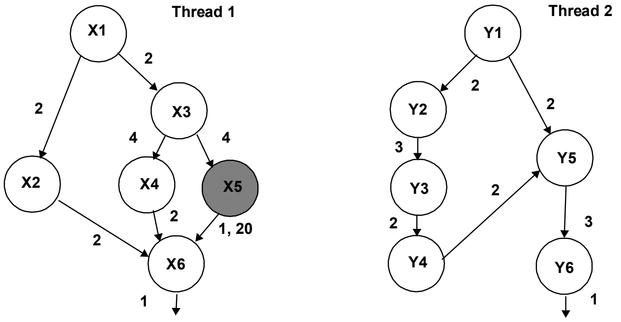
ENGN 2910A Homework 08 (60 points) – Due Date: on or before Nov 25th 2015

Professor: Sherief Reda

School of Engineering, Brown University

1. (25 points) Assume the example we taken in the class with the given data flow graph except but we will now assume X5 is a cache hit and hence it always takes one cycle to complete. For the following questions, please use the same assumption about for the superscalar OoO organization as in class.



a. Redo the speculative scheduling using interleaved multi-threading and block multi-threading. Note that when C5 is a cache hit, block multi-threading is identical to a single-threaded core, where X executes first and then Y executes.

b. How much faster (in clock cycles) is interleaved multi-threading over block multi-threading in this case?

c. Interleaved multi-threading required more hardware support for selective flossing and using thread ID for tag matching dependencies. Due to the design complexity, the clock cycle of the inter-leaned multi-threading processor is 20% slower. Does it still make sense to use interleaved multi-threading for the modified threaded core where X5 is a cache hit?

2. [25 points – from Debois] Consider a future 16-way CMP operating in a power-constrained environment. The CMP can automatically configure itself to run as 1-, 2-, 4-, 8-, 16-core CMP but always using a fixed power budget. For instance, it can run as a single-core processor by grabbing power from the other 15 cores by putting them to sleep and using the additional power to boost its frequency. Assume that sleep and wakeup times are zero, and that power and frequency have a square relationship. For instance, if one core uses the power of all 16 cores, its frequency can increase four-fold.

Consider a partially parallel application. This application starts as a single-threaded application and spends 5% of the time in sequential mode. During the following 40% of the time, the application has 16 threads, and only four threads for the next following 40% of the execution time. During the remaining execution time, the application has only one thread.

a. What is the speed of this application when it runs on this future CMP compared to running on a single-core machine that uses the same power but operates at a higher frequency using the square relationship?

b. What is the speedup of this application when it runs on this future CMP compared to running on a traditional 16-way CMP (again using the same power budget) that does not provide the reconfiguration capability?

c. Due to limitations of voltage scaling, assume that in the future power depends linearly on frequency, comment on what benefits (if any) of using the proposed reconfigurable boosting scheme for CMPs.

3. [10 points – From Hennessy] Assume a hypothetical GPU with the following characteristics:

- Clock rate 1.5 GHz
- Contains 16 SIMD processors, each containing 16 single-precision floating point units.
- Has 100 GB/sec off-chip memory bandwidth

Without considering memory bandwidth, what is the peak single-precision floating-point throughput for this GPU in GFLOP/sec, assume all memory latencies are hidden? Is this throughput sustainable given the memory bandwidth limitation?