The Rationale of the Current Optical Networking Initiatives

www.science.uva.nl/~delaat

Cees de Laat

Faculty of Science

The Rationale of the Current Optical Networking Initiatives

www.science.uva.nl/~delaat

Cees de Laat

EU

SURFnet

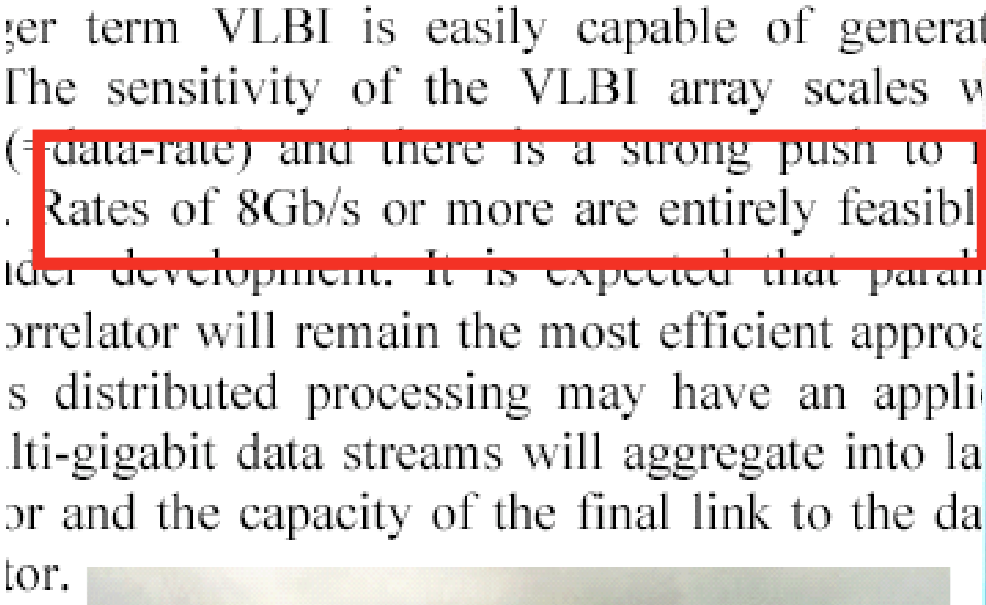
University of Amsterdam

SARA

NIKHEF

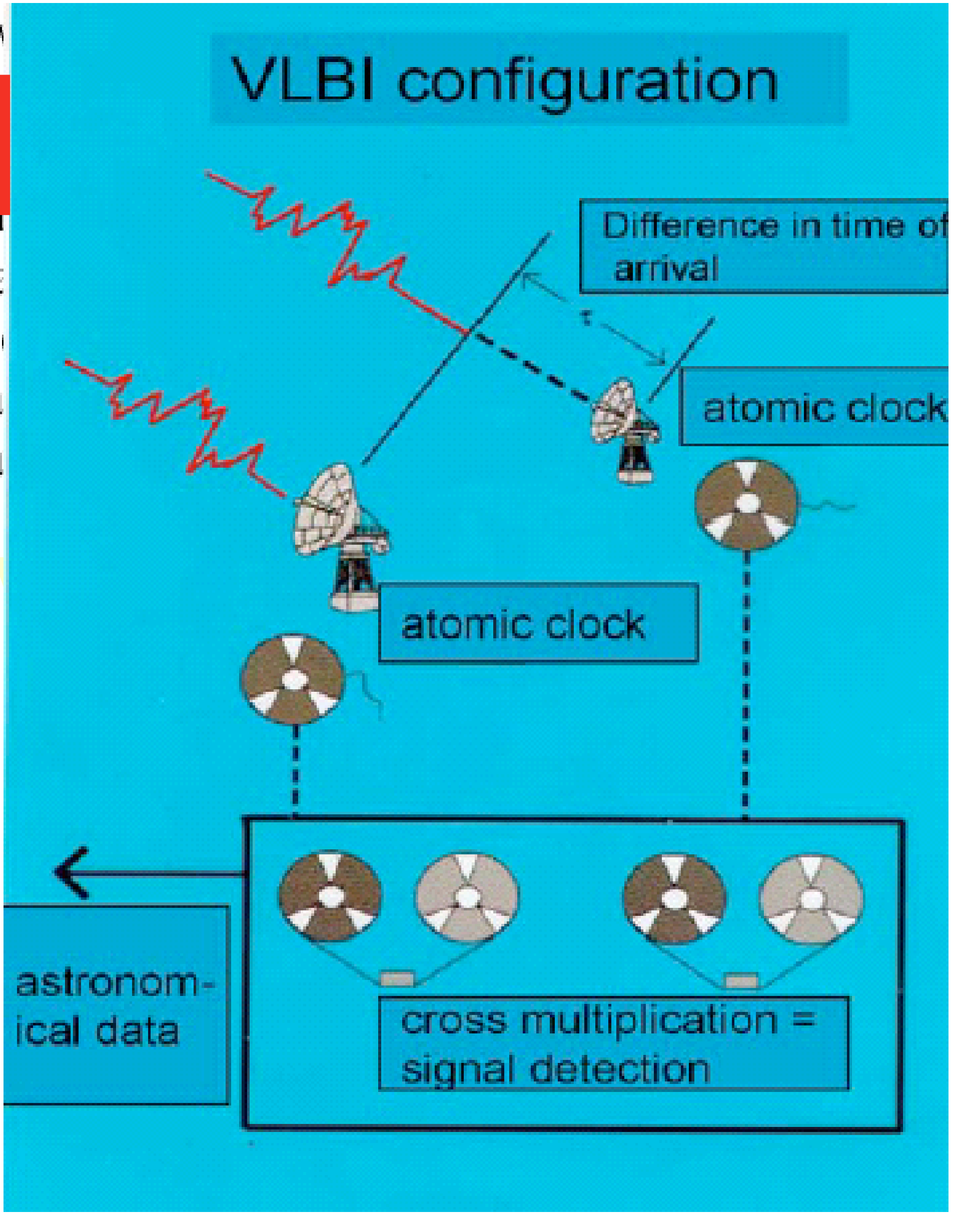
optiputer

VLBI



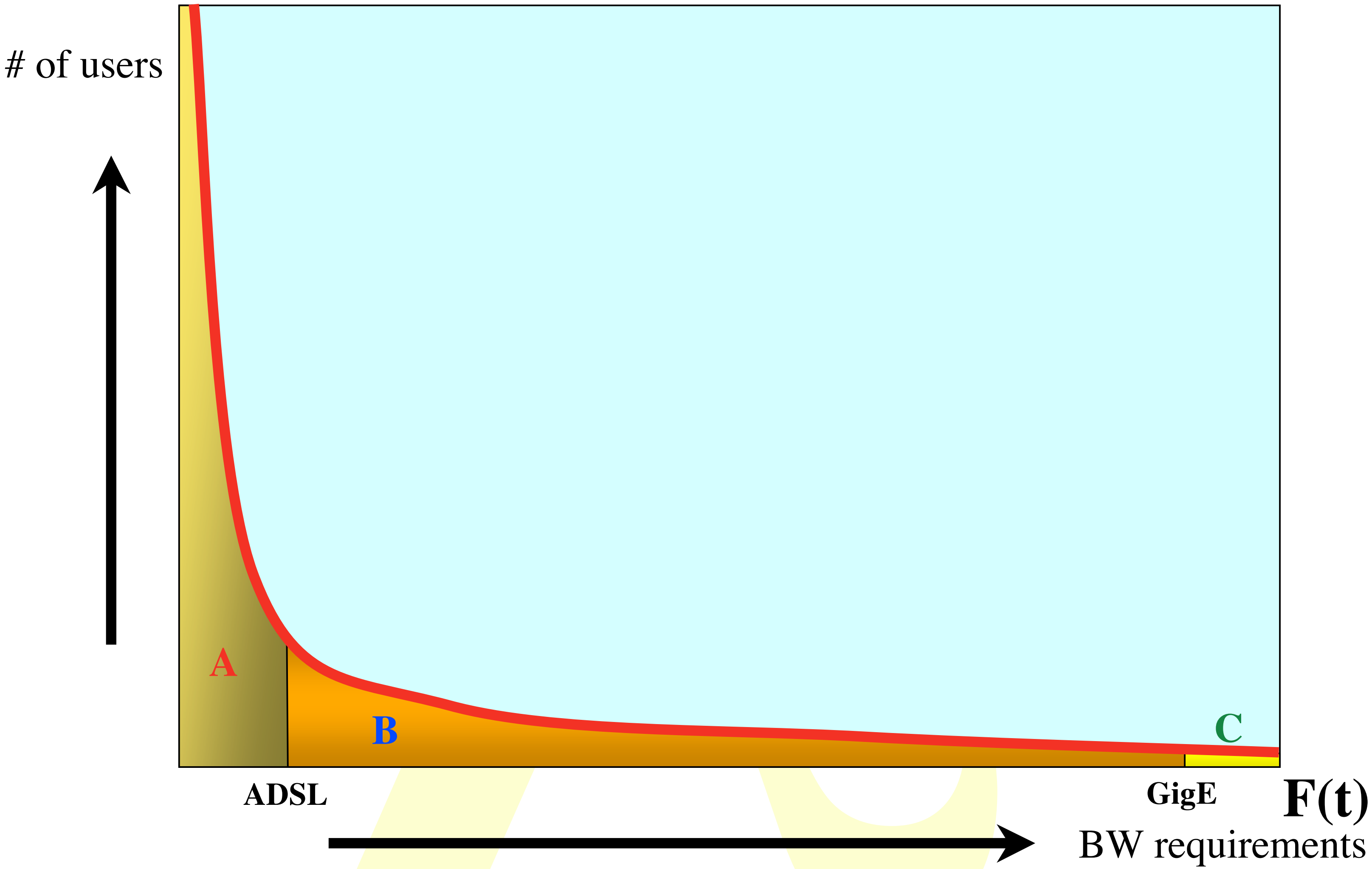
ger term VLBI is easily capable of generating many Gb of data per
The sensitivity of the VLBI array scales w
(=data-rate) and there is a strong push to i
. Rates of 8Gb/s or more are entirely feasibl
ider development. It is expected that parall
orrelator will remain the most efficient approa
s distributed 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



VLBI configuration
Difference in time of arrival
τ
atomic clock
atomic clock
astronomical data
cross multiplication = signal detection

Know the user

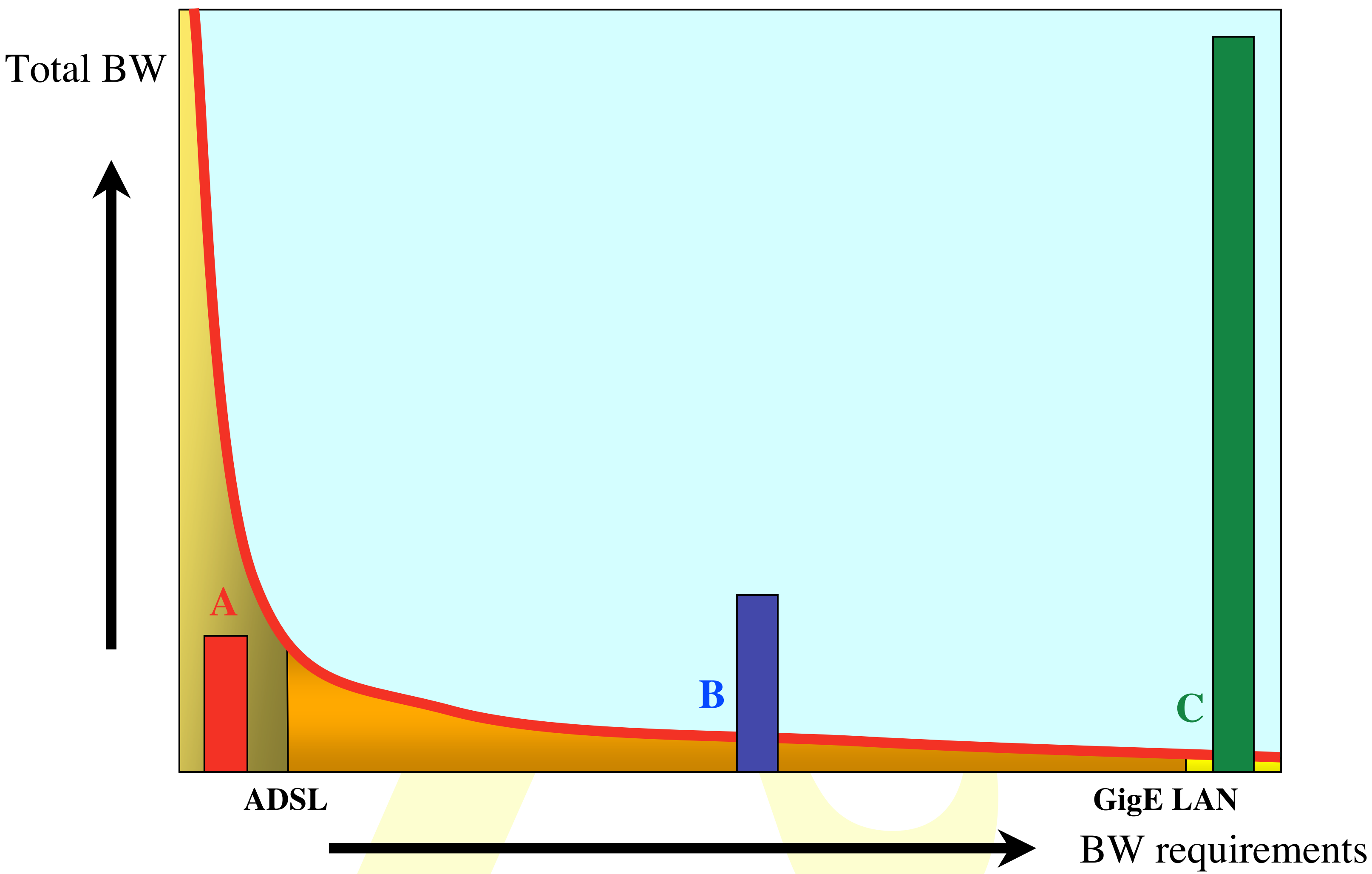


A -> Lightweight users, browsing, mailing, home use

B -> Business applications, multicast, streaming, VPN’s, mostly LAN

C -> Special scientific applications, computing, data grids, virtual-presence

What the user



A -> Need full Internet routing, one to many

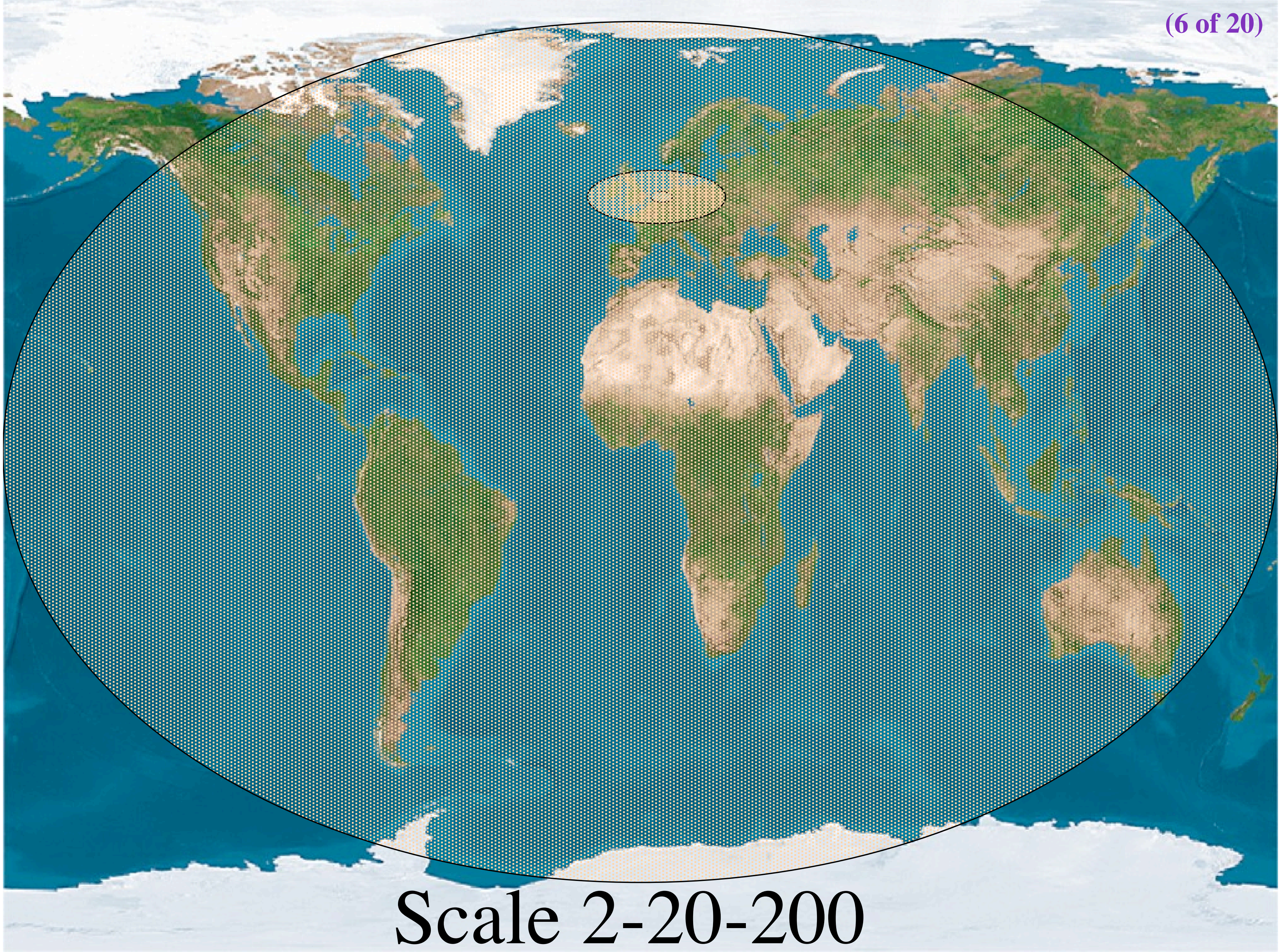
B -> Need VPN services on/and full Internet routing, several to several

C -> Need very fat pipes, limited multiple Virtual Organizations, few to few

So what are the facts

- **Costs of fat pipes (fibers) are one/third of equipment to light them up**
  - **Is what Lambda salesmen tell me**
- **Costs of optical equipment 10% of switching 10 % of full routing equipment for same throughput**
  - **100 Byte packet @ 10 Gb/s -> 80 ns to look up in 100 Mbyte routing table (light speed from me to you on the back row!)**
- **Big sciences need fat pipes**
- **Bottom line: create a hybrid architecture which serves all users in one consistent cost effective way**

(6 of 20)



Scale 2-20-200

The only formula’s

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

Now, as having been a High Energy Physicist we set

c = 1

e = 1

ħ = 1

and the formula reduces to:

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

Services

	2 Metro	20 National/ regional	200 World
A	Switching/ routing	Routing	ROUTER$
B	VPN's, (G)MPLS	VPN's Routing	Routing
C $\# \lambda \approx \frac{200 * e^{(t-2002)}}{rtt}$	dark fiber Optical switching	Lambda switching	Sub-lambdas, ethernet-sdh

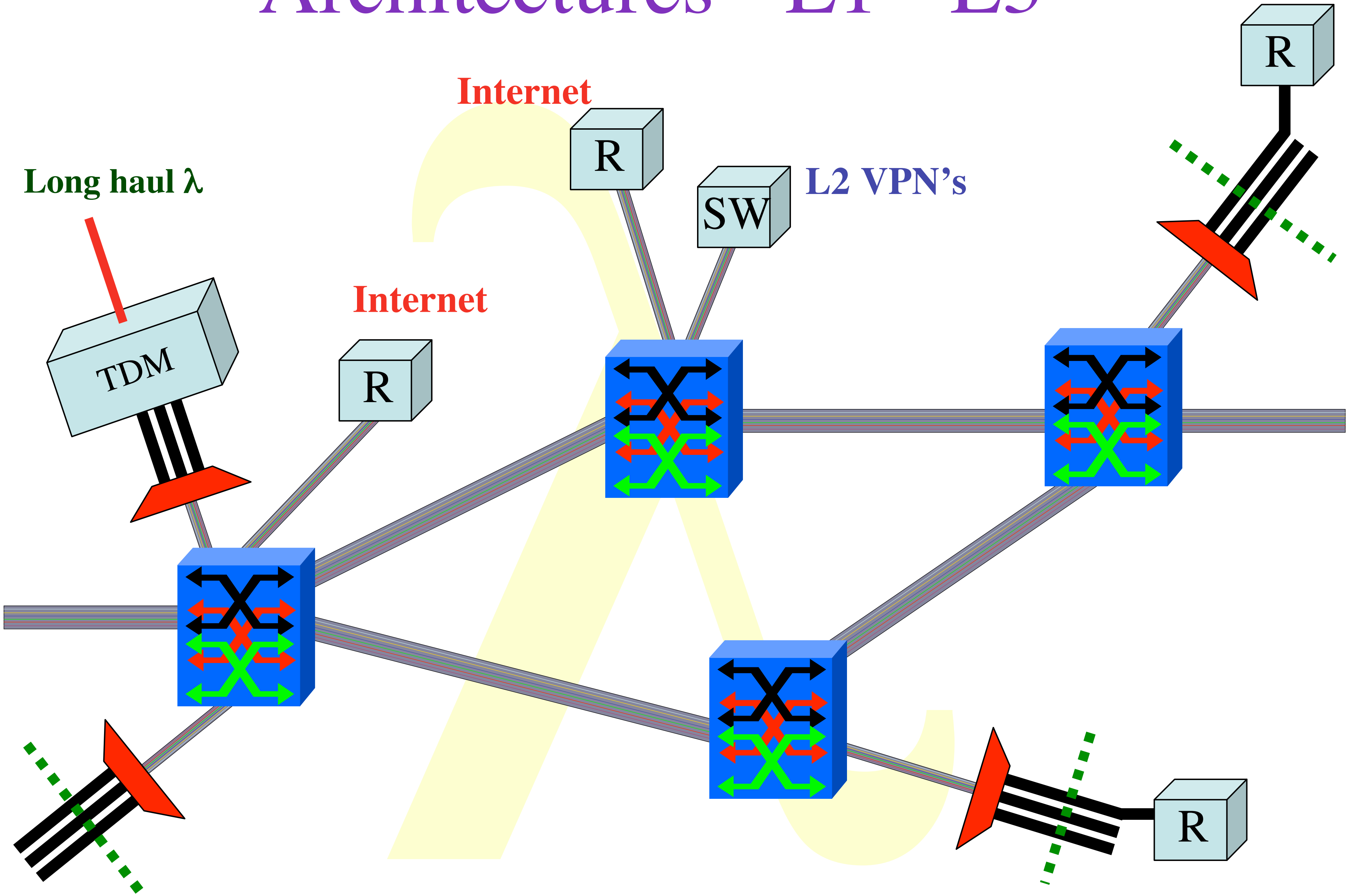
Current technology + (re)definition

- Current (to me) available technology consists of SONET/SDH switches
- Changing very soon!, optical switch on the way!
- DWDM+switching coming up
- Starlight uses for the time being VLAN’s on Ethernet switches to connect [exactly two] ports (but also routing)
- We want to understand routerless limited environments
- So redefine a λ as:

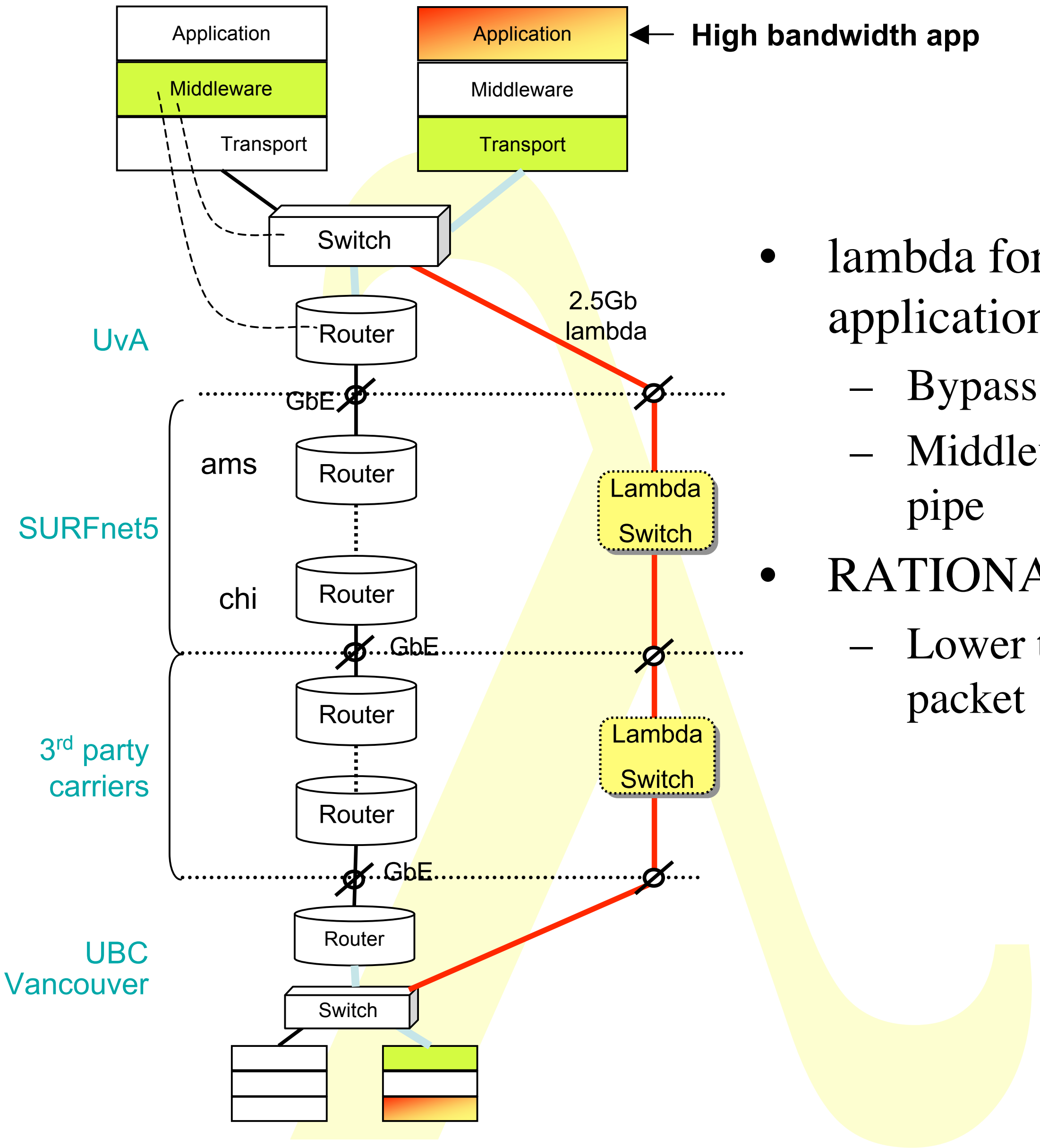
> **“a λ is a pipe where you can inspect packets as they enter and when they exit, but principally not when in transit. In transit one only deals with the parameters of the pipe: number, color, bandwidth”**

Architectures - L1 - L3

(10 of 20)

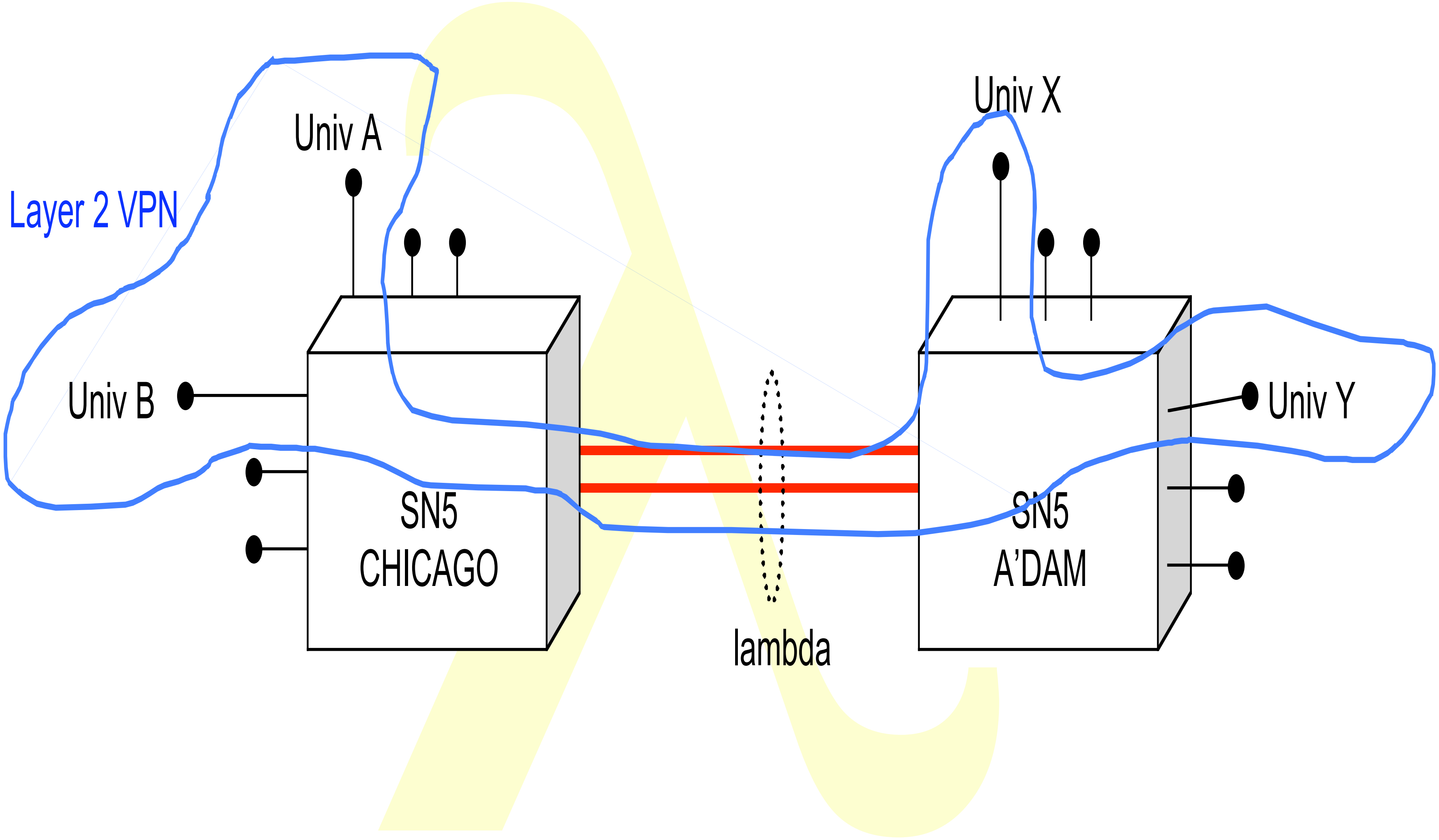


Bring plumbing to the users, not just create sinks in the middle of nowhere

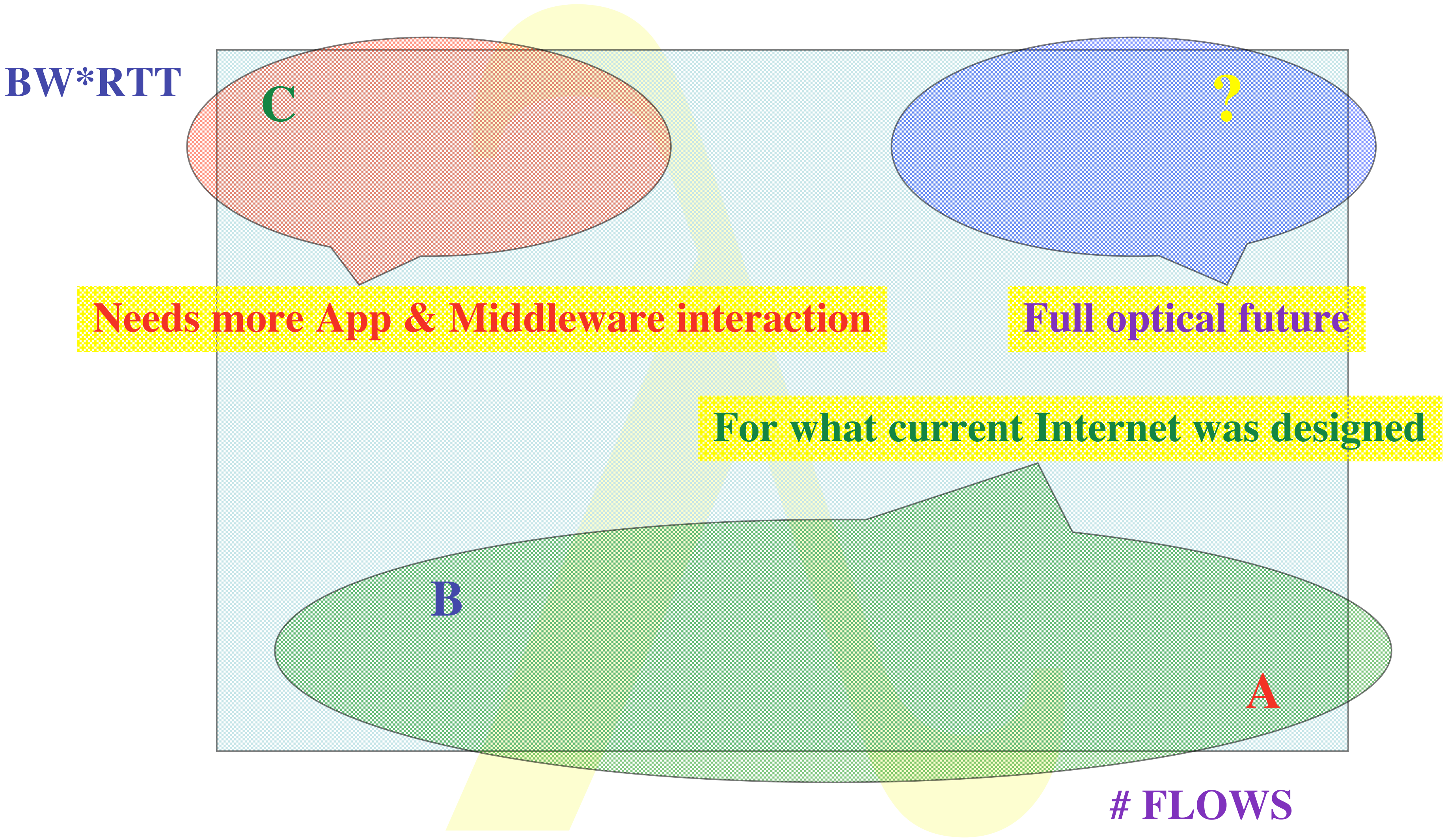


- lambda for high bandwidth applications
  - Bypass of production network
  - Middleware may request (optical) pipe
- RATIONALE:
  - Lower the cost of transport per packet

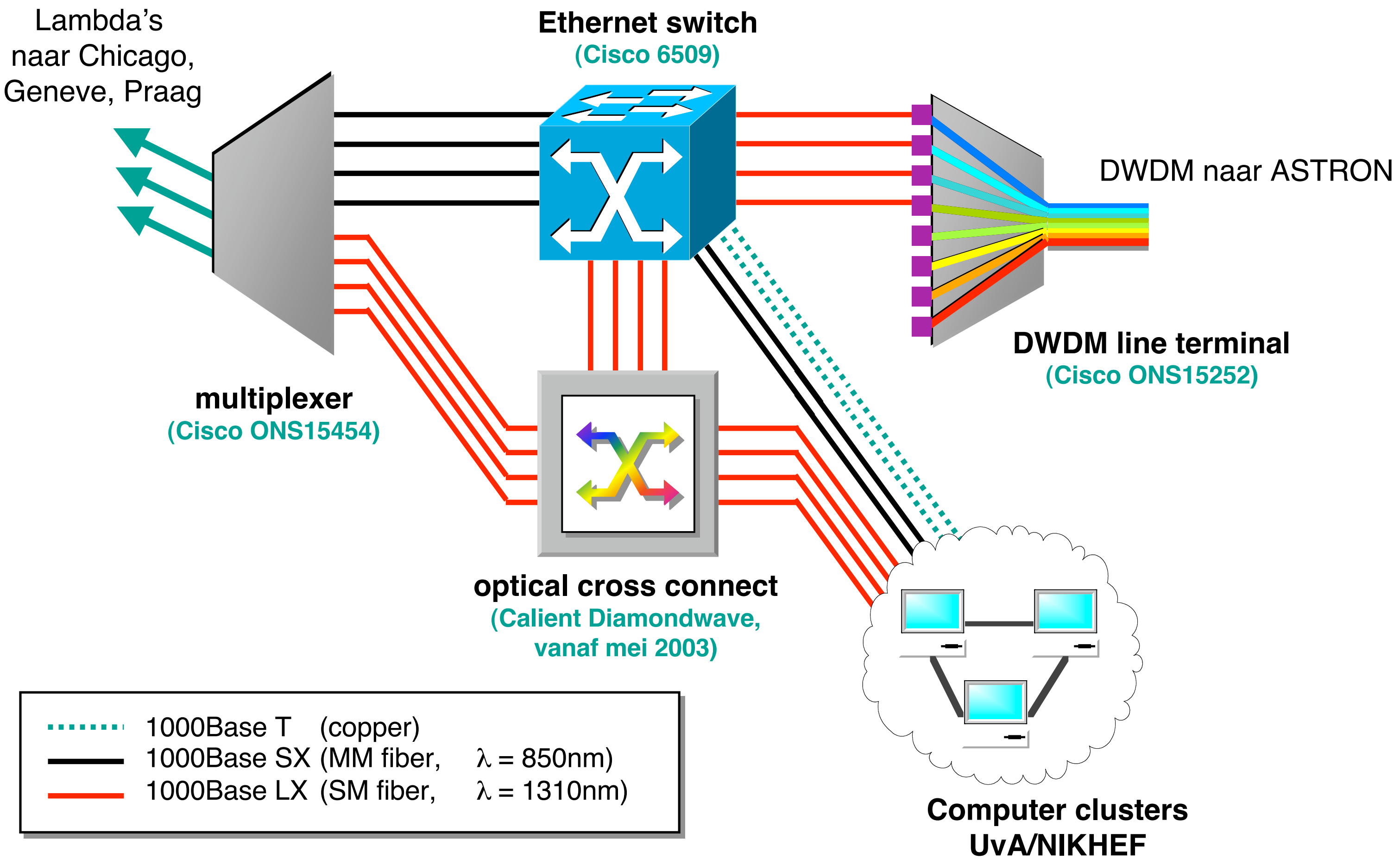
Distributed L2



Transport in the corners


BW*RTT
C
?
Needs more App & Middleware interaction
Full optical future
For what current Internet was designed
B
A
# FLOWS

NetherLight Amsterdam setup



Lambda's naar Chicago, Geneve, Praag

Ethernet switch
(Cisco 6509)

DWDM naar ASTRON

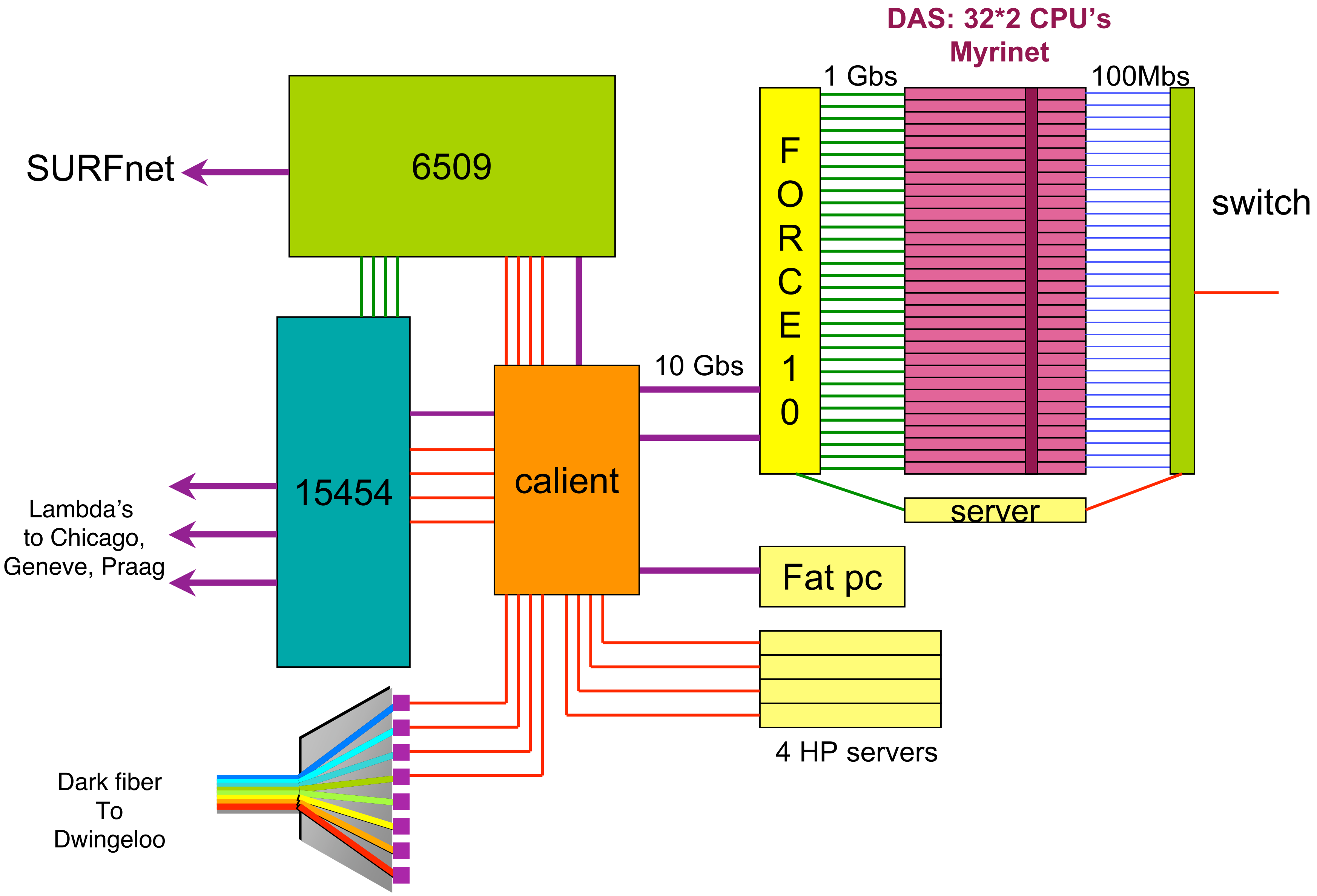
DWDM line terminal
(Cisco ONS15252)

multiplexer
(Cisco ONS15454)

optical cross connect
(Calient Diamondwave, vanaf mei 2003)

Computer clusters
UvA/NIKHEF

⋯	1000Base T	(copper)	
━	1000Base SX	(MM fiber,	λ = 850nm)
━	1000Base LX	(SM fiber,	λ = 1310nm)


DAS: 32*2 CPU's
Myrinet
1 Gbs
100Mbs
SURFnet
6509
switch
FORCE10
10 Gbs
15454
calient
server
Lambda's
to Chicago,
Geneve, Praag
Fat pc
4 HP servers
Dark fiber
To
Dwingeloo

GigaPort

SURFnet and GigaPort Next Generation

Creating the innovation engine

Kees Neggers

Managing Director SURFnet

Praha, 20 February 2003

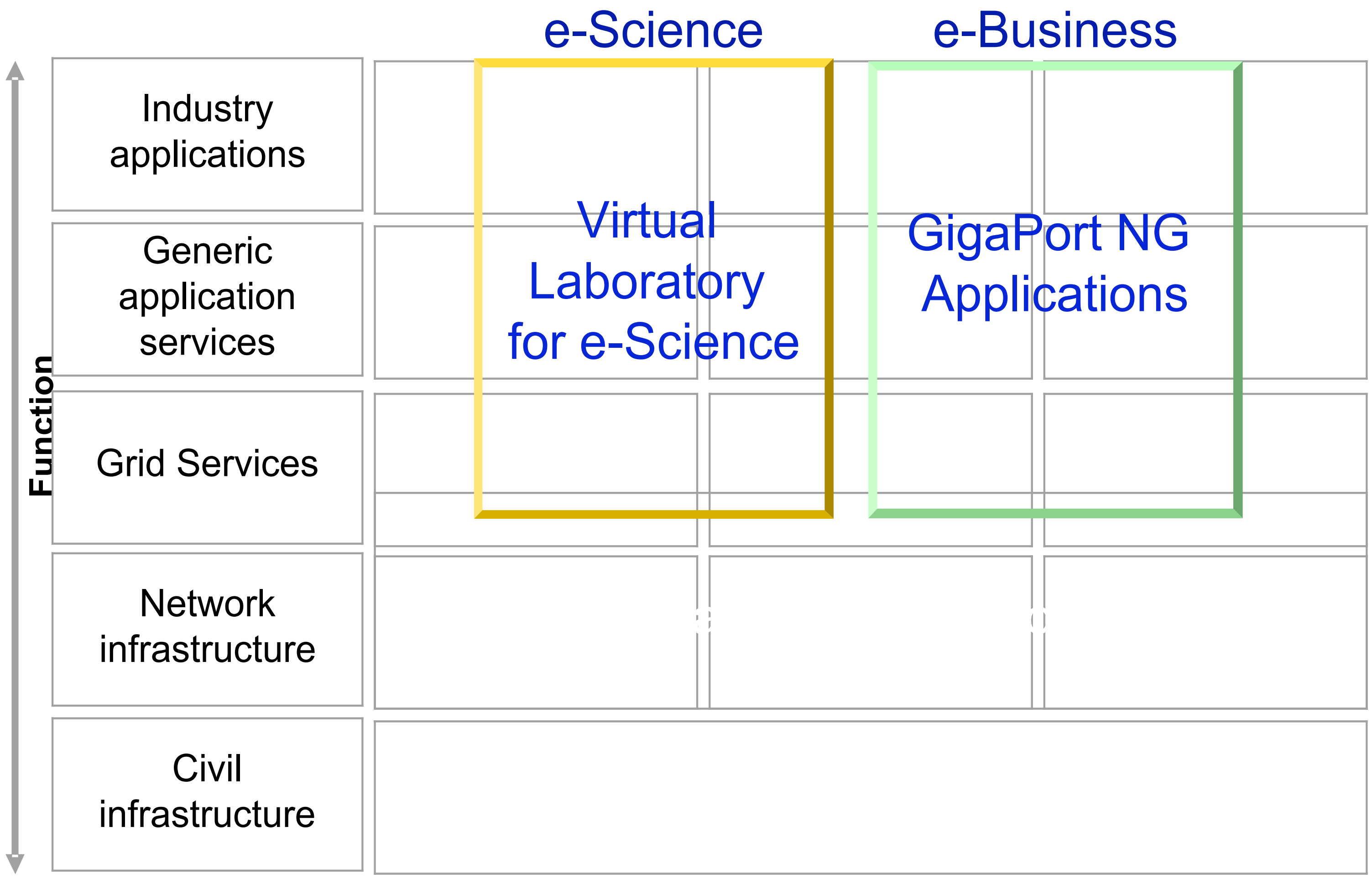
SURFnet

- Provides the Dutch National Research Network
- Not for profit company
- 200 connected organisations, 500.000 users
- Turnover (2002): 35M€
- Infrastructure services:
  - innovation paid for by government
  - cost effective exploitation for higher education and research

Effects of GigaPort

- World leading research infrastructure in NL
  - 15 PoPs connected by thirty 10Gbps lambdas
  - Dual stack IPv4 and IPv6
- Helps transition in Telecoms market
  - GigaMAN
  - Fiber to the dormitories
  - Access pilots/ mobility
- Advanced Optical Exchange: NetherLight
- Playground for new applications

Scope GigaPort Next Generation



e-Science

e-Business

Function

Industry applications

Generic application services

Grid Services

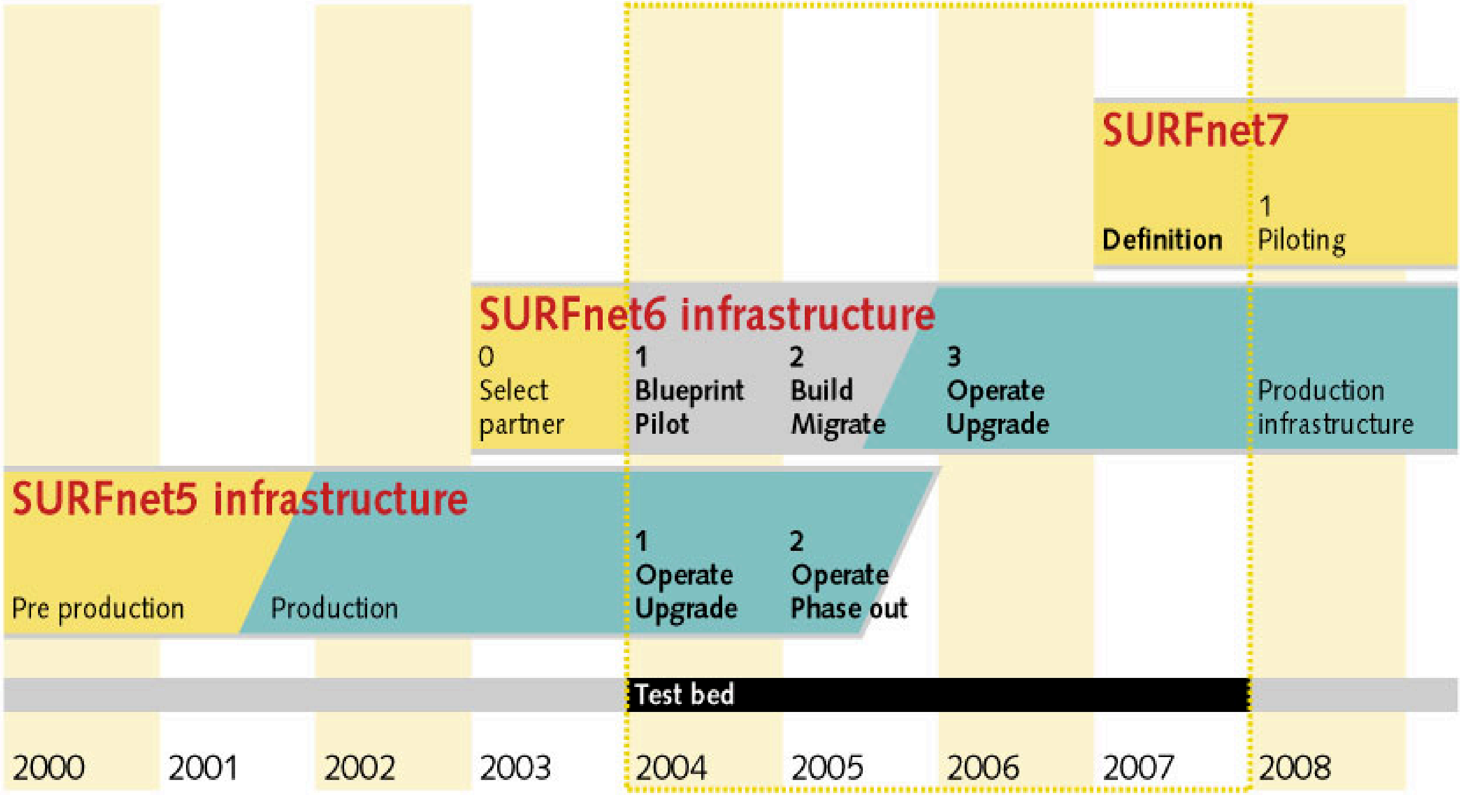
Network infrastructure

Civil infrastructure

Virtual Laboratory for e-Science

GigaPort NG Applications

Timelines SURFnet6



Research topics

- optical transmission and switching
- integrating light paths in network
- routing: new internet features and protocols
- monitoring & network management
- Testing methodology
- network access management (roaming, security, usability, personalized service provisioning)
- service grids

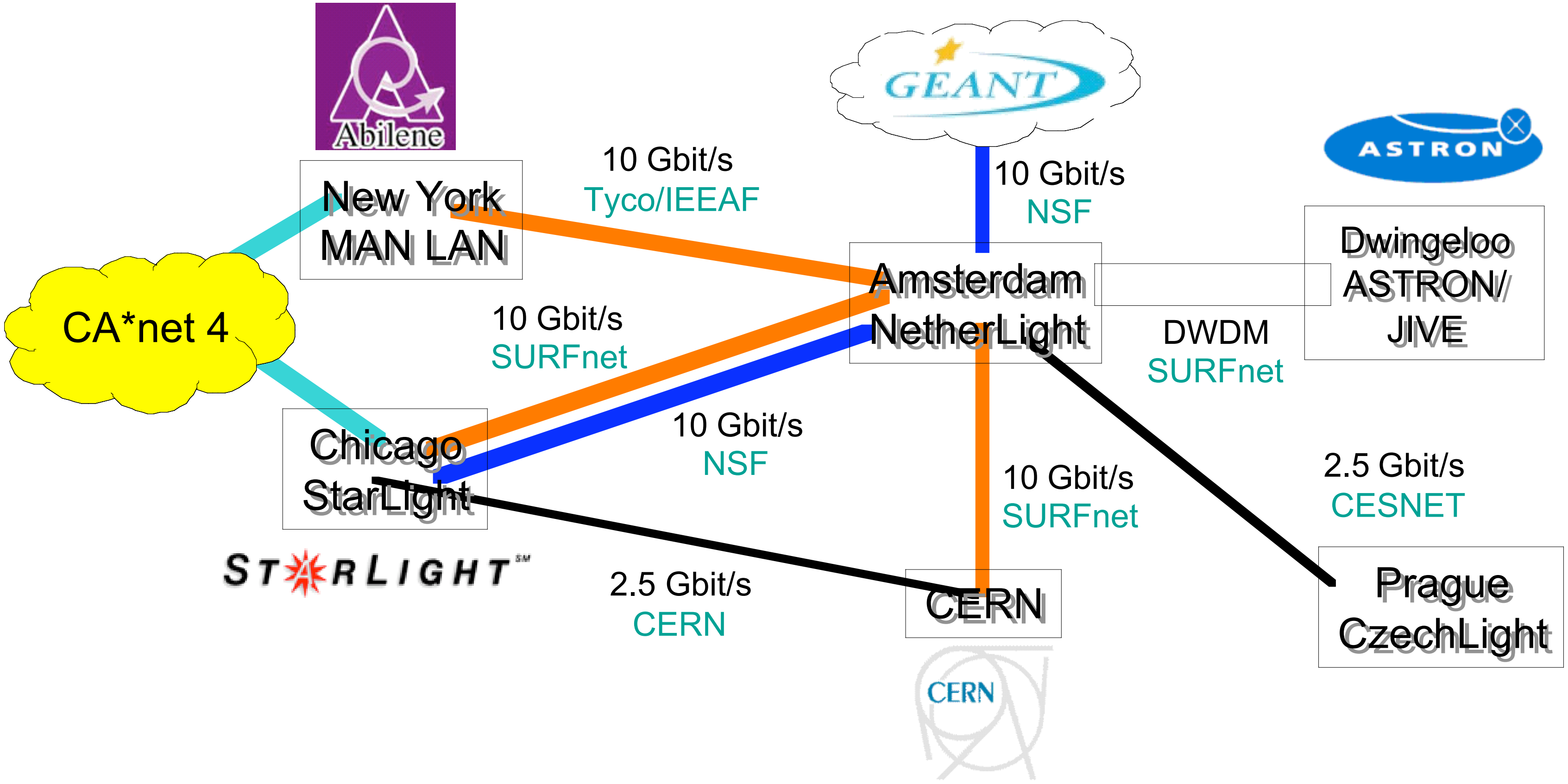
New Applications

- Grid computing
- Data mining
- Data visualization
- Virtual reality
- Remote cooperation
- Radioastronomy

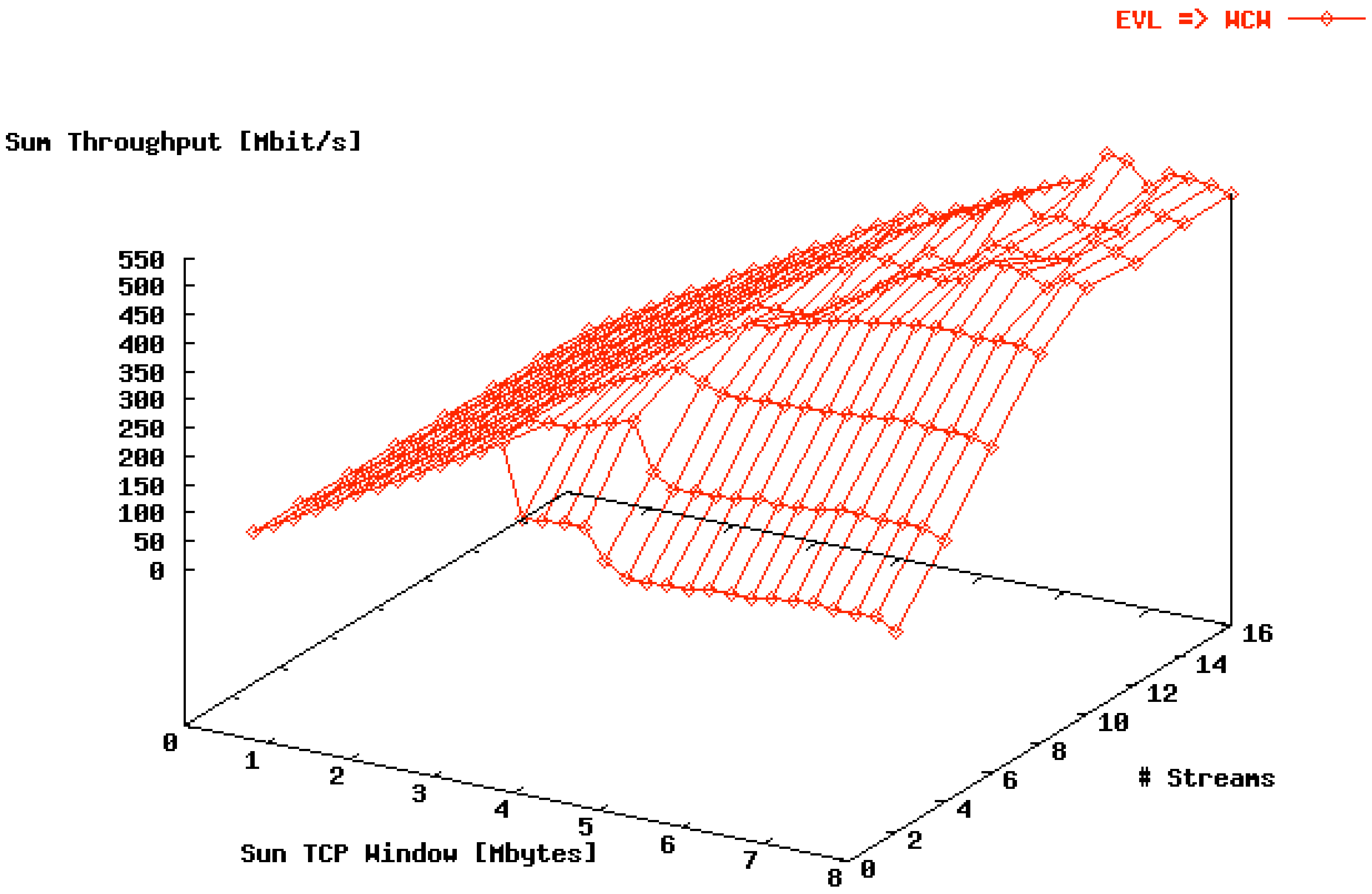
LOFAR

Telecommunication infrastructures become part of scientific instruments

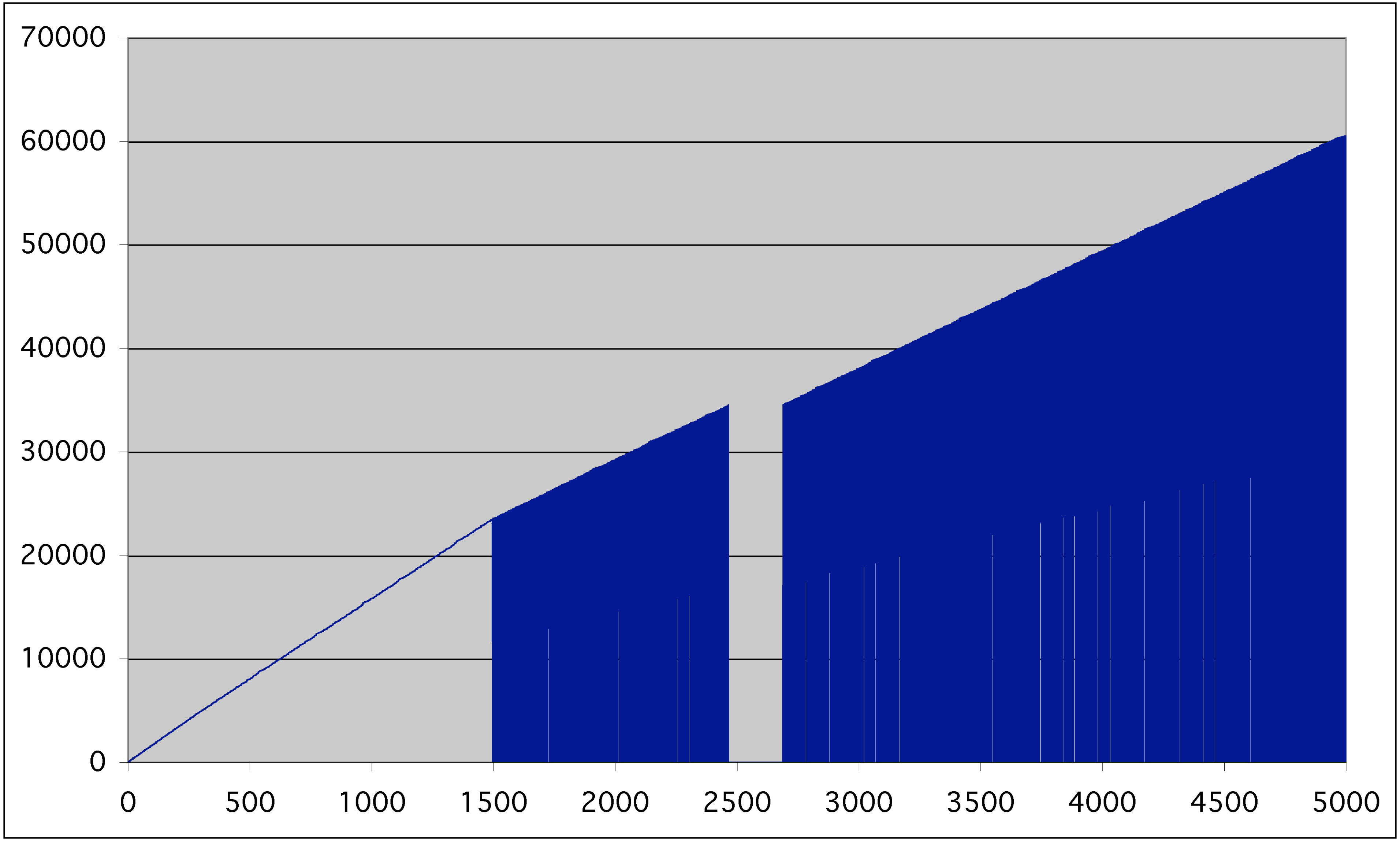
Emerging international lambda grid



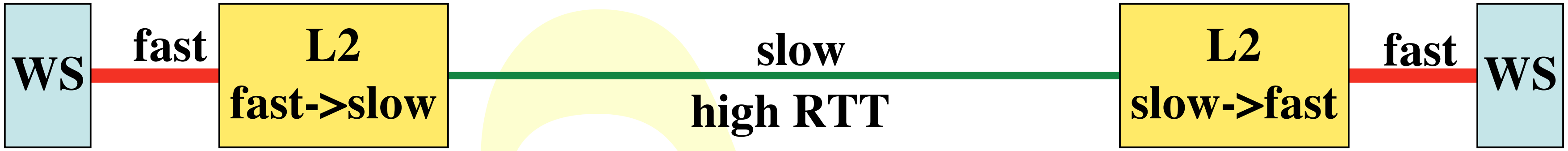
Early Lambda/LightPath usage experiences


EVL => WCW
Sum Throughput [Mbit/s]
550
500
450
400
350
300
250
200
150
100
50
0
0
1
2
3
4
5
6
7
8
Sun TCP Window [Mbytes]
0
2
4
6
8
10
12
14
16
# Streams

5000 1 kByte UDP packets



Layer - 2 requirements from 3/4



WS fast L2 fast->slow

slow high RTT

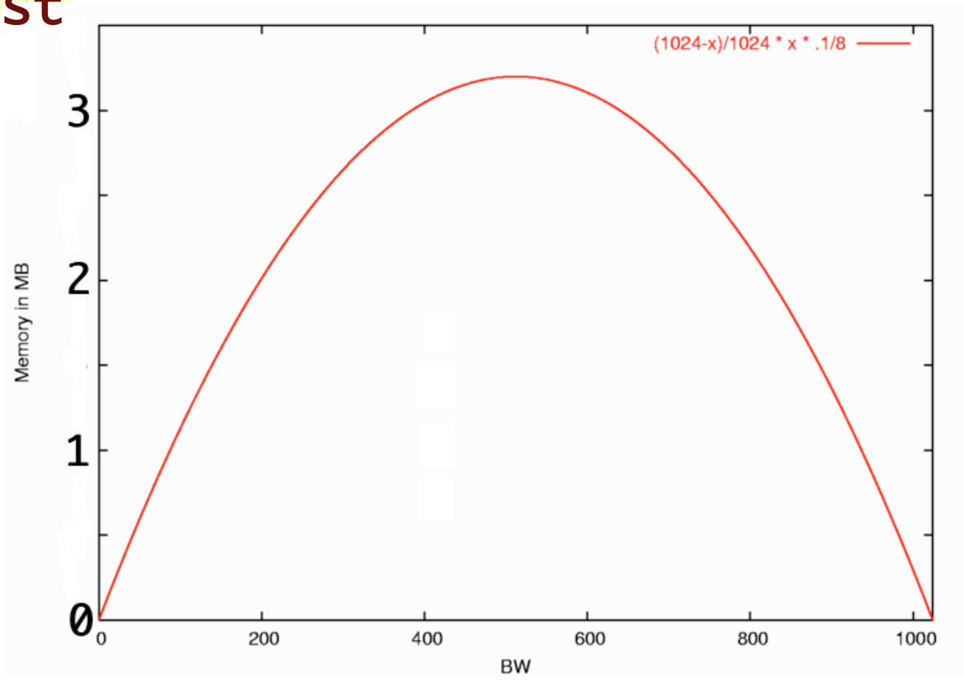
L2 slow->fast fast WS

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

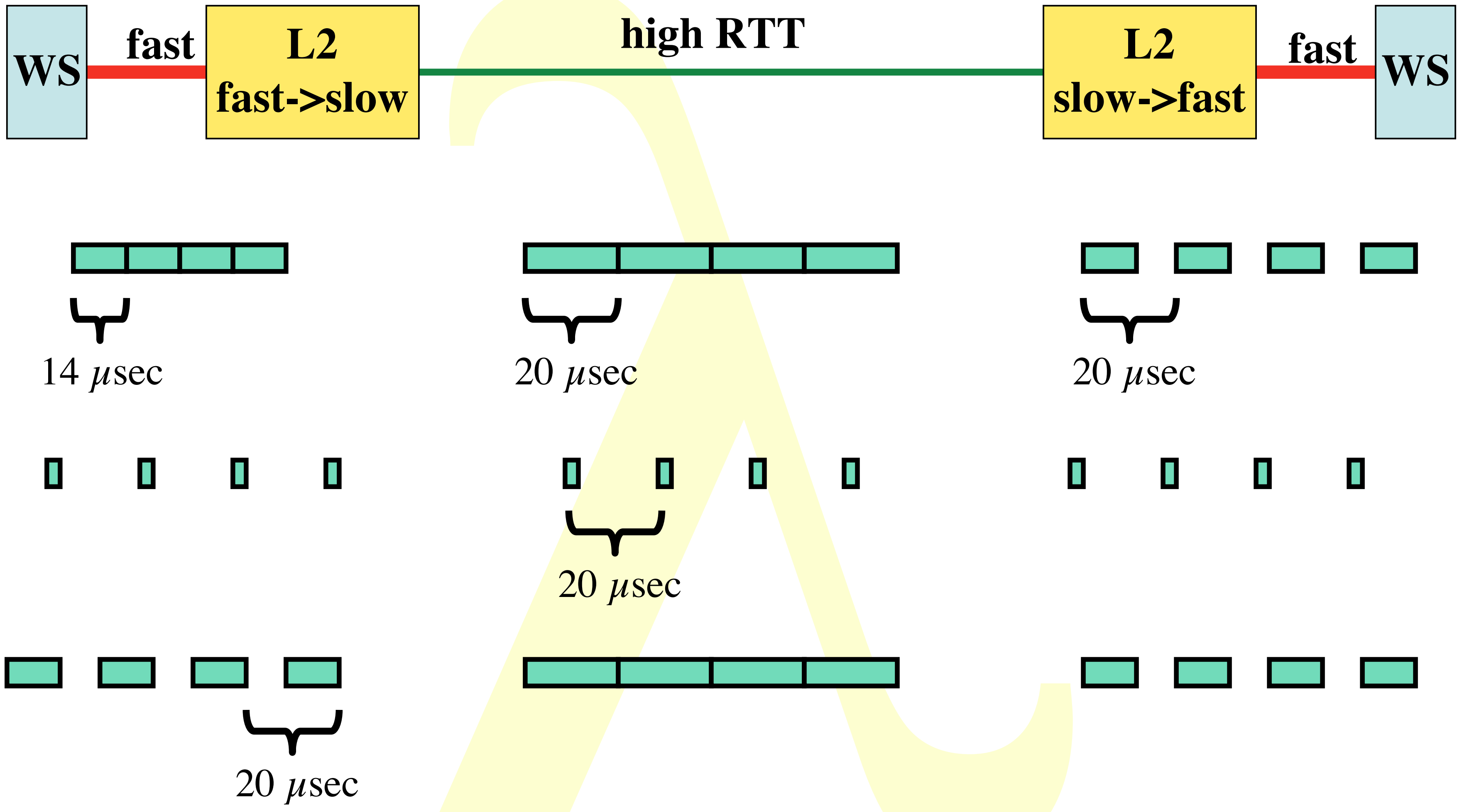
```
Window = BandWidth * RTT    &   BW == slow

                       fast - slow
Memory-at-bottleneck = ----------- * slow * RTT
                           fast
```

So pick from menu:

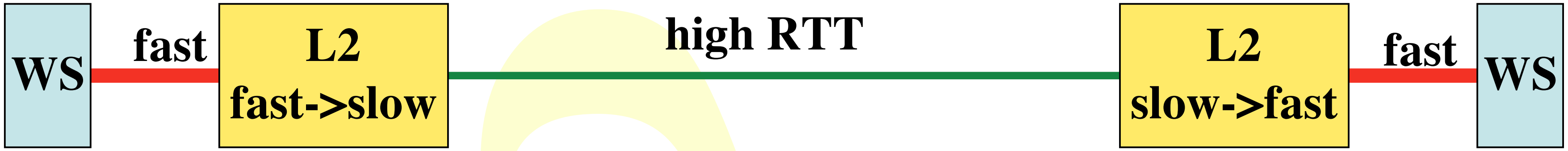
- *Flow control*
- *Traffic Shaping*
- *RED (Random Early Discard)*
- *Self clocking in TCP*
- *Deep memory*



Self-clocking of TCP



Layer - 2 requirements from 3/4



Window = BandWidth * RTT & BW == slow

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

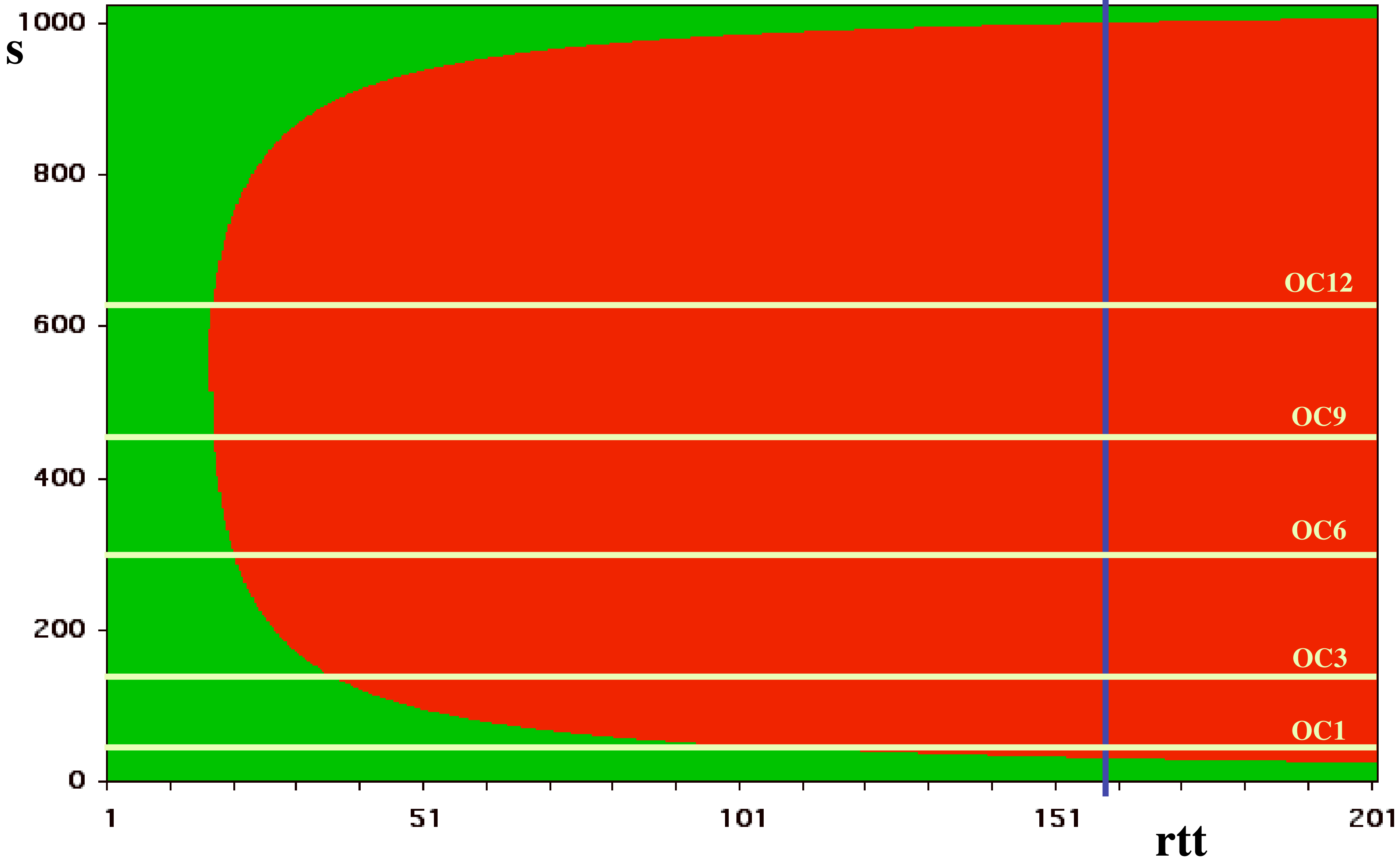
Given M and f, solve for slow ===>

$$0 = s^2 - f * s + \frac{f * M}{RTT}$$

$$s_1, s_2 = \frac{f}{2} \left(1 +/- \text{sqrt}\left(1 - 4 \frac{M}{f * RTT} \right) \right)$$

Forbidden area, solutions for s when f = 1 Gb/s, M = 0.5 Mbyte(20 of 25) AND NOT USING FLOWCONTROL

158 ms = RTT Amsterdam - Vancouver



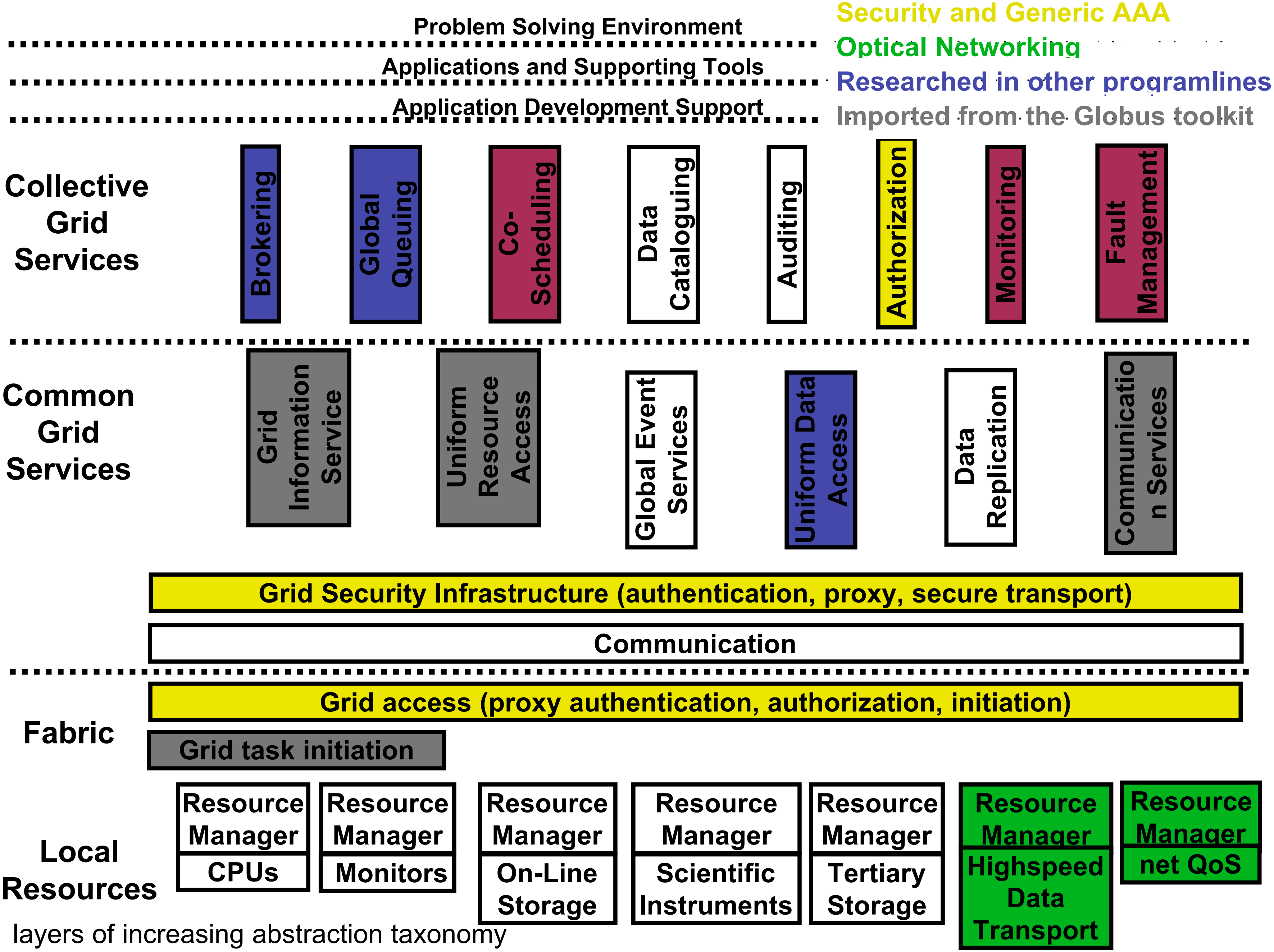
High performance computing and Processor memory co-allocation

Security and Generic AAA

Optical Networking

Researched in other programlines

Imported from the Globus toolkit



Problem Solving Environment

Applications and Supporting Tools

Application Development Support

Collective Grid Services

Brokering

Global Queuing

Co-Scheduling

Data Cataloguing

Auditing

Authorization

Monitoring

Fault Management

Common Grid Services

Grid Information Service

Uniform Resource Access

Global Event Services

Uniform Data Access

Data Replication

Communication Services

Grid Security Infrastructure (authentication, proxy, secure transport)

Communication

Fabric

Grid access (proxy authentication, authorization, initiation)

Grid task initiation

Local Resources

Resource Manager – CPUs

Resource Manager – Monitors

Resource Manager – On-Line Storage

Resource Manager – Scientific Instruments

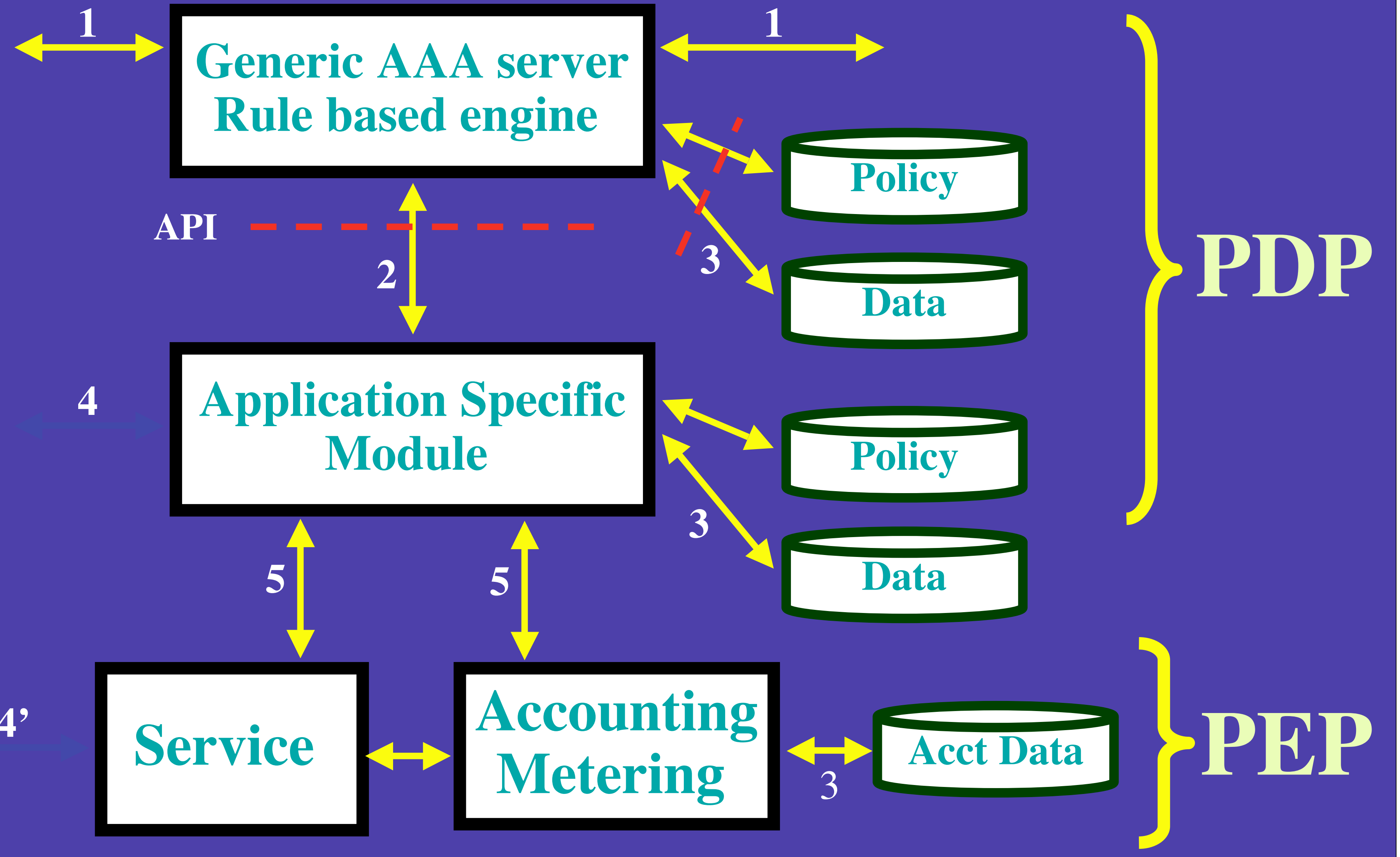
Resource Manager – Tertiary Storage

Resource Manager – Highspeed Data Transport

Resource Manager – net QoS

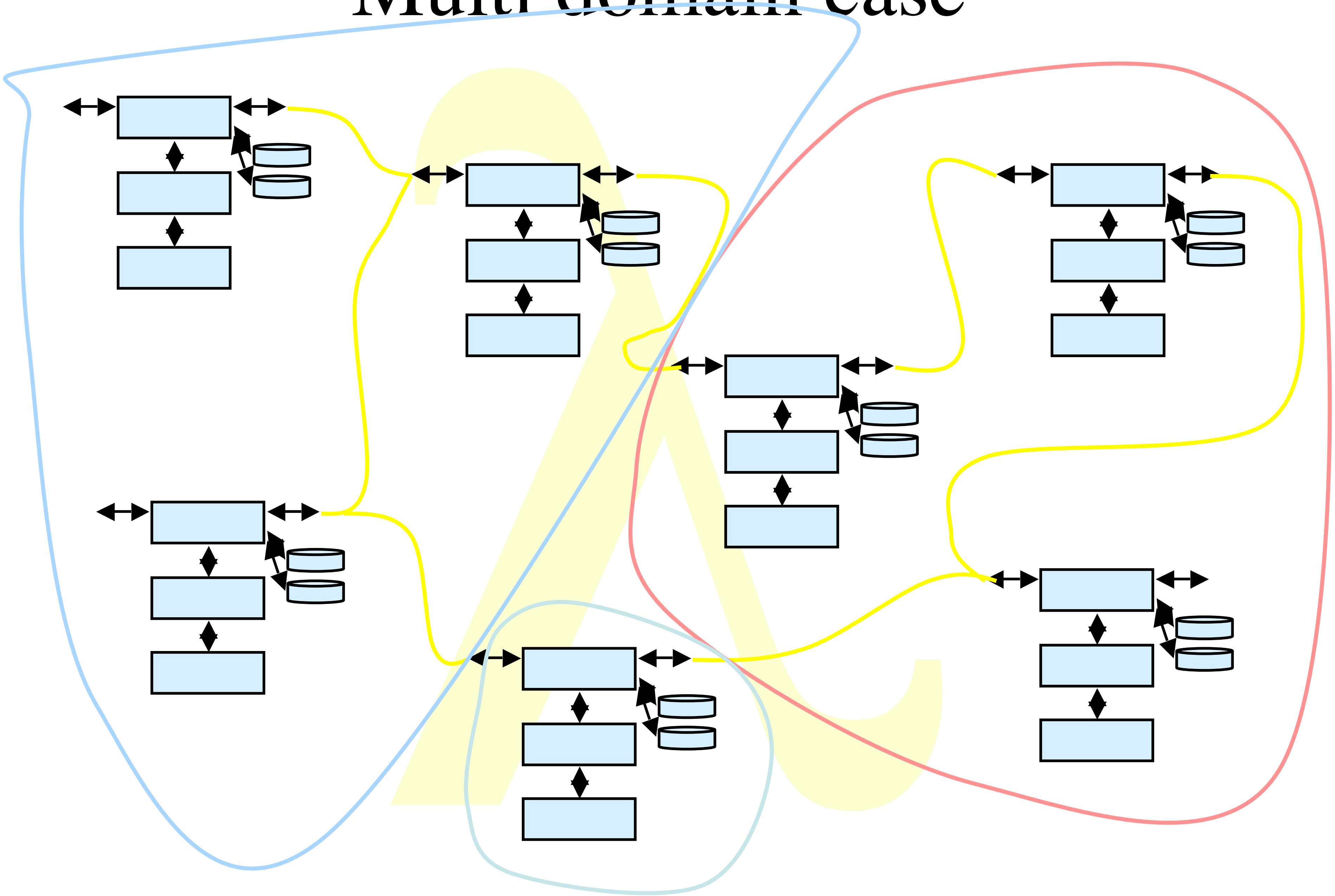
layers of increasing abstraction taxonomy

Starting point



RFC 2903 - 2906 , 3334 , policy draft

Multi domain case



intermezzo

iGrid2002

- www.igrid2002.org
- 25 demonstrations
- 16 countries (at least)
- Level3, Tyco, IEEAF Lambda’s
- CISCO, Hp equipment sponsoring
- Shipping nightmare, debugging literally
- ~30 Gbit/s International connectivity
- Huge networking collaboration
- Smelly NOC in the iGrid preparation weekend

Lessons learned

- Most applications could not cope with the network!!!
- No bottleneck whatsoever in the network
- Many got about 50 - 100 mbit/s singlestream tcp
- On Sunday evening my laptop had the highest single stream to Chicago (~ 340 Mbit/s)
- NIC’s, Linux implementation and timing problem
- Gridftp severely hit
- ~ 22 papers to be published

The END

Thanks to

SURFnet: Kees Neggers

UIC&iCAIR: Tom DeFanti, Joel Mambretti

CANARIE: Bill St. Arnaud

This work is supported by:

SURFnet

EU-IST project DATATAG