Scaling AI Models for Enterprise

White Paper

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

Table of Contents

1 Introduction

2 Serving and Scaling Large Language Models (LLMs)

3 Scaling AI Model Services in the Real World

4 Disaggregated Serving at Scale

5 Supervised Fine Tuning of LLM

6 Accelerating Structured Data Analytics

7 Retrieval-Augmented Generation (RAG)

8 Reasoning Models

9 Conclusion

10 Rivos Solutions Addressing Scaling AI for Enterprise

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

Introduction

Enterprise organizations are rapidly investing in artificial intelligence (AI) to drive innovation, improve operational efficiency, and gain competitive advantage. However, many are struggling to translate these investments into meaningful business value. A central challenge lies in the ability to scale AI effectively across the Enterprise to address real-world complexities while ensuring a strong return on investment (ROI).

High Quality Data

One of the foundational requirements for successful AI deployment is high-quality data. Training AI models rely on large volumes of accurate, consistent, and well-labeled data. Incomplete, inconsistent, or biased datasets can lead to unreliable insights and unintended consequences, undermining confidence in AI-driven decision-making.

Seamless Integration

Equally critical is the ability of AI systems to integrate seamlessly with existing Enterprise infrastructure. Legacy systems, proprietary software, and applications lacking robust APIs often inhibit smooth data flow, resulting in duplication of efforts or the inability to fully utilize available data. These integration challenges can significantly diminish the ROI of AI initiatives.

Cost Considerations

Cost considerations also present a major hurdle. The high upfront investment required for Enterprise-grade AI solutions, combined with uncertainty around long-term returns, can make it difficult to secure sufficient funding. Organizations must also weigh the trade-offs between cloud-based AI platforms, which offer scalability and lower entry costs, and on-premises solutions that provide greater control over data sovereignty and security.

Scaling

Finally, moving from proof of concept to full-scale production remains a common barrier. While pilot projects may demonstrate potential value, Enterprises often face challenges in operationalizing and scaling these solutions to reach broader business units or customer bases.

This paper explores the critical considerations and practical strategies for scaling AI in the Enterprise, with a focus on overcoming integration, data, and infrastructure challenges to unlock real and sustained business value.

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

Serving and Scaling Large Language Models (LLMs)

Serving and scaling LLMs presents unique challenges due to the tightly coupled requirements of memory capacity, bandwidth, compute performance, and high-speed interconnects. Efficient deployment also demands sophisticated load balancing and the ability to manage vast contextual information across multiple users and interactions.

Iterative Inference and KV Caching: Inference in LLMs is inherently iterative. As each new token is generated, it must draw upon information from previous steps in the process. This is enabled by a key mechanism known as Key-Value (KV) caching, a core component in transformer-based models. In this approach, the key and value matrices generated during earlier stages of processing are cached and reused for subsequent token generation, allowing the model to retain context efficiently.

Context Preservation for Extended Interactions: For extended interactions, such as multi-turn conversations or document summarization, LLMs must preserve continuity. This can be achieved either by re-supplying the entire conversation history with each new prompt or by retaining and reusing the KV cache across inference steps. The latter approach significantly reduces computational overhead and latency.

Scaling Considerations and Optimization: As usage scales, the KV cache becomes a significant driver of memory consumption. Optimizing across this memory footprint, while balancing compute and interconnect efficiency, is critical for delivering responsive and cost-effective AI services at scale.

Transformer-based Inference typically comprises two distinct phases:

Pre-fill (Prompt Processing)

This phase is compute-intensive and involves generating the initial KV cache from the user's prompt. It is resource-heavy only until the first token is produced.

Decode (Output Generation)

This phase is predominantly memory-bound. Due to the auto-regressive nature of LLMs, each token is generated sequentially, with dependencies on all preceding tokens. During this phase, the KV cache must be updated with each new token, increasing memory demand proportionally to the sequence length, number of users, and conversational turns. Performance here is typically measured by **Time per Output Tokens (TOPT).**

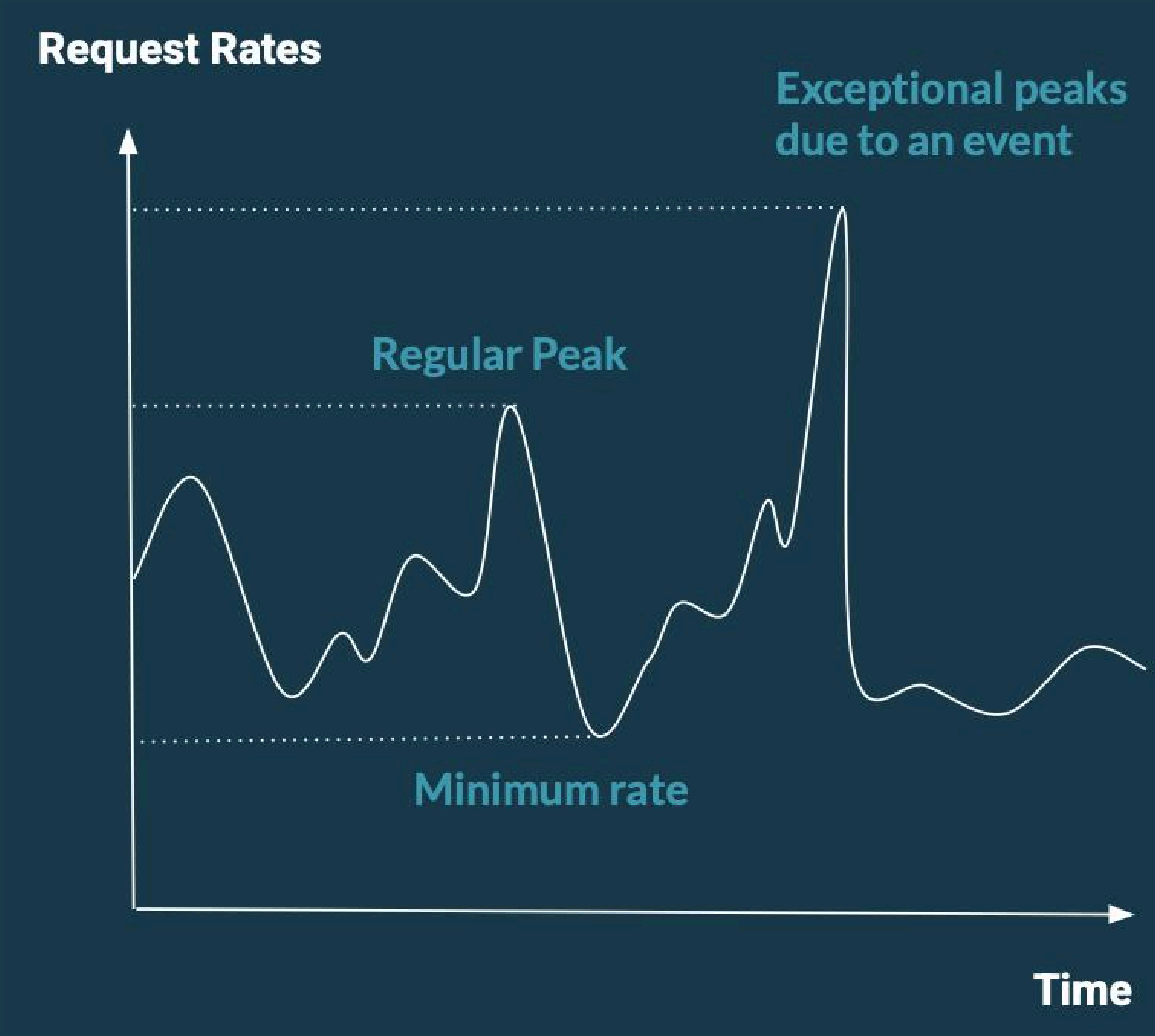
> In this paper, we will examine a variety of LLM architectures and explore how heterogeneous infrastructure, leveraging a mix of compute and memory configurations, can be strategically applied across different stages of the deep learning lifecycle. Our goal is to help identify optimal deployment strategies for maximizing performance, scalability, and ROI in real-world enterprise environments. Performance in this phase is often benchmarked by **Time to First Token (TTFT).**

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

Scaling AI Model Services in the Real World

Multiple Models with unpredictable variables: Request Size, Prompt Complexities, and Request Rates. For online processing:

- Difficult to predict specific loading
- Need to design with **scalable, flexible** approach
- **Cost effective** for minimum request rates
- **High-performance** during the peak periods



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

Disaggregated Serving at Scale

Right-Size fit performance/memory to adapt to unpredictable workload levels

The uptake of new models has brought with it new scaling challenges and how to process these requests, whilst dealing with the level of unpredictability that they have created. Systems need to be designed to be able to support multiple model types which can have their own time and request rates and complexities; whilst adapting to handling peaks. These disparate requirements require efficient and effective leveraging of the compute, memory and communications resources available both locally as well as disaggregated across multiple servers and nodes.

To achieve scalable system architecture, several key aspects must be considered:

- First is the **separation of resources**; CPU, GPGPU (or other accelerators), memory, and storage can be decoupled, sometimes being on different physical systems. This disaggregation enables greater flexibility and resource efficiency.

- However, it has communications and latency overheads so a **high-speed network** fabric is essential to interconnect these components seamlessly, enabling them to communicate as if they were part of a unified system.

- Lastly, **dynamic resource allocation** is crucial; it allows workloads to consume only the resources they need, minimizing over provisioning and improving overall system utilization.

📌 Disaggregated Serving

Architectural approach where compute, memory, and storage resources are separated rather than being tightly bundled in a single server or node.

This enables more flexible and efficient scaling, especially for AI workloads, compared to monolithic serving.

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

Disaggregation Benefits

- **Flexibility**

  Accommodates a wide range of AI models, from LLMs to smaller models, with varying compute and memory requirements.

- **Efficiency**

  Resources can be shared and reused across different workloads, improving the overall utilization.

- **Cost-Effective Scaling**

  Allowing memory and GPGPU capacity to grow independently, avoiding the expense of overprovisioning, and investing where the resources are needed.

> **Ultimately, this alignment of infrastructure with actual workload demands leads to better total cost of ownership (TCO) and a stronger return on investment.**

Disaggregation Challenges

- **Multiple Model Support**

  Varying sizes, request rates and prompt complexities. None of these can be predicted in advance, as they also fluctuate over time. This variability demands dynamic adaptation to changing workloads while maintaining consistent service quality and meeting strict QoS thresholds.

- **Pre-fill and Decode Differing Compute Resources**

  Challenging to efficiently schedule accelerators over both phases. This has resulted in a shift towards disaggregated serving where nodes are split functionally into pre-fill and decode phases.
  This brings its own challenges when it comes to the KV cache, which has to be exported to the nodes that will use it during the decode phase. This can affect the latency and power consumption due to data moving between nodes.

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

Supervised Fine Tuning of LLM

Aligning the models behavior to the use case

Supervised Fine Tuning (SFT) of a Large Language Model (LLM)

Process of adapting a pre-trained LLM to perform specific tasks using labeled datasets that were used to train a model with supervised learning.

With open source AI increasing, the ability to adapt LLMs to new domains through the use of publicly available open weights is improving. Meaning a pretrained model can be fine-tuned or analyzed without retraining it from scratch.

callScanQr()
secondButton = "No"
eventSecond = DialogInterface.OnClickListener { dialog, _ ->
dialog.dismiss()
cancelScan()
private fun observeScanQr() {
viewModel.scanQrResult().observe(viewLifecycleOwner, Observer {
it?.let { result ->
Log.d("RESULT","====> $result")
displayScanResultSuccessDialog()
Toast.makeText(requireContext(), it.message, Toast.LENGTH_SHORT).show()
1 10 1001 01010110
0 1 1 0010

Fine Tuning has less compute requirements compared to the original training, however the memory requirements for the model parameters often exceeds the current system's memory available. This can result in inefficiencies because of having idle compute functionality installed to only accommodate the memory requirements.

Multi-turn dialogue generation are used to improve the understanding of the dialogue content and enable LLMs to generate appropriate responses by incorporating historical dialogue.

Fine Tuning helps **align the model's behavior with desired outcomes**; such as answering questions, generating summaries, or following instructions, by learning from known input-output examples.

It is a form of transfer learning where a general-purpose LLM (pre-trained on vast, generic data) is trained further on a smaller, task-specific dataset, where the correct responses are known (supervised).

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

Supervised Fine Tuning Benefits

- **Enhanced Task-Specific Performance**

  Refines model’s behavior by further training it on a curated dataset of input-output pairs. making the model more helpful and better aligned with the needs of individual users or organizations.

- **Efficiency**

  Performed using relatively small datasets, making it an efficient approach for targeted improvements.

- **Domain Adaptation**

  Allowing the model to specialize in specific fields such as legal, finance, or healthcare.

Supervised Fine Tuning Challenges

- **Need for High-Quality, Labeled Data**

  Risk of overfitting when using small or narrowly focused datasets. There's also the potential for catastrophic forgetting, where the model loses its general knowledge during specialization.

- **Compute Cost**

  While the compute cost of fine-tuning is lower than pre-training, it remains significant, with memory loading often being the primary constraint.

- **Multi-Turn Historical Dialogue**

  Brings complexity of effectively leveraging conversational data and the increased training time, which can negatively impact user experience.

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

Accelerating Structured Data Analytics

Increasing the quality of the data used for AI Models

Structured Data Analytics

Data analytics functionality must prioritize efficient use of available resources.

Rather than relying solely on scale-out strategies to achieve performance, it is crucial to first **scale up**, to fully leverage the memory, bandwidth, and compute capabilities of a single socket.

This approach ensures optimal resource utilization and lays a strong foundation before expanding to multi-node architectures.

Large-scale analytics of structured data, typically stored in databases, continues to be a key driver of business value by informing smarter decision-making. As data volumes grow rapidly, organizations must not only scale their infrastructure but also **optimize for cost efficiency** to extract the maximum value from their data assets.

While structured data has long been foundational, its role has taken on new significance as high-quality data becomes critical input for AI models. Today, the focus extends beyond data quality to include maximizing return on investment from the structured data already available.

Historically, CPU-only approaches have relied on partitioning the dataset and processing it across a large number of nodes to achieve scalability. However, this method introduces significant overhead due to the costs of data partitioning and the communication required between nodes, over Ethernet. Additionally, CPU architectures are primarily optimized for high single-thread performance and are constrained by the number of available cores and memory bandwidth per socket.

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

Benefits of a Heterogeneous Compute for Structure Data Analytics

- **Enhanced Performance**

  An approach that integrates both CPU and GPGPU functionality within a single node significantly enhances performance by enabling tasks to be assigned to the most suitable architecture.

  - CPUs are well-suited for sequential, control-intensive, or low-latency operations, while GPGPUs excel at handling massively parallel workloads such as filtering, scanning, and aggregations.
  - This division of labor allows GPGPUs to process high-throughput, data-parallel operations more efficiently, resulting in faster query execution and superior performance per watt.

By increasing the computational capabilities of each node, the overall number of nodes required can be reduced, thereby decreasing inter-node communication. This leads to improved resource utilization, lower latency, and less data movement, ultimately reducing energy consumption and operational costs.

Challenges of a Heterogeneous Compute for Structure Data Analytics

- **Data Movement Costs**

  The main data movement costs incurred are due to movement to and from the GPGPU compute functional areas, so by reducing this has multiple benefits across energy consumption as well as user experience.

- **Poor Scheduling**

  Effective scheduling is critical to achieve optimal load balancing and task distribution between CPUs and GPGPUs. Poor scheduling can lead to underutilization of resources or bottlenecks. Additionally, data movement across components and nodes must be minimized, especially in scale-out architectures, to avoid increased latency and excessive energy consumption, both of which negatively affect user experience and operational costs.

- **Programming**

  Programming heterogeneous systems also presents challenges, as it typically requires expertise in both CPU and GPU programming models and APIs. Furthermore, traditional GPU-based solutions capable of supporting structured data analytics have historically been expensive to acquire and operate, due to high power consumption and infrastructure demands. This makes them less cost-effective for smaller, on-premises deployments.

However, if structured data analytics can be natively integrated with AI model execution within the same system, it offers a compelling advantage. Analytics insights can directly enhance AI outcomes without requiring separate infrastructure or data pipelines.

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

Retrieval-Augmented Generation (RAG)

Accelerating Unstructured Data

By augmenting the generative process with up-to-date and relevant data, RAG enables more **accurate and context-aware responses**, making it especially valuable for Enterprise and knowledge-intensive applications.

These strengths make RAG particularly well-suited for applications where factual precision, up-to-date knowledge, and secure handling of internal or confidential data are critical. The benefits of RAG also extend to other areas such as Agentic AI where the quality of the data is vitally important to ensure Agent decisions are made based on the best known data at the time of the decision.

RAG at Scale

As RAG scales, it introduces several key challenges that impact implementation. **Managing vast datasets** and supporting potentially thousands of concurrent users adds significant complexity, requiring a robust, distributed architecture capable of efficiently handling both processing and storage at scale. Additionally, RAG systems must account for **continuously evolving data**, ensuring that updates are promptly reflected in responses. Maintaining **low-latency performance** is also critical, particularly for interactive applications where responsiveness directly affects user experience.

📌 Retrieval-Augmented Generation (RAG)

AI framework that enhances the capabilities of large LLMs by integrating them with external knowledge sources. This approach improves the quality of AI inference by providing access to real-time, domain-specific or organization-specific information.

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

RAG Benefits

- ## Factual Accuracy & Transparency

  It improves factual accuracy by dynamically retrieving current information, and it provides transparency by allowing traceability to the original source documents used in generating a response.

- ## Flexible & Future-Proof

  The framework is also flexible and future-proof: its knowledge base can be updated independently of the model, ensuring that the system doesn't become outdated.

- ## Simplifies Data Management

  It simplifies data management, making it easier to update or remove information from the retrieval index.

> **Generative AI models are inherently limited by the static nature of their training data, which introduces a cutoff that prevents them from accessing future or real-time information. RAG addresses this by enabling access to the latest and most relevant data when forming responses.**

RAG Challenges

- ## Balancing Accuracy and Coverage

  Access to the latest data introduces new challenges, such as ensuring that the retriever accurately understands the context and nuances of user queries to return truly relevant information. Balancing accuracy and coverage through effective recall metrics is critical to delivering high-quality results. Additionally, integrating external knowledge with a model's built-in understanding can be complex, particularly in enterprise settings where data conflicts must be resolved without compromising confidentiality and security.

- ## Large Datasets & Complex Integration

  RAG systems must also manage significant operational hurdles. The rapid growth of knowledge bases increases storage requirements and adds complexity to indexing, demanding more advanced strategies to ensure fast and reliable retrieval. Updating large datasets introduces overhead in terms of both time and computational resources, especially when consistency and integrity must be maintained.

- ## Latency

  As knowledge bases expand, latency can become a bottleneck, especially for real-time applications. Furthermore, efficient parallelization and pipeline optimization are essential to scaling RAG systems, requiring careful coordination between retrieval and generation processes to maximize throughput and resource utilization.

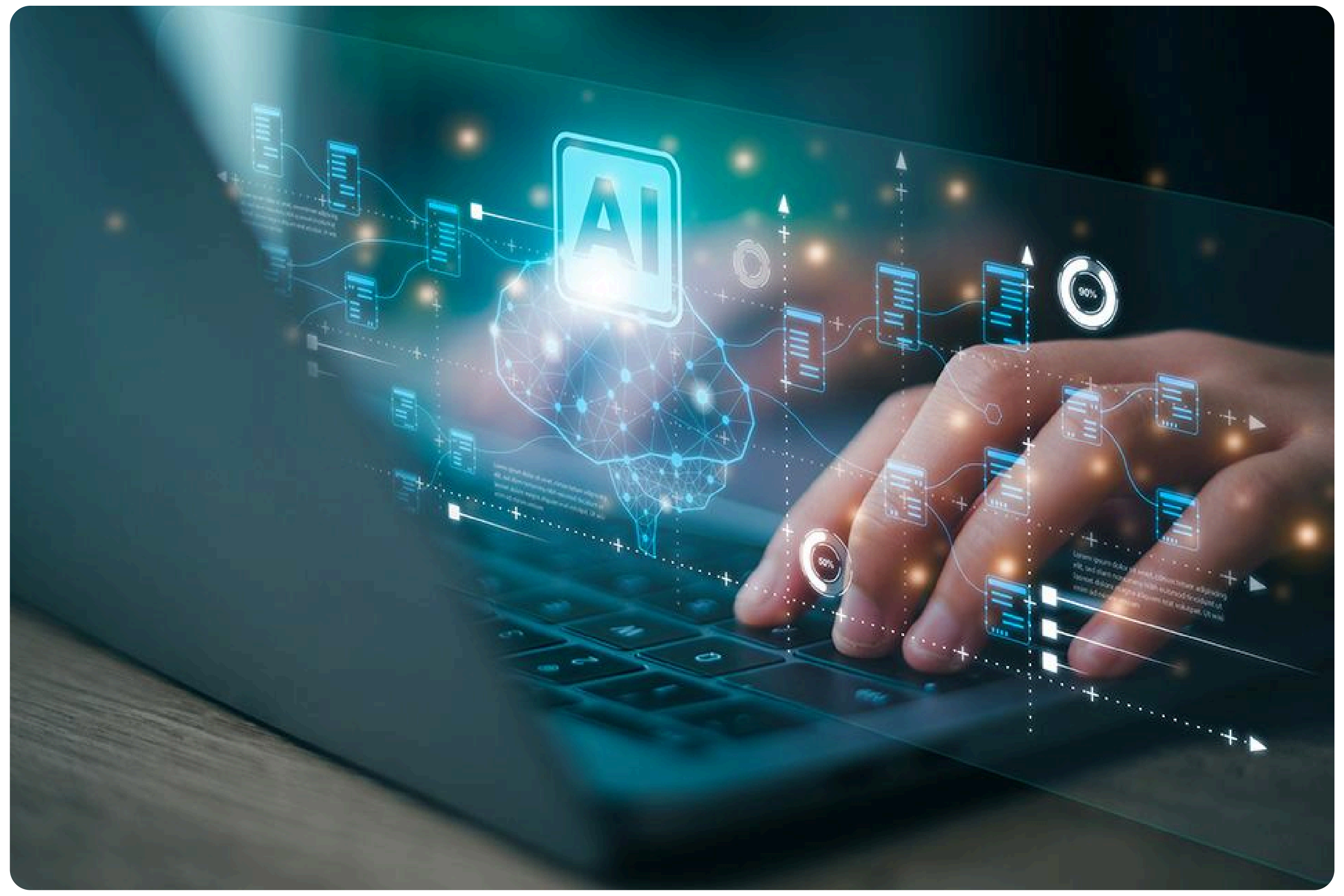
Copyright © 2025 Rivos Inc. All rights reserved. Ri logo and Rivos are trademarks of Rivos Inc.

Reasoning Models

Serving Reasoning Models at Scale

Reasoning models are a key functionality to maximize the return value on AI investments. They offer a route to improved performance and higher accuracy and to solving new problem areas, by breaking the problems down into multiple, smaller steps and solving the problem in a more comprehensive way.

The key characteristic of a reasoning model is it spends more time generating “thinking tokens” which increases the time during the inference phase, resulting in the need for larger amounts of KV cache. Thinking tokens are generated in an autoregressive manner meaning where future values in a time series are predicted based on past values of the same series, so a single request uses the resources and memory for a longer time, resulting in more memory required.



Reasoning Models

Reasoning AI LLMs are advanced large language models designed not just to generate text, but to logically analyze problems, follow step-by-step reasoning, and produce well-structured, justifiable answers. They aim to mimic human-like thinking rather than simply predicting likely words.

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

Benefits of Reasoning Models

- **Improved Decision Making**

  Reasoning models have the ability to analyze complex data sets and infer logical conclusions from the data they have. This can help users to see alternative conclusions especially under uncertain scenarios.

- **Transparency of Decisions**

  In reasoning models the rationales applied give clearer justifications to the decisions that were made. Enabling increased trust to be built as well as traceability to the route for the decision making process, where humans alone can miss critical linkages as well as biases. They can check for logical inconsistencies or incorrect assumptions.

- **Reusability**

  Logic and rule based reasoning models can often be reused or extended without requiring retraining enabling new circumstances to be applied or additional domain knowledge to be included.

Challenges of Reasoning Models

- **Reasoning Scalability & Distributed System Complexity**

  Statistical reasoning models need vast high-quality data sets. Real-time reasoning is hard to achieve at scale. Sharding brings the complexity of managing tables that are distributed across multiple servers

- **Memory Constraints**

  These challenges require not only compute functionality scalability but also the memory subsystem to enable larger amounts of KV cache closer to the computing and processing elements. If insufficient KV cache is locally available results in memory spread across a large number of chips, requiring high bandwidth requirements between the nodes impacting both the resultant performance and power consumption. Closer the KV memory is the compute elements improves responsiveness and reduces power requirements.

- **Infrastructure Cost Inefficiencies**

  Limited and localized memory access is a key barrier to delivering peak user access and performance. By being able to trade off between performance and capacity can improve a service's scalability and range of users that can be supported. Since memory capacity often determines the number of nodes required, it is expensive and inefficient to add nodes just to get access to increased memory capacity, when the computational and networking requirements are not needed.

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

To support a wide range of AI models effectively, a multi-model AI solution must incorporate several fundamental techniques to ensure optimal performance, scalability, and efficiency.

Parallelism

Parallelism plays a key role in accelerating computation and improving overall system throughput.

Efficient Scalability

Efficient scalability is essential, particularly through extended memory support that minimizes the need to distribute workloads across multiple devices.

Reduced Data Movement

Reducing data movement is equally critical, as it not only enhances performance but also significantly lowers power consumption.

Flexible Memory

A flexible memory architecture further supports this by accommodating growing model and dataset sizes without requiring additional hardware.

Comprehensive Software

Comprehensive software solutions are needed to orchestrate and manage these resources seamlessly.

Network Infrastructure

Underlying network infrastructure must be tuned to the specific characteristics of the AI workload, ensuring that frequently accessed data remains close to the compute resources to reduce latency and maximize throughput.

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

Rivos Solutions Addressing Scaling AI Challenges

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

Rivos delivers a differentiated approach to these challenges. Offering secure, scalable, energy-efficient data center solutions, based on a multi-chiplet SoC, that can be tailored to meet a business's needs, whether air or liquid cooled.

All our solutions are built on a full standard software stack, optimized for Rivos hardware. This open-source based platform offers upstream support for Rivos hardware across popular 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 Scaling AI for Enterprise

Rivos has designed a Data Center class SoC providing standard and parallel compute by combining Rivos' high-performance fully featured 64-bit RVA23 RISC-V CPU cores with cache coherence to Rivos' designed SIMT GPGPU.

A tightly integrated memory subsystem includes both on-chip high-bandwidth HBM and high capacity server class DDR5 RDIMMs, both memory types are accessible directly to both the heterogeneous compute components. Given the importance of connectivity, multiple Ultra Ethernet NICs are integrated to give high bandwidth with a standard interconnect.

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.

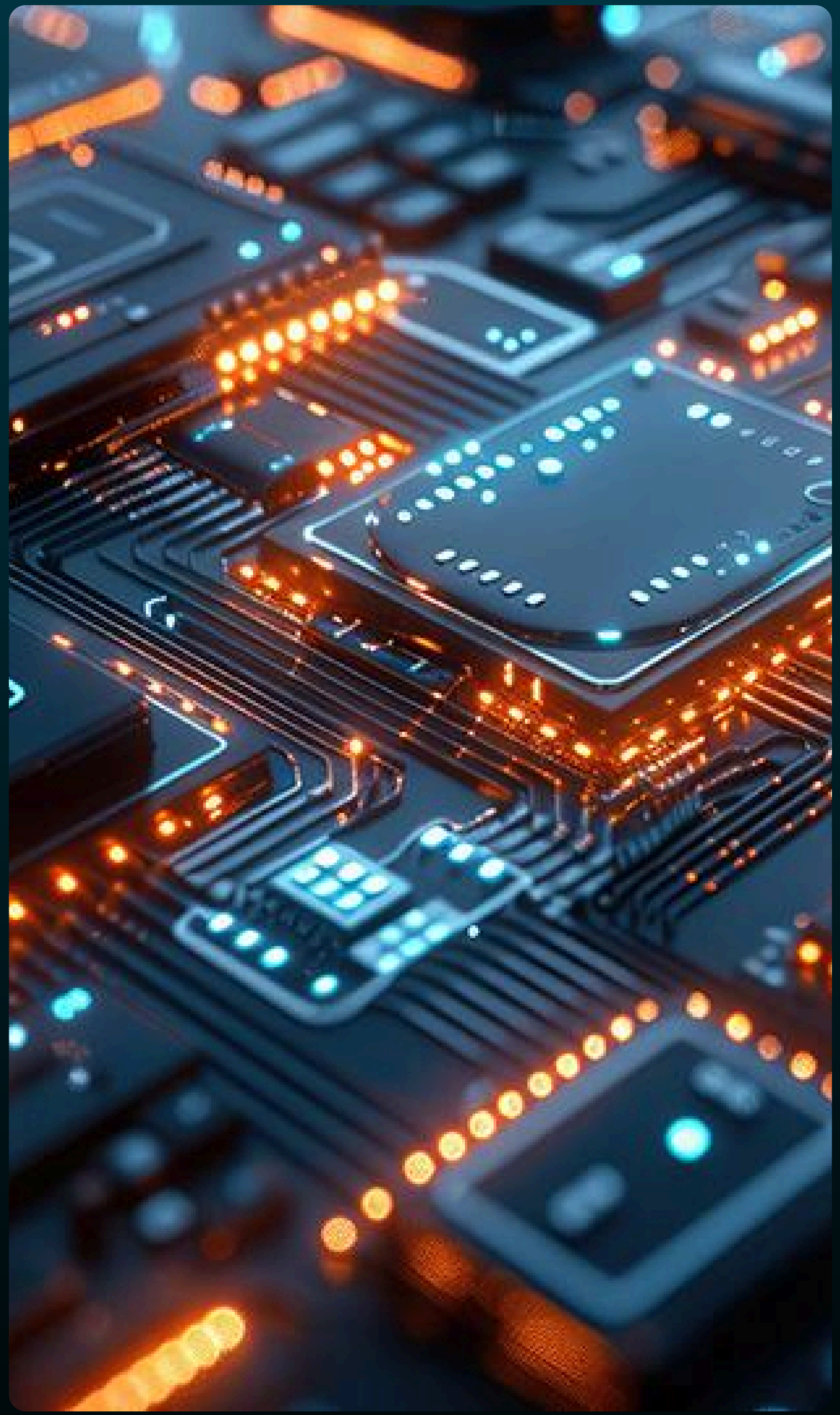
Rivos Solutions Addressing Scaling AI for Enterprise

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 Enterprises to introduce and scale their AI usage, whilst easily integrating it alongside existing air-cooled installs.



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

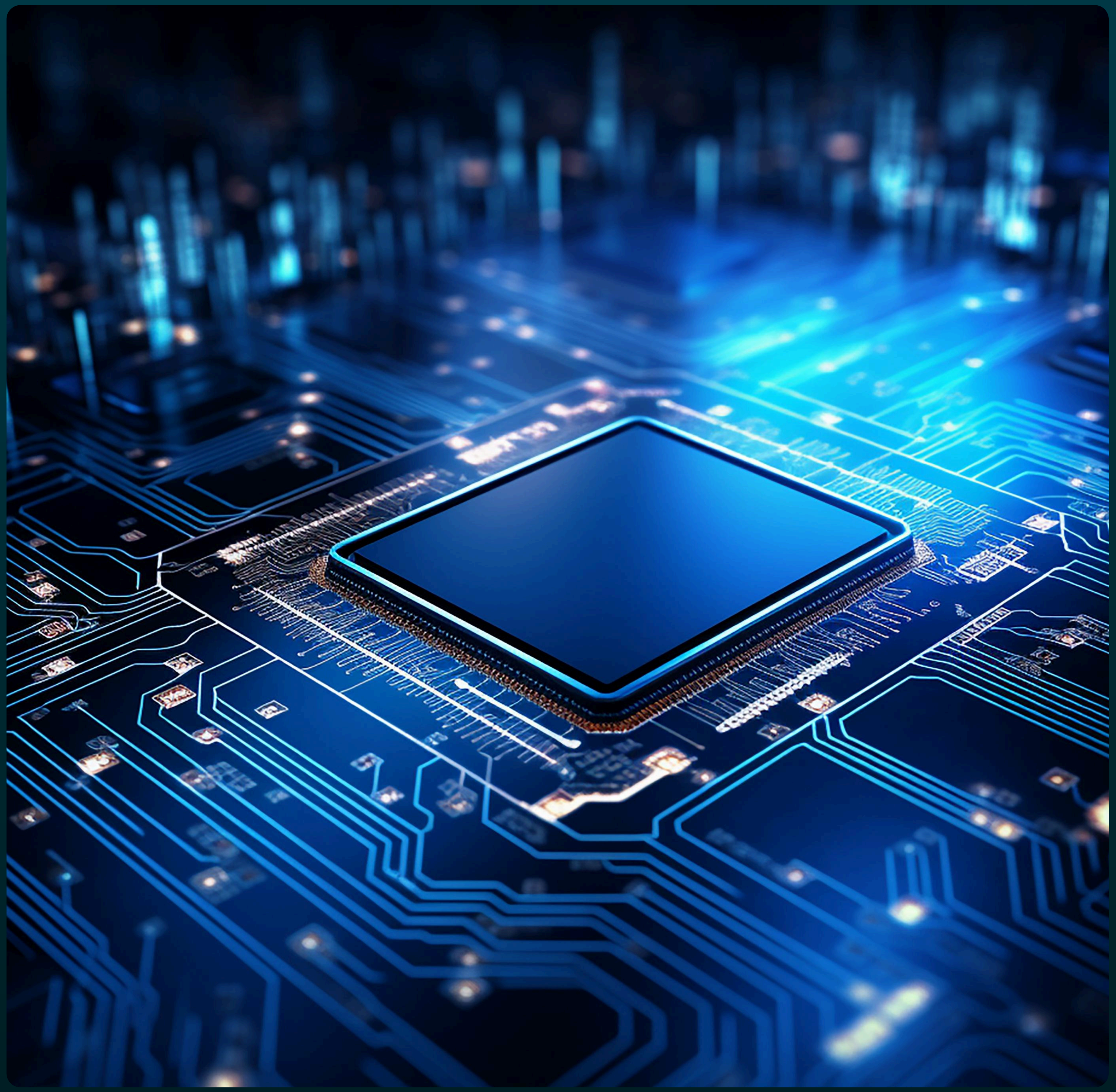
Rivos Solutions Addressing Scaling AI for Enterprise

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.

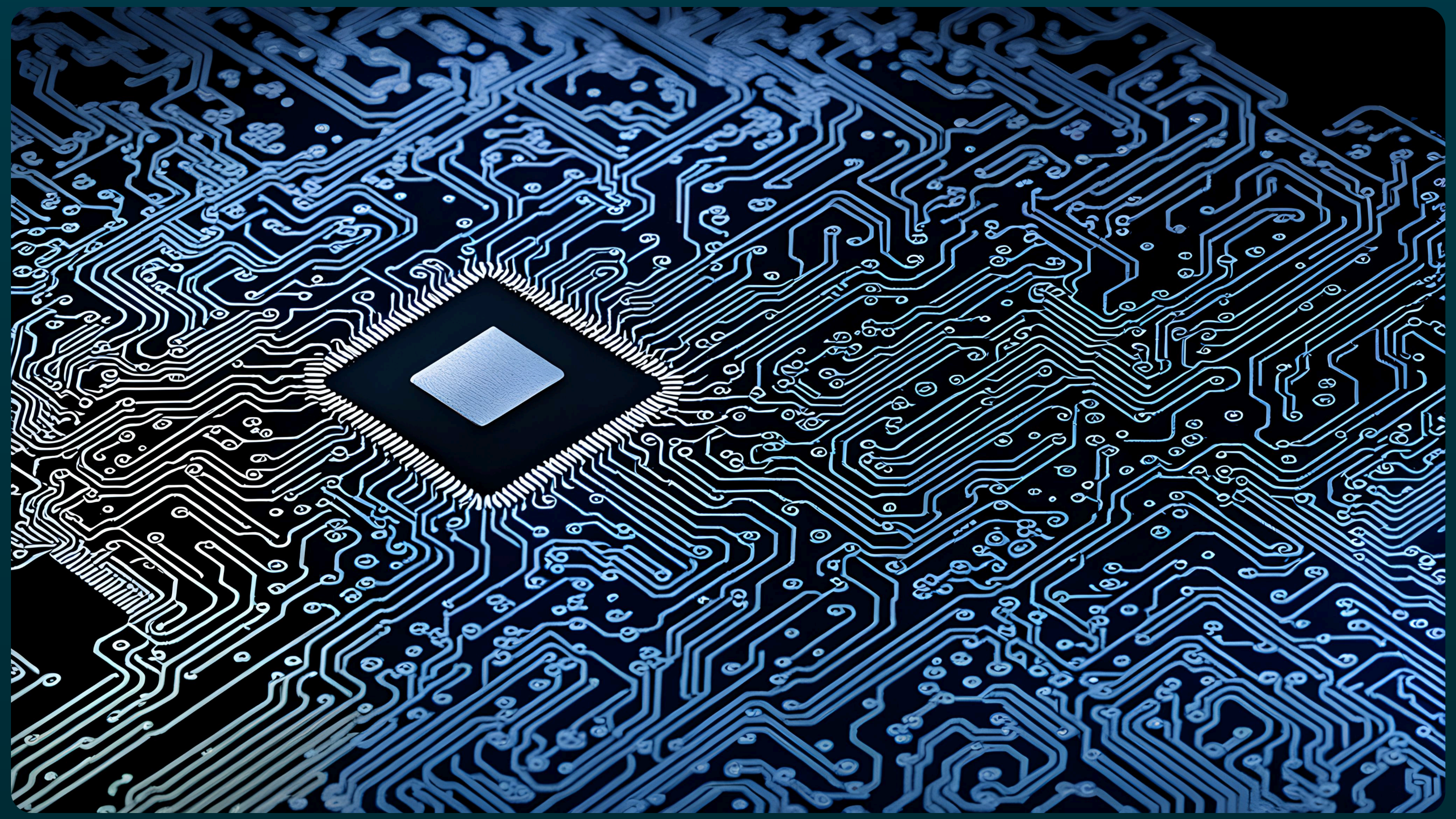
Rivos engineers are actively involved in key initiatives at RISC-V International as well as working with open source RISC-V software developers as a founding member of RISC-V Software Ecosystem (RISE).



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 data-parallel algorithms used in today's leading deep learning frameworks, simplifying adoption and 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, Enterprises can reduce their reliance on in-house expertise while benefiting 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

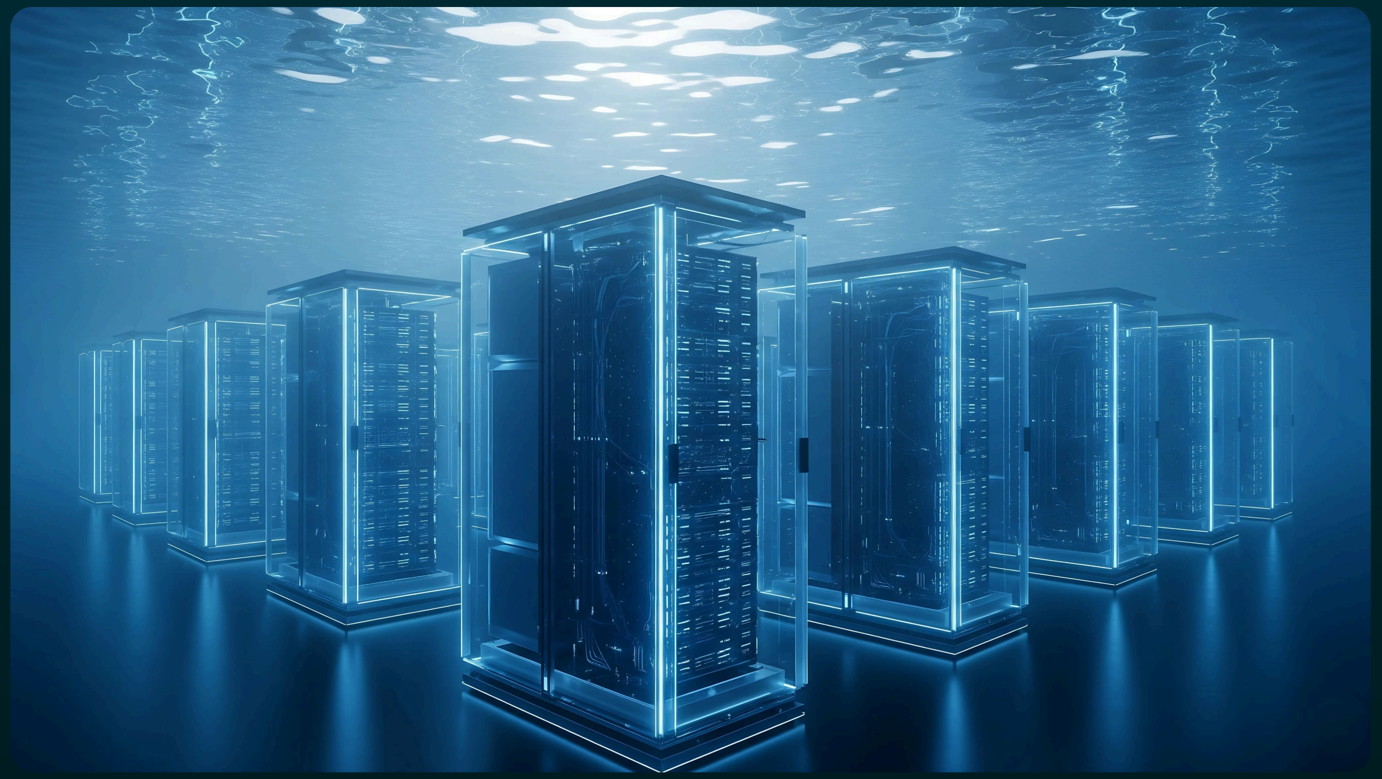
4 Scalable Air- and Liquid- Cooled Options

Liquid cooling is increasingly common for high-density AI server racks and is fully supported by Rivos. Yet many Enterprises rely on existing air-cooled infrastructure.

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.

- **PCIe CEM card fits with existing server GPU slots**: Provides accelerated compute to x86 servers with the option of using multiple cards
- **Standard AI server for standalone or rack cluster use**: UBB Style - dual x86 hosted base system with PCIe Gen 6 links to Rivos GPGPU accelerators
- **Self-hosted AI appliance server with full network and management support**: Can be part of a managed AI cluster



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

Conclusion

Scaling AI functionality within an Enterprise will be driven by the domain and the specific needs of the users of the systems. Flexibility and adaptability are key to fully enabling the benefits of AI within an Enterprise. Also the ability to scale - turning early pilots into full production installs that grow as the Enterprise needs grow in an efficient and cost effective way.

Rivos' data center-class SoC delivers a powerful and flexible solution tailored 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 HBM 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.

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 enterprises 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.