The Road towards Optical Networking

www.science.uva.nl/~delaat

Cees de Laat

EU

SURFnet

University of Amsterdam

SARA

NikHef

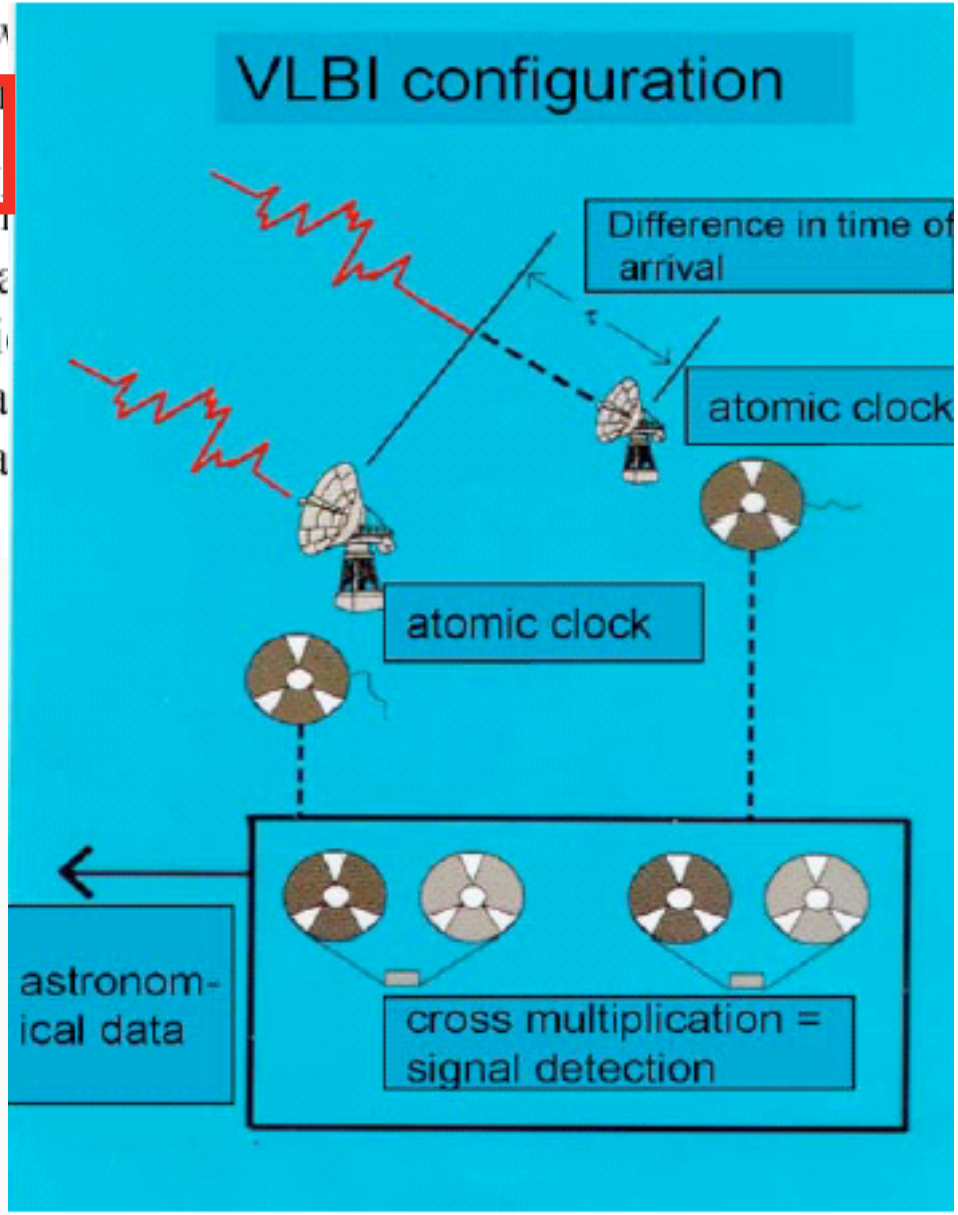
What is this buzz about optical networking

- **Networks are already optical for ages**
- **Users won't see the light**
- **Almost all current projects are about SONET circuits and Ethernet (old wine in new bags?)**
- **Are we going back to the telecom world, do NRN's want to become telco's**
- **Does it scale**
- **Is it all about speed or is it integrated services**

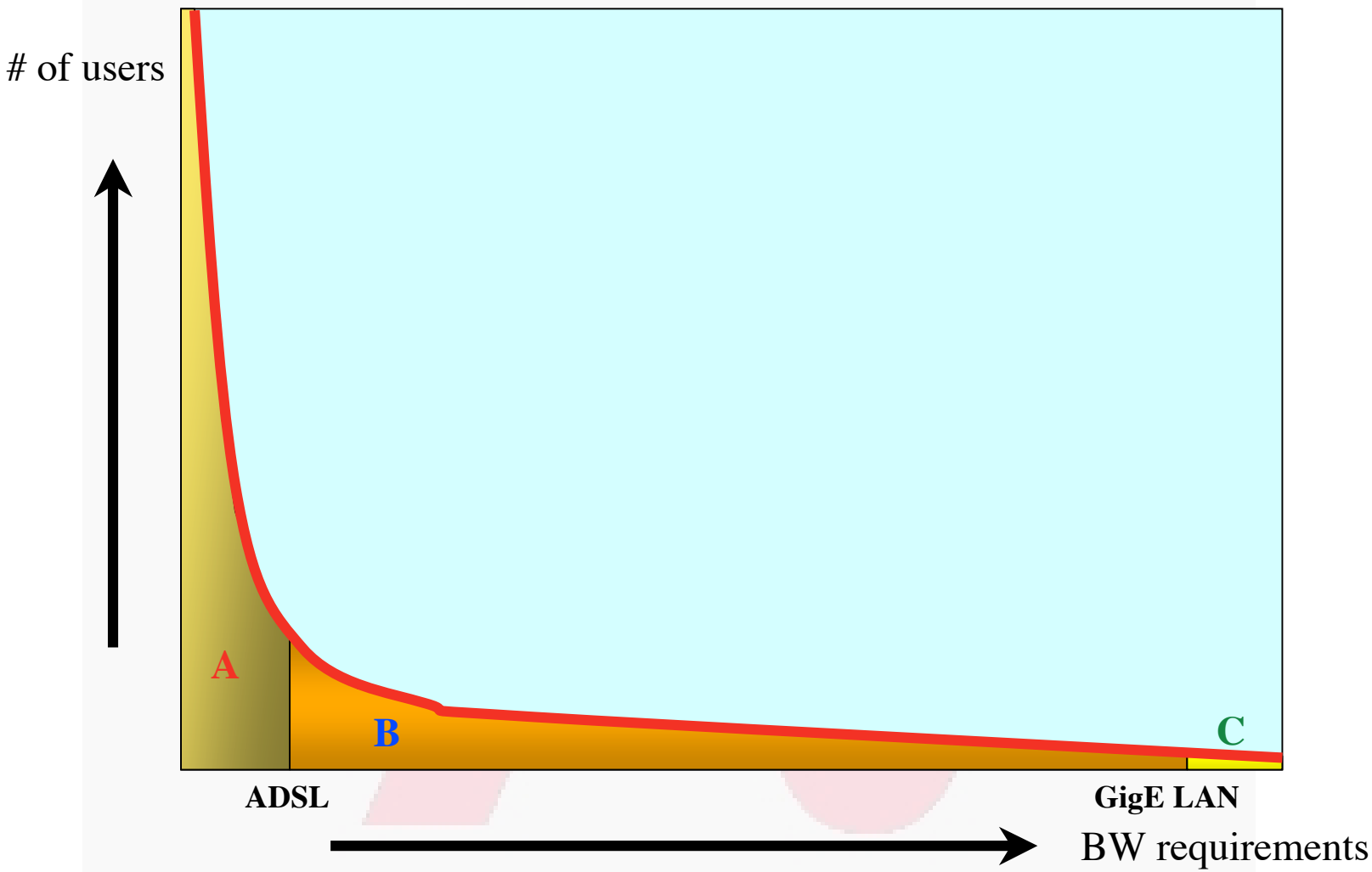
VLBI

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
. Rates of 8Gb/s or more are entirely feasibl
der 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

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

So, what's up doc

Suppose:

- **Optical components get cheaper and cheaper**
- **Dark (well, dark?) fibers abundant**
- **Number of available λ/user -> ∞**
- **Speeds of 10, 100, 1000 Gbit/s make electrical domain packet handling physically difficult**
  - **150 bytes @ 40 Gbit/s = 30 ns = 15 meter fiber**
  - **QoS makes no sense at these speeds**
- **Cost per packet forwarding lower at L1 / L2**

Then:

- **Long term view ---> full optical**
- **λ provisioning for grid applications**
- **How low can you go**

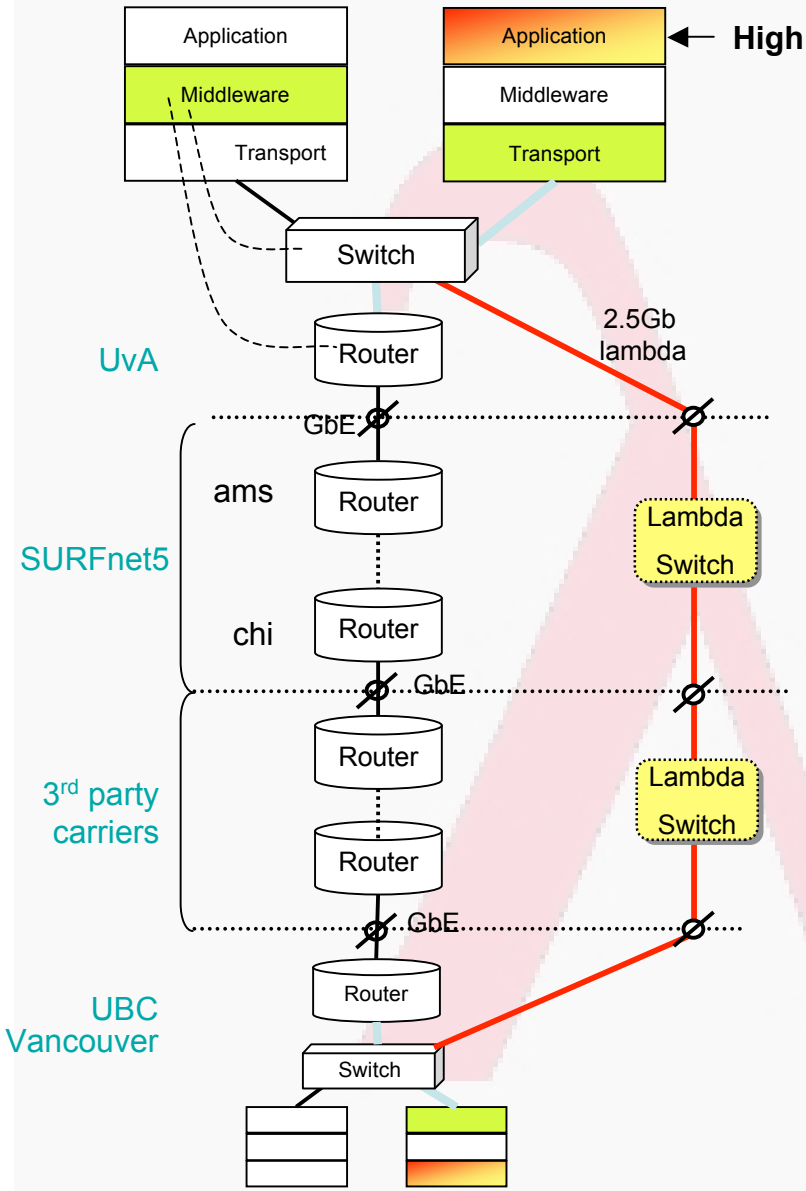
Optical networking, 3 scenarios

- **Lambdas for internal ISP bandwidth provisioning**
  - **An ISP uses a lambda switching network to make better use of its (suppliers) dark fibers and to provision to the POP's. In this case the optical network is just within one domain and as such is a relatively simple case.**
- **Lambda switching as peering point technology**
  - **In this use case a layer 1 Internet exchange is build. ISP's peer by instantiating lambdas to each other. Is a N*(N-1) and multi domain management problem.**
- **Lambda switching as grid application bandwidth provisioning**
  - **This is by far the most difficult since it needs UNI and NNI protocols to provision the optical paths through different domains.**

Current technology + (re)definition

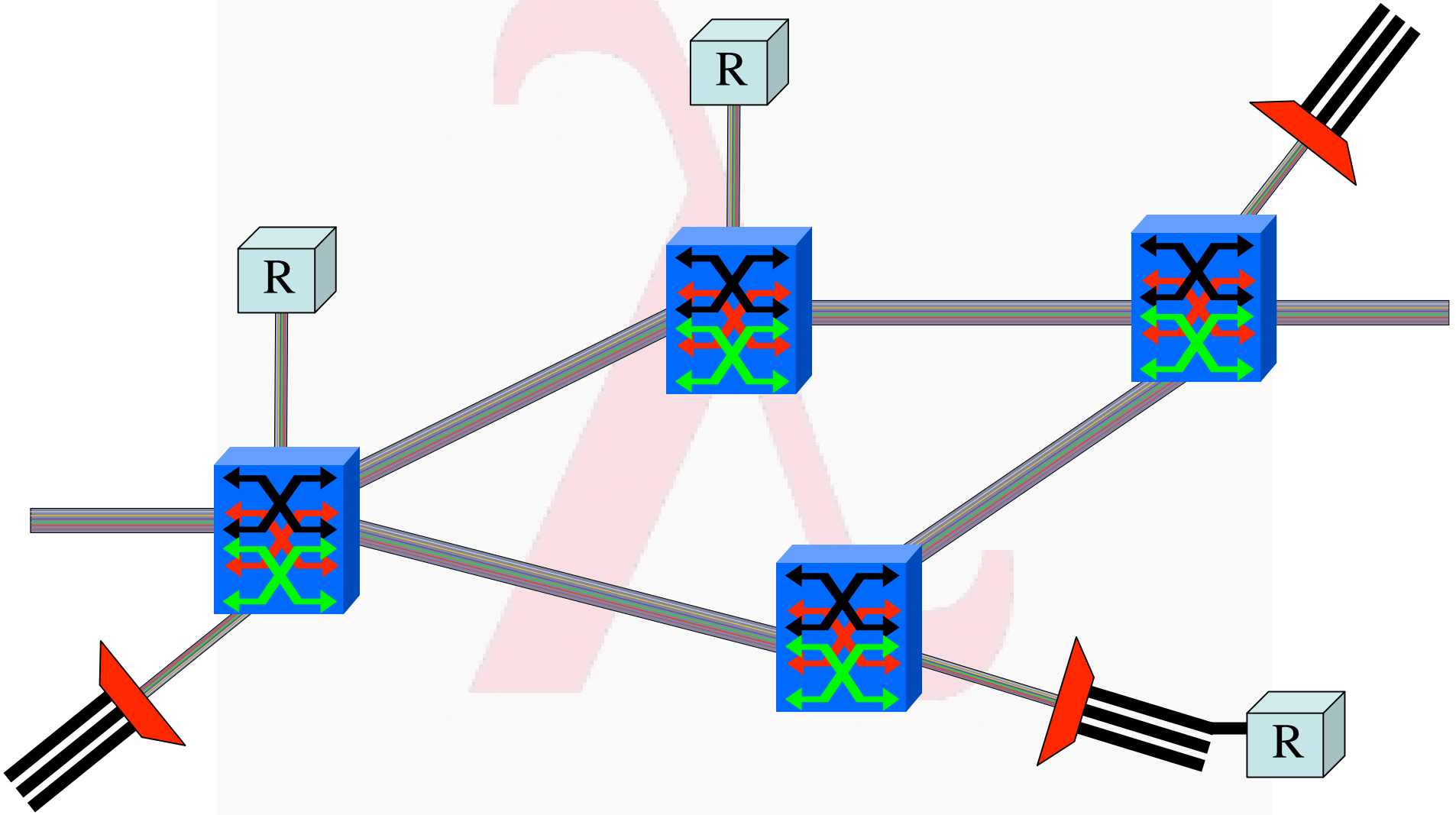
- Current (to me) available technology consists of SONET/SDH switches
- DWDM+switching coming up
- Starlight uses for the time being VLAN's on Ethernet switches to connect [exactly] two ports
- So redefine a λ as:

"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"

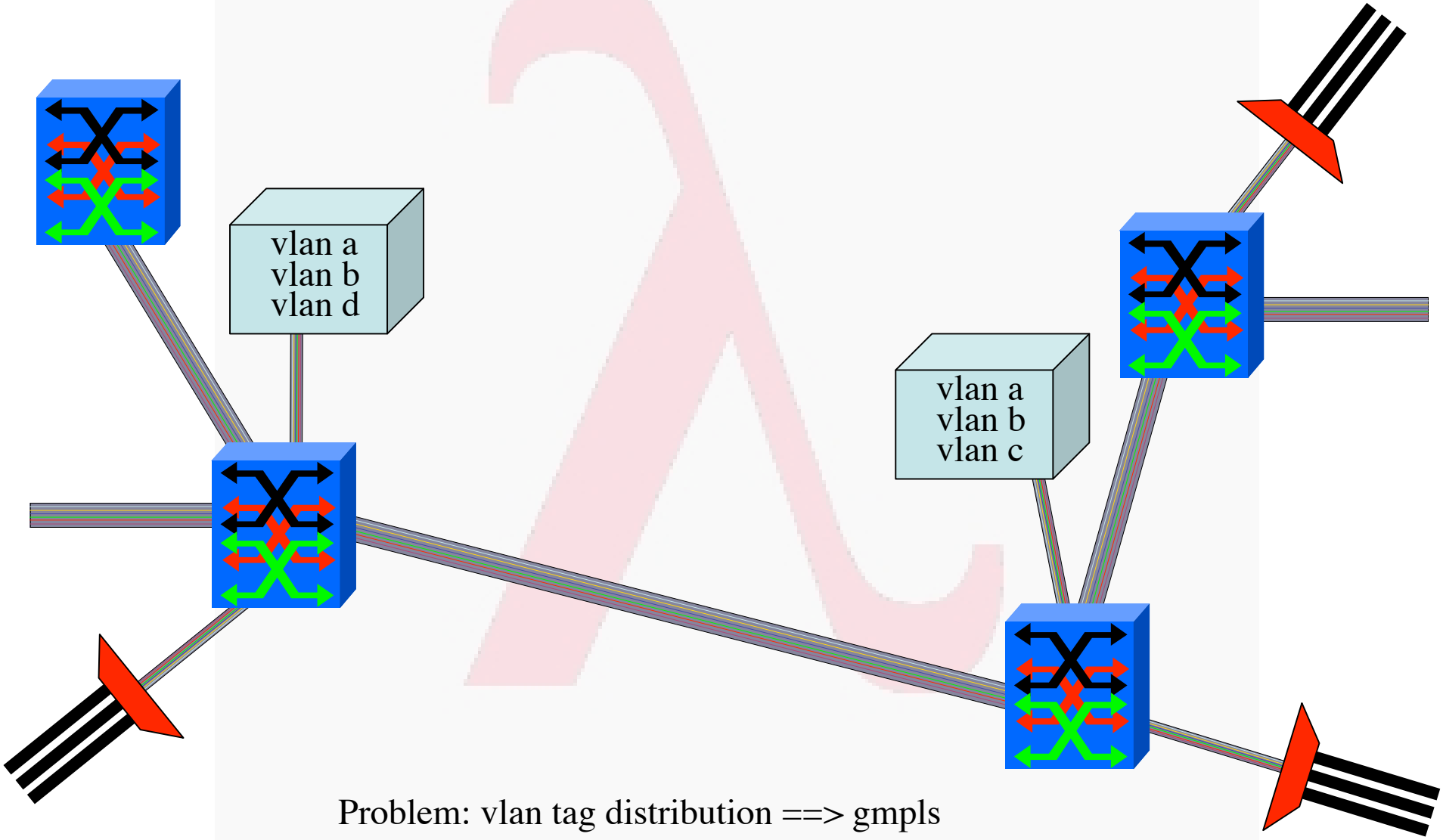


- lambda for high bandwidth applications
  - Bypass of production network
  - Middleware may request (optical) pipe

Other architectures - L1 - 3

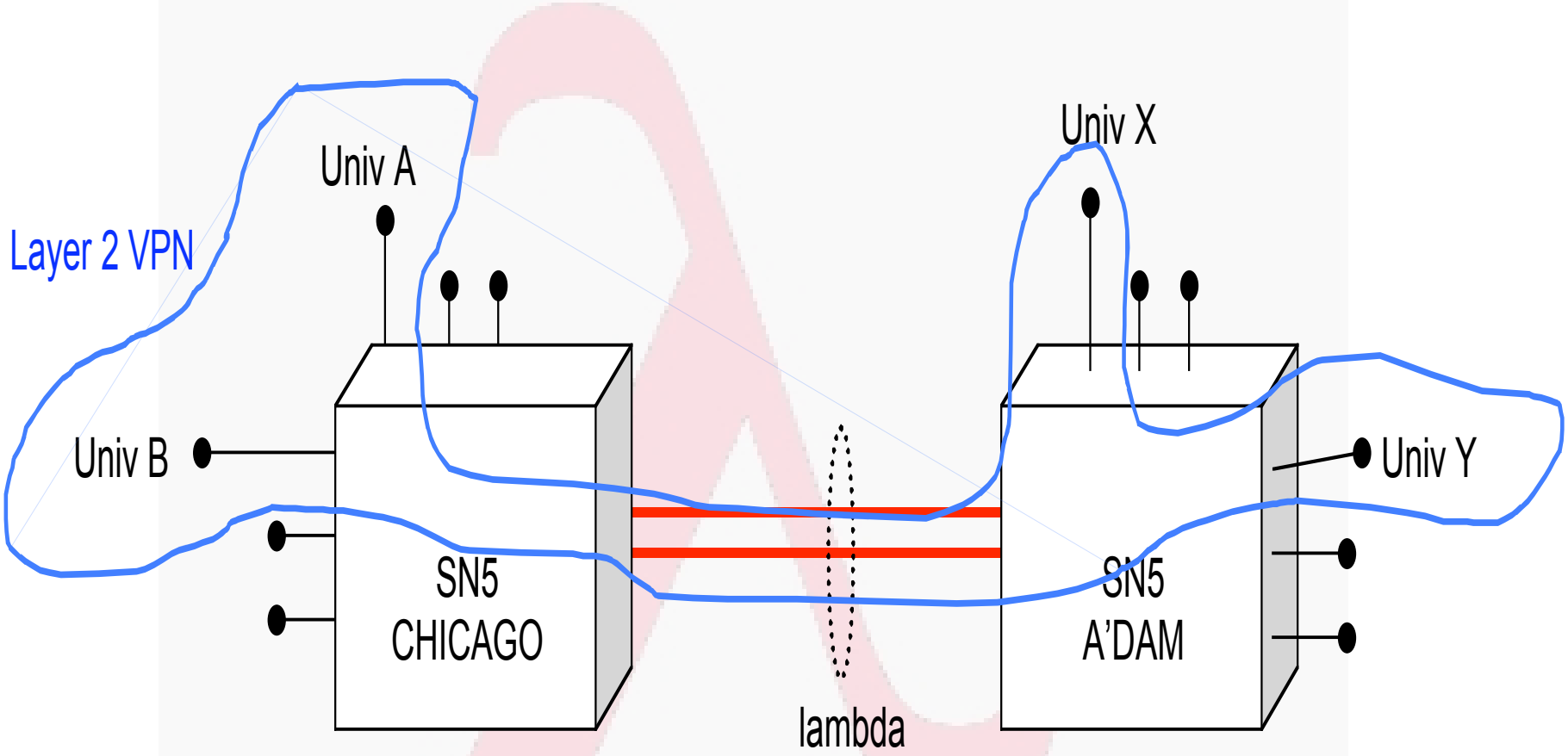


Other architectures - Distributed virtual IEX'es



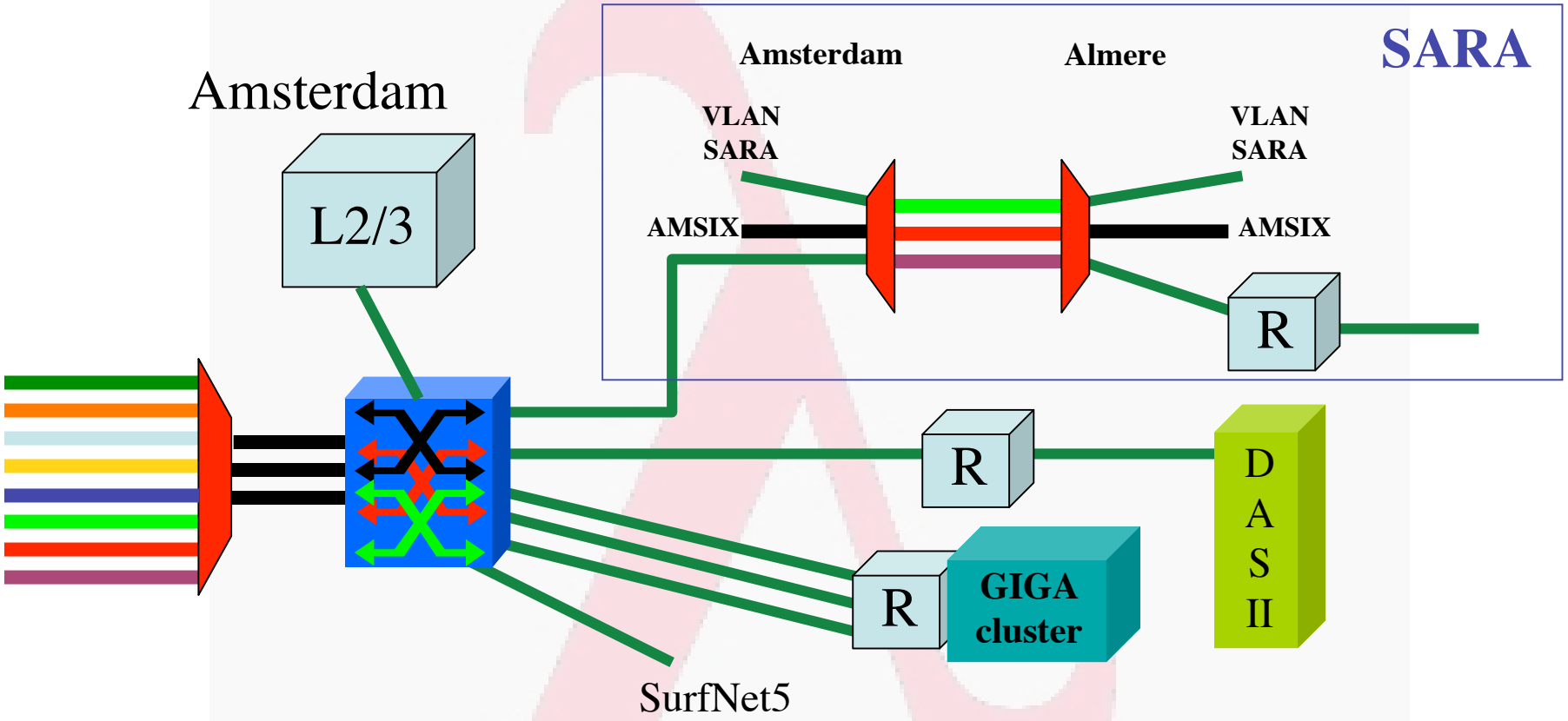
Problem: vlan tag distribution ==> gmpls

Distributed L2



2.5 Gbps SONET/SDH "Lambda"

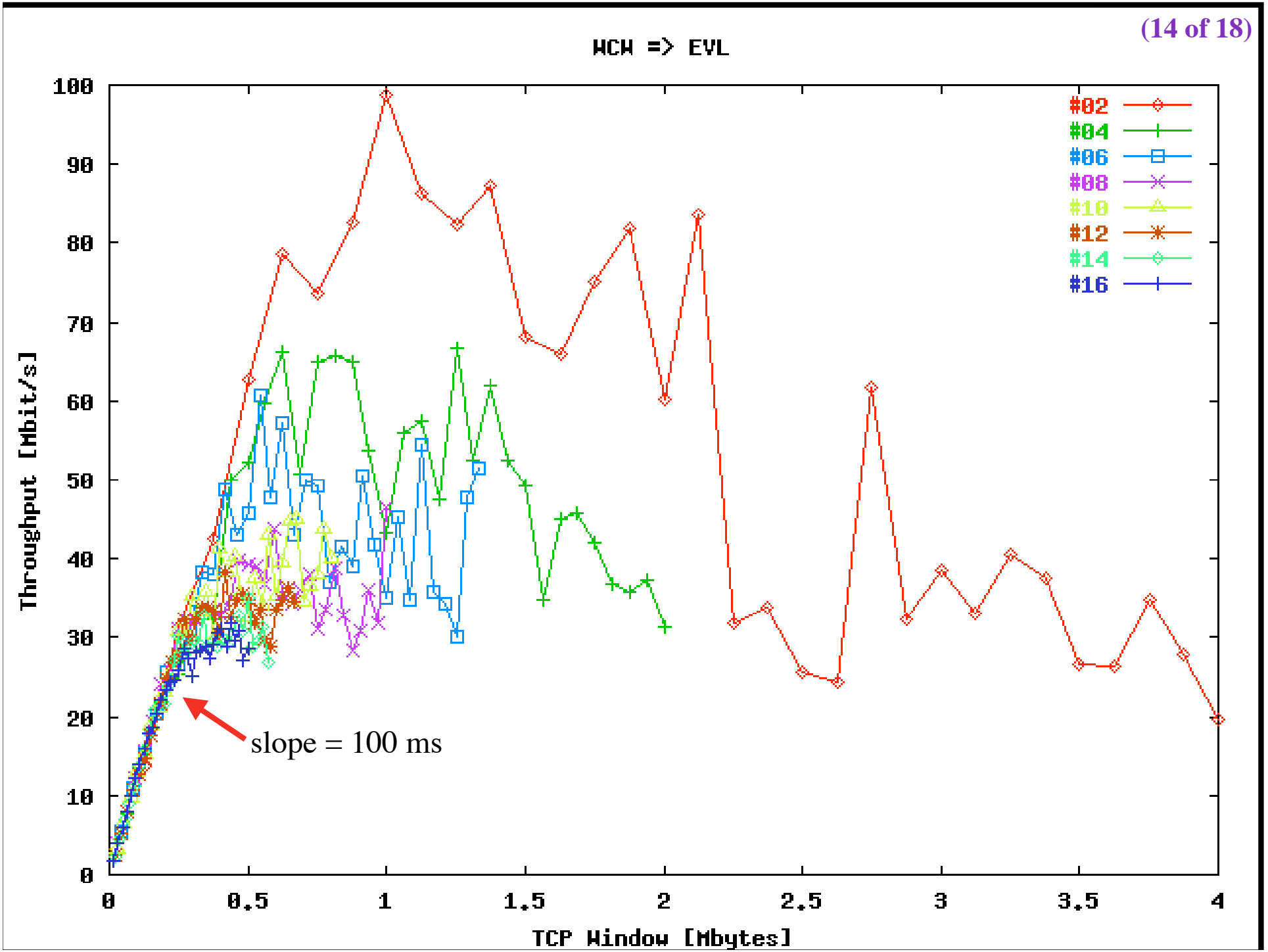
10/100/1000 Mbps Ethernet



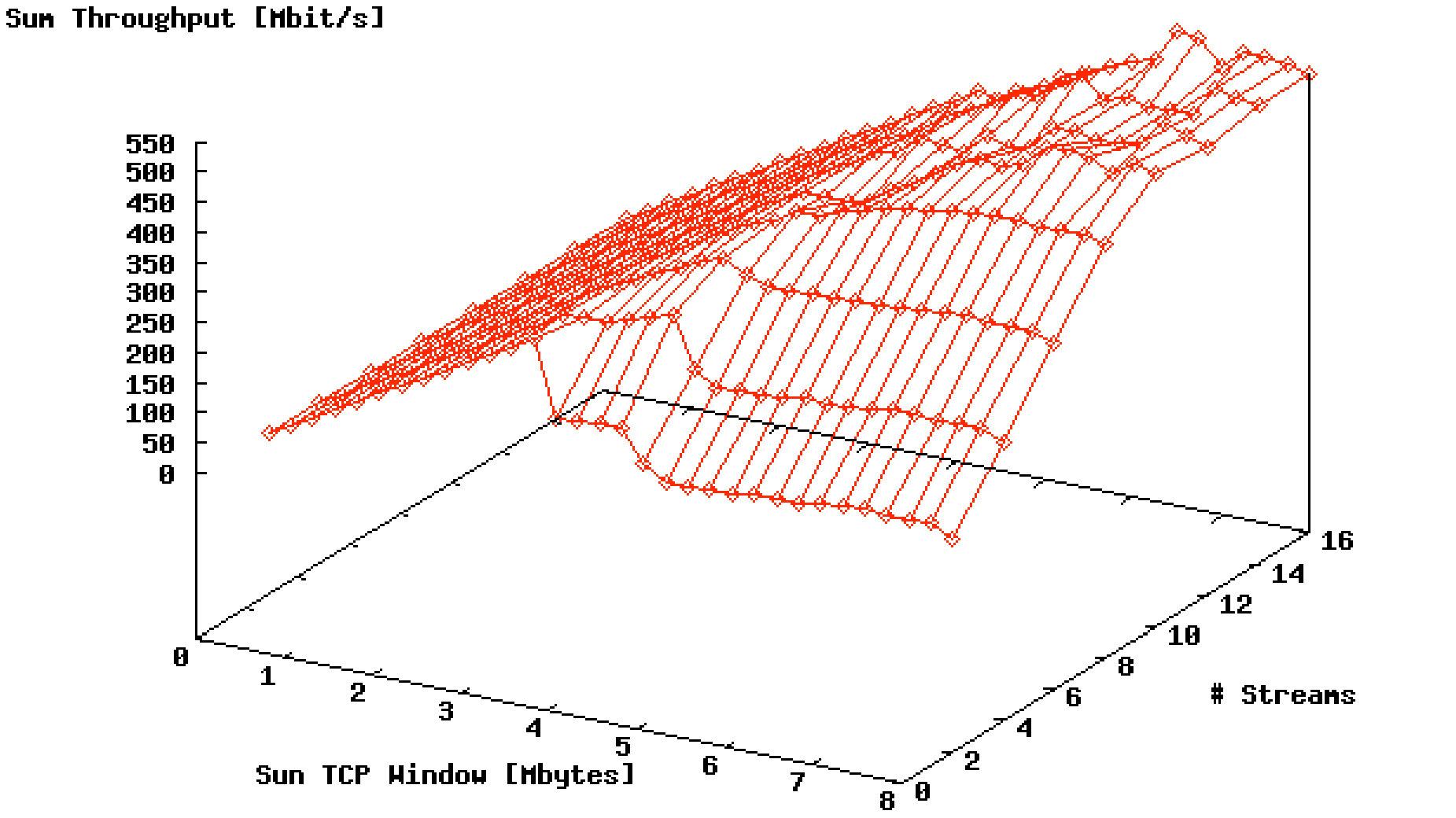
Amsterdam 2nd phase

First experiences with SURFnet pure for research Lambda

- **2.5 Gbit SONET λ delivered dec 2001**
  - **Took about 3 months, should be 300 ms**
- **First generation equipment delivered nov 2001**
- **Back to back tests => OC12 limit -> 560 Mbit/s**
- **1 unit shipped to Chicago (literally, took 3 weeks)**
- **End to end now 80 Mbit/s**
- **So, what is going on?**
- **Second generation equipment just delivered**
- **1 unit shipped to Chicago (yes, is going to take 3 weeks)**


WCW => EVL
Throughput [Mbit/s]
TCP Window [Mbytes]
0
10
20
30
40
50
60
70
80
90
100
0
0.5
1
1.5
2
2.5
3
3.5
4
#02
#04
#06
#08
#10
#12
#14
#16
slope = 100 ms

EVL => WCH


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

Layer - 2 requirements from 3/4

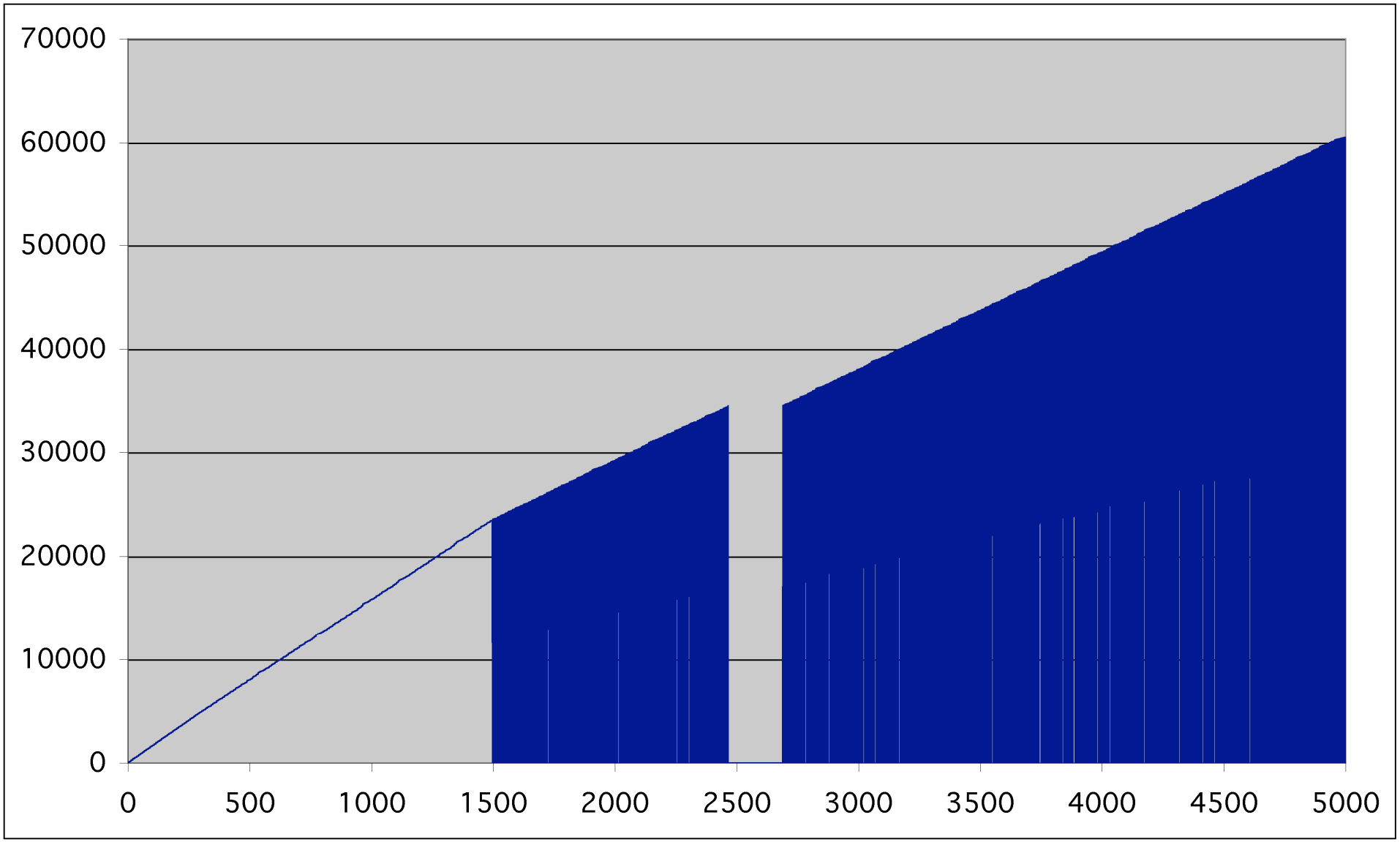
TCP is bursty due to sliding window protocol and slow start algorithm. So pick from menu:

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

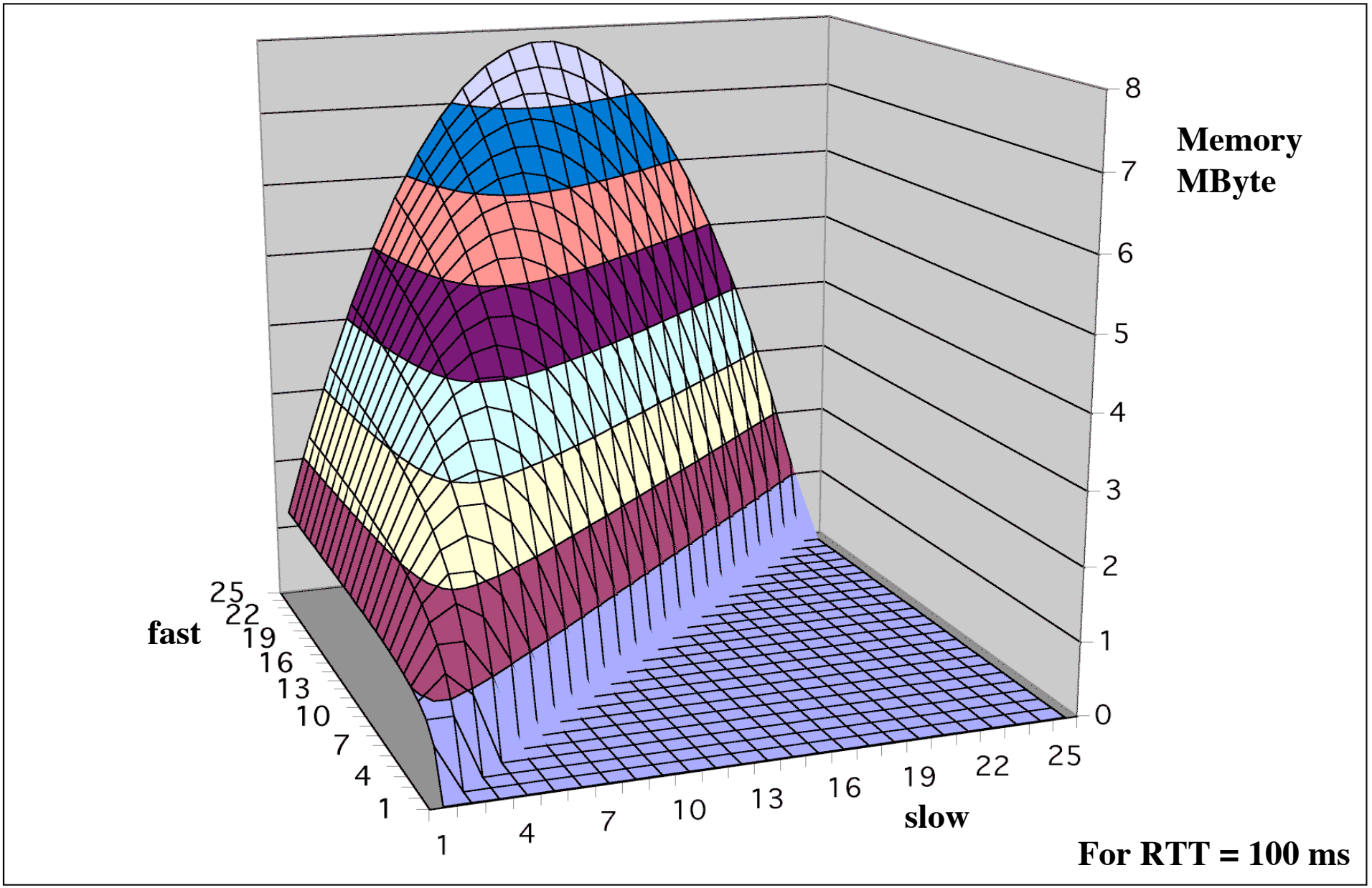
Window = BandWidth * RTT & BW == slow

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

5000 1 kByte UDP packets



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



Layer - 2 requirements from 3/4

Window = BandWidth * RTT & BW == slow

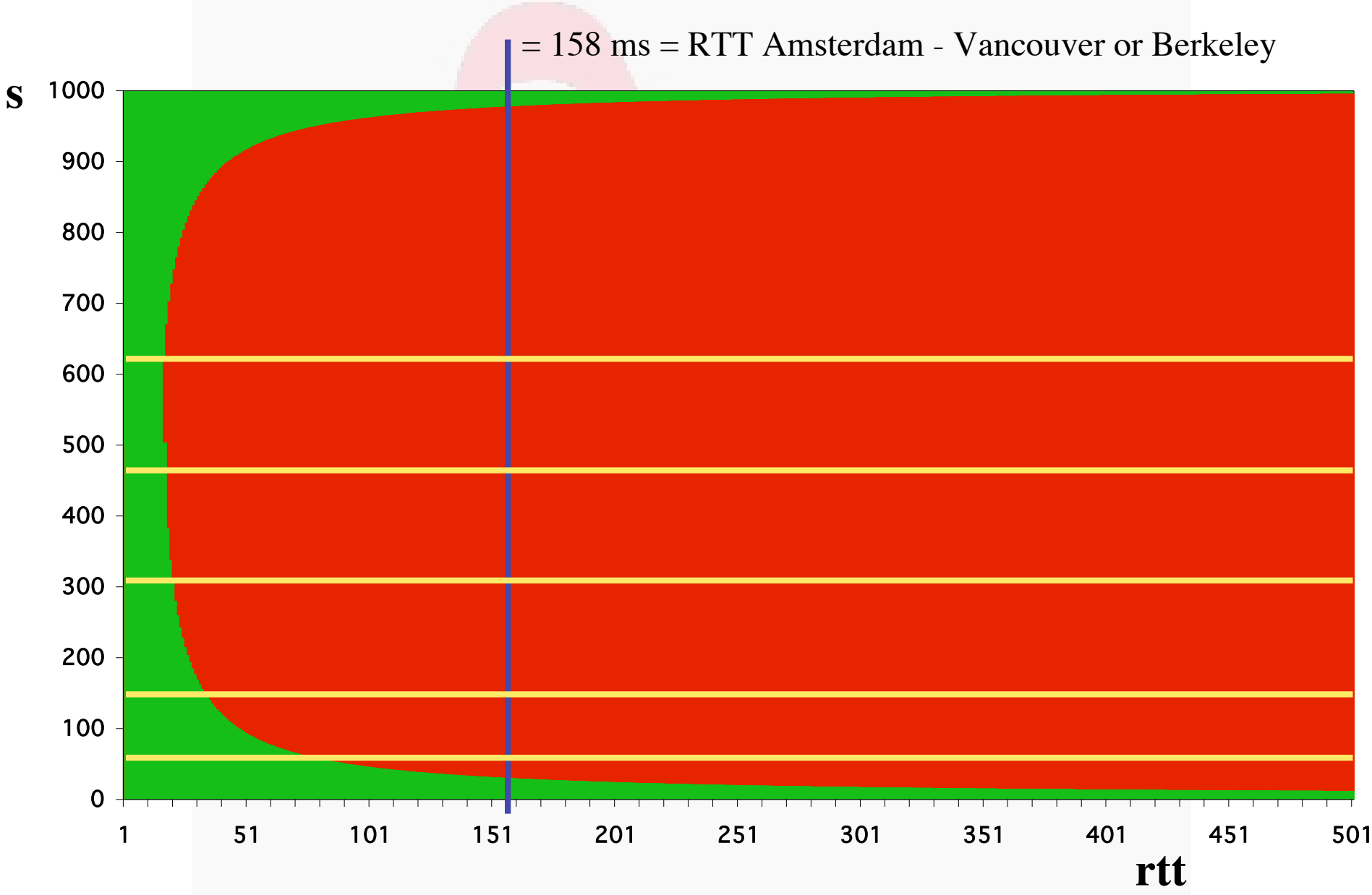
$$\text{Memory-at-bottleneck} = \frac{\text{fast} - \text{slow}}{\text{fast}} * \text{slow} * \text{RTT}$$

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

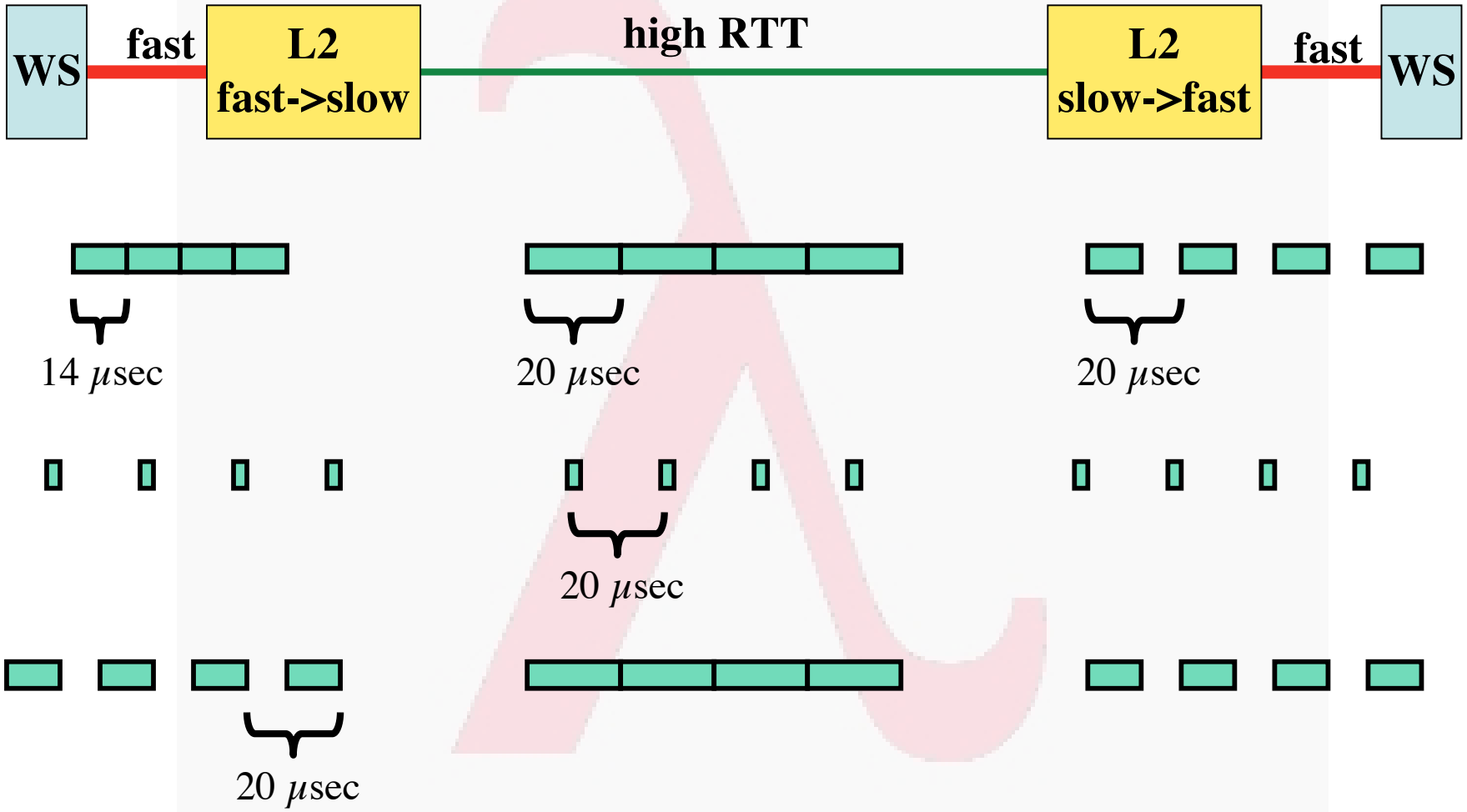
$$0 = s^2 - f * s + \frac{f * M}{RTT}$$

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

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



Self-clocking of TCP



Revisiting the truck of tapes

Consider one fiber

- **Current technology allows for 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!!!**

iGrid 2002

The International Virtual Laboratory

www.igrid2002.org

24-26 September 2002
Amsterdam Science and Technology Centre (WTCW)
The Netherlands

- A showcase of applications that are "early adopters" of very-high-bandwidth national and international networks
  - What can you do with a 10Gbps network?
  - What applications have insatiable bandwidth appetites?
- Scientists and technologists to optimally utilize 10Gbps experimental networks, with special emphasis on e-Science, Grid and Virtual Laboratory applications
- Registration is open (www.igrid2002.org)
- iGrid is not just a conference/demonstration event, it is also a testbed!!
- Contact
  - maxine@startap.net or deLaat@science.uva.nl