UNIVERSITY OF CALIFORNIA
UC RIVERSIDE
Marlan and Rosemary Bourns
College of Engineering

DREAM: Device-driven Efficient Access to Virtual Memory

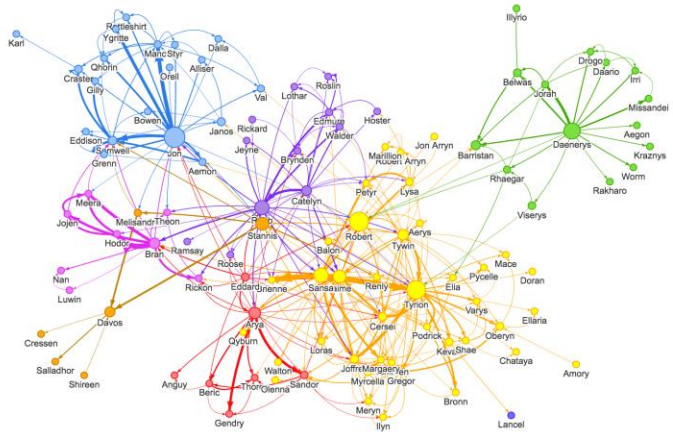
Nurlan Nazaraliyev, Elaheh Sadredini, Nael Abu-Ghazaleh

ICS'25, Salt Lake City, 11th June, 2025

Emerging Applications for Accelerators

UCR | Marlan and Rosemary Bourns College of Engineering

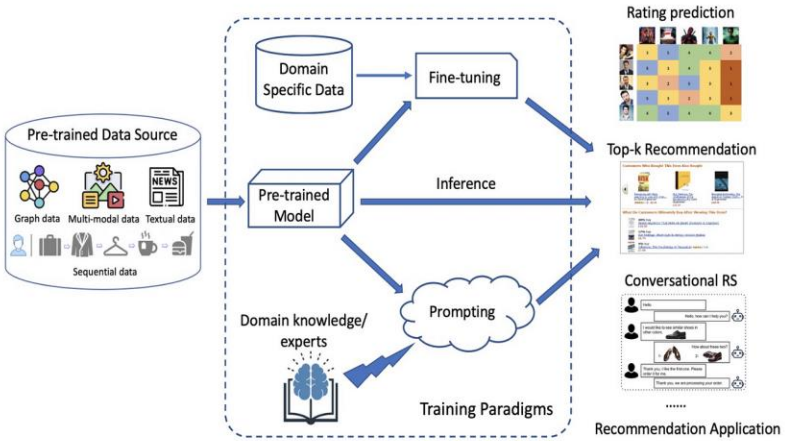
Graph analytics



Data analytics

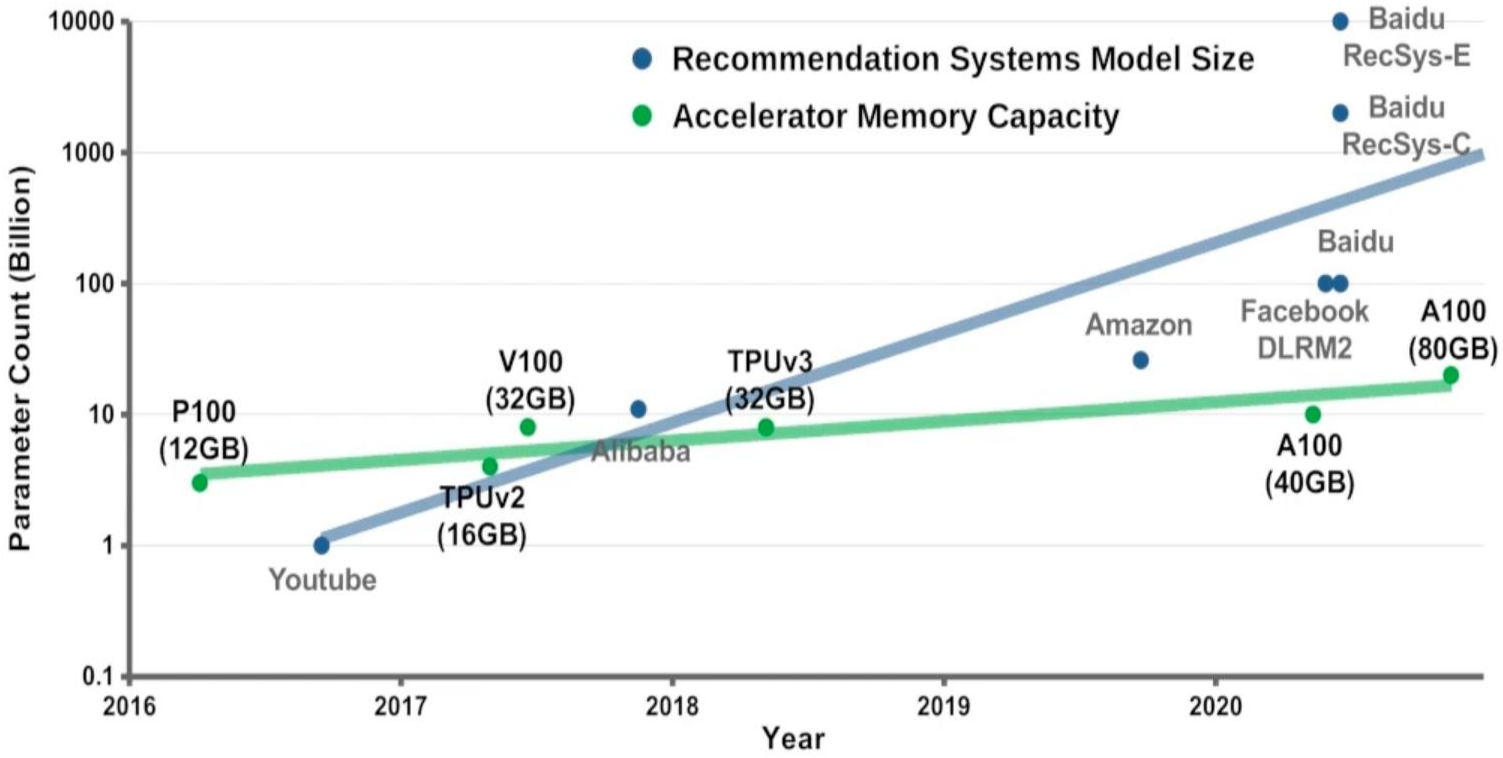
id integer	name character varying	sex charac	age charac	height charac	weight charac	team character vary	noc charac	games character varying	year integer	season character	city character var	sport character varying	event character varying	medal charac
10	Einar Ferdinand "Einari" ...	M	26	NA	NA	Finland	FIN	1952 Summer	1952	Summer	Helsinki	Swimming	Swimming Men's 400 metres ...	NA
11	Jorma Ilmari Aalto	M	22	182	76.5	Finland	FIN	1980 Winter	1980	Winter	Lake Placid	Cross Country Ski...	Cross Country Skiing Men's 3...	NA
12	Jyri Tapani Aalto	M	31	172	70	Finland	FIN	2000 Summer	2000	Summer	Sydney	Badminton	Badminton Men's Singles	NA
13	Minna Maarit Aalto	F	30	159	55.5	Finland	FIN	1996 Summer	1996	Summer	Atlanta	Sailing	Sailing Women's Windsurfer	NA
13	Minna Maarit Aalto	F	34	159	55.5	Finland	FIN	2000 Summer	2000	Summer	Sydney	Sailing	Sailing Women's Windsurfer	NA
14	Pirjo Hannele Aalto (Ma...	F	32	171	65	Finland	FIN	1994 Winter	1994	Winter	Lillehamm...	Biathlon	Biathlon Women's 7.5 kilomet...	NA
15	Arvo Ossian Aaltonen	M	22	NA	NA	Finland	FIN	1912 Summer	1912	Summer	Stockholm	Swimming	Swimming Men's 200 metres ...	NA
15	Arvo Ossian Aaltonen	M	22	NA	NA	Finland	FIN	1912 Summer	1912	Summer	Stockholm	Swimming	Swimming Men's 400 metres ...	NA
15	Arvo Ossian Aaltonen	M	30	NA	NA	Finland	FIN	1920 Summer	1920	Summer	Antwerpen	Swimming	Swimming Men's 200 metres ...	Bronze
15	Arvo Ossian Aaltonen	M	30	NA	NA	Finland	FIN	1920 Summer	1920	Summer	Antwerpen	Swimming	Swimming Men's 400 metres ...	Bronze
15	Arvo Ossian Aaltonen	M	34	NA	NA	Finland	FIN	1924 Summer	1924	Summer	Paris	Swimming	Swimming Men's 200 metres ...	NA
16	Juhamatti Tapio Aalton...	M	28	184	85	Finland	FIN	2014 Winter	2014	Winter	Sochi	Ice Hockey	Ice Hockey Men's Ice Hockey	Bronze
17	Paavo Johannes Aalton...	M	28	175	64	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Individual ...	Bronze
17	Paavo Johannes Aalton...	M	28	175	64	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Team All-A...	Gold
17	Paavo Johannes Aalton...	M	28	175	64	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Floor Exerc...	NA
17	Paavo Johannes Aalton...	M	28	175	64	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Horse Vault	Gold
17	Paavo Johannes Aalton...	M	28	175	64	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Parallel Bars	NA
17	Paavo Johannes Aalton...	M	28	175	64	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Horizontal ...	NA
17	Paavo Johannes Aalton...	M	28	175	64	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Rings	NA
17	Paavo Johannes Aalton...	M	28	175	64	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Pommelle...	Gold

Recommender Systems



Accelerators becoming more widely used for nontraditional applications

Emerging Challenge!



Models bigger than accelerator memory!

10,000x growth in model size compared to 10x growth in accelerator memory!

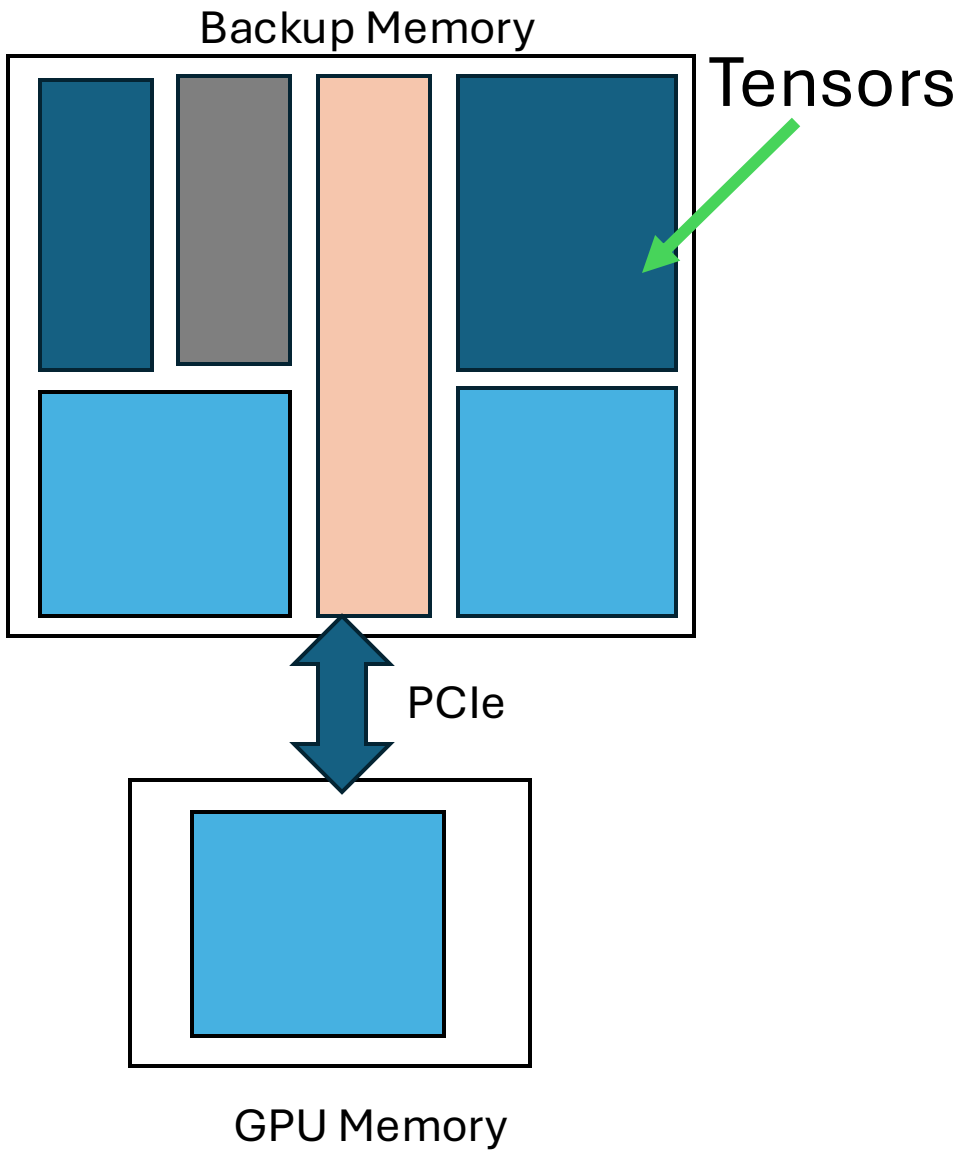
Easy Solution - Partition Workload

Example:

- Tensor-level partitioning

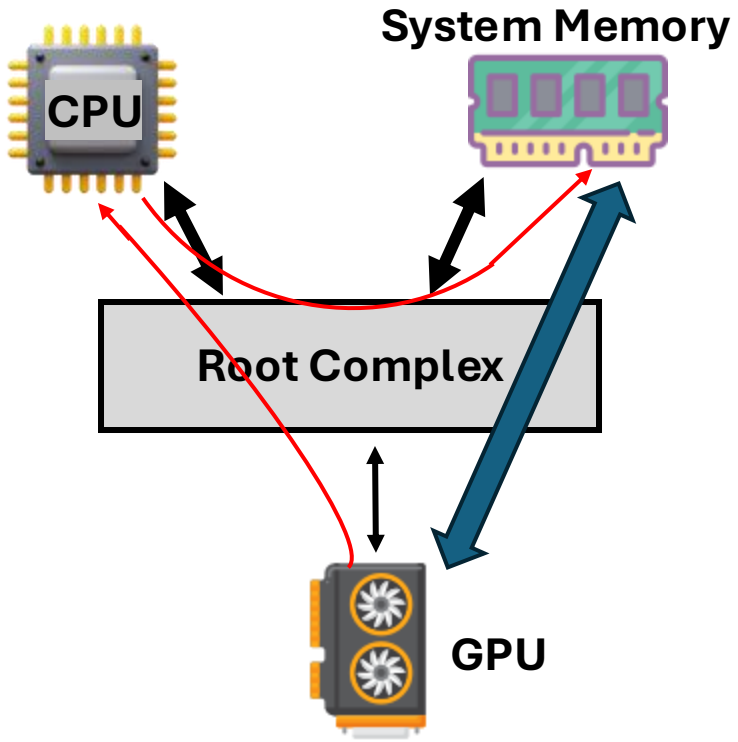
Hard to adapt for irregular applications!

High I/O amplification in face of sparsity!



Alternative: Oversubscribe GPU memory!

Existing Approach: NVIDIA Unified Virtual Memory

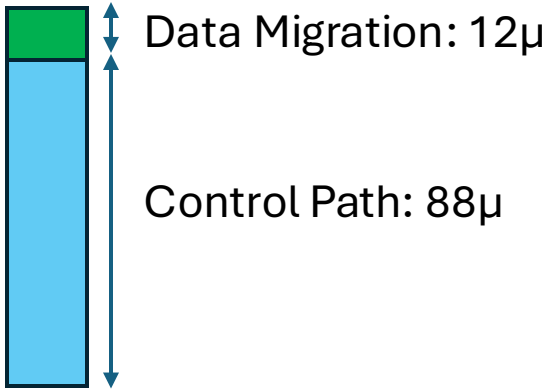


Red: control, blue: data path

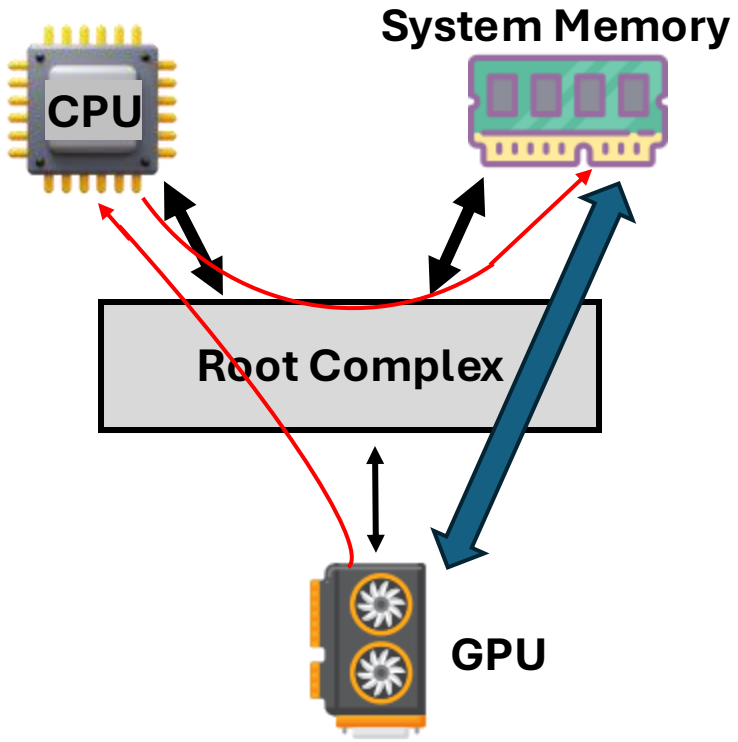
Alternative: Oversubscribe GPU Memory!

Existing Approach: NVIDIA Unified Virtual Memory

For a page fault of 64KB:



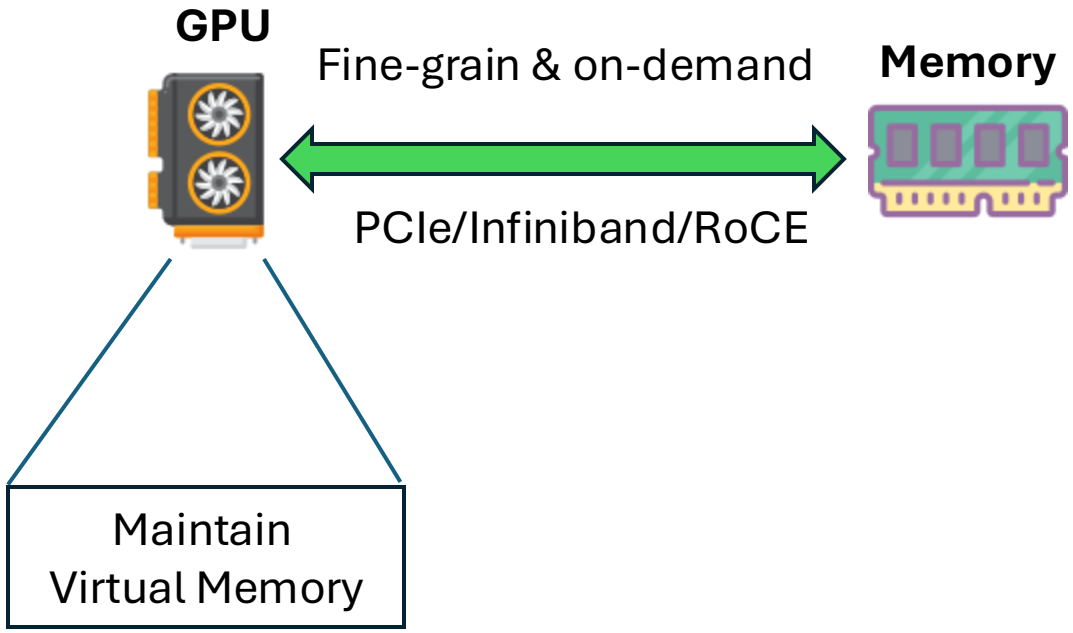
Host-side involvement for page fault handling : 7x longer than transfer!



Red: control, blue: data path

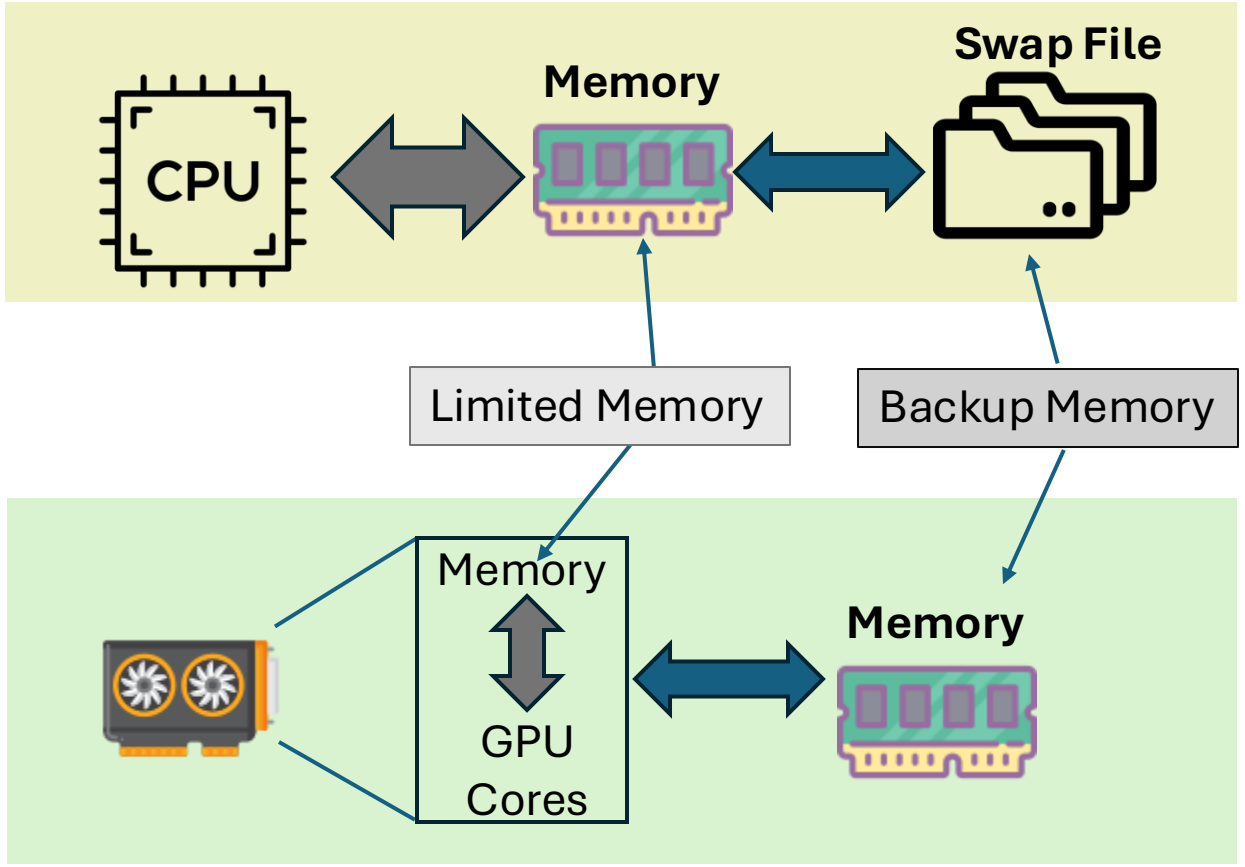
Time for Research Question!

How to enable GPUs to manage their own virtual memory?



Insight!

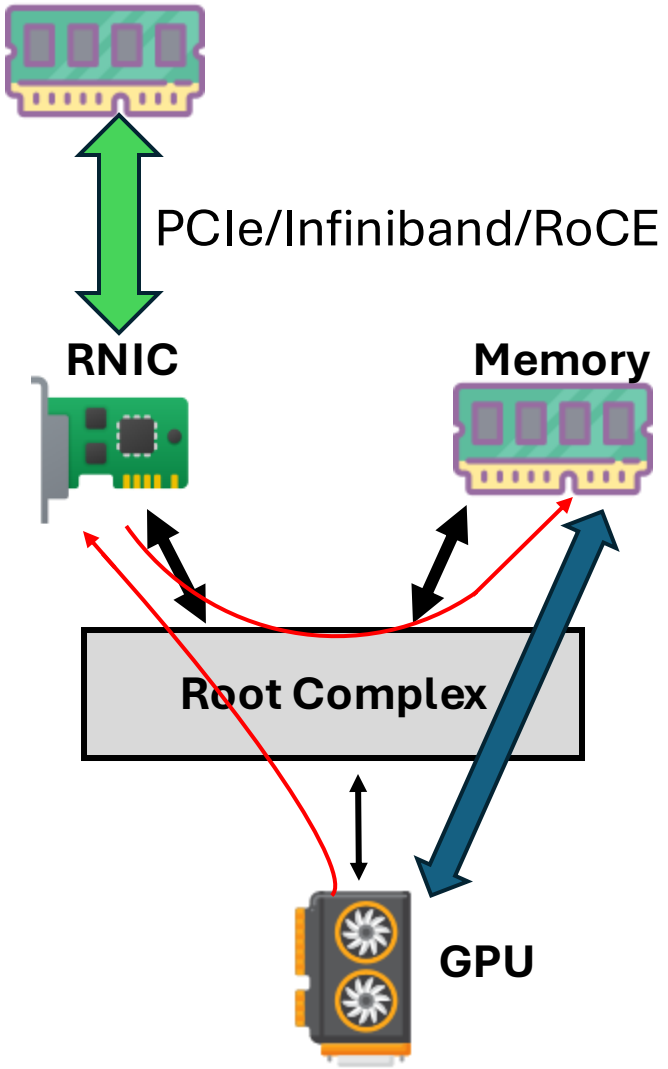
Based on insights from Linux swap system!



Enable Memory Access through Network Card

Replace the OS + UVM driver with special hardware!

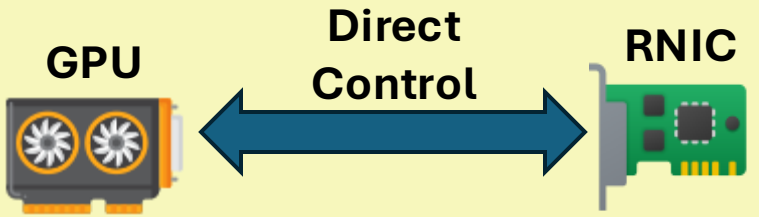
RNIC: RDMA-capable network interface card



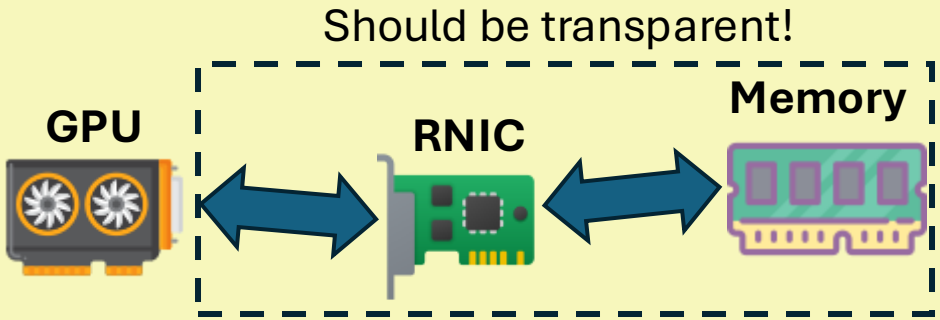
Red: control, blue: data path

Challenges to Solve!

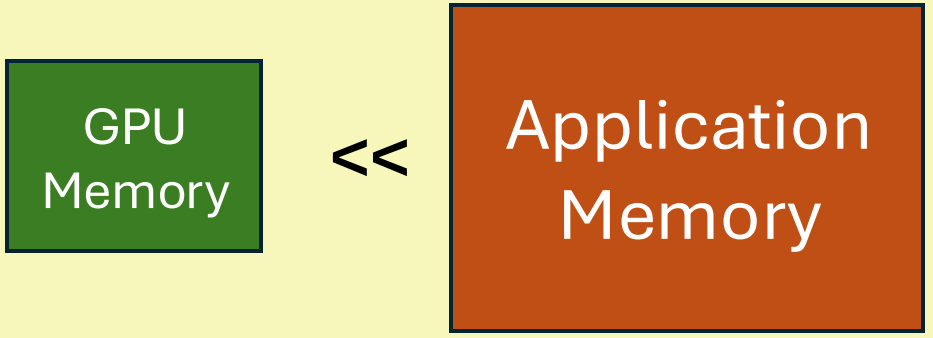
How to enable GPU to access NIC resources?



How to hide the integration complexity?

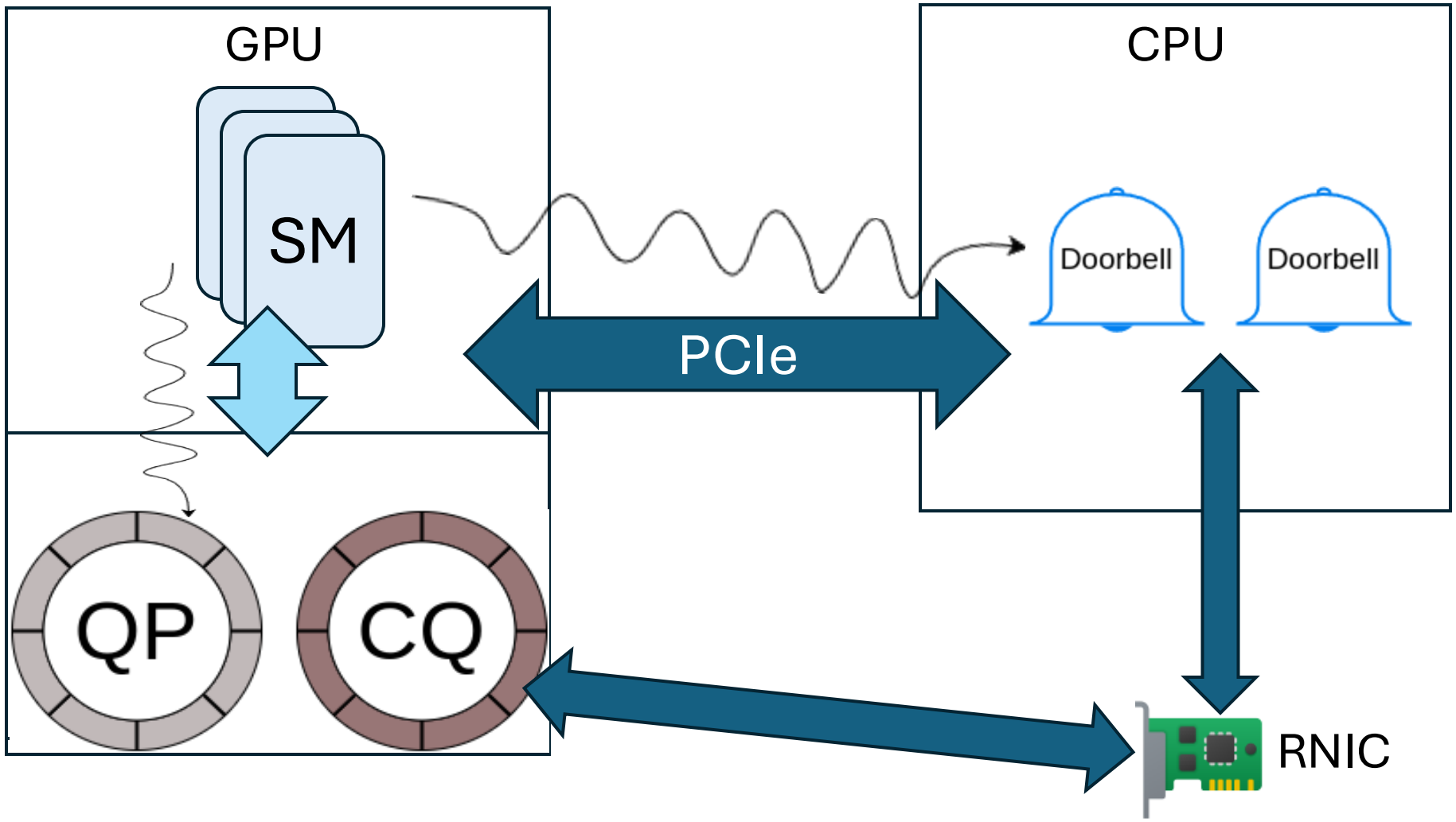


How to support oversubscribed GPU memory?



Challenge 1

How to enable GPU to access NIC resources?



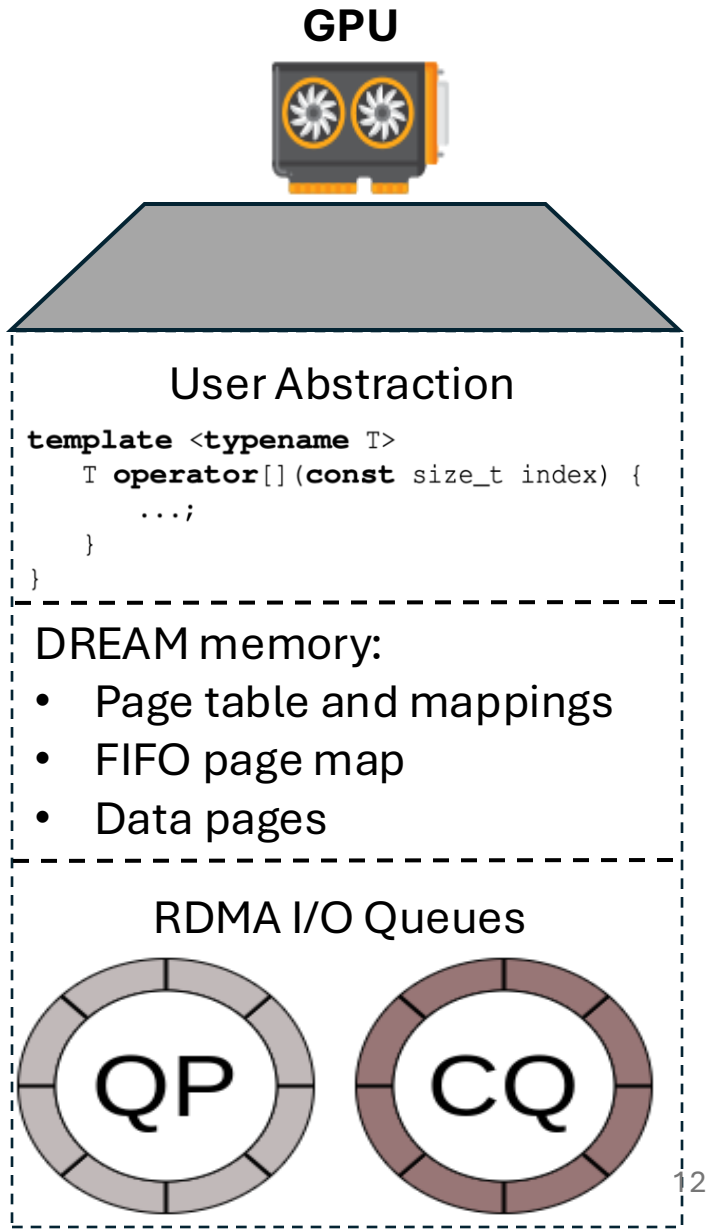
System Design

How to hide the integration complexity?

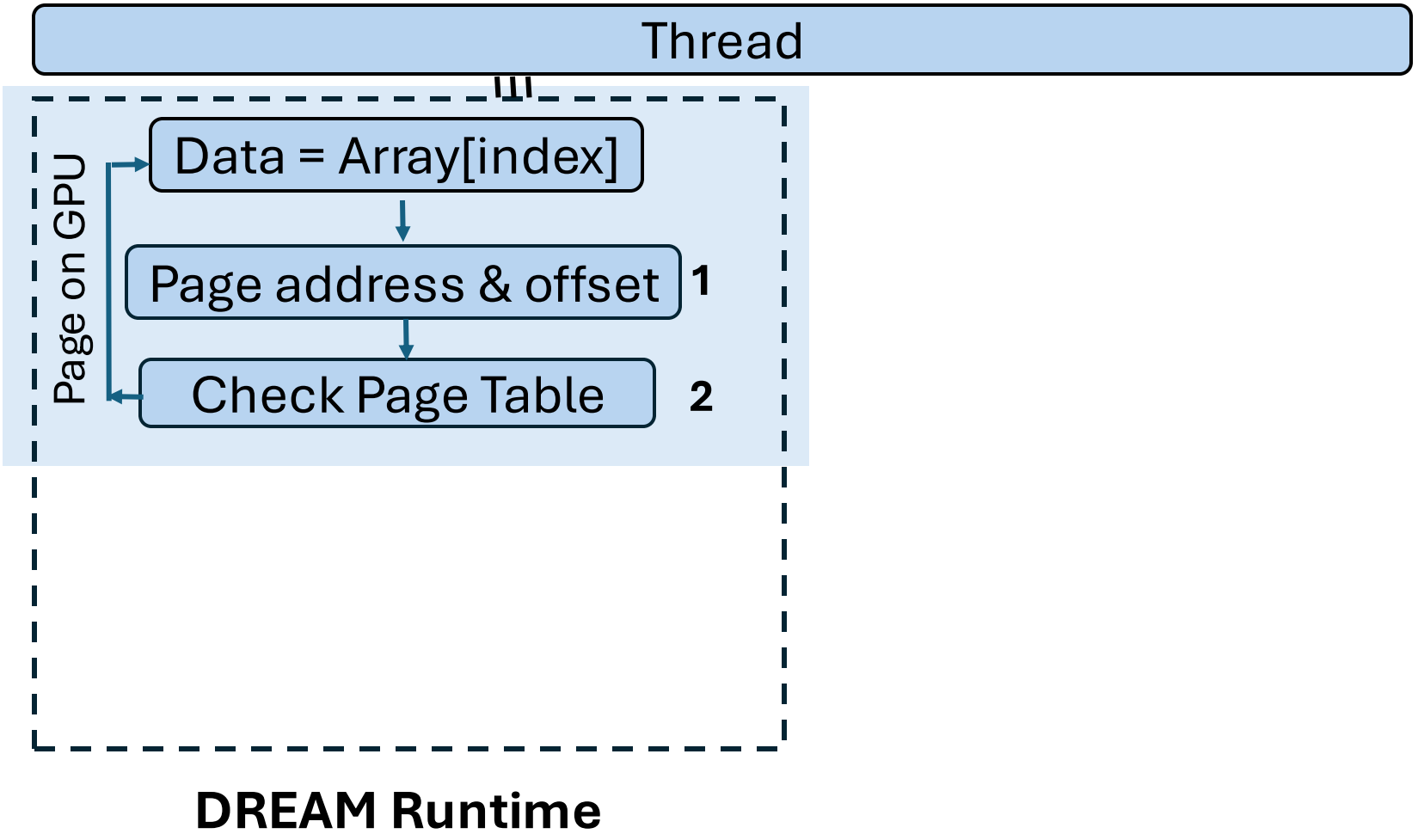
Example code:

```
__global__
void kernel(dream_ptr<float> *data, size_t n) {
    size_t tid = ...;
    ...
    if (tid < n) {
        float a = data[tid];
        ...;
    }
}
```

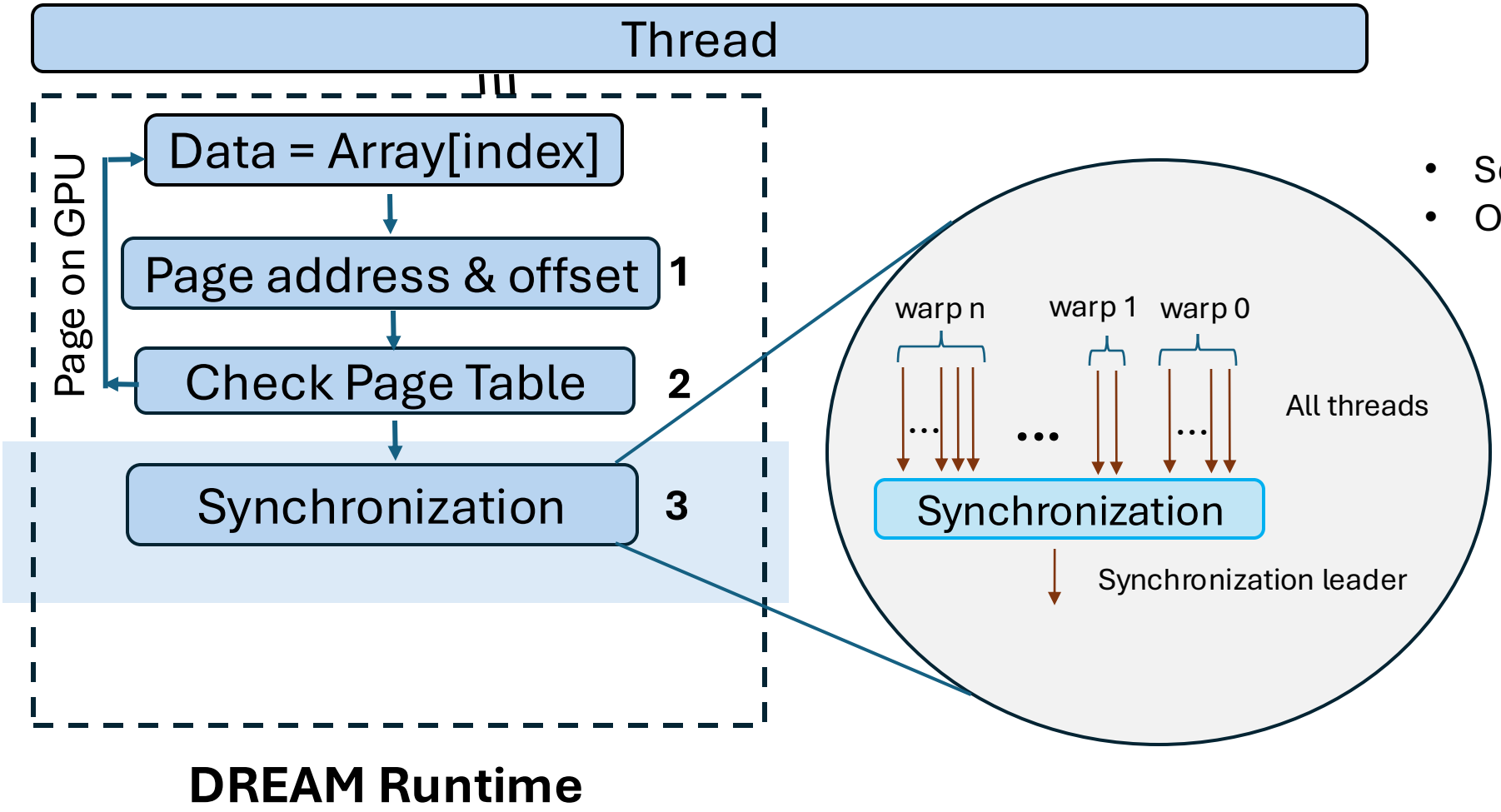
Same from CPU's view point!



GPU Threads in DREAM System

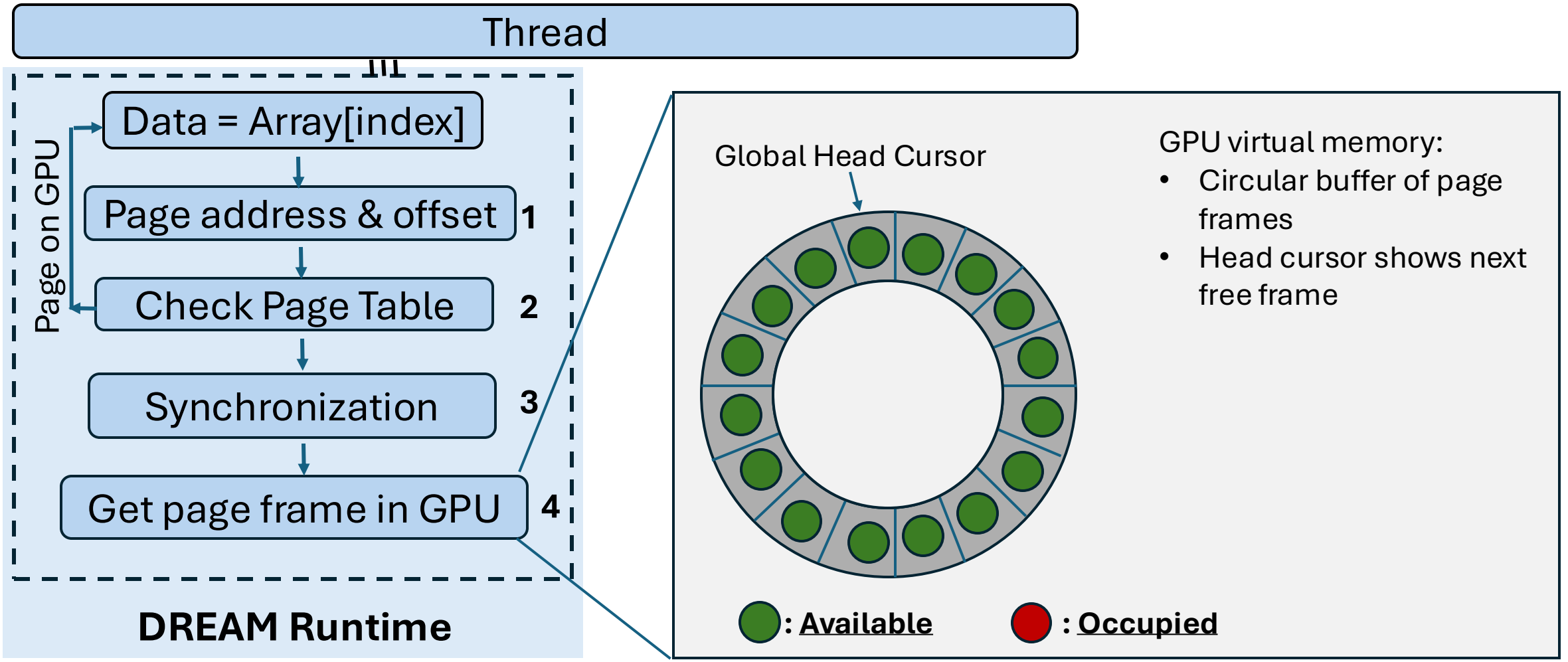


GPU Threads in DREAM System

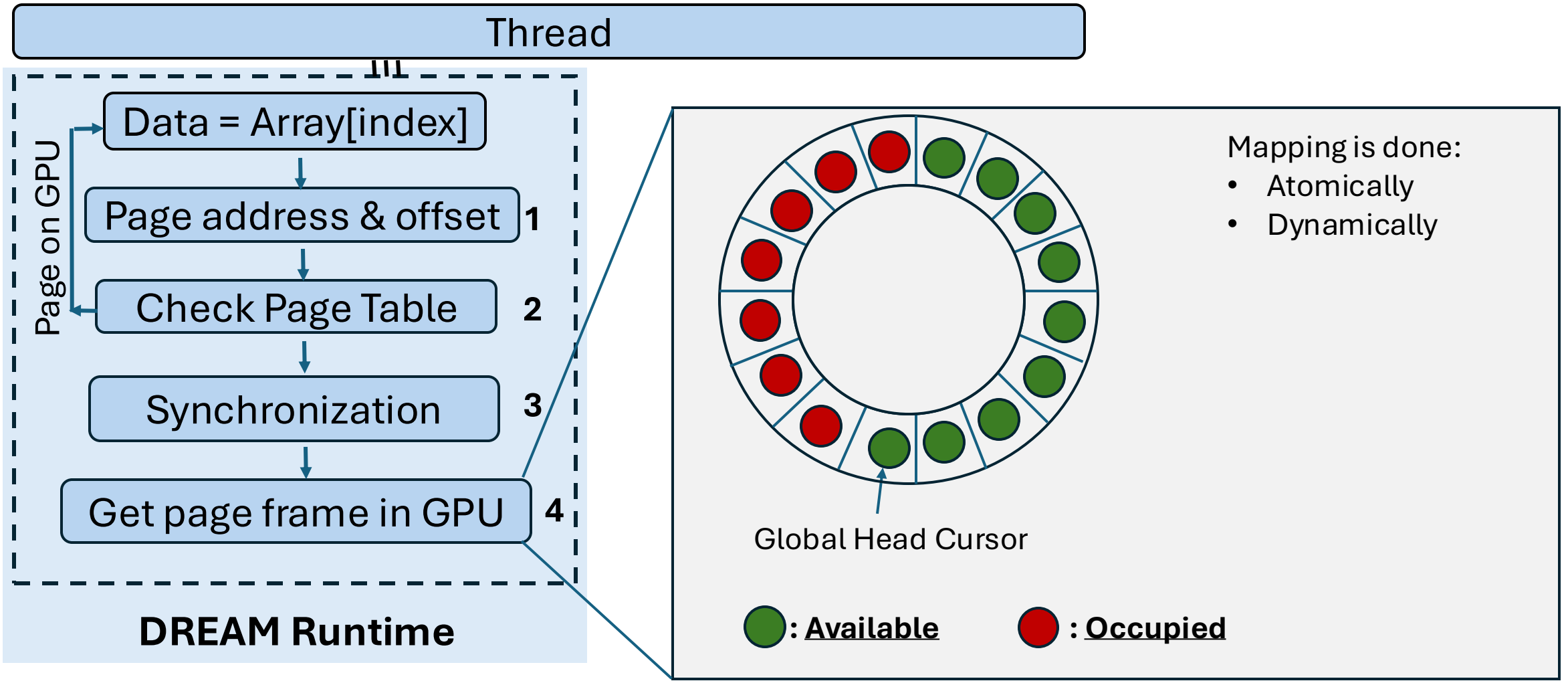


- Selection of warp leaders
- One leader out of all warp leaders

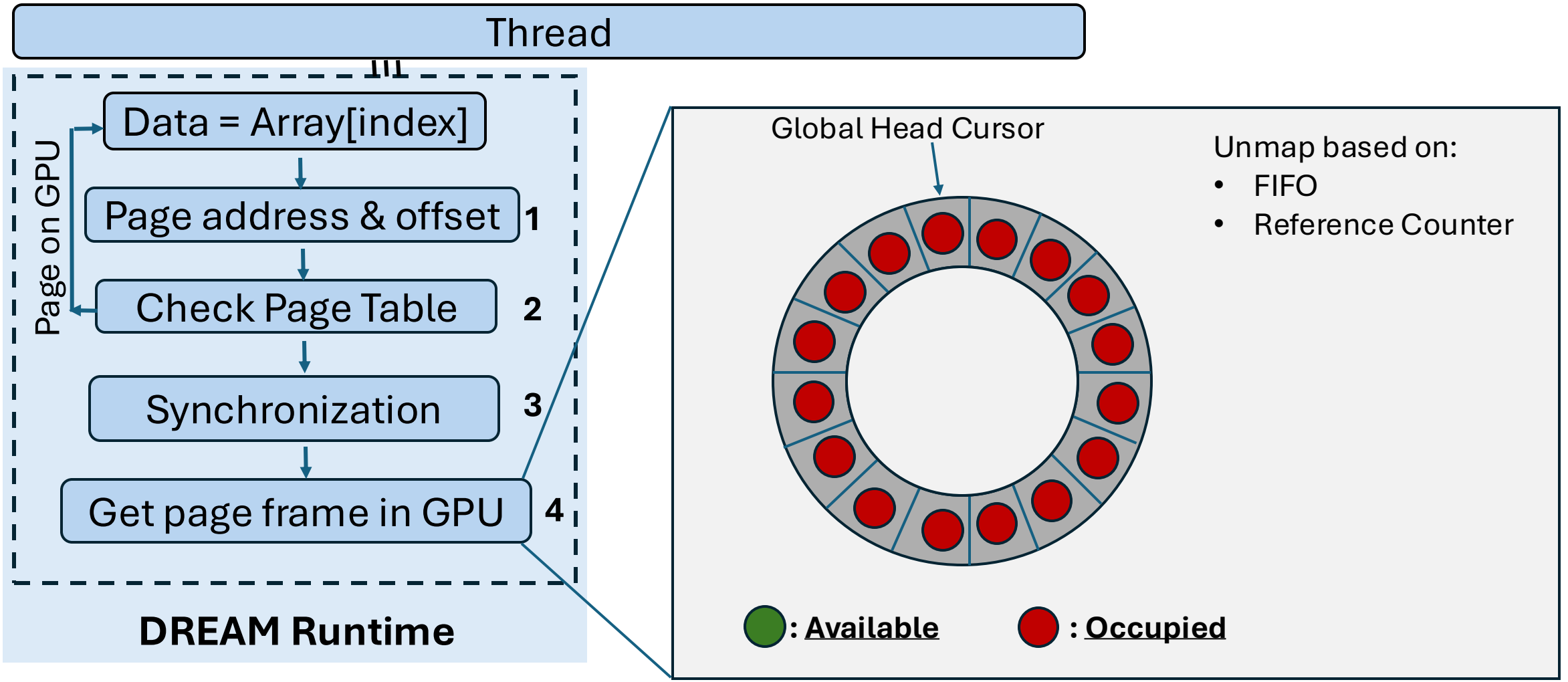
GPU Threads in DREAM System



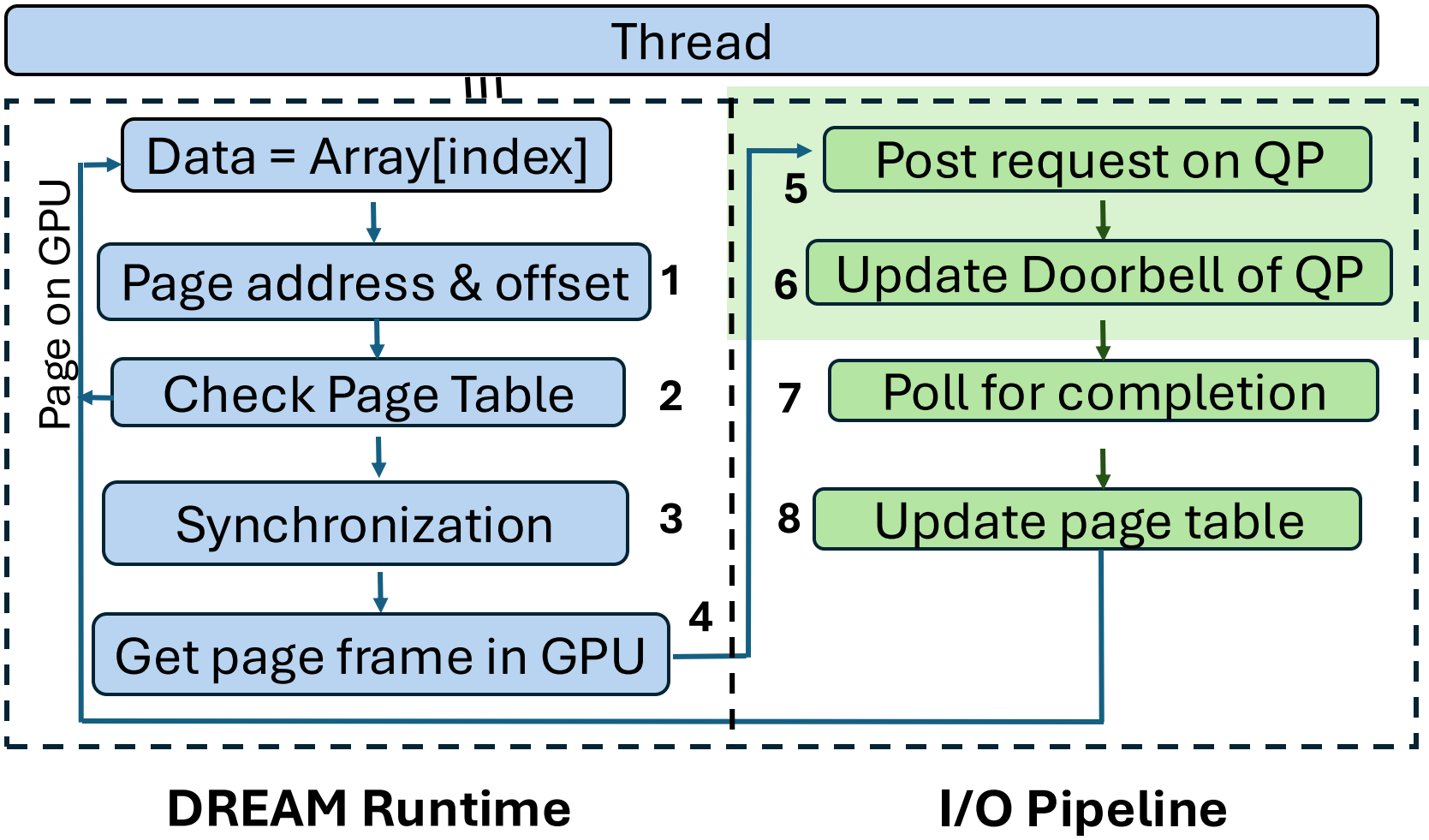
GPU Threads in DREAM System



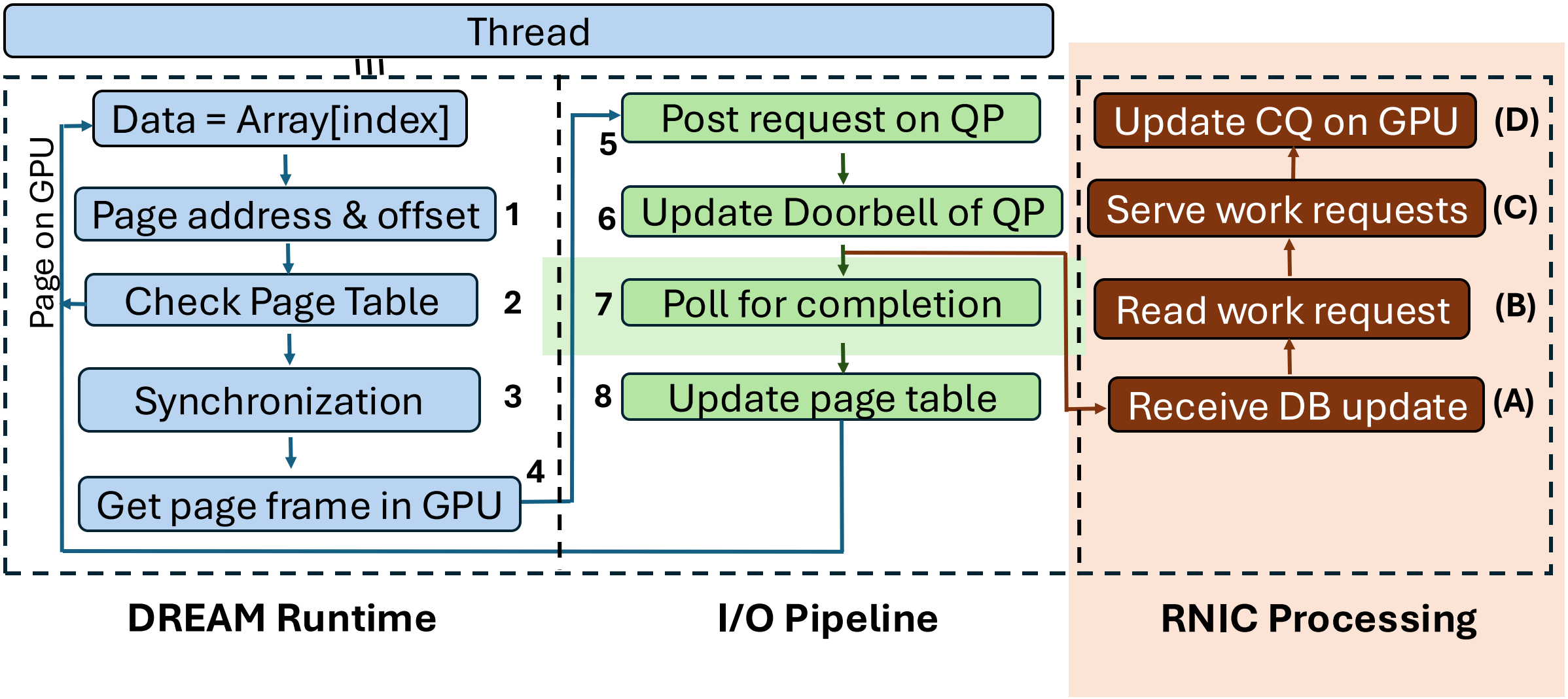
GPU Threads in DREAM System



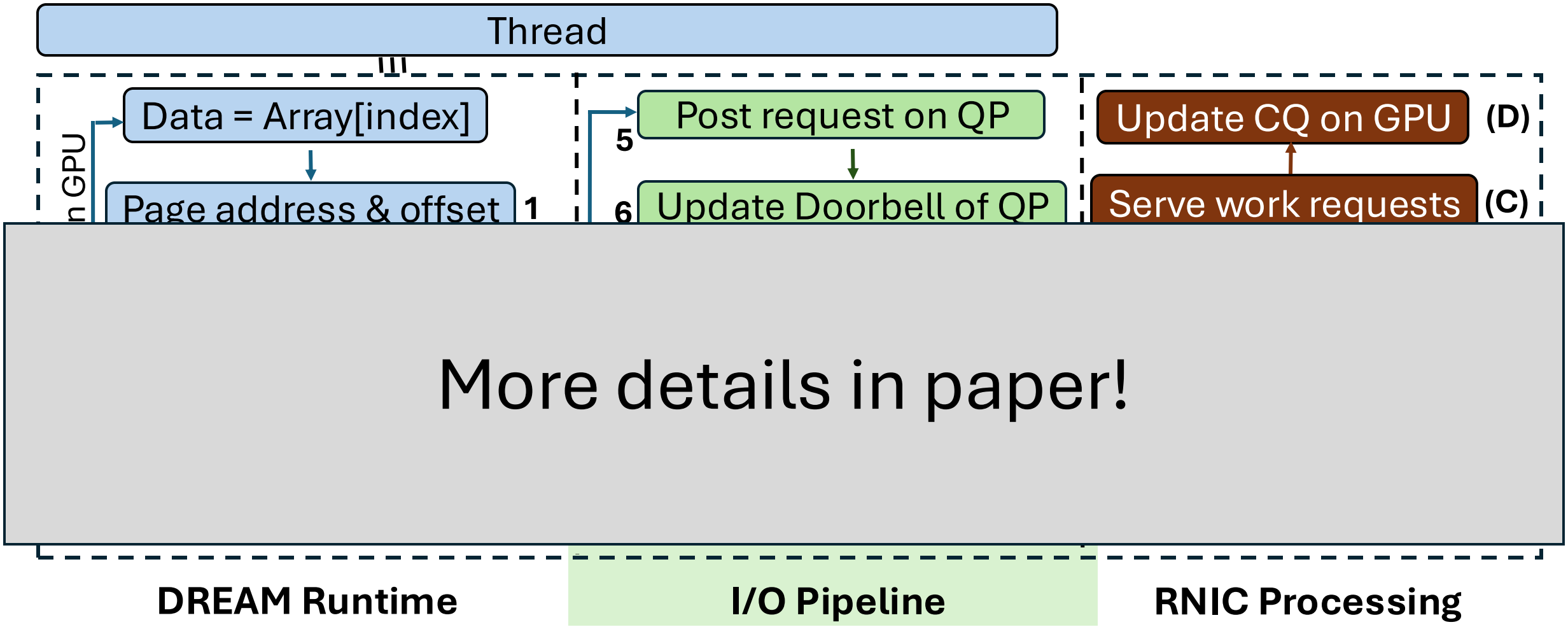
GPU Threads in DREAM System



GPU Threads in DREAM System



GPU Threads in DREAM System



Evaluation Setup

Software prototype developed on CloudLab and evaluated on:

- Microbenchmarks to measure transfer throughput
- Linear Algebra Kernels: MVT, BIGC, ATAX, etc.
- Graph Application: BFS, CC, etc.
- Query Processing

Component	Specification
CPU	2× AMD 7542 (32 cores, 2.40 GHz)
GPU	NVIDIA Tesla V100 32GB
RAM	512GB 3200MHz DDR4
NIC	NVIDIA Mellanox ConnectX-5 25Gbps & ConnectX-6 100Gbps
Software	Ubuntu 22.04 LTS, NVIDIA Driver 535.183.01, CUDA 12.2

Baselines:

➢ UVM, Rapids, Subway [EuroSys'20]

✓ How does DREAM perform in end-to-end application time?
✓ How does DREAM help reduce redundant I/O?
✓ How well does DREAM applications bigger than GPU memory size?

Graph Analytics - UVM

Largest graphs from SuiteSparse [1]

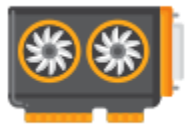
|E|: number of edges, |V|: number of vertices

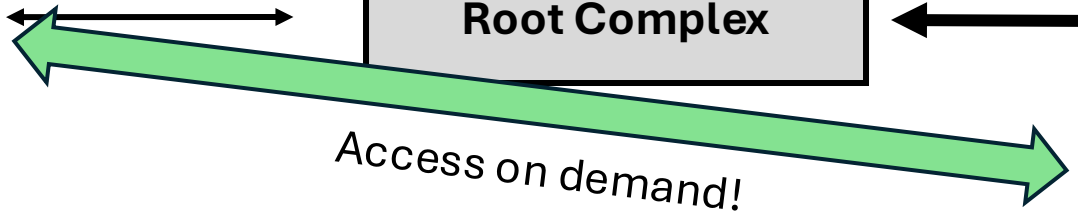
Dataset Name	Abbr	|E|	|V|	Size (GB)
GAP-Urand	GU	4.29B	134.2M	16.0
GAP-Kron	GK	4.23B	134.2M	15.7
Friendster	FS	3.61B	65.6M	13.5
MOLIERE	MO	6.67B	30.2M	24.8

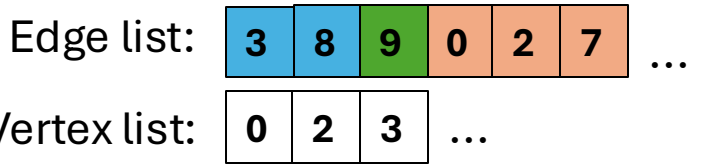
NVIDIA V100 GPU (32GB)

Target UVM baseline [2]
Well optimized solution!

Abundant CPU memory







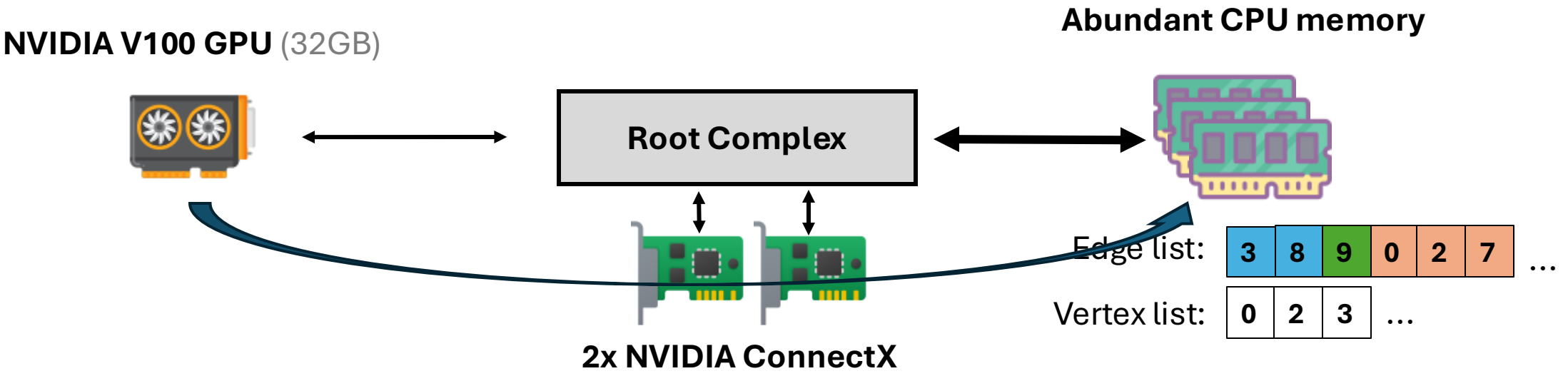
[1] The SuiteSparse Matrix Collection Website Interface: https://doi.org/10.21105/joss.01244
[2] EMOGI: Efficient memory-access for out-of-memory graph-traversal in GPUs.

Graph Analytics - DREAM

Largest graphs from SuiteSparse [1]

|E|: number of edges, |V|: number of vertices

Dataset Name	Abbr	|E|	|V|	Size (GB)
GAP-Urand	GU	4.29B	134.2M	16.0
GAP-Kron	GK	4.23B	134.2M	15.7
Friendster	FS	3.61B	65.6M	13.5
MOLIERE	MO	6.67B	30.2M	24.8

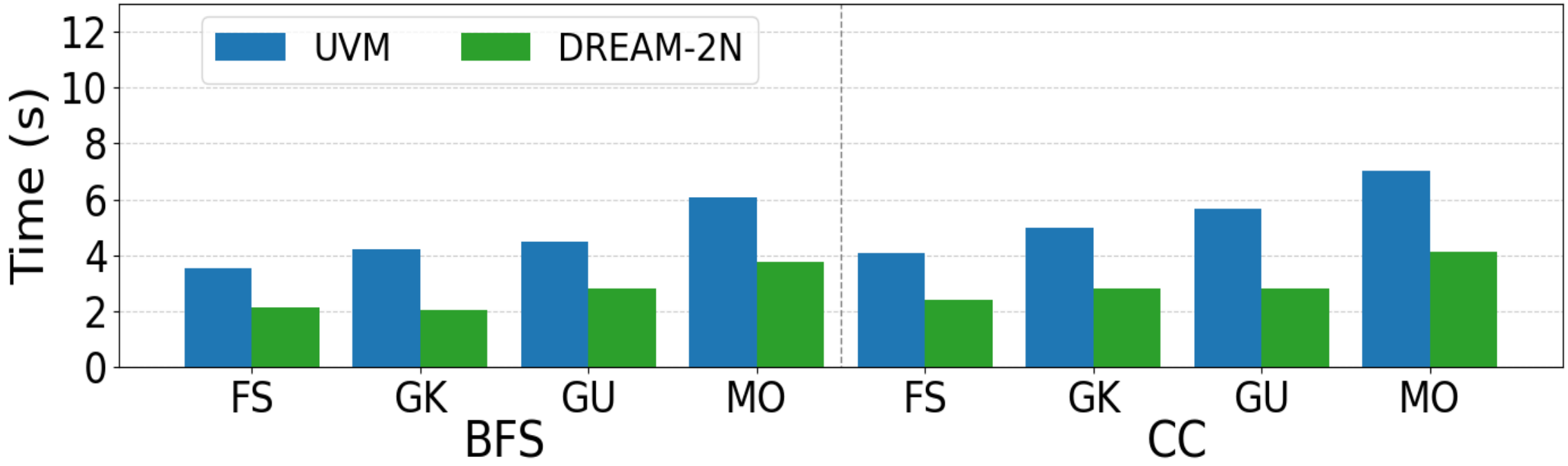


[1] The SuiteSparse Matrix Collection Website Interface: https://doi.org/10.21105/joss.01244

Graph Analytics Results

Page size: UVM: 64KB; DREAM: 4KB

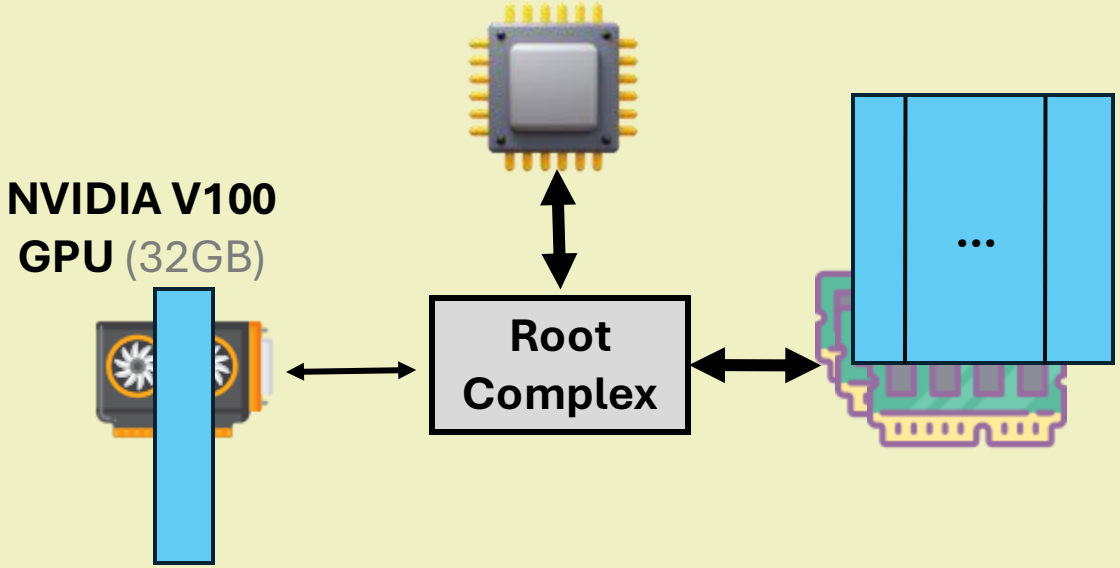
Average of 1.5x and 1.4x speedup over target UVM baseline for BFS and CC, respectively



Query Evaluation - Rapids

Chicago Taxi Trips Dataset [1]

- More that 210M trips
- Size > 80 GB
- Parquet file format – columnar storage
- Columns processed:
  - Trip Seconds
  - Miles
  - Fare
  - Extras
  - Tips
  - Tolls



Baseline RAPIDS V24.10 [2]:

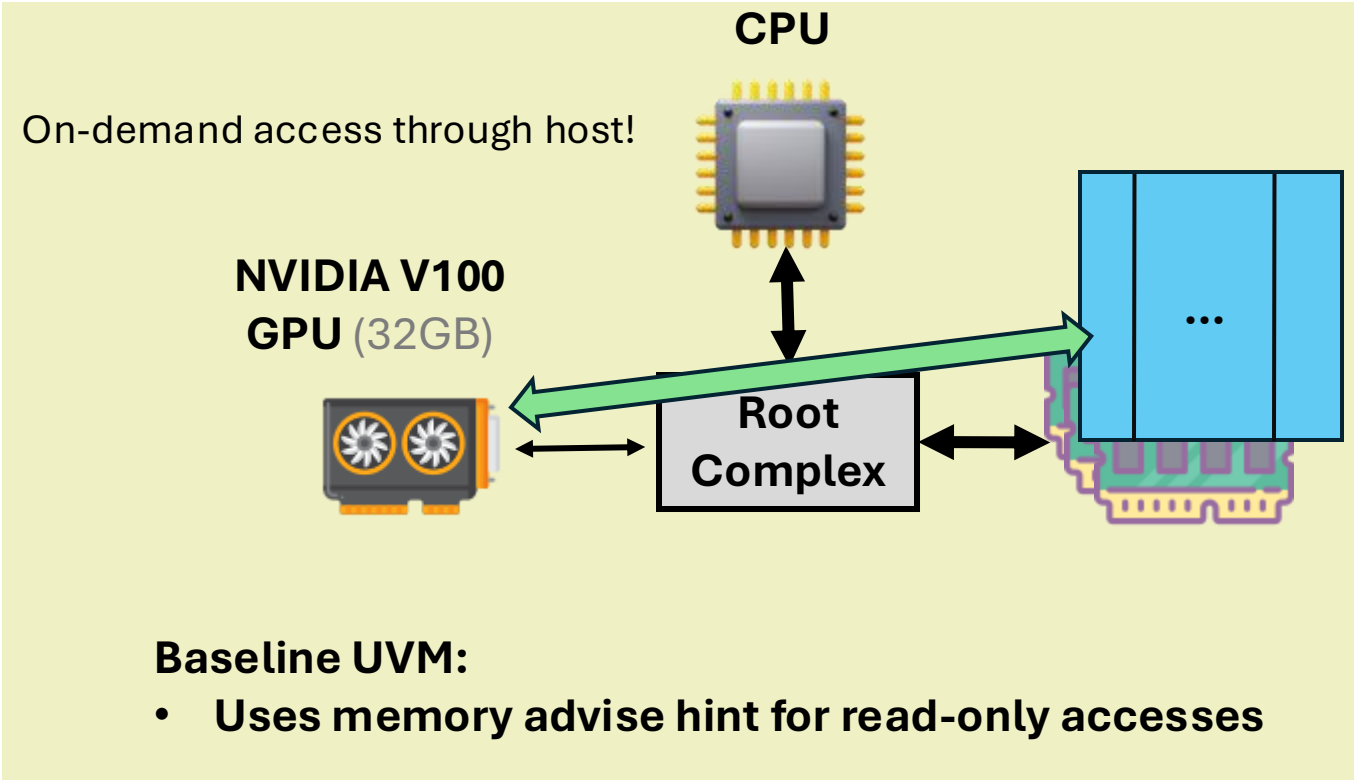
- **Pin the necessary columns in host memory**

[1] City of Chicago: https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew
[2] **NVIDIA RAPIDS**: https://developer.nvidia.com/rapids

Query Evaluation - UVM

Chicago Taxi Trips Dataset [1]

- More that 210M trips
- Size > 80 GB
- Parquet file format – columnar storage
- Columns processed:
  - Trip Seconds
  - Miles
  - Fare
  - Extras
  - Tips
  - Tolls



Baseline UVM:

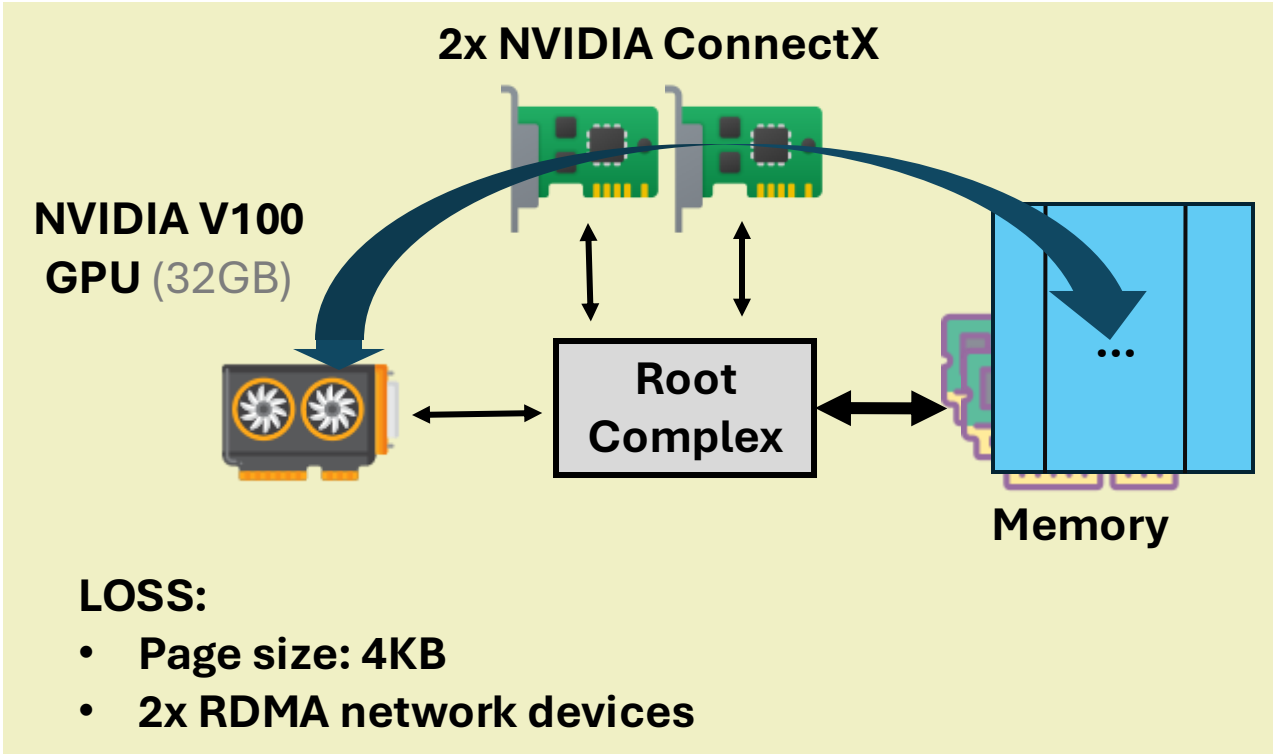
- **Uses memory advise hint for read-only accesses**

[1] City of Chicago: https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew

Query Evaluation - DREAM

Chicago Taxi Trips Dataset [1]

- More that 210M trips
- Size > 80 GB
- Parquet file format – columnar storage
- Columns processed:
  - Trip Seconds
  - Miles
  - Fare
  - Extras
  - Tips
  - Tolls

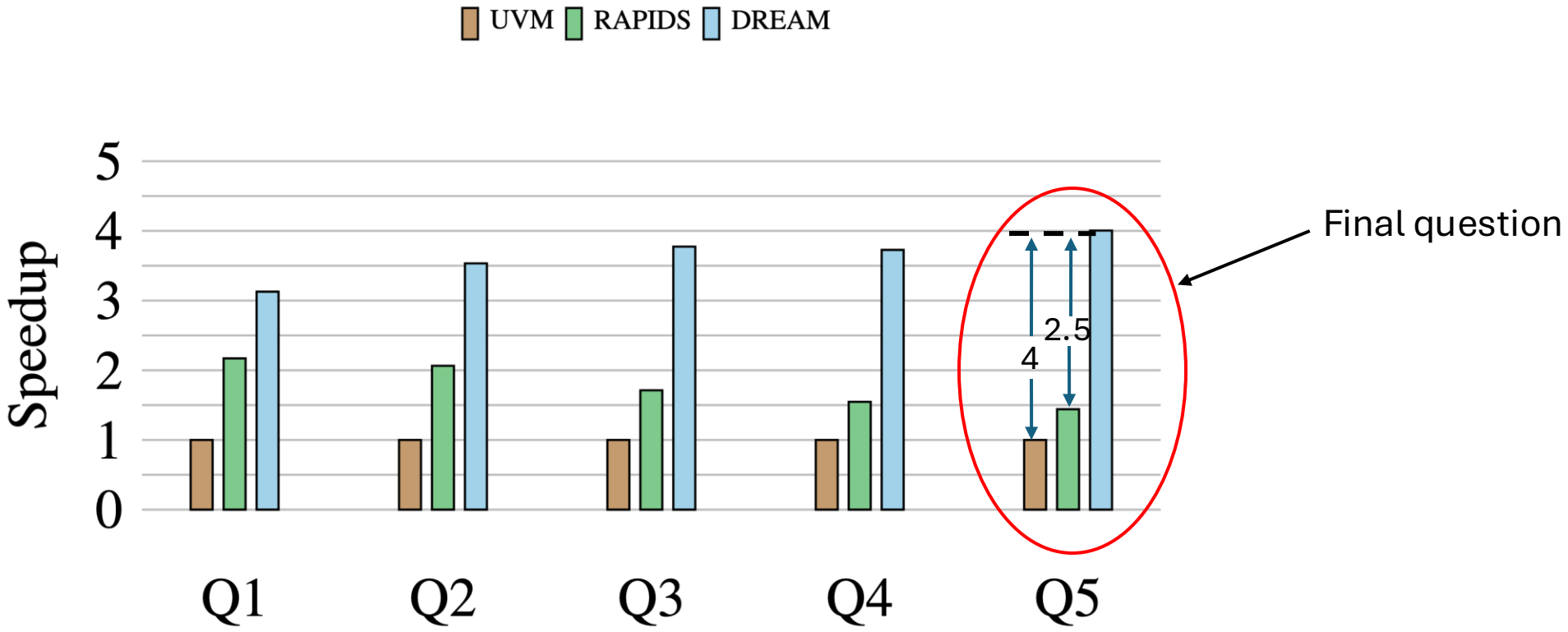


Find: How much money did driver make for trips longer than 9000 seconds?

[1] City of Chicago: https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew

Query Evaluation Results

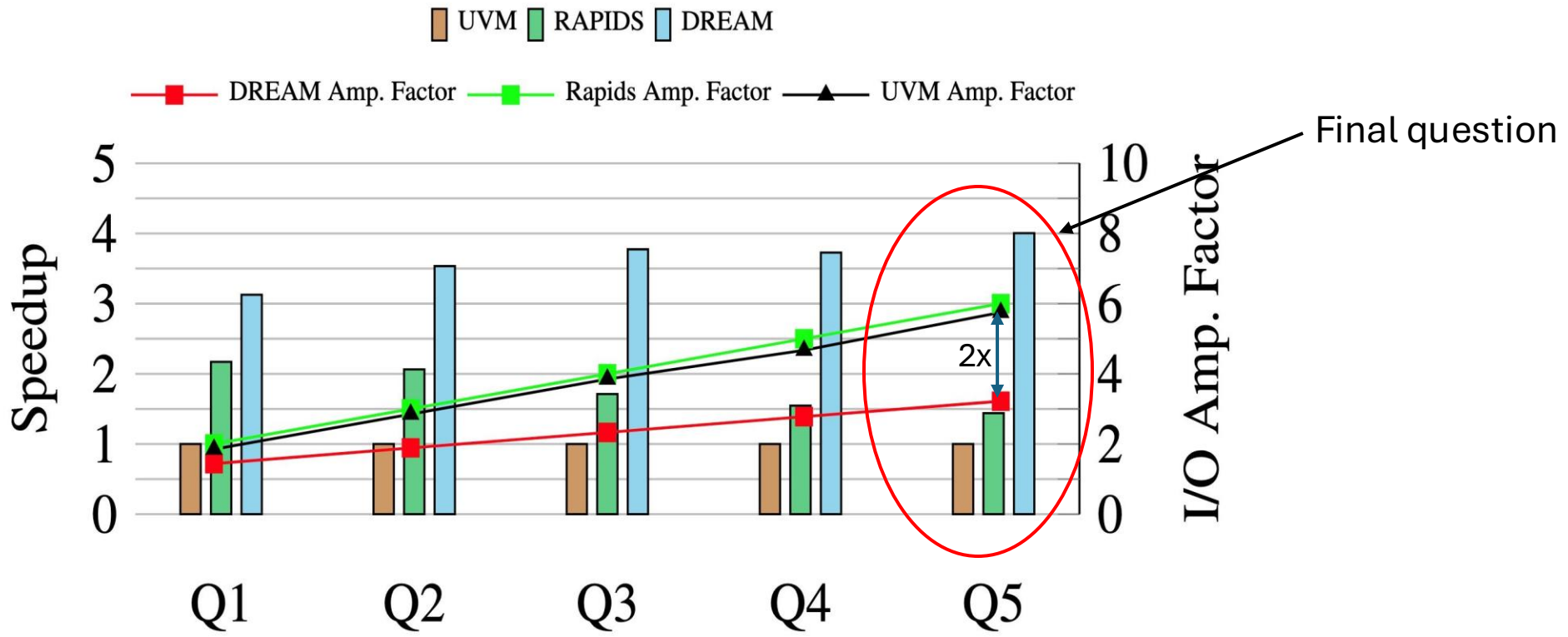
- Transfers only necessary data (Sparsity: 0.08%)



Q1: Trip Miles, Q2: Fares, Q3: Extras, Q4: Tips, Q5: Tolls; Query is cumulative.

Query Evaluation Results

- Transfers only necessary data (Sparsity: 0.08%)
- Lower I/O due to finer grain transfer



Q1: Trip Miles, Q2: Fares, Q3: Extras, Q4: Tips, Q5: Tolls; Query is cumulative.

Oversubscription Analysis

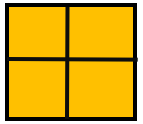
Criteria:

- ✓ Limit the available memory
- ✓ Observe the effect on performance

Oversubscription level:

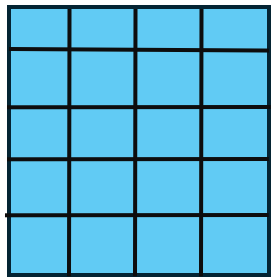
$$\frac{\text{Workload Size}}{\text{Available GPU memory}} - 1$$

Available GPU memory



<<

Workload size



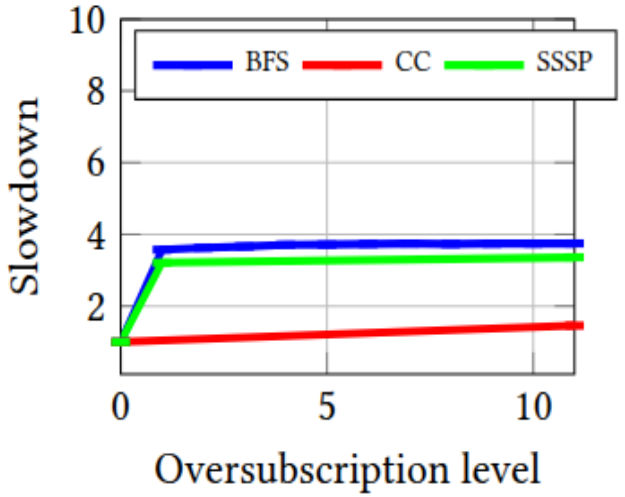
Oversubscription Analysis

Graph workloads:

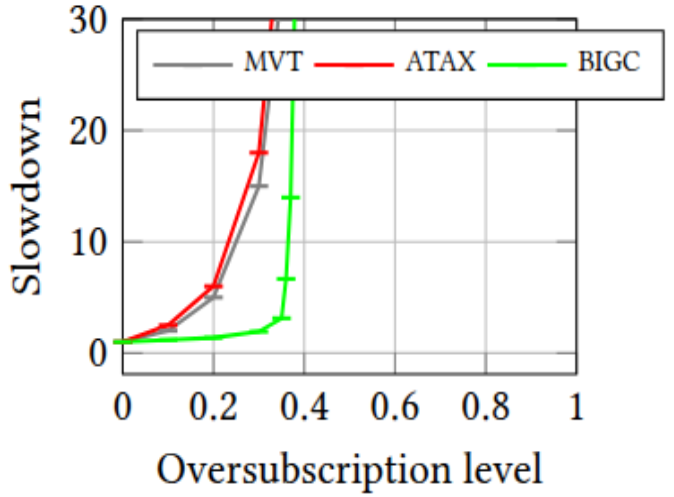
- **2x** speedup on slowdown [Figures (a) vs (c)]

Linear algebra kernels (with columnar access patterns):

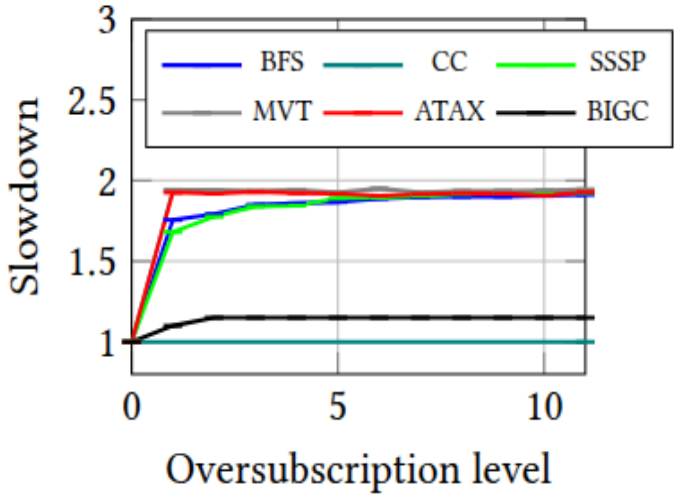
- **2x** slowdown compared to exponential slowdown [Figures (b) vs (c)]



(a) UVM for graph analytics



(b) UVM for matrix-vector operations



(c) DREAM approach

Conclusion

GPUs can now maintain their own virtual memory without depending on CPU!

- ❖ Reduction of redundant I/O by around 2x.
- ❖ Improved application performance up to 4x.
- ❖ Design can be adapted for disaggregating GPU resources with minor modifications.

❑ Has a potential to open up new research directions

Code is available at https://github.com/nnurlan008/dream

Thank you for listening!