sambanova

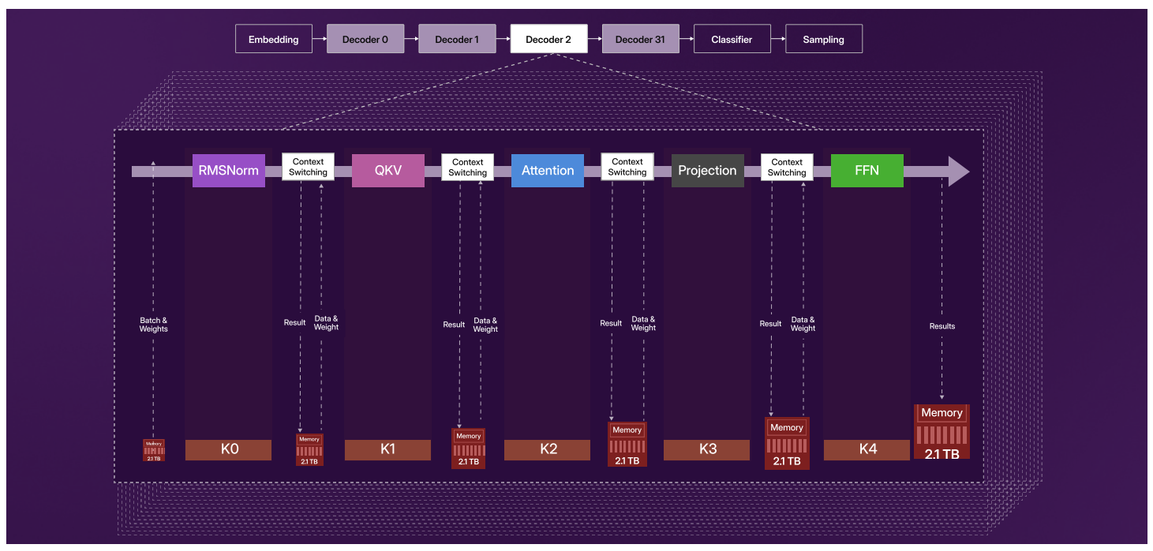
SambaNova SN40 Reconfigurable Dataflow Unit

Fast, energy-efficient inference at scale.

- Purpose-built for inference
- Dataflow Architecture
- Three-tier memory
- Efficient power profile

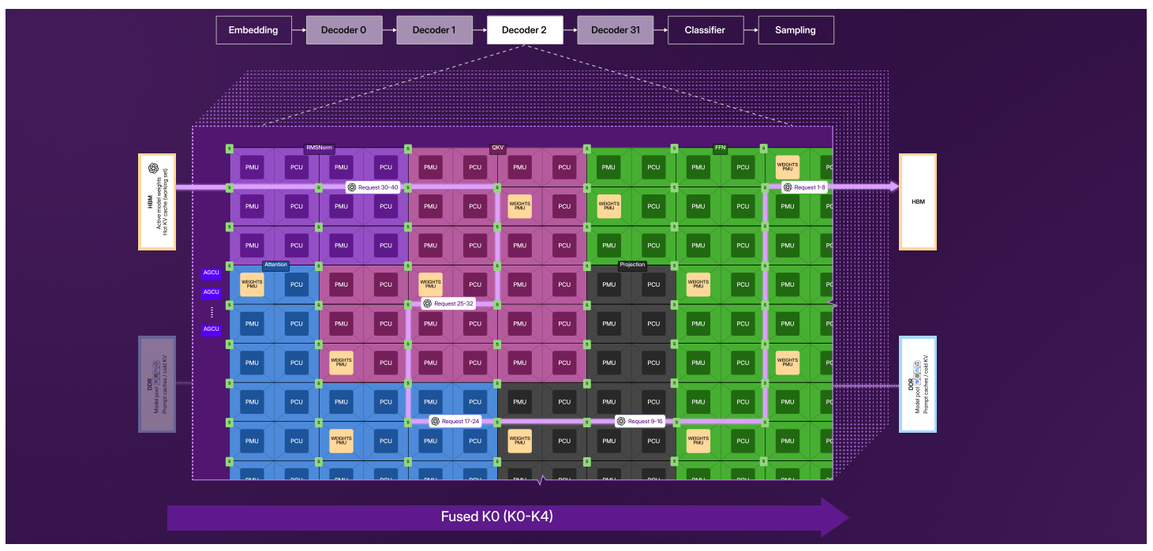
Problem Visual: AI's Data Movement Problem

Moving data back and forth between compute and memory is one of the biggest costs in inference. The SN40 is built to keep data moving forward — from one operation to the next.



16-GPU: Kernel-by-Kernel Execution (Bottleneck)

Two primary innovations that enable this are its dataflow architecture and a large, three-tier memory design.



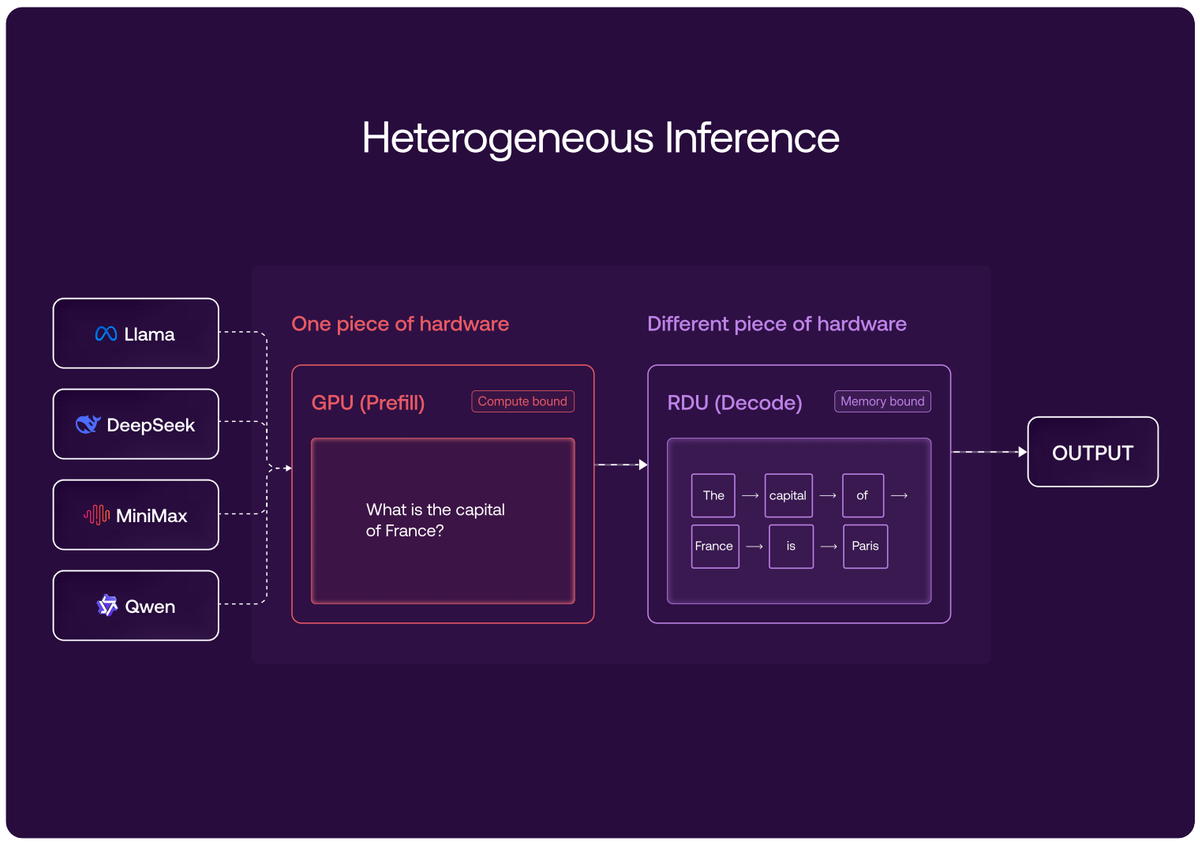
16-RDU: Spatial Execution (Zero Overhead)

Two primary innovations that enable this are its dataflow architecture and a large, three-tier memory design.

Purpose-Built For Inference

Once a model is trained, the infrastructure challenge shifts from building the model to serving it quickly and efficiently. Inference workloads are latency-sensitive, memory-intensive, and unpredictable. Each request needs fast access to model weights, active context, and the data required to generate the next token.

SN40 is purpose-built for this stage of AI. Its dataflow architecture keeps inference data moving through the system, helping reduce unnecessary memory movement and support fast, efficient token generation.

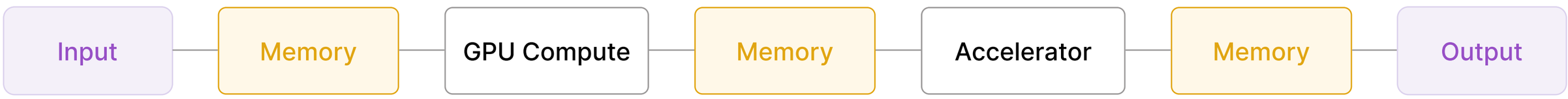


SN40L is purpose-built to solve AI's data-movement problem, delivering fast inference at scale.

Combining Many Operations Into A Single, Seamless Flow

Traditional accelerators call back to memory between operations. The SN40 chains operations into one continuous dataflow eliminating the overhead of relaunching work at every step.

TRADITIONAL EXECUTION — DETOUR THROUGH MEMORY AT EVERY STEP

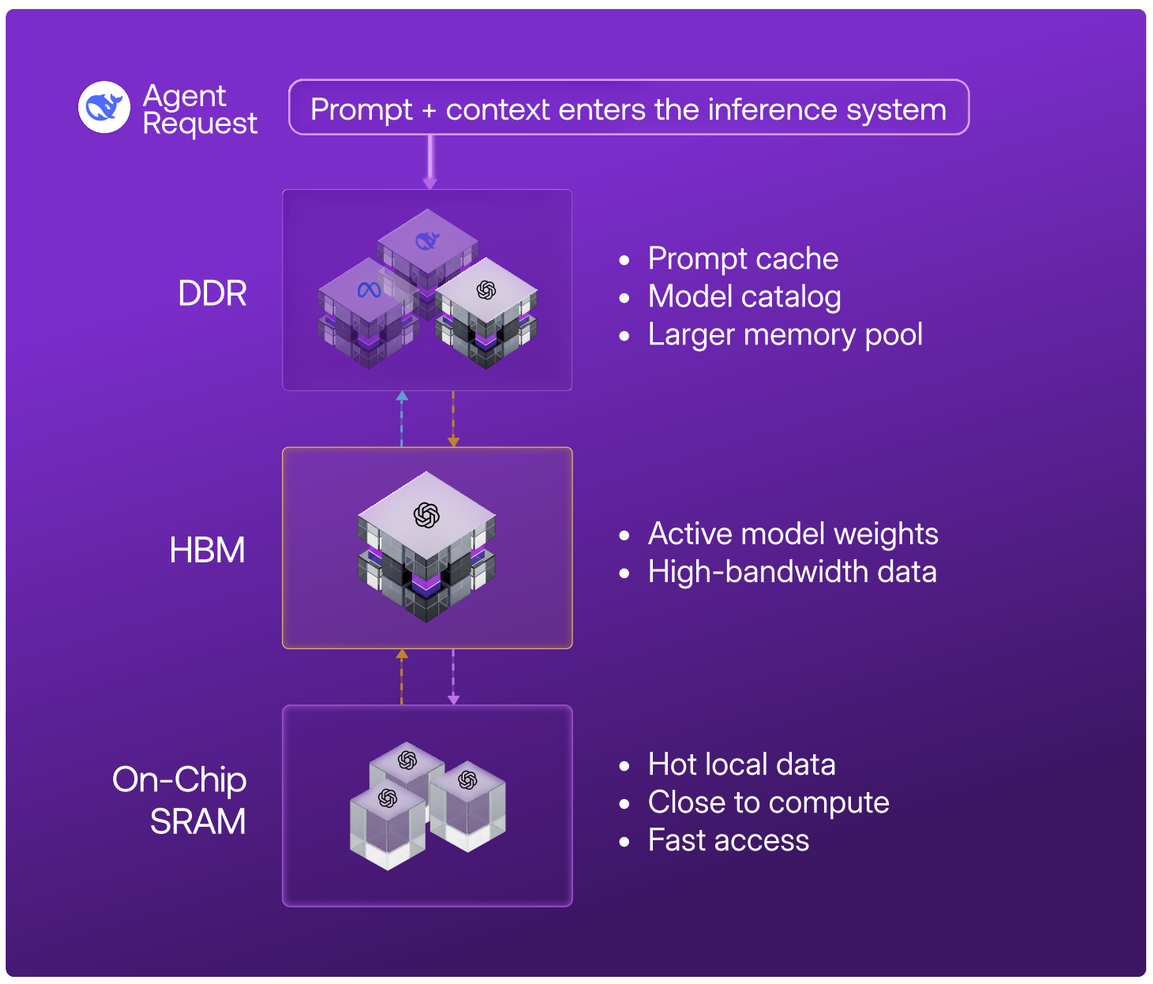


- More data movement
- More overhead
- Higher latency

SN40 DATAFLOW — OPERATIONS CHAINED IN A SINGLE KERNEL

- Seamless dataflow
- Less memory movement
- Faster inference
- Better energy efficiency

THREE-TIER MEMORY HIERARCHY



Model data stays close to compute

Three tiers keep model data near compute supporting large models and near-instant switching between many of them.

RACK-LEVEL POWER & COOLING

10 kW

Average power per SambaRack

SN40-powered SambaRack systems are designed for high-performance inference in existing air-cooled data centers, without exotic liquid-cooling infrastructure.

Fast inference

Energy-efficient operation

Low latency, high throughput

Air-cooled deployment

SN40 delivers fast, energy-efficient inference at scale.

Schedule a Meeting → www.sambanova.ai/contact

- Linkedin.com/company/sambanova
- @SambaNovaAI
- info@sambanova.ai