論文2000-37CI-6-1

비균등 트래픽을 위한 MIN의 설계 및 성능 평가
(Design and Performance Evaluation of MIN for Nonuniform Traffic)

崔 昌 勳*, 金 聖 天**
(Chang-Hoon Choi and Sung-Chun Kim)

요 약

본 논문에서는 클러스터 지향 다단계 상호 연결 망(Cluster Oriented Multistage Interconnection Network)인 COM_R을 소개한다. COM_R은 통신이 빈번하게 발생되는 프로세서-메모리 클러스터에 보다 짧은 경로를 제공하여 지역화 된 통신 형태를 갖는 병렬 응용 분야에 적합하도록 구성할 수 있다. COM_R에 대한 성능 분석은 네트워크에서의 경로 설정 성공 확률(probability of acceptance), 대역폭(bandwidth), 지역 참조성의 변화에 따른 평균 거리(weighted average distance) 및 비용-효율성(cost-effectiveness)에 대해 평가하였다. 성능 평가에 대한 분석 결과에 따르면, COM_R은 지역화의 정도가 높은 통신 형태에서 동일한 네트워크 크기를 갖는 MIN보다 높은 성능을 나타내었다. 최악의 경우(worst case)에서의 $N\times N$ COM_R의 직경(diameter)은 $n+1$로서 이것은 동일한 네트워크 크기의 MIN과 비교했을 때 단지 1개의 스테이지만을 더 가지고 있는 것이다. 따라서 COM_R은 공유 메모리 다중 프로세서 시스템(shared memory multiprocessor system)에서 지역화 된 통신 분포뿐만 아니라 균등 분포 통신를 갖는 병렬 응용 분야에 적합한 MIN으로 활용될 수 있을 것이다.

Abstract

This paper presents a Cluster Oriented Multistage Interconnection Network called COM_R. COM_R can be constructed suitable for the parallel application with localized communication by providing the shortcut path inside the processor-memory cluster which has frequent data communication. We evaluate the performance of COM_R with respect to probability of acceptance, bandwidth, cost-effectiveness and average distance under varying degrees of localized communication. According to the result of analysis for performance evaluation, COM_R shows higher performance than the regular MINs of the same network size in the highly localized communication. In the worst case, the diameter of an $N\times N$ COM_R is only $n+1$ which has only one stage more as compared the MIN with the same network size. Therefore COM_R can be used as an attractive interconnection network for parallel applications with not only the localized communication distribution but also the uniform distribution in shared-memory multiprocessor system.

* 正會員, 尙州大學校 컴퓨터工學部
(Sangju National University, School of Computer Engineering)
** 正會員, 西江大學校 컴퓨터學科
(Sogang University, Dept of Computer Science)
接受日字:1999年5月3日, 수정완료일:2000年11月7日

Ⅰ. 서 론

MIN(Multistage Interconnection Network)은 $\log_2 N$의 낮은 직경(diameter)과 확장성(scalability) 그리고 간단한 분산 자기 제어 라우팅(distributed self-routing)[6]을 사용하기 때문에 다중 프로세서 시스템

(multiprocessor system)에서의 상호연결 망(interconnection network)으로 많이 사용되고 있다.

이러한 MIN의 종류로서 Omega Network[3],[4],[10], Delta Network[3],[4], Baseline Network[3],[11]등을 예로 들 수가 있다. 본 논문에서 이러한 MIN들을 기존의 MIN이라고 언급할 것이다. 기존의 $N\times N$ MIN의 구현은 $\frac{N}{2}\times\log_2 N$개의 스위칭 소자(SE : switching element)로 구성된 상호 연결망이다. 이러한 MIN은 hypercube[3],[10]와 같은 우수한 직경(diameter)과 Crossbar Switch Network보다 적은 하드웨어(hardware)를 사용하고도 완전 접근성(full access property)을 만족할 뿐만 아니라, 간단한 분산 자기 제어 라우팅을 할 수 있어 효율적인 상호 연결 망으로 알려져 현재 많은 다중프로세서 시스템에서 널리 사용되고 있다[4],[6],[8].

그러나 기존 MIN에서의 스테이지 수는 $\log_2 N$이므로 네트워크 크기 N이 커질 경우 스테이지의 수도 계속 증가하게 되어 전체 입출력 쌍간의 통신 경로의 길이 또한 증가하게 된다. 따라서 기존의 MIN에서의 모든 프로세서는 임의의 메모리로의 접근 시간이 모두 동일한 통신 지연 시간을 갖고 있기 때문에 통신의 빈도가 높은 프로세서-메모리 쌍들에 대해서 보다 빠른 통신 경로의 제공이 불가능하다. 다시 말해서 통신이 자주 발생되는 프로세서-메모리 쌍들 간에도 항상 $\log_2 N$의 시간이 소요될 수밖에 없다. 따라서 본 논문에서는 프로세서들간에서 통신 분포의 지역화를 지역 참조성(locality of reference)이라는 표현으로 사용할 것이다. 이러한 지역 참조성의 손실은 시스템 성능을 저하시키는 중요한 요인이 될 것이다.

지역 메모리 참조에 대한 효율성을 활용하기 위하여 일반적으로 자주 사용되는 응용 프로그램에 대한 통신 확률 분포 함수(communication probability distribution function)를 찾을 수 있다면, 프로세서-메모리 쌍들에서의 통신에 대해 지연 시간을 최소로 하는 최대 크기의 프로세서-메모리 모듈 쌍들에 대한 클러스터(cluster)의 크기를 결정할 수 있을 것이다. Abraham과 Davidson[1]은 이러한 통신 분포를 통하여 통신 지연 시간을 최소화시킬 수 있는 클러스터 크기를 결정하는 연구가 진행되어 왔다. 이에 관련된 또 다른 연구로서, 병렬처리 시스템에서 자주 사용되는 프로그램에 대해서 수행시간을 최소화하기 위한 프로세서 그룹을 결정하는 연구가 Agrawal과 Jagadish[1]에 의해 발표되었다.

이렇게 통신이 빈번한 프로세서 그룹내에서의 지역 참조성을 활용할 수 있는 정적 상호연결 네트워크로서 hypercube를 예로 들 수가 있는데, 이는 임의의 두 노드간의 직경이 n으로서 안정적인 거리를 갖고 있을 뿐만 아니라, 지역 참조성이 높은 노드와는 거리가 1인 이웃(direct neighbor)으로 연결되어, 자주 발생되는 통신에 대하여 짧은 지연을 보장하게 하였다. 또한 이진 트리(binary tree)구조에서도 지역 참조성이 높은 노드와는 거리가 1인 이웃노드로 연결될 수 있어, 지역 참조 역시 활용할 수 있다. 이러한 트리 네트워크구조는 병렬처리 데이타 베이스 및 여러 응용 분야에서 많이 사용되고 있다[4],[5].

그러나 동적 상호 연결망인 MIN은 hypercube와 동일한 $\log_2 N$의 직경을 가지고 있지만, 모든 프로세서-메모리 쌍간의 거리는 항상 $\log_2 N$(N×N MIN에서 스테이지수)이고, 네트워크 크기 N이 증가함에 따라 그 거리도 증가하게 된다. 더욱이 MPP 시스템과 같이 많은 프로세서를 사용하는 시스템에서와 같이 네트워크의 크기가 커질 경우 그 지연 시간 또한 그에 비례하여 증가하게 되어 시스템 성능 저하를 초래할 수 있게 된다.

일반적으로 단일 프로세서 시스템 환경에 있어서는 대부분의 메모리 참조는 메모리의 위치상에서 아주 적은 지역에서만 빈번하게 발생된다[1]. 이러한 연구는 캐쉬를 기반으로 한 시스템(cache based system)의 발전을 성공적으로 이루게 되었다. 이와 유사하게 다중 프로세서 시스템 환경 하에서의 많은 대부분의 응용 프로그램에서는 프로세서간의 통신(interprocessor communication)은 주로 프로세서-메모리들의 작은 크기를 갖는 그룹에서 발생하게 된다[1],[2],[4],[9]. 따라서 수많은 프로세서를 갖는 대형 시스템에서 각 프로세서, 메모리 쌍간에 모두 동일한 길이의 연결 경로를 제공하기보다는 통신이 자주 발생되는 작은 그룹에 더 빠른 경로를 제공함으로써 보다 향상된 시스템 성능을 얻을 수 있을 것이다.

기존 MIN에서 발생되는 지역 참조성 활용 문제를 해결하기 위해서 본 논문에서는 클러스터(cluster) 중심의 MIN인 COM_R을 제안한다. COM_R은 통신이 자주 발생되는 프로세서-메모리 쌍들에 대해 보다 짧은 통신 경로를 제공함으로써 지역 참조성을 활용할 수 있

게 하였다. 더욱이 서로 다른 클러스터에 속하는 프로세서와 메모리들간의 통신 경로도 기존의 MIN보다 단지 1개의 스테이지만을 더 통과할 뿐이어서, 지역 참조율이 낮은 통신 분포(uniform distribution) 환경일지라도 평균 직경(diameter)은 기존의 MIN과 차이가 거의 없다.

Ⅱ. COM_R의 구조

$N\times N$ COM_R은 N/R개의 클러스터의 지역 네트워크(local MIN)와 1개의 전역 네트워크(global MIN)로 구성되어 있다. 여기서 N과 R은 각각 $N=2^n$, $R=2^r$이다. 지역 네트워크를 구성하는 각각의 클러스터들은 R개의 입력과 R개의 출력선을 가진다. 이러한 지역 네트워크는 프로세서와 메모리 모듈간의 지역 참조성을 활용할 수 있게 한다. 즉, R은 N보다 작게 유지시키어 통신이 빈번히 발생되는 프로세서들 또는 메모리 모듈들은 이들 동일한 클러스터 내부에 위치시키어 짧은 통신 경로를 제공하여 보다 빠른 통신을 가능하게 할 수 있게 한다. 또한 전역 네트워크는 기존의 MIN과 동일한 형태의 MIN을 그대로 활용하게 되는데 이것은 지역 네트워크(클러스터)들을 연결시키는 역할을 하게 된다. 이러한 전역 네트워크들은 서로 다른 클러스터에 속하는 프로세서와 메모리들간의 통신을 위한 상호 연결망인 것이다. 그림 1은 이러한 COM_R의 전체적인 네트워크 구성을 보인 것이다. 그리고 그림 2는 R의 크기가 4와 8로 클러스터화 할 수 있는 지역 네트워크의 예를 나타낸 것이다. 따라서 $N\times N$ COM_R은 통신 분포가 높은 프로세서들와 메모리들을 클러스터화 하여 구성된 지역 네트워크를 통하여 지역 참조성을 활용할

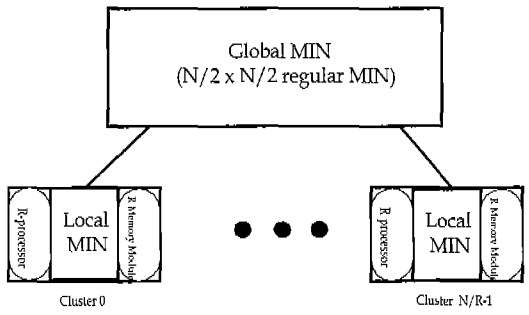

그림 1. $N\times N$ COM_R의 전체적 구성도
Fig. 1. The generic structure of an $N\times N$ COM_R.

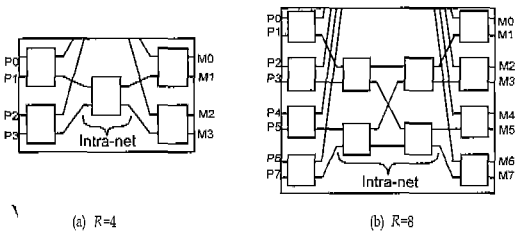

그림 2. 지역 네트워크(클러스터) 구성 예($R=4,8$)
Fig. 2. The examples of local MIN ($R=4,8$).

수 있고, 통신 빈도가 적은 외부 클러스터와의 통신을 위해서는 전역 네트워크를 사용하게 된다. 그러나 이렇게 외부 클러스터와의 통신을 위해 전역 네트워크를 통과할 경우에도 직경이 MIN에서의 직경 보다 단지 한 개의 스테이지(stage)를 더 통과하는 것에 불과할 뿐이어서 비 지역화 된 통신 분포(균등분포)를 갖는 응용 프로그램에서도 COM_R은 MIN의 성능과 비슷한 결과를 기대할 수가 있다.

$N\times N$ COM_R의 지역 네트워크는 $r+1$ 개의 스테이지로 구성된다. 지역 네트워크에 속하는 스테이지는 왼쪽 끝부터 C_stage 0, C_stage 1, ... C_stage r의 순서로 번호를 부여한다(여기서 C는 Cluster Module을 상징함). 이러한 지역 네트워크를 연결하는 전역 네트워크는 $N/2\times N/2$의 기존의 MIN과 동일한 위상(topology), 즉, Omega, Baseline 네트워크 등을 그대로 활용할 수 있게 된다. 전역 네트워크에 속하는 스테이지는 네트워크의 하단에서 상단쪽 방향으로 G_stage 0, G_stage 1, ... , G_stage(n-2)로 번호가 각각 부여된다(여기서 G는 Global Module을 상징함). 그림 3은 R의 크기가 4로 구성한 32×32 COM_4의 예를 보인 것이다.

프로세서 또는 메모리 모듈과 직접 연결된 지역 네트워크의 입력 스테이지(input stage)와 출력 스테이지(output stage)는 각각 $R/2$개의 스위칭 소자(SE : Switching Element)들로 구성되어 있다. 입력 스테이지에 있는 SE들의 상단 출력 포트(upper link port)는 전역 네트워크의 입력 링크 포트(input link port)들과 각각 연결되어 있다. 또한 출력 스테이지에 있는 SE들의 상단 출력 포트(upper link port)는 전역 네트워크의 출력 링크 포트(output link port)들과 각각 연결되어 있다. 한편 입력 스테이지와 출력 스테이지에 있는 SE들의 하단 출력 포트(lower link port)는 지역 네트워크내

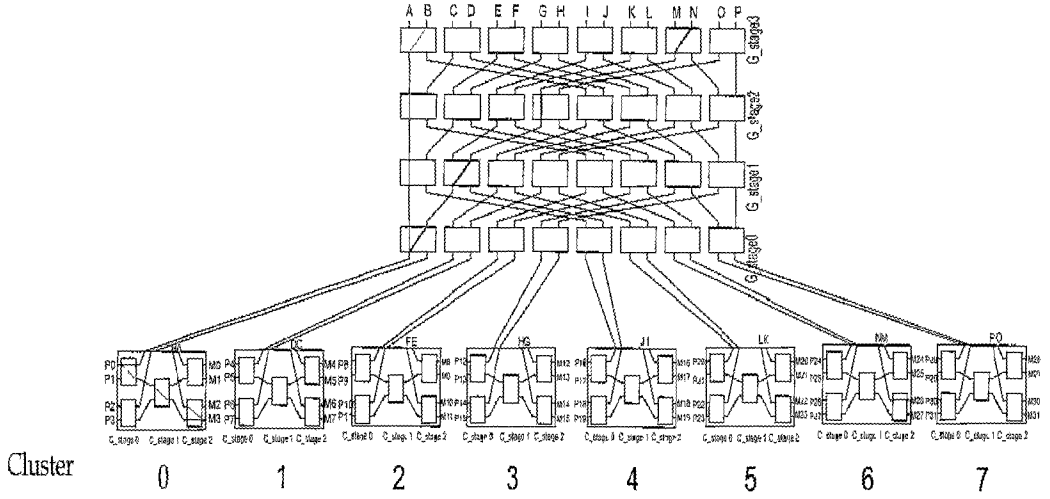

그림 3. 32×32 COM_4구조
Fig. 3. 32×32 COM_4 Structure.

의 SE들 즉, intra-net(그림 2)와 연결되어 있다. intra-net은 $R/2\times R/2$ 크기 MIN을 사용한다. 이렇게 구성함에 따라 그림 3의 32×32 COM_4에서 볼 수 있듯이 클러스터 내에서는 단지 3개의 스테이지만을 통과하게 된다. 따라서 동일 크기를 갖는 기존의 32×32 MIN에서 모든 프로세서 메모리 모듈간에 통신을 위해 통과해야하는 스테이지 수가 5인 것과 비교했을 경우에 COM_4에서는 더 짧은 경로를 제공하게 된다. 더욱이 $N\times N$의 크기가 증가하여도 기존의 MIN은 모든 쌍에 대해서 통과해야만 하는 스테이지 수는 모두 증가되어야 하는 반면, COM_R은 N값의 증가에 관계없이 $R=4$로 유지시켜서 클러스터를 구성한다면, 지역 네트워크에서 통과해야되는 스테이지 수는 3으로 항상 고정시킬 수 있다.

만약 최악의 경우로서 프로세서, 메모리들간의 통신 분포가 비 지역화 되어 다른 클러스터에 속하는 프로세서와 메모리들의 내용을 참조하는 통신 분포를 갖는 병렬 응용 프로그램의 수행을 고려해 보자. 그러나 이러한 최악의 경우일지라도 $N\times N$ COM_R에서의 직경은 $n+1$로서 동일 크기의 기존의 MIN보다 단지 1개의 스테이지를 더 통과하게 된다. 이는 기존의 MIN의 직경 n과 비교한다면 그 차이 매우 적은 것이다.

또 다른 장점으로는 기존의 MIN과 달리, COM_R에서는 클러스터 내부에 존재하는 프로세서-메모리쌍들간에 추가적인 선택 가능 통신 경로가 2개가 존재한다. 즉, 이들 쌍들은 지역네트워크를 통과하여 서로 연결시킬 수 있을 뿐만 아니라, 전역 네트워크에서도 이들간의 연결 또한 가능하다. 따라서 네트워크의 오류 및 트래픽(traffic)을 회피할 수 있는 다른 선택 경로를 가질 수 있게 된다.

III. 라우팅 전략

기존의 MIN에서와 같이 COM_R에서도 목적지 태그 라우팅 알고리즘(destination tag routing algorithm)[7]을 활용할 수 있다. 먼저 근원지(source) S와 목적지(destination) D를 고려하자. $S=s_{n-1}s_{n-2}\ldots s_0$와 $D=d_{n-1}d_{n-2}\ldots d_0$을 각각 근원지 주소와 목적지 주소라고 하자. 입력 스테이지로부터 시작하여, 만약 목적지 클러스터와 근원지 클러스터가 동일하다면 근원지 S와 직접 연결된 입력 스테이지에 있는 스위칭 소자는 입력을 하단 출력 링크로 연결시키며, 만약 목적지 클러스터와 근원지 클러스터가 서로 다를 경우에는 입력 스테이지에 있는 스위칭 소자는 입력을 상단 출력 링크로 연결시킨다. 근원지 S와 목적지 D의 표현은 아래와 같이 표현될 수 있다. 여기서 $s_{n-1}s_{n-2}\ldots s_r$ 은 근원지 클러스터 주소이며 $d_{n-1}d_{n-2}\ldots d_r$는 목적지 클러스터 주소를 나타낸다.

$$S=s_{n-1}s_{n-2}\ldots s_rs_{r-1}\ldots s_0\ ,$$
$$D=d_{n-1}d_{n-2}\ldots d_rd_{r-1}\ldots d_0$$

따라서 라우팅 테그는 아래와 같이 정의된다.

$$\begin{bmatrix} 0d_{n-1}\ldots d_0 & \text{if } s_{n-1}\ldots s_r \neq d_{n-1}\ldots d_r \\ 1d_{r-1}\ldots d_0 & \text{if } s_{n-1}\ldots s_r = d_{n-1}\ldots d_r \end{bmatrix}$$

네트워크에서의 라우팅 방법은 각 스테이지에서 만약 라우팅 태그의 $d_i=0$이면, SE의 입력 링크는 상단 출력 링크로 연결시키고, 또한 $d_i=1$이면 SE의 입력 링크는 하단 출력 링크로 연결시킨다(여기서 $0 \le i \le r \le n-1$). 예를 들어, 그림 3에서 근원지 $P_0(s_4\, s_3\, s_2\, s_1\, s_0 = 00000)$와 목적지 $M_{26}(d_4\, d_3\, d_2\, d_1\, d_0 = 11010)$과의 연결을 위한 라우팅 태그의 생성은 위에서 정의된 라우팅 태그 형식에 의해 $s_4 s_3 s_2 \neq d_4 d_3 d_2$ 이므로 $0d_4d_3d_2d_1d_0(=011010)$가 된다. 그림 3에서 굵은 선은 이들의 경로를 나타낸다. 한편, 또 다른 예로써 $P_0(s_4 s_3 s_2 s_1 s_0 = 00000)$와 $M_3(d_4 d_3 d_2 d_1 d_0 = 00011)$간의 통신 경로를 위한 라우팅 태그는 동일한 클러스터내의 통신이기 때문에 111이 된다. 그림 3에서 점선은 이에 대한 경로를 나타낸 것이다. 이렇게 지역 메모리를 참조할 경우에 보다 짧은 경로를 제공함으로써 시스템 성능을 향상시킬 수 있게 된다. 더욱이 P_0와 M_3는 선택 경로를 한 개 더 가질 수가 있어 통신 라인의 오류가 발생하거나 통신 트래픽이 심할 경우 이를 우회하여 연결할 수 있게 된다. 이때 전역 네트워크를 통과하는 P_0와 M_3에 대한 추가 선택 가능 경로는 $0d_4d_3d_2d_1d_0(=000011)$이 된다.

Ⅳ. 성능 평가

본 장에서는 제안된 $N\times N$ COM_R에 대한 성능을 분석하기로 한다. 일반적으로 기존의 MIN에서의 입출력 연결을 위한 네트워크 지연 시간은 그 네트워크의 스테이지의 수인 n으로 고정되어 있다. 그러나 COM_R에서의 네트워크 지연 시간은 통신의 지역 참조성의 정도에 의존한다. 즉, 어느 병렬 응용 프로그램의 수행에 대한 통신 분포가 지역 네트워크에서 많이 발생한다면, 네트워크 지연 시간은 매우 짧을 것이며, 반대로 병렬 응용 프로그램의 수행에 대한 통신 분포가 균등 분포이거나 클러스터내의 통신이 거의 발생하지 않고 외부 클러스터와 통신이 빈번할 경우에는 네트워크 지연시간은 이보다 길어질 수 있을 것이다.

따라서 $N\times N$ COM_R에서 최악의 경우의 거리(직경)를 생각해 보자. 임의의 한 프로세서가 다른 클러스터에 있는 메모리 모듈에 접근하려면 전역 네트워크를 통과해야 하므로 $N\times N$ COM_R의 전역 네트워크는 $N/2\times N/2$ 의 MIN이므로 총 스테이지 수는 $n-1$이 된다. 따라서 지역 네트워크에 입력 스테이지 한 개와 출력 스테이지 한 개를 추가로 통과해야 되므로 총 $n+1$개의 스테이지를 통과해야 한다. 이것은 $N\times N$ COM_R에서 최악의 경우에 대한 직경을 나타낸다. $N\times N$ COM_R에서 최선의 경우(best case)를 고려해 보자. 이러한 경우는 수행되는 응용 프로그램의 통신 분포가 클러스터 내부에 거의 밀집되어진 상태 즉, 지역 참조성 비율이 매우 높게 발생할 때를 말한다. 이때의 통신은 지역 네트워크만을 이용할 수 있기 때문에 지역 네트워크의 스테이지 수 $r+1$이 된다. 여기서 r의 크기는 네트워크의 크기에 따라 조정이 가능할 뿐만 아니라, 네트워크 크기 N이 증가에 관계없이 고정된 거리로서 유지시킬 수도 있다.

COM_R과 동일한 크기의 $N\times N$ MIN에서의 모든 프로세서는 임의의 메모리로의 접근할 수 있는 거리가 n으로 동일한 거리를 갖고 있기 때문에 지역 참조성의 활용이 불가능하게 되고, 또한 스테이지 수도 네트워크 크기 N이 증가할 경우 스테이지의 수도 역시 증가하게 되어 전체 입출력 쌍간의 통신 경로의 길이 또한 증가하게 된다. 이와 비교한다면 제안된 COM_R은 최악의 경우에도 기존의 MIN보다 거리가 단지 1만이 더 추가되었을 뿐이고, 최선의 경우에 있어서는 기존의 MIN에서는 불가능했던 통신의 발생이 빈번한 프로세서-메모리쌍들에게 빠른 경로를 제공할 수 있게 하는 지역 참조성을 활용할 수 있는 장점을 가지게 된다.

다음은 하드웨어 비용(hardware cost)에 대한 분석을 고려해 보자. 기존의 MIN과 제안된 COM_R은 2×2의 동일한 크기의 스위칭 소자의 사용을 가정하였기 때문에, 여기서 하드웨어 비용은 $N\times N$ COM_R에서 사용되는 스위칭 소자(2×2 SE)의 갯수로서 평가되었다. 이는 $N/2\times N/2$크기의 MIN를 구성하는 스위칭 소자의 갯수와 N/R개의 지역 네트워크에서 사용되는 스위칭 소자 갯수에 대한 합으로써 산출될 수 있으며 이에 대한 결과 식은 아래와 같다.

$$\frac{N}{4}\log_2\frac{N}{2}+(R+\frac{R}{4}\log_2\frac{R}{2})\frac{N}{R}=N(n+r+2)/4 \quad (1)$$

예를 들어 $R=4$인 32×32 COM_4에서는 위와 같은 식을 이용하면 72(=32(5+2+2)/4)개의 SE들이 사용됨을 알 수 있다.

다음은 $N\times N$ COM_R의 평균 거리를 분석에 대한 분석이다. 위에서도 언급했듯이 COM_R에서의 네트워크 지연 시간은 통신의 지역 참조성의 정도에 의존하게 된다. 그러나 COM_R에 대한 최악의 경우에 대해 기존의 MIN과 비교 평가 또한 필요하다. 따라서 균등 분포(uniform distribution)하에서의 COM_R의 평균 거리를 아래의 식에서와 같이 산출된다.

$$Avg.\ distance=n+1-\frac{R(n-r)}{N} \quad (2)$$

이번에는 통신의 지역 참조성 비율(α)을 고려한 가중 평균 거리(weighted avgerage distance)에 대한 식을 보이도록 한다. 여기서 α는 각 프로세서들이 동일한 클러스터에 속하는 메모리 모듈들을 참조할 확률이다(따라서 각 클러스터 외부에 있는 메모리를 참조할 확률은 $(1-\alpha)$가 됨). 이것은 클러스터 내부에 부여되는 참조성의 정도 즉, α를 변화시키어 COM_R에 대한 가중 평균 거리를 구하는 것이다.

$$Weighted\ Avg.\ Distance=(r+1)\alpha+(n+1)(1-\alpha)$$
$$where\ 0\le\alpha\le1 \quad (3)$$

위에서 언급된 COM_R에 대해 분석한 결과들을 기존의 MIN과의 비교하여 요약한 결과는 표 1과 같이 나타낼 수 있다.

표 1. *MIN* 과 COM_R의 성능 파라메터 비교
Table 1. The comparison of performance parameter between *MIN* and COM_R.

	$N\times N$ COM_R	$N\times N$ *MIN*
Network Size	$2^n\times2^n$	$2^n\times2^n$
Cluster Size	$2^r(=R)$ $(R<N)$	cluster구성 불가능
Diameter	$n+1$	n
Uniform Ave. Distance	$n+1-\frac{R(n-r)}{N}$	n
Weighted Ave. Distance	$(r+1)\alpha+(n+1)(1-\alpha)$	n
Hardware Cost	$N(n+r+2)/4$	$nN/2$

또 다른 COM_R에 대한 성능 평가로서 대역폭(bandwidth), 접근 성공 확률(probability of acceptance) 및 가격-효율성(cost-effectivencss)을 아래 조건[3]을 기초로 한 확률적 분석(probabilistic analysis) 방법을 통하여 수행하였다.

- 각 프로세서는 메모리 모듈로의 접근을 무작위적(randomly)이고 독립적 (independently)으로 요청한다.
- 각 프로세서는 확률 m을 가지고 새로운 요청을 발생시킨다.(따라서 m은 각 시스템 사이클(cycle)당 각 프로세서에 의해 발생되는 평균 메모리 접근 요청의 횟수를 의미한다.)
- 거부된 요청은 무시된다.
- 네트워크의 거리는 네트워크를 통과하는 스테이지 수로 정의한다.

통신 분포가 균등 분포를 따르는 환경하에서 전역 네트워크는 기존의 MIN과 동일하므로 이 네트워크에서 사용되는 2×2 스위칭 소자의한 출력 포트에 대한 요청 비율은 아래와 같이 나타낼 수 있다[8].

$$m_{out}=1-(1-m_{in}/2)^2 \quad (4)$$

여기서 m_{in}은 한 스위칭 소자의 입력 포트에 대한 요청 비율을 나타낸다. 그러나 균등 분포하에서 COM_R의 입력 스테이지에 있는 스위칭 소자들은 위와 같은 요청 비율을 따르지 않는다. COM_R의 입력 스테이지에 있는 2×2 스위칭 소자의 입력 포트들에 대한 요청 비율 m이 1이라고 했을 때, 단위 시간당 스위칭 소자의 상단 출력 포트(m_0^{UP})과 하단 출력 포트(m_0^{DOWN})을 통과하는 요청의 비율은 아래의 식과 같이 나타난다.

$$m_0^{UP}=2((1-R/N)R/N)+(1-R/N)^2/2 \quad (5)$$
$$m_0^{DOWN}=2((1-R/N)R/N)+(R/N)^2/2 \quad (6)$$

C_{m_i}와 G_{m_j}를 각각 C_stage에 있는 스테이지 i ($0\le i\le r$)와 G_stage에 있는 스테이지 j ($0\le j\le n-2$)의 출력 포트에 대한 요청 비율이라고 정의하자. $N\times N$ COM_R에 대한 대역폭(BW)은 Patel[8]의 2×2 crossbar switch의 성능 분석과 유사한 방법으로서 아래 식과 같이 나타낼 수 있다.

$$BW=2^nC_{m_r} \quad (7)$$

여기서

$$C_{m_r}=1-(1-\frac{\max[C_{m_{r-1}},G_{m_{n-1}}]}{2})^2$$

$$C_{m_i}=1-(1-\frac{C_{m_0^{DOWN}}}{2})^2 \text{ and } C_{m_0^{DOWN}}=m_0^{DOWN}$$

$$G_{m_r}=1-(1-\frac{G_{m_{i-1}}}{2})^2 \text{ and } G_{m_0}=1-(1-\frac{C_{m_0}}{2})^2$$

and $C_{m_0^{UP}}=m_0^{UP}$ for $0\le i\le n-1$

또한 병렬 응용 프로그램에서 프로세서 및 메모리 모듈들 간의 지역 참조 비율(degree of locality) 인자 $\alpha(0\le\alpha\le1)$를 고려한 $N\times N$ COM_R에 대한 BW는 아래와 같은 식으로 결정된다.

$$BW=2^nC_{m_r} \tag{8}$$

여기서

$$C_{m_r}=1-(1-\frac{\max[C_{m_{r-1}},G_{m_{n-1}}]}{2})^2$$

$$C_{m_i}=1-(1-\frac{C_{m_0^{DOWN}}}{2})^2 \text{ and } C_{m_0^{DOWN}}=2\alpha(1-\alpha)+\alpha^2/2$$

$$G_{m_i}=1-(1-\frac{G_{m_{i-1}}}{2})^2 \text{ and } G_{m_0}=1-(1-\frac{C_{m_0}}{2})^2$$

and $C_{m_0^{UP}}=2\alpha(1-\alpha)+(1-\alpha)^2/2$ for $0\le i\le n-1$

따라서 한 프로세서에서의 임의 메모리 모듈로의 연결 요청이 받아들여질 확률(PA : Probability of Acceptance)[8]은 아래와 같은 식을 통하여 얻을 수 있다.

$$PA=\frac{BW}{2^nm} \tag{9}$$

마지막으로 가격-효율성 비율(cost-effectiveness ratio)은 하드웨어 가격에 대한 대역폭으로 산출하였다. 또한 하드웨어 가격은 $N\times N$ 네트워크에서 사용되는 스위칭 소자의 갯수로서 정의하였다.

그림 4는 기존의 MIN과 $R=4$인 COM_4와 $R=8$인 COM_8에서의 평균 거리를 비교한 것이다. 지역 참조성 $\alpha=0$ 일 때는 기존의 MIN의 평균 거리가 더 짧게 나타났지만 $\alpha=0.2$ 이상에서 부터는 COM_R이 더 우수하게 나타나고 있다. 또한 $\alpha=1.0$에 근접할수록 COM_R의 평균 거리는 네트워크의 크기에는 크게 영향을 받지 않고 R값에 따라 거의 일정함을 보이고 있다.

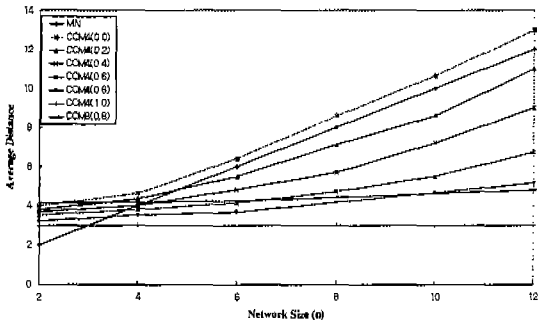

그림 4. 지역 참조율 α의 변화 따른 평균 거리의 변화
Fig. 4. Variation of average distance for the degree of locality of reference α.

그림 5는 $R=4$인 COM_4와 $R=8$인 COM_8에서의 지역 참조율 α값을 변화시키면서 Crossbar Switch 네트워크와 기존의 MIN들간의 대역폭을 서로 비교한 것이다. 물론 Crossbar Switch 네트워크가 단연 높은 성능을 나타내고 있다. 그러나 기존의 MIN과 COM_R을 비교했을 때는 $\alpha=0.2$ 이상에서 부터는 COM_R이 매우 우수한 성능을 보이고 있다. 또한 $\alpha=1.0$일 때는 기존의 MIN과 비교했을 때 n=12에서 최대 2.2배의 성능향상을 나타내고 있다.

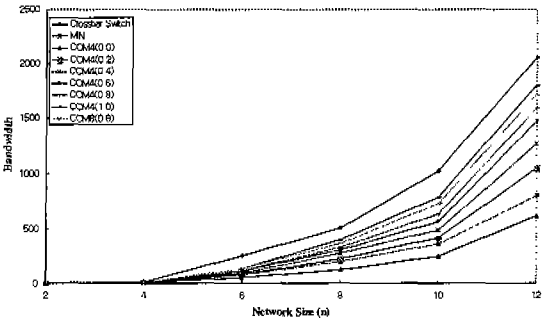

그림 5. 지역 참조율 α의 변화에 따른 대역폭의 변화
Fig. 5. Variation of bandwidth with the degree of locality of reference α.

그림 6은 네트워크 크기의 증가에 따른 네트워크들의 비용-효율성을 측정한 것이다. 제안된 COM_R ($R=4,8$)은 전체 네트워크의 크기가 작을 경우(n=4 미만)에는 기존의 MIN 보다, 비용-효율성 면이 다소 떨어지는 것을 볼 수 있으나 네트워크의 크기가 커짐에 따라서 COM_R은 기존의 MIN 뿐만 아니라 Crossbar Switch 네트워크보다도 비용-효율성이 우수하게 나타나고 있다.

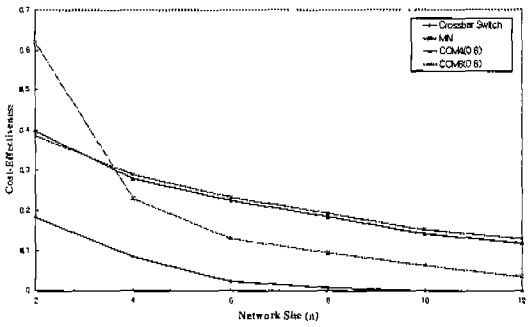

그림 6. $N\times N$ 네트워크들의 가격-효율성
Fig. 6. Cost-effectiveness for $N\times N$ networks.

그림 7은 입력 요청률 즉, m=1일때 지역 참조율 α와 네트워크 크기를 각각 변화시키면서 네트워크의 연결 요청에 대한 성공 확률(PA : Probability of Acceptance)를 분석한 것이다. 여기에서도 볼 수 있듯이 네트워크 크기가 커질수록, 또한 지역 참조율 α가 커질수록 PA은 기존의 MIN보다 최대 1.6배의 성능 향상을 보이고 있다.

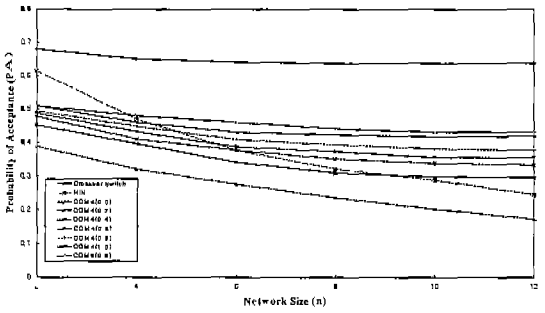

그림 7. 지역 참조율 α의 변화에 따른 PA의 변화
Fig. 7. Variation of PA with the degree of locality of reference α.

V. 결 론

본 논문에서는 공유 메모리 다중 프로세서 시스템에서 균등 분포뿐만 아니라 지역화 된 통신 분포를 갖는 병렬 응용 프로그램을 효율적으로 수행시킬 수 있는 상호 연결망으로서 COM_R을 소개하였다. COM_R은 통신 분포가 높은 프로세서들와 메모리들을 클러스터화 하여 구성된 지역 네트워크를 통하여 지역 참조성을 활용할 수 있고, 통신 빈도가 적은 외부 클러스터와의 통신을 위해서는 전역 네트워크를 사용하게 된다. 특히, COM_R은 통신이 빈번하게 발생되는 프로세서-메모리 클러스터에 보다 짧은 경로를 제공하여 지역화 된 통신 형태를 갖는 병렬 응용 분야에 적합하도록 설계되었다. 더욱이 COM_R 다른 외부 클러스터와의 통신을 위해 전역 네트워크를 통과할 경우에도 직경이 기존의 MIN에서 보다 단지 한 개의 스테이지만을 더 통과하는 것에 불과하기 때문에 비 지역화 된 통신 분포(균등 분포)를 갖는 응용 프로그램에서도 COM_R은 MIN의 성능과 거의 차이가 나지 않는 결과를 얻을 수 있었다.

성능 분석에서도 나타났듯이 COM_R은 네트워크의 크기가 커질 수록, 또한 α값이 커질수록 기존의 MIN보다 대역폭, PA, 평균 거리, 비용-효율성면 등에서 모두 성능이 매우 우수하게 나타났다. 따라서 COM_R은 공유 메모리 다중 프로세서 시스템에서 통신 분포뿐만 아니라 지역화 된 통신 분포를 갖는 병렬 응용 분야에 적합한 MIN으로 활용될 수 있을 것으로 기대된다.

참 고 문 헌

[1] S. G. Abraham, and E. S. Davidson, "A Communication Model for Optimizing Hierarchical Multiprocessor System," In Proc. Int'l Conf. on Parallel Proc., pp.467-474, 1986.

[2] R. Agrawal and H.V. Jagadish, "Partitioning Techniques for Large-Grained Parallelism," IEEE Trans. Compt., vol. C-37, pp. 1627-1634, Dec. 1988.

[3] Y. Chang and L. N. Bhuyan, "Extending Multistage Interconnection Networks for Multitasking," In Porc. Int. Conf. on Parallel Proc., vol. I, pp.151-158, 1992.

[4] A. L. Decegama, The Technology of Parallel Processing, Parallel Processing Architectures and VLSI hardware volume I, Prentice-Hall Press, 1989.

[5] J. R. Goodman and C.H. Sequin, "Hypertree: A multiprocessor Interconnection Topology," IEEE Trans. Compt., vol. C-30, pp.923-933,. Dec., 1981.

[6] K. Hwang, Advanced Computer Architecture : Parallelism Scalability Programmability, McGraw-Hill International Edition, 1993.

[7] D. H. Lawrie, "Access and Alignment of Data in a Array Processor," IEEE Trans. Computer, pp.1145-1155, Dec. 1975.

[8] J. H. Patel, "Performance of Processor-Memory Interconnections for Multiprocessors," IEEE Trans. Computer, pp.771-780, Oct. 1981.

[9] D. A. Patterson and J. L. Hennessy, Computer Architecture A Quantative Approach, Morgan Kaufmann Pub., 1996.

[10] H. J. Siegel and R. J. McMillen, "The Multistage Cube: A Versatile Interconnection Network," IEEE Computer, pp.65-76, Dec. 1981.

[11] C. L. Wu and T. Y. Feng, "On a Class of Multistage Interconnection Networks," IEEE Trans. Computer, pp.696-702, Aug. 1980.

저 자 소 개

崔 昌 勳(正會員)

1965년 10월 28일생, 1988년 2월 명지대학교 전자계산학과(학사), 1990년 2월 서강대학교 대학원 전자계산학과(공학석사), 1997년 8월 서강대학교 대학원 전자계산학과(공학박사), 1990년 1월~1990년 9월 대우통신(주) 기술개발부, 1995년 10월~1996년 2월 미국 AT&T (NCR) 기술파견 연구원, 1997년 9월~현재 국립상주대학교 컴퓨터공학부 전임강사/조교수, <주관심 분야 : 컴퓨터구조, 병렬처리시스템, 컴퓨터시뮬레이션>

金 聖 天(正會員) 第33卷 B編 第10號 參照