Three-Dimensional Integration: Analysis, Modeling and Technology Development

Personnel
A. Rahman and A. Fan (R. Reif)

Sponsorship
MARCO Focused Research Center on Interconnect

As the critical dimensions in VLSI circuits continue to diminish, system performance of Integrated Circuits (IC) will be increasingly dominated by interconnect’s performance. For the technology generations approaching 100-nm, innovative circuit designs and new interconnect materials and architecture will be required to meet the projected system performance. New interconnect material solutions such as copper and low-k dielectric offer only a limited improvement in system performance. Significant and scalable solutions to interconnect delay problem will require fundamental changes in system architecture, design, and fabrication technologies.

Three-dimensional (3-D) IC, devices are allowed to exist on more than one device layer, and they can be contacted from both top and bottom device layers. Flexibility to place devices along the third dimension allows higher device density and smaller chip area in 3-D IC. The critical interconnect paths that limit system performance can also be shortened by 3-D integration to achieve faster clock speed. By 3-D integration, active layers fabricated with different front-end processes can be stacked to form systems on a chip. A cross section of a proposed 3-D integrated circuit/system is shown in Figure 22.

System-Level Performance Modeling and Trade-off Analysis

A recently derived 2-D stochastic wire-length distribution [1] that has been used to predict interconnect delay and system performance metrics has been extended to 3-D interconnects to determine the trade off of 3-D integration. Using 3-D stochastic wire-length distribution and interconnect delay constraints, figures of merit such as critical path delay, chip area, complexity/cost etc. have been estimated. The wire-length distribution is derived using an empirical relation known as Rent’s rule which relates the number of input and output terminals of an integrated circuit to the number of logic gates within that circuit.
Based on our simulation, 3-D integration results in narrower wire-length distribution, with more local (short) wires and less global (long) wires, than the conventional planar implementation. The average and total wire-lengths in 3-D integration are also shorter than 2-D integration. Wire-length distribution of 3-D IC with 21 million transistors/3.5 million logic gates, consistent with 0.18 µm technology generation, is shown in Figure 23. In estimating the wire-length distribution, it is assumed that i) Rent’s rule can be applied iteratively throughout the system, ii) it is equally likely to form vertical interconnects between device layers as horizontal interconnects within device layers, and iii) the number of interconnects is conserved in 2-D and 3-D implementation.

Using the stochastic wire-length distribution and interconnect delay criteria, various trade-off analysis can be performed between 2-D and and 3-D implementation of integrated circuits. For example, i) the clock frequency/cycle time can be estimated for fixed total chip area and cost/complexity function and ii) the chip area can be estimated for fixed clock frequency and cost/complexity function. Similarly, the impact of integrating additional metal layers on system performance and chip area can also be evaluated for both 2-D and 3-D IC.

Based on our analysis of scaled technologies, we find that the contribution of interconnect delay on local clock frequency in high-performance circuits such as microprocessors is going to be in the range of 30%-50%. Using 3-D IC with two device layers, ~15%-25% improvement in local clock frequency can be achieved. However, the contribution of interconnect delay on global/across-chip clock frequency can be in the range of 80%-90% and a much higher improvement in across-chip clock frequency can be achieved by 3-D integration.

Fig. 22: Cross sectional view of a proposed three-dimensional integrated circuit formed by low-temperature wafer bonding.
In most of the high-performance logic circuits, the chip area is interconnect limited. Using 3-D technologies, both the reductions in chip area and wiring pitch can be achieved for fixed system performance. Since the interconnect limited chip area is proportional to \( \text{total wire-length} \times \text{wiring-pitch} \), significant reduction in total chip area can be achieved by 3-D integration. For example, for fixed clock frequency, in 3-D IC with two device layers, the reduction in total chip area is in the range of 20%-30%. Based on our analysis work, we find that due to heat removal issues, blockage of wiring tracks due to inter-device layer vias, cost/complexity, etc. there is an optimum number of device layers that can be profitably integrated. It appears to be 3-4 device layers.

Fabrication

In direct 3-D integration, active device wafers are bonded together, while all active layers are electrically interconnected using high aspect ratio vias. Prior to wafer bonding, the device wafers are assumed to contain multiple aluminum metal layers and inter-level dielectrics (ILD), thus requiring low-temperature bonding below 450°C to avoid Al degradation. To implement a structure with 3-D stacked device layers, a metal-to-metal bonding process has been proposed. From Figure 24, metal (Cu) bumps on both wafers will serve as electrical contacts between via on the top wafer and Al interconnects on the bottom wafer. At the same time, these metal bumps also function as the wafer bonding medium. In addition, dummy metal patterns can be made to increase the bonding surface area. The dummy metal films can also be auxiliary structures such as ground planes or heat conduits for different Si active layers. A scanning electron micrograph of bonded wafers is shown in Figure 24.

Successful wafer bonding was achieved using Cu-Ta (300 / 50 nm) layers on Si at 400°C for 30 min (in the EV bonder) and annealed at 400 °C for 30 min. The bonding interfaces include blanket-to-blanket or blanket-to-patterned metal films. In the latter case, Cu lines with line-widths ranging from 60-200 µm were wet-etched using \( \text{H}_2\text{O}_2:\text{HCl:H}_2\text{O} \) (1:1:125 by volume). Furthermore, Cu test structures of 1-3 µm were patterned using conventional image-reversal lift-off techniques. The Cu film does not require special pre-bonding surface preparation, such as metal CMP or ultraviolet light exposures. The bonded pairs at 400°C exhibited good bonding strength when the razor blade test was applied.

Fig. 23: Wire-length distribution of 2-D and 3-D IC of random logic networks. \( N_z \) is the number of device layers, \( N \) is the total number of logic gates, f.o. is the average fan-in/out, and \( k \) and \( p \) are Rent’s parameters. The gate pitch is a normalized unit, defined as the average separation between logic gates.
Referring to Figure 22, the proposed 3-D process involves two wafer bonding steps. The first step is to bond the top SOI device wafer to a dummy substrate prior to SOI substrate etchback. Subsequently, the thinned top device wafer is bonded to the bottom device wafer. There are three technology developments in this process: The bonding of the handle wafer, wafer thinning, and handle wafer release. To begin with, the handle wafer was attached to the SOI using a Al/polyimide/Al stack, with polyimide as the bonding adhesive at 400 °C. Next, a combination of mechanical grinding (400 µm of Si) and plasma etching (SF6/C4F8) was used for the SOI etchback, with the buried oxide layer as the etch-stop. Finally, after bonding the top / bottom device wafers with Cu, the handle wafer can be released by immersion in 50 °C HCl. The HCl attacks the aluminum in the Al/PI/Al tri-layer, thereby separating the two wafers.

Fig. 24: Cross-sectional SEM of bonded Cu/Ta - Cu/Ta wafers. Cu/Ta = 300 / 50 nm, bonded at 400°C.