Reconfigurable Processor Architectures Exploiting High Bandwidth Optical Channels

M.F. Sakr[1,3], **S. P. Levitan**[1], **C. L. Giles**[3], **D. M. Chiarulli**[2]

[1]EE Department, University of Pittsburgh, Pittsburgh, PA, 15261
sakr@ee.pitt.edu
[2]CS Department, University of Pittsburgh, Pittsburgh, PA, 15260
[3]NEC Research Institute, Princeton, NJ, 08540

1 INTRODUCTION

There is growing interest in studying the possibility of reconfigurable architectures as replacements for general purpose computing for certain application domains. Reconfigurable systems can take advantage of deep computational pipelines, perform concurrent execution and are inherently data flow in nature. Furthermore, these systems have the capability of 'on the fly' reconfiguration of all or portions of the hardware to represent all the functionality required to complete the execution of an application. However, these architectures suffer from slow run time reconfiguration (RTR) due to the fact that the configuration memory resides off-chip and hence requires high access latency. This disadvantage limits the system performance and the application domain in which reconfigurable systems could prove effective. To overcome slow RTR, recent approaches include on-chip configuration memory to cache the next possible configurations [Schmit97]. This approach trades off die area for fast RTR which diminishes the processing power of the reconfigurable processor. The high cost of adding configuration cache, up to 50% of the die area, would considerably increase the number of hardware reconfigurations required compared to architectures without on-chip cache. This paper presents an alternative reconfigurable architecture which overcomes these limitations by exploiting high bandwidth optical channels. We develop a performance model to analyze and compare the performance of cache based RTR architectures, optical based RTR architectures and hybrid optical-cache based RTR architectures.

2 PROPOSED ARCHITECTURE

Optical access of configuration memory: In order to provide maximum utilization of the die area for computation and fast RTR, we propose a reconfigurable processing system which employs high speed parallel optical channels for loading configurations. An array of optical detectors is added to the die of the reconfigurable processing unit. An array of optical transmitters is mounted onto the memory module using flip-chip bonding technology (Figure 1). Each transmitter is a Vertical Cavity Surface Emitting Laser (VCSEL), while the receivers are metal-semiconductor-metal (MSM) photo detectors [OC97]. The configuration data is transmitted optically in a 2D fashion to each of the on-chip detectors achieving very fast parallel reconfiguration of the processing elements. Since the configuration can be delivered at high data rates, this allows the system the capability of reconfiguring the hardware more often than in conventional systems, hence providing more hardware resources. Further, the photo detectors are implemented directly on the silicon with little die area overhead and minor alteration of the fabrication processes. Therefore, optically reconfigurable architectures can achieve fast RTR and offer better utilization of the die area.

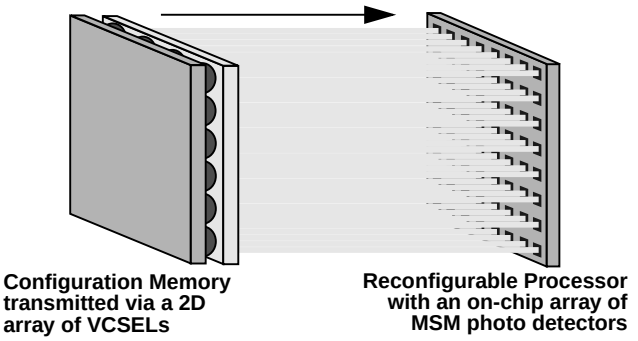

Figure 1: Reconfigurable Processor reading off-chip configuration memory using a 2D array of optical detectors.

3 PERFORMANCE ANALYSIS

We develop a performance model to measure the performance of reconfigurable architectures. Detailed analysis of the performance model can be found in [Sakr98]. We define the performance of a processor as the time a processor needs to complete the execution of an application. In general, the execution time can be defined as the total time needed per configuration of the hardware times the number of configurations required to execute the application. The time needed per hardware configuration consists of three parameters. First, is the time to load the configuration data that defines the functionality represented in the logic, T_C. Second, is the time to read/write the data needed to perform the computation, T_D. Third, is the time needed to execute the operations defined on the available data T_E. In this summary we focus only on the effects of T_C and T_E on total execution time. Therefore, for this study we define total execution time as $T = (T_C + T_E)C$, where C is the number of configurations needed. The number of configurations, C, is defined as the total number of microoperations that define the application divided by the average number of microoperations that can be represented in a single hardware configuration. This is a function of the processing area available on the chip as well as the structure of the application being executed.

For cache based RTR architectures, when reconfiguration is required the needed configuration bits could either reside in the on-chip cache or off-chip in configuration memory. The probability that a configuration can be found in the cache is denoted as P_{hit} and the miss rate as P_{miss}. Hence, $T_C = P_{hit} \times T_{onchip} + P_{miss} \times T_{offchip}$, where T_{onchip} is the time to access a single configuration from on-chip configuration cache and $T_{offchip}$ is the time to access a single configuration from off-chip configuration memory. With N electronic channels to configuration memory, $T_{offchip} = B_{config}/(N \times S_{elec})$, where S_{elec} is the bandwidth of each channel to memory and B_{config} is the maximum number of configuration bits required per configuration.

For optical based RTR architectures all configuration bits must be loaded from off-chip memory. The time needed to load configuration bits from off-chip memory depends on the number of channels (N) that connect the processor to memory and the bandwidth (S_{optic}) available per channel. Hence, $T_C = B_{config}/(N \times S_{optic})$ where B_{config} is the maximum number of bits required per hardware configuration.

4 PERFORMANCE COMPARISON

We compare the performance of three processor architectures executing a single application consisting of 10,000,000 microoperations. The processors have the same die area, G, which is 100,000 logic gate equivalents. We study the effect of adding more communication channels to configuration memory on the total execution time and the effect of adding more configuration cache on the total execution time.

The cache based RTR architecture employs on-chip configuration cache, G_{cache}, with an on-chip access time (T_{onchip}) of 2ns and N electrical channels to configuration memory each with a bandwidth (S_{elec}) of 50MHz. The cache hit rate (P_{hit}) is modeled versus cache size using the well known parachor curve [Dietel90] as follows: $P_{hit} = (1 + (G(1-L))/G_{cache})^{-1}$ where L ($0<L<1$), represents the locality of the accesses. An L value close to one indicates high locality, we use $L = 0.98$.

The optical based RTR architecture employs N optical channels to configuration memory but does not use any on-chip configuration cache. The overhead of each optical detector is 100 gate equivalents. The bandwidth (S_{optic}) of each optical channel is 500MHz. We vary N and record the effect on total execution time.

The hybrid architecture employs both on-chip configuration cache and optical channels to configuration memory. We vary the size of the on-chip cache and the number of optical channels and record the effect on total execution time.

Since we assume that all processors are fabricated using the same technology, we can assume that T_E to be the same for all architectures and can be treated as a constant (T_E=0.1ms).

In Figure 2, we show the results of the performance comparison. For the cache based architecture, we show the effect of increasing the die area used as configuration cache and the number of electronic channels on the execution time, T. The configuration time decreases as the percent of the die area that is used as configuration cache increases. However, the number of configurations, C, grows as the die area used for configuration cache increases. As illustrated by the graph, the number of configurations required for cache sizes greater than 40% dominates the advantage of the increased cache. Also, for the optically reconfigurable architecture, we plot the total execution time, versus the number of channels, N, to configuration memory. The number of optical channels is varied from 1 to 64. The more channels used to access configuration memory, the shorter the execution time. For the hybrid architecture, we plot the effect of increasing the cache size and the number of optical channels on execution time. This architecture shows the best performance when using a small cache and a few optical channels, however, we see the same effect on total execution time as in cache based designs when more of the die area is utilized for cache.

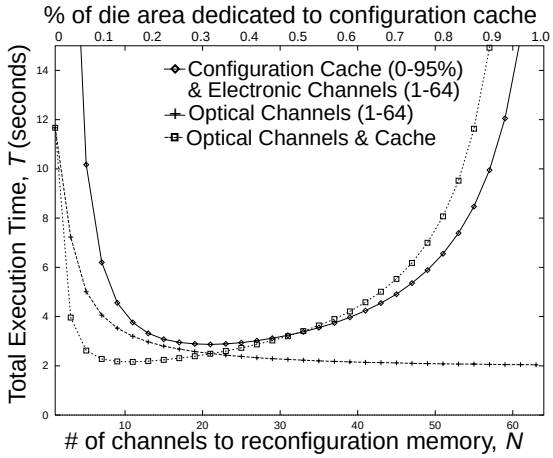

Figure 2: Total Execution vs. number of optical channels and vs. percentage of die area dedicated to configuration cache.

The analysis demonstrates that by using optical channels an improvement in performance can be achieved in comparison to systems using part of the die area for configuration cache. Further, small on-chip cache sizes prove to be very effective if high bandwidth optical channels are also employed.

References

H.M. Deitel, *Operating Systems*, Addison Wesley, 1990.

Proceedings of the International Conference on Optics in Computing OC97, Optical Society of America, Lake Tahoe 1997.

M.F. Sakr, S.P. Levitan, C.L. Giles and D.M. Chiarulli, "Reconfigurable Processor Employing Optical Channels," *Proceedings of the International Conference on Optics in Computing*, paper O40, 1998.

H. Schmit, "Incremental Reconfiguration for Pipelined Applications," *Proceedings of the IEEE Symposium on FPGAs for Custom Computing Machines*, 1997.