□ 기술해설 □

상용병렬처리 시스템

상용 대규모 병렬처리 컴퓨터의 시스템 구조

서울시립대학교 김 기 철*

● 목 차 ●

1. 서 언

본 고에서는 근자에 개발되어 많은 학자들의 관심을 끌고 있는 상용 대규모 병렬처리 컴퓨터들을 소개하고자 한다. 소개되는 시스템들은 상업적인 측면에서 비교적 성공적인 회사의 시스템이며, 그 시스템이 속하는 아키텍쳐나 그 시스템이 사용되는 응용 분야에서 대표적인 시스템이라고 여겨지는 시스템들이다. 또한 각 사의 최신 모델을 중점적으로 소개하려고 노력하였다. 소개하려는 시스템들과 그 특징은 다음과 같다. 즉,

1) MP-1, MP-2: 2-D 메쉬 SIMD 시스템
2) NCR MPP: 대규모 데이타베이스 시스템
3) Intel Paragon: 2-D 메쉬 MIMD 시스템
4) nCUBE-2, nCUBE-3: 하이퍼큐브 MIMD 시스템
5) Kendall Square Research KSR1: 계층구조의 링구조, All-cache 시스템
6) Cray Research T3D: 슈퍼컴퓨터 기술과 Alpha 칩 사용, 3-D 메쉬 MIMD
7) Tera Computer: 3-D 메쉬의 파이프라인형 패킷 스위치채용 MIMD

등이다. 이중에서 Tera Computer의 3D는 상업적으로 성공하지 못하였으나 그 시스템이 미친 영향력을 고려하여 선정하였다. 본 고에서 선정한 시스템이외에도 Thinking Machine사의 CM-5, Tandem사의 Himalaya 및 Meiko, Parsytec GmbH의 시스템들도 있으나, 지면관계상, 그리고 저자의 자료 부족으로 본 고에서는 소개하지 않고, 다른 분들에게 남겨놓는다.

본 고에서는 상기의 시스템을 소개하는 데 있어서 인터코넥션 네트워크를 중심으로한 시스템 아키텍쳐 위주로 소개할 것이다. 또한 사용된 VLSI 기술, 패키징 기술 및 소프트웨어에 관련된

*정회원

사항도 필요시에는 소개하였다.

2. MasPar

MasPar에서는 메쉬 토폴로지를 가진 SIMD 시스템인 MP-1을 1990년부터 판매하였다. 1991년에 Digital Equipment Cooperation과 OEM 관계를 맺었으며 1992년에 MP-1과 이진 호환성이 있는 MP-2를 판매하기 시작하였다. MasPar의 시스템들은 대규모 데이타 병렬성을 통한 성능향상을 추구하며 전통적인 반도체 기술과 패키징 기술을 사용하고 있다.

2.1 MasPar MP-1

MP-1 아키텍처의 기본개념은 Thinking Machine사의 CM-1, CM-2와 매우 유사하다. 즉, MP-1은 CM-1이나 CM-2와 마찬가지로 데이타 병렬성을 SIMD방식으로 추구하며, 데이타 통신에 많은 자원을 할애한다. 또한 회로설계보다는 많은 프로세서를 연결한 아키텍처의 특성을 토대로 성능향상을 도모하고 있다. MP-1의 주요 특징은 다음과 같다.

1) 프로세서의 수: 1024-16384
2) 정수연산: 30 GIPS
3) 실수연산
   단정도: 1.5 GFLOPS
   배정도: 650 MFLOPS
4) 동작방식: SIMD

MP-1의 Array Control Unit(ACU)은 Processor Array에서의 연산을 지시하며 자체적으로 연산을 수행하기도 한다. ACU는 32 bit, harvard-style RISC형 아키텍처이다. Processor array를 위한 microcoded engine을 뺀 모든 스칼라 인스트럭션은 한 클럭이 70 nsec이며, 한 클럭당 하나의 인스트럭션을 수행한다. 따라서 전체적으로 14 MIPS의 성능을 가지고 있다. Processor Array는 1-16개의 프로세서 보드로 되어 있으며 하나의 프로세서 보드는 64개의 클러스터로 되어 있다. 하나의 클러스터는 16개의 프로세서와 프로세서 메모리로 구성되어 있다. 따라서 하나의 프로세서 보드에는 1024개의 프로세서가 있으며 16개의 프로세서 보드를 사용하면 16384개의 프로세서를 탑재한 시스템이 된다. 또한, 그림 1과 같이 16개의 프로세서를 하나의 클러스터로 구성하고 각각의 프로세서를 선형배열(linear array)형태로 구성하고 있으나, 사용자에게는 4×4의 이차원 배열로 나타난다.

하나의 MP-1 프로세서 칩은 1.6 미크론의 이층 메탈 CMOS로 제작되며 칩당 32개의 프로세서(2개의 클러스터)를 내장하고 있다. MP-1 프로세서 칩의 클럭속도는 ACU와 마찬가지로 70 nsec이

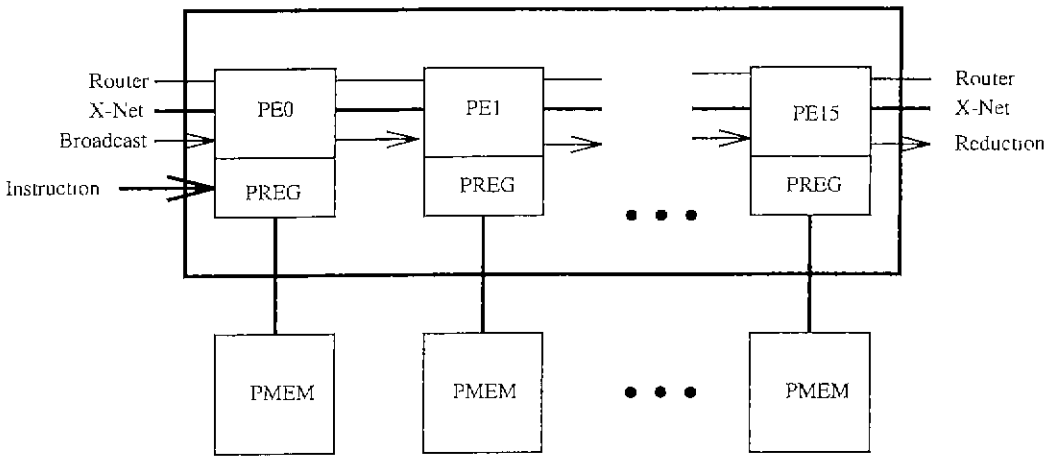

그림 1 MP-1의 클러스터 구성

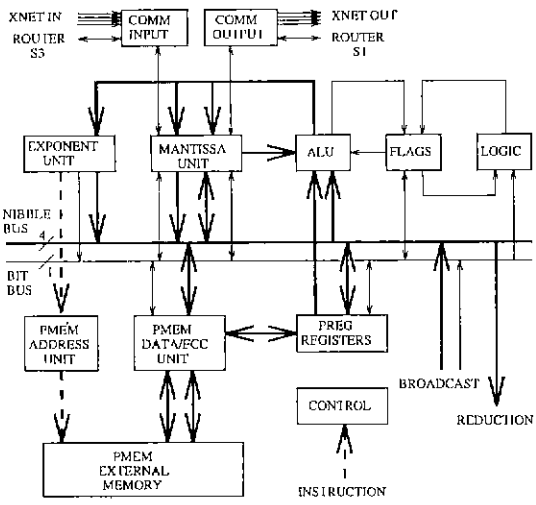

그림 2 MP-1의 프로세서 구조

다.

PE와 메모리의 구조는 그림 2와 같다. MP-1은 SIMD를 따르므로 PE는 인스트럭션 훼치 로직(instruction fetch logic)과 인스트럭션 디코딩 로직(instruction decode logic)을 가지고 있지 않으며, 데이타 경로 로직 (data path logic)만을 가지고 있다. MP-1의 PE는 CM-1과는 달리 64 비트 하드웨어를 사용함으로써 연산 능력을 크게 향상하였다. MP-1을 구성하는 PE의 각 유닛의 데이타 폭(data width)은 다음과 같다.

1) 가수부: 64 bit
2) 지수부: 16 bit
3) ALU: 4 bit
4) 논리 및 플래그 비트: 1 bit

MP-1의 또다른 특징은 내부에 커다란 레지스터를 가지고 있으며, 모든 연산은 레지스터에서 수행된다. 사용자에게는 40개의 32 비트 레지스터가 존재하는 것으로 투영되며, 이와는 별도로 내부기능 수행용으로 8개의 32 비트 레지스터가 있다. 각 레지스터는 비트 또는 바이트 단위로 이용될 수 있다. PE 내부에서의 데이타 전송은 4 비트의 니블 버스(nibble bus)와 비트 버스(bit bus-1 bit)를 통하여 이루어진다.

칩 내부에 커다란 레지스터 세트를 가짐으로 하여 MP-1은 PE와 PREG 사이에 높은 데이타 전송율을 유지하고 있다. 16 K의 PE를 가진 시스템의 경우 PE와 PREG 간의 대역폭은 117 Gbyte/sec에 이른다.

MP-1은 내부적으로 3개의 인터코넥션 네트워크를 가지고 있다. 이중 OR-reduction tree는 전체 Processor Array의 상태를 ACU가 알게 하기 위하여 사용되며 X-Net과 Multistage Clos Network은 데이타 전송을 위하여 사용된다.

X-Net은 모든 프로세서를 주위 8개의 프로세서와 연결시킨다. 모든 프로세서에서는 대각선 방향으로 4개의 연결선이 나와 있으며, 연결선이 만나는 각 X 교점에는 트라이 스테이트 노드(tri-state node)가 있어 주위 8개 프로세서와의 통신을 가능하게 한다. 모든 PE는 동일한 방향으로 통신을 수행한다. 통신방법은 비트 직렬(bit-serial)방식이며 PE의 클럭과 동기되는 통신 방식을 따른다. 16 K PE 시스템의 경우 X-Net의 사용시 총체적 통신율은 20 Gbyte/sec를 넘을 수 있다.

Multistage Clos Network은 PE-to-PE 통신을 위하여 사용되며 또한 MP-1 입출력 시스템의 기본이 된다. Multistage Clos Network은 S1, S2, S3의 세개의 단계로 구성되었으며 1024×1024의 크로스바의 역할을 한다. 16개의 PE로 구성된 클러스터마다 하나의 포트가 S1에 연결되어 있으며 또 다른 하나의 포트가 S3에 연결되어 있다. 통신방식은 비트 단위의 동기식 회로 스위칭(circuit switching)방식이다. Multistage Clos Network에 사용되는 라우터 칩은 64×64의 크로스바 스위치이며 목적지 번지를 디코딩하여 입출력을 연결한다. Multistage Clos Network의 대역폭은 16 K PE의 경우 1.5 Gbyte/sec이다. PE 배열의 입출력은 Multistage Clos Network의 마지막 단계를 입출력 장치나 입출력 메모리에 연결함으로써 이루어 진다.

2.2 MasPar MP-2

MP-2는 MP-1과 이진 호환성을 가지는 시스템으로서 다음과 같은 성능을 가지고 있다.

1) 프로세서의 수: 1024-16384
2) 정수연산: 68 GIPS
3) 실수연산

단정도: 6.3 GFLOPS
배정도: 2.4 GFLOP
4) 동작방식: SIMD

MP-2는 MP-1과 동일한 아키텍쳐를 따르고 있으며 ACU와 프로세서 칩의 재설계를 통하여 시스템의 성능을 향상하고 있다. 예를 들면, MP-2의 시스템 클럭이 비록 MP-1의 시스템 클럭인 70 nsec보다 느린 80 nsec이나, Barrel Shifter Unit을 사용하고 있으며, MP-1이 4비트 버스를 사용하는 대신 MP-2는 32비트의 버스를 사용하므로 전체적으로 성능을 두배로 향상하고 있다.

3. NCR MPP

Teradata Cooperation은 DBC/1012라는 대규모 병렬처리를 이용한 대용량 데이타베이스 시스템을 제작 판매하여 1991년에 2억2천만불의 수익을 올린 회사이다. NCR은 Teradata를 합병하여 DBC/1012를 개량한 NCR 3600 시스템을 판매하였으며 1993년에 새로운 인터코넥션 네트워크를 채택한 NCR 3700 시스템을 개발한 바 있다.

3.1 Teradata DBC/1012

DBC/1012의 'DBC'는 DataBase Computer의 약자이며, '1012'는 10의 12승 즉, 용량이 terabyte 단위임을 의미한다. DBC/1012는 LAN 또는 대형컴퓨터에 연결되어 데이타베이스 엔진의 역할을 수행한다. DBC/1012의 주요 특징은 다음과 같다.

1) 프로세서의 수: 최대 1024
2) 프로세서: 8086, 80386
3) 동작방식: MIMD
4) 통신방식: 메세지 통신(message passing)
5) 인터코넥션 네트워크: Ynet Bus(binary tree)

DBC/1012의 가장 큰 특징은 Ynet Bus에 있다. 여기서 'Bus'라는 용어를 사용하였으나 실제로 Ynet Bus는 이진 트리(binary tree) 인터코넥션 네트워크를 사용하고 있다. Ynet Bus의 각 노드(internal node)는 PE를 포함하지 않으며, Ynet의 말단(leaf node)에만 PE가 연결된다. Ynet Bus의 각 노드는 6 MHz의 단일 클럭에 동기되어 동작하며 메시지의 전달과 간단한 데이타 베이스 레코드를 다루는 작업을 수행한다. Ynet 노드가 제공하는 메시지의 전달에는 point-to-point, 방송(broadcast), 다중방송(multicast)이 있으며 데이타 베이스 레코드를 다루는 작업으로는 선택(selecting), 소트(sorting), 머지(merging) 등이 있다.

Ynet의 말단은 Interface Function Processor (IFP), Communication Processor(COP) 또는 Access Module Processor(AMP)가 있으며 8086 계열의 CPU로 구성되어 있다. COP에는 하나의 연결단자가 있으며 IFP는 SQL을 primitive step으로 바꾸는 역할을 한다. 각각의 AMP에는 4개의 디스크 장치가 연결되어 있다. AMP는 명령을 접수하면 각 디스크 장치에서 필요로 하는 연산을 수행시켜 그 결과를 IFP나 COP에 전송한다.

3.2 NCR 3600

NCR 3600 시스템은 Teradata의 DBC/1012를 개선시킨 시스템이다. 하드웨어의 주된 개선점은 i486-50을 이용한 다중프로세서 보드를 사용했다는 것이다. NCR 3600 시스템은 DBC/1012와 같이 Ynet을 사용하며 Oracle이나 Teradata의 두가지 데이타 베이스 시스템을 구성하고 있다. Oracle 구성에서는 Ynet의 말단 프로세서로 응용 프로세서(AP: Application Processor) 그룹(subsystem)만을 사용하며, Teradata 구성에서는 AP, 파싱엔진(PE: Parsing Engine) 및 AMP를 사용한다. 여기서 AP 그룹(subsystem)은 하나 또는 두개의 AP 노드로 구성되며, AP노드의 구성은 그림 3과 같다.

응용 프로세서 노드는 2~8개의 i486-50 CPU를 사용하며 2개의 서로 독립된 64 비트 프로세서/메모리 버스로 구성된 이중 시스템 버스를 가지고 있다. 이중 시스템 버스는 최대 25MHz에서 400 Mbyte/sec의 성능을 제공한다. AP 노드의 메모리 시스템은 양방향, 4×1 인터리빙 메모리로 구성되어 있다.

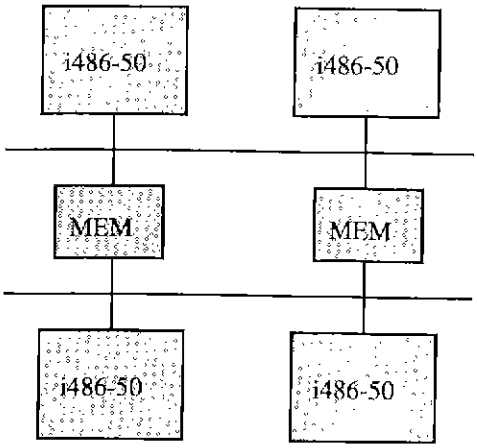

그림 3 NCR 3600 시스템의 응용 프로세서 노드의 구성

3.3 NCR 3700

NCR 3700 시스템은 4096개의 프로세서까지 확장가능한 성능을 제공하기 위하여 만들어진 시스템이다. NCR 3700 시스템은 메세지 전송 MIMD로 동작하며 BYNET라는 새로운 인터코넥션 네트워크를 사용하고 있다. BYNET는 다음과 같은 특징을 가지고 있다. 즉,

1) 개선된 뱐얀 네트웍(folded Bynan topology)의 채용
2) 4096 프로세서까지 확장가능
3) 화이버 와틱 채널구성
4) 노드당 20 Mbyte/sec 의 속도
5) 네트웍의 결함에 따른 네트웍 재구성 가능

등이다.

4. Intel Paragon XP/S

Intel의 Paragon 시스템은 Touchstone 프로젝트의 Delta 시스템을 상용화한 것이다. Touchstone Delta 시스템은 1991년 11월에 Caltech에 설치되었으며 528개의 프로세서를 이용하여 Linpack 벤치마크에서 13.9 GFLOPS의 성능을 발휘하여 1992년 봄까지 세계기록을 유지하였던 시스템이다. Paragon XP/S 시스템은 300 GFLOPS까지 확장가능한 성능을 제공한다. Intel Paragon XP/S의 시스템 개요는 다음과 같다.

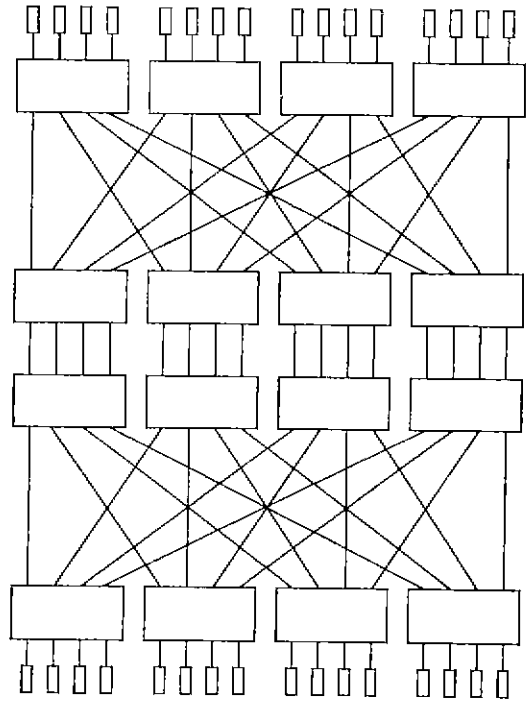

그림 4 NCR 3700 시스템의 BYNET

1) 주요 사양
   (1) 최대 실수연산성능(배정도): 5~300 GFLOPS
   (2) 최대 정수연산성능: 2.8~160 GIPS
   (3) 노드간 메세지 라우팅 성능: 200 M byte/sec (full duplex)
   (4) 메인 메모리: 1~128 Gbyte
   (5) 디스크 용량: 6 Gbyte~1 Tbyte
2) 시스템 아키텍쳐
   (1) 확장성이 뛰어난 분산메모리 시스템
   (2) 2-D 메쉬 토폴로지
   (3) 파이프라인방식으로 운영되는 하드웨어 통신장치
3) 소프트웨어 환경
   (1) 완벽한 UNIX(POSIX, System V.3, 4.3 bsd) 지원
   (2) C, FORTRAN, Ada, C++, Data-parallel FORTRAN

Paragon 시스템은 프로세싱 노드들을 2-D 메쉬를 이용하여 연결하였다. 각 노드들은 I/O, 운영체제 서비스, 계산 등의 기능을 수행한다. Pa-

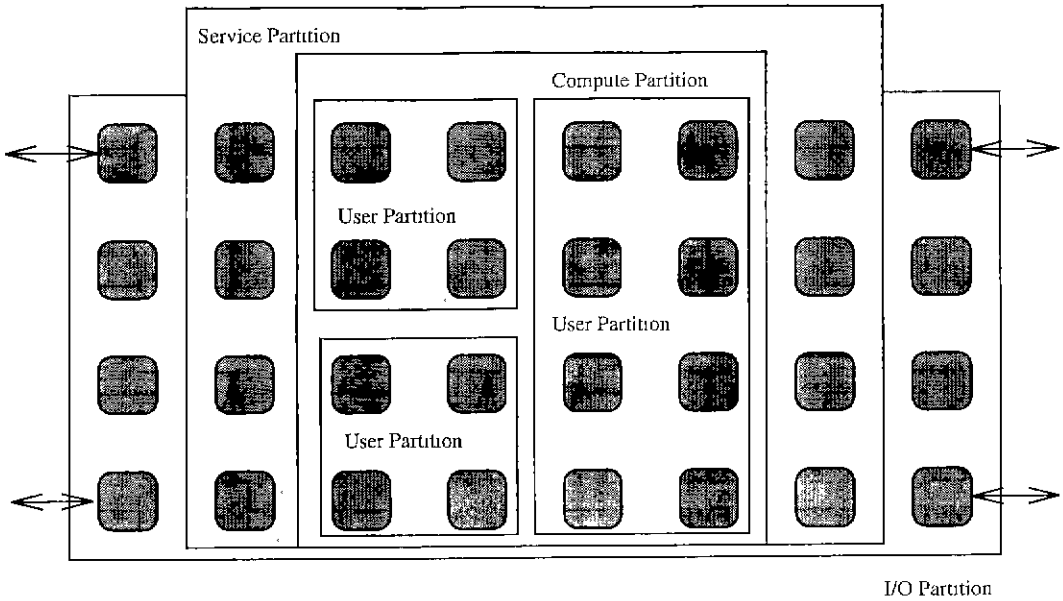

그림 5 Paragon XP/S의 시스템 분할

ragon 시스템의 특징중의 하나는 다양한 시스템 분할이 가능하다는 것이다. 프로세싱 노드들은 각각 I/O, 운영체제 서비스, 계산 등의 여러가지 작업을 위하여 분할될 수 있다. 그림 5는 시스템 분할의 한 예이다.

프로세싱 노드의 주요 구성요소는 두개의 50 MHz i860 XP 마이크로프로세서와 16~128 Mbyte의 메모리, 그리고 Network Interface Controller(NIC)이다. 두개의 i860 XP중에서 하나는 응용 프로세서 (AP: Application Processor)로 사용되며 하나는 메세지 프로세서 (MP: Message Processor)로서 NIC와 함께 통신을 전담한다. Intel은 차세대 Paragon 시스템에는 다중 프로세서 노드를 사용할 예정이다. 그림 6에 Paragon XP/S의 프로세싱 노드의 블럭도가 나타나 있다.

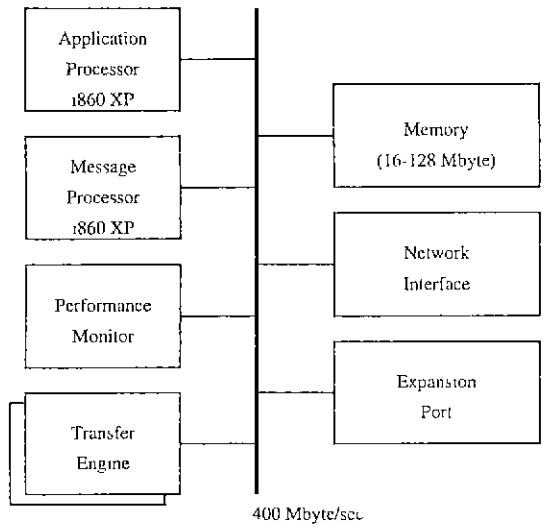

그림 6 Paragon XP/S의 프로세싱 노드

Paragon 시스템의 인터코넥션 네트워크는 그림 7에 보인 바와 같은 2-D 메쉬이며 노드간 메세지 라우팅시 전송률이 200 Mbyte/sec이며, 전송 지연시간 (trasmission latency time)이 25 microsec이다. 2-D 메쉬의 각 교점에는 Paragon Mesh Routing Chip(PMRC)이 있다. PMRC는 0.75 미크론의 삼중 메탈(triple metal) CMOS 기술로 제작되었으며 40 nsec안에 각각의 입력 메시지에 대한 라우팅을 결정하고 이를 수행할 수 있다. 각 노드와 PMRC 사이의 통신은 노드 안에 있는 Network Interface Controller(NIC)가 수행한다. NIC는 PMRC간에서 패리티 검사가 가능한 full duplex 방식으로 통신을 수행하며

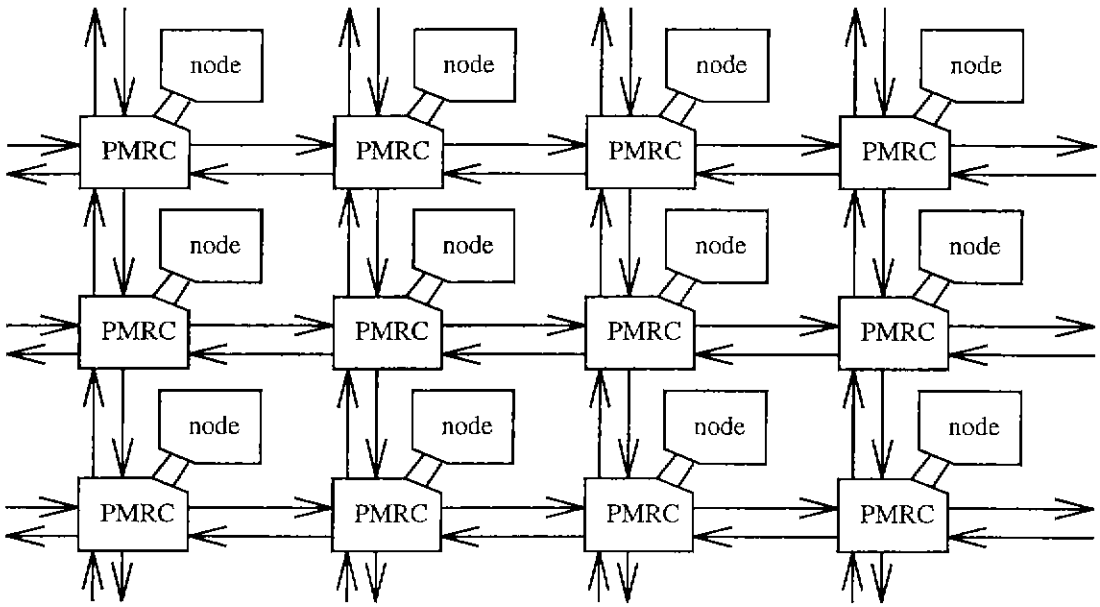

그림 7 Paragon XP/S의 2-D 메쉬 인터코넥션 네트워크

각각의 메세지 전송에 대해서 메세지별로 에러 체크를 수행한다. 200 Mbyte/sec에 달하는 NIC의 전송율을 유지시키기 위하여 각 노드에는 두 개의 Block Transfer Engine이 있다. Block Transfer Engine은 한번에 4 Kbyte의 데이타를 주고받을 수 있다. 각 노드의 메세지 프로세서(MP)로는 하나의 i860 XP 프로세서가 사용된다. MP-입력 메시지를 수신하여 AP가 사용할 수 있게 준비하며, 프로토콜과 메세지의 패킷화 등을 포함하는 메세지 전송과 관련된 작업을 수행한다.

5. nCUBE

nCUBE는 1980년대 초부터 하이퍼큐브 시스템을 판매해온 회사이다. 초기의 nCUBE/10 시스템으로부터 1989년부터 판매해온 nCUBE-2 그리고 최근에 발표된 nCUBE-3에 이르기까지 nCUBE의 시스템의 아키텍쳐에는 크게 변한 바가 없다. 그러나 nCUBE는 VLSI 기술과 패키징 기술 등을 잘 이용하여 꾸준히 경쟁력있는 시스템을 개발하여 왔다.

5.1 nCUBE-2

nCUBE-2는 1989년에 발표된 제 2세대 하이퍼큐브 시스템으로 nCUBE의 VLSI 기술이 잘 나타나 있는 시스템이다. nCUBE는 ALU와 FPU는 물론 MMU(Memory Management Unit)와 라우팅 기능을 하나의 칩에 집적시킴으로서 경쟁력있는 하이퍼큐브 시스템을 제작하는데 성공하였다. nCUBE-2의 주요 사양은 다음과 같다.

1) 프로세서 수: 8~8192(13-D 하이퍼큐브)
2) 최대성능: 27 GFLOPS, 60 GIPS
3) 메인 메모리: 최대 32 Gbyte
4) 프로세서 메모리간 대역폭: 655 Gbyte/sec
5) 프로세서간 대역폭: 252 Gbyte/sec
6) 동작모드: MIMD
7) 토폴로지: 하이퍼큐브

nCUBE-2 노드 프로세서는 다음과 같은 기능을 가지고 있다.

1) 64비트 VAX형 CPU
2) IEEE-인증 FPU
3) 14개 DMA 채널: 13개는 하이퍼큐브 연결에, 1개는 입출력에 사용
4) 메모리 관리 장치(memory management unit)

5) 라우팅 장치(routing hardware)
6) 7.5 MIPS, 3.3 MFLOPS

하나의 nCUBE-2 노드 프로세서는 1~64 Mbyte의 메인 메모리를 가질 수 있다. nCUBE-2의 구성은 노드 모듈을 기본으로 하고 있다. 노드 모듈은 노드 프로세서와 메모리로 구성되어 있으며 2.5 cm×8.25 cm의 크기이다. 하나의 프로세서 보드는 64개의 노드 모듈을 가진 6-D 하이퍼큐브로 되어 있다. 하나의 마더보드는 프로세서 보드를 위한 슬롯을 16개 그리고 입출력 카드를 위한 슬롯을 8개 가지고 있어 10-D 하이퍼큐브를 구성할 수 있다. nCUBE-2는 8개의 마더보드를 사용함으로서 최대 크기인 13-D 하이퍼큐브까지 구성할 수 있다. 하나의 입출력 카드에는 16개의 입출력 프로세서가 있으며 하나의 입출력 프로세서는 8개의 노드 프로세서에 입출력 서비스를 제공하게 되어 있다. nCUBE-2의 인터코넥션 네트워크는 잘 알려진 대로 하이퍼큐브이다. 라우팅 방법으로는 wormhole 라우팅 방법을 쓴다. nCUBE-2의 하이퍼큐브의 가장 큰 특색은 13-D 하이퍼큐브까지 구성이 가능하면서도 하나의 DMA 채널이 2.75 Mbyte/sec의 큰 대역폭을 가지는 점이다. nCUBE-2는 각각의 노드에 nCX 운영체제를 사용하고 있다. nCX는 128 Kbyte의 메모리만을 사용하며 다음과 같은 기능을 제공한다.

1) 프로세스 관리
2) 지역 메모리 관리
3) 메세지 전송
4) 멀티 태스킹
5) UNIX SVR4 시스템 콜 인터페이스
6) POSIX 신호지원

5.2 nCUBE-2S

nCUBE-2S는 새로운 VLSI 기술을 사용하여 nCUBE-2의 성능을 향상시킨 시스템으로서 nCUBE-2와 소프트웨어 호환성을 가지며 다음과 같은 사양을 가지고 있다.

1) 프로세서 수: 8~8192(13-D 하이퍼큐브)
2) 최대성능: 34 GFLOPS, 123 GIPS
3) 메인 메모리: 최대 32 Gbyte
4) 프로세서 메모리간 대역폭: 819 Gbyte/sec
5) 프로세서간 대역폭: 577 Gbyte/sec
6) 동작방식: MIMD
7) 토폴로지: 하이퍼큐브

5.3 nCUBE-3

1993년에 nCUBE는 nCUBE-2 계열의 후속으로 nCUBE-3를 발표하였다. nCUBE-3는 nCUBE-2에 비해 보다 진보한 VLSI 기술, 패키징 기술, 냉각 기술을 사용하며 주요 사양은 다음과 같다.

1) 프로세서 수: 8~65536(16-D 하이퍼큐브)
2) 성능: 6.5 TFLOPS(배정도 실수연산), 3 TIPS(64 비트 정수연산)
3) 메인 메모리: 65 Tbyte
4) 총체적 데이타 전송율: 24 Tbyte/sec

nCUBE-3의 프로세서는 0.6미크론, 3-층 메탈 CMOS 기술을 사용하여 제작되며 300만개의 트랜지스터를 가지고 있다. nCUBE-3 프로세서의 동작 속도는 50 MHz이다. nCUBE-3의 프로세서에는 메모리를 제외한 모든 프로세서 노드가 내장되어 있으며 주요 유닛들의 사양은 다음과 같다.

1) ALU: 64 bit, 가상메모리관리기능 보유, 50 MIPS
2) FPU: 100 MFLOPS
3) 인스트럭션 캐쉬, 데이타 캐쉬: 16 Kbyte, 2-way set associative cache

하나의 nCUBE-3 프로세서는 16 Mbyte에서 1 Gbyte까지의 메모리를 사용할 수 있으며 48비트의 주소영역에 256 Tbyte의 요구 페이지(demand paged) 방식의 가상 메모리를 가질 수 있다.

nCUBE-3의 인터코넥션 네트워크는 하이퍼큐브이다. nCUBE-3는 평균 처리율(average throughput rate)을 향상시키고 24 Tbyte/sec의 데이타 전송율을 제공하기 위하여 다음과 같은 특색을 가지고 있다.

1) 메세지 전송 개시시의 지연시간: 5 micro-sec 이하
2) 라우팅 방식 : cut-through 라우팅, 200 nsec/hop
3) DMA 채널 버퍼
4) 50 Mbyte/channel, full duplex
5) 16채널의 통신선로

nCUBE-3는 ParaChannel이라는 새로운 입출력 시스템을 사용하고 있다. 각각의 nCUBE-3 프로세서는 하이퍼큐브 연결에 사용되는 16개의 DMA 채널외에 두개의 독립된 입출력 채널을 가지고 있다. 이 두개의 입출력 채널은 ParaChannel 보드의 입출력 노드에 직접 연결되어 있으며, full duplex 20 Mbyte/sec의 입출력을 각각의 프로세서에 제공한다. 입출력 시스템은 최대 1024개의 ParaChannel을 가질 수 있으며 하나의 ParaChannel은 2.5 Gbyte/sec의 전송율을 가지고 있다. 하나의 ParaChannel I/O Array는 16-노드의 하이퍼큐브로 되어 있으며 400개 이상의 디스크를 연결할 수 있다. 또한, nCUBE-3는 대부분의 경우에 일반적인 패키징 기술과 공기 냉각 방식을 사용한다. 하지만 TFLOPS의 성능을 내는 시스템들은 다음과 같은 진보적인 기술을 사용한다.

1) 스택방식 메모리(stacked memory)기술
2) multi-chip modules
3) 열전도식 또는 액체 쿨링방식

6. Kendall Square Research KSR1

Kendall Square Research는 모든 메모리가 캐쉬만으로 이루어진 매우 특이한 시스템인 KSR 1을 판매하고 있다. KSR1의 가장 큰 특징은 물리적으로 분산되어 있는 메모리를 사용자가 하나의 공유메모리처럼 사용할 수 있다는 것이다. KSR1에서는 데이타가 하나의 특정 메모리에 고정되어 있지 않고 새로운 프로세서에 사용되면 그 프로세서의 메모리로 이동한다. 전체 메모리는 하나의 계층적 캐쉬로 작용한다. KSR1의 인터코넥션 네트워크는 계층구조의 링(ring)이다. 하나의 링은 32개의 연산소자(PE: processing element)를 위한 슬롯과 두개의 라우팅 셀을 위한 슬롯을 가지고 있다. 두개의 라우팅 셀은 각각 상위 링과의 uplink와 downlink를 위하여 사용된다. 연산소자는 최하위 링인 ring : 0에 연결된다. n-번째 계층인 ring : n에는 하위의 링인 ring : n-1이 연결된다. 이러한 구조는 링의 계층을 하나 증가시킴으로 시스템에 포함되는 프로세서의 수를 32배로 확장할 수 있다. 그림 8에 세개의 계층을 가지는 KSR 시스템이 나타나 있다. 현재의 KSR1 시스템은 두개의 계층만을 사용하며 ring : 1의 모든 슬롯에 ring : 0를 연결시켜 1088(32×38)까지의 프로세서 셀을 포함한다.

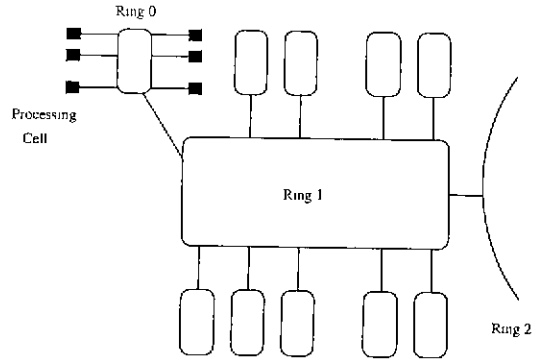

그림 8 계층구조의 링으로 구성된 KSR

프로세서 셀은 12개의 주문형 CMOS 칩으로 제작되었으며 다음과 같은 주요 요소로 구성되어 있다.

1) Co-Execution Unit(CEU): instruction fetch, data fetch and store control, address calculation등을 수행한다.
2) 정수처리부(IPU: Integer Processing Unit)
3) 실수처리부(FPU: Floating Point Unit)
4) I/O장치부(XIO: eXternal Input/Output Unit): DMA와 프로그램에 의한 입출력을 담당
5) 캐쉬제어부(CCU: Cache Control Units): 프로세서 셀에 있는 부 캐쉬(subcache)인 0.5 Mbyte의 캐쉬와 32 Mbyte의 지역 메

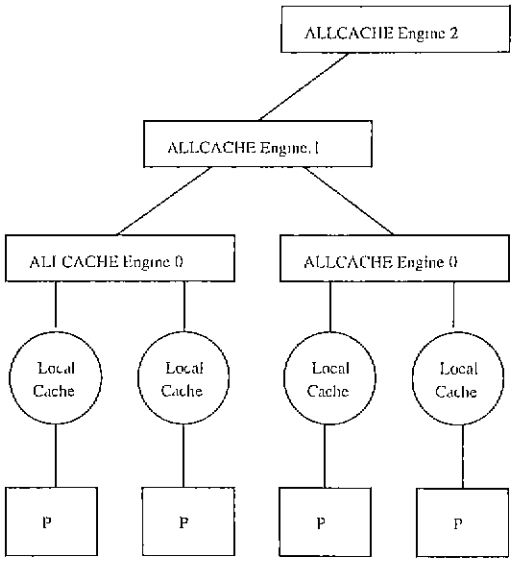

그림 9 KSR의 ALLCACHE 엔진

모리와 지역 캐쉬(local cache)간을 접속하는 장치로 4개의 유닛이 있음.

6) 셀 접속 장치(CIU: Cell Interconnect Units): 프로세서(processing cell)와 최하위 계층인 ring : 0 사이의 인터페이스를 담당하며, 모두 4개의 장치가 존재함.

프로세서에는 256 Kbyte의 인스트럭션 캐쉬와 256 Kbyte의 데이타 캐쉬가 있다. 프로세서의 캐쉬를 부캐쉬(subcache)라고 한다. 프로세서와 연결된 작은 기판에는 32 Mbyte의 메모리가 실장되며, 이를 지역 캐쉬(local cache)라고 한다. Ring : 0에는 32개의 프로세서외에도 두개의 ARD(ALLCACHE Routing and Directory)셀이 있다. 두개의 ARD 셀은 ring : 1과의 uplink와 downlink를 위하여 사용된다. 모든 지역 메모리는 인터코넥션 네트워크와 함께 ALLCACHE 메모리 시스템을 구성한다. 그림 9에 ALLCACHE 엔진의 계층적 구조가 나타나 있다. ALLCACHE 메모리 시스템은 사용자에게 프로그램과 데이타를 위하여 64비트의 균등한 번지 영역(uniform address space)를 제공할 수 있으나 실제로는 40비트의 번지 영역만이 사용된다. 이 메모리 공간을 SVA(System Virtual Address Space)라고 한다. SVA의 내용과 장소는 분산되어 저장되고 있으며 ALLCACHE 메모리 시스템은 그림 10에서와 같이 밀집도가 크며, 공유 가능한 번지 영역을 가지는 프로그래밍 모델을 가능하게 한다.

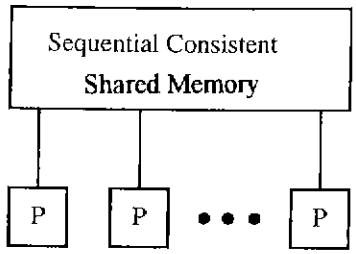

그림 10 사용자가 본 KSR의 ALLCACHE 메모리 시스템

7. Cray Research T3D

Cray Research사는 1992년부터 미국 에너지성과 3년 예정의 7천만불짜리 프로젝트를 수행하고 있다. Cray-DoE 프로젝트의 주내용은 코드네임이 T3D인 3대의 MPP를 제작하여 Livermore National Laboratory, Los Almos National Laboratory, 그리고 Cray Research에 설치한다는 것이다. 이 3대의 T3D는 고속 네트워크로 연결되어 상기한 세 기관중 어디에서도 다른 기관의 시스템을 충분히 사용할 수 있게 되어 있다. Cray Research사는 이미 T3D 시스템을 공급하는 데 성공하였으며 7개 시스템의 주문을 받아 놓은 상태이다[8].

T3D는 DEC의 Alpha 칩을 사용하며 인터코넥션 네트워크 토폴로지는 3차원의 torus를 사용하고 메모리는 물리적으로는 분산되어 있으나 논리적으로는 하나의 공유메모리를 구성한다. T3D는 32개에서 2048개 까지의 Alpha 칩을 사용할 수 있다. T3D는 Cray Research의 슈퍼컴퓨터에 연결된 후위 시스템(back-end system)으로 사용되며 단독 (stand-alone) 시스템으로의 사용도 가능하다. T3D의 내용중 가장 흥미있는 것은 패키징 기술이 될 것이다. Cray Research는 T3D에 Cray Y-MP, C90의 패키징, 냉각, 그리고 입출력 시스템 기술을 사용하였다. 이는 T3D가

8×8×8의 크기인 경우에 개당 23 W를 소비하는 150 MHz Alpha 칩을 사용하면 프로세서에서만 약 11.8 KW의 전력이 소비된다는 사실을 생각할 때 당연하다. 이러한 패키징 기술과 냉각 기술은 보수적인 회로와 패키징 기술을 사용한 CM-1, CM-2, 그리고 MasPar 시스템들과 크게 비교되며 그동안 비교적 보수적인 기술들을 사용하던 병렬처리 시스템들에도 슈퍼컴퓨터에 사용되던 진보적인 기술들이 사용되기 시작함을 보여주는 것이다.

8. Tera Computer 3D

Tera Computer 사는 3-D 메쉬 인터코넥션 네트워크와 파이프라인 방식의 패킷 스위칭기법을 이용한 공유메모리 방식의 병렬처리 컴퓨터의 개발을 추진하였었다. 일반적으로 공유메모리 방식의 병렬 컴퓨터에서는 크로스바 스위치나 다단계 인터코넥션 네트워크 등을 사용하며 프로세서와 메모리의 거리가 일정하나, Tera Computer에서는 프로세서와 메모리가 서로 섞이어 분산되어 있으며 프로세서와 메모리간의 거리가 일정하지 않다. Tera Computer의 주요 특징은 다음과 같다.

1) 프로세서: 256 VLIW 프로세서
2) 메모리 유닛: 512 (각각 128 Mbyte)
3) 입출력 프로세서: 256
4) 입출력 캐쉬 유닛(I/O cache unit): 256
5) 인터코넥션 노드: 4096
6) 클럭 주기: 3 nsec

Tera Computer의 인터코넥션 네트워크는 3-D toroidal mesh이며 각 노드는 파이프라인 방식의 패킷 스위칭 기능을 수행한다. 각 링크는 한 클럭내에 전송 번지, 목적지 번지, 연산명령어 및 64비트의 데이타를 가지는 패킷을 양방향으로 전송할 수 있다. 각 인터코넥션 노드는 프로세서, 메모리 유닛, 입출력 프로세서, 입출력 캐쉬 등의 자원과 연결될 수 있다. 256개의 프로세서를 가지는 Tera Computer의 경우 16×16×16 toroidal 메쉬에 있는 4096개의 노드중에서 1280개의 노드는 여러 자원과 연결이 되어 있으며 2816개의 노드는 패킷 스위치의 역할만을 수행한다.

Tera Computer에 쓰이는 각 프로세서는 동시에 128개의 인스트럭션을 동시에 처리할 수 있는 VLIW(Very Long Instruction Word) 아키텍쳐를 사용하고 있으며 목표로 하는 성능은 최대 400~650 MFLOPS, 일반적인 응용 프로그램에 대하여 통상 240 MFLOPS 수준이다.

Tera Computer의 메모리 유닛은 각각 128 Mbyte의 메모리를 가지며 메모리 참조시의 부하균등을 이루기 위하여 데이타가 일부 메모리 모듈에 집중되는 현상을 방지하는 기법을 시도하고 있다.

9. 시스템 기술의 동향과 고찰

본 고에서는 7개의 상용 대규모 병렬처리 시스템을 소개하였다. 비록 다루고 있는 시스템들이 소수이긴 하지만 이들로 부터 공통된 몇가지 특성과 동향을 알 수 있다. 특히 본 고에서 소개하는 시스템들이 상용화된 시스템들이거나 앞으로 이 분야에서 많은 영향력을 끼치는 시스템임을 감안하면 이들로 부터 파악한 동향은 앞으로 대규모 병렬처리 컴퓨터의 발전방향을 유추하는 중요한 단서를 제공할 것으로 판단된다.

9.1 시스템의 발전방향

9.1.1 연산소자의 고급화

병렬처리 컴퓨터에 사용되는 개별 프로세서들의 성능은 날이 갈수록 강력해지고 있다. 이는 VLSI 기술의 발달과 마이크로프로세서 기술의 발달을 앞지르고 있어서, 향후 언젠가는 병렬처리 컴퓨터용 프로세서들도 최첨단의 고급 프로세서를 채용하게 될 것이다. 기존의 비트 연산 기술을 이용하는 것으로 대표되는 SIMD 시스템에서도 이러한 현상은 나타나고 있다. 예를 들어, MasPar MP-2의 64비트 아키텍쳐는 CM-1, 2의 1비트 아키텍쳐에 비해 매우 고급화된 것이다. MIMD 시스템의 경우에도 이러한 프로세서의 고급화 현상이 더욱 뚜렷이 나타나고 있다. 예를 들어, Tera Computer는 한번에 128개의 인스트럭션을 수행할 수 있는 VLIW 프로세서를

사용하고 있으며, Cray의 T3D는 현재 상용 마이크로프로세서중에서 가장 고성능인 RISC칩인 Alpha 칩을 사용하고 있다.

nCUBE-3는 주문제작에 의하여 개발된 프로세서를 사용함에도 불구하고 100 MFLOPS, 1 Gbyte 메모리, 256 Tbyte의 요구 페이지 가상메모리 (demand-paged virtual memory)를 지원하는 강력한 프로세서를 탑재하고 있다. MIMD 병렬처리 컴퓨터의 노드에서 가상 메모리를 제공하기 시작한 것도 특기할 만하다.

9.1.2 지연시간의 최소화와 통신 대역폭의 향상

인터코넥션 네트워크에서 네트워크 대역폭과 네트워크 지연은 매우 중요한 특징들이다. 병렬처리 컴퓨터의 연산능력이 향상되면서 인터코넥션 네트워크의 성능이 따라서 향상되는 것은 당연하나, 최근의 MIMD 시스템에서는 인터코넥션 네트워크의 또다른 주요 특징인 네트워크 토폴로지에 비하여 상대적으로 높은 대역폭과 낮은 네트워크 지연을 강조하는 경향이 강하다. 이는 다음에 언급되는 hidden topology, 그리고 공유메모리와 깊은 관계를 가지고 있다.

9.1.3 네트워크 토폴로지와 무관한 프로그램환경

최근의 병렬처리 컴퓨터중에는 사용자들이 시스템의 네트워크 토폴로지와 무관하게 단순한 통신을 위한 기본함수만을 사용하여 프로그래밍을 하는 경향이 있다. 이러한 경향은 시스템의 효율성보다도 사용자의 편의를 더 강조한 결과이며 병렬처리 시스템들이 범용화를 시도하는 과정의 일부분으로 생각된다. 이는 전통적으로 병렬처리 컴퓨터의 분류기준의 하나로 사용되던 네트워크 토폴로지의 중요성이 병렬처리 시스템의 범용화 과정에서 적어도 사용자의 입장에서는 감소하고 있음을 보여준다. 이러한 경향이 극단적으로 나타나는 예가 서로 다른 네트워크 토폴로지를 가진 시스템 사이에 소프트웨어의 호환성을 주장하는 경우이다. Intel은 다음과 같이 2-D 메쉬인 Paragon과 하이퍼큐브인 iPSC/860 사이의 소프트웨어 호환성을 주장하고 있다[1].

"Intel's 7.6 GFLOPS iPSC/860 supercomputer shares with the Paragon system the same development toolset, and the UNIX operating system is also available for the iPSC/860. Applications developed on the iPSC/860 move to the higher performance of the Paragon with no code modification, making the iPSC an ideal development platform for the Paragon system."

토폴로지와 무관한 환경을 구현하려는 경향은 전항에서 언급된 네트워크의 높은 대역폭과 지연시간을 작게 해야할 필요성을 더욱 높이고 있다.

9.1.4 공유메모리

병렬처리 컴퓨터의 프로그래밍에 어려움을 느끼는 사용자들을 위하여 물리적으로는 분산되어 있는 메모리를 논리적으로는 하나의 커다란 공유메모리처럼 사용케 하려는 시도들이 나타나고 있다. 예를 들어 Kendell Square Research의 KSR1에서는 부캐쉬(subcache)를 제외한 시스템의 모든 메모리를 하나의 커다란 공유메모리로 간주하고 있다. 이와 같이 사용자에게 논리적인 공유메모리를 제공하려는 움직임은 Cray Research와 Tera Computer 등에서도 나타나고 있다.

9.2 기술분석

병렬처리 컴퓨터에 관련된 기술중에는 서로 반대되는 성향의 기술들이 혼재한다. 따라서 여러가지 기술이 복합적으로 응용되고 있는 경우에 기술들 간의 상대적인 우위성을 살펴보는 것도 흥미롭다.

9.2.1 전통적인 기술과 진보적인 기술

전통적인 기술을 사용하는 기업은 가격대 성능비를 중시하며, 개개의 프로세서에 의한 성능 향상보다는 대규모 병렬처리에 의한 성능 향상에 중점을 둔다. MasPar와 NCR이 이계열에 속한다고 할 수 있다. 이와 달리 진보적인 기술을 추구하는 회사들은 최고의 성능을 제공하는 시스템의 개발에 중점을 둔다. 특히 TFLOPS를 향한 경쟁이 종전에는 고가의 슈퍼컴퓨터에서만

사용되던 기술을 저가의 병렬처리 컴퓨터에 적용하고 있다. Cray Research, Tera Computer 및 nCUBE 등이 이계열에 속하는 대표적인 회사들이다. 전통적인 기술을 토대로 첨단의 성능을 내는 회사들과 진보적인 기술을 적용하여 고가의 첨단 시스템을 구축하는 회사들은 앞으로 상당 기간동안 공존할 것이며, 우위를 점하기 위한 피나는 연구개발 노력이 성능의 혁신으로 바뀌어 우리에게 다가올 것이다.

9.2.2 자체개발 프로세서와 상용 프로세서

MasPar, nCUBE, Kendall Square Research 및 Tera computer 등은 자체 개발한 프로세서를 사용하며 NCR, Intel, Cray Research 등은 상용 프로세서를 탑재하는 시스템을 개발하고 있다. 본 고에는 소개되지 않았지만 트랜스퓨터(transputer)를 탑재하는 회사들을 모두 고려하면 상용 프로세서를 사용한 시스템이 자체개발 프로세서를 사용하는 경우에 비하여 훨씬 많다. MasPar와 Kendall Square Research는 SIMD 아키텍쳐와 ALLCACHE 시스템 때문에 자체 개발한 프로세서를 사용하는 것이 당연해 보인다. nCUBE는 MIMD 시스템임을 고려할 때 상용 프로세서를 이용하는 것이 유리해 보이나 하이퍼큐브 토폴로지를 구현하는데 따르는 복잡도를 줄이기 위하여 자체개발 프로세서를 탑재하는 것으로 보인다. 그 이유는 상용 프로세서중에는 nCUBE의 하이퍼큐브 연결을 효과적으로 제공할 수 있는 것이 없기 때문이다. 이는 시스템의 네트워크 토폴로지를 사용자에게 보이지 않으려는 경향에도 불구하고 네트워크 토폴로지가 시스템에 끼치는 영향이 매우 크다는 것을 암시한다.

9.2.3 SIMD와 MIMD의 선택

SIMD와 MIMD 중 어떤 것이 병렬처리 컴퓨터의 주력이 될 것인가? 이에 대한 해답은 현시점에서 MIMD로 기우는 듯 하다. 본 고에서 소개한 비교적 성공적인 병렬처리 컴퓨터 중에 SIMD 시스템은 MasPar 하나 뿐이며, 상용 SIMD시스템의 선구자였던 Thinking Machine 사에서도 CM-5시스템에서는 SIMD와 MIMD를 모두 제공하여 SIMD에서 벗어나고 있음이 이러한 경향을 뒷받침하고 있다. MIMD 시스템의 강세는 병렬처리 컴퓨터의 이용이 종래의 과학계산용 위주에서 상업용 위주로 확산되고 있는 것과 밀접한 관계를 찾을 수 있을 것으로 보인다.

10. 결 언

본 고에서는 상용 대규모 병렬처리 컴퓨터에 대한 소개를 하였다. 소개하는 시스템을 선정하는 데는 시스템이 상업적으로 성공하였는가와 시스템이 속하는 아키텍쳐와 응용 분야에서 대표성이 있는가를 주로 고려하였다. 특히 본 고에서는 인터코넥션 네트워크를 중심으로 시스템을 소개하였으며, 시스템 아키텍쳐를 주로 다루도록 하였다. 본 고를 작성하면서 저자가 느낀 바는 이제 상용 병렬처리 컴퓨터도 치열한 경쟁의 시대로 접어들었다는 것이다. 초기의 상용 병렬처리 컴퓨터는 병렬성의 이용과 그에 따른 성능향상으로 인한 높은 성능대 가격비가 경쟁력의 원천이었으나, 최근에는 병렬처리의 일반화로 단순한 병렬성의 이용만으로는 충분한 경쟁력을 가질 수 없게 되었다. 앞으로는 시스템을 친숙하게 사용할 수 있는지, 풍부한 소프트웨어가 지원되는지, 사용자의 문제를 해결할 수 있는 방안이 제공되는지 및 대규모 병렬성을 보다 경제적으로 이용할 수 있는 새로운 기술이 적용되고 있는지가 향후 어느 병렬처리 컴퓨터가 상업적으로 성공할 것인지를 결정짓는 시기가 곧 도래할 것이다.

참고문헌

[1] *Paragon XP/S Product Overview*, Intel Corporation, 1991.
[2] *MasPar Application Developer's Perspective*, MasPar Computer Corporation, 1992.
[3] *NCR System 3600 Product Description*, NCR Corporation, 1992.
[4] *NCR MPP System*, NCR presentation material.
[5] *nCUBE 3 Technical Overview: The Future of Massively Parallel Computing*, nCUBE, 1993.
[6] R. Alverson, D. Callahan, A. Porterfield, D. Cu-

mmings, B. Smith, and B. Koblenz, "The Tera Computer System," *International Conference on Supercomputing*, pp. 1~6, 1990.

[7] T. Blank, "The MasPar MP-1 Architecture," *IEEE COMPCON* Spring 1990, pp. 20~24, 1990.

[8] G. Khermouch, "Large Computers," *IEEE Spectrum*, pp. 46~49, January, 1994.

[9] J. T. Kuehn and B. J. Smith, "The Horizon Supercomputing System: Architecture and Software," *Supercomputing*, pp. 28~34, 1988.

[10] J. R. Nickolls, "The Design of The MasPar MP-1: A Cost Effective Massively Parallel Computer," *IEEE COMPCON*, pp. 25~28, Spring, 1990.

[11] J. R. Nickolls, "The Design of The MasPar MP-2: A Cost Effective Massively Parallel Computer," *Manuscript, MasPar Computer Corporation*, 1992.

[12] F. Pittelli and D. Smitley, "Analysis of a 3D Toroidal Network for a Shared Memory Architecture," *Supercomputing*, pp. 42~47, 1988.

[13] G. Ramanathan and J. Oren, "Survey of Commercial Parallel Machines," *ACM Computer Architecture News*, Vol. **21**, No. 3, pp. 13~33, 1993.

[14] E. Rosti, E. Smirni, T. D. Wagner, A. W. Apon, and L. W. Dowdy, "The KSR1: Experimentation and Modeling of Poststore," *ACM Sigmetrics Conference on Measurement & Modeling of Computer Systems*, pp. 74~85, 1993.

[15] M. R. Thistle and B. J. Smith, "A Processor Architecture for Horizon," *Supercomputing*, pp. 35~41, 1988.

[16] A. Trew and G. Wilson, Ed., *Past, Present, Parallel: A Survey of Available Parallel Computing Systems*, Springer-verlag, 1991.

[17] G. Zorpette, "The Power of Parallelism," *IEEE Spectrum*, pp. 28~33, September 1992.

[18] G. Zorpette, "Large Computers," *IEEE Spectrum*, pp. 34~37, January 1993.

[19] 김기철, "상용 대규모 병렬처리 컴퓨터의 인터코넥션과 시스템 아키텍처," 병렬처리시스템연구회지 제4권 3호, pp. 33~48, 11월, 1993.

김 기 철

1982 서울대학교 전기공학과, 학사
1984 서울대학교 전기공학과, 석사
1991 University of Southern California 전기공학과 박사
1984 ~1994 한국전자통신연구소, 선임연구원
1994 ~현재 서울시립대학교 반도체 공학과

관심 분야 : VLSI 설계, 병렬처리