論文95-32B-5-7

대규모 병렬컴퓨터를 위한 교차메쉬구조 및 그의 성능해석

(Performance Analysis of the XMESH Topology for the Massively Parallel Computer Architecture)

金 鍾 晋 *, 崔 興 文 **

(Jong-Jin Kim, and Heung-Moon Choi)

요 약

대규모 병렬컴퓨터 아키텍쳐의 구현에 적합한 상호접속망으로서 교차메쉬(XMESH)구조를 제안하고 그 성능을 해석하였다. 제안된 구조에서는 토로이달메쉬(또는 토로이드)의 수평링크는 그대로 두고 수직링크대신 대각선방향 교차연결을 도입하여 기존의 토로이드가 대규모 병렬컴퓨터의 상호접속망을 집적구현하는데 있어서 갖는 장점 즉, 대칭구조, 단위 부시스템(subsystem)의 주기적 반복배열, 일정접속도 등의 장점을 그대로 가지면서도 우수한 확장성을 갖게 하였다. 성능해석에 의하여 비교한 결과, 제안된 구조는 토로이드 혹은 대각메쉬(diagonal mesh)에 비해 노드간 최대거리 및 평균노드간거리가 짧으며 메시지 처리율의 상한이 크게 나타나 성능이 더 우수함을 확인하였다. 이 결과들을 입증하기 위해 교차메쉬를 위한 최적 자기경로 알고리듬을 개발하여 각 구조의 평균지연, 최대지연 및 메시지처리율에 대해 우회경로알고리듬을 사용하여 시뮬레이션을 수행한 결과, 통신부하량의 대소나 시스템의 노드수와 관계없이 교차메쉬는 모든 측면에서 토로이드 및 대각메쉬에 비해 성능이 우수함을 확인할 수 있었다.

Abstract

We proposed a XMESH(crossed-mesh) topology as a suitable interconnection for the massively parallel computer architectures, and presented performance analysis of the proposed interconnection topology. Horizontally, the XMESH has the same links as those of the toroidal mesh(TMESH) or toroid, but vertically, it has diagonal cross links instead of the vertical links. It reveals desirable interconnection characteristics for the massively parallel computers as the number of nodes increases, while retaining the same structural advantages of the TMESH such as the symmetric structure, periodic placement of subsystems, and constant degree, which are highly recommended features for VLSI/WSI implementations. Furthermore, $n \times k$ XMESH can be easily expanded without increasing the diameter as long as $n \le k \le n+4$. Analytical performance evaluations show that the XMESH has a shorter diameter, a shorter mean internode distance, and a higher message completion rate than the TMESH or the diagonal mesh(DMESH). To confirm these results, an optimal self-routing algorithm for the proposed topology is developed and is used to simulate the average delay, the maximum delay, and the throughput in the presence of contention. In all cases, the XMESH is shown to outperform the TMESH and the DMESH regardless of the communication load conditions or the number of nodes of the networks, and can provide an attractive alternative to those networks in implementing massively parallel computers.

* 正會員, 釜山工業大學校 電子工學科
(Dept. of Elec., Pusan Nat'l Univ. of Technology)

** 正會員, 慶北大學校 工科大學 電子工學科
(Dept. of Elec., College of Eng., Kyungpook Nat'l Univ.)

接受日字: 1994年8月10日, 수정완료일: 1995年4月 29日

I. 서 론

수천 혹은 수만 개의 노드프로세서로 구성되는 고성능의 대규모 병렬컴퓨터시스템을 구현하기 위해서는 효과적인 상호연결망을 선택하는 것이 필수적이다. 따라서 병렬컴퓨터시스템에서 노드(node)상호간에 어떻게 통신하고 어떤 종류의 연결망으로 연결할 것인가 하는 상호연결망 위상구조(interconnection network topology)에 관한 연구가 활발히 진행되고 있다.

최근 반도체 기술의 발달로 고집적 병렬컴퓨터의 제조가 기술적 및 경제적으로 가능하게 되었다. 대규모 병렬컴퓨터시스템을 집적화 함에 있어서는 전체 칩(chip) 면적을 줄이는 것이 필요하며, 이를 위해서는 링크의 길이(length of link)는 짧아야하고 링크접속도가 적고 일정한 것이 바람직하다. 각 노드 혹은 단위 부시스템의 구조가 동일한 형태로 반복 배치될 경우에는 배치작업비용(layout cost)을 줄일 수 있다.[1]

현재 시판중인 많은 병렬 컴퓨터들이 하이퍼큐브 구조로 개발되었다.[2] 하이퍼큐브의 강한 연결도(strong connectivity), 정규성(regularity) 및 대칭성(symmetry)으로 인해 여러 목적의 응용에 탁월한 성능을 발휘하지만 망의 규모가 커짐에 따라 접속도의 대수적인 증가로 말미암아, 하이퍼큐브구조는 VLSI구현에는 적합하지 않은 것으로 알려져 있다. 이와 같이 대규모 병렬컴퓨터를 칩에 구현하는 VLSI 또는 WSI(wafer scale integration)에서는 인접 노드끼리 연결되는 메쉬구조가 적합하다고 알려져 있다.[11] 토로이달메쉬(toroidal mesh:TMESH) 즉 토로이드(toroid)는 대칭구조를 가지며, 부시스템이 주기적으로 반복 배치되며, 최적자기경로설정(optimal self-routing) 알고리듬을 가지므로 널리 애용되고 있다. 그러나, 노드간 최대거리 및 평균거리가 비교적 크다는 단점을 갖고 있다. Vecchia등[3]은 컴퓨터망의 확장성이 뛰어난 WK-순환위상구조를 제안하였다. 그러나 이 구조는 부시스템간의 연결링크의 수가 적어 이 링크에서 병목현상이 심각하게 되므로 국부성이 크고 노드간의 통신요구량(traffic quantity)이 적은 문제들에 제한적용이 가능하다. Yang등[4]은 기존의 토로이드 구조에 대각선방향의 원격링크를 반복적으로 추가함으로써 노드간거리가 매우 작은 RDT(recursive diagonal torus)를 제안하고 최적에 가까운(near-optimal) 경로설정 알고리듬을 개발하였다. 그러나 이 구조는 접속도가 8이나 되며 멀리 떨어진 노드까지 접속하는 링크의 길이가 너무 길고 그 수도 많아 신호의 지연이 커지고 넓은 칩 면적이 요구되어 칩생산성(chip yield)이 낮아지므로 VLSI 혹은 WSI구현에는 적합하지 않을 것으로 생각된다. Zohorjan등[5]은 균등메시지분포를 가지며 일정한 수의 메시지를 가진 시스템에서 메시지처리율의 상한을 구하는 방법을 제안하였으며 Reed등[6]은 균등메시지분포의 경우보다 더 현실적인 가정하에서 시스템의 성능을 분석하기 위해 메시지 routing분포라는 개념을 도입하였다. Arden[7]은 n×k토로이드의 수평 및 수직방향의 링크 대신 대각선방향으로 링크를 접속하여 토로이드 보다 개선된 대각메쉬(diagonal mesh:DMESH)구조를 제안하였으며 Tang등[8]은 이 구조가 토로이드에 비해 노드간 최대거리(diameter)가 작으며 이분폭(bisection width)이 큼을 증명하고 Barnes[9]가 제안한 우회경로(deflection routing)알고리듬을 이용하여 대각메쉬가 토로이드에 비해 메시지처리율이 우수함을 보였다.

본 논문에서는 대규모 병렬컴퓨터 아키텍처에 적합한 새로운 상호접속망으로서 교차메쉬(XMESH: crossed-mesh)구조를 제안하였다. 제안된 구조는 토로이드의 수직방향으로 연결된 링크들을 대각선방향으로 교차연결한 것으로서, 토로이드가 병렬컴퓨터망을 집적구현하는데 있어서 갖는 장점들 즉, 대칭구조, 부시스템의 주기적 반복배열, 일정수의 접속도 등의 장점을 그대로 가지면서도 노드간 최대거리 및 평균거리가 짧고 메시지처리율이 높도록 하였다. 그리고 대규모 병렬컴퓨터의 집적화에 적합하다고 알려진 다른 구조들과의 성능비교를 위해 노드간 최대거리, 노드간평균거리 및 이분폭(bisection width) 등을 해석적으로 분석하고, 병목해석에 의해 메시지처리율(message completion rate)의 상한을 구하여 비교 검토하였다. 그리고 교차메쉬를 위한 최적 자기 경로설정 알고리듬(optimal self-routing algorithm)을 개발하였으며, 이 알고리듬 및 토로이드와 대각메쉬에 대한 기존의 경로설정 알고리듬[8]을 이용하여 이들 구조의 성능을 우회경로알고리듬을 사용하여 시뮬레이션하고 이를 검토 비교하였다.

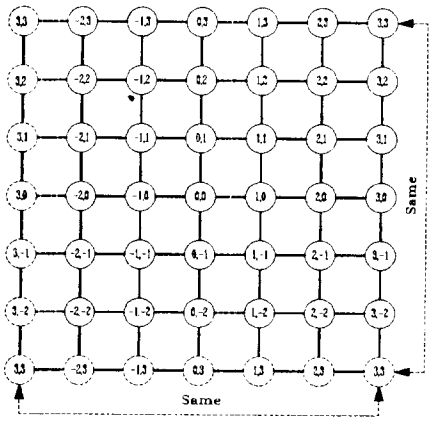
II. 교차메쉬

1. 위상구조(topology)

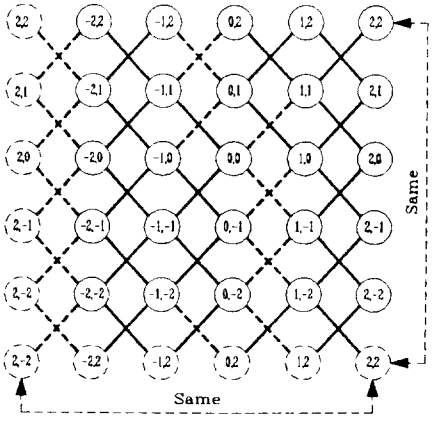
본 논문에서 제안된 교차메쉬의 노드의 수 N은 위상구조가 정방형의 한 축상의 노드의 수를 w라고 할 때 $N = w \times w$이 되며, 정방형이 아닐 때는 $N = n \times k$로 주어진다. 이 때 각 노드의 주소는 순서쌍 (i,j)로 표현된다. 여기서 i 및 j는 $(-k/2+1) \le i \le k/2$ 및 $(-n/2+1) \le j \le n/2$이며, n 및 k는 짝수이다. 제안된 구조는 수평방향의 링크는 토로이드(그림1(a))와 동일

하나. 토로이드의 수직방향 링크대신 대각선방향으로 교차연결한 것이다. 그림1(c)는 노드의 수가 6×6인 교차메쉬로서 이는 메쉬의 형태와 유사하며 일부 링크의 연결형태가 X형으로 구성되므로 XMESH구조라 명명하였다. 이 구조는 부시스템(subsystem)의 형태가 주기적으로 반복배열 되어 있으므로 대규모 시스템을 한 칩에 집적화하기가 수월하다.

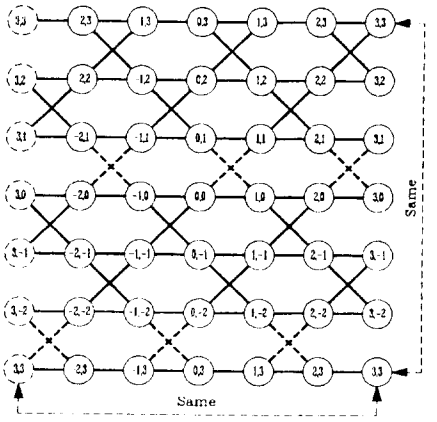
그림 1(c)에서 (-2,*)와 (3,*) 및 (*,-2)와 (*,3)노드의 wrap-around되어 있는 링크는 (3,*) 및 (*,3) 노드들을 이중표기함으로써 나타내었다. 한편 그림2에서 보는 바와 같이 임의의 노드 (i,j)에서 i+j가 짝수일 때는 $([i-1]_k,[j-1]_n)$, $(i,[j-1]_n)$, $(i,[j+1]_n)$ 및 $([i+1]_k,[j+1]_n)$노드와 직연결되며 i+j가 홀수일 때는 $(i,[j-1]_n)$, $([i+1]_k,[j-1]_n)$, $([i-1]_k,[j+1]_n)$ 및 $(i,[j+1]_n)$노드와 직연결되어 각 노드는 4개의 링크를 갖는다.



(a)



(b)



(c)

그림 1. 여러가지 위상구조
Fig. 1. Various interconnection networks.
(a) 6×6 TMESH.
(b) 5×5 DMESH.
(c) 6×6 XMESH.

여기서 이 구조의 순환적 특징 및 상대주소공간 개념에 따라 기호 $[x]_k$는 다음과 같이 정의하였다.

$$[x]_k = \begin{cases} x, & \text{if } -(k-1)/2 < x \le k/2 \\ x-k, & \text{if } x > k/2 \\ x+k, & \text{if } x \le -(k-1)/2 \end{cases} \quad (1)$$

2. 경로설정(Routing)

상호연결망에서 경로설정 알고리듬이란 한 메시지(message)를 망내의 목적지까지 안내하는 경로설정과정이며, 이 경로설정의 주 목적은 각 메시지의 전송지연이 적은 경로를 선택하는 것이다. 병렬컴퓨터시스템에서 한 노드가 다른 노드로 한 메시지를 전송하기 위해서는 그 메시지의 발생지에서 목적지까지의 경로상에 있는 각 노드는 이 전송에 참여를 하게 된다. 즉, 각 노드는 한 메시지를 받아서 다음에 이 메시지를 어디로 전송할 것인지를 결정하여야 한다. 본 위상구조의 순환구조 및 상대주소공간개념에 근거하여 최적 자기경로설정(optimal self-routing) 알고리듬을 제안하였으며, 이를 그림 3에 나타내었다. 이 알고리듬은 임의의 노드 (i,j)에서 목적지노드 (p,q)까지 최단거리로 메시지가 전송되도록 고안하였다.

발생지에서 메시지가 생성될 때 목적지노드의 주소가 메시지내에 함축된다고 가정하였다.

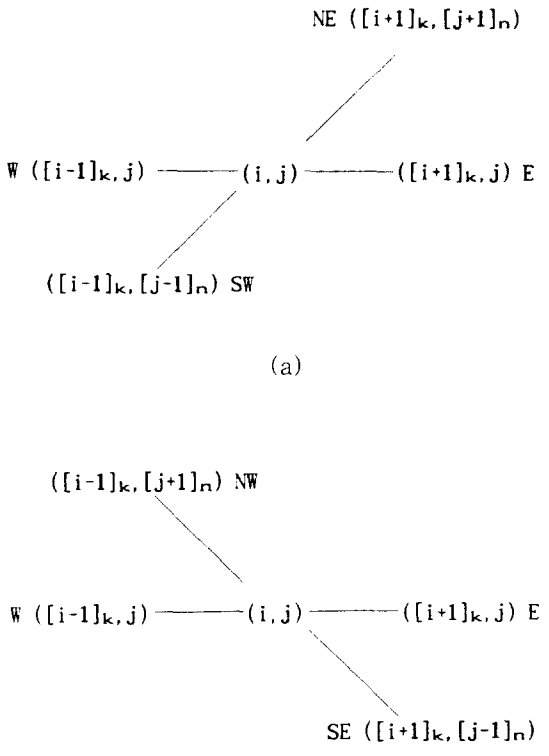


그림 2. 노드 (i,j)의 인접 노드
(a) i+j가 짝수인 경우
(b) i+j가 홀수인 경우

Fig. 2. The neighbor nodes of the node (i,j).
(a) when i+j is even.
(b) when i+j is odd.

메시지가 경로상의 중간노드에 도달했을 때, 중간노드의 주소와 목적지노드의 주소를 비교하여 다음의 진행방향을 결정한다. 최단 경로가 여러 개가 존재할 경우 가능한 경로중 임의로 하나 선택하도록 하였다. 만약 메시지의 경로충돌이나 링크결함등의 이유로 해당 링크에서 blocking이 발생하는 경우에는 다른 최단경로중 하나를 선택하게 되며 다른 최단 경로가 없다면 준최적(semi-optimal) 경로를 선택한다. 다음의 중간노드에서는 이 과정을 계속 반복함으로써 경로설정을 하여, 결국 목적지노드까지 메시지를 전송하게 된다. 그림 3은 n×k 교차메쉬의 노드(i,j)에서 노드(p,q)로의 경로설정 알고리듬을 기술한 것으로서, i+j가 홀수인 경우는 교차메쉬의 구조적 특성상 y대신 $[-y]_n$를 대치하면 i+j가 짝수인 경우와 동일한 형태가 됨에 착안하여 작성하였다. 토로이드나 대각메쉬의 경우보다는 다소 복잡하지만 간단하게 구현할 수 있음을 알 수 있다.

```
Routing between (i,j) and (p,q) in a XMESH with
N=n×k nodes. (n,k are even)

Step 1 : Evaluate x=[p-i]k and y=[q-j]n, where
                  ⎧ k,     if -(k-1)/2<x≤k/2
         [x]k =   ⎨ x - k,  if  x > k/2
                  ⎩ x + k,  if  x ≤ -(k-1)/2

Step 2: If i+j is odd, replace y with [-y]n.

Step 3: Calculate z.
        z = ⎧ x-y,    if x*y>0
            ⎩ x+y,    if x*y≤0

Step 4: Determine the optimal directions.
        If i+j is odd, NEs and SWs in step 4
          must be replaced with SEs and NWs.
  1. When x>0 and y>0
       if ⎧z≥2)                  : NE or E
          ⎨|z|≤1)                : NE
          ⎨z≤-2) and (z=even)    : NE,W or E
          ⎩z≤-2) and (z=odd )    : NE
  2. When x<0 and y>0
       if ⎧z≤-3)                 : W or SW
          ⎨-2≤z≤-1)              : W
          ⎨z=1)                  : NE or E
          ⎨z≥0) and (z=even)     : NE,W or E
          ⎩z≥3) and (z=odd)      : NE
  3. When x<0 and y<0
       if ⎧z ≤-2)                : SW or W
          ⎨|z|≤1)                : SW
          ⎨z≥2) and (z=even)     : SW,E or W
          ⎩z≥3) and (z=odd)      : SW
  4. When x>0 and y<0
       if ⎧z≥3)                  : E or NE
          ⎨1≤z≤2)                : E
          ⎨z=-1)                 : SW or W
          ⎨z≤0) and (z=even)     : SW,E, or W
          ⎩z≤-3) and (z=odd)     : SW
  5. When x≠0 and y=0
       if ⎧z≥3)                  : E or NE
          ⎨1≤z≤2)                : E
          ⎨-2≤z≤-1)              : W
          ⎩z≤-3)                 : W or SW
  6. When x=0 and y≠0
       if ⎧z=1)                  : NE or E
          ⎨z>1) and (z=even)     : NE,W or E
          ⎨z>1) and (z=odd)      : NE
          ⎨z=-1)                 : SW or W
          ⎨z<-1) and (z=even)    : SW,E or W
          ⎩z<-1) and (z=odd)     : SW
```

그림 3. 교차메쉬의 임의의 노드 (i,j)에서 노드 (p,q)로의 경로설정 알고리듬

Fig. 3. Routing algorithm from node (i,j) to node (p,q) in XMESH.

Ⅲ. 구조해석

1. 접속도 및 이분폭

한 노드에 연결되는 링크의 수를 링크접속도(link connectivity, degree)라고 하며 실제 병렬컴퓨터시스템을 구현하는데 있어 이 접속도는 상당한 제약조건이 되며 가급적이면 이 접속도는 적고 일정한 것이 바람직하다. 표 1에서는 교차메쉬와 다른 여러 구조들과 노드수, 링크접속도 및 이분폭을 비교하였다. 하이퍼큐브(Hypercube)의 경우 링크접속도는 하이퍼큐브의 차원(dimension) D와 같으며 링크의 수는 D·

2^{n-1}으로 노드의 수가 많아질수록 링크에 대한 비용이 무척 커지게 된다. 따라서, 많은 수의 노드를 포함하는 큰 시스템을 집적화 함에는 하이퍼큐브구조가 적합치 않다고 알려져 있다.[11] 반면에 교차메쉬, 대각메쉬 및 토로이드의 경우에는 네트워크가 아무리 커지더라도 접속도는 일정수로 고정되어 있으면서 구조 또한 규칙적인 형태를 가지므로, 노드의 수가 많은 큰 시스템이라도 용이하게 집적화할 수 있다는 장점이 있다.

표 1. 링크접속도, 이분폭, 노드간 최대거리 및 평균거리의 비교

Table 1. Comparison of the degree, the bisection width, the diameter and the mean internode distance.

		XMESH	TMESH	DMESH
Number of Nodes, N		w^2	w^2	w^2
Number of Links, L		$2w^2$	$2w^2$	$2w^2$
Degree		4	4	4
Bisection Width, B		min(4n, 2k)	min(2n, 2k)	min(4n, 4k)
Diameter, D_m	if n=k	w/2+2	2⌊w/2⌋	w-1
	if n < k	max{n/2+2, k/2}	⌊n/2⌋ + ⌊k/2⌋	max{n,(k-1)/2}
Mean Internode Distance, d		$\frac{4w^3+9w^2+2w-24}{12(w^2-1)}$	$\frac{w^3}{2(w^2-1)}$	$\frac{w}{2}$

상호연결망의 이분폭(bisection width)이란 망을 동일한 수의 노드로 이등분하기 위해 제거하여야 하는 최소한의 링크의 수를 말한다. 제거하여야 할 링크를 그림 1에서 점선으로 표시하였다. 효과적인 통신 및 결함허용여유(fault tolerance)를 위해서는 이분폭이 큰 것이 바람직하다. n×k(단, n 및 k는 짝수) 토로이드, n×k(단, n 및 k는 짝수) 교차메쉬 및 n×k(단, n 및 k는 홀수) 대각메쉬의 이분폭 B_t, B_x 및 B_d는 각각 다음과 같다.

$$B_t = \min(2n, 2k) \tag{2}$$

$$B_x = \min(4n, 2k) \tag{3}$$

$$B_d = \min(4n, 4k) \tag{4}$$

교차메쉬의 이분폭 B_x는 $n \le k < 2n$일 경우에는 대각메쉬의 이분폭 B_d보다 작으나 $k \ge 2n$일 경우에는 B_d와 같아지게 된다. 그리고 k>n일 경우 교차메쉬의 이분폭 B_x는 토로이드의 B_t보다 항상 큰 값을 가지며 $k \ge 2n$이면 $B_x = 2B_t$로 이분폭이 2배가 된다.

2. 노드간 거리

네트워크의 노드간 최대거리(diameter)란 임의의 두 노드간 거리(number of hops)중 최대값을 말한다. 여기서 노드간 거리란 한 노드에서 다른 노드로 가는 여러 경로중 최단거리를 의미한다. 반면, 평균노드간 거리(mean internode distance)는 한 메시지가 목적지까지 도달하기 위해 가로질러야 할 평균 링크의 수이며, 보통 메시지의 평균 지연시간 측면에서는 노드간 최대거리보다 더 우수한 척도라 할 수있다.

각 노드는 균등한 메시지 routing분포(uniform message routing distribution)를 갖는다고,즉, 임의의 노드 i에서 노드 j로 메시지를 보내는 확률이 모든 i,j(i≠j 및 i,j∈V(G))에 대하여 동일하다고 가정한다. 이와 같이 균등메시지분포를 갖는 대칭망의 임의의 한 노드에서 거리가 정확히 k인 노드의 수를 N(k) 그리고 최대거리를 k_{max}라 정의하면, 평균노드간거리 d는 다음 식으로 주어진다.

$$d = \frac{\sum_{k=1}^{k_{max}} k*N(k)}{N-1} \tag{5}$$

교차메쉬의 경우 N(k)은 식(6)과 같이 주어지며 평균노드간거리 d는 식(7)과 같이 구할 수 있다.

$$N(k) = \begin{cases} 4 & \text{for } k=1 \\ 8k-6 & \text{for } 2 \le k \le k_{max}-3 \\ 3w-6 & \text{for } k=k_{max}-2 \\ 3w/2-2 & \text{for } k=k_{max}-1 \\ w/2-1 & \text{for } k=k_{max} \end{cases} \tag{6}$$

$$d = \frac{\sum_{k=1}^{k_{max}} k*N(k)}{N-1} = \frac{4w^3+9w^2+2w-24}{12(w^2-1)} \tag{7}$$

여기서 k_{max}는 교차메쉬의 노드간 최대거리 D_{mx}이며 $k_{max} = D_{mx} = (w/2+2)$이다. 동일한 방법으로 정방형(w×w)일 경우의 각 구조들에 대해 노드간 최대거리 및 노드간 평균거리를 구하여 표 1에 나타내었다. 그리고 그림 4 및 5는 노드수에 따른 노드간 최대거리 및 평균노드간거리를 그림으로 나타낸 것이다. 네트워크의 크기가 그리 크지 않을 때에는 각 구조간에 큰 차이가 없으나 노드의 갯수가 커짐에 따라 그 차이는 점점 벌어지게 된다. 제안된 교차메쉬구조는 표 2에서 보듯이 노드의 수가 16384개 이상이 되면 노드간 최대거리 및 평균노드간거리에 있어서 토로이드의 경우보다 각각 약 0.52배 및 0.68배 그리고 대각메쉬의 경우보다 각각 약 0.52배 및 0.68배 밖에 되지 않는

다. 네트워크에서 두 노드간의 전송시간은 이 값들과 비례하는 경향이 있으므로 토로이드 및 대각메쉬에 비해 교차메쉬구조가 통신속도면에서 더 우수함을 입증한다.

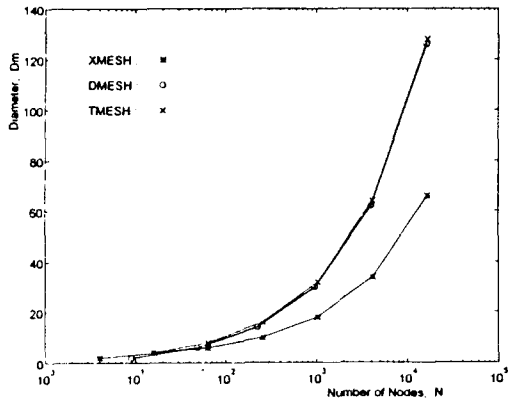


그림 4. 노드수에 따른 노드간 최대거리
Fig. 4. The diameter vs. the number of nodes.

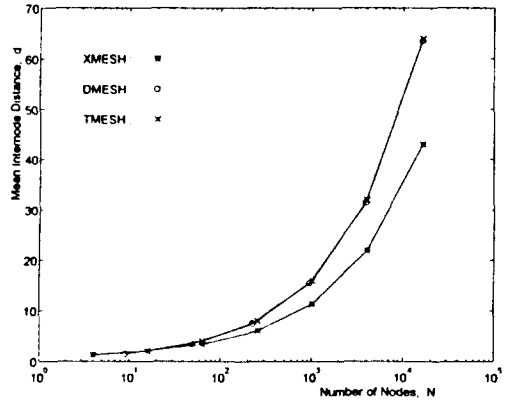


그림 5. 노드수에 따른 평균 노드간거리
Fig. 5. The mean internode distance vs. the number of nodes.

n×k 대각메쉬는 n=k일 때는 토로이드와 노드간 최대거리의 관점에서 비슷한 특성을 갖지만, k>n+2일 때는 토로이드보다 더 우수한 특성을 갖는다. 이와 같이 대각메쉬가 아주 우수한 특성을 나타내는 조건하에서 대각메쉬, 토로이드 및 교차메쉬의 노드간 최대거리를 비교하기로 한다. 그러나 교차메쉬는 n 및 k가 짝수일 경우에만 정의되고 대각메쉬는 n 및 k가 홀수일 경우에만 정의되므로 동일한 조건하에서는 비교할 수 없다. 따라서 대각메쉬보다 조금 더 작은 경우와 조금 더 큰 규모의 교차메쉬와 노드간 최대거리를 비교하여 표2에 나타내었다. 여기서 n×k 교차메쉬가 더 열악한 조건 하에서도 토로이드뿐 아니라 대각메쉬보다 더 짧은 노드간 최대거리를 갖는다는 것을 보여주고 있다. 또한, n×k 교차메쉬구조는 확장성이 뛰어나 k≤n+4인 범위 내에서는 표 2에서 보는 바와 같이 노드간 최대거리의 증가없이 확장이 가능함을 알 수 있다.

표 2. $k \geq n$일때 노드간 최대거리의 비교
Table 2. Comparison of the diameter when $k \geq n$.
(D_{mx} : Diameter of XMESH
D_{mt} : Diameter of TMESH
D_{md} : Diameter of DMESH)

	D_{mx}	D_{mt}
n=k=34	19	34
n=34, k=36	19	35
n=34, k=38	19	36
n=34, k=40	20	37
n=34, k=50	25	42
n=34, k=68	34	51

	D_{md}	D_{mt}
n=k=35	35	34
n=35, k=37	35	35
n=35, k=39	35	36
n=35, k=41	35	37
n=35, k=51	35	42
n=35, k=69	35	51

	D_{mx}	D_{mt}
n=k=36	20	36
n=36, k=38	20	37
n=36, k=40	20	38
n=36, k=42	21	39
n=36, k=52	26	44
n=36, k=70	35	53

Ⅳ. 성능평가 및 검토

1. 메시지처리율

본 장에서는 병목해석에 의한 성능평가방법[5-6,10-11]을 사용하여, 네트워크의 메시지처리율의 상한값을 구하고 이를 비교 검토하기로 한다. 네트워크의 한 노드에서 다른 노드로 메시지를 보낼 때마다 이 메시지는 몇 개의 통신 링크와 경유지의 노드들을 지나서 목적지 노드에 도달하게 되며 이 목적지에서 주어진 계산

을 하게 된다. 이 때 메시지는 목적지노드와 경유되는 링크를 "방문(visit)한다"고 하며, 메시지가 임의의 장치(링크 혹은 목적지노드) i를 방문하는 횟수의 평균값을 장치 i에 대한 방문비라 하고 이를 V_i로 표시하며 장치 i에서 이 메시지를 처리하는데 소요되는 평균 시간을 평균서비스시간이라 하고 S_i라고 표시하기로 한다. X_i를 장치 i에서의 평균 메시지처리율($X_i \leq 1/S_i$)이라 하고, 장치i가 이용(busy)되는 확률을 장치 i의 이용도(utilization)라 하고 U_i로 표시하면 다음 법칙이 성립한다.[11]

$$U_i = X_i S_i \quad \text{(utilization law)} \tag{8}$$

$$X_o = \frac{X_i}{V_i} \text{(forced flow law)} \tag{9}$$

여기서, X_o는 전체 네트워크의 메시지처리율이다. (9)식에 (8)식을 대입하면 다음 식을 얻을 수 있다.

$$X_o = \frac{U_i}{V_i S_i} \tag{10}$$

만약 네트워크에서 순환되는 메시지의 수가 점차 증가한다면 최대의 $V_i S_i$값을 갖는 장치부터 이용도가 1에 근접하게 되며, 이 장치가 전체 시스템에서 메시지의 순환을 제한하게 된다. 따라서 메시지처리율 X_o의 상한값은 다음 식으로 주어진다.

$$X_o \leq \frac{1}{V_b S_b} \quad (\text{단, } V_b S_b = \max V_i S_i) \tag{11}$$

이와 같은 해석방법을 점근해석(asymptotic analysis) 또는 병목해석(bottleneck analysis)이라 한다.[6,10] 이 때 편의상 모든 노드는 동일한 평균서비스시간 S_P를 가지며 모든 링크에서 역시 동일한 평균 서비스시간 S_C을 갖는다고 가정한다. 또한 균등한 메시지 routing분포를 갖는, N개의 노드로 구성된 대칭구조인 망에서 노드에 대한 방문비는 $V_P = 1/N$으로 주어진다.

d를 노드간 평균거리, 즉, 한 메시지가 전달되기 위해 가로지르는 평균 링크의 수라고 하고 L을 망의 링크의 수라 하면, 링크방문비 V_C 및 메시지처리율 X_o는 다음과 같다.

$$V_c = \frac{d}{L} \tag{12}$$

$$X_0 \leq \frac{1}{\max\{V_P S_P, V_c S_c\}} = \min\{\frac{1}{V_P S_P}, \frac{1}{V_c S_c}\} \tag{13}$$

w×w 교차메쉬의 $V_i S_P$, 링크의 수 L 및 노드간평균거리 d는

$$V_i S_P = \frac{S_P}{N} \tag{14}$$

$$L = 2w^2 \tag{15}$$

$$d = \frac{4w^3+9w^2+2w-24}{12(w^2-1)} \tag{7}$$

이므로 링크 방문비 V_C 및 교차메쉬의 메시지처리율 X_o는 다음과 같다.

$$V_C = \frac{d}{L} = \frac{4w^3+9w^2+2w-24}{24w^2(w^2-1)} \tag{16}$$

$$X_o \leq \min\{\frac{w^2}{S_b}, \frac{24w^2(W^2-1)}{S_c(4w^3+9w^2+2w-24)}\} \tag{17}$$

$$X_o \leq \frac{24w^2(w^2-1)}{S_c(4w^3+9w^2+2w-24)} \tag{18}$$

동일한 방법으로 정방형(w×w)일 경우의 각 구조들에 대해 메시지처리율의 상한을 계산하여 표 3에 나타내었다.

표 3. 메시지처리율의 비교
Table 3. Comparison of the message completion rate.

		XMEXH	TMESH	DMESH
bottleneck criteria	$\frac{1}{V_P S_P}$	$\frac{w^2}{S_P}$	$\frac{w^2}{S_P}$	$\frac{w^2}{S_P}$
	$\frac{1}{V_i S_i}$	$\frac{24w^2(w^2-1)}{(4w^3+9w^2+2w-24)S_i}$	$\frac{4(w^2-1)}{wS_i}$	$\frac{4w}{S_i}$
bound on Xo		$\frac{24w^2(w^2-1)}{(4w^3+9w^2+2w-24)S_i}$	$\frac{4(w^2-1)}{wS_i}$	$\frac{4w}{S_i}$

그림 6은 균등한 메시지 라우팅분포를 가정하고 노드와 링크들이 단위처리속도를 갖는다는 가정하에 여러 구조의 노드수에 따른 메시지처리율의 한계를 구하여 비교하고 이를 그림으로 나타내었다. 여기서 교차메쉬의 메시지 처리율의 상한은 토로이드 및 대각메쉬에 비해 크며 이는 노드수가 증가할수록 차이가 커져서 노드수가 16384개 정도가 되면 메시지처리율의 측면에서 공히 1.5배 정도 성능이 더 우수함을 알 수 있다.

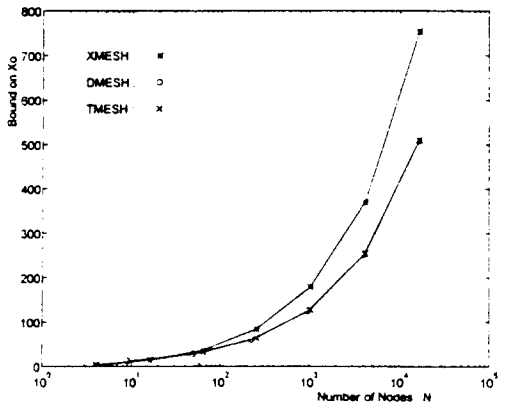


그림 6. 노드수에 따른 메시지처리율의 한계
Fig. 6. The bound on the message completion rate vs. the number of nodes.

2. 시뮬레이션 및 결과

이 절에서는 우회경로설정알고리듬(deflection routing algorithm)을 이용하여 대각메쉬, 토로이드 및 교차메쉬의 성능을 비교하였다. 시뮬레이션을 함에 있어 대각메쉬, 토로이드 및 교차메쉬는 양방향의 링크를 가지며 초기 단계에서 각 노드마다 메시지를 한개 내지 네개씩(Nmsg=1...4) 할당한다. Nmsg=1일 경우 가벼운 부하에 해당하며, 각 노드는 4개씩의 출력링크를 가지므로 Nmsg=4일 경우는 최대부하에 해당한다. 각 메시지의 목적지주소를 발생시키기 위해 균등한 메시지분포(uniform message distribution)를 갖는 난수발생기(random number generator)를 사용하였으며 본 시뮬레이션에서는 2단계 스케줄링 알고리듬을 이용하였다.

첫번째 단계에서는 각 노드에 존재하는 모든 메시지들을 정렬하여 그 우선순위에 따라 최적인 출력링크(optimal output link)를 할당한다. 최적인 경로상의 링크에 경합이 생길 경우 높은 우선순위를 가진 메시지가 먼저 최적의 링크를 할당받고 낮은 우선순위의 메시지는 다른 최적의 경로를 할당받는다. 만약 최적인 경로가 남아 있지 않다면 그 메시지는 다른 모든 메시지가 최적경로를 할당받은 후 남아 있는 경로중 하나를 할당받게 된다. 두번째 단계에서는 출력 링크에 할당된 각 메시지를 전송하게 되며 이 메시지를 받은 노드는 메시지의 목적지 주소를 조사하여 목적지노드에 도달한 메시지가 있으면 이를 시스템에서 제거하고 새 메시지를 발생시킨다. 따라서 시스템내의 메시지의 수는 항상 N*Nmsg개로 고정된다. 이 2단계 스케줄링 알고리듬은 구현하기 간단하지만 우회조건(deflection criteria)을 잘못 선정할 경우 livelock현상이 발생하여 메시지가 목적지에 도달하지 못하고 무한히 시스템내를 돌아다니는 경우가 발생할 수도 있다. 여러 우회조건의 영향을 파악하기 위해 N=36×72=2592노드를 가진 교차메쉬에 대하여 다음 4가지의 조건을 적용하였다.

i) random : 임의의 순서로 경로설정이 되는 경우
ii) age : 오래된 메시지가 높은 우선순위를 갖는 경우
iii) path_num : 최적 경로의 수가 적은 메시지가 높은 우선순위를 갖는 경우
iv) age+path_num : 오래된 메시지가 높은 우선순위를 갖는데 만약 같은 조건 이라면 최적경로의 수가 적은 메시지가 높은 우선순위를 갖는 경우

그 결과 임의경로설정조건(random)의 경우 livelock현상으로 무한지연이 발생하였으며 조건 age및 age+path_num의 경우가 가장 우수하게 나타났다. 우회조건으로 age 및 path_num 두 개를 사용함에 비해 age 하나만을 사용할 경우 스케줄링 알고리듬이 간단하면서도 성능의 손색이 거의 없으므로 본 실험에서는 우회조건 age를 사용하여 36×72교차메쉬, 35×71대각메쉬 및 35×71토로이드에 대하여 사이클이 진행됨에 따른 최대지연(maximum delay), 평균지연(average delay) 및 메시지처리율(throughput)을 조사하였다. 실험의 결과를 검토하기위해 노드간 최대거리 및 평균노드간거리의 이론치를 표 1에 의한 계산, 그림 3의 알고리듬 및 기존의 토로이드와 대각메쉬에 대한 알고리듬에 의하여 구하여 표 4에 나타내었다.

표 4. 노드간 최대거리 및 평균노드간거리
Table 4. The diameter and the mean internode distance.

	36×72 XMESH	35×71 TMESH	35×71 DMESH
Diameter, D_m	36	34	52
Mean Internode Distance, d	19.86	23.50	26.50

그림 7은 최대지연을 나타낸 것이며 Nmsg=4일 경우 즉, 최대부하일 경우에는 링크에 대한 경합이 발생하여 경로가 우회되므로 많은 지연이 있음을 알 수

있다.

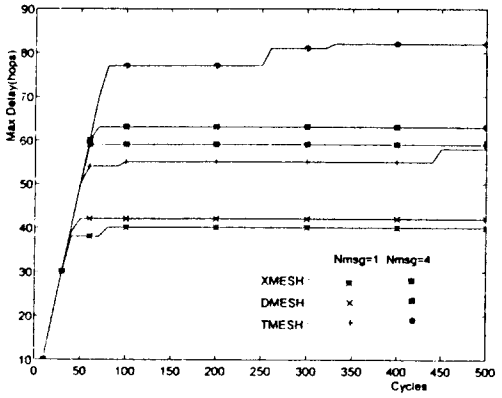


그림 7. 최대지연
Fig. 7. The maximum delay.

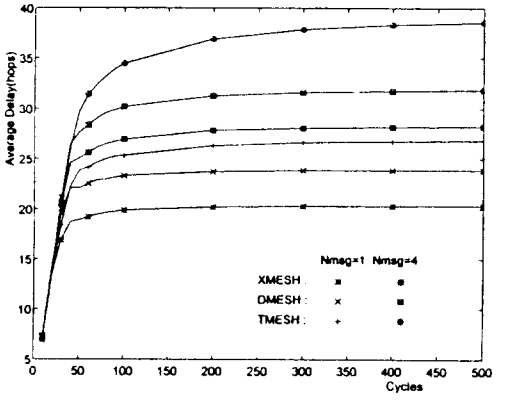


그림 8. 평균지연
Fig. 8. The average delay.

그림 8은 평균지연을 나타낸 것으로서 Nmsg= 1 일 경우에는 시뮬레이션의 결과인 평균지연이 이론치인 평균노드간거리에 거의 근접하고 있는데 이 경우는 메시지의 수가 작아 상호간 경합이 적으므로 대부분의 메시지가 우회되지 않고 최적경로를 통해 전송되기 때문이다. 메시지의 수를 많게 하면 우회되는 확률이 커져서 최대지연 및 평균지연이 커지게 되는데 시뮬레이션의 결과로 이를 확인할 수 있다. 또한 포화상태가 된 후에는 교차메쉬의 경우 대각메쉬나 토로이드에 비해 평균지연이 항상 작게 나타나며 이 차이는 부하의 양이 커질수록, 즉 메시지의 수가 많아질수록 더 커짐을 알 수 있다. 그림 9는 메시지처리율에 대한 실험결과를 나타낸 것이다. 사이클이 진행됨에 따라 메시지 처리율은 최대지연이나 평균지연과 마찬가지로 포화됨을 볼 수 있는데 이는 시스템 내부의 메시지 수를 항상 일정한 수로 고정하였기 때문이며 메시지 수가 많은 중부하(heavy load)에서는 이 포화가 늦게 일어남을 시뮬레이션을 통해 확인할 수 있다.

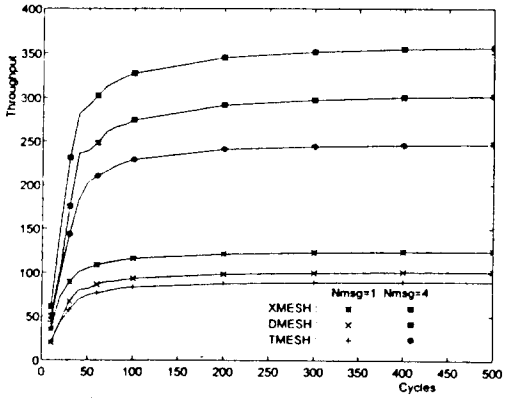


그림 9. 메시지처리율
Fig. 9. The throughput.

V. 결 론

본 논문에서는 대규모 병렬컴퓨터의 구현에 적합한 새로운 위상구조로서 교차메쉬구조를 제안하였다. 제안된 구조에서는 토로이드의 수평링크는 그대로 두고 수직링크대신 대각선방향 교차연결한 것으로서 토로이드가 갖는 장점들 즉, 대칭구조, 단위 부시스템(subsystem)의 주기적 반복배열, 일정접속도 등의 장점을 가져 대규모 병렬컴퓨터의 상호접속망을 집적구현하기 용이하게 하였다. 또한, n×k 교차메쉬구조는 확장성이 뛰어나 k가 n부터 n+4까지의 범위 내에서는 노드간 최대거리의 증가없이 확장이 가능하다.

성능해석에 의하여 비교한 결과, 제안된 구조의 노드간 최대거리 및 평균노드간거리는 토로이드 혹은 대각메쉬의 경우보다 짧으며 노드수가 많아질수록 격차가 커져서 노드의 수가 16384개 정도가 되면 노드간 최대거리 및 평균노드간거리에 있어서 토로이드의 경우보다 각각 약 0.52배 및 0.68배 그리고 대각메쉬의 경우보다 각각 약 0.52배 및 0.68배 밖에 되지 않는다. 그리고 병목해석(bottleneck analysis)에 의해 메시지 처리율의 상한(upperbound)을 구한 결과, 교차메쉬의 메시지 처리율의 상한은 토로이드 및 대각메쉬에 비해 크며 이는 노드수가 증가할수록 차이가 커져서 노드수가 16384개 정도에서는 메시지 처리율의 상한은 토로이드 및 대각메쉬에 비해, 메시지처리율의 측면에서 공히 1.5배 정도 성능이 더 우수함을 확인하였다.

이 결과들을 입증하기 위해 교차메쉬를 위한 최적자기경로 알고리듬을 개발하여 각 구조의 평균지연, 최

대지연 및 메시지처리율에 대해 시뮬레이션을 수행하였다. 이 실험 결과, 통신부하량의 대소나 시스템의 노드수와 관계없이 교차메쉬는 모든 측면에서 토로이드 및 대각메쉬에 비해 성능이 우수함을 확인할 수 있었다.

참 고 문 헌

[1] P. Mazumder, "Evaluation of On-Chip Static Interconnection Networks," IEEE Trans. Comp. Vol.C-36, pp.365-369, Mar. 1987.

[2] K. Hwang, *Advanced Computer Architecture : Parallelism Scalability Programmability*, McGraw-Hill, 1993.

[3] G.D.Vecchia and C.Sanges, "Recursively Scalable Networks for Message Passing Architectures," 12th IMACS World Congress on Scientific Computation, Paris, pp.33-40, July 1988.

[4] Y.L.Yang and H.Amano, "Recursive Diagonal Torus: An Interconnection Network for Massively Parallel Computers," Proc. of 5th Symp. on PDP, pp.591-594, Dec. 1993.

[5] Zahorjan, K.C.Sevick, D.L.Eager, and Galler, "Balanced Job Bound Analysis of Queueing Networks," Commun. Ass. Comput. Mach., Vol.25, pp.134-141, Feb. 1982.

[6] D.A.Reed and H.D.Schwetman, "Cost-performance Bound for Multimicrocomputer Network," IEEE Trans. Comput., Vol. C-32, pp.83-95, No.1, Jan. 1983.

[7] B.W.Arden and F.Li, "Simulation of interconnection networks for massively parallel systems," Tech. Rep., Dept. of Elec. Eng., Univ. of Rochester, Rochester, NY, 1991

[8] K.W.Tang, and S.A.Padubidri,"Diagonal and Toroidal Mesh Networks," IEEE Trans. on Comp., Vol.43, pp.815-826, July 1994.

[9] P.Barnes, "On Distributed Communication networks," IEEE Trans. Commun. Syst., Vol.12, pp.1-9, 1964.

[10] D.A.Reed and D.C.Grunwald, "The Performance of Multicomputer Interconnection Networks," Computer, Vol.20, pp.63-73, No.6, June 1987.

[11] P.J.Denning and J.P.Buzen, "The Operational Analysis of Queueing Network Models," Comp. Surveys, Vol.10, pp.225-261, Sept. 1978.

저 자 소 개

金 鍾 晋(正會員)

1957年 4月 28日生. 1985年 2月 한국과학기술원 전기및전자공학과 졸업(공학석사). 1990年 3月 ~ 현재 경북대학교 대학원 전자공학과 박사과정. 1987年 3月 ~ 현재 부산공업대학교 전자공학과 부교수. 주관심분야는 병렬분산처리, 상호접속망, 컴퓨터시스템구조 등임.

崔 興 文(正會員) 제 27권 제7호 참조

현재 경북대학교 전자공학과 교수