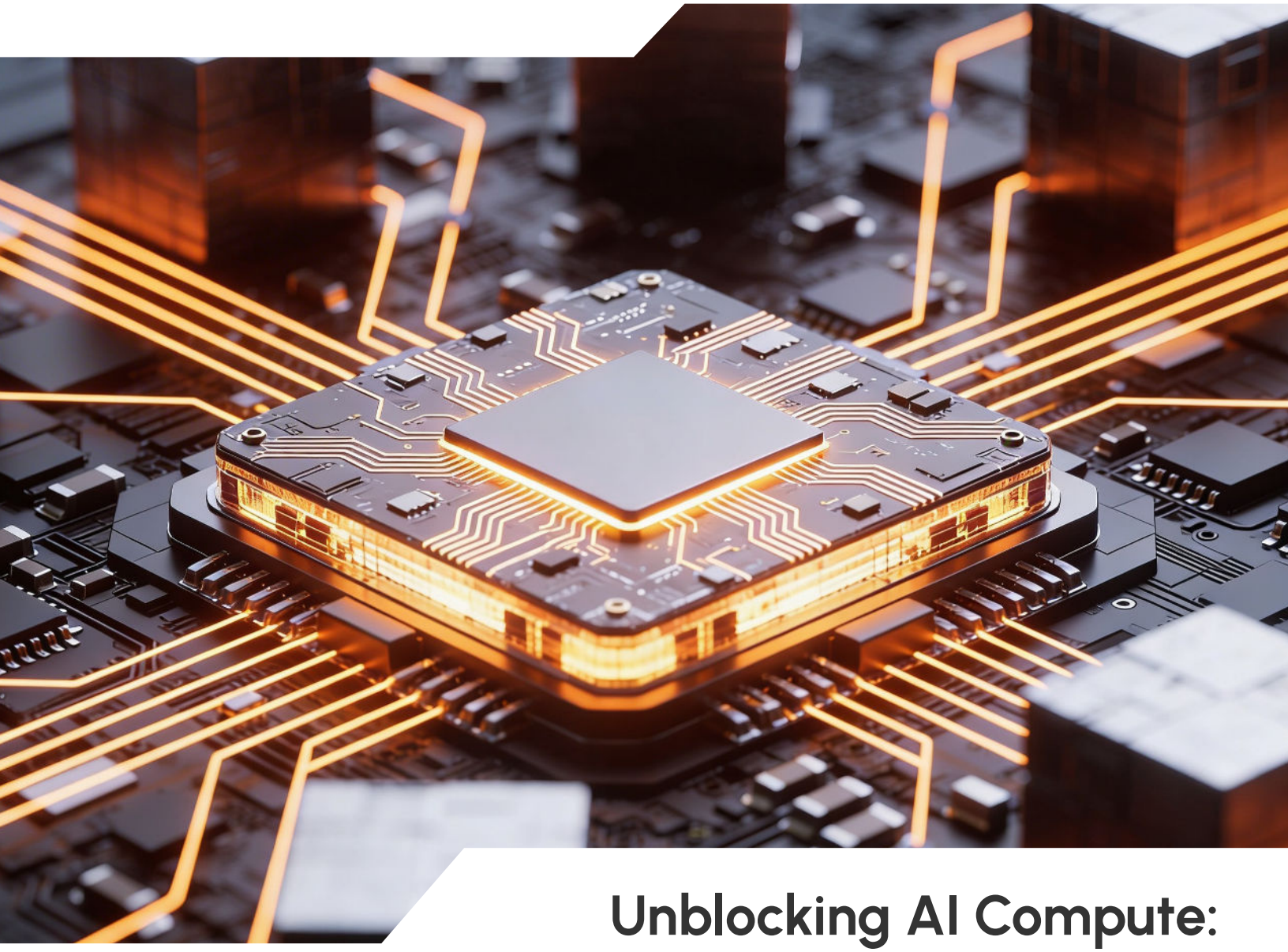
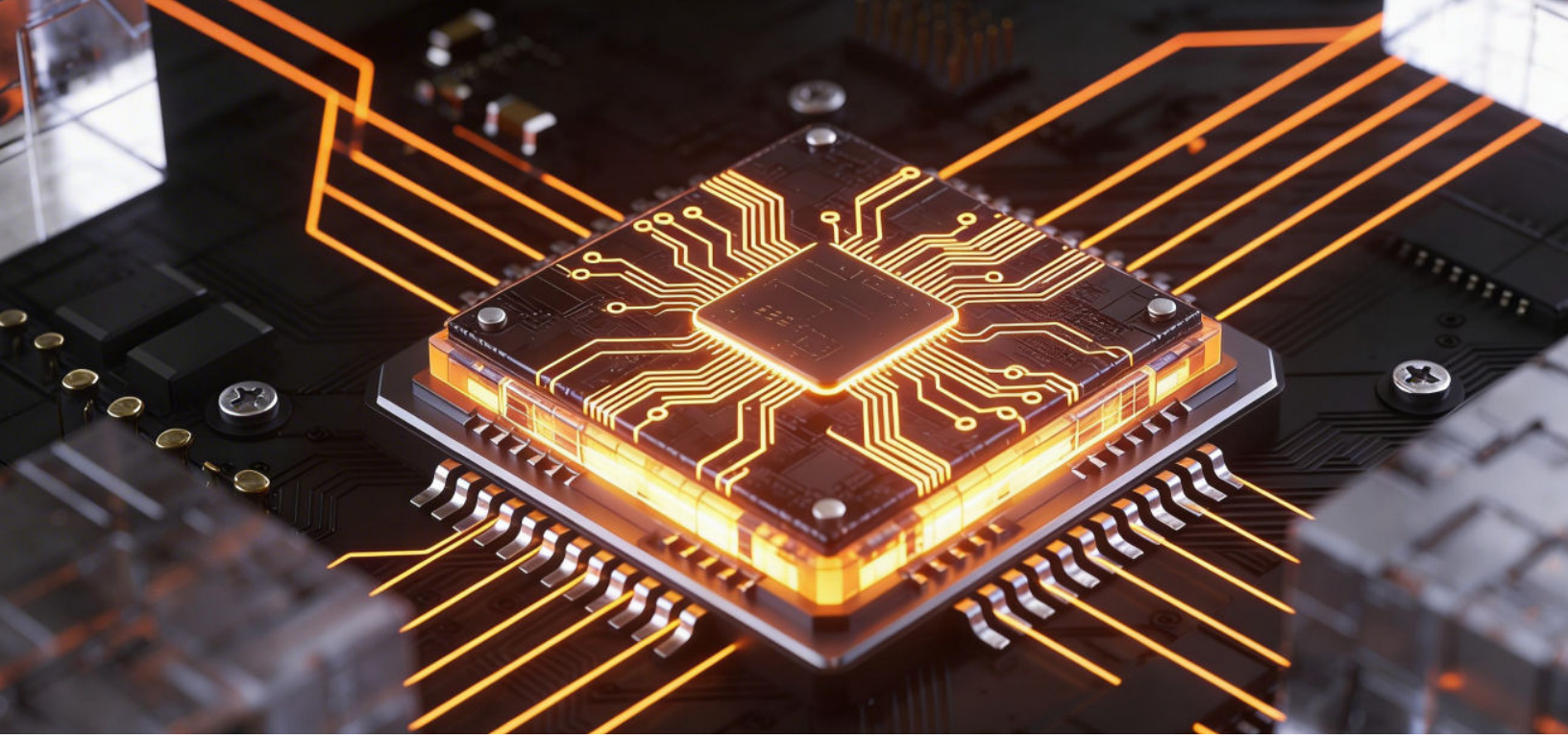
Futurum®

SiFive



Unblocking AI Compute: SiFive Intelligence's Open Solution for Edge to Cloud Scale

APRIL 2026



1. Executive Summary

The AI computing landscape is fragmented by a mix of instruction set architectures (ISAs) that force a choice between high-performance complexity and low-power simplicity. SiFive breaks this paradigm by delivering a unified family of architectures based on the RISC-V ISA spanning the entire compute spectrum from milliwatt edge sensors to megawatt hyperscale data centers. This open architectural foundation ensures software interoperability and code reuse, allowing developers to build once and deploy everywhere, even as AI models evolve continuously.

The primary bottleneck for modern AI is shifting from raw arithmetic throughput (FLOPs) to data movement and memory bandwidth. Transformer-based LLMs rely on an optimization technique known as Key-Value (KV) caching, which stores tensors for low-latency reuse. During the decode phase of inference, the size of LLM models the length of context windows means that model parameters and KV caches routinely exceed GPU high bandwidth memory capacities and then saturate off-chip bandwidth. Legacy instruction sets and memory hierarchies have not adapted to the demands of modern AI workloads due to:

- **GPU Inefficiencies**: While dominant in the market, GPUs carry a significant silicon tax in the form of massive register files and complex context management logic required to hide latency through multithreading. When complex scheduling can't keep cores fed, many units sit idle and real-world utilization drops.
- **CPU Cache Thrashing:** Massive matrix operations evict critical control data from limited caches, requiring access to off-chip memory.

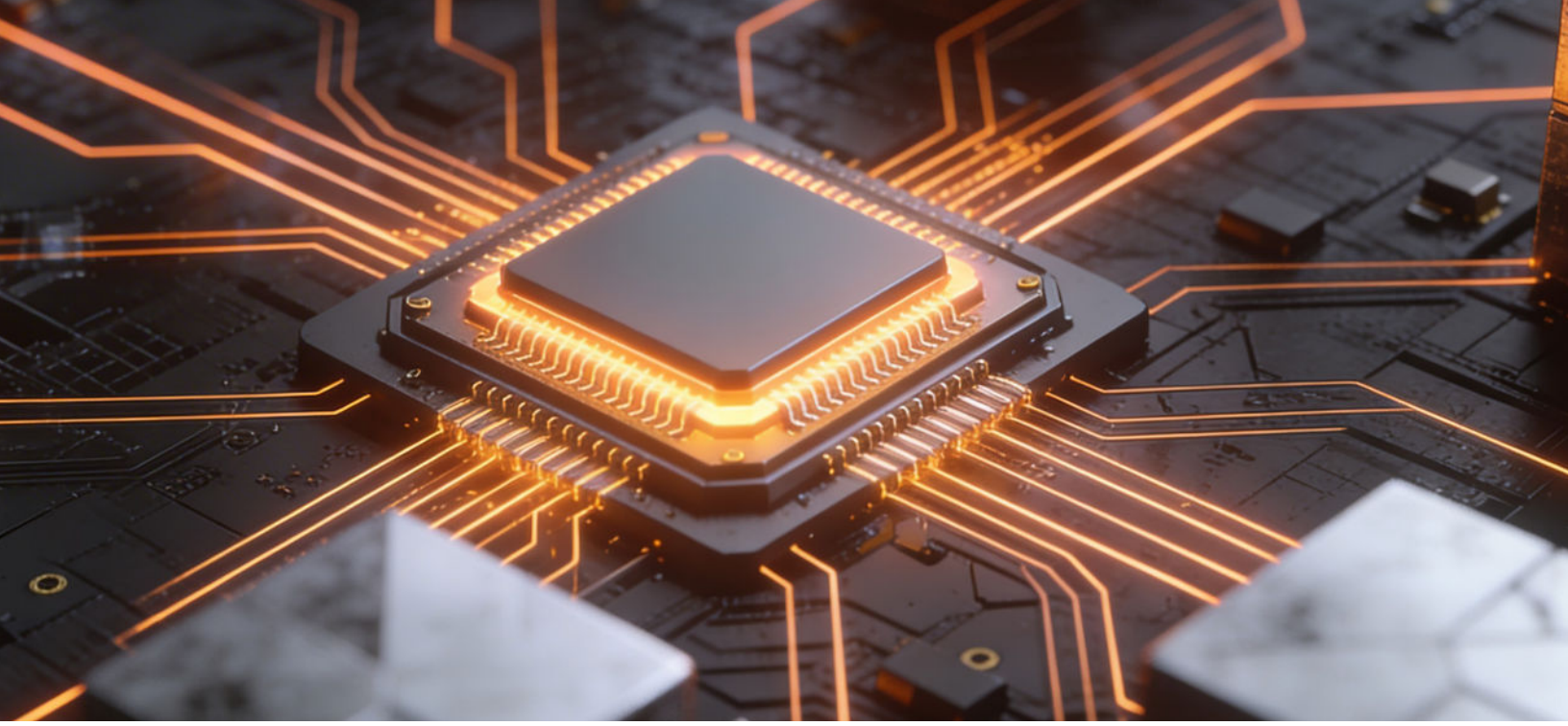
SiFive Intelligence solutions provide a fundamental breakthrough in how data is processed and delivered:

- **Loosely Coupled Vector Architecture:** By having some separation between the scalar pipeline and the vector pipeline, the scalar engine can resolve branches and unroll loops while the vector engine focuses on heavy data processing for more efficient compute.
- **Memory Latency-Hiding Queues:** A configurable Vector Load Data Queue (VLDQ) can eliminate most pipeline stalls that typically impact AI performance.

- **Cache Efficiency:** Selective cache bypassing ensures the L1 cache remains dedicated to control flow, while the vector unit achieves a perceived one-cycle load-to-use latency and the scalar pipeline can achieve zero load-to-use.
- **Standard Software Support:** Without needing to write specialized kernels or complex DMA engines to move data between memory and accelerators, developers have a familiar C/C++ programming experience.
- **Superior Edge AI Performance:** Up to 2x the inference performance of CPU competitors for common edge AI tasks

As noted by Futurum Research, the industry is in the midst of an unbundling phase of one-size-fits-all AI servers. Hyperscalers and Tier 1 design leaders are seeking to own their compute destiny via tight coupling of high-volume workloads and custom silicon. NVIDIA's recent licensing of Groq's IP shows that deterministic, single-threaded computing can be more efficient and performant for high-speed inference. SiFive's AI IP offerings represent a compelling set of new choices for AI builders seeking:

- **Performance & Power Efficiency:** Solving the memory wall at the pipeline level
- **Cost Diversity:** Driving price-performance gains for high-value accelerators with cost-effective CPUs
- **ISA-Level Interoperability:** Moving beyond closed ecosystems and one-shot designs to an open, composable toolbox for long-term ISA compatibility



2. Market Context: Why AI Compute Needs a New Architectural Path

The Scale Problem: LLMs Break Traditional CPU Architectures

The primary bottleneck for LLM inference is no longer the speed of arithmetic operations, but the bandwidth available to feed data to the compute units. The decode phase of large-batch inference remains persistently memory-bound, with DRAM bandwidth saturation emerging as the critical constraint preventing full utilization of GPU compute capabilities.

The fundamental issue lies in the sheer size of the models and the complex data structures they generate. Large Language Models require immense memory capacity not just for model weights, but for the Key-Value (KV) cache, which stores the state of the attention mechanism across the context window. When serving multiple users with long context windows, the memory requirements quickly exceed GPU capacity, compounded by a shortage of high-bandwidth memory.

Standard CPU and GPU architectures rely on complex hierarchies of caches (L1, L2, L3) managed by control logic. However, the massive matrix traversals required by LLMs often lead to severe cache thrashing, where data blocks are evicted just before they are needed again, leading to high latency and energy waste.

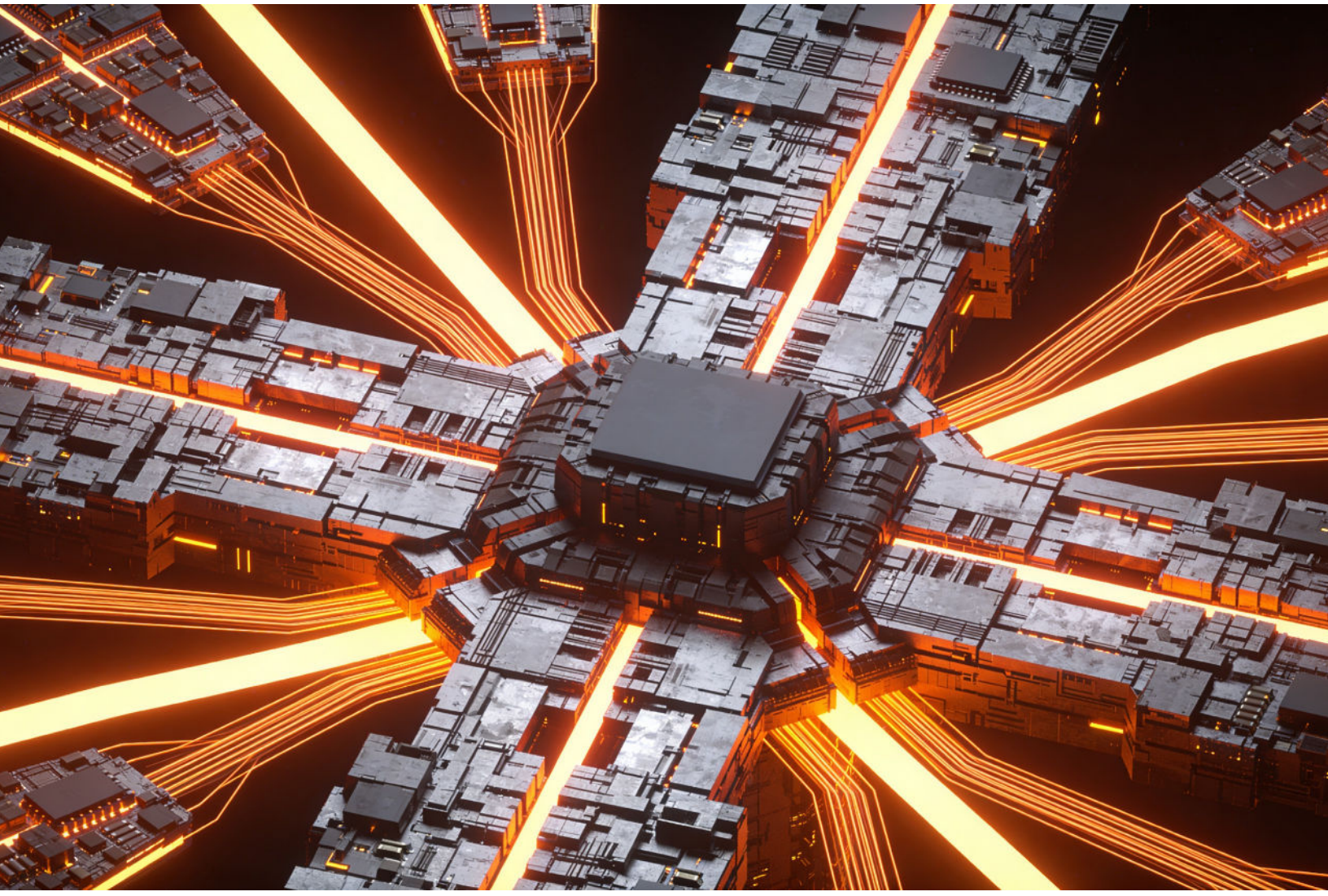
The GPU Dominance Era—and Its Limits

The dominance of the GPU was established on the premise of high-performance compiler software. However, the specific characteristics of generative AI have exposed fundamental inefficiencies in the GPU model. Mixture of experts (MoE) models only activate a small fraction of their parameters for any single input, yet the sheer size of the total model necessitates that the entire massive set of weights, most of which remain idle, reside in memory, creating a significant bandwidth bottleneck. GPUs are not optimized for the control logic needed to activate these weights, causing them to wait for instructions from the host CPU. During the decode phase, the processor must load the model weights from off-chip memory for every single token generated. During the decode phase, GPU FLOP utilization rates can fall to 10% or lower, as valuable GPUs wait for model weights to be loaded from off-chip memory.

Industry Motivation to Diversify Compute

Recognizing that off-the-shelf GPUs cannot meet the specific efficiency, scalability, and economic requirements of their internal workloads, the major hyperscalers have decoupled from the merchant silicon roadmap. In particular, Google's strategy highlights a crucial trend: the adoption of RISC-V to control custom silicon. The Google TPU utilizes the SiFive Intelligence X280 with VCIX to provide flexible control and scalar/vector processing, enabling an elegant division of labor where the custom Matrix Unit (MXU) accelerator can focus entirely on massive matrix calculations. This results in a highly efficient, custom-tailored chip that is optimized specifically for AI workloads.

RISC-V offers architectural modularity that legacy ISAs cannot match, allowing designers to add custom extensions for scalar, vector, and matrix operations without breaking software compatibility. RISC-V Vector extensions (RVV) utilize a vector-length agnostic (VLA) approach. A single instruction can process a variable number of elements controlled by the vlen register without the overhead of thread management. RVV architectures, particularly those with decoupled vector pipelines, can handle these irregular data patterns more efficiently by utilizing predication and vector masking to disable specific lanes without the penalty of thread divergence. This modularity is proving essential for AI ASICs to minimize memory latency.

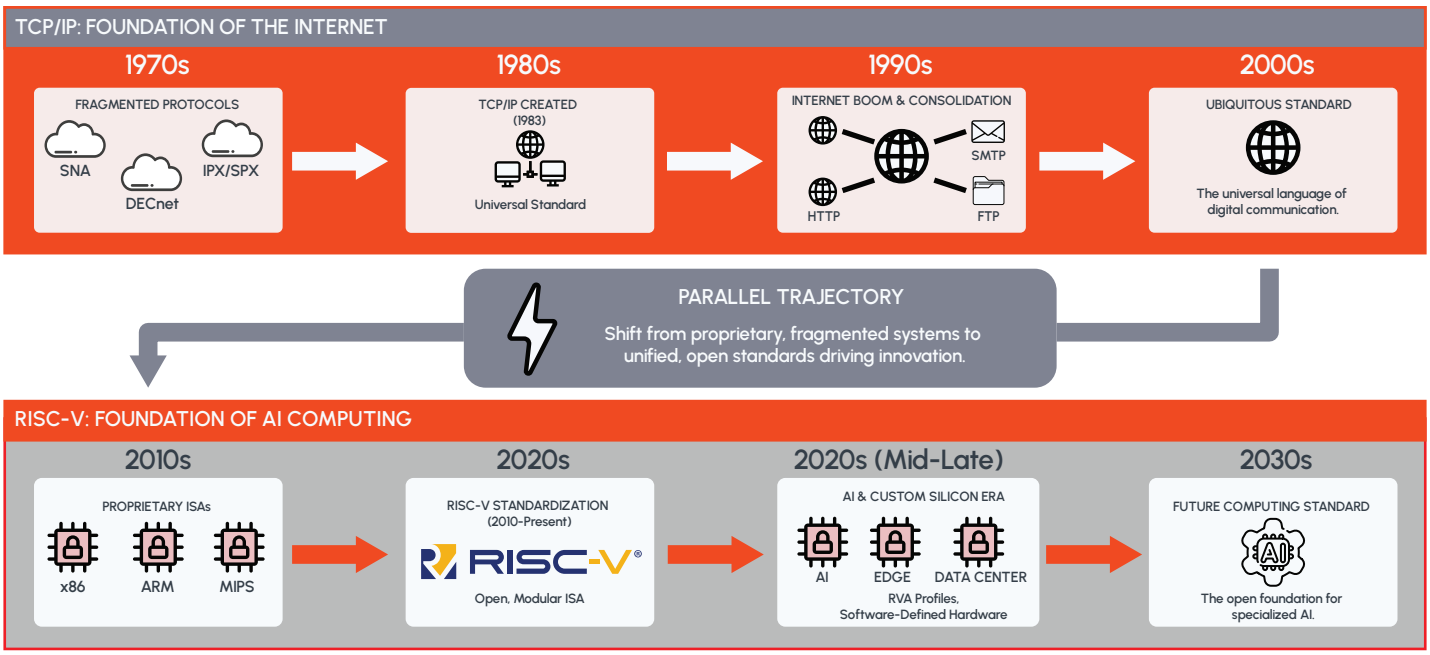


3. The RISC-V Advantage: A Foundation for AI Innovation

RISC-V is in a similar position to TCP/IP in the "protocol wars" of the early internet era. Before the open internet, digital communication was fragmented by proprietary networking protocols like IBM's Systems Network Architecture (SNA) and Novell's IPX/SPX. These systems functioned as walled gardens, forcing companies to buy entire vertical stacks from a single vendor to ensure compatibility. TCP/IP shattered those silos by creating a universal, open standard for data transmission (see Figure 1).

Like TCP/IP, RISC-V is dismantling the vendor lock-in of proprietary instruction sets in the AI era. The open-standard nature of RISC-V collapses the traditional wall between hardware and software teams. Because the ISA IP is available to software engineers long before a chip is ever manufactured, they can build production-ready toolchains that are waiting for the silicon the moment it returns from the fab. This synchronization transforms silicon from a multi-year project into an agile asset, allowing hyperscalers to deploy custom AI accelerators at the speed of software innovation rather than the slow pace of legacy hardware roadmaps.

Figure 1. The Evolution of Open Standard: TCP/IP vs. RISC-V

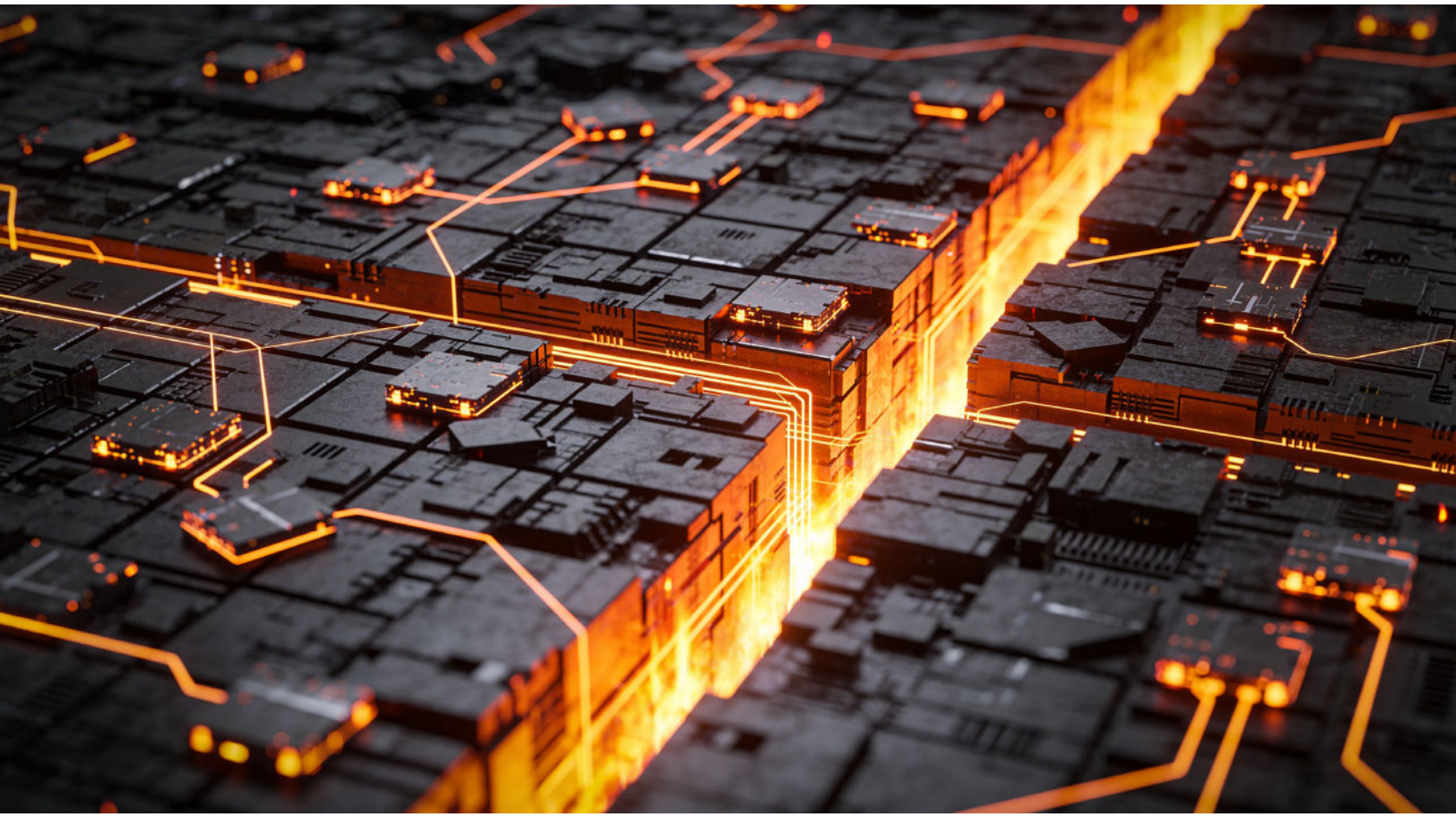


Source: Futurum Research

RISC-V is a key element of the shift from a one-size-fits-all monolithic processor design to a composable toolbox. The recently ratified RVA23 profile provides a standard software foundation that bundles essential extensions into a mandatory feature set while allowing for deep differentiation. Within this profile, the RISC-V Vector Extension (RVV) handles essential non-linear AI operations like Softmax and GELU while remaining agnostic to vector length. To build on this scalable foundation, RISC-V task groups use rigorous workload analysis to ratify new extensions, ensuring that hardware evolution addresses actual software bottlenecks, including the prefill and decode phases of AI inference. Architects can shape and select specific microarchitectural implementations for matrix math to comply with their workload requirements.

RISC-V is rapidly closing the historical gap in software maturity. The alignment of Red Hat Enterprise Linux 10 and Ubuntu with the RVA23 application profile marks the transition from experimental architecture to a stable, production-ready OS foundation for mission-critical AI applications. At the same time, new AI compilers are automating the complex math needed to bridge software and silicon. The result is a system where hardware is specialized for performance, but software remains flexible and easy to update.

Perhaps the strongest validation of this open model comes from the industry giant. NVIDIA has moved from merely observing RISC-V to actively integrating it, shipping over a billion RISC-V cores to replace their proprietary Falcon microcontrollers that manage GPU operations. Building on this usage, NVIDIA has announced intentions to port its dominant CUDA AI acceleration stack to the RVA23 profile. NVIDIA also participates in the RISE Project, working on optimizing RISC-V software with a group of industry leaders. This is a watershed moment analogous to Microsoft embracing Linux, signaling that RISC-V has graduated from a microcontroller alternative to a central orchestrator in the world's most advanced accelerated computing environments.



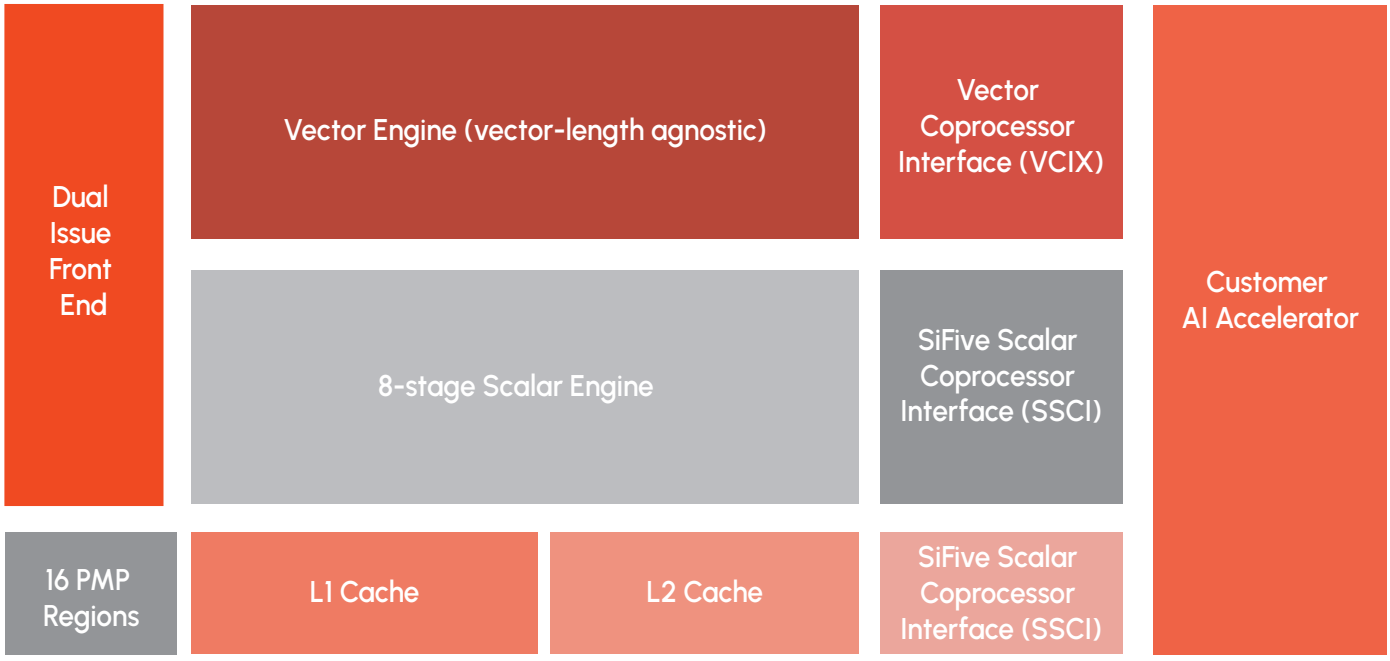
4. SiFive's Architectural Breakthroughs: Decoupled Vector Machines and Latency-Hiding Queues

Decoupled Vector Architecture: A Fundamentally Different Approach

Traditional CPU designs fail in AI once weights exceed cache capacity. AI workloads involve massive weight tensors that quickly exceed memory cache capacities, forcing the processor to fetch from slower, more distant memory layers. In standard monolithic designs, the physical distance between the execution units and the unified L2 cache creates a non-trivial wire delay. For AI, where data must be fed continuously to the pipeline, these stall cycles become a bottleneck, eliminating the benefits of high clock speeds and lowering hardware utilization.

To better utilize limited memory caches, the SiFive Intelligence architecture loosely decouples the scalar pipeline (control) from the vector pipeline (compute), allowing the scalar unit to "race ahead" and pre-fetch data. In SiFive's 2nd Generation Intelligence family (see Figure 2), the scalar unit identifies vector load instructions and issues memory requests to the L2 cache or memory system immediately, long before the vector unit is actually ready to use that data. By the time the vector unit is ready to compute, the data has already returned and is waiting in a configurable Vector Load Data Queue (VLDQ), which acts as a high-speed staging area. This results in a perceived load-to-use latency of one cycle, effectively masking the memory stalls that typically bottleneck AI performance.

Figure 2. Indicative SiFive 2nd Generation Intelligence Family Processor Architecture



Source: Futurum Research

The primary challenge of modern AI is memory latency. GPUs address this by switching between thousands of threads to hide stalls. While effective, this approach incurs a silicon tax in the form of massive register files and complex context management logic. The decoupled SiFive approach achieves similar results with a deterministic instruction flow. By allowing the scalar core to issue up vector load requests in advance, the X390 processor in a 4 core cluster can maintain 1,024 outstanding memory requests within a single-threaded model.

- **Efficiency Gain:** The single-threaded data queue achieves the latency-masking power of GPUs and up to 8x the capacity of traditional server CPUs without the silicon tax of GPU-style context switching.
- **Predictable Throughput:** By operating in a single-threaded execution environment, it removes the jitter and latency penalties common in heavily multithreaded systems.

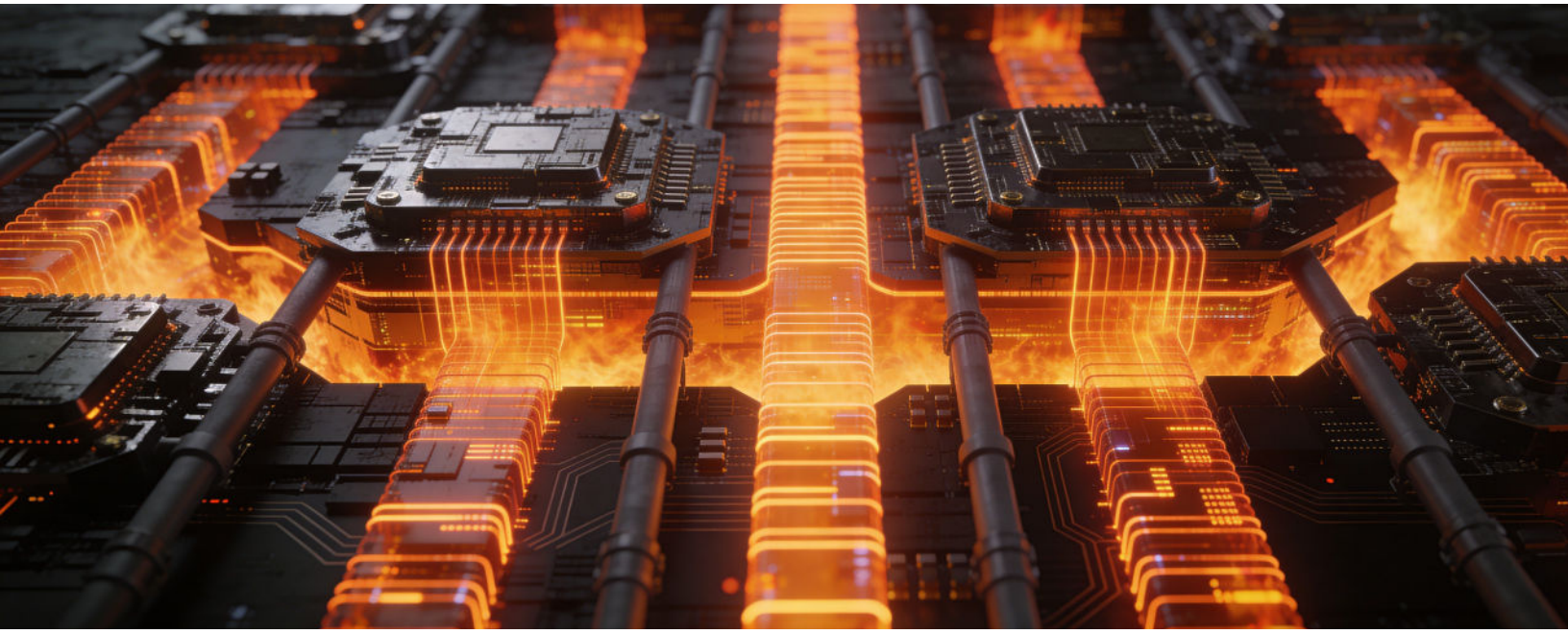
Why Queue Depth is the Silicon Architect's Lever

Queues decouple the dispatch of instructions from their execution, allowing the system to maintain a massive number of outstanding memory requests. This capability is essential for ensuring the execution units never starve.

The configurable depth of SiFive's Vector Load Data Queue (VLDQ) allows architects to size the queue specifically for specific silicon topologies, including:

- **Multi-chiplet designs:** a deep VLDQ masks the prolonged latencies of die-to-die interconnects, maintaining a consistent one-cycle latency for the vector unit.
- **Low-latency memory:** the VLDQ can be minimized to reduce silicon footprint and power consumption without altering the software-visible architectural logic.

This flexibility allows a single RISC-V IP core to scale from a power-sipping edge sensor to a high-bandwidth data center chiplet.



Zero-Cycle Memory Access & Cache Integrity

A persistent bottleneck in AI processing is "cache thrashing," a phenomenon where the continuous streaming of massive weight datasets evicts critical instructions and scalar data from the high-speed L1 cache, severely degrading performance. RISC-V's decoupled architecture solves this through Selective Cache Bypassing, which directs vector loads to the L2 or a dedicated Core Local Port (CLP). This ensures that the L1 remains dedicated to control flow data while the vector unit streams in parallel.

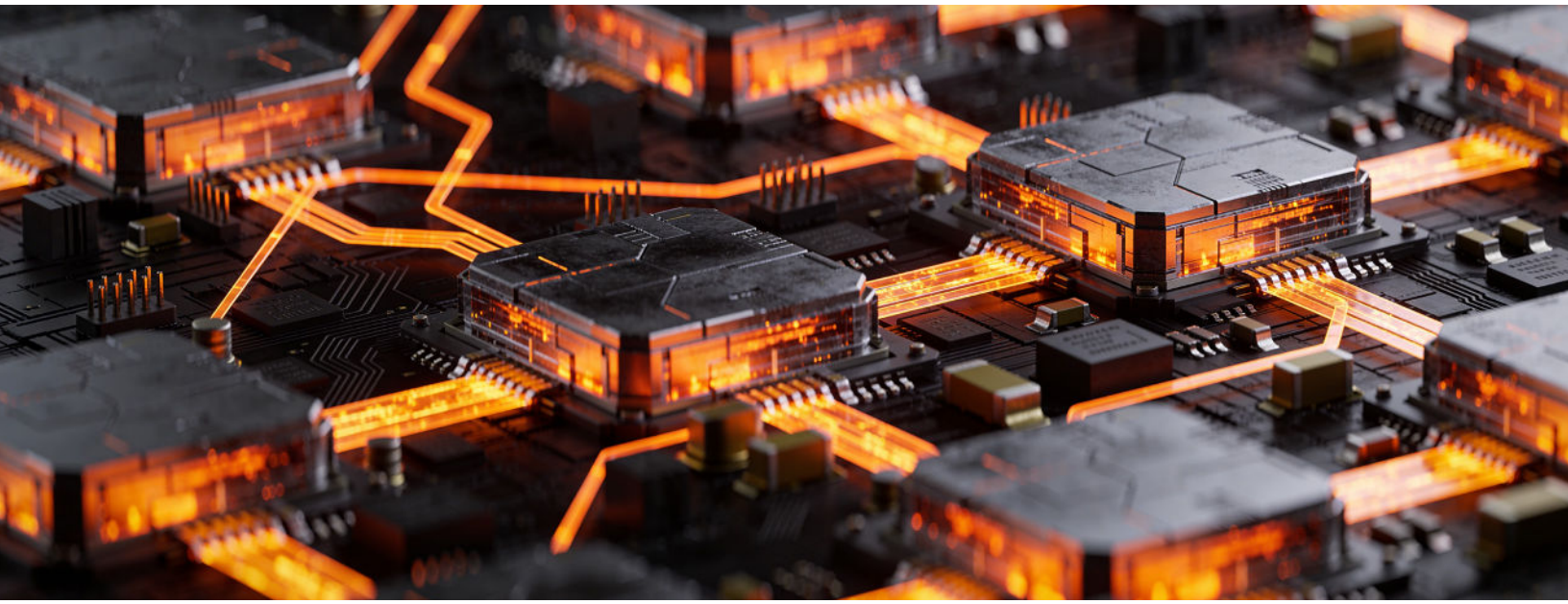
For multi-chiplet architectures, the VLDQ acts as a high-speed buffer, absorbing the variable latencies of die-to-die interconnects. By leveraging specialized ports like the CLP, architects can achieve zero-cycle perceived latency for compute-critical elements. The result is a non-blocking weight-delivery system that maintains absolute deterministic performance even as AI model sizes continue to scale.

A Traditional Programmer Experience with GPU-Class Performance

The industry is currently suffering from kernel fatigue. Developers must manage proprietary GPU kernels and/or rigid, fixed-function NPU state machines. These black boxes are notoriously difficult to debug, and custom software often becomes obsolete when AI models change, which currently occurs on a near-weekly basis.

To streamline the AI engineering lifecycle, RISC-V offers a Turing-complete vector engine that functions within a standard C/C++ programming environment. New high-bandwidth, low-latency interfaces like VCIX (Vector Coprocessor Interface eXtension) allow the CPU to push instructions directly into the accelerator pipeline, transforming hardware logic analysis into a standard software debugging session.

Unlike fixed-function accelerators, this architecture can handle complex LLM functions via dedicated hardware exponential units and adapt to new operators via software updates. As a result, the vector engine enables the processor to be used as an Accelerator Control Unit (ACU) by providing the necessary programmable compute flexibility that fixed-function accelerators lack. An ACU can manage data movement and scheduling for a customer's dedicated accelerator, as detailed on SiFive's blog.



5. Software Strategy and Ecosystem Alignment

Practical Realities of Model Porting

Standard PyTorch and TensorFlow models now run out-of-the-box on the RISC-V architecture. This compatibility is achieved through an open-source toolchain comprised of the IREE (Intermediate Representation Execution Environment) and LLVM that functions much like a universal web browser for silicon. By abstracting the hardware layer, these tools allow developers to map massive foundation models directly onto RISC-V cores without manual code rewrites. This ecosystem alignment ensures that as the AI expands, RISC-V hardware remains a first-class citizen in the global software repository, significantly lowering the barrier to entry for hyperscalers and edge developers alike.

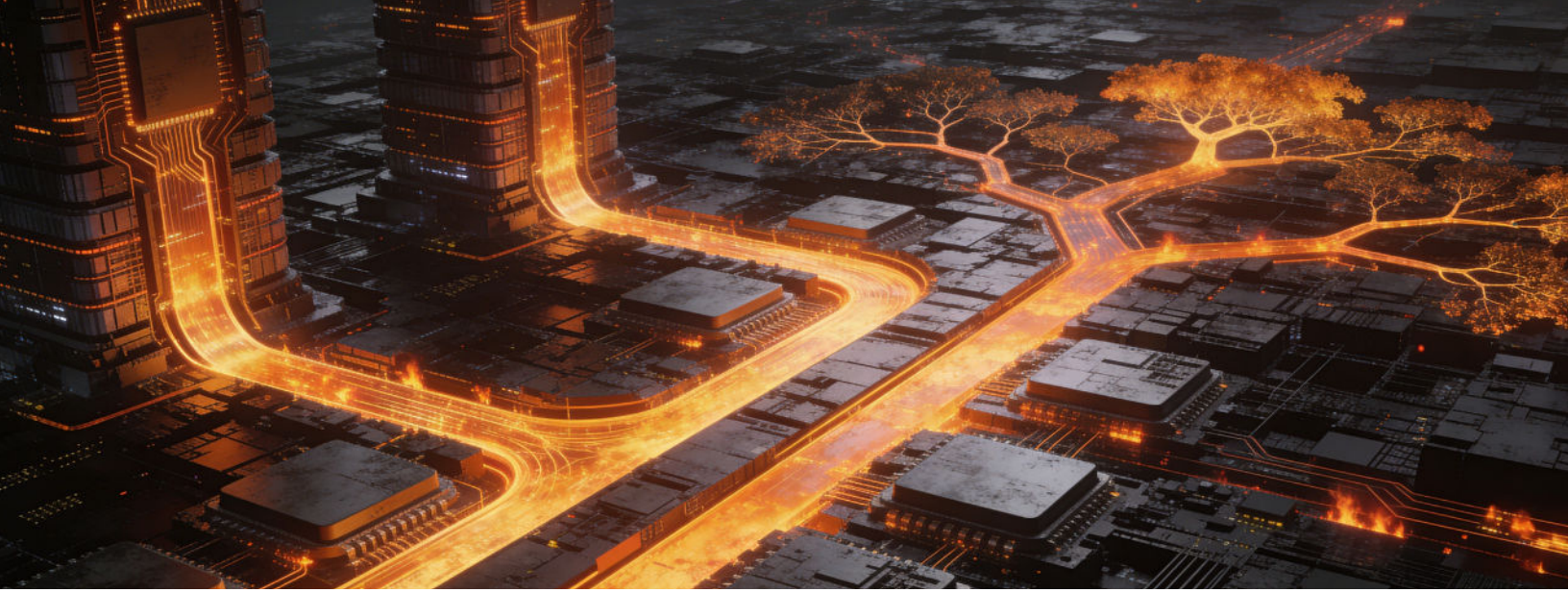
High-performance deployments still benefit from a human-in-the-loop approach. For the most compute-intensive kernels, architects can leverage hand-tuned libraries like the SiFive Kernel Library (SKL) to maximize vector unit throughput. This hybrid strategy combining automated compiler efficiency with targeted manual tuning mirrors the early days of web optimization, where standard protocols handled the bulk of the traffic while critical paths were fine-tuned for maximum speed.

Long-Term Ecosystem Path

As foundation models shift toward extreme efficiency, exemplified by innovations like DeepSeek's multi-head latent attention (MHLA) and Mixture of Experts (MoE) architectures, the hardware ecosystem must adapt. The Multi-Head Latent Attention (MHLA) innovation enables better performance on long-context reasoning, shifting the core performance bottleneck from memory bandwidth to compute efficiency by radically reducing Key-Value (KV) cache pressure. RISC-V can adapt to this innovation with features like FP8-native vector extensions to handle the new, compute-intensive requirements of these efficient models.

However, not all models will universally adopt MHLA. The AI landscape continues to diversify, with other approaches like router-based MoE models that prioritize deterministic execution and throughput efficiency over long context understanding. These designs rely on sparse matrix computation rather than dense attention, creating a different optimization target that aligns well with Integrated Matrix Extensions (IME). RISC-V's ability to expose a composable set of domain-specific extensions positions the architecture to adapt across this expanding range of specialized model types, from attention-heavy designs to sparsity-driven MoE systems.

The roadmap for RISC-V matrix extensions significantly strengthens its position as the central foundation for the next decade of AI. While proprietary architectures remain locked into rigid roadmaps, new Vector Matrix Extensions (VME) introduce dedicated architectural states to dominate the compute-heavy prefill phase of LLMs. By integrating open extensions with IP, including hardware-pipelined exponential units for non-linear math, SiFive's designs reduce critical activation functions like GELU and Softmax from 20+ cycles to a single cycle. The RISC-V community provides the open, transparent foundation needed to run these efficient models on sovereign silicon, ensuring that the next decade of AI innovation remains decentralized, programmable, and future-proof.



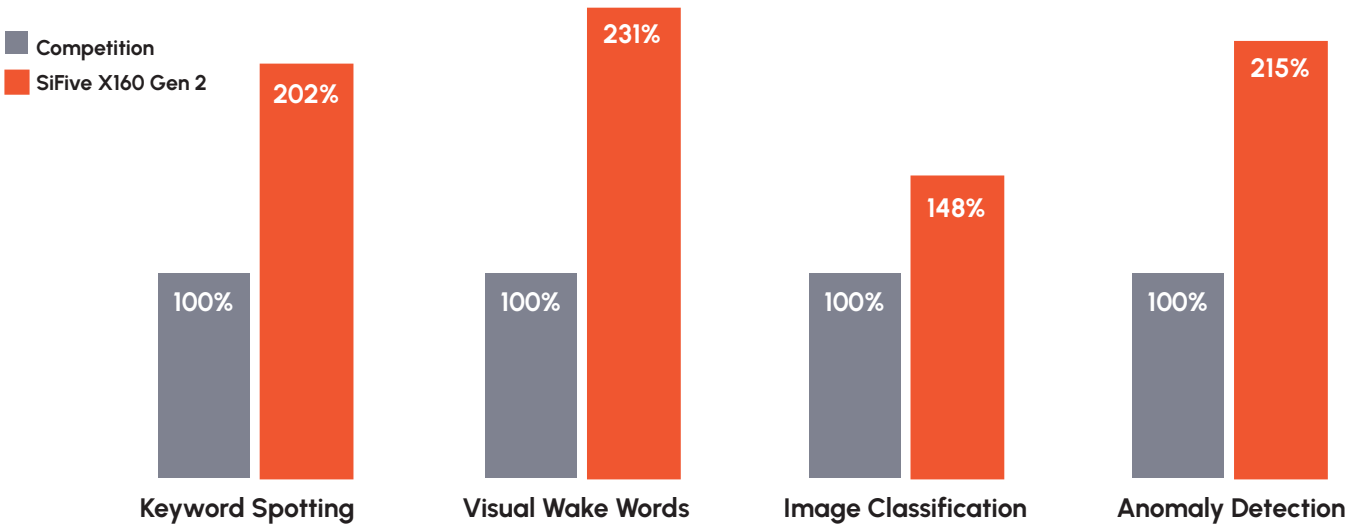
6. Workload Fit: Large Models, Small Models, and Everything Between

As workloads diversify from sub-milliwatt sensor fusion to multi-megawatt foundation model training, the industry is witnessing the collapse of the one-size-fits-all general-purpose processor paradigm. In its place, chipmakers are launching higher volumes of new designs with memory hierarchies and execution pipelines based on model size and AI lifecycle stage.

Small Models

Zero-cycle memory access is particularly potent for models under 10MB, which can reside entirely within the 1MB shared L2 cache or Tightly Integrated Memory (TIM), eliminating the energy and time penalties of external RAM access. When compared to the ARM Cortex-M85, the SiFive X160 Gen 2 delivers roughly 2x the inference performance within the same silicon footprint for key edge tasks, including keyword spotting, visual wake words, and visual anomaly detection (see Figure 3). This leap is driven by the transition from ARM's scalar-heavy approach to SiFive's vector engines and dedicated hardware for AI activation functions.

Figure 3. Normalized Inference Performance: SiFive vs. Competition



Source: Futurum Research

Medium to Large Models

For Medium to Large Models, such as Large Language Models (LLMs), model weights reach into the billions of parameters, making it impossible to fit them entirely within on-chip caches. In these scenarios, the bottleneck shifts from compute speed to memory bandwidth, requiring a system that can tolerate memory latency rather than just trying to avoid it. This is where SiFive's decoupled vector architecture shines: the scalar unit acts as a scout, processing the instruction stream and dispatching vector and matrix commands to a command queue well in advance of their execution. This decoupling allows the processor to initiate hundreds of concurrent memory requests to external DRAM, effectively hiding the massive latency of off-chip data fetches.

Reinforcement Learning for Foundation Model Fine-Tuning

LLM scaling increasingly owes to reinforcement learning (RL), where models receive rewards for strong performance on complex tasks. Verifiers are based on scalar logic, often carried out by CPUs. These workloads are notoriously difficult for standard servers because they mix heavy branching logic with dense tensor math. SiFive's decoupled architecture allows the scalar unit to handle complex branching logic while the vector engine simultaneously computes the neural network's next move. On top of this architectural advantage, developers can use the flexible RISC-V ISA to add custom instructions for specialized RL activation functions or environment-specific math, giving SiFive a performance advantage over generic CPUs.

7. Scalability: SiFive's Architecture from Edge to Data Center

From Tiny Edge Devices to Industry-Leading Data Center Designs

SiFive's Intelligence Family can power everything from basic sensors and robotics at the edge to Blackwell-class high-end designs for the data center. This range is made possible by a unified ISA that ensures code reuse and software interoperability across all tiers. Because this microarchitecture can be configured to compete directly with today's most advanced data center accelerators, it provides a useful alternative for customers looking to build next-generation training and inference chips configured to their specific use cases and operating environments.

Customization and Chiplet Configuration

A primary advantage of SiFive's RISC-V IP is the deep level of customization available to customers. SiFive allows architects to tailor the IP specifically for their target application:

- **Area and Power:** Optimize the physical footprint and energy consumption for battery-operated edge devices.
- **Performance:** Tune microarchitecture parameters to maximize throughput for specific AI workloads.
- **Memory Locality and Chiplets:** Configure memory-latency-hiding queues and data paths to support multi-chiplet arrangements with varied latencies.
- **Microarchitecture Parameters:** Adjust internal configurations, such as the use of "zero-cycle" memory paths, to prevent cache thrashing during large model execution.

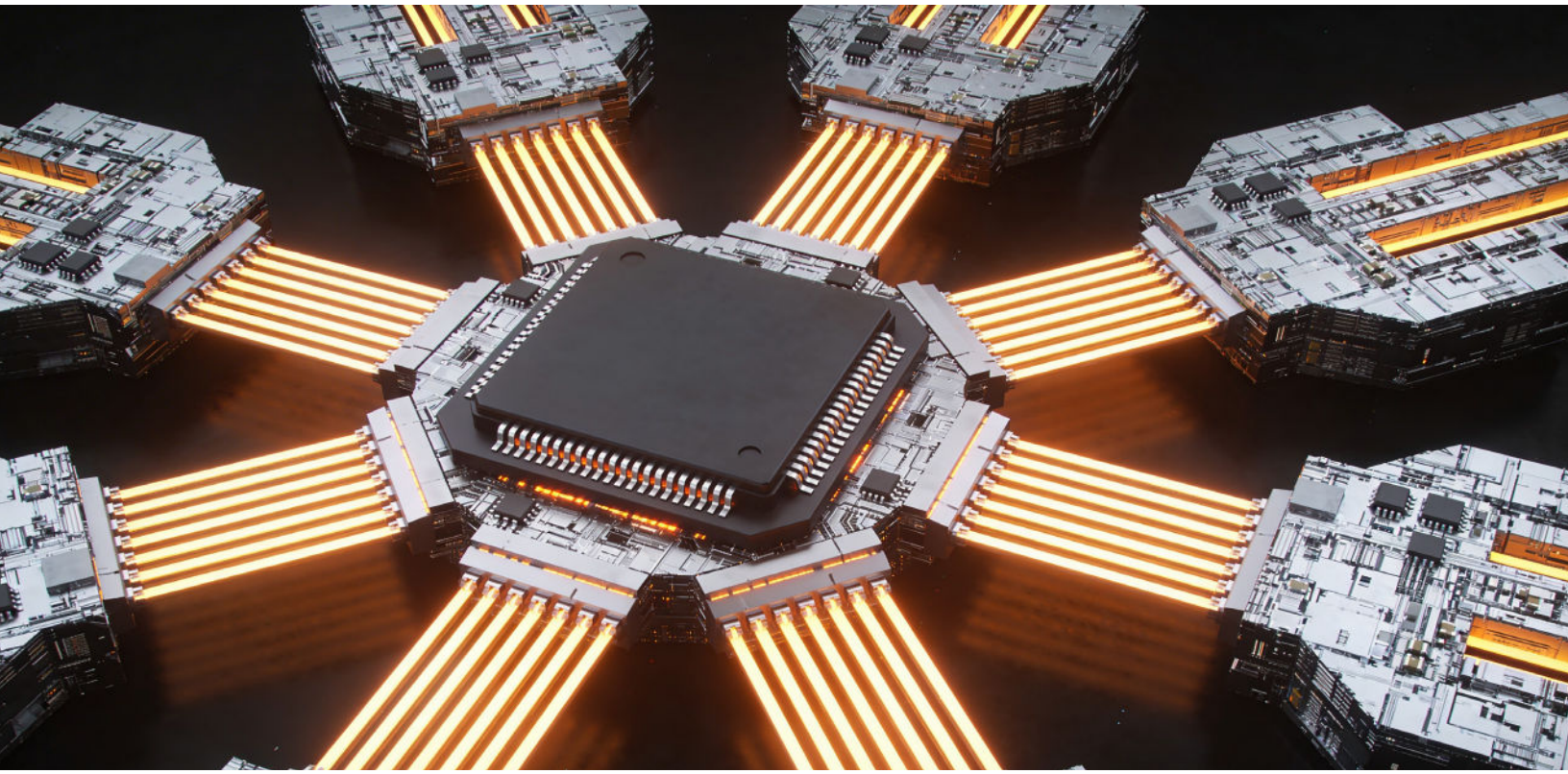
The Emerging Edge AI Opportunity

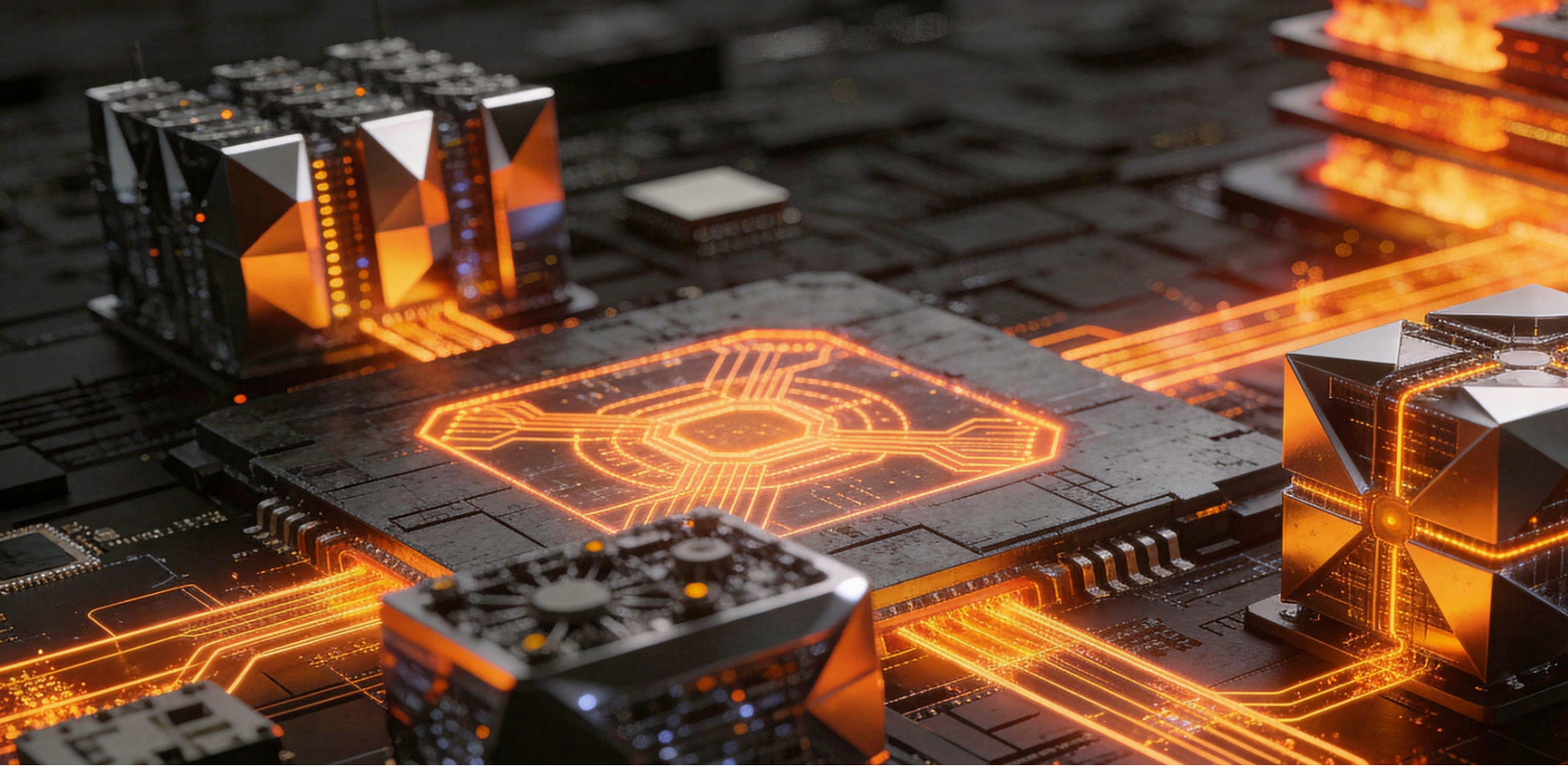
The next frontier of artificial intelligence moves beyond centralized clouds to the periphery of the network – the Edge. This shift is characterized by a transition toward local inference, where data is processed at or near the point of origin in order to meet strict requirements for latency, privacy, security, and bandwidth efficiency. This Edge AI landscape is inherently diverse, and encompasses a vast array of applications that include advanced robotics, industrial automation, high-fidelity computer vision, acoustic monitoring, and sophisticated sensor fusion.

Unlike the relatively uniform environment of the data center, edge deployments demand an architecture that can adapt to specialized workloads without sacrificing the power efficiency required for constrained environments.

SiFive is already taking a leadership role in this space by providing the silicon foundation for multiple high-profile edge AI designs. A notable example is Kinara, whose SiFive-powered technology was later acquired by NXP to bolster their edge processing portfolio. This real-world adoption highlights that SiFive's RISC-V Intelligence Family is already proving to be a practical, high-performance solution currently powering the "physical AI" devices of today—from automotive mobility systems to local decision-making engines in distributed industrial environments.

As the Edge AI market continues to mature, chip designers are expected to increasingly place a premium on architectural diversity and configurability. Compared to fixed-function and more rigid architectures, which can struggle to meet the unique "power-performance-area" (PPA) targets required by, say, a smart camera versus an industrial sensor, SiFive's modular RISC-V approach allows customers to tailor their IP to create workload-tuned silicon perfectly optimized for its specific edge application (adjusting parameters like memory locality, cache-efficiency designs, and vector unit configurations, for example). This level of flexibility positions SiFive as a particularly critical technology partner for companies building the next generation of intelligent edge nodes.





8. Competitive Positioning

Extensible Software for Real-time AI

SiFive has bridged a historical gap in RISC-V software standards gap by introducing its own equivalent to NVIDIA's cuDNN or Arm's Compute Library. The SiFive Kernel Library (SKL) provides pre-optimized mathematical primitives in a tuned C/C++ kernel library for neural networks and signal processing. SKL stands out relative to more mainstream AI-focused CPU kernel stacks for gaining high CPU utilization with tight hardware coupling. Beyond CPUs, GPU libraries can't maximize CPU utilization, leaving a data bottleneck that can starve GPUs. SiFive's move to open-source SKL allows developers to optimize for RISC-V much as they have for Arm.

SiFive is well-positioned for a future of real-time LLM inference. Traditional GPUs often struggle at the edge because their architecture carries an offload tax of millisecond-scale latency due to PCIe bus transfers. SiFive cores enable zero-copy data sharing with custom hardware, achieving nanosecond-scale latency and up to 1 TB/s of aggregate bandwidth. This combination of elastic scaling and memory latency tolerance makes SiFive the ideal foundation for on-device applications where deterministic performance and power efficiency are non-negotiable.

Edge-to-Cloud AI Enablement

As open-source models reach parity with closed-source alternatives, AI is rapidly extending from the cloud to the edge and developers now need to run memory-intensive workloads at any scale. ISA-level innovation offers silicon developers a common instruction set that abstracts memory-centric operations across low-cost hardware and the increasingly critical CPUs in the data center. By standardizing new vector, matrix, sparsity, and memory primitives into open profiles, the ISA becomes a shared language any chipmaker can implement, creating a broad, interoperable pool of edge and data center platforms. This approach disrupts XPU design vendors that monetize proprietary accelerator stacks. The incentive for customization makes "write once, run anywhere" across low-cost hardware and data center CPUs commercially unattractive for them. The SiFive approach can materially cut non-recurring software engineering, enabling faster time-to-silicon and lower costs.

9. Customer and Market Use Cases

Data Center Inference

Currently, the largest segment of SiFive's customer base is focused on inference in the data center. As Large Language Models (LLMs) transition from training to large-scale deployments, the industry requires high-efficiency hardware capable of tolerating memory latency without the power overhead of traditional GPUs. SiFive's architecture is exceptionally well-suited for this type of hyperscale inference, with at least one customer already developing data center-class chips designed to compete directly with NVIDIA for training and inference workloads.

Physical AI and Edge AI Devices

The fastest-growing base for SiFive is Physical AI, with intelligence embedded directly into edge devices. In these environments, local inference is preferable real-time response and enhanced user experience, security, safety, and privacy.

- **Robotics and Mobility:** SiFive powers autonomous systems in robotics and the automotive sector, where low-latency decision-making is critical for navigation and safety. This may be a critical high-growth direction for SiFive as software-defined vehicles and robots are expected to experience significant growth in the coming years.
- **Industrial Automation:** From smart sensors to complex factory automation, SiFive's configurable IP allows for efficient localized processing in distributed environments – another potentially high-growth segment in the coming years.
- **Vision and Audio Analytics:** The architecture's vector capabilities make it ideal for high-throughput tasks like smart camera vision and real-time audio analytics.

A key success in this area is SiFive's technology powering edge AI chips such as those from Canara (acquired by NXP), demonstrating the platform's readiness for industrial-grade deployment.

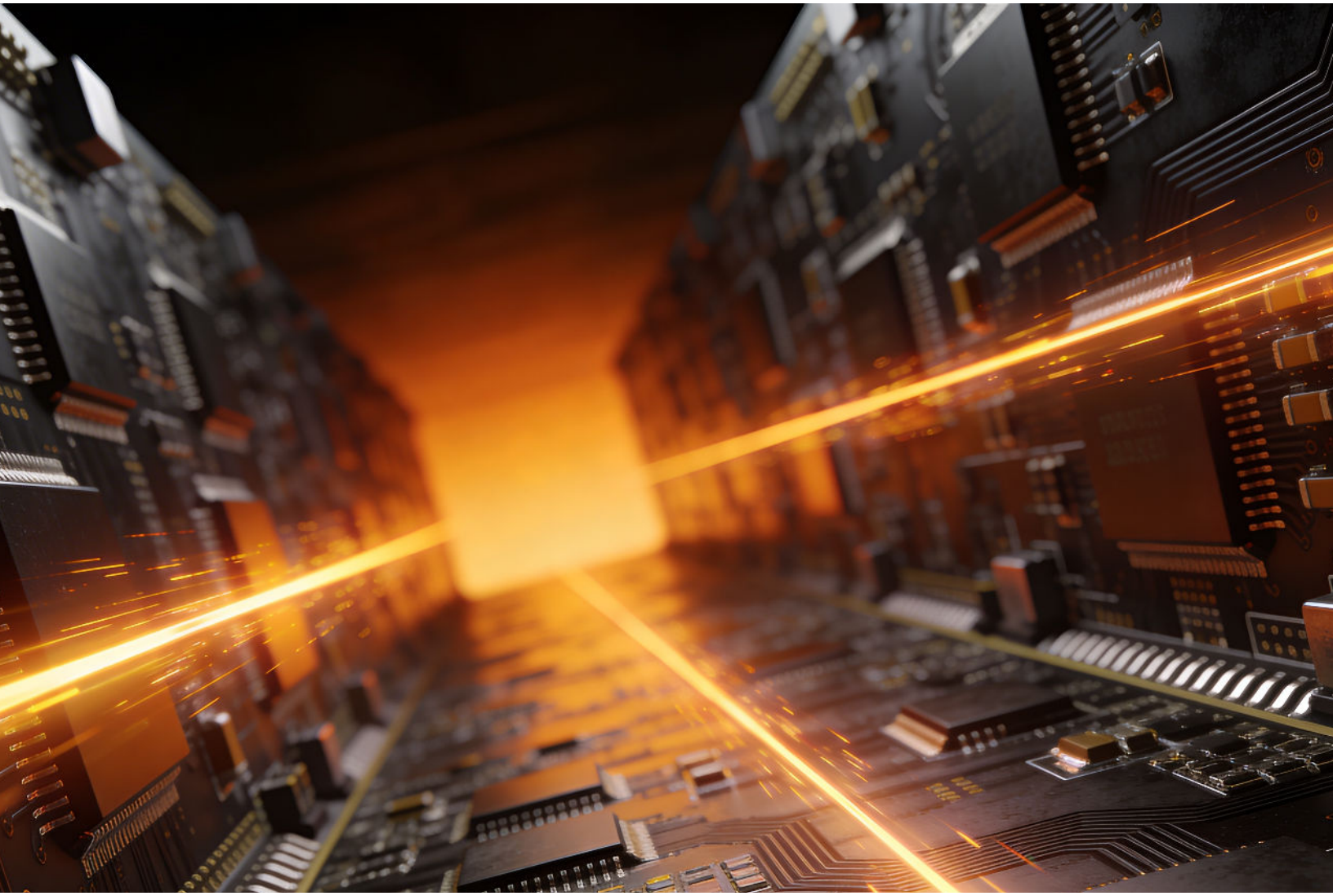
Incremental Model Tuning

Beyond raw inference, SiFive enables incremental model tuning, allowing organizations to adapt foundation models to specific datasets or domains without the prohibitive cost of full-scale training. This capability is particularly valuable for enterprises seeking to maintain specialized, up-to-date models on-premises or at the edge.

Custom Chip Design

SiFive's configurability serves hyperscalers and major semiconductor vendors who are increasingly designing custom XPUs to escape vendor concentration risks and high costs. By providing an open-standard, workload-tuned foundation, SiFive acts as a strategic technology partner for the next generation of custom AI accelerators.

Silicon design teams can leverage the SiFive Core Designer and SiFive Custom Instruction Extensions (SCIE) to rapidly integrate proprietary AI math or domain-specific accelerators directly into the processor's pipeline, effectively bypassing the rigid limitations of standard off-the-shelf silicon. This modularity enables a level of architectural tailoring where customers can optimize for specific parameters like area, power, and performance.



10. Analyst Perspective: Why SiFive's Approach Matters Now

Market Timing

In AI, the engineering community is running headlong into the memory wall. The industry has spent the last few years pursuing scaling laws by throwing more HBM (High Bandwidth Memory) and more power at the problem. But as model parameter counts explode, legacy architectures are buckling. Software teams are now picking models based on how they fit into a cache, not just how well they perform.

This is exactly why we're seeing a trend toward silicon diversification. NVIDIA has pivoted toward open architectures with NVLink Fusion and CUDA support for RISC-V, Big companies like Meta are deeply invested in RISC-V with the acquisition of Rivos, and Qualcomm appears to be aggressively pivoting toward the architecture with its acquisition of Ventana. SiFive's 2nd Generation Intelligence family is hitting the market just as this hunger for silicon freedom peaks. We're moving from a world of one-size-fits-all AI chips to a portfolio approach where you can mix and match IP to fit anything from a tiny IoT sensor to a massive hyperscale cluster.

Architectural Elegance: Hiding Latency Through Design

The decoupled vector and queueing model employed by SiFive is a foundational architectural shift. CPUs are the secret weapon of custom AI servers in feeding data to greedy AI accelerators. Mismatches between commoditized dataflow processors and advanced GPU clusters are a primary driver of low efficiency metrics. To more tightly sync these resources, we see deterministic computing for specific workloads as a critical engineering challenge. High-reliability and low-latency agentic experiences will need optimized solutions for each stage of the model pipeline to avoid the chaos of all-to-all GPU processing. SiFive can play a critical role in the deployment phase of agents.

Future-Proofing AI Compute

The biggest historical knock against RISC-V was the "software gap." That argument is officially dead. With the ratification of the RVA23 profile, we finally have a standard that guarantees binary compatibility for Linux distributions like Red Hat and Ubuntu. Against that backdrop, industry moves such as NVIDIA working to enable CUDA toolchains on SiFive RISC-V development platforms and participating in the RISE Project signal that the performance-critical AI software stack is landing on open ISAs, not just proprietary accelerators. This means a developer can write code once and have it run across a whole spectrum of hardware. With AI demand far outstripping supply, the pressure for open and customizable solutions should prove overwhelming.

11. Conclusion

The evolution of RISC-V from a niche alternative to a primary architectural standard marks a turning point in AI computing. SiFive leads this charge by bridging the gap between familiar CPU programmability and GPU-class performance efficiency. Through its unique decoupled vector architecture and latency-hiding queues, SiFive's Intelligence Family poses an open solution to the industry's most pressing challenge: the memory wall.

SiFive's customizable, unified ISA aligns with technology history's pattern of trending toward open standards. As the industry moves away from one-size-fits-all servers toward diversified, workload-tuned silicon, SiFive can serve as the essential partner for edge-to-cloud system architecture.

These microarchitectural innovations ensure that software remains portable and future-proof across the compute spectrum. Whether deploying sub-milliwatt sensors or frontier data center accelerators, an open platform can become the engine for an era of decentralized and ubiquitous intelligence.

Important Information About This Report

AUTHORS

Brendan Burke
Research Director, Semiconductors,
Supply Chain & Emerging Tech | The Futurum Group

PUBLISHER

Futurum Research

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

This document, including any supporting materials, is owned by The Futurum Group. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of The Futurum Group.

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

SiFive

ABOUT SIFIVE

SiFive is a fabless semiconductor company focused on developing high-performance processor IP based on the open-standard RISC-V instruction set architecture. The company licenses configurable CPU and accelerator cores to semiconductor manufacturers and system companies across markets including data center, automotive, aerospace, and embedded systems. SiFive's portfolio spans general-purpose, real-time, and application-class processors, along with domain-specific offerings such as the SiFive Intelligence family, which targets AI and machine learning workloads with scalable vector and matrix compute capabilities. By leveraging the flexibility of RISC-V, SiFive positions itself as an alternative to proprietary architectures, enabling customers to customize silicon designs while maintaining ecosystem interoperability.

Futurum®

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

Futurum

CONTACT INFORMATION: The Futurum Group LLC | futurumgroup.com | (833) 722-5337

© 2026 The Futurum Group. All rights reserved.