Bachelor’s Seminar in Comp Arch

- Fall 2021
- 2 credit units

- **Rigorous seminar on fundamental and cutting-edge topics in computer architecture**
  - Critical presentation, review, and discussion of seminal works in computer architecture
    - We will cover many ideas & issues, analyze their tradeoffs, perform **critical thinking** and brainstorming

- Participation, presentation, synthesis report
- You can register for the course online
  - [https://safari.ethz.ch/architecture_seminar/spring2021](https://safari.ethz.ch/architecture_seminar/spring2021)
Announcement

- If you are interested in learning more and doing research in Computer Architecture, three suggestions:
  - Email me with your interest (CC: Juan)
  - Take the seminar course and the “Computer Architecture” course
  - Do readings and assignments on your own

- There are many exciting projects and research positions, e.g.:
  - Memory systems
  - Hardware security
  - GPUs, FPGAs, heterogeneous systems, ...
  - New execution paradigms (e.g., in-memory computing)
  - Security-architecture-reliability-energy-performance interactions
  - Architectures for medical/health/genomics
  - A limited list is here: https://safari.ethz.ch/theses/
Broader Agenda

- Single-cycle Microarchitectures
- Multi-cycle and Microprogrammed Microarchitectures
- Pipelining
- Issues in Pipelining: Control & Data Dependence Handling, State Maintenance and Recovery, ...
- Out-of-Order Execution
- Other Execution Paradigms
Approaches to (Instruction-Level) Concurrency

- Pipelining
- Fine-Grained Multithreading
- Out-of-order Execution
- Dataflow (at the ISA level)
- Superscalar Execution
- VLIW
- Systolic Arrays
- Decoupled Access Execute
- SIMD Processing (Vector and array processors, GPUs)
Readings for Today

- **Required**

- **Recommended**
Readings for Next Week

- **Required**

- **Recommended**
Systolic Arrays
Systolic Arrays: Motivation

- Goal: design an accelerator that has
  - Simple, regular design (keep # unique parts small and regular)
  - High concurrency → high performance
  - Balanced computation and I/O (memory) bandwidth

- Idea: Replace a single processing element (PE) with a regular array of PEs and carefully orchestrate flow of data between the PEs
  - such that they collectively transform a piece of input data before outputting it to memory

- Benefit: Maximizes computation done on a single piece of data element brought from memory
Systolic Arrays

Figure 1. Basic principle of a systolic system.


Memory: heart
Data: blood
PEs: cells
Memory pulses data through PEs
Why Systolic Architectures?

- **Idea:** Data flows from the computer memory in a rhythmic fashion, passing through many processing elements before it returns to memory.

- **Similar to blood flow:** heart → many cells → heart
  - Different cells “process” the blood
  - Many veins operate simultaneously
  - Can be many-dimensional

- **Why?** Special purpose accelerators/architectures need
  - Simple, regular design (keep # unique parts small and regular)
  - High concurrency → high performance
  - Balanced computation and I/O (memory) bandwidth
Systolic Architectures

- Basic principle: Replace a single PE with a regular array of PEs and carefully orchestrate flow of data between the PEs
  - Balance computation and memory bandwidth

- Differences from pipelining:
  - These are individual PEs
  - Array structure can be non-linear and multi-dimensional
  - PE connections can be multidirectional (and different speed)
  - PEs can have local memory and execute kernels (rather than a piece of the instruction)
Systolic Computation Example

- **Convolution**
  - Used in filtering, pattern matching, correlation, polynomial evaluation, etc ...
  - Many image processing tasks
  - **Machine learning**: up to hundreds of convolutional layers in Convolutional Neural Networks (CNN)

\[
\text{Given} \quad \text{the sequence of weights } \{w_1, w_2, \ldots, w_k\} \\
\text{and the input sequence } \{x_1, x_2, \ldots, x_n\}, \\
\text{compute} \quad \text{the result sequence } \{y_1, y_2, \ldots, y_{n+1-k}\} \\
\text{defined by} \\
\quad y_i = w_1 x_i + w_2 x_{i+1} + \ldots + w_k x_{i+k-1}
\]
LeNet-5, a Convolutional Neural Network for Hand-Written Digit Recognition

This is a 1024*8 bit input, which will have a truth table of $2^{8196}$ entries
An Example of 2D Convolution

Structure information
Input: 5*5 (blue)
Kernel (filter): 3*3 (grey)
Output: 5*5 (green)

Computation information
Stride: 1
Padding: 1 (white)

Output Dim = (Input + 2*Padding - Kernel) / Stride + 1
An Example of 2D Convolution

Input Layer

CNN kernel

Output Layer
Convolutional Neural Networks: Demo

LeNet-5, convolutional neural networks

Convolutional Neural Networks are a special kind of multi-layer neural networks. Like almost every other neural networks they are trained with a version of the back-propagation algorithm. Where they differ is in the architecture.

Convolutional Neural Networks are designed to recognize visual patterns directly from pixel images with minimal preprocessing. They can recognize patterns with extreme variability (such as handwritten characters), and with robustness to distortions and simple geometric transformations.

LeNet-5 is our latest convolutional network designed for handwritten and machine-printed character recognition. Here is an example of LeNet-5 in action.

Many more examples are available in the column on the left:

Several papers on LeNet and convolutional networks are available on my publication page:

[LeCun et al., 1998]

[Bottou et al., 1997]
L. Bottou, Y. LeCun, and Y. Bengio. Global training of
Implementing a Convolutional Layer with Matrix Multiplication
In 2010, Prof. Andreas Moshovos adopted Professor Hwu’s ECE498AL Programming Massively Parallel Processors Class

Several of Prof. Geoffrey Hinton’s graduate students took the course

These students developed the GPU implementation of the Deep CNN that was trained with 1.2M images to win the ImageNet competition

Slide credit: Hwu & Kirk
Example: AlexNet (2012)

- AlexNet wins the **ImageNet classification competition** with ~10% points higher accuracy than state-of-the-art
- Krizhevsky et al., “*ImageNet Classification with Deep Convolutional Neural Networks*”, NIPS 2012.

---

**ImageNet Classification with Deep Convolutional Neural Networks**

Alex Krizhevsky  
University of Toronto  
kriz@cs.utoronto.ca

Ilya Sutskever  
University of Toronto  
ilya@cs.utoronto.ca

Geoffrey E. Hinton  
University of Toronto  
hinton@cs.utoronto.ca

**Abstract**

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called “dropout” that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.
Example: GoogLeNet (2014)

- Google improves accuracy by adding more network layers
  - From 8 in AlexNet to 22 in GoogLeNet
  - Szegedy et al., “Going Deeper with Convolutions”, CVPR 2015.

---

**Going Deeper with Convolutions**

Christian Szegedy\(^1\), Wei Liu\(^2\), Yangqing Jia\(^1\), Pierre Sermanet\(^1\), Scott Reed\(^3\), Dragomir Anguelov\(^1\), Dumitru Erhan\(^1\), Vincent Vanhoucke\(^1\), Andrew Rabinovich\(^4\)

\(^1\)Google Inc.  \(^2\)University of North Carolina, Chapel Hill  \(^3\)University of Michigan, Ann Arbor  \(^4\)Magic Leap Inc.

\(^1\){szegedy, jiayq, sermanet, dragomir, dmitru, vanhoucke}@google.com
\(^2\)wliu@cs.unc.edu, \(^3\)reedscott@umich.edu, \(^4\)arabinovich@magicLeap.com


Deep Residual Learning for Image Recognition

Kaiming He  Xiangyu Zhang  Shaoqing Ren  Jian Sun
Microsoft Research
{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

ImageNet experiments

First CNN

Human: 5.1%
Neural Network Layer Examples

**LeNet**

- **Image:** 28 (height) × 28 (width) × 1 (channel)
- **Convolution with 5×5 kernel+2 padding:** 28×28×6
  - **sigmoid**
- **Pool with 2×2 average kernel+2 stride:** 14×14×6
- **Convolution with 5×5 kernel (no pad):** 10×10×16
  - **sigmoid**
- **Pool with 2×2 average kernel+2 stride:** 5×5×16
  - **flatten**
- **Dense:** 120 fully connected neurons
  - **sigmoid**
- **Dense:** 84 fully connected neurons
  - **sigmoid**
- **Dense:** 10 fully connected neurons
  - **Output:** 1 of 10 classes

**AlexNet**

- **Image:** 224 (height) × 224 (width) × 3 (channels)
- **Convolution with 11×11 kernel+4 stride:** 54×54×96
  - **ReLu**
- **Pool with 3×3 max. kernel+2 stride:** 26×26×96
  - **ReLu**
- **Convolution with 5×5 kernel+2 pad:** 26×26×256
  - **ReLu**
- **Pool with 3×3 max. kernel+2 stride:** 12×12×256
  - **ReLu**
- **Convolution with 3×3 kernel+1 pad:** 12×12×384
  - **ReLu**
- **Convolution with 3×3 kernel+1 pad:** 12×12×384
  - **ReLu**
- **Convolution with 3×3 kernel+1 pad:** 12×12×256
  - **ReLu**
- **Pool with 3×3 max. kernel+2 stride:** 5×5×256
  - **flatten**
- **Dense:** 4096 fully connected neurons
  - **ReLu, dropout p=0.5**
- **Dense:** 4096 fully connected neurons
  - **ReLu, dropout p=0.5**
- **Dense:** 1000 fully connected neurons
- **Output:** 1 of 1000 classes

By Cmglee - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=104937230
Systolic Computation Example: Convolution (I)

- Convolution
  - Used in filtering, pattern matching, correlation, polynomial evaluation, etc ...
  - Many image processing tasks
  - Machine learning: up to hundreds of convolutional layers in Convolutional Neural Networks (CNN)

Given the sequence of weights \( \{ w_1, w_2, \ldots, w_k \} \) and the input sequence \( \{ x_1, x_2, \ldots, x_n \} \),

compute the result sequence \( \{ y_1, y_2, \ldots, y_{n+1-k} \} \)
defined by

\[
y_i = w_1x_i + w_2x_{i+1} + \ldots + w_kx_{i+k-1}
\]
Systolic Computation Example: Convolution (II)

- \( y_1 = w_1x_1 + w_2x_2 + w_3x_3 \)
- \( y_2 = w_1x_2 + w_2x_3 + w_3x_4 \)
- \( y_3 = w_1x_3 + w_2x_4 + w_3x_5 \)

Figure 8. Design W1: systolic convolution array (a) and cell (b) where \( w_i \)'s stay and \( x_i \)'s and \( y_i \)'s move systolically in opposite directions.
Worthwhile to implement adder and multiplier separately to allow overlapping of add/mul executions
One needs to carefully orchestrate when data elements are input to the array.

And when output is buffered.

This gets more involved when:
- Array dimensionality increases
- PEs are less predictable in terms of latency
Example 2D Systolic Array Computation

- Multiply two 3x3 matrices (inputs)
  - Keep the final result in PE accumulators

\[
\begin{bmatrix}
  c_{00} & c_{01} & c_{02} \\
  c_{10} & c_{11} & c_{12} \\
  c_{20} & c_{21} & c_{22}
\end{bmatrix} = \begin{bmatrix}
  a_{00} & a_{01} & a_{02} \\
  a_{10} & a_{11} & a_{12} \\
  a_{20} & a_{21} & a_{22}
\end{bmatrix} \times \begin{bmatrix}
  b_{00} & b_{01} & b_{02} \\
  b_{10} & b_{11} & b_{12} \\
  b_{20} & b_{21} & b_{22}
\end{bmatrix}
\]

Figure 1: A systolic array processing element

\[
P = M \\
Q = N \\
R = R + M \times N
\]
Two-Dimensional Systolic Arrays

Figure 11. Two-dimensional systolic arrays: (a) type R, (b) type H, and (c) type T.

To a given problem there could be both one- and two-dimensional systolic array solutions. For example, two-dimensional convolution can be performed by a one-dimensional systolic array\(^{24,25}\) or a two-dimensional systolic array.\(^6\) When the memory speed is more than cell speed, two-dimensional systolic arrays such as those depicted in Figure 11 should be used. At each cell cycle, all the I/O ports on the array boundaries can input or output data items to or from the memory; as a result, the available memory bandwidth can be fully utilized. Thus, the choice of a one- or two-dimensional scheme is very dependent on how cells and memories will be implemented.
Combinations

- Systolic arrays can be chained together to form powerful systems.

- This systolic array is capable of producing on-the-fly least-squares fit to all the data that has arrived up to any given moment.

Figure 12. On-the-fly least-squares solutions using one- and two-dimensional systolic arrays, with $p = 4$. 

GIVEN AN $n \times p$ MATRIX $X$ WITH $n \geq p$, AND AN $n$-VECTOR $y$, DETERMINE A $p$-VECTOR $b$ SUCH THAT $\|y - xb\|$ IS MINIMIZED.

STEP 1: ORTHOGONAL TRIANGULARIZATION
STEP 2: SOLUTION OF TRIANGULAR LINEAR SYSTEM
Systolic Arrays: Pros and Cons

- **Advantages:**
  - **Principled:** Efficiently makes use of limited memory bandwidth, balances computation to I/O bandwidth availability
  - **Specialized** (computation needs to fit PE organization/functions)
    → improved **efficiency, simple design, high concurrency/performance**
    → good to do **more with less memory bandwidth** requirement

- **Downside:**
  - **Specialized**
    → **not generally applicable** because computation needs to fit the PE functions/organization
More Programmability in Systolic Arrays

- Each PE in a systolic array
  - Can store multiple “weights”
  - Weights can be selected on the fly
  - Eases implementation of, e.g., adaptive filtering

- Taken further
  - Each PE can have its own data and instruction memory
  - Data memory → to store partial/temporary results, constants
  - Leads to *stream processing, pipeline parallelism*
    - More generally, *staged execution*
Pipeline-Parallel (Pipelined) Programs

Figure 1. (a) The code of a loop, (b) Each iteration is split into 3 pipeline stages: A, B, and C. Iteration i comprises Ai, Bi, Ci. (c) Sequential execution of 4 iterations. (d) Parallel execution of 6 iterations using pipeline parallelism on a three-core machine. Each stage executes on one core.

Stages of Pipelined Programs

- Loop iterations are divided into code segments called **stages**
- Threads execute stages on different cores

```
loop {
    Compute1
    Compute2
    Compute3
}
```
Pipelined File Compression Example

Figure 3. File compression algorithm executed using pipeline parallelism
Systolic Array: Advantages & Disadvantages

- **Advantages**
  - Makes *multiple uses of each data item* → reduced need for fetching/refetching → better use of memory bandwidth
  - **High concurrency**
  - Regular design (both data and control flow)

- **Disadvantages**
  - Not good at exploiting *irregular parallelism*
  - Relatively special purpose → need software, programmer support to be a general purpose model
Example Systolic Array: The WARP Computer

- HT Kung, CMU, 1984-1988

- Linear array of 10 cells, each cell a 10 Mflop programmable processor
- Attached to a general purpose host machine
- HLL and optimizing compiler to program the systolic array
- Used extensively to accelerate vision and robotics tasks

The WARP Computer

Figure 1: Warp system overview
The WARP Cell

Figure 2: Warp cell data path
An Example Modern Systolic Array: TPU (I)

Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

As reading a large SRAM uses much more power than arithmetic, the matrix unit uses systolic execution to save energy by reducing reads and writes of the Unified Buffer [Kun80][Ram91][Ovt15b]. Figure 4 shows that data flows in from the left, and the weights are loaded from the top. A given 256-element multiply-accumulate operation moves through the matrix as a diagonal wavefront. The weights are preloaded, and take effect with the advancing wave alongside the first data of a new block. Control and data are pipelined to give the illusion that the 256 inputs are read at once, and that they instantly update one location of each of 256 accumulators. From a correctness perspective, software is unaware of the systolic nature of the matrix unit, but for performance, it does worry about the latency of the unit.

Recall: Example 2D Systolic Array Computation

- Multiply two 3x3 matrices (inputs)
  - Keep the final result in PE accumulators

\[
\begin{bmatrix}
c_{00} & c_{01} & c_{02} \\
c_{10} & c_{11} & c_{12} \\
c_{20} & c_{21} & c_{22}
\end{bmatrix} =
\begin{bmatrix}
a_{00} & a_{01} & a_{02} \\
a_{10} & a_{11} & a_{12} \\
a_{20} & a_{21} & a_{22}
\end{bmatrix} \times
\begin{bmatrix}
b_{00} & b_{01} & b_{02} \\
b_{10} & b_{11} & b_{12} \\
b_{20} & b_{21} & b_{22}
\end{bmatrix}
\]

Figure 1: A systolic array processing element

- \( P = M \)
- \( Q = N \)
- \( R = R + M \times N \)
Figure 1. TPU Block Diagram. The main computation part is the yellow Matrix Multiply unit in the upper right hand corner. Its inputs are the blue Weight FIFO and the blue Unified Buffer (UB) and its output is the blue Accumulators (Acc). The yellow Activation Unit performs the nonlinear functions on the Acc, which go to the UB.
An Example Modern Systolic Array: TPU2

4 TPU chips vs 1 chip in TPU1

High Bandwidth Memory vs DDR3

Floating point operations vs FP16

45 TFLOPS per chip vs 23 TOPS

Designed for training and inference vs only inference

An Example Modern Systolic Array: TPU3

- 32GB HBM per chip vs 16GB HBM in TPU2
- 4 Matrix Units per chip vs 2 Matrix Units in TPU2
- 90 TFLOPS per chip vs 45 TFLOPS in TPU2

https://cloud.google.com/tpu/docs/system-architecture
Cerebras’s Wafer Scale Engine (2019)

- The largest ML accelerator chip
- 400,000 cores

Cerebras WSE
1.2 Trillion transistors
46,225 mm²

Largest GPU
21.1 Billion transistors
815 mm²

https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning
https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning
Cerebras’s Wafer Scale Engine-2 (2021)

- The largest ML accelerator chip
- 850,000 cores

Cerebras WSE-2
2.6 Trillion transistors
46,225 mm²

Largest GPU
54.2 Billion transistors
826 mm²

https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning
https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning