LLMServingSim: A Simulation Infrastructure for LLM Inference Serving Systems

Jaehong Cho

Minsu Kim

Hyunmin Choi

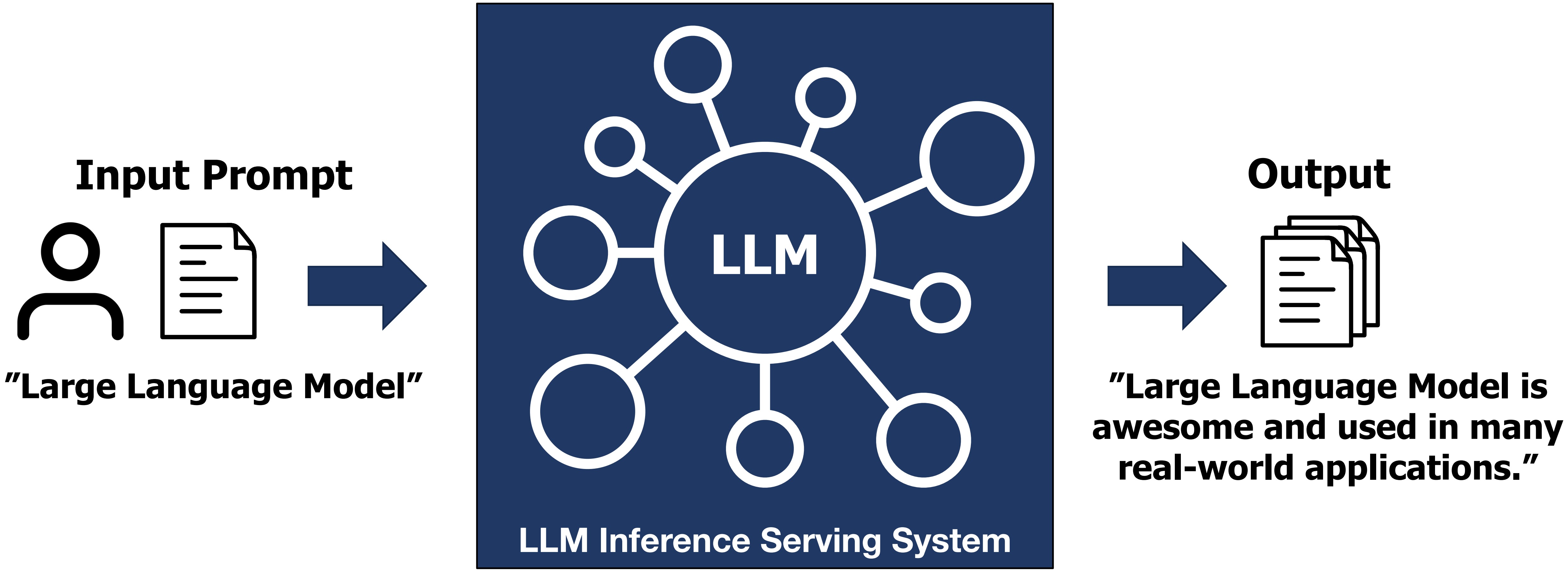
Jongse Park

KAIST

CASYS | KAIST Computer Architecture & System Lab

Large Language Model (LLM)

LLM Inference Serving System



Input Prompt

"Large Language Model"

LLM

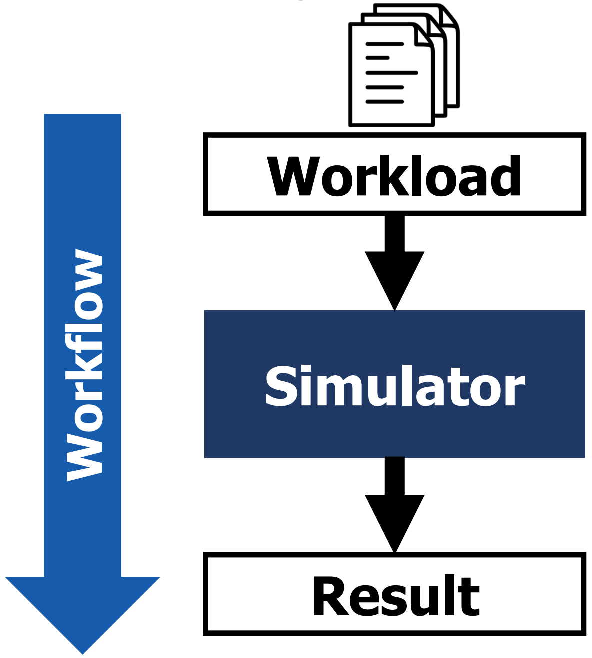
LLM Inference Serving System

Output

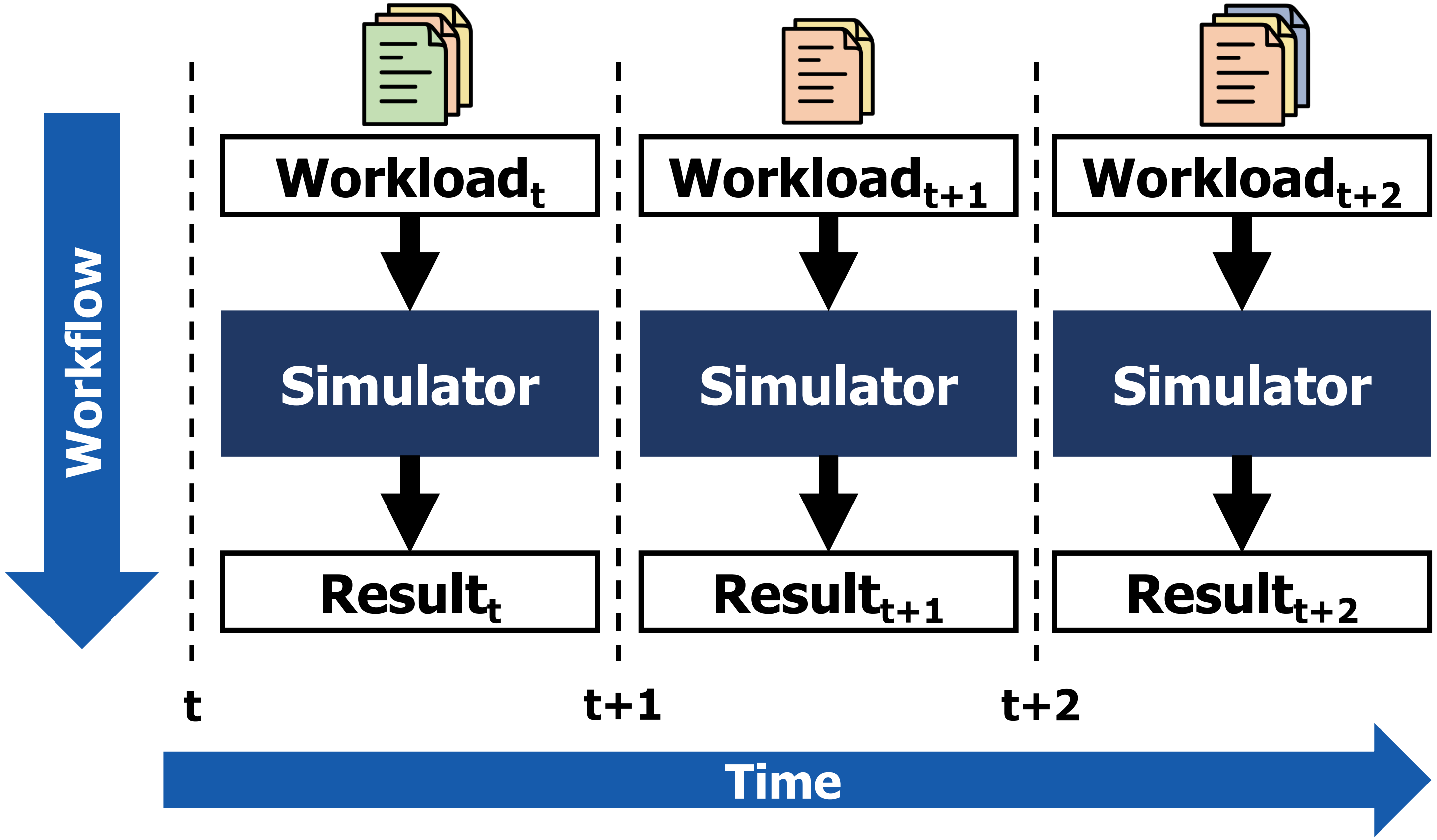
"Large Language Model is awesome and used in many real-world applications."

Challenge 1: Autoregressive LLM

- Repetitive identical iterations for ”training”

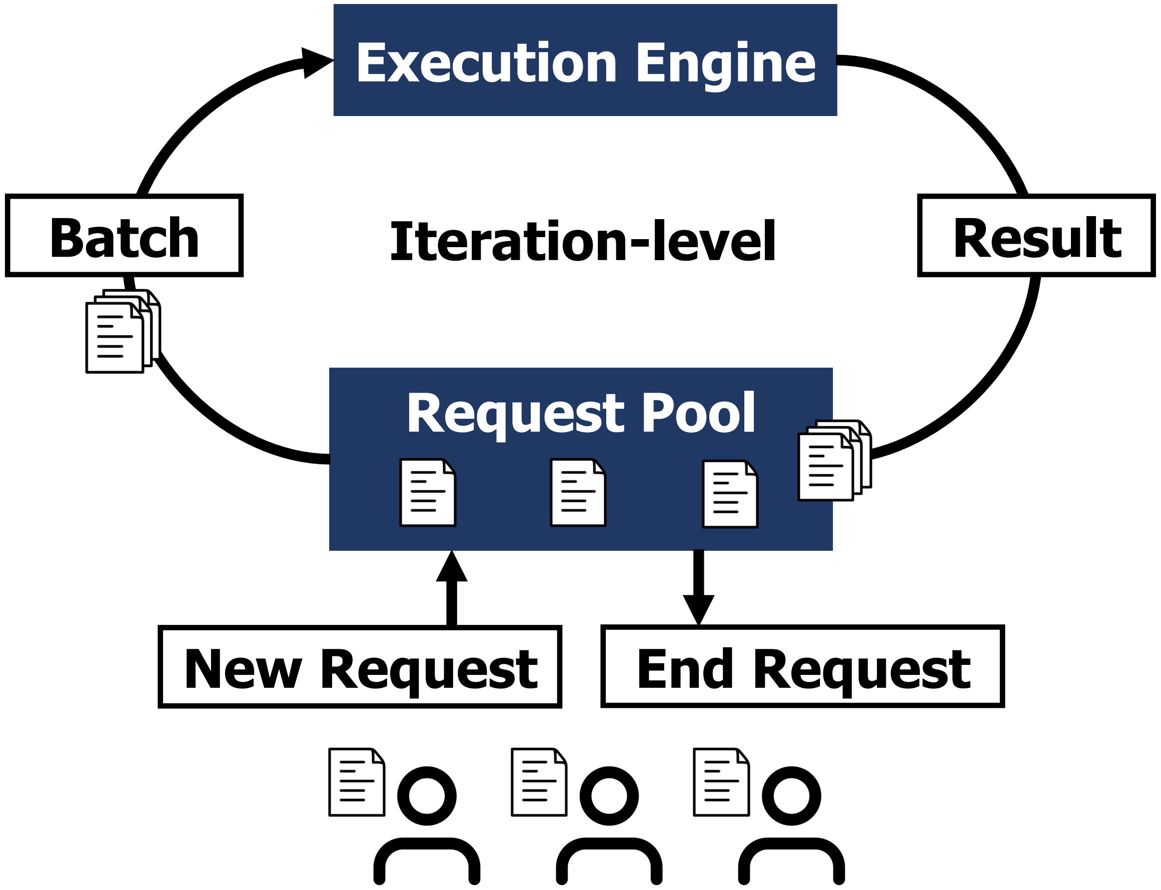


- Dynamic workloads differ in time

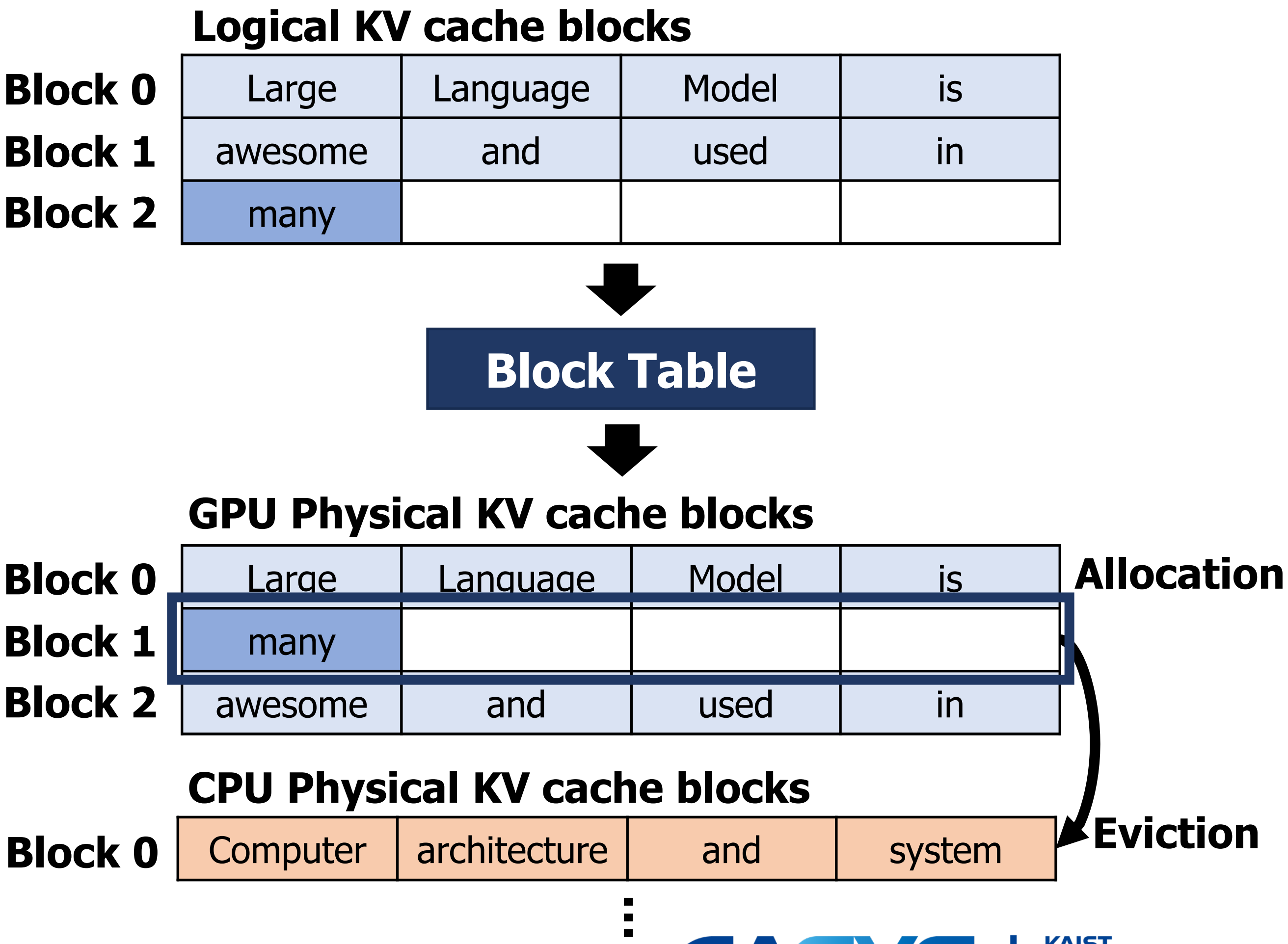


LLM Inference Serving

Iteration-level Scheduling & KV Cache Paging



Block: Unit of managing KV cache



Logical KV cache blocks

Block 0	Large	Language	Model	is
Block 1	awesome	and	used	in
Block 2	many			

GPU Physical KV cache blocks

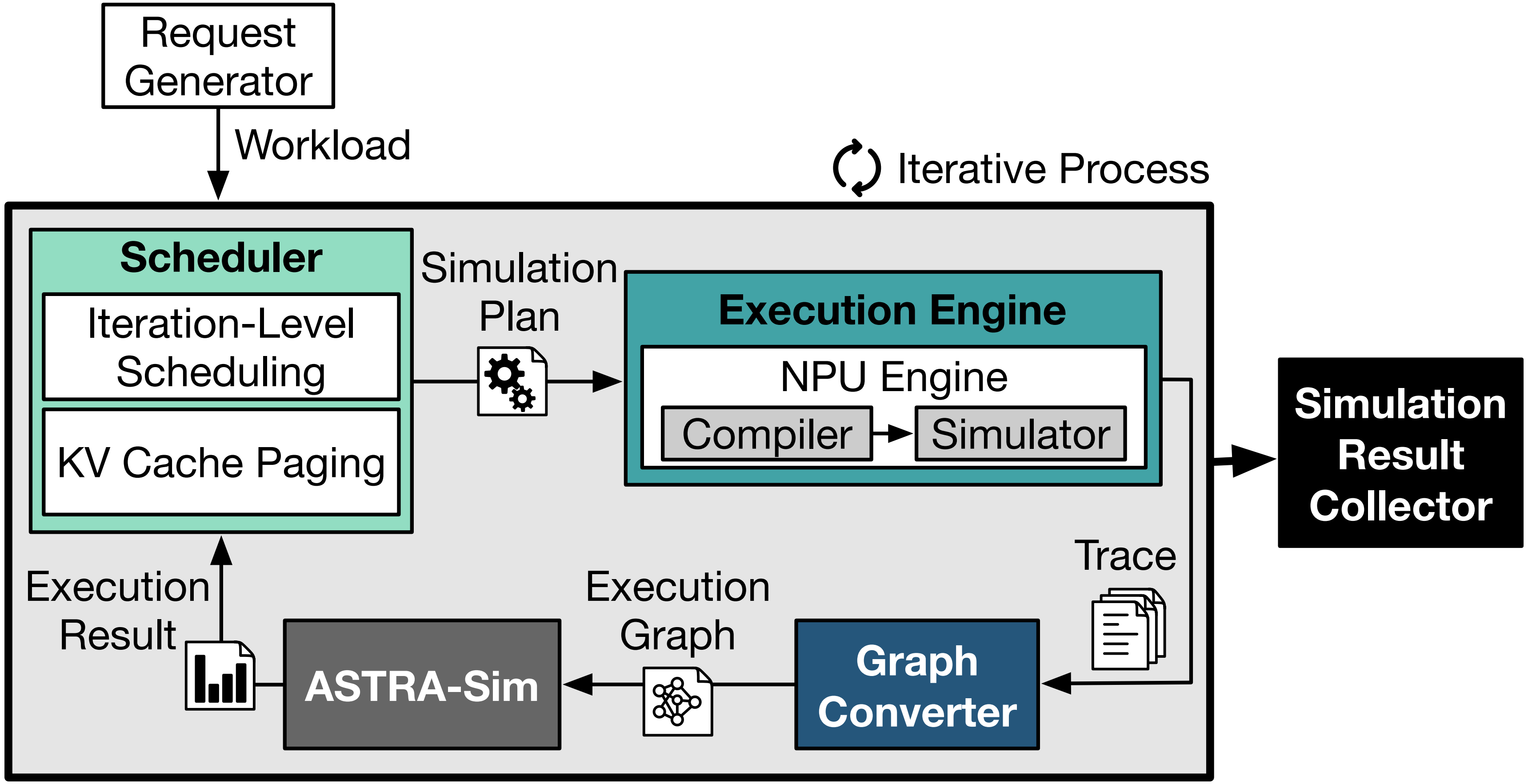
Block 0	Large	Language	Model	is
Block 1	many			
Block 2	awesome	and	used	in

CPU Physical KV cache blocks

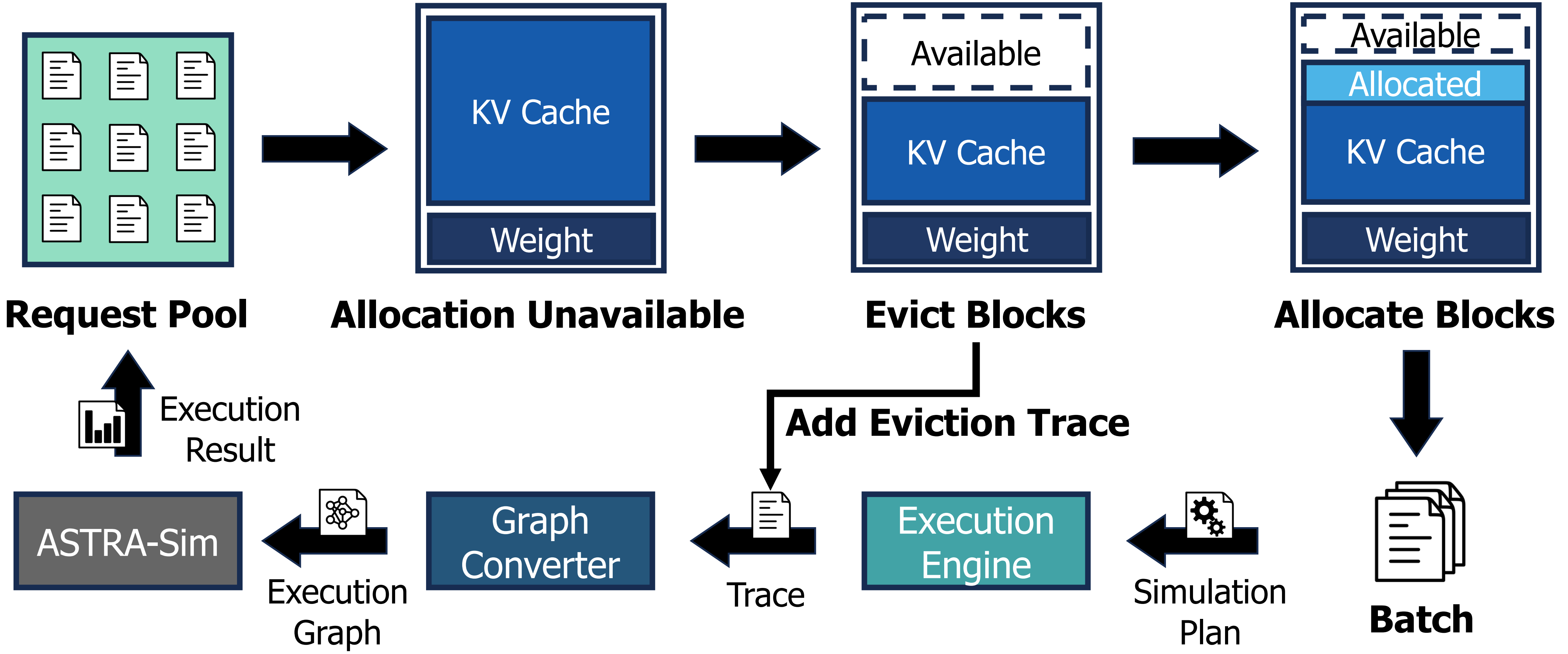
Block 0	Computer	architecture	and	system

⋮

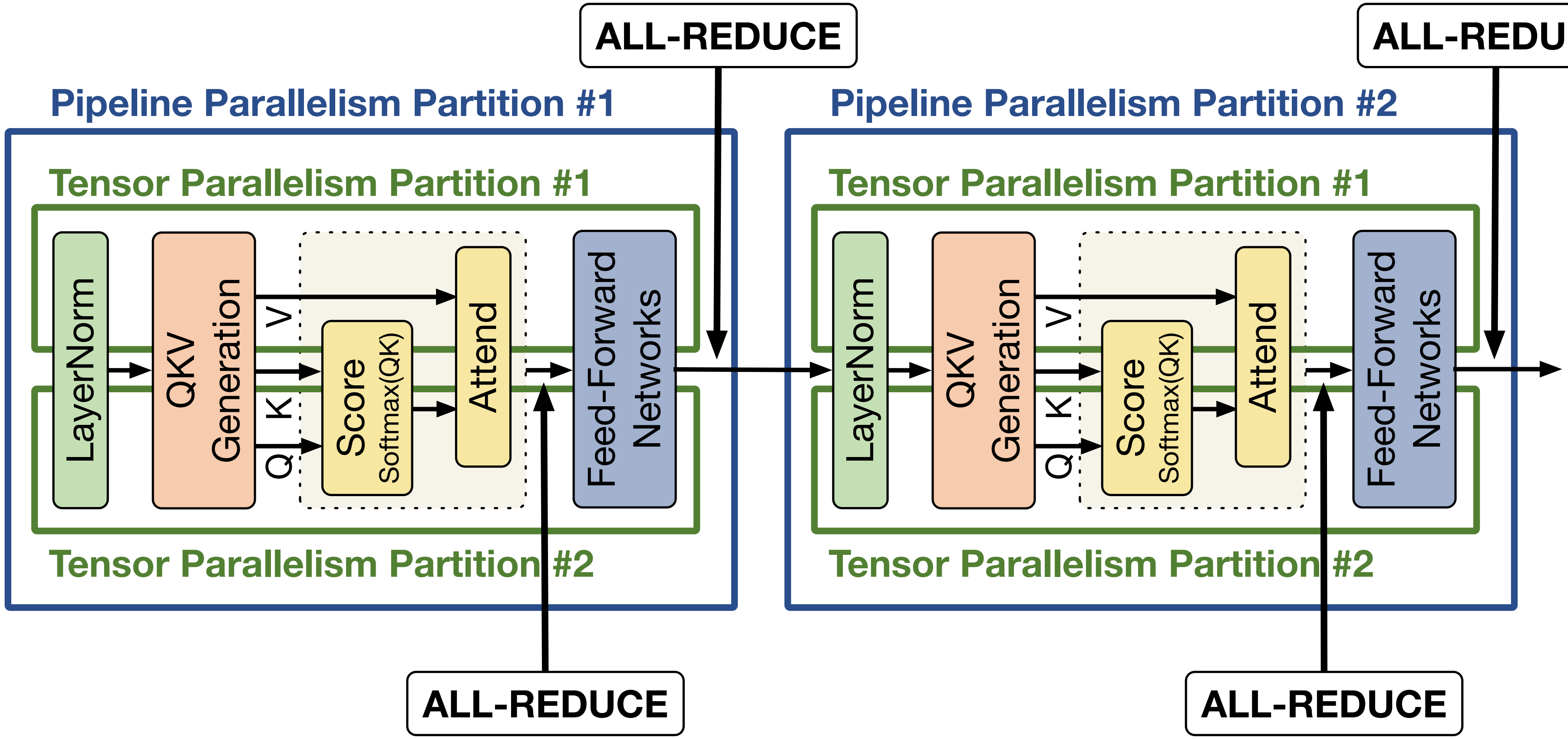
Solution 1: Iteration-level Scheduling



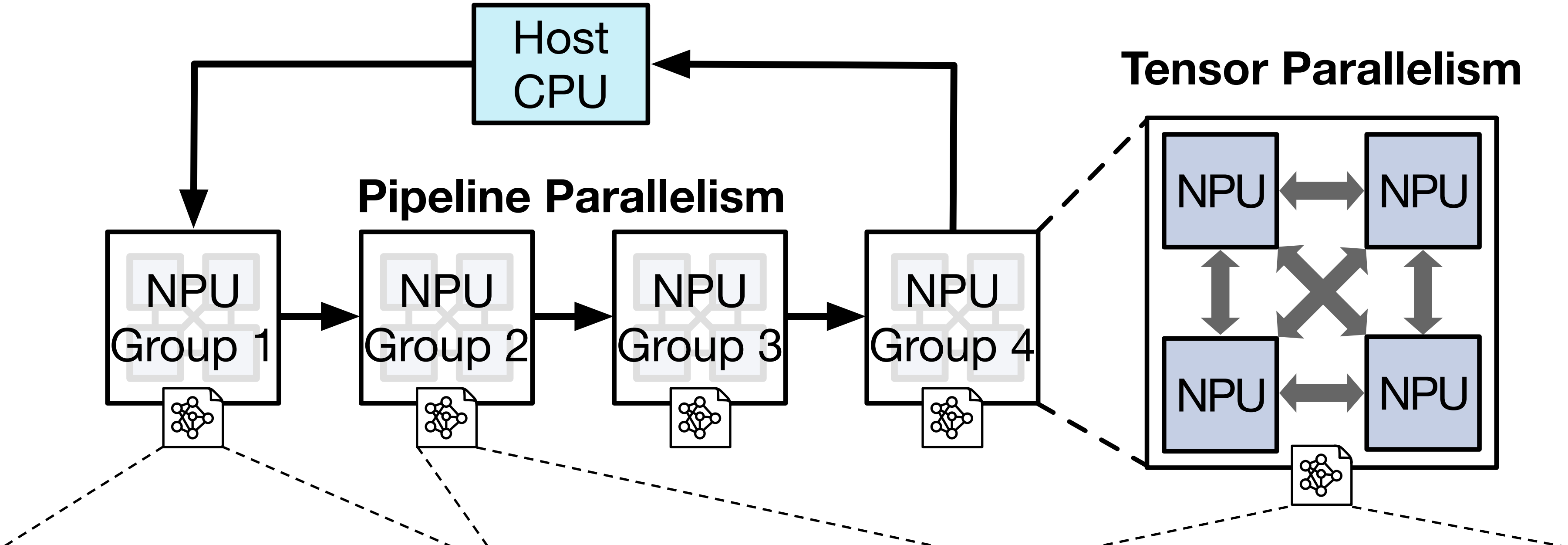
Solution 1: KV Cache Paging



Challenge 2: LLM Specific Parallelism


ALL-REDUCE
ALL-REDUCE
Pipeline Parallelism Partition #1
Pipeline Parallelism Partition #2
Tensor Parallelism Partition #1
Tensor Parallelism Partition #1
LayerNorm
QKV Generation
V
K
Q
Score
Softmax(QK)
Attend
Feed-Forward Networks
LayerNorm
QKV Generation
V
K
Q
Score
Softmax(QK)
Attend
Feed-Forward Networks
Tensor Parallelism Partition #2
Tensor Parallelism Partition #2
ALL-REDUCE
ALL-REDUCE

Solution 2: LLM Graph Converter



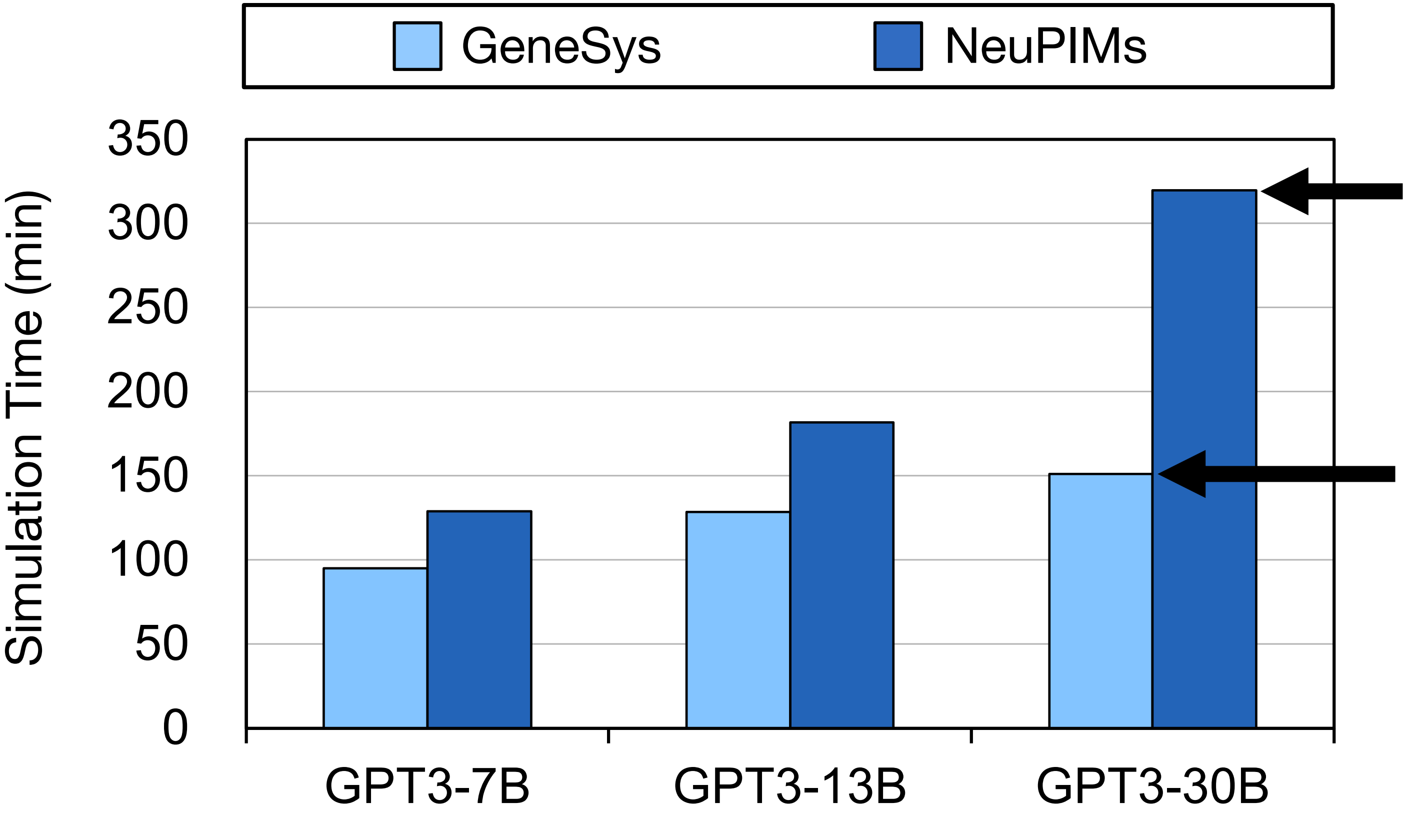
ID	Op	Src	Dest
0	Embd	-	-
1		-	-
…		-	-
N	**Send**	Group 1	Group 2

ID	Op	Src	Dest
0	**Recv**	Group 1	Group 2
1		-	-
…		-	-
N	**Send**	Group 2	Group 3

ID	Op	Src	Dest
…		-	-
k	**ALL-REDUCE**	ALL	ALL
…		-	-

Challenge 3: Slow Simulation Time

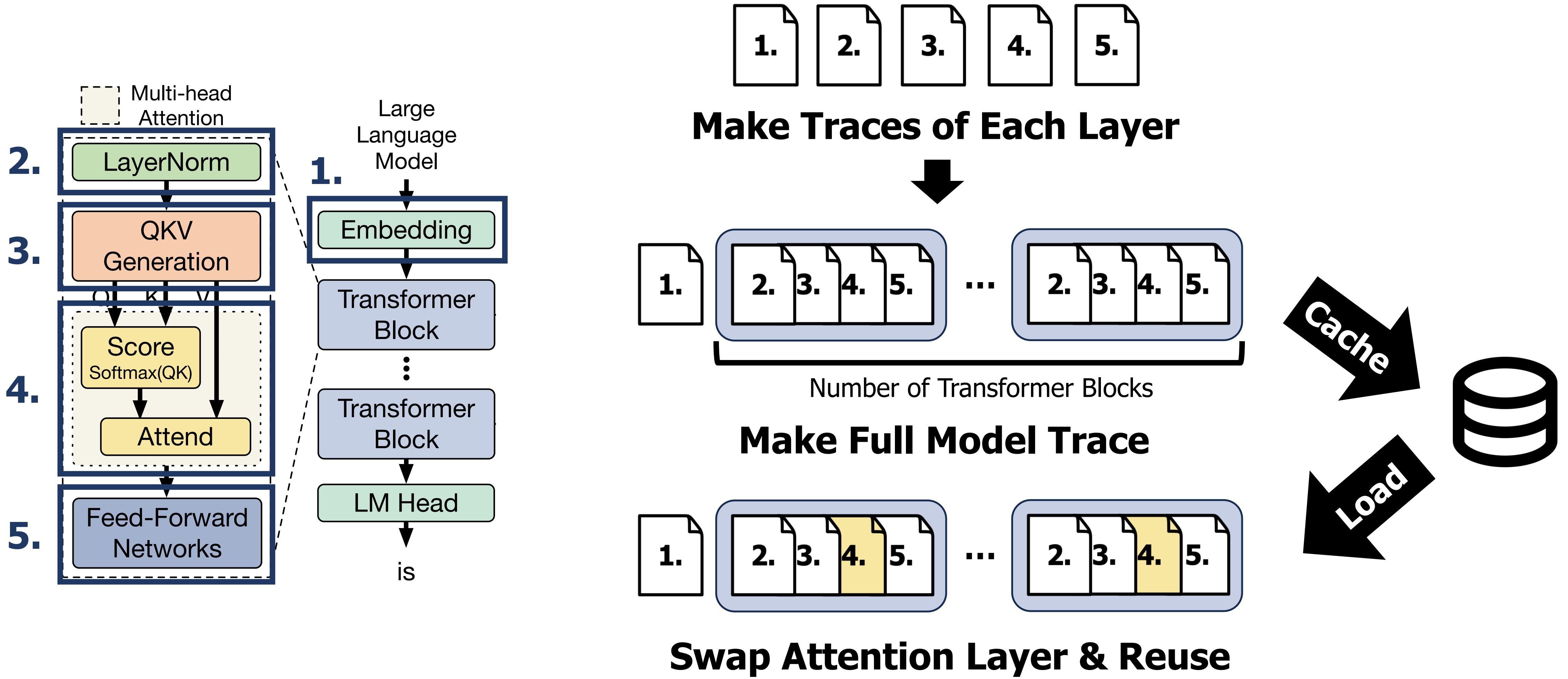
- Batch 32
- Sequence Length 512
- 1 Iteration



More than **5 hours** for 1 iteration

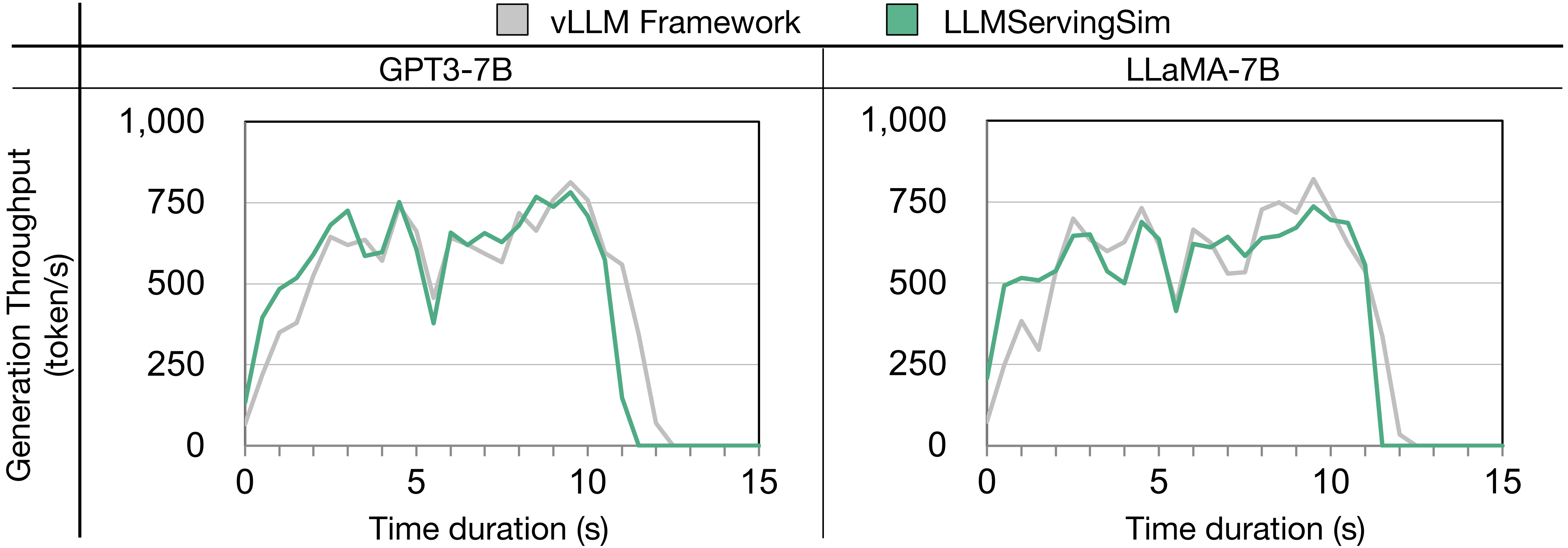
More than **2 hours** for 1 iteration

Solution 3: Computation Reuse



Evaluation

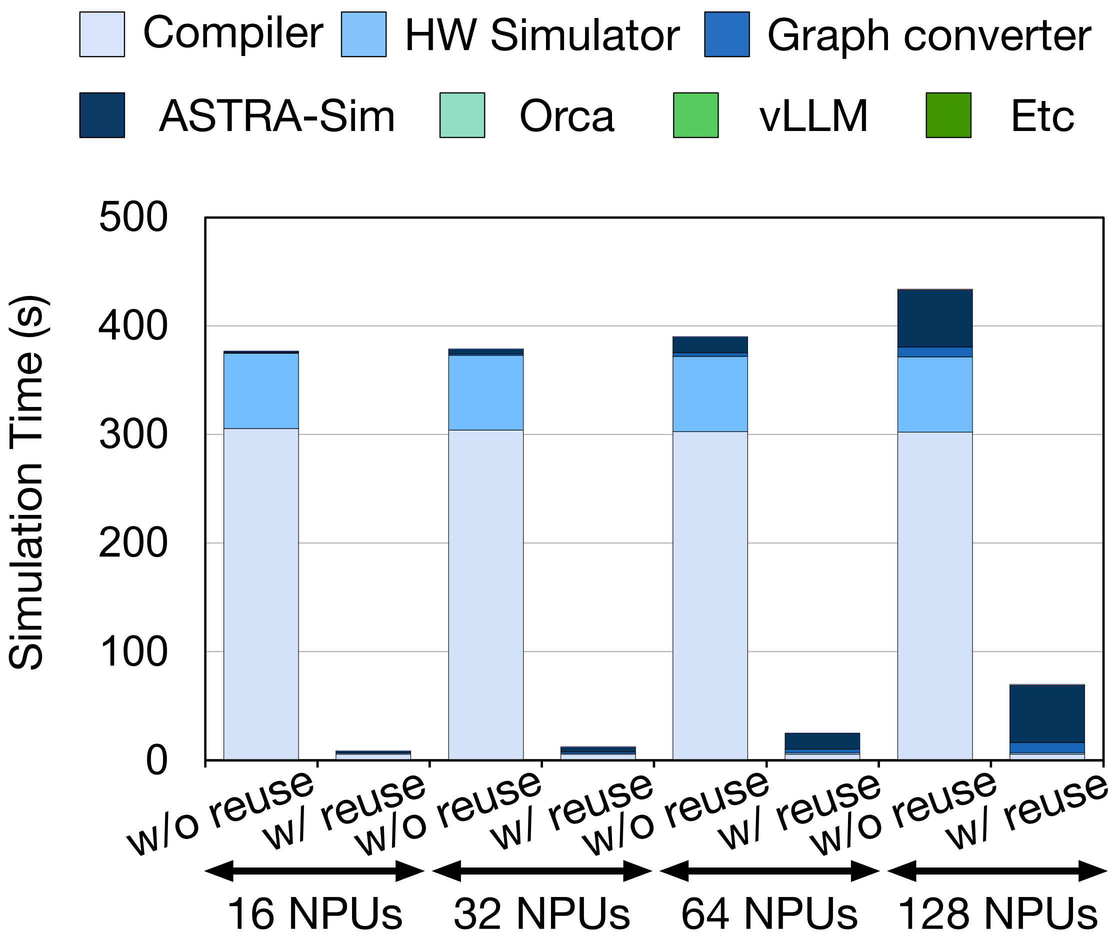
Validation



- ShareGPT sampled requests, request arrival pattern using Poisson distribution
- Average error rate **14.7%**

Evaluation

Performance of Computation Reuse



- GPT-3 175B 1 iteration
- Sequence Length: 2048

- Without reuse: **~400 sec**
- With reuse: **<1 min**

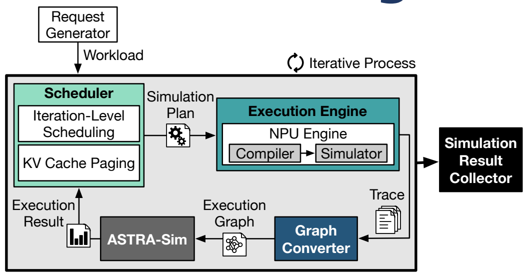
- **18.7×** speedup using computation reuse

Conclusion

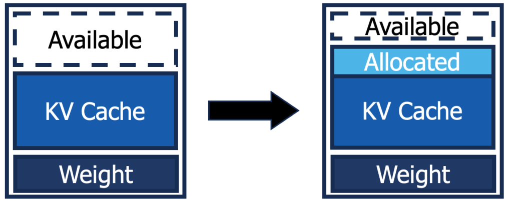
- **Large scale LLM inference serving system simulator**

- **14.7%** error rate against real GPU-based LLM serving system
- **18.7x** speedup using computation reuse

Iteration-level Scheduling

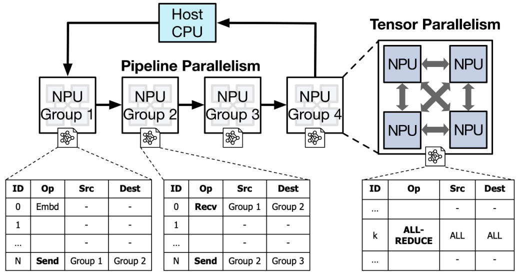


KV Cache Paging



Evict Blocks **Allocate Blocks**

LLM Parallelism

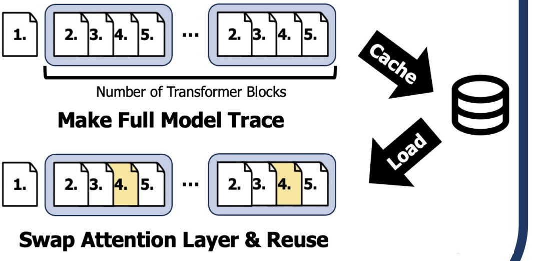


ID	Op	Src	Dest
0	Embd	-	-
1		-	-
...		-	-
N	**Send**	Group 1	Group 2

ID	Op	Src	Dest
0	**Recv**	Group 1	Group 2
1		-	-
...		-	-
N	**Send**	Group 2	Group 3

ID	Op	Src	Dest
...		-	-
k	**ALL-REDUCE**	ALL	ALL
...		-	-

Computation Reuse


1.
2. 3. 4. 5.
...
2. 3. 4. 5.
Number of Transformer Blocks
Make Full Model Trace
Cache
Load
1.
2. 3. 4. 5.
...
2. 3. 4. 5.
Swap Attention Layer & Reuse