3-D Stacked Memory System Architecture Exploration by ESL Virtual Platform and Reconfigurable Stacking Memory Architecture in 3D-DSP SoC System

Hsien-Ching Hsieh, Yi-Fa Sun, Jen-Chieh Yeh, and Po-Han Huang
ICL/Industrial Technology Research Institute
195, Sec. 4, Chung Hsing Road
Chutung, Hsinchu
Taiwan, R.O.C.
{ RandyHsieh, YFSun, jcyeh, pohan }@itri.org.tw

ABSTRACT

Three-dimensional (3-D) integration promises continuous system-level functional scaling beyond the traditional 2-D device-level geometric scaling. It allows stacking memory dies on top of a logic die using through-silicon vias (TSVs) to realize high bandwidth by deploying the vertical connections between functional blocks. In this paper, we present a design strategy using ESL virtual platform to explore 3-D memory architecture for a heterogeneous multi-core system and base on exploration results, we propose the reconfigurable stacking memory architecture for three-dimension IC. Based on the virtual platform, designers can rapidly obtain the 3-D stacking interface for better system performance and TSV utilization. A feasible stacking architecture and memory interface which meets the design constraints and performance requirements has been evaluated for the target system. To demonstrate our 3-D IC design techniques, the stacking memory approach is employed in our "3D-DSP" design. In 3D-DSP, we stack 512KB SRAM directly on top of the logic die which is heterogeneous multi-core computing platform for multimedia application purpose. The logic and memory dies are fabricated in the TSMC 90nmG 1P9M CMOS process. Finally, we use 3D-DSP EVB to demonstrate the performance improvement. Real multimedia H.264 decoding experiment shows that the stacking system can achieve about 66.4% performance improvement compared to the original 2-D system.

I. INTRODUCTION

Three-dimensional integrated circuits (3-D ICs) which utilize through-silicon vias (TSVs) for vertical interconnection have been touted as an improved alternative to the packaging solutions, such as system-in-package (SiP), which rely on wire bond. This is in view of that there are many advantages adopting the TSV-based integration, including smaller signal delay due to shorter interconnect, smaller footprint by stacking multiple dies, and heterogeneous integration to mix different technologies. Specifically, stacking memory dies on top of a logic die, to address ever-increasing bandwidth and capacity requirements for multimedia applications, like video and image processing, can lead to less energy consumption than the off-chip solutions. Although the TSV technology may overcome many limitations and drawbacks on the 2-D IC design, however, several design challenges and issues remain to be addressed. For example, a complete chain of electronic design automation (EDA) tools starting from architecture exploration, verification, to physical implementation has to be established. In addition, architecture evaluation is a necessary step to explore a variety of tradeoffs in terms of performance, power, and cost at the first design stage [1].

For the 3-D system architecture exploration, some system-level design approaches have been proposed recently. In [2], the performance of stacked memory hierarchy of a RISC-system is investigated by using an analytical model. Based on the method, L2 cache and main memory implemented by die stacking can be quickly explored at the early design stage. In addition, a cost-driven 3-D IC design flow is proposed to guide the design space exploration using cost analysis [3]. It is useful to make the decision on 3-D integration strategy from different design options (e.g., number of layers, bonding approach etc). Furthermore, Tsai et al. proposed a 3-D cache delay-energy estimation tool to explore the partitioning of a cache [4]. It also supports the designers to explore the effects of cache and technology parameters. Another three-step simulation framework is also presented to explore a network on chip (NoC) based multi-processor system on chip (MPSoC) architecture [5]. Based on the environment, an MPSoC prototyping system is required for exploring the different 3-D schemes and configurations.

In this paper, we demonstrate how the ESL virtual platform can be used to assist software developers in the early application evaluation and hardware designers in exploring suitable architecture for a 3-D IC. Based on our existing SoC design as the base logic die, a stacking memory system is considered to evaluate our TSV technology and improve the system performance. With a transaction-level virtual platform, the designers can rapidly obtain the relationship between the system performance and TSV count to explore the possible implementing of the 3-D test vehicle. Finally, a 3D-DSP SoC system comprising multi-core processors and extended stacking memory is explored. Base on the exploration results by ESL virtual platform, we propose the reconfigurable stacking memory architecture to improve system performance of 3D-DSP SoC system.

II. TARGET 3-D TEST VEHICLE

Figure 1 shows the block diagram of the original 2-D heterogeneous multi-core SoC design [6,7]. It consists of one ARM microprocessor as the main processor, and dual DSP cores dedicated for computation-intensive operations.

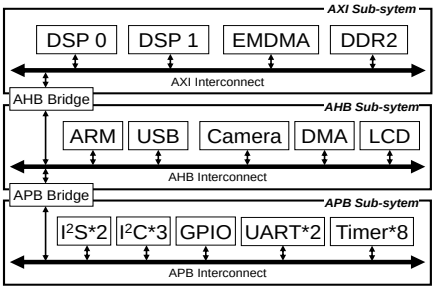


FIGURE 1. ORIGINAL 2-D HETEROGENEOUS MULTI-CORE SOC DESIGN.

For considering stacked memory on top of the original 2-D heterogeneous multi-core SoC, three possible system architectural candidates for the stacked memory have been identified. As shown in Figure 2, these are A) extending a memory space on the AXI shared bus directly, B) extending a space on each DSP's bus interface unit (BIU) as back door memory, and C) extending the local I$ and D$ of each DSP core via local instruction memory unit (IMU) and data memory unit (DMU). Try to image that if we did not adopt the ESL design methodology, these possible architecture designs should be implemented in RTL. Even the RTL design and verification efforts could take a lot of time, not to mention the

Keyword — Three-dimensional integrated circuits, Stacking, Digital signal processors, Reconfigurable architectures, Electronic Systems, System-on-chip

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

software validation and architecture evaluation. In order to explore a feasible 3-D system and architecture, we employ an existing transaction-level virtual platform which is assembled by a commercial EDA tool (Synopsys Platform Architect MCO) to verify the developed applications, explore the stacked memory architecture, and evaluate the system performance.

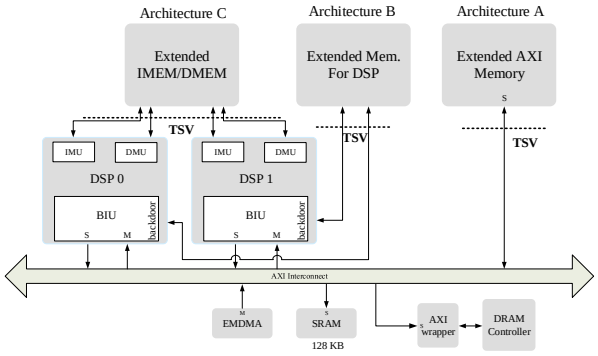


FIGURE 2. THREE STACKED MEMORY ARCHITECTURES EXTENDED FROM THE ORIGINAL DESIGN.

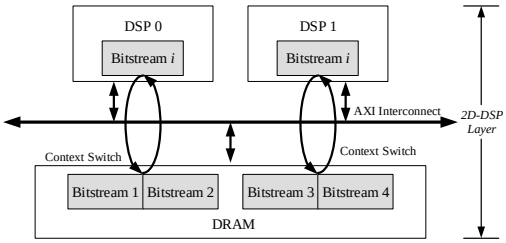
First stage, we use the virtual platform to validate the 3-D applications. According to the extended memory by stacked memory architecture, the software developers are able to access extra memory space for further system performance improvement. After the software development by the virtual platform, three possible system-level stacked memory architectures on top of base logic layer are discussed. In this stage, the total number of TSVs is one of the design constraints for the 3D-DSP SoC design. Intuitively, different stacked memory architectures lead to different levels of system performance. With the virtual platform, the designers are able to evaluate the relationship between the TSV count and the system performance rapidly. After the system-level design space exploration, the detail memory system structure exploration is considered for minimizing the number of TSV.

III. SW DEVELOPMENT BY VIRTUAL PLATFORM

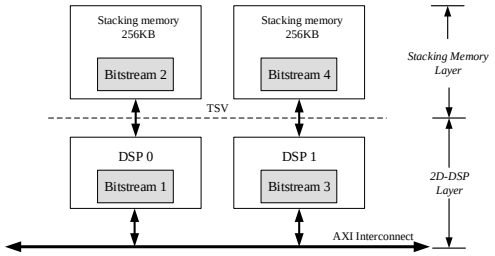
The target applications of DSP processor are mainly multimedia codec and image processing, such as H.264. Currently, few literatures addressed on the software optimization topic based on the stacking memory architecture. Most of the researches concentrate on reducing the access latency, increasing the memory bandwidth and reducing the power consumption to improve the system performance from the hardware perspective. Pan et al. present a 3-D stacking technology to bond one VLIW processor die with multiple DRAM dies [8]. On the contrary, the DSP in this paper is 5-way VLIW architecture and equipped with 64 KB instruction cache and 64 KB local data memory. The stacking SRAM is regarded as extension memory for instruction or local memory. In order to improve the system performance, the programmers have to rethink the software algorithm and the data structure to efficiently use the extended stacking SRAM.

In recent years, video surveillance system is getting more attention duo the security requests. Multi-channel video decoding and displaying is necessary for H.264-based surveillance system. Based on the stacking memory characteristic, we need to redesign software architectures to improve the performance. In this section we present a four-channel H.264 decoder to demonstrate the performance of 3-D stacking memory architecture. In this paper, the H.264 decoder is assembly coded. The decoded header information, variables, coefficients, buffers and temp memories are defined in the data structure and are about 60 KB of current implementation. This data structure is allocated in the 64 KB local data memory. For multi-channel H.264 decoder, frame-based synchronization scheme is implemented which means same frame number of all bit streams will be processed in sequence.

To extend the single-channel to the multi-channel H.264 decoder, the data structure allocation of four-channel H.264 decoder based on 2D-DSP and 3D-DSP platform are shown in Figure 3.



(A) FOUR-CHANNEL H.264 DECODER ON 2D-DSP



(B) FOUR-CHANNEL H.264 DECODER ON 3D-DSP

FIGURE 3. DATA STRUCTURE ALLOCATION OF FOUR-CHANNEL H.264 DECODER BASED ON 2D-DSP AND 3D-DSP PLATFORM

In 2D-DSP architecture, there is only 64 KB local data memory inside DSP processor, therefore the complete data structure of four bit streams must be allocated on external DRAM as shown in Figure 3(A). When processing to the corresponding bit stream, the data structure inside DSP local memory must be moved out to the external DRAM first and then copy the correlated data structure from the DRAM to the local memory before execution. The additional context switches results in extra memory accesses and decreases the system performance. For the 3D-DSP architecture in this paper, there are additional 256 KB SRAM stacked on each DSP processor. Therefore, each DSP can keep first bit stream data structure in 64 KB local data memory and another bit stream data structure on the stacked stacking memory as shown in Figure 3(B). The unwanted context switch is greatly reduced in the stacking memory architecture.

IV. STACKED MEMORY ARCHITECTURE EXPLORATION

In the original 2-D design, the embedded memory (SRAM) access interface does not pose a critical issue, because there is no special I/O requirement; all connections are internal. However, when we deal with the stacking of a memory die on top of a processor or logic die, the TSV related design rules (i.e., pitch, size, keep-out etc.) and their number have to be taken into consideration. To determine the organization of 3-D stacked memory, we need a system-level design methodology to analyze the system performance and the number of TSVs, examining their tradeoffs at the early design stage. The number of TSVs may be one of the most important indices for 3D-DSP SoC design evaluation. Our target is set to keep the total number of TSVs below 4,000 so as to occupy an

effective area no bigger than 5% of the die size (4mm×4mm). A half of the TSVs are used for signal transmission.

As mentioned before, the dual-DSP core H.264 decoders with four channels of bit streams are used as the benchmarks in this analysis. According to the 3D-DSP SoC virtual platform to explore the feasible stacked memory architectures, the experimental results are summarized in Table 1, where the number of execution cycles and the number of TSV are easily obtained by the virtual platform.

TABLE 1. EXPERIMENTAL RESULTS OF REQUIRED EXECUTION CYCLES AND TSVS FOR DIFFERENT STACKED MEMORY ARCHITECTURES.

Architectures	Number of Execution Cycles	Improvement	Number of TSVs
Original 2-D design	164,531,751	-	-
A) Stacked on AXI bus	162,312,919	1.39%	272
B) Stacked on DSP BIU	162,308,407	1.39%	544
C) Stacked on DSP IMU/DMU	121,891,167	34.89%	1,886

Architecture A) improves the execution time by 1.39%, because the extended stacked SRAM connected to the AXI bus can be used as a small and fast temporal memory for DSP status backup on context switching for different channels in the H.264 decoder. The performance improvement is minor. Because the internal status data are moved by DMA to external, traffic on the AXI bus is still very heavy in this architecture. Nevertheless, stacking memory via the AXI shared bus is easily extended and it only needs 272 signal TSVs.

Architecture B) improves the execution time only slightly by stacking memory on the DSP bus interface. In this case, the extended memory is also treated as a dedicated external memory for DSP, so it can be a temporal space for context switched data moved by the DMA in the H.264 application for different bit streams as well. Status exchange with the external memory causes huge overhead on the performance. The number of signal TSVs increases to 544, which is twice as large, compared to that by Architecture A). The above two designs, stacking the extended memory outside of the DSP core, have poor performance improvement on the multi-channel H.264 decoding. It is obvious that the design on the AXI shared bus would be the better choice for it is easily extended with little design effort for other applications.

Architecture C) is proposed by stacking memory on DSP's IMU and DMU. The system performance is improved by 34.89%, because context switch for DSP status is completely eliminated and all information can be stored in the local memory without any external data movement. The huge traffic on the system bus can be minimized and the bus contention is reduced. Although connecting such a dedicated memory to IMU and DMU takes 1,886 signal TSVs. Due to each stacked memory consists of multiple banks and each bank consists of multiple SRAM cores. The number of TSV is counted by these memory cores' data width and control signals. It is acceptable in our project as long as the performance improvement is higher than 25%. However, the TSV utilization still be considered carefully.

V. MEMORY SYSTEM STRUCTURE EXPLORATION AND RECONFIGURABLE STACKING MEMORY ARCHITECTURE

After the system-level architecture exploration, the detail memory system structure exploration will be investigated and discussed in this section. In the original design, memory organization and memory access interface of each DSP core are with eight banks and six memory access ports, respectively. These memory access ports are assigned by the following: P0 for scalar, P1 for cluster 1, P2 for cluster 2, P3 for DSP DMA, P4 for customized function unit, and P5 for BIU (Bus Interface Unit).

A straightforward method is to stack the extra 256KB SRAM in a manner similar to the embedded one, with the same numbers of banks and access ports without any modification. The bank decoder and memory access port interface are implemented on the base logic layer. However, the number of signal TSVs becomes 1,176 just for the DMU of each DSP core. Subsequently, the partial function of DSP's DMU is implemented in the stacked memory layer. The memory access interface between the stacked memory and the DSP core maintains the same arrangement of six memory access ports. The number of TSVs is reduced to 936 for the DMU of each DSP core. Both DSP architectures do not lead to system performance loss, but their TSV counts are not satisfactory, considering that the IMU requires about the same number of TSVs.

Hence, we analyzed our H.264 application in order to remove the redundant memory access port, if there is any. P4 for customized functional unit is identified, since there is no co-processor present in our application. The number of TSVs is reduced to 788; however, it is still out of our expected range. For TSV count reduction, the memory access interface has to be simplified further. The port sharing concept has come to our attention. First of all, we need to find out the ports which have less data contention if their accesses are shared. Based on the analysis by using the ESL virtual platform, Table 2 lists the data contention ratio of each port, which is derived by the extra time due to contention divided by the total system execution time.

TABLE 2. PORT CONTENTION RATIO OF DSP DATA MEMORY.

Ports	P0	P1	P2	P3	P5
P0		225,843 (0.29%)	197,991 (0.26%)	13,541 (0.02%)	18,515 (0.02%)
P1	225,843 (0.29%)		6,894,847 (9.13%)	83,068 (0.11%)	56,172 (0.07%)
P2	197,991 (0.26%)	6,894,847 (9.13%)		**26,005 (0.03%)**	53,415 (0.07%)
P3	13,541 (0.02%)	83,068 (0.11%)	**26,005 (0.03%)**		2,294 (0.00%)
P5	18,515 (0.02%)	56,172 (0.07%)	53,415 (0.07%)	2,294 (0.00%)	

As shown in the table, the most severe contention occurs between P1 and P2, which occupies about 9.13% of the system execution time. On the other hand, the data contention ratio between P3 and P5 is 0.003% which has almost no influence on the system performance at all. Thus, to merge P3 and P5 into one port seems a good choice in the 3-D stacked memory design. By doing so, the number of TSVs is reduced to 640 and there is virtually no performance loss for our application.

After the port sharing analysis, we use the virtual platform to evaluate the bank structure of DSP's DMU. According to the target application and virtual platform, the number of banks capable of parallel access, and the system performance analyses for full-interleaving/non-interleaving are evaluated. Figure 4 shows the experimental results. In the evaluation, the dual DSP cores, DMA and BIU will access the stacking memory for data computations and data movement simultaneously. Encouraged by the result, the number of bank is four is good enough. Besides, the full-interleaving addressing mode is more suitable for the target H.264 video decoding. Under the full-interleaving mode, opting for four banks increases the data contention count by 637,740 cycles, compared to eight banks, which is about 0.86% of the total

execution time. It indicates that choosing the 4-bank memory architecture needs 600 TSVs and only suffers system performance loss by 0.86%. Our experimental results by ESL virtual platform show that, for extended data memory for each DSP core, the number of signal TSVs can be reduced from the straightforward implementation of 1,176 to 600 with the minimum performance loss in the four-channel H.264 video decoding application. Both requirements on TSV usage and system performance are met.

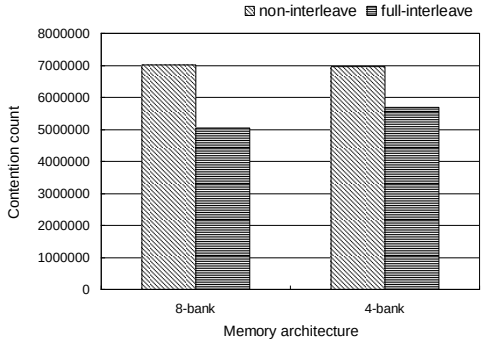


FIGURE 4. CONTENTION COUNT VERSUS NUMBER OF MEMORY BANKS WITH INTERLEAVED AND NON-INTERLEAVED MEMORY ACCESSES.

The reconfigurable stacking memory structure will be investigated and discussed in this section. There is an extended 256 KB of SRAM is allocated for each DSP processor. In order to provide high flexibility of memory architecture, memory reconfigurable features are developed. It allows users to configure different architectures of stacking memories for different applications. For example, users can configure a part of stacking memory as instruction memory and the others as data memory or treat the entire memory as data memory. The stacking memory architecture of 3D-DSP and memory bank organization is shown as Figure 5. There are two advantages of different size of memory model. First, it is more convenient for configuring the size of DMEM and IMEM. Second, the area of one 32KB memory bank is smaller than two 16KB memory banks, which mean more area for TSVs.

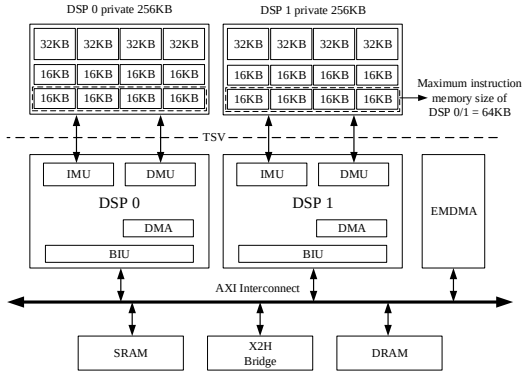


FIGURE 5. THE STACKING MEMORY ARCHITECTURE OF 3D-DSP AND MEMORY BANK ORGANIZATION

VI. EXPERIMENTAL RESULTS ABOUT PERFORMANCE AND DESIGN SPECIFICATION OF 3-DSP

A. Experimental Results

The performance improvement from 2D-DSP to 3D-DSP of multi-channel H.264 decoder is shown in Table 3. Based on stacking SRAM benefits, the performance is increased from 12.5 fps to 20.8 fps.

TABLE 3. PERFORMANCE IMPROVEMENT OF H.264 DECODER FROM 2D-DSP AND 3D-DSP ARCHITECTURE

Platform	Method	Frame Per Second (fps)	Improvement
2D-DSP	Reference data in DRAM	12.5 fps	-
3D-DSP	Reference data in stacking memory	20.8 fps	66.4%

B. Design specification

This section shows the design specification of 2D-DSP and stacking memory. 2D-DSP (Base Logic Die) and stacking memory die are fabricated in the TSMC 90nmG 1P9M process. Figure 6 shows the 3D-DSP evaluation board and photographs of the logic and memory dies. Originally, the thickness of the base logic die is 725um and after grinding is 200um. The thickness of the stacking memory die is 60nm.

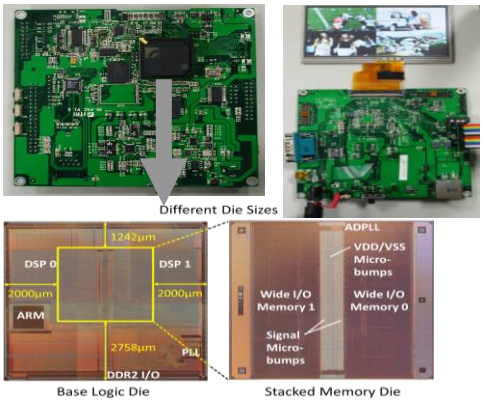


FIGURE 6. PHOTOGRAPHS OF THE LOGIC AND MEMORY DIES

VII. CONCLUSION

In this paper, we presented the 3-D IC architecture exploration framework by using the ESL design methodology. With the framework, we evaluated and explored three feasible stacking memory architectures and corresponding software applications, based on the proposed 3D-DSP SoC virtual platform. Detailed investigation inside the DSP core was performed for further improvement. Our experimental results show that, for extended data memory for each DSP core, the number of signal TSVs can be reduced from the straightforward implementation of 1,176 to 600 with the minimum performance loss in the four-channel H.264 video decoding application. Both requirements on TSV usage and system performance are met. It demonstrates that using the virtual platform is able to assist the designers to easily achieve the hardware/software co-design and architecture exploration at the early design stage. The design has been completed now. we propose the reconfigurable stacking memory architecture for 3D-DSP SoC system. The stacking memory is integrated with IMU and DMU on dual DSPs. Since stacking memory exhibit features that could reduce memory access latency and configure as instruction cache, software developers could redesign software architectures to improve the performance. Our experiments show that about 66.4% performance improvements of multi-channel H.264 decoder based on the 3D-DSP architecture. In the near future, we plan to examine other design parameters, such as DRAM timing and power, on the 3-D IC system to consider heterogeneous integration via the ESL virtual platform.

REFERENCES

[1] R. Puri, and D. S. Kung, "The dawn of 22nm era: design and CAD challenges," in Proc. VLSI Design, Jan. 2010, pp. 429-433.

[2] M. B. Kleiner, S. A. Kuhn, P. Ramm, and W. Weber, "Performance improvement of the: memory hierarchy of RISC-systems by application of 3-D technology," IEEE Trans. Components, Packaging, and Manufacturing Technology – Part B, vol. 19, no. 4, pp. 709-718, Nov. 1996.

[3] D. Xiangyu, and X. Yuan, "System-Level Cost Analysis and Design Exploration for Three-Dimensional Integrated Circuits (3D ICs)," in Proc. Design Automation Conf., pp. 234-241, Jan. 2009.

[4] Y. F. Tsai, F. Wang, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, "Design Space Exploration for 3-D Cache," IEEE Trans. Very Large Scale Integration Systems, vol. 16, no. 4, pp. 444-455, Apr. 2008.

[5] A. Richard, D. Milojevic, F. Robert, A. Bartzas, A. Papanikolaou, K. Siozios, and D. Soudris, "Fast Design Space Exploration Environment Applied on NoC's for 3D-Stacked MPSoC's," in Proc. Architecture of Computing Systems, pp. 1-6, Feb. 2010

[6] T.J. Lin, et al, "Overview of ITRI PAC project – from VLIW DSP processor to multicore computing platform", in *Proc. VLSIDAT*, Apr. 2008.

[7] Tien-Wei Hsieh, et al., "Energy-effective design & implementation of an embedded VLIW DSP," in *Proc. ISOCC*, Apr. 2008.

[8] Y. Pan and T. Zhang, "Improving VLIW processor performance using three-dimensional (3-D) DRAM stacking," IEEE International Conference on Application-specific Systems, Architectures and Processors, pp. 38-45, July 2009.