Super-FPGA: Overcoming Von Neumann to Save Moore

Venkatesh Akella, Soheil Ghiasi
Electrical and Computer Engineering
University of California, Davis
{akella,soheil}@ece.ucdavis.edu

Introduction

The consistency of Moore's law has shaped the advances in Microprocessors, among other branches of semiconductor industry. The compute performance of commodity processors has doubled every 18 months throughout the past few decades. Interestingly, recent advances in microprocessor performance are mostly due to continuing increase in clock frequency, rather than architectural enhancements. However, as process scaling approaches physical limits, the continuation of Moore's law and exponential growth in processor performance, are going to be hindered.

Hence, architecture and compilation innovation seems to be the only viable solution to further continue to increase the compute performance using the present fabrication technology to realize the processors. Exploiting parallelism and overcoming the traditional Von Neumann execution bottleneck, is one of the most efficient and practical methods to achieve this goal. Many multi-core based processors are developed with the same vision [1, 2, 3].

The performance of processors is often limited by the memory bandwidth, which does not scale as fast as the processing engine. As a result, the majority of on-chip area is dedicated to caches and the associated circuitry to alleviate the problem. On the other hand, field programmable gate arrays (FPGAs) replace the temporal execution model of the processor with spatial instantiation of the operations, which improves the performance, reduces the intermediate memory requirement, and provides more area on the chip to fabricate logic. More importantly, FPGAs exhibit a faster growth in their compute performance compared to the processors of the same generation (Figure 1). Although superior at the present time and in the near future, the compute performance of FPGAs will also be limited by Moore's law at some point.

To address this issue, we are developing a highly parallel FPGA-based architecture, called Super-FPGA, which combines the density advantage of Moore's law (up to the point it is available) with a fundamental paradigm shift in application mapping to extend the exponential growth of compute performance to future technology nodes. We expect our design to have a clock frequency of about 10X higher than the present FPGAs. Moreover, its improved density along with the revolutionized mapping methodology would realize designs of about 10X larger complexity.

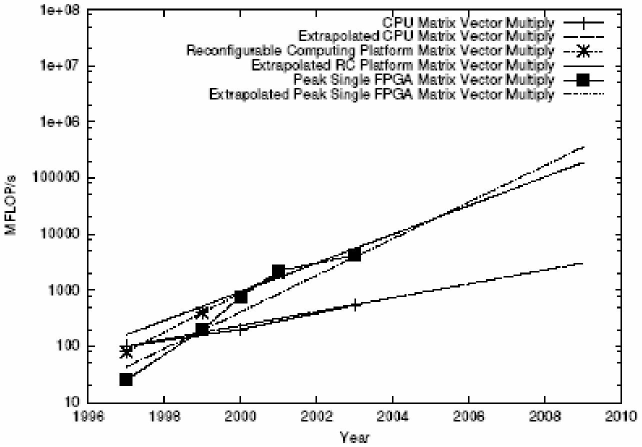

Figure 1: The performance trend of FPGA, Reconfigurable Computer and CPU for a major floating point operation [4].

Super-FPGA Architecture

Our Super-FPGA architecture is composed of a number of smaller FPGA-like tiles. The vision is that future technology nodes would allow fabrication of about several million look-up tables on the chip, which provides a medium to exploit massive parallelism. Hierarchical integration of smaller programmable tiles, would allow us to manage the complexity, scale up the clock frequency, and intelligently map a given application onto our architecture.

The Super-FPGA architecture utilizes highly optimized local interconnect for intra-tile communication, combined with chip-level network-on-the-chip for inter-tile signal transfer. The chip-level interconnect provides a virtual point to point ultra-fast connection. This allows scaling of the clock frequency despite the fact that a complex design would most likely have a number of long connections.

The appropriate complexity of each tile is currently being investigated. It is well-know in the ASIC community that typical design building blocks are about 10-30K gates in size, which indirectly represents the complexity that the engineers are comfortable with to design, after modularly breaking down a large application. We suspect that a similar complexity would be appropriate for the granularity of Super-FPGA tiles. We have researched the same problem under the multi-processor chip framework and have found out the characteristics of the optimum tile [2].

Since the applications might go through control-dominated as well as data-intensive execution patterns during their lifetime, some tiles are equipped with embedded processors that show better efficiency for control-intensive portion of the application. Memory elements are distributed inside the tiles to provide rapid data load/save for intensive intra-tile computations. Global data sharing is performed via copying the contents of the memory local to the source tile, to the memory local to the destination tile. Figure 2 visualizes a rough sketch of the Super-FPGA architecture. Each tile would be an FPGA-like fabric, and different tiles can be independently customized to allow an improved mapping of the application to the system.

Advantages and Paradigm Shifts

The Super-FPGA architecture provides a highly dense programmable fabric. The fabric can be used to exploit massive parallelism in the applications, and provide the much needed extension of the compute performance growth for processing engines. The Super-FPGA is a practical system designed with the vision that area would be free with imaginary extreme density scaling. The continuation of the Moore's law for the next 5-7 years will approach this ultimate imaginary limit.

Under this extreme assumption, a one to one mapping of application operations to physical resources on the chip is possible. In other words, the need for resource sharing at different points of application runtime would be eliminated, as area is thought to be free. A direct implication of this ultimate scaling is the fact that all of application loops can be completely unrolled. Consequently, data dependency among operations would be the only deciding factor in the runtime of the application. Clearly, this provides a hypothetical upper bound on the possible achievable performance for running an application.

The Super-FPGA provides a programmable hardware medium that provides any given application with close-to-free physical resources running at about 10X clock rates of current FPGAs. Hence, the application loops can be unrolled to a large degree (under the loose constraint of area), and the compute performance would be much higher than the processors of the same generation.

Moreover, the memory bandwidth bottleneck of the processors is overcome in this model. Extreme unrolling and imaginary one to one mapping of the application operations to physical resources on the chip eliminates the memory requirements. This model replaces each data communication with local storage equipped with direct point to point wires. Wires incur smaller delay and can be programmed to transfer the data to different locations on the chip. Overcoming the memory bottleneck, or the traditional Von Neumann execution model, would allow further extension of "compute performance growth" predicted by Moore's law.

Challenges

There are, however, a number of fundamental questions that need further investigation to allow efficient manufacturing and utilization of Super-FPGA. One the architecture side, the interconnect architecture among programmable tiles has to be studied. We envision that a packet switched network on a chip would provide an efficient communication means, however this is an area that we would like to pursue further. The choice of homogeneous tile structure, or heterogeneous selection and customization of tiles is another important area for future research. The homogeneous architecture is easier to manufacture, however, individual or group-based tile customization increases the opportunity for application specific compute performance improvement at the price of more complicated compilation/mapping methodology.

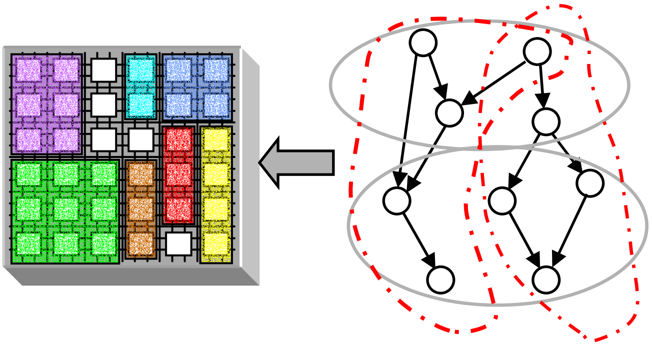

Figure 2: Heterogeneity and extreme scaling complicates the compilation task. The two mappings shown above are optimal under two different intra/inter tile communication models.

Mapping a given application onto Super-FPGA is another important area that should be researched. In fact, the 10X performance and density improvement cannot be delivered without a revolutionized virtualization mechanism and mapping technology. The Super-FPGA architecture provides a gigantic physical platform with little/no need for resource sharing, more demanding performance requirement, complicated and heterogeneous communication schemes, possibly different tile architectures, and distributed memory architecture. All of the above parameters redefine the compilation task, and call for extensive research to develop dedicated compilation tools to allow utilization of the hardware potential.

Research should be carried out to investigate the required properties, and possibly develop a dedicated application specification model and language to abstract the massive complexity of the architecture, and expose a virtualized yet powerful model to the application developer. Availability of efficient compilation tools is the key to utilization of Super-FPGA for embedded applications.

References

[1] Anant Agarwal et. al, "Evaluation of the Raw Microprocessor: An Exposed-Wire-Delay Architecture for ILP and Streams", International Symposium on Computer Architecture, 2004

[2] Venkatesh Akella, et. al " "Synchroscalar: A Multiple Clock Domain, Power-Aware, Tile-Based Embedded Processor", International Symposium on Computer Architecture, 2004

[3] Sony, Toshiba and IBM (STI), "The Cell Processor".

[4] Underwood, K.D. Hemmert, K.S, "Closing the gap: CPU and FPGA trends in sustainable floating-point BLAS performance", Symposium on Field-Programmable Custom Computing Machines, pp. 219 – 228, 2004