UNIVERSITY OF OREGON

Learning Efficient Sparse Encoding for High-Performance Tensor Decomposition

Subtitle: Can AI Beat Humans in Sparse Computation

Jee Whan Choi

University of Oregon
March 4th, SIAM PP 2026

Motivation

Sparse computation is difficult – structure of the data becomes (a big) part of the problem

Motivation

Sparse computation is difficult – structure of the data becomes (a big) part of the problem

- Non-contiguous data makes accessing data and parallel computation less efficient
- Overhead of indexing and metadata
- (Parallel) Algorithmic and performance modeling complexity
- Under-utilization of specialized hardware resources (e.g., tensor cores, prefetchers)
- And more

Motivation

What can we do?

- Specialized data structures/formats (e.g., CSR, ELL, DIA, etc.), if data is structured
- What if we (human) can't see the pattern? **Can AI come to our aid**?
- We want to answer this question for **sparse tensor computation**

Popularity of tree- and block-based formats

- Tree- and block-based formats are extensions of sparse matrix formats
- However, they are ill-suited for sparse tensors

1. Tensor computation often operate over every dimension
   - This contrasts with sparse matrix-vector multiply (*one dimension*) and sparse matrix-matrix multiply (*two dimensions*)
2. Tensor sparsity >> matrix sparsity
   - Due to their dimensionality, tensors are extremely sparse – it's difficult to find dense blocks
   - Tensor sparsity ranges from 1.5×10^{-2} to 4.3×10^{-15} [7]

Our hypothesis – *simple mode-agnostic, list-based formats are the best for sparse tensor computation*

Linearized Formats

- ALTO (Adaptive Linearized Tensor Order) for CPUs [2]
- BLCO (Block Linearized Coordinate) for GPUs [3]
- Application of linearized formats to
  - *streaming* tensor decomposition [4]
  - *on-the-fly* Khatri-Rao product for CP-APR [6]
  - non-negative sparse tensor factorization for GPUs via PLANC [5]

ALTO – Adaptive Linearized Tensor Order

Tammy Kolda

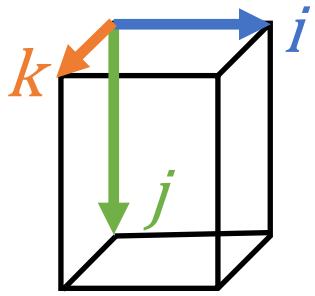
T. G. Kolda, B. W. Bader, ”**Tensor Decompositions and Applications,”** *SIAM Review*, 2009.

Harrison, Adam P., and Dileepan Joseph, "High Performance Rearrangement and Multiplication Routines for Sparse Tensor Arithmetic." *SIAM Journal on Scientific Computing,* 2018

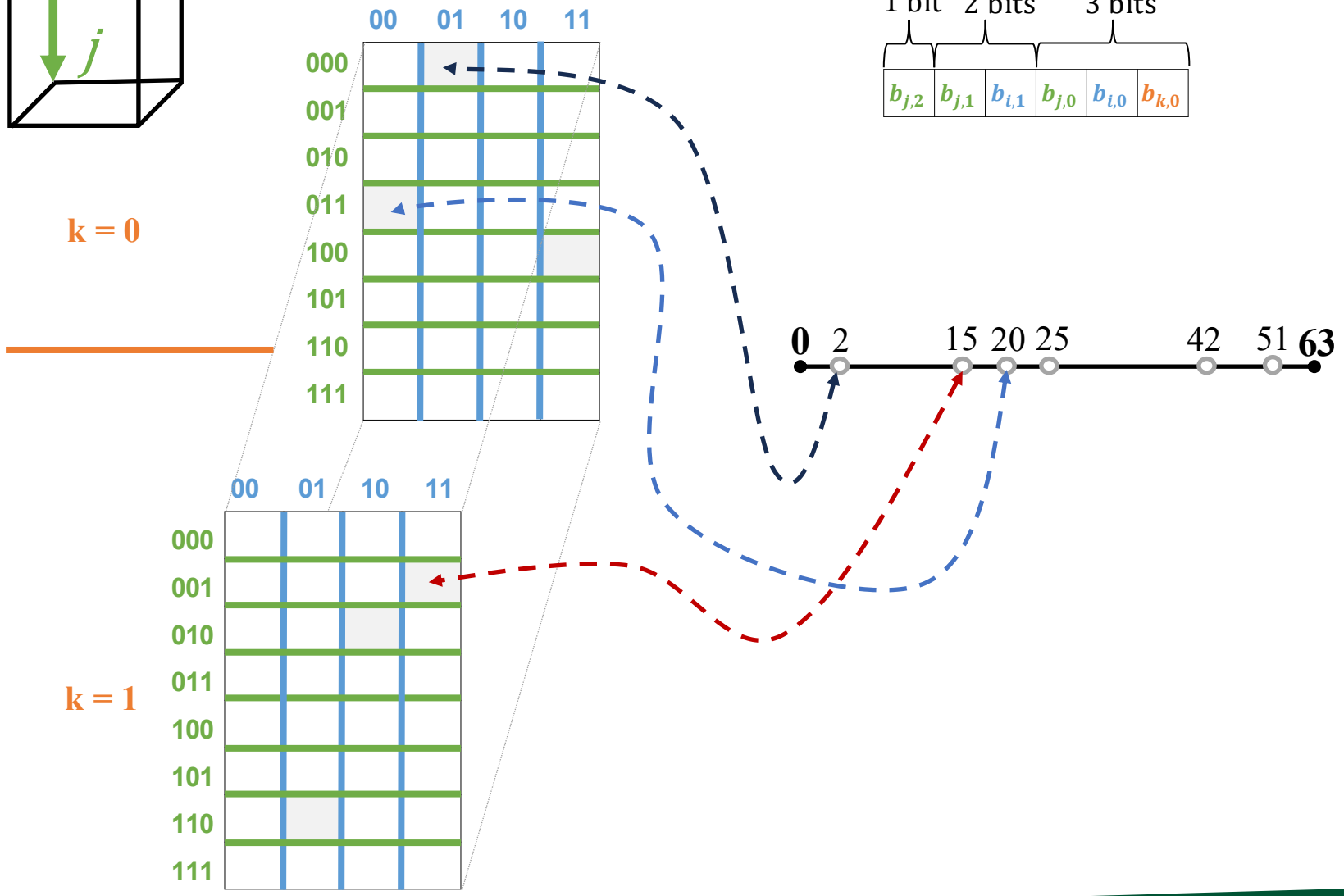
ALTO

ALTO

Value	Position
$x_{1,0,0}$	2 (000010)
$x_{3,1,1}$	15 (001111)
$x_{0,3,0}$	20 (010100)
$x_{2,2,1}$	25 (011001)
$x_{3,4,0}$	42 (101010)
$x_{1,6,1}$	51 (110011)



4x8x2 tensor



ALTO bit mask

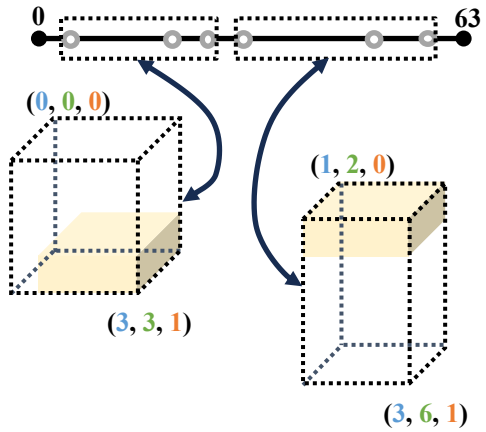
Matricized Tensor Times Khatri-Rao Product (MTTKRP)

```
for l = 1,…,L in parallel do
    for all x ∈ X_l do
        fetch i_x,j_x,k_x,v_x
        fetch v_1 = A(2)(j_x,:)
        fetch v_2 = A(3)(k_x,:)
        scratch += v_x · (v_1 * v_2)
        A(1)(i_x,:) += scratch
    endfor
endfor
```

ALTO line segment

```
fetch a_x (ALTO index),v_x
i_x,j_x,k_x = decode(a_x)
```

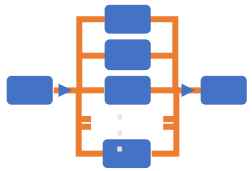
This can be done efficiently using parallel bit extract/deposit (`pext`/`pdep`) instructions on x86 CPUs



Adaptive update mechanism

- If rows (i_x) have limited reuse → update using atomic ops.
- If rows (i_x) have large reuse → use scratchpad to combine local updates and parallel reduction to merge globally
- `reuse = nnz / mode_length`
- Global reduction uses the boundary information to "pull" rows (i_x) from the appropriate scratchpad

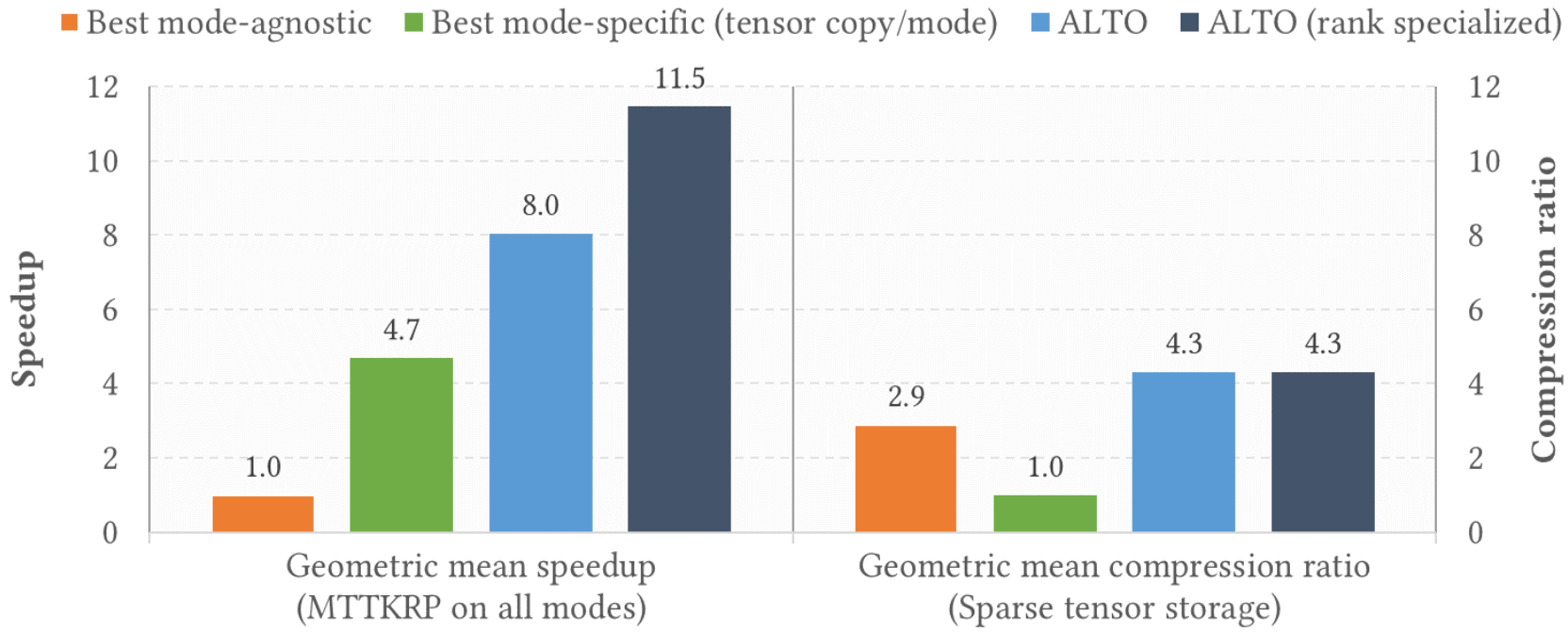
ALTO vs. State-of-the-art


Formats	Granularity	Mode Orientation	Data Locality	Parallelism	Load balance
List-based (COO)	Nonzero element	Mode-agnostic	Poor	Suffer from conflicts	Maximized
Tree-based (CSF)	Compressed tree	Mode-specific	Improved for a specific mode	Improved for a specific mode	Work imbalance (especially in short modes)
Block-based (HiCOO)	Compressed block	Mode-agnostic	Improved	Suffer from conflicts	Work imbalance across blocks
ALTO	Nonzero element	Mode-agnostic	Improved (via nearest-neighbor traversal)	Improved (via adaptive update)	Maximized

Performance Summary

- Intel Cascade Lake-X
  - 28 x 2 cores @ 1.8 GHz
- Oracle selects the best mode-agnostic and mode-specific format for each of the 15 tensors
- Mode-agnostic formats: COO and HiCOO
- Mode-specific formats: CSF and CSF with tiling with N copies



Matricized Tensor Times Khatri-Rao Product (MTTKRP)

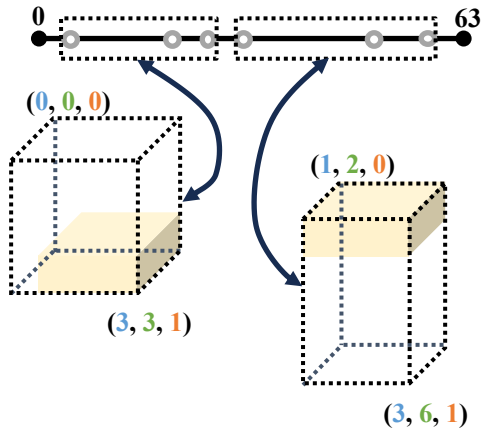
```
for l = 1,…,L in parallel do
        for all x ∈ X_1 do
                fetch i_x,j_x,k_x,v_x
                fetch v_1 = A^(2)(j_x,:)
                fetch v_2 = A^(3)(k_x,:)
                scratch += v_x · (v_1 * v_2)
                A^(1)(i_x,:) += scratch
        endfor
endfor
```

ALTO line segment

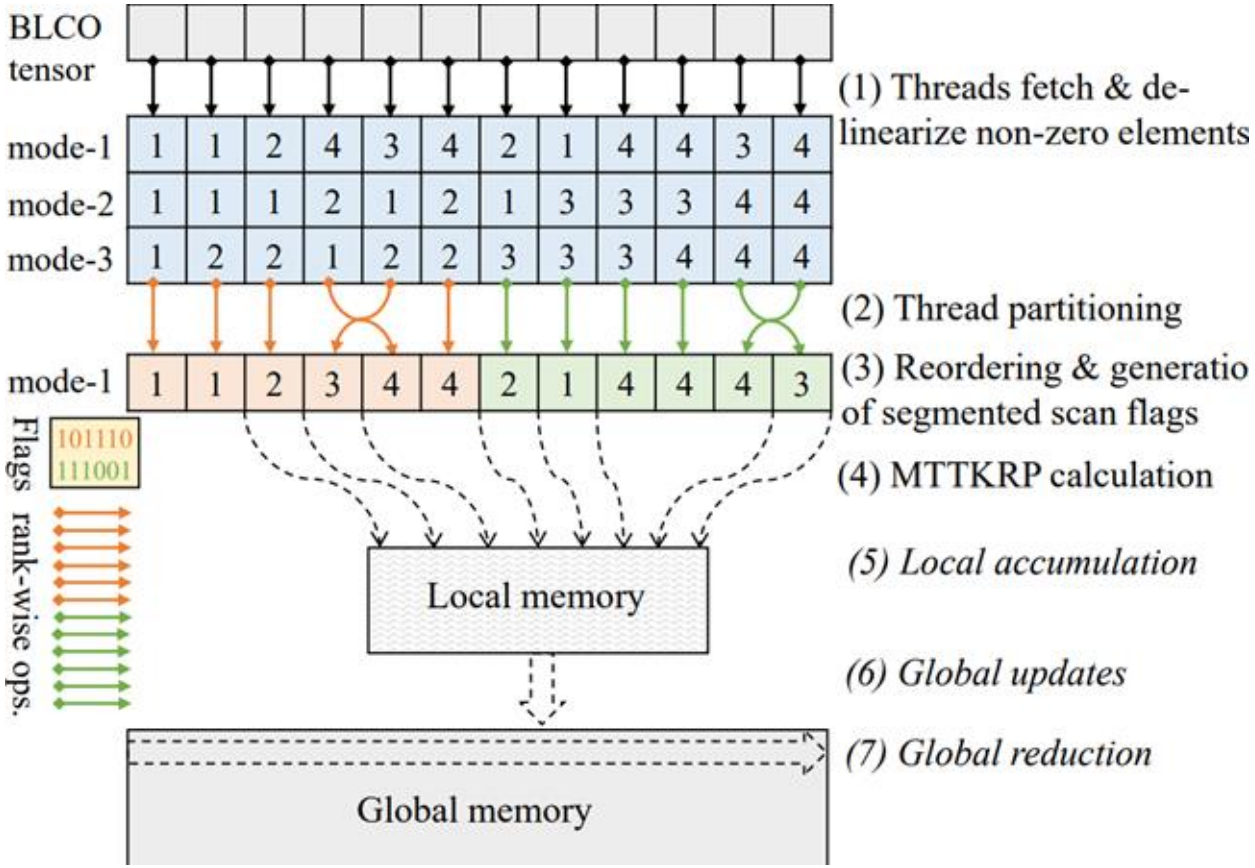
```
fetch a_x (ALTO index),v_x
i_x,j_x,k_x = decode(a_x)
```

GPU do NOT provide bit-level scatter/gather operations!

Synchronization is more expensive on GPUs!

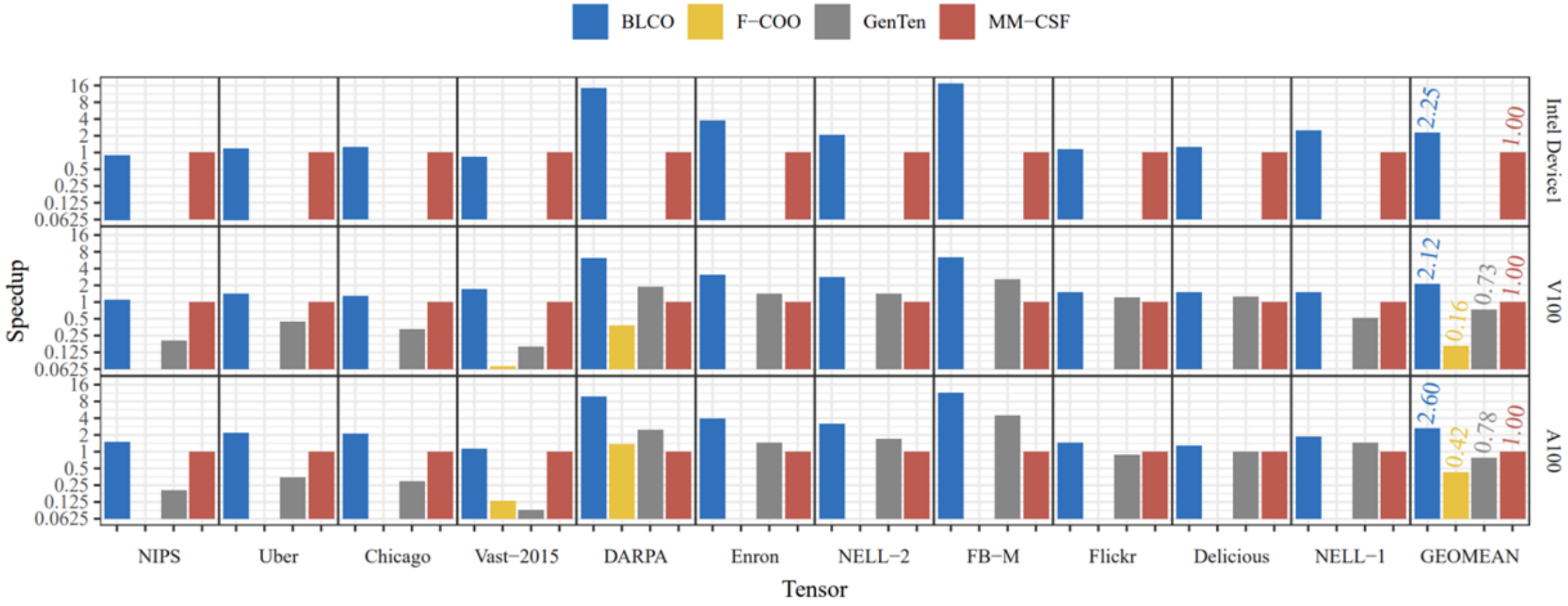


GPU - Synchronization



Performance

- MM-CSF is the baseline
- `Amazon`, `Patents`, and `Reddit` do not run on any prior frameworks
- F-COO only supports 3D tensors and `segfaults` on some tensors



Performance

- Consistent memory throughput for every mode

Data Set	Format	n	Vol[1]	TP[2]	Data Set	Format	n	Vol[1]	TP[2]
Uber	BLCO	1	2.78	3.60	Enron	BLCO	1	44.82	4.11
		2	2.75	3.61			2	46.23	4.62
		3	2.75	3.53			3	47.88	4.92
		4	2.73	2.77			4	47.22	4.70
	MM-CSF	1	1.68	1.68		MM-CSF	1	41.39	0.31
		2	1.33	2.03			2	62.83	3.16
		3	1.33	1.93			3	37.15	2.29
		4	2.12	0.32			4	37.05	3.01
Vast-2015	BLCO	1	16.91	3.92	NELL-1	BLCO	1	107.5	2.44
		2	16.73	3.77			2	104.5	2.32
		3	13.92	2.90			3	110.7	2.39
	MM-CSF	1	9.19	1.19		MM-CSF	1	123.1	2.21
		2	8.36	1.57			2	118.5	2.19
		3	8.36	1.45			3	122.1	0.86

[1] Memory volume in GB, measured by *l1tex__t_bytes.sum* in Nsight Compute [3]
[2] Memory throughput in TB/s, calculated by (Vol / total execution time)

Can we make this better?

- Different sparsity pattern would benefit from a different traversal order
  - Different space filling curves? Hilbert, Peano, etc.?
- What about AI? Can they produce a traversal order customized for the input tensor's sparsity pattern?
  - What type of AI model?
  - How do we generate/gather training data?
- What if you asked ChatGPT?

Can we make this better?

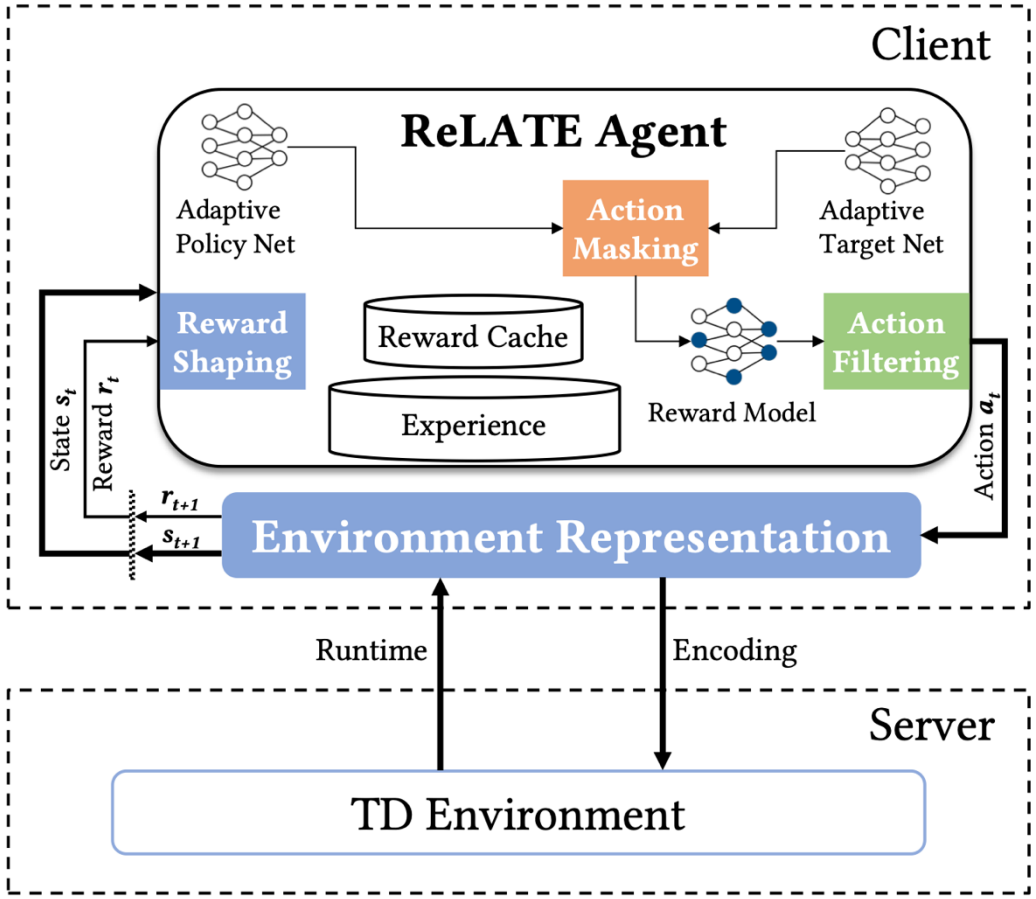
- What if you asked ChatGPT?
  - **“Can you design a sparse tensor format for MTTKRP that's better than ALTO?”**

Can we make this better?

- What if you asked ChatGPT?
  - **“Can you design a sparse tensor format for MTTKRP that's better than ALTO?”**
  - Result: A mode-specific tree-structured format like CSF with tiling (i.e., tiled CSF)
    - Admits that they are similar under specific conditions (when tiling size is well-chosen and mode ordering is optimal) and performance difference maybe within 10 – 20%.
    - Requires N copies like CSF
    - However, claims to be better than ALTO
  - It did a good job of justifying its decision choices (e.g., that majority of the data comes from the factor matrices, and grouping by common index to maximize reuse, etc.)
  - It seems to be good at analyzing ”static” metrics (i.e., flops, # of memory requests, etc.)
  - However, it was unable to account for variation in input pattern and architectural properties
  - So, maybe not quite there yet

ReLATE – Reinforcement Learned Adaptive Tensor Encoding

- Deep Reinforcement Learning (DRL)-based framework for deriving a linearized encoding for sparse tensors
- Key features
  - **Client-server model**
  - Double DQN
  - Reward cache
  - Prioritized Replay Buffer
  - **Hybrid model-free and model-based learning**
  - **Reward shaping mechanism**



Problem Formulation

- Bit interleaving
  - Different bit interleaving corresponds to a different linearized traversal of the tensor space
  - Total number of ways for I^N space = $(N \log_2(I))!$
- Bit from the index tuple -> a bit location in the linearized index
  - Sequential decision problem -> MDP -> **reinforcement learning**
- Still very expensive
  - 256 x 256 x 256 x 256 tensor -> $(4 \times 8)! = 2.63 \times 10^{35}$
  - How do we **reduce the search space**?

Encoding

- Represented as a matrix with N rows and (N $\log_2 I$) columns
  - For example, for a (4 x 8 x 2) tensor, that would be 3 rows and (2 + 3 + 1 = 6) columns
- At each state at step t, we decide which bit from the index tuple will be selected for the bit position t in the linear index
- In each column only one element can be 1 and everything else will be 0 (one-hot encoding),

The initial state

b^5	b^4	b^3	b^2	b^1	b^0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

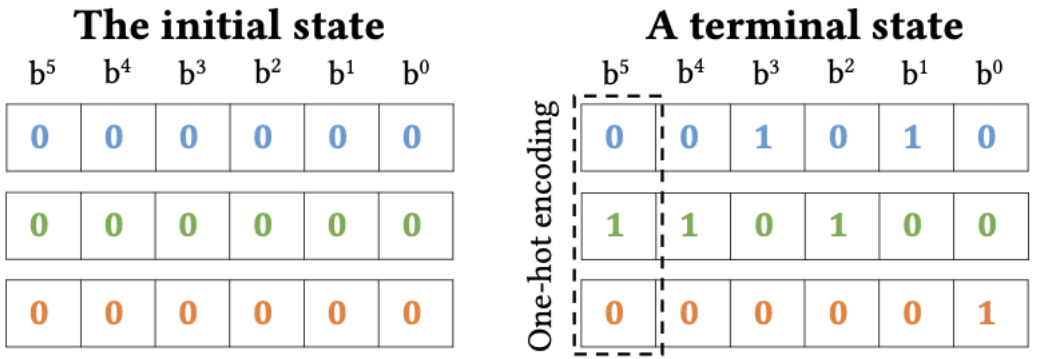
A terminal state

One-hot encoding

b^5	b^4	b^3	b^2	b^1	b^0
0	0	1	0	1	0
1	1	0	1	0	0
0	0	0	0	0	1

Encoding

- We limit the encoding so that bits from a given mode are assigned to the linearized index from most-significant to least-significant , which reduces the total search space from *(N log_2I) -> (N log_2I) / (($log_2$1)! ($log_2$2)! … (log_2N)!)*
- We can restrict how the encoding bits are selected to ensure a correct encoding is always chosen (e.g., # of bits in row n == $log_2 I_n$)



Reward calculation

- Client-server model
  - Client – trains the model
  - Server – evaluates the learned encoding by executing MTTKRP and measuring its performance
  - This allows us to leverage idle data center capacity and evaluate rewards in parallel to improve learning efficiency
  - Minimize noise/jitter
- Using raw execution time -> exploding/vanishing error gradients
  - Using speedup (over ALTO) improved stability, same hyperparameter for different input tensors

Learning Algorithm

- We use a hybrid model-free (double DQN) and model-based DRL

Learning Algorithm

- We use a hybrid model-free (double DQN) and model-based DRL
  - Double DQN improves stability by
    - decoupling action selection (online network, θ) and action evaluation (target network, θ^-)
    - using a prioritized replay buffer to store the latest significant transitions observed by the agent, and sampling mini-batches from the buffer
    - θ and θ^- are adaptive CNN – number of hidden units increases with state-action space and CNN encodes the hierarchical spatial information of the target environment

Learning Algorithm

- We use a hybrid model-free (double DQN) and model-based DRL
  - During early stages of exploration, we have high ϵ to select more random actions that are evaluated on the target environment (i.e., we physically run MTTKRP)
  - During this exploration stage, we also **train a fully connected network** to form a simple reward model
  - Once this model becomes accurate enough (> 90%), when we find an action with a high reward (no worse than the highest observed reward +- error), it is evaluated and used to further update the model
  - This allows us to skip the expensive evaluation step – expensive during the early stage of training, but becomes more efficient during the later stage

Learning Algorithm

- Reward evaluation is expensive
  - Large number of non-zero elements, large and skewed dimensions, large number of modes
  - For example, Reddit takes 10+ seconds to execute 1 MTTKRP
- We use both a model-free (double DQN) and model-based DRL
- We also use a reward cache to avoid evaluating encodings we have already seen

Learning Algorithm

- Delayed reward
  - Reward is calculated at the very end (when we have a real encoding), which results in sparse/delayed reward -> instability and slower learning
  - We introduce a **reward shaping mechanism** that allocates credit from the reward to all actions leading up to the final encoding
    - Uniform credit distribution showed the best result

Evaluation

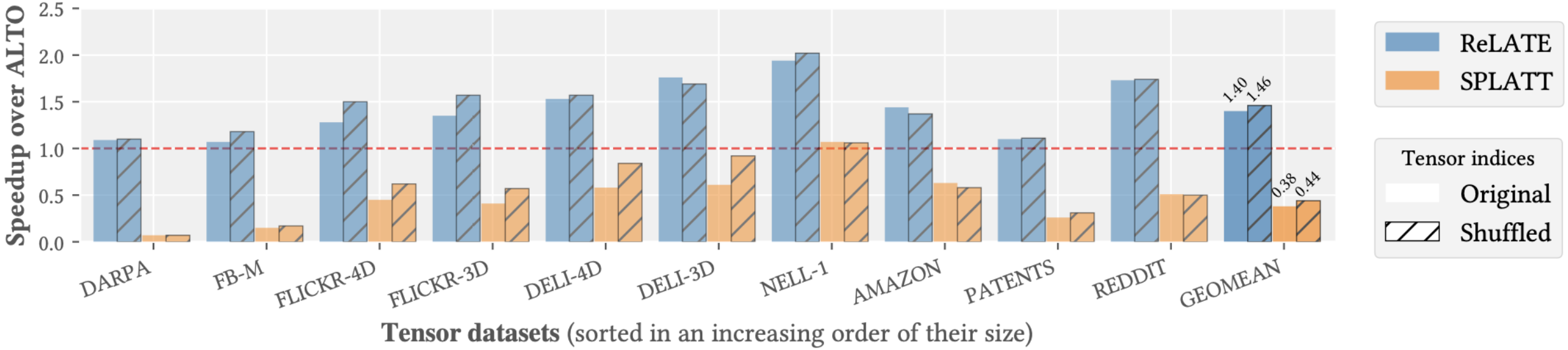


Figure 5: The speedup of our DRL-based sparse encoding (ReLATE) compared to ALTO and SPLATT on a 128-core EMR system.

Evaluation

- Speedup over ALTO (generally) improves with tensor size and sparsity
  - This suggests that DRL is effective compared to input-agnostic (human) expert-designed formats – higher improvements for more difficult problems
- Shuffling improves CSF as well
  - Improves workload balance

Learning Effectiveness

- Error = |estimated – measured|/ measured
- Averaged over 10 samples model had not seen during training
- Reward Model
  - Highly accurate
  - Learned quickly (< 100 ~ 500) episodes

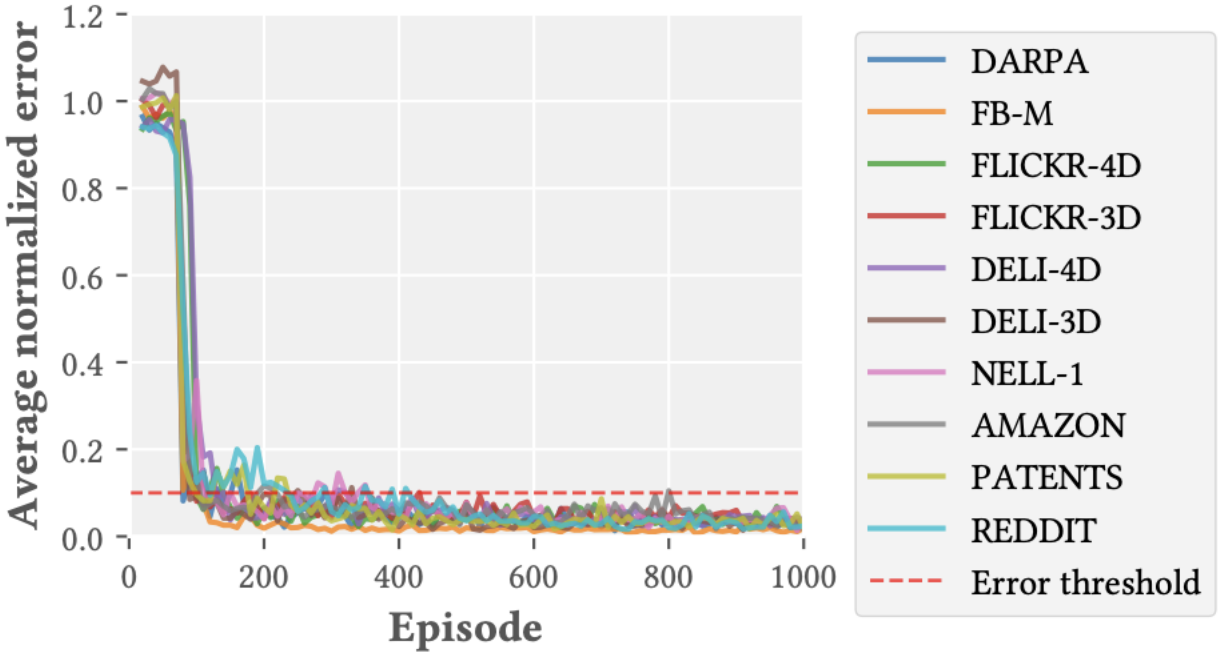


Figure 8: Average estimation error of ReLATE's reward model, relative to the actual reward, over the first thousand episodes.

Learning Effectiveness

- Error = |estimated – measured|/ measured
- Averaged over 10 samples model had not seen during training
- Reward Model
  - Highly accurate
  - Learned quickly (< 100 ~ 500) episodes
- Large % of rewards are modeled for many tensors
  - On average 34% are real

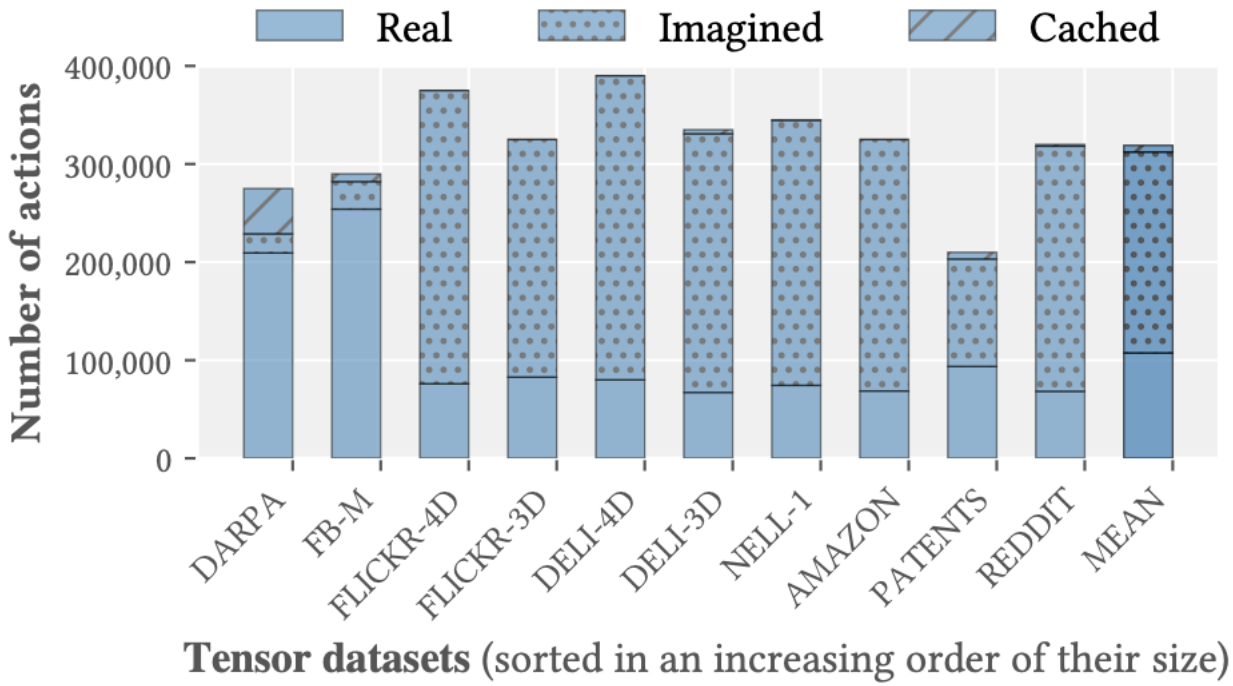


Figure 9: The number of real, cached, and imagined actions taken by the ReLATE agent across sparse tensors.

Practical?

- Yes and no
- Yes
  - In practical data analysis using tensor decomposition, you run a) on several randomly initialized factors, b) multiple factor ranks, c) different data processing methods (to handle noise, outliers, etc.) -> tens to hundreds of thousands of MTTKRP
  - Amortizable for real-world data analysis applications
- No
  - We need to train for each new input tensor (cumbersome, unless seamlessly integrated into the workflow)
  - Current speedup (1.4x geometric mean) may not seem “worth the trouble.”
- However, it demonstrates the potential for using AI for creating an input-adaptive encoding for sparse tensors

Questions?

Credit goes to my collaborators and students:

- **S Isaac Geronimo Anderson**
- Fabio Checconi
- Brian J Gravelle
- Ahmed E Helal
- Ramakrishnan Kannan
- Jan Laukemann
- **Andy Nguyen**
- Fabrizio Petrini
- Teresa Ranadive
- Piyush Sao
- Tyler Simon
- Shaden Smith
- **Yongseok Soh**
- Jesmin Jahan Tithi