Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention

Jingyang Yuan[1,2], Huazuo Gao[2], Damai Dai[2], Junyu Luo[1], Liang Zhao[2], Zhengyan Zhang[2], Zhenda Xie[2], Y. X. Wei[2], Lean Wang[2], Zhiping Xiao[3], Yuqing Wang[2], Chong Ruan[2], Ming Zhang[1], Wenfeng Liang[2], Wangding Zeng[2]

Full Paper Personal Page

[1]Peking Unviersity [2]DeepSeek-AI [3]University of Washington

Correspondence to: Ming Zhang<mzhang_cs@pku.edu.cn>, Wenfeng Liang <wenfeng.liang@deepseek.com> and Wangding Zeng <zengwangding@deepseek.com>

The Challenges of Long-Context LLMs

Softmax Attention Faces…

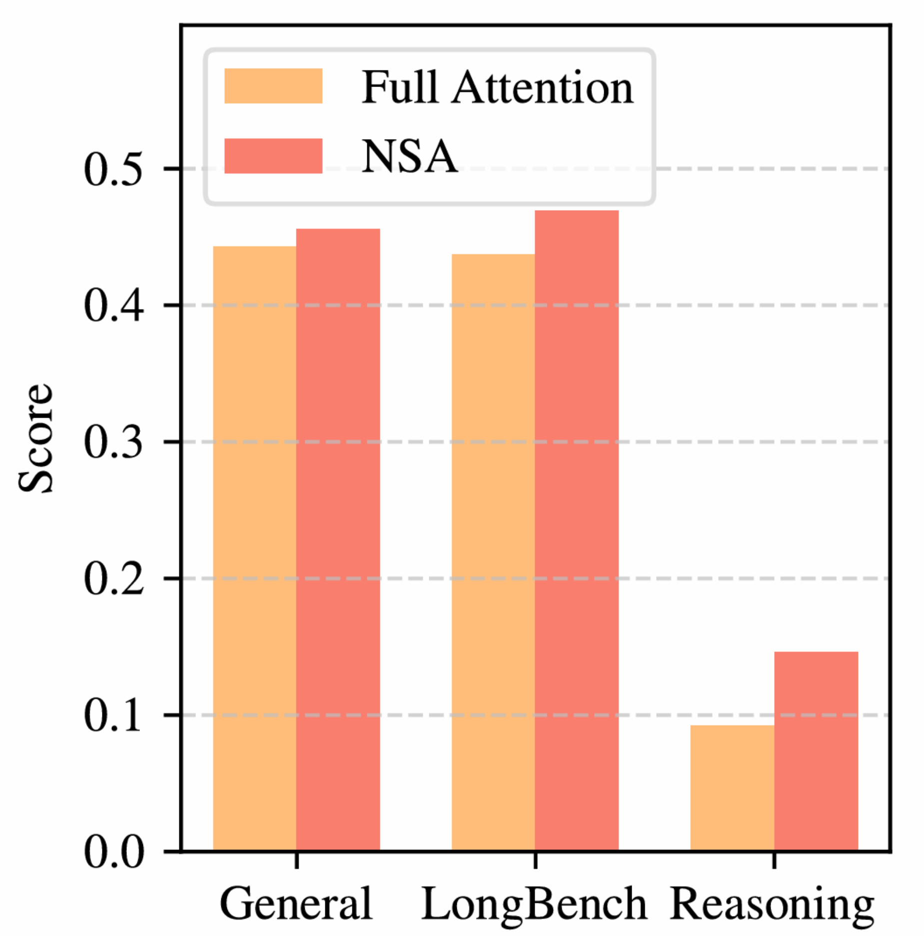
- High Computational Cost
- Latency Bottleneck

Existing Sparse Attention Faces …

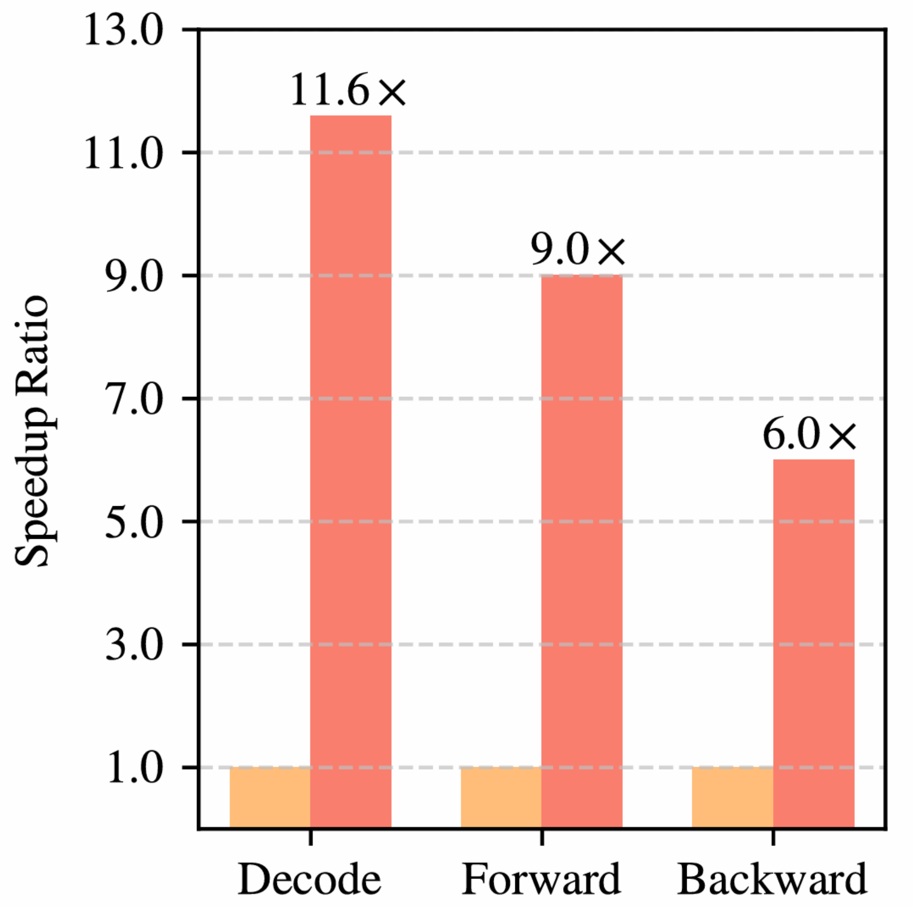
- Unable to Speedup Training
- Illusion of Inference Efficiency

The NSA Solution: A Natively Trainable Sparse Attention

High Performance



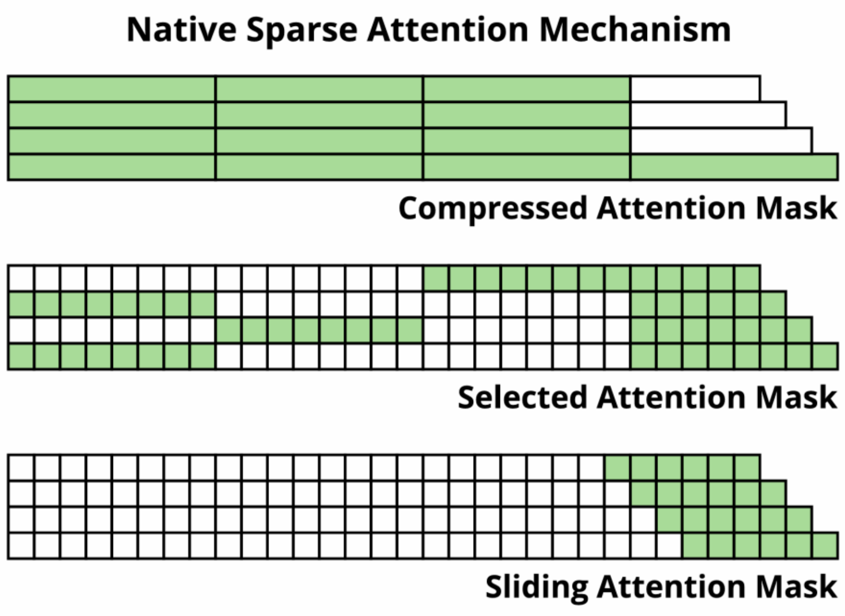
High Speed



Native Sparse Attention: Trainability & High Efficiency

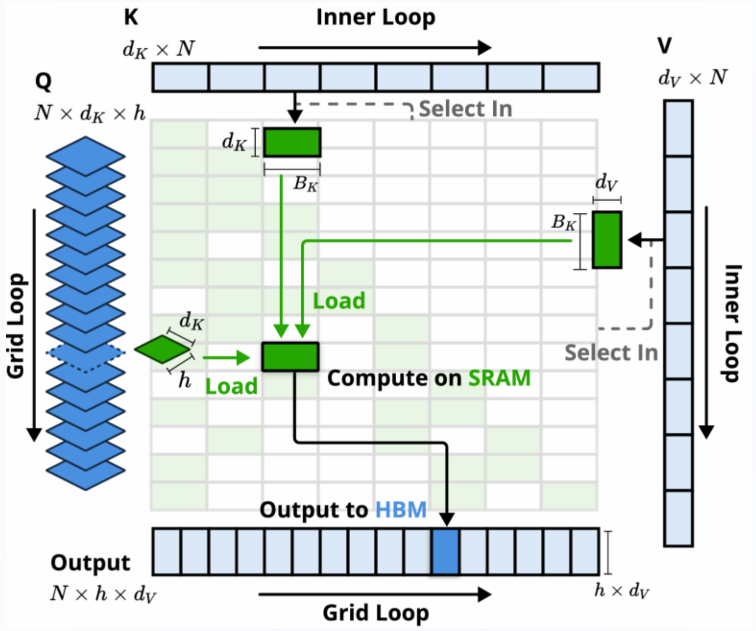
NSA =

Trainable, Sparse Architecture

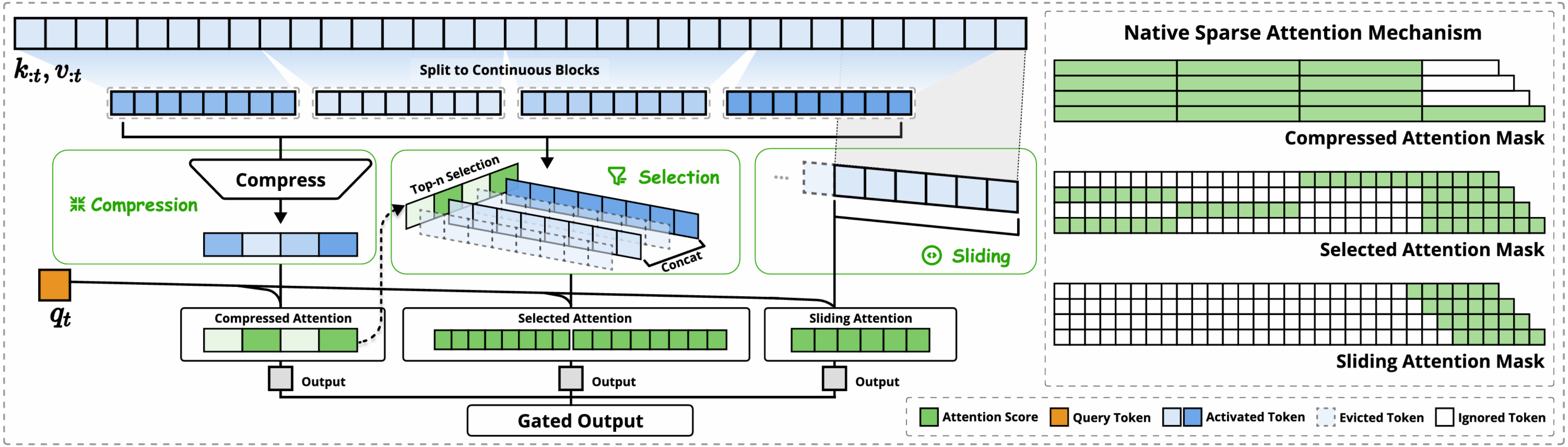


+

Hardware-aligned System Design

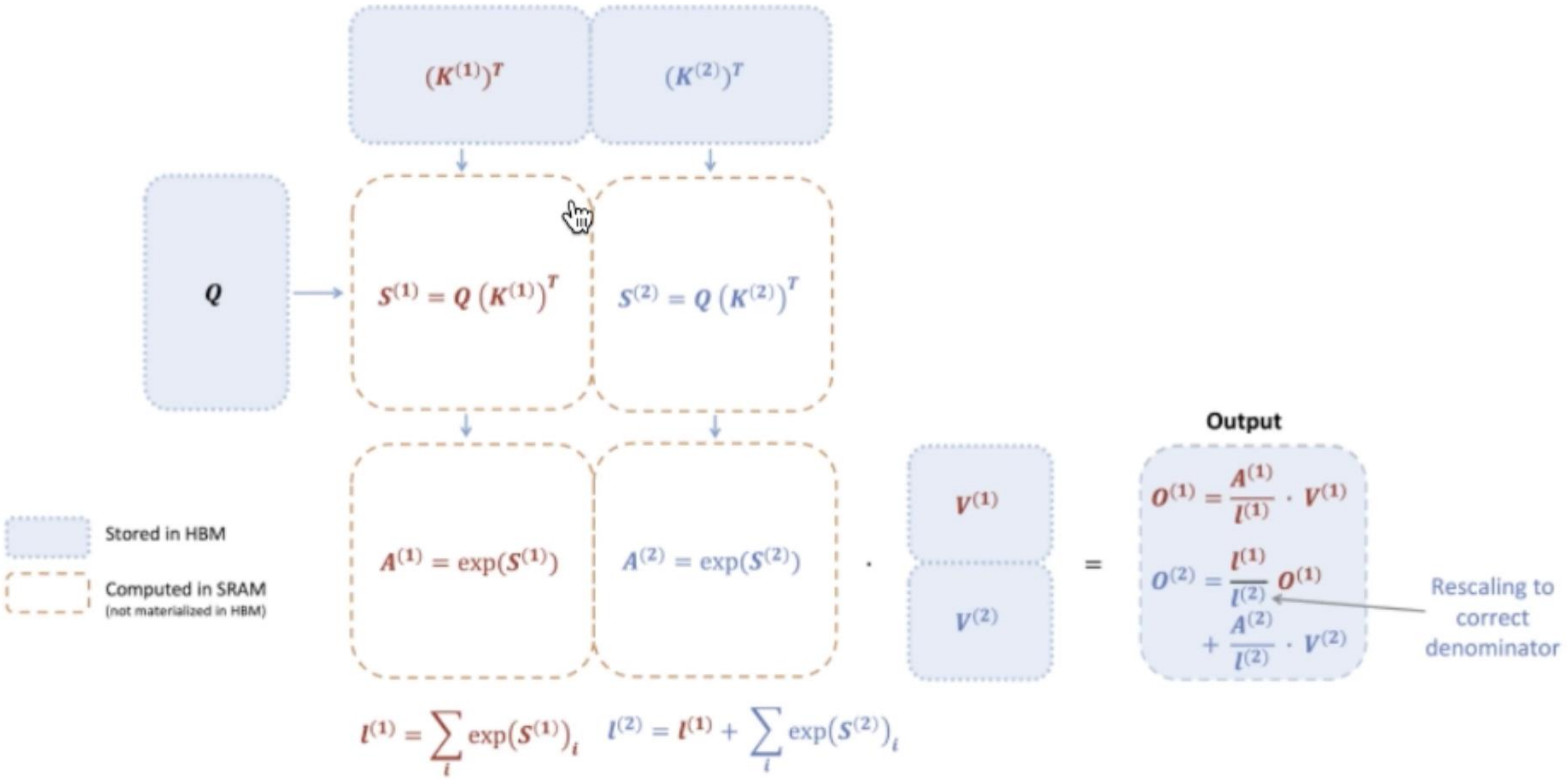


Key Innovation: Natively Trainable Design



NSA Architecture: Enable End-to-end Training

FlashAttention



Group Query Attention

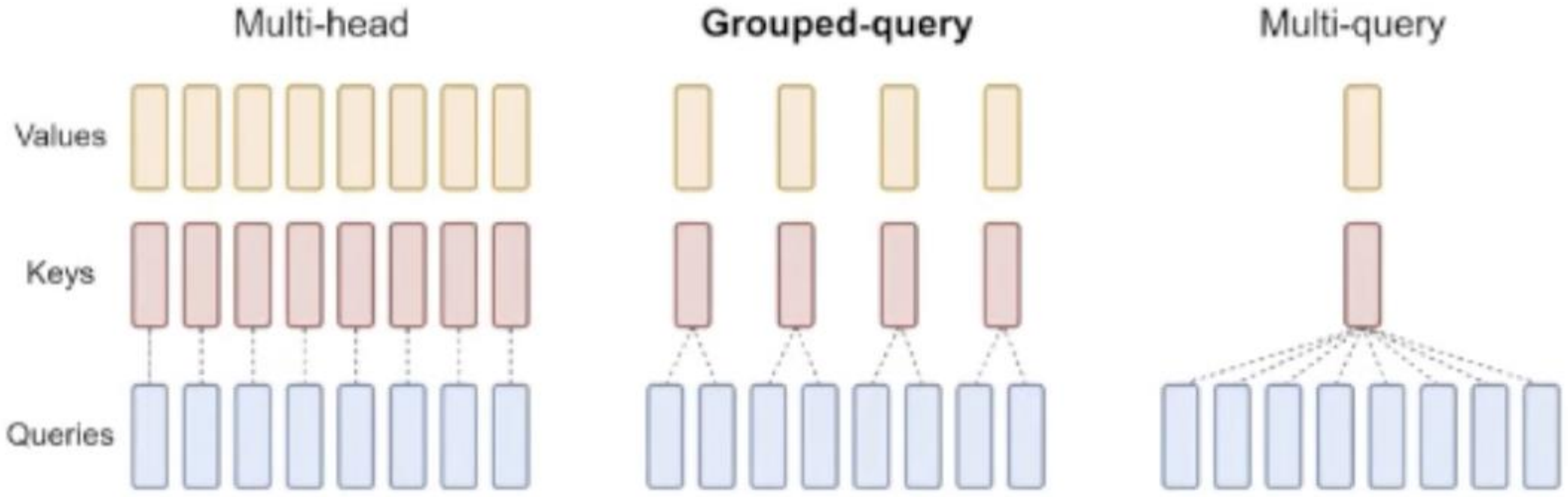
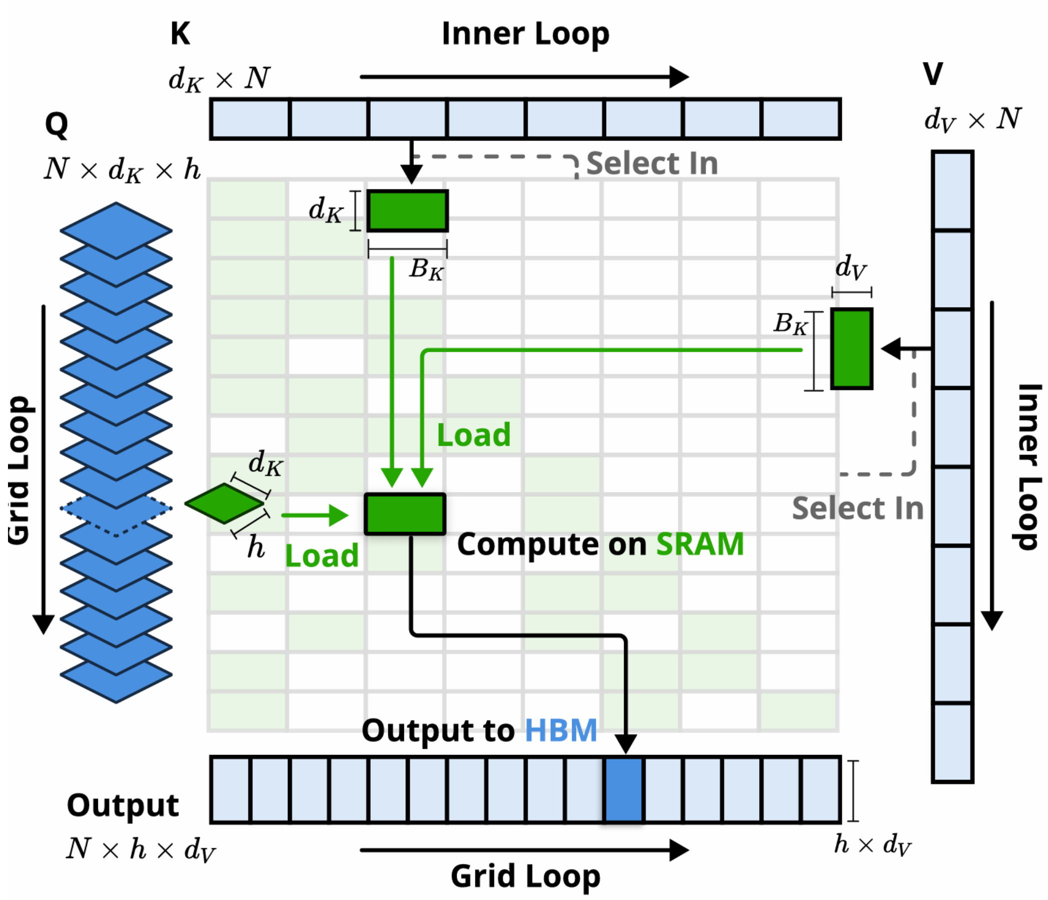


Figure 2: Overview of grouped-query method. Multi-head attention has H query, key, and value heads. Multi-query attention shares single key and value heads across all query heads. Grouped-query attention instead shares single key and value heads for each *group* of query heads, interpolating between multi-head and multi-query attention.

Key Innovation: Hardware-Aligned System

- **Hardware-Friendly Blockwise Loading**
- **Customized Head-wise Vectorized Kernel**
- **Balanced Arithmetic Intensity**



Evaluating Performance

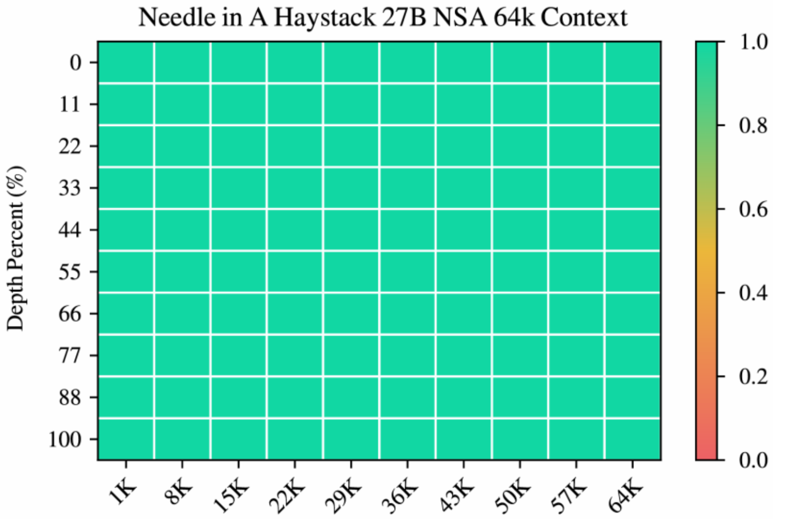
Outperforming Full Attention!

Model	MMLU Acc. 5-shot	MMLU-PRO Acc. 5-shot	CMMLU Acc. 5-shot	BBH Acc. 3-shot	GSM8K Acc. 8-shot	MATH Acc. 4-shot	DROP F1 1-shot	MBPP Pass@1 3-shot	HumanEval Pass@1 0-shot	Avg.
Full Attn	**0.567**	0.279	0.576	0.497	0.486	0.263	0.503	**0.482**	0.335	0.443
NSA	0.565	**0.286**	**0.587**	**0.521**	**0.520**	**0.264**	**0.545**	0.466	**0.348**	**0.456**

Superior General Performance

Model	SQA			MQA				Synthetic		Code	Avg.
	MFQA-en	MFQA-zh	Qasper	HPQ	2Wiki	GovRpt	Dur	PassR-en	PassR-zh	LCC	
H2O	0.428	0.429	0.308	0.112	0.101	0.231	0.208	0.704	0.421	0.092	0.303
InfLLM	0.474	0.517	0.356	0.306	0.250	0.277	0.257	0.766	0.486	0.143	0.383
Quest	0.495	0.561	0.365	0.295	0.245	0.293	0.257	0.792	0.478	0.135	0.392
Exact-Top	0.502	0.605	0.397	0.321	0.288	0.316	0.291	0.810	0.548	0.156	0.423
Full Attn	**0.512**	0.623	0.409	0.350	0.305	**0.324**	0.294	0.830	**0.560**	0.163	0.437
NSA	0.503	**0.624**	**0.432**	**0.437**	**0.356**	0.307	**0.341**	**0.905**	0.550	**0.232**	**0.469**

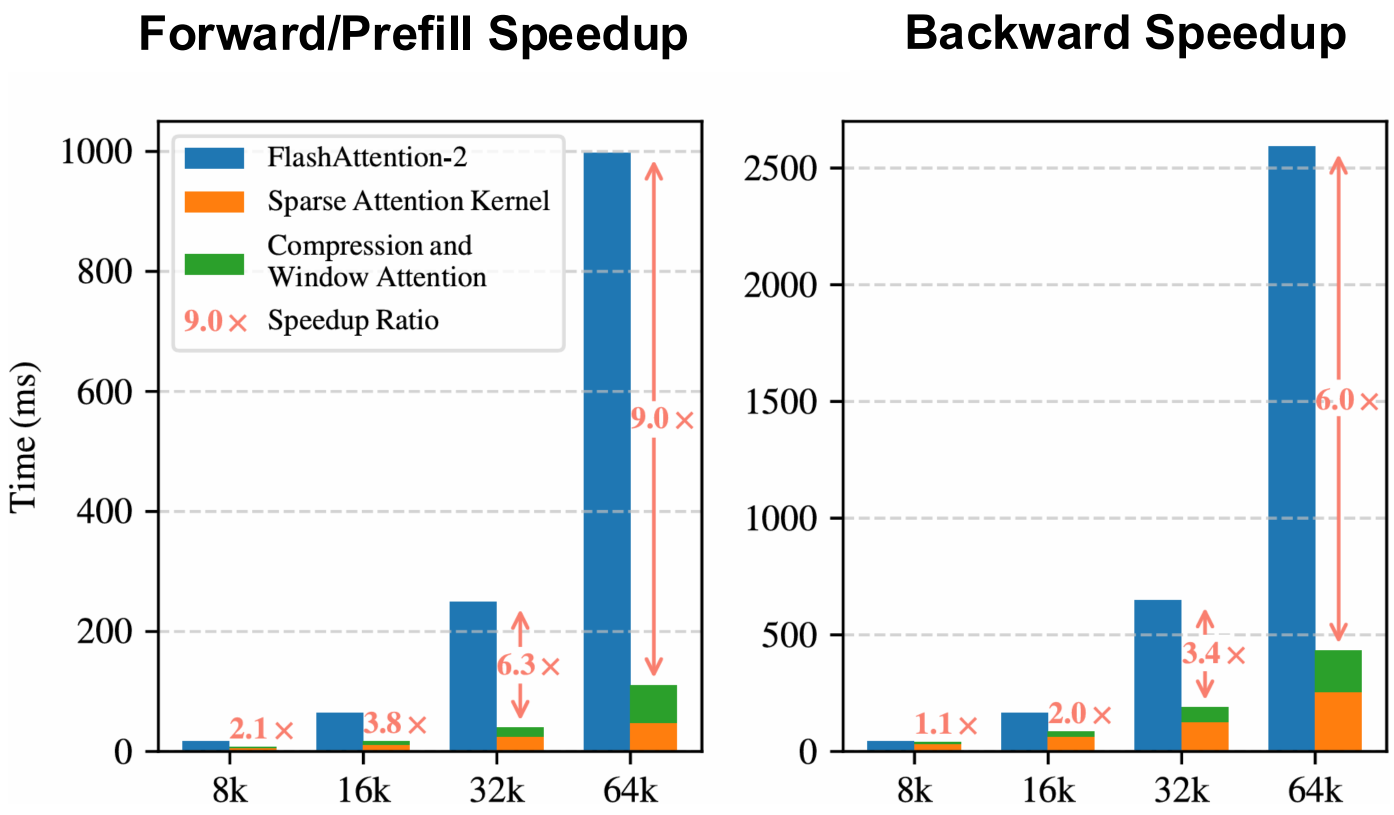
Long-Context Capability: LongBench



Generation Token Limit	8192	16384
Full Attention-R	0.046	0.092
NSA-R	**0.121**	**0.146**

Reasoning Ability

Efficiency: Substantial Speedups



Speedup in All Phases

Decoding Speedup

Context Length	8192	16384	32768	65536
Full Attention	8192	16384	32768	65536
NSA	2048	2560	3584	5632
Expected Speedup	**4×**	**6.4×**	**9.1×**	**11.6×**

Future Work

- ✓ Investigate Attention Score Patterns
- ✓ Improve Alternative Selection Strategies
- ✓ Overcome Key-Clustering Bottlenecks
- ✓ Extend Natively Sparse Training

The future is **Sparse**. NSA provides a efficient foundation for the **next generation of long-context LLMs**.

Conclusion of Our NSA

A Dedicate Hardware-Aligned System

Breaking the Performance-Cost Trade-Off

Catalyzing the next frontier of efficient LLM