ISSN(Print): 2287-5905

http://dx.doi.org/10.3745/KTSDE.2013.2.2.081

Design of MAHA Supercomputing System for Human Genome Analysis

Young Woo Kim[†] · Hong-Yeon Kim[†] · Seungjo Bae[†] · Hag-Young Kim[††]
Young-Choon Woo[†††] · Soo-Jun Park[††††] · Wan Choi[†††††]

ABSTRACT

During the past decade, many changes and attempts have been tried and are continued developing new technologies in the computing area. The brick wall in computing area, especially power wall, changes computing paradigm from computing hardwares including processor and system architecture to programming environment and application usage. The high performance computing (HPC) area, especially, has been experienced catastrophic changes, and it is now considered as a key to the national competitiveness. In the late 2000's, many leading countries rushed to develop Exascale supercomputing systems, and as a results tens of PetaFLOPS system are prevalent now. In Korea, ICT is well developed and Korea is considered as a one of leading countries in the world, but not for supercomputing area. In this paper, we describe architecture design of MAHA supercomputing system which is aimed to develop 300 TeraFLOPS system for bio-informatics applications like human genome analysis and protein-protein docking. MAHA supercomputing system is consists of four major parts - computing hardware, file system, system software and bio-applications. MAHA supercomputing system is designed to utilize heterogeneous computing accelerators (co-processors like GPGPUs and MICs) to get more performance/$, performance/area, and performance/power. To provide high speed data movement and large capacity, MAHA file system is designed to have asymmetric cluster architecture, and consists of metadata server, data server, and client file system on top of SSD and MAID storage servers. MAHA system softwares are designed to provide user-friendliness and easy-to-use based on integrated system management component - like Bio Workflow management, Integrated Cluster management and Heterogeneous Resource management. MAHA supercomputing system was first installed in Dec., 2011. The theoretical performance of MAHA system was 50 TeraFLOPS and measured performance of 30.3 TeraFLOPS with 32 computing nodes. MAHA system will be upgraded to have 100 TeraFLOPS performance at Jan., 2013.

Keywords : Genome Analysis, Bio-Informatics, Supercomputer, MAHA Supercomputer, Heterogeneous

대용량 유전체 분석을 위한 고성능 컴퓨팅 시스템 MAHA

김 영 우[†] · 김 홍 연[†] · 배 승 조[†] · 김 학 영[††] · 우 영 춘[†††] · 박 수 준[††††] · 최 완[†††††]

요 약

지난 10여년 동안 컴퓨팅 분야는 다양한 연구와 변화를 통하여 눈부신 발전을 이루어오고 있다. 반도체 기술의 발전은 프로세서 및 시스템 아키텍처, 프로그래밍 환경 등에 새로운 패러다임의 변화를 야기하고 있다. 특히 고성능컴퓨팅(HPC)분야는 첨단 기술이 집적된 분야로써, 한 국가의 경쟁력으로 간주되고 있다. 2000년대 후반부터 선진 국가들은 Exascale의 슈퍼컴퓨팅 기술의 개발에 박차를 가하고 있으나, 한국의 경우 ICT 분야에 집중하여 관련 핵심기술의 확보가 시급한 상황이다. 본 논문에서는 슈퍼컴퓨팅 기술을 확보하고 대규모 유전체 분석 및 단백질 구조 분석을 위한 고성능 컴퓨팅 시스템인 MAHA 슈퍼컴퓨팅 시스템의 아키텍처를 제시하고 설계 및 구현에 관하여 서술한다. MAHA 슈퍼컴퓨팅 시스템은 컴퓨팅 하드웨어, 파일 시스템, 시스템 소프트웨어 및 바이오 응용으로 구성되며, 성능/$, 성능/면적 및 성능/전력을 향상시키기 위한 이종 매니코어 연산장치에 기반 한 고성능 컴퓨팅 구조를 설계하였다. 대규모 데이터에 대한 빠른 처리를 위하여 SSD 및 MAID시스템에 기반 한 고성능 저전력 파일시스템과 사용자 편의성 및 이종 매니코어 자원의 효과적인 활용을 통한 바이오 응용 성능 향상을 위한 시스템 소프트웨어를 설계하였다. 2011년 12월 MAHA 슈퍼컴퓨팅 시스템은 32개의 컴퓨팅 노드에 기반 하여 이론 성능 50 테라 플롭스, 실측 성능 30.3 테라 플롭스(시스템 효율 56.2%)로 설계, 구축 되었으며, 2013년 100 테라 플롭스 규모로 확장될 예정이다.

키워드 : 유전체 분석, 바이오인포매틱스, 슈퍼컴퓨터, MAHA 슈퍼컴퓨팅 시스템, 이기종

※ 이 논문은 지식경제부 및 한국산업기술평가관리원의 IT산업원천기술 개발사업의 일환으로 수행하였음.
† 정 회 원 : 한국전자통신연구원 책임연구원
†† 정 회 원 : 한국전자통신연구원 서버플랫폼연구팀 팀장
††† 정 회 원 : 한국전자통신연구원 고성능컴퓨팅시스템연구팀 팀장
†††† 정 회 원 : 한국전자통신연구원 바이오의료IT융합연구부장
††††† 정 회 원 : 한국전자통신연구원 클라우드컴퓨팅연구부장
논문접수 : 2013년 1월 8일
수 정 일 : 1차 2013년 1월 11일
심사완료 : 2013년 1월 11일
* Corresponding Author : Young Woo Kim(bartmann@etri.re.kr)

1. 서 론

지난 수년간 컴퓨팅 기술은 다양한 관련 기술 분야의 발전에 힘입어 많은 발전을 이루어 오고 있다. 특히 2000년대 초반 100nm이었던 반도체 공정 기술은 지속적으로 무어의 법칙에 의해, 2012년 현재 22nm공정 까지 집적도의 향상 및 성능 향상을 이루어 오고 있다. 반면, 반도체 공정의 미세화가 진행됨에 따라 다양한 기술 측면에서 성능 향상의 걸림돌이 속출하고 있으며, 이를 해결하기 위한 다양한 연구가 진행 중에 있다. 특히, 고집적화로 인한 전력 소모 및 발열 문제는 컴퓨팅 기술의 근간의 하나인 프로세서 설계 및 구조에 대한 패러다임을 바꾸어 가고 있는 중이다.

슈퍼컴퓨팅 분야는 이와 같은 컴퓨팅 기술의 최첨단 기술이 종합된 분야로서, 2008년 미국 DARPA의 Exascale computing에 관한 보고서 발표 이래, 세계 유수의 국가들은 최첨단 기술을 동원하여 ExaFLOPS급의 슈퍼컴퓨터 개발에 박차를 가하고 있는 중이다[1, 2]. 반면 국내의 경우 슈퍼컴퓨터의 기술 개발보다는 도입을 통한 활용측면에 중점을 두어 기술수준의 발전이 저조한 형편이나, 최근 관련 연구에 대한 관심과 연구 활동이 증가하고 있다.

본 논문에서는 이와 같은 슈퍼컴퓨팅 관련 기술 확보를 위하여 2011년부터 개발 중인 MAHA(MAny-core HPC system for bioinformatics Applications)슈퍼컴퓨팅 시스템의 설계 및 구조에 대하여 설명한다[3]. 제 2장에서는 MAHA 슈퍼컴퓨팅 시스템의 설계에 앞서, 최근의 슈퍼컴퓨팅 기술 및 관련 기술의 변화에 대하여 간략히 설명한다. 제 3장에서는 MAHA 슈퍼컴퓨팅 시스템의 구조와 관련 설계 내용을 서술하고, 제 4장에서 MAHA 슈퍼컴퓨팅 시스템의 성능 평가 결과를 설명하도록 한다.

2. 슈퍼컴퓨팅 기술 동향

본 장에서는 슈퍼컴퓨팅 기술과 관련한 최근 동향 및 기술 변화에 대하여 설명하도록 한다.

2.1 슈퍼컴퓨팅 기술 동향

전술한 바와 같이 최근의 반도체 기술 변화는 컴퓨팅 분야에 많은 영향을 끼치고 있다. 특히 고집적에 따른 전력 소모의 증가는 반도체 회로의 속도 상승을 통한 성능 개선에 중요한 장애 요인(전력장벽)으로 작용하고 있다(Fig. 1)[1, 4, 5]. 소모 가능 전력의 제한으로 프로세서의 동작 속도는 더 이상 증가하지 않고 있으며, 멀티/매니 프로세서와 같은 프로세서 구조의 변화를 불러오고 있다.

슈퍼컴퓨터 시스템은 전통적인 벡터프로세서에서 발전하여 현재는 상용 프로세서 기반 클러스터 구조의 슈퍼컴퓨터가 널리 개발되고 있다(Fig. 2). 특히 그래픽 기술에 기반한 GPGPU(General Purpose computing on GPU), MIC (Many Integrated Core)와 같은 기술의 등장으로 인하여, 매니코어 기반의 이종 컴퓨팅 기술을 채용한 슈퍼컴퓨터 개발이 활발히 진행되고 있다[5, 6]. GPGPU를 채용한 슈퍼컴퓨터는 2010년 처음 등장하여, 2012년 현재 시스템 대수로 53대(12.4%), 성능으로는 약 31.8 PetaFLOPS(33%)를 차지하고 있다[6]. 이와 같은 매니코어 기반의 이종 컴퓨팅 기술은 시스템 구조, 병렬/시스템 소프트웨어 등의 분야에 대한 새로운 연구를 촉진하는 계기가 되었다[3, 4]. 세계 주요 선진 국가는 현재 ExaFLOPS급의 슈퍼컴퓨터 기술 개발을 진행 중이며, ExaFLOPS 슈퍼컴퓨터 개발을 통하여 매니코어, 고도병렬 처리 등 각 컴퓨팅 분야의 첨단 기술 확보에 집중하고 있다.

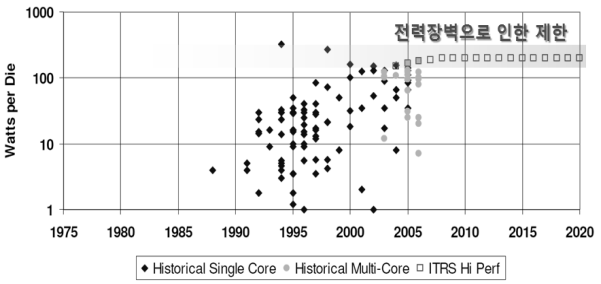

Fig. 1. Limitation due to power wall [1, 4]

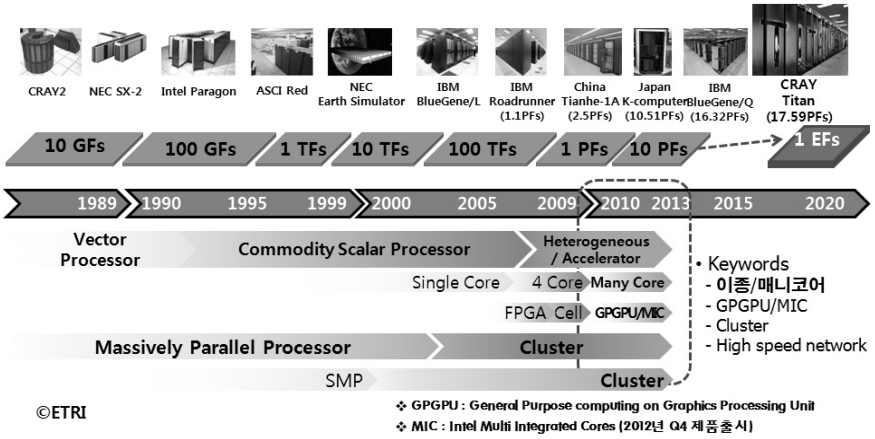

Fig. 2. Trends of supercomputing technology [5]

2.2 한국 슈퍼컴퓨팅 현황

슈퍼컴퓨팅 분야에서의 강국은 미국으로써, 2012년 말 현재 시스템 수로는 251대, 성능으로는 89.1 PetaFLOPS로서 절대적인 우위를 차지하고 있다. 반면 한국의 경우 시스템 수 4대, 성능 1.0 PetaFLOPS를 보이고 있다(Table 1)[6].

다음의 Table 2는 슈퍼컴퓨팅 기술과 관련한 주요 국가의 기술 수준을 나타낸 표로서, 일본 과학기술진흥기구에서 2011년 발표한 "과학기술·연구개발의 국제비교" 자료에 기초한 것이다[8]. Table 1과 Table 2에서 알 수 있듯이, 한국의 슈퍼컴퓨터 관련 기술의 수준은 국가 경제 규모에 비하여 상당히 낮은 수준으로 평가되고 있음을 알 수 있다.

반면 일본은 2011년 K-Computer 등의 개발을 통하여 2002년의 Earth Simulator 이후 다시 슈퍼컴 1위를 탈환하였으며, 특히 중국은 2010년 Tianhe-1A 슈퍼컴퓨터를 통하여 처음으로 세계 1위에 등극하는 등 기술개발에 박차를 가하고 있다.

Table 1. Summary of supercomputing at Top500 [6]

	Number of systems		System performance	
	Number	%	PetaFLOPS	%
US	251	50.2 %	89.1	55.0 %
Japan	32	6.4 %	19.4	12.0 %
Europe	105	21.0 %	33.3	20.5 %
China	72	14.4 %	12.3	7.6 %
India	8	1.6 %	1.1	0.7 %
Korea	4	0.8 %	1.0	0.6 %

Table 2. Comparison of supercomputing technology [5, 7, 8]

	Research		Technology		Industrial	
	Status	Trend	Status	Trend	Status	Trend
US	◎	→	◎	→	◎	→
Japan	◎	↗	◎	↗	◎	→
Europe	○	→	○	→	○	→
China	○	↗	◎	↗	○	↗
Korea	X	↘	X	→	△	→

◎ : Showing significant improvement, ○ : Improving, △ : Delaying, X : Significantly delayed
↗ : Uptrend, → : Status quo, ↘ : Downward trend

3. MAHA 슈퍼컴퓨팅 시스템 구조 및 설계

3.1 MAHA 슈퍼컴퓨팅 시스템 개요

MAHA 슈퍼컴퓨팅 시스템은 대규모의 유전체 분석을 고속으로 처리하는 것을 목표로 개발되고 있다. 최초의 인간 유전체 분석은 1990년에서 2003년까지 10여 년간 30억불의 투자를 통하여 인간 유전체 전체를 분석 완료하였다[9]. 이후 유전체 시퀀싱 기술의 발전과 컴퓨터에 기반 한 바이오인포매틱스 기술의 발전으로 인하여, 유전체 시퀀싱 비용은 점차 감소하여 1,000불미만으로 한사람의 유전체를 시퀀싱 할 수 있는 시대가 도래하고 있다. 반면, 시퀀싱 된 유전체 데이터가 증가함에 따라, 유전체 정보를 분석하기 위한 고성능 컴퓨터의 수요가 증가하고 있다(Fig. 3)[10].

MAHA 슈퍼컴퓨팅 시스템은 대규모의 유전체 시퀀스 정보를 빠르게 분석하여 인간의 유전체를 분석하기 위한 시스템이다. 대규모 유전체를 고속으로 분석하기 위해서는 고속의 데이터처리와 고속의 파일시스템, 효과적인 시스템 관리 소프트웨어와 유전체 분석 알고리즘을 필요로 한다. MAHA 슈퍼컴퓨팅 시스템은 GPGPU와 MIC기술에 기반 한 이종 매니코어 연산장치를 사용하여 이론 성능 300 TeraFLOPS를 가지도록 설계하였다. 다음의 Table 3은 MAHA 슈퍼컴퓨터의 설계 사양이다.

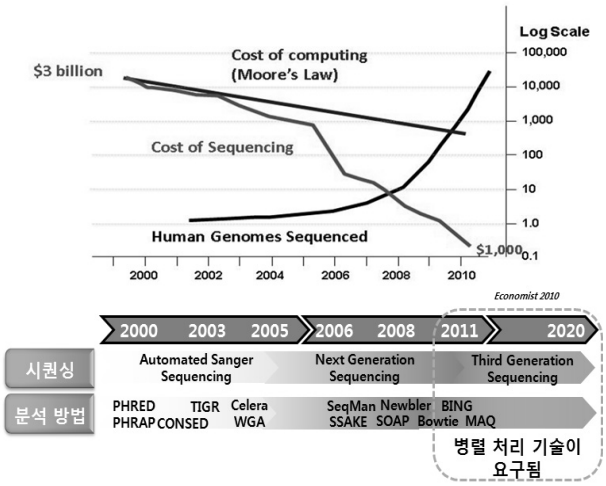

Fig. 3. Trends in genome analysis [10]

Table 4는 일반 CPU 노드와 매니코어 연산장치에 기반한 노드간의 비교를 나타낸 표이다. 매니코어 연산장치는 CPU에 비하여 많은 코어를 내장하여 단위 면적, 단위 노드당 일반 CPU 노드에 비하여 높은 성능을 가진다. 일례로서 Table 4에서 알 수 있듯이 CPU노드 대비 성능은 약 8배, 에너지 소모 약 1/3배, 면적은 약 1/8배로 CPU노드에 비하여 좋은 특성을 가진다.

MAHA 슈퍼컴퓨팅 시스템은 크게 MAHA 컴퓨팅 플랫폼, MAHA 파일 시스템, MAHA 시스템 소프트웨어, MAHA 바이오 응용의 4개 블록으로 구성된다. MAHA 컴퓨팅 플랫폼은 이종 매니코어 기반의 컴퓨팅 파워와 시스템 네트워크를 제공하며, MAHA 파일 시스템은 SSD와 HDD에 기반 한 고속, 대용량 스토리지를 공급한다. MAHA 시스템 소프트웨어와 MAHA 바이오 응용은 이종자원에 기반한 유전체 분석의 고속 처리를 위한 시스템 소프트웨어와 유전체 분석 및 단백질 구조 분석 응용 소프트웨어를 제공한다. 다음의 Fig. 4와 Fig. 5는 MAHA 슈퍼컴퓨팅 시스템의 HW 및 SW 구조를 나타낸 그림이다.

3.2 MAHA 컴퓨팅 플랫폼

MAHA 컴퓨팅 플랫폼은 매니코어 연산장치에 기반하여, MAHA 시스템의 연산능력을 제공하는 부분이다. 컴퓨팅 플랫폼을 구성하는 요소는 크게, 관리노드, 컴퓨팅 노드, 네트워크 및 스토리지노드로 구성된다. 이중 관리노드와 컴퓨팅 노드는 동일한 CPU 노드로 구성되며 컴퓨팅 노드는 매니코어 연산장치를 탑재한다.

1) 컴퓨팅 노드 구조 설계

Table 4에서와 같이 단위 컴퓨팅 노드의 연산성능을 향상시키기 위해서는 많은 수의 매니코어 연산장치를 사용하는 것이 바람직하다. 그러나 CPU 구조 및 IO 연결의 형상에 따라 실제 성능에 제약사항이 발생 할 수 있다. MAHA 컴퓨팅 플랫폼의 노드 구조 설계에는 다음과 같은 사항을 고려하였다.

Table 3. Design specifications of MAHA supercomputer

	Index	Spec.	Units
Computing	Performance	300	TeraFLOPS (Rpeak)
	CPU Cores	> 2,000	core
	Accelerator Cores	> 200,000	core
	Memory	> 4.3	TeraBytes
Storage	Performance	> 1,000,000	IOPS
	SSD Capacity	144	TeraBytes
	HDD Capacity	960	TeraBytes
Network	Computational	> 40	Gbps
	Management	1/10	Gbps
Power	Total Max. Power	230	kW

Table 4. Comparison between CPU based node and many-core based node

	CPU node(1)	Many-core node(2)	Ratio (CPU:Manycore)	Units
Perfor-mance	0.332	2.67	1 : 8.04	TeraFLOPS/node
Cores	16	5008	1 : 313	cores/node
Energy	692.7	237.3	1 : 0.34	Watt/TeraFLOPS
Space(3)	7.17	0.89	1 : 0.12	Racks (42U)
Cost(4)	8,024	3,636	1 : 0.45	$/TeraFLOPS

(1) CPU based node : Dual Intel Xeon E5 (2.6GHz) [11]
(2) Many-core based node : Dual Intel Xeon E5 (2.6GHz) + Dual NVIDIA K20 GPGPU [12]
(3) Number of racks for 100 TeraFLOPS
(4) Cost for 1 TeraFLOPS, only for CPU ($1,333/CPU), GPGPU($4,000 estimated)

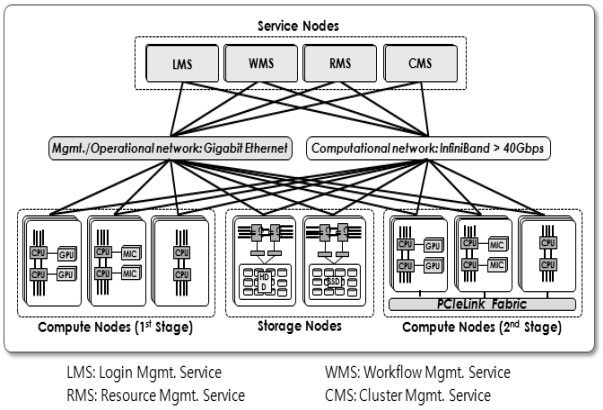

Fig. 4. Basic structure of MAHA supercomputing HW

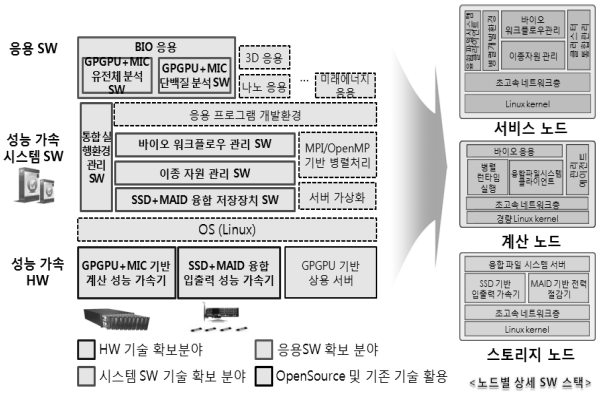

Fig. 5. Basic structure of MAHA supercomputing SW

a) 메모리 대역폭 : CPU-메모리
b) 매니코어 연산장치 대역폭 : CPU-매니코어장치
c) 네트워크 장치 대역폭 : CPU-네트워크장치

Intel CPU를 기준으로 할 때 프로세서의 세대별로 시스템의 구성 및 성능에 차이가 있다. MAHA 시스템의 설계에서는 상기의 3가지 대역폭을 기준으로 프로세서와 매니코어 연산장치 기반의 노드 구조를 설계하였다.

다음의 Fig. 6과 Table 5는 다양한 프로세서 및 매니코어 연산장치, 메모리와의 구조와 그에 따른 성능 평가를 보여준다.

Fig. 5. Comparison of various computing node structure

	Bandwidth (Memory)	Bandwidth (per manycore device)	Bandwidth (Network device)
6A	102.4 GB/s	16 GB/s	8 GB/s
6B	31.9 GB/s	8 GB/s	4 GB/s
6C	63.8 GB/s	8 GB/s	4 GB/s
6D	31.9 GB/s	4 GB/s	4 GB/s
6E	31.9 GB/s	2 GB/s	4 GB/s

SB-EP : Sandy Bridge-EP, W-EP : Westmere-EP [13], IOH : IO Hub

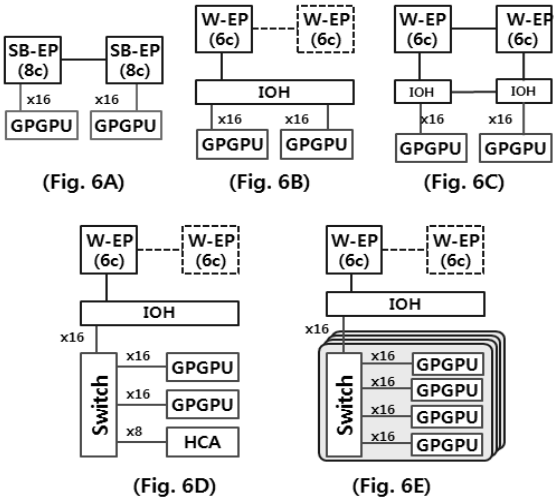

Fig. 6. Various structures of computing node

매니코어 연산장치는 CPU의 IO장치 중 하나로 사용되며 따라서 PCI Express x16 슬롯을 통하여 연결된다. 매니코어 연산장치는 CPU외부의 연산 코어를 사용하므로, 연산시 필요한 데이터를 매니코어 연산장치로 이동하여 연산을 수행하고, 그 결과를 CPU의 메모리로 전달하여야 한다. 이때 많은 량의 데이터 이동이 필요하며 이를 지원하기 위해서는 매니코어 연산장치의 대역폭을 충분히 확보하여야 한다.

Intel 계열의 프로세서에서는 1개의 CPU 혹은 IOH에서 40 lane의 PCI Express 연결을 제공한다[11, 13, 14]. 따라서 특별한 장치(스위치)를 사용하지 않을 경우, CPU당 최대 2개의 매니코어 연산장치의 연결이 가능하다. 이를 초과하여 매니코어 장치를 연결할 경우, 스위치 장치를 사용하여 PCI Express를 확장하여야 하는데, 이 경우 스위치의 up-link 연결의 대역폭 제한으로 매니코어 장치의 유효 대역폭이 스위치에 연결된 장치 수에 반비례(1:n)하여 줄어들게 된다. MAHA 시스템에서는 연산 성능을 최대한 보장하기 위하여 CPU와 매니코어 연산장치의 비율을 1:2 이하로 제한하였다.

Intel Sandy Bridge-EP 프로세서는 CPU core와 메모리 제어기, IO 버스가 통합된 8 core 프로세서로서 메모리 대역폭과 PCI Express 대역폭이 크게 증가하였다.

$$\text{Memory bandwidth} = 4 \times 8 \times 1{,}600 = 51.2 \text{ GB/s} \quad (1)$$

$$\text{IO bandwidth} = 8 \times 40 = 320 \text{ Gbps} \quad (2)$$

MAHA 컴퓨팅 노드의 구조는 상기의 3가지 대역폭을 고려하여 Fig. 7과 같은 구조로 설계하였다. 관리노드는 컴퓨팅 노드에서 매니코어 연산장치를 제외한 구조를 사용한다.

2) 시스템 구조 설계

MAHA 슈퍼컴퓨팅 시스템은 클러스터 기반의 HPC 구조를 기본으로 한다. 시스템은 관리노드, 컴퓨팅노드, 네트워크 및 스토리지 노드로 구성 된다(Fig. 4).

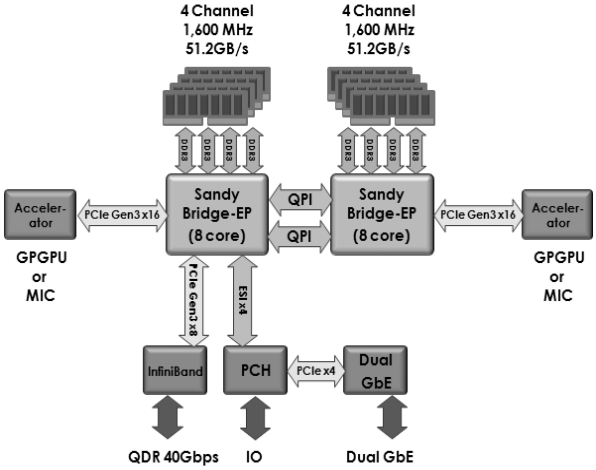

Fig. 7. Designed structure of MAHA computing node

관리노드는 1대의 로그인노드, 2대의 active-standby 관리노드의 3개의 노드가 하나의 관리노드 그룹을 이뤄 시스템을 관리하도록 설계하였다. Active-standby 형상의 관리노드는 시스템의 가용성과 안정성을 제공한다.

스토리지 노드는 고성능과 대용량을 제공하기 위하여 SSD 서버와 HDD기반의 MAID(Massive Array of Idle Disk) 서버를 설계하고, 이를 기반으로 시스템에 데이터를 공급한다.

시스템 네트워크는 관리 네트워크와 연산 네트워크로 나누어지며, 관리 네트워크는 1/10 GbE의 계층 구조로 설계되었으며 관리노드를 통하여 시스템의 상태관리, Job 스케줄링, 자원관리를 수행한다. 연산 네트워크는 고속의 연산 및 데이터 전송을 위한 InfiniBand 네트워크로 설계하였다. InfiniBand로 구성된 연산 네트워크는 고속의 데이터 전송을 위하여 Folded-Clos 토폴로지로 구성되며, non-blocking 성능을 가지도록 설계하였다. InfiniBand와 별도로 독자 네트워크인 PCIeLINK Fabric 설계 및 구현 중에 있으며, PCIeLINK 네트워크는 PCI Express 기술을 활용하여 컴퓨팅 노드간의 고속의 지역 네트워크로 활용될 예정이다.

Table 6. MAHA system network

Network	Type	Speed	Topology
Management	Ethernet	1/10 Gbps	Tree
Computational	InfiniBand	40 Gbps (QDR)	Folded-Clos
Computational	PCIeLink	128/256 Gbps	Mesh/Torus

3) 시스템 공조 설계

MAHA 슈퍼컴퓨팅 시스템은 일반 서버와 달리, 매니코어 연산장치를 고집적하여 연산성능을 향상시키도록 설계된 시스템으로써, 같은 대수의 상용 서버에 비하여 상대적으로 많은 전력을 사용한다. 서버의 운용에 따른 효과적인 공조를 위하여 MAHA 슈퍼컴퓨팅 시스템을 위한 공조 시스템을 설계하였다.

MAHA 슈퍼컴퓨팅 시스템은 공랭식 냉각 방식을 기본으로 하며, 한국의 기후 특성에 맞추어 공조시스템의 전력 소모를 최소화 하도록 설계하였다[15]. 공조는 두열의 시스템 랙을 back-to-back 배열하여 뜨거워진 공기가 중앙으로 모여 상부의 배기구를 통하여 배출되는 hot zone 방식의 구조로 설계하였다. 또한, 서버실의 내부온도와 외부온도를 비교하여 외부온도가 충분히 낮을 경우, 외부의 차가운 공기를 서버실로 공급함으로써 불필요한 냉각 전력을 절감하도록 설계하였다(Fig. 8).

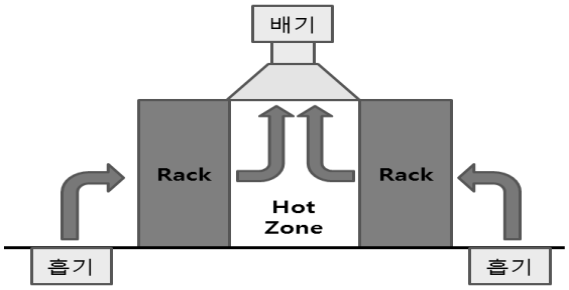

Fig. 8. Design of HVAC for MAHA supercomputing system

3.3 MAHA 파일 시스템

MAHA 파일 시스템은 독자 구조의 파일 시스템으로써, 클라우드 파일 시스템인 글로리 파일 시스템을 기반으로 하여 HPC 파일 시스템에서 요구하는 메타데이터 처리, 순차 입출력, 랜덤 입출력 성능을 향상시킨 파일 시스템이다(Fig. 9).

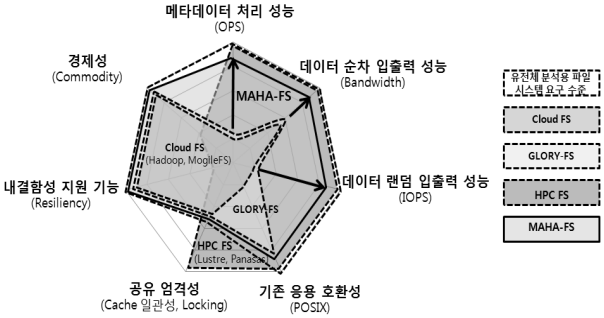

Fig. 9. Comparison between MAHA file system and general HPC file system

MAHA 파일 시스템은 메타데이터 서버, SSD와 HDD로 구성된 데이터 서버, 그리고 클라이언트 파일 시스템으로 구성된다. 데이터 서버는 범용 서버에 기반, 복수개의 RAID 제어기와 SDD 및 HDD의 장착이 가능한 백플레인으로 설계하여 높은 성능을 보장함과 동시에 시스템의 cost를 최소화 하도록 설계하였다.

MAHA 파일 시스템은 Fig. 10에서와 같이 데이터의 특성(접근 빈도, 크기)에 따라서 계층적으로 데이터 서버에 저장 및 접근되는 구조로 설계하였다. 컴퓨팅 노드에 즉시적으로 사용되며, 사용빈도가 높은 고성능 데이터는 SSD 기반의 스토리지 서버에 저장하여 시스템의 성능을 향상시키며, 데이터의 특성에 따라 HDD 병렬 파일시스템, MAID 파일 시스템으로 데이터를 동적 배치 및 이동하여 시스템의 성능, 용량을 보장함과 동시에 저전력을 구현하도록 설계하였다. 특히 MAID 파일 시스템(HW 포함)은 데이터의 특성에 기반 하여 HDD의 동작과 전력을 제어하여 시스템 수준의 전력 소모를 절감하도록 설계하였다.

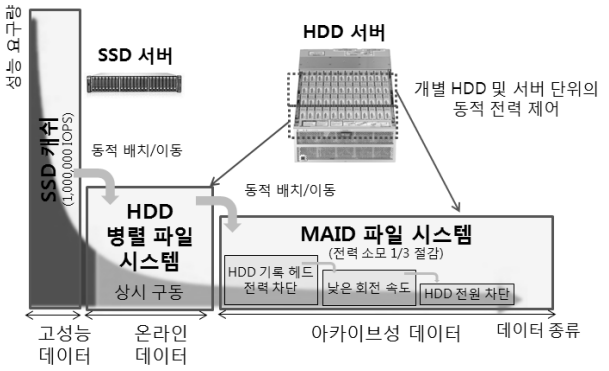

Fig. 10. Structural design of MAHA file system

3.4 MAHA 시스템 소프트웨어

MAHA 시스템 소프트웨어는 이종 매니코어 연산장치에 기반 한 MAHA 시스템에서 워크플로우 형태의 유전체 데이터 분석 및 단백질 구조 분석을 편리하고 효율적으로 수행할 수 있도록 설계되었다. 이에 따라 MAHA 시스템 소프트웨어는 HPC 환경을 위한 바이오 워크플로우 관리 기능, 바이오 응용의 특성을 반영한 자원 관리 기능, 그리고 MAHA 시스템을 위한 클러스터 통합 관리 기능을 제공한다(Fig. 11). 따라서 HPC 시스템에 대한 충분한 지식이 없는 사용자도 MAHA 시스템을 편리하게 사용할 수 있고, 또한 바이오 응용에 특화된 자원 관리 기능을 통해 성능 향상을 기대할 수 있다.

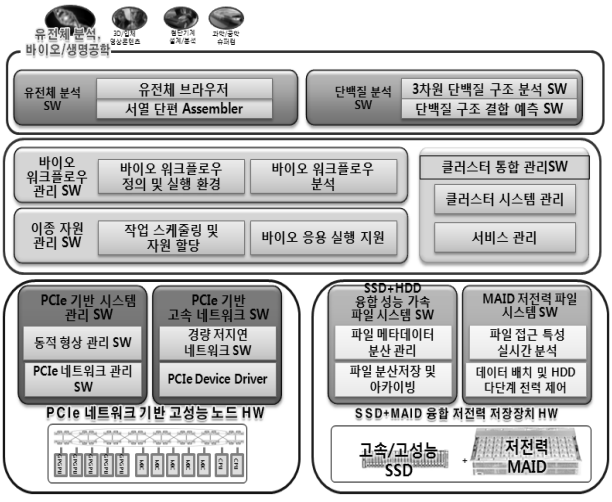

Fig. 11. Structure of MAHA system SW stack

3.5 MAHA 바이오 응용 SW

MAHA 바이오 응용 SW는 MAHA 시스템을 활용하여 고속으로 인간 유전체 분석을 수행하는 것을 목표로 한다. 유전체 분석은 생물 자료 시편(타액, 혈액 등)을 이용하여 수백만 개의 유전체단편 정보(수십 ~ 수백 개 길이의 유전체 정보를 가진 리드 데이터) 추출하여 이를 유전체 분석을 위한 원 자료로 이용한다. 각 유전체 단편 정보는 순서 재정렬, 중복성 제거 등을 거쳐 하나의 유전체 지도로 만들어진다(리드 매핑 단계). 만들어진 유전체 지도는 다양한 분석을 위한 포맷 변환을 거치고(포맷 컨버팅 단계), 전체 유전체 지도에서 유전적으로 의미 있는 정보(SNP, Single Nucleotide Polymorphism)를 추출한다(SNP Calling 단계). 추출된 SNP를 바탕으로 개개인의 특성에 따른 변이 정보를 분석하고, 이를 기반으로 유전적인 특징을 찾는 단계(SNP 해석)를 거쳐 개인에 따른 유전적인 특성을 파악하게 된다.

MAHA 바이오 응용 SW는 이와 같은 단계에 따른 바이오 파이프라인으로 구성되며, 각 바이오 파이프라인의 처리 단계에 따른 특성(compute intensive, IO intensive, memory intensive 등)에 맞도록 파이프라인의 구성과 이종자원의 활용을 최적화 하도록 설계 되었다. 다음의 Fig. 12는 이와 같은 MAHA 바이오 파이프라인을 나타낸 그림이다.

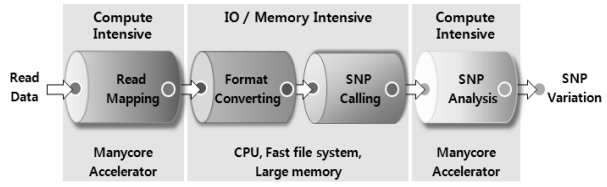

Fig. 12. Structure of MAHA Bio-pipeline

4. MAHA 슈퍼컴퓨팅 시스템 구현 및 성능 평가

본 장에서는 MAHA 슈퍼컴퓨팅 시스템의 구현과 성능 평가 결과에 대하여 서술한다.

4.1 MAHA 슈퍼컴퓨팅 시스템 구현

MAHA 슈퍼컴퓨팅 시스템은 Table 3의 사양과 같이 이론 성능 300 TeraFLOPS의 시스템 구현을 단계별로 개발하는 것을 목표로 하고 있다. 이에 따라 2011년도에 이론 성능 50 TeraFLOPS의 시스템을 설계, 구축 하였다. Fig. 13은 MAHA 슈퍼컴퓨팅 시스템의 개발 로드맵이다.

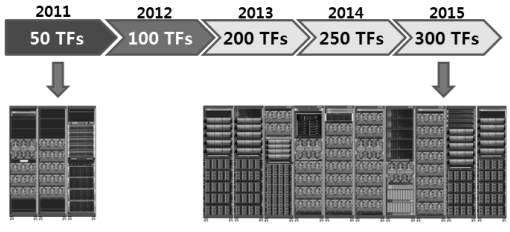

Fig. 13. Roadmap of MAHA supercomputing system

1) 50 TeraFLOPS MAHA 슈퍼컴퓨팅 시스템

2011년 50 TeraFLOPS 규모의 MAHA 시스템은 MAHA 시스템의 개발을 위한 테스트플랫폼 성격의 시스템으로, 설계한 하드웨어 플랫폼의 성능 평가와 시스템 소프트웨어 및 기술 개발을 위한 개발 플랫폼으로 활용되고 있다.

MAHA 슈퍼컴퓨팅 시스템은 Fig. 4와 Fig. 7의 구조로 설계한 컴퓨팅 플랫폼을 기반으로 하여 시스템의 구조를 구현하였다. 컴퓨팅 시스템은 집적도의 향상과 관리의 용이성 등을 고려하여 블레이드 서버 형태로 개발하였다. 다음의 Fig. 14와 Fig. 15는 50 TeraFLOPS MAHA 슈퍼컴퓨팅 시스템의 시스템 구현 개념도와 구현한 MAHA 시스템이다.

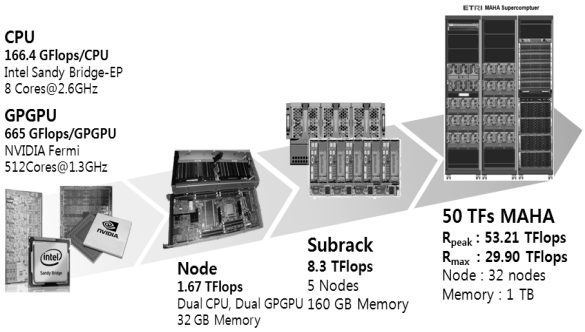

Fig. 14. Construction of 50 TeraFLOPS MAHA supercomputing system

컴퓨팅 노드는 2개의 8 코어 Intel Xeon E5 프로세서와 2개의 NVIDIA의 M2090 GPGPU를 하나의 노드에 집적하여 구현하였다. 1개의 컴퓨팅 노드의 이론 성능은 총 1.67 TeraFLOPS를 제공하며, 1,040개의 코어, 32GB의 1,600MHz DDR3 메모리를 내장하였다. 컴퓨팅 노드는 5개의 노드가 1개의 서브랙을 구성하고, 총 32개 노드, 이론 성능 53.2 TeraFLOPS, 1 TB 메모리의 시스템을 구축하였다. 시스템의 스토리지로는 SSD 기반의 14 TB, HDD기반의 70TB 용량을 제공하며, 시스템 네트워크로 1/10GbE와 40Gbps InfiniBand 네트워크를 구현하였다.

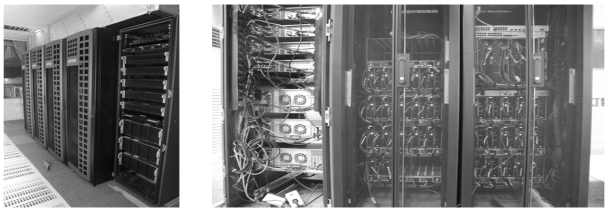

Fig. 15. Implementation of 50 TeraFLOPS MAHA supercomputing system

2) 100 TeraFLOPS MAHA 슈퍼컴퓨팅 시스템

2012년도 MAHA 시스템은 이론 성능 100 TeraFLOPS 성능을 가지도록 설계되었다. 시스템은 다양한 응용 프로그램과 바이오 파이프라인 특성에 따른 최적의 매니코어 기술을 지원하기 위하여, 50 TeraFLOPS 규모의 Intel Xeon

Phi(이전 MIC)[16] 기반의 컴퓨팅 플랫폼을 설계하였다. Fig. 16은 Xeon Phi에 기반 한 50 TeraFLOPS MAHA 슈퍼컴퓨팅 시스템의 구현 개념도이다. MAHA 슈퍼컴퓨팅 시스템은 2013년 1월 이론 성능 104.5 TeraFLOPS 시스템으로 확장될 예정이다.

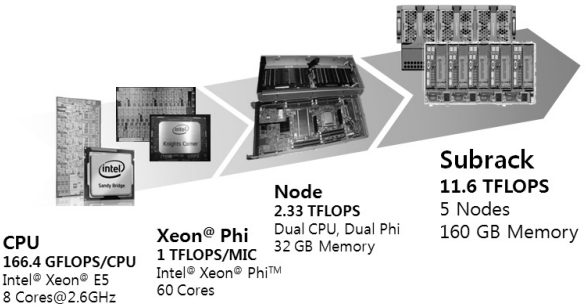

Fig. 16. Xeon Phi based MAHA supercomputing system

3) MAHA 슈퍼컴퓨팅 시스템 공조 구현

MAHA 슈퍼컴퓨팅 시스템은 전술한 바와 같이 일반 상용 서버에 비해, 상대적으로 많은 전력을 사용한다. 서버의 운용에 따른 효과적인 냉각 및 공조를 위하여 MAHA 슈퍼컴퓨팅 시스템을 위한 공조 시스템을 설계, 구현하였으며, 다음의 Fig. 17은 구현한 서버실 공조 시스템으로, 외기 입력을 통한 시스템 냉각 구조를 구현하였다.

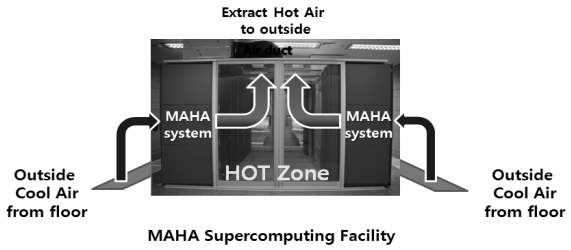

Fig. 17. HVAC implementation for MAHA supercomputing system

4.2 MAHA 슈퍼컴퓨팅 시스템 성능 측정 및 평가

2011년 개발한 50 TeraFLOPS MAHA 슈퍼컴퓨팅 시스템의 실제 성능을 평가하기 위하여, Hybrid high Performance Linpack(HPL) 벤치마크를 수행하여 성능을 평가하였다. 성능 평가는 Linux 2.6.32, Intel MKL, OpenMPI 1.4.3, SGE에 기반하여 수행하였다. 성능 측정 결과 최대 30.31 TeraFLOPS, 평균 29.9 TeraFLOPS의 성능, 시스템 효율 56.2%의 결과를 측정하였다. 다음의 Fig. 18과 Table 7은 성능 측정 결과 및 시스템 사양을 정리한 표이다.

5. 결 론

본 논문에서는 슈퍼컴퓨팅 관련 기술 확보를 위하여 2011년부터 개발 중인 MAHA 슈퍼컴퓨팅 시스템의 설계 및 구조에 대하여 서술하고, 시스템의 실측을 통한 성능 측정 결과를 제시하였다. MAHA 슈퍼컴퓨팅 시스템은 대규모의 유전체 분석을 고속으로 처리하는 것을 목표로 개발되고 있으며, 이를 위하여 MAHA 컴퓨팅 플랫폼, MAHA 파일 시스템, MAHA 시스템 소프트웨어, MAHA 바이오 응용의 4개 분야에 대한 기술을 개발하고 있다.

Table 7. Specifications and evaluation of 50 TeraFLOPS MAHA supercomputing system

	Index	Spec.	Units
Computing	Performance	53.1	TeraFLOPS (Rpeak)
	CPU Cores	512	core
	Accelerator Cores	32,768	core
	Memory	1	TeraBytes
Storage	SSD Capacity	14	TeraBytes
	HDD Capacity	70	TeraBytes
Network	Computational	40	Gbps
	Management	1/10	Gbps
Power	Total Max. Power	42.12	kW
Perfor-mance	Rmax	30.1	TeraFLOPS
	Efficiency (Rmax/Rpeak)	56.2	%
	Performance/Watt	719.61	MFLOPS/W

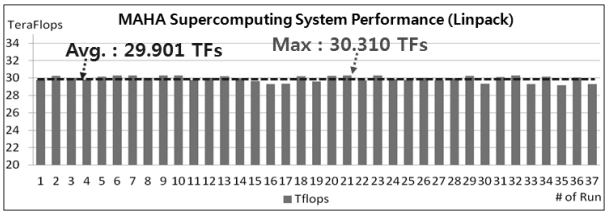

Fig. 18. Linpack results of MAHA supercomputing system

2011년 50 TeraFLOPS 규모의 MAHA 슈퍼컴퓨팅 시스템을 설계 및 구현하였으며, 구현된 시스템의 실측 결과 30.1 TeraFLOPS 성능과 56.2% 시스템 효율을 나타내었다. MAHA 슈퍼컴퓨팅 시스템은 2015년까지 총 이론 성능 300 TeraFLOPS 규모로 개발되고 있으며, 기술 확보를 위한 전용 네트워크 기술, 파일 시스템 기술, 시스템 소프트웨어 및 매니코어 바이오 응용 프로그램 기술을 개발하고 있다.

참 고 문 헌

[1] P. Kogge et al., ExaScale Computing Study: Technology Challenges in Achieving ExaScale Systems, DARPA Information Processing Techniques Office(IPTO) sponsored study, 2008.

[2] Kirk Skaugen, “Petascale to Exascale,” ISC 2010 Keynote Presentation, Intel, 2010.

[3] “Development of Supercomputing system for genome

analysis," IT industrial fusion core technology development project, MKE.
[4] YW Kim, SW Kim, "Technology and Trends of High Performance Processors," Electronics and Telecommunications Trends, Vol.25, No.5, pp.123-136, 2010.
[5] YW Kim, K Park, HY Kim, "Recent Trends on High Performance Computing System Technology," proceedings of the ITFE Summer Conference, pp.23-25, Aug., 2012.
[6] TOP500 Supercomputer sites, http://top500.org
[7] YW Kim, SW Kim, W Choi, "Summary on Worldwide HPC Developement Strategies and Status," Electronics and Telecommunications Trends, Vol.26, No.6, pp.174-188, 2011.
[8] 電子情報通信分野 科學技術·研究開發の國際比較,2011年版, 獨立行政法人科學技術振興機構研究開發戰略センター, June, 2012. http://crds.jst.go.jp/output/pdf/11ic03s.pdf
[9] Human Genome Project, Wikipedia, http://en.wikipedia.org/wiki/Human_Genome_Project
[10] Biology 2.0, Special report, The Economist, 2010, http://www.economist.com/node/16349358
[11] Intel Xeon Processor E5-1600/E5-2600/E5-460 Product Families Datasheet, Intel, 2012, http://www.intel.com/content/www/us/en/processors/xeon/xeon-e5-1600-2600-vol-1-datasheet.html
[12] NVIDIA whitepaper, "Tesla@ Kepler GPU Accelerators," NVIDIA, 2012, http://www.nvidia.com/content/tesla/pdf/Tesla-KSeries-Overview-LR.pdf
[13] Intel Intel® Xeon® Processor X5670, Intel, 2011, http://ark.intel.com/products/47920/Intel-Xeon-Processor-X5670-12M-Cache-2_93-GHz-6_40-GTs-Intel-QPI
[14] "Intelの8コア版Sandy Bridgeとモジュラー設計戰略," 後藤弘茂のWeekly海外ニュース, Impress Watch, 2011. 04, http://pc.watch.impress.co.jp/docs/column/kaigai/20110406_437481.html
[15] BY Jeong, et. al., "Data center operating cost savings for the eco-friendly air conditioning methods," Korea Patent pending, 2011.
[16] Intel® Xeon PhiTM Coprocessor, Intel, 2012, http://ark.intel.com/products/71992/Intel-Xeon-Phi-Coprocessor-5110P-8GB-1_053-GHz-60-core

김 영 우

e-mail : bartmann@etri.re.kr
1994년 고려대학교 전자공학과(학사)
1996년 고려대학교 전자공학과(석사)
2001년 고려대학교 전자공학과(박사)
2009년~2010년 과학기술연합대학원대학교 겸임교원(부교수)
2001년~현 재 한국전자통신연구원 책임연구원
관심분야: Asynchronous Circuit, Processor Architecture, High Speed System Interconnect

김 홍 연

e-mail : kimhy@etri.re.kr
1992년 인하대학교 통계학과(학사)
1995년 인하대학교 전자계산공학과(석사)
1999년 인하대학교 전자계산공학과(박사)
1999년~현 재 한국전자통신연구원 책임연구원
관심분야 : 스토리지 시스템, 파일 시스템, 데이터베이스

배 승 조

e-mail : sbae@etri.re.kr
1987년 연세대학교 전산과학과(학사)
1992년 Syracuse University 컴퓨터과학(석사)
1997년 Syracuse University 컴퓨터과학(박사)
1997년~현 재 한국전자통신연구원 책임연구원
관심분야 : High Performance Computing, Parallel Computing

김 학 영

e-mail : h0kim@etri.re.kr
1983년 경북대학교 전자공학과
1985년 경북대학교 전자공학과(석사)
2003년 충남대학교 컴퓨터공학과(박사)
2009년~현 재 UST 겸임교원(교수)
1988년~현 재 한국전자통신연구원 서버플랫폼연구팀 팀장
관심분야 : Cloud Computing, High Performance Computing

우 영 춘

e-mail : ycwoo@etri.re.kr
1986년 경북대학교 전자공학과(학사)
1988년 경북대학교 전자공학과(석사)
1988년~현 재 한국전자통신연구원 책임연구원
2011년~현 재 한국전자통신연구원 고성능컴퓨팅시스템연구팀 팀장
관심분야 : 분산처리시스템, 가상화, 고성능컴퓨팅시스템

박 수 준

e-mail : psj@etri.re.kr
1991년 University of Iowa, Biochemistry (학사)
1994년 Lehigh University, Computer Science(석사)
2011년 동국대학교 전자공학과(박사)
1994년~2012년 한국전자통신연구원 라이프테크놀로지연구팀장
2013년~현 재 한국전자통신연구원 바이오의료IT융합연구부장
관심분야 : 디지털영상처리, 유헬스, 바이오인포매틱스, 데이터마이닝

최 완

e-mail : wchoi@etri.re.kr

1981년 경북대학교 전자공학과(학사)

1985년 KAIST 전산학과(석사)

1985년~2003년 ETRI, TDX/CDMA 전전자교환기용 실시간 OS/DBMS/MW/컴파일러 개발책임자

2000년~2011년 한국정보처리기술사회 이사

2004년~2007년 ETRI, 클라우드서비스 기술 개발팀장

2007년 불교명예철학 박사

2008년~2010년 ETRI, SW콘텐츠미래기술연구부장

2011년~현 재 한국전자통신연구원 클라우드컴퓨팅연구부장

관심분야 : Cloud Computing, High Performance Computing