e-VLBI Networking Challenges

Paul Boven

JOINT INSTITUTE FOR VLBI IN EUROPE

Network status as per 2008-03-31. Image created by Paul Boven <boven@jive.nl>. Satellite image: Blue Marble Next Generation, courtesy of Nasa Visible Earth (visibleearth.nasa.gov).

What is JIVE?

Operate the EVN correlator and support astronomers doing VLBI.

A collaboration of the major radio-astronomical research facilities in Europe, China and South Africa

A 3 year program to create a distributed astronomical instrument of inter-continental dimensions using e-VLBI, connecting up to 16 radio telescopes

Radio Astronomy

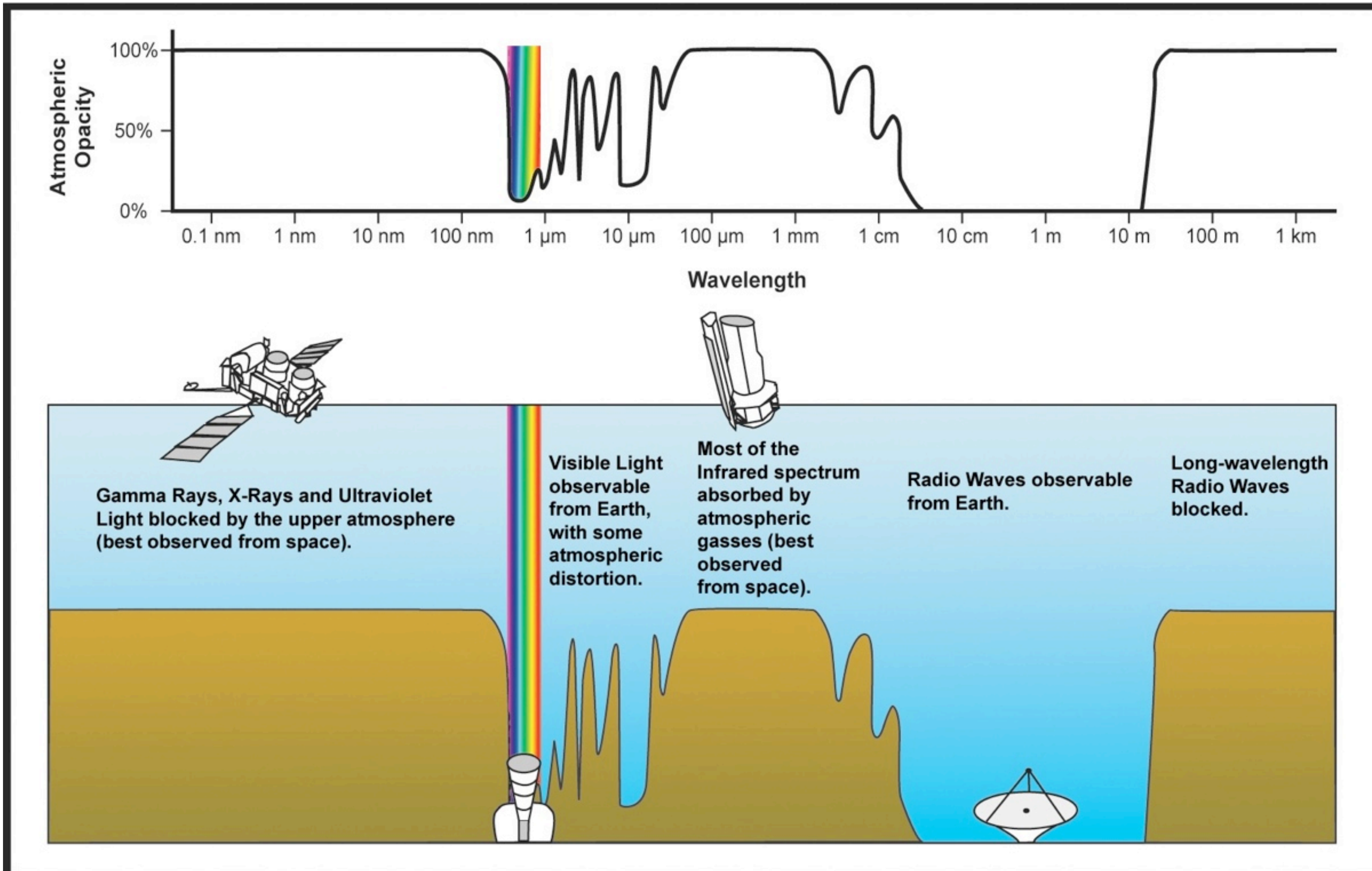
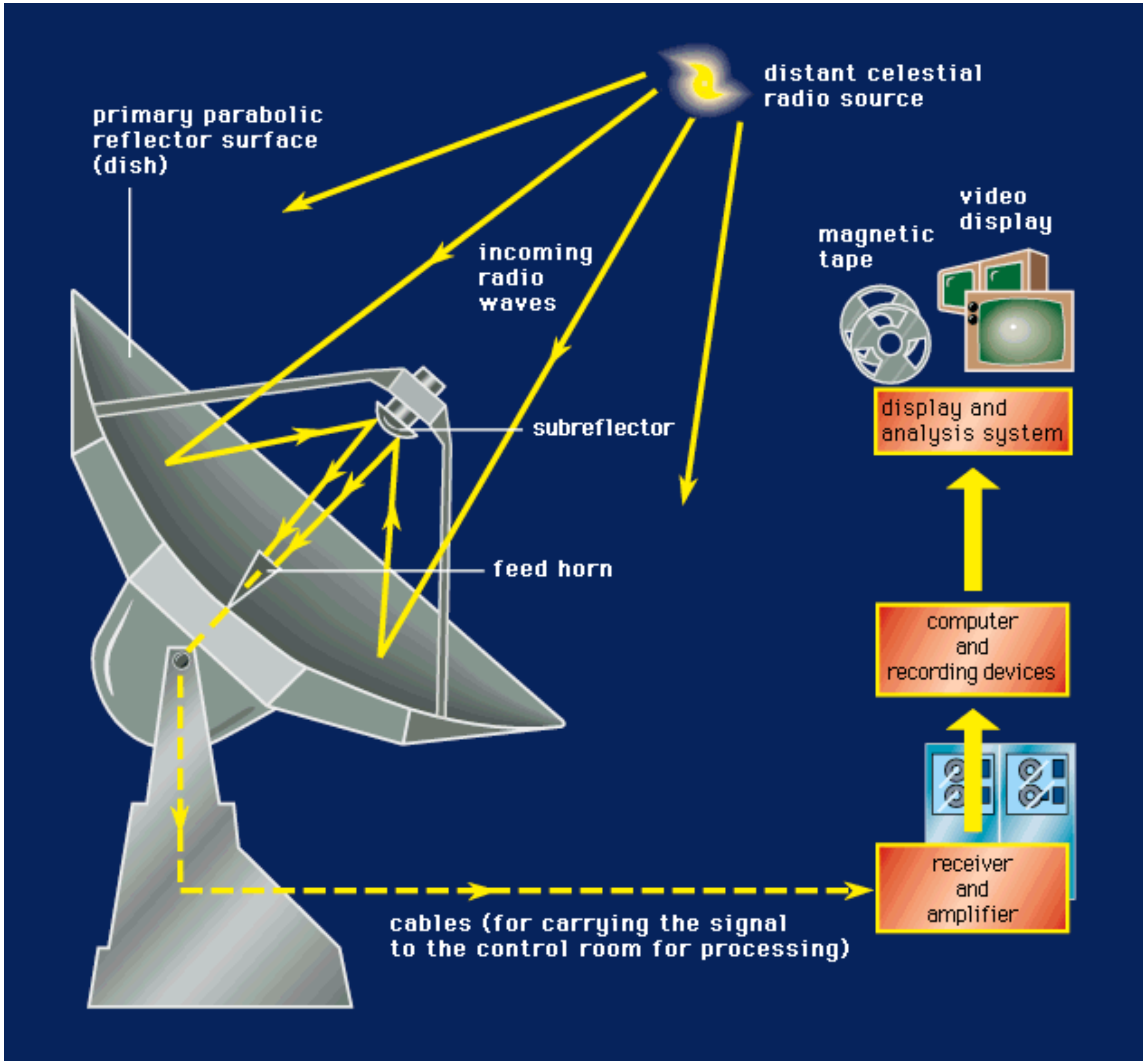


Image by NASA

Radio Astronomy



Courtesy of NRAO

Radio Astronomy

- Sun
- Milky Way
- Supernovae and their remnants
- Galaxies
- Active Galactic Nuclei
- Black Holes (candidates)
- Spacecraft

M33 in optical and radio

Image courtesy NRAO/AUI and NOAO/AURA/NSF

Radio vs. Optical astronomy

The imaging accuracy (resolution) of a telescope related to its wavelength and diameter: $\theta \approx \lambda/D$

Hubble Space Telescope:

$\lambda \approx 600$nm (visible light)

D = 2.4m

θ = 0.1 arcsecond

Onsola Space Observatory:

λ = 6cm (5GHz)

D = 25m

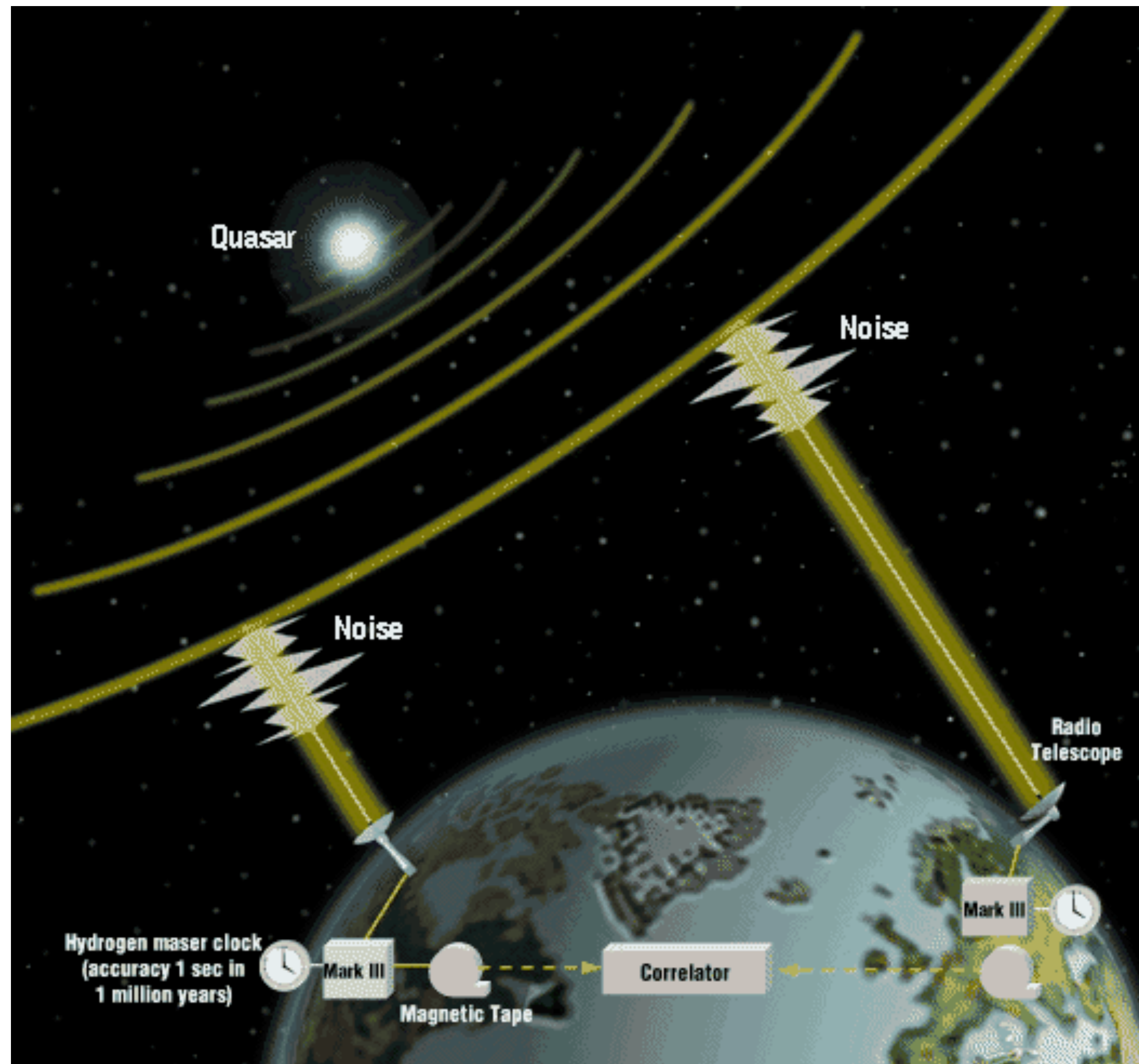
θ = 600 arcseconds

Moon: 3x3 pixels

Very Long Baseline Interferometry

• Create a huge radio telescope by using telescopes in different locations around the world at the same time

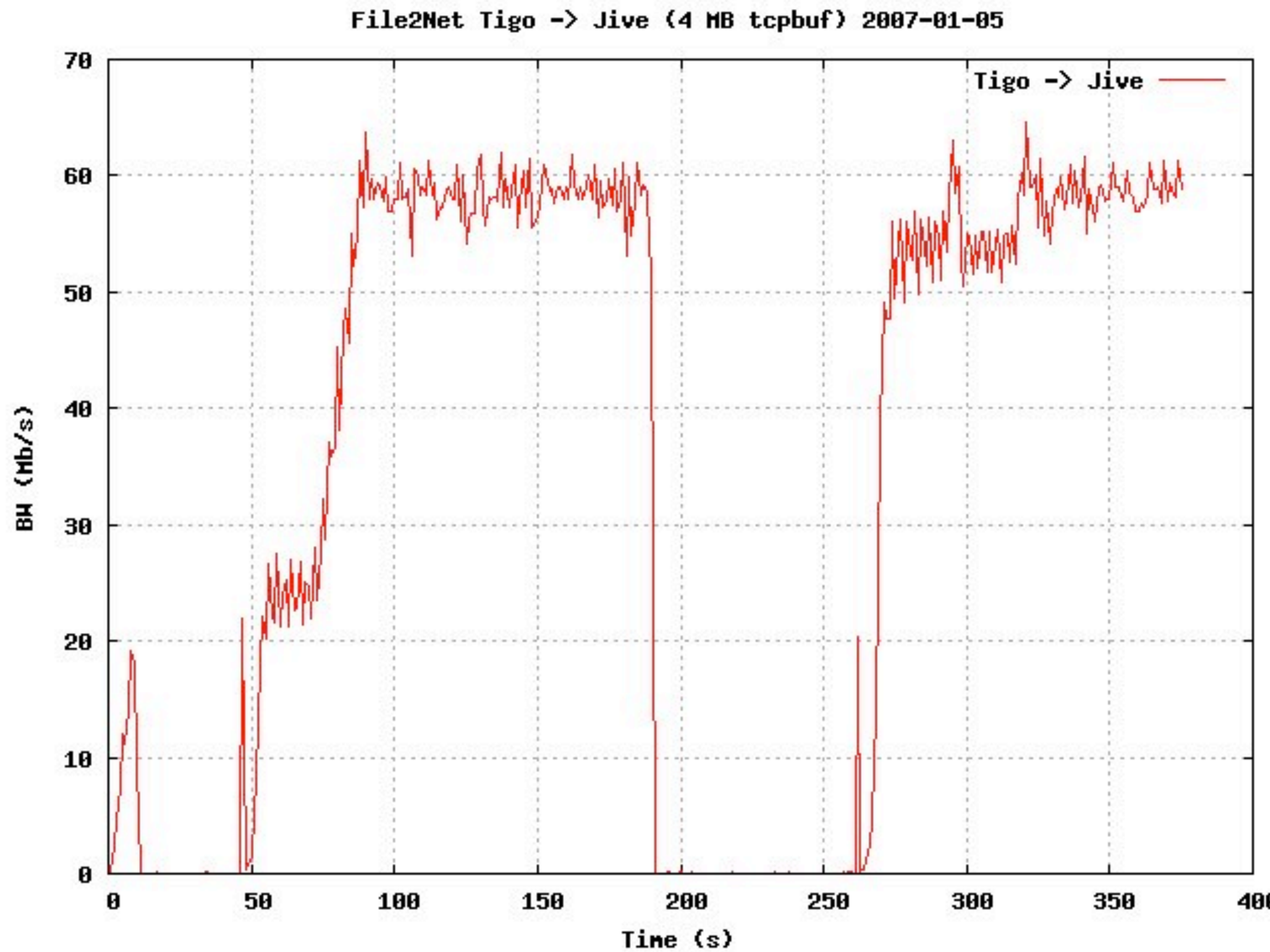
• Resolution depends on distance between dishes
• Sensitivity on dish area, time and bandwidth
• Requires atomic clock stability for timing
• Processed in a special purpose super-computer: Correlator, 16x 1024Mb/s



Very Long Baseline Interferometry

2TB sent from TIGO twice a week: 61Mb/s

Latency: 2 weeks



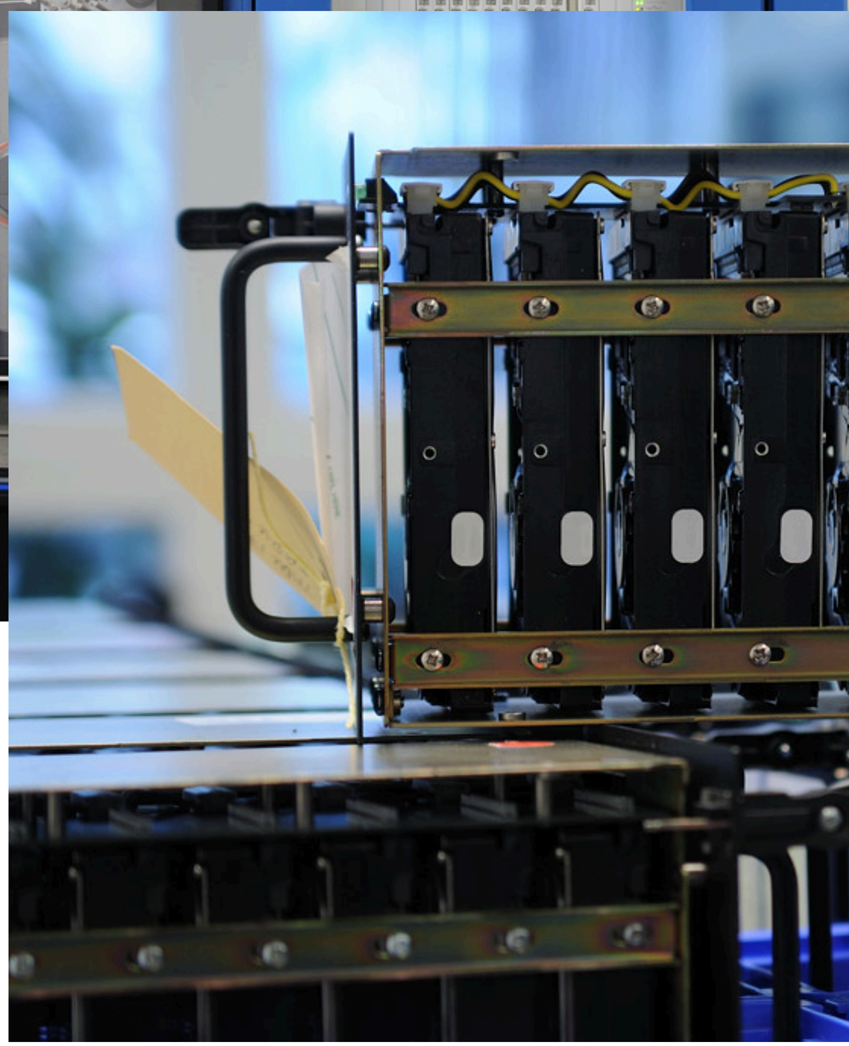
Latest e-VLBI test: 58Mb/s (max, 24Mb/s avg.)

Latency: 150ms

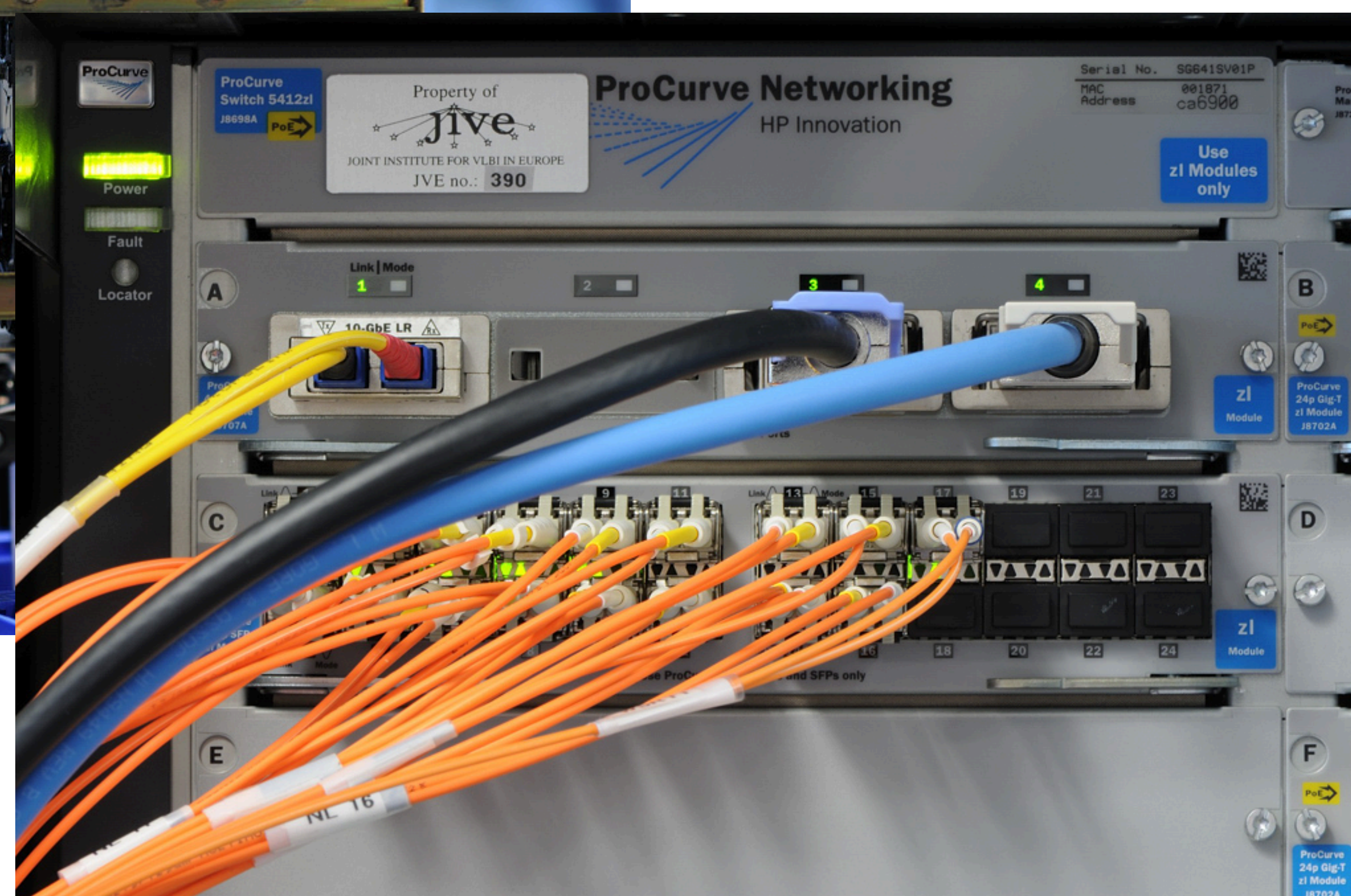
"Never underestimate the bandwidth of a station wagon laden with computer tapes hurtling down the highway"
(Andy Tanenbaum)

Very Long Baseline Interferometry

- Initially (1990) we used large single-reel tapes



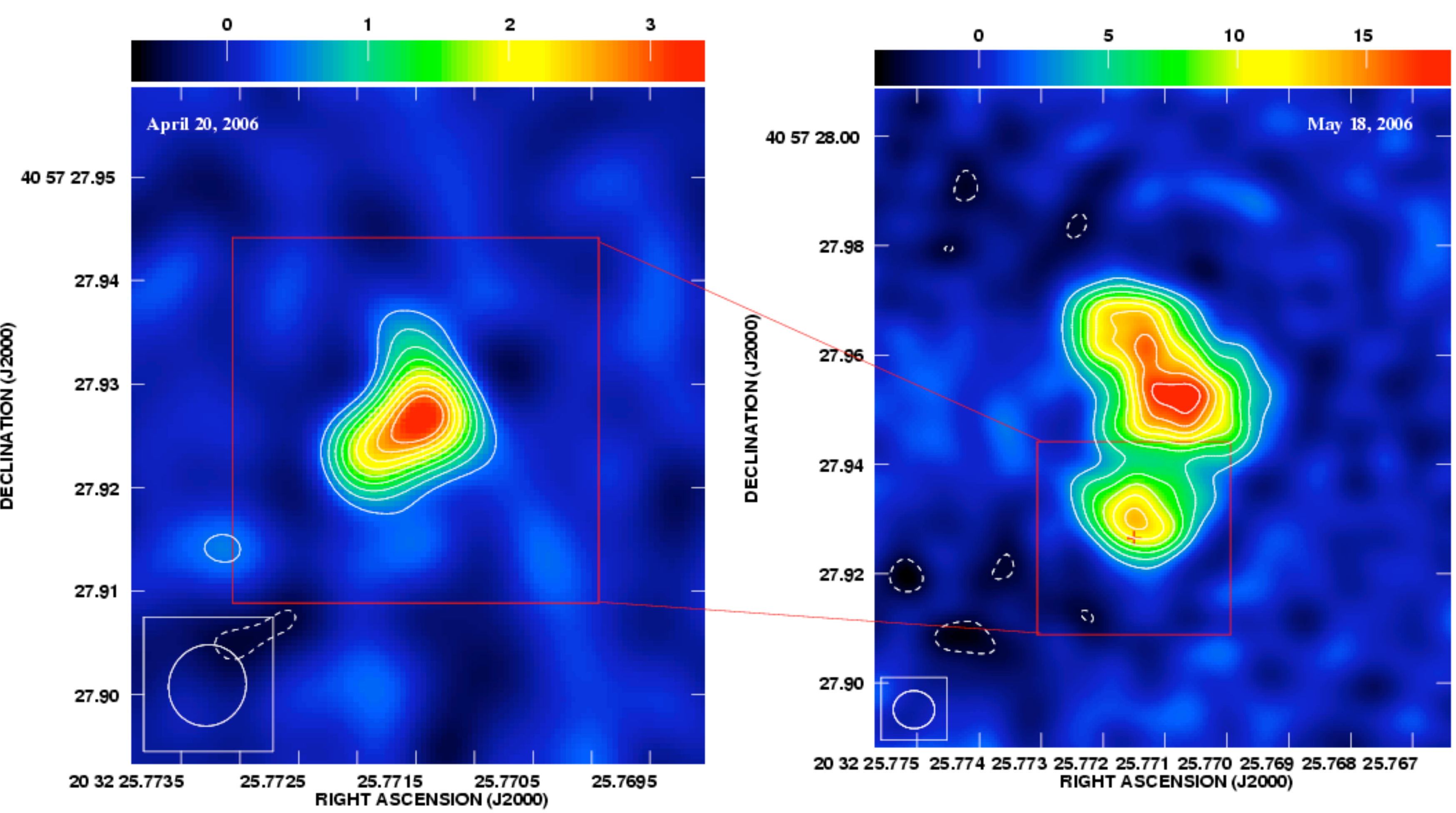
- Then came harddisk-packs



- And now: e-VLBI

Why e-VLBI

- Quick turn-around
- Rapid response
- Check data as it comes in, not weeks later
  (You can't redo just 1 telescope)
- More bandwidth
- Logistics (disks damaged/delayed/deleted...)

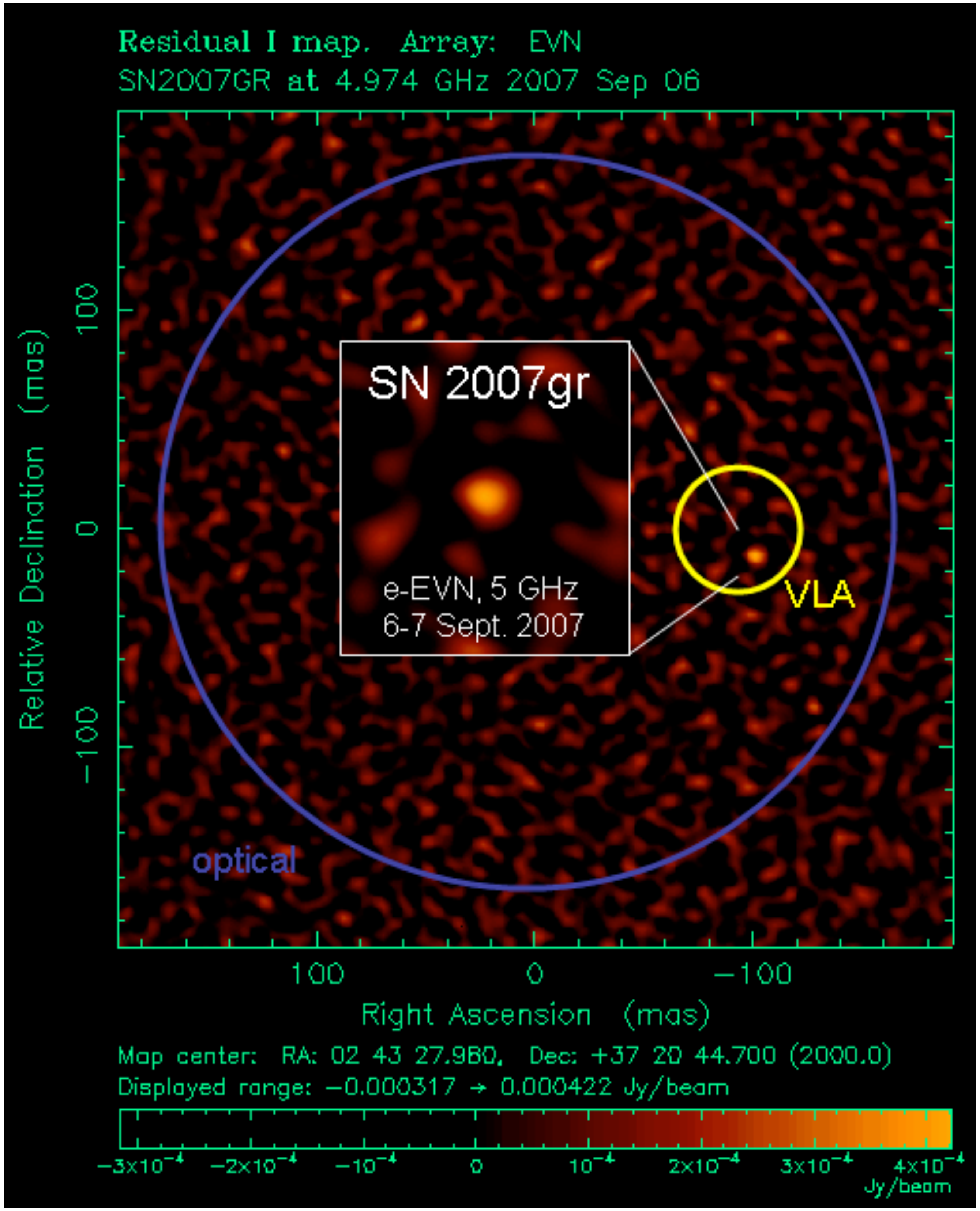


CygX-3

- Star + black hole
- Flares irregularly
- Timescale: days
- Left: 2 weeks late
- May: Observed flare with e-VLBI

e-VLBI observation of SN2007gr

- Supernova in august 2007
- This type of SN expands very rapidly
- With e-VLBI we could observe it in early sept.
- Look for: Jets, shell, neutron star, kick
- Millions of lightyears away, very faint and small
- Detection, sent Astronomical Telegram



Networking challenges

e-VLBI is:

- High Bandwidth: > 1 Gb/s
- Long Distance: Worldwide
- Near real-time
- Long duration: 12 hours
- But a little packet loss is OK
- Has to work with world-wide installed base (2.4 kernels a.o.)

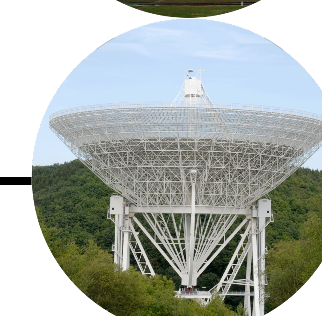
Network Overview







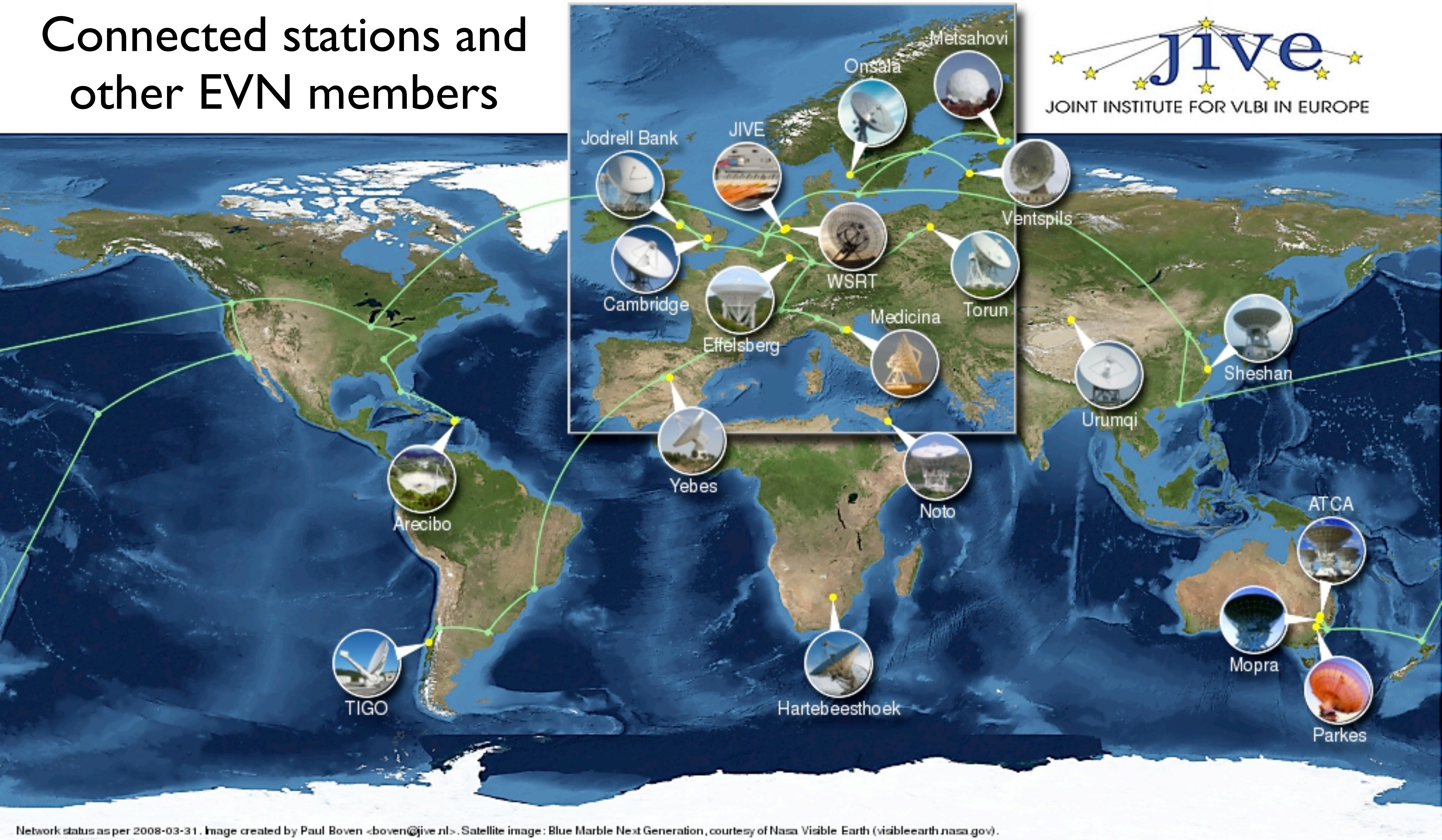


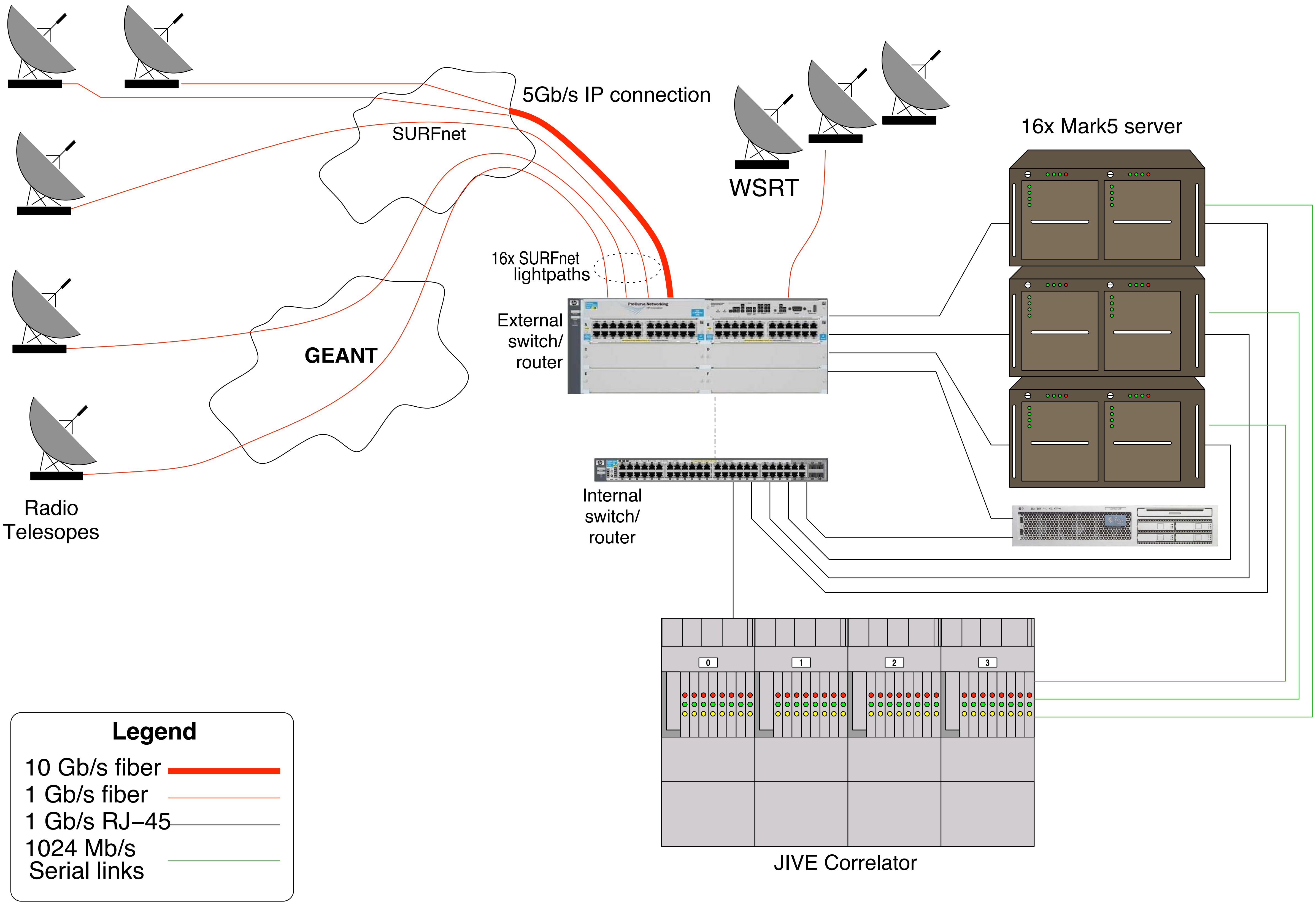
Telescope	Bandwidth	RTT
Sheshan	512 + 622 LP	180ms / 354ms
ATNF (2x)	2x 1Gb/s LP	343ms
Arecibo	512Mb/s VLAN	154ms
TIGO	95Mb/s	150ms
Medicina	1Gb/s LP	29.7ms
Onsala	1Gb/s	34.2ms
Torun	1Gb/s LP	34.9ms
Jodrell Bank	2x 1Gb/s LP	18.6ms
WSRT	1Gb/s Dark Fiber	0.57ms
Effelsberg	1Gb/s (10G)	13.5ms

The current e-VLBI network

Connected stations and other EVN members

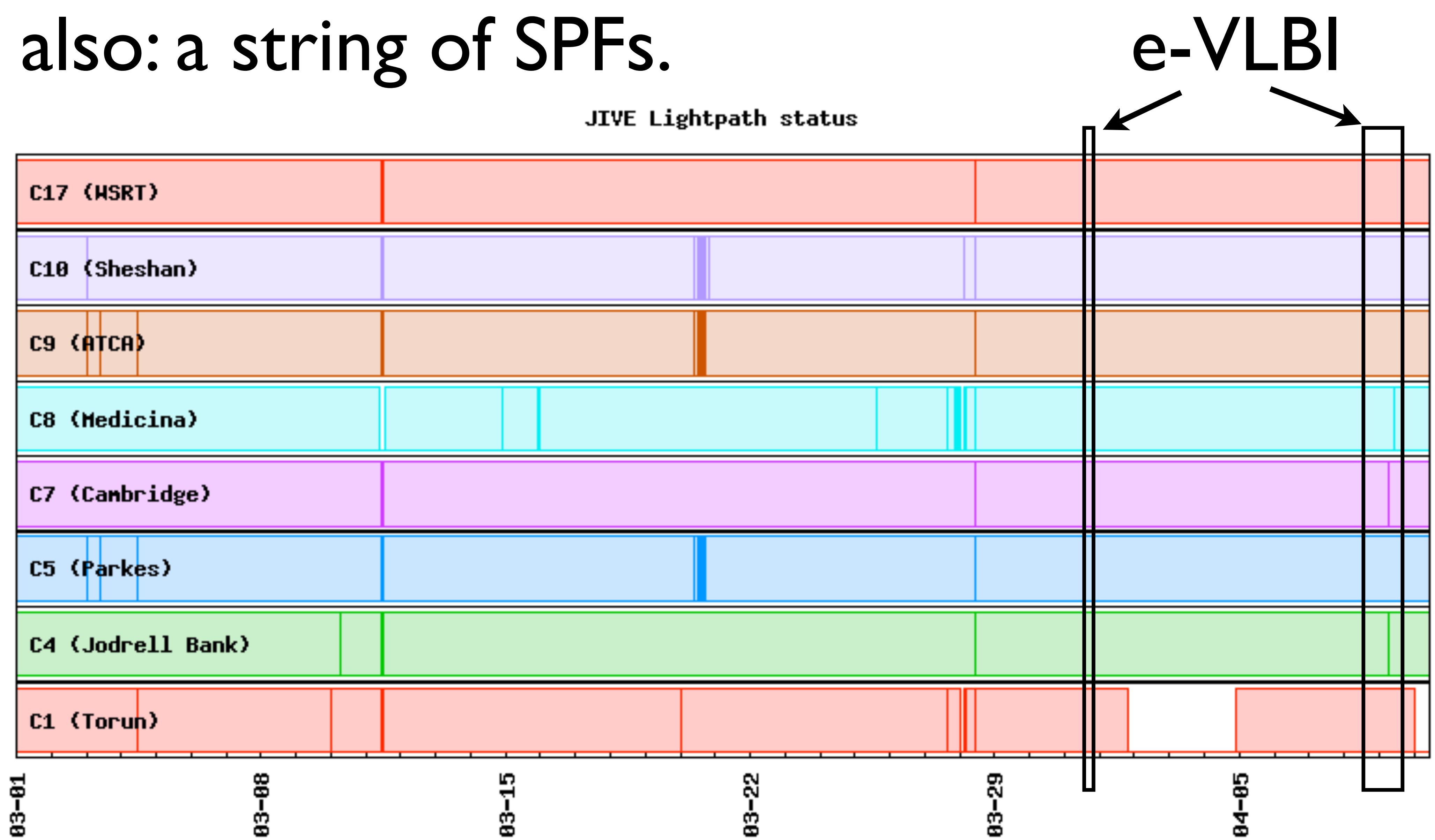


JIVE Network Setup

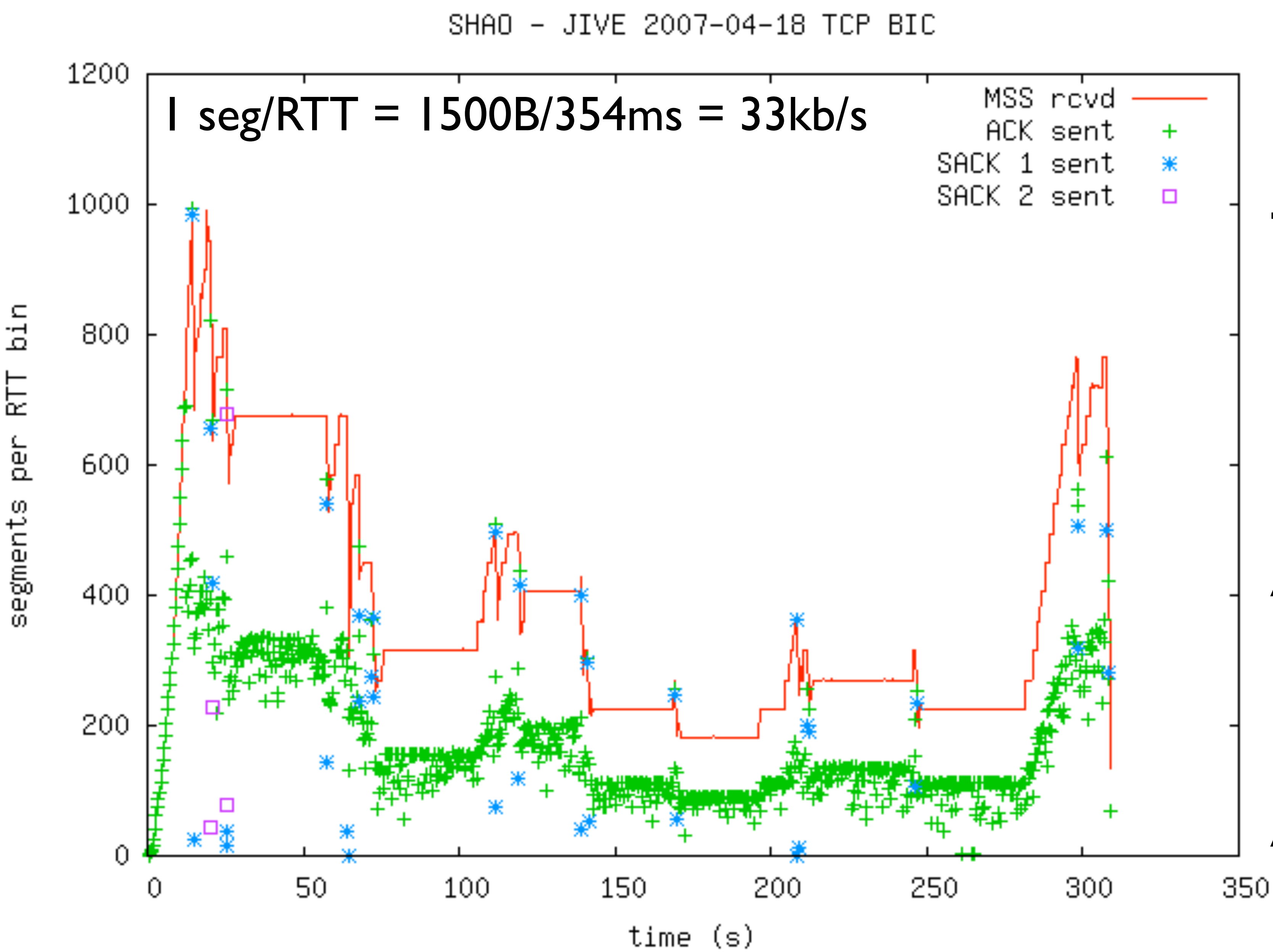


Lightpaths

- 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 SPFs.



TCP behaviour



Shanghai to JIVE

TCP-BIC stacks

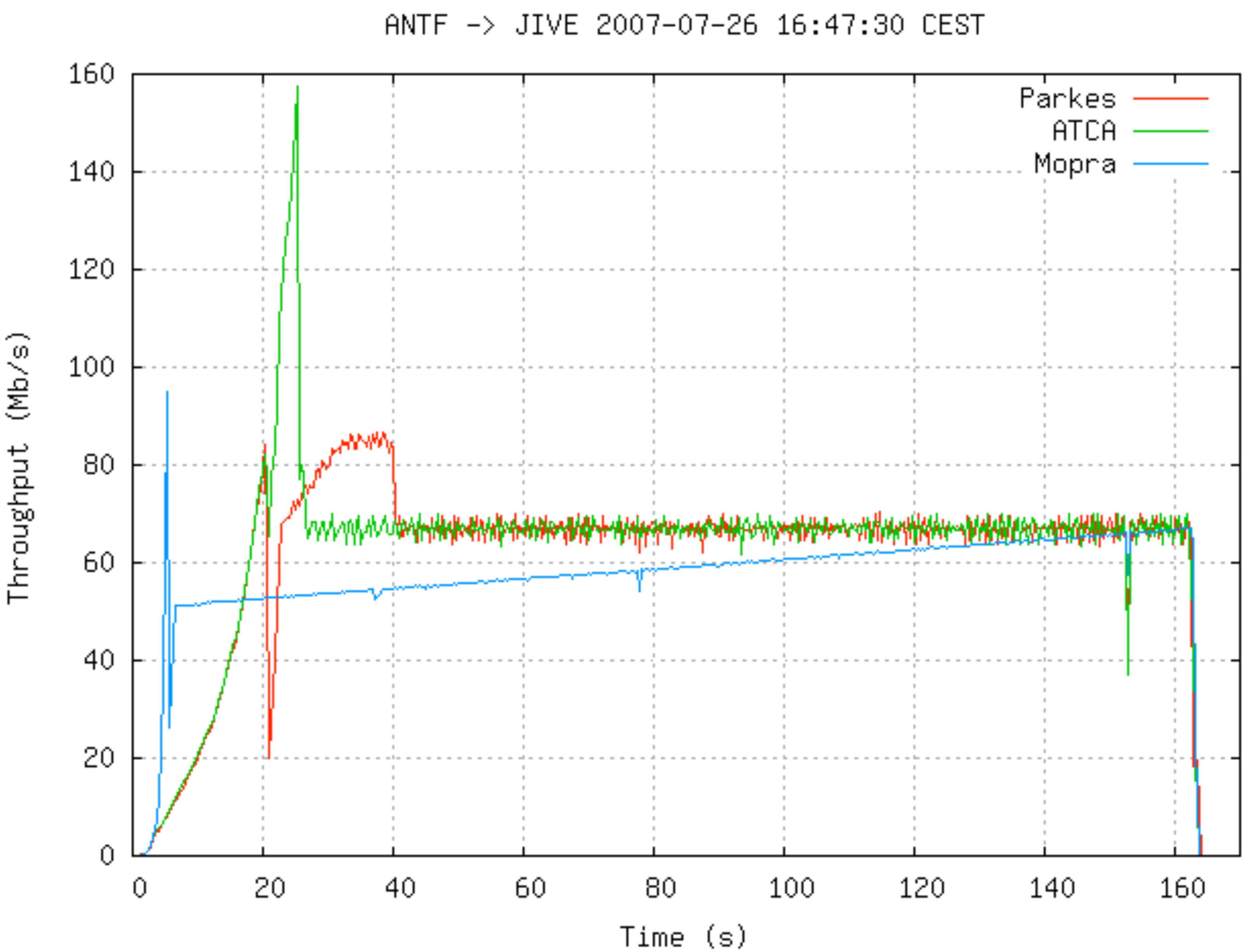
622Mb/s lightpath to HK

Apx. 25 packets lost in 5 minutes

Achieved only 33Mb/s

At large RTT, TCP cannot recover from packet loss

TCP startup/recovery



Australia to JIVE, 343ms

BIC and RENO stacks

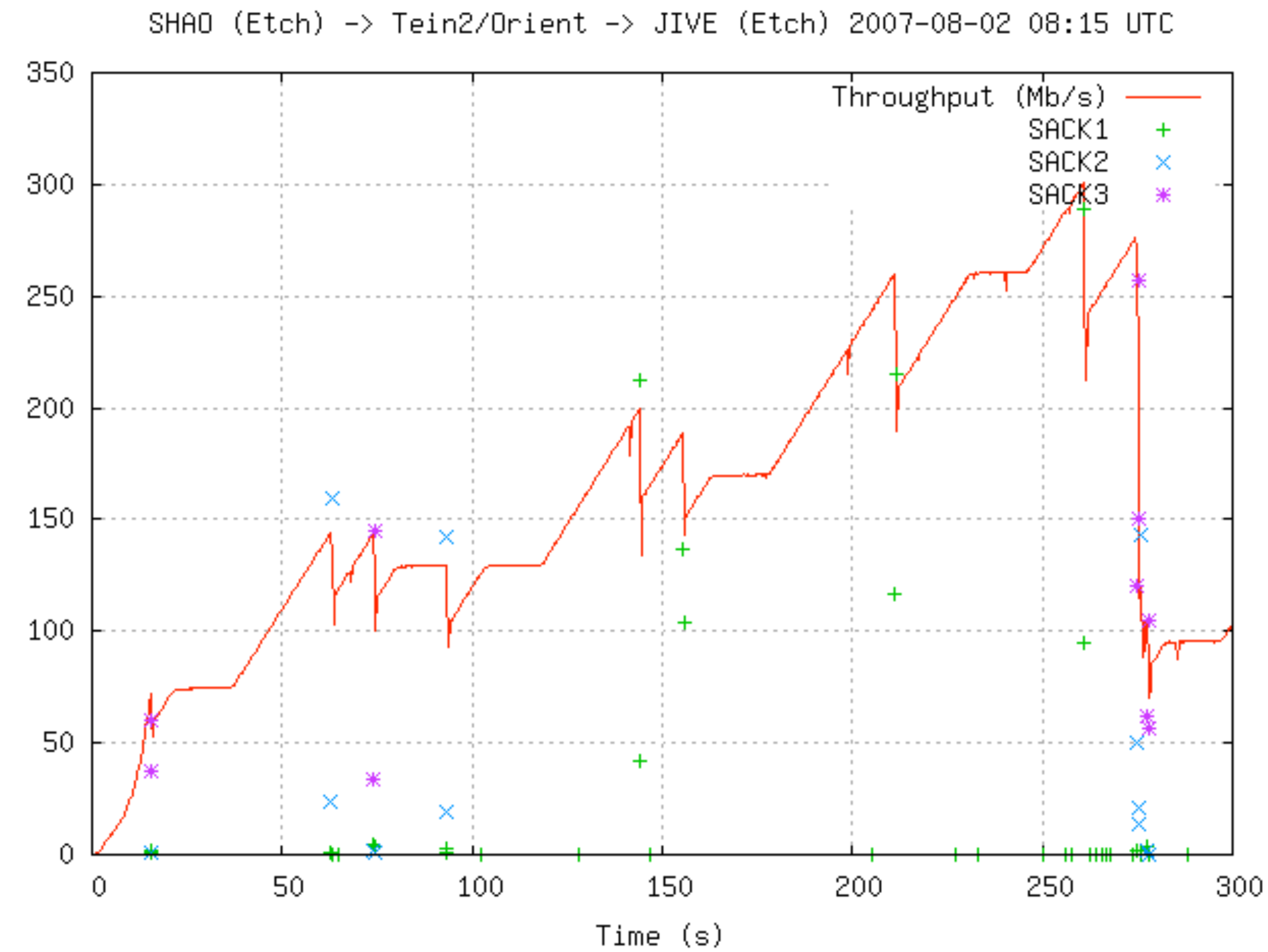
3x 1Gb/s lightpath

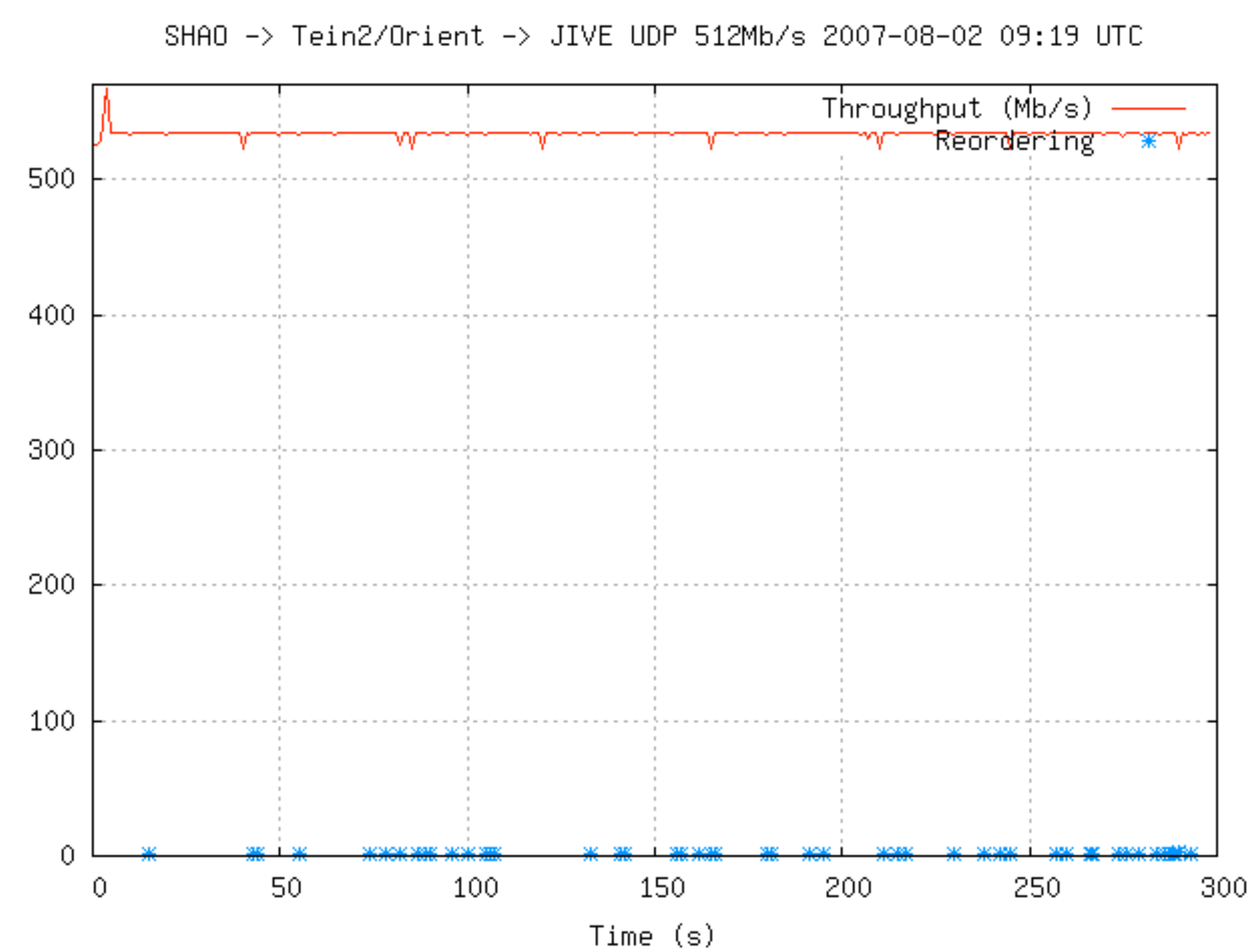
Data sent at 64Mb/s

At large RTT, TCP takes a while to get up to speed

TCP Research

- Mirror port (span)
- tcpgrok.c - analyze TCP
- eVLBI: RTT up to 354ms
- Window Size
- SACK-bugs
- Tuning defeats fairness
- Lightpath connections
- Conclusion: UDP





UDP surprises

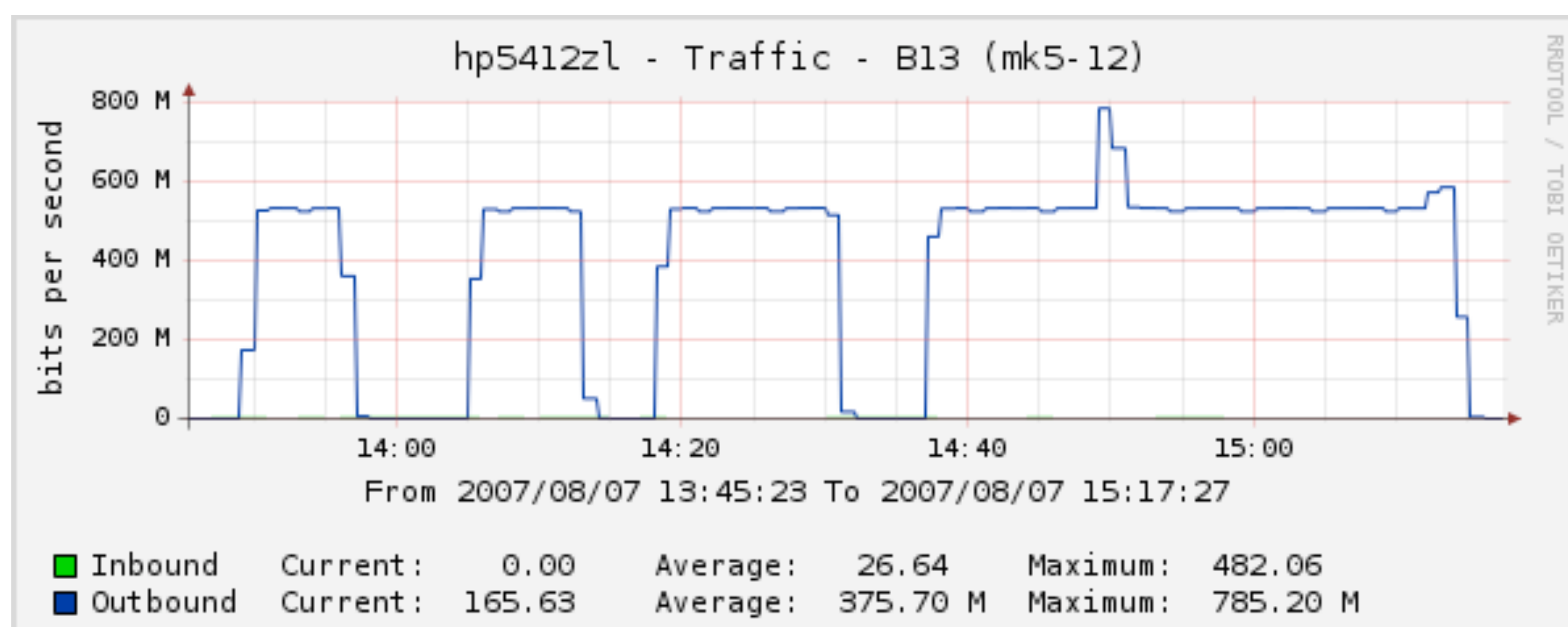
- You must prevent bursts - evenly space packets
  - Timing packets costs CPU
  - Cannot 'sleep()' because of 100Hz granularity in 2.4 Linux kernels.
- One-way lightpath UK → JIVE
  - UDP to the rescue:
  - Manually set our MAC in their ARP table
  - Used in observation of SN2007gr Supernova

UDP surprises

- Network switches learn MAC addresses by flooding
- Then listening for a reply - but our UDP is one-way

  → Run a ping to all receiving servers
- When a port goes down (e.g. crash):

  Switch forgets the MAC and starts flooding!
- Really a problem at 512Mb/s operations
- Static-Mac didn't help (bug in switch)
- 16 servers, each now has a /30



The 1Gb/s speedbump

- VLBI (tape based) comes in fixed speeds, power of 2: 128Mb/s, 256Mb/s, 512Mb/s - and 1024Mb/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
- Lightpaths come in 'quanta' of 150Mb/s, but Ethernet doesn't

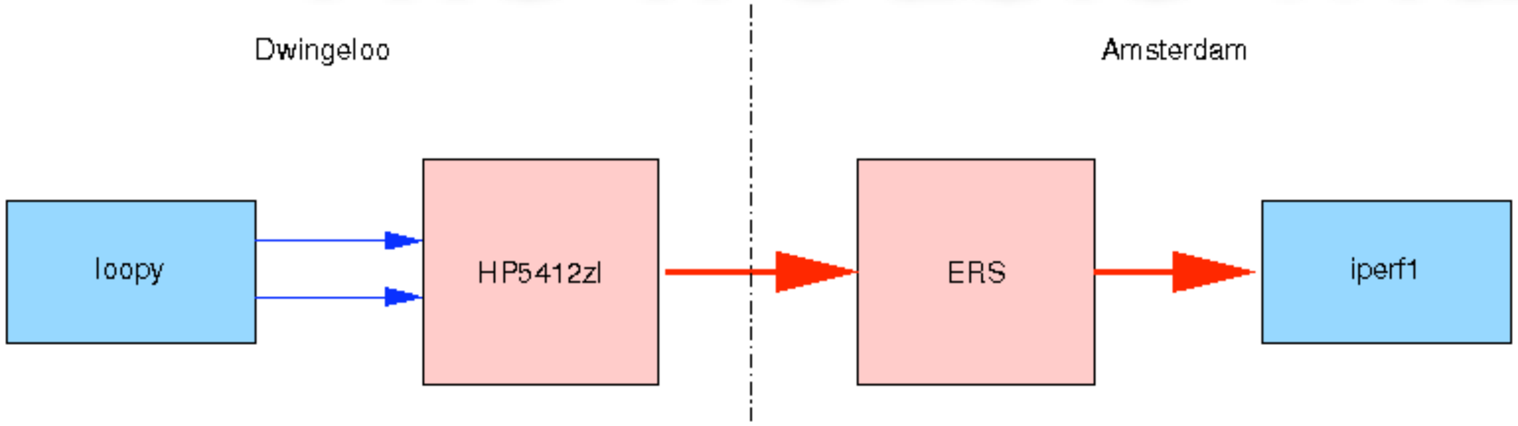
The Trouble with Trunking

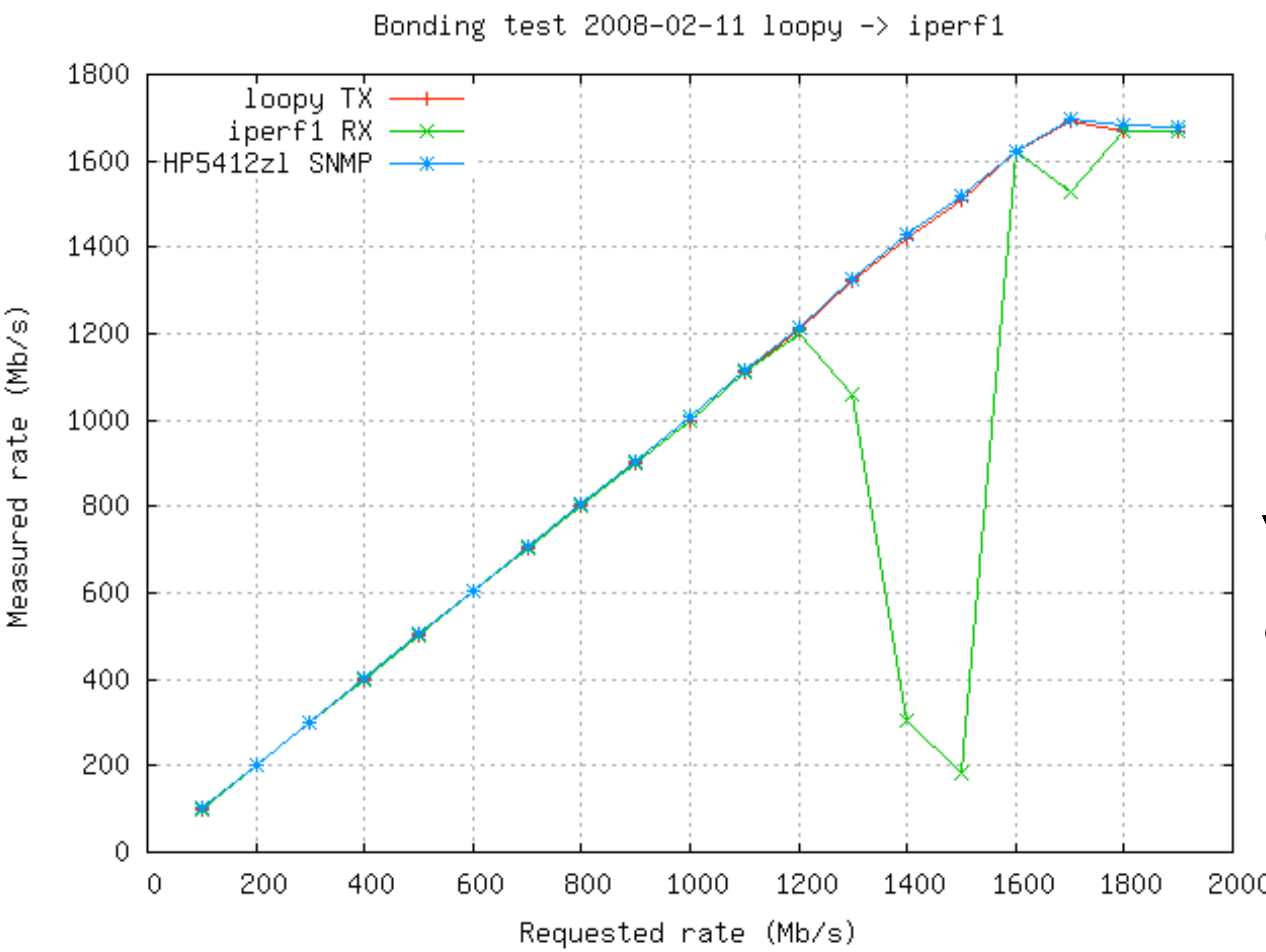
- Standard trunking: LACP (802.3ad)
  - Uses a hash of source/destination MAC, IP and/or Port to choose outgoing port
  - This is to prevent re-ordering
  - A single TCP/UDP stream will use only 1 link member!

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

"Do NOT cross the streams!"

The Trouble with Trunking

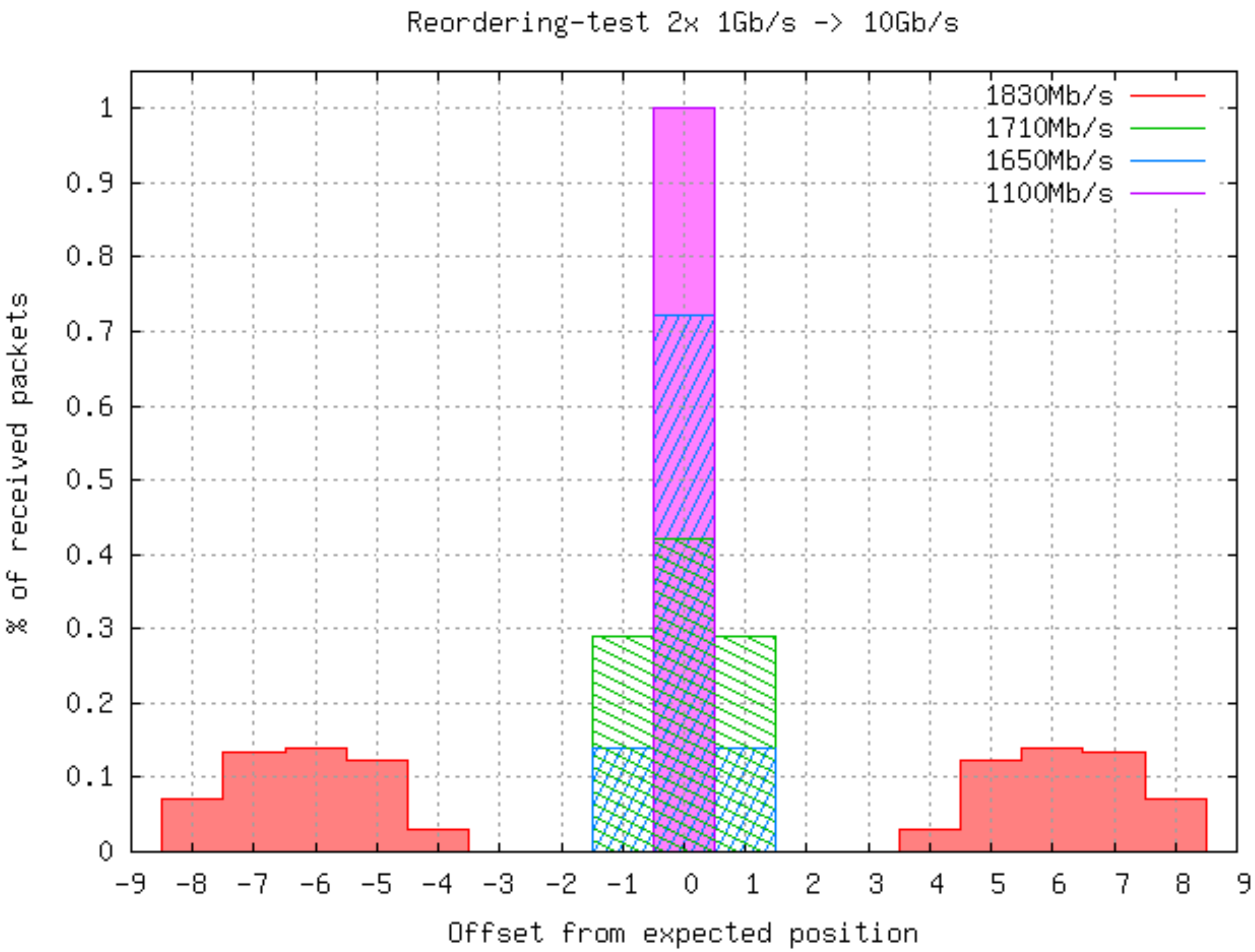




- iperf test using UDP
- Bonding driver
- To SURFnet, we didn't have 10G yet.
- Conclusion: bonding works well enough for eVLBI (1024Mb/s)
- But not as good as expected

No Trouble with Trunking!

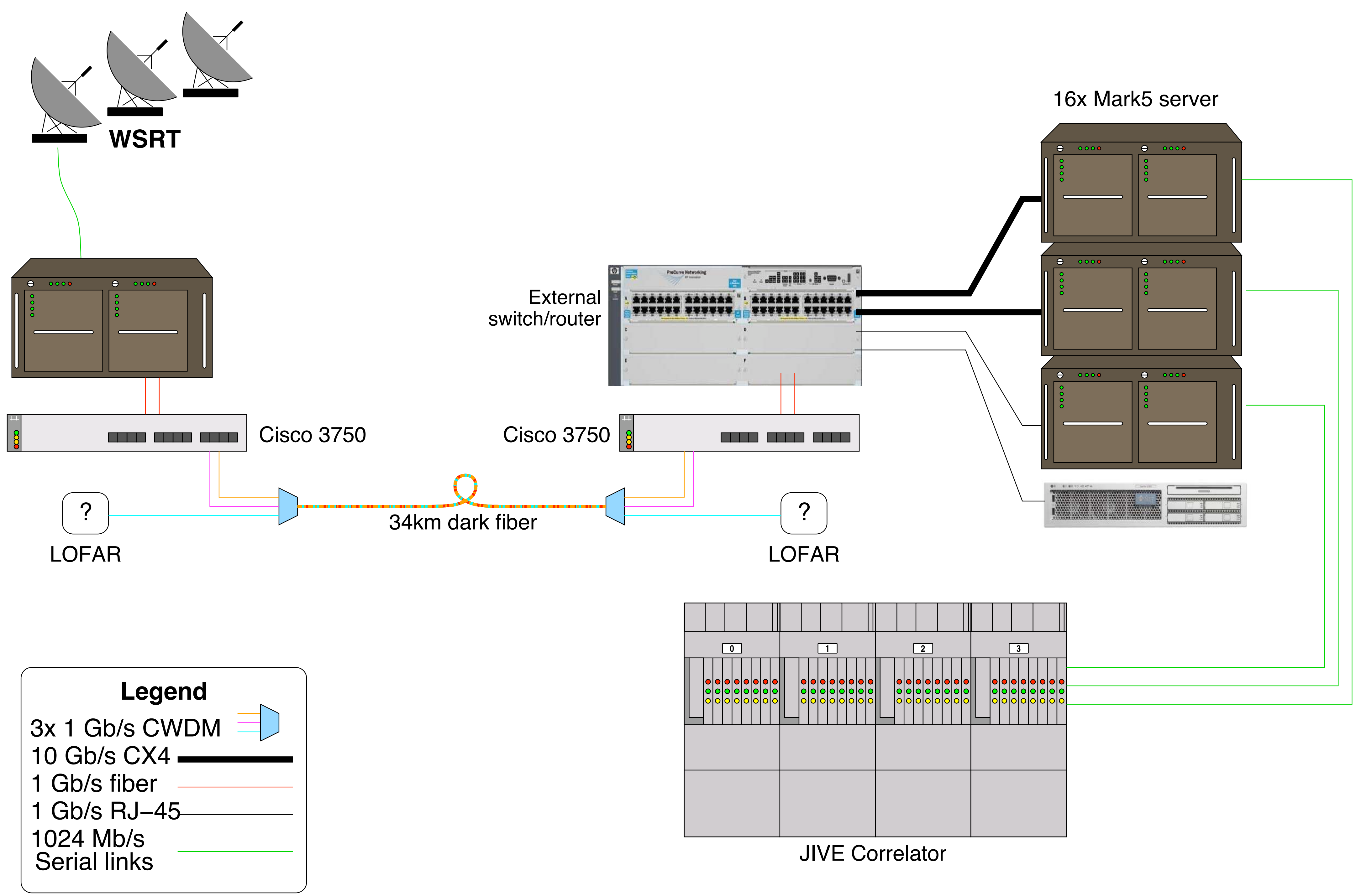
- iperf gets really confused by re-ordering of packets
- Wrote a simple re-implementation for UDP
- Store S/N to track re-ordering, post-process



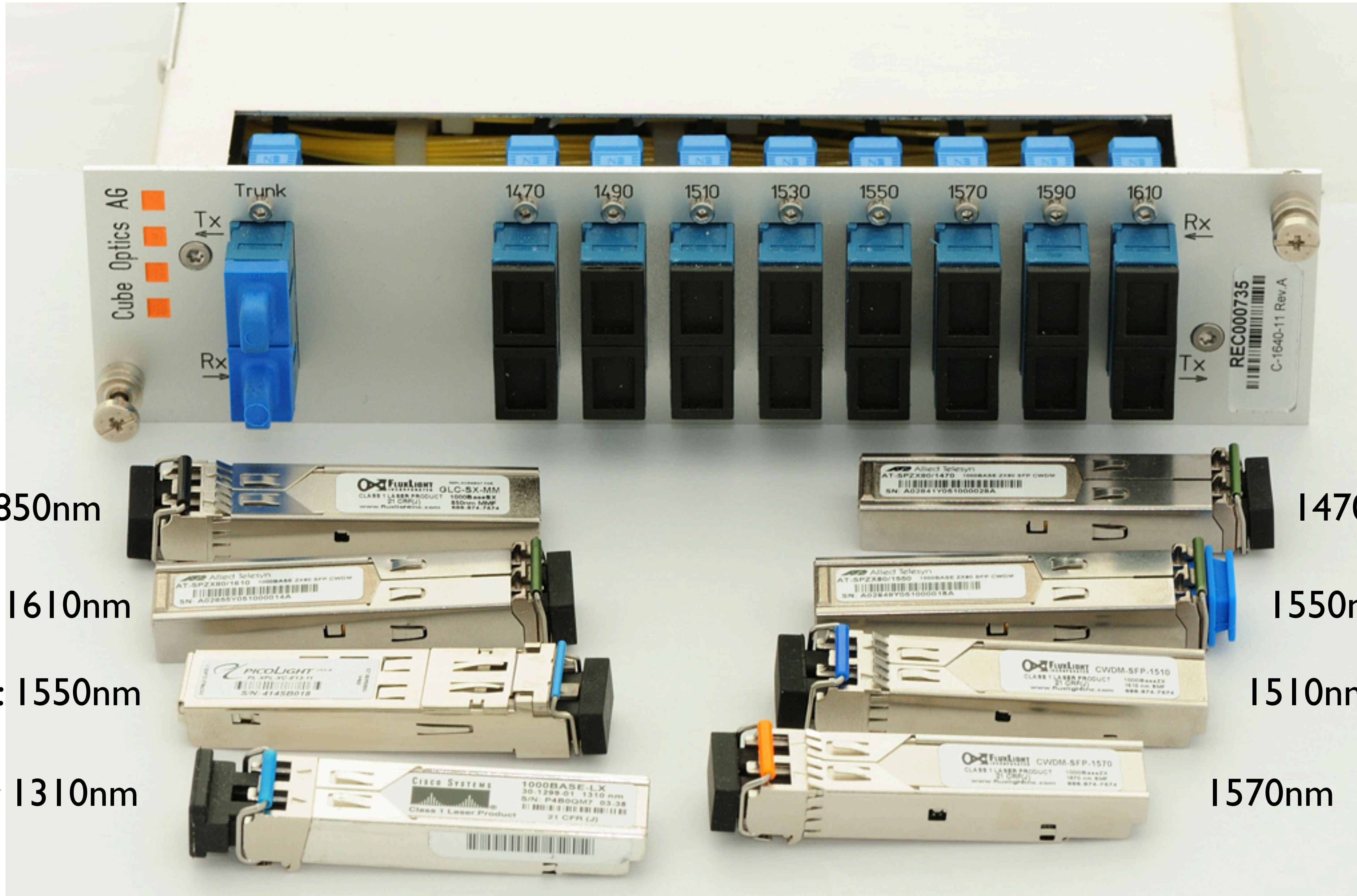
- No packet loss even at 1830Mb/s
- No re-ordering below 1100Mb/s
- Little re-ordering below 1710Mb/s

CWDM from WSRT to JIVE

Much cheaper than upgrading to 10Gb/s



All the colours of the rainbow...



SX: 850nm

1610nm

ZX: 1550nm

LX: 1310nm

1470nm

1550nm

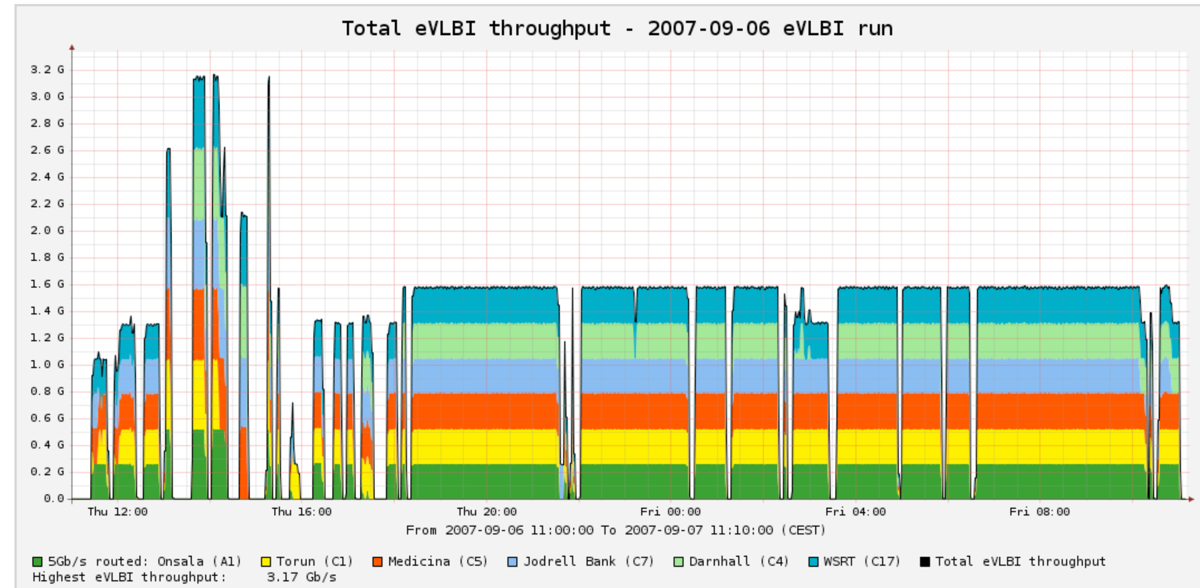
1510nm

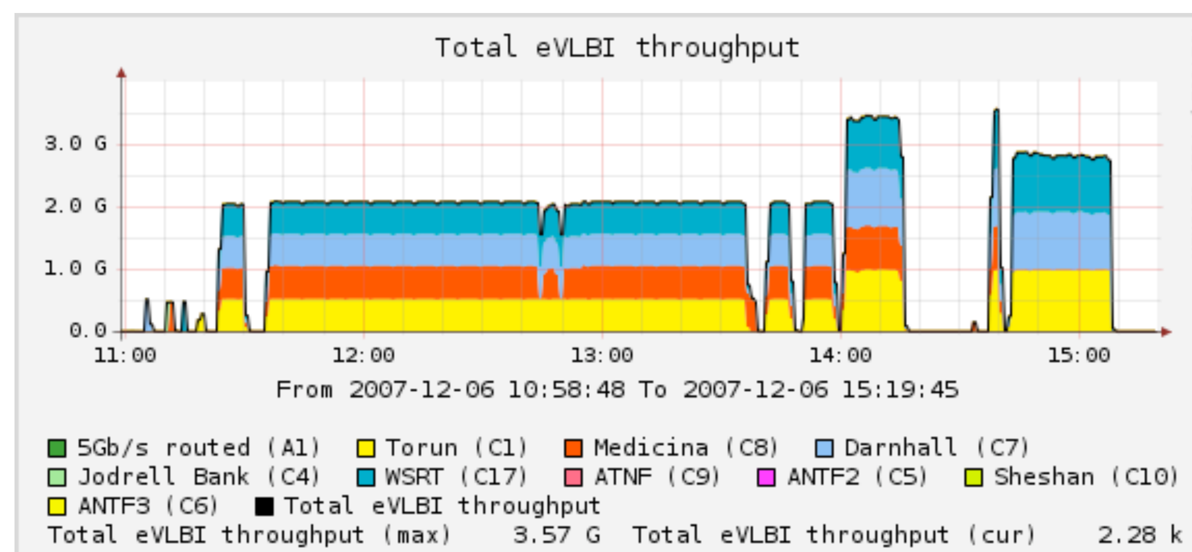
1570nm

... and then some.

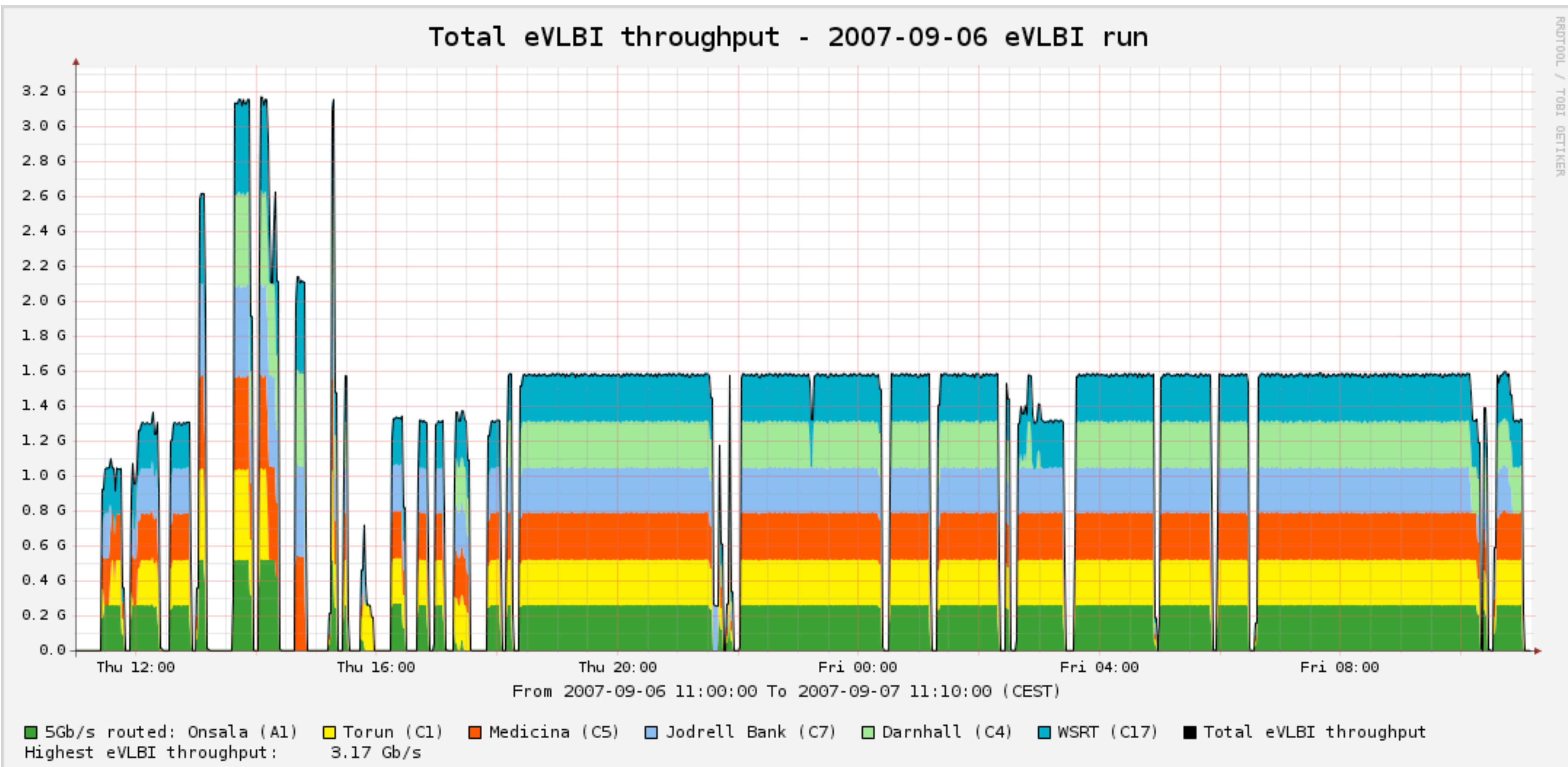
Cacti
http://graphs.jive.nl

- Cacti is an open-source network management tool
- Records SNMP counters
- Public access website
- Records once every minute
- 64 bit counters
- Secure SNMPv3

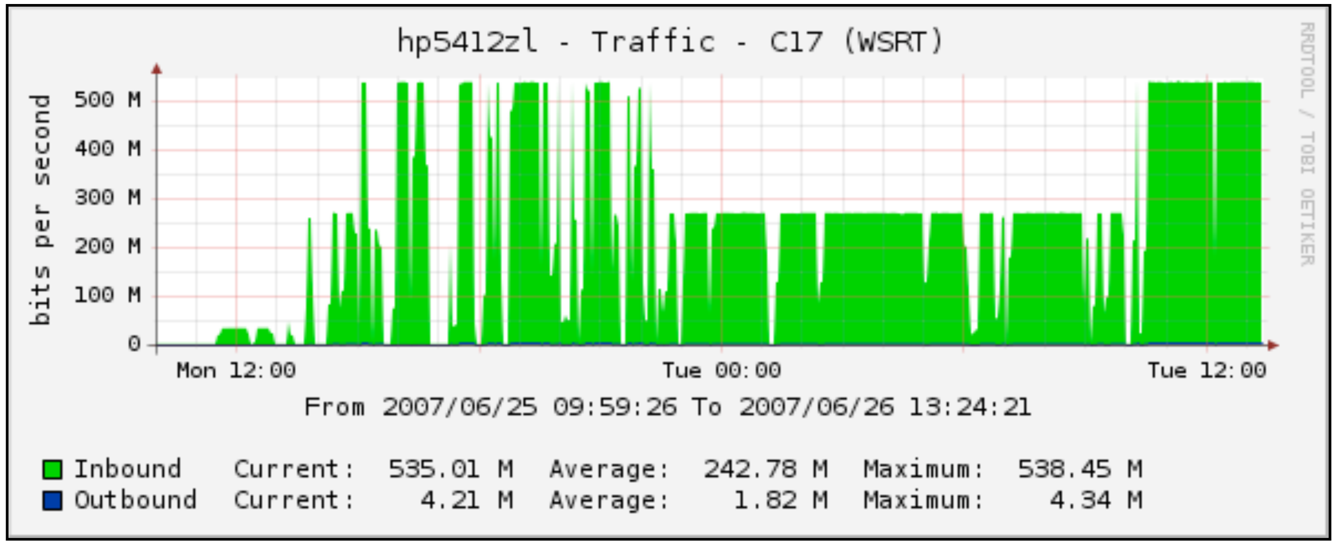




Cacti
http://graphs.jive.nl

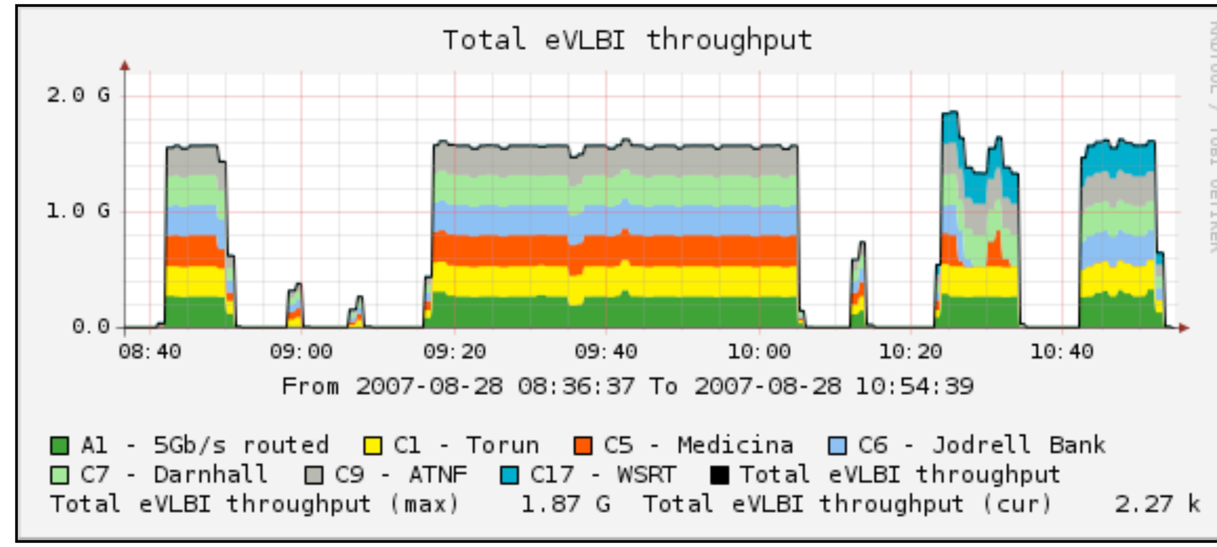


A year in graphs



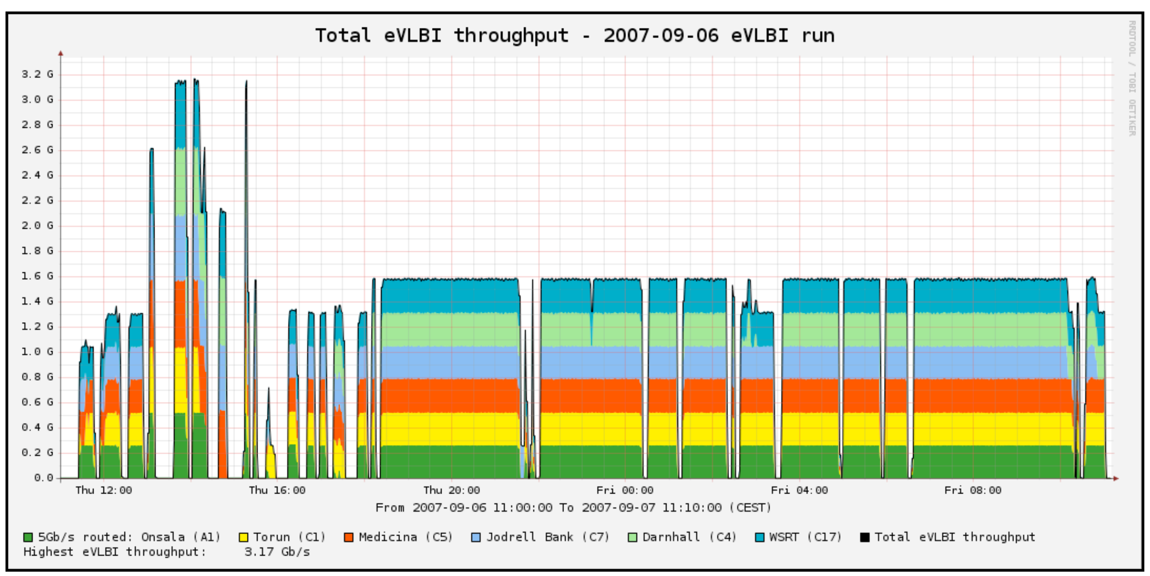
2007-06-25: 6x 256Mb/s

Calibrators near binaries



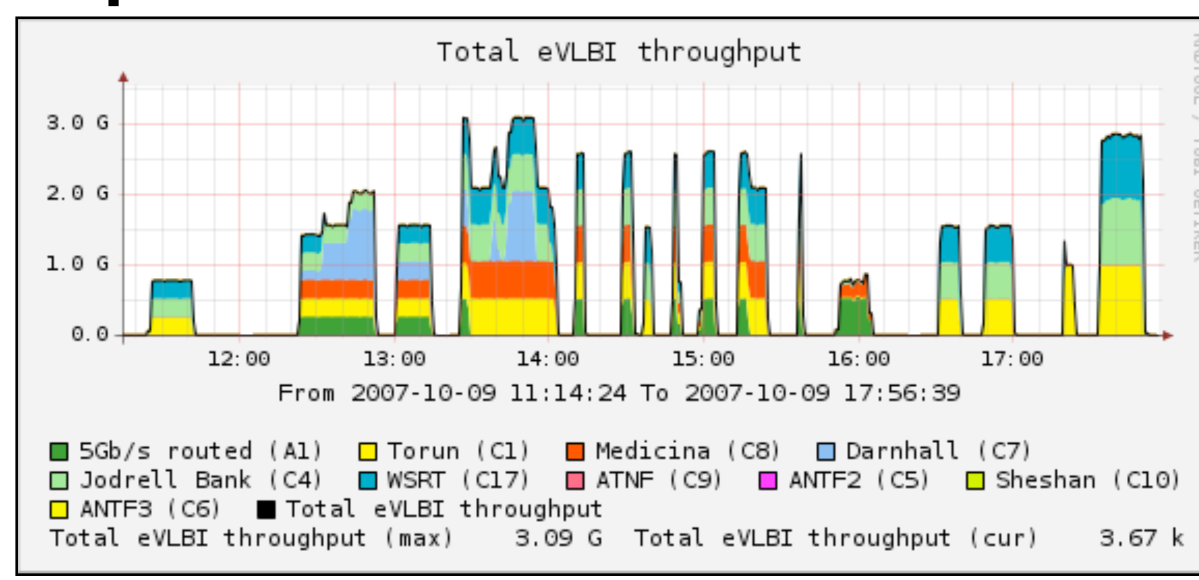
2007-08-28: 6x 256Mb/s

Apan Demo



2007-09-06: 6x 256Mb/s

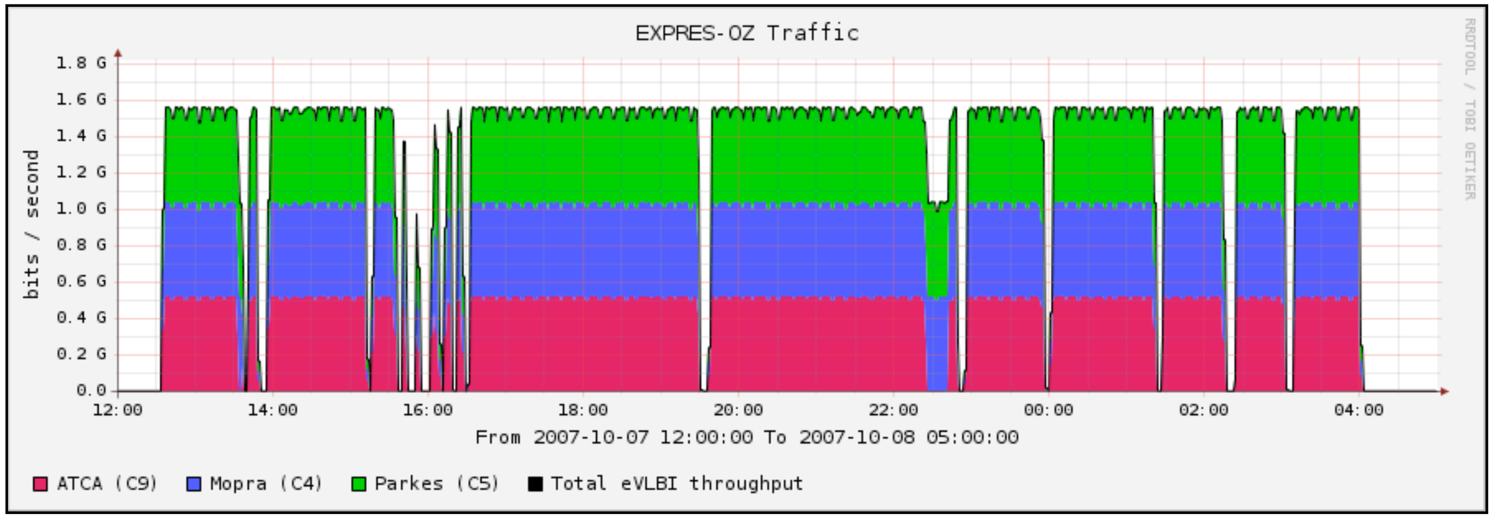
SN2007gr

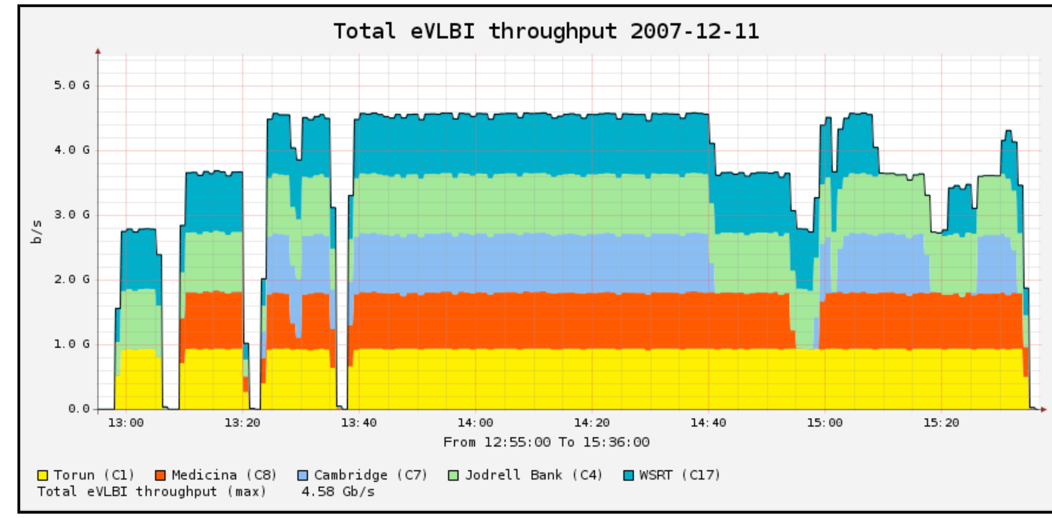


2007-10-09 3x 1Gb/s

Test session

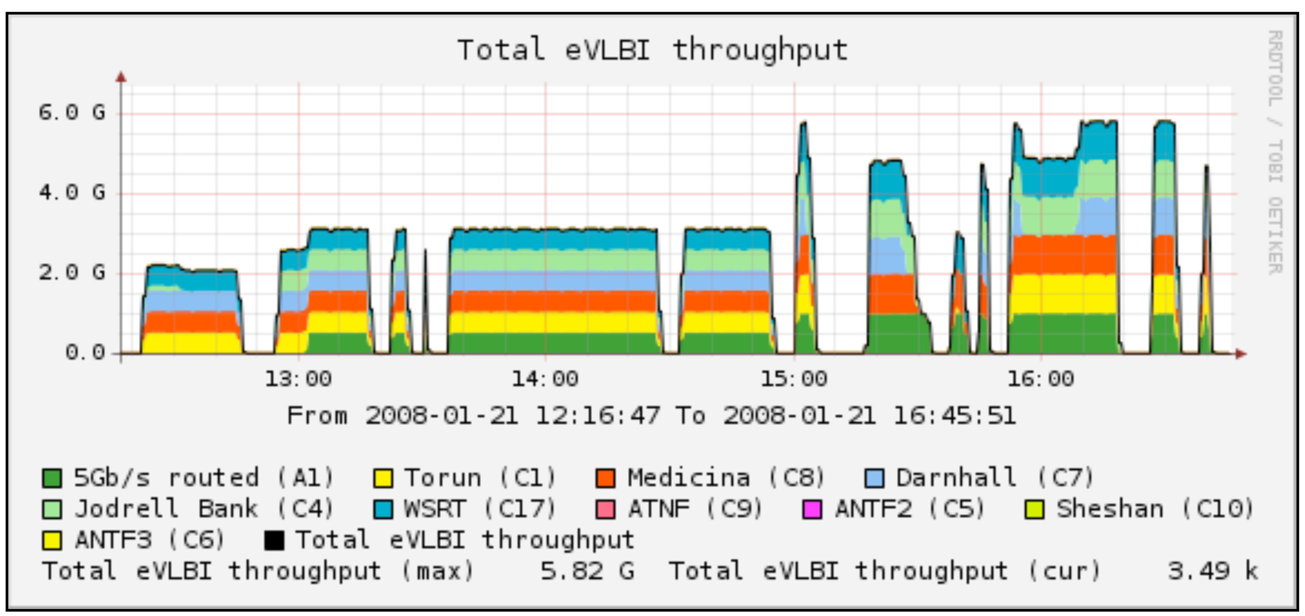
A year in graphs (2)

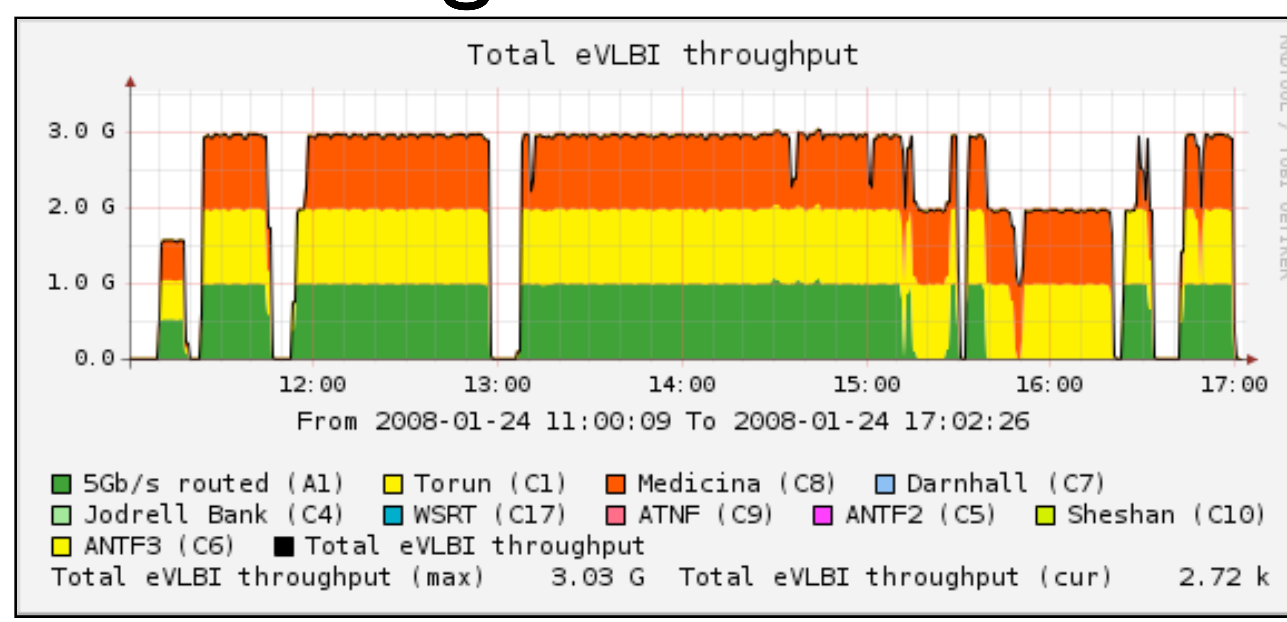




2007-10-07: 3x 512Mb/s ATNF SN1987a

2007-12-11: 5x 917Mb/s Fringes on all baselines

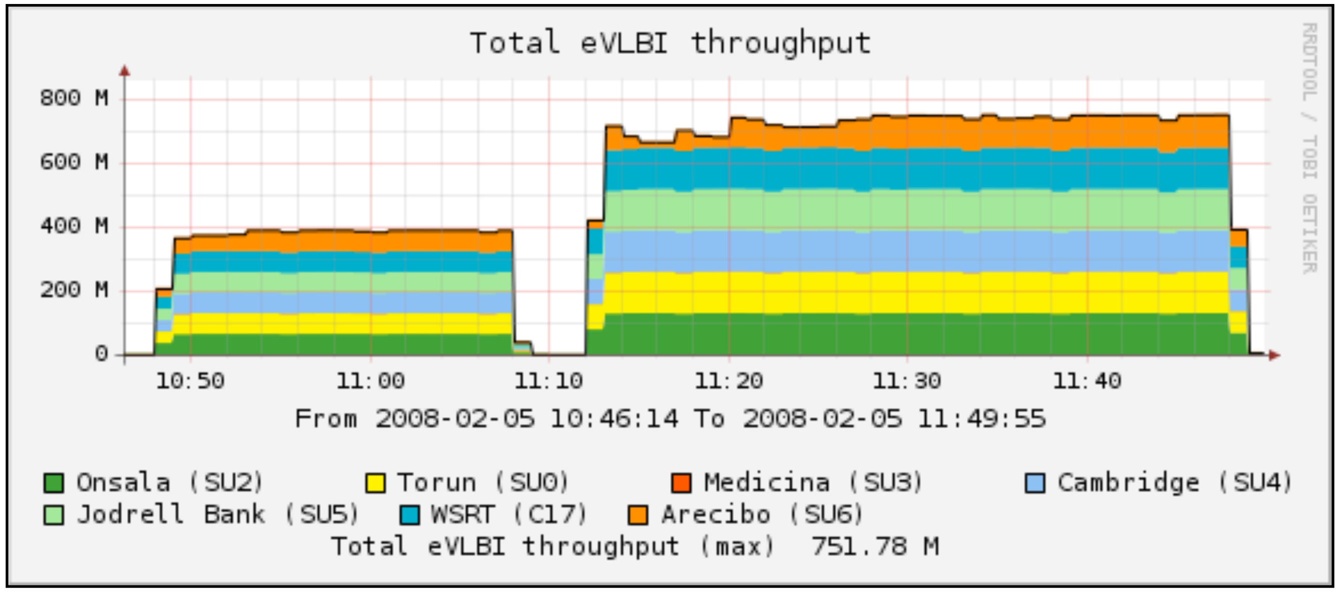




2008-01-21: 6x 970Mb/s (1:22 packet drop rate)

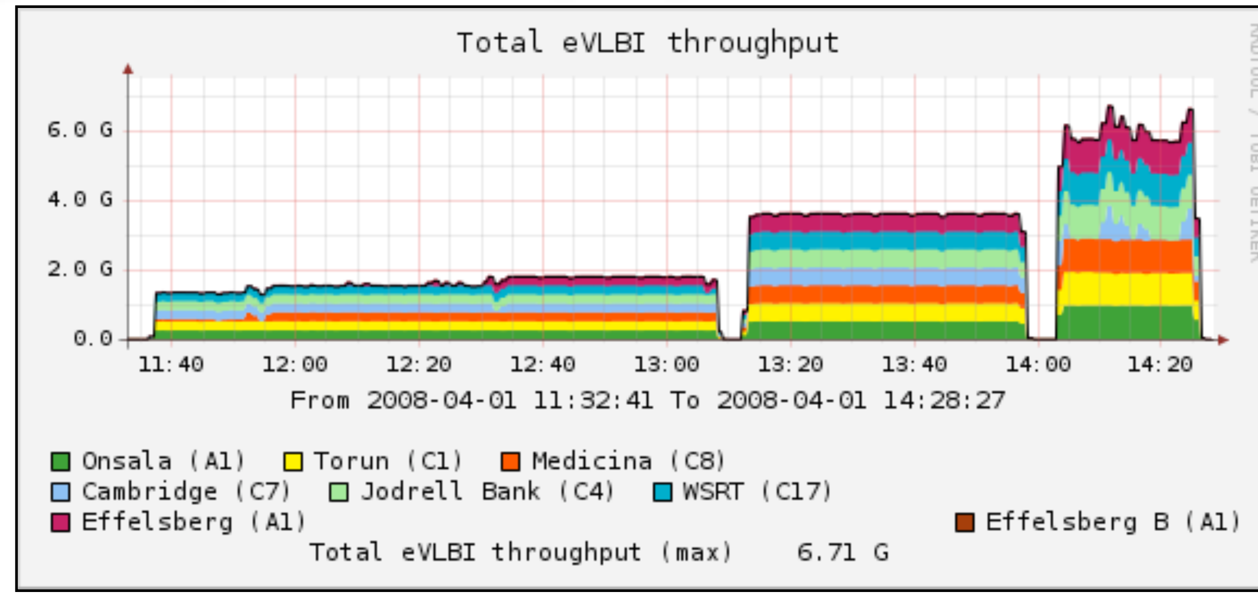
2008-01-24: 3x 970Mb/s (1:22 with all headers)

A year in graphs (3)



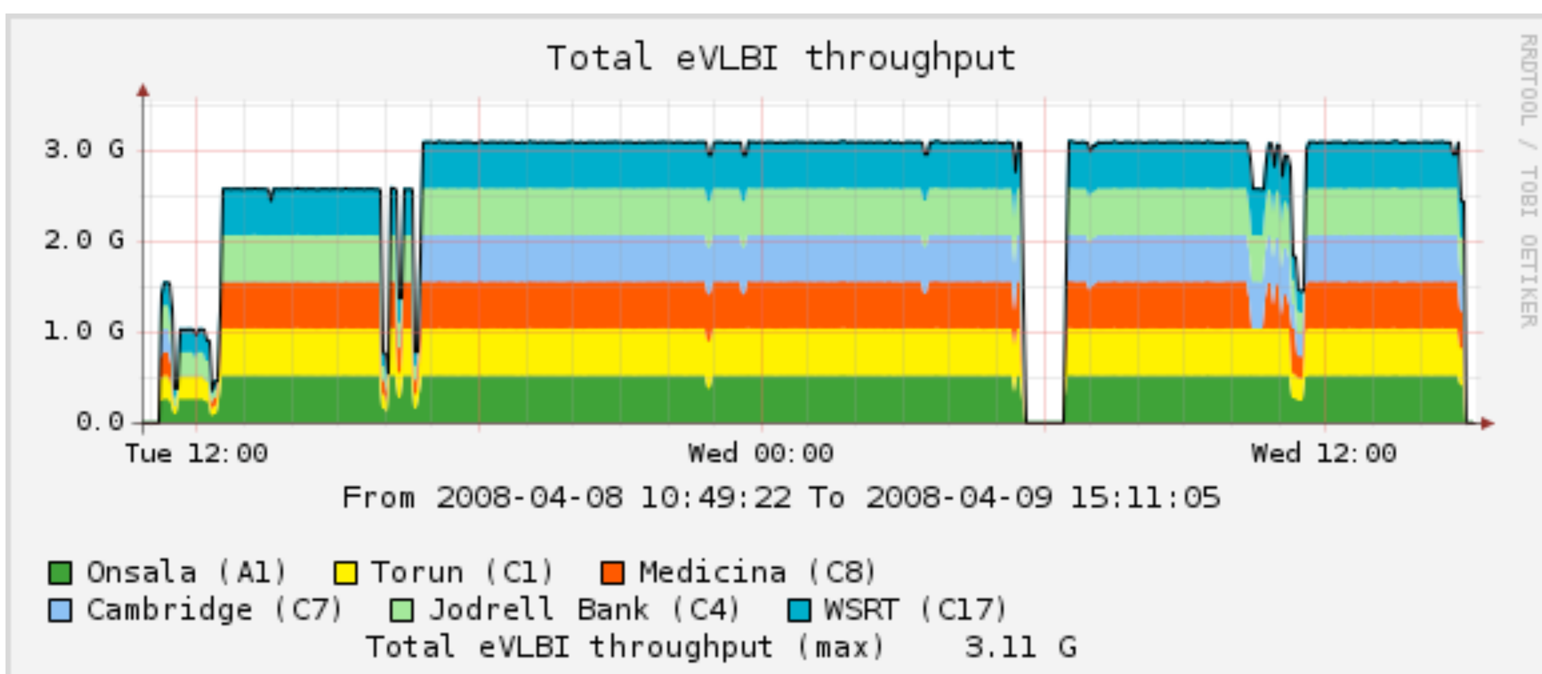
2008-02-05: 6x 128Mb/s

Test with Arecibo



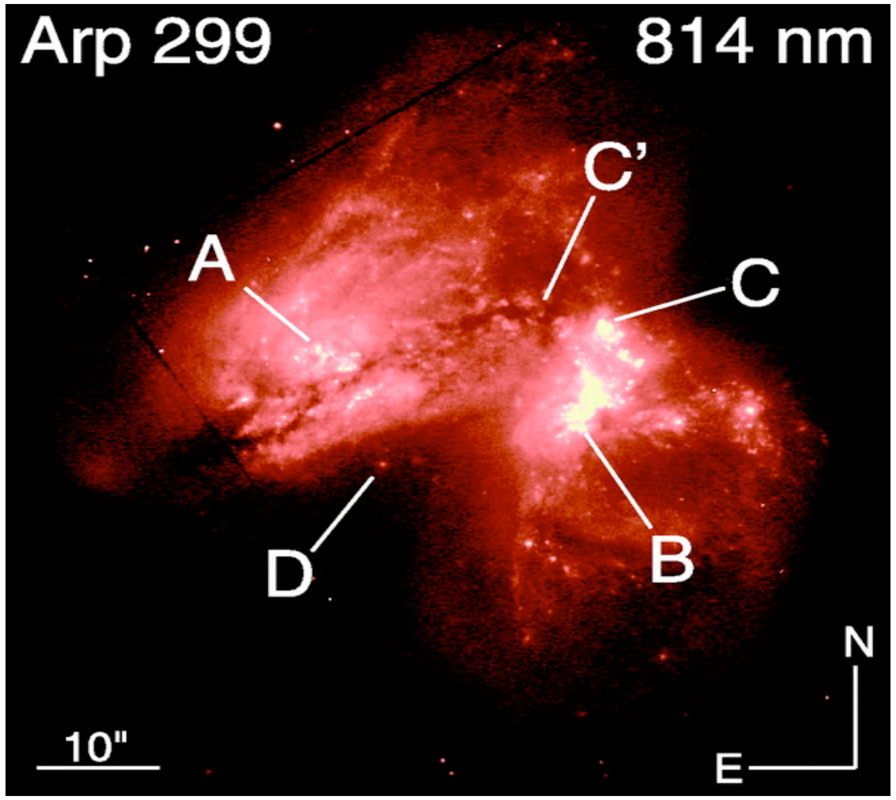
2008-04-01: 7x 970Mb/s

Test with Effelsberg



Yesterday: 6x 512Mb/s: 2 observations of 12 hours

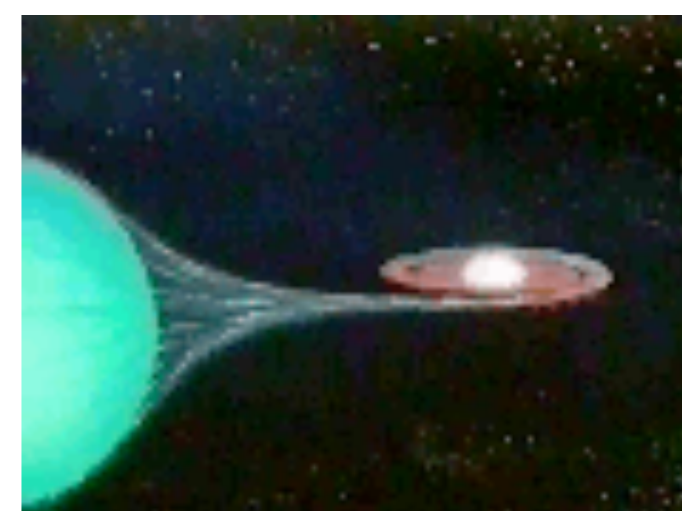
Yesterday's observations



Neff e.a., ©2004 American Astronomical Soc.

- Two colliding galaxies
- Many new large stars formed
- Large stars burn up quickly
- And turn into supernovae
- Possibly several / year
- Follow-up observations

- Xray - binary (on watchlist)
- Star + neutron star / black hole
- Matter from star falls onto companion
- Causes flare
- 'Triggered' observation
- Rapid response use of e-VLBI

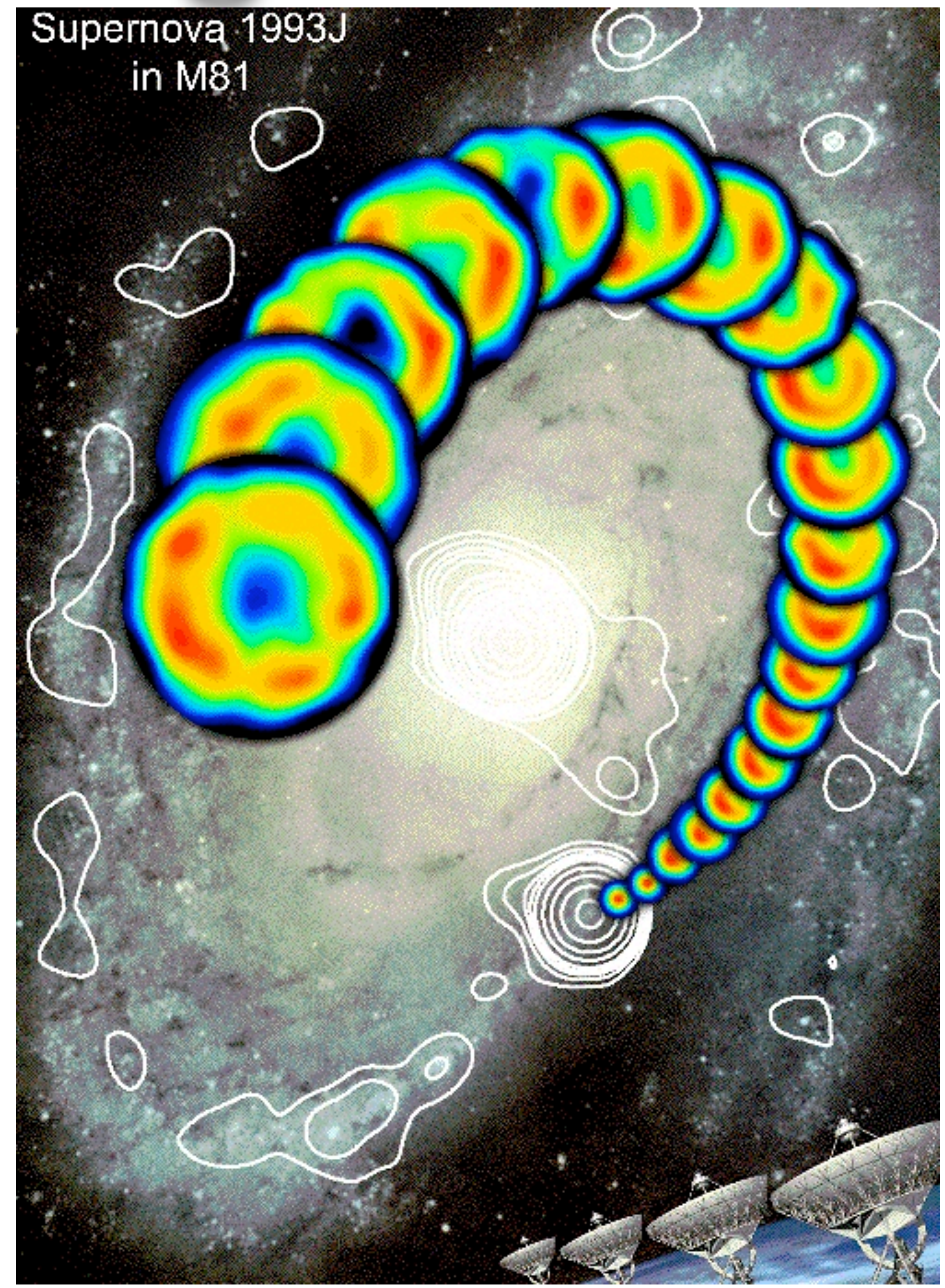

Pat Tyler / NASA

Future challenges

- New telescopes
  - Yebes, Spain
  - Sardinia, Italy
  - VSOP (Space)

- Telescopes in unconnected places
  - Hartebeesthoek, South Africa
  - Urumqi, China
  - Noto, Italy

- Higher bandwidths
  - This requires a new correlator...
    - 4 Gb/s with new telescope backends
    - 30 Gb/s - Merlin (UK)




jive
JOINT INSTITUTE FOR VLBI IN EUROPE


jive
JOINT INSTITUTE FOR VLBI IN EUROPE
Questions?