P.O.Box 8718, Beijing 100080, China	Journal of Software, Feb. 2005,16(2):165-173
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2005 by The Editorial Department of Journal of Software

可恢复的软件DSM系统JIACKPT

张福新, 章隆兵, 胡伟武, 唐志敏

张福新, 章隆兵, 胡伟武, 唐志敏
(中国科学院 计算技术研究所,北京 100080)
作者简介: 张福新(1976－),男,福建永定人,博士生,主要研究领域为机群计算,微处理器系统结构设计和性能评估;章隆兵(1974－),男,博士,主要研究领域为机群计算,微处理器设计;胡伟武(1968－),男,博士,研究员,博士生导师,主要研究领域为高性能计算机体系结构,并行处理,VLSI设计;唐志敏(1966－),男,博士,研究员,博士生导师,主要研究领域为高性能计算机体系结构,CPU芯片,网络并行计算.
联系人: 张福新 Phn: +86-10-62565533 ext 9320, E-mail: fxzhang@ict.ac.cn, http://www.ict.ac.c

Received 20034-07-07; Accepted 2004-10-10

Abstract
Software distributed shared memory (DSM) system has constructed a virtual shared memory abstract on cluster, which combines the programmability of shared memory and fine scalability of cluster. So it is widely studied. Software DSM system is easy to fail because it is a distributed system, some kinds of fault tolerance are necessary for it to be more practical. A recoverable and portable software DSM system, JIACKPT (JIAjia with CheckPoinTing), has been designed and implemented to tolerate the fault of system. JIACKPT, based on JIAJIA, has adopted the checkpointing technology. By maintaining the strict global consistent state and using some optimization techniques, JIACKPT has gotten high performance. The experimental results on an 8-node PC cluster show that the checkpoint overhead is less than 10% of the whole execution time when checkpoint is done once per minute. JIACKPT also has good portability and can run on several operating systems, such as Linux, Solaris, etc. JIACKPT is a practical recoverable software DSM system.

Zhang FX, Zhang LB, Hu WW, Tang ZM. JIACKPT: A recoverable software distributed shared memory system. *Journal of Software*, 2005,16(2):165-173.
http://www.jos.org.cn/1000-9825/16/165.htm

摘要
软件DSM(distributed shared memory)系统在机群上构造了共享存储编程环境,结合了共享存储的易编程性和机群的可扩展性,引起了广泛的研究.由于软件DSM系统是一个分布式系统,系统失败风险大,需要实现容错技术以促进其实用化.利用用户级检查点技术,在支持域存储一致模型的软件DSM系统JIAJIA的基础上,设计并实现了一个可恢复的高可移植的软件DSM系统JIACKPT(JIAjia with CheckPoinTing).由于采用适合软件DSM系统的强全局一致状态以及多种优化措施,JIACKPT易于实现且获得很好的性能.在一个8节点的PC机群上的应用测试表明,即使每分钟做一次检查点,大部分应用的检查点开销也小于10%.此外,JIACKPT还具有高可移植性.这些都表明JIACKPT已经成为一个比较实用的系统..

基金项目: Supported by the National Natural Science Foundation of China under Grant No.60303016(国家自然科学基金)

References:

[1] Bershad BN, Zekauskas MJ, Sawdon WA. The midway distributed shared memory system. In: Proc. of the 38th IEEE Computer Society Int'l Conf. 1993. 528-537. http://www.cs.cmu.edu/afs/cs/project/midway/WWW/CompCon93.ps

[2] Keleher P, Cox AL, Dwarkadas S, Zwaenepoel W. TreadMarks: Distributed shared memory on standard workstations and operating systems. In: Proc. of the 1994 Winter Usenix Conf. 1994. 115-131. http://www.cs.rice.edu/~willy/papers/wusenix94.ps.gz

[3] Keleher PJ. The relative importance of concurrent writers and weak consistency models. In: Proc. of the 16th Int'l Conf. on Distributed Computing Systems. 1996. 91-98. http://x1.cs.umd.edu/papers/writers.pdf

[4] Hu WW, Shi WS, Tang ZM. JIAJIA: A software DSM system based on a new cache coherence protocol. In: Proc. of the HPCN Europe'99. LNCS 1593, Springer-Verlag, 1999. 463-472. http://citeseer.ist.psu.edu/240766.html

[5] Richard GG, Singhal M. Using logging and asynchronous checkpointing to implement recoverable distributed shared memory. In: Proc. of the 12th Symp. on Reliable Distributed Systems. 1993. 58-67. http://citeseer.ist.psu.edu/richard93using.html

[6] Suri G, Janssens B, Fuchs WK. Reduced overhead logging for rollback recovery in distributed shared memory. In: Proc. of the 25th Int'l Symp. on Fault-Tolerant Computing. Washington DC: IEEE computer Society, 1995. 279-288.

[7] Kermarrec AM, Cabillic G, Gefflaut A, Morin C, Puaut I. A recoverable distributed shared memory system integrating coherence and recoverability. In: Proc. of the 25th Int'l Symp. on Fault-Tolerant Computing. Washington DC: IEEE computer Society, 1995. 289-298.

[8] Angkul K, Santipong T, Tzeng NF. Coherence-Based coordinated checkpointing for software distributed shared memory systems. In: Proc. of the 20th Int'l Conf. on Distributed Computing Systems. Washington DC: IEEE computer Society, 2000. 556-563.

[9] Jerzy B, Michal S. An extended coherence protocol for recoverable DSM systems with causal consistency. In: Proc. of the Int'l Conf. on Computational Science. 2004. 475-482. http://www.springerlink.com/index/YVA5RPUQDQQW8QT0.pdf

[10] Plank JS, Beck M, Kingsley G, Li K. Libckpt: Transparent checkpointing under Unix. In: Proc. of the USENIX 1995 Technical Conference. 1995. 213-223. http://www.cs.utk.edu/~plank/plank/papers/usenix-95w.html

[11] Tang ZM, Shi WS, Hu WW. Message-Passing versus shared-memory on dawning 1000A. Chinese Journal of Computers, 2000, 23 (2):134-140 (in Chinese with English abstract).

[12] Hu MC, Shi G, Hu WW, Tang ZM, Zhang FX. Comparing JIAJIA with MPI on PC cluster. Journal of Software, 2003,14(7): 1187-1194 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/1187.htm

[13] Chandy KM, Lamport L. Distributed snapshots: Determining global states of distributed systems. ACM Trans. on Computer Systems, 1985,3(1):63-75.

[14] Woo SC, M. Ohara M, Torrie E, Singh JP, Gupta A. The SPLASH-2 programs: Characterization and methodological considerations. In: Proc. of the 22th Annual Symp. on Computer Architecture. New York : ACM Press, 1995. 24-36.

附中文参考文献:

[11] 唐志敏,施巍松,胡伟武.曙光1000A上消息传递与共享存储的比较.计算机学报,2000,23(2):134-140.

[12] 胡明昌,史岗,胡伟武,唐志敏,张福新.PC机群上JIAJIA与MPI的比较.软件学报,2003,14(7):1187-1194. http://www.jos.org.cn/1000-9825/14/1187.htm