SiDA-MoE: Sparsity-inspired Data-Aware Serving for Efficient and Scalable Large Mixture-of-Experts Models

Zhixu Du[1] Shiyu Li[1] Yuhao Wu[1] Xiangyu Jiang[2] Jingwei Sun[1] Qilin Zheng[1]

Yongkai Wu[2], Ang Li[3], Hai "Helen" Li[1] Yiran Chen[1]

[1]Duke University

[2]Clemson University

[3]University of Maryland, College Park

Era of Large Models

Nov. 2022
ChatGPT
GPT-3.5

Mar. 2023
Gemini
GPT-4
LLaMA

Jan. 2024
Mistral
Mixtral

Mar. 2024
Claude 3
Opus
Sonnet
Haiku

Apr. 2024
LLaMA3-8B
LLaMA3-70B
LLaMA3-400B

Duke

Era of Large MoE Models

Nov. 2022
ChatGPT
GPT-3.5

Mar. 2023
Gemini
GPT-4
LLaMA

Jan. 2024
Mistral
Mixtral

Mar. 2024
Claude 3
Haiku
Sonnet
Opus

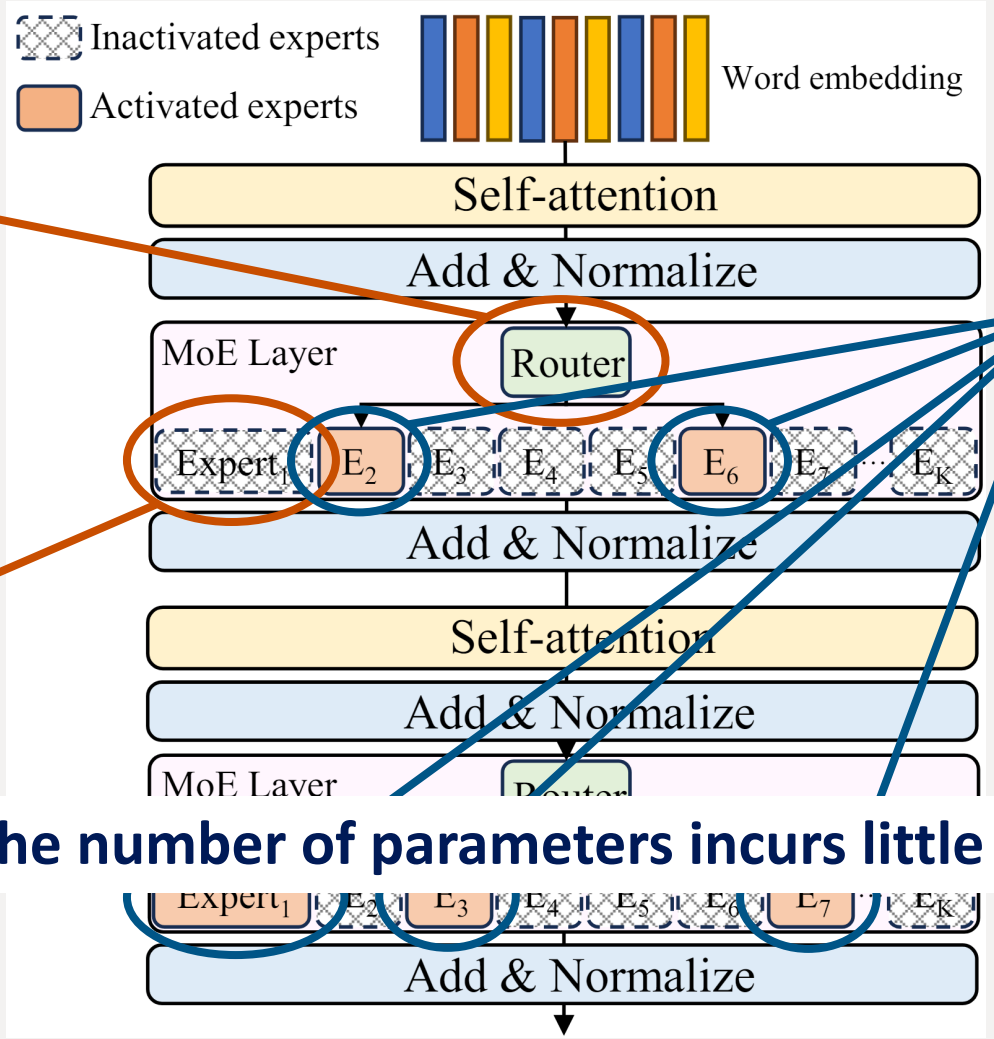
Apr. 2024
LLaMA3-8B
LLaMA3-70B
LLaMA3-400B

Many large models employs the Mixture-of-Experts (MoE) techniques.

Era of Large MoE Models

Router:
Usually, a linear classifier to decide which expert will be selected.

Expert:
Usually, a MLP model.



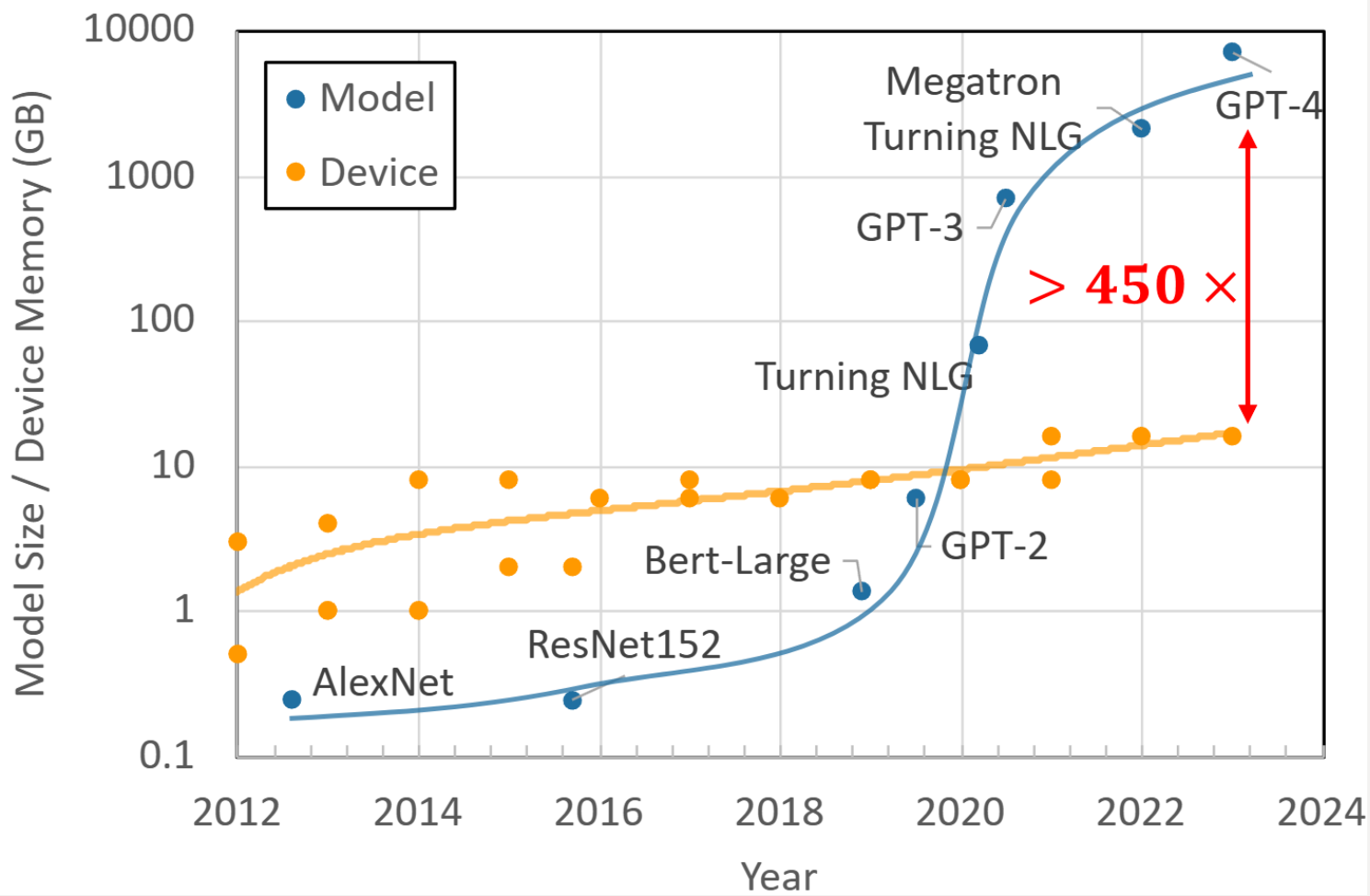
Activated experts:
Only selected experts will participate in the current round of inference, combined by a weighted sum.

$$\sum_{i \in \mathbb{I}} \alpha_i(\mathbf{x}) f_i(\mathbf{x}; \boldsymbol{\theta}_i)$$

Drastically increase the number of parameters incurs little computational overhead.

Duke

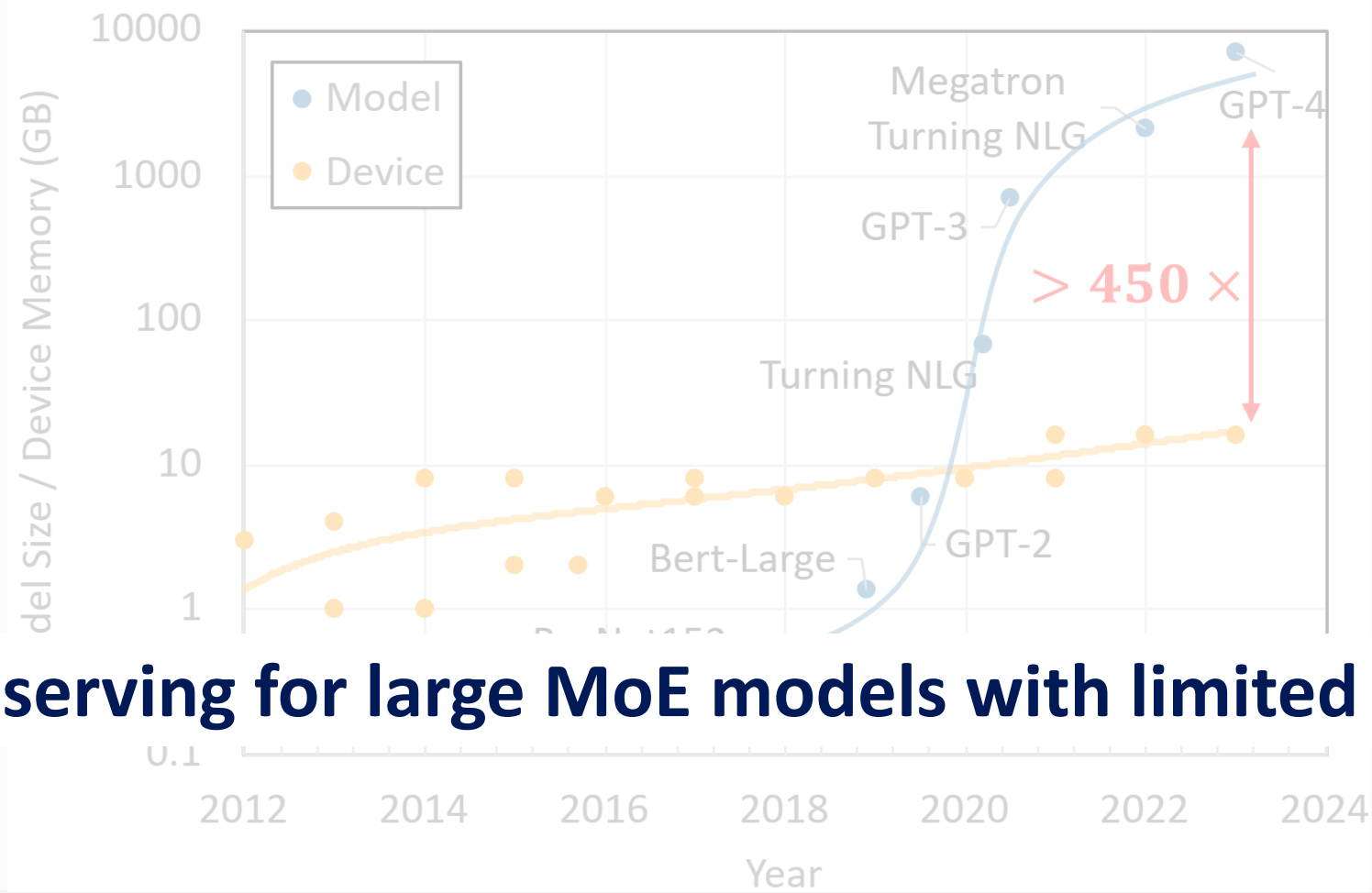
Big Models and Small Devices



Devices: Geforce GTX Mobile Series, Apple A Series, and Qualcomm Snapdragon.

Duke

Big Models and Small Devices



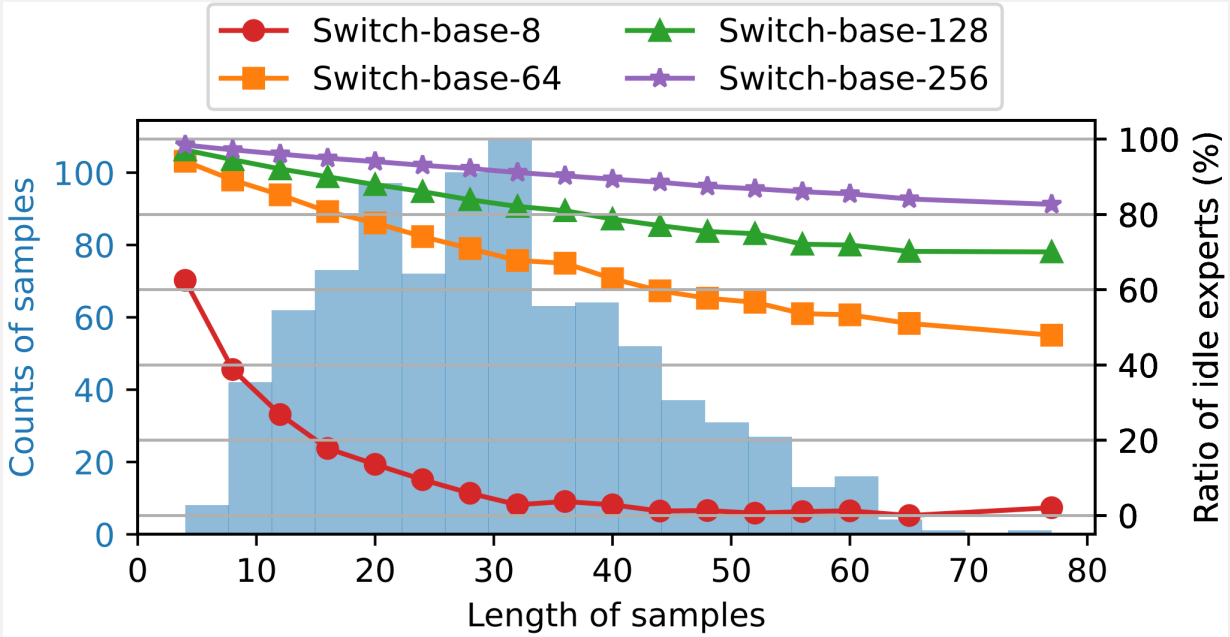
Efficient serving for large MoE models with limited resources.

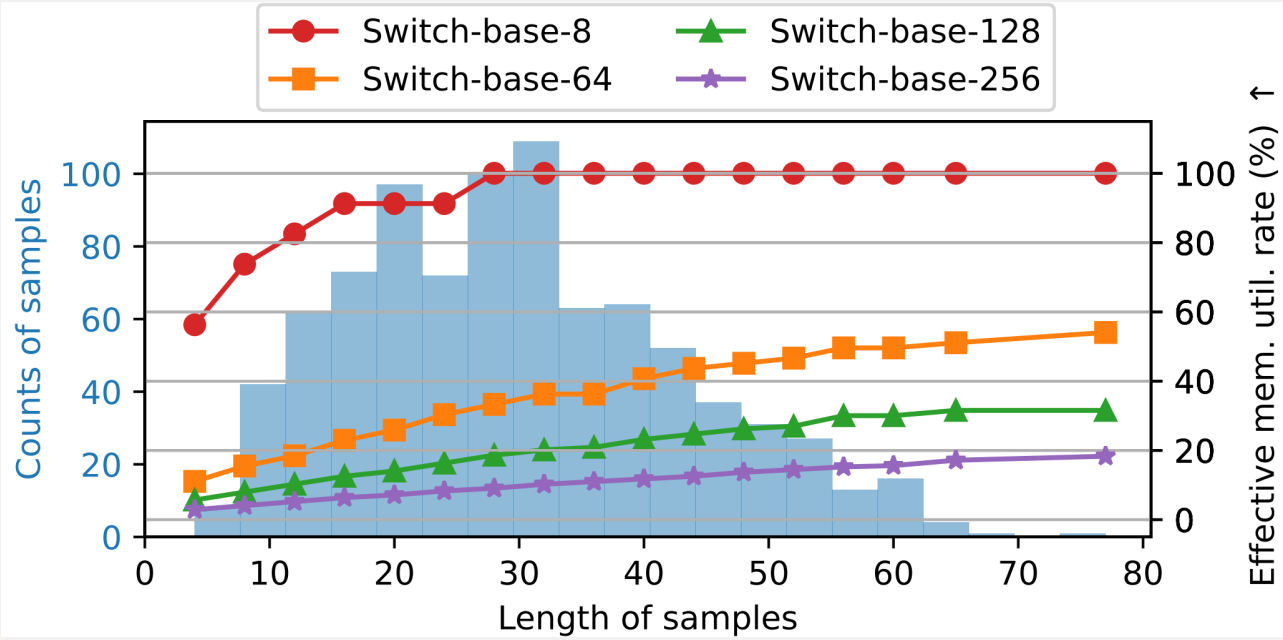
Duke

Devices: Geforce GTX Mobile Series, Apple A Series, and Qualcomm Snapdragon.

Motivations

Sparsely-activated Experts





- Teaser on SST2.
- Up to 80% experts are idle on Switch-base-256.
- Up to 80% GPU memory are ineffective for the inference on Swicth-base-256.

Duke

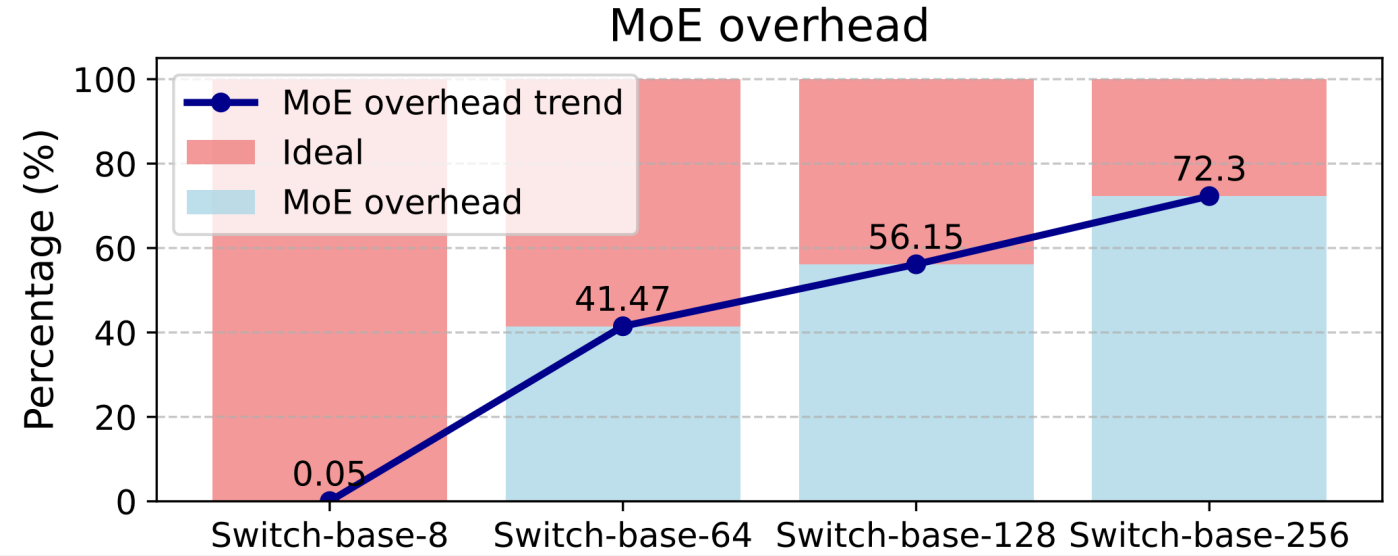
Sparsely-activated Experts

- Teaser on SS
- Up to 80% ex
- Up to 80% G wicth-base-256.

Source: https://zhuanlan.zhihu.com/p/674751021

Duke

MoE Overhead on Resource Limited Devices

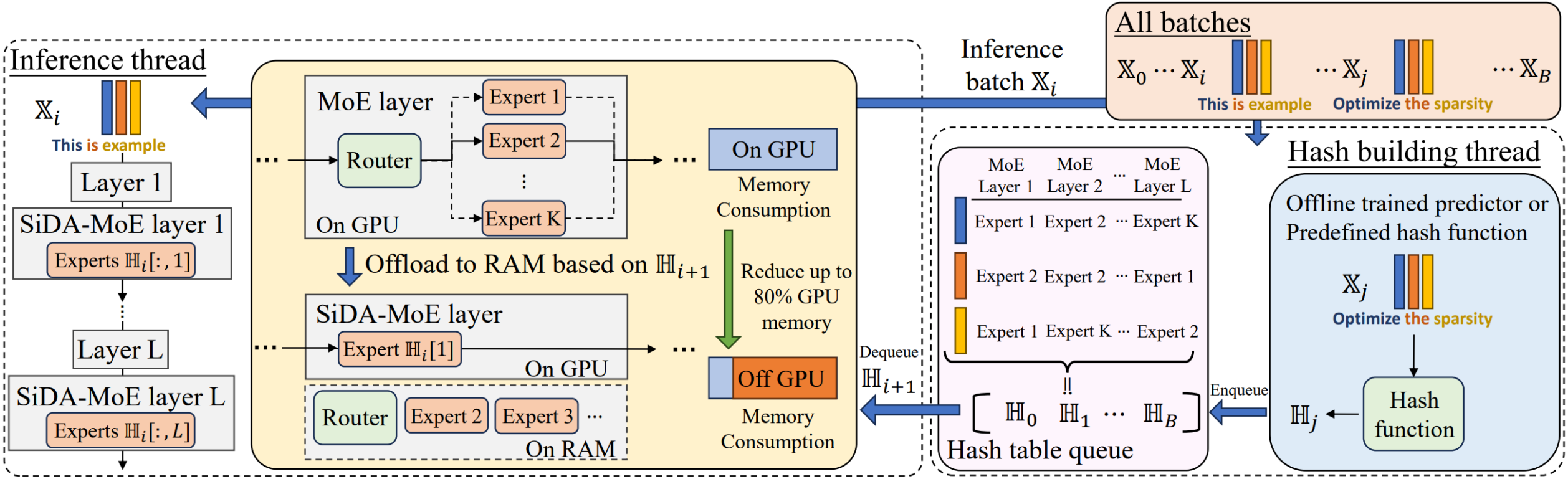


- Teaser on SST2
- MoE Overhead: expert selection, expert invocation, additional communication costs.
- In resource constrained scenarios, the invocation overhead surpasses the computation, meaning the number of experts called dominates the overall inference time.

Duke

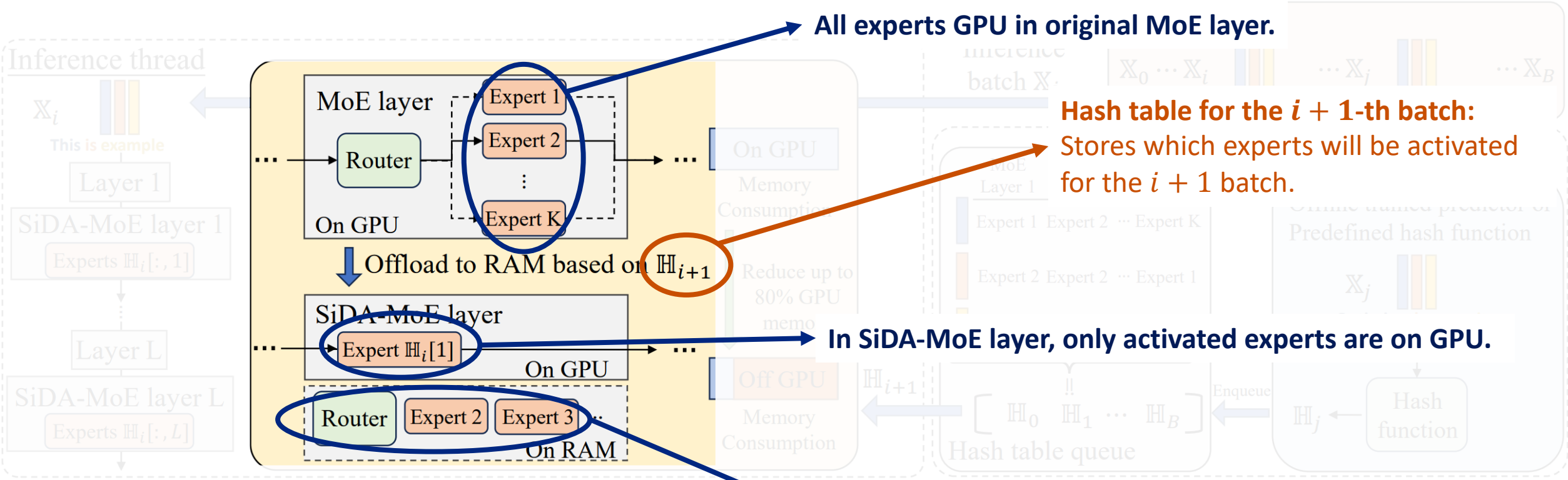
SiDA-MoE

SiDA: Sparsity-inspired Data-Aware

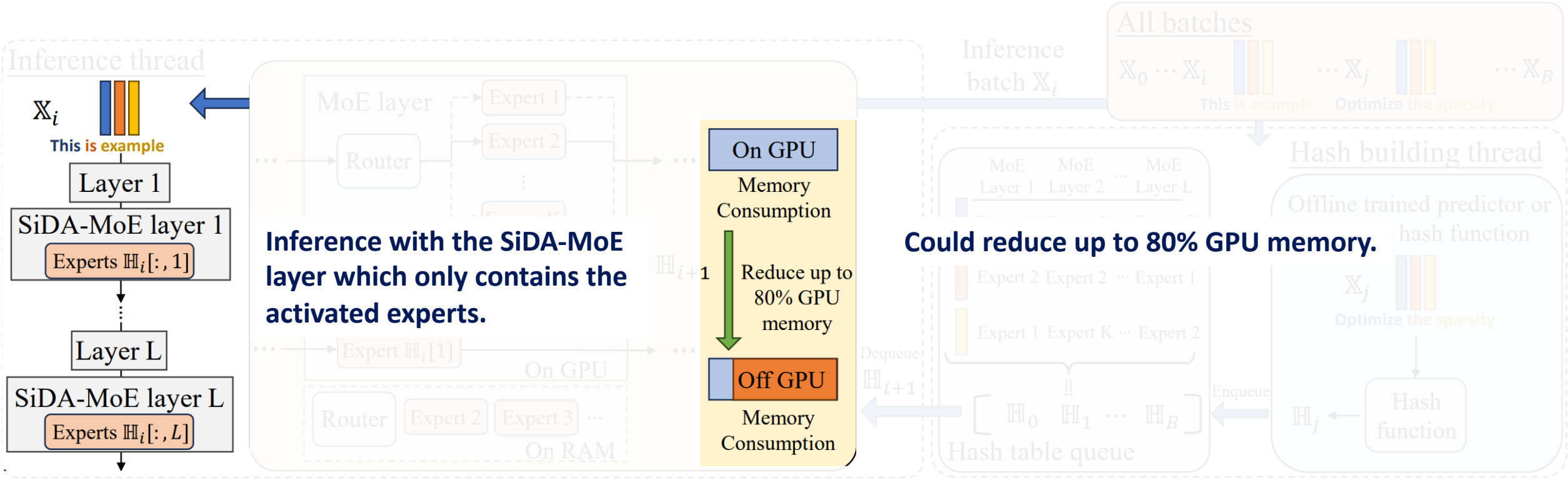


Duke

SiDA: Sparsity-inspired Data-Aware



SiDA: Sparsity-inspired Data-Aware



Inference with the SiDA-MoE layer which only contains the activated experts.

Could reduce up to 80% GPU memory.

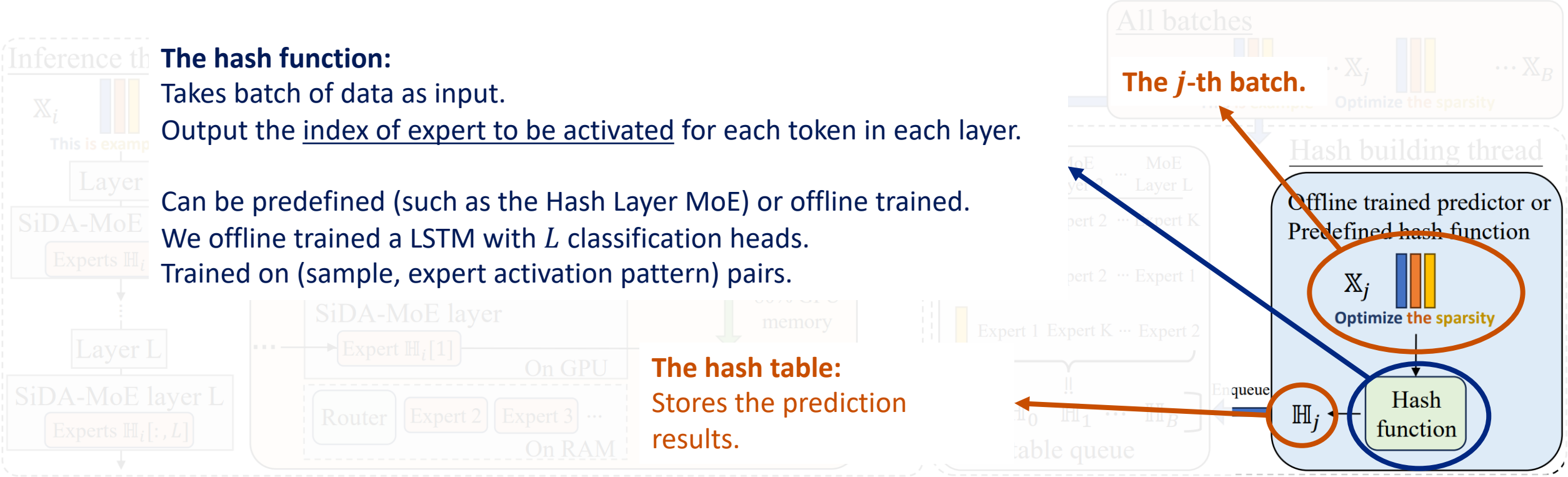
Duke

SiDA: Sparsity-inspired Data-Aware

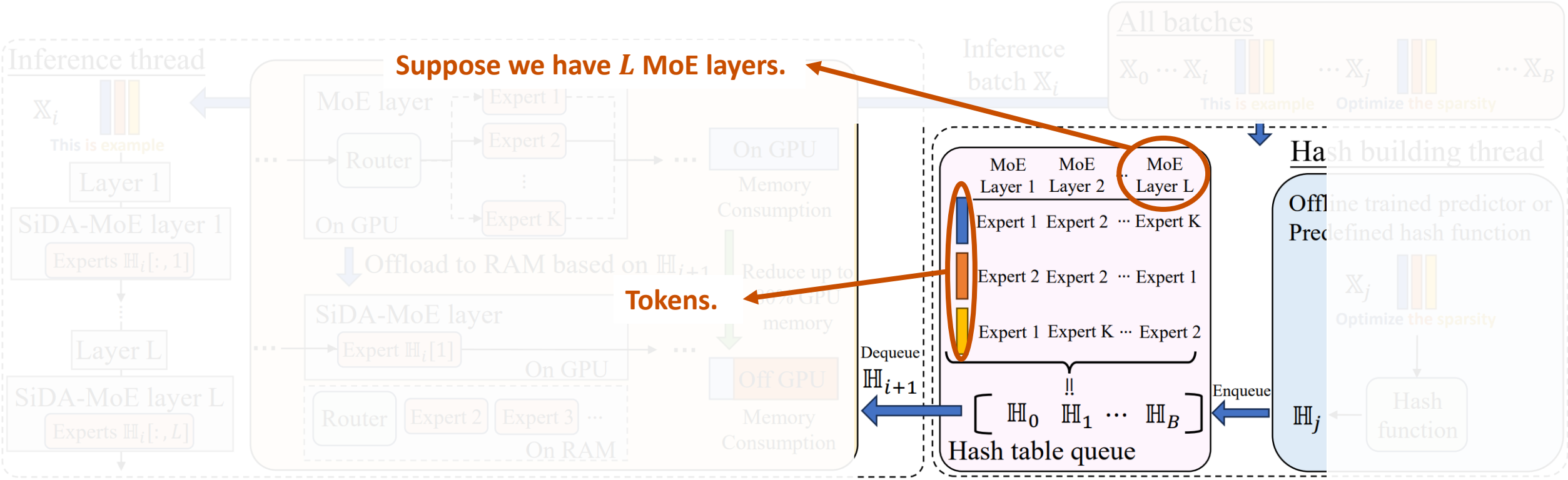
The hash function:
Takes batch of data as input.
Output the index of expert to be activated for each token in each layer.

Can be predefined (such as the Hash Layer MoE) or offline trained.
We offline trained a LSTM with L classification heads.
Trained on (sample, expert activation pattern) pairs.

The hash table:
Stores the prediction results.

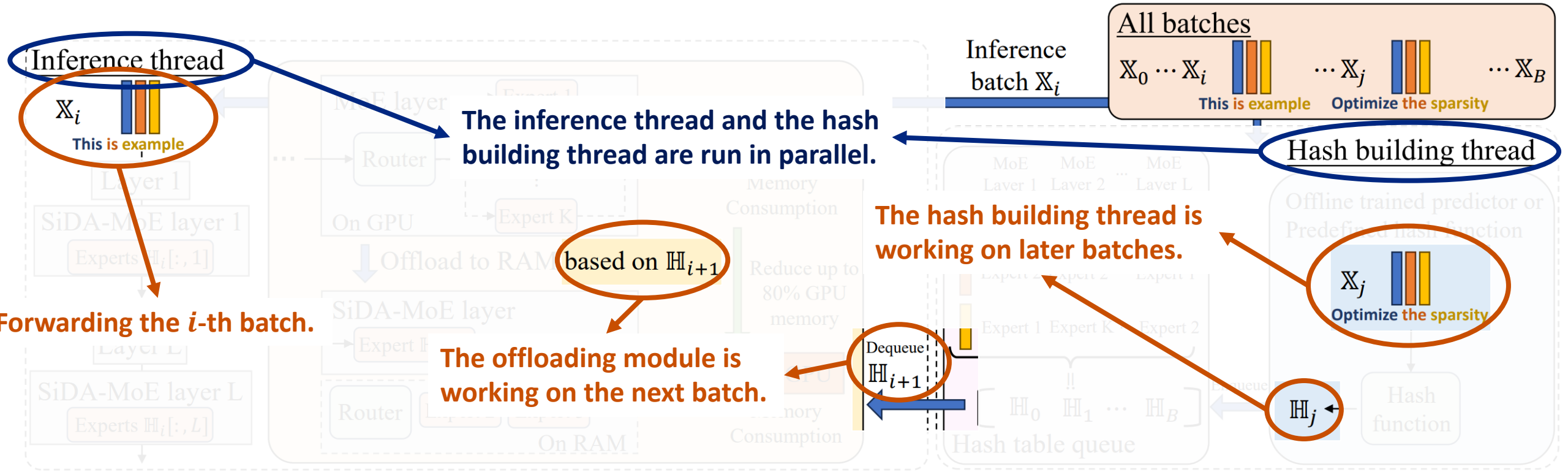


SiDA: Sparsity-inspired Data-Aware



Duke

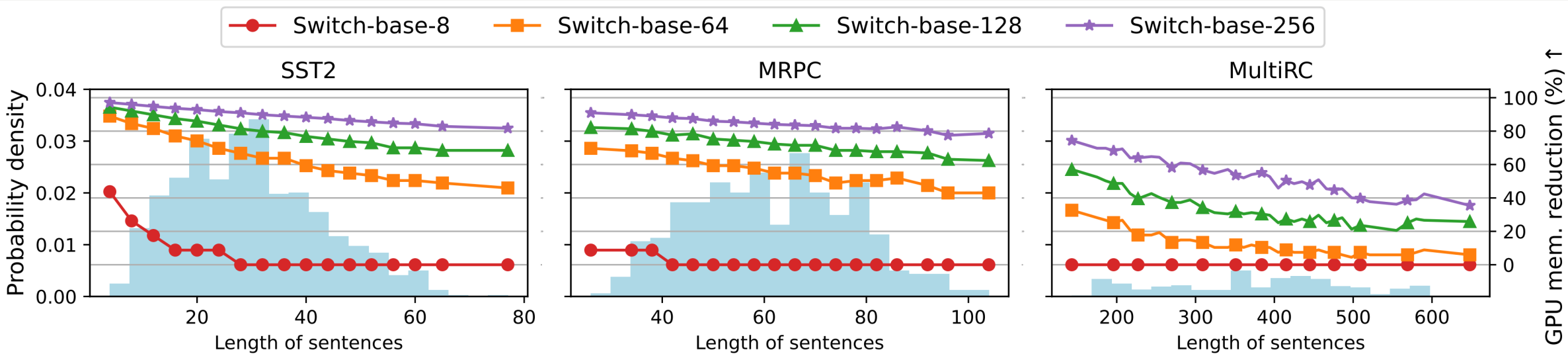
Our System



Duke

Experiments

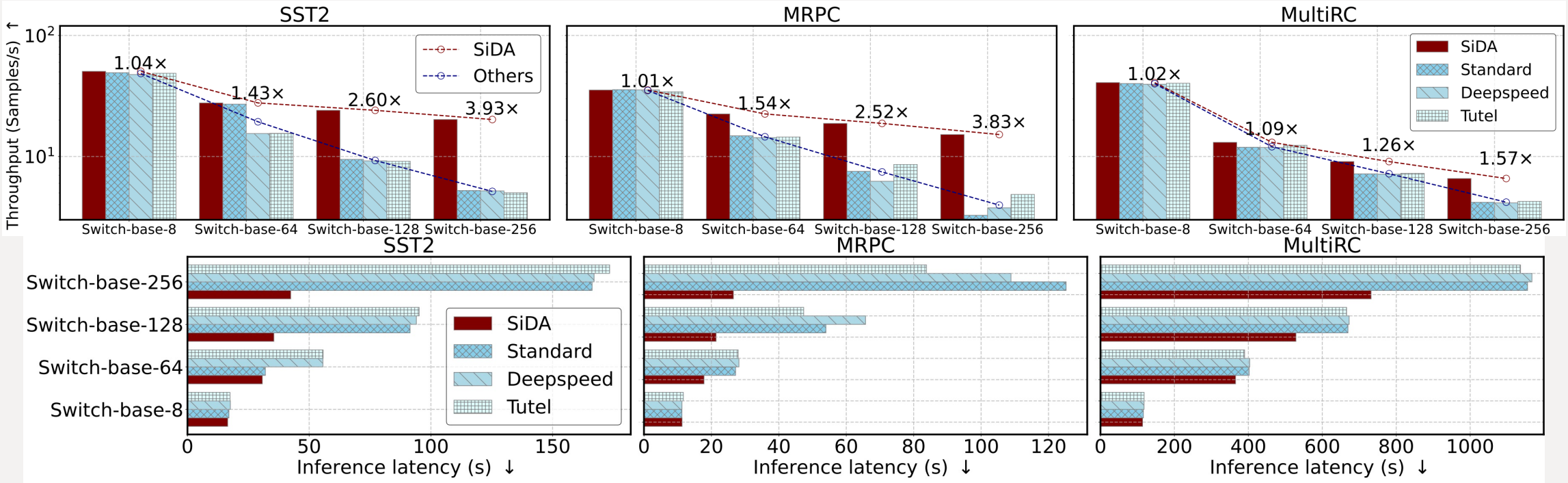
Memory



- SiDA saves up to 80% memory for Swith-base-256 on short sentences.
- SiDA saves over 40% memory for Swith-base-256 on Long sentences.

Duke

Throughput and Latency

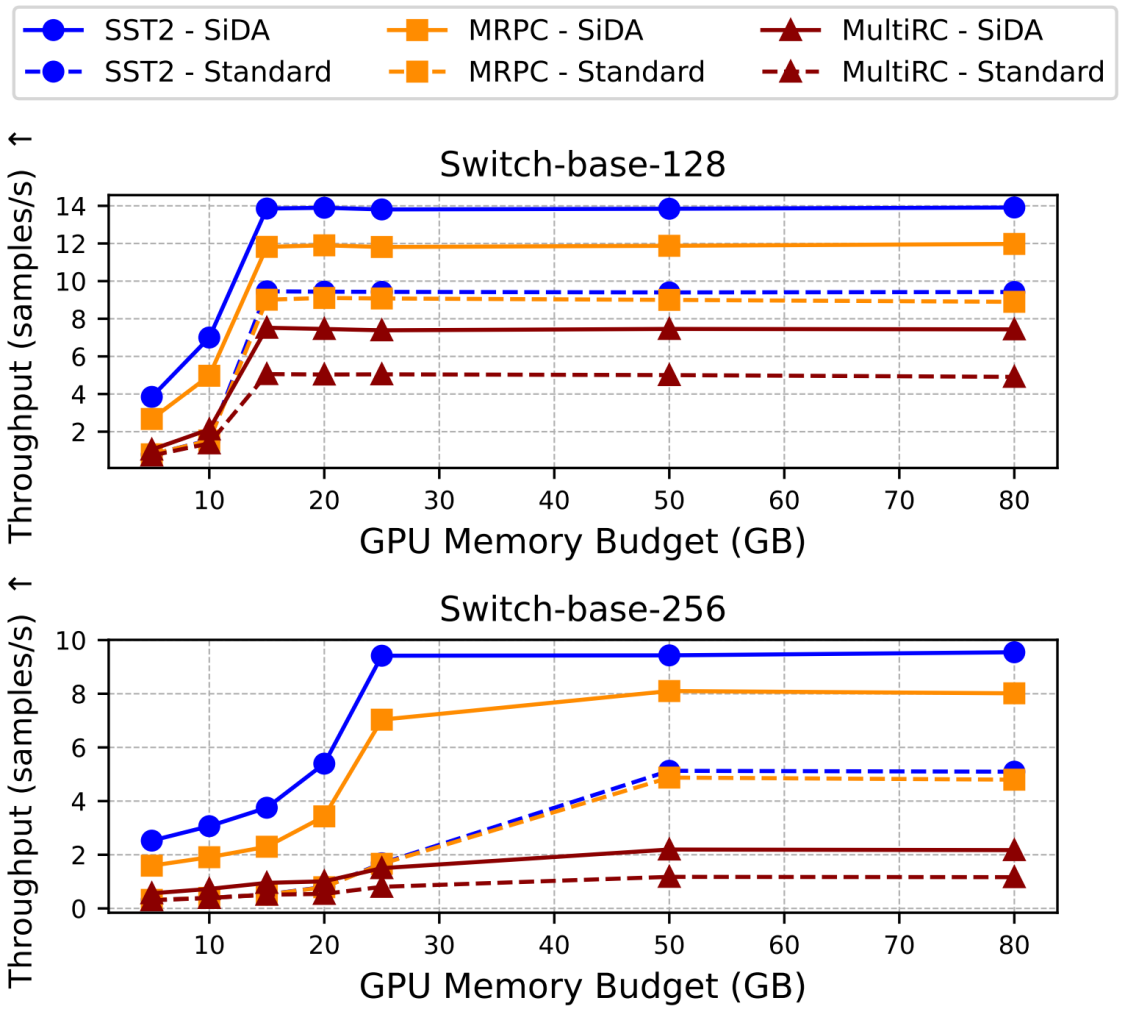


- SiDA achieves up to 3.93x speed up on throughput for short sentences.
- SiDA achieves at least 1.57x speed up on throughput for short sentences.

Duke

Memory-limited Scenarios

SiDA consistently outperforms baselines given different GPU memory budgets.



Duke

Accuracy Preservation

Backbone	SST2	MRPC	MultiRC
Switch-base-8	99.00%	97.41%	91.74%
Switch-base-128	98.78%	98.65%	90.49%

Top-3 hash hits rate (prediction accuracy of the hash function).

Backbone	Pretrained ppl. (↓)	SiDA-MoE ppl. (↓)
Switch-base-8	6.68	18.49
Switch-base-64	4.93	11.84
Switch-base-128	4.86	11.73
Switch-base-256	4.59	8.11

Accuracy compromise as a pretrained model.

Backbone		SST2	MRPC	MultiRC
Switch-base-8	Finetuned	92.20	89.14	56.70
	SiDA-MoE	90.59	86.91	56.11
	Fidelity	98.25%	97.49%	98.95%
Switch-base-128	Finetuned	93.57	89.66	59.95
	SiDA-MoE	87.04	83.01	55.49
	Fidelity	93.02%	92.59%	92.56%

Accuracy compromise on downstream tasks.

Duke

Summary

- The Mixture-of-Experts technique is getting popular in the era of large models.
- We propose SiDA-MoE, a sparsity-inspired data-aware serving for MoE models.
- SiDA-MoE achieves huge amount of GPU memory save and significant speed up, while compromising little accuracy.

Thanks for your attentions!

If you have further questions, please feel free to contact me:
zhixu.du@duke.edu

Duke