Intelligent Resource Scheduling for Co-located Latency-critical Services: A Multi-Model Collaborative Learning Approach

Abstract
Latency-critical services have been widely deployed in cloud environments. For cost-efficiency, multiple services are usually co-located on a server. Thus, run-time resource scheduling becomes the pivot for QoS control in these complicated co-location cases. However, the scheduling exploration space enlarges rapidly with the increasing server resources, making the schedulers hardly provide ideal solutions quickly. More importantly, we observe that there are “resource cliffs” in the scheduling exploration space. They affect the exploration efficiency and always lead to severe QoS fluctuations. Resource cliffs cannot be easily avoided in previous schedulers. To address these problems, we propose a novel ML-based intelligent scheduler – OSML. It learns the correlation between architectural hints (e.g., IPC, cache misses, memory footprint, etc.), scheduling solutions and the QoS demands based on a data set we collected from 11 widely deployed services running on off-the-shelf servers. OSML employs multiple ML models to work collaboratively to predict QoS variations, shepherd the scheduling, and recover from QoS violations in complicated co-location cases. OSML can intelligently avoid resource cliffs during scheduling and reach an optimal solution much faster than previous approaches for co-located LC services. Experimental results show that OSML supports higher loads and meets QoS targets with lower scheduling overheads and shorter convergence time than previous studies.

1. Introduction
Cloud applications are shifting from monolithic architectures to loosely-coupled designs, consisting of many latency-critical (LC) services (e.g., some microservices, interactive services) with strict QoS requirements [18,19,52,53]. Many cloud providers, including Amazon, Alibaba, Facebook, Google, and LinkedIn, employ this loosely-coupled design to improve productivity, scalability, functionality, and reliability of their cloud systems [2,3,18,52].

QoS-driven resource scheduling faces more challenges in this era. The cost-efficiency policy drives providers to colocate as many applications as possible on a server. However, these co-located services exhibit different behaviors across multiple interactive resources, such as CPU cores, cache, bandwidth, and main memory banks. These behaviors also can be drastically different from demand to demand and change within seconds. Moreover, with the increasing number of cores, more threads contend for the shared LLC (last-level cache) and memory bandwidth. Notably, these shared resources interact with each other [32,33,45]. All these issues make resource scheduling for co-located LC services more complicated and time-consuming. Moreover, in reality, end-users keep increasing demands for quick responses from the cloud [15,47,53]. According to Amazon’s estimation, even if the end-users experience a 1-second delay, they tend to give up the transactions, translating to $1.6 billion loss annually [1]. Briefly, unprecedented challenges are posed for resource scheduling mechanisms [8,10,31,47,52]. Some previous studies [17,31,33,45,61] design clustering approaches to allocate LLC or LLC together with main memory bandwidth to cores for scheduling single-threaded applications. Yet, they are not suitable in cloud environments, as the cloud services often have many concurrent threads and strict QoS constraints (i.e., latency-critical). Alternatively, the existing representative studies either use heuristic algorithms—increasing/decreasing one resource at a time and observing the performance variations [10] or use learning-based algorithms (e.g., Bayesian optimization [46]) in a relatively straightforward manner. The studies in [10,46] show that scheduling five co-located interactive services to meet certain QoS constraints can take more than 20 seconds on average. Existing schedulers still have room for improvement in scheduling convergence time, intelligence, and how to schedule complicated interactive resources (e.g., parallel computing units and complex memory hierarchies) in a timely fashion. Moreover, the existing schedulers can hardly avoid “resource cliffs”, i.e., decreasing a resource only slightly during scheduling leads to a significant QoS slowdown. To the best of our knowledge, our community has been expecting new directions on developing resource-scheduling mechanisms to handle co-located LC services [16,30,31,45].

To this end, we design OSML, a novel machine learning (ML) based resource scheduler for LC services on large-scale servers. In practice, using ML models significantly improves scheduling exploration efficiency for multiple co-located cloud services and can handle the complicated resource sharing, under/over-provision cases timely. ML has achieved tremendous success in improving speech recognition [54], benefiting image recognition [25], and helping the machine to beat the human champion at Go [13,24,51]. In OSML, we make progress in leveraging ML for resource scheduling, and we make the following contributions.

1) Investigation in RCliff for Multiple Resources During Scheduling. We study resource cliff (RCliff, i.e., reducing a resource only slightly leads to a significant QoS slowdown) for computing and cache resources. More importantly, we show that RCliffs commonly exist in many widely used LC services and find the heuristic schedulers can hardly avoid RCliff (Sec.3.3), always leading to a sudden and sharp QoS slowdown. Furthermore, we show ML can be an ideal approach that benefits scheduling (Sec.4.4).

2) Collaborative ML Models for Intelligent Scheduling. OSML is an ML-based scheduler that intelligently schedules multiple interactive resources to meet co-located services’ QoS targets. OSML learns the correlation between architectural hints (e.g., IPC, cache misses, memory footprint, etc.), optimal scheduling solutions, and the QoS demands. It employs MLP models to avoid RCliffs intelligently, thus avoiding the sudden QoS slowdown often incurred by the RCliffs in prior schedulers; it predicts the QoS variations and resource margins, and then delivers appropriate resource allocations. It
Table 1: Latency-critical (LC) services, including micro-/interactive services with the 99th percentile tail latency QoS target [10,18,46,52].

| LC service     | Domain                        | RPS (Requests Per Second) |
|----------------|-------------------------------|---------------------------|
| Image recognition | 2000,3000,4000,5000,6000 (Max) |                           |
| Key-value store       | 3000,3400,3800,4200,4600         |                           |
| RT translation       | 2200,2400,2600,2800,3000         |                           |
| Web server          | 60k,120k,180k,240k,300k          |                           |
| Online search       | 3600,4400,5200,6000,6800         |                           |
| Login              | 300,600,900,1200,1500            |                           |
| Online renting ads  | 10,100,1000                     |                           |

Table 2: Our platform specification vs. a server used 10 yrs. before.

| Conf. / Servers | Our Platform               | Server (10 Years Ago) |
|-----------------|----------------------------|-----------------------|
| CPU Model       | Intel Xeon E5-2697 v4       | Intel i7-860          |
| Processor Speed | 2.3GHz                     | 2.8GHz                |
| Logical Processor Cores | 36 Cores (18 phy. cores) | 8 Cores (4 phy. cores) |
| Main Memory / Channel / BW | 256GB, 2400MHz DDR4 | 8GB, 1600MHz DDR3 |
| Private L1 & L2 Cache Size | 32KB and 256KB | 32KB and 256KB |
| Disk            | 1TB,7200 RPM,HD            | 500GB,5400 RPM,HD     |
| GPU             | NVIDIA GP104               | N/A                   |

leverages an enhanced DQN to shepherd the allocations and recover from the QoS violation and resource over-provision cases. Moreover, as OSML’s models are lightweight and their functions are clearly defined, it is easy to locate the problems and debug them.

(3) An Open-sourced Data Set for Low-overhead ML. We have collected the performance traces for widely deployed LC services (in Table 1), covering 62,720,264 resource allocation cases (including the corner cases) that contain around 2-billion samples. These data have a rich set of information, e.g., the RCliffs for multiple resources; the interactions between workload load, and the mainstream architectures. Our models can be trained and generalized with these data and then used on new platforms with low-overhead transfer learning. We show our data-driven approach is promising in terms of performance and deployment efficiency. We will make our data set publicly available; people can study it and efficiently train their models without an extended period for data collection.

(4) Real Implementation and Detailed Comparisons. We implement OSML based on latest Linux. OSML is designed as a co-worker of the OS scheduler located between the OS kernel and the user layer. We compare OSML with the most related open-source studies [10,46] and show the advantages.

In practice, OSML captures the applications’ online behaviors and forwards them to the ML models running on CPU or GPU, and schedules resources accordingly. Compared with [10,46], OSML takes 36~55% less time to meet the QoS targets and support 10~50% of higher loads. OSML supports to reclaim over-provisioned resources to improve efficiency. Its ML models are with low run-time overheads. We make OSML open source as planned.

2. Background and Motivation

The cloud environment has a trend towards a new model [3,18,52], in which cloud applications comprise numerous distributed LC services (i.e., micro/interactive services), such as key-value storing, database serving, and business applications serving [18,19]. Table 1 includes some widely used ones, and they form a significant fraction of cloud applications [18]. These services have different features and resource demands.

In terms of the datacenter servers, at present, new servers can have an increased number of cores, larger LLC capacity, larger main memory capacity, higher bandwidth, and the resource scheduling exploration space becomes much larger than ever before as a result. Table 2 compares the two typical servers used at different times. On the one hand, although modern servers can have more cores and memory resources than ever before, they are not fully exploited in today’s cloud environments. For instance, in Google’s datacenter, the CPU utilization is around 45~53% and memory utilization ranges from 25~77% during 25 days, while Alibaba’s cluster exhibits a lower and unstable trend, i.e., 18~40% for CPU and 42~60% for memory in 12 hours [32,49], indicating that a lot of resources are wasted. On the other hand, the larger resource scheduling exploration space, which consists of more diverse resources, prohibits the schedulers from achieving the optimal solution quickly. Additionally, cloud applications can have dozens of concurrent threads [10,46]. When several cloud applications run on a server, they share and contend resources across multiple resource layers – cores, LLC, memory bandwidth/banks. Previous studies show they may incur severe performance degradation and unpredictable QoS violations, and propose the scheduling approaches at architecture [9,23,44], OS [31,45,50], and user-level [10,37,38]. Yet, do they perform ideally for scheduling co-located LC services on modern datacenter servers?

3. Resource Scheduling for LC Services

To answer the above question, we study the LC services (Table 1) that are widely deployed in cloud environments.

3.1. Understanding the LC Services - Resource Cliff

We study how sensitive these LC services are to the critical resources, e.g., the number of cores and LLC capacity, on a commercial platform (“our platform” in Table 2). For Moses, as illustrated in Figure 1-a, with the increasing number of cores, more threads are mapped on them simultaneously. Meanwhile, for a specific amount of cores, more LLC ways can benefit performance. Thus, we observe the response latency is low when computing and LLC resources are ample (i.e., below 10ms in the area within green color). The overall trends are also observed from other LC services.

However, we observe the Cliff phenomenon for these services. In Figure 1-a, in the cases where 6 cores are allocated to Moses, the response latency is increased significantly from 34ms to 4644ms if merely one LLC way is reduced (i.e., from 10 ways to 9 ways). Similar phenomena also happen in cases where computing resources are reduced. As slight resource re-allocations bring a significant performance slowdown, we denote this phenomenon as Resource Cliff (RCliff). It is de-
OAA is sensitive to the number of threads, i.e., if someone starts Allocation Area (OAA) in the scheduling exploration space, (denoted by the red box in Figure 1-a), there would be a sharp overall response latency can be higher (as illustrated in Figure 20/28/36) and mapped across a different number of cores, the 

more threads, will the OAA change? When more threads are started (e.g., Moses as an example, on the RCliff) (denoted by the red box in Figure 1-a), there would be a sharp performance slowdown if only one core or one LLC way (or both) is deprived. Figure 1-b and c show RCliffs for Img-dnn and Sphinx, respectively. From another angle, RCliff means that a little bit more resources will bring significant performance improvement. Figure 1-a shows that Moses exhibits RCliff for both core and LLC. Moreover, we test the services in Table 1 across various RPS and find the RCliffs always exist, though the RCliffs vary (8.8% on average) according to different RPS.

The underlying reason for the cache cliff is locality; for the core cliff, the fundamental reason is on queuing theory - the latency will increase drastically when the request arrival rate exceeds the available cores. RCliff alerts the scheduler not to allocate resources close to it because it is "dangerous to fall off the cliff" and incurs a significant performance slowdown, i.e., even a slight resource reduction can incur a severe slowdown. Notably, in Figure 1, we highlight each LC service’s Optimal Allocation Area (OAA) in the scheduling exploration space, defined as the ideal number of allocated cores and LLC ways to bring an acceptable QoS. More resources than OAA cannot deliver more significant performance, but fewer resources lead to the danger of falling off the RCliff. OAA is the goal that schedulers should achieve.

3.2. Is OAA Sensitive to the Number of Threads?

In practice, an LC service may have many threads for higher performance. Therefore, we come up with the question: *is the OAA sensitive to the number of threads, i.e., if someone starts more threads, will the OAA change?*

To answer this question, for a specific LC service, we start a different number of threads and map them across a different number of cores (the num. of threads can be larger than the num. of cores). Through the experiments, we observe - (i) More threads do not necessarily bring more benefits. Take Moses as an example, when more threads are started (e.g., 20/28/36) and mapped across a different number of cores, the overall response latency can be higher (as illustrated in Figure 2). The underlying reason lies in more memory contentions at memory hierarchy and more context switch overheads, leading to a higher response latency [20,36]. (ii) The OAA is not sensitive to the number of concurrent threads. For Moses in Figure 2, even if 20/28/36 threads are mapped to 10–25 cores, around 8/9-core cases always perform ideally. Other LC services in Table 1 also show a similar phenomenon, though their OAs are different for different applications.

In practice, if the QoS for a specific LC service is satisfied, LLC ways should be allocated as less as possible, saving LLC space for other applications. Similarly, we also try to allocate fewer cores for saving computing resources. Here, we conclude that the OAA is not sensitive to the number of threads. We should further reveal: *how do the existing schedulers perform in front of OAs and RCliffs?*

3.3. Issues the Existing Schedulers May Meet

We find three main shortcomings in the existing schedulers when dealing with OAs and RCliffs. (1) **Entangling with RCliffs.** RCliffs challenge the schedulers. Many schedulers often employ heuristic scheduling algorithms, i.e., they increase/reduce resources until the monitor alerts that the system performance is suffering a significant change (e.g., a severe slowdown). Yet, these approaches could incur unpredictable latency spiking. For example, if the current resource allocation for an LC service is in the base of RCliff (i.e., the yellow color area in Figure 1-a/b/c), the scheduler has to try to achieve OAA. However, as the scheduler doesn’t know the “location” of OAA in the exploration space, it has to increase resources step by step in a fine-grain way, thus the entire scheduling process from the base of the RCliff will incur very high response latency. For another example, if the current resource allocation is on the RCliff or close to RCliff, a slight resource reduction for any purpose could incur a sudden and sharp performance drop for LC services.
Table 3: The input parameters for ML models.

| Feature     | Description                                      | Models |
|-------------|--------------------------------------------------|--------|
| IPC         | Instructions per clock                          | A/A'/B/B'/C |
| Cache Misses| LLC misses per second                           | A/A'/B/B'/C |
| MBL         | Local memory bandwidth                          | A/A'/B/B'/C |
| CPU Usage   | The sum of each core’s utilization              | A/A'/B/B'/C |
| Virt. Memory| Virtual memory in use by an app                 | A/A'/B/B'/C |
| Res. Memory | Resident memory in use by an app                | A/A'/B/B'/C |
| Allocated Cores | The number of allocated cores                   | A/A'/B/B'/C |
| Allocated Cache | The capacity of allocated cache                 | A/A'/B/B'/C |
| Core Frequency | Core Frequency during run time                 | A/A'/B/B'/C |
| QoS Slowdown | Percentage of QoS slowdown                      | B       |
| Expected Cores | Expected cores after deprivation               | B'      |
| Expected Cache | Expected cache after deprivation               | B'      |
| Cores used by N. | Cores used by Neighbors                      | A/B/B'/C |
| Cache used by N. | Cache capacity used by Neighbors              | A/B/B'/C |
| MBL used by N. | Memory BW used by Neighbors                    | A/B/B'/C |
| Resp. Latency | Average latency of a LC service                | C       |

The previous efforts [10,32,50,53] find there would be about hundreds/thousands of times latency jitter, indicating the QoS cannot be guaranteed during these periods. Thus, RCliffs are essential and should not be neglected when designing a scheduler. (2) Unable to accurately and simultaneously schedule a combination of multiple interactive resources (e.g., core counts, LLC ways and bandwidth usage) to achieve OAA in low overheads. Prior studies [10,31,32,45] show that the core computing ability, cache hierarchy, and memory bandwidth are interactive factors for resource scheduling. Solely considering a single dimension in scheduling often leads to sub-optimal QoS. However, the existing schedulers using heuristic or model-based algorithms usually schedule one dimension resource at a time and bring high overheads on scheduling multiple interactive resources. For example, the state-of-the-art work PARTIES [10] takes around 20–30 seconds on average (up to 60 seconds in the worst cases) to find ideal allocations when 3–6 LC services are co-running. The efforts in [16,41,42] also show the heuristics inefficiency due to the high overheads on scheduling various resources with complex configurations. (3) Unable to provide accurate QoS predictions. Therefore, the scheduler can hardly balance the global QoS and resource allocations across all co-located applications, leading to QoS violations or resource over-provision.

An ideal scheduler should avoid the RCliff and quickly achieve the OAA from any positions in the scheduling space. We claim it is time to design a new scheduler, and using ML can be a good approach to handle such complicated cases in low overheads.

4. Leveraging ML for Scheduling

We build a new resource scheduler – OSML, which differs from the previous schedulers in several ways. (1) It uses data-driven static ML models and reinforcement learning model to work collaboratively to perform scheduling. Model-A predicts the OAA and the RCliff for a specific LC service. Model-B balances the QoS and resource allocations among co-located LC services. Model-C is an online reinforcement learning model that dynamically shepherds the allocations. (2) We collect extensive real traces for widely deployed LC services, making using data-driven ML technologies practical in cloud systems and providing more accurate predictions as a result.
can be fungible [10], Model-B’s outputs include three policies, i.e., <cores, LLC ways>, <cores dominated, LLC ways> and <cores, LLC ways dominated>, respectively. The tuple items are the number of cores, LLC ways deprived and reallocated to others with the allowable QoS slowdown. The term “cores dominated” indicates the policy using more cores to trade the LLC ways, and vice versa. The allowable QoS slowdown is determined according to the user requirement or the LC services’ priority and controlled by the OSML’s central logic. We denote Model-B’s outputs as B-Points.

Model-B trades QoS for resources. For example, when an LC service (E) comes to a server that already has 4 co-located services, OSML enables Model-A’ to obtain <n+, m+>, which denotes at least n more cores and m more LLC ways should be provided to meet E’s QoS. Then, OSML enables Model-B and uses the allowable QoS slowdown as an input to infer B-Points for obtaining resources from other co-located services. B-Points include the “can be deprived” resources from E’s neighbors with the allowable QoS slowdown. Finally, OSML finds the best solution to match <n+, m+> with B-Points, which has a minimal impact on the co-located applications’ current allocation state. Detailed logic is in Algo._1. Besides, we design Model-B’ (a shadow of Model-B) to predict how much QoS slowdown will suffer if a certain amount of resources is deprived of a specific service. The structure of Model-B’ is similar to Model-B.

Model-B Training. For training Model-B and B’, we reduce the allocated resources for a specific LC service from its OAA by fine-grain approaches, as illustrated in Figure 4. The reduction has three angles, i.e., horizontal, oblique, and vertical, i.e., B-Points include <cores dominated, LLC ways>, <cores, LLC ways>, <cores, LLC ways dominated>, respectively. For each fine-grain resource reduction step, we collect the corresponding QoS slowdowns and then label them as less than (≤) 5%, 10%, 15%, and so on, respectively. Examples are illustrated in Figure 4, which shows the B-Points with the corresponding QoS slowdown. We collect the training data set for every LC service in Table 1. The data set contains 65,998,227 data tuples, covering 549,987 cases.

Model-B Function. We design a new loss function:

\[
L = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_1}{y_1 + c} \times (s_i - y_i) \right)^2,
\]

in which \(s_i\) is the prediction output value of Model-B, \(y_i\) is the labeled value in practice, and \(c\) is a constant that is infinitely close to zero. We multiply the difference between \(s_i\) and \(y_i\) by \(\frac{y_1}{y_1 + c}\) for avoiding the weights during backpropagation in the cases where \(y_i = 0\) and \(\frac{y_1}{y_1 + c} = 0\) caused by some non-existent cases (we label the non-existent cases as 0, i.e., \(y_i = 0\), indicating we don’t find a resource-QoS trading policy in the data collection process).

4.3. Model-C: Handling the Changes On the Fly

Model-C Description. Model-C corrects the resource under/over-provision and conducts online training. Figure 5 shows the Model-C in a nutshell. Model-C’s core component is an enhanced Deep Q-Network (DQN) [43], consisting of two neural networks, i.e., Policy Network and Target Network. The Policy and Target Network employ the 3-layer MLP, and each hidden layer has 30 neurons. Policy Network’s inputs consist of the parameters in Table 3, and the outputs are resource scheduling actions (e.g., reducing/increasing a specific number of cores or LLC ways) and the corresponding expectations (defined as Q/action). These actions numbered 0–48 are defined as Action_Fun: \(\{<m, n> | m \in [-3, 3], n \in [-3, 3]\}\), in which a positive m denotes allocating m more cores (i.e., add operation) for an application and a negative m means depriving m of cores (i.e., sub operation); n indicates the actions on LLC ways. Figure 5 illustrates Model-C’s logic. The scheduling action with the maximum expectation value (i.e., the action towards the best solution) will be selected in ① and executed in ②. In ③, Model-C will get the Reward value according to the Reward Function. Then, the tuple <Status, Action, Reward, Status’> will be saved in the Experience Pool in ④, which will be used during online training. The terms Status and Status’ denote system’s status described by the parameters in Table 3 before and after the Action is taken. Model-C can quickly have the ideal solutions in practice (about 2 or 3 actions). Please note that in ①, Model-C might randomly select an Action instead of the best Action with a 5% chance. By doing so, OSML avoids falling into a local optimum [43].

Model-C’s Reward Function. The reward function of Model-C is defined as follows:

If \(\text{Latency}_{t-1} > \text{Latency}_t\):
\[
R_t = \log(1 + \text{Latency}_{t-1} - \text{Latency}_t) - (\Delta \text{CoreNum} + \Delta \text{CacheWay})
\]

If \(\text{Latency}_{t-1} < \text{Latency}_t\):
\[
R_t = -\log(1 + \text{Latency}_t - \text{Latency}_{t-1}) - (\Delta \text{CoreNum} + \Delta \text{CacheWay})
\]

If \(\text{Latency}_{t-1} = \text{Latency}_t\):
\[
R_t = - (\Delta \text{CoreNum} + \Delta \text{CacheWay})
\]

where \(\text{Latency}_{t-1}\) and \(\text{Latency}_t\) denote the latency in previous and current status, respectively; \(\Delta \text{CoreNum}\) and \(\Delta \text{CacheWay}\) represent the changes in the number of cores and LLC ways, respectively. This function gives higher reward and expectation to Action that brings less resource usage and lower latency. Thus, Model-C can allocate appropriate resources. Algo._2 and 3 show the logic on using Model-C in detail.

Offline Training. A training data tuple includes Status, Status’, Action and Reward, which denote the current status of a LC service, the status after these actions are conducted (e.g., reduce several cores or allocate more LLC ways) and the reward calculated using the above functions, respectively.

To create the training data set for Model-C, we resort to the data set used in Model-A training. The process is as follows. Two tuples in Model-A training data set are selected to denote Status and Status’, and we further get the differences of the resource allocations between the two status (i.e., the actions that cause the status shifting). Then, we use the reward function to have the reward accordingly. These 4 values form
a specific tuple in Model-C training data set. In practice, as there are a large number of data tuples in Model-A training data set, it is impossible to try every pair of tuples in the data set, we only select two tuples from resource allocation policies that have less than or equal to 3 cores, or 3 LLC ways differences. Moreover, we also collect the training data in the cases where LLC sharing occurs among different LC services and save them in the Experience Pool. Using them, Model-C can have the knowledge on selecting actions in resource sharing cases. To sum up, we have 1,521,549,190 tuples in Model-C training data set.

**Online Training.** Model-C collects online traces. The training flow is in the right part of Figure 5. Model-C randomly selects some data tuples (200 by default) from the Experience Pool. For each tuple, Model-C uses the Policy Network to get the Action’s expectation value (i.e., Q(′Action)) with the Status. In Model-C, the target of the Action’s expectation value is the Reward observed plus the weighted best expectation value of the next status (i.e., Status’). As illustrated in Figure 5, Model-C uses the Target Network to have the expectation values of Status’ for the actions in Action_Function and then finds the best one, i.e., Max(Q(′Action’)). We design a new Loss Function based on MSE: (Reward + γMax(Q(′Action’)) − Q(′Action’))². It helps the Policy Network predict closer to the target. The Policy Network is updated during online training. The Target Network’s weights are synchronized periodically with the Policy Network’s weights. Doing so enables the Target Network to provide stable predictions for the best expectation value of Status’ within a predefined number of time steps, thus improving the stability of the training and prediction.

**4.4. Discussions on the design of ML Models**

(1) Why using MLPs. Table 4 characterizes the ML models used in OSML. We employ three-layered MLPs in Model-A and B, because they can fit continuous functions with an arbitrary precision given a sufficient number of neurons in each layer [67], and we can use extensive training data to improve the accuracy of MLPs for predicting OAAAs and RCliffs. Moreover, after offline training, using MLPs brings negligible run-time overheads to OSML. (2) Why do we have the three models? We divide the OSML’s scheduling logic into three parts, which the three models cover, respectively. Models work in different scheduling phases, and no single model can handle all cases. Model-A predicts the RCliffs and OAAAs; Model-B predicts the QoS variations and resource margins in co-location cases. DQN in Model-C learns online to shepherd the scheduling results from Model-A/B. They are necessary and work cooperatively to cover main scheduling cases. Moreover, they are easier to generalize than other approaches, e.g., a table lookup approach (Sec.6.4).

**5. OSML: System Design**

**5.1. The Central Control Logic**

The overview of the system design of OSML is in Figure 6. The central controller of OSML coordinates the ML models, manages the data/control flow, and reports the scheduling results to the upper scheduler. Figure 7 shows its overall control logic. More details are as below.

**Allocating Resources for LC services.** Algo._1 shows how OSML uses Model-A and B in practice. Figure 7 high-

---

**Table 4: The Summary of ML models in OSML.**

| ML | Model | Features | Model Size | Loss Function | Gradient Descent | Activation Function |
|----|-------|----------|------------|---------------|-----------------|-------------------|
| A  | MLP   | 9        | 144 KB     | Mean Square Error (MSE) | Adam Optimizer | ReLU             |
| A  | MLP   | 12       | 155 KB     |                |                 |                  |
| B  | MLP   | 13       | 110 KB     | Modified MSE   |                 |                  |
| B  | MLP   | 14       | 106 KB     |                |                 |                  |
| C  | DQN   | 8        | 141 KB     | Modified MSE   | RMSProp         | ReLU             |

---

**Figure 6: The overview of the system design of OSML.** C? Model-C uses DQN that depends on the start points. It starts with Model-A/B’s outputs to avoid exploring the whole (large) scheduling space. Without the approximate OAA provided by Model-A/B for many unseen cases, only using Model-C will incur more scheduling actions (overheads). (3) Generalization for Unseen apps and New servers. (i) We use “hold-out cross validation”, i.e., the training data (70% of the whole data set) excludes the testing data (30%) for each LC service. (ii) We train models with extensive representative traces from many types of LC services (e.g., memory/CPU intensive, random memory access, etc. [10,46]), fitting the correlation between architectural hints (IPC, cache miss, memory footprint, CPU/LLC utilization and MBL are more critical parameters for affecting models’ performance), OAA, and the QoS demands. For instance, the spearman correlation coefficient is 0.571, 0.499, and -0.457 between OAA and cache miss, MBL, and IPC, respectively. On different platforms or for new/unseen applications, these numbers might be varied; however, this correlation trend is not changed, enabling OSML to generalize to other situations. (iii) Using transfer learning, collecting new traces on a new server for several hours will make OSML work well for it (refer to Sec.6.4). (iv) OSML is a long-term project open to the community; we continue adding new traces collected from new applications and new servers to the data set for enhancing models’ performance for new cases.

---

**Figure 7: OSML’s central logic.**
lights its operations. For a newly coming LC service, the central controller calls Model-A via the interface `modelA_oaa_rcliff()` to get the OAA and RCliff. Suppose the current idle resources are not sufficient to satisfy the new LC service. In that case, OSML will enable Model-B through the interface `modelB_trade_qos_res()` to deprive some resources of other LC services with the allowable QoS slowdown (controlled by the upper-level scheduler) and then allocate them to the new one. In the depriving process for a specific LC service, OSML reduces its allocated resources and gets close to the RCliff, but it will not easily fall off the RCliff unless expressly permitted (refer to Algo._4).

**Dynamic Adjusting.** Figure 7 shows the dynamic adjusting of Algo._2 and 3, in which Model-C works as a dominant role. During the run time, OSML monitors each LC service’s QoS status for every second. If the QoS violation is detected, the central controller will enable Model-C through the interface `modelB_trade_qos_res()` to deprive some resources of other LC services with the allowable QoS slowdown (controlled by the upper-level scheduler) and then allocate them to the new one. In the depriving process for a specific LC service, OSML reduces its allocated resources and gets close to the RCliff, but it will not easily fall off the RCliff unless expressly permitted (refer to Algo._4).

Moreover, Algo._4 will enable resource sharing if the default scheduling is to do hard partitioning of cores/LLC ways), if all of the co-located LC services are close to their RCliff and the upper scheduler still wants to increase loads on this server. Model-A and B work cooperatively to accomplish this goal in Algo._4. In practice, to minimize the adverse effects, resource sharing usually happens between only two applications. Note that Algo._4 might incur the resource sharing over the RCliff, and thus may bring higher response latency for one or more LC services. OSML will report the potential QoS slowdown to the upper scheduler and ask for the decisions. If the slowdown is not allowed, the corresponding actions will not be conducted.

**Bandwidth Scheduling.** OSML partitions the overall bandwidth for each co-located LC service according to the ratio $\frac{BW_j}{\sum BW_i}$. $BW_i$ is a LC service’s OAA bandwidth requirement, which is obtained from the Model-A. Note that such scheduling needs MBA support [4,5] in CPU.

### 5.2. Implementation
OSML interacts with the upper scheduler or clients to obtain the latency of all requests and checks whether they have met their QoS targets. OSML monitors the run-time parameters for each co-located LC service using performance counters for every second (default). If the observation period is too short, other factors (e.g., cache data evicted from the cache hierarchy, context switch) may interfere with the sampling results. Moreover, we find OSML performs well with other
We design OSML that works cooperatively with OS (Figure 6). As the kernel space lacks the support of ML libraries, OSML lies in the user space and exchanges information with the OS kernel. OSML is implemented using python and C. It employs Intel CAT technology [4] to control the cache way allocations, and it supports dynamically adjusting. OSML uses Linux’s taskset and MBA [5] to allocate specific cores and bandwidth to an LC service. OSML captures the online performance parameters by using the pqos tool [4] and PMU [5]. The ML models are based on TensorFlow [6] with the version 2.0.4, and can be run on either CPU or GPU.

6 Evaluations

6.1. Methodology

We evaluate OSML on our platform in Table 2. Details on LC applications can be found in Table 1. The metrics involve the QoS (similar to [10], the QoS target of each application is the 99th percentile latency of the knee of the latency-RPS curve. Latency higher than the QoS target is a violation.); Effective Machine Utilization (EMU) [10] (the max aggregated load of all co-located LC services) – higher is better. We first evaluate the scenarios where LC services run at constant loads, and the loads are from 10% - 100%. Then, we explore workload churn. We inject applications with loads from 20%-100% of their respective max load. Furthermore, to evaluate the generalization of OSML, we employ some new/unseen applications that are not in Table 1 and the new platform in our experiments. If an allocation in which all applications meet their QoS cannot be found after 3 mins, we signal that the scheduler cannot deliver QoS for that configuration.

6.2. OSML Effectiveness

We compare OSML with the most related approaches in [10,46] based on the latest open-source version.

**PARTIES** [10]. It makes incremental adjustments in one-dimension resource at a time until QoS is satisfied – “trial and error” – for all of the applications. The core mechanism in [10] is like an FSM [60].

**CLITE** [46]. It conducts various allocation policies and samples each of them; it then feeds the sampling results – the QoS and run-time parameters for resources – to a Bayesian optimizer to predict the next scheduling policy.

**Unmanaged Allocation (baseline).** This policy doesn’t control the allocation policies on cores, LLC, and other shared resources for co-located LC services. This policy relies on the original OS schedulers.

**ORACLE.** We obtain these results by exhaustive offline sampling and find the best allocation policy. It indicates the ceiling that the schedulers try to achieve.

We show the effectiveness of OSML as follow.

1. OSML achieves a high EMU with shorter scheduling convergence time in most cases. Using ML models, OSML achieves OAA quickly and can efficiently handle cases with diverse loads. We tested 104 different loads for OSML, PARTIES and CLITE, respectively. Figure 8-a shows the distributions of the scheduling results for these 312 (104*3) cases.

![Figure 8](image-url)

Figure 8: (a) The performance distributions for OSML, PARTIES, and CLITE; 104 different loads are tested for every scheduler. (b) Violin plots of convergence time for loads in (a).

Every dot represents a scheduling case for a specific workload that contains several co-located LC services with diverse RPS. The x-axis shows the convergence time; the y-axis denotes the achieved EMU. Generally, OSML can achieve the same EMU with a shorter convergence time for a specific load. Figure 8-b shows the violin plots of convergence time for these loads. On average, OSML takes 20.9 seconds to converge, while PARTIES and CLITE take 32.7 and 46.3 seconds, respectively. OSML converges 1.56X and 2.22X faster than PARTIES and CLITE. OSML performs stably – the convergence time ranges from 5.3s (best case) to 80.6s (worst case). By contrast, the convergence time in PARTIES ranges from 5.5s to 111.1s, and CLITE is from 14.0s to 140.6s. OSML converges faster mainly because the start point in the scheduling space provided by Model-A is close to OAA. PARTIES and CLITE take a longer convergence time, indicating that they require high scheduling overheads in cloud environments.

In Cloud, jobs come and go frequently; thus, scheduling occurs frequently, and longer scheduling convergence time often leads to unstable/low QoS.

We further analyze how these schedulers work in detail. Figure 9-a/b/c show the actions used in OSML, PARTIES, and CLITE’s scheduling process for case A in Figure 8. This case includes Moses, Img-dnn, and Xapian with 40%, 60%, and 50% of their maximum loads. For this load, PARTIES and CLITE take 14.5 seconds, 72.6 seconds and 8.2 seconds to converge, respectively. Figure 9 highlights scheduling actions using solid red lines to represent increasing resources and blue dotted lines to denote reducing resources. Figure 9-a shows PARTIES takes 7 actions for scheduling cores and 1 action for cache ways. It schedules in a fine-grained way by increasing/decreasing one resource at a time. CLITE relies on the sampling points in the scheduling exploration space. Figure 9-b shows CLITE repeats sampling until the “expected improvement” in CLITE drops below a certain threshold. CLITE only performs five scheduling actions according to its latest open-source version; but it takes the longest convergence time (72.6 seconds). The underlying reason is that CLITE’s sampling/scheduling doesn’t have clear targets. In practice, the improper resource partitions/allocations during sampling lead to the accumulation of requests, and the requests cannot be handled due to resource under-provision. Therefore, it brings a significant increase in response latency. Moreover, due to the early termination of CLITE’s scheduling process, CLITE cannot schedule resources to handle QoS violations in a timely manner, leading to a long convergence time. Figure 9-c shows OSML achieves OAA for each LC service with 5 actions. Compared with prior schedulers, OSML has clear aims and
schedules multiple resources simultaneously to achieve them. It has the shortest convergence time – 8.2 seconds.

Moreover, as the scheduling is fast, OSML often supports more loads. Figure 10 shows the OSML’s results on scheduling the three LC services – Moses, Img-dnn, and Xapian. For a specific scheduling phase, by using ML to achieve OAA, OSML supports 10–50% higher percentage of loads than PARTIES and CLITE (e.g., highlighted cells in Figure 10-d). All schedulers perform better than the Unmanaged (Figure 10-a), as they reduce the resource contentions.

(2) Compared with PARTIES and CLITE, OSML uses fewer resources to support identical loads to meet the QoS targets. As illustrated in Figure 9-a, PARTIES partitions the LLC ways and cores equally for each LC service at the beginning; once it meets the QoS target (using 8 actions), it stops. Thus, PARTIES drops the opportunities to explore alternative better solutions (i.e., using fewer cores or cache ways to meet identical QoS targets). PARTIES allocates all cores and LLC ways finally. CLITE also uses all cores and cache ways shown in Figure 9-b. By contrast, OSML schedules according to applications’ resource requirements instead of using all resources. Figure 9-c shows that using Model-A, OSML achieves each LC service’s OAA (the optimal solution) after 5 actions. OSML detects and reclams over-provided resources using Model-C. For example, the last action in Figure 9-c reclaims 3 cores and 2 LLC ways from Xapian. Finally, OSML saves 3 cores and 9 LLC ways. As OSML is designed for LC services that are executed for a long period, saving resources means saving budgets for cloud providers.

(3) Using ML models, OSML provides solutions for sharing some cores and LLC ways among LC services, therefore supporting higher loads. PARTIES and CLITE don’t show resource sharing in the original design. Using Algo._4, OSML lists some potential resource sharing solutions, and then enables Model-B' to predict the QoS slowdown for each case. The sharing solution with a relatively lower QoS slowdown is selected. More details refer to Figure 7. Figure 9-d shows how OSML shares resources for the highlighted case B in Figure 10-d. OSML enables Model-C to add resources for Moses in Algo._2 and uses Algo._4 to share 2 CPU cores with Xapian. Finally, the QoS is met. By enabling resource sharing, OSML can support higher loads than PARTIES and CLITE, and can even be close to ORACLE in Figure 10-e. If not OSML, however, the “trial and error” approach has to try to share core/cache way in a fine-grain way among applications, and then observes the real-time latency for making a further
We evaluate how OSML behaves with dynamically changing loads. Each LC service’s QoS is normalized to the solely running case. As illustrated in Figure 11-a, in the beginning, Moses with 60% of max load arrives; then Sphinx with 20% of max load and Img-dnn with 60% of max load arrive. We observe their response latency increases caused by the resource contentions among them. In Figure 11-b, PARTIES aids the QoS violations step by step. During the scheduling, Moses always has high latency until it ends at 80 seconds. CLITE’s scheduling relies on sampling and Bayesian optimizer. CLITE starts scheduling at time point 16, where all the three services arrive. At 32s, CLITE obtains the scheduling solution for these LC services after five sampling steps. However, it does not meet Moses and Img-dnn’s QoS targets. In Figure 11-c, Moses and Img-dnn still have high latency. By contrast, with Model-A’s OAA predictions and Model-C’s online scheduling, OSML quickly provides better scheduling solutions at time point 48 for all three services. During the identical scheduling phase (e.g., the time point 16 to 80), we can observe the lowest overall normalized latency in Figure 11-d. Moreover, Figure 11-e and f illustrate OSML’s scheduling actions for achieving ideal solutions. In short, within a few scheduling actions (scheduling overheads) that schedule multiple resources, OSML quickly meets the QoS targets.

From 180 to 228, we increase the load for Img-dnn as illustrated in Figure 11-a. OSML meets Img-dnn’s changing demands by using Model-C. PARTIES does not reflect quickly for this change, and it works for other services. Thus, as illustrated in Figure 11-b, the QoS violation is not aided until 244s, when Img-dnn’s load decreases. For CLITE, it has to sample each time when the load changes. But during the sampling, a specific service might not have sufficient resources to handle the requests; thus, the requests are accumulated, leading to QoS fluctuations/violations during the scheduling (Figure 11-c). Figure 12 highlights the scheduling actions for Img-dnn from 180 to 228. During this phase, PARTIES does not add resources for Img-dnn; but it add more resources for Specjbb and Xapian as they are with higher latency. Img-dnn’s response latency keeps increasing. CLITE samples several scheduling policies in the scheduling space, but does not converge and thus incurs QoS fluctuations. By contrast, OSML’s Model-C achieves Img-dnn’s OAA using fewer scheduling actions. Moreover, as mentioned before, OSML saves resources and thus it can serve more workloads. For example, as shown in Figure 11, Mysql (an unseen workload in training) comes at time point 180; OSML allocates the saved cores to it without sharing or depriving other LC services of resources.

### 6.4. Evaluations for New/Unseen Apps and New Platforms Generalization.

Based on our comprehensive data set, the ML models are well trained; and it is possible to skip profiling for new/unseen applications. The sensitivity will be at most 4-core error (Table 5). We evaluate OSML for unseen applications that are not in Table 1, e.g., Silo [62], Shore [62], and
The workloads contain at least one new application. The efforts in [22, 56, 34] leverage ML to optimize computer systems. We freeze the first hidden layer of the MLPs; we retrain the models. The studies in [9, 39] use ML to manage various workloads. The work in [55] employs DNN to optimize the buffer size for the database systems. CALOREE [41] can learn key control parameters to meet latency requirements with minimal energy in complex environments. The studies in [26, 31, 58, 59] optimize the OS components with learned rules or propose insights on designing new learned OS.

For new platforms, we use fine-tuning in transfer learning (TL). We freeze the first hidden layer of the MLPs; we retrain the models. The last two hidden layers and the output layer using the traces collected on two new platforms (w/ CPU Xeon Gold 6240M and E5-2630 v4, respectively). For each LC service, based on our data set, collecting new traces on a new platform for several hours will be sufficient (covering the more allocation cases, the better). The time consumption will be shorter if using multiple machines in parallel. Table 5 shows the average values of ML models' quality. The new models' prediction errors are slightly higher than the previous models on the original platforms, but OSML still handles them well. By contrast, if we use a table lookup approach instead, we have to use additional memory to store the data tuples, e.g., 60GB will be wasted for the current data set to replace Model-A. More importantly, it is difficult to generalize a table look-up approach for new/unseen applications or platforms, as their traces and the corresponding OAA don’t exist in the current data set.

Overheads. OSML takes 0.2s for each time (0.01s for ML model and 0.19s for online monitoring). As our models are light-weighted (OSML uses only one core), running them on CPU and GPU has a similar overhead. If models are on GPU, it takes an extra 0.03s for receiving results from GPU. OSML doesn’t monopolize GPU. Generally, the overhead doesn’t hinder the overall performance. In the cloud, applications' behaviors may change every second due to the diversity of user demands. Thus, OSML plays a critical role during the entire run time. For training time, using our current data set, it takes 3.3 mins, 5 mins, and 8.3 hours to train Model-A, B, and C for one epoch, respectively. One epoch means that all training samples in the data set are used to train once. We train models for ten epochs. Training can be accelerated using the popular Multi-GPU training technology - using multiple GPUs simultaneously to train one model. Doing so is practical in datacenters, and training time will not impede practice.

7. Related Work and Our Novelty

ML for System Optimizations. The work in [55] employs DNN to optimize the buffer size for the database systems. The efforts in [22, 56, 34] leverage ML to optimize computer architecture or resource management in the network to meet various workloads. The studies in [9, 39] use ML to manage on-chip hardware resources. CALOREE [41] can learn key control parameters to meet latency requirements with minimal energy in complex environments. The studies in [26, 31, 58, 59] optimize the OS components with learned rules or propose insights on designing new learned OS. In OSML, we design an intelligent multi-model collaborative learning approach, providing better co-location solutions to meet QoS targets for LC services faster than the latest work stably.

ML for Scheduling. Decima [35] designs cluster-level data processing job scheduling using RL. Resource Central [12] builds a system that contains the historical resource utilization information of the workloads used in Azure and employs ML to predict resource management for VMs. [40] uses RL to predict which subsets of operations in a TensorFlow graph should run on the available devices. Paragon [14] classifies and learns workload interference. Quasar [15] determines jobs’ resource preferences on clusters. Sinan [74] uses ML models to determine the performance dependencies between microservices in clusters. They are cluster schedulers [14, 15, 74]. By contrast, OSML deeply studies scheduling in co-location cases. Selecta [72] predicts near-optimal configurations of computing and storage resources for analytics workloads based on profiling data. CLITE [46] uses Bayesian optimization for scheduling on-node resources. The work in [48] applies ML to predict the end-to-end tail latency of LC service workflows. Twig [63] uses RL to characterize tail latency for energy-efficient task management. CuttleSys [76] leverages data mining to identify suitable core and cache configurations for co-scheduled applications. For complicated co-location cases on a specific cloud server, using the fewest scheduling actions on average compared with the latest studies, OSML can avoid RCliff and achieve the ideal allocations (OAAs) for multiple interactive resources simultaneously for LC services. Moreover, OSML performs well in generalization.

Resource Partitioning. PARTIES [10] partitions cache, main memory, I/O, network, disk bandwidth, etc., to provide QoS for co-located services. The studies in [17, 28, 57, 71] design some new LLC partitioning/sharing policies. The efforts in [23, 27, 44, 45, 73] show that considering cooperative partitioning on LLC, memory banks and channels outperforms one-level memory partitioning. However, the cooperative partitioning policies need to be carefully designed [29, 30, 37], and [16, 32] show the heuristic resource scheduling approach could be ineffective in many QoS-constrained cases. [7, 11] study the “performance cliff” on cache for SPEC CPU 2006 applications and Memcached. Caladan [75] doesn’t involve cache optimizations, and core/cache cliffs cannot be avoided, causing QoS fluctuations in some cases. By contrast, OSML is the first work that profoundly explores cache cliff and core cliff simultaneously (i.e., RCliff) for many widely used LC services in co-location cases. OSML is a representative work using ML to guide the multiple resources partitioning in co-location cases; OSML is cost-effective in new cloud environments.

8. Conclusion

We present OSML, an resource scheduler for LC services. OSML employs ML to preserve QoS for the co-scheduled services. We should OSML performs well. We also learn that straightforwardly using a simple ML model might not handle all of the processes during the scheduling. Therefore, using multiple ML models cooperatively in a pipe-lined way can be an ideal approach. More importantly, we advocate the new solution, i.e., leveraging ML to enhance resource
scheduling, could have an immense potential for OS design.
In a world where co-location and sharing are a fundamental
reality, our solution should grow in importance and benefits
our community.

References

[1] “How 1s could cost amazone $1.6 billion in sales.”
https://www.fastcompany.com/1825005/how-one-second-could-
cost-amazon-16-billion-sales

[2] “Microservices workshop: Why, what, and how to get there,”
http://www.slideshare.net/adriancockcroft/microservices-workshop-
craft-conference

[3] “State of the Cloud Report,” http://www.rightscale.com/op/state-of-the-
cloud. Accessed: 2019-01-28

[4] “Improving real-time performance by utilizing cache allocation technology,”
https://www.intel.com/content/dam/www/public/us/en/documents/white-
papers/cache-allocation-technology-white-paper.pdf; Intel Corporation,
April, 2015

[5] “Intel 64 and IA-32 Architectures Software Developer’s Manual,”
https://software.intel.com/en-us/articles/intel-sdm, Intel Corporation,
October, 2016

[6] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis,
Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving,
Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry
Moore, Derek G. Murray, Benoit Steiner, Paulsystems, Vraj Vasudevan,
Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, “Tensor-
Flow: A System for Large-Scale Machine Learning,” in OSDI, 2016

[7] Nathan Beckmann, Daniel Sanchez, “Talus: A Simple Way to Remove
Cliffs in Cache Performance,” in HPCA, 2015

[8] Daniel S. Berger, Benjamin Berg, Timothy Zhu, Siddhartha Sen, Mor
Harchol-Balter, “RobinHood: Tail Latency Aware Caching – Dynamic
Reallocation from Cache-Rich to Cache-Poor,” in OSDI, 2018

[9] Ramazan Börgen, Engin Ipek, Jose F. Martinez, “Coordinated Man-
age of Multiple Interacting Resources in Chip Multiprocessors: A Machine Learning Approach,” in Micro, 2008

[10] Shuang Chen, Christina Delimitrou, Jose F. Martinez, “PARTIES: QoS-
Aware Resource Partitioning for Multiple Interactive Services,” in ASP-
LOS, 2019

[11] Asaf Cidon, Assaf Eisenmann, Mohammad Alizadeh, Sachin Katti,
“Cliffhanger: Scaling Performance Cliffs in Web Memory Caches,”
in NSDI, 2016

[12] Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus
Fontoura, Riccardo Bianchini, “Resource Central: Understanding and
Predicting Workloads for Improved Resource Management in Large
Cloud Platforms,” in SOSP, 2017

[13] Jeff Dean, David A. Patterson, Cliff Young, “A New Golden Age in
Computer Architecture: Empowering the Machine-Learning Revolu-
tion,” in IEEE Micro 38 (2): 21-29 (2018)

[14] Christina Delimitrou, Christos Kozyrakis, “QoS-Aware Scheduling in
Heterogenous Datacenters with Paragon,” in ACM TOCS, 2013

[15] Christina Delimitrou, Christos Kozyrakis, “QoSAware Clustering in
QoS-Aware Clustering,” in ASPLOS, 2014

[16] Yi Ding, Nikita Mishra, Henry Hoffmann, “Generative and Multi-phase
Learning for Computer Systems Optimization,” in ISCA, 2019

[17] Nosayba El-Sayed, Anurag Mukkara, Po-An Tsai, Harshad Kasture,
Xiaosong Ma, Daniel Sanchez, “KPart: A hybrid Cache Partitioning-
Sharing Technique for Commodity Multicores,” in HPCA, 2018

[18] Yu Gan and Christina Delimitrou, “The Architectural Implications of
Cloud Microservices,” in IEEE Computer Architecture Letters, 2018

[19] Yu Gan, Yanqi Zhang, Kelvin Hu, Dailun Cheng, Yuan He, Meghna
Pancholi, Christina Delimitrou, “Leveraging Deep Learning to Improve
Performance Predictability in Cloud Microservices with Seer,” in ACM
SIGOPS Operating Systems Review, 2019

[20] Mark D. Hill, Michael R. Marty, “Amdahl’s Law in the Multicore Era,”
in IEEE Computers, 2008

[21] Kurt Hornik, “Approximation Capabilities of Multilayer Feedforward
Networks,” in Neural Networks, 1991

[22] Engin Ipek, Onur Mutlu, José F. Martínez, Rich Caruana, “Self-
Optimizing Memory Controllers: A Reinforcement Learning Approach,”
in ISCA, 2008

[23] Min Kyu Jeong, Doe Hyun Yoon, Dam Sunwoo, Michael Sullivan,
Ikwon Lee, Mattner Erez, “Balancing DRAM Locality and Parallelism
in Shared Memory CMP Systems,” in HPCA, 2012

[24] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav
Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhattacharjee, Nan Boden,
Al Borchers, Rick Boyle, Pierre-lec Cantin, Clifford Chao, Chris Clark,
Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara
Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert
Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt,
Dan Hurt, Julian Ibarz, Aaron Jaffrey, Alex Jaworski, Alexander Ka-
plan, Harshit Khatian, Daniel Killebrew, Andy Koch, Naveen Kumar,
Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary,
Zhuoyuan Liu, Kyle Lucks, Alan Lundin, Gordon MacKean, Adri-
ana Maggiore, Mairé Mahony, Kieran Miller, Rahul Nagarajan, Ravi
Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick,
Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir
Salek, Emad Samadlani, Chris Severn, Gregory Sizikov, Matthew Smel-
ham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory
Thorson, Bo Tian, Horia Toma, Eck Tuttle, Vijay Vasudevan, Richard
Walter, Walter Wang, Eric Wilcox, Doe Hyun Yoon, “In-Datacenter
Performance Analysis of a Tensor Processing Unit,” in ISCA, 2017

[25] Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, “ImageNet Classi-
fication with Deep Convolutional Neural Networks,” in Advances in
neural information processing systems, 2012

[26] Yanjing Li, Onur Mutlu, Sushabhis Mitra, “Operating System Schedul-
ing for Efficient Online Self-Test in Robust Systems,” in ICCAD, 2009

[27] Lei Liu, Zehan Cui, Mingjie Xiong, Chengyong Wu, “A Software Mem-
ory Partition Approach for Eliminating Bank-level Interface in Mul-
ticore Systems,” in PACT, 2012

[28] Jiang Lin, Qingda Lu, Xiaoning Ding, Zhao Zhang, Xiaodong Zhang, P.
Sadayappan, “Gaining insights into multicore cache partitioning:
bridging the gap between simulation and real systems,” in HPCA, 2008

[29] Fang Liu, Yan Solihin, “Studying the Impact of Hardware Prefetching
and Bandwidth Partitioning in Chip-Multiprocessors,” in Sigmetrics,
2011

[30] Seung-Hwan Lim, Jae-Seok Huh, YoujAIe Kim, Galen M. Shippman,
Chita R. Das, “D-Factor: A Quantitative Model of Application Slow-
Down in Multi-Resource Shared Systems,” in Sigmetrics, 2012

[31] Lei Liu, Yong Li, Chen Ding, Hao Yang, Chengyong Wu, “Rethinking
Memory Management in Modern Operating System: Horizontal,
Vertical or Random?” in IEEE Trans. on Computers, 2016

[32] David Lo, Liqun Cheng, Rama Govindaraju, Parthasarathy Ran-
ganathan, Christos Kozyrakis, “Heracles: Improving Resource Effi-
ciency at Scale,” in ISCA, 2015

[33] Lei Liu, Shengjie Yang, Lu Peng, Xinyu Li, “Hierarchical Hybrid
Memory Management in OS for Tiered Memory Systems,” in IEEE
Trans. on Parallel and Distributed Systems, 2019

[34] Hongzi Mao, Mohammad Alizadeh, Ishai Menache, Srikant Kandula,
“Resource Management with Deep Reinforcement Learning,” in HotNet-
XV, 2016

[35] Hongzi Mao, Malte Schwarzkopf, Shaileshh B. Venkatakrishnan, Zili
Meng, Mohammad Alizadeh, “Learning Scheduling Algorithms for
Data Processing Clusters,” in SHOCOMM, 2019

[36] Yashwant Marathe, Nagendra Gulur, Jee Ho Ryoo, Shuang Song, and
Liyi K. John, “CSALT: Context Switch Aware Large TLB,” in HPCA,
2012

[37] Alexander Martinez, Egin Ipek, “Dynamic multicore resource manage-
ment: A machine learning approach,” in IEEE Micro 29 (5):8-17 (2009)

[38] Jason Mars, Lingjia Tang, Robert Hundt, Kevin Skadron, Mary Lou
Soffa, “Directly Characterizing Workloads for Improved Resource Man-
agement in Large and Bandwidth Partitioning in Chip-Multiprocessors,” in Sigmetrics, 2011
