Research Challenges in Management and Orchestration of AI Deployments in Distributed Telco Cloud Infrastructure

Ravishankar Ravindran

Presentation to Nokia, Bell Labs, June, 04, 2025

Abstract

Cloud infrastructure has rapidly evolved from delivering general-purpose compute, memory, and networking services to becoming a foundational enabler of large-scale intelligent applications. At its core, the cloud is a dynamic pool of virtualized computation, communication, and storage resources made commercially viable through hyperscale data centers. These data centers have continuously transformed to support an expanding class of intelligent, data-driven applications, powered by accelerators such as GPUs, DPUs, and TPUs. These specialized processors are now essential ingredients in the construction of modern AI-native cloud platforms.

In this new era, clouds are increasingly architected as AI factories—hyper-converged infrastructures tailored for AI training and inferencing workloads. These AI factories span distributed deployments, from edge to regional and central sites, to meet the stringent requirements of latency, bandwidth, energy efficiency, and personalized services. The application of AI spans nearly every domain—from 6G networks to autonomous driving, IoT, smart cities, agriculture, financial systems, and even neuromorphic computing—making AI-native cloud infrastructure a critical research and engineering frontier.

As we know, Telco cloud infrastructure is also undergoing a foundational transformation—from virtualized network functions running in centralized data centers to fully distributed, edge-to-core cloud platforms owned and operated by telecom providers. Going forward, these operator-managed infrastructures must not only deliver traditional connectivity services but also support a growing class of intelligent, low-latency applications enabled by AI

This talk will explore research challenges in the management and orchestration of Telco AI deployments across distributed cloud infrastructures. We address key issues including:

- Automation using AI Agents for scalable orchestration and management across heterogeneous distributed cloud
- Infrastructure-level optimization to support inference pipelines
- Engineering next-generation cloud fabrics to handle large-scale training workflows

The emerging agentic paradigm—where AI agents reason, plan, and act autonomously across complex systems—requires a rethinking of orchestration and resource management mechanisms at all levels of the Telco cloud stack. The talk examines the efforts in industry forums such as AI-RAN, O-RAN, and Nephio, which are shaping the standards and reference architectures for AI deployments within telecom. Our discussion will highlight open research questions and opportunities for collaboration toward building intelligent, responsive, and resilient distributed Telco cloud platforms.

Bio

Ravishankar Ravindran has over 24 years of experience contributing to advanced data networking products and research. As a Telco Architect at F5, he led the system architecture and design for F5's Telco Cloud platform, supporting 5G vRAN and Core workloads. His work includes active participation in standards development, particularly in the O-RAN Alliance's Working Group 6 (Cloud Architecture and Orchestration), and contributions to the Nephio project under the Linux Foundation Networking (LFN), focusing on Kubernetes-based domain orchestration for Telco use cases spanning RAN, Core, and Transport networks.

Previously, as Chief Architect at Corning Inc., he focused on the architecture and design of disaggregated RAN (CU/DU) for third-party cloud platforms and contributed to the integration of vDU with third-party O-RUs based on O-RAN's Open Fronthaul (O-FH) specifications, with a focus on the M-plane. Prior to that, he served as Chief Architect at Sterlite Technologies (STL), where he was responsible for the end-to-end design of multi-tier RAN Intelligent Controllers (RICs), aimed at optimizing large-scale RAN systems through xApps such as Mobile Load Balancing, Traffic Steering, and Dynamic Spectrum Sharing.

Before this, he led the Future and Network Theory Lab at Futurewei (Huawei Technologies) as a Principal Researcher, focusing on efficient networking for cloud robotics, autonomous vehicles, and drone systems. His research emphasized next-generation networking requirements, including information-centric networking (ICN), software-defined networking (SDN), and network virtualization—particularly addressing challenges in mobility, content distribution, and content-centric routing protocols.

Prior to this role, he was part of the CTO Office at Nortel, where he was a member of the Advanced Technology Group, working on research areas such as control plane routing protocols for IP/(G)MPLS, L2/L3 Virtualization services, scheduling problems in 4G wireless, and end-to-end QoE/QoS engineering for multimedia services. He later served as a Technology Advisor at Avaya.

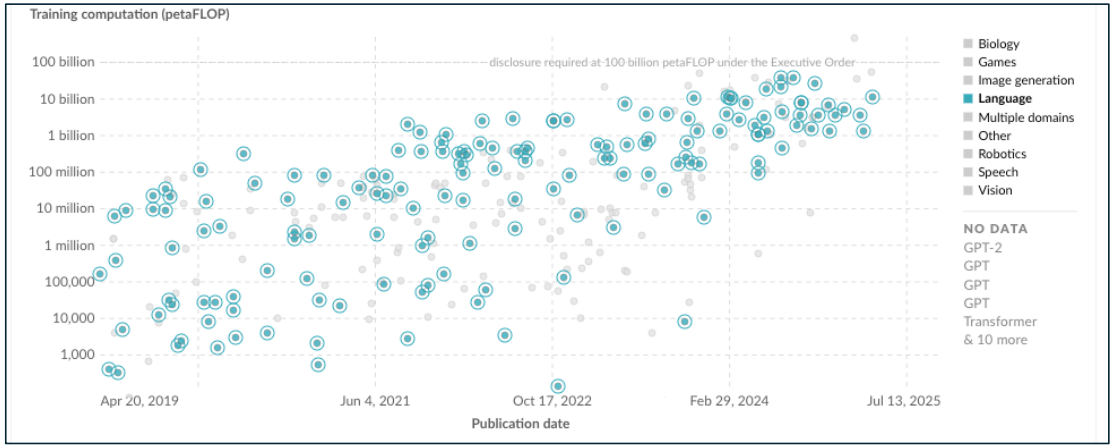
Ravindran has been an active contributor to numerous standardization bodies including the IETF, ITU, ATIS, the O-RAN Alliance, and LFN's Nephio. He participated in the ITU's Focus Group on IMT-2020, helping to define early standards for 5G. He holds a Ph.D. in Electrical Engineering from Carleton University, has served as an editor for the Springer Photonic Network Communications (PNET) journal, and has been part of technical program committees for top-tier conferences. Ravindran is a (co-)inventor on over 90 granted and filed U.S. patents (with additional patents pending) and has authored over 50 peer-reviewed papers in IEEE and ACM venues. His research and patents have been cited over 4,000 times, according to his Google Scholar profile.

Agenda

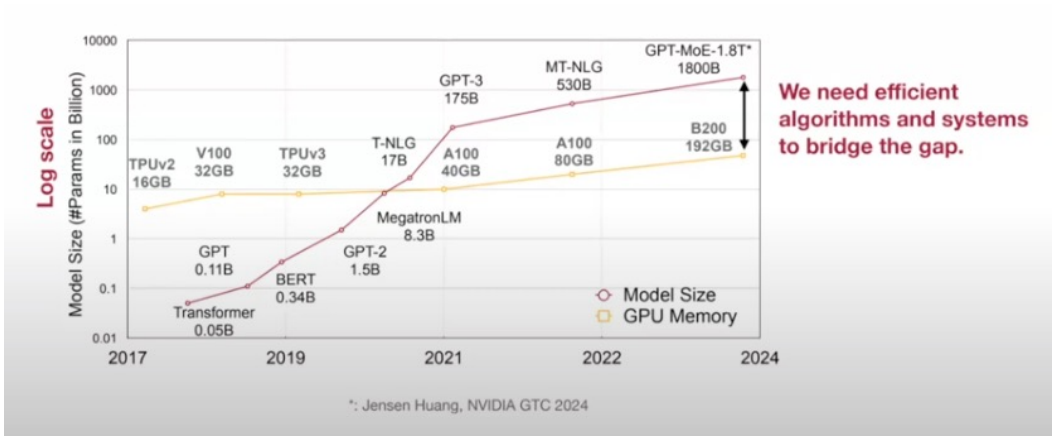
- **General Purpose Cloud to AI Factory evolution**
- **Telco Cloud Evolution and Architecture**
- **AI (GenAI) in Telco Domain**
- **Telco forums with AI Deployment focus**
  - AI-RAN
  - O-RAN
  - Nephio
- **Research opportunities in Telco AI Deployments**
  - Agentic-Driven Service Management and Multi-cloud AI Deployment
  - Cloud engineering for Inferencing pipelines
  - Cloud fabrics to handle large-scale training workflows
- **Conclusions**

Evolution from GP-Cloud to AI-Factories

- **Public Cloud success based on hyperscale-level (Compute,Storage,Networking)-as-a-Service offering.**
  - Pay-as-go-Model
  - Off-Prem Migration/SAAS hosting for Enterprise
  - CAPEX/OPEX efficiency for Telco Operators
  - Efficient Application deployment (Scale, Availability, Security (WAAF/WAAP), DNS, CDN etc.)
- **GenAI is transforming Clouds to be AI Factories**
- **Public Clouds today offer (Models, Training, and Inferencing/RAG Platforms)-as-a-Service**
  - Clouds are turning into infrastructure that hosts intelligent agents to perform domain specific Tasks (improve productive and drive down costs)
  - SAAS services are being replaced by AI Agentic implementations
  - Imperative APIs are being replaced with prompt driven to interface with Agentic based service implementations
  - Allowing easy integration of existing Services with intelligent AI model-driven services in the backend.
- **The LLMs are driving cloud infrastructure require significant GPU/TPU/DPU resources**
  - Figure shows this exponential growth in petaFlops with increasing model size)
  - Abstracted Model Libraries for Scale-up and Scale-Out
  - Efficient LLM data processing pipeline CPU/GPU/TPU/DPUs, Memory (HBM, SSD, Network Storage), NICS, Switches



Ref: https://ourworldindata.org/grapher/artificial-intelligence-training-computation?

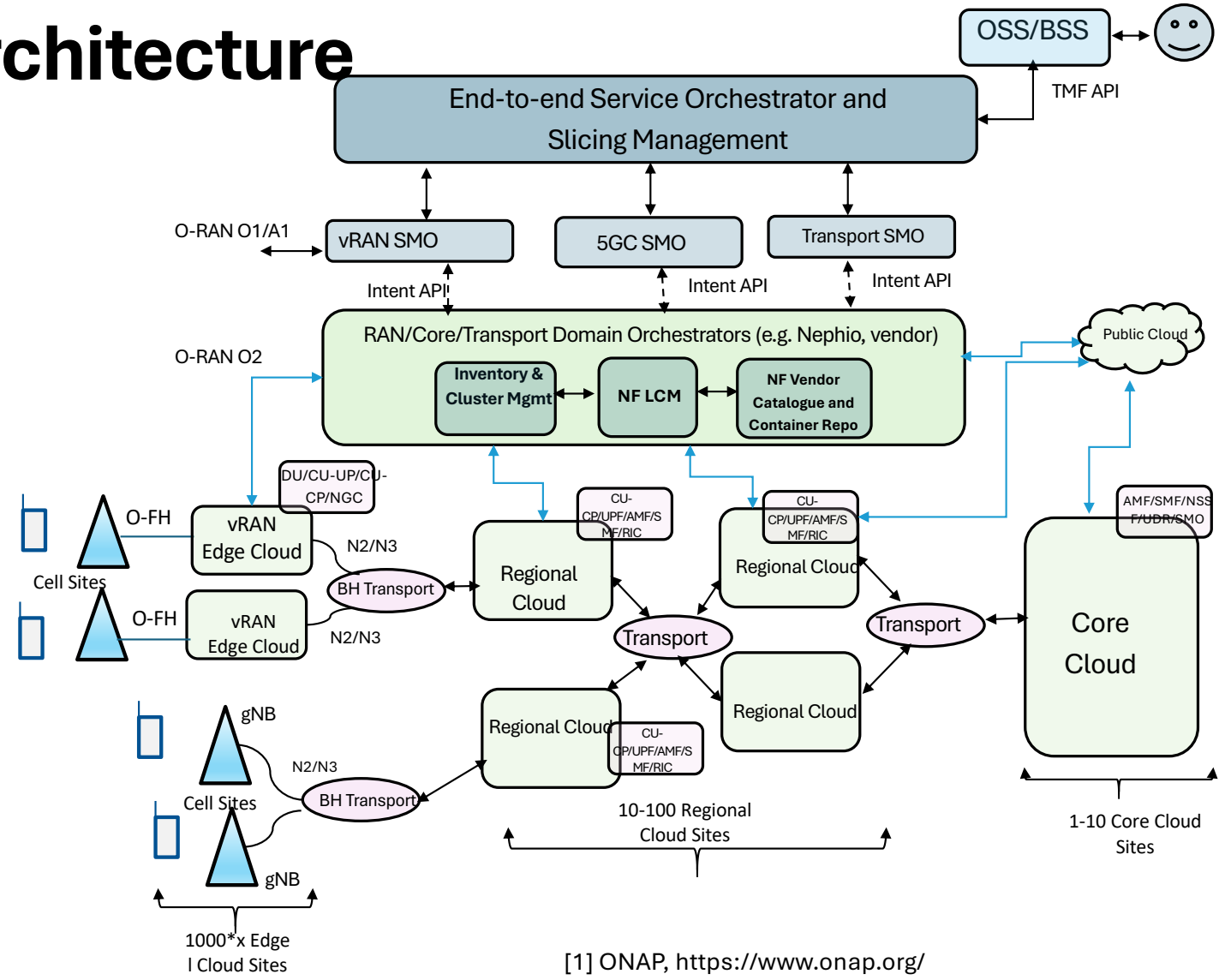


Telco Cloud Evolution

- **Began with the NFV/SDN transition in the industry ~2012**
  - Move from vertically integrated SOC + Software (HW Appliances) to COTS + Accelerators
  - Driven by - Open Source Virtualization Stacks – IAAS (e.g OpenStack), CAAS (e.g. Kubernetes) + SDN Stacks (OpenVSwitch, vRouter etc.)
- **Why Cloud ?**
  - Agile Service Management
  - Avoid Vendor lock-in
  - Reduce CAPEX and OPEX
- **Fixed/Wireless access Network functions (EPC, IMS, BNG) were virtualized first**
- **Today, 5G/NGC NFs in most deployments are virtualized, including the UPF**
- **Telcos adopted ETSI/NFV architecture (MANO ~ NFVO/VIM/VNFM)**
  - Extensions for Cloud Native CNF deployments
- **Open source Service Orchestration frameworks like ONAP have emerged**
- **RAN disaggregation standardized in 3GPP (Rel.15)**
  - CU-CP, CU-UP, DU & RU (Multiple splits proposed)
- **In, O-RAN, virtualized deployment is one of the key focus area**
  - O-RAN/O2 i/f manages the cloud infrastructure and NF LCM.
  - Operators in various stages of deploying vRAN systems

Telco Cloud Architecture

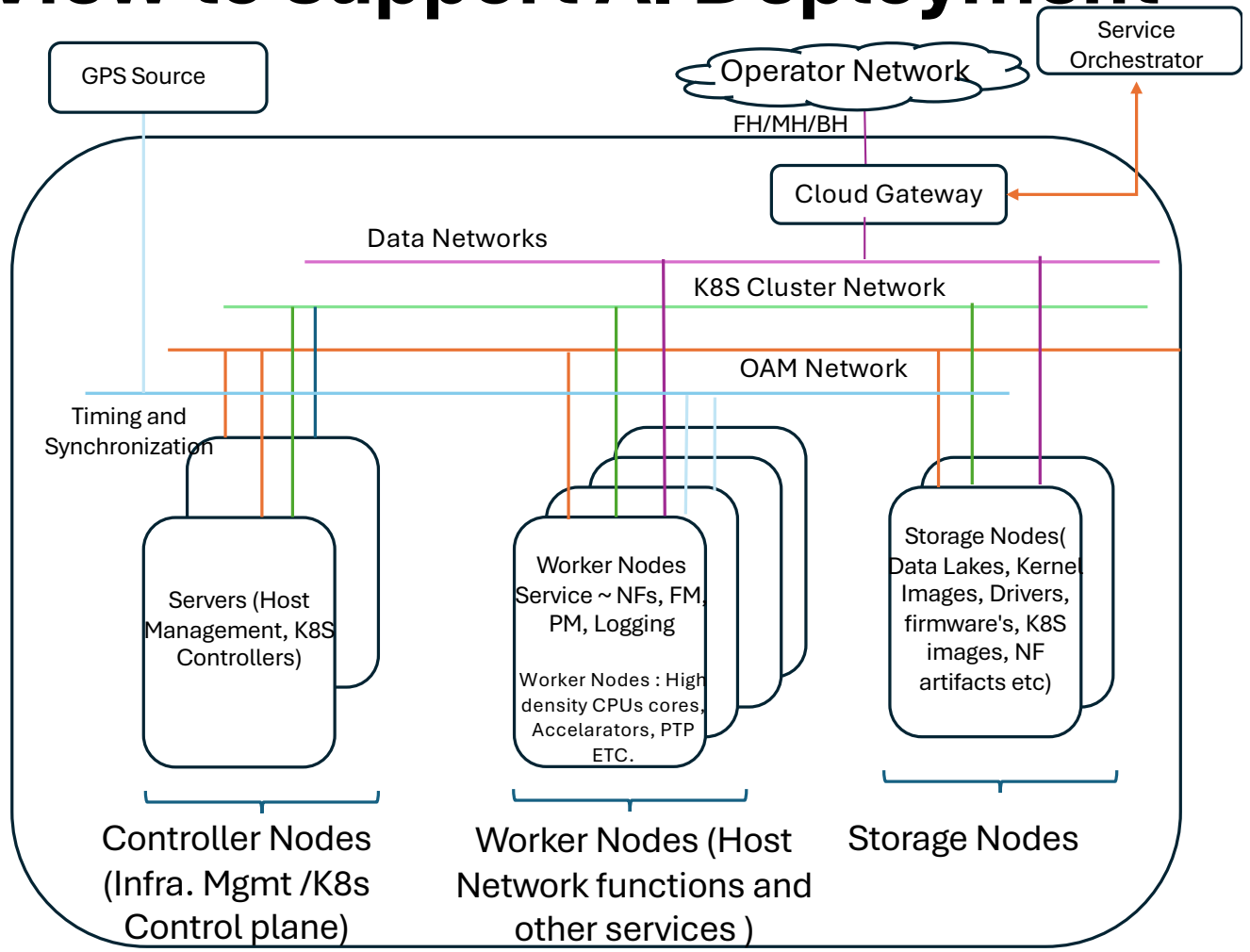
- **Distributed Telco Cloud Systems**
  - **Edge/Regional/Core Clouds**
- **OSS/BSS**
  - Portal for Customers or Service Manager to provision end-to-end Services (Telco), CSMF
- **Service Orchestrator (e.g. ONAP)**
  - Offline Service Design functions (Day 0) & Runtime functions (Service Deployment & Operations and Assurance) (Day 1+)
  - Network Slicing (NSMF/NSSMF)
  - Manages end-to-end lifecyle of 5G Services
    - Embb, URLLC, MMTC) with SLA Policy (Bandwidth, Latency, Jitter, Loss), Monitoring, Optimization, Data Management
  - Deployment scenarios – Urban/Dense/Rural/Private 5G etc.
- **Domain Orchestrator (e.g. Nephio)**
  - Service agnostic, Translated Intents to Cloud, NF, Transport level configurations
  - Multi-Cloud infra/Cluster Mgmt and NF LCM
  - Edges/Regional/Central Clouds
  - Supports NF Vendor Catalogue and Container Registry Management (CSAR/Helm/Container Binaries)



[1] ONAP, https://www.onap.org/

Telco Cloud Site View to support AI Deployment

- **Telco Cloud Site (Edge/Regional/Central) built for:**
  - High Performance (Deterministic guarantees), RT Kernel
  - Physical separation of Servers for Management/Control and NF hosting
  - High Availability (1:1, 1:N)
  - Fault Tolerance (Local Storage, Live migration)
  - Security (Air-gapped or limited Internet)
  - Multi-Networking & Accelerators
  - Timing and Synchronization
- **NF Requirements**
  - Secondary networks (DPDK/SR-IOV), NUMA aware, CPU pinning, huge pages, Accelarator slice etc.
- **Leverage Cloud-native principles (GitOps, CI/CD etc.) for NF LCM**
- **AI Deployments means**
  - These sites to support AI SW Stack on CPU, GPU, TPU, DPU resources to handle AI Deployments (inferencing and training)



[1] Starlingx, https://www.starlingx.io/

AI (GenAI) impact in Telco

- **Why AI in Telco ?**
  - Lots of data, from UE, Network Equipment's, Network operations, Customer service etc.
  - Contextualize services to specific deployment scenarios (e.g. Rural/Urban/Sub-urban etc.)
  - UE-centric Optimization (Differentiated SLA, Embb, URLLC, MMTC)
  - Capex and Opex Efficiency (Power efficient operation, Data-driven Resource Management)
- **Telco's have implemented AI based system for non-critical operational functions**
  - Network Planning (capacity planning)
  - Datacenter/Base station Energy Savings
  - Predictive maintenance (inventory management)
  - Customer Service Assurance (root cause analysis)
- **Automation of Service/Network Management is the goal (zero touch)**
  - Orchestration functions are Day/0 function where LLM can help to optimize design/provisioning time
  - Mission critical functions such as Service/Network Optimization (e.g. SON/RIC) requires handling of error-prone recommendations and hallucinations (llms)
- **Considering GenAI (LLM) use cases, LTM [1] are being proposed and developed to help with different aspects of Telco Stack and Operations**
  - Low hanging fruit - scenarios a human is required to generate/analyze textual data – Configurations, Logs, Training, Customer support
- **Unlike Internet corpus, Telco AI has domain specific hurdles:**
  - Lack of large standardized data sets (vendor specific, O-RAN tries to standardize this)
  - User data has security and data privacy (regulations)

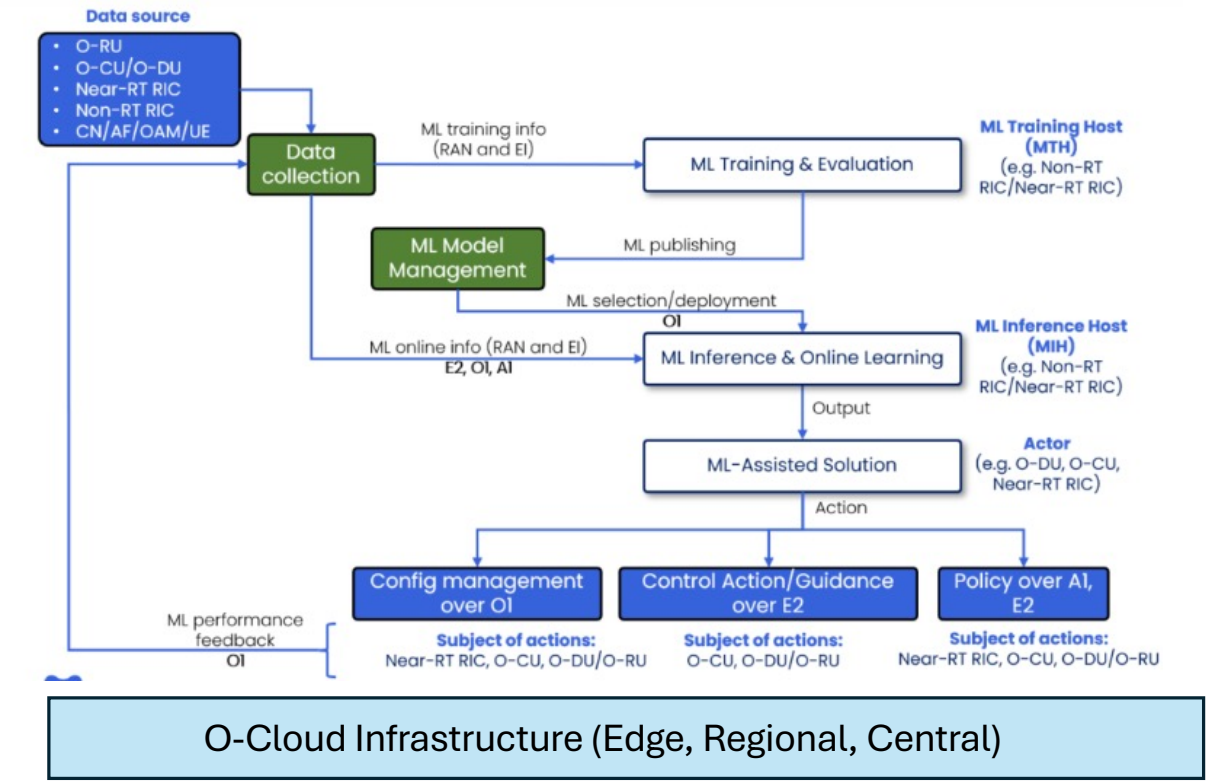
[1] GenAINet Emerging Technology Initiative, "Large Scale AI in Telecom, Charting the Roadmap for Innovation,Scalability, and Enhanced Digital Experiences", White paper, 2025

Telco Forums involved in AI Deployments
- O-RAN, AI-RAN, Nephio
- Others (3GPP, ITU, LFN etc.)

AI Deployment in O-RAN

- O-RAN Architecture principles
  - Open Control/Management Interfaces
  - Cloud Native Deployment
  - Data-centric RAN Optimization
- Envisioned on optimizing RAN deployment based on data from RAN stack and the Cloud infrastructure
- XApps/Rapps for RRM could be traditional algorithms or AI Models
  - > 1s , <1s , ~TTI closed loop optimization
- Recently using Rapps for Cloud Infrastructure optimization
- WG-2 has a blueprint to manage AI model's LCM.
- In the figure, ML Training and Inferencing ca be distributed (Non-RT/Near RT RIC)
- Actors use ML Models can be used by several Actors (O-DU, O-CU, Near-RT RIC (Xapps)

AI Model Life Cycle Management



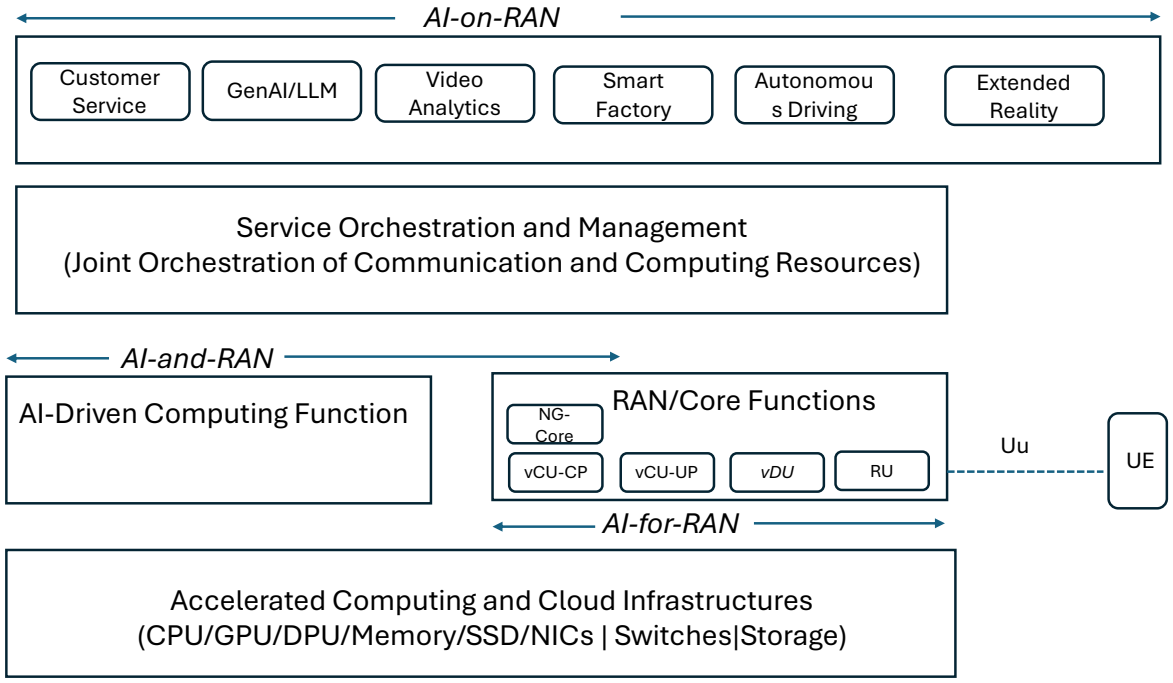
Simplified diagram from WG-2, Technical Report [1]

Source - https://rimedolabs.com/blog/ml-framework-in-o-ran/

[1] ORAN/WG-2, "AI/ML workflow description and requirements", 2021

AI-RAN – High Level View

- **Created in 2024, brings together Cloud Vendors, Operators and RAN Vendors to evolve use cases for AI-RAN**
  Primarily driven by Nvidia and Operators to evolve the use of the GPU resources in the cloud sites for other than RAN workloads
- **Three broad set of use cases:**
  AI-for-RAN
  AI Models for RAN function e.g. optimizing Spectral Efficiency, Beam forming, power management, interference mitigation, ORAN/AI
  AI-on-RAN
  AIaS - Leveraging RAN infrastructure for MEC Services, e.g. IoT, Autonomous systems,
  AI-and-RAN
  Multiplexing common GPU resource with both RAN NFs and MEC Services.

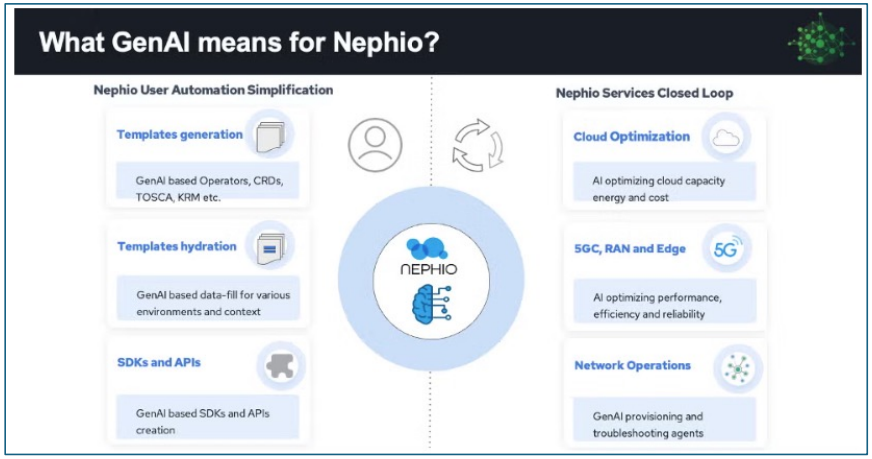


[1] Kundu et al, "AI-RAN: Transforming RAN with AI-driven Computing Infrastructure", (Vision paper from Nvidia, submitted to IEEE for possible publication)

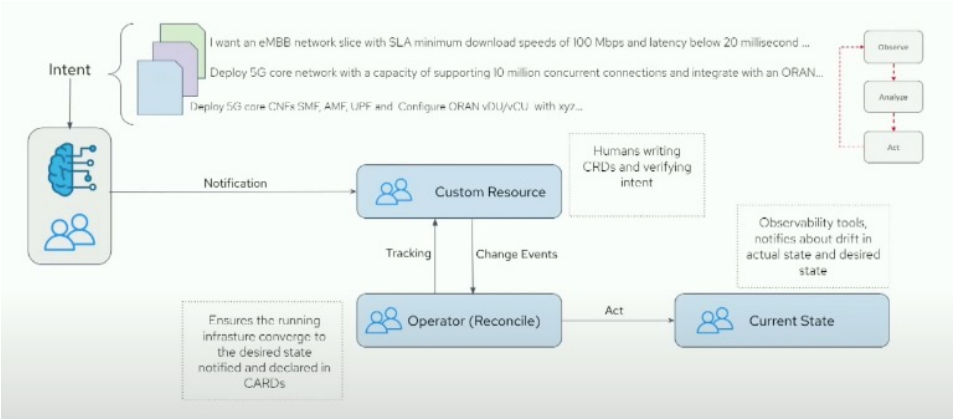
[2] Softbank, "Telco AI : Landscape, Challenges and Path Forward", Feb, 2025 (White paper)

LFN/Nephio and GenAI

- **Design Principles**
  - Intent Driven Automation (what and not how)
  - Kubernetes Pattern
  - Multi-Cloud Flexibility
- **Nephio Use cases**
  - Multi-Cloud Management/Cluster Creation
  - 5GC NF Orchestration
  - O-RAN/CU/DU Orchestration
  - O-RAN/Cloud Management (O2-ims/dms) [1]
  - Transport use cases
- **Nephio and GenAI**
  - Template Generation
    - Service, Cloud, Transport, NF level
  - Templates Hydration
    - Customize the template for deployment scenario
  - SDKs and APIs (e.g. Helm to Operator conversion)
    - Applied at the Cloud Site to map the intents to Vendor specific manifests and configuration



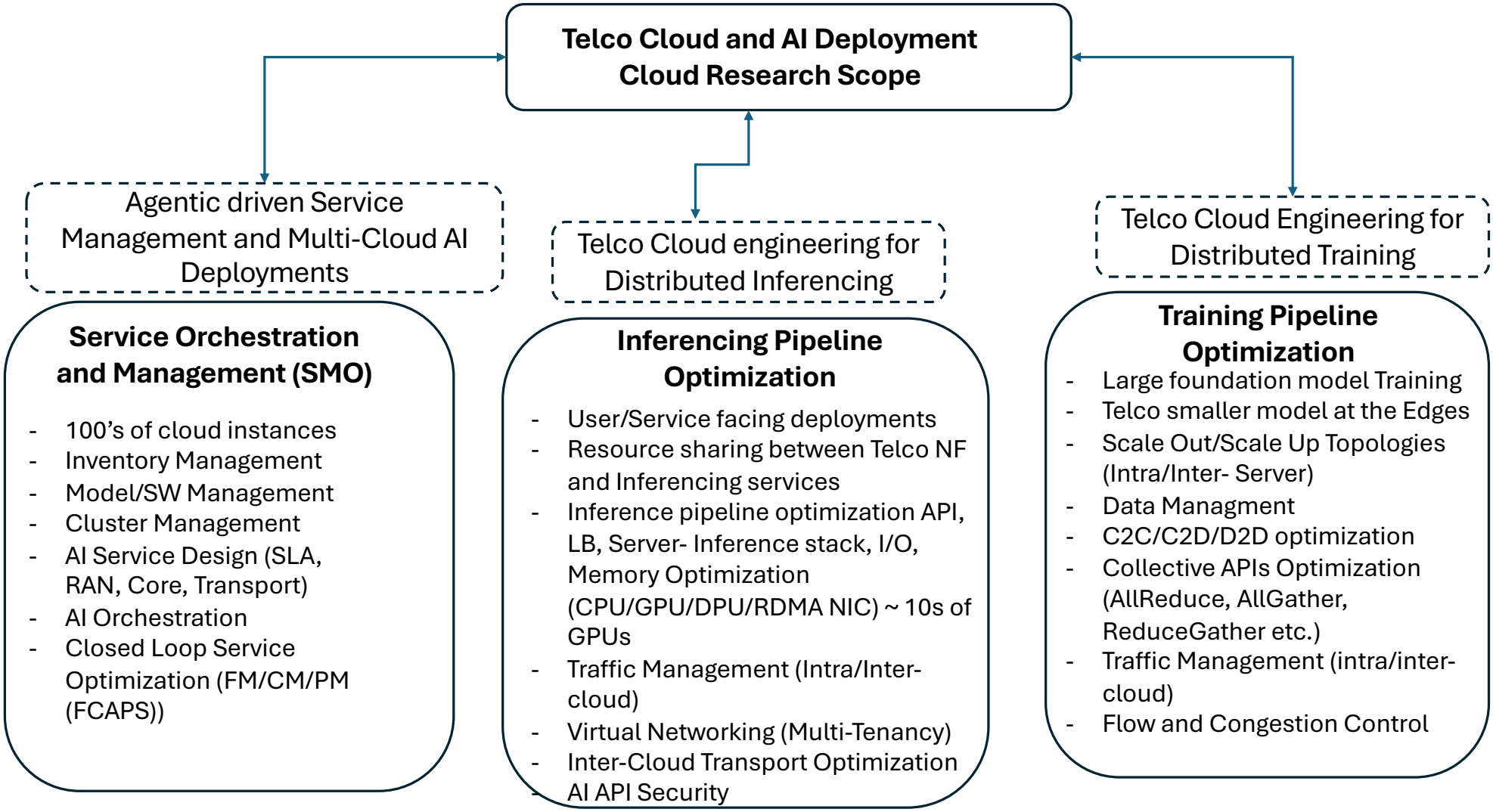
Nephio and GenAI, https://www.youtube.com/watch?v=iOTP6gzvd5w (ONS, 2024)



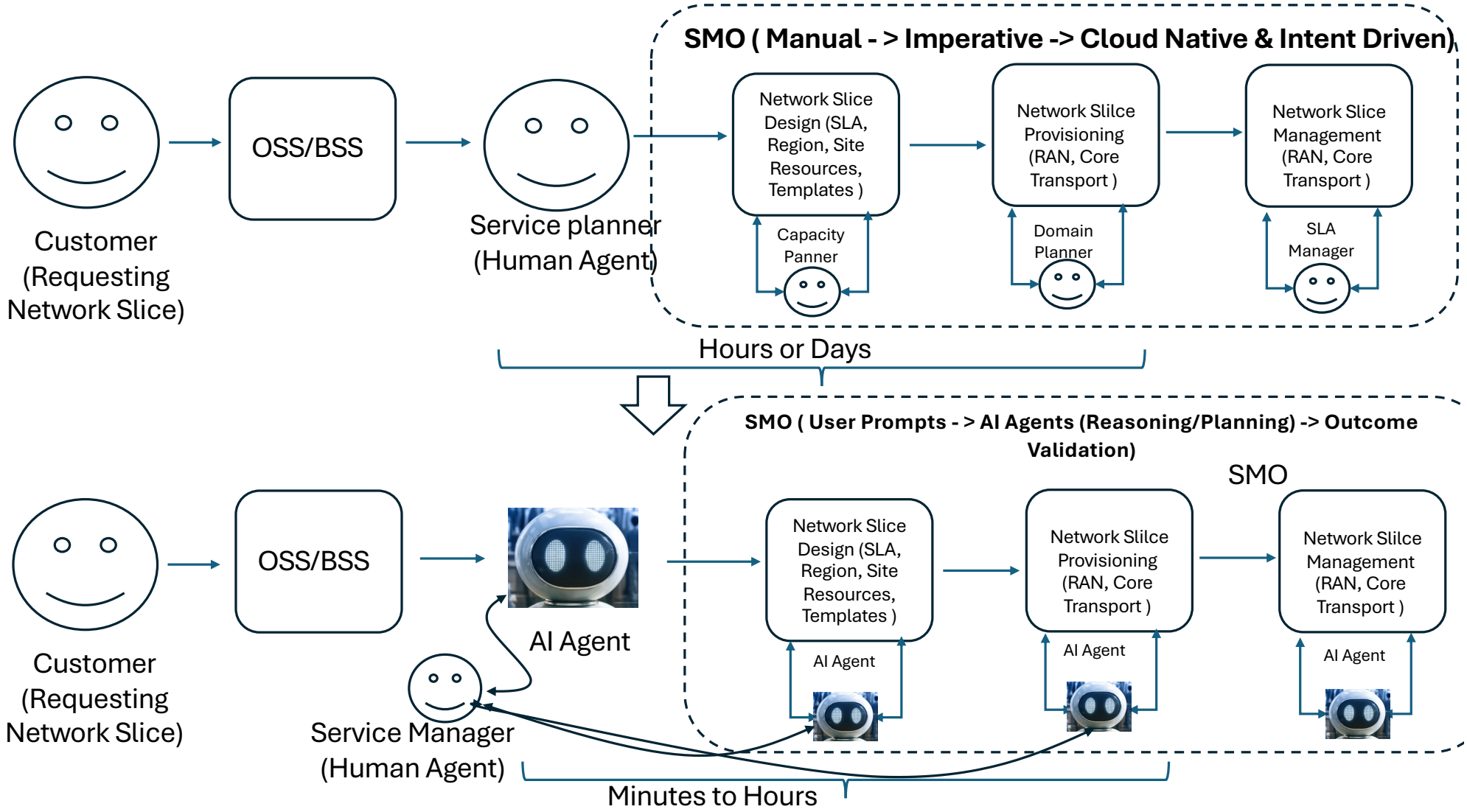
[1] Ravi Ravindran , “Enabling SMO over Nephio, a perspective”, ORAN-SC Workshop, ONE Summit, 2024, https://wiki.o-ran-sc.org/display/EV/O-RAN-SC+Workshop+At+ONE+Summit+2024

Research Challenges in AI Deployment in Telco Cloud

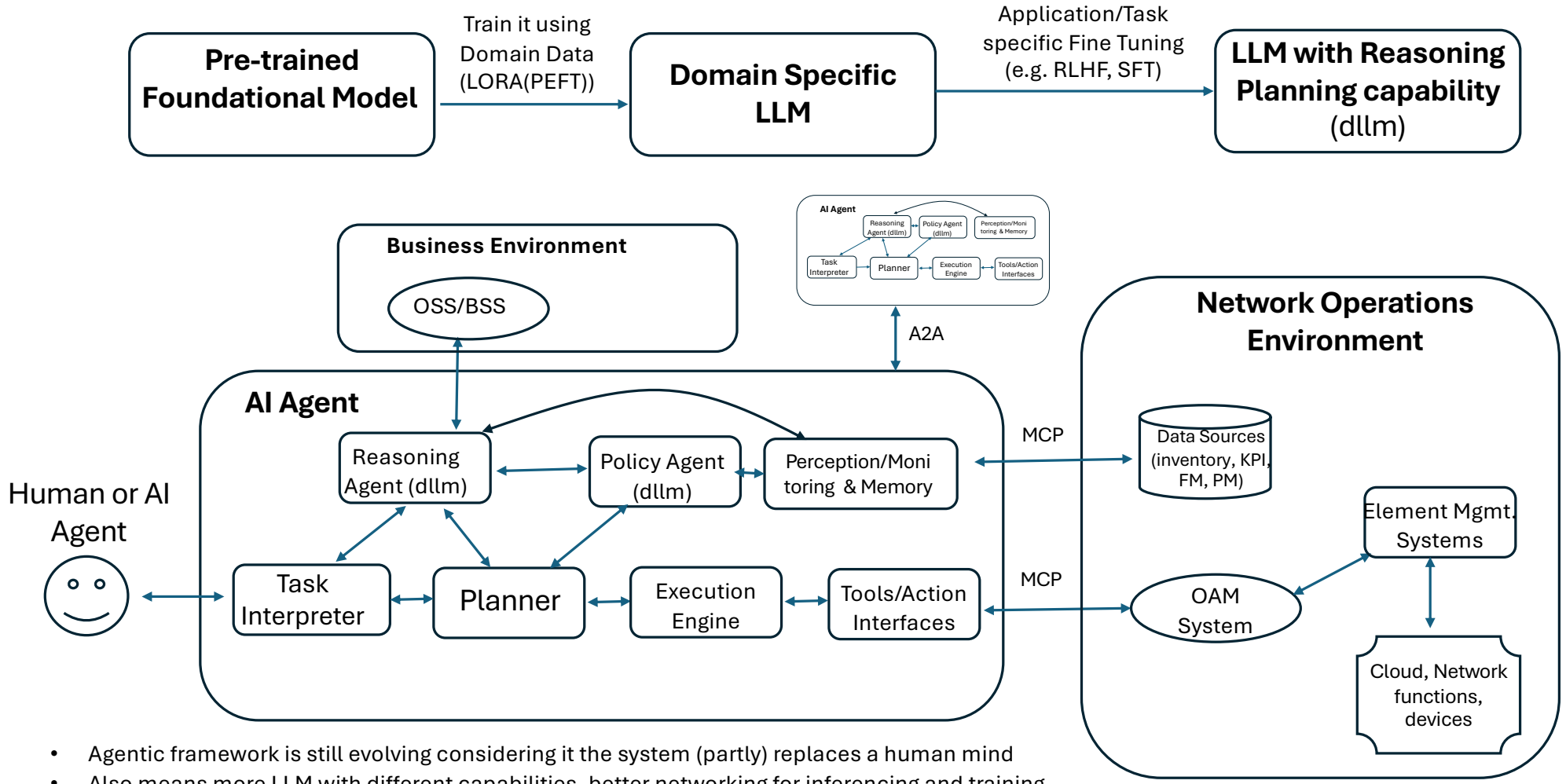
Telco Cloud and AI Deployment Cloud Research Scope



Agentic Driven Service Orchestration and Management



Agentic System View



- Agentic framework is still evolving considering it the system (partly) replaces a human mind
- Also means more LLM with different capabilities, better networking for inferencing and training.

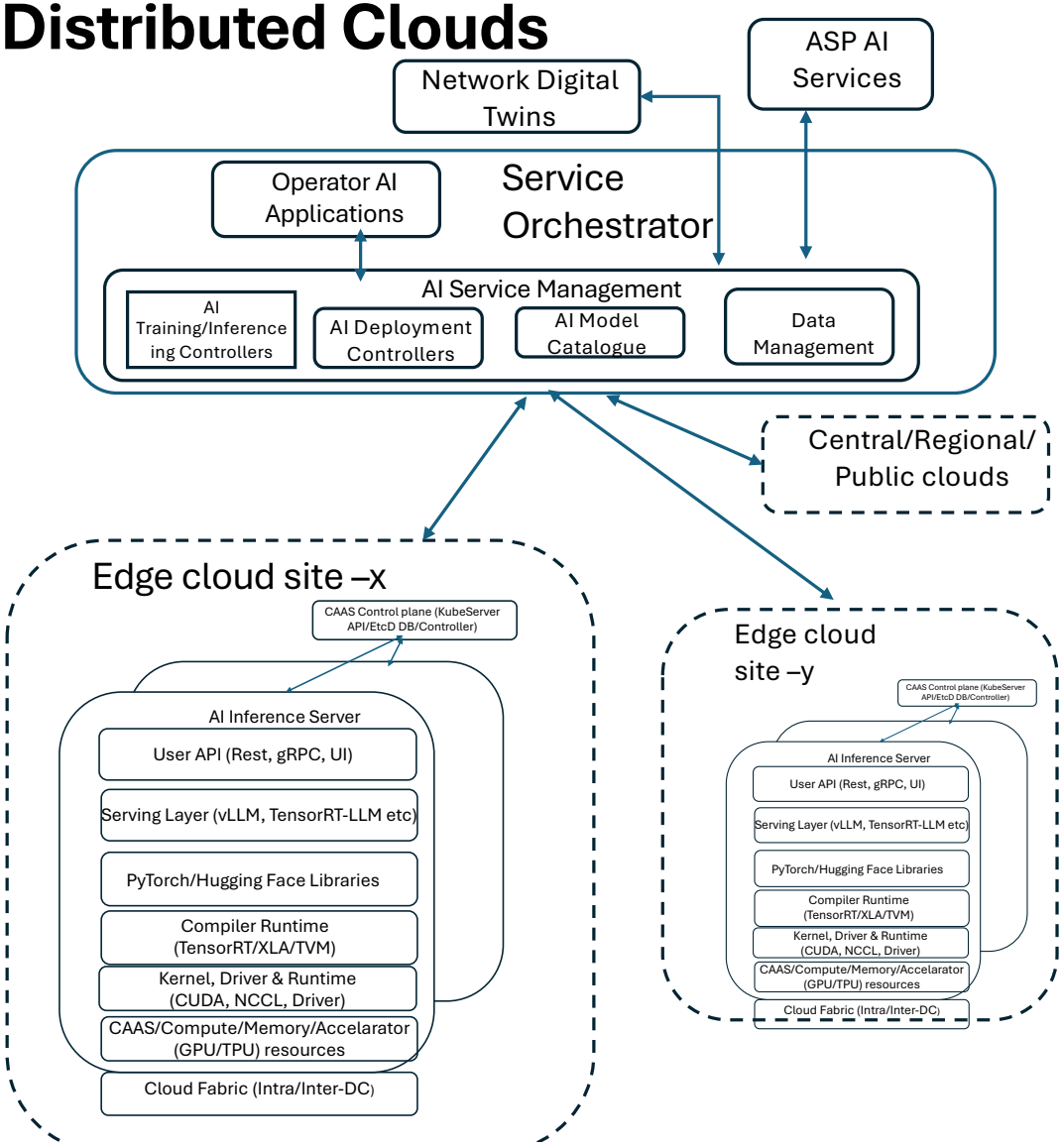
Research in Agentic driven Service Orchestration and Management

- **Agentic based OSS/BSS System**
  - Customer can request to provision on demand interacting with an Ageantic OSS/BSS system, and translate to domain level intents
- **Intent-based domain orchestration using Agents**
  - Not a CR (data model), but a user prompt based interface between SO and Domain orchestration
- **Multi-Domain orchestration Agent**
  - Agent coordinates across domain orchestrators to deploy, scale, or migrate services using vendor specific FCAPS configuration
- **Service Assurance Agent**
  - Agents can monitor SLA/KPI, plan steps to mitigate service degradation by scaling or provisioning services on-demand
- **Service Inventory Reconciliation and Topology discovery Agent**
  - Agent queries real-time systems, reconciles discovered state with intended topology

- Heart of it is an LLM
- Research challenges include:
  - Domain/Task centric Training data
  - Training + Fine Tuning + Task Training to realize Reasoning and Planning LLMs
  - Prompt engineering
  - Domain specific Agent performance benchmarking
  - In multi-agent setup (domain specific inter-agent APIs such as - MCP/A2A)
  - Agent integration with source of truth - Large Tools/API space - CM/FM/PM - Operating in multi-vendor environments

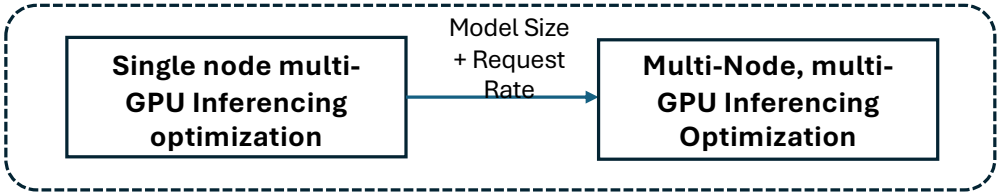
AI deployment considerations in Distributed Clouds

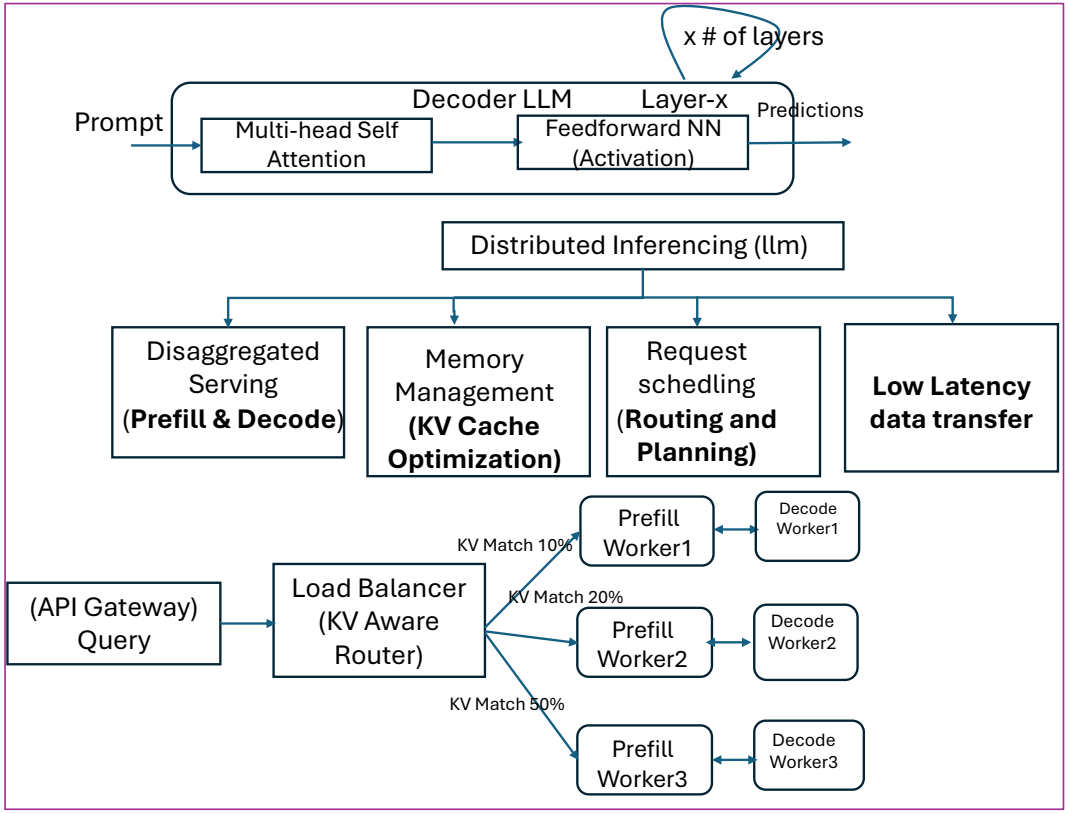
- **AI Service Management layer can be another overlay similar to NF function management functions.**
  - Data models, Service APIs across management layers NB/SB considering inter-operability
- **AI- Workloads + supporting functions could be:**
  - Edge Inferencing, federated Inferencing/training, Data Collection, supporting functions like performance management, DB, LCM of AI models etc.
- **AI Service orchestration/resource management challenges:**
  - *AI model planning* across distributed cloud considering users and applications (e.g. coordination between large and small models in the edge)
  - *Sharing Distributed cloud clusters resources* with infrastructure and AI Workloads
  - *Mission-critical AI workloads* will require dedicated CPU/GPU resources shared with infrastructure services, hence capacity resource planning and resource management is important.
  - *Dynamic resource requirement* will depend on the use cases, model size,
  - *Data exposure functions*, *Data/AI Security* (DDoS on AI models, malicious Prompt attacks etc.), Availability are all critical considerations



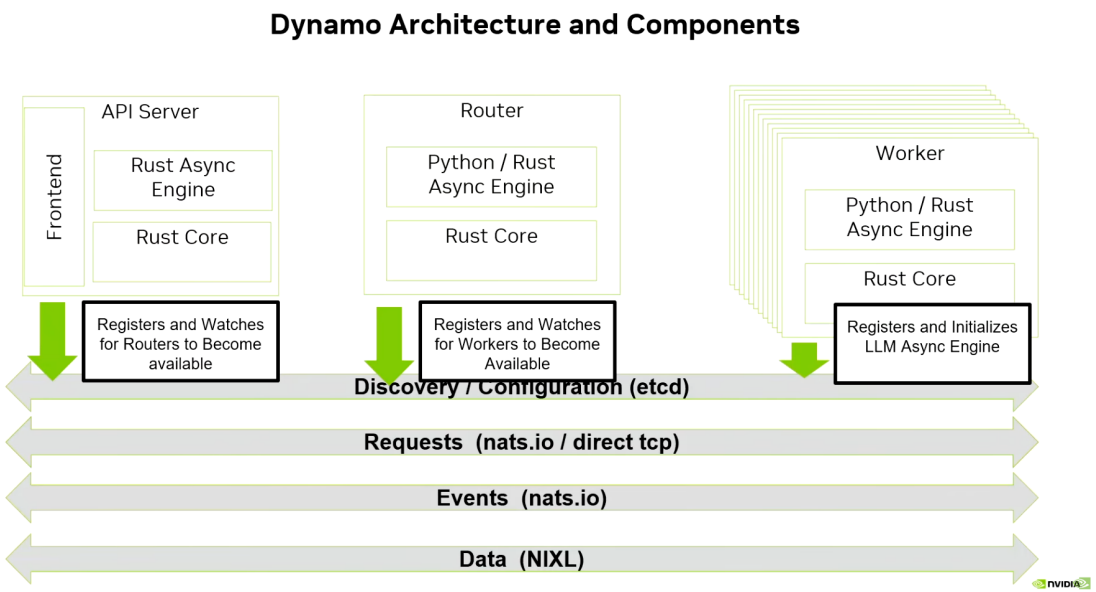
Research challenges for Distributed Inferencing

- **Main key shift is - Models for inferencing distributed across multiple GPUs**
  - DDP, Tensor, Pipeline, MOE based
- **System Optimization –API Server, Load Balancer, Inference Cluster, Network**
- **Inference Pipeline Optimization:**
  - *Characterization of the input and output sequence length*
    - Useful for multi-tenancy resource management
  - *Division of compute between the Prefill Workers vs Decode workers*
    - Metrics - Minimize Time for First Token, Inter-Token Latency , tokens/s/GPU and tokens/s/user
    - Heterogenous GPU types
  - *Key/Value based Routing*
    - API Gateway -> KV Aware Router -> LLM Server
  - *KV Memory Management*
    - HBM, Host, Local SSD, Network Storage
    - Data can be prefetched based on the context (e.g. large context lengths for Agentic interactions)
  - *MoE based models offers another dimension of inference optimization*
    - Since only a subset of parameters or part of the network is active for a given token
  - *Pruning, Quantization to improve memory requirements, provides more KV Cache memory*



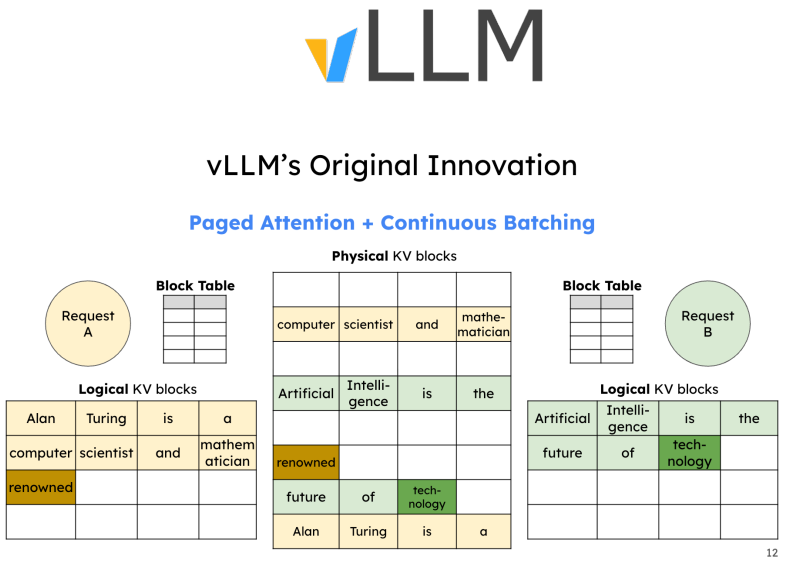


Opensource and Commercial Inferencing Platform



- Auto-discovery for API Server to discover the Router, Routers to discover the inference workers
- Requests to translate client protocols like (OpenAI) to backend standard formats
- Event Bus – Listens and update prefix cache tree based on KV Cache updates, PM/FM/logging of the different service components
- New Data protocol (NIXL) for zero copy between GPU, NIC across the pipeline
- Open source as well

[1] Nvidia Dynamo - https://www.youtube.com/watch?v=3C-6STonTLU&t=556s

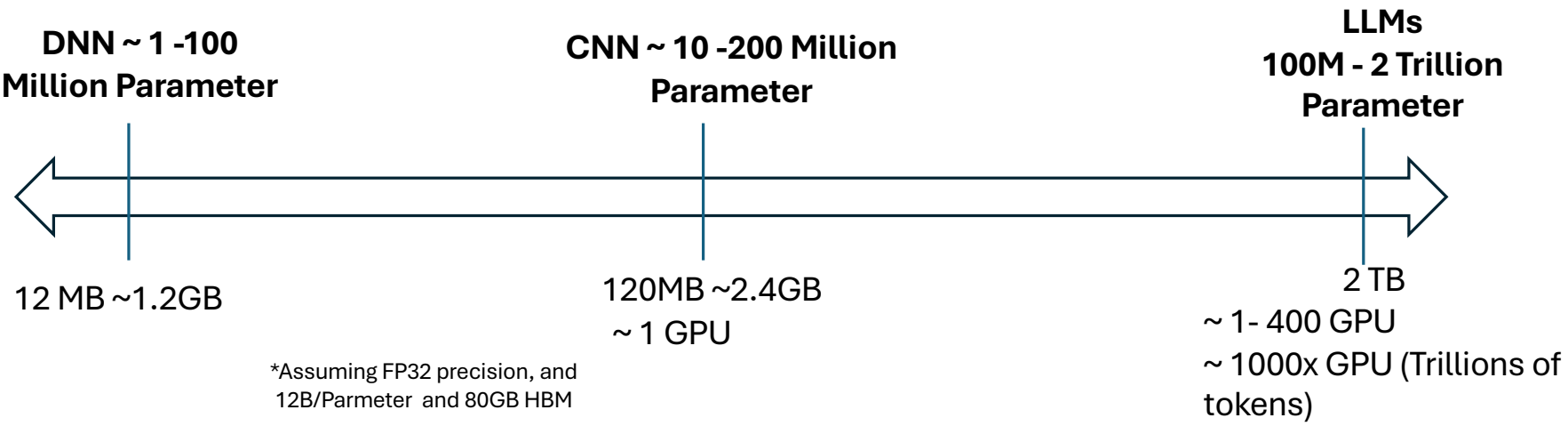


- GPU Auto-discovery and choice of parallelism
- Quantization
- Paged Attention (KV Cache Optimization)
  - Enabling scalable attention using virtual memory
- Speculative decoding
- Continuous Batching
- Supports most open-source models (Llama, Mistral etc.)
- OpenAI-compatible API

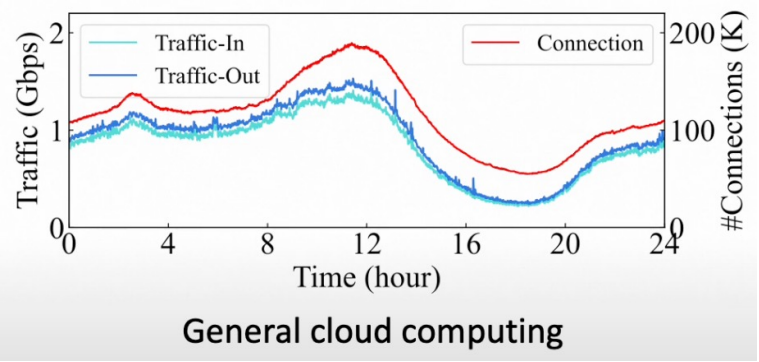
[1] https://github.com/vllm-project/vllm/blob/main/README.md

Engineering Training pipeline for Telco Cloud

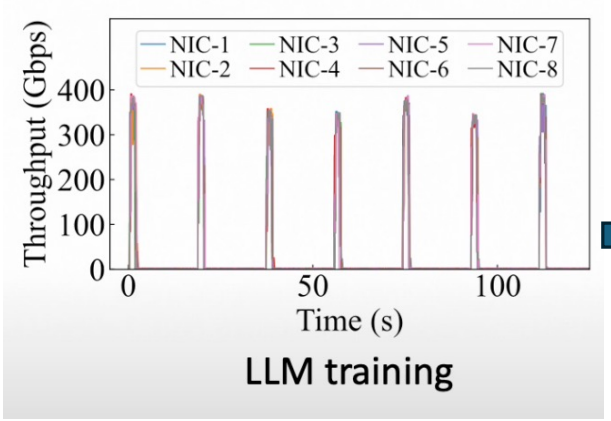
- **Cloud engineering for Training pipeline dependso on the type and size of AI models that are trained**
  - Foundational model developers are training close to trillion parameter or more on Internet data
  - Large LLM models require significantly high GPU resources as required by foundational LLM models (GPT-4.o, Gemini, Llama etc.)
  - AI/HPC environments physically separate networks for Data Management/Checkpointing using CPUs and Training loops involving 1000x of GPUs
  - UEC catering to very specific needs in this space to optimize networking
- **Users of foundational models (open source e.g. llama), fine tune distilled models or RAG pipelines (pvt data) to achieve better accuracy**
- **From Telco perspective, the requirement of large GPU cluster will depend on the end users and services they will be used for.**
  - Fig. gives an estimate for number of GPUs depending on the model type.



Collective Traffic Characteristic



General cloud computing



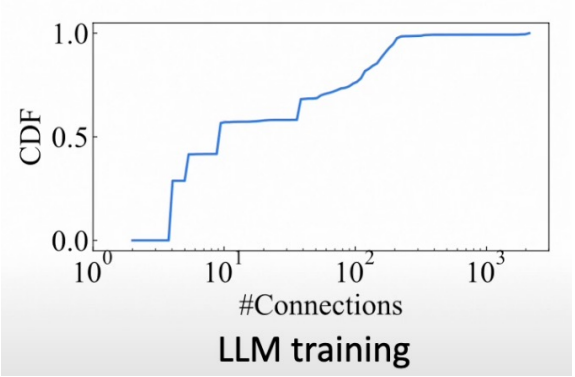
LLM training

- Link Bottlenecks or failure leads to larger JCT hence poor GPU Utilization

- Periodic Bursty Traffic Max NIC BW

Standard Data Center traffic :

- Relatively Continuous and Static traffic pattern
- High number of connection (high Entropy)



LLM training

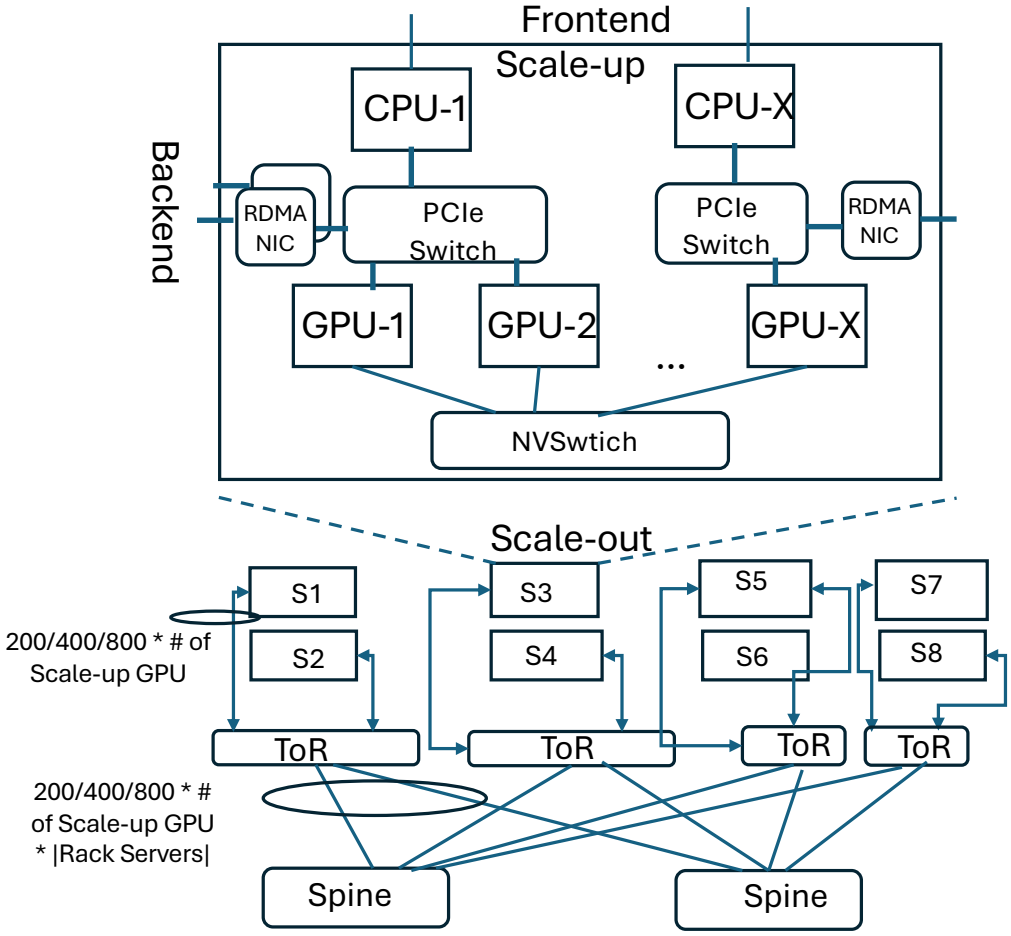
- Low Entropy means traditional ECMP techniques are not efficient

- Fewer number of Connections

[1] Kun Qian et al, “Alibaba HPN: A Datacenter Network for Large Language Model Training”, Sigcomm’24

Research Challenges - Distributed Training in Telco Clouds

- **In the Central/Regional/Edge Cloud**
  - Job involves few 10s to 1000's of GPU lasting for weeks
  - Large LLM models requires Scale-up and Scale-out topologies
  - Multi-dimension Parallelism – DDP, FSDP, Tensor and Pipeline Parallelism
- **GPU-HBM BW significantly more than CPU-Memory, PCIe switch, NIC, Network capacity**
  - Focus on intra-server and inter-server Network optimization to achieve maximum GPU utilization
- **Scale-up Challenges:**
  - Localize computation considering single GPU memory constrains, but very high intra-server Inter-connect
  - C2C, D2D, C2D zero copy data transfer optimization
  - PCI-e Switch Control/User Plan Optimization
  - Derivative protocols (NVLink, CXL, UALink)
  - UALink Forum ~ intra/RACK and inter-Rack upto 1024 GPUs
- **Scale-out Challenges:**
  - Minimize Collective APIs Tail latency (JCT)
  - Clos or other Networking Topologies to minimize hops
    - E.g. Rail optimized topologies
  - Choice of Transport
    - ROCEv2 has several drawbacks for AI/HPC [1]
    - UEC Transport
  - Transport Congestion Control and Incast Congestion
    - Sender driven/Receiver driven Credit based schemes Maximize average utilization
    - Load Balancing/ECMP
    - Packet Spraying with OOO delivery
    - Packet Trimming/In-Network computing
  - L2 Congestion Control
    - PFC
  - Minimize Session state in the NIC (Memory optimization)
    - Ephemeral Connections ~ no 3-way handshake like TCP
  - UEC Forum developing a new Transport layer which can explore multi-path more generously



[1] Hoefler et al, "Datacenter Ethernet and RDMA: Issues at Hyperscale", IEEE Computer, June 2023

Proposed Approaches

- [Meta Sigcomm/24]
  - Low flow entropy, burstiness and high-intensity elephant flows
  - Training Node ~ 8xH100 GPUs + 8x400G RDMA NICs, PCIeGen5 + GPUDirect
  - Topology
    - Physical separation between Frontend (Data source) and Backend (training)
    - Training network - Three Stage Clos Topology (RTSW/CTSW/ATSW) ~ 24000 GPUs
  - Objective
    - Achieve roofline performance (ideal minimal JCT) maximizing the average link utilization
  - Optimizations
    - Switch Buffer Sizing
      - RTSW, CTSW, ATSW
    - Routing enhancements
      - Path Pinning , Enhanced ECMP , Central TE
    - DSCP marking on the RDMA control packets
    - Congestion Management
      - DCQCN & Receiver-driven using Collective library
    - Forwarding enhancements on the switch
      - FlowLet based switching + ECMP (future exploration)

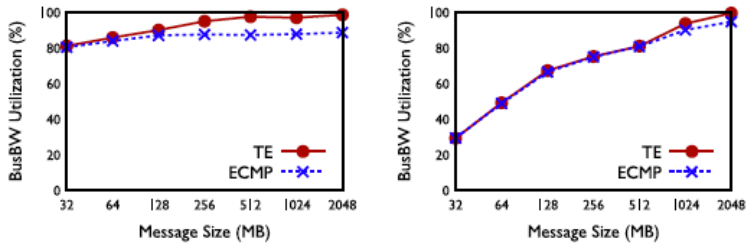
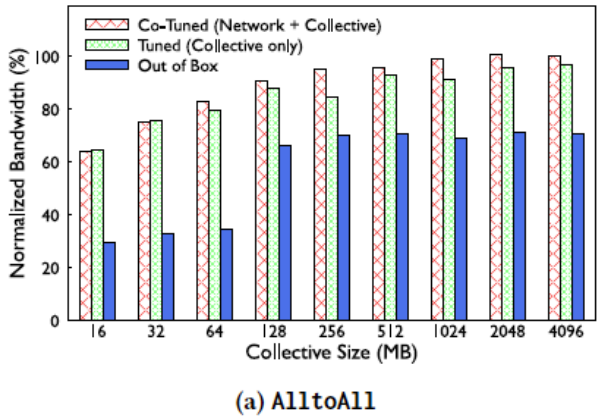


Figure 10: Collective benchmark: Normalized Performance Comparison between E-ECMP and TE (128 GPU)



A Gangidi et al, "Rdma over ethernet for distributed training at meta scale" Proceedings of the ACM SIGCOMM 2024 Conference, 2024

Conclusions

- Cloud providers are complimenting their (compute/storage/networking)-as-a-service with various types of AIaaS at large scale
- AI(GenAI)/Agentic implementations is an opportunity for Operators to monetize through AIaaS model over the national scale Telco Cloud infrastructure
- AI Service orchestration systems could be used to deploy the AI models at scale
- Telco AI-Cloud engineering for inferencing and training depends on use cases, type of models required for these services
- Several research opportunities in the context of Telco Cloud and AI - Agentic based Service Orchestration, Distributed Inferencing and Training services for ASPs.

References

[1] Zexu Li et al, "Evolving Towards Artificial-Intelligence-Driven Sixth-Generation Mobile Networks: An End-to-End Framework, Key Technologies, and Opportunities", Appl. Sci. 2025, 15, 2920

[2] GenAINet Emerging Technology Initiative, "Large Scale AI in Telecom, Charting the Roadmap for Innovation, Scalability, and Enhanced Digital Experiences", White paper, 2025

[3] ORAN/WG-2, "AI/ML workflow description and requirements", 2021

[4] Starlingx, https://www.starlingx.io/

[5] Kundu et al, "AI-RAN: Transforming RAN with AI-driven Computing Infrastructure", (Vision paper from Nvidia, submitted to IEEE for possible publication), 2024

[6] Kun Qian et al, "Alibaba HPN: A Datacenter Network for Large Language Model Training", Sigcomm'24

[7] Hoefler et al, "Datacenter Ethernet and RDMA: Issues at Hyperscale", IEEE Computer, June 2023

[8]A Gangidi et al, "Rdma over ethernet for distributed training at meta scale" Proceedings of the ACM SIGCOMM 2024 Conference, 2024

[9] Ravi Ravindran , "Enabling SMO over Nephio, a perspective", ORAN-SC Workshop, ONE Summit, 2024, https://wiki.o-ran-sc.org/display/EV/O-RAN-SC+Workshop+At+ONE+Summit+2024

[10] Shyam Parekh, Ravi Ravindran et al, "Software Defined Mobile Load Balancing in LTE and 5G Networks", WSN3, 2020

[11] Ravi Ravindran, Aytac Azgin, K.K.Ramakrishnan, "*Edge Transport (ETRA): Edge Transport Protocols for Next generation Mobile IoT Systems*", IEEE, Globecom Network 2030 Workshop, 2019

[12] Aytac Azgin, Ravi Ravindran, "*Network-assisted Consumer Mobility Support for Information Centric Networks*", IEEE, ICCCN, 2019

[13] Haiyang Qian, Fu Li, Ravishankar Ravindran, Deep Medhi, "Energy Aware Aggregation of Dynamic Temporal Workload in Data Centers", IEEE Transactions on Parallel and Distributed Systems, 2015

[14] Ravi Ravindran, Prakash Suthar et al, "Deploying ICN in 3GPP's 5G NextGen Core Architecture", IEEE, 5G World Forum, 2018.

[15] R. Ravindran et al., "5G-ICN: Delivering ICN Services over 5G Using Network Slicing," IEEE Communication., Mag., vol. 55, no. 5, May 2017, pp 101-107.

Backup