



Towards 400 Gbps throughput measurements

Raul Lopes

Jisc

2026-05-07

Motivation



Achieve 80% of available throughput with

- Not very expensive architecture
 - architecture bases on dual 8 cores Intel processors
- Using RDMA
 - qperf and perftools
- present perfSONAR tools
 - iperf, iperf3
- Over-designed (Corporate templates)

Motivation



Achieve 80% of available throughput with

- Not very expensive architecture
 - architecture bases on dual 8 cores Intel processors
- Using RDMA
 - qperf and perftools
- present perfSONAR tools
 - iperf, iperf3
- Over-designed (Corporate templates)

The solution:

- `ib_Write_rdma`
- iperf3 over four Connect6-X, each connected at 100Gbps

Get it out of the way

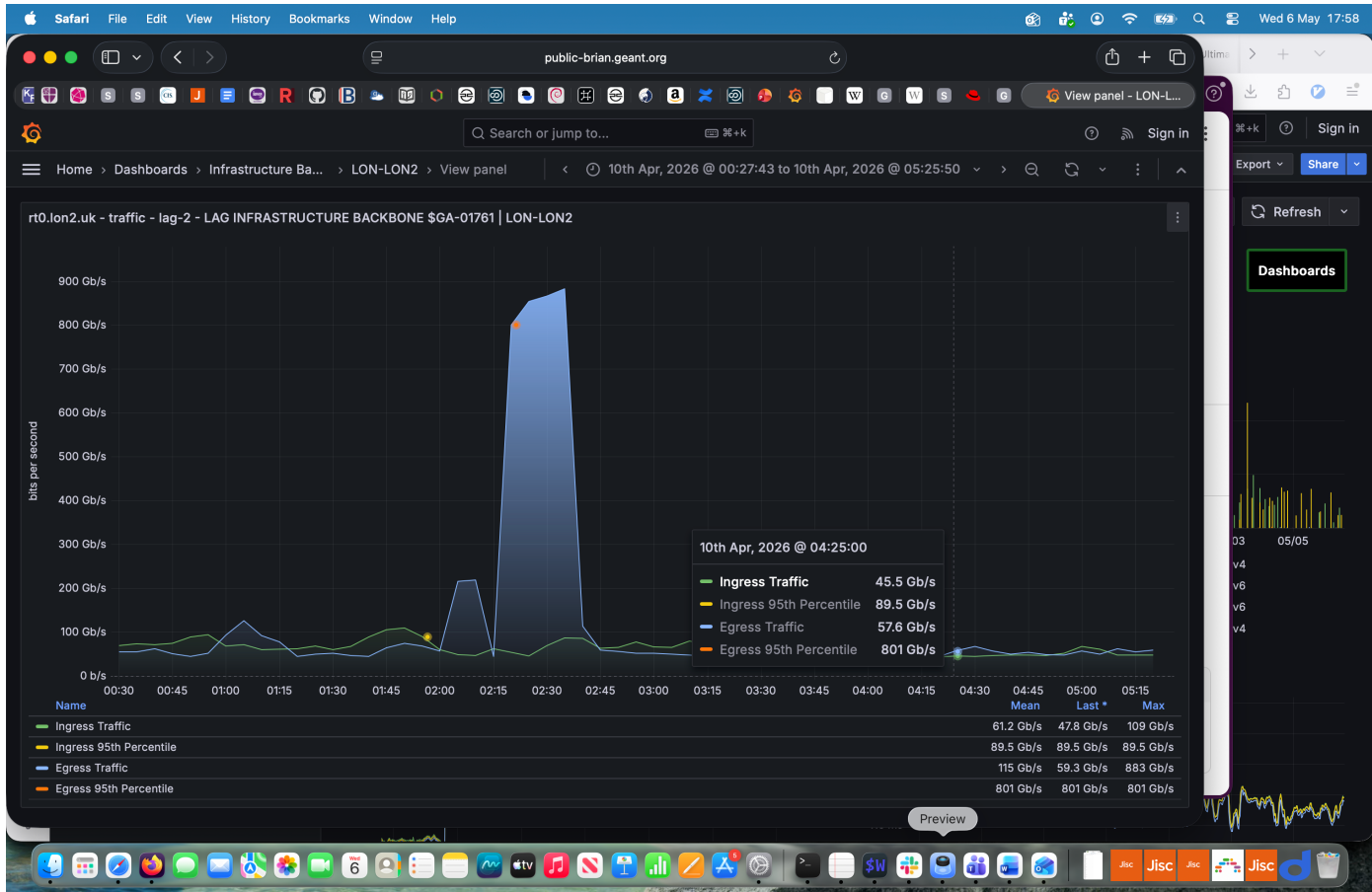


Figure 1: 790Gbps on RTT=29.5ms (test run by Niall Donaghy)

The architecture



- two 800G-capable servers
 - dual Intel Xeon 6714P
 - 16 cores per server (Hyperthreading: 32)
 - Memory: 256 GB RAM
 - two Connect7-X at 400Gbps each

The architecture



- two 800G-capable servers
 - dual Intel Xeon 6714P
 - 16 cores per server (Hyperthreading: 32)
 - Memory: 256 GB RAM
 - two Connect7-X at 400Gbps each
- two 400G-capable servers
 - dual Intel Xeon 6507P
 - 16 cores per server (Hyperthreading: 32)
 - Memory: 128GB RAM
 - four Connect6-X at 100GB each

The tools using one Connect7-X



	iperf3	qperf	ib_write_rdma
RTT=0.3ms	380Gbps	28Gbps	398Gbps
RTT=29.5ms	289Gbps	22Gbps	395Gbps

Table 1: The tools and their peak performance

Summary of that table



- Results with qperf
 - 28Gbps on RDMA and less than 10Gbps on TCP (Possibly this operator's failure)
 - Results ib_write_rdma
 - On RTT=29.5 ms
 - Tests run by Raul and Niall several times
 - One of Raul's tests showed sustained 395Gbps for one hour registered on Brian

The set of tests



Tests were run trying to achieve at least

- 320 Gbps in each Connect7-X
- 80 Gbps in each Connect6-X

Focus on iperf3 and ib_write_rdma

- low performance obtained with qerf
- iperf was also used in early stages, but the performance was best equal to iperf3

Tests with four different TCP stack tunings (via sysctl)



- Stacks tested by Raul: default install, Red Hat, Fasterdata, Cloudflare
- three CWA algorithms: reno, cubic, bbr
- Best TCP stack tested by Raul
 - based Network white papers by Red Hat and Cloudflare
 - validated in Jisc for perfSONAR and Globus DTN

Tests with default install



```
table( columns: (auto, auto, auto, auto, auto) inset: 10pt, align: center +  
horizon,
```

```
[streams], [16], [32], [64], [128]
```

```
[default], [29.2Gbps], [56Gbps], [113bps],[227Gbps] [Jisc], [186],  
[256], [269],[279]
```

```
), caption: [Median throughput rates (Gbps) ], )
```

- Numbers for 12 runs of 5 seconds each
- bbr did not show gains over cubic
- we tested bbr in the Jisc host: big gains compared to: cubic, reno, http

Changes in Jisc stack



net.ipv4.tcp_congestion_control net.ipv4.tcp_rmem net.ipv4.tcp_wmem
net.ipv4.tcp_adv_win_scale net.ipv4.tcp_notsent_lowat
net.ipv4.tcp_no_metrics_save net.ipv4.tcp_slow_start_after_idle
net.core.netdev_max_backlog net.core.netdev_budget
net.core.netdev_budget_usecs net.core.rmem_max net.core.wmem_max
net.core.optmem_max net.core.default_qdisc net.core.qdisc_max_burst
txqueuelen

perfSONAR stack: fasterdta (left of the plot) vs Jisc configura

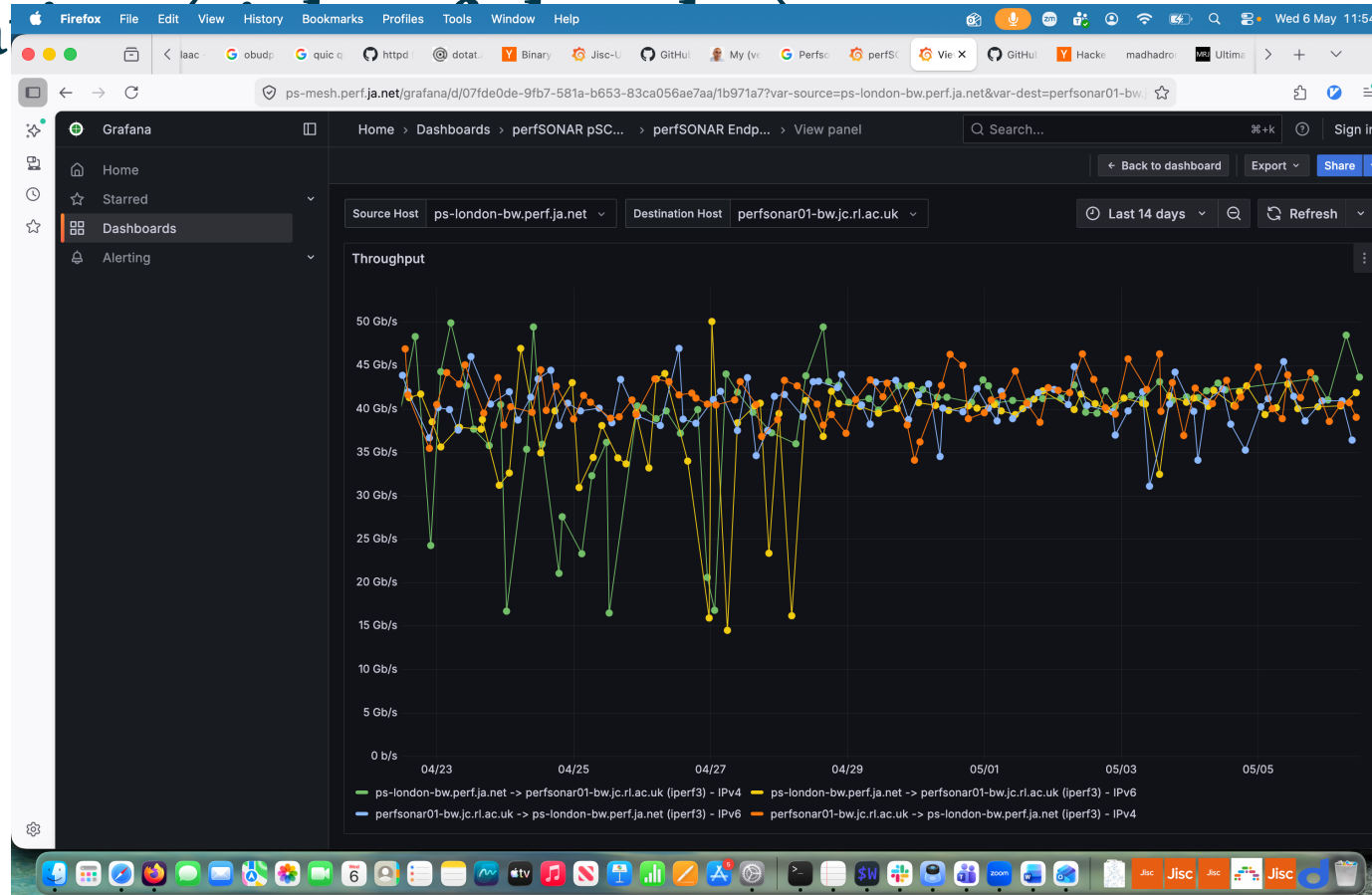


Figure 2: Jisc to Jasmine host at RAL

Comment



- Median values
- Reason: Large TCP buffers
 - fasterdata: 2GB
 - Jisc: 512MB
 - preferred by RH: 1 GB

Next test



- 1.1 Tbps with 4 servers
 - easy for RDMA
 - It may need a FReeBSD stack

Questions

