THE WALL STREET JOURNAL.

The Secret to Tech's Next Big Breakthroughs? Stacking Chips

As microchips become 3-D, there are dividends in performance, power consumption and capabilities

One of the most advanced 3-D chip packages has powered the Apple Watch since its introduction, an analyst says.
PHOTO: IFIXIT

By Christopher Mims
Nov. 19, 2017 9:00 a.m. ET

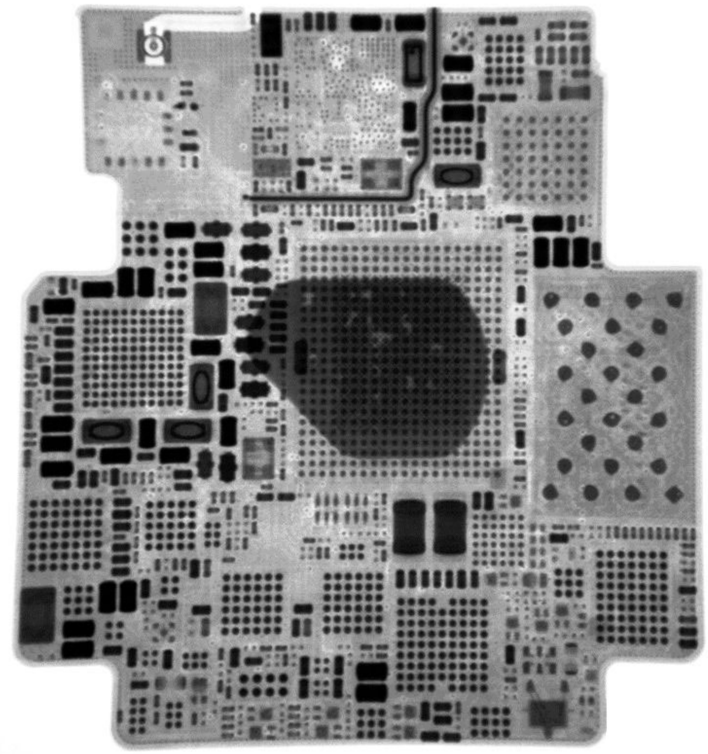
A funny thing is happening to the most basic building blocks of nearly all our devices. Microchips, which are usually thin and flat, are being stacked like pancakes.

Chip designers—now playing with depth, not just length and width—are discovering a variety of unexpected dividends in performance, power consumption and capabilities.

Without this technology, the Apple Watch wouldn't be possible. Nor would the most advanced solid-state memory from Samsung, artificial-intelligence systems from Nvidia and Google, or Sony's crazy-fast next-gen camera.

Think of this 3-D stacking as urban planning. Without it, you have sprawl—microchips spread across circuit boards, getting farther and farther apart as more components are needed. But once you start stacking chips, you get a silicon cityscape, with everything in closer proximity.

The advantage is simple physics: When electrons have to travel long distances through copper wires, it takes more power, produces heat and reduces bandwidth. Stacked chips are more efficient, run cooler and communicate across much shorter interconnections at lightning speed, says Greg Yeric, director of future silicon technology for ARM Research, part of microchip design firm ARM.



X-ray of the Apple S1 'chip' from an Apple Watch Series 1.
PHOTO: CHIPWORKS

Cross-section view of the Apple S1 chip from an Apple Watch Series 1.
PHOTO: CHIPWORKS

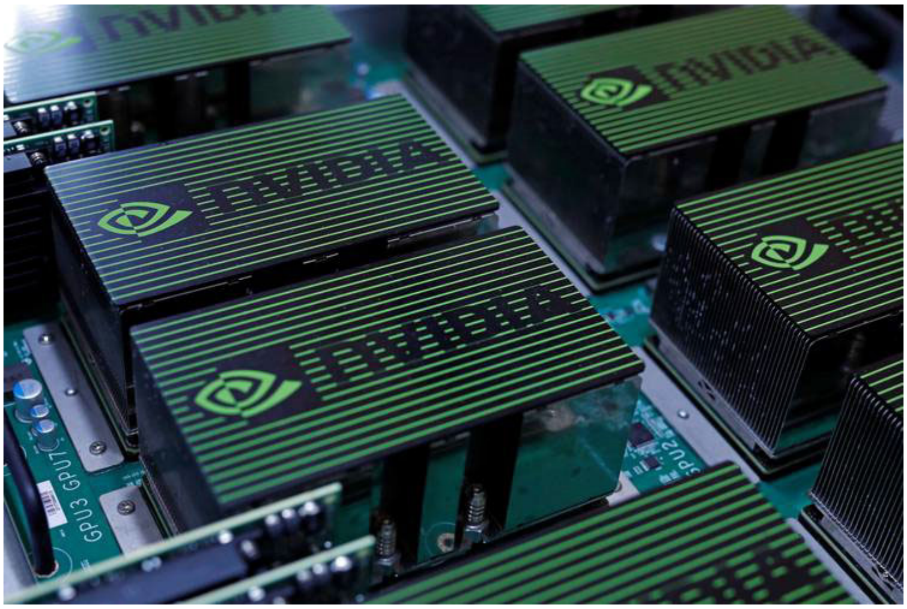
While the principles that underlie 3-D microchips are straightforward, making them is anything but. First proposed in the 1960s, the technology has sporadically appeared in high-end applications, such as military hardware, Mr. Yeric says.

But stacked-chip offerings from most major chipmakers—AMD, Intel, Apple, Samsung and Nvidia —plus smaller, specialized companies like Xilinx, have been around only five years or so, says Sinjin Dixon-Warren, an analyst at microchip research firm TechInsights. What changed?Engineers started running out of other ways to squeeze more performance out of microchips.

Stacked chips are frequently part of a "package" of other scrunched-together chips. In addition to saving space, this lets makers create many different chips—with different manufacturing processes—and then more or less literally glue them all together. The "3-D system in package" approach contrasts with the "system on a chip" approach frequently used in mobile phones, where all the different components of the phone are etched on a single piece of silicon.

One of the most advanced 3-D chip packages has powered the Apple Watch since its introduction, Mr. Dixon-Warren says. Thirty different chips are hermetically sealed inside a plastic envelope. To save space, memory is stacked on top of the logic circuit, he says. The watch couldn't be so compact without chip stacking.

But where Apple's chips are stacked only two stories high, Samsung has produced a veritable silicon high-rise. Samsung's V-NAND flash memory, used for storing data in phones, cameras and laptops, has 64 chips placed one atop the other. Samsung just announced that a future version will have 96 layers.



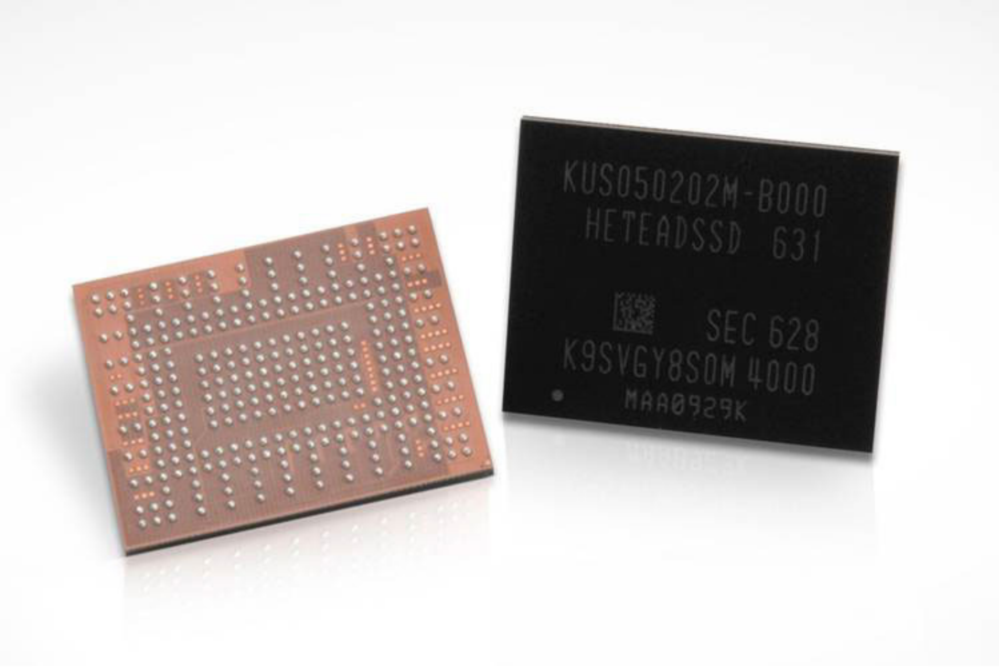
Nvidia's Volta microprocessors are built for artificial intelligence, with up to eight layers of high-bandwidth memory stacked onto the GPU. Shown, Nvidia chips exhibited at the Computex show in Taipei in May.
PHOTO: TYRONE SIU/REUTERS

Memory is a natural application for chip-stacking technology, since it solves a problem that has long plagued chip designers: Adding more cores to anything from an iPad to a supercomputer didn't translate to hoped-for speed gains because of the communications lag between logic circuits and the memory they need to do their jobs. Sticking memory right on top of chips allows for many more short connections between the two.

That's how Nvidia's built-for-AI Volta microprocessors work, says Brian Kelleher, the company's senior vice president of hardware engineering. By stacking up to eight layers of high-bandwidth memory directly on top of the GPU, these chips are breaking records in processing efficiency.

"We are power-limited," says Mr. Kelleher, referring to the amount of electricity a system is budgeted, which can be eaten up by both the power put into it and the heat it generates. "Any power we can take out of the memory system, we can put into computation."

Chip stacking enables totally new capabilities too. Some phone cameras stack an image sensor directly on top of the chip that processes the image. The extra speed means they can grab multiple exposures of an image and fuse them together, capturing more light for dim scenes.



Samsung's 64-layer vertical V-NAND chips, displayed in August 2016, which bigger data storage and faster processing.
PHOTO: YONHAP NEWS/ZUMA PRESS

A prototype camera from Sony goes further by using three layers, not just two—a Dagwood sandwich of image sensor, memory and logic circuit that allows for snapping up to 1,000 frames per second. The result is that light hits the sensor, and that data is dumped directly into memory, where it can be processed in real time. In addition to achieving better visibility in low light, this can also be used for super slow-mo video and freezing fast-moving objects in a single frame.

Right now, there are still costly barriers to getting 3-D microchips into more devices.

For starters, 3-D chips are so new, the design tools used to lay them out simply haven't evolved enough yet, Mr. Yeric says. Until simple design tools—the kind currently used for flat chips—are widely available, stacked chips will continue to be the sole purview of companies with the most engineering talent.

Another problem is that manufacturers are still learning how to physically stack chips atop one another and connect them reliably. This means lower yields, or fewer usable chips, coming from some of the manufacturing processes.

But Mr. Dixon-Warren says the spread of 3-D chips is rapid and their takeover inevitable. A decade ago, this technology was limited almost exclusively to university labs; five or six years ago, it was still hard to find commercial examples. But now it's popping up all over, in applications like networking and high-performance computing and in high-end wearables like the Apple Watch. It is also in evidence in the brains of the iPhone X, says Kyle Wiens, chief executive of iFixit, which disassembles electronics to evaluate how easy they are to repair.

Eventually, 3-D chips should make our wearables as capable as much larger devices, and allow them to operate for days even as they bristle with sensors, ARM's Mr. Yeric says. "I would not be surprised if for example someday your watch can check your blood sugar," he said.

Taking chips from Flatland into the Third Dimension is just the beginning. Soon, the layers will communicate with light instead of electricity. Further down the line, they will move off silicon entirely—perhaps to synthetic diamond—as we exchange circuit boards for glowing crystals of unprecedented processing power. Just like Superman.

Write to Christopher Mims at christopher.mims@wsj.com