Ultra-Fast Broadband Realities

Presented by Donald Love

donald@resolveware.com

www.resolveware.com

Asia Pacific Regional INTERNET Conference on Operational Technologies

Agenda

- Overview of Ultra-Fast Broadband in New Zealand
- Some issues observed in NZ production networks
  - Lack of transparency for Ethernet frames
  - Upstream loss
  - Excess padding on user end ONT
  - Poor end user throughput especially for high delay sources
- Explanation and remedial actions for poor user rates
  - Impact of Local Fibre Company burst settings
  - Challenges in mitigating the burst issues
- Open source test tools

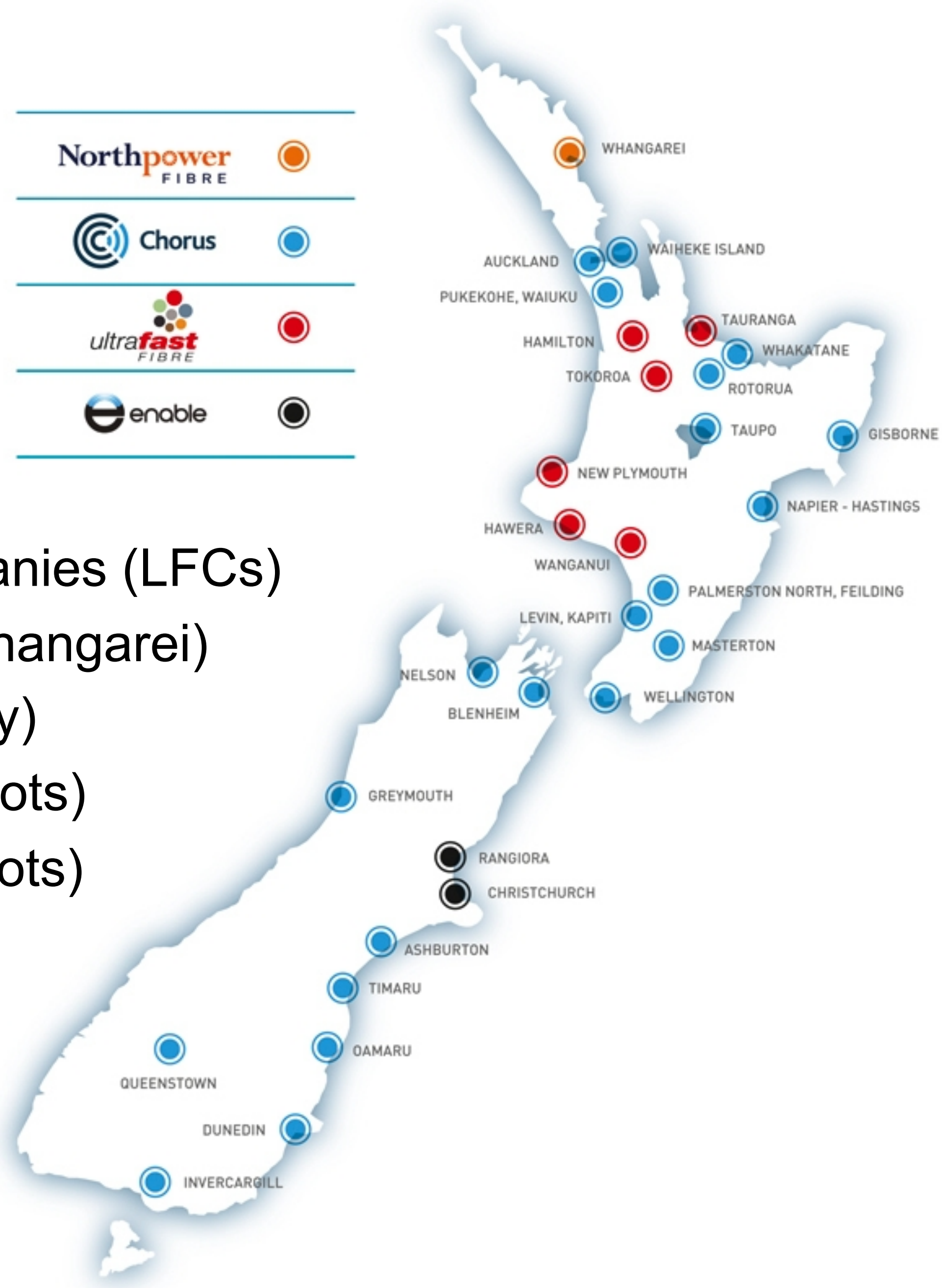
My Reality


Wellington
30 mins drive
Nearest
UFB 12km
My office
My ADSL modem
via WiFi

Introduction

- Mandated by the New Zealand Government
  - Layer 2 fibre based service to 75% of population in 33 towns and cities by 2019
  - "Open Access"
  - Availability of a minimum speed of 100 Mbps Downstream (from the Internet to the user) and a minimum of 50 Mbps Upstream (from user to the Internet)
  - Separate to the Rural Broadband Initiative
- Managed by Crown Fibre Holdings (http://www.crownfibre.govt.nz)
  - In Feb 2011 standards approved, common specifications
  - Local Fibre Companies (LFCs) build network
  - Retail Service Providers (RSPs) connect to 33+ handovers

Locations

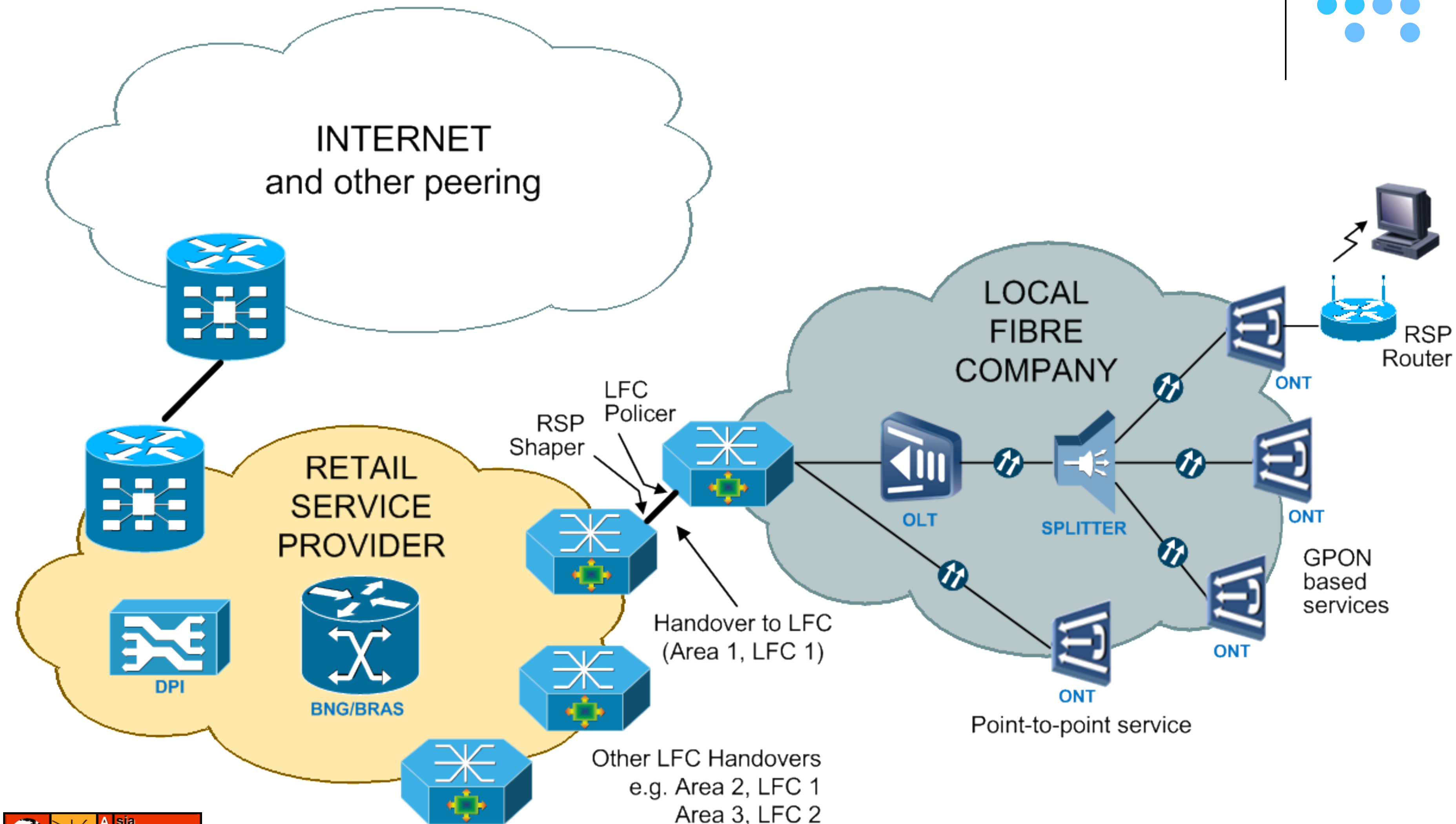


Four Local Fibre Companies (LFCs)

- Northpower (Whangarei)
- Chorus (majority)
- UltraFast (red dots)
- Enable (black dots)

Ultrafast and Enable share same vendor and configuration

Network Topology



Service Variants

Ultra-Fast Broadband (UFB) in New Zealand

Service	Low Priority EIR (Mbps)		Low Priority CIR (Mbps)		High Priority CIR (Mbps)		Target Market	Delivery
	Down	Up	Down	Up	Down	Up		
Bitstream 2	30 to 100	10 to 50	-	-	2.5 to 10	2.5 to 5	Residential and Small Business	GPON
Bitstream 3	-	-	-	-	2.5 to 100	2.5 to 100	Business and Premium Res.	GPON
Bitstream 3a	95 to 7.5	95 to 7.5	2.5	2.5	2.5 to 90	2.5 to 90	Business and Premium Res.	GPON
Bitstream 4	-	-	-	-	100 to 10000	100 to 10000	Premium business	P2P Fibre
Multicast	Delivered in combination with other Bitstream 2, 3 or 3a						Residential and Small Business	GPON
ATA Voice	Delivered in combination with other Bitstream 2, 3 or 3a							

Note: Bitstream 4 has EIR and CIR permutations beyond that shown. There are also Education profiles that can be delivered over GPON or Point-to-Point (P2P) fibre.

Some Technical Expectations

Residential Users

- Reliable
- ***Faster than ADSL !***
- If using voice, no worse than a POTS service

Business Users

- Reliable
- Transparent for their applications
- Achieve contracted speeds
- Priority traffic protected with low loss and delay

Transparency - PCP

Intended Behaviour:

ONT Input	Handover Out
PCP=4	PCP=4
PCP=0,1,2,3,5,6,7	PCP=0

This behaviour was in 2H 2012 and is now corrected. It was well known for users of the LFC and not an issue if the RSP controls the upstream PCP.

Observed Behaviour:

ONT Input	Handover Out
PCP=1,2,3,5,6,7	= no traffic =
PCP=4	PCP=4
PCP=0	PCP=0

For business use, where routers may set PCP outside the passed values, it could have been a problem.

Asia Pacific Regional INTERNET Conference on Operational Technologies

Transparency - Multicast

Intended Behaviour:

- Only block IEEE reserved bridge range, pass other multicast

Observed Behaviour:

Destination Address	ONT Out
224.0.0/24	Passed
Ethernet Multicast (no IP)	Blocked
224.0.1/24 – 239.255.255/24	Blocked

Observed in Nov 2012 and is now corrected.

When the filter was active it seemed to block any Ethernet multicast unless the IPv4 address matched the limited range.

These are smart networks and transparency can be more complex than expected.

Upstream Loss – Bitstream 3

For traffic within CIR low level loss was seen

- Loss (at 0.005%) well less than the SLA of 0.1%
- Starts when rates > 70 Mbps for 100 Mbps service

Cisco TelePresence is highly sensitive to packet loss, and as such has an end-to-end packet loss target of 0.05%.

Specifically, if packet loss exceeds 0.10% (or 1 in 1000 packets, which we call Loss Threshold 1) for several **seconds**, then:

• A warning message appears at the bottom of the on the 65” plasma display indicating that the network is experiencing congestion and that call quality may be affected. http://www.cisco.com/en/US/docs/solutions/Enterprise/Video/tpqos.pdf

- Will this loss change over time ?? Does it matter ?

APRICOT
Asia Pacific Regional INTERNET Conference on Operational Technologies

Excess Pad Issue

Minor issue found on Chorus service

- Found during EtherBERT test for frame integrity
- Handover 64 byte frame becomes 100 byte from ONT
- Any downstream frame < 100 bytes padded by ONT
- 100 Mbps @ 68 bytes becomes 140 Mbps

Is it a problem?

- Would any application or network equipment care?
- If tightly shaping to 100 Mbps towards LFC, egress from the ONT may not fit 100 Mbps interface
- Mitigation - leave small margin or use 1G interface

New UFB = Poor Performance?

Residential users with poor international speed

- Some claims of worse than ADSL (e.g. 1.5 Mbps)
- Even national speeds were lower than expected
- RSP change of network settings improved service

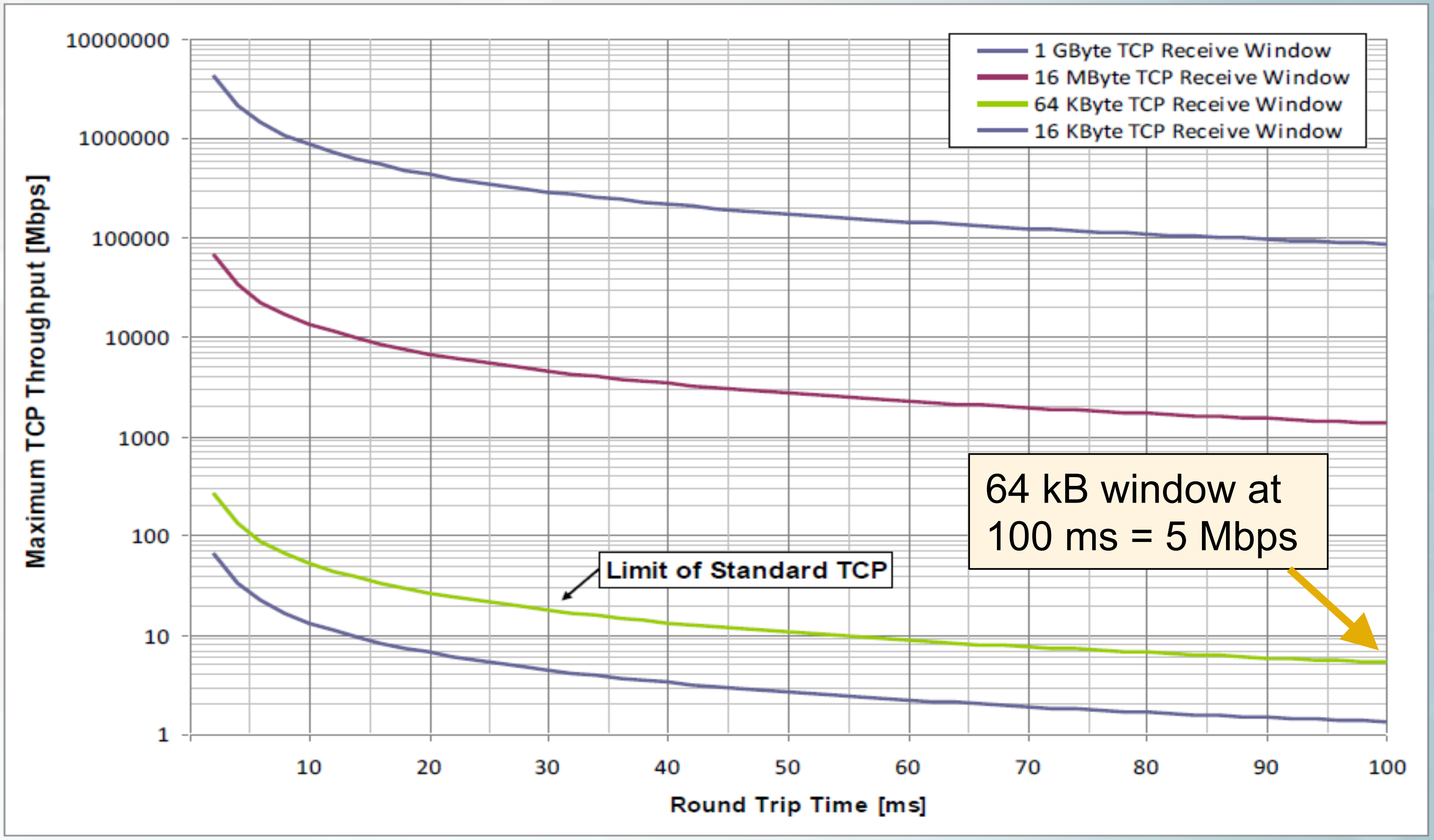
Wow, my oseas speeds have skyrocketed.

Chicago was around 1.8 dn but it has jumped to 15 dn /4.7 up now.

LA is now 15 dn/4.8 up.

- Why did this happen?

TCP Primer – Window Size



Window scaling (RFC 1323) allows TCP windows to be expanded beyond 64K

"Reproduced with permission of the Metro Ethernet Forum."

TCP Primer – CWND Phases

TCP expects loss to signal congestion

```
  /\  |
  /\  |High ssthresh  TCP CWND                        TCP
  /\  |Loss Event *   halving    3-Loss Recovery      Equilibrium
   T  |           * \   upon loss
   h  |           *  \     /  \          Time-Out      Adjusted
   r  |           *   \   /    \        +--------+    * ssthresh
 T o  |           *    \/       \      / Multiple|   *
 C u  |           * 2-Congestion\   /  Loss      |   *
 P g  |          *    Avoidance  \/    Event     |  *
   h  |         *              Half              | *
   p  |       *                TCP CWND          | * 1-Slow Start
   u  | * 1-Slow Start                        Min TCP CWND after T-O
   t  +-----------------------------------------------------------
        Time > > > > > > > > > > > > > > > > > > > > > > > > > > >
```

Don't want loss here

Bad News

```
Note:  ssthresh = Slow Start threshold.
```

From RFC 6349

High RTT dramatically slows process !

Factors Limiting Speed

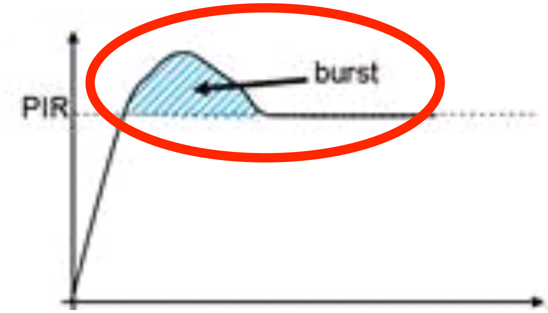
TCP over Long Fat Networks requires:

- Large window sizes to match Bandwidth Delay Product meaning window scaling essential
- Very low loss, especially during slow start
- Effective error recovery (cope with multiple events per window)
- Bursts of traffic from server should be passed without loss to end user especially if **<u>average</u>** rates less than link capacity

Burst Size Acceptance by the Local Fibre Company:

- Chorus chose small size to manage stringent jitter spec
- The Retail Service Provider (RSP) needs to manage egress bursts towards the LFC (e.g. Chorus)

RSP Shaper Response



The burst released by the RSP shaper, beyond the service rate, **must not** exceed the burst acceptance of the LFC policer if random loss is to be avoided.

- Some shapers may struggle to match the attack time required for Chorus settings for low priority (32 kB)
- For high priority (currently 8 kB) much more challenging
- CBS=8 kB equates to 0.64 ms for a 100 Mbps service !
- For low priority we are matching to a capacity which is not available 100% of time – what happens when the shaper incorrectly believes there is capacity? Does the LFC buffer?

Buffering – RSP to LFC

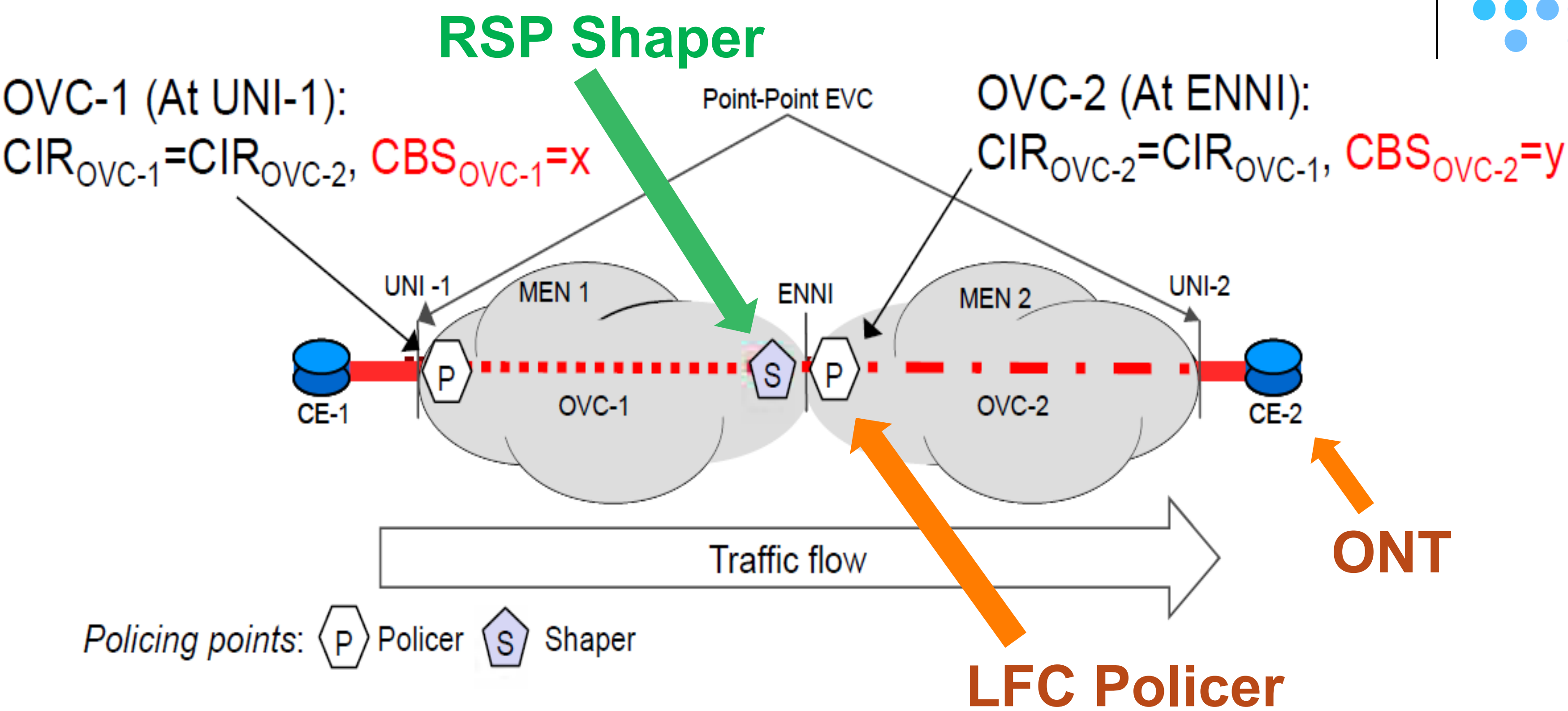
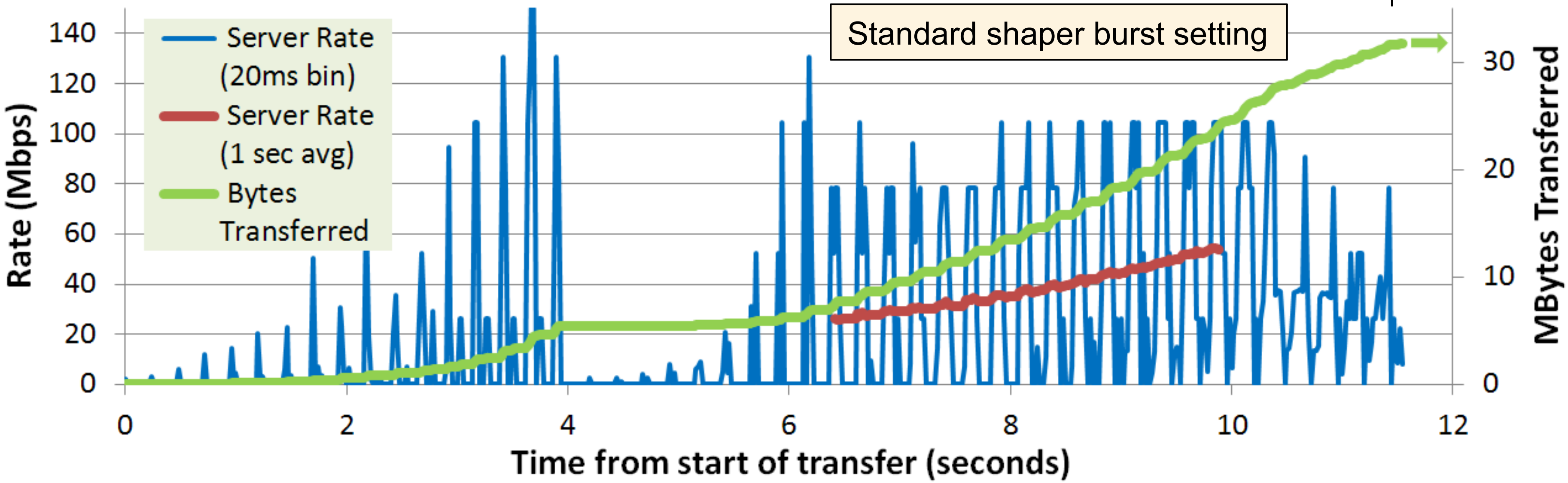
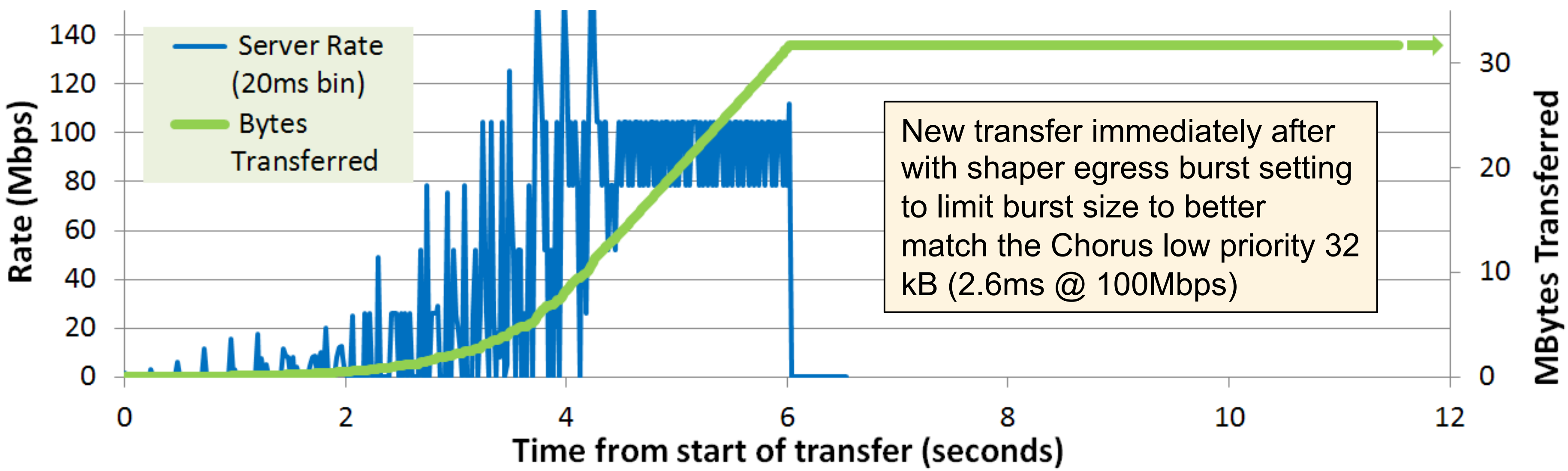


Figure 7 – Burst Alignment Example with Policing Points for Traffic Traversing the ENNI

From MEF23.1 - Section 8.7.1 Burst Size and Burst Alignment
"Reproduced with permission of the Metro Ethernet Forum."

RSP Network Change Example





FTP download of 32 MB file from Montreal, Canada to Auckland

RSP Network Change Example

```
tcptrace analysis at sending end
# before burst size tuning
e->f:                               f->e:
    total packets:          9692        unique bytes sent:   31625365
    sack pkts sent:         2439        actual data bytes:   31651429
    dsack pkts sent:           6        rexmt data pkts:           18
    max sack blks/ack:         3        rexmt data bytes:       26064
    max win adv:         3145728 bytes  throughput:          2630533 Bps
    avg win adv:         2541406 bytes  RTT max:                278.7 ms

# after shaper burst size tuning to better match LFC policer
k->l:                               l->k:
    total packets:          8282        unique bytes sent:   31625365
    sack pkts sent:            1        actual data bytes:   31625365
    dsack pkts sent:           0        rexmt data pkts:            0
    max sack blks/ack:         1        rexmt data bytes:           0
    max win adv:         3145728 byte   throughput:          4830419 Bps
    avg win adv:         2559270 bytes  RTT max:                281.3 ms
```

FTP download of 32 MB file from Montreal, Canada to Auckland

How Much Improvement?

The improvements can vary substantially

- High delay sources much more affected
- Users who had 1.5 Mbps now 20 Mbps +
- Shared pools with ADSL may match ADSL speed

Client and server TCP settings a big factor

- Larger windows essential for Long Fat Networks
- Older error recovery techniques less loss tolerant
- “Without SACK, TCP takes a very long time to recover after multiple and consecutive losses”
- SACK may have been disabled for security reasons
  https://tools.ietf.org/html/rfc6349

Useful Test Tools

Linux based tools with open source software

- Wireshark, tshark, tcptrace – capture and analyse
- Iperf, nuttcp – rate testing
  - nuttcp UDP burst mode sends defined number of at line rate – useful for assessing buffers
    - e.g. `nuttcp -u -Ri28M/50 -T2 192.168.105.1`
  - Be cautious with Iperf or nuttcp at high upstream rates as GPON can add some jitter causing brief receiver overload
- Scapy – packet crafting and analysis
  - Python based, generate ranges easily - e.g. to walk PCP values: `Dot1Q(vlan=10,prio=(0,7))`

Conclusion – NZ Situation

Various minor issues seen with each LFC

- Issues causing concerns have been fixed
- Others may or may not still exist
  - Would they be a concern for you or your customers?
  - Important that RSPs audit services and work with LFC

Burst size alignment has restored TCP (for now)

- A complex topic, especially for high delay, where user bandwidth expectations may need on-going education
- On-going process with proposed new Chorus service definitions with larger burst sizes (and delay SLAs)

LFCs professional, helpful and responsive

Final Thoughts

GPON based issues

- Service transparency, if Layer 2 services, a concern
- Upstream loss for premium services a consideration

Burst size alignment at handovers impacting TCP

- Stringent delay SLAs, perhaps defined by the regulator, can invoke design choices which may prove challenging for delivering TCP friendly services
- User expectations of high speed over Long Fat Networks may be unrealistic but what is reasonable?

Thank You

APRICOT
Asia Pacific Regional INTERNET Conference on Operational Technologies

Backup Slides

International Performance Sample

Verizon Enterprise Solutions Latency Statistics for Country Specific Metrics (ms)												
	2013	2012										
	January	December	November	October	September	August	July	June	May	April	March	February
Hong Kong to US (230.000)	152.561	156.701	148.912	148.205	149.422	148.114	149.470	150.293	148.715	147.630	149.262	152.449
Singapore to US (260.000)	202.966	199.444	196.596	200.893	199.364	201.820	201.808	201.797	192.459	202.148	202.620	201.899
Australia to US (210.000)	155.542	155.561	156.894	161.278	157.141	155.505	155.490	155.546	155.505	155.559	155.547	155.511

Verizon Enterprise Solutions Packet Delivery Statistics for Country Specific Metrics (%)												
	2013	2012										
	January	December	November	October	September	August	July	June	May	April	March	February
Hong Kong to US (99.000)	99.975	100.000	99.998	99.985	99.998	99.994	100.000	99.992	99.972	99.988	99.985	100.000
Singapore to US (99.000)	100.000	100.000	99.998	99.996	99.999	99.999	99.998	99.984	100.000	100.000	100.000	100.000
Australia to US (99.000)	100.000	100.000	100.000	99.962	99.977	99.998	99.989	99.933	100.000	100.000	99.999	99.999

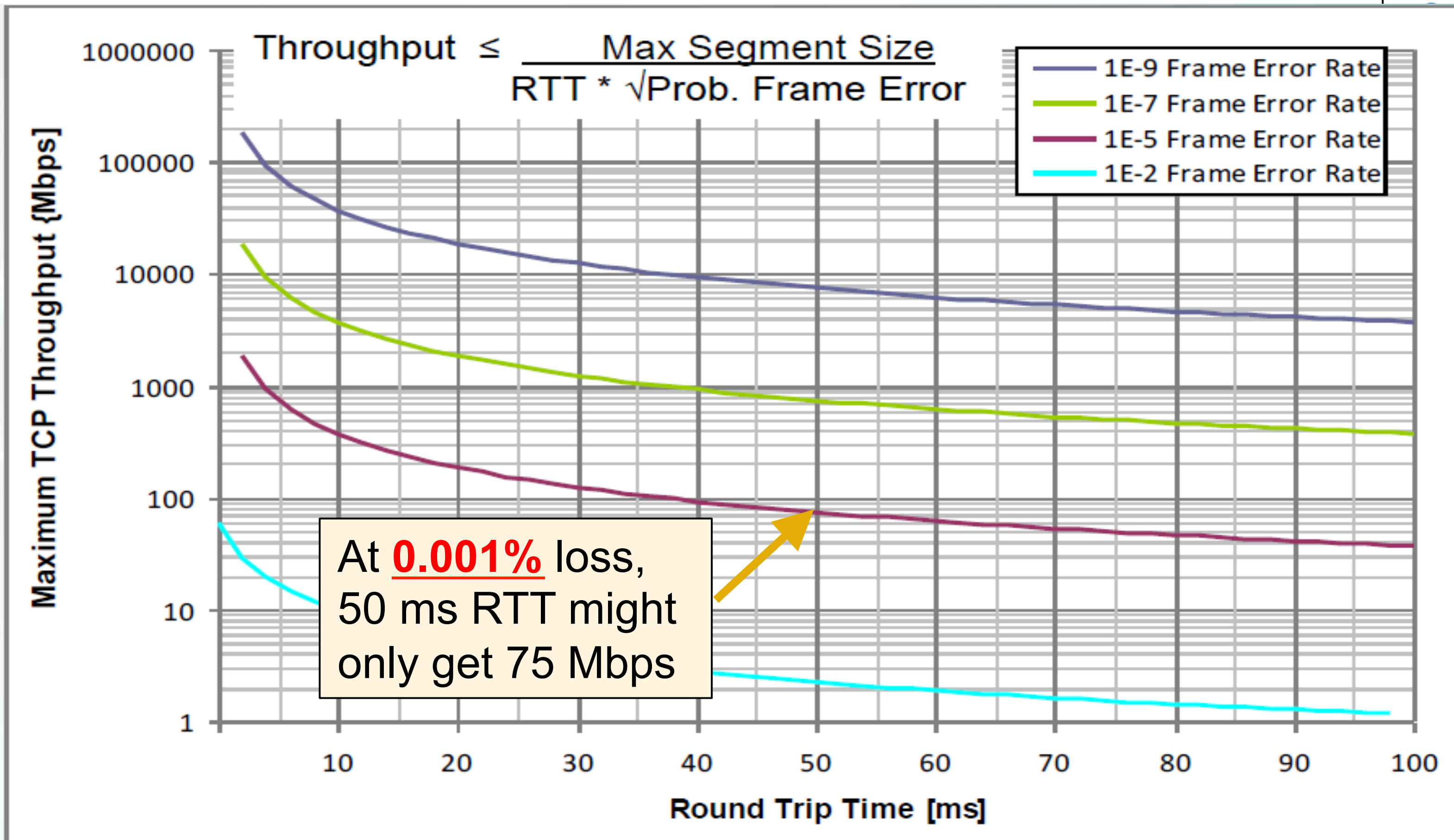
http://www.verizonenterprise.com/about/network/latency/

User Experience

http://www.geekzone.co.nz/forums.asp?forumid=82&topicid=113670

APRICOT Asia Pacific Regional INTERNET Conference on Operational Technologies

TCP Primer – Frame Errors



Round trip time (RTT) combines with frame error rate to limit achievable TCP throughput. Dropped Frames are same as errored frames. Traffic Policing is like a very high error rate.

"Reproduced with permission of the Metro Ethernet Forum."

Loss Specifications

UFB in NZ for high priority inside CIR

- 0.1% over each 5 minute interval (24 hrs / day)
- Earlier TCF draft 0.01% loss

MEF

- Tier 1 CPOs – Metro – CoS Label H = < 0.01%
- Measurement time interval 1 month !!

ITU Y.1541

- “Proposed” Class 6 – IPLR $< 1\times10^{-5}$ = < 0.001%
- Measurement time interval 1 minute

Delay Specifications

UFB in NZ for high priority inside CIR

- Frame delay < 5 ms, FDV < 1 ms
- 99% or 99.9% (varies with LFC) every 5 minutes

MEF Tier 1 CPOs – Metro – CoS Label H

- FD < 10 ms, FDV < 5ms, IFDV < 3 ms
- Measurement time interval 99% and 1 month !!

ITU Y.1541

- “Proposed” Class 6 – IPDV < 50 ms
- Measurement time interval 1 minute

CFH NIPA Agreements

6. Layer 2 Traffic

6.1 Each End User's traffic must be delivered to the POI within the following Service Levels, measured over each five minute interval (24 hours per day):

	Frame Delay must be:	Frame Delay Variation must be:	Frame Loss must be:
CIR	≤ 5 mS	≤ 1 mS	≤ 0.1%
EIR	n/a	n/a	≤ 2%

6. Layer 2 Traffic

6.1 Each End User's traffic must be delivered to the POI within the following Service Levels, measured over each five minute interval (24 hours per day):

	Frame Delay must be:	Frame Delay Variation must be:	Frame Loss must be:
CIR	≤ 5 mS	≤ 1 mS	≤ 0.1%
EIR	n/a	n/a	≤ 2%

At least 99% of the frames within the five minute measurement interval must be within the above Service Levels, otherwise the service is to be considered unavailable for that five minute interval.

SACK

RFC 6349 – TCP testing

- In networks with unknown load and error patterns, TCP SACK will improve throughput performance. On the other hand, security appliance vendors might have implemented TCP randomization without considering TCP SACK, and under such circumstances, SACK might need to be disabled in the client/server IP hosts until the vendor corrects the issue. Also, poorly implemented SACK algorithms might cause extreme CPU loads and might need to be disabled.

TCP Error Correction

Snippets

- At most 1 lost segment can be retransmitted in Reno and NewReno per round trip time
- Selective acknowledgments: acknowledges non-continuous blocks of data
- Reno performs well only if no loss or one packet drop within a window
- NewReno can deal with multiple lost segments without entering slow start