Basics - CPU Internals, Performance Measurement

Adapted from Rajeev Subramanium's Spring '16 CS6810 course (University of Utah)

14 Jan 2019

Aftab Hussain
University of California,
Irvine

Quick peek into what we are evaluating

CPU

IC

(Integrated Circuit)

Transistor

The electronic circuitry that carries out instructions of a program.

Consists of control unit, registers, an arithmetic and logic unit, the instruction execution unit, and interconnections among those components.

Traditionally referred to as the **processor**.

Processor -> Microprocessor

Microprocessor is a processor that incorporates the functions of a central processing unit on a single integrated circuit (IC), or at most a few integrated circuits.

Christensson, Per. "Processor Definition." TechTerms. Sharpened Productions, 09 April 2012. Web. 14 January 2019. <https://techterms.com/definition/processor>.

https://en.wikipedia.org/wiki/Central_processing_unit

CPU

IC
(Integrated Circuit)

Transistor

A small chip that can function as an amplifier, oscillator, timer, **microprocessor**, or even computer memory.

Usually made of silicon, that can hold anywhere from **hundreds to millions of transistors, resistors, and capacitors**.

Can perform **calculations** and **store data**.

Digital ICs use logic gates, which work only with values of **ones** and **zeros**.

A **low signal** sent to to a component on a digital IC will result in a value of 0, while a **high signal** creates a value of 1.

Digital ICs are the kind you will usually find in computers, networking equipment, and most consumer electronics.

Christensson, Per. "Integrated Circuit Definition." *TechTerms*. Sharpened Productions, 2006. Web. 14 January 2019. <https://techterms.com/definition/integratedcircuit>.

CPU

IC

(Integrated Circuit)

Transistor

A basic electrical component that **alters the flow of electrical current.**

Most transistors include three connection points, or terminals, which can connect to other transistors or electrical components. By modifying the current between the first and second terminals, the current between the second and third terminals is changed.

This allows a transistor to act as a **switch**, which can turn a signal on or off (can be represented as 0 or 1).

A series of transistors may also be used as a logic gate when performing logical operations.

CPU transistors, such as those used in Intel's Ivy Bridge processor, are separated by a distance of 22 nanometers. This microscopic size allows chip manufacturers to fit hundreds of millions of transistors into a single processor.

Christensson, Per. "Transistor Definition." *TechTerms*. Sharpened Productions, 07 October 2011. Web. 14 January 2019. <https://techterms.com/definition/transistor>.

Performance Metrics

Two primary metrics:

wall clock time (response time for a program or latency)
throughput (jobs performed in unit time)

Improving response time would increase throughput, but not vice versa necessarily.

Throughput can be increased by improving CPU utilization.

When 2 programs are using a processor, one may have to wait for another due to cpu constraints, hence worsening latency.

Benchmark suites

Performance is measured with benchmark suites: a
collection of programs that are likely relevant to the user

SPEC CPU 2006: cpu-oriented programs (for desktops)
SPECweb, TPC: throughput-oriented (for servers)
EEMBC: for embedded processors/workloads

Need to Summarize Performance

Consider 25 programs from a benchmark set – how do we capture the behavior of all 25 programs with a single number?

	P1	P2	P3
Sys-A	10	8	25
Sys-B	12	9	20
Sys-C	8	8	30

Need to Summarize Performance

Consider 25 programs from a benchmark set – how do we capture the behavior of all 25 programs with a single number?

	P1	P2	P3
Sys-A	10	8	25
Sys-B	12	9	20
Sys-C	8	8	30

>Sum of execution times (AM)
>Sum of weighted execution times (AM)
>Geometric mean of execution times (GM)

Need to Summarize Performance

Consider 25 programs from a benchmark set – how do we capture the behavior of all 25 programs with a single number?

	P1	P2	P3	
Sys-A	10	8	25	43
Sys-B	12	9	20	41
Sys-C	8	8	30	46

>Sum of execution times (AM)
>Sum of weighted execution times (AM)
>Geometric mean of execution times (GM)

Need to Summarize Performance

Consider 25 programs from a benchmark set – how do we capture the behavior of all 25 programs with a single number?

	P1	P2	P3	
Sys-A	10	8	25	43
Sys-B	12	9	20	41
Sys-C	8	8	30	46 - P3 getting too much value

>Sum of execution times (AM)

>Sum of weighted execution times (AM)

>Geometric mean of execution times (GM)

Need to Summarize Performance

Consider 25 programs from a benchmark set – how do we capture the behavior of all 25 programs with a single number?

	P1	P2	P3
Sys-A	10	8	25
Sys-B	12	9	20
Sys-C	8	8	30

>Sum of execution times (AM)
>Sum of weighted execution times (AM)
>Geometric mean of execution times (GM)

Choose a reference machine and add a weight to each program so they all have same importance.

Need to Summarize Performance

Consider 25 programs from a benchmark set – how do we capture the behavior of all 25 programs with a single number?

	P1	P2	P3
Sys-A	10	8	25
Sys-B	12	9	20
Sys-C	8	8	30

>Sum of execution times (AM)
>Sum of weighted execution times (AM)
>Geometric mean of execution times (GM)

Geometric mean of n numbers = nth root of product of n numbers

** affecting execution time of a program affects GM in the same way, regardless of the program being affected.

Need to Summarize Performance

Consider 25 programs from a benchmark set – how do we capture the behavior of all 25 programs with a single number?

	P1	P2	P3
Sys-A	10	8	25
Sys-B	12	9	20
Sys-C	8	8	30

>Sum of execution times (AM)
>Sum of weighted execution times (AM)
>Geometric mean of execution times (GM)

Geometric mean of n numbers = nth root of product of n numbers

** no normalization
** no ref machine

GM Inconsistency

	Computer-A	Computer-B	Computer-C
P1	1 sec	10 secs	20 secs
P2	1000 secs	100 secs	20 secs

Conclusion with GMs:

(i) A=B (ii) C is ~1.6 times faster

GM Inconsistency

	Computer-A	Computer-B	Computer-C
P1	1 sec	10 secs	20 secs
P2	1000 secs	100 secs	20 secs

Conclusion with GMs:
(i) A=B (ii) C is ~1.6 times faster

What's the workload to satisfy the above?
[board explanation, see note]

Two Comparison Metrics

"Speedup" is a ratio = old exec time / new exec time

"Improvement", "Increase", "Decrease" usually refer to percentage relative to the baseline
= (new perf – old perf) / old perf

Performance is the inverse of execution time.

Clocks

[voltage pulses]

Performance Equation

Clock cycle time = 1 / clock speed

CPU time = clock cycle time x
cycles per instruction (CPI) x
number of instructions

(The total execution time of a program)

Problems 4 & 5 on Rajeev's Slides (Slide nos. 18, 19, 24, 25)

[Also see note on performance equation derivation in terms of IPC (Instructions per Cycle)]

Power

The instantaneous electrical power *P* delivered to a component is given by

$$P(t) = I(t) \cdot V(t)$$

where

P(*t*) is the instantaneous power, measured in watts (joules per second)

V(*t*) is the potential difference (or voltage drop) across the component, measured in volts

I(*t*) is the current through it, measured in amperes

If the component is a resistor with time-invariant voltage to current ratio, then:

$$P = I \cdot V = I^2 \cdot R = \frac{V^2}{R}$$

where

$$R = \frac{V}{I}$$

is the resistance, measured in ohms.

https://en.wikipedia.org/wiki/Power_(physics)

Problems 1 & 2 on Rajeev's Slides (Slide nos. 3, 4, 5, 7, 8)

Thank you