Storage Class Memory is Dead, All Hail **M**anaged-**R**etention **M**emory: Rethinking Memory for the AI Era

Sergey Legtchenko, **Ioan Stefanovici**, Richard Black, Ant Rowstron,
Junyi Liu, Paolo Costa, Burcu Canakci, Dushyanth Narayanan, Xingbo Wu

Microsoft Research

Paper appeared at HotOS'25

AI Inference – The Dominant Cloud Workload

Generative AI has changed the game & inference demand is huge

NEWS | 10 April 2025

Data centres will use twice as much energy by 2030 — driven by AI

These facilities accounted for roughly 1.5% of global electricity consumption in 2024.

By Sophia Chen

AI EFFECT

Tech megacaps plan to spend more than $300 billion in 2025 as AI race intensifies

PUBLISHED SAT, FEB 8 2025·8:00 AM EST | UPDATED SAT, FEB 8 2025·11:02 AM EST

Samantha Subin
@SAMANTHA_SUBIN

SHARE

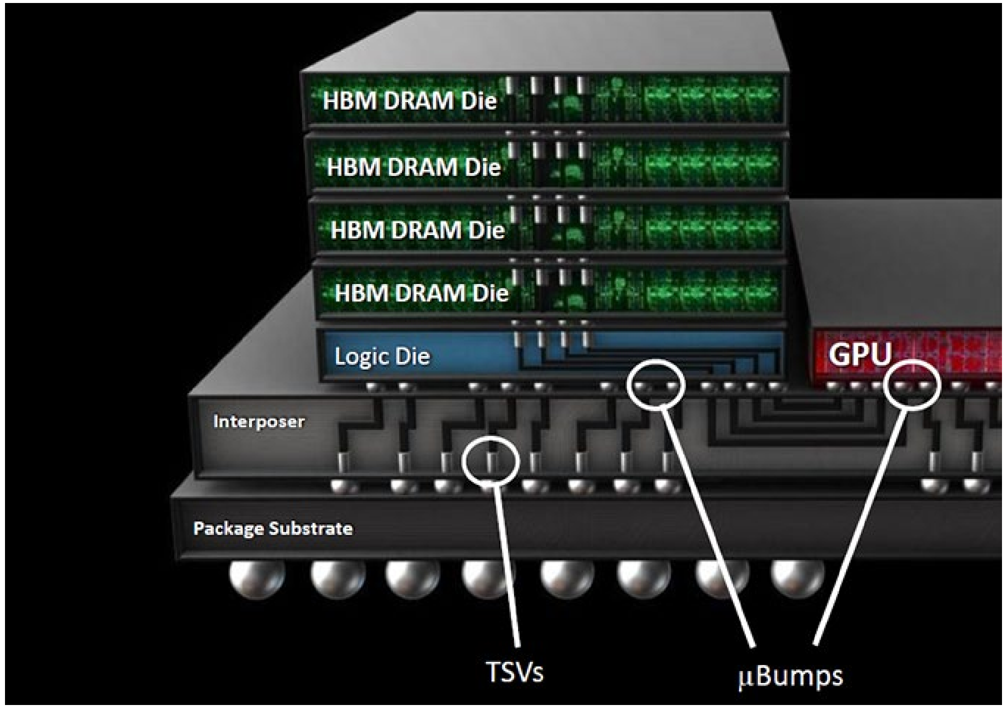
A great challenge and opportunity for the research community to rethink system hardware and software architectures

The Problem Today: The Curse of HBM

High **B**andwidth **M**emory is the **only** option today to achieve good bandwidth to AI data

But, a litany of problems…

- Complex manufacturing and packaging
- Unreliable
- Expensive: **significant** portion of GPU cost and power



(source: https://www.anandtech.com/show/9969/jedec-publishes-hbm2-specification)

What is HBM in *LLM Inference* Actually Used For?

LLM inference: the prominent workload

Two large data structures: model weights + KV cache

Model weights (~ 50% today)
- Write: once
- Read: each forward pass

KV-cache (~ 50% today):
- Write: append-only
- Read: in whole each forward pass

Observation: very large, predictable, block Reads dominate
- HBM is "overprovisioned" on write performance
- Small, random access of HBM not necessary

Can we leverage the specific properties of AI inference to design a better memory?

A New Class of Memory for AI Inference

- “New” memory technologies: STT-MRAM, ReRAM, PCM, FeRAM,...
  - Viewed through “Storage Class Memory” lens – long-term data *retention* was a goal
- For AI Inference:

Important Metrics	Less Important Metrics
Capacity / $	Write performance
Read bandwidth	Small, random access
Energy	**Long-term Retention**

Key insight: possible to trade-off *write performance & retention time* for important metrics

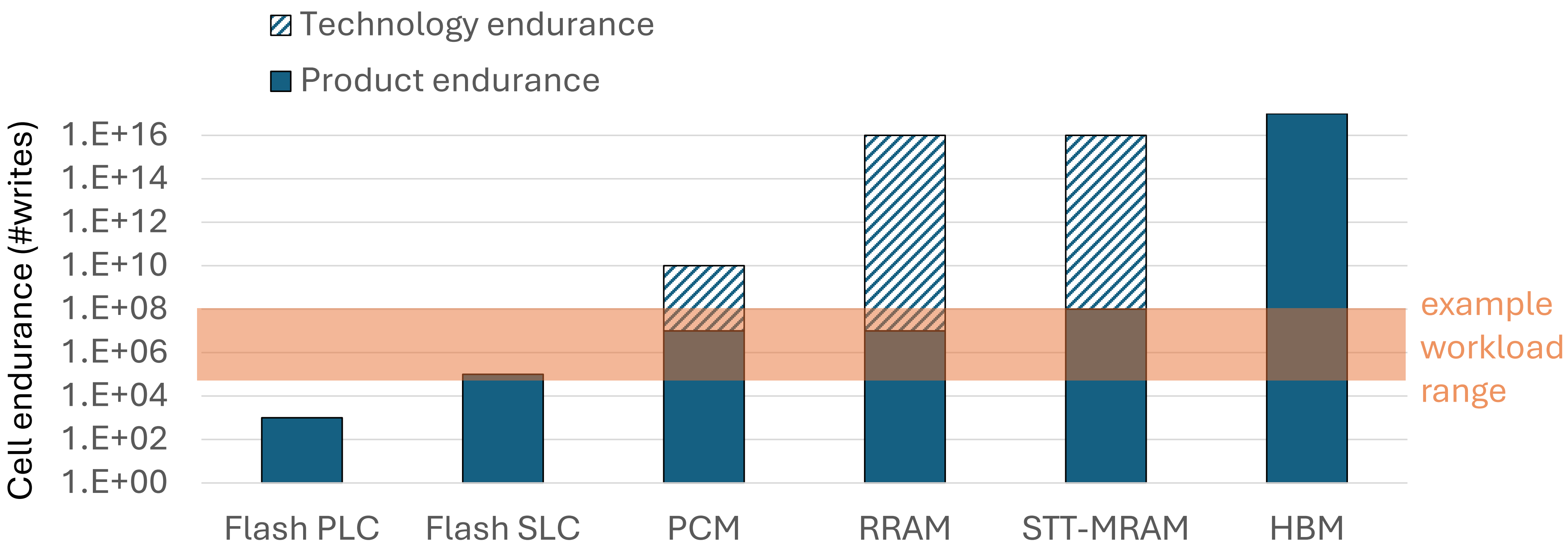
- Storage Class Memory non-volatility (10+ yr retention) is not required
- Hours-long retention time is sufficient and enables power advantage

Managed-**R**etention **M**emory: a new class of memory for AI inference

MRM: A New Opportunity for SCM Technologies

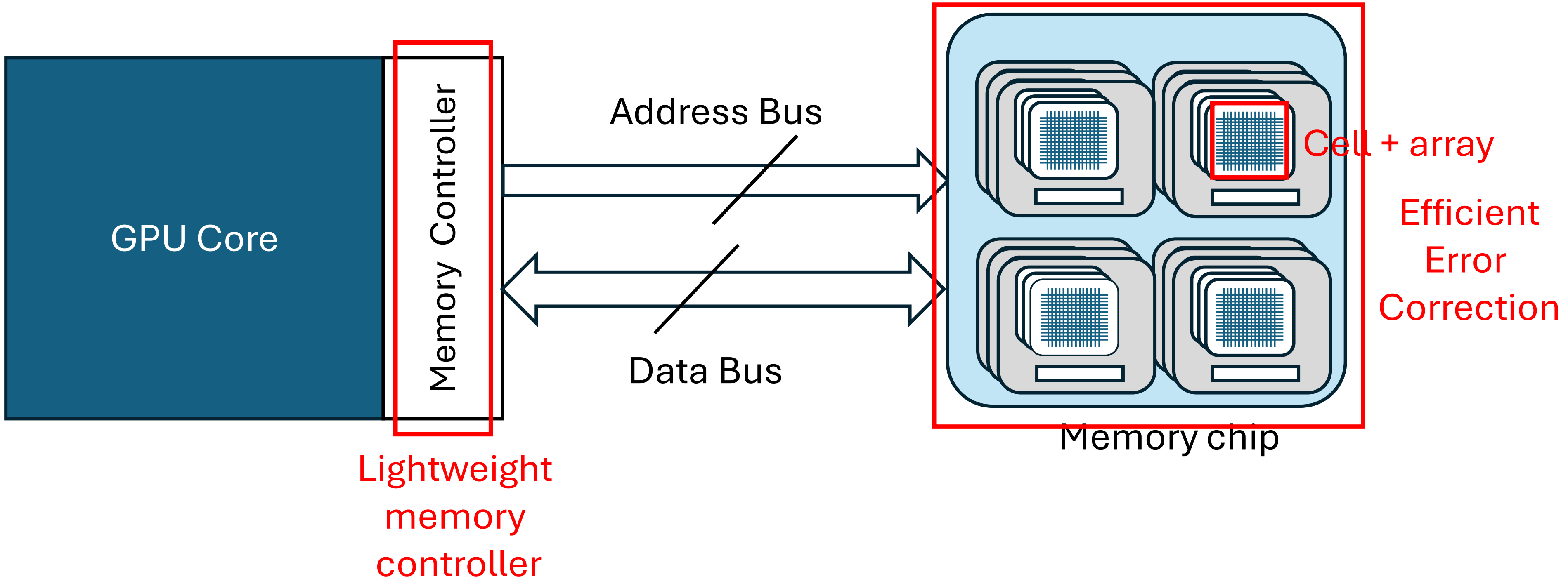
Existing memory technologies **can** inherently be optimised for MRM

example trade-off: retention ↑ endurance ↓



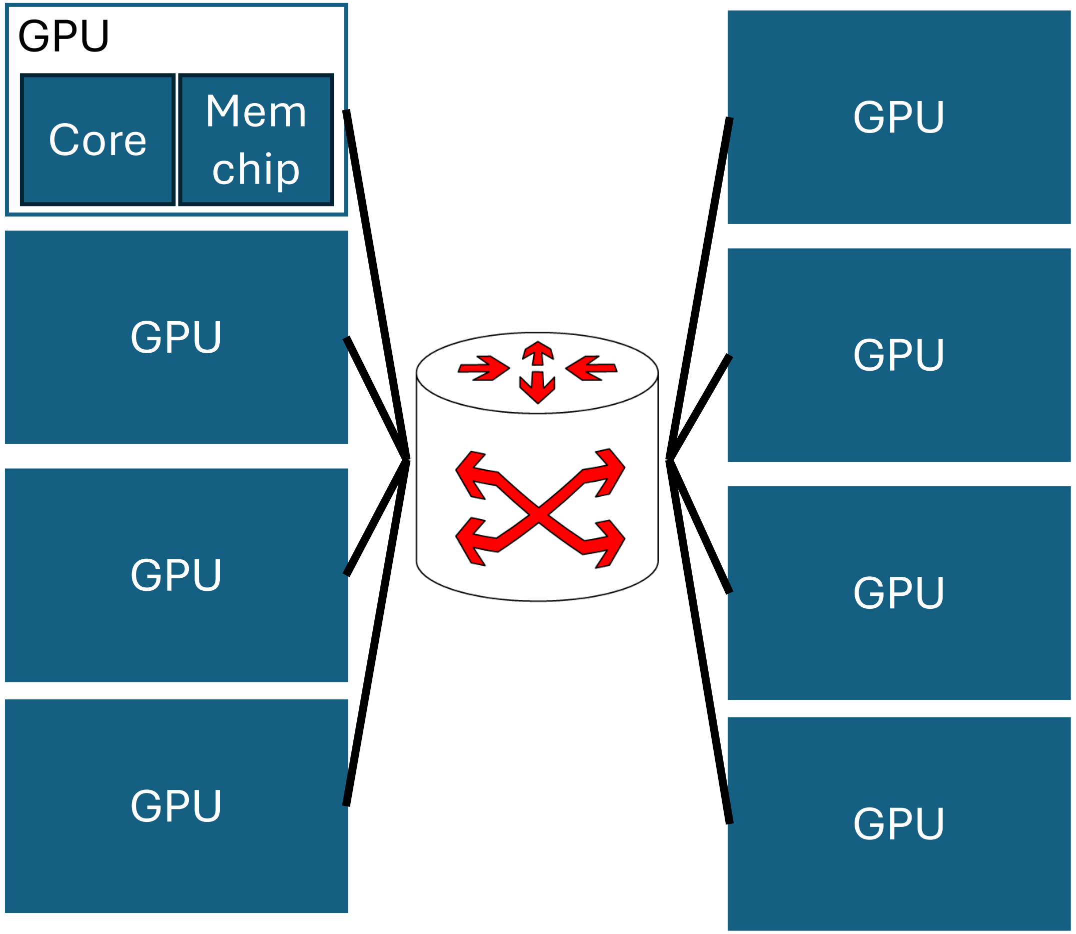
MRM Research Opportunities

- Innovation across the HW/SW stack needed
- How do we leverage workload + cell properties?
  - lack of small random access, lack of refreshes



MRM Systems Research Opportunities

- MRM abstraction: how to expose MRM to systems?
- Dynamically configurable retention
  - Should software configure retention period per write?
- Retention-aware data placement & scheduling
  - Software-driven movement



MRM: Rethinking Memory for the AI Era

1) Massive opportunity and need to disrupt HBM for AI inference

2) **M**anaged-**R**etention **M**emory: a new class of memory that trades off retention and write performance for energy, read performance, and cost

3) Ripe for innovation across cells, arrays, controllers, system abstractions, and much more!