Tarek El-Ghazawi, Professor and IEEE Fellow
The George Washington University
www.seas.gwu.edu/~tarek/

Making Heterogeneity a First Class Citizen

Abstract- In recent years, the top supercomputers in the world were built around the use of heterogeneous architectures where conventional microprocessors are accelerated using specialized ones such as gaming processors, general purpose graphics processors (GPGPUs) and field programmable gate arrays (FPGAs). Such heterogeneous systems did not only achieve unprecedented successes but also often made history. Today, Titan at ONRL, with 18,688 NVIDIA K20 GPUs, achieves 17.59 PFLOPS, i.e. nearly 18,000 trillion calculations per second. In 2008, the Road Runner at LANL, using 12960 of the Cell processor chip, built by IBM/Sony/Toshiba for the Sony PlayStation 3, reached the PetaFLOPs performance. In 2010, China was able to build, also for the first time in history, a supercomputer, Tianhe-1A, that ranked as the top supercomputer in the world. Tianhe-1A major speed boost came from 7000 NVIDIA graphical processing units (GPUs). There are, therefore, obviously substantial performance advantages for such systems. However, there are also enormous productivity and portability challenges. Developing applications for such systems requires heroic programming efforts, often using proprietary tools, and when developed for one system they are not easily portable without additional substantial effort and/or performance loss. In this talk, we examine such challenges for heterogeneous parallel computers and examine how hardware, and the software stack, can integrate heterogeneity in a seamless fashion.