AIL2 ARCHITECTURE

Decentralized AI Operating System

Document Version: 1.0.0

Status: Technical Review

Abstract:

This document outlines the core architectural components of the AIL2 network, focusing on the consensus mechanism, state channel interactions, and the specialized node topology required to support high-frequency decentralized inference.

1. Architectural Overview

The AIL2 architecture is built on a modular "Layer 2" framework designed to offload heavy compute tasks from the main consensus chain. It consists of three primary planes:

1. The Settlement Plane (L1 Bridge): Handles finality, staking, and dispute resolution.
2. The Coordination Plane (AIL2 Core): Manages node discovery, task routing, and state channels.
3. The Compute Plane (GPU Grid): The physical layer where inference and training actually occur.

2. Inference Workflow Diagram

```
+----------------+        +------------------+        +-----------------+
|  Client / DApp | -----> |  Broker Node     | -----> |  State Channel  |
+----------------+  (1)   +------------------+  (2)   +-----------------+
        ^                        |                            |
        | (Result)               | (Task Route)               | (Lock State)
        |                        v                            v
+----------------+        +------------------+        +-----------------+
|  Verification  | <----- |  Compute Node    | <----- |  GPU Cluster    |
|    Oracle      |  (4)   |  (Miner)         |  (3)   |  (Inference)    |
+----------------+        +------------------+        +-----------------+
```

(1) Request Submission: Client submits an inference request with a bounty.

(2) Channel Opening: A temporary state channel is opened between Client and Broker.

(3) Execution: The task is routed to the optimal Compute Node (Miner) based on latency and model availability.

(4) Verification: The result is returned with a SNARK proof (Proof of Inference) to the Oracle for verification before settlement.

3. Dynamic State Channels

To achieve "near-zero latency" for multi-agent systems, AIL2 utilizes ephemeral state channels (ESCs). unlike traditional state channels which are static, ESCs can dynamically re-route to different agents without closing the main on-chain connection.

Features:

- Instant Finality: Off-chain signatures allow agents to trust responses immediately.
- Liquidity Hubs: Nodes act as hubs, allowing distinct agents to pay each other without direct channels.
- Privacy Preserving: Interactions within the channel are encrypted and not visible on the public ledger until settlement.

4. Network Topology

The network is composed of heterogeneous nodes, each specialized for specific functions:

A. Validator Nodes:

- High-availability servers.
- Responsibilities: Block production, ZK-Proof verification, Challenging dishonest miners.

B. Compute Miners (GPU Nodes):

- High-throughput hardware (H100/A100 clusters).
- Responsibilities: executing the 'Universal Model' containers and generating local proofs.

C. Agent Runners:

- Lightweight nodes.
- Responsibilities: Hosting the logic for autonomous agents (LLM wrappers) that consume the compute resources.