Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 December; 6 10s(2): 2061-2082

Adaptive AI Workflows for Edge-to-Cloud Processing in Decentralized Mobile Infrastructure

Goutham Kumar Sheelam*

*IT Data Engineer, Sr. Staff, Email: gouthamkumarsheelam@gmail.com, ORCID ID: 0009-0004-1031-3710

Abstract

Deterioration in wireline/mobile communications infrastructure reliability and subsequent service outages with a snowballing trend amount to significant social costs that are likely to spiral. Service providers are reducing the attack surface by decoupling control and service functions, substituting open-source solutions for traditional proprietary service instances. Self-governing agent devices and nodes are intended to be augmented with enhanced sensing, control, and storage capabilities, notably cognitive capabilities. These architecture trends snowball a move to decentralized multicentric cloud infrastructure envisioned by new paradigms. There is a need for resilient decentralized mobile infrastructure based on Avatars that is responsive to large-scale topology dynamics, heterogeneous traffic demands, and cognitive agents with cross-layer perceptions and behaviors. They argue that adaptive AI workflows underpinned by dynamic Spatio-temporal agent/offer modeling augmented with situational spatio-temporal modeling, multi-agent MABs, and RL methods can enable edge-to-cloud processing of nodes and traffic agents for decentralized mobile infrastructure. [1] The proposed notions of situational spatio-temporal agent/offer modeling, multi-agent MAB, RL-based distributed behavioral learning, and cloud-edge creativity enable AI workflows for edge-to-cloud processing in trajectory-aware network infrastructure environments. The enhanced awareness of agents and offers with their situational relevance in dynamic interactions is enhanced, enabling smart cognitive state and intent discovery, unprecedented online autonomous adaptive processing and decision-making, and unexplored creativity at the edge and in the cloud. Low-complexity multi-agent MAB can tune the related prices and processing power in degree-matched multi-hop network architecture for edge-to-cloud decentralized power-rate allocation with learning guarantees. Knowledge transferred by simple tabular RL is discoverable and composable, realizing online autonomous decentralized emergent control even with dynamic topology and changing objective environments.

Keywords: Big Data, Generative AI, Cloud Connectors, Education IT Solutions, Sustainable Energy Technologies, Data Analytics, AI in Education, Smart Energy Systems, Cloud Computing, Digital Transformation, Intelligent Infrastructure, Predictive Analytics, Machine Learning in Education, Energy Optimization, Real-time Data Processing, Renewable Energy Technologies, IT Innovation, Adaptive Learning Systems, smart Grid Technology, Cloud Integration, Energy Efficiency, Personalized Education, Green Technology, Educational Technology (EdTech), Sustainable Innovation

1. Introduction

Emerging technologies supporting artificial intelligence (AI) are being adopted at an increasing pace, affecting a wide range of industries, including consumer electronics, automotive, healthcare, finance, education, manufacturing, and entertainment. Meanwhile, edge computing, where computation is performed at the periphery, closer to the data sources, is rising as a highly promising technology. It allows network operations to be more decentralized and efficient, providing flexibility, scalability, robustness, and a high quality of service. With AI based on deep learning algorithms and edge computing combined, a concept of edge intelligence is introduced by performing AI tasks at the edge instead of in the cloud. By enabling distributed and decentralized orchestration of smart data and computations, it can provide ultra-high-performance, massively scalable, and privacy-preserving implementations of AI tasks. Emerging AI applications such as autonomous driving and virtual and augmented reality, which require massive amounts of data processing, have boosted the research of edge intelligence, resulting in new methods and paradigms. Many AI methods and paradigms were employed at the edge, demonstrating the validity of AI-enabled edge computing. For example, to mitigate the costly data transfer and processing on the cloud, AI-driven edge caching is introduced in which data popularity prediction can be leveraged to cache the data at the edge nodes. The network communication load is reduced, while the average data retrieval latency can still be effectively minimized as it relies on multiple cooperative edge nodes to retrieve the data. For minimizing the delay of video analytics, deep reinforcement learning (DRL) based edge video analytics framework is developed following the three-layer edge architecture to distribute parts of the video analytic tasks at the edge nodes. By deciding the execution strategies, task allocation, and resource allocation of edge computing tasks, the average processing time of the tasks and the bandwidth used for data transmission can be effectively reduced. However, with evolving environments, a complicated balance is struck by each method between avoidable performance deterioration due to freed inertia and maintainable robustness against unnecessary signal noise. Synchronous self-learning and self-adaptive micromodels are also investigated.

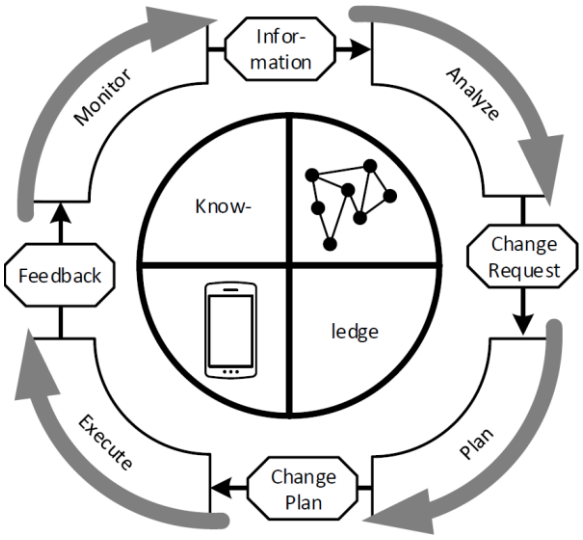

Fig 1: Self-Adaptive Distributed Systems

2. Background and Motivation

Emergence of Decentralized Mobile Infrastructure Edge-computing-based mobile infrastructure with distributed strengths is emerging. The software-driven decentralized Mobile Infrastructure is scalable and operationally elastic. However, strict Quality of Service limits and rapid evolving business scenarios are putting it at unusual challenges. Analog to Software-Defined, Infrastructure as a Service now includes device resources. There is a desire of deploying a learning functionality to dMIs which can handle workflows switching through dMIs as well as edge-cloud pathing optimization. Such approaches need to mitigate, or better utilize, the exponential trade off between accuracy and computation complexity. Inspired by cognitive networks, one consideration is developing machine learning frameworks to harness the large amount of historical data trajectory. The operand execution off-load redirections could be governed by controllable intelligence and a few learning capabilities.

Adaptive Edge-Inbox Workflow and Learning Paths A proper taxonomy of audio-visual AI computing and learning platforms is proposed, surfacing an adaptive edge-inbox workflow. AI processing should be time-series step-wise on temporal images, audio clips and their embeddings. Such analysis units cannot be batched and need to adhere frames/intervals by complexity and event time. A workflow example is presented, featuring modeling-based AI computations and a QoS-Cobb quotient methodology. Such a cognition-driven computing model is primitive background on how to compose an adaptive edge-inbox AI workflow for this pursuit. Performance-redirection maps of workflows and computing paths are also discussed.

Applications and potential cases are first introduced and followed by AI workflow architecture building blocks, with systematic proceeding starting from edge gear specifications, AI computing task portfolio, to their performance estimates. Edge device and algorithm portfolios are finally evaluated with regard to edge-inbox QoS. Network protocol layering and resource allocation are module steps, which leverage a heuristic joint optimization of revenue and best effort Service Level Agreement provisioning.

3. Decentralized Mobile Infrastructure

Network (such as in rural areas) through the use of unlicensed bands such as WLAN and LiFi. To share data with the Cloud for global perspective, neighbouring observations may cluster neighbouring drone observations into a virtual Network. Each mini-Edge Drones can be volunteerized and multi-hosted by forming Edge Cooperatives. With a cooperation-aware Ultrawideband, Wi-Fi and 5G enhanced prediction based searching can be applied to send ray-points for data cleanup and contextual clustering. The mini-Edge at a host may reduce in resources so to check for non-required address and send a data summary instead for a visualization at it, while keeping Edge collaborative tasks ongoing. Not necessitating guaranteed service delivery may also enable cooperative data collection for a rainfall/earthquakes crowd-sensing network.

Equation 1: Network Coverage and Signal Propagation

$$PL(d) = 20\log_{10}(d) + 20\log_{10}(f) + 32.44$$

Where:

- $PL(d)$: Path loss in dB
- d: Distance in kilometers
- f: Frequency in MHz

Data federation is an initial step for a collaborative Edge Infrastructure. Federated Learning would enable more contextual personalized models and support studying issues such as socio-technical deployment. It however imposed a non-overcoming in each local training at near-zero days. Addressed by introducing adaptive Centralized Learning and completion handling within a common intervention, it is yet to investigate the proper design and target deployment. Cloud as macro-Edge may facilitate decreasing Carbon Footprint via Climate-Sensitive Macrosystems gaining localized prediction for decision making. Options labeled 1-4, and colocation option with redundancy based on First-In-First-Out, may be enabled and studied for Cloud implementation preparation methods.

The marches between federated data/monitoring are a challenge not yet well addressed. Addressing it and timing smoothing mechanism for Edge modeling and Cloud planning integration is worth noting and improving service delivery and discipline behaviors. The gradual delivery of Cloud situational data via collaborative Ground-Edge-Cooperative systems by proposing and learning a cumulative Truth ranking for score aggregation and opportunity aware lighting data offloading algorithms, which still has plenty of improvement issues.

3.1. Definition and Characteristics

Edge-to-cloud processing enables massive data transfer and distributed learning by transmitting model parameters such as weight vector characterization network topology structure or other numerical data. With the assistance of AI in edge-grained features of heterogeneous smart IoT, considerable complexities arise in semantic structures of AI workflows among chain relationships in IoT and cloud destinations. Combinations with AI inference models summarized from deep learning-based data-driven sensing intelligence and performance metrics indication designing describe AI workflows of intricate processing. AI workflow enclosures for edge-to-cloud processing are discussed aiming at formulating, executing, and adapting structured AI workflows among processing edges and cloud destinations.

The decentralization of the mobile network architecture and the application of containerized solutions lead to the emergence of the multi-access edge computing environment and the evolution of new generations of radio access networks. Such emerging network architecture builds a brand new mobile infrastructure paradigm with an end-to-end provisioning of smart IoT applications. Edge-to-cloud processing of smart services aggregates streamed multimedia contents from heterogeneous IoT centers to elaborate AI-driven data-shaping dependencies with edge locality. Such grand vertex AI cognitive domains incur escalating complexities with entity diversities in heterogeneous AI processing models. Global meta-AI computation leveraging distributed learning in cloud facilities transmits the gradients among decentralized clusters of IoT and cloud splits.

Smart contracts codifying application logic enforce deterministic rules of transactional events making state transitions in a transparent way by leveraging an open and permissionless environment. Such a transparent but evolving network could introduce decision-making and trustability issues in computational AI model sharing. Cooperative AI is envisioned to address such issues by establishing trusted environments for the multi-agent system even with discretionary behaviors from participants being concerned. Agent-centric cooperation enforced by trust networks is modeled by probabilistically potential games which converge to mutual structures of action sharing and information exchange as socially optimal equilibria.

3.2. Challenges and Opportunities

The advent of decentralized mobile infrastructure provides vast opportunities and inspires technical exploration. The explosive growth in demand for mobile services and rapid increase in production of mobile data have been driving the deployment of resource-rich edge devices at the network edge. In comparison to the limited computing and storage resources of mobile devices at the edge of the network, the network edge has much richer resources, giving rise to an opportunity for a more collaborative and decentralized mobile infrastructure. Such wide deployment and rich resources

of edge devices provide unprecedented opportunities to enable large-scale edge intelligence. Edge resources can perform mobile deep learning tasks for a large number of smart mobile devices.
However, the heterogeneous network architectures, location-changing node roles, resource-dynamism, privacy-preserving demanding, and computation-intensive workloads in the decentralized edge device networks pose grand challenges for the intelligent collaboration of edge devices. Consequently, collaboratively training a mobile deep learning model across many edge devices becomes a complex task. Moreover, the diverse resource availability over edge devices demands adaptable mobile deep learning pipelines, pushing intelligence distribution and leadership beyond the cloud edge continuum. Both device resource-constrained and data-sensitive mobile use-cases emerge, demanding adaptive workflows to process data spills across the heterogeneous edge and cloud infrastructures. The vertical categorization, communication congestion, latency constraining, and data security further complicate the agility.
Leaf and Cloud distribute massive data but imbalanced loads, leading to the makespan uncertainty for newly aroused tasks on the edge processing units. The challenged scheduling scheme cannot adapt workloads from Leaf to Cloud. Node join/leave and data moving over the network edge complicate the adaptability of communication pipelines across edge devices. Hidden denoised knowledge on matured models is not fully exploited, leading to the convergence inactivity on fresh imitation models. With the true propagation frequency changed, flexibility in sparsity of model communication is maximized at each round. The dynamically growing/global network edge further complicates workflows on discovery, classification, and adaptation. Existing edge-to-cloud/remote workflow frameworks cannot instantiate policies that describe how and where to migrate edge models continuously at runtime. Detecting and response to changes in state quickly, understanding the scope of possible changes in schedule and resources, and specifying how to revise them at runtime are three essential challenges in mobile cooperative edge workloads.

4. Edge Computing Fundamentals

As computational capabilities expand throughout the network, a number of terms concerning solutions centered around middle-tier computation are being brought forward. Each group of companies devises and partly standardizes its own alternative approach on edge computing, fog computing and smartphone clouds. Multiple visions and objectives evolve within the context of each term. On top of network and IoT application optimization, companies also want to sell their products and remain in control of service provisioning [3]. The respective policy suggestions regarding resource and task management may hence go in contradictory directions. For example, carrier-controlled edge computing would mean explicitly segregating the ones wanting low latency and local data processing from those seeking privacy preservation. In general though, most first analysis question the use-value of mobile or edge infrastructure and intensively discuss whether ought to build on it at all. The job in hand is more exclusively on the sustainability of the networks and their ecosystems built around them as well as on the systems dynamics thereby elicited.
Edge computing proposes to extend cloud computing to the edge of the network. This offers high processing power directly “in the field” for both incoming data service provisioning on the cloud and for the effective processing of time-critical on-board data flows. It prohibits large amounts of flow-intensive data transfers between cloud and devices, increasing throughput efficiency and releasing precious network bandwidth. It diminishes response latencies by contributing additional tiers of computation between far-off devices and the cloud. New services for low-latency applications such as immediate augmentations of view or the preventive avoidance of events can now run. Stable LTE, 5G and Wi-Fi rollout combined with mobile computation is developing to be a new standard adopted by the automotive industry. However, edge computing is often falsely understood. Cloud services are already provisioned “at the edge” to wide area networks today, for example via local internet exchange points. This is not what is here argued for. Instead, edge computing requires a true hybrid paradigm between cloud and devices, where the high computational tier is controlled by the infrastructure owner and runs mobility-aware task allocation and flow management algorithms. Edge computing is here argued as an unresolved prospective global system.

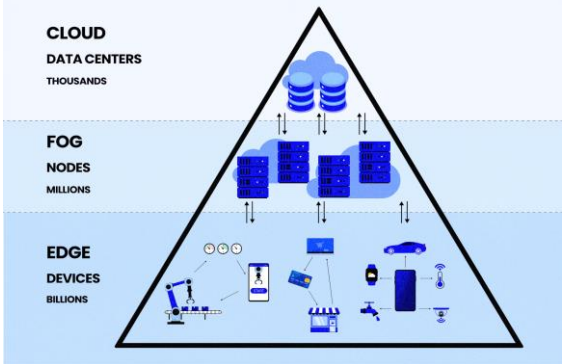

Fig 2: Edge Computing

4.1. Architecture Overview

The proposed Osmotic Architecture (OSA) is composed of a layered cloud-edge structure designed as a collection of microservices deployed on distributed nodes. Each service is lightweight and loosely coupled with the other ones through carefully defined interfaces. In this way, the stress on a single node/service can be avoided, allowing graceful operational degradation in case of overload conditions. OSA is complemented by the orchestration procedure, with the aim of creating an optimal set of services dynamically. The orchestrator aims to monitor the service utilisation in Cloud and Edge, looking for overloaded services in order to decide whether to replicate them, and checking for rarely used ones in order to terminate them. The orchestration procedure leverages a controller service that dynamically manages cloudlets and deployed services.

The OSA architecture is applied in the context of an IoT-CPS in which a collection of sensors, performing telemetry functions, is dispersed over a wide geographical area, constantly producing data measured from the relevant physical processes. The data monitored from the sensors are split in a time- and space-continuous fashion. A twofold objective of complying with both users' QoS requirements and cloud-cost efficiency concerns is addressed through the proper modelling of nodes and services. On the one hand, utility functions modelling the acceptance of modelling errors, required computing times, and allowed latencies are defined, controlling node/service types, thresholds, and switches in the architecture. On the other hand, pricing levels and variability are computed considering not only cloud- but also edge-related costs and revenues. Such cloud-cost functions are estimated through regression techniques from empirical cloud-cost profile data. The orchestration process aims to automatically configure the architecture parameters so that both objectives are optimised.

4.2. Key Technologies

The conceptual design of the proposed architecture relies on a number of cooperative and complementary technologies, such as AI workflows, automatic identification of computing and networking nodes, fast service optimization mechanisms on the edge, behind and out of the cloud. Disparate AI-supported services will be generated on the edge-to-cloud systems in 1/32 s, for example, converged intelligent tasks in healthcare, public safety and smart traffic cases. Continuous non-visual data streaming in great volume and speed is forecasted in the near future. A societal reflective and deep learning solution is suggested with incremental learning for fast service generation. Scalar service-oriented artificial data streams with ambient data streaming will be exploited to inference computational pressure.

The collaborative service optimization and mesh cloud supported by decentralized AI will efficiently distribute heavy workloads to monitored connected MEC clusters. Verification and filtering techniques in the edge-to-cloud systems can be exploited on the generated models for generalization ability and fairness. The properly trained executives at the edge-to-cloud will ensure a healthy and friendly AI in a sustainable way. A multi-typed and multi-scaled AI orchestration mechanism will administrate a hypercube swarm of AI executors to optimize AI service routing for inference accuracy and speed. A distributed and hybrid architecture of AI systems at cognitive and adaptive service. Intelligent models will be cooperatively generated at the public and proprietary clouds and migrated or transformed for fast inference at the edge. Data will be selectively held at edge nodes for learning model enhancement based on data-driven and resource-driven principles. A fine-grained resource and data trading mechanism will be developed based on the market economy to balance data security and model enhancement in a sustainable manner.

Network system resource consumption is monitored to discover out-of-scope nodes which automatically switched off to save energy. Optimal node status will be determined based on human health states in healthcare. Fast service continuity optimization or disaster recovery will significantly improve service Quality of Experience in public safety and smart traffic environments. Collaborative system provisioning optimizes heterogeneous physical and virtual network nodes across wide area networks, data center networks and edge networks. Policies based on rule-based or tree learning approaches will be enhanced with social data.

5. Cloud Computing in Mobile Infrastructure

The rapid evolution of devices, protocols, and applications in mobile infrastructure gives rise to a rich set of user-centric use cases and services. Emerging trends such as the Internet of Things and the Metaverse also increasingly demand a wealth of aware and adaptive functionalities from such a decentralized mobile infrastructure. As such, both future devices, as well as user-centric services, require stamped the importance of on-device processing, highly adaptive AI functionalities, and increased collaboration between devices. In order to enable highly versatile device-task mappings, cloud computing plays an instrumental role in mobile infrastructures as the cloud infrastructure plays a complementary role to the mobile infrastructure.

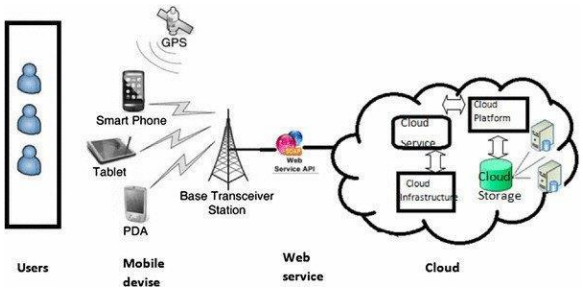

Fig 3: Mobile Cloud Computing

Cloud computing in mobile infrastructure is vital for addressing different dimensions of computation hops when developing complex services without compromising extreme real-time requirements. With the advent of 5G and beyond cellular technology, cloud computing has entered the mobile space through innovative concepts such as Mobile Cloud Computing, Mobile Edge Computing, and Fog Computing. Cloud computing in a mobile infrastructure offers new opportunities in diverse areas such as user-centric AI workloads, security workloads, and time-sensitive workloads.
Mobile edge computing: new opportunities and use cases. Increasingly powered by AI workloads, empirical workloads show that ML workloads are typically composable temporal workloads, additionally boosted by a dynamically growing number of devices or tasks. The mobile side offers new opportunities to run/accelerate such workloads by connecting distributed source and substitute data in intimate proximity and by approaching them in foresighted use-centric device-task mappings. However, highly adaptive device-edge-task mappings also impose new key challenges in offloading ML workloads across a geographically disparate topology of edge nodes, devices, and cloud infrastructure. Such visualized future waves of ML workloads, as well as newly exposed challenges by the introduction of mobile and multi-container devices, are discussed intensely.

5.1. Integration with Edge Computing

In many IoT scenarios, devices vary greatly in parameters such as computing capability, data arrival rate and accuracy. Therefore, one key is how to adapt AI workflows to heterogeneous devices, or edge-to-cloud adaptive AI workflows. This involves Mobility-Aware Deep Reinforcement Learning techniques for generic, expandable and transportable AI workflows. In this approach, edge devices replace the static locations and queues in channel-aware approaches for edge-cloud offloading. With enhanced parallelism via device collaboration, trained DDPG neural networks provide workflows to edge servers. When such an edge device has to be allocated to a new application, corrective adjustment methods with robust exploration can be applied to fine-tune the optimal scheduling policy. The extensive evaluation of the adaptive AI workflows corroborate scalable and state-of-the-art performance of transient mobile workloads. On the other hand, classical multi-agent scheduling typically focuses on the optimization objectives of general and all agents. However, for mobile offloading in the edge-cloud paradigm, since the joint decision supporters are frequently changing, learning performance may not meet expectations if all agents are considered simultaneously for training. Therefore, to avoid system confusion and abrupt failures, incumbent agents can be learned additionally after all upgrades of core agents. In addition, many deep reinforcement learning-based workloads are proposed with lots of hyperparameters to be optimized. To avoid a long costly training period before adapting to mobility, a context autoencoder with a self-attention mechanism can be devised for hyperparameter learning. After training with mobile and diverse workloads during working hours, a light and robust model can be obtained offline and transferred online as an initial step to pursue better performance, reducing the need for avoidable retraining. As such, comprehensive adaptive AI workloads with both edge-cloud capacity upgrade and online retraining efficiency can be developed, which can help shape a smart, self-adaptive and user-centric decentralized mobile internet infrastructure.

Equation 2: Energy Consumption Model

$$E_{\text{total}} = E_{\text{tx}} + E_{\text{rx}} + E_{\text{comp}}$$

Where:

- $E_{\text{tx}} = P_{\text{tx}} \cdot T_{\text{tx}}$
- $E_{\text{rx}} = P_{\text{rx}} \cdot T_{\text{rx}}$
- $E_{\text{comp}} = \kappa C f^2$ (local computation cost)

5.2. Benefits of Cloud Services

Cloud computing provides tremendous capacity to store and process data in the cloud, but not every bit of data produced and operations to be run on it benefit from such capabilities. When data is generated in the cloud, this principle is quite valid. However, with the exponential growth of portable devices, there is an increasing focus on processing and storing data on the edge of the network. Such devices produce large amounts of data, devaluing in time and becoming even more valuable with regard to context information. Having to additionally transfer such data to the cloud incurs excessive latency and costs—indeed, the success of cloud computing is partly based on high bandwidth and low-cost communication lines, which is why requirements of cloud services are constantly readjusted [3]. Nevertheless, a really big frontier lies in wireless communication: the vast majority of surface is still underexploited, meaning that at low activity levels, cloud computing is almost "free." Now consider this low activity data for which interest, e.g. in the case of urban reconnaissance, lies in patterns: using complex algorithms to compress, analyze, and categorize data in the cloud makes so much less sense than for such data to be repackaged on the edge into just one small footprint or into multiple GPS-locations for spatial representations and other more conventional format compression and only then uploaded to the cloud.

During cloud computing's explosion, nothing conclusive has surfaced with respect to cloud utilization for these edge-generated data. Not being reduced elsewhere, intensive processing requires massive storage, bound computation, and an immense contribution to investment cost—altogether compelling the need for exceedingly carefully designed, computerized automation pipelines even more to catalogue where, why, by whom, how, and at which appropriateness the data were encoded in the first place. On the other hand, the same challenges are likely to arise in the edge in the age of massive data as well, and adaptive workflow systems capable of local optimization abound. Typically, such systems are applications executed on managed planet-scale resources provisioning copious computation, control software, and storage chains. However, form wants the delivery of such systems on cheap mobile resources with limited and expensive capacity.

6. Adaptive AI Workflows

Compared to the traditional centralized cloud computing model, edge computing offers significant advantages in terms of quality of service (QoS), quality of experience (QoE), and information security. However, it also brings challenges to reliable service composition. Different attributes of edge-assisted service (EAS) candidates should be comprehensively considered by advanced composition methods. Despite the promising potential of edge-offloading composition, there are still challenges to effective service provision, mainly involving user/service composition, edge transparency, and service reliability.

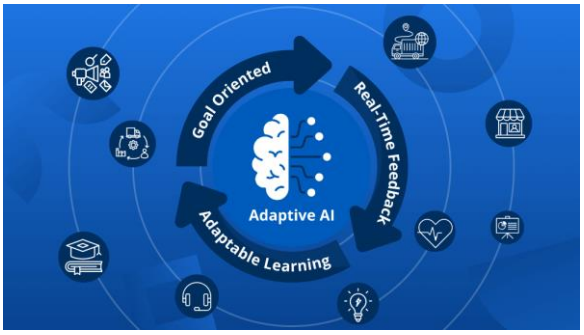

Fig 4: Adaptive AI: Operational mechanics, benefits, use cases

To tackle challenges in QoS-aware composition in the edge-assisted computing environment, an innovative dynamic QoS/QoE-aware reliable service composition framework (DREAM) is proposed. To guide its design, the multi-faceted characteristics of edge computing are recognized. To achieve adaptation to the spatio-temporal dynamically changing edge environment, portability and data-driven service metrics are examined to create transparency towards composable service candidates with mobile computation and mobile data storage. To balance edge-offloading and reliability during composition, an effective service candidate generation method is devised and a novel compositional tree structure with content addressability for integrity verification and result privacy is proposed. Advanced adaptive selection and backtracking mechanisms are designed to guarantee compositional reliability adaptively.

As a new philosophy for computing, edge-assisted computing is going to be a trend in pervasive computing systems with computationally intense IoT applications. Different from traditional cloud-oriented components that utilize centralized cloud services, edge devices offload EAS candidates in a service economy manner. Developing adaptive AI workflows for scalable, reliable, and privacy-preserving edge-offloading data IoT applications is a promising prospective. When transferring AI services from the cloud to edge devices, a new service economy regarding connected services that massively disperses AI-computerized devices are created. AI offloading can be a three-party prediction workflow composed of EAS candidates considered either a service provider or client.

6.1. Definition and Importance

The term "quality of service" (QoS) refers to the overall performance of a telecommunications or computing system and can answer several performance-related questions, such as: How fast does it transmit the data? How much is transmitted? When the data is transmitted, how long does it take? To develop the right QoS parameters for edge assistants, first, chances of non-considered application domains' parameters should be understood. For instance, in the case of applications like Amazon Alexa, voice recognition and audio streaming, response time, audio bandwidth, and jitter should be considered as QoS parameters that need to be guaranteed [5]. Adversely, for applications such as Facebook Messenger or Instagram, features related to video transmission should hold QoS parameters. These mentioned QoS parameters would be used as inputs for the service requests and service capability descriptions. Besides non-considered parameters for applications, some of the parameters should be considered as QoE ones. The parameters which answer the quality-related questions should be considered as QoE. For instance, in applications such as Instagram and WhatsApp voice messages, video transmission parameters such as color distortion, updates rate, freezes per time frame, and irresolution should be conformed as QoE parameters. After determining the domain-specific QoS and QoE parameters, their value range or domains should also be clarified. Without defining them, it would be impossible to compare the performance results with the requirements. It is not uncommon for each application domain to have a different sense of value domains for QoS parameters. This does not bring difficulties for devising the optimization functions since there exist many ways to normalize a numerical value into new values between zero and one.

6.2. Components of Adaptive AI Workflows

In the proposed modeling framework, every analysis task can be uniformly represented by AS graphs containing AI points, data points, and links, where the AS graph models the computational workflow of the task involving AI and data processing and management, namely, an Adaptive AI workflow. It is assumed that there are multiple AI points and data points, denoted by (V , E), where the nodes $vi \in V$ are the AI points carrying edge AI services and/or cloud AI services, denoted by $\{w1, \dots , wl\}$ and compatible with data points located in edge or cloud centers, respectively. The edges ej are directed edges representing the edge modes of the AI workflows, and the augmented track data stream is continuously fed into AI points, and the links $li \in E$ are the underlying communication links capable of transferring tracking data between AI and data points. It is indispensable to have a matching data point for an AI point to process data, and an AI point can access data in a portable computing mode if they are located in the same center, although it involves a high transmission delay and cost for an edge AI service to access cloud data. Formally, it has the following properties: (i) Each AI point can have at most a single data point matched and accessible; (ii) Each data point can have none or one AI point matched; and (iii) All data points must be placed in one of the centers. The AS modeling framework offers a unified representation of AI workloads and a clear understanding of the mapping decisions at the edge, where both matching and running decisions need to be made for each AI subtask based on the AI workflow and resource information. An Adaptive AI workflow is composed of multiple tasks processed in a well-defined computational workflow involving different resources and has a QoS expectation in terms of e2e latency among others, which is determined for each application type. The graph-based AS modeling framework enhances the understanding and necessary mappings of complex AI workflows modeled as AST, while the modeling and processing of dynamic workloads at the resource and workload levels could largely suggest methods for monitoring and optimizing the QoS of such workflows. Given a curation AI workflow, its QoS expectation could vary dynamically as its tasks evolve due to changing resource availability and prices in mobile environments. The limitations of handling complex AI workflows using existing additional workflow scheduling frameworks level priority processing over higher composite tasks. Two flow-based methods are derived for dynamic QoS-aware processing of Adaptive AI workflows on decentralized resource infrastructures, which consist of three parts: the task-extracting algorithm identifies the newly added workflows and tasks at the frame level. Then a low-latency service orchestrating algorithm is devised for immediate execution and provides the current-needed processing services till the end of the window. Further, the tracking-AI scheduler incrementally adapts the past creations so that they fit the changing resources while meeting the ongoing QoS expectations.

7. Workflow Design Principles

The design principles for adaptive AI workflows are identified that can migrate AI processing workloads across edge, fog, and cloud platforms considering the modifications across the implementation contexts. The design principles are first motivated by the description of the contexts.

AI workloads for adaptive AI workflows refer to the AI processing, which might consist of several processing modules or stages running in turn where the output of a stage serves as the input of the next stage. The arrival of data and tasks, the composition of the workflows, the state and parameters of AI models, running environments, etc., may all change in a dynamic manner. The design should be concerned with the aforementioned issues and indicated properties while AI workflows are adapted.

Equation 3: Resource-Constrained Scheduling

$$\sum_{v_i \in V_k} C_i \leq R_k, \quad \forall k$$

Where:

- V_k: Set of tasks on resource k
- R_k: Capacity of resource k

Stateful, while the major part or parameter set(s) of any module may change during the migration/restoration and during an ongoing execution of a module, it keeps using the originally assigned data. For example, in a training workflow model migration, the topology, weight set and trade-offs of a model may change; however, the training dataset remains the same; otherwise, it would serve as a totally different task and get classified as a re-deployment, which does not conform to the adaptive workflow. A sliding window for input streams may grow, split, move or shrink during the migration/restoration of an inference workflow, whereas the input streams keep using their valid pre-assigned and corresponding data partitions and nodes; otherwise, it would be regarded as a different execution and identified as a re-deployment.
Topology preservation: adjacency-preserving or source/sink preservation during a migration/restoration of a complete (partially designated) workflow, a workflow node is still regarded as a replica if (i) it preserves the status, which directly affects its action, (ii) otherwise, it must be cascade modified on such topological properties as the number of output/input edges. For example, in an inference workflow with an adjacency-preserving engine migration, it treats a node running a different model as a totally different replica; however, one running with a changed model [6] from a compatible family would still be regarded as the same. In a migration of a designated ODE processing node running a clustering model changed to a DBSCAN version, the outputs and affected inputs would be automatically adjusted.

7.1. Scalability

As easy as it is to constantly add more workers to a task, it's tougher when the arrangement is not uniform across and the dataset is constantly changing. Inelastic systems like these end up overprovisioning and wasting resources. Modern competitors are turning to publicly available solutions, optimizing their cloud processes, but there is little research on edge systems. Nonetheless, edge systems can significantly improve overall processing time. With the ever-rising demand for edge computing systems, it is natural for competitors to be looking for solutions in this area.
Edge computing systems are much like traditional cloud systems where data is sent to workers to be processed and results returned. However, edge systems also introduce constraints. Systems developers have little control over the network latencies between the devices, and the edges are often heterogeneous, asymmetrically limiting options about where data can be processed. With these constraints, the placement of processes becomes difficult, especially when data is constantly being added. Ideally, nodes would keep a track of the evolving streams and a suitable process placement be calculated. However, this is a huge problem with the worst case being that every device has to check against every process. Current edge solutions are designed more for individual applications and do not lend themselves to creating generalized edge systems that can be reused with relative ease on other applications.
Even with a bad lift model or complex edge locations, a well-constructed reinforcement learning-based actor-critic algorithm could probably end up within the 90th percentile on load simulation. However, it requires a huge effort to generate proper simulation data with a variety of unexpected situations, components that could be better optimised, and continuous actions. It also requires doing the same thing for large scale experiments with too many workers to set it all up to be collected in the first instance. Better etiquette, the recognition of some workers being able to handle some predictions better than others, and cloud resolution repairs during overloads could all be looked at in further research.
The data sets generated have proven to be successful in optimising metric for smaller scale simulated edges and it is continuing on the data verification piece of the problem. On real world data where edges have much less certainty and rigidity than what was generated prior it needs to see how well the edge can be optimised. Edge systems are constantly evolving and more understanding of what data could help optimise at a larger scale.

7.2. Flexibility

The IoT era has brought increasing demand for low latency processing of massive data traffic collected in various industries. Minimizing response latency could enable real time processing of emerging applications such as autonomous driving, mobile AR or smart manufacturing [4]. Given the limited network bandwidth and physical repeaters deployment costs, it will be difficult to backhaul video data streams that exceed millions of pixels per second to central cloud [6]. Edge servers locates close to IoT devices may reuse the backhaul bandwidth for multiple cameras and reduce processing latency on normal workload. This flexible data and compute resource provision solution is alarming, since non-critical workloads could be offloaded to fog servers that have stricter Quality of Experience requirements. Smart highways with low latency processing demand could be brought to edge cloud, whereas low latency processing cameras detect the physical truck

queue at the side of the highway. While all other traffic lights close, early detection and reacting could still ensure smoothness along the highway. It is necessary to adaptively offload computing workloads along with edge devices and network changes. On the other hand, with the prevailing mobile infrastructure involvement of decentralized cloud and edge computing, it will be much easier and cheaper to deploy edge service for new smart applications when considering the voluntary resource donation from the public. Non-collusion edge server clusters could be mutually managed to keep a trade off between consumption economy and service availability. All the mentioned arguments lead to the development of strong motivation for flexible adaptive cloud-edge offloading workflow execution system. There are increasing numbers of cloud-edge workflow execution systems. FLEW proposed a workflow execution framework of cloud-edge hybrid offloading. Cluster-WISE proposed a hierarchy aggregate elastic workflow engine with vertical and horizontal overlay management to share a common abstract workflow architecture. Utilizing modified scheduling methods from the able/disability detection of coalitional game theories, WISE smartly allocates workers respecting the visibility of auxiliary data nodes across its downward workflow. Both of the other systems preview potential offloading opportunities beforehand. They ignore the ad hoc interactions during or after offloading decisions, which is not compatible with privacy networks when strictly restricting the resource usage.

7.3. Resilience

Providing mobile services performance guarantees is an open challenge in a context where diverse autonomous systems cohabit with very different workloads. In this context, the decentralized processing and utilizing of computational edge resources is fundamental and must be considered to reduce delays and risks for certain critical workloads. In this work, the adaptive computing framework is further investigated with the addition of a prototype that implements safety properties on the knowledge and historical data. Stakeholders must be able to increase their privacy by increasing safety. Different levels of exposure must be made available, from knowing nothing to divulging full knowledge of all data.
Adaptive provisioning is a valuable ally, which with good tradeoffs on resource occupancy enables the inclusion of safety properties in the optimization process. Today, cloud computing deals with data providers aiming to remain private and secure while providing data to as many services as possible. The target would be to be able to collect accuracy feedback on all services, but without exposing themselves. If the loss of full knowledge is acceptable, the data could be aggregated with respect to metrics like privacy, confidence, or region. The service, afterwards, may be addressed proportionally. For stratified data evaluation, the possibility of removing clouds would need consideration, giving up system-wide accuracy. In the future, this kind of interaction would be enabled, with adaptation procedures that are much more aimed towards providing malicious intrusion resilience. Moreover, the frontier of the data providing could be expanded, rendering it more expressive for knowledge diffusion and sharing than data distribution. Finally, edge nodes could guarantee privacy properties with more restrictive methods. Today there are still possible exploitation faults that transfer security and defense during operation. Second-order toggling properties are valid formal constraints for the edge point of views, where similar behaviors are present due to weight misconfiguration and gain.

8. Data Management Strategies

Low-power wide-area (LPWA) and long-range bandwidth (LoRa Bands) technologies promise the networking of tens of millions of devices in metropolitan cities and beyond. The rapid adoption of these technologies in several countries often leads to some devices being used beyond the original design constraints, sometimes without the manufacturers' knowledge. The potential threat of unsafe, faulty, or harmful devices and the possibility that actions might be taken without human intervention give rise to concerns for action and data trust in what has often been called the Internet of Things (IoT). A decentralized approach using blockchain technology is proposed, involving multiple chains that can communicate and share data in real time. Blockchain-keeping nodes could be ideally placed in decentralized cloud computing or edge nodes across the nation. While actors or devices can market themselves to edge nodes for cache space and bandwidth, agents must diffuse reputations across chains to construct a trust model for these marketing processes.

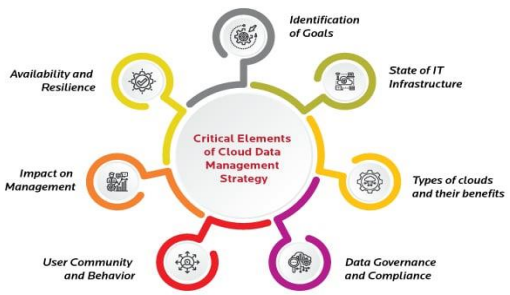

Fig 5: 7 Data Management Strategies for Business Success

Bio-inspired algorithms like cuckoo search could drive search responses to the longest lashes, while a specific auction mechanism could award rewards and penalties to search success or failure. Periodically, market allocations from the search phase would be clustered for wide dissemination and to improve short-term trust propagation consequently. Though reliable evaluations influence an agent's reputation, reducing energy usage by only responding to and diffusing reliable evaluations is ideal. Intractable queries introduce a trust factor impacting response time and model acceptance risk, and signature injection attacks targeting the underlying trust model are a serious issue.

8.1. Data Collection and Preprocessing

The data collection and preprocessing stage consists of two main components: data collection agents and virtual machines (VMs) running Kafka and Spark. The running architecture is expected to have an edge device running cloudlet and mobile devices connected to the edge device via WiFi. The cloudlet wirelessly receives data from users and connects to the cloud via a wired link, while the cloud is running HDFS and the VMs run Kafka and Spark again. The data collection agents connect to the devices and generate log data. They also allow to create intelligent persistent storage based on a fast and efficient data representation independent of the deployment platform. The agents running close to a platform collect platform data via platform-specific APIs and dump the collected data in text files in the HDFS file system. This procedure is implemented by the data collection agents deployed in the cloudlet and the cloud setup (with two VMs). Data are sent by agents via text files to external systems via TCP/UDP, which is possible in the cloud setup. An agent connected only to the end platform (closed setup) collects data but cannot send it out (DISABLE) via the data transmission method mentioned earlier due to security concerns. To retrieve data in this case, a shared folder is used, which allows to automatically send data files from the edge device to the cloud setup. Each agent is configured to run every five minutes. In total, from one hour run, some three-one-minute long recordings are created for each agent. The recording of each agent contains the data generated during the run such as the frame integral count in the 220 second interval where basic configuration changes got also implemented.

8.2. Data Storage Solutions

As the importance of data management in distributed architecture increases, new architectures provide specific solutions to these challenges in edge computing and fog computing. These new systems include new storage representations, data processing paradigms, and new protocols to access storage on devices. Solutions include using a combination of immutable, append-only logs and soft-state messages, based on the belief that the immutability of blocks would allow for the construction of very large object storage networks at the edge while still ensuring resiliency through redundancy against data corruption. This approach further lowers overhead by supporting a wide range of data replication and gossip protocols that are useful for disseminating large time-critical messages while maintaining robust fault tolerance under diverse network conditions.

New textured languages, APIs with multiplexing designs, and algorithms for allowing data processing from heterogeneous locations, caches, and logs of new data structures make cloud services available to edge devices. This approach can mobile edge operations computations through mashups of discrete operations originally defined on towers of data in the cloud. In a hybrid competency-, data-, and address-space network environment, a data warehouse-oriented architecture with a cyclic hierarchy of data organization is proposed to extract information from spatial and temporal sources and allow runs of fault-tolerant queries. While the edges of the hierarchy may include the use of cloud platforms, the interior is designed to include in router-cloud-like nodes that aggregate, predict, and model data while utilizing limited computational resources for wireless networking.

Availability and proper operation of a cloud service must at least in part depend on the operation of local edge devices, which work together in a way that wants immutable software. Centralized control is a security problem, as attackers have a single point to compromise, and the edge is a much larger space for monitoring. Efforts toward peer-to-peer approaches with an emphasis on the maintenance overhead palette used are periodically scuttled by unanticipated edge node failures or overlook the effects of mobility. Nonetheless, important considerations for a successful system include minimization of overhead and amplification of collisions in regions on top of addresses for storage.

9. AI Model Deployment

The mobile AR use case requires on-device AI inference with a small footprint, low latency, and high energy efficiency for running SqueezeNet. However, the server-side multi-device collective learning and SVC miss classification use cases require a high-energy budget for cloud AI model training, on-edge multi-device collaborative inference tasks, and server-driven AI inferences with heavy computation capacity, higher latency, and energy budget. Thus, to enable seamless service continuity between these edge, cloud, and device processing modes, the platform provides an open ecosystem-based AI model deployment manager that automatically deploys the AI model onto the most suitable processing unit at the edge, cloud, and device sides according to user policies controlling latency, energy budget, computation resource space, and data privacy. In addition, it adopts a dynamic refines mechanism to adjust the deployment of the AI model that has already been deployed and running on the processing unit to adapt to the change of user preferences. This is essential for mobile

AR use cases, as during the movement of the device side, the mobile AR application can switch from a server-driven AI model deployment to an on-device AI model deployment to achieve lower latency and energy efficiency [9] at runtime.
In the deployment architecture, the user invokes the AI model deployment manager to query its available infrastructure-related AI model deployments. The ASP side computes and returns user-requested AI model deployment results according to checking the deployed AI model compatibility, data pipeline drafting, server-side data communication ability, and edge and cloud capacity. If newly-trained AI models meet requests, the platform leverages a customized AI model packaging tool to match the deployment conditions that include input tensor and model type format filtering rules, model ID, receiver communication protocol and wired/wireless ability selection, and initial model weight saving. A compiled model is generated and securely accessed by the PB. Based on the user preferences, the deployment manager selects an optimal AI model deployment and pushes an AI model deployment application based on the K8s architecture. The application is then translated into K8s Job and Pod styles, which interact with the processing unit and instantiate pre-defined tasks injected into the provisioned processing unit with data pipeline setups for scalable computing resources-based AI inference running.

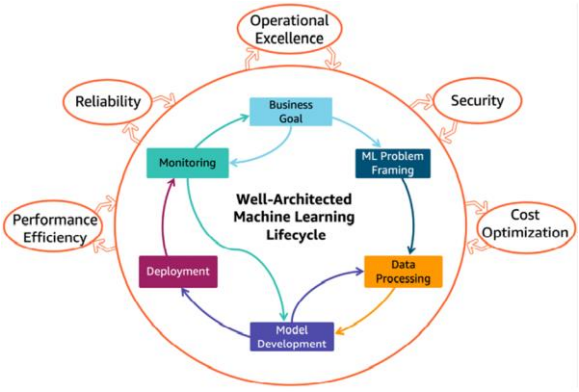

Fig 6: Best Practices for Deploying AI Models in Production

9.1. Model Training Approaches

In most existing mobile autonomous systems, the model training (MT) approaches to the observations and non-uniform experiences of the models in edge devices are often not properly considered. This results typically in a number of distributed mobile edge devices, such as IoT drones, that train their models separately and take the inference (I) from their previously trained models independently on-board. A heterogenous IoT data analysis framework consisting of cloud training and edge inference processes is proposed to hold the collaboration of the edge and cloud in performing adaptive Mobile Edge Computing (MEC) operations on non-uniform data streams in various edge devices. To protect privacy and provide real-time responses, the predicted data class and confidence are sent from the edge to the cloud instead of the raw data. Following this, proper machine learning models are pretrained on the cloud and transferred to edge devices to setup adaptive inferences for incoming unknown data streams [8].
This model training approach hurtful for their real applications, particularly, for such mobile autonomous systems in resource-constraint edge devices like drones, the MT of AI models can be unattainable due to the capacity and power limitation of the devices, and the large data volumes and model sizes in ML. Therefore, such edge devices are required to cooperate with a proper cloud, which is equipped with abundant resources and can provide a well-formed infrastructure to perform the cloud model training. Then the source filtered data groups and data integrity are requested, to which, this approach generates proper models and sends them back to the edge devices for adaptive I on the unknown non-uniform observations with the assistance of the cloud model selector and model manager. However, most existing adaptive edge-cloud processing frameworks do not take collaborations between edge devices and clouds into consideration.
As for the collaboration of our framework, on the one hand, to protect privacy, the raw data need not to be sent from edge devices to the cloud, and only the predictions before the inference block are sent to the cloud. On the other, the edge device only transfers slight structure features of its observations to the cloud, while the key requirement is how to train the models adaptively for the proper edge device models instead of flood output models. And some model instability, size, and capacity constraints also limit their collaborative abilities. Therefore, the adaptive edge-cloud processing framework is solution for such a decentralized mobile infrastructure, where the models should be able to be transferred and shared with other devices. For cloud, a generic mechanism model for a potentially broad number of edge devices is required to facilitate the I on heterogeneous edge devices.

9.2. Real-time Inference Techniques

In real life, processing images/videos often encounters inadvertent light changes [9]. APT8060 can adjust input brightness on the fly, while FAN53519 can provide dynamic power supply rail voltages to enable camera/ML core scaling. These two wide range components help provide real-time inference under changing environments. Low light detection can be

achieved with median and max pooling on hardware-complete CNNs. When there's a change in lighting condition, APT8060 monitors the captured frames, (1) If the frame brightness exceeds T1, hardware class 3 detection can continue; (2) If the frame becomes darker, both fan53519 power supply adjustment and class 2 CNNs low-light detection will be conducted. Otherwise, as the lighting's restored to normal, the latter's dynamic state will propagate to either (1) or (2) to make sure efficient inference can still happen. This captures real-world variability seamlessly. DAG implementations have flexibility & robustness from IWA to retain optimal energy efficiency for real-time inference on PULP chips. Pre-trained models are floating-point representations with an accuracy of 90.2%. Quantization reduces representation sizes by 4x, leading to negligible accuracy drops. Quantized CNNs are synthesized, placed, and routed using digital flow with SRAM as weight memory. The 5-stage pipelined architecture with dual-loop buffering is designed to maximize the use of the resource array. The input stream is stored in scratchpads for forward propagation. Kernel and MAC controllers invoke address generator hardware controllers that load weights and inputs, and output computation results from the end of PS. Post-processing includes ReLU and Hey, assuring weight buffering mechanisms are put into conscious power management. Different debugging requests output either the computation results individually or contents in wide sequential reads/sets. Depending on the triggered model, the NEG controller asserts input data type and selected weights sequentially on interfaces to transfer net parameters at the onset.

10. Performance Evaluation Metrics

Cloud computing, especially public cloud services, has been widely adopted for complex processing, thanks to its on-demand resource renting and cost-effective pricing. However, as the number of mobile devices continues to increase dramatically and the demand for seeing information becomes real-time and ubiquitous, the ensuing explosion of mobile data is overwhelming existing cloud infrastructures. The rapid growth of the Internet of Things (IoT) is bringing more devices into the computing ecosystem. Edge computing, which provides an opportunity to alleviate the burden on cloud services by offering computation services at the edge of the network, is emerging as a promising solution. Edge AI has gained interest as this technology becomes feasible. Running AI workflows at the edge can realize early insight on low-latency use cases locally and hence avoid wasting resources on irrelevant data since only filtered data will be sent to the cloud for further processing [10]. However, edge devices are more constrained regarding computing, storage, battery, and communication as they are mostly mobile. Besides, the wide variety in device performance and the complicated and dynamic environment pose challenges to running AI workloads efficiently yet robustly at the edge. Therefore, it is essential to develop adaptive AI workflows for edge-to-cloud processing based on the current resource status and low-latency requirement. AI workflows with a limited number of AI algorithms usually consist of repeated processing on historical data.

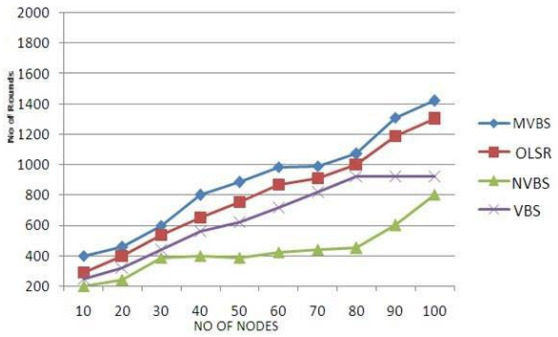

Fig: Performance evaluation graph

In comparison, with the rapid increase of IoT devices, AI workloads have become more complex, as more and more data streams simultaneously appear, and more and more workload types, such as types of detected objects or images, should be processed. In addition, with the growing number of IoT devices and rapid technological progress, constrained edge devices can be mobile and heterogeneous, and have complicated and dynamic multi-dimensional computing resources. Due to large data volume, decreasing workload concentration, and commerce consideration, edge devices are usually deployed in a decentralized manner, and each edge device can only process a subset of all workloads. All the aforementioned factors may degrade the offloading efficiency and accuracy of edge-to-cloud processing unless intelligent scheduling is adopted in a real-time manner.

To evaluate the system performance of edge-to-cloud processing on decentralized mobile infrastructure using adaptive AI workflows including edge selection and workload scheduling, various metrics need to be defined with respect to resource and workload characteristics, application requirements, and raw performance variables. In this section, these metrics are firstly reviewed, and then the metrics that are applicable to the proposed decentralized mobile infrastructure for efficient edge selection and robust workload scheduling are introduced.

10.1. Latency and Throughput

They are looking for an accurate analysis, quantization and characterization of the proposed adaptive workflows and systems. This considers latency, throughput and data management. Briefly, these metrics quantify the quality and efficiency of the proposed solutions. Latency is tangible in terms of bottleneck links and processing time. It is measurable in the workflow execution durations. Throughput is tangible at the edge in terms of how many devices can be executed in parallel. It is quantized on a per-row basis on the results table. Data management plans on data accesses and sensing adaptability are considered on the two relevant workstreams. The use of standard geo-location databases across the locations to transfer data access policy across providers and countries. Simplifying, bending data management plans can be used to reflect that edges are continuously pondering new destinations for data sends. It is hoped that each edge can continuously try to quickly exploit new nearby destinations while confirming the selection and ceasing to explore the other competitors. An approach along these lines is foreseen for wavelength flexibility systems, which use delay and contention free packet capture to improve efficacy under high traffic loads.

The approach foresees a base approximation similar to the shedding schedules used in high-volume optic systems in order to translate service constraints from edge to edge on the early buffers. However, this can be complex due to machine learning on routing instructions, including simple threshold detectors in application-specific circuits on how much to accede or stop subscribing on-to links. It has been experimented with aggregate behaviour detectors to classify or request switches, received high-level instructions identifying partitions of designs to consider migrating to an edge and event thresholds controlling how much of the hardware should keep flexible. The last terrain considers creating heterogeneous adaptive systems consisting of both deterministic and probabilistic controllers that try to either find, mislead or route efforts to attract latency-sensitive workloads or provide more aggregated resources to serve resource-sensitive workloads. This versatility is necessary in long-range applications, which account for a significant proportion of mobile services.

10.2. Energy Efficiency

The decentralized mobile infrastructure integrated with Edge platforms will consume tremendous energy for monitoring and deducing behaviors and adaptively training L-ML in edge and cloud platforms. Thus, energy efficiency is a fundamental aspect of the proposed methodology. Primarily, the work is tailored for energy efficiency in a decentralized mobile platform for real-time automatic driver behavior monitoring and deducing dangerous driver behaviors on the Edge. Additionally, there are considerable opportunities for energy-efficient programming and training of global hierarchical L-ML on cloud platforms. Finally, energy-efficient and cost-saving methodologies include using real-time objective assessments in potential automation for routine driver behavior monitoring over data fusion and transfer in the cloud infrastructure, clustering samples within the Edge, simplification of cloud programs, and sparsity of data transfer [11]. The methodologies to improve energy efficiency will rely on either programming and training provision or number and dimensionality of data inputs. Thus, the work is tailored for large-scale data inputs. The hybrid spiking neural network (hSNNs) with memristors will show superior energy efficiency in transitional Deep Learning (DL) onto the Edge and will be optimized to probe suitable computational conditions for autonomous data acquisition in the decentralized infrastructure. The optimization includes architecture optimization and adaptation to heterogeneous input applications. An energy cost model based binary search algorithm will be proposed to find optimized implementation conditions for available hardware, such as hSNN architecture optimization and weight precision.

11. Case Studies

One major challenge in new computing paradigms, like Edge AI, is dealing with the change detection in systems, while the same class of problems is studied in the field of hybrid control. Here, the state space is divided into regions of dimensional space, while control inputs in each region establish a reduced order controller that produces low complexity commands. Due to the hybrid nature of the system, arbitrary trades of regions can occur. Each time, the complete model structure remains identical (but with different parameters) and thus the limit cycles can dynamically be adjusted to the new change. Large scale models can be deployed on distributed architecture for fast inference, and it will be even necessary to perform referred services, as the model scales up toward trillion parameters. The change, thus, will be a model distribution based challenge while avoiding large communication costs.

Enabling edge-to-cloud processing in decentralized mobile infrastructure is also a major concern within the domain of architecture co-design. For distributed systems working across multiple nodes in the bus network, communication latencies due to the bus turnaround time and the effect of network re-calibration time on task management and allocation arise. This communication-centric view largely complements the edge-to-cloud processing view centering on the task distribution. An adaptation scheme meeting both communication and task distribution concerns for distributed multi-processor systems with on-chip buses is explored, while an algorithm is applied for better task management on nodes.

The proposed Edge-Cloud-Edge algorithm, taking as input the bus latency graph, low computing for cloud processing and bus dimension/resource constraints, can achieve the optimal task processing graph of a single grand task across multiple nodes in sub-exponential time. There is however a trade-off in terms of how fast a task is. If it is able to cluster large amount information processing, thus typically demanding urgent/faster hardware, allocating to edge is favorable. However

in the case that edge devices are fickle/mobile, other concerns like resource constraints and interoperability become prominent. A co-design would be for off-edge with cloud to provide enough latency supported wise decently resource allocated processing with high enough quality of service for the task to be performed. The exploration of this problem is open.

11.1. Healthcare Applications

With the rapid growth of the Internet of Medical Things (IoMT) in recent years, healthcare is becoming one of the biggest consumers of IoT technology. It is expected that there will be around 50 billion connected medical devices globally by 2025. However, intensive networked and machine learning (ML) tasks within such grids require significant computational resources, storage, and battery capacity, and may also transmit sensitive user health data. This may strain the existing transmission infrastructure and leakage of personal health data may expose users to privacy concerns. The need for performing on-premises tasks within the medical institution itself is emerging. In addition to data storage, IoT infrastructures within hospitals can be further improved by using edge servers to assist with ML tasks at the edge of the network. The desired low latency of intensive tasks such as augmented reality (AR) applications may only be satisfied if a powerful resource server is placed close to patients in the hospital or clinic. As such, many healthcare organizations, e.g., hospitals, lab offices, and multi-national companies, are progressively adopting an edge computing paradigm such that the tasks are processed at the edge of the network and data are not transferred to a public cloud. In addition to mitigating the transmission infrastructure burden, this also avoids reliability and waiting time issues associated with offloading the task to the cloud.

However, only a few hospitals or clinics may have partners equipped with such a resource service or the shareability of resources may take time, and thus hospitals may be left with computing resources only for on-device processing. In other words, the hospitals may not be equipped with enough computing resources to run state-of-the-art ML or AI tasks in practice. Also, due to the well-known health insurance portability and accountability act (HIPAA), some tasks have to avoid the cloud or resource-sharing environment altogether and should only perform the process on-device [13]. Given these complexities brought about by the evolvement of medical infrastructure from legacy devices to IoT, and also taking into consideration of latency, security, and resource availability, it becomes critical for a task to determine how the computation may have to run based on the grid of resources. Tasks that are heavy and take a long time to run may have more flexibility in resource choice whereas that are light may need to run quickly and locally in resource-constrained devices. Thus, it is critical for the IoMT system to determine for each of the published medical tasks, which of the resources provided should be used as the running platform and which of the computation should be completed locally on-device.

11.2. Smart City Initiatives

Anthropogenic and natural changes in the environment produce various types of events such as flooding, fires, landslides, extreme weather, civil unrest, etc. Smart-city applications aimed at monitoring events of this type generate a large volume of multimedia information on a continuous basis. This information is produced by a high number of different distributed sensors such as cameras, weather stations, UAVs, micro-drones, harmonized surveillance devices, and environmental and health sensors. This information is processed using a combination of edge computing technologies and a cloud data center using an architecture that has three layers. The first layer consists of video cameras and environmental sensors placed in the city that gather information on the planning scope of the smart city. The second layer consists of edge computing domain-based devices that are located at the edge of the network and that execute a preliminary processing of the data and applications based on machine learning techniques. The third layer consists of cloud applications that receive processed data from the day care and focus on tasks such as storing and centralizing data, additional processing, and application and analytic conditions requiring more computing resources or longer processing times. In addition, cloud applications can provide huge storage capabilities for massive data sets. However, with traditional computing methods and technologies, this huge amount of data cannot be processed in real time, and the information extracted from it takes a long time to be available. Large scale video monitoring local video analytics is computationally intense and time consuming. When the number of video streams is high, the bottleneck becomes the increasingly huge overall computational task of large scale video analysis.

Sensor networks using numerous types of environmental sensors have been deployed to monitor the levels of air quality parameters in the context of green smart city initiatives. Environmental sensors can consist of both outdoor and indoor node types, where outdoor nodes take measurements in different locations of the city, while indoor nodes take measurements in waiting areas, commuting areas, and so forth. The data gathered by the different types of sensors is streamed to different computational units in order to apply and execute the corresponding fusion rules, estimate the state of the environment, and assess the quality levels of the monitored parameters. These computations take place under dynamic conditions and involve the labeling of new arriving measurements and streams, updating residues, selection of streams to be transmitted toward the fusion unit, and many others, which contribute to the increasingly complex and computationally intensive batch filtering problem.

11.3. Industrial Automation

Manufacturing is undergoing a fundamental transformation in response to pressing organizational, economic, and societal challenges. The path to an Industry 4.0 era is being laid, with the digitization of a decade-long productive and process heritage. Following the COVID-19 pandemic, the urgency to reshape plants in a fast, sustainable, and resilient way is intensified. Key enabling 4.0 technologies like Artificial Intelligence (AI) and the Internet of Things (IoT) realize each block of the industrial ecosystem, including automation equipment, machines, factories, supply chains, and cities. Three intertwined trends capture a deep paradigm shift follows in business demands, computing technologies, and standards. Plants are evolving into a network of smaller, individually controllable sensing and actuation components. Collectively, these assets generate a multitude of continuous data streams, feeding numerous Decentralized Local Decision Points (DLDPs) distributed in the industrial framework. The awareness about the actual state of the processes is improves by data-driven models (a priori trained) and rules (a posteriori written). Edge-Cloud intelligence models capture the causality between input streams and output variables, which potentially enables inference, anomaly detection, classification, etc. Such a strategy is reminiscent of cloud-recognition requests for image processing. Nonetheless, modelling will be tailored on need and usage [4].

An industrial automation use case is illustrated in terms of setting, architecture, and adopted algorithms. This scenario in the mining industry consists of the prediction of impurity on iron concentrate. To preserve its economic viability, excessive ore quality degradation should be managed. Nevertheless, this is exacerbated by a constant decrement in ore concentration. To increase the recovery of ore from raw materials, tailor-made processing techniques are employed, among which flotation is an extraction process that broadly separates gangue from ore on the notions of surface chemistry. To this end, nafta, starch, and amina flow via three actuators is finely controlled over the pulp's air flow rate, gas holdup, pH, density, and froth velocity on a myriad of interconnected valves, pumps, ducts, and tanks, requiring operative proficiency from engineers and technicians [14].

12. Future Directions

AI inferencing algorithms have proliferated across various industries. By harnessing the latest developments in Mobile Edge Computing (MEC) and Decentralized Infrastructure, this increasingly complex landscape is now poised to leverage collaborative workflows of AI running at the edge, across devices. The smart mobile devices owned by users will capture information and generate models in a collaborative fashion, leveraging the in-net computation resources provided by the MEC servers. Yet yesterday, an innovative algorithm must be deployed today. But how?

The development of collaborative AI. The multiple phases of the evolution of MLS explaining the expansion of Internet connected smart devices, and the explosion of decentralized mobile infrastructure are presented. In that regard, the latest developments in AI and the potentially game-changing collaborative AI that will radically enhance the user experience of this technology and its resilience to malicious events are essential to describe, and this new paradigm needs to be adapted to large-scale decoupled infrastructures. Service-oriented multi-agent systems offer the foundation for the design of both the collaborative AI algorithms and the decentralized infrastructure they will be running on, allowing flexibility at both the application and implementation layers. In-network processing of the collaborative AI workflows. Recent agreements among industry giants around a number of open-source protocols and communication frameworks that have revolutionized the scalability, robustness, and privacy of the federated ML algorithms are described. Such algorithmic frameworks are applied to scenarios involving federated Learning, Federated Transfer Learning, Federated Model Distillation and the Decentralized GA variant of Federated Learning. Important implementation challenges and novel solutions adapting message dispatching to a decentralized infrastructure to massively reduce network overhead, or fully decentralized equivalents of the algorithms are described. Edge-to-cloud orchestration of the collaborative AI workflows. The approaches taken by different cloud providers in the edge computing domain are illustrated. An orchestration framework is designed, which will define and manage a topology of containerized compute resources on different locations and automate the deployment of workloads across them. A novel multi-layer representation of collaborative AI workflows is proposed to facilitate their deployment across the edge-to-cloud ecosystem. The composable workflows defined on a high level of abstraction are explicated into native abstractions understood by the edge and cloud platforms.

12.1. Emerging Technologies

With the development of Deep Learning, Convolutional Neural Networks (CNN) have been widely used in many fields, including image classification, face detection, and medical image processing. Nevertheless, Deep Learning places a lot of restrictions on resource-constrained IoT devices. A smart camera can be built using an Xbox camera, Raspberry Pi, and a smart tag based on an Android phone. The Android phone transmits inputs to the smart tag via a Wi-Fi Direct network, and the smart tag directs the data to the GPU-based cloud. There, CNNs are exploited in the tasks of object classification and location localization. However, it is usually costly to use binary decision tree algorithms in real-time due to their time-cost, which limits its wider application. The edge computing paradigm provides an alternative approach to approach low-latency AI modeling. A smart LED lamp is built based on a ARM Cortex A55 architecture and a smart cellphone. There, the YOLOv3-tiny model is divided into several sub-models, with a WIMbD-W device used to process the first sub-model

and a smart cellphone to cope with the others. Results show that latency is lowered from 559.279 ms to 226.459 ms and peak memory is decreased from 869.837 M to 168.928 M. To implement edge-enhanced early event burst detection, key components on general edge devices are elastically configured to offload the probing data, training, and inference.
The increasing prevalence of mobile devices can be exploited for mobile-enhanced edge cloud services. A smart IoT architecture is comprised of a mobile-enabled edge-delivering computation platform and edge service. The service can be horizontally reused by mobile devices, while the platforms with identical functionality can cooperate with each other for better resource utilization. User behaviors are modeled using the Becker model, which is an intertwined transmission network composed of mobile device service network links, edge-delivering platform links, and mobile device-plm links. This model enables the mobile-enhanced edge ecosystem to optimize and pre-simulate service flows. Remarkably, potential mobility is modeled as invisibility, which unifies the horizontal mobility of mobile devices and edge service platform mobility. The Becker model allows descriptions of previous system states to be reproduced with sample states. The Shortest Path Problem is addressed as a Markov Decision Process. To improve efficiency, state compression is studied.

12.2. Potential Research Areas

The confluence of Edge and AI techniques at the mobile edge provides opportunities for innovative paradigms, services and applications. To achieve IoT and AI for the edge, ubiquitous connectivity, edge acceleration and intelligent edge platforms are three design strategies in mobile edge computing, intelligent edge caching and edge intelligence, respectively. To seize new opportunities brought by the valuable training data stored at edge devices and for the upcoming AI and 5G era, a few promising research directions of Edge Intelligence are first discussed from different aspects, such as task allocation, deep AI, service management and the integration of multiple Edge providers.
New opportunities brought by the convergence of Edge, AI and mobile networks]. This ecosystem is different from the past fixed or cloud computing, where computation is centralized and intelligence is external to the networks. The distinctive features of edge intelligence are: (1) Decentralized due to multi-access and crowdsourcing, (2) Real-time to fulfill stringent latency and reliability constraints, (3) Ubiquitous to handle dynamic and large-scale devices and services, (4) SAM unknown due to wireless Nature of AND-Asymmetric knowledge of the provider and the mobile user. A few edge slicing mechanisms with incidental costs are proposed for DAA, Non-DAA edge services for delay-sensitive tasks are explored, and EDGE Orchestrators and Enforcers for compliance are also developed. Several edge AI frameworks across the edge cloud and mobile AI and services are proposed. Then, integrated methodologies for immediate local execution and long-term service outsourcing are proposed. Moreover, service allocation frameworks based on joint locality-aware game-theoretic schemes for DAA and Non-DAA edge service allocation are proposed, including the standards for AnYA-compliant edge services. In the new edge cloud network, training-intended service orchestration and DNN-based energy-efficient task classification are proposed.

13. Ethical Considerations

The rise of AI-enabled mobile devices brings enormous opportunities for realizing advanced and all-new services and applications. Yet, the potential of AI on mobile devices also brings challenges. One of the challenges is how to build a trustworthy AI system that respects the users' Rights, such as intended use, privacy, transparency, etc. This chapter focuses on how to design and build an end-to-end AI on mobile devices in a human-centric way. Trustworthy AI systems rely on technical means that incorporate the ethics of algorithms as well as the ethics of data.
To safeguard users' Rights, first, an ethical assessment should be conducted to understand the ethical concerns of the proposed AI system. The ethical assessment can be integrated into the design process, where representatives from multiple disciplines meet to clarify the intended use and analyze the scenario of the AI system being designed. A set of ethical guidelines pertinent to the designed AI system can then be developed. The guidelines cover high-level legal compliance as well as technology constraints that respect users' Rights. The ethical assessment and guidelines together lay a firm ground for the design of the optional technical means that ensure the trustworthy use of the AI system.
Offering general technical means to detect and mitigate ethical risks is challenging. The constraints for users' Rights may cover a wide range. The technical means are context-specific and tailor to particular use-cases. This chapter addresses how to design and engineer optional on-device and server-side technical means that ensure the trustworthy use of a federated cached AI system in a lightweight, automated, and user-controllable way. For each family of the technical means, the design constraints of optional designs are studied along with on-device and server-side implementations. It is highly practical to imitate recent successful engineering designs of the event detection model. The technical means and technology specifications that ensure the trustworthy use of federated cached AI systems are carefully designed to address users' Rights and offer users the opportunity to thoroughly understand the AI system activities and control the use of their Rights.

13.1. Privacy Concerns

Data privacy is receiving more attention with the fast development of edge-to-cloud processing in mobile environments. Motion and crowdsourcing applications raise privacy concerns about user data being collected by multiple locations and third parties via mobile devices. Content provider/service providers may impact users' lives without users' awareness. Even if consumer protection laws require the data collector and user agreement, agreements are often written in vague words, and on-site monitoring is complicated. Machine learning frameworks can learn from previous data to adaptively recommend desired services to users. However, when user data and learning models are shared by edges with clouds, private personal data may be exposed, which in turn causes the untrustworthiness of machine learning systems, and the model inversion problem. For instance, facial landmarks generated by deep learning can help reconstruct a clear image via public photos. Maintaining privacy in machine learning frameworks in cloud learning is a hot research topic. Differentially privately perturbation can guarantee the privacy of training data. However, the more public layers of a learning model, the more the private information behind layers can be reconstructed regardless of mechanisms to protect privacy. Hence, data privacy gained in a data-sharing model is not guaranteed. Edge-to-cloud processing aims to migrate on-device workloads into clouds for computation capability, power, and monetary savings. This raises fresh privacy concerns for contents and behavioral privacy. Edges should be autonomous to select what to migrate without the control of clouds. Nevertheless, the lack of understanding regarding privacy leakage limits optimal offloading strategies. Concerns including data, location, model, and algorithm privacy raised by machine learning in cloud computing and edge computing are summarized, and privacy techniques and defense mechanisms are introduced. Other aspects of privacy protection like threat models and benchmarking metrics are also discussed.

13.2. Bias in AI Models

Artificial intelligence (AI) is a logical intelligence exhibited by code. One crucial area is the development of AI services, which provide AI capabilities to users through programming interfaces. Services may be partial, providing an incomplete specification, or non-monotonic, permitting changes to beliefs. Such AI services may be learned from data and, therefore, may not behave precisely as they have been programmed. A newly learned AI service may perform well globally but poorly in a sub-area of its interpretation. As the adoption of AI services increases, the expectation of ethical decisions extends beyond humans to automated systems that act on their behalf. AI systems have successfully been applied in many areas, such as fraud detection, credit scoring, self-driving cars, and image classification. However, as the use of such systems has increased, so have concerns about their fairness, especially in safety-critical areas such as employment and criminal justice. The AI systems' unfairness generally arises from biases in the underlying data. This paper identifies and classifies the equity aspects that may apply to AI services and analyzes how a portion of these biases can be detected in completely opaque AI systems: those where the set of hypotheses or the input-output mapping is not available.

Most AI systems take into account some data-driven decision-making. Ideally, the training data should be as representative of the chosen game domain as possible to learn a useful decision-making model. Unfortunately, this is not always feasible. Bias may occur when the distribution of training data, as known in practice, is not representative of a natural phenomenon being modeled. Potentially important actions that should be taken with respect to sharing jobs or arming children, which may not have relevant input variables and are thus unobservable, may fail to be taken. reported that a hiring algorithm built for large companies produced heterogeneous results, with one firm rejecting all CVs that mentioned "women's" holidays. While this algorithm was trained on a perfectly good dataset, it learned that CVs mentioning "women's" holidays had not been relevant in the past. Given that younger candidates are less likely to mention a "women's" holiday than older candidates, there are implications for gender bias. Biases that played a role in societies have been identified in popular AI systems, such as job hiring, predictive policing, and AI systems used in the US judicial system that were biased against teenagers co-founder of even colors NN.

14. Policy Implications

No technology currently exists that can support and ensure the use of 6G new systems architecture with relevant edge-to-cloud workflows and AI platform workflows. Considering that standards are created by consensus, it is unlikely that leading companies will agree on such a common infrastructure or new standards without guarantees. The technical challenge lies in an edge infrastructure that covers mobile networks, public cloud, and hyperscale cloud that enables the provision of edge intelligence services across network layers while maintaining real-time responsiveness. The following policy recommendations provide possible approaches. This will enable the evolution of localised workflows that act on key share-and-apply intelligent functions. Prioritised use-cases should cover a wide range of data rates, latency, response times, and processing divides. Existing video security processing, augmented/reality applications, eHealth tests, and collaborative AI learning offer starting points. Trusted third parties with cross-domain oversight may be necessary to bring stakeholders together, assist with piloting and testing, and develop governance at the solution level needed across edge-to-cloud networks. Edge devices make cost decisions at speed. Public system operators in collaboration with service providers may be best placed to fund cost-effective short-term enhancements to participant devices and services. In many Mobile Edge Computing, Open Radio Access Network, and data-strategy settings, competitive pressures may drive

service coverage, quality-focus, and capabilities to be good enough for regional processing. Transparency is a key governance issue for consequential AI analytics that monitor and change behaviour. A combination of functional verification of learning models and their training data outcomes, with media accountability-based traceability, may provide mechanisms to deliver transparency through trusted oversight and auditing. Companies that can provide effective mechanisms for AI model-trust and IoT data provenance will have significant competitive advantages.

14.1. Regulatory Frameworks

Adaptive AI Workflows for Edge-to-Cloud Processing in Decentralized Mobile Infrastructure need to face an increasing set of regulation frameworks. Thus, the evolutions of regulations will be monitored and the support of the Adaptive AI Workflows to cope with them will be presented.

The first agreed-on regulation is the AI Act that attempts to regulate a wide range of AI applications based mainly on the risk behind their use, assigning to them different levels of importance. The highest risk corresponds to a list of prohibited applications. The next level down regards a tightened regulation of AI models. This risk category includes models that affect individuals by evaluating their reliability and trustworthiness and AI models directly interacting with citizens through chatbots or social robots. Optional risk categories are applications that, even if with potential high risk, are not yet so deployed or commercialized technologies.

Based on the AI Act, the Access-Only Decentralized AI Workflows are designed to allow the customer to complete the application deployment without disclosing their proprietary data or models. Pre-processing and learning are kept under control of the customer. Only the metadata are disclosed. With this architecture, even the design of the architecture is kept proprietary and unseen to competitors or adversarial parties.

Inside the methods of AdaptAware AI Workflows keeping track of trustworthiness for AI algorithms will be made available as implementation but kept with obfuscated availability. This allows to retain the processing of the style of the AI workflow, minimizing the knowledge disclosure, while encoding the different algorithms also in terms of the tools, in a way that also the number of ones processing possible a query reduces. Thus, even the sharing of the algorithms becomes unattended. Training and application of AI systems become decentralized while computational efficiency is still ensured. The sensitivity of data and difficulties in generalization are also posed as risks for further improvement of the regulation. Regulation on those grounds are in their infancy and thus the further elaboration can be part of future work where a constant monitoring together with the AI act and the AI tools is planned.

Finally, some arguments are made on the regulatory aspects touching Privacy by Design frameworks. On this risk side no further regulation tools are foreseen to be elaborated as the system will apply existing regulations.

14.2. Standards for AI Deployment

In recent years, distributed artificial intelligence (AI) has been a hot topic in academia and industry to meet the rising demand and challenges brought by widespread AI applications. Distributed AI systems are usually employed to split an AI task with high computational complexity or an AI service with many access requests among cloud-edge-end devices to improve performance. With the wide deployment of edge and end devices, such as base stations, cameras, and the Internet of Things, distributed AI systems are required to be developed and operated in a large and decentralized manner, where the edge and end devices can freely join or leave the system. For instance, since the COVID-19 outbreak, much more AI-enabled edge cameras have been deployed to help monitor social distancing and crowd density. New datasets continuously arrive at the deployed edge cameras, and previously deployed edge cameras may leave the system, such as the ones that lie idle. Therefore, it is essential to explore new distributed AI architectures and approaches tailored for decentralized mobile infrastructure.

This chapter describes a decentralized mobile AI architecture that allows edge and end devices to join or leave freely. To tackle the decentralized model training, a general framework is developed that integrates local follow-the-leader with residual message passing. To deploy lightweight and personalized models on end devices, a new federated adaptive knowledge distillation method is proposed to achieve federated adaptive model pruning and knowledge distillation: (1) personalized AI services can be facilitated for higher quality edge intelligence and lower communication efficiency; (2) heterogeneous end devices can leverage aggregated knowledge from multiple providers while keeping data samples and intermediate knowledge on-device. Finally, opportunities for further research in new distributed AI architectures and approaches are discussed [9].

15. Conclusion

For decades, mobile systems have demonstrated the power of transforming master-to-slave paradigms well supported by commodity infrastructures into decentralized systems through mobile device marketers. Mobile operators' traditional sites have continued this trend, boasting their decentralized mobile networks with cloudification for an expanded national area. This mobile decentralized infrastructure improves radio coverage, with Master 2.0 turning to Edge to handle data as a flow for improved quality of service. Yet the other extreme of the processing circuits in the cloud can cost a huge delay. The content can be totally changed from the point of collection to that of processing when viral contents are involved.

These limitations have called for adaptive AI workflows for edge-to-cloud processing, uncovering local learning opportunities as well as offloading chances via a stochastic block selection and scheduling scheme.
Combating NDA and privacy issues, mobile decentralized infrastructure, moving from a centralized scheme to an end-to-end architecture, is still hindered by the shocking requirement of training costs on massive data transferred to the cloud and the associated privacy/confidentiality issues. On-device ML unveiled possibilities of directly training/on-device inferencing the cached resources, which together with the broadcast nature of wireless channels, would lead to the adaptive MDI design to satisfy these concerns. Not only does it ease workforce recruitment, but it also paves the way for understanding time-varying social needs through deep generative models. The integration of these shallow models and MDI technologies has been studied in decentralized mobile social data implication and processing, but the cross-domain learning integrations with on-device collected networks have not been unraveled.
While mobile edge AI proved the possible complication in model communication overheads, developers nowadays lean towards ensemble modeling to avoid potential pitfalls or inject robustness, and another possible remedy to the massive model complication is to prune unimportant weights rather than filter out unimportant models or devices before communication with full models. Techniques to prune out the unimportant weights/filters have been extensively studied, but adaptive MDI with the respective aggregation in federated edge-o-cloud processing remains unexplored. The works on the use of indirect messaging are typically tailored for vertical FL, whilst the impositional understanding and settling down are still needed in the horizontal domain.

References:

[1] Venkata Krishna Azith Teja Ganti, Chandrashekar Pandugula,Tulasi Naga Subhash Polineni, Goli Mallesham (2023) Exploring the Intersection of Bioethics and AI-Driven Clinical Decision-Making: Navigating the Ethical Challenges of Deep Learning Applications in Personalized Medicine and Experimental Treatments. Journal of Material Sciences & Manufacturing Research. SRC/JMSMR-230
[2] Sondinti, K., & Reddy, L. (2023). Optimizing Real-Time Data Processing: Edge and Cloud Computing Integration for Low-Latency Applications in Smart Cities. Available at SSRN 5122027.
[3] Malempati, M., Sriram, H. K., Kaulwar, P. K., Dodda, A., & Challa, S. R. Leveraging Artificial Intelligence for Secure and Efficient Payment Systems: Transforming Financial Transactions, Regulatory Compliance, and Wealth Optimization.
[4] Chava, K. (2023). Generative Neural Models in Healthcare Sampling: Leveraging AI-ML Synergies for Precision-Driven Solutions in Logistics and Fulfillment. Available at SSRN 5135903.
[5] Komaragiri, V. B. The Role of Generative AI in Proactive Community Engagement: Developing Scalable Models for Enhancing Social Responsibility through Technological Innovations
[6] Chakilam, C. (2023). Leveraging AI, ML, and Generative Neural Models to Bridge Gaps in Genetic Therapy Access and Real-Time Resource Allocation. Global Journal of Medical Case Reports, 3(1), 1289. https://doi.org/10.31586/gjmcr.2023.1289
[7] Lahari Pandiri, Srinivasarao Paleti, Pallav Kumar Kaulwar, Murali Malempati, & Jeevani Singireddy. (2023). Transforming Financial And Insurance Ecosystems Through Intelligent Automation, Secure Digital Infrastructure, And Advanced Risk Management Strategies. Educational Administration: Theory and Practice, 29(4), 4777–4793. https://doi.org/10.53555/kuey.v29i4.9669
[8] Challa, K. Dynamic Neural Network Architectures for Real-Time Fraud Detection in Digital Payment Systems Using Machine Learning and Generative AI
[9] Mahesh Recharla, Sai Teja Nuka, Chaitran Chakilam, Karthik Chava, & Sambasiva Rao Suura. (2023). Next-Generation Technologies for Early Disease Detection and Treatment: Harnessing Intelligent Systems and Genetic Innovations for Improved Patient Outcomes. Journal for ReAttach Therapy and Developmental Diversities, 6(10s(2), 1921–1937. https://doi.org/10.53555/jrtdd.v6i10s(2).3537
[10] Phanish Lakkarasu, Pallav Kumar Kaulwar, Abhishek Dodda, Sneha Singireddy, & Jai Kiran Reddy Burugulla. (2023). Innovative Computational Frameworks for Secure Financial Ecosystems: Integrating Intelligent Automation, Risk Analytics, and Digital Infrastructure. International Journal of Finance (IJFIN) - ABDC Journal Quality List, 36(6), 334-371.
[11] Avinash Pamisetty. (2023). Integration Of Artificial Intelligence And Machine Learning In National Food Service Distribution Networks. Educational Administration: Theory and Practice, 29(4), 4979–4994. https://doi.org/10.53555/kuey.v29i4.9876
[12] Pamisetty, V. (2023). Optimizing Public Service Delivery through AI and ML Driven Predictive Analytics: A Case Study on Taxation, Unclaimed Property, and Vendor Services. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 124-149.
[13] Venkata Narasareddy Annapareddy, Anil Lokesh Gadi, Venkata Bhardwaj Komaragiri, Hara Krishna Reddy Koppolu, & Sathya Kannan. (2023). AI-Driven Optimization of Renewable Energy Systems: Enhancing Grid

Efficiency and Smart Mobility Through 5G and 6G Network Integration. Educational Administration: Theory and Practice, 29(4), 4748–4763. https://doi.org/10.53555/kuey.v29i4.9667

[14] Someshwar Mashetty. (2023). Revolutionizing Housing Finance with AI-Driven Data Science and Cloud Computing: Optimizing Mortgage Servicing, Underwriting, and Risk Assessment Using Agentic AI and Predictive Analytics. International Journal of Finance (IJFIN) - ABDC Journal Quality List, 36(6), 182-209. https://ijfin.com/index.php/ijfn/article/view/IJFIN_36_06_009

[15] Lahari Pandiri, & Subrahmanyasarma Chitta. (2023). AI-Driven Parametric Insurance Models: The Future of Automated Payouts for Natural Disaster and Climate Risk Management. Journal for ReAttach Therapy and Developmental Diversities, 6(10s(2), 1856–1868. https://doi.org/10.53555/jrtdd.v6i10s(2).3514

[16] Botlagunta Preethish Nandan, & Subrahmanya Sarma Chitta. (2023). Machine Learning Driven Metrology and Defect Detection in Extreme Ultraviolet (EUV) Lithography: A Paradigm Shift in Semiconductor Manufacturing. Educational Administration: Theory and Practice, 29(4), 4555–4568. https://doi.org/10.53555/kuey.v29i4.9495

[17] Kaulwar, P. K., Pamisetty, A., Mashetty, S., Adusupalli, B., & Pandiri, L. Harnessing Intelligent Systems and Secure Digital Infrastructure for Optimizing Housing Finance, Risk Mitigation, and Enterprise Supply Networks

[18] Srinivasarao Paleti. (2023). Data-First Finance: Architecting Scalable Data Engineering Pipelines for AI-Powered Risk Intelligence in Banking. International Journal of Finance (IJFIN) - ABDC Journal Quality List, 36(6), 403-429.

[19] Kaulwar, P. K. (2023). Tax Optimization and Compliance in Global Business Operations: Analyzing the Challenges and Opportunities of International Taxation Policies and Transfer Pricing. International Journal of Finance (IJFIN)-ABDC Journal Quality List, 36(6), 150-181.

[20] Abhishek Dodda. (2023). Digital Trust and Transparency in Fintech: How AI and Blockchain Have Reshaped Consumer Confidence and Institutional Compliance. Educational Administration: Theory and Practice, 29(4), 4921–4934. https://doi.org/10.53555/kuey.v29i4.9806

[21] Singireddy, J., & Kalisetty, S. Optimizing Tax Preparation and Filing Services: A Comparative Study of Traditional Methods and AI Augmented Tax Compliance Frameworks.

[22] Murali Malempati. (2023). A Data-Driven Framework For Real-Time Fraud Detection In Financial Transactions Using Machine Learning And Big Data Analytics. Journal for ReAttach Therapy and Developmental Diversities, 6(10s(2), 1954–1963. https://doi.org/10.53555/jrtdd.v6i10s(2).3563

[23] Malempati, M., Sriram, H. K., Kaulwar, P. K., Dodda, A., & Challa, S. R. Leveraging Artificial Intelligence for Secure and Efficient Payment Systems: Transforming Financial Transactions, Regulatory Compliance, and Wealth Optimization

[24] Phanish Lakkarasu. (2023). Generative AI in Financial Intelligence: Unraveling its Potential in Risk Assessment and Compliance. International Journal of Finance (IJFIN) - ABDC Journal Quality List, 36(6), 241-273.

[25] Ganti, V. K. A. T., Pandugula, C., Polineni, T. N. S., & Mallesham, G. Transforming Sports Medicine with Deep Learning and Generative AI: Personalized Rehabilitation Protocols and Injury Prevention Strategies for Professional Athletes.

[26] Sondinti, K., & Reddy, L. (2023). The Socioeconomic Impacts of Financial Literacy Programs on Credit Card Utilization and Debt Management among Millennials and Gen Z Consumers. Available at SSRN 5122023

[27] Hara Krishna Reddy Koppolu, Venkata Bhardwaj Komaragiri, Venkata Narasareddy Annapareddy, Sai Teja Nuka, & Anil Lokesh Gadi. (2023). Enhancing Digital Connectivity, Smart Transportation, and Sustainable Energy Solutions Through Advanced Computational Models and Secure Network Architectures. Journal for ReAttach Therapy and Developmental Diversities, 6(10s(2), 1905–1920. https://doi.org/10.53555/jrtdd.v6i10s(2).3535

[28] Kannan, S. The Convergence of AI, Machine Learning, and Neural Networks in Precision Agriculture: Generative AI as a Catalyst for Future Food Systems

[29] Sriram, H. K. (2023). Harnessing AI Neural Networks and Generative AI for Advanced Customer Engagement: Insights into Loyalty Programs, Marketing Automation, and Real-Time Analytics. Educational Administration: Theory and Practice, 29(4), 4361-4374.

[30] Chava, K. (2023). Revolutionizing Patient Outcomes with AI-Powered Generative Models: A New Paradigm in Specialty Pharmacy and Automated Distribution Systems. Available at SSRN 5136053

[31] Malviya, R. K., & Kothpalli Sondinti, L. R. (2023). Optimizing Real-Time Data Processing: Edge and Cloud Computing Integration for Low-Latency Applications in Smart Cities. Letters in High Energy Physics, 2023

[32] Challa, K. (2023). Transforming Travel Benefits through Generative AI: A Machine Learning Perspective on Enhancing Personalized Consumer Experiences. Educational Administration: Theory and Practice. Green Publication. https://doi. org/10.53555/kuey. v29i4, 9241.

[33] Pamisetty, A. (2023). AI Powered Predictive Analytics in Digital Banking and Finance: A Deep Dive into Risk Detection, Fraud Prevention, and Customer Experience Management. Fraud Prevention, and Customer Experience Management (December 11, 2023).

[34] Pamisetty, V. (2023). Intelligent Financial Governance: The Role of AI and Machine Learning in Enhancing Fiscal Impact Analysis and Budget Forecasting for Government Entities. Journal for ReAttach Therapy and Developmental Diversities, 6, 1785-1796.

[35] Pallav Kumar Kaulwar, Avinash Pamisetty, Someshwar Mashetty, Balaji Adusupalli, & Lahari Pandiri. (2023). Harnessing Intelligent Systems and Secure Digital Infrastructure for Optimizing Housing Finance, Risk Mitigation, and Enterprise Supply Networks. International Journal of Finance (IJFIN) - ABDC Journal Quality List, 36(6), 372-402. https://ijfin.com/index.php/ijfn/article/view/IJFIN_36_06_015

[36] Adusupalli, B. (2023). DevOps-Enabled Tax Intelligence: A Scalable Architecture for Real-Time Compliance in Insurance Advisory. In Journal for Reattach Therapy and Development Diversities. Green Publication. https://doi.org/10.53555/jrtdd.v6i10s(2).358

[37] Abhishek Dodda. (2023). NextGen Payment Ecosystems: A Study on the Role of Generative AI in Automating Payment Processing and Enhancing Consumer Trust. International Journal of Finance (IJFIN) - ABDC Journal Quality List, 36(6), 430-463. https://ijfin.com/index.php/ijfn/article/view/IJFIN_36_06_017

[38] Sneha Singireddy. (2023). Integrating Deep Learning and Machine Learning Algorithms in Insurance Claims Processing: A Study on Enhancing Accuracy, Speed, and Fraud Detection for Policyholders. Educational Administration: Theory and Practice, 29(4), 4764–4776. https://doi.org/10.53555/kuey.v29i4.9668

[39] Sondinti, K., & Reddy, L. (2023). Towards Quantum-Enhanced Cloud Platforms: Bridging Classical and Quantum Computing for Future Workloads. Available at SSRN 5058975

[40] Ganti, V. K. A. T., Edward, A., Subhash, T. N., & Polineni, N. A. (2023). AI-Enhanced Chatbots for Real-Time Symptom Analysis and Triage in Telehealth Services.

[41] Vankayalapati, R. K. (2023). Unifying Edge and Cloud Computing: A Framework for Distributed AI and Real-Time Processing. Available at SSRN 5048827.

[42] Annapareddy, V. N., & Seenu, A. (2023). Generative AI in Predictive Maintenance and Performance Enhancement of Solar Battery Storage Systems. Predictive Maintenance and Performance Enhancement of Solar Battery Storage Systems (December 30, 2023).

[43] Kannan, S., & Saradhi, K. S. Generative AI in Technical Support Systems: Enhancing Problem Resolution Efficiency Through AIDriven Learning and Adaptation Models.

[44] Sambasiva Rao Suura, Karthik Chava, Mahesh Recharla, & Chaitran Chakilam. (2023). Evaluating Drug Efficacy and Patient Outcomes in Personalized Medicine: The Role of AI-Enhanced Neuroimaging and Digital Transformation in Biopharmaceutical Services. Journal for ReAttach Therapy and Developmental Diversities, 6(10s(2), 1892–1904. https://doi.org/10.53555/jrtdd.v6i10s(2).3536

[45] Murali Malempati, D. P., & Rani, S. (2023). Autonomous AI Ecosystems for Seamless Digital Transactions: Exploring Neural Network-Enhanced Predictive Payment Models. International Journal of Finance (IJFIN), 36(6), 47-69.

[46] Nuka, S. T. (2023). Generative AI for Procedural Efficiency in Interventional Radiology and Vascular Access: Automating Diagnostics and Enhancing Treatment Planning. Journal for ReAttach Therapy and Developmental Diversities. Green Publication. https://doi. org/10.53555/jrtdd. v6i10s (2), 3449

[47] Koppolu, H. K. R. Deep Learning and Agentic AI for Automated Payment Fraud Detection: Enhancing Merchant Services Through Predictive Intelligence

[48] Anil Lokesh Gadi. (2023). Engine Heartbeats and Predictive Diagnostics: Leveraging AI, ML, and IoT-Enabled Data Pipelines for Real-Time Engine Performance Optimization. International Journal of Finance (IJFIN) - ABDC Journal Quality List, 36(6), 210-240. https://ijfin.com/index.php/ijfn/article/view/IJFIN_36_06_010

[49] Recharla, M., & Chitta, S. AI-Enhanced Neuroimaging and Deep Learning-Based Early Diagnosis of Multiple Sclerosis and Alzheimer's.

[50] Paleti, S. Transforming Money Transfers and Financial Inclusion: The Impact of AI-Powered Risk Mitigation and Deep Learning-Based Fraud Prevention in Cross-Border Transactions.4907-4920

[51] Moore, C. (2023). AI-powered big data and ERP systems for autonomous detection of cybersecurity vulnerabilities. Nanotechnology Perceptions, 19, 46-64.

[52] Jha, K. M., Bodepudi, V., Boppana, S. B., Katnapally, N., Maka, S. R., & Sakuru, M. (2023). Deep Learning-Enabled Big Data Analytics for Cybersecurity Threat Detection in ERP Ecosystems.

[53] Boppana, S. B., Moore, C. S., Bodepudi, V., Jha, K. M., Maka, S. R., & Sadaram, G. (2021). AI And ML Applications In Big Data Analytics: Transforming ERP Security Models For Modern Enterprises.

[54] Jha, K. M., Bodepudi, V., Boppana, S. B., Katnapally, N., Maka, S. R., & Sakuru, M. (2023). Deep Learning-Enabled Big Data Analytics for Cybersecurity Threat Detection in ERP Ecosystems.

[55] Katnapally, N., Murthy, L., & Sakuru, M. (2021). Automating Cyber Threat Response Using Agentic AI and Reinforcement Learning Techniques. J. Electrical Systems, 17(4), 138-148.

[56] Velaga, V. (2022). Enhancing Supply Chain Efficiency and Performance Through ERP Optimization Strategies.