Real Systems. Real Traction.

Bringing High-Performance RISC-V Platforms to Life

Balaji Baktha, CEO, Ventana Micro Systems

RISC-V Europe Summit, Paris 2025

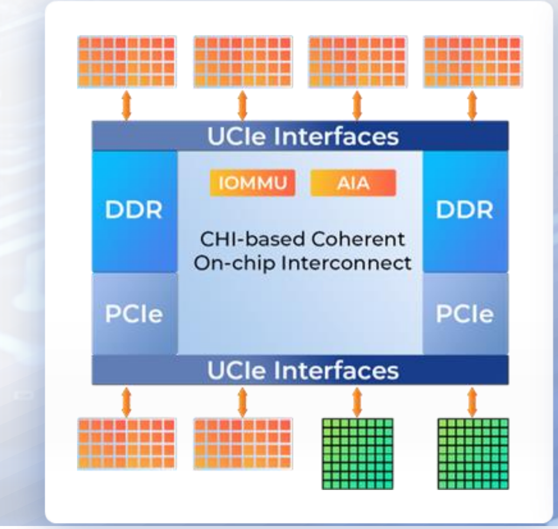
VENTANA Delivers a Full-Scale Compute Platform

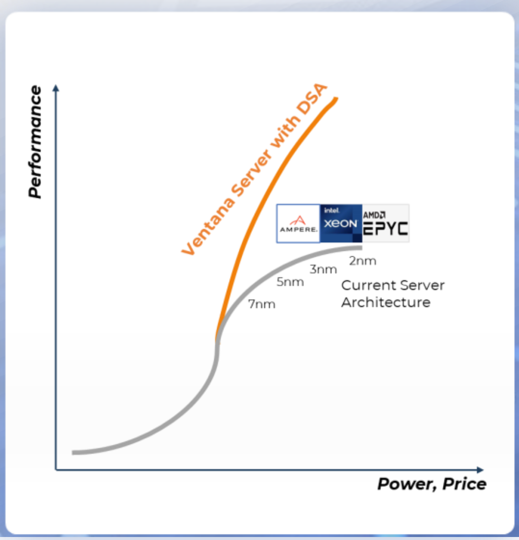
Frictionless - Making it easy for customers to adopt RISC-V in their data center applications

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

RISE

RISC-V Software Ecosystem

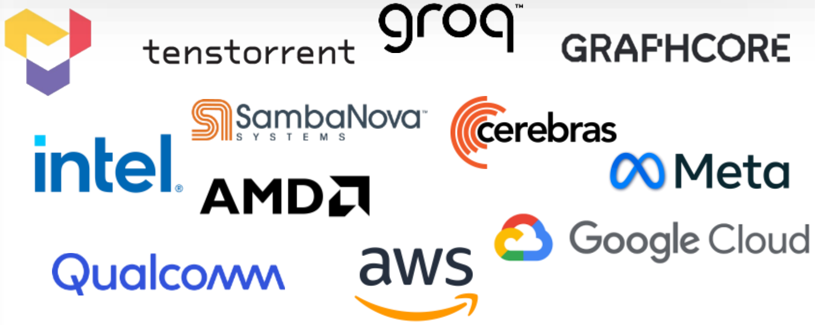




- ✓ SECURITY
- ✓ RAS
- ✓ LOW LATENCY COHERENT NOC
- ✓ HIGH PERFORMANCE CACHE ARCHITECTURE
- ✓ HIGH PERFORMANCE CPU
- ✓ FEATURE PARITY WITH LEADING CPU ARCHITECTURES

AI Acceleration is Fragmented – RISC-V Unifies It

They All Struggle Because



⃠ **Fragmented Architectures**

⃠ **Reinvented Software Stacks**

⃠ **No Shared Investment →**
No CUDA Alternative

AI Workloads Vary Compute Must Scale

Training

Attached Matrix

Integrated Matrix

Vector/Dot Product

Signal Processing/ Classical ML

But They All need a Unified Foundation

RISC-V Enables Ecosystem Scale

RISC-V®

☑ **Share ISA across CPUs and accelerators**

☑ **Scalable to fit across any AI Tier**

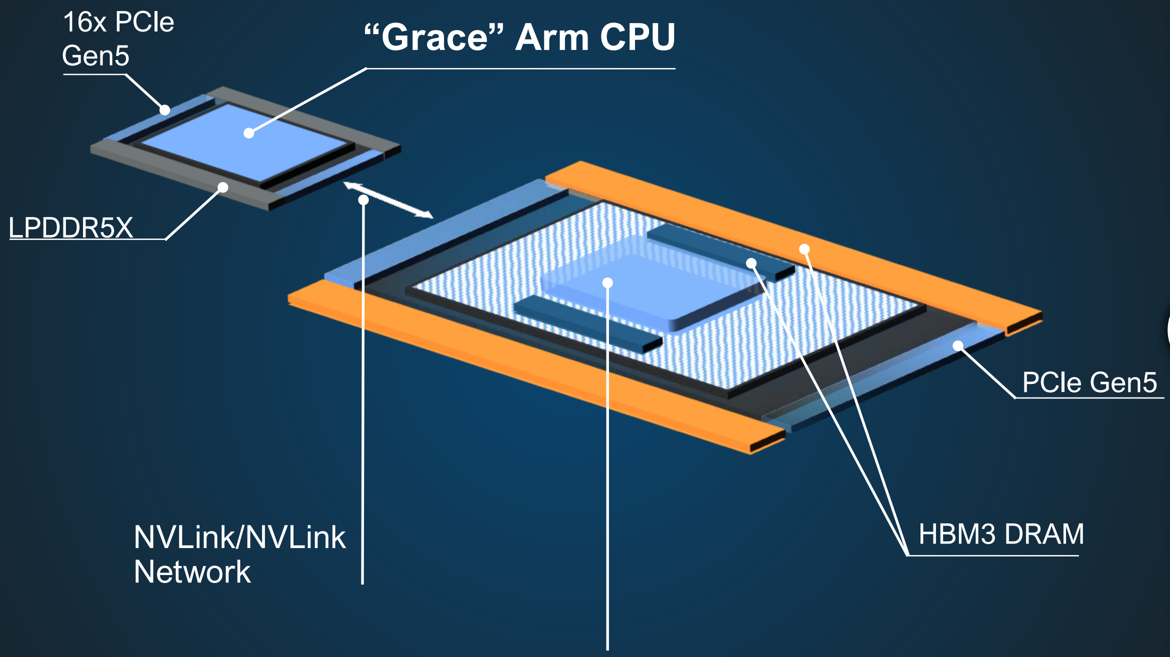
☑ **Common toolchain and compiler infrastructure**

☑ **Ecosystem leverage as industry wide scale**

Different Workloads need different compute.
But they all need a unified foundation.

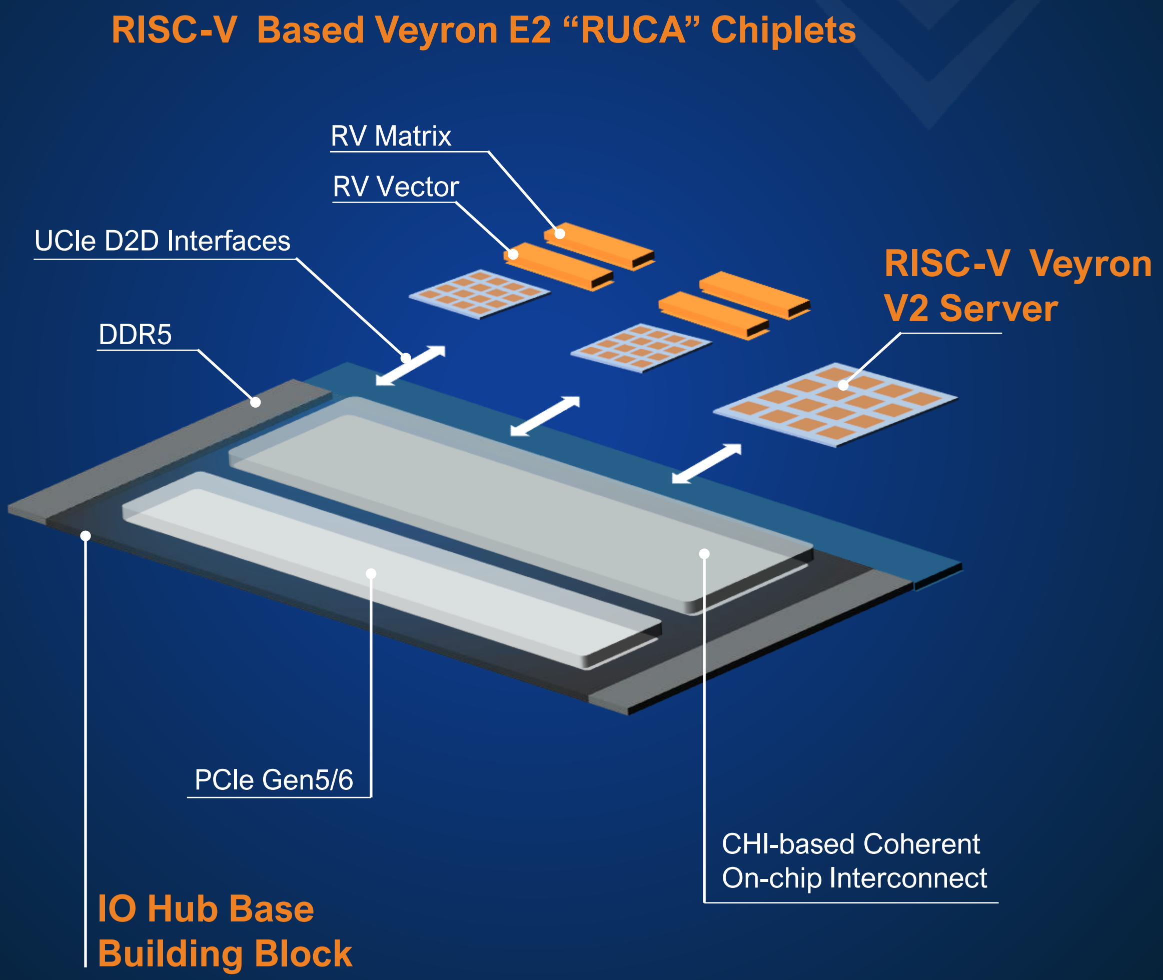
From Grace Blackwell to Open Compute: Ventana Enables the RISC-V AI Superchip

RISC-V Equivalents to Nvidia Grace Blackwell



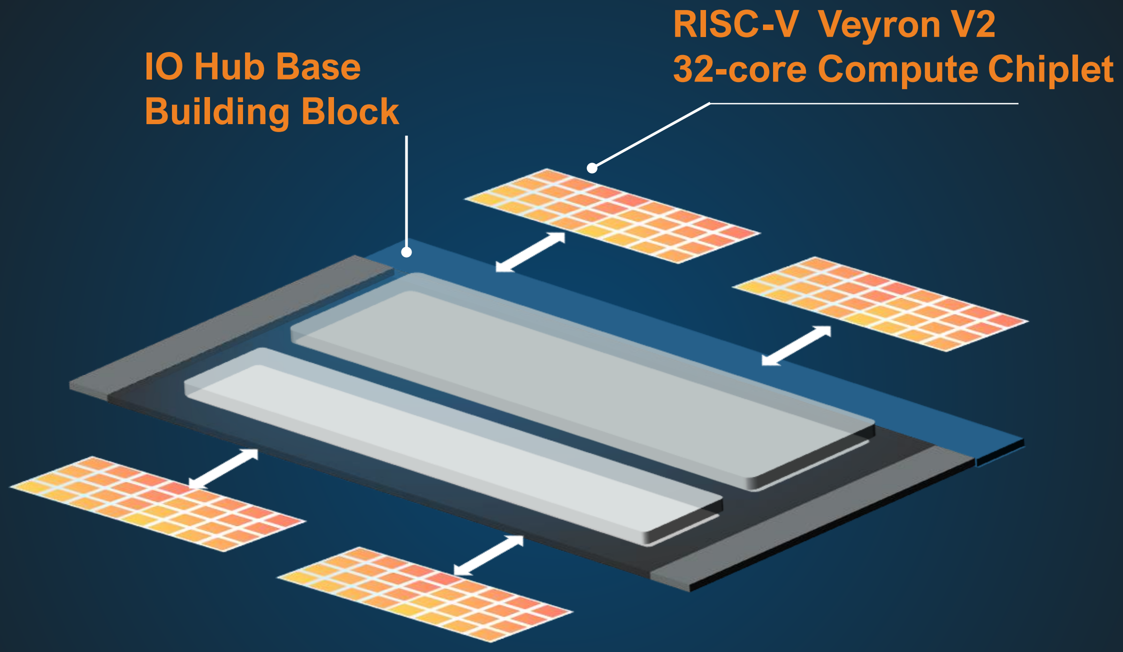
Blackwell GPU
132 Steaming Multiprocessors
5th Generation Tensor Cores

CPU+AI Equivalent Using Ventana RISC-V


RISC-V Based Veyron E2 "RUCA" Chiplets
RV Matrix
RV Vector
UCIe D2D Interfaces
RISC-V Veyron V2 Server
DDR5
PCIe Gen5/6
CHI-based Coherent On-chip Interconnect
IO Hub Base Building Block

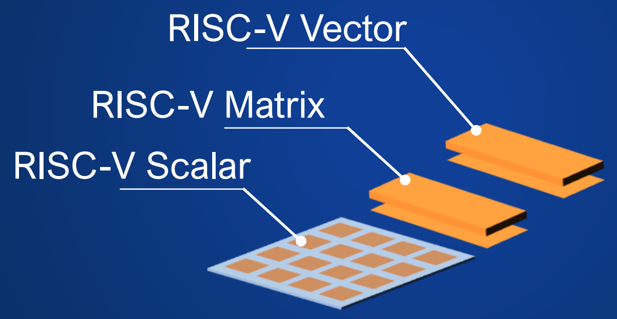
Turning Vision into Action: Real-World Execution from Compute to AI

Cloud Compute Server Partners



AI Acceleration Partners

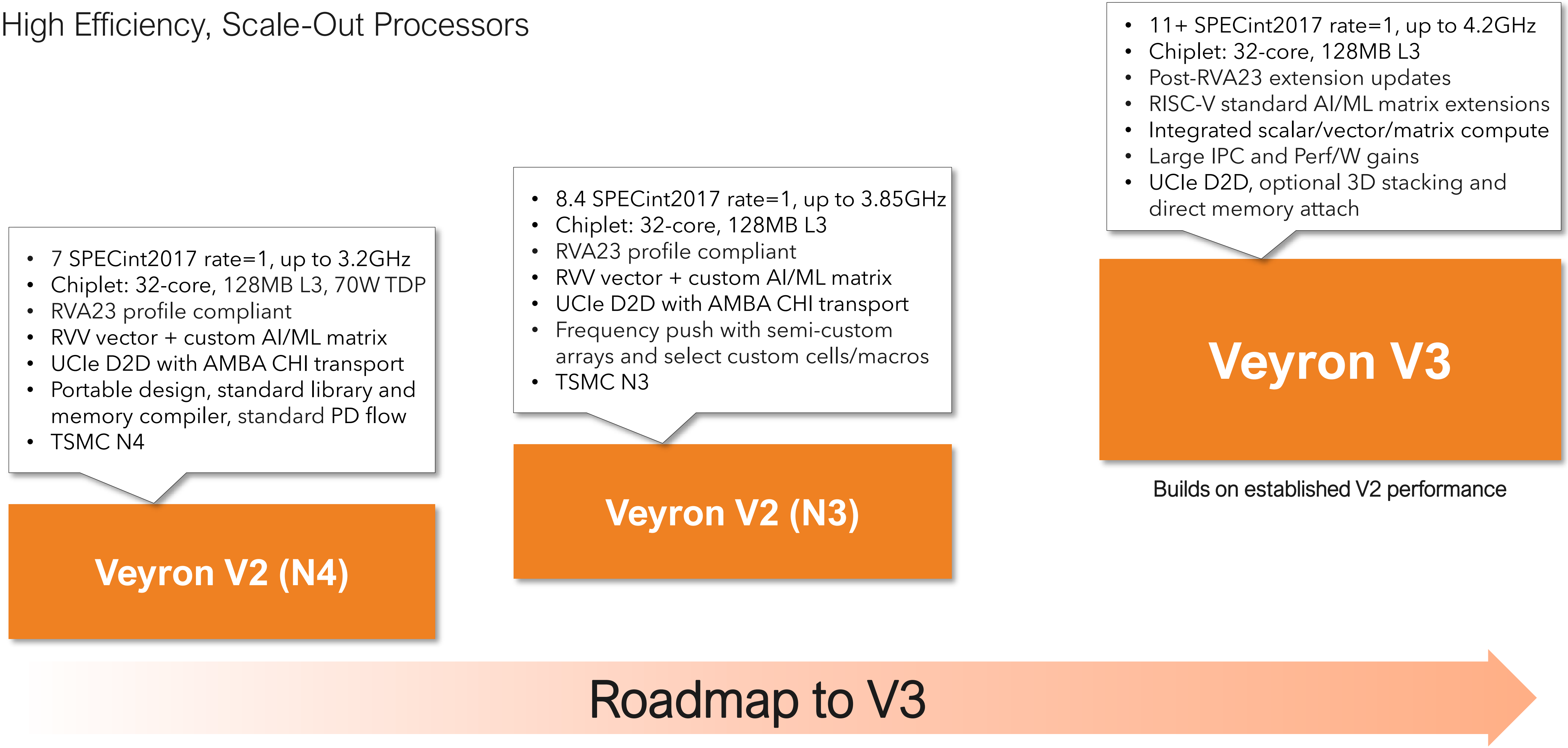
Veyron E2-based “RUCA” AI/HPC Acceleration



Ventana is the only company delivering both RISC-V compute and RISC-V AI Acceleration – in real silicon, at system scale with a high-performance platform ready for the data center

The Big Announcement!

High Efficiency, Scale-Out Processors



Veyron V2 (N4)

- 7 SPECint2017 rate=1, up to 3.2GHz
- Chiplet: 32-core, 128MB L3, 70W TDP
- RVA23 profile compliant
- RVV vector + custom AI/ML matrix
- UCIe D2D with AMBA CHI transport
- Portable design, standard library and memory compiler, standard PD flow
- TSMC N4

Veyron V2 (N3)

- 8.4 SPECint2017 rate=1, up to 3.85GHz
- Chiplet: 32-core, 128MB L3
- RVA23 profile compliant
- RVV vector + custom AI/ML matrix
- UCIe D2D with AMBA CHI transport
- Frequency push with semi-custom arrays and select custom cells/macros
- TSMC N3

Veyron V3

- 11+ SPECint2017 rate=1, up to 4.2GHz
- Chiplet: 32-core, 128MB L3
- Post-RVA23 extension updates
- RISC-V standard AI/ML matrix extensions
- Integrated scalar/vector/matrix compute
- Large IPC and Perf/W gains
- UCIe D2D, optional 3D stacking and direct memory attach

Builds on established V2 performance

Roadmap to V3

Unlike others, Ventana uses absolute SPECint2017 results, not Spec/GHz modeling.
No extrapolations. No outdated 2006 benchmarks. Just real performance.

Veyron V3 Goals and Targets

- **High Single-Thread Performance with High Power Efficiency – Optimized for Scale-Out Datacenter Compute and Maximizing Single-Socket Performance**
  - 11+ SPECint2017 rate=1
  - 24 TeraFlops per core of AI/ML FP8 matrix compute (4.5+ PetaFlops in a single 192-core SiP)
  - High-bandwidth, low-latency coherent fabric enables excellent cluster performance scaling
- **Builds on Proven Veyron V2 Foundation**
  - Large IPC gains through combination of new uarch innovations and enhancements to existing uarch
  - Same base microarchitecture with the foundation already laid for V3 uarch innovations
- **Advanced Microarchitecture for Power & Performance**
  - Dense-compute internal macro-ops, advanced multi-instruction-to-macro-op fusion, caching of macro-op sequences
  - Software-transparent hardware optimization of warm and hot macro-op sequences
  - Atypical uarch choices to achieve both high performance and low power (e.g. virtual DL1 cache eliminates complex power-hungry and timing-critical multi-ported CAM-based L1 TLB)
- **High Frequency of Operation – up to 4.2 GHz**
  - Targeted semi-custom P&R flow for frequency optimization of critical structures
  - Targeted custom high-speed standard cells and SRAM macros
  - Benefits of advanced process node

V3 Microarchitecture Overview

- **Aggressive Out-of-Order design**
  - Superscalar, decouple fetch/predict front end
  - Large caches, BTB, resource queues, schedulers, ROB, etc.
- **Advanced Branch and Memory Prediction**
  - Multiple primary and secondary branch predictors
  - Memory dependence and bypass predictors, load value predictor, ...
- **High Parallelism**
  - 16 execution schedulers and pipelines
  - 5 integer, 3 load/store, 3 scalar FP, 5 vector/matrix
  - 200+ scheduler entries
- **Macro-Op Optimized Microarchitecture → Acts Much Wider Than It Looks**
  - Dense-compute macro-ops encode multiple RISC-V instructions
  - Macro-ops magnify effective decode width, sizes of backend queues and tracking structures, and number of schedulers and execution pipelines - without growing them physically
    - Optimizes performance while saving power and area
  - **Avoids costly brute-force 10-16 wide approaches to increasing IPC**

Macro-Ops, Advanced Fusion Engine, Macro-Op Cache

- **High Compute-Density Macro-Ops**
  - Internal macro-ops encode 1-5 RISC-V instructions
  - Load/store operations with complex address calculations and 4-cycle load-to-use
  - Single-cycle execution for complex three-operand register and branch operations
  - Reduces power per RISC-V instruction across the microarchitecture

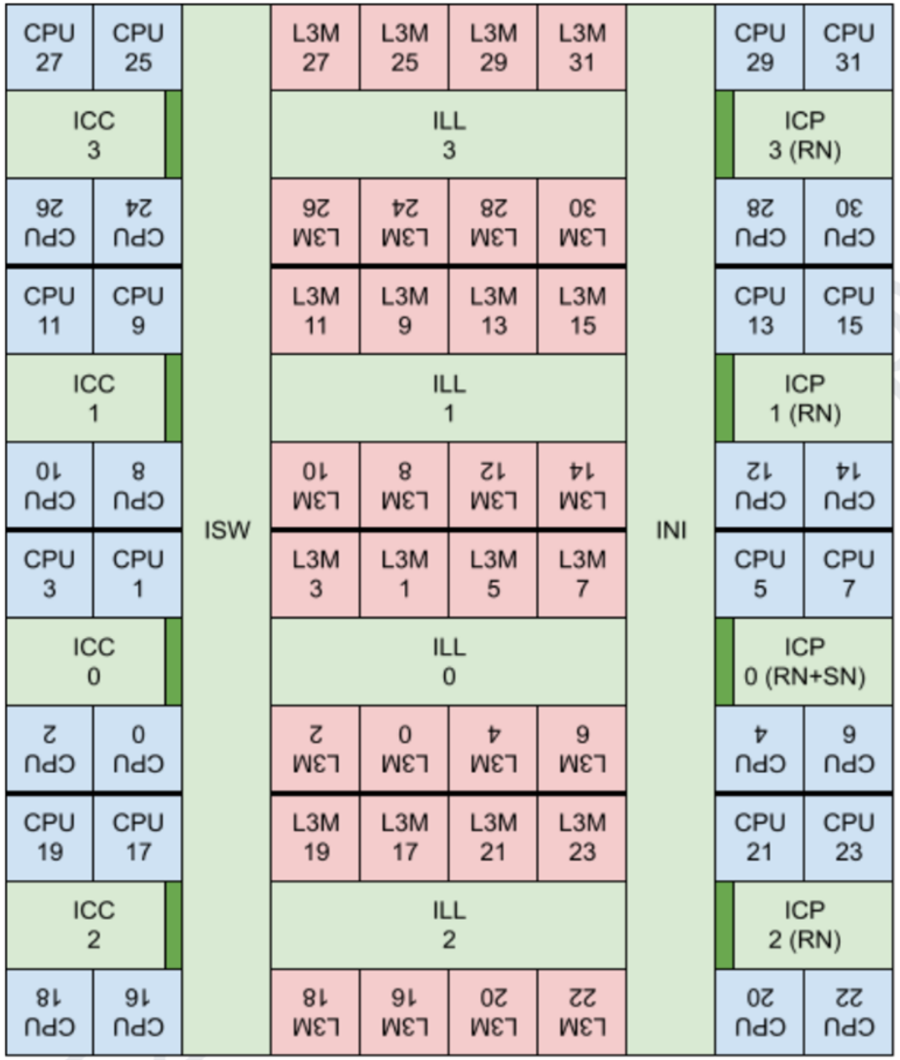
- **Advanced Fusion Engine**
  - Operates on warm and hot sections of RISC-V code
  - Dynamically fuses multiple RISC-V instructions into optimized macro-ops
  - Performs a variety of code optimizations that generally increase ILP and shorten code execution path lengths

- **"IL1" Macro-Op Cache**
  - Caches optimized variable-length macro-op sequences
  - Bypasses fetch/decode and reduces power for most code execution
  - Fully hardware-coherent and TLB-consistent (i.e. fully software transparent)

32-Core / 128 MB L3 Compute Cluster Topology

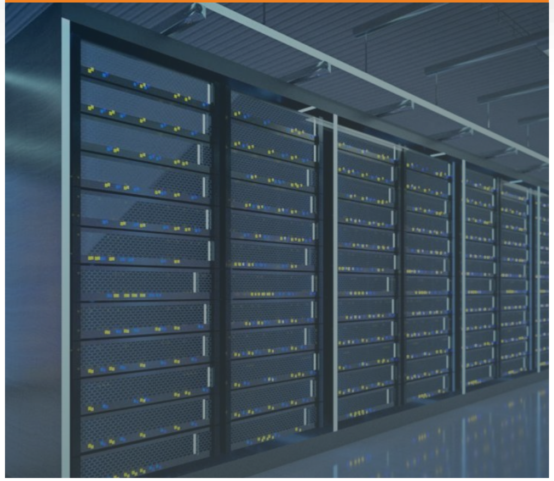
- 32 cores
- 32 L3 slices
- 2-D mesh interconnect modules

- Physically scalable topology
- Interconnect bandwidth and L3 bandwidth scales with core count
- Cluster to SoC bandwidth scales with core count



Scalable RISC-V Architecture for Every High-Performance Market

Data Center



Automotive

Intelligent Edge

Client

Unified Veyron Architecture: Scales from hyperscale compute to ultra-efficient edge and client form factors.

Power, Performance & Safety Flexibility: Data center-class throughput | Automotive-grade features including ASIL-B/D safety (ISO 26262) | Software test libraries | ECC/parity protection, fault isolation | Power-aware clusters for edge and mobile

Platform Coverage: Coherent compute + AI across segments | Shared toolchain, ISA, and optimization path | Production-ready platform delivery for Tier 1 & OEM alignment

VENTANA

- ✓ Highest-Performance RISC-V CPUs–With Production Silicon on the Way
- ✓ Delivering Real High-Performance RISC-V Systems–Hardware, Software, and Platform
- ✓ Scalable AI Acceleration–Data Center to Client, Fully RISC-V
- ✓ Enabling Momentum Across the Ecosystem and Customer Designs
- ✓ Veyron V3 Extends Product Leadership