Efficient Automatic Speech Recognition on the GPU

By: Jike Chong
November 2, 2010

Abstract:
Automatic speech recognition (ASR) allows multimedia content to be transcribed from acoustic waveforms to word sequences. This technology is emerging as a critical component in data analytics for a wealth of media data that is being generated everyday. Commercial usage scenarios are already appearing in industries such as customer service call centers for data analytics where ASR is used to search recorded content, track service quality, and provide early detection of service issues. Fast and efficient ASR enables economic employment of text-based data analytics to multimedia contents. This opens the door to unlimited possible applications such as automatic meeting diarization, news broadcast transcription, and voice-activated multimedia systems in home entertainment systems.

This talk provides speech recognition application developers with an understanding of specific implementation challenges to achieve 10x faster performance compared to sequentially processing ASR on a CPU. The techniques presented accelerate the speech inference process, the advanced weighted finite state transducer (WFST) based methods, and the Viterbi algorithm, and are illustrated with an efficient reference implementation on the GPU that could be productively customize to meet the needs of specific usage scenarios. Four generalizable techniques will be presented, including data-parallel versions of: dynamic data-gather buffer, find-unique, lock-free data structures using atomics, and hybrid global/local task queues. When used together, these techniques can effectively resolve ASR implementation challenges on a GPU. Furthermore, we encapsulate the performance optimization techniques into a speech recognition application framework and enables flexible customization that targets specific usage scenarios while preserving execution efficiency.

About the Speaker:
Jike Chong is a Ph.D. researcher at University of California, Berkeley, working on application frameworks for application domain experts to effectively utilize highly parallel platforms. The work focuses on automatic speech recognition and computational finance as two parallel application areas. Jike’s prior work in parallel computing led to several patents at Sun Microsystems, Inc, and Intel Corporation. While working at Sun, he designed micro-architectural features for the highly parallel flagship T2 microprocessor. As an Intel Ph.D. Research Fellow (2008-2010) at the Intel Application Research Labs, he developed acceleration techniques for computational finance kernels for
Intel’s upcoming Many Integrated Core (MIC) processors. At Xilinx Research Labs, he developed multi-ported memory subsystems for parallel network applications on the FPGA. Jike earned his B.S. and M.S. in Electrical and Computer Engineering at Carnegie Mellon University, and the Management of Technology (MOT) Certificate from Haas Business School at University of California, Berkeley. He is a Member of Eta Kappa Nu, Tau Beta Pi, and the IEEE.