Myrinet 상에서 VMMC를 기반으로 하는 효율적인 MPI 구현†

김호중°*, 손영철*, 장영배*, 이문상*, 김명균**, 맹승렬*

*한국과학기술원 전자전산학과 전산학전공

**울산대학교 컴퓨터,정보통신공학부

An Efficient Implementation of MPI over VMMC for Myrinet

Ho-joong Kim°*, Young-Chul Sohn*, Young-Bae Jang*, Moon-Sang Lee*, Myung-Kyun Kim**, Seung Ryoul Maeng*

*Division of Computer Sci., Department of Electric Eng. and Computer Sci., KAIST

**School of Computer Eng. and Information Tech., University of Ulsan

{hjkim, ycsohn, ybjang, mslee, maeng}@camars.kaist.ac.kr* mkkim@uou.ulsan.ac.kr**

요 약

클러스터 시스템의 성능을 향상시키기 위해서는 Myrinet과 같은 고성능 통신망 인터페이스가 필수적이다. 그러나 Myrinet에서 동작하는 저수준 통신 계층들은 각기 고유한 기작을 사용하므로 호환성이 떨어진다. 따라서 MPI와 같은 통신 프로그래밍 표준을 효율적으로 구현하여 응용프로그램 수준에서 고성능과 호환성을 동시에 제공하여야 한다. 본 논문에서는 VMMC 통신 계층을 기반으로, 늦은 위치 갱신, 선택적 무복사 전송 등의 최적화 기법을 적용하여 우수한 성능의 MPI를 구현하였다.

1. 서론

클러스터 시스템의 성능을 높이기 위해서는 통신망의 성능을 높이는 일이 중요하다. 따라서 클러스터 시스템은 Myrinet과 같은 고성능 통신망을 사용하며, 통신망 하드웨어의 성능을 최대한 활용할 수 있는 FM[1], BIP[2], VMMC[3] 등의 효율적인 통신 계층(communication layer)을 채택한다.

그러나 이러한 통신 계층들은 비교적 저수준에서 동작하며 성능을 높이기 위하여 각기 다른 통신 방식과 고유한 기작을 사용하므로 응용프로그램 수준 호환성이 떨어진다. 이 문제를 해결하기 위하여 MPI, socket 등의 통신 프로그래밍 표준을 상위 통신 계층에 구현하는 방법이 주로 사용된다. 이 때 상위 계층의 오버헤드로 인하여 응용프로그램 수준에서 얻을 수 있는 성능이 크게 감소하므로, 응용프로그램 수준에서 더 높은 성능을 얻기 위해서는 고성능의 저수준 통신 계층 상에 통신 프로그래밍 표준을 효율적으로 구현하여야 한다.

본 논문에서 구현한 MPI-VMMC[4]는 VMMC 통신 계층을 기반으로 하는 MPI 구현이다. VMMC는 Myrinet 상에서 동작하는 통신 계층 중 매우 우수한 성능을 보이지만, MPI와 다른 통신 방식을 사용하므로 두 계층 사이의 통신 방식 변환으로 인한 성능 저하가 매우 크다. MPI-VMMC는 송수신 큐 형태의 부계층을 이용하여 MPI의 Send/Recv 통신 방식과 VMMC의 직접 저장(direct deposit)[3] 방식의 메시지를 서로 변환하며, 늦은 위치 갱신(lazy pointer)[5] 기법을 사용하여 큐 관리의 오버헤드를 줄인다. 또한 메시지 크기에 따라 무복사 전송(zero-copy transfer)을 선택적으로 수행함으로써 메시지 전송 성능을 높인다.

†이 연구는 국가지정연구실 사업의 지원을 받는다.

2. 관련연구

2.1 Myrinet

Myrinet은 LAN 규모의 클러스터 시스템을 구성하는 스위치 기반의 고성능 통신망이다. Myrinet 통신망 인터페이스는 호스트와 통신망을 연결하는 역할을 담당하며 RISC 프로세서와 DMA 엔진, SRAM 메모리로 구성된다. 통신망 인터페이스는 사용자가 프로그래밍 할 수 있도록 설계되어 있으므로, 사용자의 요구에 따라 다양한 기작을 구현할 수 있다. 따라서 Myrinet 상에 구현된 FM, BIP, VMMC 등의 통신 계층은 제각기 다른 통신 방식을 사용하며, 다양한 성능을 나타낸다.

2.2 저수준 통신 계층

2.2.1 FM

Illinois 대학의 FM(Fast Messages)은 작은 크기의 메시지를 Send/Recv 방식으로 효율적으로 전송하도록 고안된 통신 계층이다. 그러나 송신자에서 programmed I/O를 사용하여 메시지를 호스트로부터 통신망 인터페이스로 전달하므로 최대 전송 성능이 비교적 낮다.

FM 2.x는 gather/scatter 기능을 지원하므로 MPI 헤더와 같이 작은 메시지를 보내는 데 유리하다. 따라서 FM 상에 구현된 MPI-FM은 FM 통신 계층의 기본 성능에 근접하는 높은 성능을 얻는다.

2.2.2 BIP

LIP-ENS Lyons 대학의 BIP(Basic Instruction-level Parallelism)는 간단한 기능과 고성능을 목표로 개발된 통신 계층이다. BIP는 랑데뷰(rendezvous) 방식으로 통신하여 메시지를 무복사 전송한다. 또한 통신 계층을 간단하게 구현함으로써, Myrinet 상에서 동작하는 통신 계층 중 가장 높은 메시지 전송 성능을 얻는다.

그러나 메모리 보호(protection)를 제공하지 않는 등, BIP의 통신 환경은 상위 통신 계층 구현에 적합하지 않다. 따라서 MPI 구현시 성능 저하가 비교적 크다.

2.2.3 VMMC

Princeton 대학의 VMMC(Virtual Memory-Mapped Communication)는 직접 저장 방식으로 메시지를 전송하는 통신 계층이다. 직접 저장 방식은 송신자가 수신자의 메모리 주소를 지정하여 메시지를 전송하는 방식이다. 수신자 내에서의 메시지 전송은 통신망 인터페이스가 담당하며 호스트 프로세서가 메시지 수신에 관여하지 않는다. 따라서 수신자 호스트에 부하(load)가 걸리지 않는다.

직접 저장 방식은 미리 지정된 수신 주소로만 메시지를 전송하므로 MPI와 같은 Send/Recv에는 적합하지 않다. VMMC-2는 무복사 전송과 Send/Recv 통신 방식을 제공하기 위하여 전송 방향 재지시(transfer redirection)[3] 기작을 사용한다. 그러나 전송 방향 재지시 기작 자체의 부하와 구현 상의 한계로 인하여, 여전히 상위 계층 지원을 위한 오버헤드가 크다.

3. 설계 및 구현

본 논문에서는 VMMC를 기반으로 MPI를 구현한다. 앞서 언급한 바와 같이 VMMC는 상위 계층 지원에 어려움이 있다. 그러나 통신 계층의 기본 성능이 우수하므로, 상위 계층을 효율적으로 설계하여 오버헤드를 줄임으로서 높은 성능의 MPI를 구현할 수 있다.

3.1 송수신 큐의 구조

MPI-VMMC를 구현하기 위하여 VMMC 계층과 MPI 계층 사이에서 메시지 전송 방식을 변환하는 부계층을 설계하였다. 이 부계층은 한 쌍의 송수신 큐의 형태로 존재한다. 그림 1은 MPI-VMMC 송수신 큐의 구조이다. 수신 큐는 수신자의 물리적 메모리(physical memory) 영역에 존재한다. 한편 송신 큐는 송신자의 물리적 메모리 영역 외부에 사상(mapping)된다.

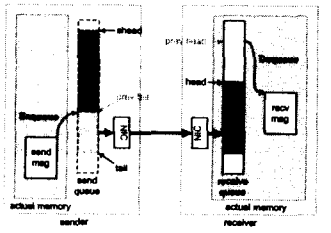

그림 1: MPI-VMMC 송수신 큐

MPI 송신 명령은 송신 큐로의 입력(enqueue) 명령으로 변환되며, 송신 큐의 입력단(tail)이 실제로 가리키는 수신 큐의 메모리 주소로 VMMC 전송 명령이 수행된다. MPI 수신 명령은 수신 큐 출력단(head)에 해당하는 메모리 주소로부터 메시지를 읽어 온다(dequeue).

3.2 성능 향상 기법

3.2.1 큐 관리의 최적화

송수신 큐가 서로 다른 두 노드에 존재하므로 이들의 입력단/출력단 값의 일관성을 유지하여야 한다. 큐의 내용이 바뀔 때마다 상대방 노드의 정보를 갱신하는 방법은 하나의 메시지마다 입출력을 위한 두 개의 관리 메시지를 추가로 발생시키므로 통신망의 성능을 크게 저하시킨다.

본 논문에서는 늦은 위치 갱신(lazy pointer)[5] 기작을 사용하여 큐 관리 메시지의 수를 감소시킨다. 늦은 위치 갱신 기작은 송신자 측에서 출력단의 값을 캐싱하는 방법이다. 송신자는 매 번 메시지 입력시마다 큐가 가득 찼는지 확인하기 위하여 입력단과 출력단의 값을 비교해야 하는데, 출력단의 값은 수신자 노드가 갱신한다. 따라서 매번 출력단 값을 얻기 위하여 수신자 노드에 요청하는 대신 얻어온 출력단 값을 캐싱하고 송신 큐가 가득찰 때만 출력단 값을 갱신함으로서 메시지의 수를 크게 줄일 수 있다. 또한 수신자는 메시지 번호와 태그를 이용하여 메시지의 타당성(validity)을 검사한다. 현재 출력단이 가리키는 메시지가 타당하지 않다면 큐가 비어있는 상태이다.

3.2.2 전송 방향 재지시의 선택적 사용

직접 저장 방식은 수신자의 지정된 기본 버퍼로만 메시지를 전송할 수 있다. VMMC에서 제공하는 전송 방향 재지시 기작은 이러한 한계를 해결하고 최종 수신 영역으로 메시지를 직접 전송할 수 있도록 한다. 그러나 전송 방향재지시 기작은 메시지가 도착하기 전에 사용하여야 성능 향상을 얻을 수 있다. 또한 기작 자체의 부하가 크다. 따라서 작은 메시지의 전송시 기본 버퍼로부터 일 회 복사하는 경우보다 전송 방향 재지시를 사용하여 무복사 전송하는 경우의 성능이 더 낮을 수 있

다. 본 논문에서는 메시지 크기에 따라 전송 방향 재지시를 선택적으로 사용하여 수신 성능을 높인다.

3.2.3 제어 메시지 전송 방식

MPI 헤더는 비교적 작은 크기의 메시지로서 데이타 메시지와 함께 전송되어야 한다. 한편 MPI 헤더는 일반적으로 데이타와 연속되지 않은 메모리 영역에 위치한다. gather/scatter를 지원하는 FM과 달리 VMMC는 하나의 메시지로는 하나의 연속된 메모리 영역만을 보낼 수 있으므로, MPI 헤더를 별도의 메시지로 보내야 한다. VMMC는 작은 메시지의 전송 성능이 낮으므로 이러한 메시지 개수의 증가로 인하여 성능이 크게 감소한다. 따라서 작은 크기의 데이타 메시지는 헤더와 함께 일회 복사하여 하나의 메시지로 포장하여 전송하는 것이 오히려 유리하다.

4. 성능평가

본 논문에서는 두 대의 Pentium III 450MHz 컴퓨터를 LANai 4.3 프로세서를 사용하는 Myrinet 통신망으로 연결하여 구성한 Windows NT 4.0 기반의 클러스터 시스템 상에서 MPI-VMMC의 성능을 측정하였다.

작은 메시지를 전송하는 경우 메시지 크기에 비례하여 전송 대역폭이 증가하므로[4] 늦은 위치 갱신을 사용하여 큐 관리를 최적화함으로써 전송 성능을 크게 개선할 수 있다. 실험 결과 1KBytes 미만의 작은 메시지에서 늦은 위치 갱신을 사용하여 메시지 개수를 $\frac{1}{3}$로 줄였을 때 전송 대역폭이 3배로 증가하였다.

그림 2는 전송 방향 재지시와 제어 메시지 별도 전송의 성능을 측정한 것이다. 전송 방향 재지시를 이용한 무복사 수신(a)과 제어 메시지를 별도로 보내는 무복사 송신(b)은 작은 크기의 메시지에 대해서는 일 회 복사하는 경우보다 오히려 성능이 낮다. 따라서 (c)는 작은 메시지에 대해서 메시지 복사를 수행한다. 한편 큰 메시지의 경우 (a)와 (b)는 각각 송신자와 수신자의 메시지 복사로 인하여 성능이 저하되지만 (c)는 무복사 전송을 하므로 높은 성능을 얻는다.

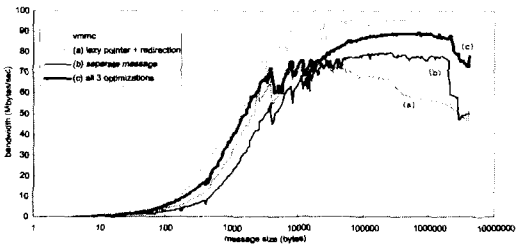

그림 2: 최적화 기작에 의한 성능 향상

그림 3은 MPI-VMMC와 MPI-FM의 성능을 비교한 것이다. 작은 메시지를 전송하는 경우 FM이 VMMC보다 높은 성능을 나타낸다. 또한 MPI-FM은 FM의 *gather/scatter* 기능을 사용하여 *MPI* 헤더를 효율적으로 전송하므로 MPI-VMMC에 비하여 오버헤드가 적다. 따라서 작은 메시지를 전송하는 경우 MPI-FM을 사용하는 것이 유리하다. 그러나 16KBytes 이상의 큰 메시지를 전송하는 경우 VMMC의 최대 전송 대역폭이 *FM*보다 높으며 *MPI* 계층에서 최소한의 오버헤드로 무복사 전송을 하므로, MPI-VMMC가 MPI-FM에 비하여 우수한 성능을 얻는다.

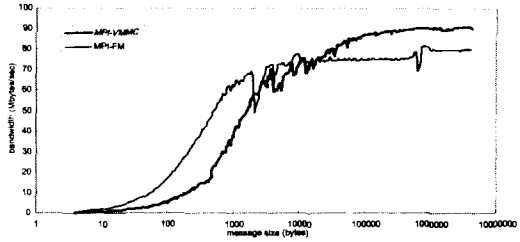

그림 3: MPI-VMMC와 MPI-FM의 성능 비교

MPI-VMMC의 최대 전송 대역폭은 90.7Mbytes/sec로서, VMMC 통신 계층의 최고 성능인 95.7Mbytes/sec의 약 95%에 달한다.

5. 결론

본 논문에서는 VMMC 통신 계층 상에 MPI를 구현하였다. 또한 늦은 위치 갱신 기작을 사용하고 메시지 크기에 따라 선택적으로 무복사 전송을 함으로써 MPI 지원으로 인한 오버헤드를 최소화하였다. MPI-VMMC는 최대 90.7MBytes/sec의 전송 대역폭을 얻으며, 이는 Myrinet 상에서 동작하는 우수한 MPI 구현이다.

참고 문헌

[1] M. Lauria, S. Pakin, and A. A. Chien. Efficient Layering for High Speed Communication: the MPI over Fast Messages (FM) Experience. *Cluster Computing, HPDC7 special issue*, 1999.

[2] L. Prylli and B. Tourancheau. Protocol Design for High Performance Networking: a Myrinet E xperience. Technical Report Research Report 97-22, LIP-ENS Lyons, France, 1997.

[3] C. Dubnicki, A. Bilas, Y. Chen, S. N. Damianakis, and K. Li. *VMMC-2: Efficient Support for Reliable, Connection-Oriented Communication.* In *Hot Interconnects V*, August 1997.

[4] Ho-joong Kim. An Efficient Implementation of MPI over VMMC for Myrinet. Master's thesis, KAIST, 2000.

[5] S. Mukherjee, B.Falsafi, M. D. Hill, and D. A. Wood. Coherent Network Interfaces for Fine-Grain Communication. In *Proc. of the 23rd Int'l Symp. on Computer Architecture*, May 1996.