Latency Challenges for 25/50/100G EPON

Rene Bonk, Thomas Pfeiffer - Nokia Bell Labs

Bill Powell - Nokia Fixed Networks CTO Group

Charlotte, NC USA
September 2017

Motivation

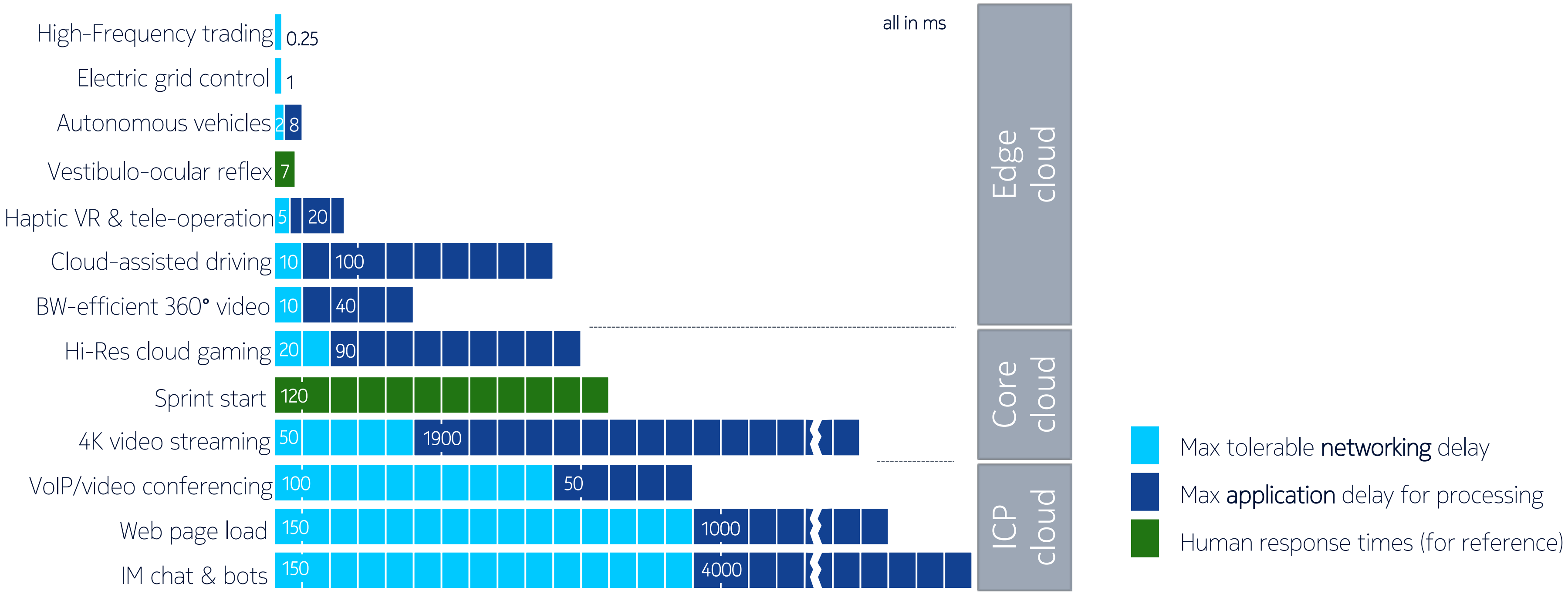
Presentation from ZTE (Jun Shan Wey, Huntington Beach, Jan/2017 wey_3ca_1_0117.pdf) suggests the examination how low latency requirements for future services would impact NG-EPON standards

In this presentation we focus on the impact of latency critical service and network realizations on the 802.3ca standard discussion

As an example where low-latency demands and next-generation PON networks coincide, we have chosen fronthaul (FH)/next-generation fronthaul interfaces (NGFI) in 4G/5G mobile networks that might be realized by 25G/50G/100G-EPON

NOKIA

E2E-service/network latency in 5G applications

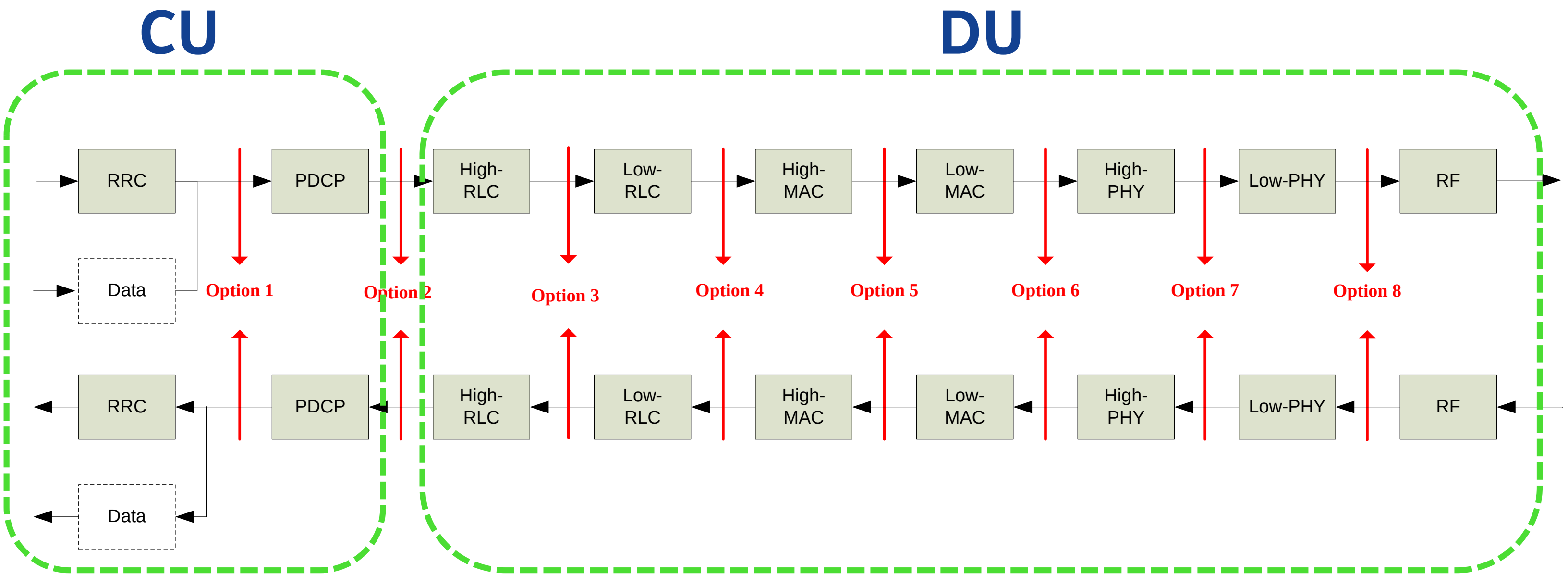


Radical shift in network architecture required to deliver required latency

NOKIA

4G & 5G wireless base stations processing chain & optional split points by 3GPP

Transport with low latency demands



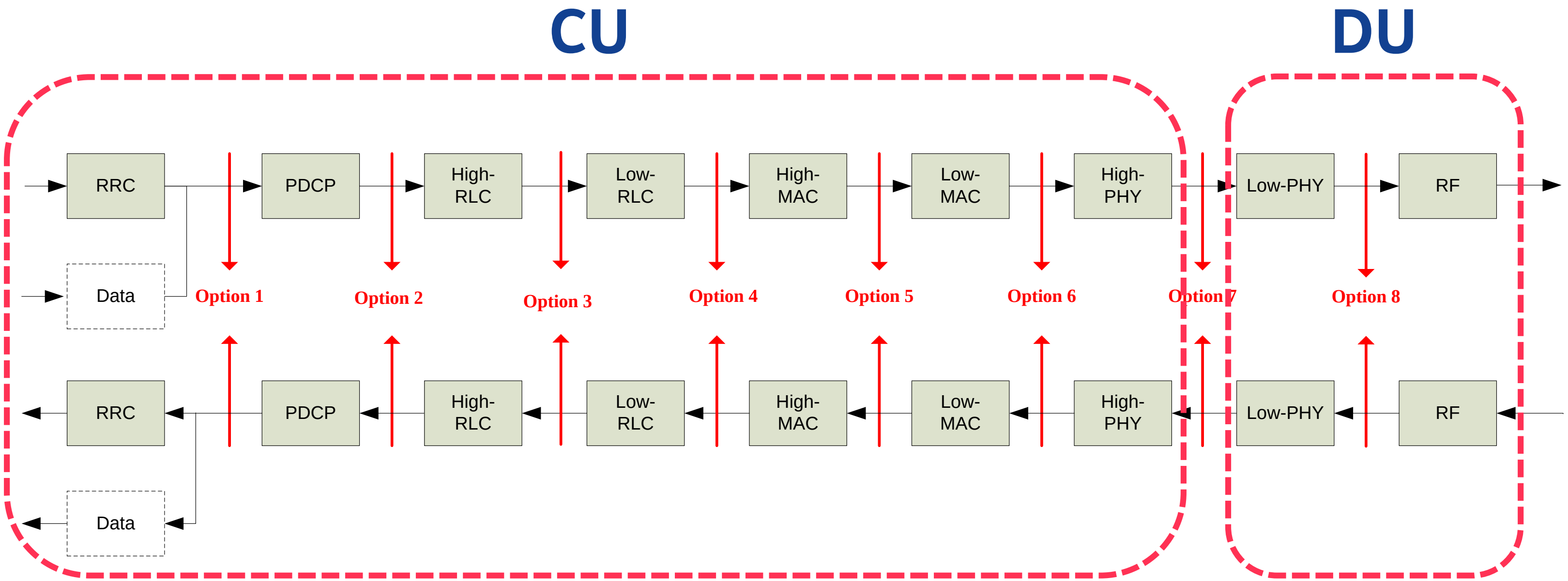
CU: Central Unit (e.g. collocated with OLT)

DU: Distributed Unit (e.g. collocated with ONU)

[1]

4G & 5G wireless base stations processing chain & optional split points by 3GPP

Transport with high latency demands



CU: Central Unit (e.g. collocated with OLT)

DU: Distributed Unit (e.g. collocated with ONU)

[1]

NOKIA

Updated

Allowed one way latency as per 3GPP next generation fronthaul architectures

Latency critical are the split options:

6, 7a, 7b, 7c and 8

→ 802.3ca should support these 4G/5G network requirements, i.e., allow the use of future PON for new services/networks

→ maximum one way latency ~250µs

→ Why 250µs one way latency?

Here: Network latency limit

HARQ (Hybrid Automatic Repeat Request) loop within the split (see next slide)

Otherwise: new services might demand low latency solutions (see over next slide)

Protocol Split option	Required bandwidth	Max. allowed one way latency [ms]	Comment
Option 1	[DL: 4Gb/s] [UL: 3Gb/s]	[10ms]	
Option 2	***[DL: 4016Mb/s] [UL:3024 Mb/s]***	***[1.5~10ms]***	***[16Mbps for DL and 24Mbps for UL is assumed as signalling]***
Option 3	[lower than option 2 for UL/DL]	[1.5~10ms]	
Option 4	[DL:4000Mb/s] [UL:3000Mb/s]	[approximate 100us]	
Option 5	[DL: 4000Mb/s] [UL: 3000 Mb/s]	[hundreds of microseconds]	
Option 6	***[DL: 4133Mb/s] [UL:5640 Mb/s]***	***[250us]***	***[133Mbps for DL is assumed as scheduling/ control signalling. 2640Mbps for UL is assumed as UL-PHY response to schedule]***
Option 7a	***[DL:10.1~22.2Gb/s] [UL:16.6~21.6Gb/s]***	***[250us]***	***[713.9Mbps for DL and 120Mbps for UL is assumed as MAC information]***
Option 7b	***[DL:37.8~86.1Gb/s] [UL:53.8~86.1 Gb/s]***	***[250us]***	***[121Mbps for DL and 80Mbps for UL is assumed as MAC information]***
Option 7c	***[DL:10.1~22.2Gb/s] [UL:53.8~86.1Gb/s]***	***[250us]***	
Option 8	[DL:157.3Gb/s] [UL: 157.3Gb/s]	[250us]	

Service latency limited

Network latency limited

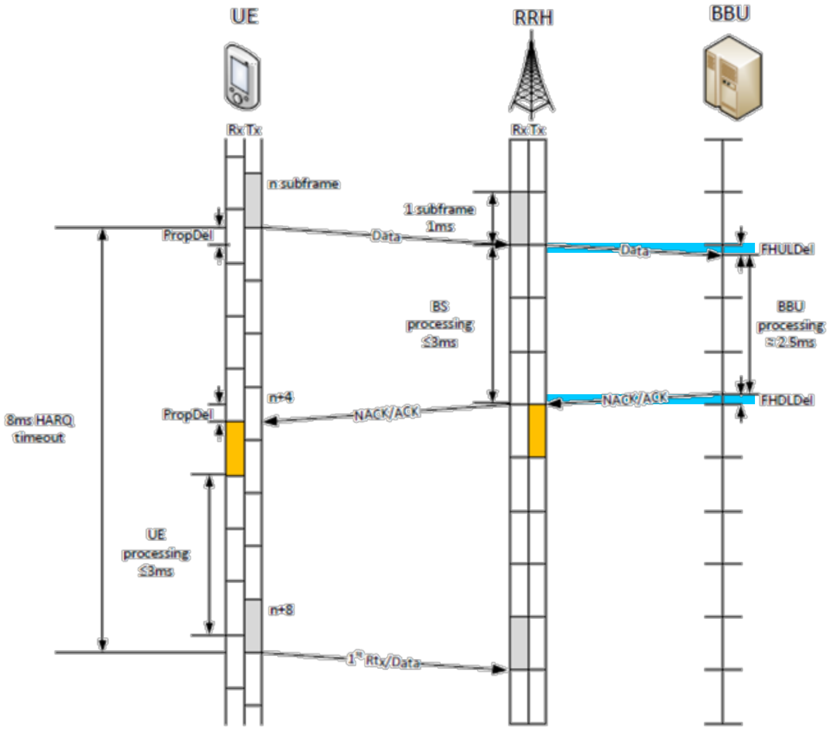
Outside HARQ loop

Inside HARQ loop

[1]

NOKIA

Signaling/control latency



Transmission Time Interval (TTI) refers to the duration of a transmission on the radio link

HARQ (Hybrid Automatic Repeat Request)

- UE – sends packet (TTI), TTI = 1ms in LTE
- BBU – responds NACK/ACK (acknowledge) after process of 2.5ms (within 3 TTI)
- UE – processes in 3 TTI

blue = fronthaul (BBU-RRH) transport delay
< 0.5ms

Fronthaul : Highly dependent on BBU process implementation (FPGA or SoC)
– vendor specific

In 5G:

Shorter and flexible TTI are under discussion

UL-HARQ architecture might be modified

NOKIA

Large variety of applications in 5G networks with substantially different traffic characteristics

Radio technology	Peak rate	Average rate	e2e delay (service level)
Enhanced Mobile Broadband (eMBB)	5 - 20 Gb/s	100 Mb/s per user in urban/suburban areas 1 – 4 Gb/s (hot spot areas)	10 msec
Ultra-Reliable Low Latency Communication (URLLC) / Critical Machine Type Communication (incl. D2D)	much lower than in eMBB: N x Mb/s	much lower than in eMBB: n x Mb/s	1 – 2.5 msec
Massive Machine Type Communication (mMTC)	much lower than in eMBB: N x Mb/s	much lower than in eMBB: n x kb/s - n x Mb/s	1 – 50 msec

(based on ITU-R M.2083 and 3GPP)

Example: CPRI fronthaul over TDM-PON for small cell applications in 4G

Take CPRI as an example to study latency critical PON transport

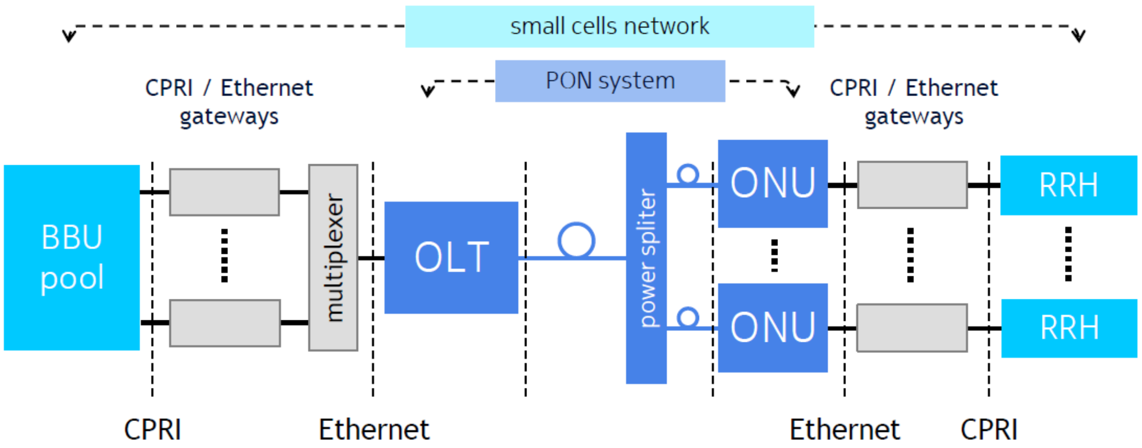
Fronthaul (split option 8 on slide 4/5): LTE 20MHz, 2x2 MIMO requires CPRI Option 3

IEEE PON: 25G/25G → 10 cells can be supported without CPRI compression

Assume fixed bandwidth assignment on the TDM-PON

Native CPRI-over-Ethernet: backward compatibility with CPRI interfaces on radio equipment

- Chopping and encapsulating the CPRI stream

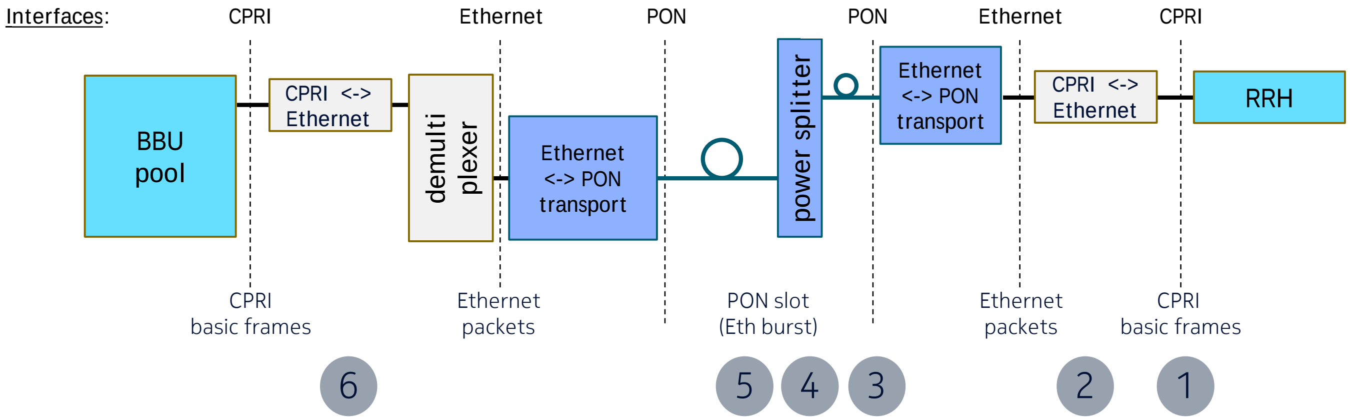


[2]

NOKIA

Upstream transmission : CPRI ← Ethernet ← TDM-PON ← Ethernet ← CPRI

Set-up



Total latency with S/F buffers, neglecting overheads in PON & analogue/digital processing times

1) Buffer time for continuous CPRI data (multiple integer of: 16 CPRI basic frames = 1 "CPRI block")
2) CPRI to Ethernet encapsulation
3) MAC scheduling delay (wait time for PON slot)
4) Buffer time for PON burst (slotted transmission)
5) Fiber propagation delay
6) Ethernet to CPRI decapsulation

Latency parameter evaluations

1) Buffer time for continuous CPRI data = **multiple integer of 4.2µs**
   - Value of multiple integer depends on # of CPRI blocks per burst
2) CPRI to Ethernet encapsulation per block= **0.33µs**
3) MAC scheduling delay (wait time for PON slot)
4) Buffer time for PON slot length = **multiple integer of 0.42µs** (matching PON cycle time to CPRI rates)
5) Fiber propagation delay for 20km (one way, maximum distance in 802.3ca) **≈ 100µs**
6) Ethernet to CPRI decapsulation per block = **0.33µs**

≈ 105µs (for 1 CPRI block) + scheduling delay

Some additional "FH-over-PON"-related latencies:

Processing delays (seems system vendor specific, not further investigated)

FEC (encoding/decoding) delays

MAC scheduling delay

Basic analysis of scheduling delay to allow further investigation of latency budget

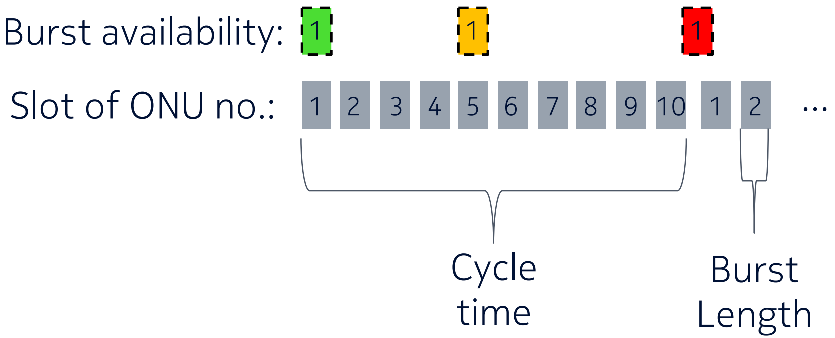
Fixed bandwidth allocation per ONU assumed (no Gate/Report message exchange, autonomous grants)

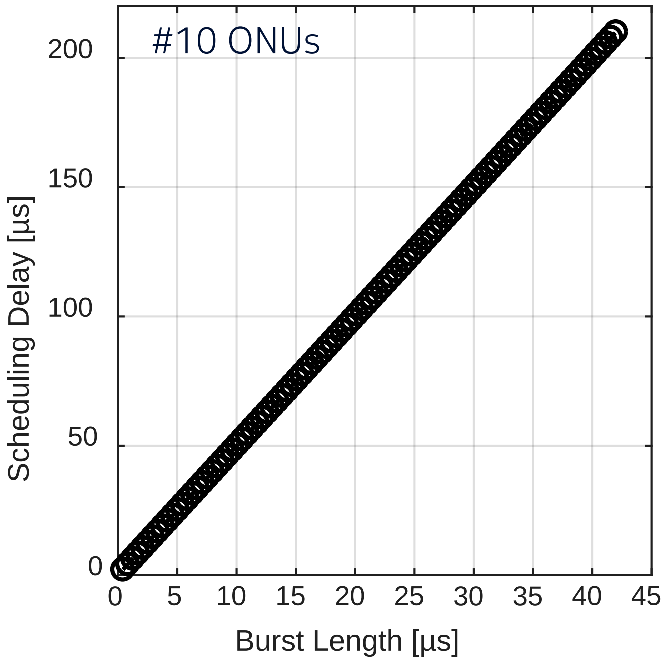
Assumption: PON system supports up to 10 ONUs/RRH

→ Average MAC scheduling delay = cycle time /2

→ Scheduling delay evaluated for different burst lengths

(no guard time between bursts, no US burst header)

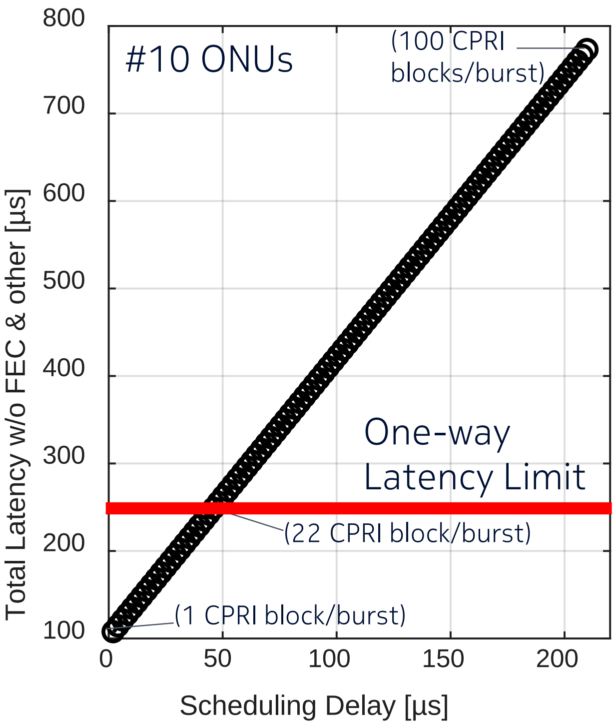


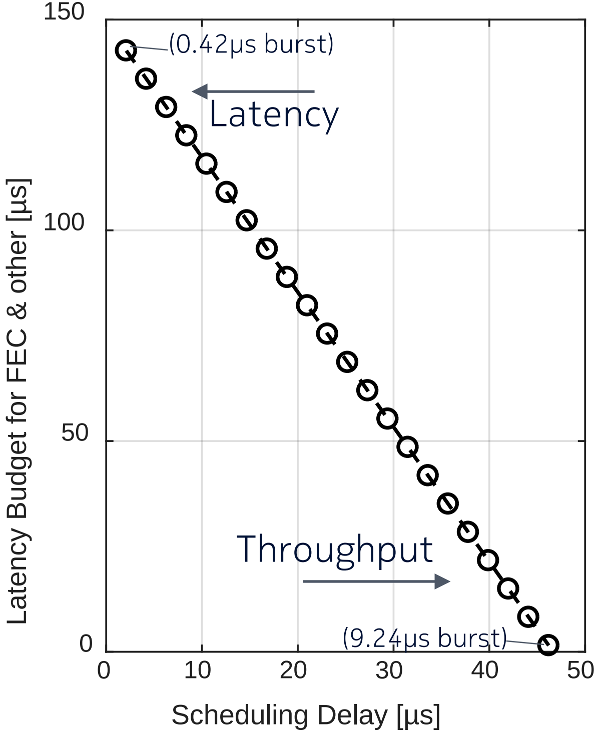


Tolerable latency for FEC and other means

Trade-off between latency and throughput (here represented by scheduling delay, burst-length)

Time budget allocable to FEC very limited (need for processing delay, header, guard time, etc…)





Summary

- 802.3ca should allow for lowest possible latency implementation to support a large variety of today's and future services and network solutions using 25/50/100G EPON as a transport medium
- MAC scheduling delay and processing delays are a system vendor implementation specific challenge
- Fiber reach could be limited to gain latency budget, but this approach seems rather operator implementation specific. 802.3ca should study low-latency implementation scenarios with up to 20km fiber reach
- FEC is required in any case to support the demanding power-budget for PR30
- Trade-off between latency and throughput as well as overheads for very short-bursts (short cycle-times)
- We suggest to adapt the use of low-latency FEC implementations in 802.3ca, i.e.:"Short" code-word length (about 2k to 10k bits)
  → PR30 power budget seems achievable
- RS codes seems generally better suited for low-latency implementations compared to LDPC codes

References

[1] 3GPP Release 14, TR 38.801 V14.0.0 (2017-03)

[2] T. Pfeiffer, "Fixed-mobile convergence in optical metro-access networks", Workshop "Role of Optics in Fronthaul and Backhaul for 5G Networks and Beyond", CLEO PR, OECC and PGC 2017, Singapore, August 2017

NOKIA