RISC-V Heterogeneous Programming Paradigm: Atomic IO Enqueue (AIOE) Extension and AIOE with Virtualization

Abstract

In the era of artificial intelligence, a single CPU architecture can no longer meet the demands of diverse intelligent computing workloads. Consequently, heterogeneous computing has emerged as the mainstream approach. Integrating different Domain-Specific Accelerators (DSAs) into computing systems could enhance overall computational efficiency. For instance, high-dimensional tensor computing tasks can be offloaded to TPUs/NPUs/GPGPUs, while data stream processing tasks are delegated to DPUs. The challenge of efficiently managing DSAs in heterogeneous systems has become a prominent industry focus, driving several technological advancements:

- PCI-e 5.0 and CXL 2.0 have introduced Deferrable Memory Write (DWMr) TLPs
- Armv8.7/9.2 has incorporated ST64BV0 instructions for 64-byte atomic I/O enqueue operations
- The x86 architecture has implemented ENQCMD instructions with comparable functionality

These innovations collectively reduce control latency and optimize system resource utilization.

To help RISC-V adapt to this trend, the presentation introduces the Atomic IO Enqueue Extension (AIOE) and AIOE with virtualization. The AIOE extension is designed for RV64 ISA, which includes one PMA definition, two U-mode instructions, two S-mode instructions, a single S-mode CSR, and two envcfg control bits. The presentation also introduces how to use AIOE under the virtualization scenario, which involves the new proposal for RISC-V IOMMU: G-stage table In Process Context (GIPC). With the help of AIOE and GIPC, RISC-V could explore a new heterogeneous programming paradigm from HPC to embedded scenarios.