Inferra Whitepaper

A non-custodial marketplace and exchange for GPU compute, on Solana

VERSION 1.0

Abstract

GPU time is bought blind. Prices are quoted per vendor, capacity is prepaid before anyone knows it works, buyers risk a provider that never delivers, and providers risk a buyer that never pays. There is no neutral price for an hour of an H100, and no way to hedge one.

Inferra adds two layers on Solana. The first is a non-custodial marketplace: buy real GPU time from verified providers, with payment locked in an on-chain escrow and released only after the machine is delivered. The second is an exchange: a live GPU-hour price index with perpetuals and futures, margined and settled in $INFERRA.

Both layers settle in one token. $INFERRA is the payment asset, the trading collateral, the provider bond, and the protocol fee sink. Real purchases give the index a real anchor, and trading the index moves the price the next buyer pays, so the market does its own price discovery.

1. How Inferra works

1.1 Buying compute

A purchase is four steps. Inferra never holds the buyer's funds, and a provider is never paid before delivery.

1. **Escrow.** `create_order` locks the buyer's $INFERRA in a vault tied to that single order. The protocol has no wallet that can move it.
2. **Deliver.** The provider serves the machine over SSH.
3. **Settle.** `confirm_delivery` is the only instruction that pays out: 98% to the provider, 2% to the protocol treasury. The buyer can sign it, or for verified providers a server arbiter signs once delivery is observed, so a normal purchase needs no second signature.
4. **Or refund.** If the machine never arrives or the deadline passes, the escrow refunds the buyer. Disputes settle on chain.

Every completed order updates the provider's on-chain record (orders delivered, volume served). Reputation is public and cannot be faked.

1.2 Delivery

Settlement is backed by a real machine. When an order settles, a node agent on the provider's box starts an isolated container with the GPU passed through, locked down (non-root, dropped capabilities, restricted egress) and reachable only with the buyer's SSH key. Usage is metered by GPU-hour and access is cut off when the purchased hours or the lease term run out. An on-chain payment maps to time-boxed access to real hardware, not a credit on a private ledger.

1.3 Providers

Supply comes from operators who put capital at risk. A provider bonds $INFERRA on chain to list capacity. Non-delivery and fraud are slashable against that bond, and delivery history accrues as on-chain reputation. A provider's stake and standing are only worth keeping by continuing to deliver.

2. $INFERRA in the ecosystem

$INFERRA is a fixed-supply SPL token on Solana: 1,000,000,000 units, mint authority revoked, so no more can ever be created. One token does four jobs:

Role	Function
Payment	Compute is bought and settled in $INFERRA through the escrow.
Collateral	The exchange is margined and cash-settled in $INFERRA.
Provider bond	Providers stake $INFERRA to list capacity; the stake is slashable.
Fee sink	A 2% fee on every settled order flows to the treasury.

2.1 The value loop

Supply is fixed, so the token accrues value from use, not issuance. Every GPU-hour bought routes $INFERRA through escrow and takes a 2% fee to the treasury. As demand grows, more of the supply sits locked in provider bonds and open escrows, and protocol revenue compounds, without minting a single new token. Treasury revenue funds buyback, liquidity, and ecosystem incentives.

2.2 No hidden mint

The mint authority is revoked. There is no lever to print new tokens to pay rewards or cover a shortfall. Incentives come from allocated pools and protocol revenue, never from dilution.

3. The exchange

Inferra trades the price of compute. Each GPU model (H100, H200, A100, MI300X, B200) carries a live price quoted in $INFERRA per GPU-hour: the Inferra Index.

3.1 The oracle

A per-model oracle publishes the GPU-hour price the exchange runs on. The index is not a fixed rate card. It moves with real activity, because every trade and every compute purchase pushes it, so it tracks live supply and demand. The same index is the price at which compute is bought, which makes trading it price discovery for compute itself: push the price up by trading and you also pay more for the compute you buy, so the loop stays honest.

3.2 Perpetuals

Inferra runs perpetuals on the Inferra Index. Go long or short on the price of compute with leverage up to 10x, margined and settled in $INFERRA, marked continuously against the oracle. How it works:

- The protocol is the counterparty to every position, so traders can always open and close without waiting for a matching order. - Trader margin and the protocol's bankroll sit in separate vaults, so one trader's collateral is never spent on another's profit. - A position that falls below maintenance margin is

liquidated, with an insurance fund backstopping the venue.

Dated futures use the same engine with a settlement date, so a buyer can lock in the cost of future GPU hours and a provider can lock in forward revenue.

3.3 Why Solana

Sub-second settlement and near-zero fees let the exchange mark, trade, and settle continuously in $INFERRA at a cost that would be impossible on a slower chain. One chain and one asset carry the marketplace and the exchange together.

4. The long-term vision

Inferra ships in phases. Each phase is useful on its own and sets up the next. The order is intentional: real commerce first, decentralization second, a mature trading market last.

Phase 1: real compute commerce. Non-custodial escrow, real SSH delivery, and verified providers, live in $INFERRA. Buy real GPU compute from a verified provider, paid in $INFERRA, delivered to your terminal.

Phase 2: the open supply network. The node agent is open-sourced so anyone can install it, bond a stake, and become a provider without permission. Supply becomes a network instead of a list, secured by delivery attestation and anti-sybil checks.

Phase 3: a real price for compute. As purchase volume builds, Inferra publishes a public price for GPU compute, the Inferra Index, with a price API, so anyone can read or trade the cost of compute.

Phase 4: compute as an asset class. A liquid index and forward curve make compute hedgeable and investable. Builder products sit on top: prepaid compute credits, an OpenAI-compatible inference gateway settled in $INFERRA, and an SDK to buy, sell, and hedge compute from code.

5. Design principles

Non-custodial by default. Value sits in on-chain escrows and vaults with enforced invariants, not in a company wallet. The contract bounds what the protocol can do.

Real before reflexive. The marketplace settles real machines, so the index is anchored to genuine compute demand rather than an invented number.

Fixed supply. One token, no inflation, value from real usage. Incentives are funded from revenue and allocation, not dilution.

6. Status and disclaimers

Inferra is under active development. Features, parameters, and timelines may change as the protocol matures.

This document is for information only. It is not an offer to sell or a solicitation to buy any token, and not financial, legal, or tax advice. $INFERRA is a utility token for transacting compute within the Inferra protocol, not an investment contract or a claim on any entity's revenue. Derivatives carry risk of total loss and are unavailable where prohibited. Nothing here promises future functionality or value.