High Performance Networking for Grid Applications

www.science.uva.nl/~delaat

Cees de Laat

Faculty of Science

High Performance Networking for Grid Applications

www.science.uva.nl/~delaat

Cees de Laat

EU

SURFnet

University of Amsterdam

SARA

NIKHEF

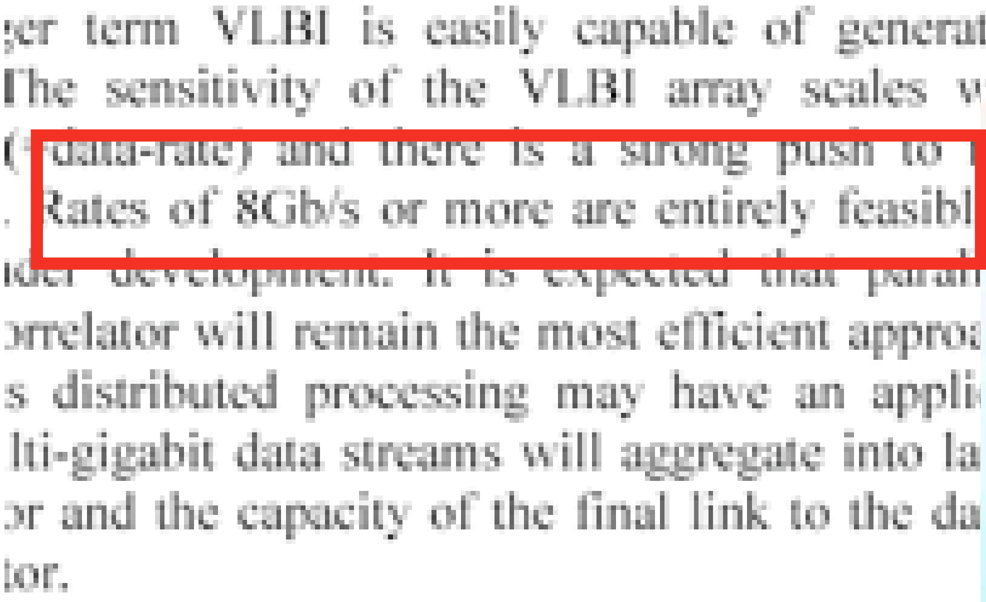
NCF

Contents of this talk

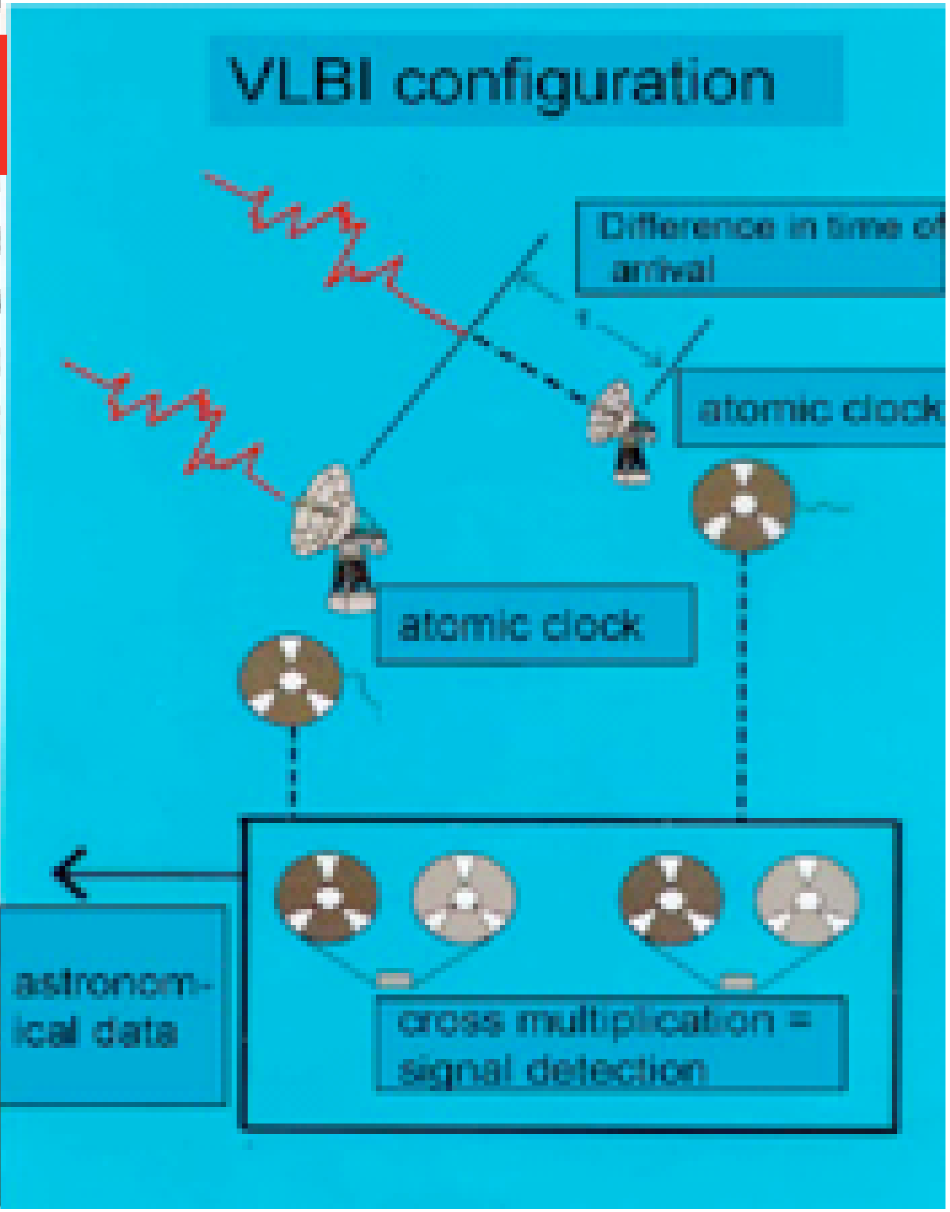
This slide is intentionally left blank

eVLBI

VLBI



ger term VLBI is easily capable of generating many Gb of data per
The sensitivity of the VLBI array scales with the square root of the
(√data-rate) and there is a strong push to i
. Rates of 8Gb/s or more are entirely feasibl
nder development. It is expected that parall
orrelator will remain the most efficient approa
s distributed processing may have an applic
lti-gigabit data streams will aggregate into la
or and the capacity of the final link to the da
tor.



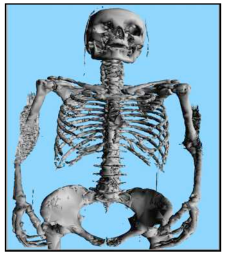
VLBI configuration

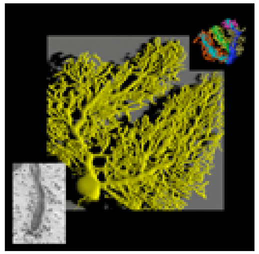
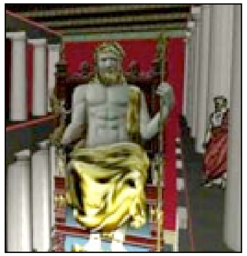
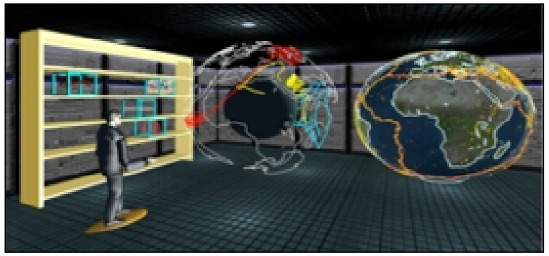
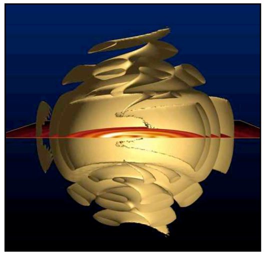
Difference in time of arrival
atomic clock
atomic clock
astronom-ical data
cross multiplication = signal detection

iGrid 2002

September 24-26, 2002, Amsterdam, The Netherlands

- 28 demonstrations from 16 countries: Australia, Canada, CERN, France, Finland, Germany, Greece, Italy, Japan, The Netherlands, Singapore, Spain, Sweden, Taiwan, United Kingdom, United States
- Applications demonstrated: art, bioinformatics, chemistry, cosmology, cultural heritage, education, high-definition media streaming, manufacturing, medicine, neuroscience, physics, tele-science



- Grid technologies demonstrated: Major emphasis on grid middleware, data management grids, data replication grids, visualization grids, data/visualization grids, computational grids, access grids, grid portals
- 25Gb transatlantic bandwidth (100Mb/attendee, 250x iGrid2000!)

www.igrid2002.org

(3b of 12)

Experimental Networks

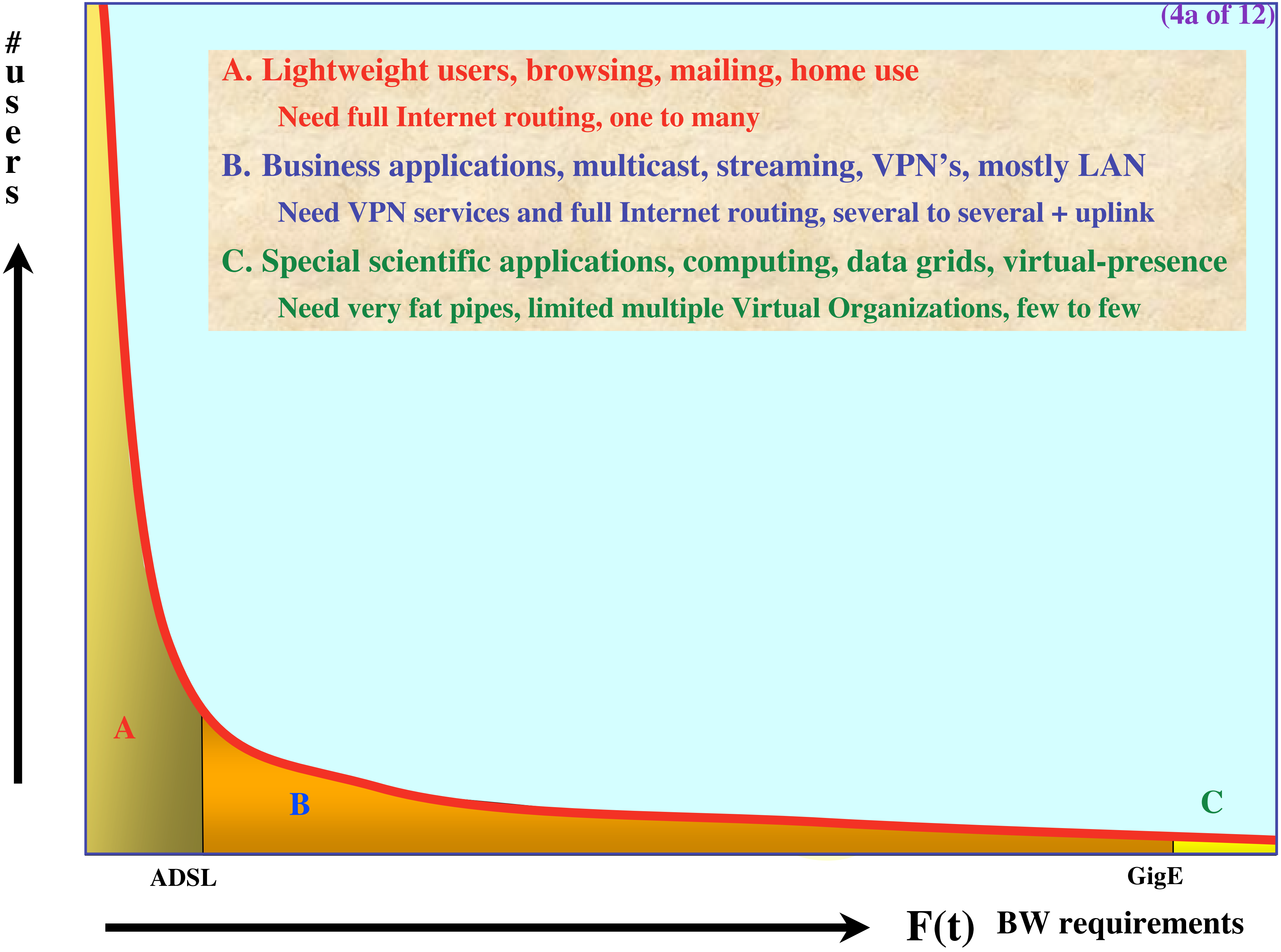
- **High-performance trials of new technologies that support *application-dictated* development of software toolkits, middleware, computing and networking.**
- **Provide *known and knowable characteristics* with deterministic and repeatable behavior on a persistent basis, while encouraging experimentation with innovative concepts.**
- **Experimental Networks are seen as the *missing link* between Research and Production Networks.**

http://www.evl.uic.edu/activity/NSF/index.html
http://www.calit2.net/events/2002/nsf/index.html

What is a LambdaGrid?

- A *grid* is a set of networked, middleware-enabled computing resources.
- A *LambdaGrid* is a grid in which the lambda networks themselves are resources that can be scheduled, like all other computing resources. The ability to schedule and provision lambdas provides *deterministic* end-to-end network performance for real-time or time-critical applications, which cannot be achieved on today’s grids.

(4a of 12)



A. Lightweight users, browsing, mailing, home use

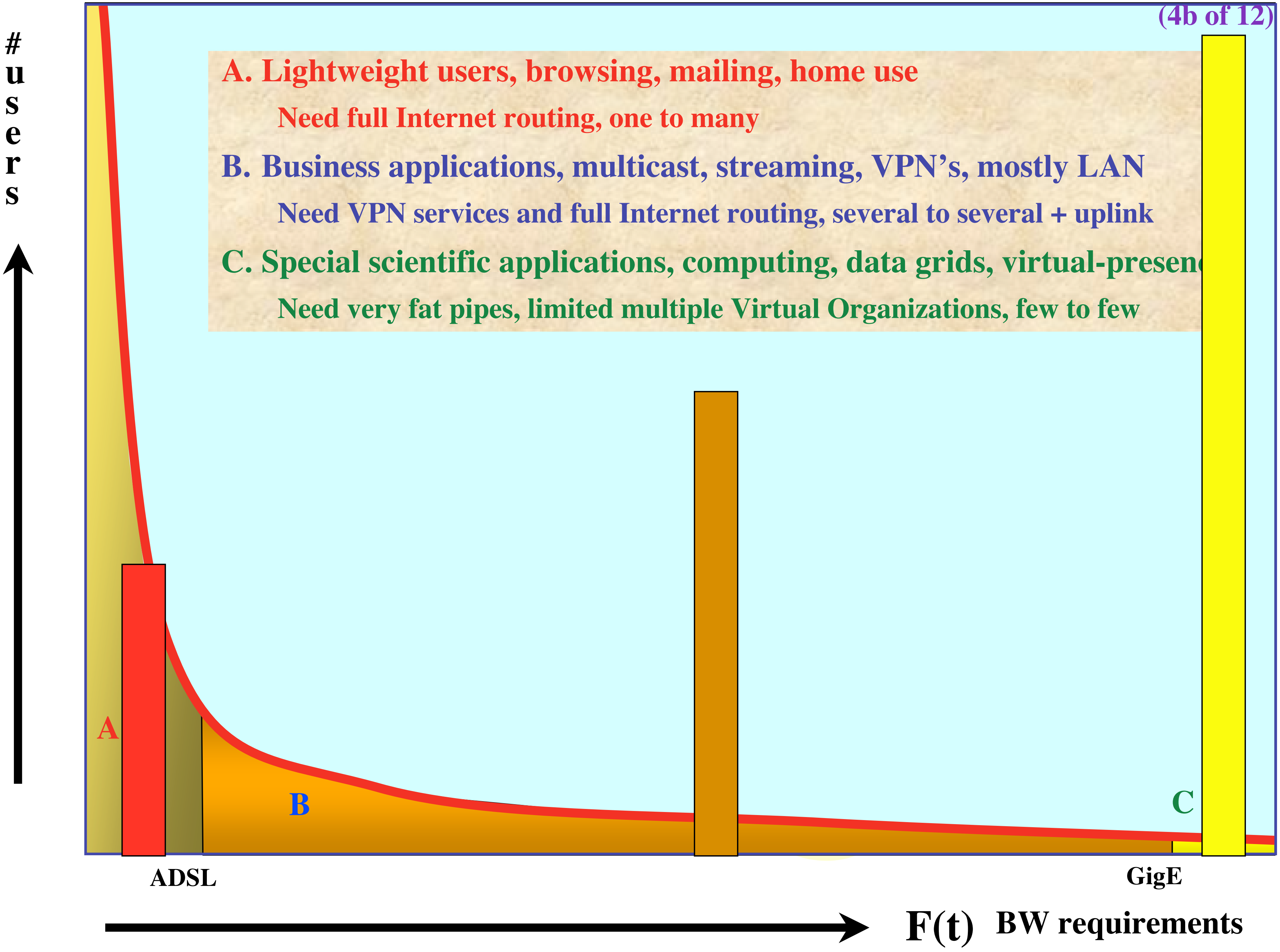
Need full Internet routing, one to many

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



A. Lightweight users, browsing, mailing, home use

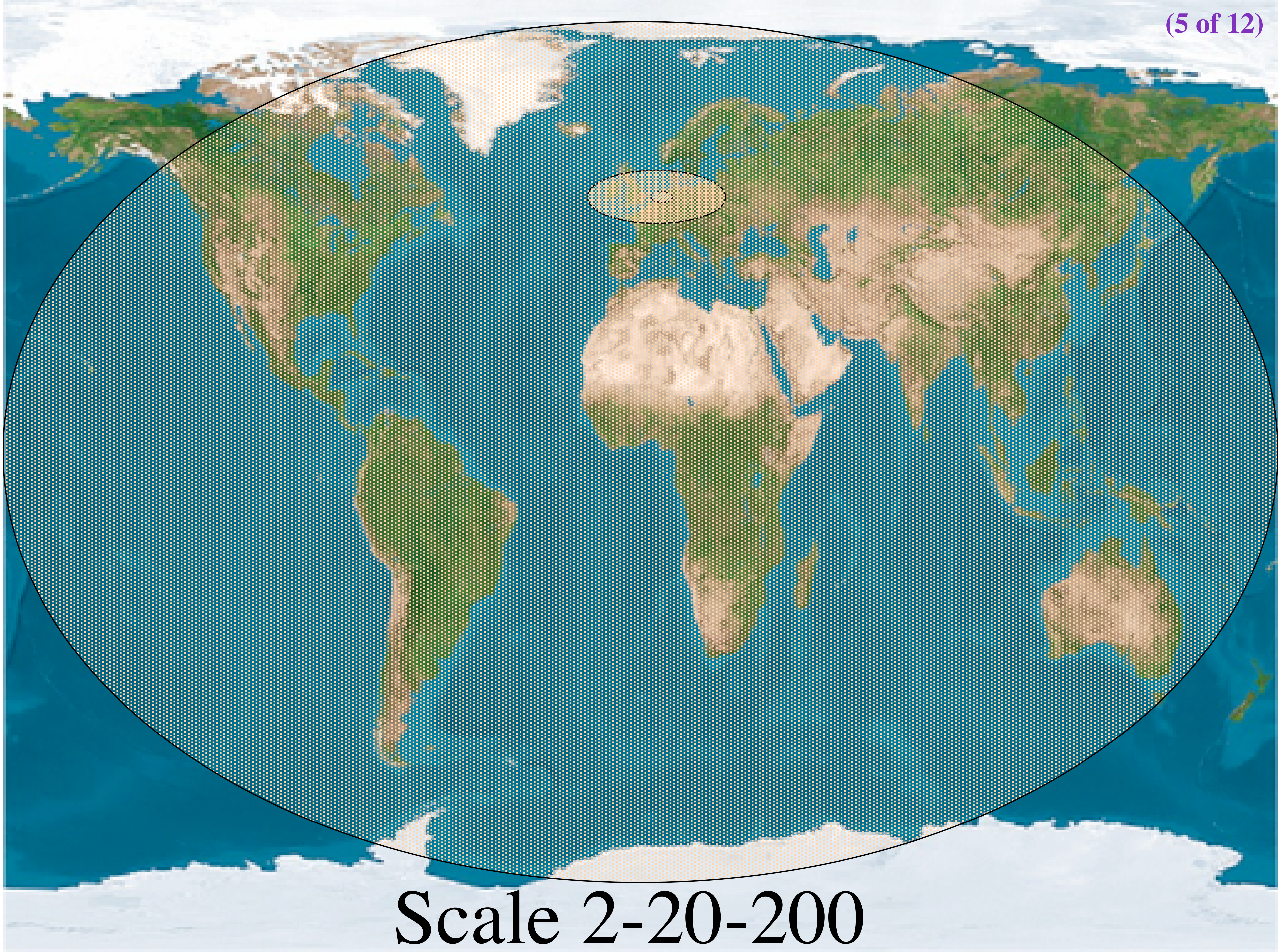
Need full Internet routing, one to many

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-presenc

Need very fat pipes, limited multiple Virtual Organizations, few to few



Scale 2-20-200

The only formula’s

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

Now, having been a High Energy Physicist we set

c = 1

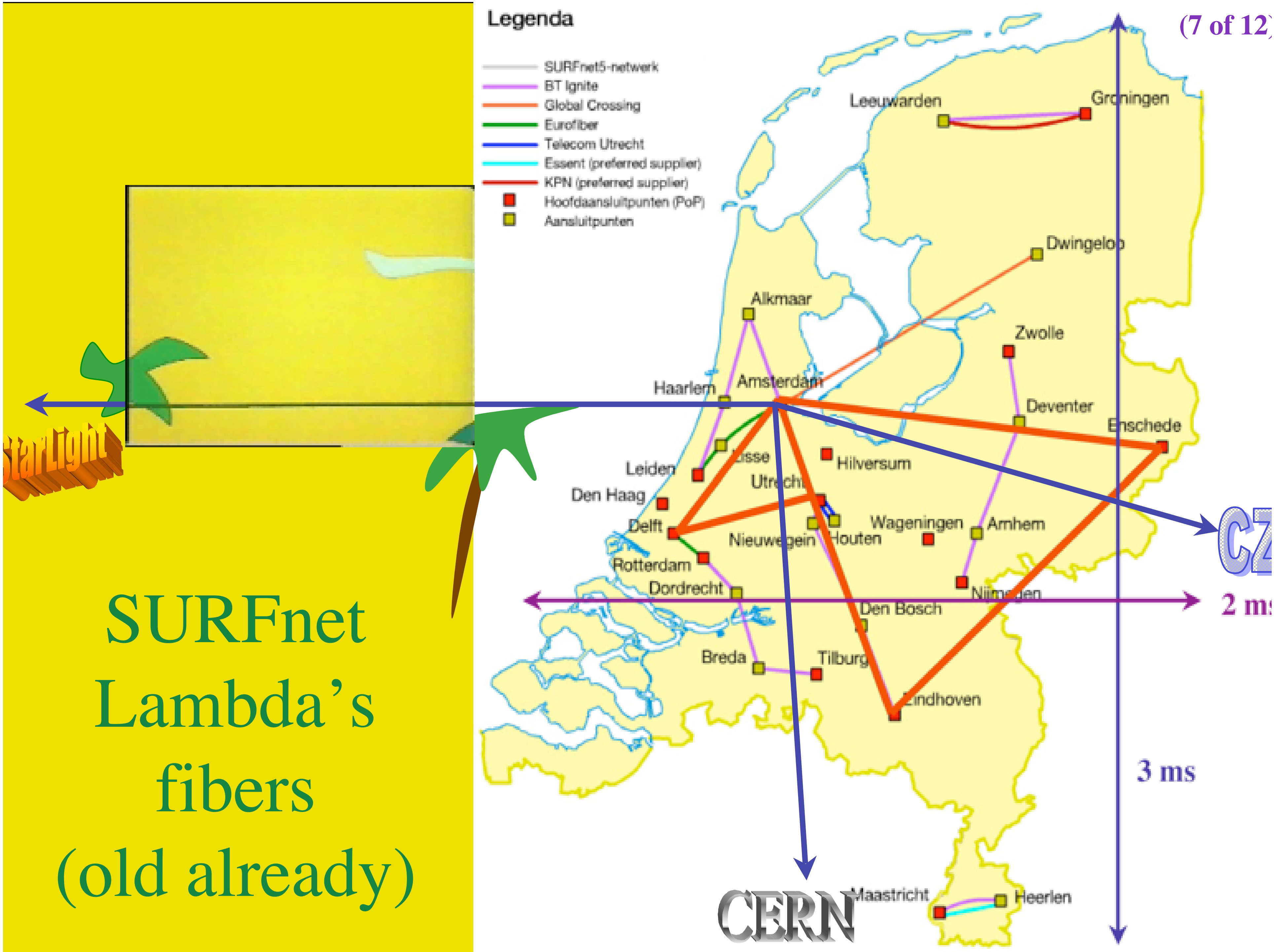
e = 1

$\bar{h}$ = 1

and the formula reduces to: $\# \lambda \approx \frac{200 * e^{(t-2002)}}{rtt}$

(7 of 12

SURFnet Lambda’s fibers (old already)



Services

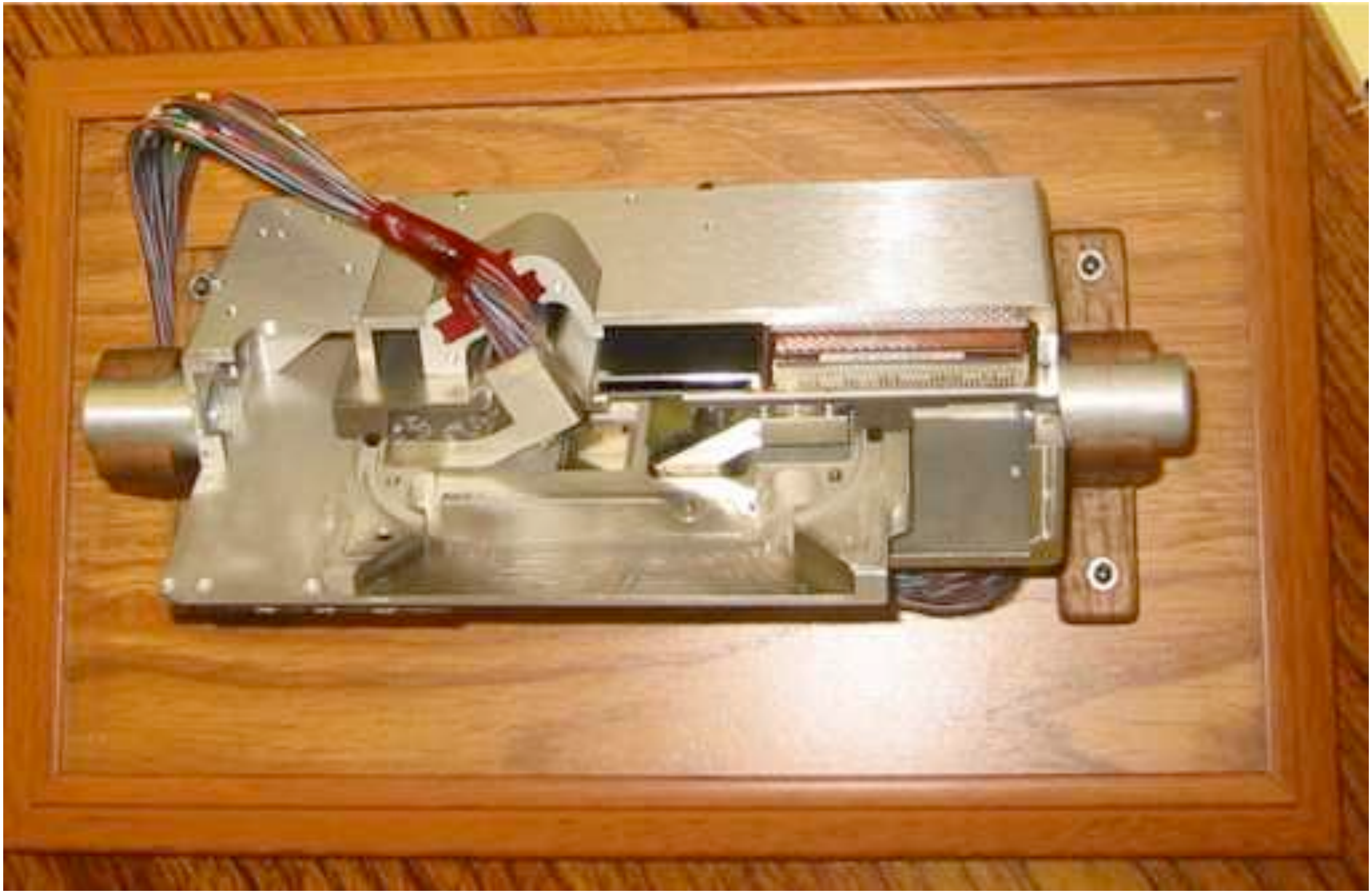
SCALE CLASS	2 Metro	20 National/ regional	200 World
A	Switching/ routing	Routing	ROUTER$
B	VPN's, (G)MPLS	VPN's Routing	Routing
C $\# \lambda \approx \frac{200 * e^{(t-2002)}}{rtt}$	dark fiber Optical switching	Lambda switching	Sub-lambdas, ethernet-sdh

Current technology + (re)definition

- Current (to me) available technology consists of SONET/SDH switches, 10 gig ethernet and dark fiber environments
- Optical switch installed (this week)!
- DWDM+switching included
- Starlight/NetherLight deploy 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”

MEMS optical switch (CALIENT)

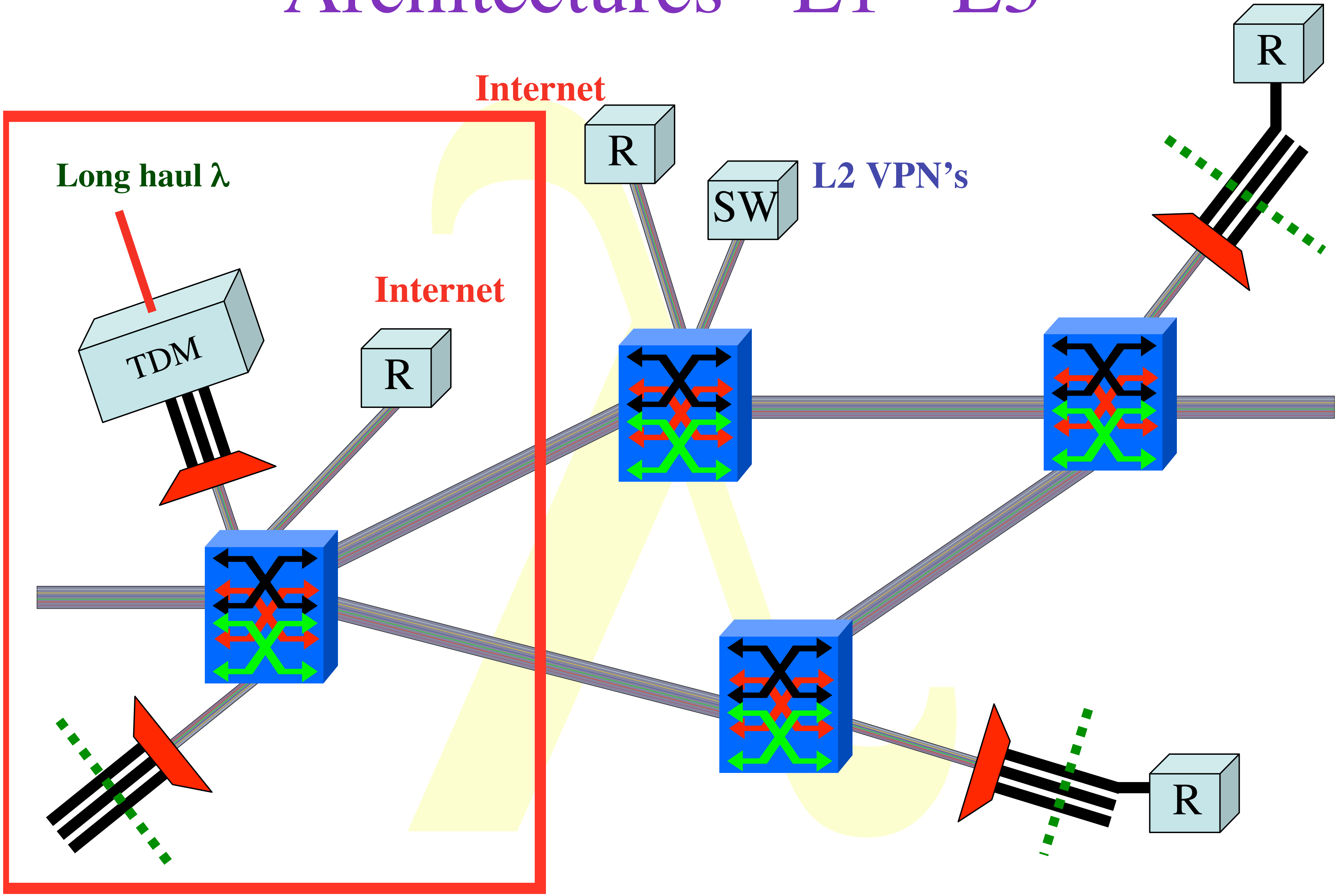


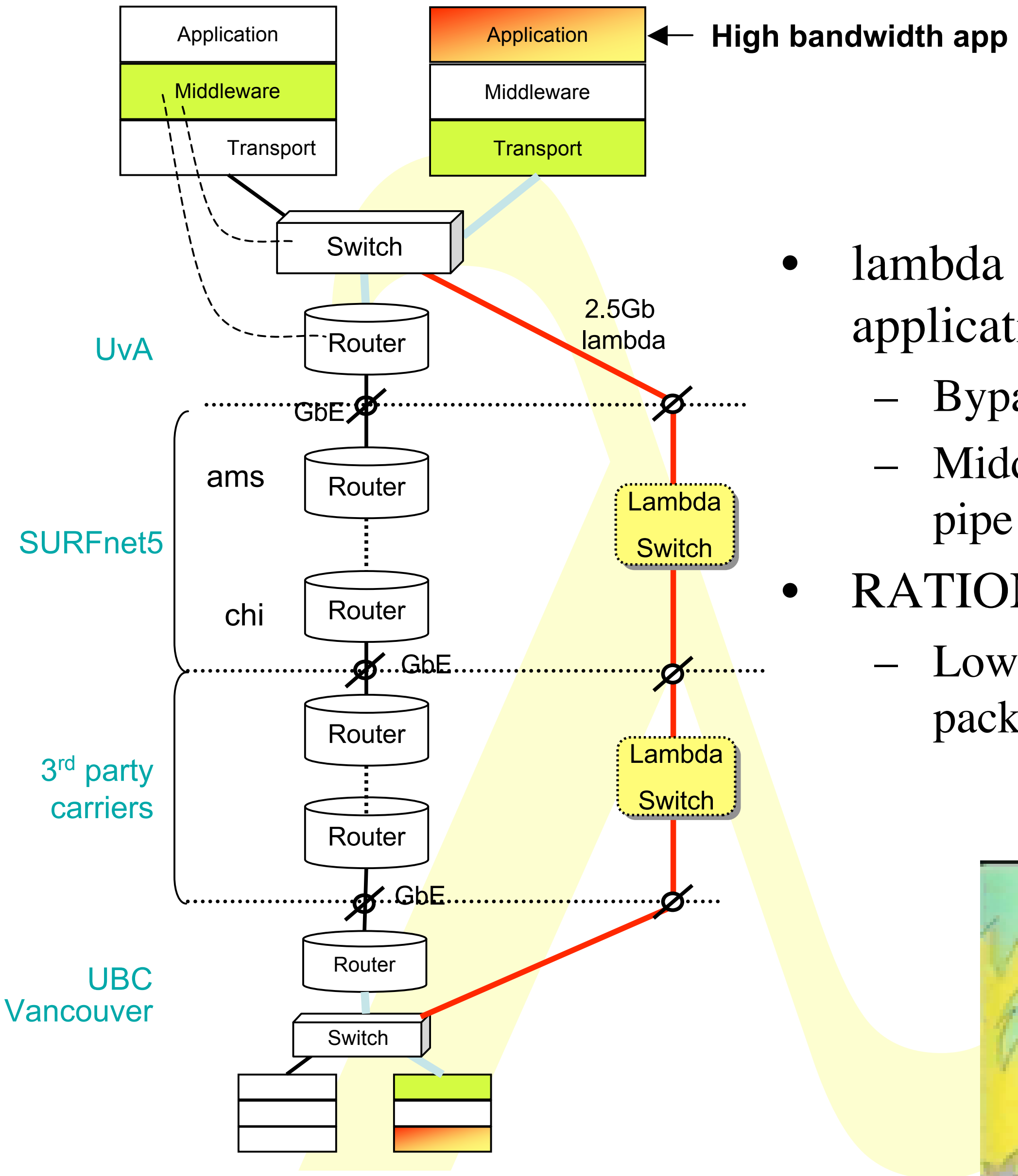
So what are the facts

- **Costs of fat pipes (fibers) are one/third of cost 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 @ 40 Gb/s -> 20 ns to look up in 140 kEntries routing table (light speed from me to you!)**
- **Big sciences need fat pipes**
- **Bottom line: look for a hybrid architecture which serves all users in a cost effective way**

Architectures - L1 - L3

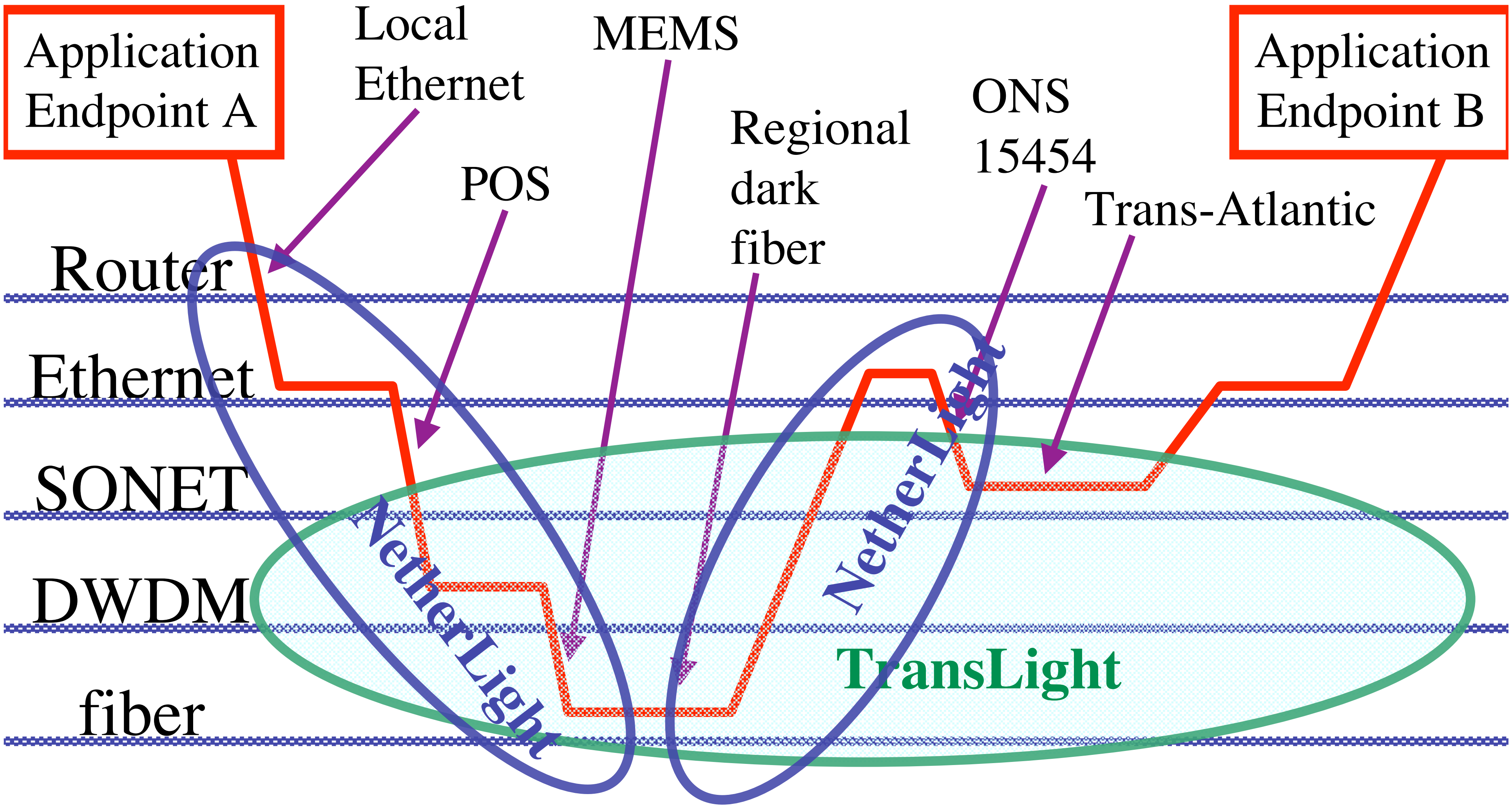
(10b of 13)



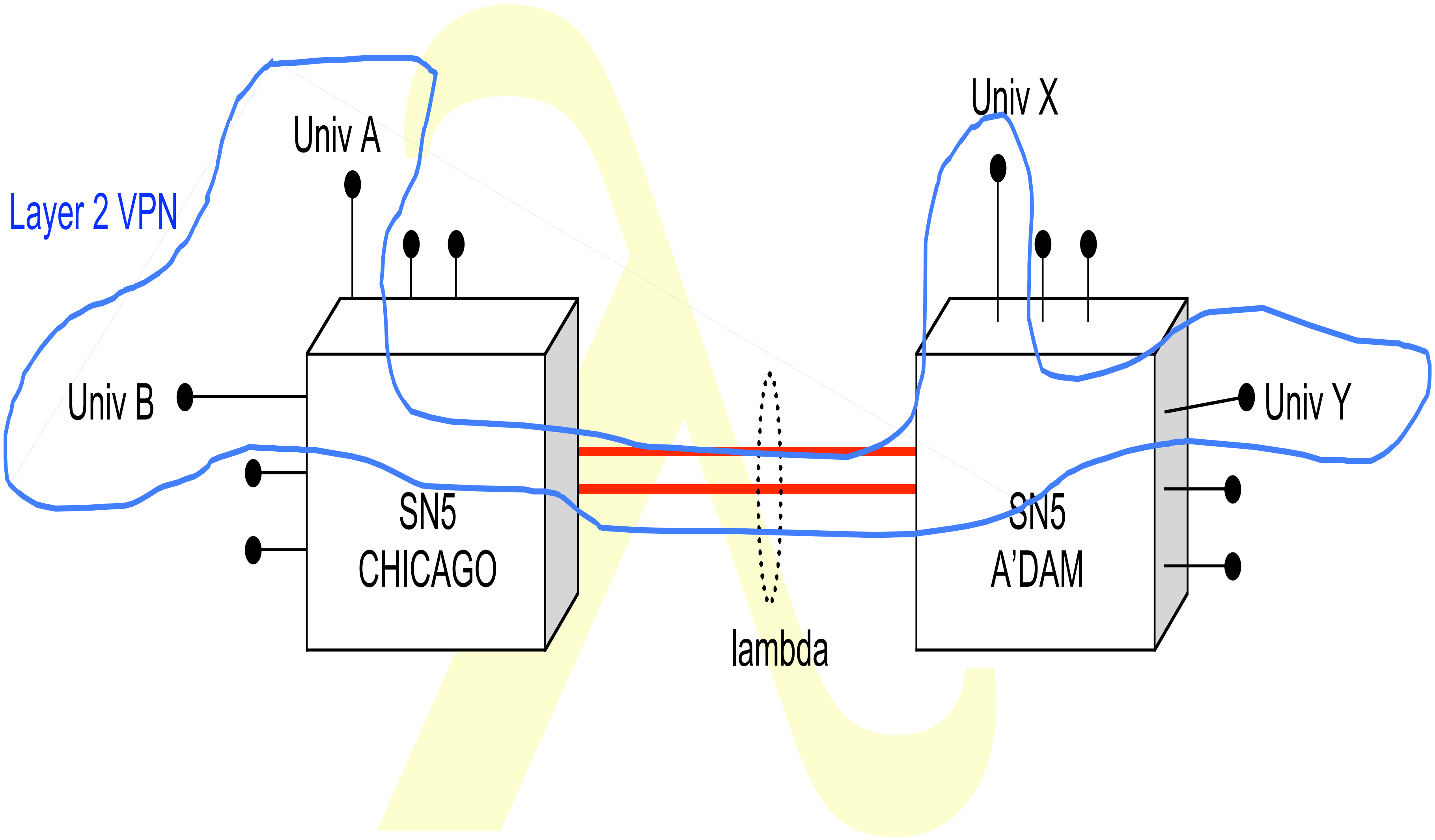


- lambda for high bandwidth applications
  - Bypass of production network
  - Middleware may request (optical) pipe
- RATIONALE:
  - Lower the cost of transport per packet

How low can you go?

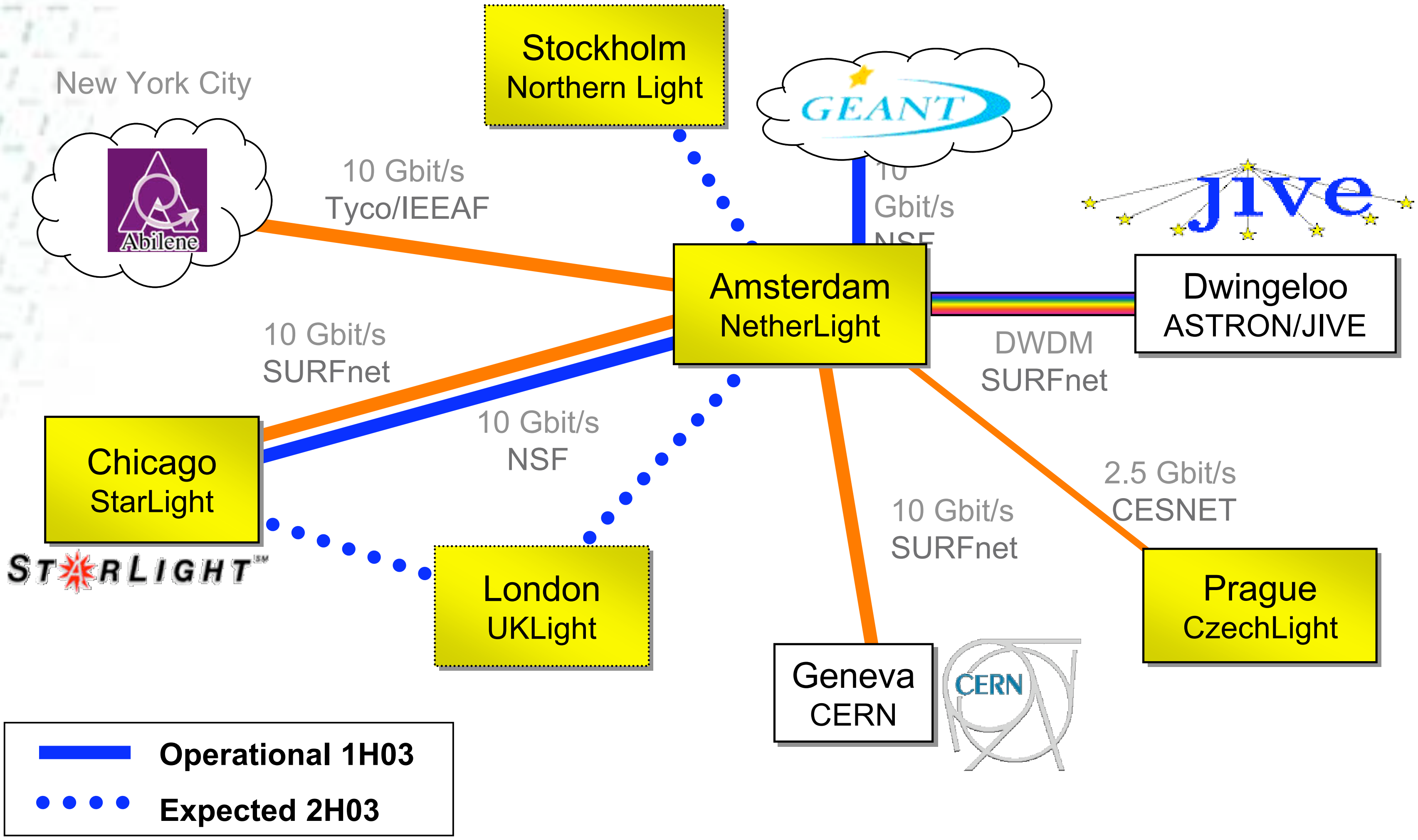


Virtual Organization on L2



NetherLight Network: 2003

Emerging international lambda grid



Stockholm
Northern Light

New York City

Abilene

GEANT

jive

10 Gbit/s
Tyco/IEEAF

10
Gbit/s
NSF

Amsterdam
NetherLight

Dwingeloo
ASTRON/JIVE

DWDM
SURFnet

10 Gbit/s
SURFnet

10 Gbit/s
NSF

Chicago
StarLight

STARLIGHT

London
UKLight

10 Gbit/s
SURFnet

2.5 Gbit/s
CESNET

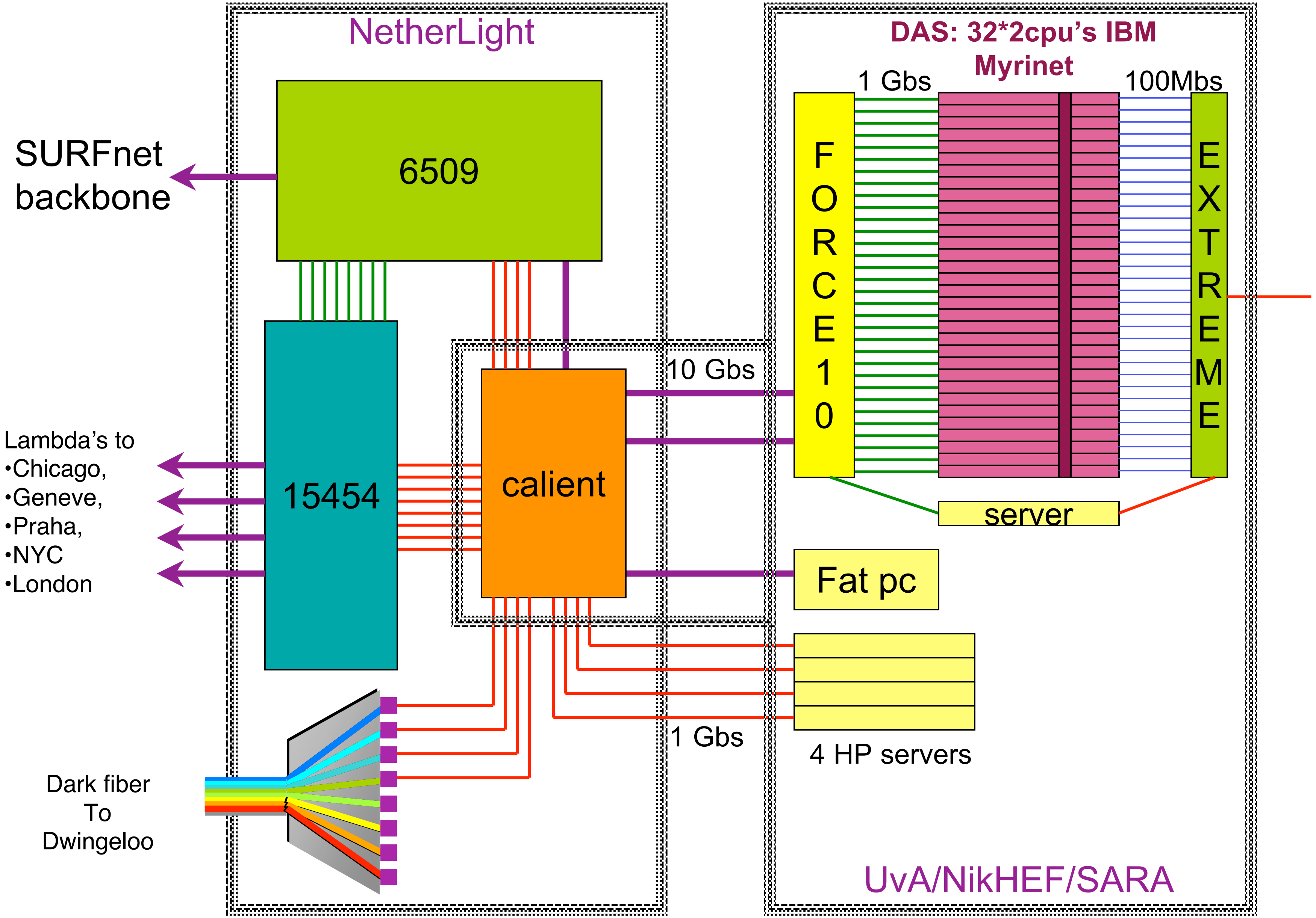
Prague
CzechLight

Geneva
CERN

CERN

▬ **Operational 1H03**

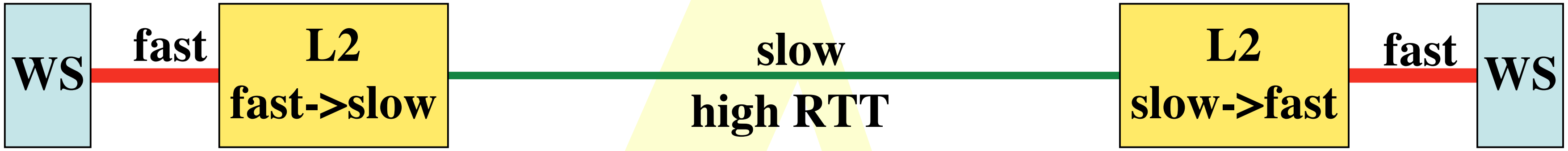
● ● ● ● **Expected 2H03**


NetherLight
6509
SURFnet backbone
15454
Lambda's to
•Chicago,
•Geneve,
•Praha,
•NYC
•London
calient
10 Gbs
1 Gbs
Dark fiber
To
Dwingeloo
DAS: 32*2cpu's IBM
Myrinet
1 Gbs
100Mbs
FORCE10
EXTREME
server
Fat pc
4 HP servers
UvA/NikHEF/SARA

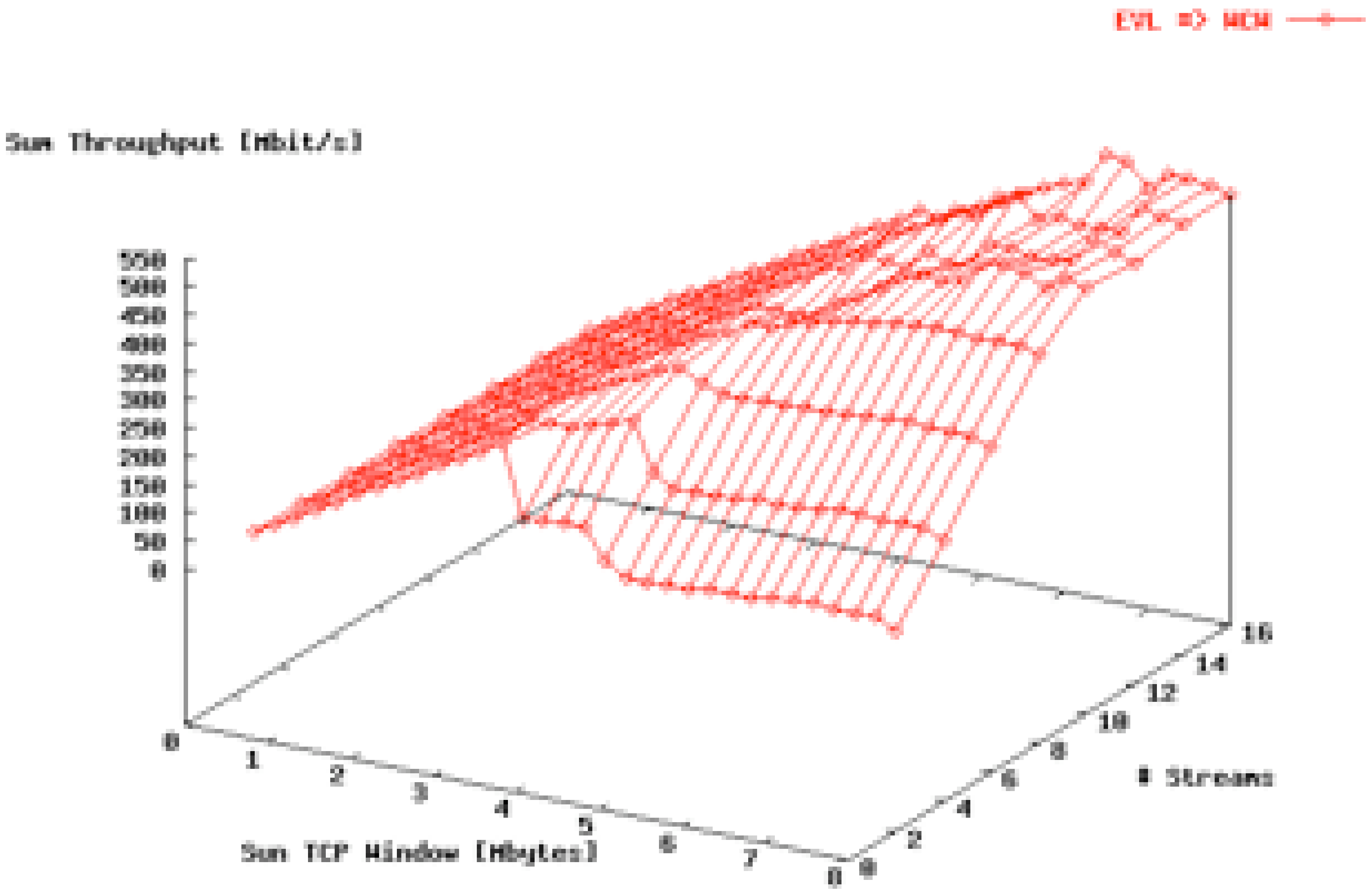
(intermezzo)

NetherLight

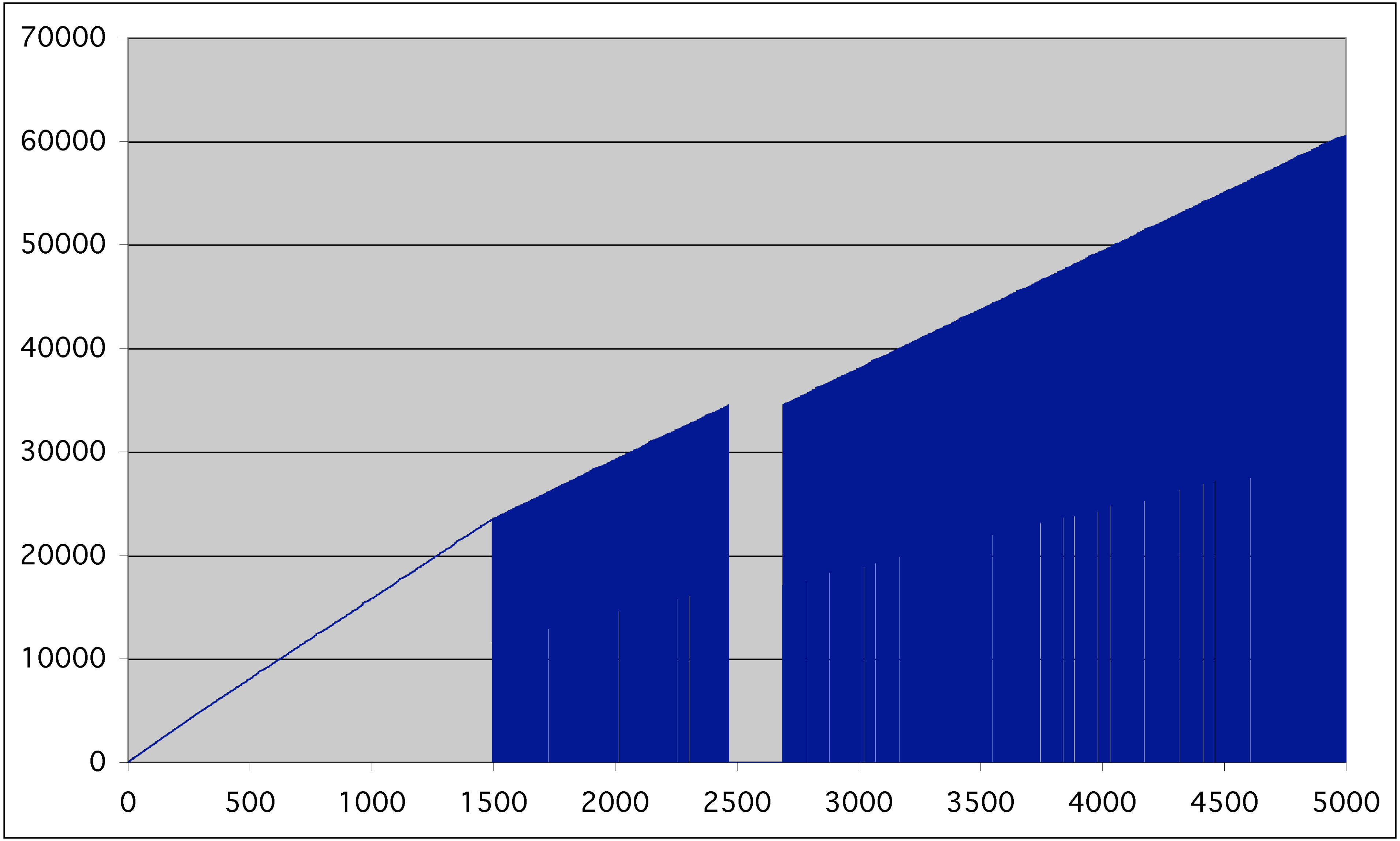
Early Lambda/LightPath TDM experiences



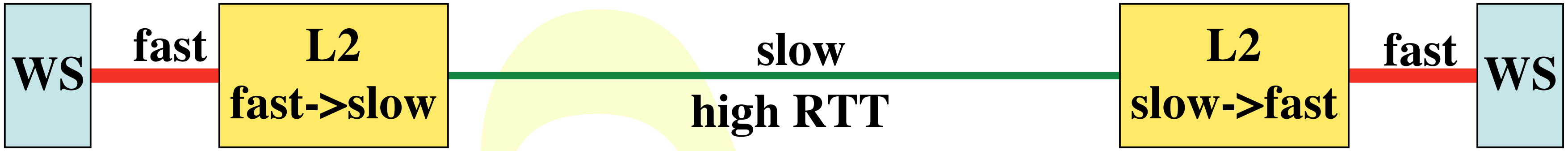
(17b of 20)


EVL => NCN
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

5000 1 kByte UDP packets



Layer - 2 requirements from 3/4

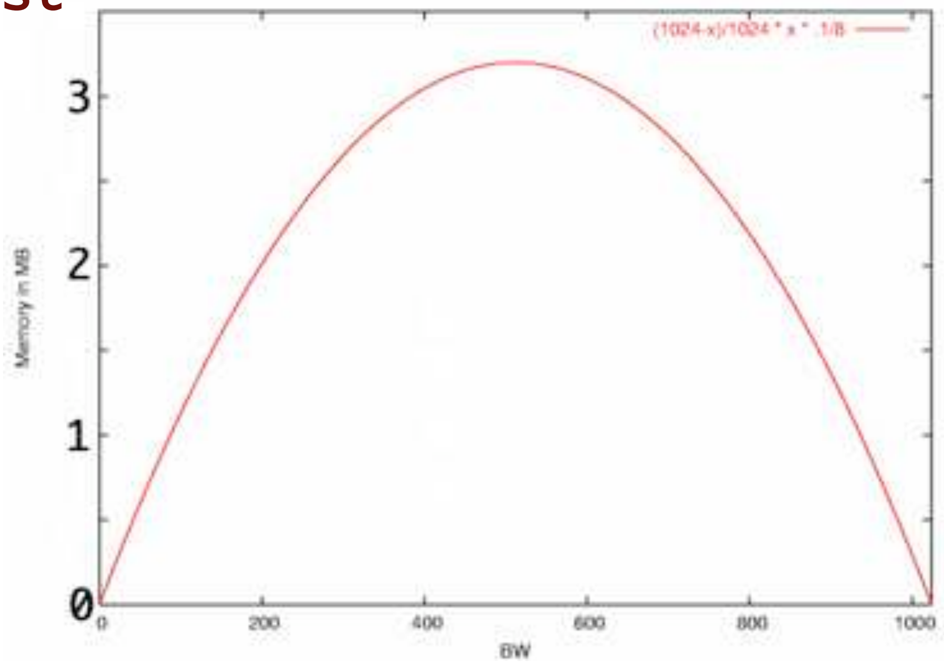


TCP is bursty due to sliding window protocol and slow start algorithm.

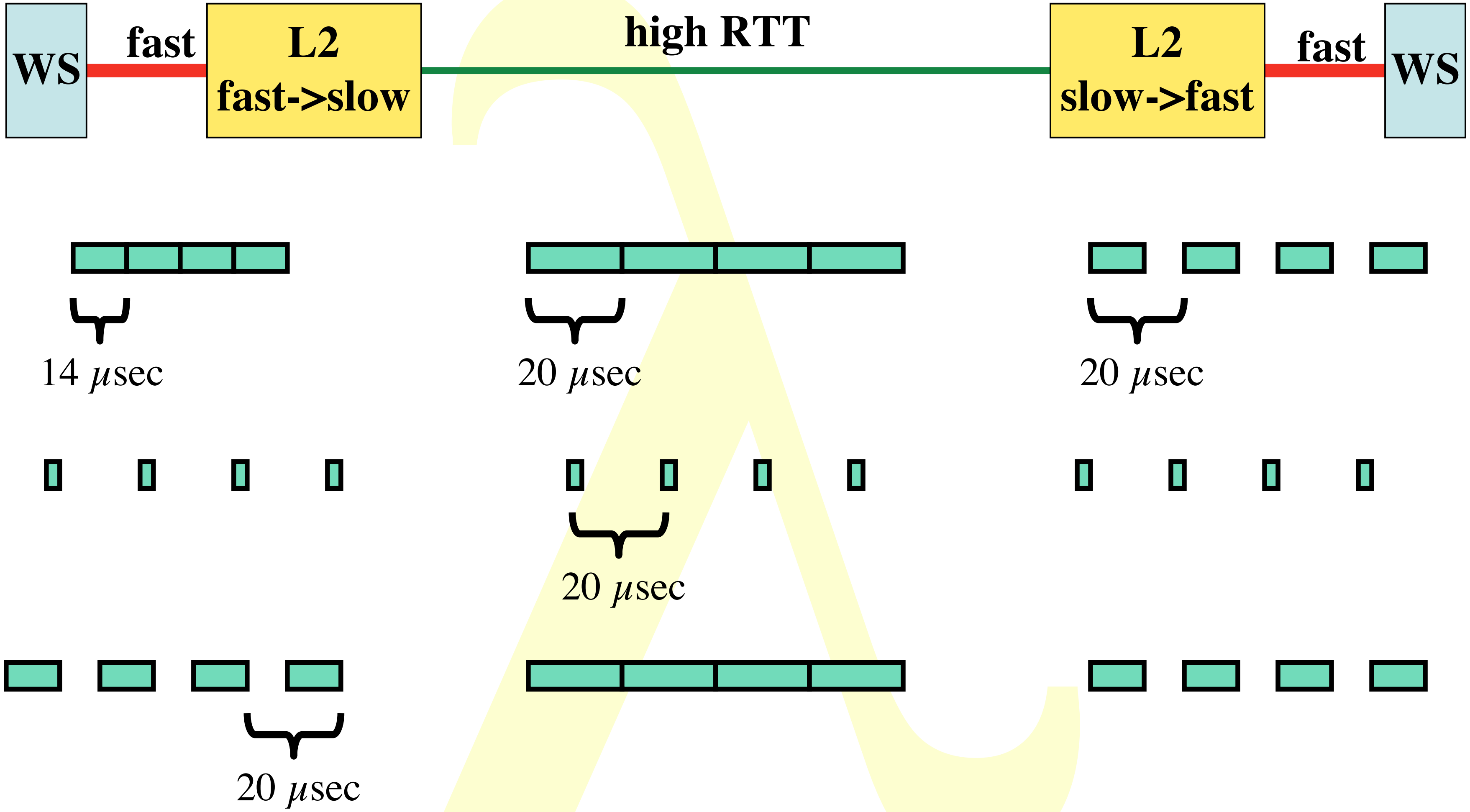
```
Window = BandWidth * RTT     &    BW == slow

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

So pick from menu:

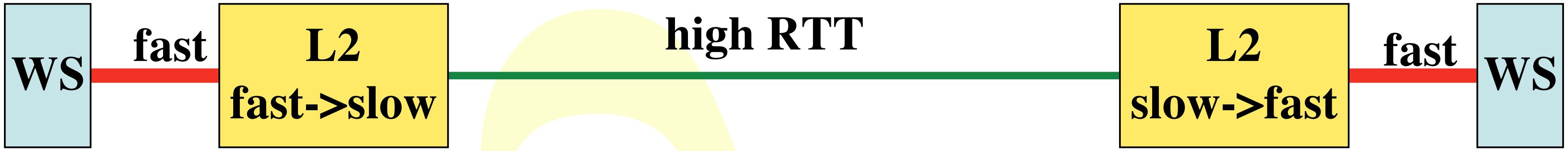
- *Flow control*
- *Traffic Shaping*
- *RED (Random Early Discard)*
- *Self clocking in TCP*
- *Deep memory*



Self-clocking of TCP



Layer - 2 requirements from 3/4



Window = BandWidth * RTT & BW == slow

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

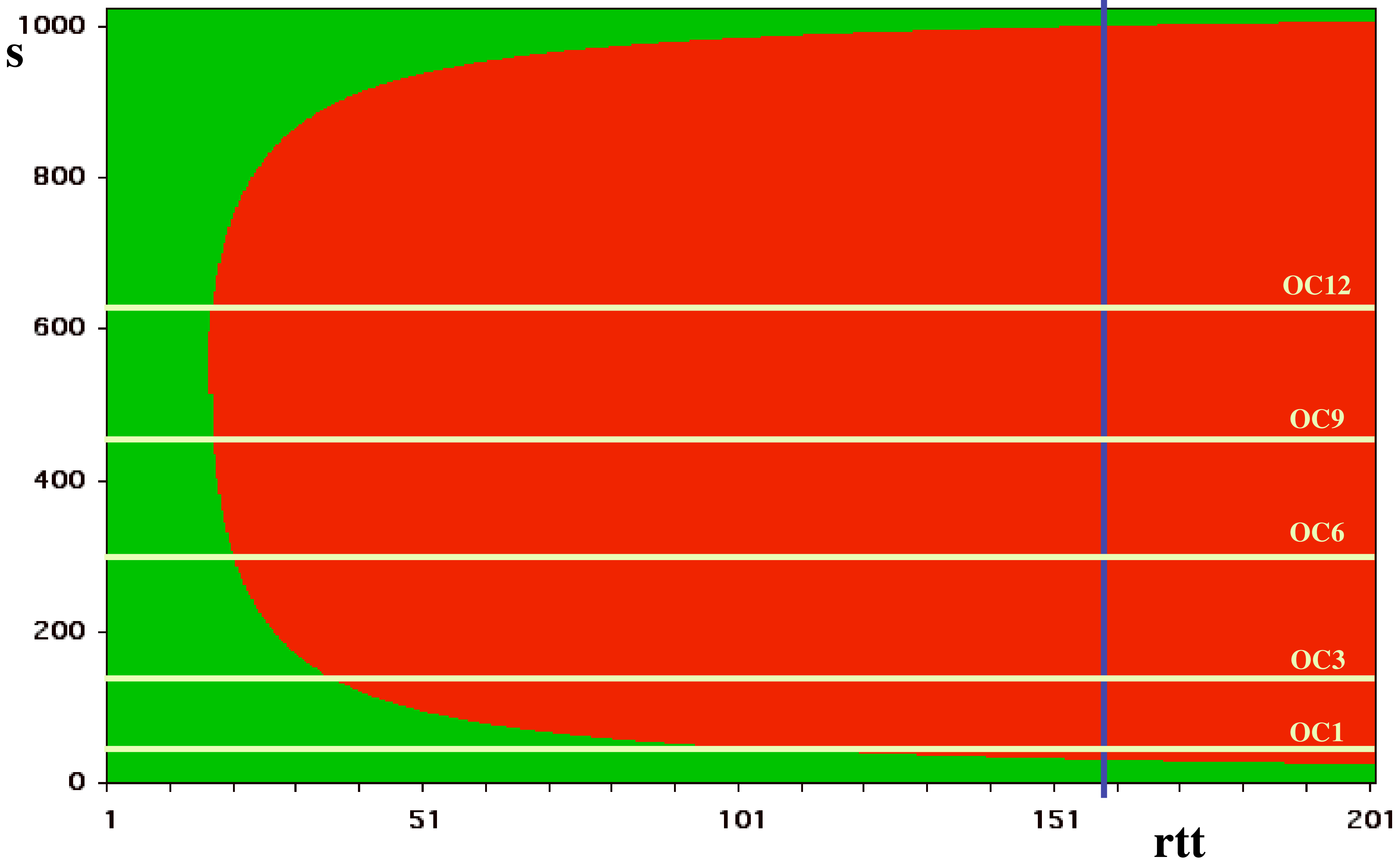
Given M and f, solve for slow ===>

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

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

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

158 ms = RTT Amsterdam - Vancouver



(19d of 20)

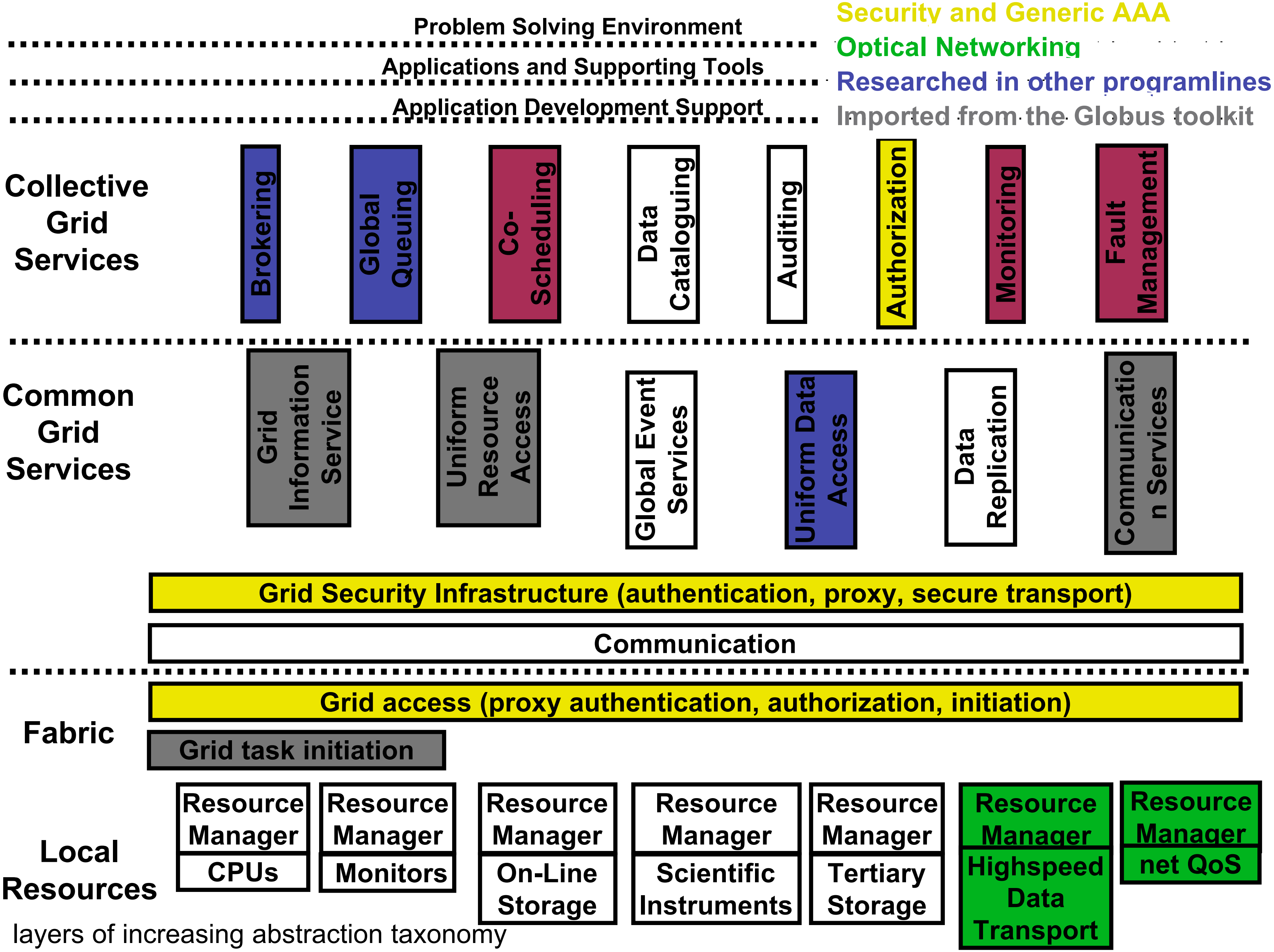
High performance computing and Processor memory co-allocation

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

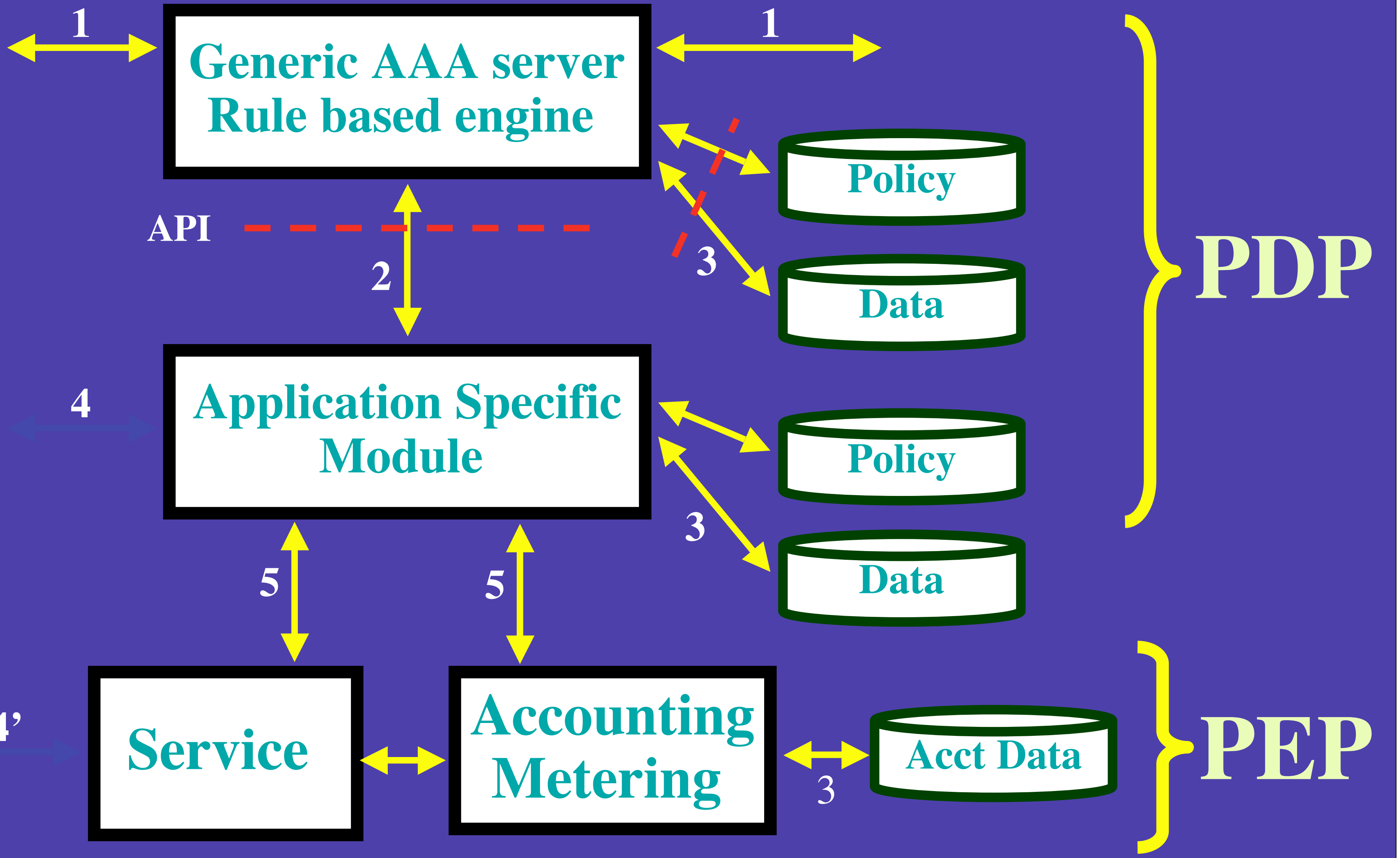
Resource Manager Tertiary Storage

Resource Manager Highspeed Data Transport

Resource Manager net QoS

layers of increasing abstraction taxonomy

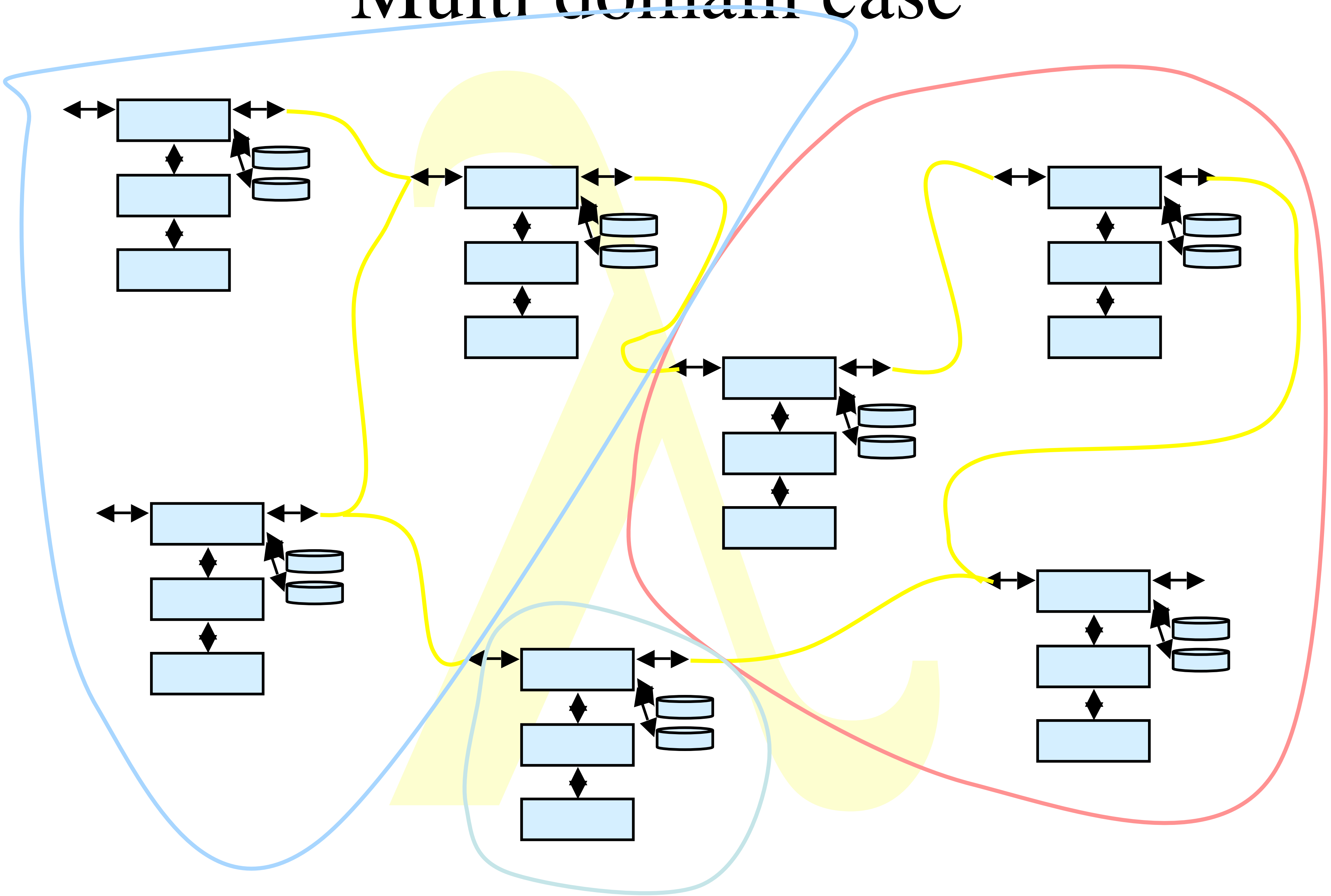
Starting point



RFC 2903 - 2906 , 3334 , policy draft

(19f of 20)

Multi domain case



(Future) Projects

- National:
  - NCF Grid project
  - VLE
  - GigaPort-NG
  - LOFAR
- European
  - DataGrid
  - DataTAG
- International
  - NetherLight
  - StarLight
  - AnyLight, LowLight, BackLight
  - Optiputer

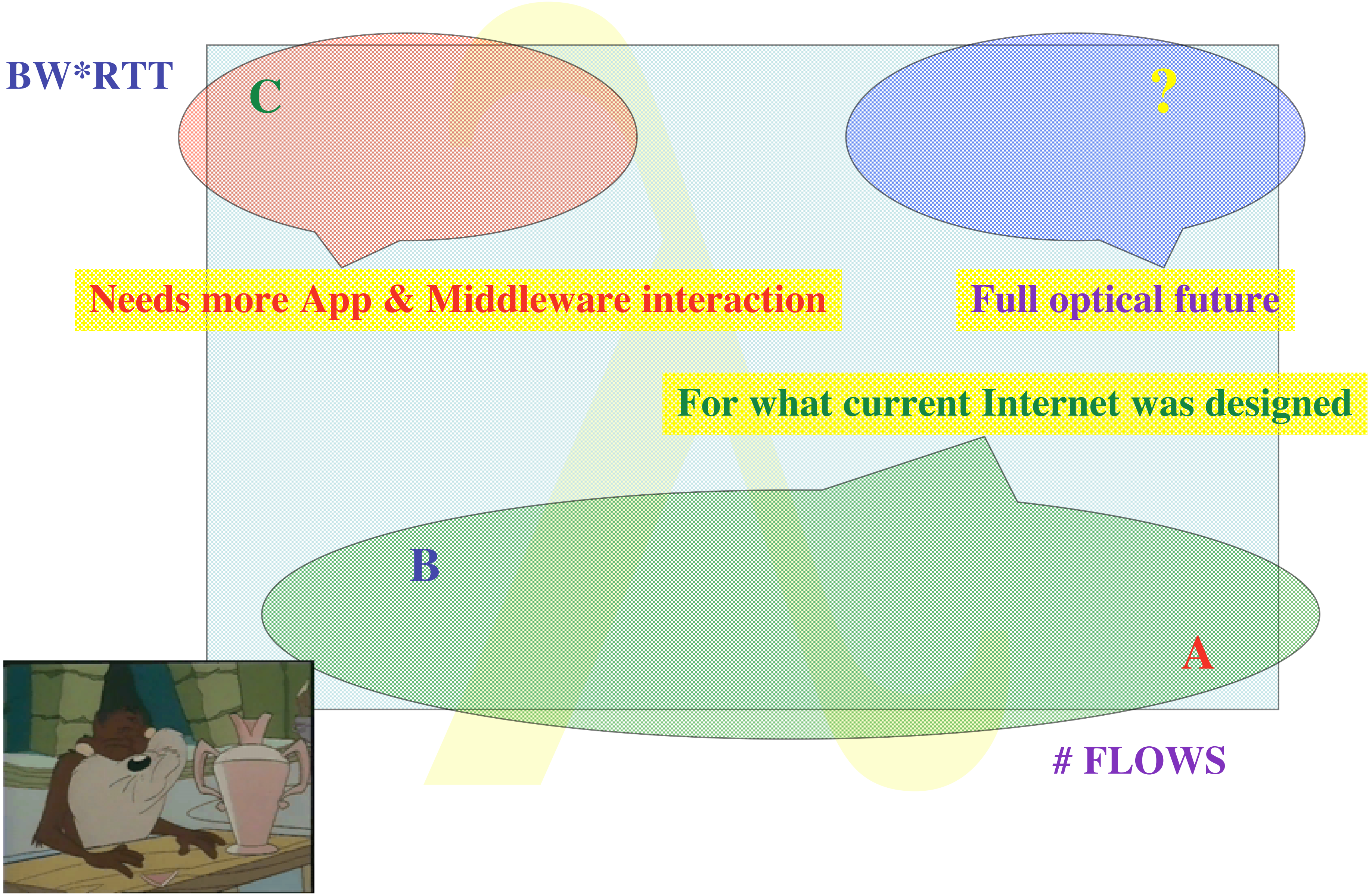
Research:

Models of Lambda networking

Transport

AAA

Transport in the corners



BW*RTT

C

?

Needs more App & Middleware interaction

Full optical future

For what current Internet was designed

B

A

# FLOWS

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