Fast and QoS-Aware Heterogeneous Data Center Scheduling Using Locality Sensitive Hashing

Mohammad Shahedul Islam, Matt Gibson, Abdullah Muzahid
University of Texas at San Antonio
Email: (mohammad.islam, matthew.gibson, abdullah.muzahid)@utsa.edu

Abstract—As cloud becomes a cost effective computing platform, improving its utilization becomes a critical issue. Determining an incoming application’s sensitivity toward various resources is one of the major challenges to obtain higher utilization. To this end, previous research attempts to characterize an incoming application’s sensitivity toward interference on various resources (Source of Interference or SoI, for short) of a cloud system. Due to time constraints, the application’s sensitivity is profiled in detail for only a small number of SoI, and the sensitivities for the remaining SoI are approximated by capitalizing on knowledge about some of the applications (i.e., training set) currently running in the system. A key drawback of previous approaches is that they have attempted to minimize the total error of the estimated sensitivities; however, various SoI do not behave the same as each other. For example, a 10% error in the estimate of SoI A may dramatically effect the QoS of an application whereas a 10% error in the estimate of SoI B may have a marginal effect. In this paper, we present a new method for workload characterization and scheduling that considers these important issues. First, we compute an acceptabletra for each SoI based on its effect on QoS, and our goal is to characterize an application so as to maximize the number of SoI that satisfy this acceptable error.

Then we present a new technique for workload characterization and scheduling based on Locality Sensitive Hashing (LSH). Given a set of $n$ points in a $d$-dimensional Euclidean space, LSH is a hashing technique such that points nearby are hashed to the same “bucket” and points that are far apart are hashed to different buckets. This data structure allows approximate nearest neighbor queries to be executed with nearly asymptotically optimal running time. This allows us to perform workload profiling quickly with high accuracy and scheduling in heterogeneous data centers with high quality of service (QoS) and utilization.

I. INTRODUCTION

Cloud has become a popular computing platform to provide flexible and cost effective services both to end users and data center operators. Public cloud providers like Amazon EC2, Microsoft Windows Azure, Google Compute Engine etc. host tens of thousands of applications each day. Cloud systems aim at providing scalable computing power at low cost. Cloud providers improve cost efficiency by using commodity servers and reducing power and cooling infrastructure cost. Similarly, they improve compute capabilities by building new data centers and relying on chip technology. These methods are approaching their point of diminishing return. Recent studies have shown that power delivery and cooling costs are less than 10% of the overhead [8]. At the same time, multithreaded scaling is coming to an end. Therefore, to improve cost-compute efficiency, it has become extremely important to increase utilization of a data center.

Increasing data center utilization requires scheduling applications together on the same server. However, applications can interfere with each other due to various shared resources. Such interference can lead to performance degradation [21]. The situation gets worsened by continuous load fluctuation, application diversity, heterogeneity of the servers [19], [8] etc. Therefore, cloud operators often disallow application co-location or use an over-provisioning of resources for high-priority tasks. Due to such scheduling approaches, data center utilization has been found to be notoriously low [19], [8]. A major step toward better utilization would be a scheduler that schedules applications by considering diversity in both applications and server configurations.

Most prior techniques [21], [23] rely on a detailed offline approach or a long term monitoring and modeling approach for characterizing recurring applications. As a result, they are not effective for large data centers that receive tens of thousands of potentially unknown and often non-recurring applications each day. Recently, there is some work [13], [24], [14] that takes a two-fold approach. First, it characterizes an application’s behavior towards various shared resources (referred to as Sources of Interference, or SoIs). Then, it schedules the application or adapts currently running applications accordingly. This is a promising direction for increasing a data center’s utilization. Inspired by this line of research, this paper also ventures into the same overall approach but aims to increase characterization accuracy and scheduling efficiency. It takes inspiration from an algorithm in the domain of computational geometry.

When an application arrives, we would like to characterize the application by measuring its sensitivity toward various SoIs. Sensitivity of an application toward a particular SoI is denoted by its Sensitivity Score. It is measured as a fraction of the total available resource (corresponding to the SoI) which is required, at least, to maintain 95% of the application’s standalone performance (i.e. QoS) in the best server [13]. Details of sensitivity score are discussed in Section II. Prior approaches treat each SoI equally; however, our experiments indicate that the same amount of error in estimating sensitivity scores for different SoIs can have significantly different effects on a scheduler’s ability to pick the right server. Hence, the impact on QoS of applications can vary too. This is shown in Figure 1. The graph is obtained by artificially injecting inaccuracy in different SoI’s sensitivity scores and measuring what fraction of applications achieve 95% or more QoS at the end. We used a state-of-the-art scheduler, Paragon [13], for this experiment. 20% error in estimating sensitivity toward Integer Processing Unit and Memory Capacity causes 9% and 11% fewer applications to achieve 95% or more QoS. On the contrary, the same amount of inaccuracy for Storage Capacity causes only 5% applications to lose the QoS threshold. Thus, the experiment suggests that we can tolerate different ranges of errors in estimating sensitivity of different SoIs. To exploit this insight, we present a novel method for workload characterization that...
considers varying error tolerance intervals for different SIs. For each sensitivity score, we determine the error interval that causes a 5% decrease in the number of applications achieving a QoS of 95% (or more), when compared to the case where the scheduler has perfect information about the sensitivity scores. As shown in Figure 1, the memoryCapacity sensitivity score can be underestimated by 20% but cannot be overestimated by more than 10%. We would like to obtain a sensitivity score that is within this error range. We call this range the Desired Error Interval (DEI). We compute the DEI for each sensitivity score. We attempt to characterize an incoming workload so that our predicted sensitivity scores fall within the DEI of as many sensitivity scores as possible. To this end, we propose a new online workload characterization technique based on Locality-Sensitive Hashing (LSH) [5]. Given a set of n points in a d-dimensional Euclidean space, LSH is a hashing technique that allows us to find nearest neighbor points for any query point with nearly asymptotically optimal running time. We maintain a Training Set of applications for which all sensitivity scores are known. When a new application arrives, we quickly profile the new application for a few SIs (the ones that need to be more accurate) and calculate sensitivity scores for those SIs. We then, use LSH to find applications in the training set with similar sensitivity scores so that we can predict the remaining sensitivity scores. We predict the remaining scores by taking the median of the similar applications’ corresponding sensitivity scores. Once the scores are obtained, we again use LSH to search for a server for the application. We encode the information regarding the current applications on a particular server as a point in a high-dimensional Euclidean space, and LSH quickly returns a point that corresponds to a server that is a good fit for the application.

**Fig. 1:** How application QoS is affected due to errors in different SIs.

We evaluated our scheme using SPEC, SPLASH2, PARSEC, PUMA Hadoop [3], CloudSuite [16] as well as multiprogrammed workloads. We experimented with 3 different data center configurations. Our approach predicts SIs with high accuracy and selects servers quickly. Compared to Paragon, our approach characterizes applications 1.42 times as much accurately. It achieves an average CPU utilization of 62% during a high load scenario for our default data center. During such a load, 60% applications achieve at least 80% QoS.

The paper is organized as follows: Section II provides background on sources of interference and locality-sensitive hashing; Section III explains the main idea; Section IV outlines the implementation issues; Section V provides the experimental results; Section VI discusses related work and finally, Section VII concludes.

**II. BACKGROUND**

**Sources of Interference.** We develop a set of kernels to characterize an application’s sensitivity toward SIs. For each SI, we calculate two metrics - one to measure how much resource interference this application can tolerate from other applications and another to measure how much resource interference this application causes to other applications. Each of them is a real number in [0, 1]. For each SI, we develop a kernel whose intensity (i.e. its consumption of the particular resource expressed as a fraction of the total resource) can be progressively tuned up. The kernels are similar to iBench [12]. To calculate the Tolerate metric, we run the kernel for the SI in parallel with the application and progressively tune up the intensity of the kernel until the application’s performance degrades to 95% of its isolated performance in the best server. An application with a high tolerance toward the SI will have a high value for this metric. To determine the Cause metric, we run the kernel in parallel with the application as before. This time, we tune up the intensity of the kernel until the kernel’s performance degrades to 95% of its original isolated performance in the best server. The metric is then set to the value of 1 – intensity. An application that consumes a lot of the resource will have a high value for this metric. We use the term Sensitivity Score to refer to both of these metrics. Since our kernels and measuring technique of sensitivity scores are similar to iBench, we are not going to describe the kernels here. Interested readers can look into that paper for more details. We consider 10 SIs. The SIs are memory capacity and bandwidth, storage capacity and bandwidth, network bandwidth, LLC capacity and bandwidth, TLB capacity, integer processing unit and floating point processing unit.

**Nearest Neighbor Search** Our approach for workload profiling utilizes techniques from computational geometry, namely nearest neighbor search. In this subsection we give our motivation for considering these techniques, and we then provide some background on these techniques.

**Motivation.** In our profiling scheme, we will maintain a training set T of n applications for which we will determine all sensitivity scores in an offline preprocessing step. As described above, for each SI we obtain sensitivity scores for two different metrics which gives us a total of twenty sensitivity scores, and suppose the applications T1, T2, ..., Tn have all six scores. For each application Ti ∈ T, we denote its sensitivity scores s1, s2, ..., sn where s1 and s2 are the tolerate and cause sensitivity scores for SI-1, s3 and s4 are similar scores for SI-2, etc. When a new application Ai arrives, we choose a small number of SI and determine the exact sensitivity scores for these SIs and approximate the remaining scores in an effort to save time. These approximate scores are determined by identifying the applications in T which are the most similar to Ai with respect to the computed scores.

To illustrate our high level idea, consider the following example. For simplicity, suppose that there are six total sensitivity scores, and suppose the applications T1, ..., T5 in Table I are the applications in T for which we know all six scores. Suppose for a new application Ai we determine s1 = 0.2 and
and $s_2^2 = 0.3$, and we now wish to approximate the remaining four scores for $a_i$. The applications in $T$ that are the most similar with respect to $s_1$ and $s_2$ are $t_1$ and $t_3$. We obtain the approximations for $a_i$ as a function of the known scores of $t_1$ and $t_3$.

<table>
<thead>
<tr>
<th>$T$</th>
<th>$s_1$</th>
<th>$s_2$</th>
<th>$s_3$</th>
<th>$s_4$</th>
<th>$s_5$</th>
<th>$s_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>0.1</td>
<td>0.3</td>
<td>0.7</td>
<td>0.6</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>$t_2$</td>
<td>0.7</td>
<td>0.6</td>
<td>0.4</td>
<td>0.5</td>
<td>0.9</td>
<td>0.7</td>
</tr>
<tr>
<td>$t_3$</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>$t_4$</td>
<td>0.5</td>
<td>0.2</td>
<td>0.4</td>
<td>0.1</td>
<td>0.8</td>
<td>0.6</td>
</tr>
<tr>
<td>$t_5$</td>
<td>0.7</td>
<td>0.6</td>
<td>0.1</td>
<td>0.2</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**TABLE I:** Training set example.

In order for this approach to be successful, we need to maintain a large enough training set $T$ so that all incoming applications will have a few applications in $T$ that are “similar” with respect to the sensitivity scores. At the same time, it is important that we can efficiently search with respect to the sensitivity scores. At the same time, it is important that we can efficiently search with respect to the sensitivity scores.

**4** approximations for a point in $\mathbb{R}^d$, where $d$ is a real number, it is natural to view each application as a point in $\mathbb{R}^d$, and thus the problem can be modeled as a geometric nearest neighbor search problem. As each application has twenty sensitivity scores, each of which is a real number, it is natural to view each application as a point in $\mathbb{R}^{20}$, where the sensitivity scores are the coordinates of the point. Formally, for each application $a_i$ we have a point $p_i = (s_{i1}, s_{i2}, \ldots, s_{i20})$ (in the example, we have $p_1 = (0.1, 0.3, 0.7, 0.6, 0.1, 0.3), p_2 = (0.7, 0.6, 0.4, 0.5, 0.9, 0.7), \ldots$). Now consider two such points $p_i$ and $p_j$. If these points are “close” to one another in the geometric space, then their sensitivity scores are also “close” to one another, and therefore the corresponding applications $a_i$ and $a_j$ perform similarly with respect to the SoI. Likewise, if $p_i$ and $p_j$ are “far apart” from one another then $a_i$ and $a_j$ perform differently with respect to the SoI. Therefore we can find applications in $T$ which are similar to $a_i$ by performing nearest neighbor queries on these points.

**Locality-Sensitive Hashing.** One technique for nearest neighbor search is Locality-Sensitive Hashing (LSH). The high level idea behind LSH is to build a hash table on the points in $\mathbb{R}^d$ such that points that are nearby in $\mathbb{R}^d$ get hashed to the same “bucket” and points that are far apart in $\mathbb{R}^d$ get hashed to different buckets. Then nearest neighbor queries can be performed by determining which bucket the query point $q$ lies in and then scanning the points of $P$ that were hashed to the same bucket. LSH was originally introduced in 1999 [17] and several improvements have since been given [4], [10]. A C++ implementation given by the authors of [4] has been made available [2], and this is the implementation that we use.

LSH can be used to solve several variants of nearest neighbor search, and the variant that will be considered in this paper is randomized R-near neighbor reporting. That is, given parameters $R > 0$ and $\delta \in [0, 1)$, we will use LSH to report points in $P$ whose distance to a query point $q$ is at most $R$, and each such point will be reported with probability $1 - \delta$. Note that every point in $P$ could be reported if $R$ is large enough and could return no points if $R$ is small enough; however, our motivation for considering this variant is that for an appropriately chosen $R$, all points whose distance to the query is at most $R$ will be strong candidates for approximating sensitivity scores. Additionally we can obtain this set of candidates very quickly as the hashing does not depend on the size of $T$, and it scales very nicely for high-dimensional data (e.g. thousands of dimensions). This is sufficient background on LSH for understanding our workload profiling and scheduling techniques (in particular, knowing how it is implemented is not important for this paper); we refer the interested reader to see [5] for a nice introduction to LSH.

**III. Characterizing and Scheduling Workloads**

In this section, we give the details of our approach to workload profiling and scheduling using LSH. We compute a training set $T$ of applications for which we know all sensitivity scores across all server configurations. When a new application $a_i$ arrives, we use LSH to quickly find applications in $T$ which will allow us to approximate the sensitivity scores for $a_i$ with lower error. Given our approximation of the sensitivity scores, we use LSH to find a server to schedule $a_i$ while maintaining the QoS for $a_i$ as well as any applications currently running on the server. We first describe how we compute approximate sensitivity scores, and then we show how we use these scores to find a good server for $a_i$.

**Workload Profiling.** We will present two procedures. The first is an offline procedure which is given a pool of applications for which we know all of the sensitivity scores, and it carefully chooses a subset of applications from the pool to serve as the training set. The procedure then outputs the associated LSH data structure. The second procedure is an online procedure which, given a new application, uses the LSH hash table given by the offline procedure to provide a fast and accurate approximation of the new application’s sensitivity scores. We first describe the profiling for a single server configuration, and later describe how to generalize the approach for a heterogeneous datacenter.

**Offline Procedure.** Let $A$ denote a set of applications for which we know all twenty sensitivity scores. The offline procedure begins by choosing a training set $T \subseteq A$ of cardinality $n$ for some parameter $n$. Recall that these applications naturally map to points in $\mathbb{R}^{20}$. Intuitively, we want the points associated with the training set applications to be distributed throughout $\mathbb{R}^{20}$, so that any new application received in the online procedure will have a few points in the training set nearby. To achieve this, we use the well-known $k$-means clustering algorithm [18] to partition $A$ into $k$ clusters for some constant $k$ (i.e., $k = 10$). We interpret the applications assigned to the same cluster as being of the same “type”, and we choose our training set to contain several applications from each of the different “types”. To do this, we randomly choose $n/k$ applications from each cluster to be in $T$.

Now that we have chosen our training set $T$ of $n$ applications, we are ready to build the LSH data structure. In the online procedure we will receive a new application $a_i$, and we choose a SoI to compute the associated $2\alpha$ sensitivity scores (for some parameter $\alpha$, e.g., $\alpha = 2$) and then approximate the remaining scores. Recall that each sensitivity score has a DEI, and our goal is to maximize the number of approximate sensitivity scores which fall within their DEI. The DEI has the form $(-\alpha Y, \alpha X)$ which implies that we can underestimate the score by at most $Y/\alpha$ and we can overestimate the score by at
most $X\%$. We define the width of a DEI to be $X + Y$. Since each SoI has two sensitivity scores, it also has two associated DEIs. Let $w_1$ and $w_2$ denote the two corresponding DEI widths of a SoI. We define the width of an SoI to be the minimum of $w_1$ and $w_2$. The SoI with the smallest widths have the least room for error, and accordingly we want to compute exact scores for the SoI with the smallest widths. To this end, we use the $\alpha$ SoI with the smallest widths.

Let $S$ denote the set of $\alpha$ SoI with the smallest widths. When a new application arrives in the online procedure, we will exactly compute the $2\alpha$ sensitivity scores associated with $S$. From this we obtain a point $p_i \in \mathbb{R}^{2\alpha}$, and therefore we want to perform nearest-neighbor queries in a $2\alpha$-dimensional space. For each application in $T$, we construct a point with $2\alpha$ dimensions to be inserted into the LSH table. The coordinates of this point consists of the $2\alpha$ sensitivity scores associated with $S$. We call this point the projection onto the SoI of $S$. Recall that for some parameters $R$ and $\delta$, the LSH table will return any point within distance $R$ from a query point with probability $1 - \delta$. We choose $R$ to be $2\alpha$ so that training set points whose distance is at most $1$ away from our query point in each coordinate (on average) are returned. We choose $\delta$ to be 0.05 so that each point within distance $R$ is returned with probability 0.95.

For example, again consider the training set given in Table I. Suppose $\alpha = 1$, and that the first SoI has the smallest width. Then we will exactly compute the first two sensitivity scores associated with this SoI in the online procedure, and therefore we want to build a LSH table based on these scores in the offline procedure. The projection of application $t_1$ onto these scores gives us the point $(0.1, 0.3)$, the projection of $t_2$ gives us $(0.7, 0.6)$, and the remaining points are constructed similarly. See Algorithm 1 for a formal description of our offline procedure. We remark that when implemented, one could execute this offline procedure several times per day (e.g. every hour) to ensure that the training set is a good representation of the applications that are being received.

Algorithm 1 Offline Procedure

Let $A$ be a set of applications for which all sensitivity scores are known. Use the $k$-means algorithm to partition $A$ into $k$ clusters. Let $C$ denote output clusters.

$T \leftarrow \emptyset$

for all clusters $C \in C$ do

Let $C' \subseteq C$ be a randomly-chosen subset of elements in cluster $C$ such that $|C'| = n/k$.

$T \leftarrow T \cup C'$

end for

Let $S$ denote the $\alpha$ SoI with the smallest widths, and let $P$ denote the $n$ points in $\mathbb{R}^{2\alpha}$ obtained by projecting the applications in $T$ onto the SoI in $S$.

Build and save the LSH hash table $P$ for parameters $R = \frac{2\alpha}{10}$ and $\delta = 0.05$.

Online Procedure. Now we assume that we have the LSH hash table stored in memory, and we are given a new application $a_i$ for which we currently do not know any of its sensitivity scores. We compute the exact sensitivity scores of $a_i$ for the $2\alpha$ sensitivity scores associated with $S$, and this gives us a point $p_i \in \mathbb{R}^{2\alpha}$. We use the hash table to obtain a set of points $N'$ in the training set within distance $R$ of query point $p_i$. We then let $N$ be the subset of $N'$ consisting of the at most $c$ points in $N'$ that are closest to $p_i$ for some parameter $c$ (e.g., $c = 5$). We take the median of the scores of the applications in $N$ to determine our estimate of the remaining scores for $a_i$. See Algorithm 2 for a formal description.

Algorithm 2 Online Procedure

Let $a_i$ denote a new application for which we have no prior knowledge of its sensitivity scores.

Let $S$ denote the $\alpha$ SoI with the narrowest DEIs, and generate the $2\alpha$ associated sensitivity scores for $a_i$. Let $p_i$ denote the associated point in $\mathbb{R}^{2\alpha}$.

Let $N'$ be the set of points returned by LSH when using the query point $p_i$, and let $N' \subseteq N'$ be the at most $c$ points from $N'$ that are closest to $p_i$, breaking ties arbitrarily. If $N' = \emptyset$ then randomly choose scores for $a_i$ and exit.

for all choices $r$ such that $s_i^r$ is unknown do

Let $N''$ denote the set of all scores $s_j^r$ for each $p_j \in N$.

Set $s_i^r$ to be the median of $N''$.

end for

Extending to Many Server Configurations. In a heterogeneous data center, the effect of interference on an application may be quite different on different server configurations. Accordingly, the sensitivity scores for an application with respect to some SoI may be quite different for different configurations. In a data center with 10 server configurations, an application will have 200 sensitivity scores (20 scores for each server configuration). Our approach is similar to our previous one for a single server configuration. First, we choose the $\alpha$ SoI with the smallest widths, and then we compute the sensitivity scores with respect to these SoI for each configuration. Our approach is similar to our previous one for three different server configurations. We choose the configurations so that we are obtaining the scores for the “best” configuration, “median” configuration, and “worst” configuration. The scheduler keeps an ordered list of “best”, “median” and “worst” server configuration for a given SoI. The intuition is sensitivity scores for “good” server configurations may not be effective for predicting the scores for “bad” server configurations. For example, two applications which do not cause much interference on the best configuration may perform quite differently on the worst server configuration, and therefore we may not be able to accurately predict the scores for the worst configuration from the score of the best configuration. Given our choice of $\alpha$ SoI and three server configurations, we compute the $6\alpha$ corresponding sensitivity scores (two scores per SoI per configuration). We then perform a LSH query to find points in the training set whose scores are similar to our computed scores, and we take the median of the scores of these points to approximate the remaining $200 - 6\alpha$ scores.

Server Selection. Given the approximate sensitivity scores for a new application $a_i$, we now describe how we select a server. We want to schedule applications on servers so that all applications will have a performance at least 95% of their standalone performance. In order to do this, we will need to schedule applications in a way so that applications running simultaneously on the same server are “compatible” with the other applications running on the server. We first give the
intuition behind our approach, and then describe how we use LSH to choose a server for $a_i$.

Scheduling Intuition. Recall that for each SoI, we have two sensitivity scores for a fixed server configuration. Without loss of generality, consider SoI-1, and consider the sensitivity scores $s_1^i$ and $s_2^i$ of application $a_i$ with respect to SoI-1, both of which are real numbers in $[0, 1]$. We interpret $s_1^i$ (i.e., tolerate) to be a measure of how much the interference of other applications with respect to SoI-1 will affect the running time of $a_i$, where intuitively $s_1^i = 0$ implies that the running time of $a_i$ is not highly influenced by other applications’ interference and $s_1^i = 1$ implies that the running time of $a_i$ is heavily influenced by other applications’ interference. We interpret $s_2^i$ (i.e., cause) to be a measure of how much interference other applications will cause to other applications with respect to SoI-1, where $s_2^i = 0$ implies it causes very little interference and $s_2^i = 1$ implies it causes a lot of interference.

Let $S$ denote the set of applications currently running on a server. We schedule applications to ensure that the following invariant holds true for each application $a_j \in S$: $s_j^1 + \sum_{k : a_k \in S \setminus \{a_j\}} s_k^2 \leq 1$. Intuitively, this invariant ensures that the sum of the interference caused by all other applications does not exceed the tolerance level of $a_j$. Indeed, if $a_j^1$ is close to 0, then we allow it to be scheduled with applications which cause a large amount of interference, while if $a_j^1$ is close to 1 then the interference caused by other applications will have to be significantly lower.

When choosing a server for application $a_i$, there are two issues to deal with: (1) the interference caused by the applications running on the server should not exceed the threshold for $a_i$, and (2) the additional interference caused by $a_i$ should not exceed the threshold for some $a_j$ currently running on the server. Let $S_{\text{caused}} = \sum_{j : a_j \in S} s_j^2$ denote the sum of the interference caused by applications on some server (not including $a_i$). Suppose we want to check (1) and (2) to determine if $a_i$ can be scheduled with the applications in $S$ while maintaining the invariant. For (1) to hold, it must be that $s_i^1 + S_{\text{caused}} \leq 1$, or equivalently $s_i^1 \leq 1 - S_{\text{caused}}$.

Now consider issue (2), and consider any application $a_j \in S$. Since the invariant holds true, we have $s_j^1 + \sum_{k : a_k \in S \setminus \{a_j\}} s_k^2 \leq 1$, but $\sum_{k : a_k \in S \setminus \{a_j\}} s_k^2$ can equivalently be stated as $S_{\text{caused}} - s_j^2$. If $a_i$ can be scheduled without exceeding the threshold for $a_j$, then it must be that $s_i^1 + S_{\text{caused}} - s_j^2 - s_j^2 \leq 1$ which implies that $s_j^2 \leq 1 - S_{\text{caused}} - (s_i^1 - s_j^2)$. It follows that if we can schedule $a_i$ with the applications in $S$, it must be that $s_j^2 \leq 1 - S_{\text{caused}} - (s_i^1 - s_j^2)$ for all $a_j \in S$. Consider the application $a_j \in S$ such that $1 - S_{\text{caused}} - (s_i^1 - s_j^2)$ is the minimum such value over all applications in $S$. Then clearly $a_j$ will satisfy issue (2) for all applications in $S$ if and only if $s_j^2 \leq 1 - S_{\text{caused}} - (s_i^1 - s_j^2)$. Note that $a_j$ is the application in $S$ that maximizes $s_j^2 - s_j^2$. This is because $1 - S_{\text{caused}}$ is fixed for any application in $S$, and therefore $1 - S_{\text{caused}} - (s_i^1 - s_j^2)$ is minimized by maximizing the last term that is subtracted from $1 - S_{\text{caused}}$.

So it now follows that we can schedule $a_i$ with the applications in $S$ while maintaining the invariant if and only if the following two properties hold: (1) $s_i^1 \leq 1 - S_{\text{caused}}$, and (2) $s_j^2 \leq 1 - S_{\text{caused}} - (s_i^1 - s_j^2)$. Note that they are stated with respect to SoI-1, but they must hold for the sensitivity scores for all SoI.

Using LSH to Choose a Good Server. Before choosing a server to schedule $a_i$, we first consider which server configuration is the best fit for the application. Let $C_k$ denote the sum of the sensitivity scores of application $a_i$ for the $k$th server configuration. Since the two properties both imply that it is good to have small sensitivity scores, we first rank the server configurations in non-increasing order according to $C_k$. We then look for a server in this order until we find one that satisfies properties 1 and 2 for all SoI. If no such server exists, we check the next configuration and repeat.

Without loss of generality, suppose we are searching for a server of the first server configuration (SC-1). We will now describe how we use LSH to search for a good server for application $a_i$. The sensitivity scores for $a_i$ associated with server configuration 1 are $s_1^1, s_2^1, \ldots, s_{10}^1$, and as before, we view these scores as the coordinates of a point $p_i \in \mathbb{R}^{20}$. For each server $S_j$ of SC-1, we construct a point $p_j \in \mathbb{R}^{20}$ where the coordinates of these points correspond to the $1 - S_{\text{caused}}$ and $1 - S_{\text{caused}} - (s_i^1 - s_j^2)$ values for each of the ten SoI. We say that a server point $p_j$ dominates an application point $p_i$ if each coordinate of $p_j$ is greater than or equal to the corresponding coordinate of $p_i$. It follows that $a_i$ can be scheduled on server $S_j$ if and only if $p_j$ dominates $p_i$. To do this, we build an LSH table on the server points. The values of $S_{\text{caused}}$ for a particular server change over time, we periodically (e.g., every 10 minutes) rebuild these hash tables with updated server values. We query this hash table to look for a point that dominates $p_i$. See Algorithm 3 for a formal description. If Algorithm 3 ends without choosing a server, we add $a_i$ to a queue of applications to be scheduled later.

Algorithm 3 Server Selection

Given a new application $a_i$ with its 200 approximate sensitivity scores, let $SC_1^1, SC_2^1, \ldots, SC_{10}^1$ denote server configurations sorted in nondecreasing order according to the sum of the scores for that configuration.

for $j = 1$ to 10 do

Let $p_j$ denote the point in $\mathbb{R}^{20}$ corresponding with the 20 sensitivity scores associated with $SC_j^1$.

For some parameter $\epsilon > 0$, let $p'_i$ denote the point in $\mathbb{R}^{20}$ obtained by adding $\epsilon$ to each coordinate of $p_i$, rounding down any coordinate greater than 1.

while some coordinate of $p'_i$ is less than 1 do

Perform an LSH search of the server points for $SC_j^1$ using query point $p'_i$.

If LSH returns a server point that dominates $p_i$, schedule $a_i$ for this server, update the coordinates of the server point, and exit Algorithm 3. Otherwise, add $\epsilon$ to all coordinates of $p'_i$, rounding down any coordinate greater than 1.

end while

end for
**IV. Implementation**

The scheduler is in charge of initial profiling as well as scheduling when an application arrives. It consists of a *Master* and multiple *Worker* components. The master component runs on a node dedicated for scheduling. A worker component runs on each of the compute nodes. When a application arrives, the master chooses 2 SoIs for profiling. Recall that 2 SoIs with the narrowest DEI are chosen. DEIs are assumed to capture the significance of various shared resources in a particular data center and hence, remain fixed for a given data center. For each SoI, the master maintains a list of server configurations from best to worst and finds the best, worst, and (somewhat) medium configuration. It randomly picks a server for those configurations. The application is profiled for that particular SoI in those servers by the worker component. Profiling is done in parallel (in a minute). Each worker sends the scores back to the master. The master then uses Algorithm 2 to predict sensitivity scores for the rest of the SoIs and server configurations. In a large data center, there can be multiple masters - each in charge of a portion of the data center. Sensitivity scores are stored in the local disk of the master. Periodically (every hour or so), the master applies Algorithm 1 on the locally stored sensitivity scores to build the training set and LSH data structures. This is done to ensure that the training set remains a good representation of the applications. Locally stored sensitivity scores for older applications are removed in every few hours to keep the storage consumption under a limit. When sensitivity scores of an incoming application are calculated, the master uses Algorithm 3 to schedule the application. The entire scheduler is implemented in Python.

**V. Evaluation**

The characterization parameters and server configurations for our experiments are given in Table II. We experimented with 3 data centers. Each of them contains 10 different server configurations and a total of 50 servers. Our high end data center contains more instances of the *Best Server* (Table II) whereas low end one contains more instances of the *Worst Server*. Our default data center contains an even mixture of different server configurations. We use all applications from SPEC CPU 2000 & 2006, Parsec, PUMA, Hadoop, and CloudSuite benchmarks. We also generate 100 multiprogrammed workloads, each consisting of 4 applications from SPEC. Thus, in total, we have 198 workloads consisting of individual applications and multiprogrammed workloads.

![Table II: Parameters for experiments. Bold values are the defaults.](image)

<table>
<thead>
<tr>
<th>Characterization parameter</th>
<th>Num. cluster, k = 5, 10, 15, 20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Num. train. set, n = 50, 100, 150, 200</td>
</tr>
<tr>
<td></td>
<td>Num. near. neighbor, c = 5, 6, 7,..., 15</td>
</tr>
<tr>
<td></td>
<td>Num. train. SoI, α = 2, 3, 4,..., 10</td>
</tr>
</tbody>
</table>

**DEI Analysis.** We experiment with different errors injected in SoI scores (SS) and measure what fraction of applications achieve at least 95% QoS. We construct DEI of an SoI as the error interval outside which 5% or more applications (compared to the ones obtained by using accurate SS) fail to reach the QoS threshold. Figure 2(f) shows the DEIs for different SoIs. *memBandwidth* and *tlbCapacity* have the smallest widths. Hence, we choose them for initial profiling.

**Prediction Ability.** Figure 2(a) shows the comparison of LSH based approach against two versions of Paragon. One version chooses the same initial SoIs as LSH while the other (Paragon(r)) chooses SoIs randomly. For implementing Paragon, we used a hand tuned version of matrix factorization algorithm available from Apache Mahout [1]. For illustration, the figure shows the case for an ideal predictor. For each approach, we calculate how many sensitivity scores (SSs) fall within the corresponding DEIs. LSH based approach consistently performs better or similar to both versions of Paragon. The first version of Paragon performs better than Paragon(r). Therefore, from now on, we will only consider the first version. For the default choice of 2 SoIs, LSH predicted 14 out of 16 SSs correctly. This is 1.42 times more accurate than Paragon. Figure 2(b) shows correct SSs as we increase the training set size. As before, we keep other parameters fixed at the default values. In LSH-based approach, with larger training set, we are likely to find many workloads that very similar to the incoming one and hence, the accuracy increases initially. After the size of 35, the accuracy remains more or less the same. For Paragon the accuracy increases slightly; however, it always remains below the one for LSH. Figure 2(c) shows the accuracy as we increase the number of nearest neighbors in LSH-based approach. The accuracy increases up to 4 neighbors and then remains the same. For the default value of 5 neighbors, it predicts 13.78 SSs per workload.

Figure 2(d) shows the number of correctly predicted SSs per workload for different server configurations. In each configuration, our approach works better than Paragon. For the best (i.e. Server 10) and worst (i.e. Server 1) server LSH predicts 13.7 and 13.98 SSs correctly. Figure 2(e) shows, for each SS, what fraction of applications has been correctly predicted with LSH based approach and Paragon. Here, we are considering the worst server. Out of 16 SSs, LSH based approach and Paragon predicts 5 scores with similar accuracy. Among the remaining scores, LSH predicts 10 scores more accurately whereas Paragon predicts only 1 score more accurately. Other servers provide similar results.

**QoS and Utilization.** We measure the QoS of different applications running in the default data center. Figure 3 shows distribution of QoS using LSH and Paragon scheduler. Each stack in Figure 3(a) and (d) corresponds to the % of applications that suffer from a certain level of QoS degradation. Figure 3(a) is for high load scenario (i.e. one application in every 0.5 sec) whereas (d) is for low load scenario (i.e. one application in every sec). During high load, 45% applications suffer from QoS degradation of at most 5% in LSH based approach. The number for Paragon is 38%. During low load, the same number is 61% and 49% for LSH based approach and Paragon respectively. So, compared to Paragon, LSH based prediction always allows more applications to achieve QoS of 95% or more.

Figure 3(b) and (c) show data center utilization map in high load whereas (e) and (f) show the same for low load. Figure 3(b) and (e) are for LSH based approach whereas (c) and (f) are for Paragon. Darker area implies higher utilization. During both high and low load, LSH based approach produces
Fig. 2: (a)-(e) show prediction error for lsh and Paragon. (f) shows DEIs.

Fig. 3: QoS and utilization achieved using different approaches.

Fig. 4: QoS in (a) higher and (b) lower end data centers.
set (Figure 5). Prediction time tends to remain the same for training set size up to 100. After that, it tends to increase slightly. For our default training set size of 150, the prediction time is 0.013 seconds for each workload. This is extremely fast. Our initial profiling takes (≈)60 seconds.

![Fig. 5: Running time analysis.](image)

**VI. RELATED WORK**

There has been significant work on data center workload characterization [11], [15], [20], [6], [7]. This line of work generates workloads with characteristics that closely resemble those of the original applications. The generated workloads are then used in system studies. Although this is a viable approach, the generated workloads sometimes cannot capture every aspect of the applications. Moreover, they are not suitable for unknown applications. Mars et al. [21], [24] designed two kernels that create tunable contention in memory capacity and bandwidth to quantify the sensitivity of a workload to memory interference. The kernels are used during either offline profiling [21] or online profiling [24]. Tang et al. [22] designed SmashBench, a benchmark suite for cache and memory contention. Delimitrou et al. [12] proposed iBench, a benchmark suite to obtain sensitivity curve of a workload for 15 shared resources. Paragon [13] is the first work to predict a workload's sensitivity based on the knowledge about existing applications. The prediction is done using Netflix [9] algorithm. Once the sensitivity scores are predicted, Paragon applies a greedy algorithm for server selection to maximize utilization while minimizing interference. The work is later extended in Quasar [14] where server selection is done based on predicted sensitivity scores and performance constraints given by a user.

**VII. CONCLUSION**

We have presented new method of evaluating and scheduling workload in a heterogeneous data center. The technique that is based on locality-sensitive hashing. Given a new application, we are interested in approximating its dependence on certain resources. Due to time constraints, we can only spend a small amount of time profiling the application. After this profiling, we are able to identify similar applications from a training set extremely quickly using locality-sensitive hashing. We then use these similar applications to approximate the remaining information for the new application. We then use LSH to select appropriate servers to schedule. We demonstrated the effectiveness of our approach with respect to our new evaluation metric by comparing our results with that of Paragon. We demonstrate that our approach predicts more accurately (by a factor of 1.42) and achieves better QoS than Paragon.

**REFERENCES**


