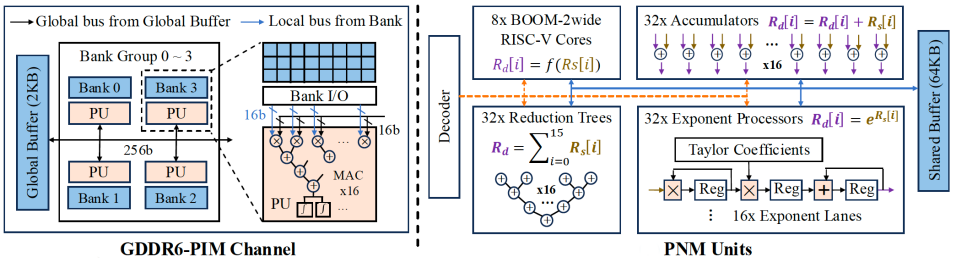
INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

Data Reuse-Aware Scalable Processing-In-Memory Architecture for Efficient Large Language Model Inference (DREAM)

TECHNOLOGY NUMBER: 2025-630

Accelerate Blue Foundry - 2025 (Physical Sciences)



OVERVIEW

DREAM is an advanced system for running large language models that replaces costly graphics cards with a new type of smart memory chip—one that not only stores data but also processes it right where it lives to deliver faster and much more affordable AI services, especially for tasks involving huge amounts of text or long conversations.

DESCRIPTION

Running powerful language models—as used in smart chatbots or AI writing assistants—demands storing and managing huge volumes of information, quickly accessing previous parts of a conversation, and supporting large “memory windows” for better understanding. Traditional computers rely on GPUs, which are excellent for heavy calculations but struggle and become costly when the main challenge is moving and storing large amounts of data, not just crunching numbers. DREAM solves this bottleneck by putting small compute units directly into memory chips. This means information can be used and reused right where it's stored, instead of being constantly shifted around. DREAM is a scalable system by building a hierarchical in/near memory platform with a network of smart memory devices. This innovation not only makes the whole process faster and far less expensive, but also supports more advanced types of AI models and longer, more complex prompts than older solutions.

VALUE PROPOSITION

Technology ID
2025-630

Category
Hardware
Engineering & Physical Sciences
Accelerate Blue Foundry - 2025/Physical Sciences

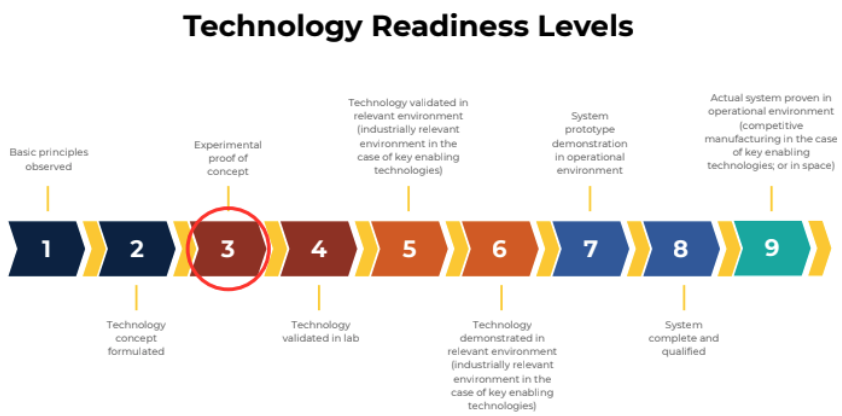
Inventor
Reetuparna Das

Further information
Joohee Kim
jooheek@umich.edu

View online

- **Massive Speed & Cost Efficiency**: Achieves 4.7× faster inference and 11.8× more tokens per dollar versus state-of-the-art GPU servers like DGX H100, directly impacting service costs.
- **Scalability for Advanced LLMs**: Accommodates huge LLMs and context windows (up to 1 million tokens), enabling advanced applications previously limited by memory bottlenecks and context size.
- **Optimization Beyond GPUs**: Eliminates GPU underutilization by matching the memory-focused nature of LLM workloads, while also supporting sophisticated model types (e.g., Mixture of Experts) better than existing PIM (processing-in-memory) designs.

TECHNOLOGY READINESS LEVEL



INTELLECTUAL PROPERTY STATUS

Patent applications pending.

MARKET OPPORTUNITY

As LLM applications proliferate—ranging from generative text/audio, coding assistants, and knowledge agents to real-time chatbots—the need for scalable, economical infrastructure grows rapidly. DREAM's approach of GPU-free, memory-centric AI inference directly addresses the cost and power inefficiencies blocking current LLM deployment for hyperscalers, cloud AI providers, and enterprise data centers, especially for models with long context requirements or flexible model compositions like MoE. Additionally, industries like finance, healthcare, and media creation that depend on high-throughput, low-latency AI stand to benefit from this efficient architecture. Trends show that demand for LLM-powered services is outstripping current GPU supply and drives infrastructure costs higher every quarter, as supported by public cloud pricing and AI hardware sales trends.