LIACS - Computerarchitectuur - 2012 - BSc 2de jaar

(Hertentamen)

U moet deel 1 beantwoorden, dan één van de 2 andere delen. Boeken en aantekeningen zijn niet toegestaan. Tentamen telt voor 30% van het eindcijfer (dan practica 50%, huiswerk 20%). Elk deel in het tentamen telt voor 15% van het eincijfer. Iedere vraag wordt aangegeven met het % van het eindcijfer tussen haakjes (bvb [x%]).

U mag vragen beantwoorden in het Engels of het Nederlands.

Tip: begin door uw tijd te verdelen tussen vragen. Bepaal dan pas de lengte van een uitleg op basis van beschikbare tijd.

Part 1 (15%)

a. Explain the difference between the units "cycles per instruction" and "instructions per cycle" when measuring the performance of a pipelined processor. [3%]

b. Explain in your own words, if possible using examples, the difference between each of the following pairs of words [3%]:

   i. hardware vs. software multithreading

   ii. out-of-order issue vs. VLIW issue

   iii. static vs. dynamic RAM (SRAM vs. DRAM)

   iv. virtually indexed caches vs. physically indexed caches

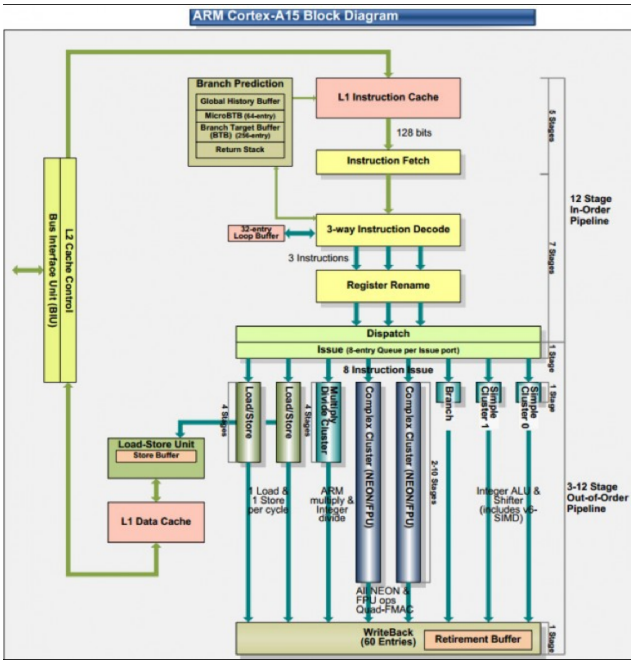
c. What is the benefit of "frequency/voltage scaling"? How can it be exploited in processor architectures? [3%]

d. Explain why, while the performance of supercomputers continues to increase exponentially over time, the performance of individual nodes in these supercomputers has stabilized since ~2004. [6%]

Part 2 (15%)

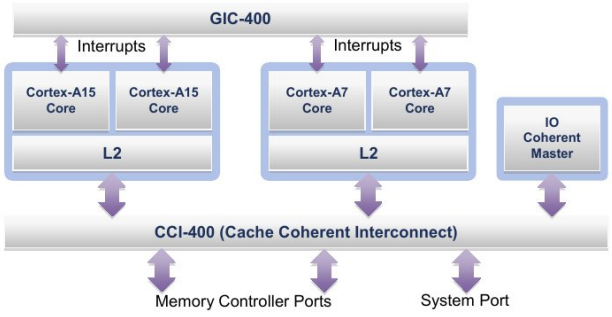
a. Explain and illustrate RAW, WAR, WAW hazards in pipelined processors. For each, explain in which kind of pipeline these hazards *cannot* be found. [5%]

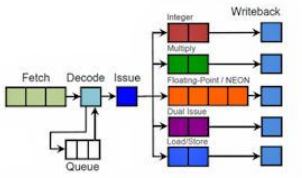
b. The ARM Cortex-A15 processor described in the diagram below uses a RISC pipeline and out-of-order execution. Determine the theoretical maximum IPC and theoretical minimum CPI for this processor. [3%].



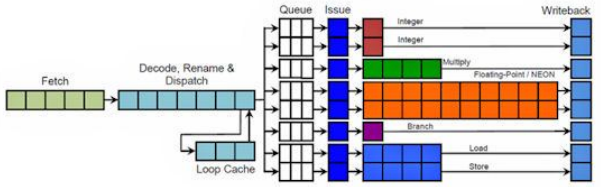
c. The Cortex-A15 core above uses instruction predication. Explain how predication works and how it helps increase execution efficiency. [2%]

d. The ARM "big.LITTLE" architecture provides two A15 and two A7 cores on the same chip (see the 2 diagrams below). The A7 core uses in-order single issue and out-of-order completion and is more energy efficient. The A15 uses out-of-order 2-way issue, out-of-order completion, with a longer pipeline and runs at a higher frequency. A video processing application is developed to run on this chip. It contains multiple software algorithms running in a data pipeline. How do you choose which algorithms to run on the A15 cores and which to run on the A7 cores? [5%]





ARM Cortex-A7 Pipeline



ARM Cortex-A15 Pipeline

Part 3 (15%)

a. For each of the following pairs of terms, describe in 1 sentence the difference between the two: [3%]

- *direct mapped* vs. *set-associative* caches;
- *write policy* vs. *replacement policy*;
- *hit* vs *miss*;
- line *index* vs *tag*.

b. A vector operation over N elements is parallelized by dividing the work among P cores, each running a `compute` function that works on N/P elements (assume that N is a multiple of P). The following are two possible implementations of `compute`:

```
void compute1(int p, int N, int P) {
  for (int i = p*N/P; i < (p+1)*N/P; ++i)
    A[i] = B[i] * C[i];
}

void compute2(int p, int N, int P) {
   for (int i = p; i < N; i += P)
      A[i] = B[i] * C[i];
}
```

(`p` is the index of the core running `compute`, `P` the total number of cores, `N` is the total number of elements in `A`, `B` and `C`).

During testing it appears `compute2` scales very poorly (the performance actually decreases as P increases) whereas `compute1` scales nearly perfectly (the performance increases nearly linearly with P). The number of write misses for `compute2` is also much larger. Explain what mechanism is causing this difference. [5%]

b. Address translation is a mechanism via which an operating system can give the appearance of separate memory address spaces to different processes. It is based on the concept of virtual (logical) address from the perspective of software, translated to a physical address towards the main memory. The component in charge of translation is the memory management unit (MMU), which keeps a cache of translated addresses called TLB. Explain the difference between *hardware-managed* and *software-managed* TLBs in your own words (1-2 sentences for each). Explain at least one advantage and one inconvenient of each. [3%]

d. Depending on where the TLB is placed in the micro-architecture, a L1 cache can be either virtually or logically addressed. Nowadays, processors with virtually indexed L1 caches extend the tag with extra bits called "process ID". Explain which problem of virtually indexed caches is solved by using the process ID in the tag. [4%]