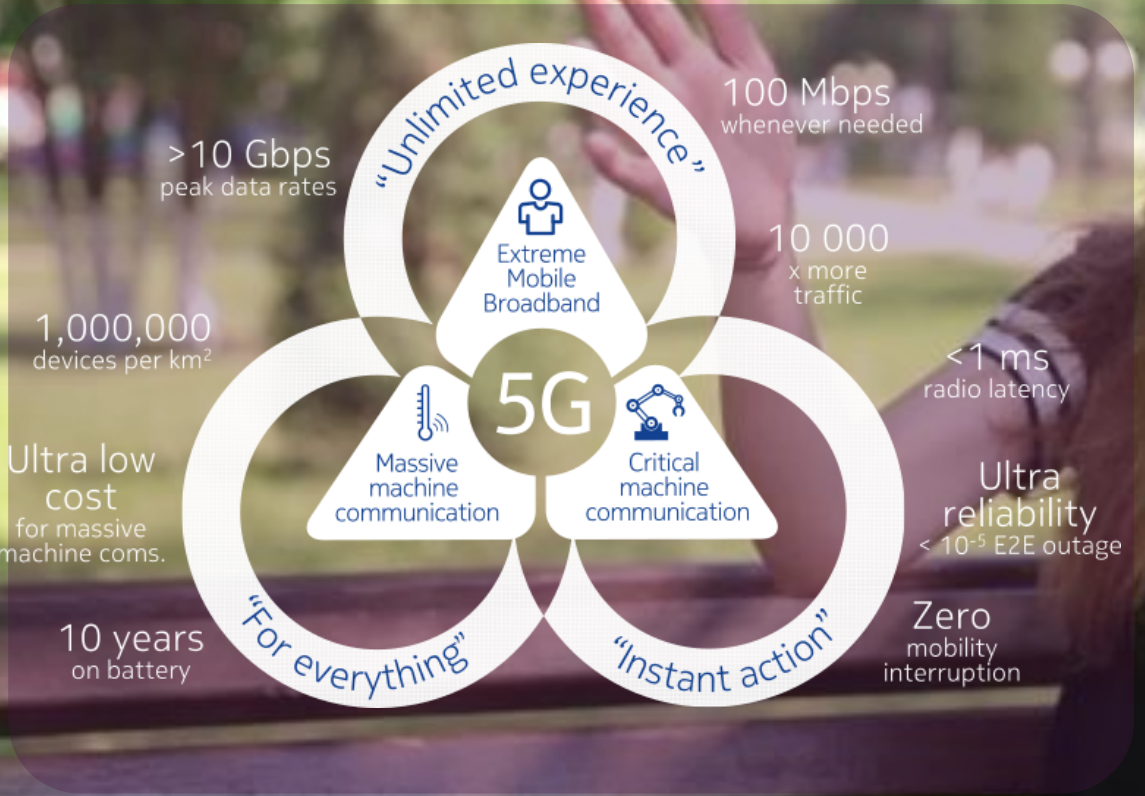
Low Latency Networking

Shivendra Panwar

With: Ish Jain, Thanos Koutsaftis and Rajeev Kumar

NYU Tandon School of Engineering

QoS Requirements



AR/VR requirements

Data Rates	100Mbps-1Gbps
Interruptions	0.1/min
Video stall (pause)	<10 ms

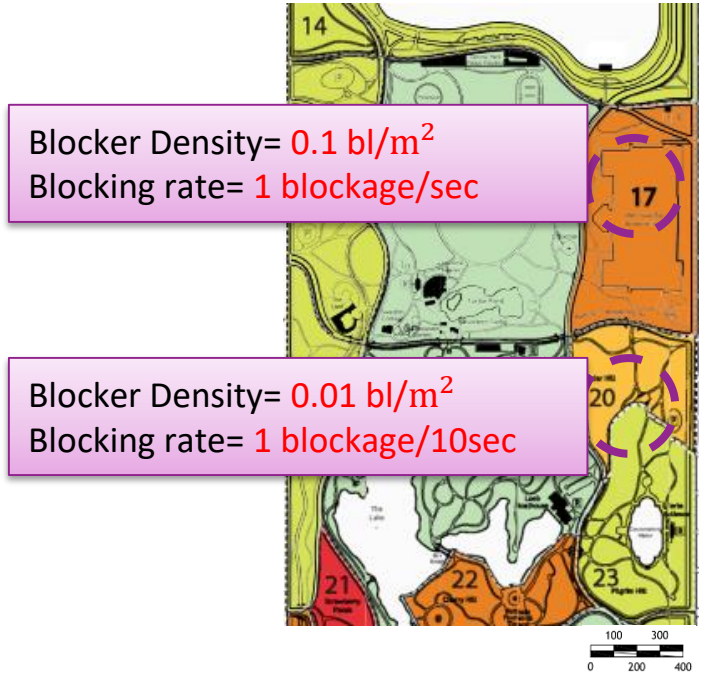
(otherwise causes nausea or sickness)

Source: Nokia, VR/AR in the 5G Era NEM Summit November 23, 2016

Image source: https://videohive.net

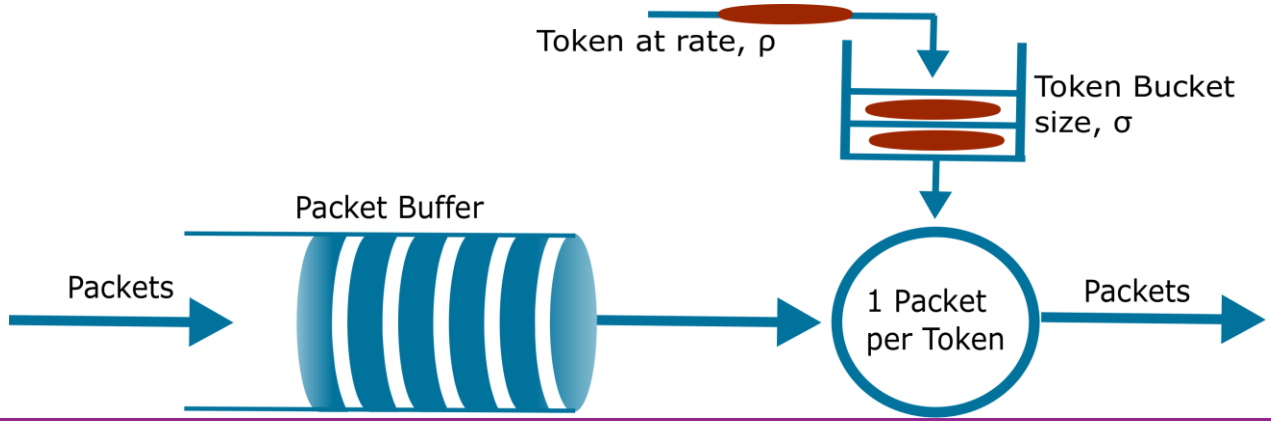
Can 5G mmWave networks meet QoS requirements?

- In an urban mmWave cellular network:
  - For blockage probability 1e-3, a high BS density (350 BS/km^2) is required.
  - NLOS paths may reduce the BS density to 270 BS/km^2, but, still the requirement is very high.
- Need for new protocols to meet the stringent QoS requirements



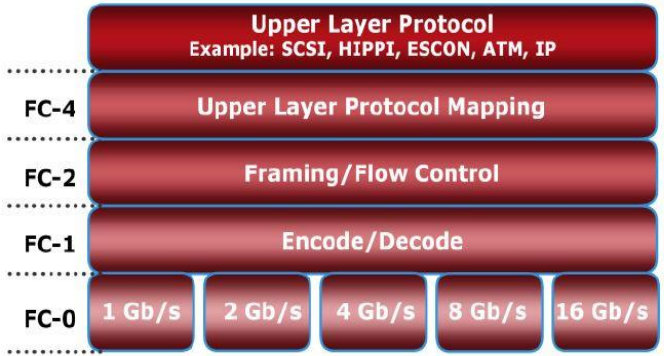
Revisiting past technologies and protocols (1)

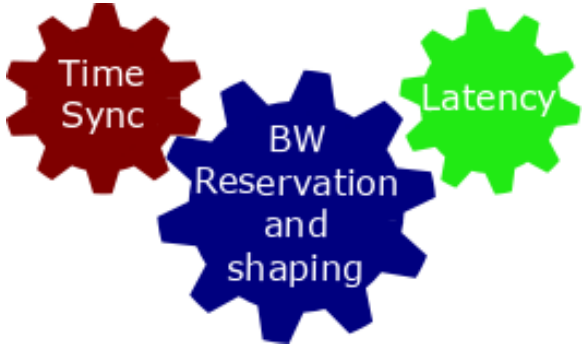
- Circuit switching (and fast circuit switching):
  - Dedicated communication channel before nodes start communicating
  - For 5G: Multiple paths will be required to counter the effect of blockages will lead to network inefficiency
- Sigma-rho (σ, ρ) Regulators: σ: maximum backlog, ρ: average access rate
  - Can obtain an upper bound on maximum network delay



Revisiting past technologies and protocols (2)

- Fibre Channel Protocol (FCP)
  - Provides in-order, lossless delivery of data (can support up to 128 Gbit/s)
  - Primarily used for connecting computer storage data to servers
  - Transport protocol that predominantly transports Small Computer System Interface (SCSI) commands over Fiber channel networks
- Time-Sensitive Networking (TSN)
  - TSN task group was formed in Nov. 2012
  - Transmissions are based on a global time and schedules slots
  - New features to standard Ethernet: time synchronization, traffic scheduling, and system configuration





Next generation networks: Future inspired by past

- Deterministic Networking (DetNet) and TSN
  - Dynamic sharing of resources instead of holding up
  - Working on guarantees instead of predictions

Image source: https://www.tttech.com

Time Sync	BW Reservation and shaping	Latency
Synchronization of all devices in the network upto sub-ms precision, for precise coordination	Devices reserve a percentage of bandwidth (e.g. IEEE 802.1), and traffic shaping (credit based, or time aware)	Worst-case guarantees on latency, earliest deadline first, priority queuing, …

P4 to achieve DetNet & TSN:
Domain-specific programming language to describe the data plane procedures

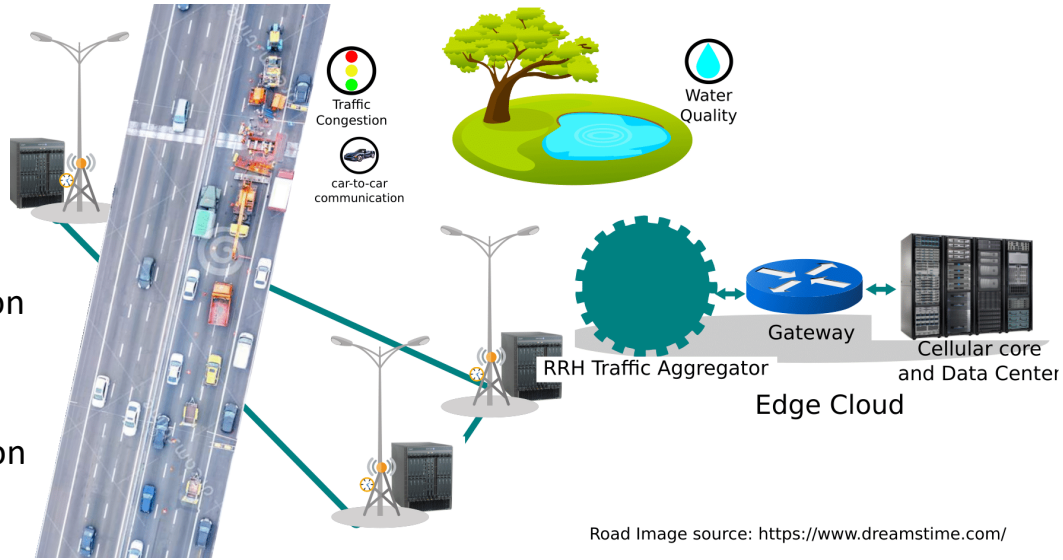
Future low latency networks

Network architecture and resource management	• Pushing computation power and storage close to end users, data center at the edge
Scheduling	• Precise timestamping of packets • Earliest due date deadline scheduling • Rate-controlled scheduling
Recent advances in transport layer protocols	• Bottleneck Bandwidth and Round trip propagation time (BBR) transport protocol • Quick UDP Internet connections (QUIC)

Network architecture and resource management (1)

Distributed computing and storage

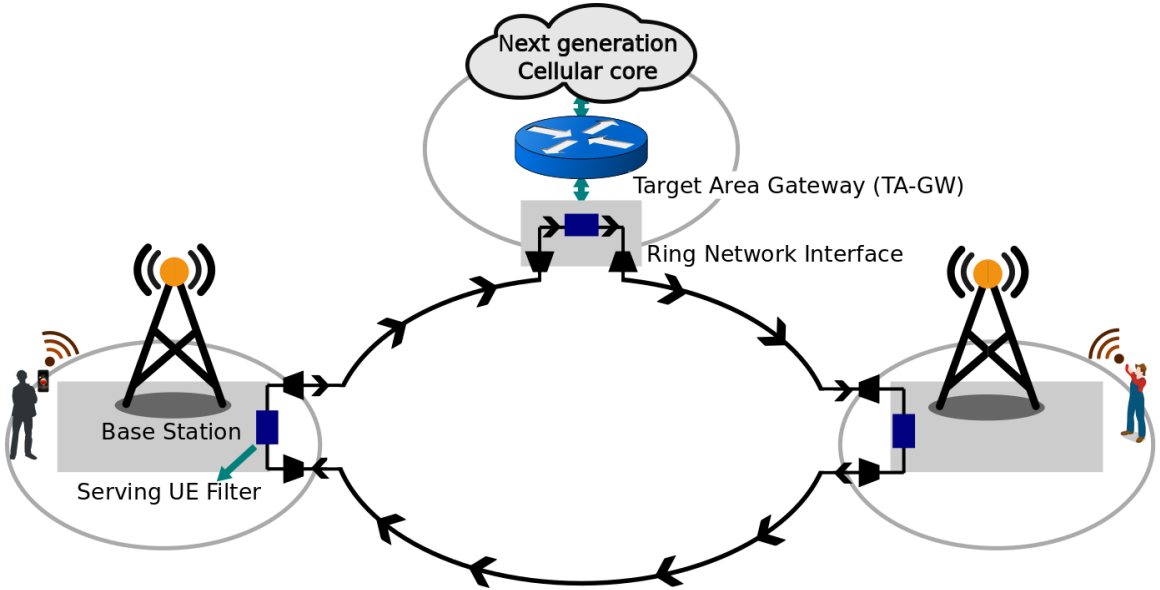
- RRH with small computing and storage resources
- Need to aggregate resources with data rack like internal delays
- Global network time synchronization
- Insufficient computing resources
  - Forward data to nearest RRH or edge-cloud for computation
    - Enhanced FCP?



Networked Robotics: migration of UE and IoT computations to the distributed computing devices in the network to achieve fast decision making and enhanced battery life

Network architecture and resource management (2)

Ring Packet Forwarding to mitigate blockage initiated handoff delays

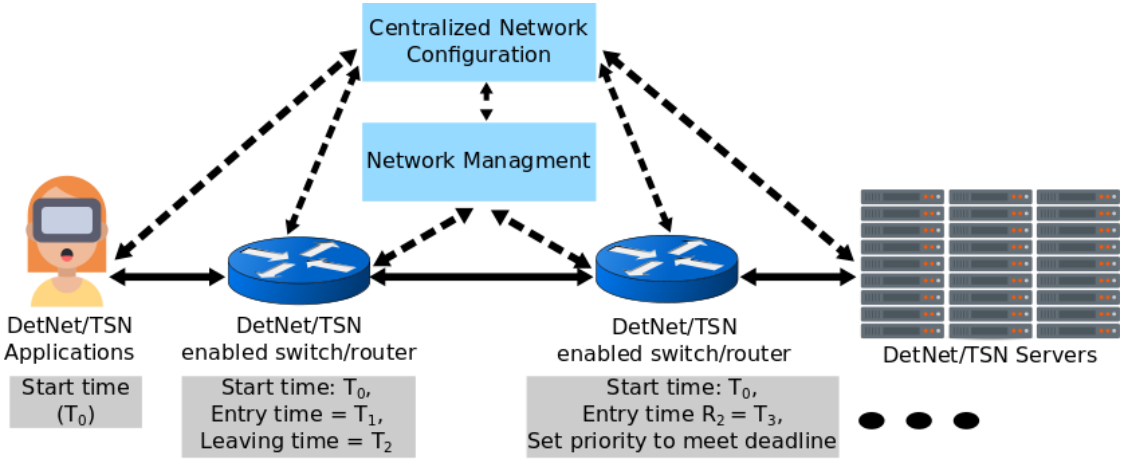


Ring Packet Forwarding: packets keep getting injected in a ring of RRH/BS nodes until it gets served

Scheduling (1)

Precise timestamping of packets

- Useful in determining the amount of time packets spend at each intermediate node
  - Information profile is passed to the network management utility
  - Network management utility configures routers/switches in the path
  - Routers/Switches update scheduling schemes to meet QoS requirements



Scheduling (2)

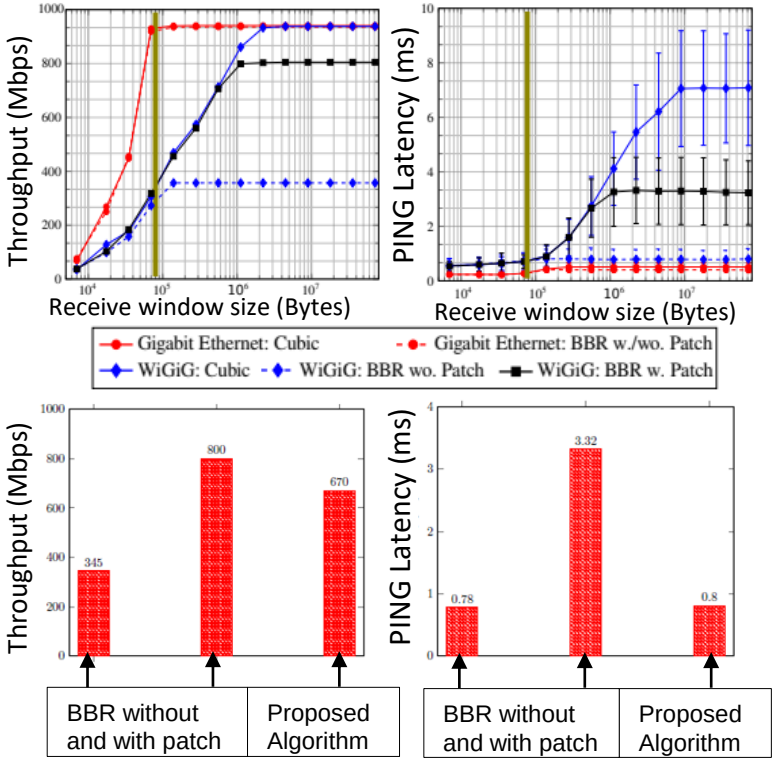
- Earliest Due Date (EDD) scheduling:
  - Packet with earliest deadline is selected
  - Each packer gets a hard delay bound independently of bandwidth requirement
  - Requires per-connection state and a priority queue

- Rate-controlled scheduling:
  - Includes two components:
    - Regulator: Traffic shaping
    - Scheduler: Performance guarantees
  - Packets are stored in the regulator until they become eligible for service
  - Scheduler selects among eligible packets for service

Recent Advances in Transport Layer Protocols (1)

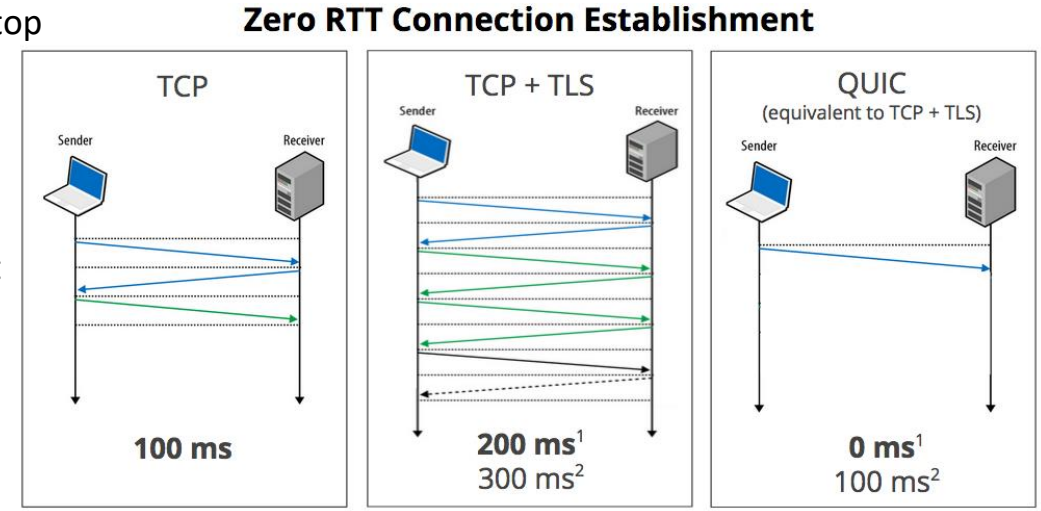
- TCP BBR promises high throughput with low latency
  - Operates around Bandwidth Delay Product (BDP)
- BBR performance degrades significantly in Wireless link due to variations in RTT
  - Google's patch: a higher throughput, at the cost of higher latency
  - No longer operates around BDP
- Recent work* on BBR in wireless links:
  - Justifies losses using mathematical analysis
    - suggests a solution based on analysis
  - Achieves high throughput with lower latency

(* under submission, available upon request)



Recent Advances in Transport Layer Protocols (2)

- QUIC: new Transport Protocol for the Internet
  - Similar to TCP+TLS+HTTP2, but on top of UDP
  - Set of multiplexed connections
  - Aims to accelerate HTTP traffic
- Benefits of QUIC over TCP
  - Reduced connection establishment time
  - Improved congestion control
  - Forward error correction
  - Connection migration
  - Equivalent (if not greater) security



References

1. Jain I. K., Kumar R., Panwar S., *"Can Millimeter Wave Cellular Systems provide High Reliability and Low Latency? An analysis of the impact of Mobile Blockers," arXiv preprint arXiv:1807.04388,* Jul. 2018.
2. M. C. Choo and R. L. Cruz, *"Approximate analysis of average performance of (sigma, rho) regulators," Proc. of IEEE INFOCOM,* Jun 1990.
3. R. W. Kembel, "*Fibre Channel A Comprehensive Introduction*," *Northwest Learning Assoc.*, 2009.
4. S. Kesahv, "*An engineering approach to computer networking: ATM networks, the Internet, and the telephone network," Vol. 1. Reading: Addison-Wesley*, 1997.
5. "Time-Sensitive Networking Task Group", [Online], Available: http://www.ieee802.org/1/pages/tsn.html
6. H. Austad, "Deterministic Networking for Real-Time Systems (Using TSN and DetNet)," Oct. 2017. [Online]. Available: https://bit.ly/2DztRw7
7. G. T. McKee, and P. S. Schenker, *"Networked robotics," Sensor Fusion and Decentralized Control in Robotic Systems III*. Vol. 4196. International Society for Optics and Photonics, 2000.
8. B. Pat *et. al.,* "P4: Programming protocol-independent packet processors," *ACM SIGCOMM Computer Communication Review* 44.3 (2014): 87-95.
9. N. Cardwell *et al., "BBR: congestion-based congestion control," ACM Queue*, vol. 14, no. 5, pp. 50:20–50:53, 2016.
10. N. Cardwell *et. al., "BBR congestion control work at Google IETF 101 update," Tech. Rep.,* Mar. 2018. [Online]. Available: https://bit.ly/2NNB1xW
11. A. M. Kakhki *et. al., "Taking a long look at QUIC: an approach for rigorous evaluation of rapidly evolving transport protocols." Proc. of ACM Internet Measurement Conference,* Nov. 2017