dMT: inexpensive throughput enhancement in small-scale embedded microprocessors with differential multithreading

K.R. Hirst, J.W. Haskins and K. Skadron

Abstract: The authors examine differential multithreading (dMT) as an attractive organisation for increasing throughput in simple, small-scale, pipelined processors like those used in embedded environments. dMT copes with pipeline stalls due to hazards and data- and instruction-cache misses by using duplicated pipeline registers to run instructions from an alternate thread. Results show that dMT boosts throughput substantially and can in fact replace dynamic branch prediction or can be used to reduce the sizes of the instruction and data caches.

1 Introduction

Previous research has demonstrated the effectiveness of multiple hardware contexts for improving throughput, hiding memory latency and supporting thread-level and instruction-level parallelism in CPU-intensive computations. Differential multithreading (dMT) is a low-cost version of hardware multithreading in which multiple instruction streams share a single pipeline, and the processor squashes pipeline stalls from one thread by executing instructions from another thread. (When we refer to a 'thread-switch', we mean a pipeline switch between its active instruction streams and not an OS-level switch among kernel threads.) These threads might be separate, independent processes or cooperating threads within a single process. A simpler version of this technique, block multithreading or BMT, was first described by Farrens and Pleszkun in [1]. dMT extends BMT by adding the ability to switch active threads in response to misses in the instruction- and data-caches.

Conventional pipelines fall short of maximum throughput because stalls in the pipeline prevent the retirement of an instruction in every cycle. The objective of dMT, like BMT, is to asymptotically approach the maximum throughput of one instruction per cycle (1 IPC) by switching among multiple instruction streams in response to stall conditions. This is in contrast to other techniques (like larger, more highly associative caches or data forwarding (bypassing data from later instructions to earlier ones to avoid data hazards) which increase pipeline utilisation for only a single instruction stream.

Although single-issue ('scalar') organisations are no longer used in high-performance processors, they remain common even in new processor designs for small-scale, embedded devices. Some embedded processors in fact omit the data-cache altogether, a configuration for which dMT is especially valuable. Overall, multithreading benefits embedded workloads for which throughput is as important as the single-thread execution time, and in particular, workloads that would run on the simpler, single-issue base architecture we consider and are prevalent in embedded environments. Specific examples include embedded applications such as video game units, portable personal organisers, and process-control systems. BMT and dMT are especially useful for multithreaded workloads. Video game systems are an example; they must simultaneously support artificial intelligence manipulation of multiple computer-controlled characters, background music playback, player input processing, and data (pre)fetch from the game ROM. Unfortunately, finding a suitable, non-proprietary multi-threaded workload has proven difficult, and our evaluations focus on the throughput of multiple, semi-independent applications.

Not only does dMT increase throughput, it does this so effectively that other processor structures, such as caches or branch predictors, can be made smaller or even eliminated. For example, we will show in Section 5 that dMT allows the use of a smaller cache or the complete elimination of branch-prediction hardware. In [2], we also found that dMT is highly effective in chip-multiprocessor configurations, more so than non-multithreaded dual-issue processor cores.

Our contribution stems from our extensions to BMT which can be thought of as combining aspects of the BMT [1] and Runahead [3] pipelines. BMT uses dual-decoder logic to perform instruction interleaving at the end of decode, where the issue logic selects one instruction to promote to the next stage of the pipeline. Placing the interleaving mechanism solely in this early stage of the pipeline allows BMT to respond to data hazards and long-latency branch delays; it cannot, however, avoid stalls that result from misses in the instruction or data caches. Our

IEE Proceedings online no. 20040185

doi: 10.1049/ip-cdt:20040185

Paper received 29th May 2002 and in revised form 26th September 2003

K.R. Hirst was with Department of Computer Science, University of Virginia, 151 Engineer's Way, Box 400740 Charlottesville, VA 22904-4740, USA and is now with Kidwell McGowan Associates, Sterling, VA

J.W. Haskins was with Department of Computer Science, University of Virginia, 151 Engineer's Way, Box 400740 Charlottesville, VA 22904-4740, USA and is now with the Institute for Defense Analyses Center for Computing Sciences in Bowie, MD

K. Skadron is with Department of Computer Science, University of Virginia, 151 Engineer's Way, Box 400740 Charlottesville, VA 22904-4740, USA

implementation, dMT, gains this ability by capturing and storing in-flight instructions when an instruction stream encounters a stall for all sources of pipeline stalls.

We also re-examine BMT in light of more modern benchmarks. The original BMT work used only the Livermore loops. Our studies use the MiBench [4], MediaBench [5], and SPECint95 [6] suites.

2 Related work

Our work is most closely related to BMT [1]. It describes three different policies for interleaving instructions: every-cycle, blocked and prioritised. Every-cycle switches threads after every clock cycle, and blocked only in response to stalls. For prioritised, the thread with priority resumes execution as soon as its stall condition resolves, regardless of the status of the other thread; this policy is useful for real-time workloads. By resuming its execution as soon as its stall condition resolves, the thread with priority will incur no penalty beyond what it would if it were executed alone, thus preserving the predictability of its execution time. We find that the blocked and prioritised policies give a fairly similar performance for both BMT and dMT, and that every-cycle is consistently the worst. We will only present data on results with the blocked policy; results for prioritised and every-cycle can be found in [7].

Traditional multithreaded architectures such as the Tera [8] achieve performance gains by every-cycle scheduling. Every-cycle scheduling among the Tera's 128 hardware contexts allows the Tera to hide latencies experienced by individual threads, and indeed, the time required to service all 128 contexts masks memory latency and permits the Tera to be completely cacheless. Simultaneous multithreaded (SMT) architectures [9, 10] take a different approach, extending wide-issue superscalar architectures by allowing multiple hardware contexts to issue instructions to the execution units: in any given cycle, a mix of instructions from several different threads might issue. Neither the Tera nor SMT is readily deployable into an embedded system that needs high performance and yet is still constrained by cost, size and power. In particular, it is not obvious how to cost-effectively extrapolate the Tera and SMT approaches to a scalar, single-issue pipeline.

Two more systems that are related to BMT are APRIL [11] and Runahead [3]. Like dMT, APRIL thread-switches on a cache miss; however, APRIL uses a more heavyweight thread switch in which the pipeline must drain. The consequent ten-cycle delay is not suitable for hiding pipeline stalls. Runahead microprocessors speculatively execute instructions past a first-level D-cache miss. These instructions are not committed; their purpose is to uncover subsequent memory instructions whose target address is calculable. Even although the result of these references is discarded, they serve as lightweight prefetches. Unlike dMT, Runahead does not attempt to fill in the stall cycles with instructions from another instruction stream.

Low-cost microprocessors are typically in-order issue and often only single-issue, yet this simplicity does not preclude their suffering from pipeline stalls. In fact, some processors contain no cache, in which case each memory reference introduces a stall. Examples include the Motorola DragonBall [12] used in Palm Pilots and the Zilog Z80 [13] used in Nintendo Gameboys.

3 dMT Design overview

This Section provides a more detailed description of the dMT organisation The baseline architecture for our design uses the classic five-stage, single-issue RISC organisation of the ARM10 [14].

For this discussion, we assume a dMT processor that allows two threads to be present on the CPU at any given time. The identity of the current active thread is held in the thread register. Simultaneously hosting two or more instruction streams requires that the program counter and register file be replicated, one per hardware context. Most importantly, (as discussed in Section 2), dMT duplicates the pipeline registers between the fetch-decode (IF-ID), decode-execute (ID-EX) and execute-memory (EX-MM) stages. These duplicated pipeline registers are used to capture in-flight instructions. If an instruction entering the WB stage will not stall, as we currently assume, then the MM-WB pipeline register does not need to be duplicated (recall that in any particular stage, only one instruction is active per cycle). The dMT pipeline organisation is depicted in Fig. 1. Not shown is the fact that the instruction in the WB stage needs a separate tag to indicate which thread it belongs to and hence which register file to write to. This is necessary because the instruction in the WB stage may be from a different thread than previous stages.

Each data and control signal entering the duplicated pipeline registers has a fanout of two: one copy into each 'half' of the pipeline register. This would probably be implemented in a bit-sliced fashion, with the two sources of each output signal co-located to minimise wiring length. This can be thought of as taking the two logical copies of the pipeline register and interleaving them. Each logical 'half' of the duplicated registers has a write-enable whose setting is determined by the active thread identified in the thread register. Only the half owned by the active thread is write-enabled; the other half holds the state of the stalled thread. Each signal entering the next stage requires a multiplexor to choose the correct pipeline input from one or the other half of the pipeline register. The thread register is also used to control these multiplexors. Processors with longer pipelines will require additional duplicated pipeline registers. This increases the total cost of dMT, since more pipeline registers must be duplicated along with the associated multiplexors. However, as pipelines grow longer, branch costs also grow, necessitating branch prediction unless dMT is used to hide these costs.

In dMT, a thread that encounters a stall condition can be thought of as having two parts: (i) a committable part; and (ii) a dependent part. Instructions in the committable part are unaffected by the stall and continue to flow through the pipeline unhindered; instructions in the dependent part cannot proceed until the stall is resolved, and are captured and held in their respective half of the pipeline registers. These frozen instructions continue execution once the offending instruction (the producer) completes and the dependence is resolved. We assume that a stall condition can be detected sufficiently early so the new thread can take

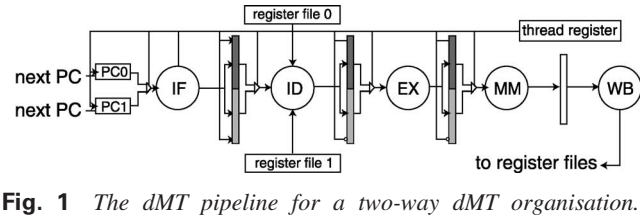

Fig. 1 *The dMT pipeline for a two-way dMT organisation. (Reproduced from [2], ©2001, IEEE by permission of IEEE). In reality, to minimise wiring cost, the pipeline register would not be partitioned as shown, but rather bit-sliced: each signal would have two possible sources that are stored adjacent in the pipeline register*

over execution immediately after the stall condition is detected, with no wasted cycles. Although stall conditions are probably detected late in each stage, the only additional time required is the propagation delay to distribute the choice of thread for the next cycle in time to set up the multiplexors at the beginning of the next stage. If this requires an additional cycle, and introduces one stall cycle each time a thread is switched, the benefits we report decline significantly. Note that no additional logic is required to detect stall conditions; the same logic that detects hazards in conventional processors is leveraged by dMT. The only exception is the possibility of crossthread accesses to the same location in the memory, which requires extra logic to possibly stall the second thread, see below.

A thread switch affects only those stages in the dependent part; all these stages switch in unison immediately. A thread switch does not affect those stages in the committable part. For example, on a data-cache miss, all instructions after the cache are potentially data dependent. The dependent part therefore consists of instructions in the IF, ID and EX stages; the committable part consists of the pending memory operation and the instruction in the WB stage. When the data cache detects a miss, dependent instructions are frozen in their respective half of the pipeline registers. As another example, a branch creates a control hazard that prevents further fetching from that thread. An I-cache miss similarly prevents further fetching. In these cases, the committable part consists of all instructions currently in the pipeline (including the branch in the case of the control hazard). The dependent part simply consists of the PC of the stalled thread; the PC can be thought of as another pipeline register preceding the fetch stage. If no thread's stall condition is resolved, the pipeline experiences a true stall until one thread or the other can proceed. A detailed itemisation of possible stalls and their treatment can be found in [2].

Since some exceptions and many external interrupts require all active threads to be suspended, it is probably easiest to always suspend all threads on any kind of exception or interrupt. This means that both register files and PCs must be saved, which will modestly increase the cost of handling an exception. Since exceptions should be rare, the extra tens of cycles per exception should be negligible.

The input multiplexors to the ID, EX and MM pipeline stages and the extra register file will inevitably place pressure on the processor cycle time. The multiplexor can actually be implemented as a wired-OR of the tri-state-enabled possible outputs, so one extra gate delay is a reasonable expectation for the associated overhead. A useful rule of thumb is that the cycle time might be lengthened by as much as 10% for every extra gate delay [15]. The detailed design of these multiplexors and their associated overhead is subject to a variety of design choices. A detailed implementation is beyond the scope of this discussion, so we focus on cycle-level simulations. Some of the increased throughput from dMT's ability to recapture stall cycles will be offset by the increase in cycle time. For example, if we indeed assume that the dMT's clock runs 10% slower, then throughput improvements of about 10% or better will likely exhibit actual speedups. In other words, the increase in throughput can be thought of as the breakeven point in terms of how much reduction in clock speed (due to the extra dMT hardware) can be tolerated by dMT before it performs worse than a conventional organisation.

We will assume that dMT configurations never include branch prediction. Our focus is on throughput and we find that the branch predictor confers a minimal benefit from this standpoint. Indeed, as we show later, a dMT configuration without branch prediction does better than a non-dMT configuration with branch prediction! Branch prediction does impact end-to-end execution time of a single thread, but misprediction handling in the dMT configuration is more complex than in a conventional pipeline; either rolling back all extant threads (contravening our throughput goal), or requiring extra hardware to squash only mis-speculated instructions in the mispredicting thread. Adding and evaluating branch prediction in a dMT pipeline is an interesting area for future work.

Finally, it is worthwhile to comment on the impact of dMT on multithreaded programs. From a correctness standpoint, a dMT implementation is no different than any other MT or CMP processor; namely consistency must be enforced by appropriate synchronisation. Coherence, on the other hand, is a non-issue here, because the dMT threads share a common cache. Accesses to the same physical address (for example, the first thread may take a cache miss when reading some location, and while that thread is stalled, the other thread may attempt to write to the same location) do need to be identified but can be treated as data hazards. From a performance standpoint, it is true that synchronisation delays will sacrifice some of the reported improvements in throughput. But even for many MT programs, these delays should not be ubiquitous to the point where all the improvement is forfeited. We will focus on the performance of independent programs on a dMT machine; evaluating multithreaded programs is an interesting area for future work.

4 Simulator and benchmarks

We model BMT and dMT by using Wisconsin's SimpleScalar 3.0a software package [16] We assume that in the absence of other stalls, all instructions take one cycle to execute. Our baseline assumption is that the cache miss penalty for the first-level cache is five cycles (to a second-level cache or to some form of embedded DRAM); later we also consider a longer miss penalty of ten cycles.

We use a mix of benchmark programs from the SPEC95 [6], MediaBench [5], and MiBench [4] suites, and we also used the Dhrystone benchmark [17]. Rather than show data averaged across all benchmarks as in [1], we chose pairs of benchmarks to run together. This lets us show a richer variety of reactions to BMT and dMT; some benchmarks have very distinctive behaviours. Naturally, it was impossible to present data for all possible pairs, so we chose a subset of the benchmarks from each suite and selected the most sensible pairs that we could derive; such as a game and an image utility (go and ijpeg) or a speech compression/decompression tool and an image compression/decompression tool (gsmencode and epicencode/gsmdecode and epicdecode). Despite its well known drawbacks, we also include the Dhrystone benchmark because it has been extensively used so for benchmarking embedded microprocessors. We chose to use SPEC95 over SPEC2000 [18] because we were specifically interested in go (a game and also a program known for its poor branch behaviour) and ijpeg (an image-processing program).

Except for epic, jpeg, pegwit, dhrystone and MiBench, which are short and were run to completion, all simulations were fast-forwarded according to the methodology in [19] (for SPEC) or 100 million instructions (for longer MediaBench programs) in order to avoid unrepresentative initial behaviour, and statistics were gathered for the next 100 million instructions. Because all programs run for approximately the same length in our simulations, we report results using simple arithmetic means.

The benchmarks were compiled with gcc 2.6.3 and -O3 optimisation for the SimpleScalar PISA instruction set.

It would also be interesting to evaluate the benefits of dMT from an energy-efficiency standpoint. Although the extra register file and pipeline register bits will increase power dissipation, a higher throughput will reduce total execution time and hence total energy consumed for a given workload. A detailed energy evaluation is another interesting area for future work.

5 Results

5.1 Experimental configurations

To identify the different configurations we explore, each is named in the pattern XXX-YYY, where XXX is either 'base' or 'MT' and YYY indicates the processor organisation. The same configuration is used for both BMT and dMT. The following abbreviations are used:

0: I-cache only
f: forwarding
C: large cache configuration (16-kB I-cache and 8-kB D-cache, both four-way)
c: small cache configuration (8-kB I-cache and 2-kB D-cache, both two-way)
b: dynamic branch predictor

So 'base-Cfb' is an ARM10-like, five-stage, single-issue processor with forwarding, a 16-kB I-cache and 8-kB D-cache, and a bimodal branch predictor; 'MT-cf' is a multithreaded processor (either BMT or dMT) with forwarding, an 8-kB I-cache and 2-kB D-cache, and no branch prediction. When a dynamic predictor is used, it is a 2-bit bimodal predictor with a 128-entry branch target buffer, as in the ColdFire v4 [20].

We present data for a total of 14 different pairings of 18 different benchmark programs. We used these pairs to make ten different comparisons of non-multithreaded and multithreaded processors. For the most interesting seven comparisons, we present a graph (Figs. 2–4) that compares their IPCs for each of the benchmark pairs. (Additional graphs for the other configurations can be found in [7].) In each comparison, various multithreaded organisations are compared to a comparably configured or superior baseline configuration. Figure 2 compares three systems with equivalent cache, branch prediction, and forwarding (f compared to f, etc.). These show how much extra throughput is obtained by simply adding dMT to an existing design. The benefits are especially notable for a configuration without data caches. Figure 3 adds the ColdFire's dynamic branch predictor to each baseline configuration, but for each of these comparisons, MT still omits branch prediction, switching threads instead. These results demonstrate MT's robustness against control hazards. Figure 4 presents one comparison where the baseline system uses the larger cache sizes listed above, while the MT systems use the smaller. This demonstrates dMT's capability to reduce the need for larger caches. Note that some configurations appear in more than one comparison; this is in order to illustrate the tradeoffs dMT permits. In addition, we present data for some of the same pairings but varying cache miss penalty in Fig. 5, and present data for six quadruplets of SPEC95 and MediaBench applications in Fig. 6 to show the performance of four-way dMT. Table 1 gives the average improvement in throughput (IPC) for each comparison. In each Figure, we evaluate the performance of only the blocked policy for BMT and dMT; results for the every-cycle and prioritised policies can be found in [7].

It is useful to compare the hardware overhead of dMT, although this is difficult to do since the area overhead is heavily dependent on a myriad of design choices. However, we can count the number of bits of each type: pipeline register, register file and cache. We estimate that each set of copies of the pipeline registers requires 1237 bits, so the

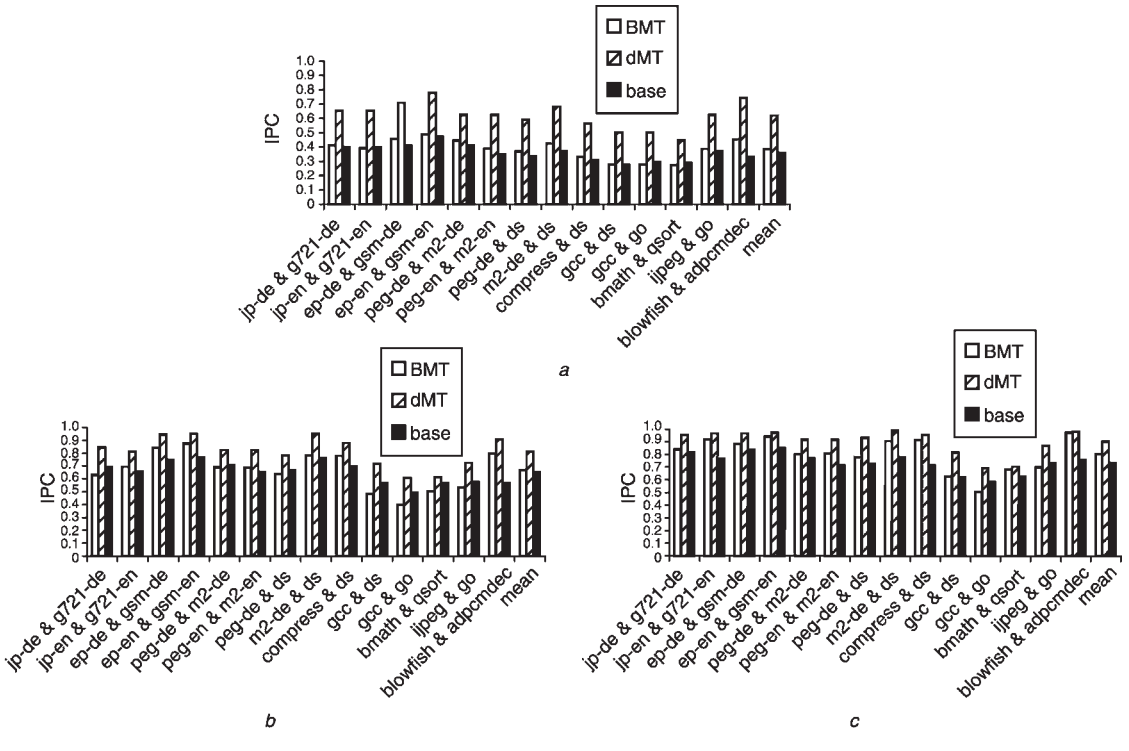

Fig. 2 *Comparison of baseline, dMT and BMT configurations with similar organisations*

a Base-f compared to MT-f
b Base-cf compared to MT-cf
c Base-Cf compared to MT-Cf

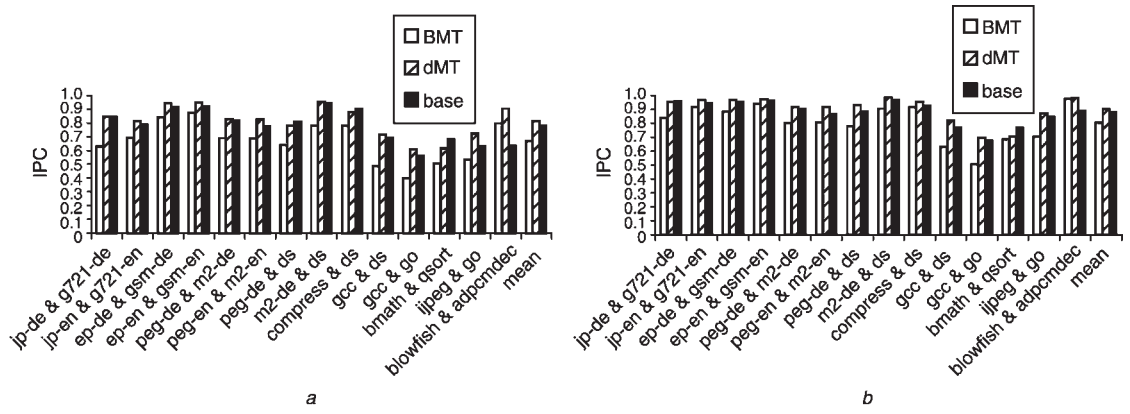

Fig. 3 *Comparison of baseline, dMT, and BMT configurations with similar organisations except lack of branch prediction for MT*

a Base-cfb compared to MT-cf
b Base-Cfb compared to MT-Cf

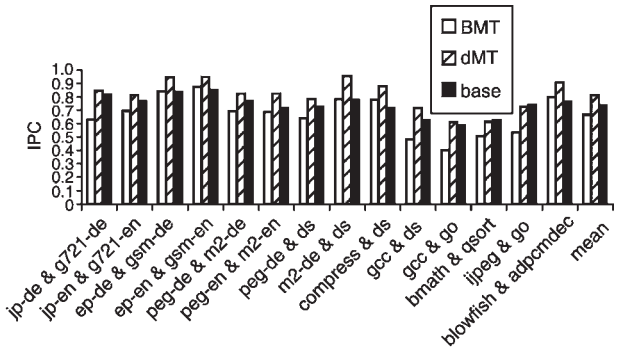

Fig. 4 *Comparison of baseline, dMT, and BMT configurations with similar organisations except that the MT configuration has less cache (base Cf vs. MT-cf)*

non-dMT implementation uses 1237 bits, the two-way dMT uses 2474, and the four-way uses 4948. For a 32-bit machine, we estimate that each copy of the register file costs 2048 bits. Among hardware that is not duplicated, the branch predictor requires 8192 bits; the 'c' cache configuration requires 91 136 bits, and the 'C' configuration requires 215 808 bits. These estimates assume MESI state bits in the cache and no ECC bits in the cache or register file.

The MT pipelines run both instruction streams simultaneously. In contrast, the baseline architecture runs the two benchmarks back-to-back. To obtain the most comparable results across organisations, we terminate the MT simulations when one instruction stream terminates or when it completes its simulation allotment of 100 million instructions. The baseline simulation then runs its two programs for exactly the same number of instructions as the MT simulation.

We also measured the frequency with which dMT (blocking policy) switches threads. Arithmetic means taken across our set of benchmark pairs are reported in Table 2. As expected, with no forwarding (the '0' configuration), thread switches are frequent (every 1.7 instructions), and as the configuration becomes more aggressive, thread switches become less frequent (reaching a level of every 5.3 instructions with 'Cf').

5.2 Discussion

For equivalent cache, branch-prediction, and forwarding configurations (Fig. 2), both dMT choices obtain a dramatically better throughput than the non-multithreaded organisation. This is perhaps not a fair comparison, because dMT has a small amount of extra hardware. However, it shows that dMT does indeed recapture stall cycles and uses them to boost throughput, and it shows that adding dMT can improve throughput substantially; dramatically so for a design without data caches. For two-way dMT, it boosts throughput by 73% in a processor with no data cache, by more than 24% in the small cache configuration, and by slightly less than 23% in the large cache configuration, all of which are certainly more than any performance loss due to extra latency introduced by the multiplexors after pipeline stages. For dMT, sometimes the blocking policy outperforms the prioritised policy and *vice-versa*, but overall, the difference is small, a few percent. This is due to specific reactions to cache contention. The prioritised policy will be useful for real-time workloads where a specific thread must complete in a specified amount of time or requires some other determinism. The prioritised policy replicates dedicated pipeline behaviour for the prioritised thread except for the absence of branch prediction (which in any case is probably not desirable for workloads requiring determinism) and except for the possibility of cache contention between the dMT threads.

Note that, in the large cache configuration, dMT comes quite close to the ideal of 1 IPC, even though dMT omits a branch predictor. Also note that dMT outperforms BMT for all these configurations.

Figures 3 and 4 compare different organisations that highlight the tradeoffs that multithreading permits. We find that dMT can be used in place of a dynamic branch predictor or to allow substantially smaller caches. Table 1 shows that dMT can also be used in place of forwarding, although this seems an unlikely design choice.

Of course, with only two threads, dMT is not able to recapture all stall cycles. This means that in the extreme cases we examine, where the MT configurations are substantially handicapped compared to the baseline, dMT's performance is outpaced by the baseline architecture. A few of these are shown in Table 1, but since the results are negative, we omit corresponding graphs in the interest of space. While dMT is not a cure-all, its ability to recapture stall cycles is impressive, and our results suggest that it makes possible many interesting hardware tradeoffs.

Figure 3 shows that block multithreading can be used to replace a small, 128-entry, 2-bit bimodal predictor. For base-cfb compared to MT-cf, dMT is superior for all but two

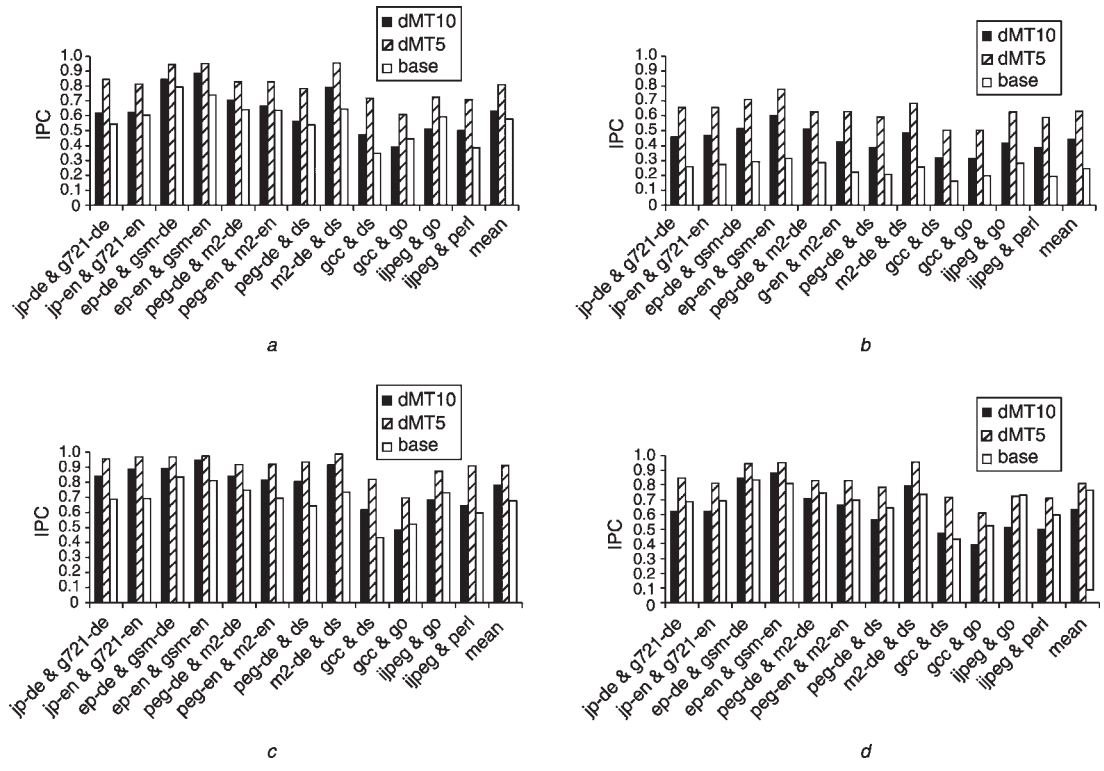

Fig. 5 *Comparison of base and dMT configurations for a larger cache miss penalty of ten cycles; dMT5 is the performance of the five-cycle miss penalty configuration from previous graphs*

a dMT-cf compared to base-cf
b dMT-f compared to base-f
c dMT-Cf compared to base-Cf
d dMT-cf compard to base-Cf

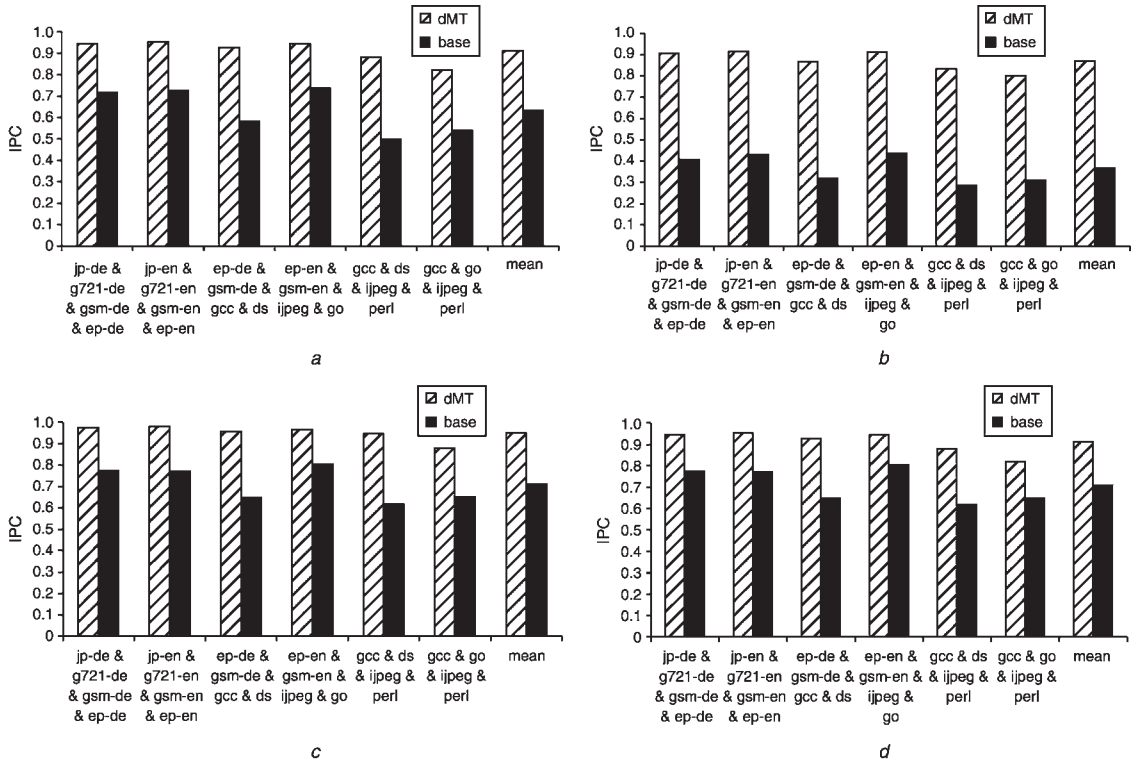

Fig. 6 *Comparison of base and dMT configurations for four threads*

a dMT-cf compared to base-cf
b dMT-f compared to base-f
c dMT-Cf compared to base-Cf
d dMT-cf compared to base-Cf

Table 1: Baseline against MT comparisons. Δ IPC columns give the throughout improvement for a two-way dMT with a five-cycle miss penalty, a two-way dMT with a ten-cycle miss penalty, and a four-way dMT with a five-cycle miss penalty (arithmetic means)

Baseline configuration	MT configuration	Graph	Δ IPC, %	Δ IPC(dmt10), %	Δ IPC(4X), %
Base-f	MT-f	Fig. 2	73.2	80.2	139.0
Base-cf	MT-cf	Fig. 2	24.5	9.4	43.9
Base-Cf	MT-Cf	Fig. 2	22.6	15.2	33.8
Base-cfb	MT-cf	Fig. 3	4.2	−6.0	23.6
Base-Cfb	MT-Cf	Fig. 3	2.6	−4.0	12.9
Base-cf	MT-f	na	−4.9	−23.7	37.8
Base-Cf	MT-cf	Fig. 4	10.4	−6.4	17.2
Base-Cfb	MT-cf	na	−7.6	−22.4	8.7
Base-f	MT-0	na	44.3	48.7	84.7
Base-cf	MT-0	na	−20.7	−37.2	12.3

Table 2: Instructions between dMT instruction-stream switches (arithmetic mean)

dMT configuration	Instructions per switch
0	1.72
f	2.22
cf	3.72
Cf	5.28

benchmark pairs. For base-Cfb compared to MT-Cf, dMT is superior for all but one benchmark pair, jpeg-decode & g721-decode, which is a tie. This means that dMT is successful in finding useful work in the alternate thread when a branch is detected. From a hardware cost standpoint, let us assume that the branch predictor and register file both use minimum-size transistors [21]. Each register-file cell will still be twice as large as than a branch predictor cell, due to the registers' extra read/write ports, so the extra register file is equivalent to 4096 branch-predictor bits. Even if each pipeline-register cell is four times larger than a branch-predictor cell, two-way dMT and the baseline end up almost the same in terms of hardware cost. Four-way dMT, however, will be somewhat more expensive, because the extra register files and pipeline registers will more than consume the hardware savings of eliminating the small branch predictor.

Figure 4 shows that dMT can be used to reduce cache sizes. 'dMT-cf' is better than 'base-Cf', usually substantially so, for all but two benchmark pairs. This means that dMT can be used to reduce, without penalising throughput, the cache configuration in an aggressive system like the ColdFire v4, with a 16-kB I-cache and 8-kB D-cache, both four-way associative, down to a more modest 8-kB I-cache and 2-kB D-cache, both two-way associative. Indeed, even after this reduction in cache, dMT's throughput is still 10.4% better. From a hardware cost standpoint, let us assume that the cache and register file use minimum-size transistors. The extra register file is equivalent to 4096 cache bits. However, the reduction in cache size is 124 672 bits. Even when the extra register files and pipeline registers are accounted for, both two-way and four-way dMT are clearly still a substantial net win from a hardware standpoint.

From a throughput standpoint, neither dMT nor BMT can be used to replace the D-cache entirely, as seen in Table 1 for 'base-cf' compared to 'MT-f', where dMT is 5% worse and BMT more so. It remains to be seen whether a very tiny D-cache might suffice for dMT. It is also interesting to note in the 'base-Cfb' compared to 'MT-cf' comparison in Table 1 that reducing both the I- and D-cache sizes and also removing the dynamic branch predictor incur only modest reductions in throughput for many of the benchmarks; on average, dMT is only 7.6% worse.

Adding the required multiplexor between the multiple pipeline registers required by dMT may slow the clock rate. Since adjusting clock rate to accommodate these multiplexors is equivalent to lowering IPC for the same clock rate, the IPC improvements in Table 1 show how much change in clock rate can be accommodated. For example, if we assume that dMT will reduce the clock rate by 10%, adding dMT instead of a branch predictor ('base-cfb' compared to 'dMT-cf' or 'base-Cfb' compared to 'dMT-Cf') may or may not be worthwhile, but even with a possible 10% penalty due to extra hardware, dMT still improves throughput substantially for similar configurations ('base-f' compared to 'dMT-f', 'base-cf' compared to 'dMT-cf', and 'base-Cf' compared to 'dMT-Cf'), and dMT allows use of smaller caches without reducing throughput ('base-Cf' compared to 'dMT-cf'). If the multiplexing can be implemented with less overhead, other configurations may also become competitive. Of course, it may be that even small reductions in throughput would be tolerated if enough hardware savings can be realised, for example the branch predictor ('base-cfb' compared to 'dMT-cf' and 'base-Cfb' compared to 'dMT-Cf').

To explore sensitivity to cache miss latency, Fig. 5 shows the performance of similar base and dMT configurations, except that dMT is evaluated with both five- and ten-cycle miss penalties. The larger miss penalty still exhibits positive, albeit diminishing returns from dMT. This Figure shows that dMT will maintain the same relative performance edge at least for modestly higher-latency L2 caches or for fast embedded DRAM.

Figure 6 gives the performance of dMT with four thread contexts. Here we formed quadruplets by selecting a mix of threads for which we observed varying performances in the two-thread experiments. These results show that running a high number of threads can come quite close to masking the penalty of an L1 cache absence, simply by switching the active thread as necessary. There is even less need for cache for this configuration, since the large cache configuration does not outperform the small cache configuration by a reasonable margin. Of course, there are other ways to run

the same four threads, for example two dMT processors each running two threads. This will give a higher per thread performance but exhibit more hardware cost. This illustrates the range of interesting tradeoffs that dMT opens up, an extensive study of which is another interesting question for future work.

6 Conclusions

We have presented dMT, an inexpensive technique for sharing a single pipeline between multiple active threads. The addition of duplicated pipeline registers has been shown to enable the capture of in-flight instructions anywhere in the pipeline. This organisation is thus able to respond not only to data hazards and branch delays, but also to misses in the primary I-cache and D-cache. This combines beneficial aspects of BMT [1] and Runahead [3], and allows attractive hardware tradeoffs. We have also shown that dMT can reclaim a significant amount of wasted cycles. For processors without data caches, dMT boosts throughput by 71.6% over a non-multithreaded organisation, and for processors with cache, dMT boosts throughput by 23–24%. Better yet, instead of using dMT to increase the throughput of more complex, single-threaded configurations, it can be used, without reducing throughput, to eliminate the dynamic branch predictor or to reduce instruction- and data-cache sizes.

Our results also show that dMT is consistently superior to BMT, because dMT handles a wider variety of stall conditions. Furthermore, like the original BMT concept, our design can take advantage of each of the thread switching policies. Of particular note is the prioritised switching policy, which returns control to the prioritised thread immediately after its stall condition is resolved, making dMT viable in real-time systems.

Our results suggest a variety of interesting avenues for future work. Incorporating and evaluating branch prediction in a dMT pipeline, and evaluating multithreaded, data-sharing programs are both interesting questions. It would also be interesting to evaluate the benefits of dMT from an energy-efficiency standpoint. Finally, dMT opens up a new dimension of design, allowing CPUs with various degrees of dMT to be combined into a multiprocessor system, and the proper balance of dMT with the number of processors is an open question.

7 Acknowledgments

This material is based upon work supported in part by the US National Science Foundation under grants. CCR-0082671 and CCR-0133634. We would also like to thank Doug Clark, Yingmin Li, and Mircea Stan for their valuable advice and assistance as we developed this work and the anonymous reviewers for their detailed and insightful comments.

8 References

1 Farrens, M.K., and Pleszkun, A.R.: 'Strategies for achieving improved processor throughput'. Proc. 18th Annual Int. Symp. on Computer Architecture (ISCA-18), Toronto, Canada, May 1991, pp. 362–369
2 Haskins, J.W., Jr., Hirst, K.R., and Skadron, K.: 'Inexpensive throughput enhancement in small-scale embedded microprocessors with block multithreading: Extensions, characterization and tradeoffs'. Proc. 20th IEEE Int. Conf. on Performance, Computing and Communications, Phoenix, AZ, April 2001, pp. 319–328
3 Dundas, J., and Mudge, T.: 'Improving data cache performance by pre-executing instructions under a cache miss'. Proc. ACM Int. Conf. on Supercomputing, Vienna, Austria, July 1997, pp. 68–75
4 Guthaus, M.R., Ringenberg, J.S., Ernst, D., Austin, T.M., Mudge, T., and Brown, R.B.: 'Mibench: A free, commercially representative embedded benchmark suite'. Proc. IEEE 4th Annual Workshop on Workload Characterization, Austin, TX, December 2001
5 Lee, C., Potkonjak, M., and Mangione-Smith, W.H.: 'Mediabench: A tool for evaluating multimedia and communications systems'. Proc. 30th Int. Symp. on Microarchitecture, Research Triangle Park, NC, December 1997, pp. 330–335
6 Standard Performance Evaluation Corporation, SPEC CPU95 Benchmarks. http://www.specbench.org/osg/cpu95, Dec. 1999
7 Hirst, K.R., Haskins, J.W., Jr., and Skadron, K.: 'dMT: Inexpensive throughput enhancement in small-scale embedded microprocessors with differential multithreading: Extended results'. Tech. Report CS-2003-18, Univ. of Virginia, Dept. of Computer Science, Oct. 2003
8 Alverson, R., Callahan, D., Cummings, D., Koblenz, B., Porterfield, A., and Smith, B.: 'The Tera computer system'. Proc. Int. Conf. on Supercomputing, Rhodes, Greece, May 1990
9 Mitchell, N., Carter, L., Ferrante, J., and Tullsen, D.: 'ILP versus TLP on SMT'. Proc. Int. Conf. on Supercomputing, November 1999
10 Tullsen, D.M., Eggers, S.J., Emer, J.S., Levy, H.M., Lo, J.L., and Stamm, R.L.: 'Exploiting choice: Instruction fetch and issue on an implementable simultaneous multithreading processor'. Proc. 23rd Int. Symp. on Computer Architecture, Philadelphia, PA, May 1996
11 Agarwal, A., Lim, B.-H., Kranz, D., and Kubiatowicz, J.: 'APRIL: A processor architecture for multiprocessing'. Proc. 17th Int. Symp. on Computer Architecture, Seattle, WA, May 1990
12 Motorola Semiconductor Products. Motorola's DragonBall(tm) integrated processor family. http://www.mot.com/SPS/WIRELESS/products/DragonBall.html, 1999
13 Scherrer, T.: Thomas Scherrer Z80-family official support page. http://www.geocities.com/SiliconValley/Peaks/3938/z80_home.htm, 2000
14 Gwennap, L.: 'ARM10 points to set-tops, handhelds', *Microprocess. Rep.*, 16 Nov. 1998
15 Borkar, S.: 'Design challenges of technology scaling', *IEEE Micro*, Jul.-Aug. 1999, pp. 23–29
16 Burger, D.C., and Austin, T.M.: 'The simplescalar tool set, version 2.0', *Comput. Archit. News*, 1997, **25**, (3), pp. 13–25
17 Weicker, R.P.: 'Dhrystone: A synthetic systems programming benchmark', *Commun. ACM*, 1984, **27**, (10), pp. 1013–1030
18 Standard Performance Evaluation Corporation. SPEC CPU2000 Benchmarks. http://www.specbench.org/osg/cpu2000, Dec. 1999
19 Skadron, K., Ahuja, P.S., Martonosi, M., and Clark, D.W.: 'Branch prediction, instruction-window size and cache size: Performance tradeoffs and simulation techniques', *IEEE Trans. Comput.*, 1999, **48**, (11), pp. 1260–1281
20 Turley, J.: 'ColdFire doubles performance with v4', *Microprocess. Rep.*, 1998
21 Steinhaus, M., Kolla, R., Larriba-Pey, J., Ungerer, T., and Valero, M.: 'Transistor count and chip-space estimation of simplescalar based microprocessor models'. Proc. Workshop on Complexity-effective Design, Göteborg, Sweden, June 2001