

Lambda-Grid developments

Global Lambda Integrated Facility

www.science.uva.nl/~deLaat

Cees de Laat

GigaPort
EU



University of Amsterdam

SARA
NCF



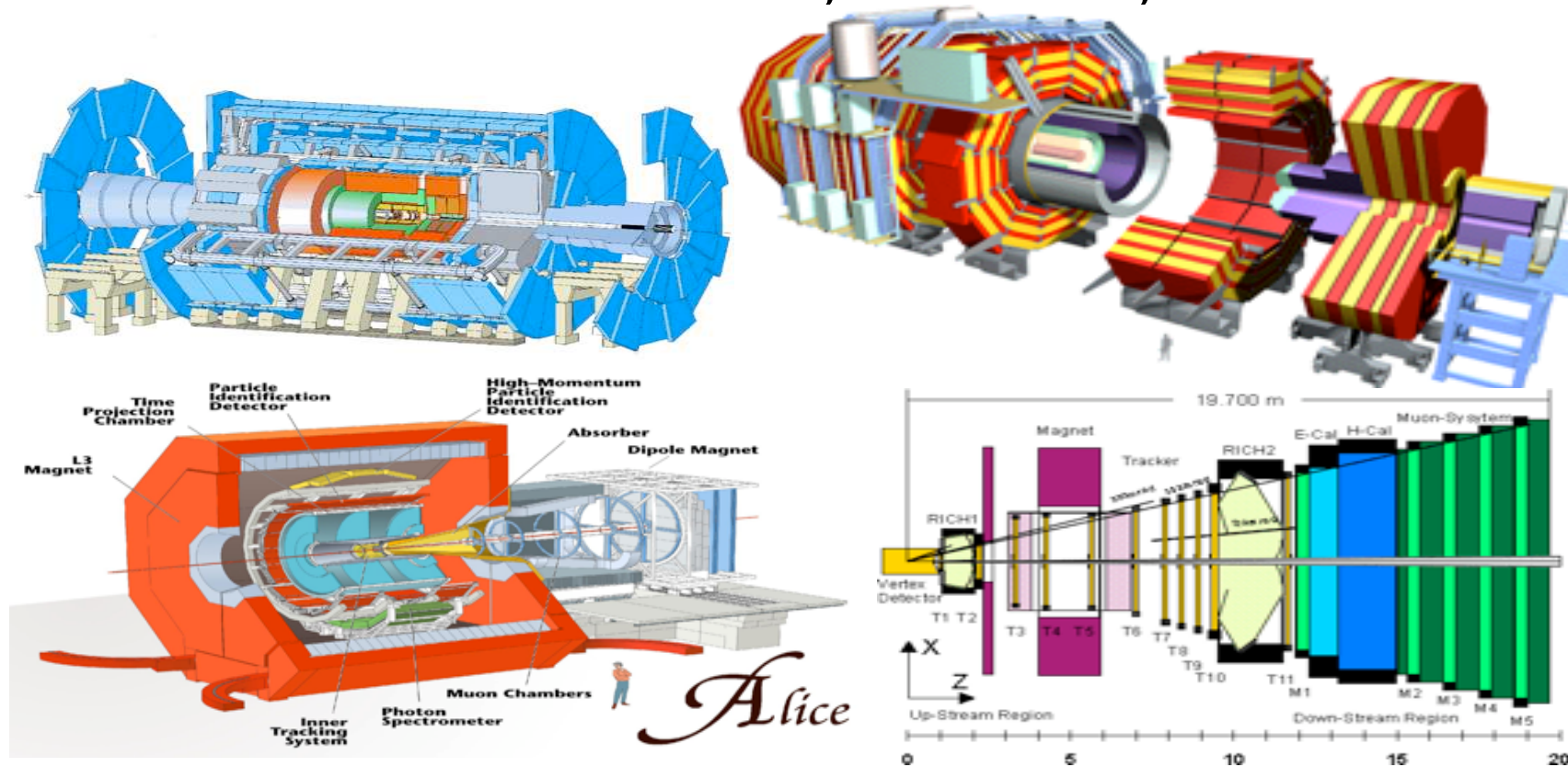
Contents

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**



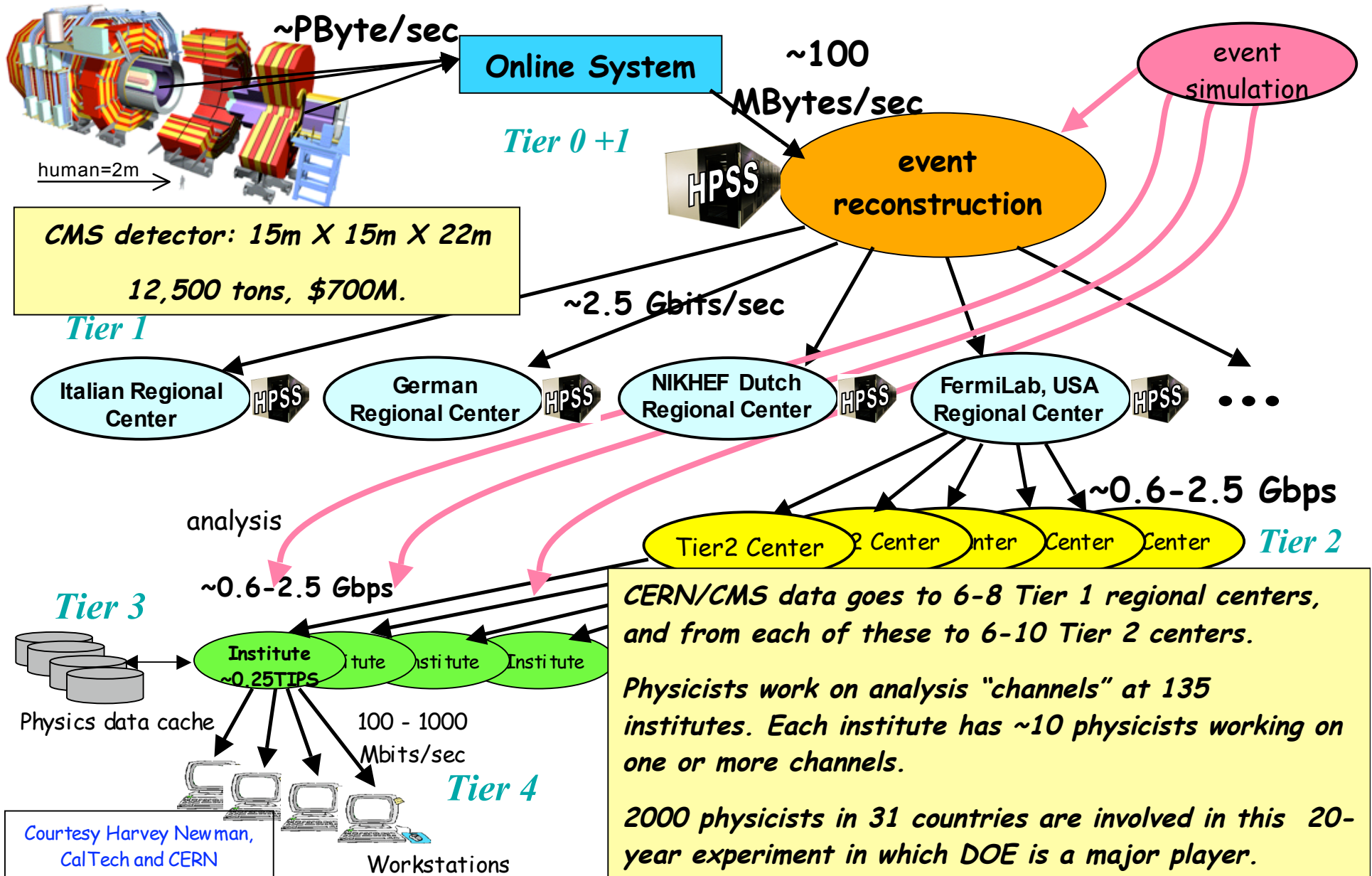
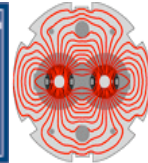
6000+ Physicists &
Engineers; 60+
Countries;
250 Institutions

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



LHC Data Grid Hierarchy

CMS as example, Atlas is similar



VLBI

VLBI is easily capable of generating many Gb of data per

The sensitivity of the VLBI array scales with

(data-rate) and there is a strong push to

Rates of 8Gb/s or more are entirely feasible

development. It is expected that parallel

correlator will remain the most efficient approach

distributed processing may have an application

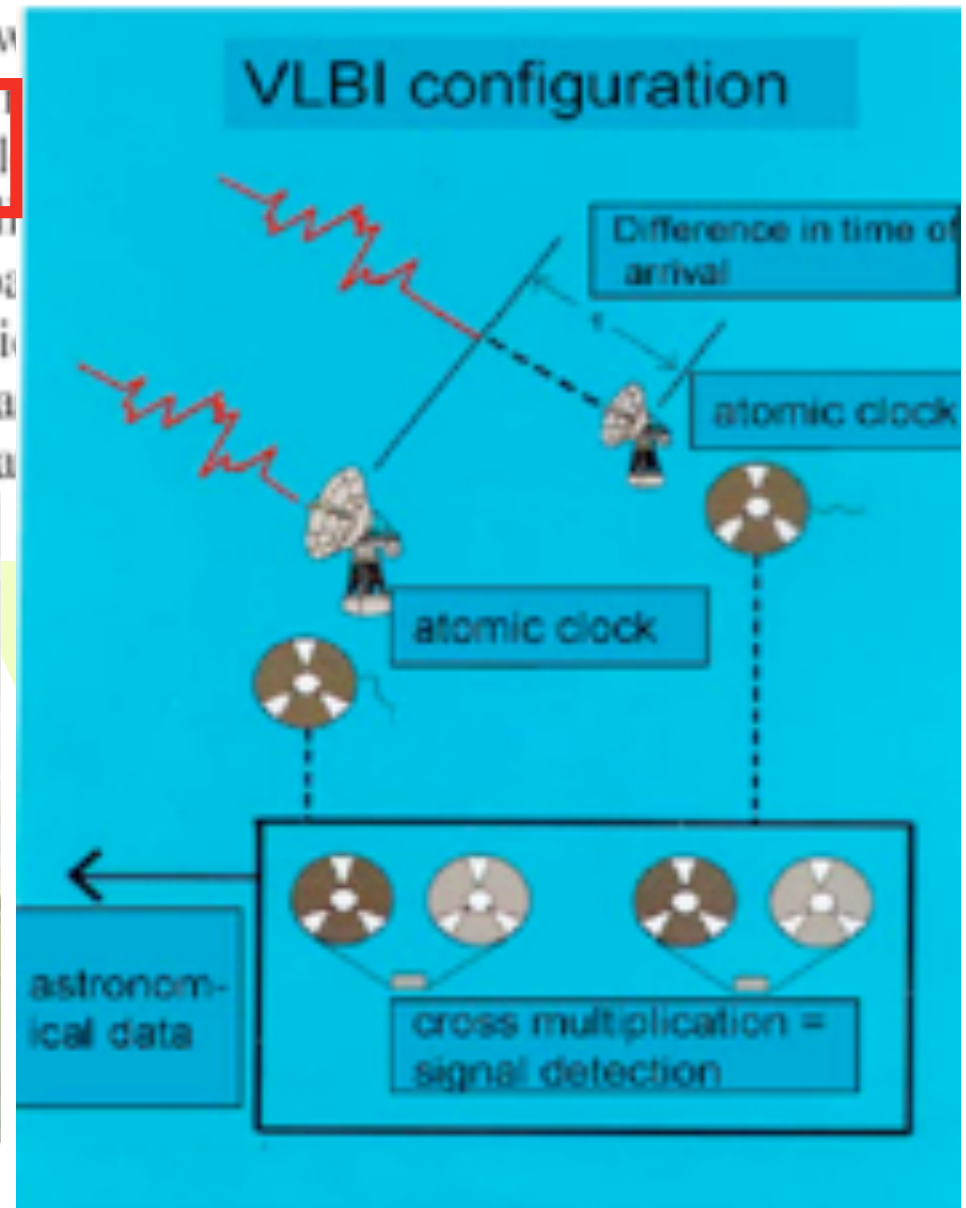
multi-gigabit data streams will aggregate into larger

and the capacity of the final link to the data

center.



Westerbork Synthesis Radio Telescope - Netherlands



Lambdas as part of instruments

GigaPort



www.lofar.org

37 Tbit/s - 116 Tops/s

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

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

SURFnet



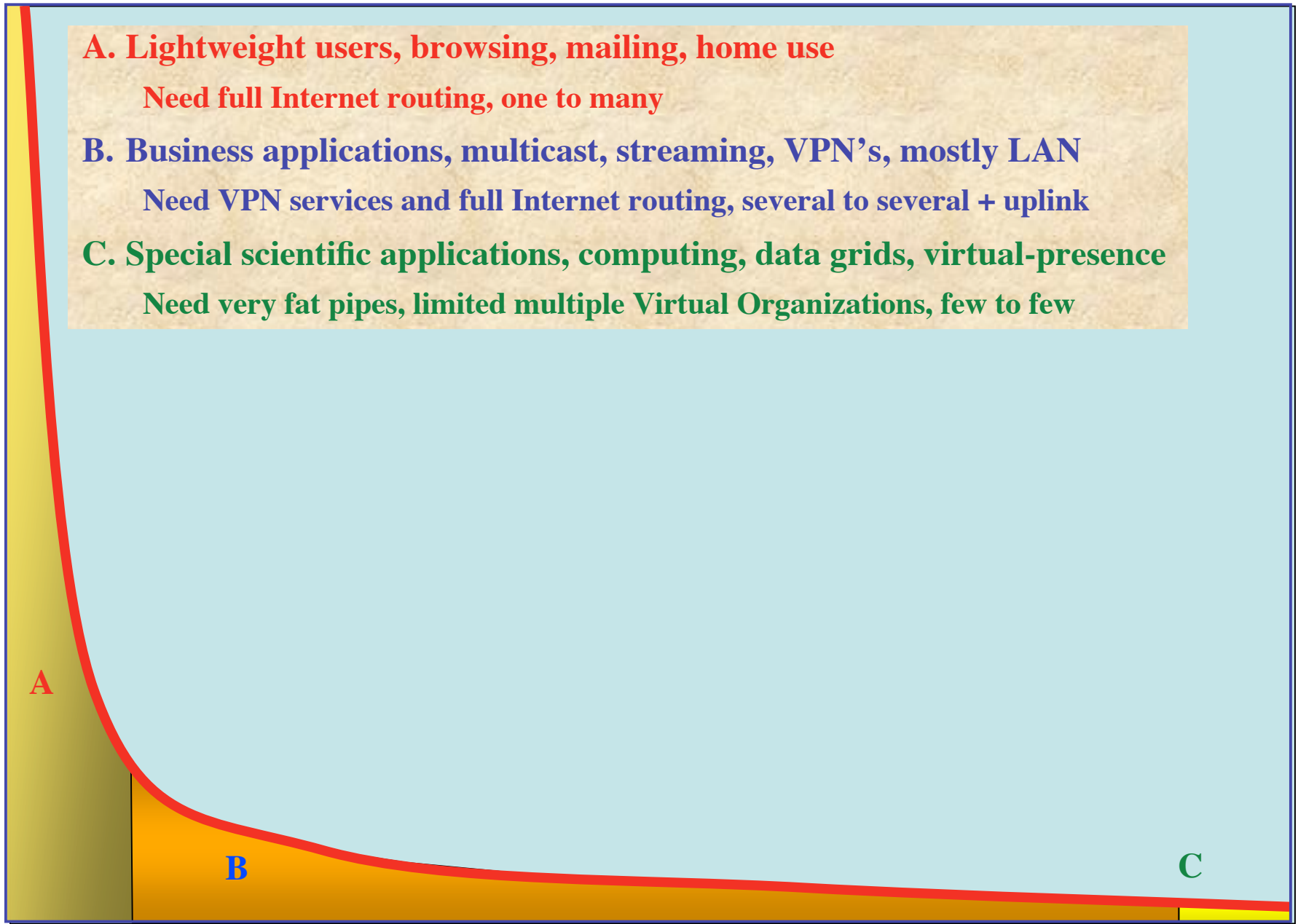
Grids

Showned 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

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



ADSL

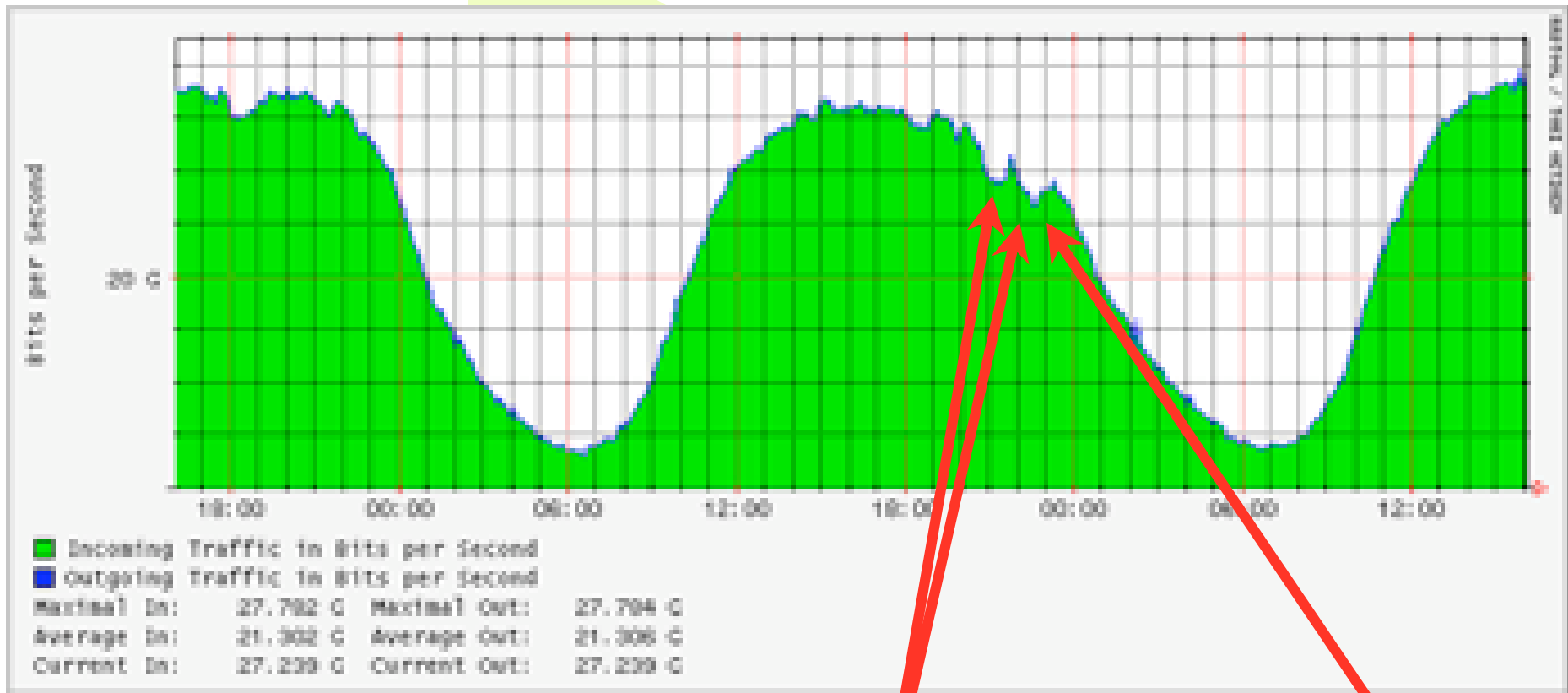
GigE

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

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

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

$\Sigma C \gg 100 \text{ Gb/s}$ →

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

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

A

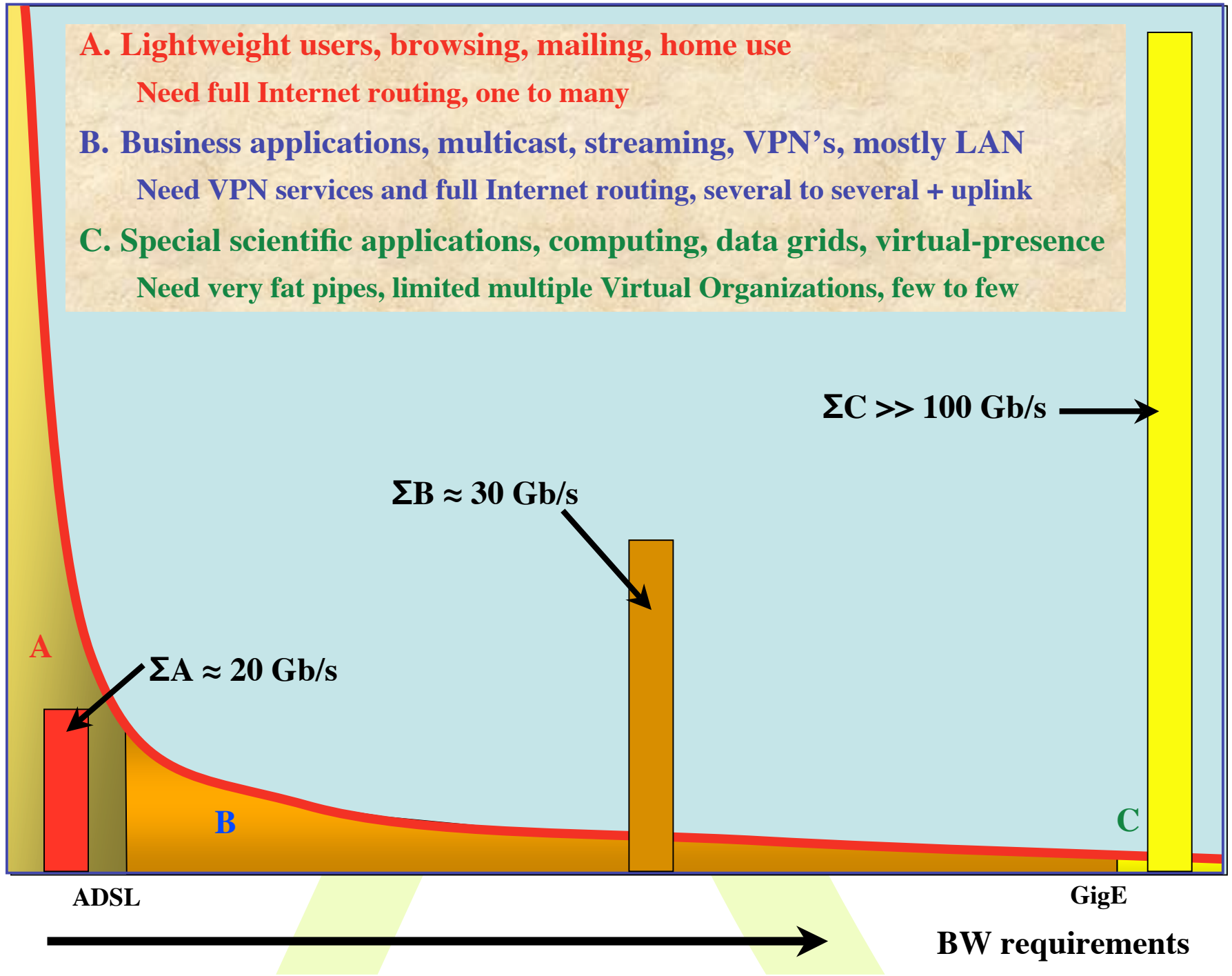
B

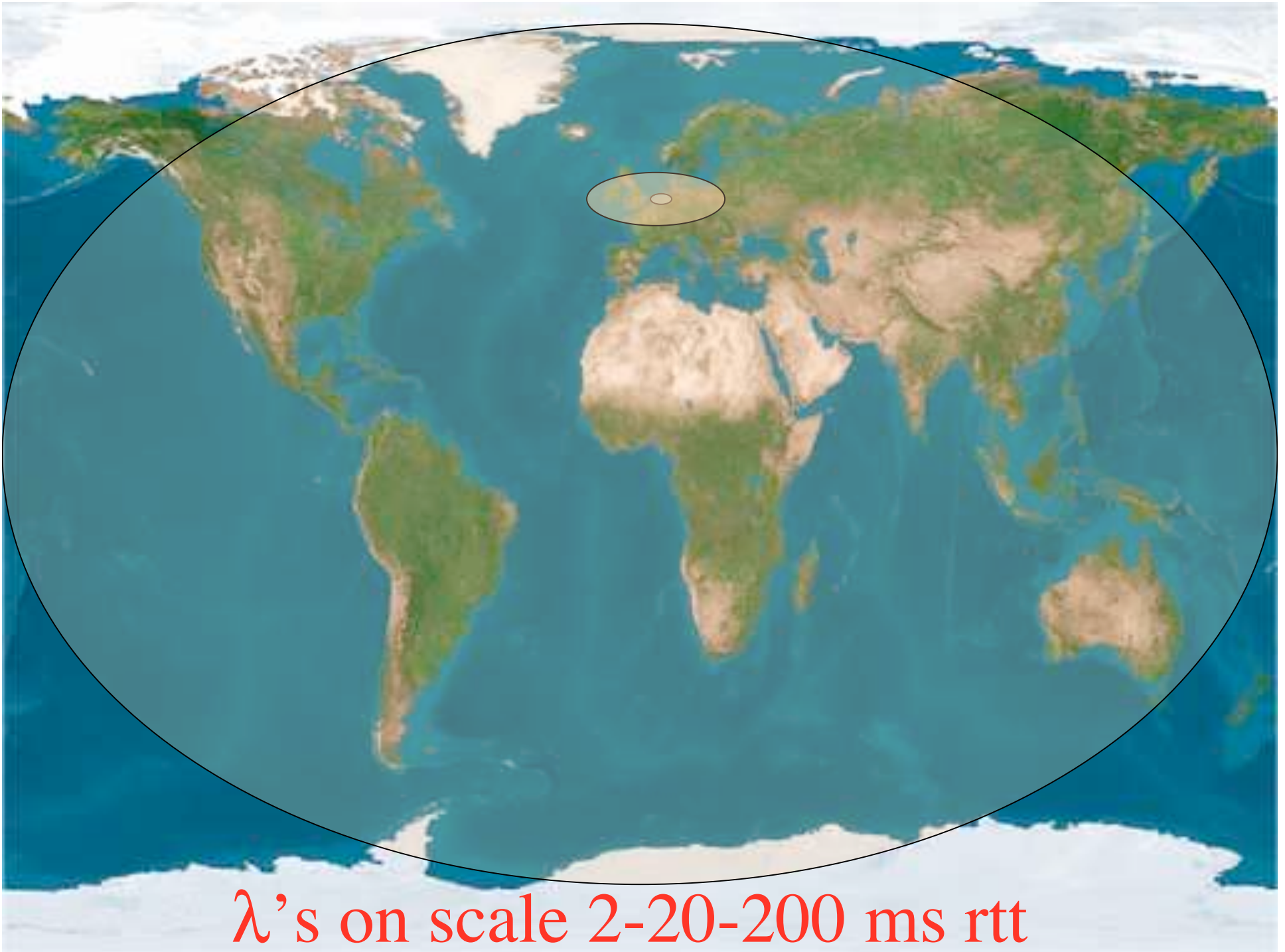
C

ADSL

GigE

→
BW requirements





SURFnet fibers

(pict outdated anytime ;-)



StarLight

NY

UK

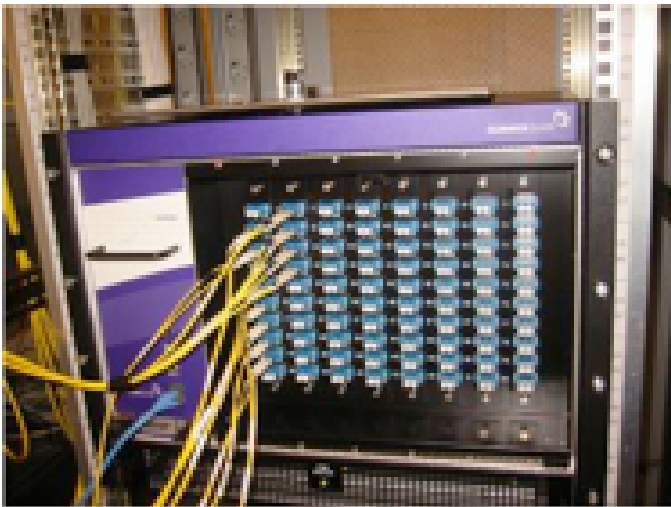
CZ

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

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

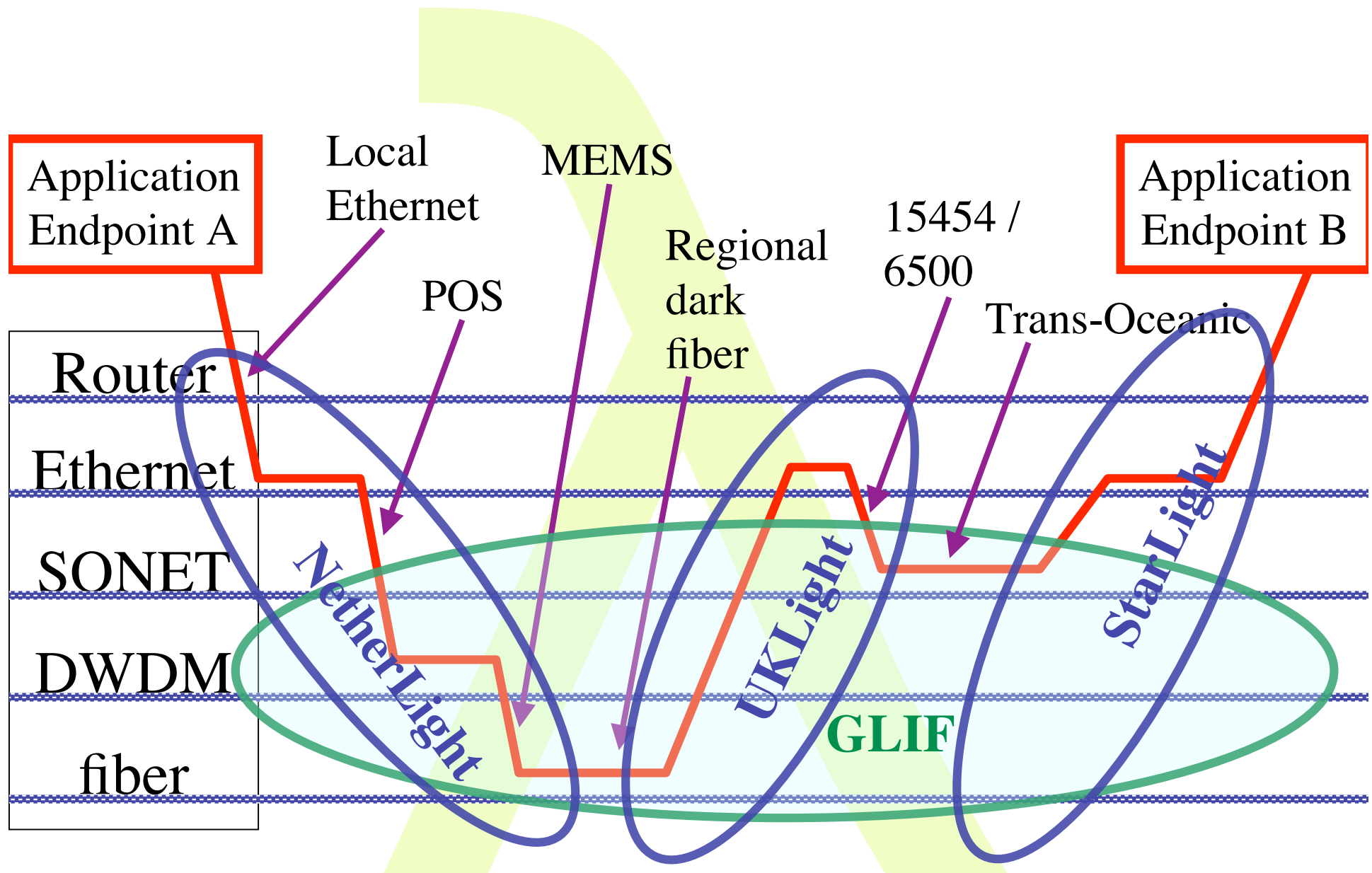


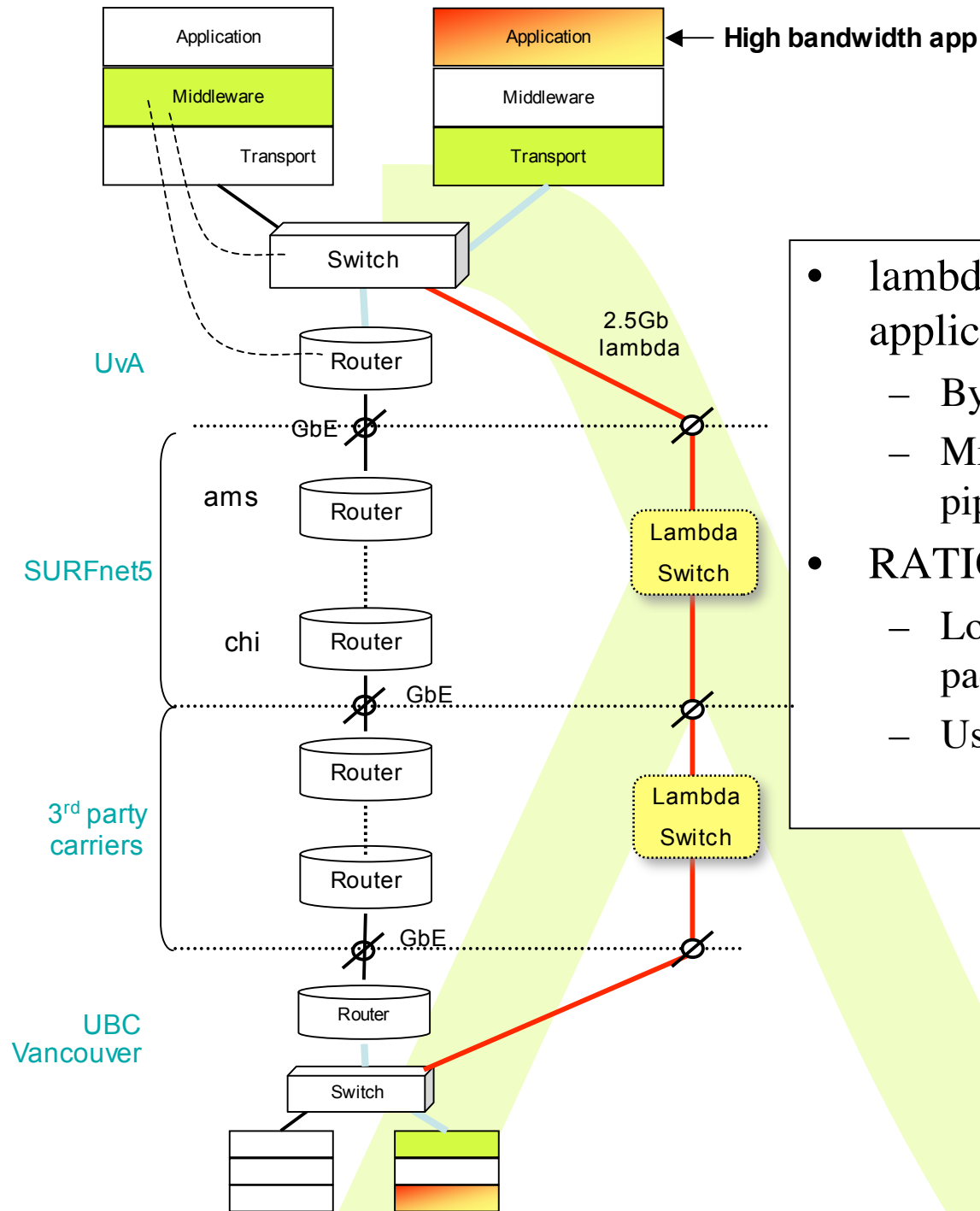
UVA/EVL's
64*64
Optical Switch
@ NetherLight
in SURFnet POP @
SARA

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



How low can you go?





- 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!

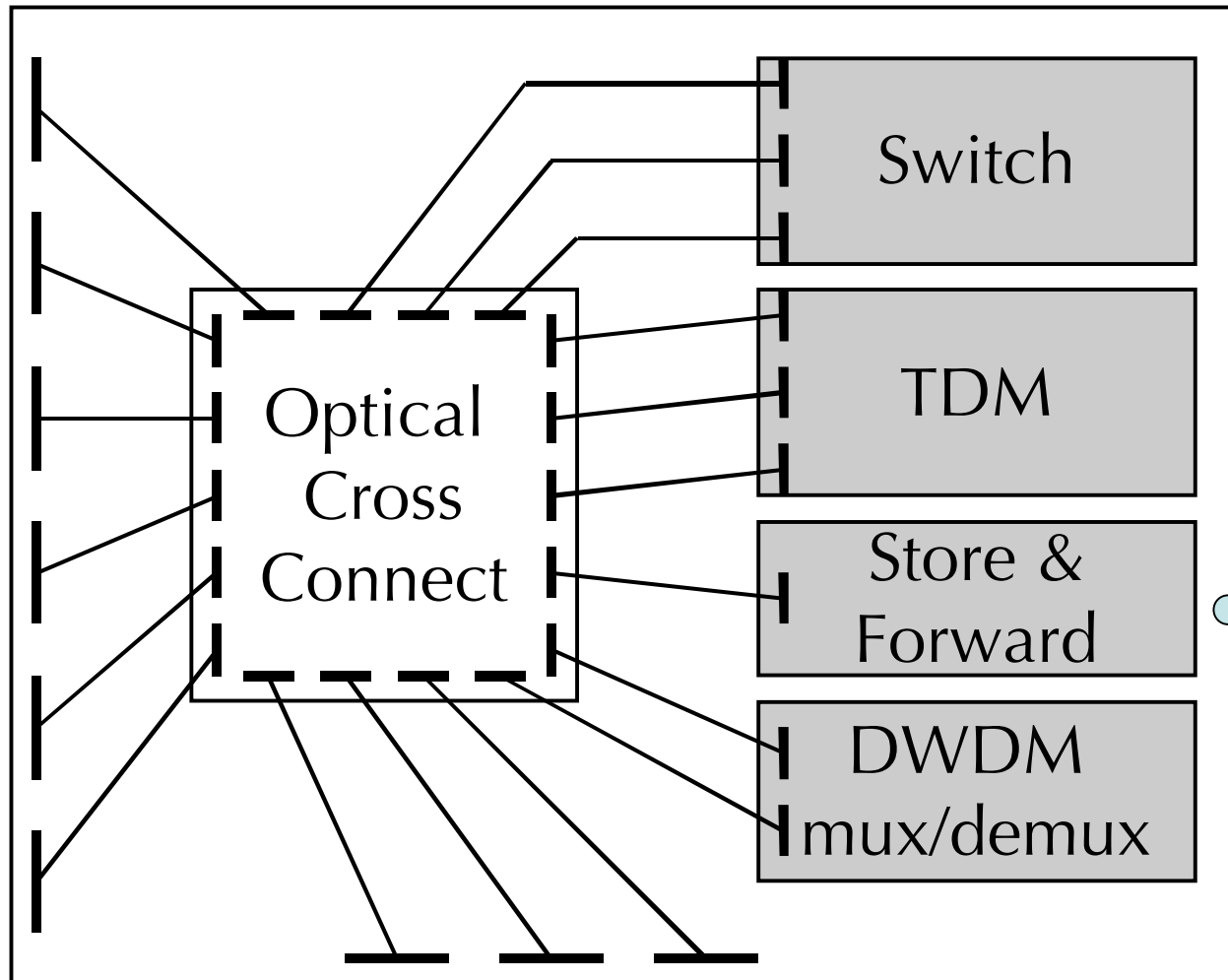


Services

<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

Optical Exchange as Black Box

Optical Exchange

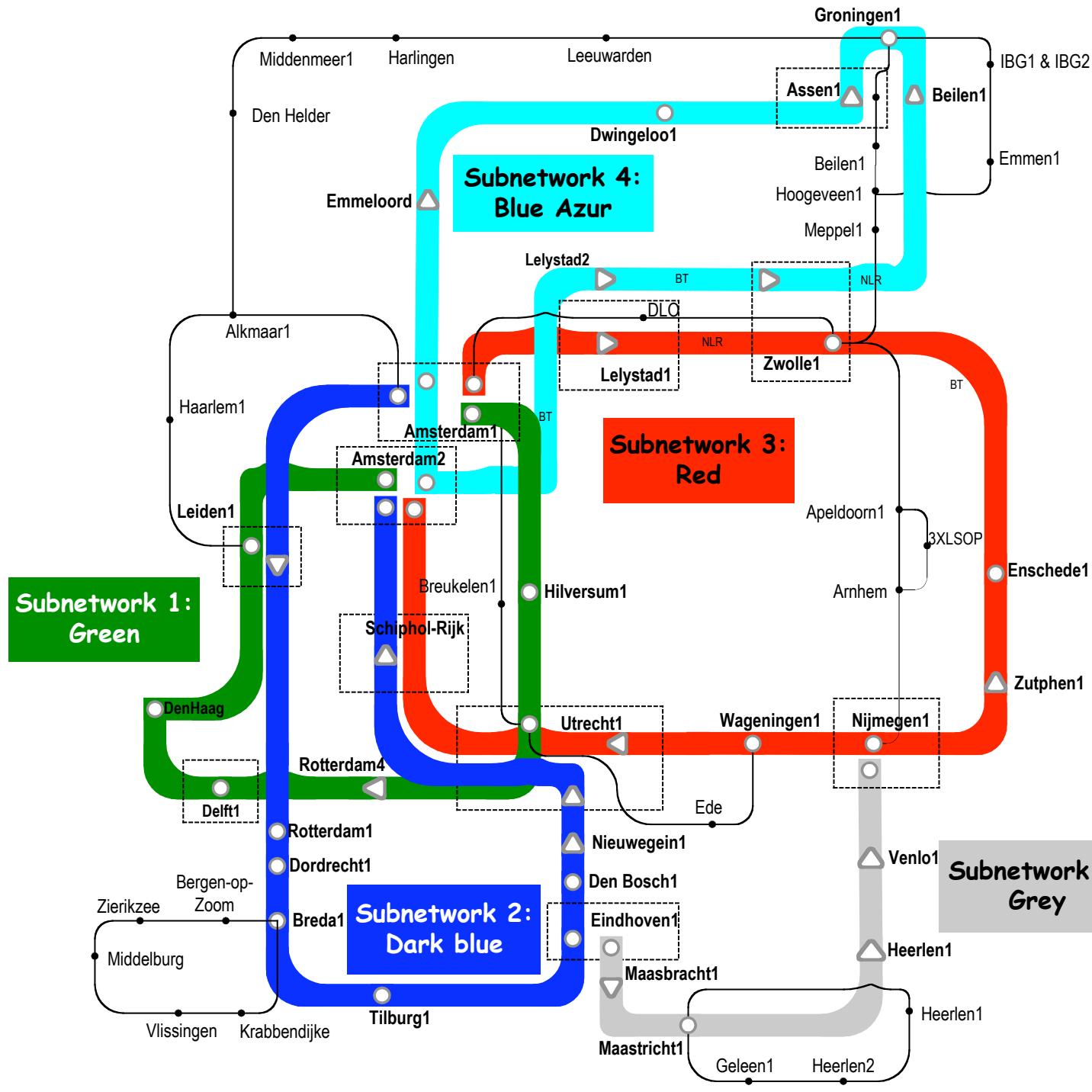


TeraByte
Email
Service

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+

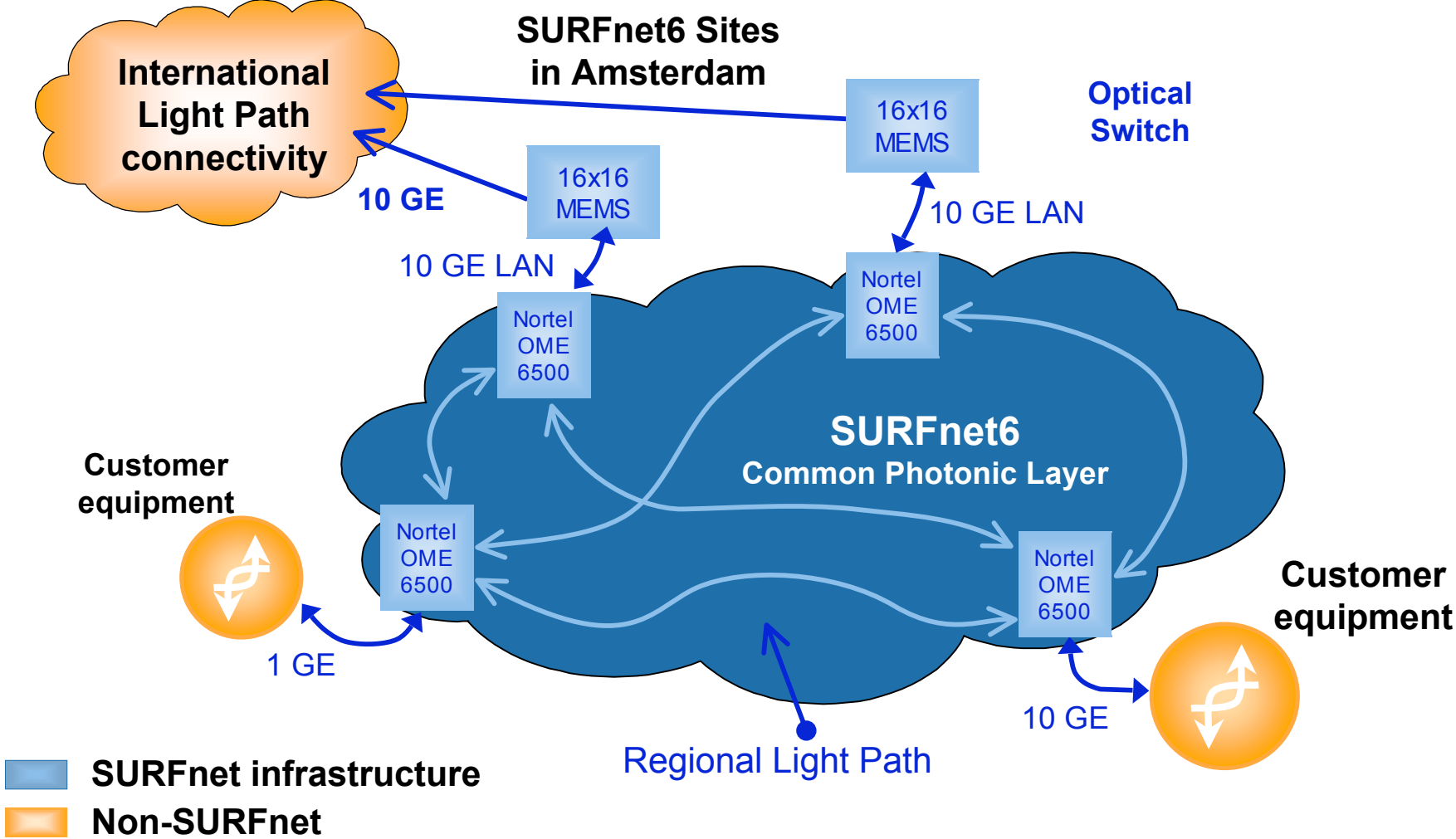
Common Photonic Layer (CPL) in SURFnet6



SURFnet on inspection in Science Park Amsterdam :-)



Light Paths provisioning implementation

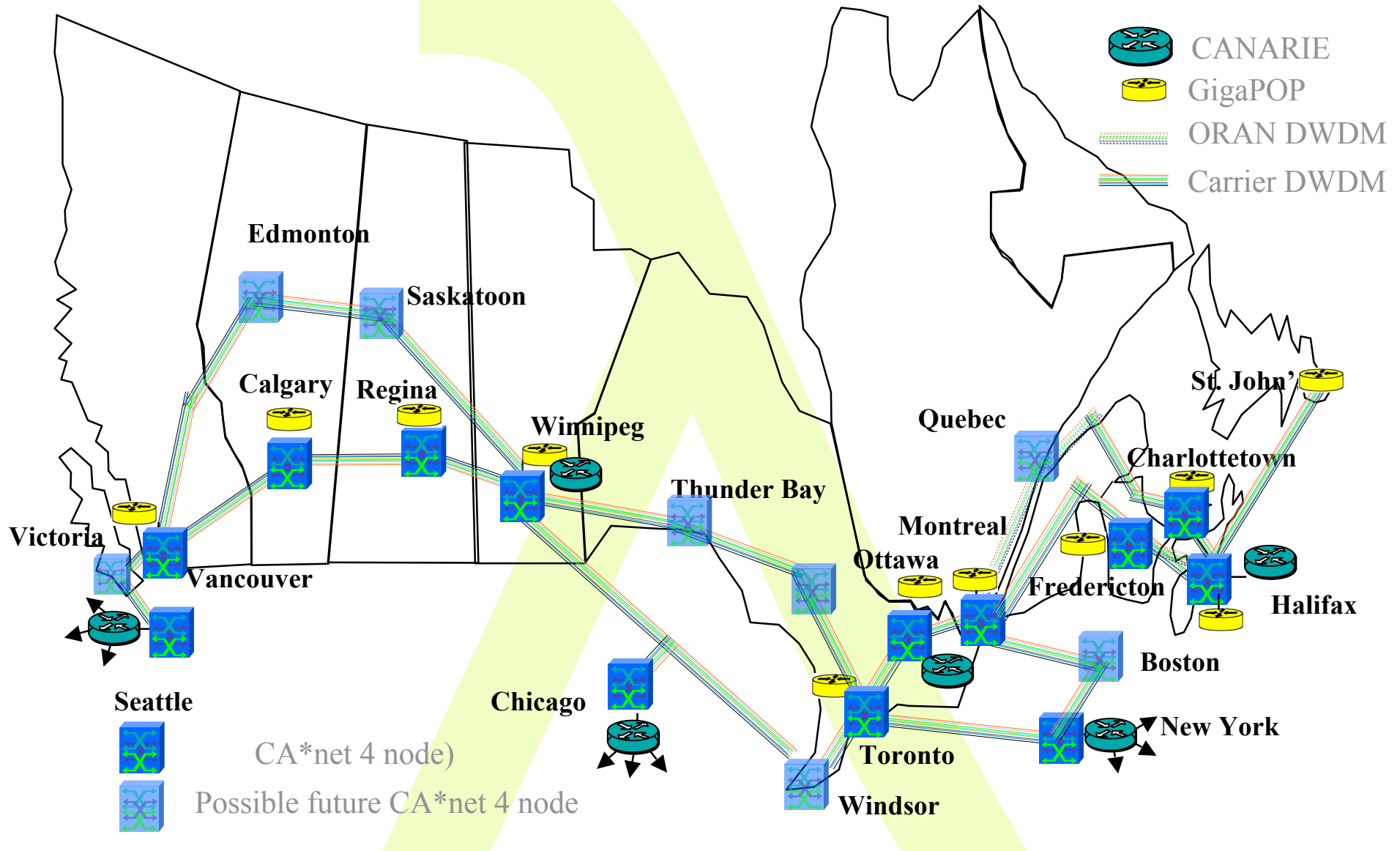


UCLP intended for projects like National LambdaRail

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

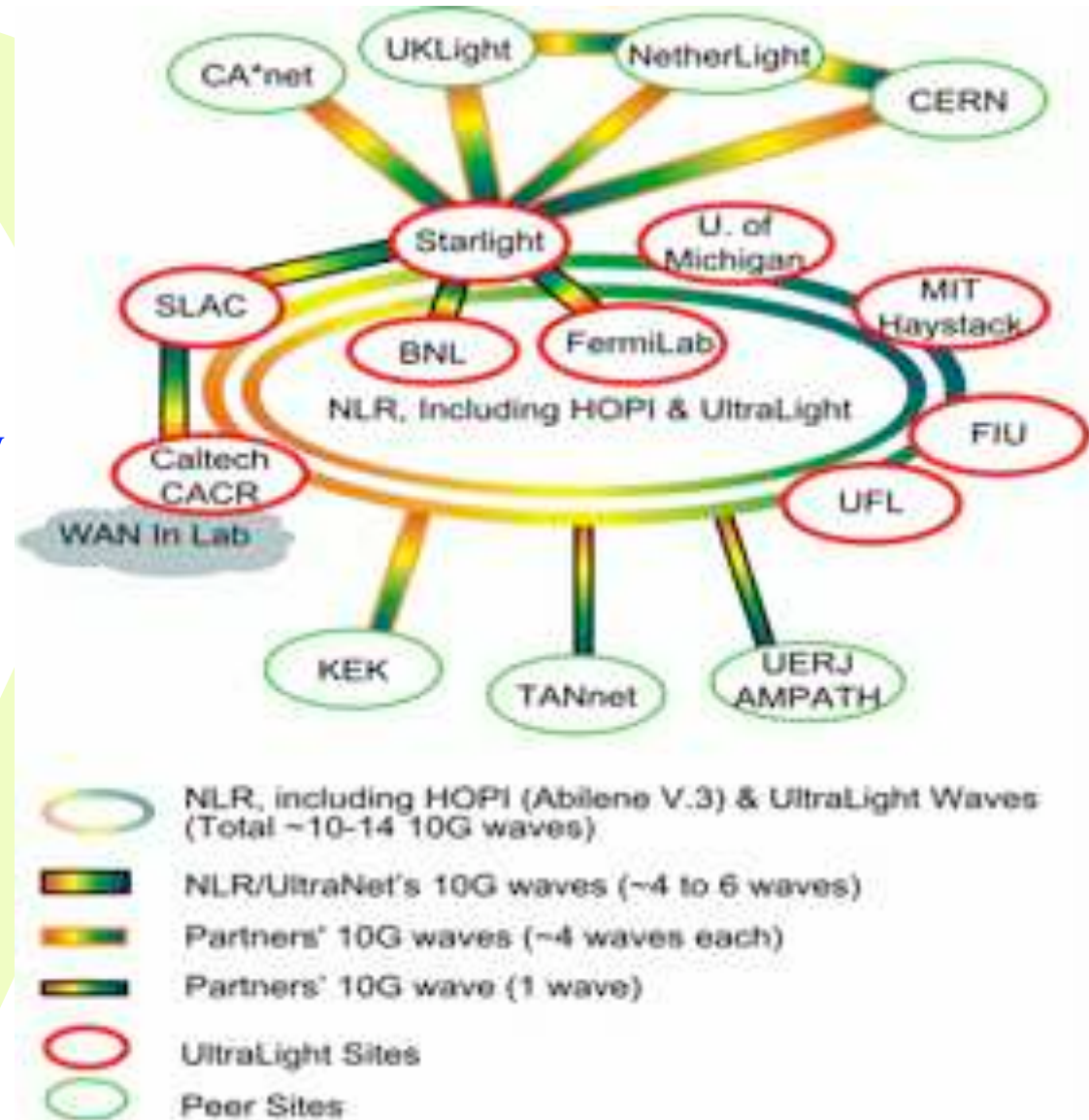


CA*net 4 Architecture

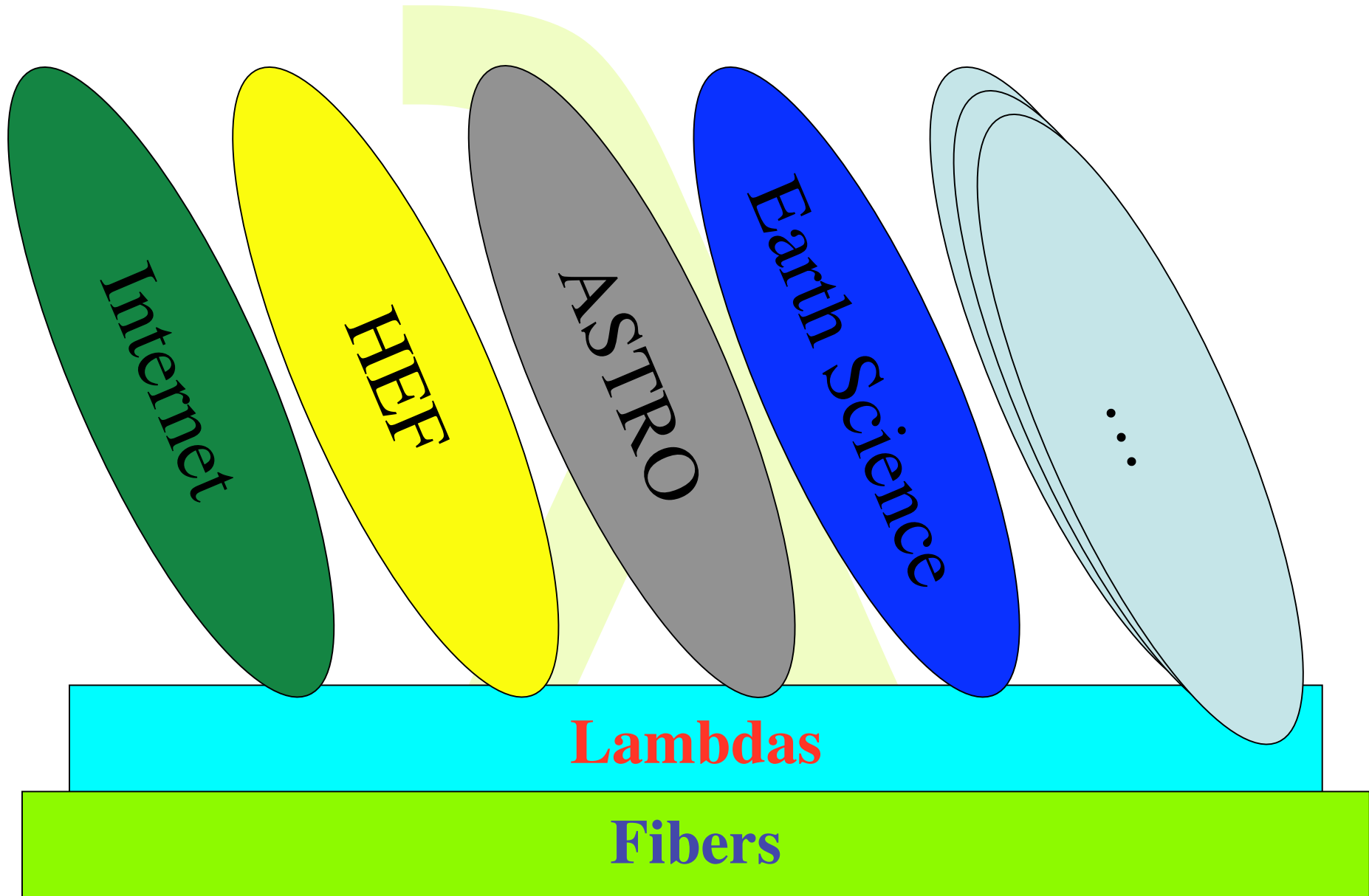


UltraLight Network: PHASE III

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

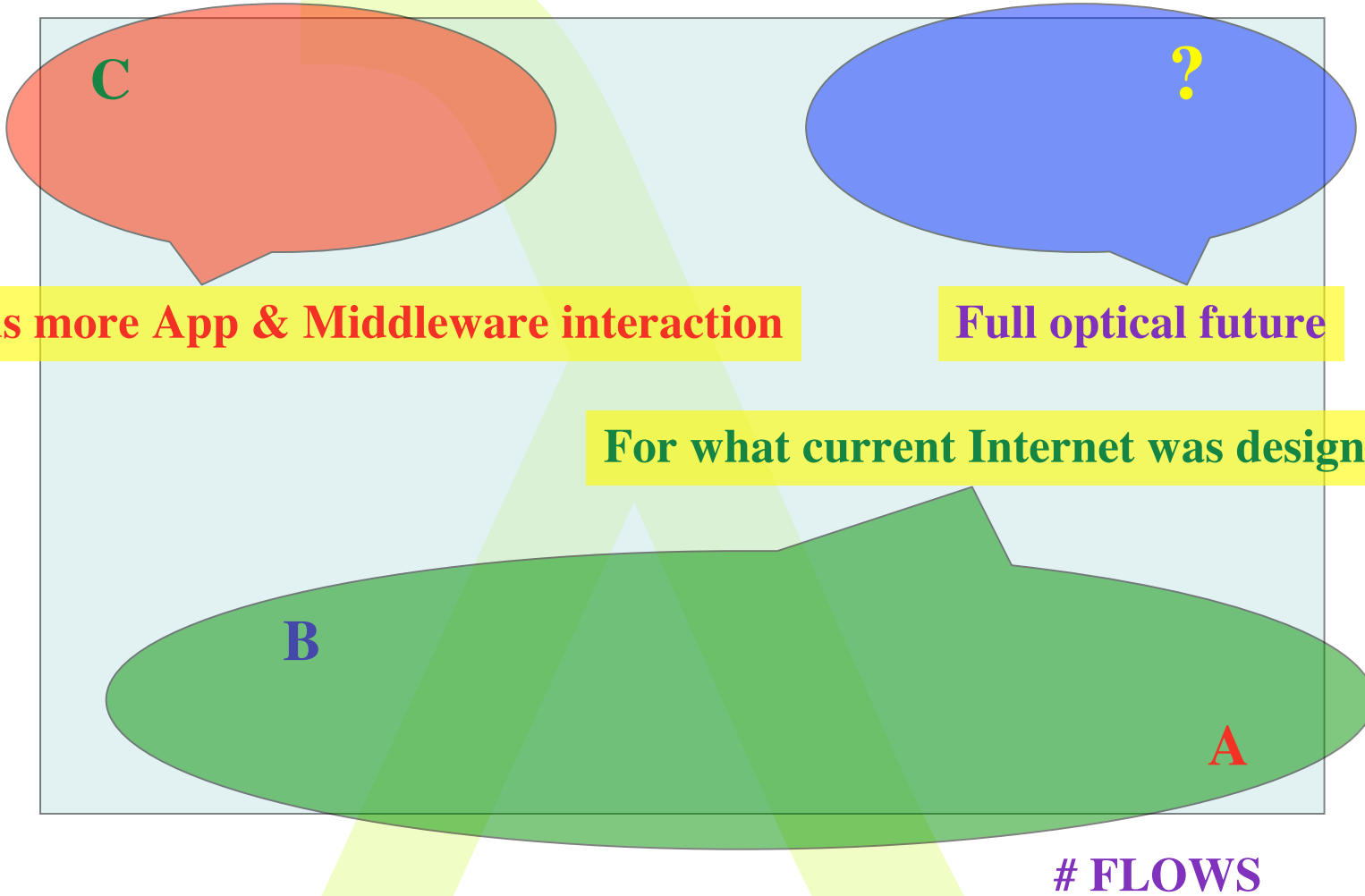


Discipline Networks



Transport in the corners

$BW * RTT$



Needs more App & Middleware interaction

Full optical future

For what current Internet was designed

$\# \text{ FLOWS}$

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!



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**

Nottingham 2004



GLIF Q3 2004



Visualization courtesy of
Bob Patterson, NCSA.

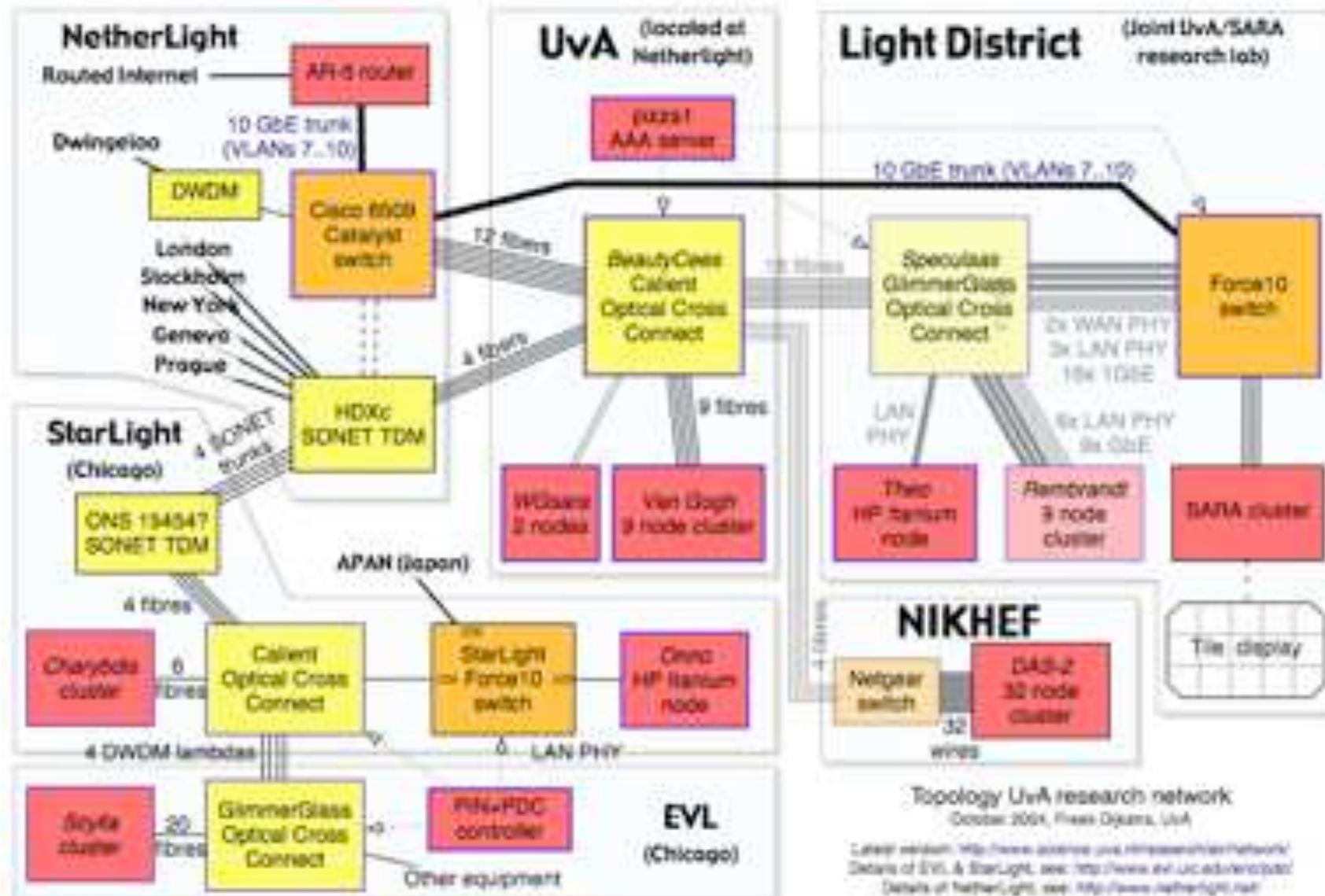
- **Optical Networking:** What innovation in architectural models, components, control and light path provisioning are needed to integrate dynamically configurable optical transport networks and traditional IP networks to a generic data transport platform that provides end-to-end IP connectivity as well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



Research topics AIR @ UvA

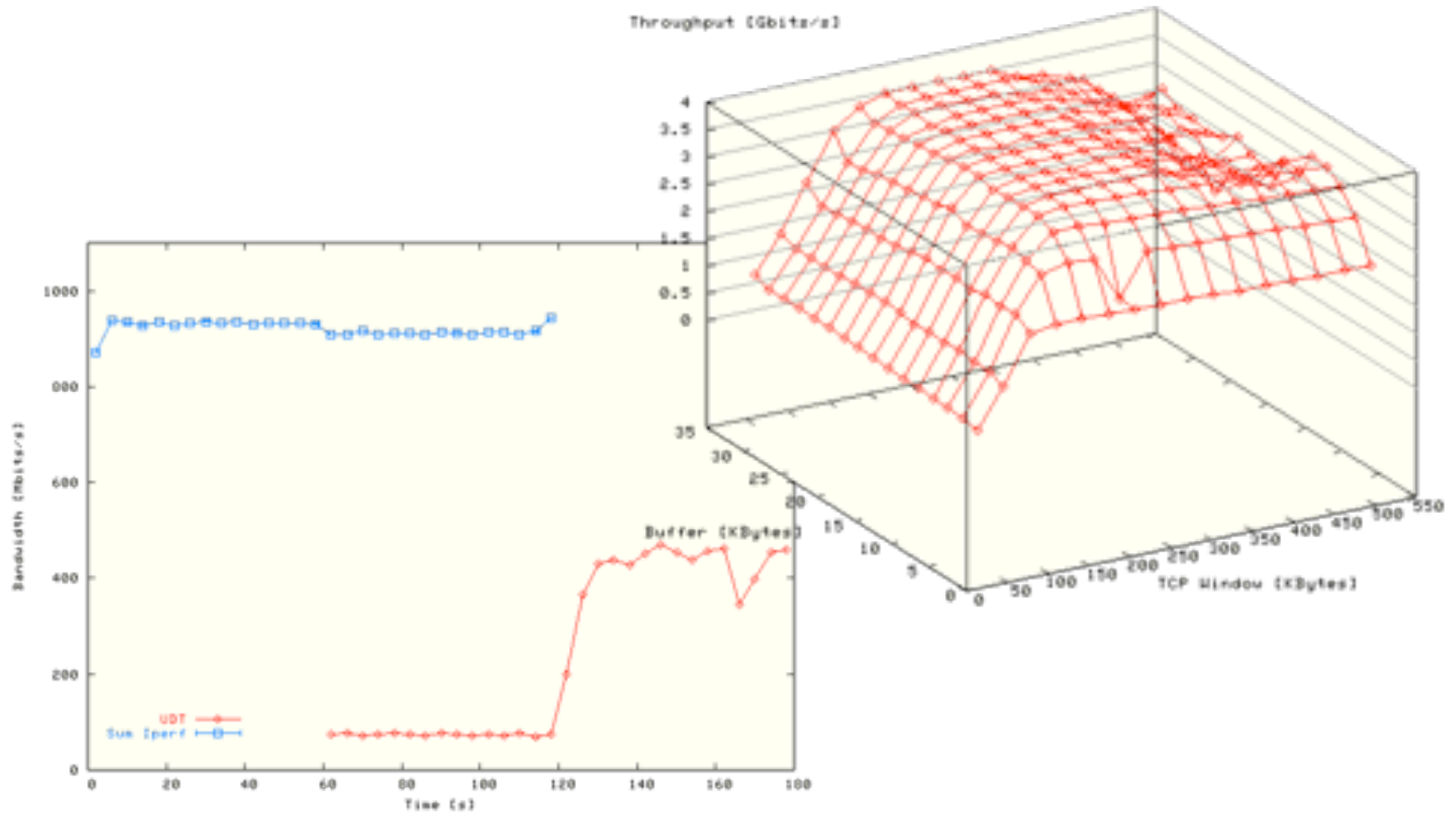
- Optical networking architectures and models for usage
- Transport protocols for massive amounts of data
- Authorization of complex resources in multiple domains
- Embedding in Grid environments

LightHouse





Example Measurements



Layer - 2 requirements from 3/4



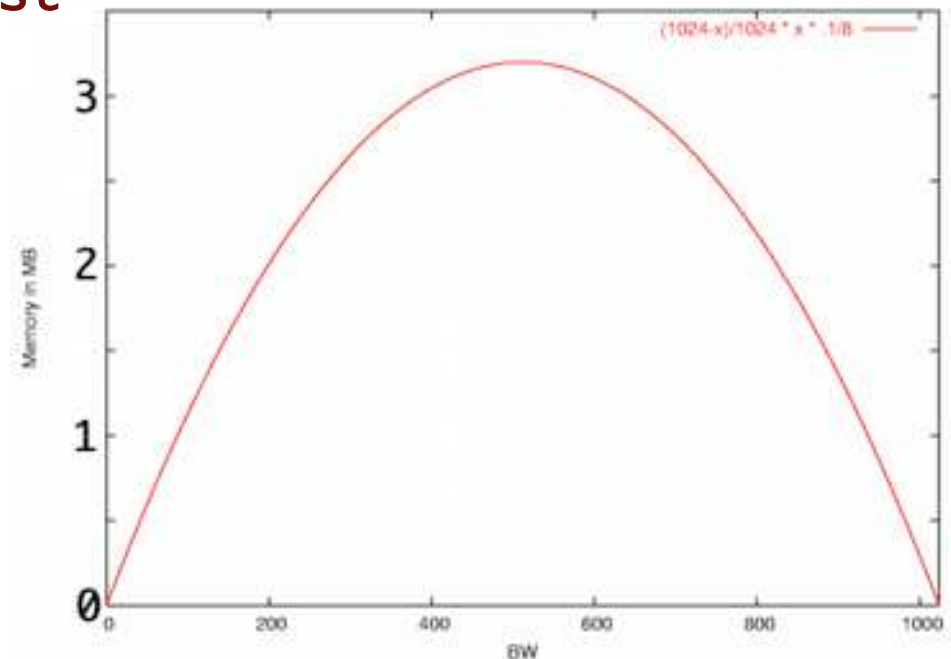
TCP is bursty due to sliding window protocol and slow start algorithm.

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

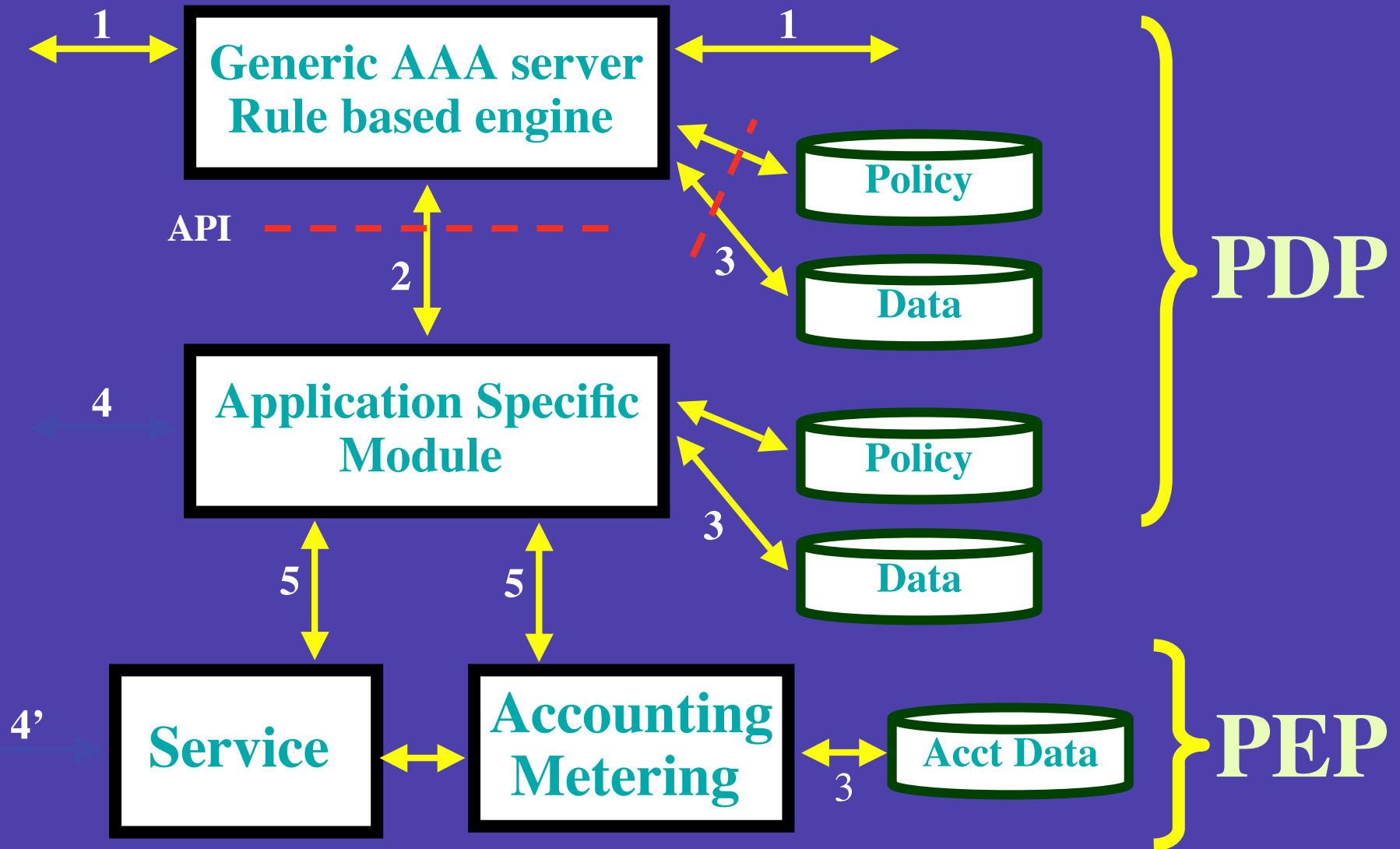
$$\text{Memory-at-bottleneck} = \frac{\text{fast} - \text{slow}}{\text{fast}} * \text{slow} * \text{RTT}$$

So pick from menu:

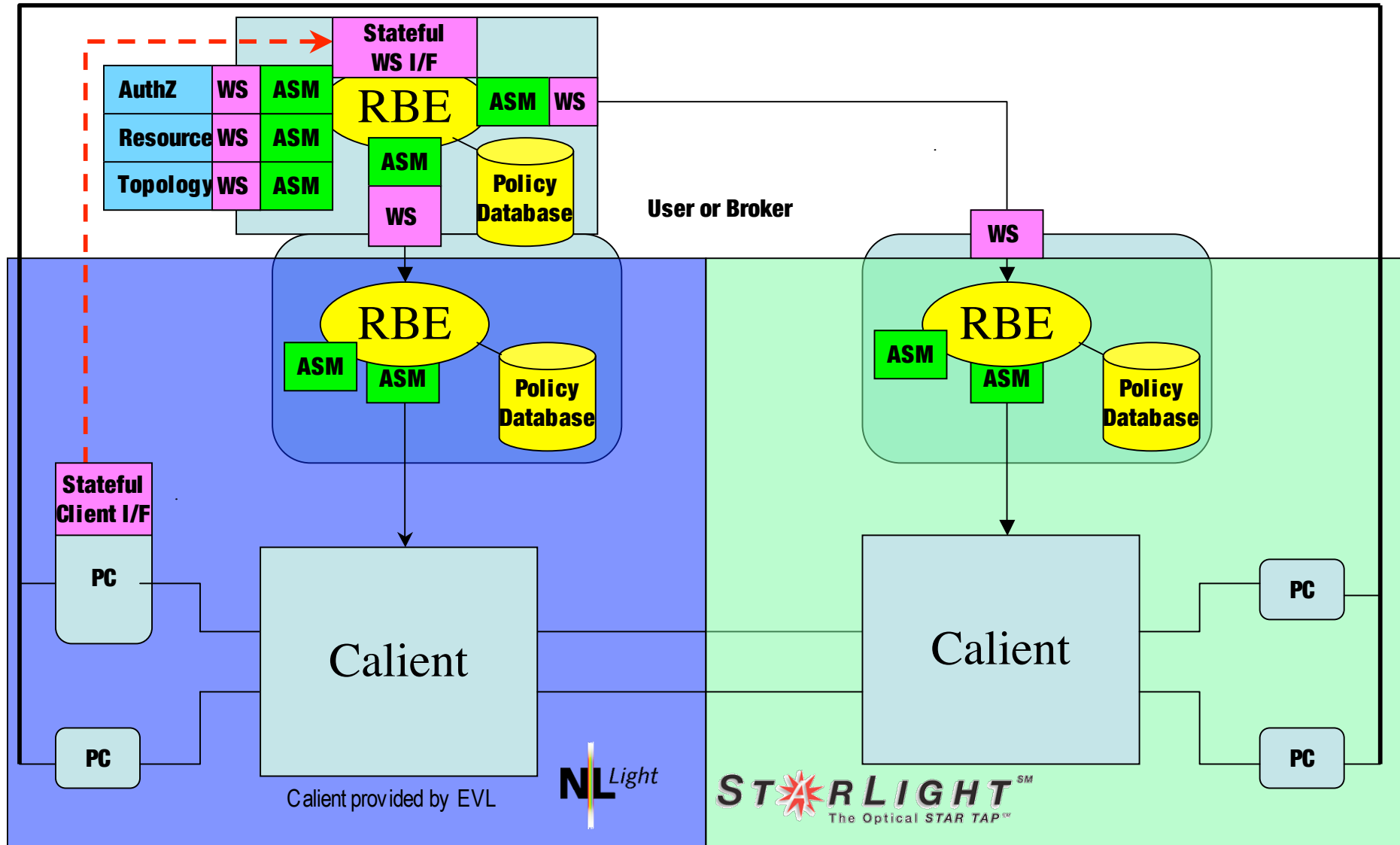
- ◆ *Flow control*
- ◆ *Traffic Shaping*
- ◆ *RED (Random Early Discard)*
- ◆ *Self clocking in TCP*
- ◆ *Deep memory*



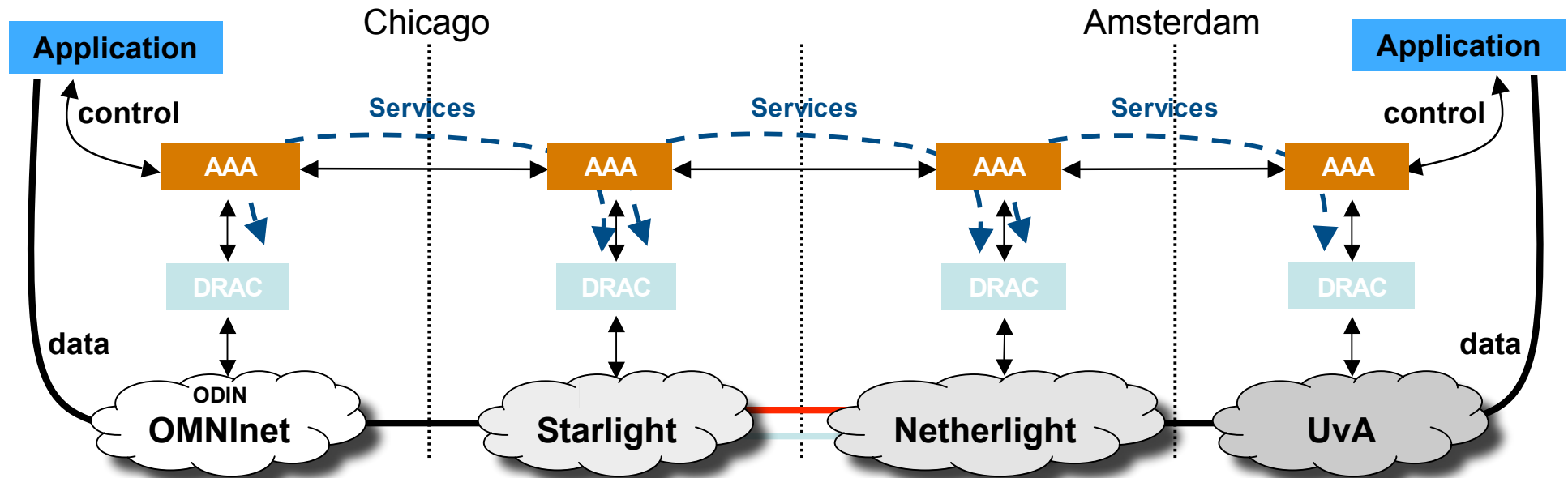
Starting point



RFC 2903 - 2906 , 3334 , policy draft

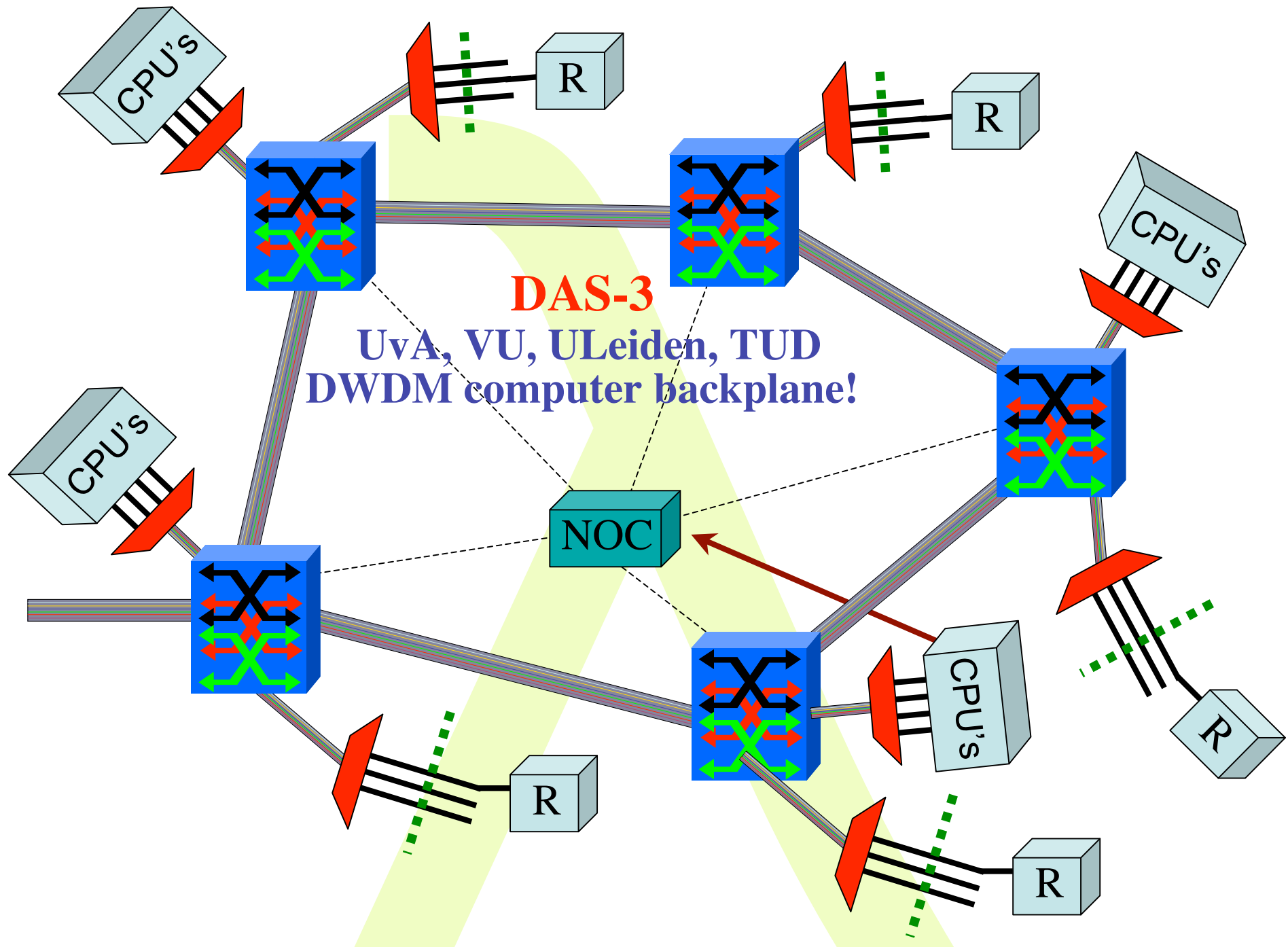


SC2004 CONTROL CHALLENGE



- 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





Important notes

- **Dark fiber based Multi Lambda infrastructures are emerging**
- **Networks are build ON TOP OF THEM !**
- **Need for a hybrid (L1, L2, L3) networks to support GRIDs, communities and science disciplines**
- **Need for open policy exchanges**
- **Need for a persistent photonic infrastructure**
- **Need for a European FiberCo!**
- **Participation in the worldwide infrastructure**

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

The END

Thanks to

SURFnet: Kees Neggens, 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

Case
Delaat

3/12/2003
9:00 AM - 3:00 PM
Wednesday

Partially complete list:

Caas
Chase
Cess
Kess
Case

