Beating AI Bottlenecks with Better Switches

RUSSELL GARCIA, Chief Executive Officer, Menlo Microsystems

Companies are now considering new ways to build switch matrices for IC testers, using relays built using microelectromechanical systems (MEMS).

SOMETHING CHANGED WHEN WE STARTED measuring computing capacity in gigawatts rather than gigaflops. As AI models grew in scale, it became clear that raw computing speed was only one part of the story—testing, routing, and infrastructure were emerging as hard limits that the industry had to solve. The ability of machine-learning strategies, enabled by large language models, to produce behaviors that could be described as artificial intelligence (AI), has prompted huge investments in massive data centers (FIGURE 1). The scale of these investments is forcing the industry to confront constraints that once seemed secondary. The interesting question is no longer how fast AI models can scale, but where the supporting infrastructure begins to limit them. Mark Zuckerberg, CEO of Meta, has described the next Meta data center as covering an area equivalent to a "significant part of Manhattan," underscoring the scale of AI infrastructure.

It is not just Meta that is building AI capacity. According to the Artificial Intelligence Index Report 2025, published by Stanford University's Human-Centered AI research group, corporate investment in AI was $252.3 billion in 2024, up 25.5% on 2023. The largest portion of this was provided by private investments, which grew 44.5% on the previous year. It seems likely that both these metrics will have been surpassed in 2025.

In this rapidly evolving environment, where vast amounts of capital meet vast amounts of technology in pursuit of large-scale technological impact and industrial advancement, a range of factors can become a bottleneck to progress. Land with access to large supplies of cooling water commands a premium. Gaining access to the power grid becomes a critical issue, rather than a simple matter of permitting. Sourcing specialized materials, such as the glass cloth used to make the substrates upon which AI chip subsystems are assembled, becomes critical to their production. Demand for AI chips starts absorbing leading-edge fab capacity, affecting production timelines and costs for other products, such as fast DRAM and even cellphone chips, down the queue and up in price.

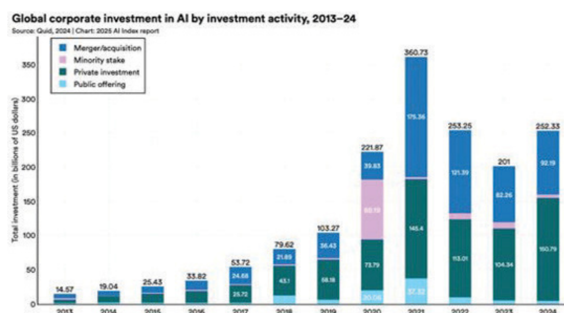
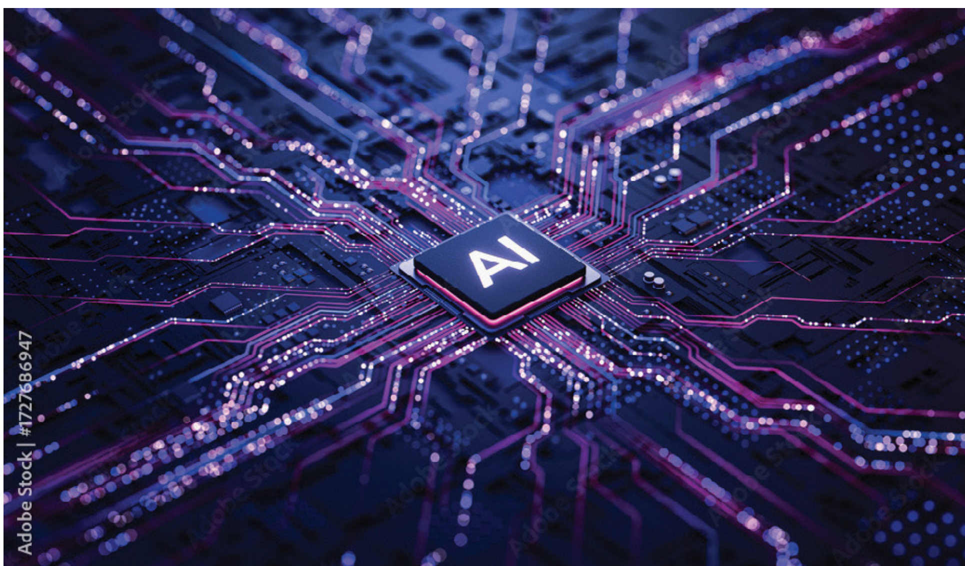


Figure 1. More than a trillion dollars have been invested in AI in the past five years. *Source: Stanford Human-Centered AI group – 2025 AI Index Report*

Testing times

Another potential bottleneck for this revolution is the testing of AI chips. Semiconductor test equipment, particularly that which tests bare die when they are still part of a wafer, is complex, expensive high-performance technology that must operate 24/7 in fast-moving production environments. The AI chips it is used to test have extremely dense pinouts, making them hard to access consistently with wafer probes. The advent of 2.5D and 3D integration technologies, such as high-bandwidth memory (HBM) stacks, or TSMC's chip-on-wafer-on-substrate (CoWoS) packaging approach, is making test access even harder. And much of the tester circuitry must run at the speed of, or faster than, the fastest signal on any die it encounters to make meaningful measurements.

Constraints on physical access to test points, signal integrity issues, the size of the measurement electronics needed for each pin, and the capital cost of such sophisticated equipment, mean that testers



Much of the tester circuitry must run at the speed of, or faster than, the fastest signal on any AI chip it encounters to make meaningful measurements.

must operate on a divide-and-conquer basis. Optimizing these systems is critical, because even small delays in testing can limit the pace at which AI innovations reach the market.

A typical automatic test system for ICs combines several major functional blocks. A central controller and timing engine coordinates test-pattern generation, triggering, and measurement sequencing across all pins. Per-channel cards provide drivers, receivers, level shifting, and protection for each pin under test. Source/measurement units, specialized RF instruments, digitizers and other functions, are shared resources that are accessed through high-density switching and routing matrices, to connect the instruments and pin electronics to the pins under test.

These matrices are often constructed using an analog crosspoint array, whose rows connect to instruments and whose columns connect to pins under test, using relays. These matrices are often expandable through an analog backplane, so that the rows and/or columns can be extended straightforwardly by adding more crosspoint boards.

The probe boards which make the connection to the pins under test have local multiplexers to select among multiple test paths, for example for DC parametric tests or for access to boundary scan chains. The boards will also have local relays to isolate pins and switch them between roles, such as functional or parametric testing.

Time for testing

The large capital cost of IC testers and the pressure to get chips through production and validation mean that test engineers must work within strict time budgets. They need to work out how best to allocate time on the tester to each type of test they may want to do, how long each type of test should run, considering how long it takes to reconfigure the tester to move between the different types of test. It turns out a significant amount of this time is taken up in switching the signal routes through the crosspoint matrix to connect pins under test to different testing resource.

The challenge is to minimize this time without affecting signal integrity or the functional density of the tester. One of the constraints that must be managed is that switches in the matrix must break-before-make, rather than make-before-break, to avoid creating short circuits. Another is to ensure that switching only happens when a signal is absent, to protect contacts that are carrying high voltages from potential arc erosion. A further source of delay is in the matrix switches themselves, which are often implemented using reed relays. These use a magnetic field to move an armature to make or break an electrical contact and so have both an actuation time and a settling time to be considered. It is also possible that reed relay armatures will vibrate even after contact has been made, due to the collapsing magnetic field of the switching coil, which can introduce noise into the signal from the pin under test.

Some companies are now considering new ways to build switch matrices for IC testers, using relays built using microelectromechanical systems (MEMS) production techniques adapted from the semiconductor industry with Menlo Micro as a leading innovator in this space. These relays have near to ideal switching characteristics. They have almost zero resistance when closed, infinite resistance when open, are isolated from all other switches in a circuit, and isolated from any drive electronics, and have very low switching and settling times. They're also small, which helps build the kind of dense circuitry that is useful for sustaining and improving the functional density of IC testers.

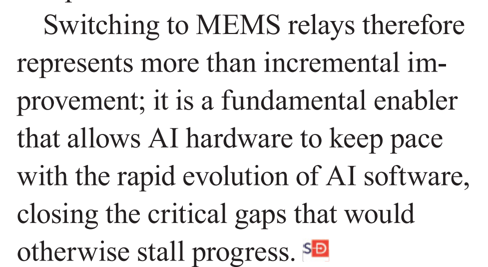
Switching to MEMS relays provides a transformative improvement, removing critical bottlenecks in testing. By speeding up signal routing and improving reliability, these relays help ensure that testing keeps pace with the rapid evolution of advanced AI devices rather than becoming a limiting factor.

What makes MEMS relays essential, rather than optional, is how comprehensively they address the bottlenecks that have been slowing AI chip testing. Traditional mechanical switch matrices impose unavoidable delays due to slow actuation, settling times, and vibration-induced noise. These delays accumulate, limiting throughput in high-volume AI chip testing, and can cascade into shortages and higher costs for other critical components like

AI chips such as high-performance GPUs are located in a variety of environments, including data center racks.

DRAM and cell phone chips. MEMS relays, by contrast, remove these mechanical and electrical limitations, enabling almost instantaneous switching with negligible signal interference.

In effect, MEMS is the only current solution that scales testing speed, density, and reliability simultaneously to match the pace of AI chip evolution. Without it, production testing would remain a choke point, slowing the deployment of increasingly complex AI models despite advances in fab capacity and design. By adopting MEMS-based switches, the industry not only overcomes the physical and bottlenecks in IC testers but also ensures that AI infrastructure—data centers, chips, and memory—can continue scaling efficiently to meet growing computational demands.



Switching to MEMS relays therefore represents more than incremental improvement; it is a fundamental enabler that allows AI hardware to keep pace with the rapid evolution of AI software, closing the critical gaps that would otherwise stall progress.

Metrology

Continued from page 37

closed-loop control system ensures precise bonding alignment and prevents defects caused by overpressure. Additionally, handheld diagnostic tools with integrated charge amplifiers enable engineers to validate process parameter verification and quality control processes anywhere.

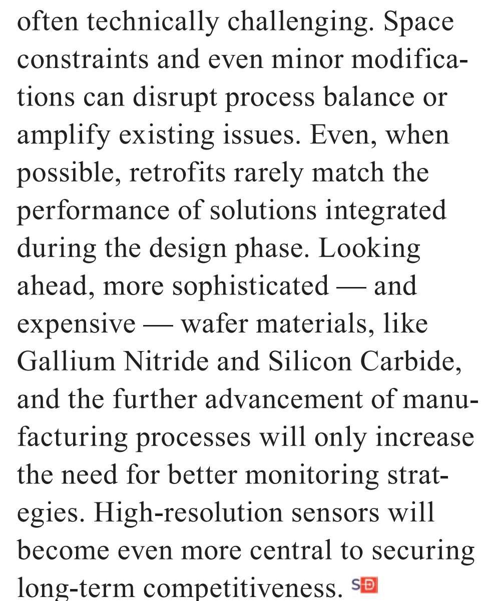
Sensor placement: there is no one-size-fits-all solution

Designing equipment for heterogeneous 3D integration means more than simply selecting the appropriate sensing technology. It also requires placing sensors exactly where they deliver maximum insight. The ideal measurement location may not offer the environmental conditions needed for long sensor life. For example, while piezoelectric sensors can tolerate high temperatures, they still have operational limits. In certain applications, it may be necessary to use a more heat-resistant sensor or adjust the placement to ensure equipment protection and reliability.

Kistler works closely with semiconductor equipment manufacturers to identify the most sensitive points within their systems, ensuring that force, pressure, and vibration sensors are embedded where they capture the most informative data. Further services such as calibration, and system validation during equipment build, ensure long-term accuracy and reliability.

Taking bold action: securing yield and strategic advantage in 3D integration

Delaying investments in advanced sensor systems and smart manufacturing technologies presents risks for Semiconductor Equipment Manufacturer. As product requirements evolve — such as miniaturization or the use of new materials — process stability can be compromised. Retrofitting sensors into completed machines is often technically challenging. Space constraints and even minor modifications can disrupt process balance or amplify existing issues. Even, when possible, retrofits rarely match the performance of solutions integrated during the design phase. Looking ahead, more sophisticated — and expensive — wafer materials, like Gallium Nitride and Silicon Carbide, and the further advancement of manufacturing processes will only increase the need for better monitoring strategies. High-resolution sensors will become even more central to securing long-term competitiveness.



About the author

As the Industry Lead Semiconductor at Kistler, Robert Hillinger drives the development of industry sensor measurement solutions and supports sales colleagues and customers globally to help them achieve higher yield and process reliability.