

Chapter 2

Resource Allocation in Serverless Computing

Ramu Kuchipudi

Chaitanya Bharathi Institute of Technology, India

Ramesh Babu Palamakula

Chaitanya Bharathi Institute of Technology, India

T. Satyanarayana Murthy

Chaitanya Bharathi Institute of Technology, India

ABSTRACT

This book chapter introduces readers to the dynamic and transformative world of serverless computing. With the focus on agility, scalability, and cost-effectiveness for developing and deploying applications, serverless computing has become a game-changer. The chapter's introduction to serverless computing provides a clear understanding of its fundamental principles and concepts, setting it apart from conventional computing models. Specifically, it explores the fundamental aspects of serverless architectures, emphasizing critical components such as functions as a service (FaaS), event driven programming, and the involvement of cloud providers in this setting. The chapter discusses the potential of serverless systems to optimize resource usage, speed up development, and adapt to changing user preferences while minimizing operational expenses. The chapter highlights practical use cases and success stories that demonstrate how serverless has been successfully implemented in various industries and domains.

1. INTRODUCTION

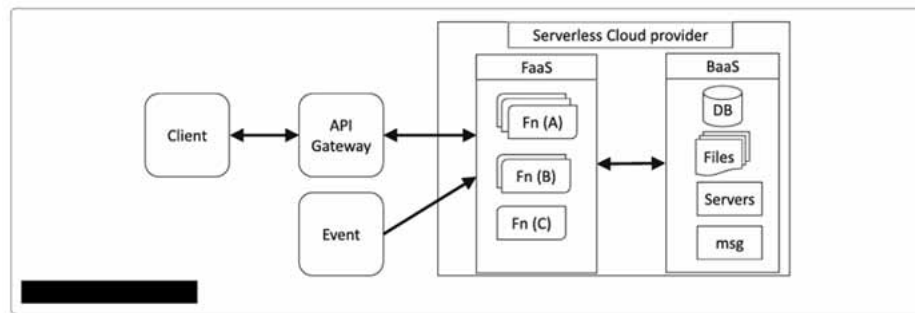
Serverless computing is a revolutionary cloud computing execution model that has transformed the way applications are built and deployed. In this model, cloud providers handle the allocation of machine resources dynamically and efficiently, relieving customers from the burdensome tasks of managing servers and infrastructure. The term “serverless” was coined by Ken Fromm in 2012, encapsulating the essence of this paradigm shift (Grobmann et al, 2019).

DOI: 10.4018/979-8-3693-1682-5.ch002

Resource Allocation in Serverless Computing

Serverless cloud computing provides backends as a service (BaaS) and functionality as a service (FaaS), such as the shown in Figure 1. BaaS includes services such as storage, messaging, and user management. FaaS, on the other hand, allows developers to deploy and run code on computing platforms. FaaS relies on services provided by BaaS, such as databases, messaging, and user authentication. FaaS is considered the most dominant model of serverless and is also referred to as “event-driven functionality.”(Villamizar et al,2017)

Figure 1. Serverless architecture



1.1 Dynamic Resource Allocation

To cope with peak loads, organizations in traditional server centric models had to provision and maintain their servers, frequently overestimating their capacity needs. Cloud providers like AWS Lambda, Azure Functions, and Google Cloud Function can allocate resources on demand through serverless computing, which enhances this approach. Moreover, this means that instead of manual capacity planning, computing resources are allocated only when a particular function or piece of code needs to be executed.

1.2 Lowering Expenses

The use of serverless computing as a pay as you go pricing model is transforming cost-conscious businesses. In standard server-based models, customers are charged for server instances regardless of whether they are actively processing requests or idling. When it comes to serverless systems, customers pay only for the computational resources used in executing functions (Kulkarni et al., 2019). The implementation of cost efficiency not only cuts down on operational expenses but also fosters experimentation and innovation.

1.3 Scalability Reduced to a Simpler Level

Scalability of applications is facilitated by serverless computing. Whenever function executions are initiated due to traffic spikes or new events, cloud providers automatically adjust the required resources to handle the load. By using elastic scaling, applications are able to adjust to changes in demand without manual adjustment. The era of adding servers or worrying about traffic spikes is gone.

8 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/resource-allocation-in-serverless-computing/343718

Related Content

Speculative Scheduling of Parameter Sweep Applications Using Job Behavior Descriptions

Attila Ulbert, László Csaba Lorincz, Tamás Kozsik and Zoltán Horváth (2009). *International Journal of Grid and High Performance Computing* (pp. 22-38).

www.irma-international.org/article/speculative-scheduling-parameter-sweep-applications/2166

Location and Provisioning Problems in Cloud Computing Networks

Federico Larumbe and Brunilde Sansò (2014). *Communication Infrastructures for Cloud Computing* (pp. 22-45).

www.irma-international.org/chapter/location-and-provisioning-problems-in-cloud-computing-networks/82529

Exploiting P2P Solutions in Telecommunication Service Delivery Platforms

Antonio Manzalini, Roberto Minerva and Corrado Moiso (2010). *Handbook of Research on P2P and Grid Systems for Service-Oriented Computing: Models, Methodologies and Applications* (pp. 937-955).

www.irma-international.org/chapter/exploiting-p2p-solutions-telecommunication-service/40834

Adaptive Routing Strategy for Large Scale Rearrangeable Symmetric Networks

Amitabha Chakrabarty, Martin Collier and Sourav Mukhopadhyay (2010). *International Journal of Grid and High Performance Computing* (pp. 53-63).

www.irma-international.org/article/adaptive-routing-strategy-large-scale/43884

Deep Analysis of Enhanced Authentication for Next Generation Networks

Mamdouh Gouda (2010). *International Journal of Grid and High Performance Computing* (pp. 37-52).

www.irma-international.org/article/deep-analysis-enhanced-authentication-next/43883