CASE STUDY

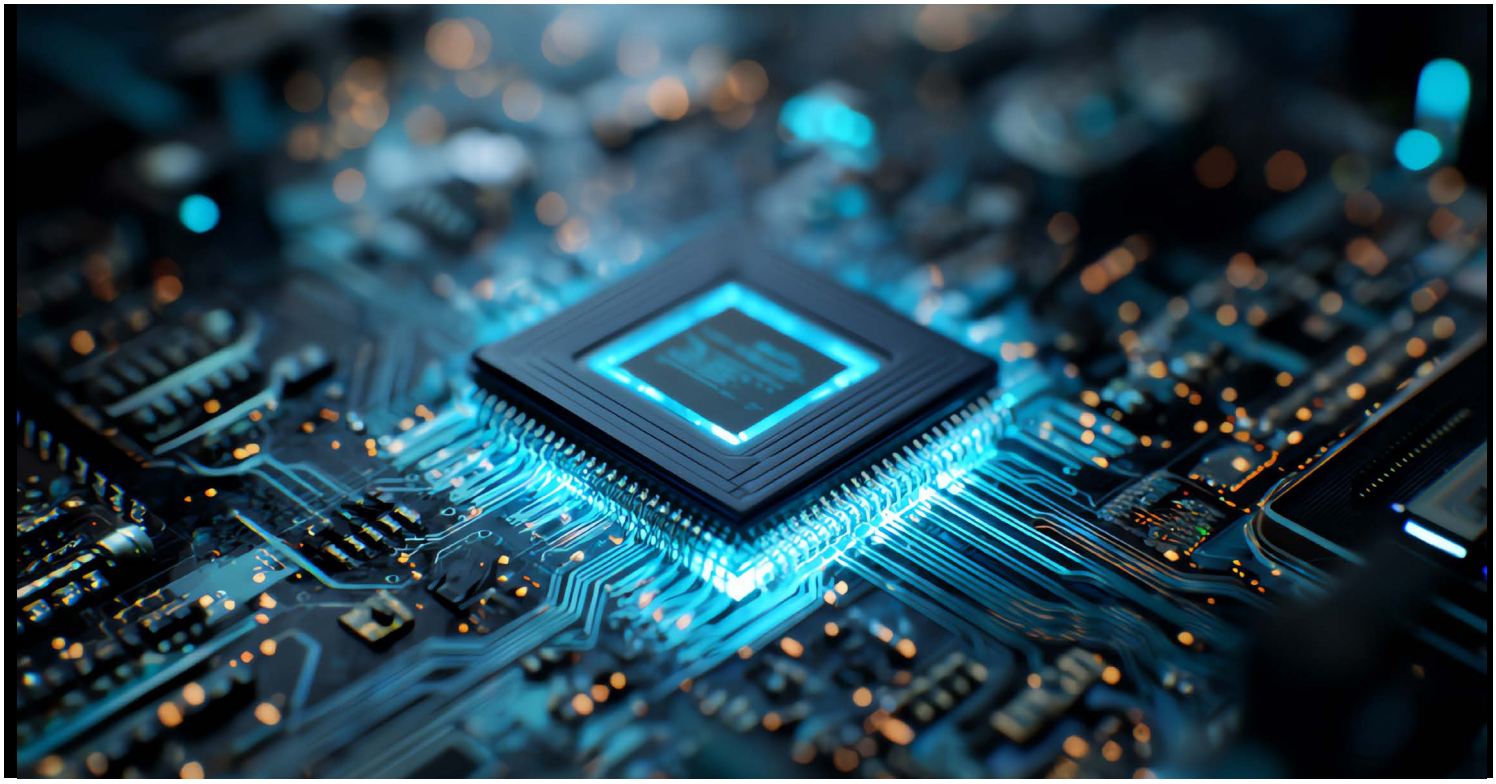
menlomicro

Addressing the Challenges for High Performance GPU Testing

Co-authors:

NVIDIA | Joe Sarmiento – Director Test Engineering
Mahmut Yilmaz – Director DFX Methodology
Derek Lee – Test Engineering Manager

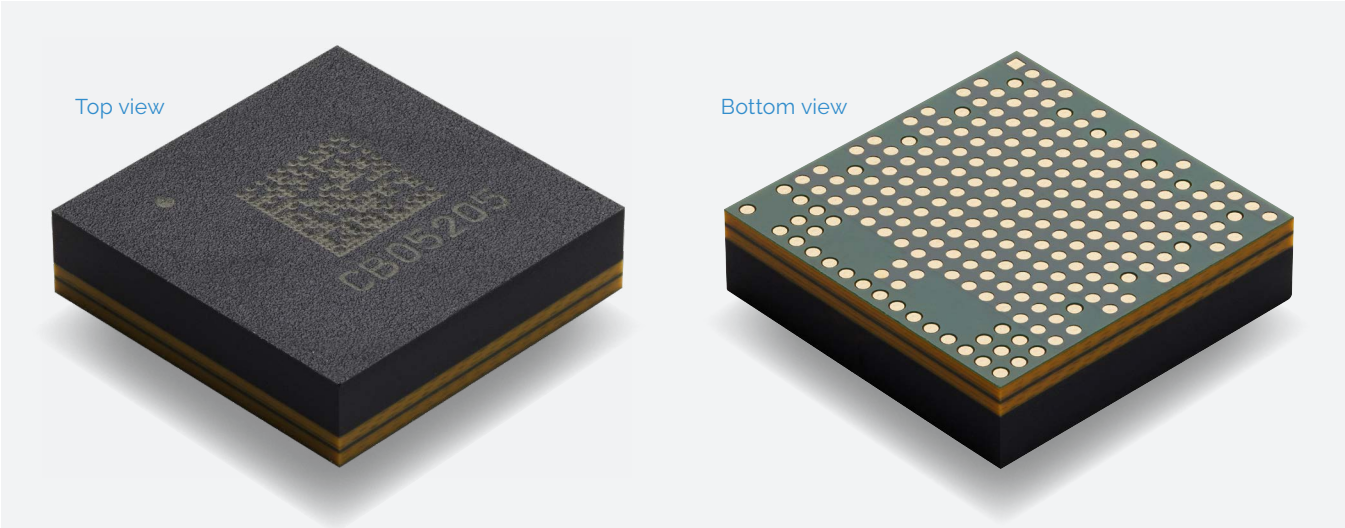
Menlo Micro | Harry Liu – Director, RF Product Marketing
Stewart Yang – Sr. Principal System Application Engineer



Introduction

The rapid advancement of artificial intelligence (AI) and machine learning (ML) is driving unprecedented demand for computing power, with GPUs (Graphics Processing Units) now serving as the backbone of modern high-performance computing. From accelerating large language models to powering scientific simulations and autonomous systems, GPUs deliver the extraordinary parallel processing performance today's workloads demand.

In tandem with the rise of AI, high-speed interconnect technologies such as PCI Express (PCIe) are evolving at a rapid pace. Each new generation—Gen4, Gen5, Gen6 and now Gen7—has doubled bandwidth, dramatically increasing the complexity of both system design and production validation. As GPUs transition from PCIe Gen5 to Gen6 and beyond, the demand for robust, scalable, and cost-effective production test strategies has become more critical than ever.



MM5620
High-Performance Loopback Switch

Production Test Challenges

Testing high-performance GPUs presents a unique set of challenges, particularly as interface speeds, power demands, and system complexity increase. These challenges span electrical, thermal, and system-level domains:

Signal Integrity (SI):

With PCIe Gen5/6 pushing data rates up to 32 GT/s and 64 GT/s, maintaining SI across test sockets, probe cards, and interconnects is exceptionally difficult. Even minor impedance mismatches or crosstalk can cause eye closure and bit errors, making it essential to replicate real-world link conditions with high fidelity.

Power Integrity (PI):

Modern GPUs draw hundreds of watts, often through multiple power rails. Test setups must provide high-current delivery with minimal voltage droop, ripple, or ground bounce while avoiding excessive thermal buildup.

Test Speed and Throughput:

Long test times impact production cost. Validating PCIe training, memory access, and firmware boot must be optimized using parallel testing, reduced pattern depth, and built-in diagnostics to maintain throughput targets.

Thermal Management:

GPUs generate significant heat even under test. Traditional handlers struggle to dissipate this effectively, prompting the use of active or liquid cooling to prevent throttling and maintain consistent results.

Scalability and Cost:

As AI adoption accelerates, GPU production volumes continue to rise. Scaling test capacity without escalating cost requires new approaches, such as loopback testing, test-on-board (ToB) strategies, and modular test architectures.

Versatility:

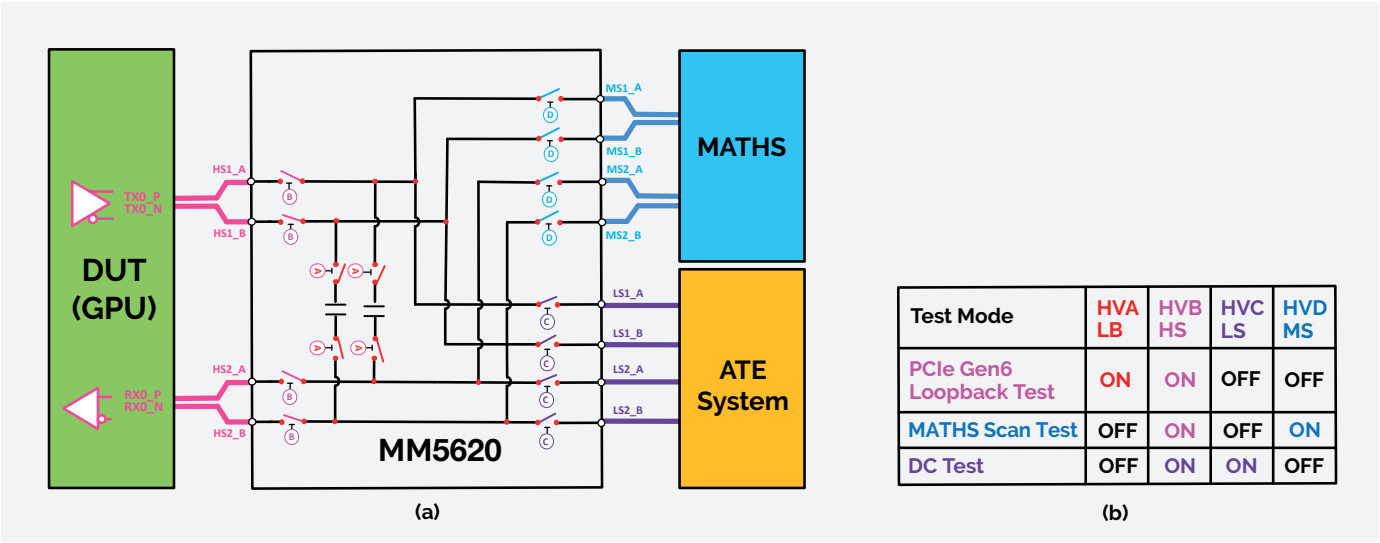
Test solutions must support a variety of SKUs, interfaces (PCIe, NVLink, HBM), and form factors (PCIe cards, SXM modules) while minimizing changeover complexity and downtime.

Customized, High Performance Loopback Switches

Modern GPU production demands test strategies that can keep pace with rapidly escalating performance requirements. To meet these critical challenges, three distinct signal paths must be supported:

- At-speed loopback test in PCIe Gen6
- PCIe Gen5 based scan test using MATHS (Mechanism to Access Test-Data over High-Speed Link) methodology
- DC measurements, including drive strength and sensitivity.

The differential SP3T switch topology is required to enable these three test paths while maximizing coverage and minimizing complexity. By collapsing multiple test modes into a single signal path, advanced switches simplify test hardware design and directly reduce cost of test. Figure 1 illustrates how a single MM5620 device supports these test modes.



Test Mode	HVA LB	HVB HS	HVC LS	HVD MS
PCIe Gen6 Loopback Test	ON	ON	OFF	OFF
MATHS Scan Test	OFF	ON	OFF	ON
DC Test	OFF	ON	ON	OFF

Figure 1. (a) High-level ATE Configuration Using a Single MM5620 Device (b) Truth Table Showing the Critical Test Modes

The development of high-performance loopback switches such as the MM5620 Loopback switch from Menlo Microsystems has enabled the adoption of innovative test methodologies like MATHS and BOB (Bolt-On-Box). These methodologies are transforming GPU production testing by leveraging existing high-speed functional interfaces, primarily PCIe, to conduct structural tests with greater bandwidth and lower cost than traditional ATE systems.

At PCIe Gen6 speeds, where preserving signal integrity becomes critical, the MM5620 provides the low-loss, high-linearity performance necessary to ensure reliable loopback validation. In addition, its versatility allows a single device to support multiple GPU SKUs and platform interfaces, reducing the need for SKU-specific test hardware and accelerating time-to-market.

MATHS, in particular, uses the native PCIe interface of the GPU to access scan and BIST logic through a dedicated MATHS DMA controller and sequencer, effectively bypassing the need for additional I/O or custom tester hardware. Its high portability and minimal reliance on pin-based interfaces allow it to be applied across ATE, SLT, BLT, and even in-field testing environments. Its architecture incorporates robust isolation zones, secure boot software, encrypted test data protocols, and asynchronous clock domain crossings to ensure reliability, safety, and security at scale.

BOB further complements MATHS by integrating a PC-based test controller with the ATE load board. This hybrid configuration allows high-speed PCIe communication directly between the GPU and BOB's system memory, offloading pattern storage from the ATE and enabling a unified test flow that supports both parametric and functional validation. The reuse of ATE load boards across test modes (PB and MATHS) significantly reduces cost while enhancing flexibility and test correlation.

Together, MATHS and BOB exemplify how next-generation switching technology meets the growing test challenges of demanding GPU performance. By collapsing multiple test modes into a single signal path, they streamline complex validation workflows while reducing hardware overhead. At PCIe Gen6 speeds, switches such as the MM5620 preserve signal integrity across loopback and scan paths, ensuring performance measurements remain accurate and repeatable.

Equally important, the versatility of this approach allows the same switch topology to support multiple GPU SKUs and platform interfaces, enabling OEMs and test providers to scale without re-engineering their hardware base. The result is a test strategy that not only accelerates development and production ramps, but also lowers total cost of test while maintaining confidence in performance at the highest data rates.

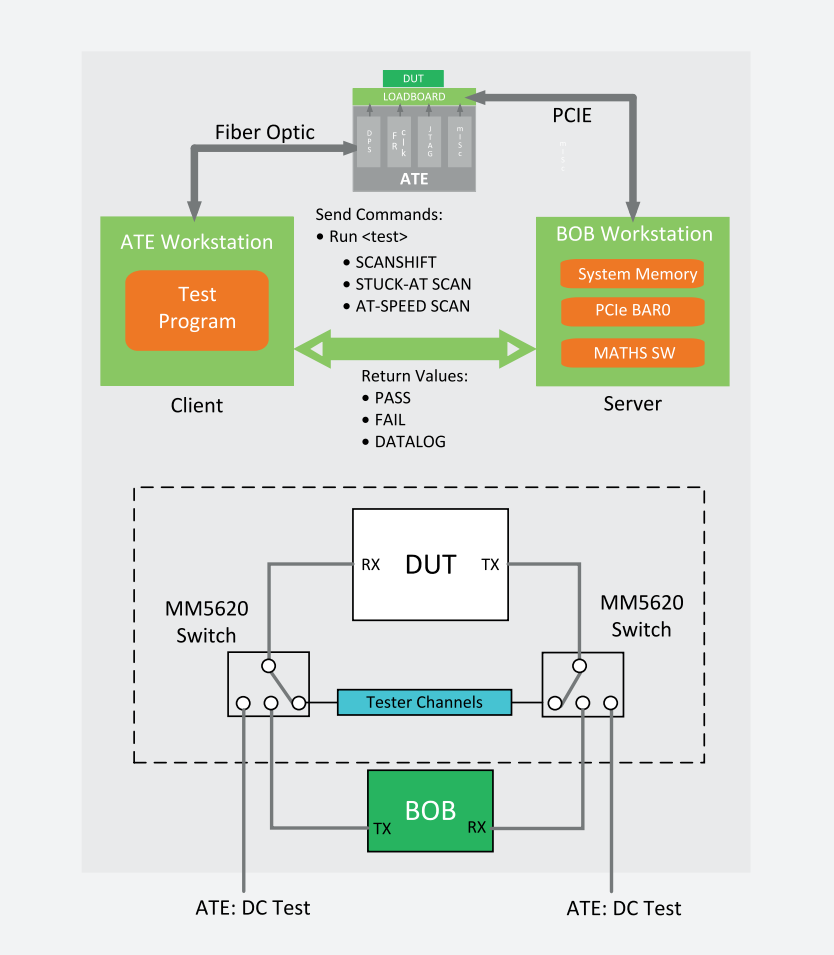


Figure 2. BOB Configuration and its Interaction with the ATE Loadboard

This convergence of structural and functional validation—enabled by robust loopback switches—provides a blueprint for how the industry can continue to test increasingly complex GPUs in a scalable, secure, and cost-effective way.

Figure 2 illustrates the BOB configuration and how it integrates with MATHS to deliver scalable, high-bandwidth testing.

Performance Results and Impact

An actual PCIe Gen 6 external loopback test passed on a production GPU load board with the switches with very healthy 20% frequency margin. The maximum data rate is close to 70 Gbps (PAM4). The real time scope showed clean eye openings when equalization was applied. Menlo MM5620's electrical performance is robust, and it has been selected and successfully used to test multiple GPU load boards.

Figure 3 shows the 64 Gbps, PAM4 eye-diagram performance on two MM5620 loopback lanes.

"As NVIDIA drives the future of accelerated computing, the demands on production test are evolving just as rapidly. Our collaboration with NVIDIA demonstrates how advanced switching can unlock new test methodologies – combining loopback, MATHS, and DC validation at PCIe Gen6 speeds. Together, we are setting a new benchmark for scalable cost-effective test strategies that will support the next generation of AI and data center platforms."

CHRIS GIOVANNIELLO
SVP OF HSD & RF BU
MENLO MICRO

(a) (b)

Figure 3. (a) Lane 1 (b) Lane 9

MATHS and BOB-enabled platforms have demonstrated substantial benefits in GPU testing, including up to 90% test-time reduction for pin-limited designs. In other cases, around 30% test-time savings were achieved even when ATPG tool constraints limited bandwidth utilization. These improvements free up test cycles for more advanced fault models like cell-aware and small-delay defect patterns.

Performance results also highlight strong correlation across platforms. Vmin and Fmax measurements between ATE, SLT, and in-field tests showed high consistency, enabling better debug and root-cause analysis of marginal failures. For RMA and field-return scenarios, MATHS allows non-destructive structural testing directly on the board—avoiding the need for desoldering and rework, thus accelerating time-to-resolution.

By using loopback switches and standard PCIe interfaces, these methods also allow greater scalability as device complexity and interface speeds grow. MATHS makes full use of available PCIe bandwidth, with Gen5 supporting over 50 GB/s duplex data transfer across 16 lanes, eliminating test-time bottlenecks imposed by legacy pin-based approaches.

High-Performance GPUs Are Located in a Variety of Environments, including Data Center Racks.

Conclusion and Future Innovation

As AI continues to push the boundaries of computing performance, GPU test methodologies must evolve to stay ahead of escalating interface speeds, thermal loads, and packaging complexity. The industry is entering a new era of chiplet-based architectures, CXL interconnects, and co-packaged optics, all of which will demand new test paradigms that are both scalable and efficient.

The MATHS architecture, with its ability to deliver structural tests via high-speed PCIe links, provides a unified, scalable solution that addresses these evolving demands. Its flexibility across ATE, SLT, BLT, and in-field testing enables deep coverage while reducing the dependency on expensive vector-based ATE systems. Combined with the BOB configuration, which retrofits existing testers with high-speed data paths and cost-effective memory offload, MATHS ensures production test strategies can adapt to future silicon complexity without requiring wholesale infrastructure changes.

Key Directions for Future Innovation

- **Higher-Bandwidth Loopback Switch Designs:** Enhancing loopback switches to support PCIe Gen7 at 128 Gbps and next-generation networking at 224 Gbps Ethernet.
- **Differential SPDT Switch:** Advancing a differential SPDT switch operating up to 70 GHz on a test chip, validated through insertion loss and return loss simulations versus measurements (see Figure 4), along with eye diagram simulations (see Figure 5) demonstrating clear signal integrity up to 200 Gbps.
- **Extended Loopback Testing:** Expanding loopback testing methodologies to multi-die and 2.5D/3D stacked devices.
- **AI-Assisted Test Data Analysis:** Leveraging AI for predictive failure detection and adaptive test optimization.
- **Dynamic Test Orchestration:** Coordinating distributed test nodes to enable chiplet-level validation.
- **Expanded Interface Support:** Extending coverage beyond PCIe to interfaces such as CXL, supporting heterogeneous system architectures.
- **Unified Test Framework:** Utilizing MATHS' secure and portable software framework to standardize structural test execution across platforms, from silicon bring-up to field diagnostics.

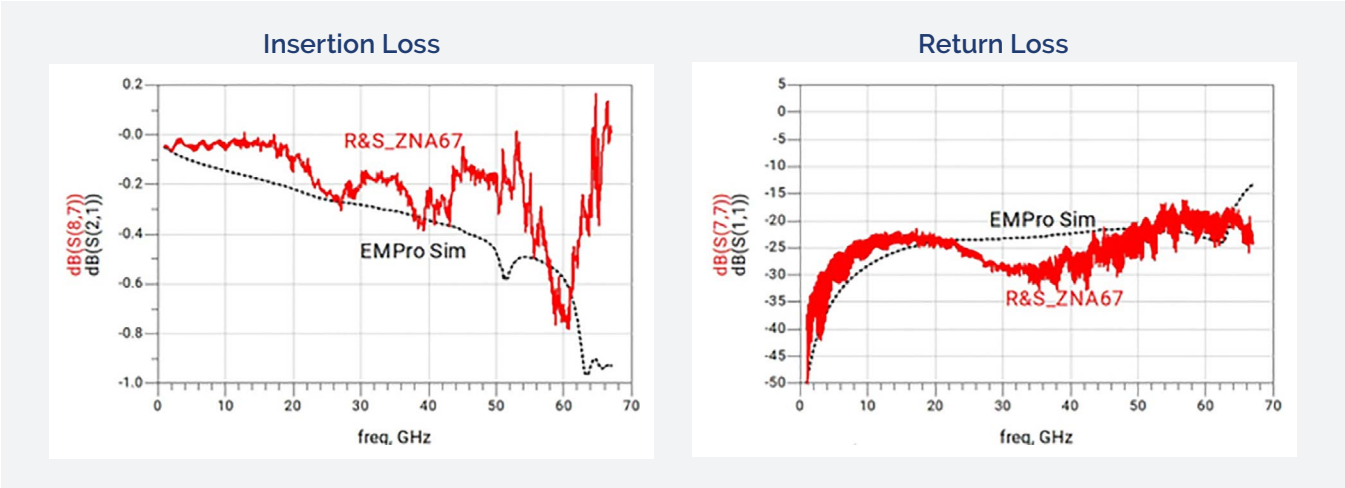


Figure 4. Measurement vs Simulation on a Differential SPDT Test Chip for Insertion Loss and Return Loss up to 70 Ghz [1]

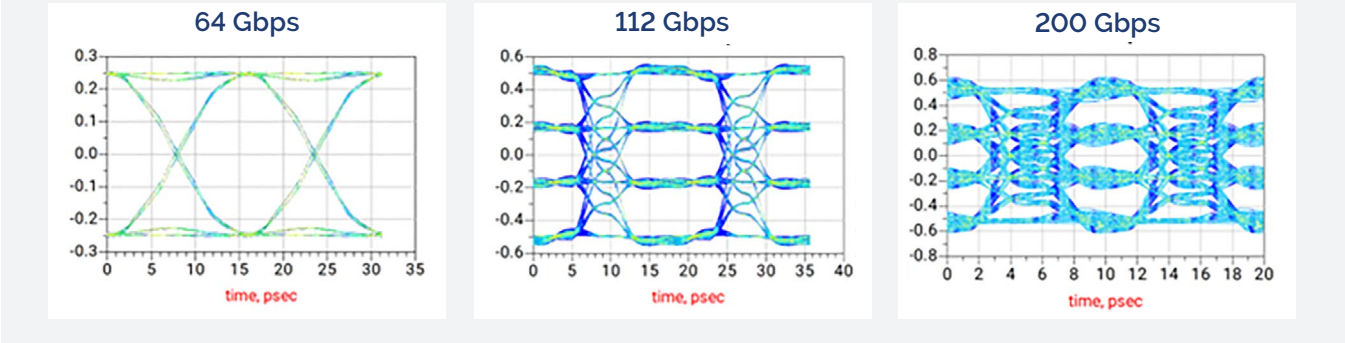


Figure 5. Eye Diagram Simulation without Equalization up to 200 Gbps on a Differential SPDT Test Chip [1]

By extending test technology beyond PCIe Gen6, the feasibility of operating at PCIe Gen7 speeds (128 Gbps and beyond) has been demonstrated, paving the way for next-generation system validation. The 70 GHz Differential SPDT switch design, supported by early test chip simulation and measurement data, highlights how insertion loss, return loss, and eye diagram analysis can ensure robust performance at extreme bandwidths, emphasizing the collective progress being made toward scalable and reliable production test strategies for emerging technologies.

In summary, customized loopback switch architectures such as those based on MATHS and BOB, anchored by advanced hardware like the MM5620, will remain central to addressing the production test challenges of high-performance GPUs. With their ability to operate at system speeds, provide deep diagnostic visibility, and scale across test environments, they are well-positioned to enable the next generation of computing platforms. Continued innovation in this space will be essential to meet the growing demands of AI, data center, and high-performance computing markets—efficiently, reliably, and at scale.

Reference:

[1] Xu Zhu, Nicholas Yost and Stewart Yang, Menlo Micro "Edge Coupled DC to 60 GHz Differential SPDT RF MEMS Switch for High-Speed Digital Applications," IEEE/MTT-S International Microwave Symposium (IMS 2025), San Franscisco, CA, USA: https://www.researchgate.net/publication/394633368_Edge_Coupled_DC-60GHz_Differential_SPDT_MEMS_Switch_for_High-Speed_Digital_Applications

menlomicro

About Menlo Micro

Menlo Micro is setting a new standard for switches in modern systems with the Ideal Switch. Legacy switches bottleneck the modern world, limiting performance, efficiency, reliability, and scalability in critical systems. Menlo Micro solves this with the Ideal Switch - the first disruptive switching technology in over 30 years and the only platform scalable across both power and frequency domains.

From milliwatts to megawatts and DC to mmWave, the Ideal Switch provides customers with a single switch platform that simplifies designs, reduces cost, and enables a wider range of applications without compromise. Built on a proprietary chip-scale platform, it enables smaller, lighter, faster, more reliable, and energy-efficient systems, while lowering total cost of ownership in high-speed digital, broadband RF, and power applications.

Menlo Micro defines modern switching across industries including AI/HPC, test & measurement, aerospace, defense, industrial automation, and telecommunications. Its technology accelerates AI GPU testing, enables beamforming in satellite communications, improves filtering in mobile radios, cuts energy use in factory automation, enhances fault detection in energy infrastructure, and more.

For more information,
visit www.menlomicro.com