SPACEMIT
进迭时空

RISC-V IO Virtualization Implementation on X100

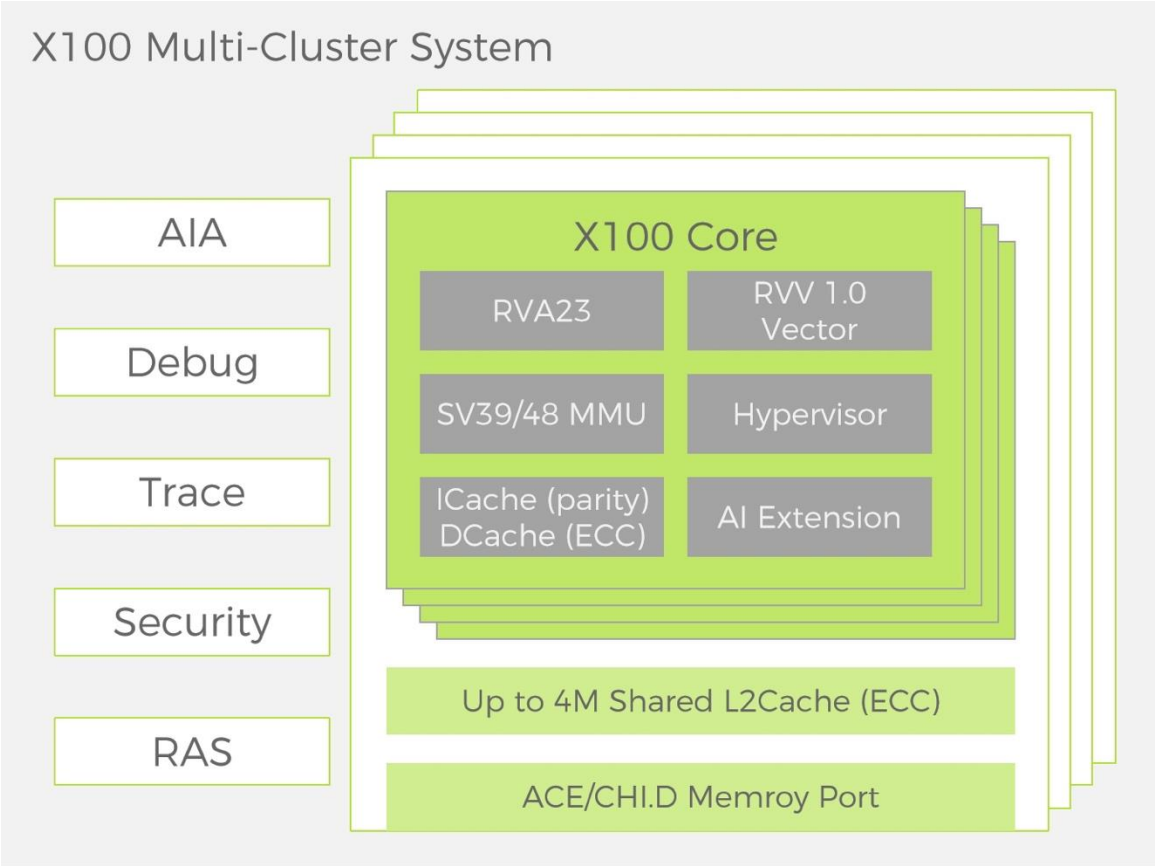
Lv 'ZETALOG' Zheng

X100 - RISC-V Virtualization Infrastructure

RISC-V High Performance Core w/ CPU Virtualization

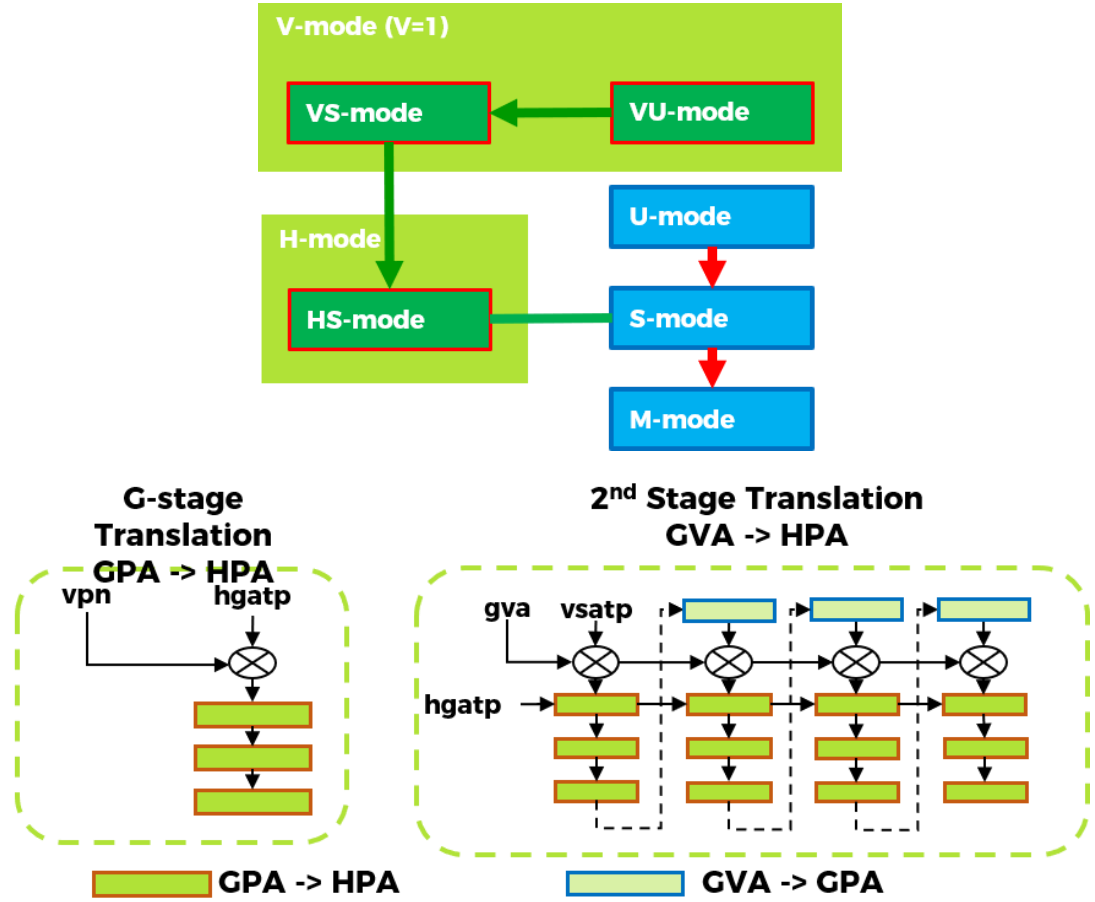
High Performance RISC-V Core

- 64-bit high performance RISC-V processor
- RVA2023 compatible RV64GCVBH
- Spec2k6Int 9.0/GHz 2.5GHz@T12
- Full RVV1.0 support
- SpacemiT IME (Integrated Matrix Extensions) , 4-core fusion AI computing power INT8 2.5TOPS 2.5GHz@T12
- Multi-core, multi-cluster, maximum 64-cores per chip
- CHI bus interface, and multi-die, multi-chip support
- Server spec security, RAS, debug facilities



Full RISC-V Hypervisor Extension Support

- Hypervisor extended S (HS) mode
  - HS-mode CSR, time delta HPM
  - Guest page/virtual instruction fault/HS-mode ECALL
  - VSEI/VSTI/VSSI
- Virtualization Mode (V)
  - Virtualized supervisor (VS) mode/Virtualized user (VU) mode
  - VS-mode background CSR
  - ECALL from VS-mode/VU-mode
- Address translations
  - hgatp pointing to G-stage translation table
  - vsatp pointing to VS-stage translation table
  - PTW/TLB with GPA support
- Instructions
  - HS-mode system barrier
  - Guest memory access

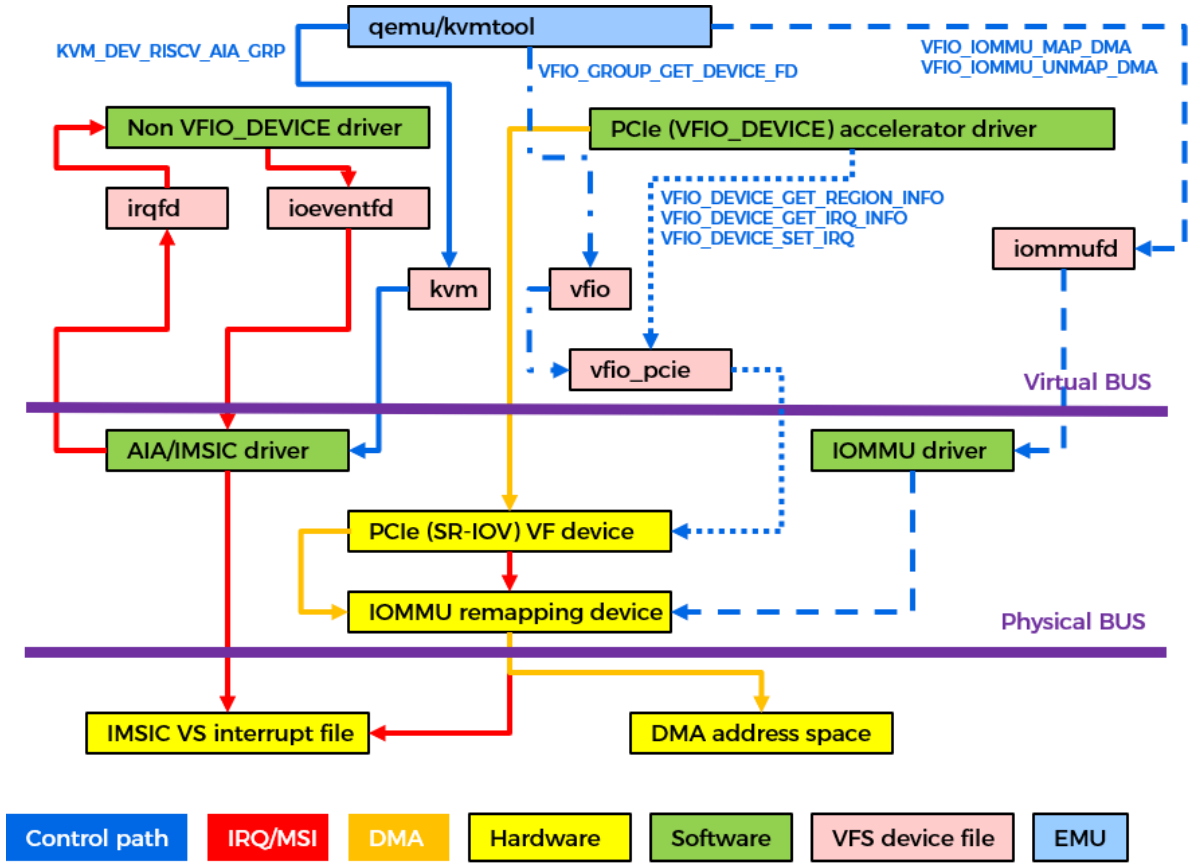


RISC-V IO Virtualization Architecture

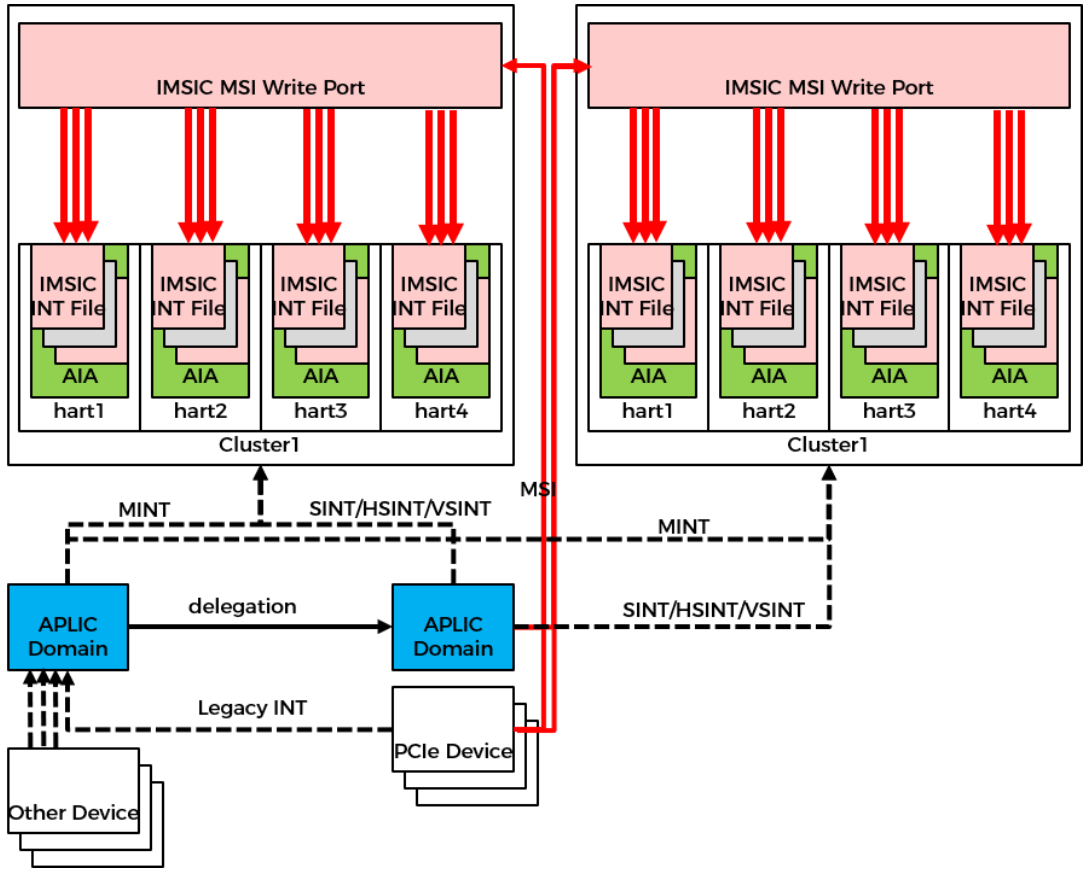
RISC-V IO Virtualization

IO Virtualization Solutions

- Community present condition
  - S2 direct DMA access in guest OS
    - VFIO supports PCI and IOMMU
    - IOMMU is a VFIO building block to create S2 remapping of VFIO device IO regions (VFIO_IOMMU_MAP/UNMAP_DMA)
    - VFIO_PCI binds SR-IOV virtual function to guest OS
  - MSI based IRQ virtualization
    - /dev/kvm setup AIA attributes
    - Associate KVM eventfd/irqfd with AIA VS interrupt file using architecture specific hooks
- Community future trends
  - KVM IRQ Bypass
  - S1+S2 support
    - VirtIO Data Path Acceleration (vDPA)
    - Guest OS owns PC, S1 tables, pointers in PC, S1 PPNs should exhibit implicit S2 translations

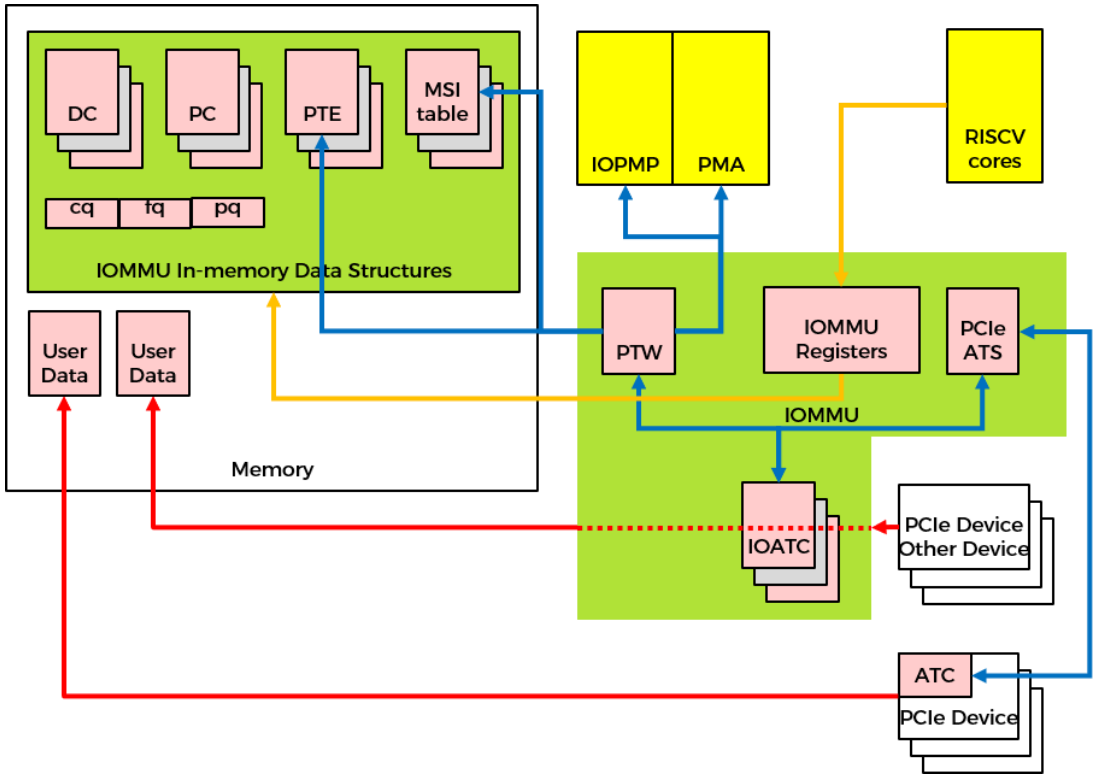


AIA Applications



- **AIA v1.0 compliant**
  - **Major IRQ priority**
  - **Virtual interrupt, VTI**
  - **GEILEN >= 8**
  - **WSI and MSI support**
  - **1023 APLIC interrupt sources**
  - **2048 IMSIC interrupt sources**
- **Prioritized platform IRQs:**
  - **RAS CE, RAS NE, NMI, debug and trace, etc.**
- **MSI support**
- **PCIe interrupt routing**
- **IRQ virtualization and remapping**
- **Interrupt domain**
- **GSI/IPI support**

IOMMU Applications



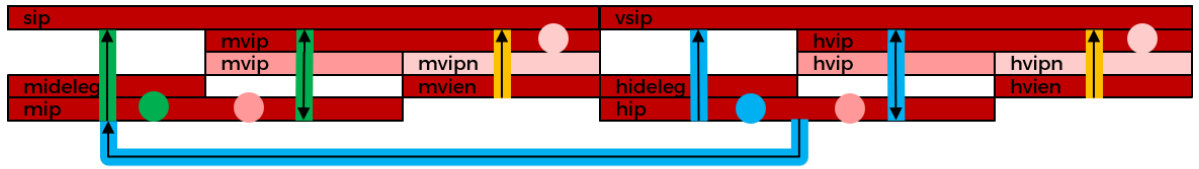
- IOMMU v1.0 Compliant
  - 20-bit DDI, 16-bit PDI, 44-bit DMA
  - Sv39/Sv48 S1 + Sv39x4 Sv48x4 S2 support
  - PCIe ATS/T2GPA/PRI support
  - MSI flat support
  - IOPMP/PMA check
  - Svpbmt/Svnapot support
- Interrupt remapping and virtualization
- DMA remapping, memory protection
- Pointer-is-a-Pointer in heterogenous computing (accelerators)
- Nested address translation in Guest OS
- Translation Cache in PCIe device
- Demand Paging from PCIe device

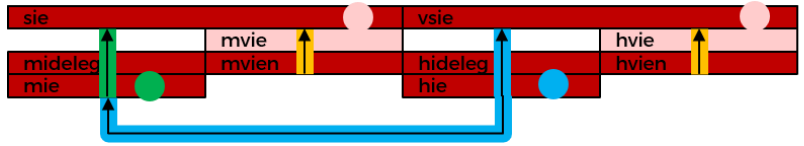
IRQ Virtualization Support inside of X100 (X100 AIA)

X100 AIA - RISC-V Advance Interrupt Architecture

AIA Register Models

- **Wired Interrupt Pending**
  - **xip bits wired from real hardware interrupt lines**
- **Interrupt Filtering (xvien=0)**
  - **xip bits delegated from lower privileged level**
  - **Aliased in xvip bits**
- **Virtual Interrupts (xvien=1)**
  - **Invisibly stored in xvipn bits when xvien=0**
  - **Non-aliased bits in xvip, aliased in xvipn bits**
- **Virtual Trap Interrupt (VTI)**
  - **Used to save registers for virtual machines**
  - **Priority specified by DPR rules**
- **Configurability**
  - **SUPER_INT_MASK/HYPER_INT_MASK: mutal exclusive platform specific major interrupts**
  - **VIRT_INT_MASK: platform specific major interrupts that can support virtual interrupt register model**



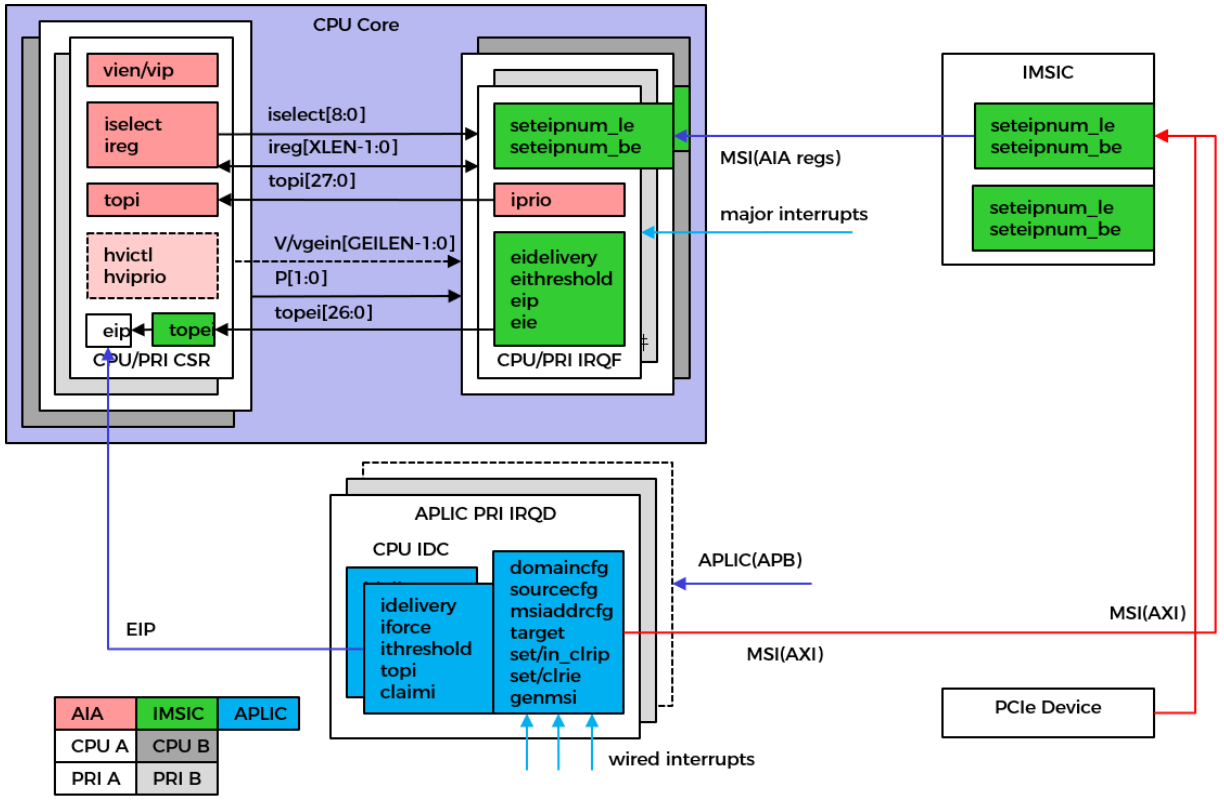


SUPER_INT_MASK	SUPER_INT_MASK/HYPER_INT_MASK: mutual exclusive
HYPER_INT_MASK	
VIRT_INT_MASK	VIRT_INT_MASK: S/VS specific, M/HS invisible virtual interrupt

Interrupt filtering: xvien=0	m|hvip: m|hvip alias bit of m|hip when m|hvien=0
Virtual interrupt: xvien=1	m|hvie: s|vsie writeable bit when m|hvien=1
	m|hvipn: m|hvip non-alias bit when m|hvien=1

AIA/IMSIC/APLIC Verification

- UT level test benches
  - AIA (in-core modules including MSI write port)
    - Fully randomizable, checkable
  - APLIC domain
    - Fully randomizable, checkable
- ST level test benches
  - IMSIC (include AXI MSI write port)
  - AIA + IMSIC + APLIC

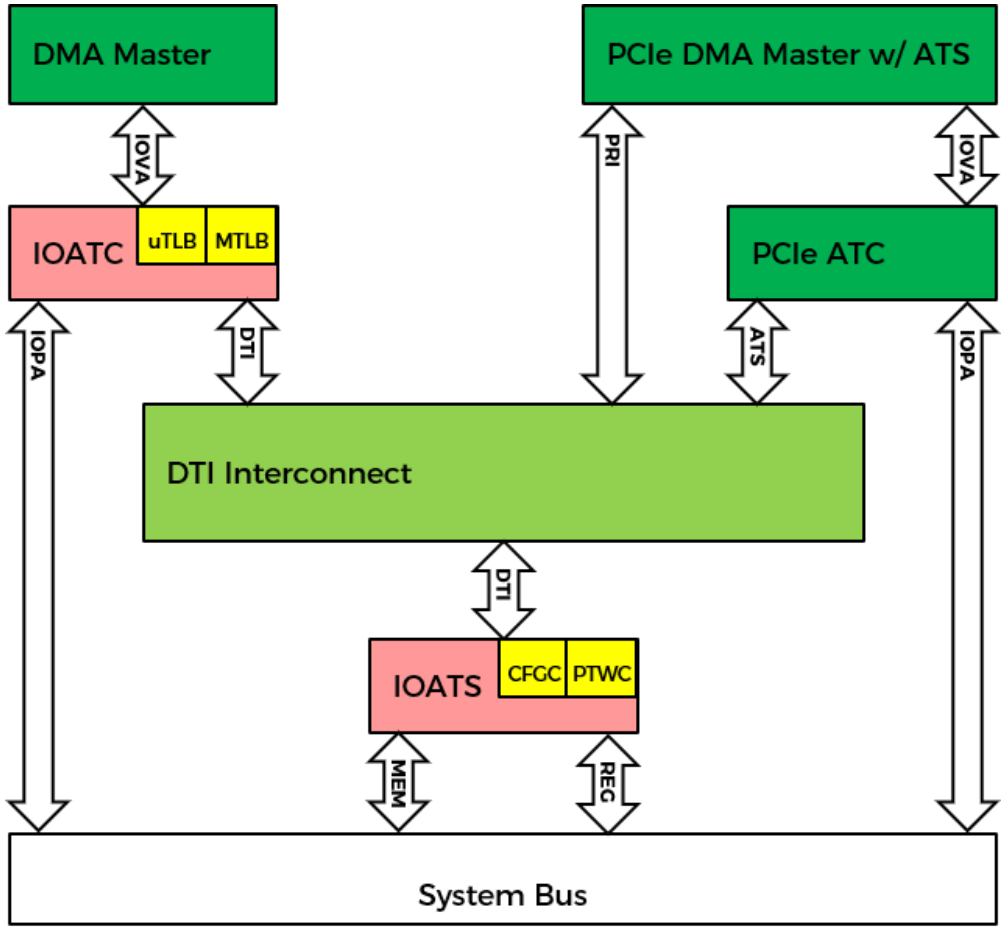


IO Virtualization Support outside of X100 (T100)

T100 IOMMU - IO Virtualization Applications

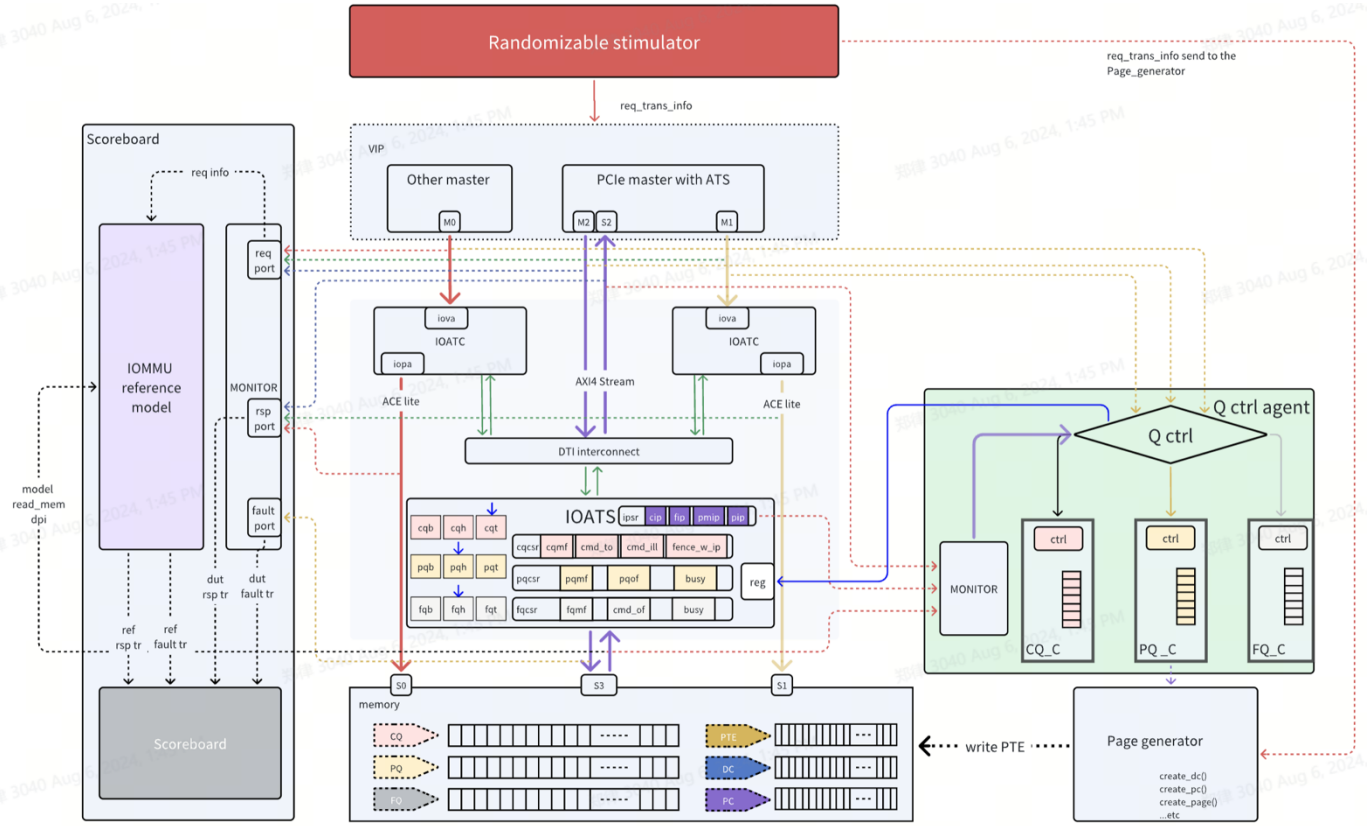
IOMMU IOATS/IOATC Module

- IOATS
  - CFG cache: 4-way set associative
  - DC+PC transaction information
  - PTW cache: 4-way set associative
  - S1/S2-L0/1/2/3 walk cache
  - DTI interface support
- IOATC
  - Micro TLB: fully associative
  - Main TLB: 4-way set associative
  - S1+S2 nested translation
  - DTI interface support



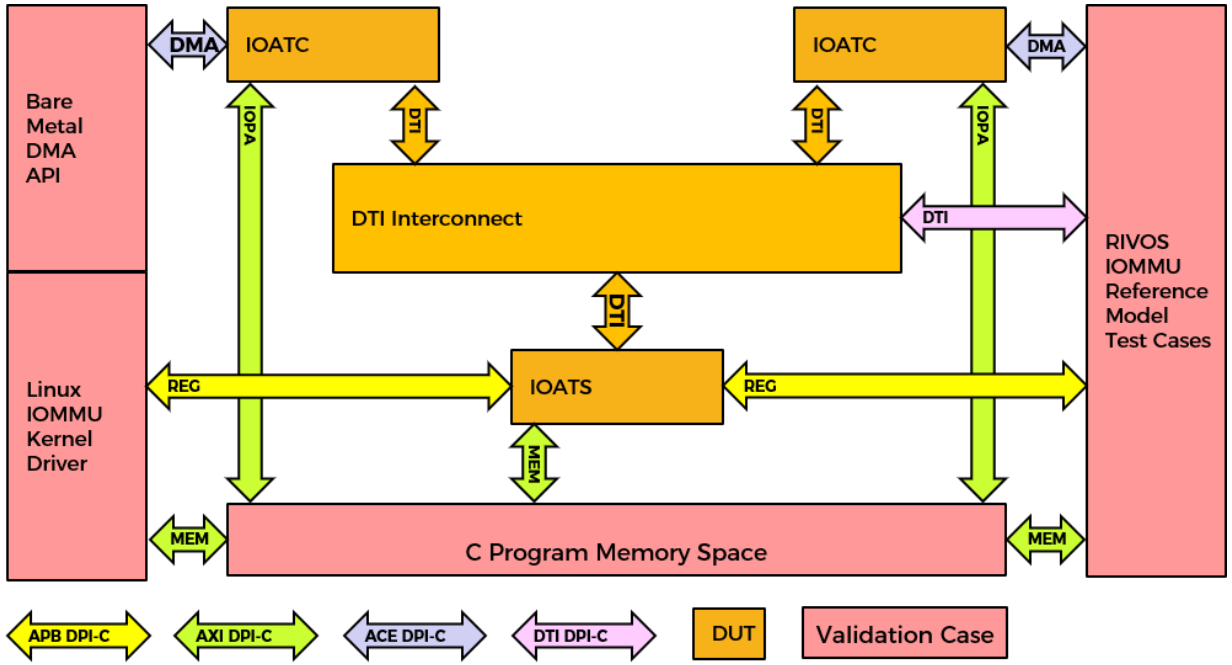
IOMMU Verification (System-Verilog)

- Page generator
  - DDT/PDT/S1 PT/S2 PT/MSI PT
- Community reference model integration
  - DTI based checker
- Randomizable stimulator
  - DDT level/PDT level
  - S1 mode/S2 mode/S1 granularity/S2 granularity
  - IOVA/IOPA/RWXU
  - Svpbmt/Svnapot
- Direct cq/fq/pq case



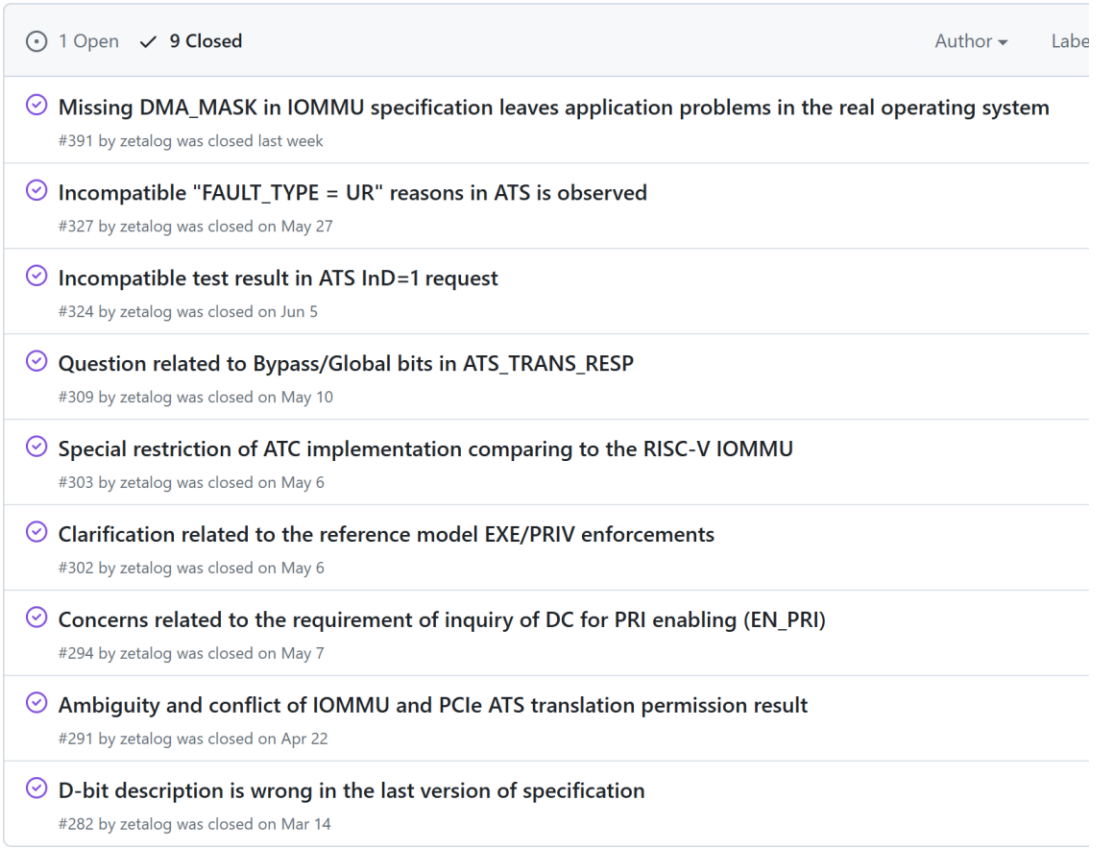
IOMMU Validation (System-C)

- Linux Kernel Driver
  - DMA request via DPI-C
  - DUT accesses programmed configuration tables, page tables, queues via DPI-C
  - IOMMU registers programmed via DPI-C
  - DUT memory accesses are monitored and validated
- RIVOS Reference Model
  - DMA/ATS request via DPI-C
  - DUT accesses programmed configuration tables, page tables, queues via DPI-C
  - IOMMU registers programmed via DPI-C
  - IOMMU register contents validated w/ reference model
  - DUT memory contents validated w/ reference model



Community Technical Leadership

- Specification clarification
  - PCIe ATS related ambiguity
  - A/D bit ambiguity
- Reference model correction
  - PCIe TLP v.s. PCIe DTI TB requirement
  - Fault type/value details
- Usage model discussion
  - EN_PRI check
  - Qemu MSI detection
  - DMA_MASK restriction



FPGA Demonstrations

IO Virtualization Implementations

X100 FPGA Implementations

- S1 Demo
  - IOMMU+DMAC
- S2 Demo
  - IOMMU+PCIe+NVMe+SR-IOV

```
root@buildroot:~# lspci
00:00.0 PCI bridge: Synopsys, Inc. DWC_usb3 / PCIe bridge (rev 01)
01:00.0 Non-Volatile memory controller: DapuStor Corporation NVMe SSD Controller DPU600
root@buildroot:~# nvme virt-mgmt /dev/nvme0 -c 1 -r 0 -n 14 -a 8
success, Number of Resources allocated:0xe
root@buildroot:~# nvme virt-mgmt /dev/nvme0 -c 1 -r 1 -n 14 -a 8
success, Number of Resources allocated:0xe
root@buildroot:~# nvme virt-mgmt /dev/nvme0 -c 1 -a 9
success, Number of Resources allocated:0
root@buildroot:~# echo 1 > /sys/class/nvme/nvme0/device/sriov_numvfs
[  328.526065] pci 0000:01:00.1: [1e3b:0600] type 00 class 0x010802
[  328.534495] pci 0000:01:00.1: enabling Extended Tags
[  328.563277] nvme 0000:01:00.1: Adding to iommu group 3
[  328.570168] domain alloc 4
[  328.574279] nvme 0000:01:00.1: domain type 11 attached w/ PSCID 4
[  328.592736] nvme nvme1: pci function 0000:01:00.1
[  328.599028] nvme 0000:01:00.1: enabling device (0000 -> 0002)
[  328.649687] nvme nvme1: 1/0/0 default/read/poll queues
[  328.706980]  nvme1n1: p1
root@buildroot:~# lspci
00:00.0 PCI bridge: Synopsys, Inc. DWC_usb3 / PCIe bridge (rev 01)
01:00.0 Non-Volatile memory controller: DapuStor Corporation NVMe SSD Controller DPU600
01:00.1 Non-Volatile memory controller: DapuStor Corporation NVMe SSD Controller DPU600
root@buildroot:~#
```

```
Run init for debuging based on ramdisk ...

~ # ls /sys/class/dma
dma0chan0   dma0chan11  dma0chan14  dma0chan3   dma0chan6   dma0chan9
dma0chan1   dma0chan12  dma0chan15  dma0chan4   dma0chan7
dma0chan10  dma0chan13  dma0chan2   dma0chan5   dma0chan8
~ #
~ #
~ #
~ #
~ #
~ # cd /sys/module/dmatest/parameters/
/sys/module/dmatest/parameters # echo 1 > iterations
/sys/module/dmatest/parameters # echo 1 > norandom
/sys/module/dmatest/parameters # echo 1024 > transfer_size
/sys/module/dmatest/parameters # echo dma0chan0 > channel
[  131.217454] dmatest: Added 1 threads using dma0chan0
/sys/module/dmatest/parameters # echo 1 > run
[  131.246057] dmatest: Started 1 threads using dma0chan0
/sys/module/dmatest/parameters # [  131.263895] iommu: map: iova 0xffffc000 pa 0x0000000004f88000 size 0x4000
[  131.273882] iommu: map: iova 0xffff8000 pa 0x0000000004f8c000 size 0x4000
[  131.282506] iommu: map: iova 0xffff7000 pa 0x0000000004f4a000 size 0x1000
[  131.300534] dmatest: dma0chan0-copy0: summary 1 tests, 0 failures 33.71 iops 33 KB/s (0)
```

```
[   70.017781] pci-host-generic 30000000.pci: host bridge /soc/pci@30000000 ranges:
[   70.020960] pci-host-generic 30000000.pci:       IO 0x0003000000..0x000300ffff -> 0x0000000000
[   70.023007] pci-host-generic 30000000.pci:      MEM 0x0040000000..0x007fffffff -> 0x0040000000
[   70.024489] pci-host-generic 30000000.pci:      MEM 0x0400000000..0x07ffffffff -> 0x0400000000
[   70.027333] pci-host-generic 30000000.pci: Memory resource size exceeds max for 32 bits
[   70.031015] pci-host-generic 30000000.pci: ECAM at [mem 0x30000000-0x3fffffff] for [bus 00-ff]
[   70.043876] pci-host-generic 30000000.pci: PCI host bridge to bus 0000:00
[   70.045058] pci_bus 0000:00: root bus resource [bus 00-ff]
[   70.045525] pci_bus 0000:00: root bus resource [io  0x0000-0xffff]
[   70.046409] pci_bus 0000:00: root bus resource [mem 0x40000000-0x7fffffff]
[   70.046873] pci_bus 0000:00: root bus resource [mem 0x400000000-0x7fffffff]
[   70.054786] pci 0000:00:00.0: [1b36:0008] type 00 class 0x060000 conventional PCI endpoint
[   70.113420] pci 0000:00:01.0: [1e3b:0600] type 00 class 0x010802 PCIe Root Complex Integrated Endpoint
[   70.127845] pci 0000:00:01.0: BAR 0 [mem 0x00000000-0x0000ffff 64bit]
[   70.152726] pci 0000:00:01.0: enabling Extended Tags
[   70.262725] pci 0000:00:01.0: BAR 0 [mem 0x400000000-0x40000ffff 64bit]: assigned
[   70.503598] nvme nvme0: pci function 0000:00:01.0
[   70.507535] nvme 0000:00:01.0: enabling device (0000 -> 0002)
[   71.517797] nvme nvme0: 1/0/0 default/read/poll queues
[   72.111146]  nvme0n1: p1
[   72.635177] printk: legacy console [ttyS0] disabled
```

THANK YOU!

SpacemiT – High Performance RISC-V Silicon

spacemit.com