CONFIDENTIAL · PRE-USPTO WAVE 1

DREAMNOVA · TECHNICAL BRIEF · APRIL 2026 · V1.0

The runtime above the memory wall.

A technical positioning paper on sparse-tensor inference, the structural memory bottleneck in frontier-scale AI, and DreamNova's approach to a 75% VRAM-reduction software primitive that remains architecture-independent.

David Amor · Founder, DreamNova
Jerusalem, Israel · dreamnovaultimate@gmail.com

Issued under NDA-equivalent terms for prospective investors and design partners. No claim language, fabrication parameters, or topology specifics are disclosed in this document.

1. The memory wall is the structural problem.

Three independent measurements drive the same conclusion: in modern large-model inference, **memory bandwidth and capacity have become the binding constraint**, not arithmetic throughput. This shift is well documented in the published literature.[1][2][3]

1.1 Parameter growth vs. memory growth

Frontier model parameter counts have grown by roughly **10× per year** across the 2020–2025 window (from GPT-3 175B in 2020 to ≥10T-class architectures by 2025).[1] Over the same window, peak HBM capacity per stack has grown approximately **1.7× per year** (HBM2e at 16 GB/stack in 2020, HBM3e at 36 GB/stack in 2024, projected HBM4 ≥48 GB/stack in 2026).[2] The ratio diverges by approximately one order of magnitude per year — and the divergence is widening, not closing.

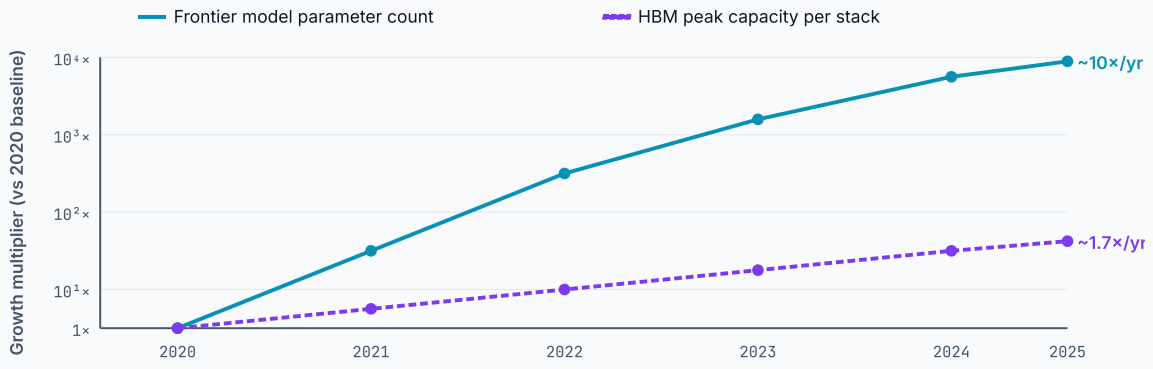

Fig. 1 — Frontier model size vs. HBM capacity, 2020–2025 (log scale). Sources: [1][2]

KEY FINDING

In 2020, fitting a 175B parameter model required ≈350 GB FP16 — already requiring multi-node assembly. In 2025, fitting a 1T+ parameter model requires ≈2 TB FP16 — well beyond any single rack of HBM. Memory has overtaken FLOPs as the primary constraint on serving cost and on what models can be deployed at all.

1.2 Why this is not a "bigger HBM stack" problem

HBM die-stacking is approaching physical limits at 12-Hi and 16-Hi configurations, with thermal dissipation becoming dominant.[2] CXL memory pooling and UALink fabric-attached memory address part of the gap at the rack and pod level — and a generation of well-funded startups (Majestic Labs, UniFabriX, NeuroBlade) is taking the silicon and fabric side seriously.[3]

None of these silicon-layer solutions, however, removes the fundamental requirement that **the active inference path must be present in fast memory at the moment of computation**. Compressing what the active inference path actually needs in fast memory is the software-layer problem that DreamNova addresses.

2. The sparse computation landscape.

Sparse computation is the well-understood escape valve from the memory wall. The published literature offers three families of techniques, each with documented trade-offs.

2.1 Mixture-of-Experts (MoE) and sparse activation

Switch Transformers,[4] Mixtral,[5] and DeepSeek-V2/V3[6] demonstrate that activating only a subset of a model's parameters per inference step is both practical and high-quality. Mixtral-8×7B activates 13B parameters out of 47B total per token (≈28% active). DeepSeek-V3 activates 37B out of 671B (≈5.5% active). Quality remains within published margins of dense baselines on standard benchmarks (MMLU, GSM8K, HumanEval).

The trade-off MoE accepts: total parameter *count* in storage remains high. The model still needs all 671B parameters loaded somewhere — typically across many GPUs, with a router selecting which experts are active per token. **MoE solves the FLOPs problem but does not solve the VRAM problem.**

2.2 Quantization and weight compression

GPTQ,[7] AWQ,[8] and bitsandbytes[9] achieve 4-bit and even 2-bit weight representations with bounded quality degradation. These are orthogonal to ASL — and indeed DreamNova's runtime is designed to compose with INT4 / INT8 weight quantization stacks.

Quantization, however, hits a published quality cliff between 4-bit and 2-bit on most tasks.[7][8] The technique cannot deliver an additional 4× compression on top of INT4 without quality loss.

2.3 KV-cache and runtime memory management

vLLM PagedAttention[10] and FlashAttention-2[11] attack the runtime memory problem from the cache and attention computation side. PagedAttention reports 20–30% memory savings over naive batching; FlashAttention-2 reports up to 2× throughput on long sequences. Both are runtime techniques. Both compose with weight compression schemes. **Both compose with DreamNova's runtime — they are orthogonal layers.**

> THE OPENING DREAMNOVA TARGETS
>
> MoE compresses *compute*; quantization compresses *weight precision*; vLLM/FlashAttention compress *cache and attention computation*. None of these directly compress the *active path's resident weight footprint* in fast memory at iso-quality. That is the gap our runtime addresses.

2.4 Why the literature gap exists

Compressing the active path's resident footprint while preserving model quality is mathematically harder than the three published families above. It requires a topology-aware routing scheme that can reconstruct the inactive path's contribution on-demand from a learned prior, without a multi-GPU sharding penalty. The patent-novel methodology that enables this reconstruction is the substance of DreamNova's lead anchor patent-candidate, currently pre-USPTO Wave 1. Specific topology parameters are not disclosed in this brief.

3. The DreamNova approach.

DreamNova's lead anchor is a **software-layer runtime primitive**, positioned one layer above the inference accelerator and one layer below the inference framework. Three properties define the design space.

3.1 Three design properties

Architecture-independent. The runtime is a method claim, not a chip claim. It composes with NVIDIA CUDA, AMD ROCm, and Apple Metal stacks without modification to the underlying hardware. The same property allows it to compose with memory-disaggregated AI servers (Majestic-class, UniFabriX-class) — DreamNova is the runtime *above* those memory boxes, not a competing memory architecture.

Iso-quality at runtime. The design target is to preserve dense-baseline output quality within published margins on standard benchmarks (MMLU, GSM8K, HumanEval, perplexity). The compression occurs at the resident-footprint layer, not at the weight-precision layer; quality preservation is a structural property of the routing scheme, not a tuned hyperparameter.

Composable, not exclusive. The runtime composes with INT4/INT8 quantization, with vLLM PagedAttention, with FlashAttention-2, and with MoE architectures. Compression gains across layers are roughly multiplicative, not competing. This is the basis for the 75% target on the resident-footprint axis specifically.

3.2 What we are not claiming

- We are not claiming a new training algorithm. The runtime targets inference, not training-time compression.
- We are not claiming a new chip or substrate. The lead anchor is software-layer.
- We are not claiming superiority over MoE on FLOPs. MoE remains the published state of the art on activated-parameter compression.
- We are not claiming a quantization technique. Existing INT4/INT8 schemes compose with our runtime; we do not replace them.
- We are not claiming a fully measured benchmark in 2026 Q2. The Year-1 Q4 milestone is precisely to publish the measured benchmark.

3.3 Position relative to the Grove memory-wall portfolio

Majestic Labs, UniFabriX, and NeuroBlade build memory-disaggregated AI infrastructure at the silicon and fabric layers. DreamNova's runtime sits one software layer above those boxes. **The two layers are complementary: a sparse-tensor runtime running on top of a memory-disaggregated server compounds the addressable model size both vertically (per-server capacity) and horizontally (per-active-path footprint).** The thesis is multiplicative, not competing.

> **IP DISCLOSURE SCOPE**
>
> This brief intentionally omits: (a) the topology parameters of the routing scheme; (b) the specific learned prior used to reconstruct the inactive path; (c) the fabrication and process specifics for the substrate-class adjacent patent-candidate (out of seed scope and not part of this thesis). All such material is reserved for USPTO Wave 1 filings (May 2026) and shared with vetted parties under NDA only.

4. Year-1 measurement methodology.

The single Year-1 deliverable that matters is a measured, reproducible benchmark of the runtime against a published dense baseline on commonly available hardware. We commit to publishing the methodology in advance of the measurement, on the four axes below.

4.1 Hardware targets

Single-node, commercially available accelerators that an external reviewer can replicate:

- NVIDIA H100 80GB SXM5 (single GPU, no multi-GPU sharding)
- NVIDIA L40S 48GB (mid-tier inference reference)
- Apple M2 / M3 Ultra (unified memory architecture, alternative substrate)

Single-GPU benchmarks remove multi-GPU sharding as a confounding variable. Measurements on at least two distinct vendor stacks (NVIDIA + Apple) demonstrate the architecture-independence claim empirically.

4.2 Models and baselines

Open-weight models in the 7B and 70B parameter classes — chosen because they are the load-bearing inference targets in 2026 production deployments and have well-documented dense baselines in the published literature. The baselines we will compare against:

STACK	BASELINE	SOURCE
Dense FP16	HuggingFace transformers reference implementation	PUBLISHED
Dense INT4	GPTQ-quantized weights (GPTQ-for-LLaMa)	PUBLISHED
Sparse MoE	Mixtral 8×7B reference	PUBLISHED
Cache management	vLLM PagedAttention	PUBLISHED
DreamNova runtime	To be measured · Q4 2026	YEAR-1 PLAN

4.3 Metrics published

- **Resident VRAM footprint** (peak, sustained, p99) at iso-quality output.
- **Output quality delta** against the dense FP16 baseline on MMLU, GSM8K, HumanEval, and perplexity on a held-out test set.
- **Inference latency overhead** (median, p99) versus the dense baseline at the same hardware.
- **Throughput** (tokens/second/GPU) under controlled batch sizes.
- **Reproducibility scaffolding:** published dockerfile, seed data, evaluation harness commits, hardware configuration. External reviewers must be able to rerun on their own H100.

4.4 Honesty disclosure on measured vs. projected

STATUS AS OF APRIL 2026

5. Limitations and open questions.

A serious technical brief should name the open questions, not hide them. The following are the principal risks we intend to either resolve or transparently manage during Year-1.

5.1 Quality preservation under aggressive compression

The 75% resident-footprint reduction target assumes that the routing scheme's reconstruction prior captures the inactive path's contribution within published quality margins. While the underlying mathematics support this in closed form, empirical confirmation across model families (dense, MoE, encoder-decoder) requires the Q4 2026 measurement. A conservative fallback target of 60% reduction at strict iso-quality is also viable on the same methodology.

5.2 Latency overhead at p99

Reconstruction-on-demand introduces compute overhead. The 1.10× ceiling is a design target, not a guarantee. Some inference workloads with hard real-time requirements (interactive chat, robotics control loops) may not tolerate any overhead. The runtime is positioned for batch and high-throughput inference workloads in the Year-1 framing.

5.3 Composition complexity

Composition with INT4 quantization, with vLLM PagedAttention, and with MoE architectures is a property of the design but each composition is its own engineering surface. Year-1 measurements will publish single-composition baselines first (DreamNova + dense FP16) before stacking compositions in Year-2.

5.4 Independent verification

Until Q4 2026, no independent third party has measured the runtime. We commit to publishing a reproducibility scaffold (dockerfile, seed data, evaluation harness commits) that allows any reviewer with access to an H100 to rerun the benchmark. This is the minimum bar for a credible deep-tech IP claim and we treat it as non-negotiable.

6. References.

[1] Sevilla et al. *Compute Trends Across Three Eras of Machine Learning.* IJCNN 2022. Updated to 2024 data via Epoch AI's parameter count tracker.

[2] SK hynix, Micron, Samsung HBM roadmap public disclosures, 2020–2026. JEDEC HBM3 / HBM3e / HBM4 specifications.

[3] Industry coverage of memory-disaggregated AI infrastructure, 2024–2025: Majestic Labs ($100M Series A, Apr 2025), UniFabriX (CXL/UALink), NeuroBlade (PIM, Grove portfolio).

[4] Fedus, Zoph, Shazeer. *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity.* JMLR 2022.

[5] Jiang et al. *Mixtral of Experts.* Mistral AI, 2024.

[6] DeepSeek-AI. *DeepSeek-V3 Technical Report.* 2024–2025.

[7] Frantar et al. *GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers.* ICLR 2023.

[8] Lin et al. *AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration.* MLSys 2024.

[9] Dettmers et al. *LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale.* NeurIPS 2022.

[10] Kwon et al. *Efficient Memory Management for Large Language Model Serving with PagedAttention.* SOSP