Extended Abstracts of the 1996 International Conference on Solid State Devices and Materials, Yokohama, 1996, pp. 824-826

A New Three-Dimensional Multiport Memory for Shared Memory in High Performance Parallel Processor System

K. Hirano, S. Kawahito, T. Matsumoto, Y. Kudoh, S. Pidin,
N. Miyakawa+, H. Itani*, T. Ichikizaki*, H. Tsukamoto* and M. Koyanagi

Dept. of Machine Intelligence and System Engineering, Tohoku University
Aramaki, Aoba-ku, Sendai 980-77, Japan
+ Fuji Xerox Co.
** Mitsubishi Heavy Industry Co.*

We propose a new multiport memory with three-dimensional (3D) structure for a parallel processor system with real shared memories. This multiport memory can act as a real shared memory without the bus-bottle neck. Therefore, a high performance parallel prcessor system with shared memories can be easily constructed using this multport memories. The simulation results for basic memory operation and the broadcast operation in this 3D multiport memory are described. Furthermore, a new 3D integration technology to fabricate this memory is also proposed and the key technologies for this 3D integration are explained.

1. Introduction

The parallel processing using multi-processors is very effective to dramatically improve the computational throughput. It is well known that a shared memory is very useful to build the high performance parallel processor system with simple configuration and architecture. However, the conventional system with shared memories has the drawback that only limited number of processors can be connected through a shared memory. This is because the processors are connected to the shared memory through the common buses in the conventional system and therefore the overall system performance is eventually limited by the data transfer speed of the buses. A high performance parallel processor system with shared memories can be constructed using the multiport memories because the multiport memory can act as a real shared memory without the bus-bottle neck. However, it is not easy to design the high speed multiport memory with many read/write ports using the conventional method because one memory cell must drive many bit-lines and the memory cell area significantly increases when many ports are connected to the memory cell. To solve this problem, each port should be driven by each memory cell. This implies that several memory cells with their own ports are stacked each other. That is, the multi-cells with multiports. Such multiport memory can be achieved by using three-dimensional integration technology. In this paper, we propose a new parallel processor system with three-dimensional multiport memories as shared memories and a new three-dimensional integration technology to achieve this system.

2. Parallel Processor System with 3D Multiport Memories

A configuration of parallel processor system with directory architecture for shared memory is shown in Fig.1. Many processors are connected to one shared memory in a cluster of this system. The computational throughput can be significantly improved within a cluster because a real multiport memory is used as a shared memory in this system and hence the bus-bottle neck is not significant any more. Several clusters are connected

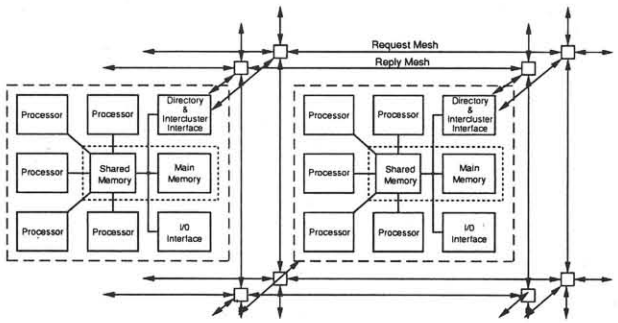


Fig.1 Parallel processor system with directory architecture for shared memory.

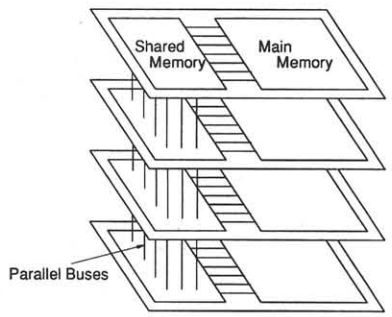


Fig.2 3D shared memory system.

by the high speed buses when it is required to connect more processors. In this case, shared data are stored in several shared memories beyond one cluster. These shared data are supervised by the shared memory directories. Therefore, the performance is slightly reduced when the data transfer among the processors and the shared memories is required beyond the clusters. Nevertheless, the overall performance of this system is significantly improved by using real multiport memories as shared memories. A configuration of three-dimensional shared memory system is shown in Fig.2 where several layers of shared memory are connected by many short buses for the broadcast in the vertical direction. A main memory is connected to the respective layer of shared memory by the internal multi-buses with very high data transfer speed and data band width. A processor in a cluster is connected to the respective

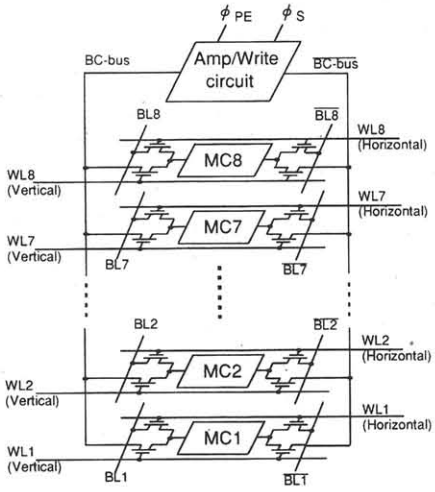


Fig.3 3D multi-port memory circuit for shared memory.

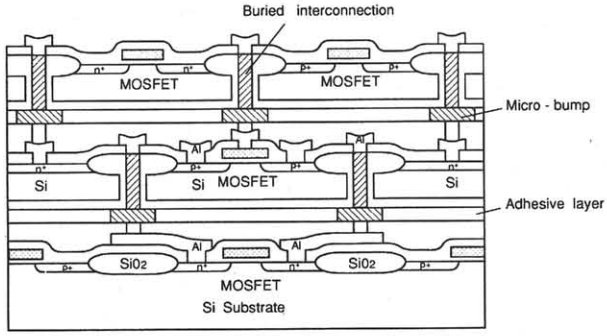


Fig.4 Cross-sectional view of 3D multi-port memory.

layer of shared memory. Data written to some layer of shared memory by the respective processor are simultaneously transferred (broadcast) to other layers of the shared memory through the vertical multi-buses (broadcast buses). Therefore, the memory cells with an identical memory address in all memory layers of shared memory have an identical data after the data transfer. This identical data can be read-out simultaneously and independently by several processors which are connected to the respective layers of shared memory. Therefore, this shared memory acts as a real shared memory without the bus-bottle neck. The circuit configuration to achieve such real shared memory is shown in Fig.3. Three-dimensional (3D) multiport memory is used to achieve the real shared memory in the figure where a part of 8 memory layers is shown. A memory cell of this multiport memory has two pairs of access transistors (two ports) for the horizontal access and the vertical access. The horizontal access transistor is used to execute the conventional memory operation within a memory layer while the vertical access transistor is used to broadcast the data in the vertical direction. The amplifier and write circuits are installed in the top layer of this multiport memory in order to amplify the data read-out to the broadcast buses and rewrite them to the memory cells in all memory layers of multiport memory. A cross-sectional view of 3D multiport memory is shown in Fig.4 where a new 3D integration technology is used to stack the memory layers. Figure 5 shows the simulated waveforms of the multiport memory with 8 memory layers (8 ports) which was designed based on 2μm CMOS design rule. In the figure, the data "1" and "0" are written to

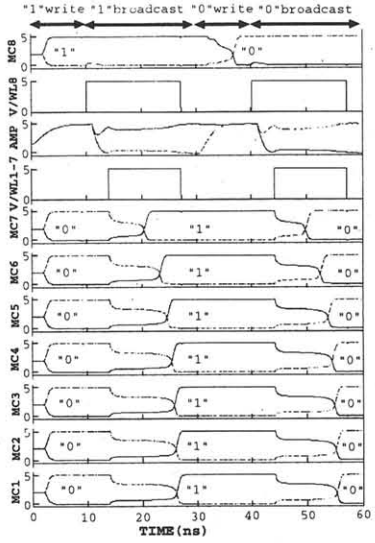


Fig.5 Simulated waveforms for 3D multi-port memory.

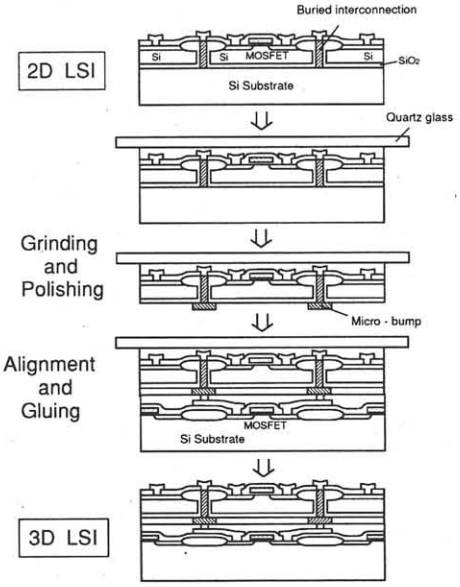


Fig.6 Fabrication sequence of 3D LSI.

the memory cell in the eighth layer of multiport memory and then transferred to the memory cells in the other memory layers in the former part and the latter part of the operation, respectively. It was confirmed from Fig.5 that the basic memory operation and the broadcast operation are successfully executed. It takes about 13ns for broadcasting the data to all layers of 3D multiport memory.

3. 3D Integration Technology for Multiport Memory

A new 3D integration technology as shown in Fig.6 has been proposed. Thinned memory layers are stacked on the bottom memory layer with the thickness of about 200μm using a wafer bonding technique. The upper memory wafer with many buried interconnections is glued to the quartz substrate using a liquid wax and then thinned down to 20μm using the grinding and the chemical-mechanical polishing (CMP) techniques. The buried interconnection consists of a highly doped poly-Si

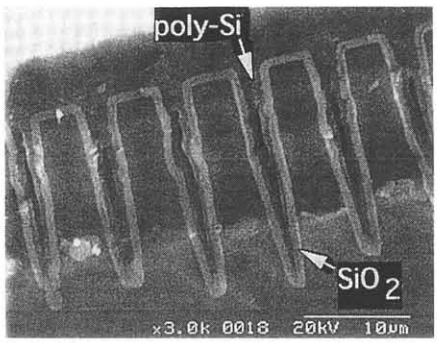


Fig.7 SEM cross-section of oxidized deep trench filled with poly-Si.

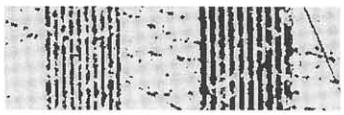


Fig.8 Photomicrograph of the back surface of silicon wafer after grinding.

which is deposited onto the oxidized deep silicon trench. The thinned memory wafer is glued to the lower memory layer on the surface of which an UV-hardening adhesive material with the thickness of 1μm is spin-coated. Three-dimensional memory structure for the multiport memory is formed by repeating these sequences. Figure 7 shows SEM cross-section of oxidized silicon trench with the depth of about 20μm which is filled with a highly doped poly-Si. This highly doped poly-Si is used as the buried interconnection. As is obvious in the figure, the buried interconnection with the size of about 2μm and the length of about 20μm is clearly formed. A photomicrograph of the back surface of silicon wafer after thinning the wafer down to 20μm by grinding and CMP is shown in Fig.8 where the bottom of silicon trench is clearly seen. The electrical contact between the upper and lower buried interconnections is implemented using In/Au micro-bumps as shown in Fig.9. This micro-bump is formed using a lift-off method. A newly developed 3D wafer aligner is used to glue the upper memory wafer to the lower memory wafer. Infra-red light is used for the alignment of two wafers in this 3D wafer aligner. In addition, the wafer stage is precisely controlled in the movement using the piezo actuators. The controllability of the wafer stage is 50nm in the x, y and z directions. Furthermore, the gap between two wafers and the contact force to glue two wafers can be monitored and controlled in-situ before and after two wafers are contacted, respectively. Infra-red images

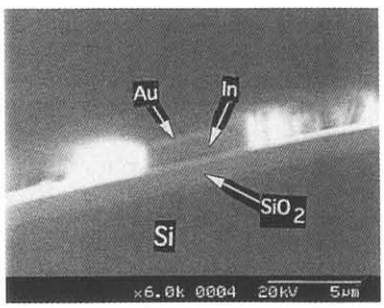
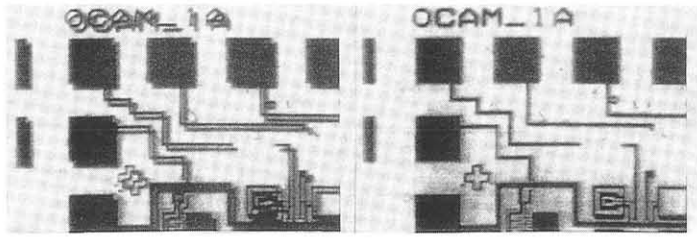


Fig.9 SEM cross-section of micro-bump.



(a)Before alignment (b)After alignment

Fig.10 Wafer alignment using 3D wafer aligner.

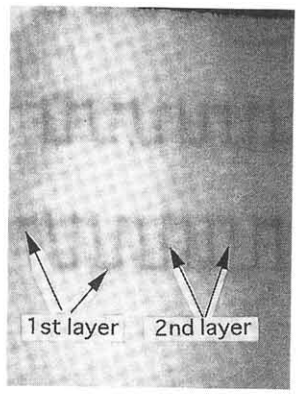


Fig.11 Photomicrograph of test structure to evaluate the contact resistance between micro-bumps.

of test wafers before and after the alignment using 3D wafer aligner are shown in Fig.10. The alignment accuracy of two wafers is around 1μm. Figure 11 shows the infra-red image of the test structure to evaluate the contact resistance between two micro-bumps in the upper and lower layers. This image was taken after bonding two wafers. The UV-hardening adhesive material is coated on the surface of the lower wafer and hence on the surface of the micro-bump in the lower wafer. This adhesive material on the surface of the micro-bump is pushed aside during applying the force to contact two wafers. It was confirmed using the test structure as shown in Fig.11 that a good electrical contact between two bumps in the upper and lower wafers can be obtained. Thus, the key technologies to fabricate 3D multiport memory have been developed.

4. Conclusion

We proposed a new multiport memory with three-dimensional structure for a parallel processor system with real shared memories. The basic memory operation and the broadcast operation in this 3D multiport memory is confirmed by the computer simulation. Furthermore, a new 3D integration technology to fabricate this memory was also proposed and the key technologies for this 3D integration were developed.

Reference

[1] T. Matsumoto et al., Ext. Abst. of SSDM, (1995)1073