Low-Latency, Concurrent Checkpointing for Parallel Programs

James S. Plank

Department of Computer Science
Univeristy of Tennessee
Knoxville, TN 37996
plank@cs.utk.edu

Kai Li

Department of Computer Science
Princeton University
Princeton, NJ 08544
li@cs.princeton.edu

appearing in:

IEEE Transactions on Parallel and Distributed Systems, 5(8), Aug 1994, pp 874–879.

Low-Latency, Concurrent Checkpointing for Parallel Programs*

Kai Li Jeffrey F. Naughton James S. Plank

August 9, 1994

Abstract

This paper presents the results of an implementation of several algorithms for checkpointing and restarting parallel programs on shared-memory multiprocessors. The algorithms are compared according to the metrics of overall checkpointing time, overhead imposed by the checkpointer on the target program, and amount of time during which the checkpointer interrupts the target program. The best algorithm measured achieves its efficiency through a variation of copy-on-write, which allows the most time-consuming operations of the checkpoint to be overlapped with the running of the program being checkpointed.

Introduction

This paper presents four algorithms for checkpointing and restarting parallel programs running on shared-memory multiprocessors. To test the efficiency of these algorithms, we implemented them on the DEC Firefly multiprocessor [29], and profiled their performance on five benchmark programs. The algorithms range from very simple to more complex, utilizing techniques such as copy-on-write [9, 30] and buffering [26] to realize good performance on the multiprocessor. In the best of these algorithms, most of the checkpoint is taken in parallel with the target program's execution, and when it does interrupt the target, the interrupts are for small, fixed periods of time (under 0.1 seconds in our implementation).

The main use of checkpointing is to provide the mechanism for performing *backward error recovery*, a general means of fault tolerance as defined in [1]. The strength of backward error recovery is its ability to provide fault tolerance in the presence of unanticipated faults – faults which were not envisioned in the design of the system. No other means of fault tolerance has this property.

Checkpointing can also be used as a means of process migration or coarse-grained job-swapping. This is the intended use of the Condor checkpointer [19]. In fact, one can view backward error recovery as merely process migration to the same machine at a different point in time.

*This research was supported in part by the National Science Foundation under grants CCR-8814265 and IRI-8909795, and by the Digital Equipment External Research Program and Systems Research Center. Kai Li's address: Computer Science Department, Princeton University. Jeffrey Naughton's address: Computer Science Department, University of Wisconsin. James Plank's address: Computer Science Department, University of Tennessee, Knoxville

All of the checkpointing algorithms presented take a "full checkpoint" – they checkpoint the entire state of the target program. The alternative would be to take "incremental" checkpoints, which save only that portion of the state that has changed since the last checkpoint. We have concentrated on full checkpoints to test the worst-case behavior of these algorithms. The work involved in taking an incremental checkpoint is a subset of the work involved in taking a full checkpoint, and therefore if the algorithms exhibit good efficiency, high concurrency, and low latency in taking a full checkpoint, then taking an incremental checkpoint can only exhibit better performance in such measures.

The Checkpointing Algorithms

The goal of a checkpointing algorithm is to establish a *recovery point* in the execution of the program, and save enough information to reconstruct the state of the program at this recovery point in the event of a failure. In the case of a uniprocessor, a checkpoint can be taken by freezing the processor and saving the state of the CPU (i.e. the registers) and the state of memory to disk. Note that in saving the state of memory, one need not save the executable code itself, as this can be reconstructed from the executable file. To restore this recovery point, one merely reads the state of memory from disk and then restores the state of the CPU. Thus checkpointing is exactly like generating a core file (which is of course a type of checkpoint). This is the approach taken by Condor [19].

To generate a checkpoint for a shared-memory multiprocessor, one can do the analogous thing: Freeze all the processors, then save the states of all the CPUs and the state of memory to disk. We implemented this simple algorithm and called it *sequential*, as it performs none of its work in parallel with the program which it is checkpointing.

We set three goals for a good checkpointing algorithm on a multiprocessor:

1. It must be reasonably efficient.
2. It must impose little overhead on the target program. In other words, it should attempt to be maximally concurrent.
3. What overhead it does impose must be low latency – that is, it should interrupt the target program for only small periods of time.

We believe that reasonable values for these goals are as follows: The overall checkpoint time should be no more than twice the optimal checkpoint time. The checkpointing should add no more than 20% to the running time of the program during the time that it is checkpointing. The latency of interrupts should be kept below a tenth of a second (we chose this value because any more might be perceived as noticeable to the user watching the program's execution).

The sequential checkpointing algorithm is clearly optimal in terms of overall checkpoint time, as it is limited solely by the duration of the disk writes. However, it has zero concurrency and is 100% latency. The second algorithm that we implemented attempts to improve the concurrency of checkpointing. We call it *main memory* checkpointing, as it freezes the processors, saves the

checkpoint into a separate address space in main memory, and then restarts the processors. After starting the processors, the algorithm forks a new thread in the new address space that writes the checkpoint to disk. In this algorithm, the concurrency of checkpointing should be improved over the sequential algorithm, as the execution of the target program is overlapped with the writing of the checkpoint to disk.

To improve the latency of the main memory algorithm, we implemented a third algorithm, which uses *copy-on-write* [9, 30] to make the main memory checkpoint. Copy-on-write is a technique which uses a processor's virtual memory page protection hardware to make a memory-to-memory copy with low latency. The copy-on-write checkpointing algorithm works as follows: First, it freezes the processors and copies their CPU states to the separate address space. Also, it sets the page protection bits of all pages in main memory to be "read-only." Next, it unfreezes the processors, and starts a separate *copier* thread that copies pages to the new address space and resets the pages' protection to "read-write." If a thread of the target program generates a page access violation, then it must write that page to the new address space before setting its protection to "read-write" and restarting. When the copier thread is done copying, the main memory checkpoint is complete, and the copier thread writes the checkpoint to disk.

Finally, we implemented a fourth algorithm which we call *concurrent, low-latency (CLL)* checkpointing. It improves upon the copy-on-write algorithm by adding buffering, a standard operating systems technique usually used to hide disk latency during file system writes. What the buffering does in this case is allow the checkpoint to be written to disk at the same time as it is being copied from memory. Specifically, the CLL algorithm allocates a fixed pool of pages (the buffer) in the second address space that the copier thread and page fault handlers fill, and that a new thread called the *writer* thread empties by writing the pages to disk. The writer and copier threads are both started immediately after the processors are unfrozen.

All four algorithms are shown graphically in Figure 1.

The improvements of the CLL algorithm over the copy-on-write algorithm should be twofold. First, the extra memory requirements of this scheme are fixed – they are the size of the page pool. The copy-on-write scheme needs extra space that is the same size as the target program's address space, and thus is more likely to cause thrashing. Second, the overall checkpoint time of the CLL algorithm should be less than that of the copy-on-write algorithm. This is because the CLL algorithm starts writing to disk as soon as the processors are restarted, instead of waiting until a complete main-memory checkpoint has been made.

A possible concern of the CLL algorithm is what happens when the page pool fills up. Then pages in the pool are freed only as fast as they can be written to disk. If the size of the page pool is chosen to be the amount of available physical memory, then the CLL algorithm should still outperform the copy-on-write algorithm for the following reason: In the copy-on-write algorithm, pages might be swapped to the swap area on disk so that the checkpoint can fit into main memory. If these pages are part of the checkpoint, then they must be swapped back into memory to be written to the checkpoint file. If they are part of the target program, then they will eventually have to be swapped back into memory when the program needs them. In either case, the copy-on-write algorithm performs an extra disk write and read for each swapped page, while the CLL algorithm needs no swapping, and therefore performs no extra disk writes. This extreme case for

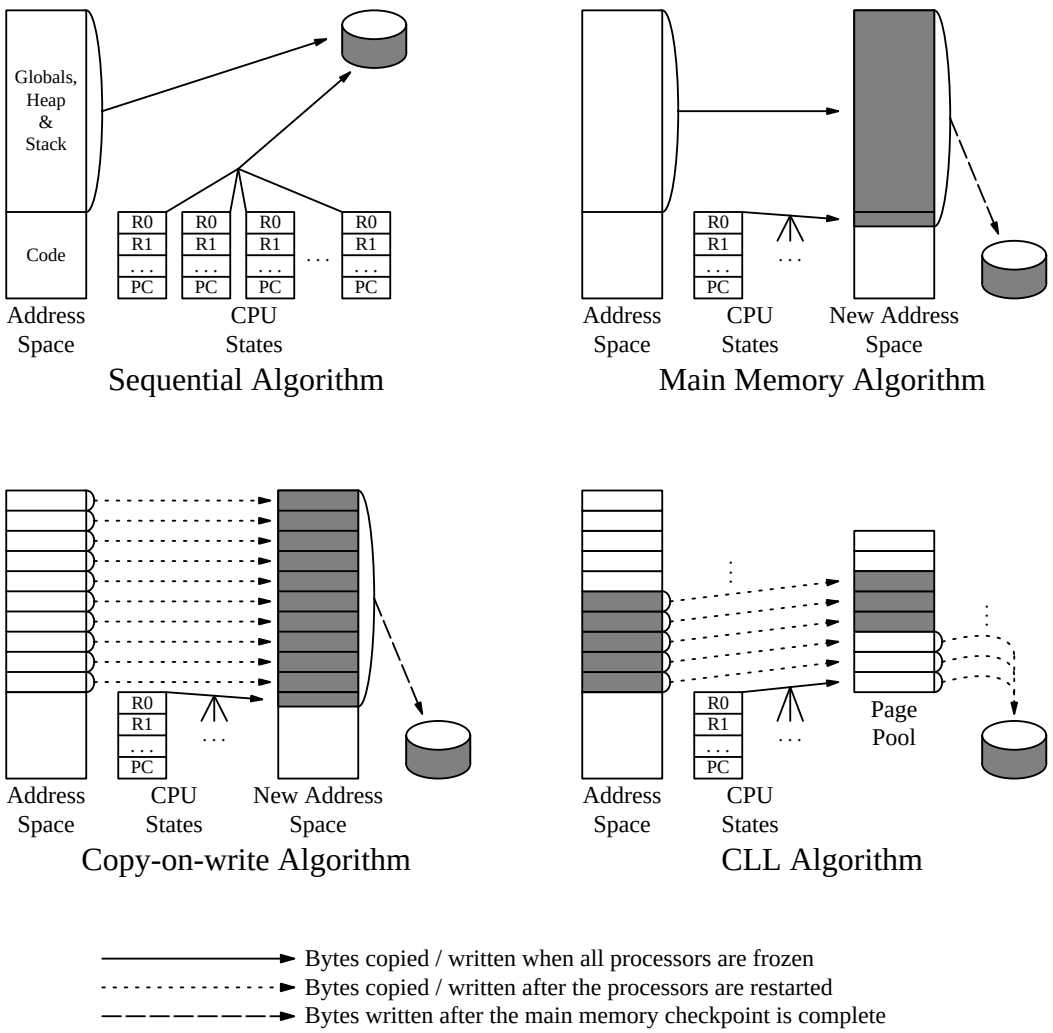


Figure 1: The four checkpointing algorithms.

both algorithms is tested in our implementation.

Recovery

Recovering from a checkpoint is straightforward, and is the same for all four algorithms. The processors are frozen, and the contents of main memory are replaced with the contents saved in the checkpoint. The states of the CPUs are restored to their states at the recovery point, which are also saved in the checkpoint. When the processors are restarted, execution of the target program continues from the recovery point.

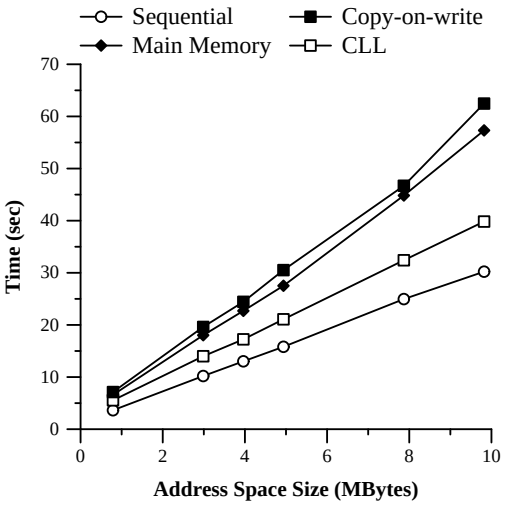
Implementation

We have implemented all four checkpointing algorithms as well as recovery on a DEC Firefly multiprocessor. The Firefly is an experimental shared-memory multiprocessor developed at the DEC System Research Center [29]. A Firefly consists of four CVAX processors, each with a floating point unit and a direct-mapped 64 Kbyte cache. The caches are coherent, so that all processors within a single Firefly see a consistent view of shared memory. The operating system for the Firefly is Taos [20], an Ultrix with threads and cheap thread synchronizations.

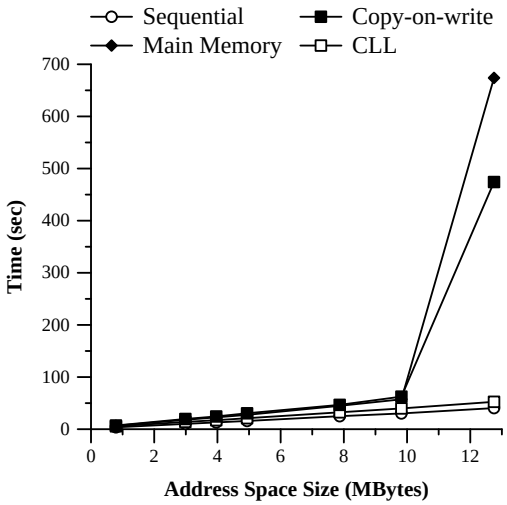
The Firefly upon which we tested our algorithms has 16 Mbytes of physical memory, and a separate I/O processor that shares memory with the other four processors but also has a separate bus connecting the Firefly to I/O devices and the outside world. In our system, the I/O processor is connected to a RD54 disk drive.

The implementation is written in Modula-2+ [24], an extension of Modula-2. The user interface to the checkpointing routines is a single call to *setup_checkpoint()* at the beginning of the user's program. This call is used either to restore the program's execution to a saved recovery point, or to specify how long the checkpointer should wait before interrupting the program to take a checkpoint. Freezing the processors is provided by Taos through a system call. In the absence of such a system call, one could freeze the processors either through interprocessor interrupts, or by protecting all the pages in memory to be "no access," and gaining control of the processors in the page fault handlers.

In taking a checkpoint only the user's address space is saved. The states of the kernel and the file system are not saved. The ramifications of this decision are that constructs that rely on kernel and external state, such as remote procedure calls and open file pointers, are not guaranteed to be recoverable. Thus, the programs which we tested were bereft of such constructs. We view these restrictions as tolerable for two reasons. First, the goal of our implementation is to examine the performance of checkpointing algorithms in regard to the metrics of speed, concurrency and latency. The goal is not to write a production-level checkpointing system for the Firefly. Second, recoverable kernels have been studied and implemented [6, 21], as have checkpointers for uniprocessors which either provide recovery for read-only and sequential read-write files [19], or



Graph 1: Checkpoint Time.



Graph 2: Checkpoint Time for Large Address Spaces.

rewrite the Unix[1] file system to be completely recoverable [28]. We couldn't justify taking the time to duplicate this work on Taos when the result is so tangential to our experiments.

One of the variables in our implementation is the page size. Although the actual page size on the Firefly's memory management unit is 512 bytes, we emulate different page sizes by varying the number of bytes which are copied to the second address space by the copier thread and by the page fault handlers. Larger pages will increase the time to handle a fault, but they will also decrease the number of faults, and because of locality of reference, they may also decrease the rapidity at which faults occur.

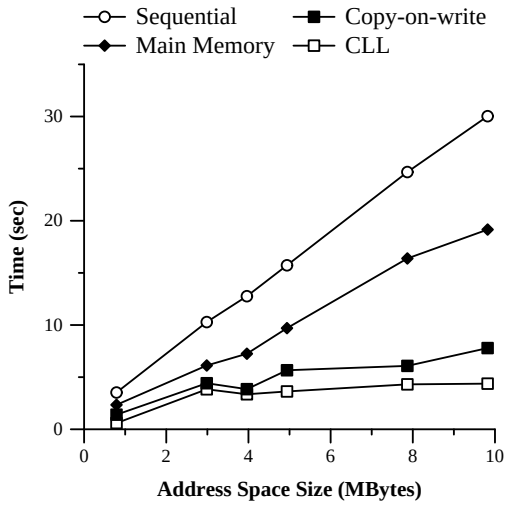
Experiments

For our initial experiments, we tested all four checkpointing algorithms on a parallel implementation of merge sort. We also checkpointed four other parallel programs: traveling salesman, matrix multiplication, pattern matching, and bubble sort. Since the results from experiments with these programs were so similar to the results with merge sort, we do not present them here. The merge sort program sorts 250,000 indexed records, where the record size can be changed to modify the size of the program's address space.

In all experiments, the four processors of the Firefly are partitioned so that the target program uses three and the checkpointer uses one. This is to measure the maximal concurrency of our checkpointing methods. The checkpointer waits for the target to run for ten seconds, and then it takes one complete snapshot. For the results presented here, the page size is eight Kbytes, and the page pool size is 1 Mbyte. All times represent wall-clock time when the target program and checkpointer have exclusive use of the system. All times are averages of five or more runs.

Graphs 1 through 4 display the overhead imposed by the four checkpointing algorithms, as

[1] Unix is a trademark of AT&T



a function of the size of merge sort's address space. The total checkpoint time displayed in Graph 1 measures the elapsed time from the start of the checkpoint to its conclusion. Graph 2 extends Graph 1 to include the checkpoint time when the address space approaches the size of physical memory. The checkpoint overhead in Graph 3 is the amount of time by which the checkpoint increases the target's running time. Graph 4 displays the overhead as a percentage of the checkpoint's running time. This is equal to checkpoint overhead divided by the checkpoint time. Concurrency can be calculated as:

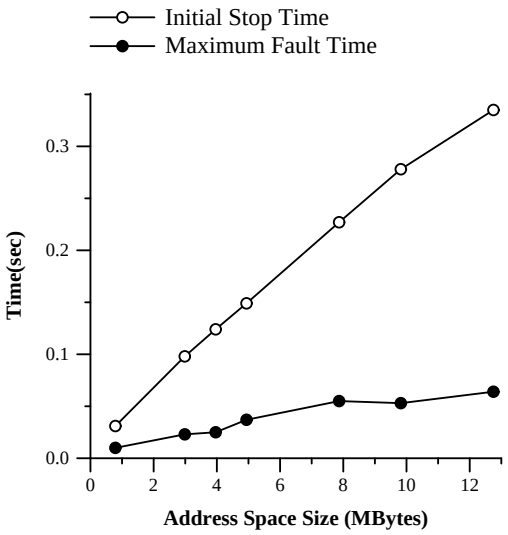
$$concurrency = 100\% - overhead.$$

It comes as no surprise that in all four algorithms, checkpoint time is proportional to the size of the address space. Also as expected, the sequential algorithm records the fastest checkpointing time, followed by the CLL algorithm. The other two algorithms take longer to record a checkpoint because they wait for a complete copy of the checkpoint to exist in main memory before writing the checkpoint to disk. Of these two algorithms, the copy-on-write algorithm takes the longest because of the extra work it spends processing page faults, and because it copies a page at a time.

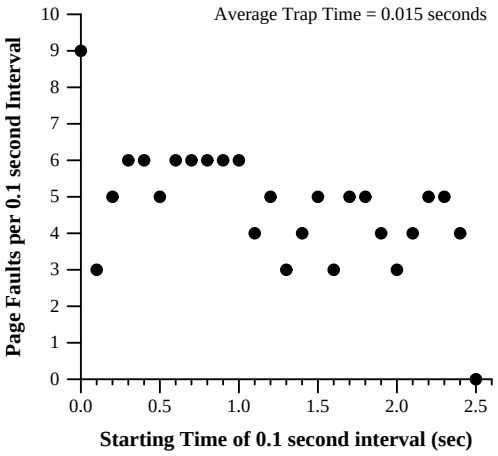
As shown in Graph 2, when the address space approaches the size of physical memory, the main memory and copy-on-write algorithms exhibit severe thrashing, as their memory needs far exceed the size of physical memory. The other two algorithms keep their memory needs below the size of physical memory, and therefore don't suffer such rash penalties. It's worth noting that for all but the smallest address space tested, the pool of pages in the CLL algorithm became completely filled. Therefore, some worst-case data is included in the graphs.

Graphs 3 and 4 show that the two algorithms based on copy-on-write display the smallest overhead, and therefore the greatest concurrency. This is because these algorithms freeze the processors for the smallest amount of time. Taken as a whole, Graphs 1 through 4 show that the CLL algorithm is clearly the best of the four in regard to the combination of checkpoint time and concurrency. The results that follow pertain only to this algorithm.

Graphs 5 and 6 show latency data for the CLL algorithm. The overhead of checkpointing is



Graph 5: Latency Data.



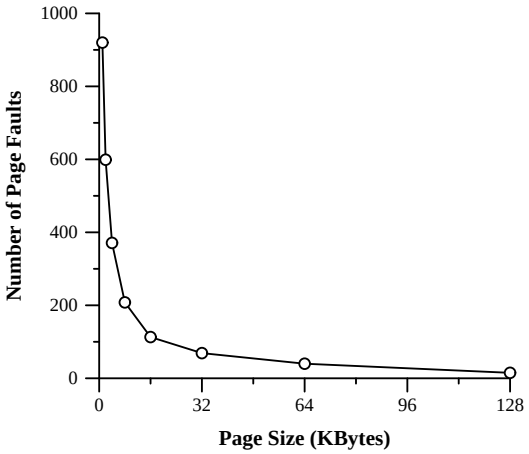
Graph 6: Frequency of Page Faults

divided into two parts – the time that all the threads are stopped initially to protect the address space and save the threads' states, and the time that the target threads are trapped, waiting to process page faults. The first curve in Graph 5 represents the initial stop time as a function of address space size, and the second represents the maximum time that any thread waits as a result of a page fault.

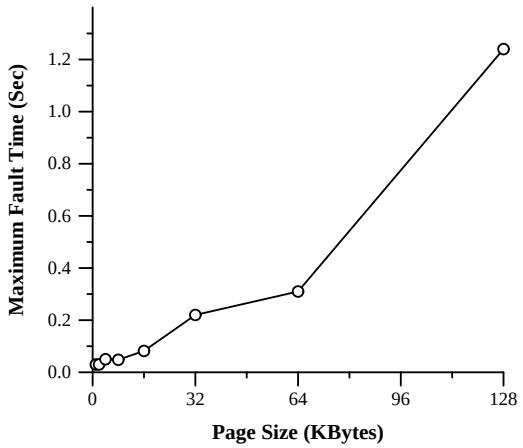
For address spaces up to three Mbytes, the initial stop time is kept below a tenth of a second. Moreover, for all address space sizes, the maximum page fault time is well below our low-latency goal of 0.1 seconds. Graph 6 displays the frequency of page faults over time for a run with a four Mbyte address space. In this graph, the checkpoint is broken into 0.1-second intervals, and the number of page faults in each interval is plotted. The purpose of the graph is to show that work is indeed being accomplished by the target threads during the initial phases of the checkpoint.

Note that after an initial burst of nine page faults, the trapping frequency steadies at six faults per tenth of a second for the first second. Then it slows to about four traps per tenth of a second until there are no more page faults. The average time to process a page fault is 0.015 seconds. Thus, during the first second of the checkpoint, the threads spend around .09 seconds per .1-second interval processing page faults. Since there are three target threads, this means that the threads spend only a third of their time processing page faults in the first second; the rest of the time is devoted to completing the merge. Thus, even at the beginning of the checkpoint, when one expects the highest frequency of page faults, the target still performs an adequate amount of work.

Graphs 7 and 8 display the results of altering the page size, again for a merge sort example with a four Mbyte address space. As would be expected, the total number of page faults is proportional to the inverse of the page size, while the maximum time to process a trap increases almost linearly with the page size. Therefore, the ideal page size is one which significantly decreases the number of page faults, while not significantly increasing the maximum page fault time.



Graph 7: Number of Page Faults vs. Page Size.



Graph 8: Maximum Fault Time vs. Page Size.

This data shows that a page size of sixteen Kbytes is ideal. The number of page faults is kept relatively small (around 120 faults), and the maximum page fault time is still below a tenth of a second. Note that this large page size has one other advantage: If the hardware were built with an actual page size of sixteen Kbytes, then upon protecting the address space, it would only have to change $\frac{1}{32}$ the number of page table entries than it currently has to change. This should reduce the initial stop time (the first curve in Graph 5) by the same factor, which would bring it to well under .1 second for all address space sizes.

We omit the data for the results of checkpointing for the other four benchmark programs (traveling salesman, matrix multiplication, pattern matching, and bubble sort), except to say that their performance in all cases was the same or better than merge sort examples with similar address space sizes.

Related Work

The bulk of work on and implementations of backward error recovery and fault tolerance in parallel and distributed systems has been in database and transaction-processing systems [5, 10, 17, 22, 25]. These schemes benefit from the fact that database computations can be viewed as consisting of atomic transactions. Since we concentrate on general purpose parallel programs, no such computational model can be assumed.

General-purpose checkpointing has been studied and implemented on both uniprocessors and distributed systems. Proposals and overviews for uniprocessor checkpointing have been provided by [1, 15, 23]. In [28] a backward recovery implementation is described which focuses on a file system for Unix which is fully recoverable. [19] describes a portable checkpointing system called "Condor," which runs on many commercial uniprocessors and successfully checkpoints a majority of Unix applications for the purpose of process migration. Neither implementation attempts to provide concurrency or low-latency.

There has been much work on designing checkpointers for distributed systems [1, 3, 11, 12, 13, 14] and for multicomputers [4, 18]. Here the focus of the work is establishing a consistent recovery point – that is, either synchronizing the processors to define a global recovery point, or post-processing the processors' checkpoints to rebuild a plausible recovery point of the system. This is not a problem in shared-memory multiprocessor checkpointing, as the memory bus provides a simple place to enforce processor synchronization.

Staknis proposed a new memory design called sheaved memory [27] for supporting checkpointing in paged systems. In a sheaved memory, physical page frames can be bundled together so that data written to one frame in the bundle is simultaneously written to all frames in the bundle. Removing a frame from its bundle would provide a snapshot of that memory page. Building such a memory would be quite costly and it probably would be used only in special purpose machines.

Of special note is a recent implementation by Elnozahy, Johnson and Zwaenepoel [7]. They implement distributed checkpointing and recovery on a network of sixteen Sun 3/60's with a centralized file server. They implement the sequential and CLL algorithms, as well as both incremental and non-incremental checkpointing. They show that incremental checkpointing can reduce the amount of data being checkpointed by up to 97%. More relevantly to this paper, they corroborate our results by showing that the CLL significantly lowers overhead for incremental checkpointing. They don't show any results concerning the CLL algorithm for non-incremental checkpointing.

The idea of using virtual-memory access protection hardware to achieve synchronization for the concurrent checkpointing was motivated by both shared virtual memory [16] and real-time, concurrent garbage collection [2].

Conclusion

We have presented and implemented a low-latency, concurrent algorithm for checkpointing parallel programs on stock shared-memory multiprocessors. The algorithm requires a constant amount of extra space, no change to the target parallel programs, and no special hardware assistance. Our experiment shows that this algorithm meets our performance goals on all five benchmarks: 80 to 97% of its checkpointing executes concurrently with the target programs; checkpoint time is always within 50% of optimal, and the latency is kept under 0.1 seconds. These goals were met by applying the techniques of copy-on-write [9, 30] and buffering [26] to the checkpointer.

Our algorithms are solely concerned with taking one snapshot with no prior history of the target's execution. For programs with large virtual address spaces, recording the changes between snapshots should be much more efficient than taking each snapshot separately. In the future, our scheme can be combined with [8] to use dirty page information and calculate snapshots incrementally. Such a method would not impose a large initial stop time. Moreover, the checkpointing time will be reduced because pages which haven't been changed since the last snapshot won't be brought into physical memory and written out to disk. Results of distributed checkpointing from [7] support these assertions.

Acknowledgements

We would like to thank Garret Swart for his help with modifying Taos, and Marvin Theimer for his helpful comments.

References

[1] T. Anderson and P. A. Lee. *Fault Tolerance Principles and Practice*. Prentice/Hall International, Inc., Englewood Cliffs, New Jersey, 1981.

[2] A.W. Appel, J.R. Ellis, and K. Li. Real-time concurrent collection on stock multiprocessors. In *ACM SIGPLAN '88 Conference on Programming Language Design and Implementation*, pages 11–20, June 1988.

[3] K. Mani Chandy and Leslie Lamport. Distributed snapshots: Determining global states of distributed systems. *ACM Transactions on Computer Systems*, 3(1):3–75, February 1985.

[4] Flaviu Cristian and Farnam Jahanain. A timestamp-based checkpointing protocol for long-lived distributed computations. In *Proceedings of the 10th Symposium on Reliable Distributed Systems*, pages 12–20, October 1991.

[5] David J. DeWitt, Randy H. Katz, Frank Olken, Lenard D. Shapiro, Michael R. Stonebraker, and David Wood. Implementation techniques for main memory database systems. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1–8, June 1984.

[6] F. Douglis and J. Ousterhout. Process migration in the sprite operating system. In *Proceedings of the 7th International Conference on Distributed Computing Systems*, pages 18–25, September 1987.

[7] Elmootazbellah Nabil Elnozahy, David B. Johnson, and Willy Zwaenepoel. The performance of consistent checkpointing. In *Proceedings of the 11th Symposium on Reliable Distributed Systems*, 1992.

[8] S. Feldman and C. Brown. Igor: A system for program debugging via reversible execution. *ACM SIGPLAN Notices, Workshop on Parallel and Distributed Debugging*, 24(1):112–123, January 1989.

[9] R. Fitzgerald and R.F. Rashid. The integration of virtual memory management and interprocess communication in accent. *ACM Transactions on Computer Systems*, 4(2):147–177, May 1986.

[10] Robert B. Hagmann. A crash recovery scheme for a memory-resident database system. *IEEE Transactions on Computers*, C-35(9):839–843, September 1986.

[11] David B. Johnson and Willy Zwaenepoel. Recovery in distributed systems using optimistic message logging and checkpointing. *Journal of Algorithms*, 11(3):462–491, September 1990.

[12] M. Frans Kaashoek, Raymond Michiels, Henfy E. Bal, and Andrew S. Tanenbaum. Transparent fault-tolerance in parallel Orca programs. Technical Report IR-258, Vrije Universiteit, Amsterdam, October 1991.

[13] Richard Koo and Sam Toueg. Checkpointing and rollback-recovery for distributed systems. *IEEE Transactions on Software Engineering*, SE-13(1):23–31, January 1987.

[14] Ten H. Lai and Tao H. Yang. On distributed snapshots. *Information Processing Letters*, 25:153–158, May 1987.

[15] B.M. Lampson. Atomic transactions. In B.M. Lampson, M. Paul, and H.J. Siegert, editors, *Distributed Systems – Architecture and Implementation*, pages 246–264. Springer-Verlag, 1981.

[16] Kai Li and Paul Hudak. Memory coherence in shared virtual memory systems. *ACM Transactions on Computer Systems*, 7(4):321–359, November 1989.

[17] Kai Li and Jeffrey F. Naughton. Multiprocessor main memory transaction processing. In *Proceedings of the International Symposium on Database in Parallel and Distributed Systems*, pages 177–187, December 1988.

[18] Kai Li, Jeffrey F. Naughton, and James S. Plank. An efficient checkpointing method for multicomputers with wormhole routing. *International Journal of Parallel Processing*, 20(3), June 1992.

[19] Michael Litzkow and Marvin Solomon. Supporting checkpointing and process migration outside the Unix kernel. In *Conference Proceedings, Usenix Winter 1992 Technical Conference*, pages 283–290, January 1992.

[20] P.R. McJones and G.F. Swart. Evolving the unix system interface to support multithreaded programs. Tech Report 21, DEC Systems Research Center, September 1987.

[21] J.K. Ousterhout, A. Cherenson, F. Douglis, M. Nelson, and B. Welch. The sprite network operating system. *IEEE Computer*, 21(2):23–36, February 1988.

[22] Calton Pu. On-the-fly, incremental, consistent reading of entire databases. In *Proceedings of the Eleventh International Conference on Very Large Databases*, pages 369–375, Stockholm, Sweden, August 1985.

[23] B. Randell. System structure for software fault tolerance. *IEEE Transactions on Software Engineering*, SE-1(2):220–232, 1975.

[24] Paul Rovner, Roy Levin, and John Wick. On extending modula-2 for building large, integrated systems. Research report 3, DEC Systems Research Center, 1985.

[25] Kenneth Salem and Hector Garcia-Molina. Checkpointing memory-resident databases. Technical Report CS-TR-126-87, Department of Computer Science, Princeton University, 1987.

[26] Abraham Silberschatz and James L. Peterson. *Operating Systems Concepts*. Addison-Wesley, Reading, Mass., 1988.

[27] Mark E. Staknis. Sheaved memory: Architectural support for state saving and restoration in paged systems. In *Third International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 96–103, April 1989.

[28] David J. Taylor and Michael L. Wright. Backward error recovery in a UNIX environment. In *16th Annual International Symposium on Fault-Tolerant Computing Systems*, pages 118–123. IEEE Computer Society, July 1986.

[29] C.P. Thacker and L.C. Stewart. Firefly: a multiprocessor workstation. In *Proceedings of Second International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 164–172, October 1987.

[30] Marvin M. Theimer, Keith A. Lantz, and David R. Cheriton. Preemptable remote execution facilities for the v-system. In *Proceedings of the Tenth ACM Symposium on Operating Systems Principles*, pages 2–11, Orchas Island, Washington, December 1985.