nsdi '26

PlanetServe: A Decentralized, Scalable, and Privacy-Preserving Overlay for Democratizing Large Language Model Serving

Fei Fang[†,*], Yifan Hua[†,*], Shengze Wang[†,*], Ruilin Zhou[†], Yi Liu[†], Chen Qian[†], Xiaoxue Zhang[‡]

* Co-primary authors

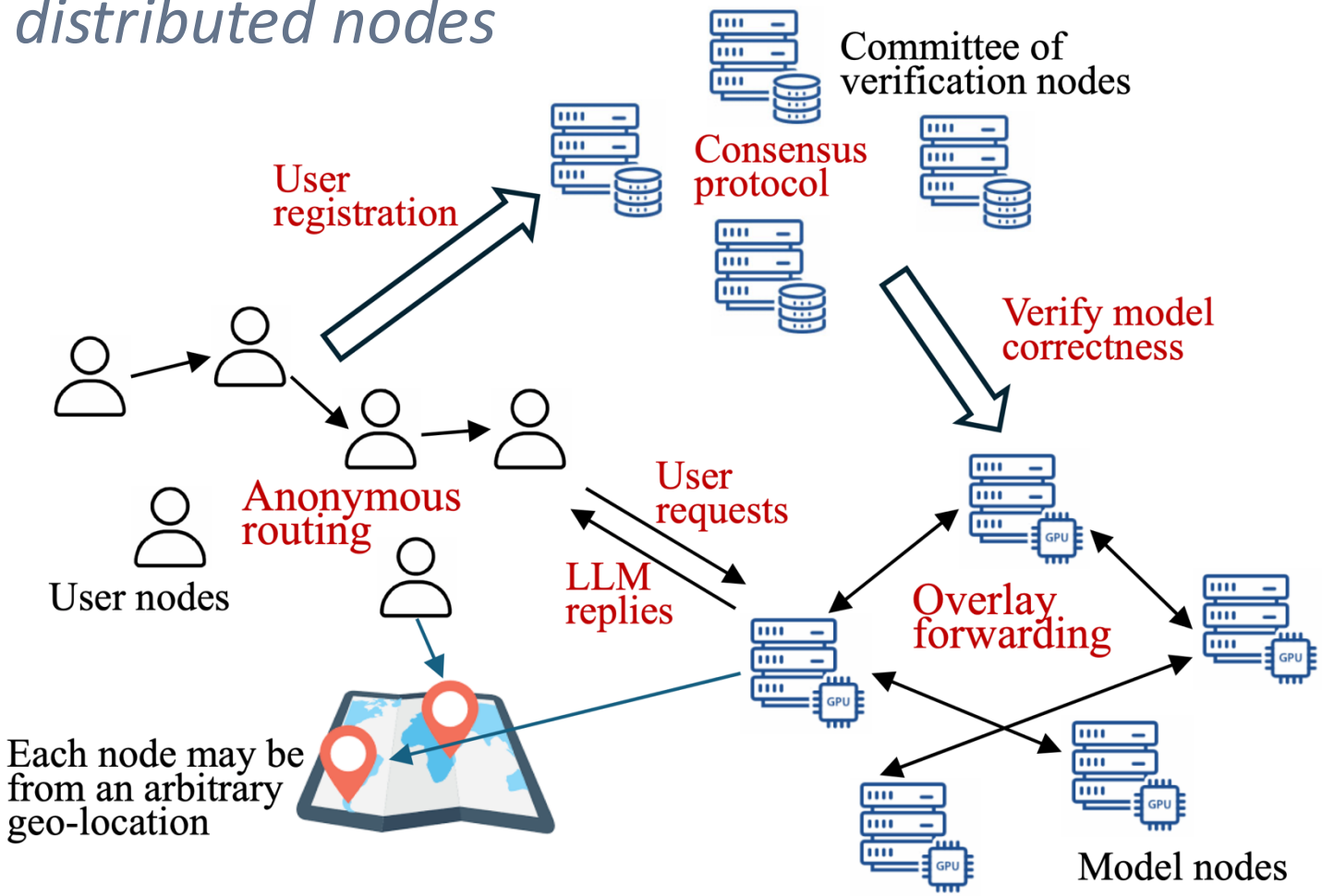
† UNIVERSITY OF CALIFORNIA SANTA CRUZ

‡ University of Nevada, Reno

PlanetServe: Core Vision

Inspired by PlanetLab — a P2P overlay allowing researchers to deploy Internet applications across distributed nodes

PLANETLAB
An open platform for developing, deploying, and accessing planetary-scale services



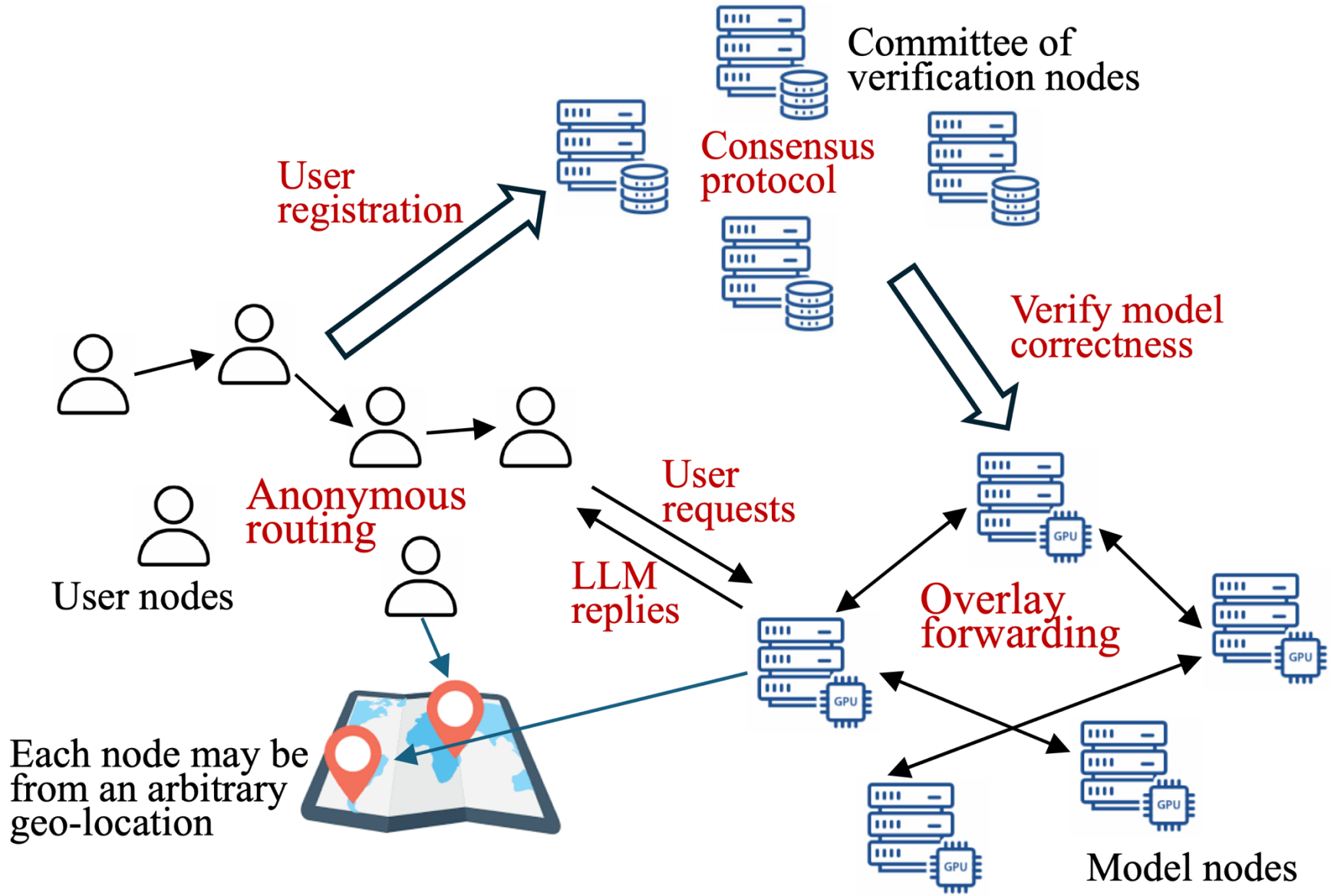
C1 Overlay Network Organization

C2 Anonymous Communication

C3 Overlay Forwarding for Efficiency

C4 Decentralized Verification

Node Roles & Architecture



User Node

- End-users seeking LLM access
- Maintains user anonymity
- Can join or leave at any time

Model Node

- Executes LLM inference
- Contributed by orgs/individuals — earns reputation

Verification Node

- Runs BFT consensus
- Issues credential checks
- Maintains reputation scores

The S-IDA Protocol: How It Works

Secure Information Dispersal Algorithm (S-IDA) — enables anonymity, confidentiality, and fault tolerance simultaneously

1

Plaintext

message, M → AES Key, K → ciphertext, K(M)

2

Rabin's IDA

k-of-n threshold

Slice 1
Slice 2
Slice 3
Slice 4

(n, k)
n = 4, k = 3

3

Shamir's Secret Sharing

k-of-n threshold

Slice 1
Slice 2
Slice 3
Slice 4

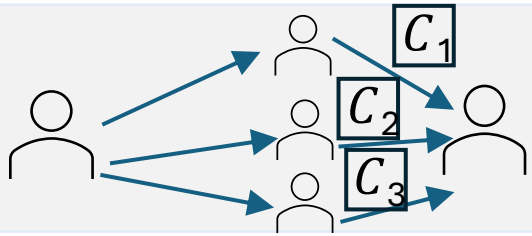
(n, k)
n = 4, k = 3

4

$K(M)_i$
K_i
clove C_i

Build Cloves

5



Multi-path Send

The S-IDA Protocol: How It Works

Secure Information Dispersal Algorithm (S-IDA) — enables anonymity, confidentiality, and fault tolerance simultaneously

Erasure coding technique optimized for efficient data dispersal.

1

Plaintext

message, M

ertext, K(M)

2

Rabin's IDA

k-of-n threshold

Slice 1
Slice 2
Slice 3

Slice 4

(n, k)
n = 4, k = 3

3

Shamir's Secret Sharing

Slice 1
Slice 2

Slice 4

(n, k)
n = 4, k = 3

information-theoretic secrecy

4

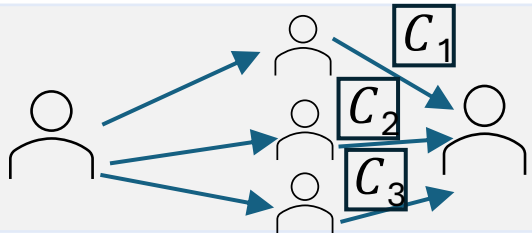
$K(M)_i$

K_i

clove C_i

Build Cloves

5

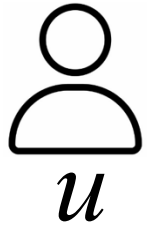


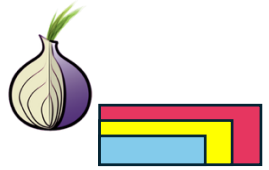
Multi-path Send

Anonymous Routing

Step 1

Preparation





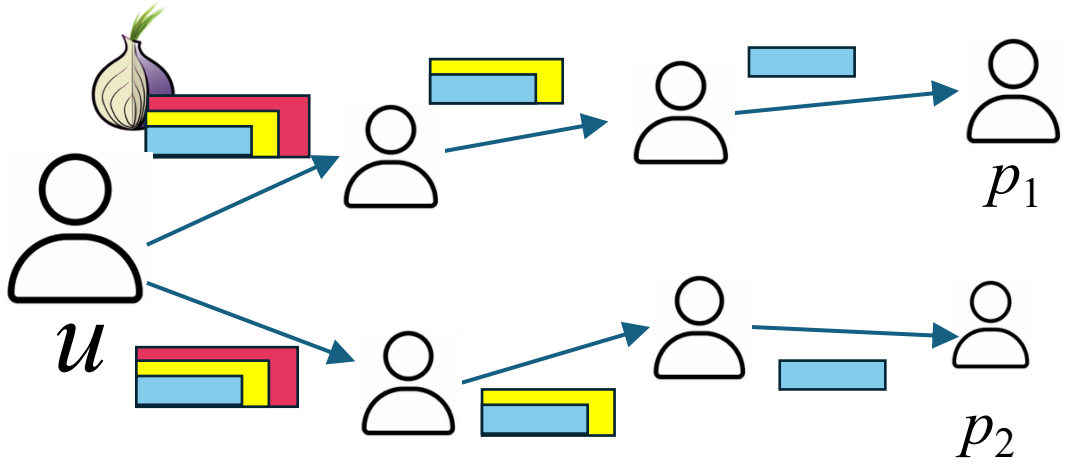
Select N proxies
N = 4

C_1 C_2
C_3 C_4

Determines (n, k)
n = 4, k =3

Step 2

Establish Proxy

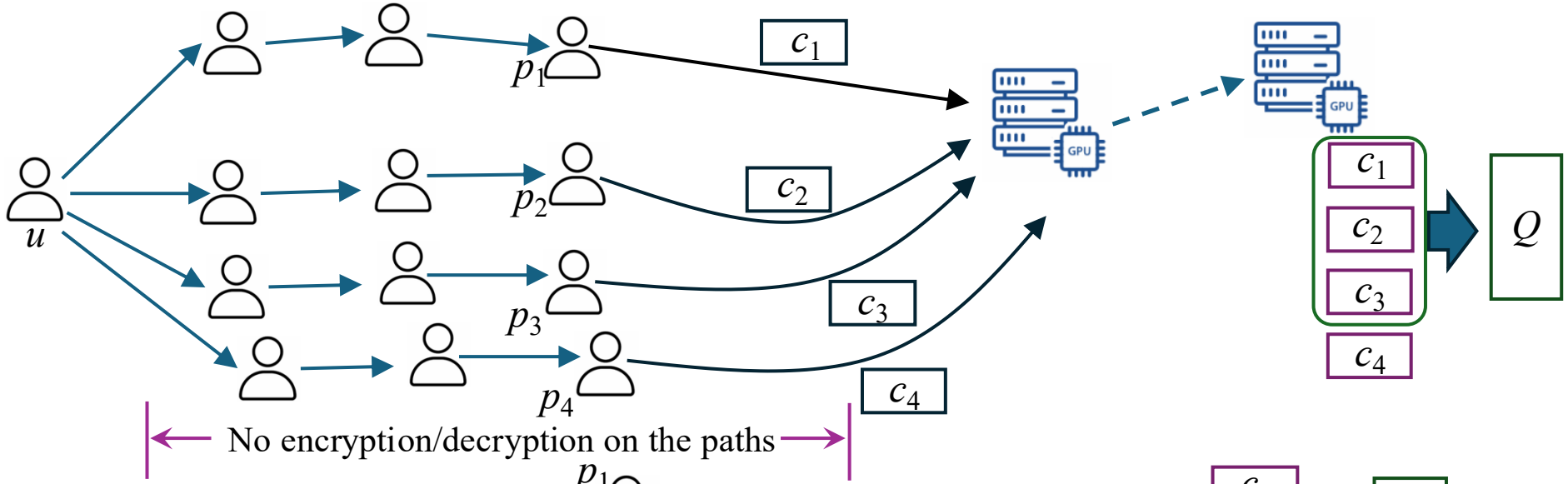


- N Onion paths
- length l=3 hops
- Path IDs & session IDs stored at each relay.
- No future public-key ops needed.

Anonymous Routing

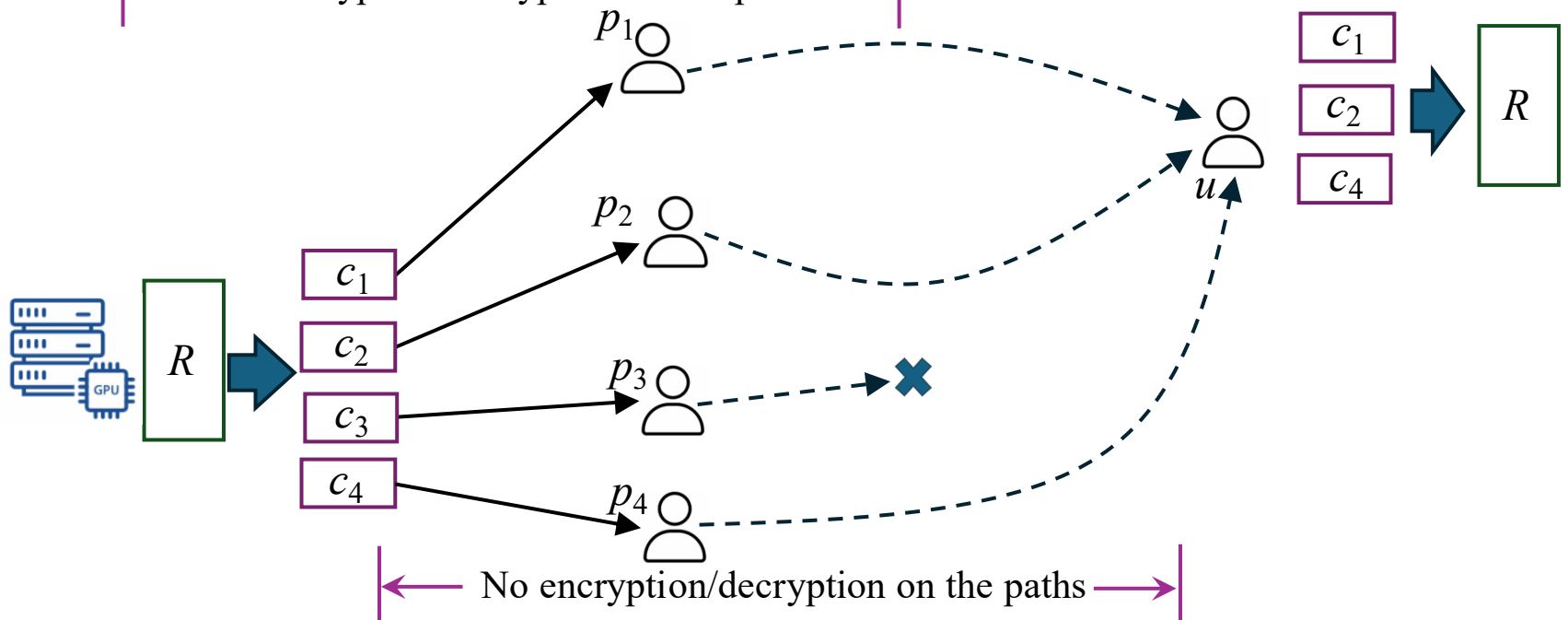
Step 3

Query

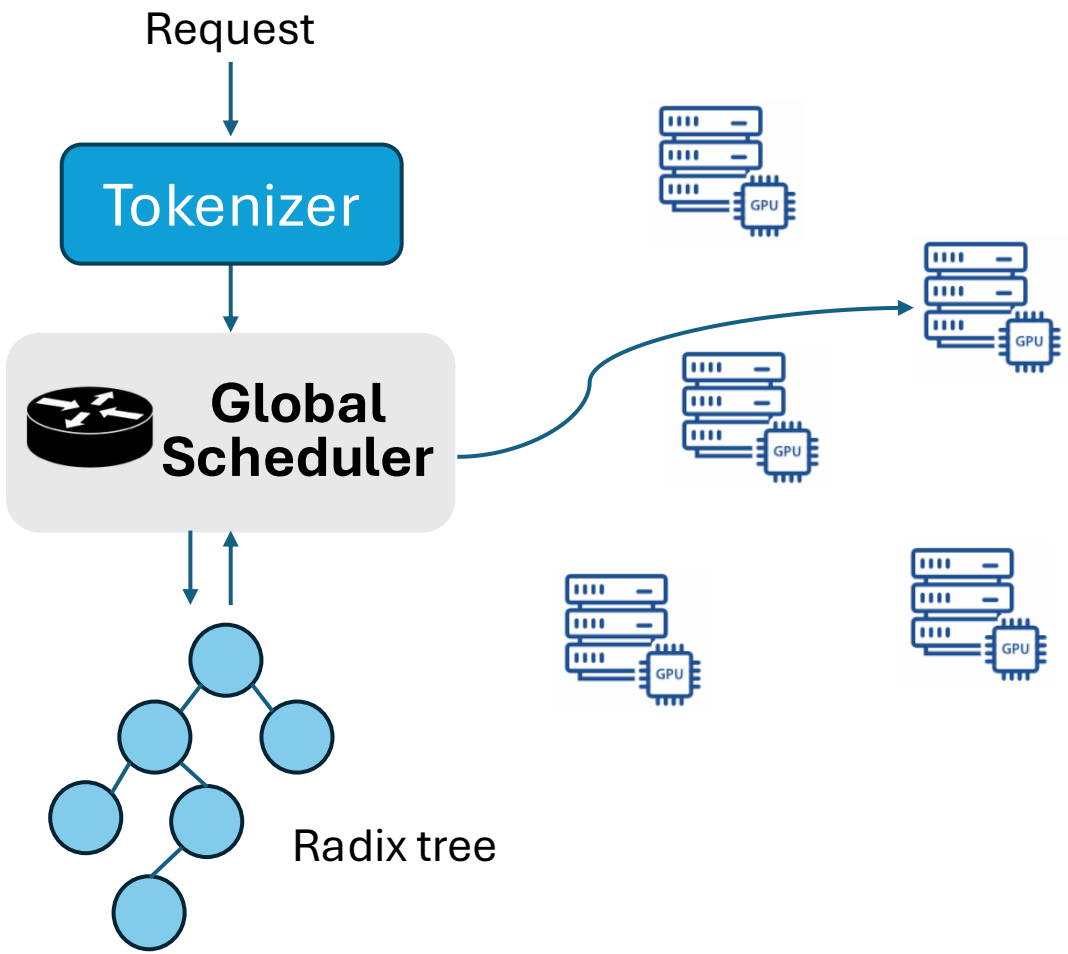


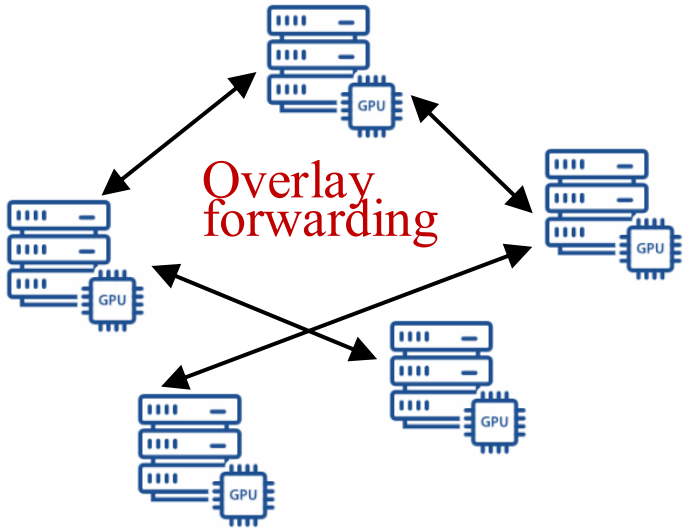
Step 4

reply



Hash-Radix Tree (HR-Tree): Design





no central router

Hash-Radix Tree (HR-Tree): Design

Req 1 The weather is nice today. Let us go surfing!
0x0A 0x8B 0x00 0x71

Req 2 The weather is bad today. The ceiling is leaking.
0x0A 0x5C 0x89 0x21

Req 3 The weather is nice today. Hope we can have a nice trip!
0x0A 0x8B 0x54 0x77

✕ High Memory Overhead

full raw data

✓ Hash Fingerprints

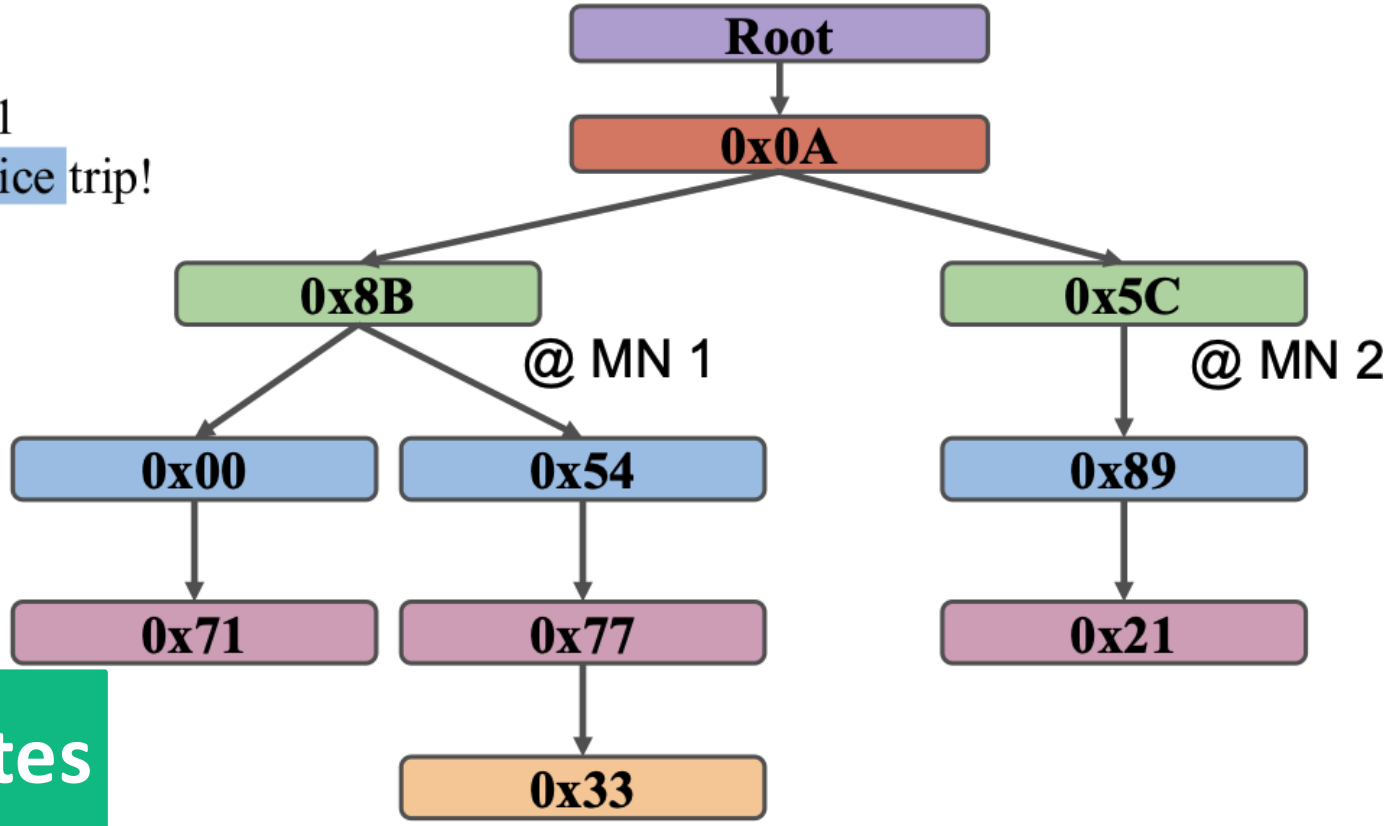
8-bit hash fingerprints

✕ High Communication Overhead

Full broadcasts

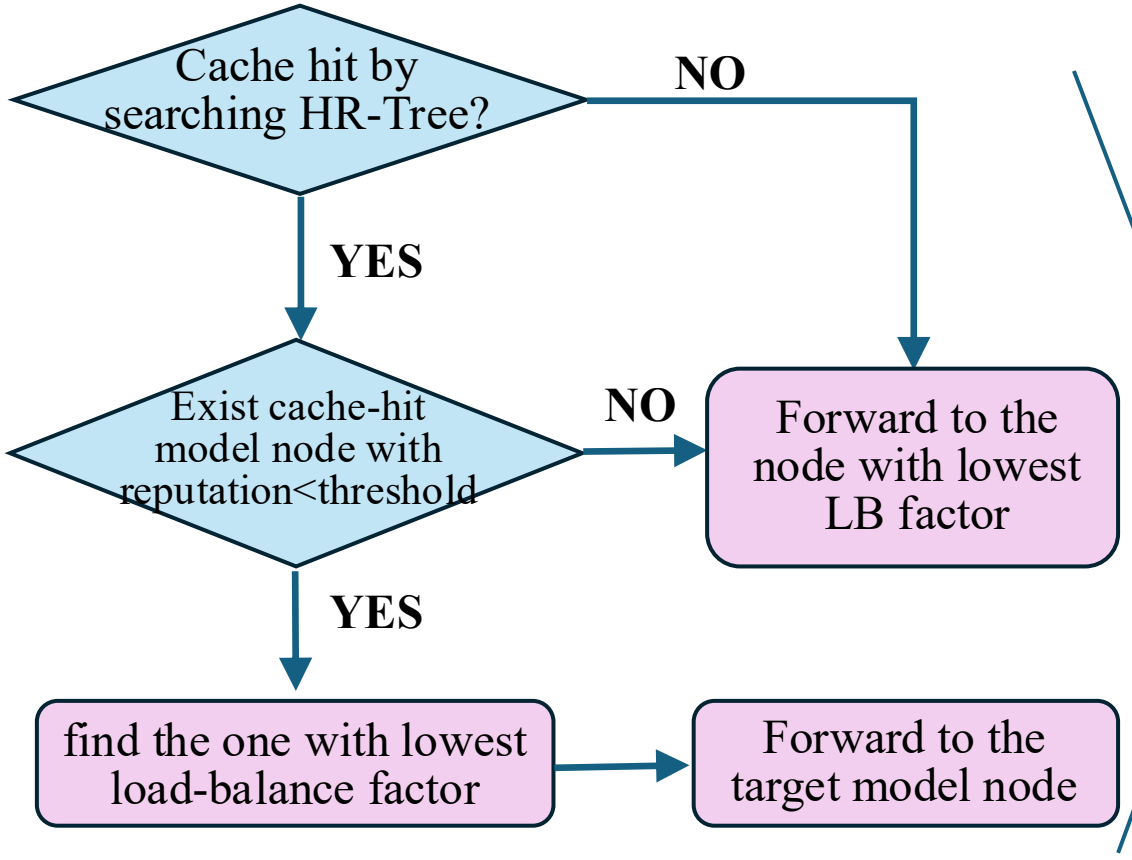
✓ Delta Updates

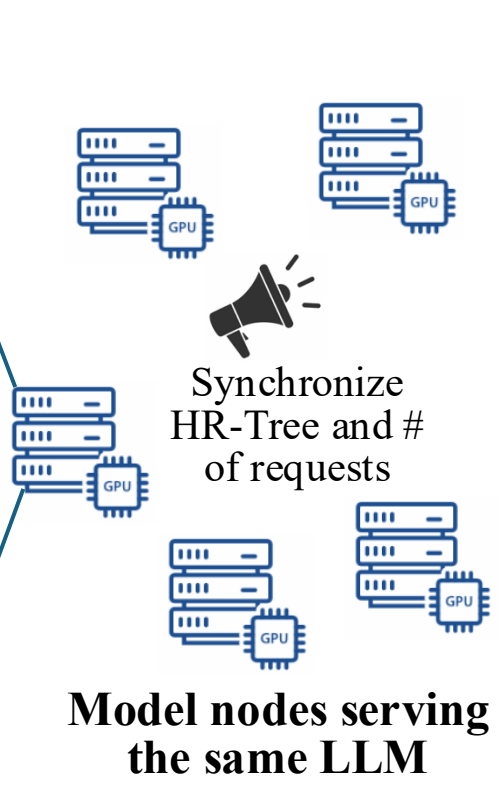
snapshot + minimal updates.



	IP address	LB factor	Reputation
Model node 1	178.3.2.1	0	0.9
Model node 2	178.3.2.2	0.5	1

Forwarding Logic & Load Balancing





Load-Balance Factor

Forward to the node with:

- ✓ low latency
- ✓ short queue
- ✓ enough capacity
- ✓ good reputation

Session Affinity Model node IP included in each response → consecutive prompts routed to the same node, maximizing KV cache reuse within a conversation.

Model Verification

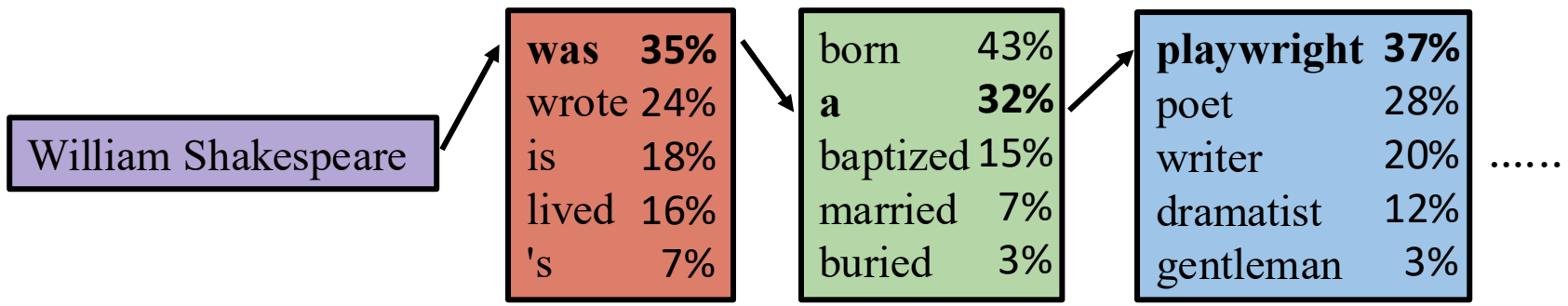
Key insight: Model nodes cannot distinguish verification probes from regular user requests (all routed through anonymous overlay)

1 Challenge

2 Collect Response

3 Compute Credit

4 Reputation Update & Punishment



Match reference → ↑ credit
Diverge repeatedly → ↓ credit

Model Verification

Key insight: Model nodes cannot distinguish verification probes from regular user requests (all routed through anonymous overlay)

1 Challenge

2 Collect Response

3 Compute Credit

4 Reputation Update & Punishment

Untrusted if R < threshold

Node marked untrusted and removed from routing

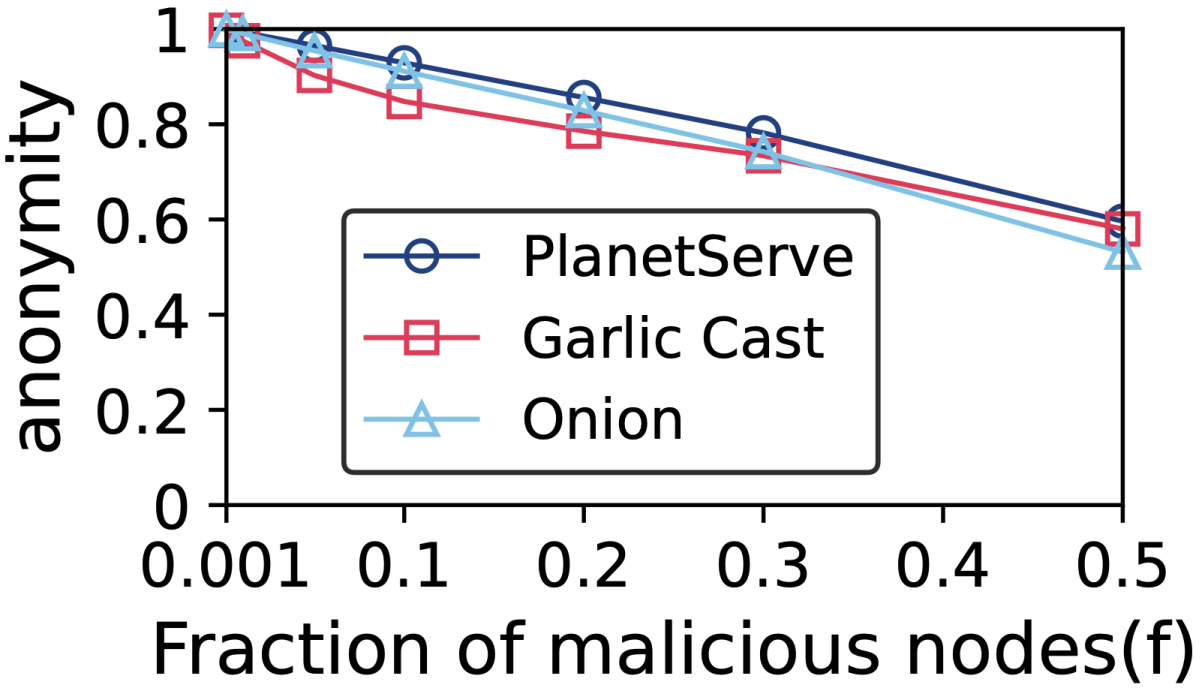
Match → Trust ↑

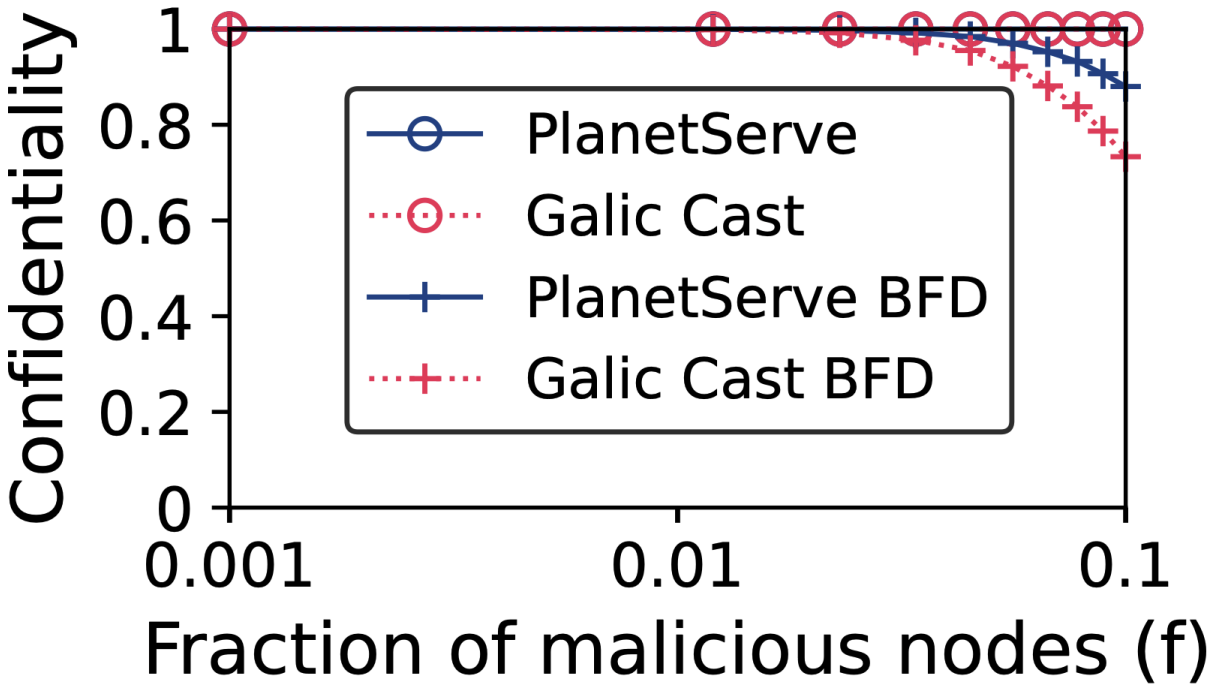
Mismatch → Trust ↓

Repeated mismatch →penalty

Security Analysis & Performance Evaluation

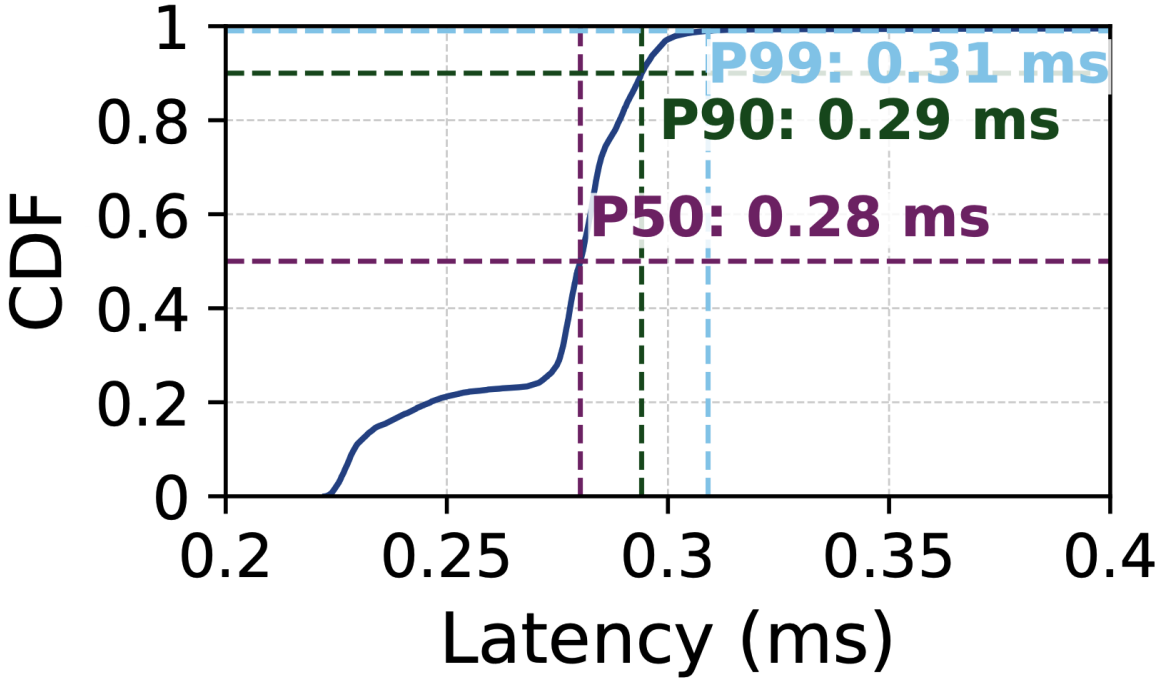
Entropy-based Anonymity and Message confidentiality



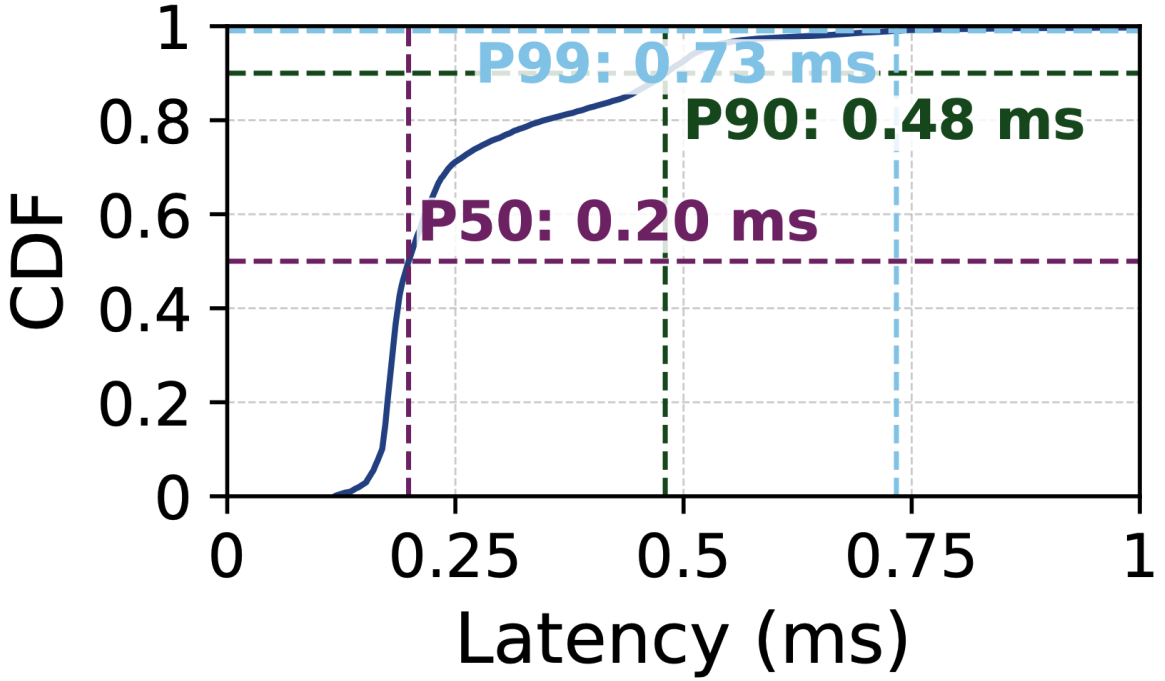


Security Analysis & Performance Evaluation

Anonymous message (clove) processing introduce negligible overhead



(a) Preparation latency

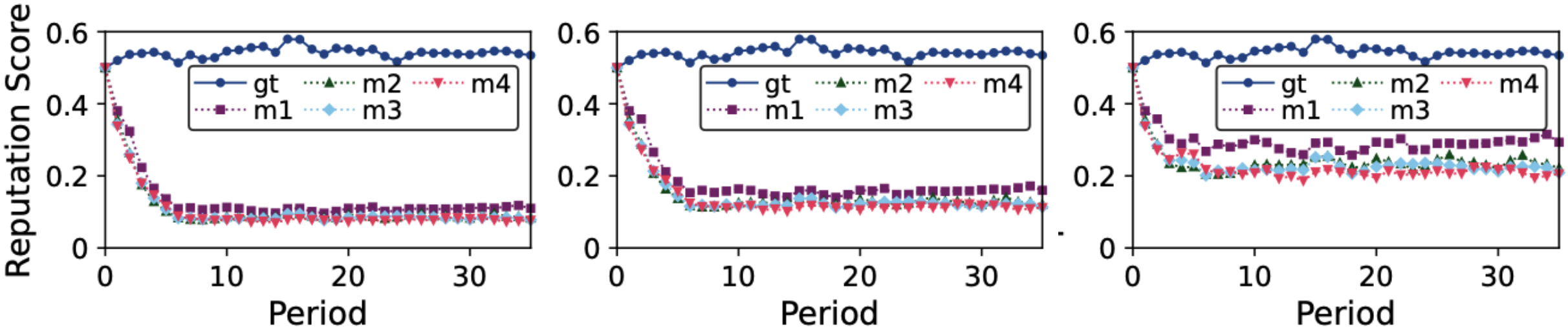


(b) Decryption latency

The anonymity overhead is negligible relative to inference time.

Security Analysis & Performance Evaluation

Dishonest Model Nodes

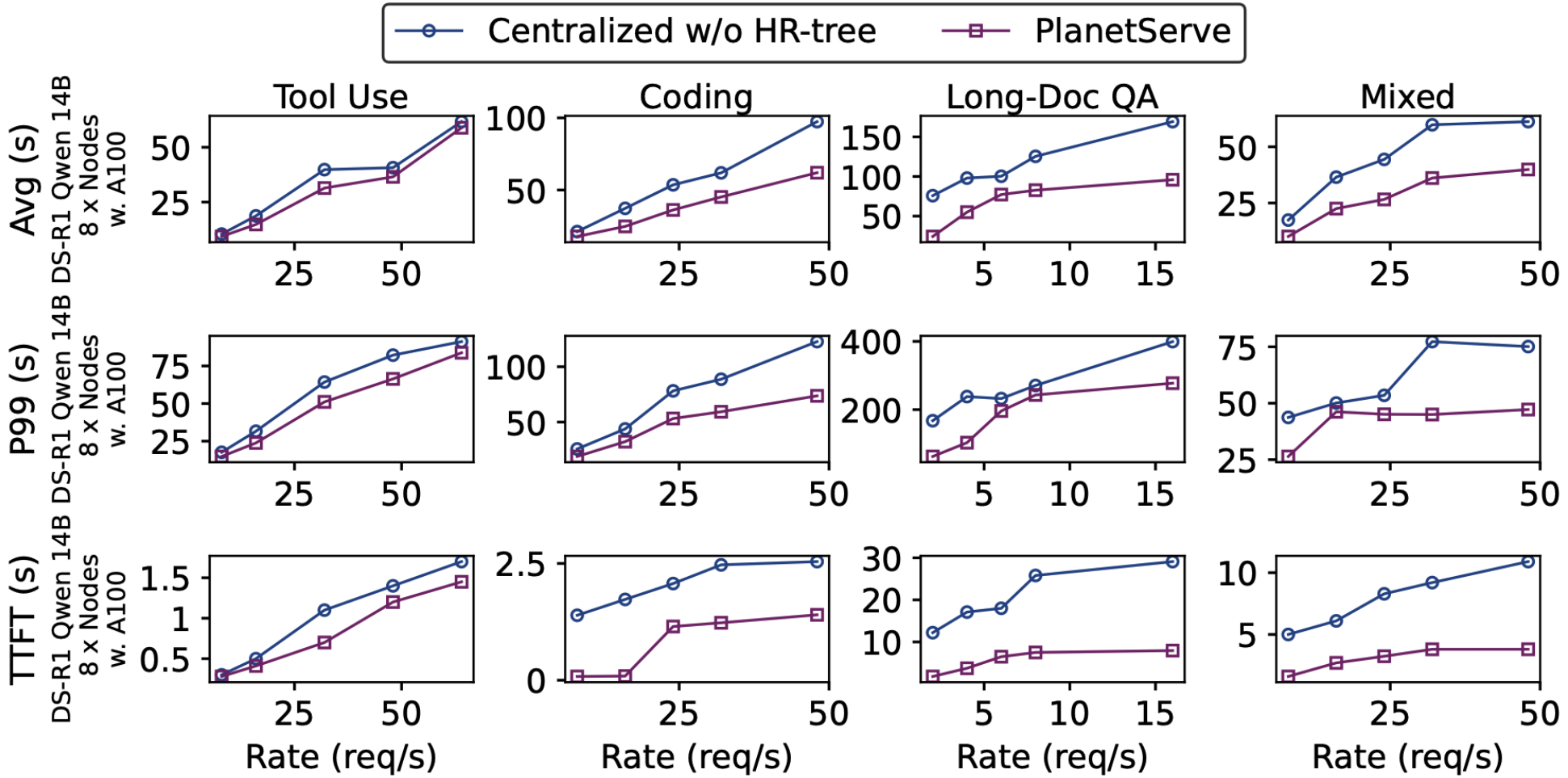


(a) Punishment threshold $\gamma = 1$. (b) Punishment threshold $\gamma = \frac{1}{3}$. (c) Punishment threshold $\gamma = \frac{1}{5}$.

Honest nodes keep high reputation, while dishonest nodes are quickly suppressed.

Security Analysis & Performance Evaluation

Overlay Forwarding



Conclusion

PlanetServe: a Decentralized, Scalable, and Privacy-Preserving Overlay for Democratizing LLM Serving

Novel anonymous routing achieves anonymity with minimal overhead

HR-Tree enables decentralized KV cache reuse + load balancing → >50% latency reduction

Model verification reliably detects dishonest nodes

Thank You

Questions & Discussion

Code: github.com/fffeifang/PlanetServe.git