EE485F Term Paper

CAPT Charles B. Cameron, USN

August 22, 2010

1 Requirements

Due Date The term paper is due in class not later than 3 December, 2010, although you may hand it in earlier if you wish to do so. A semi-randomized schedule for presentations will be posted on the web site. Time slots allow 20 minutes for the presentation and five minutes for questions and answers.

Format Papers must be double-spaced on $8\frac{1}{2}'' \times 11''$ paper using a font between 10 and 12 points in size, inclusive, with margins of at least $1\frac{1}{2}''$ on all four sides in order to facilitate my reviewing them. Papers must be between 3000 and 5000 words in length (12–20 pages), not counting the title page, headings, figures, tables, listings, or mathematics. I encourage—but do not require—you to use LaTeX to format your papers. Here is a link to get started, if you choose to go this route:

`http://www.usna.edu/Users/ee/cameronc/Miscellaneous/LaTeX/LaTeX.html`

Organization Organize your papers so that it is clear from the introduction what it is you are writing about and clear from the conclusion what the import of your research is. Use mathematics, figures, tables, and program fragments as necessary to clarify your exposition. Include a title page with your name on it. Do not put your name on any other pages. Acknowledge all your sources in your bibliography—even those that relate to oral conversations with classmates, friends, or others.

2 Topics

The permissible topics are listed below. Other topics may be used with my permission, provided you seek it in writing not later than 1 October, 2010.

- As the feature size of integrated circuits shrinks, the possibility of errors in processing may increase. For example, single-event upsets caused by gamma rays may change the contents of a memory cell in an unpredictable way. The probability of such events increases as the size of a memory cell shrinks. Another consequence of decreasing feature size is the increased possibility of a defect in a transistor when the die is produced. Even if such defects do not appear or

are not detected at the time the integrated circuit is manufactured, they may be induced by metal migration *after* manufacture. Discuss how various schemes for high reliability such as fault tolerance and redundancy might be used to compensate for such problems. Consider the effect these measures would have on the available chip real estate.

- Error-correction circuitry such as parity-bit detection and Hamming codes with single-error correction and double-error detection have been used in memory, disk storage, L1 cache, L2 cache, and even in registers. Discuss the effects that using such schemes would have on the management of a memory hierarchy and on cache policies such as write-back or write-through and write-allocate or write-no-allocate. Also discuss the trade-offs entailed by their use. How would multicore chips be affected? How feasible is it to use error-scrubbing techniques in which the entire memory or cache is repeatedly examined for single-bit errors before double-bit errors occur? Consider the rate at which such scanning can occur compared to the rate of occurrence of bit errors, having due regard to the size of the data structure being scanned.

- Broadly speaking there are two different ways in which computer architects have taken advantage of the increased number of transistors that can fit on a chip. They can increase the number of processor cores on the chip or they can increase the number of system functions such as memory controllers, I/O bus interfaces, network interfaces, and cache controllers or the amount of space devoted to level-1 and level-2 caches. When they choose to use more processor cores, some of the system functions will need to be implemented off chip. When they choose to implement the system functions on the chip, the number of processor cores will be reduced. Explore the trade-offs associated with such choices.

- Discuss the suitability of wormhole routing to the problem of getting multiple cores on a single chip and on nearby chips to cooperate with each other. How might such routing be used to support multithreading? Consider how deadlock might arise and how it might be broken when a ring-topology is used. Consider the same problem when a toroidal topology is used. Consider it, too, when a mesh-topology (like a torus, but lacking wrap-around links) is used. What existing or historically existing systems actually use or have used wormhole routing?

- The translation look-aside buffer is a high-speed cache used to translate a program's virtual addresses into corresponding physical addresses. The memory cache for holding the retrieve result uses low-order address bits for indexing into a cache line and high-order bits for determining which line of the cache should hold the addressed memory data. The cache also uses some of the high-order bits as tags for identifying the contents of a particular cache line. The tags and the line indices typically employ the same type of addresses. Either they both use virtual addresses or they both use physical addresses. Hybrid schemes are possible, however, in which one of them uses virtual addresses and the other uses physical addresses. Discuss the trade-offs associated with the various combinations of these schemes. Consider such factors as the size of the cache, whether

or not the cache must support multiprocessor coherence, interprocessor communication via shared memory, context switching, and cache policies such as write-through and write-back.

- The speed of processors has been increasing at a great rate and so has the capacity of disk drives. However, the latency and bandwidth of disk drives has not been increasing at a rate commensurate with the increase in speed of the processors that access them. Discuss techniques that mitigate the disparity between processor speed and disk-drive latency and bandwidth.

- The Standard Performance Evaluation Corporation (SPEC) maintains benchmarks for computer manufacturers to use in order to characterize the performance of their processors. The benchmarks, however, are targeted at scientific and desktop processors, not at digital signal processors and embedded microcontrollers. As a consequence it can be hard for designers to make effective comparisons between different processors in the latter two categories. Discuss the reasons for this and what might be a reasonable approach for SPEC to take to rectify this deficiency. What is the basis for comparison between such processors as things stand now? How meaningful are comparisons based on these factors?

- In the 1980s researchers explored the use of reduced instruction sets in order to facilitate the use of pipelining to obtain higher rates of throughput. At the time, such processors often had the highest execution rates of any processors then available. In the 1990s came the development of superscalar processors. Although some of these were based on the use of reduced instruction sets, others employed complex instruction sets. Today, some of the fastest microprocessors use complex instruction sets. Discuss the factors that made this shift possible. What innovations did the manufacturers of complex instruction set computers (CISC) introduce to accomplish this? What motivated them to do it? What were the trade-offs that they considered? What are the consequences for users of these processors today? How does the issue of power dissipation arise in this context? Why could the makers of reduced instruction set computers (RISC) not make even faster processors?

- Two unconventional approaches to the design of cache memories are the use of *victim caches* and *stream caches*. Describe the situations in which they might be useful and provide improved performance. What trends are there that beg for other novel approaches to the design of cache memories? Consider how these various caching schemes work both with single processors and with multiple processors trying to synchronize their caches.

- Compare and contrast the techniques and performance of *branch prediction*, *speculative execution*, and *predicative* [not *predictive*] *execution*.

- Two fundamentally different approaches to connecting processors together entail the use of busses, on the one hand, or switched interconnects, on the other hand. Discuss the advantages and disadvantages to both. Consider situations where the

processors are close together or far apart, fast or slow, clustered or widely separated, and heterogeneous or homogeneous. How does the design of processors and systems vary as a function of the choice of interconnection approach.