

Read-out from IRTFOPEN Meeting on Internetworking Challenges for AI

20260318 – IRTF Open meeting – Session 2

Internetworking Challenges for AI

IRTFOPEN Meeting

3 presentations about complementary challenges associated with distribution of AI workloads over the Internet:

1. **Disaggregated Architecture for LLM Inference** by Mingxing Zhang

KV-centric disaggregation of LLM inference engine first done inside a single datacenter realm, but paving the way for **KV-centric networking at scale** to distribute inference workloads beyond a DC's realm

2. **Reliability engineering challenges in Networking for AI** by Hong Xu

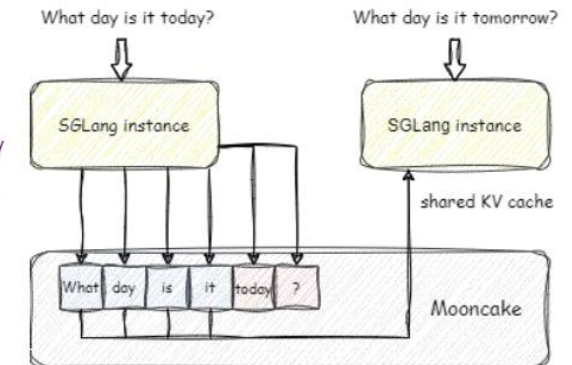
3. **On AI Agent Networking** by Lixia Zhang

Store More: Elastic Shared Multi-layer KV Cache



Key features

- **Distributed KV cache sharing:** storing one and usable by all
- **Dynamic resource scaling:** dynamically adding and removing store nodes (startup in <80s for 500GB memory and 8 RDMA NICs)
- **Multi-layer storage (WIP):** offloading cached data from RAM to SSD



SYSTEMATIC FRAMEWORK FOR AGENT-DRIVEN DIAGNOSTICS



- Standardized reproducible benchmarking
- C1: Realistic comprehensive fault datasets
- C2: Faithful, sandboxed, interactive environment
 - Agents interact with the env
 - Emulation, simulation, testbed
- C3: Integration with production tooling
 - Pingmesh, Mycroft, etc.

ADDRESSING AGENTIC AI NETWORKING CHALLENGES

Starting point: get naming right, then build security into agentic AI

- DNS as a unifying namespace for cyberspace
 - DNS: decentralized name management with TLD coordination and name delegation
 - DNS: offering globally unique, semantically meaningful names
 - Every entity — organization, user, agent, and service — have DNS names as primary identifiers
 - Semantic meaningfulness makes trust reasoning human-navigable
- Identity = Name + Key
 - Name provides semantic context; key provides cryptographic verifiability
 - Trust chains must be human-navigable; machine-verifiability alone is not adequate
- Crypto protections anchor on local trust
 - Global namespace, local trust
 - AI agents make this a scaling requirement, not an architecture option
 - Delegation chains are native infrastructure: scoped, verifiable, multi-hop

Internetworking Challenges for AI

IRTFOPEN Meeting

3 presentations about complementary challenges associated with distribution of AI workloads over the Internet:

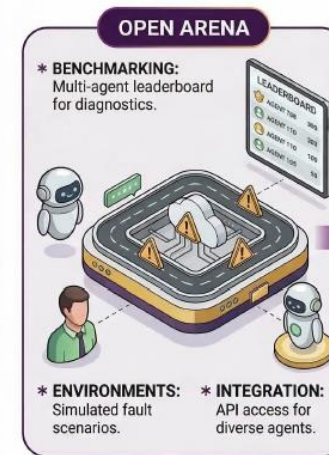
1. ***KV-centric networking at scale*** in *Disaggregated Architecture for LLM Inference* by Mingxing Zhang

2. ***Reliability engineering challenges in Networking for AI*** by Hong Xu

Research initiatives towards agentic management of (AI) DC infrastructure require ***testbed and benchmark*** to foster idea exchange, and benchmark solutions between one another

1. ***On AI Agent Networking*** by Lixia Zhang

SYSTEMATIC FRAMEWORK FOR AGENT-DRIVEN DIAGNOSTICS



Created by NanoBanana

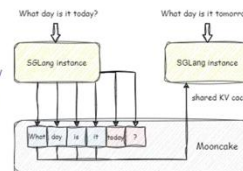
Hong Xu (CUHK)

- Standardized reproducible benchmarking
- C1: Realistic comprehensive fault **datasets**
- C2: Faithful, sandboxed, interactive **environment**
 - Agents interact with the env
 - Emulation, simulation, testbed
- C3: **Integration** with production tooling
 - Pingmesh, Mycroft, etc.

Store More: Elastic Shared Multi-layer KV Cache

Key features

- Distributed KV cache sharing: storing one and usable by all
- Dynamic resource scaling: dynamically adding and removing store nodes (startup in <80s for 500GB memory and 8 RDMA NICs)
- Multi-layer storage (WIP): offloading cached data from RAM to SSD



ADDRESSING AGENTIC AI NETWORKING CHALLENGES

Starting point: get naming right, then build security into agentic AI

- **DNS as a unifying namespace for cyberspace**
 - DNS: decentralized name management with TLD coordination and name delegation
 - DNS: offering *globally unique, semantically meaningful* names
 - Every entity — organization, user, agent, and service — have DNS names as primary identifiers
 - Semantic meaningfulness makes trust reasoning human-navigable
- **Identity = Name + Key**
 - Name provides semantic context; key provides cryptographic verifiability
 - Trust chains must be human-navigable; machine-verifiability alone is not adequate
- **Crypto protections anchor on local trust**
 - Global namespace, local trust
 - AI agents make this a scaling requirement, not an architecture option
 - Delegation chains are native infrastructure: scoped, verifiable, multi-hop

Internetworking Challenges for AI

IRTFOPEN Meeting

3 presentations about complementary challenges associated with distribution of AI workloads over the Internet:

1. ***KV-centric networking at scale*** in *Disaggregated Architecture for LLM Inference* by Mingxing Zhang

2. ***Testbeds and benchmarks*** in *Reliability engineering challenges in Networking for AI* by Hong Xu

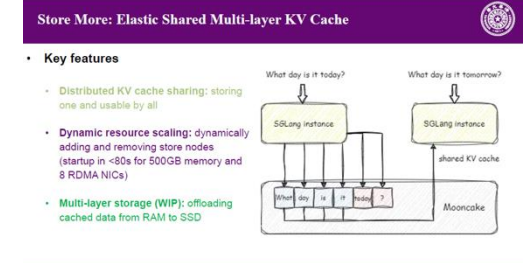
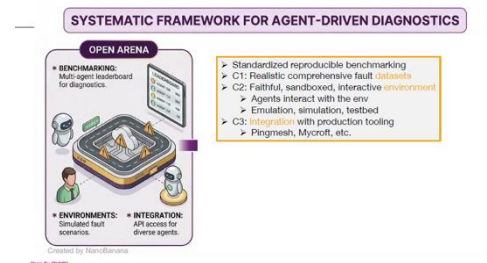
3. ***On AI Agent Networking*** by Lixia Zhang

Need to avoid rushing to a solution, get ***naming and addressing*** right in relationship with security as a foundation for trustworthy agent to agent communications

ADDRESSING AGENTIC AI NETWORKING CHALLENGES

Starting point: get naming right, then build security into agentic AI

- **DNS as a unifying namespace for cyberspace**
 - DNS: decentralized name management with TLD coordination and name delegation
 - DNS: offering *globally unique, semantically meaningful* names
 - Every entity — organization, user, agent, and service — have DNS names as primary identifiers
 - Semantic meaningfulness makes trust reasoning human-navigable
- **Identity = Name + Key**
 - Name provides semantic context; key provides cryptographic verifiability
 - Trust chains must be human-navigable; machine-verifiability alone is not adequate
- **Crypto protections anchor on local trust**
 - Global namespace, local trust
 - AI agents make this a scaling requirement, not an architecture option
 - Delegation chains are native infrastructure: scoped, verifiable, multi-hop



Research challenges for INet4AI

Topics of Interest so far

- **Testbed, benchmarks and dataset availability**
 - Enabling experiments
 - Benchmarking new systems
- **Principled naming, identity, trust delegation**
 - Solid foundations for agentic communication
 - Beyond current engineering efforts
- **AI system architecture**
 - (More decentralized) KV-cache-centric architecture
 - Unified transport

Thank you!

Contact: net4ai@irtf.org

Internetworking Challenges for AI

IRTFOPEN Meeting

3 presentations about complementary challenges associated with distribution of AI workloads over the Internet:

1. **Disaggregated Architecture for LLM Inference** by Mingxing Zhang

KV-centric disaggregation of LLM inference engine first done inside a single datacenter realm, but paving the way for **KV-centric networking at scale** to distribute inference workloads beyond a DC's realm

2. **Reliability engineering challenges in Networking for AI** by Hong Xu

Research initiatives towards agentic management of (AI) DC infrastructure require **testbed and benchmark** to foster idea exchange, and benchmark solutions between one another

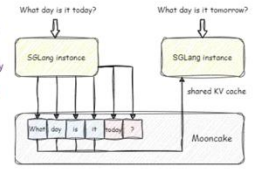
3. **On AI Agent Networking** by Lixia Zhang

Need to avoid rushing to a solution, get **naming and addressing** right in relationship with security as a foundation for trustworthy agent to agent communications

- But not only !
- Some concepts and ideas mentionned in several presentations (e.g. KV-centric communication between agents in Hong Xu's presentation)

Store More: Elastic Shared Multi-layer KV Cache

- **Key features**
 - Distributed KV cache sharing: storing one and usable by all
 - Dynamic resource scaling: dynamically adding and removing store nodes (startup in ~80s for 500GB memory and 8 RDMA NICs)
 - Multi-layer storage (WIP): offloading cached data from RAM to SSD




The diagram illustrates the architecture of the Elastic Shared Multi-layer KV Cache. It shows two SQLang instances, each receiving a query like 'What day is it today?'. These instances are connected to a shared KV cache. Below the cache is a Mooncake storage layer, which is responsible for offloading cached data from RAM to SSD. The diagram also shows a sequence of characters 'What day is it today?' and a question mark, indicating the flow of data and processing.

SYSTEMATIC FRAMEWORK FOR AGENT-DRIVEN DIAGNOSTICS

OPEN ARENA

- **BENCHMARKING:** Multi-agent testbed for diagnostics.
 - Standardized reproducible benchmarking
 - C1: Realistic comprehensive fault **testbeds**
 - C2: Faithful, sandboxed, interactive **environment**
 - Agents interact with the env
 - Emulation, simulation, testbed
 - C3: **Integration** with production tooling
 - Pingmesh, Mycroft, etc.
- **ENVIRONMENTS:** Simulated fault scenarios.
- **INTEGRATION:** API access for diverse agents.

Created by NanoGenius



The diagram shows a central 'OPEN ARENA' box containing a multi-agent testbed. To the right, a list of benchmarking criteria (C1, C2, C3) is provided, along with a list of environments and integration points. The diagram is credited to NanoGenius.

ADDRESSING AGENTIC AI NETWORKING CHALLENGES

Starting point: get naming right, then build security into agentic AI

- **DNS as a unifying namespace for cyberspace**
 - DNS: decentralized name management with TLD coordination and name delegation
 - DNS: offering *globally unique, semantically meaningful* names
 - Every entity — organization, user, agent, and service — have DNS names as primary identifiers
 - Semantic meaningfulness makes trust reasoning human-navigable
- **Identity = Name + Key**
 - Name provides semantic context; key provides cryptographic verifiability
 - Trust chains must be human-navigable; machine-verifiability alone is not adequate
- **Crypto protections anchor on local trust**
 - Global namespace, local trust
 - AI agents make this a scaling requirement, not an architecture option
 - Delegation chains are native infrastructure: scoped, verifiable, multi-hop