

*Hands-on evaluation
of new routers and switches
(activity 9.7)*

*Juniper Networks TX Matrix testing
(including STM-256 card and DWDM PIC)*

TF-NGN

Ljubljana, 4-5 July 2006

Marcin Garstka

marcinga@man.poznan.pl



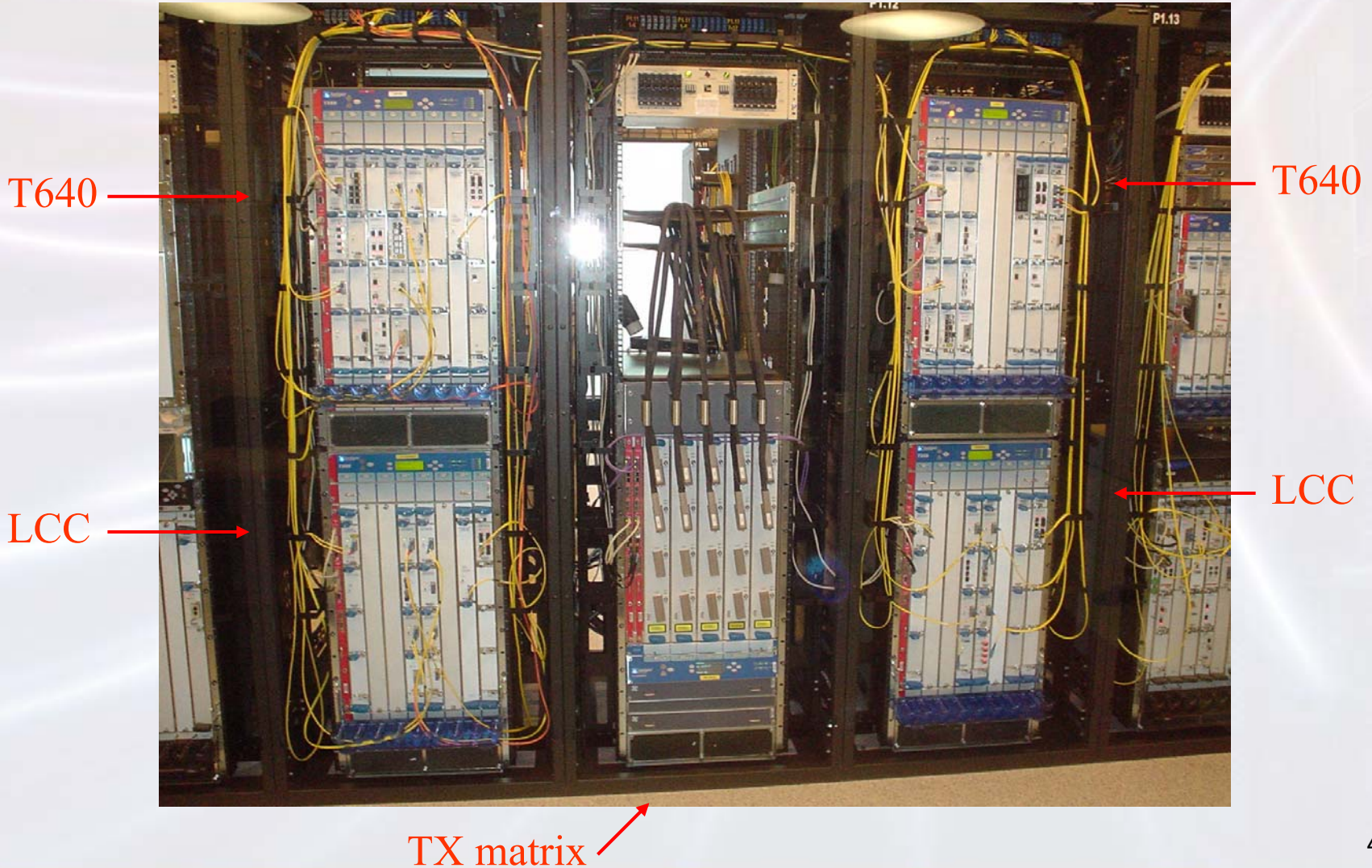
Motto:

To prove that TX is really a single
router in multiple chassis.

Juniper Networks TX Matrix testing:

- 20-21 June 2006
- Amsterdam, Juniper POC lab
- Participants:
 - Marcin Garstka (PSNC)
 - Josef Voytech (CESNET)
 - Carsten Rosche (Fraunhofer-IMK Sankt Augustin)
 - Cătălin Meiroșu (Terena)
 - Jean-Marc Uze (Juniper)
 - Holger Metschulat (Juniper)
 - Andre Stiphout (Juniper POC lab)

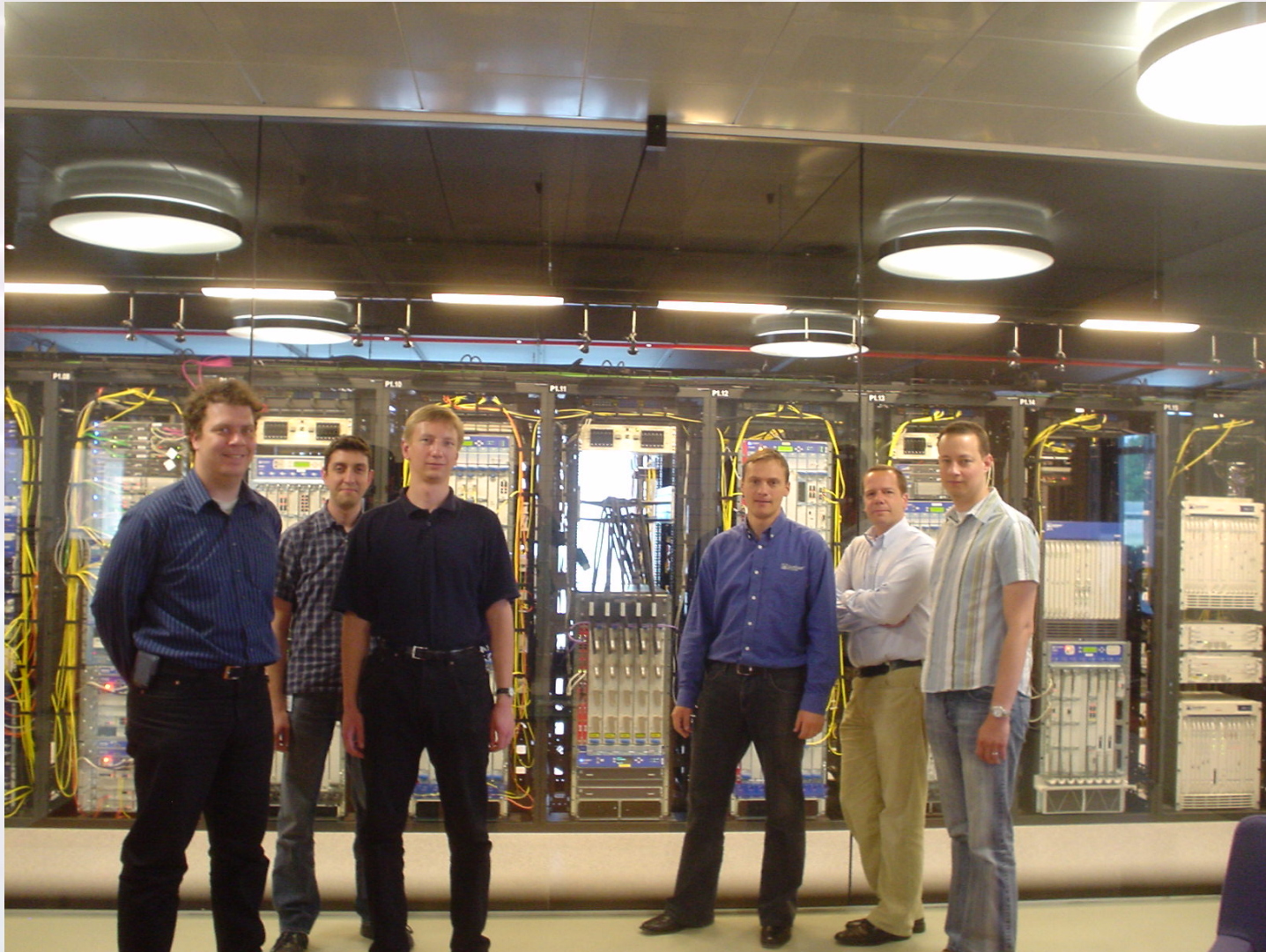
How our TX looked like



How hard we worked



Our group with TX



Test 1 – DWDM 10GE PIC

- DWDM 10 GE (LAN PHY) PIC supports 45 wavelengths (full C-band)
- Switching between wavelengths
- ISIS adjacency brought down for 3 seconds when changing wavelengths



Test 1 – DWDM 10GE PIC – configuration

```
andre@matrix# set interfaces ge-0/0/0 optics-options wavelength 1549.32
```

```
[edit]
```

```
andre@matrix# show interfaces ge-0/0/0
```

```
description matrix-jupiter-DWDM;
```

```
optics-options {
```

```
    wavelength 1549.32;
```

```
}
```

Test 1 – DWDM 10GE PIC – monitoring

```
andre@matrix# run show interfaces ge-0/0/0
```

```
Physical interface: ge-0/0/0, Enabled, Physical link is Up
```

```
(...)
```

```
Wavelength      : 1549.32 nm, Frequency: 193.50 THz
```

```
(...)
```

```
andre@matrix> show interfaces diagnostics optics ge-0/0/0
```

```
Physical interface: ge-0/0/0
```

```
Laser bias current      : 83.920 mA
```

```
Laser output power      : 1.580 mW / 1.99 dBm
```

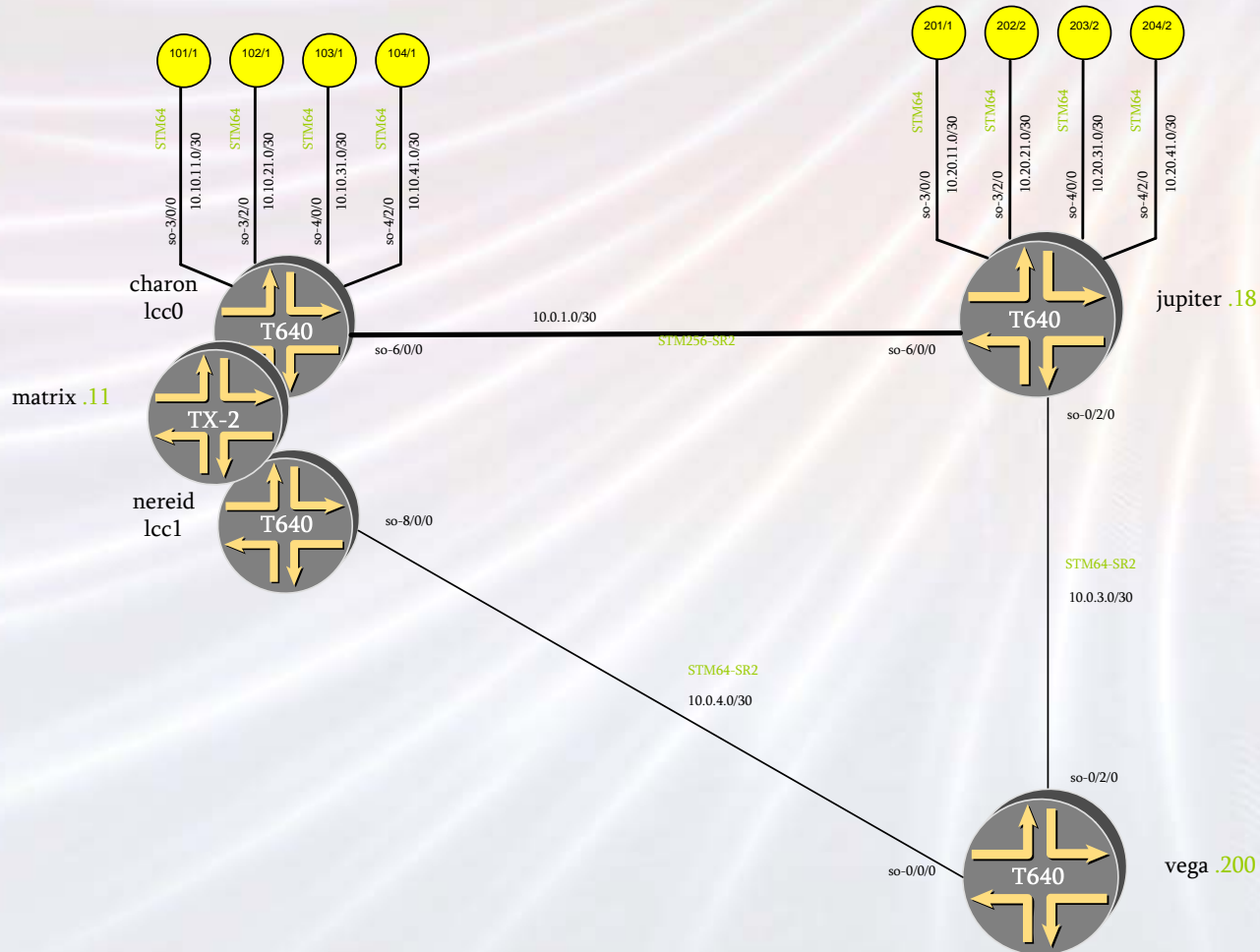
```
Receiver signal average optical power : 0.1274 mW / -8.95 dBm
```

```
(...)
```

Test 2 – STM-256 card test

- Eight 10Gps streams generated from Agilent testers, transmitted over STM-256 link between two routers (four streams in one direction, four in the other)
- IPv4 test with packet size 40B (said to be the worst case for many routers)
- IPv6 test with packet size 60B
- IPv4 in MPLS test with packet size 40B

Test 2 – lab setup



Test 2 – sample results (IPv4 test)

The screenshot shows the 'Setup - Traffic' window with a list of traffic flows. The 'Results - Realtime' window displays a table of performance metrics for various traffic flows. Two red circles highlight the 'Rx Packet Loss' column, which shows '0' for all flows, indicating no packet loss. A red box highlights the 'Tx Test Packets' and 'Rx Test Packets' columns, which show identical values for each flow, indicating that all sent traffic was received.

Port	Tx Test Packets	Rx Test Packets	Tx Test Octets	Rx Test Octets	Tx Test Throughput (Mb/s)	Rx Test Throughput (Mb/s)	Rx Packet Loss	Average Latency (us)	Minimum Latency (us)	Maximum Latency (us)	Misdirected Packets	Sequence Errors	HQ/C/R Transmit Efficiency
204/1	397531321	397531200	19081503408	19081497600	8480.67	8480.67	n/a	45.36	31.91	52.07	0	n/a	99.74
104/1->204/1, TrafficMesh 35/7	397531200	397531200	19081497600	19081497600	8480.67	8480.67	0	45.36	31.91	52.07	n/a	0	n/a
201/1	397531368	397531184	19081505664	19081496832	8480.67	8480.67	n/a	44.70	32.24	50.00	0	n/a	99.74
101/1->201/1, TrafficMesh 35/1	397531184	397531184	19081496832	19081496832	8480.67	8480.67	0	44.70	32.24	50.00	n/a	0	n/a
101/1	397531184	397531368	19081496832	19081505664	8480.67	8480.67	n/a	52.22	32.45	60.68	0	n/a	99.74
201/1->101/1, TrafficMesh 35/2	397531368	397531368	19081505664	19081505664	8480.67	8480.67	0	52.22	32.45	60.68	n/a	0	n/a
203/1	397531101	397531289	19081492848	19081501872	8480.66	8480.67	n/a	48.57	32.60	58.41	0	n/a	99.74
103/1->203/1, TrafficMesh 35/5	397531289	397531289	19081501872	19081501872	8480.67	8480.67	0	48.57	32.60	58.41	n/a	0	n/a
202/1	397531441	397531033	19081509168	19081489584	8480.67	8480.66	n/a	45.21	32.90	51.51	0	n/a	99.74
102/1->202/1, TrafficMesh 35/3	397531033	397531033	19081489584	19081489584	8480.66	8480.66	0	45.21	32.90	51.51	n/a	0	n/a
104/1	397531200	397531321	19081497600	19081503408	8480.67	8480.67	n/a	48.92	33.00	56.64	0	n/a	99.74
103/1	397531289	397531101	19081501872	19081492848	8480.67	8480.66	n/a	48.50	33.00	58.75	0	n/a	99.74
203/1->103/1, TrafficMesh 35/6	397531101	397531101	19081492848	19081492848	8480.66	8480.66	0	48.50	33.00	58.75	n/a	0	n/a
204/1->104/1, TrafficMesh 35/8	397531321	397531321	19081503408	19081503408	8480.67	8480.67	0	48.92	33.00	56.64	n/a	0	n/a
102/1	397531033	397531441	19081489584	19081509168	8480.66	8480.67	n/a	49.46	33.12	58.29	0	n/a	99.74
202/1->102/1, TrafficMesh 35/4	397531441	397531441	19081509168	19081509168	8480.67	8480.67	0	49.46	33.12	58.29	n/a	0	n/a

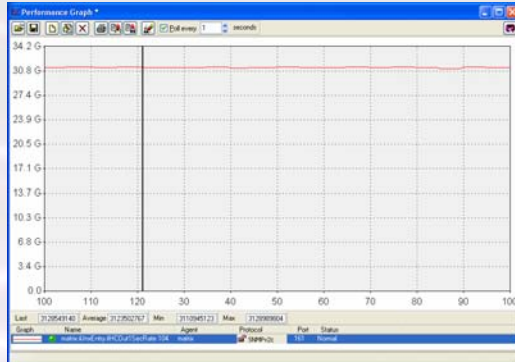
No packet loss

No packet loss

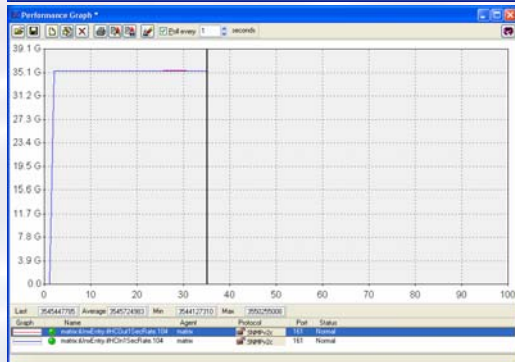
Sent and received traffic

Test 2 – sample graphs

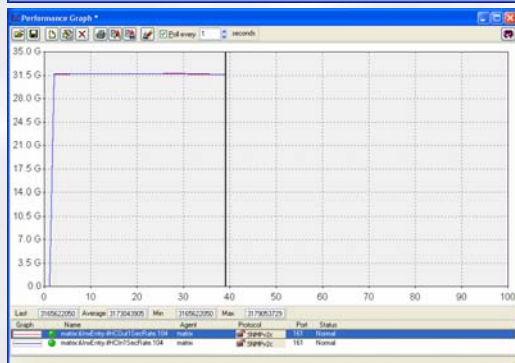
IPv4



IPv6



MPLS



Graphs generated by MG-SOFT from statistics gathered from one of the routers (statistics in one direction)

All statistics on IP or MPLS layer (without PPP/HDLC headers)

Max bandwidth in theory:

- Sonet (with overheads) – 39,813Mbps
- For IPv4 (40B packets) – 31,296 Mbps
- For IPv6 (60B packets) – 35,486 Mbps
- For MPLS (40B packets) – 31,828 Mbps

Test 2 – with filtering etc

Tests with TCP packets, variable packet size (60-256B), variable TCP ports
100,000 BGP prefixes inserted from each Agilent (800,000 prefixes total)

Tests:

- Packet filtering (1000 terms, applied in and out, only negative hits, filtering on IP addresses)
- Packet counting (1000 counters, counting on TCP ports)
- Filter based forwarding (1000 terms with only negative hits, last hit positive, filtering on IP addresses)

Results:

- No impact on bandwidth
- A few microsecond increase of latency

Test 2 – Juniper's worst case

Juniper routers chop each packets into J-cells, each J-cell is 64B long

First J-cell: 12B overhead, 52B data

Each next J-cell: 2B overhead, 62B data

Juniper's worst case – packet size of 53B (each packet consumes two J-cells, the second one carries only 1B of user data)

For 53B packets the STM-256 card is able to transmit approx 24,800Gbps (66% of line rate)

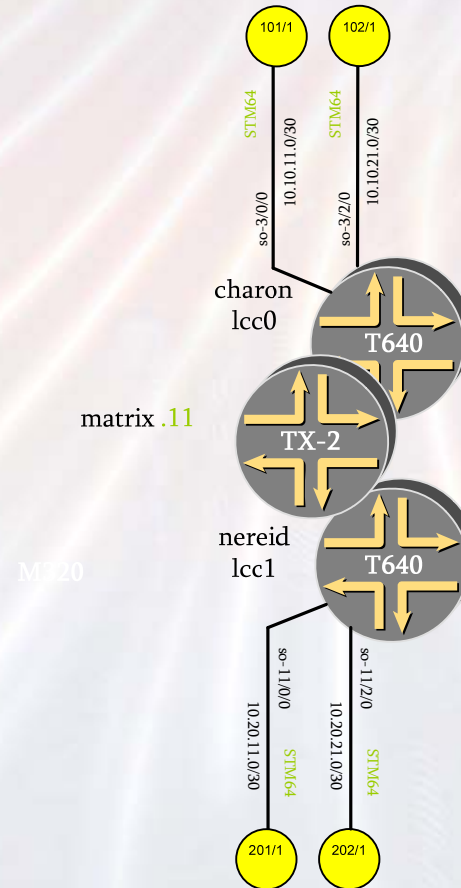
This problem will not occur in real life – we do not expect to have a 40 Gig/s interface fully loaded by precisely 53 bytes packets. Also 66% (> more than 50%) shows the headroom available in the switching fabric. This effect decreases with larger packet sizes, and there is no effect with any mix of packet-sizes.

Test 2 – results

- STM-256 card tested with IPv4, IPv6 and MPLS packets
- Line rate even for short packets
- Packet filters, packet counters and policy based routing do not affect bandwidth

Test 3 – TX reliability

- Failover between TX Routing Engines
- LCC (Line Card Chassis) switching off and on



Test 3 – TX reliability

- The active Routing Engine was removed from the TX
- The router switched to the other Routing Engine
- Packet loss was monitored during the switchover
- No packet loss detected – non-stop forwarding
- Graceful RE switchover (graceful restart must be supported by adjacent/neighbouring routers)
- All routing protocols adjacencies went down and were re-established on the other Routing Engine
- Forwarding plane was not affected – non-stop forwarding

Test 3 – TX reliability

Result – switching and LCC off and on does not affect traffic which does not passes this LCC (no packet loss detected)

Results - Realtime

Port	Tx Test Packets	Rx Test Packets	Tx Test Octets	Rx Test Octets	Tx Test Throughput (Mbps)	Rx Test Throughput (Mbps)	Rx Packet Loss	Average Latency (us)
202/1->101/1, TrafficMesh 14/1	290944445	290944445	18620444480	18620444480	283.74	283.74	0	19.75
202/1->101/1, TrafficMesh 14/2	290944445	290944445	18620444480	18620444480	283.74	283.74	0	19.87
202/1->101/1, TrafficMesh 14/3	290944445	290944445	18620444480	18620444480	283.74	283.74	0	19.94
202/1->101/1, TrafficMesh 14/4	290944445	290944445	18620444480	18620444480	283.74	283.74	0	19.97
202/1->101/1, TrafficMesh 14/5	290944445	290944445	18620444480	18620444480	283.74	283.74	0	19.97
202/1->101/1, TrafficMesh 14/6	290944445	290944445	18620444480	18620444480	283.74	283.74	0	19.99
202/1->101/1, TrafficMesh 14/7	290944445	290944445	18620444480	18620444480	283.74	283.74	0	19.99
202/1->101/1, TrafficMesh 14/8	290944445	290944445	18620444480	18620444480	283.74	283.74	0	20.11
202/1->101/1, TrafficMesh 14/9	290944444	290944444	18620444416	18620444416	283.74	283.74	0	20.18
202/1->201/1, TrafficMesh 14/10	0	18620444416	0	18620444416	0	283.74	0.00	290944444
202/1->201/1, TrafficMesh 14/11	0	18620444416	0	18620444416	0	283.74	0.00	290944444
202/1->201/1, TrafficMesh 14/12	0	18620444416	0	18620444416	0	283.74	0.00	290944444
202/1->201/1, TrafficMesh 14/13	0	18620444416	0	18620444416	0	283.74	0.00	290944444
202/1->201/1, TrafficMesh 14/14	0	18620444416	0	18620444416	0	283.74	0.00	290944444
202/1->201/1, TrafficMesh 14/15	0	18620444416	0	18620444416	0	283.74	0.00	290944444
202/1->201/1, TrafficMesh 14/16	0	18620444416	0	18620444416	0	283.74	0.00	290944444
202/1->201/1, TrafficMesh 14/17	0	18620444416	0	18620444416	0	283.74	0.00	290944444
202/1->201/1, TrafficMesh 14/18	0	18620444416	0	18620444416	0	283.74	0.00	290944444

No packet loss during RE switchover

Test 4 – forwarding plane

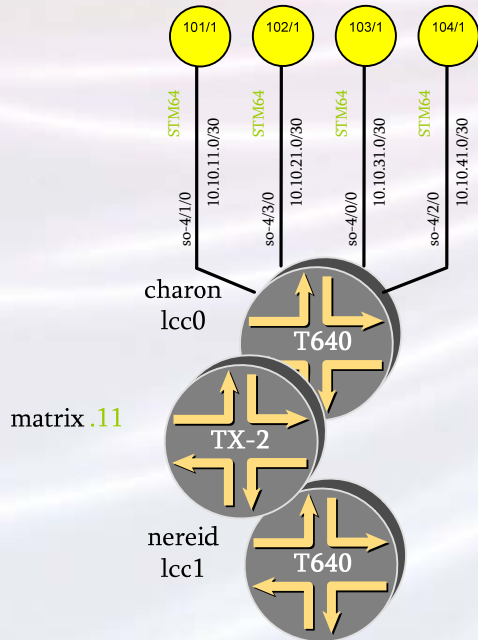
Latency measured when packets are transmitted:

- between two PICs on the same PFE
- between two PICs on different PFEs, the same FPC
- between two PICs on different FPCs, the same line-card chassis
- between two PICs on different line-card chassis

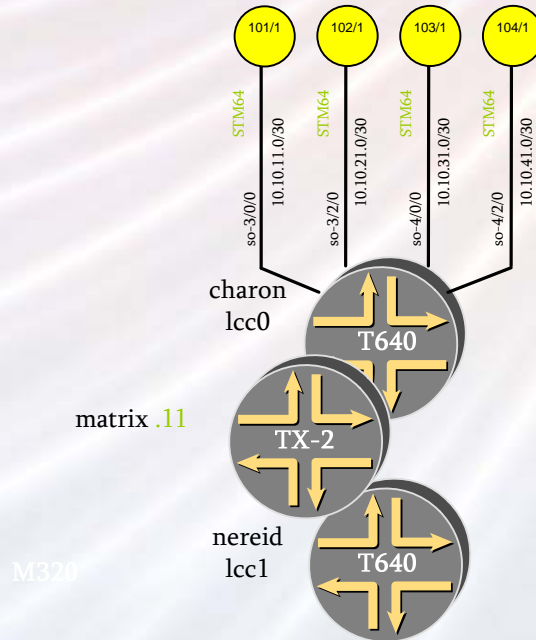
Test for IPv4, IPv6 and MPLS

Test 4 – lab setup

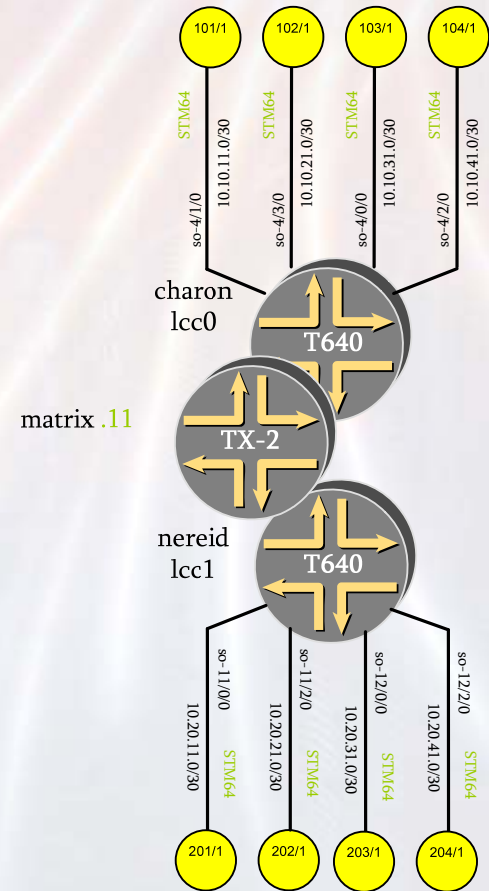
Intra-PFE, Intra-FPC



Inter-FPC



Inter-LCC



Test 4 – forwarding plane

Average latency in microseconds, measured by Agilent testers

	IPv4	IPv6	MPLS
Same PFE	15	14.2	14.6
Different PFEs, same FPC	21.6	24.3	not tested*
Different FPCs, same LCC	21.8	20.7	21.2
Different LCCs	21.3	20.8	21

Results:

- Shorter latency in the ‘Same PFE’ test is as a result of packet processing on PFE (without transmitting to the central switching matrix).
- In other cases latency do not depend on whether packets are received and transmitted on the same LCC or on different LCCs (in all cases packets are processed by the central switching matrix).

* One MPLS test was missed unintentionally.

Test 5 – convergence of routing protocols

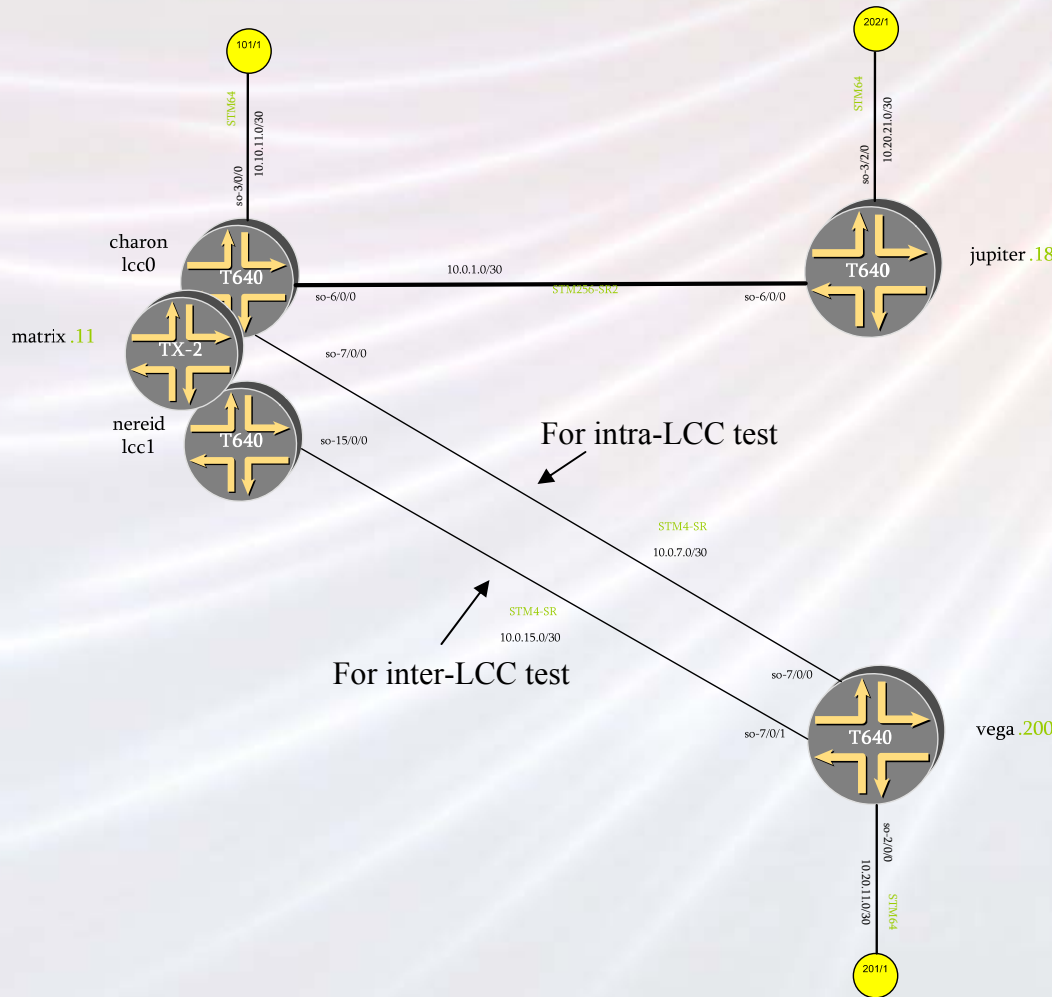
5a – BGP convergence inter- and intra-LCC

5b – ISIS convergence inter- and intra-LCC

5c –indirect next-hop for iBGP

For all tests the protocols have been configured with default behaviours, except spf-delay which was brought down from default 200ms to 50ms (minimum)

Test 5 – BGP convergence



The same prefixes injected from 202/1 and 201/1 (500,000 prefixes /16 – /24)

iBGP between routers, eBGP to Agilent testers

Traffic generated from 101/1 and monitored on 202/1 and 201/1

Matrix has higher LP for jupiter – prefers routes from jupiter and sends traffic to 202/1

Test 5 – BGP convergence

275,000 routes (/24) from 202/1 withdrawn and then readvertised

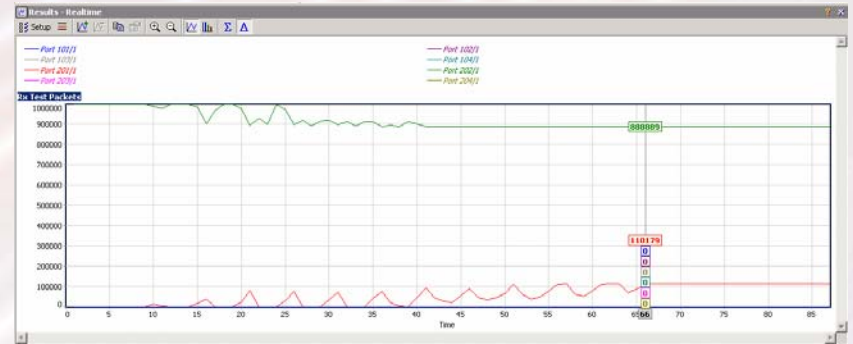
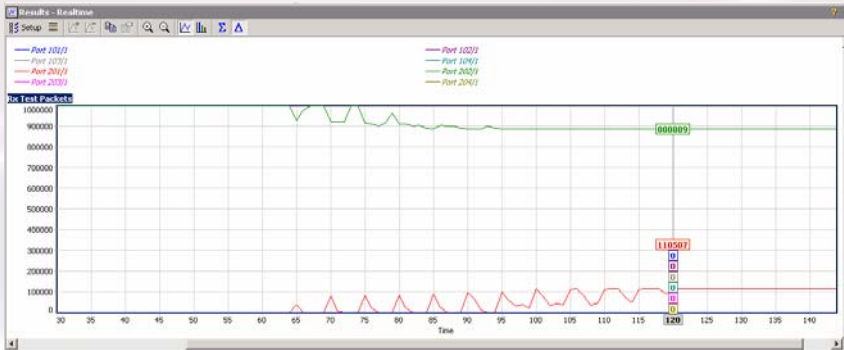
Convergence time measured

Test repeated in inter-LCC and intra-LCC configuration

Test 5 – BGP convergence

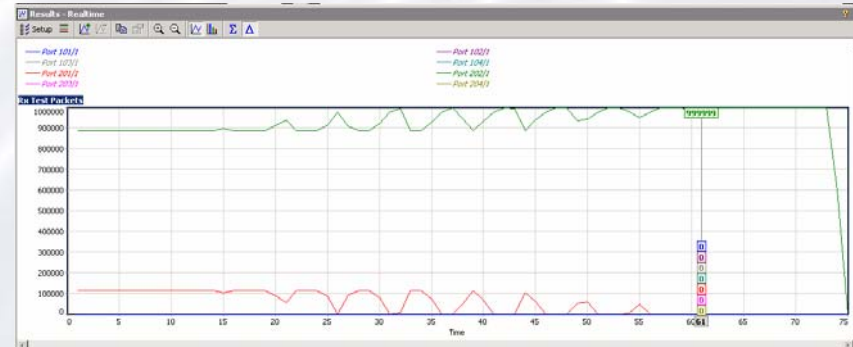
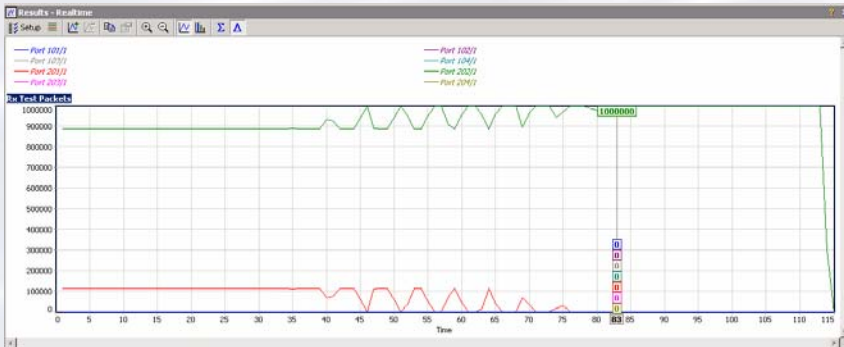
withdraw – single chassis – 56 sec

withdraw – multiple chassis – 57 sec

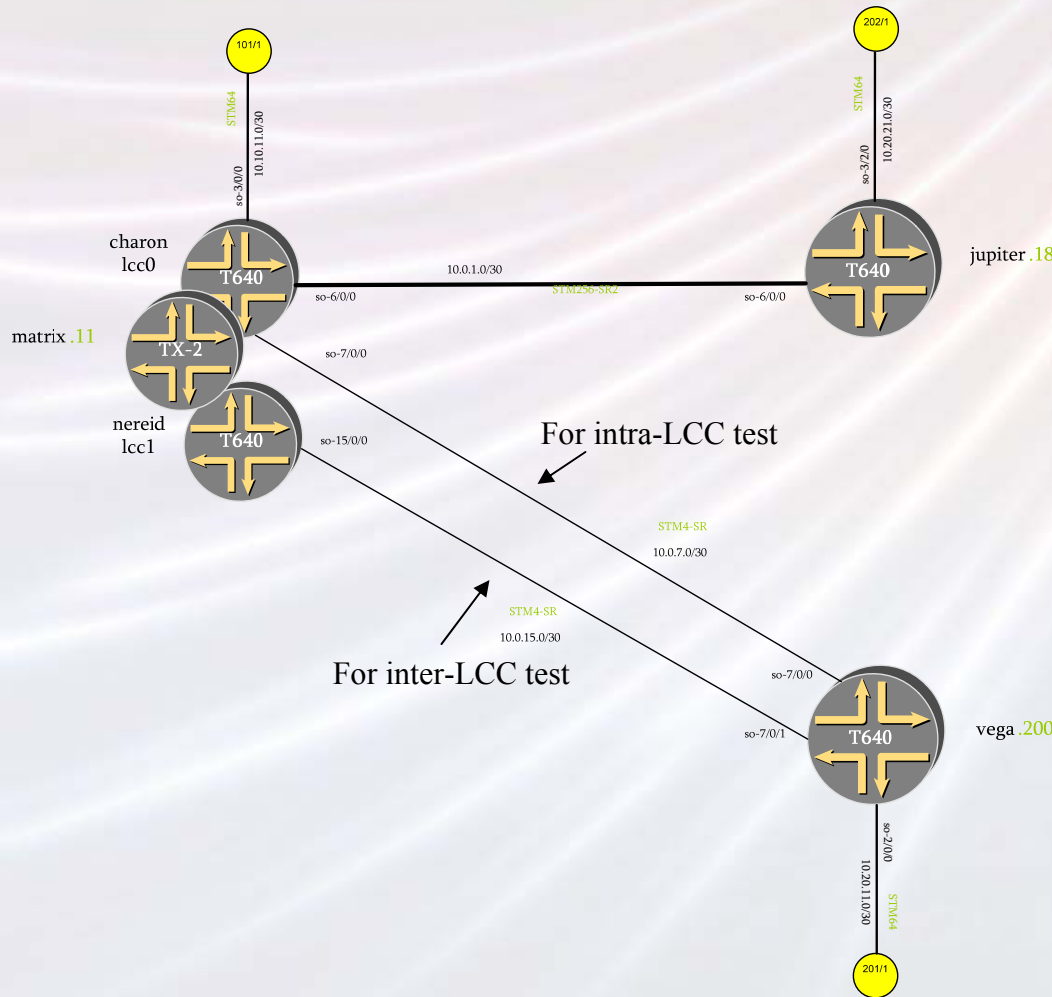


advertise – single chassis – 49 sec

advertise – multiple chassis – 47 sec



Test 5 – ISIS convergence



Testers 202/1 and 201/1 emulate 5 routers with 5000 routes each (the same routes on both testers)

Jupiter and vega prefer routes from locally connected testers

Matrix prefers routes from jupiter

Traffic generated from 101/1 and monitored on 202/1 and 201/1

Matrix sends traffic to 202/1

Test 5 – ISIS convergence

Emulate link failure between vega and tester

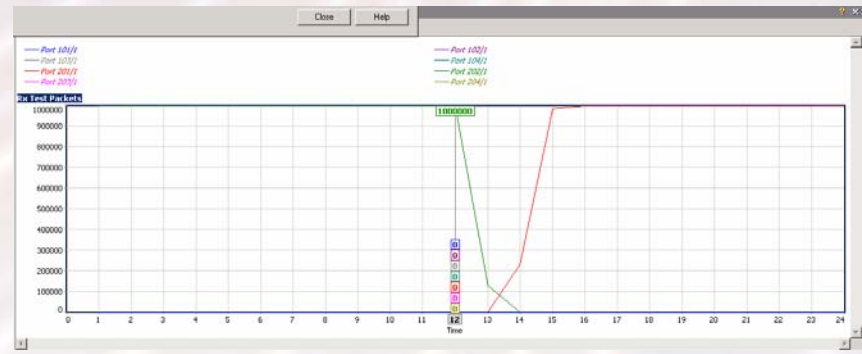
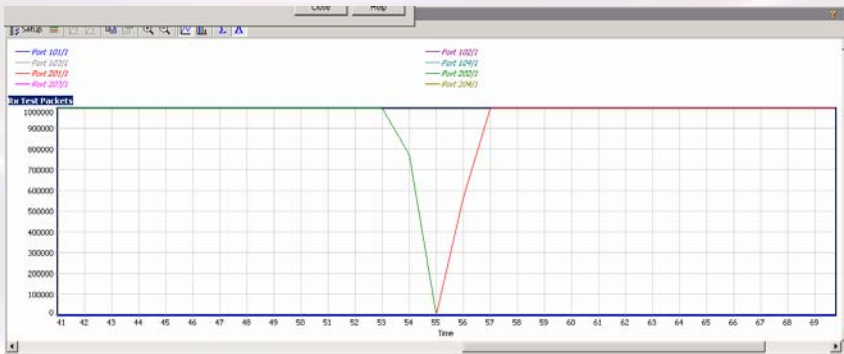
ISIS convergence time measured

Test repeated in inter-LCC and intra-LCC configuration

Test 5 – ISIS convergence

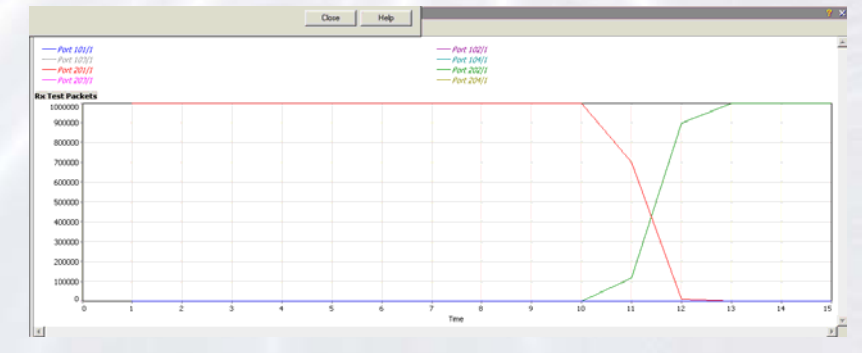
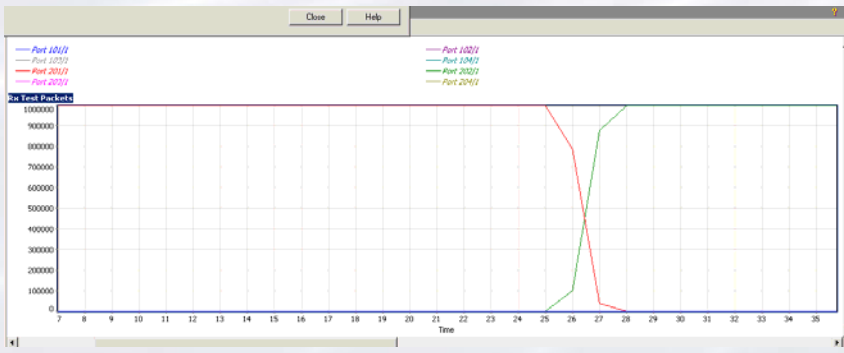
Link failure – single chassis – 4 sec

link failure – multiple chassis – 3 sec



Link up – single chassis – 3 sec

link up – multiple chassis – 3 sec



Test 5 – BGP and ISIS convergence

Results – times for intra- and inter-LCC convergence are very similar.

Explanation – routing information is always processed by TX.

Test 5 – indirect next-hop for iBGP

Indirect next-hop convergence is possible when the protocol (BGP) next-hop does not change, i.e. there is only an IGP failure, somewhere in the network

When the ISIS/OSPF next hop changes (because of topology change) the BGP routes do not need to be removed from routing table and installed again – convergence is much faster

Convergence time in our test (with 500,000 routes):

- eBGP (without indirect next-hop) – 80 sec
- iBGP (with indirect next-hop) – 10 sec

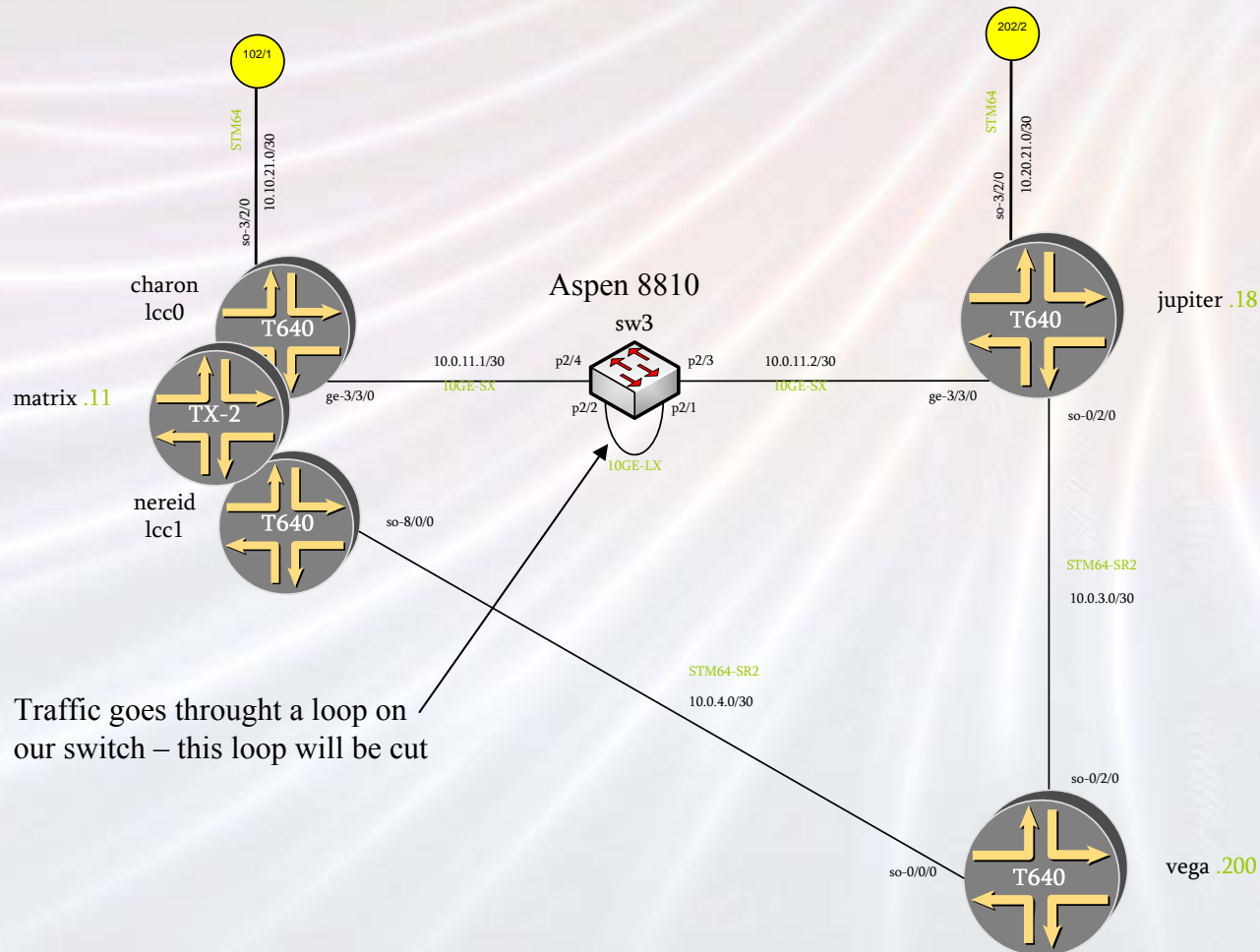
Test 6 – Bidirectional Forwarding Detection (BFD)

Faster convergence of routing protocol when interface state do not change to down, e.g. link failure between two Ethernet switches (or sometimes SDH switches with Ethernet client interfaces)

Shorter dead times than with routing protocols

Can be used with BGP (single hop eBGP only), ISIS, OSPF, static route, RIP, LDP, RSVP based LDP

Test 6 – BFD lab setup



Test 6 – BFD test

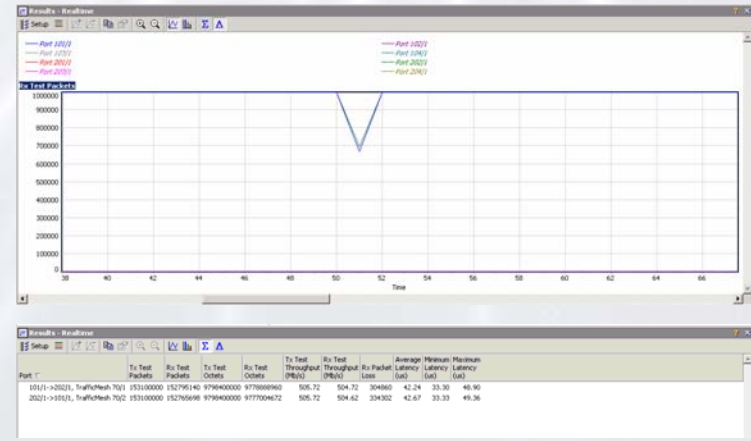
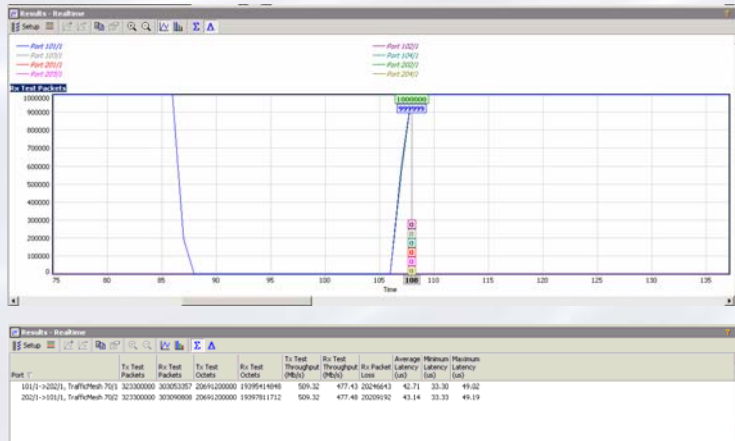
ISIS with default hello time – 9 seconds

Expected time for ISIS to detect the failure of the loop on switch – between 18 and 27 seconds

Test results:

Without BFD (22 sec recovery,
20,000,000 packets lost)

With BFD (0.3 sec recovery,
304,000 packets lost)



Test 6 – BFD test results

BFD makes routing protocol convergence much faster

One recommended value with current implementation is to use a detection-time of 250 ms for up to 100 sessions

Test 7 – QoS/CoS – forwarding classes

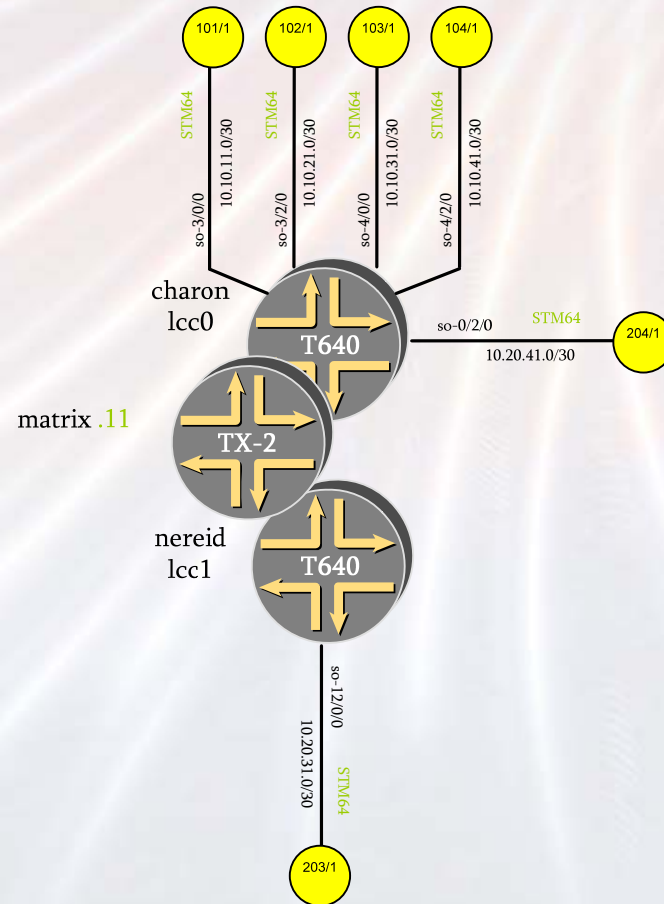
Three forwarding classes:

- Premium IP
- Best Effort
- Less than Best Effort

Traffic from 101, 102, 103 and 104 to 204 (intra-LCC) or 203 (inter-LCC)

Links to 203 and 204 congested

BGP and ISIS between TX and 203 and 204



Test 7 – QoS/CoS – forwarding classes – part 1

Premium IP
– 2Gbps

Best Effort
– 6Gbps

Less than
Best Effort
– 30Gbps

intra-LCC

inter-LCC

Port	Tx Test Packets	Rx Test Packets	Tx Test Octets	Rx Test Octets	Tx Test Throughput (Mb/s)	Rx Test Throughput (Mb/s)	Rx Packet Loss	Average Latency (us)	Minimum Latency (us)	Maximum Latency (us)
101/1->204/1, BE	323200000	323200000	487385600000	487385600000	6017.11	6017.11	0	40.65	26.02	47.38
101/1->204/1, EF	684020090	684020090	166900901960	166900901960	2060.50	2060.50	0	42.43	21.82	53.05
101/1->204/1, LBE*	78860800	5572008	118922086400	8402588064	1468.17	103.74	73288792	4032.85	89.51	7407.79
102/1->204/1, LBE*	513168120	24703072	773857524960	37252232576	9553.80	459.90	488465048	3688.35	27.45	68091.77
103/1->204/1, LBE**	513168123	24147017	773857529484	36413701636	9553.80	449.5	489021106	39597.88	28.67	68081.41
104/1->204/1, LBE**	513168117	24492384	773857520436	36934515072	9553.80	455.9	488675737	39550.88	29.88	68097.80

Port	Tx Test Packets	Rx Test Packets	Tx Test Octets	Rx Test Octets	Tx Test Throughput (Mb/s)	Rx Test Throughput (Mb/s)	Rx Packet Loss	Average Latency (us)	Minimum Latency (us)	Maximum Latency (us)
101/1->203/1, BE	60799999	60799999	91686398492	91686398492	5963.34	5963.34	0	10.77	25.60	46.87
101/1->203/1, EF	128676252	128676252	31397005488	31397005488	2042.08	2042.08	0	12.37	21.79	52.69
101/1->203/1, LBE*	14835200	1015501	2237148160	1531375508	1455.06	99.0	13819699	402.87	42.00	7430.75
102/1->203/1, LBE*	96536541	4677744	14557710328	7054037952	9468.43	438.00	91858797	3969.06	27.48	68088.86
103/1->203/1, LBE**	96536538	4649753	14557705304	7011827524	9468.43	456.0	91886785	3937.23	28.71	68099.75
104/1->203/1, LBE**	96536542	4545510	14557705336	6854629080	9468.43	445.83	91991032	3968.03	29.93	68100.36

No packet loss for PIP and BE, huge packet loss for LBE

Test 7 – QoS/CoS – forwarding classes – part 2

Premium IP
– 2Gbps

Best Effort
– 20Gbps

Less than
Best Effort
– 18Gbps

intra-LCC

inter-LCC

Port	Tx Test Packets	Rx Test Packets	Tx Test Octets	Rx Test Octets	Tx Test Throughput (Mb/s)	Rx Test Throughput (Mb/s)	Rx Packet Loss	Average Latency (us)	Minimum Latency (us)	Maximum Latency (us)
101/1->204/1, BE	10276968	44063725	154976704584	66448097300	7606.22	3261.56	58705973	76393.19	29.54	77442.01
101/1->204/1, EF	158102603	158102603	38577035132	38577035132	1893.35	1893.35	0	39.21	21.45	43.96
102/1->204/1, LBE	128450752	3985438	193703734016	6010040504	9506.93	294.17	124465314	3462.68	26.07	75241.03
103/1->204/1, BE"	25692424	7228133	38744175392	10900024564	1901.55	534.07	18464291	76335.00	30.41	77715.72
103/1->204/1, LBE"	102758327	3193892	154959557116	4816389136	7605.38	236.33	99564435	6679.08	27.31	75073.18
104/1->204/1, BE'	128450754	44358056	193703737032	66891948448	9506.93	383.04	84092698	6702.28	28.52	77498.19

Port	Tx Test Packets	Rx Test Packets	Tx Test Octets	Rx Test Octets	Tx Test Throughput (Mb/s)	Rx Test Throughput (Mb/s)	Rx Packet Loss	Average Latency (us)	Minimum Latency (us)	Maximum Latency (us)
101/1->203/1, BE	151296305	64901264	228154827940	9771106112	7605.16	3262.77	86395041	76384.42	30.37	77224.11
101/1->203/1, EF	232758123	232758123	56792982012	56792982012	1893.10	1893.10	0	39.36	22.13	44.18
102/1->203/1, LBE	189103861	5874522	285168622388	8658779176	9505.62	295.19	183229339	34702.87	26.89	75400.62
103/1->203/1, BE"	37824076	10615895	57038706608	16008769660	1901.29	533.03	27208181	7682.05	31.23	77340.96
103/1->203/1, LBE"	151279787	4704248	228129916796	7094005984	7604.33	236.41	146575539	6703.38	28.12	75246.53
104/1->203/1, BE'	189103861	65258735	285168622388	98410172380	9505.62	3280.34	123845126	6758.07	29.35	77294.11

No packet loss for PIP, huge packet loss for BE, almost all traffic lost for LBE

Test 7 – QoS/CoS – forwarding classes – results

- PIP traffic has no packet loss
- BE has priority over LBE
- No ISIS adjacency or BGP session drop detected on congested links

CoS features behave according to service definition.

Results are the same for intra-LCC and inter-LCC traffic.

Test 7 – QoS/CoS – head of line blocking

Head of line blocking occurs when traffic which should be transmitted via a congested interface blocks input queue on its ingress interface, which affects other on the same ingress interface – traffic which should not be congested

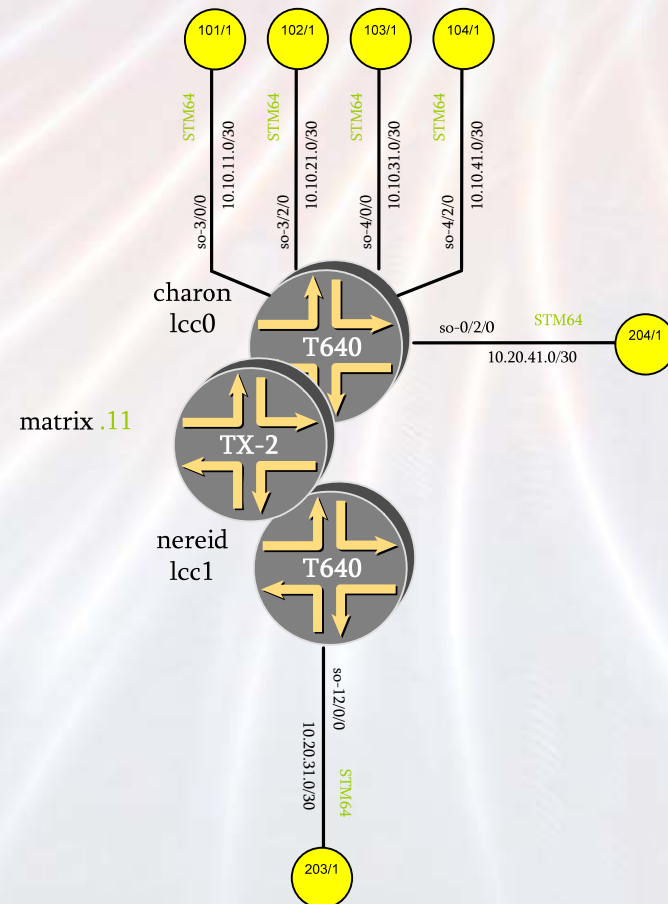
Test 7 – QoS/CoS – head of line blocking

Traffic:

- 203 to 101 – 100% of bandwidth
- 204 to 101 – 50% of bandwidth
- 204 to 102 – 50% of bandwidth

Link to 101 is congested

If TX has head of line blocking problem, the congestion on 101 will affect traffic from 204 to 102



Test 7 – QoS/CoS – head of line blocking - results

Traffic from 204 to 102 was not affected.

No head of line blocking in this test scenario.

No packet loss between 204 and 102

Port	Tx Test Packets	Rx Test Packets	Tx Test Octets	Rx Test Octets	Tx Test Throughput (Mb/s)	Rx Test Throughput (Mb/s)	Rx Packet Loss	Average Latency (us)	Minimum Latency (us)	Maximum Latency (us)
203/1->101/1, LBE	47633251	31853689	71830942508	48035363012	9420.45	6299.72	15779562	32129.11	26.68	48203.63
204/1->101/1, LBE	23816670	15817812	35915538360	23853260496	4710.23	3128.30	7998888	32129.02	25.55	48203.16
204/1->102/1, LBE	23816670	23816670	35915538360	35915538360	4710.23	4710.23	0	24.42	23.46	27.60

Many thanks to Jean-Marc Uze
for organising this testing.