AETHER COMPUTE GROUP

Distributed AI Inferencing — Technical Whitepaper

Architecture, scheduling, and operational model of Aether Compute Group's planetary-scale inference fabric.

Region / Region	Global
Document Type	Technical Whitepaper
Version	2026.Q3 · rev-4
Classification	Public — Distribution Permitted
Contact	compliance@aethercompute.io

Executive Summary

Aether Compute Group operates a three-tier distributed inference fabric comprising a 470 MW core IDC footprint, sovereign edge regions, and a hotel-based micro-edge network across 40+ countries. This document formalizes the platform's technical architecture and the regional compliance controls applicable to workloads served from the Global jurisdiction.

1. Compute Fabric Topology

The platform is organized as three concentric tiers. Tier-0 Core IDC aggregates 470 MW of liquid-cooled H100/H200/B200 capacity across purpose-built campuses. Tier-1 Sovereign Edge operates 42 regional sites averaging 6–18 MW each, deployed inside licensed carrier-neutral facilities. Tier-2 Hotel Edge comprises 3,800+ hotel-hosted micro-nodes providing last-mile inference within 40 ms of 92% of global population centers.

Workloads are placed by the Aether Scheduler, a latency- and residency-aware planner that evaluates legal jurisdiction, token throughput headroom, carbon intensity, and request affinity on every dispatch. Placement decisions are auditable and exportable as signed JSON traces.

2. Inference Runtime

The runtime supports vLLM, TensorRT-LLM, SGLang, and TGI. Speculative decoding, paged KV-cache, and continuous batching are enabled by default. Multi-tenant isolation is enforced via SR-IOV and MIG partitioning; per-tenant TEE (Intel TDX / NVIDIA CC) is available on request.

Cold-start for 70B-class models is under 9 seconds cluster-wide via pre-warmed weight tiering. P50 first-token latency is < 40 ms intra-region, < 120 ms cross-continent.

3. Cross-Border Data Handling

All customer traffic terminates within the region of origin. Model weights are cached in a residency-scoped tier; user prompts and completions never traverse a border unless the customer explicitly enables cross-region overflow. Overflow, when enabled, is subject to per-request policy tokens and produces a signed audit record.

4. SLA Framework

Standard tier: 99.9% availability, credit schedule as per MSA §7. Sovereign tier: 99.95%. Dedicated capacity: 99.99% with per-second token-throughput SLO.

Latency SLOs are measured at the ingress edge and validated by an independent third-party synthetic monitor operated from 220 vantage points.