European Journal of AI,
Computing & Informatics

Vol. 1 No. 2 2025

Article Open Access

Decision Analysis of System Architecture in Artificial Intelligence Cloud Service Model

Changming Li [1,*]

[1] Huawei Software Technologies Co. Ltd., Nanjing, Jiangsu, 210012, China
[*] Correspondence: Changming Li, Huawei Software Technologies Co. Ltd., Nanjing, Jiangsu, 210012, China

Received: 27 April 2025
Revised: 08 May 2025
Accepted: 28 May 2025
Published: 29 May 2025

Abstract: The artificial intelligence cloud service model is gradually becoming the core carrier to support the implementation of intelligent applications, and the rationality of its system architecture decision directly affects service efficiency and resource utilization. This paper focuses on the architecture design challenges in the integration scenario of artificial intelligence and cloud services, and explores the multi-dimensional balance mechanism of technology selection, cost control, and performance optimization. The research points out that system architecture decisions need to take into account algorithm complexity, data dynamics, and business flexibility requirements, and traditional architecture evaluation methods have limitations in the cloud-native environment. By introducing a multi-dimensional decision-making framework and combining cost-benefit analysis with simulation modeling tools, the comprehensive value of different architecture solutions can be quantitatively evaluated. Practical cases show that the dynamically scalable microservice architecture and serverless computing model have significant advantages in real-time inference scenarios, but one needs to be vigilant against the security risks caused by fragmented deployment. Future research should focus on the collaborative architecture design between edge intelligence and the cloud to meet the industrial needs of low latency and high concurrency.

Keywords: artificial intelligence; cloud services; system architecture; decision analysis; service models

1. Introduction

The breakthrough progress in artificial intelligence technology has spurred a profound transformation in the cloud service model, and cloud-platform-based artificial intelligence services have become the key infrastructure for enterprises' digital transformation. Currently, the surging demand for massive data processing and real-time inference scenarios makes it difficult for traditional monolithic architectures to meet the requirements of dynamic resource scheduling and elastic expansion, and system architecture design is facing dual pressures of performance bottlenecks and cost control. Although the academic community has achieved many results in areas such as cloud computing resource scheduling and distributed machine learning frameworks, the architecture decision-making methodology for artificial intelligence scenarios still shows fragmented characteristics, lacking systematic research on the coupling degree of the technology stack, the adaptability of heterogeneous hardware, and service reliability. In industrial practice, some enterprises blindly adopt cutting-edge technologies to build architectures, resulting in resource waste and increased operational complexity, which exposes the lack of scientific decision-making mechanisms [1]. This paper starts from the techno-economic perspective of architecture design, analyzes the decision-making factors under the specific constraints of artificial intelligence cloud services, aiming to construct an evaluation system that combines theoretical rigor and engineering feasibility, and provide methodological support for the efficient deployment of intelligent services.

2. Overview of Artificial Intelligence Cloud Service Model

2.1. Definition and Characteristics of Artificial Intelligence Cloud Services

A cloud-based artificial intelligence service is a technology that relies on cloud-computing infrastructure to encapsulate artificial intelligence algorithms, computing power resources, and data services into standardized module formats that can be accessed on-demand via the Internet. This model achieves dynamic resource scheduling based on distribution of computing tasks while alleviating the need for the user to understand details of diverse computing providers, thus enabling the user to flexibly call GPU clusters or Edge nodes depending on the complexity of task, overcoming limitations of local physical hardware performance and lowering the threshold for algorithm deployment. The fundamental offering relates to service scalability; the service can support the entire process from data cleaning to model training, and can also be embedded within third-party enterprise business systems as an application programming interface (API) software form to create a lightweight API calling mechanism. In terms of technical architecture, the underlying virtualization technology is decoupled from physical resources and logical services. The middle layer coordinates heterogeneous computing environments using container orchestration tools. Most importantly, the upper layer manages algorithm modules in a loosely coupled manner by utilizing the microservice design methodology, enabling aggregation of microservices. In practice, developers can quickly build inference pipelines based on pre-set machine learning frameworks, and enterprise customers can pay fees according to the actual inference volume, avoiding the problem of resource idleness due to fixed costs in traditional architectures [2].

2.2. Main Types of Artificial Intelligence Cloud Services

AI cloud services can generally be framed into three archetypal paradigms depending on their functional levels and user purposes. First, the model training platform for algorithm developers is an integrated suite that enables a complete toolchain from data annotation to hyperparameter tuning, mitigating low iteration efficiency caused by inadequate local compute access by leveraging distributed computing resources and open-source frameworks. Second, inference-as-a-service, designed for time-sensitive business responses, abstracts the demand for pre-trained models through a standardized interface, allowing enterprise customers to call image recognition or natural language processing capabilities without having to manage underlying hardware configurations, thereby minimizing product intelligence transformation cycle time. Third, automated machine-learning tools for vertical industry applications embed domain knowledge graphs and feature engineering templates to facilitate non-technical use cases, creating prediction models for fast rollout of personalized needs such as retail inventory optimization or medical image screening. The different service types complement each other in terms of resource intensity, technology coupling, and delivery speed, jointly constructing a full-chain ecosystem for the development, deployment, and application of AI.

2.3. Development Status and Trends of Artificial Intelligence Cloud Service Models

Currently, the field of artificial intelligence cloud services shows an evolutionary trend where technological iteration and market demand are deeply coupled. Mainstream cloud service providers continuously optimize the pooling capabilities of heterogeneous computing resources to support diverse scenarios ranging from the training of large models with hundreds of billions of parameters to lightweight inference at the edge. The modular encapsulation of open-source frameworks lowers the threshold for algorithm engineering. Developers can quickly build business processes suitable for financial risk control or intelligent customer service based on pre-set templates, promoting the standardization process of industry solutions. The favorable trend of hybrid-cloud architecture supports companies to adopt cross-platform model deployment strategies. The hybrid-cloud architecture assumes strong local storage for core data while enabling access to elastic computing power from the cloud to alleviate the conflict between compliance requirements and computing efficiency. At the technological front, federated learning framework has matured to change the data collaborative model, and the medical institutions and manufacturing companies can share the feature parameter under the privacy protection mechanism to break through the limitation of the data silo on the model accuracy. Concerns about energy consumption have also driven innovations in green computing technologies. Dynamic voltage and frequency scaling technologies, along with compute-sparse techniques, have been successfully adopted in the design of inference chips to alleviate the carbon footprint caused by high-density operations. The regulatory system's gradual improvement promotes the integration of trustworthy AI evaluation tools into cloud service management platform. Ideal model interpretability and bias detection function has also become one of the key elements of service providers' differentiated competition.

3. Analysis of Artificial Intelligence Cloud Service System Architecture

3.1. Basic Concepts and Classification of System Architecture

As the core framework supporting the operation of artificial intelligence cloud services, the system architecture is essentially a structured organization method of computing resources, algorithm models, and service components, and its design directly affects data processing efficiency and business response capabilities. At the technical implementation level, architecture classification usually follows the differences in service granularity and resource scheduling strategies. The monolithic architecture combines data processing and model inferences into one application process: this is fine for light inference but makes managing high-concurrency difficult; using a microservice architecture unbundles applications into independent deployed containerized modules separated according to business function, allowing the parallel scaling of image recognition and natural language processing services, utilizing hardware resources effectively, but adds complexity to operations and maintenance; and serverless architecture uses an event-driven approach to dynamically invoke function computing units of applications, with automatic scaling used for when applications are expecting heavy web traffic, such as e-commerce promotional periods or holidays to free up resources left idle for the duration, but suffers from cold-start latency. In practical application, the event-driven architecture uses messaging queues to support asynchronous communication and can address the huge amounts of time-series stream data being generated from Internet of Things devices, while the resource scheduling strategy needs to dynamically allocate GPU instances with resource scheduling and load balancing algorithms to achieve inference tasks within a finite latency threshold [3].

3.2. Components of Artificial Intelligence Cloud Service System Architecture

The system architecture of artificial intelligence cloud services is collaboratively constructed by multi-level technical components. The resource layer relies on virtualization technology to abstract the physical server cluster into an elastically deployable computing unit pool, supporting the on-demand deployment of GPU accelerators and TPU chips to meet model training requirements of varying intensities. The storage layer uses a distributed object storage system to process unstructured data streams and cooperates with an in-memory database to reduce the I/O latency during feature extraction. The middle layer manages the microservice lifecycle through a container orchestration engine, and service mesh technology handles security authentication and traffic control for inter-module communication, enabling algorithm updates without interrupting online inference services. The application layer encapsulates a standardized API gateway to connect to business systems. The natural language processing module and the computer vision engine are deployed within separate containers, allowing the development team to flexibly combine functional modules for scenarios such as medical image analysis or intelligent customer service. The horizontally expandable monitoring and alarm module tracks resource utilization and request response latency in real-time, the log analysis system captures abnormal feature values during the model inference process, and the security module integrates data encryption transmission and model watermark technology to prevent algorithm piracy.

3.3. Application Scenarios of Different System Architectures in Artificial Intelligence Cloud Services

The selection of the artificial intelligence cloud service system architecture is deeply bound to the technical characteristics and resource constraints of business scenarios. The microservice architecture, with its module decoupling and independent expansion capabilities, supports the parallel iteration of multiple models in e-commerce recommendation systems or financial anti-fraud scenarios. The containerized deployment mode allows the algorithm team to update the image recognition model version without interrupting the online service. The serverless architecture is naturally suitable for scenarios with sudden traffic surges. The social media public opinion monitoring system relies on an event-triggering mechanism to automatically scale up computing resources, maintaining the stability of sentiment analysis services during the outbreak of hot events and avoiding pushing up operating costs due to resource redundancy. The edge computing architecture deploys lightweight inference models to Internet of Things gateway devices. In the scenario of predictive maintenance of industrial equipment, sensor data is used for local anomaly detection, and only key features are uploaded to the cloud for training the global model, which alleviates bandwidth pressure and meets the demand for real-time response. The hybrid architecture shows flexibility in the field of medical image analysis. Top-tier hospitals use private clouds to handle the annotation of patients' private data and utilize public cloud supercomputing clusters for distributed model training. The trained tumor recognition model is then deployed at the edge for auxiliary diagnosis. The federated learning architecture facilitates cooperation among risk control models of cross-regional banks. Participants exchange gradient parameters in an encrypted form to update the global model, which not only complies with financial supervision requirements but also enhances the accuracy of anti-money laundering identification.

4. Influencing Factors of System Architecture Decision in Artificial Intelligence Cloud Service Mode

The decision-making for system architecture in the artificial intelligence cloud service model is influenced by three factors: technology advancement, business scenarios, and compliance risks. Technology advancement cultivates architectural design advancement. For example, container-based deployments increase the efficiency of resource scheduling, serverless architecture simplifies operation and maintenance but depends on understanding the cold-start behavior of the algorithms, and all these variations to flexibility on flexibility depend on the core problem that is being solved. Some characteristics of business scenarios influence architectural choices and future direction. High-concurrency scenarios present a preference for the elastic expanding capability whilst low-frequency task scenarios have cost control objectives to maximize their returns from cloud operational expenditures whilst still receiving quality services for their target user audience. The limitations for technology vendors are data security and privacy protection. Overall, medical services need to balance minimizing data transfer to reduce reliance on local storage of large, dense datasets with the need for timely acceptance of model updates to improve their AI algorithms. Financial applications require both end-to-end encrypted storage mechanisms and audit traceability mechanisms. Cross-regional deployment introduces the complexity of governing the data sovereignty preferences between user jurisdiction and that of the local jurisdiction. Operational sustainability at the maintenance layer is also important. Hardware choices related to edge computing nodes directly affect model inference latency at the lower layer, and the frequency of algorithm version updates relates to the rhythm of dynamic cloud resource adjustment, and the accuracy of traffic segmentation between private clusters and public services in a hybrid-cloud architecture is crucial for minimizing disaster recovery costs and ensuring effective process mitigation [4].

5. Methods for Analyzing System Architecture Decision Making in Artificial Intelligence Cloud Service Models

5.1. Multi-Attribute Decision Making Method

Multi-attribute decision-making methods in system architecture selection involve multi-dimensional trade-offs between performance indicators and constraints. The analytic hierarchy process breaks down the response speed and disaster-recovery capabilities of latency-sensitive businesses into quantifiable weight factors. When designing the architecture of a financial trading system, it is necessary to balance the priorities of a low-latency trading engine and geographically diverse and multi-active data centers. The entropy weight method automatically calculates the objective weights of resource utilization and service reliability based on historical operation and maintenance data. A video live-streaming platform dynamically adjusts the deployment strategy of edge nodes according to the correlation between bandwidth cost and stutter rate. The technique for order preference by similarity to ideal solution provides a quantitative evaluation model for architecting an e-commerce recommendation system, relating the real-time update efficiency and resource overhead of the personalized algorithm to an orthogonal decision space. The development team can weigh the sort of overall score of the microservice architecture versus the serverless architecture. The fuzzy comprehensive evaluation approach handles the non-linear indicators of a medical image cloud platform. Upgrading the PACS system in a first-class hospital requires consideration of the image's retrieval speed, the strength of data encryption, and the ability for regional medical institutions to collaborate on the same platform. Combining expert experience with Monte Carlo simulation provides a better, more definitive architecture solution. In the real implementation stage, it will be necessary to have a decision matrix constructed to track the iterative outcomes of the architecture. After an intelligent customer service system has increased its session concurrency, the threshold relationship between elastic expansion and cold-start latency would have to be recalibrated. Cloud-based decision-support tools can automatically produce optimal resource configurations under various loads.

5.2. Cost-Benefit Analysis Methods

Cost-benefit analysis focuses on the dynamic balance between resource input and business output over the entire lifecycle of the architecture. E-commerce platforms use elastic resource pools to handle traffic fluctuations during promotional activities. The automatic resource scaling strategy temporarily activates the GPU instance bidding mode during peak order periods and immediately releases resources after the event to avoid idle expenses. Video live-streaming platforms evaluate whether building their own inference clusters is more cost-effective than using cloud API calls. In scenarios with sudden traffic surges, the hybrid deployment mode saves hardware procurement costs compared to the pure private cloud solution, but the marginal costs from cross-cloud data transmission must also be considered. When deploying medical image annotation tools in a private cloud for a medical AI system, it is necessary to calculate the relationship between the compliance costs of local data storage and model iteration efficiency, and select machine configurations that support distributed training to reduce the computing power consumption of single experiments. Predictive maintenance applications for manufacturing equipment take into account the hardware investment in edge computing gateways. Local inference reduces bandwidth costs due to less frequent data transfers to the cloud, but the O&M labor expense from frequent model updates must be considered. Enterprises use the TCO model to compare the pay-as-you-go billing mode of public cloud instances with the 3-year reserved instance discount plan. Intelligent customer service organizations select the appropriate billing method based on the business growth curve, and the operations and maintenance dashboard tracks resource usage deviations from alert thresholds in real time. Dynamic cost prediction tools can simulate the cost curve based on architectural expansion. Logistics organizations input historical order data to train a resource demand prediction model and pre-expand to the economical upper limit before "Double Eleven" to mitigate the balancing act of service stability and the risk of going over budget [5].

5.3. Simulation and Emulation Methods

Simulation and emulation methods provide a verification environment that combines virtual and real elements for architecture design. Before deploying a traffic flow prediction model, the urban traffic management cloud platform constructs a digital twin road network and injects historical toll-booth data to test the throughput limits of edge computing nodes in different regions. Before the promotion season, e-commerce platforms use traffic modeling tools to conduct stress tests on the microservice architecture, simulating the collaborative efficiency between the order processing link and the inventory service in a scenario where thousands of people are concurrently snapping up products, and adjusting the backlog threshold of the message queue to prevent the avalanche effect. When the industrial Internet of Things cloud service imports equipment vibration waveform data to train a fault detection model, it uses virtualization technology to clone the production-line sensor network and verifies the inference stability of the lightweight architecture under a 5% data packet loss rate. The high-frequency financial trading system generates extreme market data streams through the Monte Carlo method to evaluate the clock synchronization accuracy between the local matching engine and the cloud-based risk control module in the hybrid-cloud architecture, and identifies potential arbitrage vulnerabilities caused by nanosecond-level delays. The medical image cloud platform creates a virtualized GPU cluster to simulate CT image retrieval requests during morning peak hours in a leading medical institution, compares the automatic expansion response speeds of traditional virtual machine deployment and containerization solutions under sudden loads, and determines the optimal instance specification ratio. The logistics scheduling system converts historical waybill data into patio-temporal simulation parameters and reproduces the intelligent path planning process of national distribution centers in the cloud to verify the degradation-handling ability of the distributed architecture when the regional network is interrupted [6].

6. Conclusion

Opting for the architecture of artificial intelligence cloud services is primarily an exercise in pursuing the Pareto optimal solution in an uncertain environment, while simultaneously considering the dynamic game between the rate of technological evolution and the impending business demands that continuously adjust the decision-making boundaries. Research suggests that while containerized deployment improves resource utilization, it may degrade cross-platform model interpretability; serverless architecture protects operational and maintenance costs, but will have to restore the cold-start algorithm module. Decision-makers must develop a matrix that aligns technological maturity with business criticality to avoid misguided decisions due to insufficient technical understanding.

Moreover, the embeddedness of emerging technologies such as federated learning and quantum computing will require future architecture design to place increased emphasis on the symbiosis between privacy protection and computation.

References

1. S. Lins, K. D. Pandl, H. Teigeler, S. Thiebes, C. Bayer, and A. Sunyaev, "Artificial intelligence as a service: classification and research directions," *Bus. Inf. Syst. Eng.*, vol. 63, pp. 441–456, 2021, doi: 10.1007/s12599-021-00708-w.
2. C. Singla, S. Kaushal, A. Verma, et al., "A hybrid computational intelligence decision making model for multimedia cloud based applications," in *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications*, Academic Press, 2018, pp. 147–157, doi: 10.1016/B978-0-12-813314-9.00007-4.
3. M. J. Kavis, Architecting the Cloud: Design Decisions for Cloud Computing Service Models (SaaS, PaaS, and IaaS). Hoboken, NJ, USA: Wiley, 2014. ISBN: 9781118826461.
4. M. Alabdulhafith, H. Saleh, H. Elmannai, Z. H. Ali, S. El-Sappagh, J. W. Hu, et al., "A clinical decision support system for edge/cloud ICU readmission model based on particle swarm optimization, ensemble machine learning, and explainable artificial intelligence," *IEEE Access*, vol. 11, pp. 100604–100621, 2023, doi: 10.1109/ACCESS.2023.3312343.
5. P. Tadejko, "Cloud cognitive services based on machine learning methods in architecture of modern knowledge management solutions," in *Data-Centric Business and Applications: Towards Software Development*, vol. 4, pp. 169–190, 2020. ISBN: 9783030347055.
6. J. Wan, J. Yang, Z. Wang, and Q. Hua, "Artificial intelligence for cloud-assisted smart factory," *IEEE Access*, vol. 6, pp. 55419–55430, 2018, doi: 10.1109/ACCESS.2018.2871724.

Disclaimer/Publisher's Note: The views, opinions, and data expressed in all publications are solely those of the individual author(s) and contributor(s) and do not necessarily reflect the views of PAP and/or the editor(s). PAP and/or the editor(s) disclaim any responsibility for any injury to individuals or damage to property arising from the ideas, methods, instructions, or products mentioned in the content.