Three-Dimensional Integration: Analysis, Modeling, and Technology Development

Personnel

A. Fan, S. Das, K.-N. Chen, R. Tadepalli, N. Checka, and C.-S. Tan (R. Reif and A. Chandrakasan)

Sponsorship

MARCO Focused Research Center on Interconnect (MARCO/DARPA)

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 50 nm, innovative circuit designs and new interconnect materials and architectures 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 the interconnect delay problem will require fundamental changes in system architecture, design, and fabrication technologies.

In three-dimensional (3-D) ICs, 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 a 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 1.

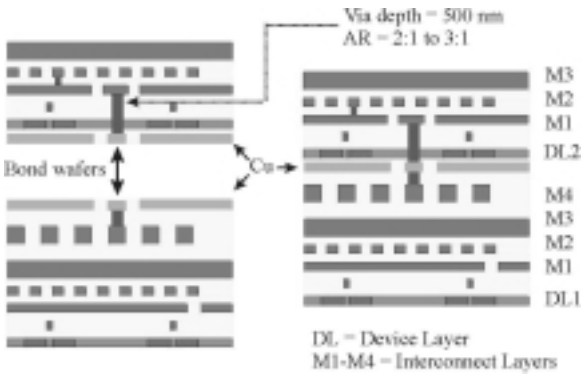

Fig 1: Cross sectional view of a proposed three-dimensional integrated circuit formed by low-temperature wafer bonding.

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 2. 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.

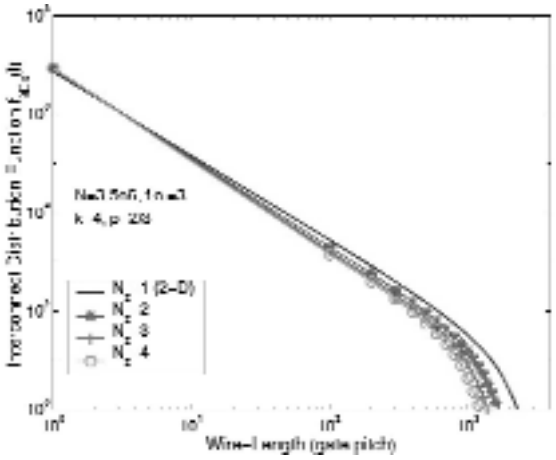

Fig. 2: Wire-length distribution of 2-D and 3-D IC of random logic networks. Nz is the number of device layers, N is the total number of logic gates, f.o. is the average fan-in/out, and k and p are Rent's parameters. The gate pitch is a normalized unit, defined as the average separation between logic gates.

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.

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 total wire-length x 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.

Recent work has involved the exploration of circuit architectures suitable for 3-D integration. For example, our work has shown that Field-Programmable Gate Arrays (FPGAs) may benefit up to 50% in terms of interconnect delay and power dissipation and as much as 40% in logic density. Also, architectures that integrate dissimilar technologies may be enabled or improved by 3-D integration. Toward this end, we have begun research into System-on-a-Chip (SoC) implementation issues that arise within the 3-D integration framework. Another aspect of our research concerns the digital design flow for 3-D integrated circuits. With a view towards full system synthesis using 3-D technology, we are developing a set of design tools for 3-D integration. This effort comprises placement, routing, layout, and verification tools for 3-D digital integrated circuits. With the use of such tools, we will be able to verify and expand upon the predictions made using our stochastic interconnect models.

Fabrication

Referring to Figure 1, the implementation of 3-D ICs involves vertical stacking of CMOS device layers using Cu-Cu wafer bonding at 400oC. All active layers are electrically interconnected using 2:1 or 3:1 aspect ratio vias. Metal (Cu) bumps on both wafers will serve as electrical contacts between the top wafer and Al interconnects on the bottom wafer. In addition, these metal bumps also function as the wafer bonding medium. Auxiliary Cu pads exclusive from inter-layer communication activities could possibly be used as ground planes or heat conduits for different Si active layers. Prior to bonding, the device wafers are assumed to contain multiple aluminum metal layers and Inter-Level Dielectrics (ILD), thus requiring low-temperature bonding below 450oC to avoid Al degradation.

Successful wafer bonding was achieved using Cu/Ta (300 / 50 nm) layers on Si at 400oC for 30 min and annealed at 400 oC in N2 for 30 min. The Cu film does not require special pre-bonding surface preparation, such as metal CMP or ultraviolet light exposures. The bonded pairs at 400 oC exhibited good bonding strength when the razor blade test was applied.

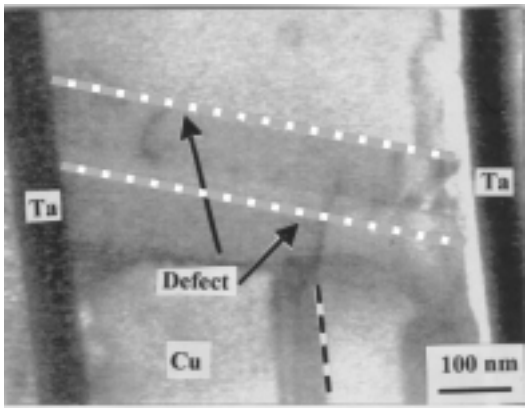

Fig. 3: TEM of bonded Cu/Ta - Cu/Ta wafers, exhibiting twins (dashed lines) perpendicular to the Cu-Cu bonding interface. Cu/Ta = 300 / 50 nm, bonded at 400 oC.

Furthermore, microstructures of the Cu-Cu interface can be examined using XTEM, as shown in Figure 3. The twin grains that span between the two Ta layers (from different wafers) suggests the existence of localized homogenous Cu-Cu bond; in other words, at the twin sites, the Cu-Cu interface is non-distinct. Also visible from Figure 4 are grains travelling parallel to the bonding interface. Presumably, these grains originated from defects at the bonding interface prior to thermocompression.