Aalto University
School of Science

Multi-continuum Computing: Service-based Applications and Systems

Hong-Linh Truong
linh.truong@aalto.fi

03.09.2025

Learning objectives

- Understand and apply computing continuum concepts
- Understand and apply service-based applications/systems
- Understand and apply composability for capabilities and components
- Understand and analyze multi-continuum computing characteristics, applications and systems

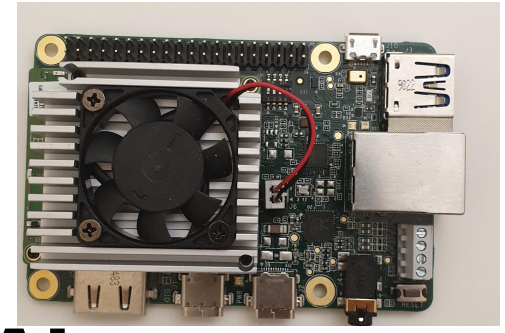
Content

- Computing continuum
- Service-based applications/systems
- Composability: capabilities and components
- Multi-continuum computing
  - dimensions and characteristics
  - quality of analytics and contracts

Computing Continuum

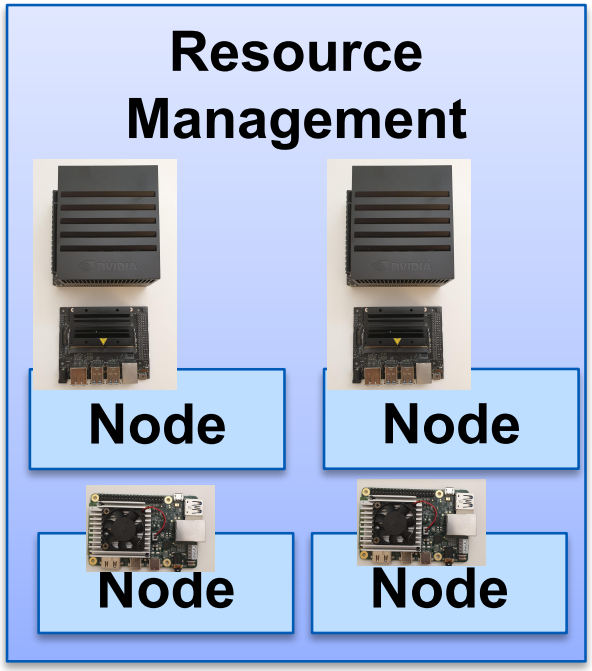
Edge computing: enable computing capabilities at the edge

Coral with Edge TPU System-on-Module, Google Edge TPU ML accelerator coprocessor



Jetson NVIDIA (GPU+CPU)

Edge systems



Our testbed for this course

Cloud computing/HPC: powerful and heterogeneous computing capabilities in (public and enterprise) data centers

- Clusters of VMs/containers
  - e.g., in Aalto we use Triton (https://scicomp.aalto.fi/triton/) and CSC (https://www.csc.fi/)
- High performance systems
  - Large-scale
  - Known accelerators
    - GPU and FPGA
  - High-bandwidth, low-latency interconnect fabrics for communication among CPU/GPU, memory and storage

Illustration of the HPE Cray EX cabinets. Copyright: Hewlett Packard Enterprise

Figure source:
https://www.lumi-supercomputer.eu/deep-dive-into-the-building-of-the-lumi-data-center/

Hardwares accelerating high performance computing, low latency communications for distributed workloads

- (New) AI Accelerators/Processing Units
  - TPU (Tensor Processing Unit)
  - Neutral Network Processor (NNP)
  - Vision Processor Unit (VPU)
  - IPU(Intelligent Processing Unit)
- Smart NICs
  - network cards with accelerator computing: offloading tasks like network functions, traffic inspection, and AI/ML data processing
- CXL (Compute Express Link)
  - high-speed, low-latency interconnect for CPU-device and CPU-memory

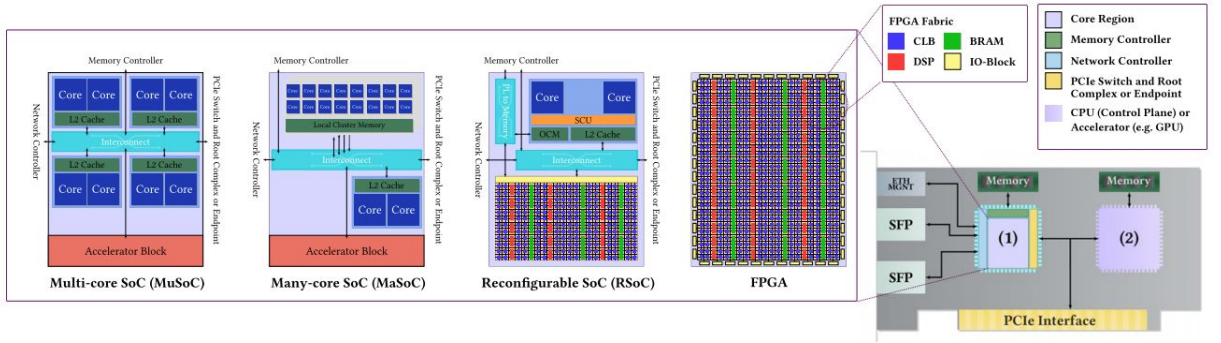
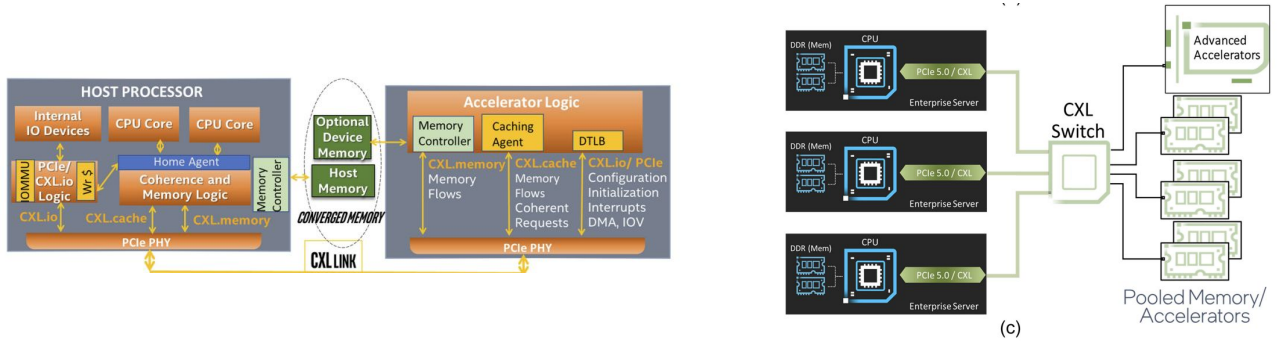
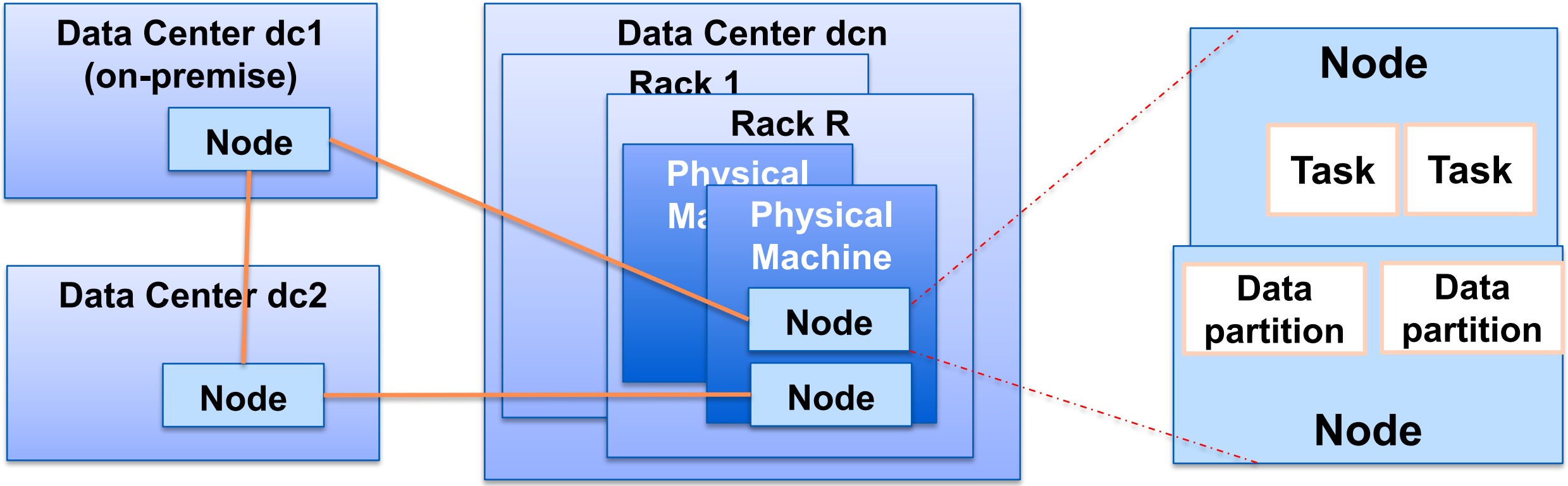


Figure source: Matthias Nickel and Diana Göhringer. 2024. A Survey on Architectures, Hardware Acceleration and Challenges for In-Network Computing. ACM Trans. Reconfigurable Technol. Syst. https://doi.org/10.1145/3699514



Figures source: D. Das Sharma. 2023. Compute Express Link: Enabling heterogeneous data-centric computing with heterogeneous memory hierarchy. *IEEE Micro*. (2023 Mar.–Apr).

Large-scale distributed infrastructures for multi-tenant, big data/AI applications

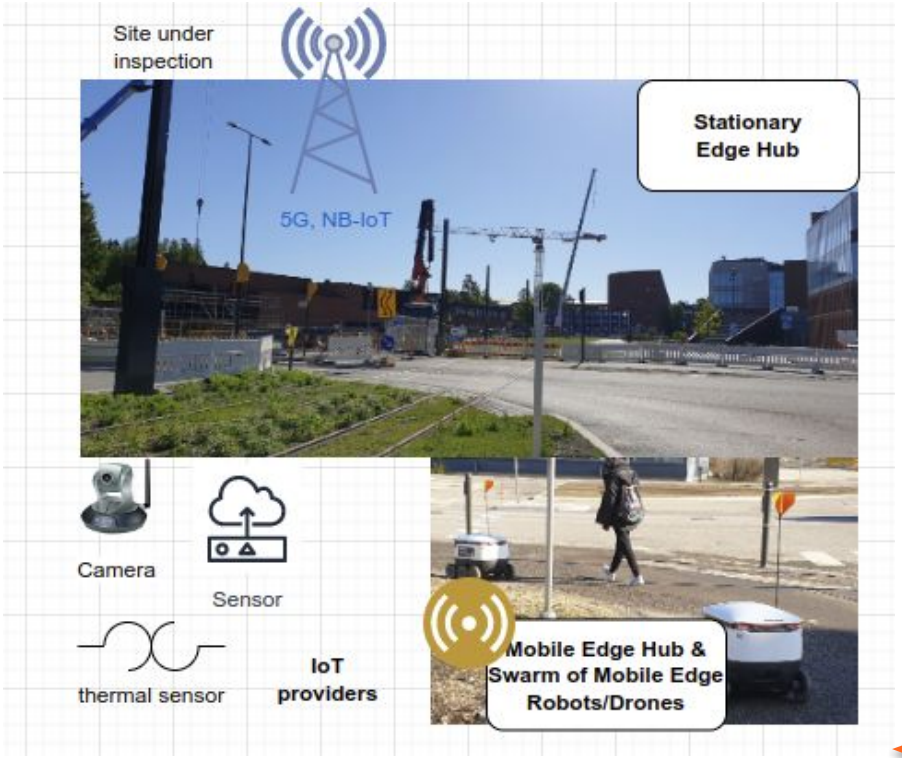


Data centers: cloud data centers, edge data centers, HPC centers

Edge-cloud continuum

- Enable widely decentralized complex workloads
  - AI/ML, data analytics, and real-time operations in distributed and heterogeneous environments
- Interconnecting edge and cloud resources and using them in similar manners
  - place and move tasks seamlessly between edges and clouds like in the same system
  - reduce issues and costs due to diverse development, deployment and operations
- From the technical viewpoint: for both edge and cloud
  - use similar enabling technologies, like containers and orchestrators
  - employ similar management techniques and methods (deployment, monitoring, policies, etc.)

Edge-cloud-HPC continuum



Edge

Cloud/HPC

Edge system (Edge Center)

Edge system (on-premise)

Cloud (big, centralized data centers)

HPC/Supercomputing (big, centralized data centers)

close (physically)

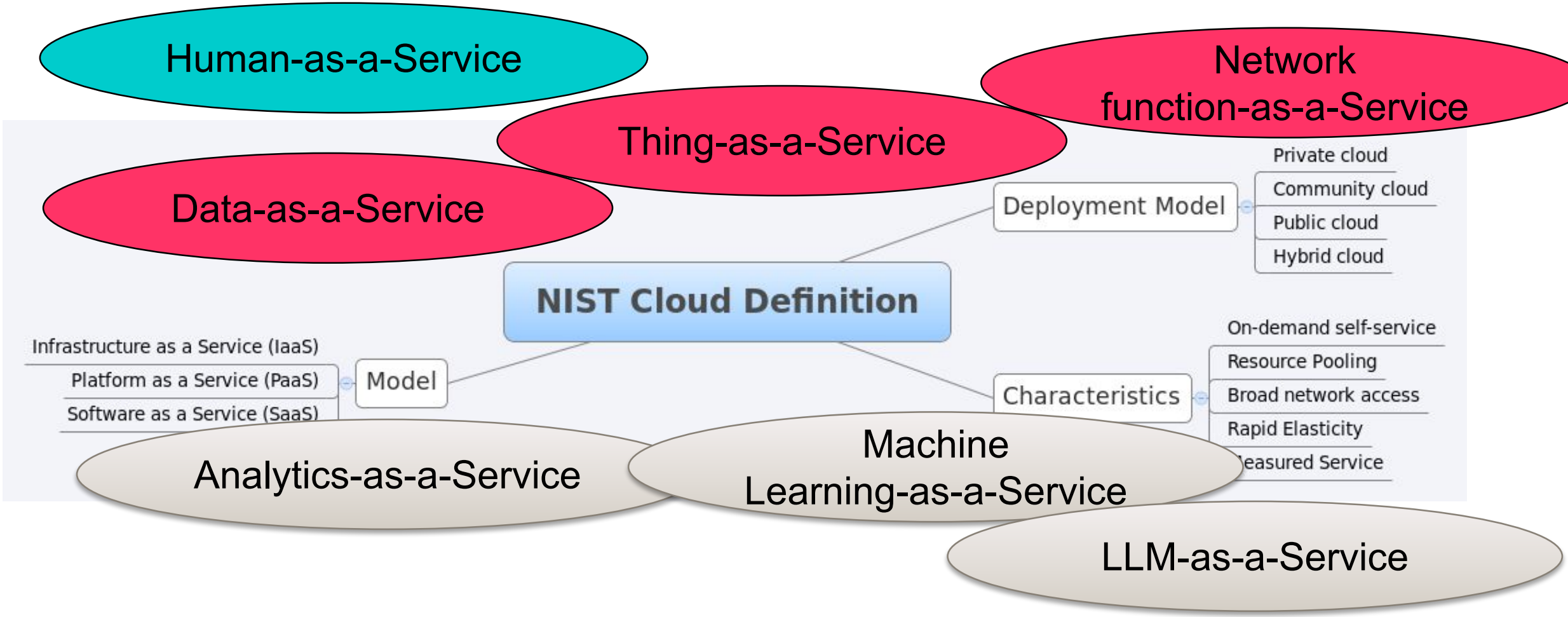
far (physically)

Service-based applications/systems

Services

- Service
  - reusalbe, independent software components, which can be accessed remotely with well-defined interfaces/APIs via standard protocols
    - e.g., REST, RPC, and AMQP
  - Microservice
- Service-based system
  - consists of multiple components; components are services
  - provisioned and deployed as services
- Separation from infrastructures/resource layers
  - resources can be provided by infrastructure services: computing, storage, memory, connectivities
    - e.g., a drone can be provided as a resource for computing/sensing

Types of Services



NIST definition: https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-145.pdf

Microservices

- Many components for data storage, data processing, ML inferences, controlling, etc.
  - act as basic elements of a complex system
- Acting as an intermediary
  - providing access to large (service) complex systems

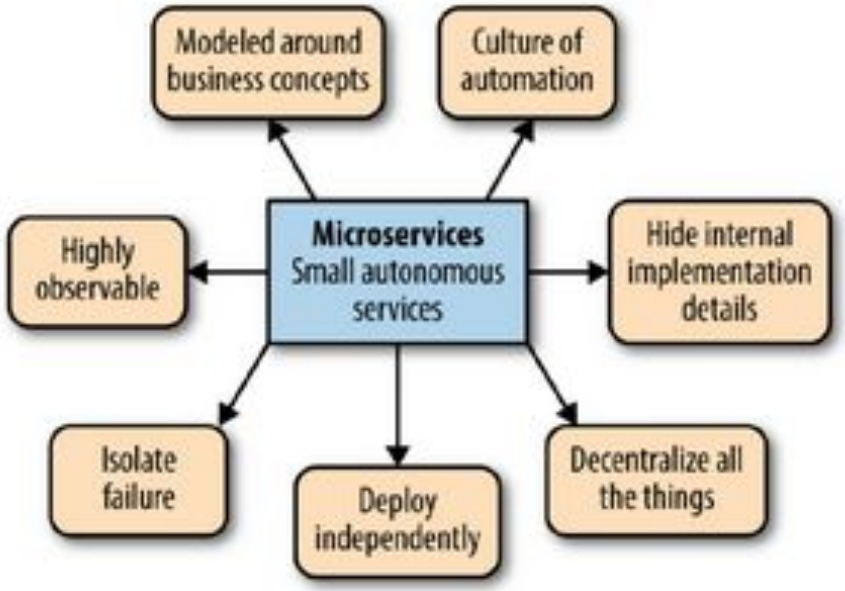
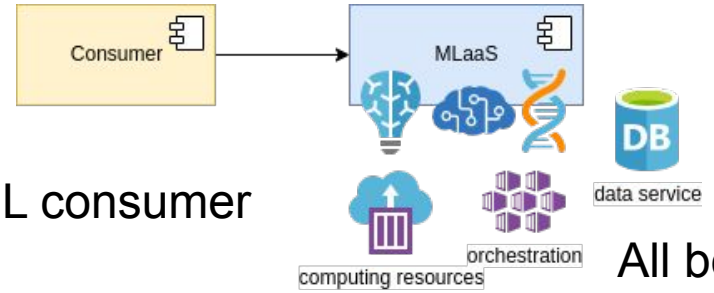


Figure source: Sam Newman, Building Microservices, 2015

Service consumer and provider models: Service coupling and deployment and service contracts

Two stakeholders engagement (Google, Microsoft, OpenAI, etc.)

Example in ML as a Service:



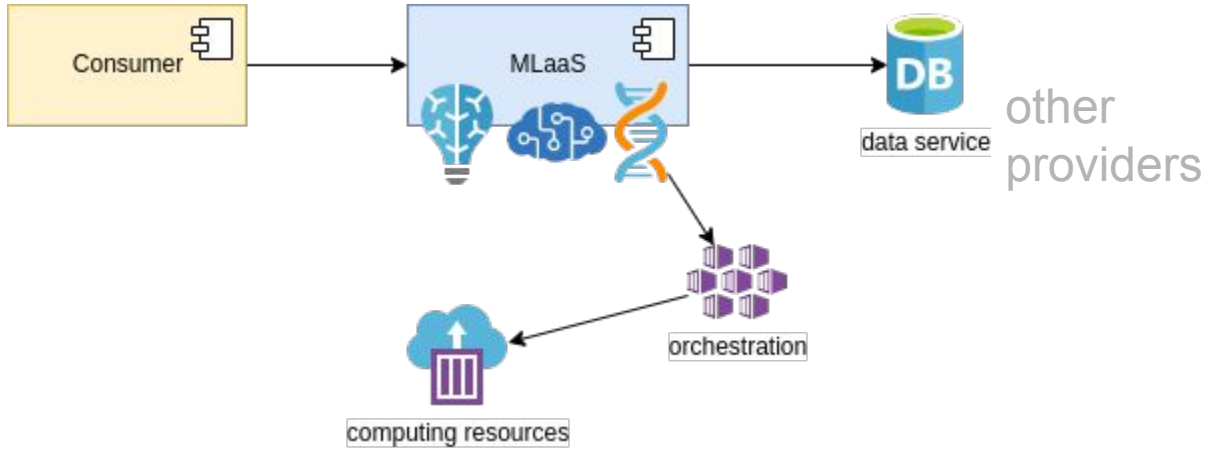
ML consumer

All belongs to the same provider

A special form: for internal uses in a large company

Three stakeholders engagement

Example in ML as a Service:



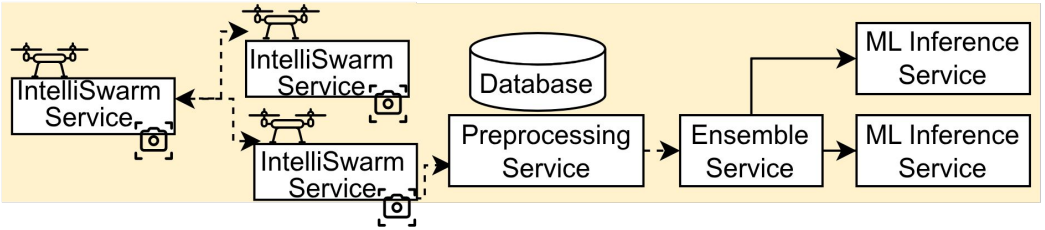
Relies on external platforms for orchestration, computing resources and other data (e.g., using Sedon, SagerML, etc.)

Service consumer and provider models: complex supply chains

Complex software supply chains and stakeholders in development and operations:

- data, AI/ML models, GenAI/LLM services and various services
- computing resources from swarm-edge-cloud continuum
- intelligence capabilities from AI/GenAI/ML and human capabilities (in-house vs external)

Example: for a swarm system



Service Provider:
drone services, computing continuum, AI/ML inference platforms

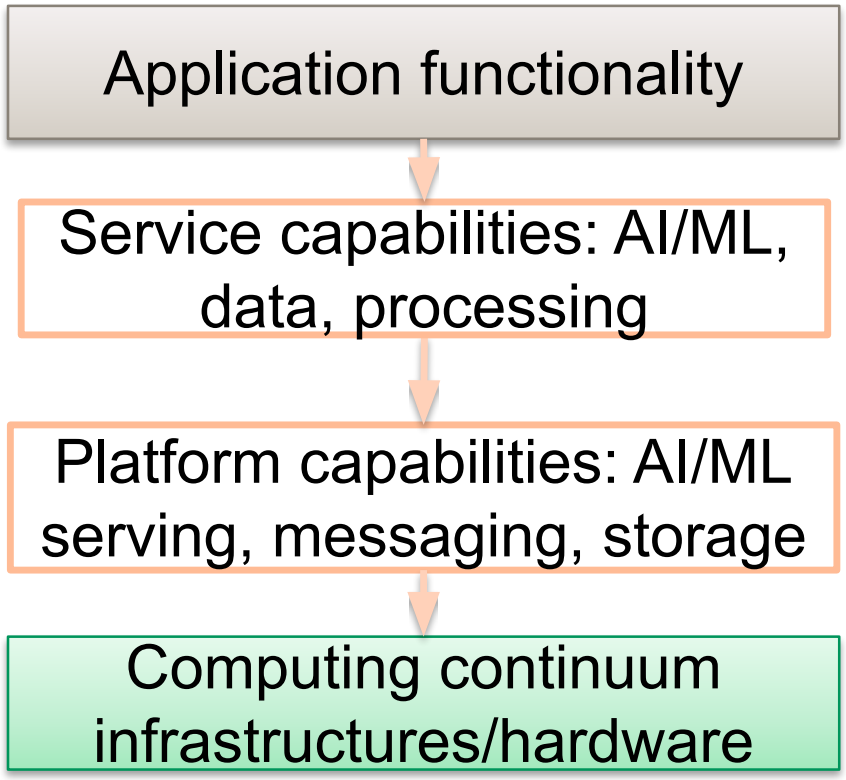
Application provider:
sensing services, AI/ML training and fine-tuning, AI/ML serving, pipelines

Third-party providers:
training data, libraries, ML/foundation models domain knowledge, programming frameworks, utilities,…

Composability: capabilities and components

Common layers

Which components do what, and where are they?



Layers

Software runtime and libraries: not a single software framework/stack

- Many **platform services** according to workloads and architectures, e.g., *messaging systems* with MQTT, NATS, CoAP, AMQP, Kafka, or Pulsar
- Multiple powerful **AI/ML programming frameworks** (PyTorch, TensorFlow, scikit-learn, Keras, Ray, etc.)
- Diverse **AI/ML serving platforms** (Seldon, BentoML, Triton, Ray, ZenML, Lightning LitServe, Ollama, vLLM, etc.)
- Multiple programming models and paradigms: Microservices, serverless, batch workflows, stream processing & reactive system

AI/ML capabilities combined with other capabilities

- Systems/Application compositions/workflows include many different services
  - AI/ML inferences are crucial but *just one* type of components
- Deployed in swarm-edge-cloud computing resources, usually as microservices
  - With other third-party platform-as-a-service
- Heterogeneous infrastructures, compute resources and connectivities

End-to-end system, with AI/ML inferences from traditional ML and edge LLMs

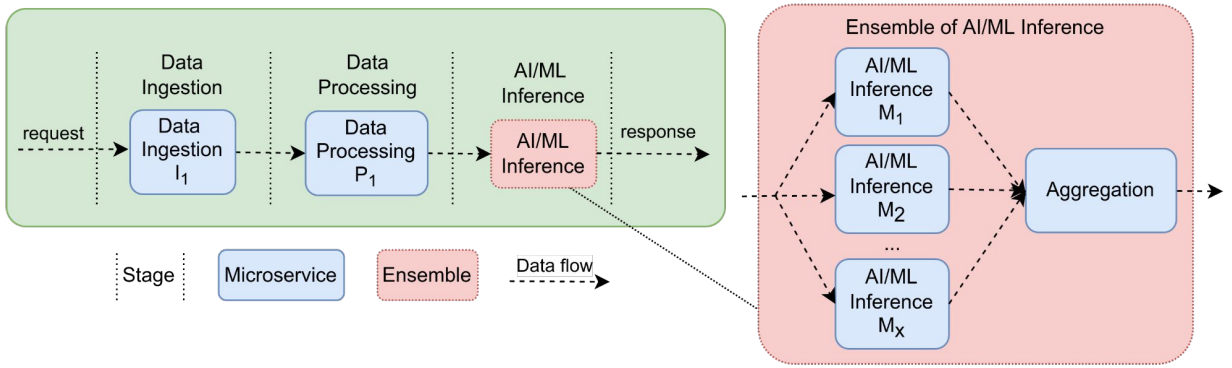
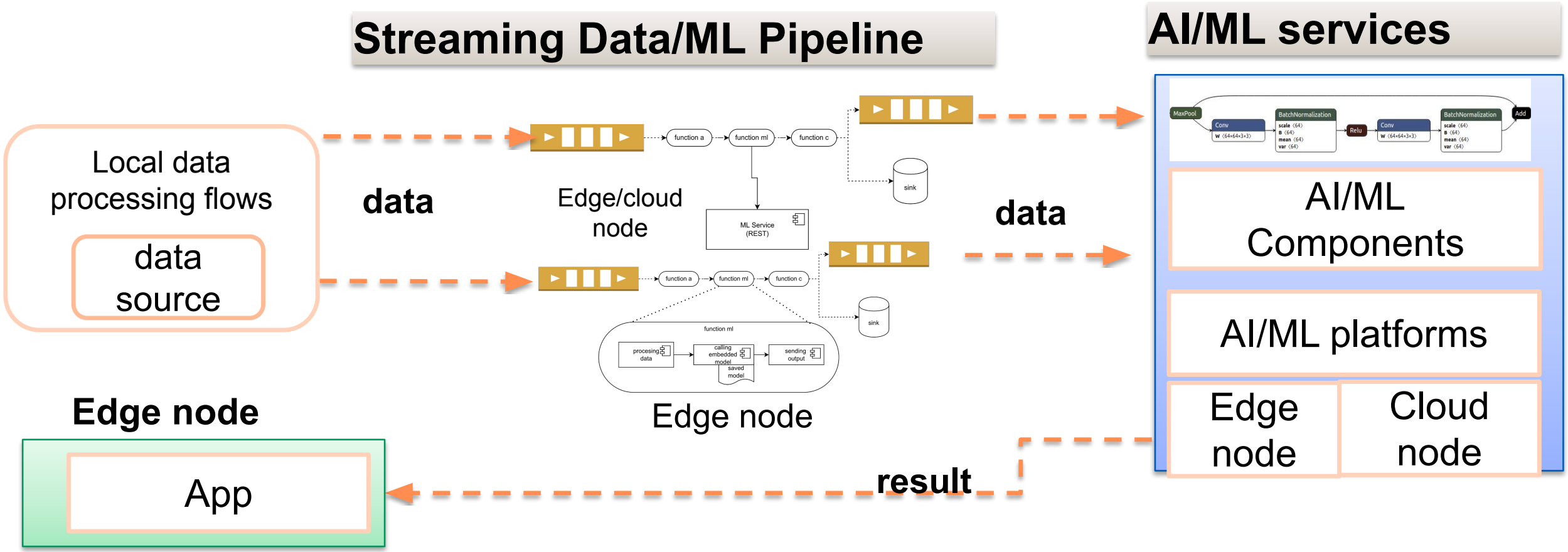


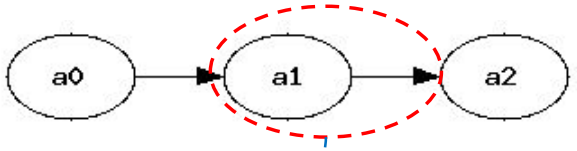
Figure source: *"Novel Contract-based Runtime Explainability Framework for End-to-End Ensemble Machine Learning Serving"*. https://doi.org/10.1145/3644815.3644964

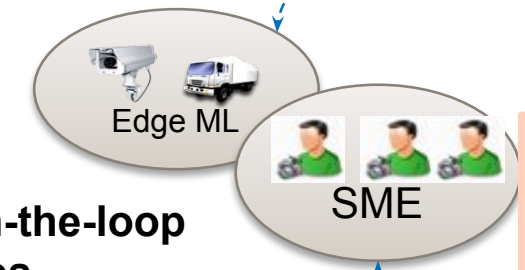
Composition and interactions in edge-cloud applications/systems



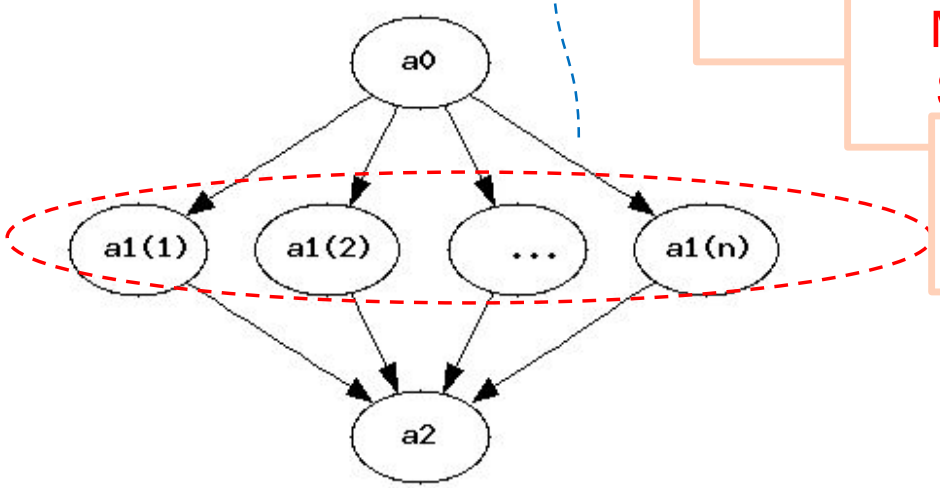
Workflows: composition of analytics and intelligence tasks

SME: Subject-Matter Expert





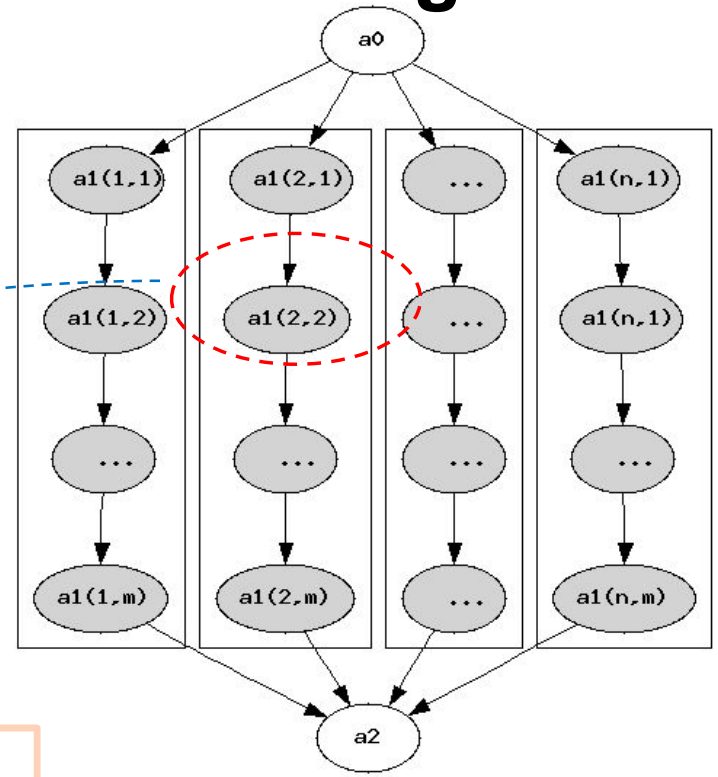
Human-in-the-loop Collectives

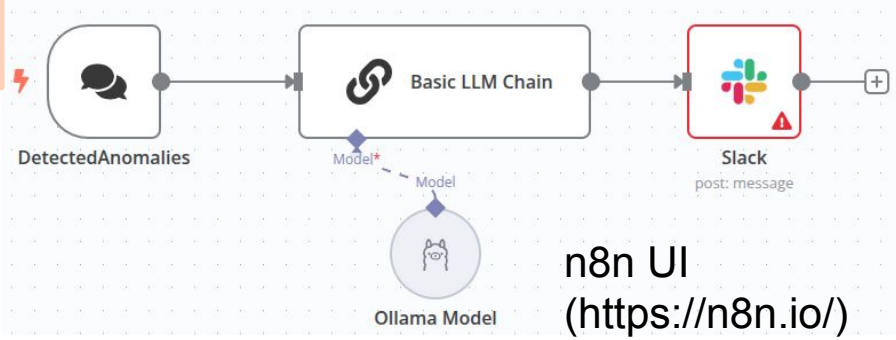


ML Training

ML/LLM Serving

AI/GenAI services





n8n UI (https://n8n.io/)

Chains of services

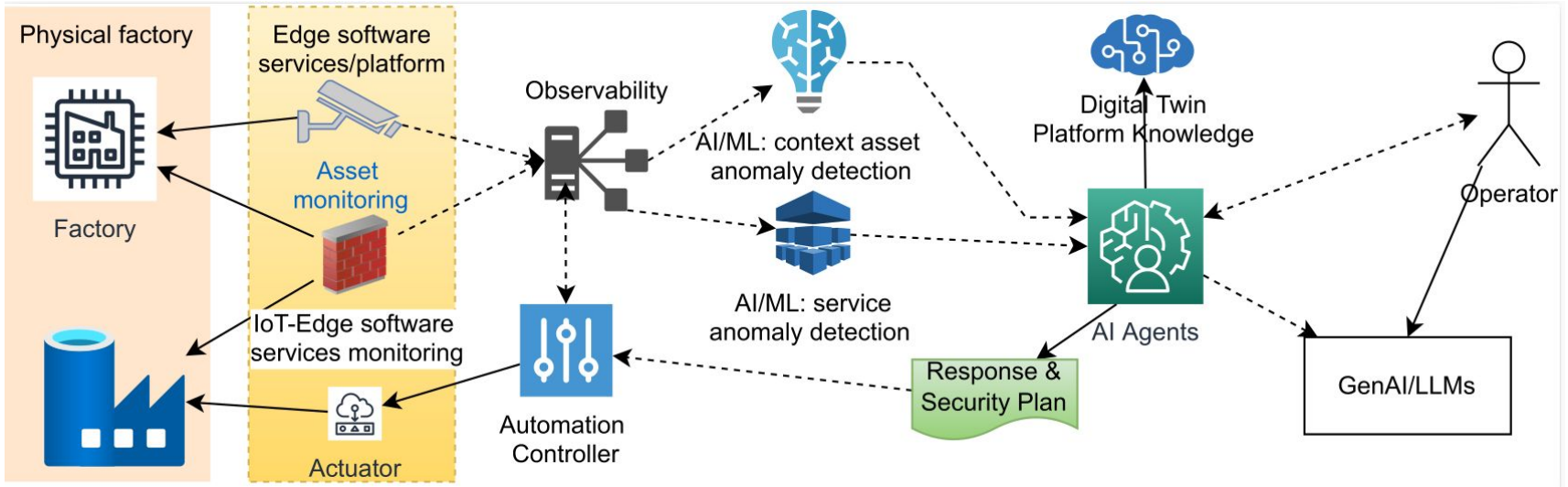


Figure source: Hong-Linh Truong, ***"New Frontiers in Service Engineering Analytics for Multi-Continuum Systems (FrontSEA)"***, working paper, 2025

- **Different services interacting via composable logics**
  - IoT sensing/monitoring, data analytics, ML inferences, etc.
  - for analysis, optimization and control functionalities
- **Compositions/ensembles of AI-human**
  - different design patterns/combinations
  - Gen AI/LLMs, AI Agents, and human
    - replacement and combination techniques enable intelligence continuum

AI

Agentic Systems

- Has long history: *"an agent is a computer system situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives.*"

  (From M. Wooldridge. "An Introduction to Multi-Agent Systems". Wiley. 2001)

- An AI agent includes many features from monitoring to reasoning to control

- Current trend
  - AI Agents are backed with GenAI/LLMs capabilities

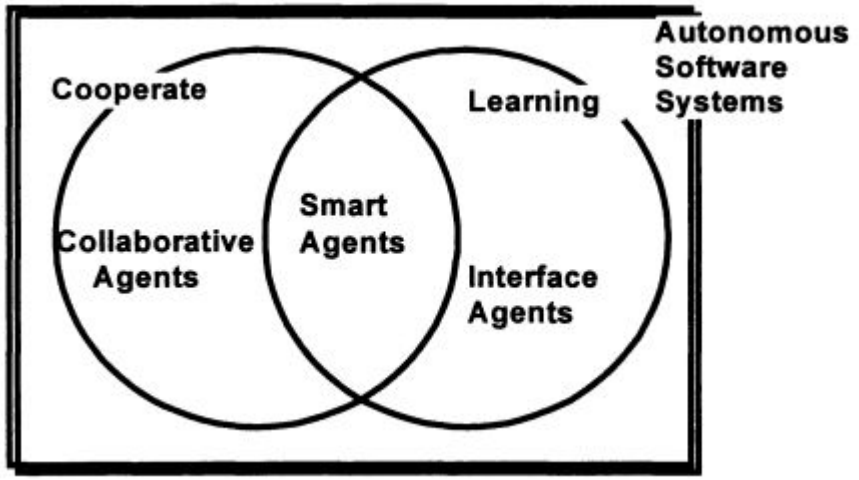


Figure source: Nwana, H.S., Ndumu, D.T. (1998). A Brief Introduction to Software Agent Technology. In: Jennings, N.R., Wooldridge, M.J. (eds) Agent Technology. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-03678-5_2

Multi-continuum computing

Multi-continuum view

Multiple combinations of computing continuum

Swarm	Edge	Cloud
Edge	Cloud	HPC
Cloud	HPC	Quantum

Application requirements for multi-continuum

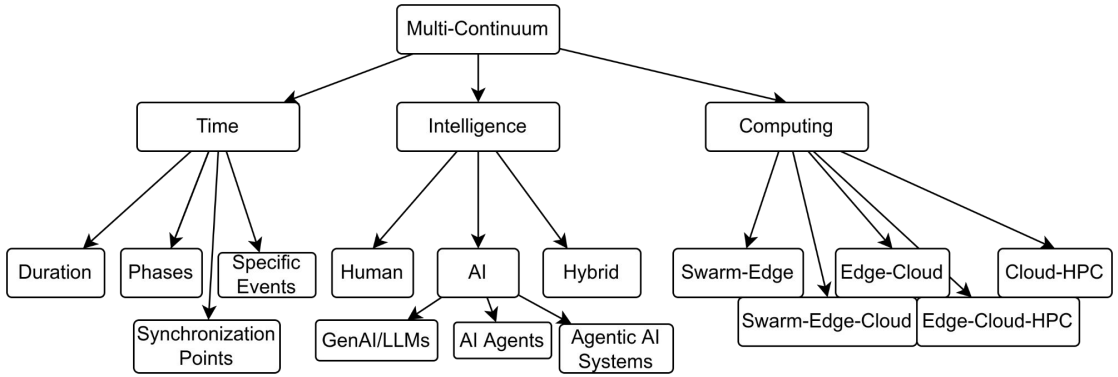


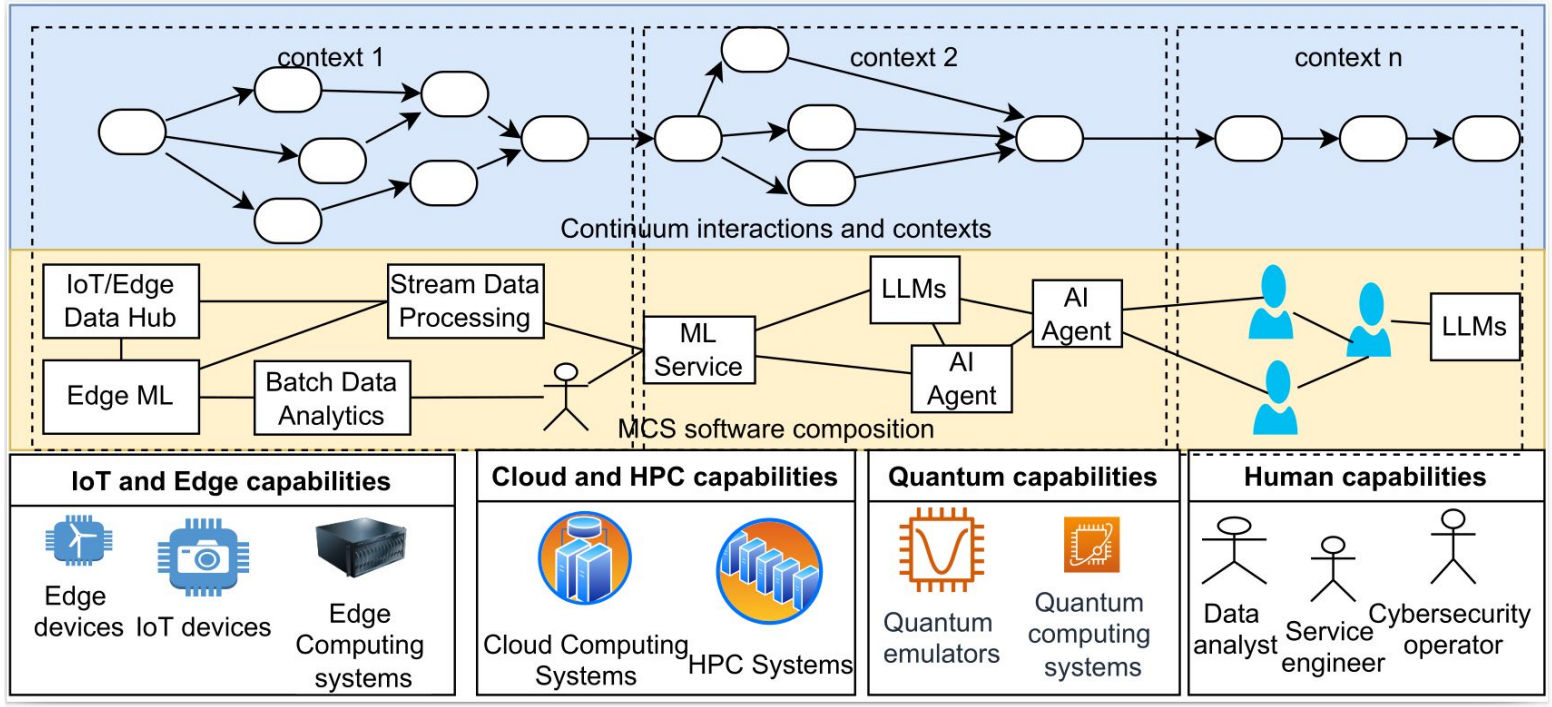
Figure source: Hong-Linh Truong, Kostas Magoutis, *"Multi-continuum view for Swarm-Edge-Cloud Service-based Applications"*, **working paper**, 2025

Why is it important?

- New computing models for emerging requirements
  - performance, data regulation
  - application needs
    - leverage best capabilities for suitable tasks
    - long-running solutions for complex problems
- New computing models leveraging advanced technologies
  - connectivities
  - computing: processing hardware and memory and advanced interconnect fabrics
  - AI

Capabilities, composition and context in multi-continuum computing

Capabilities enable complex compositions to solve problems



Context drives the way to compose capabilities to solve problems

Figure source: Hong-Linh Truong, ***"New Frontiers in Service Engineering Analytics for Multi-Continuum Systems (FrontSEA)"***, working paper, 2025

Capabilities for solving complex problems: computing, time, and intelligence

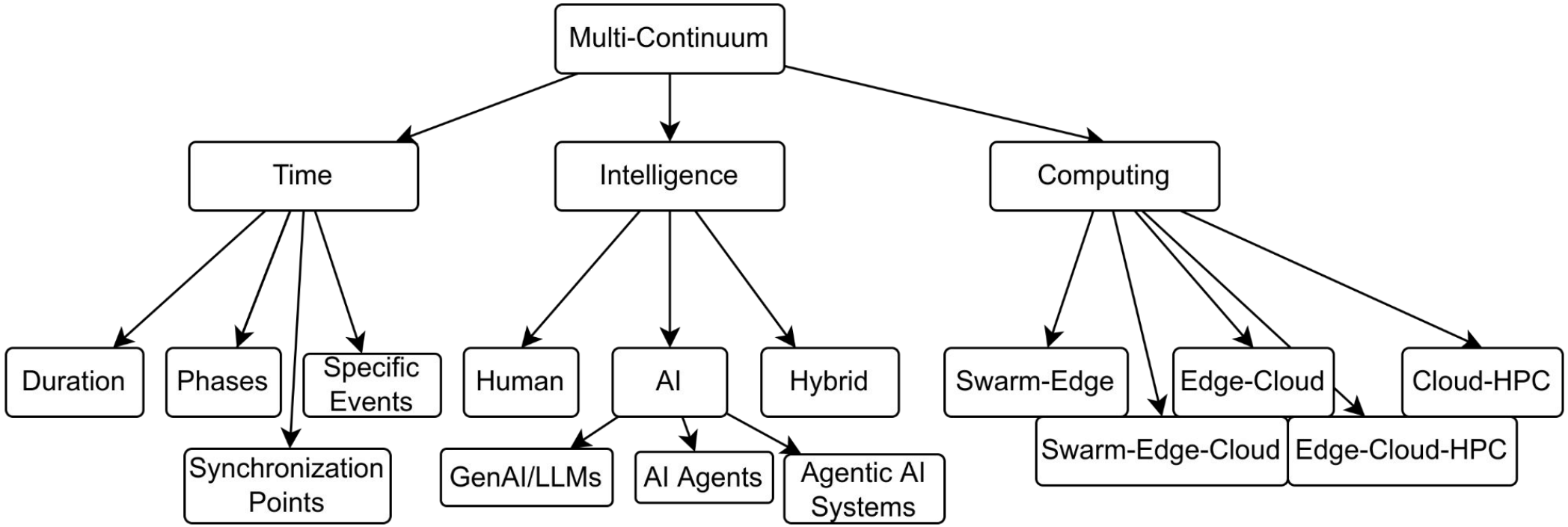


Figure source: Hong-Linh Truong, Kostas Magoutis, *"Multi-continuum view for Swarm-Edge-Cloud Service-based Applications"*, **working paper**, 2025

Computing Continuum

- Using a combination of edge, cloud, and HPC resources
- Resources can be changed
  - computing resources
  - connectivities
  - storage
  - etc.
- Provisioning resources for a long run
  - seamless resource elasticity
  - job management, placement and scheduling

Time Continuum

- Key entities in a mission from a time viewpoint
  - tasks, including input and output
  - resources performing tasks
- Continuum requirements for
  - specified durations
  - phases
  - specific events
  - synchronization points
- Continuum provisioning and management for long missions
  - no interruption between phases, specific events and synchronization points

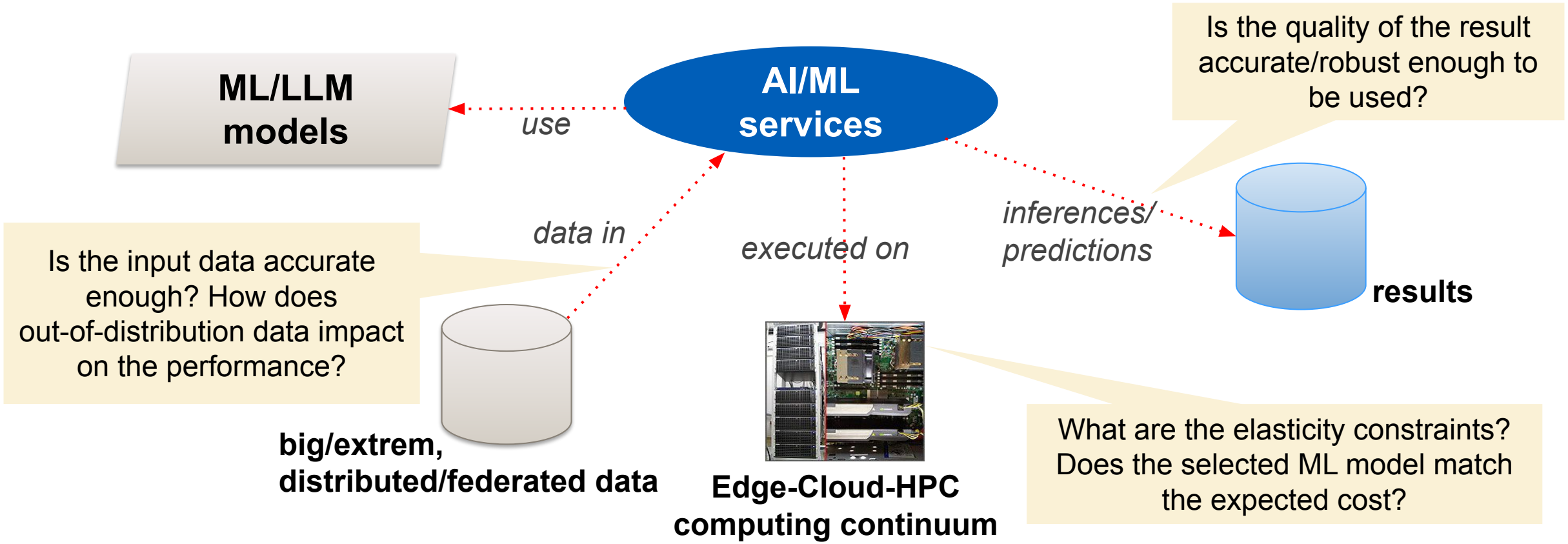
Intelligence Continuum

- Intelligence from software and humans
  - software: GenAI/LLMs, AI Agents, …
- Hybrid intelligence capabilities
  - human intelligence can take the form of an individual or a crowd
  - intelligence from software can be built from ensembles/collectives of GenAI/LLMs/ML services
- Hybrid intelligence can be provided as
  - sequence/structured pattern
  - a collective one of human and AI capabilities
- Basic operations for capability provisioning
  - add, remove, change, replacement

Intelligence continuum in the age of GenAI/LLM services

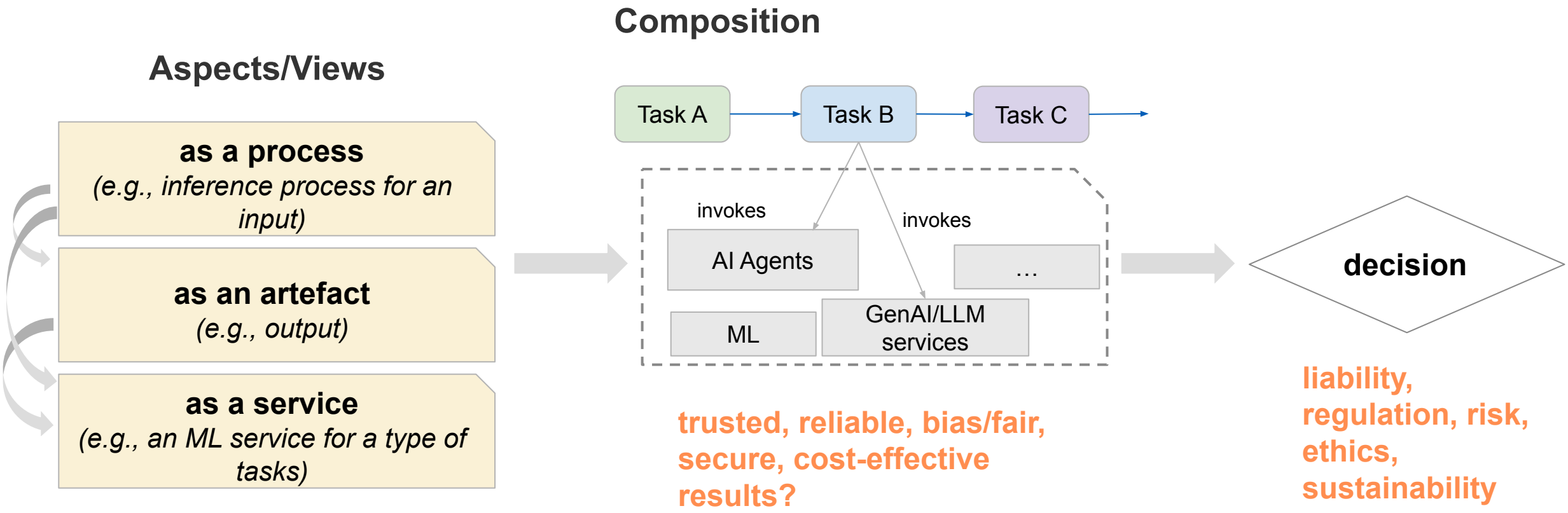
- **Which and how AI/GenAI services are used in ensembles?**
  - non-functional properties for GenAI/LLMs are determined and managed based on specific task aspects/views
  - performance, cost, reliability, risk, relevance degrees, etc. based on the type of tasks and contexts
- **Which and how AI+human intelligence capabilities are coordinated for consumer contracts and continuum requirements?**
  - expected quality attributes: not just time and cost but also, e.g., reliability, risks and sustainability (task and context specific)
- **Which and how quality attributes are linked to risks, the use of guards of AI risks?**
  - complex AI/GenAI risks ⇒ trustworthiness
  - many quality attributes cannot be automatically measured
    - combine measurement, feedback and estimation techniques

Dependencies across data, AI/ML models, computing resources and quality aspects

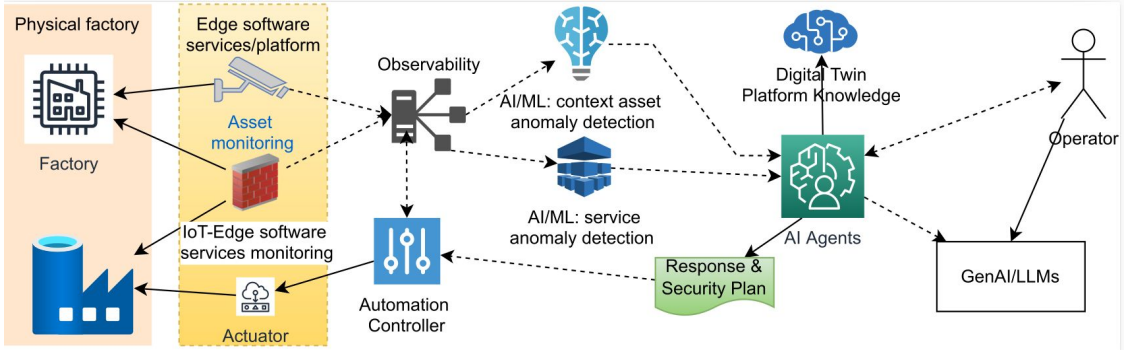


Source: **"Coordination-aware assurance for end-to-end machine learning systems: the R3E approach"**, AI Assurance, https://doi.org/10.1016/B978-0-32-391919-7.00024-X

Quality of analytics (QoA) views



Emerging runtime observability for GenAI services and AI Agents

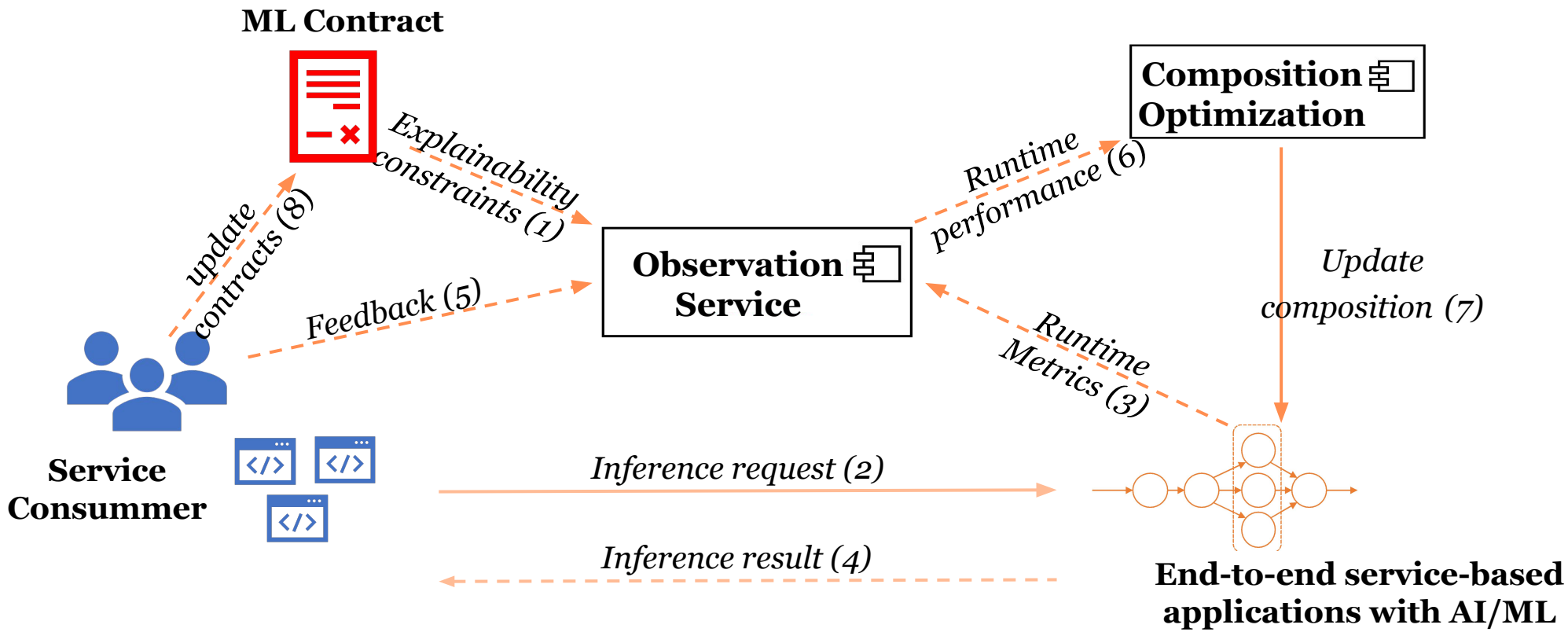


Runtime observability/explainability for AI Agents/LLMs workflows

- Observability of AI tasks
- Explainability w.r.t. AI trustworthiness and risks
- AI Agents interactions
- LLM Ensembles

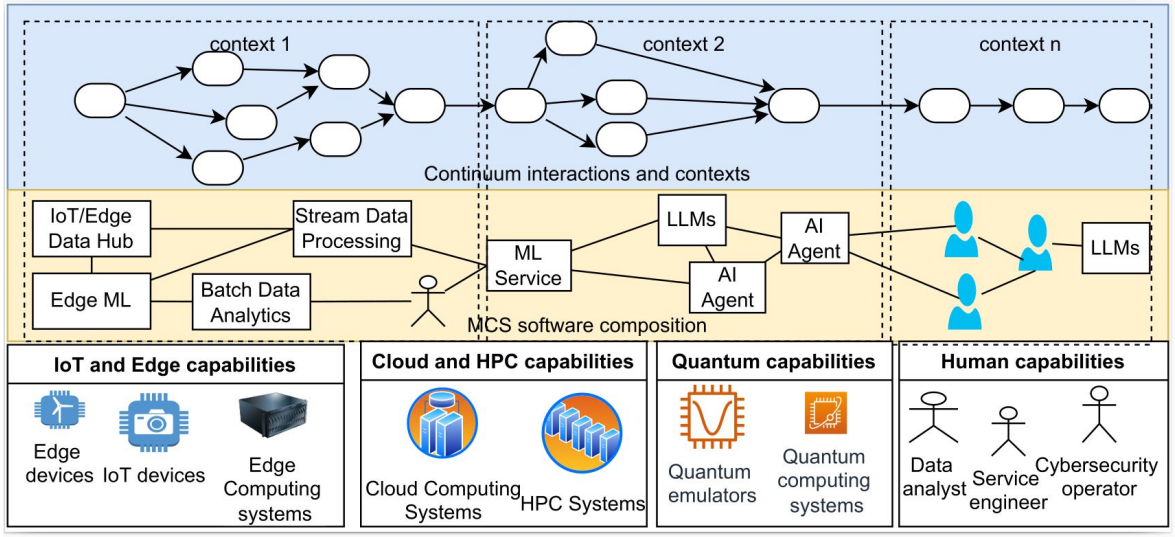
Figure source: Hong-Linh Truong, ***"New Frontiers in Service Engineering Analytics for Multi-Continuum Systems (FrontSEA)"***, working paper, 2025

Consumer-defined contracts



Based on "Novel Contract-based Runtime Explainability Framework for End-to-End Ensemble Machine Learning Serving", CAIN 2024, doi: 10.1145/3644815.3644964

Multi-continuum systems



Multi-continuum Service-based Applications/Systems

Complex characteristics! so how to build them?

- *Robustness, Reliability, Resilience, & Elasticity (R3E)*
- *Monitoring, Obervability, Vulnerability and Explainability*
- *Trustworthiness*

Study log for this week

Think about

- A scenario of service-based applications/systems in multi-continuum computing
- Read one of the papers in the reading list for today lecture

Then

- which kind of continuum characteristics, potentials or challenges do you see?
- ~1-2 page – submit it to the MyCourses for comments/feedback (keep it in your git)

Kiitos
aalto.fi