Data Aware Inference & Training Network

A unified, intelligent, multi-plane architecture for AI learning, inference, and agent interactions

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

Huawei Technologies Canada

IETF 123 - Madrid
July 2025

IETF Draft

https://www.ietf.org/archive/id/draft-akhavain-moussa-ai-network-00.txt

IEEE network magazine:
Distributed Learning and Inference System: A Network Perspective

Background

- **AI expanding footprint:**
  - Driven by LLMs: ChatGPT, Claude, Grok, DeepSeek
  - Applications: Editing, data analysis, healthcare, coding, etc.
  - **Success = robust Training + Inference.**

  - **Training:**
    - Large datasets, massive compute, high-speed interconnects.
    - Centralised vs. distributed (model-follow-data e.g. federated, sequential, etc.)

  - **Inference:**
    - Low-latency, continuous operation.
    - Single-model vs. multi-model collaboration.

Key challenges

- Data dynamics.
- Data and model mobility.
- Discovering distributed data, compute, and models.
- QoS: accuracy, latency, resource guarantees.
- Privacy, trust, ownership, billing.
- Continuous testing, versioning, upgrades.

- **Problem:** Traditional networks are not optimised for the unique requirements of AI systems (training, inference, agent interaction).
- **Solution:** Introduce the Data Aware-Inference and Training Network (DA-ITN).
- **Goal:** A unified, intelligent, multi-plane network architecture for the full spectrum of AI requirements.

DA-ITN Concept

- A unified, multi-plane infrastructure-agnostic network enabling end-to-end AI lifecycle operations.
- **Connects:**
  - Clients, data providers, compute, agents
- **Goals:**
  - Scalability, transparency, accountability

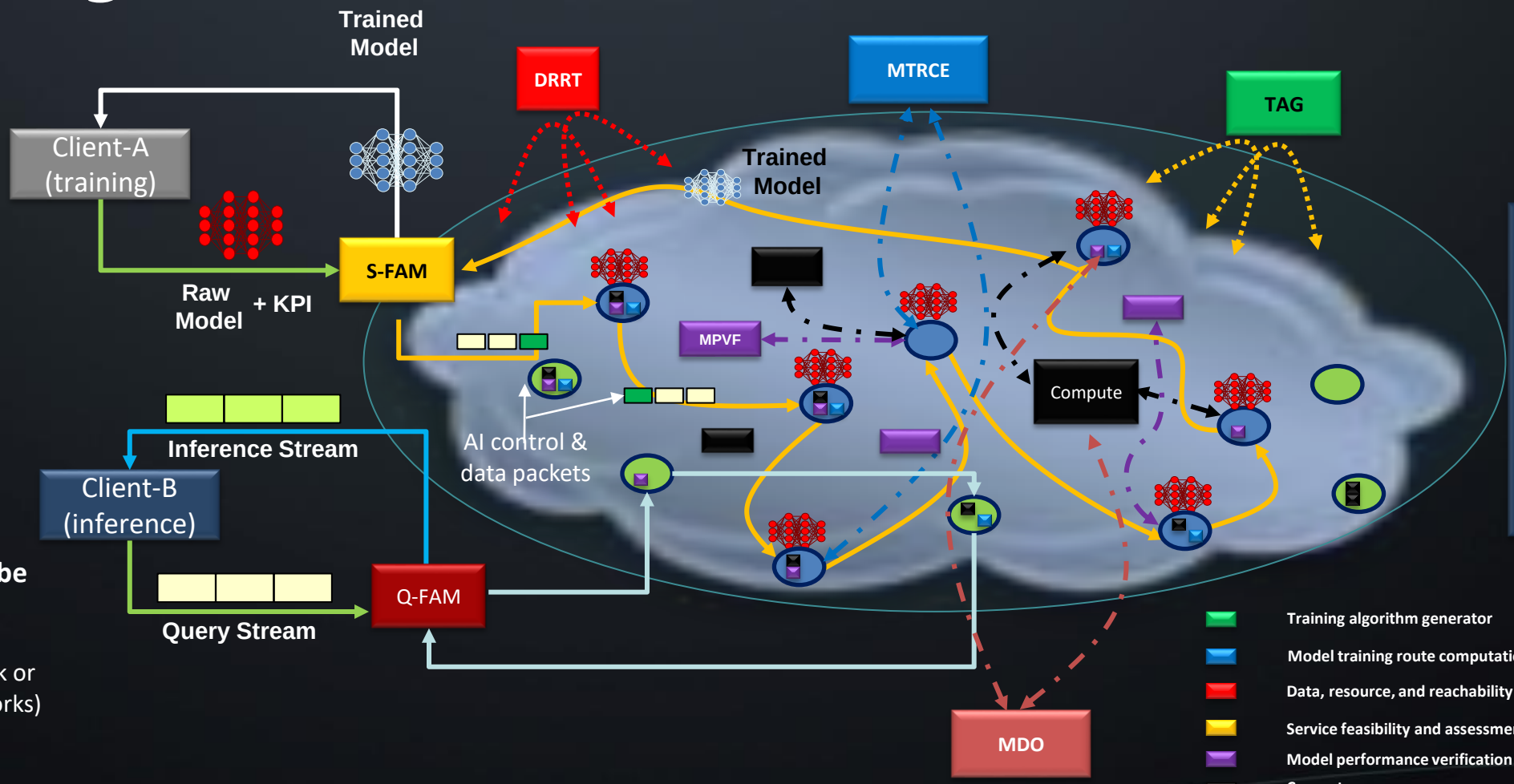
DA-ITN Multi-Plane Framework

- Control Plane + Intelligence Layer:
  - Data, model, and resource descriptors collection
  - Model training route computation engine (MTRCE).
  - Data, resource, and reachability topology engine (DRRT).
  - Discovery, orchestration.
- Data Plane:
  - Model/data mobility, rendezvous scheduling.
- OAM Plane:
  - Monitoring, fault management, lifecycle ops.

DA-ITN: High Level View

Control plane (CP)

- Supports query/response, updates & notifications
- Gathers and distributes
  - Node information
  - Data related information
    - Dataset type
    - Characteristics
    - Attributes,
    - Changes in data quality
  - etc.
- DRR topology management
- Charging
- Joining the network

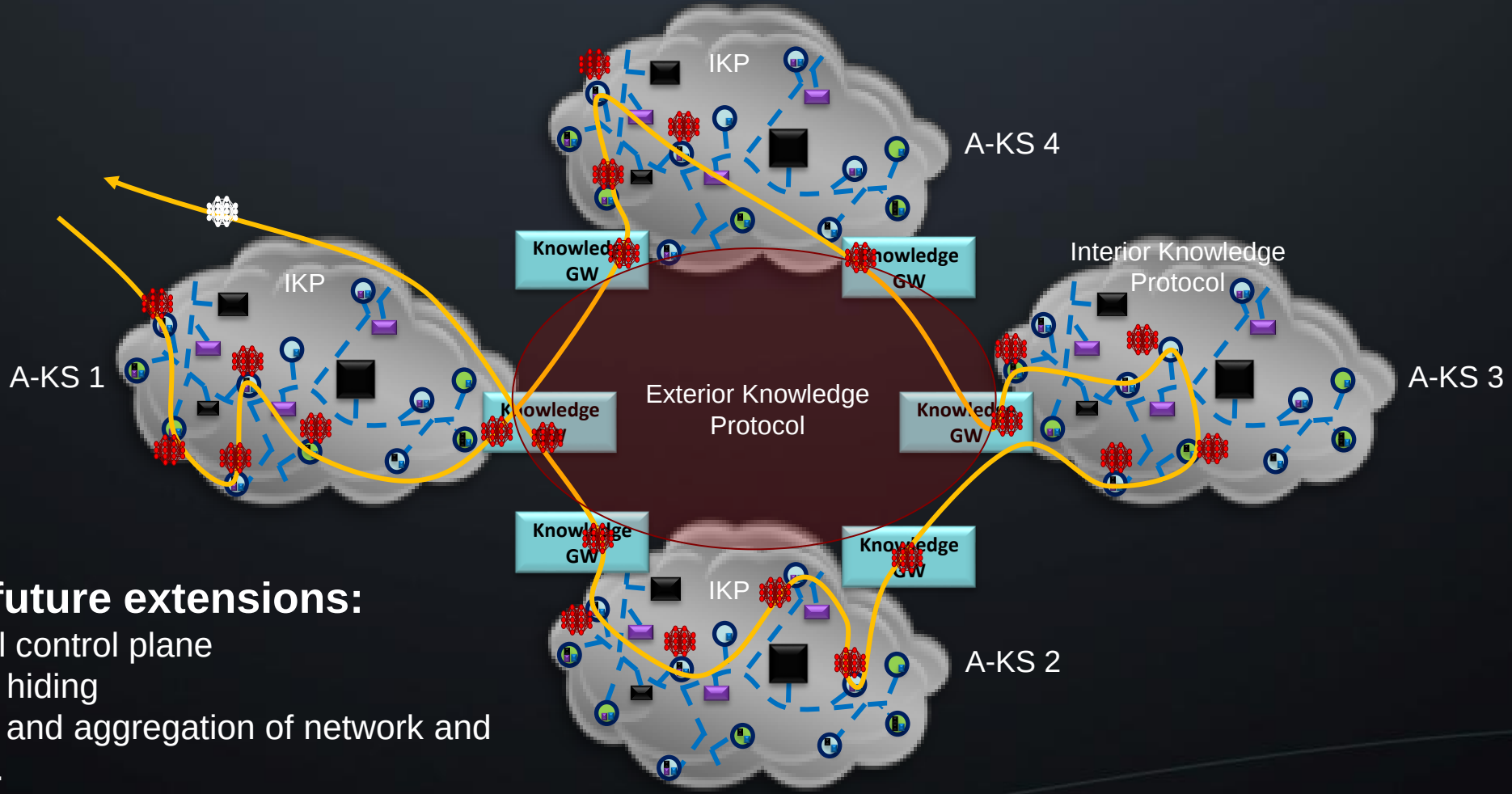


Model/Query routing can be based on:

- Pre-set schedule
- Periodicity nature of network or nodal resources (Tidal networks)
- Submitted request
- Specific KPI
- etc.

- Supports both centralised and distributed architectures.
- Allows support for different training methods (e.g. FL, SL, CL, etc.)
- Requires CP, DP, and OAM plane.

DA-ITN: Hierarchy, Abstraction, and Aggregation



Potential future extensions:

- Hierarchical control plane
- Information hiding
- Abstraction and aggregation of network and information.
- Support for Autonomous Knowledge System (A-KS).

Training Requirements

- **Data Collection & Model Dispatching**
  - Optimise large data transfers vs. model transfers
  - Minimise redundant content, manage Age of Information (AoI)
- **Data & Resource Discovery**
  - Metadata descriptors, cross-domain search
  - Relationship maps for composite datasets
- **Mobility & Service Continuity**
  - Checkpointing, rerouting, fallback resources
- **Privacy, Trust & Ownership**
  - Verifiable descriptors, anti-poisoning guarantees
- **Testing & Performance Management**
  - Distributed test sets, MPVU units
- **QoS, Charging & Billing**
  - SLA-driven KPIs, prepaid/pay-per-use, tiering

Inference Requirements

- **Query & result routing**
  - Cross-domain session mgmt., streaming support
- **Multi-modal collaboration**
  - Chaining, parallel, hierarchical, decentralised
- **Compute & resource guarantees**
  - SLA accountability, remote attestation
- **Utility Governance & QoSP**
  - Digital rights management (DRM)-style rights, usage monitoring, billing

Control Plane & Intelligence Layer

- **Continuously collects:**
  - Data/model/resource descriptors
  - Reachability & performance metrics
- **Builds dynamic topologies via DRRT engine**
- **Enables global discovery, orchestration, billing**
- **Protocol-agnostic: supports ACP, MCP, A2A**

Data Plane

- Manages high-speed transfer of models & data
- Schedules intelligent rendezvous points
- Leverages underlying 6G, edge, cloud, wireline, NTN
- Ensures suitable latency, optimised delivery

OAM Plane

- Lifecycle management for all entities
- Real-time monitoring: convergence, loss, latency
- Configuration, fault detection & recovery
- Policy enforcement, SLA audits, usage logs

DA-ITN in Training

- **Five-layer stack:**
  - Terminal Layer (data hosts, compute, MPVUs, clients)
  - Network Layer (CP/DP links over heterogeneous media)
  - DRRT Layer (global data/resource/reachability topology)
  - MTRCE intelligence layer (orchestration, discovery)
  - OAM Layer (monitoring, control APIs)

Training Workflow

- Client submits model + model descriptors
- DRRT builds candidate data & compute rendezvous points
- MTRCE calculates training path
- DP moves model/data to rendezvous node
- Distributed or centralised training executes
- MPVU validates checkpoint, performance reports
- OAM logs metrics, triggers billing events

DA-ITN in Inference

- **Mirrors five-layer architecture for inference**
  - Discovery of models via CP & descriptors
  - DP routes queries & streams results
  - Intelligence plane orchestrates multi-model flows
  - OAM tracks accuracy, latency, usage, and billing

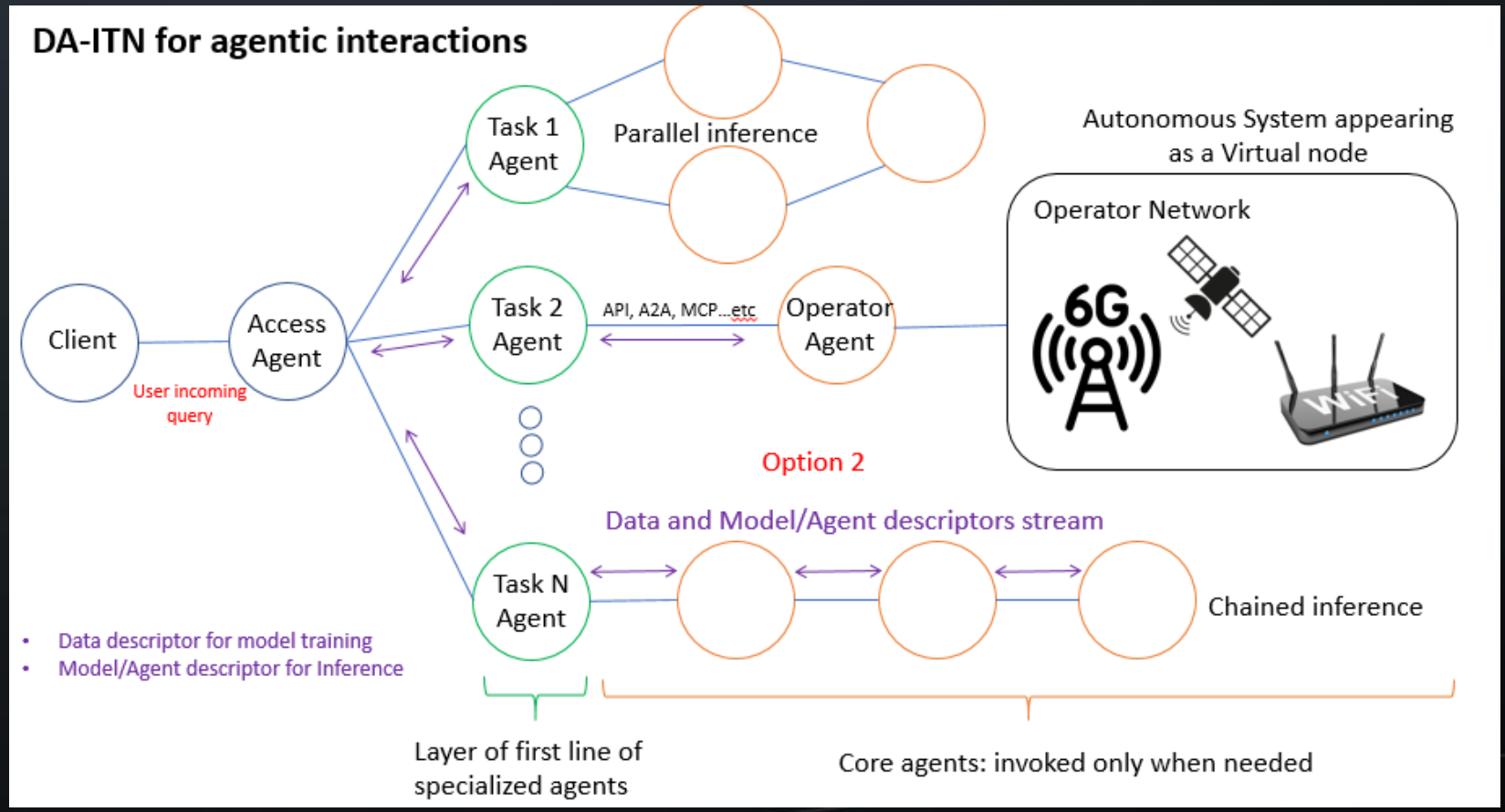
Inference Workflow

- Client or agent publishes task descriptor
- Models self-select or client selects via discovery
- Intelligence plane sets up chaining/parallel graph
- DP establishes session, streams inputs/outputs
- Models return results; orchestration aggregates
- OAM audits performance, enforces SLAs, bills usage

Agentic Networks & Collaboration

A collaborative inference Task

- Access agent recognize the need to build a team of agents to fulfill the task, so it divides the task and assigns each subtask to appropriate specialized agent
- Agent 2 is responsible for discovering the agents (CP of the DA-ITN) and invoking the operator agent to build networks (private data, control and OAM planes) among the discovered agents
- Agents start to collaborate depending on the identified nature of the inference task assigned: chained or parallel



Concluding remarks

- DA-ITN: blueprint for an AI-native Internet
- Addresses end-to-end lifecycle: training, inference, collaboration
- Open research: QoS frameworks, incentive models, security
- Inviting community feedback & prototype implementations
- Define registry for descriptor schemas (data, model, agent)
- Standardize DRRT topology formats & update triggers
- Charter protocols: MCP, A2A, ACP, etc.
- Collaboration with 3GPP, ETSI, NGMN on AI-native APIs

Thank You.