메시지 패싱 병렬 처리 시스템의 상호연결망 비교

한종석, 심원세, 한우종

한국전자통신연구원 컴퓨터.소프트웨어연구소 하드웨어구조연구팀

Comparison of Interconnection Networks for Message Passing Parallel Processing Systems

Jong-Seok Han, Won-Sae Sim, Woo-Jong Hahn

Hardware Architecture Research Team, ETRI-CSTL

jshan@computer.etri.re.kr

요 약(Abstract)

본 논문에서는 메시지 패싱 전송을 기반으로 하는 병렬 처리 시스템의 상호연결망 구조와 특성을 조사하고 비교한다. 특히, 상용 시장에서 널리 알려진 대표적인 병렬 처리 시스템의 상호연결망 특성과 ETRI 에서 개발된 고속 병렬 컴퓨터(SPAX)의 계층 크로스바 상호연결망(Xcent-Net) 특성을 상호 비교한다.

메시지 패싱 전송 기반의 상호연결망은 일반적으로 확장성이 우수하여 대규모 병렬 처리 시스템을 구축하는데 유리하다. Cray T3E 시스템, Intel ASCI TFLOPS 시스템, Tandem Himalaya S70000 시스템, IBM RS6000 SP2 시스템등은 메시지 패싱 상호연결망을 기반으로 수백개에서 수천개의 대규모 프로세서를 연결한 병렬 처리 시스템이다. ETRI SPAX 시스템은 Xcent-Net 메시지 패싱 상호연결망을 기반으로 최대 256 개 프로세서를 연결한 고속 병렬 처리 시스템으로 우수한 확장성과 높은 성능을 제공한다. 본 논문에서는 상호연결망의 구조와 함께 라우팅 스위치 구조 및 특성을 중심으로 전송 지연시간, 그리고 노드당 전송 대역폭 특성을 비교한다.

I. 서 론

병렬 처리 시스템의 상호연결망은 병렬 처리 컴퓨터의 구조와 성능을 결정하는 중요한 구성요소로서 효율적인 연결 방법과 대규모 프로세싱 노드를 연결하기 위한 우수한 확장성을 제공하여야 한다.

병렬 처리 시스템은 크게 공유 메모리 구조의 CC-NUMA(cache coherent non-uniform memory access) 병렬 처리 시스템과 메시지 패싱 구조의 병렬 처리 시스템, 그리고 공유 메모리 구조의 SMP (symmetric multi-processor) 노드들을 메시지 패싱 상호연결망으로 연결한 SMP 클러스터링 병렬 처리 시스템으로 구분 할 수 있다. 상호연결망 관점에서 보면 메시지 패싱 병렬 처리 시스템의 상호연결망과 SMP 클러스터링 병렬 처리 시스템의 상호연결망은 동일한 구조와 특성을 가지기 때문에 모두 메시지 패싱 상호연결망으로 분류한다. Tandem Himalaya S70000 시스템의 메시지 패싱 상호연결망인 Server-Net[3]은 SMP 클러스터링 병렬 처리 시스템의 일종인 NT 클러스터 시스템의 상호연결망으로 사용된다. IBM RS6000 SP2 시스템의 상호연결망인 SP2 스위치[4]는 내부의 단일 프로세서 노드와 SMP 노드들을 연결하는 메시지 패싱 상호연결망이다.

CC-NUMA 병렬 처리 시스템은 공유 메모리 구조를 기반으로 캐쉬 일관성 유지 프로토콜을 제공하기 때문에 시스템 특성상 우수한 확장성을 제공하지 못한다. 반면에 메시지 패싱을 기반으로 하는 병렬 처리 시스템은 확장성이 우수하여 대규모 병렬 처리(MPP, massively parallel processing)에 적합하다. 세계 최초의 테라플롭스 병렬 처리 시스템인 Intel ASCI TFLOPS 시스템은 이중 2 차원 메쉬 구조의 메시지 패싱 상호연결망[2]을 사용하여 최대 9216 개 프로세서와 최대 1.8 테라플롭스 성능을 제공한다. Cray T3E 시스템은 3 차원 토러스 메쉬 구조의 메시지 패싱 상호연결망[1]을 사용하여 최대 2048 개 프로세서와 최대 1.8 테라플롭스 성능을 제공한다.

ETRI 가 개발한 고속 병렬 컴퓨터인 SPAX(Scalable Parallel Architecture computer based on X-bar network) 시스템은 계층 크로스바 구조의 메시지 패싱 상호연결망인 Xcent-Net[5]을 사용하여 최대 256 개 주처리 프로세서와 최대 64 개 입출력 프로세서를 제공한다. 상호연결망은 일반적으로 연결 구조, 전송 대역폭, 전송 지연시간과 같은 특성들로 비교 될 수 있다. 상호연결망의 전송 대역폭 특성은 시스템 연결 노드의 수에 따라 달라지기 때문에 노드당 전송 대역폭을 가지고 보

다 공정하게 비교할 수 있다. 상호연결망의 핵심 구성 소자인 라우팅 스위치는 상호연결망의 구조와 특성을 결정하는 매우 중요한 소자이기 때문에 본 논문에서는 상호연결망의 구조와 함께 라우팅 스위치 구조 및 특성을 중심으로 전송 지연시간, 그리고 노드당 전송 대역폭 특성을 비교한다.

1 장의 서론에 이어 2 장에서는 상용 메시지 패싱 병렬 처리 시스템인 Cray T3E 시스템, Intel ASCI TFLOPS 시스템, Tandem Himalaya S70000 시스템, IBM RS6000 SP2 시스템의 상호연결망 및 라우팅 스위치 특성에 대해 기술하고 3 장에서는 ETRI SPAX 시스템의 상호연결망 및 라우팅 스위치 특성에 대해 기술한다. 4 장에서는 상호연결망의 연결 구조와 라우팅 스위치의 구조 및 특성을 중심으로 종합적으로 비교 분석하며, 마지막으로 5 장에서 결론을 맺는다.

II. 상용 메시지 패싱 병렬 처리 시스템

1. Cray T3E 시스템

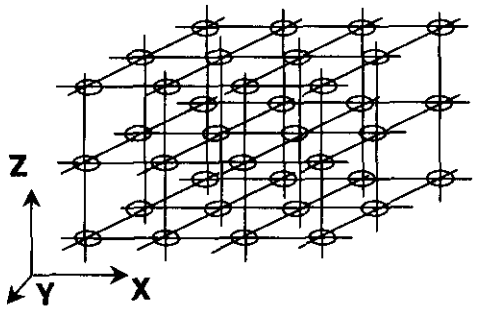
1) 시스템 개요

Cray T3E-900 시스템은 3 차원 토러스 메쉬 구조의 메시지 패싱 상호연결망을 사용하여 최대 2048 개 프로세서와 최대 1.8 테라플롭스 성능을 제공한다. 한 노드는 450 MHz 로 동작하는 하나의 DEC Alpha 21164 프로세서로 구현되며, 상호연결망은 최대 2048 개 노드를 연결할 수 있다. 마이크로 커널 기반의 UNISCO/mk 운영체제를 사용하며 단일 시스템 이미지의 컴퓨팅 환경을 제공한다. 이중 링 구조의 확장성이 높은 고유의 GigaRing 채널을 사용하여 채널당 최대 267 Mbytes/s 의 대역폭을 제공한다.

2) 상호연결망 및 라우터 특성

상호연결망은 3 차원 토러스 메쉬 구조를 갖는다. Full Duplex 전송 방식을 지원하며 결함허용 적응 경로 제어를 수행한다. <그림 1>은 3 차원 토러스 메쉬 구조를 보여준다.

상호연결망을 구성하는 라우팅 스위치는 7 X 7 스위치 구조를 가지며 1 개의 노드 연결 포트와 각 방향 2 포트씩 총 6 개의 상호연결 포트를 제공한다.



<그림 1 : Cray T3E - 3 차원 토러스 메쉬 구조>

라우팅 스위치는 최대 10 개의 플릿(flit)으로 구성된 패킷을 사용한다. 각 플릿은 70 비트 크기로서 5 개의 14 비트 핏(phit)으로 구성된다. 교착 상태를 방지하고 적응 경로제어를 수행하기 위하여 5 개의 가상 채널을 사용하며 Credit-based 흐름 제어를 수행한다. 라우팅 스위치는 내부 75 MHz, 입출력 375 MHz 로 동작하여 노드당 최대 600 Mbytes/s 의 전송 대역폭을 제공한다.

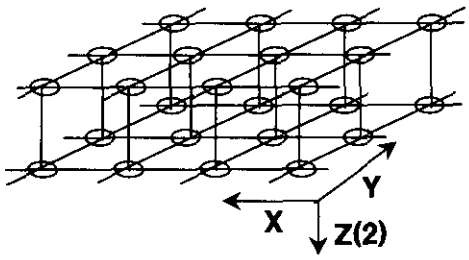
2. Intel ASCI TFLOPS 시스템

1) 시스템 개요

Intel ASCI TFLOPS 시스템은 이중 2 차원 메쉬(two xy-plane mesh)의 메시지 패싱 상호연결망을 사용하여 최대 9072 개 주처리 프로세서와 144 개 입출력 및 응용 프로세서로 구성된다. 최대 1.8 테라플롭스 성능을 제공하며, 한 노드는 200 MHz 로 동작하는 2 개의 Intel Pentium-Pro 프로세서로 구현된다. 상호연결망은 최대 4608 개 노드를 연결할 수 있다. Cougar 운영체제와 Paragon UNIX 운영체제를 사용하며 단일 시스템 이미지의 컴퓨팅 환경을 제공한다.

2) 상호연결망 및 라우터 특성

상호연결망은 이중 2 차원 메쉬 구조를 갖는다. 즉, X x Y x 2 의 연결 구조에서 먼저 Z 방향으로 경로제어를 수행하고 X 방향, Y 방향, 그리고 다시 Z 방향으로 경로제어를 수행한다. Full Duplex 전송 방식을 지원한다. <그림 2>는 이중 2 차원 메쉬 구조를 보여준다.



<그림 2 : ASCI TFLOPS 의 이중 2 차원 메쉬 구조>

상호연결망을 구성하는 라우팅 스위치는 6 X 6 스위치 구조를 가지며 1 개의 노드 연결 포트와 XY 방향 2 개씩 4 포트, Z 방향 1 포트를 총 5 개의 상호연결 포트를 제공한다. 16 비트의 데이터 폭을 가지고 4 클럭을 이용해 8 바이트 크기의 한 플릿을 전송한다. 한 포트당 4 개의 가상 레인(virtual lane)을 사용한다. 라우팅 스위치는 내부 200 MHz, 입출력 200 MHz 로 동작하여 노드당 최대 400 Mbytes/s 의 전송 대역폭을 제공하며, 라우터당 50 ns 의 전송 지연시간을 갖는다.

3. Tandem Himalaya S70000 시스템

1) 시스템 개요

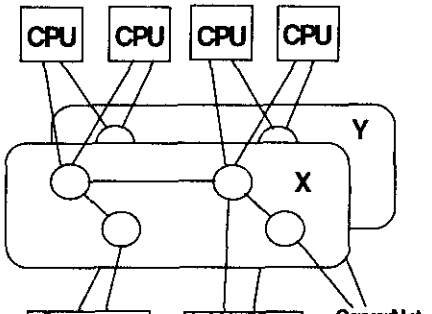
Tandem Himalaya S70000 시스템은 메쉬, 하이퍼큐브, 트리등 가변 구조의 메시지 패싱 상호연결망을 사용하여 최대 4080 개 프로세서와 최대 1.6 테라플롭스 성능을 제공한다. 한 노드는 200 MHz 로 동작하는 최대 16 개의 MIPS R10000 프로세서로 구현되며, 상호연결망은 최대 255 개 노드를 연결할 수 있다. 최소 2 개에서 최대 16 개 프로세서로 구성되는 노드는 SMP 구조를 가지기 때문에 SMP 클러스터링 병렬 처리 시스템이라 할 수 있다. NonStop Kernel 운영 체제를 사용하며 단일 시스템 이미지의 컴퓨팅 환경을 제공한다.

2) 상호연결망 및 라우터 특성

Tandem Himalaya 시스템의 상호연결망인 ServerNet 은 다양한 연결 구조를 지원한다. ServerNet 은 메시지 패싱 상호연결망으로서 Intel 과 MicroSoft 에서 NT 클러스터링 상호연결망으로 채택하였다. ServerNet 은 8 비트 데이터 폭을 갖는 ServerNet I 에 이어 1 비트 직렬 전송을 수행하는 ServerNet II 로 발전하였다. <그림 3>은 ServerNet I 상호연결망의 구조를 보여준다.

ServerNet 을 구성하는 라우팅 스위치는 6 X 6 스위치 구조를 가지며 2 개의 노드 연결 포트와 4 개의 상호연결 포트를 제공한다. 8 비트의 데이터 폭을 가지고 최대 80 바이트 패킷을 전송한다. 라우팅 스위치는 내부 50 MHz, 입출력 50 MHz 로 동작하여 이중 상호연결망 구성시 노드당 최대 100 Mbytes/s 의 전송 대역폭을 제공하며, 라우터당 300 ns 의 전송 지연시간을 갖는다.

최근 개발된 ServerNet II 라우팅 스위치는 13 X 13 스위치 구조를 가지며 직렬 전송을 수행한다. 라우팅 스위치는 내부 125 MHz, 입출력 1 GHz 로 동작하여 포트당 125 Mbytes/s 의 전송 대역폭을 제공한다.



<그림 3 : Tandem Himalaya - ServerNet 구조>

4. IBM RS6000 SP2 시스템

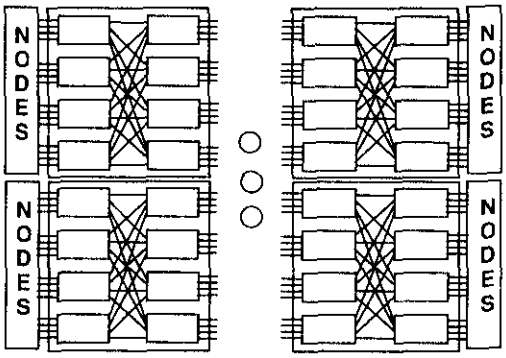
1) 시스템 개요

IBM RS6000 SP2 시스템은 다단계(multistage) 구조의 메시지 패싱 상호인결망을 사용하여 최대 2048 개 프로세서와 최대 1.3 테라플롭스 성능을 제공한다. 한 노드를 332 MHz 로 동작하는 최대 4 개의 PowerPC 604e 프로세서로 구현할 경우 상호연결망은 최대 512 개 노드를 연결할 수 있으며(thin 노드 또는 wide 노드), 한 노드를 200 MHz 로 동작하는 최대 8 개의 PowerPC 604e 프로세서로 구현할 경우 상호연결망은 최대 256 개 노드를 연결할 수 있다(high 노드). AIX 운영 체제를 사용하며 가상 공유 디스크를 제공한다.

2) 상호연결망 및 라우터 특성

IBM RS6000 시스템의 상호연결망인 SP Switch 는 Omega 다단계 상호연결망 구조를 갖는다. Full Duplex 전송 방식을 지원하며 노드간 결함허용 다중 경로를 제공한다. 모든 노드는 동일 거리에 위치하고 전역 클럭을 사용하기 때문에 전송 지연시간이 모두 동일한 특성을 가진다. <그림 4>는 다단계 구조의 SP Switch 상호연결망을 보여준다.

상호연결망을 구성하는 라우팅 스위치는 4 X 4 스위치 구조를 가지며 하나의 스위치 보드에 8 개의 라우팅 스위치가 장착된다. 각 포트는 8 비트의 데이터 폭을 가지고 있으며, 라우팅 스위치는 내부 75 MHz, 입출력 150 MHz 로 동작하여 노드당 최대 150 Mbytes/s 의 전송 대역폭을 제공한다.



<그림 4 : IBM RS6000 SP2 - SP 스위치 구조>

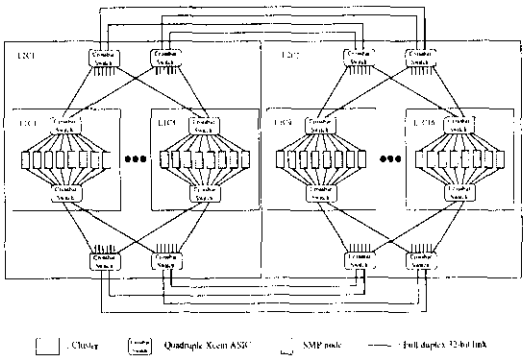
III. SPAX 메시지 패싱 병렬처리 시스템

1. SPAX 시스템 개요

ETRI SPAX 시스템은 계층 크로스바 구조의 메시지 패싱 상호연결망인 Xcent-Net 을 사용하여 최대 256 개 주처리 프로세서와 최대 64 개 입출력 프로세서를 제공한다. 최대 50 기가플롭스 성능을 제공하며, 한 노드는 200 MHz 로 동작하는 4 개의 Intel Pentium-Pro 프로세서로 구현된다. 상호연결망은 최대 128 개 노드를 연결할 수 있다. SPAX 시스템은 4 개의 프로세서로 구성된 SMP 노드를 Xcent-Net 으로 연결한 일종의 SMP 클러스터링 병렬처리 시스템이다. 마이크로 키널을 기반으로 한 MISIX 운영체제를 사용하며 단일 시스템 이미지의 컴퓨팅 환경을 제공한다.

2. Xcent-Net 상호연결망 및 라우터 특성

Xcent-Net 은 계층 크로스바 구조의 메시지 패싱



<그림 5 : ETRI SPAX - Xcent-Net 구조>

상호연결망이다. Full Duplex 전송 방식을 지원하며 결함허용 적응 경로제어를 수행한다. Xcent-Net 은 다른 상호연결망과 달리 가변 데이터 폭 구조를 지원한다. 독립적으로 분산 경로제어를 수행하는 여러 개의 라우팅 스위치를 바이트 슬라이스(byte slice)로 확장하여 라우팅 스위치 그룹을 형성한다.

Xcent-Net 을 구성하는 라우팅 스위치 그룹은 10 X 10 스위치 구조를 가지며 8 개의 노드 연결 포트와 2 개의 상호연결 포트를 제공한다. 라우팅 스위치는 최대 68 개의 플릿(flit)으로 구성된 패킷을 사용한다. 각 플릿은 32 비트 크기이다. 내부 66.67 MHz, 입출력 66.67 MHz 로 동작하여 이중 상호연결망 구성시 노드당 최대 528 Mbytes/s 의 전송 대역폭을 제공한다. 라우팅 스위치당 90 ns 의 전송 지연시간을 갖는다.

IV. 메시지 패싱 상호연결망 비교

상호연결망은 링, 트리, 메쉬, 하이퍼 큐브 구조와 같은 정적 상호연결망과 다단계 구조와 같은 동적 상호연결망으로 구분할 수 있다. 상용 병렬 처리 시스템중에서 IBM RS6000 SP2 시스템의 상호연결망인 SP 스위치는 동적 상호연결망이며 나머지 시스템은 모두 정적 상호연결망이다. <표 1>은 상호연결망 및 라우터의 특성을 종합적으로 비교한 표이다.

정적 상호연결망중 Xcent-Net 은 계층 크로스바 구조인 반면 대부분의 정적 상호연결망은 메쉬 구조를 기반으로 하고 있다. 노드 연결 포트 특성중 이중 포트(dual port)는 상호연결망의 노드 연결 링크에 발생한 결함을 허용할 수 있도록 한다. 메시지 전송 기반의 상호연결망은 확장성이 우수하기 때문에 수천개 노드까지 쉽게 확장할 수 있다. 현재 Intel TFLOPS 시스템의 경우 4608 노드를 제공한다. 상호연결망의 핵심 소자인 라우팅 스위치의 경우 최소 4 개에서 최대 10 개까지 포트를 제공하며 모두 Full Duplex 방식을 제공한다. 내부 동작 속도와 입출력 동작 속도는 스위치마다 차이를 보이고 있으며, Cray T3E 시스템의 상호연결망과 SP Switch 상호연결망의 경우 고속 입출력 채배(multiplex) 전송 기능을 제공한다. 상호연결망 비교에 있어서 가장 중요한 요소가 전송 대역 폭과 지연시간이다. Intel TFLOPS 시스템의 경우 가장 빠른 40 ns 의 전송 지연 시간을 제공한다. IBM RS6000 SP2 시스템의 경우

<표 1 : 상호연결망 및 라우터 특성 비교>

	CRAY T3E	Intel TFLOPS	Tandem Himalaya	IBM RS6000	ETRI SPAX
연결 구조	3D Torus	Dual 2D Mesh	Flexible	Multi-stage	HierCrossbar
연결망 명칭	-	-	Server Net	SP Switch	Xcent-Net
노드연결포트	Single Port	Single Port	Dual Port	Single Port	Dual Port
CC / MP	Message Passing	Message Passing	Message Passing	Message Passing	Message Passing
최대 연결	2048 노드	4608 노드	255 노드	512 노드	128 노드
스위치	7 X 7	6 X 6	6 X 6	4 X 4	10 X 10
내부	75 MHz	200 MHz	50 MHz	75 MHz	67 MHz
입출력	375 MHz	200 MHz	50 MHz	150 MHz	67 MHz
신호선	13 bits	16 bits	8 bits	8 bits	32 bits
노드당 대역폭	600 MB/s	400 MB/s	100 MB/s	150 MB/s	528 MB/s
스위치 Delay	40 / 80 ns	50 ns	300 ns	300 ns	90 ns
주요 기술	HighSpeed IO	HighSpeed Clk	General Purpose	Global Clock	Byte Slice
구현 기술	CMOS/ LVDS	BiCMOS	CMOS/ TTL	CMOS	CMOS/ TTL

전역 클럭을 제공하여 전역 동기 상호연결망을 지원하고, ETRI Xcent-Net 은 바이트 슬라이스 개념을 이용하여 가변 데이터 폭 구조를 지원한다.

V. 결 론

메시지 전송을 기반으로 하는 병렬 처리 시스템의 상호연결망 구조와 함께 라우팅 스위치 구조 및 특성을 중심으로 비교하였다. 특히 ETRI SPAX 시스템의 Xcent-Net 을 다른 상용 병렬 처리 시스템의 상호연결망과 상호 비교하므로써 노드당 전송 대역폭과 지연 시간 특성이 비슷하거나 우수함을 알 수 있었다. 현재 구현된 Xcent-Net 은 목표 규격을 상회하여 최대 512 노드까지 연결이 가능하며 최대 720 Mbytes/s 의 전송 대역폭과 70 ns 이내의 지연 시간을 제공할 수 있다.

참고 문헌

[1] S.L. Scott, et. al., ``The Cray T3E Network : Adaptive Routing in a High Performance 3D Torus", Hot Interconnect IV, pp. 147-156, Aug. 1996.

[2] J. Carbonaro, et. al., ``Cavallino : The Teraflops Router and NIC", Hot Interconnect IV, pp. 157-160, Aug. 1996.

[3] W.E. Baker, et. al., ``A Flexible ServerNet based Fault-Tolerant Architecture", Int'l Symp. On Fault-Tolerant Computing, pp. 27-30, Jun. 1995.

[4] C.B. Stunkel, et. al., ``The SP2 High-Performance", IBM System Journal, Vol. 34, No 2, 1995.

[5] J.S. Han, et. al., ``Xcent-Net Interconnection Network for a High Performance Computing Server", Int'l Conf. On High Performance Computing, Poster Session, Dec. 1997.