Small cell-assisted Group Paging for cellular mMTC

Bui Hoang Anh Tuan

Computer Communications Lab., University of Aizu

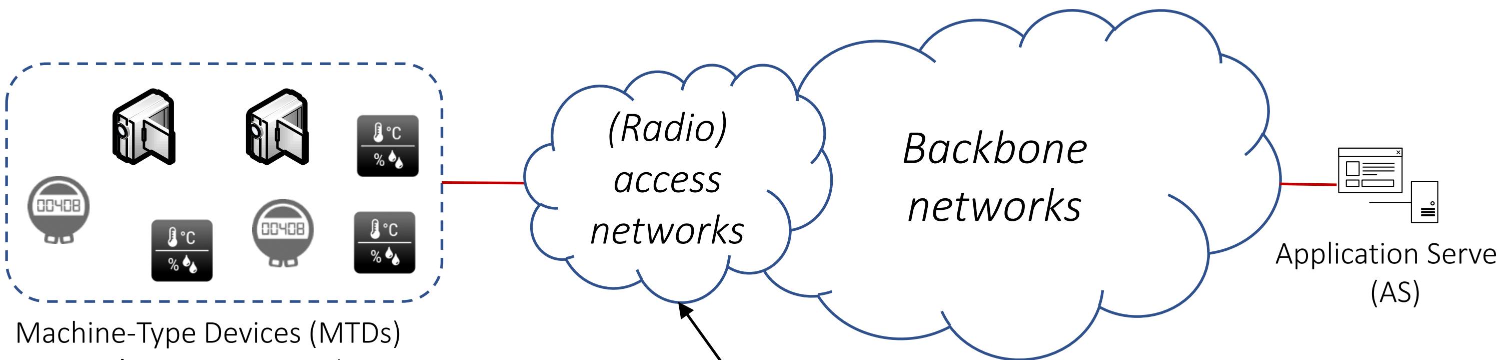
June 20th, 2019

Outline

I. Cellular massive MTC & Cellular Radio Access Network Overload
II. RAN Congestion Control Schemes
III. Small cell-assisted Group Paging
IV. Theoretical Delay Model
V. Simulation Results
VI. Conclusion

I. Cellular massive MTC

- What is massive MTC (mMTC)?



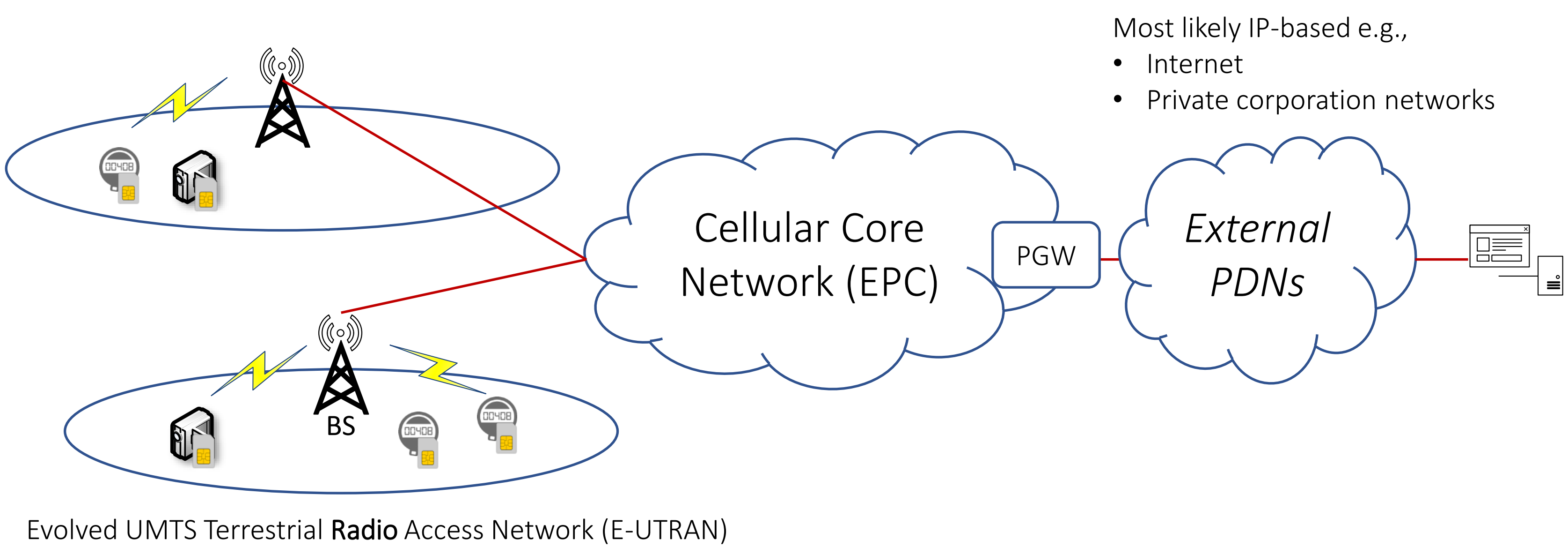
Machine-Type Devices (MTDs)
- **Low data rate:** 1-100 Kbps
- **High density:** up to 1M/km^2
- Acceptable latency: seconds to hours
- Low power: up to 15 years of battery life

Many available technologies
- Cellular (GSM, WCDMA, CDMA2000, LTE, NR)
- WLAN (WiFi, …)
- LPWAN (NB-IoT, Sigfox, LoRaWan …), etc.

- Autonomous information exchange between a massive number of low-rate MTDs and AS
- An official 5G use case

I. Cellular massive MTC

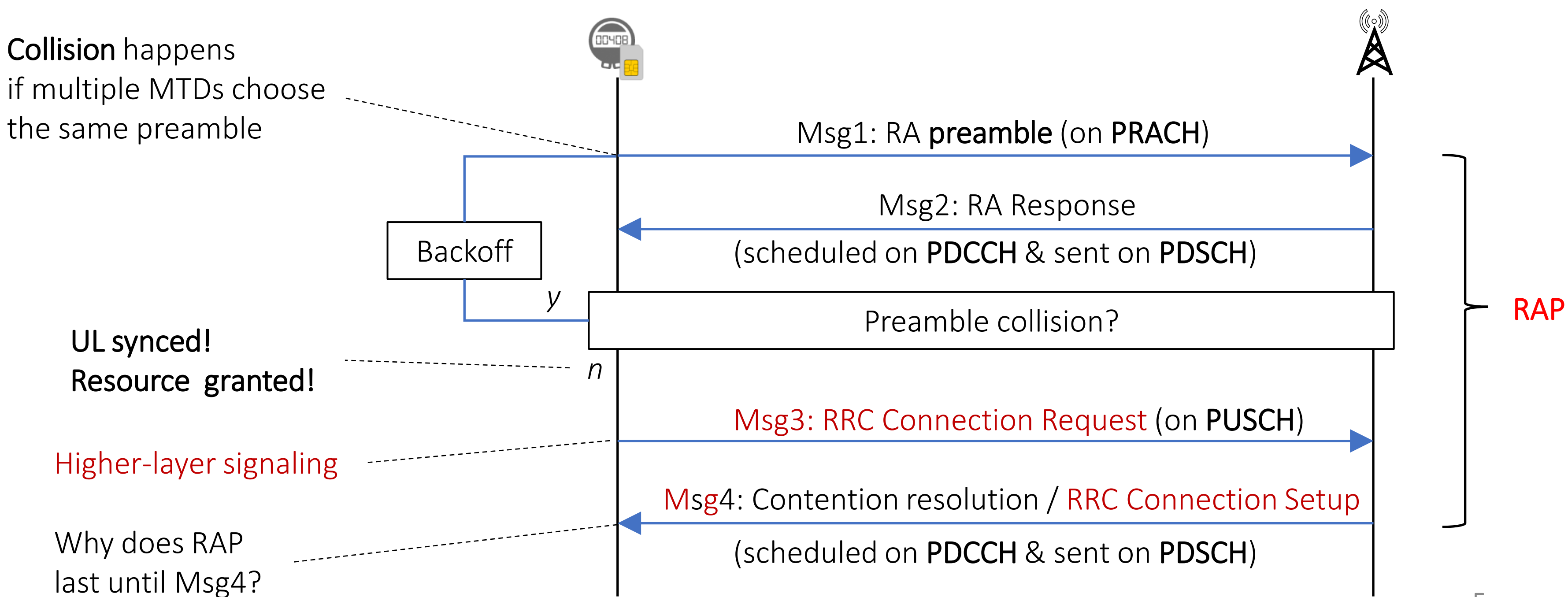
- Why go cellular e.g., LTE for MTC?
  - ➢ Wide coverage → supports MTDs' ubiquity
  - ➢ Matured & well-adopted → easy massive installation



*LTE is actually the name of the 3GPP work item concerning development of the radio access technology and E-UTRAN

I. Cellular RAN Overload

- Is LTE suitable for mMTC?
  - ➢ If an MTD wants to access, it must undergo Random Access Procedure (RAP)
  - ➢ RAP has two purposes: UL synchronization & to request radio resource for higher-layer signaling



*Hint: what if PHY phenomena cause a Msg1 collision to be undetected?

I. Cellular RAN Overload

- Identifying the bottlenecks
  - Limited number of preambles & backoff-based contention resolution → frequent preamble collisions in massive access
  - PDCCH is used for scheduling of almost everything → Msg2 or 4 may not be scheduled for successful MTDs during PDCCH resource shortage
- Consequence? MTDs quit i.e., “blocked” after consecutive failures

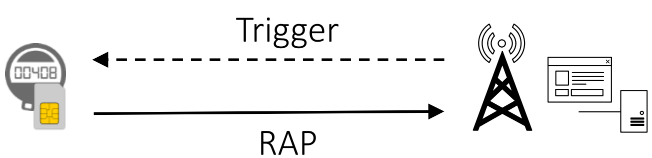
→ LTE needs enhancement to support mMTC

II. RAN Overload Control Schemes

- RAN overload in cellular mMTC is well-known, and various solutions exist
- They can be classified into push-based and pull-based schemes

Push-based (device originated)
MTD initiates RAP on its own will
e.g., upon event detections

Our focus

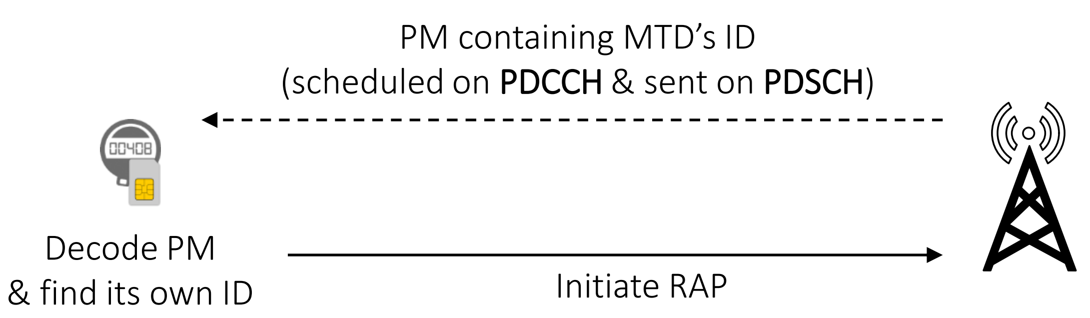


Pull-based (device terminated)
NW triggers the MTD to initiate RAP
e.g., when report request is received from AS

Provide NW with more control over access traffic

II. RAN Overload Control Schemes

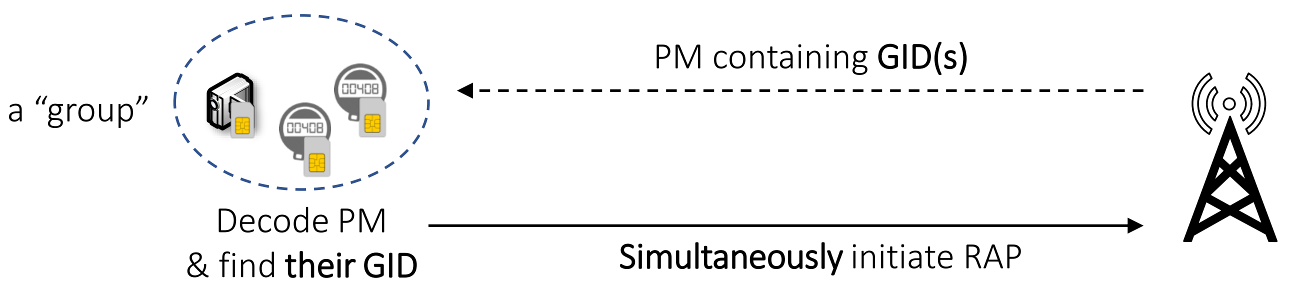
- Paging and Group Paging (GP) are two main approaches of pull-based solutions
- Paging:
  - ➢ BS calls for an MTD by sending a paging message (PM) containing the MTD's ID
  - ➢ MTD, upon receiving a PM with its ID, initiates RAP



II. RAN Overload Control Schemes

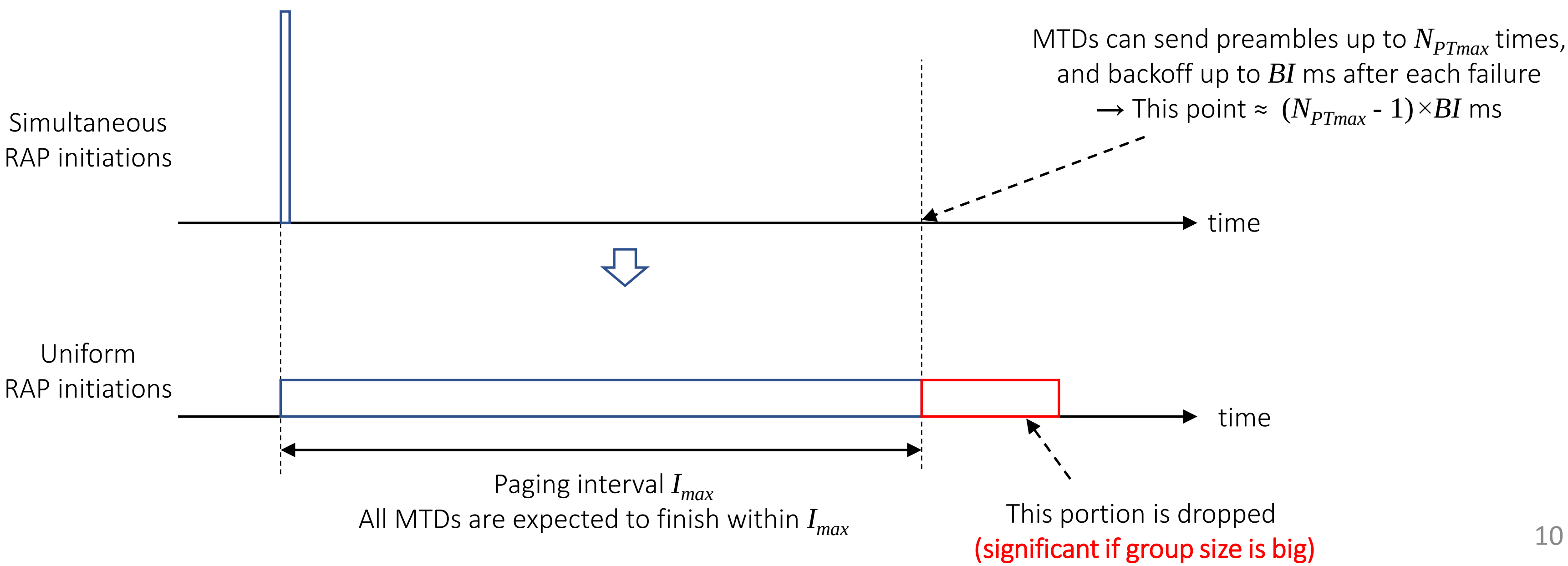
- Paging's limitations:
  - ➢Up to 4 PMs per 10ms, each carries up to 16 IDs

    → Paging all MTDs takes a long time
- Group paging (GP) is proposed to overcome this
  - ➢MTDs are divided into groups identified by Group IDs
  - ➢BS pages the MTDs **on a group basis** (using GIDs)



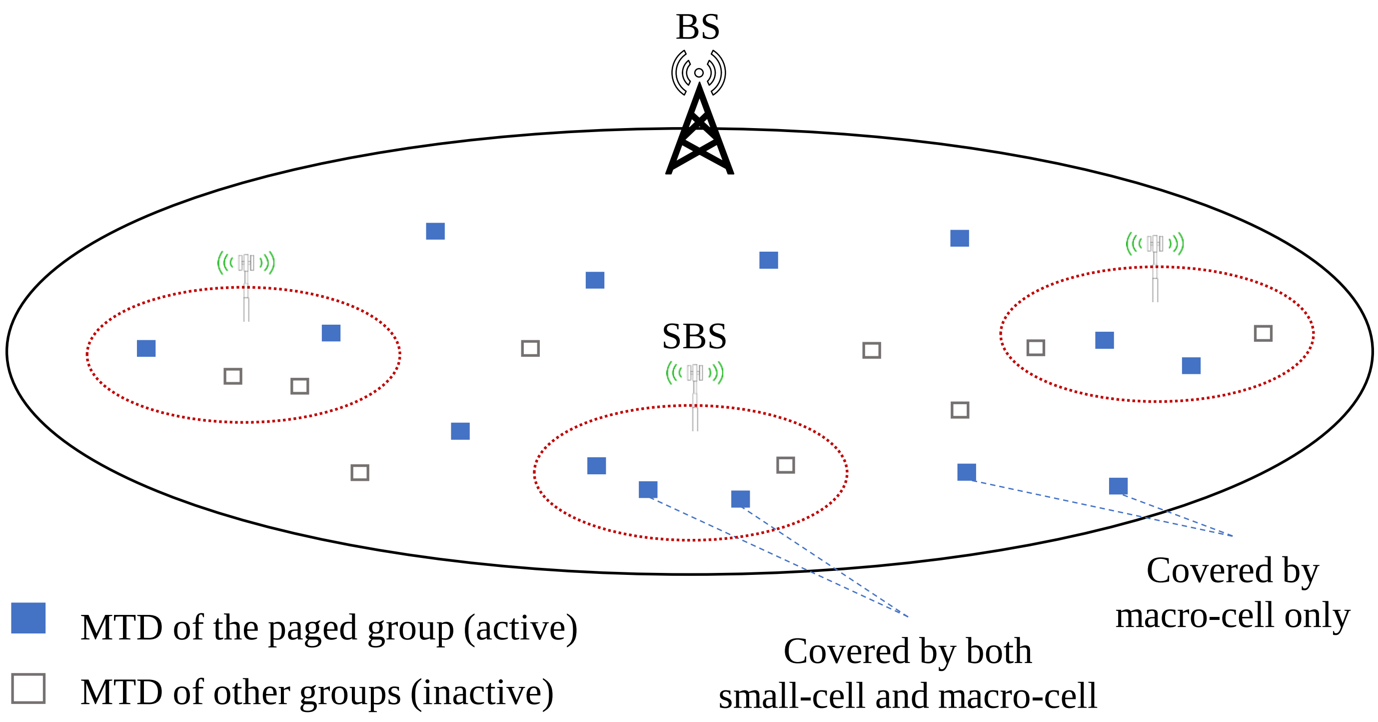
II. RAN Overload Control Schemes

- GP's limitation
  - ➢MTDs of paged group simultaneously initiate RAP → RAN overload issue easily returns
- Most current studies try to overcome this by pre-spreading MTDs over the "paging interval" [ref]



III. Small cell-assisted GP

- In the future, small-cells (SCs) will be densely deployed and cover a large portion of MTDs



*SBS = Small-cell BS

III. Small cell-assisted GP

- An SBS can act as a “representative” for multiple MTDs in its vicinity to request for resource [1]
- Once the SBS obtained resources for Msg3, its MTDs compete over those resource

→ Move part of access load on PRACH & PDCCH (two bottlenecks) to PUSCH

→ Small cell-assisted GP

PUSCH (Physical UL Shared Channel) is schedulable by macro BS and is used for both higher-layer signaling and user data transmissions

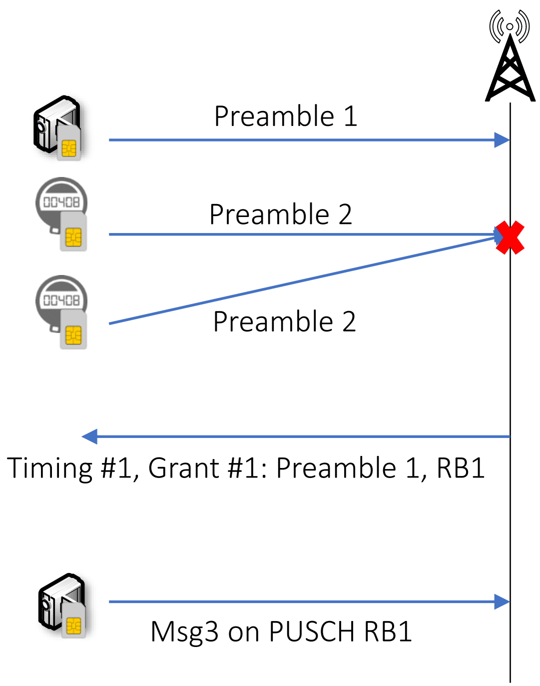
III. Small cell-assisted GP

- How to realize the proposal?
  1. SC-MTDs do not send preambles. The SBS is in charge of that
  2. BS sends a grant allocating N_b resource blocks (instead of 1) if it finds a preamble sent by the SBS
  3. Each SC-MTD decode the grant to get locations of the RBs and randomly select one to send its Msg3

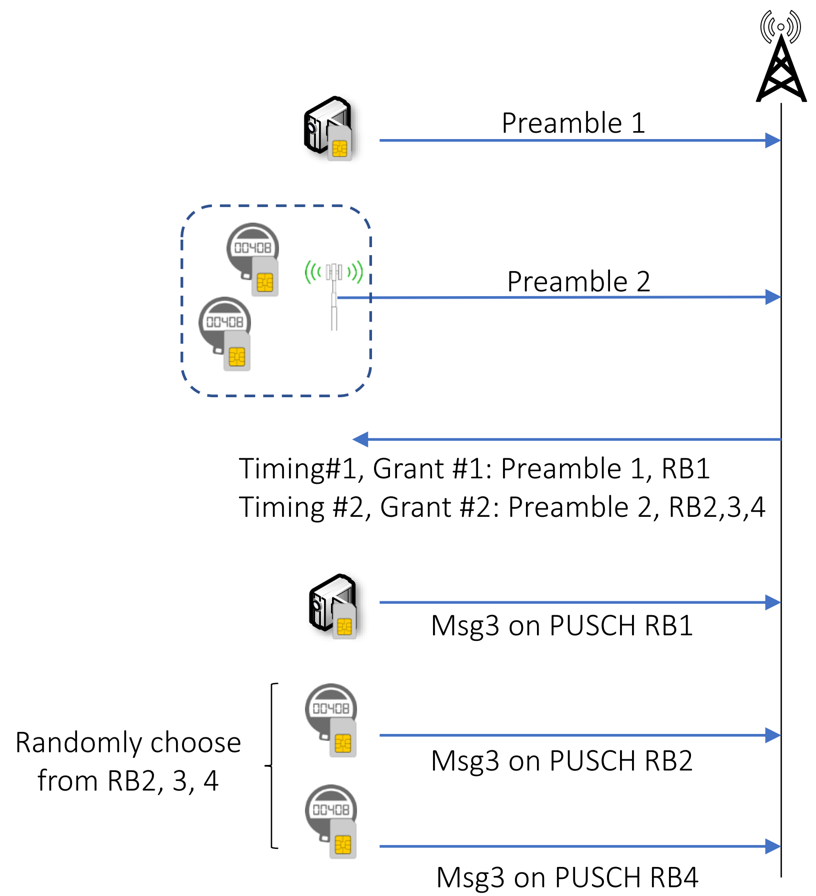
PUSCH (Physical UL Shared Channel) is schedulable by macro BS and is used for both higher-layer signaling and user data transmissions

III. Small cell-assisted GP

- Comparison to conventional RAP



Conventional RAP



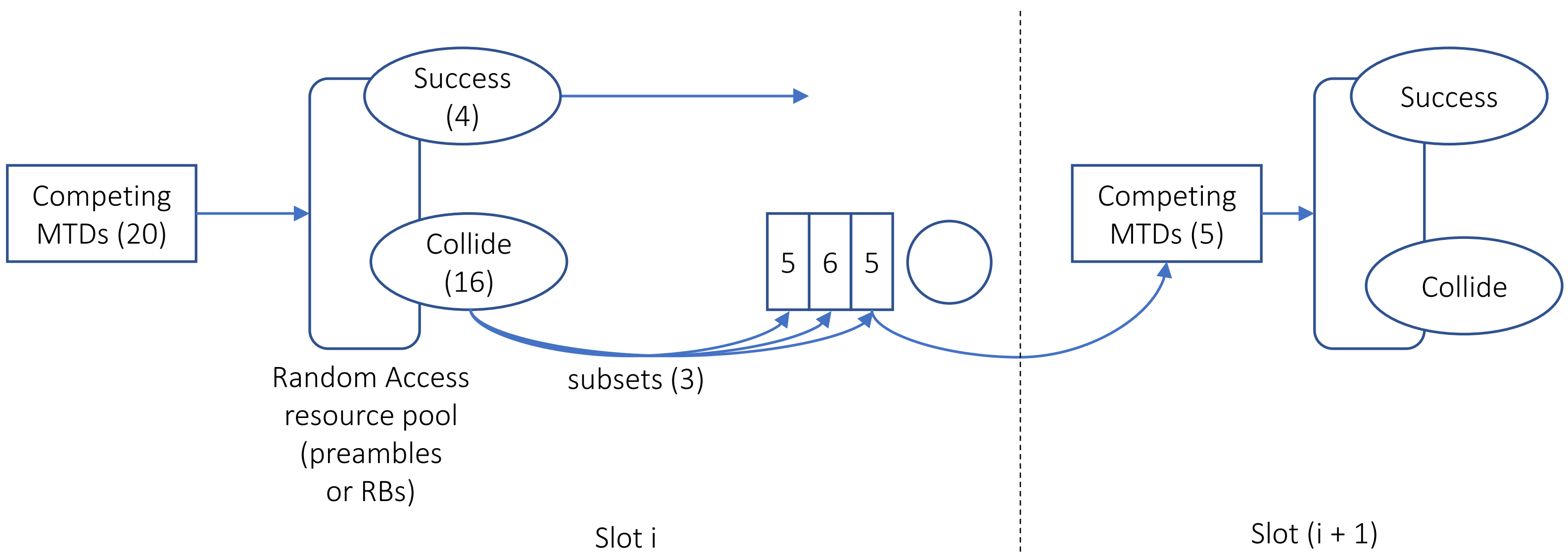
Proposal (assuming N_b = 3)

III. Small cell-assisted GP

- Remaining questions:
  - ➢How to (efficiently) resolve contention during Msg1 & Msg3 transmissions?
  - ➢How does an SBS know that there are remaining MTDs (so as to continue asking BS for resources)
- To answer both questions, we use a Distributed Queue (DQ)-based contention resolution protocol

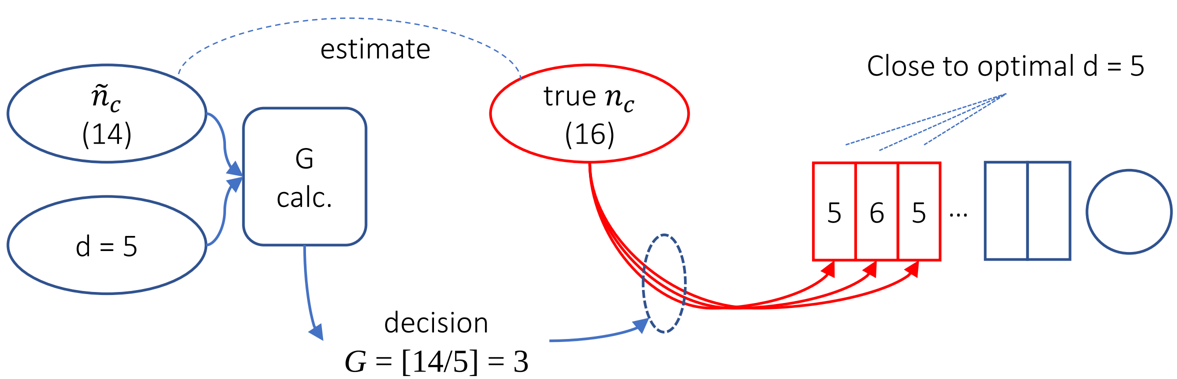
III. Small cell-assisted GP

- DQ uses a “logical queue” to resolve contentions between competing devices
  - ➢Colliding MTDs are divided into subsets and pushed to the end of a “queue”
  - ➢In each slot, only the head subset **exits** & **retransmits**



III. Small cell-assisted GP

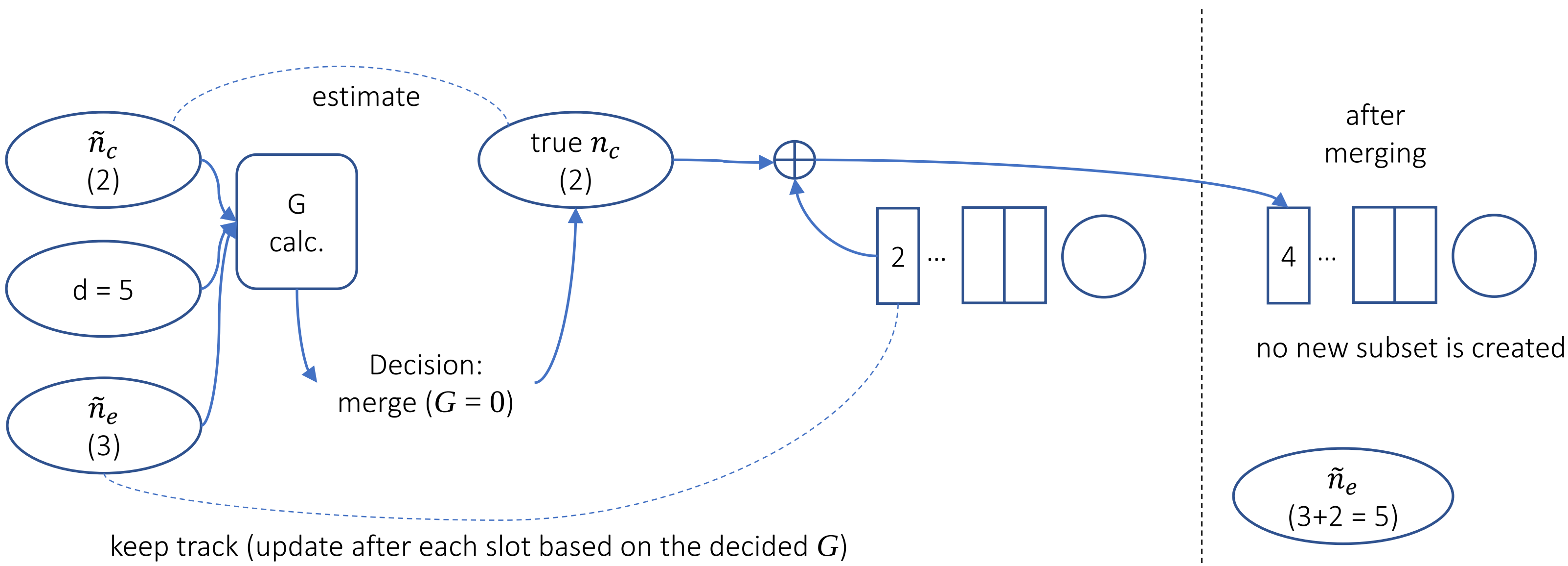
- Choosing the right number of subsets G is vital to DQ's performance
- In our previous work, G is based on 1) optimal subset size d, and 2) estimate $\tilde{n}_c$of the number of colliding MTDs
  - If $\tilde{n}_c > d$, then $G = [\tilde{n}_c/d]$
  - If $\tilde{n}_c < d$, then $G = 1$ (no further division)



The estimate nc is obtained using a MAC-layer technique using observed statuses of the resources

III. Small cell-assisted GP

- We newly notice that when the subsets' size is too low, it is better to merge them together
  - ➢The BS monitors the (estimated) tail subset's size $\tilde{n}_e$
  - ➢If $\tilde{n}_c + \tilde{n}_e \le d$, current colliding MTDs will be merged with the end subset



III. Small cell-assisted GP

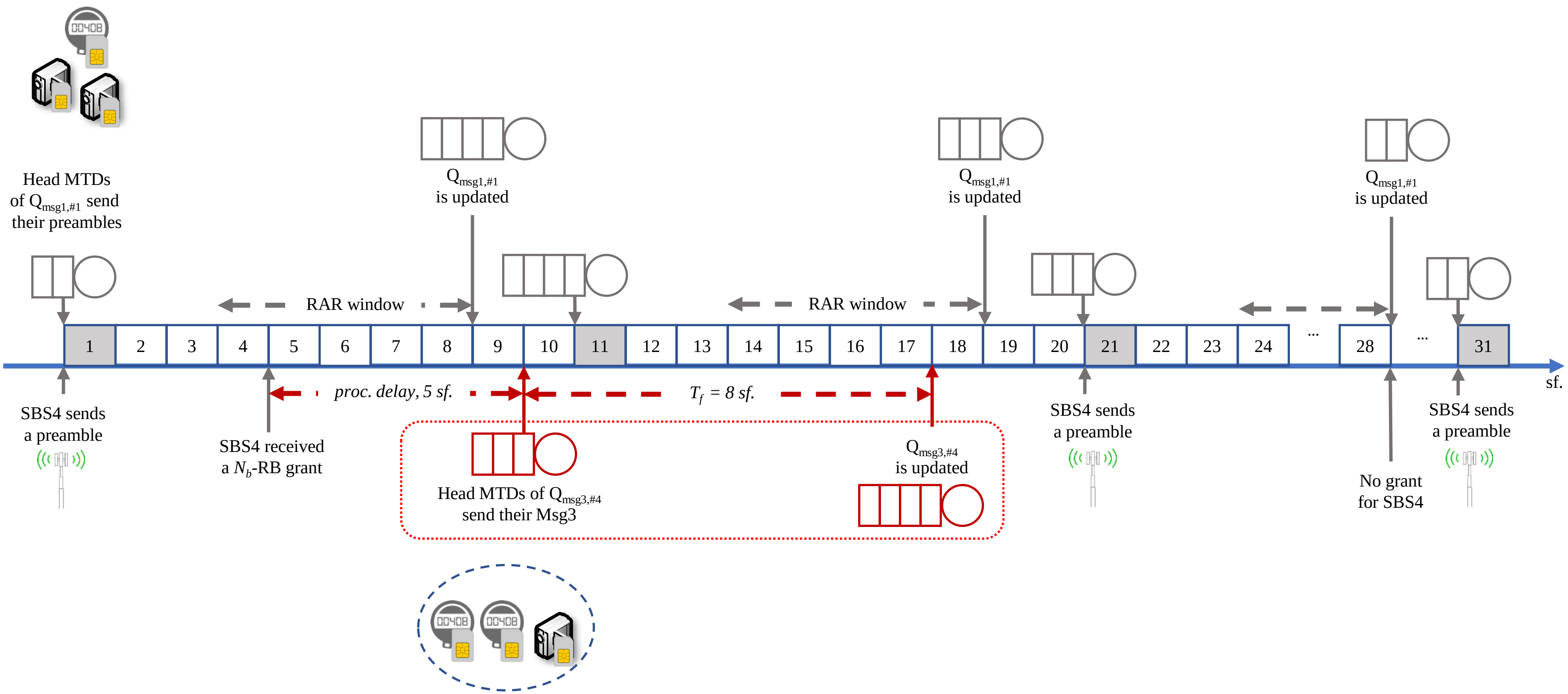
- Such DQ-based protocol is used to resolve contentions during both Msg1 and Msg3
- But there is a key difference between two DQ-based processes
  - ➢Msg1 contention is between macro-only MTDs and SBSs (different contender types)
  - ➢Msg3 contention is between local SC-MTDs of an SBS (same contender type)

III. Small cell-assisted GP

- We need to define how SBSs are treated during Msg1 DQ process
- Two options
  1. SBSs are treated equally as macro-only MTDs
  2. SBSs are prioritized over macro-only MTDs
- Option 2 slightly increases Msg1 contention rate but significantly reduces delay of SC-MTDs
  - ➢ We choose to let SBSs stay permanently at the head subset

III. Small cell-assisted GP

- Small cell-assisted GP: example



Note: Grey squares = “slots” for Msg1 DQ

IV. Theoretical Delay Model

- There are two main tasks
  1. Model the DQ-based contention resolution process in general
  2. Model the interaction between Msg3 DQ process and Msg1 DQ process

IV. Theoretical Delay Model (1)

- Let $\mathcal{N}_i[n]$ and $\mathcal{L}_i$ ~ number of devices transmitting for the n-th time and queue's length in i-th slot
- Define the (random) state vector of the system at i-th slot as

$$\overrightarrow{\mathcal{N}_i} = \langle \mathcal{N}_i[1], \mathcal{N}_i[2], \dots, \mathcal{N}_i[n_{PTmax}] \rangle$$

  and the (given) correspondent as

$$\overrightarrow{N_i} = \langle N_i[1], N_i[2], \dots, N_i[n_{PTmax}] \rangle$$

- The system can be described by the stochastic processes $\{\overrightarrow{\mathcal{N}_i}\}$ and $\{\mathcal{L}_i\}$

*Note: Calligraphy letters e.g., $\mathcal{N}_i[1]$, correspond to random quantities while normal ones e.g., $N_i[1]$, correspond to fixed (deterministic) quantities

IV. Theoretical Delay Model (1)

- Let us see how the processes evolves over time
- Denote by $\mathcal{N}_{i,S}[n]$, $\mathcal{N}_{i,C}[n]$ the number of MTDs who succeed and collide at their n-th attempt in i-th slot
- We then have a system of evolution equations

$$
\begin{cases}
\mathcal{N}_{i+\mathcal{L}_i+g}[1] = \mathcal{N}_{i+\mathcal{L}_i+g,Arrival} \\
\mathcal{N}_{i+\mathcal{L}_i+g}[2] = bino\left(\mathcal{N}_{i,C}[1], 1/\mathcal{G}_i\right) \\
\cdots \\
\mathcal{N}_{i+\mathcal{L}_i+g}[n_{PTmax}] = bino\left(\mathcal{N}_{i,C}[n_{PTmax} - 1], 1/\mathcal{G}_i\right) \\
\mathcal{L}_{i+1} = \mathcal{L}_i - 1 + \mathcal{G}_i
\end{cases}
$$

*Note: $\mathcal{G}_i$ is the number of groups ($0 \le g \le \mathcal{G}_i$) in i-th slot

IV. Theoretical Delay Model (1)

- In principle, $\mathbb{P}\left(\overrightarrow{\mathcal{N}_{i+\Delta}} = \overrightarrow{N_{i+\Delta}}, \mathcal{L}_{i+1} = L_{i+1} | \overrightarrow{\mathcal{N}_i} = \overrightarrow{N_i}, \mathcal{L}_i = L_i\right)$ can be found based on previous equation system because all other quantities are function of $\overrightarrow{\mathcal{N}_i}$

- The (joint) distributions of those quantities may not have closed form. More importantly, the number of possible state values is prohibitive large

*see Higher Order Markov Model

IV. Theoretical Delay Model (1)

- We approximate the random processes $\overrightarrow{\mathcal{N}_i}$ & $\mathcal{L}_i$ by their deterministic "mean" trajectory $\overrightarrow{N_i}$ & L_i

- In other words, we consider the deterministic trajectory $\left(\overrightarrow{N_0}, L_0\right); \left(\overrightarrow{N_1}, L_1\right) = \mathbb{E}\left[\overrightarrow{\mathcal{N}_1}, \mathcal{L}_1 \mid \overrightarrow{N_0}, L_0\right]; (\overrightarrow{N_i}, L_i) = \mathbb{E}\left[\overrightarrow{\mathcal{N}_i}, \mathcal{L}_i \mid \overrightarrow{N_{i-\Delta}}, \mathcal{L}_{i-1}\right]$ … instead of dealing with transition probabilities between myriad possible trajectories

IV. Theoretical Delay Model (1)

- Thus, we have the following evolution equations

$$\begin{cases} N_{i+L_i+g}[1] \mathrel{+}= N_{i+L_i+g,Arrival} \\ N_{i+L_i+g}[2] \mathrel{+}= N_{i,C}[1]/G_i \\ \cdots \\ N_{i+L_i+g}[n_{PTmax}] \mathrel{+}= N_{i,C}[n_{PTmax}-1]/G_i \\ L_{i+1} = L_i - 1 + G_i \end{cases}$$

where

$$\begin{cases} G_i = \left\lceil \frac{N_{i,C}}{d} \right\rceil = \left\lceil \frac{N_i - N_{i,S}}{d} \right\rceil \text{ if } N_{i+L_i} + N_{i,C} > d \\ \qquad 0, \quad \text{otherwise} \\ N_{i,S}[n] = \frac{N_i[n]}{N_i} \times N_i (1 - 1/R)^{N_i - 1} \end{cases}$$

*Note: $i + L_i = \sum_{k=1}^{i-1} G_k$, and we use the notation $N_{i(,S,C)} = \sum_n N_{i(,S,C)}[n]$

IV. Theoretical Delay Model (1)

- When $G_i = 0$, subset merging occurs and the equations differ slightly

$$\begin{cases} N_{i+L_i}[1] \mathrel{+}= N_{i+L_i,Arrival} \\ N_{i+L_i}[2] \mathrel{+}= N_{i,C}[1] \\ \dots \\ N_{i+L_i}[n_{PTmax}] \mathrel{+}= N_{i,C}[n_{PTmax} - 1] \\ L_{i+1} = L_i - 1 \end{cases}$$

- These help us to "update" future state values based on previous ones

IV. Theoretical Delay Model (1)

- How to determine process termination point i_{term}?
  1. If we iterate outside of paging interval => i_{term} = I_{max}
  2. But the process may be terminated early if all MTDs are solved before Imax elapses => i_{term} = ?

- Note that a DQ process is finished when the queue is empty i.e., L_i = 0
- So, just iterate until we find i_{term}-th slot where $L_{i_{term}}$ = 0

IV. Theoretical Delay Model (1)

- Once we know N_i for all i, we can compute the average delay as

$$\mathbb{E}[D] = \frac{\sum_{i=1}^{i_{term}} (N_{i,S} \times i)}{\sum_{i=1}^{i_{term}} N_{i,S}}$$

Number of MTDs who succeed in i-th slot

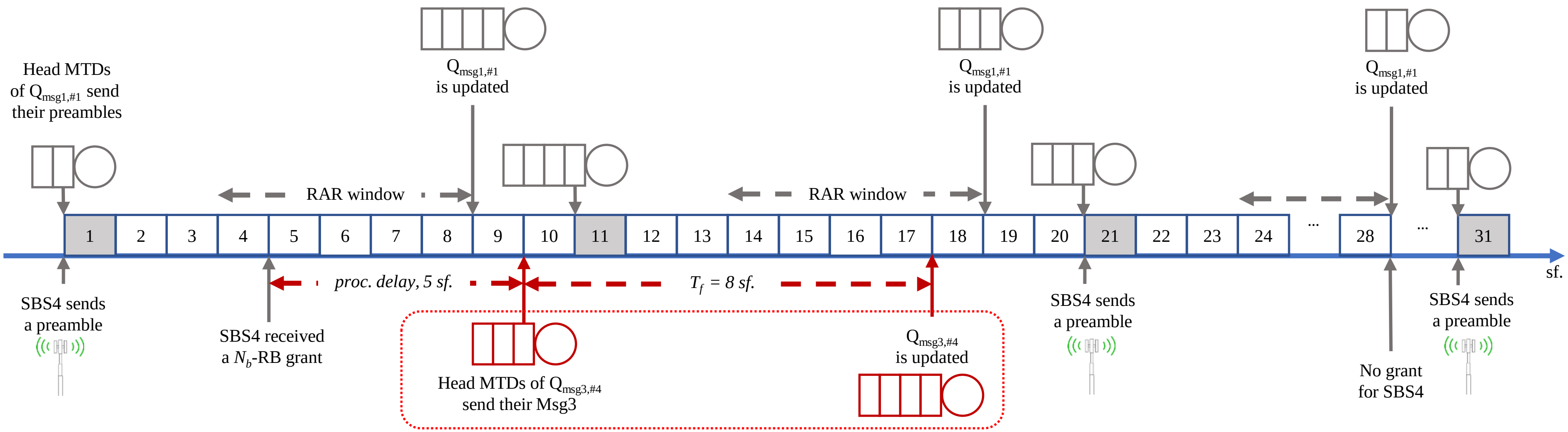
Delay of those MTDs who succeed in i-th slot

Total number of successful MTDs

and total service time (TST) it takes to resolve all MTDs simply as i_{term}

IV. Theoretical Delay Model (2)

- This model applies directly to Msg3 DQ
- For Msg1 DQ, modification is needed due to SBSs



IV. Theoretical Delay Model (2)

- An SBS only take part in Msg1 DQ process until its own Msg3 DQ process is terminated
- We assume that a Msg3 DQ process always finishes after i_{term} "Msg3 slots"
- Note that "Msg3 slots" are not periodic as Msg1 slots. They only appear when SBSs obtain grants from BS

We assume that an SBS finishes after it has obtained i_{term} grants

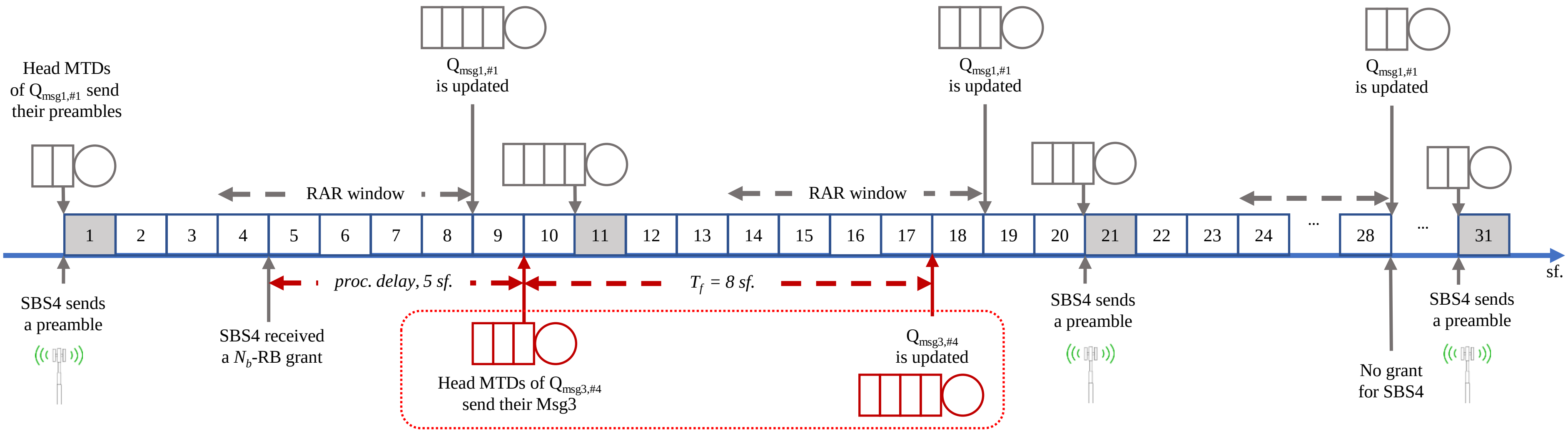
IV. Theoretical Delay Model (2)

- Now let us denote by $\mathcal{M}_i[k]$ the R.V. describing the number of SBSs that have obtained k grants up until i-th slot
- The (random) vector describing Msg1 process is thus $\langle \mathcal{N}_j^{m1}[1], \dots, \mathcal{N}_j^{m1}[n_{PTmax}], \mathcal{M}_j[1], \dots, \mathcal{M}_j[i_{term}] \rangle$

To distinguish with msg3 process

IV. Theoretical Delay Model (2)

- But the timings requires more complex modeling



SBS4 will retransmit in

- Second Msg1 slot (subframe 11) if it obtains no grant
- Third Msg1 slot (subframe 21) if it is granted in any subframe from 4-7
- Fourth Msg1 slot (subframe 31) if it is granted in subframe 8

IV. Theoretical Delay Model (2)

- SBSs who transmit in a slot will retransmit in either 1, 2, or 3 slots later, given that they have not reached i_{term} grants
- Thus, when processing the j-th slot, we need to update j+1, j+2, and (j+3)-th slots as well

IV. Theoretical Delay Model (2)

- The system of evolution equations for M_i are

$$M_{j+1}[0] \mathrel{+}= M_j[0] * P_0$$
$$M_{j+1}[i_{term}] \mathrel{+}= M_j[i_{term}] * P_0$$

Prob. of receiving no grant (retry in next slot)

$$M_{j+2}[0] = 0$$
$$M_{j+2}[1] \mathrel{+}= M_j[0] * P_1$$
$$M_{j+2}[i_{term}] \mathrel{+}= M_j[i_{term} - 1] * P_1$$

Prob. of receiving a grant in any of first 4 subframes of RAR window (retry 2 slots later)

$$M_{j+3}[0] = 0$$
$$M_{j+3}[1] = M_j[0] * P_2$$
$$M_{j+3}[i_{term}] = M_j[i_{term} - 1] * P_2$$

Prob. of receiving a grant in last subframe of RAR window (retry 3 slots later)

P0, P1, and P2 are calculable partly based on M_i and N_i^{m1}

IV. Theoretical Delay Model (2)

- The system of evolution equations for N_j and L_j are the same, except that

$$\begin{cases} G_i = \left\lceil \dfrac{M_{j,C} + N_{j,C}}{d} \right\rceil = \left\lceil \dfrac{M_j - M_{j,S} + N_j^{m1} - N_{j,S}^{m1}}{d} \right\rceil \text{ if } N_{j+L_i}^{m1} + M_{j,C} + N_{j,C}^{m1} > d \\ \qquad\qquad\qquad\qquad 0, \qquad \text{otherwise} \\ N_{j,S}[n] = \dfrac{N_j^{m1}[n]}{M_j + N_j^{m1}} \times (M_j + N_j^{m1})(1 - 1/R)^{M_j + N_j^{m1} - 1} \end{cases}$$

Same notation as before: $M_{j(,S,C)} = \sum_{k=1}^{i_{term}} M_{j(,S,C)}[k]$

IV. Theoretical Delay Model (2)

- The delay formula is a little bit different

$$\mathbb{E}[D] = \frac{\sum_{j=1}^{j_{term}} j * \left\{N_{j,S}^{m1} + \sum_{k=1}^{i_{term}} M_{j,s}[k] * N_{k+1,S}\right\}}{\sum_{j=1}^{j_{term}} \left\{N_{j,S}^{m1} + \sum_{k=1}^{i_{term}} M_{j,s}[k] * N_{k+1,S}\right\}}$$

Number of SBSs who have just obtained their (k+1)-th grant in *j*-th slot

Average number of successful SC-MTDs in a small-cell, given that the corresponding SBS just obtains its new (k+1)-th grant

- Terminal point j_{term} is found in a way similar to before

V. Simulation Results

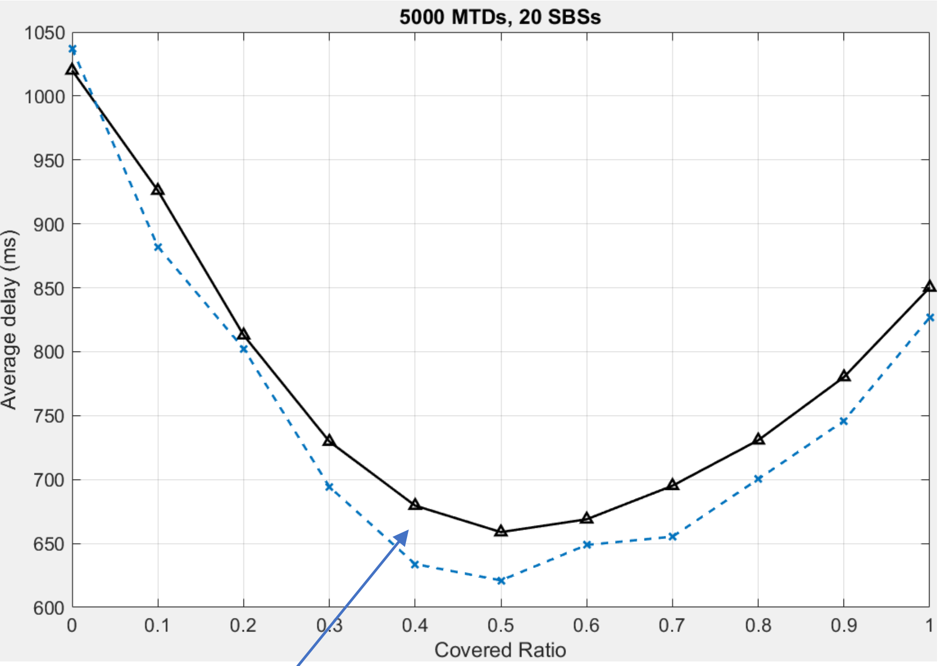
- Simulation parameters

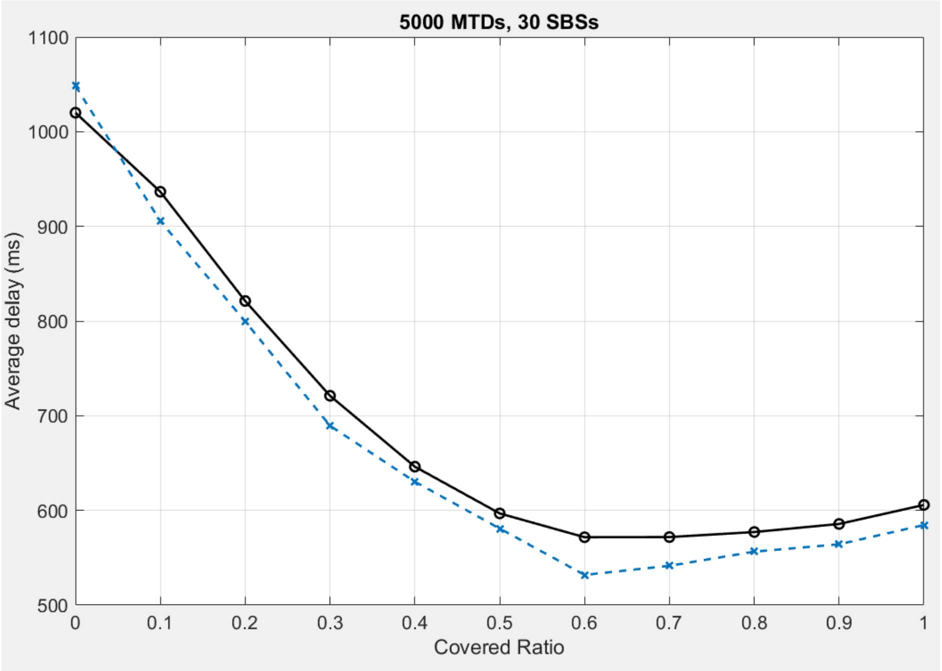
Parameters	Values
No. of MTDs in paged group	$N = 5000$
No. of SBSs	$N_{sc} = 20, 30, 40$
Covered ratio	0, 0.1, 0.2, ..., 1
RAO periodicity	$T_{RA\_REP} = 5$ ms
Subframe length	1 ms
No. of preambles	$K = 54$
Max no. of preamble trans.	$N_{PTmax} = 16$
RAR window size	$W_{RAR} = 5$ subframes
No. of grants per RAR	$N_{RAR} = 3$
No. of allocated RBs per grant for SBS	$N_b = 10$
Preamble detection prob. for i-th preamble trans.	$\left(1 - 1/e^i\right)$ for MTDs 1 for SBSs
Backoff Indicator	$BI = 120$ ms
Retrans. prob. for Msg 3 & 4	0.1
Max no. of Msg 3 & 4 HARQ trans.	5
Round-trip time of Msg 3 (Msg 4)	8 (5) subframes

V. Simulation Results

- Theoretical vs simulation: The trend matches, but underestimation level is a little high

$$(D_{sim} - D_{theo})/D_{sim} \sim 6.7\% \text{ at max}$$



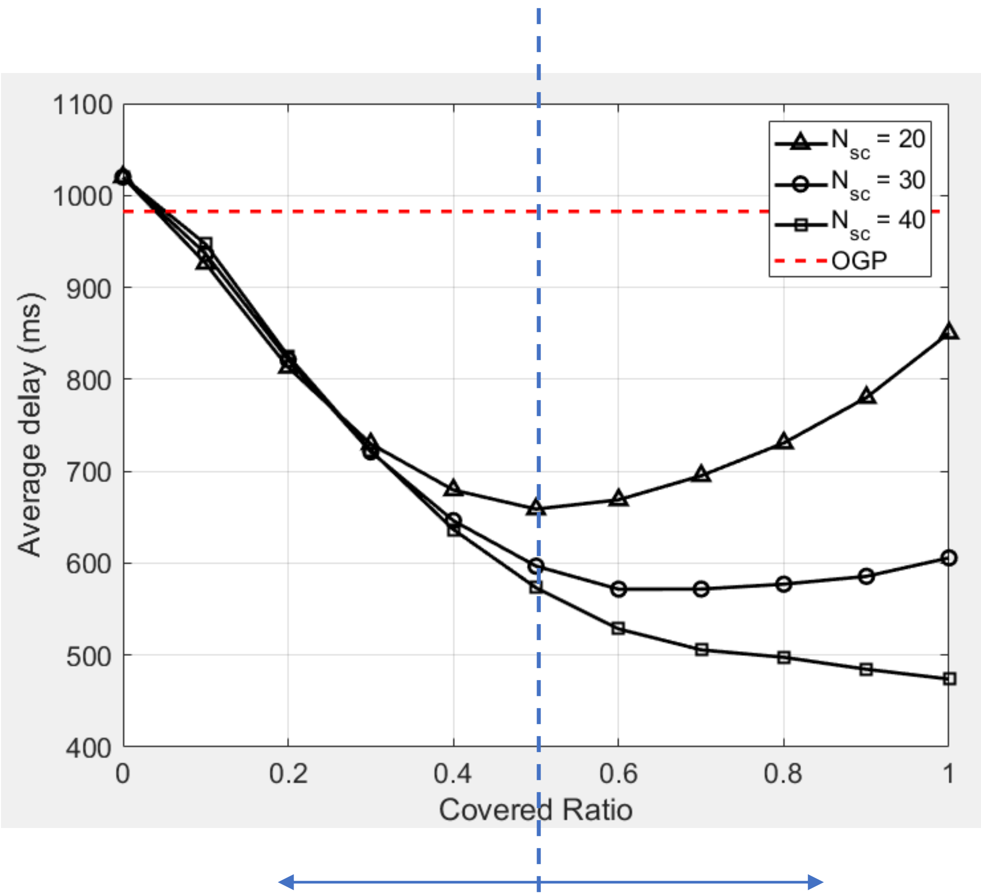


max deviation here

Black solid line = **Simulation**
Blue dashed line = Theory

V. Simulation Results

- Small-cell assisted GP vs Optimal GP [2]: delay



Heavy Msg1 contention at the macro cell

Heavy local (Msg3) contention at the small-cells

- Given **same covered ratio:**

- The more SBSs there are, the lower delay becomes

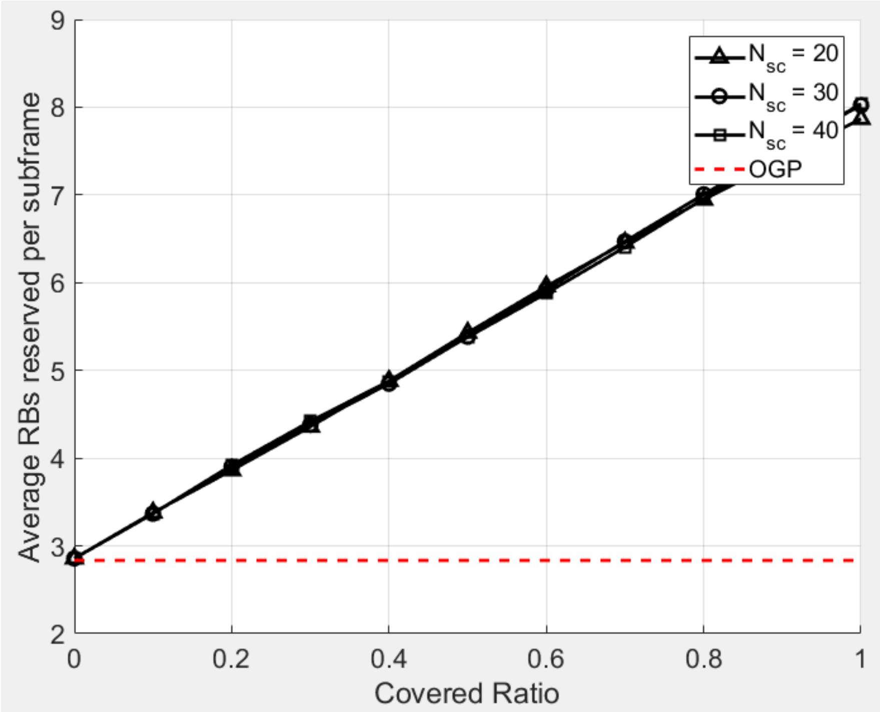
- Given **same number of SBSs:**

- When the ratio of MTDs covered is increase, delay goes down at first, then hit a breakpoint and goes up again

- Easily outperform OGP (delay-wise)

V. Simulation Results

- Avg. PUSCH resource consumption (over the whole paging interval I_{max})



- Amount of PUSCH RB consumed increases linearly with covered ratio => tradeoff for delay improvement
- Given a covered ratio, consumption is almost the same regardless of N_{sc} => increasing number of SBSs offers “real” gain
- But we cannot increase N_{sc} forever (that would bring back heavy Msg1 contentions)

VI. Conclusion

In this presentation, we have

- Introduced GP as a pull-based RAN overload control scheme in cellular mMTC
- Proposed a small-cell assisted GP scheme to share access load between PRACH and PUSCH
- Proposed an enhanced DQ-based contention resolution protocol to handle contention during both Msg1/Msg3 transmissions
- Proposed a theoretical delay model & tested its correctness as well as the effectiveness of the framework against OGP

References

[1] G. Farhadi and A. Ito, "Group-Based Signaling and Access Control for Cellular Machine-to-Machine Communication," 2013 IEEE 78th Vehicular Technology Conference (VTC Fall), Las Vegas, NV, 2013, pp. 1-6.

[2] O. Arouk, A. Ksentini, and T. Taleb, “Group paging-based energy saving for massive mtc accesses in lte and beyond networks,” IEEE J. Sel. Areas Commun., vol. 34, no. 5, pp. 1086–1102, May 2016.