

Random Access for Machine-to-Machine Communications: Challenges and Prospects

Lin Dai

Department of Electrical Engineering
City University of Hong Kong

lindai@cityu.edu.hk

March 15, 2024

- Random Access for Machine-to-Machine (M2M) Communications
- A Unified Theory of Random Access
- Example: Rate-Constrained Delay Optimization of Aloha-based M2M Communications
- Access Design for Next-Generation Communication Networks

Random Access for M2M Communications

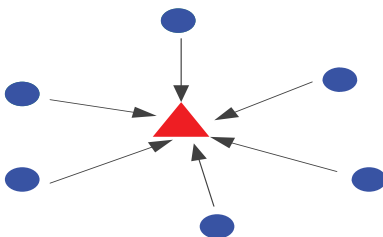
M2M Communications

- M2M communications is expected to play a dominant role in the next-generation communication networks.
 - 80 billions machine-type devices to be connected to mobile networks by 2025
 - Wide applications in various domains such as smart grid, transportation, health care, manufacturing and monitoring
- Two out of three main services of 5G networks are for M2M communications: *massive Machine-Type Communications (mMTC)* and *Ultra-Reliable Low-Latency Communications (URLLC)*.
- Four out of six usage scenarios of 6G networks are related to M2M communications: *hyper reliable and low-latency communications, massive communications, ubiquitous connectivity, integrated sensing and communications, integrated AI and communications.*

Features of M2M Communications

- A typical scenario of conventional Human-to-Human (H2H) communications: A relatively small number of users each with a large amount of data to transmit
- M2M Communications:
 - massive number of devices
 - short packet payload
 - diverse and more stringent Quality-of-Service (QoS) requirements
- *How to provide pervasive and efficient access for M2M communications?*

Multiple Access (MAC)



Multiple users transmit to a common receiver: How to share the resources?

- Centralized Access: A central controller performs resource allocation.
- **Random Access:** Each user determines when/how to access in a **distributed** manner.

- Adopted in cellular systems since the first generation: Each Base Station (BS) allocates dedicated resources to users for data transmission.
- FDMA \longrightarrow TDMA \longrightarrow CDMA \longrightarrow OFDMA
- **Extensive signalling exchange between BSs and users:** Inefficient when the number of users is large and each with a small amount of data.

- Adopted in both cellular and WiFi networks:
 - Cellular in the licensed spectrum: signalling exchange
 - Cellular in the unlicensed spectrum: data transmission
 - WiFi: data transmission
- **Small overhead and scalable:** appealing for M2M communications.

Design of Random Access Networks: Three Key Questions

For each node:

- When to start a transmission?
- When to end the transmission?
- What if the transmission fails?

Question 1: When to Start a Transmission?

- Transmit if packets are awaiting in the queue.
 - Aloha [Abramson'1970]
- A more “polite” solution: Transmit if packets are awaiting in the queue **and the channel is sensed idle**.
 - Carrier Sense Multiple Access (CSMA) [Kleinrock&Tobagi'1975]
- Examples:
 - *Aloha-based*: Random access process of cellular networks (1G-5G) in the licensed spectrum
 - *CSMA-based*: Distributed Coordination Function (DCF) of WiFi networks, 5G New Radio Unlicensed (NR-U) in the unlicensed spectrum

Question 2: When to End the Transmission?

- Stop when the packet transmission is completed.
- Any “smarter” solution? **Stop when the transmission is deemed a failure.**
 - With full-duplex (i.e., able to receive signals during transmissions):
Stop when other on-going transmissions are sensed.
 - With half-duplex (i.e., unable to receive signals during transmissions):
Send a short request to reserve the channel first before the data packet transmission.

Connection (Grant)-based Access

- Examples:
 - *Connection-based*: 4-step Random Access-Small Data Transmission (RA-SDT) in 5G networks in the licensed spectrum, Request-To-Send/Clear-To-Send (RTS/CTS) access mechanism of DCF in WiFi networks
 - *Connection-free*: 2-step RA-SDT in 5G networks in the licensed spectrum, basic access mechanism of DCF in WiFi networks

Question 3: What if the Transmission Fails?

- The definition of transmission failure depends on what type of receivers is adopted. Various assumptions on the receiver have been made, which can be broadly divided into three categories.
 - *Collision Model*: When more than one node transmit their packets simultaneously, a collision occurs and none of them can be successfully decoded. A packet transmission is successful only if **there are no concurrent transmissions**.
 - *Capture Model*: Each node's packet is decoded independently by treating others' as background noise. A packet can be successfully decoded as long as its **received signal-to-interference-plus-noise ratio (SINR) is above a certain threshold**.
 - *Joint-decoding*: **Multiple nodes' packets are jointly decoded**, e.g., Successive Interference Cancellation (SIC).

Question 3: What if the Transmission Fails?

- Resolving transmission failures: **Backoff**
 - Probability-based: Retransmit with a certain probability at each time slot.
 - Window-based: Choose a random value from a window and count down. Retransmit when the counter is zero.
- How to set the transmission probability?
 - **Adjust the transmission probability according to the number of transmission failures i** that the packet has experienced, i.e., $q_i = q_0 \cdot Q(i)$, where $Q(i)$ is an arbitrary monotonic non-increasing function of the number of transmission failures i , $i = 0, 1, \dots$
 - *Binary Exponential Backoff (BEB)*: $Q(i) = 2^{-i}$.
Adopted in DCF of WiFi networks and 5G NR-U.
 - *Constant Backoff*: $Q(i) = 1$.
Adopted in the random access process of cellular systems in the licensed spectrum.

Random Access for M2M Communications: Challenges

- Despite the simplicity in design, the network performance may significantly degrade as the number of users increases if the access parameters are not properly selected.
- To support the massive access and high QoS requirements of M2M communications, the random access schemes need to be carefully designed, with the parameters optimally tuned.

But how?

A Unified Theory of Random Access

A Long History of More than Half a Century

- Numerous random access schemes have been adopted in communication networks since the first random access network, Aloha, was developed by Abramson in 1970.
- Design Degrees of Freedom of Random Access Networks:
 - Sensing-free (Aloha) or Sensing-based (CSMA)
 - Connection-free or Connection-based
 - Backoff: Constant, Exponential, ...
 - ...

Performance Metrics of Random Access Networks

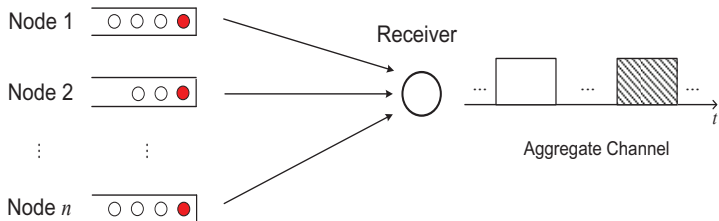
- **Network Throughput:** the average number of successfully decoded packets of the network per time slot.
- **Network Sum Rate:** the average number of successfully decoded information bits of the network per time slot.
- **Delay** of a packet
 - Access delay (service time): the time interval from the instant that it becomes the Head-of-Line (HOL) packet to its successful transmission.
 - Queueing delay (waiting time + service time): the time interval from the packets arrival to its successful transmission.
- **Age of Information (Aol):** the amount of time elapsed since the instant that the freshest delivered packet was generated.
- **Network Stability:** A network is stable if all the nodes' queues are stable.
 - Queue-stability: A queue is defined as queue-stable if the steady-state distribution of its queue length exists.
 - Throughput-stability: A queue is defined as throughput-stable if its throughput is equal to the input rate.

Lack of A Unified Theory of Random Access

- Despite a long history and wide applications, the theory of random access is much less developed than centralized access, which has been the focus of the MAC theory.
- Analytical models are usually customized for specific random access schemes to tackle specific problems.

Modeling of Random Access Networks

- Modeling approaches can be broadly divided into two categories: **Channel-centric** and **Node-centric**, where the former focuses on modeling **the aggregate service process**, and the latter focuses on modeling **nodes' queues**.



Channel-Centric Modeling

- Modeling focus: Aggregate service process
- Representative models:
 - [Abramson'1970], [Kleinrock&Tobagi'1975]
 - Assume the aggregate traffic follows the Poisson distribution.
 - [Carleial&Hellman'1975], [Kleinrock&Lam'1975]
 - Model the dynamic change of the aggregate traffic
- Capture the essence of contention among nodes and simplify the throughput analysis.
- **Ignore nodes' queues.**

Node-Centric Modeling

- Modeling focus: Nodes' queues
- Representative models:
 - [Tsybakov&Mikhailov'1979]
 - Model the queue lengths of n nodes as an n -dimensional Markov chain.
 - **Tractable only for two-node Aloha.**
 - [Rao&Ephremides'1988], [Szpankowski'1994]
[Luo&Ephremides'1999]
 - Develop **approximations and bounds** based on a hypothetical dominant system, where a node would send dummy packets when its queue is empty.
 - [Bianchi'2000], [Kwak,Song&Miller'2005]
 - Consider the symmetric scenario: Model the backoff behavior of each single node.
 - Accurate characterization of network throughput and mean access delay of packets.
 - All the nodes' queues are assumed to be **saturated**.

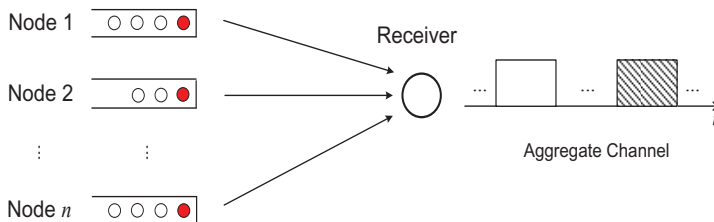
Key questions remain unanswered:

- How to characterize the **coupled** service processes of nodes' queues?
 - Delay: How to characterize and minimize the mean queueing delay of data packets?
 - Stability:
 - + How to determine the stability region of input rates, only within which the network can be stabilized?
 - + For given input rates within the stability region, how to tune the access parameters of nodes to stabilize the network?
 -
- How to characterize the effects of key factors?
 - To sense or not to sense: When is sensing beneficial?
 - Connection-based or connection-free: When is establishing a connection beneficial?
 - Constant backoff or exponential backoff: Which backoff function is the best?
 -

Toward a Unified Analytical Framework for Random Access

- Unified Analytical Framework

- Incorporate all design degrees of freedom and performance metrics
 - Analysis of different random-access schemes can all be based on the same framework.
- For modeling of a multi-queue-single-server system, the main challenge lies in characterization of the coupled service processes of queues, which are determined by the aggregate activities of their **Head-Of-Line (HOL)** packets.



Key to Establishing a Unified Analytical Framework

- Key ingredients for a unified analytical framework of random access [Dai'22] [Dai'13] [Dai'12]:
 - Modeling of HOL packets' behavior: Discrete-time Markov renewal process
 - Characterization of steady-state probabilities of successful transmission of HOL packets \mathbf{p} : Fixed-point equations of \mathbf{p}



L. Dai, "A theoretical framework for random access: Stability regions and transmission control," *IEEE/ACM Trans. Networking*, vol. 30, no. 5, pp. 2173-2200, Oct. 2022.



L. Dai, "Toward a coherent theory of CSMA and Aloha," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3428-3444, Jul. 2013.

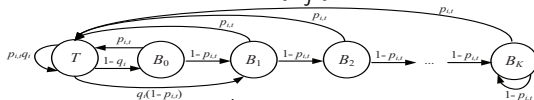


L. Dai, "Stability and delay analysis of buffered Aloha networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2707-2719, Aug. 2012.

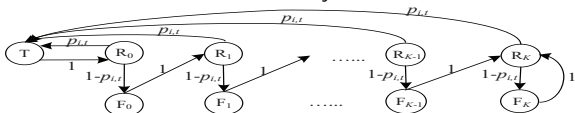
Key to Establishing a Unified Analytical Framework: Modeling of HOL Packets

- HOL packets' behavior can be modeled as a discrete-time Markov renewal process $(\mathbf{X}^h, \mathbf{V}^h) = \{(X_j^h, V_j^h), j = 0, 1, \dots\}$:

- The embedded Markov chain $\mathbf{X}^h = \{X_j^h\}$ **without Sensing**:



- The embedded Markov chain $\mathbf{X}^h = \{X_j^h\}$ **with Sensing**:



- With sensing, the sensing states need to be distinguished from the transmission states.
- The holding time of each state depends on sensing, backoff scheme, and protocol overhead.
- $p_{t,i}$: probability of success given that the HOL packet of Node i is transmitted at t . $\lim_{t \rightarrow \infty} p_{t,i} = p_i, i \in \mathcal{N}$.

Key to Establishing a Unified Analytical Framework: Characterization of \mathbf{p}

- Network performance crucially depends on the steady-state probability of successful transmission of HOL packets \mathbf{p} , which is determined by the aggregate activities of all HOL packets.
- Fixed-point equations of \mathbf{p} can be established based on specific receiver and channel models.

Based on Our Proposed Analytical Framework

• **Fundamental Limits**

- Maximum network throughput
- Minimum mean access/queueing delay
- Maximum network sum rate
- Stability region

• **Insights to Network Design**

- Optimal tuning of backoff parameters (transmission probability, backoff window size, ...) based on the long-term traffic input rates of nodes
- Effects of key factors (sensing, backoff function, connection establishment, network size, receiver design, ...) on limiting performance and performance tradeoffs
- Applications to practical networks

A Glimpse of Our Work

	Aloha	CSMA	WiFi Networks	4G/5G Networks	
				Licensed Bands	Unlicensed Bands
Network Throughput Optimization	[Dai'12] [Gao-Dai'19]	[Dai'13], [Sun-Dai'16] [Gao-Dai'19]	[Dai-Sun'13], [Gao-Sun-Dai'13], [Gao-Dai'13], [Gao-Sun-Dai'14], [Sun-Dai'15], [Sun-Dai'16], [Gao-Dai-Hei'17]	[Zhan-Dai'18] [Zhan-Dai'19]	[Sun-Dai'20]
	[Gao-Fang-Song-Dai'23]				
Delay Optimization	[Da'i12] [Li-Zhan-Dai'21] [Zhao-Dai'23]	[Dai'13], [Sun-Dai'16]	[Dai-Sun'13], [Sun-Dai'15], [Sun-Dai'16]	[Zhan-Dai'19] [Li-Zhan-Dai'21] [Zhao-Dai'23]	
Network Sum Rate Optimization	[Li-Dai'16] [Li-Dai'18]	[Sun-Dai'17] [Sun-Dai'19]	[Sun-Dai'17], [Gao-Sun-Dai'19]		
Stability Region	[Dai'22] [Yang-Dai'23]				

Rate-Constrained Delay Optimization of Aloha-based M2M Communications

Example: Rate-Constrained Delay Optimization of Aloha-based M2M Communications

- Aloha has been adopted in cellular networks, Long Range Radio Wide Area Networks (LoRaWAN), Short Range Devices (SRD) systems, Wireless Body Area Networks (WBAN), ...
- For M2M applications, the data rate and packet delay are important performance metrics.
- *How to optimize the delay performance while satisfying a certain data rate requirement?*



Y. Li, W. Zhan, and L. Dai, "Rate-constrained delay optimization for slotted Aloha," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5283-5298, Aug. 2021.

System Model

- Nodes with data buffers transmit to a single receiver over fading channels using slotted Aloha.

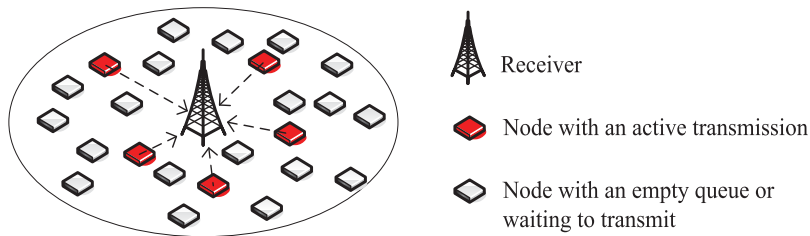


Fig. 2. With Aloha, each node transmits with a certain probability in each time slot when it has packets in its buffer.

- Symmetric setting:
 - Traffic input rate of each node: λ
 - Transmission probability of each node after the i -th failure:
 $q_i = q_0 \cdot Q(i)$, where q_0 is the initial transmission probability, $Q(i)$ is the backoff function (monotonic non-increasing function of i),
 $i = 0, 1, \dots$
 - Channel model: Rayleigh fading – Received SNR of each packet is exponentially distributed with mean ρ .
 - Receiver model: For each packet, its transmission is successful if and only if there are no concurrent transmissions and its received SNR $\eta \geq \mu = 2^{R_{in}} - 1$, where R_{in} (bit/s/Hz) is the information encoding rate.

Problem Formulation

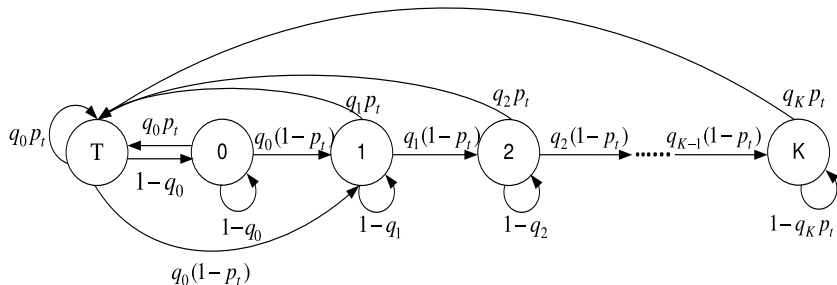
- **Access delay** D_T : The time interval from the instant that a packet becomes the HOL packet to its successful transmission.
- **Node throughput** λ_{out} : The long-term average number of successfully transmitted packets per node.
- **Effective data rate** R_{out} : The long-term average successfully transmitted information rate per node.

$$R_{out} = R_{in} \cdot \lambda_{out}$$

- $\min_{\mu > 0, 0 < q_0 \leq 1} E[D_T]$
s.t. $R_{out} \geq R_0$.

R_0 : minimum required data rate for each node.

HOL-Packet Model



- Steady-state probability of successful transmission of HOL packets:
 $p = \lim_{t \rightarrow \infty} p_t$.
- Service rate of each node's queue: $\pi_T = \frac{1}{\sum_{i=0}^{K-1} \frac{(1-p)^i}{q_i} + \frac{(1-p)^K}{p q_K}}$.
- Mean access delay: $E[D_T] = \frac{1}{\pi_T}$.
- Node throughput: $\lambda_{out} = \lambda$ if $\lambda < \pi_T$, and $\lambda_{out} = \pi_T$ if $\lambda \geq \pi_T$.

Steady-State Probability of Successful Transmission of HOL Packets p

- A packet transmission is successful if and only if
 - its received SNR η is about the threshold μ , and
 - there are no concurrent transmissions, that is, all the other $n - 1$ nodes are either idle with empty queues or busy but not requesting transmissions.
- Steady-state probability of successful transmission of HOL packets:
$$p = \Pr\{\eta \geq \mu\} \cdot (p_{emp} + (1 - p_{emp}) \cdot p_{not})^{n-1}$$
 - For each node, the probability of being busy with a HOL packet but not requesting transmission is $p_{not} = \pi_T(1 - q_0) + \sum_{i=0}^K \pi_i(1 - q_i)$.
 - For each node, the probability of being idle with an empty queue is $p_{emp} = 1 - \lambda/\pi_T$ if $\lambda < \pi_T$, and $p_{emp} = 0$ if $\lambda \geq \pi_T$.
 - $\Pr\{\eta \geq \mu\} = \exp\left(-\frac{\mu}{\rho}\right)$.

Fixed-Point Equations of p in All-Unsaturated and All-Saturated Conditions

- **All-unsaturated:** All the nodes' queues are unsaturated, i.e., with a non-zero probability of being empty.

- Fixed-point equation of p :

$$p = \exp\left(-\frac{\mu}{\rho} - \frac{\hat{\lambda}}{p}\right)$$

- **All-saturated:** All the nodes' queues are saturated, i.e., always busy.

- Fixed-point equation of p :

$$p = \exp\left\{-\frac{\mu}{\rho} - \frac{n}{\sum_{i=0}^{K-1} \frac{p(1-p)^i}{q_i} + \frac{(1-p)^K}{q_K}}\right\}$$

Rate-Constrained Minimum Mean Access Delay

Theorem 1: If $0 \leq R_0 \leq \frac{\bar{C}}{n}$, then the rate-constrained minimum mean access delay D_R^ is given by*

$$D_R^* = \begin{cases} \frac{\mathbb{W}_0 \left(-n\lambda \exp \left(\frac{2 \frac{R_0}{\lambda} - 1}{\rho} \right) \right)}{\lambda \mathbb{W}_{-1} \left(-n\lambda \exp \left(\frac{2 \frac{R_0}{\lambda} - 1}{\rho} \right) \right)} & \text{if } 0 < \lambda \leq \frac{e^{-1}}{n} \text{ and } R_0 \leq \frac{C_u}{n}, \\ n \exp \left(1 + \frac{\mu_1}{\rho} \right) & \text{if } \frac{\hat{\lambda}_\rho}{n} < \lambda < \frac{e^{-1}}{n} \text{ and } \frac{C_u}{n} < R_0 \leq \frac{C_s}{n}, \text{ or } \lambda \geq \frac{e^{-1}}{n} \text{ and } R_0 \leq \frac{C_s}{n}, \end{cases}$$

which is achieved when the SNR threshold μ is set to

$$\mu_R^* = \begin{cases} 2 \frac{R_0}{\lambda} - 1 & \text{if } 0 < \lambda \leq \frac{e^{-1}}{n} \text{ and } R_0 \leq \frac{C_u}{n}, \\ \mu_1 & \text{if } \frac{\hat{\lambda}_\rho}{n} < \lambda < \frac{e^{-1}}{n} \text{ and } \frac{C_u}{n} < R_0 \leq \frac{C_s}{n}, \text{ or } \lambda \geq \frac{e^{-1}}{n} \text{ and } R_0 \leq \frac{C_s}{n}, \end{cases}$$

and the initial transmission probability q_0 is set to (34)

$$q_{0,R}^* = \begin{cases} -\frac{1}{n} \mathbb{W}_{-1} \left(-n\lambda \exp \left(\frac{2 \frac{R_0}{\lambda} - 1}{\rho} \right) \right) & \text{if } 0 < \lambda \leq \frac{e^{-1}}{n} \text{ and } R_0 \leq \frac{C_u}{n}, \\ \frac{1}{n} & \text{if } \frac{\hat{\lambda}_\rho}{n} < \lambda < \frac{e^{-1}}{n} \text{ and } \frac{C_u}{n} < R_0 \leq \frac{C_s}{n}, \text{ or } \lambda \geq \frac{e^{-1}}{n} \text{ and } R_0 \leq \frac{C_s}{n}, \end{cases}$$

where μ_1 is the smaller root of the following equation

$$\frac{1}{n} \exp \left(-1 - \frac{\mu}{\rho} \right) \log_2(1 + \mu) = R_0.$$

Otherwise, the optimization problem (28) has no feasible solution.

Rate-Constrained Minimum Mean Access Delay

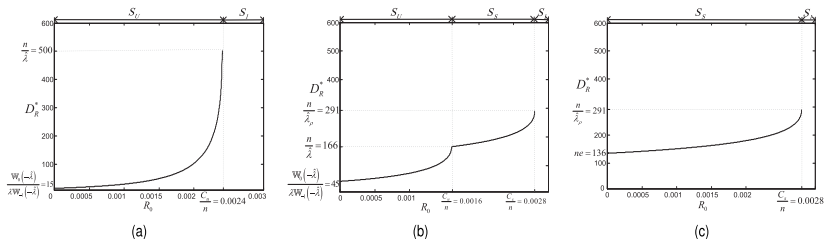


Fig. 8. Rate-constrained minimum mean access delay D_R^* (in unit of time slots) versus the minimum required data rate for each node R_0 (in unit of bit/s/Hz). $n = 50$. $\rho = 0$ dB. $\hat{\lambda}_\rho = 0.1715$. (a) $\hat{\lambda} = 0.1$. (b) $\hat{\lambda} = 0.3$. (c) $\hat{\lambda} = 0.5$.

- Rate-constrained minimum mean access delay D_R^* does not exist when the minimum required data rate R_0 is too large.
- For small traffic input rate $\hat{\lambda}$, the network operates at the all-unsaturated condition. As $\hat{\lambda}$ or R_0 increases, the network may shift to the all-saturated region.

Insights for Massive Access of M2M Communications

TABLE I
CHARACTERISTICS OF THREE TRAFFIC MODELS IN SMART GRID [32], [37]

	Payload Size	Reporting Period	Delay Requirement	Use-case
Traffic model 1 (Delay-insensitive light traffic)	500 bytes	Every 15 minutes	15 minutes	Periodical power grid state reporting
Traffic model 2 (Delay-insensitive heavy traffic)	500 bytes	Every 5 minutes	15 minutes	
Traffic model 3 (Delay-sensitive traffic)	500 bytes	Every 60 minutes	1 second	Control message exchange

- Consider LTE-M with bandwidth $B = 1.08$ MHz and time slot length 15 milliseconds.
- The minimum required data rate normalized by the system bandwidth B is $R_0 = \frac{\text{Payload Size}}{\text{Reporting Period} \times B}$ (bit/s/hz).

Insights for Massive Access of M2M Communications

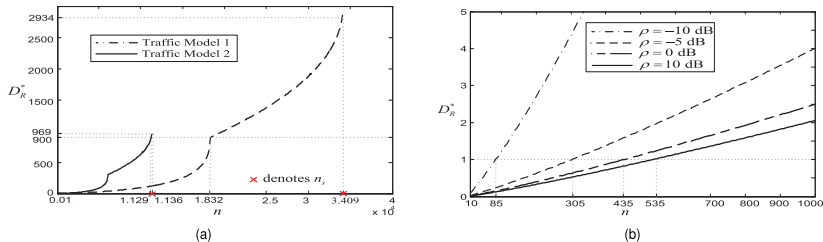


Fig. 11. Rate-constrained minimum mean access delay D_R^* (in unit of seconds) versus the number of devices n . (a) Traffic model 1 and Traffic model 2. $\rho = 0$ dB. (b) Traffic model 3. $\rho = -10$ dB, -5 dB, 0 dB or 10 dB.

- LTE-M is well suited for massive access of machine-type devices with loose QoS requirements.
- For delay-sensitive applications, the network should operate at the unsaturated region with the rate-constrained minimum mean access delay D_R^* linearly increasing with the number of devices n when n is small.

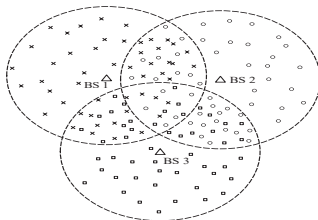
Access Design for Next-Generation Communication Networks

In the Future ...

To support *more* devices with *higher* QoS requirements:

- More BSs/APs
- More “intelligent” access design

More BSs/APs



- **Zero gain (and even worse performance)** if improperly designed.
- Inter-cell interference should be taken account of when optimizing the access design – **That requires information exchange among BSs/APs!**



Y. Yang and L. Dai, "Stability region and transmission control of multi-cell Aloha networks," *IEEE Trans. Commun.*, vol. 71, no. 9, pp. 5348-5364, Sep. 2023.



Y. Gao, L. Dai, and X. Hei, "Throughput optimization of multi-BSS IEEE 802.11 networks with universal frequency reuse," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3399-3414, Aug. 2017.

More “Intelligent” Access Design

- **Learning-based access design:** Each node independently determines when to access based on its own observations/measurements and past experience.
- **Abundant potential demonstrated:** For instance, it was shown in [Peng&Dai'2024] that by properly designing the reward and actions, the network throughput of a simple multi-armed bandit (MAB)-based slotted Aloha network with the collision receiver can surpass the well known limit of e^{-1} and reach the maximum of 1.
- **Lack of analytical framework for performance evaluation:** Effects of key learning parameters such as the learning rate may not be fully understood.



N. Peng and L. Dai, “Multi-Armed-Bandit-based Framed Slotted Aloha for throughput optimization,” to appear in *IEEE Commun. Lett.*

Acknowledgement

Collaborative work with my students:

Current PhD Students:

- Xinran Zhao
- Yunshan Yang
- Xinlong Wang
- Nian Peng

Former PhD Students:

- Xinghua Sun (Sun Yat-sen University)
- Yayu Gao (Huazhong University of Science and Technology)
- Yitong Li (Zhengzhou University)
- Wen Zhan (Sun Yat-sen University)

Thank You!

You may find more information here:

<http://www.ee.cityu.edu.hk/~lindai/>



If you have any questions, please do not hesitate to contact me:

lindai@cityu.edu.hk