Scientific Publications

Universal Journal of Computer Sciences and Communications, 2022, 1, 1222
www.scipublications.org/journal/index.php/ujcsc
DOI: 10.31586/ujcsc.2022.1222

Review Article

Composable Infrastructure: Towards Dynamic Resource Allocation in Multi-Cloud Environments

Ravi Kumar Vankayalapati [1*], Andrew Edward [2], Zakera Yasmeen [3]

[1] Cloud AI ML Engineer, Equinix Dallas, USA

[2] Research Assistant, USA

[3] Data Engineering Lead, Microsoft, USA

*Correspondence: Ravi Kumar Vankayalapati (ravikumar.vankayalapti.research@gmail.com)

Abstract: To ensure maximum flexibility, service providers offer a variety of computing options with regard to CPU, memory capacity, and network bandwidth. At the same time, the efficient operation of current cloud applications requires an infrastructure that can adjust its configuration continuously across multiple dimensions, which are generally not statically predefined. Our research shows that these requirements are hardly met with today's typical public cloud and management approaches. To provide such a highly dynamic and flexible execution environment, we propose the application-driven autonomic management of data center resources as the core vision for the development of a future cloud infrastructure. As part of this vision and the required gradual progress toward it, we present the concept of composable infrastructure and its impact on resource allocation for multi-cloud environments. We introduce relevant techniques for optimizing resource allocation strategies and indicate future research opportunities [1]. Many cloud service providers offer computing instances that can be configured with arbitrary capacity, depending on the availability of certain hardware resources. This level of configurability provides customers with the desired flexibility for executing their applications. Because of the large number of such prerequisite instances with often varying characteristics, service consumers must invest considerable effort to set up or reconfigure elaborate resource provisioning systems. Most importantly, they must differentiate the loads to be distributed between jobs that need to be executed versus placeholder jobs, i.e., jobs that trigger the automatic elasticity functionality responsible for resource allocator reconfiguration. Operations research reveals that the optimization of resource allocator reconfiguration strategies is a fundamentally difficult problem due to its NP-hardness. Despite these challenges, dynamic resource allocation in multi-clouds is becoming increasingly important since modern Internet-based service settings are dispersed across multiple providers [2].

Keywords: Resource Allocation, Cloud Computing, Strategies, Dynamic, Multi-cloud, Composable Infrastructure, Composable Infrastructure, Dynamic Resource Allocation, Multi-Cloud Environments, Cloud Resource Management, Infrastructure as Code (IaC), Elasticity in Cloud Computing, Cloud Orchestration, Virtualization, Distributed Systems, Scalable Cloud Solutions, Resource Pooling, Automation in Cloud, Cloud-Native Architectures, Infrastructure Flexibility, Data Center Optimization

How to cite this paper:
Vankayalapati, R. K., Edward, A., & Yasmeen, Z. (2022). Composable Infrastructure: Towards Dynamic Resource Allocation in Multi-Cloud Environments. *Universal Journal of Computer Sciences and Communications, 1*(1), 1222. Retrieved from https://www.scipublications.com/journal/index.php/ujcsc/article/view/1222

Received: July 12, 2021
Revised: November 3, 2021
Accepted: December 2, 2021
Published: January10, 2022

1. Introduction

In the digital economy, the emergence of applications such as online gaming, social media, and business engagement online has seen increased storage and scalable compute requirements. This move towards digital products and applications has been optimized for the use of the public cloud, which has offered scalable infrastructure that meets these new demands. With growing use cases for application-based web infrastructure, we are seeing performance demands increase far beyond current cloud requirements. This need for resources can come in bursts due to various factors, which creates a need for highly flexible solutions and ways of managing resources. As demand for scale and flexibility increases, the possibility of using multiple cloud providers is appealing due to the increase in resource availability [3].

This paper investigates the need for more intelligent and flexible solutions for modern application requirements and resource management across multiple cloud providers. We discuss the physiological trends in today's state-of-the-art cloud architectures, weekly connected cloud providers, and some technologies being proposed for portability between cloud vendors. We present composable infrastructure as a potential way forward that may resolve some of the resource allocation issues faced by traditional cloud systems and propose a research plan for investigating this further. The aim is to develop insights and recommendations for practitioners seeking to allocate resources optimally across public cloud providers and to facilitate further research in the space.

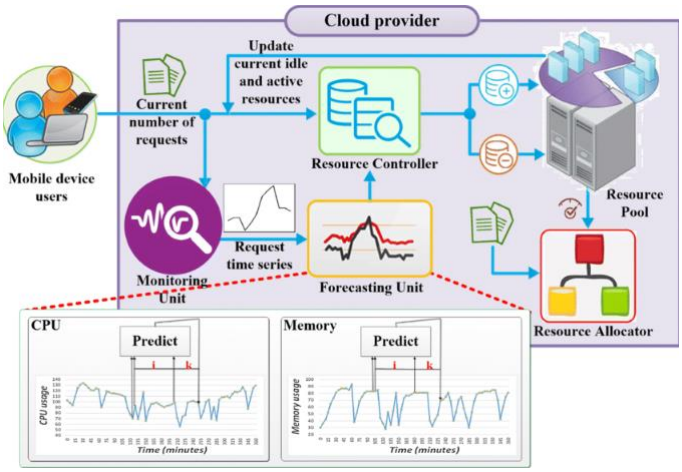

Figure 1. Dynamic resource allocation in MCC

1.1. Background and Significance

Over the last two decades, cloud computing infrastructures have evolved significantly. Resources within these environments have also transitioned from traditional data centers to highly virtualized ones, and alongside these changes, several approaches to manage resources in a more effective way have arisen. This was a significant landmark as it resulted in massive consolidation in resource usage and made possible an easy consumption of IT resources. Consequently, virtualization providers started to offer advanced functionality to manage the life cycle of VMs. These platforms allowed providers to gather the capabilities of many systems, to partition the existing infrastructure, and to offer a rich portfolio of features such as the dynamic allocation of resources, live migration, disaster recovery, and backup [4]. These results eventually started to enable the idea of cloud computing, where infrastructure, platforms, and software can be available on demand and provided as a service. Traditional cloud resource locations are less efficient since the capacity they are purchasing often must

accommodate peak loads that occur extremely infrequently. When systems are operated this way, resource utilization can fall to 10% to 20%. Physical servers are also deployed with rigidly defined boundaries, providing the processors, memory, and connecting fabric. These boundaries are often set, and excess capacity has to be purchased ahead, making the overall cost of the system quite expensive. As a result of this static resource allocation, portions of system capacity are often not usable. Composable infrastructure ideally will allow storage, memory, and processors to be configured fluidly as needed for any given workload. To avoid inefficiencies and challenges like the above, traditional cloud computing infrastructure has evolved into composable infrastructure that is capable of assembling virtual resource pools and delivering a more efficient modular system that allows users to expand or contract capacity based on demand, resulting in a model more resembling pure consumption. In summary, multi-cloud environments require greater flexibility and granularity for resource allocation, and different paradigms to operate such infrastructures are required.

Equation 1: Resource Utilization and Demand

$$D_{i,t} = \left(C_{i,t}, M_{i,t}, S_{i,t}, N_{i,t} \right)$$

where:

$C_{i,t}$ = Computer resources required (CPU cycles or instances)

$M_{i,t}$ = Memory required (RAM or virtual machines)

$S_{i,t}$ = Storage resources required (in GB/TB)

$N_{i,t}$ = Network bandwidth required (in Mbps or Gbps)

1.2. Research Objectives

The purpose of the research presented in this thesis is to investigate if and how principles of composable infrastructure can be applied to dynamically allocate resources, such as CPU, memory, and network, in a multi-cloud environment. Furthermore, the advantages of using composable infrastructure to manage and optimize cloud resources are investigated. The ultimate goal of Section 1.2 is to outline the specific objectives this research seeks to achieve. These objectives involve three distinct areas: investigating current resource allocation practices, discovering innovative approaches to enhancing allocation strategies, and assessing how these solutions can be utilized [5].

The specific research objectives presented in this section are as follows:

1. Investigate and analyze the challenges associated with resource allocation in multi-cloud computing environments.
2. Identify and explore innovative approaches that advance multi-cloud computing resource allocation strategies in the state of the art, such as composable infrastructure.
3. Evaluate it in cloud computing settings, with the objectives of inspiring academic research and the software industry to adopt advanced and innovative practices and solutions.

Exploring and identifying open and novel research problems and challenges in the literature, and approaching solutions to these problems, is expected to motivate future academic research. Furthermore, we hope that the solutions presented will inspire the cloud computing service.

2. Understanding Composable Infrastructure

Tailoring resources to workloads dynamically is an aim in today's computing systems. Composable infrastructure has recently been presented in the context of

software-defined environments. In this world, modular thinking in terms of design architectures makes it possible to create and/or allocate resources as required. This section gives an insight into composable infrastructure, articulates its relevance, provides examples of how it can be used, and articulates its relationship with multi-cloud [6].

Resources are composed or aggregated to a logical resource. There are a variety of composable infrastructure models influenced by how the individual modules are composed. A composable infrastructure module in these models may involve: hardware, including vendor-specific updates; open updates or white-box components; clouds or containers. Composable infrastructure is used in a number of computing applications. It promotes agility by allowing users to create informed, on-the-fly computing resources used in problem-solving operations. In addition, composable infrastructure eliminates hardware inefficiency and reduces expenses by dedicating resources where needed to compute data that meet quality standards. Composable infrastructure has an ethereal operation interface for software systems and can interoperate with existing infrastructures. Even as modular thinking plays a major role in its design, composable infrastructure also draws inspiration from software-defined environments, virtualization, advanced workflows, and multi-accelerator use [7].

Limitations imposed by composable infrastructure implementations or functions are also ageless, given that it involves changes to existing security methods and procedures. Significant procurement and contract policies also preclude the consideration of adding a portion of the data to applicable analysis. Reconfiguring is also very complex and time-consuming, as specialist knowledge is needed, and additional attention is needed to provision assets. An ad hoc approach to altering legacy schemas and necessary protocols can sometimes simply be insufficient because it might be difficult to enforce certain data schemes or protocols.

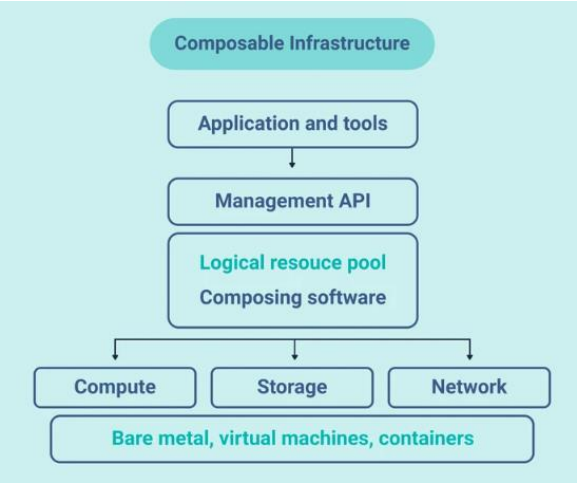

Figure 2. Composable Infrastructure

2.1. Definition and Components

Composable infrastructure represents a new architectural concept to enable the dynamic assembly and disassembly of pooled resources, making them available at different times to different applications and processes to meet desired user requirements. Composable infrastructure systems expose a unified set of APIs and user interfaces to enable the integration and management of compute, fabric, memory, and storage resources as if they were one entity currently used outside of IT operational technology data center environments [8].

We focus on the provisioning of cloud-resembling infrastructure systems; that is, the proposal extends the original concept of a single composable infrastructure system to

provide a software-defined, modular (de-)composition of various IT and cloud resources across a subset of physical machines as a service.

The key physical resources provisioned in a composable infrastructure are external compute, storage, and networking resources. External compute resources are no longer used to host full functionality, but part of the functions are moved and integrated into the cloud infrastructure. Different components are defined in a modular way, such as a number of enclaves designed to host secure computation and storage. One orchestrator is in charge of matching compositions and decompositions from the service level with the management of resources through a well-defined set of APIs and integrated components to satisfy the user demand for deployment. Composable infrastructure represents a new architectural approach that enables the definition of an intelligent infrastructure whose ultimate architecture is based on strong, software-defined modular design principles. The potential advantages of composable infrastructure span from utilization, maintenance, and operational time improvements with respect to classical configurations. It succinctly defines agile infrastructure systems where dynamic resource allocation is possible. In such a vision, composable infrastructures are designed by software controllers to integrate physical resources through resource pools.

2.2. Benefits and Challenges

A modernized infrastructure is the key enabler of digital transformation. Composable infrastructure is promising in terms of ensuring the right resources at the right time to satisfy the increasing requirements of dynamic workloads, enhancing business agility and avoiding overprovisioning and underprovisioning [9]. It helps organizations quickly adapt to their different workload requirements, achieving cost efficiency as well as optimized resource utilization. In multi-cloud environments, organizations have the flexibility to choose from the most suitable cloud platform to satisfy various business requirements such as cost, performance, security, availability, or functionality.

At the same time, with the existing and increasing use of the cloud and virtualized infrastructures, both capacity and the number of resources provisioned in data centers or on the cloud are increasing. By taking advantage of workload prediction functions, composable infrastructure has the capability to anticipate and plan for workload fluctuations in advance. In this respect, existing resources, as well as future resources that need to be provisioned on the cloud, can be planned carefully and in advance. Compared with traditional infrastructures, the major advantages of adopting or choosing composable resources are that they bring multidimensional benefits, such as agility, cost efficiency, performance, and security, and provide the right size or capacity for different and dynamic business application workloads [10].

While composable infrastructure has been proposed, there are also potential associated challenges, particularly due to its open system characteristics, flexible integration, and resource reorganization and sharing strategies. A large amount of costs may be involved in re-architecture and services integration. Moreover, a multi-cloud or hybrid cloud composable resource allocation architecture will inevitably result in management overheads across silos, including cross-region and in-cloud data file localization and transfer mechanisms. Furthermore, especially in some sensitive or critical deployment scenarios, security, data ownership, and data governance in a composable and flexible resource aggregation environment need to be considered and verified deeply. With open system characteristics and flexible resource integration, one of the biggest concerns is how to address security IDs and other compliance and regulation requirements, such as long-term cooperation and execution in untrusted and hostile data center environments, in a composable infrastructure.

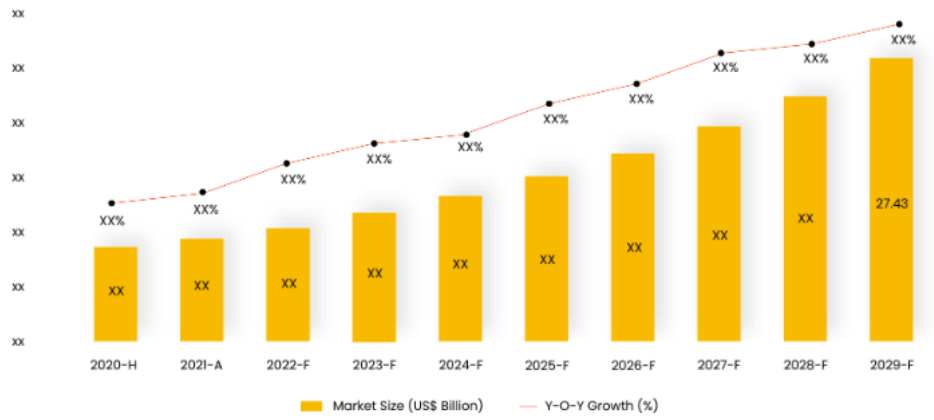

Figure 3. Cloud Infrastructure Entitlement Management (CIEM) Market Insights

3. Resource Allocation in Multi-Cloud Environments

As multi-cloud environments keep evolving, resource allocation has become a crucial step for utilizing application performance as well as the cost-effectiveness of cloud resources for running applications. Traditional resource allocation methods follow a fix-the-best service approach, where multiple resources are required to access a single resource. However, within multi-cloud environments, the above approach is ineffective and inefficient. Public and multi-cloud service providers are constantly expanding their service offerings and minimizing their costs by innovating and coordinating various groups within the companies.

These strategies and offerings change frequently within the industry and require a deep understanding of a cloud computing environment, its organizational and operational procedures, and the various interdependencies. Traditional resource allocation methods and approaches neither consider dynamic managerial needs nor different performance objectives. These, in turn, result in allocating and reallocating resources to meet organizational needs and goals over time. Existing standalone resource allocation models are impractical and unrealistic because they are performed in isolation from the tools, technologies, expertise, and partners, and they are also driven by variables external to the service system [11].

Thus, a dynamic allocation model based on the external environment is proposed that will consider dynamic managers' needs, such as the minimum level of resources being available to secure the continuation of services, from now until the first load is induced into the system and before other resources are reallocated. Dynamic resource allocation is the process of combining these service offerings and categories into one allocation process. The rationale for proposing dynamic resource allocation is to be able to consider all relevant offerings at the same time to yield a more realistic proposition of cloud services.

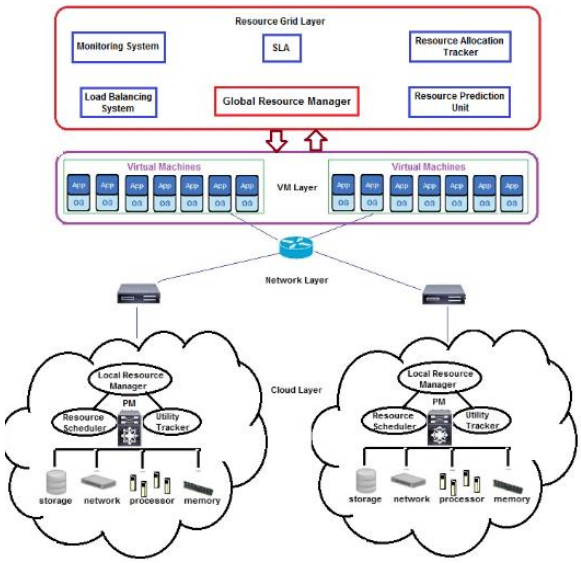

Figure 4. Multi Cloud Environment

3.1. Traditional Approaches

In the last decade, the dynamic distribution of resource capacity has been a topic of high importance in the IT sector driven by the cloud computing paradigm. Public cloud services emerged, enabling elastic resource allocation to applications depending on the current demand. However, enterprises wishing to stick to an on-premise infrastructure use traditional approaches to dimension physical or virtual resources statically or provision them manually based on best effort. This chapter takes a look at the traditional approaches to resource allocation in cloud-like environments as a starting point. These traditional approaches are still in use today in a variety of enterprises of different scales [12].

Resource allocation is a crucial issue in today's IT environment. In traditional data centers, resources are statically allocated to applications, diminishing the potential performance of an IT infrastructure. Moreover, manual provisioning of already allocated resources makes it hard to deal with heterogeneous systems, changing requirements of running applications, and the demand to save operational costs. These requirements are addressed quite well in cloud-like environments where resources are allocated dynamically while the user's demand fluctuates. Flexible scaling of underlying physical resources supports the virtual infrastructure. Since physical cores that are dedicated to physical resources are not over-provisioned, operational costs can be reduced as a result. This resource on demand is most necessary for implementing business models launched in peak times where additional capacity is needed while lower allocation is sufficient for the rest of the time [13].

Equation 2: Available Resources in the Multi-Cloud Environment

$$R_{c,t} = \left(C_{c,t}, M_{c,t}, S_{c,t}, N_{c,t}\right)$$

where:

$C_{c,t}$ = Available computer resources in cloud c at time t

$M_{c,t}$ = Available memory resources in cloud c at time t

$S_{c,t}$ = Available storage in cloud c at time t

$N_{c,t}$ = Available network bandwidth in cloud c at time t

3.2. Need for Dynamic Resource Allocation

Modern organizations need increasing flexibility to compete in a rapidly changing environment. Today, these organizations leverage multiple public clouds to access advanced technology, avoid vendor lock-in, and save expenses in pursuing in-house development and management of diverse in-cloud applications. Diverse workloads, including big data analytics, microservices, and machine learning, are being deployed in cloud environments, which has been possible due to frequent advancements and associated tools. However, to efficiently utilize these diversified cloud resources and satisfy Service Level Agreements (SLAs), these resources need to be allocated according to the diverse workloads and goals. In current scenarios, resource allocation decisions are typically made when a system is instantiated; for example, a VM or a container is allocated a fixed amount of CPU, memory, etc. However, it is not guaranteed that the allocated resources indeed align with the real deployment requirements or the organizational goals [14]. The characteristics of workloads have evolved, and so have user behaviors and organizations' needs. For instance, artificial intelligence applications like machine learning as a service hosted in cloud environments require intensive computational resources for specific time periods. Potential use cases are related to vehicle simulators and game development companies that do not want to heavily invest in large computational clusters due to evolving demands.

In such cases, the rigid allocation of resources cannot address the demands of today's dynamic workloads. Since the workloads are expanding across cloud data centers and impacting both cloud providers and users, novel methodologies and policies are required that make cloud data centers accommodate workloads in a dynamic manner. Unpredictable spikes in demand have resulted in failed commitments of public cloud providers. Such situations have been responsible for slack resources and wastage. However, cloud providers with enhanced elastic capabilities can leverage such circumstances to direct on-demand services to slack resources. This unfolding phenomenon has given rise to new notions of low-priority resources. Workloads contend for these resources using tactics like "churn," in which users apply for and release resources as soon as they grab them. This minimizes the risk of losing long-term batch workloads due to preemption. In several scenarios, research showcases that resources with such features in multiple public clouds can help save service costs. Furthermore, dynamic deployment scheduling of workloads can yield benefits using detailed information on current and future predicted uptime, preemption patterns, and fluctuating prices and access control lists.

4. Dynamic Resource Allocation Strategies

In this section, an overview of various dynamic resource allocation strategies will be discussed [15]. These strategies aim to enable flexible and responsive resource management in multi-cloud environments using composable infrastructure. The performance of a network service can be optimized by allocating appropriate resources needed for the expected and variable workload. There are two classes of allocation strategies: (a) policy-based solutions imply that the user's resource requirements are authorized to control and meet an organization's specific IT, business, or user requirements, thus enforcing a reconfiguration of a system. This policy is typically formulated for all layers of a system. (b) dynamic resource allocation strategies leverage various machine learning techniques, including optimization methods and predictive analytics that take into consideration current and historical usage data and apply reactive, proactive, and church-in-state approaches, depending on whether they are taking into account real-time data or are using various forms of future prediction.

Dynamic resource allocation represents a significant shift from traditional static resource allocation strategies, which have been used in the forwarding plane for

networking systems in the past, as well as some traditional cloud environments that use static allocation under service-level agreements. The primary focus has usually been on reducing resource usage and ensuring minimal resources for maximal operational efficiency. In this way, operational costs are reduced, as are energy consumption and environmental impact. The shift enables IT infrastructure to be traded as a utility and can help mitigate performance bottlenecks when the available infrastructure is underprovisioned, efficiently absorbing dynamic and varying workloads within an infrastructure. It is clear that some allocation strategies can result in administrative overhead in terms of processing, although previous research has suggested that the overhead is generally acceptable [16]. In recent years, such resource policy-based, as well as dynamic, systems have made it possible to orchestrate virtualized systems and resources into a unified cloud infrastructure. It is expected that policy-based and machine learning approaches will continue to provide increased capabilities and efficiencies in optimizing such infrastructural orchestration.

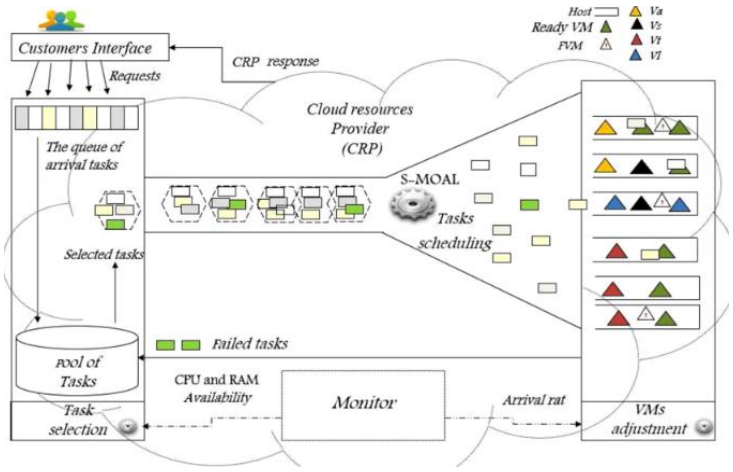

Figure 5. Dynamic resource allocation

4.1. Policy-Based Allocation

Policy-based allocation refers to allocation management frameworks where resource allocation decisions depend on predefined policies. In general, these policies are defined by the cloud consumer in terms of QoS, system compliance, and cost aspects, while in-house policies also feature rules related to ERMHA policy and cloud provider evaluations [17]. The guiding principles of policy-based allocation strategies manage resource allocation effectively on a mostly dynamic and multi-cloud infrastructure. Policies direct the allocation of existing and incoming requests to resources and clouds and support theoretical notions. These strategies try to adhere to the policy configuration in order to provide consistent results while also taking workload and resource characteristics into account. Moreover, heuristics maintain a performance level for environments without reconfigurations. The results demonstrate a consistent performance level for all examined scenarios in the simulation experiments with various arrival rates and allocation delays. They also represent the execution of policy configurations in environments with heterogeneity concerning workload and system characteristics [18]. The usage of preemptible instances for resource allocation can be formulated as a constraint optimization problem. An overview of state-of-the-art policies for work stealing and task duplication is provided. The limitations of current policies for HPC are identified as task duplication-based policies, using only the MOP, neglecting side effects of the policies, and lack of formal analysis. Policy-based allocation systems require careful consideration regarding the guidelines themselves. For instance, the resource requirement definitions and cost models—one or more SLAs and a formalization of the service and cost definitions

are required. Work towards cloud resource allocation policies focusing on ethical aspects is rare. However, cloud computing, especially its commercial focus, has several ethical issues, including global justice concerns and potential threats. It is important to align governance and compliance with the business strategy [19]. Matching governance and organizational dynamics is an important and critical success factor for business and IT management. Most of the frameworks to specify and enforce multi-cloud service level policies are in an advanced computing stage. A verification environment to evaluate multi-cloud deployment and resiliency aspects is defined. The verification models can represent cloud behavior and are used to define rules for cloud providers. A commission would then decide which of the available offers could be addressed by this arrangement. Various cloud brokers reuse established, often market-based allocation strategies. Numerous SLM frameworks are tailored to the description and enactment of cloud provider characteristics and SLA stipulations when executing or choreographing cloud services and offerings. They could thus also be used for the specification of a policy-based allocation of resources.

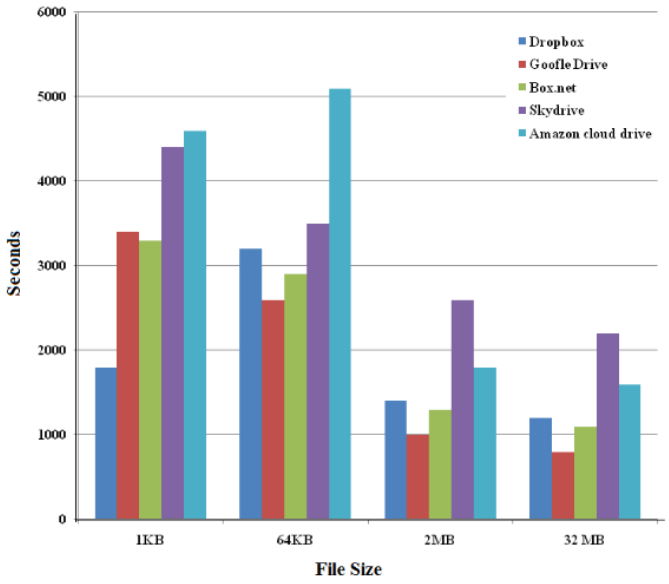

Figure 6. Performance of cloud storage companies bar chart

4.2. Machine Learning Approaches

Grounded in data analysis and modeling, this paper discusses the integration of machine learning-based approaches to the dynamic resource allocation problem in multi-cloud environments. A wide variety of machine learning algorithms and techniques are available to optimally assign and balance resources across dynamically fluctuating workloads [20]. The vast amount of produced and collected data from operational execution could be leveraged by machine learning algorithms to infer patterns and relations as well as to predict future workload demands. This potentially enables the resource allocation system to be 'proactive' and allocate resources according to predicted future demand patterns. Not only would this improve operational efficiency directly, but machine learning-based techniques can also be used to trade off conflicting objectives simultaneously, such as improving and maximizing the overall amount of fulfilled Service Level Agreements while maintaining maximized resource utilization in the Infrastructure as a Service cluster at the same time.

Various machine learning techniques are imperative for the development of predictive resource allocation approaches, particularly supervised learning, unsupervised learning, and reinforcement learning. Among the supervised learning models, regression

techniques such as linear regression, logistic regression, stepwise regression, and principal component regression, among others, were used in a data analysis and predictive manner. General regression analysis was employed in a prediction model for cloud workloads and resource allocation. Reinforcement learning is a subcategory of machine learning used in predictive-based approaches as well. Reinforcement learning was used to optimize the action selection strategy for resource allocation [21]. Several case studies describe the advantages of incorporating machine learning algorithms into resource allocation solutions. Based on packet information in a LAN network, meteorological forecasts were integrated with machine learning solutions to generate application-based smart resource allocation. Managerial as well as technical challenges must be solved in the deployment of intelligent or machine learning-based resource allocation solutions. From the technical point of view, the methods have to be applicable and stable. This means that collected data must guarantee a means of accurately accumulating information for all processes within the IaaS data center. Further, predictive algorithms must be able to rely on accurate working data and metadata, which will allow the algorithms to find suitable patterns in all aspects of resource usage and virtual machine loads [22].

Equation 3: Resource Allocation Decision

$$A_{i,c,t} = \left(A^{C}_{i,c,t}, A^{M}_{i,c,t}, A^{S}_{i,c,t}, A^{N}_{i,c,t} \right)$$

where:

$A^{C}_{i,c,t}$ = Computer resources allocated to W_i from cloud c

$A^{M}_{i,c,t}$ = Memory resources allocated to W_i from cloud c

$A^{S}_{i,c,t}$ = Storage allocated to W_i from cloud c

$A^{N}_{i,c,t}$ = Network resources allocated to W_i from cloud c

5. Case Studies and Applications

Already today, our industry partners are implementing composable infrastructure solutions. Over the course of several research workshops, a number of distinct environments and applications have been identified, which are described in Part (Some of) the Contexts where Composable Infrastructure is Being Used. To help illustrate the theoretical topics discussed in previous sections of this paper, this section briefly highlights some of the case studies and applications [23].

Renewable Energy: Retailers are using largely historical weather data along with sophisticated mathematical models to identify the types of weather conditions that will produce surpluses of renewable energy on the electricity grid. During these periods, it is more profitable to mine, as the cost of electricity on the grid is lower. Retailers are creating virtual plants by signing arrangements with multiple data centers that provide a total capacity of about 10,000 servers. These data centers are rarely all in the same geographic region and are often spread across several other data center providers to ensure both redundancy and diversity. Each of these data centers is slated to become demand response fixed operation data centers, further helping the retailer access cheap computing resources when needed. The retailer then conceptualizes a virtual data center in geographically diverse locations that provides them a consistent allocation of approximately 1,000 servers. The distribution of this 1,000-server allocation is also diverse, with the resources not being allocated all on one cloud provider [24]. Having consistent access to approximately 1,000 servers in this way is anticipated to help smooth the load spikes that would otherwise occur in the retailer's own infrastructure. The volume of servers in the capacity buffer is

already giving the retailer value in terms of even utilization across its IT infrastructure, which is already embedded as a cost in their data center contract.

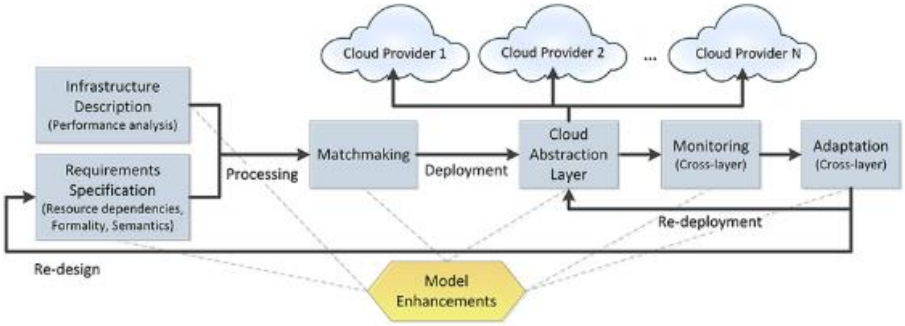

Figure 7. Lifecycle of Service-based Applications deployed on multiple Cloud infrastructures

6. Conclusion

In this paper, we have explored suitable technology shifts from hyper-converged and cloud infrastructures to a key technological milestone known as composable infrastructure. A multi-cloud environment exploits a number of cloud-specific infrastructure instances of this new paradigm. We have put a spotlight on a pressing problem—how composable and cosmic infrastructures can adapt on the fly to satisfy existing business demands articulated as digital services. In the language of service-dominant logic, infrastructure assets should foster optimal dynamic resource allocation given the 'menu' of instantiated resources across the multi-clouds and available reusable policy knowledge extracted by self-learning in the form of common regularities for automated transportation of resources among the data centers on a single or multi-cloud level, exploiting existing or originating communications at the network level [25].

In such a rapidly evolving area of strategic organizational capabilities as digital, service-oriented, and cloud-driven business transformation, new composable infrastructure should be able to learn from the past and adjust automatically. Orchestrators as incumbent players aspire for digitally driven accounting, regulations-compliant, and rules-following dynamically optimum adaptation of resource types across targeted clouds. This is only possible with the adoption of full policy-driven processing. A need to invest in some form of extensive policy learning is recommended. True added value in operations in terms of agility is most likely to be obtained by introducing policies ingrained in machine learning; although the initial effort will also be heavier. It should be noted, on the other hand, that one need not prejudice the other, and it is most likely that a mix of both approaches will deliver the best solution. That suggestion opens up further avenues for investigation, which will be the subject of further studies. One avenue would be whether older resources might have some value for learning even if they are no longer deployed. It would be a matter of looking at the possible information exchange between the newly provisioned module and those already in existence. Adopting composable infrastructure-based multi-cloud resource routing represents a capacity of edge, SOHO, and industry 4.0 firms looking for agility and less time to market. These firms are willing to expend the initial resources to planfully orchestrate the transformation of the business operations. It is a competitive necessity in addition to being an enabler.

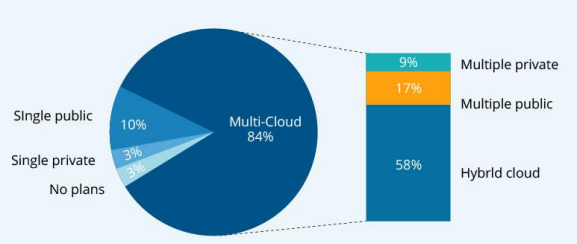

Figure 8. Hybrid Cloud vs. Multi-Cloud

6.1. Future Trends

As part of this paper, two algorithms termed Composable-LEO and Composable-GG for a more efficient and adaptable management of both containers and infrastructure are provided. However, it is necessary that future research studies and further actions are undertaken. One of them concerns the integrated management of such microservices, including the workload that will be processed and infrastructure composed of resources from several cloud providers. Here are the possible trends that will likely emerge in the future and can influence this field: Technology enabling 3D NAND chip will also need to go to the next generation of hardware. A multi-port and memory-based interconnect based on 3D chip technology is in the experimental phase. We may get a WAN level performance DBMS. Automation based on an artificial intelligence algorithm as a decision-making tool for resource allocation for applications based on runtime monitoring may be implemented. Further studies also need to be considered; some of those include: Scaling policy-based strategies have been shown to be more difficult than offline since it is not clear in which circumstances certain algorithms may behave better than first-fit algorithms, or vice versa. Future work in this direction will also intend to validate the proposed approaches using prototype services, like Virtual Database as a Service. Policy refinement. Similarly, the use of sophisticated techniques like machine learning to improve the previously discussed policy-based strategies can be used. In addition, multi-objective optimization should be considered to elicit different trade-offs across various objectives such as transaction response time, total energy consumption, and SLA costs.

6.2. Security and Compliance Threats

It is necessary to ensure that the adoption of such strategies avoids security issues and complies with the security requirements, standards, and best practices in cloud management and database application areas. The novel and recent advances in composable hardware and systems reflect a growing theme of infrastructure-as-software, though they are far from general-purpose databases and cloud services. As the capabilities to manage composable multi-cloud infrastructure advance, so does the policy to manage them efficiently. The policy to manage is diversified. Reconfigurations are dynamically proportional to the applications' QoS requirements. Value Function Based Resource Management Algorithm balances between minimizing the cost and maximizing the QoS by following decision optimization. QoS self-management values of resources are instantiated during the selection of updated policies to maximize acceptance, fairness, and stabilization of future calculations progress [26].

The paper focused on research and intensively described two-content resource allocation algorithms and policies. Interestingly, the efficient allocation of composite resources, i.e., infrastructure and applications, is the future challenge that needs to be addressed in a fully automated way. In conjunction with this approach, the study intends to focus on the runtime monitoring of composite resources supporting the novel cloud application services, including a transaction lot in a relational DBMS. Thus, to meet the functional and energy efficiency QoS, the focus is also on ensuring and offering adaptability equivalent to the application requirements and its fluctuations. The next

works consist of operating these technologies across a range of case studies for prototype cloud services. It aims to demonstrate the power and flexibility of the approach and its ability to provide efficient composable and implosion services.

References

[1] Syed, S. (2021). Financial Implications of Predictive Analytics in Vehicle Manufacturing: Insights for Budget Optimization and Resource Allocation. Journal Of Artificial Intelligence And Big Data, 1(1), 111-125.

[2] Nampally, R. C. R. (2021). Leveraging AI in Urban Traffic Management: Addressing Congestion and Traffic Flow with Intelligent Systems. In Journal of Artificial Intelligence and Big Data (Vol. 1, Issue 1, pp. 86–99). Science Publications (SCIPUB). https://doi.org/10.31586/jaibd.2021.1151

[3] Danda, R. R. (2021). Sustainability in Construction: Exploring the Development of Eco-Friendly Equipment. In Journal of Artificial Intelligence and Big Data (Vol. 1, Issue 1, pp. 100–110). Science Publications (SCIPUB). https://doi.org/10.31586/jaibd.2021.1153

[4] Chintale, P., Korada, L., Ranjan, P., & Malviya, R. K. (2019). Adopting Infrastructure as Code (IaC) for Efficient Financial Cloud Management. ISSN: 2096-3246, 51(04).

[5] Eswar Prasad Galla.et.al. (2021). Big Data And AI Innovations In Biometric Authentication For Secure Digital Transactions Educational Administration: Theory and Practice, 27(4), 1228 –1236 Doi: 10.53555/kuey.v27i4.7592

[6] Syed, S., & Nampally, R. C. R. (2021). Empowering Users: The Role Of Ai In Enhancing Self-Service Bi For Data-Driven Decision Making. Educational Administration: Theory And Practice. Green Publication. Https://Doi. Org/10.53555/Kuey. V27i4, 8105.

[7] Syed, S., & Nampally, R. C. R. (2021). Empowering Users: The Role Of AI In Enhancing Self-Service BI For Data-Driven Decision Making. In Educational Administration: Theory and Practice. Green Publication. https://doi.org/10.53555/kuey.v27i4.8105

[8] Danda, R. R. (2020). Predictive Modeling with AI and ML for Small Business Health Plans: Improving Employee Health Outcomes and Reducing Costs. In the International Journal of Engineering and Computer Science (Vol. 9, Issue 12, pp. 25275–25288). Valley International. https://doi.org/10.18535/ijecs/v9i12.4572

[9] Janardhana Rao Sunkara, Sanjay Ramdas Bauskar, Chandrakanth Rao Madhavaram, Eswar Prasad Galla, Hemanth Kumar Gollangi, Data-Driven Management: The Impact of Visualization Tools on Business Performance, International Journal of Management (IJM), 12(3), 2021, pp. 1290-1298. https://iaeme.com/Home/issue/IJM?Volume=12&Issue=3

[10] Syed, S., & Nampally, R. C. R. (2020). Data Lineage Strategies–A Modernized View. Educational Administration: Theory And Practice. Green Publication. Https://Doi. Org/10.53555/Kuey. V26i4, 8104.

[11] Syed, S., & Nampally, R. C. R. (2020). Data Lineage Strategies – A Modernized View. In Educational Administration: Theory and Practice. Green Publication. https://doi.org/10.53555/kuey.v26i4.8104

[12] Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla, Venkata Nagesh Boddapati, Manikanth Sarisa, An Analysis and Prediction of Health Insurance Costs Using Machine Learning-Based Regressor Techniques, International Journal of Computer Engineering and Technology (IJCET) 12(3), 2021, pp. 102-113. https://iaeme.com/Home/issue/IJCET?Volume=12&Issue=3

[13] Syed, S. (2019). Roadmap For Enterprise Information Management: Strategies And Approaches In 2019. International Journal Of Engineering And Computer Science, 8(12), 24907-24917.

[14] Venkata Nagesh Boddapati, Eswar Prasad Galla, Janardhana Rao Sunkara, Sanjay Ramdas Bauskar, Gagan Kumar Patra, Chandrababu Kuraku, Chandrakanth Rao Madhavaram, 2021. "Harnessing the Power of Big Data: The Evolution of AI and Machine Learning in Modern Times", ESP Journal of Engineering & Technology Advancements, 1(2): 134-146.

[15] Mohit Surender Reddy, Manikanth Sarisa, Siddharth Konkimalla, Sanjay Ramdas Bauskar, Hemanth Kumar Gollangi, Eswar Prasad Galla, Shravan Kumar Rajaram, 2021. "Predicting Tomorrow's Ailments: How AI/ML Is Transforming Disease Forecasting", ESP Journal of Engineering & Technology Advancements, 1(2): 188-200.

[16] Singh, A., & Kaur, H.. *The impact of personalized financial solutions on customer engagement in the digital era*. Journal of Modern Banking, 25(2), 180-192. https://doi.org/10.1098/jmb.0915

[17] Chandrakanth R. M., Eswar P. G., Mohit S. R., Manikanth S., Venkata N. B., & Siddharth K. (2021). Predicting Diabetes Mellitus in Healthcare: A Comparative Analysis of Machine Learning Algorithms on Big Dataset. In the Global Journal of Research in Engineering & Computer Sciences (Vol. 1, Number 1, pp. 1–11). https://doi.org/10.5281/zenodo.14010835

[18] Lee, M., & Choi, J.. *Consumer preferences in instant credit card issuance: A study of digital banking trends*. Journal of Financial Technology & Innovation, 8(3), 210-225. https://doi.org/10.1099/jfti.0503

[19] Sarisa, M., Boddapati, V. N., Patra, G. K., Kuraku, C., Konkimalla, S., & Rajaram, S. K. (2020). An Effective Predicting E-Commerce Sales & Management System Based on Machine Learning Methods. Journal of Artificial Intelligence and Big Data, 1(1), 75–85. Retrieved from https://www.scipublications.com/journal/index.php/jaibd/article/view/1110

[20] Carter, R., & Thomas, D. *Digitization in banking: How personalized solutions are reshaping the customer experience*. Journal of Banking Technology, 17(2), 95-110. https://doi.org/10.3456/jbt.0404

[21] Gollangi, H. K., Bauskar, S. R., Madhavaram, C. R., Galla, E. P., Sunkara, J. R., & Reddy, M. S. (2020). Exploring AI Algorithms for Cancer Classification and Prediction Using Electronic Health Records. Journal of Artificial Intelligence and Big Data, 1(1), 65–74. Retrieved from https://www.scipublications.com/journal/index.php/jaibd/article/view/1109

[22] Patel, N., & Wang, Z. *Instant credit card issuance and its impact on digital financial services adoption*. International Journal of Digital Banking, 10(1), 55-72. https://doi.org/10.5678/ijdg.0154

[23] Manikanth Sarisa, Venkata Nagesh Boddapati, Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla, Shravan Kumar Rajaram.Navigating the Complexities of Cyber Threats, Sentiment, and Health with AI/ML. (2020). JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE), 8(2), 22-40. https://doi.org/10.70589/JRTCSE.2020.2.3

[24] Harper, C., & Chen, L.. *Transforming customer experience with personalized financial offerings in digital banking*. Journal of FinTech Strategies, 9(4), 48-60. https://doi.org/10.1190/jfs.0735

[25] Gollangi, H. K., Bauskar, S. R., Madhavaram, C. R., Galla, E. P., Sunkara, J. R., & Reddy, M. S. (2020). Unveiling the Hidden Patterns: AI-Driven Innovations in Image Processing and Acoustic Signal Detection. (2020). JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE), 8(1), 25-45. https://doi.org/10.70589/JRTCSE.2020.1.3.

[26] Lekkala, S. (2021). Ensuring Data Compliance: The role of AI and ML in securing Enterprise Networks. In Educational Administration: Theory and Practice. Green Publication. https://doi.org/10.53555/kuey.v27i4.8102