On-the-Fly Adaptive Distillation of Transformer to Dual-State Linear Attention for Long-context LLM Serving

Tuesday, 15 July 11AM-2PM

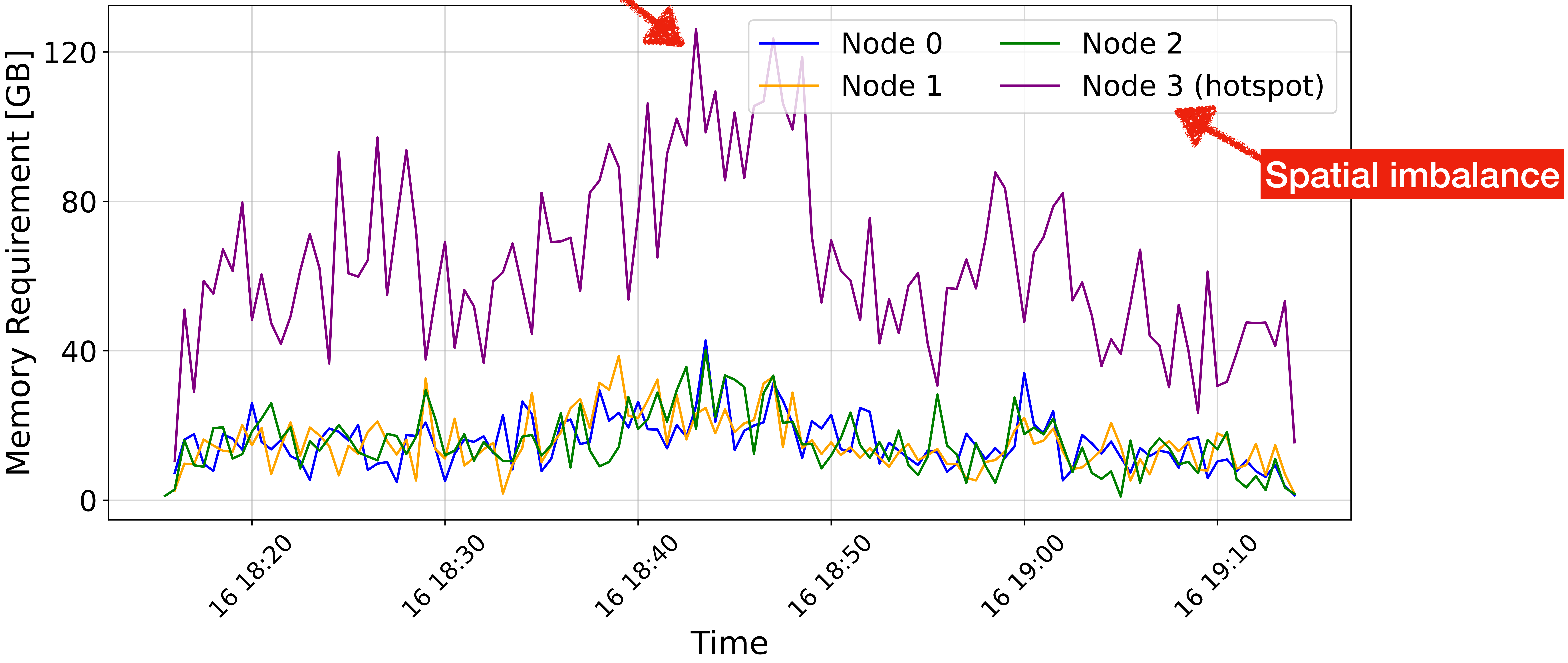
Poster Session 1

Yeonju Ro[1], Zhenyu Zhang[1], Souvik Kundu[2], Zhangyang Wang[1], Aditya Akella[1]

[1]University of Texas at Austin, USA
[2]Intel Labs, USA

Dynamic Resource Requirement

Burst Traffic and Prolonged Sessions Extend Context Length



Temporal variability

Spatial imbalance

Alternative Language Model Architecture

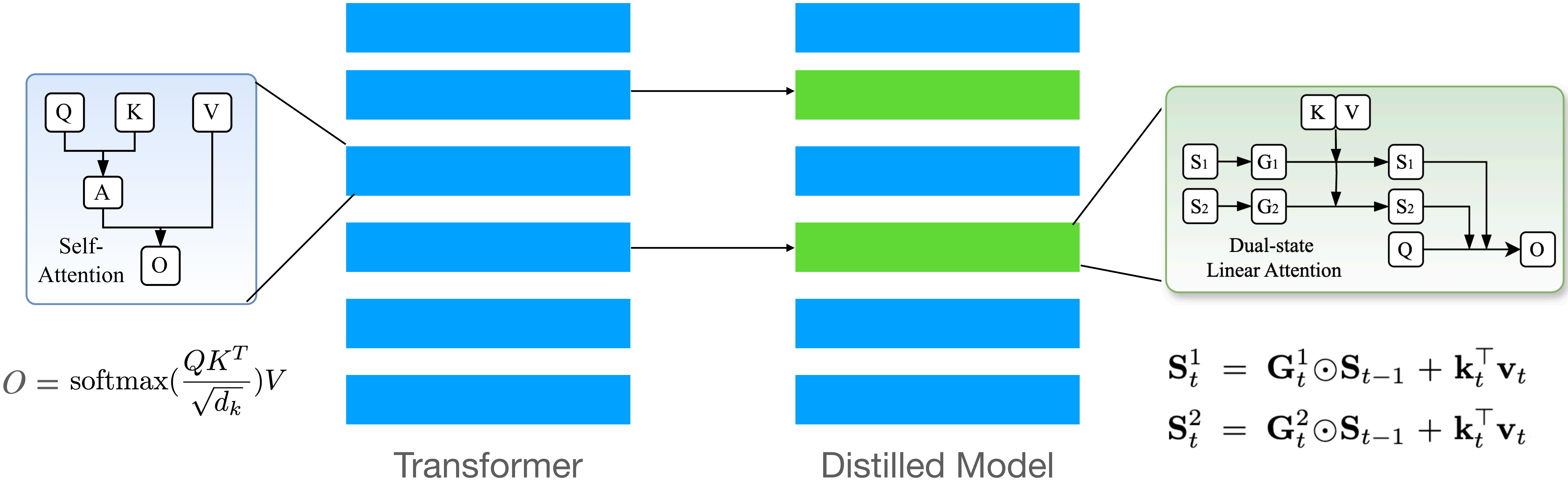
Trade-off between Self-Attention and Linear-Attention

	Computational Cost	Memory Cost (For KV cache)	Model Performance (Accuracy)
Self Attention	$O(n^2)$	$O(n)$	Great 👍
Linear Attention	$O(n)$	$O(c)$	Not Good
Mamba or Gated Linear Attention	$O(n)$	$O(c)$	Better than Linear Attn, Worse than Self Attn.

* n: context length

Leveraging Tradeoff for Adaptive Inference System

Adaptively Distilling Self Attention to Dual-State Linear Attention (DSLA)



$$O = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

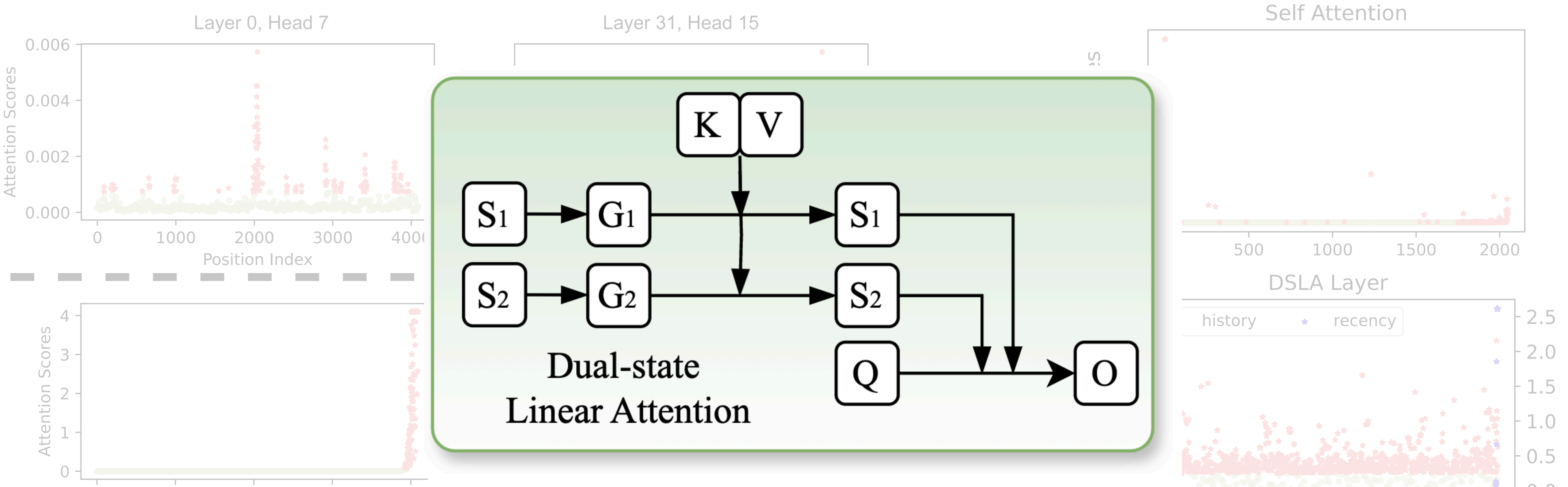
$$\mathbf{S}_t^1 \;=\; \mathbf{G}_t^1 \odot \mathbf{S}_{t-1} + \mathbf{k}_t^\top \mathbf{v}_t$$

$$\mathbf{S}_t^2 \;=\; \mathbf{G}_t^2 \odot \mathbf{S}_{t-1} + \mathbf{k}_t^\top \mathbf{v}_t$$

$$\mathbf{o}_t \;=\; \mathbf{q}_t \left(\gamma \cdot \mathbf{S}_t^1 \;+\; (1-\gamma)\, \mathbf{S}_t^2\right)$$

Dual-state Linear Attention (DSLA)

Resolve GLA's Recency Bias with Additional State

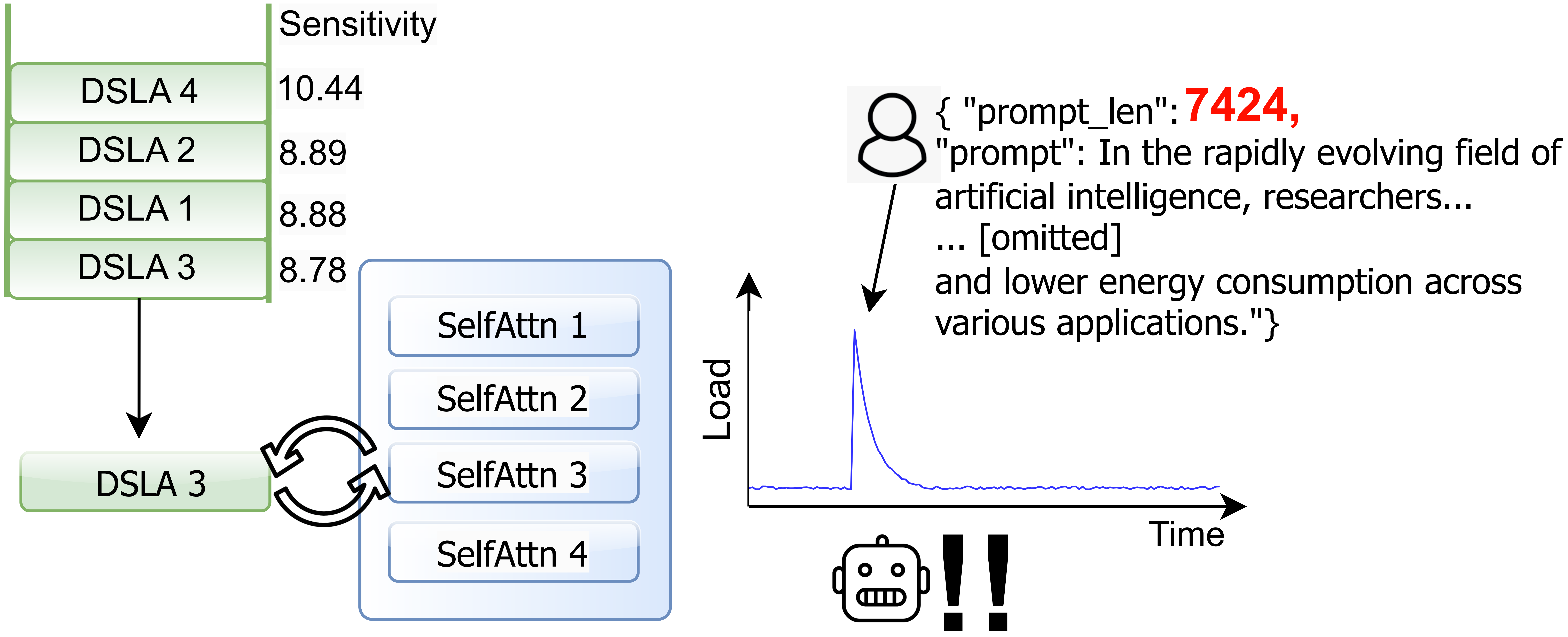


$$\mathcal{L}_{dist} = \mathcal{D}_{KL}\left(\mathbf{o}_t \,\|\, \mathbf{o}_t^{(\mathrm{gt})}\right), \quad \mathcal{L}_{cont} = \mathrm{sim}\left(\mathbf{G}_t^1, \mathbf{G}_t^2\right)$$

$$\mathcal{L} = \mathcal{L}_{dist} + \lambda \mathcal{L}_{cont};$$

Adaptive Inference with DSLA-*Serve*

Progressively Converting Self Attention to DSLA



{ "prompt_len":**7424,**
"prompt": In the rapidly evolving field of
artificial intelligence, researchers...
... [omitted]
and lower energy consumption across
various applications."}

Performance on Long Context Benchmark

LongBench

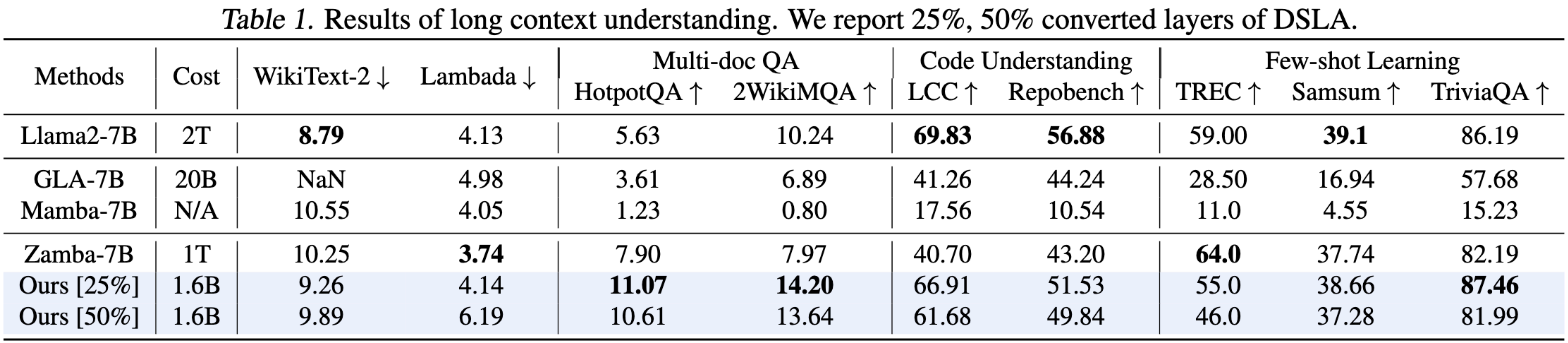


Table 1. Results of long context understanding. We report 25%, 50% converted layers of DSLA.

Methods	Cost	WikiText-2 ↓	Lambada ↓	Multi-doc QA		Code Understanding		Few-shot Learning		
				HotpotQA ↑	2WikiMQA ↑	LCC ↑	Repobench ↑	TREC ↑	Samsum ↑	TriviaQA ↑
Llama2-7B	2T	**8.79**	4.13	5.63	10.24	**69.83**	**56.88**	59.00	**39.1**	86.19
GLA-7B	20B	NaN	4.98	3.61	6.89	41.26	44.24	28.50	16.94	57.68
Mamba-7B	N/A	10.55	4.05	1.23	0.80	17.56	10.54	11.0	4.55	15.23
Zamba-7B	1T	10.25	**3.74**	7.90	7.97	40.70	43.20	**64.0**	37.74	82.19
Ours [25%]	1.6B	9.26	4.14	**11.07**	**14.20**	66.91	51.53	55.0	38.66	**87.46**
Ours [50%]	1.6B	9.89	6.19	10.61	13.64	61.68	49.84	46.0	37.28	81.99

- **Improvement over Baseline** (Llama2-7B) model that it is distilled from:
  - Multi-doc QA (Hotpot QA: 5.63 → 11.07, 2WikiMQA: 10.24 → 14.20)
  - Few-shot Learning (TREC: 59.00 → 64.0, TriviaQA: 86.19 → 87.46)
- **Outperform Hybrid Model** in Multi-doc QA, Code-understanding, Samsum/TriviaQA
- **Outperform GLA/Mamba** in all tasks

End-to-end Improvement

Replaying Augmented Azure Inference Trace

Prompt Length	Distribution	Max Conv Rate
seq_len < 2k	64.68%	12.5%
2k ≤ seq_len < 4k	16.16%	25%
4k ≤ seq_len < 8k	16.03%	37.5%
8k ≤ seq_len	3.1%	50%
Latency (Before)	93.64 ms	
Latency (After)	40.83 ms	

2.29x Improvement

Tuesday, 15 July 11AM-2PM
Poster Session 1