RAN Resource Usage Prediction for a 5G Slice Broker

Craig Gutterman*, Edward Grinshpun^, Sameer Sharma^, Gil Zussman*

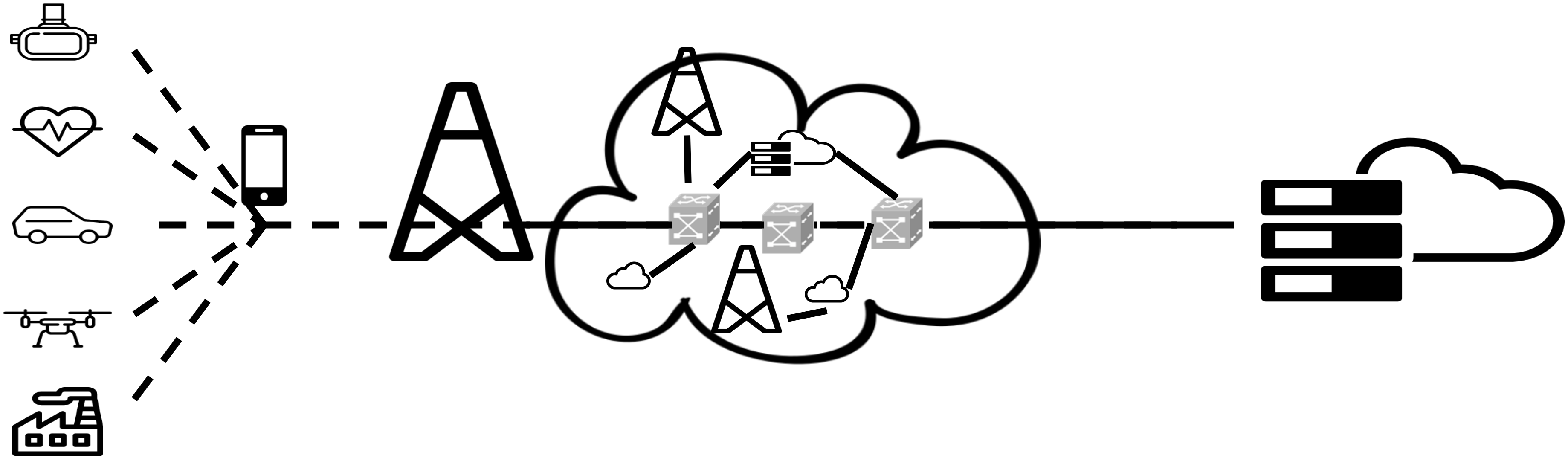
*Columbia University, ^Nokia Bell Labs

Columbia University in the City of New York

Nokia Bell Labs

wim.net Wireless & Mobile Networking Lab

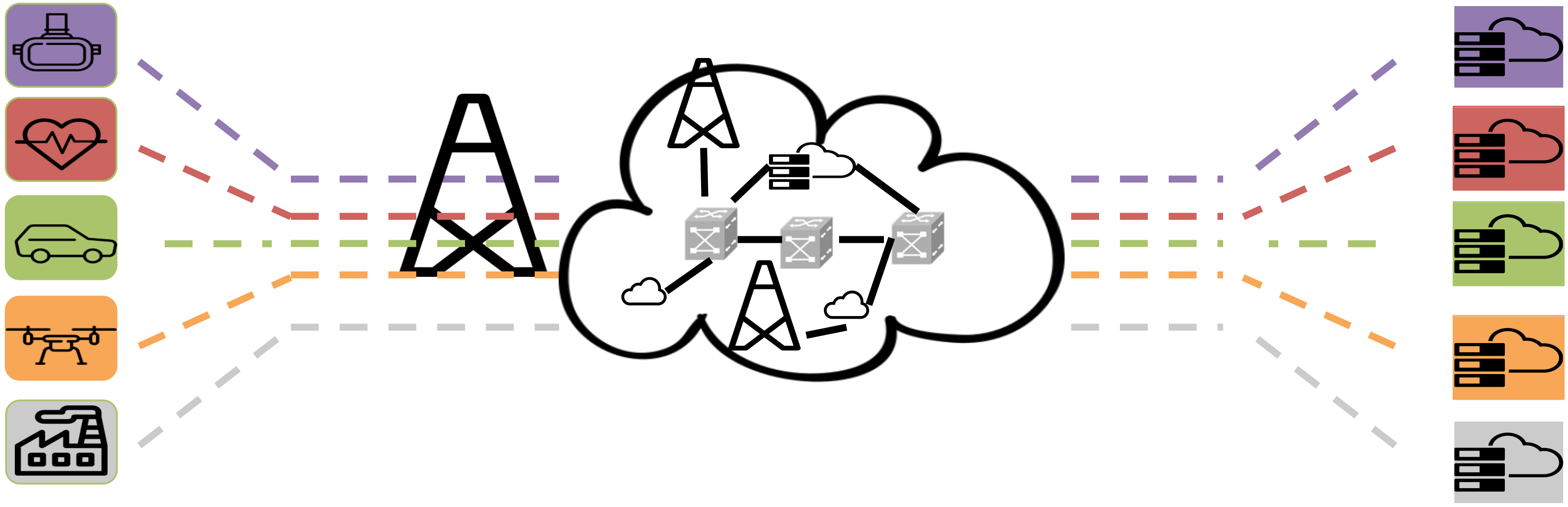
4G Network



Best Effort – All Traffic Created Equal

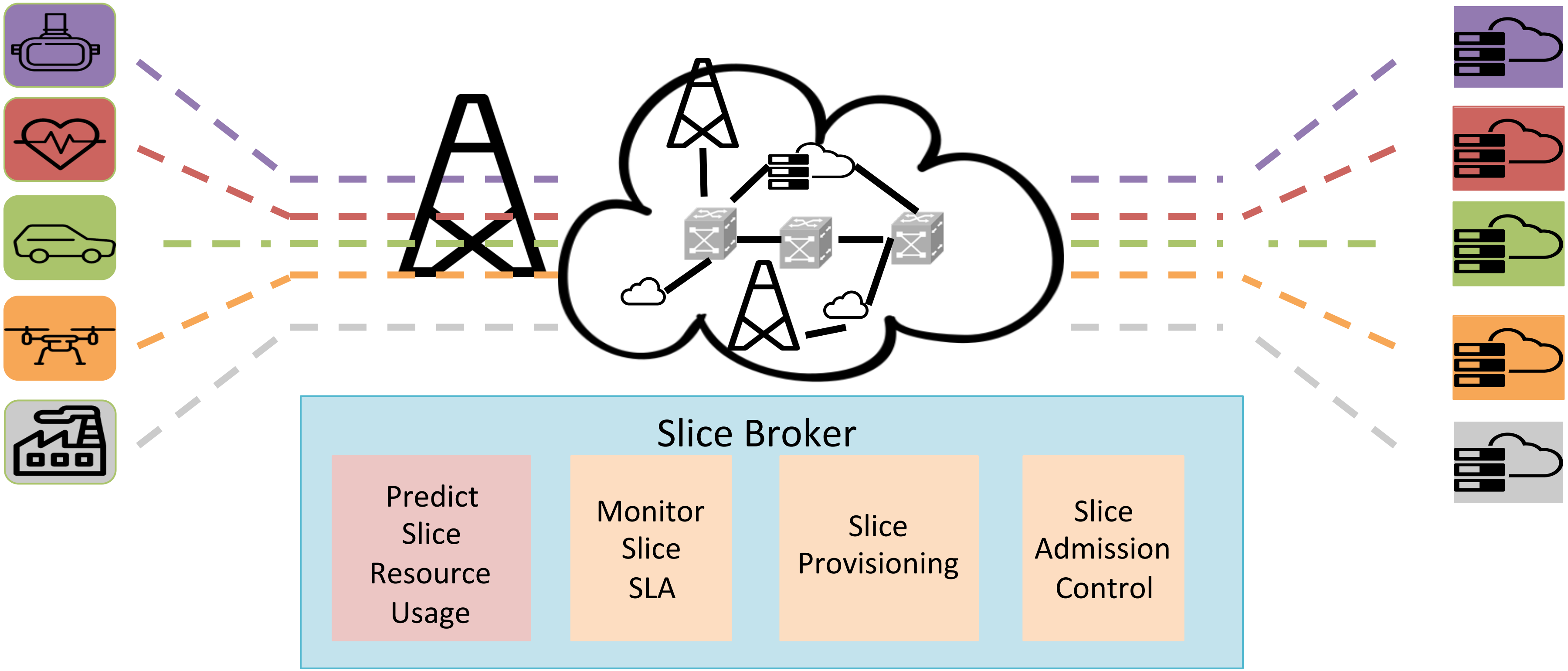
Each application has different service requirements

5G Network

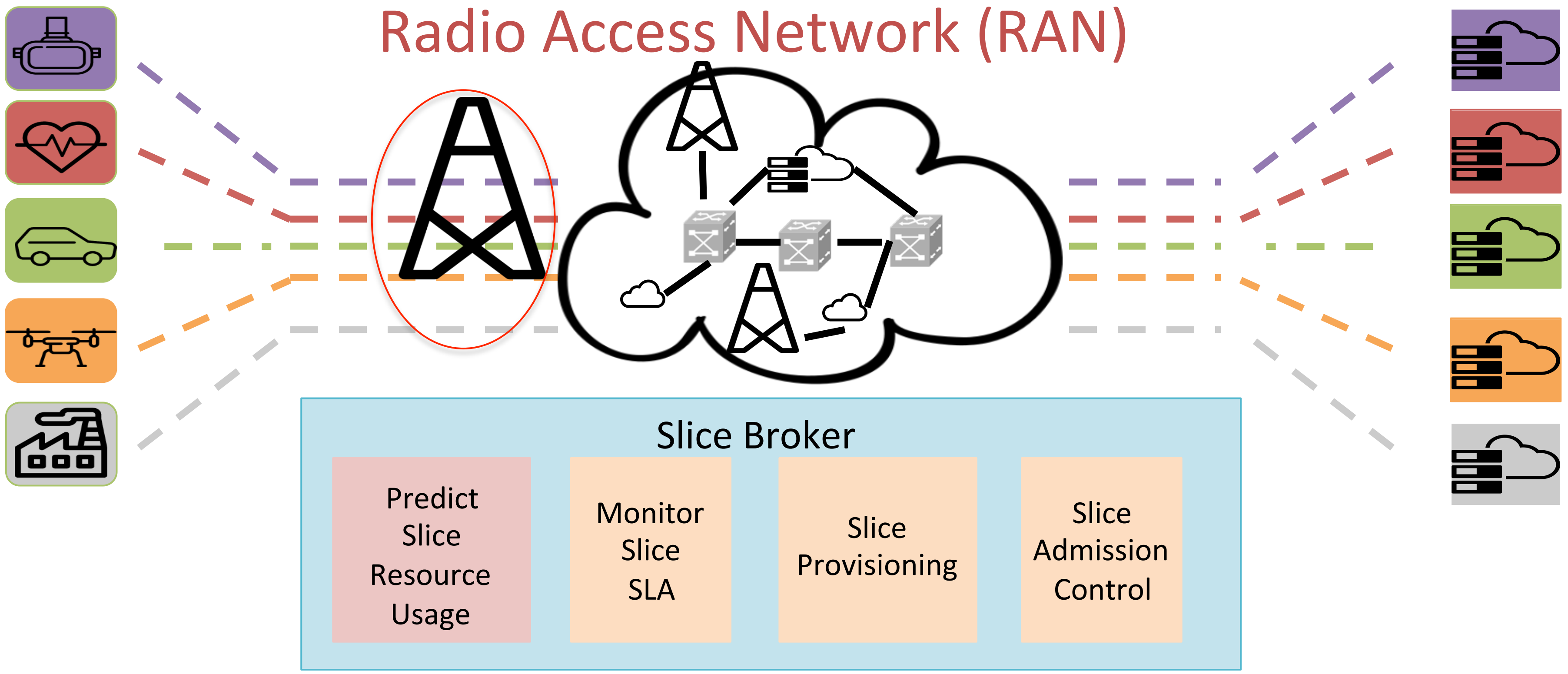


Network Slicing –physical network logically divided to deliver services

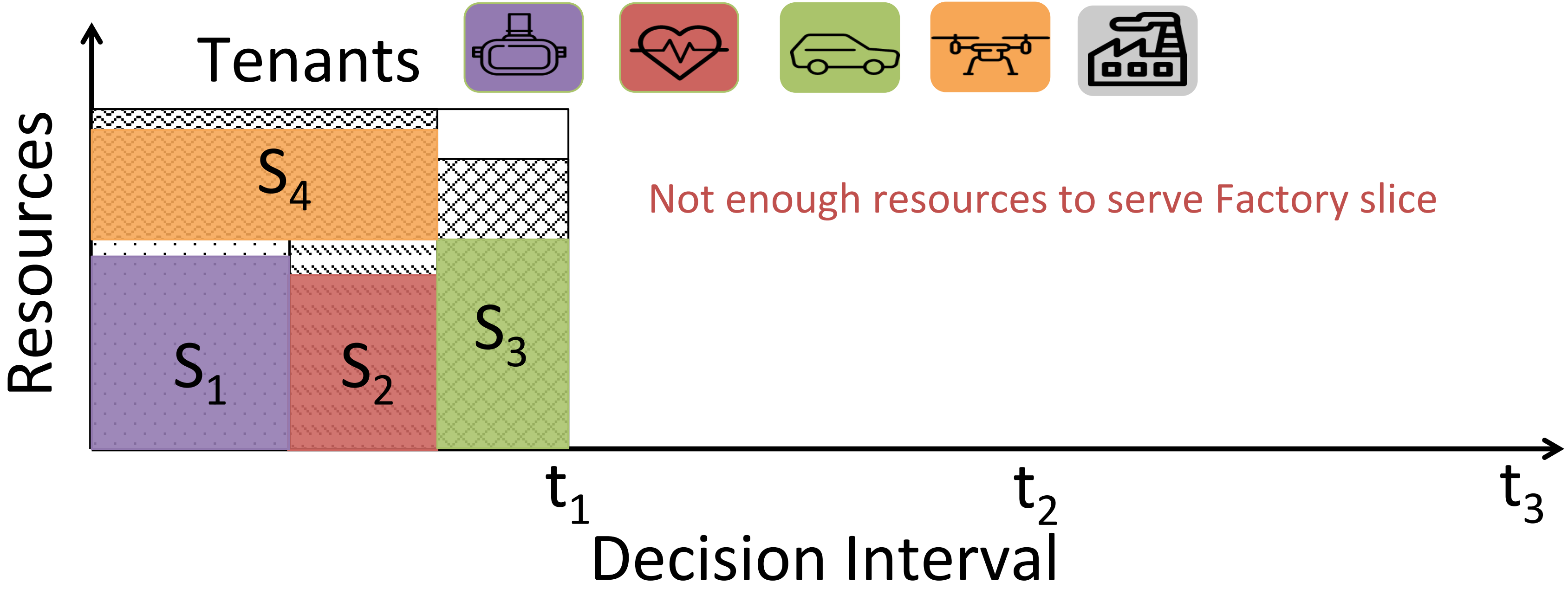
5G Network


Slice Broker
Predict Slice Resource Usage
Monitor Slice SLA
Slice Provisioning
Slice Admission Control

5G Network


Radio Access Network (RAN)
Slice Broker
Predict Slice Resource Usage
Monitor Slice SLA
Slice Provisioning
Slice Admission Control

RAN Broker

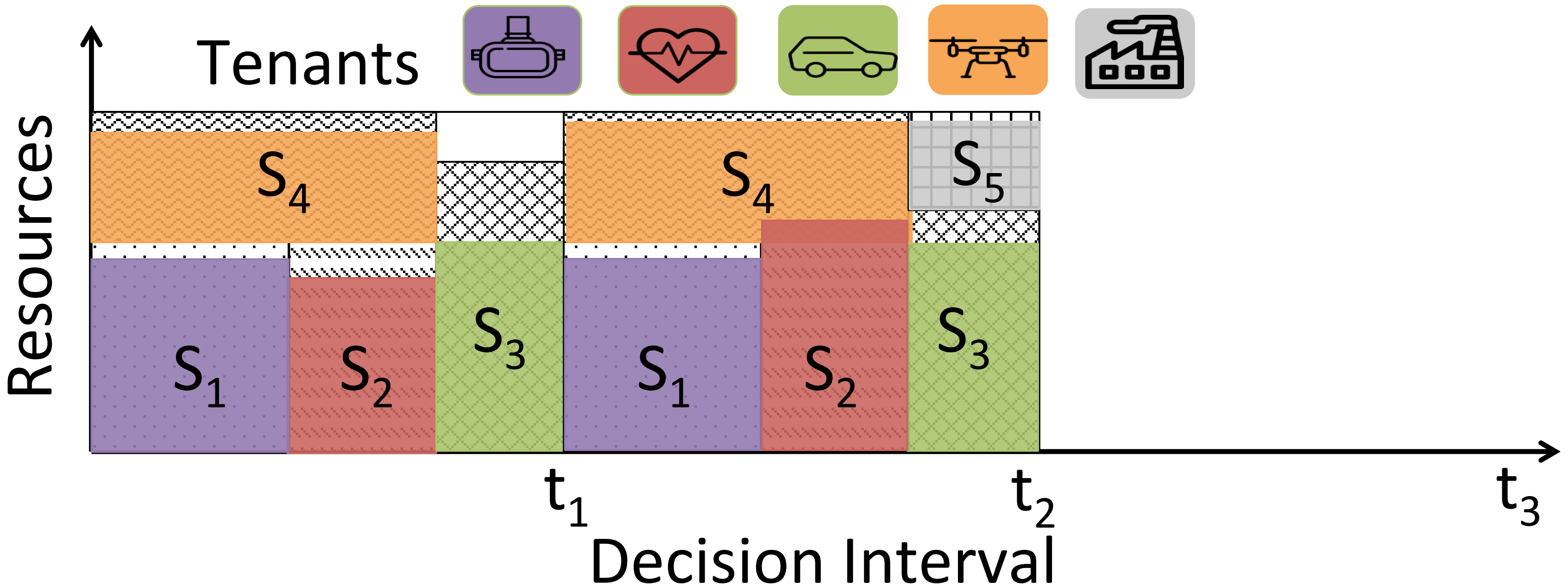


Not enough resources to serve Factory slice

Over Provision

Decreased Revenue

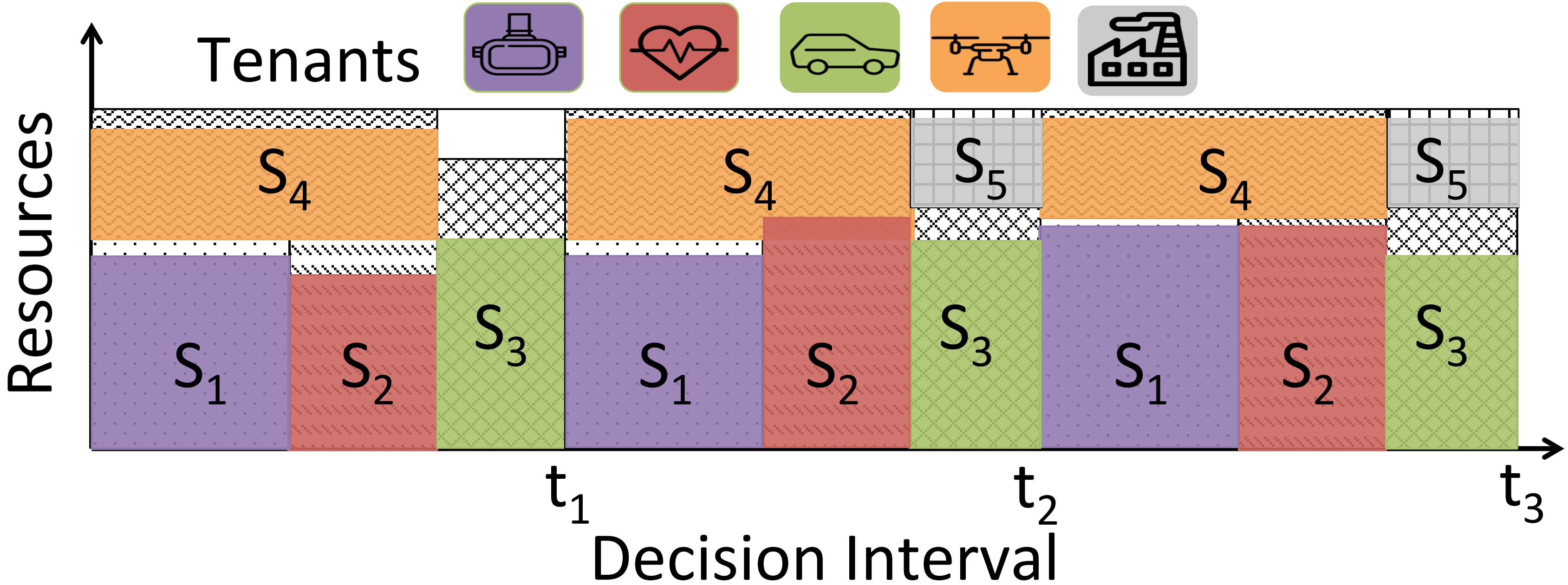
RAN Broker



Under Provision:

Service Level Agreement (SLA) Violation

RAN Broker



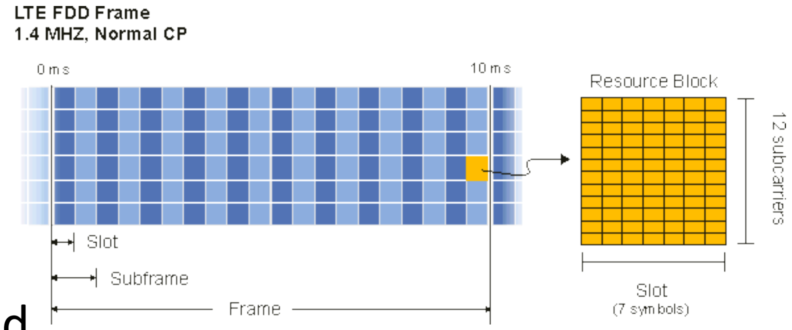
Goal: Accurate Prediction Model

Outline

- Background and Motivation
- **Radio Access Network Resource Utilization**
  - New Metric-REVA
- Prediction model
  - X-LSTM
- Evaluation
- Conclusion and Future Work

Radio Access Network (current 4G terminology)

- Bearer – IP packet flow with a defined QoS between the gateway and User Equipment (UE)
- Resources
  - Bandwidth divided into physical resource blocks (PRBs) of 180 kHz
  - Resource blocks assigned every 1 millisecond



- QoS Class Identifiers (QCI)
  - Guaranteed Bit Rate Traffic (GBR)
    - Voice Over IP
  - Non Guaranteed Bit Rate (non-GBR)
    - Email, ftp, www, streaming applications

QCI	Bearer Type	Priority	Packet Delay	Packet Loss	Example
1	GBR	2	100 ms	10^{-2}	VoIP call
2		4	150 ms	10^{-3}	Video call
3		3	50 ms		Online Gaming (Real Time)
4		5	300 ms	10^{-6}	Video streaming
5	Non-GBR	1	100 ms		IMS Signaling
6		6	300 ms		Video, TCP based services e.g. email, chat, ftp etc
7		7	100 ms	10^{-3}	Voice, Video, Interactive gaming
8		8	300 ms	10^{-6}	Video, TCP based services e.g. email, chat, ftp etc
9		9			

Alternative Radio Access Network Utilization Metrics

Metric	Issue
Aggregate percent of available PRB utilization per second	Single greedy application can utilize close to 100% of the PRBs, but the RAN is not congested
Aggregate throughput of all bearers	Single greedy user with good channel condition can have high throughput
Metrics based upon latency or throughput of individual bearers	Low throughput or high latency may result from poor channel conditions or application usage characteristics
Number of users served by the RAN	Does not take into account RAN resource consumption by individuals

Kwan et al. 2010, Wang et al. 2015, Heder et al. 2016, Zhang et al. 2017, Wang et al. 2017

Objective of new metric (REVA)

- A function of the available resources that is independent of:
  - Channel conditions of the bearers
  - The application behavior and throughput needs of individual user bearers
  - Transport protocol
  - Bearer throughput or round trip time
- The average number of PRBs used by the bearers that attempt to obtain **more than their maximal fair share of PRBs**
- Method for precise and direct computation of available throughput per bearer

$$R(b_i) = \overline{PRB_i} * C(b_i)$$

$\overline{PRB_i}$, is average PRBs for bearer i
$C(b_i)$, is the average nubmer of bits per PRB for bearer i
b_i, bearer channel conditions
$R(b_i)$, wireless throughput avaiable for bearer i

Definitions

- **Active Bearer**: Are bearers for a non-GBR QCI that use on average ϒ PRBs per second
- **Very Active (VA) Bearer:** Are bearers for a non-GBR QCI that continuously attempt to obtain more than a maximal fair share of PRBs that are available from the scheduler for a given duration of time
- **Less Active (VA) Bearer:** Are active bearers for a non-GBR QCI that are not VA
- **δ:** Fraction of control plane PRBs

REVA

- REVA determines the number of PRBs that a Very Active (VA) bearer at a given QCI can obtain
- Algorithm
  - Compute available PRB rate per QCI of the slice
  - For each QCI, classify the slice bearers into Less Active (LA) and VA
  - Iteratively eliminate bearers that use less than their fair share of the remaining resources
  - Continue until
    - No additional LA bearers are added
    - 0 or 1 non-LA bearers remain

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- $\delta = 0.02$
- Each UE has single downlink bearer at QCI 9

PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

Fair Share	2450

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- δ = 0.02
- Each UE has single downlink bearer at QCI 9

Previous Fair Share	2450
PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- $\delta = 0.02$
- Each UE has single downlink bearer at QCI 9

Bearers 11-20 use less than their fair share

↓

Less Active

Previous Fair Share	2450

PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

Fair Share	4307

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- $\delta = 0.02$
- Each UE has single downlink bearer at QCI 9

Previous Fair Share	4307
PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- $\delta = 0.02$
- Each UE has single downlink bearer at QCI 9

Bearers 7-20 use less than their fair share

↓

Less Active

Previous Fair Share	4307

PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

Fair Share	4697

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- δ = 0.02
- Each UE has single downlink bearer at QCI 9

Previous Fair Share	4697

PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- δ = 0.02
- Each UE has single downlink bearer at QCI 9

Bearers 5-20 use less than their fair share

↓

Less Active

Previous Fair Share	4697

PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

Fair Share	4845

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- $\delta = 0.02$
- Each UE has single downlink bearer at QCI 9

Previous Fair Share	4845

PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- δ = 0.02
- Each UE has single downlink bearer at QCI 9

Bearers 3-20 use less than their fair share

Less Active

Previous Fair Share	4845

PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

Fair Share	4940

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- δ = 0.02
- Each UE has single downlink bearer at QCI 9

Previous Fair Share	4940

PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

REVA Example

- 20 UEs served by 10 MHz slice (50000 PRBs/sec)
- $\delta = 0.02$
- Each UE has single downlink bearer at QCI 9

Bearers 2-20 use less than their fair share

Less Active

Bearer 1 uses more than it's fair share

Previous Fair Share	4940

PRBs Used for 20 UEs	
5000	3000
4900	180
4800	180
4700	180
4600	180
4500	180
4400	180
4300	180
4200	180
3000	180

Fair Share	4980

Outline

- Background and Motivation
- Radio Access Network Resource Utilization
  - REVA
- **Prediction model**
  - X-LSTM
- Evaluation
- Conclusion and Future Work

Time Series Forecasting

- Broker has history of T decision intervals of the series

$$\langle y_{t-1} \rangle = (y_{t-1}, y_{t-2}, \dots, y_{t-T})$$

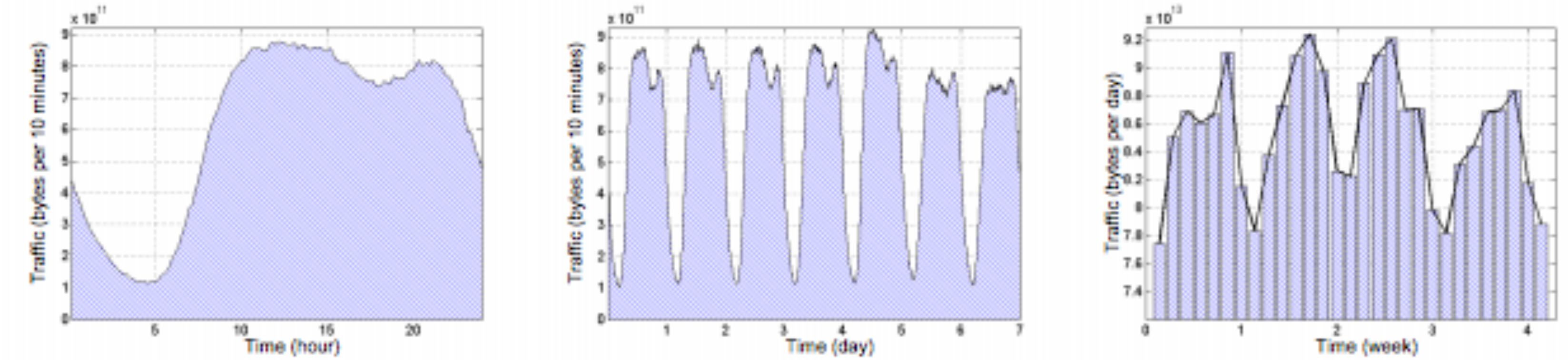
- Objective: Predict tens of seconds using multistep prediction

$$\widehat{y_t}, \widehat{y_{t+1}}, \dots, \widehat{y_{t+s-1}} = f(\langle y_{t-1} \rangle) + \varepsilon_t$$

- Approaches
  - Autoregressive Integrated Moving Average model (ARIMA)
  - Recurrent Neural Networks
    - Long Short-Term Memory (LSTM)

Problem: Do not generalize well for multistep prediction

Temporal Patterns of Cellular Traffic



(a) Hourly (b) Daily (c) Weekly

Can we improve prediction accuracy by making predictions at multiple timescales?

H. Wang, F. Xu, Y. Li, P Zhang, and D Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," in ***Proc. ACM IMC'15***.

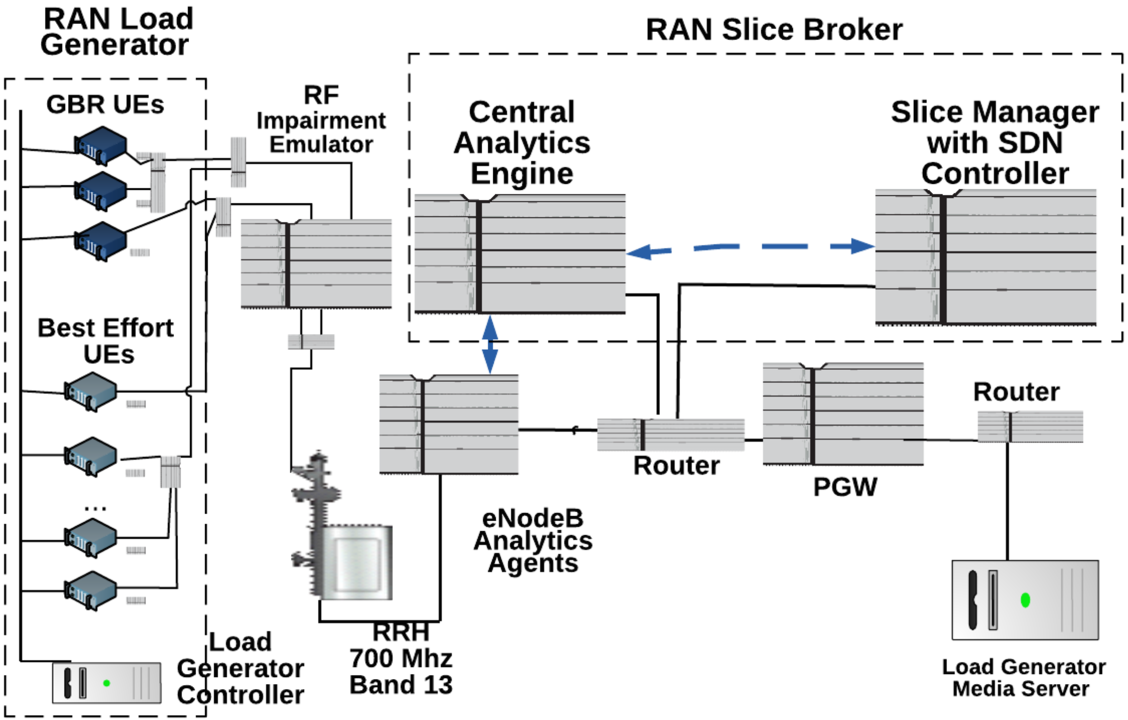
X-LSTM

- Based on the combination of LSTM and the X-11 statistical method
- Uses multiple LSTMs, each with a different time scale
- Filter out higher temporal patterns and use the residual to make additional predictions on data with a shorter time scale

Experimental Data Acquisition

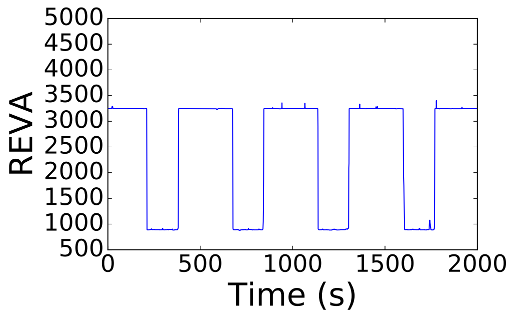
- No publically available data set with PRB distribution per bearer with <1 second granularity from deployed basestation (eNodeBs)

- Designed a lab LTE network with synthetic loads

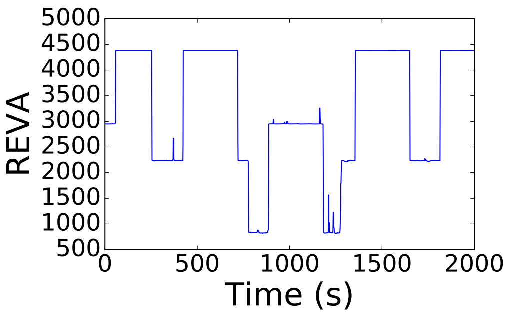


Data

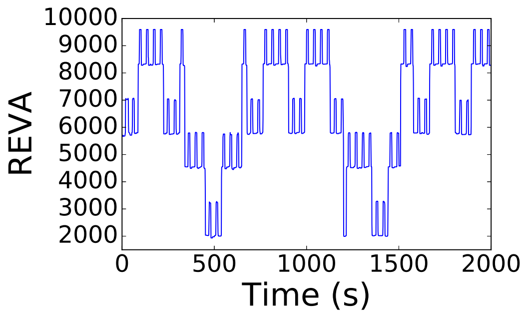
- Created traffic load using 15 UE's configured for QCI 9 (non-GBR) and 3 UE's configured for QCI 3 (GBR)
- REVA calculated at the eNodeB scheduler every 1 second
- Each experiment collected for ~18 hours



Set 1
1 periodic GBR client



Set 2
2 periodic GBR client

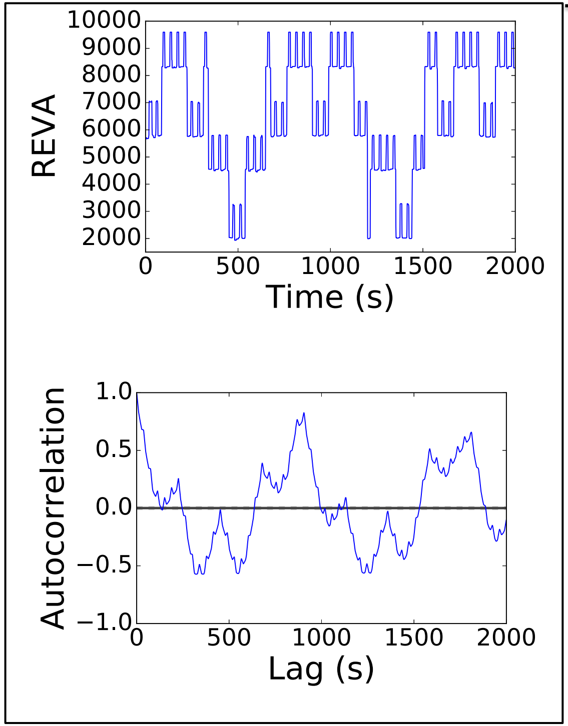


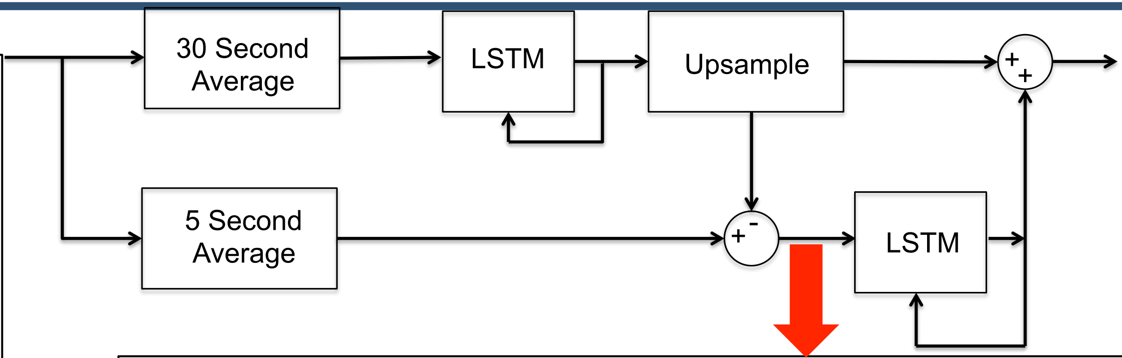
Set 3
3 periodic GBR client

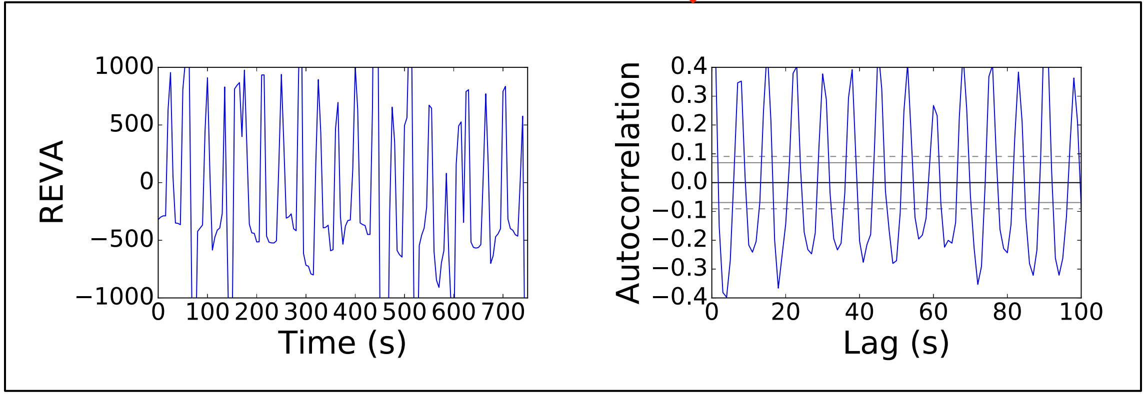
Methods for Comparison

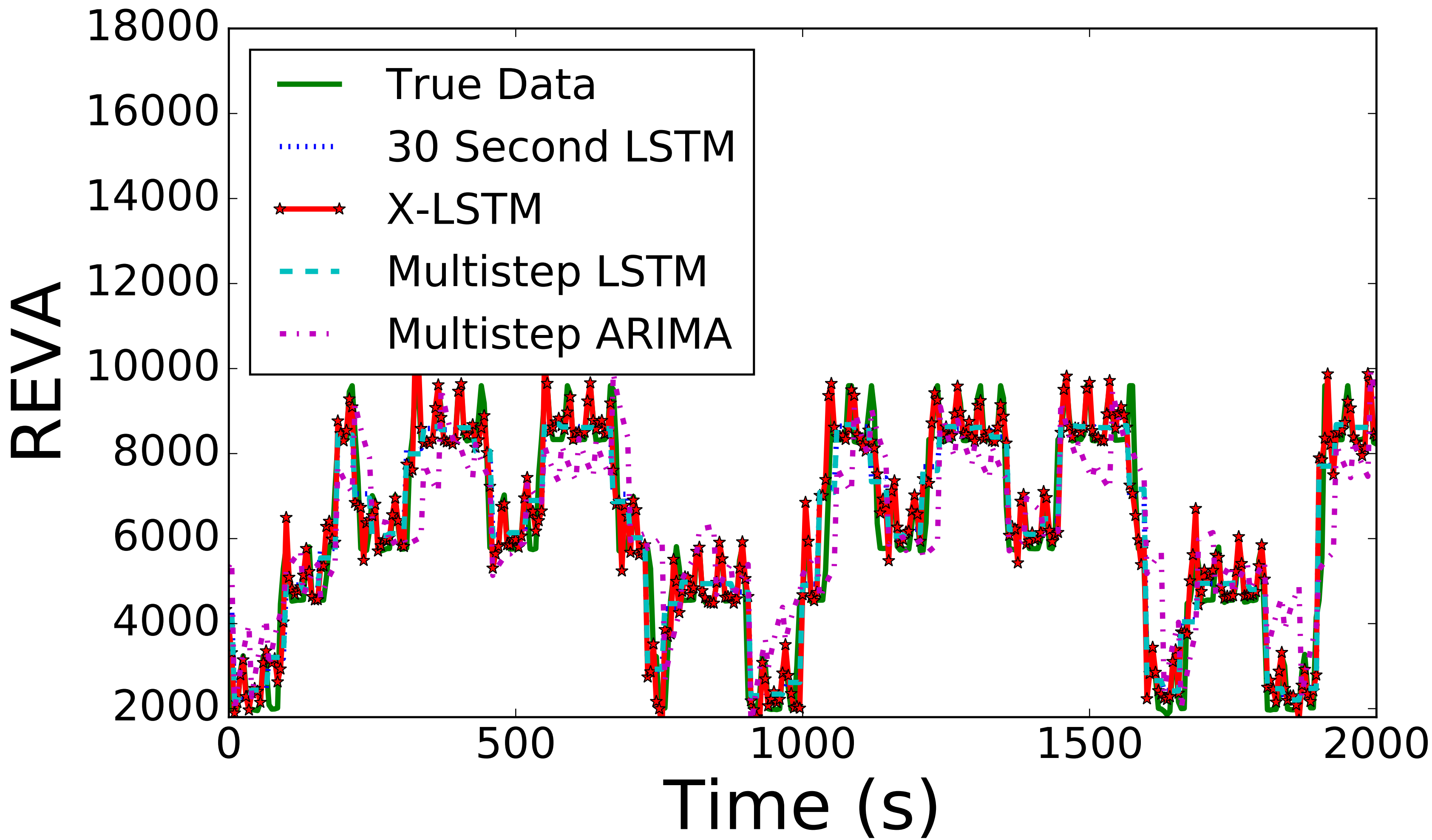
- Multistep ARIMA
  - Make 6 predictions at a time with a granularity of 5 second averages
- Predict 30 second averages using LSTM
- Multistep LSTM
  - Make 6 predictions at a time with a granularity of 5 second averages
- X-LSTM
  - Make 6 predictions at a time with a granularity of 5 second averages

X-LSTM Example





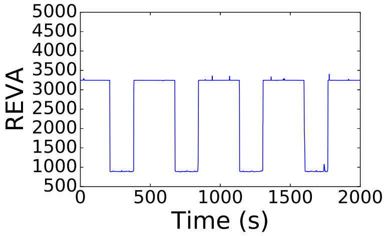




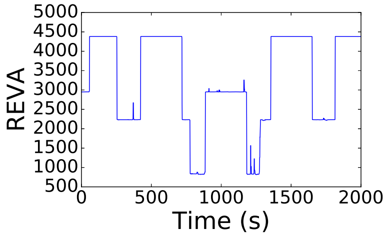
REVA
Time (s)
True Data
30 Second LSTM
X-LSTM
Multistep LSTM
Multistep ARIMA
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
500
1000
1500
2000

Evaluation

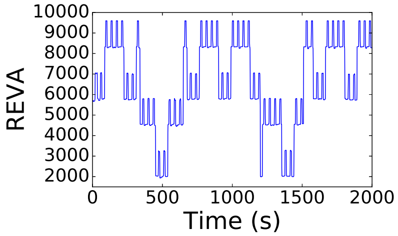
Set 1

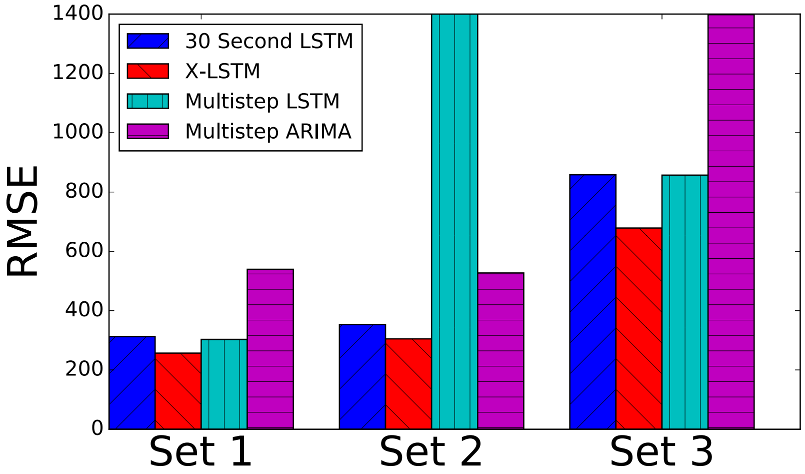


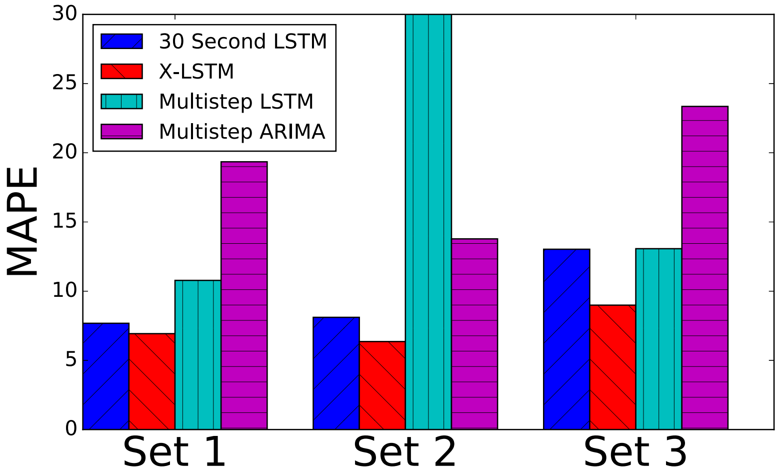
Set 2



Set 3







> 10% improvement

How does prediction accuracy relate to cost?

- Assume

$$y_t = \mathcal{N}(\hat{y_t}, \sigma^2)$$

- SLA violation has cost *k*
- One sided prediction bound *h*
- Cost function

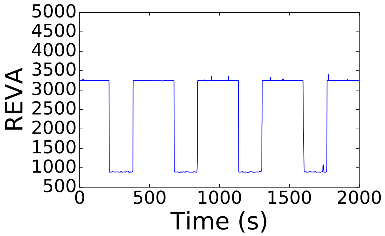
$$\Gamma(y_t) = \begin{cases} k, & if\ \hat{y_t} + h > y_t \\ y_t - h - \hat{y_t}, & if\ \hat{y_t} + h \leq y_t \end{cases}$$

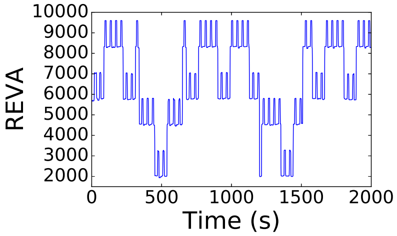
- Optimization problem

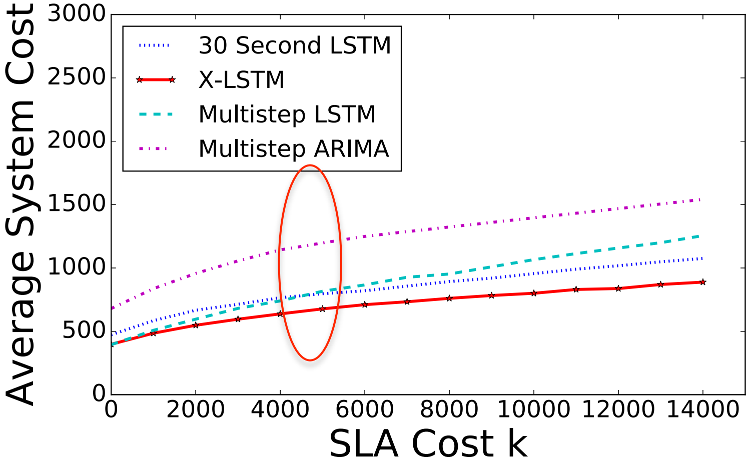
$$\underset{h}{minimize} \quad k(\hat{y_t} + h - y_t)^+ + (y_t - h - \hat{y_t})(y_t - h - \hat{y_t})^+$$

$$\underset{h}{minimize} \quad (k + \hat{y_t})\left(\Phi\left(\frac{h}{\sigma}\right)\right) - h\left(1 - \Phi\left(\frac{h}{\sigma}\right)\right) + \sigma\left(\frac{\phi(\frac{h}{\sigma})}{1 - \Phi\left(\frac{h}{\sigma}\right)}\right)$$

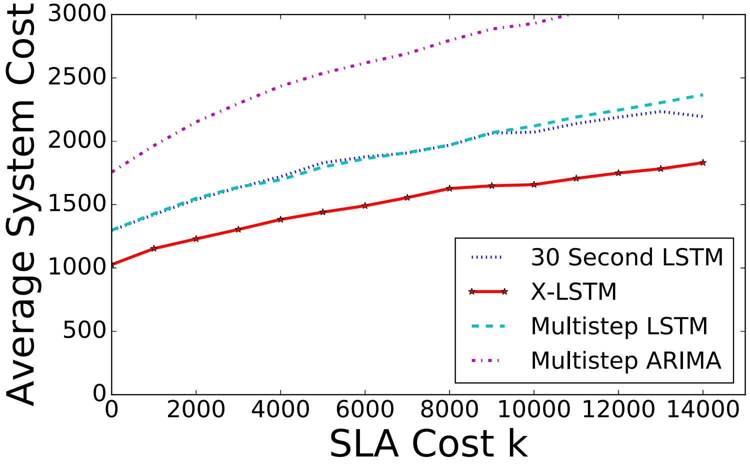
Average System Cost







Set 1



Set 3

>15% Reduction

>18% Reduction

Summary

- Define new metric, REVA, that precisely measure the amount of PRBs that the RAN scheduler can allocate to Very Active bearers
- X-LSTM provides a higher degree of prediction accuracy
  - X-LSTM provides more than 10% cost reduction per slice
- Future Work
  - Evaluate on real world races
  - Develop slice admission control algorithms for the broker

Thank You!

RAN Resource Usage Prediction for a 5G Slice Broker

Craig Gutterman*, Edward Grinshpun^, Sameer Sharma^, Gil Zussman*

*Columbia University, ^Nokia Bell Labs

clg2168@columbia.edu

Columbia University in the City of New York

NOKIA Bell Labs

wim.net Wireless & Mobile Networking Lab