Advanced Computing for Engineering Applications

Dan Negrut

Simulation-Based Engineering Lab
Wisconsin Applied Computing Center
Department of Mechanical Engineering
Department of Electrical and Computer Engineering
University of Wisconsin-Madison

Darmstadt University
August 29-September 02, 2016
A Couple of Things About Myself

- Bucharest Polytechnic University, Romania
  - B.S. – Aerospace Engineering (1992)

- University of Iowa
  - Ph.D. – Mechanical Engineering (1998)

- MSC.Software
  - Software Engineer 1998-2005

- University of Michigan, Ann Arbor
  - Adjunct Assistant Professor, Dept. of Mathematics (2004)

- DOE’s Argonne National Laboratory, Division of Mathematics and Computer Science

- University of Wisconsin-Madison, since Nov. 2005
  - Professor, Mechanical Engineering & Electrical and Computer Engineering
  - Research Focus: Computational Dynamics (Dynamics of Multi-body Systems)
  - Technical lead, Simulation-Based Engineering Lab (http://sbel.wisc.edu)
  - Co-Director, Wisconsin Applied Computing Center
Before We Get Started…

- **Goal**
  - Spend five days getting familiar with parallel computing

- **Looking ahead**
  - Cover some basics about computing at large (day 1)
  - Discuss general concepts related to parallel computing (day 2, only ½ day)
  - Discuss parallel computing on GPU cards (days 2 - 3)
  - Parallel computing with OpenMP (day 4)
  - Parallel computing with MPI (day 5)

- **The material perhaps outside your comfort zone**
  - Take it as an opportunity to break into something new
This Course: What It Is; What It Is Not

- You won’t master parallel computing in five days, but
- You will be exposed to concepts of parallel computing

- You’ll probably gain a better understanding of computing
  - Sequential and parallel

- You’ll probably be in a position to understand how to use parallel computing to solve your problem
  - To actually solve your problem you’ll require further digging

- You will understand that getting performance out of the hardware you work with requires a good understanding of how it works
Pointers for Information

- Slides emailed to you and also made available at (at end of the course) at
  http://outreach.sbel.wisc.edu/Workshops/GPUworkshop/

- This material is part of an HPC class I teach at University of Wisconsin-Madison
  - GPU Computing, OpenMP, and MPI
  - Class material available online (slides & audio streaming): http://sbel.wisc.edu/Courses/ME964/2015/

- Today’s material is based on information available here:
  http://sbel.wisc.edu/Courses/ME964/Literature/primerHW-SWinterface.pdf
My Only Recommendation
[at the beginning of five day journey]

- Ask questions
  - To understand better
  - To keep this interactive and more interesting
Today’s Computer

- Follows paradigm formalized by von Neumann in late 1940s

- The von Neumann model:
  - There is no distinction between data and instructions
  - Data and instructions are stored in memory as a string of 0 and 1 bits
    - Instructions are fetched + decoded + executed
    - Data is used to produce results according to rules specified by the instructions
From Code to Machine Instructions

- There is a difference between a line of code and a machine instruction
- Example:
  - Line of C code:
    \[
    a[4] = \text{delta} + a[3]; \quad // \text{line of C code}
    \]

- MIPS assembly code generated by the compiler (three instructions):
  
  ```
  lw $t0, 12($s2)  # reg $t0 gets value stored 12 bytes from address in $s2
  add$t0, $s4, $t0  # reg $t0 gets the sum of values stored in $s4 and $t0
  sw $t0, 16($s2)  # store at 16 bytes from address in $s2 what’s in $t0
  ```

- Set of three corresponding MIPS instructions produced by the compiler:
  
  ```
  100011100100100000000000000001100
  00000010100010001000000000100000
  10101110010010000000000000010000
  ```
From Code to Instructions

- C code – what you write to implement an algorithm
- Assembly code – intermediate step for compiler, something that humans can read
- Machine code/Instructions – what the assembly code gets translated into by the compiler and the CU understands

- Machine code: what you see in an editor like notepad or vim or emacs if you open up an executable file
- There is a one-to-one correspondence between a line of assembly code and an instruction (most of the time)
Observations:

- The compiler typically goes from C code directly to machine instructions

- People used to write assembly code

- Today coding in assembly done only for critical parts of a program by people who want to highly optimize the execution and don’t trust the compiler
Instruction Set Architecture (ISA)

- Same line of C code can lead to different sets of instructions on different computers

- This is so because two CPUs might draw on two different Instruction Set Architectures (ISA)

- ISA: defines a “vocabulary” used to express at a very low level the actions of a processor
  - ISA: the set of words that can be used to tell the CU what to do

- Example:
  - Very many embedded chips: use MIPS, which is a special version of a RISC ISA
  - Most consumer computers: use Intel chips drawing on the x86 ISA, which is a CISC ISA
Example: the same C code leads to different assembly code (and different set of machine instructions, not shown here)
RISC vs. CISC: Characteristics

- **RISC Architecture – Reduced Instruction Set Computing Architecture**
  - Usually each instruction is coded into a set of 32 bits
  - Moved to 64 bits about three years ago
  - Each executable has fixed length instruction, be it 32 or 64 (no mixing)
    - The length of the instruction is fixed
  - Promoted by: ARM Holding, company that started as ARM (Advanced RISC Machines)
    - Used in: embedded systems, smart phones – Intel, NVIDIA, Samsung, Qualcomm, Texas Instruments
    - Somewhere between 8 and 10 billion chips based on ARM manufactured annually

- **CISC Architecture – Complex Instruction Set Computing Architecture**
  - Instructions have various lengths
    - Examples: 32 bit instruction followed by 512 bit instruction followed later on by 128 bit instruction, etc.
  - Intel’s X86 is the most common example
  - Promoted by Intel and subsequently embraced and augmented by AMD
    - Used in: laptops, desktops, workstations, supercomputers
RISC vs. CISC: Pros/Cons

- RISC is simpler to comprehend, provision for, and work with
- Decoding CISC instructions is not trivial and eats up power
- A CISC instruction is usually broken down into several micro-operations (uops)
- CISC invites spaghetti type evolution of the ISA and require complex microarchitecture
  - Provide the freedom to do as you wish
“Instruction”: A Relevant Concept

- Concept of “instruction” comes into play when figuring out execution speed

- Execution speed depends on
  - The average number of instructions performed per cycle, and
  - The frequency of the processor
The FDX Cycle

- FDX stands for Fetch-Decode-Execute
- Running FDX cycle after cycle is what keeps the CU busy
  - The CU does a FDX for instruction after instruction until program completes

- Fetch: an instruction is fetched from memory
  - Recall that it will look like this (on 32 bits, MIPS, `lw $t0, 12($s2)`):
    10001110010010000000000000001100

- Decode: this strings of 1s and 0s are decoded by the CU
  - Example: here’s an “I” (eye) type instruction, made up of four fields
MIPS Instruction Types

- Arithmetic/logical/shift/comparison
- Control instructions (branch and jump)
- Load/store
- Other (exception, register movement to/from GP registers, etc.)
Instructions Formats

- Three types of instructions in MIPS ISA
  - Type I
  - Type R
  - Type J
Type I (MIPS ISA)
[I comes from “Immediate”]

- The first six bits encode the basic operation; i.e., the opcode, that needs to be completed
  - Examples: adding two numbers (000000), subtracting two numbers (000001), dividing two numbers (000011), etc.

- The next group of five bits indicates in which register the first operand is stored

- The subsequent group of five bits indicates the destination register

- The 16 bit “immediate” value, usually used as the offset value in various instructions
  - “Immediate” means that there is no need to read other registers or jump through other hoops. What you need is right there and you immediately can use it
Type R (MIPS ISA)

- Type R has the same first three fields op, rs, rt like I-type

- Packs three additional fields:
  - Five bit rd field (register destination)
  - Five bit shamt field (shift amount)
  - Six bit funct field, which is a function code that further qualifies the opcode
Instruction Set Architecture vs. Chip Microarchitecture

- ISA – can be regarded as defining a vocabulary
  - Specifies what a processor should be able to do
    - Load, store, jump on less than, etc.

- Microarchitecture – how the silicon is organized to implement the vocabulary promised by ISA

- Example:
  - Intel and AMD both use the x86 ISA yet they have different microarchitectures
The CPU’s Control Unit (CU)

- The CU/instruction interplay
  - An order comes in (this is an instruction) that the CU should handle
  - Some ingredients are needed: meat, pasta and ketchup (this is the data)
    - These ingredients – they are in the kitchen, in some registers
  - Some ready to eat product goes out the kitchen: spaghetti w/ meatballs
    - Note: what comes out of an order might not be a meal, but only an ingredient

- Bringing in the meat, bringing in the pasta, placing them in the proximity (the registers), mixing them in a certain way (op), happens in a coordinated fashion (based on a kitchen clock) that is managed by the CU

- CU manages/coordinates/controls based on the food order (the instruction)
FDX Cycle – The Execution Part: It All Boils Down to Transistors…

- How does this magic happen?

- Transistors can be organized to produce complex logical units that have the ability to execute instructions

- More transistors increase opportunities for building/implementing in silicon functional units that can operate at the same time towards a shared goal
Transistors and then some more transistors

- First discussion about how transistors:
  - Their role in implementing operations (perform tasks)

- Later discussion about transistors
  - How they are used to store data and instructions
Transistors at Work: AND, OR, NOT

- The NOT logical op. is implemented using one transistor
- AND and OR logical ops require two transistors
- Truth tables for AND, OR, and NOT

<table>
<thead>
<tr>
<th>( \text{AND} )</th>
<th>( \text{in}_2=0 )</th>
<th>( \text{in}_1=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{in}_1=0 )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \text{in}_1=1 )</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \text{OR} )</th>
<th>( \text{in}_2=0 )</th>
<th>( \text{in}_2=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{in}_1=0 )</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( \text{in}_1=1 )</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \text{NOT} )</th>
<th>( \text{in}_1=0 )</th>
<th>( \text{in}_1=1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{in}_1=0 )</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>( \text{in}_1=1 )</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Example

- Design a digital logic block that receives three inputs via three bus wires and produces one signal that is 0 (low voltage) as soon as one of the three input signals is low voltage.
- In other words, it should return 1 if and only if all three inputs are 1

<table>
<thead>
<tr>
<th>in₁</th>
<th>in₂</th>
<th>in₃</th>
<th>Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Logic Equation: \( out = \overline{in_1} + in_2 \cdot in_3 \)

- **Overbar** \( \overline{\cdot} \): negation
- **Plus** \( + \): logical OR
- **Product** \( \cdot \): logical AND
Example
[Cntd.]

- Easy to figure out the transistor setup once Logic Equation is available

<table>
<thead>
<tr>
<th>$in_1$</th>
<th>$in_2$</th>
<th>$in_3$</th>
<th>Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Logic Equation:

$$out = in_1 + in_2 \cdot in_3$$

- Solution: digital logic block is a combination of AND, OR, and NOT gates
  - The NOT is represented as a circle O applied to signals moving down the bus
Example: One Bit Adder

- Implement a digital circuit that produces the Carry-out digit in a one bit summation operation

### Truth Table

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Outputs</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i_{\text{n}} )</td>
<td>( i_{\text{n}} )</td>
<td>( \text{CarryIn} )</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Logic Equation:

\[
\text{CarryOut} = (i_{\text{n}} \cdot \text{CarryIn}) + (i_{\text{n}} \cdot \text{CarryIn}) + (i_{\text{n}} \cdot i_{\text{n}})
\]
Integrated Circuits-A One Bit Combo: OR, AND, 1 Bit Adder

- 1 Bit Adder, the **Sum** part

- Combo: OR, AND, 1 Bit Sum
  - Controlled by the input “Operation”
Integrated Circuits: Ripple Design of 32 Bit Combo

- Combine 32 of the 1 bit combos in an array of logic elements
  - Get one 32 bit unit that can do OR, AND, +
Integrated Circuits: From Transistors to CPU

Transistor → Gate → Mux (selector) → Complex Combinational Block → Array of Logic Elements

- Transistor: Take one or combine two of them
- Gate: Combine a couple of them
- Mux (selector): Requires a “Control Signal” as input
- Complex Combinational Block: Combine a couple of them
- Array of Logic Elements: Place instances in an array

Logical Unit

Examples:
- AND, OR, NOT
- Example: \( \text{out} = \text{in}_1 + \text{in}_2 \)
- Example: One bit adder
- Example: 32 bit adder

From simple to complex...
Every 18 months, the number of transistors per unit area doubles (Moore’s Law)

- Current technology (2014-2015): feature length is 10 nm (Intel)
- Next wave (2016-2017): 10 nm (Intel)
- Looking ahead (Intel)
  - 7 nm – 2017-2018
  - 5 nm – 2020-2021

No clear path forward after 2021
- Maybe Carbon Nanotubes?
Number of Transistors, on GPUs

- **NVIDIA Architectures**
  - Fermi circ. 2010:
    - 40 nm technology
    - Up to 3 billion transistors → about 500 scalar processors, 0.5 d.p. Tflops
  
  - Kepler circ. 2012:
    - 28 nm technology
    - Chips w/ 7 billion transistors → about 2800 scalar processors, 1.5 d.p. Tflops
  
  - Maxwell 2015
    - 28 nm technology
    - Chips w/ 8 billion transistors → 3072 scalar processors, 6.1 s.p. Tflops
  
  - Pascal 2016
    - 16 nm technology
    - Chips w/ 15 billion transistors → 3584 scalar processors, 10.1 s.p. Tflops (4.7 d.p. Tflops)
Registers
Registers

- Instruction cycle: fetch-decode-execute (FDX)

- CU – responsible for controlling the process that will deliver the request baked into the instruction

- ALU – does the busy work to fulfill the request put forward by the instruction

- The instruction that is being executed should be stored somewhere

- Fulfilling the requests baked into an instruction usually involves handling input values and generates output values
  - This data needs to be stored somewhere
Registers

- Registers, quick facts:
  - A register is an entity whose role is that of storing information
  - A register is the type of storage with shortest latency – it’s closest to the ALU
  - Typically, one cannot control what gets kept in registers

- The number AND size of registers used are specific to a ISA
  - Prime example of how ISA decides on something and the microarchitecture has to do what it takes to implement this design decision
  - In MIPS ISA: there are 32 registers of 32 bits; and that’s that
Register Types

- Discussion herein covers only several register types typically encountered in a CPU (abbreviation in parenthesis)
  - List not comprehensive, showing only the more important ones

- Instruction register (IR) – a register that holds the instruction that is executed
  - Sometimes known as “current instruction register” CIR

- Program Counter (PC) – a register that holds the address of the next instruction that will be executed
  - NOTE: unlike IR, PC contains an *address* of an instruction, not the actual instruction
Register Types [Cntd.]

- Memory Data Register (MDR) – register that holds data that has been read in from memory or, alternatively, produced by the CPU and waiting to be stored in memory

- Memory Address Register (MAR) – the address of the memory location in memory (RAM) where input/output data is supposed to be read in/written out
  - NOTE: unlike MDR, MAR contains an *address* of a location in memory, not actual data

- Return Address (RA) – the address where upon finishing a sequence of instructions, the execution should jump and commence with the execution of subsequent instruction
Register Types [Cntd.]

- Registers on previous two slides are a staple in most chip designs.
- There are several other registers common to many chip designs yet they are encountered in different numbers.
- Since they come in larger numbers they don’t have an acronym:
  - Registers for Subroutine Arguments (4) – a0 through a3
  - Registers for temporary variables (10) – t0 through t9
  - Registers for saved temporary variables (8) – s0 through s7
    - Saved between function calls
Register Types [Cntd.]

- Several other registers are involved in handling function calls
- Summarized below, but their meaning is only apparent in conjunction with the organization of the virtual memory
  - Global Pointer (gp) – a register that holds an address that points to the middle of a block of memory in the static data segment
  - Stack Pointer (sp) – a register that holds an address that points to the last location on the stack (top of the stack)
  - Frame Pointer (fp) - a register that holds an address that points to the beginning of the procedure frame (for instance, the previous sp before this function changed its value)
Register, Departing Thoughts

- **Examples:**
  - In 32 bit MIPS ISA, there are 32 registers
  - On a GTX580 NVIDIA card there are more than 500,000 32 bit temporary variable registers to keep busy 512 Scalar Processors (SPs) that made up 16 Stream Multiprocessors (SMs)

- Registers are very precious resources

- Increasing their number is not straightforward
  - Need to change the design of the chip (the microarchitecture)
  - Need to work out the control flow
Pipelining
Charlie Chaplin - Modern Times (1936)
Pipelining, or the Assembly Line Concept

- Henry Ford: capitalized and improved the assembly line idea on an industrial scale and in the process shaped the automotive industry (Ford Model T)

- Vehicle assembly line: a good example of a pipelined process
  - The output of one stage (station) becomes the input for the downstream stage (station)
  - It is bad if one station takes too long to produce its output since all the other stations idle a bit at each cycle of the production
  - “cycle” is the time it takes from the moment a station gets its input to the moment it output leaves the station
  - An instruction/vehicle gets executed/manufactured during each cycle
Intro

- FDX cycle: Fetch, Decode, Execute - carried out in conjunction with each instruction

- A closer look at what gets fetched (instructions and data) and then what happens upon execution leads to a generic five stage process associated with an instruction

- “generic” means that in a first order approximation, these five stages can represent any instruction, although some instructions might not have all five stages:
  - Stage 1: Fetch an instruction
  - Stage 2: Decode the instruction while reading registers
  - Stage 3: Data access
  - Stage 4: Execute the operation (Ex.: might be a request to calculate an address)
  - Stage 5: Write-back into register file
Pipelining, Basic Idea

- At the cornerstone of pipelining is the observation that the following tasks can be worked upon simultaneously when processing five instructions:
  - Instruction 1 is in the 5th stage of the FDX cycle
  - Instruction 2 is in the 4th stage of the FDX cycle
  - Instruction 3 is in the 3rd stage of the FDX cycle
  - Instruction 4 is in the 2nd stage of the FDX cycle
  - Instruction 5 is in the 1st stage of the FDX cycle

- The above is a five stage pipeline

- An ideal situation is when each of these stages takes the same amount of time for completion
  - The pipeline is balanced

- What if one stage takes a significantly longer time since it does significantly more than the other stages?
  - Break instruction into two and the length of the pipeline increases by one stage
Example: Streaming for execution 3 SW instructions

sw $t0, 0($s2) # $s2+0 = $t0

sw $t1, 32($s2) # $s2+32 = $t1

sw $t2, 64($s2) # $s2+64 = $t2

- Case 1: No pipelining – 2100 picoseconds [ps]
Example: Streaming for execution 3 SW instructions

\[
\begin{align*}
\text{sw} & \quad $t0, \ 0($s2) \\
\text{sw} & \quad $t1, \ 32($s2) \\
\text{sw} & \quad $t2, \ 64($s2)
\end{align*}
\]

- Case 2: With pipelining – 1200 picoseconds [ps]
Pipelining, Benefits

- Assume that you have
  1. A very large number of instructions
  2. Balanced stages
  3. A pipeline that is deeper than, or equal to, the number “p” of stages associated with the typical ISA instruction

- If 1 through 3 above hold, in a first order approximation, the speed-up you get out of pipelining is approximately “p”

- Benefit stems from parallel processing of FDX stages
  - This kind of parallel processing of stages is transparent to the user
    - Unlike GPU or multicore parallel computing, you don’t have to do anything to benefit of it
Pipelining, Good to Remember

- The amount of time required to complete one stage of the pipeline: one cycle

- Pipelined processor: one instruction processed in each cycle

- Nonpipelined processor: several cycles required to process an instruction:
  - Four cycles for SW, five for LW, four for add, etc.

- Important Remark:
  - Pipelining does not decrease the time to process one instruction but rather it increases the throughput of the processor by overlapping different stages of different instructions
Pipelining Hazards

Q: if deep pipelines are good, why not have them deeper and deeper?

A: deep pipelines plagued by “pipelining hazard”

- These “hazards” come in three flavors
  - Structural hazards
  - Data hazards
  - Control hazards

[Patterson, 4th edition]→ 51
Pipeline Structural Hazards [1/2]

- The instruction pipelining analogy w/ the vehicle assembly line breaks down at the following point:
  - A real world assembly line assembles the same product for a period of time
  - Might be quickly reconfigured to assemble a different product
  - Instruction pipelining must process a broad spectrum of instructions that come one after another
    - Example: A J-type instruction coming after a R-type instruction, which comes after three I-Type instructions
    - If they were the same instructions (vehicles), designing a pipeline (assembly line) is straightforward

- A structural hazard refers to the possibility of having a combination of instructions in the pipeline that are contending for the same piece of hardware
  - Not encountered when you assemble the same car model (things are deterministic in this case)
**Pipeline Structural Hazards [2/2]**

- **Possible Scenario:** you have a six stage pipeline and the instruction in stage 1 and instruction in stage 5 both need to use the same register to store a temporary variable.
  - **Resolution:** there should be enough registers provisioned so that no combination of instructions in the pipeline leads to RAW, WAR, etc. type issue
  - **Alternative solution:** serialize the access, basically stall the pipeline for a cycle so that there is no contention

- **Note:**
  - Adding more registers is a static solution; expensive and very consequential (requires a chip design change)
  - Stalling the pipeline at run time is a dynamic solution that is inexpensive but slows down the execution
Pipeline **Data Hazards** [1/2]

- Consider the following example in a five stage pipeline setup:

```assembly
add  $t0, $t2, $t4  # $t0 = $t2 + $t4
addi $t3, $t0, 16   # $t3 = $t0 + 16 ("add immediate")
```

- The first instruction is processed in five stages

- Its output (value stored in register $t0) is needed in the very next instruction

- Data hazard: unavailability of $t0 to the second instruction, which references this register

- Resolution (less than ideal)
  - Pipeline stalls to wait for the first instruction to fully complete
Pipeline Data Hazards [2/2]

```plaintext
add $t0, $t2, $t4  # $t0 = $t2 + $t4
addi $t3, $t0, 16  # $t3 = $t0 + 16 ("add immediate")
```

- Alternative [the good] Resolution: use “forwarding” or “bypassing”

- Key observation: the value that will eventually be placed in $t0 is available after stage 3 of the pipeline (where the ALU actually computes this value)

- Provide the means for that value in the ALU to be made available to other stages of the pipeline right away
  - Nice thing: avoids stalling - don’t have to wait several other cycles before the value made its way in $t0
  - This process is called a forwarding of the value

- Supporting forwarding does not guarantee resolution of all scenarios
  - In relatively rare occasions the pipeline ends up stalled for a couple of cycles

- Note that the compiler can sometimes help by re-ordering instructions
  - Out of order execution
Pipeline Control Hazards [Setup]

- What happens when there is an “if” statement in a piece of C code?

- A corresponding machine instruction decides the program flow
  - Specifically, should the “if” branch be taken or not?

- Processing this very instruction to figure out the next instruction (branch or no-branch) will take a number of cycles

- Should the pipeline stall while this instruction is fully processed and the branching decision becomes clear?
  - If yes: approach works, but it is slow
  - If no: you rely on branch prediction and proceed fast but cautiously
Pipeline Control Hazards: Branch Prediction

- Note that when you predict wrong you have to discard instruction[s] executed speculatively and take the correct execution path

- Static Branch Prediction (1\textsuperscript{st} strategy out of two):
  - Always predict that the branch will not be taken and schedule accordingly
  - There are other heuristics for proceeding: for instance, for a do-while construct it makes sense to always be jumping back at the beginning of the loop
    - Similar heuristics can be produced in other scenarios (a “for” loop, for instance)

- Dynamic Branch Prediction (2\textsuperscript{nd} strategy out of two):
  - At a branching point, the branch/no-branch decision can change during the life of a program based on recent history
  - In some cases branch prediction accuracy hits 90%
Pipelining vs. Multiple-Issue

- Pipelining should not be confused with “Multiple-Issue” as an alternative way of speeding up execution.
- A Multiple-Issue processor core is capable of processing more than one instruction at each cycle.
- Two examples to show when this might come in handy:
  - Example 1: performing an integer operation while performing a floating point operation – they require different resources and therefore can proceed simultaneously.
  - Example 2: the two lines of C code below lead to a set of instructions that can be executed at the same time:

```c
int a, b;
float c, d;
// some code a, b, c, d
a += b;
c = cos(d);
```
Pipelining vs. Multiple-Issue

- On average, more than one instruction is processed by the same core in the same clock cycle.

- Multiple-Issue can be done statically or dynamically:
  - Static multiple-issue:
    - Predefined, doesn’t change at run time.
    - Who uses it: NVIDIA - very common in parallel computing on the GPU
      - One warp of threads executes in locked-step fashion.
  - Dynamic multiple-issue:
    - Changed at run time by using hardware resources that can take additional work.
    - Who uses it: Intel, uses it heavily.

- NOTE: Both pipelining and multiple-issue are presentations of what is called Instruction-Level Parallelism (ILP).
Attributes of Dynamic Multiple-Issue

- The data dependencies between instruction being processed takes place at run time

- Checking for dependencies is complex, requires high cost in time and energy

- Checks in place to make sure result is ok

- NOTE: sometimes called a superscalar architecture
Measuring Computing Performance
Nomenclature

- **Program Execution Time** – sometimes called *wall clock time*, or elapsed time, or response time
  - Most meaningful indicator of performance
  - Amount of time from the beginning of a program to the end of the program
  - Includes (factors in) all the housekeeping (running other programs, OS tasks, etc.) that the CPU has to do while running the said program

- **CPU Execution Time**
  - Like “Program Execution Time” but counting only the amount of time that is effectively dedicated to the said program
  - Requires a profiling tool to assess (like gprof, for instance)

- On a dedicated machine; i.e., a quiet machine, Program Execution Time and CPU Execution Time would virtually be identical

[Patterson, 4th edition]
Nomenclature [Cntd.]

- Clock cycle, clock, cycle, tick – the length of the period for the processor clock; typically a constant value dictated by the frequency at which the processor operates
  - Example: 2 GHz processor has clock cycle of 500 picoseconds
- Frequency of a chip: the inverse of the clock cycle (measured in Hz)
  - IMPORTANT POINT: the frequency of chips plateaued for the last 5 years

[acm.org]→
The CPU Performance Equation

- The three ingredients of the CPU Performance Equation:
  - Number of instructions that your program executes (Instruction Count)
  - Average number of clock-cycles per instructions (CPI)
  - Clock Cycle Time

- The CPU Performance Equation reads:
  
  \[
  \text{CPU Exec. Time} = \text{Instruction Count} \times \text{CPI}\times \text{Clock Cycle Time}
  \]

- Alternatively, using the clock rate

  \[
  \text{CPU Exec. Time} = \frac{\text{Instruction Count} \times \text{CPI}}{\text{Clock Rate}}
  \]

[Patterson, 4th edition]
CPU Performance: How can we improve it?

- To improve performance the product of three factors should be reduced

- For a long time, we surfed the wave of “let’s increase the frequency”; i.e., reduce clock cycle time
  - We eventually hit a wall this way (the “Power Wall”)

- As repeatedly demonstrated in practice, reducing the Instruction Count (IC) often times leads to an increase in CPI. And the other way around.
  - Ongoing argument: whether RISC or CISC is the better ISA
    - The former is simple and therefore can be optimized easily. Yet it requires a large number of instructions to accomplish something in your C code
    - The latter is very complex but instructions are very expressive. Leads to few but expensive instructions to accomplish something in your C code
    - Specific example: ARM vs. x86
There are benchmarks used to gauge the performance of a processor

Idea: gather a collection of programs that use a good mix of instructions and flex the muscles of the chip

These programs are meant to be representative of a class of applications that people are commonly using and not favor a chip manufacturer at the expense of another one

Example: a compiler is a program that is used extensively, so it makes sense to have it included in the benchmark

Two common benchmarks:
- For programs that are dominated by floating point operations (CFP2006)
- A second one is meant to be a representative sample of programs that are dominated by integer arithmetic (CINT2006)
# SPEC CPU Benchmark:
Example, highlights AMD performance

<table>
<thead>
<tr>
<th>CINT2006 Programs</th>
<th>AMD Opteron X4 – 2356 (Barcelona)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td><strong>Name</strong></td>
</tr>
<tr>
<td>Interpreted string processing</td>
<td>perl</td>
</tr>
<tr>
<td>Block-sorting compression</td>
<td>bzip2</td>
</tr>
<tr>
<td>GNU C compiler</td>
<td>gcc</td>
</tr>
<tr>
<td>Combinational optimization</td>
<td>mcf</td>
</tr>
<tr>
<td>Go game (AI)</td>
<td>go</td>
</tr>
<tr>
<td>Search gene sequence</td>
<td>hripper</td>
</tr>
<tr>
<td>Chess game (AI)</td>
<td>sjeng</td>
</tr>
<tr>
<td>Quantum computer simulation</td>
<td>libquantum</td>
</tr>
<tr>
<td>Video compression</td>
<td>h264avc</td>
</tr>
<tr>
<td>Discrete event simulation library</td>
<td>omnitpp</td>
</tr>
<tr>
<td>Games/path finding</td>
<td>aster</td>
</tr>
<tr>
<td>XML parsing</td>
<td>xatancbmk</td>
</tr>
</tbody>
</table>

[Patterson, 4th edition]→
SPEC CPU Benchmark: Example, highlights AMD performance

- Comments:

- There are programs for which the CPI is less than 1.
  - Suggests that multiple issue is at play

- Why are there programs with CPI of 10?
  - The pipeline stalls a lot, most likely due to repeated cache misses and system memory transactions
Very Important Lesson

- The cost of memory transactions trumps by far the cost of floating point operations
  - Performing arithmetic operations is almost free
  - Fetching the data required by arithmetic is expensive
Summary of Topics Covered

- Von Neumann computational model
- From code to machine instructions
- Instruction Set Architecture (ISA)
- Transistors as building blocks for control/arithmetic/logic units
- Microarchitecture
- Registers
- Pipelining
  - Structural hazard
  - Data hazard
  - Control hazard
- Performance metrics for program execution
Memory Aspects
SRAM – Static Random Access Memory

- Integrated circuit whose elements combine to make up memory arrays
- “Element”: is a special circuit, called flip-flop
- One flip-flop requires four to six transistors
- Each of these elements stores on bit of information
- Very short access time: \( \approx 1 \text{ ns} \) (order of magnitude)
- Uniform access time of any element in the array
- Writing vs. reading incurs different cost; latter more expensive
- “Static”: once set, the element stores the value set as long as powered
- Bulky, storing element is “fat” (compared to DRAM)
- Expensive - requires four to six more transistors and different layout & support requirements

[Patterson & H]→
DRAM

- DRAM type memory: the signal is stored as a charge in a capacitor
  - No charge: 0 signal
  - Some charge: 1 signal

- The good: cheap, requires only one capacitor and one transistor

- The bad: capacitors leak, so the charge or lack of charge should be reinforced every so often, from where the name “dynamic” RAM
  - State of the capacitor should be refreshed every millisecond or so
  - Refreshing requires a small delay in memory accesses

- Is this delay incurred often? (first order approximation answer)
  - Given frequency at which memory is accessed, refreshing every millisecond (1E-3 s) means issues might appear once every million cycles
  - Turns out that 99% of memory cycles are useful; refresh operations consume less than 1% of refresh overhead

[Patterson & H]→
SRAM vs. DRAM: wrap-up

- Order of the SRAM access time: 0.5ns
  - Expensive but fast
  - Mostly used on chip for caches
  - Needs no refresh

- Order of the DRAM access time: 50ns
  - Less expensive but slow
  - Mostly used off chip (system memory)
  - Higher capacity per unit area
  - Needs refresh every 10-100 ms
  - Sensitive to disturbances

- Limit case: a 100X speedup if you can work off the SRAM

<table>
<thead>
<tr>
<th></th>
<th>Transistors per bit</th>
<th>Access Time</th>
<th>Persistent?</th>
<th>Sensitive?</th>
<th>Price</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRAM</td>
<td>6</td>
<td>1X</td>
<td>Yes</td>
<td>No</td>
<td>100X</td>
<td>Cache memories</td>
</tr>
<tr>
<td>DRAM</td>
<td>1</td>
<td>10X</td>
<td>No</td>
<td>Yes</td>
<td>1X</td>
<td>Main Memory</td>
</tr>
</tbody>
</table>
## Feature Comparison Between Memory Types

<table>
<thead>
<tr>
<th>Feature</th>
<th>SRAM</th>
<th>DRAM</th>
<th>Flash</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Speed</strong></td>
<td>Very fast</td>
<td>Fast</td>
<td>Very slow</td>
</tr>
<tr>
<td><strong>Density</strong></td>
<td>Low</td>
<td>High</td>
<td>Very high</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>Low</td>
<td>High</td>
<td>Very low</td>
</tr>
<tr>
<td><strong>Refresh</strong></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Retention</strong></td>
<td>Volatile</td>
<td>Volatile</td>
<td>Non-volatile</td>
</tr>
<tr>
<td><strong>Mechanism</strong></td>
<td>Bi-stable Latch</td>
<td>Capacitor</td>
<td>Fowler-Nordheim tunneling</td>
</tr>
</tbody>
</table>
Cost and Speed Implications

- Since SRAM is expensive and bulkier, can’t have too much
  - Plagued by Space & Cost constraints

- Compromise:
  - Have some SRAM on-chip, making up what is called the “cache”
  - Have a lot of inexpensive DRAM off-chip, making up the “main memory”

- A “good program” has a low average memory access time by hitting the cache repeatedly instead of taking costly trips to main memory
  - Some simply can’t be “good programs”
  - Some can become “good programs” but you need to know how to make them good
Fallout: Memory Hierarchy

- You now have a “memory hierarchy”

- Simplest memory hierarchy:
  - Main Memory + One Cache (typically called L1 cache)

- Today’s memory architectures typically have deeper hierarchy: L1+L2+L3
  - L1 faster and smaller than L2
  - L2 faster and smaller than L3

- Note that all caches are typically on the chip
Example: Intel Core i7 975 Extreme

- Quad core Intel CPU die that illustrates L3 cache
- For, cache hierarchy is as follows
  - 32 KB L1 cache / core
  - 256 KB L2 (Instruction & Data) cache / core
  - 8 MB L3 (Instruction & Data) shared by all cores
- 730 million transistors of which half are used for L3 cache
Memory Hierarchy

- Memory hierarchy is deep:

Moving on to talk about caches
Cache Types

- Two main types of cache

  - **Data** caches: feed processor with data manipulated during execution
    - If processor would rely on data provided by main memory the execution would be pitifully slow
      - Processor Clock faster than the Memory Clock
      - Caches alleviate this memory pressure

  - **Instruction** caches: used to store instructions
    - Much simpler to deal with compared to the data caches
      - Instruction use is much more predictable than data use

- In an ideal world, the processor’s request would be met by data that already is in cache. Otherwise, a trip to main memory is in order
Split vs. Unified Caches

- Note that in the picture below L1 cache is split between data and instruction, which is typically the case.
- L2 and L3 (when present) typically unified.
How the Cache Works

- Assume simple setup with only one cache level L1

- Purpose of the cache: store for fast access a subset of the data stored in the main memory

- Data is moved at different resolutions between p ↔ c and between c ↔ mm and
  - Between p and c: moved one word at a time
  - Between c and mm: moved one block at a time (block called “cache line”)
Cache Hit vs. Cache Miss

- The processor typically agnostic about memory organization

- Middle man is the cache controller: an independent entity enabling the “agnostic” attribute of the p↔mm interaction
  - Processor requires data at some address
  - Cache Controller figures out if data is in a cache line
    - If yes: cache hit, processor served right away
    - If not: cache miss (data brought over from main memory – very slow!)
  - Difference between cache hit and cache miss:
    - Performance hit related to SRAM vs. DRAM memory access plus overhead
  - On more advanced architectures data can be “pre-fetched”
More on Cache Misses…

- A cache miss refers to a failed attempt to read/write a piece of data from/to the cache, which results in a main memory access with much longer latency.

- There are three kinds of cache misses:
  - **Cache read miss from an instruction cache**: generally causes the most delay, because the processor, or at least the thread of execution, has to wait (stall) until the instruction is fetched from main memory.
  
  - **A cache read miss from a data cache**: usually causes less delay, because instructions not dependent on the cache read can be issued and continue execution until the data is returned from main memory, and the dependent instructions can resume execution.

  - **A cache write miss to a data cache**: generally causes the least delay, because the write can be queued and there are few limitations on the execution of subsequent instructions. The processor can continue unless the queue is full and then it has to stall for the write buffer to partially drain.
Can you control what’s in the cache and anticipate future memory requests?

- Typically not…
  - Typical system has a hardware-implemented cache controller with mind of its own

- There are ways to increase your chances of cache hits by designing software for high degree of memory access locality

- Two flavors of memory locality:
  - Spatial locality
  - Temporal locality
Spatial and Temporal Locality

- Spatial Locality for memory access by a program
  - A memory access pattern characterized by bursts of repeated requests for data that is physically located within the same memory region
  - “Bursts” because this accesses should happen in a sufficiently short interval of time (otherwise the cache line gets evicted)

- Temporal Locality for memory access by a program
  - Idea: If you access a variable at some time, then you’ll probably keep accessing the same variable for a while
  - Example: have a for loop with some variables inside the loop → you keep accessing those variables as long as you loop
Cache Characteristics
[Not covered here]

- Size attributes: absolute cache size and cache line size
- Strategies for mapping memory blocks to cache lines
- Cache line replacement algorithms
- Write-back policies

NOTE: these characteristics carry over and become more convoluted when dealing with multilevel cache hierarchies
The Concept of Virtual Memory
Motivating Questions/Issues

- Question 1: On a 32 bit machine, how come you can have 512MB of main memory yet allocate an array of 1 GB?

- Question 2: How can you compile a program on a Windows workstation with 4 GB of memory and run it later on a different laptop with 1 GB of memory?

- Question 3: How can several processes run seemingly at the same time on a processor that only has one core?
The three questions raised on previous slide answered by the interplay between the compiler, the operating system (OS), and the execution model embraced by the processor

When you compile a program there is no way to know where in the physical memory the code will get its data allocated
  - There are other “tenants” that inhabit the physical memory, and they are there before you get there

Solution: code is compiled and the executable is assumed to run in a perfect memory landscape in which it has unfettered access to 4 GB of memory (on 32 bit systems).
  - The “perfect memory landscape” is called the virtual memory space
Virtual vs. Physical Memory

- **Virtual memory**: the perfect memory landscape, made up of $2^{32}$ addresses (on 32 bit architectures) in which a process sees its data being placed, the instructions stored, etc.

- **Physical memory**: a busy place hosting at the same time data and instructions associated with many applications running on the system (your laptop, desktop, etc.)
Anatomy of the Virtual Memory

**STACK segment**
- Stores a collection of frames, each associated with one function call.
- A stack frame stores function parameters, return addresses, local variables, etc.
- Last-in-first-out (LIFO) structure; push/pop managed.

**HEAP segment**
- Used when the program allocates memory dynamically at run time.
- Managed by the OS in response to function calls like `malloc`, `free`, etc.

**BSS segment**
- Stores uninitialized global and static variables.

**DATA segment**
- Stores static variables and initialized global variables.

**TEXT segment**
- Stores instructions associated with the program.

**Virtual Address Space of a process**
- Contains fixed-size sections.
- Lowest logical address: [0x0000...]
- Highest logical address: [0xffff...]

**STACK OVERFLOW**
- If top of stack reaches beyond this logical address.

**Variable size**
- Can move this way (upon return of a function).
- Can move this way (upon a function call).
The Anatomy of the Stack

- Function `bar` and associated stack frame

```c
float bar(int a, float b) {
    int initials[2];
    float t1, t2, t3;
    //..code here..
    //..no other variables..
    return t1;
}
```
The Virtual Memory.
The Page Table

- Virtual memory allows the processor to work in a *perfect memory landscape* in which each process, when run by the processor, seems to have exclusive access to a very large memory space.

- This perfect memory landscape needs to be connected back to, or mapped back into, the physical memory.

- This is accomplished via a “page table”, which is like a translation dictionary.
Anatomy of a Virtual Memory Address

- A virtual address has two parts: the page number and the offset
Anatomy of a Virtual Memory Address

- A page of virtual memory corresponds to a frame of physical memory

- The size of a page (or frame, for that matter) is typically 4096 bytes

- $2^{12} = 4096$: 12 address bits are sufficient to relatively position each address relative to the head of the frame/page
The Translation Process

- Example: imagine that your physical memory is 2 GB
- The physical address has 31 bits: $2^{31}=2$GB
- Then the page table converts bits 12 through 31 of the virtual address into bits 12 through 30 of the physical address
Short Digression 1: The Unit of Address Resolution

- How many bits are available for data storage at each address?
- Example:
  - We have $2^{32}$ addresses that we can access
  - If each address points to a location that stores 8 bits (one byte) then we have 4 GB of addressable memory
  - However, if each address refers to a location that stores 2 bytes, we have 8 GB of addressable memory
- Intel and AMD CPUs: the unit of address resolution is 1 byte (8 bits)
- Consequence: the Intel 32 bit processors “see” a virtual memory space that can be 4 GB big and can reference each byte therein
Short Digression 2: The 32 to 64 bit Migration

- If the architecture and OS have 32 bits to represent addresses, it means that $2^{32}$ addresses can be referenced.

- If unit of address resolution is 1 byte, that means that the size of the virtual memory space can be 4 GB.

- This is hardly enough today when programs are very large and the amounts of data they manipulate can be staggering.

- This motivated the push towards having addresses represented using 64 bits: the memory space balloons to $2^{64}$ bytes, that is 16 times $1,152,921,504,606,846,976$ bytes.
Short Digression 3: 
The 32 to 64 Bit Migration

- Note that a 64 bit architecture typically calls for two things:

- From a **hardware** perspective, the size of the registers, integer size, and word size is 64 bits

- From a **software** perspective, the program “operates” in a huge virtual memory space
  - The operating system (OS) is the party managing the execution of a program in the 64 bit universe
  - Can allocate very large arrays
  - In reality, the virtual memory addresses are represented using only about 40 to 45 bits
The page table is the key ingredient that allows the translation of virtual addresses into physical addresses.

Every single process executing on a processor and managed by the OS has its own page table.

Where is this page table stored?
- Stored in the main memory.
Quiz

- How much memory does the page table need for a 32 bit operating system if the size of the virtual page is 4 MB?

- Can the page table be stored in cache?
The TLB
[Translation Lookaside Buffer]

- If Page Table stored in main memory it means that each address translation would require a trip to main memory
  - This would be extremely costly

- There is a “cache” for this translation process: TLB
  - Translation Lookaside Buffer: holds the translation of a small collection of virtual page numbers into frame IDs

- Best case scenario: the TLB leads to a hit and allows for quick translation
- Bad scenario: the TLB doesn’t have the required information cached and a trip to main memory is in order
- Worst scenario: the requested frame is not in main memory and a trip to secondary memory is in order
  - Called “page fault”
Illustration: The Role of the TLB

- A TLB provides a caching mechanism

- A TLB miss leads to substantial overhead in the translation of a virtual memory address
Memory Access: The Big Picture, Includes Cache

- A simplified version of how a memory request is serviced presented below.
Summary of Topics Covered

- Von Neumann computational model
- From code to machine instructions
- Instruction Set Architecture (ISA)
- Transistors as building blocks for control/arithmetic/logic units
- Microarchitecture
- Registers
- Pipelining
- Performance metrics for program execution
- SRAM vs DRAM
- Caches and the hierarchical setup of the memory ecosystem
- Virtual Memory
- Virtual Memory address translation and the role of the TLB
Parallel Computing: Why? & Why Now?
Acknowledgements

- Material presented today includes content due to
  - Hennessy and Patterson (Computer Architecture, 4th edition)
  - John Owens, UC-Davis
  - Darío Suárez, Universidad de Zaragoza
  - John Cavazos, University of Delaware
  - Others, as indicated on various slides
  - I apologize if I included a slide and didn’t give credit where was due
This Segment’s Main Points

- Sequential computing has been losing steam recently

- The immediate future seems to belong to parallel computing
CPU Speed Evolution

[log scale]
More Historical Data

35 Years of Microprocessor Trend Data

- Transistors (thousands)
- Single-thread Performance (SpecINT)
- Frequency (MHz)
- Typical Power (Watts)
- Number of Cores

[Sam Naffziger, AMD]→
...we can expect very little improvement in serial performance of general purpose CPUs. So if we are to continue to enjoy improvements in software capability at the rate we have become accustomed to, we must use parallel computing. This will have a profound effect on commercial software development including the languages, compilers, operating systems, and software development tools, which will in turn have an equally profound effect on computer and computational scientists.

John L. Manferdelli, Microsoft Corporation Distinguished Engineer, leads the eXtreme Computing Group (XCG) System, Security and Quantum Computing Research Group
Three Walls to Serial Performance

- Memory Wall
- Instruction Level Parallelism (ILP) Wall
- Power Wall

**Source:** “The Many-Core Inflection Point for Mass Market Computer Systems”, by John L. Manferdelli, Microsoft Corporation

http://www.ctwatch.org/quarterly/articles/2007/02/the-many-core-inflection-point-for-mass-market-computer-systems/
Memory Wall

- Memory Wall: What is it?
  - The growing disparity of speed between CPU and memory outside the CPU chip.

- Current architectures have ever growing caches to improve the average memory reference time to fetch or write instructions or data

- Memory Wall: due to *latency* and limited communication *bandwidth* beyond chip boundaries.
  - From 1986 to 2000, CPU speed improved at an annual rate of 55% while memory access speed only improved at 10%
Memory Bandwidths
[typical embedded, desktop, and server computers]

- Register reference
  - Size: 500 bytes
  - Speed: 250 ps

- Cache reference
  - Size: 64 KB
  - Speed: 1 ns

- Memory reference
  - Size: 1 GB
  - Speed: 100 ns

- Disk memory reference
  - Size: 1 TB
  - Speed: 10 ms

© 2007 Elsevier, Inc. All rights reserved.
Memory Speed:
Widening of the Processor-DRAM Performance Gap

- The processor: So fast it has left the memory behind
  - A system (CPU-Memory duo) slowed down by sluggish memory

- Plot on next slide shows on a *log* scale the increasing gap between CPU and memory

- The memory baseline: 64 KB DRAM in 1980

- Memory speed increasing at a rate of approx 1.07/year
  - However, processors improved
    - 1.25/year (1980-1986)
    - 1.52/year (1986-2004)
    - 1.20/year (2004-2010)
Memory Speed:
Widening of the Processor-DRAM Performance Gap
Memory Latency vs. Memory Bandwidth

- **Latency**: the amount of time it takes for an operation to complete
  - Measured in seconds
  - The utility “ping” in Linux measures the latency of a network
  - For **memory** transactions: send 32 bits to destination and back, measure how much time it takes → gives you latency

- **Bandwidth**: how much data can be transferred per second
  - You can talk about bandwidth for memory but also for a network (Ethernet, Infiniband, modem, DSL, etc.)

- **Improving Latency and Bandwidth**
  - The job of colleagues in Electrical Engineering
  - Once in a while, Materials Science colleagues deliver a breakthrough
  - Promising technology: optic networks and layered memory on top of chip
Memory Latency vs. Memory Bandwidth

- Memory Access Latency is significantly more challenging to improve as opposed to improving Memory Bandwidth.

- Improving Bandwidth: add more “pipes”.
  - Requires more pins that come out of the chip for DRAM, for instance.
  - Adding more pins is not simple – very crowded real estate plus the technology is tricky.

- Improving Latency: no easy answer here.

- Analogy:
  - If you carry commuters with a train, add more cars to a train to increase bandwidth.
  - Improving latency requires the construction of high speed trains:
    - Very expensive
    - Requires qualitatively new technology.
Latency vs. Bandwidth Improvements Over the Last 25 years
SK Hynix’s High Bandwidth Memory (HBM)
- Developed by AMD and SK Hynix

1st Generation (HBM1) introduced in AMD Fiji GPUs
- 1GB & 128GB/s per stack
- AMD Radeon R9 Fury X: had four stacks → 4GB & 512GB/s

2nd Generation (HBM2) will be used in NVIDIA Pascal and AMD Arctic Island GPUs
- 2 GB & 256GB/s bandwidth per stack
- NVIDIA Pascal reported to have 1TB/s memory bandwidth
3D Stacked Memory
[immediate future looks bright]

This is a big deal.
GPU w/ HBM
[cut-through]

Through Silicon Vias (TSVs), μBumps

3D engine | display controller | HBM controller
---|---|---

Silicon interposer

DRAM

dice

HBM controller die

1024 data links / HBM stack @ 500MHz

Package substrate

solder balls

Graphics card

Multi-layer Printed Circuit Board (PCB), up to 8 layers

PCI Express
Electrical current
Display connectors
3D Stacked Memory

- Shorter distances
  - Smaller latency
  - Less power to move data

- Smaller memory footprint
  - 16X smaller footprint

<table>
<thead>
<tr>
<th></th>
<th>GDDR5</th>
<th>Per Package</th>
<th>HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus Width</td>
<td>32-bit</td>
<td>Bus Width</td>
<td>1024-bit</td>
</tr>
<tr>
<td>Clock Speed</td>
<td>Up to 1750MHz (7GBps)</td>
<td>Clock Speed</td>
<td>Up to 500MHz (1GBps)</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>Up to 28GB/s per chip</td>
<td>Bandwidth</td>
<td>&gt;100GB/s per stack</td>
</tr>
<tr>
<td>Voltage</td>
<td>1.5V</td>
<td>Voltage</td>
<td>1.3V</td>
</tr>
</tbody>
</table>

![Diagram showing 3D stacked memory comparison between GDDR5 and HBM](Image)

1GB GDDR5
1GB HBM
16% less surface area

Areal, to scale

[AMD]→
Memory Wall, Conclusions

[IMPORTANT SLIDE]

- Memory trashing is what kills execution speed

- Many times you will see that when you run your application:
  - You are far away from reaching top speed of the chip
  - AND
  - You are at top speed for your memory
    - If this is the case, you are trashing the memory
    - Means that basically you are doing one or both of the following
      - Move large amounts of data around
      - Move data often

<table>
<thead>
<tr>
<th>Memory Access Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>To/From Registers</td>
</tr>
<tr>
<td>To/From Cache</td>
</tr>
<tr>
<td>To/From RAM</td>
</tr>
<tr>
<td>To/From Disk</td>
</tr>
</tbody>
</table>

© 2007 Elsevier, Inc. All rights reserved.
Instruction Level Parallelism (ILP)

- ILP: a relevant factor in reducing execution times after 1985

- The basic idea:
  - Overlap execution of instructions to improve overall performance
  - During the same clock cycle several instructions are being processed

- Two approaches to leveraging ILP
  - Dynamic: relies on hardware to discover/exploit parallelism dynamically at run time
    - It is the dominant one in the market
  - Static: relies on compiler to identify parallelism in the code and leverage it (VLIW)

- Examples where ILP expected to improve efficiency

```plaintext
for( int=0; i<1000; i++)
  x[i] = x[i] + y[i];

1. e = a + b
2. f = c + d
3. g = e * f
```
ILP: Various Angles of Attack

- **Instruction pipelining**: the execution of multiple instructions can be partially overlapped; where each instruction is divided into series of sub-steps (termed: micro-operations)

- **Superscalar execution**: multiple execution units are used to execute multiple instructions in parallel

- **Out-of-order execution**: instructions execute in any order but without violating data dependencies

- **Register renaming**: a technique used to avoid data hazards and thus lead to unnecessary serialization of program instructions caused by the reuse of registers

- **Speculative execution**: allows the execution of complete instructions or parts of instructions before being sure whether this execution is required

- **Branch prediction**: used to avoid delays (termed: stalls). Used in combination with speculative execution.
The ILP Wall

- **ILP, the good:**
  - *Existing* programs enjoy performance benefits without any modification
  - Recompiling them is beneficial but entirely up to you as long as you stick with the same ISA (for instance, if you go from Pentium 2 to Pentium 4 you don’t have to recompile your executable)

- **ILP, the bad:**
  - Improvements are difficult to forecast since the “speculation” success is difficult to predict
  - Moreover, ILP causes a super-linear increase in execution unit complexity (and associated power consumption) without linear speedup.

- **ILP, the ugly:** serial performance acceleration using ILP plateauing because of these effects
The Power Wall

- “Power, and not manufacturing, limits traditional general purpose microarchitecture improvements” (F. Pollack, Intel Fellow)

- Leakage power dissipation gets worse as gates get smaller, because gate dielectric thicknesses must proportionately decrease

Adapted from F. Pollack (MICRO’99)
The Power Wall

- Power dissipation in clocked digital devices is related to feature length through Dennard scaling
  - Dennard scaling was the practical observation that made Moore’s Law useful

- Clock speed increased by a factor of 4,000 in less than two decades
  - The ability to dissipate heat is limited though…
  - Look back at the last five years, the clock rates are pretty much flat

- Significant increase in clock speed without heroic (and expensive) cooling is not possible. Chips would simply melt

- Problem might be addressed one day by a Materials Science breakthrough
Trivia

- AMD Phenom II X4 955 (4 core load)
  - 236 Watts

- Intel Core i7 920 (8 thread load)
  - 213 Watts

- Human Brain
  - 20 W
  - Represents 2% of our mass
  - Burns 20% of all energy in the body at rest