EE 4720 **Homework 5** Solution **Due: 26 April 2002**

The following questions are based on the paper at
`http://www.intel.com/technology/itj/q12001/articles/art_2.htm`
and `http://www.ece.lsu.edu/ee4720/s/hinton_p4.pdf` (password needed off campus, will be given in class). See Homework 4 (`http://www.ece.lsu.edu/ee4720/2002/hw04.pdf`) for an introduction to the paper.

Problem 1: What is the maximum sustainable IPC of the IA-32 (in μop per cycle)? Put another way, the Pentium 4 is an n-way superscalar processor, what is n?

Maximum is 3 μPC (μop per cycle), limited by the 3-μop decode limit and also the 3-μop retire limit.

Problem 2: The Pentium 4 can decode no more than one IA-32 instruction per cycle. How then can it execute more than one IA-32 instruction per cycle (at least for small code fragments prepared by a friendly programmer)?

Decoded instructions are stored in the trace cache. A trace cache line might contain μops spanning more than one IA-32 instruction. Though it took at least two cycles to decode them, once stored in the trace cache they can be used multiple times, each time multiple IA-32 instructions are issued in one cycle.

Problem 3: One problem with superscalar systems noted in class is the wasted instructions following the delay slot of a taken branch near the beginning of a fetch group. How does the Pentium 4 avoid this?

By placing instructions in a trace cache line in dynamic order, so that the target of a branch is right after the branch, there is no need to separately fetch it.

Problem 4: The fast ($2\times$) integer ALUs have three stages, an initiation interval of 1 fast cycle ($\frac{1}{2}$ processor cycle), and a latency of zero fast cycles. Why is this surprising (not the one half part)? How does it do it?

It's surprising because normally something with three stages would have a latency of two. It works because each stage can bypass partial results (the low and high half of the result) which are enough for the dependent instruction.

Problem 5: In describing store-to-load forwarding the paper describes a special case for which data could be forwarded (bypassed) but is not because it would be too costly. Using MIPS code (or IA-32 if you prefer) provide an example of this special case.

```
sb $2, 1($3)
sb $4, 2($3)
lw $5, 0($3)
```

Problem 6: In Figure 8 the performance of a 1 GHz Pentium III is compared to a 1.5 GHz Pentium 4. Why is it reasonable for the Pentium 4 to be compared at a higher clock frequency?

Because the Pentium 4's shorter stages enable a higher clock frequency. When implemented on the same process technology, the Pentium 4 would have the higher clock frequency.