M. Hima Bindu* et al.
(IJITR) INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH
Volume No.4, Issue No.6, October – November 2016, 4702-4704.

Scalable Framework With Identical Memory Dies To Achieve High-Clock Frequency

M. HIMA BINDU
M.Tech Student, Dept of ECE
SKR College of Engineering & Technology
Nellore, Andhra Pradesh, India

G MAHENDRA
Associate Professor, Dept of ECE
SKR College of Engineering & Technology
Nellore, Andhra Pradesh, India

Abstract: **Our design implements a scalable 3-D-nonuniform memory access (NUMA) architecture according to low latency logarithmic interconnects, which enables stacking of multiple identical memory dies (MDs), supports multiple outstanding transactions, and achieves high clock frequencies because of its highly pipelined nature. We implemented our design with STMicroelectronics CMOS-28-nm low-power technology and acquired time frequency of 500 MHz, as much as eight stacked dies (4 MB) having a memory density loss. Large needed size, and ability to tolerate latency and variations in memory access time make L2 memory a appropriate choice for 3-D integration. Within this paper, we present a synthesizable 3-D-stacking L2 memory IP component, which may be mounted on a cluster-based multicore platform through its network-on-nick interfaces offering high-bandwidth memory access with low average latency. Benchmark simulation results show adding 3-D-NUMA to some multicluster system can result in a typical performance boost of 34%. In addition, experiments and estimations make sure 3-D-NUMA is energy and power efficient, temperature friendly, and it has improvements appropriate for low-cost manufacturing. Finally, improvement is quite possible in 3-D-NUMA in contrast to its 2-D counterparts, while using condition from the art through-plastic-via technologies.**

Keywords: **3-D Integration; No Uniform Memory Access (NUMA); Physical Implementation; Tightly Coupled Data Memory;**

I. INTRODUCTION

Several vertical interconnect technologies happen to be explored, including wire connecting, micro bump, contactless, and thru-plastic-via (TSV) vertical interconnect. Included in this, the TSV approach has acquired recognition, because of the high interconnection density. It's been predicted that 3-D TSV nick market will grow greater than ten occasions quicker than the worldwide semiconductor industry Nonetheless, time for adoption of three-D integration for mass production keeps shifting out to return [1]. Several technical challenges and infrastructure issues are delaying high-volume manufacturing of TSV technology for several-D ICs. Until these problems could be resolved, alternative packages will still be used. TSV plastic interposer (TSI) is a great one of methods heterogeneous dies with mixed technologies could be integrated at greater levels and greatly reduces die complexity and price. Heterogeneous integration, system miniaturization and versatility, and block level testability are the several features provided by SiP solutions. Additionally, they offer a way to integration of planar IC with 3-D-IC technology. Among the greatest motorists for top-volume adoption from the 3-D integration technologies are 3-D memory stacking with three primary classes of: 1) 3-D DRAM primary recollections 2) 3-D caches and three) 3-D scratchpad recollections (SPMs). The Three-D stacking of caches, that is a strategy still at advanced development and research stage, continues to be intensively investigated. In comparison with caches, SPMs may be seen in the machine-on-nick (SoC) memory map and therefore are appropriate for data structures, which aren't well managed through caches. L1 SPMs offer really low latency use of a cluster of tightly coupled processors. However, their 3-D stacking isn't so advantageous with current TSV technologies, as their access latency directly affects processor pipeline, and on your journey to, the 3rd dimension requires propagation of combinational signals through stacked TSVs, which aren't yet far better when it comes to speed than global on-nick wires [2]. Therefore, lower sensitivity of L2 SPMs to gain access to latency and it is variations means they are a far more interesting choice for going toward the 3rd dimension. Within this paper, we present 3-D-NUMA, an L2 memory IP created for integration like a 3-D stacked module, which may be mounted on a cluster-based multicore platform through its network-on-nick (NoC) interfaces (NIs), offering high-bandwidth memory access with low average latency. Our suggested IP is really a synthesizable and scalable NUMA architecture, which enables modular stacking of multiple memory dies (MDs) with identical layouts utilizing a single mask set, supports multiple in-flight transactions, and achieves high clock frequency, due to its highly pipelined nature. We acquired time frequency of 500 MHz, restricted to the access duration of the memory array hard macros, whereas another components can operate as much as 1 GHz. Benchmark simulation results show inclusion of this IP to some multicore NoC can provide a typical performance boost of 34%

within the situation where memory banks with similar total size are directly connected to the NoC interfaces.

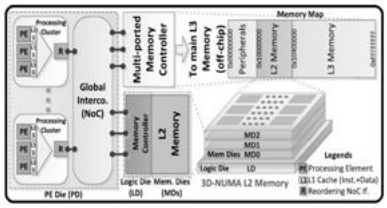

Fig.1.Framework of the proposed system

II. PROPOSED SYSTEM

The Three-D-NUMA is really a 3-D L2 memory stack made to be mounted on cluster based multicore platforms having a global NoC connecting all of the clusters, and every cluster composed by multiple tightly coupled processors. This memory IP is perfect for serving L1 cache refill/write-back instructions, since it's been made to serve load and store packets of various sizes [3]. The term-level interleaved (WLI) organization found in this memory system enables for breaking a Load64Bytes command into Load8Bytes instructions and dispatching these to eight parallel memory cones. When request packets of various sizes reach the NIs, the request engines (REs) break the input packet into smaller sized units known as chunks and issues them in parallel towards the arbitration trees (ATs), in which a pseudo round-robin arbitration is conducted one of the demands coming from various Nis. Each request travels with the memory pipeline, as well as in each MD, an incomplete address check is conducted in Fork to recognize if the request belongs compared to that MD. If matched, memory access is conducted along with a fact is came back within the response path with the Join modules. Response pathways are shared one of the read buffers (RBs), and straightforward return-address decoders issue valid signals (resp. valid component) towards the destination. Because the response chunks coming from various memory cones may arrive from order (OOO) and also at different occasions, an information structure known as RB is required to merge them, build response packets to original demands, and serialize them with the NI. It ought to be noted the access duration of the MDs increases using their indices (NUMA behavior), since all MDs are separated by pipeline registers and packets flow with these registers in every cycle. This selection enables for scalability, facilitates stacking of recent MDs having a single mask set, and modularly boosts the memory size without having affected the time frequency. The important thing property of the soft IP is configurability through several parameters highlighted. N is the amount of independent NoC interfaces, which are utilized to attach 3-D-NUMA to some NoC. C is the amount of parallel memory cones. This parameter defines the utmost possible quantity of words, which may be fetched in parallel throughout a load operation. Maximum outstanding transaction (MOT) defines the utmost permitted in-flight transactions within the memory system. This parameter directly affects the depth and complexity from the RBs. RB is among the most significant components in 3-D-NUMA, because it accomplishes different purposes [4]. First, it enables supporting multiple outstanding transactions and decouples request and response pathways completely, by accepting as much as MOT demands and granting them while their responses aren't ready yet. This can help to make use of the bandwidth from the memory pipeline more proficiently, by hiding the response latency. Finally, all of the header and control items of the input packet are kept in the RB to prevent propagating them with the whole memory pipeline. When response data returns in the memory pipeline, response packet is made by using this stored information. Flow control within the LD from the 3-D-NUMA memory system is dependent on request-grant handshaking and supports full bandwidth operation of 1 transaction per cycle, whereas the memory pipeline continues to be developed in a grant less and simple fashion. Fork is really a combinational module that transmits request chunk towards the memory array about this MD in situation of the address match, and otherwise forwards it to another MD. Join receives response chunks out of this MD and also the upper ones through small FIFOs designed for this function, choose between these questions round-robin fashion, and transmits the champion back lower within the memory pipeline. RE accounts for decoding request packets and issuing these to the memory pipeline. Physical style of 3-D-NUMA continues to be performed in line with the STM bulk CMOS-28-nm low-power technology library, having a multi VTH synthesis flow with synopsis design compiler graphical, and put and route in pedal rotation SoC encounter digital implementation. For that memory arrays, high density industrial hard macros, supplied by the STM company within the same technology, happen to be utilized. TSV fabrication yield is an important parameter in manufacturing yield and price from the final stack. Within this design, we've utilized a minimal overhead TSV repair mechanism able to supplying recovery rate by having an overhead of 1 TSV for every block of 25. This could supply the chance to enhance power delivery, IR-drops, and manufacturing yield and price. For power delivery to three-D-NUMA, several factors should be thought about and various analyses for example IR-drop, and current droops in 3-D power distribution systems ought to be performed [5]. The Three-D-NUMA can offer a great possibility for

manufacturing yield improvement in comparison to its 2-D counterpart.

III. CONCLUSION

Our design implements a scalable 3-D NUMA architecture, enables stacking of multiple identical MDs, supports multiple outstanding transactions, and achieves high clock frequencies because of its highly pipelined nature. Within this paper, we presented a synthesizable 3-D-stacking L2 memory IP component (3-D-NUMA) that could be mounted on a cluster-based multicore platform through its NIs, offering high-bandwidth memory access with low average latency. We implemented 3-D-NUMA with STM CMOS-28-nm low power technology and acquired time frequency of 500 MHz, restricted to the access duration of the memory arrays while its logic components could operate as much as 1 GHz. Benchmark simulation results show inclusion of 3-D-NUMA to some multicluster system can result in a typical performance boost. In addition, experiments and estimations confirmed that 3-D-NUMA is energy and power efficient, temperature friendly, and it has improvements appropriate for low-cost manufacturing: PCM architectural clock gating mechanism was suggested to lessen power consumption. Finally, 2.3× cost reduction was reported due to identical MD layouts together with improvement in contrast to the two-D counterpart, using the condition from the art TSV manufacturing technologies. PIMD configuration could reduce maximum temperature in comparison to the traditional memory on the top configurations.

IV. REFERENCES

[1] Y. Zhang, H. Oh, and M. S. Bakir, "Within-tier cooling and thermal isolation technologies for heterogeneous 3D ICs," in Proc. IEEE Int. 3DIC, Oct. 2013, pp. 1–6.

[2] J. H. Lau, "TSV interposer: The most cost-effective integrator for 3D IC integration," Electron. Optoelectron. Res. Lab., ITRI, Hsinchu, Taiwan, Tech. Rep. ASME InterPACK2011-52189, Sep. 2011.

[3] C. Bienia and K. Li, "Parsec 2.0: A new benchmark suite for chip-multiprocessors," in Proc. 5th Annu. Workshop Model., Benchmarking, Simul., Jun. 2009.

[4] R. G. Dreslinski et al., "Centip3De: A 64-core, 3D stacked near-threshold system," IEEE Micro, vol. 33, no. 2, pp. 8–16, Mar./Apr. 2013.

[5] V. Solberg, S. McElrea, and W. Zohni, "New 3D packaging approach for next generation high performance DRAM," Invensas Corporation, San Jose, CA, USA, Tech. Rep., 2012.

AUTHOR's PROFILE

M. Hima bindu completed her Btech in skr college of engineering and technology in 2014. Now pursuing Mtech in Electronics & Communication Engineering in SKR College of Engineering & Technology, Manubolu

G Mahendra , received his M.Tech degree, currently He is working as an Associate Professor in SKR College of Engineering & Technology, Manubolu

