Massive Random Access: Fundamental Limits, Optimal Design, and Applications to M2M Communications

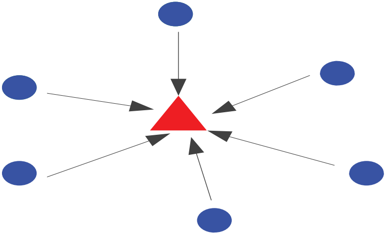
Lin Dai

Department of Electronic Engineering
City University of Hong Kong

lindai@cityu.edu.hk

December 2018

Multiple Access



Multiple nodes transmit to a common receiver: How to share the channel?

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

Design of Random-Access Networks: Three Key Questions

For each node:

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

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]

    - What if nodes cannot sense the channel correctly?
    - Send a short request and be notified by the receiver about the availability of the channel.

Question 2: 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*: 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*: 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 2: What if the Transmission Fails?

Backoff if the transmission fails.

- 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.

In general, backoff can be characterized as a sequence of transmission probabilities $\{q_t\}$, where q_t denotes the transmission probability of the head-of-line (HOL) packet in each node's buffer at time slot t.

It is usually assumed that the transmission probability is adjusted according to the number of transmission failures i that the packet has experienced by time slot t, i.e., $q_i = q_0 \cdot \mathcal{Q}(i)$, where $\mathcal{Q}(i)$ is an arbitrary monotonic non-increasing function of the number of transmission failures i, $i = 0, 1, \ldots$. With Binary Exponential Backoff (BEB), for instance, $\mathcal{Q}(i) = 2^{-i}$.

Question 3: When to End the Transmission?

- Stop when the packet transmission is completed.

- Any "smarter" solution?

  — Stop when other on-going transmissions are sensed.

Applications of Random Access

- WiFi networks
- Cellular networks
- Sensor networks
- Machine-to-machine (M2M) communications, Vehicle-to-vehicle (V2V) communications, Internet-of-things (IoT),......

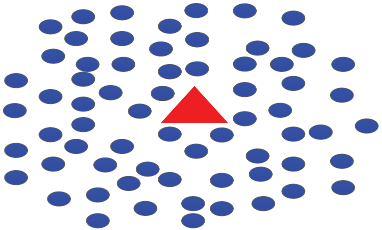
Massive Access of M2M Communications

- M2M communications is expected to play a dominant role in the next-generation communication networks with wide applications in various domains such as smart grid, transportation, health care, manufacturing and monitoring.

- Features of M2M communications
  - — Massive number of machine-type devices (MTDs)
  - — Infrequent transmissions
  - — Short packet payload

- Centralized access would lead to high cost.

Massive Random Access of MTDs



- Due to uncoordination among nodes, the access performance would quickly degrade as the network size increases if backoff parameters are not properly chosen.

- How to efficiently facilitate the massive access from MTDs?

Random access with optimized parameter setting

Modeling of Random-Access Networks

Design Freedoms

- Sense-free (Aloha) or Sense-based (CSMA)
- Packet-based (Grant-free) or Connection-based (Grant-based)
- Backoff: Constant, Exponential, ...
- Receiver: Collision, Capture, SIC, ...
- ...

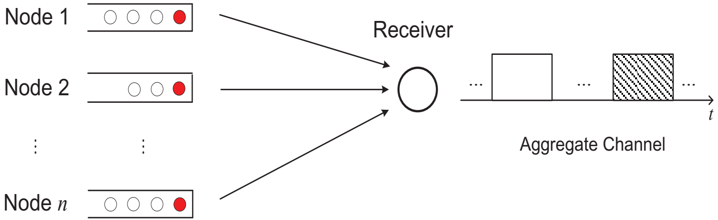
Modeling of Random-Access Networks

Performance Metrics

- Network Throughput: the average number of successfully decoded packets of the network per time slot.
- Delay
  1) Queueing delay (waiting time): the time interval from the packet's arrival to the instant that it becomes the HOL packet;
  2) Access delay (service time): the time interval from the instant that it becomes the HOL packet to its successful transmission.
- Stability
  1) The network is stable if the network throughput is equal to the aggregate input rate.
  2) The network is stable if the mean access/queueing delay is finite.
- Network Sum Rate: the average number of successfully decoded information bits of the network per time slot.
- ...

Modeling of Random-Access Networks

- Numerous models have been proposed for various random-access schemes to tackle different problems.
- Inconsistent or even contradictory results were obtained due to differences in modeling assumptions.
- The existing models can be roughly divided into two categories: channel-centric or node-centric.



Modeling of Random-Access Networks: Channel-Centric or Node-Centric

- Channel-centric modeling: to characterize the aggregate traffic of *all* the nodes.

  — Capture the essence of contention among nodes and simplify the throughput analysis.

  — Difficult to analyze the queueing performance of each node.

- Node-centric modeling: to characterize the queueing behavior of each node.

  — High modeling complexity when interactions among nodes' queues are taken into consideration.

  — Usually involve jointly solving a set of fixed-point equations (the number of equations increases with the number of nodes).

Toward a Unified and Scalable Analytical Framework

- Unified: Analysis of different random-access schemes can all be based on the same framework.
- Scalable: Modeling complexity does not increase with the network size.

- Keys to establishing a unified and scalable analytical framework for random-access networks [Dai'12] [Dai'13]:

— Treat each node's queue as an independent queueing system;

— State characterization of each individual head-of-line (HOL) packet;

— Characterization of network steady-state points based on the fixed-point equations of steady-state probability of successful transmission of HOL packets.

Performance Optimization of Random-Access Networks

Fundamental Limits

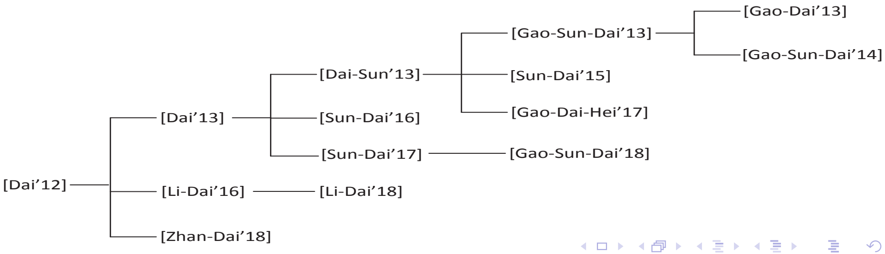
- Maximum network throughput
- Minimum mean access delay and minimum second moment of access delay
- Maximum network sum rate

Insights to Network Design

- Optimal tuning of backoff parameters
- Effects of key factors (sensing, backoff function, network size, receiver design, ...) on limiting performance
- Fundamental performance tradeoffs
- Applications to practical networks

A Glimpse of Our Work

	Aloha	CSMA	802.11 DCF	M2M in LTE
Network Throughput Optimization	[Dai'12]	[Dai'13], [Sun-Dai'16]	[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]
Delay Optimization	[Dai'12]	[Dai'13], [Sun-Dai'16]	[Dai-Sun'13], [Sun-Dai'15], [Sun-Dai'16]	
Network Sum Rate Optimization	[Li-Dai'16] [Li-Dai'18]	[Sun-Dai'17]	[Sun-Dai'17], [Gao-Sun-Dai'18]	



References

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

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 and X. Sun, "A unified analysis of IEEE 802.11 DCF networks: stability, throughput and delay," *IEEE Trans. Mobile Computing*, vol. 12, no. 8, pp. 1558–1572, Aug. 2013.

Y. Gao, X. Sun, and L. Dai, "Throughput Optimization of Heterogeneous IEEE 802.11 DCF Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 398–411, Jan. 2013.

Y. Gao and L. Dai, "Optimal Downlink/Uplink Throughput Allocation for IEEE 802.11 DCF Networks," *IEEE Wireless Commun. Letters*, vol. 2, no. 6, pp. 627–630, Dec. 2013.

Y. Gao, X. Sun, and L. Dai, "IEEE 802.11e EDCA Networks: Modeling, Differentiation and Optimization," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3863–3879, July 2014.

References

- X. Sun and L. Dai, "Backoff Design for IEEE 802.11 DCF Networks: Fundamental Tradeoff and Design Criterion," *IEEE/ACM Trans. Networking*, vol. 23, no. 1, pp. 300–316, Feb. 2015.
- Y. Li and L. Dai, "Maximum Sum Rate of Slotted Aloha with Capture," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 690–705, Feb. 2016.
- X. Sun and L. Dai, "Performance Optimization of CSMA Networks with A Finite Retry Limit," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 5947–5962, Sept. 2016.
- X. Sun and L. Dai, "Fairness-constrained Maximum Sum Rate of Multi-rate CSMA Networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1741–1754, Mar. 2017.
- 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.
- W. Zhan and L. Dai, "Massive Random Access of Machine-to-Machine Communications in LTE Networks: Modeling and Throughput Optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2771-2785, Apr. 2018.

References

Y. Li and L. Dai, "Maximum Sum Rate of Slotted Aloha with Successive Interference Cancellation," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5385-5400, Nov. 2018.

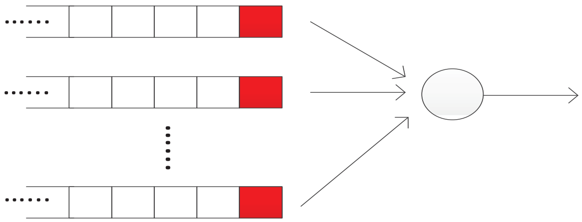
Y. Gao, X. Sun, and L. Dai, "Sum Rate Optimization of Multi-Standard IEEE 802.11 WLANs," *to appear in IEEE Trans. Commun.*

A Unified Analytical Framework

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

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

Node-Centric Modeling



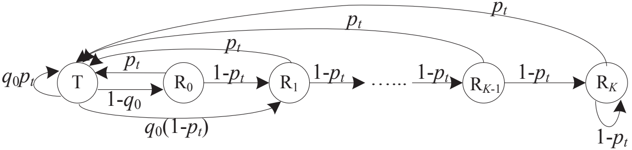
- An n-node buffered random-access network can be modeled as an n-queue-single-server system.
- Key assumption: Each node's queue can be regarded as an independent queueing system when the number of nodes n is large.
- Key steps: Characterization of each head-of-line (HOL) packet's state transition and steady-state probability of successful transmission of HOL packets.

State Characterization of HOL Packet

The state transition of each HOL packet can be modeled as a discrete-time Markov renewal process $(\mathbf{X}^h, \mathbf{V}^h) = \{(X_j^h, V_j^h), j = 0, 1, \ldots\}$.

The embedded Markov chain $\mathbf{X}^h = \{X_j^h\}$ is:

Without Sensing



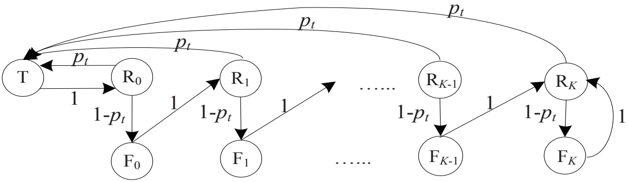
- K: cutoff phase. The HOL packet stays at State R_K if the number of transmission failures exceeds K.
- p_t: probability of successful transmission of HOL packets at time slot t. $\lim_{t\to\infty} p_t = p$.
- q_0: initial transmission probability of a fresh HOL packet.

State Characterization of HOL Packet

The state transition of each HOL packet 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 $\mathbf{X}^h = \{X_j^h\}$ is:

With Sensing



- The states of $\{X_j^h\}$ are divided into three categories: 1) waiting to request (State R_i, $i = 0, \dots, K$), 2) failure (State F_i, $i = 0, \dots, K$) and 3) successful transmission (State T).
- p_t: probability of successful transmission of HOL packets at mini-slot t given that the channel is idle at $t-1$. $\lim_{t\to\infty} p_t = p$.

State Characterization of HOL Packet

The state transition of each HOL packet 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 mean holding time in each state:

- τ_{R_i}: determined by the backoff scheme, i.e., transmission probability sequence $\{q_i\}$.
- Without sensing: $\tau_T = 1$ time slot.
- With sensing: τ_T and τ_F vary under different system settings.

Service Rate and Service Time Distribution

- For each node's queue, it has a successful transmission if and only if the HOL packet is in State T.
- Let $\{\tilde{\pi}_i\}$ denote the limiting state probabilities of the Markov renewal process $(\mathbf{X}^h, \mathbf{V}^h)$.

  — Service rate: $\tilde{\pi}_T = \frac{\pi_T \cdot \tau_T}{\sum_{j \in \mathbb{S}} \pi_j \cdot \tau_j}$, $\mathbb{S}$ is the state space of $\mathbf{X}^h$.

  — The node throughput is determined by the service rate when each node's queue is saturated.
- Let D_i denote the time spent from the beginning of State i until the service completion.

  — Service time distribution: The probability generating function of D_T, $G_{D_T}(z)$, can be calculated based on the Markov renewal process of HOL packet.

  — The service time of each node's queue is also the access delay of each packet.

Network Steady-state Points

- The network steady-state points are characterized based on the fixed-point equations of the limiting probability of successful transmission of HOL packets, $p = \lim_{t \to \infty} p_t$.
- Given that the network is in unsaturated or saturated conditions, different fixed-point equations of p can be established.
- The fixed-point equation of p in unsaturated conditions may have multiple roots, but not all of them are steady-state points.

Network Throughput

- The network throughput is defined as the average number of successfully decoded packets of the network per time slot.
-

$$\text{Network Throughput } \hat{\lambda}_{out} = \begin{cases} \text{aggregate input rate } n\lambda & \text{unsaturated} \\ \text{aggregate service rate } n\tilde{\pi}_T & \text{saturated.} \end{cases}$$

- The maximum network throughput: $\max_{\{q_i\}} \hat{\lambda}_{out}$.

- n: number of nodes;
- λ: input rate of each node's queue;
- $\tilde{\pi}_T$: service rate of each node's queue;
- $\{q_i\}$: transmission probability (backoff) sequence of each node.

Access Delay

- The access delay of each packet is the service time of each node's queue.
- Moments of access delay can be obtained from the probability generating function of service time $G_{D_T}(z)$.
- The minimum mean access delay: $\min_{\{q_i\}} E[D_T]$;
  The minimum second moment of access delay: $\min_{\{q_i\}} E[D_T^2]$.

Sum Rate

- The sum rate is defined as the average number of successfully decoded information bits of the network per time slot.
- Let R denote the information encoding rate of each packet. The network sum rate $R_s = R \cdot \hat{\lambda}_{out}$.
- The maximum sum rate: $\max_{R,\{q_i\}} R_s$.

Summary

- The key to node-centric modeling lies in proper characterization of 1) the state transition process of each HOL packet, and 2) the steady-state probability of successful transmission of HOL packets.
- Based on the proposed unified analytical framework, effects of key parameters on a wide range of performance metrics such as network throughput, access delay and sum rate, of various random-access networks can be evaluated in a systematic manner.
- The proposed analytical framework can further facilitate performance optimization to reveal the fundamental limits of random-access networks, and to show how to properly set the key system parameters to achieve the limiting performance.

Fundamental Limits of Aloha Networks

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

Y. Li and L. Dai, “Maximum Sum Rate of Slotted Aloha with Capture,” *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 690–705, Feb. 2016.

Y. Li and L. Dai, “Maximum Sum Rate of Slotted Aloha with Successive Interference Cancellation,” *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5385-5400, Nov. 2018.

Maximum Network Throughput and Maximum Sum Rate

- Maximum network throughput: The maximum average number of successfully decoded packets of the network per time slot by optimizing the transmission probability (backoff) sequence of each node.

  $\hat{\lambda}_{\max} = \max_{\{q_i\}} \hat{\lambda}_{out}$

- Maximum sum rate: The maximum average number of successfully decoded information bits of the network per time slot by optimizing the information encoding rate and transmission probability (backoff) sequence of each node.

  $C = \max_{R,\{q_i\}} R_s = \max_R R \cdot \hat{\lambda}_{\max}$

Maximum Network Throughput

- Two network steady-state points p_L and p_A:

  Unsaturated: p_L is independent of the backoff sequence $\{q_i\}$ and determined by the aggregate input rate $\hat{\lambda}$;

  Saturated: p_A is independent of the aggregate input rate $\hat{\lambda}$ and determined by the backoff sequence $\{q_i\}$.

- Network throughput $\hat{\lambda}_{out}$:

  Unsaturated: $\hat{\lambda}_{out} = \hat{\lambda}$;

  Saturated: $\hat{\lambda}_{out} = p_A \log p_A$.

- To maximize the network throughput:

  Increase the aggregate input rate $\hat{\lambda}$ to saturation and optimally choose the backoff sequence $\{q_i\}$.

Maximum Network Throughput

- Assume that the received signal-to-noise ratios (SNRs) of packets are exponentially distributed with mean received SNR ρ.
- Collision model: A packet can be successfully decoded if there is no concurrent transmission and its received SNR is above a threshold μ.

$$\hat{\lambda}_{\max}^{collision} = \exp\left(-1 - \frac{\mu}{\rho}\right). \quad (1)$$

  – Perfect channel conditions ($\rho \to \infty$): $\hat{\lambda}_{\max}^{collision} \to e^{-1}$.

- Capture model: A packet can be successfully decoded if its received SINR is above a threshold μ.

$$\hat{\lambda}_{\max}^{capture} = \begin{cases} \frac{\mu+1}{\mu} \exp\left(-1 - \frac{\mu}{\rho}\right) & \text{if } \mu \geq \frac{1}{n-1} \\ n \exp\left(-\frac{n\mu}{\mu+1} - \frac{\mu}{\rho}\right) & \text{otherwise.} \end{cases} \quad (2)$$

  – $n \to \infty$, $\hat{\lambda}_{\max}^{capture} \to \frac{\mu+1}{\mu} \exp\left(-1 - \frac{\mu}{\rho}\right)$; $\mu \to 0$, $\hat{\lambda}_{\max}^{capture} \to n$.

  – $\hat{\lambda}_{\max}^{capture}$ increases as μ decreases, suggesting a tradeoff between the information encoding rate and network throughput.

Maximum Sum Rate

- Assume that each node independently encodes its information at a constant rate R bit/s/Hz.
- Assume that each packet consists of one codeword, and each codeword lasts for one time slot, i.e., no coding over successive packets.
- Assume that the codeword length is long enough such that a packet can be successfully decoded as long as its information encoding rate does not exceed its single-user capacity, i.e., $R \leq \log_2(1+\eta)$, where η is the SNR/SINR of each packet. The SNR/SINR threshold μ is then given by $\mu = 2^R - 1$.
- The maximum sum rate can be written as $C = \max_{R,\{q_i\}} R_s = \max_{\mu} \hat{\lambda}_{\max} \log_2(1+\mu)$.

Maximum Sum Rate

- Collision model:

$$C^{collision} = \exp\left(-1 - \frac{e^{\mathbb{W}_0(\rho)} - 1}{\rho}\right) \cdot \log_2(e^{\mathbb{W}_0(\rho)}). \quad (3)$$

– $\lim_{\rho\to\infty} \frac{C^{collision}}{\log_2 \rho} = e^{-1}$; $C^{collision} \approx e^{-2} \log_2 e \cdot \rho$ for $\rho \ll 1$.

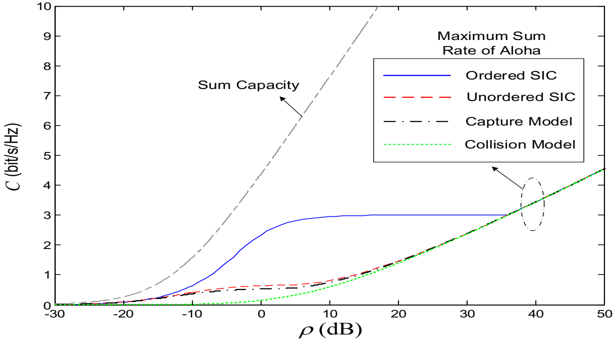
- Capture model:

$$C^{capture} = \begin{cases} \frac{\mu_h^*+1}{\mu_h^*} \exp\left(-1 - \frac{\mu_h^*}{\rho}\right) \log_2(1+\mu_h^*) & \text{if } \rho \geq \rho_0 \\ n \exp\left(-\frac{n\mu_l^*}{\mu_l^*+1} - \frac{\mu_l^*}{\rho}\right) \log_2(1+\mu_l^*) & \text{otherwise,} \end{cases} \quad (4)$$

where $\rho_0 = \frac{\frac{n}{n-1} \ln \frac{n}{n-1}}{1-(n-1)\ln \frac{n}{n-1}} \overset{n\gg 1}{\approx} 3$ dB. μ_h^* and μ_l^* are the roots of $(\mu+1)^{\frac{\mu+1}{\rho}+\frac{1}{\mu}} = e$ and $(\mu+1)^{\frac{\mu+1}{\rho}+\frac{n}{\mu+1}} = e$, respectively.

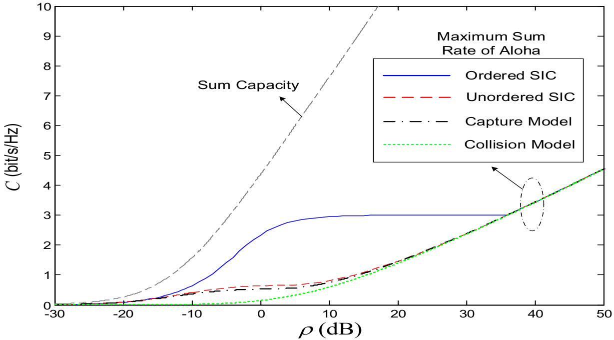
– $\lim_{\rho\to\infty} \frac{C^{capture}}{\log_2 \rho} = e^{-1}$; $\lim_{n\to\infty} C^{capture} = e^{-1} \log_2 e$ for $\rho < \rho_0$.

Maximum Sum Rate

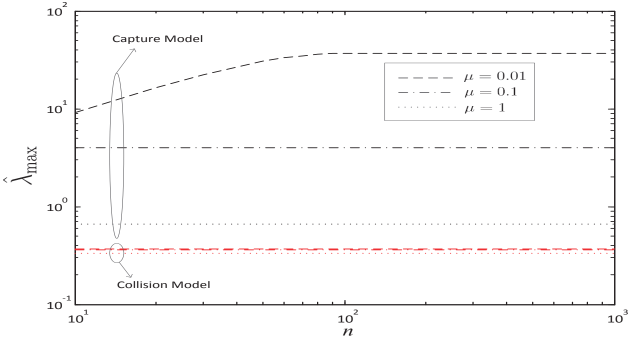


- The maximum sum rates of Aloha with collision, capture and SIC receivers all have the same high-SNR slope of e^{-1}.
- Only a small rate difference is observed between collision model and capture model.

Maximum Sum Rate



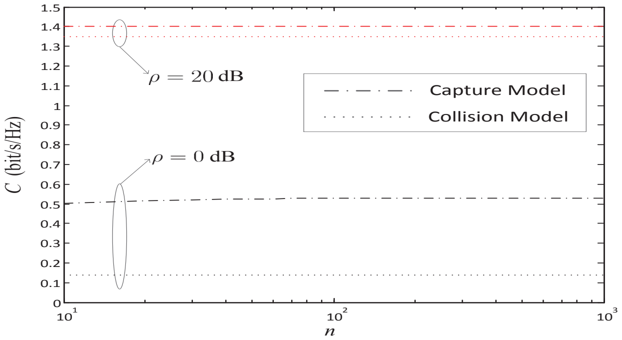
- Even with the capacity-achieving receiver structure, the ordered SIC, the maximum sum rate of Aloha is still significantly lower than the ergodic sum capacity.
- Causes of rate loss: 1) packet-based encoding/decoding; 2) no rate allocation among nodes due to no coordination.



- For large number of nodes $n \gg 1$:

$$\hat{\lambda}_{\max}^{collision} = \exp\left(-1 - \frac{\mu}{\rho}\right),\ \hat{\lambda}_{\max}^{capture} = \frac{\mu+1}{\mu} \exp\left(-1 - \frac{\mu}{\rho}\right).$$

More Insights: Massive Access



- For large number of nodes $n \gg 1$:
  - High SNR ($\rho \gg 1$):

$$C^{capture} \approx C^{collision} = \exp\left(-1 - \frac{e^{\mathbb{W}_0(\rho)} - 1}{\rho}\right) \cdot \log_2(e^{\mathbb{W}_0(\rho)});$$

  - Low SNR ($\rho < \rho_0$): $C^{capture} \approx e^{-1} \log_2 e \approx 0.53$ bit/s/Hz.

More Insights: Optimal Backoff

With the capture model, the optimal SINR threshold μ^* for achieving the maximum sum rate is

$$\mu^* = \begin{cases} \mu_h^* & \text{if } \rho \geq \rho_0 \\ \mu_l^* & \text{otherwise,} \end{cases} \quad (5)$$

and the optimal transmission probability sequence of each node $\{q_i^*\}$ is

$$q_i^* = \begin{cases} \hat{q}_0 \mathcal{Q}(i) & \text{if } \mu^* \geq \frac{1}{n-1} \\ 1 & \text{otherwise,} \end{cases} \quad (6)$$

$i = 0, \ldots, K$, where $\hat{q}_0$ is given by

$$\hat{q}_0 = \frac{\mu^*+1}{n\mu^*} \cdot \left\{ \sum_{i=0}^{K-1} \frac{\exp\left(-1-\frac{\mu^*}{\rho}\right)\left[1-\exp\left(-1-\frac{\mu^*}{\rho}\right)\right]^i}{\mathcal{Q}(i)} + \frac{\left[1-\exp\left(-1-\frac{\mu^*}{\rho}\right)\right]^K}{\mathcal{Q}(K)} \right\}.$$

- To achieve the maximum sum rate, the transmission probabilities of nodes should be adaptively tuned according to the number of nodes n and the mean received SNR ρ.

More Insights: Effect of Adaptive Backoff

What if each node transmits its packet with a constant probability q at each time slot?

With the capture model:

- Network throughput: $\hat{\lambda}_{out}^{q_i=q} = nq\exp\left(-\frac{\mu}{\rho} - \frac{nq\mu}{\mu+1}\right)$;
- Sum rate: $\tilde{R}_s^{q_i=q} = \lim_{\rho\to\infty} R_s^{q_i=q} = nq\exp\left(-\frac{nq\mu}{\mu+1}\right)\cdot\log_2(1+\mu)$.

$$\max_{\mu} \tilde{R}_s^{q_i=q} \overset{n\gg 1}{\approx} e^{-1}\log_2 e.$$

Without adaptive backoff, the network throughput exponentially decreases as the network size increases, and the sum rate converges to a limit that is much lower than 1 with the mean received SNR $\rho \to \infty$.

Summary

- Both maximum network throughput and maximum sum rate of Aloha with representative receiver structures have been obtained as functions of key system parameters including the mean received SNR ρ, the number of nodes n and the SNR/SINR threshold μ (or equivalently, the information encoding rate of each node).
- The analysis reveals that although substantial gains in network throughput can be achieved by SIC receivers and the capture model at the low SNR region thanks to multi-packet reception (MPR), the rate difference could be limited, and they all reduce to the collision model with the high-SNR slope of e^{-1} when the mean received SNR ρ is large.
- Without proper rate allocation, even with the capacity-achieving receiver structure, the maximum sum rate of Aloha could still be far below the ergodic sum capacity of fading channels, indicating that the rate loss is significant due to uncoordinated random transmissions of nodes.

Random Access of M2M Communications in LTE Networks

W. Zhan and L. Dai, "Massive Random Access of Machine-to-Machine Communications in LTE Networks: Modeling and Throughput Optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2771-2785, Apr. 2018.

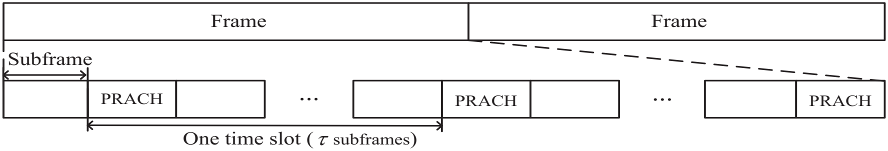
Random Access Process of LTE Networks

Connection-Based Aloha

- In LTE networks, a connection would first be established between a device and the base station (BS) before the device starts to transmit its data packets.
- The connection-based random access adopted in the LTE networks differs from the conventional packet-based random access in that the data packets do not contend for the channel individually. Instead, each device with data packets to transmit first sends a connection request to the BS, and if a device's request is successfully received, then the BS will allocate resource blocks for the device to clear its data queue.

Random Access Process of LTE Networks

Periodical PRACH Subframes



Orthogonal Preambles

- Each device randomly chooses one out of M orthogonal preambles.
- If more than one device transmits the same preamble over the same PRACH subframe, then a collision occurs and all of them fail.

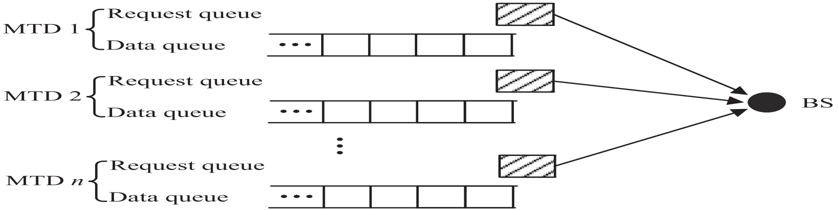
ACB Check and Uniform Backoff

- ACB check: Each access request is transmitted with probability q. That is, a random number between 0 and 1 is generated, and compared with the ACB factor $q \in (0, 1]$. If the number is less than q, then the access request is transmitted. Otherwise, it is barred temporarily.
- Uniform Backoff: If an access request is transmitted but involved in a collision, then it randomly selects a value from $[0, W]$, and counts down until it reaches zero.

Network Throughput

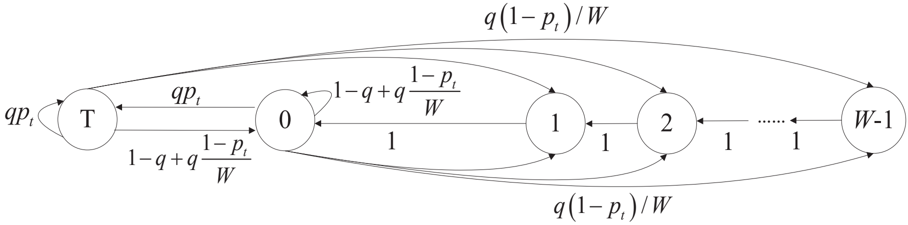
- Focus on the average number of devices that can successfully access the BS per PRACH subframe.
- Network throughput is defined as the average number of successful access requests per time slot (a time slot is defined as the interval between two consecutive PRACH subframes).
- How to properly tune the ACB factor q and the uniform backoff window size W to maximize the network throughput?

Double-Queue Model of Each Device



- Each device has one data queue and one request queue.
- Each newly arrival data packet generates an access request, but only one request can be kept since each device can have at most one ongoing access request regardless of how many data packets in its buffer.
- Upon successful transmission of the access request, the BS will assign sufficient resources for the device to clear its data buffer.
- Assume that the arrivals of data packets follow a Bernoulli process with parameter $\lambda \in (0, 1)$.

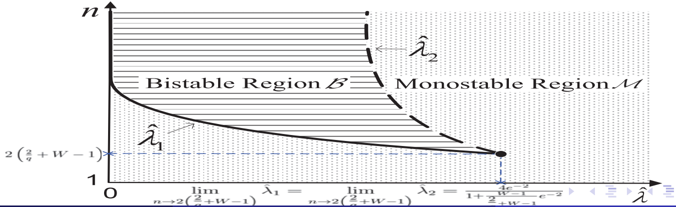
State Characterization of Access Request



- Characterization of network steady-state points is based on the fixed-point equation of limiting probability of successful transmission of access requests $p = \lim_{t\to\infty} p_t$.

Single Preamble: Network Steady-State Points

- Fixed-point equation of limiting probability of successful transmission of access requests p: $p = \exp\left(-\frac{\hat{\lambda}}{\frac{\hat{\lambda}}{n}\left(\frac{1}{q}+\frac{W-1}{2}\right)+p\left(1-\frac{\hat{\lambda}(W-1)}{2n}\right)}\right)$.
- The network has either one steady-state point or two steady-state points, depending on which stable region it operates at.
  – Bistable region $\mathcal{B} = \left\{(n, \hat{\lambda}, q, W) | n > 2\left(\frac{2}{q}+W-1\right), \hat{\lambda}_1 \le \hat{\lambda} \le \hat{\lambda}_2\right\}$, in which the network has two steady-state points p_L and p_A.
  – Monostable region $\mathcal{M} = \bar{\mathcal{B}}$, in which the network has only one steady-state point p_L.



Single Preamble: Maximum Network Throughput

- The network throughput $\hat{\lambda}_{\text{out}} = \dfrac{\hat{\lambda}}{\frac{\hat{\lambda}}{n}\left(\frac{1}{qp}+\frac{(1-p)(W-1)}{2p}\right)+1}$.
- The maximum network throughput $\hat{\lambda}_{\max} = \max\limits_{(q,W)} \hat{\lambda}_{\text{out}}$.

Theorem

The maximum network throughput $\hat{\lambda}_{\max} = e^{-1}$, which is achieved if and only if the network operates at the desired steady-state point p_L, and (q^*, W^*) together satisfy

$$\frac{1}{q^*} + \frac{1-e^{-1}}{2}\left(W^* - 1\right) = n\left(1 - \frac{e^{-1}}{\hat{\lambda}}\right). \tag{7}$$

Extension to Multi-Preamble

- The maximum network throughput $\hat{\lambda}^M_{\max} = Me^{-1}$, where M is the number of orthogonal preambles.

- The optimal ACB factor $q^{*,M}$ and uniform backoff window size $W^{*,M}$ should satisfy

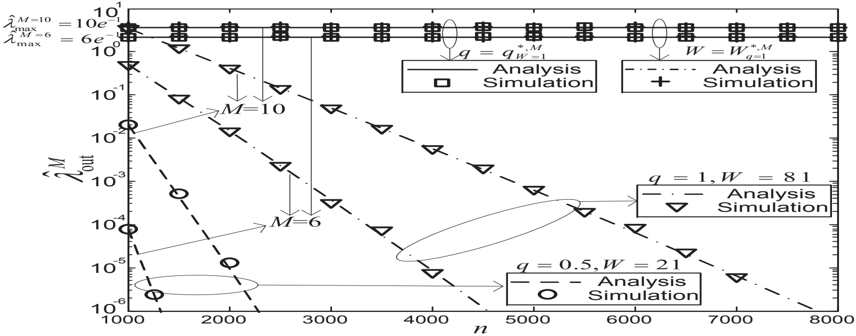
$$\frac{1}{q^{*,M}} + \frac{1-e^{-1}}{2}\left(W^{*,M} - 1\right) = \frac{n}{M} - \frac{e^{-1}}{\lambda}, \tag{8}$$

where n is the total number of devices, and λ is the traffic input rate of each device.

— $q^{*,M}_{W=1} = \frac{\lambda}{\frac{n\lambda}{M} - e^{-1}}$; $W^{*,M}_{q=1} = \frac{2\left(\frac{n}{M} - \frac{e^{-1}}{\lambda} - 1\right)}{1-e^{-1}} + 1.$

— The optimal tuning of backoff parameters is solely based on statistical information.

Network Throughput with Optimal Tuning and Standard Setting



- With the standard setting (fixed q and W): The network throughput quickly deteriorates as the number of devices n increases.
- With the optimal tuning of q or W: The maximum network throughput $\hat{\lambda}_{max}^{M}$ can always be achieved, which does not vary with the number of devices n.

Discussions: Effect of Outdated Information of n and λ

- To calculate the optimal values of backoff parameters, the BS needs to count the total number of MTDs n and collect the traffic input rate information λ from the feedback of MTDs.
- What if the information of n and λ is outdated?
- Define $\gamma_n = \frac{\tilde{n}-n}{\tilde{n}}$ and $\gamma_\lambda = \frac{\tilde{\lambda}-\lambda}{\tilde{\lambda}}$ as the relative error on n and λ, respectively, where $\tilde{n}$ and $\tilde{\lambda}$ denote the outdated information on the number of MTDs and the input rate of each MTD at the BS, respectively.
- Suppose that the optimal backoff parameters $q^{*,M}_{W=1}$ and $W^{*,M}_{q=1}$ are calculated based on the outdated information $\tilde{n}$ and $\tilde{\lambda}$.

Discussions: Effect of Outdated Information of n and λ

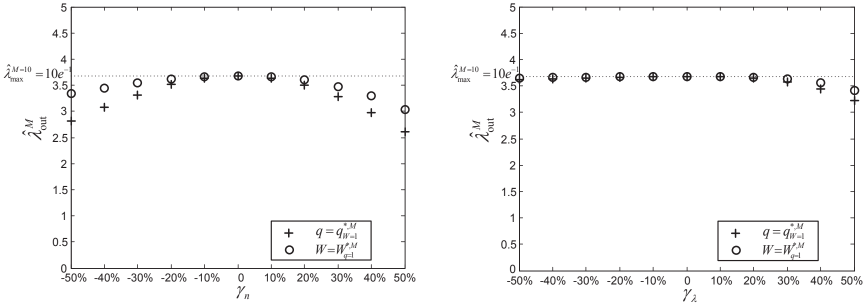


Figure : Simulated network throughput $\hat{\lambda}_{\text{out}}^{M}$ versus the relative errors γ_n and γ_λ. $M = 10$. $\tilde{n} = 1000$. $\tilde{\lambda} = 0.01$. (a) $\gamma_\lambda = 0$. (b) $\gamma_n = 0$.

- The network throughput stays at the maximum value even when the relative error γ_λ of traffic input rate is large.
- The network is more robust against the variation of network size and traffic input rate with the optimal tuning of uniform backoff window size W.

Discussions: Effect of Bursty Arrivals of Data Packets

- The analysis is based on the assumption that data packet arrivals of each MTD independently follow a Bernoulli process.
- In practical M2M communication scenarios, packet arrival processes could be bursty.

- We consider two kinds of burstiness: temporal burstiness and spatial burstiness.

  – Temporal burstiness: a continuous stream of packets is generated in a short time period for a given MTD.

  – Spatial burstiness: multiple MTDs generate packets in a synchronous manner.
- To capture the temporal burstiness, model the data packet arrival process of each MTD as a two-state Markov modulated Bernoulli process (MMBP). To capture the spatial burstiness, assume that n_s out of n MTDs have identical data packet arrival processes, i.e., they are all driven by a common MMBP.

Discussions: Effect of Bursty Arrivals of Data Packets

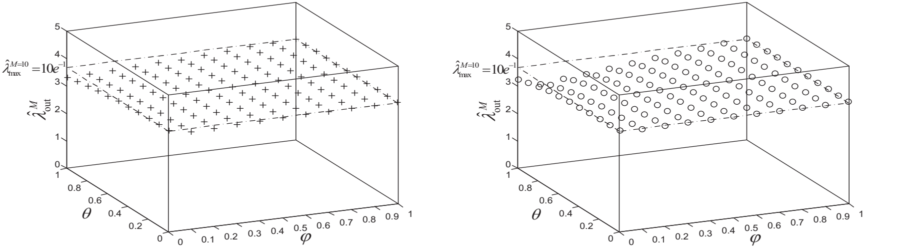


Figure : Simulated network throughput $\hat{\lambda}_{\text{out}}^{M}$ versus the synchronization ratio θ and the frequency of bursty arrivals φ. $M = 10$. $n = 1000$. $\sigma_0 = 0.005$. $\lambda_0 = 0.006$. $\lambda_1 = 1$. (a) $q = q_{W=1}^{*,M}$. (b) $W = W_{q=1}^{*,M}$.

- The network throughput is close to the maximum value within a wide range of θ and φ.
- By optimally tuning the backoff parameters, the bursty input traffic can be sufficiently randomized. Therefore, the maximum network throughput can be achieved even when the input traffic has a high degree of temporal burstiness and spatial burstiness.

Summary

- The proposed unified analytical framework is applied to LTE networks to optimize the random access performance of M2M communications.
- The modeling complexity is independent of the number of MTDs even with the queueing behavior of each MTD taken into consideration, which is highly attractive in the massive access scenario.
- The analysis reveals how to optimally tune the backoff parameters including the ACB factor and uniform backoff window size to maximize the network throughput. The optimal tuning is solely based on the statistical information such as the traffic input rate of each MTD, with which the throughput performance is found to be robust against feedback errors of the traffic input rate and burstiness of data arrivals.

The End

Thank You!