A Taxonomy of Distributed Cloud Computing

Xuan-Qui Pham, Dong-Seong Kim
ICT Convergence Research Center, Kumoh National Institute of Technology

{pxuanqui, dskim}@ kumoh.ac.kr

Abstract

Distributed cloud computing is an appealing computing model designed to expand the reach of public cloud services to the network edge. This paper furnishes a comprehensive taxonomy of distributed cloud computing, encompassing its key players, architectural layers, major enablers, cloud service models, and coupling models. We also present some open research questions for further research.

Ⅰ. Introduction

Cloud computing is the on-demand delivery of IT resources provided as a service on a pay-as-you-go basis. Public cloud is provisioned from cloud service providers (CSPs)' data centers for open use by the public. Private cloud is provisioned for exclusive use by a single customer and hosted on-premises. Hybrid cloud is a combination of private cloud and/or traditional IT infrastructure with public cloud, all of which remain unique entities but are bound together by standardized technology that enables data and application portability. And multi-cloud is a combination of public cloud services from multiple CSPs.

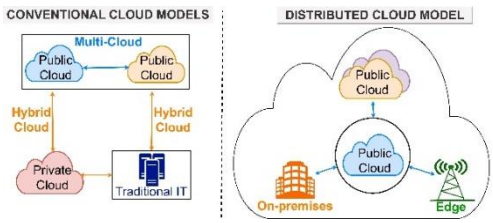

Figure 1. Distributed cloud vs. Conventional cloud

However, these deployment models have not yet delivered the expected benefits. In hybrid cloud, customers remain responsible for building, managing, operating their private cloud while hardly keeping up with the pace of innovation of CSPs. Meanwhile, multi-cloud typically encounter management complexity and fragmentation as each CSP has different management tools, technologies, and processes. Furthermore, an emerging demand exists for deploying small clouds closer to users and things to enable edge computing.

Accordingly, the emerging distributed cloud model represents an evolution from conventional cloud models [1]. Distributed cloud is defined as the distribution of public cloud services to diverse physical locations beyond the hyperscale data centers of the CSP, which retains the central responsibility for governing, operating, maintaining, and updating the distributed cloud infrastructure [2]. Meanwhile, cloud customers can access cloud services from their respective locations and manage them as a unified cloud through a single control plane. The locations include the originating CSP's public clouds, other CSPs' clouds, on-premises data centers, and network edge locations. Consequently, distributed cloud can help simplify the management of hybrid and multi-cloud and enable edge computing across diverse environments.

In this paper, we first promote the distributed cloud model in comparison to conventional cloud models, as shown in Figure 1. Subsequent sections will formulate a taxonomy of distributed cloud computing and present open research questions for further investigation.

II. A Taxonomy of Distributed Cloud Computing

A taxonomy is devised and presented in Figure 2.

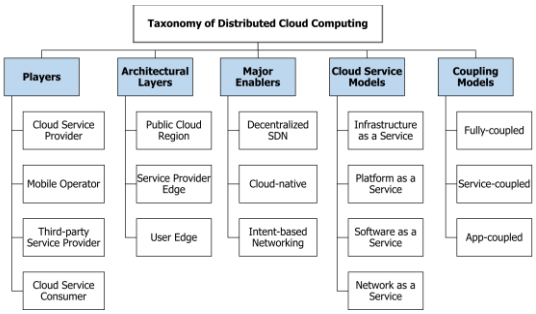

Figure 2. Distributed cloud computing taxonomy

Players: CSPs and mobile operators are the two key players in the distributed cloud ecosystem. While CSPs aim to push public cloud services as close as possible to the edge, mobile operators intend to transform their 5G network into a cloud platform to expand their offerings beyond connectivity. The distributed cloud serves as the catalyst for the partnerships between them, enabling the combination of 5G capabilities with public cloud capabilities at the edge to meet emerging application needs. Third-party service providers, such as co-location and interconnect providers, are also active participants in distributed cloud deployments. Cloud service customers refer to individual consumers, enterprises, and application developers that subscribe to and utilize distributed cloud services.

Architectural Layers: Public cloud regions serve as the starting point for the distributed cloud. Although these facilities offer economies of scale, they are often located far from most major metropolitan areas. The service provider edge refers to cloud points of presence (PoPs) where CSPs can either self-build or, more commonly, leverage existing facilities owned by its partner service providers, such as colocation providers and mobile operators. The PoPs range from the regional edge, consisting of regional data centers

Coupling models	Description	CSPs' offerings
Fully-coupled	• The CSP offers a pre-configured hardware appliance with pre-installed orchestration software. • The core cloud fully orchestrates and controls the hardware, virtualization platform, and applications of the edge cloud.	AWS Outposts, AWS Snowball Edge, Azure Stack Hub, Azure Stack Edge, and Google Distributed Cloud.
Service-coupled	• The CSP offers a portability layer typically built on Kubernetes as the foundation for services across the distributed environment. • The core cloud orchestrates and controls the virtualization environment and applications of the edge cloud. The edge cloud hardware is decoupled from the core cloud.	Amazon ECS/EKS Anywhere, Azure Arc, Azure Stack HCI, Google Anthos, and IBM Cloud Satellite
App-coupled	• The CSP offers an edge runtime for IoT applications on user edge devices. • The core cloud fully orchestrates and controls the edge cloud applications. The hardware and virtualization environment of the edge cloud are decoupled from the core cloud.	AWS IoT Greengrass and Azure IoT Edge

Table 1. Coupling models of distributed cloud infrastructure

situated adjacent to Internet exchange points, to the access edge situated in telco network facilities, such as central offices, aggregation hubs, and base stations. Meanwhile, the user edge refers to on-premises facilities, ranging from enterprise data centers to various edge devices, such as IoT gateways.

Major Enablers: Decentralized software-defined networking (SDN), cloud-native, and intent-based networking (IBN) are the major enablers. Decentralized SDN involves various architectures like hierarchical, flat, and hybrid, enabling internetworking between distributed clouds. Cloud-native relies on a microservices architecture, application programming interface, container-based infrastructure, and DevOps processes for modular and scalable application deployment across cloud environments. Meanwhile, IBN enables service orchestration in the distributed cloud by translating user intents into network policies using natural language processing and employing machine learning for optimal resource allocation and real-time issue resolution.

Cloud Service Models: Infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), and software-as-a-service (SaaS) are known as the three basic types of cloud computing service models. IaaS provides virtualized infrastructure resources, PaaS offers a platform for application development, and SaaS delivers software applications as a service. Additionally, the distributed cloud introduces network-as-a-service model that exposes network resources and functions as services through network softwarization. Furthermore, the distributed cloud drives cloud-network convergence, unifying public cloud services from CSPs and 5G network services from mobile operators into composite services delivered to end-users [2].

Coupling Models: Distributed cloud infrastructure includes the core cloud and the edge cloud. The core cloud is deployed in public cloud regions, while the edge cloud is deployed at the edge as an extension of the core cloud and is consumed in the same way. Based on the degree of coupling, such as at hardware, virtualization environment, or application and services layers, the coupling models between the core cloud and the edge cloud can be categorized into fully-coupled, service-coupled, and app-coupled. These couplings indicate how the edge cloud is orchestrated and controlled. Table 1 shows this classification with exemplified offerings from CSPs.

III. Conclusion and Open Research Questions

Distributed cloud computing is in its early stages of development, and this paper aims to devise a comprehensive taxonomy of this emerging computing paradigm. Some open research questions include: (1) What are the best strategies for ensuring effective lifecycle management of services in a distributed cloud setting? (2) How can resources be optimally allocated and managed across diverse distributed cloud environments, considering factors like varying workloads, resource availability, and performance requirements? (3) What methodologies and tools are needed for accurate performance evaluation and monitoring in distributed cloud environments? And (4) What are the best practices for securing data and communication in distributed cloud environments?

ACKNOWLEDGMENT

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-2020-0-01612) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation). It was also supported by Priority Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2018R1A6A1A03024003).

REFERENCES

[1] Gartner Top 10 Strategic Technology Trends for 2021. Accessed: Jan. 02, 2024. Available: https://www.gartner.com/smarterwithgartner/gartner-top-strategic-technology-trends-for-2021

[2] X. -Q. Pham, T. -D. Nguyen, T. Huynh-The, E. -N. Huh and D. -S. Kim, "Distributed Cloud Computing: Architecture, Enabling Technologies, and Open Challenges," IEEE Consum. Electron. Mag., vol. 12, no. 3, pp. 98-106, 1 May 2023