NYU
New York University

Multicore Processors: Architecture & Programming

Performance Evaluation

Mohamed Zahran (aka Z)
mzahran@cs.nyu.edu
http://www.mzahran.com

What Are We trying to do?

- Measure, Report, and Summarize Performance
- Make intelligent choices

Why is some hardware better than others for different programs?

What factors of system performance are hardware related?

Does performance measure depend on application type?

Let's Start with Two Simple Metrics

- **Response time** (aka Execution Time)
  - The time between the start and completion of a task
- **Throughput**
  - Total amount of work done in a given time

What is the relationship between execution time and throughput?

Computer Performance: TIME, TIME, TIME

- Response Time (latency)
  - How long does it take for my job to run?
  - How long does it take to execute a job?
  - How long must I wait for the database query?
- Throughput
  - How many jobs can the machine run at once?
  - What is the average execution rate?
  - How much work is getting done?

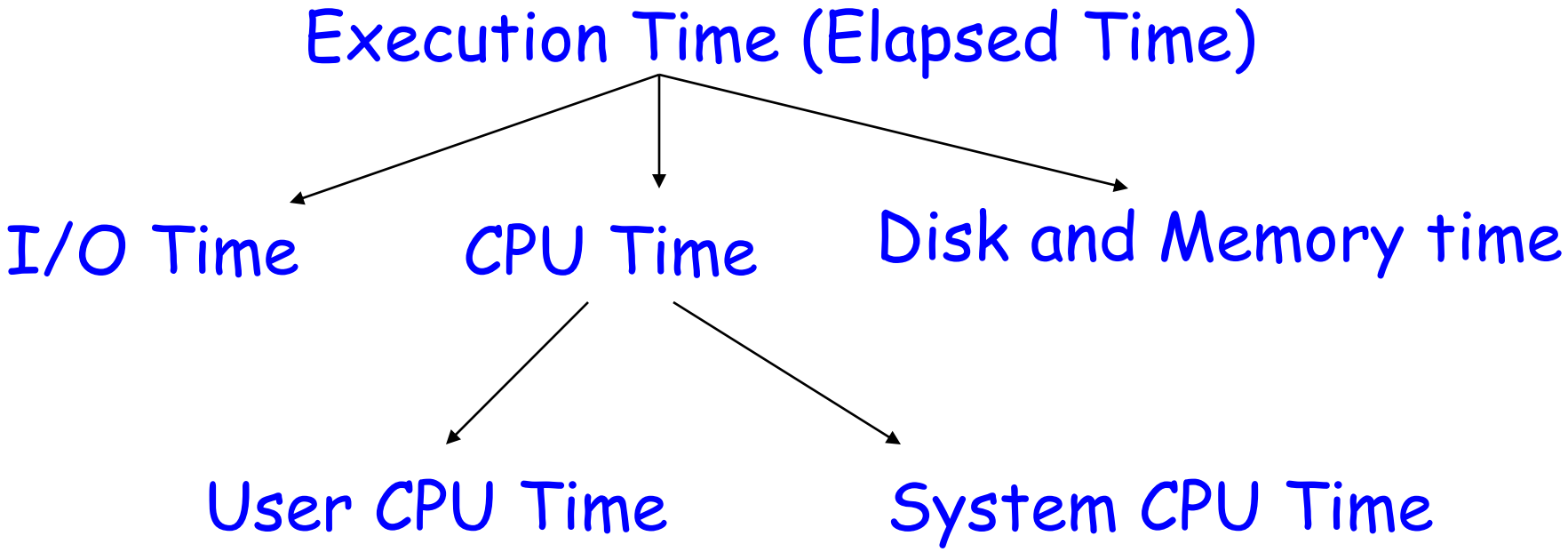
Try to solve this…

Do the following changes to the computer system increase throughput, decrease response time, or both?

- Replacing the processor with a faster version
- Adding additional processors to a system.

Execution Time

- **Elapsed Time**
  - counts everything *(disk and memory accesses, I/O , etc.)*
  - a useful number, but often not good for comparison purposes
- **CPU time**
  - doesn't count I/O or time spent running other programs
  - can be broken up into system time, and user time


Execution Time (Elapsed Time)
I/O Time
CPU Time
Disk and Memory time
User CPU Time
System CPU Time

Taking Timings

In Linux:
time prog

Returns
real Xs
user Ys
sys Zs

Inside your C program:
clock_t clock(void)returns the number of clock ticks elapsed since the program started

```
#include <time.h>
#include <stdio.h>

int main() {
        clock_t  start, end, total;
        int i;

        start = clock();

        for(i=0; i< 10000000; i++) { }

        end = clock();
        total= (double)(end – start) / CLOCKS_PER_SEC;

        printf("Total time taken by CPU: %f\n", total);
}
```

Taking Timings

In Linux:
time prog

Returns
real Xs
user Ys
sys Zs

Wall clock time
CPU time
Sys CPU time

Inside your C program:
clock_t clock(void)returns the number of clock ticks elapsed since the program started

```
#include <time.h>
#include <stdio.h>

int main() {
        clock_t  start, end, total;
        int i;

        start = clock();

        for(i=0; i< 10000000; i++) { }

        end = clock();
        total= (double)(end – start) / CLOCKS_PER_SEC;

        printf("Total time taken by CPU: %f\n", total);
}
```

Book's Definition of Performance

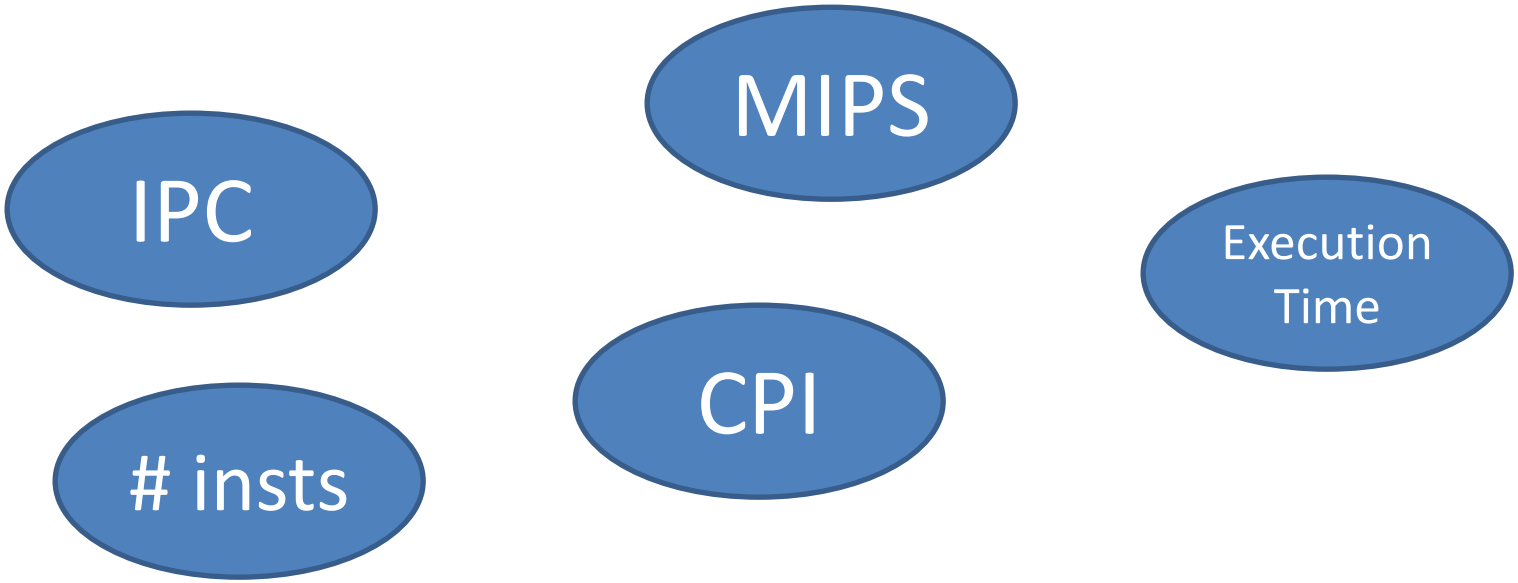
- For some program running on machine X,

$$\text{Performance}_X = 1 / \text{Execution time}_X$$

- "X is n times faster than Y" means:

$$\text{Performance}_X \ / \ \text{Performance}_Y = n$$

Let's loot at sequential programs first.



An Interesting Question

If two machines have the same ISA which of our quantities (e.g., clock rate, CPI, execution time, # of instructions, MIPS) will always be identical?

Execution time for sequential program:

$$\frac{\text{seconds}}{\text{program}} = \frac{\text{cycles}}{\text{program}} \times \frac{\text{seconds}}{\text{cycle}}$$

ET

IC * CPI

CT

ET = IC X CPI X CT

ET = Execution Time
CPI = Cycles Per Instruction
IC = Instruction Count

Example

A program runs in 10 seconds on computer A, which has a 4 GHz. clock. We are trying to help a computer designer build a new machine B, that will run this program in 6 seconds. The designer can use new (or perhaps more expensive) technology to substantially increase the clock rate, but has informed us that this increase will affect the rest of the CPU design, causing machine B to require 1.2 times as many clock cycles as machine A for the same program. What clock rate should we tell the designer to target?“

ET = IC * CPI * CT = total_cycles * CT

A: 10 = total_cycles * (1/4GHz)
B: 6 = 1.2 * total_cycles * (1/clock_rate)

10/6 = clock_rate/(1.2 * 4GHz)
Clock_rate = 8GHz

CPI Example

- Suppose we have two implementations of the same instruction set architecture (ISA).

  For some program,
  Machine A has a clock cycle time of 250 ps and a CPI of 2.0
  Machine B has a clock cycle time of 500 ps and a CPI of 1.2
  What machine is faster for this program, and by how much?

[10^{-3} = milli, 10^{-6} = micro, 10^{-9} = nano, 10^{-12} = pico, 10^{-15} = femto]

ET = IC*CPI*CT

ET_A = IC * 2*250 = 500IC

ET_B = IC * 1.2 * 500 = 600IC

#Instructions Example

- A compiler designer is trying to decide between two code sequences for a particular machine. Based on the hardware implementation, there are three different classes of instructions: Class A, Class B, and Class C, and they require one, two, and three cycles (respectively).

  The first code sequence has 5 instructions:
  2 of A, 1 of B, and 2 of C

  The second sequence has 6 instructions:
  4 of A, 1 of B, and 1 of C.

  Which sequence will be faster? How much?
  What is the CPI for each sequence?

Same machine = CT is the same for both sequences.

#Instructions Example

ET = IC * CPI * CT = total number of cycles * CT

The first code sequence has 5 instructions:
2 of A, 1 of B, and 2 of C

total cycles = (2*1) + (1*2) + (2*3) = 10

ET = 10*CT

The second sequence has 6 instructions:
4 of A, 1 of B, and 1 of C.

total cycles = (4*1) + (1*2) + (1*3) = 9

ET = 9*CT

Second sequence is faster.

Which sequence will be faster? How much?
What is the CPI for each sequence?

CPI = Cycles per Instructions = #cycles / # instructions

For 1st sequence = 10/(2+1+2) = 2

For 2nd sequence = 9/6 = 1.5

MIPS Example

- Two different compilers are being tested for a 4 GHz. machine with three different classes of instructions: Class A, Class B, and Class C, which require one, two, and three cycles (respectively). Both compilers are used to produce code for a large piece of software.

  The first compiler's code uses 5 million Class A instructions, 1 million Class B instructions, and 1 million Class C instructions.

  The second compiler's code uses 10 million Class A instructions, 1 million Class B instructions, and 1 million Class C instructions.

- Which sequence will be faster according to MIPS?
- Which sequence will be faster according to execution time?

MIPS Example

- Two different compilers are being tested for a 4 GHz. machine with three different classes of instructions: Class A, Class B, and Class C, which require one, two, and three cycles (respectively). Both compilers are used to produce code for a large piece of software.

  The first compiler's code uses 5 million Class A instructions, 1 million Class B instructions, and 1 million Class C instructions.

  The second compiler's code uses 10 million Class A instructions, 1 million Class B instructions, and 1 million Class C instructions.

- Which sequence will be faster according to MIPS?
- Which sequence will be faster according to execution time?

MIPS = Million Instructions per Second = # instructions in millions / total time in seconds

Total time in seconds = total number of cycles * cycle time = total number of cycles/Frequency

1st compiler:

total #cycles = $((5*1)+(1*2)+(1*3) * 10^6) = 10^7$

Frequency = $4* 10^9$

MIPS = $7/(10^7 / 4* 10^9) = 2800$

2nd compiler:

total #cycles = $((10*1)+(1*2)+(1*3) * 10^6)$ $= 15* 10^6$

MIPS = $12/(15* 10^6 / 4* 10^9) = 3200$

MIPS Example

- Two different compilers are being tested for a 4 GHz. machine with three different classes of instructions: Class A, Class B, and Class C, which require one, two, and three cycles (respectively). Both compilers are used to produce code for a large piece of software.

  The first compiler's code uses 5 million Class A instructions, 1 million Class B instructions, and 1 million Class C instructions.

  The second compiler's code uses 10 million Class A instructions, 1 million Class B instructions, and 1 million Class C instructions.

- Which sequence will be faster according to MIPS?
- Which sequence will be faster according to execution time?

ET = IC * CPI * CT = total number of cycles * CT = total number of cycles/frequency

1st compiler:
total #cycles = $((5*1)+(1*2)+(1*3) * 10^6) = 10^7$
Frequency = $4* 10^9$
ET = $10^7 /(4* 10^9) = 1/400$

2nd compiler:
total #cycles = $((10*1)+(1*2)+(1*3) * 10^6)$
$= 15* 10^6$
ET = $15* 10^6 /(4* 10^9) = 1.5/400$

Your Program Does Not Run in A Vacuum

- System software at least is there
- Multi-programming setting is very common in multicore settings
- Independent programs affect each other performance (why?)

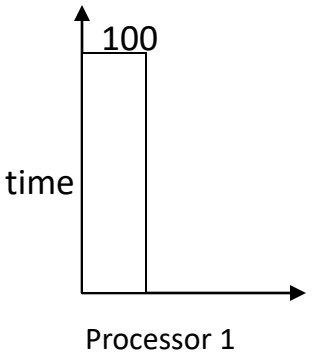
Let's look at parallel programs.

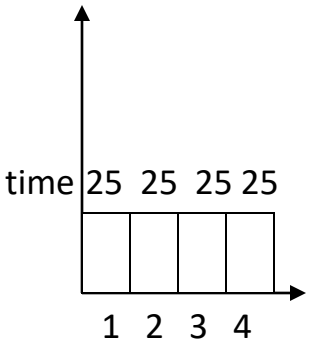
Speedup

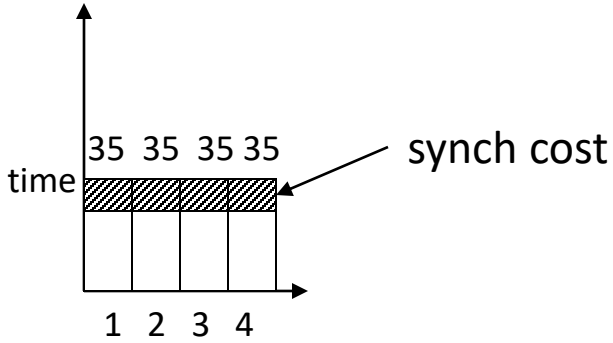
- Number of cores = p
- Serial run-time = T_{serial}
- Parallel run-time = $T_{parallel}$

$$S = \frac{T_{serial}}{T_{parallel}}$$

Example







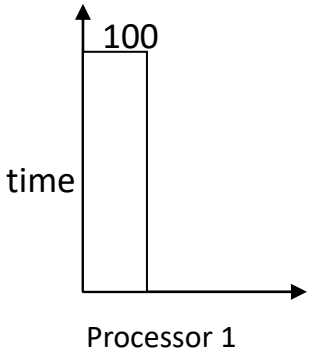
$$S_p = \frac{100}{25} = 4.0,$$

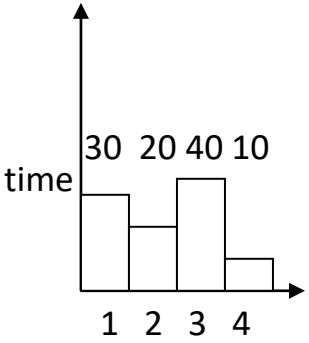
Perfect parallelization!
Does it ever occur?

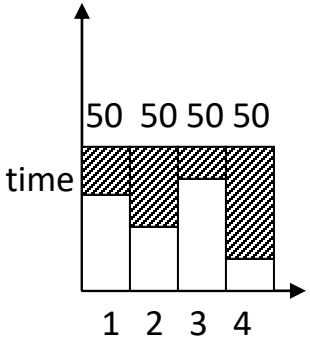
$$S_p = \frac{100}{35} = 2.85,$$

perfect load balancing

Example (cont.)







closest to
real life
parallel programs

$$S_p = \frac{100}{40} = 2.5,$$

load imbalance

$$S_p = \frac{100}{50} = 2.0,$$

load imbalance

and sync cost

Efficiency of a parallel program

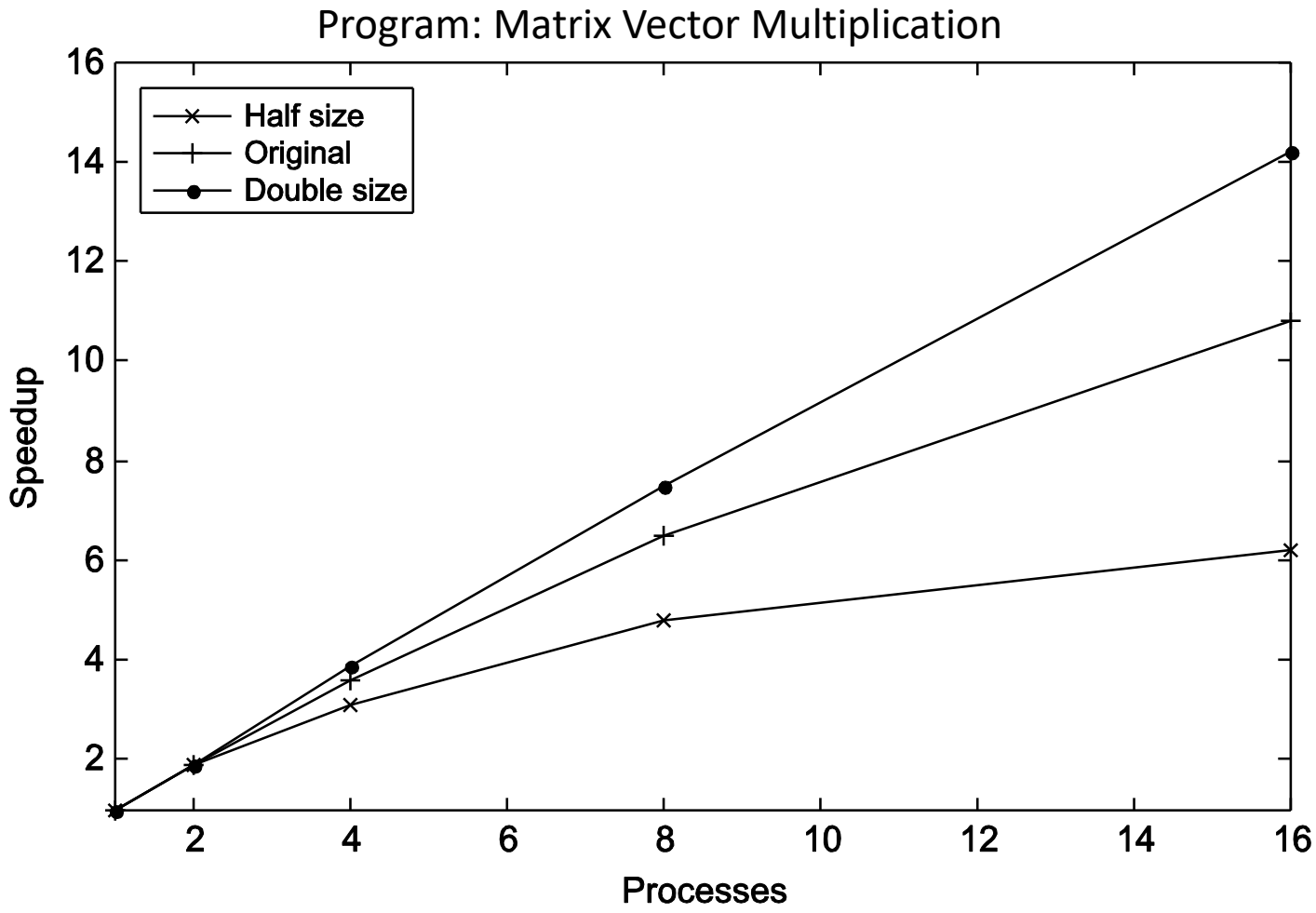
$$E = \frac{\text{Speedup}}{P} = \frac{\left(\frac{T_{serial}}{T_{parallel}}\right)}{p}$$

P = can be number of threads, processes, or cores

Be Careful about T

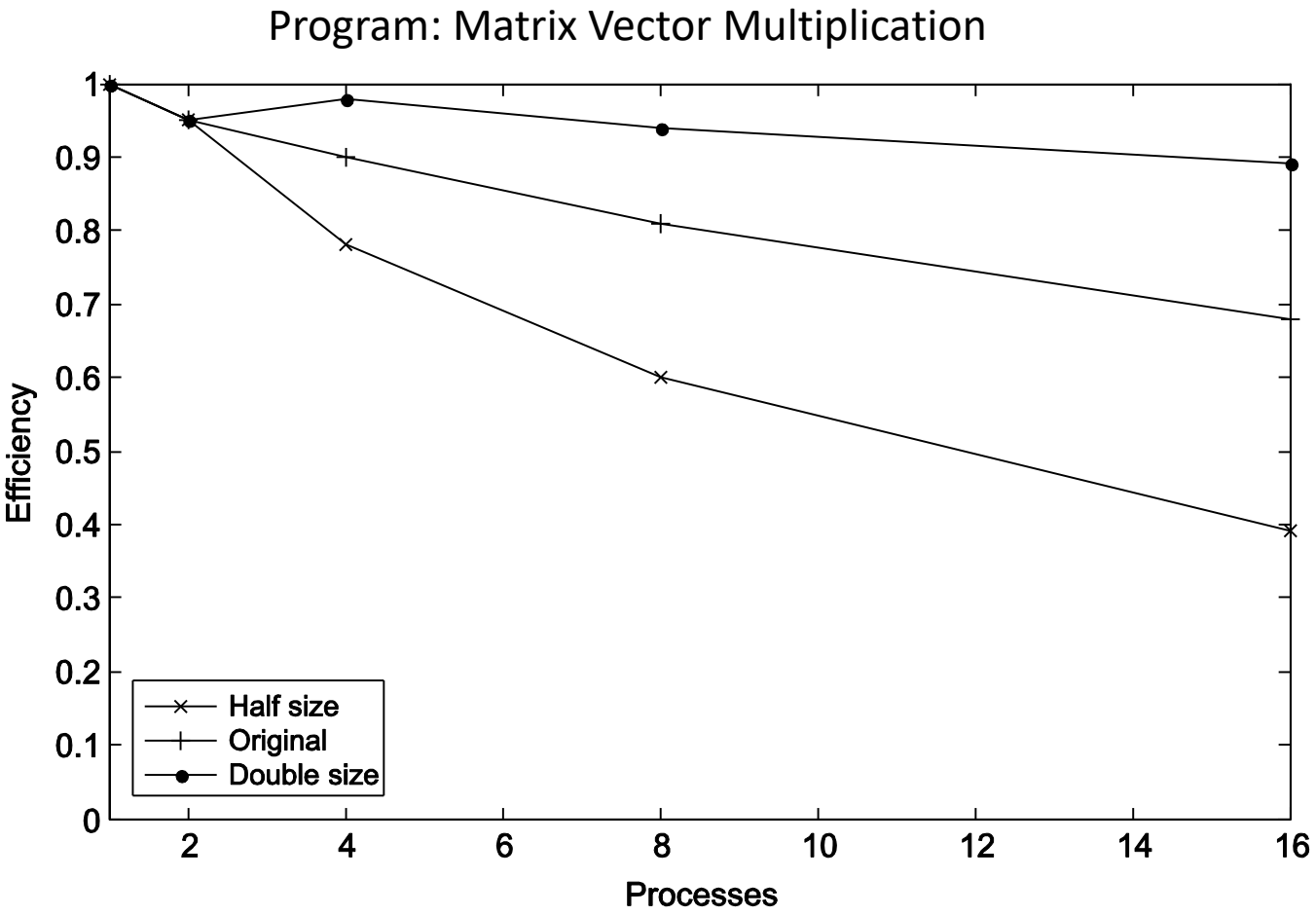
- Both T_{serial} and T_{par} are wall-clock times, and as such they are not objective. They can be influenced by :
  - The skill of the programmer who wrote the implementations
  - The choice of compiler (e.g. GNU C++ versus Intel C++)
  - The compiler switches (e.g. turning optimization on/off)
  - The operating system
  - The type of filesystem holding the input data (e.g. EXT4, NTFS, etc.)
  - The time of day... (different workloads, network traffic, etc.)

Speedup



Graph shows how speedup is affected by the number of processes and problem size.

Efficiency



Scalability

$$E = \frac{\text{Speedup}}{P} = \frac{\frac{T_{serial}}{T_{parallel}}}{p}$$

- Scalability is the ability of a (software or hardware) system to handle a growing amount of work efficiently.
- If we keep the efficiency fixed by increasing the number of processes/threads and without increasing problem size, the problem is *strongly scalable*.
- If we keep the efficiency fixed by increasing the problem size at the same rate as we increase the number of processes/threads, the problem is *weakly scalable*.

The total number of instructions executed may be different across different runs!

This effect increases with the number of cores

System-level code account for a significant fraction of the total execution time

Sources of Parallel Overheads

- Overhead of creating threads/processes
- Synchronization
- Load imbalance
- Communication
- Extra computation
- Memory access (for both sequential and parallel!)

How about multiprogramming environment?

Some Metrics About Multiprogramming

Normalized progress of program i

$$NP_i = \frac{T_i^{SP}}{T_i^{MP}},$$

T_i^{SP}: Time when running in isolation

T_i^{MP}: Time when running with other programs

System throughput

$$STP = \sum_{i=1}^{n} NP_i = \sum_{i=1}^{n} \frac{T_i^{SP}}{T_i^{MP}}$$

Higher-is-better metric

Some Metrics About Multiprogramming

Normalized Turnaround time of program i

$$NTT_i = \frac{T_i^{MP}}{T_i^{SP}}$$

Time when running with other programs

Time when running in isolation

Average normalized turnaround time

$$ANTT = \frac{1}{n}\sum_{i=1}^{n} NTT_i = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i^{MP}}{T_i^{SP}},$$

Lower-is-better metric

Other Metrics

$$IPC\_throughput = \sum_{i=1}^{n} IPC_i$$

$$weighted\_speedup = \sum_{i=1}^{n} \frac{IPC_i^{MP}}{IPC_i^{SP}}$$

$$hmean = \frac{n}{\sum_{i=1}^{n} \frac{IPC_i^{SP}}{IPC_i^{MP}}}$$

Harmonic Vs Arithmetic

- Both used to compute an average (i.e. combine several measures) of a metric.
- Assume the metric is computed A/B
  - If A is weighted equally among all the benchmarks → then harmonic mean is meaningful.
  - If B is weighted equally among all the benchmarks → arithmetic mean
- Example: Suppose we gathered IPC of several benchmarks and want to combine them
  - If we execute all benchmarks for the same amount of instructions (e.g. 1 billion instructions) → hmean
  - If we execute all the benchmarks for the same amount of cycles → arithmetic mean

Benchmarks

Benchmarks

- Performance best determined by running a real application
  - Use programs typical of expected workload
  - Or, typical of expected class of applications
  - e.g., compilers/editors, scientific applications, graphics, etc.
- Small benchmarks
  - nice for architects and designers
  - easy to standardize
- Parallel Benchmarks: PARSEC, Rodinia, SPLASH-2
- SPEC (System Performance Evaluation Cooperative)
  - companies have agreed on a set of real program and inputs
  - valuable indicator of performance (and compiler technology)

Role of Benchmarks

- help designer explore architectural designs
- identify bottlenecks
- compare different systems
- conduct performance prediction

Example: PARSEC

- Princeton Application Repository for Shared-Memory Computers
- Benchmark Suite for Chip-Multiprocessors
- Freely available at: http://parsec.cs.princeton.edu/
- Objectives:
  - Multithreaded Applications: Future programs must run on multiprocessors
  - Emerging Workloads: Increasing CPU performance enables new applications
  - Diverse: Multiprocessors are being used for more and more tasks
  - State-of-Art Techniques: Algorithms and programming techniques evolve rapidly

Example: PARSEC

Program	Application Domain	Parallelization
Blackscholes	Financial Analysis	Data-parallel
Bodytrack	Computer Vision	Data-parallel
Canneal	Engineering	Unstructured
Dedup	Enterprise Storage	Pipeline
Facesim	Animation	Data-parallel
Ferret	Similarity Search	Pipeline
Fluidanimate	Animation	Data-parallel
Freqmine	Data Mining	Data-parallel
Streamcluster	Data Mining	Data-parallel
Swaptions	Financial Analysis	Data-parallel
Vips	Media Processing	Data-parallel
X264	Media Processing	Pipeline

Example: Rodinia

- A Benchmark Suite for Heterogeneous Computing: multicore CPU and GPU
- University of Virginia

Application / Kernel	Dwarf	Domain
K-means	Dense Linear Algebra	Data Mining
Needleman-Wunsch	Dynamic Programming	Bioinformatics
HotSpot*	Structured Grid	Physics Simulation
Back Propagation*	Unstructured Grid	Pattern Recognition
SRAD	Structured Grid	Image Processing
Leukocyte Tracking	Structured Grid	Medical Imaging
Breadth-First Search*	Graph Traversal	Graph Algorithms
Stream Cluster*	Dense Linear Algebra	Data Mining
Similarity Scores*	MapReduce	Web Mining

Conclusions

- Performance evaluation is very important to assess programming quality as well as the underlying architecture and how they interact.
- The following capture some aspects of the system but do not represent overall performance: MIPS, #instructions, #cycles, frequency
- Execution time is what matters: system time, CPU time, I/O and memory time
  - To know whether your execution time is good, you need to compare it with sequential code, another parallel code, etc.
- Scalability and efficiency measure the quality of your code.