5G RAN faster

How the *Need for Speed* Impacts the xHaul

Yaakov Stein
CTO

Your Network's Edge®

About RAD

RAD
Your Network's Edge®

- Telecommunications access solutions provider
- Founded in 1981, privately owned
- Global presence, part of the $1.46 billion RAD Group

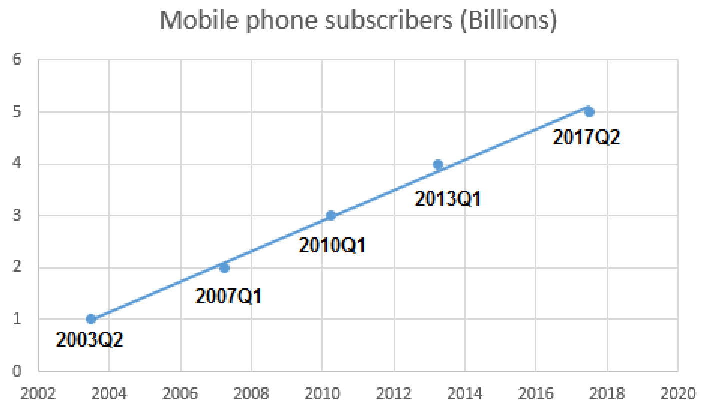
Importance of mobile communications

RAD
Your Network's Edge®

Mobile communications is consistently ranked
as one of mankind's breakthrough technologies

Annual worldwide mobile service provider revenue exceeds 1 trillion USD
and mobile services generate about 5% of global GDP

5 billion people (2/3 of the world) own at least 1 mobile phone (> 8B devices)
with over ½ of these smartphones
and over ½ of all Internet usage
from smartphones



Generations of cellular technologies

RAD
Your Network's Edge®

	1G	2G	3G	4G	5G
standards	AMPS	IS-136, GSM Groupe Spécial Mobile	UMTS 3GPP R4 - R7	LTE R8-R9, R10-R14	3GPP 15, 16
era	1980s	1990s	2000s	2010s	2020s
services	analog voice	digital voice messages	WB voice packet data	voice, video Internet, apps	*everything*
devices					
data rate	0	100 kbps (GPRS)	10 Mbps (HSPA)	100+ Mbps (LTE/LTE-A)	10 Gbps (NR)
delay		500 ms	100 ms	10s ms	5 ms

What's wrong with 4G?

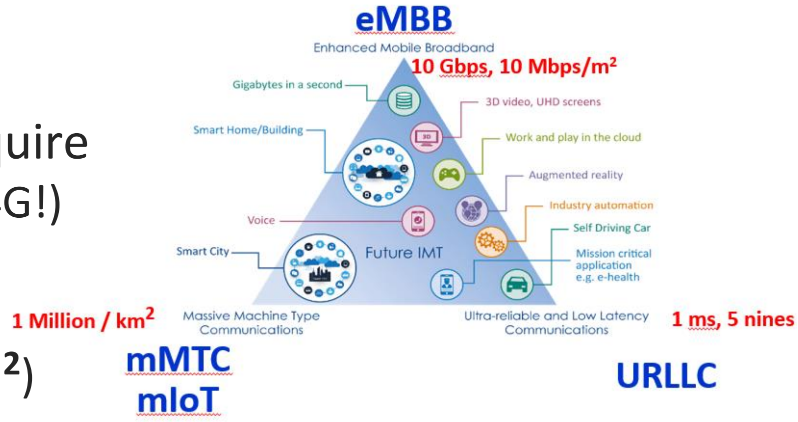
Your Network's Edge®

4G made possible:

- fast always-on Internet
- real-time video reception and creation
- apps relying on location and identity

but doesn't support new applications that require

- much higher data rates (100 times more than 4G!)
- much lower delay (as low as 1 millisecond)
- ultra high reliability (> five nines!)
- much higher connection density (1 million / km^2)



5G is designed to address all of these needs

if we get it right – there will be no need for future generations!

5G is coming fast!

RAD
Your Network's Edge®

You may think that 5G is *futuristic*, but it is coming *fast*

In June 2016, 3GPP accelerated the standardization work-plan

- 5G phase 1 (release 15) finished June 2018 for trials in 2019
- 5G phase 2 (release 16) to be finished by March 2020
- analysts predict that by 2025
  - 50% of US
  - 30% of European
  - 25% of Chinese mobile connections will be 5G

You'll see 5G in 2019 for sure

Qualcomm CEO Steve Mollenkopf

I want 5G, and even 6G, technology in the United States as soon as possible

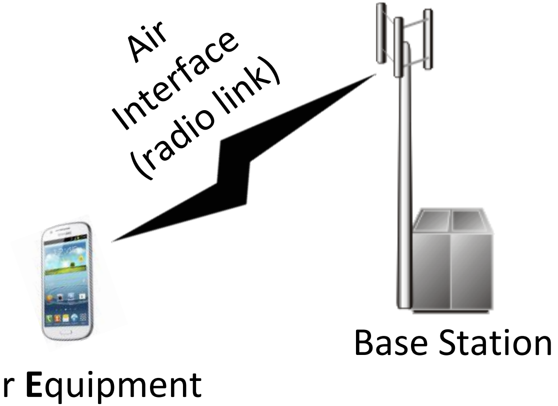
Donald Trump

Mobile communications?

When they hear *mobile* or *cellular* communications

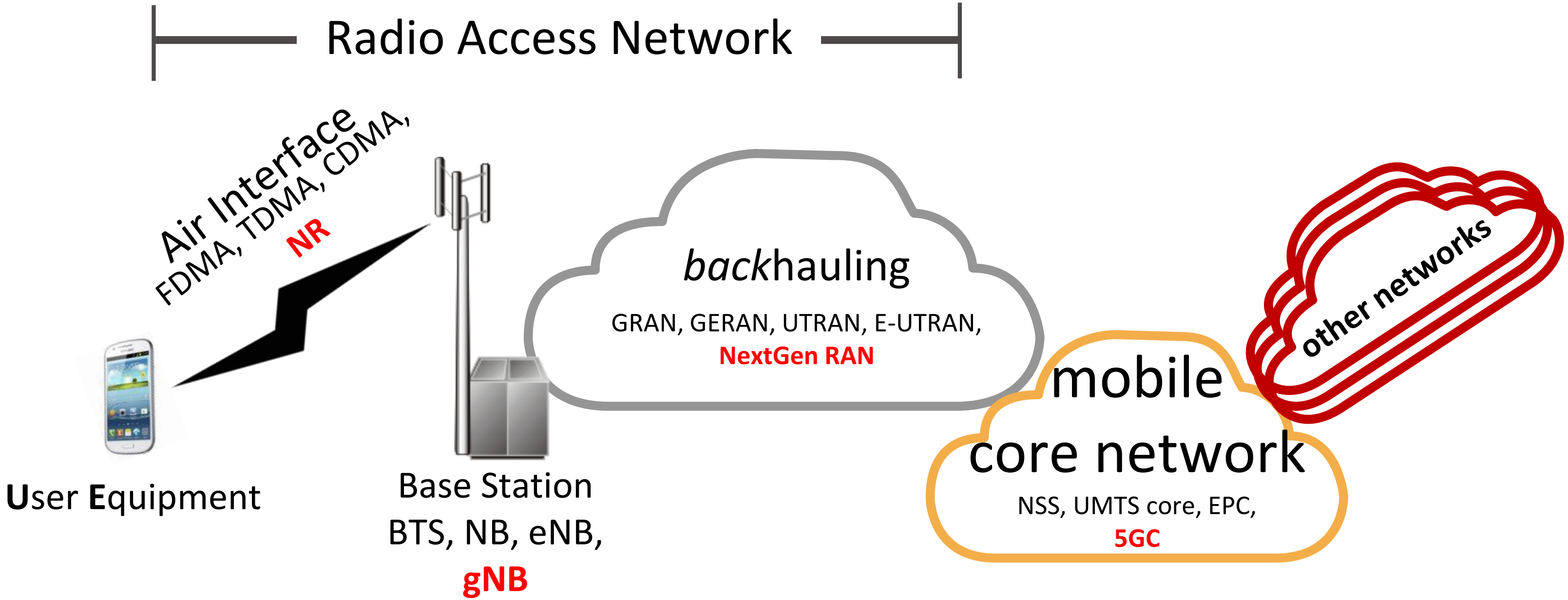
most people think only about the radio link (air interface) between

- a mobile phone (**U**ser **E**quipment) and
- a cellular base station (BTS/nodeB)



Mobile segments

In reality mobile communications involves multiple segments and 5G requires re-engineering all of them!



New Radio is only the first segment!

Your Network's Edge®

The 5G air interface uses

- more efficient modulation (**N**ew **R**adio)
- new RF bands (including mmWave bands)
- wider spectral channels (100 MHz, 1GHz)
- massive MIMO
- higher cell density

in order to attain 10-100 times higher data rates and very low latencies

But how can the backhaul segment

- keep up with these rates ?
- attain low latencies ?
- achieve ultra high reliability ?

Upgrading backhaul data rate

Your Network's Edge®

Initial expectations for Release 15 backhaul are 5 Gbps per site (plus 4G traffic) which exceeds present 1G backhaul links

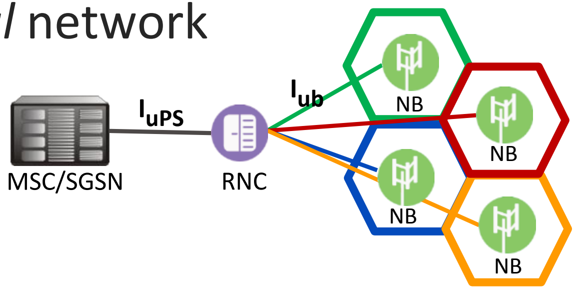
A minimalistic approach would be to upgrade backhaul networks to 10GbE

However, such a minimalist approach won't suffice for long since within a few years the capacity is expected to more than double and this approach doesn't address the latency constraints

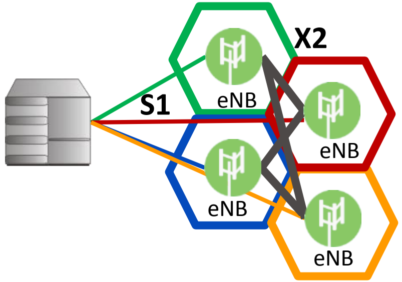
A more drastic overhaul of the backhaul segment is needed

Splitting up the 4G RAN

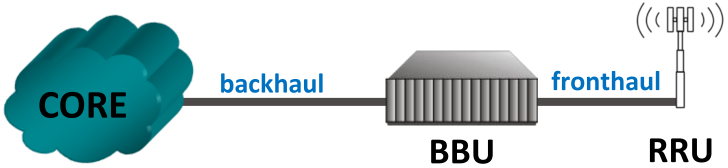
In 3G the RAN was a pure *backhaul* network



In 4G this changed in 2 ways



- the X2 interface interconnected eNBs (at least logically)
- fronthaul (e.g., CPRI) decomposes the eNB
  into Remote Radio Unit
  and BaseBand Unit

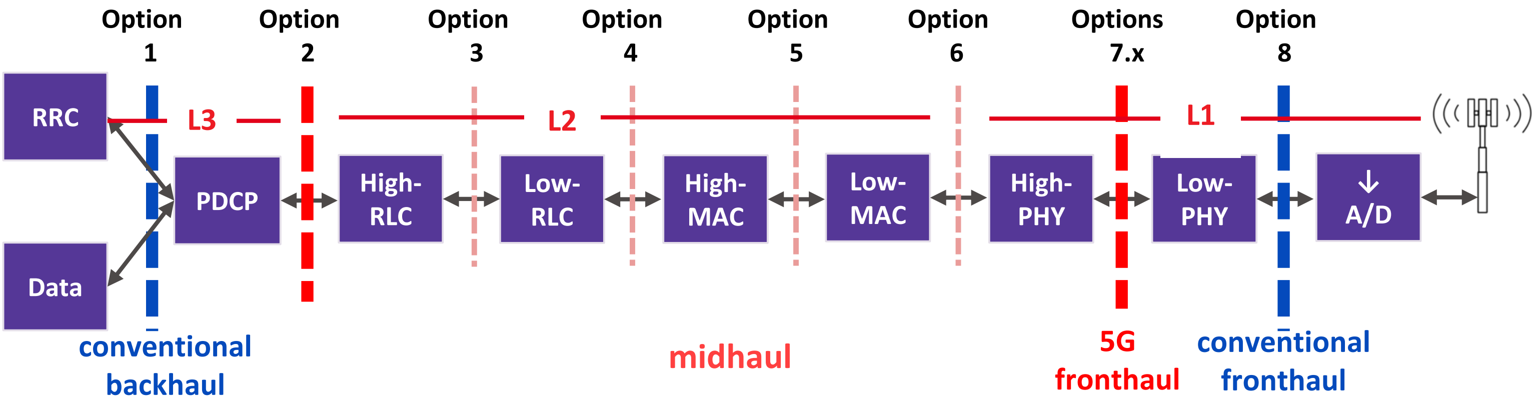


5G brings even more dramatic changes

Functional split options

Your Network's Edge®

5G allows further decomposing of the gNB
defining new xHaul *function split* options
with intermediate data-rate and latency requirements

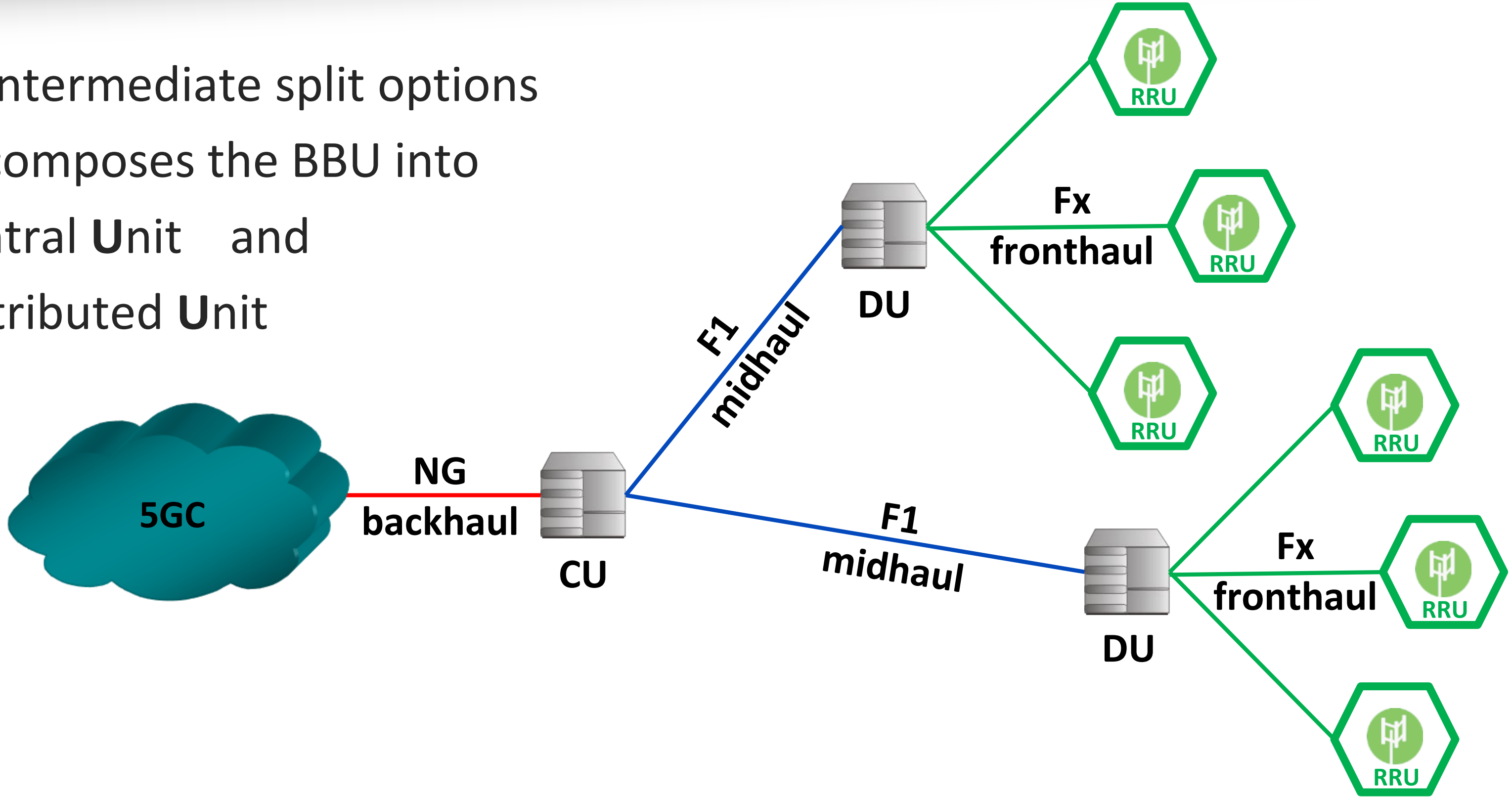


3GPP is standardizing split option 2, while ORAN is standardizing split option 7.2

5G interfaces

Using intermediate split options decomposes the BBU into

- **C**entral **U**nit and
- **D**istributed **U**nit



5G fronthaul bandwidth

Your Network's Edge®

We saw that 5G backhaul requires somewhat higher data rates
what about other functional split options?

5G fronthaul options can consume ridiculous amounts of bandwidth

Fronthaul ball-park estimate:

The *sampling theorem* tells us that we need to sample at least twice the BW
so a single **100 MHz** signal requires 200 Msps or 3.2 Gbps (without overhead)

Assuming 3 sectors each with **16 MIMO** antennas : > **150 Gbps**

Assuming 3 sectors each with **256 MIMO** antennas : > **2.5 Tbps**

Assuming **1 GHz** bandwidth, 3 sectors, **256 MIMO** antennas : > **25 Tbps**

for comparison EoY 2016 the *entire* Internet was 100 PB/month ≈ 300 Tbps

Carrying these high rates requires entirely new transport technologies

Backhaul topologies

Your Network's Edge®

Despite delay constraints of X2 interfaces between neighboring eNBs

4G backhaul networks (like other access networks) are typically

- star (for small backhaul networks)
- tree (implemented actively or using PON)
- rings (for resilience)

and not meshes

The main constraint is monetary – it is prohibitive to run fiber between cell sites

These topologies are expected to continue to dominate for now

except in certain cases, such as

- small cells
- self-backhaul

Potential 5G RAN transport technologies

RAD

Your Network's Edge®

Here are a few transport innovations that can support 5G xHaul requirements

- 10GbE, XGS-PON
  - but 10 Gbps will only be satisfactory for initial 5G deployments
- 25 GbE (802.3by), 1-lane 50 GbE (802.3cd)
  100/200/400 GbE (802.3bs)
- FlexE
- **M**obile (**M**ulti-access) **E**dge **C**omputing
- Synchronization (SyncE, IEEE 1588, DGM)
- Network slicing
- **T**ime **S**ensitive **N**etworking (and **Det**erministic **Net**working)
- Frame Replication and Elimination (IEEE 802.1CB)

high data rates

low latency

ultra high reliability

Higher rate physical layers

Your Network's Edge®

Ethernet physical layer rates are typically multiples of 10

- 10Mbps, 100Mbps, 1Gbps, 10Gbps, 100Gbps, ...

The present 100Gbps standard is based on 4 *lanes* of 25 Gbps

- so it is natural to allow the use of a single lane for rates higher than 10 Gbps

Single lane 25G has been a standard Ethernet rate since 2016

For yet more flexibility, the FlexE standard enables m*25G

IEEE is working on increasing the lane speed to 50G and eventually to 100G

- resulting in 4 lanes with capacity of 200G and eventually 400G
- which FlexE can further bond together

Time Sensitive Networking

TSN and DetNet ask – how can we improve network performance
if we have highly accurate synchronization (say, better than 1 μsec)
at network elements (Ethernet switches, IP/MPLS routers) ?

We'll see that we can

- significantly *reduce* latency
- achieve *bounded* latency

for time sensitive flows

TSN and DetNet support co-existence of sensitive and non-sensitive traffic
(sensitive traffic can be up to 75% of the total load)

TSN uses a control protocol (SRP) for configuring switch time behavior

Frame preemption

A major source of residence latency for a high priority Ethernet packet results from waiting in a queue for completion of packet transmission

For example, assuming a 1500 B packet just started transmission the high priority packet needs to wait:

- 10 Mbps 1.7 msec
- 100 Mbps : 170 μsec
- 1 Gbps : 17 μsec
- 10 Gbps 1.7 μsec
- 100 Gbps 0.17 μsec

and the situation will be much worse with jumbo packets

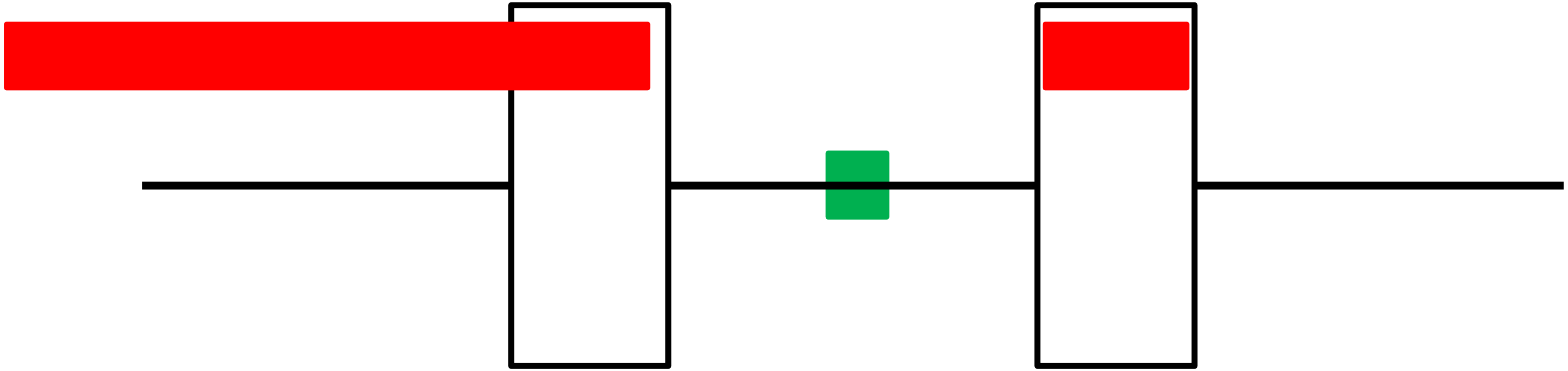
Eliminating this queuing time will greatly reduce residence latency without starving background traffic

802.1Qbu Frame preemption

Preemption is only relevant for the Ethernet physical layer
and only occurs between 2 neighboring switches
but does not require accurate synchronization

When an express frame arrives and a normal frame is being transmitted

- normal frame transmission is temporarily suspended and express frame is forwarded
- transmission of the normal frame is continued
- the neighboring switch reassembles the normal packet and forwards



802.1Qbv Scheduled Traffic

Time Aware Traffic Shaping (Time Sensitive Queues)

- requires that every network element have highly accurate time (e.g., from 1588)
- time-gated egress CoS queues transmit based on a precise *timeslot* schedule
- implemented by circular collection of *time aware* gates

Directly timing release of packets can

- support *scheduled* applications (e.g., process/vehicle control)
- provide latency and PDV guarantees
- completely avoid congestion
- return to TDM-like determinism

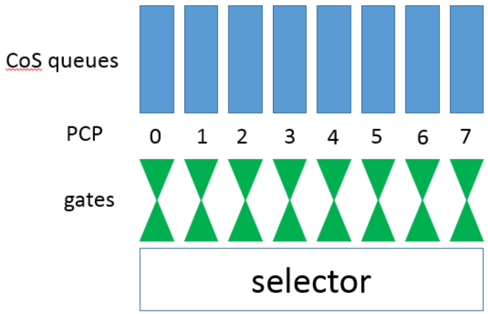
Qbv retains credit-based shapers for *non-scheduled* applications

Time Sensitive Queues

Your Network's Edge®

Queues are cyclically gated with granularity up to 1 ns
(implementations may be less precise)
thus PDV can be reduced to about 1 ns

Timeslot schedules are dynamically computed
by a centralized management system
that configures the network nodes using the **S**tream **R**eservation **P**rotocol



Gate Schedule

T	0	1	2	3	4	5	6	7
0	O	C	O	O	C	O	O	O
1	C	O	O	O	O	O	C	C
2	O	C	C	C	C	O	O	O
3	O	O	O	O	C	C	C	O
				. . .				
				. . .				
N				*repeat*				

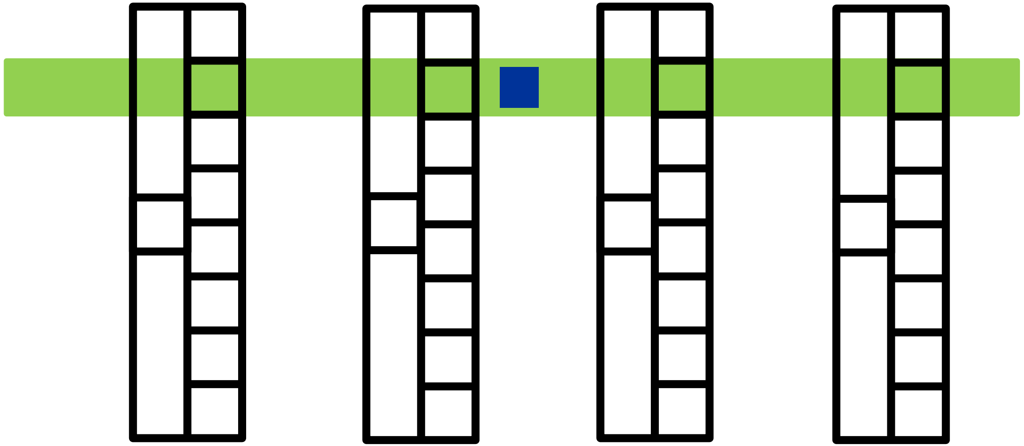
802.1Qci/Qch Cycling Queuing

Your Network's Edge®

Cyclic Queing (formerly *peristaltic shaping*) provides a (*nonoptimal*) upper latency bound

It exploits accurate timing without requiring intricate signaling

All switches release packets of the same traffic class at once
    but the PCP field is incremented when traversing the switch

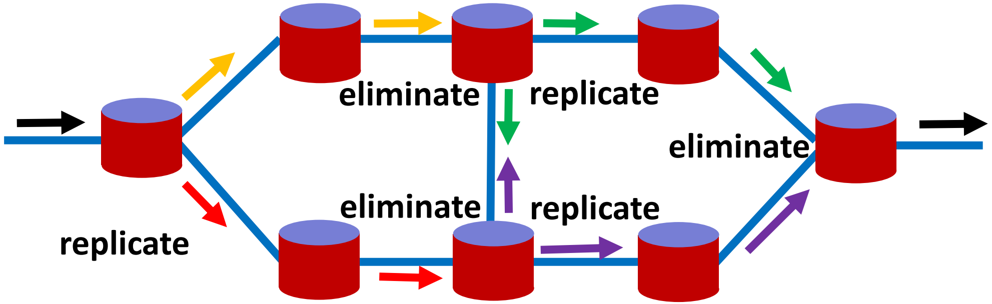


802.1CB Frame Replication and Elimination

Also developed by TSN and DetNet for Ethernet, IP, and MPLS

Achieves very low PLR (better than 10^{-6}) and ultra high reliability

- seamless redundancy by 1+1 replication + elimination
- essentially no congestion-related packet loss
- no failure detection needed
- increases network traffic load



Mobile Edge Computing

RAD
Your Network's Edge®

Mobile (**M**ultiaccess) **E**dge **C**omputing offers another solution
to both data-rate and latency requirements

MEC enables terminating traffic close to the gNB or first aggregation nodes
rather than backhauling all the way to the core

By providing processing power close to the UE
network congestion and latency are both reduced

Some MEC applications

- Internet breakout
- Content Delivery Networks
- mobile big data analytics
- DNS caching
- fog networking (IoT processing)
- connected car (V2x)

MEC concepts have been absorbed into 5G's **S**ervice **B**ased **A**rchitecture

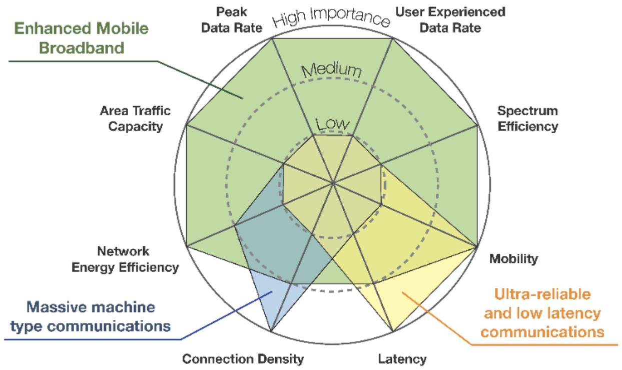
Network slicing

Your Network's Edge®

5G can't reach all of its goals simultaneously - but it doesn't have to!

For example:

- eMBB needs high data rates
  but doesn't need very low latency
- massive IoT needs high connection density
  doesn't need high data rates



So, 5G uses *network slicing*:

- *on-demand* assignment of networking/computational resources
  - bandwidth, forwarding tables, processing capacity, etc.
- resources can be physical or virtual, but
- each slices acts as a strongly isolated network or cloud
  - isolation of management, security, and performance

Implementing slicing

Your Network's Edge®

Slicing must be end-to-end - both RAN and core network must support slicing

Slicing requires programmability, flexibility, and modularity
in order to create multiple virtual networks over a common network

Different slices can be labeled using VLAN ID, VXLAN, IPv6 flow label, DSCP
but a slice is different from a separate physical network or a VPN
because of dynamic set-up / release requirements and separate management

SDN techniques (network softwarization) are used to achieve slicing

- APIs provided to specify requirements
- service definition may include NFV/MEC elements
  particularly useful for ultra low latency cases
- use of global orchestrator
  for rapid yet highly optimized slice creation

Timing for 5G

RAD
Your Network's Edge®

Frequency and time accuracy requirements are defined
to assure efficient and proper functioning of the air interface

RAN timing requirements are becoming stricter from generation to generation

Accurate time is also needed for some TSN functionalities

Base stations obtain timing from the RAN
(unless they have a local source of timing, e.g., GNSS)

5G requirements will be at least as strict as 4G
and some experts are speaking of them becoming significantly stricter

So the 5G RAN must deliver ever more accurate timing!

Delivering timing

Using a conventional centralized GrandMaster clock

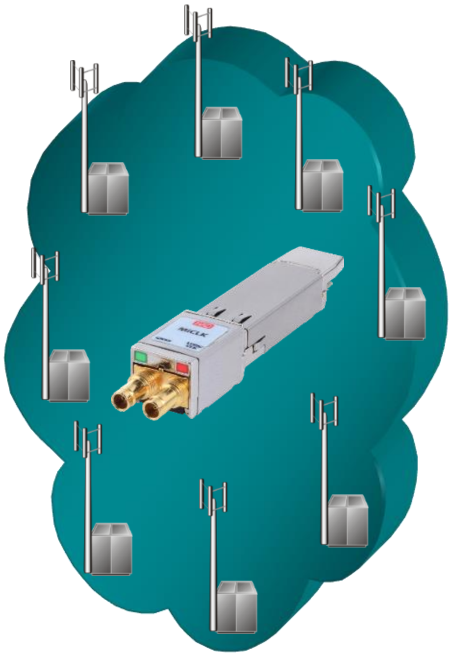
- will require PTP on-path support
- will probably require SyncE

The Distributed GM (MiCLK) approach

which brings the master clock close to the gNBs

may be advantageous

- < 50 nsec accuracy for gNBs served by one MiCLK
- ±125 nsec accuracy for gNBs served by different MiCLKs
- Note that GNSS only guarantees ± 100 nsec accuracy



Takeaways

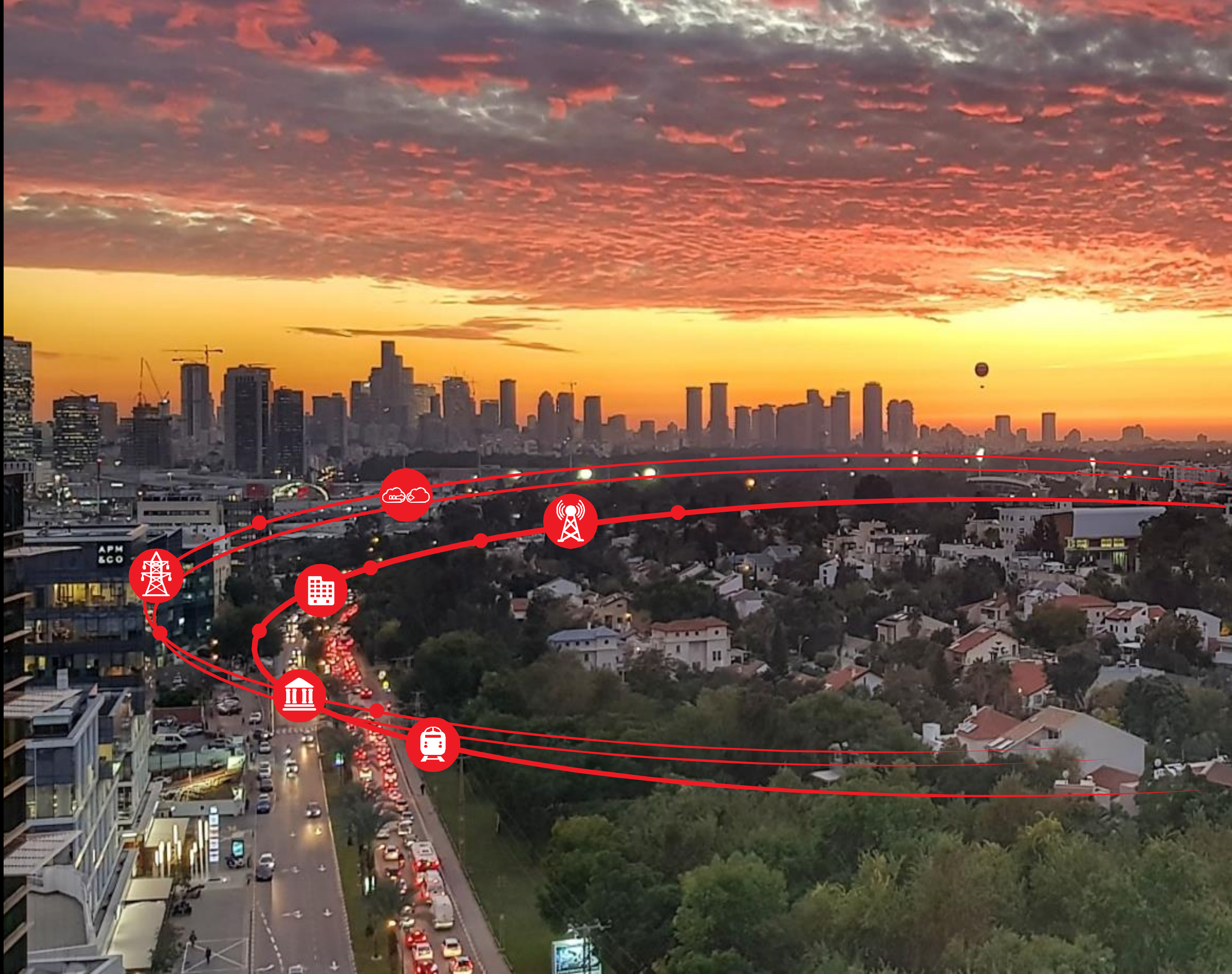
Your Network's Edge®

- For **high data rates**
  - in the short term upgrade the RAN to 10G
  - but 25G and m*25G interfaces will be needed later
- For **low delay**
  - TSN/DetNet has mechanisms that bound delay
  - MEC can help for some ultra-low delay cases
- For **ultra reliability**
  - frame replication and erasure may help
    but at the expense of bandwidth overhead
- To support differing requirements
  - use network slicing
- Timing is everything

Thank you

For your attention

Yaakov (J) Stein
CTO



Your Network's Edge®