SiFive

Enabling energy efficient Datacenters

Rocky Zhang
Principal Engineer, Field Application, SiFive

RISC-V Summit
Hangzhou, China
August 21, 2024

Summary

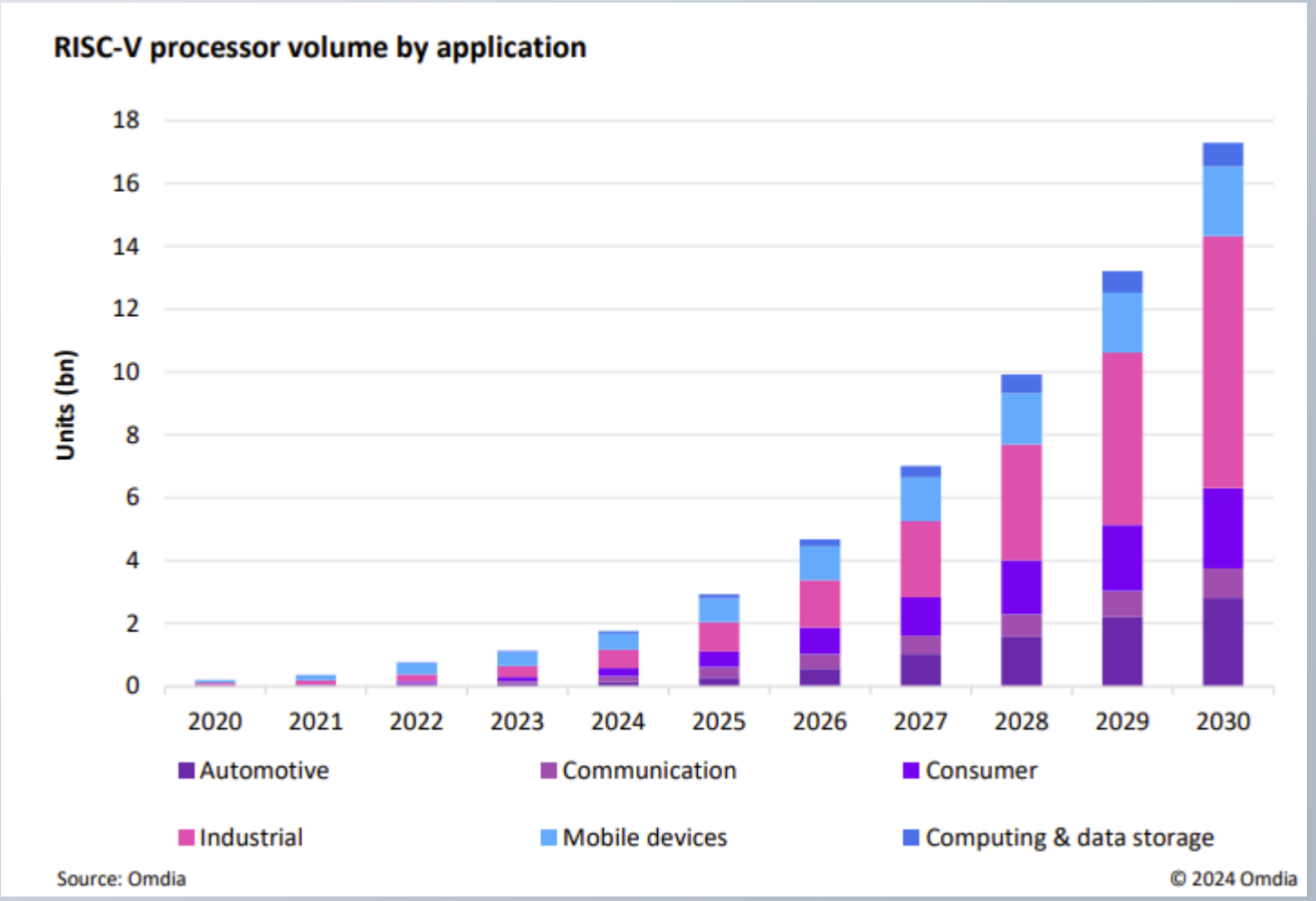
- Datacenter workloads are diversifying and an opportunity exists to create optimized-for-purpose solutions
  - Featuring a combination of general-purpose and application-specific processing elements

- The P870-D processor represents SiFive's first processor that addresses mainstream datacenter and infrastructure use case
  - Complementary to SiFive's Intelligence Processor family
  - Harnessing open-source software to provide foundational platform that addresses emerging class of AI workloads' need for cost- and power-efficiency

- Delivers size/power advantages over incumbent approaches
  - Most appropriate for workloads that benefit from high levels of parallelism

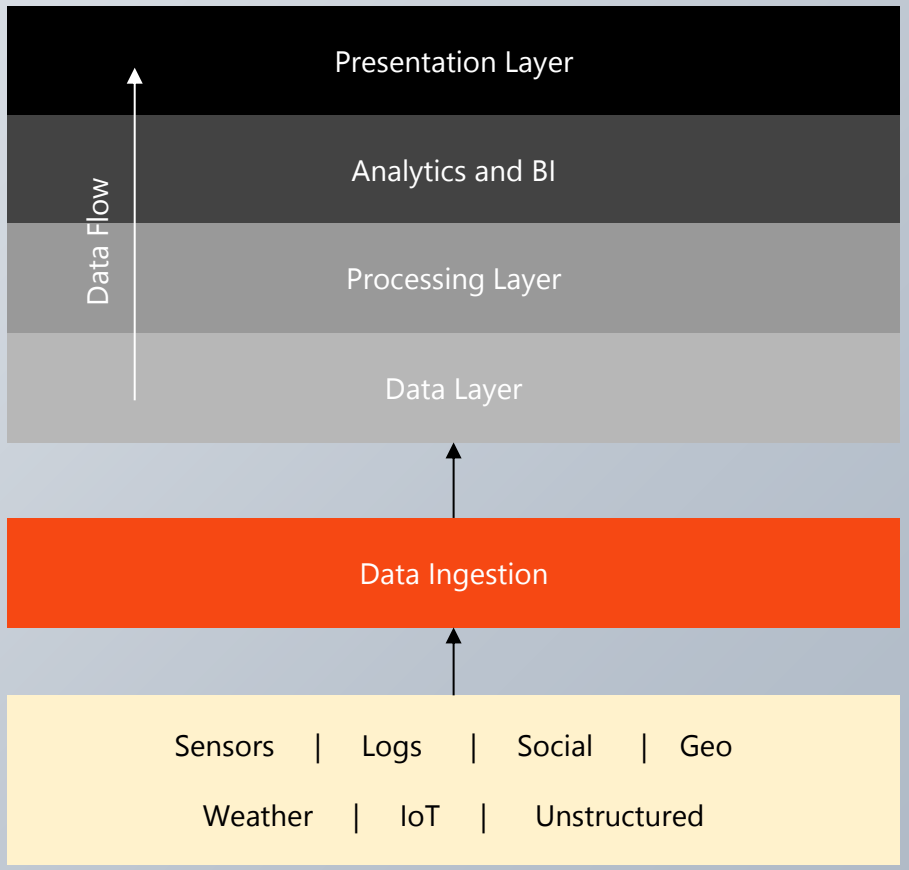
Primary Factors Driving RISC-V Into Higher Value Sockets

- Increased use of open-source APIs and foundational software across use cases
- Sustainability, cost and/or performance concerns driving need for innovation, with no one size fits all
- A desire from industry for increased resilience to supply chain disruption



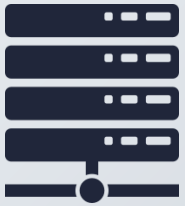
Primary Factors Driving RISC-V into Higher Value Sockets

- Web services
- Media streaming
- Storage
- Data analytics
- AI (where main training functionality is offloaded to GPGPU and/or hardware accelerator)
- High Performance Computing (HPC)



Key Infrastructure Requirements

Datacenter



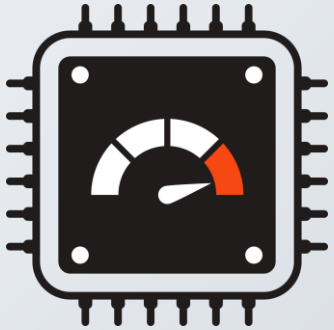
Cloud

Networking

Edge

- **TCO -** High performance with great compute density / power efficiency
- **Scalability -** Creating high core count systems, across multiple die / chiplets
- **AI Workloads -** Effective coupling of optimized-for-purpose accelerators
- **Infrastructure-specific features -** RAS, Security etc
- **Time to Market -** Simple path to use foundational open-source software
- **System approach -** Complete offering of IP beyond CPU core

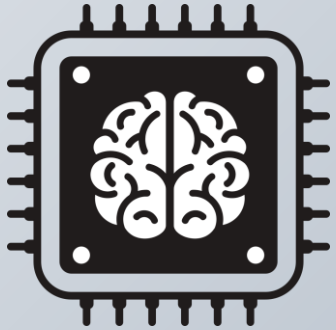
SiFive is empowering the new computing era



Large-scale high-performance general-purpose CPU

SiFive Performance Family

+



High-performance NPU

SiFive Intelligence Family
Vector Engine

Hardware Matrix Engine

Inclusion of highly capable accelerators reduces the single thread performance needs for AI use cases

3rd Generation of Performance Processors

SiFive

Datacenter

P870-D
- 6-wide OoO core
- RVA23 + Vectors
- Vector Crypto
- CHI Bridge
- Cross-cluster RAS
- Up to 256 cores

General Purpose

P870
- 6-wide OoO core
- RVA23 + Vectors
- Vector Crypto
- Up to 32 coherent cores

P670
- 4-wide OoO core
- RVA22 + Vectors
- Vector Crypto
- Up to 16 Cores

P470
- 3-wide OoO core
- Same feature set as P670 in a more efficient package

P550
- 3-wide OoO core
- RVA20

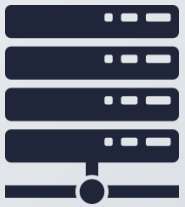
2021 | 2023 | 2024 | 2025

SoC and Early System Availability

Roadmap Subject to Change

P870-D; Grounds Up Design for Infrastructure Use Cases

Datacenter



Cloud

Networking

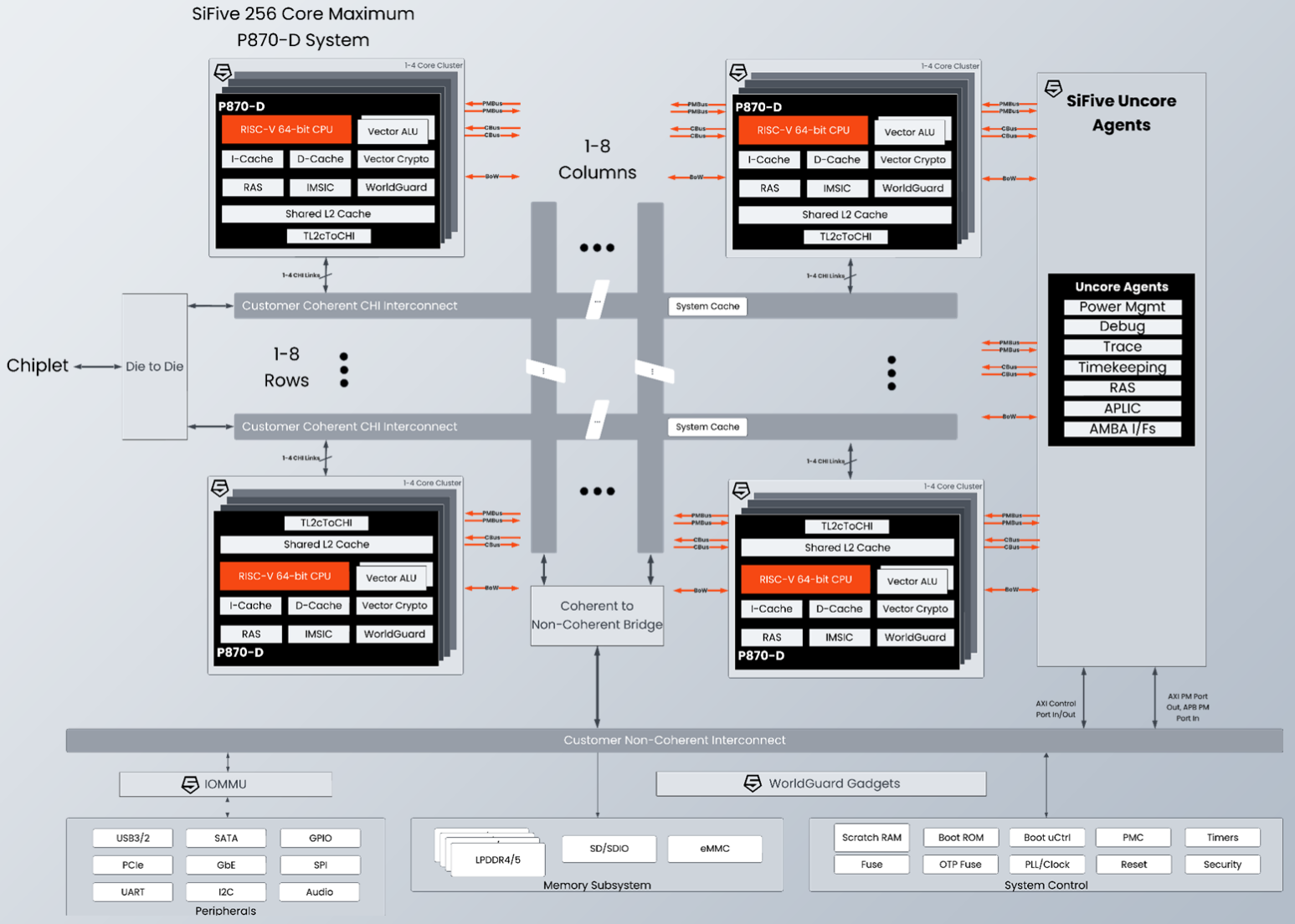
Edge

Delivering targeted benefits across a variety of usage cases

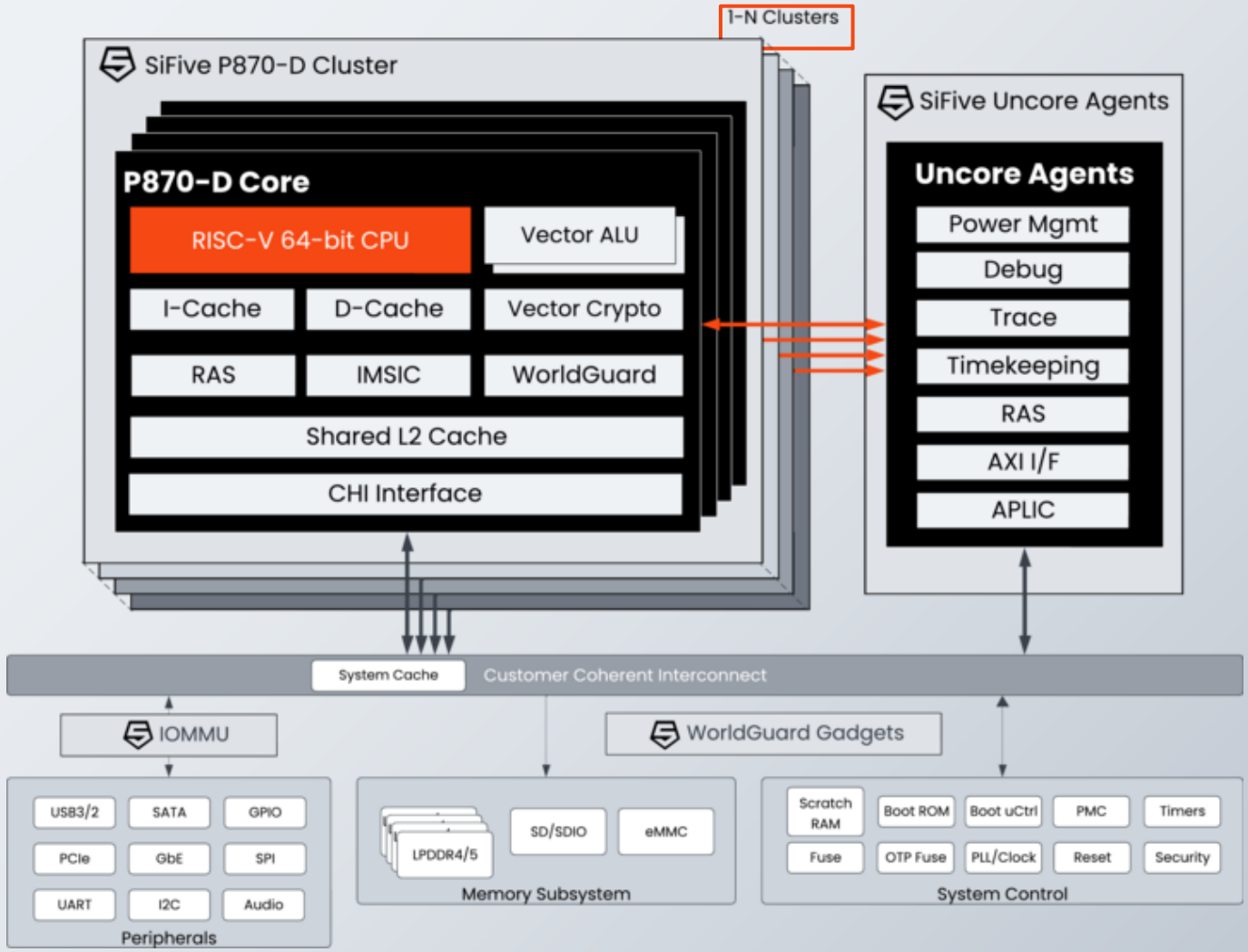
Low TCO	Architectured for efficiency with **RISC-V ISA** to significantly increase Compute density and low power consumption and deliver efficient performance
Scalable	Delivers compute on demand with solutions scalable to 256 cores A distributed **IOMMU** architecture that scales with devices Sv57 with a 5-level page table to support a larger virtual memory as demands increase 4 CHI ports/cluster supports wider link for multiple peripheral and memory device ports
8x Vector ALU/cluster of 4 cores + Portfolio of low power AI accelerators	Processors with vector instructions that can be augmented with discrete AI accelerators from **SiFive Intelligence** series for custom workloads
Leverage of existing interfaces and standards for heterogeneous compute	Support for AMBA **CHI, AXI4** coherent interfaces for scalability across heterogenous fabric
Designed for mission critical systems	**Cross-cluster RAS** protections for data integrity, maximising uptime and delivering fault tolerant systems
Secure and flexible with multiple environments	SiFive **WorldGuard** to ensure untrusted application code cannot access trusted resources.

P870-D Enables Scalable Subsystems

Core clusters, distributed IOMMU, cross-cluster RAS and WorldGuard Security



P870-D Solution Summary



SiFive Deliverable

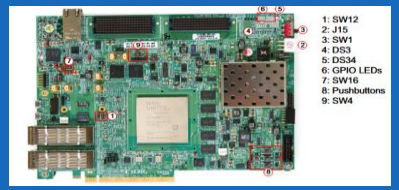
	Key Features
ISA	RVA23 compliant (RISC-V Vectors and Vector Crypto)
Decode width	Highly efficient 6-wide OoO 64 bit core, delivering 2.0 SpecINT 2k17/GHz
Cache	64KB L1$ and up to 8MB shared cluster level L2$ 3rd party SLC
Multicluster	4 cores/cluster, upto 64 clusters
Vector	2 x Vector ALU (128b VLEN/128b DLEN)
Virtualization	Sv39/Sv48/Sv57, MMU and Hypervisor support
Interrupt controller	AIA interrupt controller with APLIC.m/IMSIC or APLIC.w
RAS	Protections on memories, processor architectural states (register files, CSR) and key structures of the datapath including interconnect and shared cache controller RAS architecture to configure and reports errors compatible with RISC-V standard.
Power Management	Cluster-level DVFS and idle power modes via SQIP (wired/P-channel or MMIO)
Interconnect	Single logical CHI port/Cluster split into up to 4 physical ports CHI link width: 128b, 256b, 512b
System	System IP for system level solution:. P870-D Uncore agents with Trace and Debug, Next Gen IOMMU, WorldGuard

Extending SiFive/Arteris Partnership

P870-D to lead RISC-V based emulation systems with Arteris Ncore compliance

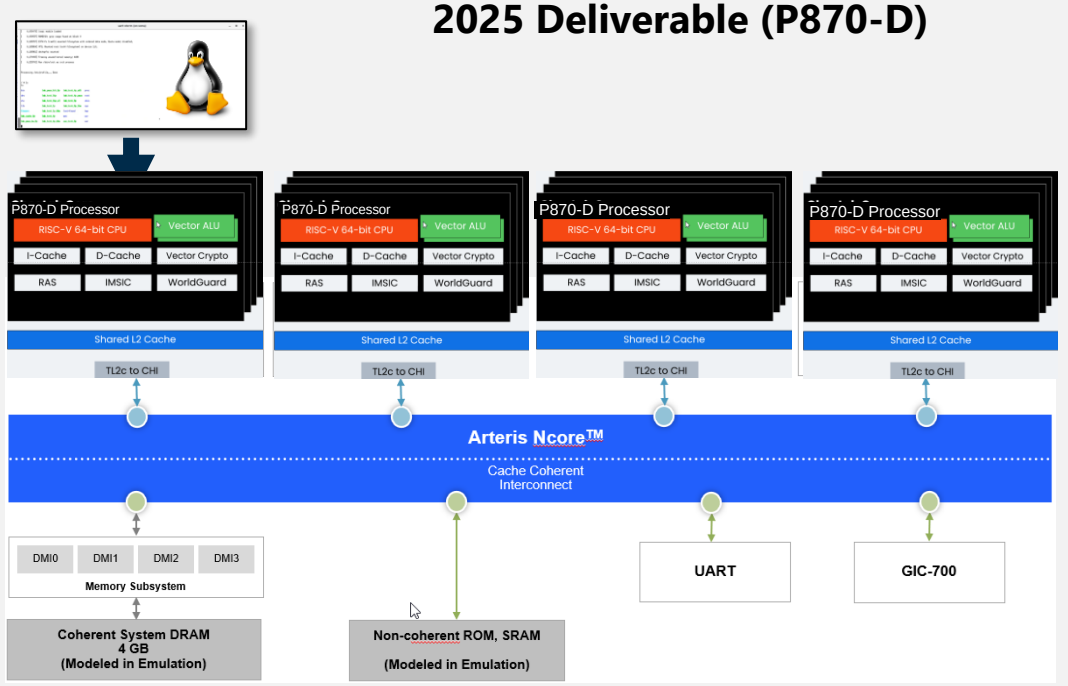
2022 Deliverable (X280)

Evaluation Kit



https://www.sifive.com/blog/sifives-risc-v-leadership-strengthens-with-new-vector

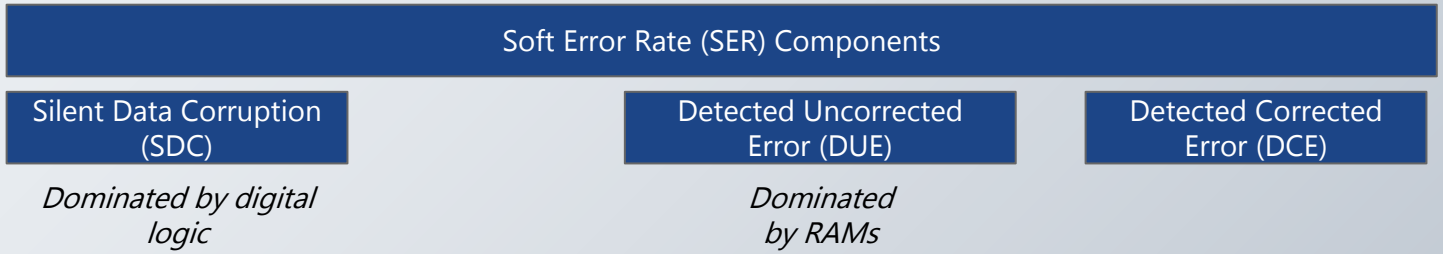
2025 Deliverable (P870-D)



Hardware-Assisted Verification and Validation System

Pre-validated SiFive cores with Ncore

P870-D Hardware Protections



P870-D RAS Framework

Protect visible states and main datapath with minimum PPA impact

Robust soft-error protection on digital logic	Robust soft-error protection on memories	SiFive RAS Architecture
Implemented on arch states, datapath to L2 cache through interconnect	Implemented on RAM Macros	Tree structure of Functions for RAS agent to identify event and access resource

SiFive RAS Benefits:

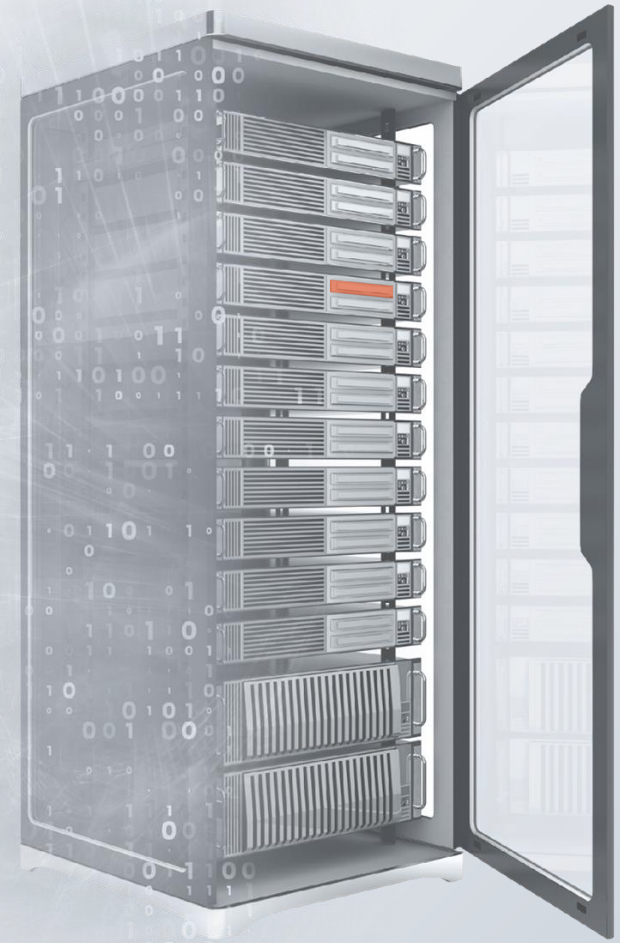
- Protections on memories, processor architectural states (register files, CSR) and key structures of the datapath including interconnect and shared cache controller
- RAS architecture to configure and reports errors compatible with RISC V standard.

Software Ecosystem Readiness

- RISC-V ecosystem delivers similar foundational technology equivalents of Arm ecosystem. Examples
  - Secure firmware runs in RISC-V M-mode, similar to ARM EL3
  - Secure OSs run in S-mode (~ ARM EL1), and secure user mode apps run in U-mode (~ ARM EL0)
  - Several options for software that executes after the boot ROM
    - U-Boot secondary program loader (SPL)
    - Coreboot
    - EFI Development Kit (EDK II)
- WorldGuard can be used to partition secure OSes from non-secure OSes

- Our target datacenter segments harness combination of open source software and first party software
  - Foundational open source software elements already exist for RISC-V
  - Standard platforms are key
  - P870-D fully aligned with RISC-V International work, enabling customers to reuse this work

RISC-V Server Platform Task Group (TG) Charter



The Server Platform Task Group is defining a specification for a **standardized set of hardware and software capabilities**, that portable system software (such as operating systems and hypervisors) can rely on being present in a RISC-V server platform.

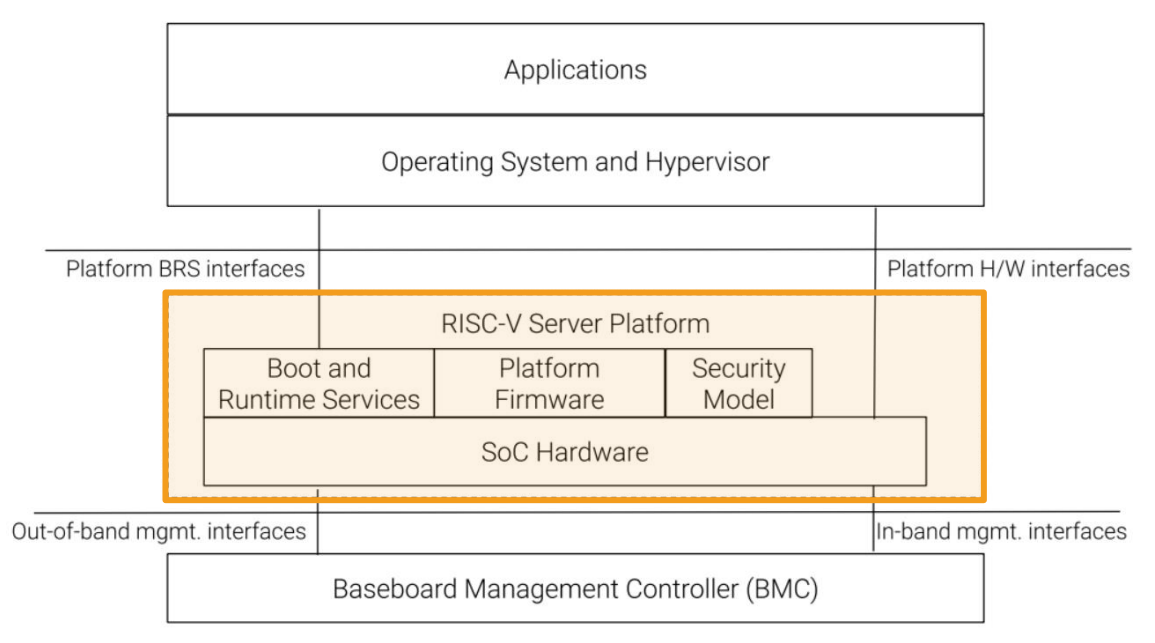
The produced specification will tie together the many hardware and software requirements for a server platform, many of which are defined in detail by other RISC-V specs, such as:

- ISA Profiles
- Boot and Runtime Services
- Server SoC
- Platform security Model

Server Platform Specification brings works together form other technical groups is the document that:

- Binds everything together
- Provides a practical architectural guide to achieve the stated goals

Server Platform specification



Enables a single binary OS image distribution model

Ties together hardware (SoC), firmware (BRS) and security specifications:

- Standardized set of SW capabilities.
  - OS loader interfaces, hardware description mechanisms, etc
- Standardized set of HW capabilities
  - harts, timers, interrupt controllers, PCIe root complexes, RAS, QoS, in-band mgmt., etc.)
- Common security model

P870-D Customer Benefits

- Low risk - Builds on proven CPU, with focus on adding system attributes needed for data center use cases
- Fastest time to market - Harnessing growing software ecosystem around RISC-V
- Ideal for workloads that benefit from parallelism

Empowering innovators

www.sifive.com