Bificurated Inference Ledgering - The Two-Phase Transaction

Nesa's inference technique employs a unique two-phase transaction mechanism that we call Bifurcated Inference Ledgering (BIL). BIL decouples the inference request from the inference response to streamline the network’s inference process and enhance system scalability.

Nesa has been engineered with this design because a major challenge of executing AI inference on the blockchain comes from the computational intensity of large model inference which can severely impact transaction throughput and block generation times.

Traditional approaches by blockchain projects, such as the method adopted by the Cortex project, involve integrating an INFERENCE command directly into the blockchain protocol. While this approach has the merit of simplicity, it is not without its drawbacks.

Specifically, it is characterized by slow performance when dealing with complex models, which subsequently leads to a bottleneck effect on the chain’s overall throughput.

Instead, Nesa takes a different, bifurcated approach.

The First Phase: Inference Request Queueing.

In the first phase, a user submits an inference request transaction, which includes the necessary details for the inference task but does not trigger the execution immediately. Instead, this transaction is registered into a queue within the blockchain ecosystem. Our system utilizes a priority queue to order these requests.

The priority for each request is directly correlated with the fee paid by the user; higher fees result in higher priority, ensuring that users with urgent needs can opt for faster processing by electing to pay a premium. This dynamic pricing model aligns resource allocation with market demand, thereby optimizing system efficiency.

Requests are registered and enqueued within smart contracts, which act as decentralized and transparent priority queue managers. These smart contracts are programmed to organize the requests according to their assigned priority, ensuring that the system's resources are allocated in a fair and economically rational manner.

This non-blocking transaction allows the blockchain to continue processing other transactions, maintaining high throughput and low latency.

The Second Phase: Inference Execution and Response

The second phase is initiated once the inference request reaches the front of the queue. Separate transactions are created by the designated inference committee, which is tasked with actually performing the inference task.

Upon completion, the results are recorded and disclosed. This phase is conducted mostly off-chain to prevent the computational load from affecting the blockchain's performance.

The separation of request and execution transactions in Nesa's design offers several advantages. Firstly, it avoids the congestion that can occur when the blockchain waits for computationally intensive inferences to complete.

Secondly, it provides flexibility in resource allocation, as the inference task can be processed in parallel with other blockchain operations.

Finally, it ensures that the blockchain maintains a consistent and fast block generation time, regardless of the complexity or size of the AI models being inferred.

Through two phases of transaction, Bifurcated Inference Ledgering ensures that Nesa can scale as more requests enter this system. This design forms the backbone of our decentralized inference system.

Previous
Decentralized Inference

Next
Robust Inference Committee Selection

Last updated 1 month ago