

An FPGA-based Massively Parallel Hardware Accelerator for SVM and CNN¹

Hans Peter Graf
hpg@nec-labs.com

Srihari Cadambi
cadambi@nec-labs.com

Igor Durdanovic
igord@nec-labs.com

Venkata Jakkula
Jakkula@nec-labs.com

Murugan Sankardadass
murugs@nec-labs.com

Eric Cosatto
cosatto@nec-labs.com

Srimat Chakradhar
chak@nec-labs.com

NEC Laboratories, America
4 Independence Way, Suite 200; Princeton, NJ 07738, USA

Abstract

We present an FPGA-based massively parallel architecture for accelerating machine learning algorithms. The architecture is based on clusters of vector processing elements (VPE) operating in single-instruction-multiple-data (SIMD) mode. The system has several key attributes that lead to high performance as well as low power dissipation. First, it efficiently uses DSP units in modern FPGAs to provide a very large number of simple vector processing elements operating in parallel. Second, the FPGA-based processor has a very high data bandwidth to off-chip memory. Third, the architecture is easily scalable since each VPE cluster is serviced by an independent memory bank and can be replicated (subject to FPGA resource constraints) to increase performance. Finally, the main data flows and interconnections are all highly local, resulting in low power dissipation.

We implement two flavors of the above architecture on an off-the-shelf FPGA-based PCI card. The card has a single Xilinx Virtex 5-LX330T FPGA serviced by 4 independent banks of DDR2-SDRAM memory totaling 1GB. Each memory bank-to-processor data bus is 32 bits wide and can be clocked at up to 333MHz, resulting in a peak processor-to-memory data bandwidth of close to 11GB/s.

The first architecture targets the kernel computation in the SVM algorithm. We implement 128 VPEs in 4 clusters of 32 each. By lowering the resolution of the computation, we obtain a speed of 9.5GMACs for SVM training. For SVM classification, we optimize the hardware further by leveraging data packing and double clocking to obtain 48GMACs. This performance is more than an order of magnitude higher than that of any FPGA implementation reported so far. It is similar to the fastest speeds published on a Graphics Processor for the MNIST problem, despite a clock rate of the FPGA that is six times lower. High performance at low clock rates

¹ A spotlight poster (T67) about this work is presented at NIPS 2008, but does not describe any details of the hardware implementation. Here we focus on hardware architecture and design that is appropriate for an audience interested in hardware.

makes this massively parallel architecture particularly attractive for embedded applications, where low power dissipation is critical.

The second architecture targets convolutional neural networks (CNNs). We implement 100 VPEs in 4 clusters of 25 each. Each cluster of VPEs operates in a systolic fashion at 120MHz, obtaining an effective speed of about 12GMACs. We tested this on a face recognition application to obtain an end-to-end processing speed of 6 frames per second, about 6 times faster than software.