International Journal of Applied Sciences, Engineering and Management
ISSN 2320 – 3439, Vol. 04, No. 05, September 2015, pp. 37 – 45

A Modular Shared L2 Memory Design for 3-D Integration

P. Aparna[1], B. Kishore Kumar[2]
[1]Mtech Scholar, JB Institute of Engineering and Technology. Hyderabad.
[2]Associate Professor, JB Institute of Engineering and Technology. Hyderabad.

Abstract: Recently, three-dimensional integration technology has allowed researchers and designers to explore novel architectures for computing systems. Due to the memory-intensive nature of signal processing systems, DSPs can greatly benefit from 3D memory integration technology realized by vertically stacking high-density memory below processing cores. In this paper, we analyze the energy and performance impacts of 3D memory integration in DSP systems by exploring a wide variety of memory configurations the technology enables. Large required size, and tolerance to latency and variations in memory access time make L2 memory a suitable option for 3-D integration. In this paper, we present a synthesizable 3-D-stackable L2 memory IP component, which can be attached to a cluster-based multicore platform through its network-on-chip interfaces offering high-bandwidth memory access with low average latency. Our design implements a scalable 3-D-nonuniform memory access (NUMA) architecture based on low latency logarithmic interconnects, which allows stacking of multiple identical memory dies (MDs), supports multiple outstanding transactions, and achieves high clock frequencies due to its highly pipelined nature.

Keywords: *3D memory integration, 3-D-nonuniform memory access, memory dies, 3-D-stackable L2 memory.*

INTRODUCTION

Three dimentional integration has been explored in academia and industry for over a decade now, and a wide variety of technologies, materials, and processes have been used for research and demonstrations. Several vertical interconnect technologies have been explored, including wire bonding, microbump, contactless (capacitive or inductive), and through-silicon-via (TSV) vertical interconnect [1]. Among them, the TSV approach has gained popularity, due to the high interconnection density. It has been predicted that 3-D TSV chip market will grow more than ten times faster than the global semiconductor industry [2]. In addition, wafer foundries such as Samsung and TSMC have been developing vertical integration offerings to meet with the demand from leading fabless companies such as Qualcomm, Broadcom, Marvell, nVidia, and Apple, along with fab-lite IC companies such as TI, STMicroelectronics (STM), and NEC/Renesas [2]. Nevertheless, the time for adoption of 3-D integration for mass production keeps shifting out into the future. Several technical challenges and infrastructure issues are delaying high-volume manufacturing of TSV technology for 3-D ICs. Until these issues can be resolved, alternative packages will continue to be used [3].

Complex system in package (SiP) solutions offered by companies, such as TESSERA, Amkor Technology, and INVENSAS, address a potentially large need in the market and are being recognized as the next industry thrust. Heterogeneous integration, system miniaturization and flexibility, and block level testability are some of the several features offered by SiP solutions. In addition, they provide a path to integration of planar IC with 3-D-IC technology [4]. TSV silicon interposer (TSI) is a good example of how heterogeneous dies with mixed technologies can be integrated at higher levels and greatly reduce die complexity and cost [5].

One of the biggest drivers for high-volume adoption of the 3-D integration technology is 3-D memory stacking with three main classes of: 1) 3-D DRAM main memories; 2) 3-D caches; and 3) 3-D scratchpad memories (SPMs) [6], [7]. DRAM dies stacked as the last level memory on the processor dies offer large capacity and high bandwidth through 3-D wide I/O interfaces. However, lack of process compatibility between DRAM dies and logic dies (LDs) has imposed several limitations and conservative design rules, which has limited so far the adoption of wide I/O memory in products [8], [9]. The 3-D stacking of caches, which is an approach still at advanced research and development stage, has been intensively investigated, as well. In contrast with caches, SPMs are visible in the system-on-chip (SoC) memory map and are suitable for data structures, which are not well managed through caches. L1 SPMs offer very low latency access (1–2 Cycles) to a cluster of tightly coupled processors. However, their 3- D stacking is not so beneficial with current TSV technologies, because their access latency directly affects processor pipeline,which are not yet much better in terms of speed than global on-chip wires. Therefore, lower sensitivity of L2 SPMs to access latency and its variations makes them a more interesting option for going toward the third dimension. In addition, most application processors and almost all mobile SoCs feature a pretty large on-chip L2 memory, which is shared by multiple cores. Snapdragon 800 processors by Qualcomm with 2 MB of L2 Cache, Exynos 5 by Samsung with 1 MB of L2

Cache, and Keystone II by Texas Instruments with 4 MB of shared L2 memory plus 4 MB of private L2 space configurable as cache or SPM, are great examples in this context.

In this paper, we present 3-D-NUMA, an L2 memory IP designed for integration as a 3-D stacked module, which can be attached to a cluster-based multicore platform through its network-on-chip (NoC) interfaces (NIs), offering high-bandwidth memory access with low average latency. Our proposed IP is a synthesizable and scalable NUMA architecture, which allows modular stacking of multiple memory dies (MDs) with identical layouts using a single mask set, supports multiple in-flight transactions, and achieves high clock frequency, because of its highly pipelined nature. We have implemented the memory IP in STM CMOS-130-nm low-power technology with up to eight stacked MDs with a memory density loss of in the MDs. We obtained a clock frequency of 500 MHz, limited by the access time of the memory array hard macros, whereas the other components can operate up to 1 GHz. Benchmark simulation results demonstrate that addition of this IP to a multicore NoC can give an average performance boost of 34% over the case where memory banks with the same total size are directly attached to the NoC interfaces. The 3-D-NUMA is able to deliver a bandwidth of 56.4 GB/s (88.2% of the theoretical limit) with an average memory access time (AMAT) of 37.2 Cycles, for maximum pressure full-bandwidth uniform-random traffic applied to its ports and with a stack of eight MDs. Furthermore, experiments confirm that 3-D-NUMA is energy efficient and temperature friendly, reducing power by over 38% (by means of architectural clock gating) and temperature by over 40° (due to its temperature friendly 3-D organization). Finally, from the manufacturing cost point of view, a 2.3× reduction is offered compared with similar designs with nonidentical MDs, and up to 22% yield improvement can be obtained compared with its 2-D flat version, with current TSV technologies.

LITERATURE SURVEY

Advanced packaging technologies provide new opportunities for heterogeneous integration, power delivery, cost optimization, and thermal management. Stacked chip scale packaging of Amkor Technology is one such example, which provides several different 2.5-D/3-D options for integration of heterogeneous dies in a package. Among other packaging technologies, dual DRAM package, dual face down, and quad face down with the main target of DRAMs provide complex forms of wire bonding, which may be adopted even for other levels in the memory hierarchy. Technologies, such as TSI [5] and wafer reconstitution [4], provide even more flexibility in hybrid 2.5- D/3-D stacking. TSIs enable stacking of different dies on both sides to achieve a better utilization of space and facilitate heat transfer of high-power chips. Wafer reconstitution provides electrical connections from the chip pads to the interconnects by means of an artificial wafer. Redistributed chip packaging (RCP) [4] developed by Freescale Semiconductor offers scalable chip-scale packaging and multidie heterogeneous integration. In addition, package-on-package stacking is supported in RCP by means of through-package vias.

For 3-D memory stacking, three main research directions have been investigated in the industry and literature: 1) 3- D DRAM main memories; 2) 3-D caches; and 3) 3-D SPMs. 3-D stacked DRAM architecture for main memory is orthogonal and complementary to this paper. Nevertheless, it has some limitations, which has prevented it from being successful in the market so far. Since DRAM dies are not process compatible with logic dies, they have to be manufactured separately and generally by different integrated device manufacturers. This will impose strict requirements and conservative standard rules on the 3-D interfaces such as large TSV size and pitch, and ultrasafe ESD protection circuits and die testing facilities. These requirements result in increased size and cost of the dies, as well as very significant supply chain set up challenges [8], [9]. On the other hand, 3-D stacking of SRAMs provides more flexibility, opportunities for process optimization, and simplified supply chain, since dies are homogeneous from a technology viewpoint (i.e., they can be manufactured in the same fab as LDs). Needless to say, given their low density and high cost, SRAM-based memories are obviously not a viable DRAM replacement for main memory, and they should be used in lower levels of the memory hierarchy. We should add here that embedding DRAM (eDRAM) memories in the lower levels of memory hierarchy is also another design alternative. Trigate CMOS eDRAM designed in 22-nm technology by Intel, and the 45-nm SOI eDRAM by IBM are two examples, which can offer better area utilization, performance, and even power consumption compared with the SRAM cells in the same technology nodes. However, these technologies require special process options and they are expensive compared with the state of the art memories. In our design, we use industrialized SRAMs, nevertheless, our proposed architecture for L2 memory can be easily adapted to use eDRAM, as well.

3-D-NUMA MEMORY IP

The 3-D-NUMA is a 3-D L2 memory stack designed to be attached to cluster based multicore platforms with a global NoC connecting all the clusters, and each cluster composed by multiple tightly coupled processors. This memory IP is well suited for serving L1 cache refill/write-back commands, since it has been designed to serve load and store packets of different sizes (up to 64 bytes). The word-levelinterleaved (WLI) organization utilized in this memory system allows for breaking a Load64Bytes command into Load8Bytes commands and dispatching them to eight parallel memory cones. This way, 3-D-NUMA can offer much higher bandwidth than simple bank-level-interleaved (BLI) memories directly attached to the NoC interfaces (see the performance analysis section for detailed results).

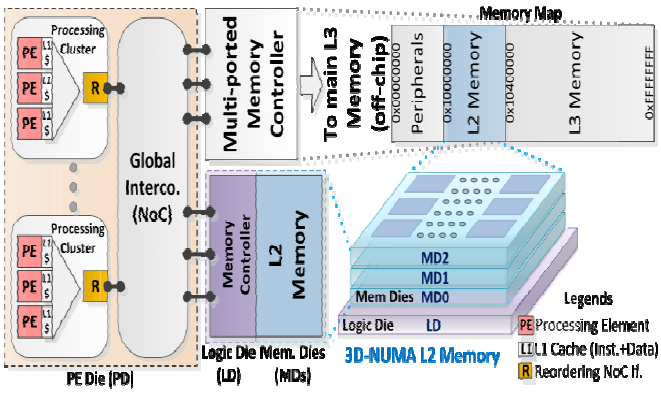

Figure 1. Overview of target memory architecture

When request packets of different sizes arrive at the NIs (Load/Store{1, 2, 4, …, 64}Bytes), the request engines (REs) break the input packet into smaller units called chunks and issues them in parallel to the arbitration trees (ATs), where a pseudoround-robin arbitration is performed among the requests arriving from different NIs. Winners enter the memory pipeline, whereas the losers wait for another cycle behind the ATs. Each request travels through the memory pipeline, and in each MD, a partial address check is performed in Fork to identify whether the request belongs to that particular MD. If matched, memory access is performed and a response is returned in the response path through the Join modules. Response paths are shared among the read buffers (RBs), and simple return-address decoders issue valid signals (resp. valid component) to the destination. Since the response chunks arriving from different memory cones may arrive out of order (OOO) and at different times, a data structure called RB is utilized to merge them, build response packets to original requests, and serialize them through the NI. It should be noted that the access time of the MDs increases with their indices (NUMA behavior), since all MDs are separated by pipeline registers and packets flow through these registers in each cycle. This feature allows for scalability, facilitates stacking of new MDs with a single mask set, and modularly increases the memory size without affecting the clock frequency (effect of the number of stacked dies on memory access time (MAT) is studied in Section IV). One should note that, such change in a flat die would require a complete silicon respin.

N is the number of independent NoC interfaces, which are used to attach 3-D-NUMA to a NoC (the design of which follows AXI bus standard). *C* is the number of parallel memory cones. This parameter defines the maximum possible number of words, which can be fetched in parallel during a load operation. Maximum outstanding transaction (MOT) defines the maximum allowed in-flight transactions inside the memory system. This parameter directly affects the depth and complexity of the RBs (described in Section III-A) while it has no effect on the memory pipeline and other components. *L* is the number of the stacked dies. *S* defines the size of each memory array, and finally, *W* and *A* define the widths of the data bus and address bus, respectively.

Dynamic RAM (DRAM)

SRAM requires a number of transistors per bit. The Difficult is to cost-effectively scale for larger memories. DRAM utilises MOSFET capacitance to store data bit. Transistor per bit cost is approx 1

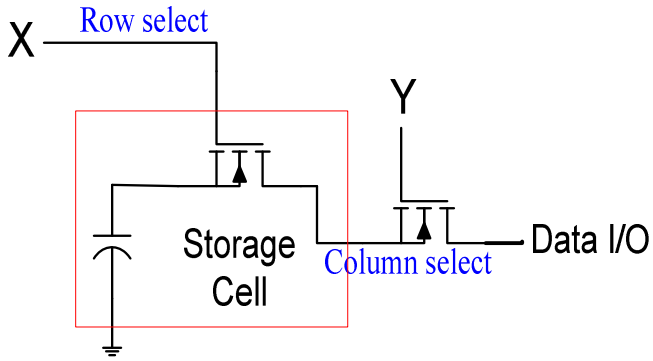

Figure 2. Storage cell

Here $Si0_2$ insulates gate and substrate . Creating dielectric capacitor between gate and substrate. Data bit is stored in this capacitance. Each bit now only requires 1 MOSFET per bit. However the charge stored in cell dissipates over time and must be recharged over time to avoid corruption.

DRAM Refresh

- Must read data bit and write value back to cell.
- JEDEC standardises DRAM row refreshes at least every 64 ms.
  - All bits in row must be refreshed.
- Dedicated hardware control DRAM refresh
  - Refresh is transparent to user

- Above 64 Kbits, DRAM more economic than SRAM logic
  - Even with refresh.

Write Operation

X	Y	Data I/O	C
0	**X**	**X**	**-**
X	**0**	**X**	**-**
1	**1**	**0**	**0**
1	**1**	**1**	**1**

Read Operation

X	Y	Data I/O	C
0	**X**	**X**	**C**
X	**0**	**X**	**C**
1	**1**	**0**	**0**
1	**1**	**1**	**1**

DRAM Organisation

DRAM is organised as “row by column” matrix. Matrix stores *n* 1-bit words. *N* is determined by the number of address lines available. Each matrix is parallelised to create word size memories. i.e : 8 parallel 4Kx1-bit DRAM matrices creates an 4K * 8-bit RAM module

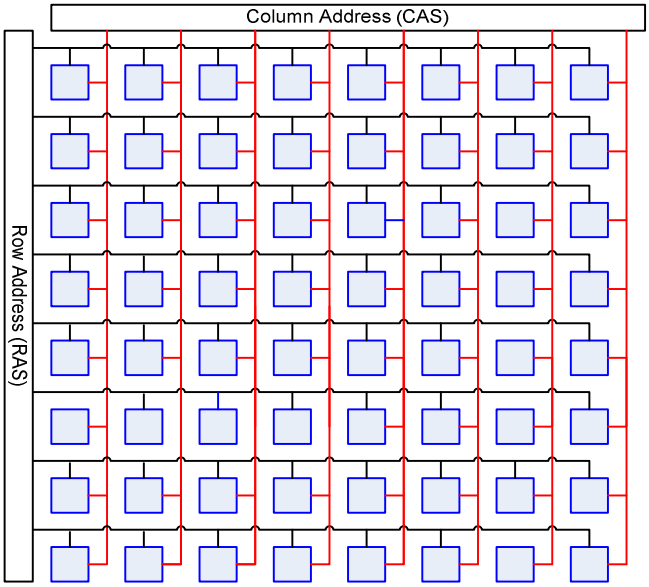

Figure 3. An 8x8 array forms a 64 x 1 dynamic RAM

The row and column select logic are comprised of address decoders. 8-rows and 8-columns need 3-address bits each. Above block is 64x1-bit DRAM. Diagram omits but matrix has 1 data I/O line. Row and Column address control which bit is active.

This block can be parallelised to create larger data word

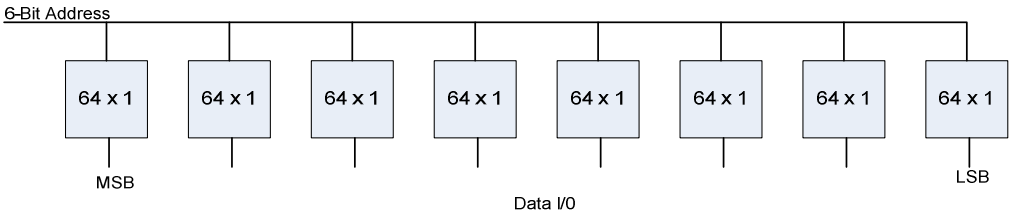

Figure 4. Memory arrangement

Each bit of data word is read/wrote in parallel

DESIGN IMPLEMENTATION

Physical design of 3-D-NUMA has been performed based on the STM bulk CMOS-28-nm low-power technology library, with a multi V_{TH} synthesis flow with LTspice design compiler graphical, and place and route in electric binary implementation.

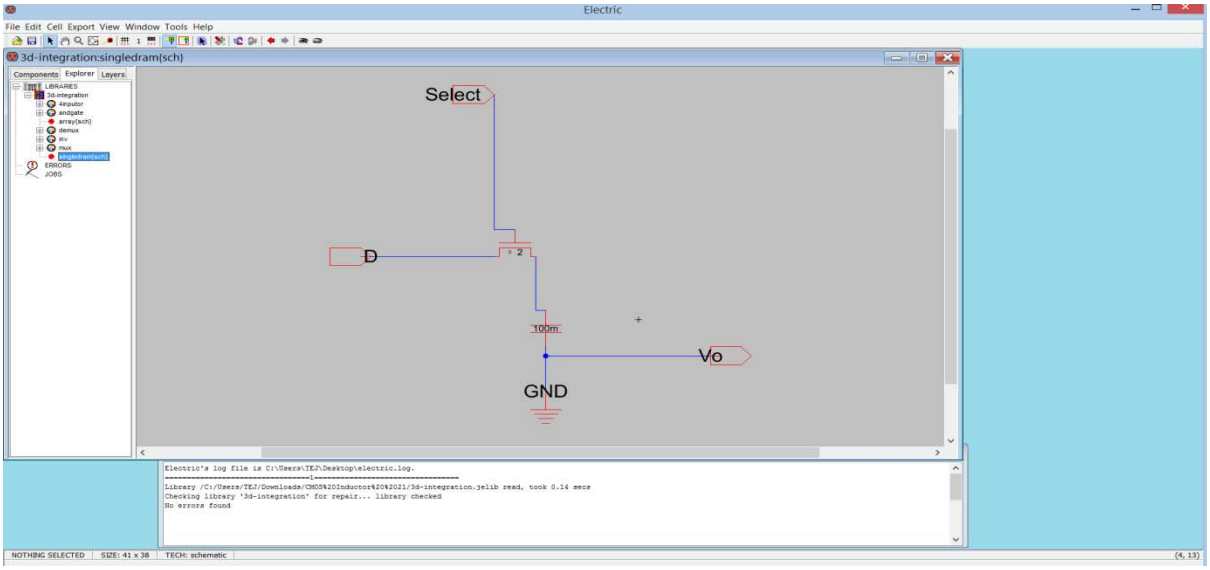

Figure 5. Single memory cell

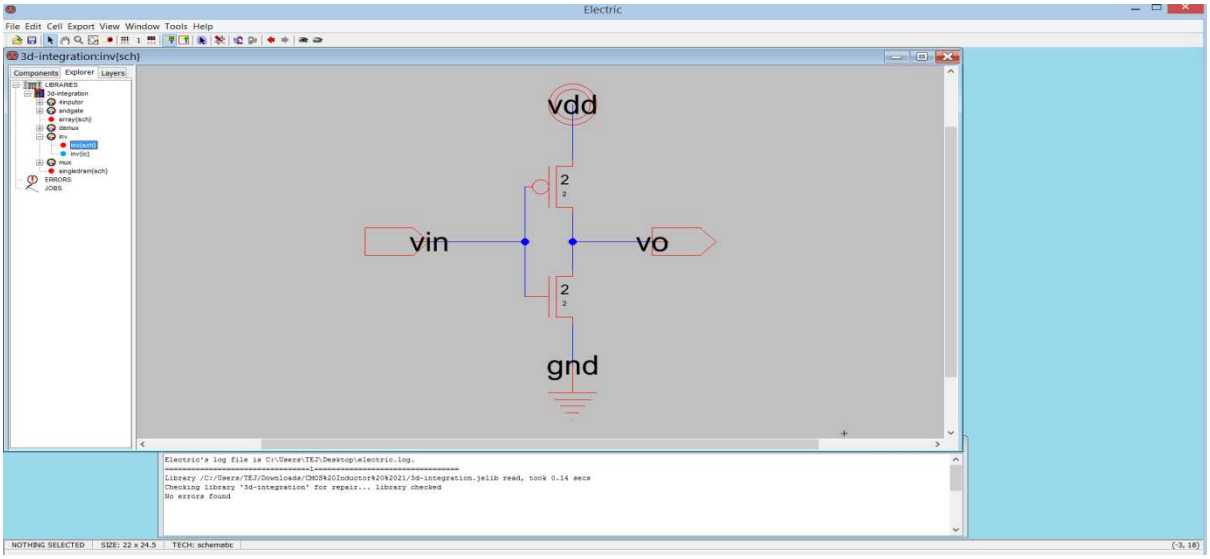

Figure 6. Inverter circuit block

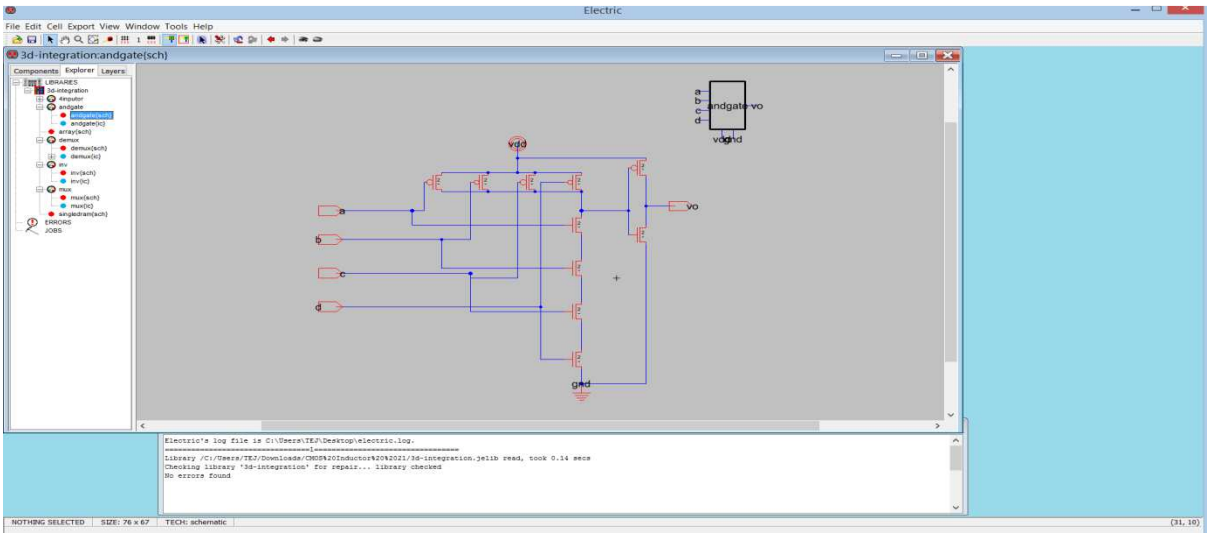

Figure 7. And gate Circuit block

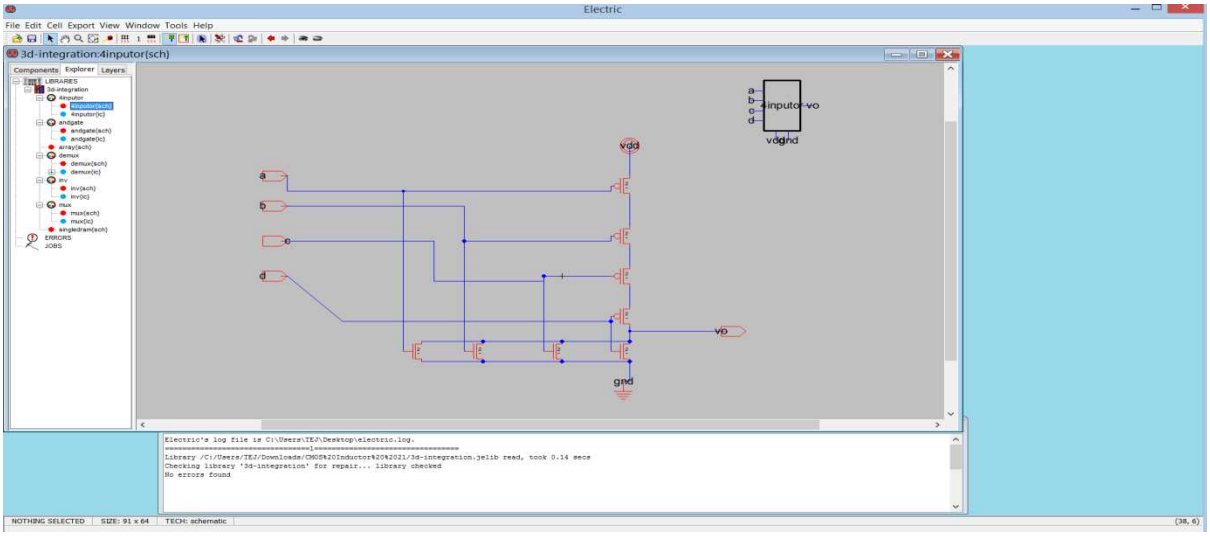

Figure 8. Four input switch

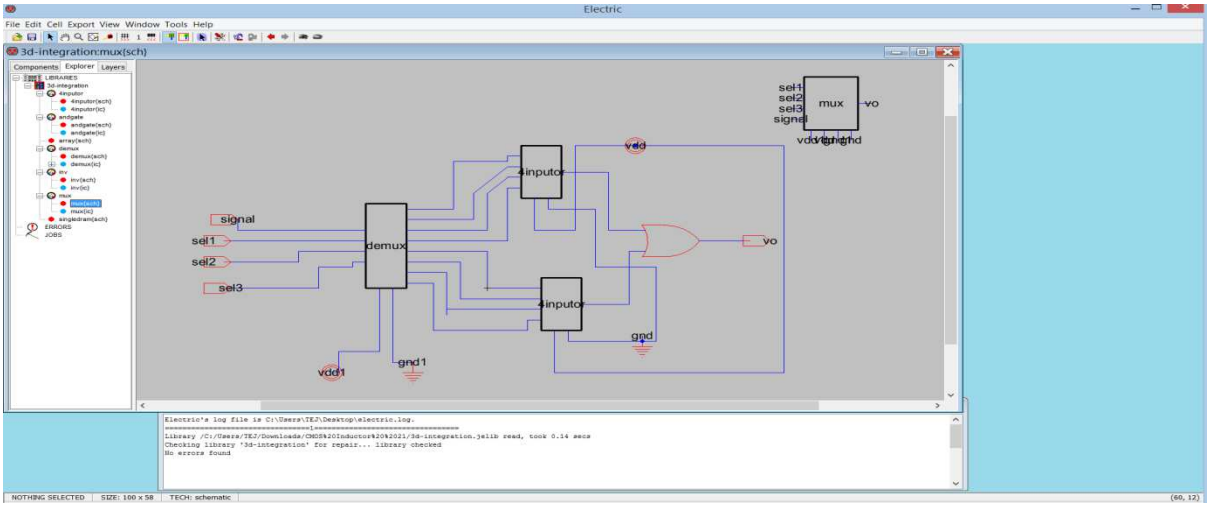

Figure 9. Multiplexer circuit

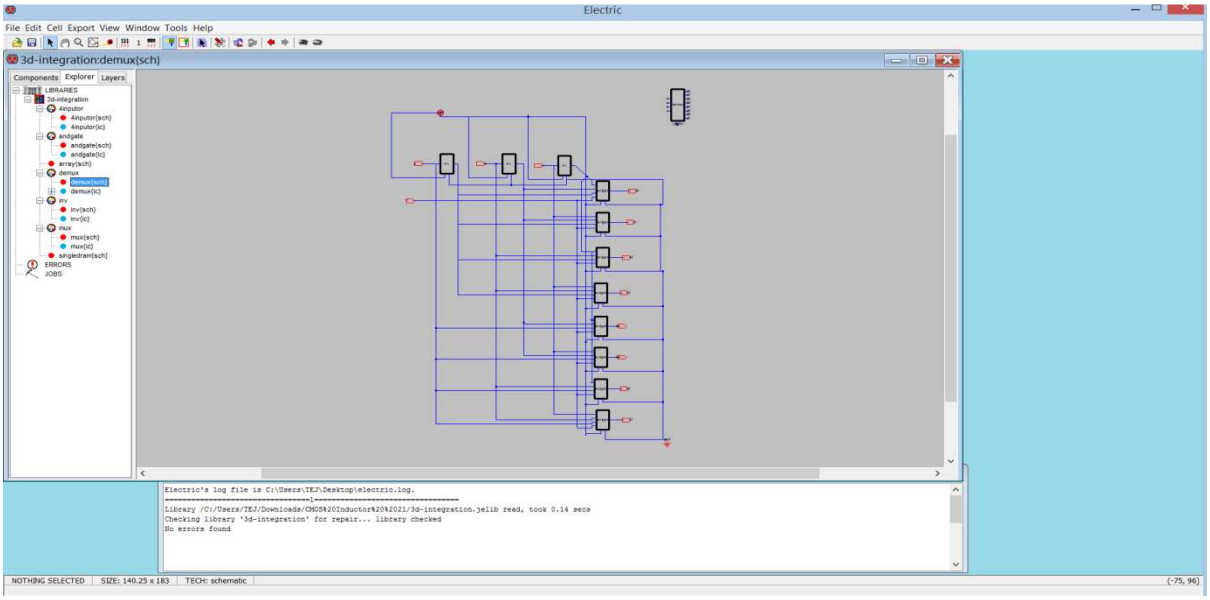

Figure 10. De Multiplexer circuit

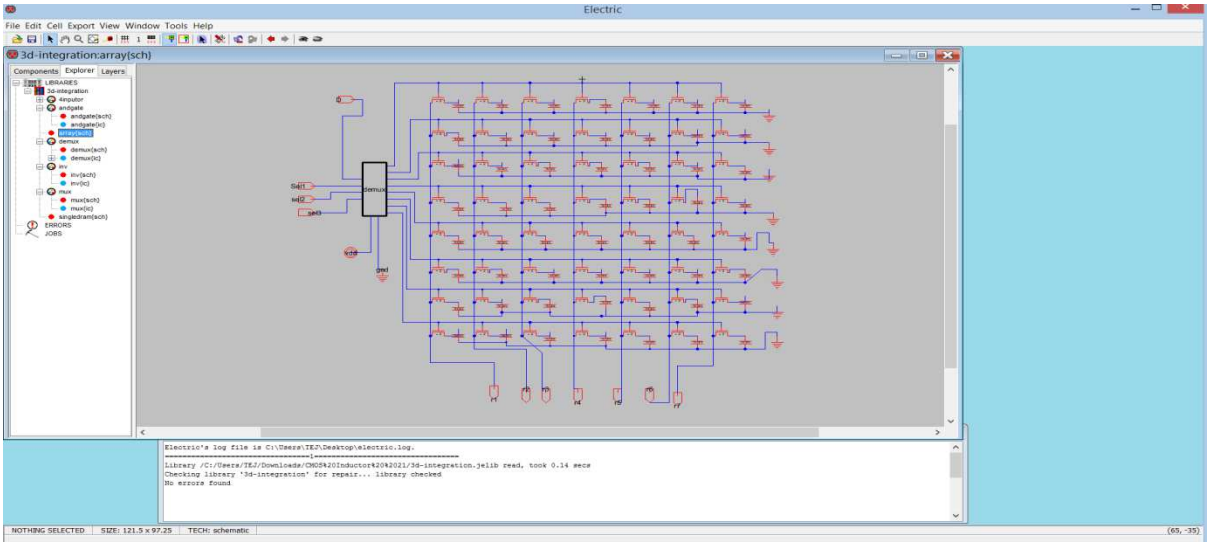

Figure 11. Memory array

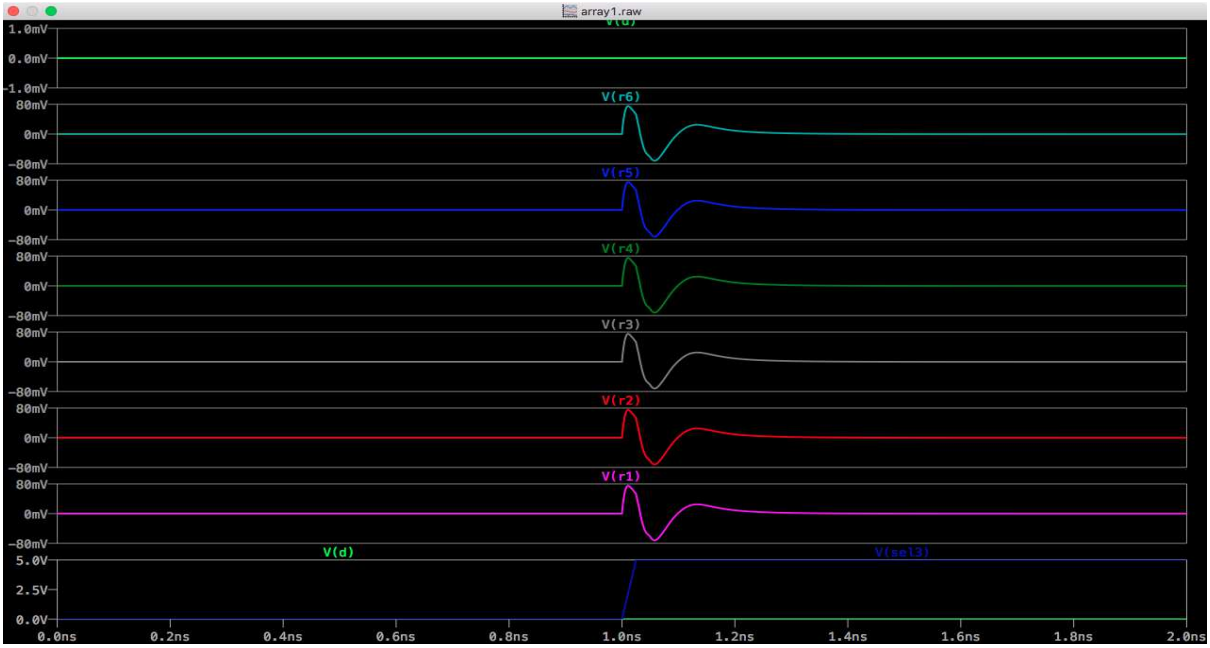

Figure 12. Voltage values for read write cycles

CONCLUSION

In this paper, we presented a synthesizable 3-D-stackable L2 memory IP component (3-D-NUMA), which could be attached to a cluster-based multicore platform through its NIs, offering high-bandwidth memory access with low average latency. Our design implements a scalable 3-D NUMA architecture, allows stacking of multiple identical MDs, supports multiple outstanding transactions, and achieves high clock frequencies due to its highly pipelined nature. We implemented 3-D-NUMA with STM CMOS-130-nm low power technology and obtained a clock frequency of 500 MHz, limited by the access time of the memory arrays while its logic components could operate up to 1 GHz (up to 4 MB in eight stacked dies with a memory density loss of 9%).

REFERENCE

[1] W. Davis et al., "Demystifying 3D ICs: The pros and cons of going vertical," IEEE Design Test Comput., vol. 22, no. 6, pp. 498–510, Nov./Dec. 2005.

[2] Yole-Développement. (2012). 3D IC & TSV Interconnects 2012 Business Update. [Online]. Available: http://www.i-micronews.com/reports/3dictsv-interconnects-2012-business%-update/8/302/

[3] E. J. Vardaman. (Mar. 2013). 3D IC with TSV: Status and Developments. [Online]. Available: http://connection.ebscohost.com/c/articles/86024505/3d-ic-tsv-status-developments

[4] Freescale-Semiconductor. (Jan. 2013). Freescale's Redistributed Chip Packaging. [Online].Available: http://www.freescale.com/files/shared/doc/reports_presentations/rcppresentation.pdf

[5] A-Star-IME. (Nov. 2010). TSV Silicon Interposer for High I/O Applications. [Online]. Available: http://www.ime.a-star.edu.sg/uploadfiles/3_proposal-tsv-interposer.pdf

[6] E. Azarkhish, I. Loi, and L. Benini, "A high-performance multiported L2 memory IP for scalable three-dimensional integration," in Proc. IEEEInt. 3D Syst. Integr. Conf. (3DIC), Oct. 2013, pp. 1–8.

[7] F. Clermidy, D. Dutoit, E. Guthmuller, I. Miro-Panades, and P. Vivet, "3D stacking for multi-core architectures: From WIDEIO to distributed caches," in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), May 2013, pp. 537–540.

[8] F. Ferro. (Aug. 2013). DRAM Remains the Status Quo. [Online]. Available: http://semiengineering.com/dram-remains-the-status-quo/

[9] Yole-Dévelopment. (Feb. 2013). A Reassessment of the Use of Wide-I/O Memory in Smartphones. [Online].Available:http://www.i-micronews.com/news/reassessment-use-wide-iomemorysmartphones.10096.html