Low-cost, scalable process for integrating GaN transistors onto silicon CMOS chips

An MIT-led team shows how Intel 16 22nm FinFET metallization and passive options enable the incorporation of components such as neutralization capacitors.

Gallium nitride (GaN) will likely be key for the next generation of high-speed communication systems and the power electronics needed for state-of-the-art data centers, but its high cost and the specialization required to incorporate it into conventional electronics have limited its use in commercial applications, notes the USA's Massachusetts Institute of Technology (MIT).

Now, researchers from MIT and elsewhere have developed a new fabrication process that integrates high-performance GaN transistors onto standard silicon CMOS chips in a way that is low-cost and scalable, and compatible with existing semiconductor foundries.

Their method involves fabricating many transistors on the surface of a GaN chip, cutting out each individual transistor, and then bonding just the necessary number of transistors onto a silicon chip using a low-temperature process that preserves the functionality of both materials.

The cost is minimal, since only a small amount of GaN material is added to the chip, but the resulting device can gain a significant performance boost from compact, high-speed transistors. In addition, by separating the GaN circuit into discrete transistors that can be spread over the silicon chip, the new technology is able to reduce the temperature of the overall system.

The researchers used this process to fabricate a power amplifier that achieves higher signal strength and efficiencies than devices consisting of silicon transistors. In a smartphone, this could improve call quality, boost wireless bandwidth, enhance connectivity, and extend battery life, it is reckoned.

Because their method fits into standard processes, it could improve existing electronics as well as future technologies. In the future, the new integration scheme could even enable quantum applications, as gallium nitride performs better than silicon at the cryogenic temperatures essential for many types of quantum computing.

"If we can bring the cost down, improve the scalability and, at the same time, enhance the performance of the electronic device, it is a no-brainer that we should adopt this technology," says Pradyot Yadav, an MIT graduate student and lead author of a paper on this method. "We've combined the best of what exists in silicon with the best possible gallium nitride electronics. These hybrid chips can revolutionize many commercial markets."

He is joined on the paper by fellow MIT graduate students Jinchen Wang and Patrick Darmawi-Iskandar; MIT postdoc John Niroula; senior authors Ulrich L. Rohde, a visiting scientist at the Microsystems Technology Laboratories (MTL), and Ruonan Han, an associate professor in the Department of Electrical Engineering and Computer Science (EECS) and member of MTL; and Tomás Palacios, the Clarence J. LeBel Professor of EECS, and director of MTL; as well as collaborators at Georgia Tech and the Air Force Research Laboratory. The research was presented in the paper 'RTu2C-4: 3D-Millimeter Wave Integrated Circuit (3D-mmWIC): A Gold-Free 3D-Integration Platform for Scaled RF GaN-on-Si Dielets with Intel 16 Si CMOS' at the 2025 IEEE Radio Frequency Integrated Circuits Symposium (IMS) in San Francisco, CA, USA (15–17 June).

Swapping transistors

Gallium nitride is the second most widely used semiconductor after silicon, and its unique properties suit applications such as lighting, radar systems and power electronics. But to access its maximum performance, it is important for GaN chips to be connected to digital chips made of CMOS silicon.

To enable this, some integration methods bond GaN transistors onto a CMOS chip by soldering the connections, but this limits how small the GaN transistors can be. The smaller the transistors, the higher the frequency at which they can work.

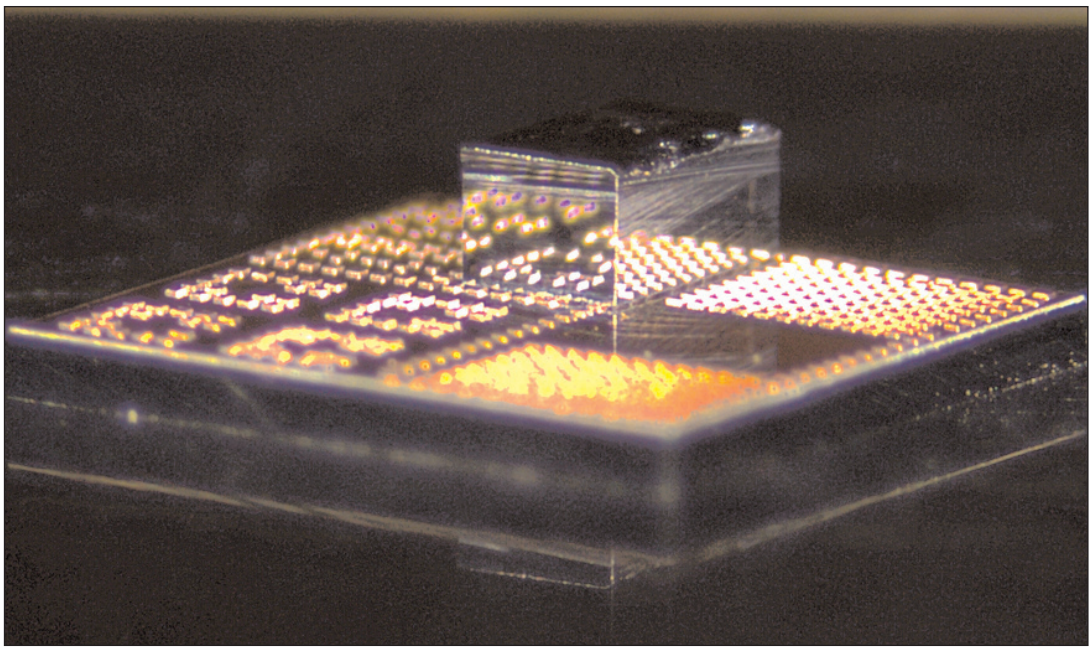
Other methods integrate an entire GaN wafer onto a silicon wafer, but using so much material is extremely costly, especially since the GaN is only needed in a few small transistors. The rest of the material in the GaN wafer is wasted.

"We wanted to combine the functionality of GaN with the power of digital chips made of silicon, but without having to compromise on either cost or bandwidth. We achieved that by adding super-tiny discrete gallium nitride transistors right on top of the silicon chip," Yadav says.

The new chips are the result of a multi-step process.

First, a tightly packed collection of transistors is fabricated across the entire surface of a GaN wafer. Using very fine laser technology, they cut each one down to just the size of the transistor, which is 240μm x 410μm, forming a dielet.

Each transistor is fabricated with copper pillars on top, to bond directly to the copper pillars on the surface of a standard silicon CMOS chip. Copper-to-copper bonding can be performed at temperatures below 400°C, which is low enough to avoid damaging either material.



A new fabrication process integrates high-performance GaN transistors onto standard silicon CMOS chips in a way that is low-cost and scalable.

Existing GaN integration techniques require bonds that use expensive gold material, which needs much higher temperatures and stronger bonding forces than copper. Since gold can contaminate the tools used in most semiconductor foundries, it typically requires specialized facilities.

"We wanted a process that was low-cost, low-temperature and low-force, and copper wins on all of those related to gold. At the same time, it has better conductivity," Yadav notes.

A new tool

To enable the integration process, the researchers created a specialized new tool that can carefully integrate the extremely small GaN transistor with the silicon chips. The tool uses a vacuum to hold the dielet as it moves on top of a silicon chip, zeroing in on the copper bonding interface with nanometer precision.

They used advanced microscopy to monitor the interface and then, when the dielet is in the right position, they apply heat and pressure to bond the GaN transistor to the chip.

"For each step in the process, I had to find a new collaborator who knew how to do the technique that I needed, learn from them, and then integrate that into my platform. It was two years of constant learning," Yadav says.

Once the researchers had perfected the fabrication process, they demonstrated it by developing power amplifiers for boosting wireless signals. Their devices achieved higher bandwidth and better gain than devices made with traditional silicon transistors. Each compact chip has an area of less than half a square millimeter.

In addition, because the silicon chip they used in their demonstration is based on Intel 16 22nm FinFET metallization and passive options, they were able to incorporate components often used in silicon circuits, such as neutralization capacitors. This greatly improved the gain of the amplifier, bringing it one step closer to enabling the next generation of wireless technologies.

"To address the slowdown of Moore's Law in transistor scaling, heterogeneous integration has emerged as a promising solution for continued system scaling, reduced form factor, improved power efficiency, and cost optimization," says Atom Watanabe, an IBM research scientist who was not involved with the published paper. "Particularly in wireless technology, the tight integration of compound semiconductors with silicon-based wafers is critical to realizing unified systems of front-end integrated circuits, baseband processors, accelerators and memory for next-generation antennas-to-AI platforms," he adds. "This work makes a significant advancement by demonstrating 3D integration of multiple GaN chips with silicon CMOS and pushes the boundaries of current technological capabilities."

This work is supported, in part, by the US Department of Defense through the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program and CHIMES, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation Program by the Department of Defense and the Defense Advanced Research Projects Agency (DARPA). Fabrication was carried out using facilities at MIT.Nano, the Air Force Research Laboratory, and Georgia Tech. ■

https://rfic-ieee.org/technical-program/rfic-technical-sessions?date=2025-06-17

www.mtl.mit.edu