switches + E-WANPHY (G)MPLS	ROUTER\$
C	dark fiber DWDM MEMS switch	DWDM, TDM / SONET Lambda switching	Lambdas, VLAN's SONET Ethernet

How low can you go?



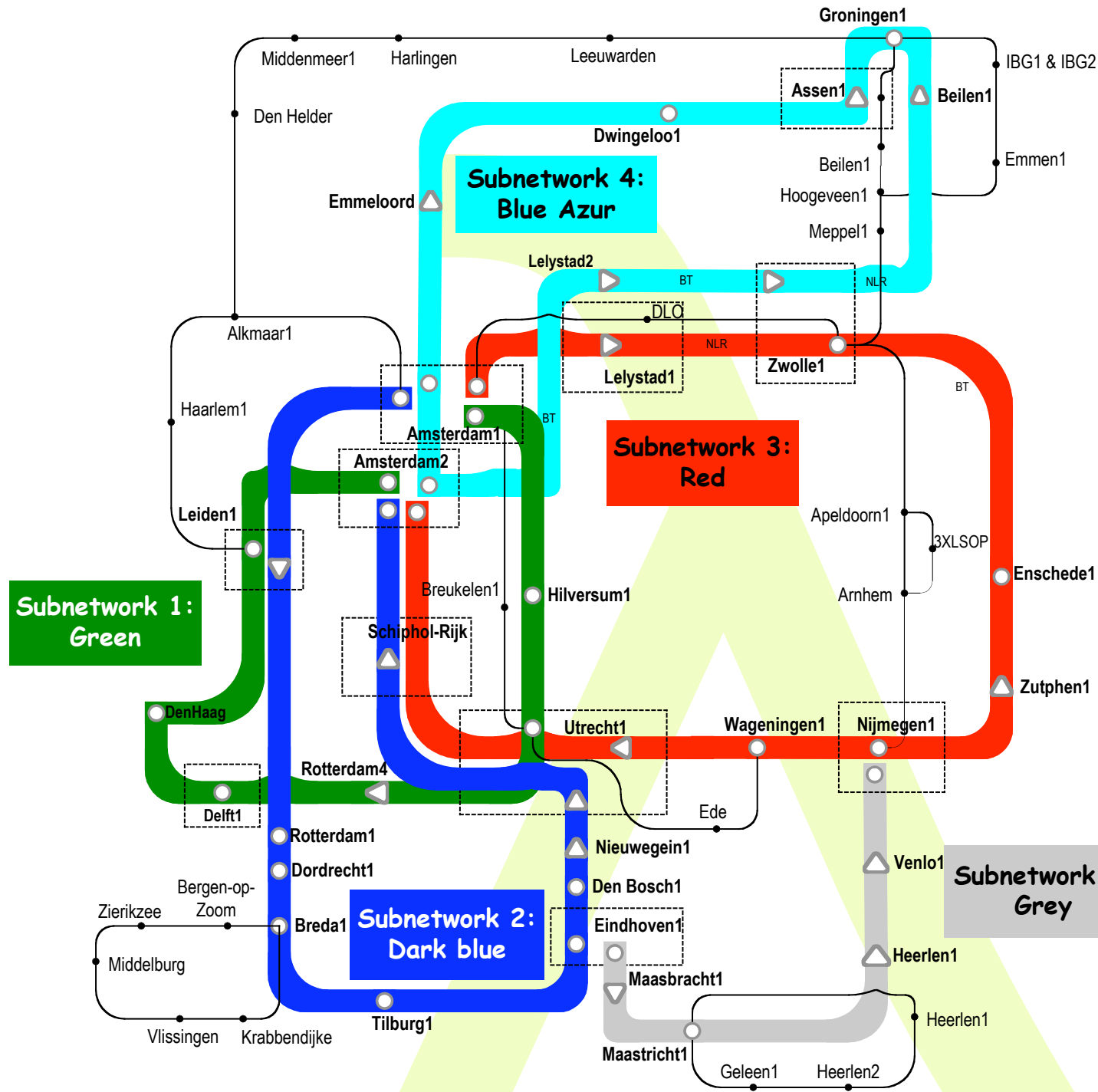
Optical Exchange as Black Box

Optical Exchange



TeraByte
Email
Service

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 :-)

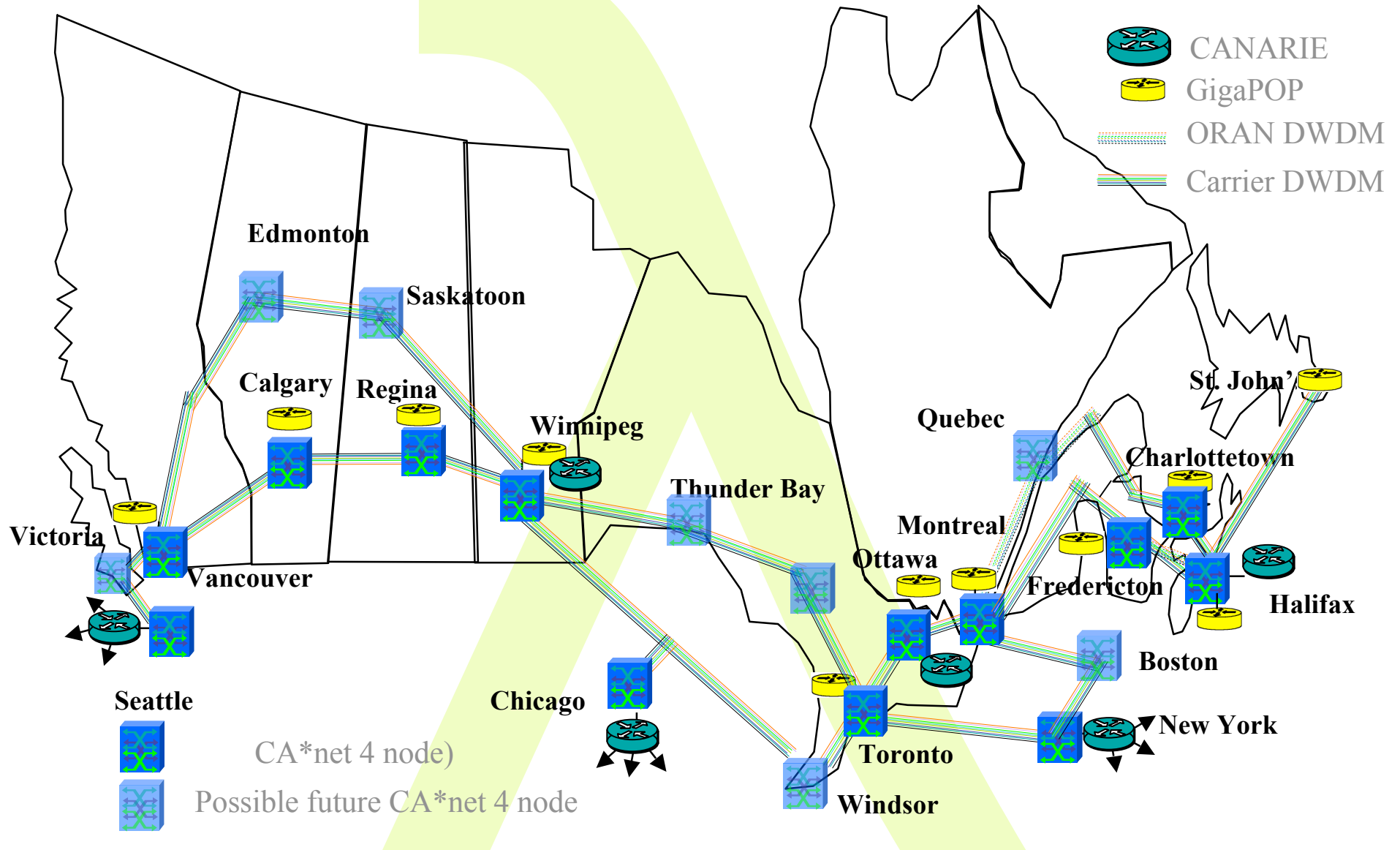


UCLP intended for projects like National LambdaRail

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



CA*net 4 Architecture

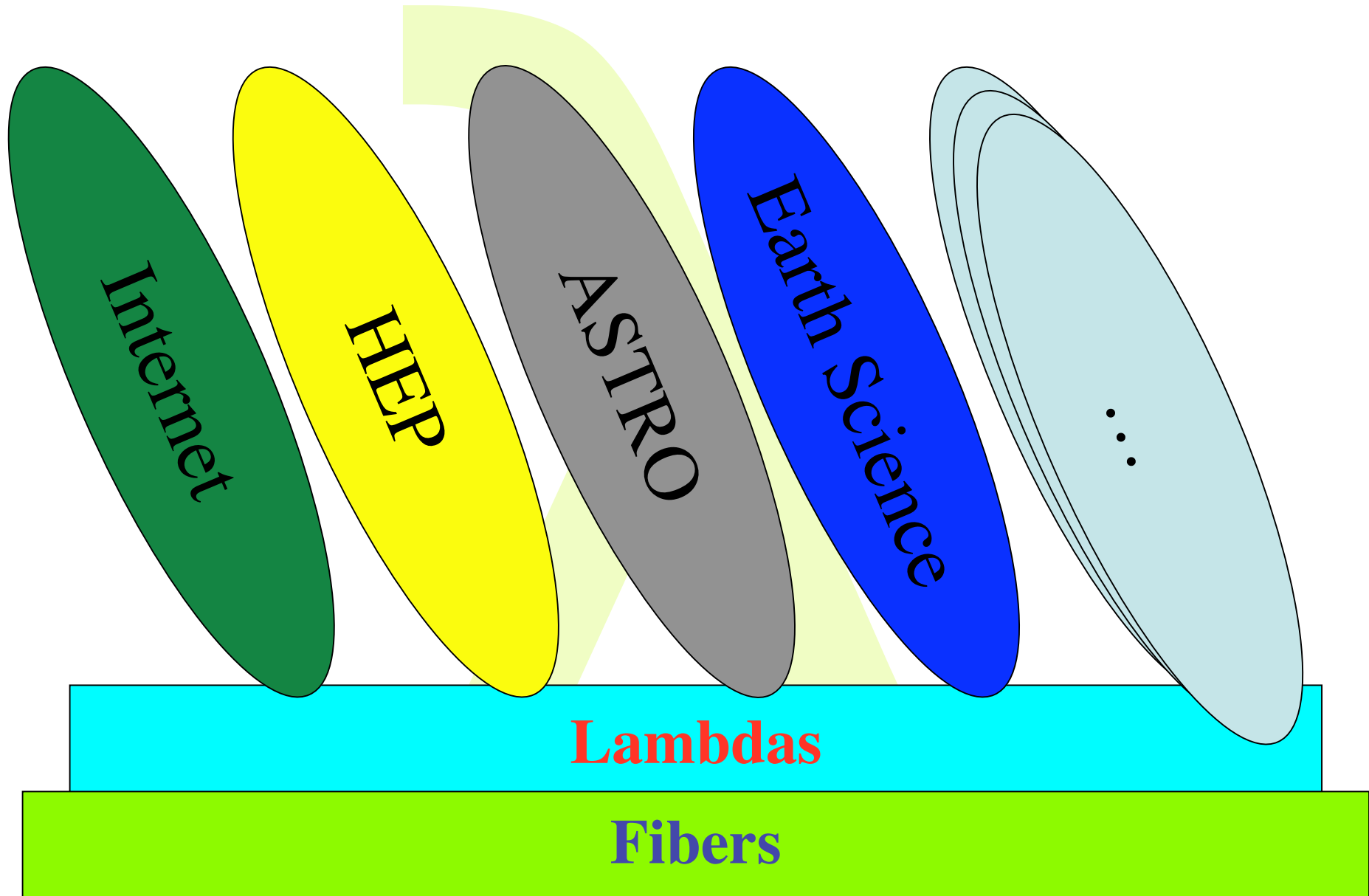


GLIF Q4 2004



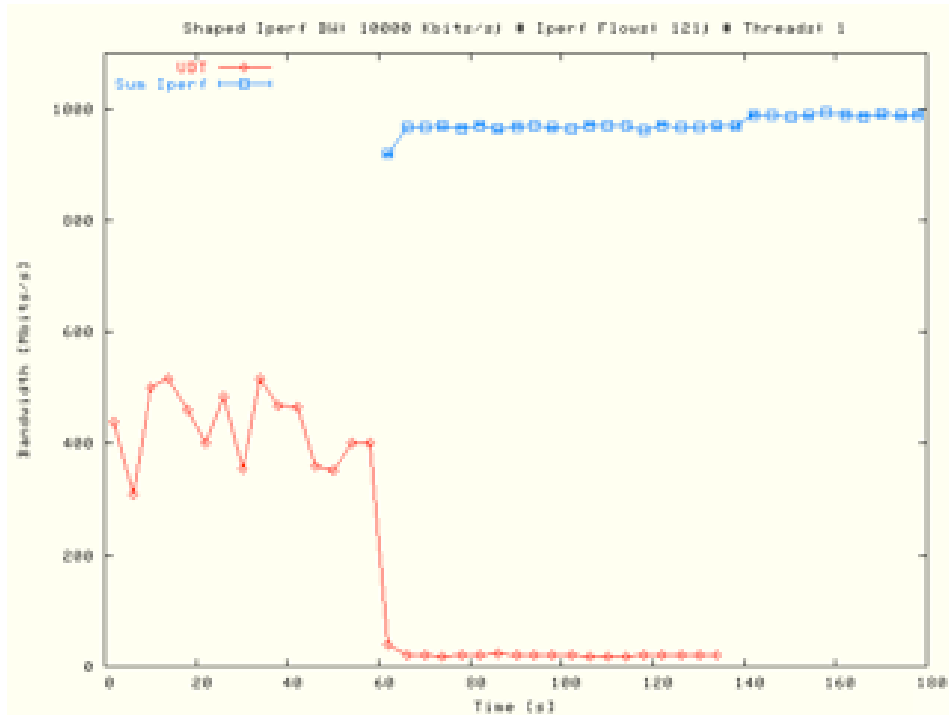
Visualization courtesy of
Bob Patterson, NCSA.

Discipline Networks

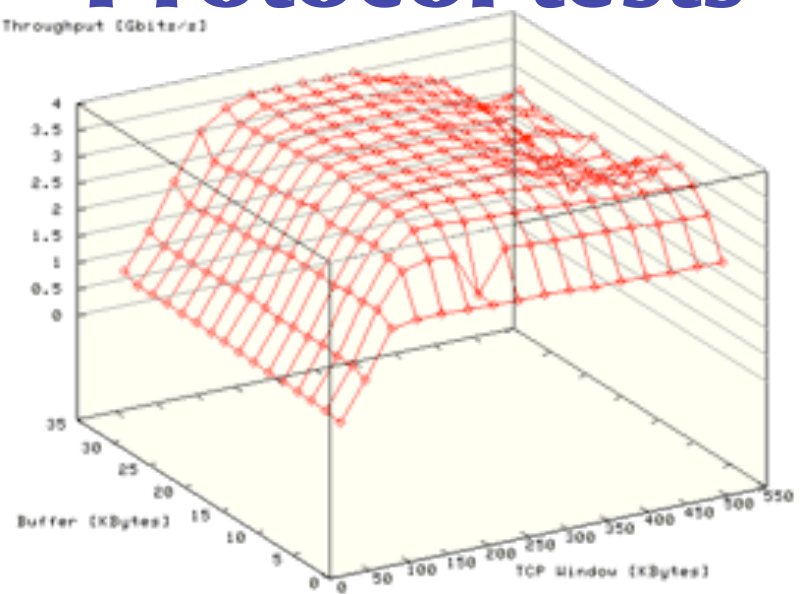




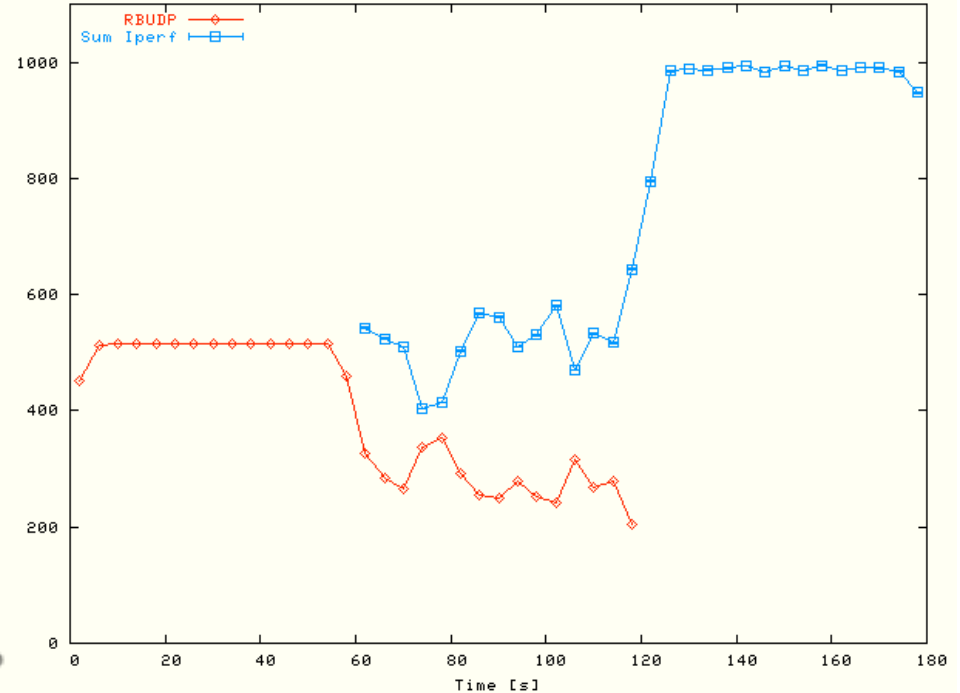
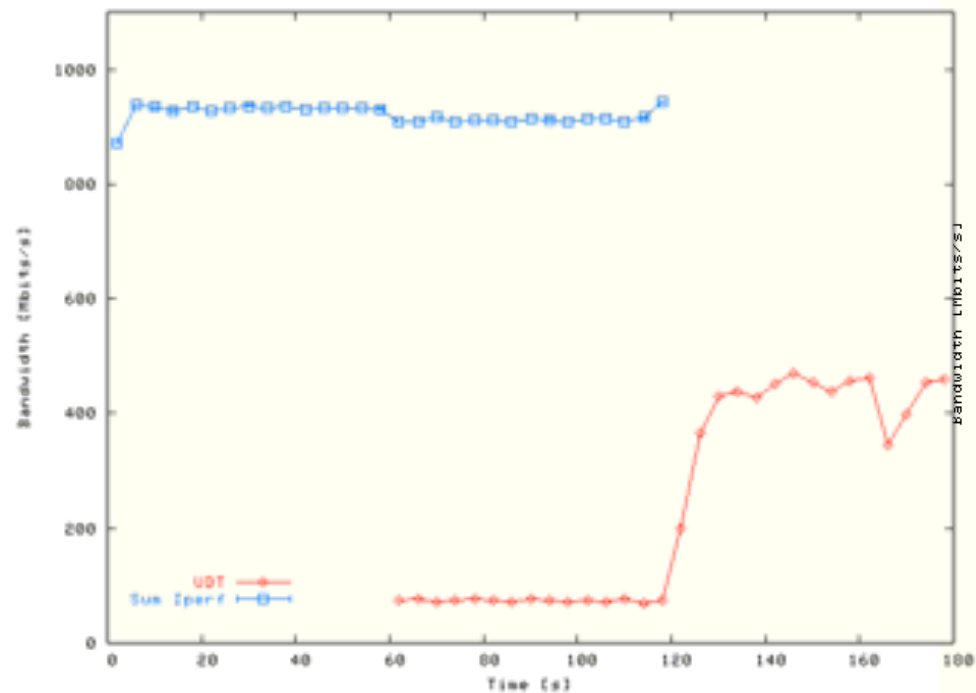
Protocol tests



Throughput (Gbits/s)



RBUDP Data Size: 32 MByte; Shaped Iperf BW: 10000 Kbits/s; # Iperf Flows: 121



Layer - 2 requirements from 3/4



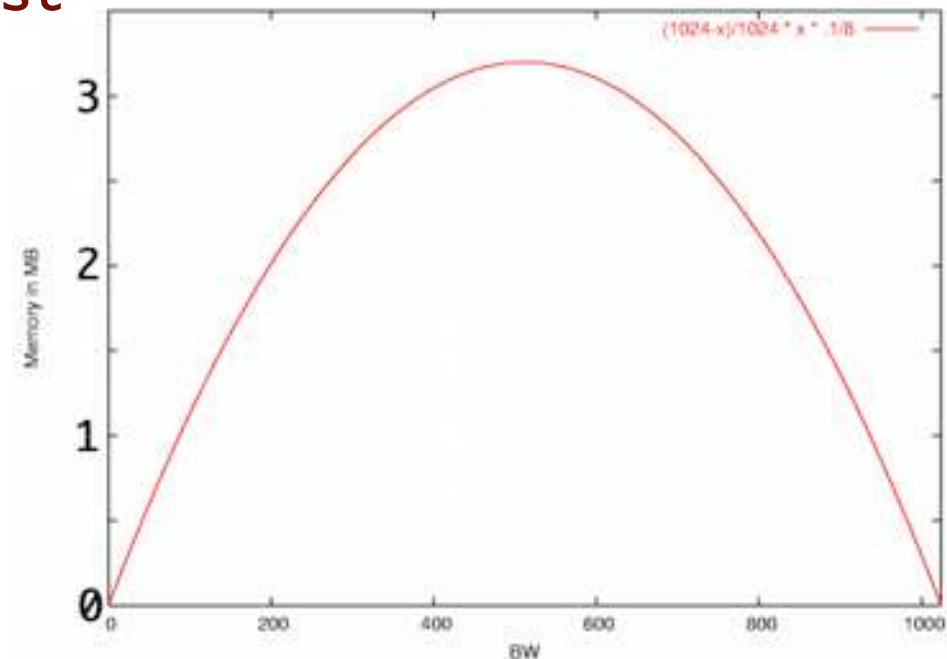
TCP is bursty due to sliding window protocol and slow start algorithm.

$$\text{Window} = \text{BandWidth} * \text{RTT} \quad \& \quad \text{BW} == \text{slow}$$

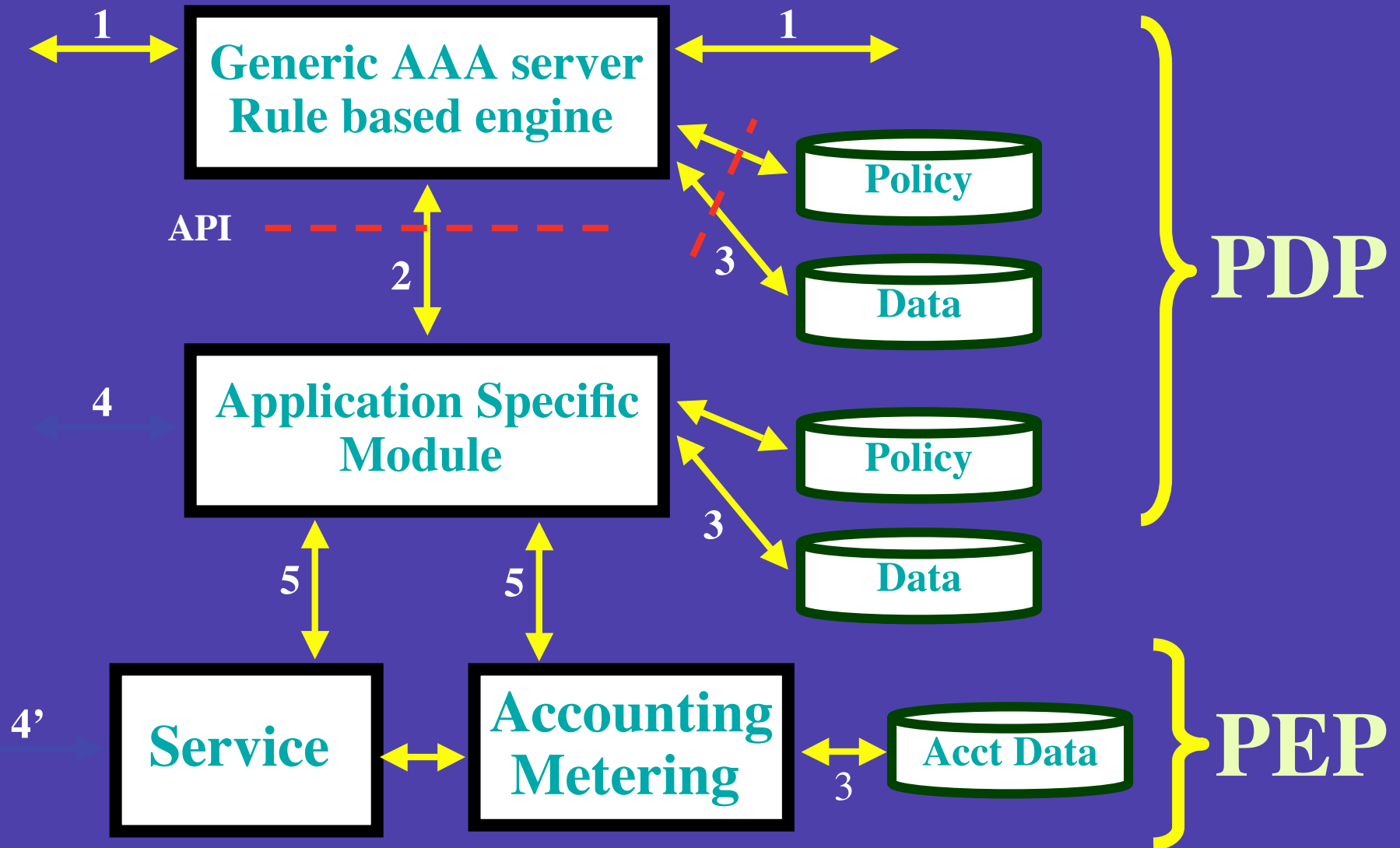
$$\text{Memory-at-bottleneck} = \frac{\text{fast} - \text{slow}}{\text{fast}} * \text{slow} * \text{RTT}$$

So pick from menu:

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

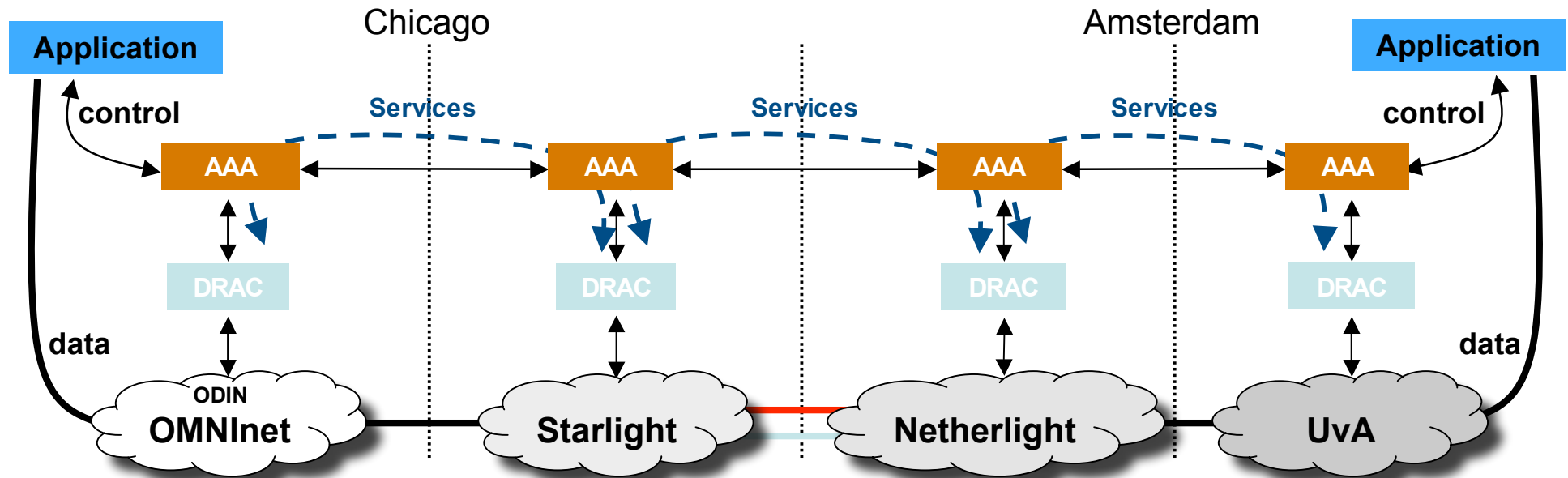


Starting point



RFC 2903 - 2906 , 3334 , policy draft

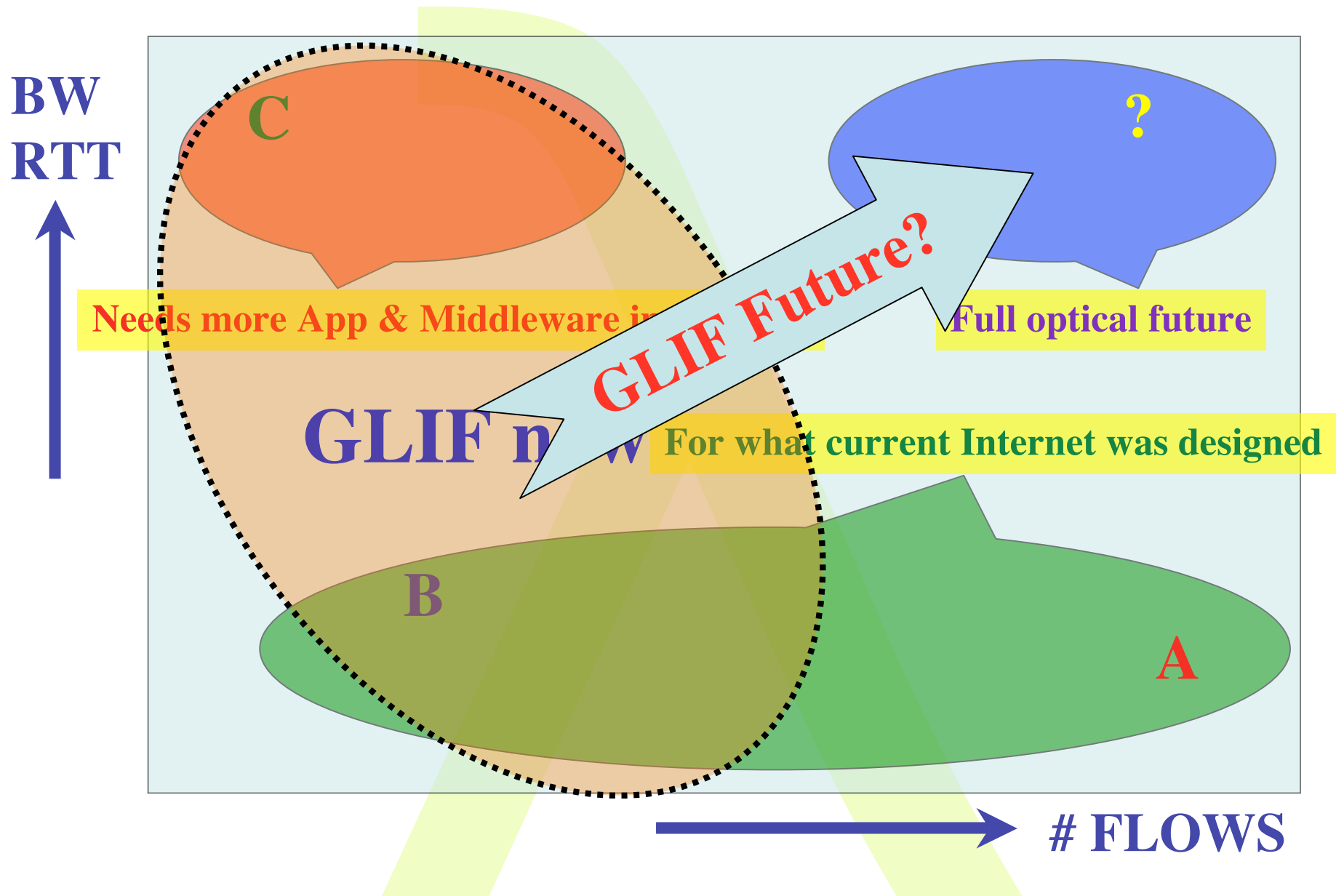
SC2004 CONTROL CHALLENGE



- 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



Transport of flows



Not quite ~~ENDING~~

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

