Navigating the AI Revolution: Addressing the Challenges Facing Cloud Service Providers

White Paper

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Table of Contents

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Executive Summary

As AI and next-generation workloads grow in complexity and scale, data center operators and cloud service providers need compute solutions that deliver high performance, energy efficiency, and seamless integration. Rivos meets this demand with purpose-built, data center-class System-on-Chip (SoC) solutions that unify out-of-order and parallel compute capabilities within a flexible architecture. Designed for AI services where cost per token and dollar per GPU hour matters, Rivos enables providers to deliver maximum user experience while minimizing the cost of delivering those tokens.

At the core of Rivos' solution is the balance between performance and energy efficiency; high-performance, fully featured Rivos-designed 64-bit RVA23 RISC-V CPU cores tightly integrated with a Rivos-designed SIMT GPGPU. These heterogeneous compute units share a unified memory subsystem with both high-bandwidth on-chip HBM3e and high-capacity DDR5 RDIMMs, ensuring efficient data access and minimal latency. Integrated Ultra Ethernet NICs deliver robust, high-throughput connectivity over standard interfaces, enabling seamless vertical and horizontal scaling.

Designed with the full software stack in mind, the Rivos SoC supports current widely-adopted AI frameworks and allows for the seamless reuse of existing data-parallel algorithms and code-bases, significantly accelerating deployment. This combination of energy-efficient design, open-source software alignment, and seamless memory and compute integration position Rivos SoCs as a compelling choice for modern data centers looking to power AI innovation.

This white paper explores the key obstacles CSPs face in enabling AI workloads, including hardware limitations, software stack complexity, data governance, pricing models, and regulatory hurdles. It outlines actionable strategies these providers can adopt to stay competitive in the AI-powered digital era whilst looking at how Rivos solutions can help to address these key challenges.

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Introduction

Artificial intelligence is fundamentally transforming the computing landscape. The deployment and training of AI models, particularly large language models (LLMs), demand specialized hardware, optimized software stacks, and vast amounts of data. While Hyperscalers have responded with purpose-built AI chips, vertically integrated platforms, and multi-billion-dollar investments, the broader ecosystem of cloud providers must adopt a more strategic and differentiated approach to stay competitive in this rapidly evolving environment.

This white paper focuses on Cloud Service Providers (CSPs), including localized service vendors, infrastructure specialists, and vertically integrated service platforms. These providers often excel in specific markets or regions, leveraging flexibility and strong local relationships as competitive advantages. Yet, they face increasing risk as AI workloads consolidate around a small number of providers with the scale and resources to build and maintain AI-first infrastructure.

For AI to reach its full potential and deliver value across the broadest spectrum of users, cloud service providers must offer a diverse range of services that can cost-effectively adapt to evolving customer needs. This flexibility is essential not only to support varied AI workloads but also to enable providers to generate sustainable, profitable revenue from AI-driven offerings.

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Core Challenges Facing CSPs

Delivering AI services at scale requires navigating a complex web of technical, operational, and regulatory hurdles. While the promise of AI is driving unprecedented demand, the realities of infrastructure, software, talent, compliance, and market pressures can significantly slow or limit adoption.

This section examines the most pressing constraints facing cloud service providers and enterprises as they scale AI capabilities, from hardware shortages and data center limits to software integration challenges, talent gaps, governance requirements, cost pressures, and geopolitical uncertainties.

1. **Infrastructure Constraints**
2. **Software Stack Complexity**
3. **Ecosystem & Talent Gaps**
4. **Data Sovereignty & Governance**
5. **Cost & Pricing Pressures**
6. **Regulatory & Geopolitical Uncertainties**

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Core Challenges Facing CSPs

Infrastructure Constraints

Providing AI services at scale requires high-performance and specialized accelerators. Securing sufficient GPUs, custom AI chips, or ASICs is often the first major barrier to growth.

Key barriers include:

- **Global scarcity and high capital costs** associated with advanced GPU architectures
- **Power and cooling limitations** in existing data centers, along with restricted infrastructure expansion and rising operational costs
- **Networking bottlenecks**, especially for training large models, which require high-throughput interconnects that often exceed current capabilities or depend on proprietary solutions

Software Stack Compatability

Many cloud providers are looking for new solutions to address the users needs, but want to avoid duplication of effort and costs incurred by supporting parallel codebases in-house, which lead to higher integration complexity and maintenance costs.

AI enablement depends on a supported, optimized, layered software stack that includes:

- **Machine learning frameworks** such as PyTorch and TensorFlow
- **Container orchestration tools** like Kubernetes and Kubeflow
- **Scalable environments** for model training and inference
- **Support for open-source** and proprietary MLOps tools

Ecosystem & Talent Gaps

One of the most critical barriers to successful AI adoption is the shortage of highly skilled and experienced AI engineering talent.

This talent gap hampers organizations' ability to:

- **Efficiently deploy, integrate, and manage AI services,** which often rely on complex and evolving software stacks.
- **Businesses face greater integration challenges,** increased operational and maintenance costs, and delays in bringing AI solutions to market.

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Core Challenges Facing CSPs

Data Sovereignty & Governance

AI models often rely on large volumes of sensitive data, raising significant regulatory and operational concerns, especially CSPs serving regulated industries or national markets.

These providers must ensure:

- **Strict compliance** with data residency and privacy regulations (e.g. GDPR)
- **Secure data handling** throughout the AI model training lifecycle
- **Transparency and accountability** in AI model usage, including bias detection and mitigation

CSPs need large legal teams to meet complex compliance demands, increasing operational overhead and legal risk when delivering AI services.

Cost & Pricing Pressures

AI workloads are resource-intensive and costly to run, while customers increasingly expect flexible, usage-based pricing.

For cloud providers, this creates significant challenges:

- **Limited economies of scale** make it difficult to offer competitive GPU pricing
- **Infrastructure inefficiencies,** such as over- or under-provisioning, drive up costs
- **Tight operating margins** restrict the ability to adopt aggressive pricing strategies

As the AI market moves decisively toward consumption-based models, cloud providers must adapt their pricing structures or risk becoming uncompetitive.

Regulatory & Geopolitical Uncertainties

The evolving geopolitical landscape is exerting growing influence over AI and cloud infrastructure.

Export controls, national AI strategies, and region-specific regulations are reshaping the competitive environment in significant ways:

- **Export restrictions** on advanced semiconductors limit access to critical AI hardware in certain regions
- **Localized AI regulations** (e.g., data sovereignty, algorithmic transparency) may require region-specific compliance infrastructure
- **Conflicting interests** between competing national interests, navigating complex cross-border pressures

As geopolitics increasingly intersects with AI, CSPs must strategically position themselves to remain compliant, resilient, and relevant.

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Critical Success Areas for CSPs

Addressing these multifaceted challenges requires a fundamentally different approach to how cloud service providers source, operate, and manage both their infrastructure and software stacks. Flexibility, adaptability, and strategic alignment are essential to effectively navigate the shifting demands of AI workloads, regulatory environments, and evolving user expectations.

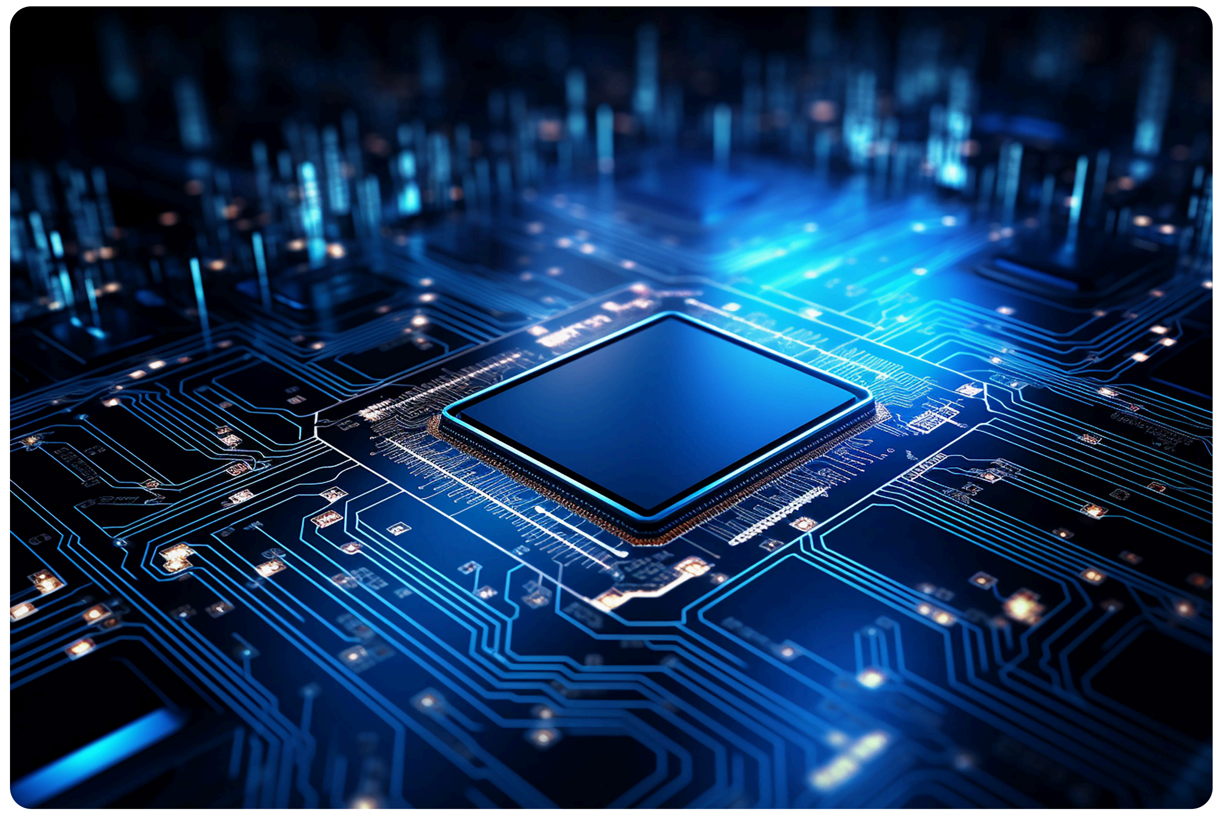
1. **Diversify Hardware Partnerships**
2. **AI Specialization and Verticalization**
3. **Utilizing Open-Source and Community Models**
4. **Emphasize Sovereignty and Trust**

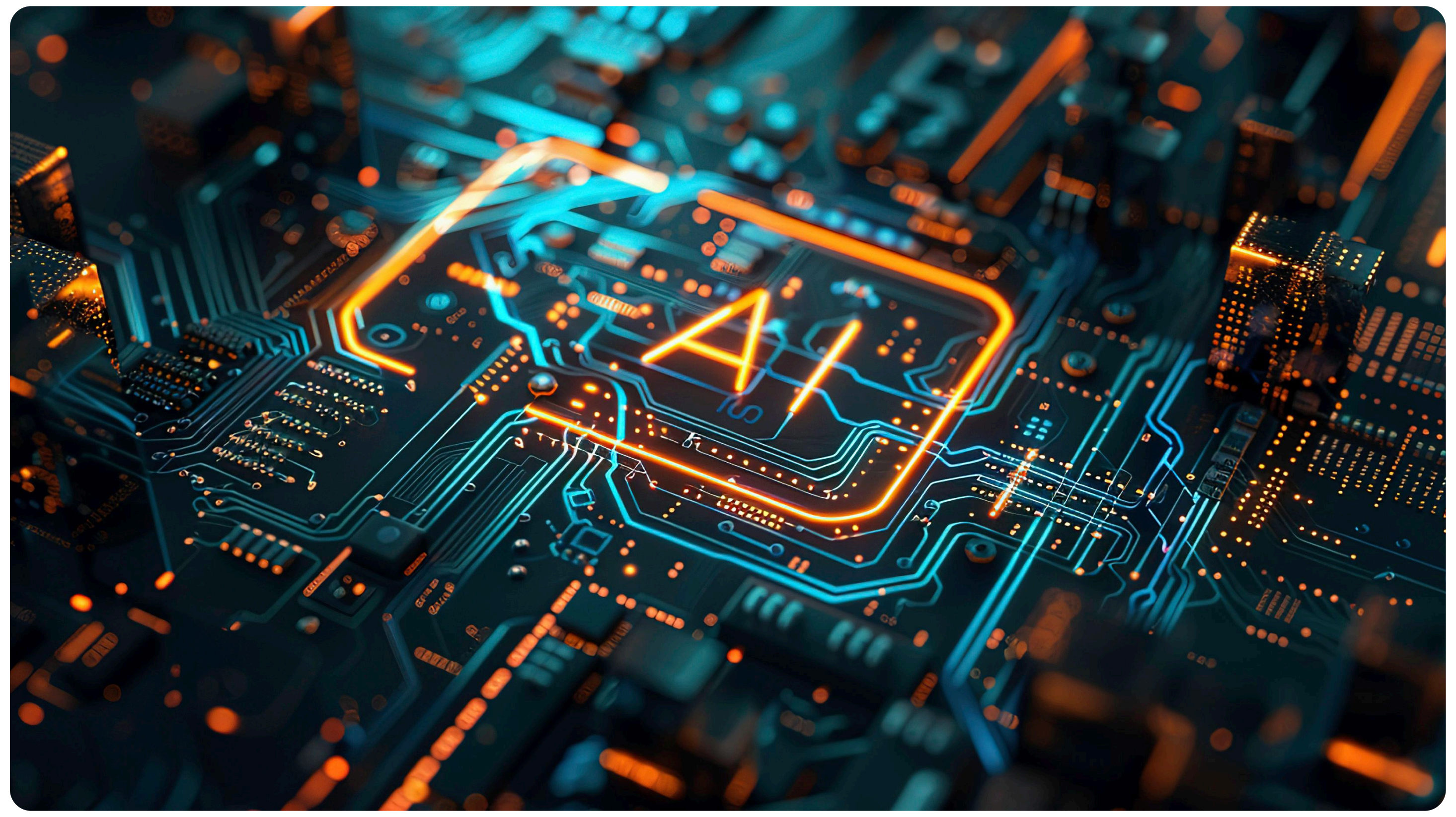
Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Critical Success Areas for CSPs

Diversify Hardware Partnerships

The market now offers a wider array of solutions, from programmable architectures like Rivos, NVIDIA, and AMD to specialized accelerator designs optimized for specific model types, such as Groq. Programmability enables flexibility to adapt to models in the future. This diversification allows for the use of more customizable hardware, fostering differentiation and expanding access to various hardware solutions. Consequently, this enables the provision of more diverse services, including GPU as a service, which helps reduce capital expenditure and operational costs. Running workloads on the most optimal architecture ensures peak performance with maximum energy efficiency.





AI Specialization and Verticalization

As AI models evolve and new model architectures emerge, we will see more domain-specific models tailored to sectors - like healthcare, finance, or manufacturing continue to advance and gain adoption. Solutions that offer the flexibility to adapt to evolving model requirements are crucial. This adaptability allows for changes in model needs without altering the underlying hardware, maximizing ROI and accelerating time to service these market opportunities. This specialization also has the added benefit of attracting customers who require localized or compliant AI solutions.

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Critical Success Areas for CSPs

Utilizing Open-Source and Community Models

Open-source software has been foundational to the pace of innovation in AI, and plays a vital role in the growing capabilities and breadth of model support for the future. Open-weight large language models (LLMs) such as Meta's LLaMA and Mistral enable CSPs to offer a wide range of AI capabilities from inference and fine-tuning services to advanced reasoning, without the need for massive upfront investments. By collaborating closely with these open-source communities and leveraging open source tools and frameworks, CSPs can accelerate time-to-market, foster innovation, and reduce development costs.

Emphasize Sovereignty and Trust

CSPs can differentiate themselves by emphasizing trust and compliance. By promoting sovereign cloud solutions, ensuring strict data locality, and providing explainable AI tools, they can effectively capture and grow their presence in highly regulated markets. Navigating between domain compliance and national requirements.

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Rivos Solutions Addressing Critical Cloud & AI Demands

To fully realize the potential of AI, businesses require flexible yet purpose-built infrastructure; seamlessly integrating high-performance programmable parallel computing (GPGPUs), scalable and resilient storage, and advanced networking. This foundation must be secure, reliable, and energy-efficient to support the growing demands of AI.

Rivos delivers a differentiated approach to these challenges. Offering secure, scalable, high-performance energy-efficient data center solutions, based Rivos' portfolio of SoCs, designed to meet both CSPs and their customers needs.

All our solutions are fully featured with upstream and validated support for most standard software stacks, libraries and frameworks, optimized for Rivos hardware. This open-source based platform offers upstream support for Rivos hardware across popular distributions, APIs and frameworks. Paired with the Rivos SDK and integrated libraries, it delivers a seamless, out-of-the-box experience for developers and CSPs alike.

1. **Energy-Efficient High-Performance SoC Solutions**
2. **Open Hardware Adoption**
3. **Flexible Open Software**
4. **Scalable Air- and Liquid- Cooled Options**

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Rivos Solutions Addressing Critical Cloud & AI Demands

Rivos has designed a Data Center class SoC providing out-of-order and parallel compute by combining Rivos-designed, fully featured 64-bit RVA23 RISC-V CPU cores with Rivos’ high-performance SIMT GPGPU.

The tightly integrated unified memory subsystem includes on-chip high-bandwidth memory(HBM) and high capacity server class DDR5 RDIMMs, both memory types are directly accessible to both the heterogeneous compute components. Multiple Ultra Ethernet NICs are integrated to deliver high bandwidth interconnect with standard protocols.

Rivos SoC has been designed to deliver exceptional performance and energy efficiency on the most demanding AI and next-generation workloads.

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

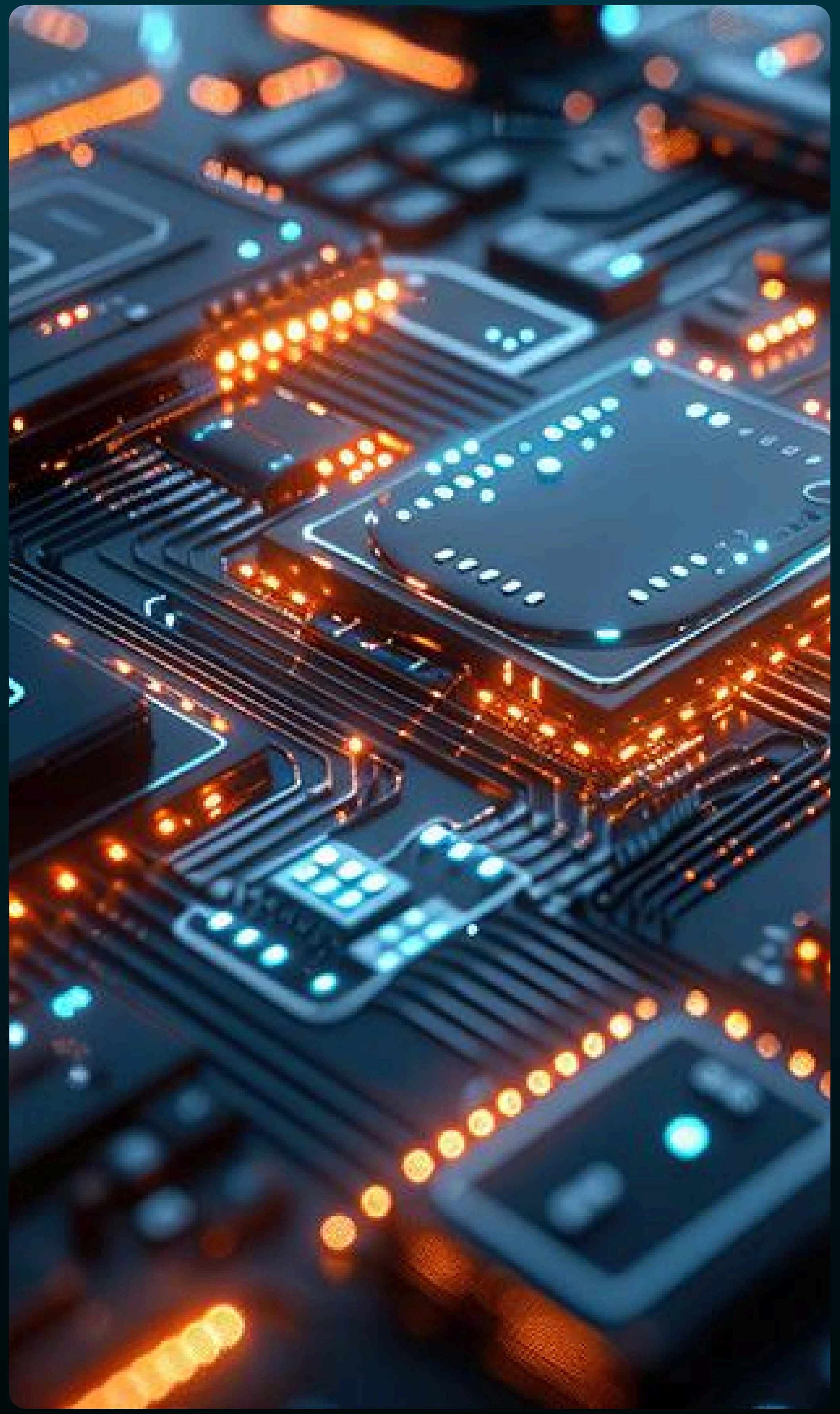
Rivos Solutions Addressing Scaling AI for Enterprise

1 Energy-Efficient High-Performance SoC Solutions

Rivos multi-chiplet SoC leverages a unified memory architecture that integrates high-bandwidth memory (HBM) in close proximity to both the CPU and GPGPU compute functionality, while offering flexible memory expansion via DDR.

This differentiated solution delivers scalable, energy-efficient performance across a wide range of AI LLM workloads; from distributed training to LLM inference and reasoning.

This architecture eliminates imbalances between compute and memory resources by minimizing external data movement and reducing power consumption. It avoids the common mismatch where additional GPU units are deployed solely to compensate for memory limitations, leading to inefficient resource utilization. Rivos' approach results in a flexible, programmable platform that helps CSPs enhance user experience, extend hardware lifespan and maximizes return on investment for next-generation AI deployments.



Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

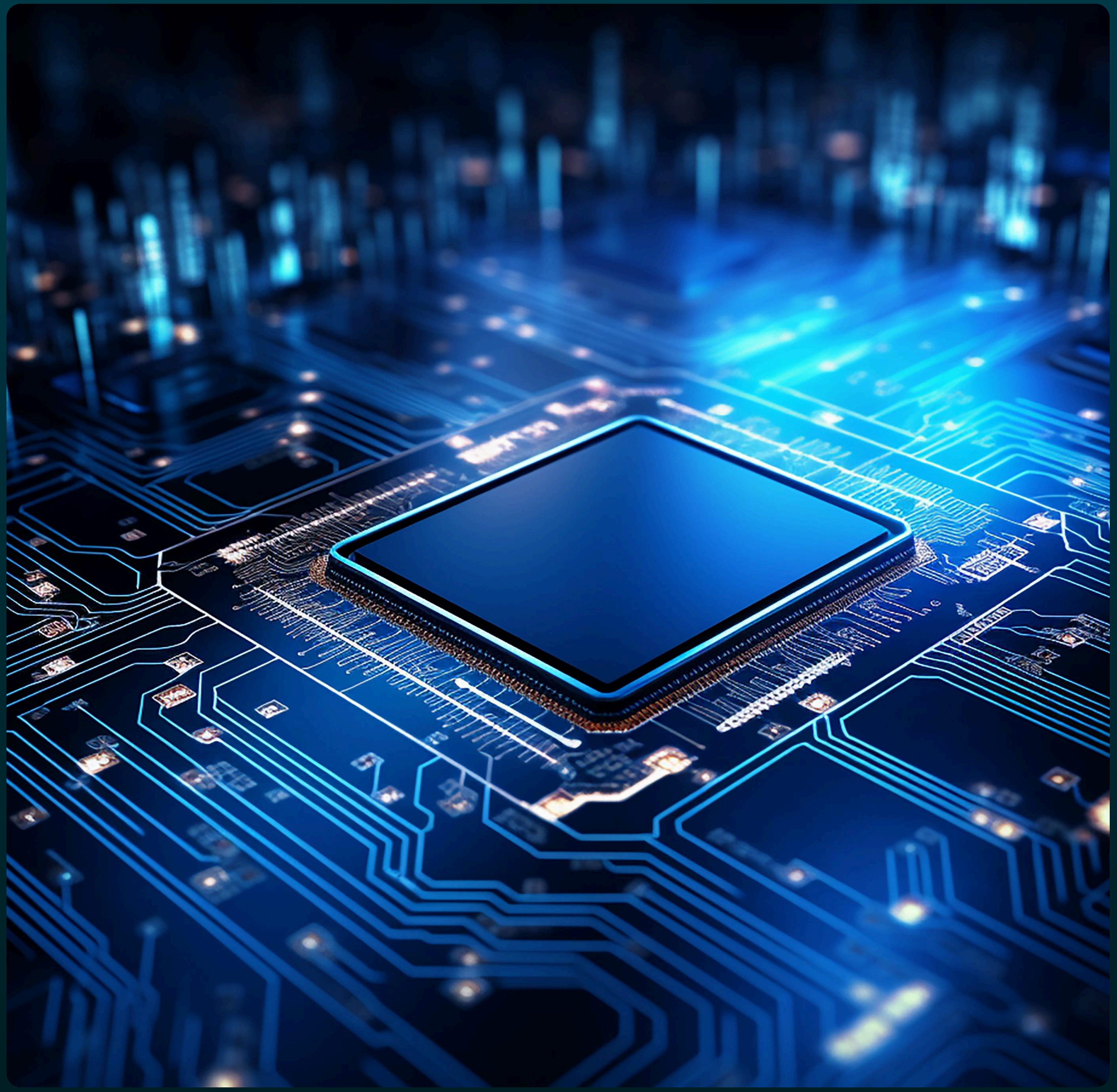
Rivos Solutions Addressing Scaling AI for Enterprise

2 Open Hardware Adoption

Rivos has based our Data Center solutions on the open standard RISC-V Instruction Set Architecture. This brings a wealth of benefits since it is a collaborative architecture that has contributors from across the semiconductor and OEM space.

As a strong believer in supporting the wider open ecosystem, for both hardware and software, Rivos is central to driving and adopting open standards.

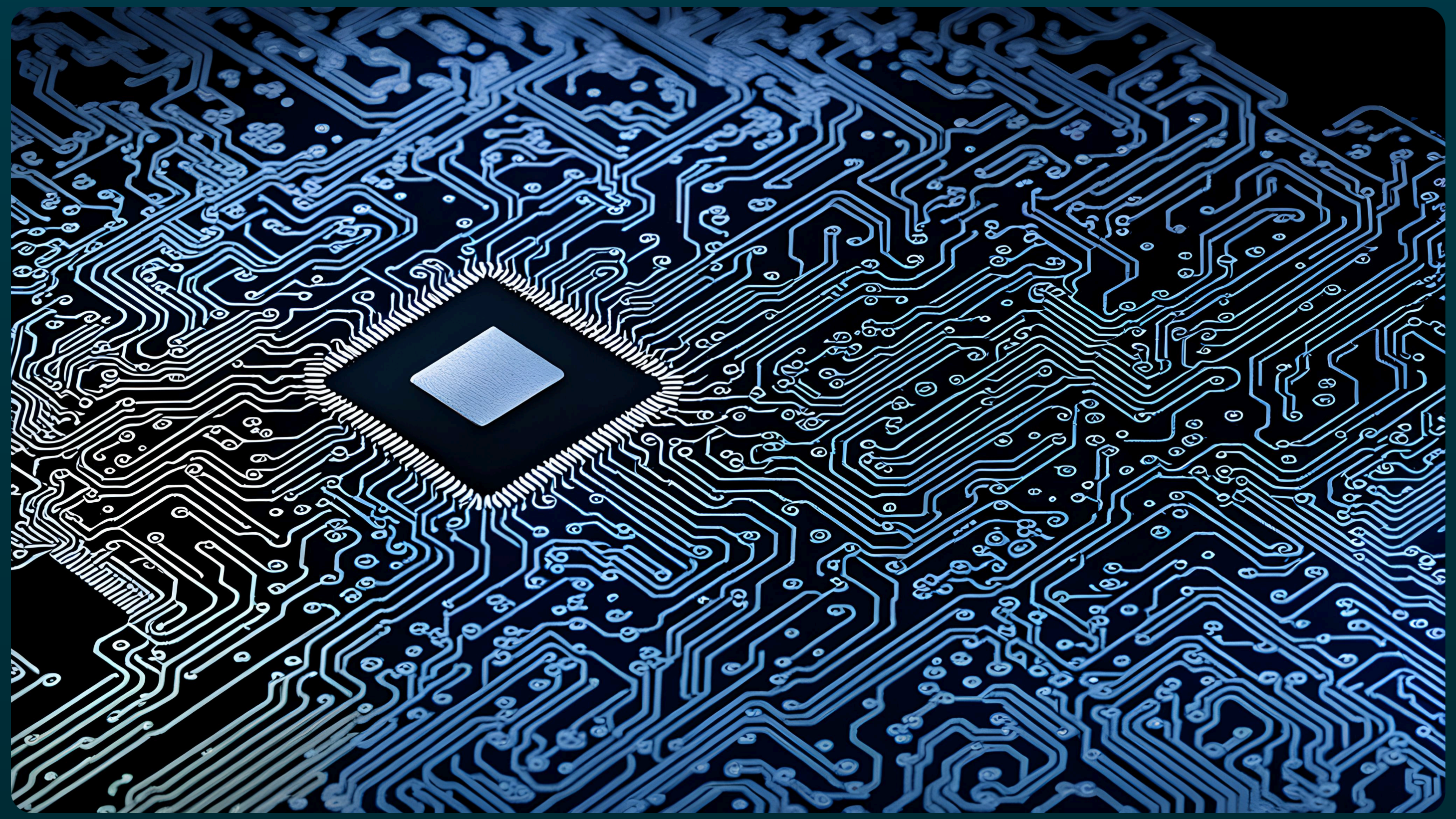
Rivos engineers are actively involved in key initiatives at RISC-V International as well as working with open source RISC-V software developers as a founding member of RISC-V Software Ecosystem (RISE).



Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Rivos Solutions Addressing Scaling AI for Enterprise

3 Flexible Open Software



Our hardware is purpose-built with the complete software stack in mind, enabling easy integration and optimized performance. A key focus is ensuring compatibility with existing libraries and algorithms used in today’s leading deep learning frameworks - simplifying adoption, reducing engineering time and costs accelerating time to value.

To support both current workloads and future AI innovations, Rivos reduces software complexity by embracing a flexible, programmable, and open-source strategy. This approach not only meets the demands of current AI workloads but also provides the adaptability needed for evolving models and frameworks.

By leveraging widely adopted open-source software components, organizations can benefit from the collective knowledge and continuous innovation driven by the broader open-source community.

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

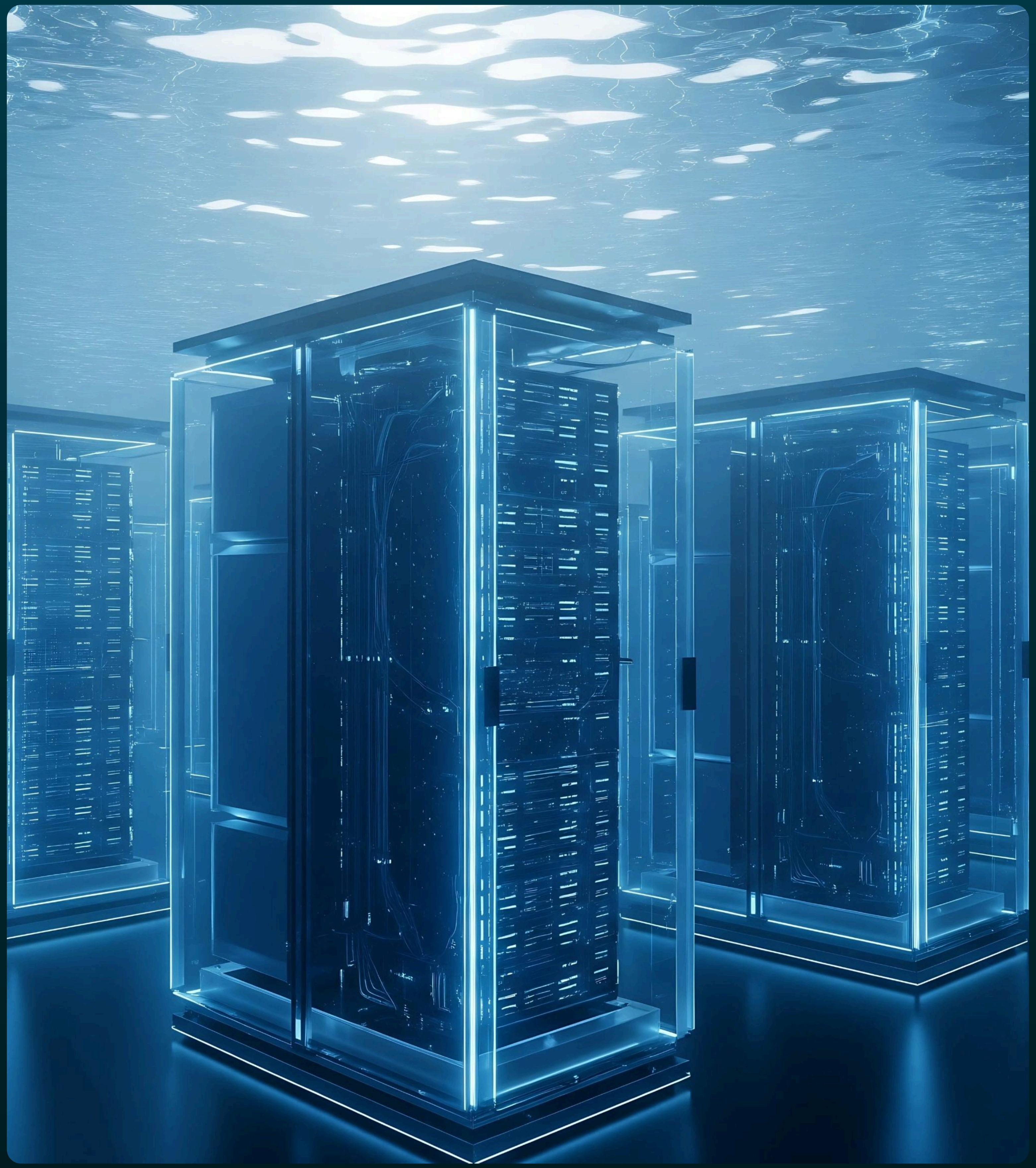
Rivos Solutions Addressing Scaling AI for Enterprise

4 Scalable Air- and Liquid- Cooled Options

IT equipment for AI needs to accommodate a growing variety of infrastructure environments and needs; liquid cooling is increasingly common for high-density AI server racks while many organizations rely on existing air-cooled infrastructure. Rivos supports options for both.

Rivos delivers solutions optimized for these environments, enabling increased infrastructure longevity, expanding the range of supported AI model types while meeting the energy-efficiency needs of air-cooled installations.

Rivos supports the traditional solutions with an x86 host, as well as the more optimized self-hosted AI appliance configurations.



Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Conclusion

Rivos' data center-class SoC delivers a powerful and flexible solution designed for the evolving demands of AI and next-generation workloads. It combines high-performance RISC-V RVA23 CPU cores with a Rivos' SIMT GPGPU and a unified memory architecture, featuring both on-chip HBM3e and DDR5 RDIMMs. Rivos enables exceptional performance, energy efficiency, and low-latency data access across a range of AI tasks, from training to inference and reasoning.

Beyond compute, Rivos recognizes the practical infrastructure needs of its customers. The platform supports both **air- and liquid-cooled configurations**, enabling deployment in traditional data centers as well as high-density AI environments. While **liquid cooling is fully supported** for modern high-performance AI server racks, **air-cooled options** are available for organizations with existing infrastructure, making AI more accessible across diverse operating environments.

Deployment flexibility is further enhanced through a range of system integration options:

- **PCIe CEM card**: Plug-and-play solution that fits standard GPU slots in x86 servers, allowing scalable compute acceleration with multi-card support
- **High Density HPC/AI Server (UBB style)**: Dual x86-hosted base system connected via PCIe Gen 6 to Rivos GPGPU accelerators, ideal for standalone rack to large cluster deployments
- **Self-hosted AI appliance server**: An optimized, fully integrated solution with built-in networking and management capabilities for turnkey AI cluster deployment

Rivos also supports both **traditional x86-hosted** configurations and **self-hosted appliance models**, giving customers the flexibility to optimize for performance, manageability, or existing infrastructure compatibility.

> **By offering a complete, standards-based platform that combines advanced compute, versatile cooling, and deployment options, Rivos is enabling cloud service providers to build energy-efficient, high-performance solutions that scale with future AI demands, without compromising compatibility, flexibility, or time to market.**

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

About Rivos

Rivos is democratizing AI with affordable, high-performance Data Center solutions built on an open, scalable, and energy-efficient SoC architecture. Its programmable hardware handles the full AI pipeline, from Training to Reasoning, and Data Analytics, on one unified platform. With a software-first, open-software approach, Rivos supports today’s AI models and is programmable enabling it to flex as AI models change in the future.

Purpose-built for large-scale workloads, Rivos solutions combine RISC-V CPUs with Rivos-developed GPGPU accelerators to deliver high performance with energy efficiency. Founded in 2021 and based in Santa Clara, Rivos raised $250M in Series A-3 funding in April 2024 and is rapidly growing its global presence and engineering team.

Contact Us:

in **https://www.linkedin.com/company/rivos-inc/**

✉ **https://rivosinc.com/company/contact-us**

Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.