기존 슈퍼컴퓨터 분석을 통한 엑사스케일 컴퓨팅 시스템 전망

Prospect of Exascale Computing System through Analysis of Supercomputers

박성규, 박기웅, 박규호
Sung Kyu Park, Ki-Woong Park, Kyu Ho Park

(305-701) 대전광역시 유성구 대학로 291 KAIST 전기및전자공학과 컴퓨터공학연구실
{skpark, woongbak, kpark}@core.kaist.ac.kr

요 약

신약개발, 기상예측, 공학 시뮬레이션 등 과학 및 공학 분야에서 주로 활용되던 슈퍼컴퓨팅은 고객 분석, 경제예측, 멀티미디어 등 활용범위가 극대화 되고 있으며, 최근 슈퍼컴퓨팅 수요 산업이 증대됨에 따라 대규모 컴퓨팅 자원의 충족을 위해 슈퍼컴퓨팅 시스템 개발의 필요성이 증가하고 있다. 슈퍼컴퓨팅 기술은 단일 기술이 아니라 프로세서, 메모리, 네트워크, 스토리지에 이르기까지의 첨단 하드웨어를 기반으로 최고의 성능을 내기 위한 운영체제와 계산 라이브러리, 그리고 그 응용 프로그램에 이르는 소프트웨어 기술의 총 집합체로써 각각의 세부 기술들은 슈퍼컴퓨팅의 실현에 필수불가결한 핵심 요소라 할 수 있다. 본 논문에서는 기존 슈퍼컴퓨터 기술 분석을 통해 향후 미래의 고성능 엑사스케일 컴퓨팅 시스템의 구조 및 이를 위한 효율적인 자원 관리 기술을 전망하여, 슈퍼컴퓨팅 기술개발 정책 및 연구방향을 제시한다.

Abstract

There is a growing need about high-performance computing system in order to support large-scale computing resources. Hige-performance computing technology consists of hardware components which are processors, interconnect, memory, network, and storage system and software components which are operating systems and applications. These technologies critically affect the advance of computing areas. In this paper, we establish a road map for exascale high-performance computing system by analyzing core technology trend and issues.

키워드: 엑사스케일 시스템, 슈퍼컴퓨터, 매니코어, 차세대 비휘발성 메모리, 자원 관리
Keyword: Exascale Computing System, Supercomputer, Manycore, Non-Volatile Memory, Resource Management

논문투고일자 : 2012. 01. 13 심사일자 : 2012. 01. 27 게재확정일자 : 2012. 02. 21

1. 서론

최근 슈퍼컴퓨팅 수요 산업이 증대됨에 따라 대규모 컴퓨팅 자원의 충족을 위해 슈퍼컴퓨팅 시스템 개발의 필요성이 증가하고 있다 [1]. 신약개발, 기상예측, 공학 시뮬레이션 등 과학 및 공학 분야의 전유물로 인식되어 왔던 과거와 달리, 최근 들어 슈퍼컴퓨팅은 고객 분석, 경제예측, 멀티미디어 등 활용범위가 극대화 되고 있는 것이다. 이와 같은 슈퍼컴퓨팅 기술은 단일 기술이 아니라 프로세서, 인터커넥팅, 메모리, 네트워크, 스토리지에 이르기까지의 첨단 하드웨어를 기반으로 최고의 성능을 내기 위한 운영체제와 계산 라이브러리, 그리고 그 응용 프로그램에 이르는 소프트웨어 기술의 총 집합체로써 각각의 세부 기술들은 컴퓨팅 분야의 발전에 크게 영향을 미치고 있다. 이뿐만 아니라 고성능 컴퓨팅 기술은 유전자 분석, 지진 분석, 기상 분석 등의 기초 과학 분야의 연구에서부터 산업에서 필요한 신약 개발, 자동차 또는 비행체 모의 실험, 고객 분석, 경제예측, 멀티미디어 등의 대규모 시뮬레이션에 이르기까지의 많은 응용에 적용함으로써 이익을 얻을 수 있어서 타 기술 발전에 크게 이바지 할 수 있는 잠재력을 가지고 있다.

고성능 컴퓨팅 기술을 확보하기 위한 장기적인 안목의 연구 투자를 비롯하여 컴퓨팅 센터를 구축하고 유지하기 위한 막대한 비용을 감당하기 위해서는 기본적으로 국가 단위의 지원이 필요한데, 이미 국외의 경우 고성능 컴퓨팅 시스템의 높은 시장성을 인식하여 이를 위한 자원 관리 및 분배를 위한 연구가 진행되고 있다. 세계 500위 안의 슈퍼컴퓨터 보유 국가 중 미국이 총 225개로 가장 많이 보유하고 있으며, 중국이 61개, 독일이 30개, 영국이 27개, 일본이 26개의 슈퍼컴퓨터를 보유 중에 있다 [2]. 이에 반해 국내 기술은 세계적인 기술에 비해 상당히 뒤떨어져 있다. <표 1>에서 알 수 있듯이 한국은 세계 500위 안의 슈

〈표 1〉 2011년 6월 기준 Top 10 슈퍼컴퓨터 및 국내 슈퍼컴퓨터 보유 현황

Rank	Site	Manufacturer	Computer	Country	Cores	Rmax (Pflops)	Power (MW)
1	RIKEN Advanced Institute for Computational Science	Fujitsu	K Computer SPARC64 VIIIfx 2.0GHz, Tofu	Japan	548,352	8.162	9.90
2	National Supercomputing Center in Tianjin	NUDT	Tianhe-1A NUDT TH MPP, Xeon 6C, Nvidia	China	186,368	2.566	4.04
3	Oak Ridge National Laboratory	Cray	Jaguar Cray XT5, HC 2.6 GHz	USA	224,162	1.759	6.95
4	National Supercomputing Centre in Shenzhen (NSCS)	Dawning	Nebulae TC3600 Blade, Intel X5650, Nvidia	China	120,640	1.271	2.58
5	GSIC, Tokyo Institute of Technology	NEC/HP	TSUBAME-2 HP ProLiant, Xeon 6C, Nvidia	Japan	73,278	1.192	1.40
6	DOE/NNSA/LANL/SNL	Cray	Cielo Cray XE6, 8C 2.4GHz	USA	142,272	1.110	3.98
7	NASA/Ames Research Center/NAS	SGI	Pleiades SGI Altix ICE 8200EX/8400EX	USA	111,104	1.088	4.10
8	DOE/SC/LBNL/NERSC	Cray	Hopper Cray XE6, 6C 2.1 GHz	USA	153,408	1.054	2.91
9	Commissariat a l'EnergieAtomique (CEA)	Bull	Tera 100 Bull bullx super-node S6010/S6030	France	138,368	1.050	4.59
10	DOE/NNSA/LANL	IBM	Roadrunner BladeCenter QS22/LS21	USA	122,400	1.042	2.34
20	Korea Meteorological Administration	Cray	Cray XE6 12 core 2.1 GHz	Korea	45,120	0.3164	20
21	Korea Meteorological Administration	Cray	Cray XE6 12 core 2.1 GHz	Korea	45,120	0.3164	21
26	KISTI Supercomputing Center	Oracle	Sun Blade x6048, X6275 Sun HPC Stack Linux Edition	Korea	26,232	0.2748	1.276
423	National Fusion Research Institute	HP	HP BL685cG7, Opteron 12C, 2.2 GHz, InfiniBand	Korea	6,912	0.0432	

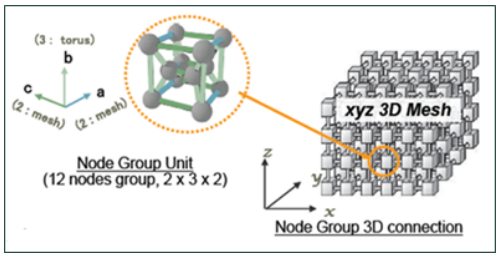
퍼컴퓨터 중 4개를 보유 중에 있으나 이 슈퍼컴퓨터들 역시 기존 해외의 슈퍼컴퓨터 관련 기술들을 그대로 들여와 구축하고 적용하는 수준에 머물러 있으며, 현재 최고 수준의 시스템에 비해 성능이 수십 분의 일에도 미치지 못하는 실정이다. 본 논문에서는 세계 Top3 슈퍼컴퓨터의 구조 및 자원관리 기법의 분석을 통해 향후 미래의 고성능 엑사스케일 컴퓨팅 시스템의 구조 및 자원 관리 방법을 전망할 것이다. 또한 현재 활발하게 개발 중인 이기종 매니코어 프로세서, 3D-Stacked 메모리, 및 차세대 비휘발성 메모리가 엑사스케일 컴퓨팅 시스템에 어떻게 적용이 될 수 있으며, 이에 대한 자원 관리 방법에 대한 논의가 이루어질 것이다.

본 논문은 2장에서 세계 Top 3 슈퍼컴퓨터에 대한 분석을 하고, 3장에서는 엑사스케일 컴퓨팅 시스템의 개발 이슈에 관해 정리를 할 것이다. 4장 및 5장에서는 엑사스케일 컴퓨팅 시스템의 구조 및 자원 관리 방법에 관해 전망을 하고, 마지막 6장에서 결론을 끝으로 논문을 마무리 할 것이다.

2. Top 3 슈퍼컴퓨터 분석

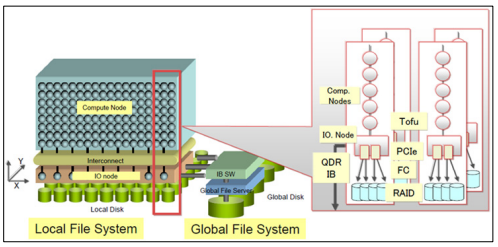
2.1 K Computer [3]

일본 Fujitsu 사에서 개발한 K Computer는 2011년 6월 기준 세계에서 가장 빠른 슈퍼컴퓨터이다. K Computer는 총 68,644개의 Fujitsu SPARC64 VIIIfx 프로세서로 이루어져 있고, 각 프로세서는 2 GHz로 동작하는 8개의 코어로 구성이 되어 총 549,152개의 코어로 구성이 된다. 프로세서의 성능 향상을 위해 각 프로세서에 256개의 부동 소수점 레지스터를 추가하고, SIMD (Single Instruction Multiple Data) 명령어 집합을 추가하였다. 또한 고속, 고정밀의 삼각함수 연산을 위한 특수 명령어 집합을 추가하여 성능을 최적화시키기 위해 노력하였다. 이를 통해 K Computer는 총 8.162 PFLOPS의 성능을 보여 세계에서 가장 빠른 컴퓨터가 되었다. 또한 수냉식 쿨링을 통해 58W의 저전력 설계를 달성하였다. Tianhe-1A의 0.64GF/W와 비교해서 0.83GF/W로 단위 전력 당 성능이 더 좋은 편이다.



(그림 1) K Computer의 Tofu 연결망 구조

K Computer는 그림 1에서 보이는 바와 같이 'Tofu' 라고 불리는 연결망 구조를 사용하여 구성이 되어 있다. Tofu는 3D 토러스 구조를 바탕으로 하고 있으며, 총 6차원 구조로써, x, y, z, a, b, c 의 좌표를 사용해서 각 노드를 선택하도록 하였다. 각 컴퓨팅 노드는 ICC (Interconnect controller)를 가지고 있고, ICC는 총 10개의 링크를 가지고 6차원 토러스 구조를 형성한다. 이 링크는 6개의 3D 토러스 링크와 4개의 tofu 유닛의 링크로 구성이 되고, 각 링크는 최대 10GB/s의 대역폭을 가지고, 100GB/s의 처리속도를 보인다.

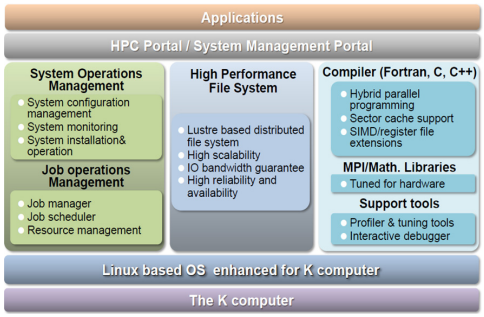


(그림 2) Fujitsu Exabyte 파일 시스템

K Computer의 스토리지 시스템은 (그림 2)에서 보이는 바와 같이 FEFS (Fujitsu Exabyte File System) 이라는 분산 파일 시스템을 사용하고, Lustre 파일 시스템을 기반으로 구성된다. FEFS는 100PB에서 1EB 정도의 매우 큰 용량의 파일과 단일 스트림으로는 GB/s, 병렬 스트림으로는 TB/s의 처리 속도 및 높은 신뢰성과 활용성을 목적으로 설계되었다. FEFS는 글로벌 파일 시스템과 로컬 파일 시스템으로 구분하여 사용이 되는데, 로컬 파일 시스템의 경우 /work 폴더에 마운트 되고, core에서 처리되는 태스크에 맞춰진 고속의 파일 시스템으로 사용된다. 이

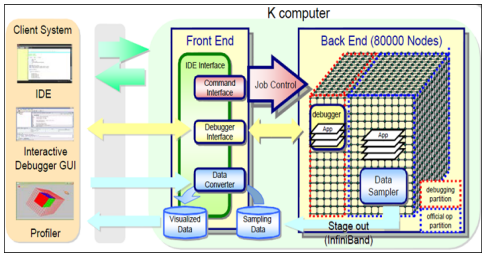
는 컴퓨팅 노드를 위하여 배타적 용도로 사용이 된다. 글로벌 파일 시스템의 경우는 /data 폴더에 마운트 되어 사용되고, 공유를 목적으로 고용량과 중복성을 가진 파일 시스템으로 구성되어 많은 서버 및 시스템과 공유가 될 수 있다.

K Computer의 소프트웨어 구조는 그림 3에서 보이는 바와 같이 리눅스 기반의 기반 운영체제를 기반으로 관리 모듈, 고성능 파일시스템, 컴파일러 내부 모듈이 적재가 되며, 중앙 관리 시스템을 통해 통제를 받는 구조를 취한다.



(그림 3) K Computer의 소프트웨어 아키텍쳐

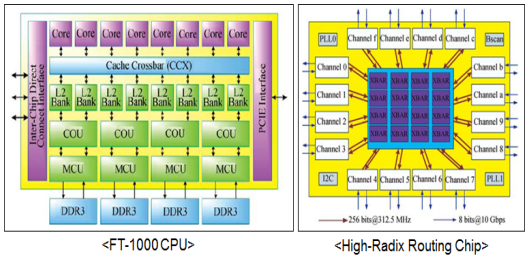
K Computer는 계층적 프로그래밍 모델을 통해 태스크 수행의 효율성을 극대화하고 있다. (그림 4)에서 보이는 바와 같이 Front End 계층에서는 유저 프로그래밍 환경을 위한 IDE 인터페이스를 제공하고 전용 디버깅 파티션인 인터랙티비브 디버거를 제공한다, Backend 계층은 샘플 데이터 프로파일러 및 튜닝 도구를 제공하고 복수개의 태스크에 대한 디버깅 파티셔닝을 제공하고 각 실행 상황을 인피니밴드의 고속 채널을 통해 Front End계층에 전달해 주도록 함으로서 디버깅 용이성 제공한다.



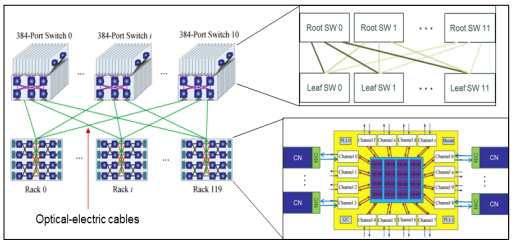
(그림 4) K Computer의 프로그래밍 모델

2.2 Tianhe-1A [4]

Tianhe-1A는 중국 NUDT에서 개발한 슈퍼컴퓨터로 현재 슈퍼컴퓨터 랭킹 2위에 위치하고 있다. Tianhe-1A는 14,336개의 Intel Xeon X5670 프로세서와 2,048개 FT-1000 CPU, 7,168개의 Nvidia Tesla M2050 GPGPU로 구성이 된다. 각 프로세서는 2.93 GHz로 동작하는 6개의 코어로 구성이 되고, 총 86,016개의 코어로 구성이 된다. 각 GPU는 1.15 GHz로 동작하는 448개의 코어로 구성이 되고, 총 3,211,264개의 GPU 코어로 구성이 된다. 각 컴퓨팅 노드는 2개의 Intel CPU와 1개의 NVIDIA GPU로 구성이 되고, 각 서비스 노드는 2개의 FT-1000 CPU로 구성이 된다. Tianhe-1A는 이기종의 프로세서와 customized 프로세서를 활용하여 총 2.57 PFLOPS의 성능을 보인다. 그림 5에서 보이는 것처럼 성능 향상을 위해 자체적으로 멀티코어 CPU를 개발하고, 대규모 interconnection network에서 통신 latency를 줄이기 위해 Routing Chip을 개발하였다.



(그림 5) Tianhe-1A의 CPU 및 연결망 칩 구조도

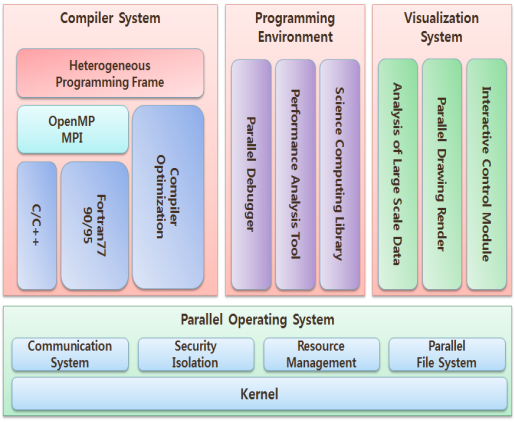


(그림 6) Fat Tree 기반 연결망

Tianhe-1A는 (그림 6)에서 보이는 것처럼 Fat Tree 기반의 연결망을 사용을 한다. 이는 토러스 구조보다 높은 성능을 보이고, 운용성 및 가용성 측면에

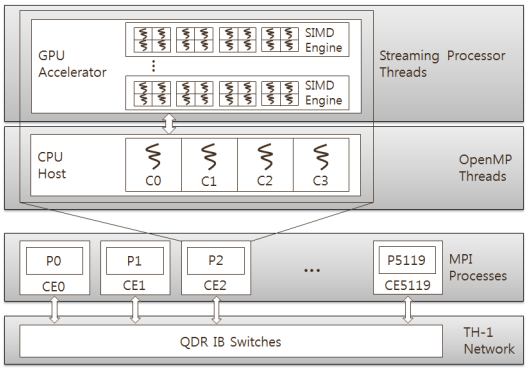
서 뛰어나지만 확장성이 떨어진다는 단점을 가지고 있다. 각 링크의 연결은 NUDT에서 제작한 High-Radix Routing Chip을 사용하고, 각 채널은 20 GBps의 성능을 보인다.

Tianhe-1A의 스토리지 시스템은 Lustre 파일 시스템 기반으로 구성이 되고, 6개의 I/O 관리 노드와 128개의 스토리지 노드로 구성이 된다. 스토리지 시스템과 컴퓨팅 시스템의 연결은 Tianhe-1에서는 InfiniBand를 사용하였으나, Tianhe-1A에서는 High-Radix Network Routing Chip을 사용하고 optic-electronic hybrid fat tree network으로 구성하였다.



(그림 7) Tianhe-1A 소프트웨어 아키텍처

Tianhe-1A 소프트웨어 아키텍처는 (그림 7)에서 보이는 바와 같이 병렬 (클러스터)형태의 운영체제를 기반으로 하고 있으며, 각 커널에는 통신, Security Isolation, 자원관리, 파일시스템 등의 내부 모듈을 탑재하고 있다. 운영체제를 기반으로 3개의 주요 컴포넌트가 적재가 되는데, 컴파일러 시스템은 OpenMP를 기반으로 다양한 언어를 한의 통합된 프레임웍으로 컴파일을 하고 최적화하는 기능을 담당하고, 개발환경 시스템은 병렬적인 디버깅 모듈과, 성능 분석 및 과학 계산용 라이브러리를 통해 주어진 태스크의 효율적인 처리를 위한 모듈 탑재하였다. 마지막으로 시각화 시스템은 각 컴퓨터 노드간의 상방 통신 및 병렬 파일시스템의 효율적인 데이터 통신 및 공유를 위한 계층이다.



(그림 8) Tianhe-1A 프로그래밍 모델

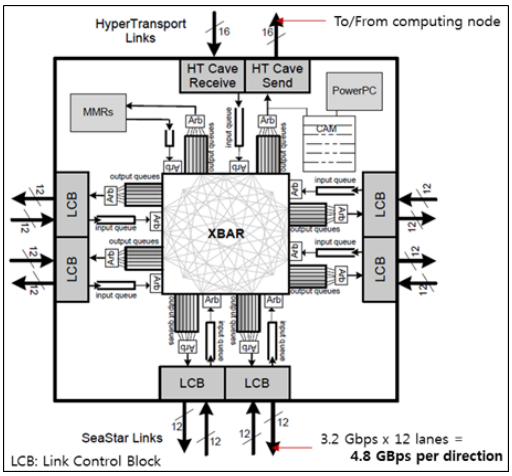
Tianhe-1A의 가장큰 특징은 GPU기반의 가속과 기능이 포함되어 있다는 것이다. 따라서 GPU내부의 SIMD 엔진을 통해 주어진 태스크의 병렬화를 극대화해야 한다. 이를 위해 OpenMP기반 프로그래밍 환경을 제공하여 복수개의 쓰레드들의 효율적인 실행을 가능하게 하였고, 각각의 물리적 코어는 TH-1 이라 불리는 자체적인 내부 네트워크로 통신을 가능하게 하여 동기화 및 내부 데이터 통신의 속도를 극대화시켰다.

2.3 Jaguar [5]

Jaguar는 미국 Cray 사에서 만든 슈퍼컴퓨터로써 현재 슈퍼컴퓨터 랭킹 3위에 위치하고 있다. Jaguar는 45,203개의 AMD Opteron 2356 프로세서로 구성이 되고, 각 프로세서는 2.3 GHz로 동작하는 4개의 코어로 구성이 되어, 총 180,832개의 코어로 구성이 된다. 각 코어마다 128 KB L1 캐쉬와 512 KB 크기의 L2 캐시로 구성이 되고, 한 프로세서 내의 4개의 코어는 2 MB 크기의 L3 캐쉬를 공유 한다. 이를 통해 총 1.64 PFLOPS의 성능을 보인다.

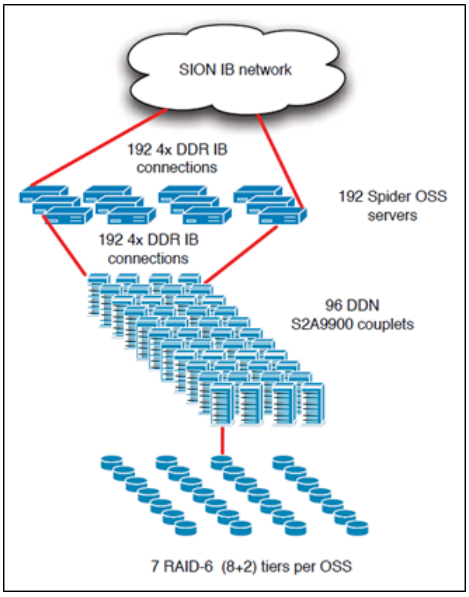
Jaguar의 연결망은 그림 9에서 보이는 것처럼 SeaStar라고 불리는 구조를 사용한다. SeaStar Router는 6개의 LCB (Link Control Block)과 2개의 HyperTransport 링크들로 구성이 됨. 각 LCB는 최대 9.6 GBps의 처리속도를 보여 준다.

스토리지 시스템은 (그림 10)에서 보이는 바와 같이 Spider 라는 파일시스템을 사용하는데, 이 역시 Lustre 파일 시스템을 기반으로 구성된 대용량 스토



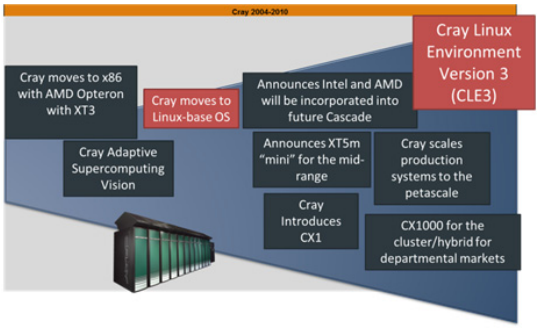
(그림 9) SeaStar Router 구조

리지 클러스터이다. SAN 기반의 객체 기반 스토리지로서 13,440개의 1 TB SATA 디스크를 사용하여 RAID 6로 구성되어, 총 10 PB 이상의 가용 용량을 제공한다. 특징으로는 Metadata를 2대의 LSI XBB2 디스크 어레이에 별도로 저장하고, 데이터들은 192개의 Object Storage Server (OSS)를 통하여 접근되며, 각 OSS는 1.25 GB/s I/O 대역폭을 제공한다.



(그림 10) Spier 파일 시스템

Jaguar의 소프트웨어 아키텍쳐는 (그림 11)에서 보이는 바와 같이 Cray Linux Environment(CLE)로 구성이 된다. 이는 크고 복잡한 어플리케이션 실행 및 240,000 개 코어 이상 스케일 가능 하도록 리눅스 기반의 OS 디자인하였고, 높은 확장성을 가진 사용자 어플리케이션으로 Lightweight 모드에서 동작하고, OS가 어플리케이션의 확장을 방해하지 않는다.



(그림 11) Cray Linux Environment (CLE)

Jaguar의 프로그래밍 모델은 공유 메모리 다중 처리 프로그래밍 API인 OpenMP를 수정하여 사용한다. Node를 유저들을 위한 어플리케이션의 visualization을 담당하는 부분과 연산을 위한 부분, 파일시스템을 관리하는 부분으로 나누어 해당 역할에 맞추어 OS를 재구성하였다. 마지막으로 각 Node 별로 담당 역할에 최적화시키기 위해 리눅스 기반의 별도의 OS를 구현하여 탑재하였다.

3. 엑사스케일 컴퓨팅 시스템 개발 이슈

System attributes	2010	"2015-2018"		"2018-2020"	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1 day)	

(그림 12) Potential System Architecture Targets in USA [7]

(그림 12)에서 알 수 있듯이 현재 슈퍼컴퓨팅 발전 속도에 미루어보아 엑사스케일 시스템은 2018년에서 2020년 사이에 달성될 것으로 전망하고 있다. 엑사스케일 시스템을 달성하기 위한 중요한 이슈가 저전력, 고성능, 고신뢰성이다. 본 절에서는 이를 달성하기 위한 개발 이슈들을 살펴볼 것이다.

3.1 저전력 설계

엑사스케일의 연산 속도를 가지기 위하서는 많은 컴퓨팅 노드들의 집합이 필요하여 컴퓨팅 연산에 소모되는 전력은 비약적으로 증가한다. 2006년 인텔사의 컴퓨터 코어의 단일 명령어 수행 당 전력 소모량은 10000 pJ 이며, 이 코어로 엑사스케일 급의 컴퓨터를 만들 경우 연간 10 BW를 소모하게 될 것이며, 추정치 10B$를 매년 소모하게 될 것이다. 크기가 큰 컴퓨터일수록 대부분의 전력 소모는 캐쉬, 메모리, 디스크들 사이의 리모트 데이터 전송에서 발생하게 되므로, 리모트 데이터 전송을 최소화 하는 것이 핵심 기술로 요구된다. 따라서 현재까지의 컴퓨터 회로의 초점은 연산 속도에 초점이 맞춰졌던 것에 반하여 앞으로의 회로의 디자인은 전력 소모를 최소화 하는 방향으로 맞춰줘야 한다. 칩 내의 통신 회로를 구성하거나, 새로운 메모리 구조 등이 그 예가 될 것이다. 데이터 전송에 사용되는 전류량에서도 가장 효율적인 대역폭을 찾아야 한다. 최근 연구 동향을 살펴보면, 컴퓨터에 공급되는 전압을 반도체의 문턱 전압에 맞추어 적용시키는 기술 연구가 진행 되고 있다. 이 기술을 사용하면 전력의 소모량을 줄 일 수 있지만, 연산 과정 중 발생되는 하드웨어적 지연이 5 배 정도 더 길어지기 때문에 이 문제를 해결해 나가는 것이 필요하다.

또한 효율적인 파워 사용을 위해 파워 소모량 모니터링 관련 기술이 활발하게 개발이 되고 있고, 각 단계에서의 power convention loss를 줄이는 노력이 계속해서 이루어지고 있다. 또한 HVAC (Heating, Ventilating, and Air Conditioning) 관련 개선 노력이 저전력 달성을 위해 이루어지고 있다.

3.2 고성능 디자인

엑사스케일 컴퓨팅 시스템은 컴퓨팅 자원들을 병렬화하여 최대한의 성능을 얻고자 하는 시스템으로, 현재 구축되어 있는 슈퍼컴퓨팅의 1000배까지의 용량 및 성능 증대를 목표로 하고 있다. 이를 위해 프로세서 개발에 있어서는 현재 단일 코어 당 클럭 속도 증가에 대한 한계로 인해, 수백 개의 코어를 한 칩 내에 집적한 매니코어 형태의 하드웨어 기술로 개발이 되고 있다 [8]. 또한, CPU의 집적에 의한 성능 향상이 아닌 GPU의 집적에 의한 성능 향상을 통해 더 높은 성능을 보다 적은 파워를 소모하면서 사용하고자 하는 노력이 이루어지고 있다.

메모리 성능에는 대역폭과 지연시간의 두 가지 이슈가 있다. 상용 DRAM은 대역폭 향상에 초점을 맞추고 있어 지연시간에서는 성능 저하가 일어나는 문제가 발생한다 [9]. 실제 DRAM이 DDR에서 DDR2, DDR3로 발전함에 따라 대역폭을 증가시키기 위해 더 복잡한 논리 회로를 사용하게 되었고, 이에 따라 지연시간이 함께 증가하는 결과가 나타났다. 이는 지연 시간뿐만 아니라 추가 논리 회로에 의한 전력 문제와 공간 효율성 문제도 야기 시킨다. 엑사스케일 시스템에서는 전력과 공간 사용을 감소시키면서 성능을 향상시키는 방법이 요구되는데, 입출력 파이프라인의 논리 회로의 간소화, 입출력 버스 넓이 증가, 혹은 반도체 소자의 성능 향상이 필요하다. DRAM 프로세스는 메모리 누수를 줄이고 생산성을 증가시키는 방향으로 최적화되고 있다. 하지만 메모리 누수를 줄이기 위한 반도체 소자는 고성능을 위해 설계된 반도체 소자와 다소 성능 차이가 있다. 또한 비용 절감을 위해 핀 개수를 제한하는 경우가 많고, 이는 같은 양의 데이터를 전송하는 데 필요한 시간을 늘리고 전력 소모를 증가시키고 있다. 따라서 엑사스케일 시스템에서의 메인 메모리 구성을 위해 메모리 구조의 변경 및 이를 위한 관리가 필요하다.

디스크의 데이터 전송 속도 자체는 꾸준히 증가하는 추세이나, 디스크 상에서 데이터를 찾는 횟수나 찾는 속도에 의해 그 성능이 상당히 감소하며, 크기가 작은 데이터의 입출력이 일어날 때 그 감소폭은 더 커지고 있다. 결국 엑사스케일 급 컴퓨터에서도 데이터 찾는 시간을 줄일 필요가 있다 [9].

엑사스케일 시스템의 성능에 있어서 이슈가 되는 부분은 각 자원들의 개별적 성능 최대화와 더불어, 수

많은 자원들을 어떻게 Scalable하게 사용할 수 있을 것인가이다. 실제로 각 자원 컴포넌트들의 성능 최대화는 하드웨어와 밀접한 관련이 있으며, Scalable한 시스템의 구현은 소프트웨어적으로 디자인되어야 할 것이다.

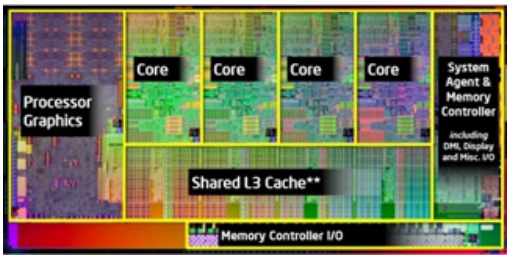
3.3 고신뢰성 시스템 구축

기존의 낮은 단계의 컴퓨팅 시스템에서는 찾을 수 없던 희박한 확률의 에러 및 실패들이 수많은 노드를 사용하는 엑사스케일 컴퓨팅 시스템에서는 매번 발생하는 이벤트가 될 수 있다. 엑사스케일급 시스템에서 이러한 시스템 인터럽트가 발생하는 현상은 더욱 가중될 것이므로, 각종 시스템 에러에 대한 관리는 시스템 가용성 및 내성에 있어서 매우 중요한 이슈가 될 것이다 [9]. 엑사스케일 시스템은 매우 많은 수의 컴퓨팅 자원을 기반으로 이루어져 있으며, 이들이 모여 큰 규모의 네트워킹을 구성하고 있고, 이러한 자원들이 소프트웨어적으로 관리가 되어야 하는 시스템이다. 따라서, 여러 영역 및 레이어에 걸쳐서 에러를 발생시키는 가변적인 원인이 존재한다. 엑사스케일급 슈퍼컴퓨팅 시스템에서 예상되는 에러의 원인 및 발생 위치는 수백만에서 수십억개에 달하는 공간적 메모리 및 디스크의 하드웨어상 에러와, 에이징에 따른 시간적 하드웨어 에러, 같은 하드웨어 서킷 사이에서도 존재하는 시공간적 에러, 전력 감소를 위한 저전력화 정책에 따른 에러, 데이터 트랜스미션 시의 비트 에러, 동시수행연산능력의 증가로 인한 데드락 문제 등이 있다. 에러 특성에 따른 구분으로는 시스템이 멈추거나 고장나도록 하는 영구적인 에러, 복구 가능하거나 짧은 시간 동안에는 나타나는 트랜지언트 에러, 시스템 멈춤이나 고장은 없으나 시뮬레이션 및 계산에 Corruption을 발생시키는 발견되지 않은 에러 등이 존재한다.

4. 엑사스케일 컴퓨팅 시스템 구조 전망

4.1 프로세서 구조 전망

단일 코어 당 클럭 속도 증가에 대한 한계로 인해 수백 개의 코어를 집적한 매니코어 형태의 하드웨어 기술이 개발되고 있다. 칩내부 코어 수가 증가하게 될 것이며, 그림 7의 인텔 샌디브릿지와 같이 CPU만이 아닌 GPU가 결합된 이기종의 프로세서가 차세대 고성능 혼합형 시스템에서 사용 될 것으로 전망 된다. 또한 코어 간 연결에 있어서 확장성, 적은 파워 소모, 성능 향상을 위해 Router 기반으로 연결망이 이루어질 것으로 보인다.



(그림 13) 인텔 샌디브릿지 내부 구조 〔10〕

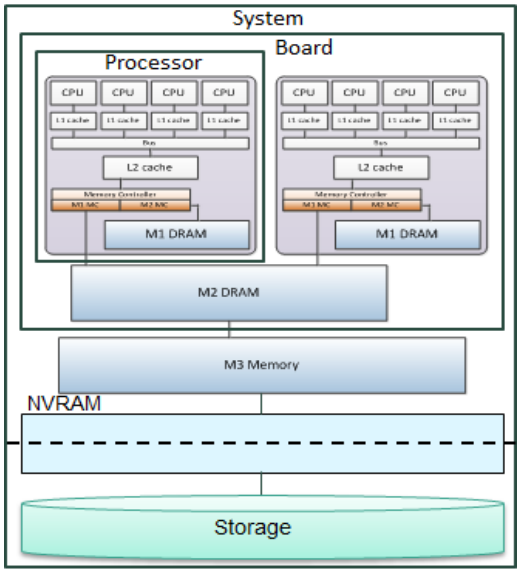
4.2 메인 메모리 구조 전망

차세대 고성능 컴퓨팅 시스템의 메인 메모리 구조는 그림 8과 같이 계층적 메인 메모리 구조가 될 것이다. 계층적 메인 메모리는 3D-Stacked 메모리 [11]와 DRAM 및 차세대 비휘발성 메모리 [12][13]로 구성이 되어 있으며 이 구조 구성을 통해 메인 메모리 접근 속도를 증가시키고 고성능 시스템에 맞는 대용량 메모리 구성이 가능하다. 계층적 메인 메모리는 M1, M2, M3 메모리로 구성이 되어 있으며, M1은 on-chip 메모리로 3D stacked 메모리로 구성이 되고, 빠른 접근 속도를 보이는 특성을 가지고 있다. M2는 on-board 메모리로 기존 메인 메모리 역할을 하며 차세대 비휘발성 메모리와 함께 하이브리드 메모리로 구성이 가능하다. M3는 off-board 메모리로 PCI-e 인터페이를 통해 메모리 용량을 확장시켜 대용량 메모리 구성이 가능하도록 한다. 차세대 비휘발성 메모리는 메인 메모리와 스토리지로 동시에 사용이 가능하다.

4.3 스토리지 시스템 구조 전망

현재 파일시스템 개발 동향은 Lustre 파일시스템 기반으로 고성능 컴퓨팅 시스템 파일시스템이 구축되

고 있다. Lustre 파일시스템은 분산, 병렬, 클러스터 파일시스템으로 객체 저장단위 및 병렬 I/O를 지원한다. 현재 Top 100 슈퍼컴퓨터 중 65개의 슈퍼컴퓨터가 사용 중이며, Top 10에서는 8개의 슈퍼컴퓨터가 이를 사용한다. 차세대 고성능 슈퍼컴퓨팅 시스템에서는 새로운 스토리지 구조에 맞는 파일시스템의 수정이 필요하다. 우선 각 노드 당 로컬 파일시스템이 존재하고, Lustre 파일시스템 기반의 글로벌 파일시스템이 존재하는 멀티 레벨 파일시스템이 사용될 것이다. 또한 InfiniBand 혹은 고성능 상호연결망을 통한 I/O 연결이 이루어질 것이고, 차세대 비휘발성 메모리 역시 스토리지로 사용될 수 있기 때문에 이를 고려한 파일시스템의 메타데이터 관리가 필요할 것이다. 또한 엑사스케일에서는 시스템의 내성이 중요하기 때문에 스토리지 시스템에서도 에러가 존재해도 시스템이 문제없이 진행될 수 있도록 디자인이 필요하다.

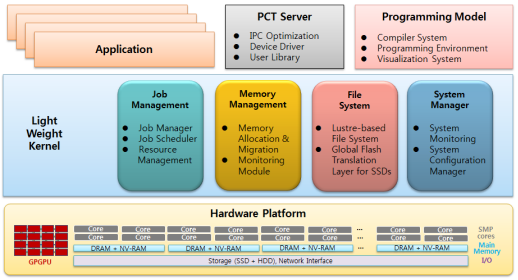


(그림 14) 계층적 메인 메모리 구조

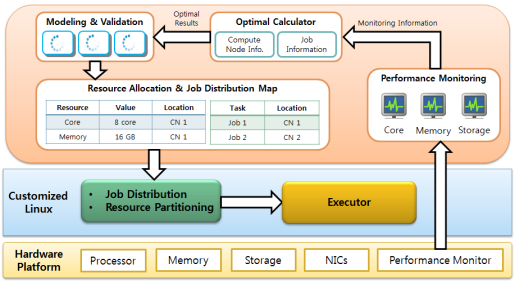
4.4 슈퍼컴퓨터의 연결망 구조의 전망

기존 병렬 버스 연결은 하나의 장비가 많은 대역폭을 소모하면 다른 장비의 사용이 어렵고 계층적 스위치 구조의 오버헤드가 존재한다. 따라서 고성능 혼합형 컴퓨팅 시스템 연결에 있어서는 InfiniBand가 사용될 것으로 전망된다 [14]. InfiniBand는 P2P 방식으로 장비와 장비간의 연결을 지원하고, 현재 4X 속도로 10G/20G까지 지원된다. 특징으로는 모듈 타입으로 giga, SAN 등의 기존 네트워크와 통합 가능하고, RDMA (Remote Direct Memory Access)를 지원하기 때문에 CPU 오버헤드를 감소시킨다. 또한 기존의 MPI 지원을 하는 특징을 가지고 있다. 현재 Top 100 슈퍼컴퓨터에서도 InifiBand가 주로 사용되고 있고, 또한 사용이 증가되는 추세이기 때문에 고성능 컴퓨팅 시스템에서의 연결망에 있어서 보다 발전된 InfiniBand가 주로 활용될 것으로 전망된다.

5. 엑사스케일 컴퓨팅 시스템 자원관리 전망



(그림 15) 컴퓨팅 노드의 경량화 된 Linux



(그림 16) 서비스 노드의 customized Linux

5.1 Operating System (OS)

기존 고성능 컴퓨팅 시스템을 위한 OS 개발 동향은 컴퓨팅 노드와 서비스 노드로 나뉜다. 그림 15에서 보이는 바와 같이 컴퓨팅 노드에는 경량화 된 Linux를 사용하여 OS가 응용 프로그램 실행에 있어서 간섭을 안 하도록 하였다. 커널에는 최소한의 기능만을 구

현하고, 기존의 커널 기능을 유저 영역에서 실행을 시킴으로써 커널 오버헤드를 줄일 수 있고, 신뢰성을 높일 수 있게 된다.

(그림 16)에서 보이는 바와 같이 서비스 노드에서는 Customized Linux를 사용하여 각 컴퓨팅 노드로 태스크를 분배하는 역할을 담당하게 될 것이다. 태스크를 분배하기 전에 모델링과 시뮬레이션을 통해 최적화 된 자원 분배 정책을 정하여 성능을 최적화 시키는 역할을 하게 될 것이다.

차세대 고성능 혼합형 시스템의 OS에서는 새로운 구조를 지원하기 위한 자원 관리 기법이 추가되어야 할 것이다. 대표적으로 CPU와 GPU가 결합된 프로세서, 계층적 메인 메모리, 차세대 비휘발성 메모리 등을 고려한 자원 관리가 이루어져야 할 것이다. 또한 태스크 스케줄러 설계에 있어서 액세스 패턴 및 병렬성 등의 태스크의 특성을 고려하여 CPU 및 GPU에 할당하여야 할 것이고, 메모리 관리에 있어서도 계층적 메인 메모리에서 각 메모리의 특성을 고려하여 할당이 이루어져야 할 것이다. 또한 스토리지와 메인 메모리로 사용되는 차세대 비휘발성 메모리를 이를 위한 관리가 필요할 것이다.

5.2 메인 메모리 관리

메모리 관리에 있어서 계층적 메인 메모리를 효율적으로 사용하기 위한 관리 기법이 필요하다. 이를 위해 우선 M1, M2, M3 메모리 특성을 고려한 효율적인 메모리 할당 기술 개발 및 메모리 레벨에서 메모리 충돌 문제 해결 기술 개발이 필요하다. 또한 효율적인 메모리 할당 및 마이그레이션을 위해서 페이지 접근 모니터링 기법 개발이 필요하다. 이를 통해 페이지 별 접근 횟수 정보 및 페이지 주소 저장하고 각 페이지 내에서 관리가 가능한 범주를 파악할 수 있다. 마지막으로는 메모리 페이지 마이그레이션 기법 개발이 필요하다. 모니터링을 통한 동적 Hot/Cold 페이지 구분에 따른 M1, M2, M3 간 페이지 마이그레이션을 통해 메인 메모리 접근 속도를 증가시킬 수 있다.

또한 여러 노드의 메모리를 통합적으로 관리하는 방법이 필요하게 되는데, 한 노드에서 메모리 부족 시, 이웃 노드의 메모리를 활용하기 위한 관리가 이루어져야 한다. 이를 위해 메모리 모니터링 및 메모리 할당 정책이 필요하게 되고, 이를 통해 메모리 사용량을 높일 수 있게 될 것이다.

5.3 스토리지 관리

스토리지 관리에 있어서는 다른 compute node 상의 같은 job에서의 I/O request를 한 node에서 담당할 수 있도록 개발이 되어야 할 것이다. 이를 통해 버퍼 캐시의 pollution 제거 및 utilization 증가 효과를 볼 수 있게 된다. 또한, 차세대 스토리지를 SSD로 사용 시, 이는 HDD와는 다르게 특성에 대한 예측이 가능하기 때문에 SSD 특성을 고려하여 I/O reordering 및 I/O 재분배 기술을 적용하여 성능 향상을 가져올 수 있을 것이다.

또한, 차세대 비휘발성 메모리를 메인 메모리와 스토리로 동시에 사용하기 위해서는 우선 메모리 데이터와 파일시스템 데이터의 구분이 필요하고, 워크로드 특성에 따른 메인 메모리와 스토리지 사이의 동적인 용량 조절이 필요하다 [15]. 이를 통해 Page fault로 인한 스토리지로의 스왑 오버헤드를 줄여 성능 향상 가능하다. 이런 자원 관리를 통해 차세대 비휘발성 메모리 사용률 증가를 통한 성능 향상 및 신뢰성 증가시킬 수 있다.

5.4 프로그래밍 모델

우선 OpenMP [16]를 수정하여 다수의 노드들 간의 병렬적 프로그램 수행을 지원해야 할 것이고, 노드들 간의 역할 분담 및 해당 역할에 맞춘 OS 재구성이 필요하다. 또한 새로운 하드웨어를 고려한 프로그래밍 모델 설계가 필요하다. 새로운 하드웨어 고려사항으로는 CPU와 GPU가 통합된 프로세서가 있을 것이고 이를 고려하여 MPI (Message Passing Interface) [17] 수정이 필요하다. 계층적 메인 메모리, 차세대 비휘발성 스토리지를 지원하기 위한 프로그래밍 모델이 개발되어야 할 것이다.

6. 결론

본 논문에서는 기존 슈퍼컴퓨터 구조 및 자원 관리 분석을 통해 향후 미래의 고성능 엑사스케일 컴퓨팅

시스템의 구조 및 이를 위한 효율적인 자원 관리 방법을 전망하였다.

프로세서는 CPU만이 아닌 GPU가 결합된 이기종의 프로세서가 차세대 고성능 혼합형 시스템에서 사용 될 것이고, 프로세서 내 코어간, 컴퓨팅 노드 간 연결은 InfiniBand를 통해 연결됨으로써 높은 성능을 얻을 것으로 전망된다. 스토리지 시스템 구조는 분산 파일시스템인 Lustre FS 기반으로 새로운 스토리지 구조에 맞는 파일시스템 수정이 필요하고, 메인 메모리 구조는 3D-Stacked 메모리와 차세대 비휘발성 메모리가 결합되어 고성능, 대용량의 계층적 메인 메모리 구조가 구성될 것으로 보인다.

이 구조를 효율적으로 활용하기 위해, OS 디자인에 있어서 이를 서포트 하기 위한 자원 관리 기법이 추가되어야 할 것이다. 대표적으로 CPU와 GPU가 결합된 프로세서, 계층적 메인 메모리, 차세대 비휘발성 메모리 등을 고려한 자원 관리가 이루어져야 할 것이다. 또한 스케줄러 설계에 있어서 엑세스 패턴 및 병렬성 등 태스크의 특성 을 고려하여 CPU 및 GPU에 할당하여야 할 것이고, 메모리 관리에 있어서도 계층적 메인 메모리에서 각 메모리의 특성을 고려하여 할당이 이루어져야 할 것이다. 스토리지와 메인 메모리로 사용되는 차세대 비휘발성 메모리를 이를 위한 관리가 필요할 것이다. 프로그래밍 모델에서도 새로운 구조를 지원하기 위해 OpenMP 및 MPI의 수정이 필요하다.

Acknowledgement

본 연구는 지식경제부 및 한국산업기술평가관리원의 IT산업원천기술개발사업의 일환으로 수행하였음. [10038768, 유전체 분석용 슈퍼컴퓨팅 시스템 개발]

참고문헌

〔1〕 J. Dongarra et al. "The International Exascale Software Project Roadmap", IESP Workshop 2011.
〔2〕 Top500 Supercomputer Sites, http://www. top500.org.
〔3〕 K Computer, http://www.fujitsu.com/global/ about/ tech/k.
〔4〕 Tianhe-1A, http://nscc-tj.gov.cn/en.
〔5〕 Jaguar, http://www.nccs.gov/jaguar.
〔6〕 J. Appavoo et. al., "Project Kittyhawk: Building a Global -Scale Computer", SIGOPS 2008.
〔7〕 R. Brightwell et. al., "Why Nobody Should Care About Operating Systems for Exascale", ROSS 2011.
〔8〕 K. H. Park et. al., "MN-Mate: Resource Management of Manycores with DRAM and Nonvolatile Memories", HPCC 2010.
〔9〕 K. Bergman et. al., "Exascale Computing Study: Technology Challenges in Achieving Exascale Systems", 〔Online〕. Available: http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf
〔10〕 Intel SandyBridge, http://www.intel.com.
〔11〕 G. H. Loh et. al.,"3D-stacked Memory Architectures for Multi-Core Processors",ISCA 2008.
〔12〕 Y. Park et. al., "Linux Kernel Support to Exploit Phase Change Memory", OLS 2010.
〔13〕 Y. Park et. al., "Power-Aware Memory Management for Hybrid Main Memory", ICNIT 2011.
〔14〕 A. Bechtolsheim "Technologies for Data-Intensive Computing," 14th International workshop on high performance transaction systems (HPTS) 2009.
〔15〕 J.-Y. Jung et. al., "Dynamic Co-Management of Pers55istent RAM Main Memory and Storage Resources", CF 2011.
〔16〕 OpenMP, http://openmp.org.
〔17〕 The Message Passing Interface (MPI) Standard, http://www.mcs.anl.gov/research/ projects/mpi.

저자소개

◆ 박 성 규

- 2007년 KAIST 전기및전자공학과 학사
- 2009년 KAIST 전기및전자공학과 석사
- 2009년~현재 KAIST 전기및전자공학과 박사과정
- 관심분야: 플래시 메모리, 차세대 비휘발성 메모리 등

◆ 박 기 웅

- 2005년 연세대학교 전산학과 학사
- 2007년 KAIST 전기및전자공학과 석사
- 2012년 KAIST 전기및전자공학과 박사
- 관심분야: 클라우드 컴퓨팅, 보안 프로토콜, 모바일 시스템 등

◆ 박 규 호

- 1973년 서울대학교 전자공학과 학사
- 1975년 KAIST 전기및전자공학과 석사
- 1983년 Parix XI 대학 컴퓨터공학과 박사
- 1983년 ~ 현재 KAIST 전기 및 전자공학과 교수
- 관심분야: 운영체제, 매니코어 시스템, 차세대 비휘발성 메모리, 플래시 메모리 등