How and When the Chip Shortage Will End
Fabs using older process nodes are the key

IEEE Spectrum

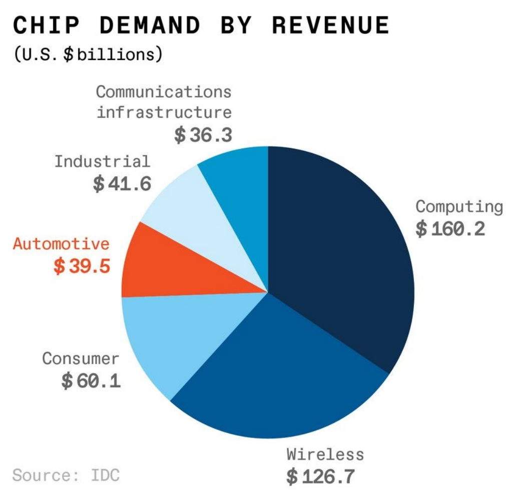
At the beginning of the COVID-19 pandemic, the auto industry cut back on production, thereby cutting back on the chips that operate their cars. Now as chips are in high demand—for cars and many other electronics—chipmakers are facing challenges to fulfill their backlog of orders.

Historians will probably spend decades picking apart the consequences of the COVID-19 epidemic. But the shortage of chips that it is caused will be long over by then. A variety of analysts agree that the most problematic shortages will begin to ease in the third or fourth quarter of 2021, though it could take much of 2022 for the resulting chips to work their way through the supply chain to products. The supply relief will not be coming from the big, national investments in the works right now by South Korea, the United States, and Europe but from older chip fabs and foundries running processes far from the cutting edge and on comparatively small silicon wafers.

Before we get into how the shortage will end, it is worth summing up how it began. With panic, lockdowns, and general uncertainty rolling across the globe, automakers cancelled orders. However, those conditions meant, among others;

- A big fraction of the workforce recreated the office **at home**, purchasing computers, monitors, and other equipment.
- At the same time entire **school systems switched** to virtual learning via laptops and tablets. And more time at home also meant more spending on home entertainment, such as TVs and game consoles.
- The **5G** rollout, and
- Continued growth in **cloud computing** quickly hoovered up the capacity automakers had unceremoniously freed.

By the time car makers realized people still wanted to buy their goods they found themselves at the back of the line for the chips they needed.

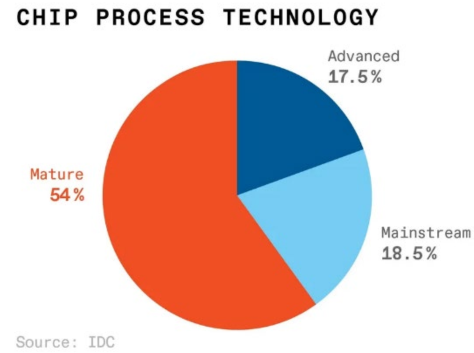


At $39.5 billion, the auto industry makes up less than 9 percent of chip demand by revenue, according to market research firm IDC. That figure is set to increase by about 10 percent per year to 2025. However, the auto industry— which employs more than 10 million people globally— is something both consumers and politicians are acutely sensitive to, especially in the United States and Europe.

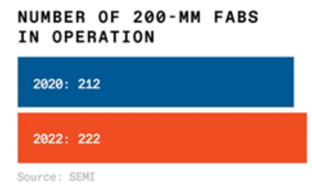
Chips for the automotive sector are made using processes intended to meet safety criteria that are different from those meant for other industries. But they are still fabricated on the same production lines as the analog ICs, power management chips, display drivers, microcontrollers, and sensors that go in everything else. “The common denominator is the process technology is 40 nanometers and older," says Mario Morales, vice president, enabling technologies and semiconductors at IDC.

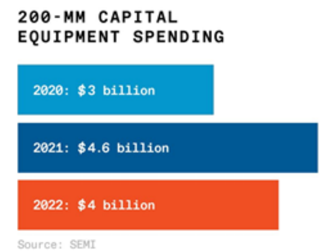
This chip manufacturing technology was last cutting edge 15 years ago or earlier, lines producing chips at these old nodes represent a full 54 percent of installed capacity, according to IDC. Today these old nodes are typically used on 200-mm wafers of silicon. To reduce cost, the industry began moving to 300-mm wafers in 2000, but much of the old 200-mm infrastructure continued and even expanded.

Despite the auto industry's desperation, there is no great rush to build new 200-mm fabs. “The return on investment just isn't there," says Morales. What is more, there are already many legacy-node plants in China that are not operating efficiently right now, but “at some point, they will," he says, further reducing the incentive to build new fabs. According to the chip manufacturing equipment industry association SEMI, the number of 200-mm fabs will go from 212 in 2020 to 222 in 2022, about half the expected increase of the more profitable 300-mm fabs.



More than 40 companies began increasing capacity by more than 750,000 wafers-per-month from the beginning of 2020 to the end of 2022. The long-term trend to the end of 2024 is for a 17 percent increase in capacity for 200-mm facilities. Spending on equipment for these fabs is set to rise to $4.6 billion in 2021 after crossing the $3-billion mark in 2020 for the first time in years, SEMI says. But then spending will drop back to $4 billion in 2022. In comparison, spending to equip 300-mm fabs was expected to hit $78 billion in 2021.





The chip shortage is happening simultaneously with national and regional efforts to boost advanced logic chip manufacturing. South Korea announced a push worth $450 billion over ten years, the United States is pushing legislation worth $52 billion, and the EU could plow up to $160 billion into its semiconductor sector.

Chipmakers were already on a spending spree. Globally, capital equipment for semiconductor production grew 56 percent year-on-year through April 2021, according to SEMI. SEMI's 3 June 2021 World Fab Forecast indicated that 10 new 300-mm fabs will start operation in 2021 with 14 more coming up in 2022.

“The push for building IC capacity around the world will certainly drive fab investment of the current decade to a new high," says Christian Gregor Dieseldorff, senior principal for semiconductors at SEMI. “We expect to see record spending and more new fab announcements in the next few years."

One potential hiccup on the road to ending the shortage is that some of the skyrocketing demand appears to be from customers that are double ordering to bulk up on inventory, says Jim Feldhan, president of Semico Research. “I don't know of any product that needs twice the amount of analog" as the year before, he says. But manufacturers “don't want a 12-cent part to hold up a 4K television," so they are stocking up.

The auto industry needs to do more than just stock up, according to Bharat Kapoor, lead partner, Americas, in the high-tech practice of global strategy and management consulting firm, Kearney. To keep future shortages at bay, the chip industry and auto executives need a more direct connection going forward so signals about supply and demand are clearer, he says.

The Chip Shortage, Giant Chips, and the Future of Moore’s Law

With COVID-19 shaking the global supply chain like an angry toddler with a box of jellybeans, the average person had to take a crash course in the semiconductor industry. And many of them didn't like what they learned. Want a new car? Tough luck, not enough chips. A new gaming system? Same. But you are not the average person, dear reader. So, in addition to learning why there was a chip shortage in the first place, you also discovered that you can—with considerable effort—fit more than 2 trillion transistors on a single chip. You also found that the future of Moore's Law depends as much on where you put the wires as how small you make the transistors, among many other things.

So, to recap the semiconductor stories you read most this year, we've put together this set of highlights:

How and When the Chip Shortage Will End (See above)

This year you learned the same thing that some carmakers did: Even if you think you've hedged your bets by having a diverse set of suppliers, those suppliers—or the suppliers of those suppliers—might all be using the output of the same small set of semiconductor fabs.

To recap: Carmakers panicked and canceled orders at the outset of the pandemic. Then when it seemed people still wanted cars, they discovered that all of the display drivers, power-management chips, and other low-margin stuff they needed had already been sucked up into the work/learn/live-from-home consumer frenzy. By the time they got back in line to buy chips, that line was nearly a year long, and it was time to panic again.

Chipmakers worked flat out to meet demand and have unleashed a blitz of expansion, though most of that is aimed at higher-margin chips than those that clogged the engine of the automotive sector. The latest numbers, from the chip manufacturing equipment industry association SEMI, show sales of equipment are set to cross US $100 billion in 2021—a mark never before reached.

As for carmakers, they may have learned their lesson. At a gathering of stakeholders in the automotive electronics supply chain this summer at GlobalFoundries Fab 8 in Malta, N.Y., there was enthusiastic agreement that carmakers and chip makers needed to get cozy with each other. The result? GlobalFoundries has already inked agreements with both Ford and BMW.

Next-Gen Chips Will Be Powered From Below Transistors

You can make transistors as small as you want, but if you can't connect them to each other, there's no point. So, Arm and the Belgian research institute Imec spent a few years finding room for those connections. The best scheme they found was to take the interconnects that carry power to logic circuits (as opposed to data) and bury them under the surface of the silicon, linking them to a power-delivery network built on the backside of the chip. This research trend suddenly became news when Intel said what sounded like "Oh yeah. We're definitely doing that in 2025."

Cerebras's New Monster AI Chip Adds 1.4 Trillion Transistors

What has 2.6 trillion transistors, consumes 20 kilowatts, and carries enough internal bandwidth to stream a billion Netflix movies? It's generation 2 of the biggest chip ever made, of course! (And yes, I know that's not how streaming works, but how else do you describe 220 petabits per second of bandwidth?) Last April, Cerebras Systems topped its original, history-making AI processor with a version built using a more advanced chipmaking technology. The result was a more than doubling of the on-chip memory to an impressive 40 gigabytes, an increase in the number of processor cores from the previous 400,000 to a speech-stopping 850,000, and a mind-boggling boost of 1.4 trillion additional transistors.

Gob-smacking as all that is, what you can do with it is really what's important. And later in the year, Cerebras showed a way for the computer that houses its Wafer Scale Engine 2 to train neural networks with as many as 120 trillion parameters. For reference, the massive—and occasionally foul-mouthed—GPT-3 natural-language processor has 175 billion. What's more, you can now link up to 192 of these computers together.

Of course, Cerebras's computers aren't the only ones meant to tackle absolutely huge AI training jobs. SambaNova is after the same title, and clearly Google has its eye on some awfully big neural networks, too.

IBM Introduces the World's First 2-nm Node Chip

IBM claimed to have developed what it called a 2-nanometer node chip and expects to see it in production in 2024. To put that in context, leading chipmakers TSMC and Samsung are going full-bore on 5 nm, with a possible cautious start for 3 nm in 2022. As we reminded you last year, what you call a technology process node has absolutely no relation to the size of any part of the transistors it constructs. So, whether IBM's process is any better than rivals will really come down to the combination of density, power consumption, and performance.

The real importance is that IBM's process is another endorsement of nanosheet transistors as the future of silicon. While each big chipmaker is moving from today's FinFET design to nanosheets at their own pace, nanosheets are inevitable.

RISC-V Star Rises Among Chip Developers Worldwide

The news hasn't all been about transistors. Processor architecture is increasingly important. Your smartphones' brains are probably based on an Arm architecture, your laptop and the servers it's so attached to are likely based on the x86 architecture. But a fast-growing cadre of companies, particularly in Asia, are looking to an open-source chip architecture called RISC-V. The attraction is to allow startups to design custom chips without the costly licensing fees for proprietary architectures.

Even big companies like Nvidia are incorporating it, and Intel expects RISC-V to boost its foundry business. Seeing RISC-V as a possible path to independence in an increasingly polarized technology landscape, Chinese firms are particularly bullish on RISC-V. Only last month, Alibaba said it would make the source code available for its RISC-V core.

New Optical Switch Up to 1,000x Faster Than Transistors

Although certain types of optical computing are getting closer, the switch researchers in Russia and at IBM described in October is likely for a computer that's far in the future. Relying on exotic stuff like exciton-polaritons and Bose-Einstein condensates, the device switched at about 1 trillion times per second. That's so fast that light would manage only about one third of a millimeter before the device switches again.

New Type of DRAM Could Accelerate AI

One of AI's big problems is that its data is so far away. Sure, that distance is measured in millimeters, but these days that's a long way. (Somewhere there's an Intel 4004 saying, "Back in my day, data had to go 30 centimeters, uphill, in a snowstorm.") There are lots of ways engineers are coming up with to shorten that distance. But this one really caught your attention:

Instead of building DRAM from silicon transistors and a metal capacitor built above it, use a second transistor as the capacitor and build them both above the silicon from oxide semiconductors. Two research groups showed that these transistors could keep their data way longer than ordinary DRAM, and they could be stacked in layers above the silicon, giving a much shorter path between the processor and its precious data.

Intel Unveils Big Processor Architecture Changes

In August 2021 Intel unveiled what it called the company's biggest processor architecture advances in a decade. They included two new x86 CPU core architectures—the straightforwardly named Performance-core (P-core) and Efficient-core (E-core). The cores are integrated into Alder Lake, a "performance hybrid" family of processors that includes new tech letting Windows 11 OS run CPUs more efficiently.

"This is an awesome time to be a computer architect," senior vice president and general manager Raja Koduri said at the time. The new architectures and SoCs Intel unveiled "demonstrate how architecture will satisfy the crushing demand for more compute performance as workloads from the desktop to the data center become larger, more complex, and more diverse than ever."

If you want, you could translate that as: "In your face, process technology and device scaling! It's all about the architecture now!" But I don't think Koduri would take it that far.

U.S. Takes Strategic Step to Onshore Electronics Manufacturing

A bit alarmed by just how geographically close China is to Taiwan and Samsung, the only two countries capable of making the most advanced logic chips, U.S. lawmakers got the ball rolling on an effort to boost cutting-edge chipmaking in the United

States. Some of that has already started with TSMC, Samsung, and Intel making major fab investments. Of course, Taiwan and South Korea are also making major domestic investments, as are Europe and Japan.

It’s all part of a broader economic and technological nationalism playing out across the world, notes geopolitical futurist Abishur Prakash, with the Center for Innovating the Future, in Toronto. Some see these “shifts in geopolitics as short term, as if they’re by-products of the pandemic and that things on a certain timeline will calm down if not return to normal,” he told IEEE Spectrum in May. “That’s wrong. The direction that nations are moving in now is the new permanent North Star.”

Semiconductor Industry Association, the U.S. trade group, says government incentives will accelerate construction. The SIA calculates that a $20-billion incentive program over 10 years would yield 14 new fabs and attract $174 billion in investment versus 9 fabs and $69 billion without the federal incentives. A $50-billion program would yield 19 fabs and attract $279 billion.

The NDAA specifies a cap of $3-billion per project unless Congress and the President agree to more, but how much money actually gets spent in total on microelectronics capacity will depend on separate “appropriations” bills.

“The next step is for leaders in Washington to fully fund the NDAA’s (National Defense Authorization Act of 2021) domestic chip manufacturing incentives and research initiatives,” said Bob Bruggeworth, chair of SIA and president, CEO, and director of RF-chipmaker Qorvo, in a press release.

Beyond financial incentives, the NDAA also authorizes microelectronics-related R&D, development of a “provably secure” microelectronics supply chain, the creation of a National Semiconductor Research Technology Center to help move new technology into industrial facilities, and establishment of committees to create strategies toward adding capacity at the cutting edge. It also authorizes quantum computing and artificial intelligence initiatives.

Event-Based Camera Chips Are Here. What’s Next?

Hey, remember all that brain-based processing stuff we've been banging on about for decades? Well, it's here now, in the form of a camera chip made by French startup Prophesee and major imager manufacturer Sony. Unlike a regular imager, this chip doesn't capture frame after frame with each tick of the clock. Instead, it notes only the changes in a scene. That means both much lower power—when there's nothing happening, there's nothing to see—and faster response times.

Semiconductor fabrication plant (fab) cost

Semiconductor fabrication involves various processing steps, starting from the collection of raw materials, process gases, selection of wafers, epi process of semiconductors, device fabrication, packaging and integration of devices. Huge amount of money is needed to perform each of these processing steps - **Krishna Yaddanapudi**, Materials Scientist at University of California, Davis.

About fifteen years ago, the cost of setting up a new state-of-the-art semiconductor fab was estimated to be around 4~5 billion USD. Adjusting for inflation today (Inflation Calculator) it corresponds to a 50% increase, so it would be about 7.5 billion USD - **Jose Soares Augusto**, Professor at University of Lisbon in 2019.

Looking to the future, the costs of keeping up with technology are rising. Designing a latest-generation chip is 10 times more expensive than older ones. Meanwhile, according to a report published by the Semiconductor Industry Association in 2020, it now takes as much as $20 billion to build a big chip plant with even more cutting-edge equipment. That’s more than a new aircraft carrier or power plant. Over a decade, a semiconductor factory can cost as much as $40 billion although state-led incentives could cut that by up to $13 billion. (**Bloomberg Opinion**).

China Plans to Create Committee to Collaborate With Intel, AMD

China plans to create a special committee to collaborate with big-name foreign chipmakers, such as Intel and AMD. With this new move, China aims to establish a domestic chip supply chain to circumvent the sanctions that the U.S. has imposed on the country.

The organization is called the "cross-border semiconductor work committee." If Nikkei's information is accurate, China may launch the committee in the first half of 2022. The Ministry of Commerce will oversee this special committee in conjunction with the Ministry of Industry and Information Technology.

According to Nikkei's sources, the committee's objective is to fortify the relationship between Chinese and foreign companies. The end game is to build a chip supply chain by securing semiconductor technologies from Japan, Europe, and the U.S. In addition, the committee wants to woo overseas companies into establishing development and manufacturing sites in China with the promise of cooperation with the local governments and funding.

Nikkei's acquired documents allegedly revealed that China targets Intel, AMD, and Infineon Technologies of Germany. In addition, an industrial group from the Netherlands, including ASML, is also on the list. The publication's sources claimed that a few companies had expressed their desire to jump on China's initiative. However, Intel and the Dutch group refused to comment when Nikkei contacted them.

It's not the first time China has set out to forge a partnership with a foreign chipmaker. For example, AMD had struck a joint venture with China via the Tianjin Haiguang Advanced Technology Investment Co. Ltd. (THATIC) to license some of its x86 and SoC IP. Then there's also the emergence of Zhaoxin, a joint venture between Via Technologies and the Chinese government, to produce x86 chips.

Semiconductor Manufacturing International Co. (SMIC), China's number one chipmaker, Advanced Micro-Fabrication Equipment Inc. China (AMEC), a semiconductor equipment maker and smartphone giant Xiaomi will reportedly participate in the program. The other participants include Tsinghua University, Peking University, and the Chinese Academy of Sciences.

In 2015, China announced the "Made in China 2025" $20 billion campaign to bolster the domestic semiconductor industry. The plan was to achieve 70% tech self-sufficiency by 2025. However, things aren't going as planned, according to IC Insights. The U.S. firm estimated that China had reached 16% in 2020, and even China's optimistic numbers had it at 30%.

The U.S. has already blocked the major Chinese players, including Huawei and SMIC, from obtaining U.S.-designed equipment, materials, and software. In addition, Chinese processor developers Phytium and Sunway and several quantum computing entities are also on the blacklist.

China's latest act could serve as a steppingstone for achieving its goals. However, other countries are also offering subsidies and incentives to attract foreign investment in the semiconductor industry so that China will have stiff competition. Additionally, IP theft continues to be a concern, such as a Micron - UMC fiasco, so foreign companies will be reluctant to bring their technology over to Chinese soil.