Throughput-Oriented LLM Inference via KV-Activation Hybrid Caching with a Single GPU

Sanghyeon Lee
Hongbeen Kim
Soojin Hwang
Guseul Heo
Minwoo Noh
Jaehyuk Huh

School of Computing

KAIST

KAIST

ICCD 2025
43rd IEEE International Conference on Computer Design

Classification of LLM Workloads

Latency Critical

50~100ms / token [1]

Multi-GPUs

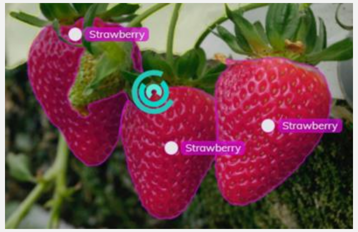
Chatbot

Voice chat

Latency Tolerant

Cost-aware SLO

Single ~ Multi-GPUs



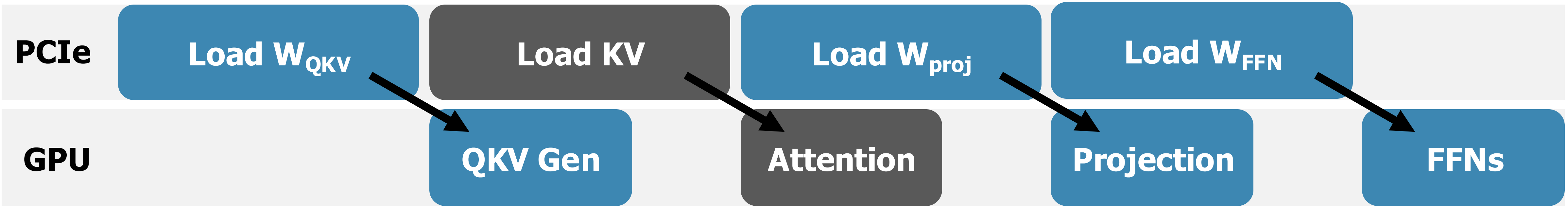
Data labeling

Benchmarking

[1] Jacoby, Derek, et al. "Human Latency Conversational Turns for Spoken Avatar Systems." *UIST*, 2024

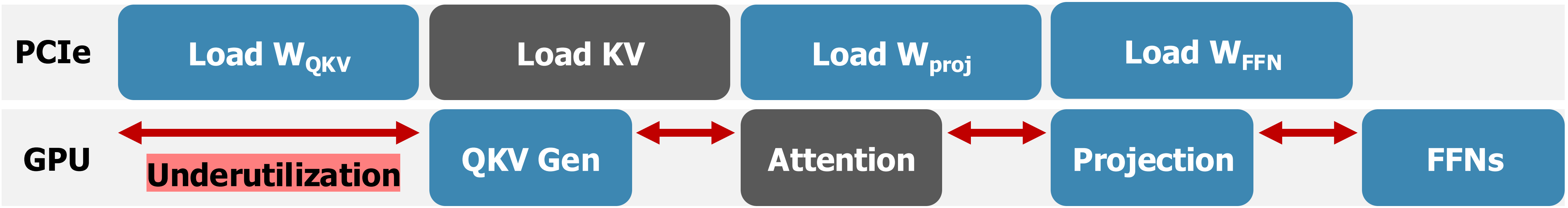
Host Memory Offloading for LLM

- **Problem:** Using multiple GPUs to serve large LLMs is extremely costly.
- **Solution:** Offload LLM weights & KV Cache to host memory.



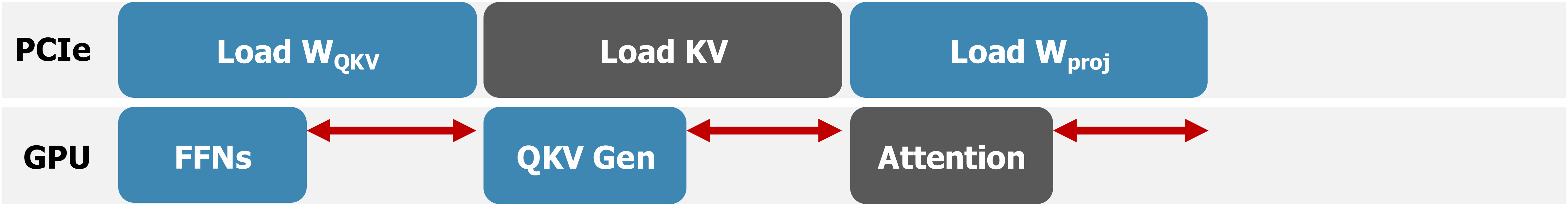
Host Memory Offloading for LLM

- **Problem:** Using multiple GPUs to serve large LLMs is extremely costly.
- **Solution:** Offload LLM weights & KV Cache to host memory.



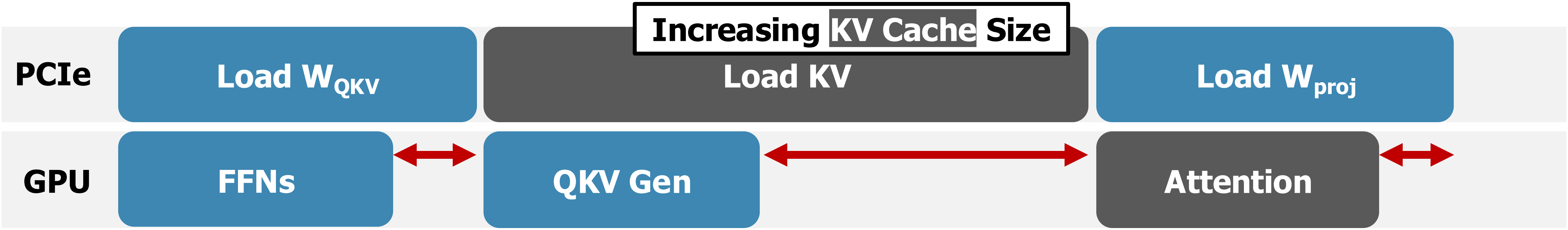
Limited Benefit of Large Batch Size

- Increasing the batch size offers **diminishing returns**.
- Larger batches lead to a massive KV Cache, creating a new PCIe bottleneck.



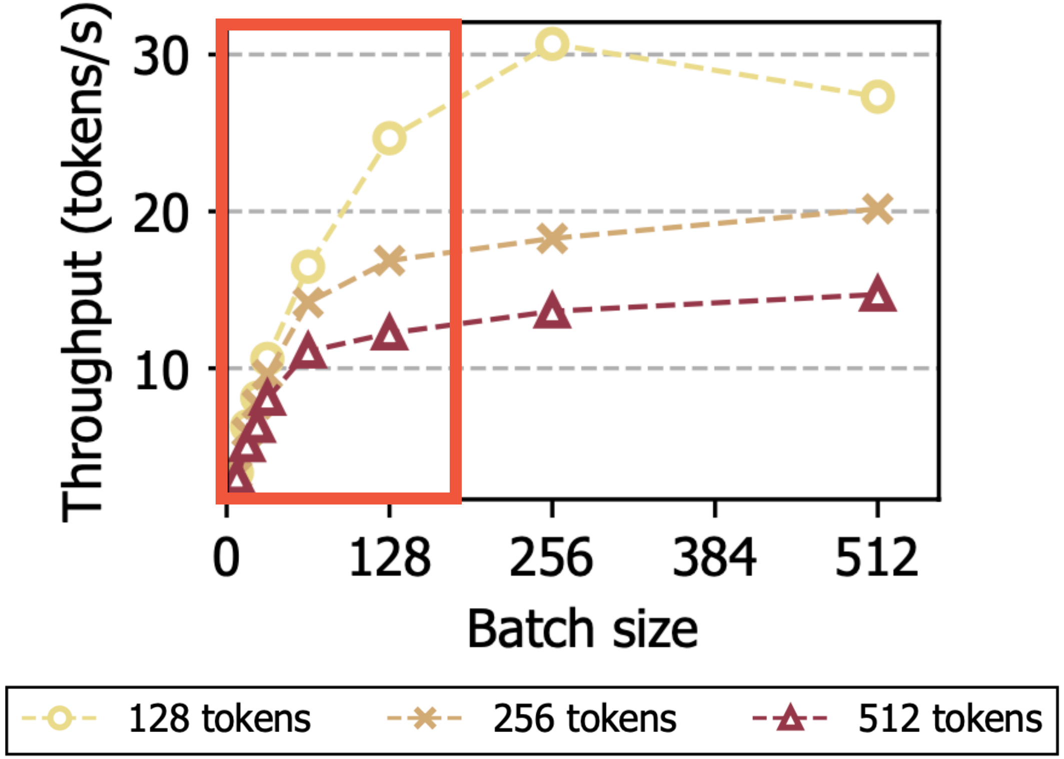
Limited Benefit of Large Batch Size

- Increasing the batch size offers **diminishing returns**.
- Larger batches lead to a massive KV Cache, creating a new PCIe bottleneck.



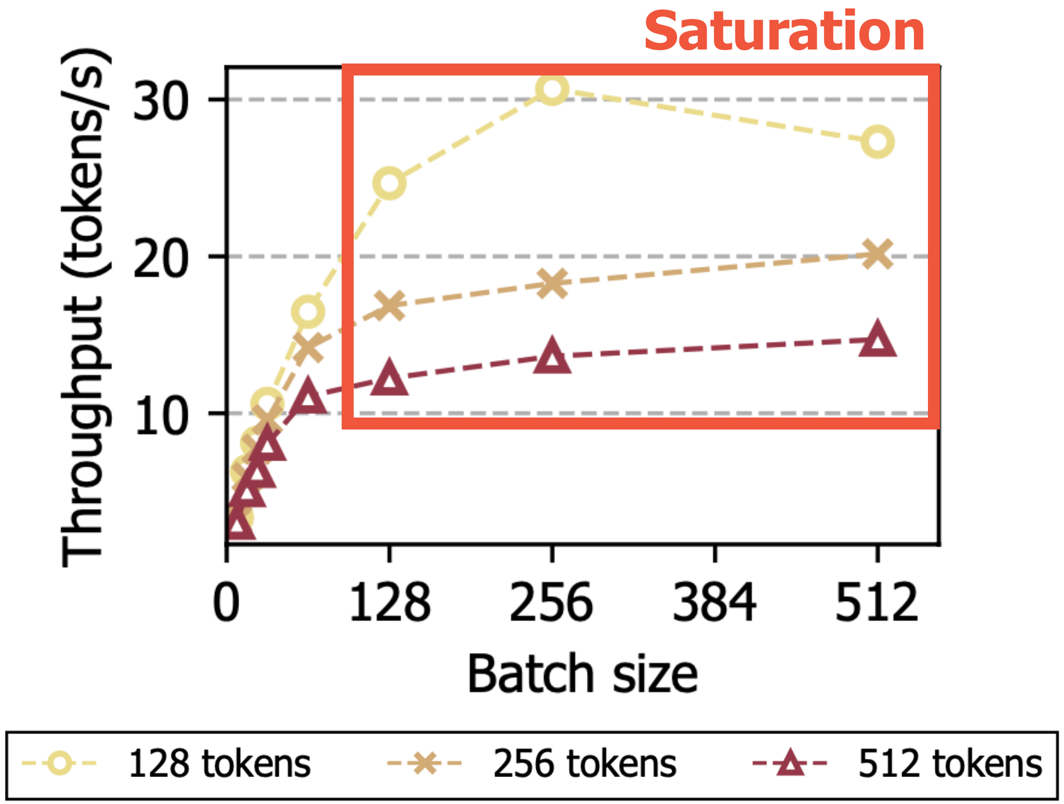
Limited Benefit of Large Batch Size

- Increasing the batch size shows diminishing returns on throughput.
- KV Cache size grows linearly with the batch size.



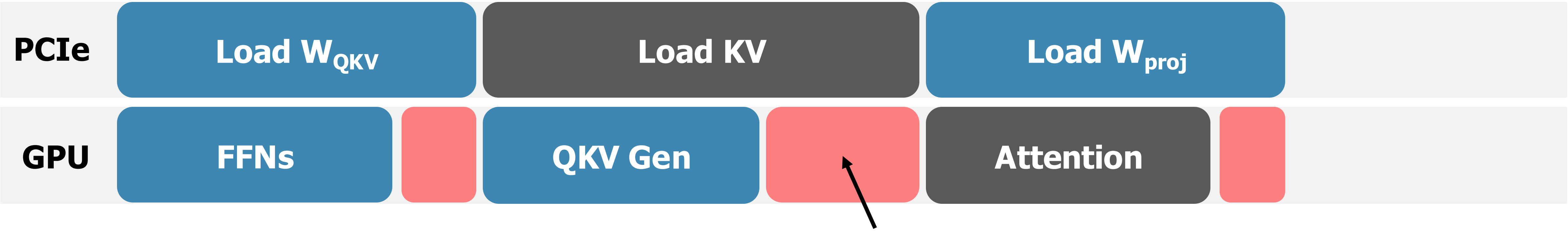
Limited Benefit of Large Batch Size

- Increasing the batch size shows diminishing returns on throughput.
- KV Cache size grows linearly with the batch size.



Opportunity: KV Cache Recomputation

- **Key Insight:** Trade slow communication (PCIe) for fast computation (GPU).
- **Solution:** Recompute the KV Cache on-the-fly, avoiding the data transfer.



PCIe

Load W_{QKV}

Load KV

Load W_{proj}

GPU

FFNs

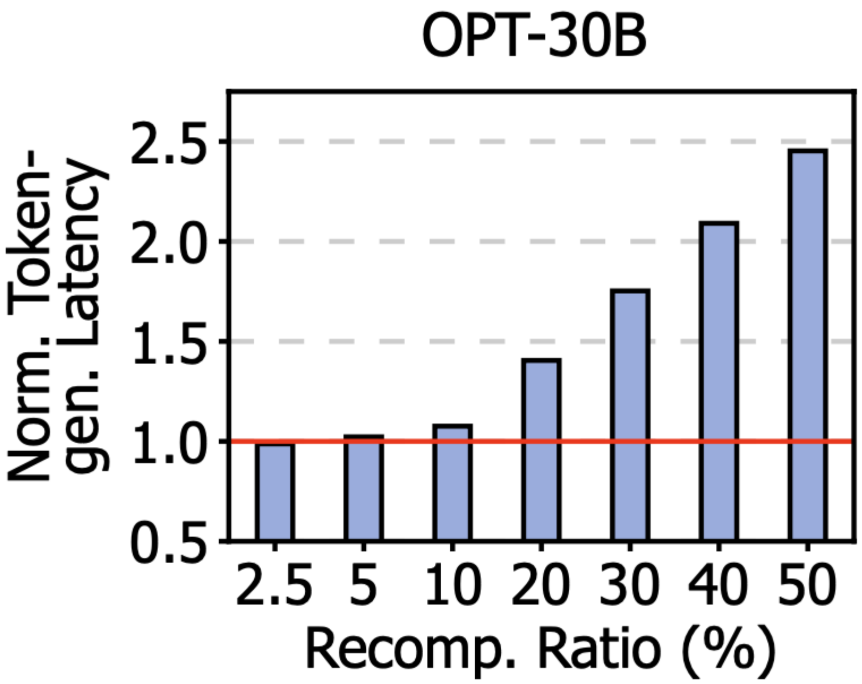
QKV Gen

Attention

Recompute KV cache

Limitation of KV Cache Recomputation

- KV recomputation is computationally expensive.
- Even a **20%** recomputation leads to a **1.45x** slowdown for OPT-30B



OPT-30B

No Recomputation Latency

Problems

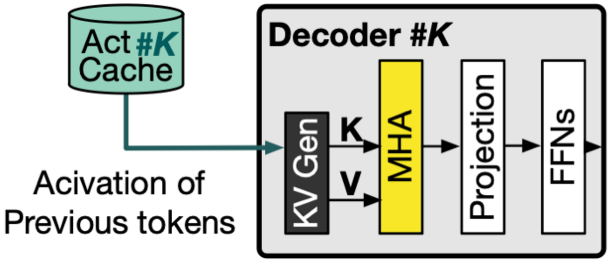
- Host memory offloading is a low-cost solution.
- But it suffers from severe GPU underutilization.
- And KV recomputation offers only marginal gains.

We need a more intelligent approach

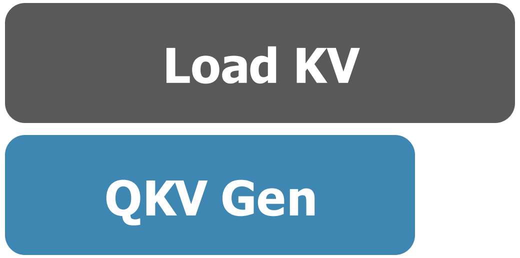
to the computation-communication trade-off

Capture: Overview

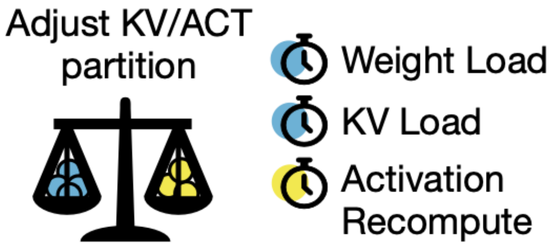
KV-Activation Hybrid Caching



Asynchronous Engine

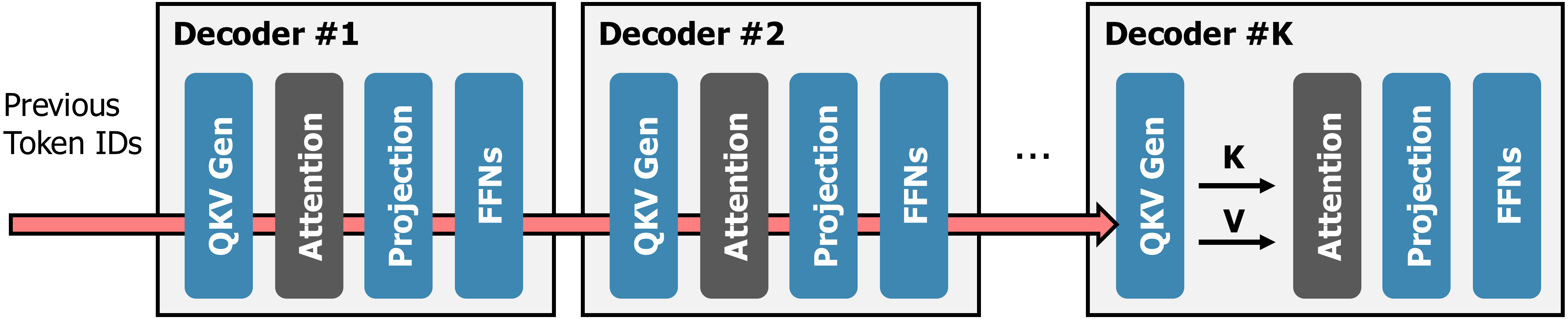


Cache Management Policy



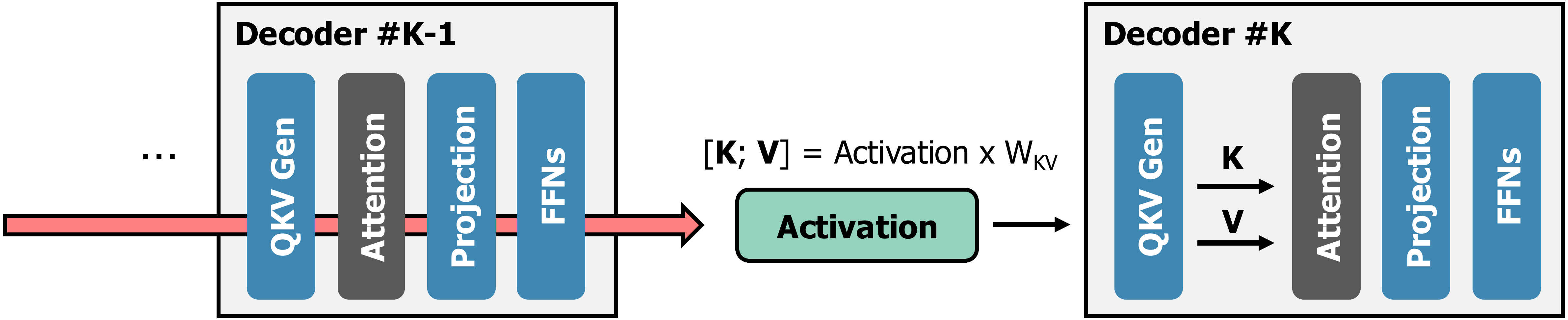
Potential of Activation Cache

- Token recomputation always re-executes the entire chain.
- This leads to massive redundant computation.



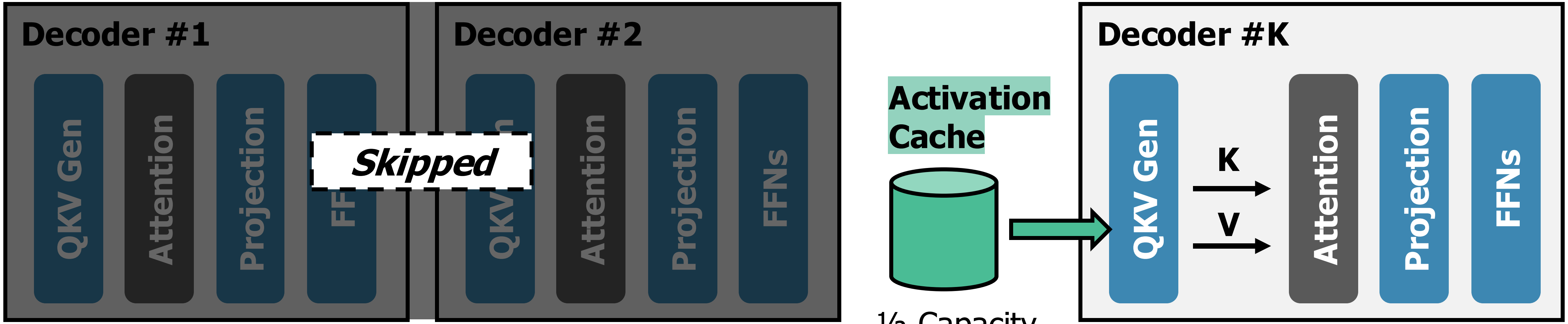
Potential of Activation Cache

- The input activation for Decoder #K is the output of Decoder #K-1



Potential of Activation Cache

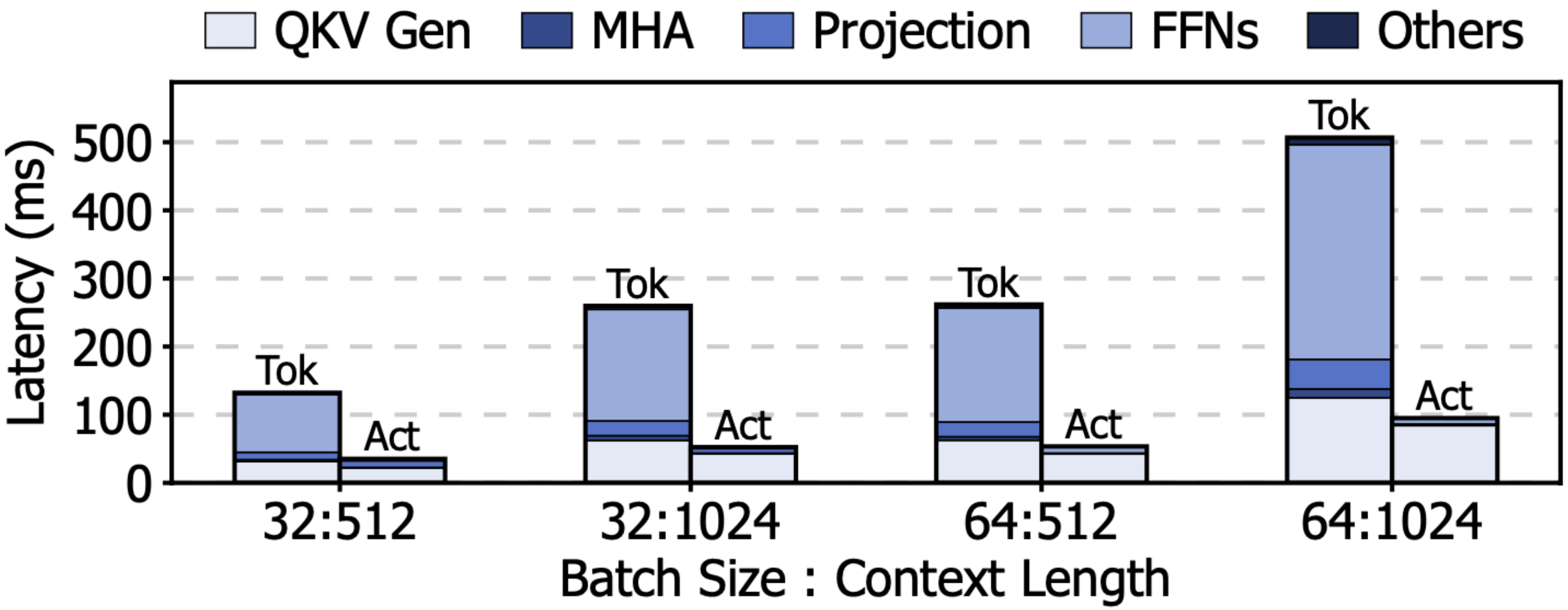
- Activation caching can skip Attention, Projection, FFNs to recompute
- Activation uses only half the memory compared to KV Cache.



½ Capacity
Skip previous decoders

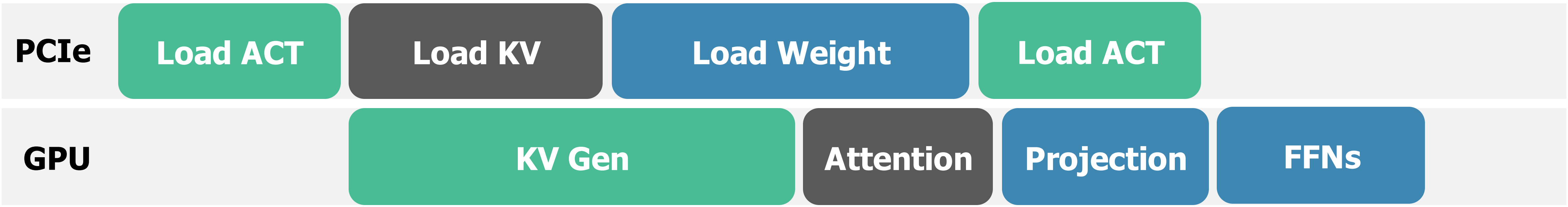
Potential of Activation Cache

- **78% faster** activation recomputation compared to token recomputation.
- Use KV-Activation Hybrid Caching to maximize PCIe and GPU overlap



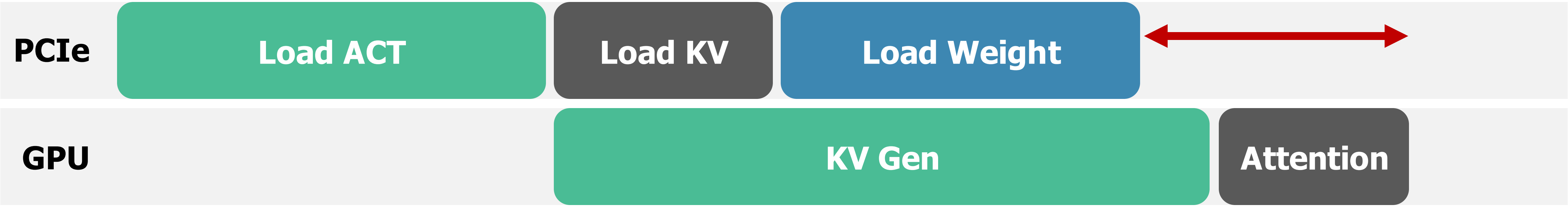
Asynchronous Inference Engine

- Leverage double buffering to hide data transfer latency.
- Asynchronously overlap PCIe transfers, recomputation, and the forward pass.



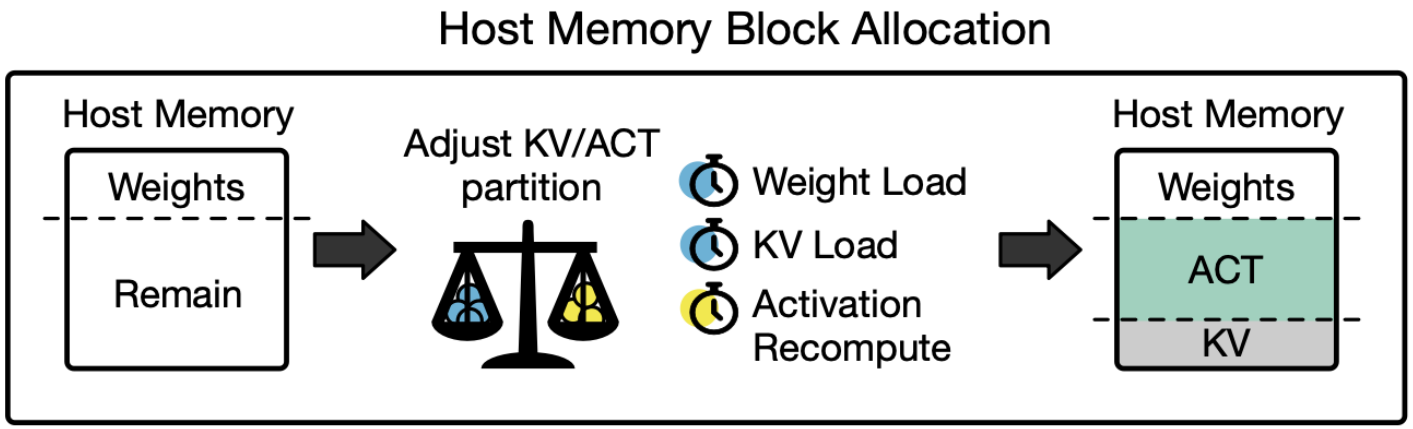
Need for Cache Management Policy

- Excessive ACT Cache idles the PCIe bus.



Cache Management Policy

- Partition the KV/Activation cache to co-optimize PCIe and GPU usage.
- Allocate GPU memory to the ACT cache, and split host memory into KV:ACT



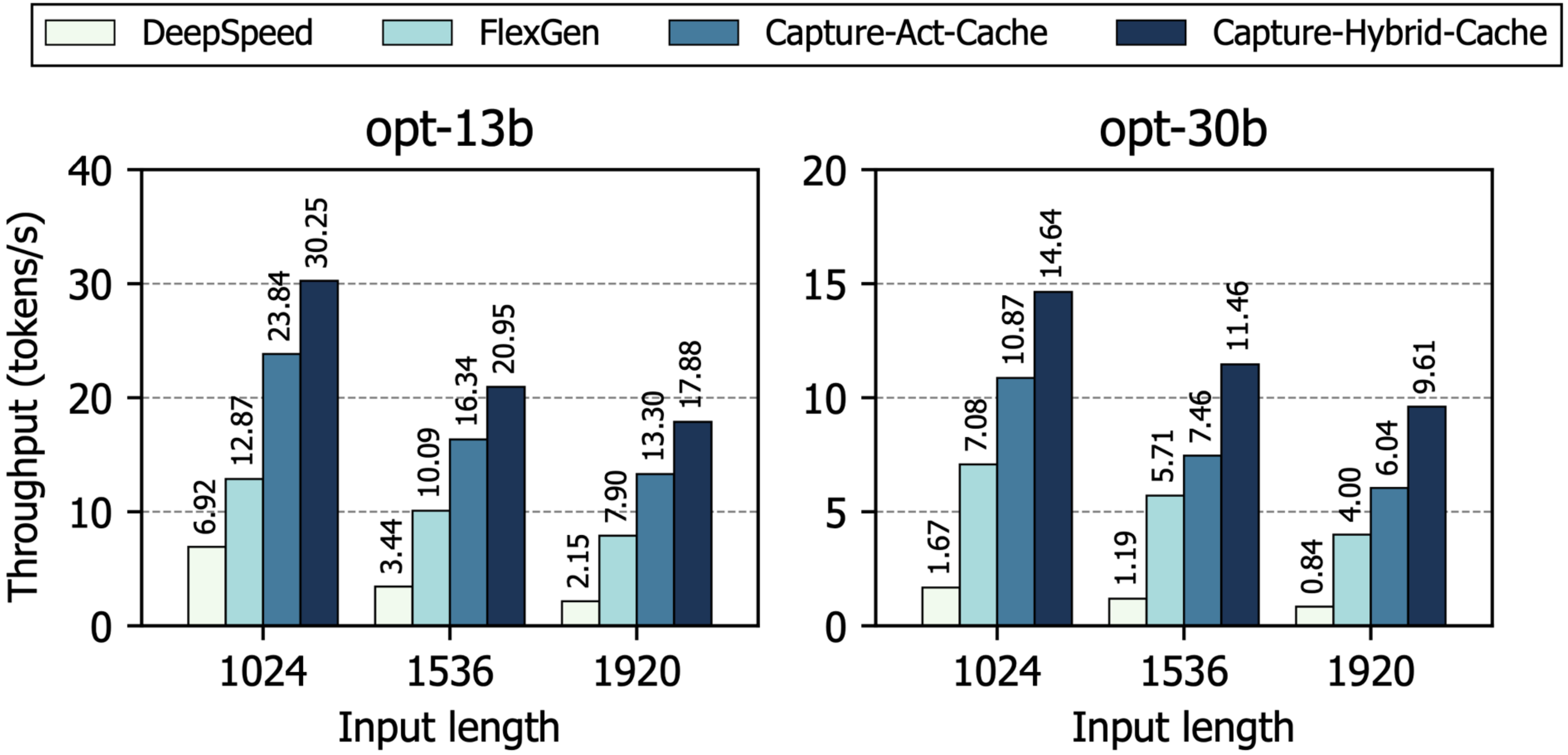
Evaluation Methodology

- **Environment**
  - NVIDIA RTX 4090 GPU, equipped with 24GB of GDDR6X via PCIe 4.0 x16
- **Models**
  - OPT-6.7B, 13B, 30B, 66B
- **Baseline**
  - DeepSpeed Inference [1]
  - FlexGen [2]
  - Activation-only-Cache System

[1] Aminabadi, et al. "Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale." SC, 2022.
[2] Sheng, Ying, et al. "FlexGen: High-throughput generative inference of large language models with a single gpu." ICML, 2023.

Eval: Throughput Improvement

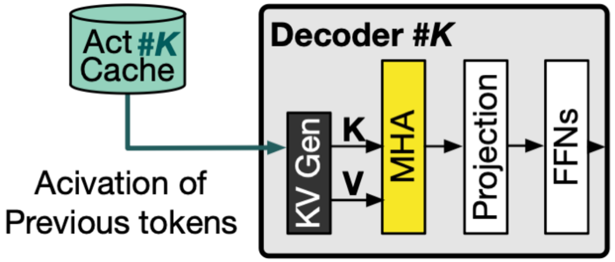
- Hybrid-Cache achieves **2.19x** higher throughput over FlexGen
- **1.35x** higher throughput over Act-Cache system



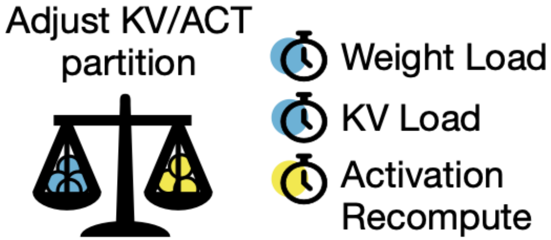
Conclusion

- **Capture**
  - Efficient Host-memory Offloading LLM Inference System
- **Contributions**
  - Solve the KV cache bottleneck caused via KV-Activation Hybrid Caching.
  - Propose a framework for efficient computation-communication overlap.

KV-ACT Hybrid Caching



Efficient PCIe-GPU Overlap



Throughput improvement

2.1×

over FlexGen

1.3×

over Activation-only