

LFOC: A Lightweight Fairness-Oriented Cache Clustering Policy for Commodity Multicores

Adrian Garcia-Garcia
Complutense University of Madrid
adriagar@ucm.es

Fernando Castro
Complutense University of Madrid
fcastror@ucm.es

Juan Carlos Saez
Complutense University of Madrid
jcsaezal@ucm.es

Manuel Prieto-Matias
Complutense University of Madrid
mpmatias@ucm.es

ABSTRACT

Multicore processors constitute the main architecture choice for modern computing systems in different market segments. Despite their benefits, the contention that naturally appears when multiple applications compete for the use of shared resources among cores, such as the last-level cache (LLC), may lead to substantial performance degradation. This may have a negative impact on key system aspects such as throughput and fairness. Assigning the various applications in the workload to separate LLC partitions with possibly different sizes, has been proven effective to mitigate shared-resource contention effects.

In this article we propose LFOC, a clustering-based cache partitioning scheme that strives to deliver fairness while providing acceptable system throughput. LFOC leverages the Intel Cache Allocation Technology (CAT), which enables the system software to divide the LLC into different partitions. To accomplish its goals, LFOC tries to mimic the behavior of the optimal cache-clustering solution, which we could approximate by means of a simulator in different scenarios. To this end, LFOC effectively separates streaming aggressor programs from cache sensitive applications, which are then assigned to separate cache partitions.

We implemented LFOC in the Linux kernel and evaluated it on a real system featuring an Intel Skylake processor, where we compare its effectiveness to that of two state-of-the-art policies that optimize fairness and throughput, respectively. Our experimental analysis reveals that LFOC is able to bring a higher reduction in unfairness by leveraging a lightweight algorithm suitable for adoption in a real OS.

CCS CONCEPTS

- **Computer systems organization** → **Multicore architectures;**
- **Software and its engineering** → **Scheduling.**

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICPP '19, August 05–08, 2019, Kyoto, Japan

© 2019 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

KEYWORDS

Multicore processors, cache partitioning, clustering, fairness, Intel Cache Allocation Technology, Linux kernel.

ACM Reference Format:

Adrian Garcia-Garcia, Juan Carlos Saez, Fernando Castro, and Manuel Prieto-Matias. 2019. LFOC: A Lightweight Fairness-Oriented Cache Clustering Policy for Commodity Multicores. In *Proceedings of ICPP '19: International Conference on Parallel Processing (ICPP '19)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Today, chip multicore processors (CMPs) constitute the dominant architecture choice for modern general-purpose computing systems and will likely continue to be dominant in the near future. Despite their benefits, these processors pose a number of challenges to the system software. One of the biggest challenges is how to mitigate the effects that come from contention on shared resources [35]. Contention occurs due to the fact that cores in a CMP are not truly independent processing units but instead typically share a last-level cache (LLC) and other memory-related resources with the remaining cores, such as a DRAM controller and a memory bus or interconnection network [2, 8]. Applications running simultaneously on the various cores may compete with each other for these shared resources, which could degrade their performance unevenly [8, 23].

Partitioning of the shared LLC (i.e., assigning a separate cache partition with a certain size to each application in a workload) has been proven effective to mitigate shared-resource contention effects [3, 14, 16, 19, 23, 30]. Recently, cache-partitioning hardware support has been adopted in commodity Intel processors via the Intel Cache Allocation Technology (CAT) [18], which enables the system software to assign a certain number of cache ways to each application. Several resource management schemes that leverage this technology have been recently proposed to accomplish different objectives such as system throughput optimization [3], delivering fairness [23], or to improve client satisfaction in virtual environments [7].

Our work primarily explores how to leverage Intel CAT, at the OS level, to deliver system-wide fairness, which contributes to reducing a number of undesirable effects on the system. For example, shared-resource contention may cause an application's completion time to differ significantly across executions, depending on its co-runners in the workload [6, 35]. Moreover, equal-priority applications may

not experience the same performance degradation when running together relative to the performance observed when each application runs alone on the CMP [2, 17]. These issues make priority-based scheduling policies ineffective [2], reduce performance predictability [32] and may lead to wrong billings in commercial cloud-like computing services [6], where users are charged for CPU hours. Notably, unfairness also leads to uneven progress of the various threads in HPC multithreaded applications [26], which seriously limits scalability.

To deliver fairness while providing acceptable system throughput, we propose LFOC, a Lightweight Fairness-Oriented Cache-clustering OS-level scheme. By using Intel CAT hardware support, LFOC dynamically creates a number of last-level-cache partitions (*clusters*) according to the features of the workload, and maps applications to different clusters based on their degree of cache sensitivity and contentiousness.

The main contributions of our work are as follows:

- In order to guide the design of LFOC, we approximated –by means of a parallel simulator– the solution to the cache-clustering problem that optimizes fairness for different workload scenarios. The exhaustive analysis of the optimal solution reveals that the key to enforce fairness lies in identifying contentious cache-insensitive (aka *streaming*) applications and confine them to a reduced set of small cache partitions, so as to devote the vast majority of space in the LLC to cache-sensitive applications.
- Based on the insights provided by the previous analysis, we proceeded to design LFOC, which attempts to approximate the cache clustering enforced by the optimal solution. Our approach continuously monitors applications’ runtime metrics with hardware performance counters and classifies applications into different categories based on cache behavior. The collected performance information is used as input to an efficient clustering algorithm. LFOC leverages a lightweight online mechanism to approximate the degree of cache sensitivity of an application that avoids costly periodic monitoring operations (i.e. measuring application performance for different cache sizes at runtime) used by other approaches [3], whenever possible.
- We implemented LFOC in the Linux kernel and evaluated it on a real system featuring an Intel Skylake processor. In our experiments, we compare its effectiveness to that of two previously proposed policies –Dunn [23] and KPart[3]–, which optimize fairness and throughput, respectively. Our analysis reveals that LFOC is able to deliver up to a 20.5% reduction in unfairness (9% on average) relative to Dunn (fairness-oriented clustering), and delivers higher throughput and fairness than every analyzed scheme for the vast majority of the workload scenarios considered.

The remainder of the paper is organized as follows. Section 2 presents background on cache partitioning and also discusses related work. Section 3 presents the analysis of the optimal solution that motivates our proposal. Section 4 outlines the design and inner workings of LFOC. Section 5 covers the experimental evaluation, and Section 6 concludes the paper.

2 BACKGROUND AND RELATED WORK

In this section we first describe the metrics we considered to assess the degree of fairness and throughput provided by cache partitioning strategies. Then, we formally introduce the notion of cache partitioning and cache clustering and present some related issues. Finally, we discuss related work.

2.1 Metrics

To measure the performance degradation of an individual application in a multi-program workload we consider the *Slowdown* metric, defined as follows:

$$Slowdown_{app} = \frac{CT_{part,app}}{CT_{alone,app}} \quad (1)$$

where $CT_{part,app}$ denotes the completion time of application *app* when it runs sharing the system under a given cache-partitioning scheme, and $CT_{alone,app}$ is the completion time of the application when running alone on the CMP system.

The slowdown of a single-threaded application can be also defined in terms of the average number of instructions per cycle observed when it runs alone with all cache space available ($IPC_{alone,app}$) and that achieved when it runs with other applications in the workload under a certain cache-partitioning scheme ($IPC_{part,app}$):

$$Slowdown_{app} = IPC_{alone,app} / IPC_{part,app} \quad (2)$$

Previous research on fairness for multicore systems [2, 23] defines a scheme as fair if equal-priority applications in a workload suffer the same slowdown as a result of sharing the system. To cope with this notion of fairness, we employ the *unfairness* metric, which has been extensively used in previous work [2, 8, 20, 28]. For a workload consisting of n applications, this metric (lower-is-better) is defined as follows:

$$Unfairness = \frac{MAX(Slowdown_1, \dots, Slowdown_n)}{MIN(Slowdown_1, \dots, Slowdown_n)} \quad (3)$$

Notably, according to the definition of the unfairness metric, we could improve its value just by slowing down certain applications to achieve similar but potentially high slowdown figures. Clearly, this is unacceptable, as it could come at the expense of high throughput degradation, which our proposal also tries to reduce. Therefore, the value of the unfairness metric must be always reported along with system throughput figures, as we do in this article. Specifically, to quantify throughput, we used the *System ThroughPut* (STP) metric [5, 23], –also referred to as *Weighted Speedup* in [3]–, defined as follows:

$$STP = \sum_{i=1}^n \left(\frac{CT_{part,i}}{CT_{alone,i}} \right) = \sum_{i=1}^n \left(\frac{1}{Slowdown_i} \right) \quad (4)$$

2.2 Cache partitioning vs. Cache clustering

Two major strategies exist to distribute cache space among applications in a shared LLC that supports way-partitioning: *cache partitioning* and *cache clustering*.

Cache-partitioning entails assigning a separate cache partition with a certain size (a specific number of ways) to each application in a workload. Let A be a workload consisting of n applications $\{a_1, a_2, \dots, a_n\}$ and let S be a system that features a k -way last-level

cache with $k \geq n$. A feasible cache partitioning of the LLC for A on S is formally defined as a set $\{w_1, w_2, \dots, w_n\}$ (with $\sum_{i=1}^n w_i = k$) where w_i denotes the number of ways assigned to application a_i ($1 \leq w_i \leq k - n + 1$).

In recent years, many cache-partitioning proposals [1, 14, 19, 30] have been proposed to target different optimization objectives, such as maximizing throughput, reducing energy consumption or improving fairness. These proposals are equipped with heuristic approximate algorithms specifically tailored to the target objective they pursue. In general, determining the solution to the optimal cache partitioning problem for a certain optimization objective is known to be NP-hard [14], so determining the best solution via an extensive exploration of the (vast) search space is largely impractical. For example, determining the optimal cache-partitioning solution for an 8-application workload ($n = 8$) on a platform with an 11-way LLC, requires the exploration of 120 solutions; the number of possible options to consider for the same workload size on a 20-way platform rises up to more than 50K.

When the number of applications exceeds the number of cache ways to share ($n > k$), partitioning the LLC is unfeasible. Two or more applications must share at least one cache way. Previous work has pointed out that even when $n \leq k$, the coarse granularity of partitions (in the order of MBs) available in modern processors featuring Intel CAT hardware extensions makes strict cache partitioning (no way sharing among applications) inappropriate in some cases [3, 23]. Due to the finer grained cache distribution that naturally results from sharing a set of cache ways between two or more applications, we may observe better performance (and sometimes better fairness) when allowing shared ways among applications (aka. *Cache Clustering*) than by means of way partitioning [1, 3].

Formally, we define a *Cache Clustering* as follows. Let A be a workload consisting of n applications $\{a_1, a_2, \dots, a_n\}$, and let S be a system that features a k -way shared cache. A cache clustering for A on S is defined as the set $T = \{C_1, C_2, \dots, C_m\}$ and the associated set W of assigned ways for each C_i in T , $W = \{w_1, w_2, \dots, w_m\}$ where each C_i is a disjoint subset of A ($C_i \subseteq A$), subject to the following restrictions (i) $1 \leq m \leq \min(n, k)$, (ii) $C_1 \cup C_2 \cup \dots \cup C_m = A$, (iii) $\forall i, j, 1 \leq i, j \leq m \wedge j > i, C_i \cap C_j = \emptyset$ and (iv) ($1 \leq w_j \leq k - m + 1$) $\wedge \sum_{i=1}^m w_i = k$. For simplicity, we will refer to each item in T as a *cluster* (i.e., group of applications). According to the earlier definition, each cluster has a certain number of cache ways assigned to it, as indicated by the W set (e.g. C_1 has w_1 ways assigned to it), which constitutes one of the possible ways to distribute the available cache ways across clusters.

As in the case of the cache partitioning, several cache clustering approaches have been recently proposed [3, 7, 23] to pursue different optimization objectives by means of approximate algorithms. Notably, by considering the size of the search space, finding the optimal cache clustering solution for a certain objective constitutes even a harder problem to solve compared to optimal cache partitioning. Specifically, to determine the optimal solution via extensive exploration of the search space in the former case, for every possible clustering of the A set –with $\min(n, k)$ items at the most–, we have to determine the distribution of ways among clusters that optimizes a given objective. It is worth noting that the number of possible solutions grows exponentially with n and k . So, for example, on

a system featuring a 20-way LLC the number of different cache clustering options for an 8-application workload is roughly 9M, whereas for an 11-application workload more than 5000M possible solutions exist.

2.3 Related Work

Many researchers have attempted to mitigate the contention problem in the LLC via software and hardware techniques [3, 12, 14, 21, 23, 34, 36]. A large body of work has addressed this problem via cache-partitioning or cache-clustering approaches equipped with approximate algorithms [3, 16, 19, 30]. A recent survey [14] discusses the most effective solutions available to target various optimization objectives.

Cache partitions can be created on systems with specific hardware support (such as Intel CAT) or by means of software-based solutions, most of which rely on page-coloring [22, 24, 29, 33]. Page-coloring can be applied to off-the-shelf multicore platforms [31], but is known to be subject to a number of limitations, which can be overcome with hardware-based cache partitioning support [23]. Among the different hardware alternatives, the main differences essentially lie on how to manage the number of ways for the different applications: some proposals are based on the cache replacement policy [11, 13, 27] while others use set sampling and duplicate cache tags to guide cache partitioning [19, 25]. In this work we propose an OS-level (also extensible to the virtual machine monitor) cache-clustering scheme that leverages hardware-based way-partitioning.

In the remainder of this section we discuss the cache-partitioning and cache-clustering policies closer to our LFOC approach.

2.3.1 Cache partitioning proposals. UCP [19] is a cache-partitioning scheme that aims to improve system throughput by minimizing the total number of misses incurred by all applications in the workload on the shared last-level cache. UCP does not attempt to determine the optimal solution but instead employs an approximate algorithm referred to as *lookahead* [19], which uses as input the MPKI table of each application. This table stores the application’s MPKI (LLC Misses Per 1K Instructions) value for any possible cache size (i.e., number of assigned cache ways under way-partitioning). In the original proposal [19], UCP relies on hardware extensions to determine per-application MPKI tables at runtime. Unfortunately, more than a decade later of the original proposal, these hardware extensions have not yet been adopted in commercial platforms. Our LFOC approach relies on the *lookahead* algorithm to distribute the vast majority of the space in the LLC among cache-sensitive applications, by using the per-application slowdown tables (i.e., slowdown for different number of ways relative to using the total way count) as input to the algorithm instead of MPKI tables; this is known to provide a fairer cache distribution [9]. The cache-partitioning algorithm proposed by Yu and Petrov [30] strives to reduce system bandwidth pressure. To this end, it partitions the LLC so as to minimize the total bandwidth. In doing so, the algorithm relies on bandwidth consumption measurements with different cache sizes gathered offline for the various applications. Note that in our evaluation we just consider those partitioning approaches that do not require offline collected data to function.

2.3.2 Cache clustering proposals. More recently, different cache-clustering algorithms have been proposed [3, 23] as an alternative to strict cache-partitioning. The KPart [3] scheme constitutes a cache clustering approach specifically designed for throughput optimization. KPart implements an iterative algorithm that creates and merges application clusters via hierarchical clustering. To decide which clusters must be merged in each iteration of the loop, and how to distribute the available ways among clusters (inter-cluster way-partitioning), the scheme leverages the distance metric proposed in [16] as well as UCP’s *lookahead* algorithm [19]. The application of *lookahead* and the evaluation of the distance metric relies on the ability to determine MPKI tables and IPC tables (i.e. number of Instructions Per Cycle for different cache sizes) online for each application. As explained in Sec. 4, LFOC requires to gather a smaller amount of performance information than KPart, and avoids to perform costly cache way sweeps periodically, thus effectively reducing the overhead.

In [23], Selfa et al. propose the Dunn cache-partitioning policy, designed to improve fairness. This approach groups applications into clusters by applying the *k-means* clustering algorithm [10], and using the fraction of core stalls caused by L2 cache misses incurred by the applications as the metric to guide clustering. (In our experimental platform this information can be obtained with the STALLS_L2_MISS performance counter event.) We should highlight that, according to the definition provided in Sec. 2.2, this strategy does not strictly constitute a pure cache-clustering approach, since the cache partitions it creates may overlap with each other. This overlapping can create unpredictable interactions between applications that belong to different clusters [3].

In our experimental evaluation, we compare the effectiveness of our proposal (LFOC) to the Dunn and KPart approaches, and demonstrate that LFOC delivers higher reductions in unfairness than these approaches across the board. We should also highlight that Dunn and KPart are user-level clustering solutions, as opposed to our approach, implemented in the OS. User-level solutions like these may incur higher overheads since they make extensive use of system calls to access privileged resources such as performance monitoring hardware and cache partitioning facilities, which are handled directly by the OS. LFOC by contrast accesses these facilities directly by using a lightweight kernel-level API.

3 ANALYSIS OF THE OPTIMAL CACHE-CLUSTERING SOLUTION

As stated earlier, the design of our approach is inspired by the behavior of the optimal cache-clustering solution that optimizes fairness. In this section we provide an analysis on the optimal solution, which we could approximate for different workloads scenarios by using the PBBCCache [9] simulator. This simulation tool relies on offline-collected application performance data obtained for different cache sizes (e.g., instructions per cycle, memory bandwidth consumption, etc.) to determine the degree of throughput, fairness and other relevant metrics for a workload under a particular partitioning algorithm on a given platform. A key feature of the PBBCCache simulator is its ability to determine the optimal cache-partitioning and optimal cache-clustering solutions for different

Table 1: Classification of applications based on cache behavior

Type	Criterion
Streaming	$(Slowdown \leq 1.03 \text{ and } LLCMPKC \geq 10)$ in at least one way assignment, and $Slowdown < 1.06$ in all way assignments
Sensitive	If not streaming and $Slowdown \geq 1.05$ for a number of ways ≥ 2
Light sharing	Not streaming and not sensitive

optimizations objectives (e.g. throughput or fairness) using a parallel branch-and-bound algorithm. In approximating the application slowdown, which is necessary to determine the degree of fairness, PBBCCache accounts for the performance degradation due to both cache sharing and memory-bandwidth contention (to this end it uses a variant of the probabilistic model proposed in [15]).

To carry out our analysis with the simulator, we used performance counters to gather the average value of different runtime metrics with varying cache sizes for applications from the SPEC CPU2006 and CPU2017 suites running alone on a real system featuring an Intel Skylake processor with an 11-way 27.5MB LLC. (More information on this platform can be found in Sec. 5.) The offline-collected metric values, which correspond to the execution of the first 1500 billion instructions of the aforementioned benchmarks, are used as input to the simulator. This information is used to approximate the optimal clustering solution for fairness, namely, the solution to the optimal cache-clustering problem that obtains the optimal (minimal) unfairness value for the maximum throughput (STP) attainable.

For our experiments we considered randomly-generated multiprogram workloads including different number of SPEC CPU applications (from 4 to 16). According to the performance data collected offline we classify applications into three classes based on their degree of cache sensitivity and contentiousness: *Cache-sensitive*, *light sharing* and *streaming* programs. At high level, the cache-sensitive category is used for those programs that experience high performance drops as we reduce the number of cache ways allotted to them; this is not the case for *light sharing* and *streaming* applications. Streaming programs are characterized by exhibiting a low slowdown for almost all ways allocations, while incurring a high number of LLC misses per cycle. Applications of this kind are cache insensitive, and typically act as aggressor programs to cache-sensitive applications co-located on the same cache cluster, as the performance of the latter can be degraded substantially. Light-sharing programs are neither cache sensitive nor aggressive to others (the working set typically fits in the per-core private cache levels). Table 1 summarizes the criteria we followed to make this classification on our experimental platform, which is based on two offline-collected metrics: the application slowdown—relative performance with respect to using the entire LLC space—and the number of LLC Misses Per Kilo Cycles (LLCMPKC). As an example that illustrates the differences in the behavior between a streaming benchmark (lbm) and a cache-sensitive one (xalancbmk), Fig. 1 shows how the slowdown and the LLCMPKC varies with the amount of ways allocated to each application.

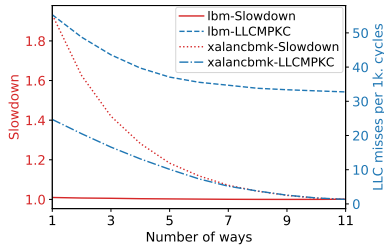


Figure 1: Slowdown and LLCMPKC for different way counts

After a thorough analysis of the optimal cache-clustering and optimal cache-partitioning solutions provided by the simulator for the various workloads, we draw the following major insights:

- In most cases, the cache-clustering solution that optimizes fairness isolates all streaming applications in a reduced set of ways (no greater than 2 in any workload). In many scenarios, a single 1-way cluster is used to confine all streaming programs.
- This same solution maps light-sharing programs onto different clusters following a hardly predictable pattern. More importantly, by conducting additional analysis with the simulator, we observed that moving individual light-sharing applications to different clusters has very little impact on throughput and fairness.
- As expected, by catering to the definition of the unfairness metric, the amount of ways assigned to cache-sensitive applications is critical for both throughput and fairness. Recall that the unfairness metric factors in the maximum slowdown observed across applications in the workload, and sensitive benchmarks typically incur a very high performance degradation if their cache size requirements are not fulfilled.
- The benefit that comes from assigning separate cache partitions (even optimally) to individual applications decreases dramatically as the number of applications gets closer to the number of cache ways. As an illustrative example, Fig. 3 shows the average unfairness delivered by the optimal partitioning solution, normalized to that of the optimal clustering solution for different workload sizes. As observed, optimal cache-partitioning suffers from increased unfairness as the workload size grows. When the application count matches the number of ways, each application can be assigned only one way under strict cache-partitioning -this is the only feasible option-, which gives rise to high unfairness in most workloads. Our overarching conclusion is that cache-clustering policies are clearly superior to cache-partitioning approaches as the ratio of the number of ways to the number of applications decreases.

To further illustrate the general behavior of the optimal clustering solution, Fig. 2 reports the average application count per cluster size, as well as the total number of clusters –grouped by its size (in ways)– that the solution builds for a subset of the workloads we explored: 20 randomly selected program mixes made up of 10 applications each. The data reported in the figure confirms the first three aforementioned observations. First, streaming applications

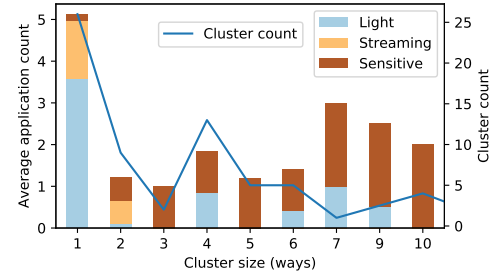


Figure 2: Cluster count and breakdown of applications into the different categories for each cluster size.

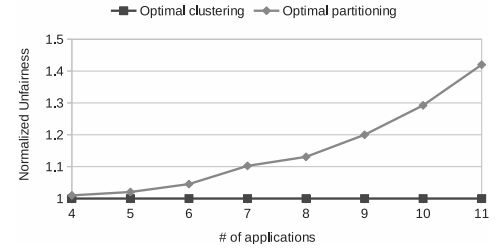


Figure 3: Comparison of optimal clustering vs optimal partitioning.

are typically confined in clusters with just one way allocated to them. In relative numbers, more than 87% of streaming application instances are assigned to this kind of clusters, while the remaining ones are allocated to 2-way clusters. Second, we can observe that light sharing applications are mapped to clusters with very different size; however, the vast majority of these programs are mapped to 1-way clusters. Third, the results reveal that cache-sensitive applications are predominantly present in big cache clusters. Specifically, more than 77% of the sensitive application instances are assigned to clusters with 4 or more ways. Finally, as is evident, 1-way clusters with a high number of applications are often present in the optimal solutions for the various workloads.

4 DESIGN AND IMPLEMENTATION

In this section we begin by describing how our clustering algorithm works at a high level. Then we proceed to indicate how applications are classified at runtime by leveraging data from hardware performance monitoring counters.

4.1 Algorithm outline

LFOC has been implemented on Linux as an extension of the OS scheduler. Specifically, it has been bundled in a loadable kernel module as a *monitoring plugin* of the PMCTrack tool [20]. This tool provides a kernel-level API to access privileged hardware facilities such as performance monitoring counters (PMCs) and Intel CAT features (i.e., HW support for cache way partitioning).

LFOC classifies applications at runtime into three classes based on its cache behavior –*light sharing*, *streaming* and *sensitive*– and assigns each application to a certain cache partition whose size is determined dynamically based on the properties of the workload.

When an application enters the system its cache behavior is unknown. To this end, a special *unknown* class is assigned to the

Algorithm 1: Cache-clustering algorithm used by LFOC

Input: ST , CS , and LS represent the sets of streaming, cache-sensitive and light-sharing applications, respectively; $max_streaming_way$ and $gaps_per_streaming$ are configurable parameters of LFOC (default value 5 and 3, respectively), nr_ways is the number of ways of the LLC.

function LFOC_partitioning(ST, CS, LS, nr_ways):
 $Clusters \leftarrow \emptyset$;
 $ways_for_streaming \leftarrow |ST|/max_streaming_way$;
for $i \leftarrow 1$ **to** $ways_for_streaming$ **do**
 Add a new 1-way cluster C to $Clusters$;
 Map up to $max_streaming_way$ apps from ST to C ;
 Remove assigned apps from ST ;
end
{ Use slowdown tables of CS apps. as input to lookahead }
 $W \leftarrow lookahead(CS, nr_ways - ways_for_streaming)$;
for $i \leftarrow 1$ **to** $|CS|$ **do**
 Add a new cluster C with $W[i]$ ways to $Clusters$;
 Map application i in CS to C ;
end
 $idx \leftarrow 0$;
while $|LS| > 0$ **and** $idx < ways_for_streaming$ **do**
 $TargetC \leftarrow Clusters[idx]$
 $gaps_available \leftarrow max_streaming_way - |TargetC| * gaps_per_streaming$;
 if $gaps_available > 0$ **then**
 Map up to $gaps_available$ apps from LS to $TargetC$;
 Remove assigned apps from LS ;
 end
 end
 Distribute remaining applications in LS in a round-robin fashion among non-streaming clusters;
return $Clusters$

application right after being spawned. At the beginning of the execution, each thread has to go through a warm-up period (3 sampling intervals in our experimental setting). Any performance information gathered with hardware counters during the warm-up period is not used to classify applications, so as to mitigate mispredictions associated with cold-start effects (e.g. the number of cache misses typically spikes intermittently at the beginning of the execution).

Periodically, our scheduler extension activates the partitioning scheme depicted in Algorithm 1, which relies on the conclusions of the analysis presented in Sec. 3. Overall, the algorithm reserves up to two cache ways to map streaming programs. The remaining cache ways are distributed among cache sensitive applications, which are then assigned to separate cache partitions. The size of these partitions is determined by means the lookahead algorithm [19], using as input the slowdown curve for each application (i.e., slowdown registered for different cache ways) built by using IPC values obtained online. With this cache-way distribution for cache-sensitive applications, LFOC attempts to fulfill their cache requirements based on the degree of cache sensitivity. Finally, light sharing applications are distributed among the various partitions, by attempting to populate partitions with streaming applications first, as the optimal solution typically does.

4.2 Application Classification

Once the warm-up period has elapsed for a particular application, LFOC enters a *sampling mode* whose goal is to determine the application class based on its performance sensitivity to the amount of space assigned in the LLC. This is crucial to decide on the share of the total cache space to be allotted to the application, as well as to determine what co-runners in the workload (if any) must be assigned to the same cache partition [14].

The sampling mode is inspired by the technique proposed in [3], which operates as follows. Two non-overlapping complementary cache partitions covering the entire LLC space are created; the first one, referred to as the *sampling partition*, is reserved for the application that triggered the transition into sampling mode, and the other one is devoted to the remaining applications. To determine the application class –based on the classification criteria presented in Sec. 3– the value of various hardware events (i.e., number of instructions retired, cycles and LLC misses) is gathered with PMCs as we increase the size of the sampling partition. Notably, for sensitive applications we also obtain the slowdown curve, which is required to create partitions for these applications, as depicted in Algorithm 1. Once the sampling process terminates, LFOC transitions back into the normal operating mode described earlier.

In the original approach [3], the size of the first partition is varied from 1 to the number of ways minus one, whereas the space of the second cache partition (complementary) decreases accordingly. This full sweep is required by the dynamic version of the KPart clustering algorithm [3], which relies on the ability to accurately determine the number of Instructions Per Cycle (IPC) and the number of LLC Misses Per Kilo Instructions (LLCMPKI) for each way count and for *every application* in the workload. We observed that this approach introduces substantial overheads due to the fact that the cache assignment enforced during the sampling mode is typically suboptimal. The sampling application receives a progressively higher amount of cache space, while the remaining applications share a increasingly smaller cache cluster. This usually leads to performance/fairness degradation especially when cache sensitive applications and streaming programs are included in the workload.

To overcome these shortcomings, LFOC immediately puts a stop to the sampling process in scenarios where increasing the size of the sampling partition further provides no additional information to the clustering algorithm. Firstly, when LLC miss rate falls below a certain low threshold, performance does not increase much when allotting more cache space to the application, so we expect IPC values –used to construct slowdown tables– to remain very close beyond that point. Secondly, streaming applications typically exhibit a very low increase in performance when granting more cache space to them. In these scenarios, LFOC interrupts the sampling process and proceeds to determine the application class. In practice, to successfully identify many streaming and light sharing applications –for which slowdown curves are not needed for clustering– only a few way counts for the sampling partition must be explored. When the application class is sensitive and the sampling process was canceled (due to the first criterion), LFOC uses the last IPC sample gathered to approximate the performance with higher ways counts, which is necessary to build the entire slowdown table.

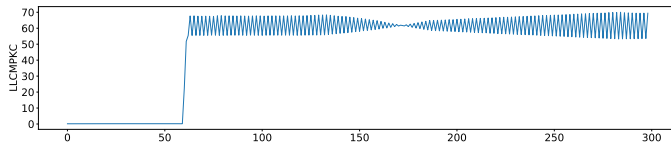


Figure 4: LLCMPKC captured at the beginning of the execution of `fotonik3d`.

Because an application may exhibit different program phases at runtime, the initial classification may not be representative throughout the execution. For example, Fig. 4 shows the LLCMPKC of the streaming `fotonik3d` application over time, where a short light-sharing phase precedes the streaming behavior that the program exhibits for the vast majority of the execution. Failing to determine application classes accurately could lead to suboptimal cache-partitioning for certain time periods, and hence to unfairness. Triggering the sampling mode periodically helps to mitigate this issue, but, unfortunately, it backfires by introducing substantial overheads. To determine application classes at runtime in a lightweight manner, LFOC triggers a transition into the sampling mode only in the event that the application class has likely changed. To this end, the OS continuously monitors for each application the value of the LLCMPKC metric and the fraction of pipeline stall cycles incurred due to long-latency memory accesses (approximated via the `STALLS_L2_MISS` performance counter event, also used in [23]), and leverages a few heuristics to capture class changes. In particular, a class change is signaled for a light sharing application if it enters a memory-intensive phase, namely, the average LLCMPKC measured over the last five monitoring periods exceeds a `high_threshold` (10 in our experimental setting, as reported in Table 1 for streaming-like behavior) or the average fraction of long-latency memory-access stalls is greater than 25%. This approach filters out spikes in the aforementioned metrics while effectively identifying memory-intensive phases. Conversely, for streaming programs, which LFOC typically assigns to cache clusters consisting of one way, the sampling mode is engaged if its average LLCMPKC falls below a `low_threshold` (defined as 30% of `high_threshold`). Finally, for sensitive applications, LFOC associates a *critical size*, defined as the amount of cache space where the slowdown falls below 1.05. The *critical size* is determined during the last sampling period triggered by the application. Essentially, a class change is signaled for sensitive applications when they enter a stable non-memory intensive phase (inverse of the criterion presented earlier for light-sharing applications) for effective cache allocations¹ smaller than the critical size, or when the average LLCMPKC is higher than `high_threshold` for an amount of cache space bigger than the critical size.

5 EXPERIMENTS

To assess the effectiveness of our OS-level cache-clustering approach we implemented it in the Linux kernel v4.9.160. For the experiments we used a server platform featuring a Xeon Gold 6138 “Skylake” multicore processor where cores run at 2Ghz. This processor integrates an 11-way 27.5MB last level (L3) cache that supports

¹The amount of cache space used by an application is gathered by leveraging the Intel Cache Monitoring Technology.

way-partitioning; each core features a 1.25MB L1 and a 20MB L2 caches (private levels).

On this platform we carried out an experimental comparison of LFOC with the stock Linux kernel –it does not partition the LLC–, and with the Dunn [23] and KPart [3] cache-partitioning policies, specifically designed to optimize fairness and system throughput respectively. To perform a fair comparison with previous approaches we used a similar methodology as that described in the corresponding articles [3, 23]. Essentially, we conduct experiments with HPC multiprogram workloads consisting of a mix of single-threaded benchmarks from SPEC CPU, and run each program for a fixed number of instructions (150 billion instructions in our setting). Specifically, we ensure that all applications in the mix are started simultaneously, and when one of them completes the corresponding instructions, the program is restarted repeatedly until the longest application in the set completes three times. We then measure unfairness and STP (throughput), by using the geometric mean of the completion times for each program.

Fig. 5 depicts the composition of the 36 randomly generated workloads we used in our experiments, which are made of benchmarks from the SPEC CPU2006 and CPU2017 suites. Note that we selected applications from both suites to experiment with a wider range of streaming and cache-sensitive programs, as most benchmarks in both suites exhibit a light sharing, cache-insensitive execution profile on our platform. This is caused in part due to the coarse granularity of the cache partitions we can create on this system: the smallest partition is as big as 2.5MB. As is evident, we considered workloads of 8, 12 and 16 applications each, so as to analyze the impact that the workload size has on the fairness improvement achieved by each approach.

In this section we first evaluate the effectiveness of the cache-clustering algorithms associated with the KPart, Dunn and LFOC approaches. We then proceed to analyze how well dynamic clustering approaches capture the time-changing behavior of the applications in different workloads.

5.1 Evaluation of Clustering Algorithms

Our goal is to measure the degree of fairness and throughput delivered by a certain clustering strategy alone (i.e. how applications are grouped into shared or separate clusters according to their runtime properties) putting aside the associated overheads due to algorithm execution at runtime, performance monitoring and cache allocation. We do account for these overheads in the experiments of the next section.

To assess the effectiveness of each clustering algorithm, we consider workloads consisting of applications whose behavior falls in a clear class (cache sensitive, streaming or light sharing) for the vast majority of the execution (*S_i* workloads in Fig. 5). For the analysis, we implemented the clustering algorithms used by KPart, Dunn and LFOC on top of the simulator described in Sec. 3, which accepts as input the average value of different performance metrics gathered offline for different cache sizes. To conduct the corresponding experiments, we launch the simulator prior to the execution of each workload to retrieve the cache-partitions and application-to-partition mappings imposed by a certain clustering strategy. Then, we enforce the corresponding cache partitions on a per-process

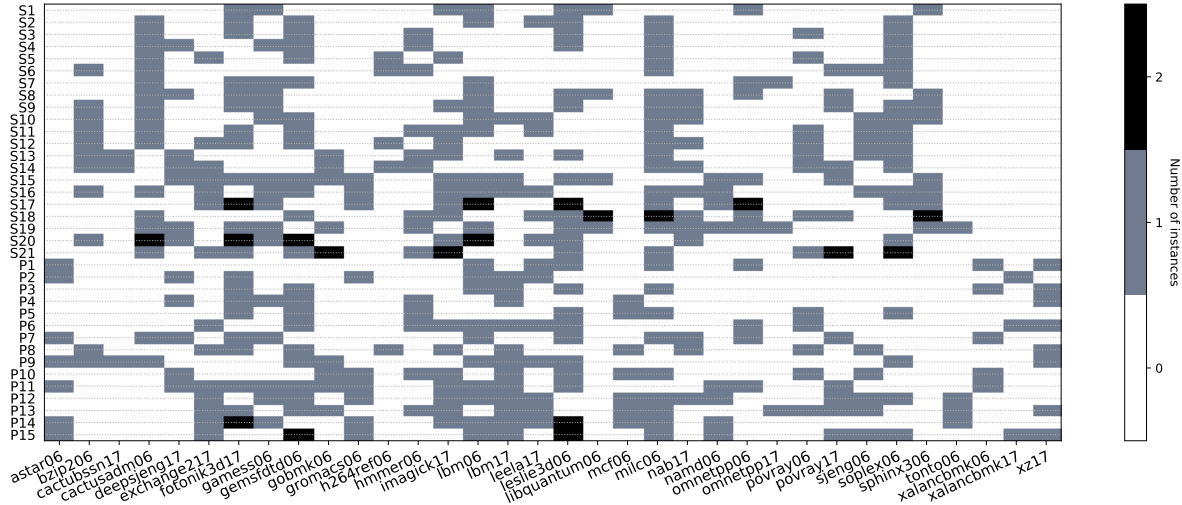


Figure 5: Multiprogram workloads used for our experiments. Each matrix cell indicates the number of instances of a benchmark (x-axis) in a workload (y-axis).

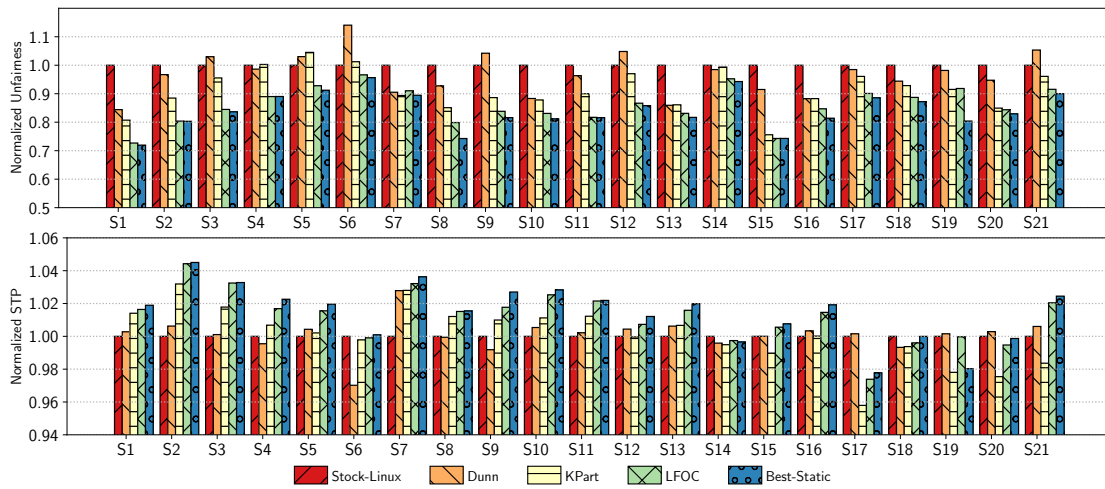


Figure 6: Normalized unfairness and STP values obtained by the static version of the various clustering algorithms.

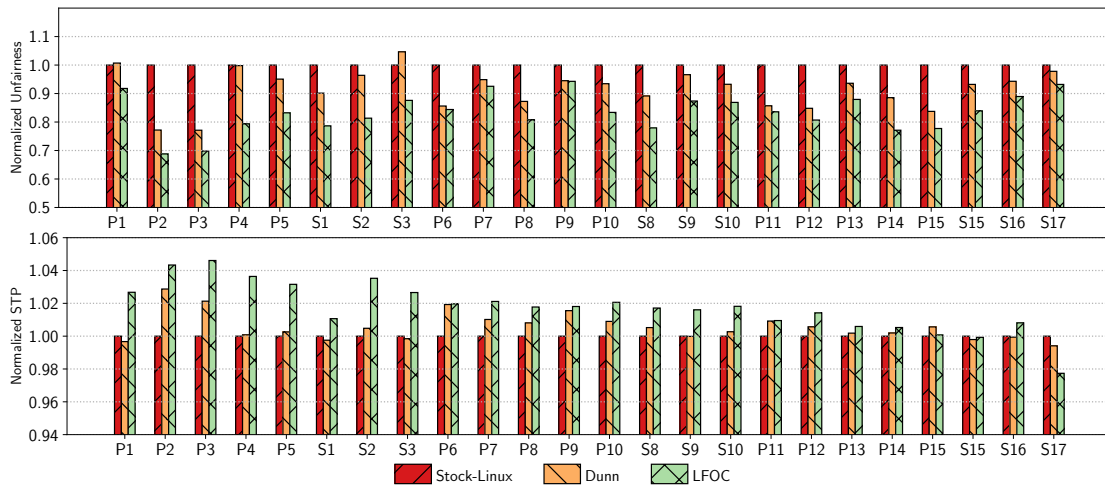


Figure 7: Normalized unfairness and STP values delivered by the dynamic cache clustering approaches

manner from user-space, using the PMCTrack tool [20], and proceed to launch the workload, which will use the same *static* cache configuration throughout the execution. For comparison purposes, we have also gathered the results of an ideal cache-clustering policy, referred to as *Best-Static*, which establishes the cache-partitions and application-to-cluster mappings based on the optimal fairness solution determined by the simulator.

Fig. 6 shows the degree of unfairness and throughput delivered by the different clustering strategies; the values have been normalized to the results of *Stock-Linux* (no cache partitioning). The results reveal that the *Dunn* approach, designed to optimize fairness, exhibits a non-uniform behavior across workloads; for some program mixes it is capable to reduce unfairness up to 15.5%, but for others it causes substantial fairness degradation (by a factor of up to 1.14x) relative to *Stock-Linux*. We found that this is due to its exclusive reliance on the `STALLS_L2_MISS` performance event; the higher the value of this event, the higher the number of cache ways allotted by *Dunn* to the application [23]. More specifically, we observed that some streaming (aggressor) cache-insensitive applications, such as *GemsFDTD* or *fotonik3d* exhibit high values of this event, as their performance is greatly affected by memory accesses. These applications can be mapped together to the same (or overlapping) cache partitions with highly sensitive programs, such as *soplex* or *omnetpp*, for which the event reaches similar figures, leading to performance degradation and unfairness. Based on this insight, we conclude that using the `STALLS_L2_MISS` event alone is not enough to drive fairness-aware partitioning policies.

We also observe that *KPart*'s clustering algorithm, designed to optimize throughput, brings modest throughput gains² in the workloads we explored (up to 3%). However, this approach does bring substantial unfairness reductions (8.6% on average). Nonetheless, we observe that *LFOC*'s simpler and more lightweight partitioning algorithm provide substantially better fairness than *KPart* for the vast majority of the workloads (up to 27.3%, and 14% on average relative to *Stock Linux*). At the same time, *LFOC* achieves higher throughput than *KPart* across the board, and performs in a close range (1.8% on average) of the *Best Static* approach (our approximation to the optimal policy in these workload scenarios).

5.2 Study of the dynamic policies

For the evaluation in this section we used our OS-level implementation of *LFOC*, and also created a user-level implementation of *Dunn*, as it was originally proposed [23] as a user-level cache clustering approach. A good property of *Dunn* is the fact that it only requires the continuous monitorization of the `STALLS_L2_MISS` performance event for the various applications over time. The simplicity of the *Dunn* approach stands in contrast to the higher complexity of *KPart*, which relies on the ability to accurately gather a substantial amount of performance information online for each application (LLCMPKI and IPC values for every possible cache-way count) in order to apply the clustering algorithm.

In an attempt to evaluate the dynamic version of *KPart* –referred to as *KPart-Dynaway* [3]– we considered the user-level implementation created by the authors [4]. Unfortunately, this implementation,

Table 2: Average execution time (in ms) of the *KPart* and *LFOC* algorithms

#Apps.	4	5	6	7	8	9	10	11
<i>LFOC</i>	0.00151	0.00154	0.00163	0.00174	0.00174	0.00182	0.00191	0.00216
<i>KPart</i>	0.51800	0.79600	1.21800	1.48100	2.01200	2.74200	3.32000	4.14000

which consists of roughly 4K lines of C++ code and makes intensive use of the *Armadillo* linear algebra library, was specifically tailored to the hardware platform where the authors conducted the experiments [3], and makes numerous assumptions that do not apply in our experimental setting (e.g., the number of cache ways should be no smaller than the number of applications in the workload). Due to these platform-specific assumptions and other issues –as yet unidentified–, the execution of *KPart-Dynaway* crashes shortly after the partitioning algorithm is executed for the first time, preventing us from launching any of the workloads we considered for the evaluation. We leave for future work the adaptation of this somewhat complex implementation for our platform, and its complete evaluation. Nevertheless, to highlight the enormous difference between the complexity of *KPart*'s partitioning algorithm and the one used by our approach for different number of applications, Table 2 shows the execution time for both algorithms (compiled with aggressive optimizations) for different workload sizes. We were able to gather that information from *KPart*'s implementation by instrumenting the code of the partitioning algorithm that completes successfully (for workloads with less than twelve applications) right before the program's crash. As is evident *LFOC*'s execution time (2 μ s) is up to three orders of magnitude smaller than *KPart*'s, which can take over 4ms to complete for 11 applications (slightly longer than the default timer tick in the Linux kernel). As we showed in the previous section, this increased complexity does not enable *KPart* to provide better fairness than our lightweight approach.

In our OS-level implementation of *LFOC* we sample performance counters every 100M instructions during the normal operation mode (see Sec. 4) and every 10M instructions during the sampling mode. Using a shorter instruction window for the sampling mode makes it possible to reduce the amount of time required for a full cache way sweep (worst-case scenario). The time required by the sampling mode cycle is 6.2ms on average. Notably, we observed that in most cases full cache-way sweeps are not performed under *LFOC* as the partitioning algorithm (as explained in Sec. 4) does not require detailed per-way metrics for all applications, as opposed to *KPart*'s algorithm. In our experiments, the partitioning algorithm for both *Dunn* and *LFOC* is executed every 500ms, as this is the setting used in the original evaluation of the *Dunn* approach [23].

Fig. 7 shows the normalized unfairness and throughput values delivered by the *Dunn* and *LFOC* (dynamic) approaches for different workloads. Note that in this case, we considered additional program mixes (*Pi* workloads) that include applications such as *xz*, *astar*, *mcf* or *xalancbmk*, which exhibit distinct long-term program phases with varying degree of memory intensity. Some of these applications go through highly cache sensitive phases, so *Stock-Linux* delivers higher unfairness values in these scenarios. That is the reason why *Dunn* exhibits a slightly fairer behavior under these circumstances relative to the scenario considered in the previous section. Yet, *LFOC* is capable to provide better throughput

²In the original paper, the authors report a 24% average increase in throughput on a different platform, with workloads whose composition was not disclosed.

than Dunn, and improves fairness over Dunn across the board (up to 20.5% for P4, and 9% on average). With respect to Stock-Linux, LFOC reduces unfairness by 16.7% on average.

6 CONCLUSIONS

In this article we have presented LFOC, an OS-level cache-clustering approach that leverages cache-partitioning support in Intel-CAT enabled multicore processors to improve fairness while maintaining acceptable throughput. LFOC classifies applications into three classes according to their degree of memory-intensity and cache sensitivity, and ensures that streaming aggressor benchmarks are confined to small cache partitions so they are effectively isolated from cache-sensitive benchmarks, which are assigned an amount of cache space in accordance to its sensitivity. In doing so, LFOC tries to mimic the behavior of the optimal cache clustering solution, which we approximated by means of a simulator. We implemented LFOC in the Linux kernel and assessed its effectiveness on a commercial multicore platform featuring an Intel Skylake processor. Our experiments reveal that LFOC is able to deliver an average 16.7% fairness improvement relative to stock Linux. At the same time, LFOC clearly outperforms two other existing cache-clustering approaches, one of which was specifically designed to deliver fairness [23].

Key aspects of LFOC are its lightweight clustering algorithm, the online heuristics it leverages to classify applications online, and its ability to fairly share the space on the LLC among applications by using limited monitoring information, which can be obtained at runtime without collecting performance data for every possible way count, unlike other cache clustering algorithms [3].

ACKNOWLEDGMENTS

This work has been supported by the EU (FEDER), the Spanish MINECO and CM, under grants TIN 2015-65277-R and S2018/TCS-4423. Adrian Garcia-Garcia is supported by a UCM fellowship grant.

REFERENCES

- [1] Jacob Brock et al. 2015. Optimal Cache Partition-Sharing. In *Proceedings of the 2015 44th International Conference on Parallel Processing (ICPP) (ICPP '15)*. 749–758.
- [2] Eiman Ebrahimi et al. 2010. Fairness via source throttling: a configurable and high-performance fairness substrate for multi-core memory systems. In *15th Int'l Conf. Architectural Support Programming Lang. and Oper. Syst. (ASPLOS 10)*. 335–346.
- [3] N. El-Sayed et al. 2018. KPart: A Hybrid Cache Partitioning-Sharing Technique for Commodity Multicores. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 104–117.
- [4] N. El-Sayed et al. 2018. Source Code of KPart. <https://github.com/Nosayba/kpart>. Accessed: 2019-02-20.
- [5] S. Eyerhan and L. Eeckhout. 2008. System-Level Performance Metrics for Multi-program Workloads. *IEEE Micro* 28, 3 (May 2008), 42–53.
- [6] J. Feliu et al. 2016. Perf & Fair: a Progress-Aware Scheduler to Enhance Performance and Fairness in SMT Multicores. *IEEE Trans. Comput. PP*, 99 (2016).
- [7] Liran Funaro, Orna Agmon Ben-Yehuda, and Assaf Schuster. 2016. Ginseng: Market-driven LLC Allocation. In *Proceedings of the 2016 USENIX Annual Technical Conference (USENIX ATC '16)*. 295–308.
- [8] A. Garcia-Garcia, J. C. Saez, and M. Prieto-Matias. 2018. Contention-Aware Fair Scheduling for Asymmetric Single-ISA Multicore Systems. *IEEE Trans. Comput.* 67, 12 (Dec 2018), 1703–1719. <https://doi.org/10.1109/TC.2018.2836418>
- [9] Adrian Garcia-Garcia, Juan Carlos Saez, and Manuel Prieto-Matias. 2019. PBB-Cache: A parallel branch-and-bound based cache-partitioning simulator. *Submitted for review to International Journal of Computational Science* (2019).
- [10] JA Hartigan and MA Wong. 1979. Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics* (1979), 100–108.
- [11] Samira Manabi Khan et al. 2014. Improving cache performance using read-write partitioning. In *20th IEEE International Symposium on High Performance Computer Architecture, HPCA 2014*. 452–463.
- [12] David Lo et al. 2015. Heracles: improving resource efficiency at scale. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture*. 450–462.
- [13] R Manikantan, Kaushik Rajan, and R Govindarajan. 2012. Probabilistic Shared Cache Management (PriSM). In *Proceedings of the 39th Annual International Symposium on Computer Architecture (ISCA '12)*. 428–439.
- [14] Sparsh Mittal. 2017. A Survey of Techniques for Cache Partitioning in Multicore Processors. *ACM Comput. Surv.* 50, 2, Article 27 (May 2017), 27:1–27:39 pages.
- [15] Tomer Y. Morad et al. 2016. EFS: Energy-Friendly Scheduler for memory bandwidth constrained systems. *J. Parallel and Distrib. Comput.* 95 (2016), 3 – 14.
- [16] Anurag Mukkara, Nathan Beckmann, and Daniel Sanchez. 2016. Whirlpool: Improving Dynamic Cache Management with Static Data Classification. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '16)*. 113–127.
- [17] Onur Mutlu and Thomas Moscibroda. 2007. Stall-Time Fair Memory Access Scheduling for Chip Multiprocessors. In *40th Ann. IEEE/ACM Int'l Symp. on Microarchitecture (MICRO 07)*. 146–160.
- [18] K. Nguyen. 2016. Introduction to Cache Allocation Technology in the Intel Xeon Processor E5 v4 Family. <https://software.intel.com/en-us/articles/introduction-to-cache-allocation-technology>. Accessed: 2019-03-20.
- [19] Moinuddin K. Qureshi and Yale N. Patt. 2006. Utility-Based Cache Partitioning: A Low-Overhead, High-Performance, Runtime Mechanism to Partition Shared Caches. In *Proceedings of MICRO 06*. 423–432.
- [20] Juan Carlos Saez et al. 2017. PMCTrack: Delivering Performance Monitoring Counter Support to the OS Scheduler. *Comput. J.* 60, 1 (2017), 60–85.
- [21] Juan Carlos Saez, José Ignacio Gomez, and Manuel Prieto. 2008. Improving Priority Enforcement via Non-Work-Conserving Scheduling. In *ICPP '08: Proceedings of the 2008 37th International Conference on Parallel Processing*. 99–106.
- [22] Alberto Scolari, Davide Basilio Bartolini, and Marco Domenico Santambrogio. 2016. A Software Cache Partitioning System for Hash-Based Caches. *ACM Trans. Archit. Code Optim.* 13, 4, Article 57 (Dec. 2016), 57:1–57:24 pages.
- [23] V. Selfa et al. 2017. Application Clustering Policies to Address System Fairness with Intel's Cache Allocation Technology. In *2017 26th International Conference on Parallel Architectures and Compilation Techniques (PACT)*. 194–205.
- [24] Timothy Sherwood, Brad Calder, and Joel Emer. 1999. Reducing Cache Misses Using Hardware and Software Page Placement. In *Proceedings of the 13th International Conference on Supercomputing (ICS '99)*. 155–164.
- [25] Lavanya Subramanian et al. 2015. The Application Slowdown Model: Quantifying and Controlling the Impact of Inter-application Interference at Shared Caches and Main Memory. In *Proceedings of the 48th International Symposium on Microarchitecture (MICRO-48)*. 62–75.
- [26] K. Van Craeynest et al. 2013. Fairness-aware scheduling on single-ISA heterogeneous multi-cores. In *22nd Int'l Conf. Parallel Arch. Compilation Techniques (PACT 13)*. 177–187.
- [27] Ruisheng Wang and Lihong Chen. 2014. Futility Scaling: High-Associativity Cache Partitioning. In *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-47)*. 356–367.
- [28] Di Xu et al. 2012. Providing Fairness on Shared-memory Multiprocessors via Process Scheduling. In *Proc. ACM Int'l Conf. Measurement and Modeling Comp. Syst. (SIGMETRICS 12)*. 295–306.
- [29] Ying Ye, Richard West, Zhuoqun Cheng, and Ye Li. 2014. COLORIS: A Dynamic Cache Partitioning System Using Page Coloring. In *Proceedings of the 23rd International Conference on Parallel Architectures and Compilation (PACT '14)*. 381–392.
- [30] Chenjie Yu and Peter Petrov. 2010. Off-chip Memory Bandwidth Minimization Through Cache Partitioning for Multi-core Platforms. In *Proceedings of the 47th Design Automation Conference (DAC '10)*. 132–137.
- [31] Heechul Yun et al. 2014. PALLOC: DRAM bank-aware memory allocator for performance isolation on multicore platforms. In *20th Real-Time Embedded Tech. and Applications Symp. (RTAS 14)*. 155–166.
- [32] H. Yun et al. 2016. Memory Bandwidth Management for Efficient Performance Isolation in Multi-Core Platforms. *IEEE Trans. Comput.* 65, 2 (Feb 2016), 562–576.
- [33] Xiao Zhang, Sandhya Dwarkadas, and Kai Shen. 2009. Towards Practical Page Coloring-based Multicore Cache Management. In *Proceedings of the 4th ACM European Conference on Computer Systems (EuroSys '09)*. 89–102.
- [34] Haishan Zhu and Mattan Erez. 2016. Dirigent: Enforcing QoS for Latency-Critical Tasks on Shared Multicore Systems. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '16)*. 33–47.
- [35] Sergey Zhuravlev et al. 2012. Survey of Scheduling Techniques for Addressing Shared Resources in Multicore Processors. *ACM Comput. Surv.* 45, 1, Article 4 (Dec. 2012), 28 pages.
- [36] Sergey Zhuravlev et al. 2012. Survey of Scheduling Techniques for Addressing Shared Resources in Multicore Processors. *ACM Comput. Surv.* 45, 1, Article 4 (Dec. 2012), 28 pages.