

A LITTLE HISTORY

Back in the late 1940s *Von Neumann* came up with the compute model that has been used by the computer industry up to the turn of the century. Basically it is: fetch a number out of memory, use the processor to do a computation, and then place the answer back into memory; simple, straight forward and sequential (one thing at a time). There have been other compute models that have been used but none ever became main-stream. The beauty of the Von Neumann compute model was that the only way to increase performance was to increase the speed of the processor. The semiconductor industry was nice enough to provide that increase by moving to a new smaller geometry node ever two years (Moore's Law). Whenever a new computer showed up on the market, using a new compute modes (usually a parallel execution model), the Von Neumann computers would outperform it in two to four years (one to two node changes).

THREADS AND PIPE-LINING

There were attempts to do some parallel computing while still using the single processor Von Neumann compute model. In C that is called threading. Threading is asking the processor to do two or more things at once. There are problems with threading but for a small number of threads it works. The microprocessor industry came up with pipe-lining. They basically provided separate processing capability in a pipe (a series of hard wired special functions) that could rapidly process some of the more time consuming calculations so that the various threads would be completed around the same time; that is the shorter calculations wouldn't have to wait for the longer calculations before completing the program.

THE POWER CRISIS

At the 130 nm node the semiconductor industry ran into the power problem. We could no longer speed up our processor without going into thermal run-a-way. We were exceeding one GHz at the time. You need to understand that IC Designers are use to using parallel logic to speed up an operation. Therefore the obvious solution was to slow down the processor, generally to 800MHz, and add more processors. Voila, we had solved our problem. This then is called multi-core processing. Unfortunately we didn't realize we were abandoning the Von Neumann compute model by doing so.

SMP & AMP

Actually that is SMP, Symmetric Multi-Processing; using more than one of the same core. We actually had been playing with APM, A-Symmetric Multi Processing for quite awhile. First we started using Co-Processors attached to the General Purpose Processor. The main processor would hand off some domain specific processing that was taking up too much of the GPU's time. This evolved to companies providing separate processing boards to off load the graphics so that a PC could be used for high performance games. That then lead to the Graphics Processor (GPU) and the PC became an AMP architecture. That also had been going on in the cell phone as they kept adding features to the phone. Some phones have six processor cores in them now.

TODAY

Presently we are struggling to come up with a new Model of Computation so that we can develop the software architecture and tools needed to program these new Multi-Core ICs. This is spawning the greatest change in computation since the 1940s.