Contention-Aware Scheduling on Multicore Systems

SERGEY BLAGODUROV, SERGEY ZHURAVLEV, and ALEXANDRA FEDOROVA
Simon Fraser University

Contention for shared resources on multicore processors remains an unsolved problem in existing systems despite significant research efforts dedicated to this problem in the past. Previous solutions focused primarily on hardware techniques and software page coloring to mitigate this problem. Our goal is to investigate how and to what extent contention for shared resource can be mitigated via thread scheduling. Scheduling is an attractive tool, because it does not require extra hardware and is relatively easy to integrate into the system. Our study is the first to provide a comprehensive analysis of contention-mitigating techniques that use only scheduling. The most difficult part of the problem is to find a classification scheme for threads, which would determine how they affect each other when competing for shared resources. We provide a comprehensive analysis of such classification schemes using a newly proposed methodology that enables to evaluate these schemes separately from the scheduling algorithm itself and to compare them to the optimal. As a result of this analysis we discovered a classification scheme that addresses not only contention for cache space, but contention for other shared resources, such as the memory controller, memory bus and prefetching hardware. To show the applicability of our analysis we design a new scheduling algorithm, which we prototype at user level, and demonstrate that it performs within 2% of the optimal. We also conclude that the highest impact of contention-aware scheduling techniques is not in improving performance of a workload as a whole but in improving quality of service or performance isolation for individual applications and in optimizing system energy consumption.

Categories and Subject Descriptors: D.4.1 [**Operating Systems**]: Process Management—*Scheduling*

General Terms: Management, Measurement, Performance

Additional Key Words and Phrases: Multicore processors, scheduling, shared resource contention

ACM Reference Format:
Blagodurov, S., Zhuravlev, S., and Fedorova, A. 2010. Contention-aware scheduling on multicore systems. ACM Trans. Comput. Syst. 28, 4, Article 8 (December 2010), 45 pages.
DOI: 10.1145/1880018.1880019. http://doi.acm.org/10.1145/1880018.1880019.

The authors thank Sun Microsystems, National Science and Engineering Research Council of Canada (in part the Strategic Project Grants program), and Google for supporting this research.
Author's address: S. Blagodurov, S. Zhuravlev, and A. Fedorova, Simon Fraser University, 8888 University Drive, Burnaby, B.C., Canada V5A 1S6; email: {sergey_blagodurov, sergey_zhuravlev, alexandra_fedorova}@sfu.ca

1. INTRODUCTION

Multicore processors have become so prevalent in both desktops and servers that they may be considered the norm for modern computing systems. The limitations of techniques focused on extraction of instruction-level parallelism (ILP) and the constraints on power budgets have greatly staggered the development of large single cores and made multicore systems a very likely future of computing, with hundreds to thousands of cores per chip. In operating system scheduling algorithms used on multicore systems, the primary strategy for placing threads on cores is load balancing. The OS scheduler tries to balance the runnable threads across the available resources to ensure fair distribution of CPU time and minimize the idling of cores. There is a fundamental flaw with this strategy which arises from the fact that a core is not an independent processor but rather a part of a larger on-chip system and hence shares resources with other cores. It has been documented in previous studies [Cho and Jin 2006; Liedtke et al. 1997; Lin et al. 2008; Qureshi and Patt 2006; Suh et al. 2002; Tam et al. 2009] that the execution time of a thread can vary greatly depending on which threads run on the other cores of the same chip. This is especially true if several cores share the same last-level cache (LLC).

Figure 1 highlights how the decisions made by the scheduler can affect the performance of an application. This figure shows the results of an experiment where four applications were running simultaneously on a system with four cores and two shared caches. There are three unique ways to distribute the four applications across the four cores, with respect to the pairs of co-runners sharing the cache; this gives us three unique schedules. We ran the threads in each of these schedules, recorded the average completion time for all applications in the workload, and labeled the schedule with the lowest average completion time as the best and the one with the highest average completion time as the worst. Figure 1 shows the performance degradation that occurs due to sharing an LLC with another application, relative to running solo (contention-free). The best schedule delivers a 20% better average completion time than the worst one. Performance of individual applications improves by as much as 50%.

Previous work on the topic of contention-aware scheduling for multicore systems focused primarily on the problem of *cache contention* since this was assumed to be the main if not the only cause of performance degradation [Dhiman et al. 2009; Fedorova et al. 2007; Knauerhase et al. 2008; Tam et al. 2007, 2009]. In this context cache contention refers to the effect when an application is suffering extra cache misses because its co-runners (threads running on cores that share the LLC) bring their own data into the LLC evicting the data of others. Methods such as utility cache partitioning (UCP) [Qureshi and Patt 2006] and page coloring [Cho and Jin 2006; Tam et al. 2009; Zhang et al. 2009] were devised to mitigate cache contention.

Through extensive experimentation on real systems as opposed to simulators we determined that cache contention is not the dominant cause of performance degradation of threads co-scheduled to the same LLC. Along with cache contention other factors like memory controller contention, memory bus

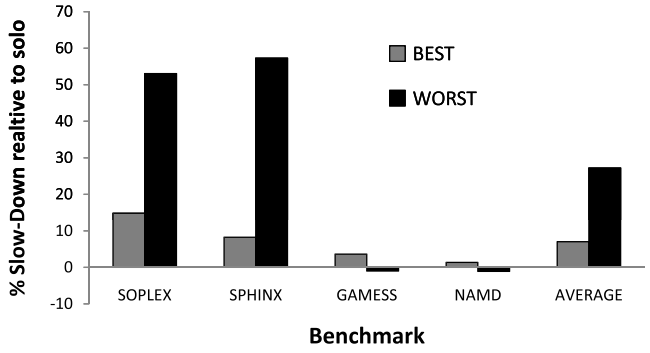


Fig. 1. The performance degradation relative to running solo for two different schedules of SPEC CPU2006 applications on an Intel Xeon X3565 quad-core processor (two cores share an LLC).

contention, and prefetching hardware contention all combine in complex ways to create the performance degradation that threads experience when sharing the LLC.

Our goal is to investigate contention-aware scheduling techniques that are able to mitigate as much as possible the factors that cause performance degradation due to contention for shared resources. Such a scheduler would provide speedier as well as more stable execution times from run to run. Any contention aware scheduler must consist of two parts: a classification scheme for identifying which applications should and should not be scheduled together as well as the scheduling policy that assigns threads to cores given their classification. Since the classification scheme is crucial for an effective algorithm, we focused on the analysis of various classification schemes. We studied the following schemes: Stack Distance Competition (SDC) [Chandra et al. 2005], Animal Classes [Xie and Loh 2008], Solo Miss Rate [Knauerhase et al. 2008], and the Pain Metric. The best classification scheme was used to design a scheduling algorithm, which was prototyped at user level and tested on two very different systems with a variety of workloads.

Our methodology allowed us to identify *the last-level cache miss rate*, which is defined to include all requests issued by LLC to main memory including prefetching, as one of the most accurate predictors of the degree to which applications will suffer when co-scheduled. We used it to design and implement a new scheduling algorithm called Distributed Intensity (DI). We show experimentally on two different multicore systems that DI performs better than the default Linux scheduler, delivers much more stable execution times than the default scheduler, and performs within a few percentage points of the theoretical optimal. DI needs only the real miss rates of applications, which can be easily obtained online. As such we developed an online version of DI, DI Online (DIO), which dynamically reads miss counters online and schedules applications in real time. Our schedulers are implemented at user-level, and although they could be easily implemented inside the kernel, the user-level implementation was sufficient for evaluation of these algorithms' key properties.

The key contribution of our work is the analysis demonstrating the effectiveness of various classification schemes in aiding the scheduler to mitigate shared resource contention. Previous studies focusing on contention-aware scheduling did not investigate this issue comprehensively. They attempted isolated techniques, in some cases on a limited number of workloads, but did not analyze a variety of techniques and did not quantify how close they are to optimal. Therefore, understanding what is the best we can do in terms of contention-aware scheduling remains an open question. Our analysis, in contrast, explores a variety of possible classification schemes for determining to what extent the threads will affect each other's performance, and we believe to cover most of the schemes previously proposed in literature as well as introducing our own. In order to perform thorough analysis we devised a new methodology for evaluating classification schemes independently of scheduling algorithms. Using this methodology we compare each classification scheme to the theoretical optimal, and this provides a clear understanding of what is the best we can do in a scheduler. Further, we analyze the extent of performance improvements that can be achieved for different workloads by methodically categorizing the workloads based on the potential speedup they can achieve via contention aware scheduling. This enables us to evaluate the applicability of cache-aware scheduling techniques for a wide variety of workloads. We believe that our work is the first to comprehensively evaluate the potential of scheduling to mitigate contention for shared resources.

The primary application of our analysis is for building new scheduling algorithms that mitigate the effects of shared resource contention. We demonstrate this by designing and evaluating a new algorithm Distributed Intensity Online. Our evaluation leads us to a few interesting and often unexpected findings. First of all, we were surprised to learn that if one is trying to improve average workload performance, the default contention unaware scheduler already does a rather good job if we measure performance of a workload over a large number of trials. The reason is that for a given workload there typically exists a number of good and bad scheduling assignments. In some workloads, each of these assignments can be picked with a roughly equal probability if selecting uniformly at random, but in other workloads a good assignment is far more likely to occur than the bad one. A contention unaware default scheduler runs into good and bad assignments according to their respective probabilities, so over time it achieves performance that is not much worse than under a contention-aware algorithm that always picks the good assignment. However, when one is interested in improving performance of individual applications, for example to deliver quality of service guarantees or to accomplish performance isolation, a contention-aware scheduler can offer significant improvements over default, because this scheduler would almost never select a bad scheduling assignment for the prioritized application.

We also found that DIO can be rather easily adjusted to optimize for system energy consumption. On many systems, power consumption can be reduced if the workload is concentrated on a handful of chips, so that remaining chips can be brought into a low-power state. At the same time, clustering the workload on a few chips forces the threads to share memory-hierarchy resources more

intensely. This can lead to contention, hurt performance and increase system uptime, leading to increased energy consumption. So in order to determine whether threads should be clustered (to save power) or spread across chips (to avoid excessive contention) the scheduler must be able to predict to what extent threads will hurt each other's performance if clustered. We found that DIO was able to make this decision very effectively, and this lead us to design DIO-POWER – an algorithm that optimizes both performance and power consumption. As a result, DIO-POWER is able to improve energy-delay product over plain DIO, by as much as 80% in some cases.

The rest of the article is organized as follows. Section 2 describes the classification schemes and policies that we evaluated, the methodology for evaluating the classification schemes separately from the policies and provides the evaluation results for the classification schemes. Section 3 attempts to quantify the effects of different factors resulting from contention for shared on-chip resources on performance on multicore CPUs in order to better explain the results of Section 2. Section 4 describes the contention aware scheduling algorithms that were implemented and tested on real systems. Section 5 provides the experimental results for the contention aware scheduling algorithms. Section 6 presents and evaluates DIO-POWER. Section 7 discusses the related work and Section 8 concludes.

2. CLASSIFICATION SCHEMES

2.1 Methodology

A conventional approach to evaluate new scheduling algorithms is to compare the speedup they deliver relative to a default scheduler. This approach, however, has two potential flaws. First, the schedule chosen by the default scheduler varies greatly based on stochastic events, such as thread spawning order. Second, this approach does not necessarily provide the needed insight into the quality of the algorithms. A scheduling algorithm consists of two components: the information (classification scheme, in our case) used for scheduling decisions and the policy that makes the decisions based on this information. The most challenging part of a cache-aware scheduling algorithm is to select the right classification scheme, because the classification scheme enables the scheduler to predict the performance effects of co-scheduling any group of threads in a shared cache. Our goal was to evaluate the quality of classification schemes separately from any scheduling policies, and only then evaluate the algorithm as a whole. To evaluate classification schemes independently of scheduling policies, we have to use the classification schemes in conjunction with a perfect policy. In this way, we are confident that any differences in the performance between the different algorithms are due to classification schemes, and not to the policy.

2.1.1 *A Perfect Scheduling Policy*. As a perfect scheduling policy, we use an algorithm proposed by Jiang et al. [2008]. This algorithm is guaranteed to find an optimal scheduling assignment, that is, the mapping of threads to cores, on a machine with several clusters of cores sharing a cache as long as the *co-run*

Table I. Co-Run Degradations of Four Obtained on a Real System

	mcf	milc	gamess	namd
mcf	48.01%	65.63%	2.0%	2.11%
milc	24.75%	45.39%	1.23%	1.11%
gamess	2.67%	4.48%	-1.01%	-1.21%
namd	1.48%	3.45%	-1.19%	-0.93%

Small negative degradations for some benchmarks occur as a result of sharing of certain libraries. The value on the intersection of row *i* and column *j* indicates the performance degradation that application *i* experiences when co-scheduled with application.

degradations for applications are known. A co-run degradation is an increase in the execution time of an application when it shares a cache with a co-runner, relative to running solo.

Jiang's methodology uses the co-run degradations to construct a graph theoretic representation of the problem, where threads are represented as nodes connected by edges, and the weights of the edges are given by the sum of the mutual co-run degradations between the two threads. The optimal scheduling assignment can be found by solving a min-weight perfect matching problem. For instance, given the co-run degradations in Table I, Figure 2 demonstrates how Jiang's method would be used to find the best and the worst scheduling assignment. In Table I, the value on the intersection of row *i* and column *j* indicates the performance degradation that application *i* experiences when co-scheduled with application *j*. In Figure 2, edge weights show the sum of mutual co-run degradations of the corresponding nodes. For example, the weight of 90.4% on the edge between MCF and MILC is the sum of 65.63% (the degradation of MCF when co-scheduled with MILC) and 24.75% (the degradation of MILC co-scheduled with MCF).

Although Jiang's methodology and the corresponding algorithms would be too expensive to use online (the complexity of the algorithm is polynomial in the number of threads on systems with two cores per shared cache and the problem is NP-complete on systems where the degree of sharing is larger), it is acceptable for offline evaluation of the quality of classification schemes.

Using Jiang's algorithm as the perfect policy implies that the classification schemes we are evaluating must be suitable for estimating co-run degradations. All of our chosen classification schemes answered this requirement: they can be used to estimate co-run degradations in absolute or in relative terms.

2.1.2 *An Optimal Classification Scheme.* To determine the quality of various classification schemes, we not only need to compare them with each other, but also to evaluate how they measure up to the optimal classification scheme. All of our evaluated classification schemes attempt to approximate relative performance degradation that arbitrary tuples of threads experience when sharing a cache relative to running solo. An optimal classification scheme would therefore have the knowledge of *actual* such degradation, as measured on a real system. To obtain these measured degradations, we selected ten representative benchmarks from the SPEC CPU2006 benchmark suite (the methodology

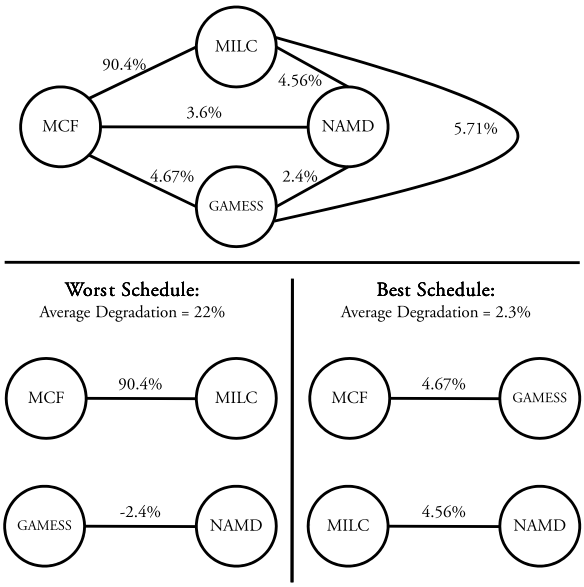


Fig. 2. An overview of using Jiang's method for determining the optimal and the worst thread schedule. Edges connecting nodes are labeled with mutual co-run degradations, that is, the sum of individual degradations for a given pair. The average degradation for a schedule is computed by summing up all mutual degradations and dividing by the total number of applications (four in our case).

for selection is described later in this section), ran them solo on our experimental system (described in detail in Section 5), ran all possible pairs of these applications and recorded their performance degradation relative to solo performance. In order to make the analysis tractable it was performed based on pairwise degradations, assuming that only two threads may share a cache, but the resultant scheduling algorithms are evaluated on systems with four cores per shared cache as well.

2.1.3 *Evaluating Classification Schemes.* To evaluate a classification scheme on a particular set of applications, we follow these steps.

(1) Find the optimal schedule using Jiang's method and the optimal classification scheme, that is, relying on measured degradations. Record its average performance degradation (see Figure 2).
(2) Find the estimated best schedule using Jiang's method and the evaluated classification scheme, that is, relying on estimated degradations. Record its average performance degradation.
(3) Compute the difference between the degradation of the optimal schedule and of the estimated best schedule. The smaller the difference, the better the evaluated classification scheme.

To perform a rigorous evaluation, we construct a large number of workloads consisting of four, eight and ten applications. We evaluate all classification

schemes using this method, and for each classification scheme report the average degradation above the optimal scheme across all workloads.

2.1.4 *Benchmarks and Workloads.* We selected ten benchmarks from the SPEC2006 benchmark suite to represent a wide range of cache access behaviors. The cache miss rates and access rates for every application in the SPEC2006 benchmark suite were obtained from a third-party characterization report [Hoste and Eeckhout 2007] and a clustering technique was employed to select the ten representative applications.

From these ten applications, we constructed workloads for a four-core, six-core, eight-core, and ten-core processor with two cores per LLC. With the ten benchmarks, we selected, there are 210 unique four-application workloads, 210 unique six-application workloads, 45 unique eight-application workloads, and 1 unique ten-application workload to be constructed on each system. There are three unique ways to schedule a four-application workload on a machine with four cores and two shared caches. The number of unique schedules grows to 15, 105, and 945 for the six, eight, and ten-core systems, respectively. Using Jiang's methodology, as opposed to running all these 9450 schedules, saves a considerable amount of time and actually makes it feasible to evaluate such a large number of workloads.

2.2 The Classification Schemes

For any classification scheme to work, it must first obtain some raw data about the applications it will classify. This raw data may be obtained online via performance counters, embedded into an application's binary as a signature, or furnished by the compiler. Where this data comes from has a lot to do with what kind of data is required by the classification scheme. The SDC algorithm proposed by Chandra et al. [2005] is one of the best known methods for determining how threads will interact with each other when sharing the same cache. The SDC algorithm requires the memory reuse patterns of applications, known as stack distance profiles, as input. Likewise, all but one of our classification schemes require stack distance profiles. The one exception is the Miss Rate classification scheme which requires only miss rates as input. The simplicity of the Miss Rate Scheme allowed us to adapt it to gather the miss rates dynamically online making it a far more attractive option than the other classification schemes. However, in order to understand why such a simple classification scheme is so effective as well as to evaluate it against more complex and better established classification schemes we need to explore a wide variety of classification schemes and we need to use stack distance profiles to do so.

A stack distance profile is a compact summary of the application's cache-line reuse patterns. It is obtained by monitoring (or simulating) cache accesses on a system with an LRU cache. A stack distance profile is a histogram with a "bucket" or a position corresponding to each LRU stack position (the total number of positions is equal to the number of cache ways) plus an additional position for recording cache misses. Each position in the stack-distance profile counts the number of hits to the lines in the corresponding LRU stack position. For example, whenever an application reuses a cache line that is at the top of

the LRU stack, the number of hits in the first position of the stack-distance profile is incremented. If an application experiences a cache miss, the number of items in the miss position is incremented. The shape of the stack-distance profile captures the nature of the application's cache behavior: an application with a large number of hits in top LRU positions has a good locality of reference. An application with a low number of hits in top positions and/or a large number of misses has a poor locality of reference. For our study we obtained the stack-distance profiles using the Pin binary instrumentation tool [Luk et al. 2005]; an initial profiling run of an application under Pin was required for that. If stack-distance profiles were to be used online in a live scheduler, they could be approximated online using hardware performance counters [Tam et al. 2009].

We now discuss four classification schemes which are based on the information provided in the stack distance profiles.

2.2.1 *SDC.* The SDC[1] classification scheme was the first that we evaluated, since this is a well known method for predicting the effects of cache contention among threads [Chandra et al. 2005]. The idea behind the SDC method is to model how two applications compete for the LRU stack positions in the shared cache and estimate the extra misses incurred by each application as a result of this competition. The sum of the extra misses from the co-runners is the proxy for the performance degradation of this co-schedule.

The main idea of the SDC algorithm is in constructing a new stack distance profile that merges individual stack distance profiles of threads that run together. On initialization, each individual profile is assigned a current pointer that is initialized to point to the first stack distance position. Then, the algorithm iterates over each position in the profile, determining which of the co-runners will be the winner for this stack-distance position. The co-runner with the highest number of hits in the current position is selected as the winner. The winner's counter is copied into the merged profile, and its current pointer is advanced. After the Ath iteration (A is the associativity of the LLC), the effective cache space for each thread is computed proportionally to the number of its stack distance counters that are included in the merged profile. Then, the cache miss rate with the new effective cache space is estimated for each co-runner, and the degradation of these miss rates relative to solo miss rates is used as a proxy for the co-run degradation. Note that miss rate degradations do not approximate absolute performance degradations, but they provide an approximation of relative performance in different schedules.

2.2.2 *Animal Classes.* This classification scheme is based on the animalistic classification of applications introduced by Xie and Loh [2008]. It allows classifying applications in terms of their influence on each other when co-scheduled in the same shared cache. Each application can belong to one of the four different classes: *turtle* (low use of the shared cache), *sheep* (low miss

[1]Chandra suggested three algorithms for calculating the extra miss rates. However, only two of them (FOA and SDC) are computationally fast enough to be used in the robust scheduling algorithm. We chose SDC as it is slightly more efficient than FOA.

rate, insensitive to the number of cache ways allocated to it), *rabbit* (low miss rate, sensitive to the number of allocated cache ways) and *devil* (high miss rate, tends to thrash the cache thus hurting co-scheduled applications).

We attempted to use this classification scheme to predict contention among applications of different classes, but found an important shortcoming of the original animalistic model. The authors of the animalistic classification proposed that devils (applications with a high miss rate but a low rate of reuse of cached data) must be insensitive to contention for shared resources. On the contrary, we found this not to be the case. According to our experiments, devils were some of the most sensitive applications, that is, their performance degraded the most when they shared the on-chip resources with other applications. Since devils have a high miss rate they issue a large number of memory and prefetch requests. Therefore, they compete for shared resources other than cache: memory controller, memory bus, and prefetching hardware. As will be shown in Section 3, contention for these resources dominates performance, and that is why devils turn out to be sensitive.

To use the animalistic classification scheme for finding the optimal schedule as well as to account for our findings about "sensitive" devils we use a *symbiosis table* to approximate relative performance degradations for applications that fall within different animal classes. The symbiosis table provides estimates of how well various classes co-exist with each other on the same shared cache. For example, the highest estimated degradation (with the experimentally chosen value of 8) will be for two sensitive devils co-scheduled in the same shared cache, because the high miss rate of one of them will hurt the performance of the other one. Two turtles, on the other hand, will not suffer at all. Hence, their mutual degradation is estimated as 0. All other class combinations have their estimates in the interval between 0 and 8.

The information for classification of applications, as described by Xie and Loh [2008], is obtained from stack-distance profiles.

2.2.3 *Miss Rate.* Our findings about sensitive devils caused us to consider the miss rate as the heuristic for contention. Although another group of researchers previously proposed a contention-aware scheduling algorithm based on miss rates [Knauerhase et al. 2008], the hypothesis that the miss rate should explain contention contradicted the models based on stack-distance profiles, which emphasized cache reuse patterns, and thus it needed a thorough validation. We define the miss rate to include all cache line requests issued by the LLC to DRAM. On our Intel Xeon system with L2 as the LLC, this quantity can be measured using the L2_LINES_IN hardware counter.

We hypothesized that identifying applications with high miss rates is beneficial for the scheduler, because these applications exacerbate the performance degradation due to memory controller contention, memory bus contention, and prefetching hardware contention. To attempt an approximation of the best schedule using the miss rate heuristic, the scheduler will identify high miss rate applications and separate them into different caches, such that no one cache will have a much higher total miss rate than any other cache. Since no cache will experience a significantly higher miss rate than any other

cache the performance degradation factors will be stressed evenly throughout the system.

In addition to evaluating a metric based on miss rates, we also experimented with other metrics, which can be obtained online using hardware counters, such as cache access rate and IPC. Miss rate, however, turned out to perform the best among them.

2.2.4 *Pain.* The Pain Classification Scheme is based on two new concepts that we introduce in this work, *cache sensitivity* and *cache intensity*. Sensitivity is a measure of how much an application will suffer when cache space is taken away from it due to contention. Intensity is a measure of how much an application will hurt others by taking away their space in a shared cache. By combining the sensitivity and intensity of two applications, we estimate the "pain" of the given co-schedule. Combining a sensitive application with an intensive co-runner should result in a high level of pain, and combining an insensitive application with any type of co-runner should result in a low level of pain. We obtain sensitivity and intensity from stack distance profiles and we then combine them to measure the resulting pain.

To calculate sensitivity S, we examine the number of cache hits that will most likely turn into misses when the cache is shared. To that end, we assign to the positions in the stack-distance profile *loss probabilities* describing the likelihood that the hits will be lost from each position. Intuitively hits to the Most Recently Used (MRU) position are less likely to become misses than hits to the LRU position when the cache is shared. Entries that are accessed less frequently are more likely to be evicted as the other thread brings its data into the cache; thus we scale the number of hits in each position by the corresponding probability and add them up to obtain the likely extra misses. The resulting measure is the sensitivity value which is shown in equation (Eq. 1). Here $h(i)$ is the number of hits to the ith position in the stack, where $i = 1$ is the MRU and $i = n$ is the LRU for an n-way set associative cache. We use a linear loss probability distribution. As such the probability of a hit in the ith position becoming a miss is $\frac{i}{n+1}$.

$$S = \left(\frac{1}{1+n}\right) \sum_{i=1}^{n} i * h(i). \tag{1}$$

Intensity Z is a measure of how aggressively an application uses the cache. As such, it approximates how much space the application will take away from its co-runner(s). Our approach to measuring intensity is to use the number of last-level cache accesses per one million instructions.

We combine sensitivity S and intensity Z into the Pain metric, which is then used to approximate the co-run degradations required by our evaluation methodology. Suppose we have applications A and B sharing the same cache. Then, the Pain of A due to B approximates the relative performance degradation that A is expected to experience due to B and is calculated as the intensity of B multiplied by the sensitivity of A (Eq. (2)). The degradation of

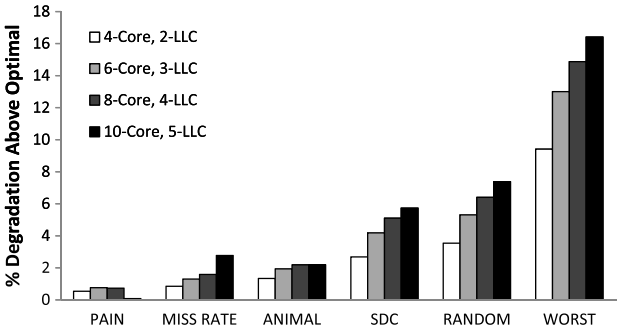


Fig. 3. Degradation relative to optimal experienced by each classification scheme on systems with different numbers of cores.

co-scheduling A and B together is the sum of the Pain of A due to B and the Pain of B due to A (Eq. (3)).

$$Pain(A_B) = S(A) * Z(B) \tag{2}$$

$$Pain(A, B) = Pain(A_B) + Pain(B_A). \tag{3}$$

2.2.5 *Classification Schemes Evaluation.* For the purposes of this work, we collected stack distances profiles offline using Intel's binary instrumentation tool Pin [Hoste and Eeckhout 2007], an add-on module to Pin MICA [Shelepov and Fedorova 2008], and our own module extending the functionality of MICA. The stack distance profiles were converted into the four classification schemes described above: SDC, Pain, Miss rates, and Animal. We estimate the extra degradation above the optimal schedule that each classification scheme produces for the four-core, six-core, eight-core and ten-core systems. Additionally, we present the degradations for the worst and random schedules. A random schedule picks each of the possible assignment for a workload with equal probability.

Figure 3 shows the results of the evaluation. Lower numbers are better. The Pain, Miss Rate and Animal schemes performed relatively well, but SDC surprisingly did only slightly better than random. Pain performed the best, delivering only 1% worse performance than the optimal classification scheme for all the systems. Interestingly we see that all classification schemes except Pain and Animal do worse as the number of cores in the system grows. In systems with more cores, the number of possible schedules grows, and so imperfect classification schemes are less able to make a lucky choice.

The Animal scheme did worse than Pain. Animal classes are a rather rough estimation of relative co-run degradations (a lot of programs will fall to the same class), and so the Animal scheme simply cannot achieve the same precision as Pain which takes into account absolute values. The Miss Rate scheme

performs almost as well as Pain and Animal scheme and yet is by far the easiest to compute either online or offline.

SDC performed worse than Pain and Animal for the following reasons. The first reason is that SDC does not take into account miss rates in its stack distance competition model. So it only works well in those scenarios where the co-running threads have roughly equal miss rates (this observation is made by the authors themselves [Chandra et al. 2005]). When the miss rates of co-running threads are very different, the thread with the higher miss rate will "win" more cache real estate – this fact is not accounted for by the SDC model. Authors of SDC [Chandra et al. 2005] offer a more advanced (but at the same time computationally more expensive) model for predicting extra miss rates which takes into account different miss rates of co-running applications. We did not consider this model in the current work, because we deemed it too computationally expensive to use in an online scheduler.

The second reason has to do with the fact that SDC models the performance effects of cache contention, but as the next section shows, this is not the dominant cause for performance degradation and so other factors must be considered as well. We were initially surprised to find that SDC, a model extensively validated in the past, failed to outperform even such a coarse classification heuristic as the miss rate. In addition to SDC, many other studies of cache contention used stack-distance or reuse-distance profiles for managing contention [Chandra et al. 2005; Qureshi and Patt 2006; Suh et al. 2002; Tam et al. 2009]. The theory behind stack-distance based models seems to suggest that the miss rate should be a poor heuristic for predicting contention, since applications with a high cache miss rate may actually have a very poor reuse of their cached data, and so they would be indifferent to contention. Our analysis, however, showed the opposite: miss rate turned out to be an excellent heuristic for contention.

We discovered that the reason for these seemingly unintuitive results had to do with the causes of performance degradation on multicore systems. SDC, and other solutions relying on stack distance profiles such as cache partitioning [Qureshi and Patt 2006; Suh et al. 2002; Xie and Loh 2008], assumed that the dominant cause of performance degradation is contention for the space in the shared cache, that is, when co-scheduled threads evict each other data from the shared cache. We found, however, that cache contention is by far not the dominant cause of performance degradation. Other factors, such as contention for memory controllers, memory bus, and resources involved in prefetching, dominate performance degradation for most applications. A high miss rate exacerbates the contention for all of these resources, since a high-miss-rate application will issue a large number of requests to a memory controller and the memory bus, and will also be typically characterized by a large number of prefetch requests.

In the next section, we attempt to quantify the causes for performance degradation resulting from multiple factors, showing that contention for cache space is not dominant; these results provide the explanation why a simple heuristic such as the miss rate turns out to be such a good predictor for contention.

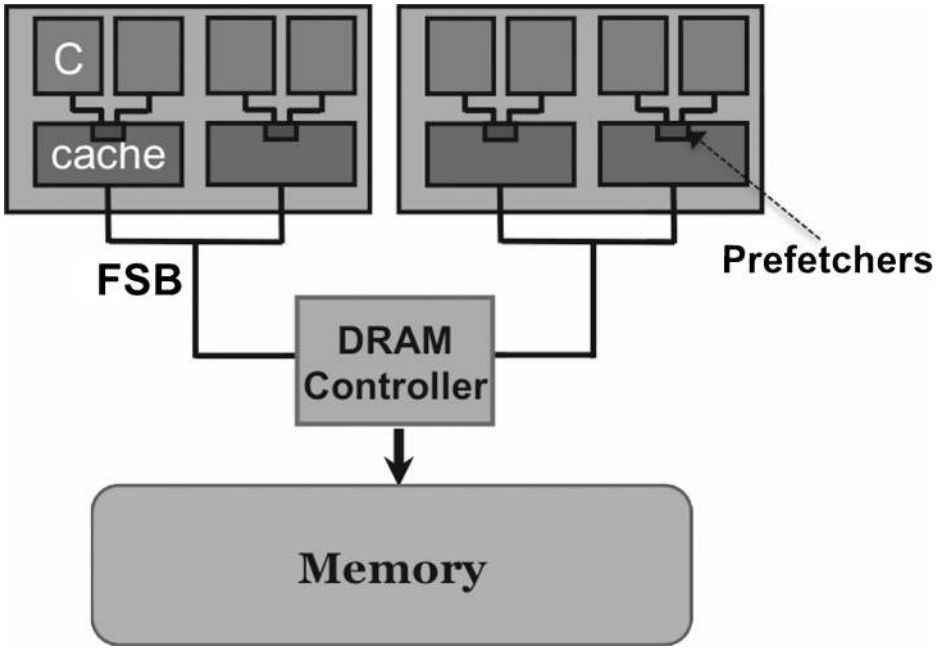


Fig. 4. A schematic view of a system assumed in this section. Each of 4 chips has 2 cores sharing a LLC cache. Chips are grouped into 2 sockets (physical packages). All cores are equidistant to the main memory.

3. FACTORS CAUSING PERFORMANCE DEGRADATION ON MULTICORE SYSTEMS

Recent work on the topic of performance degradation in multicore systems focused on contention for cache space and the resulting data evictions when applications share the LLC. However, it is well known that cache contention is far from being the only factor that contributes to performance degradation when threads share an LLC. Sharing of other resources, such as the memory bus, memory controllers and prefetching hardware also plays an important role. Through extensive analysis of data collected on real hardware we have determined that contention for space in the shared cache explains only a part of the performance degradation when applications share an LLC. In this section we attempt to quantify how much performance degradation can be attributed to contention for each shared resource assuming a system depicted on Figure 4 (eight Intel Xeon X5365 cores running at 3GHz, and 8GB of RAM).

Estimating the contribution that each factor has on the overall performance degradation is difficult, since all the degradation factors work in conjunction with each other in complicated and practically inseparable ways. Nevertheless, we desired a rough estimate of the degree to which each factor affects overall performance degradation to identify if any factor in particular should be the focus of our attention since mitigating it will yield the greatest improvements.

We now describe the process we used to estimate the contributions of each factor to the overall degradation. Our experimental system is a two-sockets server with two Intel X5365 "Clovertown" quad-core processors. The two sockets share the memory controller hub, which includes the DRAM controller. On each socket there are four cores sharing a front-side bus (FSB). There are two L2 caches on each socket, one per pair of cores. Each pair of cores also shares prefetching hardware, as described below. So when two threads run on different sockets, they compete for the DRAM controller. When they run on the same socket, but on different caches, they compete for the FSB, in addition to the DRAM controller. Finally, when they run on cores sharing the same cache, they also compete for the L2 cache and the prefetching hardware, in addition to the FSB and the DRAM controller. To estimate how contention for each of these resources contributes to the total degradation, we measured the execution times of several benchmarks under the following eight conditions.

Solo_PF_ON: Running SOLO and prefetching is ENABLED
Solo_PF_OFF: Running SOLO and prefetching is DISABLED
SameCache_PF_ON: Sharing the LLC with an interfering benchmark and prefetching is ENABLED
SameCache_PF_OFF: Sharing the LLC with an interfering benchmark and prefetching is DISABLED
DiffCache_PF_ON: An interfering benchmark runs on a different LLC but on the same socket and prefetching is ENABLED
DiffCache_PF_OFF: An interfering benchmark runs on a different LLC but on the same socket and prefetching is DISABLED
DiffSocket_PF_ON: An interfering benchmark runs on a different socket and prefetching is ENABLED
DiffSocket_PF_OFF: An interfering benchmark runs on a different socket and prefetching is DISABLED

As an interfering benchmark for this experiment we used MILC. MILC was chosen for several reasons. First, it has a very high solo miss rate which allows us to estimate one of the worst-case contention scenarios. Second, MILC suffers a negligible increase in its own miss rate due to cache contention (we determined this via experiments and also by tracing MILC's memory reuse patterns, which showed that MILC hardly ever reuses its cached data) and hence will not introduce extra misses of its own when co-run with other applications. We refer to MILC as the *interfering benchmark* and we refer to the test application simply as the *application*. In the experiment where MILC is the test application, SPHINX is used as the interfering benchmark.

Estimating Performance Degradation due to DRAM Controller Contention. We look at the difference between the solo run and the run when the interfering benchmark is on a different socket. When the interfering benchmark is on a different socket any performance degradation it causes can only be due to DRAM controller contention since no other resources are shared. Equation (4)

shows how we estimate the performance degradation due to DRAM controller contention.

$$DRAM\_contention = \frac{DiffSocket\_PF\_OFF - Solo\_PF\_OFF}{Solo\_PF\_OFF} \quad (4)$$

There are several complications with this approach, which make it a rough estimate as opposed to an accurate measure of DRAM controller contention. First, when the LLC is shared by two applications, extra evictions from cache cause the total number of misses to go up. These extra misses contribute to the DRAM controller contention. In our experimental technique, the two applications are in different LLCs and hence there are no extra misses. As a result, we are underestimating the DRAM controller contention. Second, we chose to disable prefetching for this experiment. If we enabled prefetching and put two applications into different LLC, then they would each have access to a complete set of prefetching hardware. This would have greatly increased the total number of requests issued to the memory system from the prefetching hardware as compared to the number of requests that can be issued from only one LLC. By disabling the prefetching, we are once again underestimating the DRAM controller contention. As such, the values that we measure should be considered a lower bound on DRAM controller contention.

Estimating Performance Degradation due to FSB Contention. Next, we estimate the degree of performance degradation due to contention for the FSB. To that end, we run the application and the interfering benchmark on the same socket, but on different LLCs. This is done with prefetching disabled, so as not to increase the bus traffic. Equation (5) shows how we estimate the degradation due to FSB contention.

$$FSB\_Contention = \frac{DiffCache\_PF\_OFF - DiffSocket\_PF\_OFF}{Solo\_PF\_OFF} \quad (5)$$

Estimating Performance Degradation due to Cache Contention. To estimate the performance degradation due to cache contention we take the execution time when an application is run with an interfering co-runner in the same LLC and subtract from it the execution time of the application running with the interfering benchmark in a different LLC of the same socket. This is done with prefetching disabled so as not to increase bus traffic or contend for prefetching hardware. The difference in the execution times between the two runs can be attributed to the extra misses that resulted due to cache contention. Equation (6) demonstrates how we estimate performance degradation due to cache contention.

$$Cache\_Contention = \frac{SameCache\_PF\_OFF - DiffCache\_PF\_OFF}{Solo\_PF\_OFF} \quad (6)$$

Estimating Performance Degradation due to Contention for Resources Involved in Prefetching. Contention for resources involved in prefetching has received less attention in literature than contention for other resources.

We were compelled to investigate this type of contention when we observed that some applications experienced a decreased prefetching rate (up to 30%) when sharing an LLC with a memory-intensive co-runner. Broadly speaking, prefetching resources include all the hardware that might contribute to the speed and quality of prefetching. For example, our experimental processor has two types of hardware that prefetches into the L2 cache. The first is the Data Prefetching Logic (DPL) which is activated when an application has two consecutive misses in the LLC and a stride pattern is detected. In this case, the rest of the addresses up to the page boundary are prefetched. The second is the adjacent cache line prefetcher, or the streaming prefetcher. The L2 prefetching hardware is dynamically shared by the two cores using the LLC. The memory controller and the FSB are also involved in prefetching, since they determine how aggressively these requests can be issued to memory. It is difficult to tease apart the latencies attributable to contention for each resource, so our estimation of contention for prefetching resources includes contention for prefetching hardware as well as additional contention for these two other resources. This is an upper bound on the contention for the prefetching hardware itself.

We can measure the performance degradation due to prefetching-related resources as the difference between the total degradation and the degradation caused by cache contention, FSB, and DRAM controller contention. Equation (7) calculates the total degradation of an application when the LLC is shared by looking at the difference when the interfering benchmark shares the LLC and when the application runs alone. Equation (8) shows the calculation of the prefetching degradation.

$$Total\_Degradation = \frac{SameCache\_PF\_ON - Solo\_PF\_ON}{Solo\_PF\_ON} \tag{7}$$

$$Prefetching\_Contention = Eq.(7) - Eq.(6) - Eq.(5) - Eq.(4) \tag{8}$$

Finally, we calculate the degradation contribution of each factor as the ratio of its degradation compared to the total degradation. Figure 5 shows the percent contribution of each factor (DRAM controller contention, FSB contention, L2 cache contention, and prefetching resource contention) to the total degradation for six SPEC2006 benchmarks.

The six applications shown in Figure 5 are the applications that experience a performance degradation of at least 45% chosen from the ten representative benchmarks. We see from Figure 5 that for all applications except SPHINX contention for resources other than shared cache is the dominant factor in performance degradation, accounting for more than 50% of the total degradation.

In order to understand whether similar conclusions can be made with respect to other systems, we also performed similar analysis of degradation-contributing factors on a system that is very different from the Intel system examined so far. We used an AMD Opteron 2350 (Barcelona) system. The

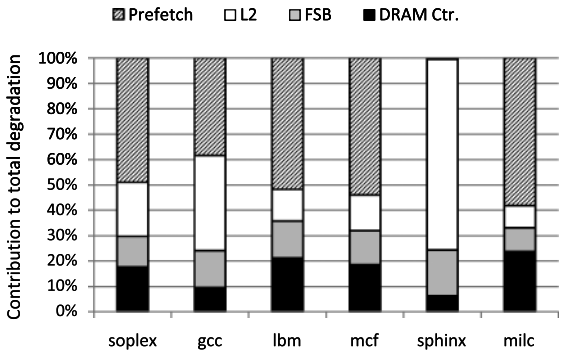


Fig. 5. Percent contribution that each of the factors have on the total degradation on a UMA multicore system.

AMD machine has four cores per processor, each core with private L1 instruction and data caches and a private unified L2 cache. There is an 2MB L3 cache shared by all four cores on the chip.

Since there are important differences between the Intel and AMD systems, the degradation-contributing factors are also somewhat different. For instance, while the Intel system is UMA (UMA stands for Uniform Memory Access), the AMD system is NUMA (Non-Uniform Memory Access). This means that each processor on an AMD system has its own DRAM controller and a local memory node. While any core can access both local and remote nodes (attached to other processors), local accesses take less time. Remote accesses take longer and require using the DRAM controller of the remote processor. Because of these differences, there are two additional degradation-contributing factors on this AMD system that were not present on the Intel system: (1) interconnect contention (IC), which would occur if a thread is accessing remote memory and competes for inter-processor interconnects with other threads, and (2) remote latency overhead (RL), which occurs if a thread is accessing remote memory and experiences longer wire delays.[2] Furthermore, prefetching works differently on the AMD system, so we were unable to isolate its effects in the same way that we were able to do on the Intel system. In summary, we identified four performance-degrading factors on the NUMA AMD system that a thread can experience relative to the scenario when it runs on the system alone and accesses the memory in a local node: (1) L3 cache contention, (2) DRAM controller contention, (3) interconnect contention and (4) remote latency. Experiments similar to those that we designed on the Intel system were used to estimate the effect of these factors on performance. One difference is that we used different benchmarks in order to capture the representative trends in degradation breakdown. We provide the detailed explanation of our methodology for estimating the breakdown of degradation factors on the AMD system in the Appendix.

[2]Note that under our definition, RL does not include IC.

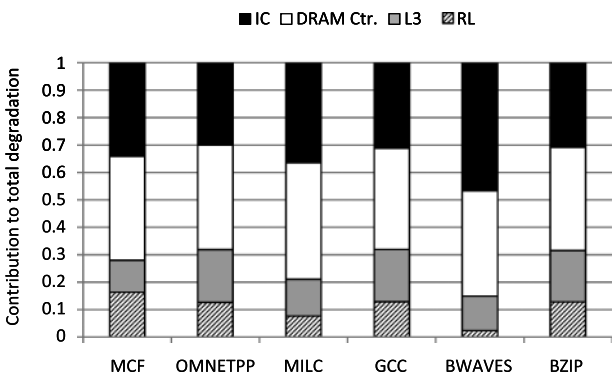


Fig. 6. Contribution of each factor to the worst-case performance degradation on a NUMA multicore system.

Figure 6 shows the contribution of each factor to the performance degradation of the benchmarks as they are co-scheduled with three instances of MILC. Although this system is different from the Intel system examined earlier, the data allows us to reach similar conclusions: shared cache contention (L3) is *not* the dominant cause for performance degradation. The dominant causes are the DRAM controller and interconnect contention.

As in case with the UMA system, this is only an approximation, since contention causing factors on a real system overlap in complex and integrated ways. For example, increasing cache contention increases other types of contention: if two threads share a cache and have a memory access patterns that result in a lot of extra cache misses, that would stress all the memory hierarchy levels placed on the path between the shared cache and main memory of the machine. The results provided are an approximation that is intended to direct attention to the true bottlenecks in the system.

3.1 Discussion of Performance-Degrading Factors Breakdown

While cache contention does have an effect on performance degradation, any scheduling strategy that caters to reducing cache contention exclusively cannot and will not have a major impact on performance. The fact that contention for resources other than cache is dominant, explains why the miss rate turns out to be such a good heuristic for predicting contention. The miss rate, which we define to include all DRAM-to-LLC transfers, highly correlates with the amount of DRAM controller, FSB, and prefetch requests as well as the degree of interconnect usage and the sensitivity to remote access latency, and thus is indicative of both the sensitivity of an application as well as its intensity.

4. SCHEDULING ALGORITHMS

A scheduling algorithm is the combination of a classification scheme and a scheduling policy. We considered and evaluated several scheduling algorithms

that combined different classification schemes and policies, and in this section, we present those that showed the best performance and were also the simplest to implement. In particular, we evaluate two algorithms based on the Miss Rate classification schemes, because the miss rate is very easy to obtain online via hardware performance counters. The scheduling policy we used was the *Centralized Sort*. It examines the list of applications, sorted by their miss rates, and distributes them across cores, such that the total miss rate of all threads sharing a cache is equalized across all caches. For the evaluation results of other heuristics and policies, we refer the reader to our technical report [Blagodurov et al. 2009].

While the Pain classification scheme gave the best performance, we chose not to use it in an online algorithm, instead opting to implement one using the miss rate heuristic. This was done to make the scheduler simpler, thus making more likely that it will be adopted in general-purpose operating systems. Using Pain would require more changes to the operating system than using the miss rate for the following reason. Pain requires stack distance profiles. Obtaining a stack distance profile online requires periodic sampling of data addresses associated with last-level cache accesses using advanced capabilities of hardware performance monitoring counters, as in RapidMRC [Tam et al. 2009]. Although RapidMRC can generate accurate stack-distance profiles online with low overhead, there is certain complexity associated with its implementation. If a scheduler uses the miss rate heuristic, all it has to do is periodically measure the miss rates of the running threads, which is simpler than collecting stack-distance profiles. Given that the miss rate heuristic had a much lower implementation complexity but almost the same performance as Pain, we thought it would be the preferred choice in future OS schedulers.

In the rest of this section, we describe two algorithms based on the Miss Rate classification scheme: *Distributed Intensity* (DI) and *Distributed Intensity Online* (DIO). DIO does not rely on stack-distance profiles, but on the miss rates measured online. DI estimates the miss rate based on the stack-distance profiles; it was evaluated in order to determine if any accuracy is lost when the miss rates are measured online as opposed to estimated from the profiles.

4.1 Distributed Intensity (DI)

In the Distributed Intensity (DI) algorithm all threads are assigned a value which is their solo miss rate (misses per one million instructions) as determined from the stack distance profile. The goal is then to spread the threads across the system such that the miss rates are distributed as evenly as possible. The idea is that the performance degradation factors identified in Section 3 are all exacerbated by a high miss rate and so we avoid the situation where the concentration of threads with high cumulative miss rate results in increased contention for shared memory resources with the consequent performance bottleneck on the system.

The algorithm uses the notion of *memory hierarchy entities* when trying to spread the workload's intensiveness across the system. Memory hierarchy

Table II. Entities on each Level of Memory Hierarchy

level of memory hierarchy	type of entity on this level	entities per machine / entities per container	
		Intel Xeon	AMD Barcelona
3	core	8/2	8/4
2	chip	4/2	2/1
1	physical package	2/2	
0	machine	1/-	

entities are distinct hardware modules (e.g., cores, chips, packages) each of which is located on its own level of memory hierarchy. The entities on the upper level are grouped into the larger entities of the lower level which are called *containers*. For example, several cores are organized into a chip (chip serves as a container for cores), whereas several chips form a physical package (the package is a container for chips).

The presence of the requested data is first checked against the very fast and small caches that are located close to the core itself. These caches are typically devoted exclusively to its closest core (and are called *private* because of that), hence they can be considered the same entity with the core. If the data is missing in the private caches, it is then checked in the larger on-chip cache shared between all cores on the chip. The latency of servicing the request from the shared on-chip cache is higher than from the private caches. However, the shared cache is still significantly faster than the main memory. If the data is absent in the shared cache, the request goes down the memory hierarchy into the Front Side Bus that is shared between several chips on the same physical package (Figure 4). There could be many physical packages on the system. Each of the packages in this case will have its own FSB shared between all the chips belonging to this package. The requests from each of the on-package buses then compete for the access of the DRAM controller. On the UMA system there is only one such controller per machine. However, the NUMA systems have several DRAM controllers, one per every memory node on the system.

For example, on our testing machines equipped with Intel Xeon and AMD Opteron processors (described in Sections 3 and 5.1) the entities on each level of the memory hierarchy are defined as shown in Table II. During the initialization stage (at boot), the system determines the number of memory hierarchy levels and the number of distinct entities on each level (as is specified in *proc* pseudo filesystem at /sys/devices/system/cpu). DI (described in blocks of pseudo code 1 and 2) then tries to even out the miss rate on all levels of memory hierarchy.

To further justify the validity of this method, we performed a study on all possible 4-thread workloads that can be constructed from the 10 representative SPEC2006 benchmarks. We computed the percent difference between average co-run degradations for each workload achieved with the optimal solution relative to the worst solution (we refer to this value as the speedup). The highest average speedup relative to the worst schedule was 25%, and the lowest was less than 1%. We broke up the workloads based on the speedup range into which they fit: (25%-15%), (15%-10%), (10%-5%), and (5%-0%), and studied which types of applications are present in each range. For simplicity we

Algorithm 1 Invocation at the beginning of every scheduling interval

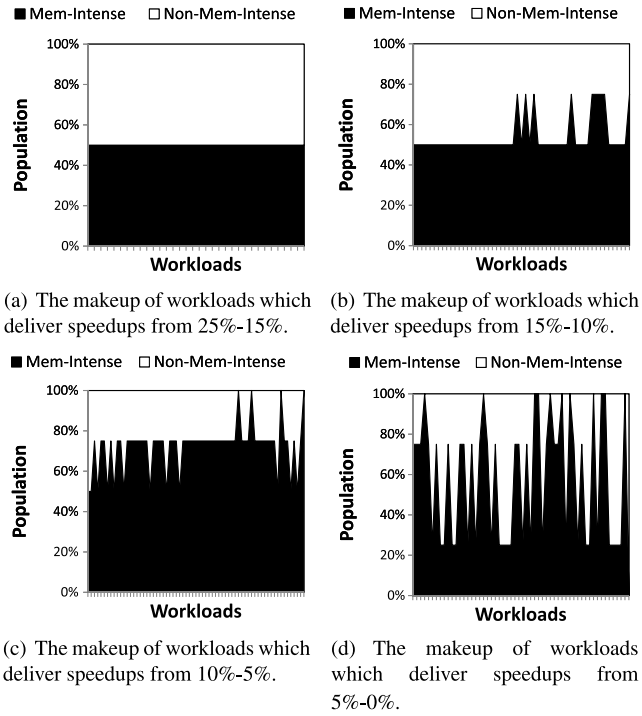
```
1: // initialization of the boolean global array Order[l], every member specifies the order of browsing entities on the corresponding level of the memory hierarchy. Each entity has its ID that was given to it by OS. Order[l]=0 means that entities are browsed from the smallest ID to the biggest, Order[l]=1 – vice-versa.
2: for l = 0; l < number_of_memory_hierarchy_levels >; l++ do
3:    Order[l]:=0;
4: end for
5: sort the threads according to the miss rate in descending order
6: let T be the array of sorted threads
7: // spread the threads across the machine
8: while T ≠ ∅ do
9:    take the first (the most aggressive) t ∈ T
10:   // invoke DI() to assign the thread
11:   DI(t, < machine >, 0)
12: end while
```

Algorithm 2 DI (thread to schedule t, container $e\_//parent$, memory hierarchy level of the container l)

```
1: Let E_all be the array of entities on hierarchy level l + 1
2: Let E_children be the array of entities on hierarchy level l+1 whose container is e_parent
3: browse the entities in E_children in order Order[l+1] and determine the first entity with the minimum number of allocated threads: e_min ∈ E_children
4: if e_min is a < core > then
5:    // we have reached the bottom of the memory hierarchy
6:    assign thread t to core e_min
7: else
8:    increment the number of threads allocated to e_min
9:    // recursively invoke DI() to assign the thread on the lower hierarchy level
10:   DI(t,e_min,l + 1)
11: end if
12: if number of threads allocated to each entity e ∈ E_all is the same then
13:   // reverse the order of browsing on this level
14:   Order[l + 1] := NOT Order[l + 1]
15: end if
```

define two categories of applications: intensive (above average miss rate), and nonintensive (below average miss rate).

We note that according to the ideas of DI, workloads consisting of two intensive and two non-intensive applications should achieve the highest speedups. This is because in the worst case these workloads can be scheduled in such a way as to put both intensive applications in the same cache, creating a cache with a large miss rate, and a performance bottleneck. Alternatively, in the best case these workloads can be scheduled to spread the two intensive applications across caches thus minimizing the miss rate from each cache and avoiding bottlenecks. The difference in performance between these two cases (the one with a bottleneck and the one without) should account for the significant speedup of



(a) The makeup of workloads which deliver speedups from 25%-15%.

(b) The makeup of workloads which deliver speedups from 15%-10%.

(c) The makeup of workloads which deliver speedups from 10%-5%.

(d) The makeup of workloads which deliver speedups from 5%-0%.

Fig. 7. The makeup of workloads.

the workload. We further note that workloads consisting of more than two or fewer than two intensive applications can also benefit from distribution of miss rates but the speedup will be smaller, since the various scheduling solutions do not offer such a stark contrast between creating a major bottleneck and almost entirely eliminating it.

Figure 7 shows the makeup of workloads (intensive vs. nonintensive) applications and the range of speedups they offer. The (unlabeled) x-axis identifies all the workloads falling into the given speedup range. We see that the distribution of applications validates the claims of DI. The other claim that we make to justify why DI should work is that miss rates of applications are relatively stable. What we mean by stability is that when an application shares the LLC with a co-runner its miss rate will not increase so dramatically as to make the solution found by DI invalid.

DI assigns threads to caches to even out the miss rate across all the caches. This assignment is done based on the solo miss rates of applications. The real miss rate of applications will change when they share a cache with a co-runner, but we claim that these changes will be relatively small such that the miss rates are still rather even across caches. Consider Figure 8, which shows the solo miss rates of the 10 SPEC2006 benchmarks as well as the largest miss rate observed for each application as it was co-scheduled to share a cache with all other applications in the set.

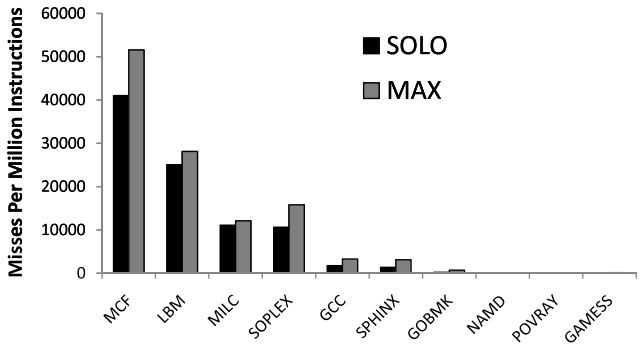


Fig. 8. The solo and maximum miss rate recorded for each of the 10 SPEC2006 benchmarks.

We see that if the applications were sorted based on their miss rates their order would be nearly identical if we used solo miss rates, maximum miss rates, or anything in between. Only the applications MILC and SOPLEX may exchange positions with each other or GCC and SPHINX may exchange positions depending on the miss rate used. The DI algorithm makes scheduling decisions based on the sorted order of applications. If the order of the sorted applications remains nearly unchanged as the miss rates changes, then the solutions found by DI would also be very similar. Hence, the solution found by DI with solo miss rates should also be very good if the miss rates change slightly. Through an extensive search of all the SPEC2006 benchmark suite and the PARSEC benchmark suite, we have not found any applications whose miss rate change due to LLC contention would violate the claim made above.

The DI scheduler is implemented as a user level scheduler running on top of Linux. It enforces all scheduling decisions via system calls which allow it to bind threads to cores. The scheduler also has access to files containing the solo miss rates. For all the applications, it uses solo miss rate estimated using stack distance profiles as the input to the classification scheme.

4.2 Distributed Intensity Online (DIO)

DIO is based on the same classification scheme and scheduling policies as DI except that it obtains the miss rates of applications dynamically online via performance counters. This makes DIO more attractive since the stack distance profiles, which require extra work to obtain online, are not required. The miss rate of applications can be obtained dynamically online on almost any machine with minimal effort. Furthermore, the dynamic nature of the obtained miss rates makes DIO more resilient to applications that have a change in the miss rate due to LLC contention. DIO continuously monitors the miss rates of applications and thus accounts for phase changes. To minimize migrations due to phase changes of applications, we collect miss rate data not more frequently than once every billion cycles and we use a running average for scheduling decisions. Every billion cycles DIO measures the new miss rate and re-evaluates the thread assignments based on the updated miss rate running average values for the workload.

The DIO scheduler, like DI, manages the assignment of threads to cores using affinity interfaces provided in Linux. As such, it mirrors the actions that would be performed by a kernel scheduler. The key difference is that the kernel scheduler would directly manipulate the runqueues in order to place a thread on a particular core, but a user-level prototype of the scheduler uses affinity-related system calls for that purpose. For example, to swap thread A on core i with thread B on core j we set affinity of A to j and affinity B to i. The kernel performs the actual swapping.

5. EVALUATION ON REAL SYSTEMS

5.1 Evaluation Platform

We performed the experiments on two systems:

Dell-Poweredge-2950 (*Intel Xeon X*5365) has eight cores placed on four chips. Each chip has a 4MB 16-way L2 cache shared by its two cores. Each core also has private L1 instruction and data caches. In our first series of experiments we used only two chips out of four. This enabled us to verify our analytical results for the 4 thread workloads directly. After that, all eight cores with eight-thread workloads were used.

*Dell-Poweredge-R*805 (*AMD Opteron* 2350 *Barcelona*) has eight cores placed on two chips. Each chip has a 2MB 32-way L3 cache shared by its four cores. Each core also has a private unified L2 cache and private L1 instruction and data caches. All eight cores with eight thread workloads were used. Although this is a NUMA system, we did not study NUMA effects in detail and did not set up our experiments to explicitly stress NUMA effects. Therefore, addressing contention on NUMA systems must be investigated in future work.

The experimental workloads for Section 5.2 were comprised of the 14 scientific benchmarks from SPEC CPU 2006 suite chosen using the clustering technique as described in Section 2 (see Table III). For the eight-core experiments we created eight-thread workloads by doubling the corresponding four-thread workloads. For example, for the four-thread workload (SOPLEX, SPHINX, GAMESS, NAMD), the corresponding eight-thread workload is (SOPLEX, SPHINX, GAMESS, NAMD, SOPLEX, SPHINX, GAMESS, NAMD). The user-level scheduler starts the applications and binds them to cores as directed by the scheduling algorithm.

We chose to use LAMP, a solution stack of open source software consisting of Linux, Apache, MySQL and PHP, as our testing environment in Section 5.3. LAMP is widely used in many areas where efficient storage and retrieving of data is required, for example, website management and data mining.

Since we are focused on CPU-bound workloads, which are *not* likely to run with more threads than cores [An Mey et al. 2007; van der Pas 2005], we only evaluate the scenarios where the number of threads does not exceed the number of cores. If the opposite were the case, the scheduler would simply re-evaluate the mapping of threads to cores every time the set of running threads changes. The decision of which thread is selected to run would be made in the same way as it is done by the default scheduler. While in this case there are

Table III. The Workloads Used for Experiments (Devils Are Highlighted in Bold)

	Workloads			
	2 memory-bound, 2 CPU-bound			
1	**SOPLEX**	**SPHINX**	GAMESS	NAMD
2	**SOPLEX**	**MCF**	GAMESS	GOBMK
3	**MCF**	**LIBQUANTUM**	POVRAY	GAMESS
4	**MCF**	**OMNETPP**	H264	NAMD
5	**MILC**	**LIBQUANTUM**	POVRAY	PERL
	1 memory-bound, 3 CPU-bound			
6	**SPHINX**	GCC	NAMD	GAMESS
	3 memory-bound, 1 CPU-bound			
7	**LBM**	**MILC**	**SPHINX**	GOBMK
8	**LBM**	**MILC**	**MCF**	NAMD

also opportunities to separate competing threads *in time* as opposed to in space, we do not investigate these strategies in this work.

Both systems were running Linux Gentoo 2.6.27 release 8. We compare performance under DI and DIO to the default contention-unaware scheduler in Linux, referring to the latter as DEFAULT. The prefetching hardware is fully enabled during these experiments. To account for the varied execution times of benchmark we restart an application as soon as it terminates (to ensure that the same workload is running at all times). An experiment terminates when the longest application had executed three times.

5.2 Results for Scientific Workloads

Intel Xeon 4 Cores. We begin with the results for the four-thread workloads on the four-core configuration of the Intel Xeon machine. For every workload, we first run the three possible unique schedules and measure the aggregate workload completion time of each. We then determine the schedule with the optimal (minimal) completion time, the worst possible schedule (maximum completion time) and the expected completion time of the random scheduling algorithm (it selects all schedules with equal probability). We then compared the aggregate execution times of DI and DIO with the completion times of OPTIMAL, WORST and RANDOM. We do not present results for the default Linux scheduler because when the scheduler is given a processor affinity mask to use only four cores out of eight, the migrations of threads across cores become more frequent than when no mask is used, leading to results with an atypically high variance. (We do report results under default Linux when we present our experiments using all eight cores.) Figure 9 shows the performance degradation above the optimal for every workload with DI, DIO, RANDOM and WORST. The results show that DI and DIO perform better than RANDOM and are within 2% of OPTIMAL.

Intel Xeon 8 Cores. Since this setup does not require a processor affinity mask, we evaluated the results of DI and DIO against DEFAULT as in this case DEFAULT does not perform excessive migrations. Figure 10 shows the percent aggregate workload speedup (the average of speedups of all the programs in the workload) over DEFAULT.

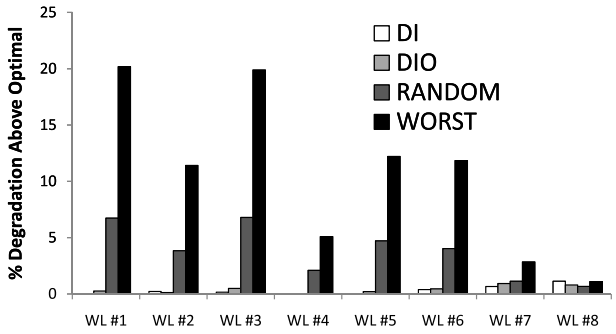


Fig. 9. Aggregate performance degradation of each workload with DI, DIO, RANDOM and WORST relative to OPTIMAL (low bars are good) for the Intel machine and 4 threads.

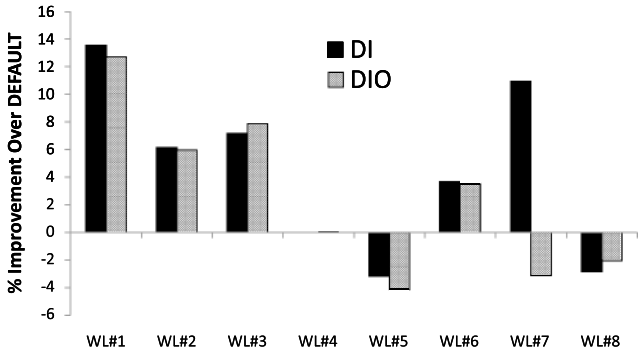


Fig. 10. Aggregate performance improvement of each workload with DI and DIO relative to DEFAULT (high bars are good) on the Intel system.

We note that although generally DI and DIO improve aggregate performance over DEFAULT, in a few cases they performed slightly worse. However, the biggest advantage of DI and DIO is that they offer much more stable results from run to run and avoid the worst-case thread assignment. This effect is especially significant if we look at performance of individual applications. Figure 12(a) shows relative performance improvement for individual applications of the worst-case assignments of DI and DIO over the worst case assignments under DEFAULT. Higher numbers are better. Worst-case performance improvement is obtained by comparing the worst-case performance (across all the runs) under DI and DIO with the worst-case performance under DEFAULT. The results show that DEFAULT consistently stumbles on much worse solutions than DI or DIO and as such there are cases when the performance of individual applications is unpredictably bad under DEFAULT. What this means is that if an application was repeatedly executed on a multicore system under the default scheduler, it could occasionally slowdown by as much as 100%(!) relative to running solo. With DIO, on the other hand, the slowdown would be much smaller and more predictable.

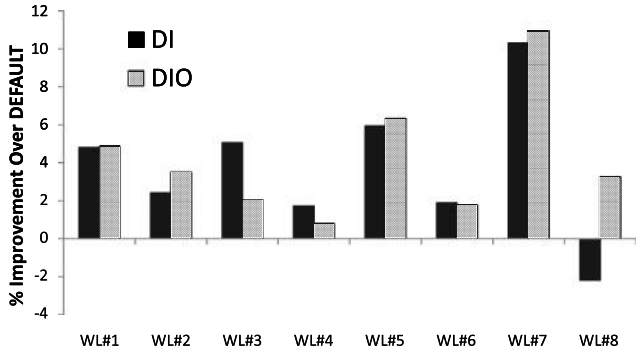
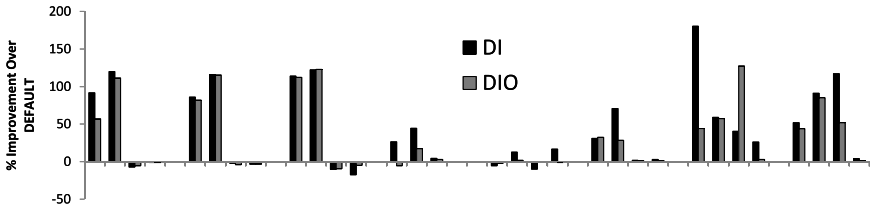


Fig. 11. Aggregate performance improvement of each workload with DI and DIO relative to DEFAULT (high bars are good) on the AMD system.

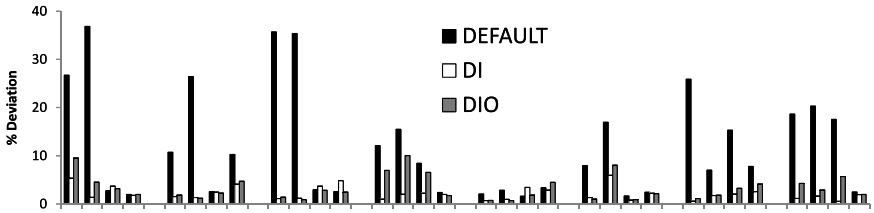
Figure 12(b) shows the deviation of the execution time of consecutive runs of the same application in the same workload with DI, DIO and DEFAULT. We note that DEFAULT has a much higher deviation from run to run than DI and DIO. DIO has a slightly higher deviation than DI as it is sensitive to phase changes of applications and as a result tends to migrate applications more frequently.

AMD Opteron 8 Cores. Finally, we report the results for the same eight-thread workloads on the AMD system. The results for the percent aggregate workload speedup over DEFAULT (Figure 11), relative performance improvement of the worst case assignments over DEFAULT (Figure 12(c)) and the deviation of the execution times (Figure 12(d)) generally repeat the patterns observed on the Intel Xeon machine with eight threads.

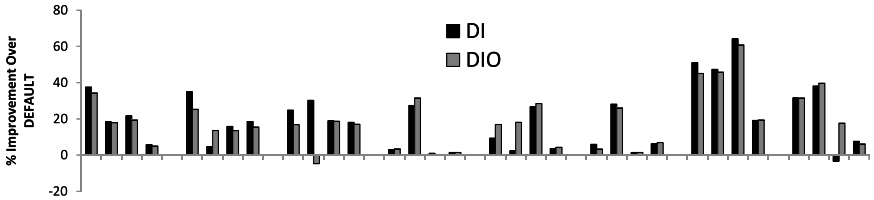
We draw several conclusions from our results. First of all, the classification scheme based on miss rates effectively enables to reduce contention for shared resources with scheduling. Furthermore, an algorithm based on this classification scheme can be effectively implemented online as demonstrated by our DIO prototype. Using contention-aware scheduling can help improve overall system efficiency by reducing completion time for the entire workload as well as reduce worst-case performance for individual applications. In the former case, DIO improves performance by up to 13% relative to DEFAULT and in the isolated cases where it does worse than DEFAULT, the impact on performance is at most 4%, far smaller than the corresponding benefit. On average, if we examine performance across all the workloads we have tried DEFAULT does rather well in terms of workload-wide performance – in the worst case it does only 13% worse than DIO. But if we consider the variance of completion times and the effect on individual applications, the picture changes significantly. DEFAULT achieves a much higher variance and it is likely to stumble into much worse worst-case performance for individual applications. This means that when the goal is to deliver QoS, achieve performance isolation or simply prioritize individual applications, contention-aware scheduling can achieve much larger performance impacts, speeding up individual applications by as much as a factor of two.



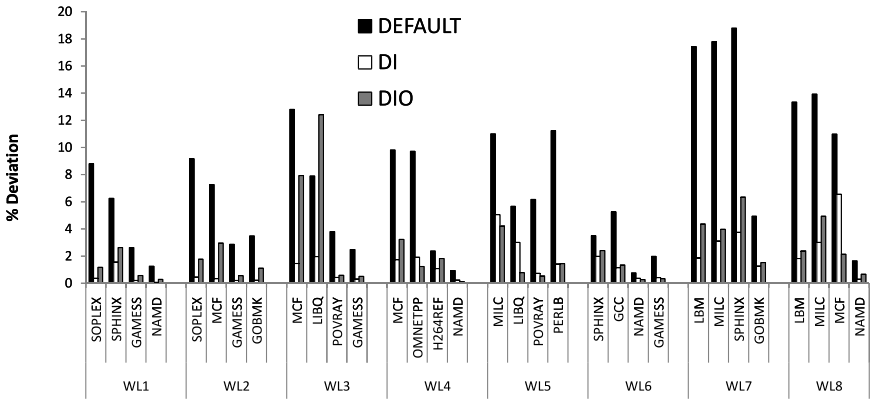
(a) Relative performance improvement of the worst case DI and DIO over the worst case DEFAULT on the Intel system (workload labels are provided in subfigure d).



(b) Deviation with DI, DIO and Default (low bars are good) on the Intel system.



(c) Relative performance improvement of the worst case DI and DIO over the worst case DEFAULT on the AMD system.



(d) Deviation with DI, DIO and Default (low bars are good) on the AMD system.

Fig. 12. Relative performance improvement and deviation.

To understand why DEFAULT performs relatively well on average, let us discuss several examples. Consider a four core machine where each pair of cores shares a cache. If the workload to be executed on this machine involves two intensive applications and two nonintensive applications and if the threads

are mapped randomly to cores (which is a good approximation for DEFAULT) then there is *only* a 1/3 probability of stumbling onto the *worst* solution where the intensive applications share the same cache. If there are three intensive applications in the workload and only one nonintensive application, then all mappings are relatively equivalent *on average* since two of the intensive applications will experience performance degradation and one will not (the one paired with a nonintensive one). Similarly, workloads with no intensive applications, one intensive applications, and all intensive applications show no real difference between solutions. As such, DEFAULT is able to perform well on average. Therefore, we believe that future research must focus on the performance of individual threads, which can vary greatly under the DEFAULT scheduler, as opposed to trying to improve average performance. This point is further highlighted in our experiments with the LAMP workloads.

5.3 Results for the LAMP Workloads

In the previous section, we showed the effectiveness of Distributed Intensity Online for several scientific workloads consisting of the benchmarks from SPEC CPU 2006 suite. In this section, we demonstrate performance improvements that DIO can provide to real multithreaded applications from the LAMP stack.

The main data processing in LAMP is done by Apache HTTP server and MySQL database engine. Both are client-server systems: several clients (website visitors in case of Apache) are concurrently connecting to the server machine to request a webpage or process the data in the database. The server management daemons *apache2* and *mysqld* are then responsible for arranging access to the website scripts and database files and performing the actual work of data storing/retrieval. In our experimental setup we initiated four concurrent web requests to the Apache 2.2.14 web server equipped with PHP version 5.2.12 and four remote client connections to the database server running MySQL 5.0.84 under OS Linux 2.6.29.6. Both apache2 and mysqld are multithreaded applications that spawn one new distinct thread for each new client connection. This client thread within a daemon is then responsible for executing the client‘s request. In our experiments, all eight cores of Intel and AMD machines with eight client threads of *apache2* and *mysqld* were used.

In order to implement a realistic workload, we decided to use the data gathered by the web statistics system for five real commercial websites as our testing database. This data includes the information about website‘s audience activity (what pages on what website were accessed, in what order, etc.) as well as the information about visitors themselves (client OS, user agent information, browser settings, session id retrieved from the cookies, etc.). The total number of records in the database is more than 3 million.

The memory intensity of the client threads varies greatly depending on what webpage or database operation is being requested by the client. We chose to request webpages that can be found in a typical web statistics system and database operations that are executed in everyday maintenance and analysis of a website‘s activity. Table IV describes the webpages and database operations

Table IV. Client Operations Requested

Type of operation (where it is used)	Client request ID	Client request description
Retrieval of visitor activity (essential for the analysis of the target audience, part of the web-statistics system). *All the web server threads executing retrieval requests are memory intensive (i.e., devils).*	Apache #0	Obtain the *session chart*: the top 100 popular sessions on the website.
	Apache #1	Obtain the *browser chart*: what browsers are used most frequently by the visitors of this website.
	Apache #2	Get the *visitor loyalty*: the rating of the most committed visitors and how many pages every visitor have accessed before.
	Apache #3	Get the *country popularity*: how many pages were accessed from each country.
Maintenance of the website database: inserts (archiving, updating and merging data from several databases), indexing. *All the database server threads executing maintenance requests are CPU intensive (i.e., turtles). Also, since the threads are working with the tables that are stored on the local hard drive, the continuous phases of CPU intensity had sudden drop downs of CPU activity due to accesses to the hard drive*	Mysql #0 - #5	Inserting rows into the database tables with the subsequent creation of the full-text table index.

that were performed in our experiments. A single Apache request corresponds to the retrieval from the server a large dataset corresponding to visitor activity. The time of servicing this lengthy request is about three minutes. A single MySQL request corresponds to performing a batch of operations typical in maintenance of website databases – on our platform this request took about five minutes. Once the request is fulfilled, it is sent again until all requests finish at least three times.

We next report the results for the four workloads consisting of the requests from Table IV on AMD and Intel systems. Aggregate workload speedup over DEFAULT is shown in Figures 13 and 14. Figures 15(a) and 15(c) show performance improvement over DEFAULT of the worst-case execution of each request type over a large number of trials. Figures 15(b) and 15(d) show the deviation of the execution times.

To help understand the results, we explain the nature of the workload. The workload executed by Apache is extremely memory-intensive. It issues off-chip request at the rate of 74 misses per 1000 instructions – this is more than three times higher (!) than the most memory-intensive application that we encountered in the SPEC CPU2006 suite. The implication of such unprecedented memory intensity is that on our Intel system, where there is only one memory controller, the memory controller becomes the bottleneck, and our scheduling algorithm, which on the Intel system alleviates contention only for front-side

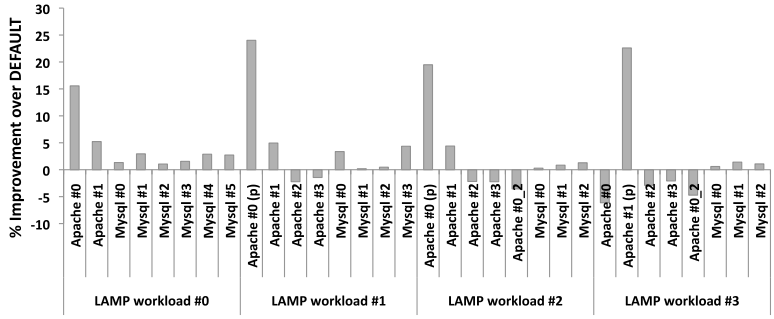


Fig. 13. Performance improvement per thread of each LAMP workload with DIO relative to DEFAULT (high bars are good) on the AMD system.

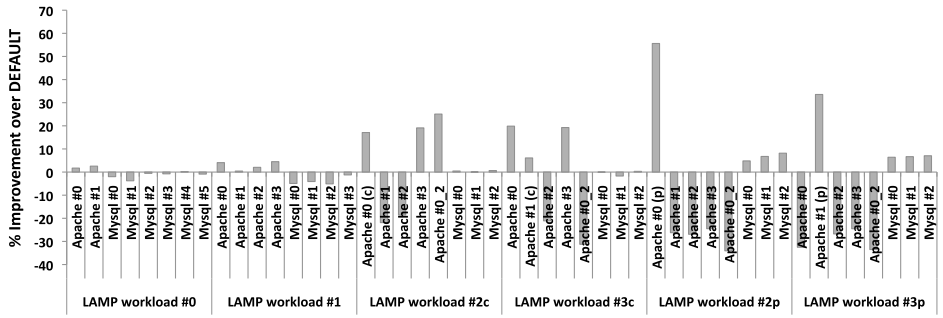
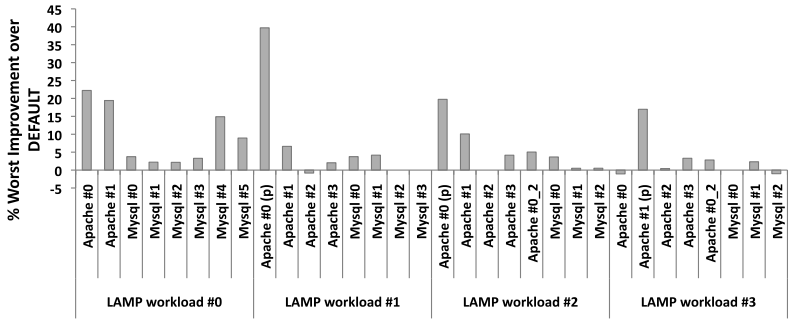


Fig. 14. Performance improvement per thread of each LAMP workload with DIO relative to DEFAULT (high bars are good) on the Intel system.

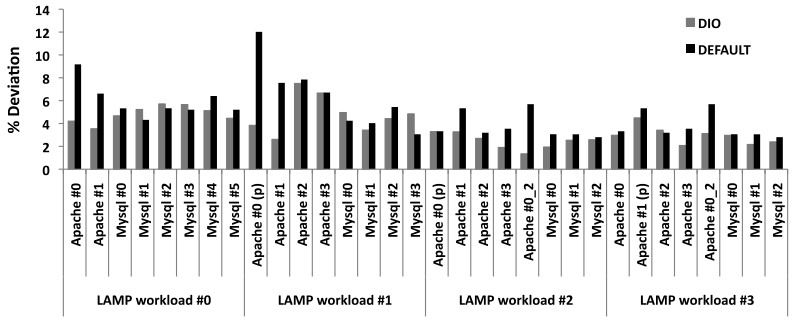
bus controllers, caches, and prefetching units (Figure 4), cannot deliver measurable performance improvements. Figure 14 illustrates this effect, showing that DIO does not improve performance on that system.

Turning our attention to Figure 13, on the other hand, we observe that on the AMD system with multiple memory controllers, DIO *does* improve performance. Apache threads, in particular, significantly benefit from DIO. (MySQL threads are not memory-intensive, so they do not and are not expected to benefit.)

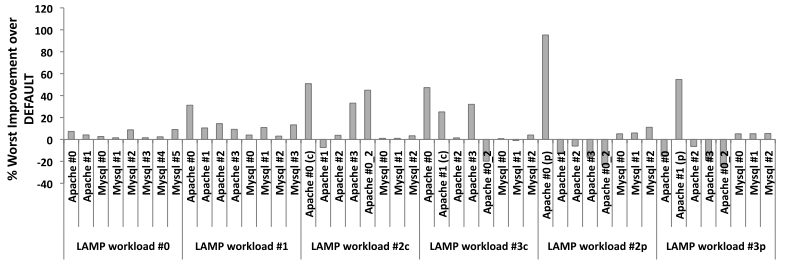
At the same time, as we further increase the number of Apache threads in the workload, we observe the same situation as on Intel systems: it is not possible to improve performance for all Apache threads simultaneously due to contention. This situation occurs when the number of Apache threads is greater than the number of memory domains. Apache executes a very memory-intensive workload and the only way to completely isolate an Apache thread from contention is to schedule it without any other Apache threads in the memory domain. This is not possible when the number of Apache threads is greater than the number of memory domains. However, what we *can* do is configure DIO to achieve performance isolation for *chosen* threads, as we demonstrate next.



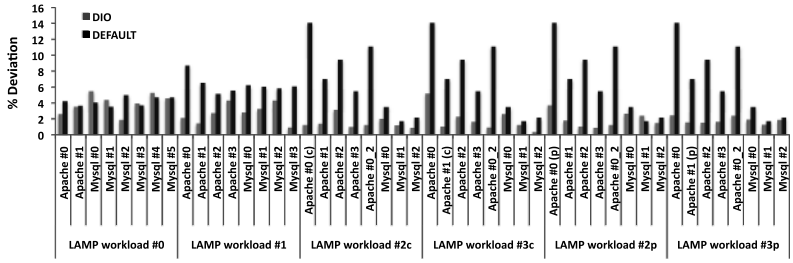
(a) Relative performance improvement of the worst case DIO over the worst case DEFAULT of each LAMP workload on the AMD system.



(b) Deviation with DIO and Default (low bars are good) of each LAMP workload on the AMD system.



(c) Relative performance improvement of the worst case DIO over the worst case DEFAULT of each LAMP workload on the Intel system.



(d) Deviation with DIO and Default (low bars are good) of each LAMP workload on the Intel system.

Fig. 15. Relative performance improvement of DIO over Default for LAMP workloads.

In this prioritized mode, DIO will ensure that a chosen memory-intensive thread will run in a memory domain without any other memory-intensive threads. For our LAMP workload, this will result in one Apache thread running in a domain only with "peaceful" (low missrate) MySQL threads. On the Intel system, a thread can be isolated either on a separate shared cache or on a separate physical package (recall Figure 4). On the AMD system, a thread isolated in a memory domain will be isolated both in the physical package and in the shared cache.

Workloads #1-3 in Figures 13, 15(a), and 15(b) are the workloads where one Apache thread is prioritized, executed on an AMD system. Workloads #2c-3p in Figures 14, 15(c) and 15(d) are similar "prioritized" workloads on the Intel system. The prioritized thread is marked with a "(p)" in scenarios where it is isolated in a physical package, and with a "(c)" when it is isolated on the separate cache but not on a separate package (on the Intel system).

We observe that the marked threads – especially those isolated on a separate physical package – achieve a very significant performance improvement over the DEFAULT scheduler on average. The improvements in the worst-case execution times are even more dramatic. DIO delivers these performance improvements, because it is able to identify threads that will suffer from contention if co-scheduled. If one such "sensitive" thread is also assigned a high priority by the user, DIO will be able to isolate it from contention. Note, this is not the same as statically assigning high-priority threads to run on a separate domain – this will unnecessarily waste resources. DIO's ability to identify contention-sensitive threads enables it to deliver isolation only when this is likely to improve performance.

DIO also reduces the deviation in execution times of Apache threads: execution times become much more stable with DIO. Deviation for MySQL threads remains unchanged, because MySQL threads are partially I/O-bound. I/O creates a large variability in execution times, which is not (and cannot be) mitigated by a contention-aware scheduler.

Our overall conclusions drawn from the results with the LAMP workloads are as follows.

—Contention-aware scheduling is especially effective on systems that have multiple contention bottlenecks, such as multiple memory controllers. In that case, a scheduler can spread the load across these several contention points and improve overall performance. On systems where performance is dominated by a single bottleneck (e.g., a single memory controller on the Intel system), effectiveness of contention-aware scheduling diminishes.

—In cases where the workload contains multiple memory-intensive threads, it becomes more difficult to improve performance for *every* thread in the workload. This is an inherent limitation of contention-aware scheduling. When contention is high only that much can be done to reduce it by means of shuffling threads across contention points. In that case, the biggest practical benefit of contention-aware scheduling is in providing performance isolation for selected threads. For instance, in a multithreaded workload, these could

be threads responsible for handling latency-sensitive queries (e.g., search threads as opposed to indexing threads in a Google workload).

6. MINIMIZING POWER CONSUMPTION ON MULTICORE SYSTEMS WITH RESOURCE CONTENTION

The Distributed Intensity Online algorithm has an inherent ability to predict when a group of threads co-scheduled on the same memory domain will significantly degrade each other's performance. We found that this ability can be successfully exploited to build a power-aware scheduler (DIO-POWER) that would not only mitigate resource contention, but also reduce system energy consumption.

The key observation behind DIO-POWER is that clustering threads on as few memory domains as possible reduces power consumption. That is because, on modern systems, memory domains correspond to power domains (i.e., chips) and so leaving an entire chip idle reduces the power that the system consumes. At the same time, clustering threads together may hurt performance. As a result, the workload's execution time may increase, leading to increased system uptime and increased consumption of energy (power multiplied by time) despite reduced consumption of power. While a conventional scheduler does not know how to make the right decision (to cluster or not to cluster), DIO-POWER does, because it relies on DI's ability to predict whether a group of threads will hurt each other's performance when clustered.

We now present the data (Table V) that shows reduction in power consumption due to clustering memory-intensive and CPU-intensive threads on the Intel machine for three workloads consisting of SPEC CPU 2006 benchmarks. The power consumed in every scenario is given relative to the power consumed by a completely idle machine when no benchmarks are running. Power is measured using power analyzer Extech 380803.

The results for the first workload show the power savings from clustering. Four instances of NAMD were first scheduled together on the same physical package (socket) so that another package can be turned off. The increase in power consumption relative to the idle state in this case was then compared with the increase when four NAMDs are spread across the system: one instance per each shared cache. Note that this is exactly how the DIO algorithm will assign threads to cores (it will spread them). As can be seen from the table, the relative power consumption in this case increased by 18W (24%).

Similar results were obtained for the workloads consisting of instances of two different applications: one memory-intensive and one CPU-intensive. Here the corresponding increase in power consumption was as high as 81W (90%) for Power Workload #2 and 50W (48%) for Power Workload #3.

Two more interesting observations can be made from the results in Table V.

—Putting devils on the same physical package results in lower power consumption of the system;

Table V. Power Consumption on the Intel System for Three Power Workloads from SPEC CPU 2006 Suite and 2 LAMP Power Workloads (Devils are Highlighted in Bold)

Workload	**Physical package 0**				**Physical package 1**				**Power increase (W)**	**Which cores are busy**
	cache 0		**cache 1**		**cache 2**		**cache 3**			
	core 0	**core 2**	**core 4**	**core 6**	**core 1**	**core 3**	**core 5**	**core 7**		
Power workload #1	NAMD1	NAMD2	NAMD3	NAMD4					76	all cores on one physical package
	NAMD1		NAMD2		NAMD3		NAMD4		94	one core in every shared cache
Power workload #2	**MCF1**	NAMD1	**MCF2**	NAMD2					90	all cores on one physical package (good co-scheduling)
	MCF1	**MCF2**	NAMD1	NAMD2					94	all cores on one physical package (bad co-scheduling)
	MCF1		**MCF2**		NAMD1		NAMD2		126	one core in every shared cache (2 MCFs on one package)
	MCF1		NAMD1		**MCF2**		NAMD2		171	one core in every shared cache (2 MCFs on different packages)
Power workload #3	**MILC1**	POVRAY1	**MILC2**	POVRAY2					104	all cores on one physical package (good co-scheduling)
	MILC1	**MILC2**	POVRAY1	POVRAY2					101	all cores on one physical package (bad co-scheduling)
	MILC1		**MILC2**		POVRAY1		POVRAY2		149	one core in every shared cache (2 MILCs on one package)
	MILC1		POVRAY1		**MILC2**		POVRAY2		154	one core in every shared cache (2 MILCs on different packages)
LAMP Power workload #1	***Apache #0***		***Apache #1***		***Apache #2***		***Apache #3***		118	DIO
	Apache #0	***Apache #1***	***Apache #2***	***Apache #3***					76	DEFAULT with power savings
	Apache #0		***Apache #1***		***Apache #2***		***Apache #3***		113	DIO-POWER
LAMP Power workload #2	***Apache #0***		***Apache #1***		*Mysql #0*		*Mysql #1*		131	DIO
	Apache #0	***Apache #1***	*Mysql #0*	*Mysql #1*					90	DEFAULT with power savings
	Apache #0	*Mysql #0*	***Apache #1***	*Mysql #1*					90	DIO-POWER

Algorithm 3 Invocation at the beginning of every scheduling interval

1: // *initialization of the boolean global array Order[l], every member specifies the order of browsing entities on the corresponding level of the memory hierarchy. Each entity has its ID that was given to it by OS. Order[l]=0 means that entities are browsed from the smallest ID to the biggest, Order[l]=1 – vice-versa.*
2: **for** $l = 0; l <$ *number_of_memory_hierarchy_levels* $>; l++$ **do**
3: *Order*[*l*]:=0;
4: **end for**
5: sort the threads according to the *miss rate* in descending order
6: **classify the threads into devils and turtles (we loosely define devils as those with at least 2000 shared cache misses per million instructions)**
7: **determine the number of devils** D
8: **mark the first** D **shared caches along with their containers for scheduling (the rest will remain idle and so can be turned into lower power state)**
9: let T be the array of sorted threads
10: // *spread the threads across the machine* ***while preventing placing devils together in the same shared cache***
11: **while** $T \neq \emptyset$ **do**
12: take the first (the most aggressive) $t \in T$
13: // *invoke DIO-POWER() to assign the thread*
14: DIO-POWER($t,<$ *machine* $>$, 0)
15: **end while**

—When threads are clustered on the same package, shared cache contention between devils does not contribute to the power consumption increase.

Remember, however, that power savings do not necessarily translate into energy savings: when devils are clustered on the same chip they will extend each other's execution time and, as a result, increase overall energy consumption. Therefore, a power-aware algorithm must be able to determine exactly when clustering is beneficial and when it is not.

We present DIO-POWER, a contention-aware scheduling algorithm that minimizes power consumption while preventing contention. Its structure described in blocks of pseudo-code 3 and 4 is similar to that of DIO with the changes highlighted in bold.

DIO-POWER reduces power consumption on the system by clustering the workload on as few power domains as possible while preventing the clustering of those threads that will seriously hurt each other's performance if clustered. Mildly intensive applications may be clustered together for the sake of energy savings and may thus suffer a mild performance loss. To capture the effect of DIO-POWER on both energy and performance, we used Energy Delay Product [Gonzalez and Horowitz 1996; Thomas et al. 1994]:

$$EDP = \frac{Energy\_Consumed}{Instructions\_per\_Second} = \frac{Average\_Power * Time * Time}{Instructions\_Retired} \tag{9}$$

EDP shows by how much energy savings (relative to the idle energy consumption of the machine) outweigh performance slowdown due to clustering.

In the next experiment, we show how DIO-POWER is able to achieve optimal EDP for the dynamic workload mix relative to conventional Linux schedulers.

Algorithm 4 DIO-POWER(thread to schedule t, container *e_parent*, memory hierarchy level of the container l)

```
1: Let E_all be the array of entities on hierarchy level l + 1
2: Let E_children be the array of entities on hierarchy level l+1 whose container is e_parent
   and that were marked for scheduling
3: browse the entities in E_children in order Order[l+1] and determine the first entity with
   the minimum number of allocated threads: e_min ∈ E_children
4: if e_min is a < core > then
5:    // we have reached the bottom of the memory hierarchy
6:    if assigning thread t to core e_min would not result in 2 devils in the same
      shared cache then
7:       // we have reached the bottom of the memory hierarchy
8:       assign thread t to core e_min
9:    end if
10: else
11:   increment the number of threads allocated to e_min
12:   // recursively invoke DIO-POWER() to assign the thread on the lower hierarchy
      level
13:   DIO-POWER(t,e_min,l + 1)
14: end if
15: if number of threads allocated to each entity e ∈ E_all is the same then
16:   // reverse the order of browsing on this level
17:   Order[l + 1] := NOT Order[l + 1]
18: end if
```

The Linux scheduler can be configured in two modes. The DEFAULT mode, used throughout this article, balances load across memory domains (chips) when the number of threads is smaller than the number of cores. DEFAULT-MC is the default scheduler with the *sched_mc_power_savings* flag turned on – in this mode the scheduler attempts to cluster the threads on as few chips as possible, but without introducing any unwarranted runqueue delays.

Although each version of the Linux scheduler is able to deliver optimal EDP for a specific group of workloads – DEFAULT for predominantly devil workloads where it makes sense to spread applications across domains, and DEFAULT-MC for predominantly turtle workloads where it makes sense to cluster applications, they cannot make the right decision across all workloads. So the user cannot configure the scheduler optimally unless the workload is known in advance. In the next experiment we will show that DIO-POWER is able to match the EDP of the best Linux scheduler for any workload, while each Linux scheduler only does well for a workload that happens to suit its policy.

Figure 16 shows how DIO-POWER works for LAMP threads described in the previous section on the Intel system. The workloads consist of four threads (to make the power savings due to clustering possible) with the varying number of Apache threads per workload.

We can see that while the DEFAULT scheduler does well in terms of EDP for the first two workloads with two and four Apache "devil" threads, where spreading these threads across chip makes sense, it does not do well for the third workload where only one Apache thread is present and so it makes sense

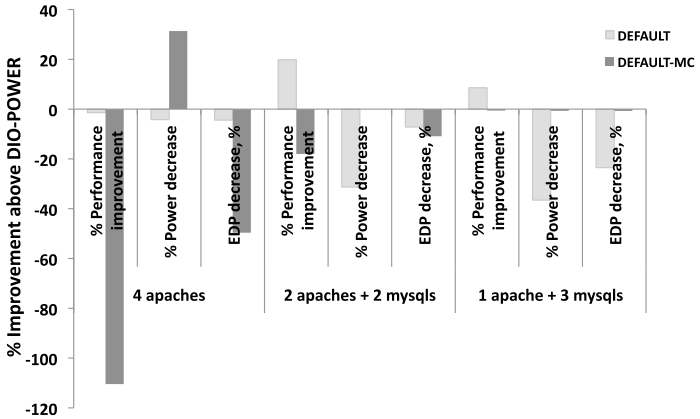


Fig. 16. The comparison in terms of performance, power consumption and EDP for 3 LAMP workloads with DEFAULT and DEFAULT-MC relative to DIO-POWER (high bars are good) on the Intel system. DIO-POWER is able to provide the best solution in terms of EDP trade-off every time.

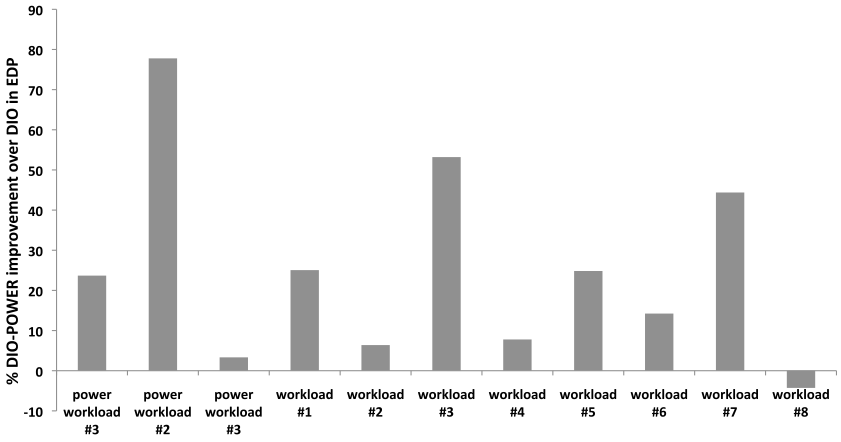


Fig. 17. Improvement in terms of EDP for each workload with DIO-POWER relative to DIO (high bars are good) on the Intel system.

to cluster threads. Similarly, DEFAULT-MC does well for the third workload, but not for the first two. DIO-POWER, on the other hand, achieves good EDP across all three workloads, matching the performance of the best Linux scheduling mode in each case, but without requiring manual configuration and advance knowledge of the workload.

Finally, Figure 17 shows the improvement in terms of EDP for each scientific workload from SPEC CPU 2006 with DIO-POWER relative to DIO on the Intel system. On average, DIO-POWER is able to show a significant improvements in EDP (up to 78%). In the case of workload #8 DIO-POWER was slightly worse than DIO. The reason is that clustering slowdown was able to outweigh power savings in this case.

We are not presenting any results for AMD Opteron systems due to almost absolute insensitivity of AMD's machines to the power savings resulting from leaving chips idle. The power consumption stayed relatively the same regardless of the number of idle chips in the system. The reason for that is probably a lack of a power savings mode for the entire processor package as opposed to any particular core (on Intel machines there are the processor/package C-states, or PC-states, where the entire processor enters lower power mode when all cores on it are idle).

We conclude that the ability of DIO to identify threads that will hurt each other's performance can be used to implement an effective power-savings policy. Existing "power-aware" schedulers in Linux can only be configured statically to cluster threads or spread them apart. DIO-POWER, on the other hand, can decide whether to spread or to cluster dynamically, depending on the properties of the workload. The same heuristic that helps DIO determine whether two threads will hurt each other's performance can help it decide whether energy savings as a result of tighter resource sharing will outweigh the energy loss as a result of contention.

7. RELATED WORK

In the work on Utility Cache Partitioning [Qureshi and Patt 2006], a custom hardware solution estimates each application's number of hits and misses for all possible number of ways allocated to the application in the cache (the technique is based on stack-distance profiles). The cache is then partitioned so as to minimize the number of cache misses for the co-running applications. UCP minimizes cache contention given a particular set of co-runners. Our solution, on the other hand, decides which co-runners to co-schedule so as to minimize contention. As such, our solution can be complementary to UCP and other solutions relying on cache partitioning [Suh et al. 2002].

Tam et al. [2009] similarly to several other researchers [Cho and Jin 2006; Liedtke et al. 1997; Lin et al. 2008; Zhang et al. 2009] address cache contention via software-based cache partitioning. The cache is partitioned among applications using page coloring. Each application is reserved a portion of the cache, and the physical memory is allocated such that the application's cache lines map only into that reserved portion. The size of the allocated cache portion is determined based on the marginal utility of allocating additional cache lines for that application. Marginal utility is estimated via an application's reuse distance profile, which is very similar to a stack-distance profile and is approximated online using hardware counters [Tam et al. 2009]. Software cache partitioning, like hardware cache partitioning, is used to isolate workloads that hurt each other. While this solution delivers promising results, it has two important limitations: first of all, it requires nontrivial changes to the virtual memory system, a very complicated component of the OS. Second, it may require copying of physical memory if the application's cache portion must be reduced or reallocated. Given these limitations, it is desirable to explore options like scheduling, which are not subject to these drawbacks.

Moscibroda and Mutlu [2007] demonstrated the problem with shared resource contention in memory controllers. They showed that shared memory controller contention on real dual-core and quad-core systems can be a significant problem, causing largely unfair slowdowns, reduced system performance and long periods of service denial to some applications. They later addressed the problem by designing fair memory controllers that reduce and manage interference in the main memory subsystem [Mutlu and Moscibroda 2007, 2008; Kim et al. 2010]. Several researchers addressed shared resource contention due to prefetchers as well as in on-chip networks [Das et al. 2009; Ebrahimi et al. 2009; Grot et al. 2009; Lee et al. 2008]. The solutions proposed offer enhancements to certain parts of the memory hierarchy rather than target the shared resource contention problem in the whole system via scheduling. Hence, we see this work as complementary to ours.

Herdrich et al. [2009] proposed rate-based QoS techniques, which involved throttling the speed of the core in order to limit the effects of memory-intensive applications on its co-runners. Rate-based QoS can be used in conjunction with scheduling to provide isolation from contention for high-priority applications.

Several prior studies investigated the design of cache-aware scheduling algorithms. Symbiotic Jobscheduling [Snavely and Tullsen 2000] is a method for co-scheduling threads on simultaneous multithreading processors (SMT) in a way that minimizes resource contention. This method could be adapted to co-schedule threads on single-threaded cores sharing caches. This method works by trying (or sampling) a large number of thread assignments and picking the ones with the best observed rate of instructions per cycle.

The drawback of this solution is that it requires a sampling phase during which the workload is not scheduled optimally. When the number of threads and cores is large, the number of samples that must be taken will be large as well.

Fedorova et al. [2007] designed a cache-aware scheduler that compensates threads that were hurt by cache contention by giving them extra CPU time. This algorithm accomplishes the effect of fair cache sharing, but it was not designed to improve overall performance.

The idea of distributing benchmarks with a high rate of off-chip requests across different shared caches was also proposed in earlier studies [Dhiman et al. 2009; Knauerhase et al. 2008]. The authors propose to reduce cache interference by spreading the intensive applications apart and co-scheduling them with nonintensive applications. Cache misses per cycle were used as the metric for intensity. Our idea of DI is similar, however we provide a more rigorous analysis of this idea and the reasons for its effectiveness, and also demonstrate that DI operates within a narrow margin of the optimal. The other papers did not provide the same analysis, so it was difficult to judge the quality of the algorithm. Further, previous work did not resolve the controversy between two approaches to model contention: Chandra et al. [2005] and Suh et al. [2002] demonstrated that reuse-distance models provide some of the most accurate measures of contention, but this contradicted the hypothesis that the rate of off-chip requests can be a good heuristic for predicting contention. Our

work reconciled these two approaches, explaining that Chandra's model is suitable for systems where cache contention is the dominant cause of performance degradation, but on real systems, where other contention factors are equally important, LLC miss rates turn out to be a very good heuristic for contention.

8. CONCLUSIONS

In this work, we identified factors other than cache space contention that cause performance degradation in multicore systems when threads share the memory hierarchy. We estimated that other factors like memory controller contention, memory bus contention and prefetching hardware contention contribute more to overall performance degradation than cache space contention. We predicted that in order to alleviate these factors it was necessary to minimize the total number of memory requests issued from each cache. To that end we developed scheduling algorithms DI and DIO that schedule threads such that the miss rate is evenly distributed among the caches.

The Miss Rate heuristic, which underlies the DI and DIO algorithms was evaluated against the best known strategies for alleviating performance degradation due to cache sharing, such as SDC, and it was found to perform near the theoretical optimum. DIO is a user level implementation of the algorithm relying on the Miss Rate heuristic that gathers all the needed information online from performance counters. DIO is simple, can work both at the user and kernel level, and it requires no modifications to hardware or the non-scheduler parts of the operating system. DIO has been shown to perform within 2% of the oracle optimal solution on two different machines. It performs better than the default Linux scheduler both in terms of average performance (for the vast majority of workloads) as well as in terms of execution time stability from run to run for individual applications.

Upon evaluating performance of DIO we concluded that the highest impact from contention-aware scheduling is in its ability to provide performance isolation for "important" applications and to optimize system energy consumption. DIO improved stability of execution times and lowered worst-case execution time by up to a factor of two for scientific workloads and by up to 40% for a commercial Apache/MySQL workload. Furthermore, DIO's ability to predict when threads will degrade each other's performance inspired the design of DIO-POWER, an algorithm that catered to system energy-delay product and improved the EDP relative to DIO by as much as 80%.

APPENDIX

A. METHODOLOGY FOR MEASURING THE BREAKDOWN OF DEGRADATION FACTORS ON THE NUMA SYSTEM

As described in Section 3, we identified four sources of performance degradation on our experimental AMD Opteron system: L3 cache, DRAM controller, inter-processor interconnects and remote latency. These factors are further abbreviated as L3, DR, IC and RL respectively. To quantify the effects of performance degradation caused by these factors we use the methodology, which is depicted in Figure 18. We run a target application, denoted as T with a set

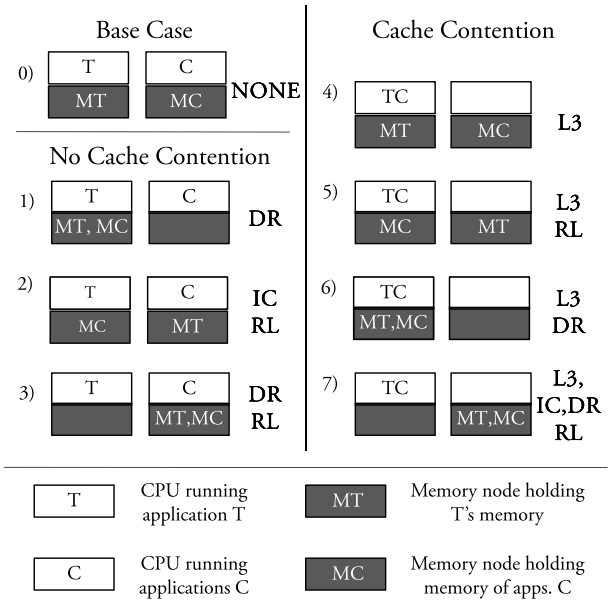


Fig. 18. Placement of threads and memory on all experimental configurations.

of three competing applications (MILC, in this case), denoted as *C*. The memory of the target application is denoted *MT*, and the memory of the competing applications is denoted *MC*. We vary (1) how the target application is placed with respect to its memory, (2) how it is placed with respect to the competing applications, and (3) how the memory of the target is placed with respect to the memory of the competing applications. We used *sched_setaffinity* system call to bind threads to cores and *numactl* package to specify the memory placement for every application. The methodology is essentially similar to the one used for the Intel system, except here we tested with three competing applications instead of one, since our NUMA AMD system has four cores per shared cache, while the Intel system has only two.

Figure 18 summarizes the relative placement of memory and applications that we used in our experiments. Next to each scenario we show factors affecting the performance of the target application: L3, DR, IC, or RL. For example, in Scenario 0, an application runs contention-free, because the competing applications and their memory are in a different domain. Additionally, the application's memory is local in this case. This is the optimal scenario performance-wise. We term it the *base* case and compare to it the performance in other cases. In Scenario 2, the target application competes with contending applications for the interconnect (IC), because all applications use the same interconnect for fetching their memory from a remote bank, and also experiences remote access latency (RL). Scenario 7 is the worst in terms of performance degradation, because all four factors are present. The scenarios where there is cache contention are shown on the right and the scenarios where there is no cache contention are shown on the left. By comparing the performance in

different scenarios we were able to estimate the degradation factors shown in Figure 6.

REFERENCES

AN MEY, D., SARHOLZ, S., TERBOVEN, C., VAN DER PAS, R., AND LOH, E. 2007. The RWTH Aachen SMP-Cluster User's Guide, Version 6.2.

BLAGODUROV, S., ZHURAVLEV, S., LANSIQUOT, S., AND FEDOROVA, A. 2009. Addressing contention on multicore processors via scheduling. *Tech. Rep., Simon Fraser University 2009-16*.

CHANDRA, D., GUO, F., KIM, S., AND SOLIHIN, Y. 2005. Predicting inter-thread cache contention on a chip multi-processor architecture. In *Proceedings of the 11th International Symposium on High-Performance Computer Architecture (HPCA'05)*. 340–351.

CHO, S. AND JIN, L. 2006. Managing distributed, shared l2 caches through os-level page allocation. In *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'39)*. 455–468.

DAS, R., MUTLU, O., MOSCIBRODA, T., AND DAS, C. R. 2009. Application-aware prioritization mechanisms for on-chip networks. In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'42)*. 280–291.

DHIMAN, G., MARCHETTI, G., AND ROSING, T. 2009. vGreen: A system for energy efficient computing in virtualized environments. In *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*.

EBRAHIMI, E., MUTLU, O., LEE, C. J., AND PATT, Y. N. 2009. Coordinated control of multiple prefetchers in multicore systems. In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'42)*. 316–326.

FEDOROVA, A., SELTZER, M. I., AND SMITH, M. D. 2007. Improving performance isolation on chip multiprocessors via an operating system scheduler. In *Proceedings of the 16th International Conference on Parallel Architectures and Compilation Techniques (PACT'07)*. 25–38.

GONZALEZ, R. AND HOROWITZ, M. 1996. Energy dissipation in general purpose microprocessors. *IEEE J. Solid-State Circ. 31*, 1277–1284.

GROT, B., KECKLER, S. W., AND MUTLU, O. 2009. Preemptive virtual clock: A flexible, efficient, and cost-effective qos scheme for networks-on-chip. In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'42)*. 268–279.

HERDRICH, A., ILLIKKAL, R., IYER, R., NEWELL, D., CHADHA, V., AND MOSES, J. 2009. Rate-based QoS techniques for cache/memory in cmp platforms. In *Proceedings of the 23rd International Conference on Supercomputing (ICS'09)*. 479–488.

HOSTE, K. AND EECKHOUT, L. 2007. Microarchitecture-independent workload characterization. *IEEE Micro 27,* 3, 63–72.

JIANG, Y., SHEN, X., CHEN, J., AND TRIPATHI, R. 2008. Analysis and approximation of optimal co-scheduling on chip multiprocessors. In *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques (PACT'08)*. 220–229.

KIM, Y., HAN, D., MUTLU, O., AND HARCHOL-BALTER, M. 2010. Atlas: A scalable and high-performance scheduling algorithm for multiple memory controllers. In *Proceedings of the 41st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'41)*.

KNAUERHASE, R., BRETT, P., HOHLT, B., LI, T., AND HAHN, S. 2008. Using OS observations to improve performance in multicore systems. *IEEE Micro 28,* 3, 54–66.

LEE, C. J., MUTLU, O., NARASIMAN, V., AND PATT, Y. N. 2008. Prefetch-aware dram controllers. In *Proceedings of the 16th IEEE International Symposium on High-Performance Computer Architecture (HPCA'08)*. 200–209.

LIEDTKE, J., HAERTIG, H., AND HOHMUTH, M. 1997. OS-controlled cache predictability for real-time systems. In *Proceedings of the 3rd IEEE Real-Time Technology and Applications Symposium (RTAS'97)*. 213.

LIN, J., LU, Q., DING, X., ZHANG, Z., ZHANG, X., AND SADAYAPPAN, P. 2008. Gaining insights into multicore cache partitioning: Bridging the gap between simulation and real systems. In *Proceedings of the International Symposium on High Performance Computer Architecture (HPCA'08)*. 367–378.

LUK, C.-K., COHN, R., MUTH, R., PATIL, H., KLAUSER, A., LOWNEY, G., WALLACE, S., REDDI, V. J., AND HAZELWOOD, K. 2005. PIN: building customized program analysis tools with dynamic instrumentation. In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI'05)*. 190–200.

MOSCIBRODA, T. AND MUTLU, O. 2007. Memory performance attacks: denial of memory service in multicore systems. In *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium (SS'07)*. 1–18.

MUTLU, O. AND MOSCIBRODA, T. 2007. Stall-time fair memory access scheduling for chip multiprocessors. In *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'40)*. 146–160.

MUTLU, O. AND MOSCIBRODA, T. 2008. Parallelism-aware batch scheduling: Enhancing both performance and fairness of shared dram systems. In *Proceedings of the 35th Annual International Symposium on Computer Architecture (ISCA'08)*. 63–74.

QURESHI, M. K. AND PATT, Y. N. 2006. Utility-based cache partitioning: A low-overhead, high-performance, runtime mechanism to partition shared caches. In *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'39)*. 423–432.

SHELEPOV, D., AND FEDOROVA, A. 2008. Scheduling on heterogeneous multicore processors using architectural signatures. In *Proceedings of the Workshop on the Interaction between Operating Systems and Computer Architecture (WIOSCA)*.

SNAVELY, A. AND TULLSEN, D. M. 2000. Symbiotic jobscheduling for a simultaneous multithreaded processor. *SIGARCH Comput. Archit. News 28,* 5, 234–244.

SUH, G. E., DEVADAS, S., AND RUDOLPH, L. 2002. A new memory monitoring scheme for memory-aware scheduling and partitioning. In *Proceedings of the 8th International Symposium on High-Performance Computer Architecture (HPCA'02)*. 117.

TAM, D., AZIMI, R., AND STUMM, M. 2007. Thread clustering: sharing-aware acheduling on smp-cmp-smt multiprocessors. In *Proceedings of the 2nd ACM European Conference on Computer Systems (EuroSys'07)*.

TAM, D. K., AZIMI, R., SOARES, L. B., AND STUMM, M. 2009. Rapidmrc: Approximating l2 miss rate curves on commodity systems for online optimizations. In *Proceeding of the 14th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'09)*. 121–132.

THOMAS, M. H., INDERMAUR, T., AND GONZALEZ, R. 1994. Low-power digital design. In *Proceedings of the IEEE Symposium on Low Power Electronics*. 8–11.

VAN DER PAS, R. 2005. The OMPlab on sun systems. In *Proceedings of the 1st International Workshop on OpenMP*.

XIE, Y. AND LOH, G. 2008. Dynamic classification of program memory behaviors in CMPs. In *Proceedings of CMP-MSI, (held in conjunction with ISCA-35)*.

ZHANG, X., DWARKADAS, S., AND SHEN, K. 2009. Towards practical page coloring-based multicore cache management. In *Proceedings of the 4th ACM European Conference on Computer Systems (EuroSys'09)*. 89–102.

Received May 2010; accepted October 2010