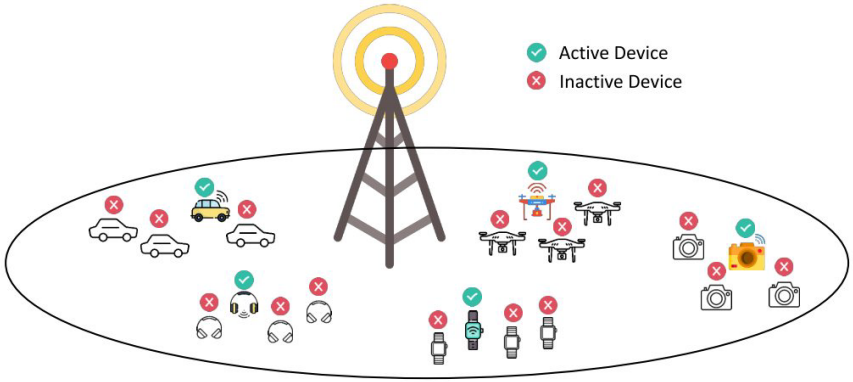
Massive Random Access with Massive MIMO: Contention versus Scheduling

Wei Yu

Joint Work with Justin Kang

University of Toronto

Massive Connectivity



- Massive connectivity is a crucial requirement for Internet-of-Things (IoT)
- Requires up to $10^5 \sim 10^6$ devices connected per base station (BS).
  - Sporadic traffic, making device identification & scheduling challenging.
  - Assigning each user an orthogonal resource requires coordination.
- Activity Detection is a first step toward coordination.
- Equally importantly, we need to schedule users to transmission slots.

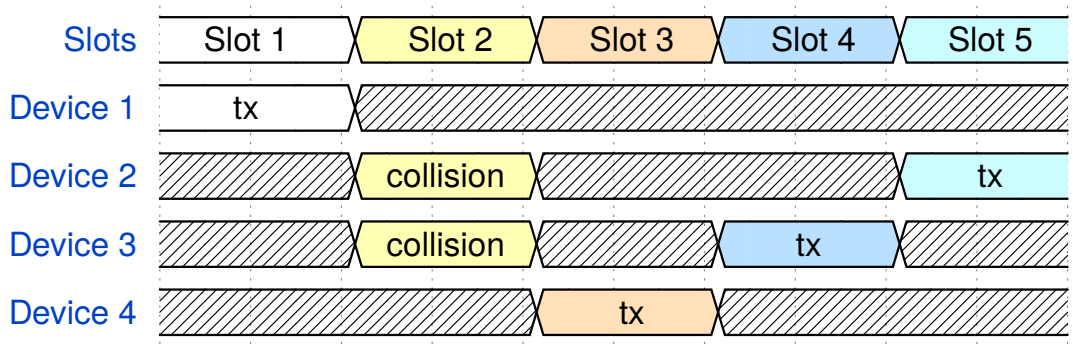
What is the cost of coordinated scheduling?

Contention-Based vs Coordinated Scheduling

- **Uncoordinated Random Access:**
  - Classic Slotted ALOHA: Contention-based uncoordinated scheduling.
  - Coded ALOHA can alleviate some of the inefficiencies of classic ALOHA.
- **Coordinated Random Access:**
  - Coordinated scheduling requires feedback from the BS to the users.
  - What is the minimum feedback rate for scheduling?
- **Massive Random Access with Massive MIMO:**
  - Coded Pilot Access vs. Scheduled Random Access

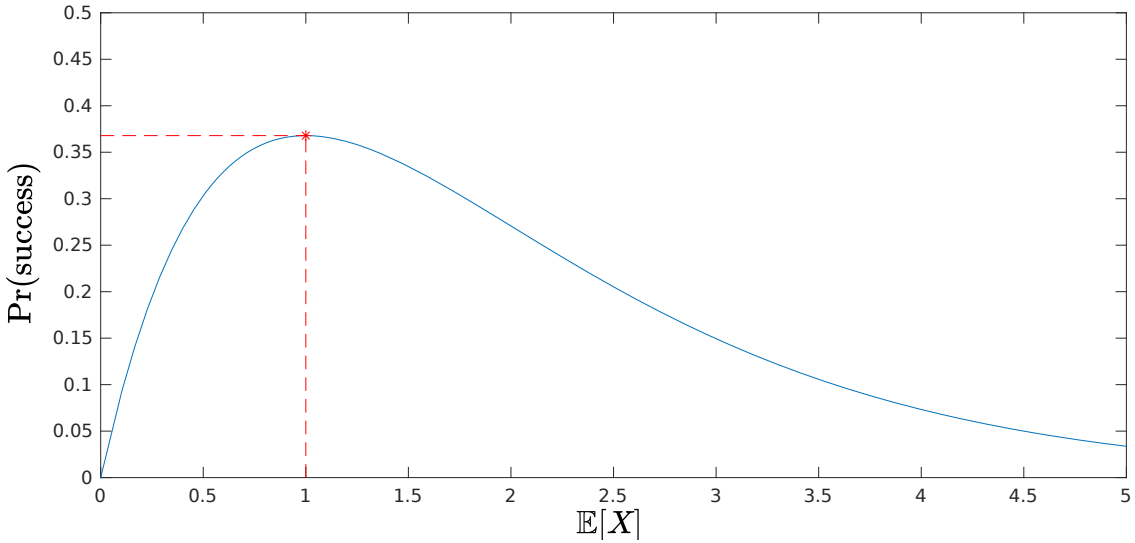
Classic Solution: Slotted ALOHA

Slotted ALOHA involves contention and is uncoordinated involving no communication between BS and users.



- Users become active and transmit at random with probability p.
- Transmission is successful only if a single user transmits in a slot.
- If there is a collision, users must re-transmit their payload.

Slotted ALOHA: Analysis

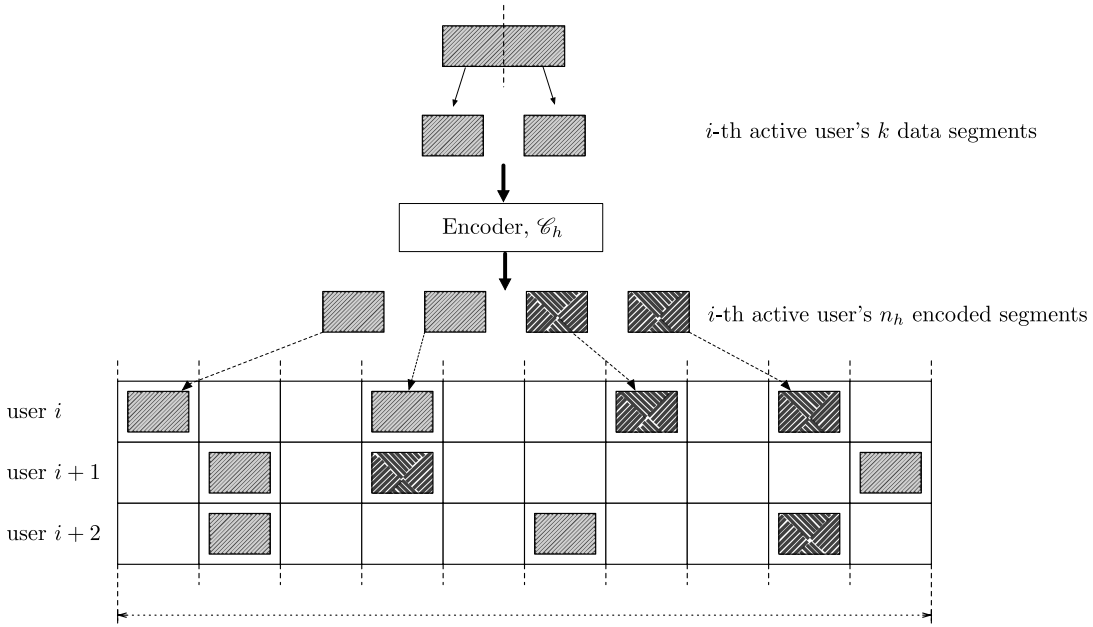


- Let X be the number of users that transmit in a slot.
- Since X is sum of independent Bernoulli trials, it follows Poisson distribution

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda k}}{k!}, \quad \text{where} \;\; \mathbb{E}[X] = \lambda. \tag{1}$$

- Successful transmission only when $k = 1$, with probability $\lambda e^{-\lambda}$.
- Optimize over λ. Throughput is maximized when $\lambda = 1$ with P(success) $= \frac{1}{e}$.
- Slots with collision or slots with no transmission (i.e., 63% slots) are wasted.

Coded Slotted ALOHA



- Coded Slotted ALOHA: Use packet-level erasure codes and successive interference cancellation (SIC) to extract information from collisions.
- Each user chooses an (n_h, k) erasure code $\mathscr{C}_h$ to encode their k segments.
- Code is chosen from a finite set $\{\mathscr{C}_h\}_{h=1}^{\theta}$ according to some p.m.f., and the n_h packets are transmitted randomly over a fixed frame.

E. Paolini, G. Liva, and M. Chiani, "Coded Slotted ALOHA: A Graph-Based Method for Uncoordinated Multiple Access," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6815–6832, 2015.

Coded Slotted ALOHA: Graph Representation

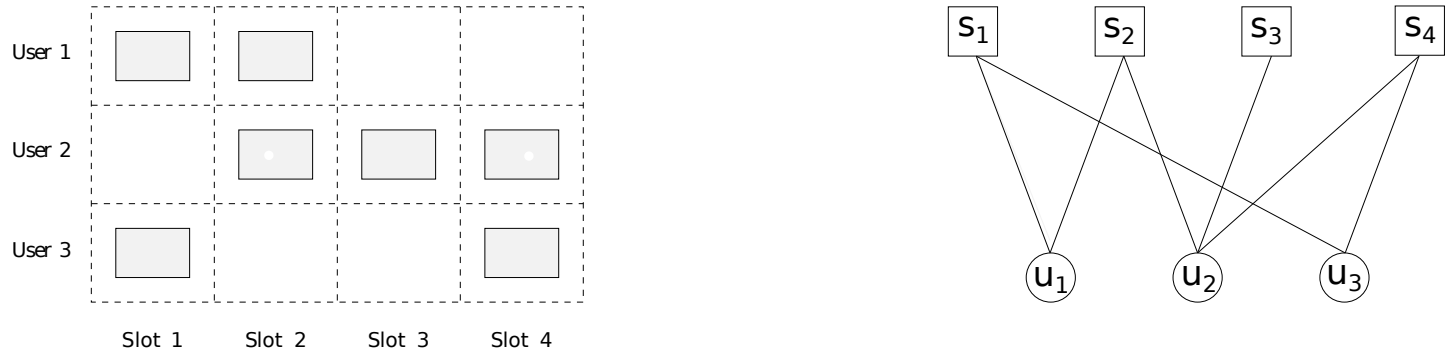


Figure: Bipartite graph model for contention resolution

- Users are represented by variable nodes, slots by check nodes.
- A user node u_i is connected to slot node s_j if user i transmits in slot j.
- Decoding process is identical to the peeling decoder for erasure channel.
- If users select repetition codes, this is known as Contention Resolution Diversity Slotted ALOHA (CRDSA).

Coded Slotted ALOHA: Decoding Example

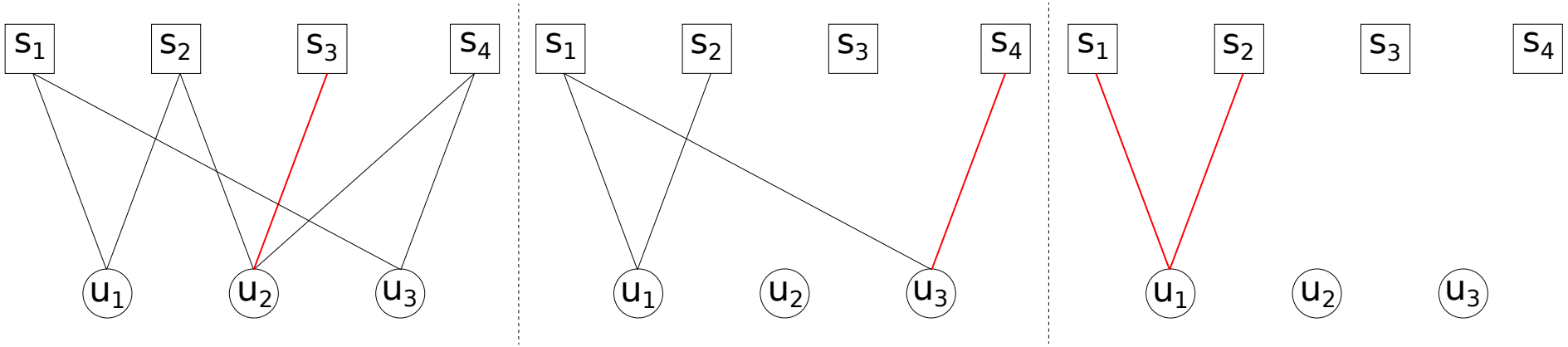
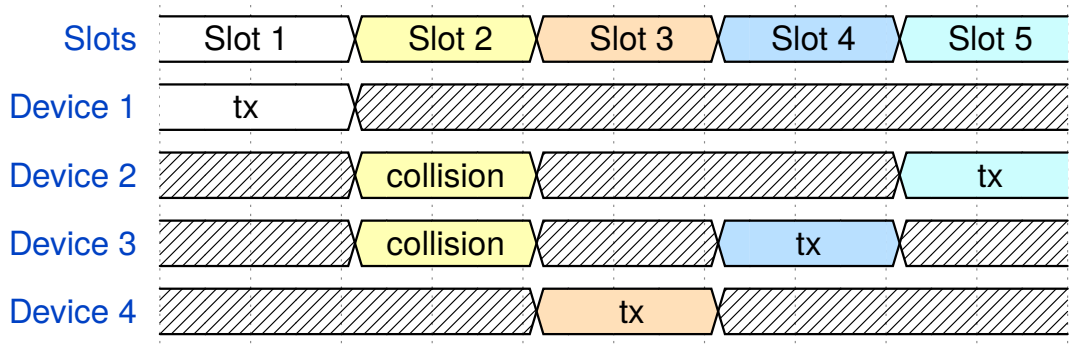


Figure: Peeling decoding for CRDSA on a bipartite graph.

- Decoding procedure for CRDSA is similar to Fountain code or LT code.
- This connection allows us to show that the optimal user-node degree distribution is the soliton distribution [Narayanan-Pfister'12].
- With this degree distribution, the throughput $\triangleq \frac{\#\text{ of decoded users}}{\#\text{ of slots}} \to 1$ asymptotically as the number of users and slots go to infinity.

Contention vs. Scheduling

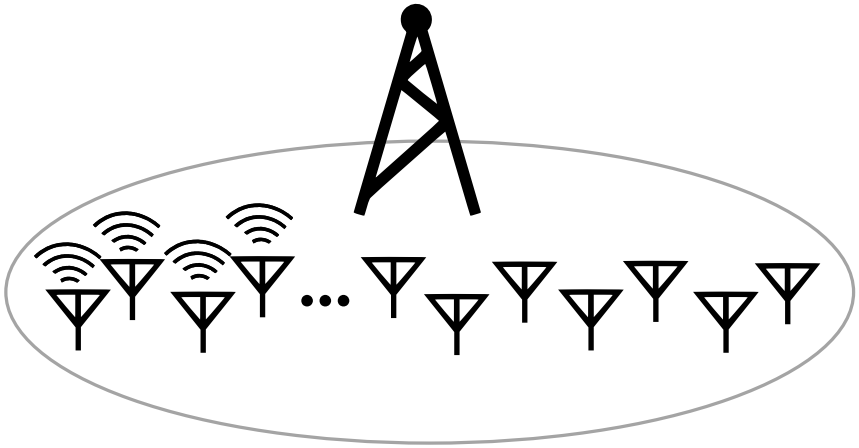


- Slotted ALOHA based schemes all involve contention and collision resolution
  - Multiple transmissions increases power consumption.
  - Collision resolution increases delay.
  - Practical schemes cannot operate at optimal throughput.
- Scheduling is an alternative approach to contention.
- Contention-based schemes are often justified based on the assumption that the cost of coordination is too great.

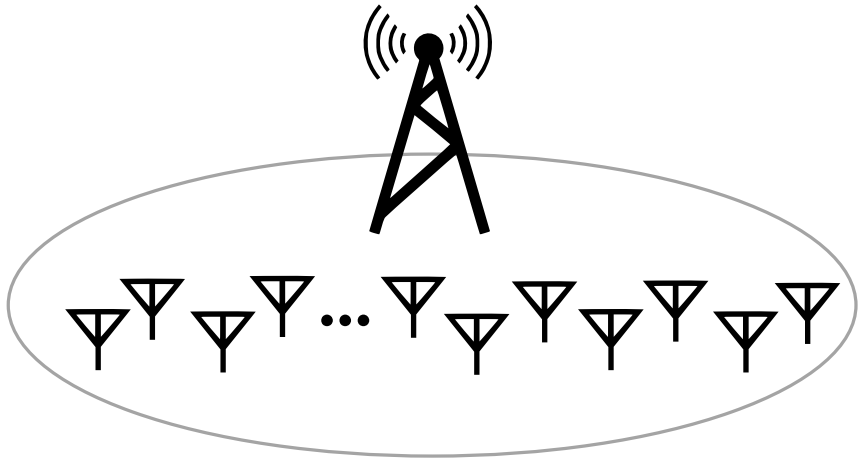
What is the cost of scheduling?

Feedback-Based Scheduling for Random Access

Each of n potential users is assigned a unique non-orthogonal pilot.

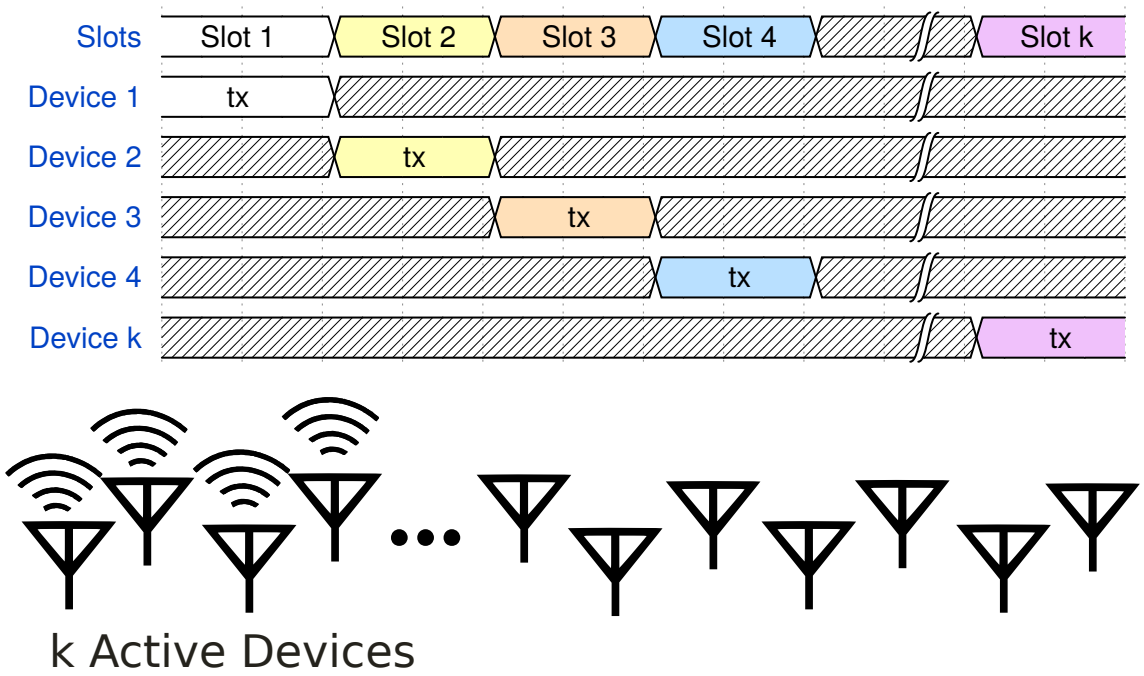


Phase 1 (Activity Detection):
The k active users ($k \ll n$) send their pilots synchronuously to the BS.



Phase 2 (Downlink Feedback): BS sends a *common* feedback message to schedule the data transmissions of k active users.

Feedback-Based Scheduling for Random Access



Phase 3 (Uplink Payload Transmission): The k active users transmit their payload in the k slots based on the schedule provided by the BS, while avoiding collision.

What is the minimum feedback needed to ensure collision-free scheduling?

Straightforward Feedback Scheme

- A naive scheme to schedule k out of n users:
  - Assign a unique index to each of the n users;
  - The BS detects the k active users based on the pilots;
  - The BS lists the k users in the order in which they should transmit;
  - Each active user finds its index in the list, waits for its turn to transmit.

- The feedback overhead of this scheme is $k \log(n)$ bits.
  - When $n = 10^6$, the cost of identification is $\log(n) = 20$ bits per user.

Can we do better?

Why Can We Do Better?

- The naive $k \log(n)$ feedback scheme is not optimal.
- There is flexibility in the order that users are scheduled.

  Example: Users $1, \ldots, k$ are to be scheduled. The BS can schedule according to any of the $k!$ permutations of these users, e.g. $\{1, \ldots, k\}$ or $\{k, \ldots, 1\}$.

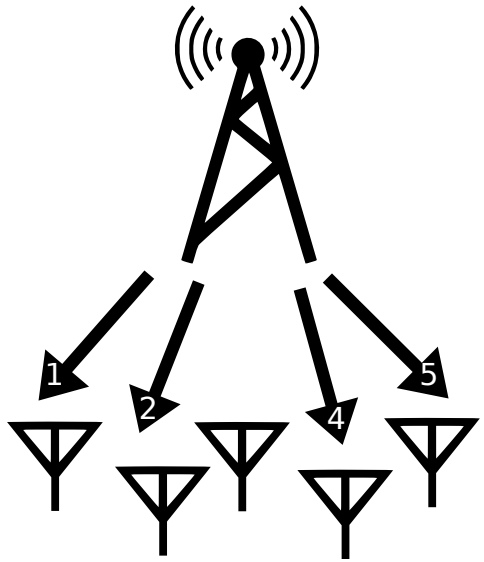
  We can remove this extraneous cost via *enumerative source coding*.

  This still requires $\log \binom{n}{k}$ bits feedback, which scales as $O(\log(n))$ for fixed k.
- Each user only needs to know its own slot, and NOT the other users' slots. Removing this extraneous information is the key to further reducing feedback.

G. K. Facenda and D. Silva, "Efficient Scheduling for the Massive Random Access Gaussian Channel," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7598–7609, Aug. 2020.

Identification Capacity

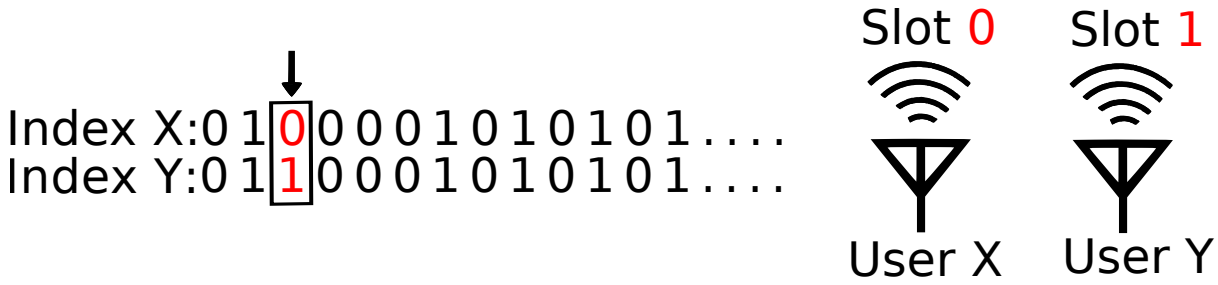
- *Identification via channels* [Ahlswede-Dueck, 1992] says that identifying one out of n users only requires $O(\log\log(n))$ bits! — This eliminates the extraneous information as users no longer know which other users are active.
- Identification codes lead to a feedback rate of $O\left(k\log\log(n)\right)$.



- A careful construction can beat even this scheme!

Two-User Case

- Consider the case of two active users $(k = 2)$, out of a total of n users:
- Any two distinct binary vectors differ in at least one index:



- BS simply transmits the location in which the user indices differ.
- The user with 0 transmits first, and the user with 1 transmits second.
- This requires only $R = \lceil \log \lceil \log (n) \rceil \rceil$ feedback with a fixed-length encoding.

Optimal for $k = 2$!

Feedback Scheduling Code for Arbitrary (n, k)

- Notation: $[n] = \{1, \cdots, n\}$. $\binom{[n]}{k} \triangleq$ set of all k-element subsets of $[n]$.
- The BS encodes the "activity pattern" into an index t

$$f : \binom{[n]}{k} \to \{1, 2, \ldots, T\} \triangleq [T].$$

- Each user "decodes" its scheduled slot using

$$g_i : [T] \to [k], \quad i \in [n].$$

(We consider k slots here, but having more slots can decrease feedback.)

- In order for no collisions between active users, we must have:

$$\forall \mathbf{A} \in \binom{[n]}{k}, \ \exists t \in [T] \ \text{ s.t. } \ \forall i \neq j \in \mathbf{A} \ \ g_i(t) \neq g_j(t).$$

Scheduling via Set-Partitioning

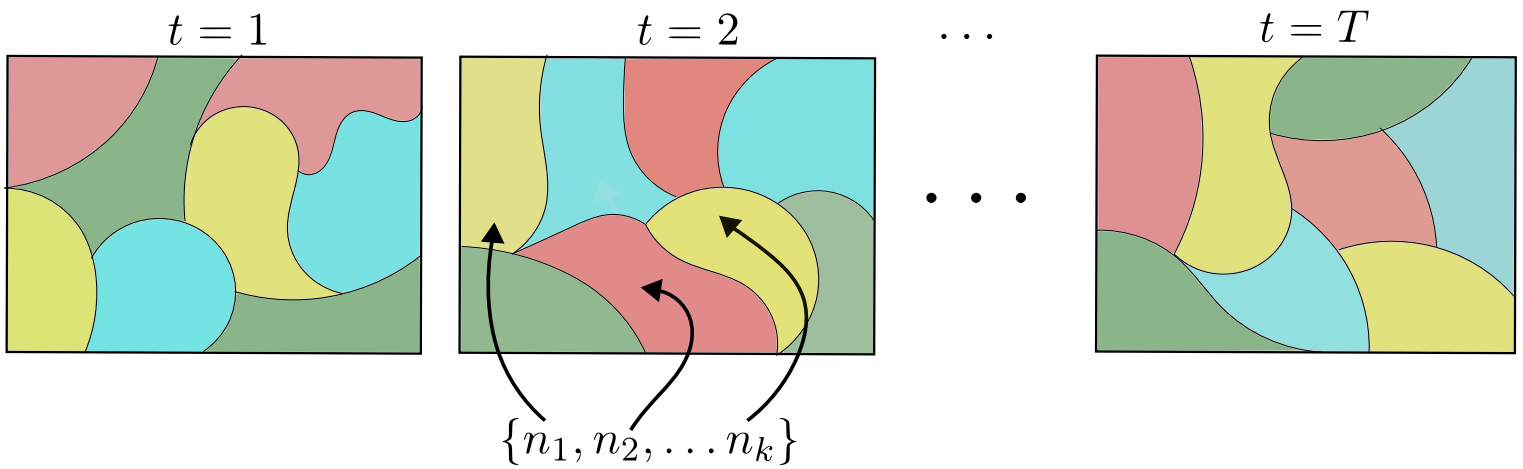
- Define a k-partition of a set $[n]$ to be a tuple of subsets $\bar{\mathbf{X}} = (\mathbf{X}_1, \ldots, \mathbf{X}_k)$ such that $\mathbf{X}_i \bigcap \mathbf{X}_j = \emptyset$, $\forall i, j$, and $\bigcup_{i=1}^{k} \mathbf{X}_i = [n]$.
- Define the set of activity patterns that can be covered by $\bar{\mathbf{X}}$ as

$$\mathbf{C}\left(\bar{\mathbf{X}}\right) = \left\{\{x_1, \ldots, x_k\} \mid x_i \in \mathbf{X}_i, i = 1, \ldots, k\right\}.$$

  i.e., there is exactly one active user in each distinct subset of the partition $\bar{\mathbf{X}}$.
- Example: For the set $[4]$, if $\bar{\mathbf{X}} = (\{1, 2\}, \{3, 4\})$, then

$$\mathbf{C}\left(\bar{\mathbf{X}}\right) = \left\{\{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}\right\}.$$

Set-Partition Encoding



- To cover all activity patterns, we construct T partitions $\bar{\mathbf{X}}^{(1)}, \ldots, \bar{\mathbf{X}}^{(T)}$ s.t.

$$\bigcup_{t=1}^{T} \mathbf{C}\left(\bar{\mathbf{X}}^{(t)}\right) = \binom{[n]}{k}.$$

- For activity pattern $\mathbf{A}$, the following encoder/decoders ensure no collision:

$$f(\mathbf{A}) = t \quad \text{s.t.} \quad \mathbf{A} \in \mathbf{C}\left(\bar{\mathbf{X}}^{(t)}\right);$$

$$g_i(t) = j \quad \text{if} \quad i \in \mathbf{X}_j^{(t)}.$$

Tetra Code: An Example for $(n, k) = (9, 3)$

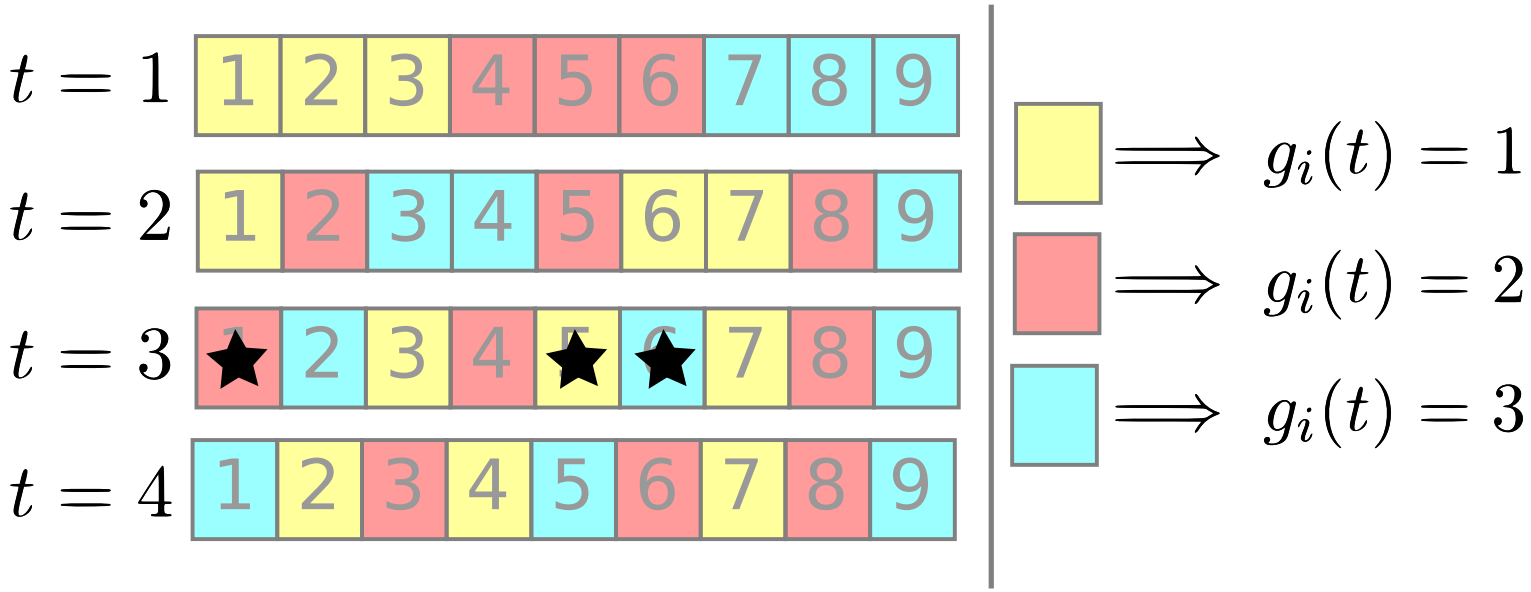


Figure: The tetra code can be used to define 4 partitions.

- Example: For the activity pattern $\mathbf{A} = \{1, 5, 6\}$, the $t = 3$ partition has all three active users in separate subsets, thus $f(\mathbf{A}) = 3$ ensures no collision.
- Only 2 bits of feedback as required! Optimal [Körner and Marton, 1988].

Set-Partition Encoding

- Any set of collision-free encoding and decoding function can be described with the set-partition framework.
- Given the decoding functions $g_i' : [T] \to [k]$, we can define T partitions $\bar{\mathbf{X}}^{(t)} = \left(\mathbf{X}_1^{(t)}, \ldots, \mathbf{X}_k^{(t)}\right)$, $t \in [T]$, where

$$\mathbf{X}_j^{(t)} = \{i \mid g_i'(t) = j, \quad i \in [n]\}.$$

- For a fixed-length feedback code, we define the feedback rate as

$$R_f^*(n,k) \triangleq \log(T^*)$$

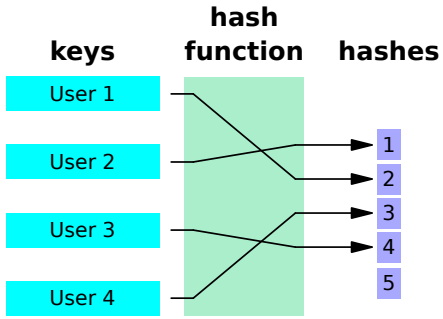
where T^* is the minimum number of partitions needed to cover all activity patterns.

Finding the minimum-rate zero-collision feedback code
now reduces to finding T^*.

Perfect Hashing Families

Finding T^* is equivalent to the *perfect hashing families* problem.

- An (n, b, k)-family of perfect hash functions is a family of functions from $[n] \to [b]$ for $n \geq b \geq k$ such that for every $\mathbf{A} \subset [n], |\mathbf{A}| = k$, there exists a function in the family that is injective on $\mathbf{A}$.



- We can view our decoding functions as a (n, k, k)-family perfect hash functions from $[n] \to [k]$ if we swap the argument and the subscript.

Theorem (Fredman and Komlós, 1984, Körner and Marton, 1988)

The minimum size T^* of an (n, b, k) perfect hash family is bounded as:

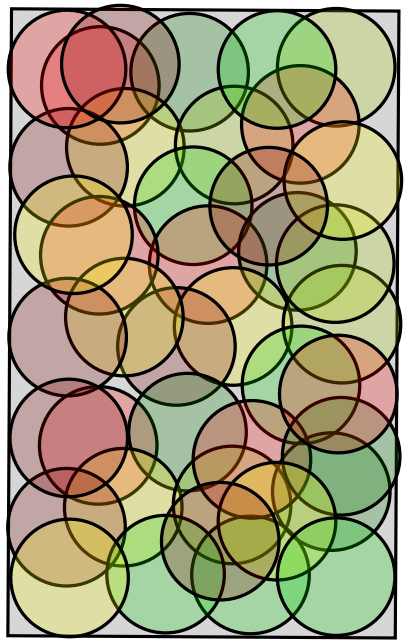
$$\frac{\log n}{\min_{1 \leq s \leq k-1} \frac{b^{\underline{s}}}{b^s} \log \frac{b-s+1}{k-s}} \lesssim T^* \lesssim \frac{(k-1)\log n}{\log \frac{1}{1-\frac{b^{\underline{k}}}{b^k}}}.$$

- The proof uses a notion of hypergraph entropy, but we can derive simpler, but still instructive bounds. Here, $b^{\underline{k}} \triangleq \frac{b!}{(b-k)!}$ is the falling factorial.

Random Partition Construction

- Take T random partitions $\bar{\mathbf{X}}^{(1)}, \ldots, \bar{\mathbf{X}}^{(T)}$, then the probability an activity pattern $\mathbf{A}$ is not covered is

$$\Pr\left(\mathbf{A} \notin \bigcup_{t=1}^{T} \mathbf{C}\left(\bar{\mathbf{X}}^{(t)}\right)\right) = \left(1 - \frac{k!}{k^k}\right)^T.$$



- By the union bound we have:

$$\Pr\left(\bigcup_{t=1}^{T} \mathbf{C}\left(\bar{\mathbf{X}}^{(t)}\right) \neq \binom{[n]}{k}\right) \leq \binom{n}{k}\left(1 - \frac{k!}{k^k}\right)^T.$$

- If the RHS of the above falls below 1, it means that there exists a family of partitions that cover all activity patterns.

Achievability Bound on Minimum Feedback Rate

- Using the fact $1 - x < e^{-x}$, we can show that the RHS falls below 1 for:

$$T \geq \left(\ln \binom{n}{k}\right)\left(\frac{k^k}{k!}\right).$$

Proposition

The minimum rate for a fixed-length collision-free feedback code must be upper bounded as:

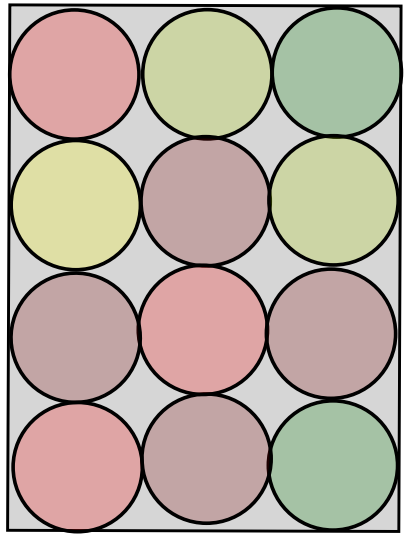
$$R_f^*(n,k) \triangleq \log(T^*) \leq k\log(e) + \log\left(\ln\left(\frac{n}{k}\right)+1\right) + \frac{1}{2}\log\left(\frac{k}{2\pi}\right).$$

Key observation: $R_f^*(n,k) \leq O(\log\log(n))$, plus a linear term in $k\log(e)$.

Converse: Volume Bound

- Since each partition can cover at most only a small fraction of the activity patterns, we can also place a volume bound on the covering:

$$T^* \geq \frac{\binom{n}{k}}{\left\lceil \frac{n}{k} \right\rceil^{n \bmod k} \left\lfloor \frac{n}{k} \right\rfloor^{k-n \bmod k}}.$$



Proposition

The minimum rate for a fixed-length collision-free feedback code must be lower bounded as:

$$R_f^*(k,n) \geq k \log(e) - \log\left(\frac{n^k}{n(n-1)...(n-k+1)}\right) - \frac{1}{2}\log(2\pi k) - \frac{\log(e)}{12k}.$$

Thus, $R_f^*(n,k) \geq O(k)$.

Converse: Exclusion Bound

- A partition $\bar{\mathbf{X}}^{(1)}$ cannot have covered any activity pattern which has all its elements drawn from $\mathbf{S_1} = [n] - \mathbf{X}_j^{(1)}$, as

$$\mathbf{C}\left(\bar{\mathbf{X}}^{(1)}\right) \cap \binom{[n] - \mathbf{X}_j^{(1)}}{k} = \emptyset, \quad j = 1, \ldots, k.$$

  i.e., activity patterns with indices exclusively drawn from $\mathbf{S_1}$ are *excluded*.

- Since one of the partitions $\mathbf{X}_j^{(i)}$ is at most size $\left\lfloor \frac{n}{k} \right\rfloor$, we have:

$$|\mathbf{S_1}| = m_1(n,k) \geq n\left(1 - \frac{1}{k}\right).$$

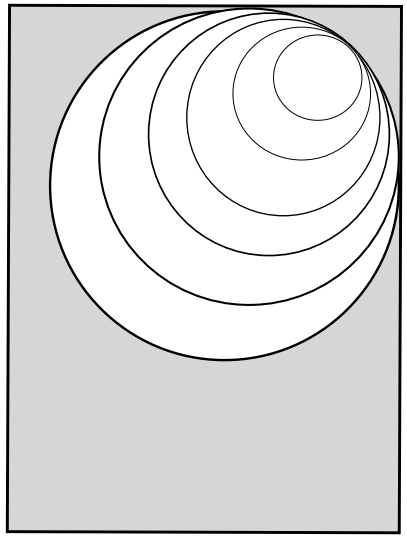
Exclusion Bound

- By repeated application of the exclusion argument:

$$m_t(n,k) \geq n\left(1-\frac{1}{k}\right)^t.$$

- For this exclusion set to shrink down to the null set (not containing any activity pattern), we need

$$n\left(1-\frac{1}{k}\right)^T \leq k-1.$$



With each partition, the exclusion region shrinks.

Proposition

The minimum rate for a fixed-length collision-free feedback code must be lower bounded as:

$$R_f^*(n,k) \geq \log\log\left(\frac{n}{k-1}\right) + \log(k) - 1.$$

From Fixed to Variable Length Feedback Code

Fixed-length collision-free feedback code:

- Random Partition: $R_f^*(n,k)$ scales at most as $k\log(e)$ plus $O(\log\log(n))$.
- Volume Bound: $R_f^*(n,k)$ scales at least as $k\log(e)$ for large n.
- Exclusion Bound: $R_f^*(n,k)$ scales at least as $\Omega(\log\log(n))$ for fixed k.

Thus, rate of fixed-length code scales linearly as $k\log(e)$ and as $\Theta(\log\log(n))$.

Can we do better?

Variable-length collision-free feedback code:

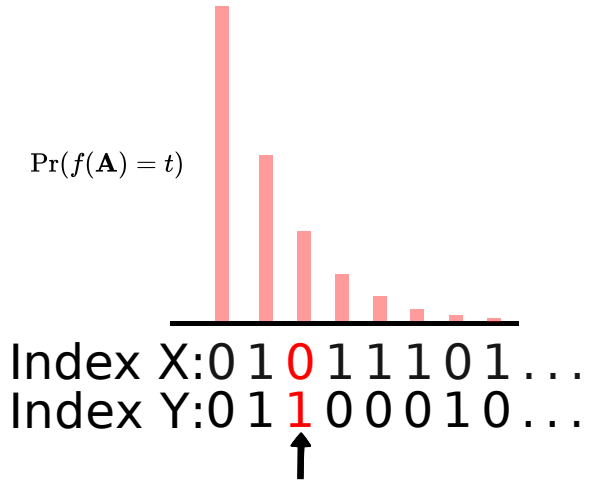
- Treat $\mathbf{A}$ as a random variable with distribution $Q(\mathbf{A})$ and define $R_v(n,k) \triangleq \mathrm{H}(f(\mathbf{A}))$, corresponding to optimal entropy coding.
- Focusing on the worst-case activity distribution, define the optimal rate as:

$$R_v^*(n,k) \triangleq \sup_{Q(\cdot)} \mathrm{H}(f(\mathbf{A})).$$

- It turns out we can remove even the $\Theta(\log\log(n))$ growth in n.

Greedy Encoding for $k = 2$

- Consider the index based feedback strategy for $k = 2$, but now greedily choose the first position where user indices differ.
- If the user activity is the worst-case uniform distribution, $f(\mathbf{A})$ follows a truncated geometric distribution.



- A direct application of Huffman Coding results in a code of rate:

$$R_v(n, 2) = 2 - \frac{\log(n) + 1}{n - 1}.$$

- This implies $\lim_{n\to\infty} R_v^*(n, 2) \leq 2$, thus the achievable feedback rate remains bounded as n tends to infinity.

Greedy Encoding for $k > 2$

- We again use the concept of greedy encoding strategy. Given a family of T k-partitions $\underline{\mathbf{B}} = (\bar{\mathbf{X}}^{(1)}, \ldots, \bar{\mathbf{X}}^{(T)})$, define the greedy encoder $f_{\underline{\mathbf{B}}}$:

$$f_{\underline{\mathbf{B}}}(\mathbf{A}) = \min_{t \in [T]} t, \quad \text{s.t.} \quad \mathbf{A} \in \mathbf{C}\left(\bar{\mathbf{X}}^{(t)}\right), \text{ else } \quad T+1,$$

  and the resulting distribution $p_{\underline{\mathbf{B}}}(t) \triangleq \Pr(f_{\underline{\mathbf{B}}}(\mathbf{A}) = t)$.
- Denote the set of all families of k-partitions of size T as $\mathcal{B}$, regardless of whether each of them covers all activity patterns, or not.
- Consider an encoder that chooses $\underline{\mathbf{B}}$ uniformly at random from $\mathcal{B}$. Define $p_{\mathcal{B}}(t) \triangleq \mathbb{E}_{\mathcal{B}}\left[p_{\underline{\mathbf{B}}}(t)\right]$. The first T terms in this distribution are:

$$p_{\mathcal{B}}(t) = \frac{k!}{k^k}\left(1 - \frac{k!}{k^k}\right)^{t-1}, \quad t = 1, \ldots, T,$$

  with the remainder of the mass at $T+1$, regardless of the distribution of $\mathbf{A}$.

Variable-Length Feedback Bounds

- With Jensen's inequality, this implies the following bound independent of T:

$$\mathbb{E}_{\mathcal{B}}\left[\mathrm{H}\left(p_{\underline{\mathbf{B}}}(t)\right)\right] \leq \mathrm{H}\left(p_{\mathcal{B}}(t)\right) \leq (k+1)\log(e).$$

- For families of partitions of size T, let $1-\epsilon$ be the fraction of collision-free families in $\mathcal{B}$, then the rate for collision-free feedback can be bounded as:

$$R_v^*(n,k) \leq \frac{1}{1-\epsilon}(k+1)\log(e)$$

- Now, we can let $T \to \infty$, so $\epsilon \to 0$, implying $R_v^*(n,k) \leq (k+1)\log(e)$.
- The volume bound converse can also be extended to variabe-length codes.

Theorem

The minimum rate for variable-length collision-free feedback code is bounded as

$$(k+1)\log(e) \;\geq\; R_v^*(n,k) \;\geq\; k\log(e) - \log\left(\frac{n^k}{n^{\underline{k}}}\right) - \frac{1}{2}\log(2\pi k) - \frac{\log(e)}{12k}.$$

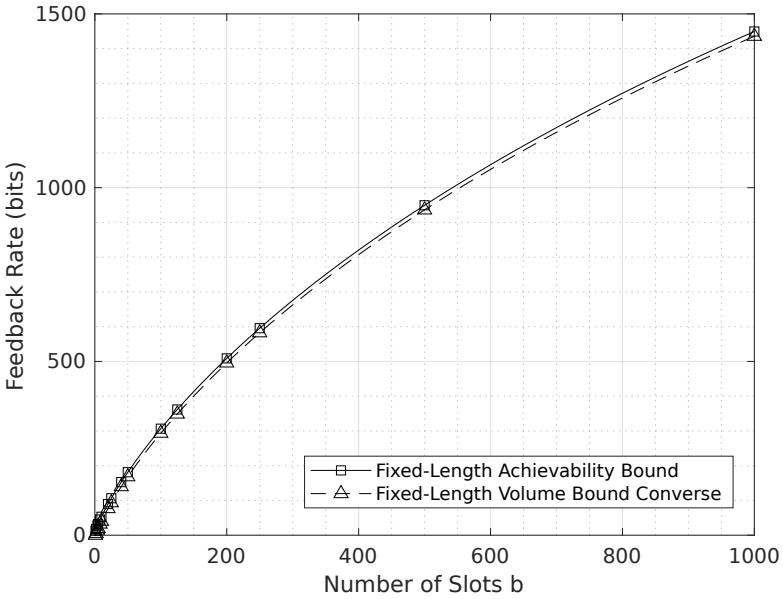
Practical Implementations

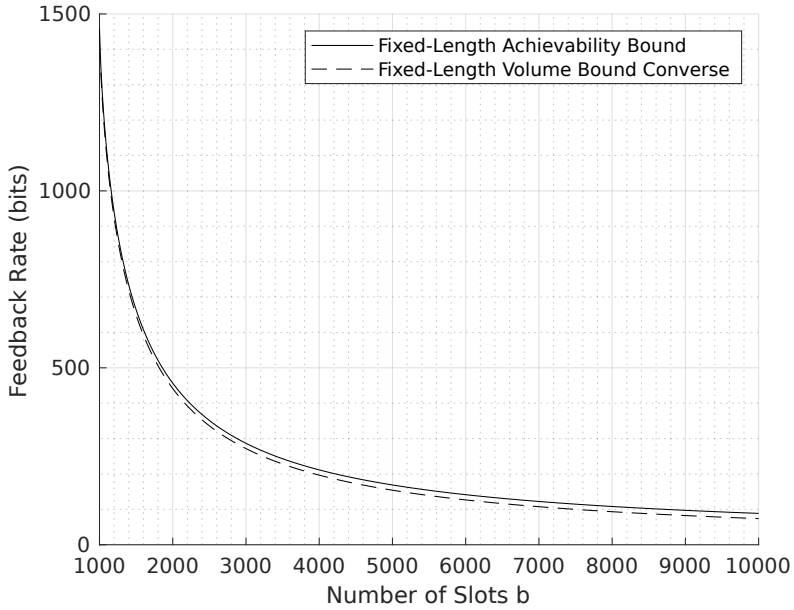
- Consider a system with $n = 10^6$ potential users and $k = 10^3$ active users:
  - Naive scheme would require **20 kbits**
  - Enumerative source coding requires **11.5 kbits**.
  - Optimal feedback only needs approximately **1.5 kbits**.
- Some practical schemes come close to achieving the $k \log(e)$ linear scaling:

Table: Practical Hashing/Feedback Algorithms

Method	Bits Per User
Random Coding	1.44
Boolean SAT	1.83
Compress-Hash-Displace	2.07

More Slots and Multiple Users per Slot





- These bounds can be extended to the case of:
  - $b \geq k$ slots (over-provisioned system), and
  - $b \leq k$ slots for systems where the BS can decode multiple users per slot.

Summary

What is the cost of coordinating collision-free scheduling?

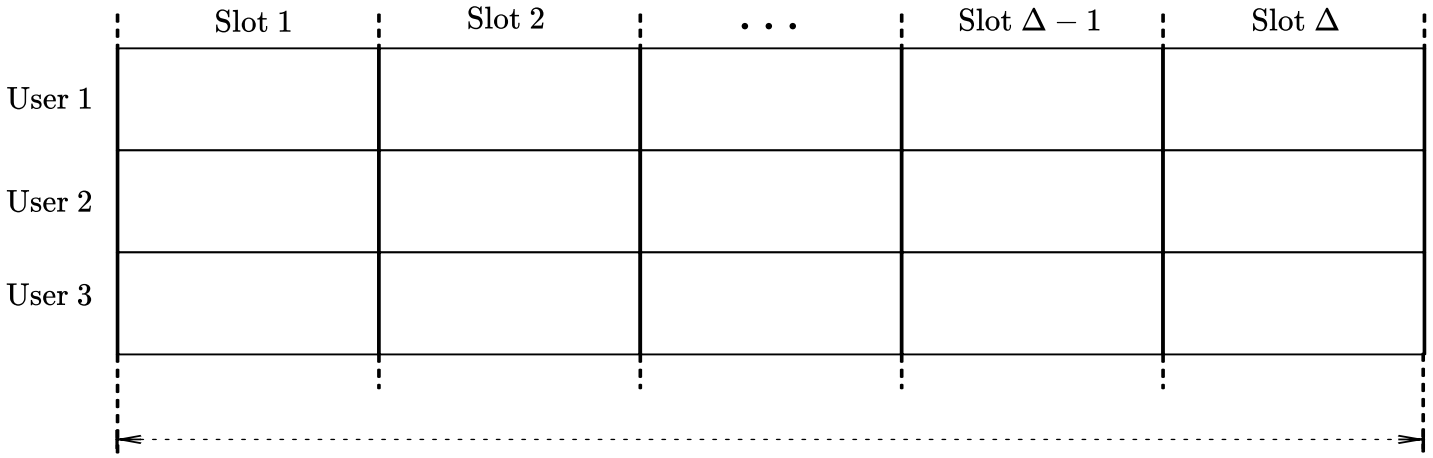
- Fixed-length feedback codes for collision-free scheduling of k active users among n potential users into k slots requires a rate of approximately $k \log(e)$ bits, plus a $\Theta(\log \log(n))$ term.
- Using variable-length feedback codes can reduce the required feedback rate for collision-free scheduling to $(k+1) \log(e)$ bits, independent of n.
- If $b \geq k$ slots are available, or more than one user can be decoded per slot, feedback can be further reduced.

Random Access for Massive MIMO Systems

1. Uncoordinated Random Access for Massive MIMO
   - Channel estimation and data transmission must both be without coordination.
   - Coded ALOHA can be adapted to Massive MIMO systems to enable uncoordinated communication.
   - We will consider a variant of coded ALOHA known as *Coded Pilot Access*.

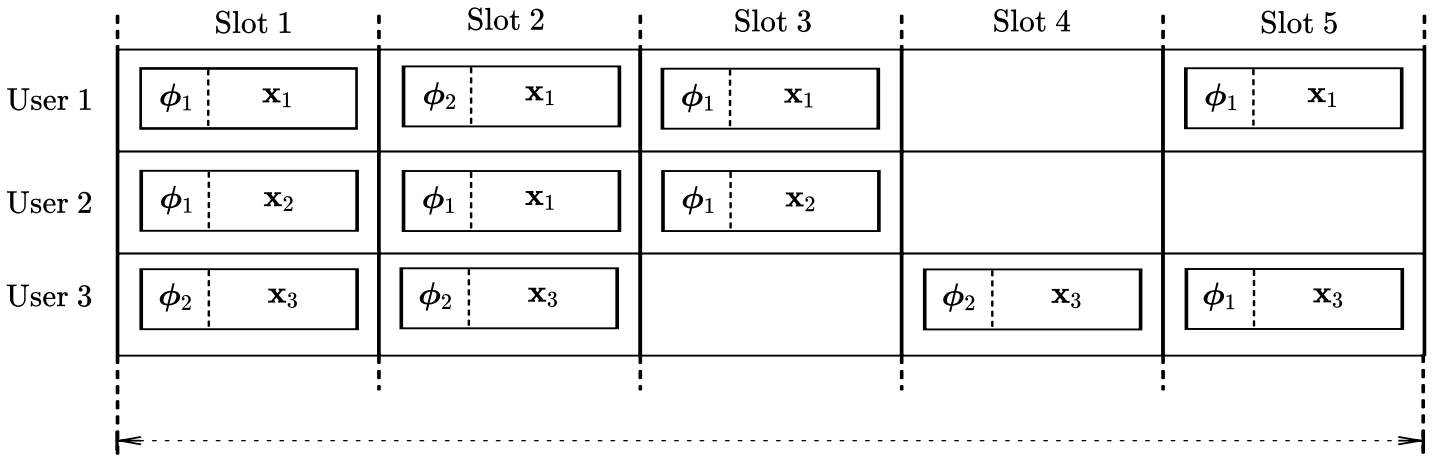
2. Scheduled Random Access for Massive MIMO
   - Activity detection can serve as an initial step for scheduled random access.
   - A relatively small amount of feedback can be used to ensure collision-free scheduling for the users.
   - Users are assigned orthogonal pilots for channel estimation.

Slotted Random Access



- The BS is equipped with M antennas.
- There are n single-antenna devices k of which are active.
- Active users transmit across Δ temporal slots each containing L symbols.
- The channels $\mathbf{h}_{d,i} \sim \mathcal{CN}(0,1)$ is i.i.d for each user i in the d^{th} slot. We assume users apply inverse power control to compensate for large scale fading.
- The BS uses the received signal $\mathbf{Y}_d$ over Δ slots to decode the messages of k active users.

Coded Pilot Access

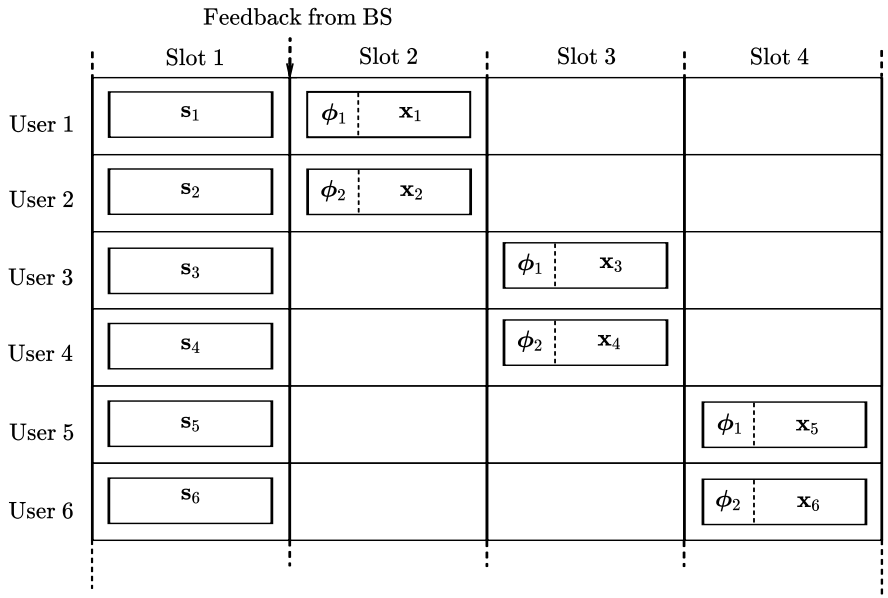


- Users transmit their payload $\mathbf{x}_i$ multiple times, each time preceded by a pilot randomly selected from a set of orthogonal pilots $\{\boldsymbol{\phi}_t\}_{t=1}^{\tau}$.
- In cases with no collision, the BS can perform channel estimation and data decoding for that user.
- The data contains the location of the other slots where the user has transmitted, allowing the BS to perform SIC.

J. H. Sørensen, E. De Carvalho, Č. Stefanović, and P. Popovski, "Coded Pilot Random Access for Massive MIMO Systems", *IEEE Trans. Wireless Commun.*, vol.17, no.12, pp.8035–8046, 2018.

Scheduled Random Access for Massive MIMO

- Users first transmit non-orthogonal pilots $\mathbf{s}_i \in \mathbb{C}^L$ for activity detection.
- BS sends scheduling message.
- Each user is assigned a unique (slot, orthogonal pilot) pair based on common feedback from the BS.



- The BS performs channel estimation using the orthogonal pilots, and then maximum ratio combining to reconstruct the payload.
- Each user is only required to transmit twice, in contrast to Coded ALOHA.

Scheduled Random Access vs. Coded Pilot Access

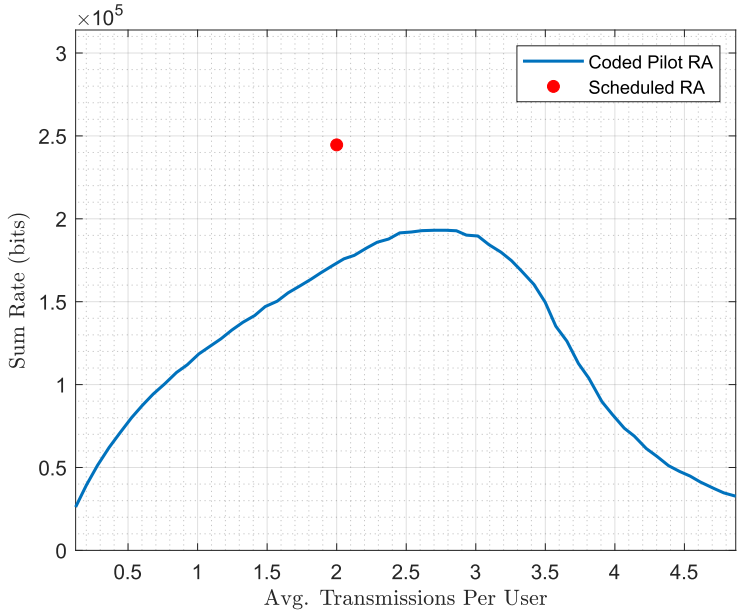


Figure: Throughput of Coded Pilot Access and Scheduled Random Access with $k = 1000$, $n = 10000$, SNR $= 10$dB, $M = 400$ BS antennas, $\tau = 64$ orthogonal pilots

- Each slot consists of $L = 300$ symbols.
  - Number of slots $\Delta = 20$ for coded pilot access;
  - Number of slots $\Delta = 17$ for scheduled random access.
- Activity detection is done via covariance method over one slot for SRA.
- Sum rate calculation assumes MRC beamforming and perfect SIC for CPA.
- Sum rate gain of 50 kbits at moderate cost of 1.44 kbits of feedback.

Conclusions

- Classic random access is contention based.
- Coded random access can alleviate some of the loss due to collision.
- If feedback is available from BS to the users:
  - BS can first detect the active users using sparse recovery methods;
  - BS can then schedule orthogonal pilots to users for channel estimation;
  - Finally, the users transmit their data to the BS.
- Significant performance improvement can be obtained at moderate feedback of 1.44 bits/user for scheduling.

Further Information

Justin Kang and Wei Yu,

"Minimum Feedback for Collision-Free Scheduling in Massive Random Access",

Submitted to IEEE Transactions on Information Theory, 2020.

[Online] available: https://arxiv.org/abs/2007.15497.