IPDPS 2026 • New Orleans, USA

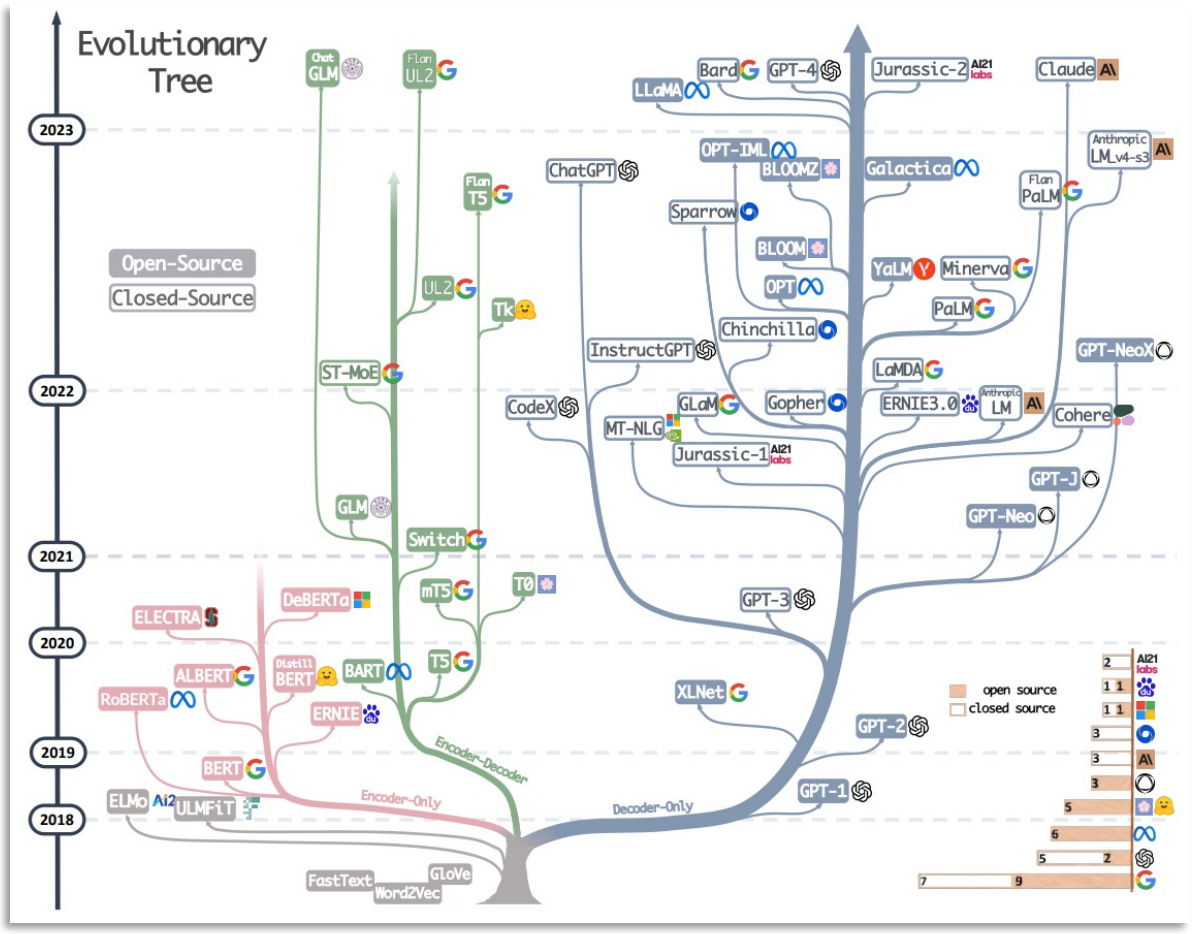
Gnosis Research Center

I/O-Aware PIM Acceleration for Long-Sequence LLM Inference with Hybrid Sparse Attention

Xiaoyang Lu, Lihan Hu
Hongrui Huang, Peng Jiang, Xian-He Sun

ILLINOIS TECH
IOWA
COLUMBIA UNIVERSITY

The Era of Large Language Models



Source: Yang et al., Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond

ChatGPT

Claude

Gemini

Grok

GitHub Copilot

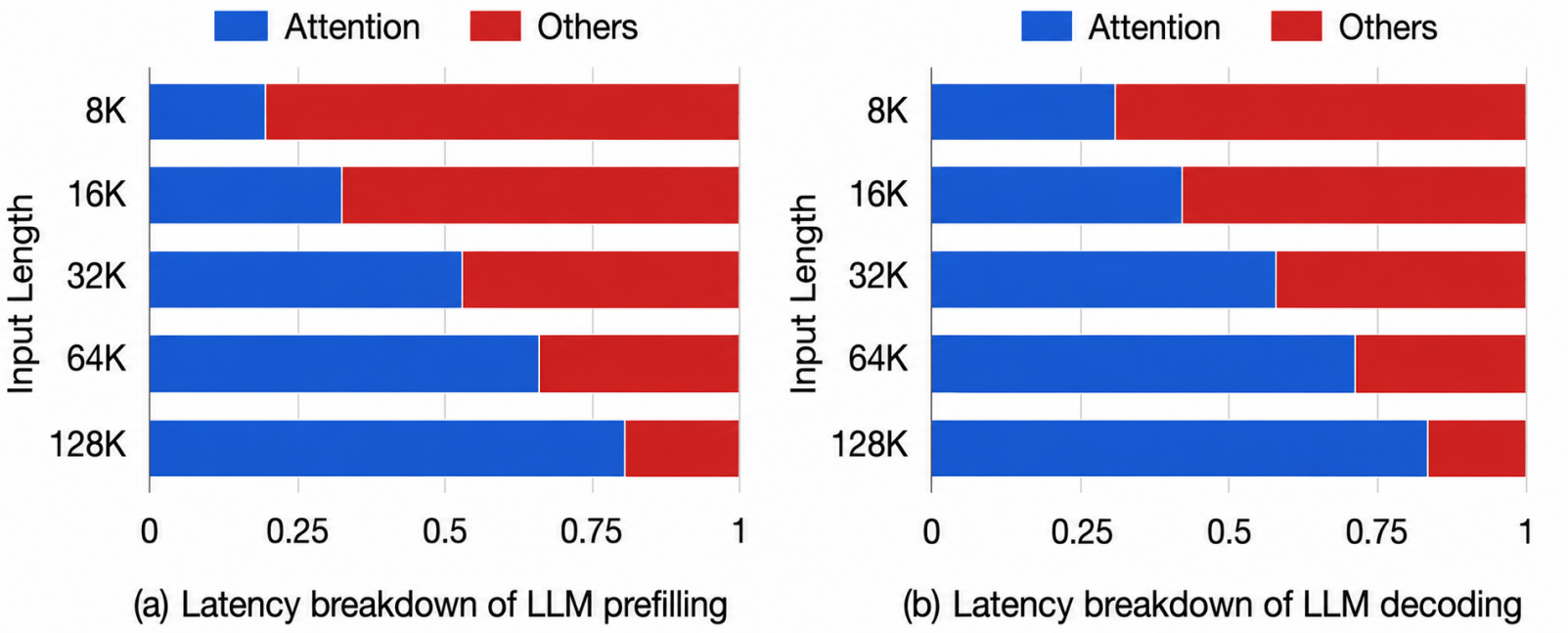
LLMs increasingly require long-sequence processing

How can we efficiently scale LLM inference to longer sequences?

Attention is Important

Attention is the core operation in Transformer-based LLMs

- Inputs: Q, K, V
- $O = \text{Softmax}(QK^T)V$
- Attention becomes LLM serving bottleneck as sequence gets longer

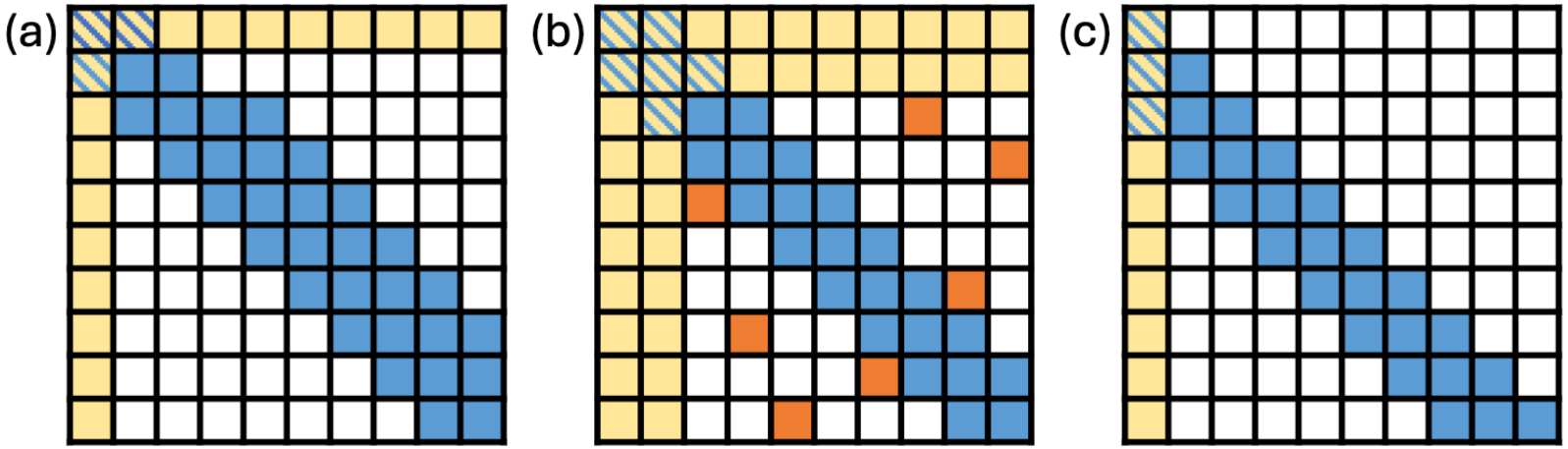


Attention dominates long-sequence inference latency

Attention dominates LLM inference as sequence length grows.
Measurements obtained with Llama-3-8B on NVIDIA A100 GPU.

Sparse Attention

Full attention: every token attends to all tokens
Sparse attention: each token attends to selected tokens
Hybrid sparse pattern: combines local, global, and random/retrieval attention
Goal: reduce computation for long-sequence inference

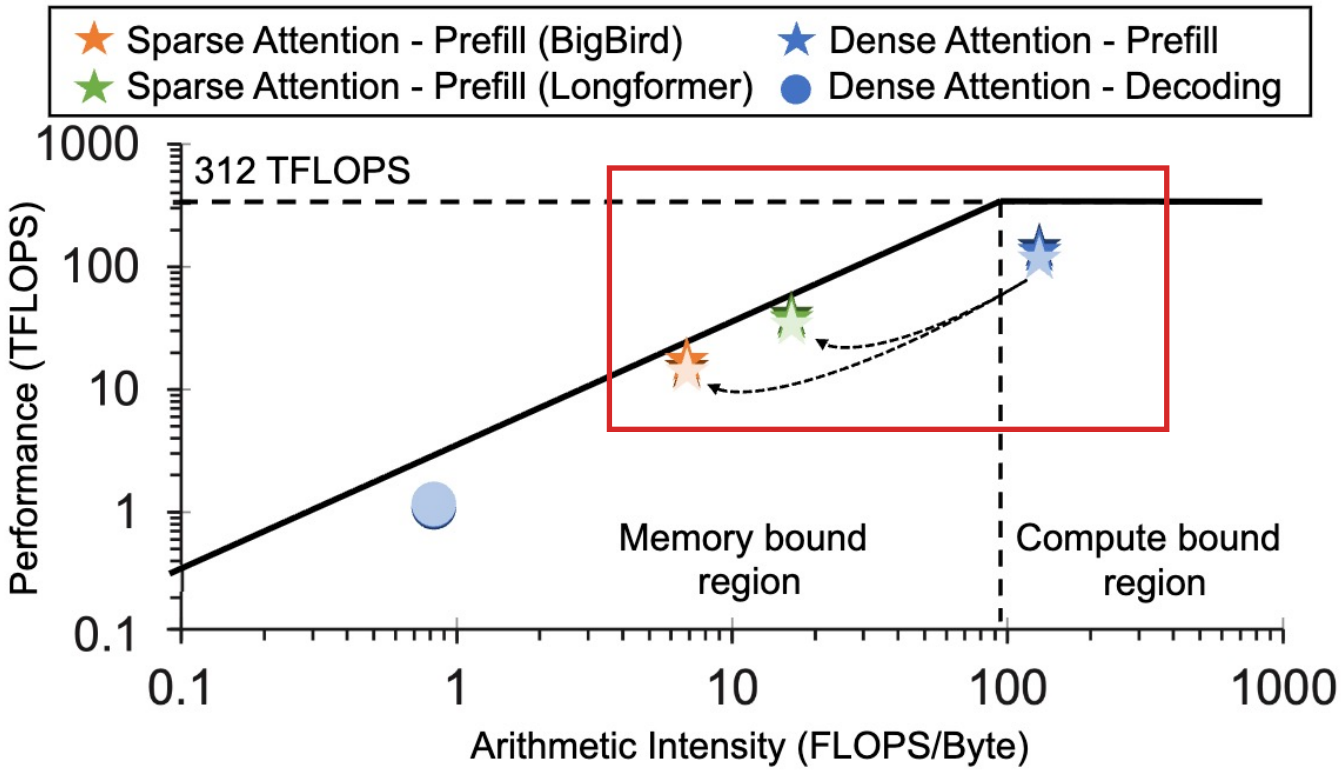


(a) Longformer: sliding-window + global tokens
(b) BigBird: sliding-window + global + random tokens
(c) DuoAttention: streaming heads with window + global tokens

Performance Characterization

Challenge: Memory-Bound

- Sparse attention reduces computation, but sparse prefill attention becomes **memory-bound**.

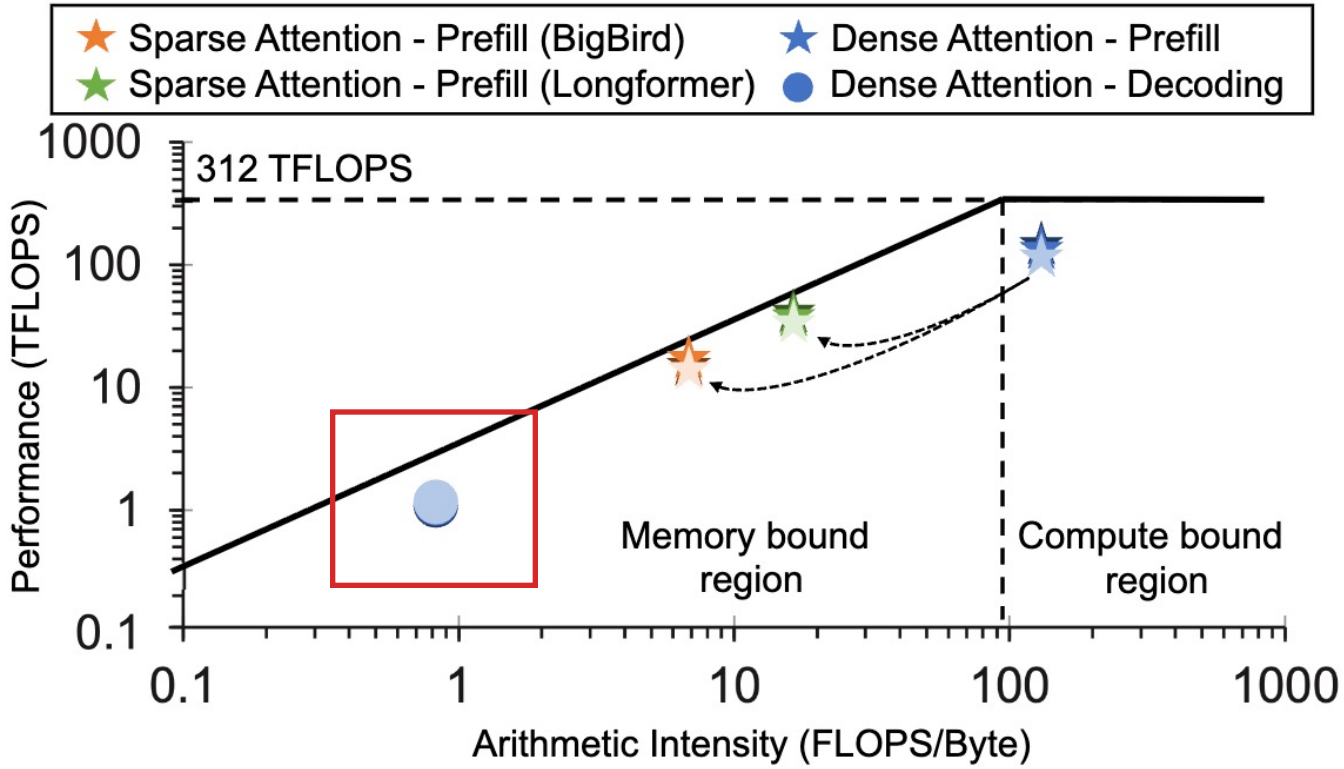


Roofline analysis of attention computations in LLM inference on an NVIDIA A100 GPU. Brightness indicates batch sizes: 4 (bright), 16 (moderate), and 32 (dark).

Performance Characterization

Challenge: Memory-Bound

- Sparse attention reduces computation, but sparse prefill attention becomes **memory-bound**.
- Decoding attention is inherently **memory-bound**.



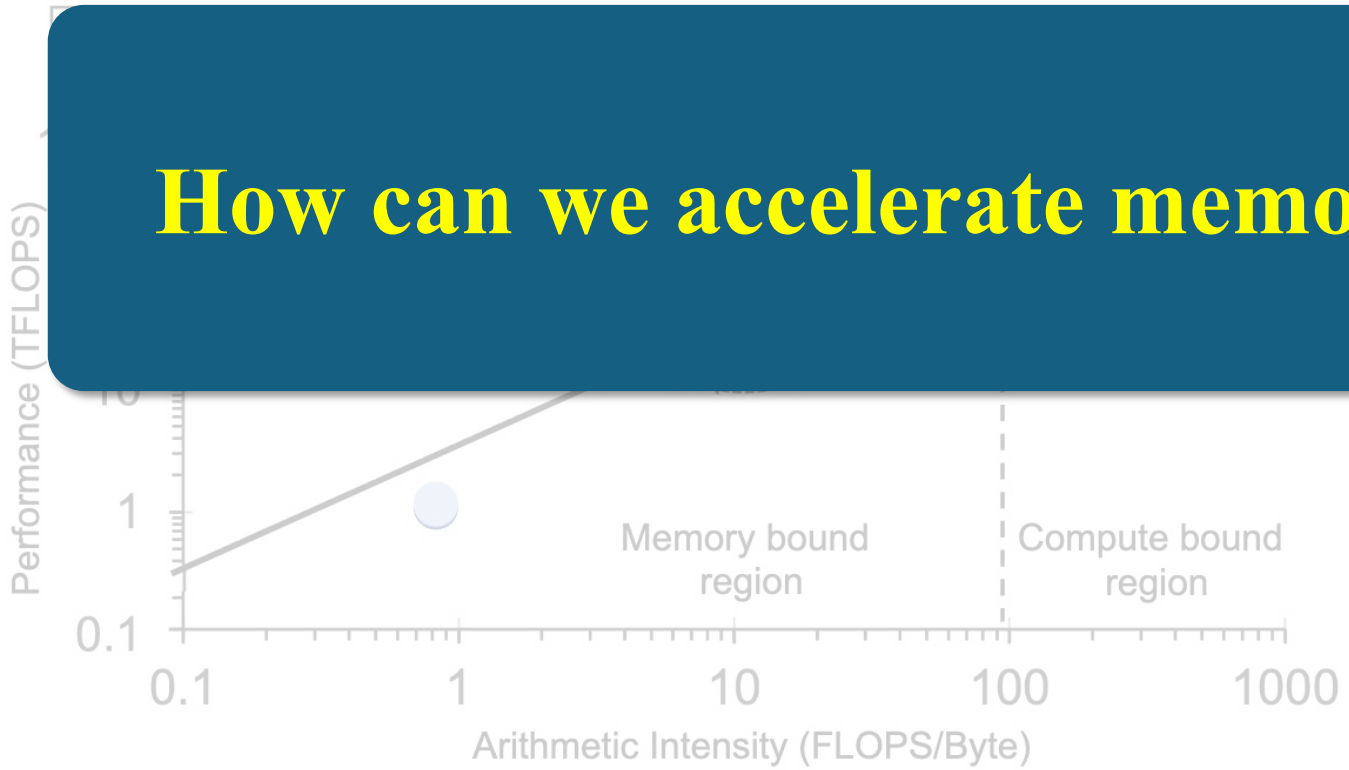
Roofline analysis of attention computations in LLM inference on an NVIDIA A100 GPU. Brightness indicates batch sizes: 4 (bright), 16 (moderate), and 32 (dark).

Performance Characterization

Challenge: Memory-Bound

- Sparse attention reduces computation, but sparse prefill attention becomes **memory-bound**.
- Decoding attention is inherently **memory-bound**.

How can we accelerate memory-bound attention?

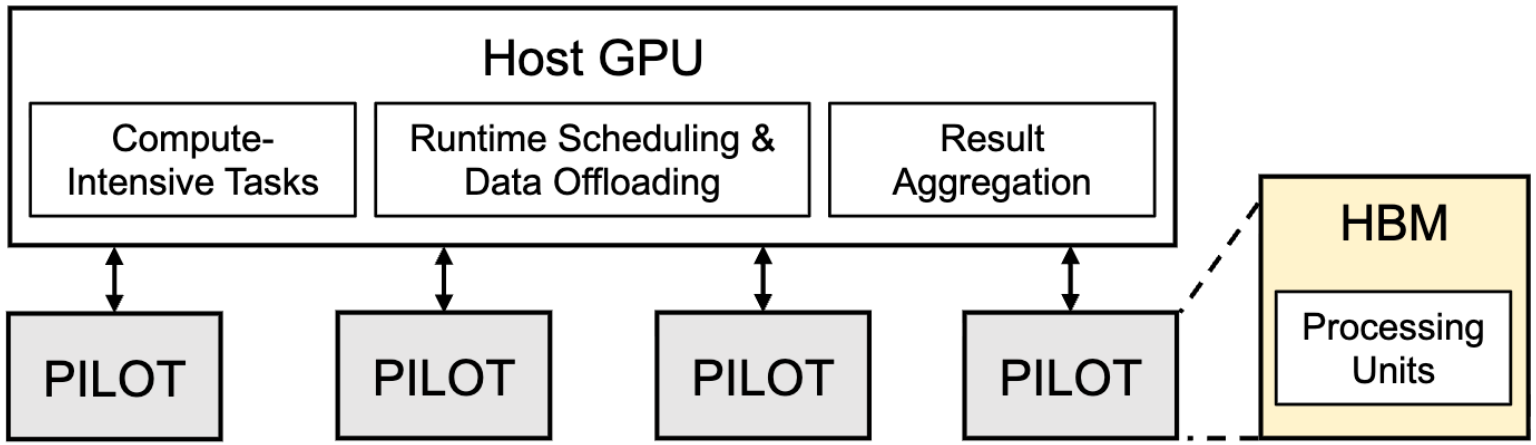


Roofline analysis of attention computations in LLM inference on an NVIDIA A100 GPU. Brightness indicates batch sizes: 4 (bright), 16 (moderate), and 32 (dark).

Heterogeneous System with Attention Accelerators (PILOTs)

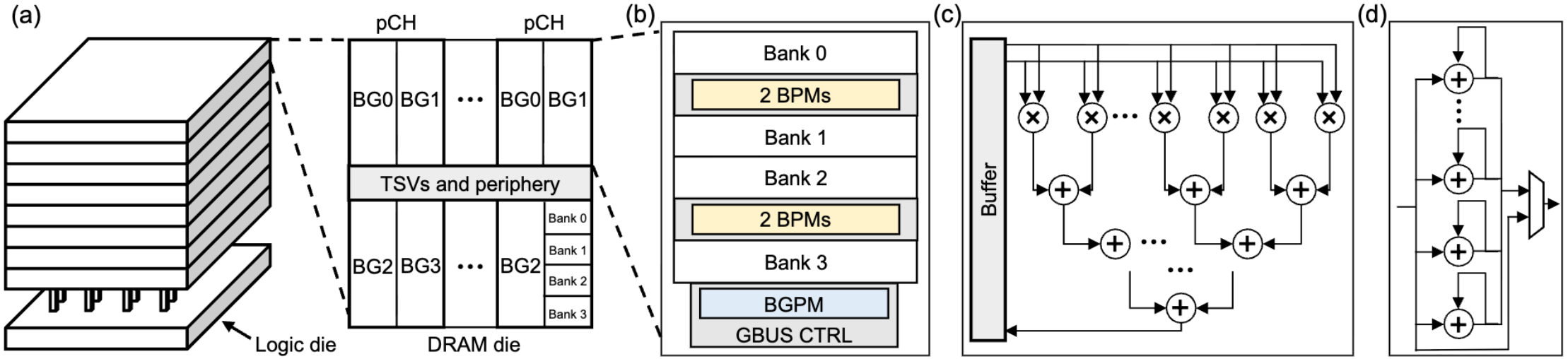
PILOT: PIM + I/O-Aware Co-Design

- **Key idea:** accelerate memory-bound attention by reducing data movement
- **PIM:** moves attention computation near HBM to reduce external I/O (between GPU and HBM)
- **I/O analysis:** guides mapping, tiling, and scheduling to reduce internal I/O (inside the PIM hierarchy)
- **GPU-PILOT system:** GPU handles compute-intensive tasks; PILOT handles memory-bound attention



PILOT Architecture

HBM-based PIM accelerator for memory-bound attention
BPM: bank-level processing module for local MAC and softmax-related operations
BGPM: bank-group-level module for lightweight inter-bank reductions
Goal: exploit HBM internal bandwidth and reduce external I/O between GPU and HBM.

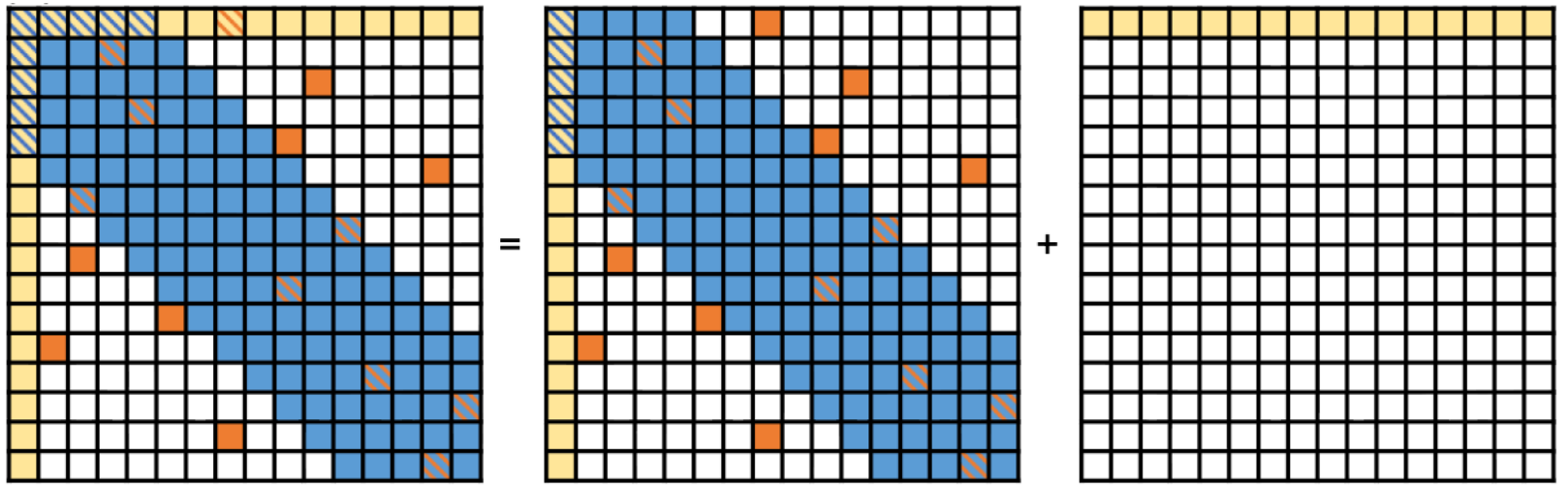


The architecture of PILOT. (a) Organization of an HBM stack, (b) processing units within HBM (bank-level and BGlevel modules), (c) Bank-level processing module (BPM), (d) BG-level processing module (BGPM).

Data Mapping for Hybrid Sparse Attention

Goal: reduce cross-bank communication and improve bank-level parallelism

(1) Hybrid sparse attention:
non-global rows + global rows

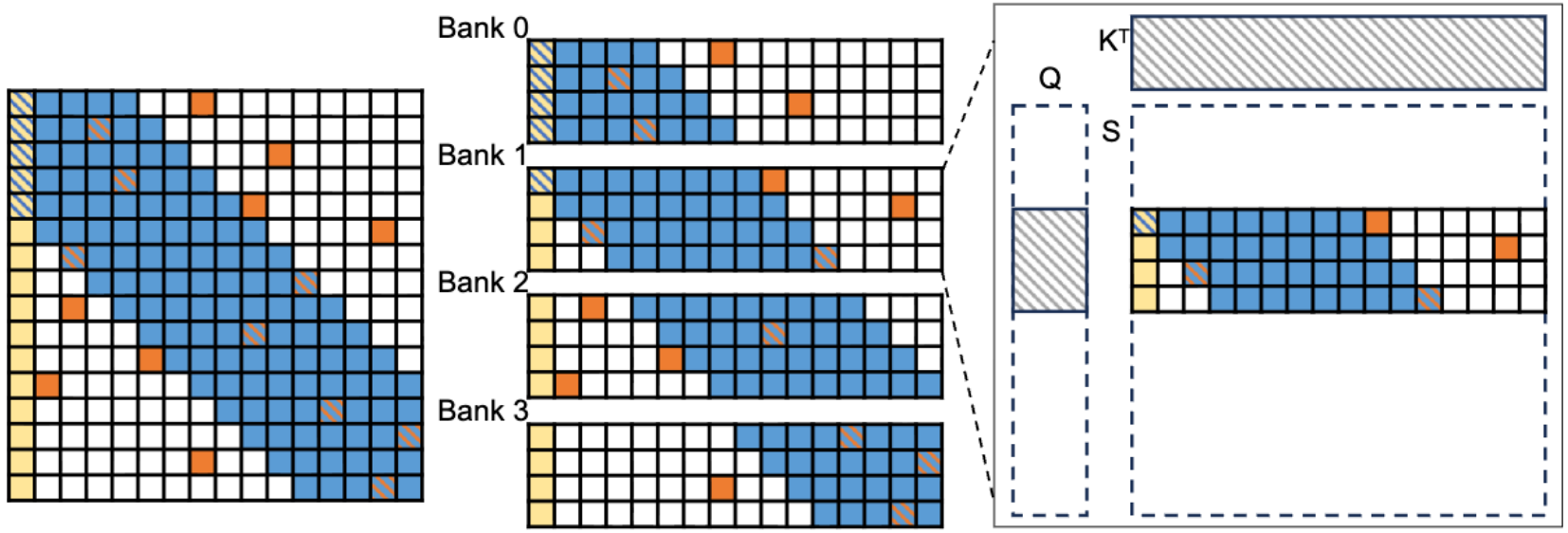


Hybrid sparse attention Non-global rows Global rows

Data Mapping for Hybrid Sparse Attention

Goal: reduce cross-bank communication and improve bank-level parallelism

(1) Hybrid sparse attention:
non-global rows + global rows
(2) Non-global rows: partition Q rows across banks; replicate K/V



Assign one attention head to one bank group and distribute rows across banks

Data Mapping for Hybrid Sparse Attention

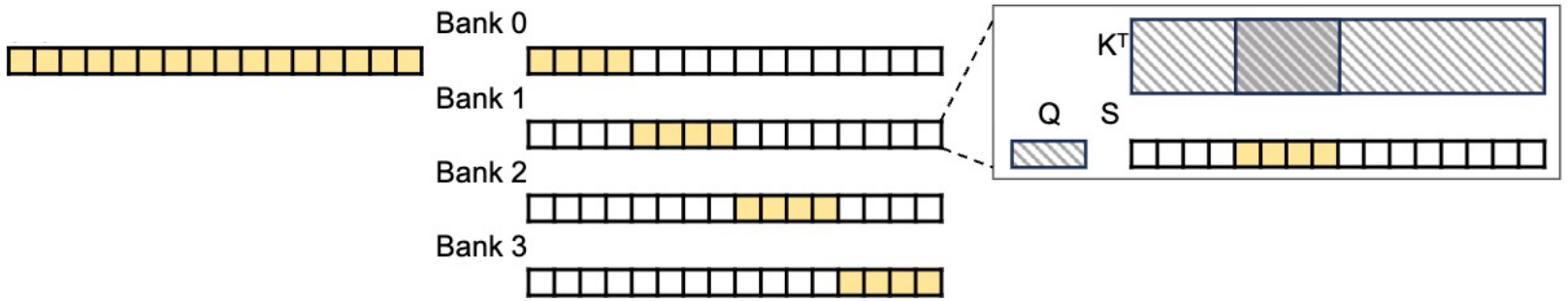
Goal: reduce cross-bank communication and improve bank-level parallelism

(1) Hybrid sparse attention: non-global rows + global rows

(2) Non-global rows: partition Q rows across banks; replicate K/V

(3) Global rows: replicate Q across banks; use local K/V data

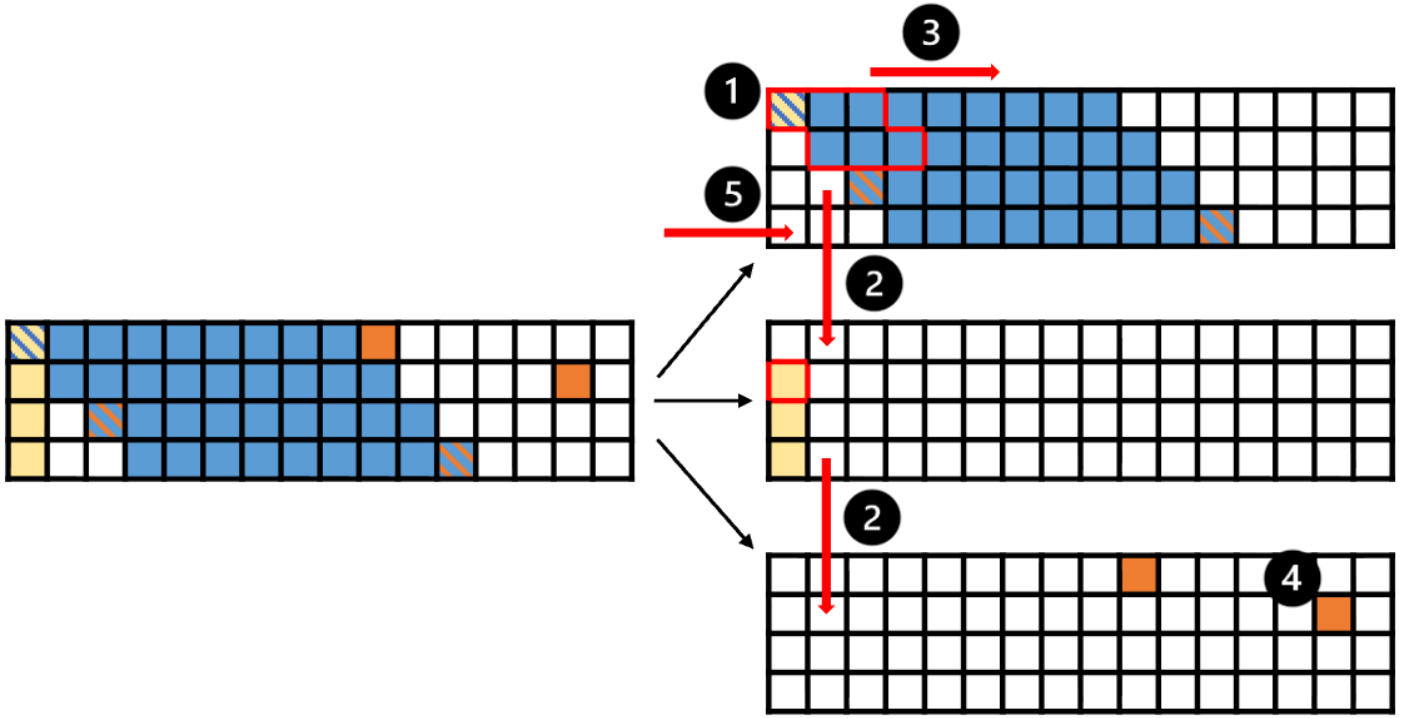
(4) Decoding: handled like global rows



Replicate Q across banks

I/O-Aware Tiling and Scheduling

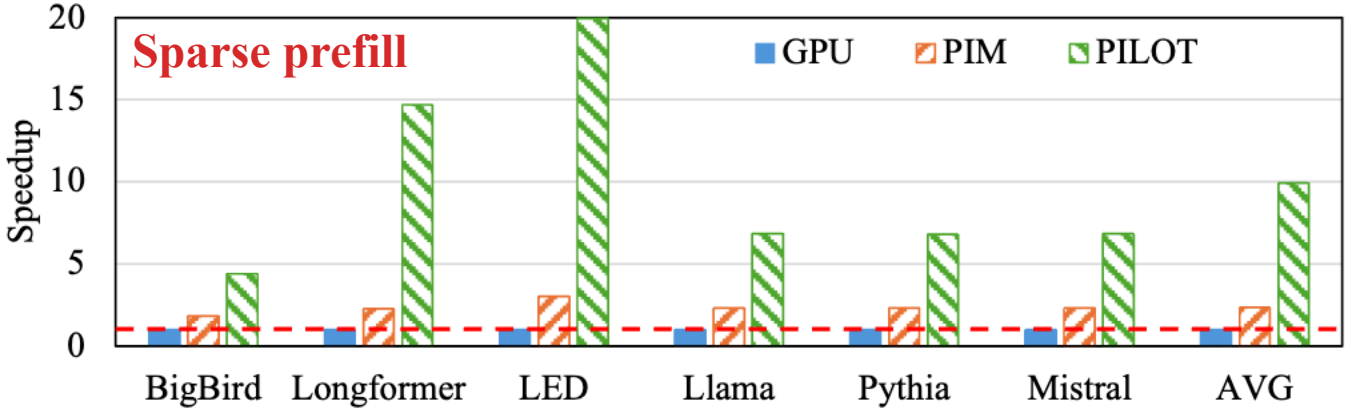
- Keep reusable tiles in the local SRAM buffer
- Stream other tiles through the buffer
- Determine tile sizes using **I/O analysis** under local buffer capacity
- **Maximize data reuse and compute-to-I/O ratio**
- Apply to sparse prefill, global rows, and decoding



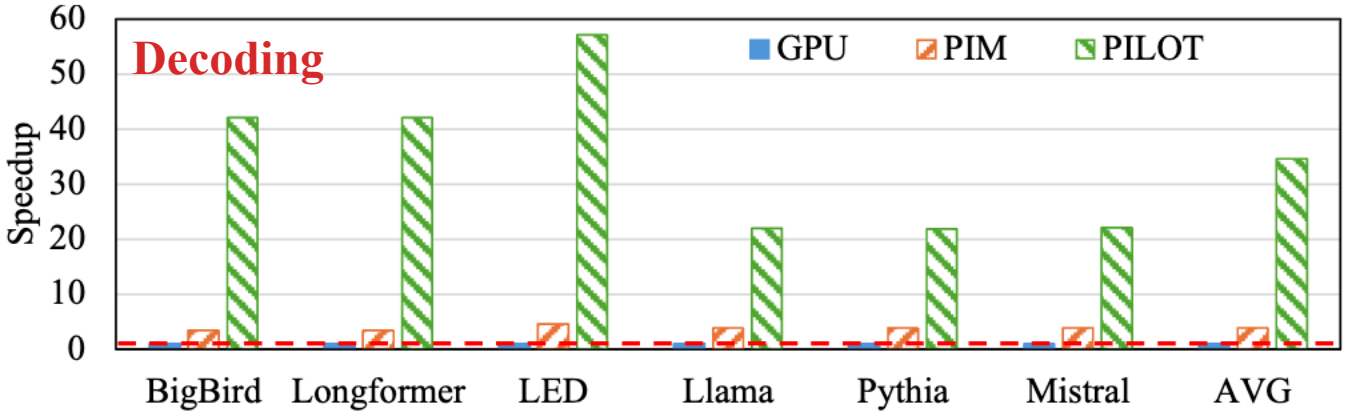
Per-bank Q-stationary scheduling for SDDMM in sparse attention during the prefill stage for non-global rows.

PILOT uses I/O analysis to reduce internal I/O inside the PIM hierarchy

PILOT Accelerates Memory-Bound Attention



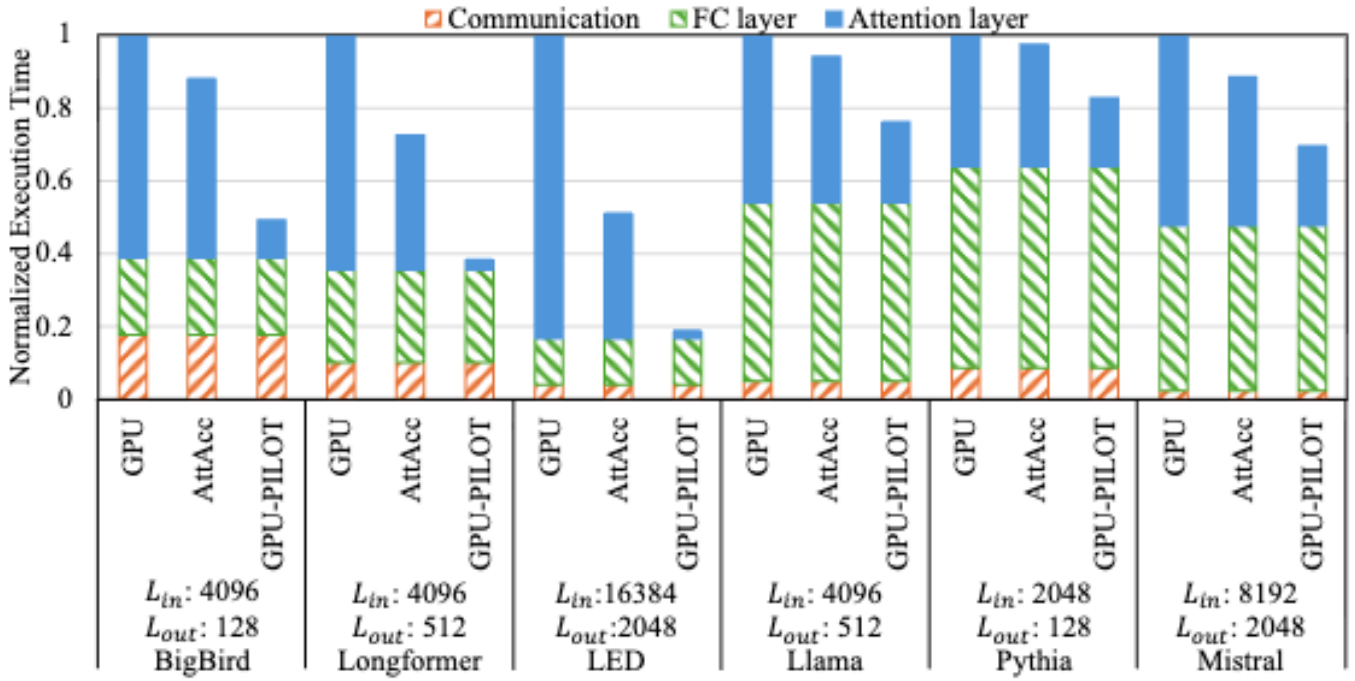
(a) Sparse attention performance for different models in prefill stage.



(b) Attention performance for different models in decoding stage.

- **Sparse prefill:** up to **19.96×** over GPU; **9.93×** on average
- **Decoding:** up to **57.11×** over GPU; **34.59×** on average
- PILOT also outperforms **vanilla PIM** without I/O-aware optimization
- **Key reason:** reduces both external I/O and internal I/O

GPU-PILOT Improves End-to-End LLM Inference



Normalized execution time of various long-sequence models on GPU, AttAcc, and GPU-PILOT systems.

- **GPU-PILOT** outperforms GPU and **AttAcc** across all models
- Up to **5.31×** over GPU
- Up to **2.71×** over AttAcc
- Benefit comes from accelerating **memory-bound sparse prefill + decoding attention**

Conclusion

Sparse prefill attention and decoding attention are **memory-bound**

PIM reduces external I/O by moving attention computation closer to memory

I/O analysis reduces internal I/O through mapping, tiling, and scheduling

PILOT reduces I/O and improves end-to-end long-sequence LLM inference

IPDPS 2026 • New Orleans, USA

Gnosis Research Center

I/O-Aware PIM Acceleration for Long-Sequence LLM Inference with Hybrid Sparse Attention

Xiaoyang Lu, Lihan Hu, Hongrui Huang, Peng Jiang, Xian-He Sun

xlu40@illinoistech.edu

ILLINOIS TECH

IOWA

COLUMBIA UNIVERSITY