컴퓨터학콜로키움 **(COSE-405) :: 8**주차 구건재 교수님

Modern Computer Architecture Research

강의 : 2022-10-26 / 작성 날짜 : 2022-10-27

고려대학교 컴퓨터학과 2017320108

고재영

기본적으로 컴퓨터 시스템은 하드웨어와 소프트웨어의 구성으로 나누어 볼 수 있다. 물질적으로 컴퓨터를 구성하는 기계적인 장치로 하드웨어를 볼 수 있고, 소프트웨어는 이러한 하드웨어에 대한 동작을 지시하거나 제어하는 명령어의 집합으로 볼 수 있다. 최근에 IT 분야안에서도 특히나 각광을 받고 많은 붐을 일으키는 것이 AI 인공지능이나 머신러닝인 만큼, 사실 어느 정도 컴퓨터 구조란 분야에 관해서 많은 사람들이 등한시하고 관심도가 상대적으로 떨어지는 것은 사실이다. 하지만 컴퓨터 아키텍쳐란 분야 자체가 컴퓨터의 등장과 관련하여 모태가 되는 이른바 근본적인 부분에서부터 출발했기 때문에 절대 간과해서는 안 될 분야임에 틀림이 없다. 오늘 이 시간에는 구건재 교수님께서 컴퓨터 아키텍쳐 분야에 관한 최근 연구 동향에 대해 강연을 하셨다.

컴퓨터 공학도라면 당연하게도 최초의 컴퓨터를 물어본다면 에니악 (ENIAC)이라고 쉽게 말할 수 있을 것이다. 이 에니악의 형태는 무지막지하게 큰데, 이건 곧 하드웨어가 매우 복잡한 것을 의미한다. 당시에는 운영체제란 개념도 없었기 때문에, 구멍이 뚫려있는 형태의 단순한 카드를 넣고 처리하는 형태로 그쳤다. 그런 기술적인 한계만큼 소프트웨어는 굉장히 간단하게 설계되었다고 볼 수 있고, 규격화된 instruction도 존재하지 않았기에 오늘날과 같은 형태의 프로그래밍이란 불가능한 환경이었다. 하지만 이후에 기계어 어셈블리어등에 대한 연구등을 통해 명령어 기반의, ISA 개념이 등장하면서 많은 발전이 일어났다.

구건재 교수님께서 강의에서 설명하신 부분 중 큰 줄기로 나누어 보면, 첫 번째에 해당한 topic은 Domain Specific Architecture에 대한 것이었다. 우선 구조적인 부분과 관련하여 연구나 실제 리얼 월드와 관련한 현황을 말씀하셨다. 컴퓨터 구조를 학부에서 배울 때 아마 초창기에 먼저 나오는 용어 중

하나로 대표적인 것이 바로 무어의 법칙일 것이다. 반도체칩 기술의 발전속도에 관련해서, 반도체 칩에 집적할 수 있는 트랜지스터에 대해 매 18개월마다 두 배씩 증가를 한다는 것이다. 이 발전 속도라는 것이 단순히 Linear한 상승을 보이는 것이 아니라, exponential한 수준의 상승곡선을 그려온 것으로 하드웨어를 포함한 발전이 여태껏 굉장한 규모로 이루어졌다. 하지만 불행하게도 최근 들어 이에 대한 전망이 마냥 좋지 않다. 많은 회사를 보면 알 수 있듯이, 싱글 코어 자체의 성능을 발전시키는 데에는 한계에 이르렀기 때문에 멀티 코어를 적용하는 방향으로 노력을 기울이고 있다. 이런 예를 포함해서 어느새 무어의 법칙이 지속될 것 같아보이지 않다고 연구자들이 생각하기 때문에, 기존과 다른 새로운 아키텍쳐를 찾는 방법을 강구하기 시작했다. 이 중 하나가 바로 DSA로 특정 도메인 분야에 관련한 기능에 관한 프로세싱을 가속화하는 것이 메인 아이디어로 볼 수 있다. 대표적으로 찾아 볼 수 있는 것이 바로 그래픽 성능을 위한 GPU라거나, neural network의 응용을 처리하는 데에 가속화하는 구글의 TPU와 같은 것이다.

두 번째로 말씀하신 주제는 Near Data Processing이다. 이전에 머신러닝이나 컴퓨터 비전과 관련하여 누누이 언급하듯이 현대에는 정말 많은 데이터의 범람을 겪고 있다. 데이터의 규모가 엄청나게 늘어났기 때문에 방대한 데이터 프로세싱을 위한 프로세서, 또 저장공간에도 관련하여 저장하는 데에 훌륭한 성능의 메모리의 필요성도 함께 대두되었다. 이 때문에 등장한 패러다임은, 물론 기존 컴퓨터 구조의 방향을 완전히 바꾸기도 한 이것은 바로 In-Storage Computing 이다. 이 개념은 전통적인 Data Hierarchy 모델을 떠올린다면 생각치 못한 부분이다. In-Storage processing은 SSD 내부에서 연산을 수행하고 이후 결과 데이터만을 전송하는 Near Data Processing 기법 중 하나이다. 그 결과로 데이터를 전송하는 데에 소요되는 시간을 효과적으로 감소시키면서 전체 수행시간에 대해 이점을 가져 연산 오버헤드를 줄여주는 데에 의의가 있다. 결국에 현재에 있어서 컴퓨터 분야에서는 트랜지스터가 작아질수록 전력 사용량이 면적에 비례하여 유지되는 Dennard Scaling과 반도체칩에 집적가능한 트랜지스터 개수의 exponential한 증가를 보이는 Moore’s law 두 개의 근간 법칙이 곧 끝나감과 함께 빅 데이터를 다루는 현 국면과 응용분야 도메인에 직접 연계하는 데에 관심도가 높아짐이 맞물려 돌아가는 것이 현황이라고 볼 수 있다.