Tiny Chips, Big Headaches

As the largest computer networks continue to grow, some engineers fear that their smallest components could prove to be an Achilles' heel.

Imagine for a moment that the millions of computer chips inside the servers that power the largest data centers in the world had rare, almost undetectable flaws. And the only way to find the flaws was to throw those chips at giant computing problems that would have been unthinkable just a decade ago.

As the tiny switches in computer chips have shrunk to the width of a few atoms, the reliability of chips has become another worry for the people who run the biggest networks in the world. Companies like Amazon, Facebook, Twitter and many other sites have experienced surprising outages over the last year.

The outages have had several causes, like programming mistakes and congestion on the networks. But there is growing anxiety that as cloud-computing networks have become larger and more complex, they are still dependent, at the most basic level, on computer chips that are now less reliable and, in some cases, less predictable.

In the past year, researchers at both Facebook and Google have published studies describing computer hardware failures whose causes have not been easy to identify. The problem, they argued, was not in the software — it was somewhere in the computer hardware made by various companies. Google declined to comment on its study, while Facebook, now known as Meta, did not return requests for comment on its study.

"They're seeing these silent errors, essentially coming from the underlying hardware," said Subhasish Mitra, a Stanford University electrical engineer who specializes in testing computer hardware. Increasingly, Dr. Mitra said, people believe that manufacturing defects are tied to these so-called silent errors that cannot be easily caught.

Researchers worry that they are finding rare defects because they are trying to solve bigger and bigger computing problems, which stresses their systems in unexpected ways.

Companies that run large data centers began reporting systematic problems more than a decade ago. In 2015, in the engineering publication IEEE

Spectrum, a group of computer scientists who study hardware reliability at the University of Toronto reported that each year as many as 4 percent of Google's millions of computers had encountered errors that couldn't be detected and that caused them to shut down unexpectedly.

In a microprocessor that has billions of transistors — or a computer memory board composed of trillions of the tiny switches that can each store a 1 or 0 — even the smallest error can disrupt systems that now routinely perform billions of calculations each second.

At the beginning of the semiconductor era, engineers worried about the possibility of cosmic rays occasionally flipping a single transistor and changing the outcome of a computation. Now they are worried that the switches themselves are increasingly becoming less reliable. The Facebook researchers even argue that the switches are becoming more prone to wearing out and that the life span of computer memories or processors may be shorter than previously believed.

There is growing evidence that the problem is worsening with each new generation of chips. A report published in 2020 by the chip maker Advanced Micro Devices found that the most advanced computer memory chips at the time were approximately 5.5 times less reliable than the previous generation. AMD did not respond to requests for comment on the report.

Tracking down these errors is challenging, said David Ditzel, a veteran hardware engineer who is the chairman and founder of Esperanto Technologies, a maker of a new type of processor designed for artificial intelligence applications in Mountain View, Calif. He said his company's new chip, which is just reaching the market, had 1,000 processors made from 28 billion transistors.

He likens the chip to an apartment building that would span the surface of the entire United States. Using Mr. Ditzel's metaphor, Dr. Mitra said that finding new errors was a little like searching for a single running faucet, in one apartment in that building, that malfunctions only when a bedroom light is on and the apartment door is open.

Until now, computer designers have tried to deal with hardware flaws by adding to special circuits in chips that correct errors. The circuits automatically detect and correct bad data. It was once considered an exceedingly rare problem. But several years ago, Google production teams began to report errors that were

maddeningly difficult to diagnose. Calculation errors would happen intermittently and were difficult to reproduce, according to their report.

A team of researchers attempted to track down the problem, and last year they published their findings. They concluded that the company's vast data centers, composed of computer systems based upon millions of processor "cores," were experiencing new errors that were probably a combination of a couple of factors: smaller transistors that were nearing physical limits and inadequate testing.

In their paper "Cores That Don't Count," the Google researchers noted that the problem was challenging enough that they had already dedicated the equivalent of several decades of engineering time to solving it.

Modern processor chips are made up of dozens of processor cores, calculating engines that make it possible to break up tasks and solve them in parallel. The researchers found that a tiny subset of the cores produced inaccurate results infrequently and only under certain conditions. They described the behavior as sporadic. In some cases, the cores would produce errors only when computing speed or temperature was altered.

Increasing complexity in processor design was one important cause of failure, according to Google. But the engineers also said smaller transistors, three-dimensional chips and new designs that create errors only in certain cases all contributed to the problem.

In a similar paper released last year, a group of Facebook researchers noted that some processors would pass manufacturers' tests but then began exhibiting failures when they were in the field.

Intel executives said they were familiar with the Google and Facebook research papers and were working with both companies to develop new methods for detecting and correcting hardware errors.

Bryan Jorgensen, vice president of Intel's data platforms group, said that the assertions the researchers had made were correct and that "the challenge that they are making to the industry is the right place to go."

He said Intel had recently started a project to help create standard, open-source software for data center operators. The software would make it possible for them to find and correct hardware errors that the built-in circuits in chips were not detecting.

The challenge was underscored last year when several of Intel's customers quietly issued warnings about undetected errors created by their systems. Lenovo, the world's largest maker of personal computers, informed its customers that design changes in several generations of Intel's Xeon processors meant that the chips might generate a larger number of errors that couldn't be corrected than earlier Intel microprocessors.

Intel has not spoken publicly about the issue, but Mr. Jorgensen acknowledged the problem and said it had been corrected. The company has since changed its design.

Computer engineers are divided over how to respond to the challenge. One widespread response is demand for new kinds of software that proactively watch for hardware errors and make it possible for system operators to remove hardware when it begins to degrade. That has created an opportunity for new start-ups offering software that monitors the health of the underlying chips in data centers.

One such operation is TidalScale, a company in Los Gatos, Calif., that makes specialized software for companies trying to minimize hardware outages. Its chief executive, Gary Smerdon, suggested that TidalScale and others faced an imposing challenge.

"It will be a little bit like changing an engine while an airplane is still flying," he said.