FMS 2025

the Future of Memory and Storage

PLiOPS
LightningAI

Beyond Prefix Caching

Moshe Twitto
CTO & FOUNDER

Expert in advanced data management and coding algorithms. Prior to co-founding Pliops, served as CTO of Samsung's SSD Controller Development Center in Israel, holds MSEE, BSEE degrees from Technion University, Summa Cum Laude.

Booth #840

Better GenAI Inferencing
w/ FusIOnX Optimizations

KV-Cache Reuse Matters

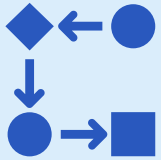
Expensive Inference

LLM Cost per/ Query

Efficient KV Reuse

Reused Cache Blocks
Reduced Inference Time

Prefix-Based Dominance



Prefix-Based Reuse is dominating

LLM inference is costly → KV-cache reuse is the key → Prefix-based reuse leads today

Emerging technologies are breaking the prefix assumption

LLM Applications w/ Compositional Context

Retrieved chunks and their order vary per query

Must-have for *60%* of LLM systems, *86%* of organizations

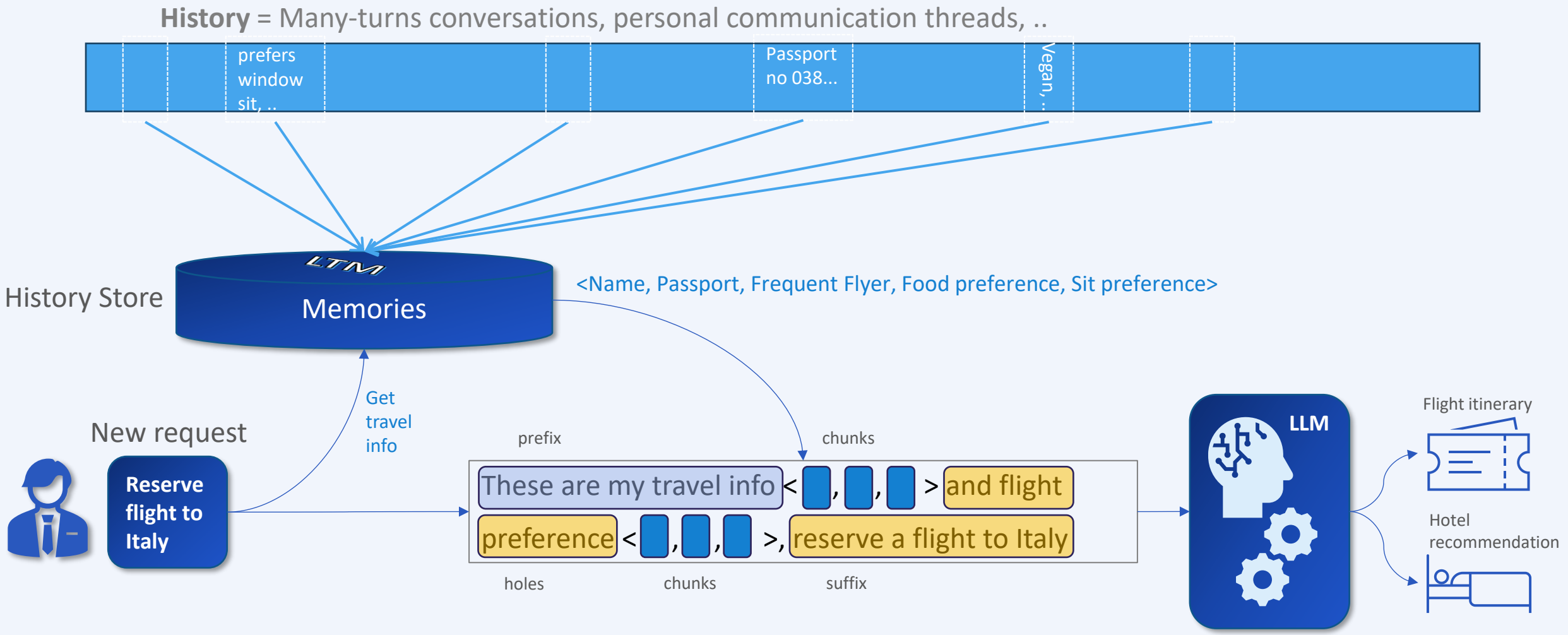
Dynamic reasoning and sub-queries
Balancing transparency with memory efficiency

Personal assistant
Partial, summarized, reordered personal context history

PLiOPS
LightningAI

Monolithic KV-cache reuse is not guaranteed
BUT full recompute is costly and slow

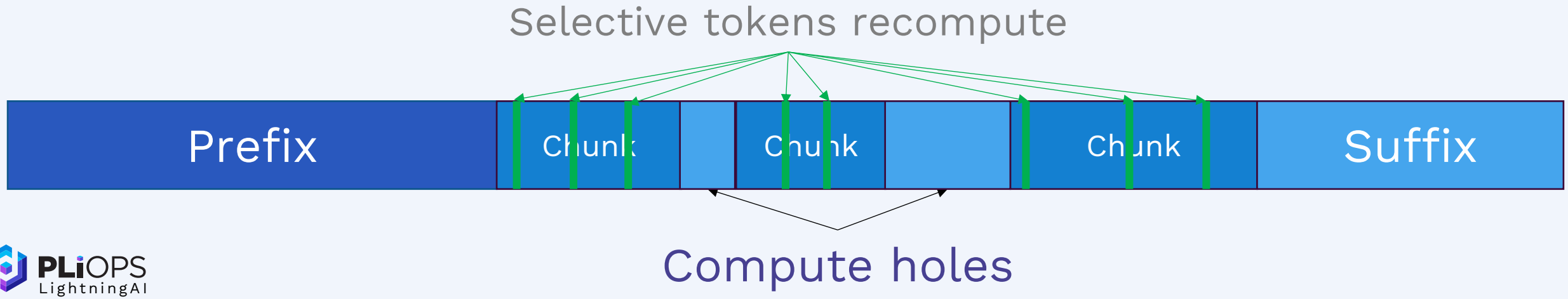
Long-Term Memory In Action



Optimization Opportunity: Compositional Reuse

Opportunity

- Wider applicability of KV-cache offloading
- Reuse pre-generated prompt *chunks*
- *Selective* token recompute to adjust chunks to new context



Pliops Solution Highlights

Content-Based Chunking

- Eliminates ingestion phase - no offline processing
- No need for explicit chunk marking
- Includes dynamically generated chunks, like summarization

Minimal application friction

Maximal use-cases coverage

Chunk De-Noising

- Need to clean the impact of initial context
- Low online recompute overhead per chunk

Minimal performance overhead

Chunk Eviction

- Storing “everything” may explode storage requirements
- Smart automatic eviction of non-reusable chunks

Optimal storage capacity

PLiOPS LightningAI

Summary

- Emerging prompt composing techniques break the prefix assumption
- We suggest a new KV-cache offloading paradigm to enable the reuse of “randomly” chosen chunks of pre-generated KV cache
- Initial traces analysis demonstrates a hit-ratio increase from ~60% to ~90%
- Results in ~4x average TTFT reduction compared to prefix caching
- Increases capacity and IOPS demand from the storage system
  - Very low temporal locality
  - Chunks range 32-128 tokens on average
- Pliops FusIOnX solution meets the requirements

FMS
the Future of Memory and Storage

PLiOPS
LightningAI

Thank You

Booth #840

Santa Clara Convention Center
August 5-7, 2025
19th Annual Conference and Exhibition