ESnet

ENERGY SCIENCES NETWORK

Switch Buffers Experiments: How much buffer do you need to support 10G flows?

Michael Smitasin (mnsmitasin@lbl.gov), Lawrence Berkeley National Laboratory
Brian L Tierney (bltierney@es.net), ESnet

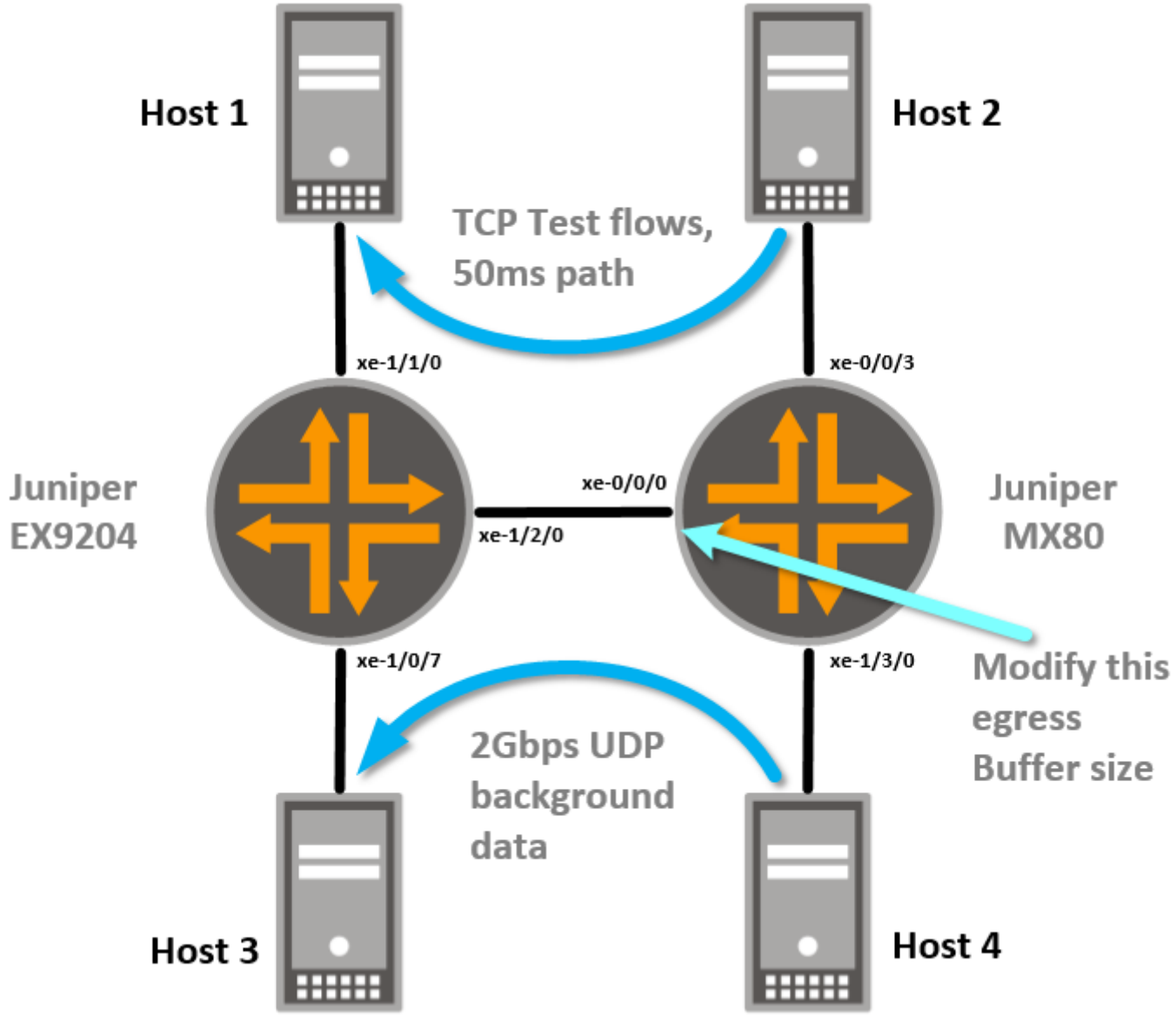
2014 Technology Exchange, Oct 29, 2014

U.S. DEPARTMENT OF ENERGY

Office of Science

BERKELEY LAB

Buffer Experiment #1: Juniper MX80



Experiment #1 Setup

- Try various buffer size on Juniper MX80 using 'scheduler-map'
- Maximum queue buffer = 125MB
- 2Gbps UDP background traffic from host 4 to host 3; 9000 B MTUs
- Added 50ms latency to TCP flows from host 2 using 'tc':
  - tc qdisc add dev ethN root netem delay 50ms
-
- Paced traffic using 'tc', eg:
- `tc qdisc add dev ethN handle 1: root htb`
- `tc class add dev ethN parent 1: classid 1:1 htb rate 7gbit`
- `tc qdisc add dev ethN parent 1:1 handle 10: netem delay 50ms`
- `tc filter add dev ethN protocol ip prio 1 u32 match ip dst 10.1.2.0/24 flowid 1:1`

Results: Small Buffers Kill Performance!

Buffer Size	Packets Dropped	TCP Throughput
120 MB	0	8Gbps
60 MB	0	8Gbps
36 MB	200	2Gbps
24 MB	205	2Gbps
12 MB	204	2Gbps
6 MB	207	2Gbps

30 Second test, 2 TCP streams

Results with Pacing, 12MB Egress Buffer
(note: many low cost switches have 9MB buffers)

2Gbps background traffic

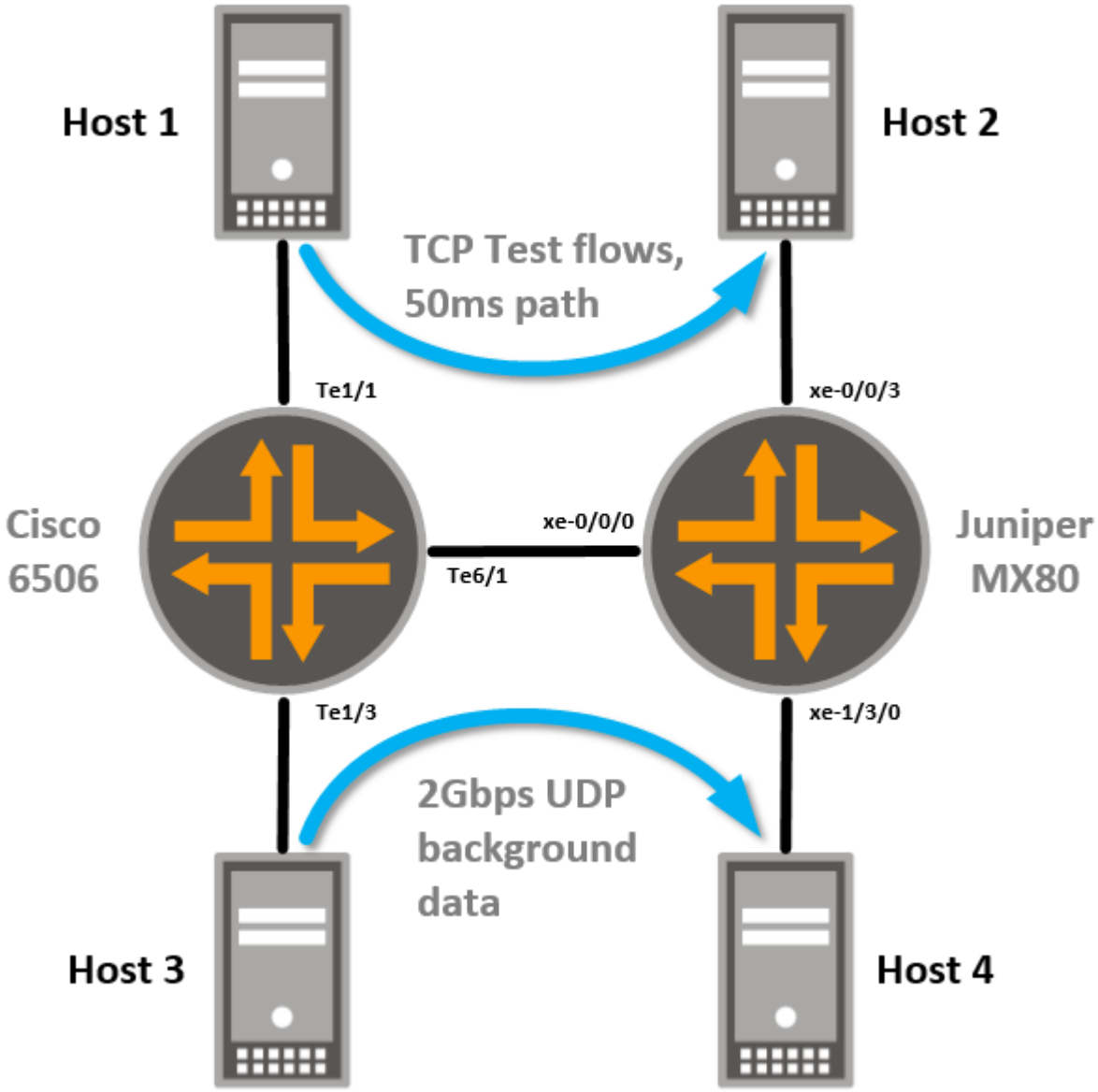
Paced Send Rate	Packets Dropped	TCP Throughput
3 Gbps	0	8 Gbps
4 Gbps	0	8 Gbps
5 Gbps	0	8 Gbps
6 Gbps	212	1.5 Gbps

3Gbps background traffic

Paced Send Rate	Packets Dropped	TCP Throughput
3 Gbps	0	8 Gbps
4 Gbps	0	8 Gbps
5 Gbps	53	2.5 Gbps
6 Gbps	203	1.5 Gbps

ESnet

Buffer Experiment #2: Cisco 6506



Experiment #2 Setup

- Try various 'hold queue' buffer size on Cisco 6506 Maximum queue buffer = 125MB
- 2Gbps UDP background traffic from host 3 to host 4, 9000 Byte MTU
- Added 50ms latency to TCP flows from host 1 to host 2 using 'tc':
  - tc qdisc add dev ethN root netem delay 25ms (on both sender and receiver)
-
- Paced sender traffic using 'tc', eg:
- `tc qdisc add dev ethN handle 1: root htb`
- `tc class add dev ethN parent 1: classid 1:1 htb rate 7gbit`
- `tc qdisc add dev ethN parent 1:1 handle 10: netem delay 25ms`
- `tc filter add dev ethN protocol ip prio 1 u32 match ip dst 10.1.2.0/24 flowid 1:1`

Results for 6506

Hold Queue Size	Packets Dropped	TCP Throughput
2000 packets (18 MB)	700-800	5 Gbps
4096 packets (36 MB)	250-550	8.5 Gbps

With Pacing; hold-queue = 4096

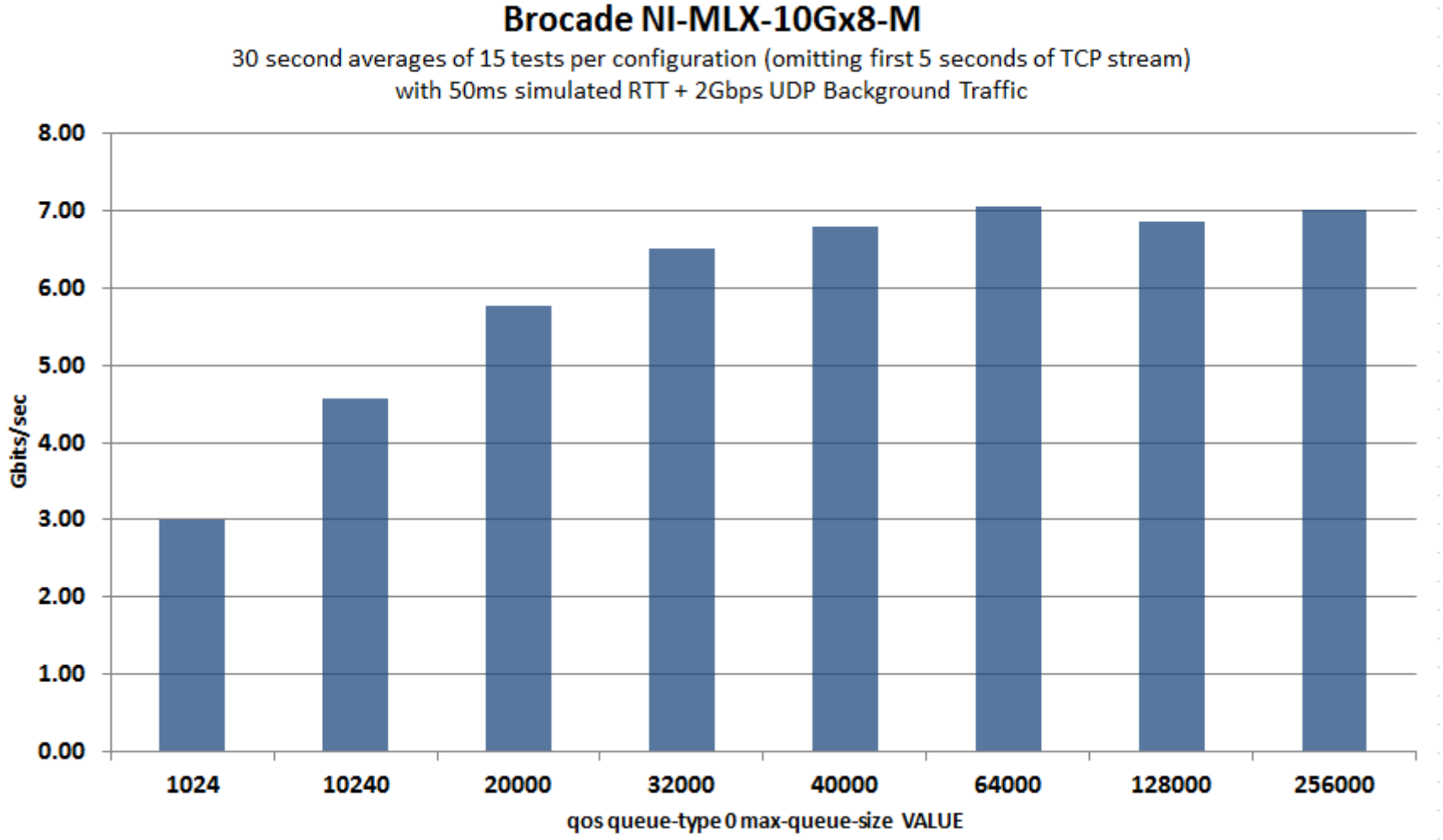
Paced Send Rate	Packets Dropped	TCP Throughput
5 Gbps	5-10	8 Gbps
6 Gbps	200-400	8 Gbps
7 Gbps	200-400	8 Gbps
8 Gbps	200-400	1.5 Gbps

Conclusion: 32MB buffer is not enough, but pacing helps

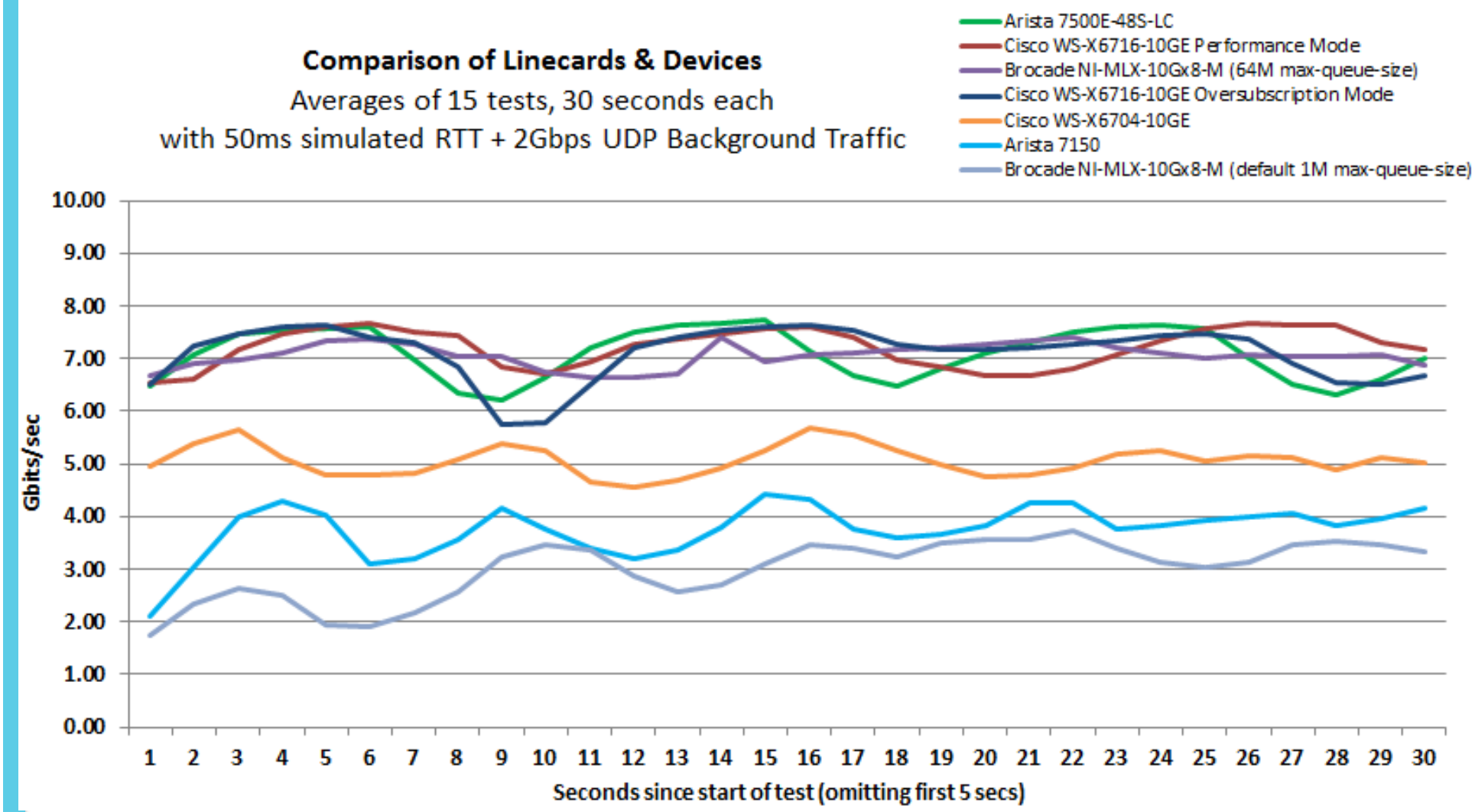
Brocade MLX results

Configuration:

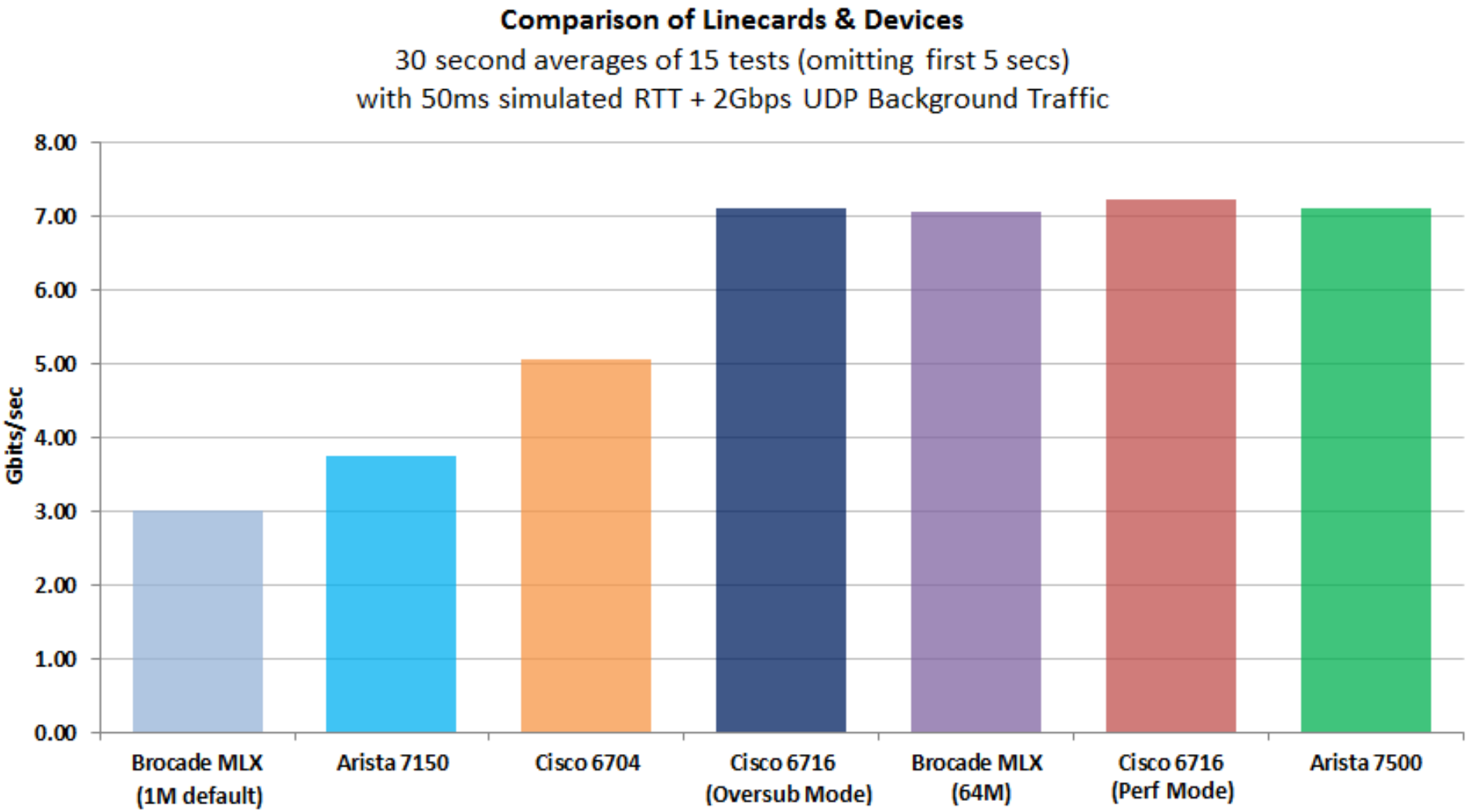
```
default-max-frame-size 9212
qos queue-type 0 max-queue-size VALUE
```



Full LBL test results



Full LBL test results



More Information

- Page describing test methodology:
  - http://fasterdata.es.net/network-tuning/router-switch-buffer-size-issues/switch-buffer-testing/
- Page with known device buffer sizes:
  - http://people.ucsc.edu/~warner/buffer.html
- email: mnsmitasin@lbl.gov, BLTierney@es.net