Homework 1 Solution
CPSC 418, Fall 2001

1 Scaling laws (30 points)

Compound interest is computed $F_n = Pe^{rn}$, where n is the number of periods, F_n is the future value after n periods, r is the rate per period, and P is the present value. Rewriting to solve for the rate: $r = \frac{\log(F_n/P)}{n}$. And, given a rate, the number of periods to achieve a certain growth can be determined: $n = \frac{\log(F_n/P)}{r}$.

a) Memory bandwidth (5 points)

"Over the same period [1989 to 1998], the peak burst bandwidth of the DRAM sub-system has also improved by **12x**…." [4] That gives a rate $r = \frac{\log 12}{9 \text{ years}} = 0.276/\text{year}$.

$2\times$	2.5 years
$10\times$	8.3 years
$1000\times$	25 years

b) Memory latency (5 points)

Also from *Tom's Hardware* [4], in the same period for latency $F_n = (400 \times 28)^{-1}$ ms and $P = (33 \times 5)^{-1}$ ms, giving $r = -0.469/\text{year}$.

$1/2$	1.5 years
$1/10$	4.9 years
$1/1000$	15 years

c) Population of Canada (5 points)

According to Statistics Canada [6], the population has grown from 29.67 million in 1996 to 30.75 million in 2000, for a rate of 1.92% per year. According to the CIA [7], the growth rate is 0.99% per year.

	StatCan	CIA
$2\times$	36 years	70 years
$10\times$	120 years	230 years
$1000\times$	360 years	700 years

d) Population of [insert low immigration country here] (5 points)

Again according to the CIA [7], China currently has a population growth rate of 0.88% per year, and the United Kingdom has a growth rate of 0.23% per year.

	China	U.K.
$2\times$	79 years	300 years
$10\times$	520 years	1,000 years
$1000\times$	780 years	3,000 years

e) Compare 1.a to 1.b (5 points)

Memory bandwidth is increasing at roughly the same rate as processor frequency. Memory latency is also increasing, although slowly. How an architecture deals with memory latency will be the most important factor affecting memory system performance. It is important that cache miss rates continue to decline, through improved caching and/or improved code. Avoiding pipeline stalls due to unresolved data dependencies will be critical.

f) Regression from 1.d (5 points)

Country	Pop. [7]	Year of Pop. 1
China	1.27 B	382 BCE
U.K.	59.6 M	5784 BCE

Clearly China must be growing more quickly today than it has in the past, as China was populated well before 382 BCE. On the other hand, England is growing more slowly today than it has in the past, as English civilization does not go back 7,800 years.

Another conclusion that can be drawn from this exercise is not to take growth rates too seriously. Extrapolating to thousands of years from a fit based on one or only a few years is obviously problematic. Growth occurs for a reason, and those reasons inevitably change over time.

2 Game consoles (30 points)

Answer at least three:

a) X-Box availability

The X-Box was to be released on November 8, 2001. Recently (but before the assignment was due) that date was changed to November 15, 2001.

b) Current PS2 market

Interpreting "market" as "those who will buy it," the PS2 is a top of the line game console (so people wanting the newest and best toys), which is also capable of playing older PS games (so people who already have an investment in PS titles, but want to play newer titles as well).

Other reasonable interpretations of "market," e.g. "market share," will also be accepted.

c) Change since papers?

The X-Box specs are still changing slightly, including processor frequencies (both CPU and graphics), disk capacity, and DVD speed. The paper [2] only speculates on many of the details of the X-Box. The PS2 on the other hand was in basically final form when the PS2 paper was published.

It is unlikely that the X-Box will undergo significant changes after its initial release, nor is any significant change in the PS2 likely. In the console game market, software makers expect the hardware to be an appliance which is the same everywhere. Add-ons may be introduced, but even seemingly minor changes, such as an increase in CPU clock rate, may adversely affect some existing software.

From the PS2 paper,"The console cannot change its functions or performance during its lifetime" [5]. Similarly for the X-Box, "...all such systems must behave exactly like a standard X-Box when playing X-Box games to ensure software compatibility" [2]. For example,"When a DVD containing an X-Box game is inserted, the system will be initialized from a memory image stored on the disc..." [2]. For that memory image to be valid, the underlying hardware must always be the same.

d) Good architecture $\Rightarrow$ market success?

Relative to previous generation consoles, the architectural advances have been significant, with a direct impact on performance. Many owners of older consoles have purchased or will purchase the newer consoles, because of improved performance. Faster hardware means cooler games, and people will pay to play cooler games. The performance gap between the PS2 and X-Box is much narrower, and is unlikely to affect relative sales significantly.

e) Anything else interesting?

Credit will be given for anything both nontrivial and true.

3 PowerPC III vs Pentium 4 (35 points)

Compare the IBM RS64 III to the Pentium 4.[1]

a) Workload (5 points)

The Pentium 4 is tailored for multimedia type applications [3], while the IBM RS64 III targets more traditional productivity and commercial type software [1]. These targets are obviously motivated by the class of computers the respective chips will power. Pentium chips can be found in a large percentage of desktop machines, while the IBM chips are destined for commercial compute servers.

[1]The answers to this question should look extremely familiar to one student in the class. Thank you for "allowing" me to borrow them.

b) Affect on pipeline design (10 points)

The pipeline in the Pentium 4 is deep, allowing for very fast clocking. Since multimedia applications typically enter loops that execute many iterations, the cost of the branch mispredict in the long pipeline is seldomly incurred. This is contrasted by the RS64 III, which has a pipeline of only 5 stages, a quarter that of the Intel chip. The short pipeline coupled with some extra control logic allows for a branch misprediction penalty of at most 1 clock cycle. This is an asset given that the target productivity/commercial software contain a high proportion of conditional (rather than loop control) branches, which are hard for branch predictors.

c) Affect on memory system design (10 points)

The Pentium 4's L2 cache has 128 byte lines, though there is a valid bit for each half line. This supports the wide load and store operations found in multimedia apps.

In multiprocessor systems, transaction processing manifests copious data sharing between processors. In support of such processing the RS64 III provides a mechanism for transfer of cache lines between processors, which lowers cache miss latencies. Also, the L2 cache in the RS64 III is 4-way set associative, which enjoys a high hit rate in commerical apps.

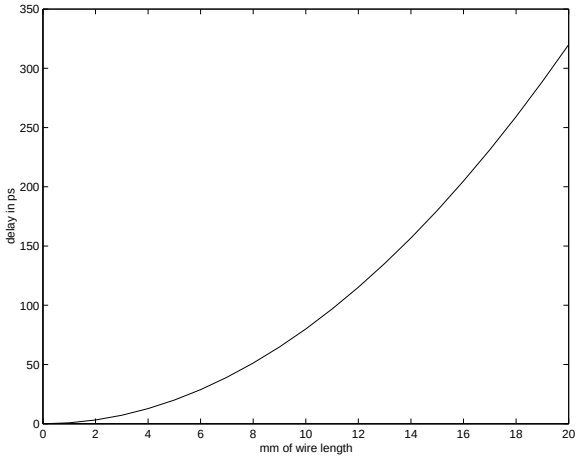
d) Other affects (10 points)

- The RS64 III provides flexibility and scalability with respect to varying multiprocessor architectures.
- The RS64 III implements a single-error correcting code in its on-chip memory arrays, including the L1 cache. This protection is motivated by the importance of data integrity in commercial computing.
- The Pentium 4 has added 144 new SSE2 (Streaming Single instruction multiple data Extension engine) instructions, which extend the existing multimedia-enhancing MMX and SSE instructions to 128 bits.

4 Wire delay (35 points)

a) Delays for lengths (10 points)

1 mm	0.8 ps
5 mm	20 ps
10 mm	80 ps
15 mm	180 ps



b) Buffers on a 15 mm wire (10 points)

You should use one buffer. See next part for derivation, or observe:

bufs	delay
0	180 ps
1	140 ps
2	160 ps
3	195 ps

c) Buffers on wires of arbitrary length (10 points)

Given R, C, x, and n as defined, let B be the buffer delay (given as 50 ps). We observe that a wire with n buffers will have $n+1$ segments. The delay D of a buffered wire of length x is therefore:

$$D(x,n) = nB + 0.4RC\left(\frac{x}{n+1}\right)^2(n+1)$$

We want to minimize D, so take the derivative:

$$\frac{\partial}{\partial n}D(x,n) = B - \frac{0.4RCx^2}{(n+1)^2}$$

Setting $\frac{\partial}{\partial n}D = 0$ and solving for n, requiring $n \geq 0$:

$$n(x) = \max\left(\frac{x\sqrt{10BRC}}{5B} - 1, 0\right)$$

Let's plug in the constants to simplify that a bit:

$$n(x) = \max\left(\sqrt{0.016}\cdot x - 1, 0\right)$$

n must also be integer. Rounding $n(x)$ to the nearest integer gives the minimum delay.

d) Proof of minimization (10 points EXTRA CREDIT)

By induction[2] on m; let L be the total wire length, and envision the wire lying on a horizontal line. When $m = 0$, the proposition is trivially true. Now assume that the proposition holds for some $m \geq 0$, and consider the placement of $m + 1$ buffers providing the minimal delay, let x be the distance from the start of the wire and the first buffer. The delay of the first wire segment and the first buffer is:

$$cx^2 + b$$

where $c = 2 \times 10^{-9}$ and $b = 0.05$ are constants. Now from our assumption the remaining m buffers must be evenly spaced on the segment of length $L - x$ between the first buffer and the end of the wire, else doing so would yield a reduced delay. We may thus express the total delay as a function of x:

$$\begin{aligned} \partial(x) &= \frac{c(L-x)^2}{m+1} + bm + cx^2 + b \\ &= \left(\frac{c}{m+1} + c\right) x^2 - \left(\frac{2cL}{m+1}\right) x + (m+1)b \end{aligned}$$

In the first line, the first term is the total delay of the $m + 1$ congruent wire segments separating the evenly spaced m buffers, while the second term is the total delay of the buffers themselves. We compute the root of the derivative:

$$\begin{aligned} \frac{d}{dx}\partial &= 2\left(\frac{c}{m+1} + c\right) x - \frac{2cL}{m+1} = 0 \\ \Rightarrow x &= \frac{L}{m+2} \end{aligned}$$

Noting that this is precisely the location of the first buffer when the $m+1$ buffers are evenly spaced, we complete the proof.

5 MOSFETs (35 points)

a) Number of silicon atoms across channel (5 points)

The density of silicon is $2.33\ \mathrm{g/cm^3}$ and the atomic weight is $28.0855\ \mathrm{g/mol}$. That gives $0.08296\ \mathrm{mol/cm^3}$, or $5 \times 10^{22}\ \mathrm{atoms/cm^3}$, for $3.68 \times 10^{7}\ \mathrm{atoms/cm}$. The channel is $0.15\ \mu\mathrm{m}$ long, or $1.5 \times 10^{-5}\ \mathrm{cm}$, for $552\ \mathrm{atoms}$ across the channel.

b) Ratio of silicon to dopant in channel (5 points)

Per $\mathrm{cm^3}$ in the p-channel region, there are $(5 \times 10^{22} - 5 \times 10^{17})$ atoms of silicon and 5×10^{17} holes, for a silicon to hole ratio of $10^5 : 1$.

[2]This solution should look familiar to the same student as in question 3.

c) Leakage probability (20 points)

Not yet available.

d) Dopant concentraion in the channel (5 points)

The dopant concentration has increased by a factor of $\frac{5\times10^{17}}{3\times10^{15}} = 167$. The dopant concentration has been increased to maintain a low level of electron leakage across the now much shorter channel.

e) Analytical solution to 5.e (10 points EXTRA CREDIT)

Not yet available.

References

[1] John Borkenhagen and Salvatore Storino. 5th generation 64-bit PowerPC-compatible processor design. `http://www.rs6000.ibm.com/resource/technology/pulsar.html`.

[2] Peter N. Glaskowsky. Microsoft weighs in with x-box: Software giant to do battle with sony, sega, and nintendo. *Microprocessor Report*, 14(4):1, 8–11, April 2000.

[3] Peter N. Glaskowsky. Pentium 4 (partially) previewed. *Microprocessor Report*, 14(8), August 2000.

[4] Bert McComas. DRAM performance: Latency vs. bandwidth. `http://www6.tomshardware.com/mainboard/98q3/980710/dram-02.html`, July 1998.

[5] Masaaki Oka and Msakazu Suzuoki. Designing and programming the emotion engine. *IEEE Micro*, 19(6):20–28, Nov/Dec 1999.

[6] *Canadian Statistics – Population, Canada, the provinces and territories*. Statistics Canada, October 2001.

[7] *The World Factbook*. United States Central Intelligence Agency, 2001.