

GPU Dominance

Some years ago, I wrote a series of articles for this magazine predicting that the GPU would become the dominant computing component in the PC with the CPU increasingly relegated to tertiary I/O processing. This prediction is of course being borne out as modern GPUs assume increasing responsibility for the computing load generated by consumer applications. Emphasis on CPU GHz has declined over the years as greater emphasis has been placed on I/O bandwidth and data throughput. The CPU and primary system memory has essentially been allocated almost entirely to running the Windows OS, while the GPU increasingly offloads the processing work from the only truly computation-intensive applications most consumers actually use their PCs for: playing games.

In fact, the two largest categories of consumer software purchases for the PC are games at No.1 and security utilities at No.2. When consumers are asked, "What are the most important things you use your PC for?" they overwhelmingly point to games, followed by the usual litany of computationally un-intensive applications. Of course nobody ever says that they buy a new PC to run security software, although that will be their largest software expense after buying games.

Going forward, the only purpose for your CPU will be to load Windows, whose only purpose will be to constantly check itself for malware and boot your games, the latter of which will increasingly be offloaded to the GPU. The other popular uses of the PC, such as email, chat, IM/text messaging, playing music, and photo management will probably increasingly migrate to your cell phone. The only applications you'll really need a PC for in the future are ones that require a big screen and big input devices, such as mice and keyboards.

Of course, Microsoft and Intel are institutionally oblivious to this trend. Microsoft doesn't see any problem with the idea that the PC's only purpose/value proposition in the home is to run a vestigial OS and some antiquated productivity applications. Intel makes most of its money selling chips for enterprise use and as servers. The extent to which they can repurpose their enterprise chips to the consumer market is just incremental money to them. It's easier to try to convince consumers that they need a media file

server in their living room for some reason than it is to specifically design CPUs that run games well instead of databases. This persistent state of delusion on Wintel's part has driven AMD-ATT's and Nvidia's fortunes as the companies that dedicate themselves to creating chips that specifically enhance the consumer experience.

There is another important reason that GPUs have momentum for taking over for the CPU. GPUs are better designed for accelerating tomorrow's nonenterprise computational problems. You're probably used to hearing about how much better games will work when developers hand-optimize them for multithreading or multicore processing. I've scoffed at this assertion in the past because most games derive little or no performance benefit from being threaded or multicore optimized. One of the reasons for this is that the heavy computational lifting and parallelism is already largely handled by the GPU.

The real world we live in is an extremely complex computing environment in which vast numbers of quantum objects are interacting with each other at the speed of light to produce the world we perceive. We evolved massively parallel computing brains to cope with the enormous complexity of the environment we live in. Traditional CPU architectures, on the other hand, were designed with serial processing in mind. Although parallelism was added to them over time to speed them up, a modern computer is essentially an extremely fast serial processing device. Although very powerful, it still requires "unnatural acts" to get modern CPUs to solve real-time parallel processing problems.

The GPU's native architectural advantage for future computing derives from its physics roots. It's easy to forget that a 3D game is an extraordinary achievement in real-time simulation of optical physics. Early GPUs were really just highly specialized massively parallel physics engines that have evolved greater processing flexibility over time. To stretch the point, a modern GPU has a great deal more in common with the structure of a human brain than the CPU does. It's no coincidence that the GPU's ability to "visualize" an interactive real-time 3D world for us has resulted in an architecture with a lot in common with the one we use to visualize an interactive 3D world for ourselves. ▲



Alex St. John was one of the founding creators of Microsoft's DirectX technology. He is the subject of the book "Renegades Of The Empire" about the creation of DirectX and Chromeffects, an early effort by Microsoft to create a multimedia browser. Today Alex is President and CEO of WildTangent Inc., a technology company devoted to delivering CD-ROM quality entertainment content over the Web.

Going forward,
the only
purpose for
your CPU
will be
to load
Windows. . . .

Send your feedback to thesaint@cpumag.com