Announcements (I)

- Practice exercises I solutions
  - To be posted
  - Remember that these are to facilitate your learning. They are not example exam questions.

- Practice exercises II
  - To be posted

- Example exam questions
  - To be posted

- Keep monitoring the course website for these
Next Week

- Our last week
- June 1/2: Review session
- Exact date depends on what we cover today
Announcement

If you are interested in learning more and doing research in Computer Architecture, three suggestions:

- Email me with your interest
- Take the “Computer Architecture” course in the Fall
- Do readings and assignments on your own

There are many exciting projects and research positions available, spanning:

- Memory systems
- GPUs, FPGAs, heterogeneous systems, ...
- New execution paradigms (e.g., in-memory computing)
- Security-architecture-reliability-energy-performance interactions
- Architectures for medical/health/genomics
- ...


We Are Almost Done With This...

- 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
Other Approaches to Concurrency (or Instruction Level Parallelism)
Approaches to (Instruction-Level) Concurrency

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


Memory: heart
PEs: cells
Memory pulses data through cells

Figure 1. Basic principle of a systolic system.
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

\[
\text{Given the sequence of weights } \{w_1, w_2, \ldots, w_k\} \\
\text{and the input sequence } \{x_1, x_2, \ldots, x_n\}, \\
\text{compute the result sequence } \{y_1, y_2, \ldots, y_{n+1-k}\} \\
\text{defined by} \\
\begin{align*}
y_i &= w_1 x_i + w_2 x_{i+1} + \cdots + w_k x_{i+k-1}
\end{align*}
\]
Systolic Computation Example: Convolution

- $y_1 = w_1x_1 + w_2x_2 + w_3x_3$
- $y_2 = w_1x_2 + w_2x_3 + w_3x_4$
- $y_3 = w_1x_3 + w_2x_4 + w_3x_5$

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
Systolic Computation Example: Convolution

- 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
Systolic Arrays: Pros and Cons

- **Advantage:**
  - 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

```plaintext
loop {
    Compute1
    Compute2
    Compute3
}
```
Figure 3. File compression algorithm executed using pipeline parallelism
Systolic Array: Advantages & Disadvantages

- Advantages
  - Makes multiple uses of each data item $\rightarrow$ reduced need for fetching/refetching
  - High concurrency
  - Regular design (both data and control flow)

- Disadvantages
  - Not good at exploiting irregular parallelism
  - Relatively special purpose $\rightarrow$ 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

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.

An Example Modern Systolic Array

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.

An Example Modern Systolic Array

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.
Decoupled Access/Execute (DAE)
Decoupled Access/Execute (DAE)

- Motivation: Tomasulo’s algorithm too complex to implement
  - 1980s before Pentium Pro

- Idea: Decouple operand access and execution via two separate instruction streams that communicate via ISA-visible queues.

Decoupled Access/Execute (II)

- Compiler generates two instruction streams (A and E)
  - Synchronizes the two upon control flow instructions (using branch queues)

```
q = 0.0
Do 1  k = 1, 400
  x(k) = q + y(k) * (r * z(k+10) + t * z(k+11))
```

Fig. 2a. Lawrence Livermore Loop 1 (HYDRO EXCERPT)

<table>
<thead>
<tr>
<th>Access</th>
<th>Execute</th>
</tr>
</thead>
<tbody>
<tr>
<td>A7 = -400</td>
<td>AEQ + z + 10, A2</td>
</tr>
<tr>
<td>A2 = 0</td>
<td>AEQ + z + 11, A2</td>
</tr>
<tr>
<td>A3 = 1</td>
<td>AEQ + y, A2</td>
</tr>
<tr>
<td>X2 = r</td>
<td>A7 + A7 + 1</td>
</tr>
<tr>
<td>X5 + t</td>
<td>x, A2 + EAQ</td>
</tr>
<tr>
<td>loop: X3 + z + 10, A2</td>
<td>A2 + A2 + A3</td>
</tr>
<tr>
<td>X7 + z + 11, A2</td>
<td></td>
</tr>
<tr>
<td>X4 + X2 *f X3</td>
<td></td>
</tr>
<tr>
<td>X3 + X5 *f X7</td>
<td></td>
</tr>
<tr>
<td>X7 + y, A2</td>
<td></td>
</tr>
<tr>
<td>X6 + X3 +f X4</td>
<td></td>
</tr>
<tr>
<td>X4 + X7 *f X6</td>
<td></td>
</tr>
<tr>
<td>A7 + A7 + 1</td>
<td></td>
</tr>
<tr>
<td>x, A2 + X4</td>
<td></td>
</tr>
<tr>
<td>A2 + A2 + A3</td>
<td>Branch if A7 &lt; 0</td>
</tr>
</tbody>
</table>

Fig. 2b. Compilation onto CRAY-1-like architecture

Fig. 2c. Access and execute programs for straight-line section of loop
Decoupled Access/Execute (III)

- Advantages:
  + Execute stream can run ahead of the access stream and vice versa
  + If A takes a cache miss, E can perform useful work
  + If A hits in cache, it supplies data to lagging E
  + Queues reduce the number of required registers
  + Limited out-of-order execution without wakeup/select complexity

- Disadvantages:
  -- Compiler support to partition the program and manage queues
    -- Determines the amount of decoupling
  -- Branch instructions require synchronization between A and E
  -- Multiple instruction streams (can be done with a single one, though)
Astronautics ZS-1

- Single stream steered into A and X pipelines
- Each pipeline in-order
Loop Unrolling to Eliminate Branches

- **Idea:** *Replicate loop body multiple times within an iteration*
  - + Reduces loop maintenance overhead
    -   - Induction variable increment or loop condition test
  - + Enlarges basic block (and analysis scope)
    -   - Enables code optimization and scheduling opportunities

-- What if iteration count not a multiple of unroll factor? (need extra code to detect this)
-- Increases code size
A Modern DAE Example: Pentium 4

Intel Pentium 4 Simplified

Approaches to (Instruction-Level) Concurrency

- Pipelining
- Out-of-order execution
- Dataflow (at the ISA level)
- Superscalar Execution
- VLIW
- SIMD Processing (Vector and array processors, GPUs)
- Decoupled Access Execute
- Systolic Arrays
We Are Now Done With This…

- 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
An Example GPU Exercise

We define the *SIMD utilization* of a program run on a GPU as the fraction of SIMD lanes that are kept busy with *active threads* during the run of a program.

The following code segment is run on a GPU. Each thread executes a *single iteration* of the shown loop. Assume that the data values of the arrays A, B, and C are already in vector registers so there are no loads and stores in this program. (Hint: Notice that there are 4 instructions in each thread.) A warp in the GPU consists of 64 threads, and there are 64 SIMD lanes in the GPU.

```c
for (i=0; i<1,024,768; i++) {
    if (A[i] > 0) {
        A[i] = A[i] * C[i];
        B[i] = A[i] + B[i];
        C[i] = B[i] + 1;
    }
}
```

(a) How many warps does it take to execute this program?
(b) When we measure the SIMD utilization for this program with one input set, we find that it is \(67/256\). What can you say about arrays A, B, and C? Be precise.

A:

B:

C:
An Example GPU Exercise (III)

(c) Is it possible for this program to yield a SIMD utilization of 100% (circle one)?

YES   NO

If YES, what should be true about arrays A, B, C for the SIMD utilization to be 100%? Be precise.

A:

B:

C:

If NO, explain why not.
(d) Is it possible for this program to yield a SIMD utilization of 25% (circle one)?

YES       NO

If YES, what should be true about arrays A, B, and C for the SIMD utilization to be 25%? Be precise.

A:

B:

C:

If NO, explain why not.
Approaches to (Instruction-Level) Concurrency

- Pipelining
- Out-of-order execution
- Dataflow (at the ISA level)
- Superscalar Execution
- VLIW
- SIMD Processing (Vector and array processors, GPUs)
- Decoupled Access Execute
- Systolic Arrays
We Are Now Done With This…

- 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