3SWAI Smarter. Faster. Greener.

White Paper: NASWAI Execution Layer

Summary

Today's AI execution is limited by hardware-specific runtimes, rigid data structures, and fixed computational flows. Our innovation reimagines the entire execution layer — introducing a groundbreaking software stack that connects AI models to hardware in a highly optimized, hardware-agnostic manner.

This architecture delivers:

- Multi-fold performance gains on standard CPUs and AI accelerators

- Significant reductions in memory usage and bandwidth requirements

- Advanced quantization methods achieving better-than-FP16 precision

- Dramatically improved energy efficiency and scalability

A provisional patent has been filed; full patent issuance is in progress.

Innovation Overview

Our execution layer is built on a fundamental rethinking of how neural computations are performed — from basic multiplications to data layout and memory flow.

Core Innovations

- Redesigned Operations:
  Standard tensor multiplications are replaced with custom routines. Activation functions are embedded directly into multiplications, effectively eliminating their runtime cost.

- Proprietary Data Types:
  Internally, the system utilizes bitfield-like custom types that enable:
  - Adaptive precision control at every chokepoint in the data pipeline
  - Seamless casting between types
  - Parallel execution of multiple operations within scalar and vector units

- Matrix Reshaping for Memory Optimization:
  By reorganizing execution (while remaining mathematically equivalent), we achieve:
  - One input/output vector access per 128–1024 weight accesses
  - Quantization-friendly weight grouping with high compression rates
  - ~6× smaller memory footprint and ~10× lower memory bandwidth consumption

3SWAI Smarter. Faster. Greener.

- Advanced Quantization Techniques:
  Our variable-length quantization retains large "superweights" with zero loss relative to FP32, while aggressively compressing smaller weights:
  - Precision surpasses FP16/BF16
  - Enables sub-2-bit effective quantization with high accuracy
  - Zero decompression overhead during execution

- Hardware-Level Optimization:
  On standard x86-64 CPUs, our system delivers:
  - Over 64 mathematical operations per cycle per thread — dramatically outperforming conventional runtimes

Performance Benchmarks

We benchmarked our solution against leading GPU and AI platforms using a 32G model. Results show competitive — often superior — performance, especially in memory efficiency and power scaling.

Applications & Use Cases

Our execution layer is applicable across a wide range of demanding AI environments:

- Scalable inference for LLMs and Transformers

- Edge and embedded AI where power and memory are limited

- Enterprise AI with full-stack customization needs

- Cloud-native inference acceleration

- High-context LLMs with modest compute requirements

- Enterprise knowledge integration

- Private, offline, or air-gapped deployments

- Next-gen abstraction layer for silicon partners

Benefits of Large Memory Capacity

This system unlocks new AI use cases that were previously impractical due to hardware constraints — particularly where contextual depth and memory capacity outweigh raw throughput.

High-Context LLMs with Modest Compute

GPU clusters (e.g., 16× B200s) offer fast inference but demand massive infrastructure. Our system allows models the size of ChatGPT-5 to run on a single CPU by:

- Compressing memory to enable terabytes of effective RAM

- Preserving FP32-level accuracy through quantization

Ideal for low-concurrency, high-context tasks:

- Engineering design assistance

- Software architecture and planning

- Long-form editorial or strategic writing

Enterprise Knowledge Integration

Most enterprise AI systems are constrained by token limits. Our architecture enables loading entire corpora into live context, such as:

- Technical manuals and documentation

- Inventory and part databases

- Regulatory compliance libraries

- Customer-specific configurations

- Historical project insights

This enables persistent, high-context LLMs to operate as fully informed internal agents, improving:

- Design consistency

- Cross-team collaboration

- Speed of onboarding and decision-making

Private & Air-Gapped Deployments

Our system runs on off-the-shelf hardware, making it ideal for:

- On-premises secure environments

- Air-gapped networks

- Highly regulated industries (defense, healthcare, infrastructure)

Applicable scenarios:

- Internal assistants with sensitive data access

- Autonomous tagging and classification of documents

- Secure R&D intelligence

- AI-enhanced interfaces for legacy systems

3SWAI Smarter. Faster. Greener.

Intellectual Property

- A provisional patent is filed; full issuance is pending.

- Patent covers:
  - Execution logic and architectural design
  - Custom data representations
  - Compression and quantization methodologies

- Additional technical details are held as trade secrets or under further patent filings.

Roadmap

We are finalizing performance optimizations and preparing for a phased product rollout.

Phase 1 — Now

- Core engine prototype complete

- Execution layer and data model validated internally

- Ongoing tuning and benchmarking

Phase 2 — December 2025

- First Viable Product (FVP)

- Launch of a large-scale Mixture of Experts (MoE) model running at RTX 5090 speed on a laptop — leveraging our breakthrough memory and execution efficiencies

Phase 3A — 2026

- Universal Quantizer

- Tooling to automatically convert third-party AI models into our optimized format

Phase 3B — 2026

- GPU Platform Expansion

- Rollout to NVIDIA/AMD platforms, replicating our CPU-level breakthroughs in GPU environments; initial GPU vendor selection driven by early partner demand

Partnering & Investment Opportunity

We are not offering an incremental upgrade — we are building a new category of AI infrastructure.

Our software execution layer transforms general-purpose hardware into high-performance AI accelerators, with capabilities rivaling or surpassing top-tier GPUs — at a fraction of the power and cost. This opens the door to a new generation of AI applications across edge, enterprise, cloud, and custom silicon.

3SWAI Smarter. Faster. Greener.

For Strategic Partners and Early Adopters

We are currently onboarding select partners to:
• Integrate the execution layer into enterprise AI stacks
• Deploy high-context LLMs without cloud dependency or GPU clusters
• Develop domain-specific, secure, or regulated use cases
• Validate and optimize real-world workloads

Early adopters benefit from:
• Exclusive access to upcoming releases
• Hands-on engineering support and integration assistance
• Co-authorship in case studies and public benchmarks
• Influence over roadmap priorities

For Investors

We are positioned at the intersection of three massive trends:
1. AI inference demand scaling faster than hardware availability
2. Enterprise and government need for sovereign, secure AI infrastructure
3. The search for post-GPU alternatives that scale across edge and datacenter

We offer:
• Defensible IP (patent-pending architecture and trade secrets)
• Breakthrough benchmarks vs industry leaders on cost, performance-per-watt, and memory scaling
• Clear monetization pathways — from enterprise licensing to OEM integrations and developer tooling

What We’re Building Together

With the right partners, we can define the next decade of AI execution — one where models are not bound by memory, context, or platform limitations.

We invite forward-thinking partners and investors to join us in bringing this infrastructure shift to the global AI landscape.

📩 Contact us: hello@3swai.com