Timesintelli Technology

TiMESiNTELLi 时擎科技

高性能RISC-V SOC架构进展与展望

时擎科技 倪潇飞

TiMESiNTELLi 时擎科技

目录

CONTENTS

Part.1 高性能RISC-V SOC的崛起

为何需要高性能RISC-V SOC?

算力需求爆发

AIoT
边缘计算
数据中心
AI大模型训练和推理

对比ARM/x86的差异化优势

模块化
可定制化
免授权费

自主可控与生态开放性的战略价值

地缘政治安全屏障
全栈技术自主权
模块化架构激发定制创新
全球协作生态降低创新门槛

定义“高性能”指标

TiMESiNTELLi 时擎科技

01 计算性能

通用CPU：多核设计、乱序执行、高主频

GPU/NPU：FLOPS/TOPS

02 内存和存储支持

高带宽

大容量

DDR5/HBM

定义“高性能”指标

能效比

- 成本： 数据中心电费是运营的主要成本之一，高能效意味着更低的运营成本。
- 散热： 高功耗带来散热挑战，影响系统设计和可靠性。
- 移动/边缘设备： 电池续航和散热空间有限，能效比是核心指标。
- 环境： 降低碳排放。

安全性和可靠性

- 硬件安全根 (Hardware Root of Trust)
- 硬件安全引擎 (Hardware Security Engines)
- 硬件强制隔离 (Hardware-Enforced Isolation)
- 调试安全 (Debug Security)
- 容错设计 (Fault Tolerance)
- 错误检测与报告 (Error Detection and Reporting)
- 鲁棒性设计 (Robust Design)

定义“高性能”指标

生态系统支持

- 完善工具链
- 适配操作系统
- IP生态成熟

高速接口支持

- PCIe
- USB
- HDMI
- Ethernet
- ...

特定应用场景

- 边缘 AI 计算与轻量化大模型推理
- 工业物联网（IIoT）与苛刻环境应用
- 端侧 AIoT 与多模态感知
- 智能视觉处理与安防监控
- 可穿戴与微型化设备

Part.2 高性能RISC-V SOC的现状与挑战

高性能RISC-V SOC的现状

TiMESiNTELLi 时擎科技

01

CPU核心架构升级

- 乱序执行与多核扩展
- AI原生融合设计
- 异构计算架构

02

性能突破

- 桌面和服务器级芯片涌现
- 边缘AI芯片：大模型端侧推理落地
- 工业级处理器：高可靠与低功耗兼顾

03

内存和接口发展

- LPDDR5
- DDR5
- PCIe4.0/5.0
- USB3.0

04

系统级优化

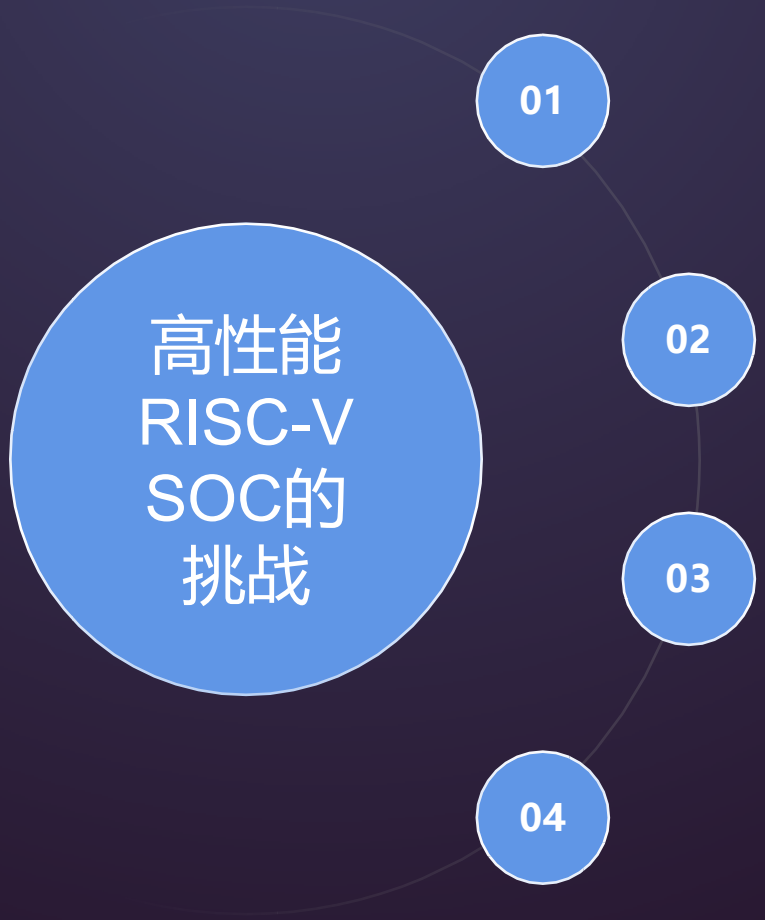
- 软硬件协同
- 虚拟化与安全性待优化

05

生态进展

- 标准化与兼容性完善中
- 软件栈有所发展
- 产业链协作创新

高性能RISC-V SOC的挑战



性能与能效瓶颈

架构深度优化不足
能效比仍有差距

安全与可靠性欠缺

内存保护不足
隔离机制薄弱

高带宽内存

HBM依赖先进封装与接口
HBM 接口 IP 待发展

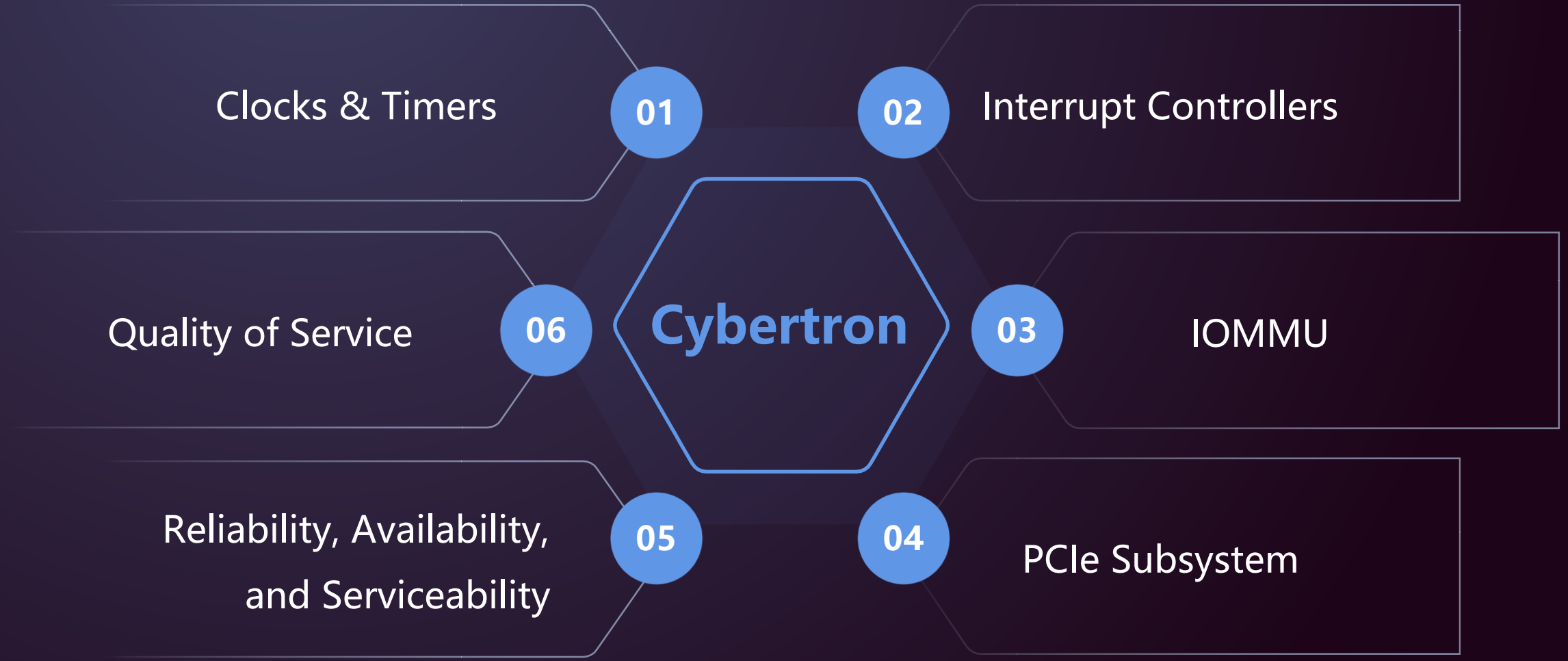
生态碎片化与软件短板

操作系统与中间件支持薄弱
工具链成熟度滞后
第三方IP兼容性风险
碎片化风险加剧

Part.3 时擎科技Cybertron高性能RISC-V SOC平台技术介绍

Outline of Cybertron RISC-V SoC Platform

TiMESiNTELLi 时擎科技



Clocks & Timers

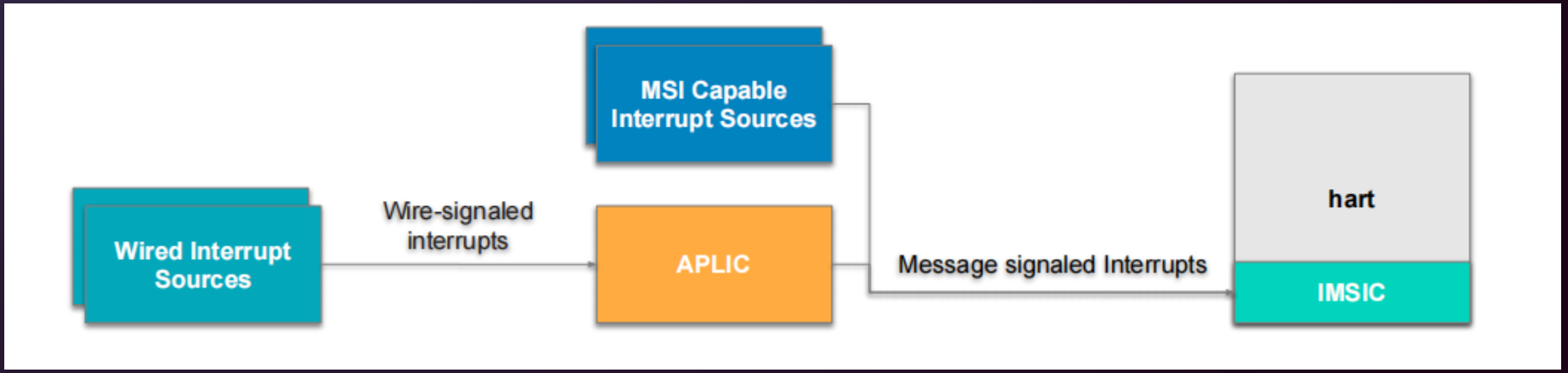
Standardize the unit of time as 1ns.

The frequency at which the CSR provides an updated time value is at least 100MHz.

The time counter appears to be always on and appears to not lose its count across hart low power idle states, including when the hart is powered off.

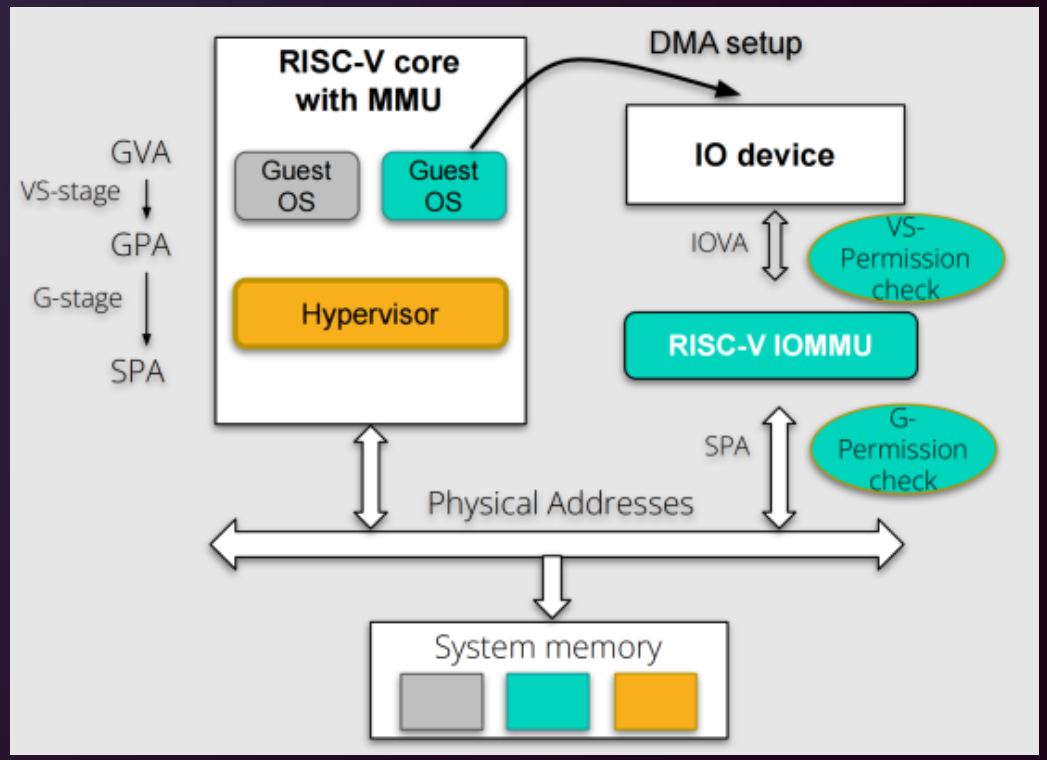
Interrupt Controllers

- Support RISC-V Advanced Interrupt Architecture (AIA)
- Message signaled interrupt (MSI/MSI-X) to the hart
- The Incoming Message-signaled Interrupt Controller (IMSIC) implement an interrupt file for S-mode.
- The IMSIC support at least 5 VS-mode interrupt files
- The S-mode interrupt file support at least 255 interrupt identities
- The VS-mode interrupt files support at least 63 interrupt identities
- IMSIC interrupt files memory regions: not cacheable, non-idempotent, coherent, strongly-ordered (I/O ordering) channel 0 I/O region

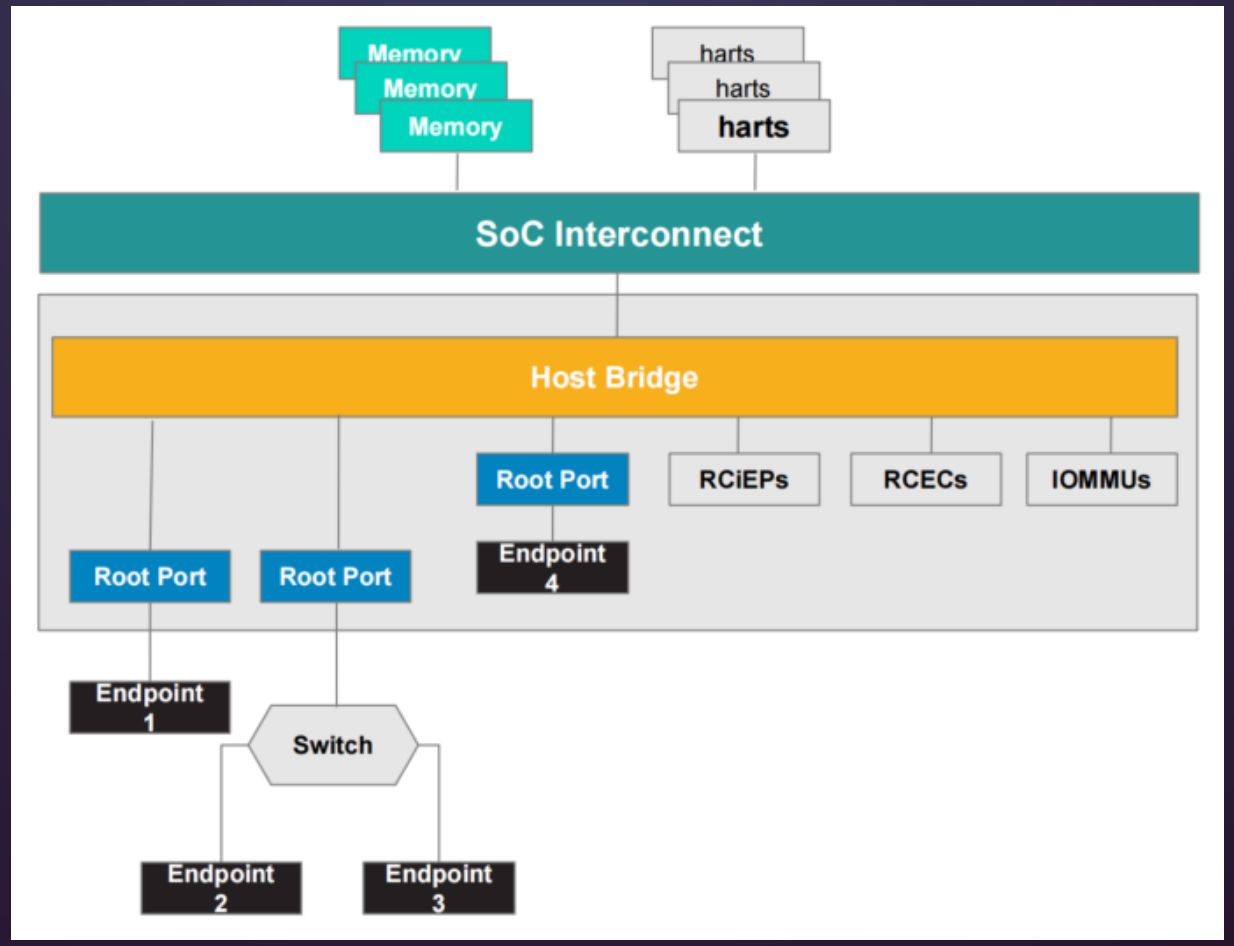


Input-Output Memory Management Unit (IOMMU)

- IOMMU provide
  - Memory protection
  - Virtual address translation
  - Virtual address space sharing between devices and CPU
  - Interrupt remapping and virtualization
- All DMA capable peripherals (RCIeP, non-PCIe, PCIe Root Ports) to be governed by a RISC-V IOMMU
- Support at least 16-bit wide Device IDs
- Support at least 20-bit wide PASID, if support PCIe PASID capability
- Support for PCIe Address Translation Services (ATS)
- Support for hardware performance monitor (HPM)
- The host bridge enforce the physical memory attribute checks and physical memory protection checks on memory accesses originated by the IOMMU and signal detected accessviolations to the IOMMU



PCIe Subsystem



- Root complex is a collection of root ports, root complex event collectors (RCEC), and root complex integrated endpoints (RCiEP)
- Root complex uses a host bridge to connect to the CPU and system memory
- Devices may be integrated into the SoC as either RCiEP or as an EP connected to a PCIe root port (endpoint 4 in this example)
- One or more IOMMUs used to provide address translation and protection

PCIe Subsystem

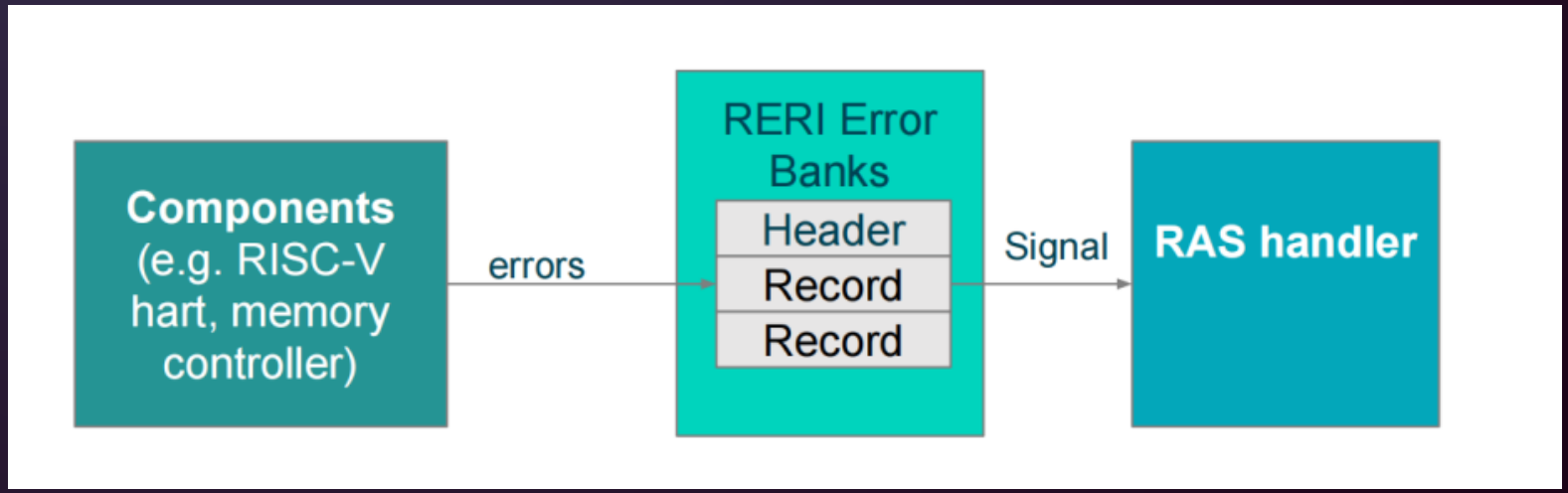
- Access configuration registers using Enhanced Configuration Access Method (ECAM)
- Memory Space
  - Support all aligned and unaligned access sizes
  - Support atomics, instruction fetch, and page walks
- Support access control services (ACS)
  - ACS source validation
  - ACS translation blocking
  - ACS I/O request blocking
- For P2P support
  - ACS P2P request redirect
  - ACS P2P completion redirect
  - ACS upstream forwarding
  - ACS direct translated P2P
  - ACS P2P egress control
- Address Routed Transactions
  - Support IOMMU translations for addresses
  - Enforce physical memory attribute checks and physical memory protection checks

PCIe Subsystem

- ID Routed Transactions
  - Support P2P routing of PCIe VDM between root ports within or across hierarchies
- Cacheability and Coherence
  - Enforce PCIe memory ordering rules
  - Support the relaxed ordering (RO) and ID-based ordering (IDO)
  - The host bridge implements hardware enforced cache coherency
  - TLP processing hints (TPH) for future extension
- Support message signaled interrupts(MSI or MSI-X)
- Error/Event Reporting
  - Support advanced error reporting (AER)
  - Support the downstream port containment (DPC)
  - Support the RP PIO controls
- Support vendor specific registers
- SoC-integrated PCIe devices Implements all software visible rules

Reliability, Availability, and Serviceability

- **Reliability** - probability that a system provides correct service
- **Availability** - measure of tolerance of errors
- **Serviceability** - time to restore the service to correct operation
- Level of RAS depends on the reliability goals of the product - typically measured using metrics such as failure-in-time (FIT) and defects-per-million (DPM)
- Error detection and correction mechanisms for caches, system memory, and interconnects
- Periodic scrubbing of system memory for errors
- Error containment techniques such as data poisoning and rules for handling poisoned data
- Support the RISC-V RAS error record register interface (RERI) for error logging and signaling



Quality Of Service

- **Modern SoCs feature tens of CPU cores, multiple levels of shared caches, shared interconnects, shared memory controllers**
  - Noisy neighbor problem - leads to non-deterministic workload performance
    - Interference among co-located tasks due to shared resource contention
  - Underutilization problem
    - Difficulty in consolidating latency critical applications without compromising service level objectives (SLO)
- **Contention for shared cache capacity (e.g., LLC) or memory bandwidth significant source of interference**
  - Allocating dedicated capacity and bandwidth to workloads based on their SLO helps address these problems
- **RISC-V Ssqosid and RISC-V Capacity and Bandwidth QoS Register interface (CBQRI) specifications provide**
  - Methods to associate IDs for capacity/bandwidth allocation with workloads
  - Configure capacity and bandwidth allocations in resource controllers
- **Server SoC specification provides guidelines on**
  - Mitigating unwarranted perf. interference by resource contention through capacity/bandwidth allocation mechanisms
  - Integrating Ssqosid and RISC-V CBQRI extensions in the SoC
    - Significant caches and memory controllers, IOMMU
  - Minimum number of resource control IDs and monitoring counter IDs

Manageability

- **Guidelines for the RISC-V server SoC to incorporate a standardized set of protocols and standards for server management**
  - Monitoring of sensors (temperature, power, etc.)
  - Parameter control (power limits, etc.)
  - Logging (RAS error records, etc.)
- **Using standards such as**
  - DMTF Redfish
  - Platform level data model (PLDM)
  - Management Component Transport Protocol (MCTP)
- **Guidelines on securing the management interface through use of standard protocols such as DMTF Security Protocol and Data Model (SPDM) for attestation and message encryption.**
- **Guidelines on hardware interfaces for in-band and out-of-band management**
  - PCIe to BMC to facilitating uses such as remote KVM and management network
  - I2C IPMI SSIF
  - UART

Performance Monitoring

Incorporating hardware performance monitors

- Significant caches
- Interconnects
- PCIe root ports

Support collection of commonly used performance metrics such as bandwidth and latency， to help guide workload placement and tuning.

If the SoC supports NUMA configurations，support filtering the counting based on whether the request is tolocal memory or to remote memory.

Security

- Implement a hardware RoT as the primary root of trust.
- Support PCIe Integrity and Data Encryption (IDE) capability.
- Support encryption of off-chip DRAM using a transient memory encryption key that has at least 256-bit key lengths.
- The cryptographic modules used to implement PCIe and off-chip DRAM encryption complies with security requirements specified by relevant security standards from national standards laboratories.
- Have the capability of interfacing with a Trusted Platform Module (TPM) that adheres to the TPM 2.0 Library specification

Part.4 高性能RISC-V SOC 未来展望

未来演进方向

AI算力深度融合：从“通用计算”到“异构智能”

原生AI指令集进一步扩展

DSA（领域专用架构）定制化

虚拟化与高可靠架构：支撑复杂系统级应用

硬件级虚拟化与安全隔离

功能安全与实时性增强

模块化与异构集成

Chiplet 异构集成

一致性互联协议标准化

HBM集成

生态标准化：从碎片化到统一兼容

RVA Profile成为生态基线

全栈开源工具链成熟

产业落地展望

TiMESiNTELLi 时擎科技

≒边缘服务器（ARM替代路径）

≒高性能AI训练芯片（对标NVIDIA Grace）

≒智能汽车域控制器（网关/座舱）

≒通用服务器CPU市场份额突破

≒数据中心协处理器（DPU/IPU）

≒RISC-V主导的泛在计算生态

TiMESiNTELLi
时擎科技

THANK YOU

感谢在座各位的聆听