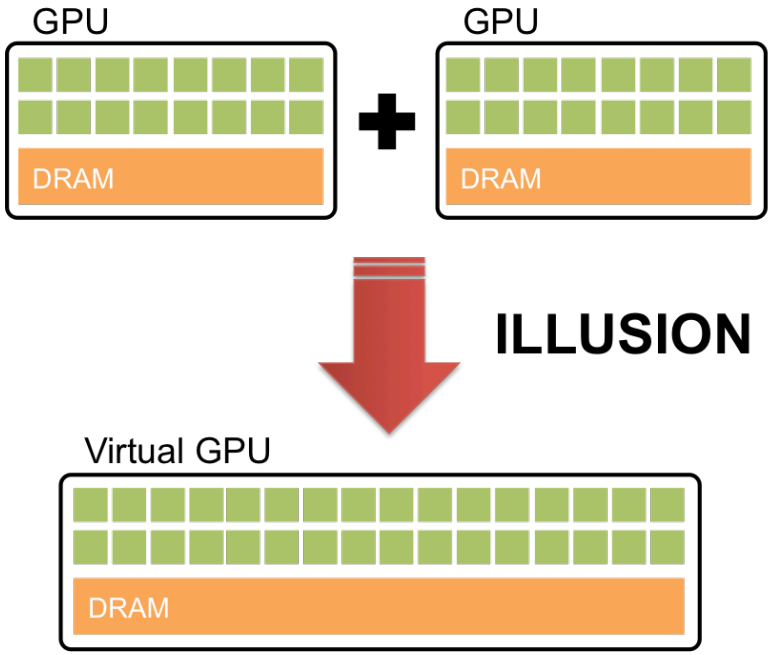
Supporting Single-GPU Abstraction through Transparent Multi-GPU Execution for Real-Time Guarantees

Wookhyun Han, **Hoon Sung Chwa**, Hwidong Bae,
Hyosu Kim, and Insik Shin

KAIST

Scope of this talk

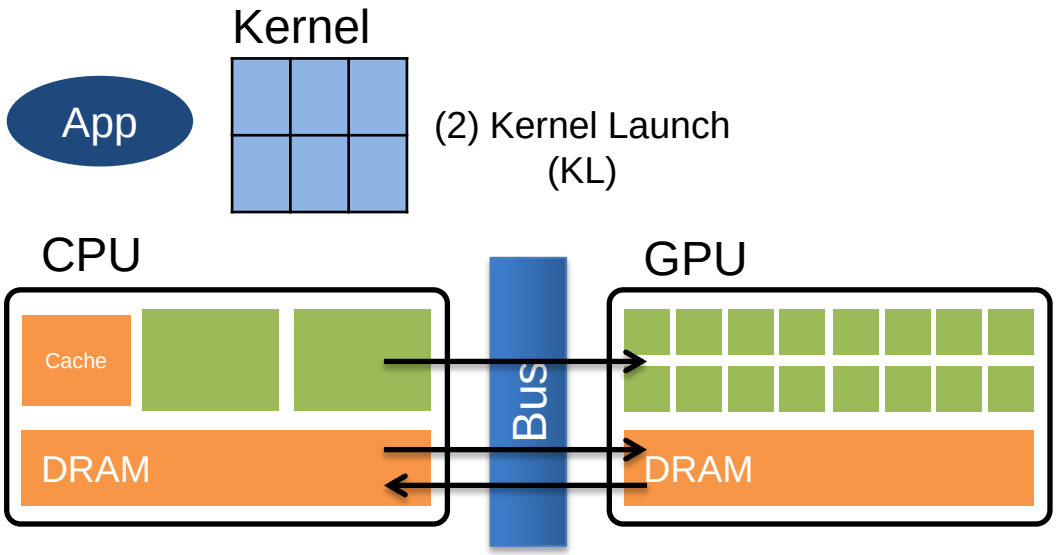
- Real-time multi-GPU system
  - Advantages of leveraging multiple GPUs
    - potential for higher performance beyond single-GPU system
  - Goal: provide an illusion of a single-GPU system



Real-time Single GPU Abstraction Framework

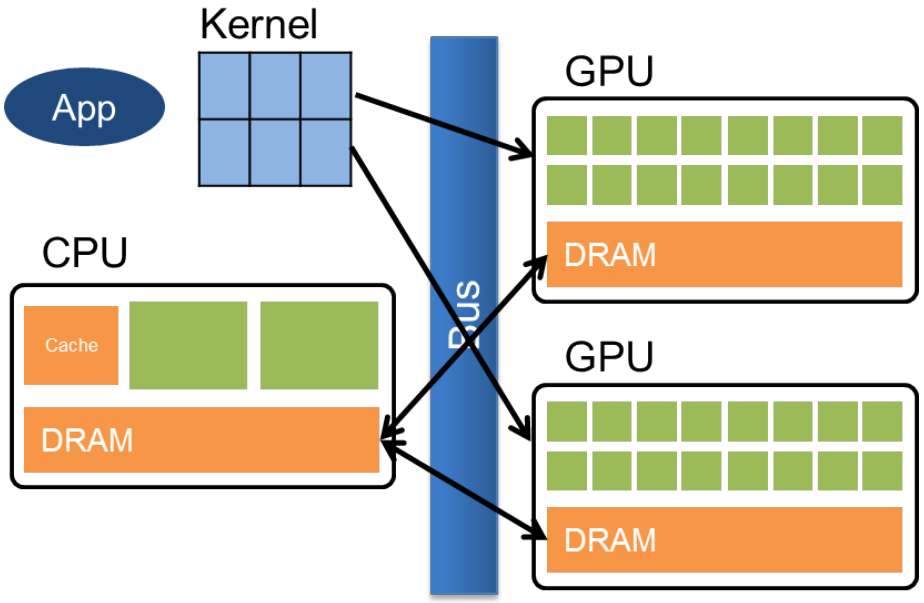
Motivation

- Current GPGPU programming model (OpenCL, CUDA)
  - GPGPU application programming is more complicated than CPU



Motivation

- Real-time Multi-GPU system
  - Multi-GPU system is much more complex than single-GPU system



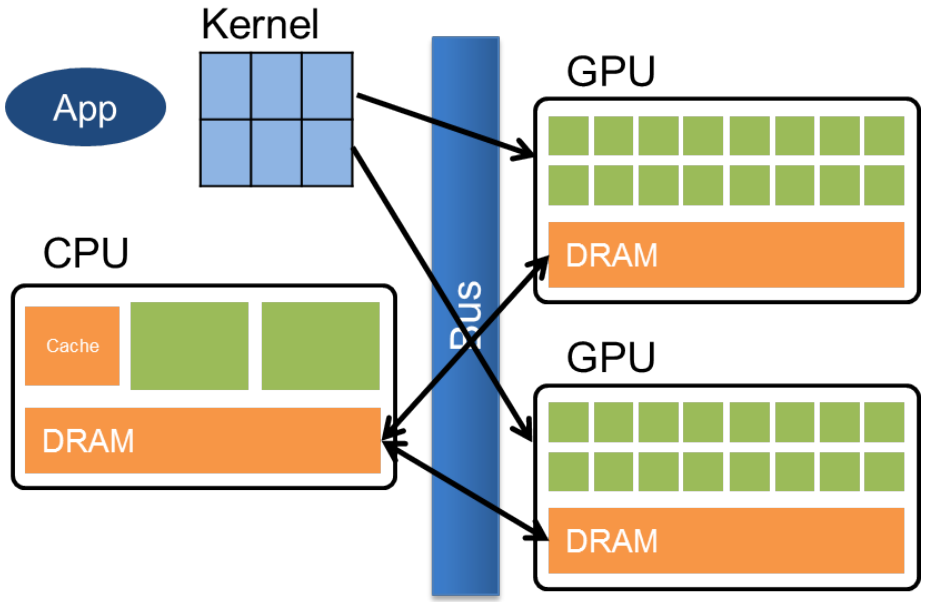
Complicated

Difficulties
- How many GPUs to use
- Workload distribution

depend on
- Hardware characteristics
- Input data set
- Behavior of other applications

Motivation

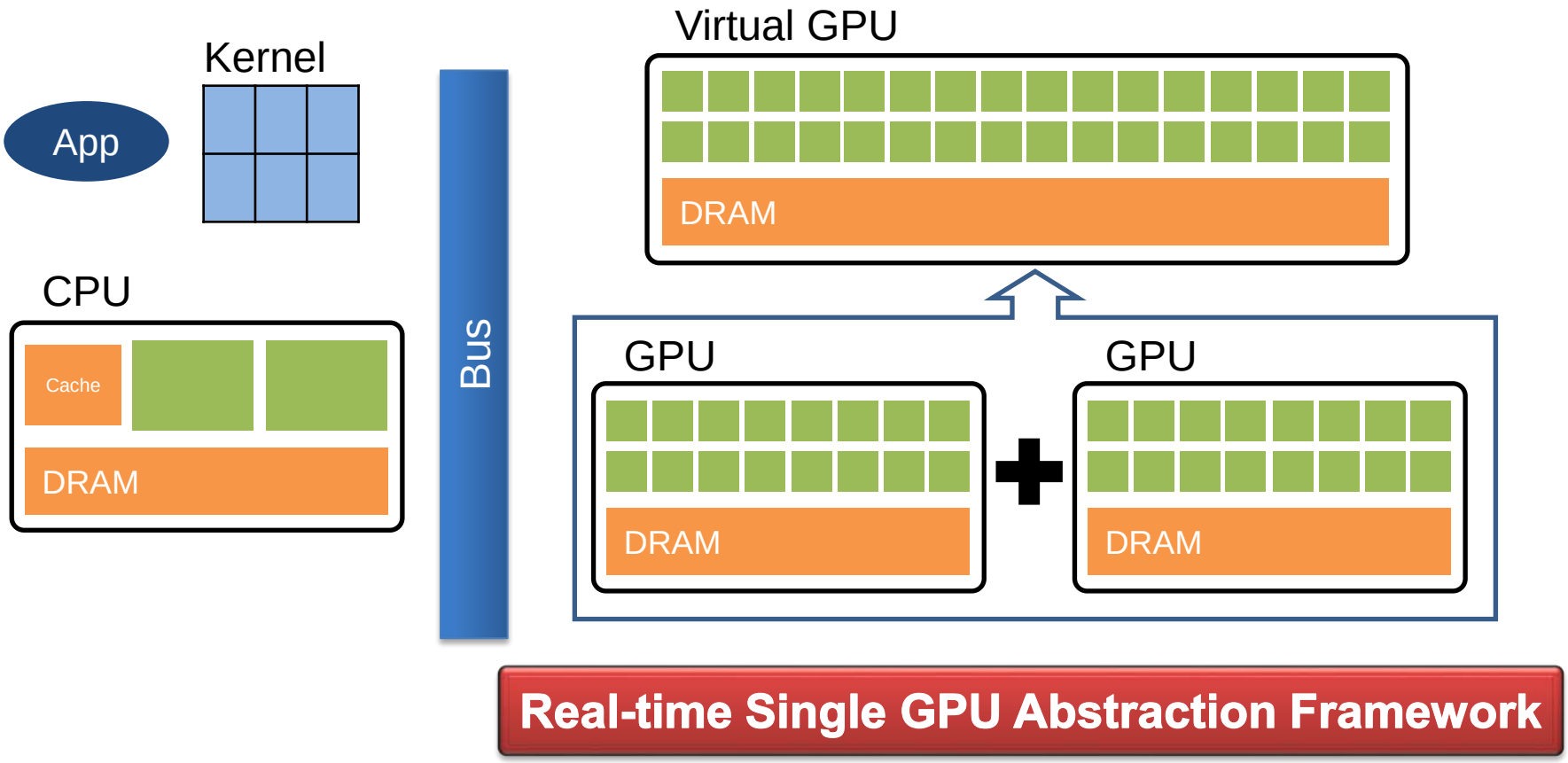
- Limitation on current GPGPU programming models
  - Single-GPU-per-kernel restriction



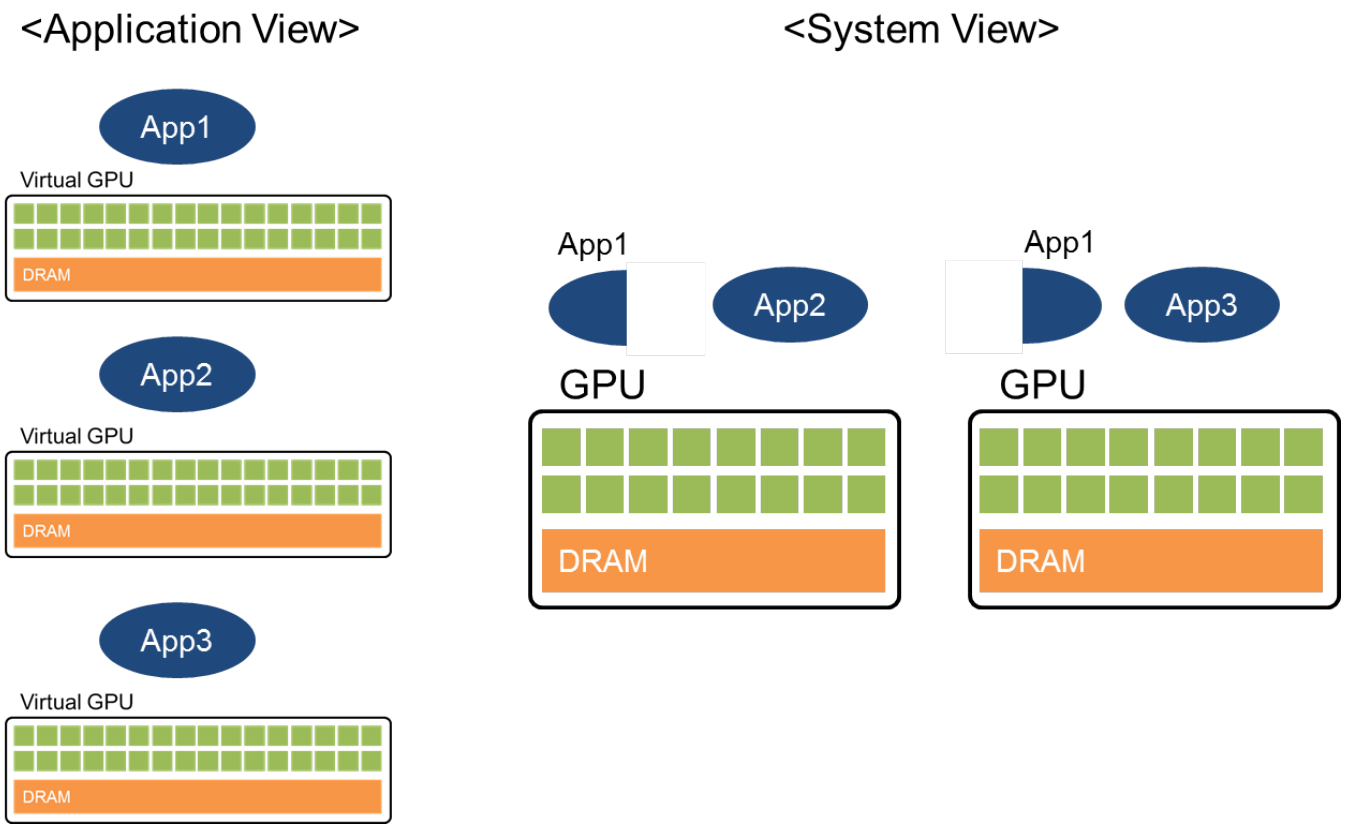
Do not provide a proper abstraction over multiple GPUs

Goal

- Real-time single-GPU abstraction Framework



Real-Time Single-GPU Abstraction Framework



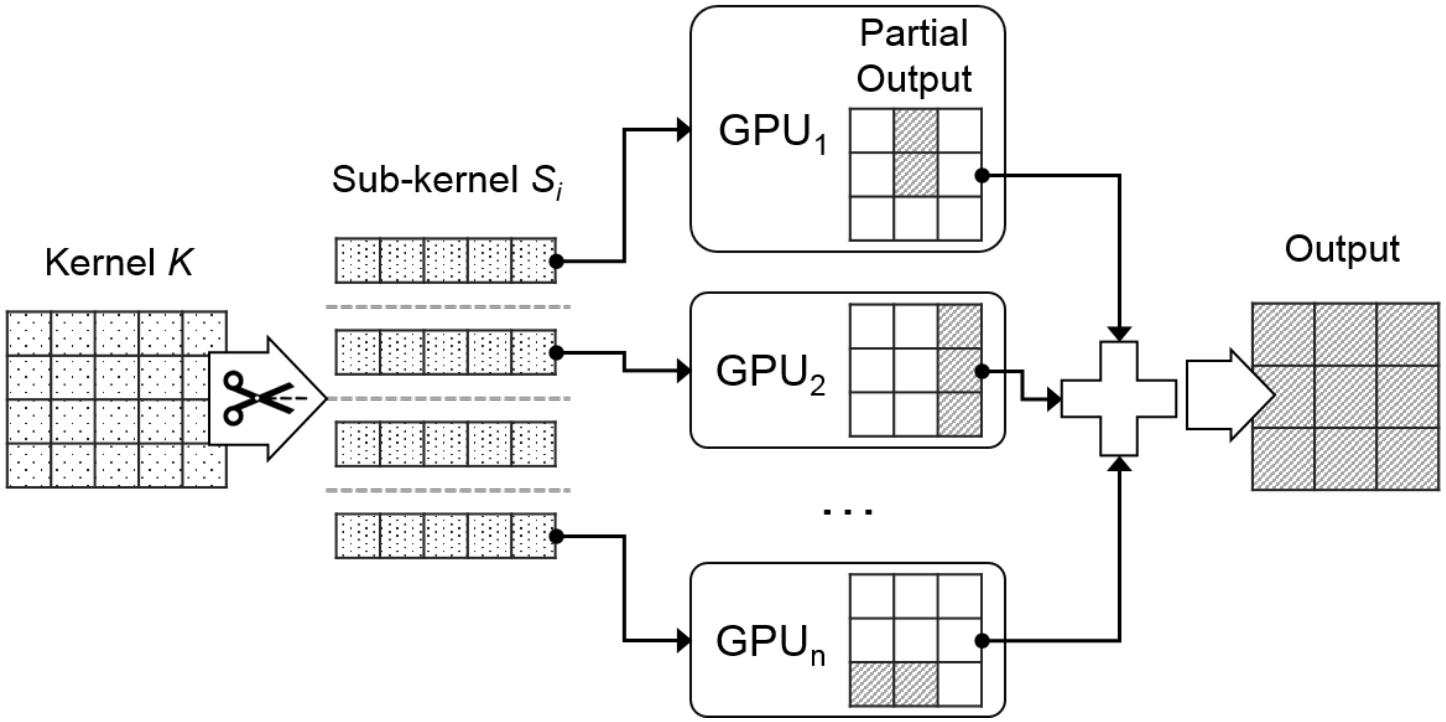
- Requirements
  - Correct single-GPU programming abstraction
  - Timing guarantees & composability

Open problems

- Mechanism for automatically distributing workload across multiple GPUs
- GPU execution mode assignment algorithm
  - The number of GPUs to use
  - Workload allocation
- Timing guarantees and composability

On-going work

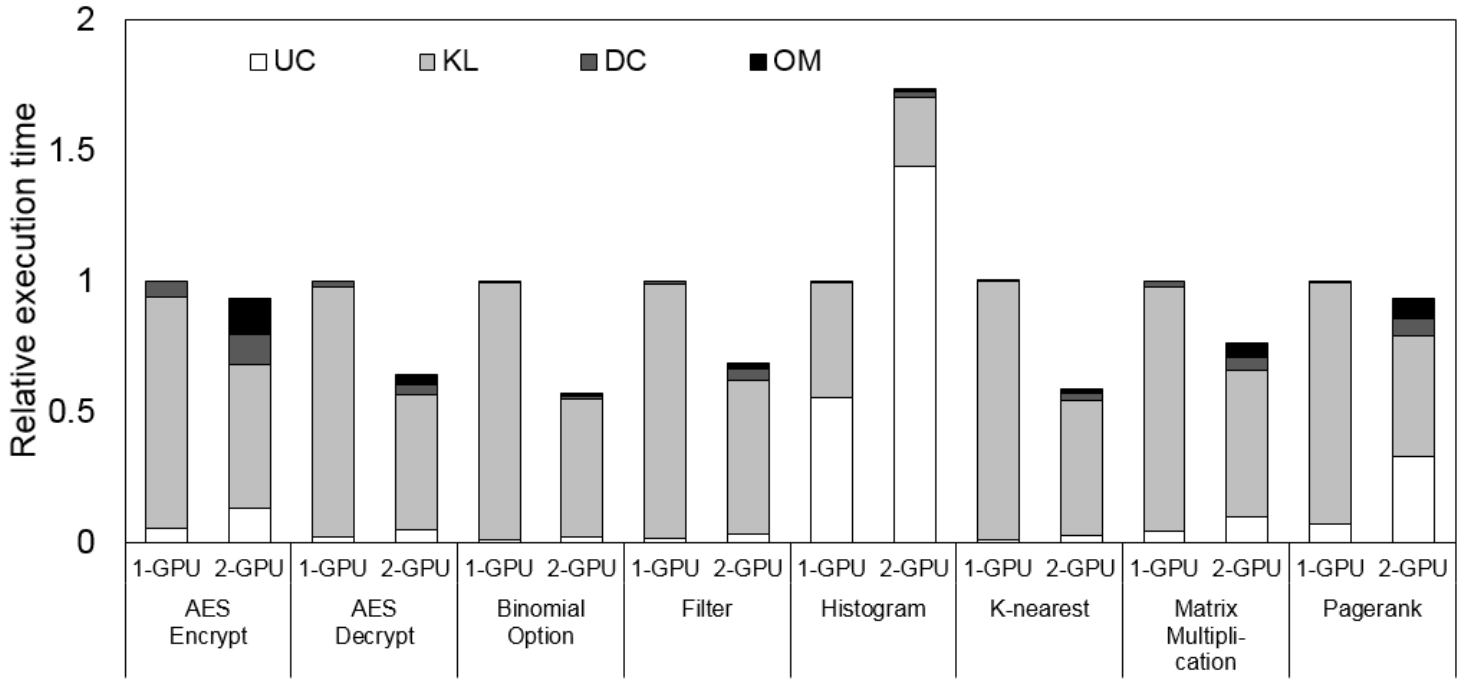
- Provide Split-and-Merge execution mechanism



  - Automatically partition the workload across multiple GPUs
  - Execute the partial workloads in parallel
  - Merge the partial results into a complete one

On-going work

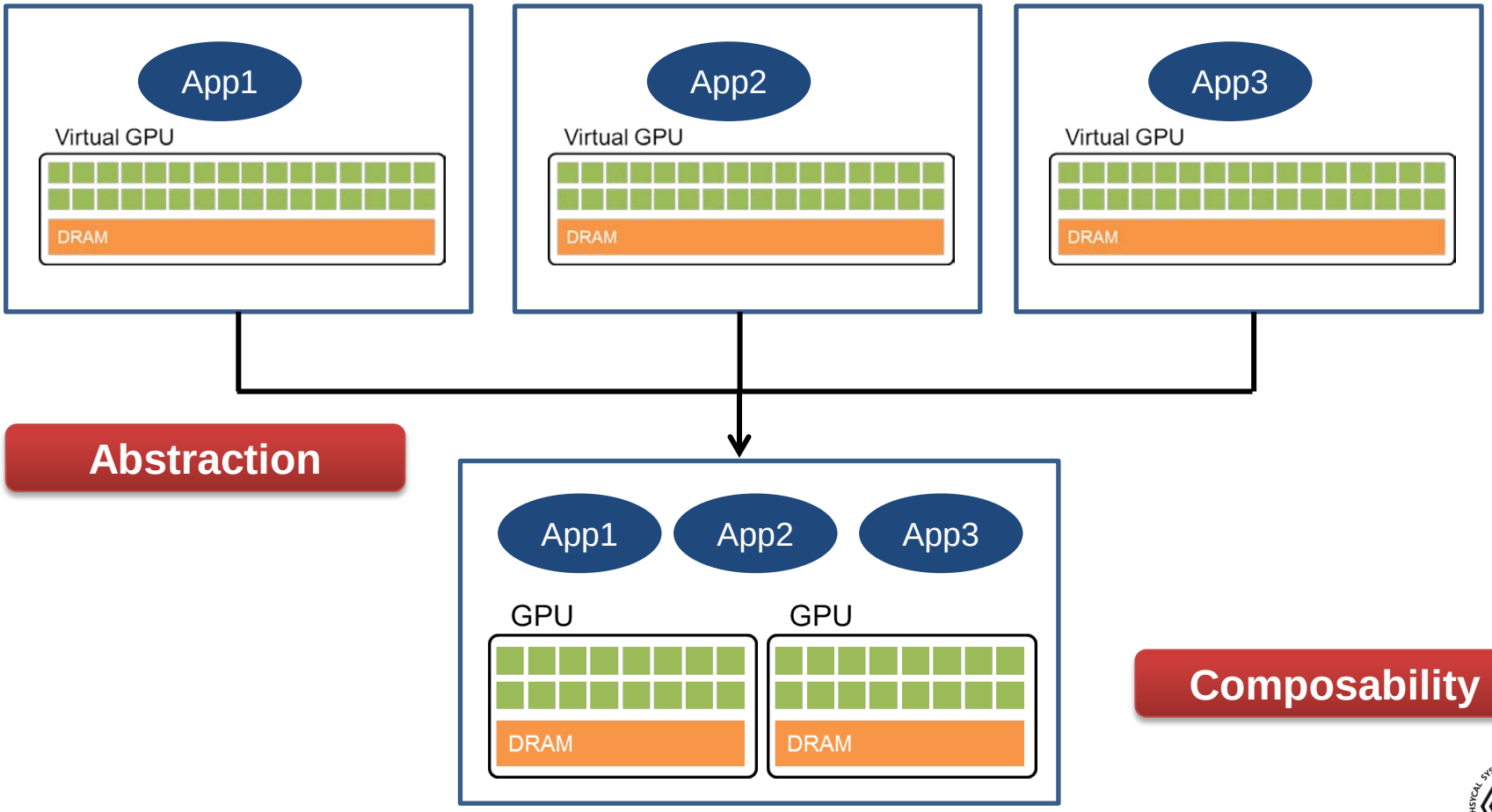
- Modeling benefits and overheads of Split-and-Merge execution



  - Predict execution time for each GPU execution mode

Future work

- Timing guarantees and composability



SUPPORTING SINGLE-GPU ABSTRACTION THROUGH TRANSPARENT MULTI-GPU EXECUTION FOR REAL-TIME GUARANTEES

Q&A?