Distributed Inference Network(DIN) and Internet of Agents (IOA)

Motivation and Requirements

IETF 123 HotRFC

China Mobile

Motivation: The Imperative for DIN (Distributed Inference Network)

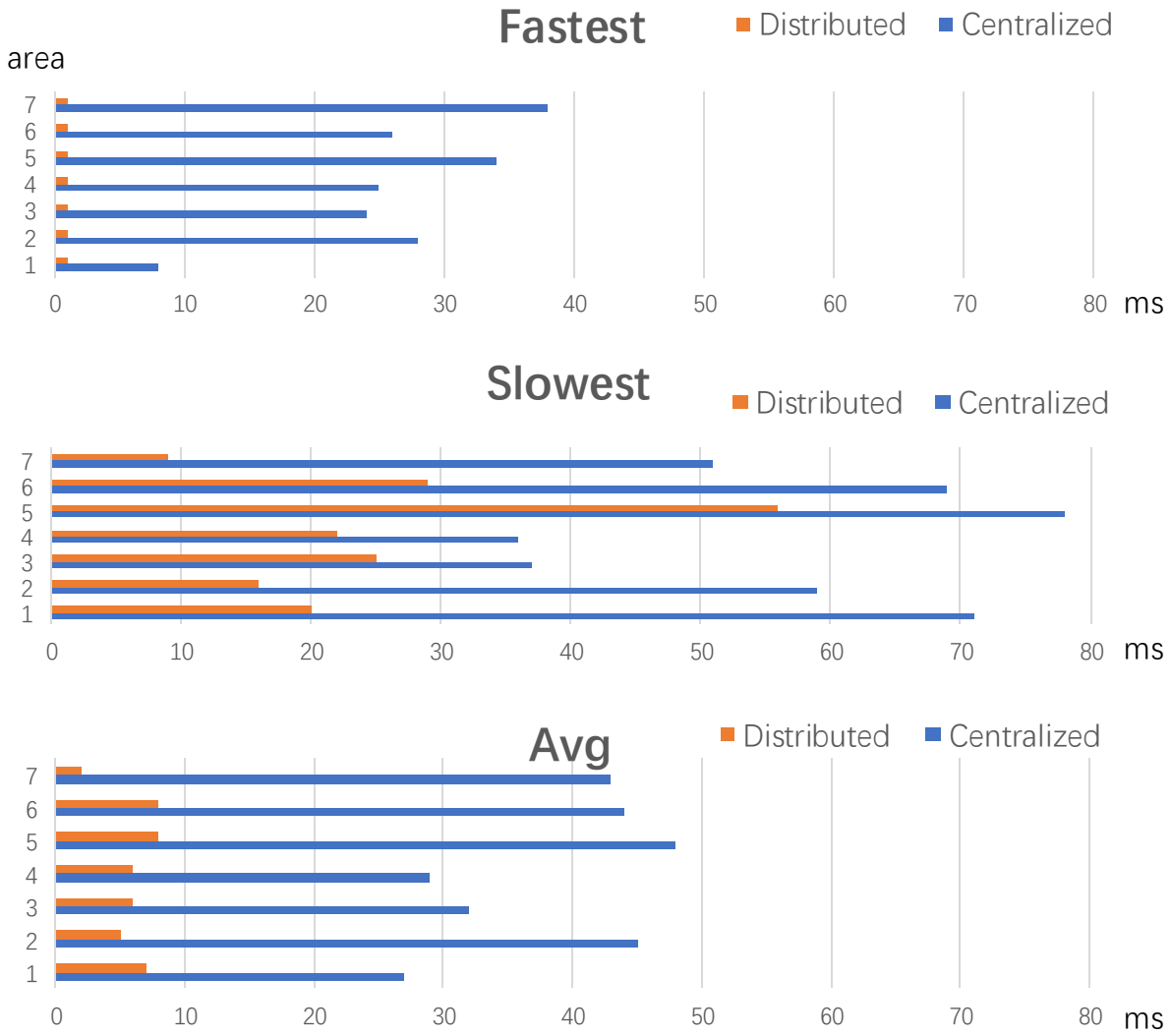
- **Era shift**
  - The explosive growth in AI inference service marks the *inclusive AI inference era*
  - AI evolves: *Human-model → Multi Agent System → Internet of Agents (IOA)*
  - presenting both challenges and opportunities for networking infrastructure

- **Centralized inference architecture hits limits**
  - Cannot scale for exploding concurrency (users/API/M2M/Agents)
  - Fail to meet ultra-low latency demands for M2M/A2A

- **DIN necessity**
  - requires novel Distributed Inference Network (DIN) architectures to handle extreme concurrency and latency, and research in scenarios, requirements and technologies

Centralized vs Distributed inference deployment delay comparison

- tested on July 17, 2025
- 142 probes, different locations and operator networks
- Centralized: 1 DNS -> 2 IPs in Shanghai
- Distributed: 1 DNS -> 133 IPs over China



IOA (Internet of Agents) and DIN

- **IOA: the paradigm change**
  - Billions of autonomous agents as new “users” driving M2A/A2A dominance
  - Workloads shift from training-centric to agent-driven inference

- **IOA’s network demands**
  - “Latency is Survival”: sub-ms A2A decisions
  - Concurrency explosion: 1000x+ sessions, bursty agent-spawned task chains
  - New traffic patterns: unpredictable & locality-sensitive flows

- **How DIN as IOA’s Enabler**
  - Scalability: Edge inference, dynamic load-aware routing, …
  - Latency by design: computing near agents to accelerate A2A loops
  - Co-evolution: agent-aware scheduling & topology adaptation

- **DIN: Not "faster pipes" – but the substrate for emergent agent intelligence**

We hope to reach a consensus on requirements within the IETF and further conduct research on problem definition and solutions.

Welcome to join us

Distributed Inference Network(DIN) and Internet of Agents (IOA)

Side Meeting

Segovia room

9:30-11:00am, Wednesday, July 23