proteanTecs

WHITE PAPER

Scaling GenAI Training and Inference Chips with Runtime Monitoring

Table of Contents

1 | The Unforgiving Reality of Scaling Cloud AI

The shift to GenAI has outpaced the infrastructure it runs on. What were once rare exceptions are now daily operations: high model complexity, non-stop inference demand, and intolerable cost structures. The numbers are no longer abstract. They're a warning.

Training a model like GPT-4 (Generative Pre-trained Transformer) reportedly consumed 25,000 GPUs over nearly 100 days, with costs reaching $100 million [1]. GPT-5 is expected to break the $1 billion mark [2]. Energy usage is just as daunting. Training GPT-4 drew an estimated 50 GWh, enough to power over 23,000 U.S. homes for a year [3]. Even with all that investment, reliability is fragile. A 16,384-GPU run experienced hardware failures every three hours, posing a threat to the integrity of weeks-long workloads [4].

Note: Internal and external memory bottlenecks, such as bandwidth limits and data transfer overhead, remain critical constraints, though they are outside the scope of this analysis.

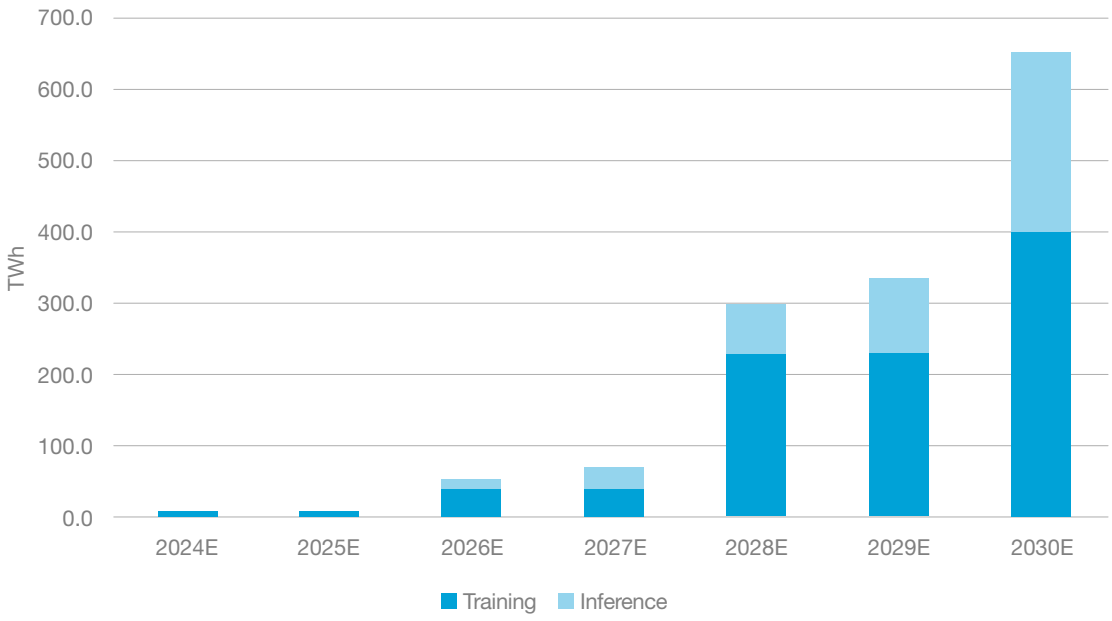

Figure 1: Projected AI power consumption grows from 8 TWh in 2024 to 652 TWh by 2030 (8,050%), driven by both training and a rapidly growing share of inference. Based on Wells Fargo data via IO Fund [5].

Inference isn't easier. ChatGPT now serves more than one billion queries daily, with operational costs nearing $700K per day [6]. Each response, priced at just fractions of a cent, adds up to an infrastructure bill that outpaces most business models. That pressure is made worse by performance gaps. Users frequently report over 20-second delays for answers [7]. At this scale, even slight inefficiencies multiply into real dollars and degraded user experience.

These are not isolated incidents. They are signs of systemic strain. Massive training runs, crushing query volumes, rising failure rates, and mounting electricity costs—this is the environment GenAI must thrive in. What's needed isn't incremental optimization. It's a way to reclaim control and scale effectively.

The table below outlines the core challenges behind these risks. Each is backed by hard data. Together, they show just how steep the hill has become.

CATEGORY	CHALLENGE	DESCRIPTION	QUANTITATIVE EXAMPLES
TRAINING	**Cost Efficiency**	Extremely costly due to hardware, power, and infrastructure demands.	GPT-4 training cost ~$100M. GPT-5 training is projected to exceed $1B.
	Reliable Scalability	Failures in one node can corrupt others across tightly coupled compute clusters.	GPT-4 training used ~25,000 A100 GPUs over 90-100 days. Another run with 16,384 GPUs had a failure every ~3 hours.
	Power & Heat	Chips consume immense energy and generate heat, driving up costs and cooling needs.	GPT-4 training consumed ~50 GWh, which is equal to the annual use of ~23,000 U.S. households.
INFERENCE	**Cost Efficiency**	Inference loads serving hundreds of millions of users generate massive infrastructure bills, threatening profitability.	Operating ChatGPT costs ~$694K/day using ~28,936 GPUs (~$0.0036 per query).
	Power Efficiency	At GenAI inference scale, even small inefficiencies lead to massive energy and cost waste.	AI data centers may reach 90 TWh/year by 2026. That's 10x the levels of 2022.
	Latency	Chips must respond fast despite compute needs growing exponentially.	Many ChatGPT users report delays of 20-30+ seconds per prompt.

Table 1: Key operational challenges in cloud AI workloads.

2 | Mastering the GenAI Arms Race: Why Node Upgrades Aren't Enough

Moore's Law predicts that the number of transistors in an IC doubles approximately every two years with minimal cost increase. The law was accurate for decades, yet recent fabrication challenges slowed it to around 2.5 years for each new node [8]. More importantly, even the original rate couldn't keep up with GenAI's computational requirements, which double much faster than transistor density.

Even the original rate of Moore's Law couldn't keep up with GenAI's computational requirements, which double much faster than transistor density.

It took 2.6 years to move from 5nm to 3nm, yet the reported performance gain at the same power was only about 10-15%, with 25-30% improvements in power efficiency at the same speed [9]. **Meanwhile, GenAI workload demands are growing orders of magnitude faster.**

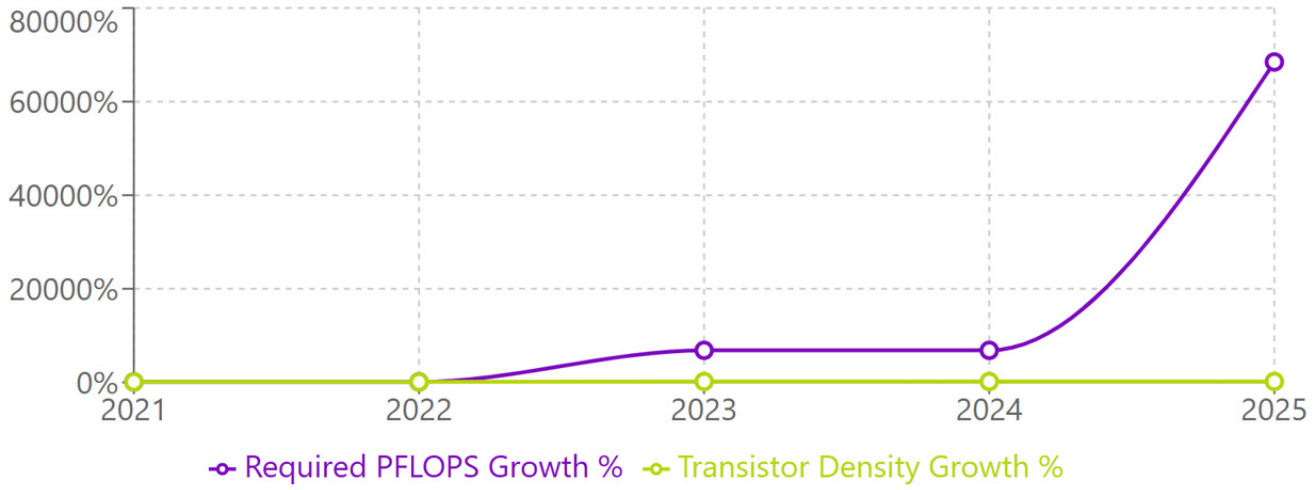

Figure 2: Growth in transistor density versus the PFLOPS required to train AI models from a 2021 baseline. By 2024, AI compute requirements surged by 6847%, while transistor density grew by only 183%. 2025 value is based on the projected PFLOPS required to train GPT-5 [10].

Still, chipmakers manage to keep up with GenAI advancements, which marks a departure from the traditional scaling model. In some cases, a chip can be 30 times faster than its predecessor, which was announced less than a year earlier [11]. Such relentless demands force chipmakers to constantly seek new ways to optimize their products.

3 | Critical Optimization Factors for GenAI Chipmakers

Today's GenAI arms race is fought with novel chip architectures and packaging. Specialized hardware designs are proliferating in the form of GPUs, TPUs, NPUs, and more, all tuned for parallelism and matrix-heavy AI math.

In this hyper-competitive landscape, chip vendors scramble to differentiate their products on multiple fronts. They promise some mix of better performance, efficiency, or scalability, but the specific strategies vary widely:

Performance

Standard Metrics: Key indicators in this domain include **PFLOPS, tokens per second, and ms/token**. These figures directly affect training time and inference latency.

Some chipmakers aim to outgun the competition with sheer performance. Flagship GPUs, for example, focus on FLOPS and huge memory throughput. While memory is a critical factor in GenAI performance, this paper focuses on compute throughput bottlenecks.

One approach that chipmakers employ to win this category is advanced packaging, connecting multiple silicon chiplets in a single heterogeneous device to increase performance density.

Even a 10% speed improvement will have a profound impact due to the immense scale. For example, training a model like LLaMA 3.1 405B involved 16,000 GPUs, consumed approximately 27 megawatts, and required an estimated 40 billion PFLOPS [12]. That level of optimization can **reduce training time by several weeks** and eliminate the need for thousands of GPU-days, **translating to millions of dollars** in infrastructure savings.

A 10% throughput optimization can reduce training time by several weeks and eliminate the need for thousands of GPU-days, translating to millions of dollars. It would also reduce inference latency, which is a critical factor in user experience.

In large-scale AI inference operations, even modest throughput enhancements can lead to significant cost reductions. For instance, OpenAI's GPT-4 processes approximately 50 billion queries annually, incurring an estimated $144 million in compute costs [13]. Implementing a 10% throughput improvement could decrease the number of required servers, resulting in an estimated $14.4 million in annual savings.

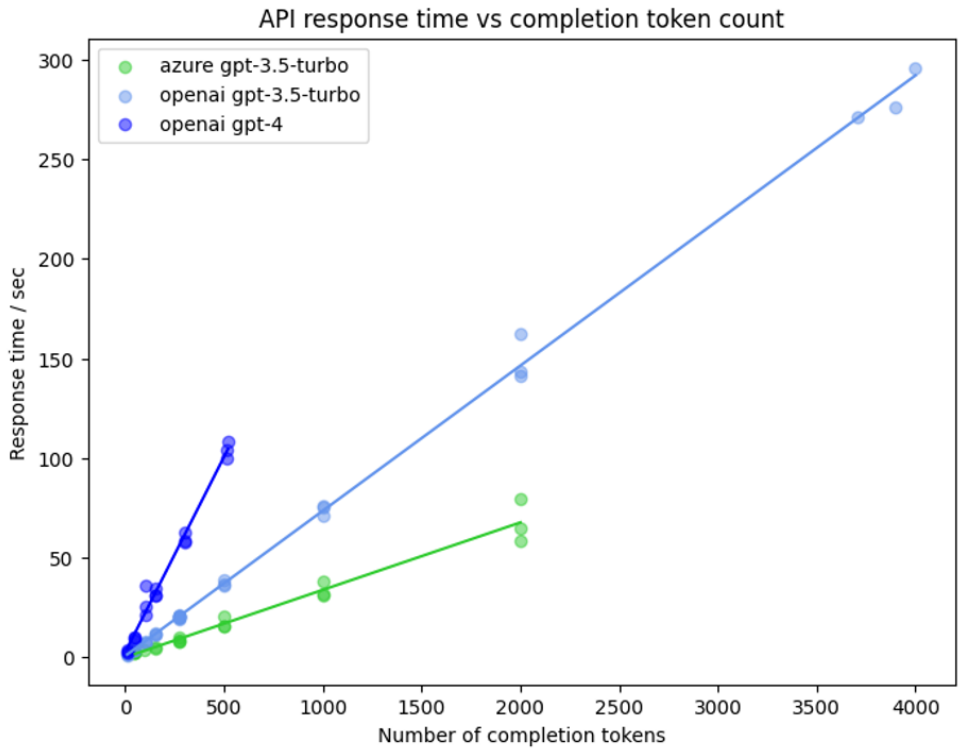

Figure 3: A dramatic increase in inference latency from 73 ms/token in OpenAI GPT-3.5-turbo to 196 ms/token in OpenAI GPT-4 [14].

Throughput optimization also reduces inference latency, which is a critical factor in user experience. For example, the response time of OpenAI's GPT-4 model has been measured at approximately 196 milliseconds per generated token [14]. Enhancing throughput by 10% could proportionally **reduce this latency**, leading to faster response times and improved user satisfaction.

Performance improvements typically begin with design-time architecture exploration and RTL optimization, such as pipeline depth, compute unit allocation, and dataflow design. On top of that, chipmakers apply techniques like standard Adaptive Frequency Scaling (AFS) to push efficiency under dynamic conditions in the field.

However, these runtime methods are generally static and not workload-aware, leading to suboptimal performance in real-world deployments. Frequency scaling is also done conservatively to preserve thermal and functional stability. While these approaches help extract more performance within safe limits, they may fall short of what GenAI workloads demand.

These runtime methods are generally static and not workload-aware, leading to suboptimal performance in real-world deployments. Frequency tuning is also done conservatively to preserve thermal and functional stability.

Power Efficiency

Standard Metrics: In this category, chipmakers optimize for **Performance-Per-Watt** (PPW), **performance-per-dollar**, or **Tokens-Per-Second-per-Watt** (TPS/W).

GenAI's exponential growth in computational requirements urges chipmakers to pay closer attention to power consumption. Beyond immediate consequences, such as thermal problems, excessive wattage has severe implications for customers' operational costs.

GenAI's exponential growth in computational requirements urges chipmakers to pay closer attention to power consumption, as excessive wattage has severe implications for customers' operational costs.

As a consequence, design wins increasingly revolve around **Total Cost of Ownership** (TCO). This metric factors in not only the upfront hardware cost but also ongoing expenses like power, cooling, and infrastructure. Solutions that deliver more compute per watt can significantly reduce TCO and make large-scale AI deployments more sustainable.

Furthermore, reducing the power consumption of individual devices directly expands infrastructure performance. Every watt saved per chip frees up headroom within the data center's fixed power budget, enabling higher system utilization across the fleet.

This power reduction allows operators to run more workloads, serve more users, or deploy additional systems without breaching energy limits. Improving PPW at the chip level becomes a strategic lever for maximizing performance within existing power constraints.

To explore how this dynamic plays out across real data center deployments, read the full blog post here.

$$PPW = \frac{\text{TFLOPS}}{\text{Watt}}$$

PPW can grow by increasing performance within the power envelope or by reducing wattage without impacting FLOPS.

Power efficiency is typically optimized through a combination of design-time techniques and runtime control. Clock gating, power gating, and multi-voltage domains are widely used at the architecture and implementation levels to reduce dynamic and leakage power.

At runtime, methods like Dynamic Voltage and Frequency Scaling (DVFS) and Adaptive Voltage Scaling (AVS) are applied to adjust power consumption based on static models or basic telemetry, such as temperature or process variation. These standard techniques are not workload-aware and typically apply uniform guard bands across all chips to ensure stability across all devices and workloads.

As a result, they leave significant excess guard bands that cause unnecessary power consumption, undermining PPW. This inefficiency calls for more precise, real-time approaches that optimize power without compromising performance or reliability.

Reliability

Standard Metrics: Defective Parts Per Million (**DPPM**), Silent Data Corruption (**SDC**) rate, and Mean Time to Failure (**MTTF**).

A chip's reliability at large scales is just as critical as its raw performance. DPPM measures the fraction of chips that exhibit failures post-manufacturing, directly impacting system uptime and operational costs. While semiconductor testing filters out detectable defects, latent issues stay hidden until real workloads expose them. As GenAI compute infrastructure scales to millions of deployed chips, even a low DPPM might translate to frequent failures with substantial consequences.

A single undetected error can compromise an entire training process as it distorts model weights across multiple interdependent nodes.

Furthermore, Silent Data Corruption (SDC) has emerged as a critical reliability threat to scaling GenAI training, as it corrupts computations without triggering alerts. Unlike memory bit flips, for example, mitigated by error correction codes (ECC), SDCs originate from subtle timing violations, aging effects, or marginal defects that escape standard semiconductor testing.

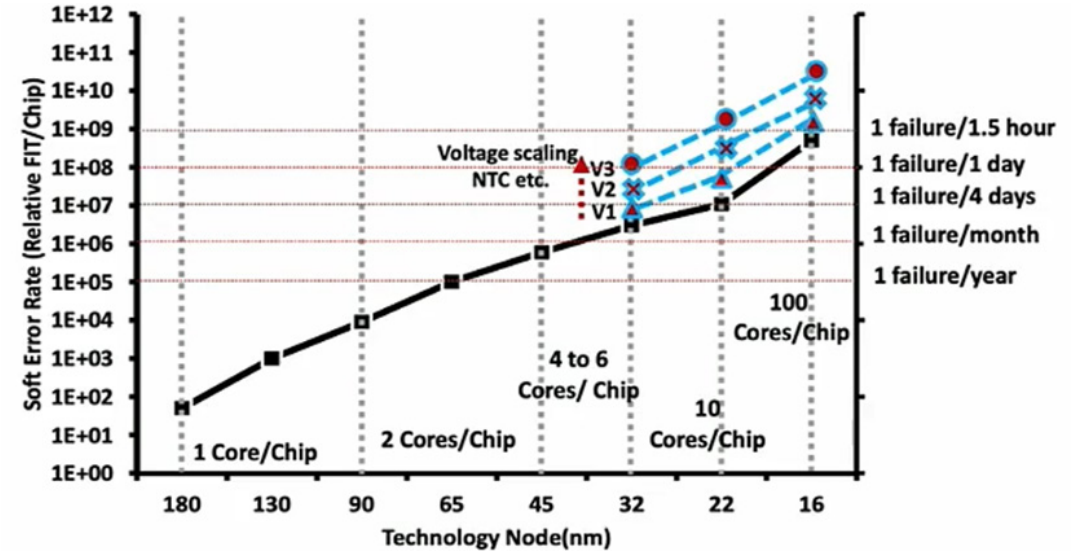

Figure 4: Soft errors such as SDC have increased from one failure per year @65nm to one failure every 1.5 hours @16nm [15]

These errors leave no trace, yet a single one can distort model weights across interdependent nodes, quietly derailing a training run that may span weeks, involve over 25,000 GPUs, and cost more than $100 million [1]. In training clusters, even a single faulty processor can jeopardize the entire job. These workloads run across tightly coupled systems, each contributing to shared model parameters. If one chip introduces a silent error during synchronization, that corruption spreads throughout the cluster.

SDCs are also far more frequent than engineers would expect. Meta reported failures every three hours in a 16,000-GPU cluster [4]. Alibaba measured 361 DPPM in its cloud infrastructure [16]. Worse still, IEEE studies show a dramatic rise in soft error rates, from one failure per year at 65nm to one every 1.5 hours at 16nm [15]. Even with massive investments in validation and testing, undetected faults challenge the foundational assumption of silicon reliability in fleet-level AI deployments.

Ensuring reliability has traditionally relied on periodic field testing to uncover potential failures. While effective for basic quality assurance, these methods may miss latent defects, workload-driven faults, accelerated aging, and SDCs. They are also time-consuming and difficult to streamline within data center environments running high-intensity GenAI. The limitations of these offline techniques point to the need for continuous, in-situ monitoring to maintain reliability at hyperscale.

Despite these diverse optimization strategies, all chipmakers share a common challenge. They must set conservative operating guard bands to ensure reliability. This necessity presents an overlooked opportunity for significant optimization that can shape who wins the GenAI race.

4 | Maximizing Performance, Power, and Reliability Gains with Workload-Aware Monitoring On-Chip

Current methods for optimizing performance, power, and reliability all share the same blind spot: they don't see how chips behave under actual workloads in the field. GenAI cloud operators pay for this lack of real-time visibility through higher power draw, lower throughput, and increased risk of failure.

Performance tuning relies on static margins. Power controls are triggered by basic telemetry. Reliability checks happen too late, after failure is already underway. None of these approaches adapts to actual stress and environmental conditions during live operation.

That's the gap.

To close it, chipmakers need a way to observe what's really happening inside the chip as it runs GenAI workloads. Not averages. Not design-time assumptions. Accurate real path behavior under actual workload, voltage, temperature. In every chip. In every rack.

This level of visibility makes it possible to optimize with precision instead of caution.

With accurate in-situ monitoring, excessive voltage and frequency guard bands can be safely reduced, freeing up power and performance that would otherwise be wasted. The same visibility allows for early detection of marginal behavior before it causes faults, SDC or system-level failures.

Excessive voltage and frequency guard bands designed for worst-case scenarios can weigh down power and performance.

Standard techniques like canary circuits fall short here. They attempt to detect potential timing violations by replicating critical paths in fixed locations, but those replicas don't age like the real design. They don't follow actual switching activity or reflect real workload stress.

To compensate for these limitations, chipmakers are forced to add excessive guard bands designed for worst-case scenarios. The resulting safety margin may prevent failures, but it also weighs down power and performance.

That tradeoff is no longer viable at GenAI's relentless pace.

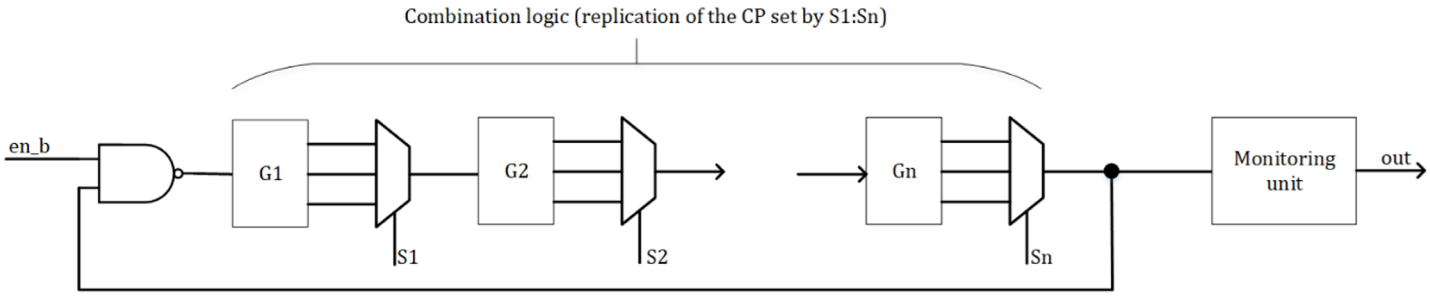

Figure 5: A canary circuit that monitors design margins is a critical path replicator, which cannot provide accurate data about actual performance limiter paths' timing.

What's needed is accurate, on-chip timing margin monitoring that adapts to real workloads in real time. That's the foundation for dynamic guard-band control. And it's the missing link for achieving reliable scalability while reclaiming trapped performance and power.

5 | proteanTecs Real-Time Monitoring for Scalable GenAI Chips

proteanTecs real-time monitoring solutions dynamically track and reclaim the guard bands of each individual chip to unlock hidden opportunities for power, performance, and reliability gains. Unlike canary circuits, proteanTecs uses on-chip Agents that provide parametric measurements in-situ and in functional mode, to detect timing issues, operational and environmental effects, aging and application stress. Among the suite of Agents are the Margin Agents that monitor timing margins of millions of real paths for more informed decisions.

proteanTecs tracks and reclaims the guard bands of each individual chip in real time to unlock hidden opportunities for power, performance, and reliability gains.

Margin Agents provide very high coverage of the design's logic and monitor the real performance-limiting paths that traditional methods often miss. This method allows precise action based on real workloads, aging, and IR drops.

The Agents are lightweight hardware IP blocks distributed across the chip. They require negligible silicon area and power, ensuring the monitoring does not interfere with the chip's operation or cost.

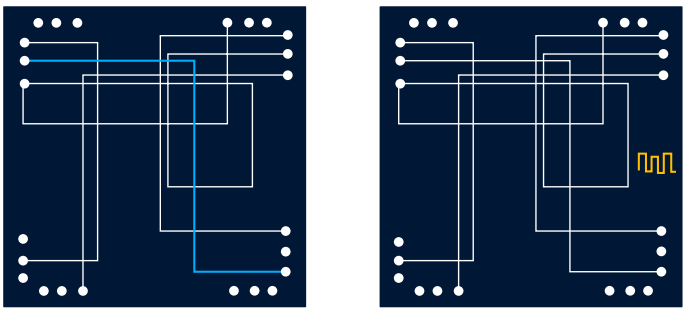

Figure 6: Unlike canary circuits (right, in yellow), proteanTecs uses on-chip Margin Agents (left, in blue) that monitor true critical paths.

proteanTecs provides a suite of AI-tuned applications to monitor infrastructure in the field:

proteanTecs AVS Pro™ Dynamic Voltage Scaling with a Safety Net

Monitors timing margins of performance-limiting paths to safely reduce voltage guard bands.

KEY BENEFITS: Higher PPW, energy cost reduction, lower TCO, reduced heat, and less cooling with up to 14% less power.

proteanTecs RTHM™ Performance and Health Monitoring

Ensures continuous performance and health in real time, providing failure avoidance and predictive maintenance.

KEY BENEFITS: Shorter downtime and reduced need for redundancy by avoiding chip failure.

proteanTecs AFS Pro™ Adaptive Frequency Scaling

Safely reduces frequency guard bands by dynamically adjusting clock speeds under actual workloads.

KEY BENEFITS: Higher FLOPS and more headroom with up to 8% better performance.

6 | proteanTecs AVS Pro™ - Dominating PPW Through Safer Voltage Scaling

Advanced chips often carry excessive voltage guard bands to withstand worst-case conditions, which can increase energy consumption. Yet GenAI accelerators must deliver ample PPW to handle demanding workloads within the power envelope. To help meet that goal, proteanTecs AVS Pro safely reclaims excess guard bands, which would otherwise weigh down overall efficiency.

AVS Pro improves PPW and reduces TCO by saving power with zero performance loss. It knows exactly how much voltage headroom each chip needs, at any given time, and cuts down the rest.

AVS Pro improves PPW and reduces TCO by saving power with zero performance loss. It knows exactly how much voltage headroom each chip needs, at any given time, and cuts down the rest.

Unlike standard canary circuits and fixed voltage guard bands, proteanTecs AVS Pro employs accurate in-chip, in-situ timing margin monitoring. This solution combines on-chip agents with dedicated algorithms for informed decisions. AVS Pro allows precise guard-band tuning based on current workloads, real aging, and actual IR drops to reduce more power in real-life scenarios while ensuring reliability. It dynamically finds the lowest stable voltage for the current conditions in real time as workloads and environmental factors change.

By employing AVS Pro, leading cloud hyperscalers **have seen an 8-14% power reduction** in their electronics, resulting in significant cost reduction and enhanced performance.

To learn more about improving PPW and reducing TCO by reclaiming excess voltage guard bands, read the full white paper here.

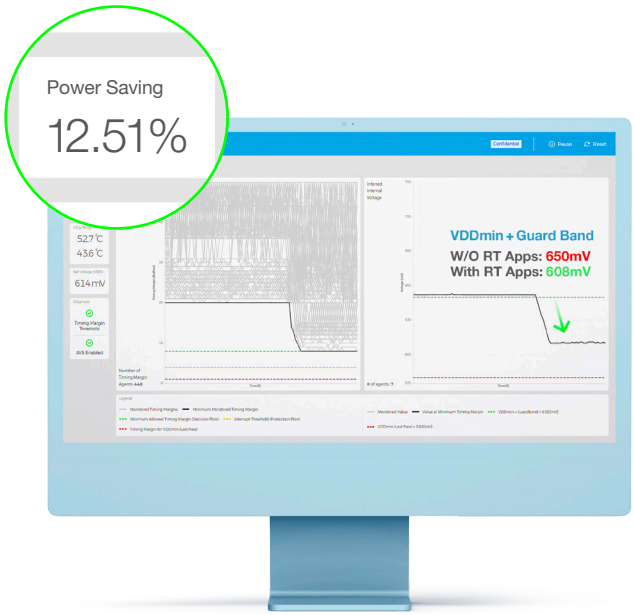

Figure 7: AVS Pro safely reduces voltage guard bands, enabling substantial PPW improvement.

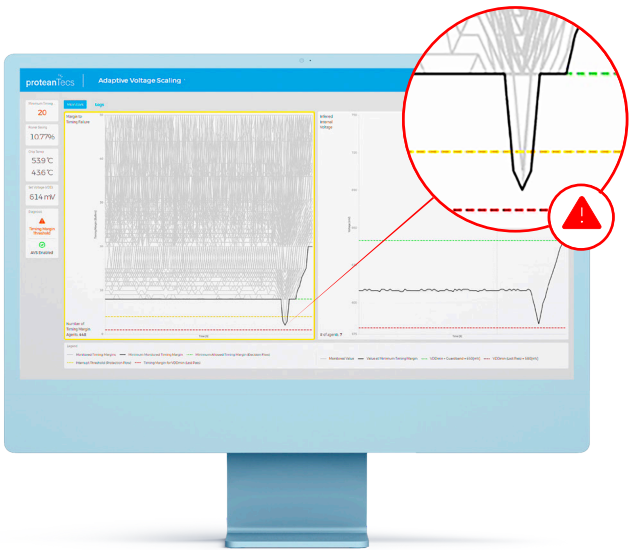

Figure 8: The AVS Pro protection layer provides real-time failure prevention.

proteanTecs AVS Pro can detect risk conditions that require raising the voltage to prevent it from dropping too low. This interrupt-based safety net gives confidence as AVS Pro continuously tunes the voltage to the lowest level the chip can safely sustain. The result is reliable execution with lower power draw and reduced thermal stress, improving TCO through savings on electricity and cooling.

In-situ monitoring of the true paths per chip is paramount, as the performance-limiting paths can change over time according to workload stress, operating conditions, and aging of individual devices.

AVS Pro also delivers long-term value by slowing silicon wear-out. By reducing the nominal voltage throughout a chip's life, AVS Pro lowers power consumption and temperature, reducing the stress on the SoC and prolonging its lifespan. As the chip ages, the margins are optimized accordingly to ensure reliability while maximizing efficiency. In one deployment, a 5nm data center communications SoC running AI workloads extended its useful life by 18% thanks to AVS Pro's voltage optimizations. This optimization directly translated into fewer hardware replacements and lower TCO for the operator.

To learn more about reducing TCO by extending chip lifetime through safer voltage scaling, read the full white paper here.

7 | proteanTecs RTHM™ - Flagging Cluster Risks Before Failure

GenAI clusters can be derailed by a single failing chip that injects errors. A subtle error in one GPU during a large model training run can silently corrupt results. Traditional methods often miss these silent data corruptions, which can then propagate for weeks. By the time an issue surfaces, it may take months of tedious debugging to find the root cause of the problem.

proteanTecs RTHM continuously tracks the health of each chip using agents that measure timing margins of millions of real paths. It monitors timing margin measurements affected by both intrinsic and extrinsic reliability issues, as well as unanticipated voltage drops and usage patterns, to assign each device a dynamic "performance index." If that index falls below a certain threshold, RTHM flags the chip long before it fails, allowing preventive action.

RTHM detects the slightest timing margin erosion due to aging, unexpected workload stress, or latent defects, catching anomalies before they cascade. This predictive monitoring can alert operators days or weeks before a crash or SDC occurs. It is much more accurate than canary circuits or periodic built-in self tests (BIST). Canary circuits fall short as they do not monitor the real paths while BISTs are destructive pass/fail tests that cannot run in real-time while the device is operating. They are not predictive, as the failure is detected after the fact.

With training runs that can cost $1 billion or more, the early detection that RTHM provides is essential.

The result is improved fleet-level reliability in large-scale AI training clusters. Even a single silent data error can spread across compute nodes and corrupt model weights. Such errors can invalidate weeks of work, wasting thousands of GPU hours and millions of dollars. With training runs that can cost $1 billion or more [2], the early detection that RTHM provides is essential.

To learn more about enabling reliable scalability by real-time health monitoring, read the full white paper here.

Figure 9: Visualization of RTHM in a 5nm chip: real-time indication of a severe margin drop in a device that might cause SDC.

8 | proteanTecs AFS Pro™ - Capturing Frequency Headroom for Higher FLOPS

proteanTecs AFS Pro enables safely increasing clock speed without compromising reliability. This capability is crucial because chips are shipped with conservative frequency guard bands that account for many worst-case effects, leaving performance on the table.

This unused headroom is a missed opportunity, especially for GenAI workloads that are intensely FLOPS-bound. Every extra MHz translates directly to faster model training and inference.

AFS Pro determines how much spare timing margin exists under current conditions, giving each device a tailored frequency that safely maximizes its potential.

Based on real-time timing margin monitoring, AFS Pro raises the clock speed dynamically to push the device closer to its unique threshold when needed, effectively reclaiming excess frequency guard bands. **If timing margins fall below a predefined threshold, AFS Pro readjusts instantly to ensure stability.**

AFS Pro raises the clock speed dynamically to push the device closer to its unique threshold, effectively reclaiming excess frequency guard bands.

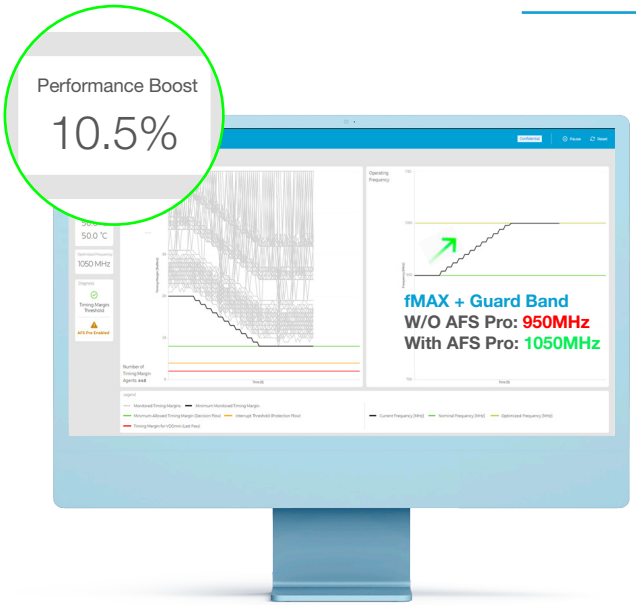

Figure 10: Reclaiming frequency headroom: Without AFS Pro the chip's max stable frequency was 950MHz. With AFS Pro it safely reached 1050MHz, enabling a ~10% FLOPS boost.

This method safely extracts more FLOPS from existing silicon. One cloud chipmaker saw AFS Pro achieve a 10.5% frequency increase in production. That gain delivered a proportional jump in throughput with the same hardware while maintaining perfect stability. In GenAI terms, that means about 10% more PFLOPS, allowing faster execution, more concurrent users, or reduced cost by meeting targets with fewer chips.

9 | System-Wide Workload and Operational Monitoring

In addition to optimizing individual chips during runtime operation, proteanTecs enables system-wide offline orchestration.

The company provides applications for on-board performance and health monitoring, usage profile monitoring, and operational monitoring. These enable advanced local and remote diagnostics with pinpoint root cause analysis, performance degradation monitoring with historical logs for predictive maintenance, and HW/SW correlation for co-design optimization.

Figure 11: proteanTecs on-board software running in a 5nm Datacenter chip.

Users gain logs that track degradation trends – based on time intervals, thresholds, and a Health Index. The solution sets smart, configurable thresholds to trigger diagnostics before a failure occurs, and provides remote diagnostics to reduce on-site service requirements, with detection of the probable source for field debugging. It suggests actionable responses, including system replacement or load balancing.

10 | Conclusion

GenAI's explosive pace is shattering the semiconductor landscape. Winning this gold rush is not defined by a single category, such as process nodes or FLOPS. Chipmakers must compete on multiple fronts, including performance, efficiency, and reliable scalability.

As this white paper shows, those who monitor, optimize, and predict chip health and performance gain a clear advantage. proteanTecs' in-field monitoring applications were built to provide that edge.

Systems are monitored in real-time to assure optimal power efficiency, latency, and resilience. In addition, operators can proactively shift workloads across servers based on chip health and predicted performance degradation, balancing stress to offload devices nearing failure. This allows for smarter load balancing, early intervention, and improved fleet reliability, before performance issues or faults occur.

GenAI design-wins favor silicon that runs at its true limits, making every PFLOPS and kW count. proteanTecs helps chipmakers achieve that by turning visibility from within the device into competitive gains.

11 | References

[1] GenSpark AI. (2024). How many GPUs are needed to train GPT-4-O?

[2] Amodei, D. (2024). The Billion-Dollar Price Tag of Building AI. TIME.

[3] Cohen, A. (2024). AI Is Pushing The World Toward An Energy Crisis. Forbes.

[4] Meta AI. (2024). The Llama 3 Herd of Models.

[5] IO Fund. (2024). AI Power Consumption Rapidly Becoming Mission Critical. Forbes.

[6] Business Insider. (2023). How Much Does ChatGPT Cost to Run? $700K/day, Per Analyst. Business Insider.

[7] Lonebull. (2025). Slow Performance and Unresponsiveness of ChatGPT Web Browser Version. OpenAI Developer Community.

[8] Kumparak, G. (2015). Moore's Law stutters as Intel's tick-tock skips a beat. The Verge.

[9] Tom's Hardware. (2024). TSMC N3P & N4X on Track with Density and Power Gains.

[10] Mollick, E. (2024). Scaling: The state of play in AI. One Useful Thing.

[11] NVIDIA. (2024). GB200 NVL72: HPC & AI GPU for Data Centers.

[12] Epoch. (2024). Can AI Scaling Continue Through 2030?

[13] Al Rifaei, A. (2024). Uncovering the Hidden Costs of AI Queries. LinkedIn.

[14] Pungas, T. (2023). GPT-4 API response time measurements. OpenAI Developer Forum.

[15] Upasani, G., Vera, X., & Gonzalez, A. (2015). A case for acoustic wave detectors for soft errors. IEEE Transactions on Computers, 65(1), 5–18.

[16] Wang, S., Zhang, G., Wei, J., Wang, Y., Wu, J., & Luo, Q. (2023). Understanding Silent Data Corruptions in a Large Production CPU Population. In Proceedings of the 29th Symposium on Operating Systems Principles (pp. 216–230).

CONTACT US TODAY

To learn more about our health and performance monitoring solutions

→ Schedule a Demo

proteanTecs

proteanTecs Ltd.
www.proteanTecs.com