Stacking Technology: Achieving Next Generation Memory Densities Today

As a result of miniaturization and performance trends, designers are constantly looking to achieve the highest possible electrical functionality and performance in the least possible amount of space. The two key limiting factors are typically the level of integration and the I/O pad limitation.

Silicon space and connectivity limitations can be addressed on two levels. One involves higher integration through process shrinks at the die level. The second offers higher integration via the stacking of multiple dies, stacked packages or stacked boards.

What is Stacking Technology?

Stacking technology is the mechanical and electrical assembly of dies, packaged components or cards for the purpose of increasing depth, width and/or functionality of electronic designs within a limited surface area. It is an optimal solution for addressing space and connectivity limitations.

There are two types of stacking methods. Custom stacking focuses on customized die or package solutions and is fairly expensive to design and manufacture. A custom stacked DRAM implementation cost can range from $500 to $1000 per GByte of memory storage. A standard or commercial approach has a completely different cost model, with average premiums of 15-20% for stacked memory modules versus regular memory modules using monolithic packaged parts.

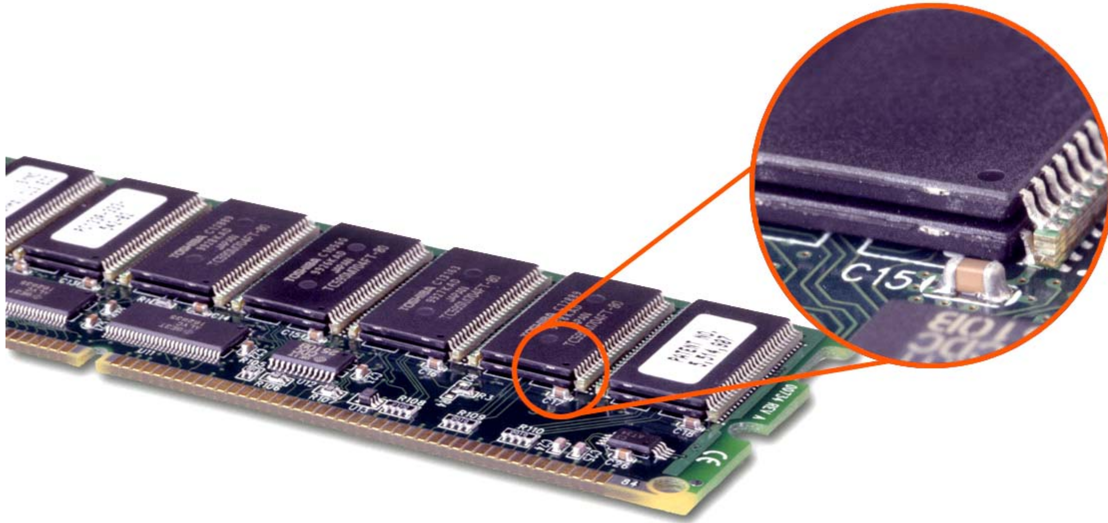
In standard stacking, several methods offer approaches to stacked dies, ICs and boards:

a) Card-on-card:
- Flex circuit (connecting stacked multiple cards using special wiring)
- Connect-on-connector (connecting multiple cards using connectors) or

b) Package-on-package
- Chip-on-chip (stacking chips on top of one another, using interchip traces on small boards to de-mux signals for control and data pins of stacked chips)
- Die-on-die (stacking multiple dies in the same multi-chip package (MCP) or chip scale package (CSP)

Chip-on-chip implementation is the most cost-effective way of stacking packages as the same commercial equipment is used to manufacture the stacked parts, modules, or cards.

The most popular chip-on-chip stacks are dual or quad stacks primarily used for DRAM, flash and SRAM module builds to increase depth or density. Typically, TSOP packages are stacked to offer over 50 percent and 77 percent board space savings, respectively. At the dual stack level, the height of the stacked devices allows for meeting JEDEC height guidelines for memory modules. Therefore, no mechanical restriction is placed on system designers.



Stacked DIMM

Issues and Benefits

Card-on-card stacking achieves higher density by using existing memory chip technology one way or the other (brute force). Disadvantages include:

1) Non symmetric - air flow and space problems, (multiple PLLs, longer signal traces)
2) Monolithic memory components waste board space at the chip level
3) Not as cost-effective (multiple PCB boards, connector cost)
4) Both flex circuit or connectors problematic for manufacturing
5) Flex circuit susceptible to lead damage
6) Connector on connector not good for shock and vibration constraints

Chip-on-chip technology, while it may raise thermal issues or be susceptible to lead and package damage for 2+ stacking at the device level, offers mitigating benefits:

1) one PLL
2) standard Gerber
3) standard manufacturing/test procedures
4) shorter signal trace
5) less costly to manufacture

In total, then, chip-on-chip stacking is the preferred method for using current memory technology to access next-generation memory densities. That is the method preferred by SimpleTech, which developed a patented IC Tower™ stacking technology.

To enable inter-chip connectivity, special PCB sideboards make the connection from bottom chip to top chip. The PCB sideboards directly connect common signals, such as data pins between the top and bottom chips. In the case of chip selects, the top chip pins are routed to an internal layer on the sideboard PCB and onto the unused pin on the bottom chip. Selectively rerouting to chip select pins turns on the top and bottom chips at different times, allowing both stacked parts to share the same data pins.

SimpleTech's stacking process utilizes standard surface mount equipment. As a result, over a million stacked devices per month may be produced utilizing only a single surface mount manufacturing line. IC Tower ® Stacking Technology is applied to build large quantities of stacked DRAM, flash, and SRAM-based stacked memory solutions.

Applications drive the Need for Stacking

Applications driving the need for stacking are low profile, small footprint server appliances as well as high-end servers and workstations that need to maximize system memory availability with minimum space impact. Key operating systems such as Unix, Linux, and Windows XP can benefit from more stacked server DRAM in the system.

Added DRAM capacity support by new and enhanced operating systems means the high-speed 3GHz+ multi-processing system can better balance CPU cycles with added high-speed physical memory. SimpleTech addresses the exponential demand for high-density memory by offering DDR-based as well as DDR2-based memory solutions, utilizing its IC Tower for peak density per memory socket. Instead of waiting for 1 Gb monolithic chips to be readily available and cost-effective, SimpleTech stacks today's highest possible densities of SDRAM, DDR or DDR2, using 512 Mb chips to achieve 1 Gb densities in the same footprint.

Given the increase in the clock rate of DDR and DDR2 memories, setup and hold times for data access are being shaved off. Having parts stacked on top of each other reduces the trace length on the board for the same density by half, allowing easier time for system and board layout designers with the signal trace management on the boards. This is achieved by eliminating the wire trace between two monolithic components, allowing minimum trace between stacked parts.

Other applications include networking and telecom applications requiring the highest possible DRAM density in the lowest possible surface area on embedded boards. To sustain high bandwidth throughputs at the control and data planes of RISC processor and network processor-based designs, for example, designers may opt to offer multiple channels of memory access utilizing high density stacked DDR2 SDRAM as the memory subsystem. As typical router or telecom gateway DRAM usage moves from 512 MB to 1GB or 2 GB per memory channel, IC Tower stacking may be the only way to achieve the higher density.

In summary, when a system design requires the highest possible memory density/ functionality per surface area, IC Tower stacking can make solutions available in viable cost models today without waiting for next generation memory components.