The Death of ILP

Steven Swanson Andrew Petersen Mark Oskin
Computer Science and Engineering
University of Washington

Recently, many analytical models of optimum processor pipeline depth have been proposed [MICRO'02,03,ISCA'02]. At a high level, these models attempt to balance available ILP against technological trends. Lacking from these models, however, is the concept of *worth*; i.e., while building a 40 stage pipelined processor may be performance optimal, is it worth the effort?

Our Wild and Crazy proposal is that we should abandon the search for parallelism, give up what we have found so far, and embrace processors built without any parallelism at all – no issue width and no pipelining. Why you ask? Our simple analytical model suggests that, unless the memory wall is demolished, processors will spend so much time waiting for main memory that there will be insufficient performance gains to justify building complex processors. Amdhal's Law is going to put a bunch of architects on the job market! Now that we have your attention, we describe the model.

The key component of our analytical model that differentiates it from prior work is the concept of *worth*: Whether or not an architecture is worth the effort it takes to create. A simple inequality embodies the notion of worth:

$$T(\rho) \leq \frac{1}{s} T(0) \tag{1}$$

$T(0)$ is execution time for a baseline architecture that does not exploit parallelism at all (i.e., one pipe stage, 1 wide issue), $T(\rho)$ is execution time for a new architecture that exploits parallelism of degree ρ (defined shortly), and s is the minimum speedup that is worth pursuing. If this inequality holds for measured values of $T(\rho)$ and $T(0)$ and a target s, then it is worth building the new architecture.

Our simple model divides execution time into computation time and time spent waiting for memory. Let I be the number of instructions in an application, c be the clock cycle (in seconds), R be the fraction of instructions that access main memory (i.e., L2 cache misses), and M be the time required to service a cache miss. Execution time is then:

$$T(\rho) = c\frac{I}{\rho} + RIM \tag{2}$$

Next, we combine equations (1) and (2) and express the clock cycle of our new machine in terms of the old machine using an idealized pipeline model $C_\rho = C_0/P_s$, where P_s is the number of pipeline stages in the new machine. We further assume an idealized utilization of issue width. Given a machine of width W all issue slots are assumed used, and hence, the degree of parallelism exploited by the machine is simply $\rho = P_s \times W$. Substituting and solving for the miss penalty and clock cycle time yields:

$$\frac{1 - \frac{s}{\rho}}{R(s-1)} \geq \frac{M}{C_0} \tag{3}$$

This equation shows that for a new architecture to be worthwhile, the "memory wall" (the ratio between memory access time and computation speed) must be bounded by a relationship between the desired speedup s, the likelihood of a cache miss R, and the amount of parallelism the processor can exploit ρ.

The latency to main memory and logic speed both decrease exponentially with time, though at different rates. We can express the cost of a cache miss as function of time $M = \alpha_m e^{k_m t}$ and the speed of logic as $C_0 = \alpha_c e^{k_c t}$. Folding α_m and α_c together and substituting in equation (3) yields:

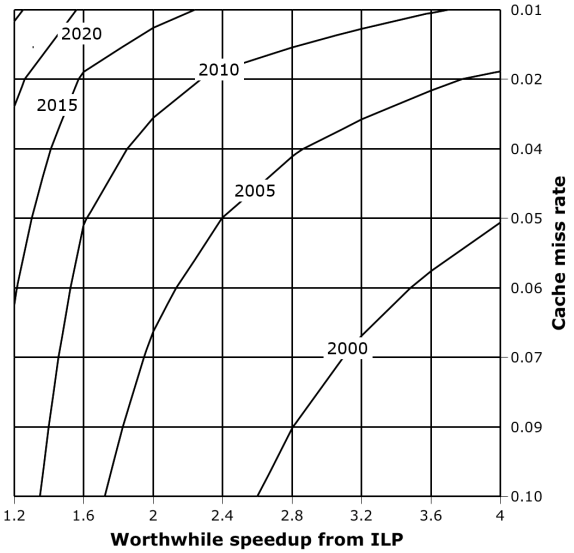

Figure 1: Each line shows the year at which pursuing parallelism become fruitless as the desired speedup and cache miss rate varies.

$$\frac{ln\frac{1-s/\rho}{\alpha R(s-1)}}{k_m - k_c} \geq t \tag{4}$$

Now for the trip down the rabbit hole. Suppose we let ρ go to infinity; then, we can ask the question, "When is it valuable to build an ILP-exploiting architecture?" By assuming a 7% decrease in main memory access latency yearly and a 20% increase in computational logic speed, we derive Figure 1, which shows the output of our model for a range of cache miss rates and desired speedups. Each line is label with a year. The area above each line represents the cache miss rates and target speedups for which a parallel processor is worthwhile in that year. For example, with 6 misses per hundred instructions in the year 2000, even an infinite issue machine with an infinitely deep, idealized pipeline will be unable to achieve a 3.6× speedup.

The implications of the graph are startling. By 2005, a 2% miss rate (across all instruction) implies that a pipelined processor can achieve no better than 3.7× the performance of a non-pipelined machine. The potential benefit of parallelism drops to 2.4× by 2010 and to less than 60% by 2015. Given the astronomical cost of the die area, power, complexity, and pipe latch delay that parallelism-chasing processors require, the outlook for cost-effective, parallel processors is dim at best after 2020. The ILP party is over then.

Caveat emptor: There are many simplifications in the model and ways in which the world may change. The most glaring omission of our model is the lack of memory parallelism. To a first order, allowing n outstanding cache misses should reduce the average perceived miss latency by $1/n$. Incorporating this into our model with $n = 4$ (a generous estimate of the amount of memory parallelism available in a typical integer application) delays parallel processors' passage from this veil of tears by about 7 years.

While our model is not perfect, the trends it reveals show that the golden age of ILP is ending. It may be the case that we should not only stop pursuing additional ILP in processors, but we should stop building processors to exploit any parallelism at all. Hey, you asked for wild and *crazy*.