



Chapter 1



The Role of Performance Measurement

Performance

What does it mean?

- ▶ Purchasing perspective
 - ▶ given a collection of machines, which has the
 - ▶ best performance ?
 - ▶ least cost ?
 - ▶ best performance / cost ?
- ▶ Design perspective
 - ▶ faced with design options, which has the
 - ▶ best performance improvement ?
 - ▶ least cost ?
 - ▶ best performance / cost ?
- ▶ Both require
 - ▶ basis for comparison
 - ▶ metric for evaluation
- ▶ Our goal is to understand cost & performance implications of architectural choices

Two Notions of performance

Airplane	Passenger Capacity	Cruising range	Cruising speed	Passenger Throughput
Boeing 777	375	4630	610	228750
Boeing 747	470	4150	610	286700
BAC/Sud Concorde	132	4000	1350	178200
Douglas DC-8-50	146	8720	544	79424

- ▶ Which has higher Performance?
- ▶ Response Time
 - ▶ Time to do a task
 - ▶ execution time, response time, latency
- ▶ Throughput
 - ▶ Task per time
 - ▶ throughput, bandwidth
- ▶ Response Time and Throughput are often in opposition

The winner?

Airplane	Passenger Capacity	Cruising range	Cruising speed	Passenger Throughput
Boeing 777	375	4630	610	228750
Boeing 747	470	4150	610	286700
BAC/Sud Concorde	132	4000	1350	178200
Douglas DC-8-50	146	8720	544	79424

- ▶ If we define performance by speed, we have two possibilities:
 - ▶ Highest cruising speed -> Concorde wins
 - ▶ Taking a single passenger with the least time -> 747 wins
- ▶ Performance is defined by many parameters
- ▶ The same with computers
 - ▶ Reduce response time
 - ▶ Increase throughput

Example

- ▶ Do the following changes to a computer system increase throughput, decrease response time, or both?
 - ▶ Replacing with faster processor
 - ▶ Adding an additional processor
- ▶ Case 1: reducing response time will increase throughput
 - ▶ -> Both
- ▶ Case 2: adding throughput reducing waiting time (response time)
 - ▶ -> Both

Definition

- ▶ Performance is in units of things-per-second
 - ▶ bigger is better
- ▶ If we are primarily concerned with response time

$$\text{Performance}_x = \frac{1}{\text{Execution time}_x}$$

- ▶ How to read:
 - ▶ Performace of Machine X
 - ▶ Execution time of Machine X

Performance Comparison

- ▶ Greater Than or Less Than

$$\text{Performance}_x < \text{Performance}_y$$

$$\frac{1}{\text{Execution time}_x} < \frac{1}{\text{Execution time}_y}$$

$$\text{Execution time}_x > \text{Execution time}_y$$

Example

Airplane	Passenger Capacity	Cruising range	Cruising speed	Passenger Throughput
Boeing 777	375	4630	610	228750
Boeing 747	470	4150	610	286700
BAC/Sud Concorde	132	4000	1350	178200
Douglas DC-8-50	146	8720	544	79424

- ▶ Time of Concorde vs. Boeing 747?
 - ▶ Concorde is $1350 \text{ mph} / 610 \text{ mph} = 2.2$ times faster
- ▶ Throughput of Concorde vs. Boeing 747 ?
 - ▶ Concorde is $178,200 \text{ pmph} / 286,700 \text{ pmph} = 0.62$ “times faster”
 - ▶ Boeing is $286,700 \text{ pmph} / 178,200 \text{ pmph} = 1.6$ “times faster”
- ▶ Boeing is 1.6 times (“60%”) faster in terms of throughput
- ▶ Concorde is 2.2 times (“120%”) faster in terms of flying time
- ▶ We will focus primarily on execution time for a single job

Performance Relation

- ▶ Machine X is n times faster than Machine Y

$$\frac{\text{Performance}_x}{\text{Performance}_y} = n$$

$$\frac{\text{Performance}_x}{\text{Performance}_y} = \frac{\text{Execution time}_y}{\text{Execution time}_x} = n$$

Example

- ▶ Machine P runs a program in 20 seconds and Machine Q runs the same program in 15 seconds
 - ▶ How much faster is machine Q than machine P?

- ▶ We know Q is n times faster than P

$$\frac{\text{Performance}_Q}{\text{Performance}_P} = n$$

$$\frac{\text{Execution time}_P}{\text{Execution time}_Q} = n$$

- ▶ Thus the performance ratio is $20/15 = 1.33..$
 - ▶ And Q is 1.33.. Times faster than P

Measuring Performance

- ▶ **Time is the measure of computer performance**
- ▶ The computer that perform the same amount of work in the least time is the fastest
- ▶ Program *execution time* is seconds per program
- ▶ The most straightforward is
 - ▶ Wall clock time
 - ▶ Response time
 - ▶ Elapsed time

What is execution time or elapsed time?

- ▶ **Problem: Computer are often time shared**
- ▶ Distinguish between elapsed time and CPU time.
 - ▶ CPU time is the time the processor is working on our program (does not include time spent on I/O or other program)
 - ▶ CPU time can be divided into
 - ▶ User CPU time
 - ▶ System CPU time
 - ▶ Difficult to measure
- ▶ Performance
 - ▶ CPU performance
 - ▶ System performance

Example

- ▶ Unix time for a task or program
 - ▶ 90.7u 12.9s 2:39 65%
 - ▶ User CPU time is 90.7 seconds
 - ▶ System CPU time is 12.9 seconds
 - ▶ Elapsed time is 2 minutes 39 seconds (159 seconds)
 - ▶ The percentage of the elapsed time that is the CPU time is 65%

$$\frac{90.7 + 12.9}{159} = 0.65$$

- ▶ 35 % is spent on I/O and other programs

Clock cycle

Almost all computer runs at a constant rate clock

- ▶ Other name for **clock cycles** : ticks, clock ticks, clock periods, clock, cycles.
- ▶ Clock period is the inverse of clock cycle
 - ▶ Ex: 2 ns clock period is 500 MHz clock cycle

Relating the metric

CPU execution time = CPU clock cycle × clock cycle time

$$\text{CPU execution time} = \frac{\text{CPU clock cycle}}{\text{clock rate}}$$

or

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

- ▶ Hardware designer can improve performance by reducing
 - ▶ the length of the clock cycle or
 - ▶ the number of clock cycle per program

Example 1

Improving performance

- ▶ Machine A which has 500 MHz clock runs a program in 5 seconds
 - ▶ What is the CPU cycle of machine A?
 - ▶ We improve machine A with a new machine B which has 750 MHz clock. Assuming the same clock cycle, how long does the same program runs on B?
 - ▶ We improve machine A with a new machine C which has 1000 MHz clock but the number of cycle is 1.3 times the number of cycle of machine A. How long does the same program runs on C?

Answer

$$CPU\ time_A = \frac{CPU\ clock\ cycle_A}{Clock\ rate_A}$$

$$5 = \frac{CPU\ clock\ cycle_A}{500 \times 10^6}$$

$$CPU\ clock\ cycle_A = 5 \times 500 \times 10^6 = 2500 \times 10^6\ cycle$$

Answer

- ▶ CPU time for machine B:

$$CPUtime_B = \frac{CPU\ clock\ cycle_A}{Clock\ rate_B}$$

$$CPUtime_B = \frac{2500 \times 10^6}{750 \times 10^6} = 3.3333 \text{ seconds}$$

- ▶ CPU time for machine C:

$$CPUtime_C = \frac{1.3 \times CPU\ clock\ cycle_A}{Clock\ rate_C}$$

$$CPUtime_C = \frac{1.3 \times 2500 \times 10^6}{1000 \times 10^6} = 3.25 \text{ seconds}$$

Example 2

- ▶ Machine A which has 500 MHz clock runs a program in 10 seconds.
 - ▶ We want to build a machine that will run the same program in 6 seconds. What is the clock rate of a new machine D if the clock cycle is increased by 1.2 times?

Example 2

- ▶ Machine A which has 500 MHz clock runs a program in 10 seconds.
- ▶ We want to build a machine that will run the same program in 6 seconds. What is the clock rate of a new machine D if the clock cycle is increased by 1.2 times?

$$CPUtime_A = \frac{CPU\ clock\ cycle_A}{Clock\ rate_A}$$

$$10 = \frac{CPU\ clock\ cycle_A}{500 \times 10^6}$$

$$CPU\ clock\ cycle_A = 5000 \times 10^6\ cycles$$

$$CPUtime_D = \frac{1.2 \times CPU\ clock\ cycle_A}{Clock\ rate_D}$$

$$6 = \frac{1.2 \times 5000 \times 10^6}{Clock\ rate_D}$$

$$Clock\ rate_D = 1000 \times 10^6 = 1\ GHz$$

Hardware Software Interface

- ▶ **Execution must depends on the number of instruction per program**
- ▶ Compiler generated the instructions to be execute and the machine had to execute the instructions to run the program

$$\text{CPU clock cycle} = \text{Instructions for a program} \times \text{Average clock cycle per instruction}$$

- ▶ The average number of cycles per instruction is abbreviated as CPI - clock cycles per instruction

Example

- ▶ **Suppose we have two machine with the same ISA**
 - ▶ Machine A: clock cycle 1.5 ns and CPI 2
 - ▶ Machine B: clock cycle 2ns and CPI 1.75
 - ▶ Which one is faster and by how much?

Answer

- ▶ CPU cycles

$$CPU\ clock\ cycle_A = 1 \times 2$$

$$CPU\ clock\ cycle_B = 1 \times 1.75$$

- ▶ CPU time

$$CPU\ time_A = CPU\ clock\ cycle_A \times Clock\ cycle\ time_A$$

$$CPU\ time_A = 2 \times 1 \times 1.5\ ns = 3.0 \times 1\ ns$$

$$CPU\ time_B = 1.75 \times 1 \times 2\ ns = 3.5 \times 1\ ns$$

Answer

- ▶ Comparison

$$\frac{CPU\ performance_A}{CPU\ performance_B} = \frac{Execution\ time_B}{Execution\ time_A}$$

$$\frac{Execution\ time_B}{Execution\ time_A} = \frac{3.5 \times Ins}{3.0 \times Ins} = 1.167$$

- ▶ Machine A is 1.167 faster than machine B for this program

Performance equation

$$\text{CPU time} = \text{Instruction count} \times \text{CPI} \times \text{Clock cycle time}$$

$$\text{CPU time} = \frac{\text{Instruction count} \times \text{CPI}}{\text{Clock rate}}$$

$$\text{Time} = \frac{\text{Instructions}}{\text{Program}} \times \frac{\text{Clock cycles}}{\text{Instruction}} \times \frac{\text{Seconds}}{\text{Clock cycle}}$$

Aspect of CPU performance

$$Time = \frac{Instructions}{Program} \times \frac{Clock\ cycles}{Instruction} \times \frac{Seconds}{Clock\ cycle}$$

	Instruction Count	CPI	Clock rate
Program	x		
Compiler	x	x	
ISA	x	x	
Organization		x	x
Technology			x

How do we obtain these numbers?

- ▶ We can measure CPU execution time
- ▶ We can get clock cycle time
- ▶ Instruction count and CPI are very difficult to obtain
- ▶ Instruction count:
 - ▶ Profiler
 - ▶ Trace
 - ▶ Simulator
- ▶ CPI
 - ▶ Detail simulation
 - ▶ Hand count clock cycle for each instruction

CPI

- ▶ **Several different classes of instructions**
- ▶ **n** many instruction classes
- ▶ C_i is the count of the number of instructions of class **i** executed
- ▶ CPI_i is the average number of cycles per instruction in class **i**
- ▶ CPU clock cycles

$$\begin{aligned} \text{CPU clock cycle} &= \sum_{i=1}^n CPI_i \times C_i \\ &= CPI_1 \times C_1 + CPI_2 \times C_2 + CPI_3 \times C_3 + \dots + CPI_n \times C_n \end{aligned}$$

Example

- ▶ Machine facts

Instruction Class	CPI for this class
A	1
B	2
C	3

- ▶ A compiler generates two code sequence

Instruction Count for instruction class			
Code	A	B	C
1	2	1	2
2	4	1	1

- ▶ Which code sequence has the most instructions?
- ▶ Which one is faster?
- ▶ What is the CPI?

Answer

- ▶ Code Sequence

- ▶ Sequence 1 : $2 + 1 + 2 = 5$ instructions
- ▶ Sequence 2 : $4 + 1 + 1 = 6$ instructions
- ▶ Sequence 2 has more instructions

- ▶ CPU clock cycles

$$\begin{aligned} \text{CPU clock cycle} &= \sum_{i=1}^3 CPI_i \times C_i \\ &= CPI_1 \times C_1 + CPI_2 \times C_2 + CPI_3 \times C_3 \end{aligned}$$

- ▶ Sequence 1 : $(2 \times 1) + (1 \times 2) + (2 \times 3) = 10$ cycles
- ▶ Sequence 2 : $(4 \times 1) + (1 \times 2) + (1 \times 3) = 9$ cycles
- ▶ Sequence 2 is faster

Answer

► CPI

$$CPI = \frac{CPU\ clock\ cycles}{Instruction\ Count}$$

$$CPI_1 = \frac{CPU\ clock\ cycles_1}{Instruction\ Count_1} = \frac{10}{5} = 2$$

$$CPI_2 = \frac{CPU\ clock\ cycles_2}{Instruction\ Count_2} = \frac{9}{6} = 1.5$$

A Simple Example

- ▶ $C_i = \text{Frequency}$

Operation	Freq	CPI_i	Freq x CPI_i
ALU	50%	1	0.5
Load	20%	5	1
Store	10%	3	0.3
Branch	20%	2	0.4
		Σ	2.2

A Simple Example

▶ Machine A:

- ▶ How much faster would the machine be if a better data cache reduced the average load time to 2 cycles?

▶ Machine B:

- ▶ How does this compare with using branch prediction to shave a cycle off the branch time?

	Original Machine			Machine A		Machine B	
Operation	Freq	CPI _i	Freq x CPI _i	CPI _i	Freq x CPI _i	CPI _i	Freq x CPI _i
ALU	50%	1	0.5				
Load	20%	5	1				
Store	10%	3	0.3				
Branch	20%	2	0.4				
		Σ =	2.2				

A Simple Example

▶ Machine A:

- ▶ How much faster would the machine be if a better data cache reduced the average load time to 2 cycles?

▶ Machine B:

- ▶ How does this compare with using branch prediction to shave a cycle off the branch time?

	Original Machine			Machine A		Machine B	
Operation	Freq	CPI _i	Freq x CPI _i	CPI _i	Freq x CPI _i	CPI _i	Freq x CPI _i
ALU	50%	1	0.5	1	0.5	1	0.5
Load	20%	5	1	2	0.4	5	1
Store	10%	3	0.3	3	0.3	3	0.3
Branch	20%	2	0.4	2	0.4	1	0.2
		Σ =	2.2	Σ =	1.6	Σ =	2

Choosing Programs to Evaluate Performance

- ▶ Workload is a set of application programs that the machine runs to measure performance
- ▶ Benchmark is a set of programs specifically chosen for measuring performance

	PROs	CONs
Actual Target Workload	representative	very specific non-portable difficult to run, or measure hard to identify cause
Full Benchmarks	Portable Widely used Improvements useful in reality	Less representative
Small “Kernel” Benchmarks	Easy to run early in design cycle	Easy to fool
Micro benchmarks	Identify peak capability and potential bottleneck	Peak may be a long way from application performance

SPEC Benchmarks www.spec.org

Integer benchmarks		FP benchmarks	
gzip	compression	wupwise	Quantum chromodynamics
vpr	FPGA place & route	swim	Shallow water model
gcc	GNU C compiler	mgrid	Multigrid solver in 3D fields
mcf	Combinatorial optimization	applu	Parabolic/elliptic pde
crafty	Chess program	mesa	3D graphics library
parser	Word processing program	galgel	Computational fluid dynamics
eon	Computer visualization	art	Image recognition (NN)
perlbmk	perl application	equake	Seismic wave propagation simulation
gap	Group theory interpreter	facerec	Facial image recognition
vortex	Object oriented database	ammp	Computational chemistry
bzip2	compression	lucas	Primality testing
twolf	Circuit place & route	fma3d	Crash simulation fem
		sixtrack	Nuclear physics accel
		apsi	Pollutant distribution

SPEC CPU Benchmark


- ▶ Programs used to measure performance
 - ▶ Supposedly typical of actual workload
- ▶ Standard Performance Evaluation Corp (SPEC)
 - ▶ Develops benchmarks for CPU, I/O, Web, ...
- ▶ SPEC CPU2006
 - ▶ Elapsed time to execute a selection of programs
 - ▶ Negligible I/O, so focuses on CPU performance
 - ▶ Normalize relative to reference machine
 - ▶ Summarize as geometric mean of performance ratios
 - ▶ CINT2006 (integer) and CFP2006 (floating-point)

$$\sqrt[n]{\prod_{i=1}^n \text{Execution time ratio}_i}$$

CINT2006 for Opteron X4 2356

Name	Description	ICx10 ⁹	CPI	Tc (ns)	Exec time	Ref time	SPECratio
perl	Interpreted string processing	2,118	0.75	0.40	637	9,777	15.3
bzip2	Block-sorting compression	2,389	0.85	0.40	817	9,650	11.8
gcc	GNU C Compiler	1,050	1.72	0.47	24	8,050	11.1
mcf	Combinatorial optimization	336	10.00	0.40	1,345	9,120	6.8
go	Go game (AI)	1,658	1.09	0.40	721	10,490	14.6
hmmer	Search gene sequence	2,783	0.80	0.40	890	9,330	10.5
sjeng	Chess game (AI)	2,176	0.96	0.48	37	12,100	14.5
libquantum	Quantum computer simulation	1,623	1.61	0.40	1,047	20,720	19.8
h264avc	Video compression	3,102	0.80	0.40	993	22,130	22.3
omnetpp	Discrete event simulation	587	2.94	0.40	690	6,250	9.1
astar	Games/path finding	1,082	1.79	0.40	773	7,020	9.1
xalancbmk	XML parsing	1,058	2.70	0.40	1,143	6,900	6.0
Geometric mean							11.7

High cache miss rates



SPEC Power Benchmark

- ▶ Power consumption of server at different workload levels
 - ▶ Performance: ssj_ops/sec
 - ▶ Power: Watts (Joules/sec)

$$\text{Overall ssj_ops per Watt} = \left(\sum_{i=0}^{10} \text{ssj_ops}_i \right) / \left(\sum_{i=0}^{10} \text{power}_i \right)$$

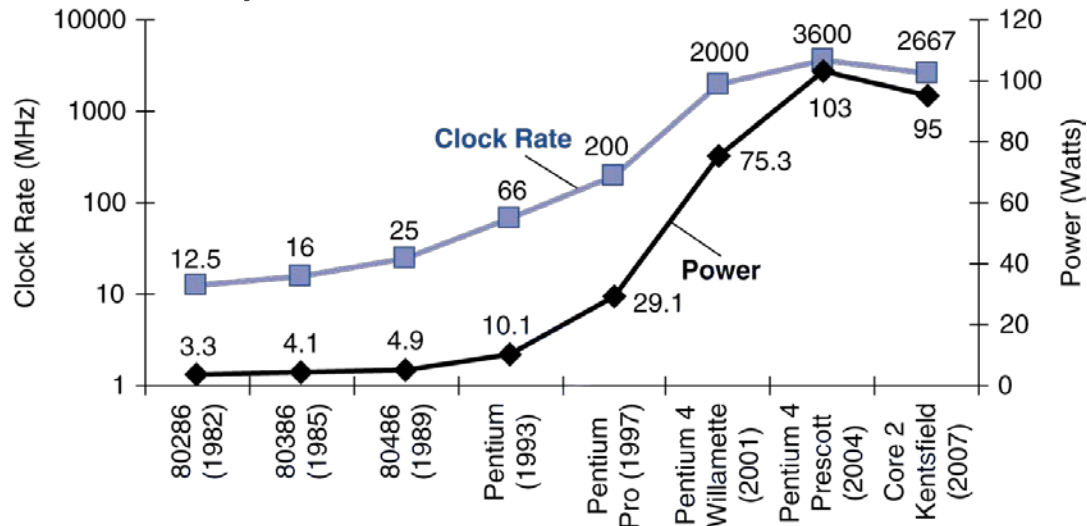
SPECpower_ssj2008 for X4

Target Load %	Performance (ssj_ops/sec)	Average Power (Watts)
100%	231,867	295
90%	211,282	286
80%	185,803	275
70%	163,427	265
60%	140,160	256
50%	118,324	246
40%	92,035	233
30%	70,500	222
20%	47,126	206
10%	23,066	180
0%	0	141
Overall sum	1,283,590	2,605
Σ ssj_ops/ Σ power		493



Power Trends

- ▶ Power consumption – especially in the embedded market where battery life is important (and passive cooling)
- ▶ For power-limited applications, the most important metric is energy efficiency



$$\text{Power} = \text{Capacitive load} \times \text{Voltage}^2 \times \text{Frequency}$$

x 30

5v → 1V

x 1000

Reducing Power

- ▶ Suppose a new CPU has
 - ▶ 85% of capacitive load of old CPU
 - ▶ 15% voltage and 15% frequency reduction

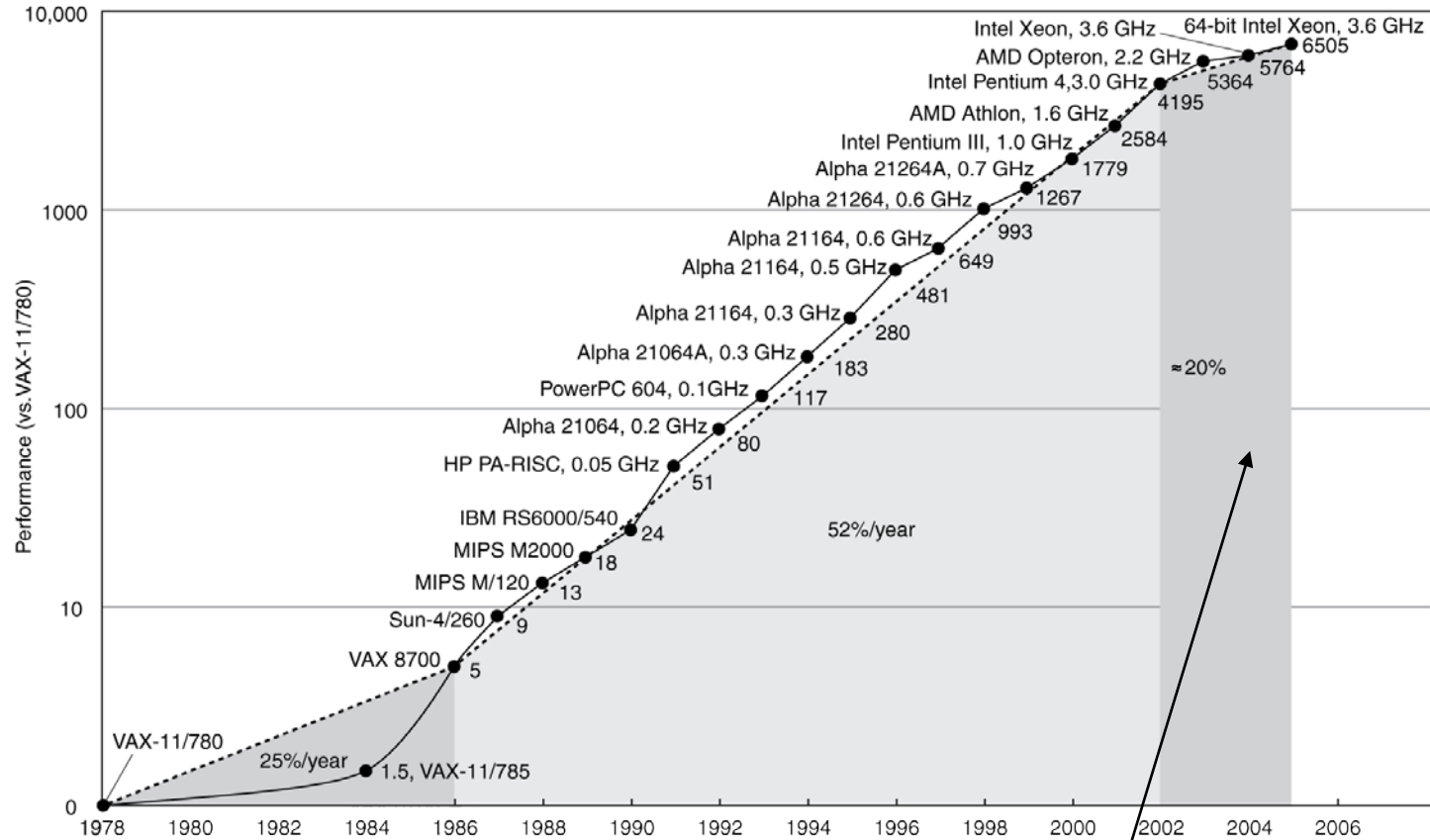
$$\frac{P_{\text{new}}}{P_{\text{old}}} = \frac{C_{\text{old}} \times 0.85 \times (V_{\text{old}} \times 0.85)^2 \times F_{\text{old}} \times 0.85}{C_{\text{old}} \times V_{\text{old}}^2 \times F_{\text{old}}} = 0.85^4 = 0.52$$

- ▶ The power wall
 - ▶ We can't reduce voltage further
 - ▶ We can't remove more heat
- ▶ How else can we improve performance?

Fallacy: Low Power at Idle

- ▶ Look back at X4 power benchmark
 - ▶ At 100% load: 295W
 - ▶ At 50% load: 246W (83%)
 - ▶ At 10% load: 180W (61%)
- ▶ Google data center
 - ▶ Mostly operates at 10% – 50% load
 - ▶ At 100% load less than 1% of the time
- ▶ Consider designing processors to make power proportional to load

Uniprocessor Performance



Constrained by power, instruction-level parallelism, memory latency

Multiprocessors

- ▶ **Multicore microprocessors**
 - ▶ More than one processor per chip
- ▶ **Requires explicitly parallel programming**
 - ▶ Compare with instruction level parallelism
 - ▶ Hardware executes multiple instructions at once
 - ▶ Hidden from the programmer
 - ▶ Hard to do
 - ▶ Programming for performance
 - ▶ Load balancing
 - ▶ Optimizing communication and synchronization

Amdahl's Law

- ▶ Speedup : how a machine performs after enhancement
- ▶ Law of diminishing returns

$$\text{Speedup}(E) = \frac{\text{Performance with } E}{\text{Performance without } E} = \frac{\text{Execution time without } E}{\text{Execution time with } E}$$

$$\text{Execution time}(E) = \text{Execution time unaffected} + \frac{\text{Execution time with } E}{\text{Amount of Improvement}}$$

Example 1

- ▶ A program runs on a machine for 10 seconds. 50 % of the time is doing multiplications. If we improve the multiplication unit so that it runs twice as fast, how big is the speedup?

Example 1

- ▶ A program runs on a machine for 10 seconds. 50 % of the time is doing multiplications. If we improve the multiplication unit so that it runs twice as fast, how big is the speedup?

$$Extime(E) = \frac{\textit{Affected extime}}{\textit{improvement}} + \textit{unaffected extime}$$

$$Extime(E) = \frac{5s}{2} + 5s = 7.5s$$

$$Speedup(E) = \frac{10s}{7.5s} = 1.3333$$

- ▶ Not two times faster

Example 2

- ▶ A program runs for 10 seconds. 70% of the time is doing additions. How much improvement on the additions if we want to reduce the running time to 3 seconds?

Example 2

- ▶ A program runs for 10 seconds. 70% of the time is doing additions. How much improvement on the additions if we want to reduce the running time to 3 seconds?

$$Extime(E) = \frac{\textit{Affected ex\ time}}{\textit{improvement}} + \textit{unaffected ex\ time}$$

$$3s = \frac{7s}{n} + (10 - 7)s$$

$$3s = \frac{7s}{n} + 3s$$

$$0 = \frac{7s}{n}$$

- ▶ No amount of improvement can reduce the running time to 3 seconds.

MIPS

- ▶ Instruction Rate

$$MIPS = \frac{InstructionCount}{Executiontime \times 10^6}$$

- ▶ Faster machine have higher MIPS rating (?)

Example

Instruction Class	CPI for this class
A	1
B	2
C	3

	Instruction count (billions)		
Code from	A	B	C
Compiler 1	5	1	1
Compiler 2	10	1	1

- ▶ Assume the machine is running at 500 Mhz.
 - ▶ Which one is faster according to execution time?
 - ▶ Which one is faster according to MIPS?

Answer

▶ Execution Time

$$\text{execution time} = \frac{\text{CPU clock cycle}}{\text{clock rate}}$$

$$\text{CPU clock cycle} = \sum_{i=1}^n \text{CPI}_i \times C_i$$

- ▶ CPU clock cycle₁ = (5x1)+(1x2)+(1x3)x10⁹ = 10x10⁹
- ▶ CPU clock cycle₂ = (10x1)+(1x2)+(1x3)x10⁹ = 15x10⁹
- ▶ Execution time₁ = (10x10⁹)/(500x10⁶) = 20 s
- ▶ Execution time₂ = (15x10⁹)/(500x10⁶) = 30 s
- ▶ Compiler 1 produces a faster program

Answer

▶ MIPS

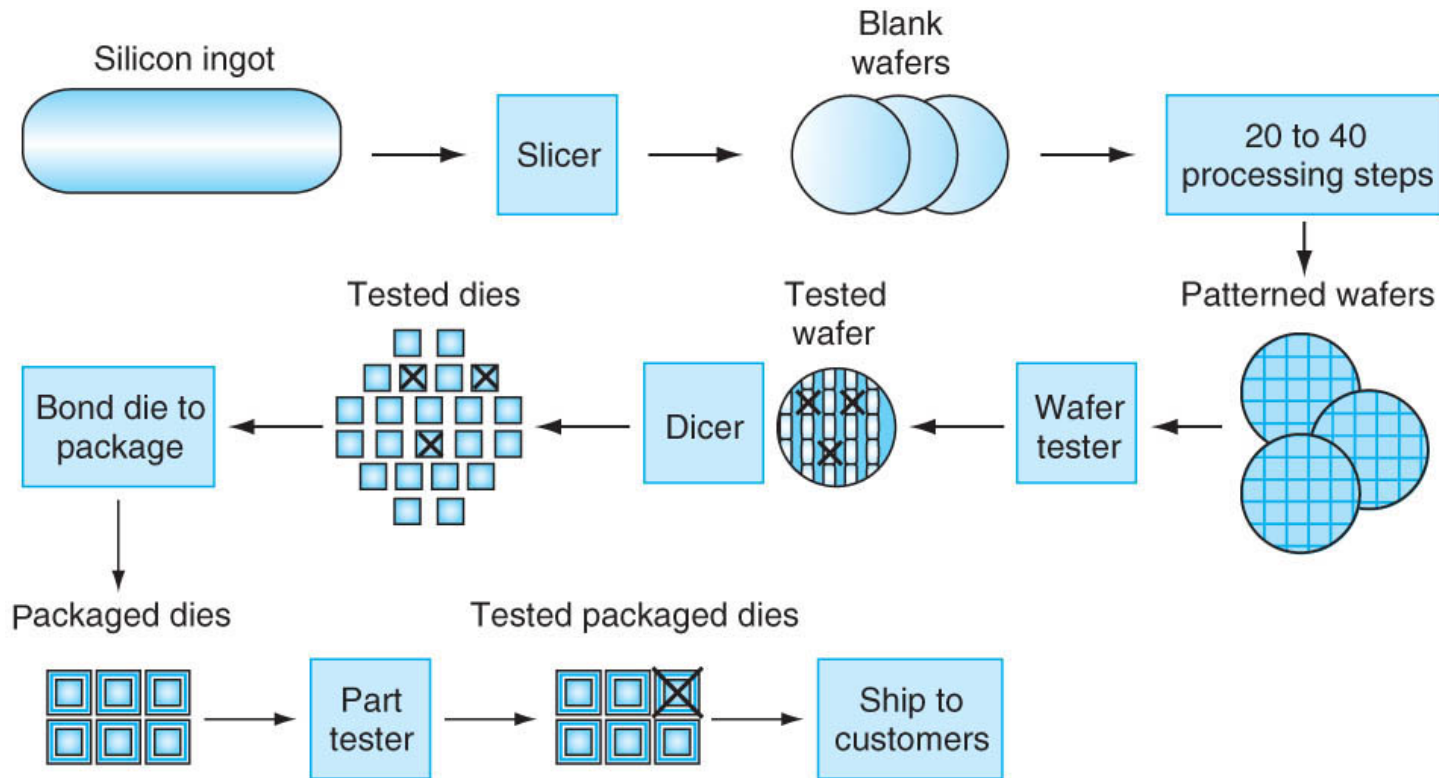
$$MIPS = \frac{InstructionCount}{Executiontime \times 10^6}$$

$$MIPS_1 = \frac{(5 + 1 + 1) \times 10^9}{20 \times 10^6} = 350$$

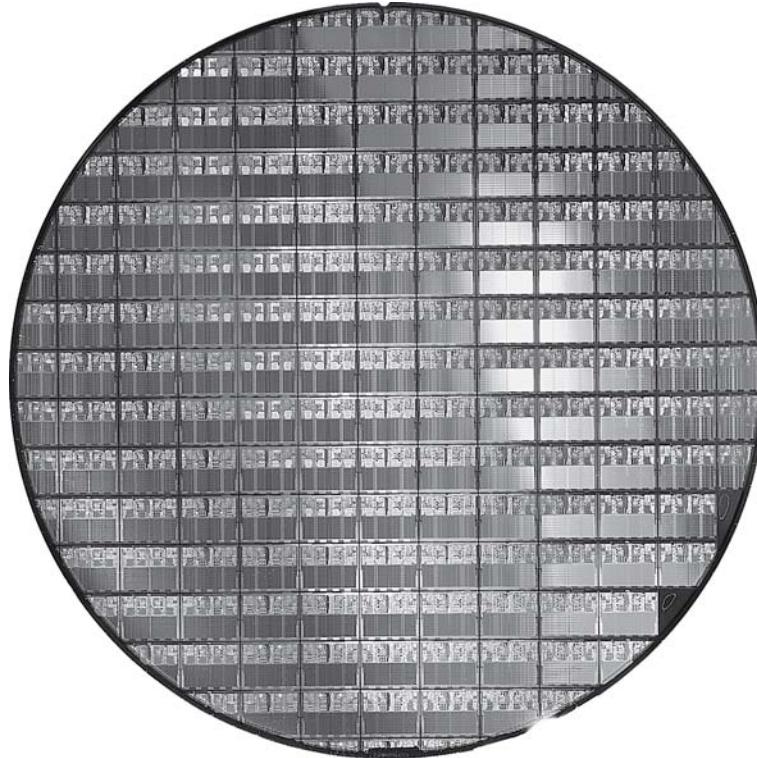
$$MIPS_1 = \frac{(10 + 1 + 1) \times 10^9}{30 \times 10^6} = 400$$

▶ Compiler 2 is faster -> MIPS fails

The chip manufacturing process



AMD Opteron X2 Wafer



- ▶ X2: 300mm wafer, 117 chips, 90nm technology
- ▶ X4: 45nm technology

Real Stuff: Manufacturing Pentium

- ▶ There are 196 Pentium dies in an 8-inch wafer
- ▶ Only 76 Pentium Pro dies in an 8-inch wafer
- ▶ Fewer dies -> higher cost
- ▶ Cost increased further because larger die is much more likely to contain defects

$$\text{Cost per die} = \frac{\text{Cost per wafer}}{\text{Dies per wafer} \times \text{yield}}$$

$$\text{Dies per wafer} = \frac{\text{Wafer area}}{\text{Die area}}$$

$$\text{Yield} = \frac{1}{\left(1 + \left(\text{Defects per area} \times \frac{\text{Die area}}{2}\right)\right)^2}$$