Executive Summary: Neural Forest: Energy-Efficient AGI Infrastructure for Petabyte-Scale Social AI

The Track: Deployment Scaling & Infrastructure

The Problem: Meta's massive AI/ML footprint, specifically within Kubernetes-orchestrated environments, currently faces a "Crisis of the Monolith" where centralized, deep networks (like Llama) drive unsustainable energy intensity and memory waste. Landmark 2026 data indicates that AI/ML pods often waste 6.1x the memory they actually consume, with standard industry practices requiring 4x memory padding to avoid the "OOMKilled" (Out of Memory) state.

The Solution: The Neural Forest (NF) Paradigm The Neural Forest replaces monolithic parameter blocks with a distributed ecosystem of thousands of **Neural Trees**—shallow, task-specific neural networks (MLPs, CNNs, or RNNs). This shift from brute-force scaling to "forced specialization" allows Meta to dismantle the traditional "Memory Wall".

- **Dynamic Resource Pruning:** Managed by a meta-cognitive **Conductor**, the system routes tasks only to relevant specialized modules. This "Inference-First" routing enables the transition from a 16.4 GiB average request to a ~2.7 GiB actual need by only activating the specific silicon required for the query.
- **NF-Core Accelerator:** To solve the hardware heat-density bottlenecks of current H200/B200 clusters, the NF-Core utilizes distributed on-chip memory (SRAM/eDRAM). This places weights directly adjacent to micro-cores, eliminating energy-intensive DRAM trips and providing a sustainable path for the social graph (TAO).
- **Audit-Ready Reporting (The "Enterprise Audit Shield"):** The NF stack, implemented via FastAPI and Kubernetes, provides verifiable, mechanistic interpretability logs. This enables per-request energy intensity reporting, ensuring Llama-integrated environments (Instagram/WhatsApp) remain compliant with new EU and US AI transparency and sustainability acts.

Deployment Roadmap

1. **Software Integration:** Deploy the Neural Forest as an orchestration layer within Meta's existing Kubernetes infrastructure to provide immediate OOMKilled prevention without the need for traditional memory padding.
2. **Silicon Transition:** Move toward integrating NF-specific logic into future chip revisions (targeting Intel 18A or custom Meta silicon) to achieve 10x-50x gains in inference

efficiency.

Conclusion By adopting the Neural Forest, Meta can lead the industry's transition from brute-force training to sustainable, distributed inference-at-the-edge, meeting its 2030 Net Zero goals while securing its infrastructure against the "Memory Wall".