Speeding-up multiprocessors running DBMS workloads through coherence protocols

Pierfrancesco Foglia*
Dipartimento di Ingegneria dell'Informazione,
University of Pisa, Via Diotisalvi 2, Pisa 56100, Italy
E-mail: foglia@iet.unipi.it
*Corresponding author

Roberto Giorgi
Dipartimento di Ingegneria dell'Informazione,
University of Siena,
Via Roma 56, Siena 53100, Italy
E-mail: giorgi@unisi.it

Cosimo Antonio Prete
Dipartimento di Ingegneria dell'Informazione,
University of Pisa,
Via Diotisalvi 2, Pisa 56100, Italy
E-mail: prete@iet.unipi.it

Abstract: In this work, it is shown how a DBMS workload, running on a shared-bus shared-memory multiprocessor, can be accelerated by adding simple support to the MESI coherence protocol. As a DBMS workload, we choose the TPC-D benchmark running on the PostgreSQL DBMS. Results show that, for a DSS workload, the use of a WU protocol with a selective invalidation strategy for private data improves performance because of the access pattern to shared data and the lower bus utilisation due to the absence of invalidation miss, when the contribution of passive sharing is eliminated. In the 16 processor case, the advantage can be quantified in a 20% of increased performance. Finally, it is shown how results can be extended to other DBMS workloads.

Keywords: shared-bus multiprocessors; DBMS and DSS systems; cache memory; coherence protocol; sharing analysis.

Reference to this paper should be made as follows: Foglia, P., Giorgi, R. and Prete, C.A. (2004) 'Speeding-up multiprocessors running DBMS workloads through coherence protocols', *Int. J. High Performance Computing and Networking*, Vol. 1, Nos. 1/2/3, pp.17–32.

Biographical notes: Pierfrancesco Foglia is an Assistant Professor at the Department of Information Engineering, University of Pisa, Italy. He received PhD and MS in computer engineering from University of Pisa, Italy. His current interests involve computer architecture themes such as embedded systems, memory system performance, and high-performance systems for web, and database applications. He is also interested in network protocols and network management systems. He is member of IEEE, IEEE Computer Society, and ACM.

Roberto Giorgi is an Assistant Professor at the Department of Information Engineering, University of Siena, Italy. He was Research Associate at the University of Alabama in Huntsville, USA. He received his PhD in computer engineering from University of Pisa, Italy, and his MS in electronics engineering, Magna cum Laude from University of Pisa,

Italy. His current interests involve computer architecture themes such as embedded systems, memory system performance, high-performance systems for web, and database applications. He is member of IEEE, IEEE Computer Society, and ACM.

Cosimo Antonio Prete is Full Professor of computer systems at the Department of Information Engineering at the University of Pisa, Italy. His research interests include multiprocessor architectures, cache memory, performance evaluation, and embedded systems. He has performed research in programming environments for distributed systems, commit protocols for distributed transactions, cache memory architecture, and coherence protocols for tightly coupled multiprocessor systems. He is also an expert on the Open Microprocessor Systems Initiative for the Commission of the European Communities. He earned his undergraduate degree in electronic engineering cum laude in 1982 and PhD from the University of Pisa in 1989.

1 INTRODUCTION

AN ever-increasing number of multiprocessor server systems shipped today run commercial workloads (Keeton et al., 2003). These workloads include database applications such as online transaction processing (OLTP) and decision support system (DSS), file servers, and application servers (Stenström et al., 1997). Nevertheless, technical workloads were widely used to drive the design of current multiprocessor systems (Stenström et al., 1997; Cvetanovic and Bhandarkar, 1994; Keeton et al., 2003) and different studies have shown that commercial workloads exhibit different behaviour from technical ones (Maynard et al., 1994; Keeton et al., 1998).

The simpler design for a multiprocessor system is a shared-bus shared-memory architecture (Tanenbaum, 2001). In shared-bus systems, processors access the shared memory through a shared bus. The bus is the bottleneck of the system, since it can easily reach a saturation condition, thus limiting the performance and the scalability of the machine. The classical solution to overcome this problem is the use of per-processor cache memories (Hennessy and Patterson, 2002). Cache memories introduce the coherency problem and the need for adopting adequate coherence protocols (Tomasevic and Milutinovic, 1994a, 1994b). The main coherence protocol classes are write-update (WU) and write-invalidate (WI) (Tomasevic and Milutinovic, 1993). WU protocols update the remote copies on each write involving a shared copy, whereas WI protocols invalidate remote copies in order to avoid updating them. Coherence protocols generate several bus transactions, thus accounting for a non-negligible overhead in the system (coherence overhead). Coherence overhead may have a negative effect on the performance and, together with the accesses pattern to application data, determines the best protocol choice for a given workload (Foglia et al., 1998; Eggers and Katz, 1989b; Eggers and Katz, 1988). Different optimisations to minimise coherence overhead have been proposed (Eggers and Jeremiassen, 1991; Tomasevic and Milutinovic, 1993; Torrellas et al., 1994; Jeremiassen and Eggers, 1995; Giorgi and Prete, 1999), also acting at compile time and architectural level (as the adoption of adequate coherence protocols (Tomasevic and Milutinovic, 1996; Giorgi and Prete, 1999)).

When the performance achieved by shared-bus shared-memory multiprocessor is not sufficient, and this is typical for DBMS applications, an SMP (symmetrical multi-processing, which includes shared-bus shared-memory architectures) or a NUMA (non-uniform memory access) approach can be utilised (Culler and Singh, 1998; Tanenbaum, 2001; Hennessy and Patterson, 2002). In the first case, a crossbar switch interconnects the processing elements, in the second an interconnection network. Such solutions increase the communication bandwidth among elements, thus allowing more CPUs to be added to the system, but at the cost of more expensive and complex communication networks. In both the designs, the basic building bock (node) may be a single processor system or, better, a shared-bus shared-memory multiprocessor (examples are the HP V-Class and the SGI Origin families of multiprocessors (Yu et al., 2002)). In this way, by adding high-performance nodes, we can achieve the desired level of performance with only a little number of elements, simplifying the crossbar or IC network design, and lowering the price of the whole system. Unfortunately, due to limited bus bandwidth, only a small number of CPU (max four for the Pentium family of CPU (Shanley and Mindshare Inc., 1999) may be included in the single node.

The aim of this paper is to analyse the scalability of shared-bus shared-memory multiprocessors running commercial workloads and DBMS applications in particular, and to investigate solutions, as concerns the memory subsystem, which can increase the processing power of such architectures. In this way, we can meet the performance requirement of commercial applications with a single shared-bus shared-memory machine, or we can adopt more performing nodes in an SMP or NUMA design, allowing simpler and cheaper design of switch or IC networks.

Important database applications include online transaction processing (OLTP) and decision support systems (DSS). DSS applications are utilised to extract management information from databases of historical data; they are characterised by long-running read-only queries (Keeton et al., 2003). OLTP applications are utilised by users to execute transactions against a database (for instance, in banking systems, air flight reservation systems, commercial order-entry environment, etc). Different

from DSS systems, OLTP applications include a large number of concurrent, relatively short, and updating queries (Keeton et al., 2003). In consequence, performance of OLTP queries depends significantly also on the locking mechanism implemented in the DBMS system. Only commercial DBMS exhibits adequate locking scheme to run OLTP application at an adequate level of performance; hence, every analysis of OLTP system should consider commercial DBMS. Unfortunately, the methodology of evaluation at our disposal requires application source code, which is not available for commercial DBMS system. So, in the following, we restrict our analysis to DSS system. Nevertheless, we draw, at the end of the paper, some consideration on how our results can change when analysing an OLTP system.

A typical DSS activity is performed via a query to look for ways to increase revenues, such as an SQL query to quantify the amount of revenue increase that would have resulted from eliminating a company discount in a given percentage in a given year (Ballinger, 2001). As mentioned earlier, such queries are, in most part, read-only queries. Consequently, as already observed in Trancoso et al. (1997) and Yu et al. (2002), coherence activity and transactions involve essentially DBMS metadata (i.e., data structures that are needed by the DBMS to work rather than to store the data) and particularly data structure needed to implement software locks (Trancoso et al., 1997; Yu et al., 2002). The type of access pattern to those data is one-producer/ many-consumers. Such pattern may advantage solution based on the write-update family of coherence protocols, especially if there is a high number of processes (the 'consumers') that concurrently read shared data (Eggers and Katz, 1988). These considerations can suggest the adoption of other coherence protocols, with respect to the usually utilised MESI WI protocol (Sweazey and Smith, 1986).

In our evaluation, the DSS workload of the system is generated by running all the TPC-D queries (Transaction Processing Performance Council, 1995) (through the PostgreSQL (Yu and Chen, 1995) DBMS) and several Unix utilities, which both access the file system and interface the DBMS with system services running concurrently. Our methodology relies on trace-driven simulation, by means of the 'Trace Factory' environment (Giorgi et al., 1997), and on specific tools for the analysis of coherence overhead (Foglia, 2001).

The performance of the memory subsystem – and therefore of the whole system – depends on cache parameters, coherence management techniques, and it is influenced by operating system activities like process migration, cache affinity scheduling, kernel/user code interference, and virtual memory mapping. The importance of considering operating system activity has been highlighted in previous works (Chapin et al., 1995; Chandra et al., 1994; Torrellas et al., 1992). Process migration, needed to achieve load balancing in such systems, increases the number of cold/conflict misses and also generates useless coherence overhead, known as passive sharing overhead (Giorgi and Prete, 1999). As passive sharing overhead dramatically decreases the performance of pure WU protocols (Giorgi and Prete, 1999; Prete et al., 1997), we considered in our analysis a hybrid WU protocol, PSCR (Giorgi and Prete, 1999), which adopts a selective invalidation strategy for private data and a hybrid WI protocol, AMSD (Cox and Fowler, 1993; Stenström et al., 1993), which deals with migration of data and, then, of processes.

Our results show that, in the case of a four-processor configuration, performance of DSS workload is moderately influenced by cache parameters and the influence of coherence protocol is minimal. In 16-processor configurations, the performance differences due to the adoption of different architectural solutions are significant. In high-end architectures, MESI is not the best choice and workload changes can nullify the action of affinity-based scheduling algorithms. Architectures based on a write-update protocol with a selective invalidation strategy for private data outperform the ones based on MESI of about 20%, and adapt better to workload.

The rest of the paper is organised as follows. In Section 2, we discuss architectural parameters and issues related to the coherence overhead. In Section 3, we report the results of studies related to the analysis of workloads similar to ours, differentiating our contribution. In Section 4, we present our experimental setup and methodology. In Section 5, we discuss the results of our experiments. In Section 6, we draw conclusions.

2 ISSUES AFFECTING COHERENCE OVERHEAD

Cache coherency maintaining involves a number of bus operations. Some of them are overhead that adds up to the basic bus traffic (the traffic necessary to access main memory). Three different sources of sharing may be observed:

- true sharing (Torrellas et al., 1990a, 1994b), which occurs when the same cached data item is referenced by different processes concurrently running on different processors
- false sharing (Torrellas et al., 1990a, 1994b), which occurs when several processors reference a different data item belonging to the same memory block separately
- passive (Prete et al., 1997) or process-migration (Agarwal, 1989) sharing, which occurs when a memory block, though belonging to a private area of a process, is replicated in more than one cache, as a consequence of the migration of the owner process.

The main coherence protocol classes are write-update (WU) and write-invalidate (WI) (Tomasevic and Milutinovic, 1993). WU protocols update the remote copies on each write involving a shared copy, whereas WI protocols invalidate remote copies in order to avoid updating them. An optimal selection for the coherence protocol can be made by considering the traffic induced by

the two approaches in case of different sharing and its effects on performances (Foglia et al., 1998). The coherence overhead induced by a WU protocol is due to all the operations needed to update the remote copies, whereas a WI protocol invalidates remote copies and processors generate a miss on the access to the invalidated copy (invalidation miss). The access patterns to shared data determine the coherence overhead. In particular, fine-grain sharing denotes high contention for shared data; sequential sharing is characterised by long sequences of writes to the same memory item performed by the same processor. A WI protocol is adequate in the case of sequential sharing, whilst, in general, a WU protocol performs better than WI for program characterised by fine-grain sharing (Eggers and Katz, 1988, 1989a, 1989b) (alternatively, a metric based on the write-run and external re-reads model can be used to estimate the best protocol choice (Eggers, 1991)). In our evaluation, we considered MESI protocol as baseline, because it is used in most of the high-performance processors today (like AMD K5 and K6, PowerPC series, SUN UltraSparc II, SGI R10000, Intel Pentium, Pentium Pro series (Pro, II, III), Pentium 4, and IA-64/Itanium). Then, based on the previous considerations on the access pattern of DSS workloads, we included a selective invalidation protocol based on a WU policy for manage shared data, which limits most effects of process migration (PSCR (Giorgi and Prete, 1999)) and a WI protocol specifically designed to treat the data migration (AMSD (Cox and Fowler, 1993; Stenström et al., 1993)).

Our implementation of MESI uses the classical MESI protocol states (Sweazey and Smith, 1986) and the following bus transactions: read-block (to fetch a block), read-and-invalidate-block (to fetch a block and invalidate any copies in other caches), invalidate (to invalidate any copy in other caches), and update-block (to write back dirty copies when they need to be replaced). This is mostly similar to the implementation of MESI in Pentium Pro (Pentium II) processor family (Shanley and Mindshare Inc., 1999). The invalidation transaction used to obtain coherency has, as a drawback, the need to reload a certain copy, if a remote processor uses again that copy, thus generating a miss (Invalidation Miss). Therefore, MESI coherence overhead (that is the transactions needed to enforce coherence) is due both to Invalidate Transactions and Invalidation Misses.

PSCR (Passive Shared Copy Removal) adopts a selective invalidation scheme for private data, and uses a write-update scheme for shared data, although it is not a purely write-update protocol. A cached copy belonging to a process private area is invalidated locally as soon as another processor fetches the same block (Giorgi and Prete, 1999). This technique reduces coherence overhead (especially passive sharing overhead), otherwise dominant in a pure WU protocol (Prete et al., 1997a). Invalidate transactions are eliminated and coherence overhead is due to Write Transactions and Invalidation Misses caused by local invalidation (Passive Sharing Misses).

AMSD (Cox and Fowler, 1993; Stenström et al., 1993) is designed for Migratory Sharing, which happens when the control over shared data migrates from one process to another running on a different processor. The protocol identifies migratory-shared data dynamically to reduce the cost of moving them. The implementation relies on an extension of a common MESI protocol. Though designed for migratory sharing, AMSD may have some beneficial effects also on passive sharing. AMSD coherence overhead is due to Invalidate Transactions and Invalidation Misses.

The process scheduling strategy, cache parameters, and the bus features also influence the coherence overhead and, therefore, the multiprocessor performance. The process scheduling guarantees the load balancing among the processors by scheduling a ready process on the first available processor. Although cache affinity is used, a process may migrate on a different processor rather than on the last used processor, producing at least two effects:

- some misses when it restarts on a new processor (context-switch misses)
- useless coherence transactions due to passive sharing.

These situations may become frequent due to dynamicity of a workload driven by the user requests, like DSS ones. Process migration influences also access patterns to data and, therefore, the coherency overhead.

All the factors described above have important consequences on the global performance that we evaluate in the Section 5.

3 RELATED WORK ON DBMS SYSTEMS

In this section, we consider a summary of main results of evaluations for DSS and similar workloads on several multiprocessor architectures and operating systems. The research of a realistic evaluation framework for shared-memory multiprocessor evaluations (Stenström et al., 1997; Keeton et al., 2003) motivated many studies that consider benchmarks like TPC series (including DSS, OLTP, WEB-server benchmarks) representative of commercial workloads (Trancoso et al., 1997; Barroso et al., 1998; Cao et al., 1999; Ranganathan et al., 1998; Lovett and Clapp, 1996; Lo et al., 1998; Keeton et al., 1998).

Barroso et al. (1998) evaluate an Alpha 21164-based SMP memory system through hardware counter measurements and SimOS simulations. Their system was running Digital-UNIX and Oracle-7 DMBS. The authors consider a TPC-B database (OLTP benchmark), TPC-D (DSS queries), and the Altavista search engine. They found that memory accounts for 75% of stall time. In the case of TPC-D and Altavista workloads, the size and latency of the on-chip caches influences mostly the performance. The number of processes per processor, which is usually kept high in order to hide I/O latencies, also significantly influences cache behaviour. When using 2-, 4-, 6-, and 8-processor

configurations, they found that coherency miss stalls increase linearly with the number of processors. Beyond an 8M-byte outer level cache, they observed that true sharing misses limit performance.

Ranganathan et al. (1998) consider both an OLTP workload (modelled after TPC-B (Transaction Processing Performance Council, 1994)) and a DSS workload (query 6 of TPC-D (Transaction Processing Performance Council, 1995; Ranganathan et al., 1998). Their study is based on trace-driven simulation, where traces are collected on a four-processor AlphaServer4100 running Digital Unix and Oracle 7 DBMS. The simulated system is a CC-NUMA shared-memory multiprocessor with advanced ILP support. Results, on a four-processor system and an ILP configuration with four-way issue, 64-entry instruction window, four outstanding misses, provide already significant benefits for OLTP and DSS workload. Such configurations are even less aggressive than ILP commercial processor like Alpha 21264, HP-PA 8000, MIPS R10000 (Yeager, 1996). The latter processor, used in our evaluation, makes us reasonably safe that this processor architecture is sound for investigation in the memory subsystem.

Another performance analysis of OLTP (TPC-B) and DSS (query 6 of TPC-D) workloads via simulation is presented in Lovett and Clapp (1996). The simulated system is a commercial CC-NUMA multiprocessor, constituted by four-processor SMP nodes connected using a Scalable Coherent Interface based coherent interconnects. Coherence protocol is directory-based. Each node is based on a Pentium Pro processor with a 512K-byte, four-way set associative cache. Results show that TPC-D miss rate is much lower and performance is less sensitive to L2 miss latency than in the OLTP (TPC-B) experiments. They analyse also scalability exhibited by such workloads. The speed-up of the DSS workload is near to the theoretical ones. Results show that scalability strongly depends on miss rate.

Trancoso et al. (1997) study the memory access patterns of a TPC-D-based DSS workload. The DBMS is Postgres95 running on a simulated four-processor CC-NUMA multiprocessor. For cache block sizes ranging from 4 to 128 bytes and cache capacities from 128 K-bytes to 8 M-bytes, the main results are that both large cache blocks and data prefetching help, due to spatial locality of index and sequential queries. Coherence misses can be more than 60% of total misses in queries that uses index scan algorithms for select operations.

Cao et al. (1999) examine a TPC-D workload executing on a Pentium-Pro four-processor system, with Windows NT and MS SQL Server. Their goal is to characterise this DSS system on a real machine, in particular, regarding processor parameters, bus utilisation, and sharing. Their methodology is based on hardware counters. They found that kernel time is negligible (less than 6%). Major sources of processor stalls are instruction fetch and data miss in outer level caches. They found lower miss rates for data caches in comparison with other studies on TPC-C (Barroso et al., 1998; Keeton et al., 1998). This is due to the smaller working set of TPC-D compared with TPC-C.

Lo et al. (1998) analyse the performance of database workloads running on simultaneous multithreading processors (Eggers et al., 1997) – an architectural technique to hide memory and functional units latencies. The study is based on trace driven of a four-processor AlphaServer4100 and Oracle 7 DBMS as in Barroso et al. (1998). They consider both an OLTP workload (modelled after the TPC-B (Transaction Processing Performance Council, 1994) benchmark) and a DSS workload (query six of the TPC-D (Transaction Processing Performance Council, 1995) benchmark). Results show that while DBMS workloads have larger memory footprints, there is a substantial data reuse in a small working set.

Summarising, these studies considered DBMS workloads but they were mostly limited to four-processor systems, did not consider the effects of process migration, and did not correlate the amount of sharing to the performance of the system. As we have more processors, it becomes crucial to characterise further the memory subsystem. In our work, we investigated both 4- and 16-processor configurations, finding that larger caches may have several drawbacks due to coherence overhead. This is mostly related to the use of shared structures like indices and locks. In Section 5, we classify the sources of this overhead and propose solutions to overcome limitations to the performance related to process migration.

4 METHODOLOGY AND WORKLOAD

The methodology that we used (Trace Factory, Giorgi et al. 1997) is based on trace-driven simulation (Stunkel et al., 1991; Prete et al., 1995; Uhlig and Mudge, 1997) and on the simulation of the three kernel activities that most affect performance: system calls, process scheduling, and virtual-to-physical address translation.

The approach used is to produce a source trace – a sequence of memory references, system-call positions (and synchronisation events if needed) – by means of a tracing tool. Trace Factory then models the execution of complex multiprogrammed workloads by combining multiple source traces and simulating system calls (which could also involve I/O activity), process scheduling, and virtual-to-physical address translation. Finally, Trace Factory produces the references (target trace) furnished as input to a memory-hierarchy simulator (Prete et al., 1995). Trace Factory generates references according to an on-demand policy: it produces a new reference when simulator requests one, so that the timing behaviour imposed by the memory subsystem conditions the reference production (Giorgi et al., 1997). Process management is modelled by simulating a scheduler that dynamically assigns a ready process. Virtual-to-physical address translation is modelled by mapping sequential virtual pages into non-sequential physical pages. A careful evaluation of this methodology has been carried out by Giorgi et al. (1997).

The workload considered in our evaluation includes DB activity reproduced by means of an SQL server, namely PostgreSQL (Yu and Chen, 1995), which handles the TPC-D (Transaction Processing Performance Council, 1995) queries. We also included Unix utilities that access the file system, interface the various programs running on the system, and reproduce the activity of typical Unix daemons.

PostgreSQL is a public domain DBMS, which relies on server-client paradigm. It consists of a front-end process that accepts SQL queries, and a back-end that forks processes, which manage the queries. TPC-D is a benchmark for DSS developed by the Transaction Processing Performance Council (1995). It simulates an application for a wholesale supplier that manages, sells, and distributes a product worldwide. Following TPC-D specifications, we populated the database via the dbgen program, with a scale factor of 0.1. The data are organised in several tables and accessed by 17 read-only queries and two update queries.

In a typical situation, application and management processes can require the support of different system commands and ordinary applications. To this end, Unix utilities (ls, awk, cp, and rm) and daemons (telnetd, syslogd, crond) have been added to the workload. These utilities are important because they model the 'glue' activity of the system software. These utilities:

- do not have shared data and thus they increase the effects of process migration, as discussed in detail in Section 5
- they may interfere with shared data and code cache-footprint of other applications.

To take into account that requests may be using the same program at different times, we traced some commands in shifted execution sections: initial (beg) and middle (mid).

In our experiments, we generated two distinct workloads. The first workload (DSS26 in Table 2) includes the TPC-D18 source trace (Table 1). TPC-D18 is a multiprocess source trace taking into account the activity of 18 processes. One process is generated by DBMS back-end execution, whilst the other processes are generated by the concurrent execution of the 17 read-only TPC-D queries (TPC-D18 in Table 2). Since we wished to explore critical situations for the affinity scheduling – when the number of process changes, – we also generated a second workload (DSS18 in Table 2) that includes a subset of TPC-D queries (TPC-D12). Table 1 contains some statistics of the uniprocess and multiprocess source traces used to generate the DB workloads (target traces, Table 2). Both source traces related to DBMS activity (TPC-D18, TPC-D12) and the resulting workloads (DSS26 e DSS18) present similar characteristics in terms of read, write, and shared accesses.

Table 1 *Statistics of source traces for some Unix commands and daemons and for multiprocess source traces (PostgreSQL, TPC-D queries) both in case of 64-byte block size and 10,000,000 references per process*

Application	No. of processes	Distinct blocks	Code (%)	Data (%)		Shared blocks	Shared data (%)	
				Read	Write		Access	Write
awk (beg)	1	4,963	76.76	14.76	8.48	n/a	n/a	n/a
awk (mid)	1	3,832	76.59	14.48	8.93	n/a	n/a	n/a
cp	1	2,615	77.53	13.87	8.60	n/a	n/a	n/a
rm	1	1,314	86.39	11.51	2.10	n/a	n/a	n/a
ls – aR	1	2,911	80.62	13.84	5.54	n/a	n/a	n/a
telnetd	1	463	82.75	12.96	4.29	n/a	n/a	n/a
crond	1	2,464	75.86	16.35	7.79	n/a	n/a	n/a
syslogd	1	2,848	80.41	14.96	4.63	n/a	n/a	n/a
TPC-D18	18	139,324	73.05	16.89	10.06	7224	1.58	0.43
TPC-D12	12	93,657	73.04	16.91	10.05	5662	1.55	0.41

Table 2 *Statistics of multiprogrammed target traces (DSS26, 260,000,000 references; DSS18, 180,000,000 references) in case of 64-byte block size*

Workload	No. of processes	Distinct blocks	Code (%)	Data (%)		Shared blocks	Shared data (%)	
				Read	Write		Access	Write
DSS26	26	179,862	74.59	16.26	9.15	7806	1.76	0.53
DSS18	18	124,268	74.00	16.11	8.89	6242	1.63	0.48

5 RESULTS

In this section, we show the memory subsystem performance, for our DSS workload, with a detailed characterisation of coherence overhead and process-migration problems. To this end, we included results for several values of the most influencing cache-architecture parameters. Finally, we considered critical situations for the affinity scheduling and we analysed how the result may be extend to other DBMS workloads. Our results show that solutions that allow us to achieve a higher scalability are possible for this kind of machine – compared to standard solutions – and consequently a greater performance at a reasonable cost.

5.1 Design space of our system

The simulated system consists of *N* processors, which are interconnected to a 128-bit shared bus for accessing shared memory. The following coherence schemes have been considered: AMSD, MESI, and PSCR (more details are in Section 2). We considered two main configurations: a basic machine with four processors and a high-performance one with 16 processors. The scheduling policy is based on cache-affinity; scheduler time-slice is 200,000 references. Cache size has been varied between 512 K bytes and 2 M bytes, while for block size we used 64 bytes and 128 bytes. The simulated processors are MIPS-R10000-like; paging relays on 4 K -page size; the bus is pipelined, supports transaction splitting, and processor-consistency memory model (Gharachorloo et al., 1991); up to eight outstanding misses are allowed (Kroft, 1981). The base case study timings and parameter values for the simulator are summarised in Table 3.

5.2 Performance metrics

To investigate the effects on the performance due to memory subsystem operations, we analysed the causes influencing memory latency and the total execution time. The memory latency depends on the time necessary to perform a bus operation (Table 3) and on the waiting time to access bus (bus latency). Bus latency depends on the amount and kind of bus traffic. Bus traffic is constituted by read-block transactions (issued for each miss), update transactions and coherence transactions (write transactions or invalidate transactions, depending on the coherence protocol). Consequently, miss and coherence transactions affect performance because they affect the processor waiting-time directly (in the case of read misses) and contribute to bus latency. Therefore, to investigate the sources of performance bottlenecks, we reported a breakdown of misses – which includes invalidation misses and classical misses (sum of cold, capacity, and conflict misses) – and the 'number of coherence transaction per 100 memory references' – which includes either write-transactions or invalidate transactions depending on the coherence protocol. The rest of traffic is due to update transactions. Update transactions are a negligible part of bus-traffic (lower than 7% of read-block transactions or about 1% of bus occupancy) and thus they do not influence greatly our analysis.

Table 3 *Input parameters for the multiprocessor simulator (timings are in clock cycles)*

Class	*Parameter*	*Timings*
CPU	Read cycle	2
	Write cycle	2
Cache	Cache size (bytes)	512K, 1M, 2M
	Block size (bytes)	64, 128
	Associativity (number of ways)	1, 2, 4
Bus	Write transaction (PSCR)	5
	Write for invalidate transaction (AMSD, MESI)	5
	Invalidate transaction (AMSD)	5
	Memory-to-cache read-block transaction	72 (block size 64 bytes), 80 (block size 128 bytes)
	Cache-to-cache read-block transaction	16 (block size 64 bytes), 24 (block size 128 bytes)
	Update-block transaction	10 (block size 64 bytes), 18 (block size 128 bytes)

In our analysis, we differentiated between Cold Misses and Capacity+Conflict Misses and Invalidation Misses (Hennessy and Patterson, 2002). Cold Misses include first access misses by a given process when it is scheduled either for the first time or it is rescheduled on another processor (the latter are also known as context-switch misses). Capacity+Conflict Misses include misses caused by memory references of a process competing for the same block (intrinsic interference misses in Agarwal (1989)), and misses caused by references of sequential processes, executing on the same processor and competing for the same cache block (extrinsic interference misses in Agarwal (1989)). Invalidation Miss is due to accesses to data that are to be reused on the same processor, but that have been invalidated in order to maintain coherence. Invalidation Misses are further classified, along with the coherence transaction type, by means of an extension of an existing classification algorithm (Hyde and

Fleisch, 1996). Our algorithm extends this classification to the case of passive sharing, finite size caches, and process migration (Foglia, 2001). In particular, Invalidation Misses are differentiated as true sharing misses, false sharing misses, passive sharing misses. True sharing misses and false sharing misses are classified according to an already known methodology (Torrellas et al., 1994; Eggers and Jeremiassen, 1991; Dubois et al., 1993). Passive sharing misses are invalidation misses generated by the useless coherence maintaining of private data (Giorgi and Prete, 1999). Clearly, these private data could appear as shared to the coherence protocol, because of the process migration. Similarly, the coherence transactions are classified as true, false, and passive sharing transactions either in the case they are invalidation or write transactions (Foglia, 2001).

5.3 ANALYSIS OF THE REFERENCE SYSTEM

We started our analysis from a four-processor machine similar to cases used in literature (Trancoso et al., 1997; Barroso et al., 1998; Cao et al., 1999; Ranganathan et al., 1998), running the DSS26 workload (Table 2).

In detail, the reference four-processor machine has a 128-bit bus and 64-byte block size. We varied cache size (from 512 K to 2 M byte) and cache associativity (1, 2, 4).

Execution Time (Figure 1) is affected by the cache architecture. It decreases with larger cache sizes and/or more associativity. Anyway, this variation is limited to an 8% between the less (512 K byte, one way) and most performing (2 M byte, four way) configuration. In this case, the role of the coherence protocol is less important, with a difference among the various protocols, for a given setup, which is less than 1%.

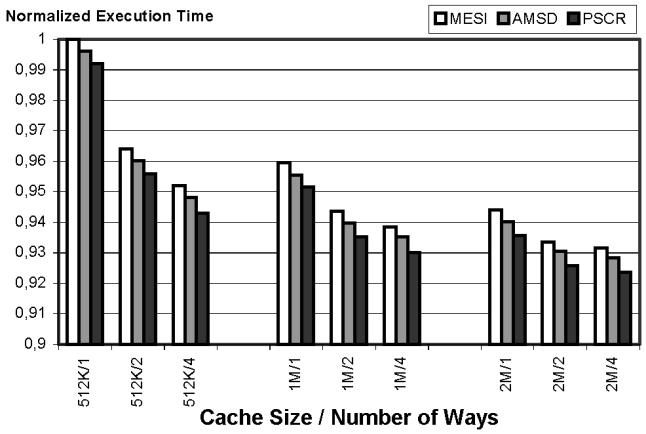

Figure 1 *Normalised execution time vs. cache size (512K, 1M, 2M bytes), number of ways (1, 2, 4) and coherence protocol (AMSD, MESI, PSCR). Data assume four processors, 64-byte block size, and a cache affinity scheduler. Execution Times are normalised with respect to the MESI – 512K, direct access architecture. PSCR presents the lowest Execution Time, whilst MESI the highest*

We analysed the reason for this performance improvement (Figure 2) by decomposing the miss rate in term of traditional (cold, conflict, and capacity) and invalidation misses. In Figure 3, we show the contribution of each kind of sharing to the Invalidation Miss Rate and, in Figure 4, to the Coherence-Transaction 'Rate' (i.e., the number of coherence transactions per 100 memory references). We also differentiated between kernel and user overhead.

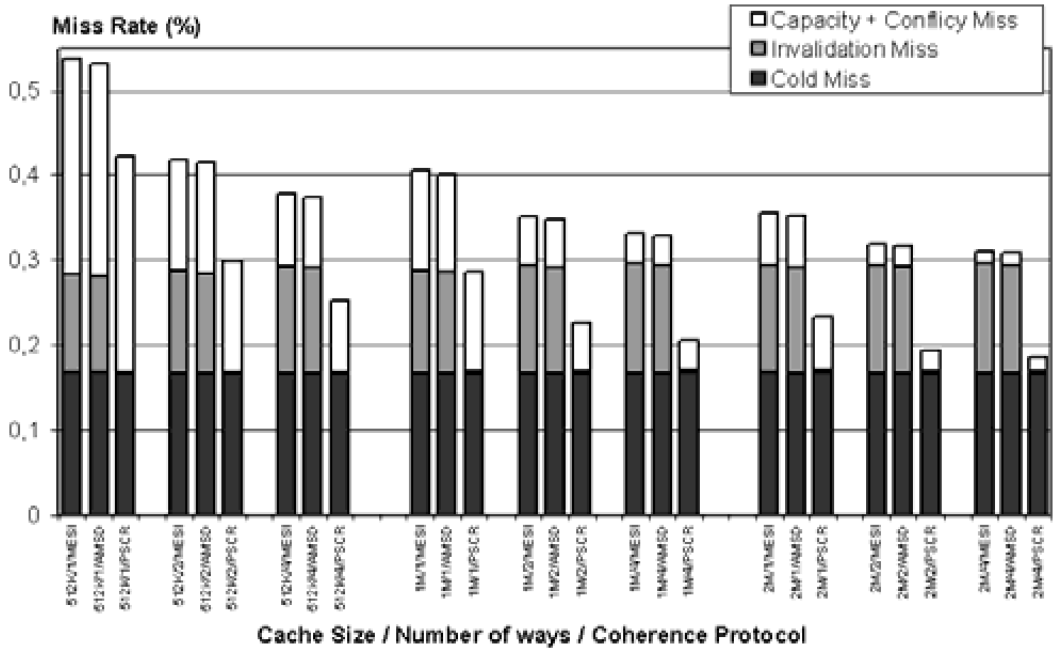

Figure 2 *Breakdown of miss rate vs. cache size (512K, 1M, 2M bytes), number of ways (1, 2, 4) and coherence protocol (AMSD, MESI, PSCR). Data assumes four processors, affinity scheduling and 64-byte block*

As expected, cache size mainly influences capacity misses. Invalidation misses increase slightly when increasing cache size or associativity (Figure 2). For cache sizes larger than 1 *M* bytes, cold misses are the dominating part and invalidation misses weigh more and more (at least 30% of total misses). This indicates that, for our DSS workload, caches larger than 2 M bytes already capture the main working set. In fact, for cache sizes larger than 2 M bytes, Cold Misses remain constant, invalidation misses increase, and only Capacity+Conflict misses may decrease, but their contribution to miss rate is already minimal. The solution based on PSCR presents the lowest miss rate and negligible invalidation misses.

Our analysis of coherence overhead (Figures 3 and 4) confirms that the major sources of overhead are invalidation misses for WI protocols (MESI and AMSD) and write-transactions for PSCR. Passive sharing overhead, either as invalidation misses, or coherence transactions, is minimal: this means that affinity scheduling performs well. In the user part, true sharing is dominant. In the kernel part, false sharing is the major portion of coherence overhead.

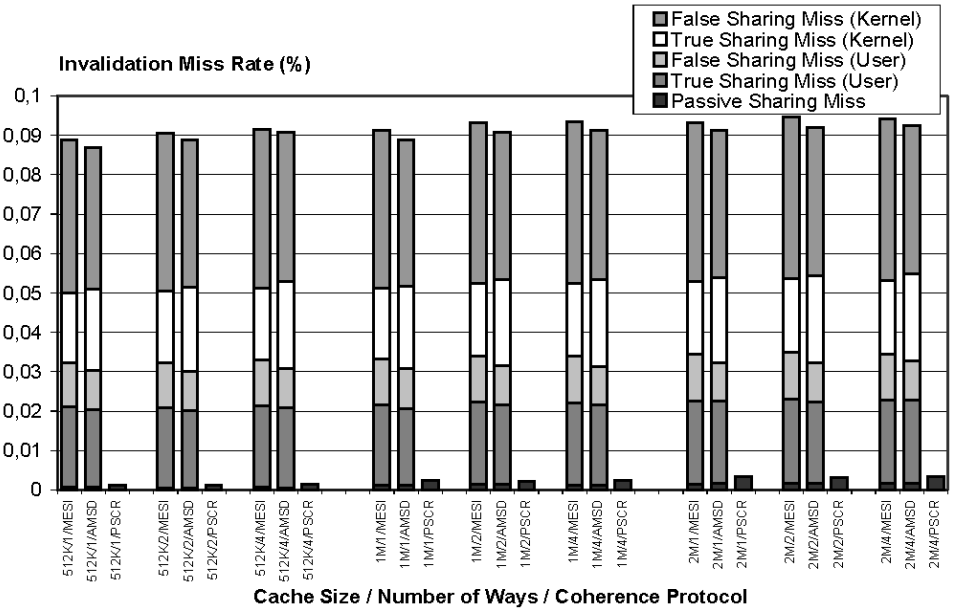

Figure 3 *Breakdown of miss rate vs. cache size (512K, 1M, 2M bytes), number of ways (1, 2, 4), and coherence protocol (AMSD, MESI, PSCR). Data assume four processors, affinity scheduling, and 64-byte block*

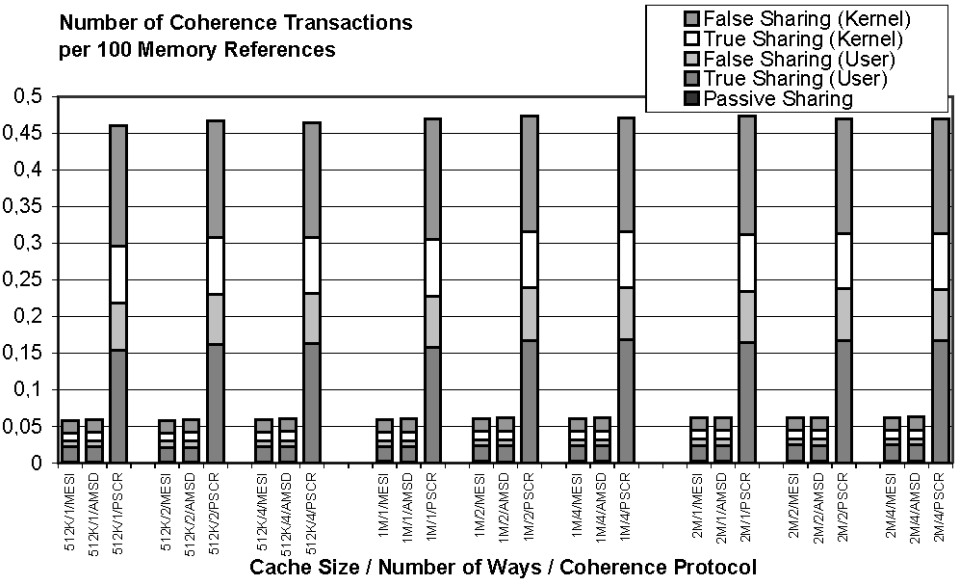

Figure 4 *Number of coherence transactions vs. cache size (512K, 1M, 2M bytes), number of ways (1, 2, 4), and coherence protocol (AMSD, MESI, PSCR). Coherence transactions are invalidate transactions in MESI, and AMSD, write transactions in PSCR. Data assume four processors, 64-byte block size, and an affinity scheduler*

The cost of misses is dominating the performances and indeed we show in Figure 1 that PSCR is able to achieve the best performance compared with the other protocols. The reason is the following: what PSCR loses in terms of extra coherence traffic, is then gained as saved misses. Indeed, misses are more costly in terms of bus timings and read-block transactions may produce a higher waiting time for the processor. Also, AMSD performs better than MESI due to the reduction of Invalidation Miss Rate overhead (Figure 2).

Our conclusions, for the four-processor configuration, agree with previous studies as for the analysis of miss rate and the effects of coherence maintaining (Barroso et al., 1998; Cao et al., 1999; Trancoso et al., 1997).

5.4 Analysis of the high-end system

In the previous baseline case analysis, we have seen that the four-processor machine is efficient: architecture variations do not produce further gains (e.g., the performance differences among protocol are small). This kind of machine may not satisfy the performance needs of DSS workloads, and more performing systems are demanded. Given current processor-memory speeds, we considered a 'high-end' 16-processor configuration. A detailed analysis of the limitations of this system shows that this architecture makes sense, i.e., a high-end SMP system results efficient, if we solve the problems produced by the process migration and adapt the coherence protocol to the access pattern of the workload. This architecture has not been analysed in literature, and still represents a relatively economic solution to enhance the performance.

In the following, we will compare our results directly with the four-processor case and show the sensitivity to classical cache parameters (5.4.1). We will consider separately the case of different block size (5.4.2), the case of a different workload pressure in terms of number of processes (5.4.3) and possible extension of results to other DBMS workloads (5.4.4).

5.4.1 Comparison with four-processor case and sensitivity to cache size and associativity

In Figure 5, we show the Execution Time for several configurations. Protocols designed to reduce the effects of process migration achieve better performance. In particular, the performance gain of PSCR over MESI is at least 13% in all configurations. The influence of cache parameters is stronger with a 30% difference between the most and less performing configuration.

We clarify the relationship between execution Time and process migration, by showing the Miss Rate (Figure 6) and the Number of CoherenceTransactions (Figure 8). In detail, we show the breakdown of the most varying components of Miss Rate (Invalidation Miss, Figure 7) and breakdown of coherence transactions (Figure 8). In the following, for the sake of clearness, we assume a 1M byte-cache size, a 64-byte block size, and two-ways.

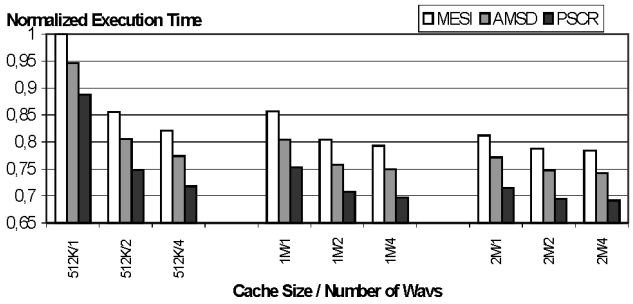

Figure 5 *Normalised execution time vs. cache size (512K, 1M, 2M bytes), number of ways (1, 2, 4), and coherence protocol (AMSD, MESI, PSCR). Data assume 16 processors, 64-byte block size, and an affinity scheduler. Execution Times are normalised with respect to the execution time of the MESI, 512 K, direct access configuration*

The Execution Time is mainly determined by the cost of read-block transactions and the cost of coherence-maintaining transactions (see also Section 5.2). Let us take MESI as baseline. AMSD has a lower Execution Time because of its lower Miss Rate (Figure 6, and in particular Invalidation Miss Rate, Figure 7). The small variation in the Number of Coherence Transactions (Figure 8) does not weigh much on the performance. PSCR gains strongly from the reduction of Miss Rate, while it is not penalised too much by the high Number of Coherence Transactions (Figures 6 and 8). In detail

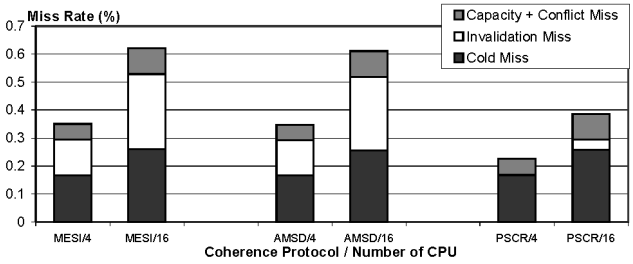

Figure 6 *Breakdown of miss rate vs. coherence protocol (AMSD, MESI, PSCR) and number of processors (4, 16). Data assume an affinity scheduler, 64-byte block, 1M-cache size two-way set associative. The higher number of processor causes more coherence misses (false plus true sharing) and more capacity and conflict misses. It is interesting to observe that while the aggregate cache size increases (from 4M to 16M bytes), Capacity and Conflict misses also increase. This situation is different from the uniprocessor case, where Capacity + Conflict misses always decrease when cache size increases. This effect is due to process migration*

- Classical misses (Cold and Conflict+Capacity) do not vary with the coherence protocol, although they increase while switching from 4 to 16 processors because of the higher migration of the processes: for the Cold Miss portion, because a process may be scheduled on a higher number of processors, and for the Conflict+Capacity, because a process has less opportunities of reusing its working set and it destroys the cache footprint generated by other processes.

- The much lower Miss Rate in PSCR is due to its low Invalidation Miss Rate (Figure 6). In particular, PSCR does not have invalidation misses due to true and false sharing, neither in the user nor in the kernel mode (Figure 7). On the contrary, in the other two protocols the latter factors weigh very much. This effect is more evident in the 16-processor case as compared to the four-processor case because of the higher probability of sharing data, produced by the increased number of processors.
- The DSS workload, in the PostgreSQL implementation, exhibits an access pattern to shared data, which speed-up the WU strategy for maintain coherence among shared data with respect to the WI strategy. We can explain such pattern with the following considerations: the TPC-D DSS queries are read-only queries and consequently, as already observed in Trancoso et al. (1997) and Yu et al. (2002) coherence misses and transactions involve PostgreSQL metadata (i.e., data structures that are needed by the DBMS to work rather than to store the data) and particularly data structure needed to implement software locks (Trancoso et al., 1997; Yu et al., 2002). The type of access pattern to those data is one producer/many consumers. Such pattern advantages WU protocols, especially if there are a high number of processes (the 'consumers') that concurrently read shared data. When the number of processors switches from 4 to 16, the number of consumer processes increases, due to the higher number of processes concurrently in execution and, consequently, true sharing Invalidation Misses and Transactions (in the user part) increase especially for WI protocols (Figures 7 and 8), thus speeding-up the performance of the WU solution (PSCR). However, we cannot utilise a pure WU protocol, as passive sharing dramatically decreases performances (Prete et al., 1997).

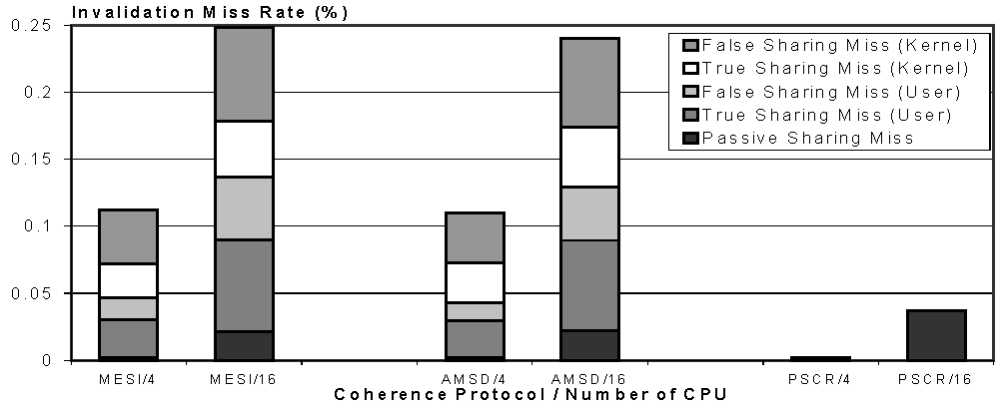

Figure 7 *Breakdown of invalidation miss rate vs. coherence protocol (AMSD, MESI, PSCR) and number of processors (4, 16). Data assume, an affinity scheduler, 64-byte block, 1M-byte 2-way set associative caches. Having more processors causes more coherence misses (false and true sharing). Passive sharing misses are slightly higher in PSCR compared to the other two protocols. This is a consequence of the selective invalidation mechanism of PSCR. In fact, as soon as a private block is fetched on another processor, PSCR invalidates all the remote copies of that block. In the other two protocols, the invalidation is performed on a write operation on shared data, thus less frequently than in PSCR*

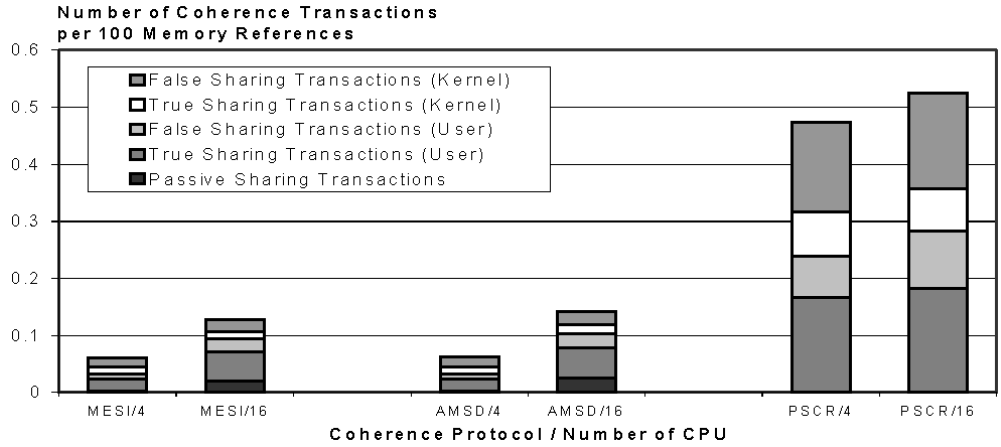

Figure 8 *Number of coherence transactions vs. coherence protocol (AMSD, MESI, PSCR) and number of processors (4, 16). Data assume, an affinity scheduler, 64-byte block, 1M-byte 2-way set associative caches. There is an increment in the sharing overhead in all of its components. This increment is more evident in the WI class protocols, also because there is more passive sharing overhead*

5.4.2 Analysing the effects of a larger block size in the 'high-end' system

We started our analysis from the 64-byte block size for references comparison with other studies. When switching from 64 to 128 bytes, PSCR has further advantages in respect of the other two considered protocols (Figure 9). This is due to the following two reasons. First, we observe a reduction of Capacity+Conflict miss component (Figure 10), a small reduction of coherence traffic (Figure 12), and Invalidation Miss Rate (Figure 11). Secondly, in the case of 64-byte block, the system is in saturation[1] (Giorgi and Prete, 1999) for all configurations of Figure 9. In the case of 128-byte blocks, an architecture based on PSCR is not saturated, and thus we can use configurations with a higher number of processors efficiently. When switching from 64 to 128 bytes, the decrease of the Execution Time is 27% for PSCR, and only a 20% for the other protocols. We observe that a block size larger than 128 bytes produces diminishing returns, because the increased cost of read-block transaction is not compensated by the reduction of the number of misses. Similar result has been obtained for a CC-NUMA machine running a DSS workload (Lovett and Clapp, 1996). In that work, the number of 'Effective Processors' for a 16-processor CC-NUMA system was almost the same as that obtained for our cheaper shared-bus shared-memory system (figure not showed).

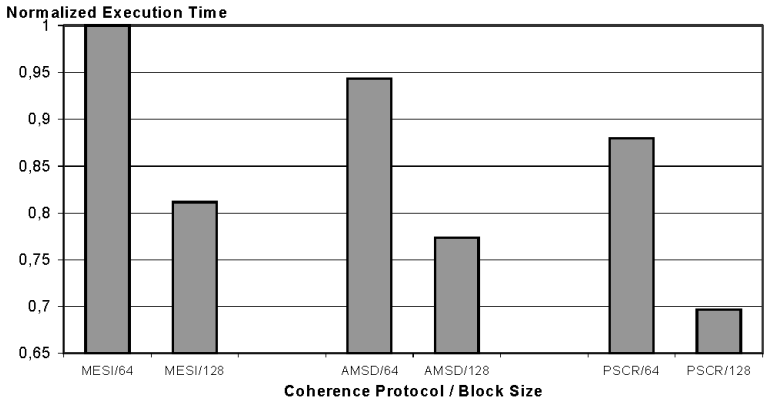

Figure 9 *Normalised execution time vs. block size (64, 128 bytes) and coherence protocol (AMSD, MESI, PSCR). Data assume, an affinity scheduler, 1M-byte 2-way set associative caches. Execution Times are normalised with respect to the execution time of the MESI, 64 byte block configuration*

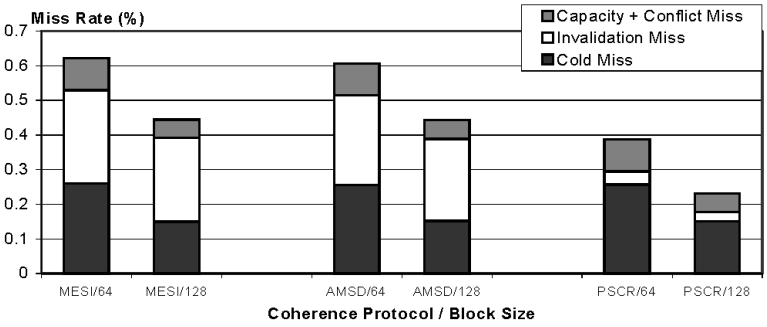

Figure 10 *Breakdown of miss rate vs. coherence protocol (AMSD, MESI, PSCR) and block size (64 byte, 128 byte). Data assume an affinity scheduler, 1M-byte 2-way set associative caches. There is a decrease of Cold, Capacity+Conflict miss components, and a little decrease of invalidation miss component*

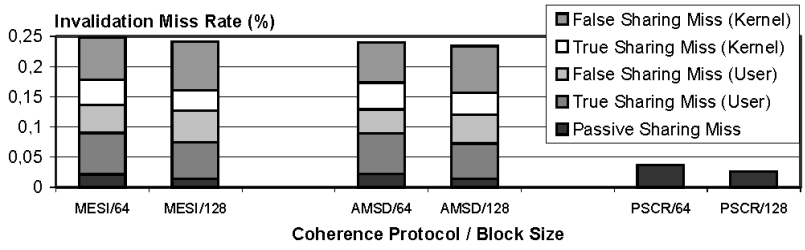

Figure 11 *Breakdown of invalidation miss rate vs. coherence protocol (AMSD, MESI, PSCR) and block size (64 byte, 128 byte). Data assume an affinity scheduler, 1M-byte 2-way set associative caches. Passive Sharing Misses decrease when increasing block size because the invalidation unit is larger*

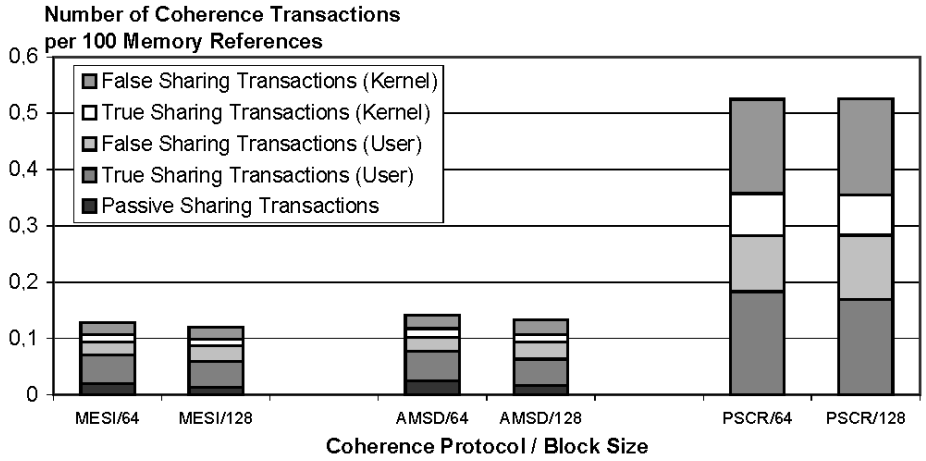

Figure 12 *Number of coherence transactions vs. coherence protocol (AMSD, MESI, PSCR) and block size (64, 128 byte). Data assume an affinity scheduler, 1M-byte 2-way set associative caches*

5.4.3 Analysing the effects of variations in the number of processes of the workload

We considered another scenario where the number of processes in the workload may vary and thus the scheduler could fail in applying affinity. The affinity scheduling could fail when the number of ready processes is limited. We defined a new workload (DSS18, Table 2) having characteristics similar to DSS26 workload that was used in the previous experiments, but constituted of only 18 processes. The machine under study is still the 16-processor one. In such a condition, the scheduler can only choose between at most two ready processes. The measured miss rate and number of coherence transactions (Figures 13, 14, 15) shows an interesting behaviour. The miss rate, and in particular Cold, Conflict+Capacity miss rate, increases with respect to the DSS26 workload. This is consequence of process migration, and it is determined by the failure of the affinity requirement: as the number of processes is almost equal to the number of processors, it is not always possible for the system to reschedule a process on the processor where it last executed. In such cases, PSCR can reduce greatly the associated overhead and it achieves the best performance (figure not shown).

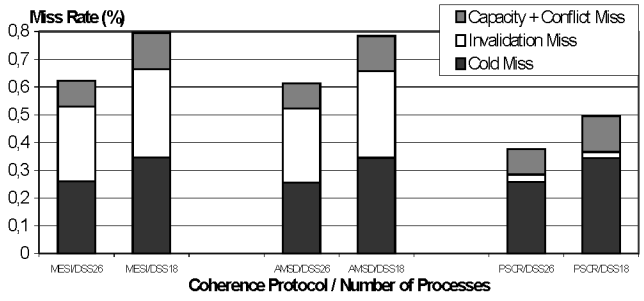

Figure 13 *Breakdown of miss rate vs. coherence protocol (AMSD, MESI, PSCR) and workload (DSS26–DSS18). Data assume an affinity scheduler, 64-byte block, 1M-byte 2-way set associative caches. The workload DSS18 exhibits the higher miss rate, due to an increased number of Cold, Capacity, and Conflict Miss. This is a consequence of process migration, which affinity fails to mitigate*

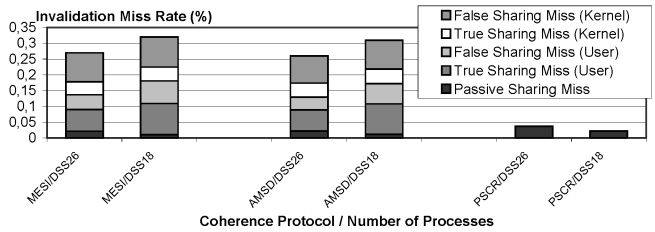

Figure 14 *Breakdown of invalidation miss rate vs. coherence protocol (AMSD, MESI, PSCR) and workload (DSS26 – DSS18). Data assume an affinity scheduler, 64-byte block, 1M-byte 2-way set associative caches. Passive sharing misses decrease, while true and false sharing misses increase. This is consequence of the failure of affinity scheduling: in the DSS18 execution. Processes migrate more than in DSS26 execution, thus generating more reuse of shared data (and more invalidation misses on shared data) but lower reuse of private data (and lower number of passive sharing misses)*

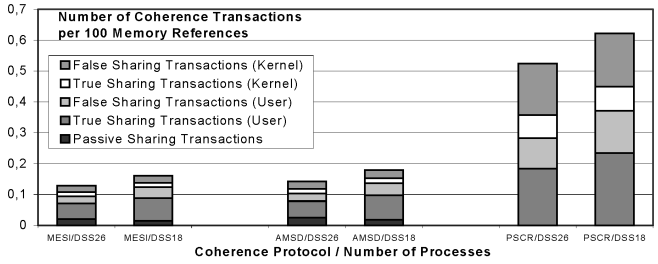

Figure 15 *Breakdown of Coherence Transactions vs. coherence protocol (AMSD, MESI, PSCR) and workload (DSS26 – DSS18). Data assume affinity, 64-byte block, 1M-byte 2-way set associative caches*

The main conclusion here is that PSCR maintains its advantage also in different load conditions, while the other protocols are more penalised by critical scheduling conditions.

5.4.4 Extending our results to OLTP systems

As also observed in other works (Barroso et al., 1998; Lovett and Clapp, 1996; Transaction Processing Performance Council, 1994; Keeton et al., 2003), the main

differences among OLTP and DSS applications are the following:

- execution time of OLTP queries is shorter
- the number of concurrent queries running against the DB is larger
- the reuse of data cached in DBMS shared memory is smaller
- queries are updating queries.

As a consequence, we can aspect that:

- passive sharing effects are less important in OLTP systems than in DSS ones, due to the shortness of queries, and then to the lower probability of reuse of private data
- affinity scheduling is much more effective in OLTP, as there is a higher number of running processes
- OLTP systems present much more sharing in the user part of the workload, due to the higher number of updating queries.

All these facts suggest that, in the case of OLTP workloads, the differences among the protocols become smaller than those in DSS workloads, and the access pattern to user shares data (i.e., the number of queries updating the same row in the database) is crucial to decide which is the best performing protocol.

6 CONCLUSIONS

We evaluated the memory performance of a shared-bus shared-memory multiprocessor running a DSS workload, by considering several different choices that could improve the overall performance of the system. We considered different architectures based on the following coherence protocols: MESI – a pure WI protocol, widely used in high-performance multiprocessors, AMSD – a WI protocol designed to reduce effects of data migrations – and PSCR – a coherence protocol using a hybrid strategy, that is WU for shared data and WI for private data, designed to reduce the effects of process migration. The DSS workload was setup using the PostgreSQL DBMS executing queries of the TPC-D benchmark and typical Unix shell commands and daemons. We considered kernel effects that are more relevant to our analysis like process scheduling, virtual memory mapping, user/kernel code interactions.

Our conclusions, for the four-processor case, agree with previous studies as for the analysis of miss rate and the effects of coherence maintaining. Our analysis outlines also:

- cache sizes larger than 2 *M* bytes already capture the working set of such workload
- the kernel effects account for 50% of the coherence overhead.

Previous studies that considered DSS workloads were mostly limited to four-processor systems, did not consider the effects of process migration, and did not correlate the amount of sharing to the performance of the system.

Our analysis of a 'high-end' machine considered a 16-processor SMP. We analysed variations of classical cache parameters and variations in the workload pressure on the scheduler due to a different number of processes. We found that in the high-end systems some factors, which were less noticed in the four-processor case, become more evident.

MESI protocol is not the best choice in high-end SMP architectures: AMSD improves the performance of a DSS system of about 10% compared to MESI; PSCR improves the performance of about 20% compared to MESI. DSS workloads running on SMP architectures generate a variable load. The affinity scheduler may fail to deliver the affinity requirements. The use of PSCR allows us to build systems, whose performance is less influenced by the load condition. Finally, in OLTP systems, we expect that the effect of migration are less evident, due to the short life of executing processes.

REFERENCES

Agarwal, A. (1989) *Analysis of Cache Performance for Operating Systems and Multiprogramming*, Kluwer Academic Publishers, Norwell, MA.

Ballinger, C. (2001) *Relevance of the TPC-D Benchmark Queries: The Questions You Ask Every Day*, NCR parallel System, http://www.tpc.org/information/other/articles/TPCDart_0197.asp.

Barroso, L.A., Gharachorloo, K. and Bugnion, E. (1998) 'Memory system characterisation of commercial workloads', *Proc. of 25th Intl. Symp. on Computer Architecture*, Barcelona, Spain, June, pp.3–14.

Cao, Q., Torrellas, J., Trancoso, P., Pey, L.J., Knighten, B. and Won, Y. (1999) 'Detailed characterisation of a quad Pentium Pro server running TPC-D', *Proc. of Intl. Conf. on Computer Design*, October, pp.108–115.

Chandra, R., Devine, S., Verghese, B., Gupta, A. and Rosenblum, M. (1994) 'Scheduling and page migration for multiprocessor compute servers', *Proc. of 6th ASPLOS*, October, pp.12–24.

Chapin, J., Herrod, S., Rosenblum, M. and Gupta, A. (1995) 'Memory system performance of UNIX on CC-NUMA multiprocessors', *ACM Sigmetrics Conf. on Measurement and Modeling of Computer Systems*, May, pp.1–13.

Cox, A.L. and Fowler, R.J. (1993) 'Adaptive cache coherency for detecting migratory shared data', *Proc. 20th Intl. Symp. on Computer Architecture*, San Diego, California, May, pp.98–108.

Culler, D. and Singh, J. (1998) *Parallel Computer Architecture: A Hardware/Software Approach*, Morgan Kaufmann, San Francisco, CA.

Cvetanovic, Z. and Bhandarkar, D. (1994) 'Characterisation of alpha AXP performance using TP and SPEC workloads', *Proc. 21st Intl. Symp. on Computer Architecture*, April, pp.60–70.

Dubois, M., Skeppstedt, J., Ricciulli, L., Ramamurthy, K. and Stenström, P. (1993) 'The detection and elimination of useless miss in multiprocessor', *Proc. of 20th Intl. Symp. on Computer Architecture*, San Diego, CA, May, pp.88–97.

Eggers, S.J. (1991) 'Simplicity versus accuracy in a model of cache coherency overhead', *IEEE Transactions on Computers*, August, Vol. 40, No. 8, pp.893–906.

Eggers, S.J. and Jeremiassen, T.E. (1991) 'Eliminating false sharing', *Proc. 1991 Intl. Conf. on Parallel Processing*, August, pp.I:377–381.

Eggers, S.J. and Katz, R.H. (1988) 'A characterisation of sharing in parallel programs and its application to coherency protocol evaluation', *Proc. of 15th Annual Int. Symp. on Computer Architecture*, Honolulu, Hl, May, pp.373–382.

Eggers, S.J. and Katz, R.H. (1989a) 'Evaluating the performance of four snooping cache coherency protocols', *Proc. of 16th Annual International Symp. on Computer Architecture*, Jerusalem, Israel, pp.2–15.

Eggers, S.J. and Katz, R.H. (1989b) 'The effect of sharing on the cache and bus performance of parallel programs', *Proc. 3rd ASPLOS,* Boston, MA, April, pp.257–270.

Eggers, S.J., Emer, J.S., Levy, H.M., Lo, J.L., Stamm, R.L. and Tullsen, D.M. (1997) 'Simultaneous multithreading: a platform for next-generation processors', *IEEE Micro*, October, Vol. 17, No. 5, pp.12–19.

Foglia, P. (2001) 'An algorithm for the classification of coherence related overhead in shared-bus shared-memory multiprocessors', *IEEE TCCA Newsletter*, January, pp.40–46.

Foglia, P., Giorgi, R. and Prete, C.A. (1998) 'Analysis of sharing overhead in shared memory multiprocessors', *31st IEEE Hawaii Int. Conf. on Systems*, Kohala Coast, HL, January, Vol. 7, pp.776–778.

Gharachorloo, K., Gupta, A. and Hennessy, J. (1991) 'Performance evaluation of memory consistency models for shared-memory multiprocessors', *Proc. Fourth ASPLOS*, Santa Clara, California, April, pp.245–357.

Giorgi, R. and Prete, C.A. (1999) 'PSCR: a coherence protocol for eliminating passive sharing in shared-bus shared-memory multiprocessors', *IEEE Trans. on Parallel and Distributed Systems*, Vol. 10, No. 7, pp.742–763.

Giorgi, R., Prete, C., Prina, G. and Ricciardi, L. (1997) 'Trace factory: a workload generation environment for trace-driven simulation of shared-bus multiprocessor', *IEEE Concurrency*, Vol. 5, No. 4, pp.54–68.

Hennessy, J. and Patterson, D.A. (2002) *Computer Architecture: a Quantitative Approach,* Morgan Kaufmann Publishers, San Francisco, CA.

Hyde, R.L. and Fleisch, B.D. (1996) 'An analysis of degenerate sharing and false coherence', *Journal of Parallel and Distributed Computing*, Vol. 34, No. 2, May, pp.183–195.

Jeremiassen, T. and Eggers, S. (1995) 'Reducing false sharing on shared memory multiprocessors through compile time data transformations', *ACM SIGPLAN Notices*, Vol. 30, No. 8, August, pp.179–188.

Keeton, K., Clapp, R. and Nanda, A. (2003) 'Evaluating servers with commercial workloads', *IEEE Computer*, February, Vol. 36, No. 2, pp.29–32.

Keeton, K., Patterson, D., He, Y., Raphael, R. and Baker, W. (1998) 'Performance characterisation of a quad pentium pro SMP using OLTP workloads', *Proc. 25th Intl. Symp. on Computer Architecture*, June, pp.15–26.

Kroft, D. (1981) 'Lockup-free instruction fetch/prefetch cache organisation', *Proc. 8th Intl. Symp. on Computer Architecture*, June, pp.81–87.

Lo, L., Barroso, A., Eggers, S.J., Gharachorloo, K.G., Levy, H.M. and Pareck, S. (1998) 'An analysis of database workload performance on simultaneous multithreaded processors', *Proc. 25th Annual Intl. Symp. on Computer Architecture*, Barcelona, Spain, June, pp.39–50.

Lovett, T. and Clapp, R. (1996) 'StiNG: a CC-NUMA computer system for the commercial marketplace', *Proc. of the 23rd Intl. Symp. on Computer Architecture*, May, pp.308–317.

Maynard, G.A.M., Donnelly, C.M. and Olszewski, B.R. (1994) 'Contrasting characteristics and cache performance of technical and multi-user commercial workloads', *Proc. of the 6th ASPLOS*, October, pp.158–170.

Prete, C.A., Prina, G. and Ricciardi, L. (1995) 'A trace driven simulator for performance evaluation of cache-based multiprocessor system', *IEEE Trans. on Parallel and Distributed Systems,* Vol. 6, No. 9, pp.915–929.

Prete, C.A., Prina, G., Giorgi, R. and Ricciardi, L. (1997) 'Some considerations about passive sharing in shared-memory multiprocessors', *IEEE TCCA Newsletter*, March, pp.34–40.

Ranganathan, P., Gharachorloo, K., Adve, S.V. and Barroso, L. (1998) 'Performance of database workloads on shared-memory systems with out-of-order processors', *Proc. of the 8th ASPLOS*, San Jose, CA, pp.307–318.

Shanley, T. and Mindshare *Inc.* (1999) *Pentium Pro and Pentium II System Architecture*, Addison Wesley, Reading, MA.

Stenström, P., Brorsson, M. and Sandberg, L. (1993) 'An adaptive cache coherence protocol optimised for migratory sharing', *Proc. of the 20th Annual Intl. Symp. on Computer Architecture*, May, pp.109–118.

Stenström, P., Hagersten, E., Li, D.J., Martonosi, M. and Venugopal, M. (1997) 'Trends in shared memory multiprocessing', *IEEE Computer*, December, Vol. 30, No. 12, pp.44–50.

Stunkel, B., Janssens, B. and Fuchs, K. (1991) 'Address tracing for parallel machines', *IEEE Computer*, Vol. 24, No. 1, January, pp.31–45.

Sweazey, P. and Smith, A.J. (1986) 'A class of compatible cache consistency protocols and their support by the IEEE futurebus', *Proc. of the 13th Intl. Symp. on Computer Architecture*, June, pp.414–423.

Tanenbaum, A.S. (2001) *Structured Computer Organisation*, 4th Edition, Prentice-Hall, Inc.

Tomasevic, M. and Milutinovic, V. (1993) *The Cache Coherence Problem in Shared-Memory Multiprocessors – Hardware Solutions,* IEEE Computer Society Press, Los Alamitos, CA, April.

Tomasevic, M. and Milutinovic, V. (1994a) 'Hardware approaches to cache coherence in shared-memory multiprocessors', *IEEE Micro*, October, Vol. 14, No. 5, pp.52–59.

Tomasevic, M. and Milutinovic, V. (1994b) 'Hardware approaches to cache coherence in shared-memory multiprocessors', *IEEE Micro*, December, Vol. 14, No. 6, pp.61–66.

Tomasevic, M. and Milutinovic, V. (1996) 'The word-invalidate cache coherence protocol', *Microprocessors and Microsystems*, Vol. 20, pp.3–16.

Torrellas, J., Gupta, A. and Hennessy, J. (1992) 'Characterising the caching and synchronisation performance of a multiprocessor operating system', *Proc. 5th ASPLOS*, September, pp.162–174.

Torrellas, J., Lam, M. and Hennessy, J.L. (1990) 'Share data placement optimisations to reduce multiprocessor cache miss rates', *Proc. Intl. Conf. on Parallel Processing*, Urbana, IL, August, pp.266–270.

Torrellas, J., Lam, M.S. and Hennessy, J.L. (1994) 'False sharing and spatial locality in multiprocessor caches', *IEEE Transactions on Computer*, June, Vol. 43, No. 6, pp.651–663.

Trancoso, P., Pey, L.J.L., Zhang, Z. and Torrellas, J. (1997) 'The memory performance of DSS commercial workloads in shared-memory multiprocessors', *Proc. of the 3rd Intl. Symp. on High Performance Computer Architecture*, Los Alamitos, CA, February, pp.250–260.

Transaction Processing Performance Council (1994) *TPC Benchmark B (Online Transaction Processing) Standard Specification,* June, http://www.tpc.org.

Transaction Processing Performance Council (1995) *TPC Benchmark D (Decision Support) Standard Specification,* December, http://www.tpc.org.

Uhlig, R. and Mudge, T. (1997) 'Trace-driven memory simulation: a survey', *ACM Computing Surveys*, June, pp.128–170.

Yeager, K.C. (1996) 'The MIPS R10000 superscalar microprocessor', *IEEE Micro*, August, Vol. 16, No. 4, pp.42–50.
Yu, A. and Chen, J. (1995) *The POSTGRES95 User Manual,* Computer Science Div., Dept. of EECS, UCB, July.
Yu, R., Bhuyan, L. and Iyer, R. (2002) 'Comparing the memory system performance of DSS workloads on the HP v-class and SGI origin 2000', *Proc. of the Int. Parallel and Distributed Processing Symposium*, Fort Lauderdale, FL, April, pp.31–37.

NOTE

[1]We recall that a shared-bus shared-memory SMP system is in saturation (Giorgi and Prete, 1999) when the performance does not increase at least of a given quantity, when we add one processor to the machine.