Prolog to

Decisive Aspects in the Evolution of Microprocessors

An introduction to the paper by Sima

This paper traces progress in high performance commercial microprocessors by focusing on the evolution of their microarchitectures. As shown, the main road of the evolution of microarchitectures is marked by an increasing utilization of available instruction level parallelism (ILP) achieved by subsequently introducing three dimensions of ILP—called temporal, issue, and intrainstruction parallelism—in successive processor generations. This evolution happened in such a way that after thoroughly exploiting ILP along one dimension in the microarchitecture, it became inevitable to begin to utilize available ILP along a new dimension as well to further increase performance.

But while a new dimension of ILP is opened by introducing an appropriate basic technique such as pipelining, superscalar instruction issue, or single-instruction, multiple data (SIMD) execution into the microarchitecture, certain bottlenecks arise in particular subsystems of the microarchitecture. So the introduction of each new basic technique triggered the evolution of associated auxiliary techniques to eliminate the induced bottlenecks. With these auxiliary techniques included into the microarchitecture, however, the potential of parallel instruction execution in the considered ILP dimension becomes by and large exhausted, and the incessant demand for a further performance increase gives rise to utilizing ILP parallelism along a new dimension as well.

In particular, temporal parallelism was the first to make its debut with pipelined processors, allowing ideally to begin processing a new instruction every new clock cycle. The emergence of pipelined instruction processing stimulated the introduction of caches and of branch prediction. With these techniques added, enhanced (second-generation) pipeline processors arrived and approached the limits of temporal parallelism. For further performance increase, a second dimension of ILP—i.e., issue parallelism—had to be introduced. So superscalar processors issuing more than one instruction per clock cycle appeared. Superscalars evolved in three generations. First-generation superscalars made use of the direct (unbuffered) instruction issue. The bottleneck inherent to this principle, however, limits the microarchitecture to a two to three-wide reduced instruction set computer (RISC) or a two-wide complex instruction set computer (CISC) design. The demand for still higher throughput called then for widening the microarchitecture. In this way, second-generation superscalar emerged featuring dynamic instruction scheduling (buffered instruction issue), register renaming, and several additional techniques to eliminate the issue bottleneck of the preceding generation. Finally, having exhausted the extent of issue parallelism available in general purpose programs, a third dimension of ILP, called intrainstruction parallelism, was introduced in microarchitectures. Accordingly, third-generation superscalars appeared executing multiple operations per instruction by means of SIMD instructions. This enhancement, effective primarily in emerging multimedia and three-dimensional applications, already required an extension of the instruction set architecture (ISA) as well as the introduction of on-chip L2 caches.

In this way, microarchitectures evolved at the instruction level basically in three consecutive cycles, where each cycle can be attributed to introducing a new dimension of ILP into the microarchitecture. In the first cycle, temporal parallelism has been introduced in form of first- and second-generation pipelined processors. The second cycle is devoted to utilizing issue parallelism as well and was implemented in first- and second-generation superscalars, while in the third cycle interinstruction parallelism was introduced in the third-generation superscalars.

The sequence of introducing possible dimensions of ILP has been determined basically by two aspects: first, the drive to keep hardware complexity as low as possible and second, the goal of maintaining backward compatibility with preceding models.

All the decisive aspects mentioned above constitute a framework that explains the main road of the evolution of microarchitectures at the instruction level, including the sequence of major innovations encountered.

Nevertheless, due to the outlined developments, the potentials of ILP processing are becoming more and more exhausted, so any further substantial performance increase requires the utilization of available parallelism at a higher than instruction level—that is, at the thread level as well. Therefore the next cycle of the evolution is devoted clearly to the thread level. However, this step of the evolution is already beyond the scope of the paper.

Digital Object Identifier 10.1109/JPROC.2004.837615

—Howard Falk