Exploiting Locality: A Flexible DSM Approach

Håkan Zeffer, Zoran Radović, and Erik Hagersten

Department of Information Technology, Uppsala University
P.O. Box 337, SE-751 05, Uppsala, Sweden
{hakan.zeffer, zoran.radovic, erik.hagersten}@it.uu.se

Abstract

No single coherence strategy suits all applications well. Many promising adaptive protocols and coherence predictors, capable of dynamically modifying the coherence strategy, have been suggested over the years.

While most dynamic detection schemes rely on plentiful of dedicated hardware, the customization technique suggested in this paper requires no extra hardware support for its per-application coherence strategy. Instead, each application is profiled using a low-overhead profiling tool. The appropriate coherence flag *setting, suggested by the profiling, is specified when the application is launched.*

We have compared the performance of a hardware DSM (Sun WildFire) to a software DSM built with identical interconnect hardware and coherence strategy. With no support for flexibility, the software DSM runs on average 45 percent slower than the hardware DSM on the 12 studied applications, while the flexibility can get the software DSM within 11 percent. Our all-software system outperforms the hardware DSM on four applications.

1. Introduction

While hardware-based shared-memory systems have been successfully built for many years, the cost in terms of design and verification for each new generation is ever increasing. Meanwhile, the advance in semiconductor technology have set the shared-memory server trend towards multiple cores per die (CMP) and multiple threads per core (SMT) [20]. We believe that this technology shift forces a reevaluation of the way to interconnect multiple such chips to form larger systems.

This paper presents a highly flexible all-software shared-memory proposal with a very low system design cost and short time-to-market. We extend the DSZOOM system [26] with several novel optimization options and add a low-overhead profiling mode that suggests appropriate *coherence flags* for the 12 applications studied. When compared with a hardware DSM system built from identical node hardware, interconnect and coherence strategy, the base system is trailing by 45 percent on average (slowdown-factor range is 0.98–3.02) for these 12 applications. The profiled coherence flags brought the numbers down to 17 percent on average (0.69–1.85) and hand-tuned coherence flags down to 11 percent (0.69–1.66). It was a bit surprising, but very encouraging, to note that the profile-based coherence flags propelled the software DSM to outperform the hardware DSM for four of the applications.

The technology presented in this paper can easily be incorporated in many parallel execution environments, such as OpenMP [7] or UPC compilers [2], providing a low cost but high performance execution platform for high-performance computing (HPC) applications.

2. Basic DSZOOM System

This section gives an overview of the basic DSZOOM system [26], a sequentially consistent [22] software-based DSM implementation that is inspired by three fine-grained software coherence proposals: Blizzard-S [30], Shasta [27, 28], and Sirocco-S [32]. DSZOOM relies on code instrumentation to maintain fine-grain coherence (load and store operations to shared memory are augmented with coherence checks). The provided protocols assume a high bandwidth, low latency cluster interconnect, supporting fast user level mechanisms for *put*, *get*, and *atomic* operations to remote nodes' memories, such as InfiniBand [18] or Sun Fire Link [34]. The system further assumes that the write order between any two endpoints in the network is preserved. These network assumptions make is possible to remove interrupt- and/or poll-based asynchronous protocol processing found in the majority of software DSM

implementations [26, 1]. A processor that has detected the need for global coherence activity will first acquire a lock associated with the coherence unit before starting the coherence activity. A requesting processor can independently lock a remote directory entry and obtain read/write permissions.

2.1. The Invalidation-Based Protocol

The invalidation-based protocol states, modified, shared and invalid (MSI), are explicitly represented by global data structures in the nodes' memories. Bits of a memory operation's effective address determine the location of a coherence unit's directory location, i.e., its "home node." All coherence units in invalid state store a "magic" data value, as independently suggested by Scales et al. [27] and Chiou et al. [6] (Schoinas et al. [31] use the same technique in the Blizzard-S system). This significantly reduces the number of directory accesses caused by load operations, since the directory only has to be consulted on a read miss. (The directory also has to be consulted in the rare case when the real data value is equal to the magic value [31, 27].)

To reduce the number of accesses to remote directory entries caused by global store operations, each node has one byte of local state (MTAG) per global coherence unit (similar to Shasta's *private state table* [28]), indicating if the coherence unit is locally writable. Before each global store operation, the MTAG byte is checked. The directory only has to be consulted if the MTAG indicates that the node currently does not have write permission to the coherence unit. The directory will assume the role of MTAG in home nodes, and hence, no extra MTAG state is needed for home nodes. To avoid race conditions, the corresponding MTAG entry has to be locked before a write permission check is carried out. Otherwise, a coherence unit can be downgraded between the consultation and the point in time where the store is performed. More details and coherence miss examples are given in [40].

2.2. Write Permission Cache

DSZOOM's access control checks for stores represent the largest part of the total instrumentation cost [39]. Most of this overhead comes from the fact that the locally cached directory entry (MTAG) must be checked atomically for each global store operation. This section describes *write permission cache* (WPC) that hides some of the instrumentation cost for stores [39]. While Shasta reduces its instrumentation overhead by statically merging coherence actions at instrumentation time [27] (*batching*), a WPC dynamically merges store coherence checks at runtime. Instead of releasing the MTAG lock after a store is performed, a thread holds on to the write permission and the MTAG lock, hoping that the next store will be to the same coherence unit. The identity of the coherence unit is stored in a dedicated register, which is consulted before the next store is performed. (DSZOOM reserves UltraSPARC's application registers [36] for fast WPC checks.) If indeed the next store is to the same coherence unit, the store overhead is reduced to a few ALU operations and a conditional branch instruction. When a store to another coherence unit appears, a *WPC miss* occurs. Only then, a new lock release followed by a lock acquire must be performed.

The WPC hit rate for SPLASH-2 benchmarks [37] varies greatly depending on the application, the number of WPC entries and the coherence unit size [40]. For example, two entries demonstrate much better hit rate, which is most significant for `fft` and `ocean` applications.

3. Extending DSZOOM's Flexibility

One of the key observations of this paper is that the WPC technology can be used as an efficient software-based *store buffer* in an update-based system. To be more specific, multiple stores could be merged before the MTAG lock is released and the data is distributed to other nodes. In this section, we extend DSZOOM's flexibility with a new update-based protocol and multiple bandwidth reduction techniques.

3.1. The Update-Based Protocol

The update-based protocol is based on write permission. All nodes have read permission to all data whereas only one has read-write permission for each coherence unit. The states read-write (W) and read (R or !W) are explicitly represented in the nodes' memories. Remote directory traffic is tamed with an update version of the MTAG optimization described in Section 2.1.

3-hop Write Miss Example: Figure 1 shows coherence activity caused by an update 3-hop write miss. *atomic1* and *put1* correspond to a 1 byte remote atomic operation and a remote 1 byte put. The state transitions for coherence unit D can be found below the nodes. The requestor, *reqD*, first checks its local MTAG for write permission of coherence unit D. Since node2 does not has write permission, the store protocol is called and the coherence activity is started. *reqD* locks the directory located at the home node (`D1`). The directory indicates that the write permission is located

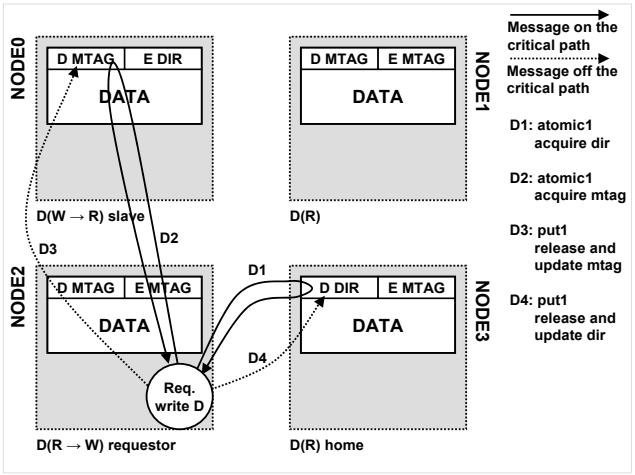

Figure 1. 3-hop write miss example for the update-based protocol.

on node0. Hence, `D2` locks node0's MTAG. This removes write permission on that node. `D3` and `D4` updates the MTAG (!W) and the directory (node2's id) respectively. Node2 is now in state read-write and has correct data (data has to be distributed before a lock is released, the network order is preserved).

3.2. Bandwidth Reduction Techniques

Update-based coherence protocols have to deal with the potential bandwidth problem introduced by excessive data pushing. In this section, we present two mechanisms (filtering strategies) that address the bandwidth problem: *dirty-data* and *private-data* filtering.

Dirty-Data Filtering. Scaling the coherence unit size has two potential benefits: (1) it can reduce the number of coherence misses and (2) the WPC hit rate is improved. Hence, the number of locks taken and the instrumentation overhead are reduced. However, update-based coherence protocols can be very sensitive to coherence unit size scaling. A large coherence unit wastes bandwidth when exposed to write-write false sharing or if processors only write parts of the coherence unit before distributing the data. We address this problem with a *dirty-data* WPC that tracks modifications of a current cache line. Hence, only modifications (dirty data) are distributed to other nodes. A coherence unit is divided into smaller parts. The lock is obtained per coherence unit whereas the WPC points to one part of the coherence unit, the "hot spot." Data are marked dirty when the hot spot is moved.

Private-Data Filtering. Benchmarks often show a large amount of stores that are to node-private data. For some applications, this is true even at page granularity. For example, 89 (77) percent of the global stores in `lu-c` (`ocean-c`) are to private pages. We exploit this application property to reduce the global update bandwidth consumed with a *private-data* filter. The virtual-memory system is used to keep track of private-to-shared state changes and the page state is used to omit updates to node-private pages.

Our private-data filter is implemented with a page-permission check from a locally-cached page directory before each global update. In addition, at read/write page faults to the shared memory segment, our signal handler updates the page directory with new permission information through a remote-atomic and a remote-put operation and uses `mprotect(2)` to set up local memory mappings. If the node does not has data, a page fetch might be needed. While multiple schemes are possible, our system only allows page-permission upgrades.

4. Profiling and Classification

Choosing optimal coherence flags may be a cumbersome task. This section presents a simple classification algorithm for fast finding of appropriate coherence settings. The classification heuristic is based on feedback from a low-overhead profile run. Our classification algorithm is not general, it is intended to show that also a simple heuristic can be used to achieve appropriate coherence settings, and hence, high performance.

We have tested our profiling mode and classification algorithm with both small and large working set sizes. Our results are almost identical. Thus, for the applications studied, it is possible to use the small working set size during the profile run and reuse the same coherence strategy for result runs. While scaling down the workload size on a uniprocessor system can heavily affect the cache performance, the profile mode tracks the entire shared memory and especially coherence traffic. Hence, it is not heavily dependent on machine parameters, such as cache sizes.

4.1. Low-Overhead Profiling

Our low-overhead profiling is capable to collect 1- and 2-entry WPC hit rate, global update bandwidth and coherence unit size information in a single run with less than 30 percent overhead (avg.). The estimation of coherence unit size uses *virtual coherence units*. We slice the global memory space into different segments (currently, 2048 bytes each). These segments use different virtual coherence unit sizes. When a coherence miss occurs, permission for the entire virtual coherence

unit size is acquired. DSZOOM's runtime system collects the number of misses for all different coherence sizes in a single run. Of course, it is important to divide the memory space in a representative way. We have tried multiple schemes (omitted because of space) and found that a simple modulo scheme works satisfactorily for our conservative classification algorithm and the applications studied. The virtual-memory system collects non-private store information. When a WPC miss occurs, a local page directory lookup classifies the store as private or non-private.

4.2. Simple Classification Method

This section describes the proposed classification algorithm. First of all, the memory consistency model of the application has to be taken into account. (DSZOOM can be run with multiple memory consistency models further discussed in Section 6.) We use the memory consistency model and global update bandwidth (estimated with profile run execution time) to select between invalidate/update.

The global update bandwidth with and without the private-data filter is used to make a choice for update filtering techniques. For example, `ocean-c` consumes more than 400 Mbyte/s without private-data filter. This number is reduced to less than 10 Mbyte/s when the filter is enabled (we enable the private-data filter if it reduces the bandwidth with more than 10 percent). Currently, we do not have a good metric for the dirty-data filter. However, it is reasonable to use this filter when a large coherence unit size is used with a 1-WPC update-based configuration.

We use the number of coherence misses to different virtual coherence unit sizes to select coherence unit size for an application. Applications with a significant amount of spatial locality (e.g., `fft` and `lu-c`) are easily recognized because the number of misses is reduced by 50 percent each time the coherence unit size is doubled. Applications that expose false sharing are also easily recognized since the number of misses increase. However, applications that exploit some locality and at the same time introduce some false sharing are harder to classify. We use a conservative approach for these applications by choosing a small coherence unit size. In addition, an upper limit of 512 bytes for update-based protocols is applied. Finally, the number of WPC entries has to be selected. The classification algorith uses the 1- and 2-entry WPC hit rate that the profile run provides.

Program	Large (Small) Problem Size
`fft`	4M (64k) points
`lu-c`	2048×2048 (512×512) matrices, 16×16 blocks
`lu-nc`	2048×2048 (512×512) matrices, 16×16 blocks
`radix`	32M (2M) integers, radix 1024
`barnes`	128k (16k) particles
`fmm`	128k (32k) particles
`ocean-c`	1026×1026 (258×258)
`ocean-nc`	1026×1026 (258×258)
`radiosity`	largeroom (room), -ae 5000.0 -en 0.050 -bf 0.10
`raytrace`	car (teapot)
`water-nsq`	4913 (2197) molecules, 2 time steps
`water-sp`	32768 (2197) molecules, 2 time steps

Table 1. SPLASH-2 benchmarks. The small working set is used together with the profiling mode described in Section 4. All performance results are based on the large data set sizes.

5. Performance Evaluation

Table 1 shows data set sizes for all of the SPLASH-2 applications studied [37]. The reason why we cannot run `volrend` is that shared variables are not correctly allocated with the *G_MALLOC* macro. `cholesky` is not run because we were not able to find large enough working sets.

5.1. Compiler and Instrumentation

All experiments in this paper use the GCC 3.3.4 compiler. To simplify instrumentation, we use GCC's `-fno-delayed-branch` flag that avoids loads and stores in delay slots, and `-mno-app-regs` that reserves UltraSPARC's thread-private registers [36]. These two flags slow down SPLASH-2 applications with less than 3 percent (avg.). Note that only the DSZOOM system uses those flags. All benchmarks are compiled with optimization level 3.

We extend DSZOOM's instrumentation tool with a simplified version of Shasta's batching technique [27]. The tool implements a *read-modify-write* batching, which merges load and store coherence checks (to the same effective address) by replacing the load check with the store's WPC check. We also schedule application instructions into coherence checking code to increase instruction-level parallelism (inspired by EEL [23]).

5.2. Hardware Setup

Most of the experiments are measured on a Sun Enterprise E6000 server [33]. The server has 16 UltraSPARC II (250 MHz) processors and 4 Gbyte uniformly

shared memory with an access time of 330 ns (*lmbench* latency [25]) and a total bandwidth of 2.7 Gbyte/s. Each processor has a 16 kbyte on-chip instruction cache, a 16 kbyte on-chip data cache, and a 4 Mbyte second-level off-chip data cache. The sequential experiments run on two processor types: a 250 MHz UltraSPARC II (USII) and a 900 MHz UltraSPARC III (USIII). The USIII processor has a 32 kbyte instruction cache, a 64 kbyte data cache, a 2 kbyte write cache and a 2 kbyte prefetch cache. The second-level cache is 8 Mbyte and off-chip.

The hardware DSM results have been measured on a 2-node Sun WildFire system built from two E6000 nodes connected through a hardware-coherent interface with a raw bandwidth of 800 Mbyte/s in each direction [15, 16]. The WildFire system has been configured as a traditional cache-coherent, non-uniform memory access (CC-NUMA) architecture with its data migration capability activated while its coherent memory replication (CMR) has been disabled. The Sun WildFire access time to local memory is the same as above, 330 ns, while accessing data located in the other E6000 node takes about 1700 ns (lmbench latency). WildFire runs the Solaris 2.6 operating system.

All software DSM implementations run in user space on the Sun WildFire system. The WildFire interconnect is in that case used as a "non-coherent" cluster interconnect between E6000 nodes. Non-cacheable block load, block store and regular SPARC atomic memory operations (`ldstub`) are used as remote put, get and atomic operations.

5.3. Instrumentation Overhead

Table 2 shows sequential-execution time in seconds for non-instrumented programs (second column). It also reports the factor increase in execution time when both load and store instrumentation is inserted for three invalidation-based configurations: the invalidation-based protocol without WPC (`inv`), the invalidation-based protocol with a 1-entry WPC (`inv-swpc`) and the invalidation-based protocol with a 2-entry WPC (`inv-dwpc`) (see Table 4 for abbreviations). All experiments run on both USII and USIII processors with a coherence unit size of 512 bytes. On average, instrumentation overhead for the slower processor (USII) is lowered from 66 percent for the `inv` protocol to 33 percent when a 2-entry WPC is used (`inv-dwpc`). For the faster processor, this reduction is even more significant (from 104 percent to 35 percent). The store instrumentation overhead for WPC implementations can be reduced even further if larger coherence unit sizes are used. For example, the store instrumentation overhead for `fft` is reduced from 178 to 27 percent for the USIII target when a 2-entry WPC is added and the coherence unit is scaled from 64 to 8192 bytes. Note that the instrumentation techniques based on WPC (such as `inv-swpc` and `inv-dwpc`) only add ALU instructions to the fast path of the execution when instrumenting loads and stores. This results in an instrumentation overhead which is fairly independent of processor technology, as can be seen in Table 2.

Program	**Time [s]** USII (III)	`inv` USII (III)	`inv-swpc` USII (III)	`inv-dwpc` USII (III)
`fft`	*17.4 (9.7)*	2.67 (2.89)	2.18 (2.08)	1.65 (1.52)
`lu-c`	*132.1 (42.9)*	3.56 (5.83)	1.48 (1.79)	1.51 (1.84)
`lu-nc`	*270.4 (87.4)*	2.15 (3.30)	1.53 (1.45)	1.60 (1.49)
`radix`	*57.7 (22.8)*	1.52 (1.75)	1.70 (1.79)	1.66 (1.69)
`barnes`	*161.0 (52.5)*	1.09 (1.15)	1.07 (1.13)	1.10 (1.13)
`fmm`	*155.0 (48.0)*	1.15 (1.28)	1.10 (1.17)	1.10 (1.18)
`ocean-c`	*84.5 (56.5)*	1.71 (1.72)	1.49 (1.30)	1.33 (1.18)
`ocean-nc`	*132.2 (91.4)*	1.45 (1.42)	1.38 (1.26)	1.31 (1.19)
`radiosity`	*39.8 (14.7)*	1.08 (1.19)	1.11 (1.18)	1.11 (1.18)
`raytrace`	*80.5 (23.9)*	1.24 (1.36)	1.24 (1.36)	1.24 (1.35)
`water-nsq`	*162.8 (68.2)*	1.18 (1.28)	1.16 (1.27)	1.12 (1.27)
`water-sp`	*115.7 (48.3)*	1.16 (1.28)	1.15 (1.23)	1.17 (1.24)
Avg.		1.66 (2.04)	1.38 (1.42)	1.33 (1.35)

Table 2. Sequential instrumentation overhead for 250 MHz UltraSPARC II and 900 MHz UltraSPARC III processor runs.

5.4. Low-Overhead Profiling

Table 3 shows performance of the profiling mode for 16-processor runs. Numbers for both training (small) and large data input sets are shown. On average, the profiling mode runs 29 percent slower than the base mode for large input sets. The performance of a profile mode is significantly increased for training input sets. For example, the most extreme case (`lu-c`) runs about 50 times faster than the base mode with a large input set.

5.5. Parallel Performance

Figure 2 shows the performance impact of various coherence protocols and optimizations for 16-processor runs when compared to DSZOOM's base protocol (`inv-64`). Selecting the most optimal coherence unit size improves `fft`'s performance with 25 percent. Adding the most appropriate WPC strategy

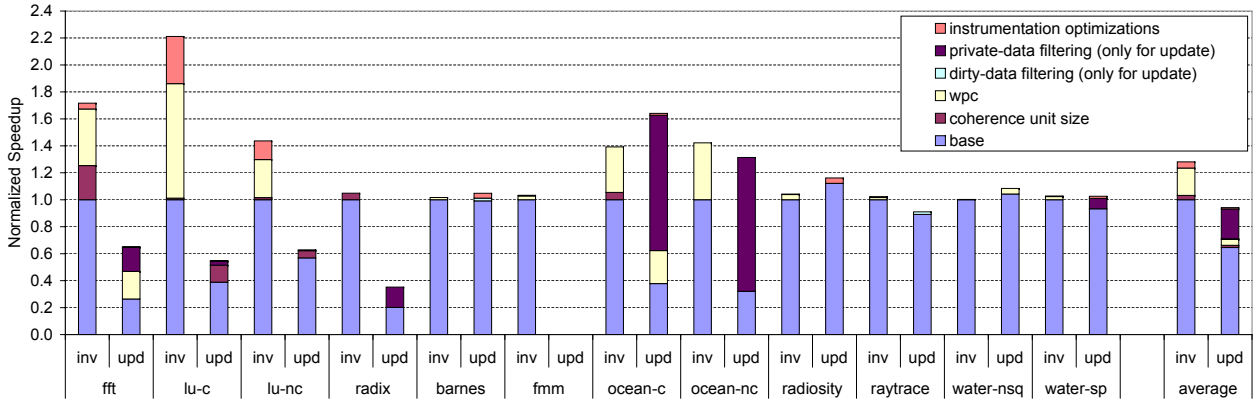

Figure 2. Invalidation- (`inv`) and update-based (`upd`) speedup when normalized to DSZOOM's invalidation-based system with a coherence unit size of 64 bytes (`inv-64`).

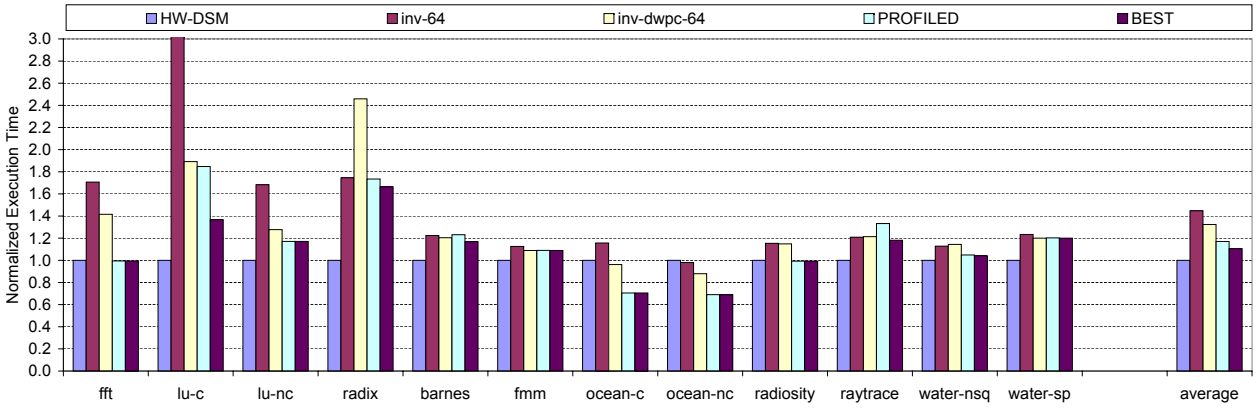

Figure 3. Hardware DSM vs. four software DSM experiments. (16-processor runs.)

amounts for an additional 0.4 speedup and instrumentation scheduling adds another 0.05. The base update protocol (`upd-swpc-64`) results in a slowdown compared with the `inv-64` protocol. However, tuning the WPC setting and enabling the private-data filtering improves `fft`'s update performance with more than 100 percent. Only performance improvements are shown in Figure 2, which is why some of the optimizations are not visible for all applications.

There is a large variation among the applications as to which class of optimization is the most important. `ocean-c`'s update protocol is greatly improved by the private-data filtering, which makes it outperform the best invalidate-based protocol, while `lu-c` enjoys a great boost to its invalidate protocol from its most optimal WPC setting. While Figure 2 can help understanding the importance of the different optimizations, it should be pointed out that the different performance improvements reported are somewhat dependent on each other, why the orders in which they are presented do effect their individual contributions.

Figure 3 shows parallel performance for 16-processor runs. Here, the software DSM performance can be compared to the execution time of a 2-node Sun Wild-Fire (HW-DSM) and DSZOOM's base configuration (`inv-64`). As a comparison, the execution time for `inv-dwpc-64` as well as the configuration suggested by the profiling tool (PROFILED), described in Section 4, are also reported. The rightmost bar (BEST) shows the best performance obtained by testing all coherence flag settings. (See Table 5 for (PROFILED) and (BEST) configurations.) To ensure that our results are not affected by application scaling characteristics, we also run all applications with four processors per node (eight in total). These results are almost identical to the ones presented in this paper (when compared to

Program	Base Mode Large Set	Profile Mode Large Set	Profile Mode Training Set
fft	5.99 s	6.13 s (1.02)	0.18 s (0.03)
lu-c	28.79 s	22.90 s (0.80)	0.55 s (0.02)
lu-nc	46.58 s	96.49 s (2.07)	2.90 s (0.06)
radix	13.09 s	16.86 s (1.29)	3.56 s (0.27)
barnes	14.62 s	19.64 s (1.34)	2.77 s (0.19)
fmm	16.26 s	26.73 s (1.64)	7.66 s (0.47)
ocean-c	9.37 s	9.04 s (0.96)	1.51 s (0.16)
ocean-nc	17.71 s	17.44 s (0.98)	2.18 s (0.12)
radiosity	4.89 s	9.16 s (1.87)	0.23 s (0.05)
raytrace	15.27 s	18.69 s (1.22)	14.02 s (0.92)
water-nsq	12.85 s	13.74 s (1.07)	3.41 s (0.27)
water-sp	9.12 s	10.62 s (1.16)	1.65 s (0.18)
Avg.		(1.29)	(0.23)

Table 3. Performance of the profiling mode for large and training (small) input data sets. Normalized execution time is shown inside parenthesis.

Abbreviation	DSZOOM Configuration
inv	invalidation-based protocol
upd	update-based protocol
swpc	single (1-entry) WPC
dwpc	double (2-entry) WPC
df	dirty-data filtering
pf	private-data filtering

Table 4. Protocol abbreviations.

the hardware DSM) and are omitted because of space.

On average the inv-64 protocol is 45 percent slower than the hardware DSM system. This overhead is reduced to 32 percent when the inv-dwpc-64 protocol is used. Optimal coherence unit size, number of WPC entries and instrumentation optimizations further improve the invalidation-based DSZOOM performance with almost 30 percent. The invalidation-based protocol actually outperforms the hardware DSM system when run with fft, ocean-c and ocean-nc. While an invalidation-based coherence protocol together with "the best" coherence unit size offers stable performance, some applications show peak performance when run in an update-based environment as long as bandwidth usage is kept low. The update-based protocol is able to outperform the hardware DSM system for ocean-c, ocean-nc and radiosity. (The reason why inv-64 is better than the hardware DSM when run on ocean-nc is that the -fno-delayed-branch actually improves performance on this particular application.) Number of WPC entries and the private-data filter are the most important optimizations while in update mode.

Program	PROFILED	BEST
fft	inv-dwpc-2048	inv-dwpc-2048
lu-c	inv-swpc-2048	inv-dwpc-2048
lu-nc	inv-swpc-128	inv-swpc-128
radix	inv-64	inv-128
barnes	upd-swpc-64	upd-df-swpc-512
fmm	inv-swpc-64	inv-swpc-64
ocean-c	upd-pf-dwpc-512	upd-pf-dwpc-1024
ocean-nc	inv-dwpc-2048	inv-dwpc-1024
radiosity	upd-swpc-64	upd-swpc-64
raytrace	upd-swpc-64	inv-swpc-64
water-nsq	upd-pf-swpc-64	upd-pf-swpc-256
water-sp	upd-pf-swpc-64	upd-pf-swpc-64

Table 5. Classification results from the profile feedback run and the best coherence setting.

Maybe the most notable performance feature is the similarity between the performance of the best DSZOOM protocol (BEST) and the Sun WildFire system (HW-DSM). The performance of the two systems is within 30 percent of each other for all applications except radix and lu-c. Note also that the continuous and non-continuous versions of lu, ocean and water all achieve a similar performance compared with the hardware DSM. This is typically not the case for traditional software DSMs. On average, DSZOOM is 11 percent slower than the hardware DSM. When radix, the application with the worst locality, is omitted, this slowdown is reduced to only 5 percent.

6. Consistency, Deadlock, and Scalability

6.1. Memory Consistency

The invalidation-based protocol of the base architecture (without a WPC implementation) maintains sequential consistency (SC) [22] by requiring all acknowledges from the sharing nodes to be received before a global store request is granted. Introducing the WPC in an invalidation-based environment will not weaken the memory model. The WPC protocol still requires all the remotely shared copies to be destroyed before granting the write permission. Of course, if the memory model of each node is weaker than sequential consistency, it will dictate the memory model of the system.

For an update-based system without load instrumentation, such as the one we present in Section 3.1, the sequential consistency property is sacrificed. Our update-based software DSM system with a 1-entry WPC and a 64 bytes coherence unit size implements processor consistency (PC) [13, 12]. The memory con-

```
01: /* P0's code */      11: /* P1's code */
02: a = 1;               12: b = 1;
03: while (flag != 1)    13: flag = 1;
04:     ; /* wait */     14: ...
05: ...
```

Figure 4. WPC-deadlock code example.

sistency model gets more relaxed if more than one WPC entry or a coherence unit size larger than 64 bytes is used. This consistency model is similar to weak-ordering (WO) [8]. Our PC and WO systems do not implement causal correctness [29].

6.2. Deadlock Avoidance Mechanisms

To avoid WPC related deadlocks, our runtime system releases a processor's WPC entries at synchronization points, at failures to acquire MTAG/directory entries and at thread termination. However, since SC and PC are supported, flag synchronization not visible to the runtime system can occur. The code in Figure 4 can for example lead to deadlock. *a*, *b* and *flag* are all global variables initially assigned the value zero. *a* and *b* are both located on the same coherence unit *u*. *flag* is located on coherence unit $v \neq u$. Let two processors, *P0* and *P1*, execute the code shown in Figure 4. *P0* enters the code first (executes line 02) and assigns *a* the value one. This implies that *P0* puts *a*'s coherence unit *u* in its WPC. When *P0* reaches line 03, it starts to spin on the shared variable *flag*, waiting for *P1*. *P1* enters the code (line 12) and tries to obtain write permission for *b*. However, since the directory state for *b*'s coherence unit *u* is locked and cached by *P0*'s WPC, we have a deadlock!

These WPC related deadlocks are easily avoided with extra runtime system support. We have in an earlier study discussed three possible mechanisms [39]. However, we are convinced that the simplest and best solution is to implement WPC deadlock avoidance in the instrumentation tool. A WPC FIFO replacement policy together with simple basic-block analysis can be used to guarantee that all MTAG locks are released before flag synchronizations (not currently implemented). For simplicity, our instrumentation tool is manually guided in the two applications (*barnes* and *fmm*) that use flag synchronization.

6.3. Protocol Scalability

This paper only presents data for a 2-node system since our WildFire machine only contains two E6000 nodes. We have used "virtual clustering" [38] to show that our invalidation-based protocol scales with number of nodes. This data is omitted since this paper is focused on the hardware comparison and because of limited space. However, we do not believe that our update-based protocol will scale for a large number of nodes. The reason why we have not used virtual clustering emulation while testing the update-based protocols is because it is very difficult to model bandwidth in an accurate way. It would be very interesting to test how a WPC-based store buffer and bandwidth filters will affect update scalability. We consider this evaluation as future work.

7. Related Work

Traditional implementations of software-based shared memory rely on virtual memory hardware to detect when coherence activity is needed. Early page-based systems [24] suffer from false sharing that arises from fine-grain sharing of data within a page. Two main research directions have evolved to improve the performance of software shared memory implementations: relaxing consistency models [3, 19, 41, 27] and providing fine-grained access control [30, 27].

Page-based systems often rely on week memory consistency models and multiple writer protocol to manage the false sharing introduced by their large coherence unit [10, 35, 1]. Carter et al. [3] introduce the release consistency (RC) model in shared virtual memory. Lazy release consistency (LRC) was introduced by Keleher et al. [19] and home based lazy released consistency (HLRC) by Zhou et al. [41]. The majority of systems implement numerous coherence strategies/protocols. For example, Munin [3] implements both invalidate- and update-based protocols (including delayed-update and write-shared protocols).

First of all, our DSM proposal differs from these systems because it is a fine-grain approach that shows much more predictable performance for applications with fine-grain synchronization. In addition, our system can run multiple memory consistency models, including SC, PC and WO, with reasonable performance while page-based systems often rely on RC, LRC or HLRC protocols. Our private-data bandwidth filter is similar to Munin's timeout mechanism that makes it possible to only update nodes that actually use data. However, contrary to Munin, our private-data filter is completely synchronous, and hence, removes all asynchronous protocol messaging. Moreover, it is designed to be used with a fine-grain system and can be run in processor consistency mode whereas Munin relies on release consistency. We compare our system with a hardware DSM while running unmodified applications

with and without fine-grain sharing and synchronization patterns.

Fine-grained software DSMs maintain coherence by instrumenting memory operations in the programs [32, 27, 28]. These systems usually provide stable and predictable performance. However, the instrumentation cost for most of the systems is not negligible. An interesting comparative study of two mature software-based systems from the late 90s shows that the performance gap between fine- and coarse-grain software DSMs can be bridged by adjusting coherence unit size, program restructuring and relaxing memory consistency models [9].

In a early version of Blizzard [11], application specific software-based protocols, which provided very high performance, were implemented and evaluated. Shasta [27, 28] implements support for multiple coherence granularities within a single application. This mechanism is exposed to the programmer through multiple memory allocation functions. Zhou et al. [42] presents performance tradeoffs for relaxed consistency and coherence granularity on a platform that provides access control in hardware but runs coherence protocols in software. Their study focuses on coherent shared memory systems with a fixed coherence granularity (64, 256, 1,024, and 4,096 bytes). The results show that no single combination of protocol and granularity performs the best for all SPLASH-2 [37] applications studied.

Our DSZOOM system differs from all these systems because it uses a synchronous directory protocol and since the virtual memory system is used to enhance performance. Where the other fine-grain systems use user-level hand optimized coherence protocols or application rewrite to enhance coherence protocol performance, we propose the use of coherence profiling and coherence flags. This is the first comparison (to our knowledge) with a real hardware DSM machine. It is also the first study in which an all-software system is able to outperform an all-hardware DSM!

The Stanford FLASH [21] project addresses concerns with hardwired protocols by migrating the entire protocol-engine to software handlers executed on a separate processor. SMTp is a more recent proposal [5] in which the coherence protocol is run by one SMT thread. Multiple systems implement a simple hardware directory protocol backed up with software handlers. The protocol described by Hill et al. [17] uses a single hardware pointer. In addition, the programmer or compiler can annotate programs with Check-In/Check-Out (CICO) directives to minimize the number of software traps. Chaiken and Agarwal [4] describe performance and cost of software extended coherence shared memory as implemented in Alewife. Grahn and Stenström [14] extends Chaiken and Agarwal's work. While most of the findings in this paper can be implemented in such systems, they rely on modified memory controllers [5], protocol processors [21] and/or hardware support for n pointers in hardware [4, 14, 17].

8. Conclusions

This paper presents a highly flexible all-software distributed shared memory system that combines code instrumentation and page protection mechanisms. Fine-grain access control checks applied at shared loads and stores avoid false sharing without any application rewriting or memory model weakening. The paper also presents two protocol classes that are based on classical invalidate/update schemes. The page protection mechanism is applicable in both cases to minimize unnecessary global memory replication and, in particular, as an efficient bandwidth-reduction technique for update-based protocols. Several other reduction techniques are presented, such as all-software store buffering, dirty- and private-data filtering.

The paper demonstrates the flexibility of this approach with two simple invalidate/update synchronous protocols that support more then 50 different combinations of coherence flags. A very simplistic low-overhead profiling mode of the system is capable to find appropriate coherence flags for the studied applications. This is an automatic single-run process based on the profile run feedback.

The system demonstrates stable and predictable performance for all applications studied. In fact, several applications run faster with this system than with a much more expensive hardware-based DSM with an identical interconnect. On average, our software DSM is 11 percent slower than the hardware DSM. When `radix`, the application with the worst locality, is omitted, this slowdown is reduced to only 5 percent.

References

[1] A. Bilas et al. Using Network Interface Support to Avoid Asynchronous Protocol Processing in Shared Virtual Memory Systems. In *ISCA '99*, pages 282–293, May 1999.

[2] W. W. Carlson et al. Introduction to UPC and Language Specification. Technical Report CCS-TR-99-157, The George Washington University, May 1999.

[3] J. B. Carter et al. Implementation and Performance of Munin. In *SOSP '91*, pages 152–164, Oct. 1991.

[4] D. Chaiken and A. Agarwal. Software-Extended Coherent Shared Memory: Performance and Cost. In *ISCA '94*, pages 314–324, Apr. 1994.

[5] M. Chaudhuri and M. Heinrich. SMTp: An Architecture for Next-generation Scalable Multi-threading. In *ISCA '04*, pages 124–135, June 2004.

[6] D. Chiou et al. StarT-NG: Delivering Seamless Parallel Computing. In *Euro-Par '95*, pages 101–116, Aug. 1995.

[7] L. Dagum and R. Menon. OpenMP: An Industry-Standard API for Shared Memory Programming. *IEEE Computational Science and Engineering*, 5(1):46–55, Jan.-Mar. 1998.

[8] M. Dubois et al. Memory Access Buffering in Multiprocessors. In *ISCA '86*, pages 434–442, June 1986.

[9] S. Dwarkadas et al. Comparative Evaluation of Fine- and Coarse-Grain Approaches for Software Distributed Shared Memory. In *HPCA-5*, pages 260–269, Jan. 1999.

[10] A. Erlichson et al. SoftFLASH: Analyzing the Performance of Clustered Distributed Virtual Shared Memory. In *ASPLOS-VII*, pages 210–220, Oct. 1996.

[11] B. Falsafi et al. Application-Specific Protocols for User-Level Shared Memory. In *SC '94*, pages 380–389, Nov. 1994.

[12] K. Gharachorloo et al. Memory Consistency and Event Ordering in Scalable Shared-memory Multiprocessors. In *ISCA '90*, pages 15–26, May 1990.

[13] J. R. Goodman. Cache Consistency and Sequential Consistency. Technical Report 61, SCI Committee, Mar. 1989.

[14] H. Grahn and P. Stenström. Efficient Strategies for Software-Only Protocols in Shared-Memory Multiprocessors. In *ISCA '95*, pages 38–47, June 1995.

[15] E. Hagersten and M. Koster. WildFire: A Scalable Path for SMPs. In *HPCA-5*, pages 172–181, Jan. 1999.

[16] J. L. Hennessy and D. A. Patterson. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann, 3rd edition, 2003.

[17] M. D. Hill et al. Cooperative Shared Memory: Software and Hardware for Scalable Multiprocessors. *ACM Transactions on Computer Systems*, 11(4):300–318, Nov. 1993.

[18] InfiniBand Trade Association, InfiniBand Architecture Specification, Release 1.2, Oct. 2004. Available from `http://www.infinibandta.org`.

[19] P. Keleher et al. Lazy Release Consistency for Software Distributed Shared Memory. In *ISCA '92*, pages 13–21, May 1992.

[20] K. Krewell. Best Servers of 2004: Where Multicore Is the Norm. In *Microprocessor Report*, Jan. 2005.

[21] J. Kuskin et al. The Stanford FLASH Multiprocessor. In *ISCA '94*, pages 302–313, Apr. 1994.

[22] L. Lamport. How to Make a Multiprocessor Computer That Correctly Executes Multiprocess Programs. *IEEE Transactions on Computers*, C-28(9):690–691, Sept. 1979.

[23] J. R. Larus and E. Schnarr. EEL: Machine-Independent Executable Editing. In *PLDI '95*, pages 291–300, June 1995.

[24] K. Li. *Shared Virtual Memory on Loosely Coupled Multiprocessors*. PhD thesis, Department of Computer Science, Yale University, Sept. 1986.

[25] L. W. McVoy and C. Staelin. lmbench: Portable Tools for Performance Analysis. In *USENIX Technical Conference '96*, pages 279–294, Jan. 1996.

[26] Z. Radović and E. Hagersten. Removing the Overhead from Software-Based Shared Memory. In *SC '01*, Nov. 2001.

[27] D. J. Scales et al. Shasta: A Low-Overhead Software-Only Approach to Fine-Grain Shared Memory. In *ASPLOS-VII*, pages 174–185, Oct. 1996.

[28] D. J. Scales et al. Fine-Grain Software Distributed Shared Memory on SMP Clusters. In *HPCA-4*, pages 125–136, Feb. 1998.

[29] C. Scheurich. *Access Ordering and Coherence in Shared Memory Multiprocessors*. PhD thesis, University of Southern California, May 1989.

[30] I. Schoinas et al. Fine-grain Access Control for Distributed Shared Memory. In *ASPLOS-VI*, pages 297–306, Oct. 1994.

[31] I. Schoinas et al. Implementing Fine-Grain Distributed Shared Memory On Commodity SMP Workstations. Technical Report #1307, Computer Sciences Department, University of Wisconsin–Madison, Mar. 1996.

[32] I. Schoinas et al. Sirocco: Cost-Effective Fine-Grain Distributed Shared Memory. In *PACT '98*, pages 40–49, Oct. 1998.

[33] A. Singhal et al. Gigaplane: A High Performance Bus for Large SMPs. In *Proceedings of IEEE Hot Interconnects IV*, pages 41–52, Aug. 1996.

[34] S. J. Sistare and C. J. Jackson. Ultra-High Performance Communication with MPI and the Sun Fire Link Interconnect. In *SC '02*, Nov. 2002.

[35] R. Stets et al. Cashmere-2L: Software Coherent Shared Memory on a Clustered Remote-Write Network. In *SOSP '97*, pages 170–183, Oct. 1997.

[36] D. L. Weaver and T. Germond, editors. *The SPARC Architecture Manual, Version 9*. PTR Prentice Hall, 2000.

[37] S. C. Woo et al. The SPLASH-2 Programs: Characterization and Methodological Considerations. In *ISCA '95*, pages 24–36, June 1995.

[38] D. Yeung et al. Multigrain Shared Memory. *ACM Transactions on Computer Systems*, 18(2):154–196, May 2000.

[39] H. Zeffer et al. Exploiting Spatial Store Locality through Permission Caching in Software DSMs. In *Euro-Par '04*, pages 551–560, Aug. 2004.

[40] H. Zeffer et al. Flexibility Implies Performance. Technical Report 2005-013, Department of Information Technology, Uppsala University, Apr. 2005.

[41] Y. Zhou et al. Performance Evaluation of Two Home-Based Lazy Release Consistency Protocols for Shared Virtual Memory. In *OSDI '96*, pages 75–88, Oct. 1996.

[42] Y. Zhou et al. Relaxed Consistency and Coherence Granularity in DSM Systems: A Performance Evaluation. In *PPOPP '97*, pages 193–205, June 1997.