MemoryDAX: What's Next

Philip Levis
First MemoryDAX/DAM Winter Workshop
1/10/2025

Thank You

In One Slide

- Memory is increasingly the bottleneck in computing systems
  - ML accelerators: SRAM, HBM bandwidth and capacity
  - Cloud and datacenter servers: DDR bandwidth and cost
- Further gains in performance require transformative changes to memory
- Understanding *which* changes requires understanding many layers
  - Devices and circuits: what memories are possible and what are their tradeoffs?
  - Architecture: how do we organize memories and maintain coherence?
  - Software: what do we need memory to do?
- We've just started a 5-year project to explore and help define the future of memory: 6 months in

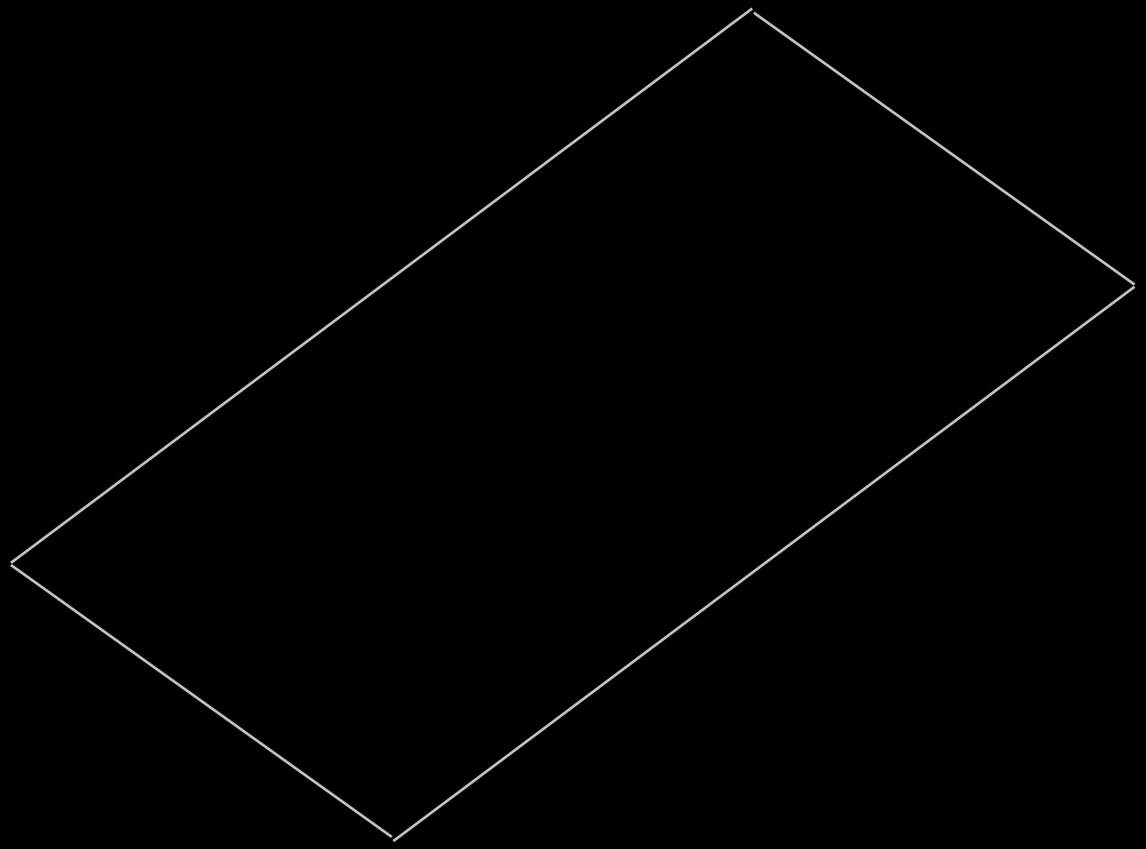
Work Presented Today

Software, A Memory Perspective	Massive, Diverse, Tightly Integrated with Compute – from Device to Software
	Memory Access Pattern Classification
	Data Lifetime and Its Refresh Implications
Architectural Primitives and Protocols	MemGlue for Heterogeneous Architectures
	Conserving Memory Bandwidth with Virtual Gather
Memory, Safe at any Speed	Integration: Performance, Power, and Thermal Constraints
	Gain Cell Compiler
	Synthesizing High Level Models from RTL

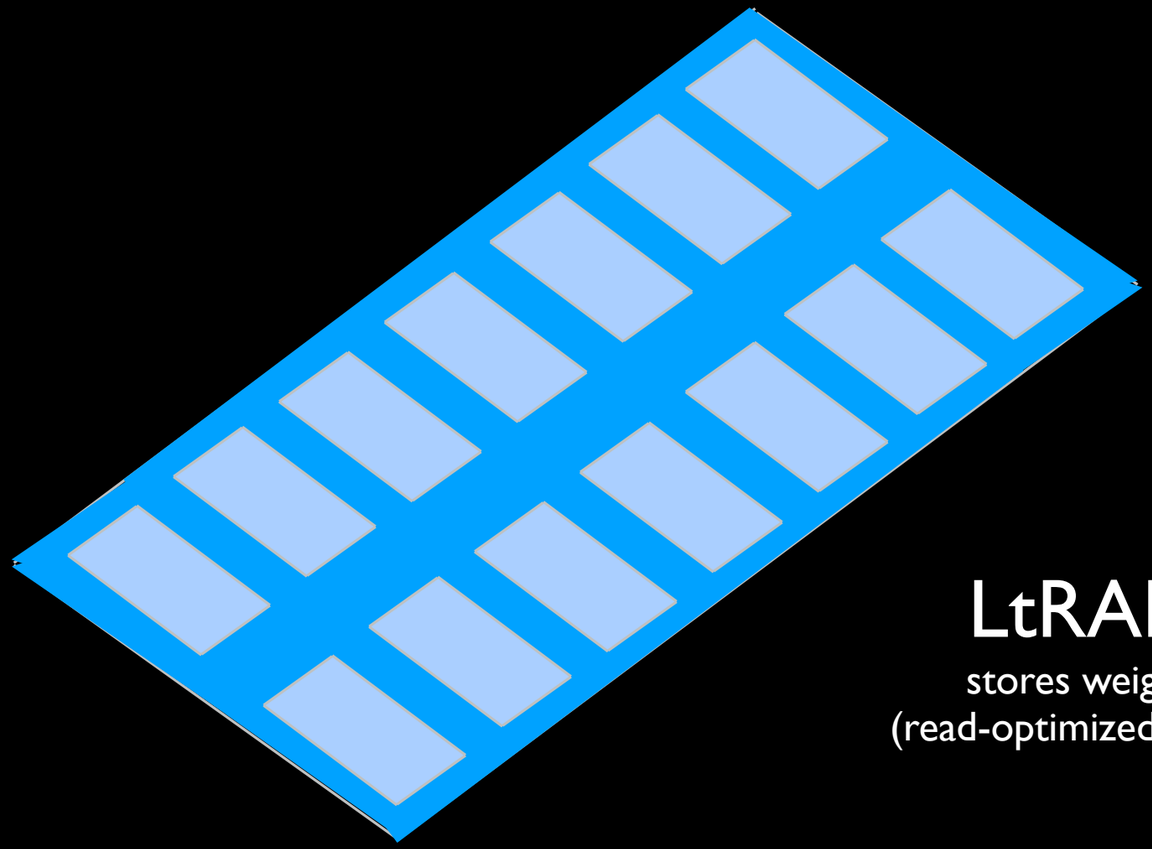
Ongoing Work

- LtRAM: design, tradeoffs, and use cases
- StRAM: design, tradeoffs and use cases
- How can programming frameworks guide memory use?
  - ML frameworks (PyTorch, TensorFlow, Ray)
  - Analytics frameworks (Spark)
  - Database engines/optimizers
  - Functional operating systems (Fixpoint)
- Processor designs (instruction caches, I/O caches)
- ML accelerator designs

Inference Accelerator



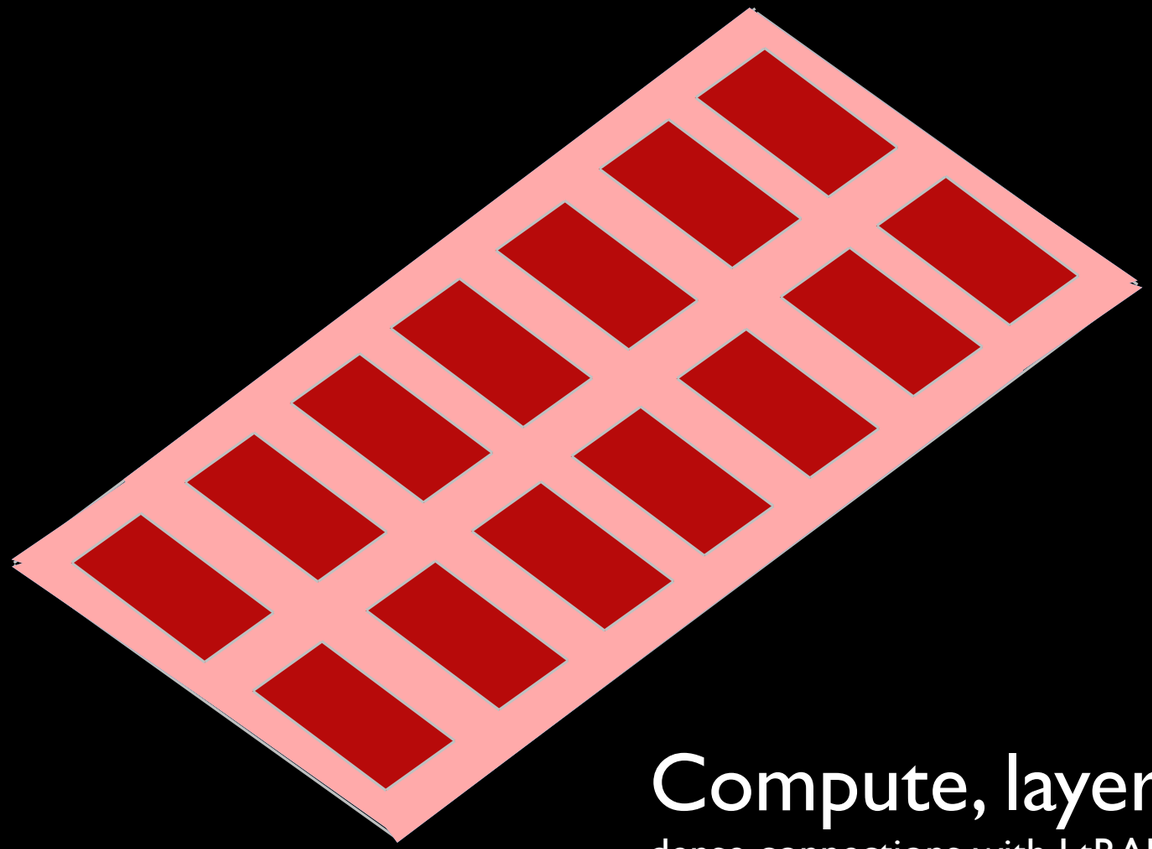
Inference Accelerator



LtRAM

stores weights
(read-optimized, dense)

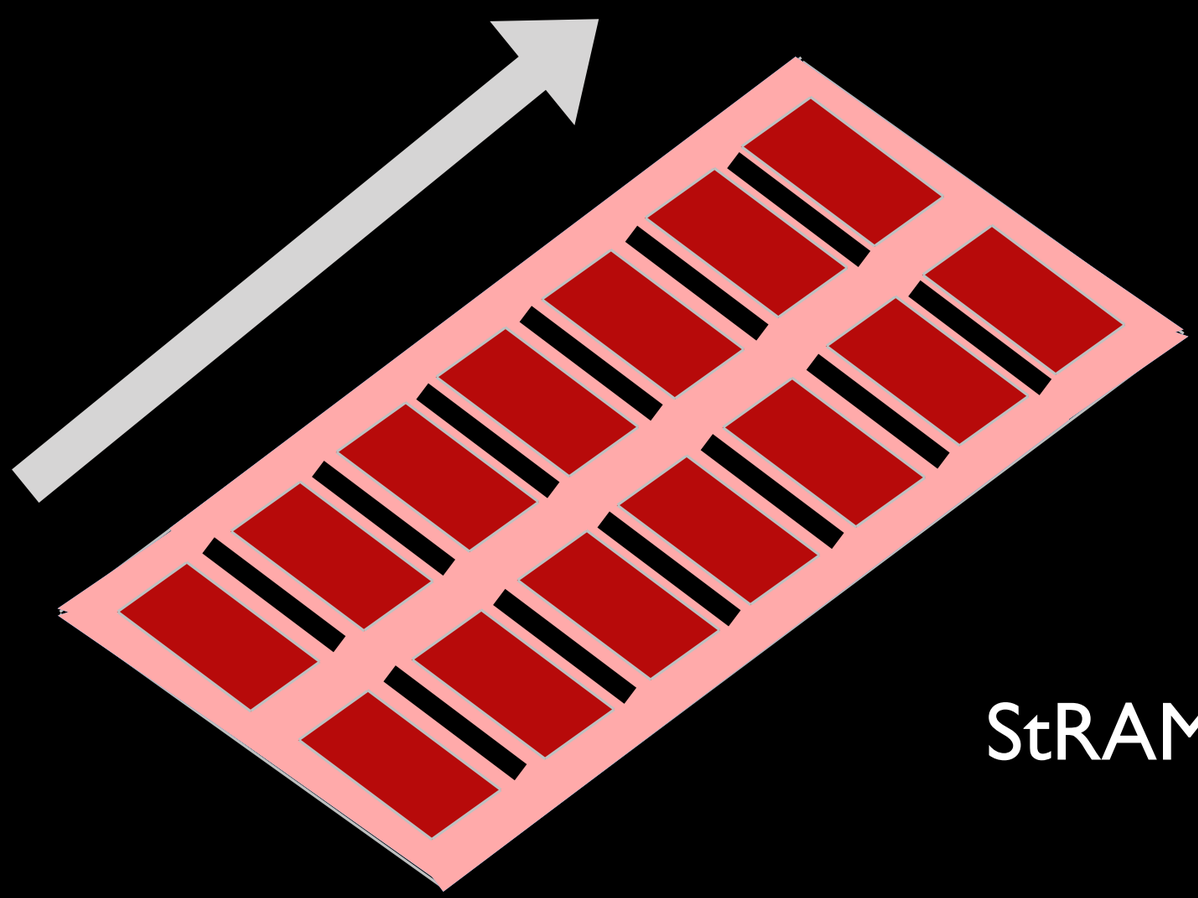
Inference Accelerator



Compute, layered on top

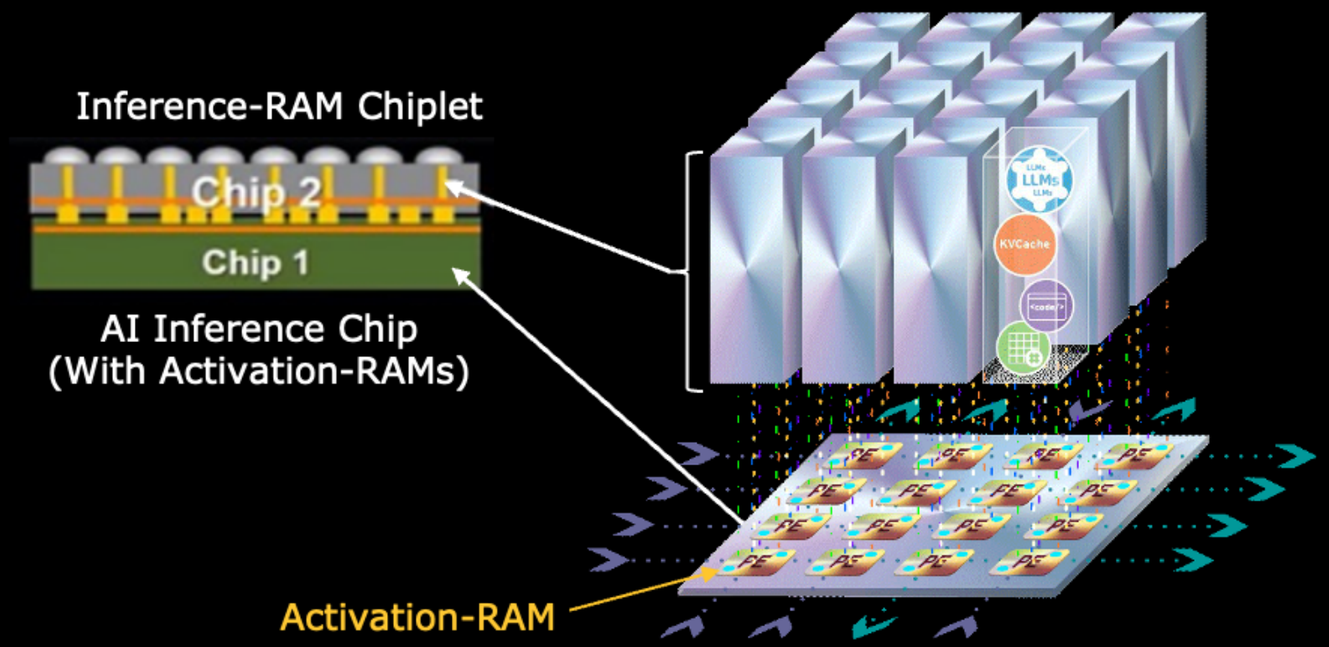
dense connections with LtRAM below

Inference Accelerator



StRAM for Activations

Collaboration with Industry (EMD)



- Read-Optimized LtRAM chiplet houses long-term intelligence, comprising
  - "Inference-RAM"
  - Plentiful model parameters
  - Plentiful KV caches
- Density-Optimized StRAM
  - "Activation-RAM"
  - Dual-ported on-chip short-term memory with just-enough-retention for activation dataflow

Faculty Team

Philip Levis

Caroline Trippel

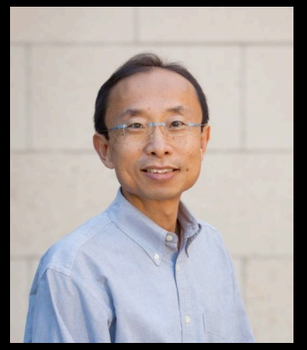
Chris Gregg

Mark Horowitz

Subhasish Mitra

Thierry Tambe

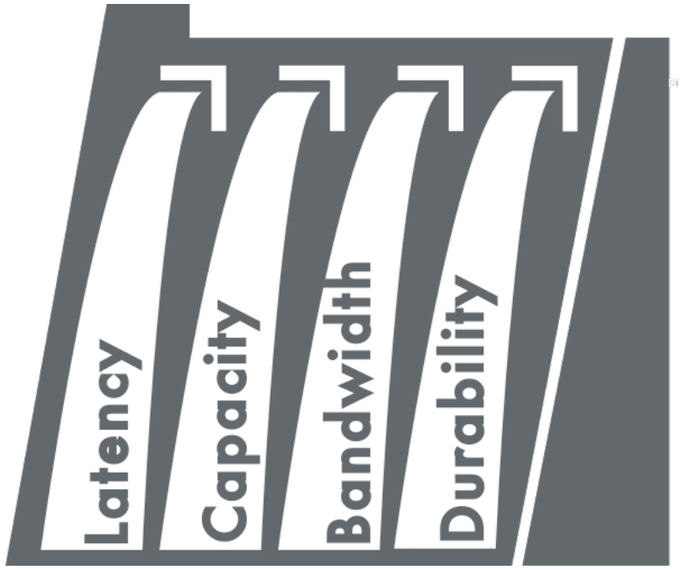
Keith Winstein



H.-S. Philip Wong

Mary Wootters

What *should* our next steps be?



MemoryDAX

Differentiated Access Memories

Summer Retreat
Monterey, Early July 2025