Lab Introduction

Jae W. Lee (jaewlee@snu.ac.kr)
Department of Computer Science and Engineering
Seoul National University

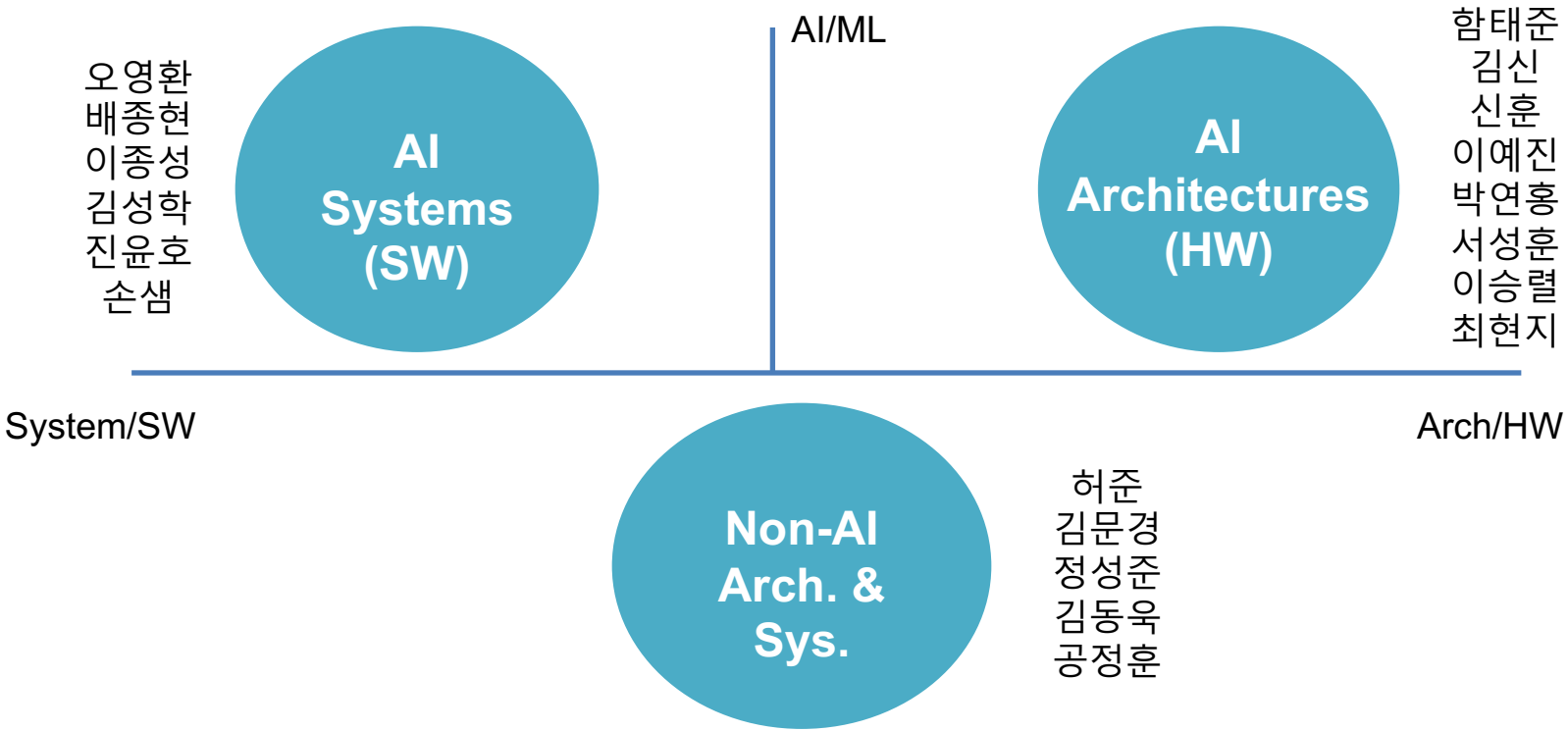
April 3th, 2020

연구실 소개 (1): Who Are We?

- **Research Group: Architecture and Code Optimization (ARC) Laboratory**
  - Rebooted in September 2016
  - Research: How to design a computer (both HW and SW) to make it faster and more energy-/cost-effective.
  - Members: 1 Postdoc, 10 Ph.D. students, 7 M.S. students, 1 Admin (as of April 2020)

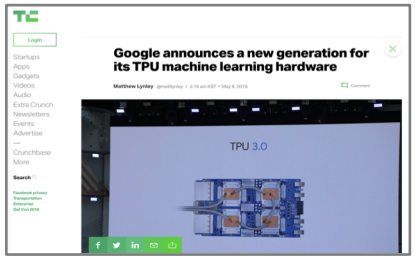
연구실 소개 (2): What Do We Do?

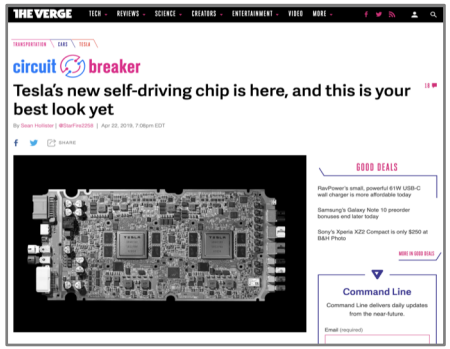
- **Three focus areas (as of April 2020)**

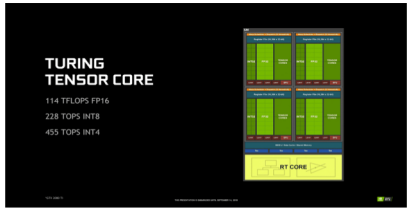


What We Do (1): AI Architectures

- **문제: ML 응용의 폭발적인 확산에 따라 기존의 general-purpose 컴퓨터로 성능/전력/비용 요구를 달성할 수 없음**
- **접근: Specialized HW + SW co-design**
  - Rise of specialized AI chips: Google, Nvidia, Microsoft, Xilinx, Tesla, …
  - Example: A^3 attention mechanism accelerator [HPCA'20]

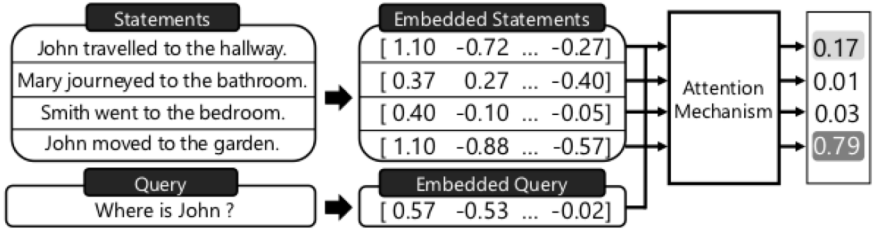




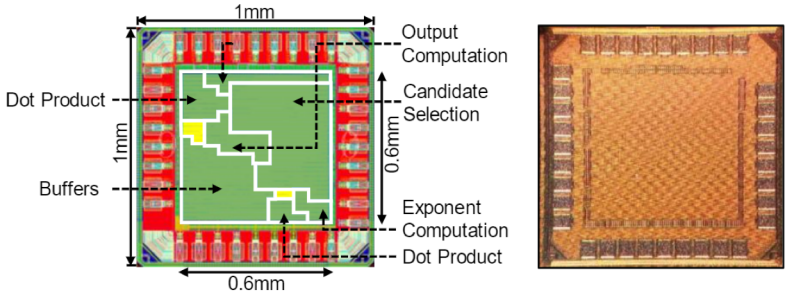


What We Do (2): AI Architectures

- **Example: A³ attention mechanism accelerator [HPCA'20]**
  - Attention determines what is relevant to the input query through content-based similarity search
    - Decides what to *attend:* Used in NLP (like BERT), Q&A systems (like MemN2N), Vision, etc.



Q&A system example



A3 test chip (TSMC40nm)

  - A specialized hardware chip that accelerates attention mechanism.
    - >100x more Performance/Watt than CPU and GPU

- **Demo video for our test cool chip is available @ YouTube! https://youtu.be/6mYXLLP0DhE**

[HPCA'20] T. Ham, et al., "A3: Accelerating Attention Mechanism in Neural Networks with Approximation", IEEE HPCA 2020.

What We Do (3): Non-AI Architectures

- **문제: 데이터센터, 모바일/엣지/IoT에서도 technology scaling의 둔화로 인해 general-purpose 컴퓨터로 성능/전력/비용 요구를 달성할 수 없음**
- **접근: (Again) Specialized HW + SW co-design**
  - Genesis: Cloud FPGA accelerator of genomic data analysis [ISCA'20a]
  - Cereal: Specialized architecture for object serialization [ISCA'20b]
  - IIU: Inverted index search accelerator [ASPLOS'20]
  - Charon: Specialized architecture for garbage collection [MICRO'19]
  - Specialized CPU for high-level PLs like JavaScript, Python, etc. [ASPLOS'17] [ISCA'16]

빅데이터 타겟

모바일, 엣지 타겟

[ISCA'20a] T. Ham, et al., "Genesis: A Hardware Acceleration Framework for Genomic Data Analysis", IEEE/ACM ISCA 2020. (To appear)
[ISCA'20b] J. Jang, et al., "A Specialized Architecture for Object Serialization with Applications to Big Data Analytics", IEEE/ACM ISCA 2020. (To appear)
[ASPLOS'20] J. Heo, et al., "IIU: Specialized Architecture for Inverted Index Search", ACM ASPLOS 2020.
[MICRO'19] J. Jang, et al., "Charon: Specialized Near-Memory Processing Architecture for Clearing Dead Objects in Memory", IEEE/ACM MICRO 2019.
[ASPLOS'17] T. Ham, et al., "A3: Accelerating Attention Mechanism in Neural Networks with Approximation", ACM ASPLOS 2020.
[ISCA'16] C. Kim, et al., "Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors", IEEE/ACM ISCA 2016.

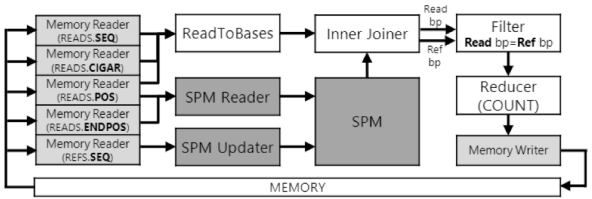
What We Do (4): Non-AI Architectures

- **Example:** Genesis, a cloud FPGA accelerator of genomic data analysis **[ISCA'20a]**
  - Gene sequencing pipeline consists of multiple stages including data manipulation operations
  - Genesis lets users to describe their target data manipulation operation in SQL and constructs a hardware accelerator for the operation using pre-constructed HW modules

SQL query representing target operation

```
/* I1: Extract Reads and Reference Partition P */
CREATE TABLE ReadPartition AS
SELECT POS, ENDPOS, CIGAR, SEQ
FROM READS PARTITION (P)
CREATE TABLE ReferenceRow AS
SELECT POS, SEQ
FROM REF PARTITION (P)
/* I2: posExplode on ReferenceRow */
CREATE TABLE RelevantReference AS
PosExplode (ReferenceRow.SEQ, ReferenceRow.POS)
FROM ReferenceRow
DECLARE @rlen int
/* Iterate over Rows */
FOR SingleRead IN ReadPartition:
  SET @rlen = SingleRead.ENDPOS - SingleRead.POS)
  /* Q1: ReadExplode to convert a read into multi-row
       table where each row represents a base pair */
  CREATE TABLE #AlignedRead AS
  ReadExplode (SingleRead.POS, SingleRead.CIGAR,
       SingleRead.SEQ)
  FROM SingleRead
  /* Q2: Inner Join two tables with the base pair's
       corresponding position as a key */
  CREATE TABLE #ReadAndRef AS
  SELECT #AlignedRead.SEQ, RelevantReference.SEQ
  FROM #AlignedRead
  INNER JOIN (SELECT * FROM RelevantReference LIMIT
       SingleRead.POS, rlen)
  ON #AlignedRead.POS = RelevantReference.POS
  /* Q3: Find the sum of matching base pairs */
  INSERT INTO Output
  SELECT SUM(#AlignedRead.SEQ == RelevantReference.SEQ)
  FROM #ReadAndRef
END LOOP;
```

Constructed hardware pipeline for the operation



Synthesized and deployed to Amazon EC2 cloud FPGA

[ISCA'20a] T. Ham, et al., "Genesis: A Hardware Acceleration Framework for Genomic Data Analysis", IEEE/ACM ISCA 2020. (To appear)

What We Do (5): Systems (Software) Research

- **문제: ML, Big Data응용의 확산으로 메모리/스토리지 수요의 폭발적으로 증가하고 있으나, 기존의 SW 스택은 성능/전력/비용 요구를 감당하지 못하고 있음**
- **접근: Customizing SW stack for key applications (possibly augmented with HW support)**
  - Hardware-assisted demand paging [ISCA'20c]
  - FlashNeuron: SSD-enabled large-batch training of very deep neural networks [USENIX ATC'20]*
  - Practical erase Suspension for modern low-latency SSDs [USENIX ATC'19a]
  - Asynchronous I/O stack: Low-latency kernel I/O stack for ultra-Low latency SSDs [USENIX ATC'19b]
  - Portable, automatic data quantizer for deep neural networks [PACT'18]

* Currently under review

FAQs

- **랩에 관심이 있는데 어디서 시작할까요?**
  - 관련 과목 수강: 컴퓨터구조, 시스템SW(시스템프로그래밍, OS, 컴파일러, 임베디드시스템, 멀티코어등), 어플리케이션(머신러닝, 데이터베이스등)
  - 인턴쉽
    - 겨울보다 여름방학 선호 (풀 타임)
    - 인턴쉽이 진학의 요구조건은 아니지만, 인턴 경험이 있는 학생의 비율 높음
    - 많은 학생들이 실제적인 기여를 하여 탑 컨퍼런스 논문 (공)저자가 됨: ISCA, ASPLOS, MICRO등
- **장학금 지원은 어떻게 되나요?**
  - 학비 + 생활비 + 인센티브 지원합니다.
  - 학생들이 학업에 전념할 수 있는 환경을 제공하려 노력하고 있습니다.