Intra- and Inter-FPGA Programmable Multiprocessor Designs with Emphasis on Large-Scale Matrix Operations*

Sotirios G. Ziavras, Xiaofang Wang and Mohammad Z. Hasan
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
Newark, NJ 07102
ziavras@njit.edu

Recent advances in multi-million-gate platform FPGAs facilitate the implementation of complex parallel systems on a programmable chip (PSOPCs). Many FPUs also can now be incorporated in these systems. Additionally, on-the-fly resource reconfiguration can be carried out by a combination of hardware-software means. In contrast to the majority of the FPGA community that still employ reconfigurable logic to develop algorithm-specific circuitry, our FPGA-based multiprocessor designs and implementations take advantage of all the aforementioned advances to deliver programmable multiprocessors. *Three representative systems of ours* are discussed.

(1) HERA (HEterogeneous Reconfigurable Architecture) is a mixed-mode reconfigurable machine that can implement simultaneously the SIMD, MIMD and M-SIMD execution modes. Each PE is centered on a single-precision IEEE 754 FPU, and can be switched between SIMD and MIMD at runtime. Mixed-mode parallelism can better match subtask characteristics in applications, thus resulting in sustained high performance. HERA is implemented on two XC2V6000 Virtex II FPGAs and currently contains 36 PEs. We evaluate HERA's performance in matrix multiplication and LU decomposition of large sparse matrices. Results with real electric power network matrices show that mixed-mode scheduling can result in speedups of 15-20% compared to strict SIMD and MIMD implementations.

(2) Matrices representing electric power networks are huge and very sparse. LU factorization techniques are often used in power flow analysis. We study block-based coarse-grain parallel LU factorization for a multiprocessor residing in an APEX20KE FPGA. We employ a clustered binary tree layout for the processors, where immediate neighbors share the same local data memory. Due to large variance in the block size, the number of non-zero elements in each block and some architecture constraints, a dynamic scheduling strategy is proposed. We test and analyze the performance for matrices up to 10279x10279. Our research demonstrates that FPGA-based configurable parallel systems can serve as low-cost alternatives in high-performance scientific computing.

(3) The solution to a set of linear sparse equations can be obtained using the W-matrix method. A vector processor is proposed for the efficient implementation of this method on two XC2V6000 FPGAs. We analyze the effect on the performance of pipelined functional units, pre-fetching, instruction chaining, multiple data buses, hardware synchronization and pipelined scattering.

* This work was supported in part by the U.S. Department of Energy under grant DE-FG02-03CH11171.