홈기반 분산공유메모리 상에서 결함복구후 향상된 재실행 알고리즘

김용국[0] 하금숙 유은경 이성우 유기영
경 북 대 학 교 컴 퓨 터 공 학 과
{ykkim, swlee}@purple.knu.ac.kr, yook@knu.ac.kr

Advanced reactivation algorithm after recovering on Home-based Distributed Shared Memory

Yong-Kuk Kim[0] Sung-Woo Lee Kee-Young Yoo
Dept. of Computer Engineering, Kyungpook National University

요 약

홈기반의 분산 공유메모리 모델은 지금 현재 가장 적은 외부 통신비용을 가진 프로토콜 모델이다. 본 논문에서는 기존의 Recoverpoint와 Recoverpoint Server를 이용한 결함허용모델(Checkpoint Model)을 보다 향상시키기 위하여 향상된 결함복구후 재실행 알고리즘을 제안한다. 이 알고리즘은 피기백(Piggyback)방식과 복수개의 Checkpoint를 사용하며 기존의 Vector Time Stamp 기법시스템보다 더 낳은 확장성과 실행속도를 제공한다.

1. 서론

병렬형 슈퍼컴퓨터가 빠르게 보급이 되고, 이러한 빠른 보급에는 Cluster형 모델이 많은 기여를 해왔다. 비단 계산용 Cluster가 아니더라도 고 가용성 모델로서의 Cluster 모델도 현재 많은 각광을 받고 있다. 하지만 고가용성 모델이 아닌 계산전용의 고성능형 Cluster(HPC)는 아직도 고가용성 Cluster 정도의 작동신뢰도를 주고 있지 못하다. 현재 이런 고성능 Cluster중 DSM 모델에서의 작동신뢰성을 향상시키기 위해 많은 결함허용 모델이 연구중에 있다. 이 결함허용모델중 본 논문은 Checkpoint와 Checkpoint를 이용한 홈기반 결함허용모델의 기존 재실행 방법중 Vector Time Stamp에 의한 재실행이 아닌 Vector 인덱스에 의한 오류복구후 재실행을 제안하고, 검증하며 전체 시스템에 미치는 영향을 분석한다.

2. 관련연구

홈 프로토콜[4]은 홈을 가짐으로 생기는 장점과 많은 양의 자료구조가 빨리 삭제될 수 있다는 장점이 있다. 차이본 생성과 차이본 적용에 대한 비용을 고려하지 않더라도 차이본들은 프로토콜이 사용하는 메모리 오버헤드의 대부분을 차지한다. 홈 노드가 자신의 페이지를 접근할 때 외부로부터 차이본을 받아 올 필요도 없고 홈의 차이본은 다른 프로세서에게 전송될 필요도 없다. 따라서 외부통신을 발생시키지 않는다. 결과적으로 홈기반의 DSM시스템은 자료의 삭제와 저장측면에서 적은 외부통신요소를 가지므로 외부상황(Network stats)에 보다 강하기(robustness) 때문에 결함허용 시스템을 구성하는데 상당히 유리한 환경이다.

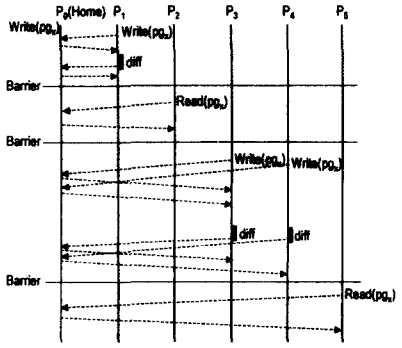

그림.1 홈기반 프로토콜의 작동

Checkpoint기법에 동기식,비동기식,통신유발식의 접근들이 있으며 최근의 결함허용모델은 복합(hybrid)타입의 Checkpoint를 사용하는 추세다.

각 Checkpoint는 안정화된 저장장치에 저장됨을 원칙으로 하며 Checkpoint사용량의 축소를 위해 적층타입의 Checkpoint를 많이 사용하고 있다.적층타입의 Checkpoint 모델은 페이지에서 저장된 Checkpoint 이후의 주소영역만 변화시킴으로 전체적인 데이터통신의 사용량을 감소시킨다. Hybrid타입의 Checkpoint는 동기식과 적층식을 동시에 사용하는데 복귀(rollback)가 자주일어날때만 어느정도의 통신량 감소효과가 있다.

Checkpoint 모델은 시스템이 결함으로부터 복구한 후에는 프로세서의 상태에 Checkpoint를 설정할수 없을 경우도 생길수 있다는 아주 근본적인 문제점이 있다. 이러한 문제점을 해소하기 위해 이전 프로세서 상태의 저장과 정확한 시점에서의 재실행이 필요해 지게 된다.

3.Checkpoint의 구조와 생성

본모델의 Checkpoint 생성시 적층식 Checkpoint와 일반 로컬 Checkpoint의 복수개의 Checkpoint를 사용한 Hybrid 타입의 Checkpoint모델을 사용한다. 이용한 적층식 Checkpoint는 프로세서가 메시지를 보낼 때 자동적으로 증가하며, 일반 로컬 Checkpoint는 프로세서가 메시지를 받을 때 증가한다. Checkpoint의 구조는 프로세서를 P_i라고 하고, 홈이 따로 존재하며 이때 홈프로세서가 P_0이며 메시지보낸횟수를 N으로했을때 [i,N,N_h]로 표시한다. 여기서 Nh는 I번째 프로세서가 홈으로 보낸 총메시지횟수이다.

4.홈기반 LRC상에서 결함허용(Fault-Tolerance) 모델

홈 기반 프로토콜의 결함허용을 위해 본 논문에서는 recoverpoint 서버에 기반한 모델을 제시한다. 여기서 또 recoverpoint는 지역적 recoverpoint와 전역적 recoverpoint로 나뉘는데 이렇게 나누는 것은 단일 오류와 다중오류에 모두 대처할 수 있게 하기 위함이다.

결함허용을 위한 시나리오는 그림2와 같다. Recoverpoint를 이용한 결함허용은 항상 barrier의 동기시점부터 결함에 대한 복구가 진행된다. 모든 프로세서는 barrier에서 동기화 되며, 이 때 각 프로세서는 각자의 페이지 복사본을 갱신한다. 각 프로세서는 홈에 위치한 recoverpoint 서버에게 각자의 recoverpoint를 저장하도록 요청(rec_request)메시지를 보내고 응답(rec_reply)메시지를 받는다. 모든 프로세서가 recoverpoint를 서버로 보내 저장한 후 recoverpoint 서버는 승낙(rec_ack)메시지를 모든 프로세서에게 보내며, 이후 각 프로세서는 자신의 주어진 작업을 계속해 나간다.

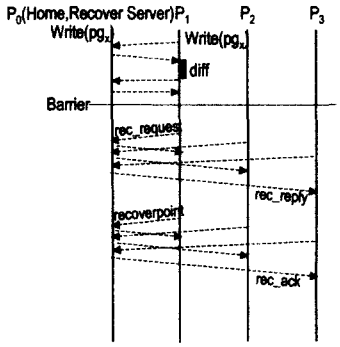

그림.2 recoverpoint의 작동

홈기반 프로토콜은 각각의 공유페이지에 있는 프로그램에 의해 정적으로 프로세서가 할당되므로 각 프로세서들은 독립적으로 recoverpoint를 가질 수 있다. Recoverpoint의 최적화를 위해 새로운 recoverpoint가 생성되면 이전의 recoveroint는 폐기된다. 이 특성은 메모리 사용 효율성을 증가시킨다. 홈기반 프로토콜에 의한 전역적인 불필요정보처리(garbage collection)로 일관된 전역적 recoverpoint의 생성이 가능해진다.이 전역적 recoverpoint는 1개 이상의 다중결함(fault)이 발생했을 때 사용된다.

Recoverpoint의 상태는 동일 차이본(diff)에 의해 복구가 가능하므로 기본적으로 모든 공유페이지들은 recoverpoint에서부터 실행할 수 있다.

Recoverpoint에 의한 단일 프로세서의 복구시 한 프로세서의 오류가 감지되면 오류가 난 프로세서는 자신의 가장 최신의 recoverpoint로부터 복구를 시작하며, 복구중인 이 프로세서는 복구하고 있다는 메시지를 기타 다른 모든 프로세서들에게 알린다. 이 메시지는 프로세서의 현재 시간(T_{recpnt})과 마지막으로 지역적 자료가 생성된 시간(T_{local})으로 구성된다. 모든 다른 프로세서들은 자신의 Vector Time과 T_{local}을 비교하여 자기의 Vector Time이 T_{local}보다 크다면 자신의 Vector Time을 홈으로 보내서 현재 복구중인 프로세서의 차이본과 비교한후 차이본을 재생성하여 전달받고 지역 메모리에 저장한다.복구중인 프로세서는 가장 큰 Vector Time을 가진 프로세서에게 T_{local}보다 큰 모든 interval을 요구한다.홈은 모든 프로세서로부터 T_{local}을 받아서 차이본을 생성하고 차이본과 T_{local}을 해당 프로세서에게 보낸다. 차이본과 T_{local}을 받은 프로세서들은 그 정보를 지역메모리에 저장한다. 홈은 마지막 Recoverpoint 기간부터 오류가 발생한 기간까지의 모든 interval을 리스트화 하여 RecIntList[p]에 저장한다.다중오류가 발생하면 모든 프로세서는 마지막으로 변경된 전역recoverpoint로 복귀(rollback)하며, 전역 recoverpoint 역시 개별recoverpoint와 마찬가지로 새로 생성될때나 마지막 복구작업이 끝난뒤에는 메모리 효율을 위해 폐기된다.

5. 성능평가

본 논문에서 제안한 홈기반 DSM 결함허용시스템의 성능평가를 위해 홈프로토콜기반의 변형된 CVM[1]을 사용하였고, IBM-SP2 슈퍼컴퓨터상에서 4노드에 4개의 응용프로그램을 대상으로 실험하였다.

응용 프로그램	프로세서 갯수	홈기반 DSM (msec)	결함허용홈기반 DSM (msec)
Water	1	10802	10900
	2	5826	5984
	4	3276	3482
Barnes	1	14132	14150
	2	7401	7301
	4	4025	4055
Tsp	1	9420	9435
	2	4752	4915
	4	2605	2754
SOR	1	2950	3017
	2	3274	3152
	4	2072	2702

표.1 실행시간 비교

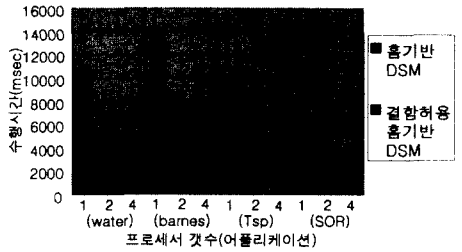

그림.3 수행시간 비교

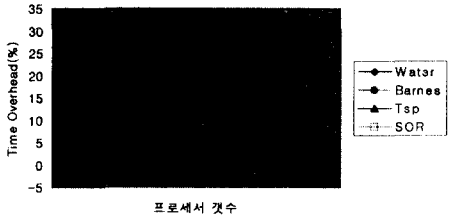

그림.4 Time Overhead 비교

6. 결론

본 논문에서는 홈의 특성에 따라 홈을 recoverpoint 서버로 하는 결함허용 시스템을 제안 하였다. 제안된 시스템은 시스템의 부하없이 단일 프로세서의 결함복구뿐만 아니라 다중 결함복구기능도 가능하다. 본 논문에서는 경쟁상황이 배제된 환경을 가정하였다. 만약 checkpoint를 저장하는 도중에 하나의 프로세서가 오류를 일으킨다면 전체 프로세서들이 recoverpoint로 복귀(rollback)해야 하고 이것으로 인해서 전체적인 성능감소가 있을 수 있다. 이문제는 저장장치의 분리와 RAID화를 통해 해결할 수 있다.

참고 문헌

1. P.Keleher," Distributed Shared Memory Using Lazy Release Consistency" ,PhD dissortation,Rice Univ.1994
2. L.Iftode, " Home-based Shared Virtual Memory" ,PhD thesis, Rice Univ.1998
3. R.Samanta, A.Bilas, L.Iftode, and J.P Singh," Home-based SVM Protocols for SMP clusters Design and Performance" In Proceedings of the 4th IEEE Symposium on High-Performance Computer Architecture, 1998
4. P.Keleher," Symmetry and Performance in consistency Protocols" In the 13rd Int. conference on supercomputing,999
5. M.Stumm, S.Zhou," Fault Tolerent Distributed Shared Memory Algorithms" ,IEEE,1990