RISC-V for AI/ML

Progress, Innovation & the Road Ahead

Wednesday, September 24, 2025

RISC-V®

RISE
RISC-V Software Ecosystem

RISC-V for AI/ML

Progress, Innovation & the Road Ahead

Agenda:

- (5 min) RISC-V Developer Updates - Amber Huffman
- (10 min) RISC-V International Remarks - Andrea Gallo
- (10 min) Yocto RISC-V Progress - Valentina Fernandez Alanis
- (20 min) Deep Dive on AI/ML Plans - Ludovic Henry
- (15 min) Q&A

RISC-V North America Summit

Join us at the RISC-V North America Summit October 22-23 in Santa Clara, CA – there are lots of ways to get involved and learn more!

What to expect at the Summit

- Tracks for Hardware and Software
- Separate ticket from Summit - $150 attendee ticket, some $75 tickets available for those that need registration assistance
- $5,000 event sponsorships still available - includes logo event related materials (website, email and signage onsite). These sponsorships are available to companies that are a sponsor of RISC-V Summit North America.

What to expect for Developer Day

- October 22, 2025 at the Santa Clara Convention Center
- Ideal for developers interested in learning about RISC-V or RISC-V Developers looking to expand their knowledge
- Register at: https://events.linuxfoundation.org/riscv-summit/features/risc-v-developer-workshops

Want to learn more about RISE? Stop by our booth or hang out in the lounge!

Get Involved: Gemini Credits + More!

Gemini Credits Several Gemini credit grants for academics to accelerate AI-driven software porting to RISC-V are available. Nominations are open now through October 2. Winners will be announced October 22 at the RISC-V NA Summit.

Nominate a Developer: RISE Developer Awards – Deadline Extended, Closes TODAY!
Developers should have made a notable contribution to the space and or is a generally wonderful collaborator, simply making things easier! Nominate now

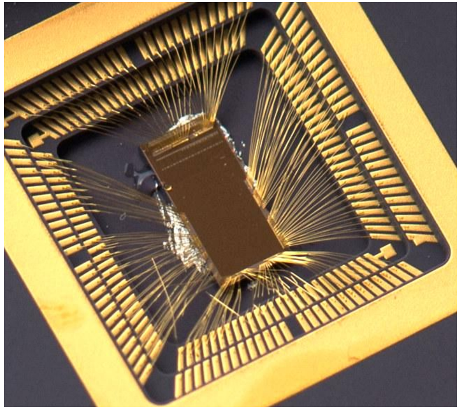
Get involved in a RISE Working Group
Want to learn more or start contributing? Join a RISE Working Group – contribute to more than 10 different projects or focus areas!

RISC-V Developer Appreciation Program, recognizing and rewarding developers who help expand RISC-V adoption through open-source contributions. Whether you're porting a small project or a larger one, contribute to the RISC-V ecosystem.

Global standards are a catalyst to accelerate technical innovation

Standards have been critical to technology innovation, adoption, and growth for decades

Standards create access to opportunities and spur growth for a wide range of stakeholders



RISC-V is a standards-defined Instruction Set Architecture developed by a global community

Expanding the base ISA architecture

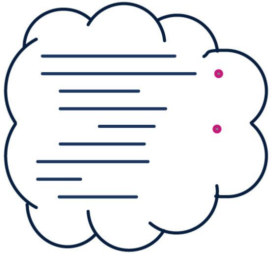
RISC-V enables specific compute blocks to be developed for specific workloads

Start with a basic CPU the add extensions

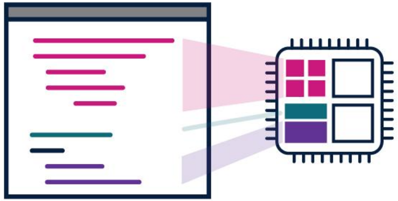
RISC-V customization can add new flexibility and performance to memory access

Enables innovation, exploration, fast development and reusability

AI Specific Design



Extensible ISA enables a software-focused approach to AI hardware

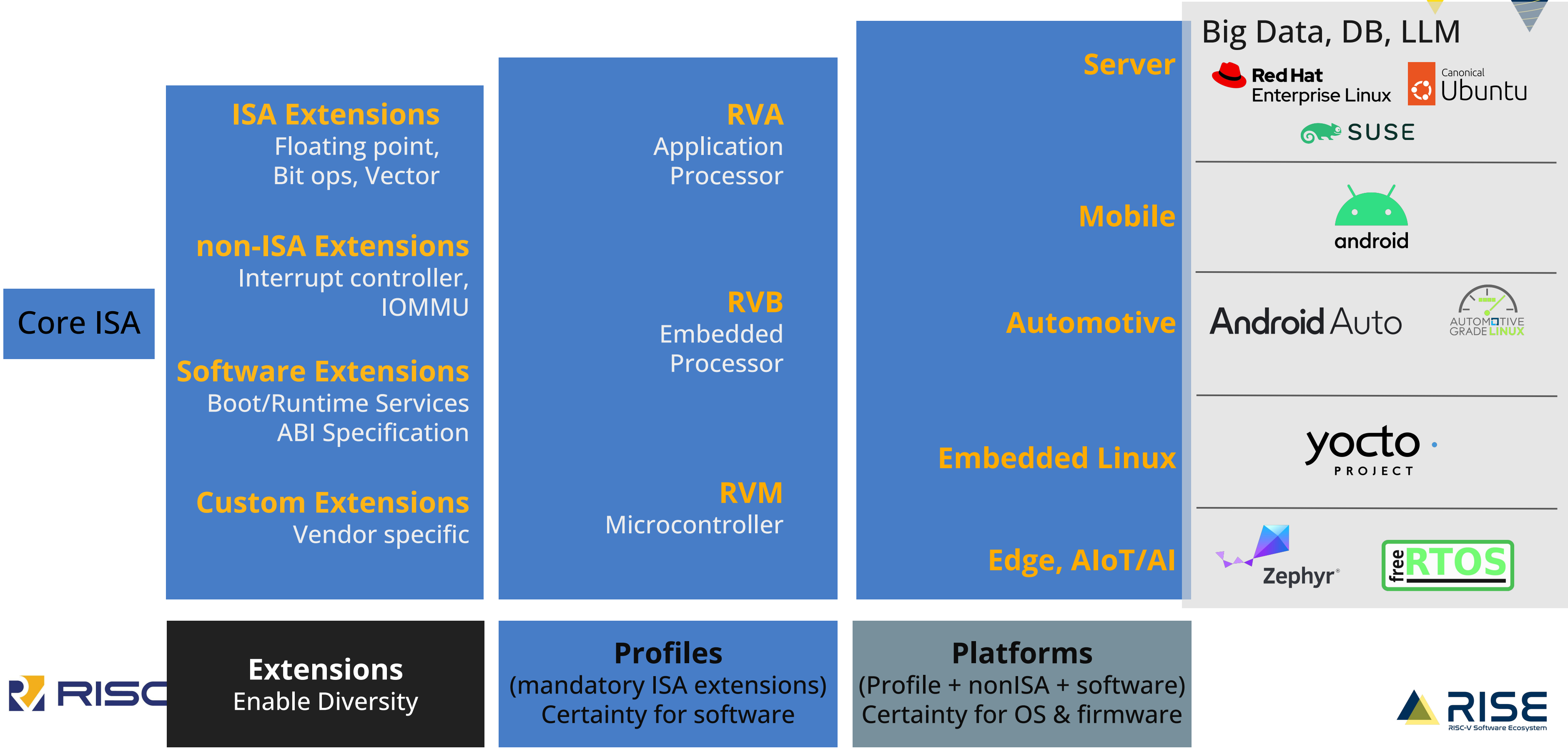


Develop custom instructions and accelerators targeted at your software workload

Unified programming model across AI workloads running on CPU, GPU & NPU

Latest advancements in AI/ML algorithms can be quickly integrated into hardware designs

Profiles, Platforms and software ecosystem



Core ISA

ISA Extensions
Floating point,
Bit ops, Vector

non-ISA Extensions
Interrupt controller,
IOMMU

Software Extensions
Boot/Runtime Services
ABI Specification

Custom Extensions
Vendor specific

RVA
Application
Processor

RVB
Embedded
Processor

RVM
Microcontroller

Server

Mobile

Automotive

Embedded Linux

Edge, AIoT/AI

Big Data, DB, LLM

Red Hat Enterprise Linux

Canonical Ubuntu

SUSE

android

Android Auto

AUTOMOTIVE GRADE LINUX

yocto PROJECT

Zephyr

freeRTOS

Extensions
Enable Diversity

Profiles
(mandatory ISA extensions)
Certainty for software

Platforms
(Profile + nonISA + software)
Certainty for OS & firmware

RISC

RISE
RISC-V Software Ecosystem

Main Specification Ratification in 2024

RVA23 Profile Ratified

Major release of RISC-V Application Processor Profile

Gives application compatibility across vendors

Major added extensions: Vector, Hypervisor

CUDA
blog post

Build Target for Linux OS distro vendors as well as critical middleware and libraries
like Yocto, RHEL, Ubuntu, CUDA

Yocto Project / RISE Progress

Key Milestones

- **May**
  - RISC-V International joined the **Yocto Project** as a **Platinum member**
- **July / August**
  - **BayLibre** partnered with **RISE** to support RISC-V development in the Yocto Project
- **September**
  - A talk on RISC-V International, RISE Project, and BayLibre's participation in Yocto Project was presented at **Yocto Developer Day**

Goal

- Make **RISC-V** a **fully supported architecture** in the **Yocto Project** and provide **ongoing maintenance** within the community

Yocto Project / RISE Progress

Current Status

- Focusing on triage/maintenance
    - Addressing issues in **Bugzilla**
        - AB-INT failures: e.g., glib2.0, libpng, libxml, rust, etc.
        - Other issues involving kernel builds, multilib, etc.
- Build and Test Infrastructure
    - RISC-V results featured in **5.3 M2 release test reports**
    - Foundations set for RISC-V as a **primary architecture**
    - Currently relying on **qemuriscv64** for automated testing

Yocto Project / RISE Progress

Future Plans

- Hardware Support Strategy
  - Evaluating the addition of **RISC-V hardware** to the **autobuilder**
  - RVA23 vs RVA22 and **available hardware**
  - Identifying widely adopted RVA22-compliant boards
- Upstreaming certain **meta-riscv** components to **openembedded-core**
- Expanding **ptest support** and **package coverage**

AI/ML on RISC-V: Why it is critical

- To be a competitive, RISC-V must deliver **high-performance, out-of-the-box** compatibility with essential AI tools
- Significant progress being made on key projects:
  - ***PyTorch***: Foundational work to enable official support and high-performance RVV support.
  - ***Llama.cpp***: Supporting strong community momentum
- Make RISC-V a **first-class citizen** for AI development and deployment
  - Wide area of work: the project themselves, dependencies (direct and indirect), CI/CD, development resources

PyTorch: Path to Official Support

- The ultimate goal is to **build, test, and distribute** PyTorch on riscv64 **directly from upstream**.

- This effort is built on several key milestones:
  - ***Accelerate ATen Operators & Dependencies*****:** Using RISC-V Vector (RVV) extension as the foundation for all performance optimization.
  - ***Upstream CI Integration*****:** Building and testing on RISC-V hardware
  - ***Strategic Partnership*****:** Active participation to the PyTorch community, establishing ourselves as a trusted partner for everything RISC-V.

```
pip3 install torch --index-url https://download.pytorch.org/whl/cpu
```

PyTorch: Accelerate ATen Operators

- **Enable PyTorch to leverage the RVV extension**
  - Extends PR #135570 for supporting the RISC-V Vector (RVV) extension
  - RISE is funding Project RP013 to accelerate optimizing PyTorch
    - ***Partnering with BayLibre*** to deliver the work upstream
    - ***Goal:*** **integrating RVV support**, and targeting **RVA23**
  - This project moves beyond simple enablement to focus on vector-length agnostic (VLA) optimization

- **Improve dependencies OpenBLAS and oneDNN**
  - Faster matrix multiplication operators like `aten::mm` or `aten::addmm`
  - Benefits PyTorch and the whole ecosystem at large

PyTorch: Upstream CI Integration

- PR #143979 introduces an optional RISC-V build to the official PyTorch CI/CD pipeline.

- **Why this matters:**
  - Makes **`riscv64` a visible, recognized architecture**
  - Automatically tests changes to prevent future build breakages
  - A **crucial first step** to becoming an officially supported platform

- **Next steps:**
  - Actively **integrating GitHub RISC-V runners** to streamline development and testing
  - Challenges with availability of faster hardware

Llama.cpp: Continue on Community Momentum 🚀

- Llama.cpp has seen an incredible wave of **organic, community-driven contributions** for RISC-V.
  - Demonstrates a **strong and active developer interest** for RISC-V

- RISE is building on this success by funding RP014
  - **Expand RVV support**, ensuring it is highly optimized for VLEN from 128 to 1024 bits with VLA optimizations
  - **Functional and performance testing on RISC-V hardware**, boosting `riscv64` to a top-tier supported platform in the upstream project

IREE (Future): Compiler and Runtime Optimizations

- A MLIR-based compiler that **optimizes AI models from frameworks like PyTorch and TFLite** to run efficiently on diverse hardware.

- **RISC-V support in IREE is under-invested** compared to ARM and x86. Optimizing this layer is crucial for competitive performance.

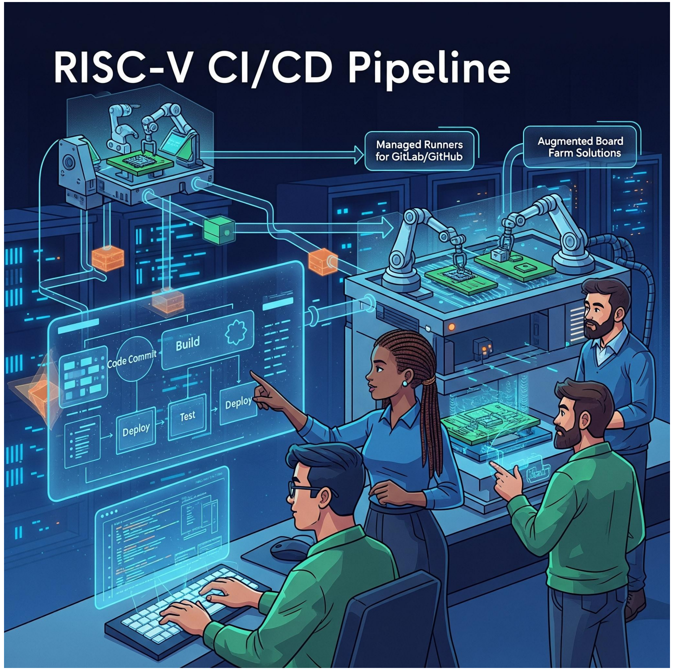
- Proposed Project:
  - ***Goal:*** Optimize the IREE compiler to generate high-performance, RVV-enabled code
  - ***Initial Target:*** Focus on the end-to-end performance of key models like YOLOv7/v8
  - ***Outcome:*** All work is upstreamed, providing the entire ecosystem with a more powerful, MLIR-based toolchain

AI/ML: Summary & Vision for the Future

- **Summary:**
  - A **coordinated, well-resourced effort** is underway to **make RISC-V a first-class citizen for AI/ML**.
  - Tackling the ecosystem at multiple levels:
    - **native framework support**: PyTorch, Llama.cpp
    - **compiler toolchain**: IREE
    - **dependencies**: python packages, native libraries
  - Focused on **robust, upstream-first** RVV and RVA23 support

- **Vision:**
  - A future where developers have a **choice of high-performance tools** for deploying AI models on RISC-V, enabling a competitive and **thriving software and hardware ecosystem**

CI Enablement Program Overview


RISC-V CI/CD Pipeline
Managed Runners for GitLab/GitHub
Augmented Board Farm Solutions
Code Commit
Build
Test
Deploy
Deploy

- Facilitating developer access to RISC-V CI/CD
- **Providing managed runners for Gitlab/Github**
- Streamlining RISC-V build and test processes
- Augmenting current board farm solutions

Q&A