Memory Technology for Extended Large-Scale Integration in Future Electronics Applications

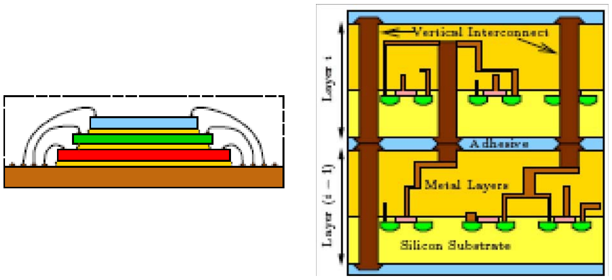
Dinesh Pamunuwa
Centre for Microsystems Engineering
Department of Engineering
Lancaster University, Lancaster, UK
d.pamunuwa@lancaster.ac.uk

Abstract

Extending 2-D planar topologies in integrated circuits (ICs) to a 3-D implementation has the obvious benefits of reducing the overall footprint and average interconnection length, with associated improvements in cost, and delay and energy consumption, while also providing an opportunity to integrate disparate technologies. Such advances are very much technology driven, and early research into 3-D integration has now crystallised into commercially viable options that are being pursued by many companies. Being able to position memory in closer proximity to processing elements in a NoC architecture as afforded by a 3-D physical architecture has the potential to improve the memory bandwidth and mitigate the general nature of delay constrained performance in IC design. Understanding the nature of the opportunities and constraints provided in such a 3-D physical architecture is crucial in realising the true benefits of 3-D integration in future applications.

1. Introduction

Device integration on a single die has kept broadly to the Moore predicted rate over the past four decades due to concerted industry efforts, and this trend is set to continue in the near future, with the DRAM 1/2 pitch being set a target of 18-20nm in 2018 by the ITRS [1]. For true system-level integration of heterogeneous elements including specialised and general purpose digital processing blocks, analogue, mixed-signal and RF functions and storage, the IC industry has increasingly been looking at 3-D integration. Multi-chip module type arrangements where packaged chips are situated on a single substrate or across a board and package-level integration options such as System-In-Package where stacked dies are connected by bond wires have been augmented by technologies capable of die-to-die and die-to-wafer integration using through-Si-vias (TSV) (see Fig. 1). These different options generally provide a trade-off between cost, performance, functionality and footprint [2].



1a: System-In-Package (Die-stacking using wire bonds)

1b: Wafer-level integration with die stacking

Figure 1. 3-D Integration Options (from [2])

The generalised NoC architecture has been proposed as being suitable for integrating heterogeneous elements and managing the communication between them efficiently. Sophisticated system-level analyses identify a memory-related bottleneck for high-throughput, demanding applications such as multi-media applications. A 3-D stacked memory system provides an interesting storage option for a general purpose NoC, potentially easing the memory bottleneck.

2. Memory Integration Technology

The main potential benefits of 3-D integration include reduced overall footprint, and reduction of overall wiring length for a given system configuration, with the associated improvements in propagation delay and energy per

978-3-9810801-3-1/DATE08 © 2008 EDAA

interconnect switching transfer, as much as 54% and 51% respectively by one estimate [4]. It also allows disparate technologies to be integrated, which for a general purpose NoC can mean the chip-level integration of DRAM and FLASH with CMOS logic. The reduced parasitics for interconnects can further significantly simplify the circuit and power distribution network design for high performance applications.

In particular, massive amounts of inexpensive storage currently supplied by Magnetic Hard Disks can potentially be supplied by flash memory located within the same chip, eliminating the multiple communication hierarchies from the chip to off-chip caches to board-level traces to cables and back. This can potentially improve the raw data bandwidth by several orders of magnitude. Using this improvement to gain a true improvement in system-level functionality discernible by the end user will require a shift from the traditional architecture with multiple hierarchical caches.

3. Challenges in 3-D Integration

Reliability of such a system is a concern, since flash cells have an activity depended life-span, being subject to a higher voltage stress during the programming phase, and also during reading, for multi-level cells [5]. Wear-levelling algorithms which spread the load evenly across cells, error-correction and redundancy to account for manufacturing variances are common in Flash memories, but may need to be augmented with innovative techniques.

The main obstacle to 3-D integration is poor thermal conductivity and heat dissipation and the resultant temperature rise due to the high power density [6]. Thermal vias can alleviate this problem, but again innovative techniques are called for, such as micro-channel cooling [7] and dynamic activity management with built-in sensors and sensing functions to monitor the operating temperature.

A major consideration in 3-D integration is the fabrication cost. Yield is very sensitive to chip stacking, and can decrease exponentially with no. of layers in wafer-to-wafer integration. This is because even if the wafers individually have acceptable yield, when combined in a vertical stack, the probability of a good die on one layer combining with a bad one on another layer is high. Mechanical stress during the assembly process, and combined mechanical and thermal stress can cause failures as well as parametric shifts. Testing is significantly more complicated, and the turn-around time is higher. However, these issues are currently being heavily researched, and integrated test functions and an inexpensive test-insert to sit in-between dies in the stack may be one solution [8].

4. Improvements in Performance

Three principal operations of a digital system can be identified: the binary switching transfer, communication of a bit and storage of a bit. Each of these operations can be characterised by a metric, such as gate delay for the binary switching transfer or interconnect latency for communication. High-level performance metrics such as MIPS, bandwidth and throughput can be calculated from these metrics. By considering limitations imposed on each of these operations from physical considerations at different hierarchical levels, ranging from the fundamental to the system level, an exploration space for performance can be defined. The key metrics relating to each of these operations for the 3-D stacked memory system will identify its location within the performance space, and highlight the possibility of potential improvements.

Such an analysis can be used to identify future opportunities for NoC based systems in demanding applications.

5. References

[1] SEMATECH. (2006). International technology semiconductor roadmap. [Online]. Available: http://www.itrs.net/Links/2006Update/2006UpdateFinal.htm

[2] R. Weerasekera, L-R. Zheng, D. Pamunuwa, and H. Tenhunen, “Extending Systems-on-Chip to the Third Dimension: Performance, Cost and Technological Trade-offs,” in *Proc. IEEE International Conference on Computer-Aided Design*, pp 212-219, San Jose, CA, Nov 2007.

[3] F. Catthoor, N. D. Dutt, and C. E. Kozyrakis, “How to solve the current memory access and data transfer bottlenecks: at the processor architecture or at the compiler level,” *in Proc. Conference on Design, Automation and Test in Europe*, pp. 426–435, 2000.

[4] M. Bamal, S. List, M. Stucchi, A. Verhulst, M. Van Hove, R. Cartuyvels, G. Beyer, and K. Maex, “Performance comparison of interconnect technology and architecture options for deep submicron technology nodes,” in *Proc. International Interconnect Technology Conference*, pp. 202–204, 2006.

[5] Ken Takeuchi et. al., “A 56-nm CMOS 99-mm^2 8-Gb Multi-Level NAND Flash Memory With 10-MB/s Program Throughput,” *IEEE J Solid-State Circuits*, vol. 42, no. 1, Jan. 2007, pp. 219-232.

[6] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, “3-d ics: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration,” *Proceedings of the IEEE*, vol. 89, no. 5, pp. 602–633, May 2001.

[7] Koo, S. Im, L. Jiang, and K. Goodson, “Integrated microchannel cooling for three-dimensional circuit architectures,” *ASME Journal of Heat Transfer*, vol. 127, 2005, pp. 49–58.

[8] Andrew Richardson et.al., “System in package technology – design for manufacture challenges,” *Circuit World*, vol. 33 No. 1, 2007, pp. 36-46.