Towards Performance Modeling of 3D Memory Integrated FPGA Architectures

Shreyas G. Singapura, Anand Panangadan and Viktor K. Prasanna[*]

University of Southern California, Los Angeles CA 90089, USA,
{singapur, anandvp, prasanna}@usc.edu

Abstract. Recent advances in three dimensional integrated circuits have enabled large amounts of memory to be stacked in layers and accessed by a logic unit using high bandwidth vertical interconnects. Several 3D architectures have been proposed with different organizations of memory and logic layers. In particular, 3D stacks of memory dies can be interfaced with a reconfigurable logic layer such as FPGA to enable highly optimized implementation of memory-intensive applications. We refer to these as 3D Memory Integrated FPGAs. Mapping algorithms to such architectures is a challenging task due to the complex interaction between memory and logic and the relation between energy consumption and memory access. Performance modeling of these architectures can enable the design space to be systematically explored while mapping a specific algorithm. In this paper, we analyze the current landscape of 3D Memory Integrated FPGAs and identify the key parameters that have a significant impact on bandwidth and energy. We specify an "abstract architecture" that captures the features of such architectures and provide a parameterization of the design space with the eventual goal of developing a performance model for optimizing algorithm implementation.

Keywords: 3DIC, vertical interconnect, Hybrid Memory Cube

1 Introduction

VLSI technologies are reaching their limits on their ability to pack more transistors on 2-dimensional areas of silicon. Three dimensional integrated circuits (3DIC) technology is one of the most promising ways to increase device densities in accordance with Moore's law. 3DICs add a dimension to the planar IC by stacking layers vertically. These layers communicate with each other using vertical interconnects such as high bandwidth Through Silicon Vias (TSVs). 3DIC technology is most advanced in the design of memory chips. For instance, in 3D Stacked DRAMs [1], layers of DRAM memory banks are stacked on top of each other with TSVs used for vertical interconnects. A further advance is the stacking of layers of logic and memory vertically and connecting these layers

with high-bandwidth vertical interconnects. These 3D Memory Integrated Architectures can enable processing near memory with high bandwidth access to memory [2–8]. The large number of interconnects between memory and logic coupled with frequent accesses to memory results in significant energy consumption. The goal of performance model-based optimization is to enable algorithms that can exploit the resources of 3D memory integrated architectures in an energy-efficient manner.

In this paper, we focus on 3D Memory Integrated Architectures with FPGA as the logic layer and consider the issues in developing performance models of 3D Memory Integrated FPGA. A challenge for developing an efficient implementation of an algorithm on a 3D Memory Integrated Architecture is the large number of architecture features that need to be optimized. These include parameters such as number of memory banks, vaults, layers and interconnects. The complex interactions between logic and memory require more control parameters such as access cost for a row, bank, vault in terms of both time and energy. A *design* refers to the mapping of an application onto an architecture. The *design space* is the set of all possible designs that are obtained by instantiating all possible values of the different parameters. The design space of 3D Memory Integrated FPGAs is large but must be explored within a reasonable amount of computation time. It is infeasible to perform a detailed evaluation of every possible design in the design space using simulation tools since the size of the design space grows exponentially with the number of parameters. Hence, it is important to identify the specific parameters which have a significant impact on performance and develop a high-level performance model based only on these parameters. We study the current 3D Memory Integrated FPGA landscape, identify these parameters, and propose an abstract architecture consisting of the parameters that significantly affect bandwidth and energy-efficiency.

The contributions of this paper are:

– a study of the current landscape of 3D Memory Integrated FPGAs
– an "abstract architecture" that captures the features of such architectures as the basis of a high-level performance model
– identification of a set of design space parameters for the abstract architecture

The rest of the paper is organized as follows. Section 2 presents the overview of 3D Memory Integrated FPGAs. In Section 3, we identify the critical parameters required to explore the design space of 3D Memory Integrated FPGAs and we conclude in Section 4.

2   3D Memory Integrated FPGA (3D MI-FPGA)

3DIC technology is being applied to develop the next generation of FPGAs [9–12]. These include 3D FPGAs (stacking multiple layers of FPGA to benefit from the high bandwidth interconnections in 3DICs) and 3D MI-FPGAs (stacking memory and FPGA). In this paper, we focus on 3D MI-FPGAs and provide

an overview of the different 3D MI-FPGAs. We then use these architectures to identify the potential parameters for a performance model.

The 3D architecture in [11], "Reconfigurable System-in-Stack" stacks memory layers on top of FPGA layers. The architecture is made up of multiple memory layers and the logic layer is split into two FPGA layers, with one layer acting as Datapath unit (accelerator) and the other as Control unit. The datapath unit layer of FPGA is made up of application specific logic and BRAMs with interconnections to the DRAM and the control unit layer. The control unit layer implements a finite state machine to manage the datapath unit.

Recently, a 3D demonstration board developed by Altera and Micron [12] features one 3D memory device from Micron connected to four Altera Stratix V FPGAs. The 3D memory is a vertical stack of DRAM memory dies connected using TSVs [13]. Although, this board does not have the memory stacked vertically on an FPGA, it enables designers to test 3D memory models and its communication with multiple FPGAs.

2.1   Features and Performance

Reconfigurable System-in-Stack [11]: In this architecture, each memory layer has 8 ports with 256 Mb per bank. TSVs connect the memory and FPGA layers. The maximum operating frequency of 3D memory is 2 GHz and this can sustain a bandwidth of 512 GB/s with a latency of 9 ns. The interconnects between FPGA layers use 3D interconnects. The maximum frequency of the FPGA is approximately 200 MHz.

Altera + Micron Demonstration Board [12]: The number of memory layers in the 3D memory is 8 and the number of logic layers in the demonstration board is 4. There are 512 interconnects between layers of the 3D memory. Banks stacked vertically one above the other form a *rank* or *vault*. Each layer of 3D memory is made of 16 vaults and 1 bank per layer per vault. The 3D memory interface is made of four links with each link connected to one FPGA. The total bandwidth from the 3D memory is approximately 130 GB/s and approximately 32 GB/s between one FPGA and the 3D memory.

3   Parameterization of 3D MI-FPGA

One of the promising characteristics of 3D architectures is that they address the memory wall problem and hence we focus on parameters which affect the bandwidth of 3D MI-FPGAs. On the other hand, 3D architectures can be energy intensive due to heavy utilization of the memory layers. Therefore, we also specify parameters affecting power or energy consumption of 3D MI-FPGAs. We present an *abstract architecture* comprising of layers of memory and layers of logic interacting through an interconnect layer. This is illustrated in Figure 1. The design space for this abstract architecture consists of parameters defining the Memory, FPGA Logic, and Logic-Memory Interconnect (LMI) layers. These parameters are described next.
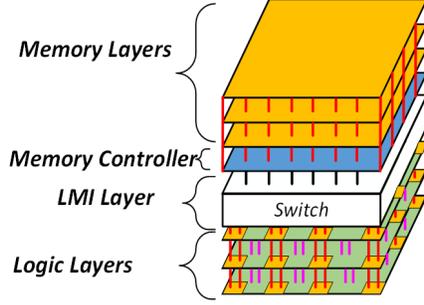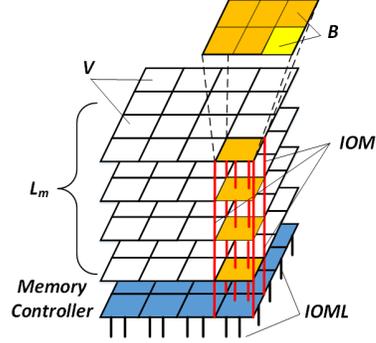
Fig. 1: 3D MI-FPGA Architecture

Fig. 2: 3D Memory

3.1 Memory Layer

Bandwidth Parameters: The bandwidth of memory is a function of parameters including its operating frequency (f_{mem}), size (M), and the number of memory layers (L_m). In the case of 3D memory, other parameters also affect bandwidth. In addition to the above, we define the following fine-grained parameters:

– Banks per vault per layer (B)
– Number of interconnections (N_{iom})
– Number of vaults (V)
– Activation overhead (t_{row})
– Interconnect latency (t_{iom})

The organization of each layer has a significant impact on the bandwidth. A representation of the 3D memory is shown in Figure 2. Each layer of memory is partitioned into banks. Group of banks on each layer share a set of interconnections to other layers. We denote this set of interconnects as *IOM*. This *group of banks* which are stacked vertically one above the other in different layers form a *vault*. All the banks belonging to the same vault share the IOM with other layers. The number of these connections, N_{iom} define the bandwidth available between the layers. In Figure 2, the memory layer can be visualized as being made of V vaults of memory sharing a group of interconnections and each vault made of B banks in each of the L_m layers. The latency of accessing memory depends on the time needed to activate a bank in a vault and on the data transfer time on the IOM. Let t_{row} denote the time to access a row in one bank of one layer and t_{iom}, the transfer time on IOM. t_{row} consists of the activation time of the bank and the data movement from the bank to IOM. The latency of accessing two rows in the same bank is higher than accessing two rows in different banks because the activation time of banks can be reduced by overlapping the activation. This pipelining of activation of banks can be extended to different layers. But, the vaults do not share the IOM and hence, all the vaults can be active at the same time.

Energy Parameters: Energy consumption of memory is affected by latency parameters (defined above) and the power parameters. In a memory, accessing two rows consumes more power than accessing one row. We denote by P_{row} the power to activate one row in a bank. In addition, the banks in memory can be set to standby mode or power-down mode. In each of these modes, there are two states: *active* and *precharge*. Keeping the bank in precharge power-down state (P_{down}) consumes the lowest power whereas active standby state (P_{act}) consumes the highest. Thus, data layout and number of row activations in a bank significantly affect the total power consumption in memory. The power consumption due to the IOM is defined by the parameter P_{iom}, the power consumed per interconnect per unit data transferred. Other parameters include read power (P_{read}) and write power (P_{write}) which are dependent on the amount of data accessed and refresh power ($P_{refresh}$), a constant.

3.2 Logic Layer

Bandwidth Parameters: An FPGA-based logic layer in our architecture consists of logic blocks, DSP blocks, memory blocks, and memory controllers (Figure 3). A functional unit can be replicated in multiple layers or a logic unit can be divided across multiple layers. We denote the set of vertical interconnects between these logic blocks by *IOL*. In addition, memory controllers on different layers share a separate set of vertical interconnects which are used to communicate with each other and the memory layers through the Logic-Memory Interconnect Layer (denoted *IOLM* and described in Section 3.3). Parameters such as the number of LUTs and slices occupied by the application determine if the logic fits on the FPGA. Similarly, the maximum frequency of operation is dependent on the application and specific implementation. The FPGA parameters that significantly impact the memory bandwidth are:

– Frequency of operation (f_{logic}) and number of layers (L_l)
– number of interconnects between logic layers (N_{iol})
– interconnect latency (t_{iol})
– configuration of on-chip memory (N_{bram}, S_{bram})
– number of memory controllers (N_{mcl})

The frequency of operation of FPGA logic determines the amount of data that can be processed by the logic layer and hence the bandwidth between logic and memory increases with f_{logic}. The number of layers of logic (L_l), size of IOL (N_{iol}), and delay from routing through IOL (t_{iol}) determine the number of parallel functional units that can run together. The IOL reduces the routing and improves the pipeline with output of logic block on one layer acting as input to logic block on another layer. The structure of on-chip memory affects the performance of FPGAs significantly. BRAMs are dedicated memory blocks with the mode of operation (read/write), the size of each block and width of each data location as configuration parameters. In each of the N_{bram} BRAMs of size S_{bram}, there are two separate read and write ports per block. The BRAMs are
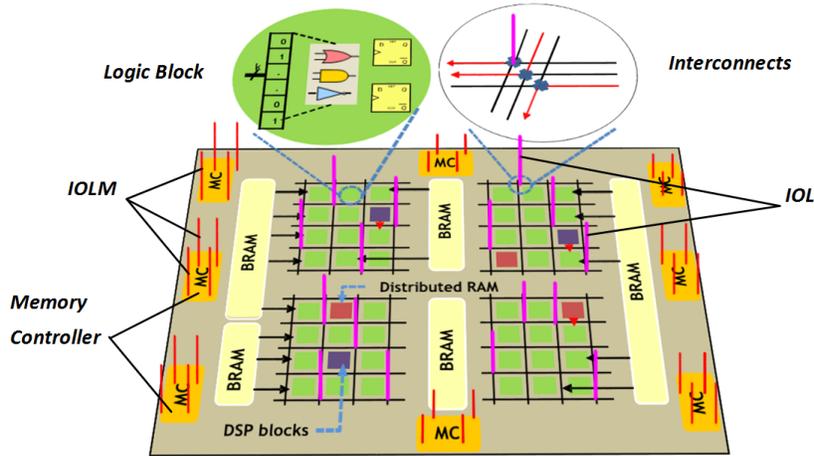
Fig. 3: FPGA-based logic layer

connected to the memory layer through N_{mcl} memory controllers. Note that we do not consider Distributed RAMs (Dist. RAMs) in this work as they are dependent on the implementation and are not dedicated memory blocks.

Energy Parameters: In an FPGA, memory is one of the major consumers of energy. We define the parameter P_{bram} as the power consumed by one block of BRAM. The total power consumption in BRAMs can be reduced by implementing a memory activation schedule that switches off blocks that are not in use. Dist. RAMs can be used as a local buffer so that the BRAM can be switched off to save energy. The power consumption of Dist. RAM, P_{dist}, is lower than P_{bram} because of its smaller size. The power consumption due to routing through IOL is determined by P_{iol}. Other components that consume power in an FPGA are slice logic, functional units, and signals. Such blocks have their own power consumption parameters, for example, Multiply and Accumulate block (P_{mac}), adder (P_{add}), slice logic (P_{slice}) and signals (P_{sig}).

3.3 Logic-Memory Interconnect (LMI) Layer

The interconnect topology shown in Figure 4 acts as the communication link between logic and memory layers carrying the address of data in the memory and the data itself. The organization of these interconnects affect the bandwidth. The parameters of the interconnect layer specify the number of interconnects, latency, and the topology of the interconnect. The aggregate bandwidth from all the vaults in the memory should be made available to the logic layers. This is determined by the topology, latency (t_{lmi}) and number of interconnects *between* the memory controller layer in memory and memory controllers in the logic layer. Denote by N_{ioml} the number of links from one vault in the memory layer to be connected to the logic layers through LMI. Then the total number of links from the memory layer (*IOML*) is $V \times N_{ioml}$.
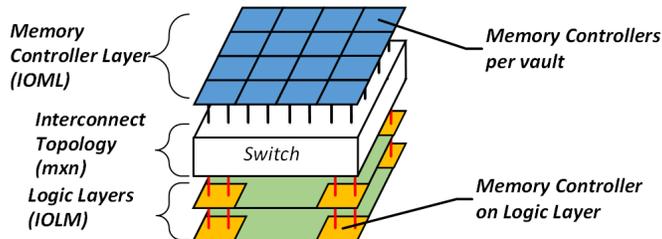
Fig. 4: Logic-Memory Interconnect

Table 1: Memory Layer Parameters [11, 12]

Parameter	L_m	M	V	B	N_{iom}
Reconfigurable System-in-Stack	2	2 GB	8	1	~ 310
Altera + Micron	8	2-4 GB	16	1	512

Similarly, denote by N_{iolm} the number of links from each memory controller on a logic layer to be connected to the memory layers through LMI. Then, the total number of links from the logic layers ($IOLM$) are $N_{mcl} \times N_{iolm}$. For example, a crossbar switch of size $m \times n$, can be used to form the interconnection topology, where $m = V \times N_{ioml}$ and $n = N_{mcl} \times N_{iolm}$. In addition, we denote by t_{lmi} to include both the setup time to configure the LMI and the data transfer time between memory controllers on memory and logic layers. P_{lmi} denotes the power consumption due to the LMI layer per connection made between the memory and logic layers.

3.4 Examples of the Abstract Architecture

The parameters defined in this section can be used to describe the architectures in Section 2. The memory layer parameters for both the architectures are tabulated in Table 1. In Altera + Micron Demonstration Board, the FPGA is not connected to the memory via 3D interconnects and hence, the parameters of the logic and the LMI layer is limited to Reconfigurable System-in-Stack architecture in Table 2.

4 Conclusion

We summarized the current state of development of 3D MI-FPGAs and identified architectural features which are critical for performance modeling. These features

Table 2: Logic and LMI Layer Parameters [11]

Parameter	L_l	S_{bram}	N_{bram} per logic unit	N_{iol}	N_{iolm}	N_{ioml}
Reconfigurable System-in-Stack	2	4 KB	5	880	~ 310	~ 310

were used in the design of an abstract architecture whose parameters can form the basis for design space exploration of 3D MI-FPGAs with memory bandwidth and energy efficiency as the primary metrics. For future work, we intend to develop a detailed performance model for a 3D MI-FPGA. We will compare the data throughput and energy consumption of representative memory-intensive applications based on performance modeling with that of detailed simulations.

References

1. G. H. Loh, "3D-Stacked Memory Architectures for Multi-Core Processors," in *ACM SIGARCH computer architecture news*, vol. 36, no. 3.   IEEE Computer Society, 2008, pp. 453–464.
2. A. Papanikolaou, D. Soudris, and R. Radojcic, "Introduction to Three-Dimensional Integration," in *Three Dimensional System Integration*.   Springer, 2011, pp. 1–12.
3. A. W. Topol, D. La Tulipe, L. Shi, D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini *et al.*, "Three-Dimensional Integrated Circuits," *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 491–506, 2006.
4. G. H. Loh, Y. Xie, and B. Black, "Processor Design in 3D Die-Stacking Technologies," *Micro, IEEE*, vol. 27, no. 3, pp. 31–48, 2007.
5. W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. D. Franzon, "Demystifying 3D ICs: The Pros and Cons of Going Vertical," *Design & Test of Computers, IEEE*, vol. 22, no. 6, pp. 498–510, 2005.
6. R. Patti, "Homogeneous 3d integration," in *Three Dimensional System Integration*. Springer, 2011, pp. 51–71.
7. B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. Mc-Cauley, P. Morrow, D. W. Nelson, D. Pantuso *et al.*, "Die Stacking (3D) Microarchitecture," in *Microarchitecture, 2006. MICRO-39. 39th Annual IEEE/ACM International Symposium on*.   IEEE, 2006, pp. 469–479.
8. S. Das, A. Fan, K.-N. Chen, C. S. Tan, N. Checka, and R. Reif, "Technology, Performance, and Computer-Aided Design of Three-Dimensional Integrated Circuits," in *Proceedings of the 2004 international symposium on Physical design*.   ACM, 2004, pp. 108–115.
9. A. Gayasen, V. Narayanan, M. Kandemir, and A. Rahman, "Designing a 3-D FPGA: Switch Box Architecture and Thermal Issues," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 16, no. 7, pp. 882–893, 2008.
10. C. Ababei, P. Maidee, and K. Bazargan, "Exploring Potential Benefits of 3D FPGA Integration," in *Field programmable logic and application*.   Springer, 2004, pp. 874–880.
11. P. Gadfort, A. Dasu, A. Akoglu, Y. K. Leow, and M. Fritze, "A Power Efficient Reconfigurable System-in-Stack: 3D Integration of Accelerators, FPGAs, and DRAM," in *System-on-Chip Conference (SOCC), 2014 27th IEEE International*. IEEE, 2014, pp. 11–16.
12. Altera and Micron. Hybrid Memory Cube Demonstration Platform. http://www.altera.com/technology/memory/serial-memory/hybrid-mem-cubes/mem-cubes.html.
13. Hybrid Memory Cube Consortium. Hybrid Memory Cube Specification. http://hybridmemorycube.org/files/SiteDownloads/HMC_Specification%201_0.pdf.