Carnegie Mellon University

Electrical & Computer ENGINEERING

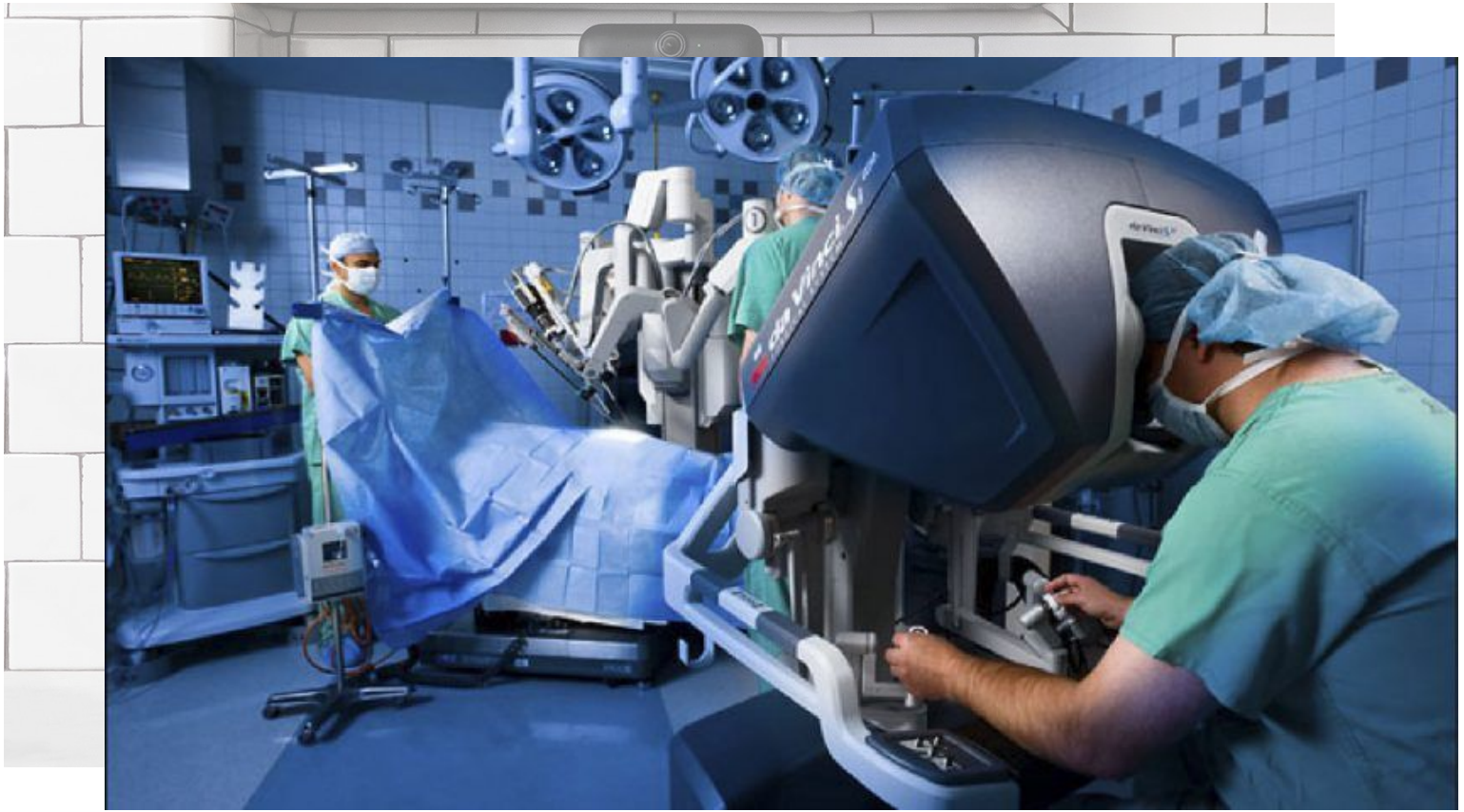
Network Slicing as an Ad-Hoc Service

Opportunities and Challenges in Enabling User-Driven Resource Management in 5G

Madhumitha Harishankar
Patrick Tague
Carlee Joe-Wong

Computing is Connected, Pervasive and Realtime

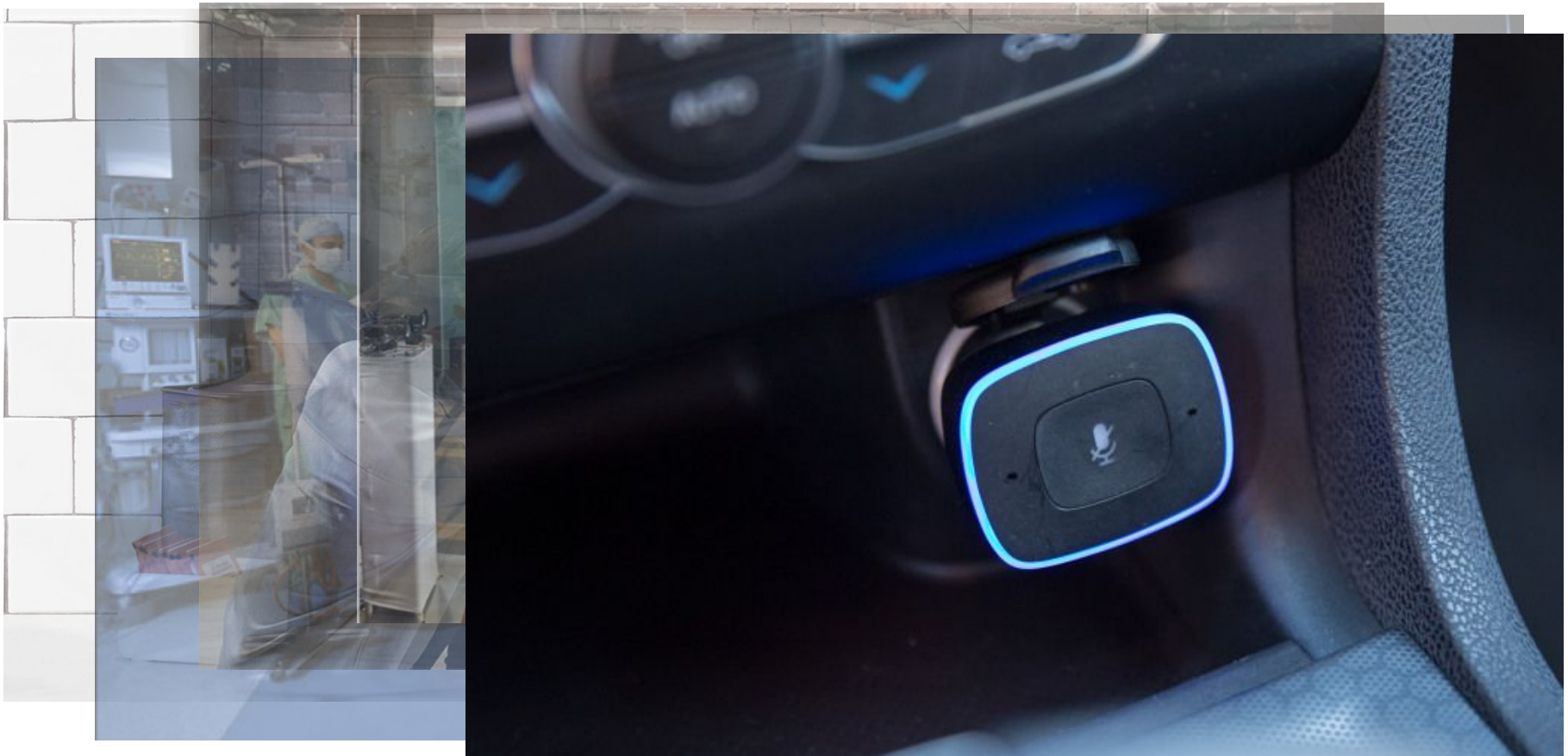
Computing is Connected, Pervasive and Realtime



Computing is Connected, Pervasive and Realtime

Computing is Connected, Pervasive and Realtime

Computing is Connected, Pervasive and Realtime



Yet Connectivity is Uncertain

- Users increasingly rely on ubiquitous connectivity
- But limited influence over their Mobile Experience
  - Erratic streaming quality
  - Unpredictable and disruptive lags
  - Call drops
  - Typically worsens with mobility

The user is unable to influence the resources allocated to their session and must contend with complete uncertainty about session performance.

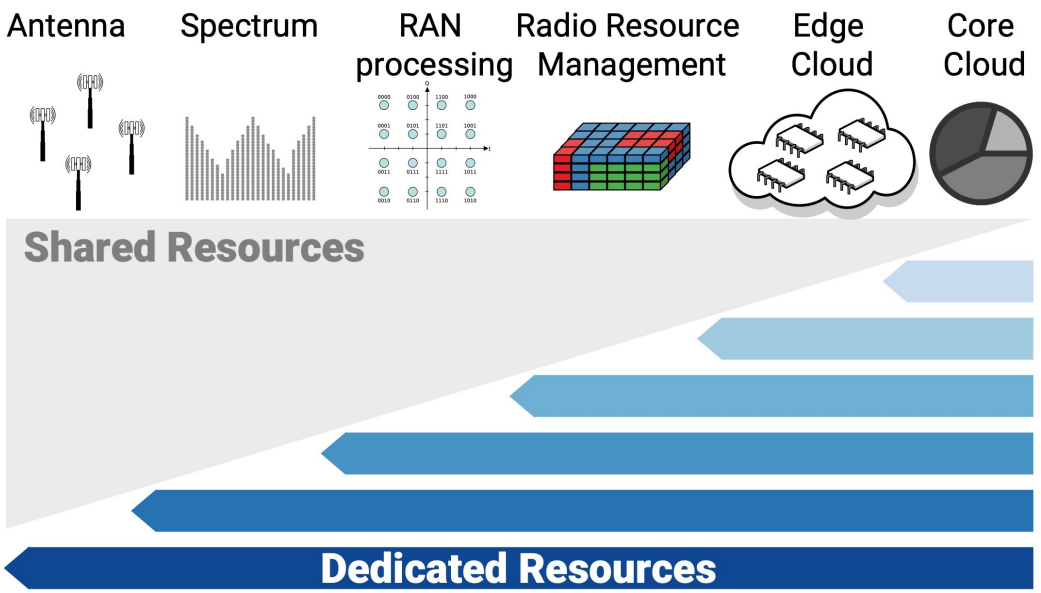
5G aims to provide Reliable Connectivity

- Highly diversified network requirements of emerging applications
- E.g. Videoconferencing <150ms, Tactile Internet <5ms
- Key Architectural Advancements in Networking
  - Virtualized Network Functions (Mobility Management, C-RAN etc)
  - Separation of Control/Data plane resulting in dynamic programmability (SDN)

Emergence of Network Slicing as an Architectural Solution

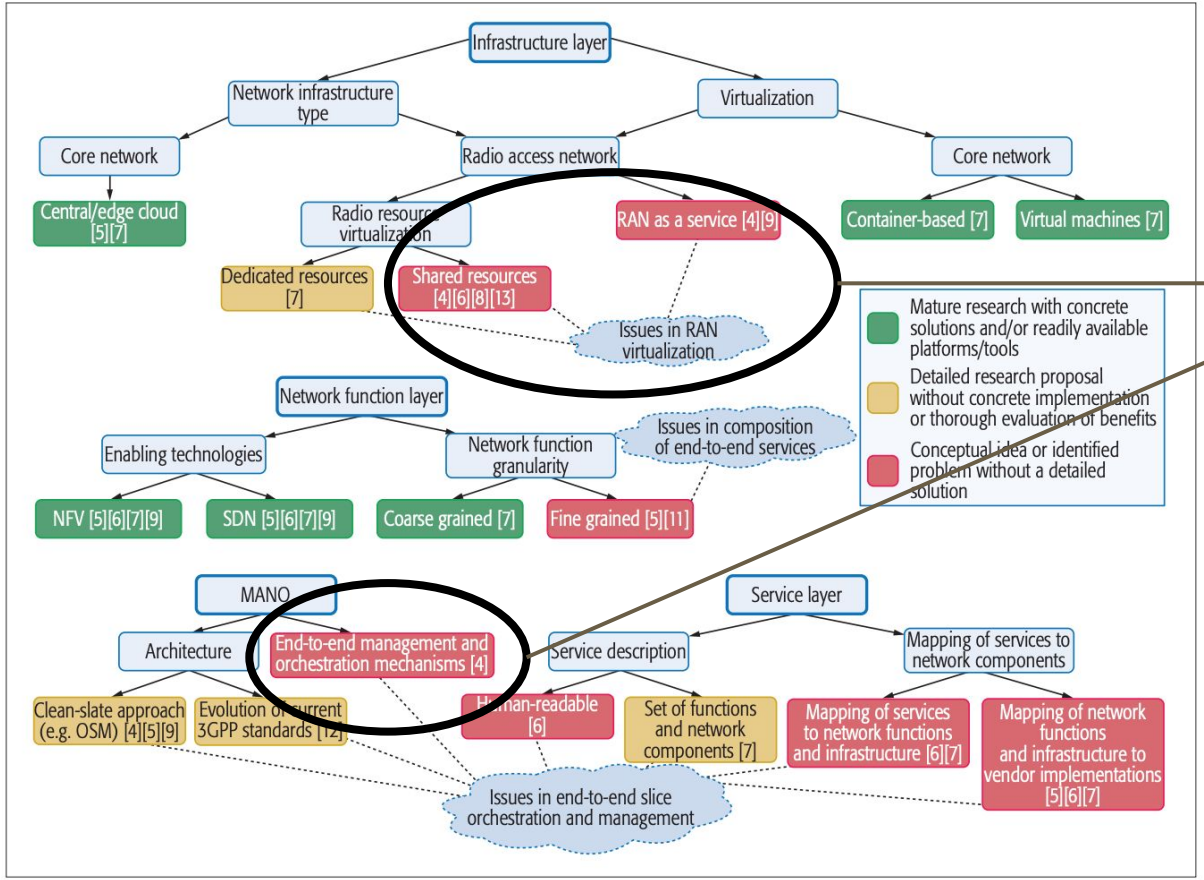
The Slicing Paradigm

- Virtual network slices corresponding to different Service Level Agreements
- Slices of varying depth, with corresponding physical resources allocated
- Note that physical resource limitations still exist



Source: Marquez, Cristina, et al. "How Should I Slice My Network?: A Multi-Service Empirical Evaluation of Resource Sharing Efficiency." *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 2018.

Maturity of Network Slicing Research



This talk

Source: Foukas, Xenofon, et al. "Network slicing in 5G: Survey and challenges." *IEEE Communications Magazine* 55.5 (2017): 94-100.

Current Slicing Models

- Slices corresponding to SLAs
  - Business to Customer (B2C)
  - For instance, IoT slice, V2I slice, Videoconferencing slice etc.
- Slices corresponding to Content Providers
  - Business to Business (B2B)
  - For instance, Netflix slice, Youtube slice, Skype slice etc.
- Access policy
  - Physical resources backing a slice are finite
  - **Who gets admitted to a slice?**
  - ***Reliable connectivity* depends on this**

Slice Admittance and Service Reliability

- B2C - SLA based slices
  - Must forecast applicable slice (i.e. QoS needs) up-ahead which can be infeasible (Realtime applications, IoT use-cases)
- B2B - Content-provider model
  - Users rely on content-providers to procure slices for their applications

In both cases, the centralized entity solely prioritizes users for admittance to a slice. Users hence exert limited influence over their connectivity.

Approach: Incentive-Aware Slice Admission

- Admit users to slices based on who needs it most
  - Align users' utility for desired SLA with their willingness to pay via incentive mechanism design
  - Capture user valuations for SLA *spontaneously* as and when they engage in application use
  - *Monetize this as a form of admission control*
  - User-Operator win-win

Users drive their slice admittance (and thereby quality of experience) with their real-time utility and stated valuations.

Slicing as a Service for Edge Entities

- Offer slicing as a *realtime service for session-oriented SLAs*
  - Directly to end-users
  - Arbitrary application specific SLAs may be accomodated
  - Users' value for resources influences their valuations
  - Valuations influence slice admittance
  - Preference-based slice admittance instead of generic network policy
  - *Made possible today because of maturity of virtualization*

Ask and Pay for a session's required resources in real time.

Incentive Mechanism Design

- Session-level slice characteristics include
  - Bandwidth/Latency required
  - Duration of session
  - Location of access (mobility)
- Incentivize *truthful* declaration
- Compute slice allocations that *maximize collective welfare*
- Compute allocations quickly for usability
- Account for opportunity cost of realtime admittance
- Charge users for *what they actually use*

A Combinatorial Auction Approach for Reducing Last-Mile Uncertainty in the Performance of Realtime Applications

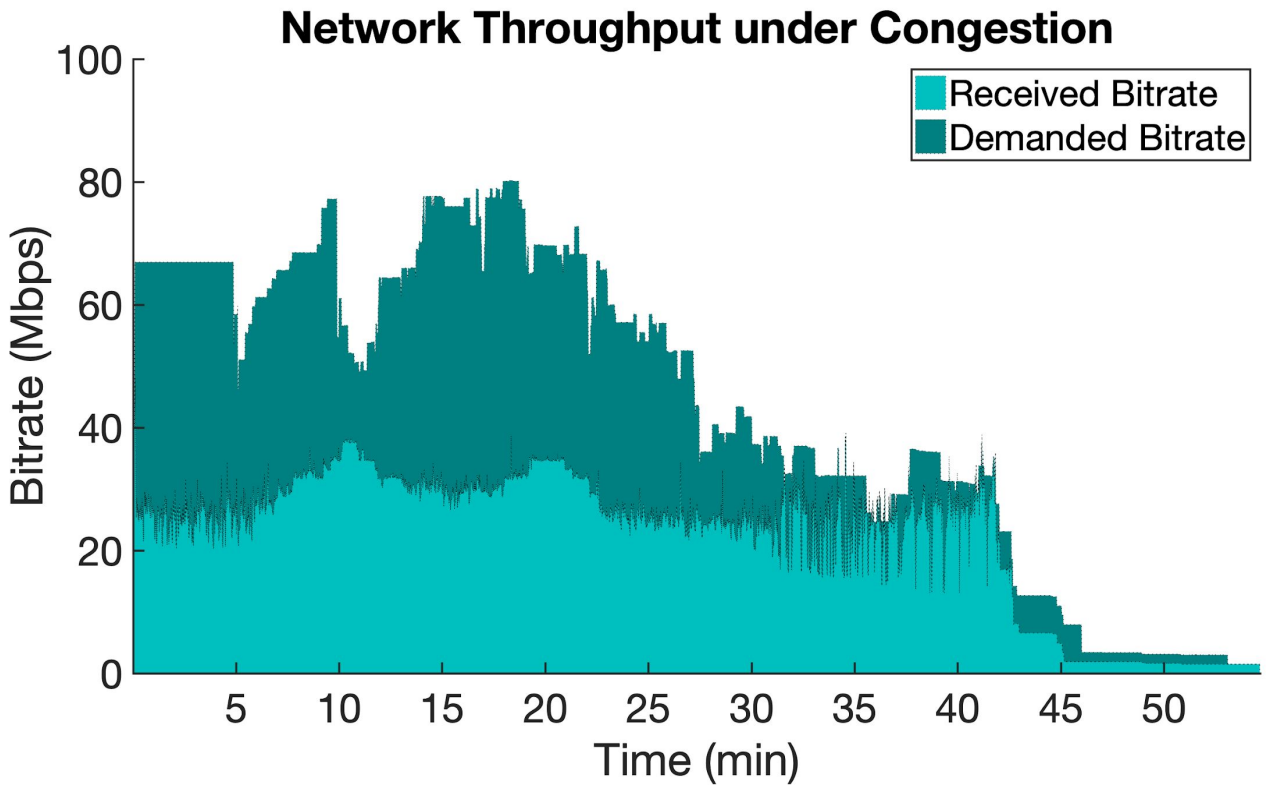
- Realtime applications especially benefit from SLA guarantees
  - Inflexible resource needs (low latency, no buffering)
- Reconcile changing resource demand and limited supply with periodic and fast auctions
  - Combinatorial auction modeling time and location as dimensions
  - Exploit real-time nature to compute winning slice allocations quickly
  - Incentivize truthfulness in user-stated slice requirements and valuations by accounting for expected future bids

But is it feasible to make application-oriented performance guarantees in the wireless link?

Assessing Practical Viability of RAN Slicing

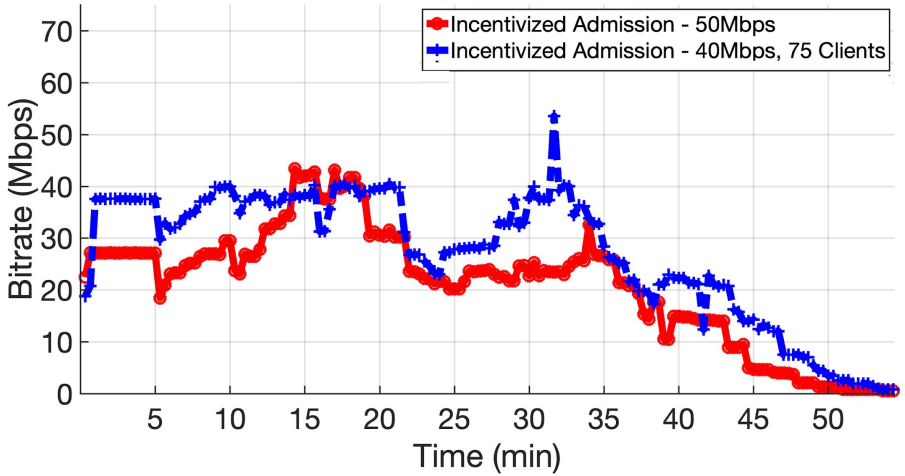
- How can we forecast “available resources” and ensure that SLA guarantees are met for sessions of admitted users?
  - Uncertainty inherent in wireless links (e.g. fading effects)
  - Links are further affected by presence of other links (congestion externalities)
  - How many users and slices can we accommodate while ensuring performance isolation?
- WiFi experiment - Cafe scenario
  - One AP - 802.11g
  - ~50 clients initiating multimedia traffic on the network (Video/Audio/Realtime Conferencing) and web browsing

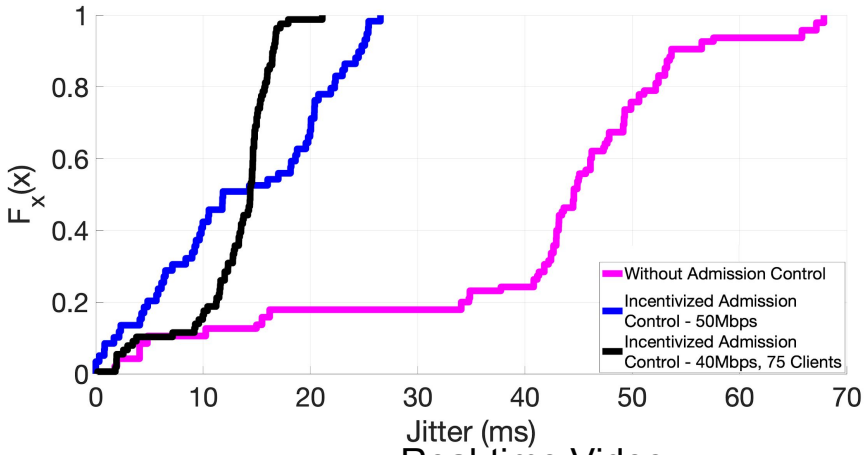
Preliminary Results - Congestion without AC



Preliminary Results - Feasibility of meeting guarantees

- With incentive-aware admission control:
  - Number of admitted clients in the network *increase* =>IA
  - Network throughput *increases* => AC
  - SLAs of Realtime Applications are entirely met => AC

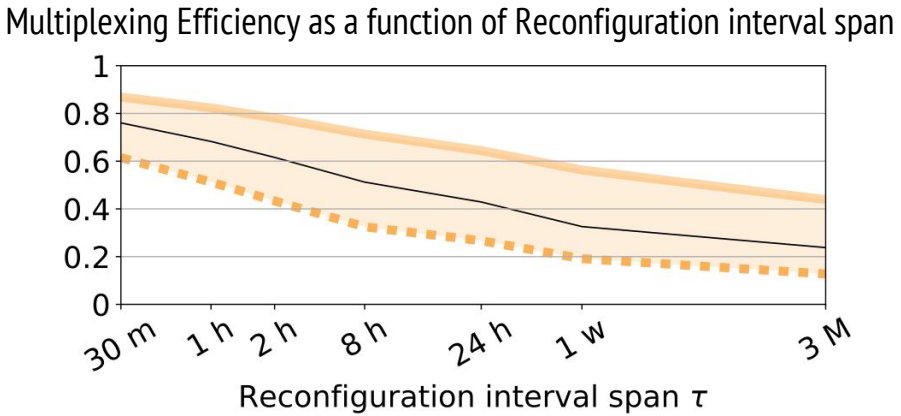




Real-time Video

Open Questions

- Slicing reduces traffic multiplexing efficiency
  - Resources of a slice are multiplexed only between traffic demand for that slice (rather than between all demand)
- Impact of Ad-hoc slicing on resource efficiency?
- Mobility and slicing?



Source: Marquez, Cristina, et al. "How Should I Slice My Network?: A Multi-Service Empirical Evaluation of Resource Sharing Efficiency." *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 2018.

Thank you.
Questions?