CPU 기술과 미래 반도체 산업 (II)

CPU Technology and Future Semiconductor Industry (II)

박상기 (Sahnggi Park, sahnggi@etri.re.kr) 소재부품원천연구본부 책임연구원

ABSTRACT

Knowledge of the technology, characteristics, and market trends of the latest CPUs used in smartphones, computers, and supercomputers and the research trends of leading US university experts gives an edge to policy-makers, business executives, large investors, etc. To this end, we describe three topics in detail at a level that can help educate the non-majors to the extent possible. Topic 1 comprises the design and manufacture of a CPU and the technology and trends of the smartphone SoC. Topic 2 comprises the technology and trends of the x86 CPU and supercomputer, and Topic 3 involves an optical network chip that has the potential to emerge as a major semiconductor chip. We also describe three techniques and experiments that can be used to implement the optical network chip.

KEYWORDS Photonic and optical interconnect, Optical network-on-chip, Optically interconnected CPU, Supercomputer architecture, CPU design and fabrication, Smartphone SoC, x86 CPU technology, Package on package, 2.5D package, 3D package

Ⅰ. 서론

컴퓨터 CPU는 인간이 도달할 수 있는 기술의 최고 수준을 나타내는 척도라 할 수 있다. 인간이 상상하는 어떠한 종류의 지능형 모델과 프로그램도 CPU 성능에 구애 받지 않을 수 없다. 4차 산업뿐만 아니라 전체 IT 산업의 발전을 이끌어 가는 정점에 CPU가 있다. 2019년 출시된 스마트폰과 컴퓨터, 슈퍼컴퓨터에 사용된 최신 CPU의 기술과 특성 및 시장동향, 그리고 미국 주요 기업과 대학교 전문가들의 연구방향, CPU 광인터커넥션 신기술을 3부에 걸쳐 기술 하였다.

Ⅰ부에 이어 Ⅱ부에 인텔 CPU의 내부 구조, register와 cache 간 그리고 L1 L2 L3 cache 서로 간 통신라인과 통신속도를 설명하고 CPU 내 통신시스템인 ring bus와 mesh network에 대해 설명하였다. 슈퍼

* DOI: https://doi.org/10.22648/ETRI.2020.J.350209

©2020 한국전자통신연구원

컴퓨터의 서버 CPU 구성과 3D torus 및 fat tree 광 네트워크 구조에 대해 설명하였다. 그리고 IT 기술 독립을 목표로 중국이 개발하고 있는 스마트폰 SoC, x86 CPU, 슈퍼컴퓨터 제작에 대한 기술 수준과 특징, 보유현황에 대해 기술하였다[1-8].

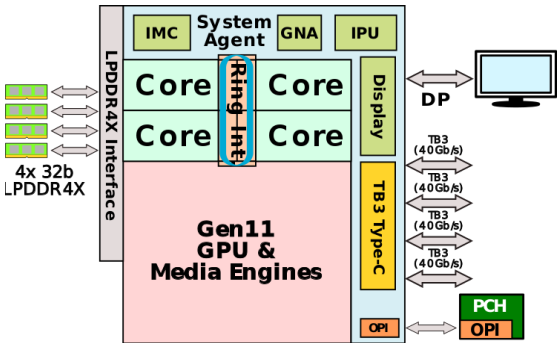
II. x86 CPU 기술과 동향

x86 CPU를 설계, 제작하는 기업은 현재 Intel, AMD, VIA의 3개가 있다. 과거에는 약 10여 개였으나 2000년대 들어 위 3개 기업에 인수 합병되거나 시장 경쟁력을 상실하고 생산을 중단하였다. VIA는 대만의 석유화학 재벌 왕영경(王永慶, 2008년 세계 178위 부자)의 딸 왕설홍(王雪紅, 1958년생)이 1987년에 설립한 기업으로 주로 motherboard에 들어가는 chipset를 주로 설계 제작하였으나 1990년대 후반부터 IDT, Cyrix 등 미국 processor 기업을 합병하여 x86 CPU 사업을 시작하였다. VIA는 2013년 중국 상해 시정부와 합작으로 벤처기업, Zaoxin을 설립하고 2014년, ZX-A를 시작으로 2019년 ZX-E, KX-6000을 생산하고 있다. KX-6000의 성능은 Intel이 2016년 생산한 7세대 Intel i5 정도이고, 중국 본토에 주로 판매되고 있는 것으로 알려져 있다. 중국이 미국으로부터 IT 기술 의존을 벗어나기 위해 국가 차원의 전략 사업으로 추진되고 있다. x86계열의 instruction set은 미국 기업 인수 합병을 통해 VIA가 개발하였고 생산은 TSMC에서 하는 것으로 관측된다. Statista 2019 통계에 의하면 2019년 1분기 기준 Intel과 AMD의 CPU 시장점유율 77% vs 23%이다. VIA CPU의 시장점유율은 찾아보기 어렵다. IBM도 x86 계열의 CPU를 설계 제작하였으나 결국 중단한 것처럼 중국 정부 지원 없이 시장원리에 의해서 생존하기 어려운 수준으로 추정된다.

표 1 2019년형 x86 CPU 특성

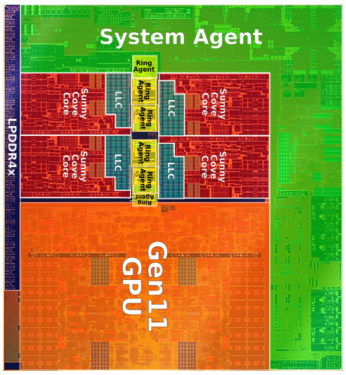
특성	Intel		AMD	
	i7(Laptop) 1065G7	Xeon Platinum 9282	Ryzen7 3780U	EPYC 7742
Core (Thread)	4(8)	56(112)	4(8)	64(128)
Clock (overclock)	1.3GHz (3.9GHz)	2.6GHz (3.8GHz)	2.3GHz (4.0GHz)	2.25GHz (3.4GHz)
L1/L2/L3	I32KB, D48KB/ 512KB/ 8MB	32KB/ 1MB/ 77MB	I64KB, 32KB/ 512KB/ 4MB	I64KB, D32KB/ 512KB/ 256MB
GPU	Intel Gen11	external	AMD Vega11	external
IPU/NPU	4th Gen/ GNAv1.0	external	external	external
TDP	15W	400W	15W	225W
Tech. node(Die mm^2)	10nm (122.52)	14nm(2 X 694 for 28 core)	12nm (Global Foun.)	TSMC 7nm
Launched	2019. 8.	2019. Q2.	2019. 10.	2019. 8.

Intel과 AMD 홈페이지에는 성능 별로 laptop, desktop, server용 CPU가 올려져 있고 그 중 desktop과 server용 CPU는 core 개수 외 큰 차이가 없으므로 laptop과 server용 CPU의 특성과 내부 통신 구조를 분석한다. 일반적인 특성은 표 1과 같다. Intel의 mobile CPU i7는 그림 1과 같이 하나의 칩 위에 다수의 소자가 집적된 SoC이다. 4개의 core와 GPU, system agent가 ring bus로 연결되어 있다. System agent에는 IPU, NPU(GNA), integrated memory



출처 https://en.wikichip.org/wiki/WikiChip

그림 1 Intel mobile CPU i7의 내부 구조도



출처 https://en.wikichip.org/wiki/WikiChip

그림 2 Intel mobile CPU i7의 실물 사진

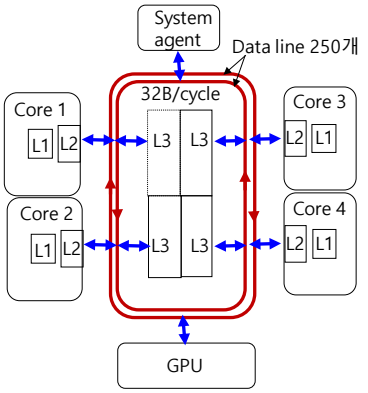


그림 3 ring bus 구조도

controller(IMC), display 및 thunder bolt 연결 포트가 있다. 그림 2는 실재 die 사진에 각 소자의 배치를 나타내기 위해 색을 덧씌운 모습이다. die 크기는 122.52mm^{2}이다.

그림 3은 ring bus를 설명하기 위한 구조도이다. Ring bus는 양방향으로 진행되는 두 개의 링형 구조이다. 각 링에는 256개 선으로 구성되는 data bus 외 snoop bus, request bus, acknowledgement bus의 세부 통신선이 있다. 양방향 총 통신선의 개수는 약 1,000개인 것으로 알려져 있다. Ring bus는 칩 제조 공정에서 13개의 metal layer 중 상부 층에 배치되어 있다. Core와 ring bus는 동일한 clock 주파수를 사용하고 data bus는 한 cycle당 32B(256bit)를 전송하므로 3.0GHz CPU의 경우 한 링당 96GB/s의 전송속도를 갖는다. Cache는 한 주소 block당 64byte씩 저장되고 data를 전송할 때도 block 단위로 전송하므로 최소 전송 단위는 64byte의 data가 2개 cycle에 걸쳐 전송된다. Core가 4개인 경우 링 위에는 6개의 접속 지점이 있고, 한 cycle당 1개의 접속 지점을 이동한다. Core 1이 core 4의 L3에 data를 전송할 경우 3개의 전송 구간을 건너뛰어야 하므로 3 cycles의 전송 latency(지연시간)를 갖는다. Ring bus는 core 수가 많아질수록 latency가 계속 증가할 뿐만 아니라 전송 위치에도 의존하는 단점을 지니게 된다. 2015년 이후 출시된 모든 Intel의 mobile용 CPU는 어느 한 core의 L2에서 L3, 또는 L3에서 L2, 또는 GPU에서 CPU로 data가 전송될 때 항상 ring bus 내 256개선을 통해 32B/cycle 속도로 전송된다. L3는 총 8MB의 저장용량이 있고 core당 2MB씩 할당된다. Instruction cache L1은 32KB, data cache L1은 48KB, L2 cache는 512KB의 저장용량을 갖고 있으며 해당 core에 소속되어 있다.

각 core의 내부 구조는 그림 4와 같다. 인텔은 mobile용에서 server용까지 동일한 core architecture를 사용하기 때문에 core의 내부 구조는 CPU 성능과 관계없이 같은 시기의 제품군 내에서 거의 동일하다. 프로그램이 실행되면 하드디스크에서 main memory(DRAM)로 프로그램이 전달되고 프로그램에 적혀 있는 순서에 따라 instructions이 L2 또는 L3를 거쳐 L1 instruction cache로 넘어오게 된다. L1에는 수백 개의 instruction이 저장되어 대기상태에 있게 된다. L1에서 predecoder로 전달되는 통신라인은 128개 선으로 구성되어 있고 16B(128bit)/cycle 속도로 전달된다. x86 instruction은 길이가 1byte부터 16byte까지 다양하므로 최소 1~16개 instruction이 1cycle에 전달된다. Predecoder

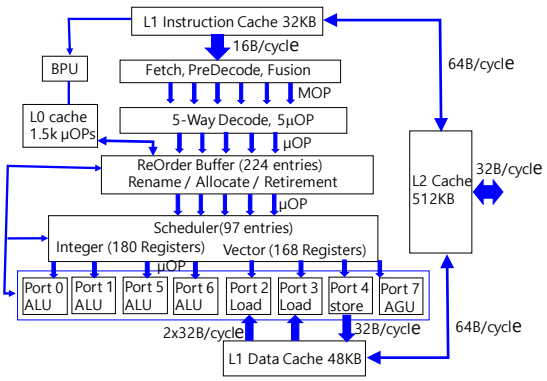


그림 4 Core 내부 구조도

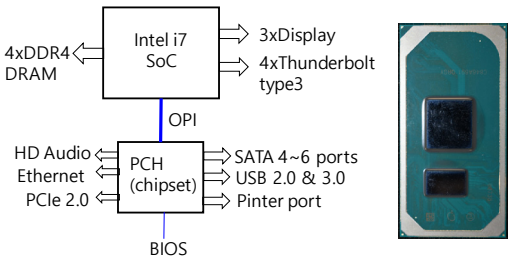
는 instruction을 불러오기(fetch), 다수의 instruction 간 경계를 확인하기(predecode), 두 개를 합쳐서 한 번에 연속으로 실행 가능한 instruction들을 결합하기(fusion) 작업을 한다.

작업이 완료된 16B 신호는 몇 개의 macro-operation(MOP)으로 나누어져 decoder로 전달된다. 이후 진행되는 단계별 작업은 여러 갈래로 나누어져 전달되고 전선수와 통신속도는 확인하기 어렵다. 다만 각 단계별 전달은 1cycle 내에 이루어진다. Decoder에서는 MOP를 읽고 해석(translate)한 후 5개의 micro-operation(μOP)로 변환하여 다음 단계로 전달한다.

모든 μOP은 길이가 일정하고 규칙적인 신호체계를 갖는다. ARM instruction set는 길이가 4byte로 일정한 데 반해 x86은 길이가 다양하므로 decode 단계에 이르기까지 그림 4에 생략된 많은 복잡한 과정을 거쳐야 한다. 2개의 instruction이 결합된 fused-μOP도 많이 포함되어 있다. Broadwell 제품에서는 4개였으나 Skylake 제품부터는 5개의 μOP으로 구성되어 ReOrder Buffer로 전달된다. 그리고 decode 단계까지는 2개의 threads가 cycle별로 번갈아 가며 독립적으로 진행된다. 즉, predecoder와 decoder는 2개씩 존재하고 한 cycle에 thread 1의 instruction이 첫 번째 predecoder와 decoder로 진행되면 다음 cycle에는 thread 2의 instruction이 두 번째 predecoder와 decoder로 진행된다. ReOrder buffer부터 두 threads는 단계별 하드웨어를 공통으로 사용한다. Decode 작업이 완료된 일부 μOp은 L0 cache에 저장하였다가 branch predictor unit(BPU)의 판단에 따라 정해진 순서에 ReOrder buffer에 합류하게 된다. L0 cache와 ReOrder buffer 간 통신라인은 64B/cycle로 높은 우선순위를 갖는다.

L1 cache에서 decode 단계까지는 program에 적혀 있는 순서대로 각 instruction이 들어가지만 ReOrder buffer에서 224개 μOp을 모으고 우선순위를 다시 정해(out-of-order) 진행한다. 5개 μOP가 ReOrder Buffer에 들어오면 entry에 저장(rename)하고 진행 순서와 방향을 정하며(allocate) 수행이 완료된 μOp은 entry에서 제거(retirement)하게 된다. Retirement가 발생할 때만 새로운 μOp이 들어올 수 있다.

새롭게 정해진 순서에 따라 5개씩 μOP이 scheduler에 전달되면 scheduler는 97개까지 저장하고 시간 배분과 순서에 따라 각 resister를 통해 arithmetic logic unit(ALU)에 보내어 계산을 수행한다. Data가 필요할 경우 L1 data cache에서 data를 불러온다. 계산이 완료되면 결과 data를 재사용 여부에 따라 scheduler의 register에 저장해 두거나 L1 data cache에 저장한다. 수행이 완료된 μOP는 재사용이 필요할 경우 ReOrder Buffer에 남겨두고 아닐 경우 retirement하게 된다. L1 data cache에서 logic part로 data를 불러오기 위한 통신선은 총 512개이고 통신속도는 64B/cycle이다(2017년형 서버용 CPU는 1,028개, 128GB/cycle 임). Data를 저장하기 위한 통신선은 256개, 통신속도는 32B/cycle이다(2017년형 서버용 CPU는 512개, 64GB/cycle 임). 그리고 L2 cache에서 각 L1 cache로 가는 통신선은 512개, 통신속도는 64B/cycle이다. L1 instruction cache에서 출



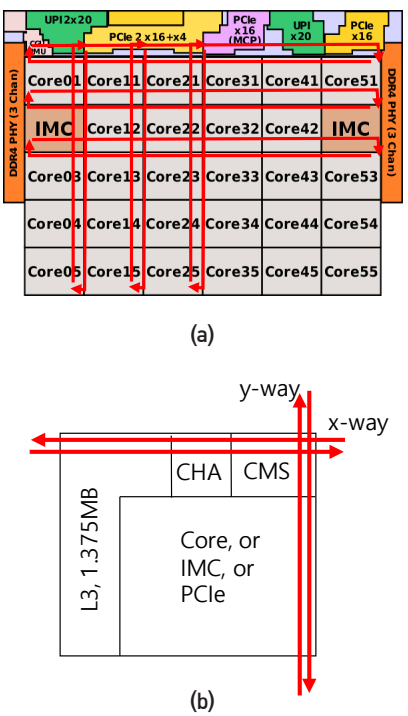
출처 https://en.wikichip.org/wiki/WikiChip

그림 5 Intel i7 SoC 인터페이스 (a) 구조도, (b) 패키지 실물

발하여 ALU의 계산 결과를 L1 data cache에 저장할 때까지 실행 과정은 pipeline 형태로 진행된다. 그림 4에 나타낸 단계는 pipeline stage를 나타내며 전체 stage 수는 가장 빠른 경우 14개, 가장 느린 경우 19개까지 있다. 그림 4는 다수의 stages를 각 단계에 축약하여 나타내었다. CPU의 연산 속도를 높이기 위해서 통상 두 가지 방법을 동원한다. 첫 번째는 그림 4에 표시된 병렬 라인 수를 증가시켜 IPC를 증가시킨다. 두 번째는 CPU의 clock 주파수를 높인다. 14~19개의 pipeline stages에서 가장 느린 stage가 CPU clock 주파수의 최고치를 결정하게 되므로 모든 단계의 실행 속도가 균일하게 높아져야 한다. 이를 위해서는 단계별 통신라인의 속도와 transistor의 스위칭 속도가 높아져야 한다. Transistor의 스위칭 속도는 gate에 전하가 충전되는 속도에 의존하므로 gate 면적에 반비례한다. Technology node가 작아질수록 스위칭 속도가 높아진다. 2019년 현재 인텔은 자체 fab에서 최고 10nm 공정으로 최신 제품을 제작하고 AMD는 TSMC 7nm 공정으로 최신 제품을 제작하고 있다.

CPU외 display와 DRAM interface를 포함하는 north bride와 기타 주변 장치를 연결하는 south bridge가 있었으나 north bridge는 CPU 칩 안에 집적되고 south bridge(PCH, platform control hub)만 남겨진 상태이다. Mobile 용 제품의 경우 그림 5와 같이 CPU와 PCH를 동일한 기판에 올려 package 공정을 실시한다. 그리고 CPU와 PCH 사이에는 기판 상에 있는 연결 라인(OPI: On-Package Interconnect)을 통해 신호가 전송된다. OPI는 PCIe3.0의 4개 lanes과 동일한 전송속도를 갖는다. 각 lane은 8Gb/s 전송속도를 가지고 있다. CPU에서 PCH로 4개 lanes, 반대 방향으로 4개 lanes씩, 각 방향 32Gb/s의 최대 전송속도이다. PCH에는 인터넷, 오디오, SATA, USB, 프린터 등이 연결된다. 그림 5(b)에 보는 바와 같이 크기는 53.76mm^{2}이다.

Intel의 최신 서버용 CPU, Xeon Platinum 9282의 배치도는 그림 6과 같다. Core 28개, DDR4 2개(각 3채널), IMC 2개, PCIe3.0 4개(각 16lanes), UPI 3개(각 20lanes), PCIe3.0 1개(4lanes, OPI에 사용)가 한 die에 그림과 같이 배치되어 있고 2개의 die를 하



출처 https://en.wikichip.org/wiki/WikiChip

그림 6 Intel Server용 CPU, Xeon Platinum (a) 배치도, (b) 코어 상세도

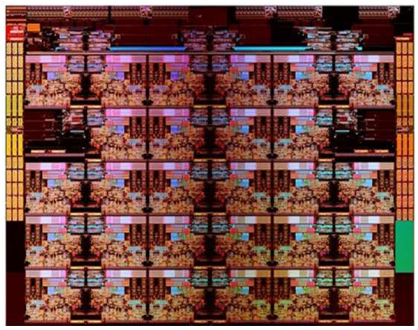
나의 패키지에 포함시켜 총 56개 core의 CPU이다. Intel은 core의 수가 10개 이상인 desktop과 server용 제품에 대해 ring bus 대신 그림 6(a)와 같은 mesh interconnect 구조를 사용하고 있다. 각 소자의 경계를 따라 x, y 방향으로 bi-directional half ring이 배치되어 있다. Mesh line은 13개 metal layer 상층부에 있고 각 소자는 그 아래 layer에 있어 서로 겹치지 않으므로 실재로 mesh line은 각 소자의 위를 지나가고 있다. 그림 6(b)에 나타낸 바와 같이 모든 소자의 접속지점에는 신호 접속과 routing을 관리하는 CMS(converged mesh stop)가 있다. 각 core는 1.375MB씩 할당된 L3 cache와 CHA(caching/home agent)를 가지고 있다. L3의 data를 자체 코어에서 요청할 경우 CHA로 보내어 core에 전달되고 다른 소자에서 요청할 경우 CHA를 경유하여 CMS로 전달된다. Cache coherency control과 routing 기능에 의해 CMS는 정해진 시간 슬롯 안에 도착 지점의 주소로 가장 빠른 경로를 선택하여 data를 보낸다. 예를 들어, core05가 core51에 data를 보낼 경우 y방향으로 4지점을 이동한 후 경로를 스위칭하여 x방향으로 5지점을 이동한다. Core를 제외하고 mesh에 연결된 접속 지점은 총 7개이고 이들의 경우 L3와 CHA가 없고 신호는 곧바로 CMS에 보내어 mesh에 올려진다. Intel은 2017년에 mesh 구조를 제품에 도입하였고, 통신라인 폭(전선의 수)과 통신속도 등 자세한 내용은 아직 공개하지 않은 것으로 알려져 있다. 그림 7은 die 사진을 나타내며 크기는 694mm^{2}이다.

AMD와 ARM사의 2019년형 CPU의 core 내 통신속도를 Intel CPU와 비교하여 표 2에 나타내었다. Instruction cache IL1에서 predecoder 간 통신속도는 Intel과 ARM core가 16GB/s, 128 lines인데 비해 AMD core는 그 두 배이다. Scheduler register에서 data cache(DL1)으로 data를 저장하는 통신은 인텔의 서버용 CPU가 64GB/s, 512lines인데 반해 AMD와 ARM core는 그 반이다. DL1에서 ALU로 Data를 불러오는 통신은 Intel의 경우 128GB/s, 1,024lines인데 비해 AMD와 ARM core는 그것의 ¼이다. 그 외 다른 통신라인의 속도와 통신 구조는 표 2에 나타낸 바와 같다. Intel에 비해 AMD와 ARM core의 내부 구조(Microarchitecture)와 통신라인(Interconnect architecture)은 자세히 공개되어 있지 않다.

Intel 서버용 CPU는 하나의 die에 core가 10개(LCC: Low Count Cores), 18개(HCC: High Count Cores), 28개(XCC: eXtreme Count Cores)의 세 종류가 있다. 세 종류의 CPU die 중 어느 하나를 이용하여 2개 die를 하나의 패키지 기판에 올리고 소켓 1개를 부착한다. 그림 8은 4개 die를 두 개씩 묶어

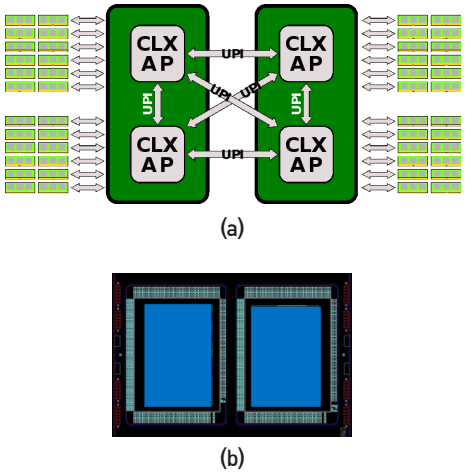
표 2 2019년형 CPU core 내부 통신

2019년형 CPU Core 내 통신	Intel(Server)	AMD(Server)	ARM
IL1 to Predecode	16GB/s	32GB/s	16GB/s
DL1 to ALU or Register	Store 64GB/s Read 128GB/s	32GB/s 32GB/s	32GB/s 32GB/s
IL1 to L2 DL1 to L2	64GB/s 64GB/s	32GB/s 32GB/s	32GB/s 64GB/s
L2 to L3	32GB/s	32GB/s	64GB/s
Interconnect	Ling Bus Or Mesh	Infinity Fabric	DynamIQ
Number of lines	16GB/s: 128, 32GB/s: 256, 64GB/s: 512		



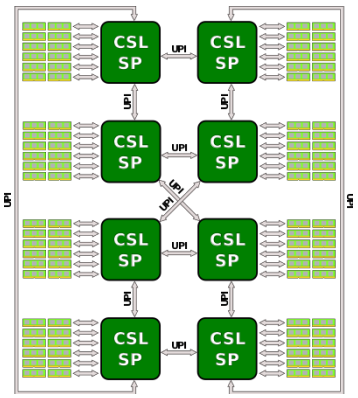
출처 https://en.wikichip.org/wiki/WikiChip

그림 7 Intel Server 용 CPU, Xeon Platinum 실물 사진



출처 https://en.wikichip.org/wiki/WikiChip

그림 8 서버용 CPU의 2 소켓 연결 (a) 구도, (b) 사진



출처 https://en.wikichip.org/wiki/WikiChip

그림 9 서버용 CPU 8개 die 연결 구도

socket 2개를 만들고 4개 die를 UPI links로 연결하는 모습을 나타낸다. 그림 6에서 하나의 die에는 UPI links 3개(각 20 lanes)와 DDR4 6채널(3채널 2개)이 있다. 그림 8에 총 4개 die를 연결하기 위해 각 die의 3개 UPI links를 이용하고 있다. 그리고 DDR4 6채널이 mother board를 통해 DRAM과 된다. 그림 8(b)는 실제 패키지 된 사진을 나타낸다. 동일한 방법으로 총 8개 die를 연결하는 방법을 그림 9에 나타내었다. Server computer 또는 Supercomputer에는 최대 28×8=224개의 Intel core가 하나의 board에 올려 질 수 있음을 나타낸다. 이와 같은 mesh interconnect와 UPI link는 현재 전기통신으로 연결되어 있지만 광통신으로 연결하는 것이 가장 효과적이고 지난 15여 년간 관련 분야의 목표이다.

UPI(ultra path interconnect)는 Intel이 이전에 개발한 QPI(quick path interconnect)를 업그레이드하여 2017년부터 칩에 적용한 일대일(point-to-point) 통신 구조이다. QPI의 통신 특성은 다음과 같다. 하나의 QPI link에는 42개의 lanes가 있고, 각 lane은 2개의 통신선으로 구성되어 있다. 두 선에 같은 신호를 보내고 신호의 차이(differential signal)를 구해 noise를 제거하는 방식을 취한다. 따라서 하나의 QPI link에는 총 84개의 신호선이 있다. 42개의 lanes 중 21개는 송신에, 21개는 수신 사용된다. 단방향 21개 중 1개 lane은 clock 신호에 사용되고 4개 lanes는 control signal에 사용되며 실제 data 전송에는 16개 lanes이 사용된다. 그림 6의 UPI는 단방향 20개 lanes를 가지고 있지만 특성은 QPI와 거의 동일한 것으로 추정된다. Clock 신호가 0에서 1로 바뀔 때와 1에서 0으로 바뀔 때 각 1bit를 보내므로 1cycle당 2bit(double data rate)을 전송한다. 한 lane의 최대 전송속도는 10.4GT/s(Giga transfer per second)이다. 한 cycle당 2회 전송(transfer)이므로 최대 link 주파수가 5.2GHz까지 가능함을 의미한다. CPU 주파수가 3.2GHz이고 link가 CPU 주파수에 동기화 되어 있을 경우 lane당 전송속도는 6.4Gb/s이다. 16개 lanes의 단방향 전송속도는 102.4Gb/s이고 동시에 송수신이 가능하다. CPU die에는 4개의 PCIe 3.0이 있고 각 PCIe에는 단방향 16개 lanes(양방향 32개 lanes)가 있다. 전송 방식은 QPI와 동일하고 lane당 최대 전송속도는 8GT/s 이다. 즉, 최대 link 주파수가 4.0GHz까지 임을 의미하며 link 주파수를 CPU 3.2GHz에 동기화시킬 경우 QPI와 전송속

도가 같다. 2019년 현재 상용화 되어 판매되고 있는 VCSEL의 직접 변조 속도는 50GHz이고 differential signal을 사용할 경우 120Gb/s가 전송 가능하므로 QPI에 비해 약 12배 빠른 속도이다. 안정성이 더 높은 25GHz VCSEL을 사용할 경우에도 6배 빠른 전송이 가능하다. 열발생률은 전기 신호가 통상 Gb/s당 2mW인데 비해 광 신호는 0.08mW 정도이므로 약 20배 이상 차이가 있다.

Intel CPU i7에 사용된 socket(socket model: LGA2011-3)의 구조도를 그림 10에 나타내었다. Socket 밑면에는 2,011개의 solder ball(납땜을 위한 작은 볼)이 있고 같은 수의 접촉 배열이 있는 PCB와 전기적으로 접촉을 한다. Socket 윗면에는 밑면과 연결된 2,011개 접촉 배열이 있다. Socket은 PCB에 SMT(surface mount technology) 공정을 통해 접착된다. Solder ball은 은(Ag) 3%, 구리(Cu) 0.5%, 주석(Sn) 96.5%로 이루어진 물질이고 용융점은 217℃이다. 열처리(thermal reflow)를 통해 용융점보다 높은 온도를 가해 접착된다. 통상 패키지 공정은 CPU die를 패키지 기판(processor substrate)에 전기적으로 접속하고 보호막을 입히는 과정을 가리킨다. Intel의 경우 CPU die 표면을 뒤집어 기판 표면과 맞댄 상태에서 solder pad가 있는 기판 전극과 CPU전극을 열처리 공정으로 접착한다(flip-chip bond). 기판에는 필요한 전기회로가 있고 밑면에는 socket 접촉점과 대응되는 지점에 전극 pad가 있다. CPU die는 뚜껑과 열전도 물질(TIM: Thermal Interface Material)로 접촉되어 있고 뚜껑은 보호 케이스 역할 외에 열 확산(IHS: Integrated Heat Spreader) 기능을 하도록 설계되어 있다. 뚜껑의 위 면은 heat sink 블록과 TIM으로 접촉되어 있고 그 위에는 통상 환풍팬(fan)이 있다. 기판의 아래 전극과 socket의 윗면 전극은 기계적인 힘을 가해 접촉을 하고 기계적인 힘은 일종의 꺽쇠인 ILM(Independent

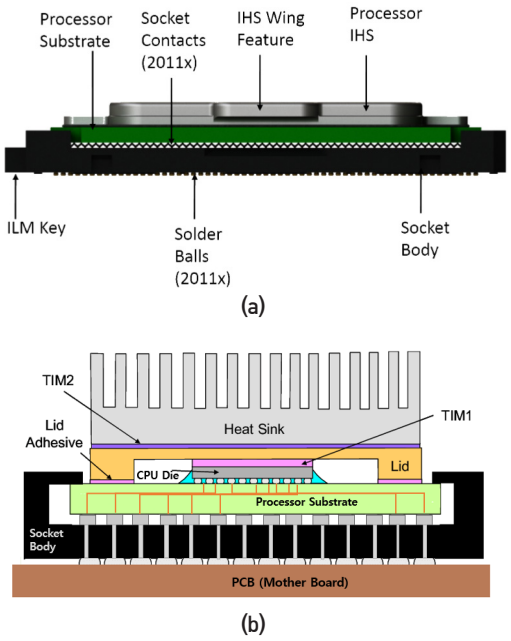


그림 10 Intel CPU 패키지 구조 (a) 외부 구조도, (b) 단면도

Loading Mechanism) 장치로 가해진다. ILM은 힘을 가하는 것 외 기판이 socket의 정확한 위치에 오도록 정렬하는 기능을 가지고 있다. 패키지 기판은 플라스틱 재질(fiber reinforced resin)이고 뚜껑은 니켈이 도금된 구리(nickel plated copper)이며, TIM을 포함하여 각각 면적과 열 적 특성에 대한 규격을 정해 두고 있다.

CPU 성능을 떨어뜨리거나 fail이 발생하는 가장 큰 이유 중 하나가 CPU의 온도이다. 온도를 적정하게 유지하기 위해 몇 가지 기술을 사용하고 있다. CPU die에는 DTS(digital thermal sensor)가 최소 4개 이상 집적되어 있다. 그리고 CPU의 케이스 뚜껑(IHS) 중앙에 thermo-couple(금속 물질의 온도 센서)를 연결하여 온도(T_{case})를 모니터링한다. CPU가 사용하는 전력(power)과 DTS온도(T_{DTS})및 케이스 온도는 Intel i7 제품에 대해 실험적으로 다음과 같다. T_{case}=0.17×P+43.3, T_{DTS}=0.398×P+43.3.

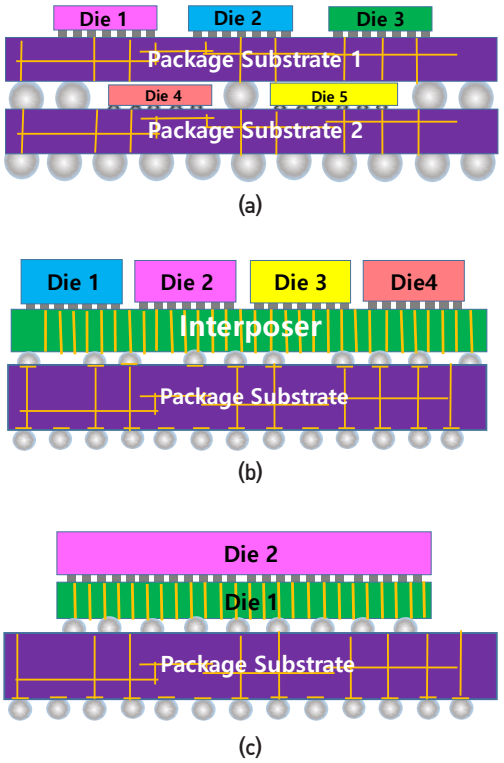


그림 11 SoC 패키지 구조 (a) POP, (b) 2.5D, (c) 3D

CPU의 TDP(Thermal Design Power) 값은 CPU 설계 당시에 설정한 전력 소모량으로 CPU의 모든 core가 설계 시 설정한 clock 주파수로 실행할 때 사용하는 전력량이다. Turbo boost를 사용하지 않는 일반적인 상태에서 CPU를 최고 성능으로 사용할 때 소모하는 전력에 해당한다. Intel i7-5960X Extreme Edition의 TDP는 140W이다. 이 경우 die 온도는 T_{DTS}=99.02℃이고 케이스 뚜껑은 T_{case}=67.1℃이다. 즉, die 온도가 약 100℃ 일 때 케이스 뚜껑은 67.1℃이고 안정적인 수명과 계산 오류가 발생하지 않기 위해 CPU는 항상 이 온도 이하로 유지되어야 한다. Intel CPU에는 TCC(thermal control circuit)와 프로그램 실행을 위한 register가 내장되어 있다. 온도가 설정된 값보다 높아질 경우 fan을 최대한 회전시키는 것 외에 두 가지 방법을 이용하여 강제적으로 온도를 내리는 프로그램이 실행된다. 첫 번째 방법은 CPU의 clock 주파수를 떨어뜨려 실행 속도를 낮추고 두 번째 방법은 CPU clock을 약 32ms 단위로 On 상태와 Off 상태를 반복한다. 이 경우에도 실패할 경우 CPU를 강제로 정지시킨다 [1-5].

최근 컴퓨터 CPU에 비해 스마트폰 SoC 패키지 기술이 크게 발전하고 있고 투자도 많이 이루어지고 있다. 대표적으로 PoP(Package on Package) 기술과 2.5D 및 3D 패키지 기술을 들 수 있다(그림 11 참조). PoP는 현재 대부분의 스마트폰 SoC 제조사들이 사용하고 있는 기술로써 그림 11(a)와 같다. SoC die와 메모리(SDRAM) die를 각각 패키지 기판에 전기적으로 접속하고 보호막 처리를 한 후 두 패키지 기판을 상하로 적층하여 BGA(Ball Grid Array) 방법으로 전극을 연결한다. 통상 SoC는 패키지 기판에 flip-chip bond로 접속하고 메모리 die는 패키지 기판과 wire-bonding 방법으로 접속한다. 2.5D 패키지 기술은 패키지 기판과 die 중간에 실리콘 삽입(interposer) 평판이 있고 SoC die와 메모리 die를 평판 위에 flip-chip bond로 접속한다. 그리고 BGA 방법으로 interposer 평판을 패키지 기판과 접속한다. 플라스틱 기판과 달리 실리콘 평판은 CMOS 공정을 사용하므로 훨씬 미세한 전선을 적층으로 형성할 수 있다. 아랫면 전극과는 TSV(Through Silicon Via) 공정을 사용하여 연결하며 이것은 3D 패키지와 동일하다. 3D패키지는 die1과 die2를 flip-chip bond 공정으로 접속하고 die1에 TSV를 형성하여 die1의 아랫면 전극과 연결한다. 그리고 die1는 BGA 방법으로 패키지 기판과 접속한다. 패키지 밀도가 높고 연결선이 짧아 통신속도에 유리하지만 단점은 서로 다른 제조사 간 metal line 설계가 공유되어야 한다. TSV 공정은 단순히 hole을 형성하는 것 외에 웨이퍼를 얇

게(대략 650 → 100μm) 만드는 공정, 얇은 웨이퍼를 CMOS 공정 라인에 적용하는 것 등 제작 원가를 높이는 공정이 수반되므로 2.5D와 3D 패키지는 아직 보편적으로 사용되지 않고 있다.

III. 슈퍼컴퓨터 기술과 동향

2019년 11월 기준 최고 성능 순위 슈퍼컴퓨터는 1, 2위에 IBM이 제작한 Summit(148.6PFlops)와 Sierra(94.64PFlops)가 올라가 있고 3위는 중국의 Sunway TaihuLight(93.01PFlops, 神威·太湖之光)가 올라가 있다. 전체 500위 중 62%인 310기를 중국 기업인 Lenovo, Inspur, Sugon의 3사가 제작하고 18.8%인 94기를 미국 기업인 HP, Cray, Dell, IBM사가 제작하였다. 중국은 IT기술 독립을 위한 대규모 예산($200B, 약 200조 원) 투입과 2001년부터 시작된 5년 단위의 단계별 기술 개발 전략에 따라 슈퍼컴퓨터뿐만 아니라 독자적인 CPU 설계 및 제작에도 상당한 기술을 보유한 상태이다. 특히 Sunway TaihuLight에 탑재된 processor는 중국 연구진이 약 15년간 기술 축적을 통해 개발한 제품(Sunway SW26010 1.45GHz)을 사용하였다. 이에 대한 자세한 성능과 구조는 다음 절에 논의한다. IBM은 세계최대 슈퍼컴퓨터 공급 기업이었으나 서버 제작 사업 중 Intel processor 사용 사업부를 2014년 중국 회사인 Lenovo에 매각하고 자체 브랜드인 IBM Power9계열의 processor만 사용함에 따라 전체 비중은 크게 낮아진 상태이다. 500위 내 슈퍼컴퓨터 중 Intel processor(Xeon E5, Xeon Gold 등)의 비중은 약 94.8%로 절대 다수가 Intel 제품을 사용하고 있다. 한국의 processor 설계 기술은 ARM의 license를 통해 제작하는 삼성의 Exynos가 전부이고 이 또한 미국 AMD와 Intel의 전직 엔지니어가 주도하였고 2019년 삼성은 Exynos 사업 중단을 발표한 바 있다. 나라별 500위 내 슈퍼컴퓨터는 중국이 228기, 미국 117기, 일본 29기, 프랑스 18기, 독일 16기 등을 보유하고 있다. 한국은 3기의 슈퍼컴퓨터를 보유하고 있고 Cray사가 3기를 공급하였다.

슈퍼컴퓨터에는 약 10~100만 개의 컴퓨터 서버가 광통신 네트워크로 연결되어 있고 서버에 있는 CPU는 통상 10~100개의 core로 구성되어 하나의 슈퍼컴퓨터는 대략 100~1,000만 개의 core를 가지고 있다. 각 서버 CPU는 자체적으로 cache coherence가 가능하도록 설계되어 있고 각 서버 메모리(DRAM)는 다른 모든 서버가 접근 가능하며 서로 data를 공유할 수 있도록(NUMA: Non-Uniform Memory Access) 네트워크가 설계되어 있다. 슈퍼컴퓨터의 내부 통신 구조는 대개 공개하고 있지 않으나 일부 공개된 자료를 근거로 유추할 수 있는 두 슈퍼컴퓨터의 내부 구조를 분석한다.

슈퍼컴퓨터에 사용되는 통신 구조(network topology)는 fat tree network 구조와 3D torus network(mesh network의 일종) 구조를 거의 대부분 채택하고 있다. 더 많은 서버를 연결하는 데 유리한 fat tree 구조는 상위 순위 대부분이 채택하고 있다. 임의의 두 서버 간 통신 경로 길이가 동일해서 다른 서버의 data를 불러오는 데 걸리는 시간(latency)이 균일한 특징이 있는 반면, 나중에 슈퍼컴퓨터를 업그레이드하기 위해 서버를 추가하는(flexibility) 데 어려움이 있고, 상위 계층의 스위칭 장비가 고장 날 경우 전체에 미치는 영향이 큰 특징이 있다. 3D torus 구조는 fat tree 구조와 대비되는 장점 외에 인접한 서버 간 data 교환(locality)이 중요한 프로그램에 유리하고 통신 구조가 효율적이어서 작은 수의 core에도 더 높은 계산 속도를 나타낼 수 있다.

NSF(National Science Foundation)에서 $20M(약 2000억 원)의 Grant를 받아 San Diego Supercomputer Center(SDSC)에 설치한 슈퍼컴퓨터 “Gordon”

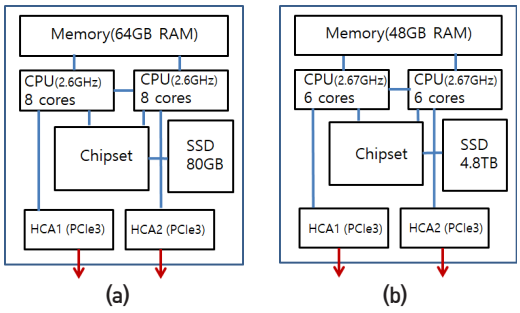


그림 12 Gordon의 서버 (a) Compute node, (b) I/O node

표 3 서버 내부 통신 폭과 속도

종류	Compute node	I/O node
CPU	Intel Xeon ES-2670, 2.6GHz	Intel Xeon X5650, 2.67GHz
DRAM	DDR3 1600, 8채널, 64GB	DDR3 1333, 6채널, 48GB
Hard disk	SSD 80GB	SSD 4.8TB
CPU to CPU	QPIx2, 20.8GB/s	QPIx2, 21.36GB/s
CPU to DRAM	32lines/ch, 8ch 102.4GB/s	32lines/ch, 6ch 64GB/s
CPU to HCA	PCIe3 16ch 10.4GB/s	PCIe3 16ch 10.68GB/s
CPU to chipset	OPI 4ch, 2.6GB/s	OPI 4ch, 2.67GB/s

은 2011년말부터 서비스를 시작하였고 당시 48위(285.8TFlops)에 올랐다. Cray사가 이전 모델(CS300-AC)을 업그레이드하여 제작하였고 3D torus network 구조를 채택하고 있다. Gordon에는 1024개의 compute node(프로그램 연산을 주목적으로 하는 서버)와 64개의 I/O node(data 입·출력을 주목적으로 하는 서버)가 있고 각 node는 64개의 스위칭 장비(Infiniband network switch)에 연결되어 있다.

그림 12는 compute node와 I/O node의 내부 구조를 나타내고 있다. Compute node는 하나의 mother board 위에 그림과 같이 8개 core를 가진 CPU socket 2개가 QPI 2채널로 연결되어 있다. 각각의 CPU는 4채널로 64GB DRAM과 연결되어 있고 OPI에 의해 chipset에 연결되어 있다. 80GB SSD(hard disk memory)는 SATA를 통해 chipset에 연결되어 있다. HCA(host channel adapter)는 mother board 슬롯에 끼워 넣는 카드 형태이고 PCIe3을 통해 CPU에 연결되어 있으며 2개의 광통신 접속 단자가 있다. I/O node 역시 비슷한 구조이고 세부적 특성은 그림에 표시된 바와 같다. 각 소자 간 통신라인 수와 통신속도 그리고 세부 특성은 표 3과 같다.

Network 스위치는 그림 13(b)와 같은 36 ports의 Mellanox 제품을 사용하였고, 36개의 광통신 단자 간 자유롭게 스위칭 가능하며 90ns switching latency를 가지고 있다. 내장되어 있는 dual core x86 CPU와 스위치 칩에 의해 작동된다. 광능동 케이블(AOC: Active Optical Cable)은 현재 200Gb/s가 가장 높은 통신속도를 가진 제품이다(표 4 참조). 56Gb/s의 변조속도를 가진 VCSEL 4개를 4개의 광섬유(multi-mode fiber)에 접속하여 신호를 보낸다. 양방향(duplex)이므로 실제로 8개의 광섬유와 8개의 VCSEL, 8개의 PD로 구성되어 있고 각 방향으로 최대 200Gb/s 속도가 가능하다. 이와 같은 광케이블은 표준 규격에 따라 패키지 크기와 pin 구조를 갖추고 있다. 표준 규격은 SFP(small form-factor pluggable, 1개 lane) 또는 QSFP(quad small form-factor pluggable, 4개 lanes)로 명명되어 있다. Gordon 시스템은 제작 당시 최신 제품인 QDR 40Gb/s QSFP를 사용하였다. 광연결은 QSFP를 HCA card와 Infiniband 스위치 port에 꽂으면 광섬유 반대편에 있는 전기 신호 핀에 의해 연결된다. 2019년 11월 기준 500위 슈퍼컴퓨터 중 140기는 Infiniband 스위치를 사용하고, 259기는 Gigabit Ethernet 스위치를 사용하고 있다. 성능별 비율은 Infiniband가 40.3%, Gigabit Ethernet이 26.1%를 차지하여 높은 성능일수록 Infiniband가 유리함을 알

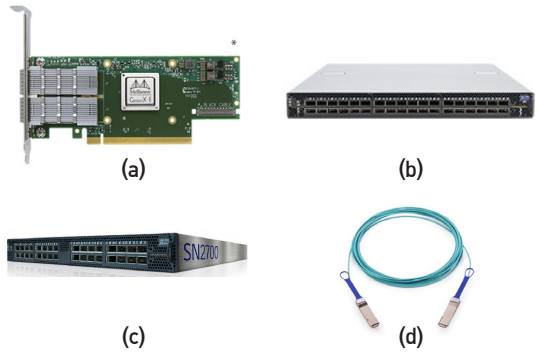


그림 13 (a) HCA, (b) Infiniband 36 ports 스위치, (c) Gigabit Ethernet 32 ports 스위치, (d) HDR 200Gb/s AOC.

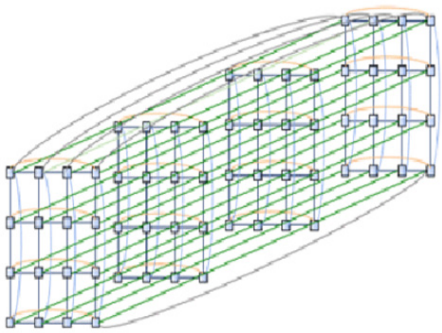


그림 14 Gordon 3D Torus network

수 있다. Mellanox 제품의 경우 Gigabit Ethernet의 switching latency가 300ns로 Infiniband보다 불리하다. 그러나 내 외부 망과의 연결이 용이한 장점이 있다. 광케이블은 동일한 QSFP 표준으로 연결 가능하고 HCA 대신 Ethernet card를 사용한다.

Gordon의 3D torus network은 그림 14, 15와 같다. Infiniband의 36개 ports 중 18개에는 compute nodes가 16개, I/O nodes가 2개, 연결되고 나머지 18개 ports는 Infiniband 서로 간에 연결된다. 18개 중 +x, −x, +y, −y, +z, −z의 6개 방향으로 각 3개 ports가 할당된다. 각 ports는 QDR 40Gb/s QSFP가 연결되므로 각 node와 infiniband 사이에는 40Gb/s의 통신속도를 가지고 infiniband와 infiniband 사이에는 120Gb/s의 통신속도를 갖는다. 3D torus의 가

표 4 AOC 통신 속도

종류	SFP	QSFP	nm/거리
SDR	2.5Gb/s	10Gb/s	850/500m
DDR	5Gb/s	20Gb/s	850/400m 1310/10km
QDR	10Gb/s	40Gb/s	850/300m 1310/10km
FDR	14Gb/s	56Gb/s	850/300m 1310/10km
EDR	25Gb/s	100Gb/s	850/100m 1310/10km
HDR	50Gb/s	200Gb/s	850/100m

장자리 +x는 반대편 가장자리 −x와 연결되고 y와 z 방향도 같은 방법으로 연결되어 모든 infiniband가 동일한 접속 구조를 갖는다. 따라서 모든 node도 위치에 관계없이 동일한 접속 구조를 갖는다. 이와 같은 연결을 2D에서 하면 torus 구조(속이 비어 있는 도넛 모양)가 되어서 붙여진 이름이다.

Gordon 슈퍼컴퓨터는 각 node의 HCA1이 하나의 3D torus를 구성하고 HCA2이 다른 하나의 3D torus를 구성하여 두 개의 독립적인 3D torus 구조를 가지고 있다. 이는 node 간 통신속도가 2배로 증가할 뿐만 아니라 하나의 torus network에 fail이 발생할 경우 다른 torus network이 작동하여 사고에 대한 안정성을 높이게 된다. 3D torus 구조는 mesh network 구조이고, 이는 두 node 간 통신이 최단 경로에 있는 다수의 스위칭 지점을 하나씩 건너뛰면서 전달된다. 따라서 두 node가 인접할 경우 통신 시간이 유리한 반면 멀리 떨어질 경우 불리하게 된다. 인접한 32개의 compute nodes와 2개의 I/O nodes가 1개의 supernode를 구성하도록 하여 하나의 supernode가 단독으로 프로그램을 실행하거나 총 32개의 supernode가 병렬적으로 계산을 수행하도록 하였다. 그림 15와 같이 64개의 I/O node는 4PB의 system storage와 연결되고 각 node는 다수의 ethernet 서버를 통해 외부 ethernet network과 연결된

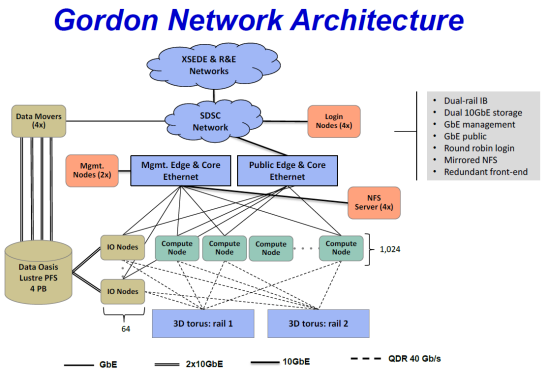


그림 15 Gordon 네트워크 구조

다. Gordon 슈퍼컴퓨터의 이론적 최대 계산 속도는 336.1TFLOPs이고 실제 측정값은 285.8TFLOPs이다.

중국은 ARM사 license의 스마트폰 SoC(Kirin990)와 대만 기업과 합작으로 생산하는 x86 CPU(ZX-E, KX-6000) 외에 순수 중국 기술로 개발한 CPU는 대표적으로 Loongson(龍芯, 용심)과 Sunway(神威, 신위)가 있다. 중국은 자국 기술의 CPU 개발을 위해 제10차 5개년(2001~2005)부터 시작하여 제12차 5개년(2011~2015)이 끝나는 2015년경에 위 두 processor의 성능을 상용 제품에 적용할 수준까지 높인 것으로 알려졌다. Loongson은 Chinese Academy of Sciences 산하 Institute of Computing Technology(ICT)에서 국가 프로젝트(863 program)로 개발하였고 현재는 민관 합작으로 설립된 Loongson Technology사에서 생산 판매하고 있다. 2019년 발매한 최신 제품은 Loongson 3B4000, 2.0GHz, 64bit(4개 core)이다. 보고된 이론적 최고 성능(peak FLOPS)은 128GFLOPS이다. 비교를 위해 Gordon에 사용된 2011년 형 Intel Xeon E5-2670(8개 core)의 경우 4개 core로 환산할 경우 83.25GFLOPS이다. Loongson processor는 MIPS instruction set을 기초로 하여 자체 개발한 LoongISA을 사용하였고, 이 중 부족한 4개 instructions은 MIPS Technologies로부터 license를 구매한 것으로 알려져 있다. MIPS을 기반으로 하기 때문에 x86 CPU의 Window OS나 ARM CPU의 Android OS를 사용하지 못하고 Linux를 사용하여야 하며 이러한 단점을 보완하기 위해 LoongISA에 translation instructions를 다수 추가하여 Linux OS에서 Window와 Android OS program을 불러올 수 있도록 하였다. 2015년경부터 국방, 통신, e-정부, 운송 등 보완이 중요한 중국 국가 시설에 보급되어 사용되는 것으로 알려졌다. 또한 인공위성에 사용하기 위해 방사선에 강한 설계(RHBD: radiation hardening by design)를 적용한 제품(Loongson X-CPU)을 개발하여 2015년 3월 30일 발사한 BeiDou 인공위성에 탑재한 것으로 알려져 있다. Sunway processor는 National High Performance IC(Shanghai) Design Center에서 주로 국방분야에 사용할 목적으로 개발되었고 2016년형 최신 제품 Sunway SW26010은 슈퍼컴퓨터 'Sunway TaihuLight'의 processor에 사용되고 있다.

Sunway TaihuLight는 중국의 한 연구소, NRCPC(National Research Center of Parallel Computer Engineering & Technology)에서 제작하였고, 상해 소재 National Supercomputing Center에 설치되어 있다. 3개 계층의 fat tree network 구조를 기반으로 40,960nodes(10,649,600cores)를 가지고 있다. 각 node는 그림 16과 같이 4개의 core groups(CG), 260개 cores를 가진 Sunway SW26010 프로세서로 구성되어 있다. 각 core group은 그림 16(b)와 같이 1개의 big core로 된 MPE(management processing element)와 64개의 작은 cores로 된 CPE(computing processing elements), MC(memory controller)가 있다. MPE core에는 L1 data cache와 instruction cache가 32KB씩 있고 L2 cache가 256KB 있다. CPE는 64개 cores에 대해 공동으로 L1 data cache 64KB, L1 instruction

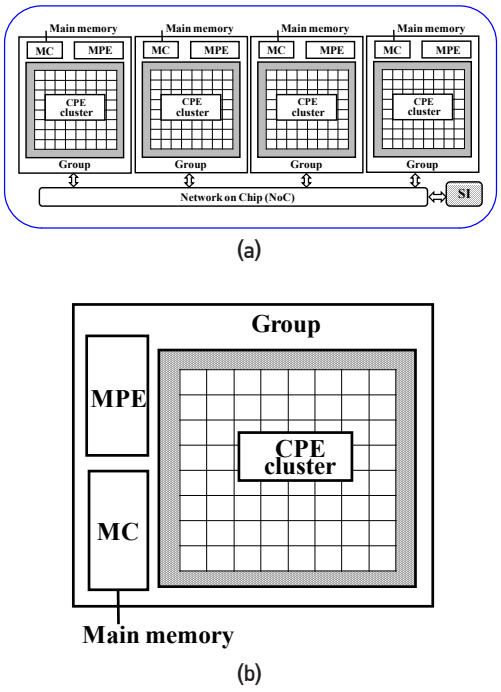


그림 16 Sunway SW26010 프로세서로 구성
(a) 1개 die에 260개 core가 있음, (b) 개별 group 구조

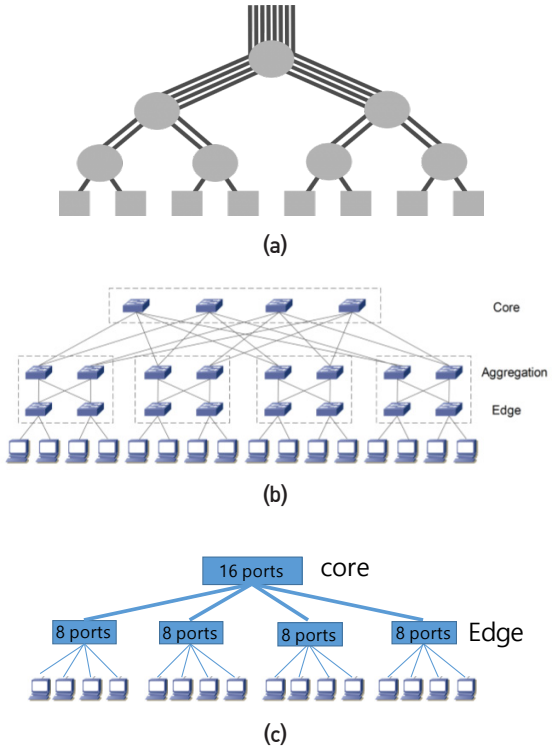


그림 17 Fat tree 네트워크 구조

cache 16KB가 있다. 각 core group당 8GB SDRAM이 있다.

4개의 core groups은 network-on-chip bus(128bit)로 연결되어 있고 SI(system interface)을 통해 광통신 단자와 연결된다. SI는 PCIe 3을 사용하며 최대 전송속도는 16GB/s이다. Sunway의 주파수 1.45GHz에 대해 전송속도는 5.8GB/s이다. Instruction set은 RISC를 기반으로 하나 자체 개발한 ISA와 Architecture를 사용하고 있다고 주장하며 구체적인 내용은 공개하지 않고 있다. MPE core는 2개의 pipelines이 있고 CPE core는 single pipeline이며, 둘 다 64-bit, SIMD, out of order 의 특성을 지니고 있다. 성능과 구조를 Intel processor와 비교하면 quad core CPU 1개와 processing element가 256개로 구성된 GPU를 조합하고 RISC ISA를 기반으로 하여 변화된 Architecture에 필요한 instruction set을 추가한 것으로 추정할 수 있다. 하나의 pipeline당 8FLOPS/cycle의 계산 속도를 가지고 있다. core의 주파수가 1.45GHz이므로 1개 node당 이론적 최대 계산 속도는 (16FLOPS×4+8 FLOPS ×256) ×1.45=3.0624TFLOPS이다. 전체 40,960nodes에 대해 125.4PFLOPS의 값을 가지며 Linpack Benchmark를 통해 측정한 값은 93.01PFLOPS이다. 2018년도에 1위에 올랐고 현재는 3위에 올라 있다.

Fat tree network 구조는 그림 17(a)와 같이 tree 모양의 연결구조를 가지나 상위 계층으로 갈수록 통신라인 수가 비대해지는 모양에서 붙여진 이름이다. 일반적인 tree 구조는 계층에 관계없이 동일한 통신라인 수를 가지므로 상위 계층으로 갈수록 통신속도가 부족하게 된다. Non-blocking network은

각 스위치에서 하향 라인 수와 상향 라인 수가 같다. 이 경우 스위치에 입력된 모든 신호는 buffer에 대기하는 시간 없이 동시에 출력된다. 슈퍼컴퓨터와 데이터 센터는 대부분 non-blocking 구조이다.

각 스위치가 4개 ports를 가지고 있는 3개 계층(three layers)의 fat tree 구조를 그림 17(b)에 나타내었다. 통상 상위 layer를 core, 중간 layer를 aggregation, 하위 layer를 edge, 서버를 leaf라 부른다. 왼쪽 2개 node는 서로 간에 edge 스위치, 4개는 aggregation 스위치만 거쳐서 도달할 수 있고 나머지는 모두 core 스위치를 거쳐야만 도달할 수 있다. Core 스위치를 거치는 경우 모든 경로는 동일한 길이를 가지고 있고 이는 fat tree의 특징이다. 하나의 경로가 fail 이 날 경우 다수의 우회 경로를 가지고 있고 경로 길이가 모두 같다. Ports가 n개인 동일한 스위치로 three layers fat tree를 구성할 경우 일반적으로 core 스위치는 최대 $(n/2)^2$, aggregation과 edge 스위치는 최대 $n^2/2$, 서버는 최대 $2(n/2)^3$까지 연결 가능하다. 만약 64개 ports의 스위치를 사용할 경우 서버는 6만5천 개 이상 가능하다. Fat tree는 많은 수의 서버를 연결하는 데 유리한 반면, 완성 후 새로운 서버를 추가하는 데 매우 불리하다. 3D torus의 경우 상하 계층이 없으므로 맨 끝의 라인을 열고 새로운 서버를 추가한 후 이전과 동일한 방법으로 라인을 연결하면 된다. 점선으로 표시된 부분 안의 스위치 4개를 내부적으로 연결하여 16ports 스위치 1개와 8ports 스위치 4개를 만들면 그림 17(c)와 같이 two layers fat tree network를 만들 수 있다. 외부적으로 다르게 보일 뿐 동일한 network 구조이다. Network 장비 업체는 16ports, 32ports 또는 36ports 스위치 다수를 내부적으로 연결하여 64ports, 128ports, 256ports 등 많은 수의 ports를 가진 스위칭 장비를 판매하고 있다.

Sunway는 Mellanox사로부터 스위치 칩과 HCA를 구매하였다. 수입한 스위치 칩과 HCA를 이용하여 custom switch 장비를 제작한 것으로 추정된다.

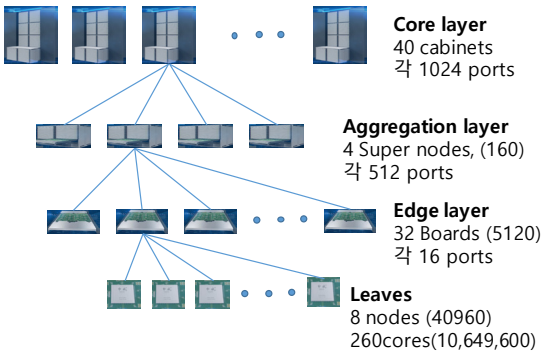


그림 18 Sunway TaihuLight 네트워크 구조

Sunway TaihuLight는 그림 18과 같은 fat tree 계층 구조를 가지고 있다. 상위 core layer는 40개의 cabinets이 있고 각 cabinet에는 4개의 supernodes가 있다. 각 supernode에는 32개의 boards가 있고 각 board에는 8개의 nodes가 연결된다. Non-blocking 구조를 지니기 위해 각 board는 16개 ports의 switch를 가지고 있고 그 중 8개는 nodes와 나머지 8개는 super node의 스위치와 연결된다. Super node의 스위치는 512개 ports를 가지고 256개는 하향으로, 나머지 256개는 상향으로 cabinet 스위치와 연결된다. Cabinet 스위치는 1,024개 ports가 있고 하향으로 160개의 super nodes와 분산하여 연결된다. Supernode의 하향 256개 ports도 128개 boards와 분산하여 연결된다. 전체 네트워크를 central switch network, management network, storage network으로 세분화하고 각 계층의 switch들을 management server 또는 storage system과 연결시킨다(더 자세한 내용은 참고문헌[8] 참조). 그리고 management sever를 통해 외부 Ethernet과 연결된다. 광케이블은 EDR 100Gb/s QSFP를 사용하였다. 이 슈퍼컴퓨터를 제작하는 데 $270M USD(약 3000억 원)이 소요되었

고, 중국 중앙정부(Central Chinese government), 장쑤성(Province of Jiangsu), 그리고 Wuxi시(City of Wuxi)가 나누어 분담한 것으로 알려져 있다[6-8].

IV. 결론

x86 CPU를 설계, 제작하는 기업은 Intel, AMD, VIA 가 있다. Intel의 mobile CPU i7는 하나의 칩 위에 다수의 소자가 집적된 SoC이다. 4개 core와 GPU, system agent가 ring bus로 연결되어 있다. Intel의 최신 서버용 CPU는 28개 코어가 한 die에 배치되어 있고 두 개 die를 하나의 패키지에 포함시켜 총 56개 core CPU이다. Intel은 core 수가 10개 이상인 데스크톱과 서버용 제품에 대해 ring bus 대신 mesh 구조를 사용하고 있다.

2019년 11월 기준 최고 성능 순위 슈퍼컴퓨터는 1, 2위에 IBM이 제작한 Summit와 Sierra가 올라가 있고 3위는 중국의 Sunway TaihuLight가 올라 있다. 슈퍼컴퓨터에는 약 10~100만 개의 컴퓨터 서버가 광통신 네트워크로 연결되어 있고 서버는 통상 10~100개 core로 구성되어 슈퍼컴퓨터는 대략 100~1,000만 개의 core를 가지고 있다. 슈퍼컴퓨터에 사용되는 통신 구조는 fat tree와 3D torus 구조를 거의 대부분 채택하고 있다.

중국은 ARM사 license의 스마트폰 SoC(Kirin990)와 대만 기업과 합작으로 생산하는 x86 CPU(ZX-E, KX-6000) 외에 순수 중국 기술로 개발한 CPU는 대표적으로 Loongson(龍芯, 용심)과 Sunway(神威, 신위)가 있다. 중국은 자국 기술의 CPU 개발을 위해 제10차 5개년(2001~2005)부터 시작하여 제12차 5개년(2011~2015)이 끝나는 2015년경에 위 두 processor의 성능을 상용 제품에 적용할 수준까지 높인 것으로 알려졌다.

참고문헌

[1] Jim Turley, "Introduction to Intel Architecture," White Paper, Intel Corporation, 2014.

[2] Intel, "An Introduction to the Intel QuickPath Interconnect," White Paper, Intel Corporation, 2009.

[3] Akhilesh Kumar, "New Intel Mesh Architecture: The 'Superhighway' of the Data Center," White Paper, Intel Corporation.

[4] Michael Berktold, "CPU Monitoring with DTS/PECI," White Paper, Intel Corporation, September 2010.

[5] Intel, "Thermal/Mechanical Specification and Design Guide (TMSDG); Intel Core i7 Processor Family for the LGA2011-3 Socket," White Paper, Intel Corporation, August 2014.

[6] Shawn Strande et al., "Building a Data - Intensive Supercomputer Architecture for the National Research Community," White Paper, Cray Inc, WP-CCS-Gordon01-0413, 2013.

[7] Weiwu Hu et al., "An introduction to CPU and DSP design in China," SCIENCE CHINA Information Sciences, vol.59, January 2016, pp. 012101:1-012101:8.

[8] Jack Dongarra, "Report on the Sunway TaihuLight System," Tech Report UT-EECS-16-742, University of Tennessee, Department of Electrical Engineering and Computer Science, June 2016.