Data & Agent Aware Inference & Training Network (DA-ITN)

draft-akhavain-moussa-ai-network-00

Arashmid Akhavain arashmid.akhavain@huawei.com
Hesham Moussa hesham.moussa@huawei.com

Huawei Technologies Canada

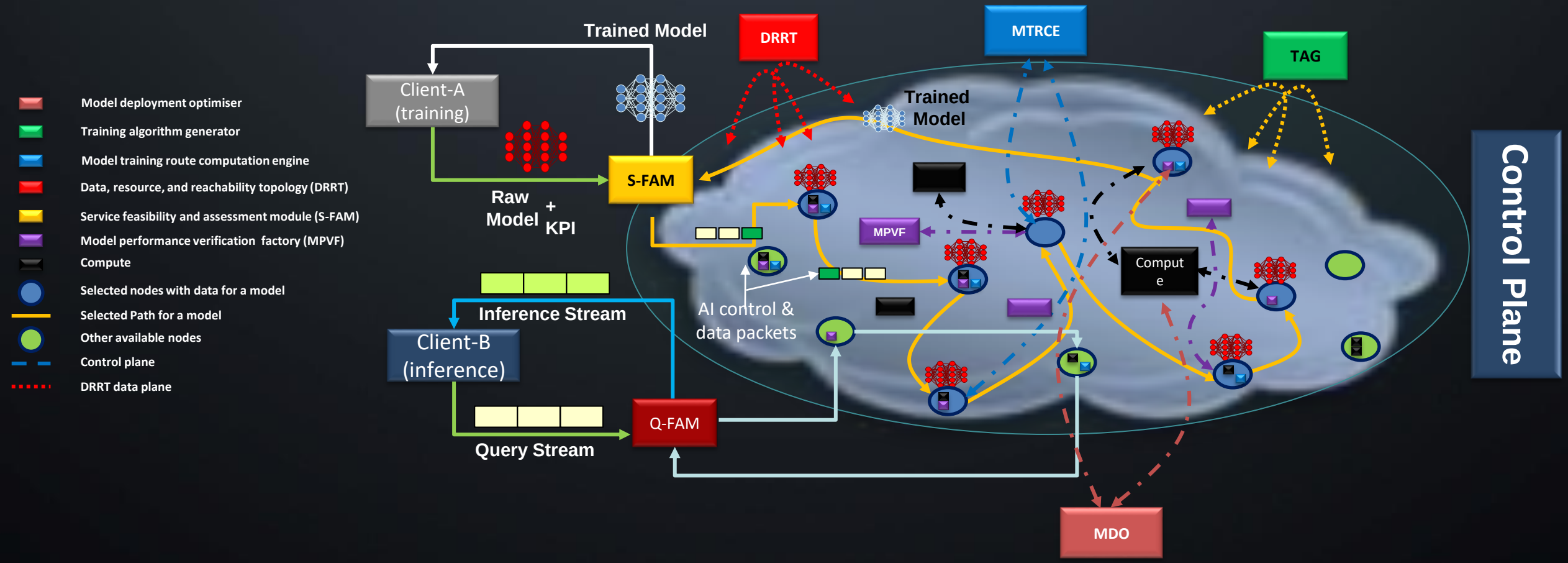
IETF 124 - Montreal
Nov 2025

IETF Draft

https://datatracker.ietf.org/doc/draft-akhavain-moussa-ai-network/

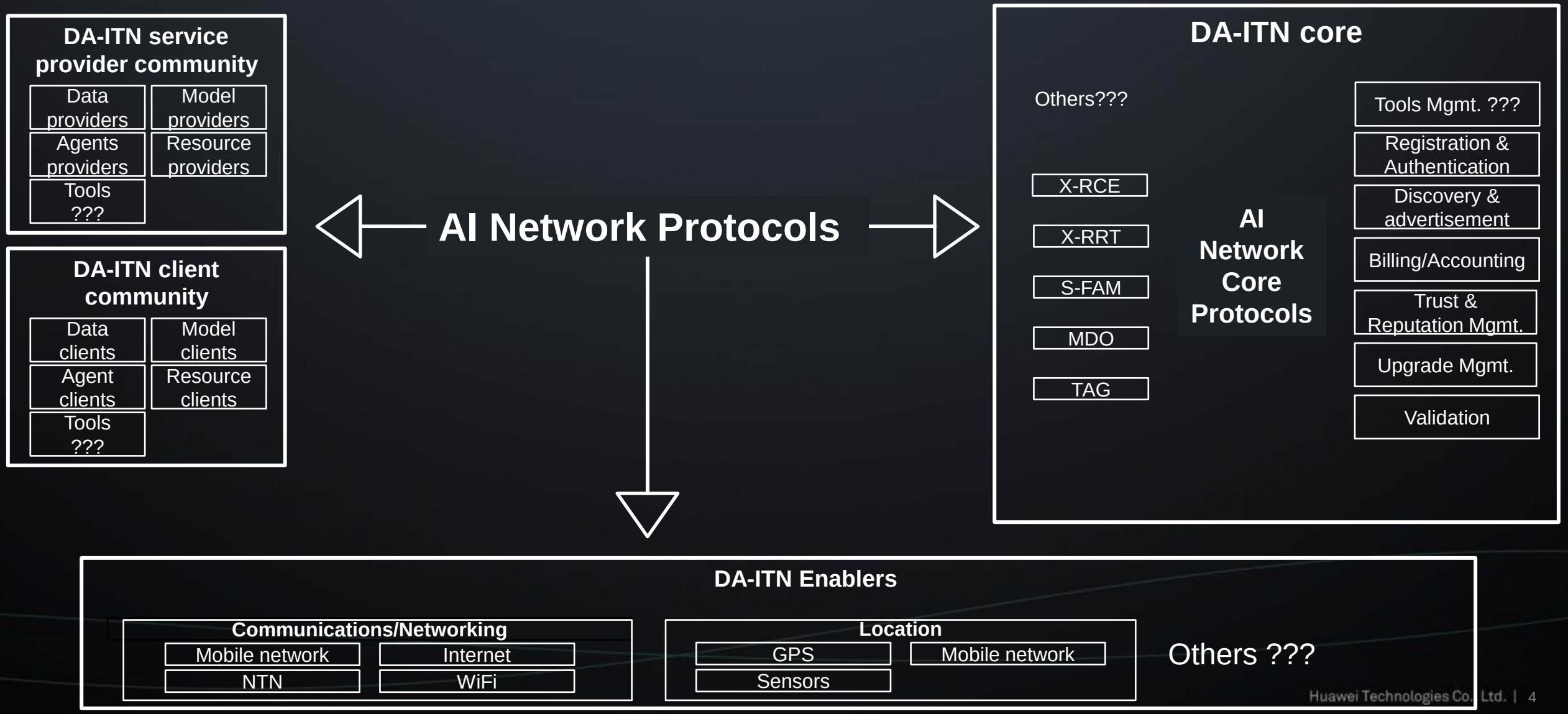
Version 01

DA-ITN: High Level View



- DA-ITN delivers a unified control, data and OAM planes to support different AI related tasks.
- Training, inference, and agent-to-agent collaboration/interaction can potentially have specific sub-planes within DA-ITN control, data, and OAM planes.

DA-ITN architecture, building blocks, and protocols



DA-ITN architecture and building blocks

- **DA-ITN core**
  - Consists of network's main modules, functions, and services.
  - Dedicated logical planes handle interactions between DA-ITN different core modules and functions. **Focus of AI Network Core Protocols.**
  - DA-ITN core interaction may or may not be visible/accessible to other blocks.

- **DA-ITN enablers**
  - This layer provides services that DA-ITN itself requires. Some examples are
    - A device/equipment hosting dataset, model, or an agent in mobile network might become idle over time. DA-ITN uses the underlay service to wake the hosting node before dispatching a job to it.
    - Tracking services offered by mobile networks or GPS.

DA-ITN architecture and building blocks

- **DA-ITN service provider community**
  - Different providers reside within this block.
  - Providers join the network and become members via a registration and authentication process offered by DA-ITN core.
  - They use DA-ITN to advertise their services, capabilities, etc. across the network. They can register for notifications to get updates e.g. arrival of new models, training data, agents, etc.
  - DA-ITN dispenses revenue to providers for the services rendered via its billing module.

DA-ITN architecture and building blocks

- **DA-ITN client community**
  - This block consists of clients requiring training, inference, agent-to-agent interactions, and those who need access to resources such as storage, compute, etc.
  - Clients access services offered by members of the DA-ITN service provider community and/or DA-ITN core via different DA-ITN interfaces.

AI Network Protocols hand the interactions between DA-ITN core, communities, and enablers

DA-ITN architecture and building blocks

NOTE:

- A node/entity can be both a provider and or a client.
- A data node provides data, but might need access to a resource provider service.
- A model provider enables inference for clients and at the same might use data providers for Retrieval-Augmented Generation (RAG).
- A node can host an agent that requires further training to finetune its skills.

AI Network challenges

- AI network poses unique challenges. Those related to agents interactions and collaborations have been outlined in a proposed charter in agent-2-agent mailing list.
- The followings are some considerations related to training and inference aspects of AI network.

Training

- Advertisement and discovery service:
  - Providers need to be able to advertise information about their data and resources.
  - Clients need to able to discover this advertised information.
  - Clients can even advertise their needs and discover providers.
    e.g. Advertised dataset information can be used by training clients to identify relevant datasets which can optimise centralised or distributed training. This is applicable to resources(compute, storage) as well.

AI Network challenges

Training

- Data utility and governance:
  - Data providers can express the way their data should be utilised. This can be of interest to AI Pref and ETSI DATA TC.
- Privacy protection and trust establishment :
  - Discovery and advertisement may convey sensitive and private information that must be adequately protected.
  - The protocols might require to convey trust levels and ratings associated with models and datasets. Clients and providers can use this information to conduct risk assessment and express approval or denial of service.
- Session management:
  - AI training clients and service providers need exchange configurations such as hyperparameters, prompts, and performance data as well as large objects such as datasets, models, and outputs, which might result in lengthy and continuous interactions. Reliability and data transport related aspect should be considered.

AI Network challenges

Inference

- Advertisement and discovery service:
  - Model providers need to discover resources for initial model deployment and subsequent instantiations of models if required. Resource providers of course need to be able to advertise their capabilities.
  - Models need to advertise their capabilities for query routing. Frontend entities (e.g. load balancers) need to be able to discover this information for scalability, reliability, and perhaps QoE.

- Session management and reliability:
  - Inference can be a lengthy process. Clients and model status need to be monitored as during this process either side might drop out (e.g. equipment failure, com link issues, etc.).

AI Network challenges

Inference

- Privacy protection and trust establishment:
  - Discovery information, advertisement, queries, and query responses may contain sensitive and private information that must be adequately protected.
  - The protocols might require to convey trust levels and ratings associated with models and queries.

- Dataset discovery:
  - Models providing inference services might need to discover new data to ensure uptodate responses (e.g. Retrieval-Augmented Generation).
  - The above point is also applicable to agents.

Seeking feedback

- The proposed AI-Agent Protocols Working Group focuses on the agentic dimension of the AI Network.
- Training and inference in AI ecosystem have requirements similar to those outlined for the agents in the proposed charter.
- Protocols such as SLIM can be extended to support training and inference.
- Potential options:
  - Broaden the scope of the proposed WG by branding it as AI Network Protocols WG to cover training and inference in addition to the agentic aspect of the AI network.
  - Write a different charter for a potential BOF and WG to cover training and inference aspects of the AI network specifically.
    - New mailing list?
- What are some potential deliverables to cover training and inference aspects of the AI network?

Other potential discussion points

- DA-ITN is envisioned to consider DATA from the AI point of view ONLY.
  - Do we need to extend this view to an overall data governance point of view?
  - That is, should DA-ITN act as data brokerage for non AI applications? (ETSI DATA-TC)

- Tools require further investigation. Need to understand the scope.
- Agentic protocols such as MCP for example provide access to MCP tools.
  - Would DA-ITN require to provide additional tools? What would they be?
  - Do training and inference require tools that DA-ITN can offer/provide?
  - Does DA-ITN have a “tool community” block?

- What about mobility? Does it require special handling at the AI network layer?
- What type of interaction with the underlay might be required?
  - Wake up dormant/idle models, agents, data nodes?

Future topics

- Identify IETF and IRTF related work items.
- Control, data, and OAM planes.
- Core components interactions.
- Discovery and advertisement mechanisms.
- Data, model, agent, and resource descriptors.
- Required interfaces enabling inter block interactions.
- Registration and authentications
- Trust, reputation, and validation.
- Billing, and accounting.
- Upgrades.
- Etc.

Thank You.