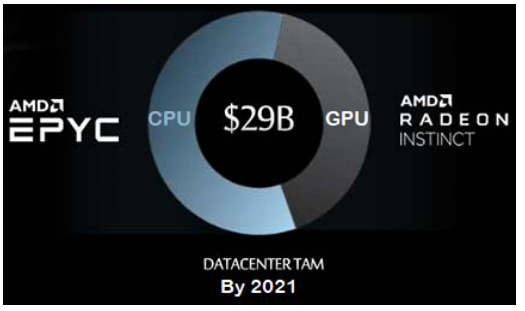
최신 ICT 이슈

Ⅱ. 무어의 법칙 종언 시대, CPU-GPU-DSA의 서버 프로세서 3파전

■ **인텔의 CPU 'Xeon(제온)'의 라이벌은 AMD나 암(Arm)의 서버 프로세서만이 아니며, 급성장 중인 AI(인공지능) 분야에서도 GPU에 밀린 지 오래이며, 새로운 경쟁자도 계속 나타나고 있음**

- AMD의 리사 수 CEO에 따르면 2021년에는 데이터센터용 프로세서 시장에서 GPU의 점유율이 CPU에 육박하는 수준으로 성장할 것이라고 전망하였음
- 리사 수는 데이터센터용 프로세서 시장 규모가 2018년 200억 달러에서 2021년 290억 달러로 성장할 것으로 내다보는데, 특히 GPU는 연간 수십 % 성장하여 전체 시장의 약 40%를 차지할 것으로 예상하고 있음
- 향후 성장이 기대되는 대규모 시뮬레이션이나 딥러닝 분야 등은 CPU보다 GPU가 더 적합하기 때문이라는 것이 그 이유
- 이런 판단 하에 AMD는 2018년 11월 제품 발표회에서 서버 프로세서 "EPYC(에픽)"의 차기 버전인 "ROME(롬, 개발 코드명)"뿐만 아니라, 새로운 데이터센터용 GPU 제품으로 "Radeon Instinct MI60"과 "MI50"을 발표하였음



<자료> AMD

[그림 1] 2021년 데이터센터용 프로세서 시장 전망

- 리사 수 CEO는 새로운 GPU 제품들이 7nm(나노미터) 공정에서 처음 생산되는 데이터센터용 GPU라고 설명하며, AMD의 데이터센터 시장 공략은 CPU와 GPU라는 쌍두마차를 통해 전개될 것임을 강조하였음
- AMD의 행보에는 약간의 조바심이 느껴지는데, 그도 그럴 것이 AMD보다 먼저 데이터센터 시장에 진입한 엔비디아(NVIDIA)는 이미 이 분야에서 GPU로 눈부신 성공을 거두었기 때문
- 엔비디아의 2018년 8~10월 기간 결산을 보면 데이터센터 사업부문의 매출은 7억 9,200만

---

* 본 내용과 관련된 사항은 산업분석팀(☎ 042-612-8296)과 최신ICT동향 컬럼리스트 박종훈 집필위원(soma0722@naver.com ☎ 02-576-2600)에게 문의하시기 바랍니다.

** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.

달러로 2년 전인 2016년 8~10월 매출에 비해 3.3배 증가했으며, 이러한 실적은 데이터센터용 GPU 시장이 얼마나 빠르게 성장하고 있는지를 잘 보여주고 있음

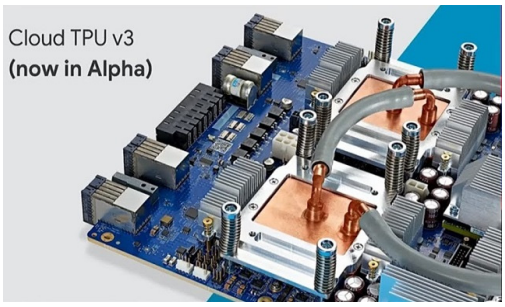
- 사업의 규모가 달라 성장률을 단순 비교하기는 어렵지만 비슷한 시기에 인텔의 데이터센터 사업부문 매출은 45억 4,200만 달러에서 61억 3,900만 달러로 1.4배 증가하였음

■ **향후 성장이 기대되는 AI 부문의 수요를 노리고 있는 것은 GPU만이 아닌데, 최근 주목받고 있는 기술은 구글의 "클라우드 TPU"임**

- 2018년 11월 스탠퍼드 대학에서 개최된 로봇 심포지엄 Bay Area Robotics Symposium (BARS) 2018에 등단한 유명 로봇 스타트업 안키(Anki)의 공동 창업자 마크 팔라투치는 구글의 AI 클라우드인 "Cloud TPU(클라우드 TPU)"를 사용 중이라고 밝혔음
- 안키가 2018년에 출시한 '벡터(Vector)'는 사용자의 친구가 되어 대화 및 게임 등 커뮤니케이션을 통해 즐거움을 주는 장난감 로봇으로, 카메라로 촬영한 이미지에서 사용자의 얼굴을 인식하는 기능과 사용자의 음성을 인식하는 기능 등 고급 AI 기능을 갖추고 있음
- 안키에 따르면 기계학습 추론 처리 시, 이미지 인식과 같은 가벼운 작업은 벡터 로봇에 내장된 퀄컴의 스마트폰용 프로세서인 Snapdragon(스냅드래곤)으로도 처리할 수 있음
- 그러나 음성 인식과 같이 중요한 추론은 스마트폰용 프로세서로 처리하기는 어렵기 때문에 클라우드 측에 맡기는데, 그 클라우드 백엔드에 구글의 클라우드 TPU를 사용 중이라고 함

■ **클라우드 TPU는 구글이 자체 개발한 딥러닝 전용 프로세서인 'TPU(Tensor Processing Unit)'를 종량제 방식으로 이용할 수 있는 구독형 서비스임**

- TPU의 2세대인 "Cloud TPU v2"는 ALU (Arithmetic and Logic Unit: 산술논리연산장치)를 32,768개 탑재하여 딥러닝에 필요한 연산을 초당 180테라 회(180T FLOPS) 실행할 수 있으며, 3세대 "Cloud TPU v3"에서는 초당 420테라 회 연산이 가능함
- 엔비디아의 데이터센터용 GPU의 최신 버전인 "Tesla V100"이 딥러닝의 연산을 초당 125테라 회 실행 가능한 것과 비교해 본다면, 최소한 딥러닝의 성능에 관해서라면 여기에 특화된 구글의 클라우드 TPU가 범용 목적의 GPU를 크게 앞선다고 볼 수 있음



<자료> Artificial Intelligence Videos

[그림 2] 수냉식을 채택한 Cloud TPU 3세대

- 안키와 같은 고도의 기술력을 가진 스타트업이 클라우드 TPU를 채택하고 있다는 점은 클라우드 TPU의 압도적 성능을 방증하는 것임
- 구글이 클라우드 TPU를 선보이자, 뒤이어 화웨이 테크놀로지와 아마존웹서비스(AWS) 역시 딥러닝 전용 프로세서를 개발 중에 있음
- 화웨이가 2018년 10월 발표한 'Ascend 910'은 딥러닝에 필요한 연산을 초당 최대 256테라회 실행할 수 있다고 하는데, Ascend 910은 2019년 2분기에 정식 출시될 예정임
- AWS도 2018년 11월 딥러닝 전용 프로세서로 "AWS Inferentia"를 발표했으며, 2019년 하반기부터 클라우드에서 이용할 수 있게 될 이 프로세서는 딥러닝의 연산을 초당 수백 테라회 실행할 수 있다고 알려져 있음
- 구글에 이어 주요 클라우드 사업자들이 동참함에 따라, 향후 AI의 워크로드 처리를 놓고 GPU와 딥러닝 전용 프로세서 사이에 경쟁이 본격화될 것으로 예상됨

■ **TPU와 같이 특정 용도에 특화된 프로세서를 "도메인 특화 아키텍처(Domain Specific Architecture: DSA)"라고 부르는데, 하드웨어 성능 개선의 방법으로 주목받고 있음**

- DSA의 대표적인 옹호론자는 RISC(축약 명령어 세트 컴퓨터) 프로세서를 고안한 2명의 컴퓨터 과학자로, 스탠퍼드 대학 학장을 역임한 존 헤네시와 버클리 캘리포니아 대학의 교수를 역임한 데이빗 패터슨은 DSA만이 하드웨어의 성능을 향상시키는 길이라고 적극 주장하고 있음

<자료> James Hamilton's Blog

[그림 3] 2017 튜링 어워드 수상자

- "Computer Architecture: A Quantitative Approach(컴퓨터 아키텍처: 정량적 접근)"와 "Computer organization and design(컴퓨터 구성과 설계)"의 저자로 알려진 두 사람은 2017년 튜링 어워드를 공동 수상한 컴퓨터 과학의 권위자들임
- ASIC(주문형 반도체)이 "애플리케이션 특화(Application Specific)", 즉 하나의 응용 분야에 최적화된 칩인 반면, DSA는 응용프로그램보다 범위가 넓은 도메인(산업 및 기술영역)에 특화된 칩이라고 할 수 있음
- 현재 구글 모기업인 알파벳(Alphabet)의 회장을 맡고 있기도 한 존 헤네시는 2018년 5월에 열린 "Google I/O" 컨퍼런스의 강연에서 DSA에 대한 관심이 증가하는 이유가 "무어의 법칙의 종언"이 다가오고 있기 때문이라고 설명

- CPU는 지금까지 집적회로의 트랜지스터 수가 1.5~2년마다 2배로 늘어난다는 무어의 법칙에 따라 성능을 향상시켜 왔으나, 현재는 반도체 제조 공정의 미세화가 한계에 다다르고 있음
- 인텔도 제조 공정의 미세화에 고전하고 있고, AMD나 IBM의 반도체 제조부문을 인수한 글로벌파운드리즈도 2018년에 반도체 제조 공정 신규 개발에서 철수한 바 있음
- 무어의 법칙이 붕괴되면 단순히 트랜지스터의 수를 늘림으로써 성능을 향상시킬 수는 없게 되며, 그렇다면 향후 어떻게 성능을 향상시킬 지가 반도체 산업의 화두가 되고 있음

■ **존 헤네시는 시간이 지나면 "오래된 아이디어가 다시 새로운 것이 되는 법(Everything old is new again)"이라며, 과거에는 작동하지 않던 DSA가 이제는 성능을 발휘하게 되었다고 주장**

- 과거 범용 CPU 시대에 시도되었으나 당시에는 잘 작동하지 않았던 기술들이 있으며, DSA도 그 중 하나로 용도를 제한함으로써 이제는 오히려 성능을 발휘할 수 있게 되었음
- 일례로 "VLIW (Very Long Instruction Word)" 아키텍처를 들 수 있으며, 이는 과거 인텔이 64비트 프로세서인 Itanium(이타늄)에 채택했으나 성공하지 못했던 기술로, 하나의 명령어를 통해 여러 개의 명령어를 처리하도록 하는 방식임
- 이 방식은 범용 CPU에 VLIW를 채용할 경우 프로그램에 포함된 복수의 명령어를 병렬로 실행시키기 위한 컴파일러를 개발하는 것이 어려워 잘 작동하지 않았음
- 그런데 용도를 한정한다면 여러 명령어의 병렬 실행이 용이하게 될 수 있으며, 존 헤네시에 따르면 DSA에서는 VLIW에 의한 성능 향상이 실현될 수 있다고 함
- 그 밖에도 DSA에서는, 높은 범용성을 목표로 하는 CPU에서는 채택하지 못했던 기술, 가령 부동소수점 연산의 정확도를 낮추는 대신 연산 횟수를 늘리는 등의 기술 적용이 가능한데, 실제 이 기술은 구글의 TPU와 화웨이의 Ascend 910 등 딥러닝용 DSA가 채택하고 있음
- 지금까지는 범용성이 높은 CPU의 성능 개선이 순조롭게 진행되어 왔기 때문에, 틈새시장 전용의 프로세서가 활약할 기회가 주어지지 않았지만, CPU의 성능 향상 방법이 막히면서 다양한 도메인에 다양한 DSA가 등장하게 될 것으로 존 헤네시는 전망하고 있음

■ **실제로 딥러닝 이외 영역에서도 DSA가 등장하고 있으며, 베어풋 네트웍스(Barefoot Networks)는 SDN(소프트웨어 정의 네트워크) 전용 DSA인 'Tofino(토피노)'를 개발하였음**

- 네트워크 기기에서는 패킷 처리의 구조를 하드웨어 단으로 떨어뜨린 ASIC을 사용하는 것이 일반적이었으나, 2010년대에 들어서면서 패킷 처리를 소프트웨어를 통해 정의하는 'Open Flow(오픈플로우)'와 같은 SDN 기술이 주목받게 되었음

- SDN 기술에서 소프트웨어 처리는 지금까지 CPU가 담당해 왔으나, 베어풋 네트웍스는 토피노로 CPU를 대체하려 하고 있음
- 베어풋 네트웍스는 토피노가 초당 테라비트급 패킷 처리가 가능하고, CPU로 처리하는 것에 비해 레이턴시(지연)를 500분의 1 이하로 할 수 있다며, 토피노의 성능은 ASIC과 맞먹지만 ASIC과 달리 토피노는 프로그래밍을 통한 처리가 가능한 장점이 있다고 주장
- 서버 프로세서의 분야에서 시작된 혼전은 이제 CPU 사이의 싸움만이 아니며, CPU대GPU, GPU대DSA, DSA대CPU 등 서로 다른 아키텍처 간 경쟁으로 발전해 나갈 조짐을 보이고 있어, 앞으로 시장 전개 상황을 주시할 필요가 있음

[참고문헌]

[1] ITPro, 12. 3, https://nkbp.jp/2stwzeA
[2] Inverse, 1. 9, https://bit.ly/2FDwY5X