Veyron V1 Data Center-Class RISC-V Processor

August 2023
Veyron V1 Target Markets

Data Center

Automotive

5G Edge & Generative AI

Client

Company confidential
Veyron V1: Server Class RISC-V IP + Chiplets

Veyron High Performance RISC-V CPU IP

- Up to 16 cores
- 512 KB L1-cache
- 64 KB L1 D-cache
- 512 KB L2 D-cache
- 512 KB I-cache
- 48 MB Shared L3 (sliced per core)
- 64 KB L1 D-cache
- Coherent Cluster Interconnect
- AMBA® CHI (Coherent Interconnect)

Veyron Chiplet Solutions

- Rapid productization with chiplets
- Veyron compute chiplets
  - In latest process node technology
  - Scalable CPU performance/count
- IO Hub
  - Implemented in lower-cost process node of choice
  - Customized for application requirements
- Custom Domain Specific Acceleration

• Superscalar aggressive out-of-order instruction pipeline
• High core count multi-cluster scalability (up to 192 cores)
• Comprehensive RAS features
• IOMMU & Advanced Interrupt Architecture (AIA) system IP
Veyron V1 Overview

16 High Performance RISC-V Cores
- Decode, dispatch, and execute up to eight instructions per cycle
- Symmetric execution of any mix of integer Reg/Ld/St/Br ops per cycle
- Decoupled predict/fetch front-end with advanced branch prediction

High Performance Cache Hierarchy
- 1MB L2 cache per core
- Up to 48MB of globally shared cluster-level L3 cache

Coherent CHI System Integration
- Cluster/chiplet compliant with AMBA Coherent Hub Interface (CHI) system
- ODSA-compliant BoW die-to-die interface covering cost-effective organic to advanced package integrations with Ventana-supplied D2D IP

Server-Class Product
- Full architectural support to run virtualized workloads
- RAS protection of all caches / functional RAMs, with end-to-end data poisoning and background cache scrubbing
- Ground-up microarchitecture with side-channel attack resilience
Veyron V1 Arch/Uarch Overview
RISC-V Architecture Support

- RV64GC plus many additional User, Supervisor, and Machine level architecture extensions
- Hypervisor extension
  - Type 1 and 2 hypervisors; nested virtualization
- Advanced Interrupt Architecture (AIA)
  - Including native MSI handling and interrupt virtualization
- 48-bit virtual addressing and 52-bit physical addressing
- External and self-hosted debug; trace-to-memory
- Rich set of performance events and perf counters
Core Microarchitecture Highlights

• Superscalar, aggressive out-of-order design

• Innovative microarchitecture focused on ...
  • Power-efficiency and high performance
  • Efficient physical implementation and high frequency without custom memory macros

• Decoupled predict / fetch front-end
  • Predict fetch stream ahead of actual just-in-time fetch to keep decode pipe fed
  • Advanced branch prediction of direction and target address
  • High capacity BTB and predictors
  • Fetch up to 64B per cycle; decode up to eight instructions per cycle
  • Code decompression (16b-to-32b) and fusion of common instruction-pair code idioms

• Decode, dispatch, issue, execute, and commit all operate in terms of “ops” (fused and unfused)
Core Microarchitecture Highlights

• Four symmetric integer execution pipes
  • Execute any mix of four register / load / store / branch ops per cycle
  • Int mul/div, pcnt, clmul, and CSR accesses execute via a separate shared execution unit
  • Large associated schedulers – 128-entry scheduling window in total

• Constant register loads pre-executed at dispatch
  • Effective zero-cycle latency and no back-end resources consumed

• Scalar FP execution pipe and int/FP transfer/conversion pipe (and associated schedulers)

• Cache and TLB hierarchies optimized for large code and data working sets, and for low latency
  • 512 KB Instruction L2 with power-efficient L0 cache/loop buffer
  • 64 KB Data L1 / 512 KB Data L2 closely coupled for low latency
  • Separate 3K+ entry main Instruction TLB and Data TLB (including caching clusters of similar PTEs)
Predict, Fetch, and Decode Units

• Predict fetch stream of sequential runs of instructions up to 64B long
  • Single-level 12K-entry BTB and similarly large collection of branch predictors
  • Fully-pipelined, driven by single-cycle Next Lookup Predictor
    • Predicts lookup hashes and history updates
    • Three-cycle redirect on mispredict

• IL2 + ITLB (large single-level instruction cache and instruction TLB)
  • 512 KB IL2
  • Physical I/D partitioning allows separate I and D cache hierarchy optimizations for latency and power, and eliminates code/data conflicts on large footprint workloads
  • Fully pipelined misaligned fetch of up to 64B per cycle
  • Two-cycle latency for overlapped ITLB, IL2 tag, and IL2 data accesses

• First instruction decode pipe stage does ...
  • Decompress 16-bit ‘C’ instructions to equivalent 32-bit instructions
  • Pre-decode instruction length and find next 8 instruction boundaries
  • Pre-decode instruction pair fusion opportunities
  • Combine all this together to set up muxes to extract instructions from instruction buffer
Load/Store Unit

- Can execute any mix of up to four loads and/or stores per cycle
- Closely-coupled L1/L2 data cache hierarchy for low latency
- DL1
  - 64 KB virtual cache (VIVT)
  - Four-cycle load-to-use latency
  - Large single-level DTLB accessed on cache misses (on the way to DL2)
  - Hardware synonym handling – multiple read-only synonyms can be co-resident
  - Hardware coherent based on inclusion wrt DL2
  - Hardware TLB consistent wrt TLB invalidates
- 512 KB DL2
  - Pipelined 64B-wide fills into DL1
- Hardware data prefetchers
  - Next line, sequential, strided, and multi-stride patterns
  - Prefetch next line from DL2 into DL1
  - Prefetch much farther ahead from L3/DRAM into DL2 as staging
Processor Cluster Highlights

• Support for up to 16 cores

• Cluster-level shared L3 cache
  • Support for up to 48 MB
  • Victim cache with respect to DL2's
  • Non-inclusive (exclusive except for selective shared code/data optimizations)
  • Advanced reuse-based and scan-resistant replacement policies

• N-way sliced L3 / snoop filter organization
  • Each slice responsible for 1/Nth of address space
  • “Core + L3/SF slice” physical building block
  • Per-core (non-shared) cluster-level snoop filters for IL2 and DL2 caches
Processor Cluster Highlights (cont.)

• Standard CHI-compatible external interface from cluster to SoC
  • Enables direct connect to 3rd party SOC interconnect IP

• Enhanced intra-cluster cache coherency protocol
  • Comparable to CHI plus features to support various caching optimizations within a cluster
    • Exclusive / non-inclusive cache allocation
    • Data sharing
    • Enhanced L3 replacement policy

• Bidirectional “race track” interconnect topology
  • Equivalent to dual counter-rotating rings with ends cut off
  • Best PPA for up to 16 cores
  • 160 GB/s of bisection data bandwidth at 2.5 GHz
Veyron V1: World’s First Server Class RISC-V Processor

**Highest Performance RISC-V CPU**

3.6GHz in 5nm process technology

**ASSP Based on High Performance Chiplet Architecture**

Significant reduction in development Time and Cost compared to prevailing monolithic SoC model

Disruptive ROI: Highest Single Socket Performance at Compelling Perf/Watt/$
Veyron V1 Reference Implementation PPA

- **TSMC 5nm**
  - Standard TSMC 5nm metal stack
  - Width linearly scales with tiled dual core+L3 slice
  - Highly portable design across processes and foundries

- **Veyron V1 cluster structure**
  - Up to 16 cores with fixed private 512 KB IL2 and 64 KB DL1 / 512 KB DL2
  - Up to 48 MB shared L3, physically sliced per core
  - Configurable 2/4/8/16 Core+L3 slices, 3.0/1.5/0.75 MB L3 per core

- **5-6 SPECint2017 @ 3.6GHz with 40W total cluster power**
  - Excellent multi-core scalability with high bandwidth interconnect and large L3 cache
  - Dedicated core per thread provides superior multi-core performance compared to large SMT2 cluster (equal threads, same area, twice the cores)

- **Per-core power under max “TDP” workloads**
  - <0.9W @ 2.4 GHz
  - 1.9W @ 3.2 GHz

- **Active “Turbo” power management**
  - Per cluster DVFS, per core DFS
  - Accurate digital power model for all components of cluster
  - Temp sensor coverage across entire cluster
  - Configurable TDP
Thank You