OrionX.net

Event

The OrionX Constellation: Data Center
High Performance Computing

The Return of Vector Processing

Shahin Khan

Digital transformation, cloud computing and application elasticity, open source software, and new app areas like AI and IoT are driving a renaissance in system architecture. This is an area of interest and research for us at OrionX.

Vector processing was an interesting topic to re-emerge recently. First during the International Supercomputing Conference (ISC), and again in various announcements in implicit and explicit ways.

On the Monday of the ISC conference, a new leader on the TOP500 list was announced. The Sunway TaihuLight system uses a new processor architecture that is Single-Instruction-Multiple-Data (SIMD) with a pipeline that can do eight 64-bit floating-point calculations per cycle. 10,649,600 computing cores comprising 40,960 nodes produce 93 petaflop/s running the LINPACK benchmark.

Later that day, at the "ISC 2016 Vendor Showdown", NEC had a presentation about its project "Aurora". This project aims to combine x86 clusters and NEC's vector processors in the same high bandwidth system. NEC has a long history of advanced vector processors with its SX architecture. Among many achievements, it built the Earth Simulator, a vector-parallel system that was #1 on the TOP500 list from 2002 to 2004. At its debut, it had a substantial (nearly 5x) lead over the previous #1.

Vector Processing and Parallelism

Vector processing is a time-honored system architecture that started the supercomputing market, starting with the CDC STAR-100 system (STrings and ARrays, performing at 100 million floating point operations per second), and then led by the legendary Seymour Cray and his superb team. It is an efficient way to take advantage of certain kinds of parallelism in applications.

Parallelism means you can replicate the resources that can be used simultaneously, and enables scalability. If those resources are arithmetic functional units, you could get vector processors, long instruction word architecture (LIW), or graphic processing units (GPU) which can accelerate a substantial portion of the right applications. Replicating full CPUs led to various forms of multiprocessing systems: symmetric multiprocessing (SMP), Non-Uniform Memory Access (NUMA) systems, etc.

When microprocessors advanced enough to enable massively parallel processing (MPP) systems and then Beowulf clusters, it became economical

to replicate an entire system, and the supercomputing industry moved away from vector processing and led the scale-out model.

Close integration of accelerator technologies with the main CPU is, of course, a very desirable objective. It improves programmability and efficiency. Certain vector instructions have been available in several CPU architectures. For example, Visual Instruction Set (VIS), is a SIMD instruction set extension for SPARC V9 microprocessors. Advanced Vector Extensions (AVX) are extensions to the x86 instruction set architecture for CPUs from Intel and AMD, and with compiler support in recent years.

Another example is Virtual Vector Architecture (ViVA), a technology from IBM for coupling together CPUs to act as a single vector processor, which it supported in hardware with the POWER6 processor. Along those lines, we should also mention the Convey system, now part of Micron, which goes all the way, extending the x86 instruction set, and performing the computationally intensive tasks in an integrated FPGA.

A big advantage of vector processing is that it is part of the CPU with full access to the memory hierarchy. In addition, compilers can do a good job of producing optimized vector code. For many apps (for example in climate modeling, automotive, aerospace, seismic processing, ...) vector processing can be quite the right architecture.

Vector parallel systems combined vector processing and multiprocessing and reigned supreme for many years, for very good reasons. But MPPs pushed vector processing back, and general purpose GPUs (GP-GPUs) pushed it further still. GPUs leverage the high volumes that the graphics market provides but can offer more general purpose numerical acceleration with some incremental engineering.

But when you scale, you scale not just capability, but also complexity, and hope that capability scales better. Little inefficiencies start adding up until they become a serious issue. At some point, you need to revisit the system and take steps, perhaps drastic steps. You could say the Sunway TaihuLight system atop the TOP500 list is an example of this. And there are other efforts towards building new CPUs for exascale-class systems, such as the "Neo processor" that Rex Computing is developing.

Summary

Vector processing is poised to make a comeback as a lightweight and efficient high performing architecture.