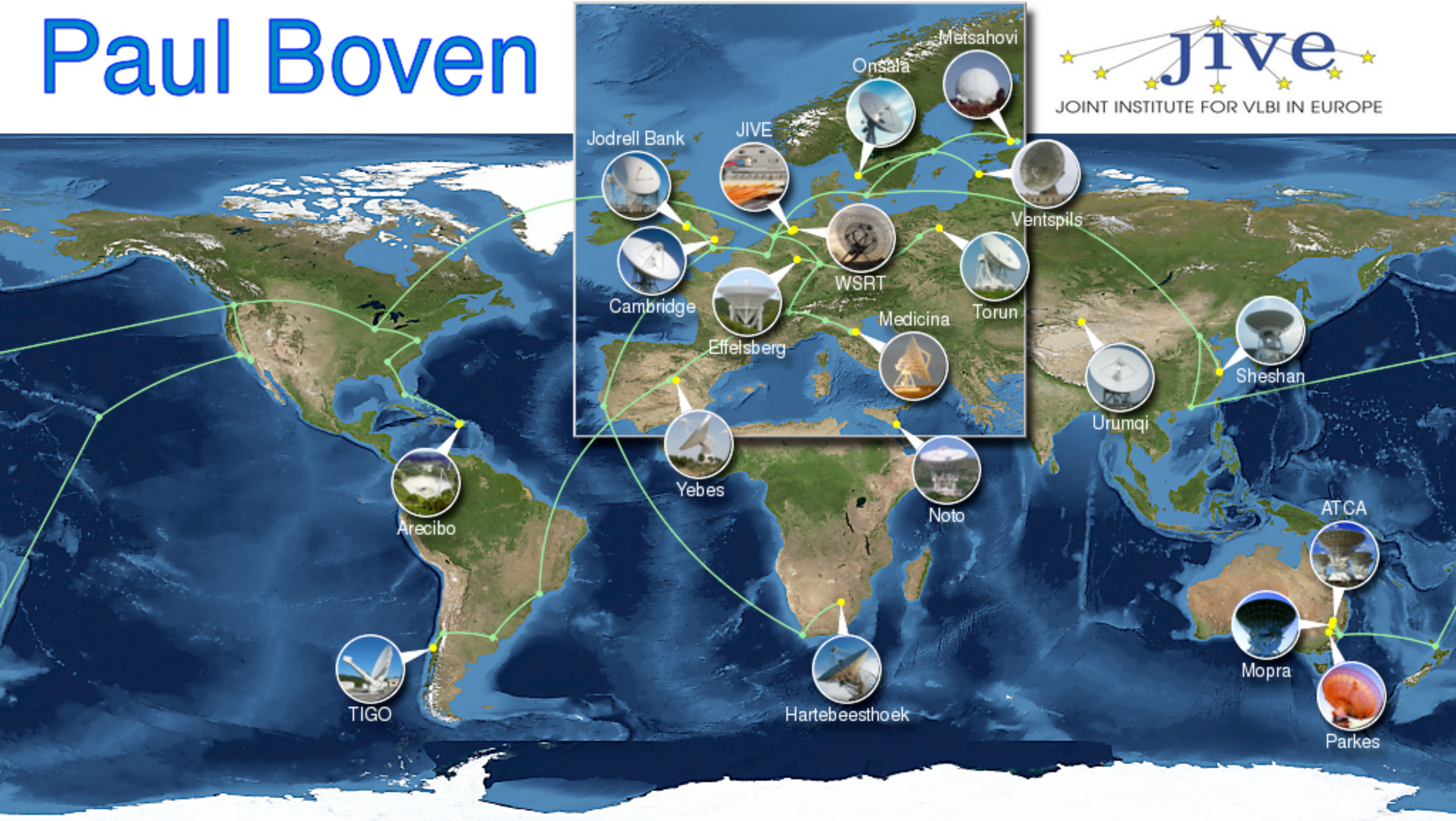
e-VLBI beyond the 1Gb/s speedbump

Paul Boven

JIVE
JOINT INSTITUTE FOR VLBI IN EUROPE



Network status as per 2008-05-02. Image created by Paul Boven <boven@jive.nl>. Satellite image: Blue Marble Next Generation, courtesy of Nasa Visible Earth (visibleearth.nasa.gov).

Introduction

- ★ Sensitivity ≈ $\sqrt{}$Bandwidth, nbr of Telescopes
- ★ Resolution ≈ Distance
- ★ Observations >12h
- ★ Production rate is 512Mb/s per telescope
- ★ Current EVN correlator capacity is 16x 1024Mb/s

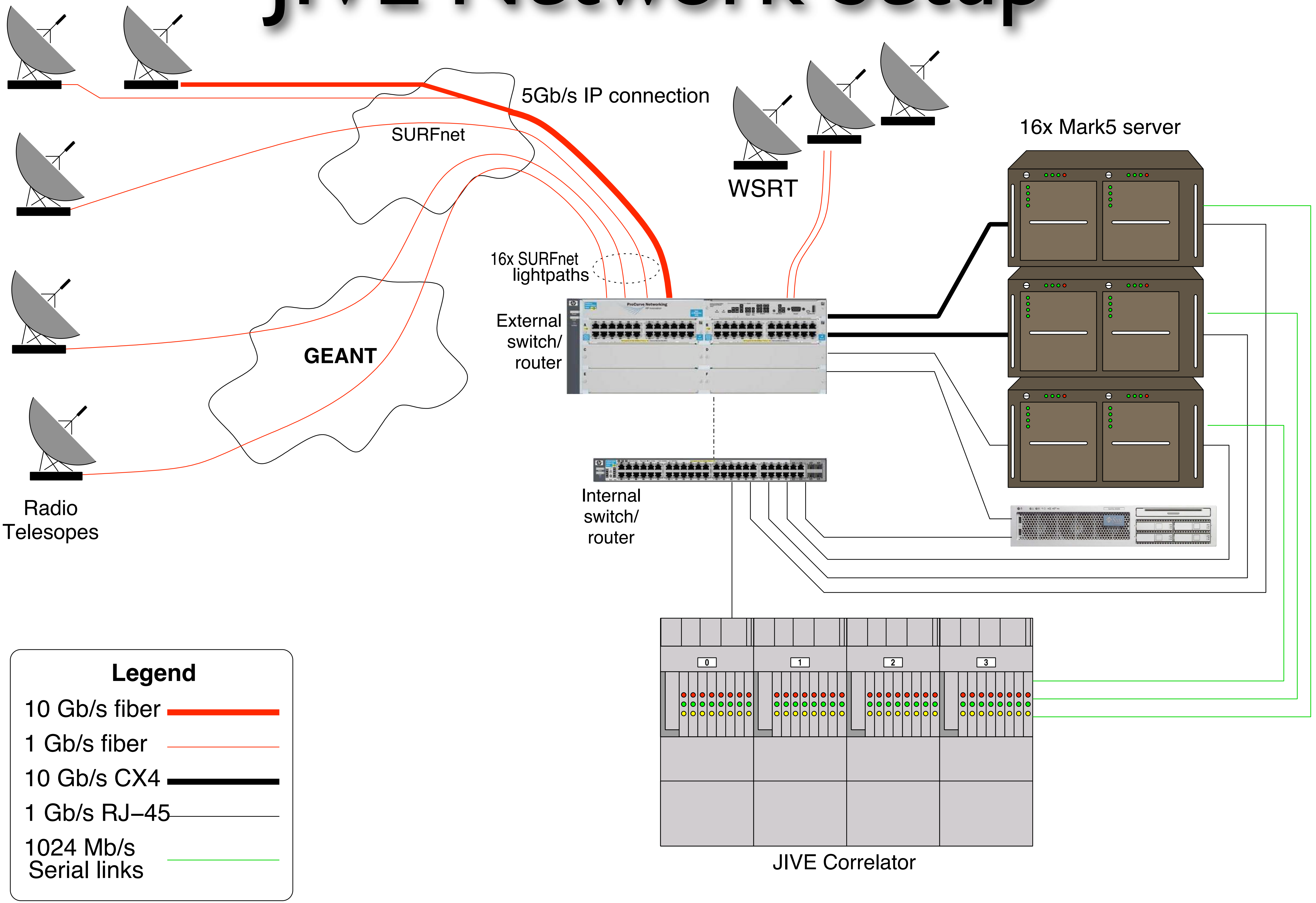
Network Overview (1)

Telescope	Bandwidth	RTT
Sheshan	512 + 622 LP	180ms / 354ms
ATNF (3)	2x 1Gb/s LP	343ms
Hartebeesthoek	64Mb/s SAT-3	181ms
Arecibo	512Mb/s VLAN	154ms
TIGO	95Mb/s	150ms
Metsahovi	10Gb/s (?)	
Torun	1Gb/s LP	34.9ms

Network Overview (2)

Telescope	Bandwidth	RTT
Onsala	1Gb/s routed	34.2ms
Medicina	1Gb/s LP	29.7ms
Jodrell Bank	1Gb/s LP	18.6ms
Cambridge/Merlin	Each 128Mb/s	16.9ms
Effelsberg	1Gb/s routed	13.5ms
WSRT	2x 1Gb/s CWDM	0.57ms
Yebes	Under construction	

JIVE Network Setup



5Gb/s IP connection

SURFnet

16x SURFnet lightpaths

GEANT

Radio Telesopes

WSRT

16x Mark5 server

External switch/ router

Internal switch/ router

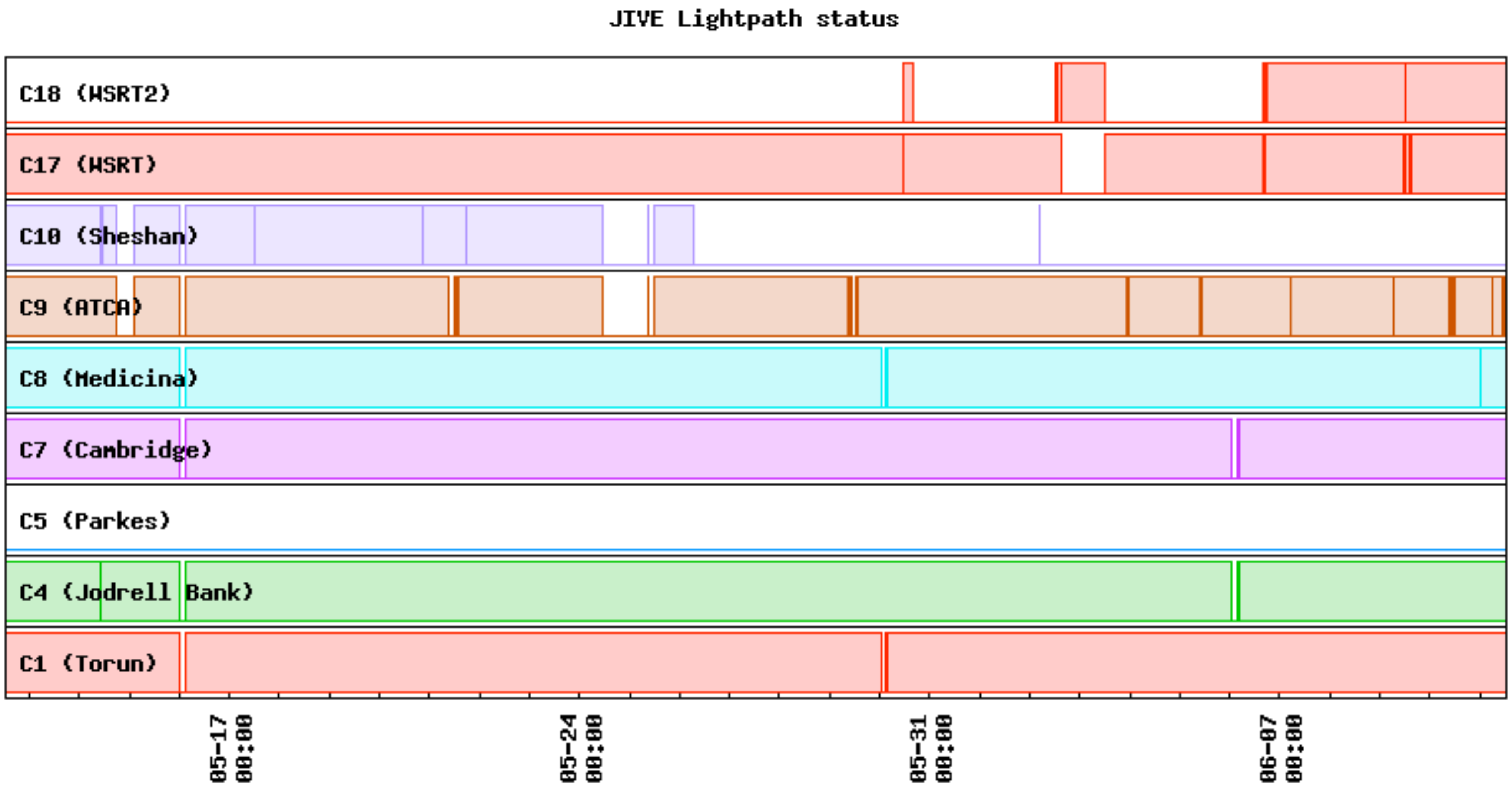
JIVE Correlator

Legend

10 Gb/s fiber	▬
1 Gb/s fiber	─
10 Gb/s CX4	▬
1 Gb/s RJ–45	─
1024 Mb/s Serial links	─

Lightpaths

- Dedicated point-to-point circuit
- Based on SDH/Sonet timeslots (NOT a lambda)
- Stitched together at cross-connects
- Guaranteed bandwidth
- But also: a string of Single Points of Failure



The 1Gb/s speedbump

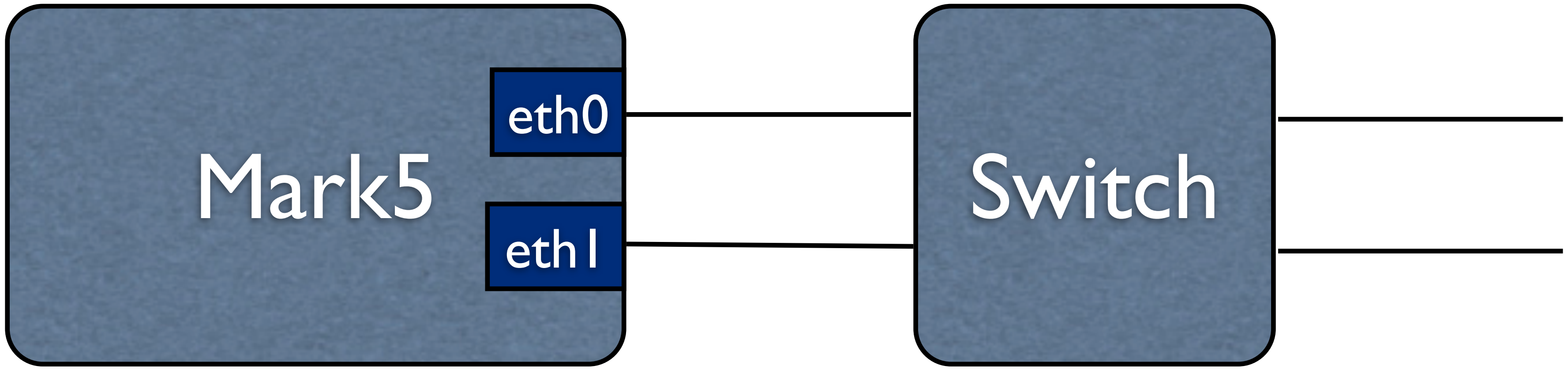
- VLBI (tape based) comes in fixed speeds, power of 2:
  64 Mb/s,128Mb/s, 256Mb/s, 512Mb/s - and 1024Mb/s
- Ethernet comes in 10, 100, 1000 and 10000 Mb/s.
- 1024Mb/s > 1Gb/s (with headers it's more like 1030)
- Dropping packets works but is sub-optimal
- Dropping 'tracks' to <1Gb/s: Takes a LOT of CPU work

Trunking

- Use two channels instead of upgrading link to 10Gb/s

- Each link carries apx. 515Mb/s
  e.g. two 622Mb/s lightpaths

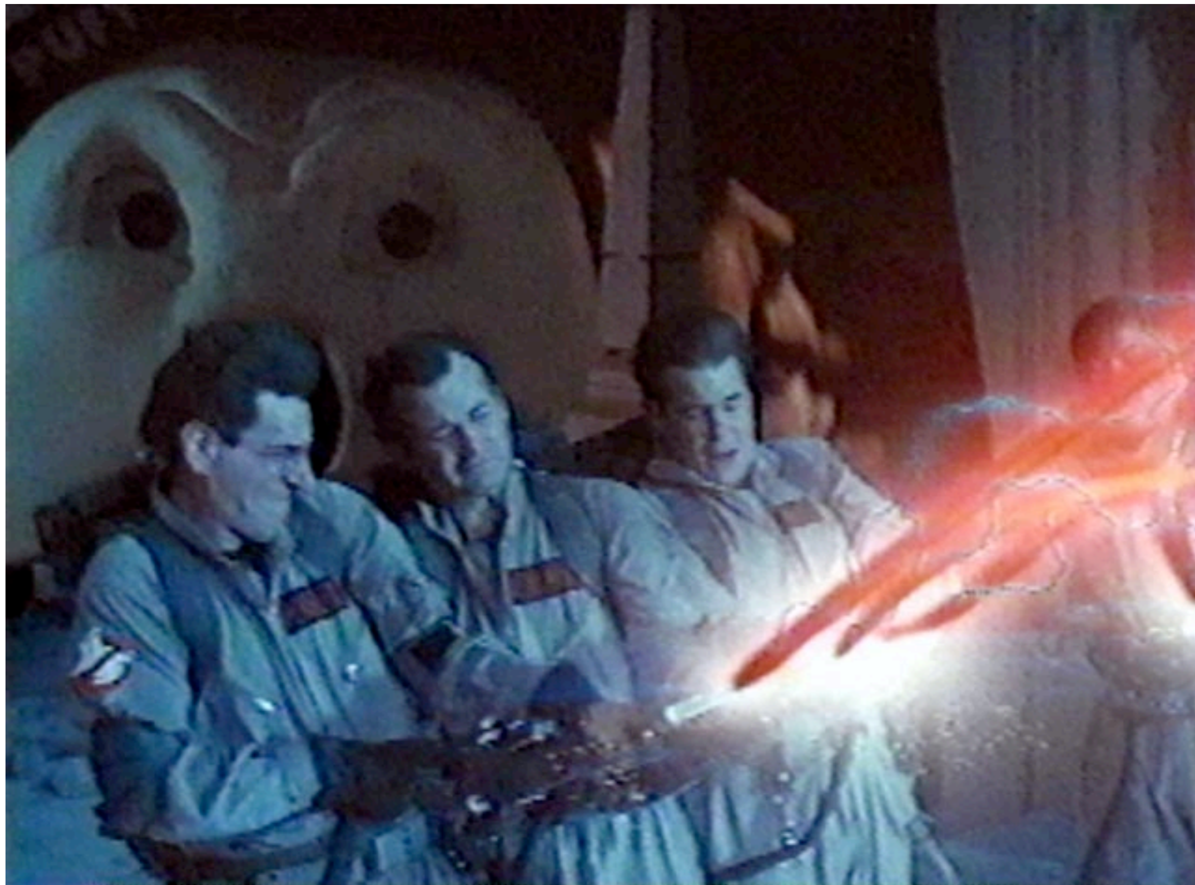
- Two ethernet interfaces in Mark5



The Trouble with Trunking

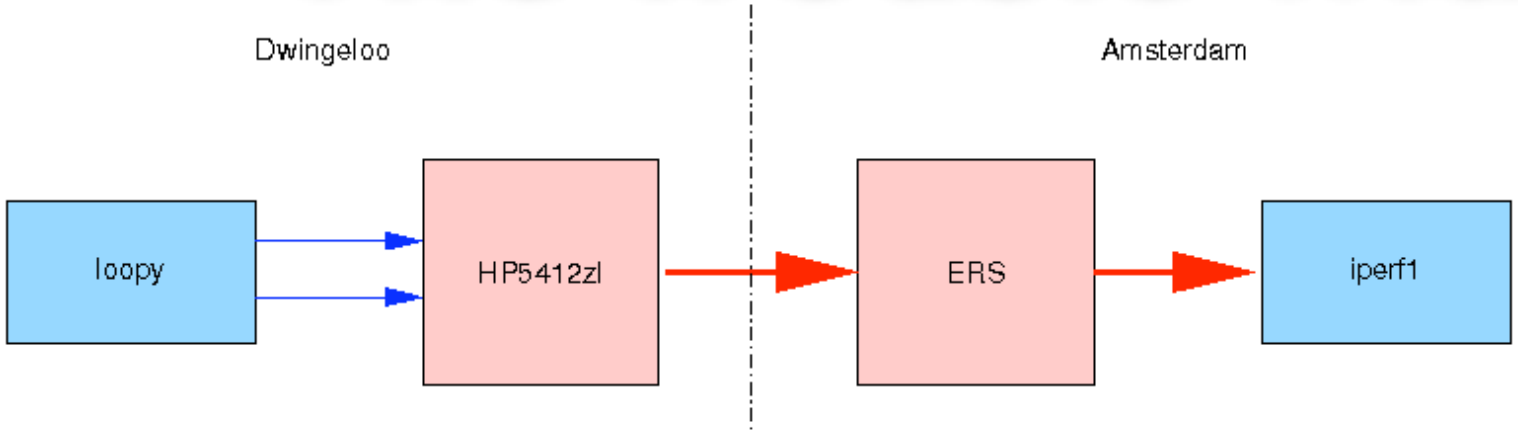
- Standard trunking: LACP (802.3ad)
  - Uses a hash of source/destination MAC, IP and/or Port to choose outgoing port
  - This is to prevent re-ordering
  - A single TCP/UDP stream will use only 1 link member!

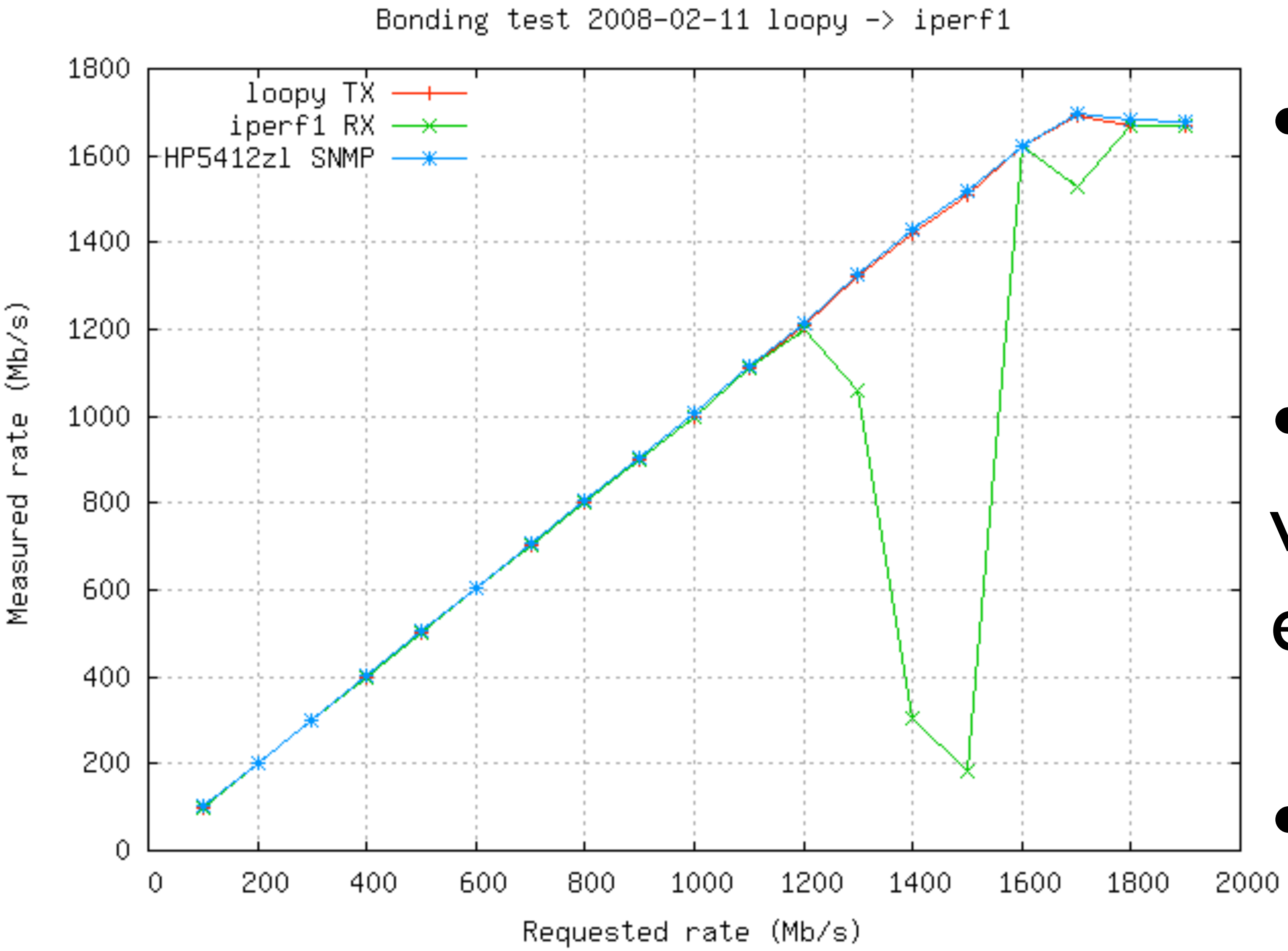
- Linux kernels come with bonding, 'ifenslave'
  - Round Robin traffic distribution
  - Keep both halves in separate VLANS/Lightpaths all the way as switches in between only speak LACP



"Do NOT cross the streams!"

The Trouble with Trunking

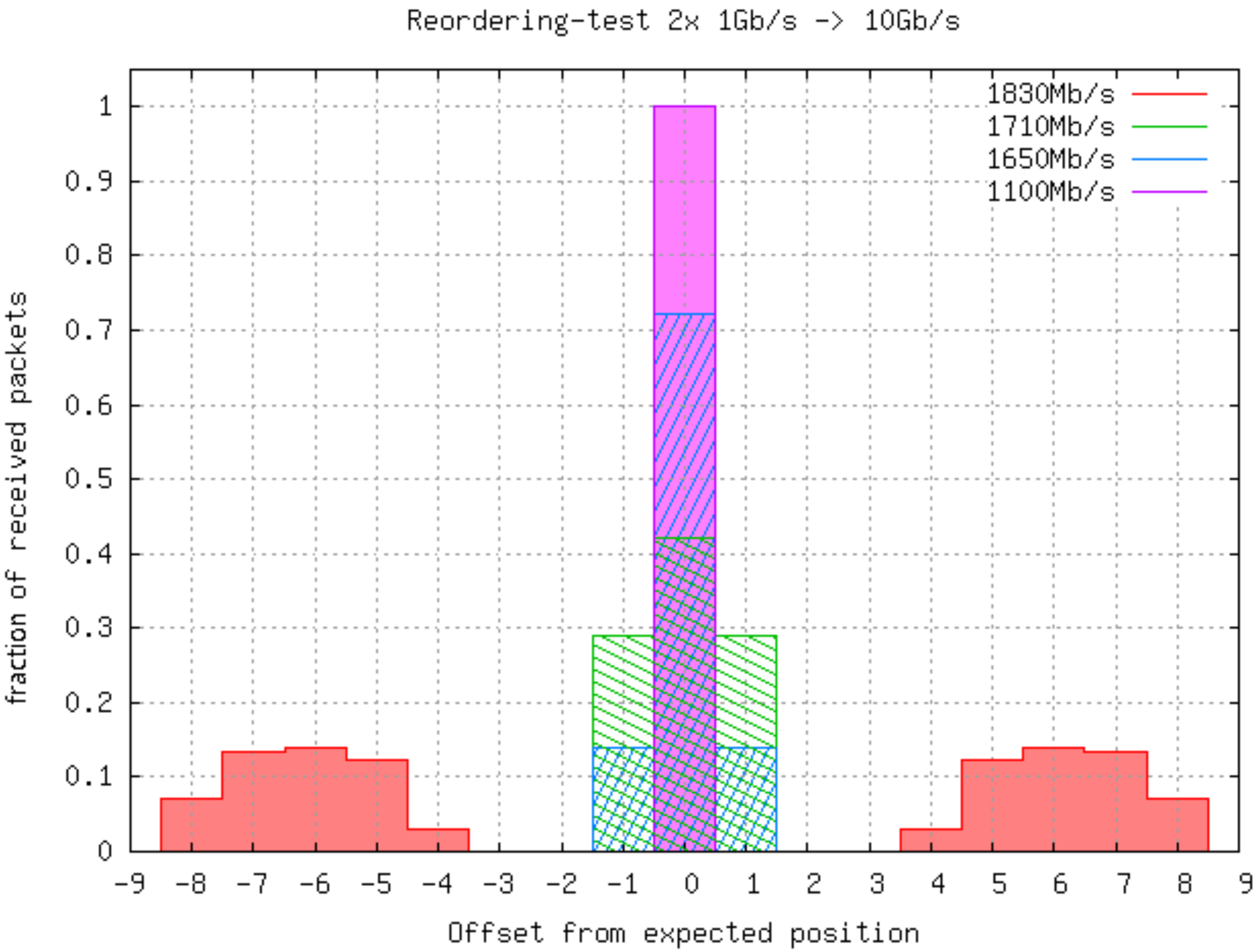




- iperf test using UDP
- Bonding driver
- To SURFnet, JIVE didn’t have 10G yet.
- Conclusion: bonding works well enough for e-VLBI (1024Mb/s)
- But not as good as expected?

No Trouble with Trunking!

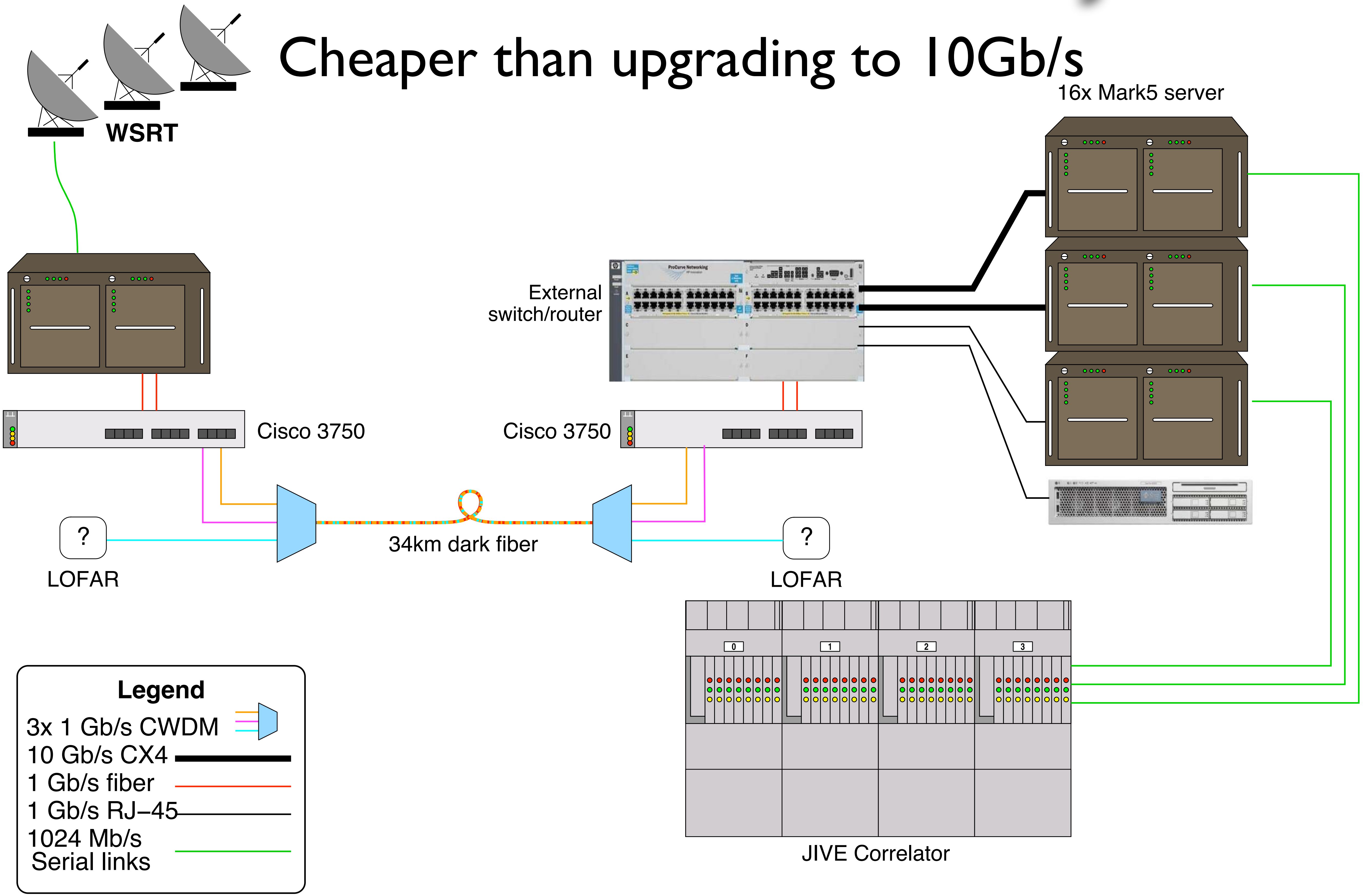
- iperf gets really confused by re-ordering of packets
- Wrote a simple re-implementation for UDP
- Store S/N in memory to track re-ordering, post-process



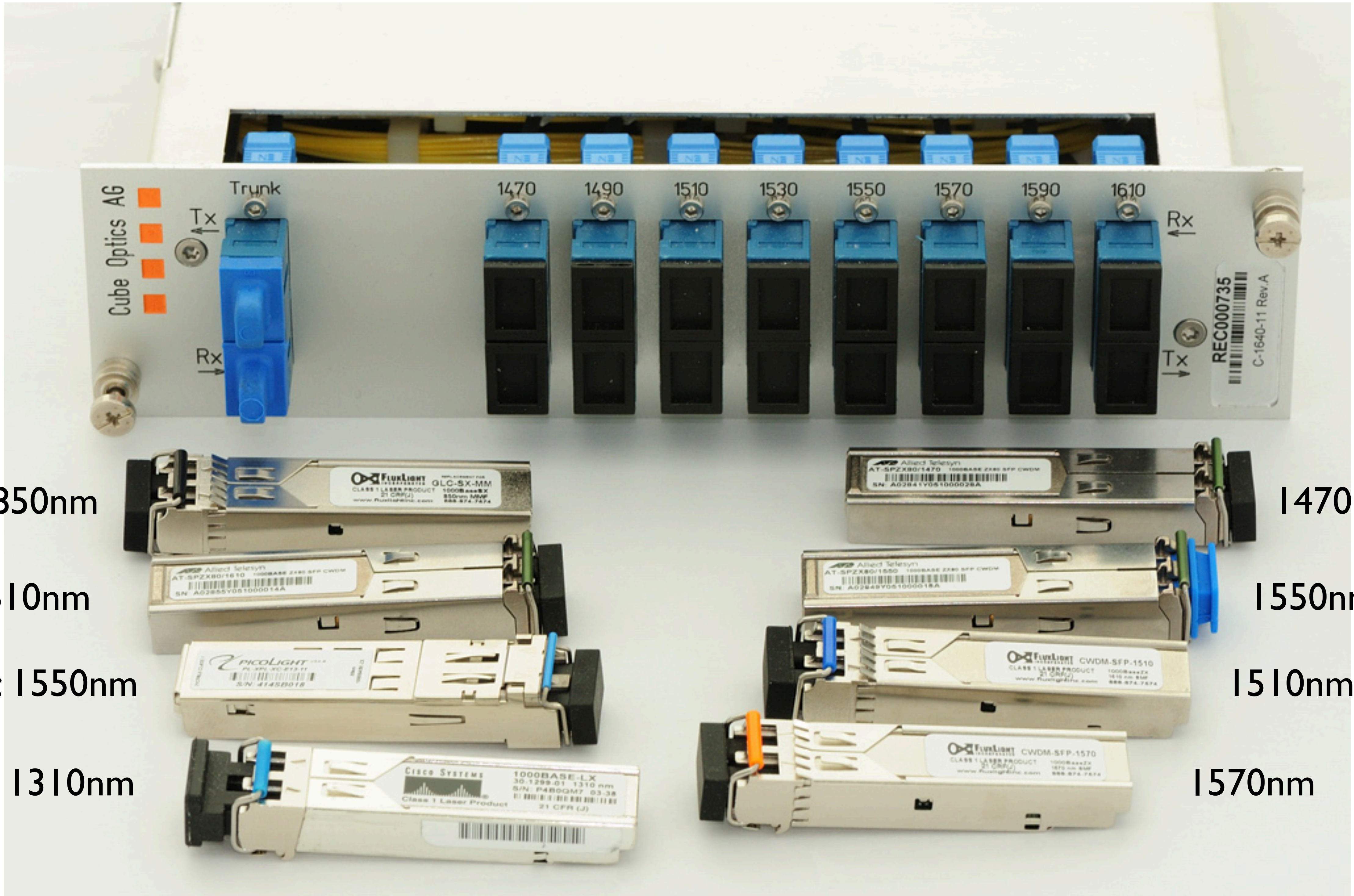
- No packet loss even at 1830Mb/s
- No re-ordering below 1100Mb/s
- Little re-ordering below 1710Mb/s

CWDM from WSRT to JIVE

Cheaper than upgrading to 10Gb/s



All the colours of the rainbow...



SX: 850nm

1470nm

1610nm

1550nm

ZX: 1550nm

1510nm

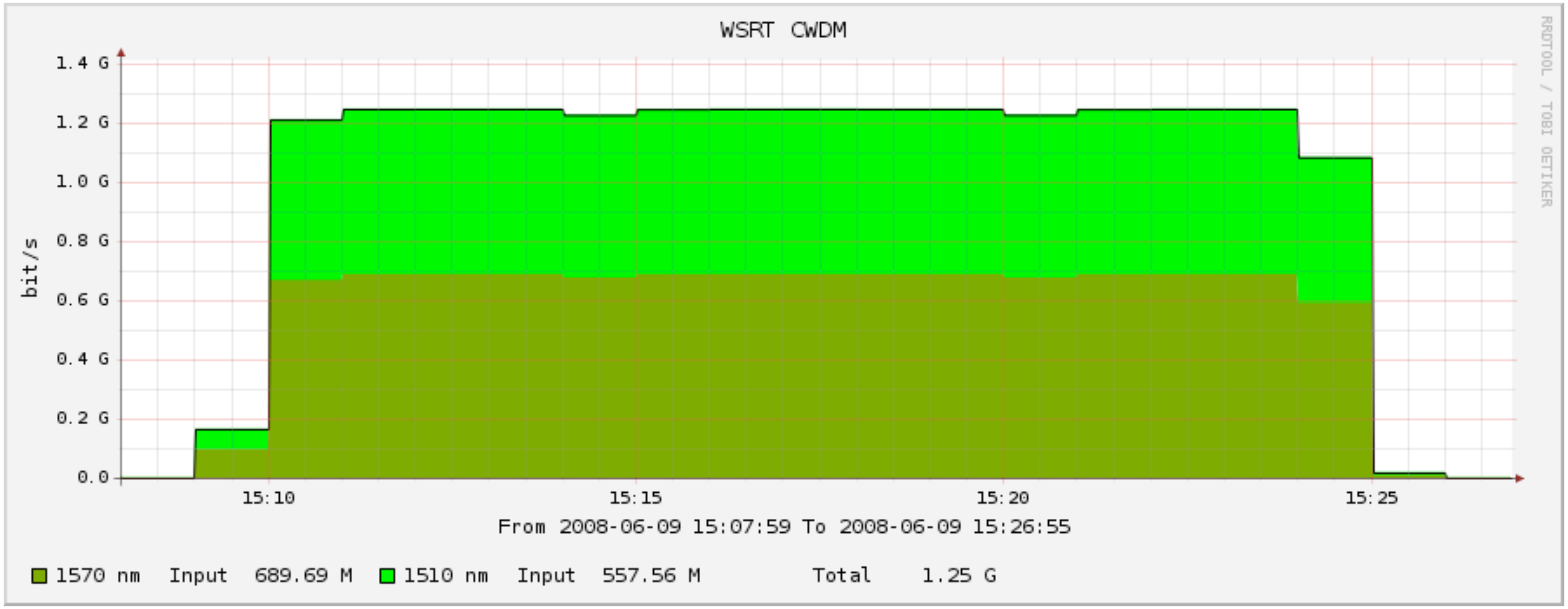
LX: 1310nm

1570nm

... and then some

1200Mb/s from WSRT to JIVE

- Requested 1200Mb/s
- Each interface carries apx. 600Mb/s
- Currently one-way



Putting it all together

- Per telescope: Trunk or 10Gb/s connection to JIVE switch/router

- Several 1024Mb/s links on our 10Gb/s to SURFnet

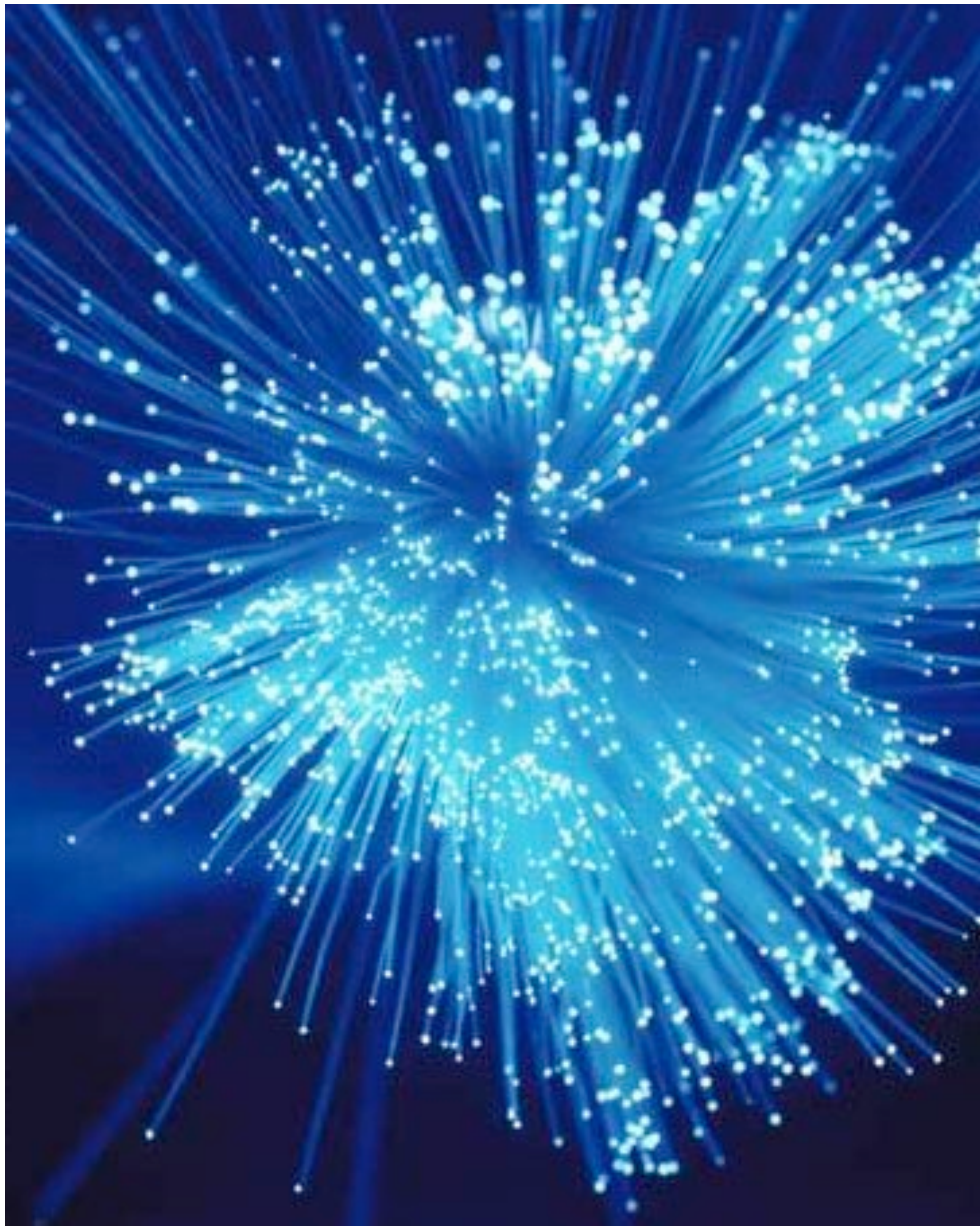
- Up to 16x 10Gb/s ethernet copper (CX4 or 10Gbase-T) on JIVE switch/router to JIVE Mark5's

- 10Gb/s ethernet in JIVE Mark5's
  - This requires a recent kernel (Debian Etch)
    - Which requires SDK8.1

- Coming soon: formatter tests, then fringe tests

An 1024Mb/s e-VLBI sub-network

- WSRT: 2 × 1Gb/s CWDM
- Onsala: 10Gb/s switched LP through NORDUnet (partly shared with e-LOFAR)
- Effelsberg: 10Gb/s VLAN (partly shared with e-LOFAR)
- Jodrell Bank: 2x 1Gb/s LP (plus N × 128Mb/s for Merlin)
- Please join!



Questions ?