State of CAD and Engineering Workstation Technologies

White Paper

TN Chan
System Architect
Compucon New Zealand
tn@compucon.co.nz
www.compucon.co.nz

V010 June 2011

CAD is Computer Aided Design
CAE is Computer Aided Engineering
CEW is Computer aided design and Engineering Workstation
CPU is Central Processing Unit
GPU is Graphics Processing Unit

CAD and CAE are specialist professional spaces. Correspondingly, CEW should be professionally produced as against the amateur approach of adding a graphics card to a generic computer. Some CAD applications are CPU intensive and some are GPU intensive. This situation emerged not too long ago and is still transitioning. This paper examines the state of hardware technology for CAD applications and assists professional designers and engineers to appraise the effectiveness of CEW in the market.

Table of Content

Hardware for CPU-Intensive Applications

Computer hardware is designed to support software applications and it is a common but simplistic view that higher spec hardware will enable all software applications to perform better. Up until recently, the CPU was indeed the only device for computation of software applications. Other processors embedded in a PC or workstation were dedicated to their parent devices such as a graphics adapter card for display, a TCP-offloading card for network interfacing, and a RAID algorithm chip for hard disk redundancy or capacity extension. However, the CPU is no longer the only processor for software computation. We will explain this in the next section.

Legacy software applications still depend on the CPU to do computation. That is, the common view is valid for software applications that have not taken advantage of other types of processors for computation. We have done some benchmarking and believe that applications like Maya 03 are CPU intensive.

For CPU-intensive applications to perform faster, the general rule is to have the highest CPU frequency, more CPU cores, more main memory, and perhaps ECC memory (see below).

Legacy software was not designed to be parallel processed. Therefore we shall check carefully with the software vendor on this issue before expecting multiple-core CPUs to produce higher performance. Irrespectively, we will achieve a higher output from executing multiple incidences of the same application but this is not the same as multi-threading of a single application.

ECC is Error Code Detection and Correction. A memory module transmits in words of 64 bits. ECC memory modules have incorporated electronic circuits to detect a single bit error and correct it, but are not able to rectify two bits of error happening in the same word. Non-ECC memory modules do not check at all – the system continues to work unless a bit error violates pre-defined rules for processing. How often do single bit errors occur nowadays? How damaging would a single bit error be? Let us see this quotation from Wikipedia in May 2011, "Recent tests give widely varying error rates with over 7 orders of magnitude difference, ranging from 10^{-10}–10^{-17} errors/bit-hour, roughly one bit error per hour per gigabyte of memory to one bit error per century per gigabyte of memory."

Hardware for GPU-Intensive Applications

The GPU has now been developed to gain the prefix of GP for General Purpose. To be exact, GPGPU stands for General Purpose computation on Graphics Processing Units. A GPU has many cores that can be used to accelerate a wide range of applications. According to GPGPU.org, which is a central resource of GPGPU news and information, developers who port their applications to GPU often achieve speedups of orders of magnitude compared to optimized CPU implementations.

Many software applications have been updated to capitalize on the newfound potentials of GPU. CATIA 03, Ensight 04 and Solidworks 02 are examples of such applications. As a result, these applications are far more sensitive to GPU resources than CPU. That is, to run such applications optimally, we should invest in GPU rather than CPU for a CEW. According to its own website, the new Abaqus product suite from SIMULIA – a Dassault Systemes brand – leverages GPU to run CAE simulations twice as fast as traditional CPU.

Nvidia has released 6 member cards of the new Quadro Fermi family by April 2011, in ascending sequence of power and cost: 400, 600, 2000, 4000, 5000 and 6000. According to Nvidia, Fermi delivers up to 6 times the performance in tessellation of the previous family called Quadro FX. We shall equip our CEW with Fermi to achieve optimum price/performance combinations.

The potential contribution of the GPU to performance depends on another issue: CUDA compliance.

State of CUDA Developments

According to Wikipedia, CUDA (Compute Unified Device Architecture) is a parallel computing architecture developed by Nvidia. CUDA is the computing engine in Nvidia GPU accessible to software developers through variants of industry-standard programming languages. For example, programmers use C for CUDA (C with Nvidia extensions and certain restrictions) compiled through a PathScale Open64 C compiler to code algorithms for execution on the GPU. (The latest stable version is 3.2 released in September 2010 to software developers.)

The GPGPU website has a preview of an interview with John Humphrey of EM Photonics, a pioneer in GPU computing and developer of the CUDA-accelerated linear algebra library. Here is an extract of the preview: “CUDA allows for very direct expression of exactly how you want the GPU to perform a given unit of work. Ten years ago I was doing FPGA work, where the great promise was the automatic conversion of high level languages to hardware logic. Needless to say, the huge abstraction meant the result wasn't good.”

Quadro Fermi family has implemented CUDA 2.1 whereas Quadro FX implemented CUDA 1.3. The newer version has provided features that are significantly richer. For example, Quadro FX did not support “floating point atomic additions on 32-bit words in shared memory” whereas Fermi does. Other notable improvements are:

- Up to 512 CUDA cores and 3.0 billion transistors
- Nvidia Parallel DataCache technology
- Nvidia GigaThread engine
- ECC memory support
- Native support for Visual Studio

State of Computer Hardware Developments

HDD is Hard Disk Drive
SATA is Serial AT Attachment
SAS is Serial Attached SCSI
SSD is Solid State Disk
RAID is Redundant Array of Inexpensive Disks
NAND is memory based on “Not AND” gate algorithm

Bulk storage is an essential part of a CEW for processing in real time and archiving for later retrieval. Hard disks with SATA interface are getting bigger in storage size and cheaper in hardware cost over time, but not getting faster in performance or smaller in physical size. To get faster and smaller, we have to select hard disks with SAS interfaces, with a major compromise on storage size and hardware price.

RAID has been around for decades for providing redundancy, expanding the size of volume to well beyond the confines of one physical hard disk, and expediting the speed of sequential reading and writing, in particular random writing. We can deploy SAS RAID to address the large storage size issue but the hardware price will go up further.

SSD has turned up recently as a bright star on the horizon. It has not replaced

HDD because of its high price, limitations of NAND memory for longevity, and immaturity of controller technology. However, it has found a place recently as a RAID Cache for two important benefits not achievable with other means. The first is a higher speed of random read. The second is a low cost point when used in conjunction with SATA HDD.

Intel has released Sandy Bridge CPU and chipsets that are stable and bug free since March 2011. System computation performance is over 20% higher than the previous generation called Westmere. The top CPU model has 4 editions that are officially capable of over-clocking to over 4GHz as long as the CPU power consumption is within the designed limit for thermal consideration, called TDP (Thermal Design Power). The 6-core edition with official over-clocking will come out in June 2011 timeframe.

Foreseeable Future

Semiconductor manufacturing technology has improved to 22 x 10^{-9} meters this year 2011 and is heading towards 18 nanometers in 2012. Smaller means more: we will get more cores and more power from a new CPU or GPU made on advancing nanotechnology. The current laboratory probe limit is 10^{-18} and this sets the headroom for semiconductor technologists.

While GPU and CUDA are having big impacts on performance computing, the dominant CPU manufacturers are not resting on their laurels. They have started to integrate their own GPU into the CPU. However, the level of integration is a far cry from the CUDA world and integrated GPU will not displace CUDA for design and engineering computing in the foreseeable future. This means our current practice as described above will remain the prevailing format for accelerating CAD, CAE and CEW.

End