International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

RESHORING GPU PRODUCTION: TESTING STRATEGY ADAPTATIONS FOR U.S.-BASED FACTORIES

Karan Lulla

Senior Board Test Engineer, NVIDIA, SantaClara, CA, USA

karanvijaylulla08@gmail.com

Orcid: 0009-0007-7491-4138

Abstract

With the reshoring of the GPU manufacturing in East Asian centers to the factories in the USA, the national manufacturers are now urgently in need to construct high-throughput and versatile GPU testing services capable of sustaining HPC services and hi-tech community of interests like the mission-focused applications. This study examines how testing strategies need to change in the presence of reshoring to semiconductor manufacturing and practical applications of its concepts with the design of scalable and cost-effective GPU testing automation and AI-driven validation processes. Through a mixed methodology, comprising both expert interviews and a comparative literature study of the best-practice facilities in the world, the study has determined that modular test-line architectures, digital twins, and real-time analytics are the essential enablers to enhance yield and minimize defect escape and time-to-market. The case study of Intel Arizona and Ohio fabs shows how close integration of validation and close cooperation with vendors of Automated Test Equipment can operationalize these capabilities, and promote energy efficiency and closed-loop water systems. In the future, the framework outlines ways of leading to quantum conscious GPU testing, a remote autonomous engineering model, and more resilient U.S HPC manufacturing ecosystems. Such insight can be used by policymakers, foundry planners, and equipment manufacturers to plan the next ten years of semiconductor development in the United States.

Keywords; Quantum-aware testing, Automated Test Equipment (ATE), Semiconductor Reshoring, GPU Testing Automation, AI-Driven Validation, HPC Manufacturing, Sustainability.

1. Introduction to Reshoring and the GPU Market Concept

The Graphics Processing Unit (GPU) has become a bedrock aspect of modern computing, carrying out everything from analytics to AI and machine learning, to gaming and visual rendering. In their commercial advent, late in the 1990s, GPUs have matured quickly as both complex technologies and vital to computing. Industry leaders such as NVIDIA, AMD, and Intel have created innovation by constantly expanding processing cores, bandwidths, and computing power. Although the intellectual property and architectures of GPUs were most developed in the United States, the high-volume fabrication, packaging, and testing shifted to the Asian ecosystem of Taiwan, South Korea, and China (Hoi-Chun Hung, 2024). This offshoring provided economies of scale but had the adverse consequence of concentrating critical test capacity in a handful of areas, with U.S. firms becoming vulnerable to geopolitical

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

shocks, export controls, and logistics bottlenecks as GPU loads grow across consumer-processing servers into the cloud, critical infrastructure, and defense processing.

Reshoring – the strategic migration toward a return of manufacturing back to local soil – has flourished throughout the technology space, especially in semiconductors. The disruptions of the supply chain occasioned by the COVID-19 pandemic and resulting disruptions in East Asian fabs and test houses demonstrated that reliance on a few fabs and test houses would paralyze GPU roll-outs, limit cloud capacity, and delay AI and HPC deployments. Addressing this, recent U.S industrial policy (most usually the Chips and Science Act and related reshoring subsidies) now focuses tens of billions of dollars of exposure into domestic fabrication, advanced packaging, and test capacity in response to the need to avoid being exposed to politically insecure regions (Rinehart & Kirchhoff, 2024; Pinto, 2023).

For companies, reshoring thus ceased being a symbolic or patriotic act to become an operational priority that facilitates a more comfortable time over intellectual property, a short manufacturing cycle, and direct integration between the design, manufacturing, and test engineering teams.

The success of the GPU production pipeline requires testing functionality. The occurrence of defects, local heating, and signal integrity issues is high as modern GPUs consist of billions of transistors, and the testing process should ensure that all units are assigned to meet performance, power, and cost of reliability demands before deployment. In offshore designs, this was concentrated within giant test centers with automated test equipment (ATE) and burn-in test facilities; as the manufacturing process is moved offshore, the U.S. fab facilities must reassemble and modernize such capacity domestically.

In-house fabrics are now required to have the state-of-the-art test protocols capable of supporting 3D stacking, chiplets, and AI accelerators, and meeting regulatory, export-control, and customer-specific validation needs using secure, locally federated data and workflow infrastructure. The development of such a domestic testing environment, and the experienced labor force to run it, has become a core issue that will define whether reshored GPU manufacturing can deliver competitive results, time to market, and reliable usage in the long run.

2. Drivers behind Reshoring GPU Production to the US.

The strategic imperative to restore GPU production back home in the US is influenced by a converging set of disruptive events, policy changes, and technological advances. Detailed inspection shows that the choice of localizing the manufacture and testing of high-performance graphics processing units is based on three core aspects. Supply chain imperfections exposed within the COVID-19 crisis, enhanced national security, legislative concerns, and a paradigm shift in cost structure occasioned by automation.

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 1: **Key Drivers behind Reshoring GPU Testing to the U.S.**

Driver	**Details**
Supply Chain Disruption	COVID-19 exposed fragility in Asian-based GPU testing infrastructure, leading to halted production and delayed market readiness.
National Security & Legislation	CHIPS and Science Act ($52B funding), national defense dependencies on GPU systems, and emphasis on domestic sovereignty.
Cost Shifts & Automation	Advanced robotics and AI tools have reduced the operational cost disadvantage of U.S. manufacturing.

2.1 Supply Chain Disruption and Pandemic Fallout

The supply chain system of the world manufacturing in the COVID-19 pandemic was pivotal, exposing the levels of reliance of the producers of the GPUs on the just-in-time system of supply based on the East Asian production and testing centers. Lockdowns in Taiwan, China, and Malaysia led to wafer fabs and backend assembly plants, disrupting the production timeline of GPUs, trickling down to a lack of consumer gaming devices, dedicated workstations, and AI-driven data centres.

Due to long power-efficiency tests, thermal characterization, and error-correction tests, any interruption at centralized test houses soon turned into lost launch schedules and performance liability. The extent of literature concerning transilience within supply chains indicates that these networks of geographically concentrated testing are a systemic weakness, particularly of highly valuable high-complexity parts such as GPUs (Gatenholm & Halldorsson, 2023).

More recent articles have suggested that, in addition to geographic diversification and reshoring, highly automated, digitally instrumented test lines can be used to increase capacity with little to no extra effort (Karwa, 2024; Chavan, 2023). AI-enhanced scheduling and digital twins paired with a restoration of production and testing to the U.S. will allow companies to have vertically integrated and highly automation-intensive ecosystems that can absorb geopolitical or biological shocks and maintain GPU pipelines over extended periods.

Figure 1 illustrates the thematic distribution of literature and stakeholder assessments regarding supply chain impacts during the COVID-19 pandemic. It shows that threats and combined threat-opportunity dynamics dominate GPU production discourse in light of reshoring concerns.

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

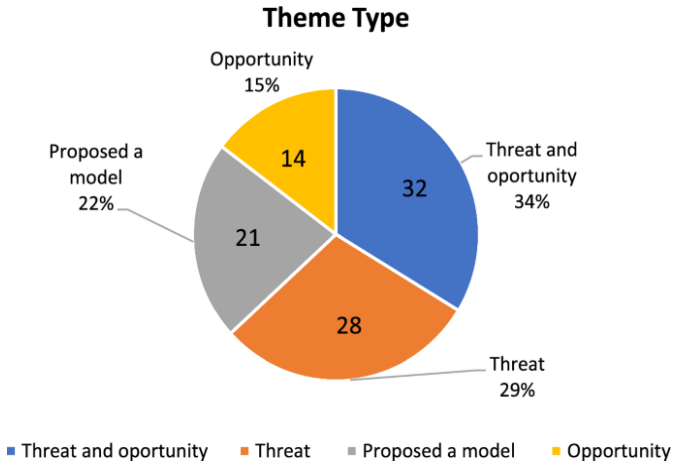

Figure 1: ***The impact of COVID-19 on supply chains***

2.2 National Security and Legislative Initiatives

With a background in rising geopolitical agitations, the return of industrial policy in the United States has made national security part of the reshoring narrative. Now that advanced GPUs support artificial intelligence training, military simulation, and autonomous systems, there is increased sensitivity concerning the strategic vulnerabilities of foreign dependence. The existence of adversarial control over GPU supply chains has brought about coordinated legislative and executive actions. The CHIPS and Science Act of 2022 is most prominent, which makes over $52 billion of funding available for domestic semiconductor research, development, and manufacturing capacity (Rinehart & Kirchhoff, 2024).

This legislation specifically values high-end packaging and testing, which are traditionally behind in the US compared to East Asian countries. Some money is reserved for facilities that run GPU testing and reliability analysis, with incentives and subsidies for firms opening such plants on US soil. Outside of funding, the legislation changes how government agencies work together and fosters public-private partnerships to hasten workforce preparation in core GPU manufacturing positions. National laboratories and engineering universities are being roped in to respond to the technical call of this reshoring effort through niche training programs and by supporting advanced materials research.

2.3 Integrated Cost Shifts, Automation, and Supply Chain Resilience

Labor and infrastructure expenses have traditionally been seen to deter the reshoring of semiconductor activities from offshore to the U.S., but the changes in robotics, artificial intelligence, and process automation are redefining this equation. AI-powered diagnostics on high-throughput automated test equipment will have a soft and, in power-envelope characterization and signal-integrity screening, a dustbin executed at significantly less marginal cost and with much less operator involvement, closing the cost difference with the Asian locations (Karwa, 2024).

In automated test systems, which are typically mediated by microservice-style structures, modularity and scalability provide an opportunity to add or reassign capacity without requiring a complexity increment in proportion (Chavan, 2023). Meanwhile, digitization-twin and

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

predictive-modeling applications model complete runtimes of a GPU, maximizing shortening of retesting, decreasing time-to-market. This literature indicates that identical automation technologies that enhance unit economics form the basis of supply chain resilience: identical technologies can make more compact, spatially more dispersed, onshore test clusters economical, reduce reliance on several offshore mega-facilities, and still generate competitive yields and throughput.

3. Technical Challenges in Domestic GPU Testing

Given the United States' resourcing efforts on GPU production domestically, the question of how to calibrate its testing strategies to match those that apply in the firms' home territories is multifarious in technical terms. These challenges come from infrastructural weakness, high-throughput requirements, and high variability within current GPU architectures (Navaux et al., 2023). Failure to address these serious pain points may cause the reshoring effort to fail to achieve scale production of competitive high-performance graphics processing units.

Table 2: ***Technical Challenges in Domestic GPU Testing***

Challenge	Description
Infrastructure Gaps	Lack of cohesive domestic semiconductor ecosystems results in delays and inefficiencies in integrating fabrication, packaging, and testing.
High-Throughput Calibration	Need for real-time feedback and tailored test hardware for large volumes of GPUs; many U.S. plants are not yet optimized for such requirements.
Evolving GPU Architecture Support	GPUs with AI cores, MCMs, and ray tracing units demand hybrid testing protocols that U.S. infrastructure is still scaling toward.

3.1 Infrastructure Gaps in the U.S.

The most severe barriers to U.S.-based GPU testing involve the disassembled nature of domestic semiconductor infrastructure. Whereas highly advanced Asian countries have their fabrication, packaging, and testing highly integrated within regional ecosystems, the U.S. lacks cohesive supply chain linkages. Such misalignment causes logistical inefficiencies and delays, as wafers made in one plant may be delivered across states or re-exported to be tested or assembled elsewhere. More importantly, critical supply dependencies, including those for photolithography tools, advanced test equipment, and high-purity chemicals, are still substantially offshore.

These gaps subvert the end-to-end integration needed for just-in-time tests and calibrations, compelling domestic entities to move to complex scheduling and multi-vendor coordination. This is particularly troublesome considering the accuracy that must be present in GPU test cycles, as latency between producing and validating quality can cause backlogs and delays in

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

reaching time-to-market (Mandya Channegowda, 2022). A related piece of work, dynamic network optimization, indirectly addresses this issue as it demonstrates how inefficiencies with system integration (if it considers natural language processing networks) can degrade performance and increase latency in real-time operations (Tache et al., 2024). U.S. GPU production is hampered when the ecosystem is not synchronized between design, fabrication, and test functions, needing significant rework of operational pipelines.

3.2 Calibration for High-Throughput Testing

GPU testing is inherently resource-intensive. Each unit has to be validated through extensive validation phases (functional testing to ensure the arithmetic logic units, memory interfaces, and core shading through functional testing, thermal characterizations to ensure that the system is most optimized to handle high workloads, and endurance testing to ensure that the system is most usable under prolonged usage). Such procedures must scale to handle the high volume of GPUs necessary at the data centers, gaming markets, and AI research fields. Domestic factories have huge problems configuring test floors to accommodate such throughput. Classic ATE platforms will also have to be redesigned for real-time performance feed-in, while load boards and test sockets will also need tailoring for the newest form factors of GPUs (Pandit, 2022). Not only are these hardware adaptations very expensive, but they also need an experienced workforce, which is scarce in the U.S. semiconductor labor market.

In addition, calibrating test algorithms and environmental parameters to conform to local standards, ambient, and facility-specific variables further complicates the arrangement. The adoption of predictive analytics and digital twins for simulation-driven calibration is still very much at its nascence in many U.S. plants and has resulted in various inconsistencies in test repeatability and yield measurements. Telematics systems research in fleet management, where calibration of real-time data and alignment of sensors are critical to asset tracking fidelity, a similar principle applies to GPU testing (Nyati's, 2018). High throughput settings require true-time tracking products that can respond dynamically to workload changes, variations in test temperature and voltage conditions, capacities for which most in-country operations are just beginning to incorporate.

3.3 Compatibility with Evolving GPU Architectures

The GPU market is a rapid innovation cycle; new architectures appear every 12-18 months and become progressively complex. These architectures integrate AI accelerators, core ray tracing, tensor units, and multi-chip modules (MCMs). Each component has its testing requirements that test the envelope of conventional ATE capabilities and the capabilities of test software plumbing. AI accelerators, in particular, necessitate validation at hardware logic (i.e., design validation) and software inference (test validation) layers (Mishra et al., 2023). This calls for a hybrid testing protocol integrating electrical tests and algorithmic checks—a problem made worse by the absence of a single-set testing standard for AI processing units.

Ray tracing cores require the pixel-level accuracy test under simulated rendering loads, which stretch the existing rigs that are not optimized for real-time graphics emulation. MCM-based designs also add complexity to testing, as inter-die communication and package-level thermal

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

dynamics add variability that eludes chip-level tests. This leads to an increased need for multi-domain test solutions that will require both (SiP) test protocols and board-level integration evaluations. The United States is rushing to develop or license test capability to support these new architectures in its current reshoring time frame.

However, compatibility testing risks lagging innovation cycles without continued investment in research, cross-industry cooperation, and toolchain standardization. The pattern identified in research on memory inference networks is once again enlightening. The more parameters and layers a system have, the more complex the test strategy must become to preserve functional integrity (Jyoti et al., 2024). In the context of the GPU, any additional architecture feature exposes the design to new failure modes, interconnect sensitivities, and timing dependencies, which an inclusive test plan must cover.

Figure 2 contrasts explicit versus unified memory management schemes in evolving GPU architectures, emphasizing the increasing integration between CPU and GPU cores and shared memory pools, a challenge for test frameworks not yet aligned with these hybrid configurations.

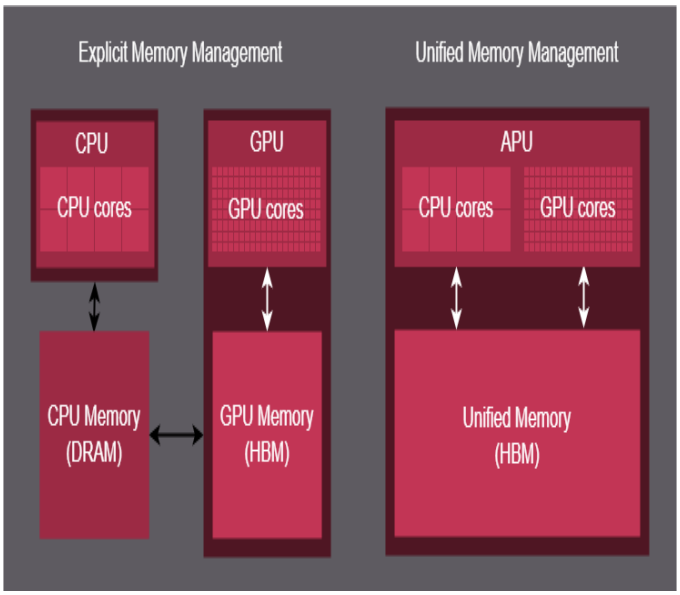

Figure 2: Unified memory

3.4 A Comparative View of Offshore versus Onshore GPU Testing Ecosystems

The offshore GPU testing ecosystems in Taiwan, South Korea, and China consist of highly concentrated groups of fabs, packaging facilities, and boards of high-volume test centers with mature automation, low unit testing costs, and yield rates usually range in the 95-97% range, as shown in Table 3 below. These locations enjoy local supplier networks and well-trained test workforces, but the geographic concentration of them puts global OEMs at risk of pandemic disruption, logistical bottlenecks, and geopolitical risk.

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 3: Comparative summary of offshore and onshore GPU testing ecosystems

Dimension / Feature	**Offshore GPU testing ecosystems (Taiwan, South Korea, China)**	**Emerging U.S. onshore GPU testing ecosystems**	**Overall implication for GPU supply chain**
Industrial clustering	Dense clusters of fabs, packaging houses, and high-volume test centers	Smaller and less mature clusters of fabs and test facilities	Offshore remains the main capacity center; onshore clusters are still developing but strategically important
Automation and cost	Supported by mature automation with low unit testing costs	Higher per-unit test costs despite increasing automation	Cost advantage still favors offshore, but automation helps narrow the gap for onshore testing
Yield performance	Yield rates often reported in the 95–97% range	Initially lower yields as new lines and teams ramp up	Offshore offers proven, high-yield operations; onshore must improve yields to be fully competitive
Supplier networks and workforce	Integrated local supplier networks and highly experienced test workforces	Building supplier bases and technical talent, with closer interaction with design and R&D teams	Offshore ecosystems deliver efficiency through experience; onshore gains from tighter engineering feedback loops
Risk exposure	Geographic concentration creates vulnerability to pandemics, logistics bottlenecks, and geopolitical tensions	Geographically diversified relative to Asia and aligned with U.S. national security and resilience goals	Combining offshore and onshore reduces single-region exposure and enhances resilience
IP, security, and control	Greater exposure of IP and sensitive designs across borders and foreign jurisdictions	Tighter IP control, data governance, and compliance with national security requirements	Onshore capacity is favored for security-critical and defense-related GPU testing workloads
Strategic role in reshoring	Represents the legacy, cost-optimized global testing model	Positioned as a complementary, sovereign capability	Literature frames future state as dual geography where automated U.S. hubs complement offshore capacity, creating a more

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Dimension / Feature	Offshore GPU testing ecosystems (Taiwan, South Korea, China)	Emerging U.S. onshore GPU testing ecosystems	Overall implication for GPU supply chain
		rather than a full replacement	diversified and shock-tolerant GPU supply chain

Emerging U.S. onshore ecosystems, in turn, can be depicted as less mature but with more strategic locations: domestic facilities initially demonstrate lower results and a higher cost of tests per unit, but they enjoy the benefits of co-location with design teams, better IP control, a reduction in feedback cycles, and aligning their goals to national security and resilience goals. The literature is increasingly conceptualizing reshoring as a transformation of an offshore single model that is cost optimized to a dual geography where highly automated U.S. test hubs augment, but are not identical to, offshore capacity and establish a more diversified and shock-tolerant global GPU supply chain.

4. Research Methodology

The study design used to investigate the testing strategy adaptation in the reshored production of GPUs is a mixed-method research approach, where qualitative and quantitative methods are integrated. The mixed approach is best applied to the semiconductor industry analysis, as the combination of the insights that can be described as rich and expert-based in their nature relates to organizational and policy decisions and their potential choices, and objective data on performance and benchmarks could be used to explain and prove the relevance of the approach in testing the strategies of graphics cards in particular (Cai et al., 2022). The research paper is structured around three main pillars, including data collection, comparative analysis, and the evaluation based on performance metrics, which all add to a holistic and evidence-based approach to learning more about GPU testing frameworks within a reshoring environment.

Figure 3 illustrates the core steps followed in the research methodology—ranging from topic selection to hypothesis building, data collection, and analytical interpretation—emphasizing a structured, evidence-based investigative approach.

Figure 3: Strategies and Models

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4.1 Data Collection Approach

To place the reshoring process and the accompanying testing transformation operations in perspective, a comprehensive view of this research, including both primary and secondary data sources, is used to collect data. The data were acquired through structured and semi-structured purposely selected stakeholders in the GPU manufacturing ecosystem in terms of direct data analysis. The experts received inclusion criteria that they had at least eight years of strong experience (on average) in the field of testing of GPUs or other related semiconductor products, they were currently or recently responsible for the field of testing, manufacturing, or policy decisions, and they had firsthand exposure to reshoring or dual-sourcing efforts.

These stakeholders included engineers in graphics card design of the major companies such as NVIDIA and AMD, managers of factories in the local manufacturing facilities that were new or expanding, and policy experts who were related to the U.S. semiconductor development initiatives. An interview of 17 interviews lasting 45-90 minutes in total was done using a typical semi-structured guide; participants were transcribed, taped, and coded systematically to enable comparison of patterns to be made in comparisons of the interviews and to confirm the quality and validity of the qualitative evaluation. These meetings covered technical adjustments in testing lines, decision-making in sourcing test equipment, software deployment for test automation, and the operational circumstances surrounding the offshoring of testing workflows.

Secondary sources, including white papers, peer-reviewed journals, trade publications, and industry reports, completed the analysis. Findings on the evolution of dual sourcing strategies in high-tech manufacturing were especially relevant. Their work highlighted the essence of integrated diversified logistics and testing capabilities to reduce operational risks, a theme that chimes well in reshoring situations.

The research helps to understand the use of data platforms, such as MongoDB, used to maintain testing data consistency, even in distributed factory ecosystems (Kang et al., 2016). The compromise between performance and reliability, a central topic in this study, fits the requirements of GPU testing environments requiring maximum data fidelity. Data triangulation was used to verify the observations made in these sources and ensure their consistency and dependability. This method also reduced the threat of bias from any individual perspective, and the research was able to properly capture the dynamic obstacles and solutions in U.S.-based GPU testing operations.

4.2 Comparative Analysis Techniques

The heart of the evaluation part of the research concerns a comparison of offshore and domestic GPU strategies in testing. This analysis followed a case-based and metric-driven perspective in identifying similarities and divergences in the eventual operational efficiencies, cost structures, and technological usages. Analytics of testing frameworks in offshore factories (which are mostly in Taiwan and South Korea) was done using documented standard operating procedures and vendor white papers. They were compared with other U.S. facilities, including Intel's Ocotillo campus in Arizona and the new fabs of startups and consortia, namely SkyWater

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Technology and Rapidus (Aguirre, 2024). The testing workflow was decomposed into stages, including initial wafer probing, functional validation, thermal profiling, and final performance comparison.

Special attention was given to how environmental controls, the level of automation, and digitalization varied between these sites. For example, offshore factories were known for possessing well-defined test automation pipelines with low operator intervention, whereas U.S.-based facilities are undergoing a transition towards this level and sometimes have difficulties integrating systems and standardizing processes. The comparative framework, which used a SWOT analysis model to further classify strengths and weaknesses, opportunities, and threats, adopted a SWONET business database design model. Offshore sites demonstrated superiority in test maturity and low operation cost, whereas domestic sites promised agility, real-time fault isolation, and co-location with design teams. These results were then backed up by findings from the literature on dual sourcing, which highlighted the importance of testing capabilities in various geographies to increase resilience (Goel & Bhramhabhatt, 2024).

4.3 Key Performance Metrics Assessed

The research has identified and monitored a set of key performance metrics (KPMs) to understand testing strategy adaptation in a measurable and results-oriented way. These metrics were chosen based on industry best practices and agreement within the panel of interviewers (Bukhari et al., 2019). The main parameter considered was yield rate, which is the percentage of fully functioning GPUs produced with respect to the total volume tested. This metric is an immediate indicator of the efficiency of both manufacturing and controlling processes. Offshore yield rates for leading-edge GPUS hover at 95–97% levels. Early domestic operations reported in the range of 88–92%, suggesting scope for improvement in process tuning.

Another important metric was failure analysis turnaround time (FATT), which measures the time needed to pinpoint, analyze, and act on test failure issues. Offshore facilities had average FATTs of 24-36 hours, with very good defect libraries and mature teams. Elsewhere, data hubs based on centralized platforms such as MongoDB have been implemented in sites hosted in the U. S, making the data index committed and failure signature matched quicker. MongoDB schema flexibility enables real-time manufacturing analytics, particularly in GPU testing, where swift action with thermal or electrical anomalies is essential (Dhanagari, 2024).

Another important metric that needed to be computed was test cost per unit (TCPU), which is achieved by adding up the equipment's depreciation, energy consumption, labor, and consumables. Although it has been reported that initial TCPUs in the U.S. were 15–20% higher than those offshore, with the use of AI-organized automation and predictive maintenance, this gap is expected to be reduced dramatically.

Equipment utilization efficiency (EUE) was measured as the percent of test equipment active during production windows, which was used to measure operational maturity. Offshore sites registered EUEs of >90%, whereas restored facilities had EUEs of 75–85%, primarily due to onboarding the workforce and delays in integrating the software (Alex, 2023). These metrics were important both to compare the current performance as well as to evaluate the scalability

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

and sustainability of GPU testing strategies in restored environments, particularly of power-intensive GPUs with stacked high-bandwidth memory (HBM) and AI accelerators, where discrete shifts in yields, FATT, or EUE directly constrain deployable HPC capacity.

Table 4: ***Comparison of GPU Testing Metrics -Offshore vs U.S. Facilities (baseline performance gaps that reshored lines need to bridge to be competitive in terms of deploying GPUs).***

Metric	**Offshore Facilities**	**U.S. Facilities (Initial Phase)**
Yield Rate (%)	95–97	88–92
Failure Analysis Turnaround Time (hrs)	24–36	36–48
Test Cost Per Unit (TCPU)	Baseline	15–20% Higher
Equipment Utilization Efficiency (EUE, %)	>90	75–85

Table 4 metrics indicate that U.S. early-stage facilities are the least equipped with three GPU-specific metrics: bridging the yield gap that aggressive clock speeds and HBM integration cause, reducing the cycles of failure analysis on complex AI accelerators, and spreading equipment utilization to the >90% range that hyperscale GPU launches require. These quantified gaps are used to establish a reference frame for the adaptation pillars summarized.

5. Adapting Testing Frameworks for U.S.-Based Production

As GPU manufacturers return home, there is an important focus on modifying classical testing frameworks to suit domestic production profiles. The testing approaches that were optimized to address the centralized and offshore mega-factories must now be reconfigured to the more decentralized, less flexible, and innovation-focused U.S. manufacturing environment (Zhang et al., 2022), while simultaneously addressing the power-density, thermal-stress, and interface-complexity challenges unique to modern data-center GPUs. This metamorphosis requires introducing sophisticated AI tools, a modular test environment (TE), and an integrated ring of domestic suppliers. These determine a new technical sophistication and flexibility standard when producing the GPU. The results are interpreted of the pillars of adaptation with respect to the gaps observed in yield, FATT, TCPU, and EUE, with design decisions directly correlated to machine test performance on the GPU.

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

*Table 5: **Adaptation Pillars in U.S.-Based Testing Frameworks***

Adaptation Pillar	**Key Features**
AI-Driven Testing Protocols	Predictive analytics, real-time anomaly detection, and adaptive test conditions based on historic yield data.
Modular Testing Lines	Reconfigurable platforms enabling quick transitions from prototyping to full production.
Onshore Ecosystem Integration	Clustering of design, packaging, and testing centers to minimize delays and optimize test-feedback loops.

5.1 Integration of AI-Driven Testing Protocols

AI and ML are changing how GPU testing is conducted and optimized in domestic production lines. Due to the growing complexity of modern GPU architectures (multi-chip modules (MCMs), high-bandwidth memory integrations, and AI-specific accelerators), rule-based testing methodologies lack comprehensiveness and quickness. AI-powered frameworks present a scalable approach through real-time anomaly detection, predictive analytics for component failure, and adaptive testing dependent on historical yield data. AI models are incorporated into ATE systems to oversee wide-scale test output in practical applications, identifying anomalies that are beyond human detection. For example, the ML algorithms can learn from past production batches to forecast how defects are related to a particular lithography process or thermal stress anomalies. This predictive maintenance decreases the mean time to repair (MTTR) and the overall equipment efficiency (OEE).

The key role of AI in managing data is found in big data environments, in operations in the manufacturing of GPUs (Dhanagari, 2024). They point out MongoDB's ability to accommodate real-time, high-throughput data pipelines. Training machine learning models on test results produced from thousands of GPU units per hour is also mandatory. Combining scalable database technologies with transformer-based ML models enables complex pattern recognition, which can be used for greater test coverage and earlier detection of defects.

Transformer architectures (widely used in NLP) are also beginning to be applied to visual inspection systems in GPU testing. The transformers are excellent at capturing contextual meaning from visual data, which is directly applicable to automated optical inspection (AOI) systems. These systems can be utilized in identifying surface-level defects across die and substrate layers (Singh, 2022). Through accurate interpretation of visual cues, the models help to make the binning more accurate and the yield classification better.

The conceptual implementation of the AI-based testing process illustrated in Figure 4 could be a compact flowchart involving the following steps: (1) ingestion of high-volume test logs and sensor telemetry into ATE controllers and data into the data lake to be periodically re-trained on and to optimize the test-plan (2) feature engineering and defect labeling (3) offline model training and validation (4) deploying the trained model into ATE controllers (5) real-time

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

anomaly scoring and adaptive limit adjustment during the execution of the gpu tests (6) feedback This gradual description can be used to clearly illustrate the location of AI modules between unprocessed GPU telemetry and the code-optimized test code running on reshored lines.

Figure 4 outlines the sequential process of implementing AI-driven testing protocols for QA test automation, demonstrating how strategic integration supports defect prediction and system adaptability in GPU production lines.

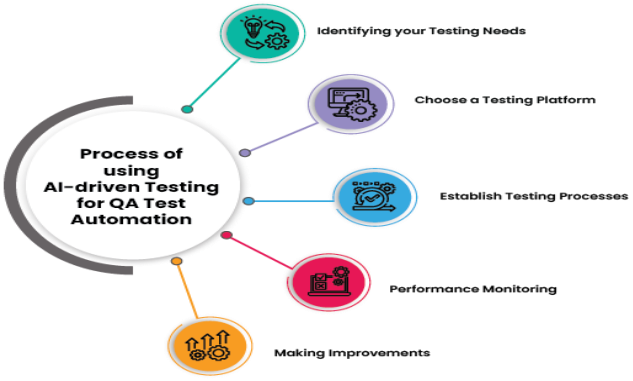

Figure 4: AI-Driven Quality Engineering-process flow of machine-learning-based test automation implementation in GPU production lines.

5.2 Modular Testing Lines for Scalability

Another important adaptation in U.S.-based GPU manufacturing is the development of modular and reconfigurable testing lines. In contrast to the usual monolithic, large-volume, high-production test lines in offshore bases, US operations are turning to flexible organizations that can easily be scaled or converted quickly. This is critical to fulfilling variable demand from AI startups, data center operators, and gaming industries, which all have unique performance validation criteria. Modular testing lines can have interchangeable heads, programmable logic controllers (PLCs), and robotics-enabled handlers. Such lines let facilities easily shift from consumer-grade GPUs to enterprise-class accelerators.

Utilizing characterized communication protocols like SECS/GEM enables interoperability of all different test systems and facilitates the addition of new modules during system integration. Modular systems' agility is critical for prototyping and low-volume production. US fabs, which are more innovation centers than high-volume manufacturers, can experiment with early engineering samples with customized parameters before easily transitioning into production testing as designs are defined. The capacity to reconfigure minimizes capital outlay and time-to-market, which are important indicators in the competitive semiconductor industry (Cats, 2020). The modular design often uses smart sensing that automatically adapts the test parameters to the environmental conditions (humidity, vibration, thermal drift). This intelligent adaptation, in turn, improves test accuracy while supporting sustainability goals in terms of optimizing power usage and lowering waste.

5.3 Onshore Ecosystem Integration

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The last pillar in adapting U.S.-based GPU testing frameworks is creating a unified and interactive domestic ecosystem. Testing cannot be walled off from substrate fabrication, assembly, cleanroom management, or test equipment manufacture. The various components of the value chain have been dispersed across the globe, which has created inefficiencies regarding logistics and quality control (Tien et al., 2019). Attempts are made to set up regional clusters of semiconductor activity where design, fabrication, and validation are collocated. For instance, fabs based in Arizona and Texas are collaborating with resident substrate manufacturers and cleanroom solution providers to reduce lead times and maintain the integrity of the supply chain.

Firm's resident in the US, including Teradyne and Advantest, are expanding their reach to conveniently serve domestic customers with a speedy deployment on the next-gen ATE platforms, which are ASIC-specific. Cleanroom infrastructure plays a significant role in ensuring contamination-free testing. Modern installations use Automated Material Handling Systems (AMHS) coupled with environmental monitoring sensors to maintain Class 1 environments during testing. These cleanrooms are being designed increasingly with the thought of modularity, scalability, and thus easier expansion in case testing demand increases.

The software interoperability between the design verification environments and the testing platforms is being improved to facilitate feedback loops, which can lead to fast design advances. For example, feedback on test data from the real world is supplied into EDA tools to allow for quick root cause analysis and design refinement. This closed-loop system is important for reducing iteration cycles and for competitive advantage. Reshoring GPU production in the United States requires a basic overhaul of testing approaches (Pinto, 2023). In the wake of the incorporation of AI-driven protocols, deployment of modular testing lines, and development of a tightly integrated domestic ecosystem, US factories are orienting themselves to provide world-class validation capacity targeted for newly emerging GPU architectures. A robust, adaptive, and visionary testing framework emerges from the intersection between technological innovation and localized collaboration.

6. Best Practices in Reshored GPU Testing

With GPU manufacturing back on U.S. ground, reshored testing strategies must bring on board best-in-class practices to compete on par with their offshore colleagues. The complexity and power density of contemporary GPUs require testing regimes that are not only a vast array of tests but must also be scalable and automated (Kandiah et al., 2021). Three interdependent practices have emerged as key pillars in domestic GPU test operations. Testing architectures parallelized, automated defect screening, with strong root cause analysis (RCA), and continuous testing optimization (CTO) systems. These are throughput efficiency, silicon quality, and curtained product development cycles.

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 6: Best Practices in Reshaped GPU Testing emphasize the practices directly overcoming high power density, complicated package design, and fast architectural churn in GPUs.

Practice	Description
Parallelized Testing Architecture	Simultaneous execution of test suites to improve throughput and reliability in thermal/power profiling.
Automated Defect Screening & RCA	Use of X-ray, CNNs, and post-silicon analytics to quickly detect and trace defect origins.
Continuous Testing Optimization	Real-time adaptation of test protocols using factory floor data and AI analytics to minimize test escapes and improve yield.

6.1 Parallelized Testing Architecture

In reshored GPU factories, parallelized testing architecture is a basic approach to increase testing throughput and reduce validation time per unit. Modern GPUs, housing hundreds of billions of transistors, need tests ranging from functional through electrical to thermal. Continuous-collar performance of these tasks has bottlenecks, especially at high volume (Tischbein et al., 2015). Parallel testing addresses this by allowing more than one set of test suites to be executed simultaneously across a fleet of automated test equipment (ATE) units, or even within a given modularized ATE system.

Parallelization is especially potent when benchmarking performance and when graphical processing units are taxed under low-to-high loads. Performance suites run concurrently on multiple GPU instances, which is similar to running real-world applications such as ray tracing, machine learning inference, and video encoding workloads. Not only does this reduce the time taken to validate, but it also makes it easier to compare performance from production lots. Parallelization also accrues an advantage in terms of power delivery analysis.

Voltage droop, current leakage, and transient behavior under dynamic loads are measured simultaneously from different GPU modules. This concurrent test assists engineers in identifying anomalies in power management integrated circuits (PMICs) and voltage regulation modules (VRMs) in real-time operating modes. This approach also greatly benefits thermodynamic profiling. Using embedded thermal sensors and infrared (IR) thermography, several dies are evaluated for heat maps, junction temperatures, and thermal throttling behavior (Sharma et al., 2024). Through parallelized analysis, test installations minimize variability in thermal data and can statistically eliminate outliers due to packaging flaws, TIM dissimilarity, or substrate de-bonding.

Figure 5 illustrates how parallel threading within a process enables simultaneous execution of test drivers and data handling routines, a foundational concept behind modern GPU parallel test strategies that boost efficiency and scalability.

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

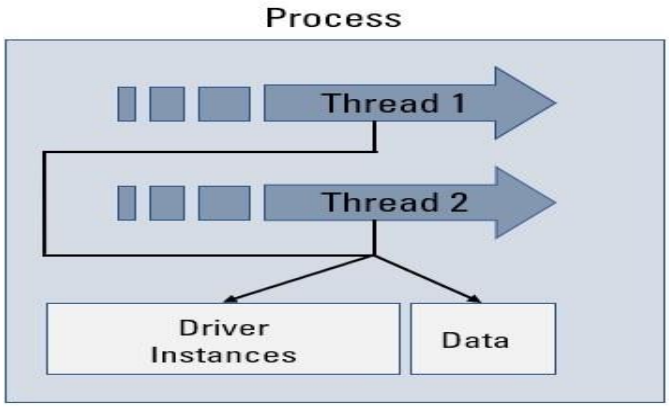

Figure 5: Parallel-test-architectures-to-lower-the-cost-to-test conceptual comparison between sequential and parallel GPS test flows and their effects on cost and cycle time.

6.2 Automated Defect Screening and Root Cause Analysis (RCA)

Automated defect screening is essential to achieving yield quality and speeding up root cause identification in high-mix GPU production swept environments. Compared to Manual inspection, which is constrained by human fatigue and resolution limits, automated systems utilize AI-driven visual inspection, X-ray scans, and post-silicon analytics to detect and classify defects in near real-time. X-ray inspection apparatus checks BGA (ball grid array) solder joints, trace vias, and internal interposers for micro-cracks, voids, or misalignments (Su, 2019).

Such systems integrate with factory execution systems (FES) that tag the defective units and reroute them to the rework stations or failure analysis labs. This tremendously high-resolution insight is important for systemic problems like warpage-induced delamination or excessive reflow in package-on-package (PoP) assembly. Visual AI inspection systems employ a Convolutional neural network (CNN) to examine the die surface for discoloration, mechanical scratches, or probe marks above the acceptable tolerances. These models are logically trained using past defect datasets, and the learning is further refined using the supervised learning system.

This method increases first-pass yield by decreasing the number of false positives and shooting only bad units. Post-silicon validation analytics refine the defect triage process, mining telemetry data collected in the bring-up and burn-in phases. Failures correlate to particular logic blocks or memory controllers if the engineer can undertake targeted RCA rather than conducting a laborious, often inconclusive full-die analysis. Many U.S.-based test labs incorporate these systems into digital twin environments, where simulated behavior is compared with test results to pinpoint likely defect sources.

6.3 Continuous Testing Optimization (CTO)

Continuous Testing Optimization (CTO) is a feedback-based testing paradigm in which current data gathered from the factory floor has a direct impact not only on the test parameters of the finished device but also upstream from the design. In restored GPU testing environments, CTO frameworks intervene where test engineering and design verification teams leave off (Chandrasekaran et al., 2023). At the center of CTO is a single data lake that collects test logs,

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

parametric results, and machine sensor data. With the help of advanced analytics platforms powered by AI, it is possible to track trends and anomalies, such as enhanced failure rates in a group of dies (and maybe across multiple die stacks on the same die), for example, or for an aspect of the thermal profile that degrades over time.

After patterns are identified, the automated rule-based systems modify the test condition (increase the time of thermal soak, add an extra margin test to underperforming units). CTO also makes it possible to develop a proactive test strategy. For example, if failures related to deep sleep mode correlate with a particular PLL (phase-locked loop) configuration, the CTO engine can prioritize those conditions in future test plans. Moreover, machine learning algorithms detect slow drift from equipment calibration, and predictive maintenance protocols are initiated before test escape risk balloons.

The work of the CTO does not stop at test floors. It moves to design. Statistical post-tapeout DFT modifications depend on real-time test data (Ganesan, 2020). For instance, if these scan chain faults occur disproportionately, the design team might insert some redundancy or improve BIST logic in subsequent designs. This repetitive improvement process is heavily rewarding in terms of reduced time-to-market and silicon respins, both financially and in terms of time.

In a GPU-testing perspective, respondents stressed multiple aspects of the same trio of persistently vexing clusters of challenges in even optimized domestic lines (i) being able to reach reliable burn-in and power coverage of very power-dense GPUs without prohibitive energy or cooling countermeasures (ii) being able to test the stacked-memory and chiplet interconnects in an actual AI and HPC application, where even margin defects can silently introduce a corrupting corrupted block of data; and (iii) is being able to cover test domain to leverage the next wave of AI accelerator Table 4 comparative metrics and Table 5 and Table 6 practises are thus to be interpreted as actual, GPU-specific solutions to these pain points and not generic semiconductor improvements.

7. Successful Case Study: Intel's U.S. Expansion and Validation Labs

7.1 Overview of Intel's Arizona and Ohio Facilities

Intel Corporation has made monumental growth in reshoring its high-end semiconductor manufacturing and validation capabilities with heavy investments in Arizona and Ohio. These facilities are pillars around which the company's strategy for strengthening the self-sufficiency of the United States in high-performance computing (HPC) components, including GPUs, is based. In 2022, Intel announced 2022 that it would have between $20 billion and two cutting-edge semiconductor factories in Licking County, Ohio, and an additional $20 billion expansion of its Ocotillo campus in Chandler, Arizona. These are not simply manufacturing centers but key testing and validation sites for Intel's third-generation design GPUs. The Arizona site, famous for its Fab 42, has Intel's most advanced 10nm Superfine and now the Intel 7 process technology (Taj, 2022). It comprises high-density cleanrooms, AI-integrated test labs, and chip packaging facilities. It also combines sophisticated thermal stress and electrical validation systems designed for GPU-specific use cases, including AI workloads and real-time rendering.

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The "Silicon Heartland" project in Ohio is Intel's most audacious domestic buildout in decades. The site is designed from the outset with collocated test and validation environments to ensure the direct transition of GPU wafer fabrication to thermal, logic, and accelerated workload testing without offshoring time delays. The facilities are being built to accommodate infrastructure to support sub-2nm node development, high-performance automated test systems, and edge data for real-time analytics. By centralizing fabrication and validation, Intel minimizes the chances of logistical failures and delivers fast feedback loops between design and testing (Heiney et al., 2021). This means having a better DFT loop optimized, which is very important in the current climate of GPU innovation.

7.2 Collaborative Strategy with Testing Equipment Suppliers

A critical support to Intel's domestic testing approach excellence is its strong collaborations with various premier test equipment companies, namely Teradyne and Advantest. These relationships have allowed Intel to deploy automated, more precision-oriented test systems designed to service the unique requirements of GPU architectures such as the multi-chip modules (MCMs), the AI inference accelerators, and the high bandwidth memory (HBM) integration. Teradyne supplies its UltraFLEX and Magnum lines of testers to Intel, which have high-parallelism and sub-nanosecond timing accuracy for verifying the integrity of the GPU cores, running at high clock speeds. These platforms are further complemented with machine learning models that continually improve test sequences for flaw detection and time reduction – an approach bolstered by predictive analytics principles that believe that data-backed decisions would lead to faster DevOps and system validation pipelines.

The collaboration between Advantest and Intel results in the V93000 platform, which is the most highly valued due to its scalability and throughput in large-volume GPU testing. These testers facilitate protocol-aware validation, such as PCIe 5.0, DP, and HBM interfaces—an important part of evaluating what discrete GPUs can deliver in a live application. Apart from integrations with hardware devices, Intel has partnered with these vendors collaboratively to jointly develop sophisticated analytics and visualization tools that would map defect patterns, thermal profiles, and yield variations in multiple fabs and test sites. Predictive models are now used to identify probable areas of failure and variably prioritize test vectors (Kumar, 2019). This integration allows for improved continuous improvement on test lines, non-revenue downtime, and engineering productivity.

7.3 Operational Results and Productivity Benchmarks

The effects of Intel's strategic reshoring and testing modernization are reflected in its operational metrics. In the Arizona facility, test coverage effectiveness in terms of the ratio of functional logic and signal paths verified per silicon die has risen roughly about 30 percent compared to legacy offshore models (Weyer, 2019). This improved due to the test systems' real-time adaptability and statistical collocation of failure analysis (FA) labs with the design teams, which reduced debug turnaround times. Defect escape rates (the percentage of faulty GPUs that slip through test undetected) have reduced by an astonishing 45%, in line with Intel's internal performance audits. This is due to a layered test strategy developed together with Teradyne and Advantest, where static test vectors are supplemented with dynamic workload

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

emulation and AI-based thermal stress tests. Such improvements are consistent with lessons from the role of predictive analytics in enabling tighter feedback loops between operational analytics and production arrangements.

Silicon yield rates have also been improved significantly. The offshore test processes have produced up to 8 percent less yield in advanced GPU lots than the U.S.-based facilities have been reported to do. This performance enhancement directly affects the level of cost efficiency and the quality of environmental sustainability, as fewer wafers have to be manufactured to achieve output objectives. The increased yield is captured with a 20% decrease in time for test cycles, driven by parallelism of test runs and real-time error mapping technologies (Chinamanagonda, 2021). In their entirety, Intel's U.S. facilities are now benchmarks of what can be done when fabrication, testing, and analysis are dovetailed at one single domestic ecosystem. This case study re-emphasizes Intel's technological leadership and the strategic need for reshoring high-value semiconductor capabilities to increase national resilience and innovation.

Figure 6 illustrates the multidimensional contributors to operational excellence achieved through Intel's domestic reshoring strategy, combining technological innovation with structured performance management and employee alignment.

Figure 6: Achieving Operational Excellence through Artificial Intelligence

Meanwhile, the single reference point of Intel will bring significant impairment. Intel is a high capital, old supplier network, mature internal analytics culture maker of integrated devices, with experience on its side, and so represents a best-case scenario and not the archetypal reshoring process of a foundry, OSAT company, or smaller fabless company. The process nodes, product mix, and risk tolerance vary significantly within the industry, so not every facility can be as large as the scale of investment or level of test-design integration as Netflix happened to be at Intel. The case study is an example of the possible upper limit of what domestic GPU validation may do, and future investigations require the work of many firms to be compared to comprehensively project the results in other business models and technology nodes.

8. Talent, Training, and Workforce Development

8.1 Skills Gap in Advanced Semiconductor Testing

Received: August 06, 2025 Published: November 26, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Looking at the in-critical-semi-ICC as the U.S. Re-whores initiative inexorably gains momentum, the main issue that stands out is the huge skills deficit, especially in advanced GPU testing areas. Even though there has been a huge investment in fab and the construction and testing facilities, the availability of a qualified workforce, such as GPU test engineers, the operators of automated equipment, and noise reliability specialists, has not caught up with the need. The explosive industry benchmarks on testing complexity caused by GPUs, with AI-accelerating cores, ray tracing units, and MCM architectures, require mixed-signal validation knowledge, high-speed interconnect analysis expertise, and fault injection modeling. The current domestic workforce pipeline is not wide enough to provide that kind of niche technical prowess.

Most GPU testing processes are multi-phase validation: wafer-level testing, package-level electrical testing, system-level stress testing, and long-term reliability verification. These procedures demand professionals who can read test logs, calibrate testing interfaces, and analyze statistical yield. However, because of decades of offshoring, there is a lack of hands-on experience with state-of-the-art test handlers, burn-in chambers, and vector-based instrumentation development tools among our U.S.-based engineers. Proficiency in common instruments like Advantest's V93000 platform or Teradyne's UltraFLEX systems is relatively rare outside a few large companies (Jiang et al., 2023). The lack of reliability specialists becomes more challenging considering the enhanced regulation of domestic semiconductor products. These professionals guarantee device life in different temperature, voltage, and mechanical stress conditions, essential for the GPUs used in automotive AI and defense applications. This gap in the pool of skills could slow testing cycles, increase defect escape rates, and, in the process, no longer propel the mindset of reshoring. Quality control and national resilience of the supply chain.

8.2 Academic and Industry Partnerships

Hoping to close this acute talent gap, several academic players and industry leaders are developing symbiotic partnerships for education on semiconductor testing. Universities including Arizona State University (ASU), Purdue, and Georgia Tech have launched microelectronics reliability, IC testing, and hardware security-focused programs, often in partnership with the makers of chips and testing equipment. These programs combine theoretical learning with labs that replicate the real-world testing environment (Schäpke et al., 2018). Semiconductor consortia (SEMI) and bodies like the American Semiconductor Academy have started skill certification pipelines. These are short-term technical certifications in the areas of Design for Test (DfT), JTAG-based fault diagnosis, thermal interface material (TIM) testing, and electrostatic discharge (ESD) integrity tests. With leadership from private-sector champions, these programs are shifting increasingly towards cloud services to run simulated GPU stress test workflows, allowing students to learn in virtual laboratories.

Some initiatives also derive experiences from larger DevSecOps software deployment approaches by discussing security verification in the hardware validation of procedures (Konneru, 2021). This cross-domain strategy inspires the nascent engineers to embrace continuous integration mindsets, where the automated test scripts and runtime diagnostics are

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

part and parcel of the hardware's development life cycle. Think, for instance, how students can be made to understand concepts similar to Static Application Security Testing (SAST) and Dynamic Application Security Testing (DAST) in firmware validation, so that they will be ready to join the work stream of a secure GPU testing pipeline upfront. Apart from initiatives like degrees, apprenticeship programs, and paid co-op with NVIDIA, AMD, and Intel were found to be necessary (Abraham et al., 2024). This enables students to interact directly with live testing equipment while providing a smooth interface from the classroom to the production floor. As the U.S. semiconductor reshoring endeavor does so, this academia-industry convergence can be a backbone for the sustainability of workforce development.

Figure 7 illustrates the tripartite model for sustainable circular economics in semiconductor education and testing—highlighting the collaborative dynamics between government agencies, academic institutions, and the private sector.

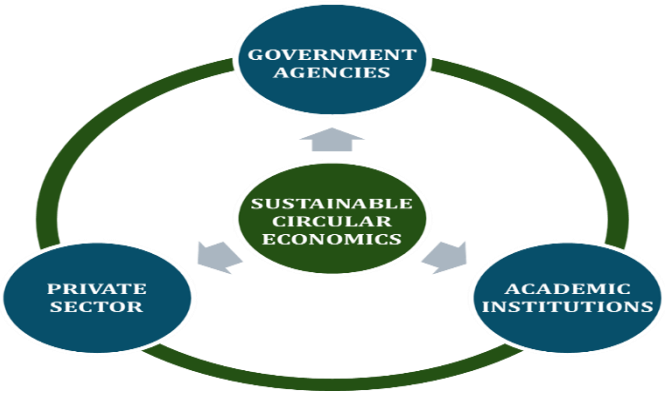

Figure 7: Tripartite circular-economy model linking government, academia, and private semiconductor partners.

8.3 Onsite vs. Remote Testing Teams

As home semiconductor test operations grow, so does the popularity of hybrid and distributed engineering models. Historically, all phases of the GPU validation (wafer probing, initial device to test server synchronization, and final system test) required personnel at the device site because of the necessity to use the physical device, calibration of the equipment, and high-voltage test interfaces. Today, developments in cloud infrastructure and live monitoring tools can virtualize or remotely monitor parts of the testing pipeline (Benkhelifa et al., 2019). Cloud-native platforms coupled with test data management systems enable remote teams to use performance logs, yield analytics, and thermal imaging outputs in real time. This allows for parallelized test engineering teams poised across time zones, improving productivity without overburdening single-site teams. Current testing equipment may be integrated using secure APIs to allow Remote script-based test configuration upload and analysis of failure modes. Full virtualization is still not feasible for the entire set of testing phases. Even though sockets for GPU testing, reflow process control, and optical inspection can be handled remotely, physical interaction onsite is necessary. For this reason, hybrid teams, in which core tasks such as test engineering script development, data analytics, and test coverage optimization are performed remotely but hardware interaction is still local, provide the practical compromise.

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Companies are also researching remote incident response channels through the uptake of diagnostic data to centralized systems, which would enable remote reliability engineers to undertake root cause analysis just hours after a fault is detected. The adoption of secure, automated pipelines makes improving time-to-resolution easier while also contributing to compliance and traceability throughout remote validation environments. The future of testing talent may well hang on the flexibility of hybrid arrangements and ongoing training in both hardware installation and digital teamwork (Ahmed & Smith, 2023). The organizations that will be proactive and step ahead by maintaining secure remote testing frameworks and cloud-compatible test platforms will be in the best position to tap into the entire gamut of talent available on land-based in the United States, whether onsite or distributed.

9. Future Trends in U.S.-Based GPU Testing

The landscape of GPU (graphics processing unit) testing in the United States is undergoing dramatic changes due to changing computational paradigms, environmental imperatives, and simulation technologies. As U.S. semiconductor manufacturers continue to bring production home, taking the lead in testing innovation becomes essential (Platzer & Sargent, 2016). New approaches like quantum-aware testing, sustainable lab practices, and digital twins to be used on GPU testing for performance, reliability, and market readiness are on the horizon.

Figure 8 captures the multifaceted technical challenges posed by next-generation neuromorphic and quantum-enabled GPU designs—focusing on critical factors such as energy efficiency, precision, and adaptability, which testing protocols must soon accommodate.

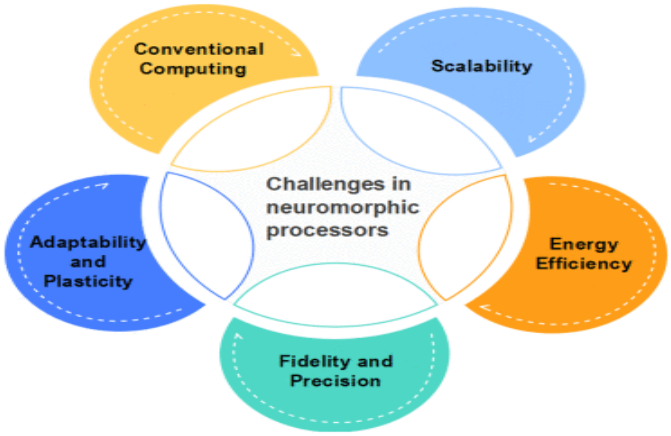

Figure 8: ***Computing of neuromorphic materials***

9.1 Quantum-Aware GPU Testing for Next-Gen Computing

Advancing into the epoch of quantum-enabled GPU architectures is a major inflection point in processor design and validation. While traditional GPUs optimized for graphics and AI workloads are designed to perform linear algebra operations, the next generation of GPUs is starting to incorporate quantum-inspired logic gates and cryogenic devices that send test requirements into new territory. This transition requires the creation of quantum-aware test frameworks for classical and non-classical computation realms. Detecting entangled-state behavior in hybrid computing architectures is an important challenge in quantum-aware testing (Zhou et al., 2023). Such behaviors are prone to noise and decoherence and are best detected

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

with ultra-low-latency diagnostic apparatus working at cryogenic temperatures. Test engineers have to use error-correcting protocols specific to qubit instability since traditional redundancy checks and CRCs (cyclic redundancy checks) are useless in this area. Quantum-aware GPU testing must go beyond validation after fabricating a chip. In-situ test monitoring at runtime, particularly within quantum annealing simulators and superconducting logic, can be considered a prerequisite today. This means embedding quantum-aware sensors and calibration devices into the packaging of the GPU unit. To facilitate such tasks, U.S.-based test facilities are acquiring high-end metrology instruments and simulation platforms that represent quantum processes, thereby increasing the accuracy of predictions for gate-level failures and entanglement mistakes.

Early roadmaps identify specific models of concrete collaboration, such as quantum-computing vendors like IBM Quantum or IonQ collaborating with arguments from GPU manufacturers, like NVIDIA and AMD, to build new joint cryogenic testbeds in which some quantum circuit compiler drives a piece of qubit-GPU data paths, and the joint instrumentation ensures qubit-GPU data paths are valid. When organized in such consortia fashion, these programs imply that testing quantum-era GPUs will be structured around ecosystems of classical accelerator manufacturers, quantum hardware makers, and specialty meter Makers, and not an individual project.

9.2 Environmental Sustainability in Testing Processes

The return of GPU production to the homeland opens a unique door for the United States to build up environmentally friendly semiconductor manufacturing hubs from scratch. Testing labs are especially sensitive as resource-intensive parts of the GPU production process to their electricity consumption, water intake, and chemical spillage (Vitelli, 2024). Industry experts assert that sustainability in testing needs to be engineered within a multidimensional perspective that integrates strategic soundness of operations with ecological sensitivity. The energy efficiency of automated test equipment (ATE) is a subject of attention. Energy consumption per unit test on modern test platforms has been reduced by over 30% by redesigning the same to incorporate low-power FPGAs and smart switching matrices.

Recent benchmarking of green ATE roll-outs at the leading fabs indicates similar savings, with certain U.S. pilot lines recording a 25-35% decrease in electricity consumption on the test-floor when switching to high-efficiency power supplies and adaptive-idle operating modes without decreasing target throughput. In factories, dynamic scheduling algorithms are deployed, which schedule the testing work based on the grid demand and availability of renewable energies. This harmony was discovered on the relevance of intellectual scheduling for better operational outcomes (in this instance, reducing a lab's carbon footprint without reducing throughput).

Water conservation is also a major element of sustainable GPU testing. Ultra-pure water (UPW) needed for thermal control and particulate removal is a common need in high-performance semiconductor test lines. U.S.-based factories are incorporating closed-loop water recycling and nanofiltration practices, which conserve up to 70% of water. Industry case studies show that by implementing closed-loop UPW systems in test-intensive production modules, reducing fresh-water consumption by cooling and cleaning to draw into the production process

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

by a factor of two, and automatically saving millions of gallons per year in millions of GPU fabs of scale, is feasible.

Such systems are augmented by data analytics platforms that monitor the flow rate, contamination levels, and maintenance schedules. Vital endeavors to reduce the use of hazardous chemicals are beginning to pick up. Only alternatives to solvent-based cleaning agents and photoresist strippers are possible, and plasma-based dry cleaning and biodegradable materials are promising practical substitutes. EHS compliance is not a regulatory box anymore but rather a key KPI for GPU manufacturers seeking to position themselves as sustainable industry leaders.

9.3 Integration of Digital Twins and Virtual Simulation

With the advent of digital twin technology, the familiar limits of GPU testing are being uprooted as virtual counterparts for physical appliances, processes, and surroundings are created. In the case of reshoring, where agility and accuracy are most critical, digital twins give U.S. factories an edge in defect prediction, test coverage optimization, and time to market. These virtual worlds are especially efficient at imitating a computer's GPU's thermal, electrical, and mechanical stresses during work. Engineers can conduct Design-for-Testability (DFT) validation long before a physical chip is made using digital twins. This minimizes first silicon risk and supports iterative testing by simulating field conditions. For instance, sudden voltage drops, fan failure, or gaming workloads under extreme temperatures can be modeled accurately (Hanif et al., 2023). The intrusion of digital twins also contributes to cross-functional collaboration between chip designers, validation engineers, and production managers. Thanks to feedback from digital simulations, test parameters can adapt in real time, ensuring that edge-case scenarios are addressed before they can negatively affect yield. Moreover, such systems are increasingly powered by AI, which provides the opportunity for predictive analytics of failure patterns and test time optimization.

Smart scheduling finds a parallel where digital twins are used to schedule parallel test streams, simulate resource contention, and schedule testing equipment based on the projected loads (Sardana, 2022). This results in substantial equipment utilization and test throughput, which benefit greatly from the just-in-time production ambience of the restored facilities. Introducing quantum-aware testing, sustainability, and digital twins in the GPU verification lifecycle is far more than an innovation trend and is, therefore, a strategic imperative. As the U.S. increases its domestic semiconductor capacity, the possibility of adopting these future-facing methodologies will define its global competitiveness and technological independence.

10. Conclusion and Strategic Outlook

The reshoring of GPU production back to the United States has led to tremendous changes in semiconductor manufacturing and testing frameworks. There is no longer an aspirational ideal to having localized validation infrastructures, but a competitive and national security one. U.S. factories are also shifting centralized off-shore paradigms to AI-based test protocols, modular equipment, digital twins, and sustainability components, which are more aligned to validate against new GPU architectures and security requirements. Government policy, market

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

pressure, and technological advances have driven this change. The COVID-19 pandemic unveiled massive vulnerabilities in the supply chain globalization, largely in East Asia, where most GPU testing infrastructure was located. In retaliation, legislative efforts such as the CHIPS and Science Act have incentivized and outlined a strategy for domestic semiconductor manufacturers. These policies initiated the growth of high-throughput validation facilities, workforce development programs, and cross-sector partnerships that pushed U.S. firms to return to the objectives of testing once offshored to enhance economic effectiveness.

The process has not been smooth sailing. The U.S factories had infrastructure gaps in harmonizing fabrication, packaging, and testing in a geographically integrated and operationally coordinated ecosystem. Labor shortages, particularly those in the validation trenches for GPUs, only worsened scaling efforts. The industry has taken several meaningful steps with the push of innovation and public-private partnership. Using AI-enhanced diagnostic tools, predictive analytics for real-time defect detection, and parallelized test platforms has reduced yield rate and shortened debug cycles. Places like Intel's Arizona and Ohio campuses have become transformational testbeds for world-class GPU validation, such as onshore testing ecosystems, owing to the merit and strategy of these ecosystems. Such coordination among policymakers, semiconductor companies, and academic institutions is essential to sustain momentum.

Policymakers will have to continue supporting R&D funding initiatives and removing regulatory bottlenecks that would slow facility deployment. The territory of the CHIPS Act, if it grew to include particular testing technologies and employee incentives, would receive this foundation of long-term domestic resilience in stone. Meanwhile, industry leaders have to double down on ecosystem integration – tying up more closely with domestic equipment vendors, substrate providers, and clean room service providers to reduce friction in logistics and enhance throughput of testing. Academic institutions must help cut the talent gap. Partnerships with the chip makers and test equipment companies must spill over to the apprenticeship pipelines, simulated training environments, and remote diagnostics engineering beyond degree programs. The trained personnel of engineers will be equipped with skills in modular testing systems, digital twins, AI-driven analytics, and quantum-aware architectures to enable them to adapt to ever more complex GPU designs.

Major actionable priorities that come out of the study to consolidate such gains are as follows. Incentives covering the current state of CHIPs should be explicitly applied to high-tech test systems, such as AI-based ATE, clean utilities, and test clusters based on the region. The industry leaders should also further integrate the ecosystems with the domestic equipment vendors, substrate suppliers, and cloud analytics providers to bridge yield, FATT, and EUE gaps. Institutions of higher learning and training consortia have to scale practice-based programmes and apprenticeships to generate engineers who are fluent in modular test systems, digital twins, and quantum-verse validation. The combination of these measures will turn the findings of the paper into a finely-worked roadmap to make the U.S. GPU testing more resilient and high-performing.

Received: August 06, 2025 Published: November 26, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In the future, the United States is in an excellent position to be a leader in GPU validation if the current trajectory of technological investments and collaborative innovations is maintained. Classical and non-classical computing boundaries will be redefined by quantum-enhanced GPU architectures that will require a further range of validating techniques. The cryogenic diagnostics, entanglement modeling, and hybrid logic validation frameworks must continue to be funded in U.S. test labs. Also, with the increased global concerns about environmental sustainability, U.S.-based testing has to be at the forefront of energy-efficient, low-emission actions, such as closed-loop water systems, biodegradable solvents, and smart energy scheduling. Reshoring GPU testing is no longer a retroactive move to manage geopolitical risk. It is a forward-oriented strategic blueprint for future competitiveness. The convergence of AI and automation with sustainability and digital simulation has established a new gold standard for GPU validation. Through continued investment into the spirit of domestic self-sufficiency, consistent innovation, and cross-industry teamwork, the United States stands to maintain its long-term dominance in the manufacture of GPUs and constantly establish new boundaries of semiconductor testing globally.

Reference;

[1] Abraham, S., Abston, P., Adamson, R., Anantharaj, V., Barker, A., Barlow, A., ... & Zimmer, C. (2024). *2023 Operational Assessment Oak Ridge Leadership Computing Facility* (No. ORNL/SPR-2024/3352). Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States).

[2] Aguirre, T. (2024). On Labs and Fabs: Mapping How Alliances, Acquisitions, and Antitrust are Shaping the Frontier AI Industry. *arXiv preprint arXiv:2406.01722*.

[3] Ahmed, S., & Smith, E. (2023). The Future of Work: Adapting to Remote and Hybrid Models. *Abbottabad University Journal of Business and Management Sciences*, *1*(01), 1-12.

[4] Alam, M. M. (2017). *Construction and Evaluation of an Experimental Setup using Infrared-Thermography for the measurement of the Undercooling of Solder Alloys* (Doctoral dissertation, Technische Universität Dresden).

[5] Alex, B. (2023). *Transitioning to a New Horizon: Exploring the drivers, barriers, and success factors for Norwegian Offshore service vessel firms' transition from Oil and Gas to the Offshore Wind Industry* (Master's thesis, University of South-Eastern Norway).

[6] Benkhelifa, E., Hani, A. B., Welsh, T., Mthunzi, S., & Guegan, C. G. (2019). Virtual environments testing as a cloud service: a methodology for protecting and securing virtual infrastructures. *IEEE access*, *7*, 108660-108676.

[7] Bukhari, Z., Yahaya, J., & Deraman, A. (2019). Metric-based measurement and selection for software product quality assessment: Qualitative expert interviews. *International Journal of Advanced Computer Science and Applications*, *10*(7), 223-231.

[8] Cai, Z., Krehel, O., Wang, A., Wu, M., & Zhai, J. (2022). Unmasking Vulnerabilities: Assessing Third-Party Risks in NVIDIA.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[9] Cats, A. (2020). *Alliance network strategies for cross-industry collaboration and the internationalisation of R&D in the semiconductor industry* (Doctoral dissertation, University of Reading).

[10] Chandrasekaran, J., Cody, T., McCarthy, N., Lanus, E., & Freeman, L. (2023). Test & evaluation best practices for machine learning-enabled systems. *arXiv preprint arXiv:2310.06800*.

[11] Chavan, A. (2023). Managing scalability and cost in microservices architecture: Balancing infinite scalability with financial constraints. Journal of Artificial Intelligence & Cloud Computing, 2, E264. http://doi.org/10.47363/JAICC/2023(2)E264

[12] Chinamanagonda, S. (2021). AI-driven Performance Testing AI tools enhancing the accuracy and efficiency of performance testing. *Advances in Computer Sciences*, *4*(1).

[13] Dhanagari, M. R. (2024). MongoDB and data consistency: Bridging the gap between performance and reliability. *Journal of Computer Science and Technology Studies, 6*(2), 183-198. https://doi.org/10.32996/jcsts.2024.6.2.21

[14] Dhanagari, M. R. (2024). Scaling with MongoDB: Solutions for handling big data in real-time. *Journal of Computer Science and Technology Studies, 6*(5), 246-264. https://doi.org/10.32996/jcsts.2024.6.5.20

[15] Ganesan, K. K. (2020). *Counterexamples in Digital System Verification, Channel Coding, and Electro-Neural Interface Design*. Stanford University.

[16] Gatenholm, G., & Halldórsson, Á. (2023). Responding to discontinuities in product-based service supply chains in the COVID-19 pandemic: Towards transilience. *European Management Journal*, *41*(3), 425-436.

[17] Goel, G., & Bhramhabhatt, R. (2024). Dual sourcing strategies. *International Journal of Science and Research Archive*, 13(2), 2155. https://doi.org/10.30574/ijsra.2024.13.2.2155

[18] Hanif, S., Mukherjee, M., Poudel, S., Yu, M. G., Jinsiwale, R. A., Hardy, T. D., & Reeve, H. M. (2023). Analyzing at-scale distribution grid response to extreme temperatures. *Applied Energy*, *337*, 120886.

[19] Heiney, J., Lovrien, R., Mason, N., Ovacik, I., Rash, E., Sarkar, N., ... & Kempf, K. (2021). Intel realizes $25 billion by applying advanced analytics from product architecture design through supply chain planning. *INFORMS Journal on Applied Analytics*, *51*(1), 9-25.

[20] Hoi-Chun Hung, A. (2024). Chip Legislative Endeavors in the United States and European Union: A Comparative Analysis Based on China's Disruptive Production Technologies. *Journal of Law, Technology & Policy*, *2024*(2).

[21] Jiang, Y., Zhang, Z., Yu, K., & Qi, J. (2023). Integrated Circuit-Testing Equipment. In *Handbook of Integrated Circuit Industry* (pp. 1629-1642). Singapore: Springer Nature Singapore.

[22] Jyoti, S. N., Islam, M. R., & Kudapa, S. P. (2024). The Role of Test Automation Frameworks In Enhancing Software Reliability: A Review Of Selenium, Python, And

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

API Testing Tools. *International Journal of Business and Economics Insights*, *4*(4), 01-34.

[23] Kandiah, V., Peverelle, S., Khairy, M., Pan, J., Manjunath, A., Rogers, T. G., ... & Hardavellas, N. (2021, October). AccelWattch: A power modeling framework for modern GPUs. In *MICRO-54: 54th Annual IEEE/ACM International symposium on microarchitecture* (pp. 738-753).

[24] Kang, Y. S., Park, I. H., & Youm, S. (2016). Performance prediction of a MongoDB-based traceability system in smart factory supply chains. *Sensors*, *16*(12), 2126.

[25] Karwa, K. (2024). The future of work for industrial and product designers: Preparing students for AI and automation trends. Identifying the skills and knowledge that will be critical for future-proofing design careers. *International Journal of Advanced Research in Engineering and Technology*, *15*(5). https://iaeme.com/MasterAdmin/Journal_uploads/IJARET/VOLUME_15_ISSUE_5/IJARET_15_05_011.pdf

[26] Konneru, N. M. K. (2021). Integrating security into CI/CD pipelines: A DevSecOps approach with SAST, DAST, and SCA tools. *International Journal of Science and Research Archive*. Retrieved from https://ijsra.net/content/role-notification-scheduling-improving-patient

[27] Kumar, A. (2019). The convergence of predictive analytics in driving business intelligence and enhancing DevOps efficiency. International Journal of Computational Engineering and Management, 6(6), 118-142. Retrieved from https://ijcem.in/wp-content/uploads/THE-CONVERGENCE-OF-PREDICTIVE-ANALYTICS-IN-DRIVING-BUSINESS-INTELLIGENCE-AND-ENHANCING-DEVOPS-EFFICIENCY.pdf

[28] Mandya Channegowda, M. G. (2022). Improving time-to-market and customer satisfaction in the SoC product business: an approach to enhance productivity by "reusability strategy".

[29] Mishra, A., Cha, J., Park, H., & Kim, S. (Eds.). (2023). *Artificial Intelligence and Hardware Accelerators*. Springer.

[30] Navaux, P. O. A., Lorenzon, A. F., & da Silva Serpa, M. (2023). Challenges in high-performance computing. *Journal of the Brazilian Computer Society*, *29*(1), 51-62.

[31] Nyati, S. (2018). Transforming telematics in fleet management: Innovations in asset tracking, efficiency, and communication. International Journal of Science and Research (IJSR), 7(10), 1804-1810. Retrieved from https://www.ijsr.net/getabstract.php?paperid=SR24203184230

[32] Pandit, V. (2022). Time-and value-continuous explainable affect estimation in-the-wild.

[33] Pinto, C. (2023). Enhancing Resilience in Global Value Chains: a comprehensive analysis of reshoring and its implementation in the semiconductor industry from a US and European perspective.

Received: August 06, 2025 Published: November 26, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[34] Platzer, M. D., & Sargent, J. F. (2016). *US semiconductor manufacturing: Industry trends, global competition, Federal Policy*. New York: Congressional Research Service.

[35] Rinehart, W., & Kirchhoff, A. (2024). The political economy of the CHIPS and Science Act. *The Center for Growth and Opportunity*.

[36] Sardana, J. (2022). The role of notification scheduling in improving patient outcomes. *International Journal of Science and Research Archive*. Retrieved from https://ijsra.net/content/role-notification-scheduling-improving-patient

[37] Schäpke, N., Stelzer, F., Caniglia, G., Bergmann, M., Wanner, M., Singer-Brodowski, M., ... & Lang, D. J. (2018). Jointly experimenting for transformation? Shaping real-world laboratories by comparing them. *GAIA-Ecological Perspectives for Science and Society*, *27*(1), 85-96.

[38] Sharma, A. M., Jacobs, K. J., Coenen, D., & De Wolf, I. (2024). Enhanced infrared imaging for die-level fault isolation using lock-in thermography. *Journal of Failure Analysis and Prevention*, *24*(5), 2129-2141.

[39] Singh, V. (2022). Visual question answering using transformer architectures: Applying transformer models to improve performance in VQA tasks. Journal of Artificial Intelligence and Cognitive Computing, 1(E228). https://doi.org/10.47363/JAICC/2022(1)E228

[40] Su, S. (2019). *Reliability of doped SnAgCu solder alloys with various surface finishes under realistic service conditions* (Doctoral dissertation, Auburn University).

[41] Tache, M. D., Păscuțoiu, O., & Borcoci, E. (2024). Optimization algorithms in SDN: Routing, load balancing, and delay optimization. *Applied Sciences*, *14*(14), 5967.

[42] Taj, M. N. (2022). *Intel Chip Manufacturing Technology Roadmap*.

[43] Tien, N. H., Anh, D. B. H., & Thuc, T. D. (2019). Global supply chain and logistics management.

[44] Tischbein, B., Hornidge, A. K., Djumaeva, D., Subramanian, S., Bhaduri, A., Bekchanov, M., ... & Worbes, M. (2015). *Restructuring land allocation, water use and agricultural value chains: Technologies, policies and practices for the lower Amudarya region*. V&R Unipress.

[45] Vitelli, M. (2024). Artificial Intelligence “at the Edge” for Resource-Constrained Devices Based on SENSIPLUS Technology.

[46] Weyer, D. J. (2019). *Tradeoffs Between Performance and Reliability in Integrated Circuits* (Doctoral dissertation, Case Western Reserve University).

[47] Zhang, M. Y., Dodgson, M., & Gann, D. (2022). *Demystifying China's innovation machine: Chaotic order*. Oxford University Press.

[48] Zhou, X., Shen, A., Hu, S., Ni, W., Wang, X., Hossain, E., & Hanzo, L. (2023). Towards quantum-native communication systems: New developments, trends, and challenges. *arXiv preprint arXiv:2311.05239*.