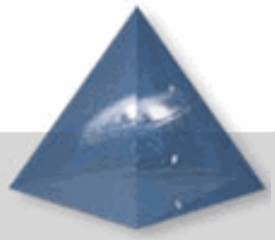
Optical Networking / Experiences @ iGrid2002

www.science.uva.nl/~delaat

Cees de Laat



Faculty of Science

Optical Networking / Experiences @ iGrid2002

www.science.uva.nl/~delaat

Cees de Laat

EU

SURFnet

University of Amsterdam

SARA

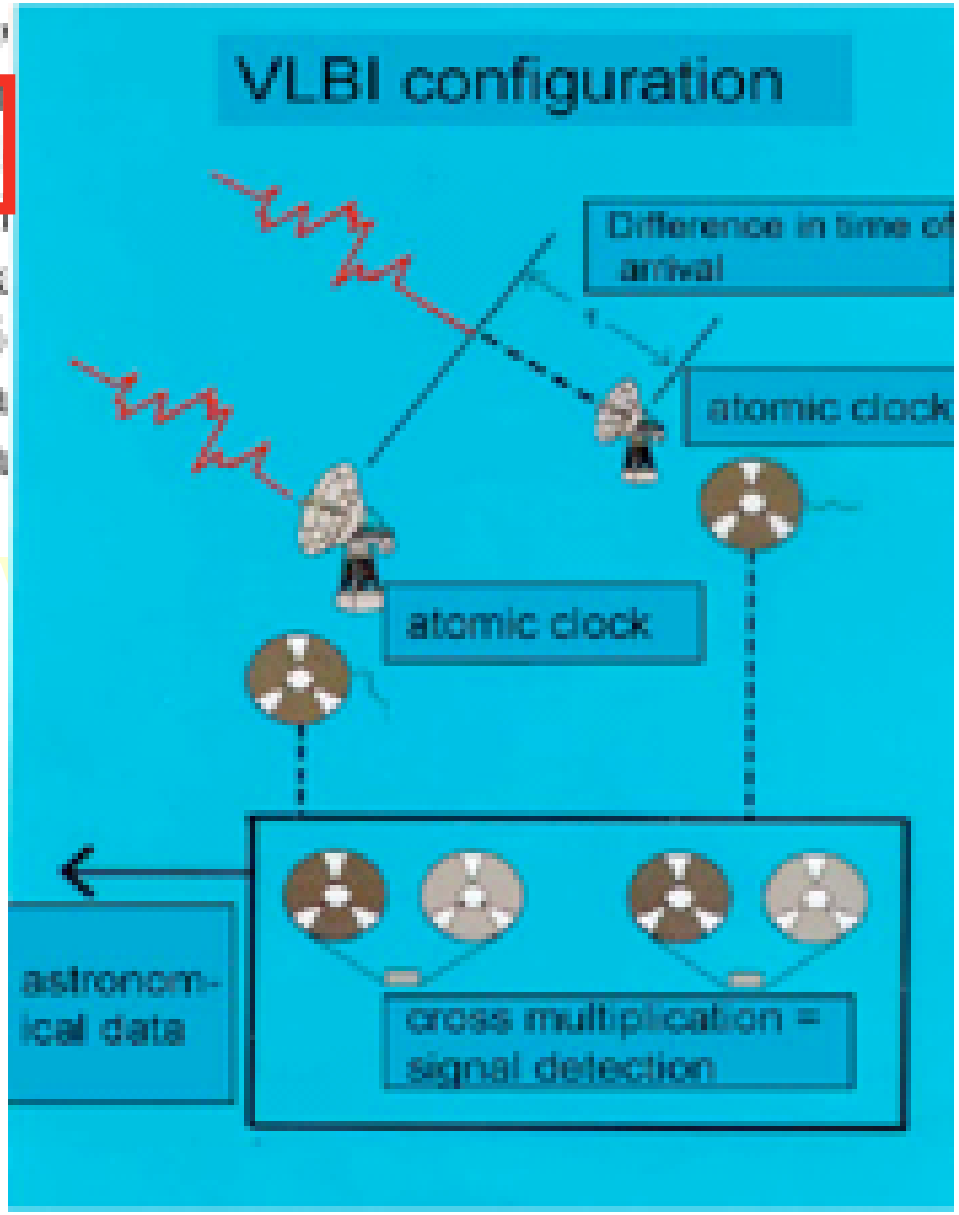
NIKHEF

optiputer

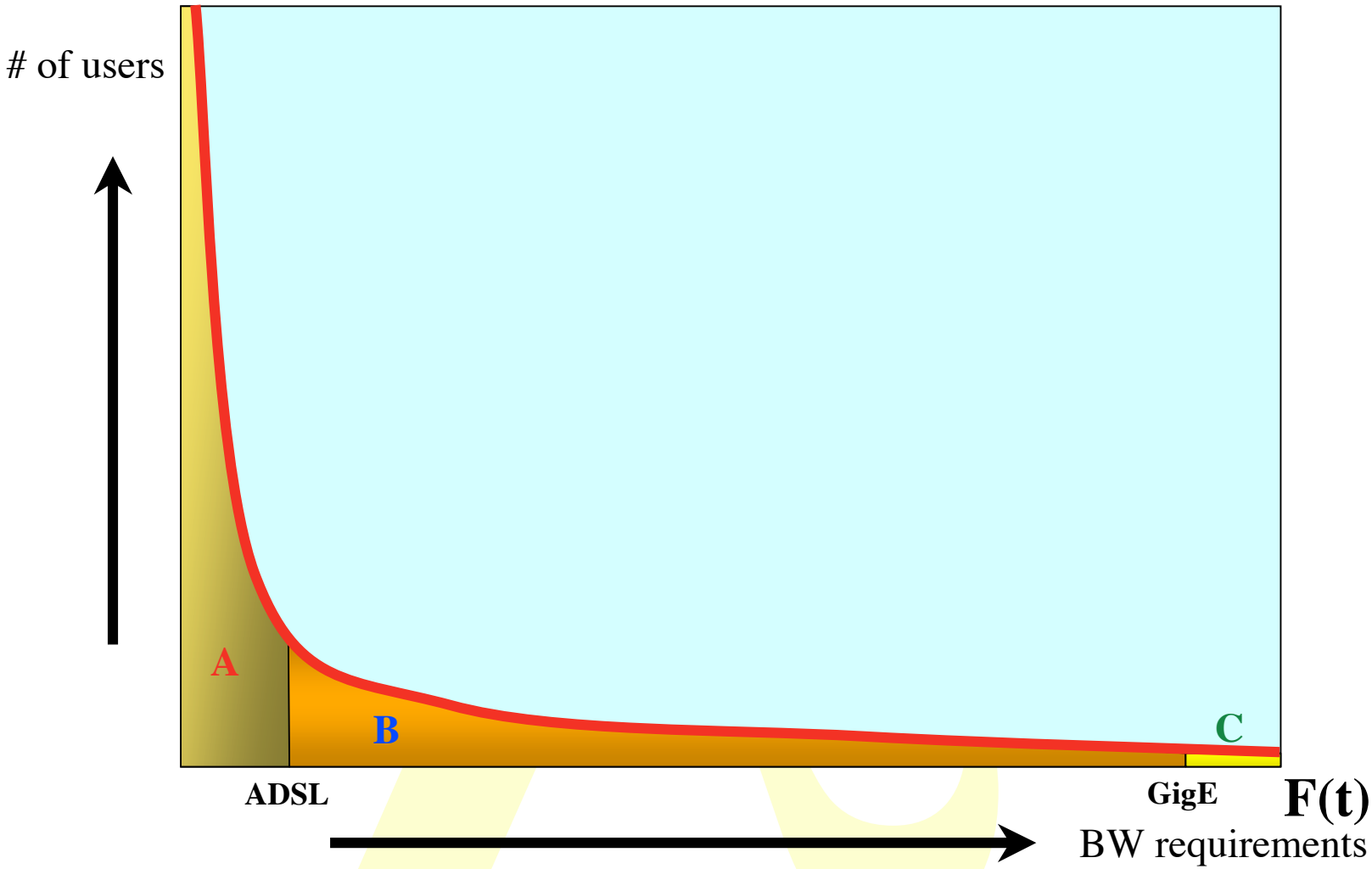
VLBI

ger term VLBI is easily capable of generating many Gb of data per
The sensitivity of the VLBI array scales v
(data-rate) and there is a strong push to
. Rates of 8Gb/s or more are entirely feasibl
ider 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



Know the user

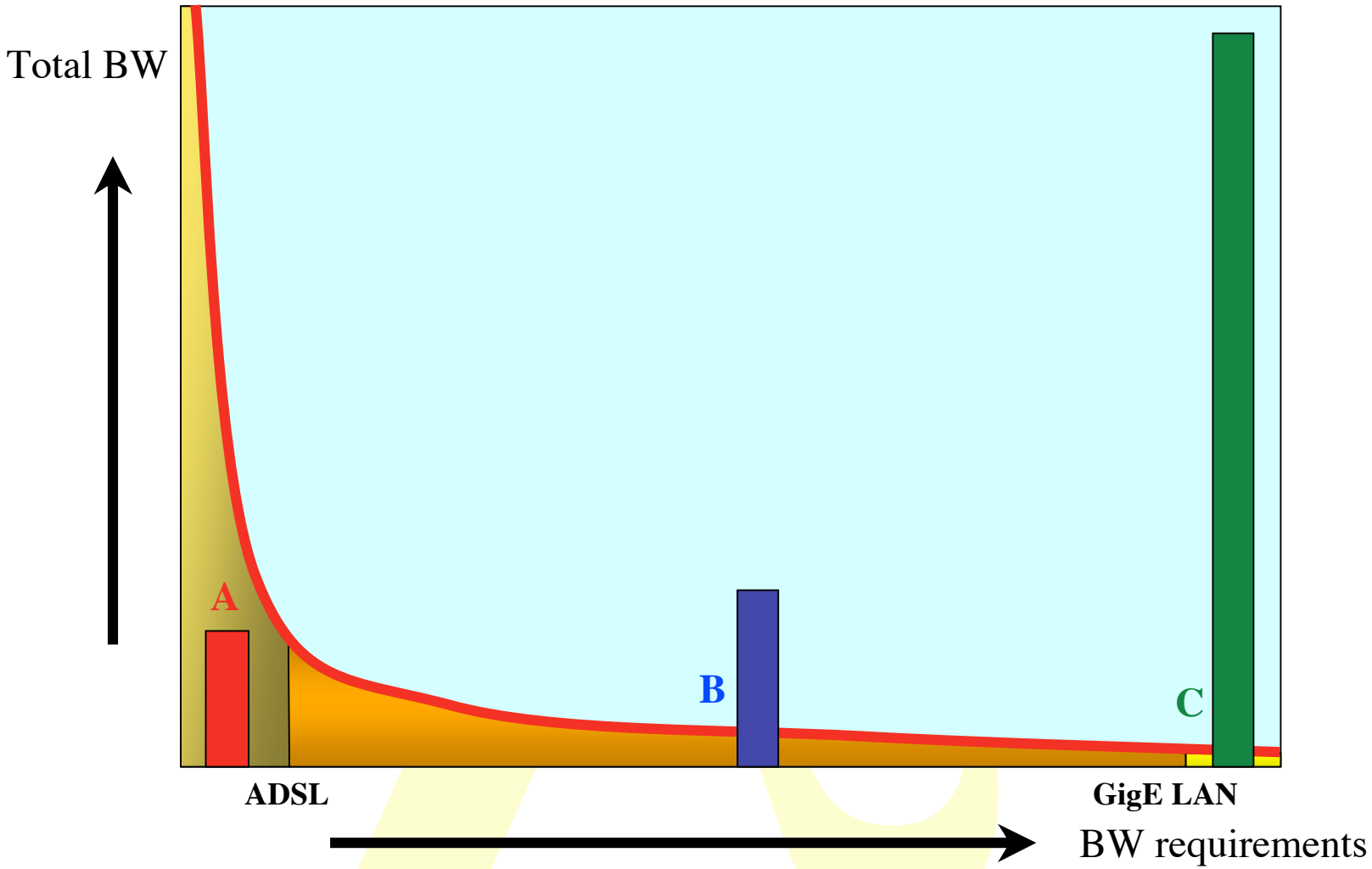


A -> Lightweight users, browsing, mailing, home use

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

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

What the user



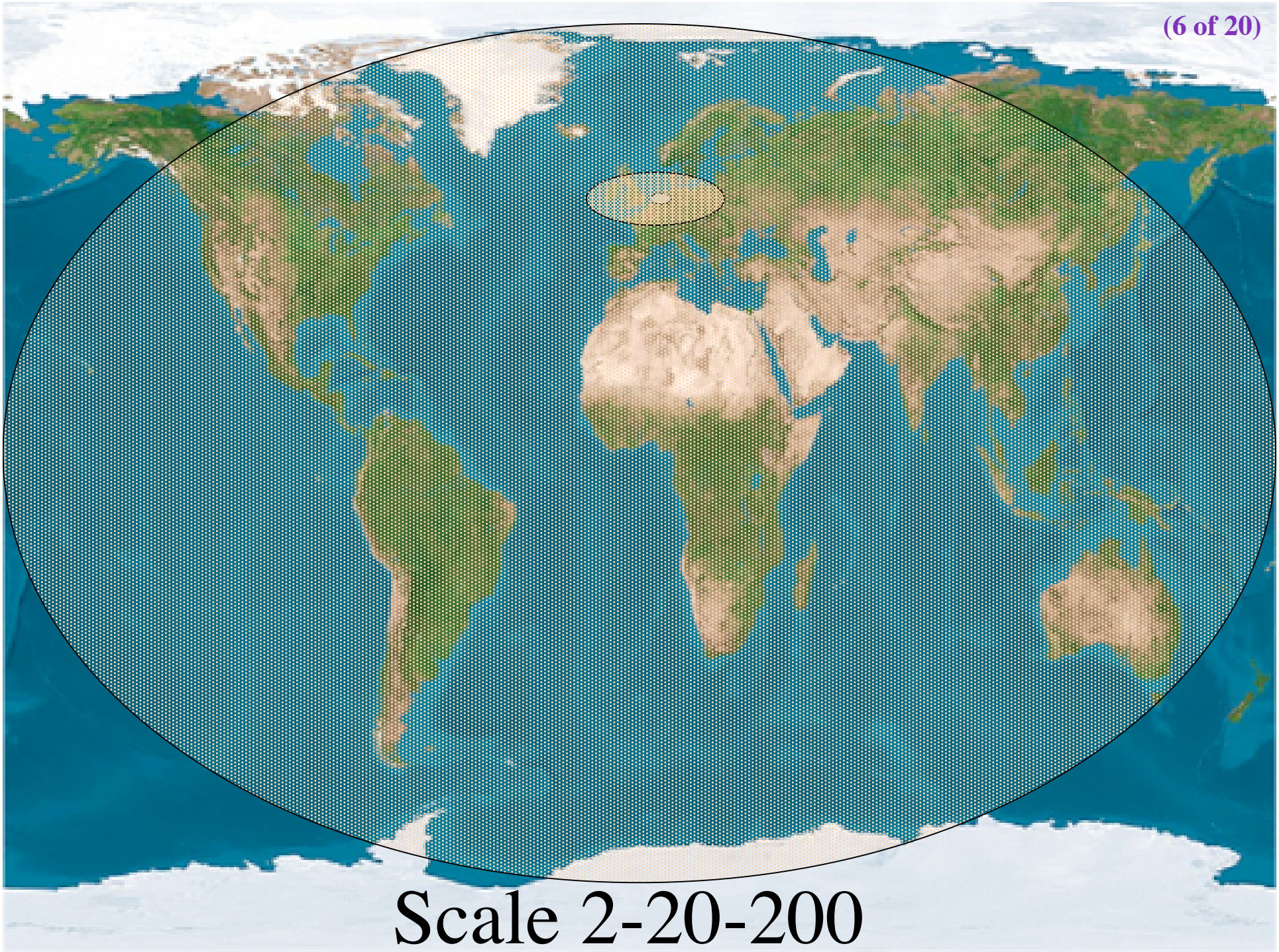
A -> Need full Internet routing, one to many

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



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

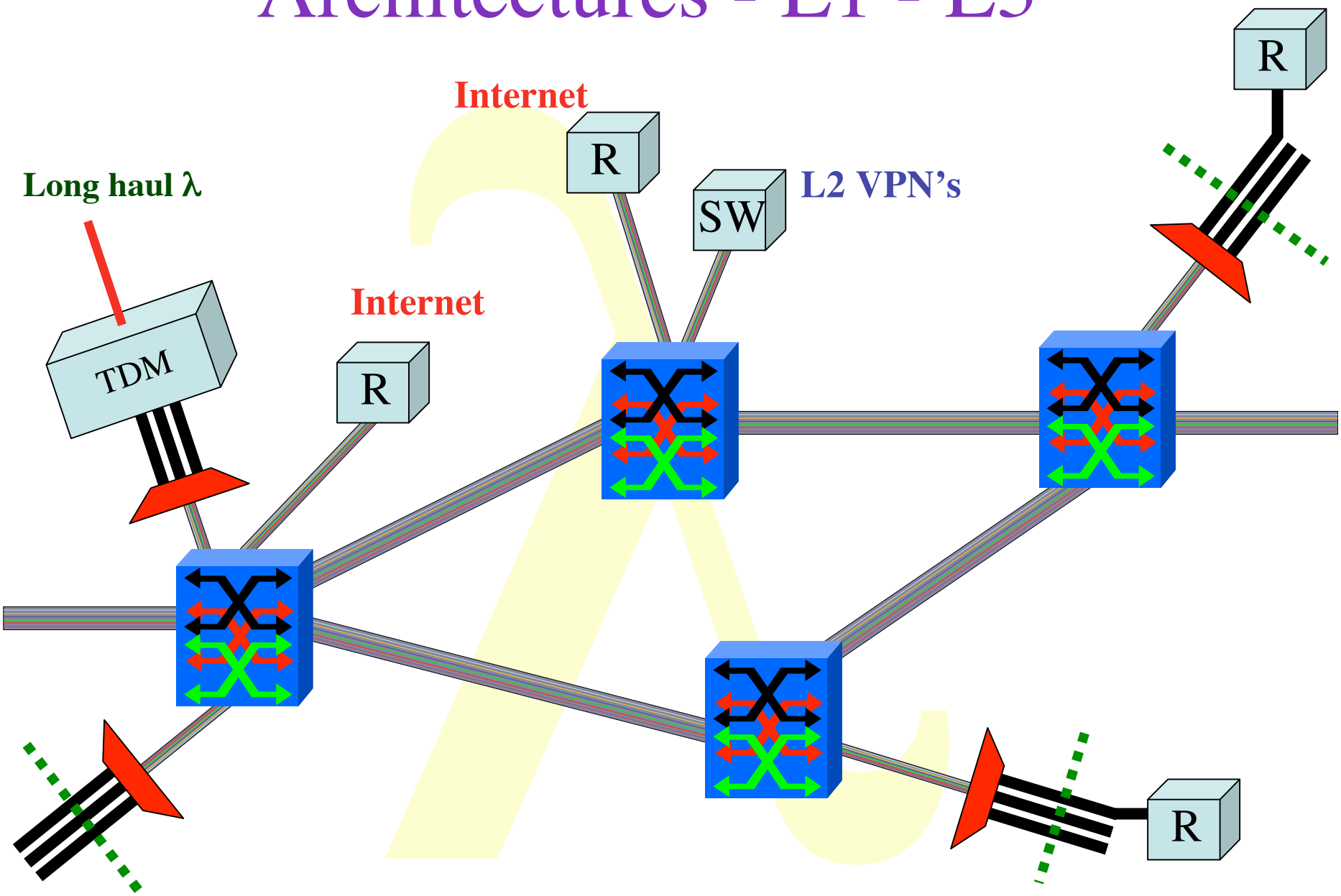
	2 Metro	20 National/ regional	200 World
A	Switching/ routing	Routing	ROUTER$
B	VPN's, (G)MPLS	VPN's Routing	Routing
C $\#\lambda \approx \frac{200 * e^{(t-2002)}}{rtt}$	dark fiber Optical switching	Lambda switching	Sub- lambdas, ethernet- sdh

Current technology + (re)definition

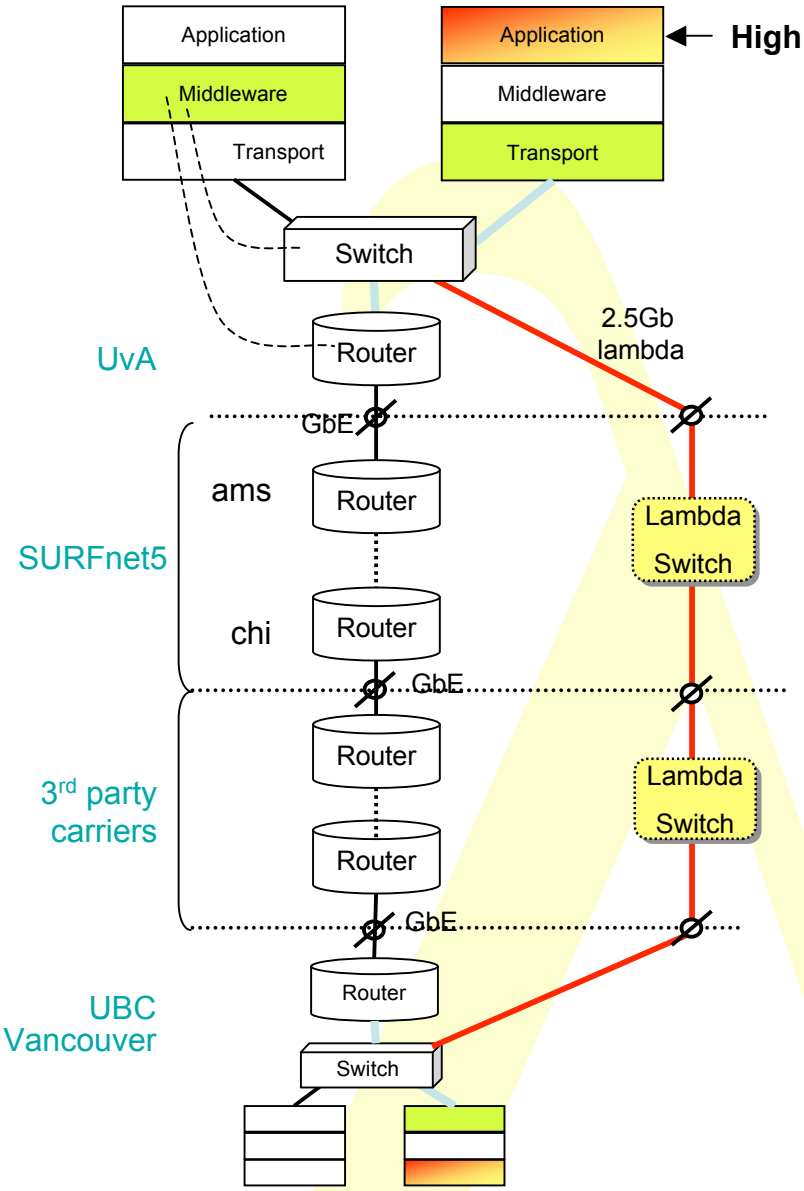
- Current (to me) available technology consists of SONET/SDH switches
- Changing very soon!, optical switch on the way!
- DWDM+switching coming up
- Starlight uses for the time being VLAN's on Ethernet switches to connect [exactly two] ports (but also routing)
- So redefine a λ as:

"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

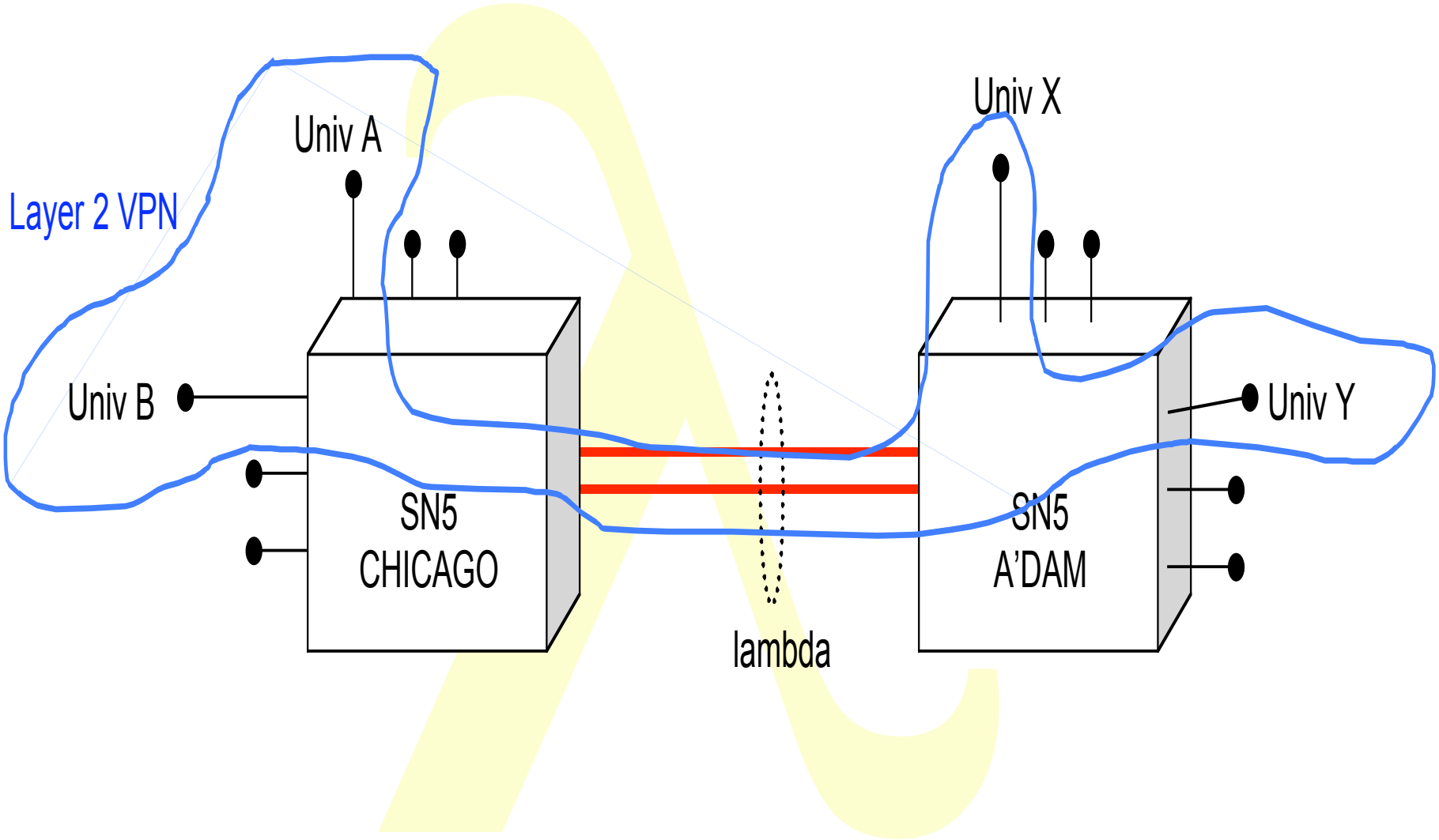


Bring plumbing to the users, not just create sinks in the middle of nowhere

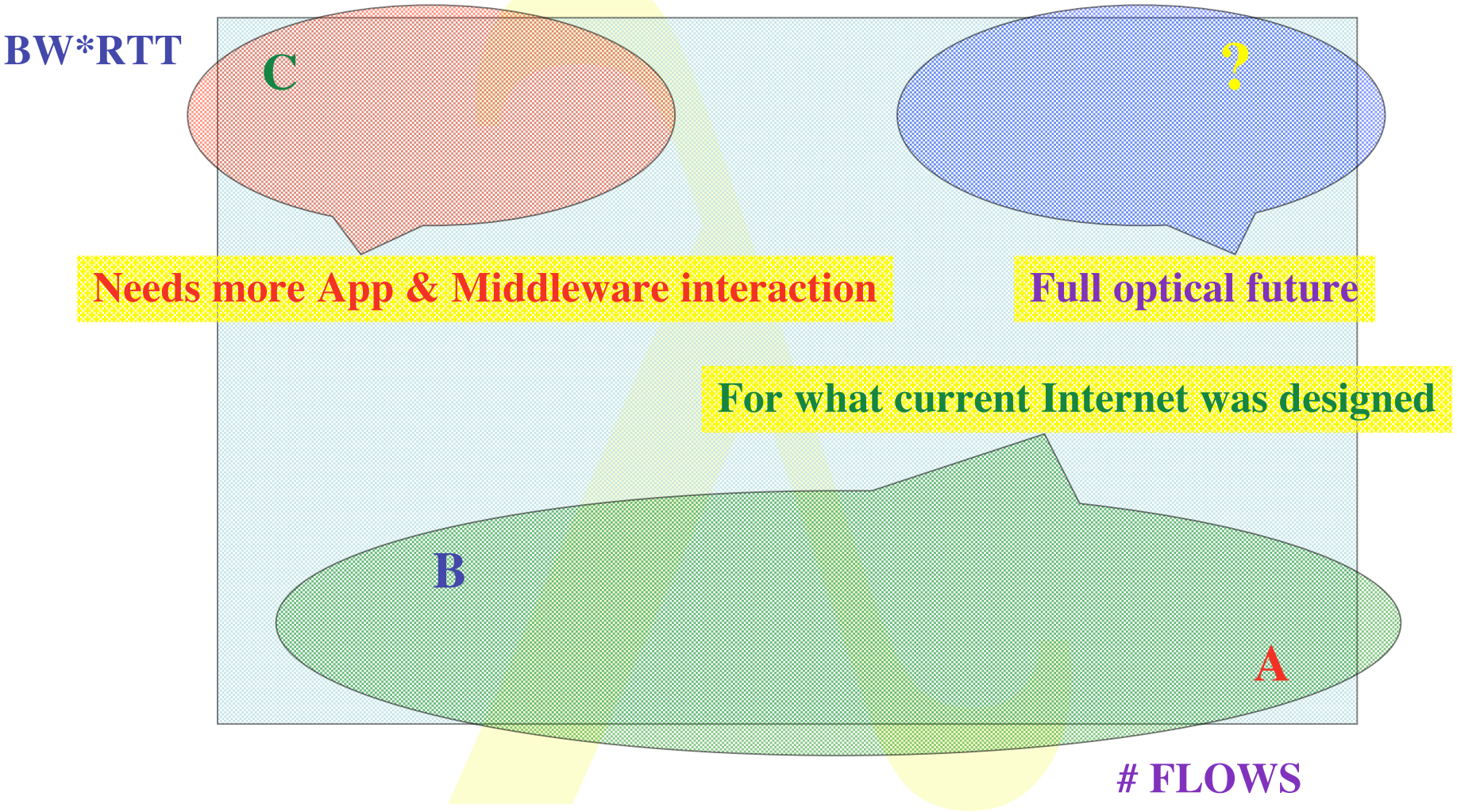


- lambda for high bandwidth applications
  - Bypass of production network
  - Middleware may request (optical) pipe
- RATIONALE:
  - Lower the cost of transport per packet

Distributed L2

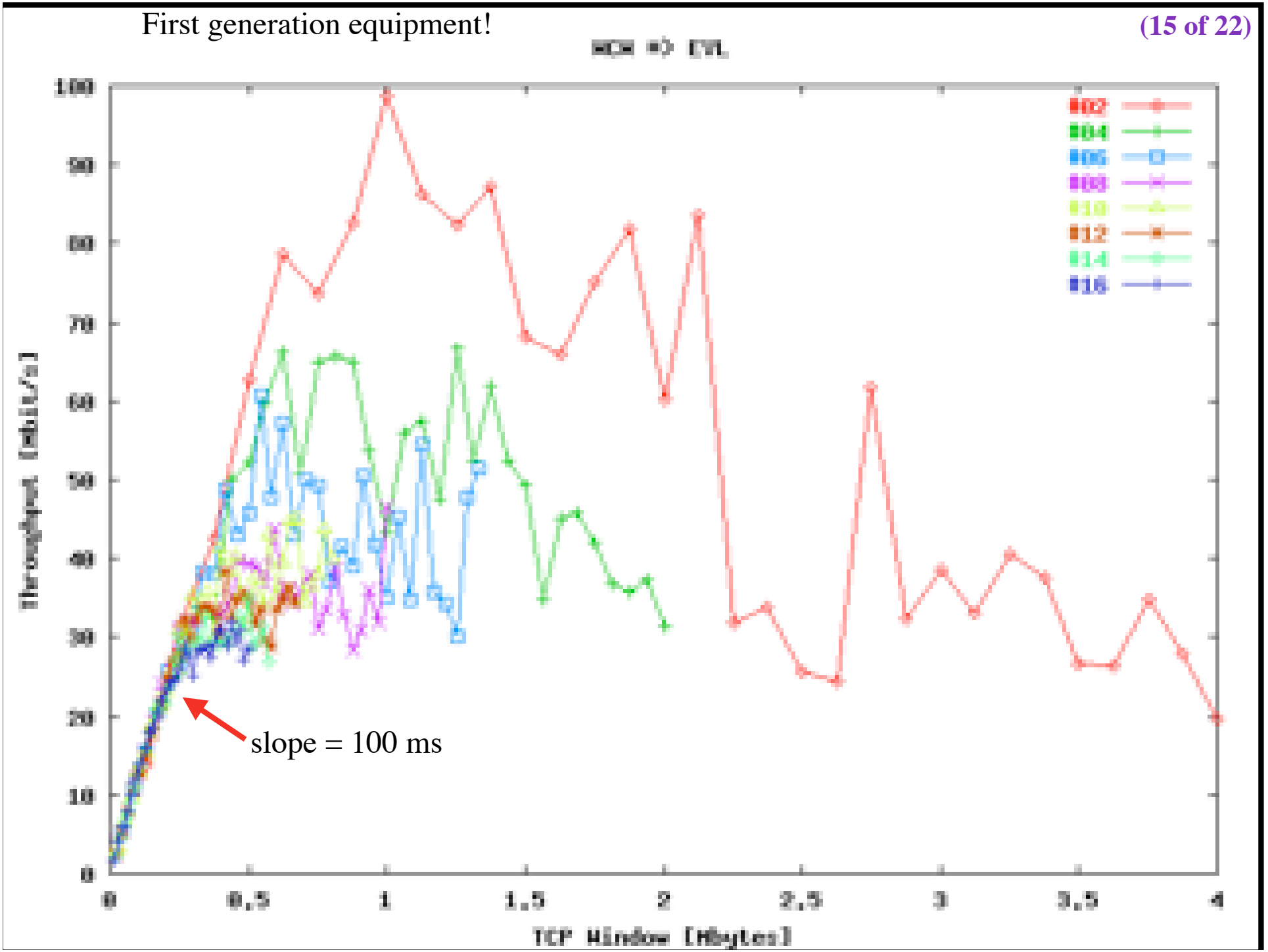


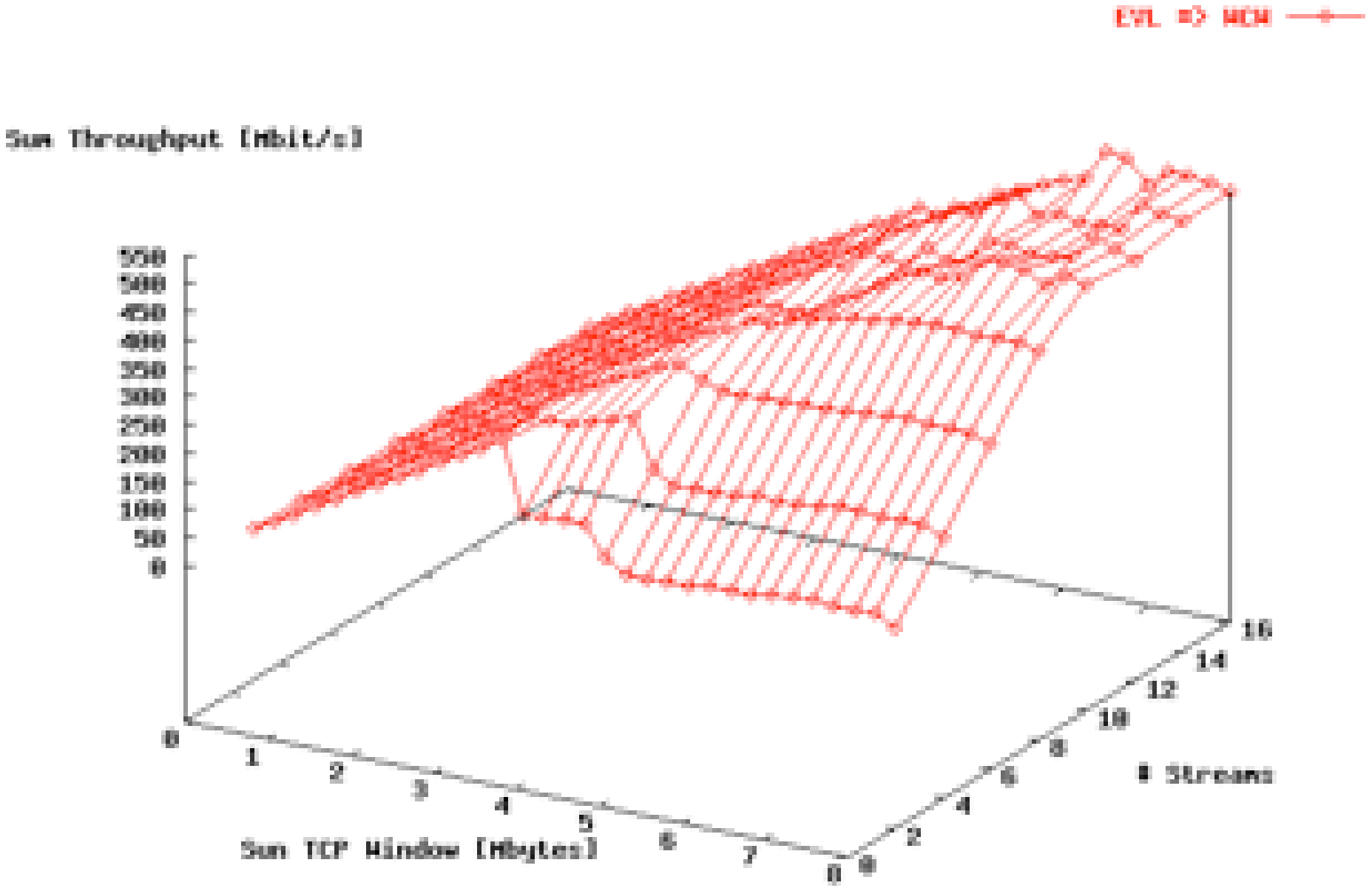
Transport in the corners



Early Lambda/LightPath usage experiences

First generation equipment!




EVL => MEM
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
Sum TCP Window [Mbytes]
0
2
4
6
8
10
12
14
16
Streams

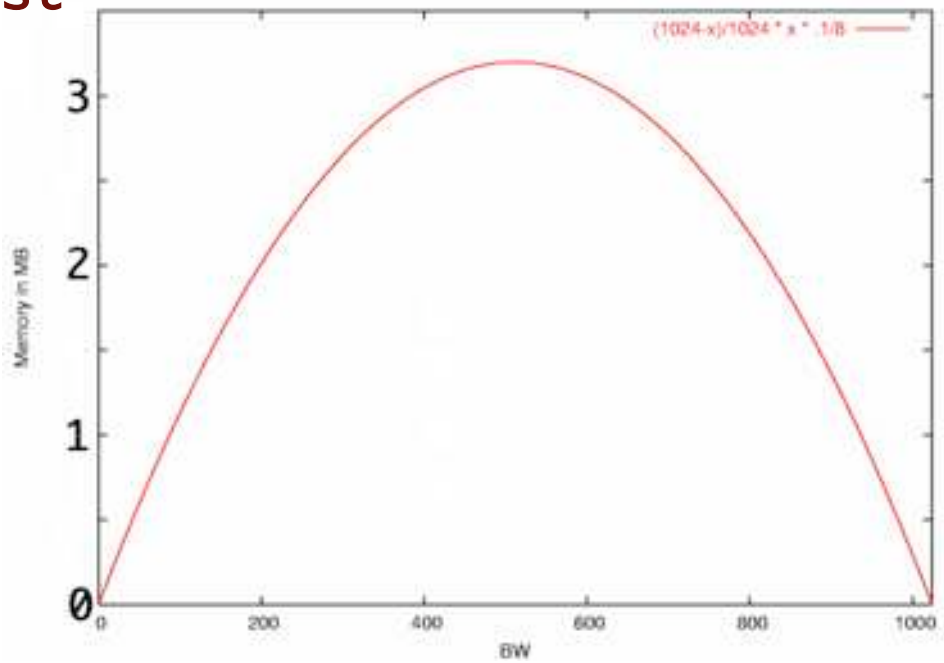
Layer - 2 requirements from 3/4

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

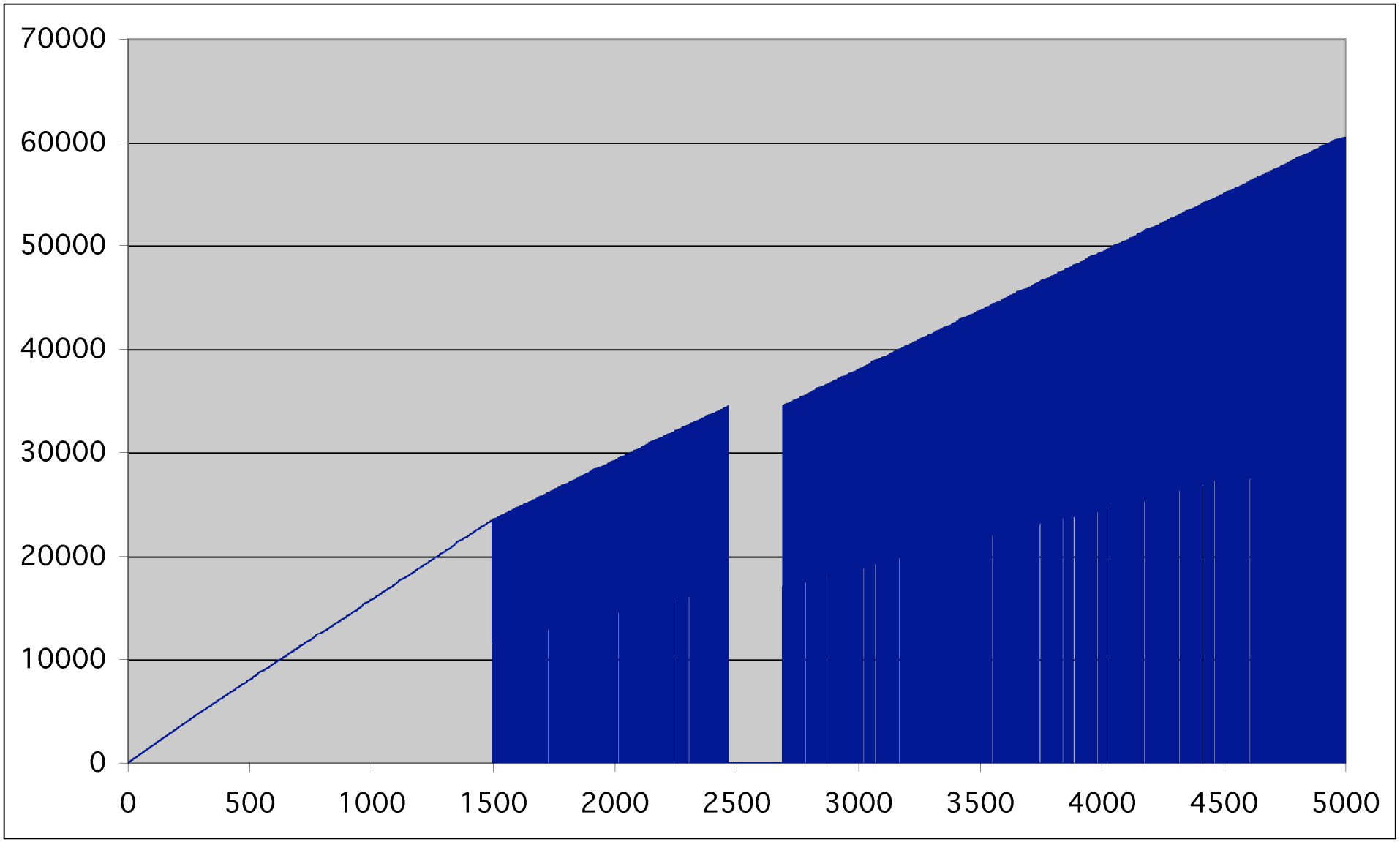
```
Window = BandWidth * RTT     &    BW == slow

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

So pick from menu:

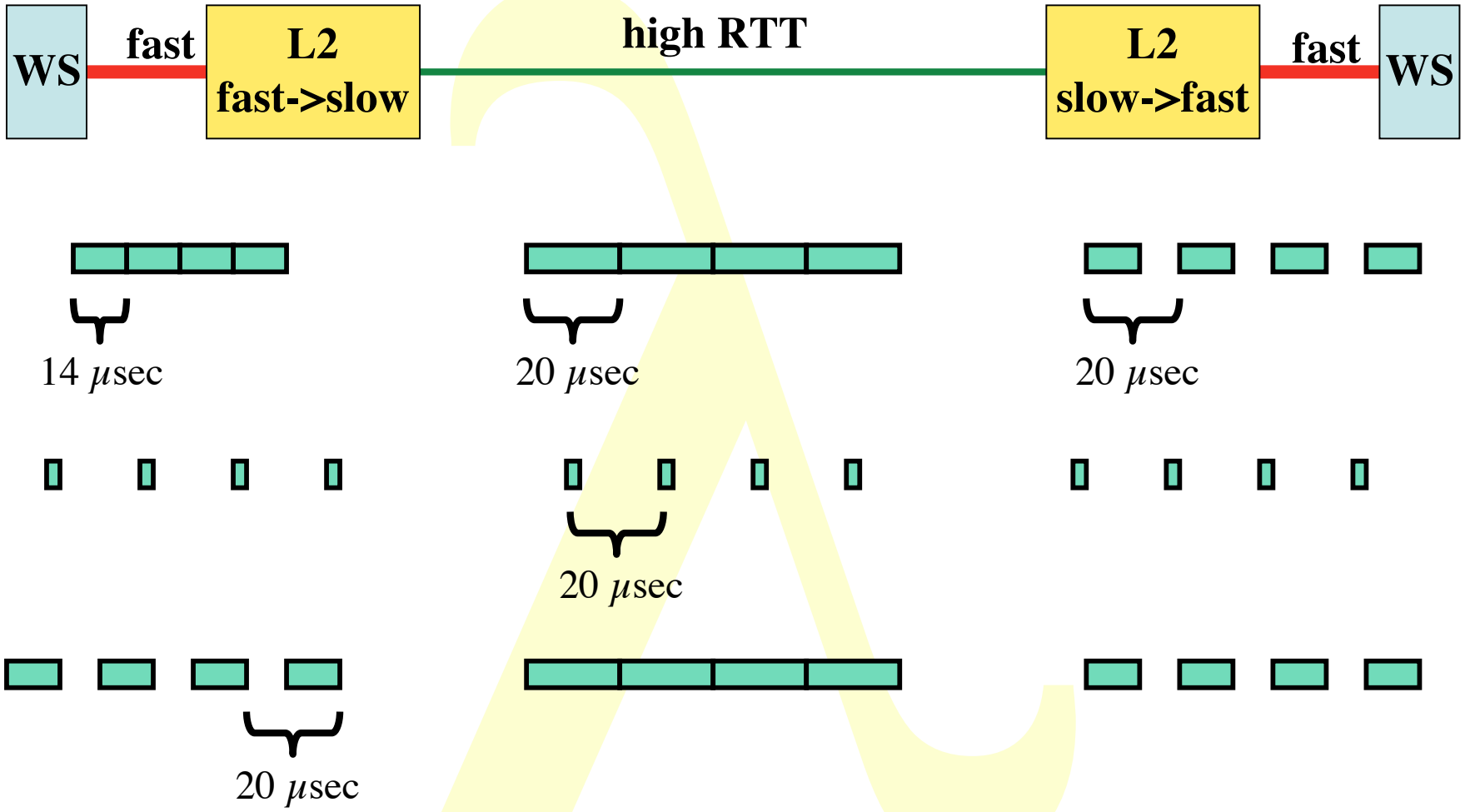
- *Flow control*
- *Traffic Shaping*
- *RED (Random Early Discard)*
- *Self clocking in TCP*
- *Deep memory*



5000 1 kByte UDP packets



Self-clocking of TCP



Layer - 2 requirements from 3/4

Window = BandWidth * RTT & BW == slow

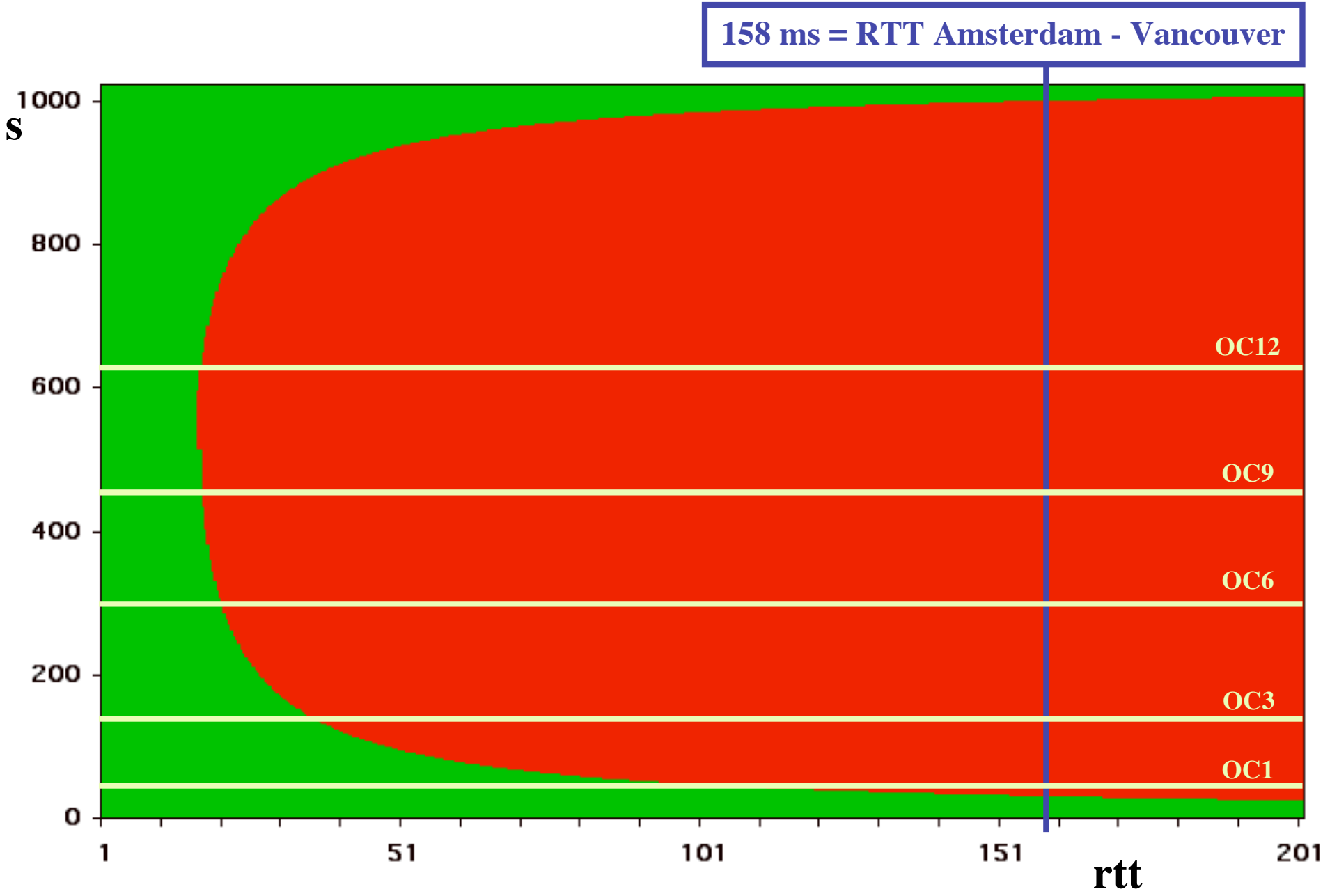
$$\text{Memory-at-bottleneck} = \frac{\text{fast} - \text{slow}}{\text{fast}} * \text{slow} * \text{RTT}$$

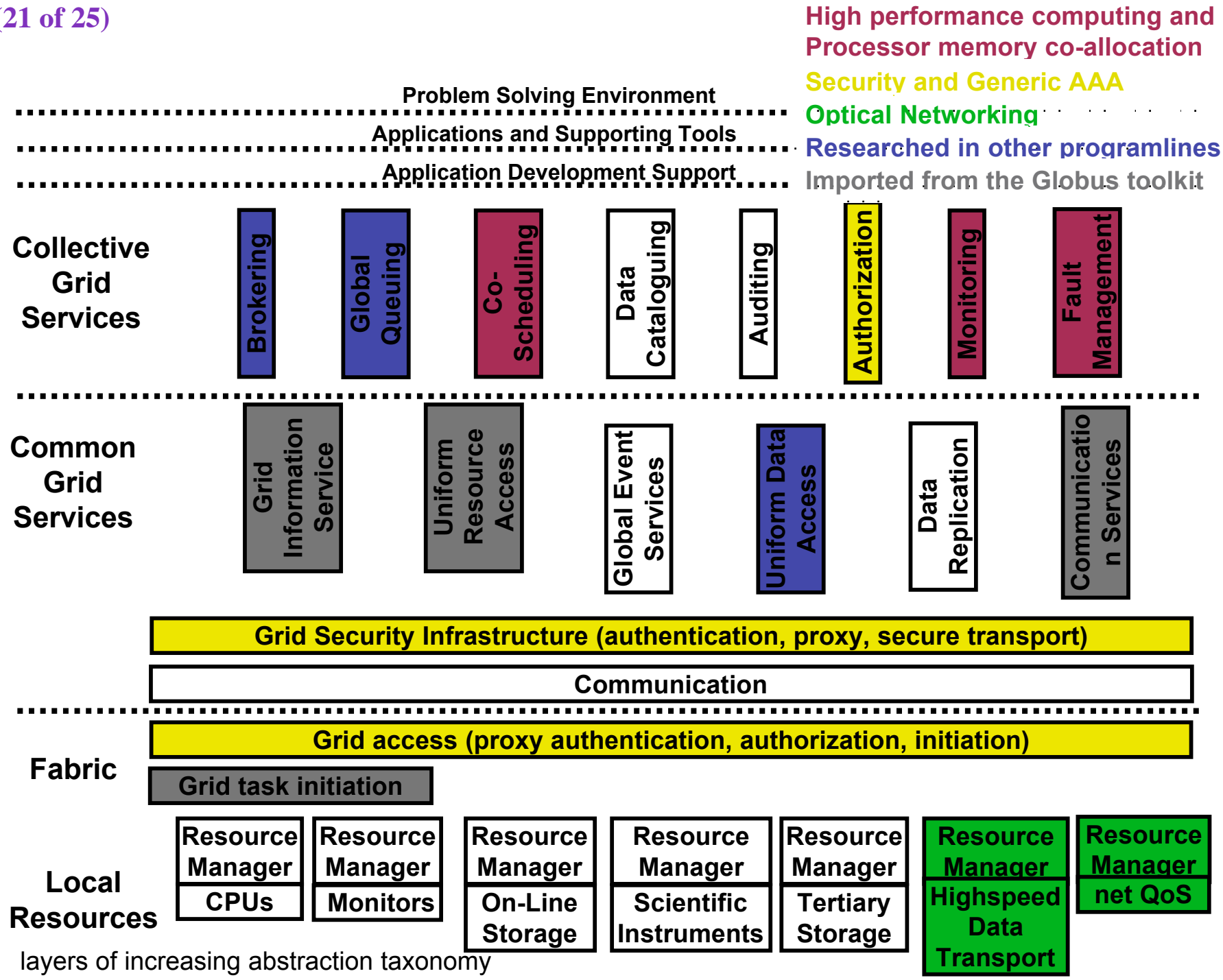
Given M and f, solve for slow ===>

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

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

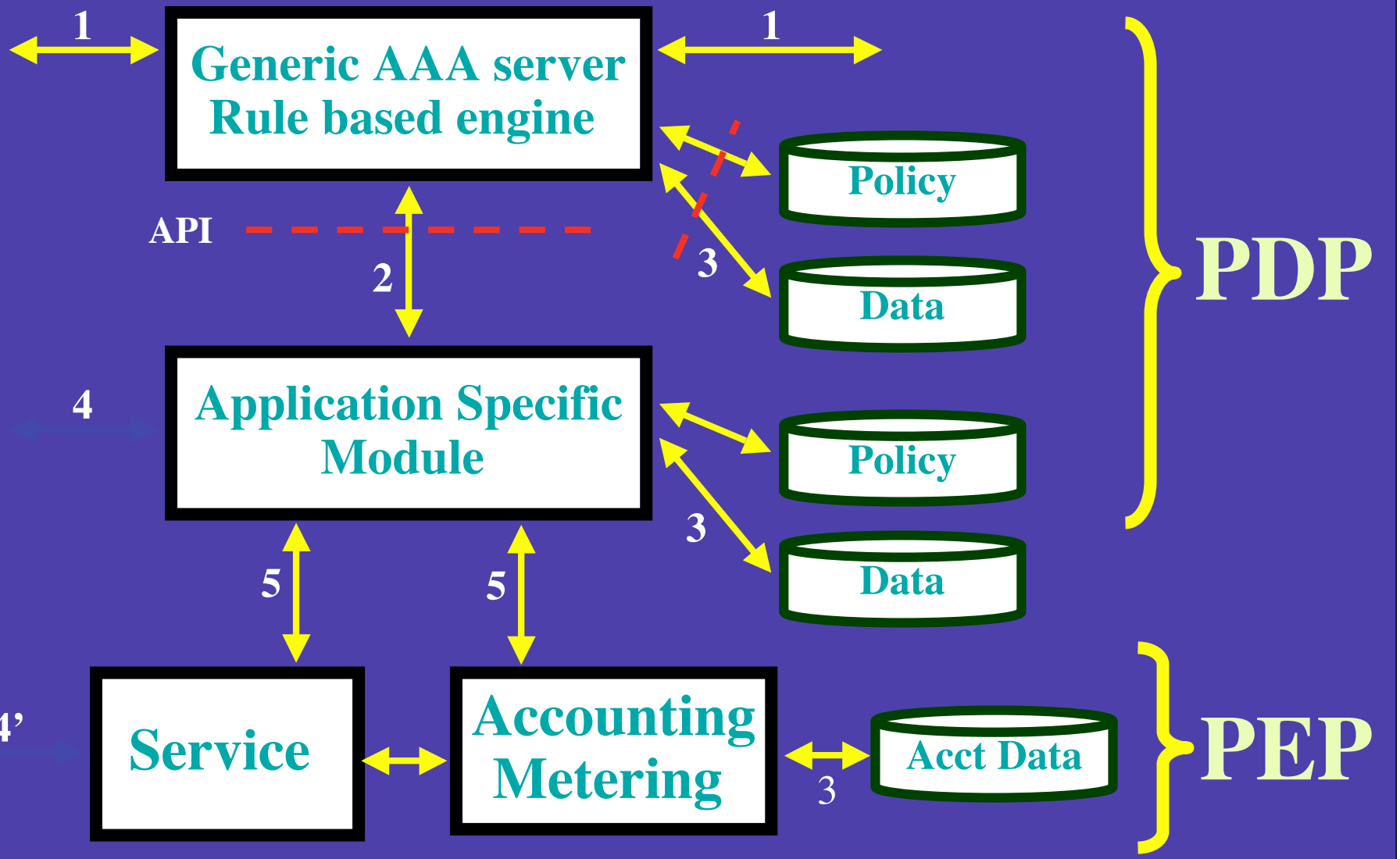
Forbidden area, solutions for s when f = 1 Gb/s, M = 0.5 Mbyte AND NOT USING FLOWCONTROL





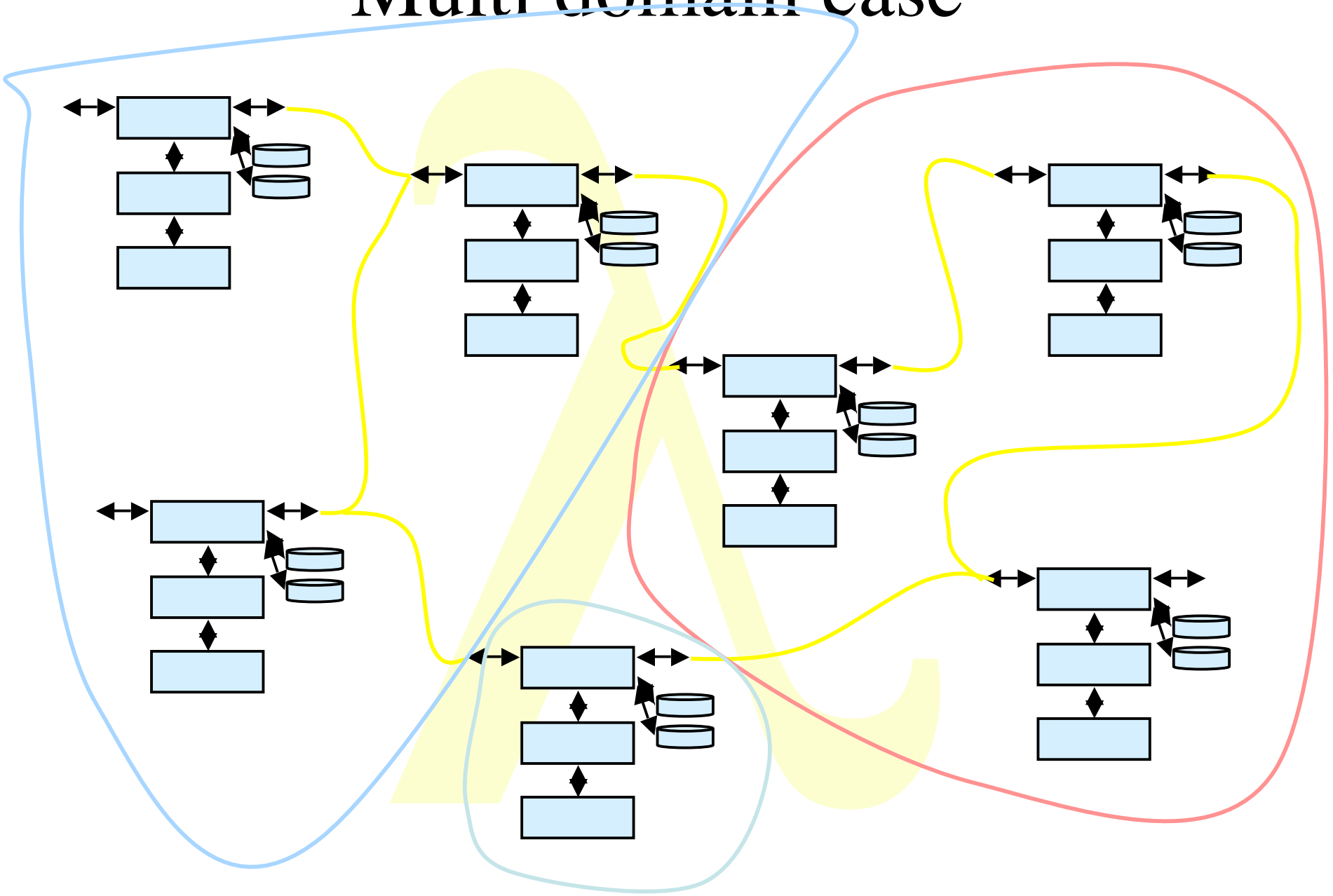
layers of increasing abstraction taxonomy

Starting point



RFC 2903 - 2906 , 3334 , policy draft

Multi domain case



iGrid2002

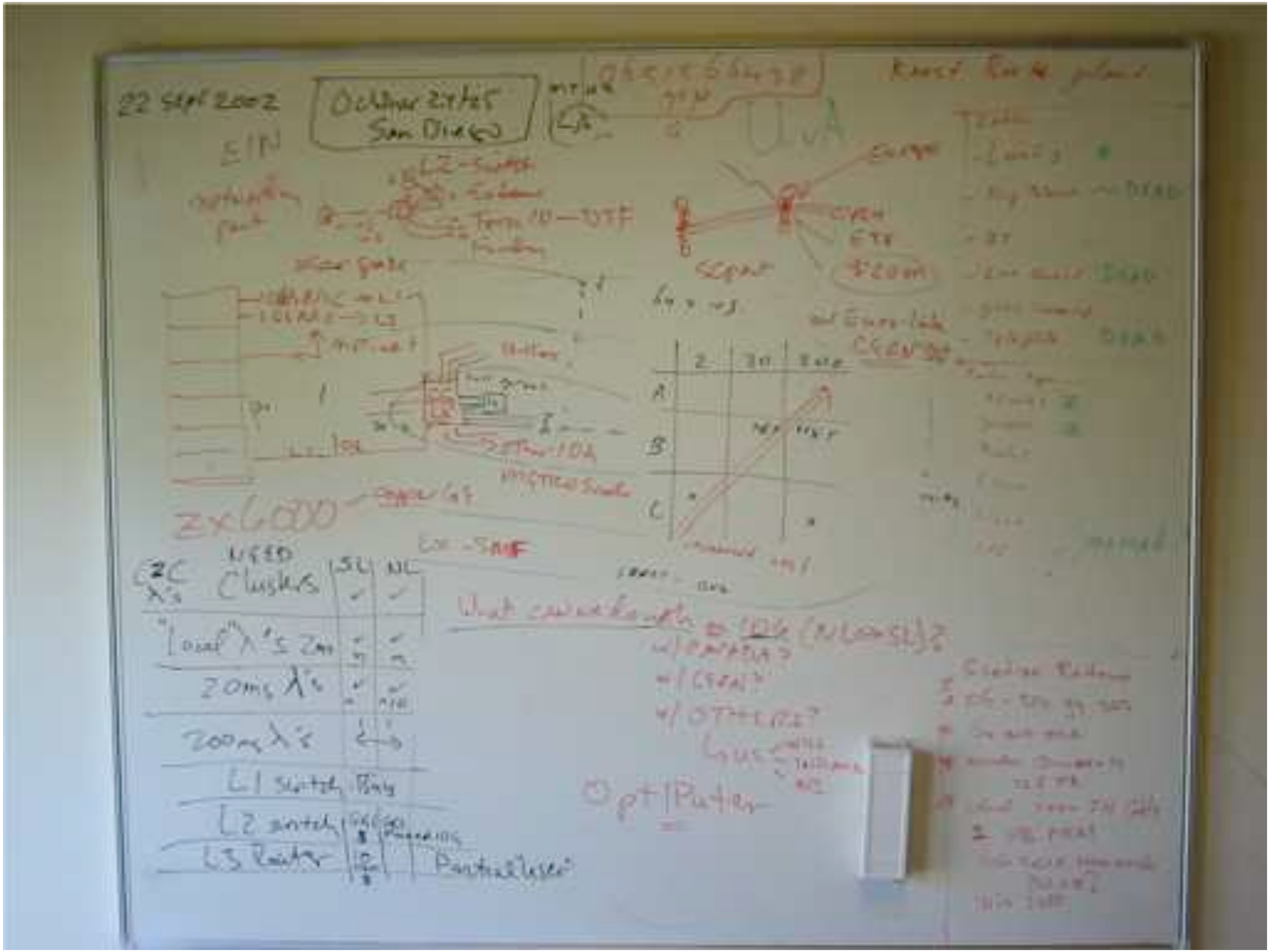
- www.igrid2002.org
- 25 demonstrations
- 16 countries (at least)
- Level3, Tyco, IEEAF Lambda's
- CISCO, Hp equipment sponsoring
- Shipping nightmare, debugging literally
- ~30 Gbit/s International connectivity
- Huge networking collaboration
- Smelly NOC in the iGrid preparation weekend

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

NOC

GridFTP testcluster


22 Sept 2002
October 24 to 25
San Diego
ZX6000
Clusters
SL
NL
L1 switch
L2 switch
L3 Router

American tourist in Amsterdam

Same guy spotted in San Diego

Revisiting the truck of tapes

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 PByte / 1600 GByte/sec = 3125 s ≈ 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 Neggers

UIC&iCAIR: Tom DeFanti, Joel Mambretti

CANARIE: Bill St. Arnaud

This work is supported by:

SURFnet

EU-IST project DATATAG