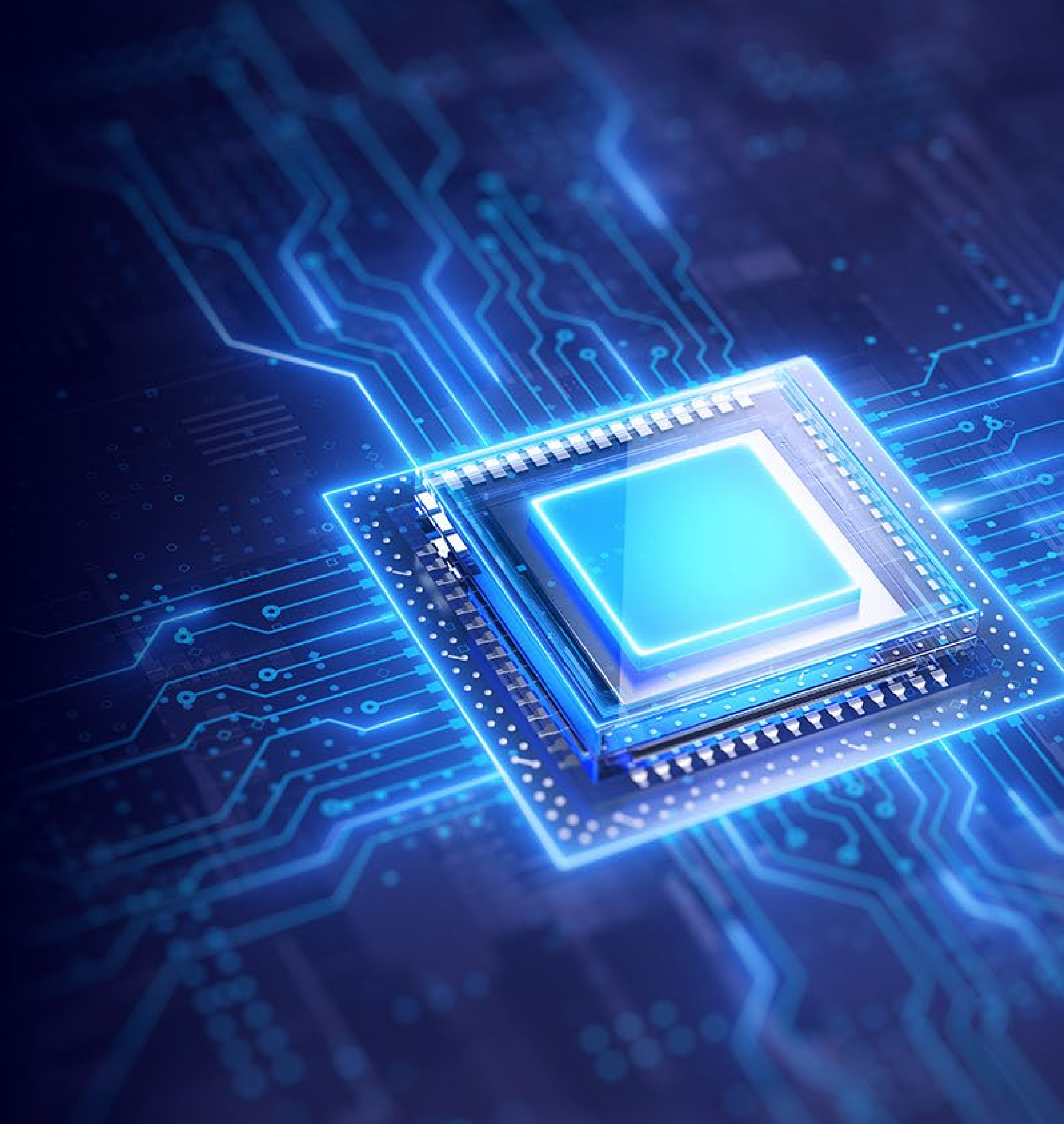
VENTANA

RISC-V, the next generation architecture for AI

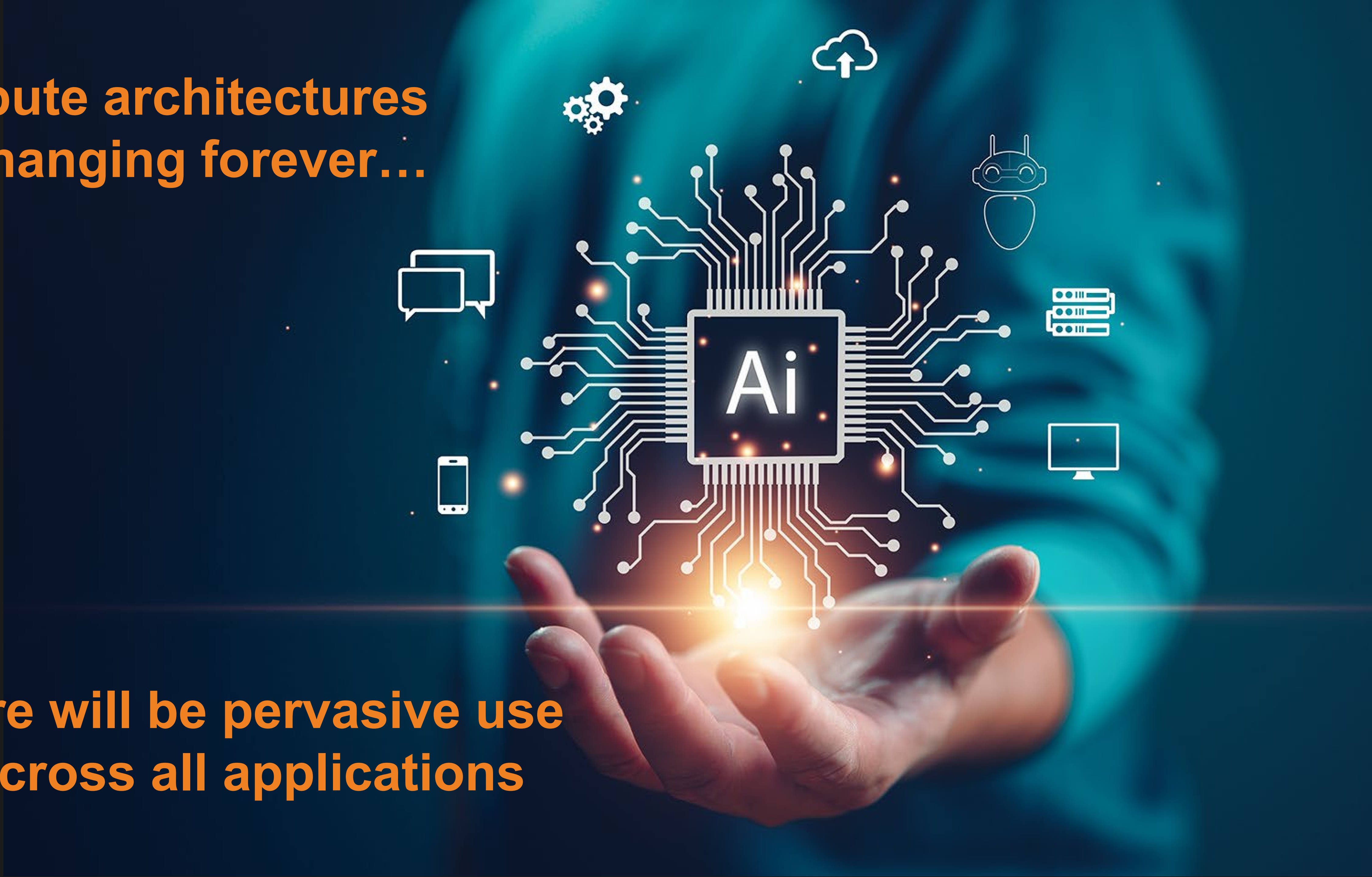
October 2024
GSA Executive Conference



Ventana RISC-V AI/HPC Engagements

Common Themes Have Emerged In Discussions With Hyperscalers, Sovereigns, OEMS, And End Customers

Compute architectures are changing forever…



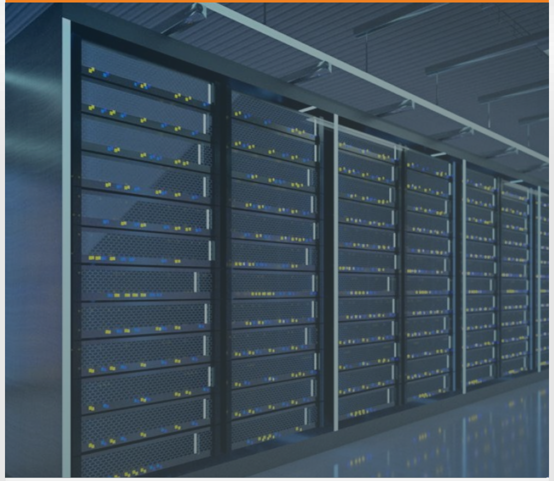
… there will be pervasive use of AI across all applications

AI Will Be Pervasive Across All Applications and Markets

Generative AI

- ✓ **Software Development:** Developer productivity boosted significantly through automatic code writing, refactoring, and documentation,
- ✓ **Content Creation:** Automation of content creation enhancing marketing and sales productivity
- ✓ **Customer Service:** Virtual assistants to automate tasks and personalize responses

Data Center



- ✓ **Technology Investment:** AI accelerators and specialized silicon will be necessary to support AI-driven operations and enhance efficiency
- ✓ **Power and Coiling:** AI workload demands require advanced power systems and liquid cooling
- ✓ **Sustainability:** Energy costs, efficiency and availability drive architecture and locationre of data centers

Automotive

- ✓ **ADAS and Driver Assist Systems:** AI powering the next jump autonomous driving systems
- ✓ **Predictive Maintenance:** anticipates vehicle issues before they occur, reducing downtime and maintenance costs, while optimizing fuel efficiency and performance through real-time data

Intelligent Edge

- ✓ **Laptops:** AI processing capability now a requirement for next generation PCs
- ✓ **Mobile:** AI longstanding capability for features such as photo enhancement
- ✓ **Industrial Manufacturing:** Efficient, scalable, low latency AI required for next generation robotics manufacturing capabilities

$22.5T Global Economic Impact Estimated For Generative AI

ILLUSTRATIVE MARKET SIZING (A TOP-DOWN BUILD)

		Economic Impact[1]	Representative Vendors
1	Content Workflow Management Apps	$1,951B	DocuSign, servicenow, walkme, UiPath
2	Vertical Software	$1,413B	BigCommerce, Matterport, coursera, Sabre
3	Collaboration Apps	$1,193B	asana, slack, Microsoft Teams, zoom
4	Enterprise Resource Management Applications	$464B	BLACKLINE, ORACLE, docebo, workday
5	Customer Relationship Management Apps	$269B	Five9, HubSpot, ORACLE, salesforce
6	Application Development Software	$265B	Couchbase, ORACLE, GitHub, PagerDuty
7	Production Apps	$60B	Adobe, WiX, GoDaddy, ZETA
8	Supply Chain Management Apps	$14B	ORACLE, SAP

US Total	$5.6T
Global Adj. (÷)	25%
Global Total	$22.5T

Source: Equity Research.

[1] Represents the cumulative wages that can be impacted by GenAI. Calculated by estimating % of employment impacted in 3 years (the estimated number of years for household adoption) by the average wage by occupation.

Pain Points With Current AI Acceleration Solutions

Current solutions take a “one-size-fits-all” approach to hardware development

- **Overoptimization For Yesterday’s Architectures**
  - AI models evolve: AlexNet→ResNet→Transformer→LLMs→?
  - Many ResNet accelerators underperform on Transformer workloads
- **Inability to Right-Size for Workloads**
  - Flexibility to adjust Compute, Memory, and IO to specific applications
  - Tight CPU-AI integration required
- **Need to Keep Up with rate of AI Innovation**
  - Current solutions lock users into specific vendors
  - Customization for changing AI architectures is critical
  - Open Hardware and Software Stacks required

RISC-V Solutions are Key to Overcoming AI Chip Challenges

Current AI Hardware

- Lack of instruction set standardization
- Lack of software support
- Inability to adapt to changing workloads
- Lack of scalability

RISC-V Solutions

- **Standard software baseline:** RISC-V Base ISA, RISC-V Vector Extensions, RISC-V Matrix/Tensor extensions (Coming Soon)
- **Open Source/Open Standards momentum**
- **Easy to modify for custom needs**
- **Match silicon to workload:** Flexibility to build processors with the right ratio of Compute, Memory and IO

Varied ISAs, Fragmented SDKs, and Limited Software Are One of the Biggest Limitations of AI Hardware Usefulness

VENTANA has the Highest Performance RISC-V CPU in the World

VENTANA'S NEW VEYRON V2 IS THE HIGHEST PERFORMANCE RISC-V PROCESSOR AVAILABLE TODAY AND IS OFFERED IN THE FORM OF CHIPLETS AND IP

- ✓ **Veyron V2** showcases **up to 40% improvement** in performance
- ✓ Improved RISE ecosystem support enables V2 to **quickly deploy open, scalable, and versatile solutions**
- ✓ Chiplet-based solutions improve unit economics, accelerating **time to market by up to two years** and **reducing development costs by up to 75%**
- ✓ Domain specific accelerator designed to **enhance workload efficiency**

3.6GHz

4nm
Process technology

32 Cores
Per cluster

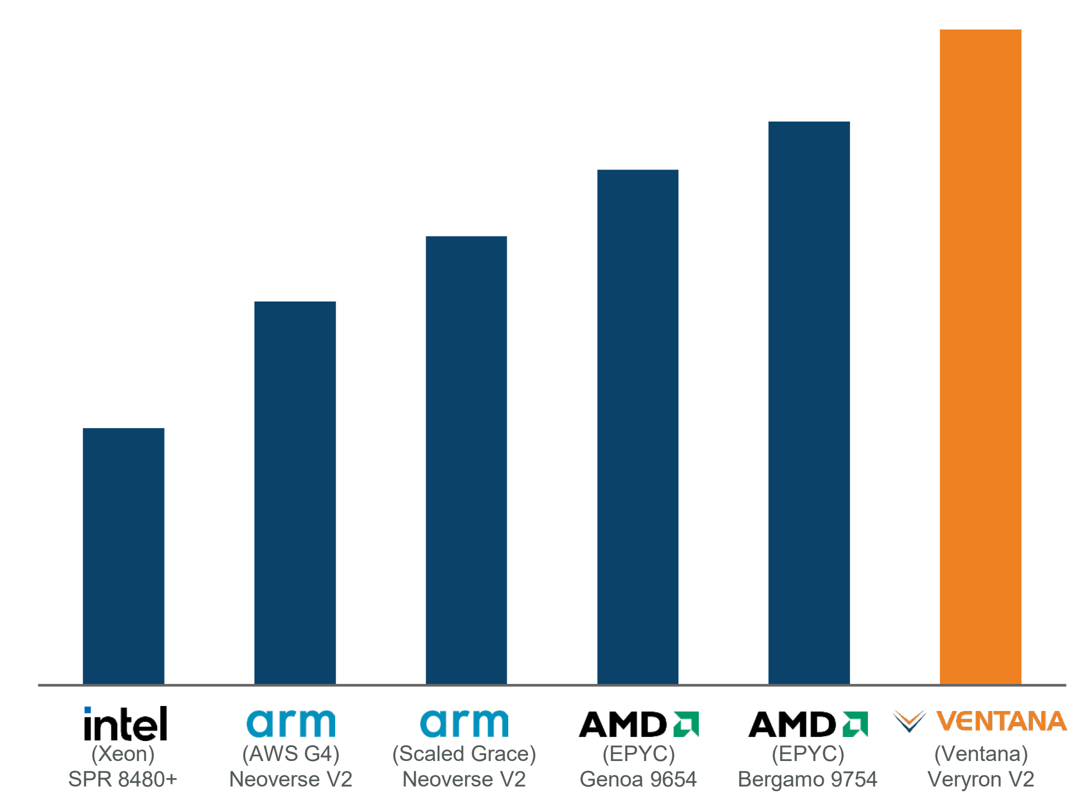
Up to
192 Cores
Multi-cluster scalability

128MB
Shares L3 cache per cluster

512b
Vector unit

Highest Performance Server-Class RISC-V Processor

(SPEC CPU 2017 Rate per Socket – Scaled to 350W)

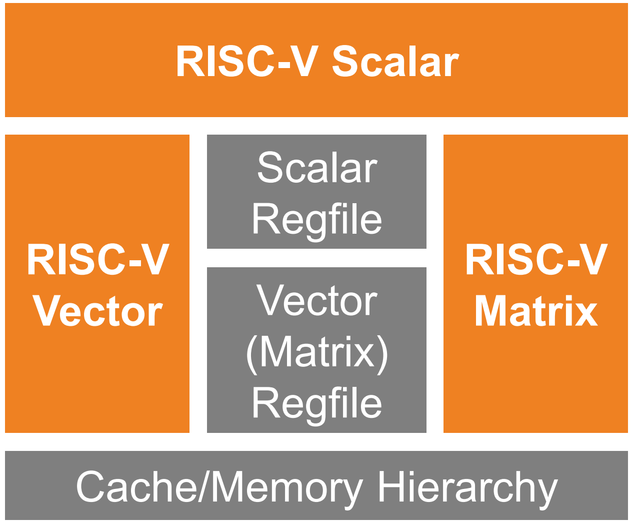


VENTANA

The Solution: RUCA RISC-V Unified Compute Architecture

- **RUCA Cores Use Standard RISC-V Scalar-Vector-Matrix Operations**
  - Open Standard Software Target
  - Unified Instruction Set
  - Tightly Coupled Shared Register Files and Cache Hierarchy
- **RUCA Cores Have Ability to Add Custom Extensions to Enhance the RISC-V Base ISA**
- **RUCA Architecture Advantage**
  - RUCA Cores contain the entire Host(Scalar) – Accelerator profile: Avoids Expensive Data Transfers and Compute Hand-Offs
  - Typical GPU/AI Offload Accelerators Require Extensive Shuffling of Data Across Board/Fabric

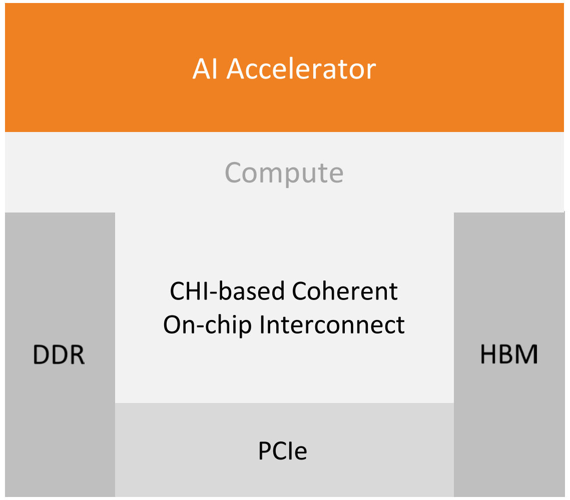
Example RUCA Core



Incubate AI Innovation with RISC-V Custom Extensions, Drive to RVI Ratified as Value Proven

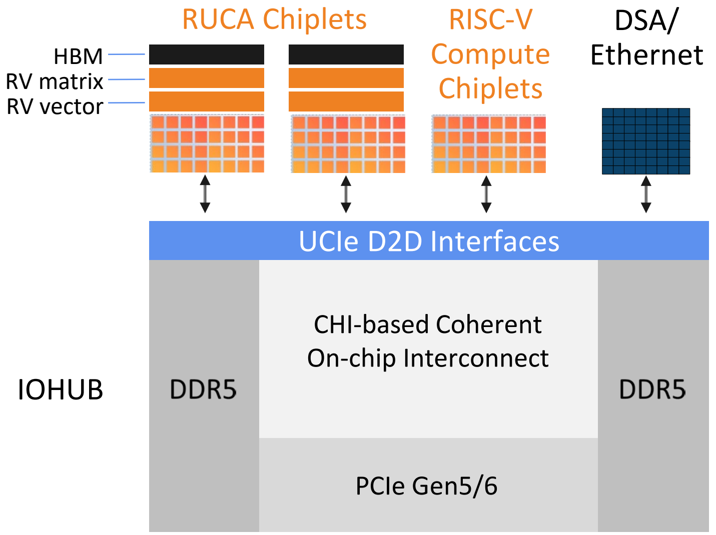
Chiplets Deliver Efficient and Scalable AI

CURRENT APPROACH: Monolithic SoC



- ⚠ **Longer time to production, TTM: 3+ years**
- ⚠ **Higher development cost: $200M-300M+**
- ⚠ Fixed solution, less flexibility
- ⚠ Proprietary accelerator ISA
- ⚠ Proprietary software libraries

FUTURE: RISC-V + Chiplets



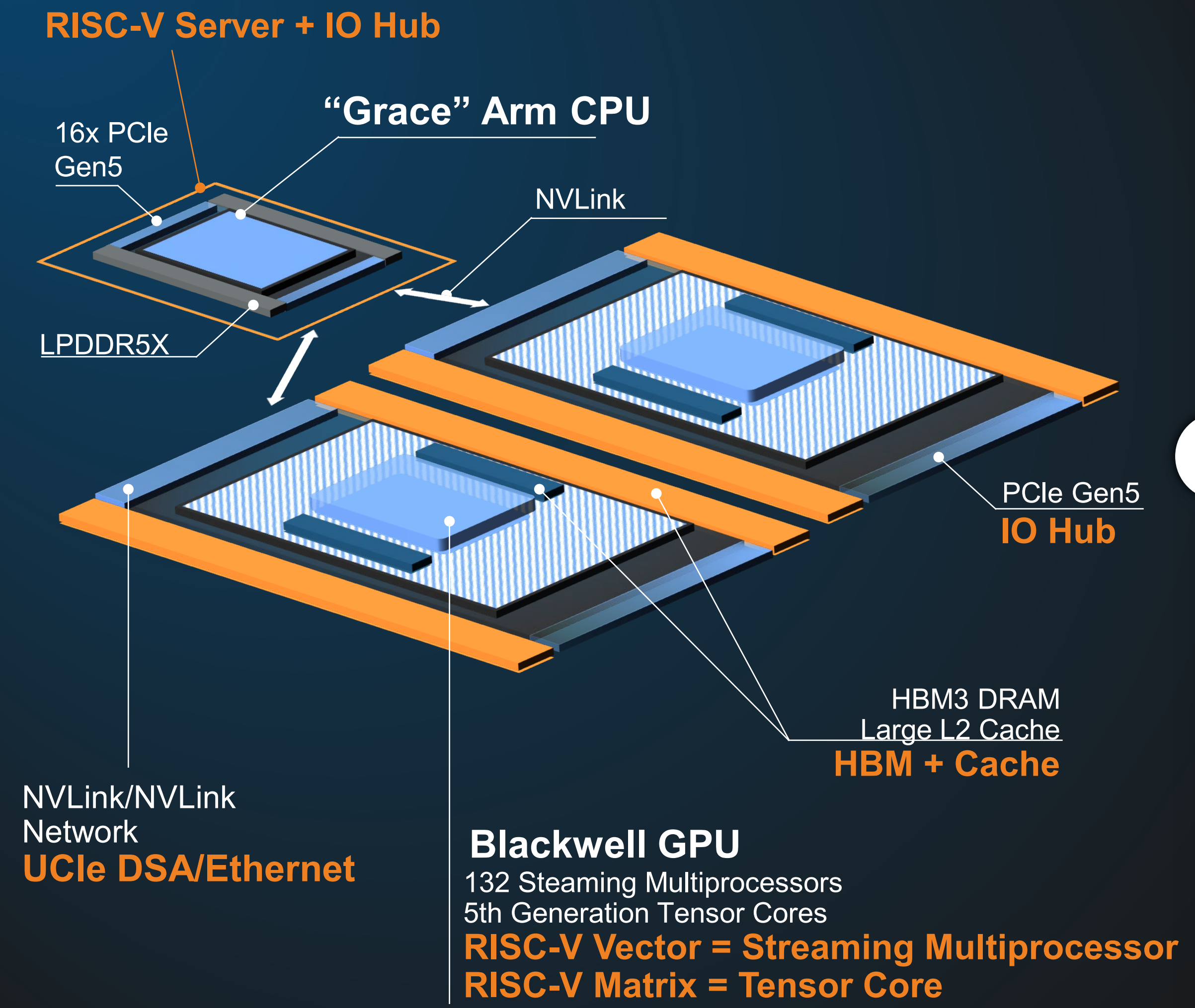
Chose number and type of RISC-V AI chiplets to match compute needs

Scale-out addressed with DSA ethernet chiplets

- ✓ **Faster production time, TTM: < 1 year**
- ✓ **Lower development cost: < $25M**
- ✓ Scalable and configurable compute, vector / matrix and memory
- ✓ Open standard RISC-V ISA
- ✓ Leverages open-source software libraries

Nvidia Grace Blackwell Deconstructed Into RUCA

RISC-V Equivalents to Nvidia Grace Blackwell

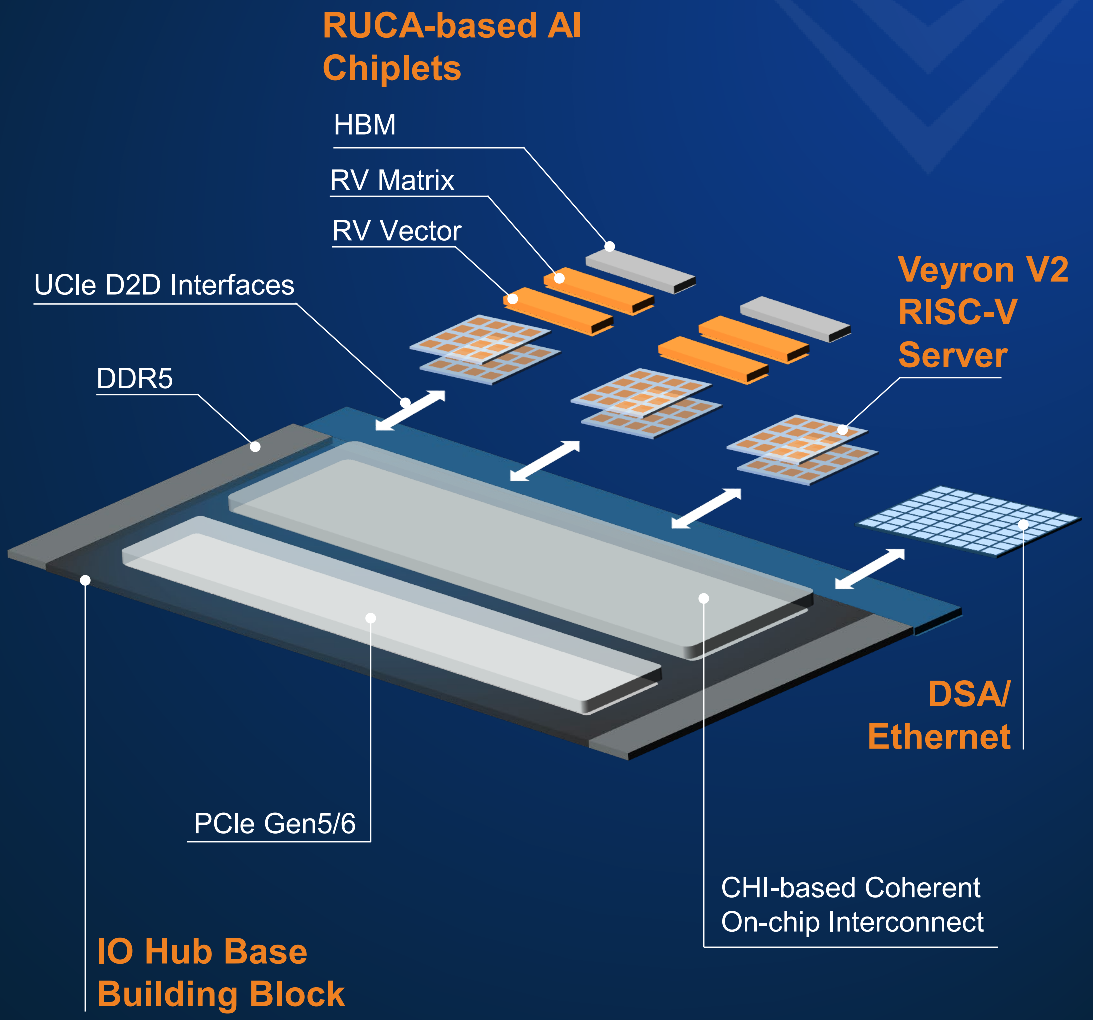


Blackwell GPU

132 Steaming Multiprocessors
5th Generation Tensor Cores

RISC-V Vector = Streaming Multiprocessor
RISC-V Matrix = Tensor Core

AI Compute Using RISC-V



RISC-V AI/ML Software Ecosystem

Work has already begun to enable the RISC-V software ecosystem

Applications	Computer Vision	Speech	Natural Language Processing	Autonomous Systems	Recommendations	Finance
Models	ResNet50 VGGNet YOLO	HMM LSTM	GPT BERT	SLAM ControlNet	Content Filter Gradient Boosted	ARIMA Monte Carlo
APIs, Libraries	OpenAI Whisper KALDI	Keras	PyTorch Lightning	TorchVision	NumPy	Transformers
Deploy, Serve	TFServe	KServe	kubernetes Kubeflow	TorchServe	docker	KVM
Frameworks	TensorFlow Lite	TensorFlow	PyTorch		ONNX	
Runtimes	TFRT			GLOW	ONNX RUNTIME	
Platform Interface Libraries, Extensions, SDK	NVIDIA CUDA		OpenAI Triton		Ventana ML SDK	
Operating System	Linux		ubuntu		debian	
Firmware Early Boot, BIOS	OpenSBI		tianocore uefi		ACPI	
Platform Ventana Veyron AI/ML Server						

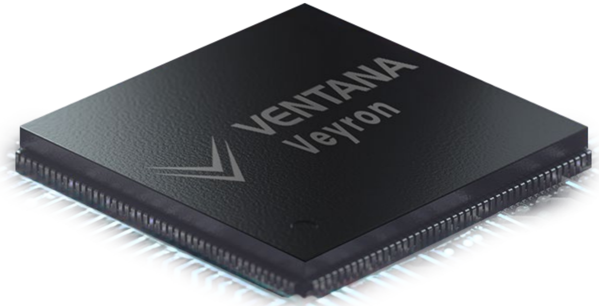
Tools & SDK

- python
- C++
- GCC
- LLVM COMPILER INFRASTRUCTURE
- MLIR
- glibc
- GDB The GNU Project Debugger
- TRACE32
- OpenOCD

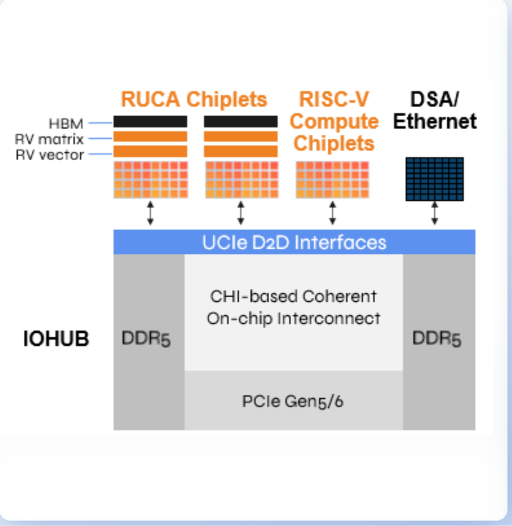
The Complete RISC-V Veyron Platform

Tensor accelerators and custom instructions are not enough

- HIGH PERFORMANCE CPU ✓
- SOFTWARE STACK ✓
- SYSTEM IP ✓
- GENERATIVE AI ACCELERATION ✓
- DOMAIN SPECIFIC ACCELERATION ✓
- CHIPLET TECHNOLOGY ✓



RISE
RISC-V Software Ecosystem



HBM
RV matrix
RV vector
RUCA Chiplets
RISC-V Compute Chiplets
DSA/ Ethernet
UCIe D2D Interfaces
IOHUB
DDR5
CHI-based Coherent On-chip Interconnect
DDR5
PCIe Gen5/6

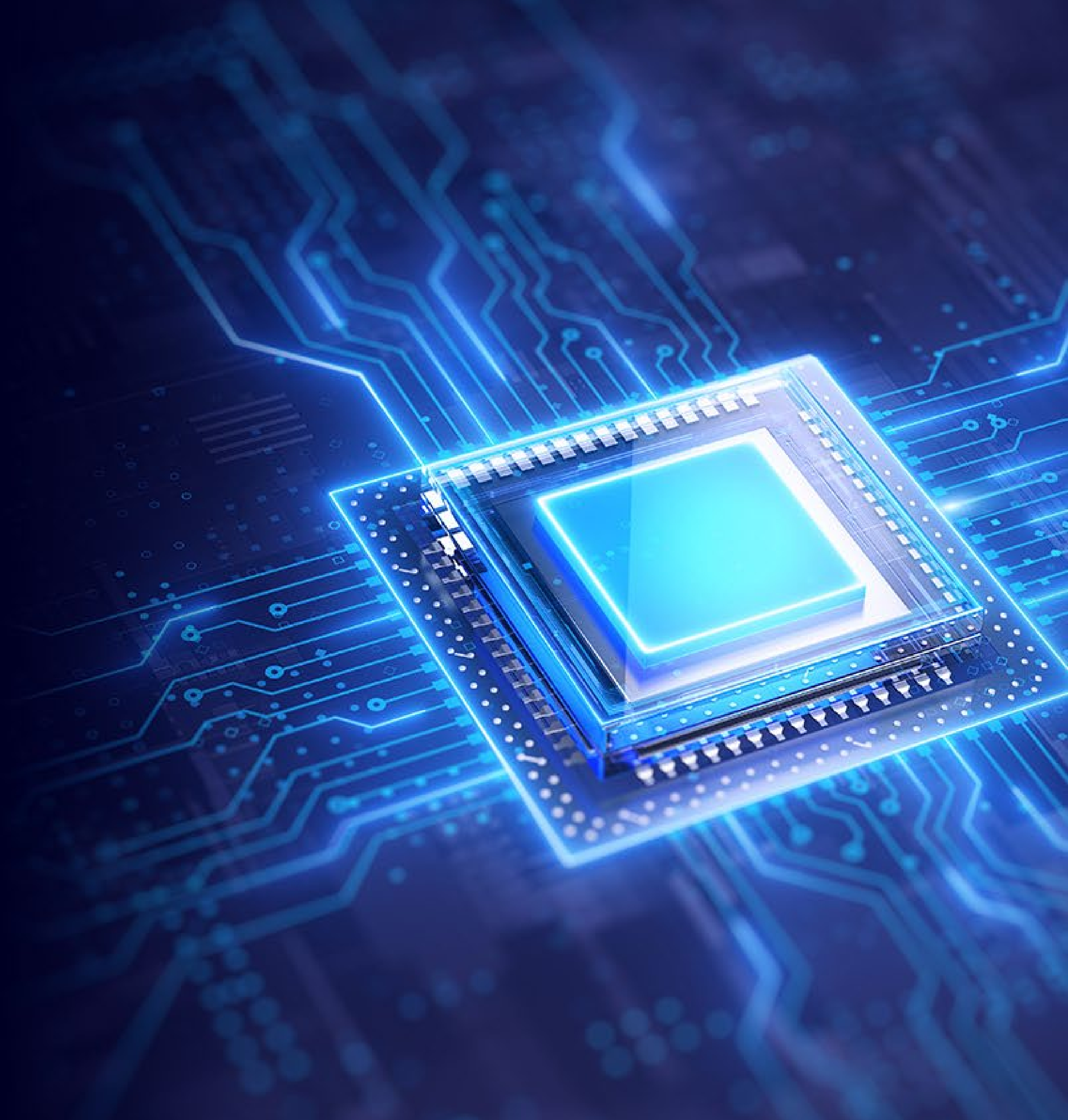
- ✓ AUTOMOTIVE SAFETY CERTIFICATIONS
- ✓ SECURITY
- ✓ RAS
- ✓ LOW LATENCY COHERENT NOC
- ✓ HIGH PERFORMANCE CACHE ARCHITECTURE
- ✓ FEATURE PARITY WITH LEADING CPU ARCHITECTURES

In Closing…

- ✓ AI will be pervasive across all tiers of computing
- ✓ A common architecture and software base is required for mass deployment
- ✓ The RISC-V Unified Compute Architecture enables the efficient development of AI software and libraries
- ✓ Achieving this vision requires a complete platform
- ✓ Ventana, through its platform, will partner to make RISC-V a driving force in the advancement of AI

VENTANA

Thank You !



Ventana Founding Team Overview

Balaji Baktha

FOUNDER AND CEO

Pioneer in Data Center semiconductors:

- **33+ years experience**
- World’s first 64-bit ARM with Veloce (Acquired by AppliedMicro)
- Led Marvell BU delivering Data Center class Networking, Communications, Compute, Storage and Wireless infrastructure products
- World’s first iSCSI with Platys, (Acquired by PMC-Sierra (Adaptec))

Greg Favor

CO-FOUNDER AND CHIEF ARCHITECT

One of the world’s leading CPU architects:

- **35+ years experience**
- Architected K6 processor at startup NexGen, acquired by AMD
- Chief Architect at Siara Systems, acquired by RedBack
- Architected first successful 64-bit ARM CPU

THE SAME TEAM THAT DESIGNED AND SHIPPED THE WORLD’S 1ST 64-BIT ARM SERVER

EXTENSIVE EXPERIENCE FROM PIONEERING PROCESSOR COMPANIES

arm Qualcomm Apple SAMSUNG intel. AMD digital NVIDIA. IBM. AMPERE Transmeta CORPORATION Sun microsystems