6.823 Computer System Architecture

Instructor: Daniel Sanchez
TA: Hyun Ryong (Ryan) Lee

The processor you built in 6.004*

What you’ll understand after taking 6.823
Computing devices then...
Computing devices now
A journey through this space

• What do computer architects actually do?

• Illustrate via historical examples
  – Early days: ENIAC, EDVAC, and EDSAC
  – Arrival of IBM 650 and then IBM 360
  – Seymour Cray – CDC 6600, Cray 1
  – Microprocessors and PCs
  – Multicores
  – Cell phones

• Focus on ideas, mechanisms, and principles, especially those that have withstood the test of time
Abstraction layers

Original domain of the computer architect (‘50s–‘80s)

Application
Algorithm
Programming Language
Operating System/Virtual Machine
Instruction Set Architecture (ISA)
Microarchitecture
Register-Transfer Level (RTL)
Circuits
Devices
Physics

Parallel computing, specialization, security, ...

Domain of computer architecture (‘90s)

Reliability, power

Expansion of computer architecture, mid-2000s onward.
Computer Architecture is the design of abstraction layers

• What do abstraction layers provide?
  – Environmental stability within generation
  – Environmental stability across generations
  – Consistency across a large number of units

• What are the consequences?
  – *Encouragement to create reusable foundations:*
    • *Toolchains, operating systems, libraries*
  – Enticement for application innovation
Technology is the dominant factor in computer design

Technology
- Transistors
- Integrated circuits
- VLSI (initially)
- Flash memories, ...

Technology
- Core memories
- Magnetic tapes
- Disks

Technology
- ROMs, RAMs
- VLSI
- Packaging
- Low Power
But Software…

As people write programs and use computers, our understanding of *programming* and *program behavior* improves.

*This has profound though slower impact on computer architecture*

Modern architects must pay attention to software and compilation issues.
Architecture is engineering design under constraints

Factors to consider:
• Performance of whole system on target applications
  – Average case & worst case
• Cost of manufacturing chips and supporting system
• Power to run system
  – Peak power & energy per operation
• Reliability of system
  – Soft errors & hard errors
• Cost to design chips (engineers, computers, CAD tools)
  – Becoming a limiting factor in many situations, fewer unique chips can be justified
• Cost to develop applications and system software
  – Often the dominant constraint for any programmable device

At different times, and for different applications at the same point in time, the relative balance of these factors can result in widely varying architectural choices
Course Information

All info kept up to date on the website:
http://www.csg.csail.mit.edu/6.823
Contact times

- Lectures Tuesdays and Thursdays
  - 1:00pm to 2:30pm

- Tutorial on Fridays
  - 1:00pm to 2:00pm
  - Attendance is optional
  - Additional tutorials will be held in evenings before quizzes

- Quizzes on Friday (*except last quiz*)
  - 1:00pm to 2:30pm
  - Attendance is NOT optional

- Instructor office hours
  - After class or by email appointment

- TA office hours
  - Wednesday 4-5:30pm
Lectures and tutorials

• Lectures/tutorials are synchronous, through Zoom
  – Video recordings will be available on the website

• Two ways of asking questions:
  – Unmute yourself and ask – for questions specific to lecture
  – Zoom chat – for relevant but less direct questions

• If you can, please enable video 😊
  – Helps you stay engaged, helps us get to know you and get nonverbal feedback like in an in-person lecture
  – Your video won’t appear on recordings, and recordings won’t be publicly available
Online resources & help

• We use Piazza extensively
  – Fastest way to get your questions answered
  – Links to lecture & tutorial videos will be posted on Piazza
  – All course announcements are made on Piazza

• This is not a normal term; if you need help, let us know!
  – We can be accommodating
The course has four modules

<table>
<thead>
<tr>
<th>Module 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Instruction Set Architecture (ISA)</td>
</tr>
<tr>
<td>- Caches and Virtual Memory</td>
</tr>
<tr>
<td>- Simple Pipelining and Hazards</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Module 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Complex Pipelining and Out of Order Execution</td>
</tr>
<tr>
<td>- Branch Prediction and Speculative Execution</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Module 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Multithreading and Multiprocessors</td>
</tr>
<tr>
<td>- Coherence and consistency</td>
</tr>
<tr>
<td>- On-chip networks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Module 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>- VLIW, EPIC</td>
</tr>
<tr>
<td>- Vector machines and GPUs</td>
</tr>
</tbody>
</table>
Textbook and readings

- “Computer Architecture: A Quantitative Approach”, Hennessy & Patterson, 5\textsuperscript{th} / 6\textsuperscript{th} ed.
  - 5\textsuperscript{th} edition available online through MIT Libraries
  - Recommended, but not necessary

- Course website lists H&P reading material for each lecture, and optional readings that provide more in-depth coverage
Grading

• Grades are not assigned based on a predetermined curve
  – Most of you are capable of getting an A
• 75% of the grade is based on four closed book 1.5 hour quizzes
  – The first three quizzes will be held during the tutorials; the last one during the last lecture (dates on web syllabus)
  – We’ll have distant-timezone quizzes and makeups if needed
• 25% of the grade is based on four laboratory exercises
• No final exam
• No final project
Problem sets & labs

- **Problem sets**
  - One problem set per module, not graded
  - Intended for private study and for tutorials to help prepare for quizzes
  - Quizzes assume you are very familiar with the content of problem sets

- **Labs**
  - Four graded labs (Lab 0 is introductory)
  - Based on widely-used PIN tool
  - Labs 2 and 4 are open-ended challenges

- **You must complete labs & quizzes individually**
  - Please review the collaboration & academic honesty policy
Self evaluation take-home quiz

• Goal is to help you judge for yourself whether you have prerequisites for this class, and to help refresh your memory
• We assume that you understand digital logic, a simple 5-stage pipeline, and simple caches
• Please work by yourself on this quiz – not in groups
• Remember to complete self-evaluation section at end of the quiz
• Due by Friday (on recitation or send answers to TA mailing list)

Please email us if you have concerns about your ability to take the class
Early Developments: From ENIAC to the mid 50’s
Prehistory

• 1800s: Charles Babbage
  – Difference Engine (conceived in 1823, first implemented in 1855 by Scheutz)
  – Analytic Engine, the first conception of a general purpose computer (1833, never implemented)

• 1890: Tabulating machines

• Early 1900s: Analog computers

• 1930s: Early electronic (fixed-function) digital computers
Electronic Numerical Integrator and Computer (ENIAC)

- Designed and built by Eckert and Mauchly at the University of Pennsylvania during 1943-45
- The first, completely electronic, operational, general-purpose analytical calculator!
  - 30 tons, 72 square meters, 200KW
- Performance
  - Read in 120 cards per minute
  - Addition took 200 μs, Division 6 ms
- Not very reliable!

**Application:** Ballistic calculations

angle = f (location, tail wind, cross wind, air density, temperature, weight of shell, propellant charge, ... )
Electronic Discrete Variable Automatic Computer (EDVAC)

- ENIAC’s programming system was external
  - Sequences of instructions were executed independently of the results of the calculation
  - Human intervention required to take instructions “out of order”

- EDVAC was designed by Eckert, Mauchly, and von Neumann in 1944 to solve this problem
  - Solution was the stored program computer
    \[ \Rightarrow \text{“program can be manipulated as data”} \]

- First Draft of a report on EDVAC was published in 1945, but just had von Neumann’s signature!
  - Without a doubt the most influential paper in computer architecture
Stored Program Computer

Program = A sequence of instructions

How to control instruction sequencing?

manual control

automatic control
  external (paper tape)  
  internal
    plug board
    read-only memory
    read-write memory

– The same storage can be used to store program and data

EDSAC 1950 Maurice Wilkes

Harvard Mark I, 1944
Zuse’s Z1, WW2

ENIAC 1946
ENIAC 1948
EDVAC 1947 (concept)
The Spread of Ideas

ENIAC & EDVAC had immediate impact

*brilliant engineering:* Eckert & Mauchly
*lucid paper:* Burks, Goldstein & von Neumann

<table>
<thead>
<tr>
<th>Location</th>
<th>Institution</th>
<th>Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAS</td>
<td>Princeton</td>
<td>46-52</td>
</tr>
<tr>
<td>EDSAC</td>
<td>Cambridge</td>
<td>46-50</td>
</tr>
<tr>
<td>MANIAC</td>
<td>Los Alamos</td>
<td>49-52</td>
</tr>
<tr>
<td>JOHNIAC</td>
<td>Rand</td>
<td>50-53</td>
</tr>
<tr>
<td>ILLIAC</td>
<td>Illinois</td>
<td>49-52</td>
</tr>
<tr>
<td></td>
<td>Argonne</td>
<td>49-53</td>
</tr>
<tr>
<td>SWAC</td>
<td>UCLA-NBS</td>
<td></td>
</tr>
</tbody>
</table>

**UNIVAC** - the first commercial computer, 1951

*Alan Turing’s direct influence on these developments is often debated by historians.*
Dominant Technology Issue:  
**Reliability**

ENIAC  \[\Rightarrow\]  EDVAC

- 18,000 tubes
- 20 10-digit numbers

- 4,000 tubes
- 2000 word storage
- mercury delay lines

Mean time between failures (MTBF)

*MIT’s Whirlwind with an MTBF of 20 min. was perhaps the most reliable machine!*

Reasons for unreliability:

1. Vacuum tubes
2. Storage medium
   - Acoustic delay lines
   - Mercury delay lines
   - Williams tubes
   - Selections

CORE  J. Forrester  1954
Computers in the mid 50’s

- Hardware was expensive
- Stores were small (1000 words)
  ⇒ No resident system-software!
- Memory access time was 10 to 50 times slower than the processor cycle
  ⇒ Instruction execution time was totally dominated by the memory reference time
- The ability to design complex control circuits to execute an instruction was the central design concern as opposed to the speed of decoding or an ALU operation
- Programmer’s view of the machine was inseparable from the actual hardware implementation
Accumulator-based computing

- **Single Accumulator**
  - Calculator design carried over to computers

Why?

Registers expensive
### The Earliest Instruction Sets

**Burks, Goldstein & von Neumann ~1946**

- **LOAD**
  - $x$
  - $AC \leftarrow M[x]$

- **STORE**
  - $x$
  - $M[x] \leftarrow (AC)$

- **ADD**
  - $x$
  - $AC \leftarrow (AC) + M[x]$

- **SUB**
  - $x$

- **MUL**
  - $x$
  - Involved a quotient register

- **DIV**
  - $x$

- **SHIFT LEFT**
  - $x$
  - $AC \leftarrow 2 \times (AC)$

- **SHIFT RIGHT**

- **JUMP**
  - $x$
  - $PC \leftarrow x$

- **JGE**
  - $x$
  - if $(AC) \geq 0$ then $PC \leftarrow x$

- **LOAD ADR**
  - $x$
  - $AC \leftarrow \text{Extract address field}(M[x])$

- **STORE ADR**
  - $x$

**Typically less than 2 dozen instructions!**
Programming: Single Accumulator Machine

\[ C_i \leftarrow A_i + B_i, \quad 1 \leq i \leq n \]

<table>
<thead>
<tr>
<th>LOOP</th>
<th>LOAD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>JGE</td>
<td>DONE</td>
<td></td>
</tr>
<tr>
<td>ADD</td>
<td>ONE</td>
<td></td>
</tr>
<tr>
<td>STORE</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>F1</td>
<td>LOAD</td>
<td>A</td>
</tr>
<tr>
<td>F2</td>
<td>ADD</td>
<td>B</td>
</tr>
<tr>
<td>F3</td>
<td>STORE</td>
<td>C</td>
</tr>
<tr>
<td>JUMP</td>
<td>LOOP</td>
<td></td>
</tr>
<tr>
<td>DONE</td>
<td>HLT</td>
<td></td>
</tr>
</tbody>
</table>

Problem?

How to modify the addresses A, B and C?
Self-Modifying Code

Each iteration involves total bookkeeping: instruction fetches, operand fetches, stores.

Most of the executed instructions are for bookkeeping.

\[ C_i \leftarrow A_i + B_i, \quad 1 \leq i \leq n \]

modify the program for the next iteration

LOO...

LOAD LOAD N
JGE N
ADD N
STORE DONE
ONE

F1 LOAD F1
ADD A
STORE F1
A

F2 ADD F2
STORE B
B

F3 STORE F3
C

LOAD ADR F1
ADD ONE
STORE ADR F1

LOAD ADR F2
ADD ONE
STORE ADR F2

LOAD ADR F3
ADD ONE
STORE ADR F3

JUMP LOOP
DONE HLT

February 16, 2021
MIT 6.823 Spring 2021
L01-30
Processor-Memory Bottleneck: Early Solutions

- **Indexing capability**
  - to reduce bookkeeping instructions

- **Fast local storage in the processor**
  - 8-16 registers as opposed to one accumulator
  - to reduce loads/stores

- **Complex instructions**
  - to reduce instruction fetches

- **Compact instructions**
  - implicit address bits for operands
  - to reduce instruction fetch cost
Index Registers
Tom Kilburn, Manchester University, mid 50’s

One or more specialized registers to simplify address calculation

Modify existing instructions
- LOAD x, IX AC ← M[x + (IX)]
- ADD x, IX AC ← (AC) + M[x + (IX)]
...

Add new instructions to manipulate index registers
- JZi x, IX if (IX)=0 then PC ← x
  else IX ← (IX) + 1
- LOADi x, IX IX ← M[x] (truncated to fit IX)
...

Index registers have accumulator-like characteristics
Using Index Registers

\[ C_i \leftarrow A_i + B_i, \quad 1 \leq i \leq n \]

<table>
<thead>
<tr>
<th>LOAD ( i )</th>
<th>( N, IX )</th>
</tr>
</thead>
<tbody>
<tr>
<td>JZ ( i )</td>
<td>( DONE, IX )</td>
</tr>
<tr>
<td>LOAD</td>
<td>( LASTA, IX )</td>
</tr>
<tr>
<td>ADD</td>
<td>( LASTB, IX )</td>
</tr>
<tr>
<td>STORE</td>
<td>( LASTC, IX )</td>
</tr>
<tr>
<td>JUMP</td>
<td>( LOOP )</td>
</tr>
<tr>
<td>DONE</td>
<td>( HALT )</td>
</tr>
</tbody>
</table>

- **Program does not modify itself**
- **Efficiency has improved dramatically (ops / iter)**
  
<table>
<thead>
<tr>
<th>Instruction fetch</th>
<th>with index regs</th>
<th>without index regs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>17 (14)</td>
<td>10 (8)</td>
</tr>
<tr>
<td>Operand fetch</td>
<td></td>
<td>5 (4)</td>
</tr>
<tr>
<td>Store</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Costs?**

\( N \) starts with \(-n\)
Operations on Index Registers

To increment index register by $k$

\[
\begin{align*}
AC & \leftarrow (IX) \quad \text{new instruction} \\
AC & \leftarrow (AC) + k \\
IX & \leftarrow (AC) \quad \text{new instruction}
\end{align*}
\]

also the AC must be saved and restored

It may be better to increment IX directly

\[
\text{INCI} \quad k, \; IX \quad IX \leftarrow (IX) + k
\]

More instructions to manipulate index register

\[
\text{STOREi} \quad x, \; IX \quad M[x] \leftarrow (IX) \quad \text{(extended to fit a word)}
\]

\[
\text{IX begins to look like an accumulator}
\]

$\Rightarrow$ several index registers

$\Rightarrow$ several accumulators

$\Rightarrow$ General Purpose Registers
Evolution of Addressing Modes

1. Single accumulator, absolute address
   LOAD x

2. Single accumulator, index registers
   LOAD x, IX

3. Indirection
   LOAD (x)

4. Multiple accumulators, index registers, indirection
   LOAD R, IX, x
   or LOAD R, IX, (x)
   the meaning?
   \[ R \leftarrow M[M[x] + (IX)] \]
   or \[ R \leftarrow M[M[x + (IX)]] \]

5. Indirect through registers
   LOAD R_I, (R_J)

6. The works
   LOAD R_I, R_J, (R_K)
   \[ R_J = \text{index}, \ R_K = \text{base addr} \]
Variety of Instruction Formats

- **Three address formats**: One destination and up to two operand sources per instruction

  \[
  \begin{align*}
  &\text{(Reg op Reg) to Reg} \quad R_I \leftarrow (R_J) \text{ op } (R_K) \\
  &\text{(Reg op Mem) to Reg} \quad R_I \leftarrow (R_J) \text{ op } M[x]
  \end{align*}
  \]

  - x can be specified directly or via a register
  - effective address calculation for x could include indexing, indirection, ...

- **Two address formats**: the destination is same as one of the operand sources

  \[
  \begin{align*}
  &\text{(Reg op Reg) to Reg} \quad R_I \leftarrow (R_I) \text{ op } (R_J) \\
  &\text{(Reg op Mem) to Reg} \quad R_I \leftarrow (R_I) \text{ op } M[x]
  \end{align*}
  \]
More Instruction Formats

• **One address formats:** Accumulator machines
  - Accumulator is always other implicit operand

• **Zero address formats:** operands on a stack
  - Stack can be in registers or in memory
    - usually top of stack cached in registers

```
load      M[sp] ← M[M[sp]]
```

Many different formats are possible!
Instruction sets in the mid 50’s

- Great variety of instruction sets, but all intimately tied to implementation details

- Programmer’s view of the machine was inseparable from the actual hardware implementation!

Next Lecture:
Instruction Set Architectures: Decoupling Interface and Implementation