Massive Multiple Access in LTE-based Machine-Type Communications: Recent Results

Bui Hoang Anh Tuan

Computer Communications Lab, University of Aizu, Japan

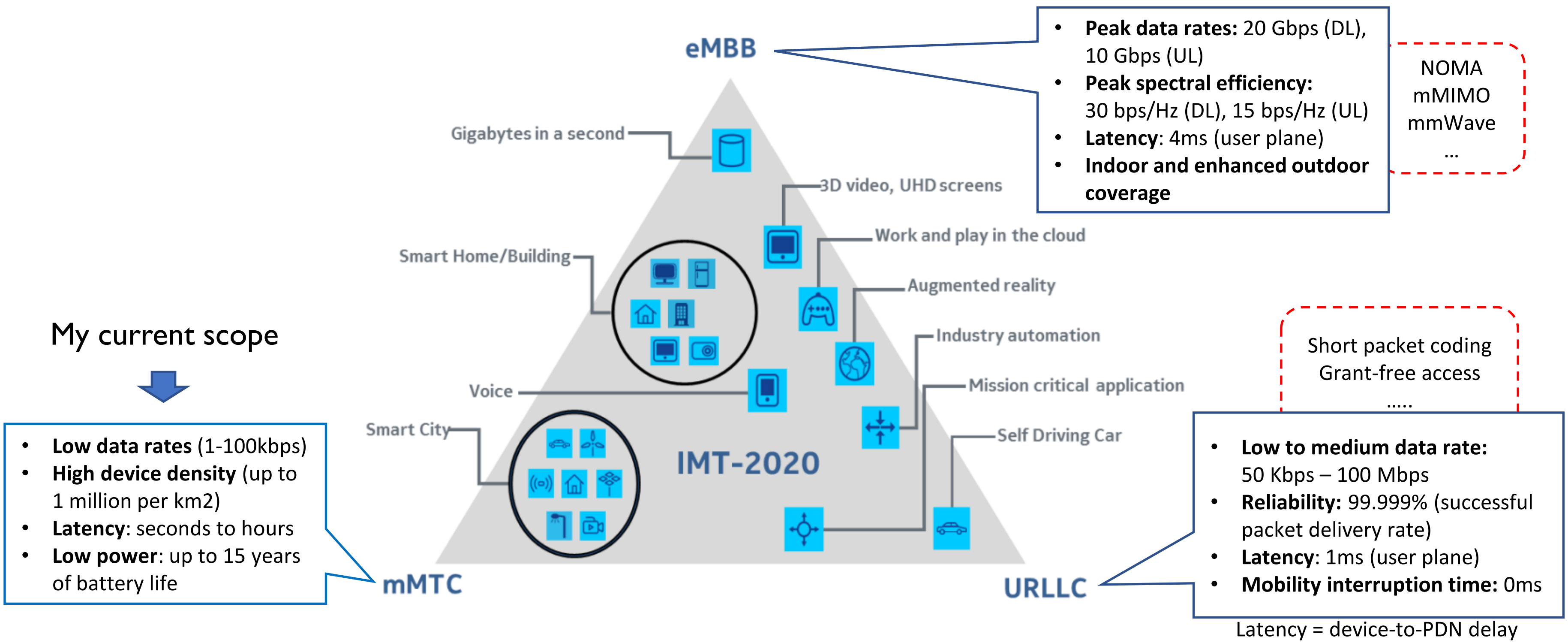
January 23rd , 2019

Outline

- Massive Machine-Type Communications (mMTC) - a 5G use case
- mMTC in 3GPP LTE cellular network
  - Signaling congestion issue
- Recent results
  - Push-based approaches
  - Pull-based approaches
  - Other approaches
- My current (very rough) idea

mMTC - a 5G use case

- What is 5G? A wireless network that supports

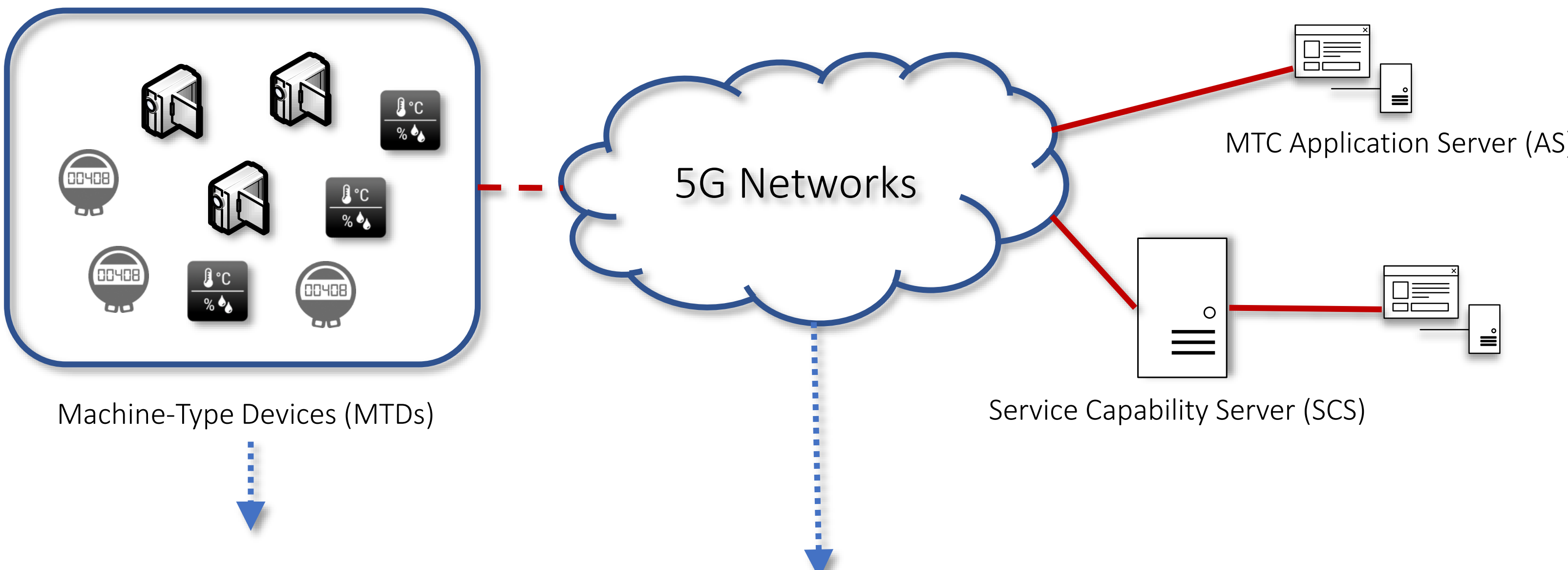


- One technology can't meet all requirements → 5G will be realized using both old & new technologies/networks

http://foothill.ieee-bv.org/2018/08/ieee-foothill-comsoc-moving-from-4g-to-5g-june-27-2018-talk/

mMTC - a 5G use case

- mMTC as a 5G use case? Autonomous communications between billions MTDs and servers via 5G networks

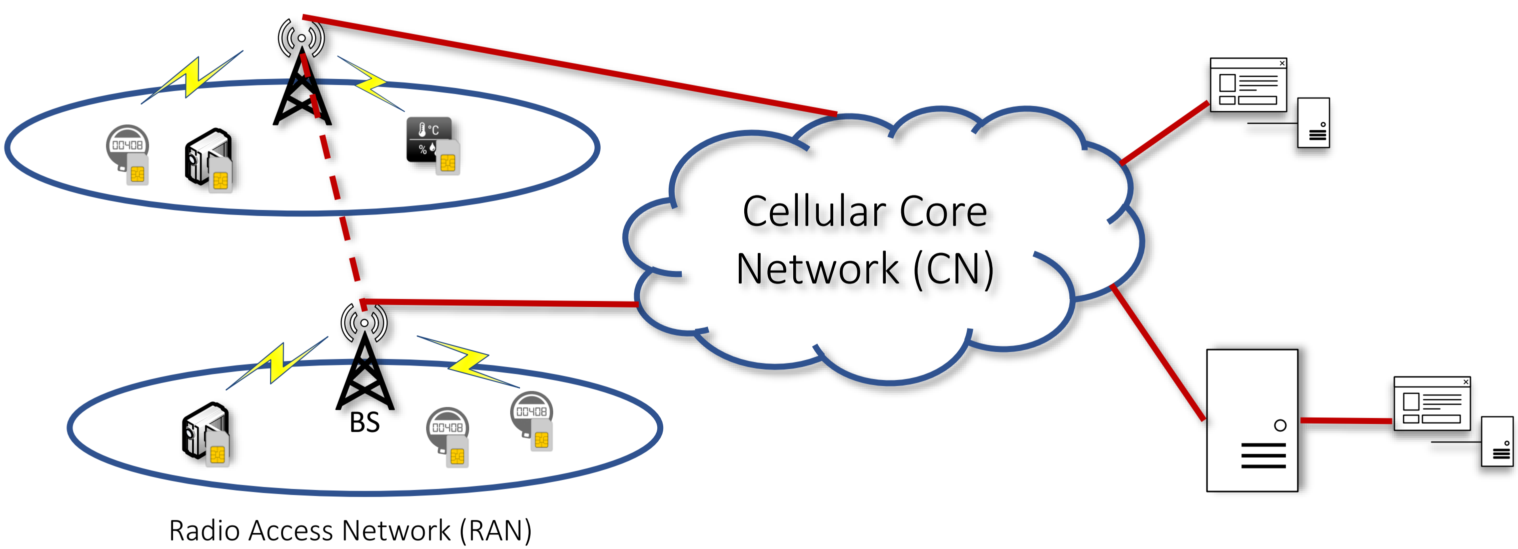


- ➢ Ubiquitously deployed
- ➢ Massive population
- ➢ Diverse range of applications

Which technology for mMTC in 5G?
WiFi, or VLC, or ZigBee, or *cellular*… ?

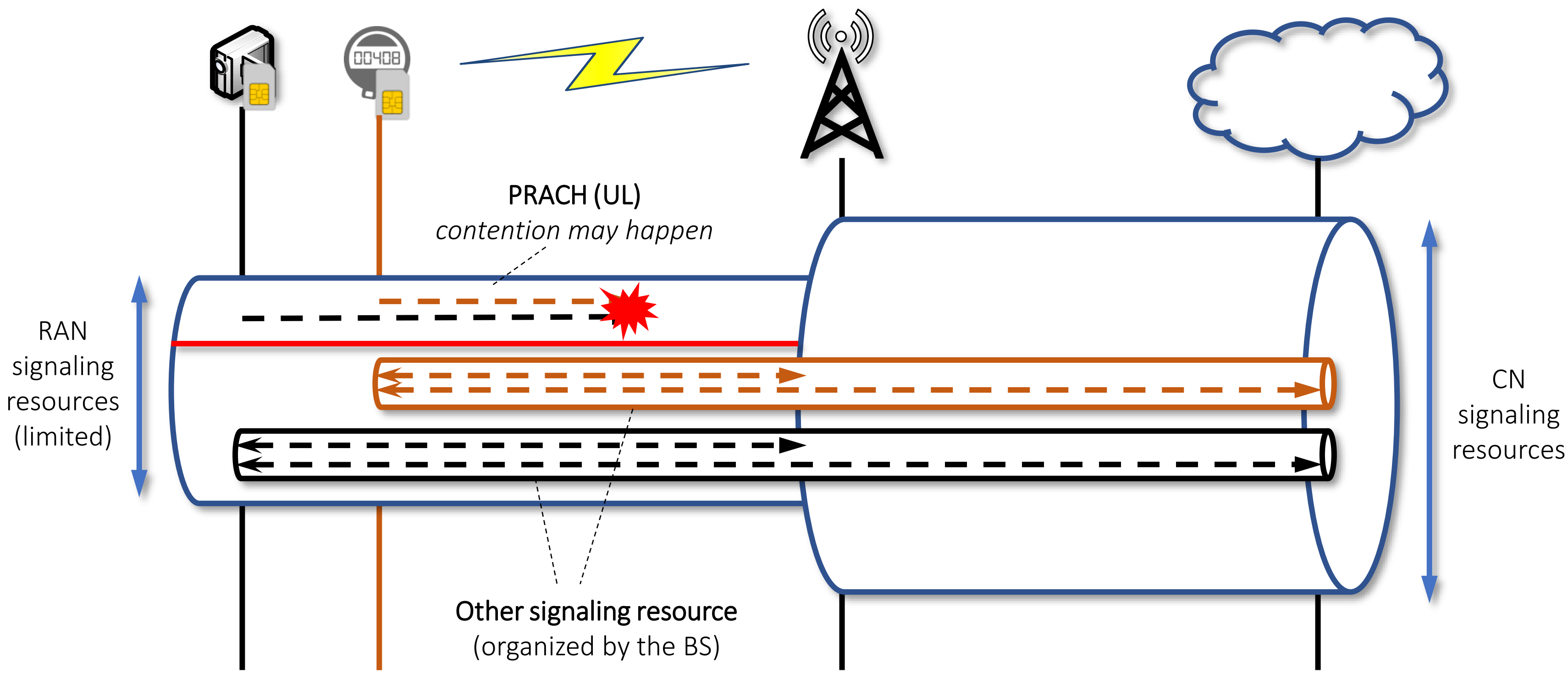
mMTC in 3GPP LTE cellular network

- Cellular network (LTE) is widely considered as one of the best choices to accommodate mMTC
  - ➢ Huge coverage → supports MTDs' ubiquity
  - ➢ Matured and well-adopted → easy massive installation



mMTC in 3GPP LTE cellular network

- *But* challenges arise because LTE wasn't design for mMTC
  - ➢ Complex connection establishment procedure over limited signaling BW → signaling overload in mMTC context



mMTC in 3GPP LTE cellular network

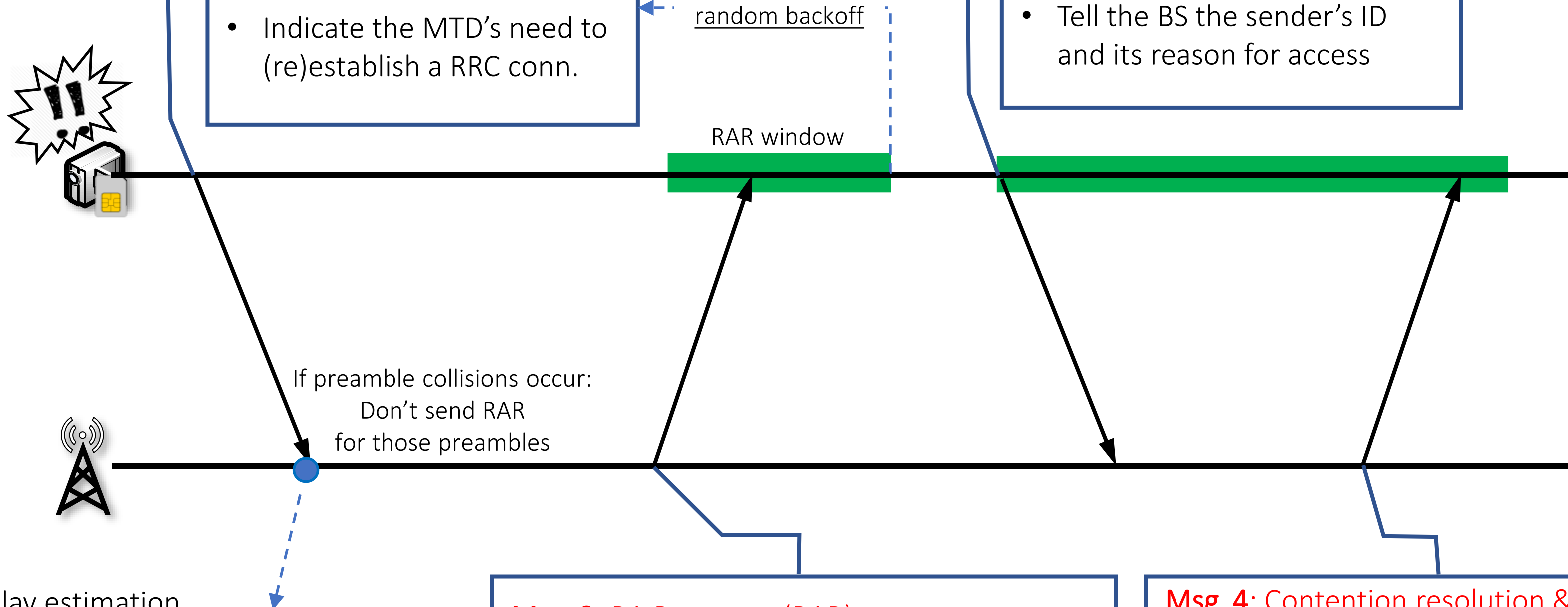
- Main purposes of RA procedure?
  1) Achieve UL synchronization (how?)
  2) Obtain dedicated UL resource for subsequent messages
- When does MTD invoke RA procedure?
  a) Initial access from RRC_IDLE
  b) RRC conn. re-establishment (radio-link / handover / integrity check… failures)
  c) Handover
  d) DL data arrives during RRC_CONNECTED and PRACH is needed (e.g., MTD is OUT_OF_SYNC)
  e) UL data arrives during RRC_CONNECTED and PRACH is needed (e.g., MTD is OUT_OF_SYNC or has no UL resource for sending "Scheduing Request")

mMTC in 3GPP LTE cellular network

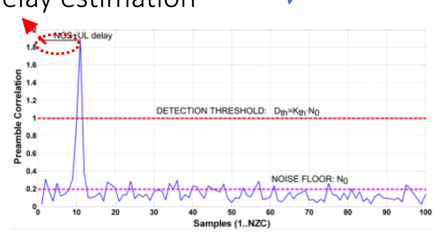
- Let's take a look at the RA procedure

Msg. 1: RA preamble
Sent over: **PRACH**
- Indicate the MTD's need to (re)establish a RRC conn.

Msg. 3: RRC connection request
Sent over: UL **data** resource
- Tell the BS the sender's ID and its reason for access



UL delay estimation

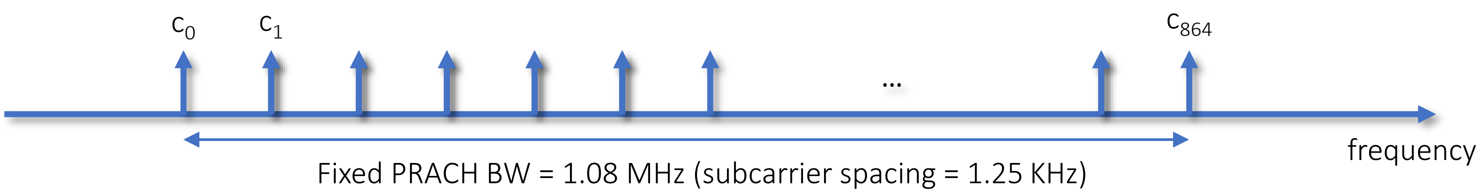


Msg. 2: RA Response (RAR)
"Pointer": sent over **DL signaling resource**
- Assign dedicated UL for Msg. 3
- Inform the UL delay of the sender for synchronization purpose
- Acknowledge upto N_{RAR} preambles per RAR

Msg. 4: Contention resolution & RRC conn. setup
"Pointer": sent over **DL signaling resource**
- Acknowledge the successful UE
- Specify the *configurations* needed by the MTD to actually establish a connection

mMTC in 3GPP LTE cellular network

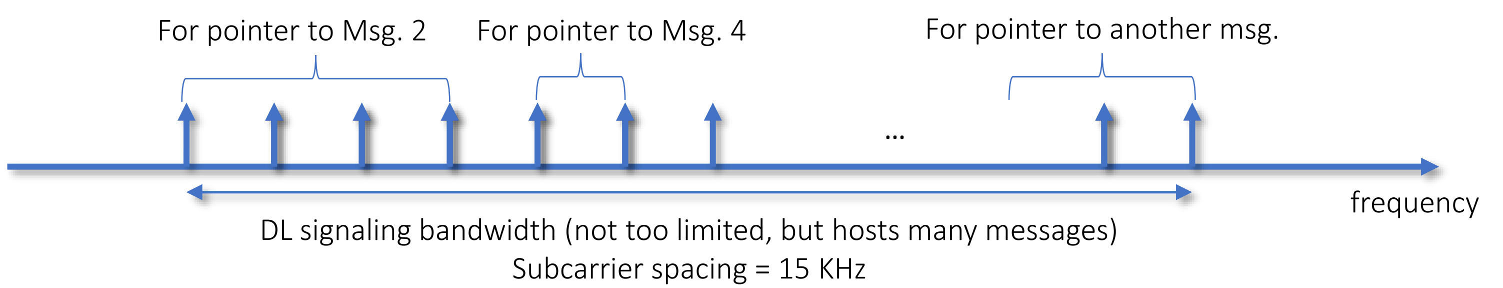
- Msg. 1, 2, and 4 are bottlenecks of this procedure
  - Msg. 1 is a preamble sequence mapped on PRACH subcarriers
  - MTDs sending different preambles may still be separated since preambles are orthogonal
  - Number of orthogonal sequences R that can be constructed on PRACH is limited



→ severe preambles collisions when number of competing MTDs in a slot is high

mMTC in 3GPP LTE cellular network

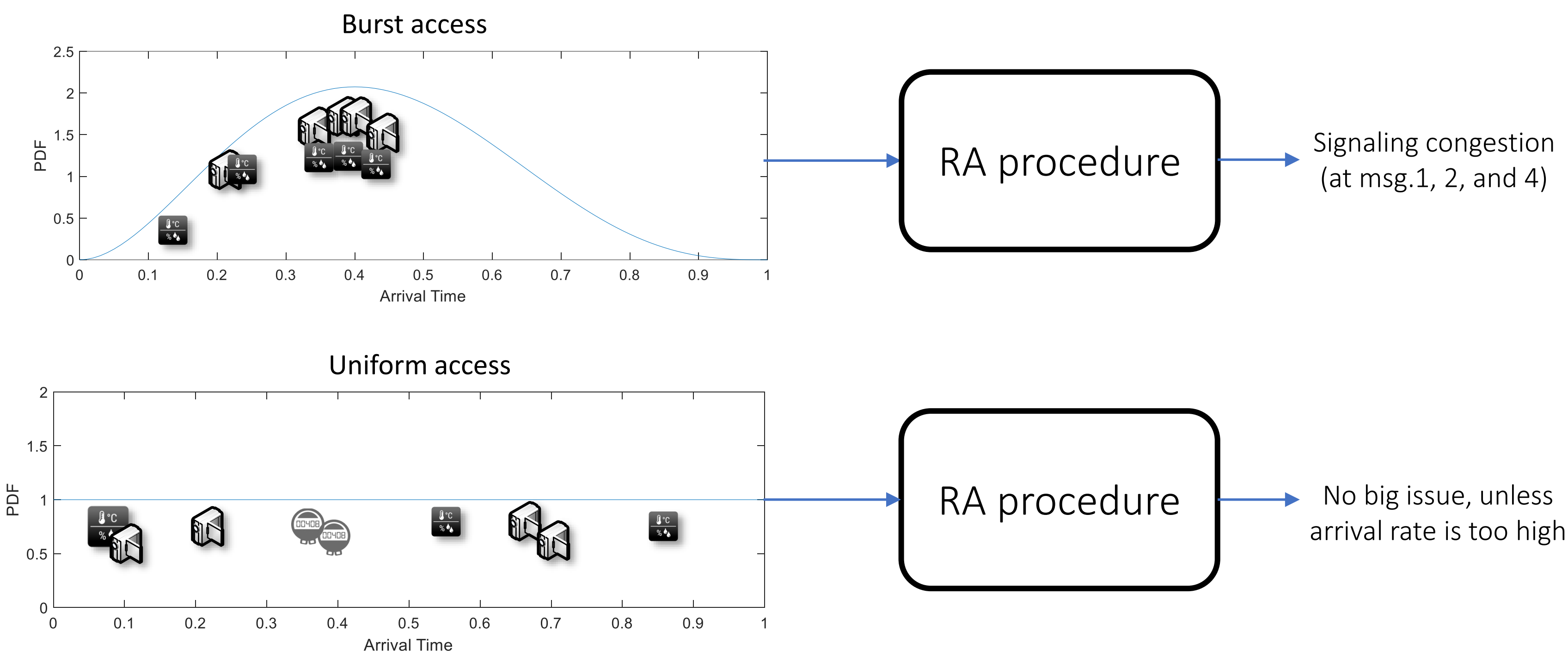
- Msg. 1, 2, and 4 are bottlenecks of this procedure (cont.)
  - "Pointers" to Msg. 2, 4, and all other DL messages are scheduled on the same DL signaling channel
  - When DL signaling resource is insufficient, some messages may be dropped
  - What happen when Msg. 2 or Msg. 4 is dropped?



→ insufficient DL signaling resource when BS need to send messages to many MTDs

mMTC in 3GPP LTE cellular network

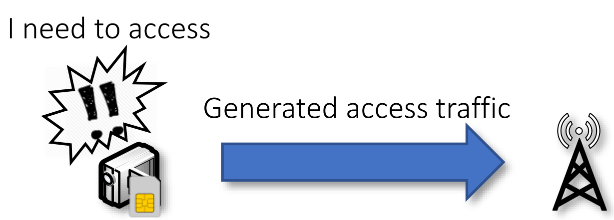
- Signaling congestion is bound to occur under high access intensity (simultaneous or burst access)



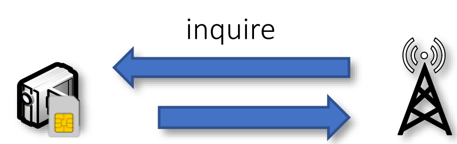
When signaling congestion occur, MTDs are likely to exceed the number of allowed attempts and gets “blocked”

Recent results

- How many approaches? Depends on how solutions are classified
- Most common way is to classify based on how access traffic is generated
  - Push-based: solutions assuming that MTDs proactively generate access traffic (we are assuming push-based until now)
  - Pull-based: solutions assuming that MTDs only generate access traffic when probed by the network



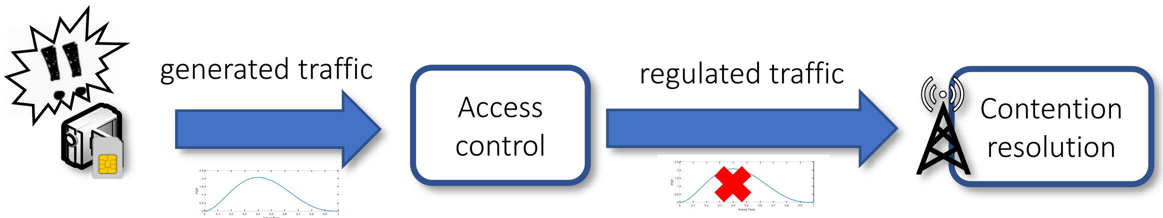
MTD “pushes” its traffic toward network



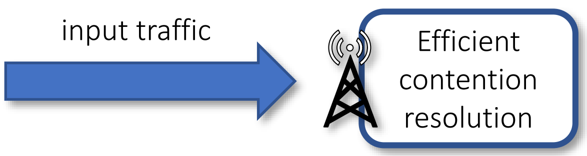
Network “pulls” traffic toward itself

Recent results

- Push-based can then be divided into sub-categories (not mutually exclusive):
  a. Solutions that try the control the access traffic generated by the MTDs



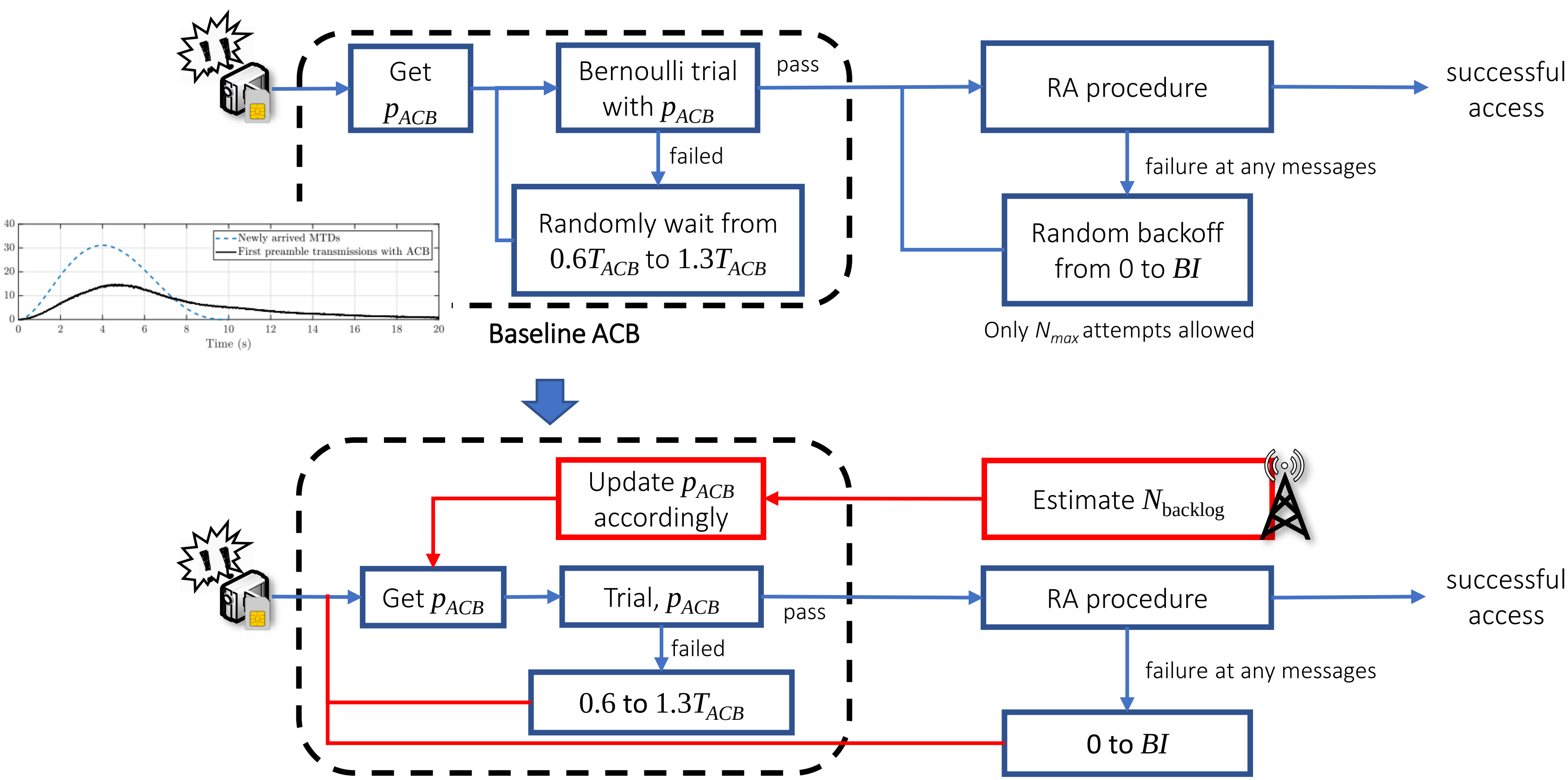
  b. Solutions that try to efficiently resolve contentions caused by the generated access traffic (either via better contention resolution mechanisms or utilizing additional information)



  c. Other solutions

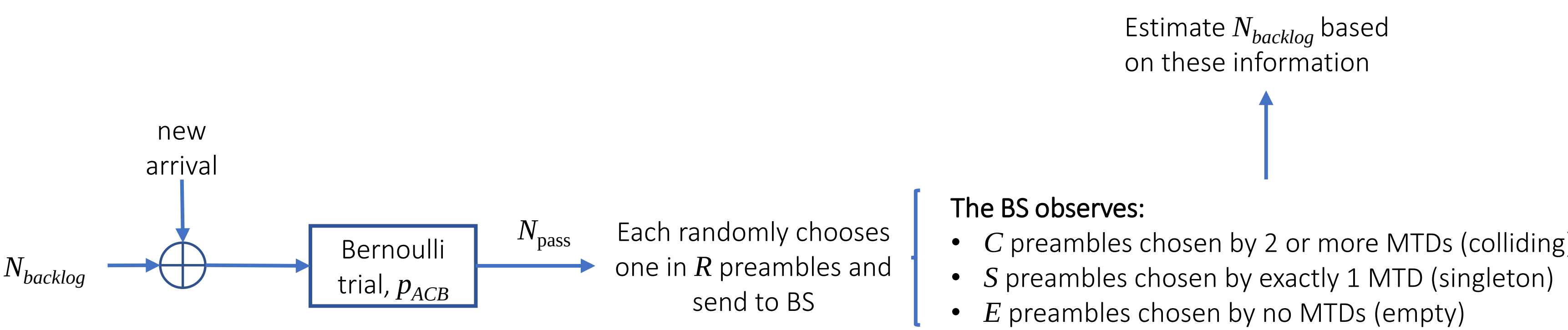
Recent results

- Push-based (a.) recent works
  - ➢ Most related papers try to improve baseline Access Class Barring (ACB) by adaptively adjusting the barring factor p_{ACB}



Recent results

- Push-based (a.) recent works (cont.)
  - Most schemes assume that MTDs failing the trial will retry in next slot i.e., $T_{ACB} = 0$
  - How do these adaptive scheme estimate current $N_{backlog}$? Based on observed status of the preambles in this slot



  - Different papers use different estimation technique

Recent results

- Push-based (a.) recent works (cont.)
  - Maximum Likelihood Estimation (MLE)
    - [1] derive close-form of $P(C = c, S = s \mid N_{pass} = n)$ and estimate $\tilde{N}_{pass}$ as the *n* that maximize this prob.
    - [2] derive close-form of $P(E = e, S = s \mid N_{pass} = n)$ and estimate $\tilde{N}_{pass}$ as the *n* that maximize this prob.
    - [1] proves that estimating $\tilde{N}_{pass}$ using only *E* yields poor result when N_{pass} > 50
  - Bayesian estimation
    - [1] also uses a Bayes estimator to find $\tilde{N}_{pass}$ that minimizes the expected relative estimation error
    - [3] uses Bayes rule to estimate $\tilde{N}_{backlog}$, given *E* = e and assume Poisson as a priori distribution of $N_{backlog}$
  - Other techniques
    - [4] assumes that current p_{ACB} is close to optimal and approximate the true optimal of the slot $p_{ACB\_opt} = f(p_{ACB\_current}, C, R)$, then estimate current $\tilde{N}_{backlog} = f(p_{ACB\_opt})$

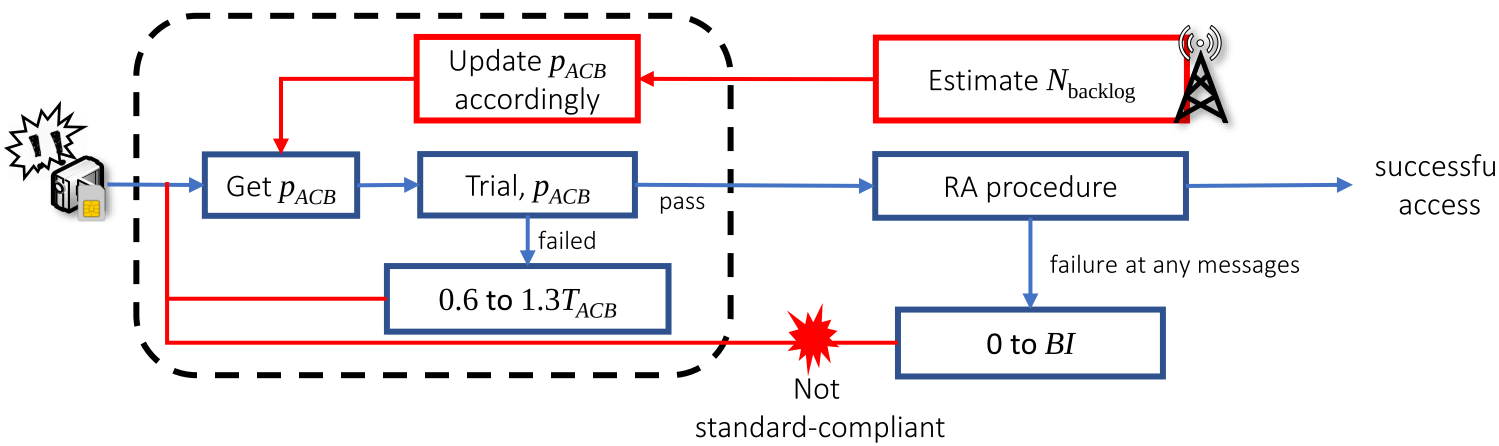
Recent results

- Push-based (a.) recent works (cont.)
  - After obtaining the estimates $\widetilde{N}_{\text{backlog}}$, how does the BS decide p_{ACB} for next slot?
  - Since the number of devices subjected to ACB in next slot is $N_{backlog\ (next\ slot)} = N_{backlog\ (current\ slot)} - S + N_{arrivals\ (next\ slot)}$, BS must "predict" the number of new arrivals in next slot as well
  - Most studies assume that N_{arrivals} [2] or the "rate at which N_{backlog} varies" [4] can't change quickly between consecutive slots and "predict" that
    - $\widetilde{N}_{\text{arrivals (next slot)}} = \widetilde{N}_{\text{arrivals (this slot)}}$
    - Or $\widetilde{N}_{\text{backlog (next slot)}} = \widetilde{N}_{\text{backlog (this slot)}} + \{\widetilde{N}_{\text{backlog (this slot)}} - \widetilde{N}_{\text{backlog (prev. slot)}}\}$

  - Then p_{ACB} is updated so that $\mathbb{E}[N_{pass\ (next\ slot)}]$ = # of preambles R

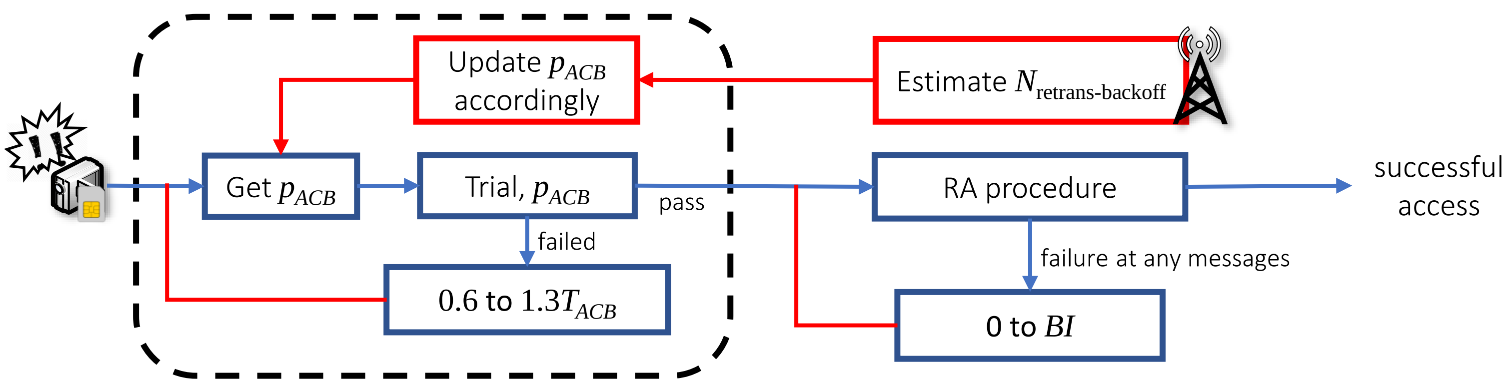
Recent results

- Push-based (a.) recent works (cont.)
  - Pros of adaptive ACB?
    - Success rate ~1, near-optimal delay performance (almost optimal p_{ACB})
    - Easy to select pACB once $\widetilde{N}_{\mathrm{backlog}}$ are obtained
  - Cons of adaptive ACB?
    - Not standard-compliant: ACB does not apply to backoff devices according to 3GPP's specification
    - High energy consumption: backlogged MTDs need to listen to update p_{ACB} in every slot (since T_{ACB} is usually set to 0 for ease of p_{ACB} calculation)



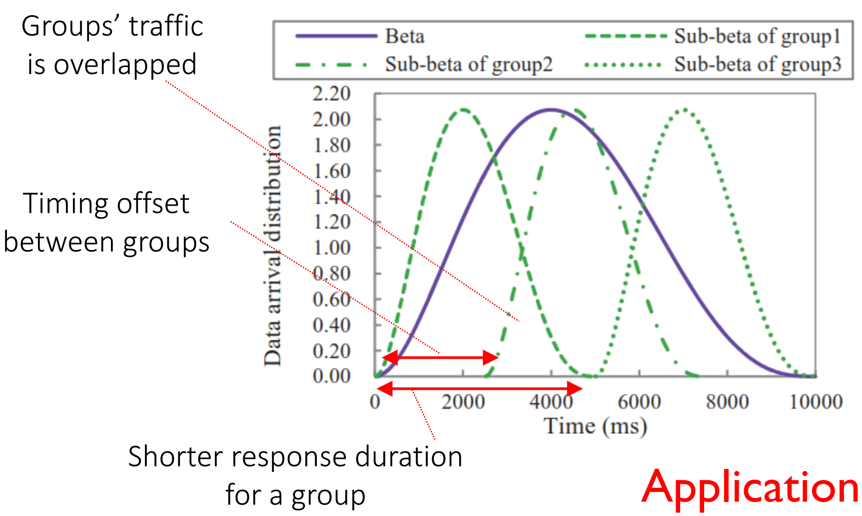
Recent results

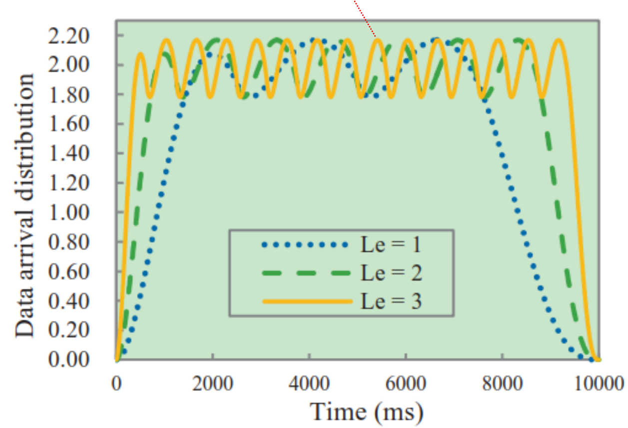
- Push-based (a.) recent works (cont.)
  - [5] is one adaptive ACB work that is standard-compliant
    - ACB only applies to new MTDs who haven't initiated RA procedure
    - Estimate the number of MTDs in **backoff state** (due to failure in RA procedure) who retransmit in this slot in a recursive manner
    - If $N_{\text{retrans-backoff}} > R$, then $p_{ACB} = 0$ (barred all new MTDs)
      If $N_{\text{retrans-backoff}} = 0$, then $p_{ACB} = 1$ (let all new arrivals in)
      Otherwise p_{ACB} is a cubic function of $N_{\text{retrans-backoff}}$ (chosen empirically)
    - Their p_{ACB} is non-optimal compared to non-compliant works (optimal p_{ACB} is hard to determine if ACB doesn't apply to backoff MTDs)



Recent results

- Push-based (a.) recent works final note
  - ➢ There are approaches in push-based (a.) other than ACB
  - ➢ [6] splits MTDs into groups, then assign shorter response durations & different timing offset for each group
  - ➢ Although each group still access in burst, the shorter response and timing offset cause the traffic to appear as uniform over the same period

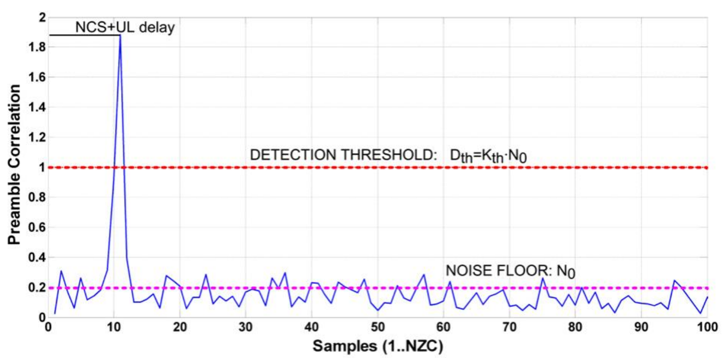




Application level traffic reshaping

Recent results

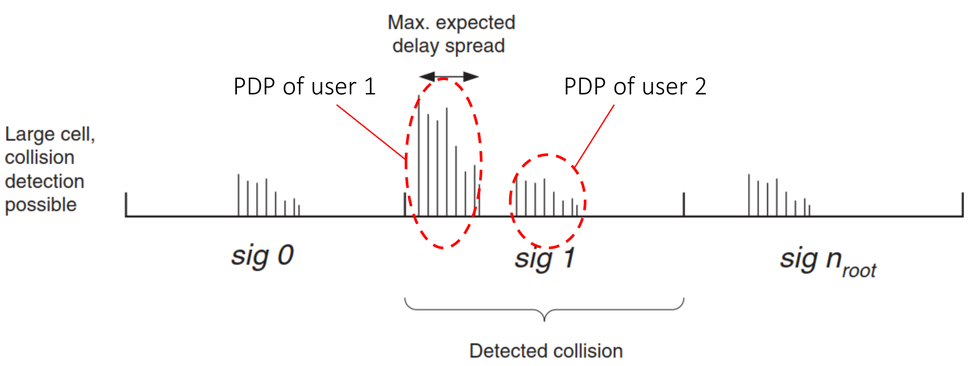
- Push-based (b.) recent works
  - ➢ Most papers try to exploit distance information as an additional "domain" for efficient contention resolution
  - ➢ One way is through the use of Timing Advanced (TA) in Msg.2
    - In LTE, BS detects the presence of a preamble by cross-correlating received signal with corresponding reference sequence
    - If a preamble is detected, its delay is also found at the same time
    - This delay is included in Msg. 2's TA field so relevant device can time Msg. 3 properly



What if multiple MTDs send the same preamble?

Recent results

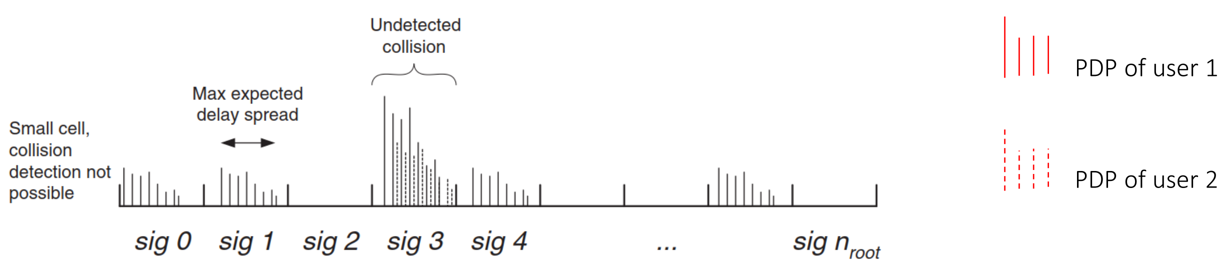
- Push-based (b.) recent works
  - ➢ There are two hypotheses to that, but no definite answer
  - ➢ Hypothesis 1: BS can detect collision
    a. In large cell, BS may be able to detect preamble collision after decoding
    b. In 3GPP's simulation, they assume that the BS can't decode the preamble → collision detection (but in another sense)



In large cell, propagation delay difference can be >> delay spread
The BS may thus be able to detect collision if the devices are spaced far enough

Recent results

- Push-based (b.) recent works
  - Hypothesis 2: BS can’t detect collision
    - In small cell, collision detection may not be possible, thus BS still sends UL grant for that preamble
    - When multiple devices receive the same UL grant, they will use the same resource and TA for Msg. 3
    - Even if TA is applicable to only one MTD, misaligned Msg. 3 from others on same resource causes interference and no Msg. 3 goes through
    - In some cases Msg. 3 can be decoded despite multiple Msg. 3 transmissions

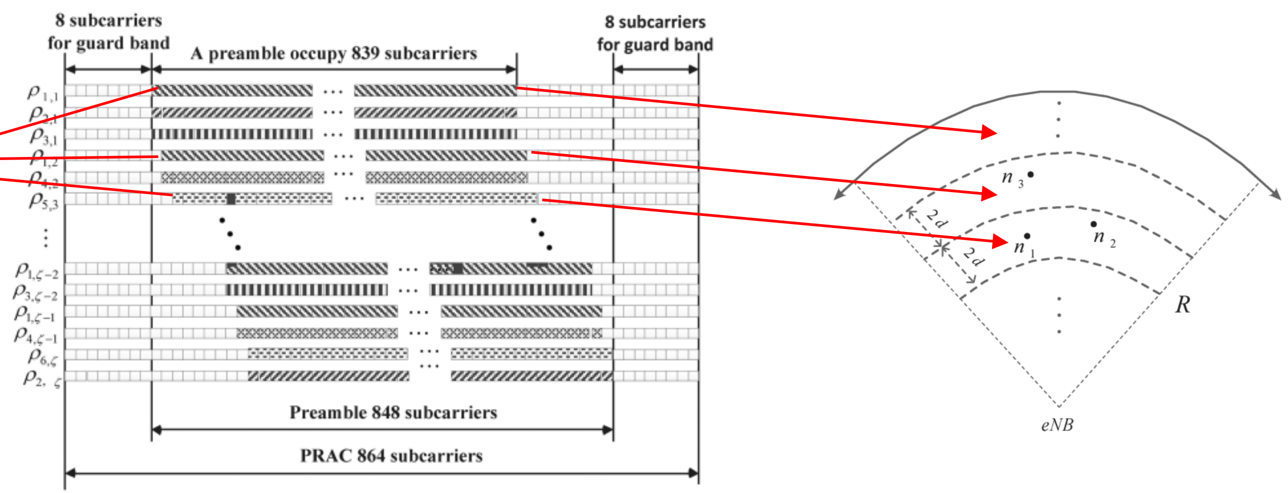


In large cell, propagation delay difference is small compared to delay spread
The BS thus can’t decide if multiple peaks are due to multipath or multi-transmission

Recent results

- Push-based (b.) recent works assuming hypothesis 1a
  - [7] divide BS's coverage into annuluses and assign different set of PRACH subcarriers for each annulus to transmit preambles on
    - Use a PHY-layer estimation method to estimate the number of MTDs selecting a certain preamble in a certain annulus
    - Only send Msg. 3 to a detected preamble in an annulus if there is only one device transmitting it

Preambles of different annuluses are no longer orthogonal?



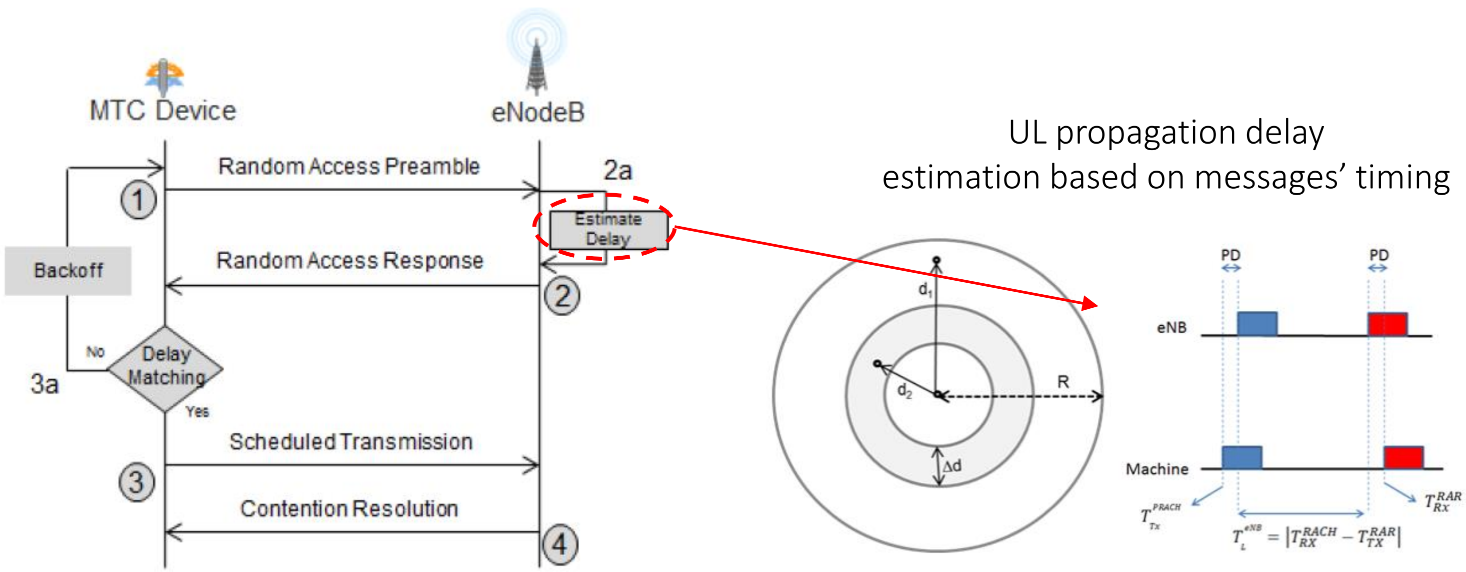
Recent results

- Push-based (b.) recent works (hypothesis 1a, cont.)
  - [8] assumes that preamble collision can be detected if maximum difference in distance between MTDs selecting that preamble exceed a threshold
    - If collision is detected, BS still sends grant but indicate that fact to MTDs in the RAR
    - An MTD, upon receiving collision indicator, estimate the number N of MTDs colliding with it
      It is assumed that BS can keep the # of contending MTDs per slot in check via optimal, non-compliant ACB
    - It then proceeds to Msg. 3 with prob. $1/(N+1)$ so that expected number of devices transmitting Msg. 3 on the granted resource is 1

$\rightarrow$ Even if an MTD collides in Msg. 1, its Msg. 3 may still be delivered

Recent results

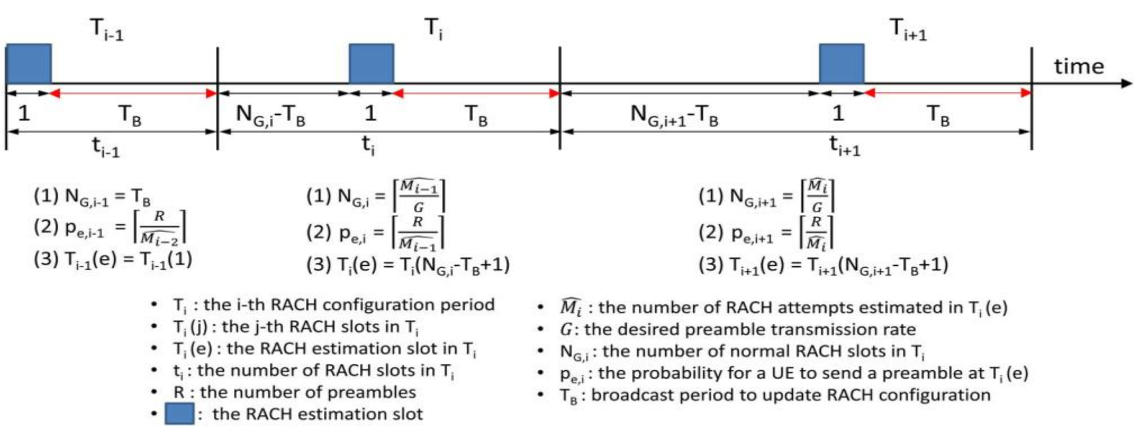
- Push-based (b.) recent works assuming hypothesis 2
  - ➢ In [9], stationary MTDs compare their own estimated UL delay (assuming UL delay = DL delay) with TA and only proceed to Msg. 3 if the two are close enough
    - ▪ An MTD may still succeed even if its preamble is also sent by others if its timing advance is unique among them



UL propagation delay estimation based on messages' timing

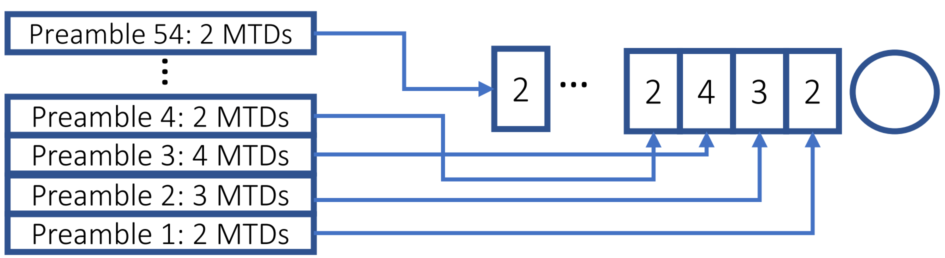
Recent results

- Push-based (b.) recent works: improving contention resolution mechanisms
  - [10] divides time axis into configuration periods. Period i consists of one estimation slot and $(t_i - 1)$ normal slots.
    - MTDs joining i-th period transmits its Msg. I in estimation slot with prob. $p_{e,i}$, and also randomly transmits Msg. I in one of the $t_i - 1$ normal slots
    - MTDs failing in the i-th period will have to wait till next period to rejoin
    - BS estimates number of devices contending in i-th period (via preamble statuses in estimation slot) and set number of normal slots $(t_{i+1} - 1)$, and $p_{e,i+1}$ for the next period accordingly

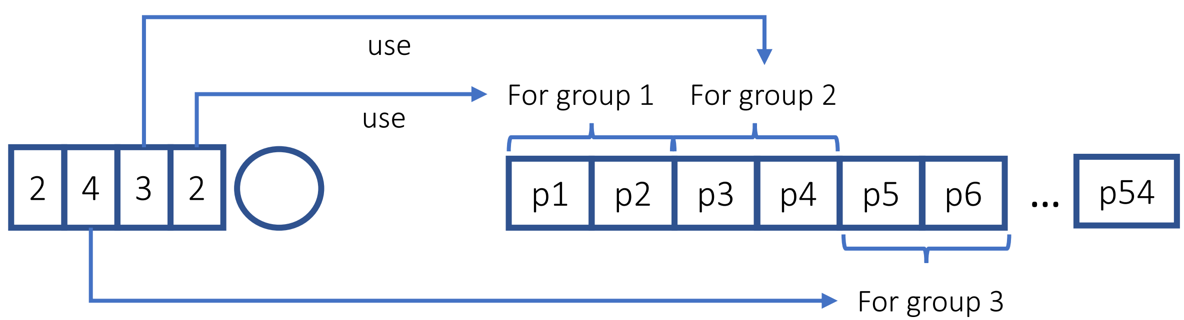


Recent results

- Push-based (b.) recent works: improving contention resolution mechanisms (cont.)
  - [11] is a DQ-based approach
    - Similar to CRQ i.e., divided into groups based on chosen preambles



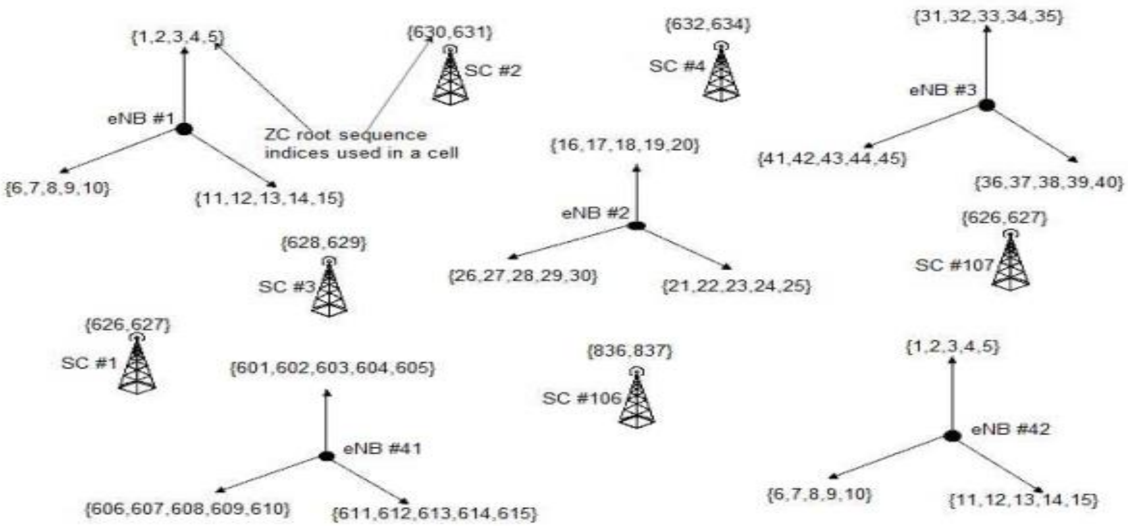
    - However, multiple groups can retransmit in the same timeslot using different subsets of preamble



Mitigate impact of over-division

Recent results

- Push-based (c.) recent works: Other approaches
  - [12] concerns about preamble reuse in a micro cells – macro cells setting
    - Multiple preambles can be generated from a single root sequence. Number of root sequences are also limited
    - The smaller the cell size, the less root sequences needed for generating a predefined number of preambles
    - A centralized root sequences allocation scheme can enhance preamble usage efficiency



Recent results

- Push-based (c.) recent works: Other approaches (cont.)
  - [13] assumes TDD-LTE and low-cost (LC) MTDs scenario
    - Multiple PRACHs and multiple narrowband DL signaling channels. LC-MTDs can only monitor one signaling channel at a time
    - Undetectable preamble collisions (hypothesis 2): BS always send a grant for a detected preamble
    - Key point: BS sents **different** UL grants for one detected PRACH preamble on different signaling channels, and an LC-MTD randomly chooses a signaling channel to obtain grant
    - Collision in Msg. 3 only occurs when multiple LC-MTDs sending the same preamble on the same PRACH also select the same signaling channel to obtain grant

Issues:
1) very high DL signaling usage
2) Even if some MTDs of the same PRACH preamble select different UL grants, the TA in those grants would only works with one of those MTDs, not all of them

Recent results

- Push-based (c.) recent works: Other approaches (cont.)
  - ➢ [14] improves the feedback structure in Msg. 2 to reduce feedback load (or serve more MTDs per Msg. 2)

Assume that preamble 1, 2, 3 are detected in a slot

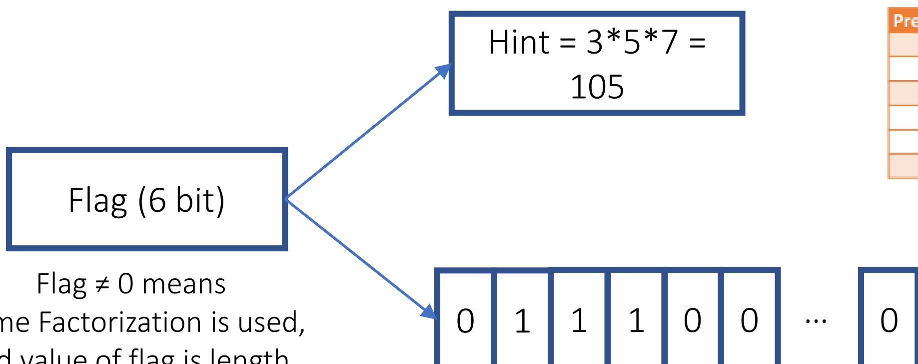
Preamble 1’s ID (000001)	Preamble 2’s ID (000010)	Preamble 3’s ID (000011)

Normal feedback message’s header

Preamble-to-prime mapping table

Preamble	Prime
0	2
1	3
2	5
3	7
⋮	⋮
63	311

MTDs then perform prime factorization on this hint i.e., 105 = 3*5*7
=> according to the table, preamble 1,2,3 are acknowledged



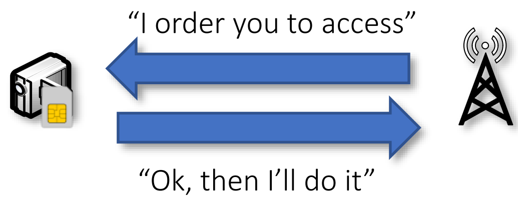
Flag ≠ 0 means Prime Factorization is used, and value of flag is length of the “hint” subheader

Use when there are many preambles to acknowledged (factorization costs too much)

Each bit corresponds to state of a preamble (1 = detected)
Maximum 64 bits (because there are max. 64 preambles)

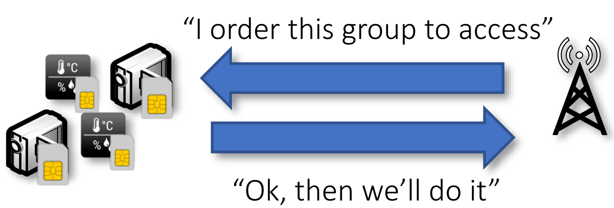
Recent results

- Pull-based schemes: MTDs generate access traffic only when inquired i.e., paged



BS may either specify the preamble to be used by the MTD (contention-free) or not. In latter case, MTD initiates contention-based RA procedure

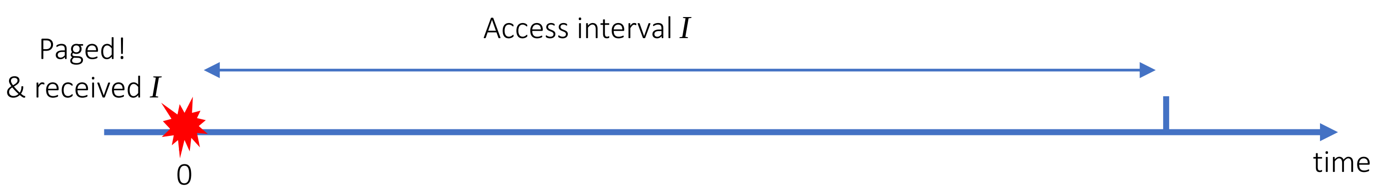
- The dominating pull-based approach is “group paging” i.e., pages a group of MTDs instead of a single MTD
  - Normally only up to 16 MTDs can be paged by a message → group paging overcomes this



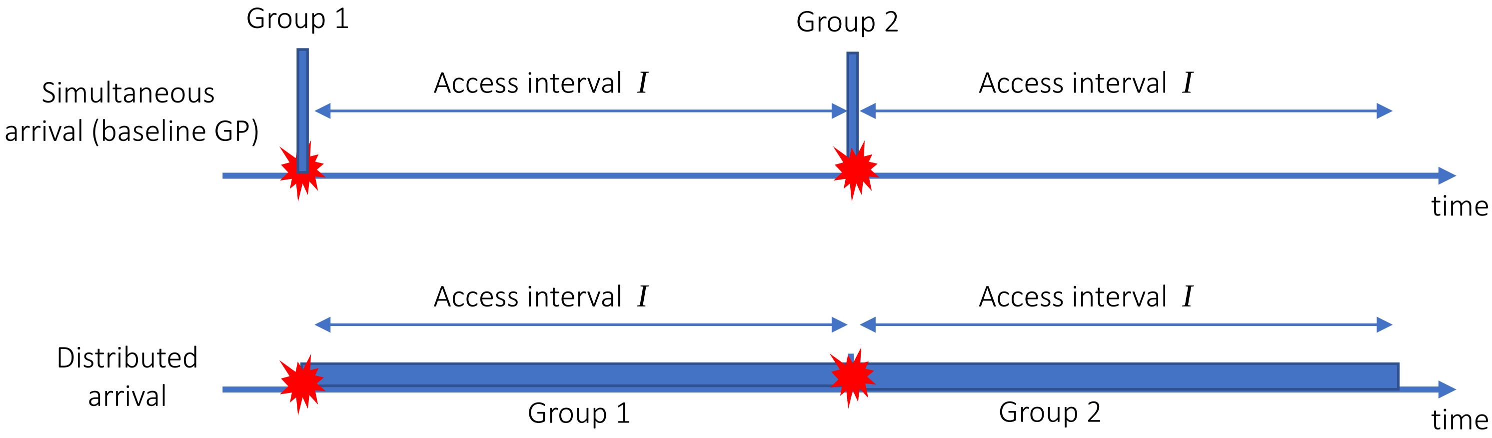
paging does not mean contention-free access
there can be contentions between paged devices

Recent results

- Pull-based schemes: Group paging (GP)
  - ➢ In general, GP schemes allocate an "access interval" for devices in a paged group to transmit in

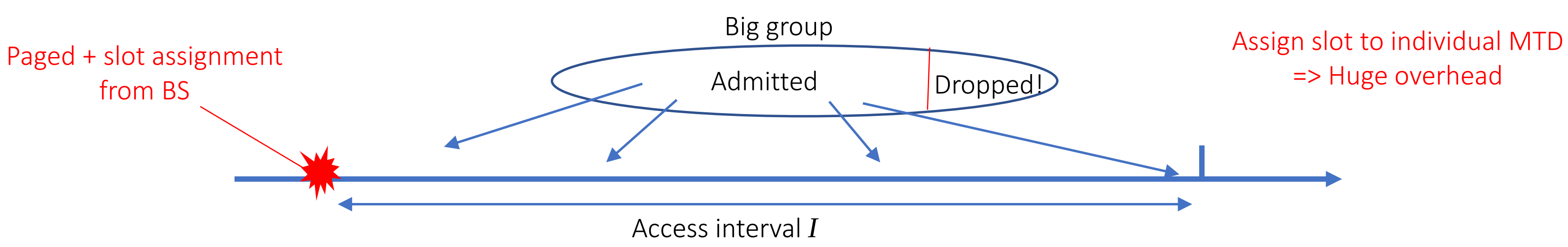


  - ➢ Most works try to distribute MTDs of a group over I i.e., pre-backoff, according to some criteria



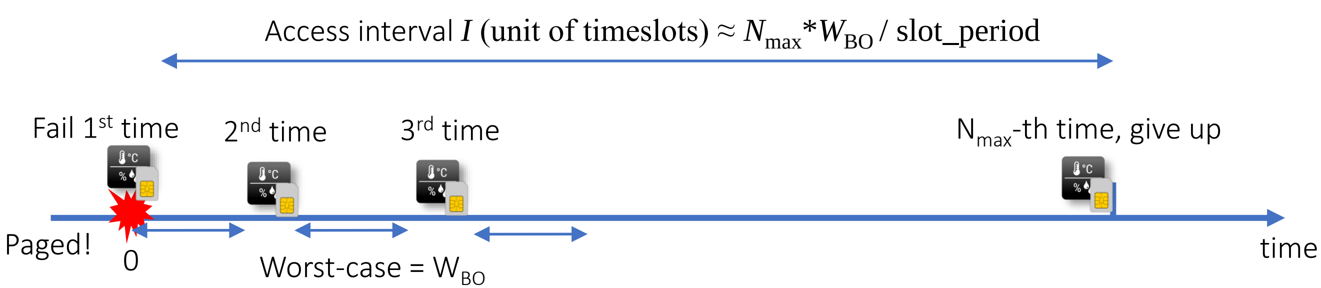
Recent results

- Pull-based schemes: Group paging (cont.)
  - [15] assigns preamble transmission slots in the interval to individual MTDs. With big group, only a portion is admitted into the interval to ensure success prob.
    - How does BS determine which MTD transmits in which slot or which one is not admitted?
    - Authors assume that each device has a minimum access success probability (ASP) requirement
    - The assignment is designed to maximize total ASP of admitted MTDs, constraint to the condition that ASP of admitted MTD must also be satisfied i.e., optimization problem



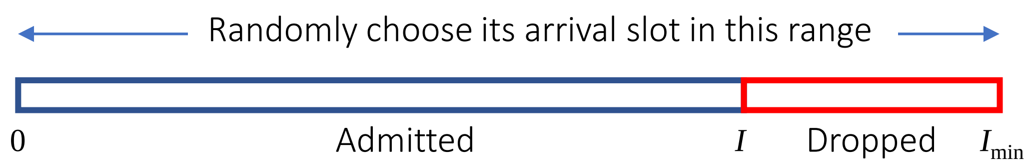
Recent results

- Pull-based schemes: Group paging (cont.)
  - [16] uses mathematic and simulation to find the num. of optimal arrivals per slot $M_{arv} = f$(num. of preambles R, num. of acknowledgeable MTDs via Msg. 2)
    - Authors assume that access interval I is bounded by the time for a worst-case MTD to consecutively fail N_{max} times and give up
    - N_g devices per group, M_{arv} per slot $\rightarrow$ we need $I_{min} = N_g / M_{arv}$ slots, but this may exceed I



Recent results

- Pull-based schemes: Group paging (cont.)
  - [16] (cont.)
    - An MTD randomly chooses an arrival slot from $[0 \longrightarrow I_{\min}]$. If its slot falls outside $[0 \longrightarrow I_{\min}]$, then the MTD is dropped. Otherwise it transmits in the chosen (valid) slot
    - Although there may be dropped MTDs, the condition M_{arv} arrivals per slot is satisfied



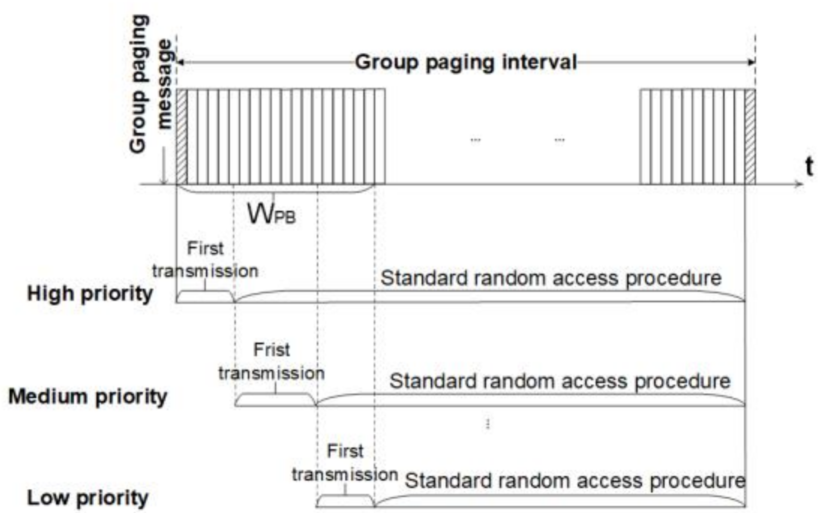
Recent results

- Pull-based schemes: Group paging (cont.)
  - [17] assumes that the MTDs of a group are in connected mode but not synchronized (case e., slide 7)
    - In this case, the UE still has its C-RNTI (cell-specific ID)
    - BS reserves a continuous range of C-RNTI for MTDs and sequentially assigns a C-RNTI in the range to MTDs of a group
    - The BS uses a rule to map an MTD's C-RNTI into (slot,preamble) combination
    - Since C-RNTI of different MTDs are unique, their assigned slots and preambles are also unique → contention-free access

This idea was copied in a 2018 ICC's paper

Recent results

- Pull-based schemes: Group paging (cont.)
  - [18] uses pre-backoff to spread the device over a sub-interval of I, but concern about device classes differentiation
    - Pre-backoff window for MTDs of higher priority occur sooner
    - Once RA procedure is initiated, there is no differentiation for an MTD
    - The pre-backoff window size for a group is set according to target access success probability in slots during that window



Recent results

- Pull-based schemes: Group paging (cont.)
  - [19] concerns with how to change priority of the classes on-the-fly
    - A class is assigned multiple IDs
    - The portion of preambles available to a class is inversely proportional to the number of matched IDs with the paging message
    - IDs are assigned to a class in a binary mapping scheme

	ID_1	ID_2	ID_3	ID_4	ID_5
Class 1	✓		✓		✓
Class 2		✓	✓		
Class 3				✓	✓
	001	010	011	100	101

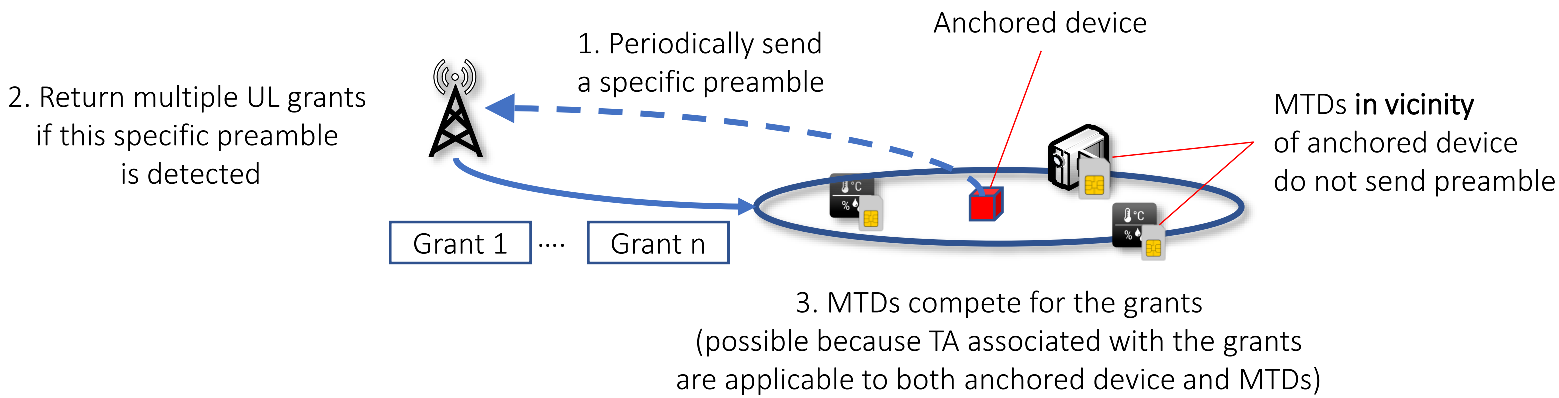
Ex: If paging msg. contains ID1, ID3, ID4, then
Class 1 can use $R/2$ preambles
Class 2 can use R preambles
Class 3 can use R preambles

Recent results

- Another way to classify solutions is to based on the MTDs' operation mode (from BS's viewpoint)
  - Non group-based: solutions assuming that MTDs always act individually
    - Most push-based schemes are non group-based
  - Group-based: solutions assuming that MTDs always act as groups
    - Often assume that there are devices acting as "group coordinator" (GC)
    - Can't really be further classified into push-pull
  - Hybrid: solutions assuming that the MTDs act as groups during certain phase and individually otherwise
    - Pull-based GP is a hybrid solution: MTDs act as groups only in paging message, then act individually afterwards

My current (very rough) idea

- There is a paper from 2012 [20] had a rather interesting hybrid group-based idea



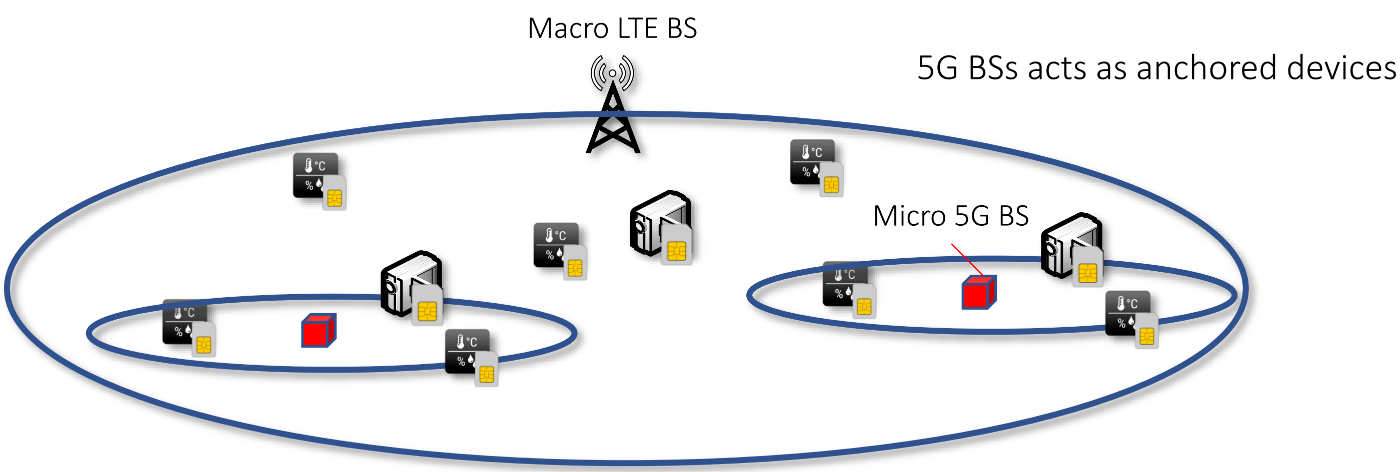
In other word, this move contentions from signaling channels to data channels (which are less congested)

However, doing this will create many unused grants under light load

Also, since the number of UL grants per RAR is limited, the multi-grants-per-single-preamble trick may not work well

My current (very rough) idea

- My current idea: In a 5G micro-macro cells setup



1. Normally, all MTDs access direct to the LTE BS (contention is resolved using DQ + estimation)

2. When the LTE BS detects congestion, it issue an indicator

3. Upon realizing the congestion indicator:
- 5G BSs start sending their dedicated preambles to the BS in each slot
- MTDs in 5G BSs' coverage stop sending preamble
- Similar operation to [20] (multiple-grants-per-dedicated preamble) → May not work well
- Contentions are resolve using (?)

4. When the LTE BS estimate that congestion is over, it issue another indicator
- Upon realizing the indicator, everything switches back to normal

References

[1] 2018.TVT.Efficient Random Access Channel Evaluation and Load Estimation in LTE-A With Massive MTC

[2] 2015.ICC.Congestion Control for Bursty M2M Traffic in LTE Networks

[3] 2017.TVT.Recursive Pseudo-Bayesian Access Class Barring for M2M Communications in LTE Systems

[4] 2016.TVT.D-ACB Adaptive Congestion Control Algorithm for Bursty M2M Traffic in LTE Networks

[5] 2018.Med-Hoc-Net.Dynamic access class barring parameter tuning in LTE-A networks with massive M2M traffic

[6] 2016.VTC.Distribution Reshaping for Massive Access Control in Cellular Networks

[7] 2018.IET.A novel random access scheme for stationary machine-type communication devices

[8] 2018.TVT.Collision-Aware Resource Access Scheme for LTE-Based Machine-to-Machine Communications

[9] 2016.GC-WS.DERA Augmented Random Access for Cellular Networks with Dense H2H-MTC Mixed Traffic

[10] 2016.TVT.Estimation and Adaptation for Bursty LTE Random Access

References

[11] 2018.Access.An Efficient Contention Resolution Scheme for Massive IoT Devices in Random Access to LTE-A Networks

[12] 2015.ICC.A Random Channel Access Scheme for Massive Machine Devices in LTE Cellular Networks

[13] 2017.TVT.Efficient Random-Access Scheme for Massive Connectivity in 3GPP Low-Cost Machine-Type Communications

[14] 2017.PIMRC.r-Hint A message-efficient random access response for mMTC in 5G networks

[15] 2018.TCOM.Differentiated Service-Aware Group Paging for Massive Machine-Type Communication

[16] 2016.JSAC.Group Paging-Based Energy Saving for Massive MTC Accesses in LTE and Beyond Networks

[17] 2014.ICC.On improving the group paging method for machine-type-communications

[18] 2017.GC-WS.Pre-Backoff Based Random Access with Priority for 5G Machine-Type Communication

[19] 2017.GC.Hybrid Group Paging for Massive Machine-Type Communications in LTE Networks

[20] 2012.ICC.A Group-based Communication Scheme Based on the Location Information of MTC Devices in Cellular Networks

Thank you for listening