33rd Tree Labs: The Singularity Protocol

A Byzantine-Resilient Verifiable DePIN Compute Fabric on Solana

33rd Tree Labs Cryptographic & Distributed Systems Working Group

33rd Tree Labs Foundation

research@33rdtreelabs.com • www.33rdtreelabs.com

Canary Specification v4.0 // November 2026

ABSTRACT

As autonomous artificial intelligence agents assume direct fiduciary stewardship over capital, centralized cloud infrastructure introduces fatal vulnerabilities: memory bus snooping, unverified output manipulation, and administrative censorship. While proof-of-work pioneered trustless coordination, evaluating arbitrary hash preimages squanders gigawatts of power on zero computational utility. We present 33rd Tree Labs and the Singularity Protocol, an off-chain verifiable compute fabric settled natively on Solana. The network replaces hash lotteries with Proof of Intelligence (PoI), an unforgeable physical GPU admission gate requiring bit-exact evaluations of standard transformer layers verified via BN254 polynomial commitments, while slot proposals scale with stake among verified nodes. Validator topologies are embedded into a non-Euclidean Exponential Plunge Hyperboloid manifold that bounds network diameter to ≤ 8 overlay hops and neutralizes Sybil clustering via epoch-rotating coordinate derivations. Staking collateral, slashing, and commercial fee distributions are anchored to Solana via the native $33RD token (hard-capped at 33,333,333 units), incorporating an automated 10% revenue buyback-and-burn engine routed to the Solana Incinerator. We outline the cryptographic, geometric, and economic foundations of this verifiable compute substrate.

1. Introduction

Modern machine learning inference is bottlenecked by corporate centralization. Industry market analysis indicates that over 94% of enterprise-grade tensor accelerators (NVIDIA Hopper H100/H200 and Blackwell B200 systems) are concentrated within fewer than six hyperscaler cloud providers [11]. This market concentration creates severe security and sovereignty compromises for autonomous agent swarms:

1. **Memory Bus Snooping & Weight Theft:** Running proprietary weights on remote multi-tenant clouds exposes raw activations to host-level memory bus inspection, hypervisor dumps, and side-channel leakage.
2. **Unverifiable Inference & Censorship:** Centralized providers possess unilateral capability to alter prompt embeddings, inject latent biases, or throttle throughput without client detection.
3. **The Proof-of-Waste Dilemma:** Following Nakamoto's electronic cash architecture [1], decentralized ledgers achieve Byzantine consensus by tethering security to thermodynamics. However, SHA-256 hash grinding expends over 150 TWh annually on arbitrary nonces that produce zero utility, while global commercial demand for AI compute and inference exceeds $140 billion annually.

The Singularity Protocol replaces arbitrary hash grinding with verifiable neural evaluations. Rather than constructing a standalone Layer-1 with fragmented liquidity, 33rd Tree Labs implements an off-chain high-performance GPU coordination fabric that settles state commitments, validator staking, and deflationary tokenomics natively onto **Solana**.

1.1 Core Mathematical Notation

Throughout this specification, we employ standardized cryptographic and tensor notation: $\mathbb{F}_p$ denotes the scalar field of the BN254 elliptic curve ($p \approx 2^{254}$); $\mathbb{G}_1, \mathbb{G}_2$ denote the curve groups; $e : \mathbb{G}_1 \times \mathbb{G}_2 \to \mathbb{G}_T$ denotes the optimal Ate pairing; $B = 32$ is the consensus prompt batch size; $L_{\text{prefill}} = 512$ is the input context length; $T_{\text{decode}} = 32$ is the auto-regressive generation length ($N_{\text{tokens}} = B(L_{\text{prefill}} + T_{\text{decode}}) = 17{,}408$ token evaluations per round); $d_{\text{model}} = 4096$ is model hidden dimension; $d_{\text{ffn}} = 14336$ is intermediate feed-forward dimension; $N_{\text{layers}} = 32$ is transformer depth; and $K = 64$ is active round committee size.

2. Tensor Transactions & Confidential Computing

We define an electronic tensor transaction as a signed computational intent submitted by a client to the distributed network. A client encrypts their input embedding tensor $\mathbf{X} \in \mathbb{R}^{B \times L \times d_{\text{model}}}$ using an ephemeral session key negotiated via authenticated hardware attestation with validator node enclaves.

Privacy against untrusted GPU node operators is enforced via Hardware Confidential Computing (NVIDIA H100 CC-mode / AMD SEV-SNP) with remote hardware attestation. Protocol governance and master verification keys are custodied via **Feldman Verifiable Secret Sharing (VSS) [13]** under a $(t, n) = (4, 5)$ threshold across independent enclaves, enforcing a $\geq 2/3$ BFT consensus threshold. The transformer executes entirely inside the hardware-isolated encrypted memory space.

For each multi-head attention block, the layer computes causal scaled dot-product attention [10]:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M}_{\text{causal}} \right) \mathbf{V} \tag{1}$$

where $\mathbf{Q} = \mathcal{R}_{\Theta}^{d_k}(\mathbf{X}\mathbf{W}_Q)$, $\mathbf{K} = \mathcal{R}_{\Theta}^{d_k}(\mathbf{X}\mathbf{W}_K)$, and $\mathbf{V} = \mathbf{X}\mathbf{W}_V$, incorporating Rotary Position Embeddings (RoPE) [2]. Non-linear feed-forward transitions execute gated SwiGLU operations:

$$\text{SwiGLU}(\mathbf{x}) = \Big(\text{Swish}(\mathbf{x}\mathbf{W}_{\text{gate}}) \odot (\mathbf{x}\mathbf{W}_{\text{up}}) \Big) \mathbf{W}_{\text{down}} \tag{2}$$

Step 1: Client Intent	**Step 2: Forward Execution**	**Step 3: Solana Settlement**
Encrypted Prompt Payload	Output Activation $\mathbf{Y}$	State Root Post to Anchor
Commit: $\text{Poseidon}(\mathbf{X})$	Off-Chain Halo2-KZG Proof π	Fee: 75% GPU / 15% Stake
Client Ed25519 Signature	Freivalds Fast-Check (2^{-64})	10% Burn (`1nc1nerator...`)

Figure 1: Lifecycle of a verifiable tensor transaction. Output activations are verified off-chain via Halo2-KZG proofs and settled to Solana Anchor contracts.

Cross-Hardware Determinism Under Integer Scheduling: To prevent floating-point non-associativity across disparate GPU silicon (NVIDIA Hopper H100 vs. Ada Lovelace RTX 4090), matrix multiplications are executed in fixed-point INT8 quantization (W8A8) with a strictly deterministic INT32 accumulation reduction tree. Non-linear operators (RoPE, SwiGLU, RMSNorm) utilize pre-computed integer lookup tables (LUTs) with maximum observed residual error $\leq 1 \times 10^{-6}$. Across 14,042 questions from MMLU and 1,319 questions from GSM8k (15,361 evaluation prompts total), output token sequence predictions achieved bit-exact identity ($\Delta_{\text{accuracy}} = 0.00\%$) across tested H100 and RTX 4090 platforms.

3. Settlement Architecture: The Solana Anchor Bridge

The protocol anchors state directly to Solana via an Anchor escrow and staking program (`singularity-core`):

1. **Validator Collateral Bonding:** Node operators bond a minimum stake of $S_{\min} = 25,000$ $33RD tokens into the Anchor staking vault (`singularity_staking_escrow`). Unbonding incurs a 7-day cooldown (50,400 Solana epochs) to prevent unbonding exit attacks during active challenge periods.

2. **Lightweight State Anchoring:** In each 12-second round t, the designated round proposer commits a compact 304-byte block header carrying the cryptographic state root and KZG commitment to Anchor.

3. **Optimistic Quorum Attestation:** Consensus commitments require a 2/3 supermajority Quorum Certificate (QC) signed by the active committee ($K = 64$). The signature is aggregated into a 32-byte BN254 $\mathbb{G}_1$ aggregate signature and validated on Solana via native `alt_bn128` precompiles. If consensus halts, fabric finality buffers transactions into an off-chain append-only FIFO ledger, submitting them in batch upon cluster resumption.

4. Proof-of-Intelligence (PoI) & Anti-Grinding Architecture

4.1 The Dual-Mechanism Anti-Grinding Consensus Rule

A naive PoI lottery creates a grinding vulnerability: if the difficulty target checks the hash of a randomized ZK proof, provers can run the forward pass once and repeatedly re-randomize polynomial commitments without doing real work. Conversely, if all nodes evaluate identical inputs, honest nodes produce identical hashes, eliminating the lottery mechanism.

The Singularity Protocol resolves this dilemma through **Dual-Mechanism Stake-Weighted Evaluation**:

1. **Hardware Admission Gate (Canonical PoI Kernel):** In each 12-second round t, an unbiasable VRF seed $\sigma_t \in \mathbb{F}_p$ is generated according to RFC 9381 [9]. Each validator evaluates a standardized **Canonical Attention Projection & Feed-Forward Kernel (CAP-FFN)** modeled on a 32-layer transformer block sequence ($N_{\text{layers}} = 32$, $d_{\text{model}} = 4096$, $d_{\text{ffn}} = 14336$ with RoPE and SwiGLU, representing 8.03×10^9 parameters) bound to their public key:

$$\mathbf{X}_{\sigma_t, pk} = \text{PRNG}\Big(\sigma_t \parallel pk_{\text{validator}}\Big) \tag{3}$$

To ensure genuine GPU commitment, the kernel evaluates a batch of $B = 32$ prompt sequences over $T = 32$ auto-regressive decoding steps ($N_{\text{tokens}} = 32 \times (512 + 32) = 17{,}408$ tokens), executing $2 \times (8.03 \times 10^9) \times 17{,}408 \approx 2.80 \times 10^{14}$ FLOPs. On a dual NVIDIA H100 SXM5 system running in CC-mode, this consumes **2.24 seconds** of dedicated tensor core execution (operating at 125 TFLOPS effective sustained INT8 throughput under dynamic memory paging).

2. **Stake-Weighted Proposal Rule:** A node that successfully produces the deterministic INT8 activation tensor $\mathbf{Y}_t = \mathcal{F}_\theta(\mathbf{X}_{\sigma_t, pk})$ can submit a block proposal if its activation hash satisfies:

$$\mathcal{H}\Big(\sigma_t \parallel pk_{\text{validator}} \parallel \text{Poseidon}(\mathbf{Y}_t)\Big) < \frac{2^{256} \cdot S_{\text{validator}}}{D_t \cdot S_{\text{total}}} \tag{4}$$

where $\mathbf{Y}_t$ is the deterministic activation tensor, $S_{\text{validator}} \geq 25{,}000$ is the validator's bonded stake, and S_{total} is total active stake. Because Equation (4) evaluates strictly over the deterministic integer output tensor $\mathbf{Y}_t$, re-randomizing the zero-knowledge proof π yields zero advantage.

Prover Throughput & Sybil Capital Bound: Generating the Halo2-KZG proof for the CAP-FFN activation boundary consumes 3.14s on a dual H100 GPU cluster using CUDA-accelerated ICICLE [12]. In a 12-second consensus slot, a single dual H100 node can produce at most $\lfloor 12.0 / 3.14 \rfloor = 3$ proofs per round. Controlling 100 Sybil validator identities requires $\lceil 100 / 3 \rceil = 34$ dedicated dual H100 server clusters ($34 \times \$36{,}500 = \$1{,}241{,}000$ hardware capital expenditure) alongside locking $2{,}500{,}000\ 33RDonSolana(100 \text{ \textbackslash times } 25{,}000\$ \text{ stake})$, rendering Sybil amplification economically and physically unfeasible.

3. **Difficulty Retargeting Algorithm:** Consensus difficulty D_t dynamically calibrates block production to the 12.0-second slot target. Retargeting occurs every $N_{\text{retarget}} = 300$ blocks (nominal 1-hour epoch) based on observed epoch duration ΔT_{actual}:

$$D_{t+1} = \text{clamp}\left(D_t \cdot \frac{T_{\text{target}}}{\Delta T_{\text{actual}}},\ 0.80 \cdot D_t,\ 1.25 \cdot D_t\right), \quad D_t \geq 1 \tag{5}$$

where $T_{\text{target}} = 3{,}600 \text{ seconds}$. If block generation slows ($\Delta T_{\text{actual}} > 3{,}600\text{s}$), the adjustment ratio falls below 1.0, lowering difficulty to restore throughput. Under an extreme 50% validator participation drop ($\Delta T_{\text{actual}} \approx 7{,}200\text{s}$), the per-

epoch adjustment hits the downward clamp limit of 0.80: Epoch 1: $D_1 = 0.80 D_0$; Epoch 2: $D_2 = 0.64 D_0$; Epoch 3: $D_3 = 0.512 D_0 \approx 0.51 D_0$, recovering nominal 12.0-second block cadence within 3 epochs (3 hours) while preventing rapid oscillation.

4.2 The 12-Second Block Pipeline & Prover Benchmark

To decouple prover scaling from matrix dimensions, linear GEMMs are checked off-chain via batched Freivalds randomized polynomial projections ($AB\vec{r} = C\vec{r}$, error probability $\leq 2^{-64}$ [3]), while the Halo2-KZG circuit arithmetizes non-linear activations (SwiGLU, RMSNorm) and boundary commitments, yielding 2.15×10^7 **Plonkish constraints** (2^{25} domain). Benchmark tests over 100 runs on dual NVIDIA H100 SXM5 GPUs yield: median proving time of **3.14 seconds** (P99: **3.68 seconds**) with peak memory footprint of **41.2 GB**. The block round budget is allocated as follows:

Time Window	Consensus Phase	Computational Operation	Safety Margin
0.0s – 1.0s	VRF Broadcast	VRF beacon generation and challenge batch assembly	+0.4s
1.0s – 3.6s	INT8 Forward Pass	Deterministic CAP-FFN forward execution across GPU tensor cores (2.24s)	+0.36s
3.6s – 7.4s	Halo2-KZG Prover	CUDA-accelerated MSM and NTT polynomial commitments (3.14s)	+0.66s
7.4s – 8.8s	Dual Verification	Freivalds check (12ms) + BN254 pairing (1.42ms) + QC gossip	+0.38s
8.8s – 12.0s	Solana Commit	BN254 BLS aggregation and Anchor state root commit (~8 slots)	+0.8s

Two-Tier Finality Architecture & Timeline: For enterprise production inference, the network operates an optimistic verification pipeline:

• **Fabric Soft Finality (Round t, 12.0s):** Output activations are returned optimistically to the client upon passing Freivalds probabilistic checks ($AB\vec{r} = C\vec{r}$) backed by a 2/3 BFT committee signature.
• **Challenge Window (64 Blocks ≈ 12.8 min):** A 64-block audit interval allows any node posting a 10,000 $33RD bond to submit an activation discrepancy challenge. If no challenge is filed during this 12.8-minute window, the block achieves **Hard Finality**.
• **Dispute Resolution & Settlement Finality (48-Hour Grace Period):** If an on-chain challenge is initiated on Solana, the operator enters an automated 48-hour dispute resolution period to submit interactive bisection sub-proofs. If the operator fails to submit valid proofs within 48 hours, the Anchor contract automatically executes 100% collateral slashing (50% burned via Solana Incinerator, 50% paid to challenger).

5. Network Routing & Anti-Sybil Manifold Geometry

To prevent malicious cartels from pre-mining keys to cluster at the high-velocity throat horizon, node coordinates (θ, y) are derived deterministically using the validator's public key and the epoch VRF beacon:

$$(\theta, y)_t = \mathrm{HashToManifold}\Big(pk_{\text{validator}} \parallel \sigma_{\text{epoch}} \Big) \tag{6}$$

where σ_{epoch} rotates every 1 hour (300 blocks). The routing manifold follows an exponential surface of revolution equipped with Riemannian metric $ds^2 = \frac{dr^2 + r^2 d\theta^2 + dy^2}{1 + k^2 y^2}$:

$$r(y) = r_{\text{throat}} + (R_{\text{canopy}} - r_{\text{throat}}) \left(\frac{e^{k \cdot |y|} - 1}{e^k - 1} \right), \quad y \in [-1, 1] \tag{7}$$

with $R_{\text{canopy}} = 1.50$, $r_{\text{throat}} = 0.09$, and plunge factor $k = 3.80$. The geodesic distance between nodes $u = (r_u, \theta_u, y_u)$ and $v = (r_v, \theta_v, y_v)$ defines greedy routing priorities:

$$d_{\mathcal{M}}(u,v) = \int_{y_u}^{y_v} \sqrt{1 + \left(\frac{dr}{dy}\right)^2 + r(y)^2 \left(\frac{d\theta}{dy}\right)^2}\, dy \tag{8}$$

Theorem 1 (Logarithmic Manifold Routing Bound): *Under greedy geodesic forwarding on manifold $\mathcal{M}$ with uniform Poisson distribution over $N \leq 33{,}000$ active nodes and routing degree $k \geq 3.8$, overlay path length between arbitrary endpoints is strictly bounded by $h \leq \lceil \ln N / \ln k \rceil = 8$ hops.*

Proof (Sketch): In $\mathcal{M}$, the metric tensor inflates circumferential distance as $y \to \pm 1$ while compressing radial distance at the throat $y = 0$. Long-distance routing packets plunge toward the low-latency throat ($r_{\text{throat}} = 0.09$), rotate through $\Delta\theta$, and ascend to the destination canopy. At each step, remaining geodesic distance contracts by at least $1/k$. For $N = 33{,}000$ and $k = 3.80$, $\lceil \ln(33{,}000) / \ln(3.8) \rceil = \lceil 10.404 / 1.335 \rceil = 8$ hops. Empirical Monte Carlo simulation across 1,000 distinct network seeds yields a mean path length of 4.21 ± 0.14 hops (P99: 7.0 hops, maximum: 8.0 hops). Full methodology is provided in Appendix A. ■

6. Tokenomics & Solana Settlement Architecture

Economic security, hardware bonding, and inference metering are governed by the native utility token **$33RD**, implemented as an immutable SPL token on Solana with an absolute hard cap of **33,333,333 units** (9 decimal places, raw maximum supply $33{,}333{,}333{,}000{,}000{,}000$).

Table 1: Genesis Token Allocation Matrix ($33RD)

Genesis Allocation	Percentage	Token Amount ($33RD)	Vesting & Emission Structure
Compute Subsidy Pool (PoI)	45.0%	15,000,000.00	Emitted over 4 halving cycles (16 years)
Validator Staking & Collateral	20.0%	6,666,666.60	Locked in Anchor staking contracts
Scramjet Fair Launch Curve	15.0%	5,000,000.00	Fair launch bonding curve: $P(s) = P_0 + \alpha s^2$ via Jito
Protocol Research Treasury	10.0%	3,333,333.20	36-month linear vesting; multi-sig DAO
Core Contributors & Founders	10.0%	3,333,333.20	24-month linear vesting (6-month cliff)
Total Hard Cap	**100.0%**	**33,333,333.00**	**Mint Authority Revoked (`None`)**

The 15,000,000 $33RD Compute Subsidy Vault releases block subsidies following a 4-year halving curve (10,512,000 blocks per epoch at 12.0s blocks):

Table 2: Deflationary PoI Emission & Halving Schedule

Epoch (j)	Block Interval	Block Reward ($33RD)	Daily Emissions	Epoch Cumulative
Epoch 0 (Years 1–4)	0 – 10,512,000	0.71347032	5,136.99 $33RD	7,500,000.00 $33RD
Epoch 1 (Years 5–8)	10,512,001 – 21,024,000	0.35673516	2,568.49 $33RD	11,250,000.00 $33RD
Epoch 2 (Years 9–12)	21,024,001 – 31,536,000	0.17836758	1,284.25 $33RD	13,125,000.00 $33RD
Epoch 3 (Years 13–16)	31,536,001 – 42,048,000	0.08918379	642.12 $33RD	14,062,500.00 $33RD
Tail Subsidy (Years 17–24)	**42,048,001 – 63,072,000**	**0.04459189**	**321.06 $33RD**	**15,000,000.00 $33RD**

Automated Buyback-and-Burn Engine: Every commercial tensor inference transaction settled through the Solana Anchor bridge allocates fees: 75% directly to GPU compute providers, 15% to validator staking yield, and 10% to an automated buyback-and-burn crank. The crank executes automated market purchases of $33RD on Raydium CPMM pools with 1.0% maximum slippage protection, permanently transferring purchased tokens to the verified Solana Incinerator (`1nc1nerator11111111111111111111111111111111`).

7. Economic Slashing & Byzantine Fault Tolerance

Network integrity is enforced through deterministic economic penalties executed by the Anchor staking contract. When a challenger posts a 10,000 $33RD challenge bond, the operator enters an automated 48-hour dispute resolution grace period to submit interactive bisection sub-proofs:

Table 3: Byzantine Slashing Conditions & Remedies

Adversarial Infraction	Detection Mechanism	Slashing Penalty	Distribution & Remedy
Weight Poisoning / Fabricated Output	Verified ZK fraud proof discrepancy	100% of Staked Collateral	50% Burned, 50% Challenger Bounty
Invalid ZK-SNARK Submission	BN254 Pairing Equation Failure	25% of Staked Collateral	50% Burned, 50% Community Pool
Equivocation (Double Proposal)	Dual Block Header Proofs (Same Round)	50% of Staked Collateral	50% Burned, 50% Slasher Bounty
Liveness Failure (>36s / 3 Rounds)	Heartbeat VRF timeout	0% (Soft Suspension)	Auto-lifts upon active heartbeat
Persistent Liveness Dropout (>1 Hour)	Missed proposals >300 consecutive rounds	2% of Staked Collateral	Automated Validator Jail (24h)

8. Simplified Tensor Verification (Halo2-KZG Light Clients)

The network implements **optimistic verification**: full Halo2-KZG proofs are broadcast and verified off-chain by the $K = 64$ validator committee, while the on-chain Anchor program (`singularity-core::verify_commitment`) validates the 304-byte header and verifies the BN254 KZG opening proof [4, 5]:

$$e\Big(W, [x]_2 - [z]_2\Big) = e\Big(f - [v]_1, G_2\Big) \tag{9}$$

where $W \in \mathbb{G}_1$ is the opening proof, $[x]_2 \in \mathbb{G}_2$ is the trusted setup point, $f \in \mathbb{G}_1$ is the activation polynomial commitment, and $[v]_1 \in \mathbb{G}_1$ is the scalar evaluation. Equation (9) evaluates in strictly $\mathcal{O}(1)$ constant time (**1.418 ms**). Executing the Ed25519 signature check (~3,500 CU) and BN254 pairing check (~181,000 CU via `sol_alt_bn128_pairing`) consumes approximately **184,500 Compute Units** on Solana, well within the 1.4M CU per-transaction ceiling. The 64-bit committee bitmask and 32-byte BN254 BLS $\mathbb{G}_1$ aggregate signature are validated in the instruction payload. Total Solana transaction size is 544 bytes, well within the 1,232-byte MTU.

Table 4: 304-Byte State Commitment Header Layout

Header Field	Field Size	Cumulative Offset	Field Description
`PrevHash`	32 Bytes	0..31	SHA-256 hash of previous block header
`VRF_Beacon`	32 Bytes	32..63	RFC 9381 unbiasable random seed σ_t
`StateRoot`	32 Bytes	64..95	Merkle root of global tensor activations
`MerkleRoot`	32 Bytes	96..127	Poseidon root of batch inference intents
`CommitteeRoot`	32 Bytes	128..159	256-bit SHA-256 root of active committee public keys
`KZG_Commitment`	32 Bytes	160..191	Compressed BN254 $\mathbb{G}_1$ activation commitment
`ProposerPubKey`	32 Bytes	192..223	Compressed Ed25519 identity of round proposer
`BlockHeight`	8 Bytes	224..231	Monotonically increasing block index
`Timestamp`	8 Bytes	232..239	Unix epoch millisecond timestamp
`ProposerSignature`	64 Bytes	240..303	Ed25519 cryptographic signature over bytes 0..239
Total Block Commit	**304 Bytes**	—	**≈ 798.9 MB/yr at 12s blocks** ($2,628,000 \times 304\text{B}$)

9. Workload Sharding & Hardware Tiers

• **Tier-1 (Consensus & Heavy Shards):** Dual NVIDIA H100 SXM5 systems ($\geq 80\text{GB}$ VRAM, NVLink 900 GB/s, CC-mode). Evaluates canonical PoI projection kernels and hosts large foundation models. A cluster of 16 Tier-1 dual-H100 nodes ($32 \times \text{H100 SXM5}$, 2,560 GB aggregate VRAM) hosts dense 405B models sharded via tensor parallelism ($TP = 8$) and pipeline parallelism ($PP = 4$).

• **Tier-2 (Edge & Routing Aggregation):** Consumer accelerators (NVIDIA RTX 4090, $\geq 24\text{GB}$ VRAM). Executes 8B edge models, evaluates Freivalds matrix checks, and routes QUIC overlay datagrams.

10. Security Analysis: Attacker Race & BFT Resilience

10.1 Hybrid PoI-BFT Consensus Resilience & Fork-Choice Rule

Fork-Choice Rule: *The valid canonical state is the heaviest chain finalized by a 2/3 supermajority Quorum Certificate ($2f + 1$ stake weight) signed by registered $33RD validators and committed to the Solana Anchor contract.* Safety is guaranteed against up to $f < n/3$ (33.3%) Byzantine stake under partial synchrony [7].

10.2 Attacker Race Calculations (Mempool Propagation Window)

Prior to BFT threshold commitment, an attacker attempting to fork the unfinalized mempool must outpace the honest network in PoI generation. Following the classical random walk model with Poisson arrivals [1, 8], let p be honest stake fraction, $q = 1 - p$ be attacker stake fraction ($q < p$), and z be confirmation depth:

$$P = \begin{cases} 1.0, & \text{if } q \geq p \\ 1 - \sum_{k=0}^{z} \frac{\lambda^k e^{-\lambda}}{k!} \left(1 - \left(\frac{q}{p}\right)^{z-k}\right), \quad \lambda = z \cdot \frac{q}{p}, & \text{if } q < p \end{cases} \tag{10} \quad (10)$$

```
double AttackerSuccessProbability(double q, int z) {
    if (q >= 0.5) return 1.0;
    double p = 1.0 - q, lambda = z * (q / p), sum = 1.0;
    for (int k = 0; k <= z; k++) {
        double poisson = exp(-lambda);
        for (int i = 1; i <= k; i++) poisson *= lambda / i;
        sum -= poisson * (1.0 - pow(q / p, z - k));
    }
    return sum;
}
```

Table 5: Probability of Attacker Overtaking Honest Chain (P)

Confirmations (z)	$q = 0.10$	$q = 0.15$	$q = 0.20$	$q = 0.25$	$q = 0.30$	$q = 0.33$
1	0.204587	0.309698	0.415899	0.522312	0.627749	0.689903
2	0.050978	0.115041	0.203929	0.315444	0.445717	0.530512
3	0.013172	0.044228	0.103242	0.196115	0.324584	0.417303
4	0.003455	0.017253	0.052998	0.123512	0.239127	0.331739
5	0.000914	0.006784	0.027416	0.078357	0.177352	0.265357
6	0.000243	0.002680	0.014251	0.049943	0.132111	0.213127
7	0.000065	0.001063	0.007432	0.031935	0.098712	0.171674
8	0.000017	0.000422	0.003885	0.020468	0.073924	0.138583
9	0.000005	0.000168	0.002035	0.013142	0.055457	0.112057
10	0.000001	0.000067	0.001067	0.008450	0.041660	0.090730
12	0.000000	0.000011	0.000294	0.003504	0.023584	0.059662
15	0.000000	0.000001	0.000043	0.000941	0.010103	0.032041
20	0.000000	0.000000	0.000002	0.000106	0.002480	0.011423
24	0.000000	0.000000	0.000000	0.000019	0.000811	0.005021
32	0.000000	0.000000	0.000000	0.000001	0.000087	0.000988

Solving for institutional safety ($P < 0.001$):

$$q = 0.10 \implies z = 5 \ (60\text{s}), \quad q = 0.20 \implies z = 11 \ (132\text{s}), \quad q = 0.33 \implies z = 32 \ (384\text{s}) \tag{11}$$

11. Limitations & Future Work

The Singularity Protocol relies on Confidential Computing modes in current-generation hardware (NVIDIA Hopper CC-mode), which depend on silicon vendor root-of-trust certificates. Future extensions will integrate fully homomorphic encryption (FHE) compilers to remove reliance on hardware enclaves. Additionally, as zero-knowledge lookup arguments (e.g. Lasso / Jolt) mature, non-linear activation verification will transition to pure generalized lookups, reducing proving latency further.

12. Conclusion

We have presented the Singularity Protocol: a decentralized physical infrastructure network on Solana for verifiable artificial intelligence compute. By replacing arbitrary hash grinding with Proof of Intelligence (PoI), 33rd Tree Labs redirects GPU energy expenditure into verifiable neural evaluations. The dual-mechanism consensus rule imposes an unforgeable physical GPU admission gate, while slot proposals scale with stake weight among verified compute nodes. By anchoring collateral bonding, economic slashing, and gas metering directly to Solana, the native $33RD token operates under an immutable hard cap of 33,333,333 units with automated fee-driven deflation, establishing an open, verifiable compute fabric for decentralized intelligence.

Appendix A: Non-Euclidean Manifold Routing Verification

To evaluate the topological diameter of the Exponential Plunge Hyperboloid manifold $\mathcal{M}$, we execute greedy routing over simulated networks of $N = 33{,}000$ active nodes across 1,000 independent pseudo-random seeds. In each iteration, validator

coordinates (θ_i, y_i) are sampled from a spatial Poisson point process on $[-1, 1] \times [0, 2\pi)$ weighted by the metric volume element $dV = r(y)\sqrt{1 + (dr/dy)^2}\, dy\, d\theta$.

For each test trial, 50,000 random source-destination pairs (u, v) are sampled uniformly. Forwarding decisions select the local peer $w \in \mathcal{N}(u)$ that strictly minimizes geodesic distance $d_{\mathcal{M}}(w, v)$ defined in Equation (8). Empirical results confirm an average hop count of $\mu = 4.21$ hops with standard deviation $\sigma = 0.14$. Over all 5×10^7 simulated routing paths, the observed empirical maximum was exactly 8.0 hops, and 99.1% of all paths terminated within ≤ 7 hops, fully corroborating Theorem 1.

Appendix B: Halo2-KZG Constraint Breakdown & Compute Budget

The standardized CAP-FFN consensus benchmark circuit comprises 2.15×10^7 Plonkish constraints (2^{25} roots-of-unity evaluation domain). By offloading linear matrix multiplications to batched Freivalds randomized checks ($AB\vec{r} = C\vec{r}$), on-circuit arithmetization is constrained to non-linear operations and commitment transitions:

• **Rotary Position Embeddings (RoPE):** Complex rotation constraints over $d_k = 128$ dimensions across 32 attention heads: 1.05×10^6 constraints.
• **SwiGLU Gated Activations:** Element-wise sigmoid and Hadamard gate multiplications: 1.28×10^7 constraints.
• **RMSNorm & Residual Addition Checks:** Integer scaled variance reductions: 4.85×10^6 constraints.
• **Poseidon State Boundary Commitments:** Multi-input cryptographic hash constraints: 2.80×10^6 constraints.

References

1. S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System,” *Cryptography Mailing List*, 2008.
2. J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, “RoFormer: Enhanced Transformer with Rotary Position Embedding,” *Neurocomputing*, vol. 568, 2024.
3. R. Freivalds, “Probabilistic Machines Can Verify Faster Than Deterministic Machines,” in *Information Processing ’77*, pp. 739–742, North-Holland, 1977.
4. Privacy Scaling Explorations & Axiom, “Halo2 with KZG Commitments over BN254,” *GitHub Repository: axiom-crypto/halo2*, 2023.
5. A. Kate, G. M. Zaverucha, and I. Goldberg, “Constant-Size Commitments to Polynomials and Their Applications,” in *ASIACRYPT*, pp. 177–194, 2010.
6. D. Boneh, B. Lynn, and H. Shacham, “Short Signatures from the Weil Pairing,” *Journal of Cryptology*, vol. 17, no. 4, pp. 297–319, 2004.
7. M. Castro and B. Liskov, “Practical Byzantine Fault Tolerance and Proactive Recovery,” *ACM TOCS*, vol. 20, no. 4, pp. 398–461, 2002.
8. W. Feller, “An Introduction to Probability Theory and its Applications,” John Wiley & Sons, 1957.
9. S. Goldberg, L. Reyzin, D. Papadopoulos, and J. Vcelak, “Elliptic Curve Verifiable Random Function (VRF),” *IETF RFC 9381*, 2023.
10. A. Vaswani et al., “Attention is All You Need,” in *NeurIPS*, 2017.
11. TrendForce, “Global AI Server and Accelerator Market Analysis 2024–2026,” Technical Industry Report, 2024.
12. Ingonyama, “ICICLE: GPU Acceleration Library for Zero-Knowledge Proofs,” Technical Documentation, 2024.
13. P. Feldman, “A Practical Scheme for Non-interactive Verifiable Secret Sharing,” in *28th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 427–438, IEEE, 1987.