Datasheet

sambanova

SambaNova SN50 Reconfigurable Dataflow Unit

Purpose-built for fast decode and premium AI inference.

Purpose-built for fast decode

Optimized for fast token generation.

Dataflow architecture

Keeps inference execution moving continuously.

Three-tier memory

Keeps the right data close to compute.

Fifth-generation RDU

Built for higher performance and efficiency.

PREMIUM INFERENCE

Premium inference for large intelligent models and agents

SN50 combines fast token generation with support for large intelligent models, delivering premium inference for demanding user experiences and agentic workloads.

Large intelligent models

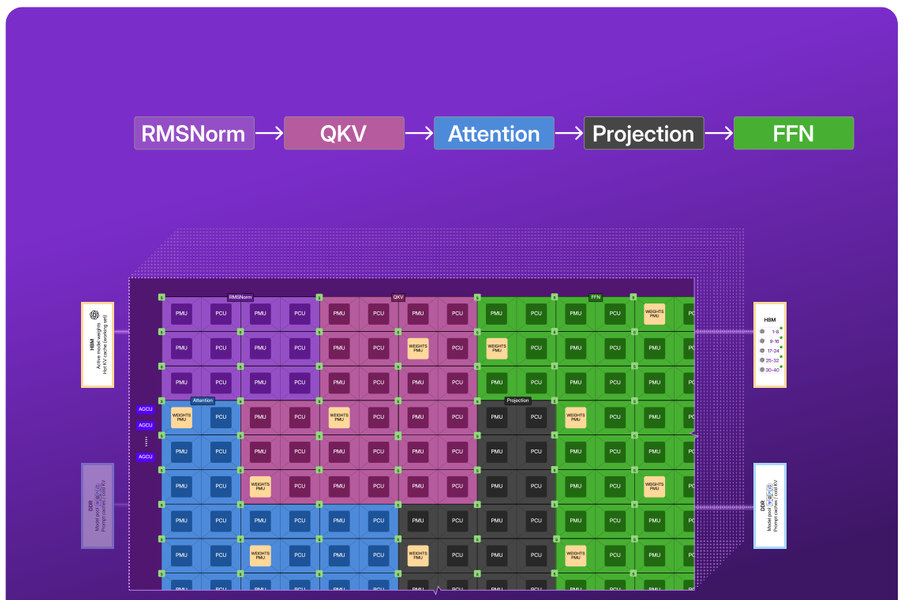
Fast token generation on the largest models

Interactive AI agents

Fast token generation for agentic workloads.

Purpose-built for fast decode

SN50 is designed for the memory-bound phase of inference. Its reconfigurable dataflow architecture keeps execution moving continuously through the processor, reducing unnecessary data movement between operations.
Combined with a three-tier memory hierarchy, SN50 keeps active model weights and inference data close to compute, supporting fast token generation and sustained throughput for coding and agentic workloads.



SN50 RDU dataflow execution

SN50 is purpose-built for fast decode, combining low latency, high throughput, and efficient inference.

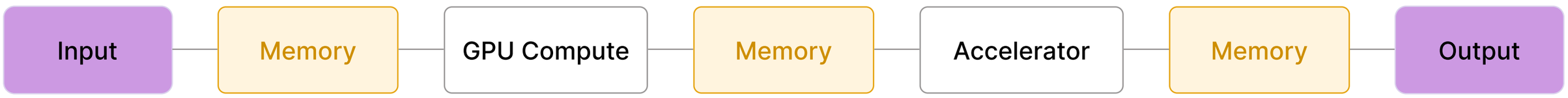
Schedule a Meeting → www.sambanova.ai/contact

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

Combining Operations Into One Continuous Dataflow

Traditional accelerators frequently return to memory between operations. SN50 chains operations into a continuous dataflow, reducing repeated memory movement and the overhead of relaunching work at every stage.

TRADITIONAL EXECUTION — DETOUR THROUGH MEMORY AT EVERY STEP



- More data movement
- More overhead
- Higher latency

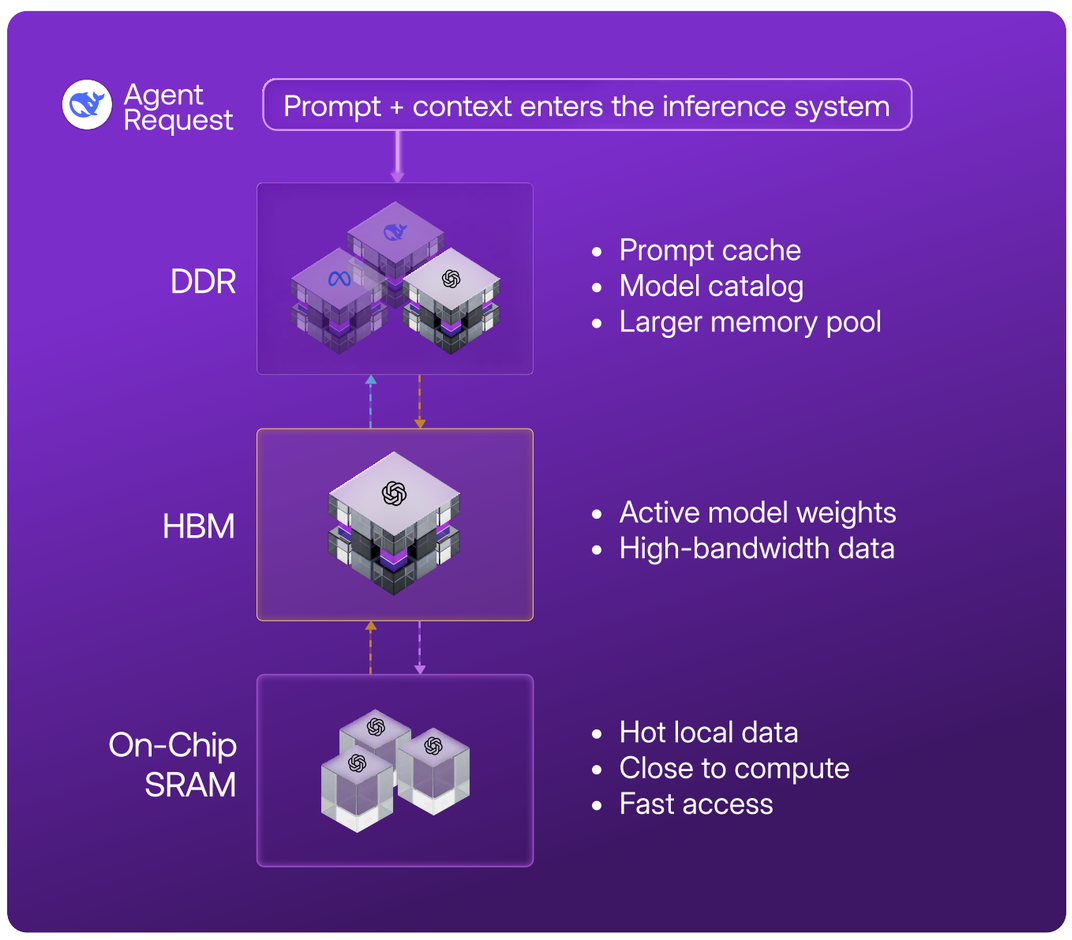
SN50 DATAFLOW — OPERATIONS CHAINED IN A SINGLE KERNEL

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

THREE-TIER MEMORY HIERARCHY

Keep the right data close to compute

SN50 uses three levels of memory to place data according to how quickly and frequently it needs to be accessed.



SN50 RDU Specifications

Process node	5 nm TSMC
Architecture	Reconfigurable Dataflow Architecture
Memory architecture	Three-tier memory: SRAM, HBM2E, and DDR5
On-chip SRAM	432 MB
HBM2E	64 GB
DDR5	Up to 2 TB

Fast decode

Fast token generation for premium inference.

Strong throughput

Supports concurrent users and demanding workloads.

Efficient data movement

Reduces repeated movement between compute and memory.

Built for agentic workloads

Designed for coding and other complex inference applications.

Schedule a Meeting → www.sambanova.ai/contact • Linkedin.com/company/sambanova • @SambaNovaAI • info@sambanova.ai