

Hard Latency-Constraints for High-Throughput Random Access: SICQTA

H. Murat Gürsu, Fuqi Guan, Wolfgang Kellerer

Chair of Communication Networks, Technical University of Munich, Munich, Germany

E-mail: {murat.guersu, fuqi.guan, wolfgang.kellerer}@tum.de

Abstract—Enabling closed control loops via wireless communication has attracted a lot of interest recently and is investigated under the name cyber-physical systems. Under cyber-physical systems one challenging scenario is multiple loops sharing a wireless medium, and the age of the control information has to be minimized without sacrificing reliability to guarantee the control stability. The number of transmitting devices depends on the control parameters thus, it is stochastic. Wireless uplink resource allocation given low latency constraints for unknown number of devices is a hard problem. For this problem, random access is the most prominent way to minimize latency, but reliability is sacrificed. However, as reliability is also critical for such applications, improved random access algorithms with hard latency guarantees are needed. Currently available random access algorithms with hard latency guarantees have low throughput and some of them are limited to low number of active devices. In this work, we provide a high-throughput random access algorithm with hard latency-constraints (SICQTA) that scales to any number of active devices. This algorithm, making use of feedback, has a varying throughput between 0.69 and 1 depending on the number of devices, which is unprecedented in the state of the art up to our best knowledge.

I. INTRODUCTION & BACKGROUND

One typical problem with latency-reliability constraints is uplink resource allocation for cyber physical systems [1]. In this problem, multiple control loops share the wireless medium. Each loop is composed of a controller, actuator and a sensor. The controller is located at a central entity while actuator and sensor are both located at the device. The closed control loops outputs actuation decisions in the controller from the input of the sensing information. The devices transmit the sensing information through uplink communication and get actuation decision as downlink communication.

The downlink is broadcast to all actuators without the need of coordination. However, depending on the state of the control loop, only some of the sensors transmit state information through uplink communication. As the transmission depend on the state of the control, the number of devices transmitting at a certain time is unknown. Thus, we have M active sensors at a certain time out of N total sensors which have to be allocated resources to optimize the control performance. This problem is previously investigated with LTE scheduling consisting of a scenario with multiple inverted pendulums in [2]. However, the solution assumes the information of device activity to overcome the over-dimensioning of scheduling. This information is not available in reality and the inefficiency to obtain this information has actually called for a new design of

LTE uplink resource allocation mechanism called as grant-free [3], reusing the state of the art in random access area.

Grant-free focuses on a scenario where devices transmit a single packet or multiple replicas to achieve the latency-reliability constraints. This requires over-dimensioning of resources to fulfill tight reliability constraints as it lacks the information that is the number of active devices [4]. As a solution to over-dimensioning, successive interference cancellation (SIC) is integrated to the random access schemes.

SIC enables recovery of overlapping packets through signal processing. This has increased the throughput of random access algorithms from 0.5 packets per slot up to 1 packet per slot with asymptotic number of devices, reaching the efficiency of scheduling based solutions. The trade-off is the decoding complexity. Through edge-cloud processing and distributed computing, complexity is expected to be dealt with for radio access algorithms [5].

Successive interference cancellation is initially explored for tree algorithms in [6]. Through that work the throughput for tree algorithms is increased to 0.69 from 0.35. In [6] the clean packet for cancellation is guaranteed with feedback, forcing devices to split from each other. However, too much structure is inefficient and in [7] it is shown that the same structure can be built through random decisions. The random decisions are shaped with a degree distribution tailored to the number of devices. It is shown that the algorithm reaches a throughput of 1 in the asymptotic region when M goes to infinity.

Another work [8] adapts that work to a frameless structure where the degree distribution is replaced with setting a Binomial probability to transmit at each slot. Compared to framed structure the results show that, [8] has a better performance in the non-asymptotic region. However, neither of these algorithms can provide a hard guarantee on the latency. Also both of them are susceptible to varying number of active devices. The hard guarantees can be provided via setting the decisions uniquely for each device.

This problem is initially investigated by Massey under the name "protocol sequences" for de-synchronized devices in [9]. These algorithms are too pessimistic to be applied to tight latency constraints as the offset between devices is the main issue there and it is not the main problem anymore thanks to improvement in hardware design. Recently, the unique decisions for each device for hard guarantees is investigated in [10] under the name "access codes", where each device

transmits packets with respect to a unique code. The design of these codes is of combinatorial complexity. The results are limited, as we detail on later parts of the paper. Moreover, the use of feedback is neglected in this work.

Uniqueness of the access decisions can be guaranteed through feedback to overcome the complexity of proposed protocol. Using addresses for such limitation is initially proposed by [11] and adapted for RFID tags with Query Tree Algorithms in [12]. However, this algorithm lacks behind in throughput compared to SIC-capable algorithms. The idea to use Interference Cancellation for Query Tree Algorithms is introduced in [13]. However, the explanation of the algorithm in [13] is unclear. The throughput they have shown is capped to 0.69 which have already been shown by [6] for TA with SIC capabilities. Hard guarantees for performance are not investigated and the difference to [6] is unclear.

In our work we propose a novel Successive Interference Cancellation for Query Tree Algorithm, SICQTA. We provide analytical hard upper and lower bounds to latency and compare it with simulations to show the validity. It is shown that the algorithm easily extends to any number of active devices unlike *access codes*, and it provides a higher throughput compared to previous SIC based works. On top of that, hard latency guarantees make it a suitable candidate as a solution of the uplink resource allocation problem with unknown number of active devices.

Our paper is organized as follows: In Sec. II we explain the scenario and provide the problem formulation for reliable access with latency constraints. In Sec. III we introduce shortly the Query Tree Algorithm and Successive Interference Cancellation Query Tree Algorithm. In Sec. IV the latency bounds are given and we compare our solution to the *access codes* while comparing the bounds with simulations. Further discussions are given in Sec. V. Finally, the paper is concluded with possible extensions of future work in Sec. VI.

II. SCENARIO & PROBLEM

We consider a star topology where the central entity is called the gateway and leaf entities of the star are called devices. We consider an uplink scenario where only devices transmit a packet to the gateway. There are N devices attached to the gateway. Considered resources in the system are slots of a single channel with a TDM scheme.

Two different channel models are considered with and without SIC. First one is a collision channel model where perfect reception is assumed. If there is no contention, there is no loss of packets [14]. Second one is for SIC scenario, we assume perfect cancellation is possible if clean packets are received. These assumptions are common in MAC layer research to focus on a layer 2 based solution. Impact of more practical channel models are discussed in Sec. V. Each device is synchronized perfectly to the slots defined by the TDM structure. The devices are randomly and sporadically activated and the number of active devices at any slot is M , such that $M \leq N$. The devices have a homogeneous radio latency constraint L and reliability constraint R . We investigate the multiple access problem of

maximizing throughput that we abstract as maximizing number of successfully used slots.¹

We define a frame structure consisting of d subsequent slots. We investigate the problem of designing codes \mathbf{c} that represents the binary access decision of a device. The code \mathbf{c} is of size d , i.e., $\mathbf{c} = \{c_1, c_2, \dots, c_d\}$ where $c_i \in \{0, 1\} \forall i \in \{1, \dots, d\}$. The device that has the code $c_i = 1$ will transmit its packet at slot i .

The codebook \mathbf{C} is a collection of all codes and is a matrix with d columns and N rows where each row represents a unique code for each device. An example is as follows:

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

with $N = 3$ devices and a frame size $d = 4$. Each device is sporadically active and the activity of all devices is represented with a vector \mathbf{n} with N elements, i.e., $n_j \in \{0, 1\}$ where $n_j = 1$ represents that the device j is active. We assume that codebook \mathbf{C} is ordered such that code of device j is in the j^{th} row of \mathbf{C} . This assumption allows us to define a frame outcome \mathbf{f} as in,

$$\mathbf{n} \cdot \mathbf{C} = \mathbf{f}. \quad (1)$$

The frame outcome \mathbf{f} represents the number of packets at each slot of the frame. However, receiver is unaware of this information such that \mathbf{f} should be converted to MAC layer success outcome \mathbf{s} . An example for collision channel would be,

$$f_i \begin{cases} = 1 & s_i = 1 \\ \text{o.w.} & s_i = 0. \end{cases} \quad (2)$$

Using the previous definitions we can define an optimization problem for codebook design.

Given frame size d and number of devices N , maximize the total success per frame through the codebook design \mathbf{C} :

$$\arg \max_{\mathbf{C}} \|\mathbf{s}\|^2, \quad (3)$$

$$\text{s.t. } \mathbf{n} \in \mathcal{N}, \quad (4)$$

$$d \leq L, \quad (5)$$

$$\mathbb{E} [\|\mathbf{s}\|^2] \geq \mathbb{R} \mathbb{E} [\|\mathbf{n}\|^2]. \quad (6)$$

where \mathcal{N} is the set of all possible activation combinations of N devices, L and R are the latency and reliability constraint respectively. The $\|\cdot\|^2$ operation is the autocorrelation operation that also gives the summation of binary vectors. This is naturally a combinatorial problem and hard to solve, as \mathbf{n} can take any value. We can write $\|\mathbf{n}\|^2 = M$, where M is the number of simultaneously active devices per frame.

The problem definition is shared here for formalism. In the following part of the paper we show that SICQTA solves this problem with a distributed algorithm that is guided via a central feedback. Optimality of the algorithm is not proven is an open issue for future work.

¹For simplicity we assume that the constraint can be expressed in terms of slots. The reliability constraint here is the radio layer reliability, that can be input to the end to end reliability model.

Algorithm 1 Query Tree Algorithm

```
1: procedure GENERATE QUERY
2:    $Q \leftarrow \{‘0’, ‘1’\}$   $\triangleright$  Initialize Q list with ‘0’ and ‘1’
3:   while  $Q$  is not empty do
4:      $q \leftarrow Q[0]$   $\triangleright q$  is the first element of  $Q$ 
5:     Transmit query at the beginning of time-slot
6:     Save received packets as  $\mathbf{r}$ 
7:      $f \leftarrow |\mathbf{r}|$   $\triangleright$  Number of received packets
8:      $Q.\text{pop}$   $\triangleright$  Delete  $Q[0]$ 
9:     if  $f = 0$  or  $f = 1$  then  $\triangleright$  Idle or success slot
10:      pass
11:     else if  $f > 1$  then  $\triangleright$  Collision slot
12:        $Q.\text{append}(‘q0’, ‘q1’)$ 
End
```

III. ALGORITHMS WITH FEEDBACK

A. Query Tree Algorithm

First, we shortly introduce the Contention Tree algorithm. At the start of the algorithm, in binary contention tree algorithm [11] any active device sets $c_1 = 1$ and transmit. If more than 1 device is active, the gateway sends a feedback to devices, informing that a collision has happened, and all the active devices do a uniform random selection whether to set $c_2 = 1$ and $c_3 = 0$ or vice-versa. The devices that have set $c_2 = 1$ transmit at slot 2. If again a collision is reported, only those that have transmitted at slot 2 do a random uniform selection for c_3 and c_4 . Meanwhile, the devices that have previously set $c_3 = 1$, change the values via setting $c_3 = 0$ and $c_4 = 1$. Thus, postponing their transmission. The process goes on until all devices have transmitted successfully. Even though this process stochastically guarantees that all access codes are unique, the distribution, representing the latency of devices, has a long tail and is not efficient for high reliability constraints.

To overcome this issue, Query Tree Algorithm (QTA) is suggested in [12]. In QTA every device has a unique id formed of u bits. This limits the total number of devices attached to the gateway to $N = 2^u$. In QTA, queries are used instead of feedback but the overhead is the same. In QTA devices are queried with respect to their id bits. The queries start with an empty query. A single bit is appended to the list of queries after each collision, starting from the left-most bit. Each new collision append a new bit. As each device has a unique id, this guarantees that two devices have a unique access decision in worst-case after u transmissions (if all previous $u - 1$ bits are the same for two devices). The gateway implementation of QTA is given in Alg. 1, where the device implementation is only answering to the queries matching its id.

A detailed example is given for $M = 4$ in Fig. 1a. We have named the 4 devices as {A,B,C,D} with ids {000,001,100,101} respectively. Each circle denotes a slot in the tree. The time-wise progression of the tree is given with slots above the tree. The id size, u is fixed to 3.

In the first slot, 4 devices transmit at the same time and collide. Next slot, the address $0xx$ is queried. Only, A and B transmit. It is again a collision. On the following slot, the

Algorithm 2 SICQTA

```
1: procedure GENERATE QUERY
2:    $Q \leftarrow []$ ,  $q \leftarrow ‘0’$ ,  $k \leftarrow ‘0’$   $\triangleright$  Initialization
3:   while  $k - 1 \neq |Q|$  do  $\triangleright$  End condition
4:     Transmit query at the beginning of time-slot
5:     Save received packets as  $\mathbf{r}$ 
6:      $q_b \leftarrow [q_1 \dots q_{n-1} \bar{q}_n]$   $\triangleright$  Invert last bit of  $q$ 
7:      $f \leftarrow |\mathbf{r}|$   $\triangleright$  Number of received packets
8:     if  $f = 0$  then  $\triangleright$  Idle slot
9:        $q \leftarrow ‘q_b0’$   $\triangleright$  Skipping collision
10:    else
11:       $Q.\text{append}(q_b)$ 
12:      if  $f > 1$  then  $\triangleright$  Collision slot
13:         $q \leftarrow ‘q0’$ 
14:      else  $\triangleright$  Cancel clean packet and skip.
15:         $q \leftarrow Q[-k]+‘0’$ 
16:         $Q \leftarrow [Q[0], \dots, Q[-k-1]]$ 
17:       $\triangleright$  Skip most recent  $k - 1$  queries thanks to SIC,  $k \geq 1$ .
End
```

query for address $1xx$ is also a collision so the algorithm moves one level down. The address $00x$ is queried and both devices transmit. The query for $01x$ results in an idle slot. Queries for address 001 and 000 is done on slot 6 and 7, respectively and both are successes. The algorithm is completed after the process is repeated for right branch.

B. Query Tree Algorithm with SIC (SICQTA)

SIC allows recovery of packets from a slot where a collision is observed. If for instance device A and B have transmitted a packet in slot 1, due to collision channel model, the outcome "A+B", is treated as a collision and slot is considered wasted. However, if device B has transmitted its packet in slot 2, the SIC model let us subtract B from "A+B" and enables recovery of A from slot 1. Instead of breadth first, the SICQTA goes depth-first. After the initial success, it checks if it can cancel the clean packet from previous collisions. If the packet is successfully cancelled then the algorithm skips the direct siblings of those slots. The algorithmic description of SICQTA is given in Alg. 2.²

A detailed example for the worst-case behavior of SICQTA is given in Fig. 1b for $M = 4$. In the first slot, all the devices are queried and it is a collision. On the second and third slot, addresses $0xx$ and $00x$ are queried, respectively. Both are collisions. The following slot, 000 is queried and it is a success. 001 is not queried, as the gateway recovered the packet from slot 2 and 3. This results in $k = 3$ as 2 slots are successfully recovered and this slot is a success. Ids in query list Q : 001 and $01x$ is not queried and skipped. Thus, $10x$ is queried, that results in a collision. Following, 100 is queried and is a success. The gateway recovered D from slot 5 and the algorithm is terminated.

²Open source Python implementation of the algorithms is available at: <https://github.com/tum-lkn/sicqta>

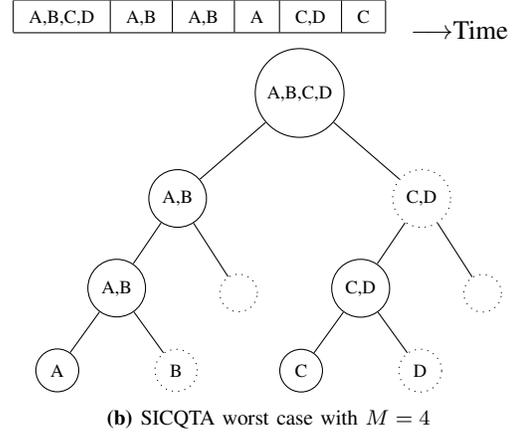
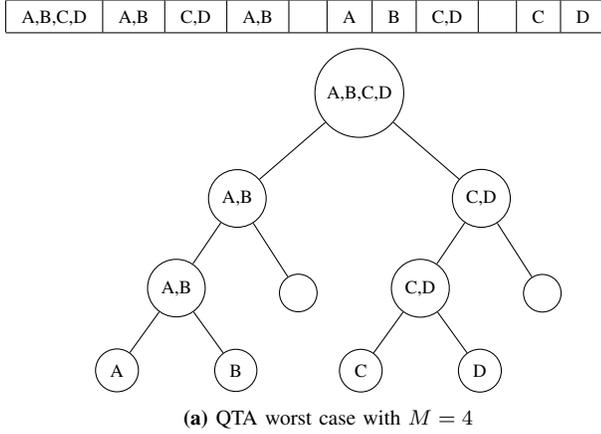


Fig. 1: Worst-case example for Query Tree Algorithms with and without SIC with $M = 4$. $u = 3$ is set such that maximum number of devices is $N = 2^u = 8$.

IV. ANALYSIS & EVALUATION

In this section we will evaluate the latency of QTA and SICQTA and give bounds to its performance. We will also compare the performance of our work and [10] as we share the same problem definition. Finally, mean delay is compared with state of the art in tree algorithms to show that the stability region is extended.

A. QTA

An upper-bound for latency y of QTA is given in [15]:

$$y \leq M(u + 2 - \log M), \quad (7)$$

where M is the number of active devices. This is a tight bound for $M \ll N$ where with increasing M it has a slack. Using the tree structure we can provide a tighter upper-bound for latency y as,

$$y \leq \left\lfloor \frac{M}{2} \right\rfloor 2 \left(u + 1 - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) - 1 \quad (8)$$

Similarly, the tree structure can be used to provide a lower-bound of latency as:

$$y \geq 2M - 1. \quad (9)$$

The proofs are given in App. A and B, respectively.

We explain why the example in Fig. 1a is the worst-case of a QTA with $M = 4$ also shedding light on the proof of the bounds. Four devices are separated into 2 groups of 2 as close as possible to the root of the tree, so they cover as much as non-overlapping slots as possible. Following, devices have repeated the same collision, until the last level of the tree. We observe that for this scenario the total number of slots is $y = 11$. Using Eq. (7) we get 13. This shows that the bound is valid and tight for this setting.

B. SICQTA

Intuitively, the efficiency of the [6] comes from the possibility to skip some slots in the tree. As it is shown in [6], the throughput of BTA is doubled. However, the throughput is the expected number of slots and this result cannot be directly translated to worst-case latency of SICQTA from

QTA. We have to adapt the Eq. (7) for SICQTA using the skipping capability of SIC. The total number of skipped slots S compared to worst-case of QTA, given M active devices can be written as,

$$S = \left\lfloor \frac{M}{2} \right\rfloor \left(u - 1 - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) + \sum_{i=1}^{\lfloor \log_2 M \rfloor} \left\lfloor \frac{M}{2^i} \right\rfloor. \quad (10)$$

The proof is given in C.

We can use this finding to provide an upper-bound for latency of SICQTA using Eq. (8) and removing the skipped slots,

$$y \leq \left\lfloor \frac{M}{2} \right\rfloor (u + 4 - \lfloor \log_2 M \rfloor) - 1 - \sum_{i=1}^{\lfloor \log_2 M \rfloor} \left\lfloor \frac{M}{2^i} \right\rfloor. \quad (11)$$

Intuitively, the algorithm needs at least M slots for M active devices and a lower-bound for latency of SICQTA can be given as $y \geq M$.

This is given without any proof, as in best-case no repetition occurs such that every slot is recoverable from another.

The upper-bound for latency can be used for the throughput calculation of the SICQTA. If number of active devices is the same as the number of total devices, i.e., $M = N = 2^u$. Then we expect SICQTA to have a throughput of 1, as each slot in the tree should be different from one another.

Eq. (11) is a relaxed bound, but it becomes tight for integer values of $\log_2 M$. Plugging in $M = 2^u$ we get,

$$2^u \leq y \leq 2^{u+1} - 1 - 2^{u+1} + 2^u + 1 = 2^u. \quad (12)$$

Thus, we have a throughput of 1 as expected. The proof is given in App. D.

We can check the bound via the example in Fig. 1b. We see that in total 6 slots are used for SICQTA in the example. Using Eq. (11) we get 6 showing that the bound is valid and tight for this scenario.

In Tab. I we have compared the number of devices N supported by CAC-SIC with SICQTA. The number of active devices are fixed to $M = 3$ for CAC, because these are the only

Constraint	L= 4	L= 5	L= 6	L= 7
CAC-SIC [10]	7	11	—	—
SICQTA $M = 3$	8	16	32	64
SICQTA $M = 4$	4	8	8	16

TABLE I: Number of devices supported by CAC-SIC for fixed number of active devices $M = 3$, with varying latency constraint, compared to SICQTA.

available results in [10]. For SICQTA, we see that with relaxed delay constraint the number of devices supported increases exponentially. And even though the results are similar for low latency constraints, the difference increases with increasing L . Also the results for SICQTA is easily extensible to other M values, while an exhaustive search is required to build codes for CAC-SIC. On the other hand effect of feedback is neglected in this analysis.

C. Simulations

We have done Monte Carlo experiments on a python based discrete event simulator 10^6 samples for each experiment varying the number of active devices.

In Fig. 2 we have plotted the bounds versus simulation for SICQTA. x-axis depicts the varying active number of devices M and the y-axis presents the latency. We have set $u = 6$ so implicitly $N = 64$, and we have varied the number of active devices M . We see that with 10^4 iterations for each data point in simulations the bounds are never surpassed and the difference between the lower and the upper bound is quite low.

As we deal with worst-case latency, this is the latency of the last device. In Fig. 2b we have evaluated the throughput with varying active number of devices M . Mean throughput is almost always above 0.8 while the tail is also quite constrained, especially with increasing M .

In Fig. 3 we extend the delay vs throughput comparison in [6] with SICQTA. In this simulation scenario continuous arrivals are considered. If a device gets a packet to transmit while there is an on-going resolution, the device is queued until the end of that resolution, reflecting the setting in [6]. We see that SICQTA enables a new throughput region that extends to throughput of 0.93 with $u = 4$. Also with $u = 6$ the throughput with stable latency is around 0.86. Of course SICQTA becomes similar to SICTA with increasing u value. This is logical as SICTA can be considered as a special setting of SICQTA with $u = \infty$. Here, it is shown that with $u = 10$ the behavior is almost the same as SICTA. It is worth mentioning that the average resolution time is increased as we see a shift on the y-axis compared to SICTA. We have also simulated higher values of u , i.e., $u = 16$ and did not observe any difference so they are not plotted here to avoid clutter. For decreasing u the throughput is expected to increase further reaching 1.

V. DISCUSSIONS

One important point for SICQTA compared to QTA is that the knowledge of M does not improve the upper-bound of latency. The knowledge of M would be used in this case to skip to level $\lceil \log_2 M \rceil$. However, in the worst-case all collisions

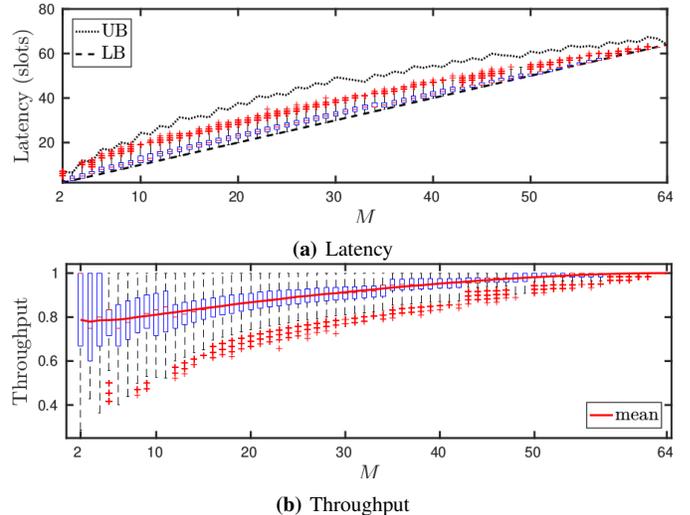


Fig. 2: Excessive simulations show the validity of the bounds. The maximum number of levels is set to $u = 6$, $N = 64$ and M is varied (x-axis).

happening before this level consist of different devices, and under a SIC framework, they can all be recovered from each other to obtain useful slots. So the number of skipped slots with knowledge of M would be equal to those skipped due to SIC. However, application of knowledge of M to QTA can improve the worst-case performance and bring it close to SICQTA.

We have compared the feedback based algorithms to non-feedback based algorithms here. However, we assumed that the feedback is instantaneous and costless. In reality that is not the case. The latency incurred due to transmission and reception may even involve hardware delays such as switching from transmit to receive and vice-versa. We leave this open for future work.

We are also working on prototyping this algorithm through IEEE 802.15.4 capable sensors and SDRs. One observation we have is that depending on the quality of the sensor device, the phase noise accumulates through successive interference cancellation and this makes collisions of 6 packets, a wasted slot as cancellation fails due to accumulated phase noise. Algorithmic solutions, such as starting the queries from level $u - 3^3$, should be considered to overcome such hardware constraints. Curious reader can refer to [16] for a theoretical model that incorporates variances in the hardware to the SIC capacity and to [17] for practical characterization of causes for hardware variances.

VI. CONCLUSION

In this work we have evaluated the problem of uplink resource allocation for unknown number of active devices. We believe that this problem represents the important uplink resource allocation problem for multiple control loops sharing the same wireless network. As a solution we present the algorithm Successive Interference Cancellation Query Tree Algorithm (SICQTA). The advantage of the algorithm compared

³This will cap the maximum number of collided devices to 8.

to previous algorithms is the high-throughput performance and the hard latency guarantees. The bounds for the performance are proven analytically and further validated with simulations.

Future work can investigate relaxing the assumptions made for easy investigation of the protocol. Firstly, the feedback is assumed instantaneous and costless, accumulation of feedback messages should be considered to decrease this bottleneck as much as possible. Secondly, we assumed that SIC works perfectly. However, due to accumulated phase noise some collisions cannot be recovered via SIC and indeed result in wasted slots. This should be evaluated and incorporated into the protocol design. Thirdly, even though it is intuitive that decreasing latency and increasing reliability helps for the cyber-physical systems, an integrated evaluation of control and communication should be done to provide concrete results.

APPENDIX

A. Proof for upper-bound for latency of QTA

The worst-case for QTA is illustrated in Fig. 4. An intuitive explanation is as follows: A device can re-transmit at maximum u times in the worst-case as that is the size of addresses and every device has a unique address. In this case the device is successful with the u^{th} transmission and it has experienced $u - 1$ collisions. In order to have a collision we need at least 2 devices, and at the worst-case all devices are grouped into two, thus $\lfloor \frac{M}{2} \rfloor$ groups. Each group collides separately for $u - 1$ times, where there will be idles on the unexplored slots so $2 \cdot (u - 1)$ slots, followed with 2 transmissions for success of each device, we get

$$y \leq \left\lfloor \frac{M}{2} \right\rfloor 2 \cdot (u - 1 + 1) \quad (13)$$

slot uses in total. We take into account, the activity of the groups of two only after the level $\lfloor \log_2 \frac{M}{2} \rfloor$. As the initial levels have a lot of overlap, we can remove these levels and consider them separately as

$$y \leq \left\lfloor \frac{M}{2} \right\rfloor 2 \cdot \left(u - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) + \mathbf{R}, \quad (14)$$

where \mathbf{R} represents the overlapping slots. The number of overlapping slots can be calculating by summing the total number of slots up to level $\log_2 \frac{M}{2} \geq \lfloor \log_2 \frac{M}{2} \rfloor$ of the tree. We can calculate the total number of nodes in this upper part of the tree as,

$$\mathbf{R} \leq 2^{m+1} - 1 = 2^{\log_2 \frac{M}{2} + 1} - 1 = M - 1. \quad (15)$$

Plugging this in Eq. (14) we get,

$$y \leq \left\lfloor \frac{M}{2} \right\rfloor 2 \cdot \left(u - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) + M - 1, \quad (16)$$

$$\leq \left\lfloor \frac{M}{2} \right\rfloor 2 \cdot \left(u + 1 - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) - 1. \quad (17)$$

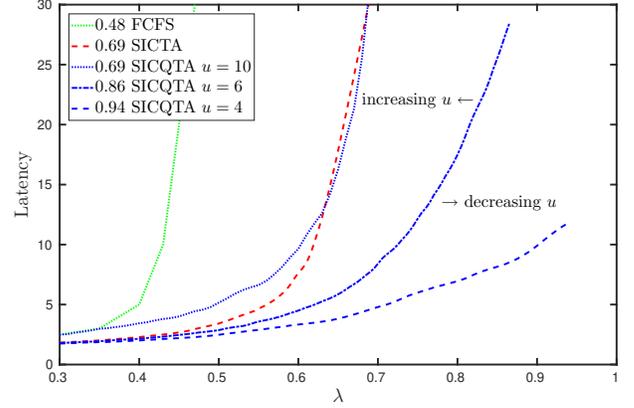


Fig. 3: Delay vs throughput of feedback based random access algorithms.

B. Proof for lower bound for latency of QTA

The best-case in the tree with M devices, is that they are organized as a triangle, guaranteeing they are as close as possible to the root. So the level of the successes are almost the same. However, the level of the devices can be the same only if $\log_2 M$ is an integer. If it is not an integer, the best-case would be some of the devices are successful at level $l_1 = \lceil \log_2 M \rceil$ and the others are at $l_2 = \lfloor \log_2 M \rfloor$. In order to have a complete triangle we would need that devices at level l_2 would each have 2 children at l_1 . So the number of slots at l_1 is equal to the sum of number of devices at l_1 plus twice the number of devices at l_2 . The number of slots at a level can also be written as 2^l so we can write,

$$2^{l_1} = M_{l_1} + 2 \cdot M_{l_2} \quad (18)$$

where M_{l_1} and M_{l_2} is the number of devices successful in l_1 and l_2 respectively. We know that the total number of devices is $M = M_{l_1} + M_{l_2}$. so we can re-write Eq. (18) as

$$M_{l_1} = 2 \cdot M - 2^{l_1}. \quad (19)$$

If we do not consider the level l_1 , the tree is a full triangle up to level l_2 . We can calculate the total number of slots in the tree for the best-case y_{LB} through calculating the number of slots for the full tree up to l_2 and adding M_{l_1}

$$y_{LB} = 2^{l_2+1} - 1 + M_{l_1}. \quad (20)$$

By definition of flooring and ceiling operation $l_2 + 1 = l_1$ if $\log_2 M$ is not an integer. And we can plug Eq. (19) in to get,

$$y_{LB} = 2^{l_1} - 1 + 2 \cdot M - 2^{l_1} = 2 \cdot M - 1. \quad (21)$$

When $\log_2 M$ is an integer the lower-bound is directly given with $2^{\log_2 M + 1} - 1$, which is equal to the result so we do not mention it separately.

C. Proof for number of skipped slots \mathbf{S}

The skipping in SICQTA consists of two different parts. First part is skipping the idles S_I and second part is skipping the canceled slots S_C . So we can write $\mathbf{S} = S_I + S_C$.

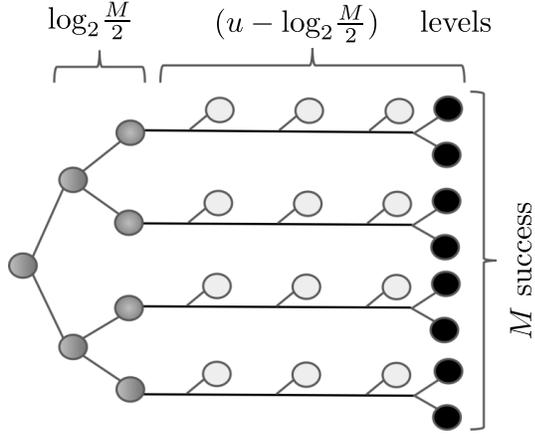


Fig. 4: The worst-case tree structure for Query Tree Algorithm.

The upper-bound for latency of QTA is derived using groups of 2 devices sticking together until the last level of the tree. At the last level they transmit separately, each as a success. The idles occur after separation from the top triangle until the end of the tree. We have $\lfloor \frac{M}{2} \rfloor$ collisions and the number of levels until the end of the tree gives us the number of skipped idle slots as

$$S_I = \left\lfloor \frac{M}{2} \right\rfloor \left(u - 1 - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right). \quad (22)$$

Thanks to SIC, after one success the other device does not have to transmit anymore, as after one success the other device can be recovered from the previous collision. Thus, at least $\frac{M}{2}$ slots are skipped for the last level of the tree.

This skipping can be applied to also formation of groups of 2. Groups of 2 are formed from groups of 4. Thus, for the first group formed out of 4 devices, the other group can be recovered from the collision, so one slot can be saved for each separation. In this step we can save $\frac{M}{4}$ slots. This logic can be extended up to $\lfloor \log_2 M \rfloor$ separations as we have a binary splitting process. This gives us,

$$S_C = \sum_{i=1}^{\lfloor \log_2 M \rfloor} \left\lfloor \frac{M}{2^i} \right\rfloor. \quad (23)$$

Finally, we can write,

$$S = \left\lfloor \frac{M}{2} \right\rfloor \left(u - 1 - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) + \sum_{i=1}^{\lfloor \log_2 M \rfloor} \left\lfloor \frac{M}{2^i} \right\rfloor. \quad (24)$$

D. Proof for number of skipped slots with $M = 2^u$

We plug in $M = 2^u$ to Eq. (23)

$$S = \left\lfloor \frac{2^u}{2} \right\rfloor \left(u - 1 - \left\lfloor \log_2 \frac{2^u}{2} \right\rfloor \right) + \sum_{i=1}^{\lfloor \log_2 2^u \rfloor} \left\lfloor \frac{2^u}{2^i} \right\rfloor = \sum_{i=1}^u 2^{u-i}. \quad (25)$$

$$S = \sum_{i=1}^u 2^{u-i} = 2^u \left(\sum_{i=0}^{u-1} 2^{-i} - 1 + 2^{-u} \right) = 2^u \left(\frac{1 - 2^{-u}}{1 - 2^{-1}} - 1 + 2^{-u} \right) = 2^{u+1} - 2^u - 1, \quad (26)$$

is what we get, as u is the number of maximum levels and is an integer we can remove the floor operation. So we can plug Eq. (26) in Eq. (11) to get,

$$y \leq \left\lfloor \frac{2^u}{2} \right\rfloor (u + 4 - \lfloor \log_2 2^u \rfloor) - 1 - 2^{u+1} + 2^u + 1. \quad (27)$$

REFERENCES

- [1] P. Tabuada, "Event-triggered real-time scheduling of stabilizing control tasks," *IEEE Transactions on Automatic Control*, vol. 52, no. 9, pp. 1680–1685, 2007.
- [2] M. Vilgelm, O. Ayan, S. Zoppi, and W. Kellerer, "Control-aware uplink resource allocation for cyber-physical systems in wireless networks," in *European Wireless 2017; 23th European Wireless Conference; Proceedings of VDE*, 2017, pp. 1–7.
- [3] "3GPP RP-181477: SID on Physical Layer Enhancements for NR URLLC; NR eURLLC L1," 2018.
- [4] M. Gürsu, B. Köprü, S. Coleri Ergen, and W. Kellerer, "Multiplicity estimating random access protocol for resource efficiency in contention based noma," in *Personal, Indoor and Mobile Radio Communications (PIMRC 18)*, 2018.
- [5] H. El-Sayed, S. Sankar, M. Prasad, D. Puthal, A. Gupta, M. Mohanty, and C.-T. Lin, "Edge of things: the big picture on the integration of edge, iot and the cloud in a distributed computing environment," *IEEE Access*, vol. 6, pp. 1706–1717, 2018.
- [6] Y. Yu and G. B. Giannakis, "Sicta: a 0.693 contention tree algorithm using successive interference cancellation," in *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, vol. 3. IEEE, 2005, pp. 1908–1916.
- [7] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted aloha," *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 477–487, 2011.
- [8] C. Stefanovic, P. Popovski, and D. Vukobratovic, "Frameless aloha protocol for wireless networks," *IEEE Communications Letters*, vol. 16, no. 12, pp. 2087–2090, 2012.
- [9] J. Massey and P. Mathys, "The collision channel without feedback," *IEEE Transactions on Information Theory*, vol. 31, no. 2, pp. 192–204, 1985.
- [10] C. Boyd, R. Vehkalahti, and O. Tirkkonen, "Interference cancelling codes for ultra-reliable random access," *International Journal of Wireless Information Networks*, vol. 25, no. 4, pp. 422–433, Dec 2018. [Online]. Available: <https://doi.org/10.1007/s10776-018-0411-6>
- [11] J. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE transactions on information theory*, vol. 25, no. 5, pp. 505–515, 1979.
- [12] J. H. Choi, D. Lee, and H. Lee, "Query tree-based reservation for efficient rfid tag anti-collision," *IEEE Communications Letters*, vol. 11, no. 1, 2007.
- [13] R. Kumar, T. F. La Porta, G. Maselli, and C. Petrioli, "Interference cancellation-based rfid tags identification," in *Proceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*. ACM, 2011, pp. 111–118.
- [14] S. Ghez, S. Verdu, and S. C. Schwartz, "Stability properties of slotted aloha with multipacket reception capability," *IEEE Transactions on Automatic Control*, vol. 33, no. 7, pp. 640–649, 1988.
- [15] C. Law, K. Lee, and K.-Y. Siu, "Efficient memoryless protocol for tag identification," in *Proceedings of the 4th international workshop on Discrete algorithms and methods for mobile computing and communications*. ACM, 2000, pp. 75–84.
- [16] S. P. Weber, J. G. Andrews, X. Yang, and G. de Veciana, "Transmission capacity of wireless ad hoc networks with successive interference cancellation," *IEEE Transactions on Information Theory*, vol. 53, no. 8, pp. 2799–2814, Aug 2007.
- [17] G. Zhou, T. He, S. Krishnamurthy, and J. A. Stankovic, "Models and solutions for radio irregularity in wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 2, no. 2, pp. 221–262, 2006.