How Different are the Cloud Workloads? Characterizing Large-Scale Private and Public Cloud Workloads

Xiaoting Qin*, Minghua Ma*, Yuheng Zhao*, Jue Zhang*, Chao Du*, Yudong Liu*, Anjaly Parayil*, Chetan Bansal*, Saravan Rajmohan*, Íñigo Goiri*, Eli Cortez*, Si Qin*, Qingwei Lin*, and Dongmei Zhang*
*Microsoft
{xiaotingqin, minghuama, yuhengzhao, jue.zhang, chaodu, yudongliu, aparayil, chetanb, saravan.rajmohan, inigog, eli.cortez, si.qin, qlin, dongmeiz}@microsoft.com

***Abstract*—With the rapid development of cloud systems, an increasing number of service workloads are deployed in the private cloud and/or public cloud. Although large cloud providers such as Azure and Google have published workload traces in the past, prior work has not focused on analyzing and characterizing the differences between private and public cloud workloads in detail. Based on our experience working with Azure, one of the most widely used cloud platforms in the world, we find that the workload characteristics are different between the private and public cloud workloads. Specifically, compared with the public cloud workloads, the private cloud workloads tend to be more homogeneous in both deployment sizes and utilization patterns, more static with occasional bursts in deployment characteristics, and more region-agnostic regarding the sensitivity to deployed regions. Our findings gain several insights and implications on cloud management and motivate us to build a centralized workload knowledge base.**

***Index Terms*—Cloud workloads, workload characteristics, resource management**

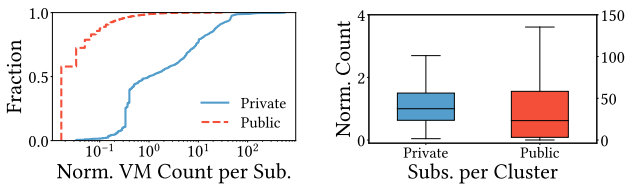
I. Introduction

Cloud computing has become increasingly popular over the years with a booming number of businesses migrating their workloads to big public cloud platforms, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform [1], [2]. Depending on business needs, customers may place their workloads on the *private cloud* hosting their own services, or they may choose the *public cloud* to share resources with other customers. According to the cloud computing trend report [3], 89% of customers adopt a multi-cloud strategy with 80% of users taking a hybrid approach for placing their workloads on both private and public clouds. This also applies to the above large cloud providers, who also operate their own services, such as Amazon.com in AWS and Microsoft 365 in Azure, as the *first-party workloads* on both the public and private clouds, in addition to providing cloud services to external customers for hosting the *third-party workloads* on the public cloud.

As the main target customers are different for the private and public cloud platforms, one may wonder whether the characteristics of hosted workloads are different as well. It is known that better understanding on workload characteristics can be helpful for making optimal decisions on cloud management, such as resource allocations, batch jobs scheduling, disaster recovery [4], and Virtual Machine (VM) migration [5], [6]. For instance, to avoid service interruption, the cloud platform could choose to migrate out VMs from nodes with unhealthy signals that may indicate hard disk failure [7]. With knowledge of the lifetime of VMs running on this node [8], the cloud platform can optimize this procedure by only migrating out VMs with long remaining time.

While there exist several studies providing extensive information on the general characteristics of workloads in cloud platforms (see Section VI for more details), those works tend to focus on workloads belonging to a single platform, either private or public cloud [5], [9]. The closest study to our work is the characterization of first-party and third-party workloads running on the public cloud Azure [8]. Although the first- and third-party workloads are distinguished, that work focuses more on their similarity rather than differences. Moreover, as first-party services can selectively place workloads between the private and public clouds, the obtained characteristics of first-party workloads can be different between these two cloud platforms. Therefore, correct characterization of workloads of each cloud platform is needed, and it is likely that workloads belonging to different groups display distinctive characteristics and need to be managed by cloud platforms under different rules. Applying uniform management polices based on the general characteristics of workloads could fail to deliver expected results or yield suboptimal decisions on each individual cloud platform.

In this study, we aim to explore the characteristics of workloads presented in the private and public clouds so that better design and management solutions can be obtained for each of them. In consideration for various optimization aspects, we study the workload characteristics of VM deployment and resource utilization, dig out insights and explore the implications brought for different opportunities on resource management in different cloud platforms. Our investigations are performed over the data involving millions of VMs collected over two large-scale private and public cloud platforms belonging to Azure. By distinguishing the private cloud workloads and public cloud workloads, our study firstly reveals the key



(a) VM Count Per Subscription (b) Sub. Count Per Cluster

Fig. 1. (a) CDFs of the normalized number of VMs per subscription for private and public cloud workloads at a particular time. (b) Box-plots of the normalized number of subscriptions per cluster. The labels on the left (right) y-axis correspond to the private (public) cloud. The boundaries of the whiskers are based on the 1.5 interquartile range.

differences in characteristics between them. Specifically, it shows that the deployment and utilization patterns of private cloud workloads are more homogeneous compared to those of public cloud workloads. Moreover, the temporal deployment patterns of private cloud workloads tend to be more static most of the time with occasional bursts. In contrast, more prominent diurnal deployment patterns are often found among public cloud workloads. Finally, private cloud workloads are more likely to be region-agnostic, allowing them to be deployed in any region without compromising user experience. These findings can help us to devise management strategies more customized for each individual cloud platform.

We have made the following contributions:

- We present a large-scale study to identify the differences between private and public cloud workloads in terms of deployment characteristics and resource utilization, and highlight the dependence of efficient and robust cloud operations on such workload characteristics.
- We point out that such differences can lead to different management strategies and optimization opportunities in different private or public cloud.
- We confirm the importance of leveraging workload characteristics through real practice and motivating us to build a centralized workload knowledge base in the future.

II. Background and Dataset

Azure. Azure is one of the most widely used cloud providers in the world, which contains both private and public cloud platforms. In our analysis, the private and public cloud workloads are deployed in separate *clusters*, which contain thousands of nodes with identical Stock Keeping Unit (SKU) configurations. These clusters are hosted in *datacenters* located over different *regions* (geo-locations).

Terminology. Each user, either the internal user or the external user, can create one or more Azure *subscriptions*. A subscription deploys VMs into a *region* (one or more datacenters) selected by the users. Allocation services in Azure will then place requested VMs into physical *nodes* (servers) [10]. Nodes are often stacked in *racks*, which may be served as fault domains in the cloud platforms.

Dataset. Results presented in this article are based on a representative dataset on the activities of workloads on Azure over an ordinary one-week period without any holiday[1]. To ensure that our findings are statistically meaningful and consistent across time, we obtain millions of VMs owned by tens of thousands of subscriptions. Specifically, we collect data of all clusters in the private cloud and sample a similar number of clusters in public cloud at random to make them comparable. The resulting number of VMs in public cloud is similar to the total number of VMs in private cloud. The dataset includes the detailed information of VMs, (*e.g.*, subscription, VM size, *etc.*) and the average resource utilization of VMs (reported every 5 minutes). The private cloud workloads are first-party workloads, i.e., Microsoft workloads, which in terms of the functionality are dominated by web application services, data analytic services, and real time communication services. The public cloud workloads consist of first-party workloads as well as third-party workloads i.e., customer workloads, thus are more opaque to the cloud platform and diverse by intuition. Both private and public cloud workloads have Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) VMs.

III. Deployment Characteristics

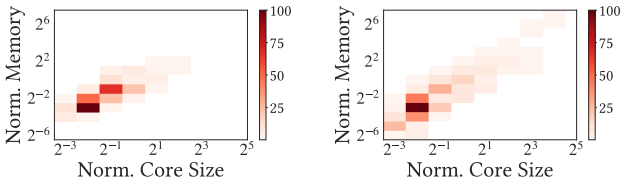
In this section, we discuss the deployment characteristics of private and public cloud workloads from several aspects. We first summarize the basic properties of VM deployment including the deployment size and VM size (Section III-A). We then dive into the detailed deployment behavior over time by studying the VM number variation at different levels (Section III-B). Finally, we study the deployment characteristics in the spatial domain (Section III-C).

A. Basic Properties

VM deployment size. We start by analyzing the deployed number of VMs at the subscription level. Figure 1(a) presents the Cumulative Distribution Functions (CDFs) of the normalized number of VMs per subscription for the workloads in private (blue solid curve) and public (red dashed curve) cloud at one time point on a weekday. Similar results (not shown) are also observed at other time points in the studied week. It shows that private cloud workloads are deployed in larger groups than public cloud workloads. As clusters in private and public cloud have similar sizes, the difference in deployment sizes suggests that clusters in the public cloud would host more subscriptions compared to clusters in the private cloud. This is confirmed by the observation shown in Figure 1(b), which indicates that a public cloud cluster hosts about 20 times more subscriptions than a private cloud cluster at the median level.

VM Size. We then explore the distributions of the number of CPU cores and the amount of memory used by each VM. Figure 2 shows the heatmaps for the normalized number of

[1]Due to the confidential policies of Azure, we omit certain exact numbers and times of the dataset, instead, we provide more relevant workload statistics and trends through normalization. Normalization units refer to quantities in the private cloud with specific choices depending on the contexts of analysis.



(a) Private Cloud (b) Public Cloud

Fig. 2. Heatmaps of normalized core and memory sizes per VM for private (left) and public (right) cloud workloads.

CPU cores and normalized amount of memory per VM for private (left) and public (right) cloud workloads. While the distributions of VMs' core and memory sizes are largely similar between the private and public cloud workloads, the distribution of the public cloud workloads extends to both the top right and bottom left corners of Figure 2(b), which suggests a non-negligible demand for relatively large and small VMs (both in the number of cores and the amount of memory) among the public cloud customers.

> **Insight 1:** Tendency to create larger deployment sizes is observed for the private cloud workloads, and workloads in the public cloud clusters are more diverse in terms of the number of subscriptions and the range of VM sizes.

Implication. The presence of high-level workload homogeneity in the private cloud clusters poses several challenges for the management of cluster efficiency and capacity. Firstly, the large deployment size makes private cloud workloads more prone to allocation failures, especially when clusters are reaching capacity limits. Secondly, for better fault tolerance, cloud platforms often spread workloads of the same service over multiple fault domains (e.g., racks). As a result, in private cloud clusters with less diverse workloads, it is often harder to place additional workloads on the same fault domain due to the higher chances of encountering workloads belonging to the same service. Considering the additional observation that the private cloud workloads are less diverse in terms of VM sizes, the efficiency of private cloud clusters can be quite sensitive to the placements and management of workloads, and more sophisticated and holistic approaches based on workload profiles, which are more accessible in the private cloud, would be needed.

B. VM Deployment in Temporal Domain

VM lifetime. As VMs are created and removed in both high quantity and frequency in the cloud platform, the deployment characteristics in the temporal domain can offer valuable information for improving efficiency. Figure 3(a) displays the CDFs of normalized VM lifetimes for private and public cloud workloads in a week. The VM lifetime is defined as the time between the creation and termination of a VM. Note that we only include the VMs started and ended in the week to be consistent with the time span of the dataset. It shows that 49% of private cloud VMs fall in the shortest lifetime bin, as compared to 81% of public cloud VMs in the same bin. The trend continues over the whole range of the x-axis. This observation suggests that the public cloud customers deploy more short-lived VMs (in percentage) compared to the private cloud customers.

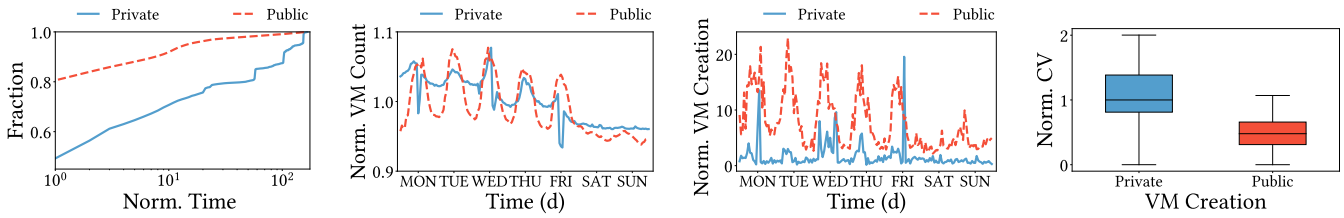
VM number. We then dive into the detailed VM deployment behavior by studying the change of VM counts over time. Figure 3(b) shows the normalized VM counts per hour in one sampled region over the selected week. For both private and public cloud workloads, the temporal changes of VM count largely follow a diurnal pattern during weekdays and exhibit a significant decrease over weekends. Moreover, compared to the pattern of VM counts for the public cloud workloads, the pattern of VM counts of the private cloud workloads tends to be less regular with occasionally large spikes. These spikes are not due to data quality issues but are mainly caused by the deployment behavior of some large services.

We further study the temporal variation in the numbers of VMs created per hour, as shown in Figure 3(c). For the public cloud workloads, the number of VMs created per hour follows a clear and stable diurnal pattern. In contrast, for the private cloud workloads, while the number of VMs created per hour usually stays at a low amplitude with little variation, bursts in which a large number of new VMs are created occasionally are observed. This burst is consistent with the spikes observed in the VM number variation of private cloud workloads. VM removal behavior is also studied and the observed temporal pattern is similar to that of VM creation.

To further examine whether the above temporal patterns of deployment can be observed in other geographic regions, we first quantify the temporal variation of VM number creation per hour with the coefficient of variation (CV) variable. The CV is defined as the ratio of standard deviation to the mean, and in the current context, it is computed over the distribution of the VM number creation per hour over one week. Namely, for each curve in Figure 3(c), we compute its CV by aggregating over the time dimension and obtain a larger value for the private cloud case due to the presence of the bursts. After obtaining the values of CV for other regions, we display their region-level distributions for both private and public clouds as box-plots in Figure 3(d). It shows that the hourly VM creations of private cloud workloads tend to have larger values of CV compared to that of public cloud workloads, indicating that similar bursty temporal patterns also present in other regions.

> **Insight 2:** The temporal deployment patterns of private cloud workloads mainly consist of low-amplitude deployments with occasional bursts, while more prominent and regular diurnal deployment patterns are found in public cloud workloads.

Implication. For the public cloud workloads, the observed

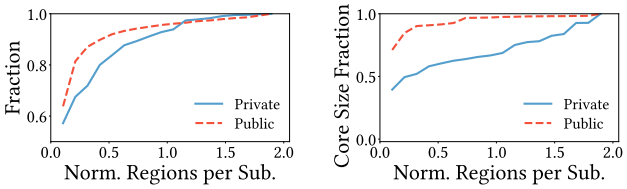


(a) CDF of VM lifetimes (b) VM count (c) VM creation (d) CV of VM creation

Fig. 3. (a) CDFs of normalized VM lifetimes in a week. (b) Variations of normalized VM counts per hour at one region. (c) Variations of normalized numbers of VMs created per hour at one region. (d) Box-plots of CVs (over the temporal distributions of VM number creation per hour) across regions.

diurnal deployment patterns are mostly due to the auto-scaling features provided by the cloud platform that automatically adjust the number of VMs based on business needs. As the platform resources are under-utilized during the valley of such a diurnal pattern, for short-lived VMs hosting public cloud workloads, one may consider adopting the spot VMs [11], [12], [13], [14] to reduce cost and improve platform resource utilization, especially during valley hours. The previous observation that 81% of public cloud VMs fall into the shortest lifetime bin shows the considerable number of candidate VMs for this adoption. This also has motivated us to develop more advanced technology (e.g., spot VM eviction rate prediction [15] and dynamic mixture of spot and on-demand VMs [16]) for better support spot VM adoption.

For the private cloud workloads, however, due to the less regular deployment patterns and the large deployment size observed above, opting for auto-scaling features for the private cloud workloads can often create large variations in the counts of deployed VMs, which may increase the chances of allocation failures especially when the capacity is running low. Moreover, the irregular deployment patterns of private cloud workloads also do not match well with their actual resource utilization patterns, which are mostly diurnal as will be discussed in Section IV. Correspondingly, a better workload-aware allocation failure prediction method and dynamic resource over-subscription system [6], [17] can be critical for improving the efficiency of capacity management for the private cloud workloads. For example, for dynamic resource over-subscription system, by leveraging the varying usage patterns of VMs, resources can be assigned at a lower level than peak usage without affecting performance. As such, over-subscription assigns fewer resources to each VM than requested, but allows VMs to use more resources if the physical machine has spare capacity. Note that there is a small chance that all VMs may reach their peak usages simultaneously. It is therefore necessary to consider the risk when coordinating resource utilization. This problem can be addressed through chance-constrained optimization framework, which has been shown to improve utilization by 20% to 86% [17] in Azure compared to baseline methods, depending on the level of safety constraint.



(a) Regions per subscription (b) Core size fraction

Fig. 4. (a) CDFs of the normalized deployed regions per subscription. (b) CDFs of the normalized deployed regions per subscription in terms of its allocated core size.

C. VM Deployment in Spatial Domain

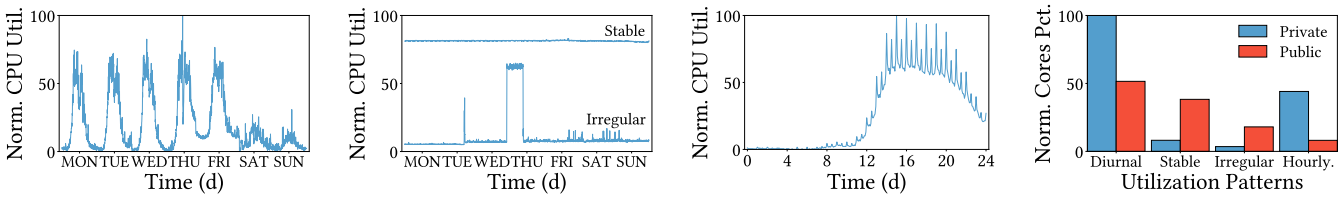
Finally, we study the deployment characteristics in the spatial domain, which has been briefly discussed in Section III-A through the study on the number of subscriptions at the cluster level (see Figure 1(b)). Here we extend the analysis to the region level. Figure 4(a) presents the CDFs of the normalized numbers of deployed regions per subscription for private and public cloud workloads. It shows that although more than 50% of subscriptions for both types of workloads make deployment in a single region, private cloud workloads tend to deploy over more regions in the rest of the subscriptions. In terms of the subscriptions' actual resource usage, as shown in Figure 4(b), the subscriptions deployed in a single region make up 40% of core usage for the private cloud workloads, so the majority of the cores are used by the subscriptions with multiple-region deployment in the private cloud. On the contrary, 70% of cores in the public cloud are used by single-region subscriptions. We will discuss the implication of the cross-region feature of the private cloud workloads in Section IV-B.

IV. Resource Utilization

In this section, we first classify the CPU utilization patterns into four typical types and analyze the CPU utilization distributions across time, comparing public and private cloud workloads (Section IV-A). Then, we discuss the spatial distribution of resource utilization (Section IV-B).

A. Temporal Utilization Pattern

According to our observation, the VM CPU utilization patterns can be classified into four types, *i.e.*, diurnal, stable, irregular, and hourly-peak. We adopt this categorization



(a) Diurnal (b) Stable and Irregular (c) Hourly-Peak (d) Percentage of each category

Fig. 5. Typical utilization patterns of workload and their distribution. (a)-(c) Sample of typical utilization patterns. (d) Normalized percentage of each typical utilization pattern in private and public cloud respectively at a particular time.

to leverage these patterns for optimization, for example, as diurnal workloads introduce peak and valley hour of resource usage, it calls for specific optimization strategy for resource management. Figure 5(a-c) illustrate four typical temporal patterns of CPU utilization for some sampled VMs. The first three patterns, the diurnal, stable, and irregular patterns are all displayed over a one-week period, while the hourly-peak pattern is shown over a one-day period. Figure 5(d) illustrates the relative proportions of these patterns in private and public clouds.

Diurnal. As shown in Figure 5(a), CPU utilization exhibits a daily periodic pattern (high during daytime and low during nights) in CPU utilization, which is related to user activity and can be detected using the approach discussed in [18]. This pattern also shows a clear difference between weekdays and weekends. Specifically, the peak value of CPU utilization is around 60% during weekdays but reaches only about 20% during weekends. VMs with such a utilization pattern would present great challenges for the cloud platform to balance resource utilization between daytime and night and between weekdays and weekends. As shown in Figure 5(d), the diurnal patterns are the most common patterns in both private and public cloud workloads. Private cloud workloads have roughly double the diurnal patterns as of public cloud workloads, implying that private cloud workloads are more user-facing.

Stable. The stable pattern is extracted by restricting the standard deviation of CPU utilization, as illustrated in top of Figure 5(b). Such type of VMs provides an opportunity for harvesting idles CPU resources with options such as over-subscription. The share of workloads with the stable pattern in public cloud is higher than that in private cloud, suggesting that the public cloud workloads are more stable and suitable to adopt over-subscription.

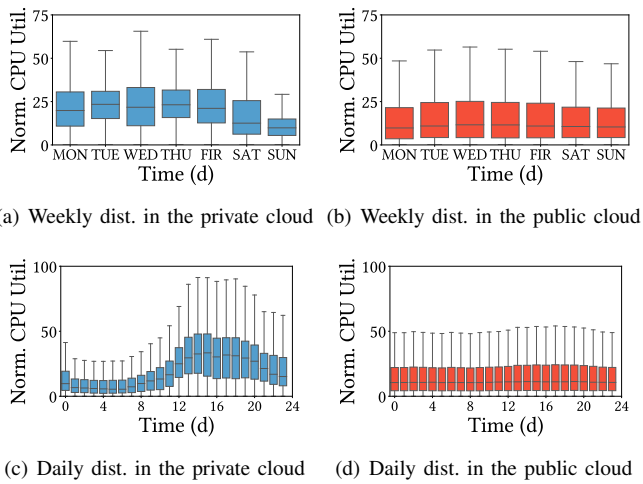
Irregular. Apart from diurnal and stable patterns, the remaining pattern is the irregular one, as shown in the bottom of Figure 5(b), while the CPU utilization is lower than 10% most of the time, it can raise to over 60% for a short time with no apparent sign. Due to this irregular resource usage behavior, it can be difficult and risky to impose aggressive resource management strategies on such type of VMs. In both private and public cloud workloads, this type of utilization pattern is relatively rare, as shown in Figure 5(d).

Hourly-Peak. Hour-peak is a special diurnal pattern, which is also extracted using period detection approach [18] (period equal to one hour). In the pattern shown in Figure 5(c), regular peaks at the beginning of the hour/half-hour can be observed. This pattern often appears in private cloud workloads supporting various work-related activities, as shown in Figure 5(d). For instance, as online meetings are often scheduled to start at the hour or half-hour marks, high CPU utilization peaks in supporting VMs can appear when users are joining the scheduled meetings.

Since the diurnal pattern is the dominant one in both public cloud and private cloud, we narrow down to characterize the CPU utilization distributions across time, with the comparisons between private and public cloud workloads. To explore the temporal distribution of resource utilization, we analyze CPU utilization of private and public cloud workloads over one week, as shown in Figure 6(a) and Figure 6(b). According to the 75-percentile, we observe that CPU utilization for both private and public clouds is lower than 30%. In addition, considering the values of different percentiles, CPU utilization of public cloud is more stable than private cloud. This may be also related to that work-related activities account for more proportions in the private cloud workloads, and thus the CPU utilization relatively drops during the weekends for private cloud. We further characterize the daily CPU utilization distribution for private and public cloud workloads, as shown in Figure 6(c) and Figure 6(d). We observe that CPU utilization for private cloud changes across the whole day, following a roughly working-hour pattern at the hourly granularity, while for public cloud is almost constant. This observation shares the same insight as from Figure 5(d).

> **Insight 3:** As utilization patterns can vary significantly among workloads, correct characterization of them can be critical for determining the right management strategy that fits with the resource usage of workloads.

Implication. Hour-peak is a unique pattern which brings different opportunities in resource management and calls for appropriate management strategies in private cloud, such as predictive resource pre-provisioning [19] and leveraging over-clocking techniques to absorb utilization peaks [20]. Workloads with hourly-peaks are more common in the private cloud while relatively rarer in the public cloud, as work-



(a) Weekly dist. in the private cloud (b) Weekly dist. in the public cloud

(c) Daily dist. in the private cloud (d) Daily dist. in the public cloud

Fig. 6. CPU utilization distribution of the private and public cloud workloads, over a week (a, b) and within a day (c, d).

related activities (e.g., Microsoft 365 services) account for more proportions in the private cloud workloads.

As the private cloud is dominated by diurnal workloads, more workloads of other utilization patterns need to be imported to reduce under-utilized resource during the valley hour. For example, identifying deferrable workloads and schedule them to the valley hour would be a feasible way to leverage the observed utilization pattern in private cloud for resource management optimization.

B. Spatial Distribution of Resource Utilization

In this spatial distribution analysis of resource utilization, we aim to answer two questions: a) whether workloads co-located on the same node have similar utilization patterns; b) how utilization patterns differ among deployed regions for workloads under the same subscription. The first question is to address how resources are shared by leveraging statistical multiplexing in cloud platforms, and the second one would help us to identify the so-called region-agnostic workloads, which are not sensitive to deployed regions and can play an important role in the region-level resource management.

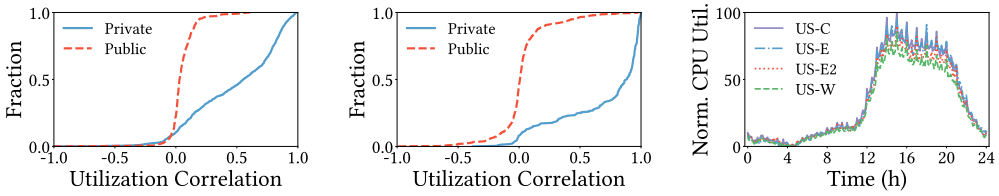
Workloads similarity at the node level. To characterize the similarity of workloads in each node, we calculate the Pearson correlation [21] of CPU utilization between VMs and their host node. As the node CPU utilization mostly originates from the usage of VMs, its higher correlation with hosted VMs indicates that the utilization patterns among hosted VMs are similar. Please note that we filter out the trivial case that nodes only host one VM, which is a small percentage of both private and public cloud nodes. Figure 7(a) shows the CDFs of the correlation for private and public cloud workloads. We observe that the median value of the correlation score of private cloud workloads (0.55) is higher than that of public cloud (0.02). It indicates that the utilization patterns of VMs in each node are more similar in the private cloud, suggesting that private cloud is more homogeneous in terms of both deployment and resource usage. This also agrees with the previous observation that the private cloud workloads are more homogeneous than the public cloud workloads.

Workloads similarity at the region level. According to Section III-C, there exist a large fraction of subscriptions that have deployments over multiple regions. It would be interesting to explore the similarity of workload utilization patterns across regions for those subscriptions. This is achieved by calculating the Pearson correlation of CPU utilization in each pair of deployed regions. To reduce the number of combinations, we restrict the deployed regions to those in the US, which has about 10 regions spreading over 9 time zones in our dataset (enough number of regions for this analysis). Figure 7(b) shows the CDFs of the correlation of workload utilization in each region pair. Note that the utilization pattern used in the correlation study is the averaged utilization computed at the region level for each studied subscription. From Figure 7(b) we observe a higher correlation of utilization across regions for private cloud workloads, indicating those subscriptions in the private cloud tend to have the same utilization pattern across different regions.

Moreover, those subscriptions with similar utilization patterns across regions may belong to the region-agnostic workloads, i.e., not sensitive to the deployed regions. For example, Figure 7(c) gives the average CPU utilization of *ServiceX* in different regions in one day. This service has prominent diurnal and hourly-peak patterns and thus it is likely to be user-facing. Although these regions are in separate time zones, the service CPU utilization patterns are roughly peaked at the same time points. This is contrary to the expectation for region-sensitive user-facing workloads, for which one would observe shifted peaks across different regions with different time zones. The workload owner of *ServiceX* also confirmed our conjecture on the region-agnostic nature of *ServiceX* by pointing out that a geo-level load-balancer is adopted for routing users' requests across multiple regions.

> **Insight 4:** Utilization patterns of workloads in each node are more similar in the private cloud than in the public cloud, showing a more homogeneous nature of private cloud workloads. There exist a large portion of subscriptions in the private cloud whose utilization patterns are similar across regions, suggesting that they are likely to be region-agnostic.

Implication. The similar utilization pattern at the node level for the private cloud indicates that workloads in the same node are likely to have utilization peaks at the same time. That would limit the room for oversubscribing resources to achieve better utilization. The above utilization pattern analysis alone is not sufficient to identify region-agnostic workloads as other factors such as data locality, compliance issues, geo-specific hardware constraints, and workload dependencies must also be considered. However, the above case study suggests that a



(a) VMs and the belonging node (b) Regions under a subscription (c) Utilization of *ServiceX*

Fig. 7. (a) CDFs of CPU utilization correlation between VMs and their host node. (b) CDFs of CPU utilization correlation of different regions under the same subscription. (c) Average CPU Utilization of *ServiceX* in different regions.

substantial number of region-agnostic workloads exist in the private cloud. Leveraging their region-agnostic characteristics, we can optimize their placement to achieve better capacity management and sustainability for the cloud in various ways.. For instance, region-agnostic workloads can be relocated from hot to cold regions (with regard to capacity usage) to balance the capacity usage globally, reduce underutilized clusters, and save cost. We may also shift more region-agnostic workloads to regions that are more accessible to renewable energy, which would benefit both the customer and cloud provider in achieving better sustainability.

To comprehend the actual impact of region-agnostic workloads, we piloted a few experiments in Azure. In one of the experiments, we focused on the Canadian regions, where one of the regions had a high percentage of underutilized cores. Using utilization data from these regions, we recommended shifting the workload of *Service-X* from *Canada-A* to *Canada-B*. As a result of this regional workload shift, the underutilized core percentage of *Canada-A* decreased from 23% to 16%, and the core utilization rate reduced from 42% to 37%, indicating an improvement in the overall health of the source region. *Canada-B*, which has sufficient idle capacities, showed minor changes in terms of core utilization and underutilized core percentage.

V. Discussion

Extensive characterization for private and public cloud workloads can bring considerable opportunities for improvements in the cloud resource management. Currently, there exists a gap between the characteristics of both public and private cloud workloads and using them to formulate appropriate management and optimization strategies for cloud platforms, especially for large scale private cloud. This work serves as a precursor for devising such strategies and shares the practices adopted by Azure to improve resource management. We hope our work will stimulate and inspire more studies from cloud providers and benefit the community.

Although the high-level mechanisms of these optimization methods are often generalizable over different cloud platforms, they highly rely on the critical input of workload knowledge to maximize their performances, which require a comprehensive study on the differentiation of the workload characteristics in different cloud platforms. Therefore, in order to obtain a scalable solution that can easily extend to different cloud platforms, one first needs to abstract out the common optimization policies and then build a centralized workload knowledge base, which continuously extracts workload knowledge from telemetry signals (e.g., CPU utilization, VM lifetime) and feeds them into the aforementioned optimization policies. A workload knowledge base will then be the key pillar of the future workload-aware intelligent cloud platform, and it allows the cloud provider to maximally optimize the platform's performance by tailoring to its hosted workloads.

VI. Related Work

There exist a plethora of works aiming for characterizing workloads in cloud platforms. Related works in this area often study workload characteristics related to resource utilization and workload deployment such as lifetime or task duration [9], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31] and explore the heterogeneity in those characteristics [32], [33], [34], [35]. By providing insights on the heterogeneity and disparity in the characteristics of workload deployment and resource utilization, those works substantiate the need for new cloud resource schedulers and provide critical design insights.

In addition to the aforementioned works, existing studies also explore other workload characteristics including failure distribution, correlation, arrival rate, and interference between resources [36], [37], [38], volatility of resource demand and usage by region and customer segments [39], [40], microservice dependencies [41], [42]. For example, authors of [38] study resource virtualization in data centers and shed light on optimizing configuring VMs' virtual resources, and [40] focus on the time evolution of workload demands and resource availability, providing a solid basis for capacity planning of the cloud considering seasonality shift.

While most existing studies in this area focusing on workload traces from public clouds, some researchers also explore the characterization of specific workload traces not originated from public clouds, such as deep learning workloads from GPU data centers [43], grid systems [44], and production Hadoop cluster [45], [46]. There is limited work on characterizing the private cloud. The most closed one to our work is [47], however it focus only on one private enterprise cloud without comparing to public cloud.

Different from the above works, we characterize the differences between workloads in two large-scale private and public cloud platforms belonging to the same cloud provider

comprehensively, gain several insights and implications on cloud management, and emphasize on the goal of maximally optimizing the platform's performance by tailoring to its hosted workloads. For instance, we categorize the utilization patterns and call for tailored optimization strategy. The node level study and insights on region-agnostic workload are also an unexplored area in current cloud management.

VII. Threats to Validity

In this section, we address potential limitations and threats to the validity of our study. We acknowledge that our dataset has certain limitations, which may impact the generalizability of our findings.

Our dataset focuses on private and public cloud provided by Azure. Therefore, the findings that are highly dependent on the workload types may not be directly applicable to other cloud providers. The dataset covers a time period of one week, specifically chosen without any holiday to minimize the impact of external factors and extract common temporal patterns. Consequently, our results may not fully capture the effects of seasonality and holiday patterns.

VIII. Conclusion

Characterizing cloud workloads is important to create more optimal resource provisioning and allocation techniques. In this paper, we examine millions of VMs from a large-scale real-world cloud provider, segregate private and public cloud workloads and provide key findings to improve cloud resource management.

With the study of characterising VM deployment and resource utilization, we found that the private cloud workloads tend to be more homogeneous in both deployment sizes and utilization patterns, more static with occasional bursts in deployment characteristics, and more region-agnostic regarding the sensitivity to deployed regions, which would motivate the cloud provider to adopt for proactive resource provisioning and global workload balancing. The implications from studying the public cloud workloads suggest different optimization opportunities such as spot VMs. Our insights will be useful for improving the reliability, efficiency and sustainability of cloud platforms.

In order to tailor management algorithms and systems by cloud platforms for the optimal operation, a detailed understanding of the unique characteristics of workloads belonging to different groups is critically needed. In our future work, we will design and implement a centralized workload knowledge base in order to better exploit workload insights for cloud management.

References

[1] Kenton McDonough, Xing Gao, Shuai Wang, and Haining Wang. Torpedo: A fuzzing framework for discovering adversarial container workloads. In *Proceedings of the 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 402–414. IEEE, 2022.

[2] Ningfang Mi, Giuliano Casale, and Evgenia Smirni. Scheduling for performance and availability in systems with temporal dependent workloads. In *Proceedings of the IEEE International Conference on Dependable Systems and Networks With FTCS and DCC (DSN)*, pages 336–345. IEEE, 2008.

[3] Brian Adler. Cloud computing trends: Flexera 2022 state of the cloud report. https://www.flexera.com/blog/cloud/cloud-computing-trends-2022-state-of-the-cloud-report/, 2022.

[4] Long Wang, Harigovind V Ramasamy, Richard E Harper, Mahesh Viswanathan, and Edmond Plattier. Experiences with building disaster recovery for enterprise-class clouds. In *Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 231–238. IEEE, 2015.

[5] Abhishek Verma, Luis Pedrosa, Madhukar Korupolu, David Oppenheimer, Eric Tune, and John Wilkes. Large-scale cluster management at google with borg. In *Proceedings of the Tenth European Conference on Computer Systems (EuroSys)*, pages 1–17. IEEE, 2015.

[6] Maria Carla Calzarossa, Luisa Massari, and Daniele Tessera. Workload characterization: A survey revisited. *ACM Computing Surveys (CSUR)*, 48(3):1–43, 2016.

[7] Tianyi Yang, Jiacheng Shen, Yuxin Su, Xiaoxue Ren, Yongqiang Yang, and Michael R Lyu. Characterizing and mitigating anti-patterns of alerts in industrial cloud systems. In *Proceedings of the 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 393–401. IEEE, 2022.

[8] Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In *Proceedings of the Symposium on Operating Systems Principles (SOSP)*, pages 153–167. ACM, 2017.

[9] Wenyan Chen, Kejiang Ye, Yang Wang, Guoyao Xu, and Cheng-Zhong Xu. How does the workload look like in production cloud? analysis and clustering of workloads on alibaba cluster trace. In *Proceedings of the International Conference on Parallel and Distributed Systems (ICPADS)*, pages 102–109. IEEE, 2018.

[10] Ori Hadary, Luke Marshall, Ishai Menache, Abhisek Pan, Esaias E Greeff, David Dion, Star Dorminey, Shailesh Joshi, Yang Chen, Mark Russinovich, et al. Protean: Vm allocation service at scale. In *Proceedings of the Symposium on Operating Systems Design and Implementation (OSDI)*, pages 845–861. USENIX, 2020.

[11] Microsoft Azure. Azure spot virtual machines. https://azure.microsoft.com/en-us/services/virtual-machines/spot/#overview. Accessed: 2022-06-13.

[12] Alibaba Cloud. Alibaba preemptible instances. https://www.alibabacloud.com/help/doc-detail/52088.htm. Accessed: 2022-06-13.

[13] Google. Google spot vms. https://cloud.google.com/compute/docs/instances/spot. Accessed: 2022-06-13.

[14] Amazon. Amazon ec2 spot instances. https://www.amazonaws.cn/en/ec2/spot-instances. Accessed: 2022-06-13.

[15] Fangkai Yang, Bowen Pang, Jue Zhang, Bo Qiao, Lu Wang, Camille Couturier, Chetan Bansal, Soumya Ram, Si Qin, Zhen Ma, Íñigo Goiri, Eli Cortez, Senthil Baladhandayutham, Victor Rühle, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. Spot virtual machine eviction prediction in microsoft cloud. In *Companion Proceedings of the Web Conference (WWW)*, pages 152–156. ACM, 2022.

[16] Fangkai Yang, Lu Wang, Zhenyu Xu, ue Zhang, Liqun Li, Bo Qiao, Camille Couturier, Chetan Bansal, Soumya Ram, Si Qin, Zhen Ma, Iñigo Goiri, Eli Cortez, Terry Yang, Victor Rühle, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. Snape: reliable and low-cost computing with mixture of spot and on-demand vms. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 631–643. ACM, 2023.

[17] Junjie Sheng, Lu Wang, Fangkai Yang, Bo Qiao, Hang Dong, Xiangfeng Wang, Bo Jin, Jun Wang, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. Learning cooperative oversubscription for cloud by chance-constrained multi-agent reinforcement learning. https://arxiv.org/abs/2211.11759, 2022.

[18] Michail Vlachos, Philip Yu, and Vittorio Castelli. On periodicity detection and structural periodic similarity. In *Proceedings of the International Conference on Data Mining (ICDM)*, pages 449–460. IEEE, 2005.

[19] Chuan Luo, Bo Qiao, Xin Chen, Pu Zhao, Randolph Yao, Hongyu Zhang, Wei Wu, Andrew Zhou, and Qingwei Lin. Intelligent virtual machine provisioning in cloud computing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1495–1502. International Joint Conferences on Artificial Intelligence Organization, 2020.

[20] Majid Jalili, Ioannis Manousakis, Íñigo Goiri, Pulkit A. Misra, Ashish Raniwala, Husam Alissa, Bharath Ramakrishnan, Phillip Tuma, Christian Belady, Marcus Fontoura, and Ricardo Bianchini. Cost-efficient overclocking in immersion-cooled datacenters. In *Proceedings of the ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pages 623–636, 2021.

[21] Wikipedia. Pearson correlation coefficient. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient. Accessed 2022-06-17.

[22] Asit K Mishra, Joseph L Hellerstein, Walfredo Cirne, and Chita R Das. Towards characterizing cloud backend workloads: insights from google compute clusters. *ACM SIGMETRICS Performance Evaluation Review*, 37(4):34–41, 2010.

[23] Qi Zhang, Joseph Hellerstein, and Raouf Boutaba. Characterizing task usage shapes in google compute clusters. In *Proceedings of the 5th International Workshop on Large Scale Distributed Systems and Middleware (LADIS)*, 2011.

[24] Yanpei Chen, Archana Sulochana Ganapathi, Rean Griffith, and Randy H Katz. Analysis and lessons from a publicly available google cluster trace. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2010-95*, 94, 2010.

[25] Peter Garraghan, Paul Townend, and Jie Xu. An analysis of the server characteristics and resource utilization in google cloud. In *Proceedings of the International Conference on Cloud Engineering (IC2E)*, pages 124–131. IEEE, 2013.

[26] Charles Reiss, Alexey Tumanov, Gregory R Ganger, Randy H Katz, and Michael A Kozuch. Towards understanding heterogeneous clouds at scale: Google trace analysis. *Intel Science and Technology Center for Cloud Computing, Tech. Rep*, 84:1–21, 2012.

[27] Zitao Liu and Sangyeun Cho. Characterizing machines and workloads on a google cluster. In *Proceedings of International Conference on Parallel Processing Workshops (ICPP Workshops)*, pages 397–403. IEEE, 2012.

[28] Sheng Di, Derrick Kondo, and Franck Cappello. Characterizing cloud applications on a google data center. In *Proceedings of the International Conference on Parallel Processing (ICPP)*, pages 468–473. IEEE, 2013.

[29] Yue Cheng, Zheng Chai, and Ali Anwar. Characterizing co-located datacenter workloads: An alibaba case study. In *Proceedings of the Asia-Pacific Workshop on Systems*, pages 1–3, 2018.

[30] Congfeng Jiang, Guangjie Han, Jiangbin Lin, Gangyong Jia, Weisong Shi, and Jian Wan. Characteristics of co-allocated online services and batch jobs in internet data centers: a case study from alibaba cloud. *IEEE Access*, 7:22495–22508, 2019.

[31] Ji Xue, Feng Yan, Robert Birke, Lydia Y. Chen, Thomas Scherer, and Evgenia Smirni. Practise: Robust prediction of data center time series. In *11th International Conference on Network and Service Management (CNSM)*, pages 126–134. IEEE, 2015.

[32] Charles Reiss, Alexey Tumanov, Gregory R Ganger, Randy H Katz, and Michael A Kozuch. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In *Proceedings of the symposium on cloud computing (SoCC)*, pages 1–13. ACM, 2012.

[33] Bingwei Liu, Yinan Lin, and Yu Chen. Quantitative workload analysis and prediction using google cluster traces. In *Proceedings of the Conference on Computer Communications Workshops (INFOCOM Workshops)*, pages 935–940. IEEE, 2016.

[34] Md Rasheduzzaman, Md Amirul Islam, Tasvirul Islam, Tahmid Hossain, and Rashedur M Rahman. Task shape classification and workload characterization of google cluster trace. In *Proceedings of the International Advance Computing Conference (IACC)*, pages 893–898. IEEE, 2014.

[35] Chengzhi Lu, Kejiang Ye, Guoyao Xu, Cheng-Zhong Xu, and Tongxin Bai. Imbalance in the cloud: An analysis on alibaba cluster trace. In *Proceedings of the International Conference on Big Data (Big Data)*, pages 2884–2892. IEEE, 2017.

[36] Jing Guo, Zihao Chang, Sa Wang, Haiyang Ding, Yihui Feng, Liang Mao, and Yungang Bao. Who limits the resource efficiency of my datacenter: An analysis of alibaba datacenter traces. In *Proceedings of the International Symposium on Quality of Service (IWQoS)*, pages 1–10. IEEE, 2019.

[37] Muhammad Tirmazi, Adam Barker, Nan Deng, Md E Haque, Zhijing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. Borg: the next generation. In *Proceedings of the European conference on computer systems (EuroSys)*, pages 1–14, 2020.

[38] Robert Birke, Andrej Podzimek, Lydia Y. Chen, and Evgenia Smirni. State-of-the-practice in data center virtualization: Toward a better understanding of vm usage. In *43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 1–12. IEEE, 2013.

[39] Huangshi Tian, Yunchuan Zheng, and Wei Wang. Characterizing and synthesizing task dependencies of data-parallel jobs in alibaba cloud. In *Proceedings of the Symposium on Cloud Computing (SoCC)*, pages 139–151. ACM, 2019.

[40] Robert Birke, Lydia Y. Chen, and Evgenia Smirni. Data centers in the cloud: A large scale performance study. In *Proceedings of the Symposium on Cloud Computing (SoCC)*, pages 336–343. ACM, 2012.

[41] Shutian Luo, Huanle Xu, Chengzhi Lu, Kejiang Ye, Guoyao Xu, Liping Zhang, Yu Ding, Jian He, and Chengzhong Xu. Characterizing microservice dependency and performance: Alibaba trace analysis. In *Proceedings of the Symposium on Cloud Computing (SoCC)*, pages 412–426. ACM, 2021.

[42] Cinar Kilcioglu, Justin M Rao, Aadharsh Kannan, and R Preston McAfee. Usage patterns and the economics of the public cloud. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 83–91. ACM, 2017.

[43] Qinghao Hu, Peng Sun, Shengen Yan, Yonggang Wen, and Tianwei Zhang. Characterization and prediction of deep learning workloads in large-scale gpu datacenters. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pages 1–15. IEEE, 2021.

[44] Siqi Shen, Vincent Van Beek, and Alexandru Iosup. Statistical characterization of business-critical workloads hosted in cloud datacenters. In *Processings of the International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pages 465–474. IEEE, 2015.

[45] Zujie Ren, Xianghua Xu, Jian Wan, Weisong Shi, and Min Zhou. Workload characterization on a production hadoop cluster: A case study on taobao. In *Proceedings of the International Symposium on Workload Characterization (IISWC)*, pages 3–13. IEEE, 2012.

[46] Zujie Ren, Jian Wan, Weisong Shi, Xianghua Xu, and Min Zhou. Workload analysis, implications, and optimization on a production hadoop cluster: A case study on taobao. *IEEE Transactions on Services Computing*, 7(2):307–321, 2013.

[47] Ignacio Cano, Aiyar, Srinivas, and Arvind Krishnamurthy. Characterizing private clouds: A large-scale empirical analysis of enterprise clusters. In *Proceedings of the Symposium on Cloud Computing (SoCC)*, pages 29–41. ACM, 2016.