华为鲲鹏 920 处理器

技术白皮书

文档版本	01
发布日期	2020-02-10

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： https://e.huawei.com

前言

概述

华为鲲鹏920处理器为面向ICT领域的ARM v8指令集64bit多核处理器芯片。本文档简单介绍华为鲲鹏920处理器系列芯片的架构和功能特性，以及芯片的各个子系统和组件之间的交互关系。

读者对象

本文档主要适用于以下工程师：

- 产品架构师
- 产品系统方案设计人员

符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

符号	说明
危险	表示如不避免则将会导致死亡或严重伤害的具有高等级风险的危害。
警告	表示如不避免则可能导致死亡或严重伤害的具有中等级风险的危害。
注意	表示如不避免则可能导致轻微或中度伤害的具有低等级风险的危害。
须知	用于传递设备或环境安全警示信息。如不避免则可能会导致设备损坏、数据丢失、设备性能降低或其它不可预知的结果。 “须知”不涉及人身伤害。
说明	对正文中重点信息的补充说明。 “说明”不是安全警示信息，不涉及人身、设备及环境伤害信息。

修改记录

文档版本	发布日期	修改说明
01	2020-02-10	第一次正式发布。

目 录

1 概述

本文档中，华为鲲鹏920系列涵盖表1-1列出的所有芯片型号，相关的重要概念如表1-2所示。

表 1-1 华为鲲鹏 920 系列芯片概览

芯片型号	计算能力	内存支持	网络能力	存储能力	PCIe接口	平台特性	加速
华为鲲鹏920 7265/7260/5255/5250	48核/64核 Armv8.2架构； 单核支持512KB L2 Cache； 单核支持1MB L3 Cache。	8个DDR控制器	2*100G； 4*25GE； 2*50G； 支持RoCEv2和SR-IOV。	2-port AHCI接口SATA控制器； x8 SAS 3.0控制器，支持STP协议。	40个PCIe 4.0通道； 多达20个根端口； 支持x16接口； 支持Peer2Peer和ATS； 支持CCIX。	最高支持4颗芯片互联； 内置引擎； 片内外设采用PCI拓扑结构。	压缩/解压缩引擎； 安全算法引擎； RSA算法引擎。
华为鲲鹏920 5220/3210	24核/32核 Armv8.2架构； 单核支持512KB L2 Cache； 单核支持1MB L3 Cache。	4个DDR控制器	2*100G； 4*25GE； 2*50G； 支持RoCEv2和SR-IOV。	2-port AHCI接口SATA控制器； x8 SAS 3.0控制器，支持STP协议。	40个PCIe 4.0通道； 多达20个根端口； 支持x16接口； 支持Peer2Peer和ATS； 支持CCIX。	内置引擎； 片内外设采用PCI拓扑结构。	压缩/解压缩引擎； 安全算法引擎； RSA算法引擎。

表 1-2 华为鲲鹏 920 处理器系列芯片的重要概念

术语	定义
Chip	华为鲲鹏920处理器系列芯片集成了多个CPU核和其他架构组件。 一颗芯片可以包含一个或多个SCCL，以及一个或多个SICL。 多颗芯片可以通过Hydra接口实现片间互联，组成具有Cache一致性的多片系统。
Chip ID	在通过Hydra接口实现的多片系统中，Chip ID指单颗芯片的ID。对芯片进行编号，便于地址译码及设备区分。
CCL	Core Cluster（CCL），即内核集群。每个CCL包含4个Arm内核（及各自的L1 Cache），为华为鲲鹏920提供专用的L2 Cache。
Cluster	华为鲲鹏920将处理器内核和相关设备组成共享公共资源或逻辑关联紧密的集群组件。
Core	当物理上不支持多线程能力时，单处理器内核对软件可见。
DAW	Dynamic Address Window（DAW），即动态地址窗口，是芯片内部的地址译码机制。通过配置动态地址窗口参数，可以修改分配给内存控制器和设备的起始地址和空间大小。
Device	设备指除内核外的物理单元，如I/O集群、片内加速设备、管理设备等。
DeviceID	DeviceID是芯片分配给每个设备的唯一标识号。 在中断系统中，DeviceID用于中断上下文索引及区分各个上报中断设备。
Dispatch	Dispatch是片上总线的组件之一，用于对物理地址进行最后一级译码，提供各设备地址空间的访问通道。
HA	Home Agent是华为鲲鹏920 Hydra协议中定义的组件之一。在多片系统中，Home Agent用于保持L3 Cache之间的Cache一致性。
Hydra接口	Hydra接口是实现片间互联的物理接口。用于多芯片扩展，满足Cache一致性和统一地址空间的要求。
Hydra	Hydra协议和Hydra接口。 Hydra协议是华为定义的片间互联Cache一致性协议。Hydra接口是遵循Hydra协议的高带宽、低时延物理接口。
ICL	I/O Cluster（ICL），即I/O集群，由物理上距离相近且共享系统总线接口和内部接口的多个设备组成。
IMU	Intelligent Management Unit（IMU）即智能管理单元，对华为鲲鹏920处理器进行管理和监督。IMU具有完整的SoC组件，完全独立于应用处理器系统。

术语	定义
MTPT	存储转换及保护表实现虚拟地址（Virtual Address，VA）到物理地址（Physical Address，PA）的转换，控制远程直接数据存取（Remote Direct Memory Access，RDMA）协议栈处理过程中的访问权限。 虚拟地址是的到物理地址的转换由RoCE引擎完成，不由系统内存管理单元（System Memory Management Unit，SMMU）完成。
Outstanding	在前一次总线读写完成之前，并行发起多个新的总线读写请求，从而提高总线访问内存的性能。Outstanding值越大，总线带宽越高。
PCIe EP模式	华为鲲鹏920的PCIe控制器可工作于Endpoint（EP）模式。在该模式下，PCIe控制器作为标准的EP设备连接到外部CPU。此时，PCIe控制器具备完整的EP配置空间和中断机制。华为鲲鹏920的所有功能都可以用作单个PCIe EP设备的扩展功能。华为鲲鹏920上可以运行不同的应用，为EP设备提供不同的功能。
SAW	Static Address Window（SAW），即静态地址窗口，是华为鲲鹏920内部的地址译码机制。用物理地址中的一些固定比特，对位于ICL中具有特定大小的物理地址空间的设备进行译码。
Scheduler	Scheduler是片上总线的组件之一。Scheduler用于调度设备对系统物理地址空间的DMA访问。
SCCL	Super Core Cluster（SCCL），即超级内核集群，由物理上距离相近且共享其他资源的多个集群组成。根据关联关系的不同，一个SCCL还可能包含L3 Cache、多个DDR控制器和一个I/O集群。
SICL	Super I/O cluster （SICL），即超级I/O集群，由3个物理上接近的ICL和1个Hydra接口模块组成。SICL还提供I/O接口，及加速和平台管理功能。根据关联关系的不同，一个SICL还可能包含一个内核集群。由于SICL负责提供PCIe和Hydra接口功能，因此华为鲲鹏920系列的所有芯片都必须包含SICL。
SCL	Super Cluster（SCL）即超级集群，包括SCCL和SICL。
SCL ID	对芯片内的SCL进行编号，便于地址译码及区分设备。

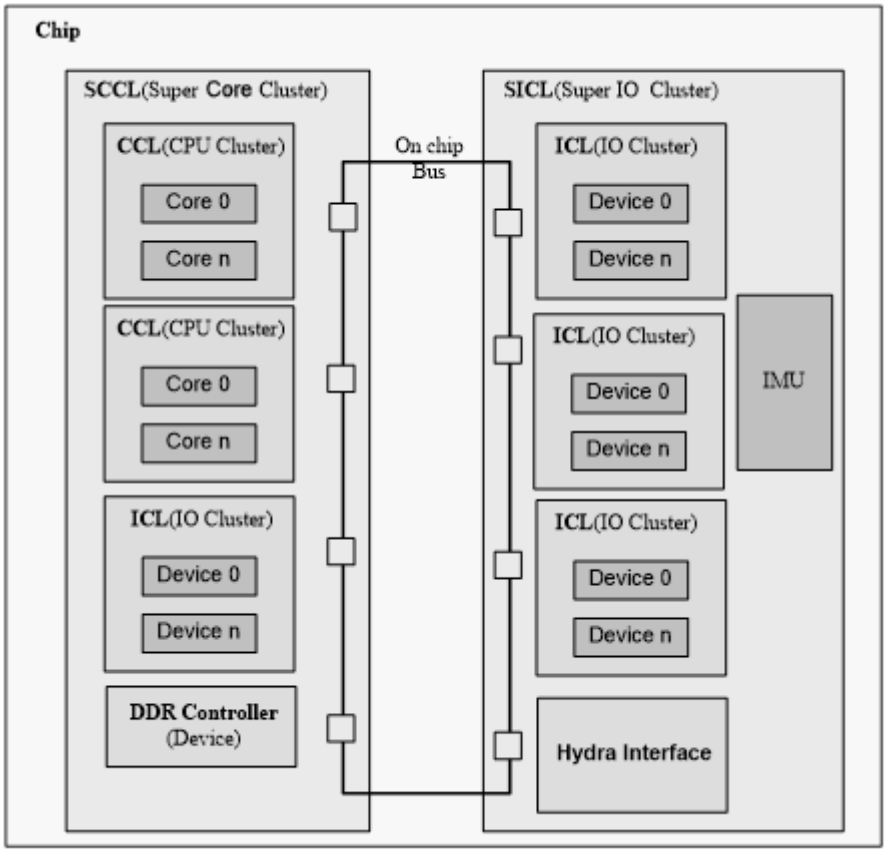
2 芯片架构

2.1 重要概念

本节对本文档中一些核心概念进行约定。

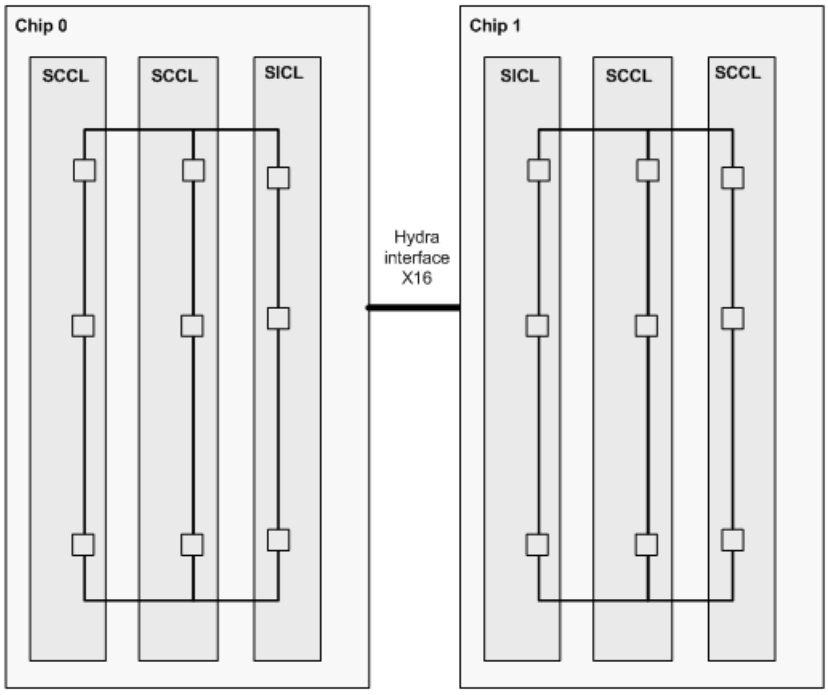
芯片内部功能模块按物理亲和性或逻辑亲和性划分为CCL和ICL。华为鲲鹏920处理器系列芯片的组成单元如**图2-1**所示。

图 2-1 华为鲲鹏 920 处理器系列芯片物理架构图



2P服务器处理器多片互联方案示例如图2-2所示。每颗芯片各提供2个SCCL和1个SICL。芯片之间通过片间Cache一致性接口连接，片间带宽高达480Gbps，即X16 Hydra接口。

图 2-2 标准 2P 服务器处理器多片互联方案示例



4P服务器处理器多片互联方案示例如图2-3所示。每颗芯片各提供2个SCCL和1个SICL。芯片之间通过片间缓存一致接口连接，片间带宽高达240Gbps，即X8 Hydra接口。此外，由于没有额外的Hydra接口，4P互联系统无法支持I/O（PCIe）扩展。

图 2-3 标准 4P 服务器处理器多片互联方案示例

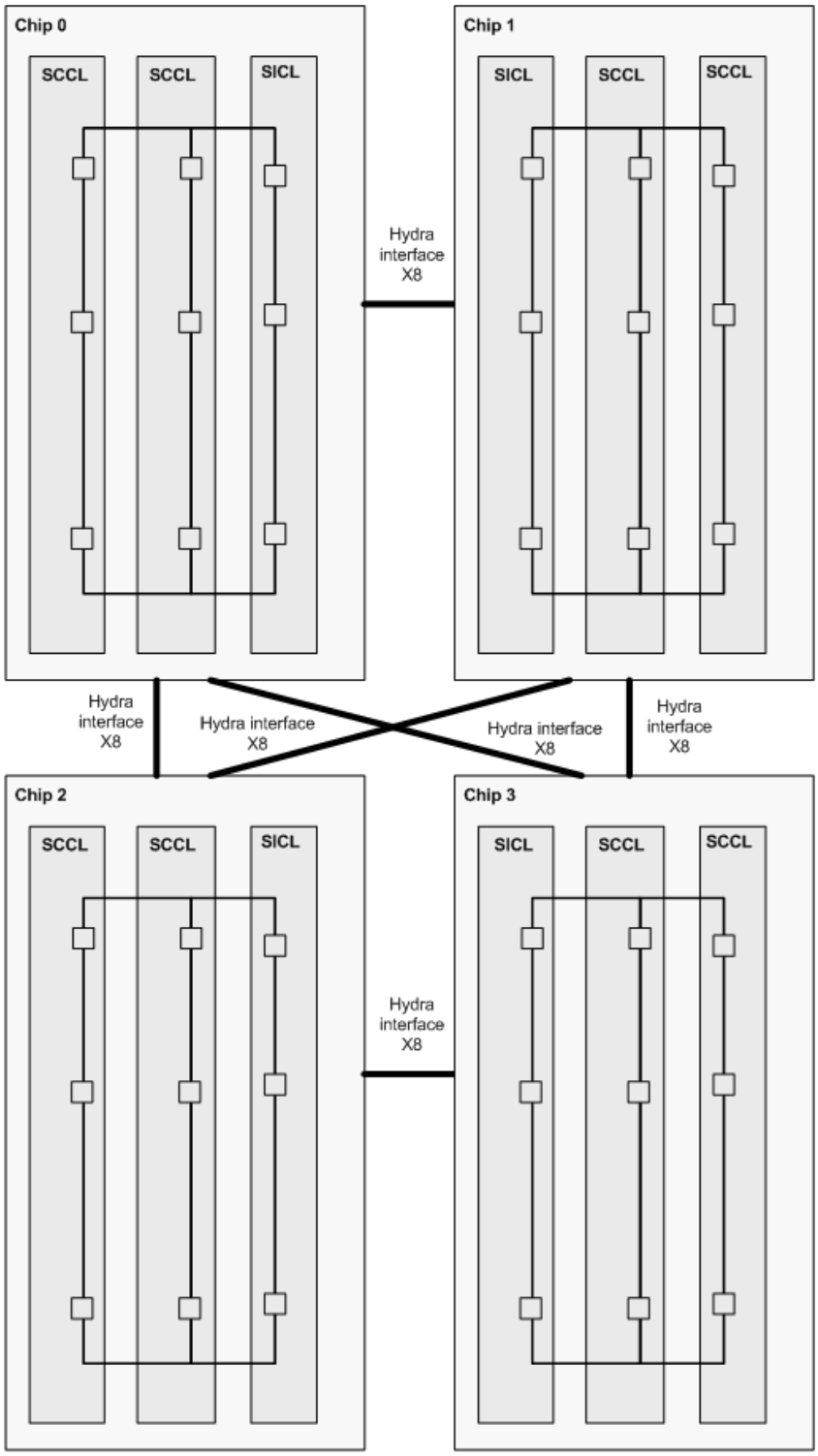


图2-4所示服务器处理器多片系统实现了PCIe扩展。本系统中有2个处理器芯片和1个主要用于PCIe接口扩展的I/O扩展的I/O桥片。3颗芯片通过Hydra接口连接（本场景中Hydra接口仅用于CPU和I/O桥的连接）。本方案多用于对PCIe插槽有扩展需求的2P服务器场景。此外，I/O桥通过地址配置来区分处理器的DDR空间。

由于华为鲲鹏920的7265/7260/5255/5250/5245/5240/5235/5230只能为Hydra接口提供24-lane的SerDes，因此4P（即，有4个CPU槽位）多芯片系统无法支持I/O扩展功能。

图 2-4 带 PCIe 扩展的服务器处理器多片系统架构示意图

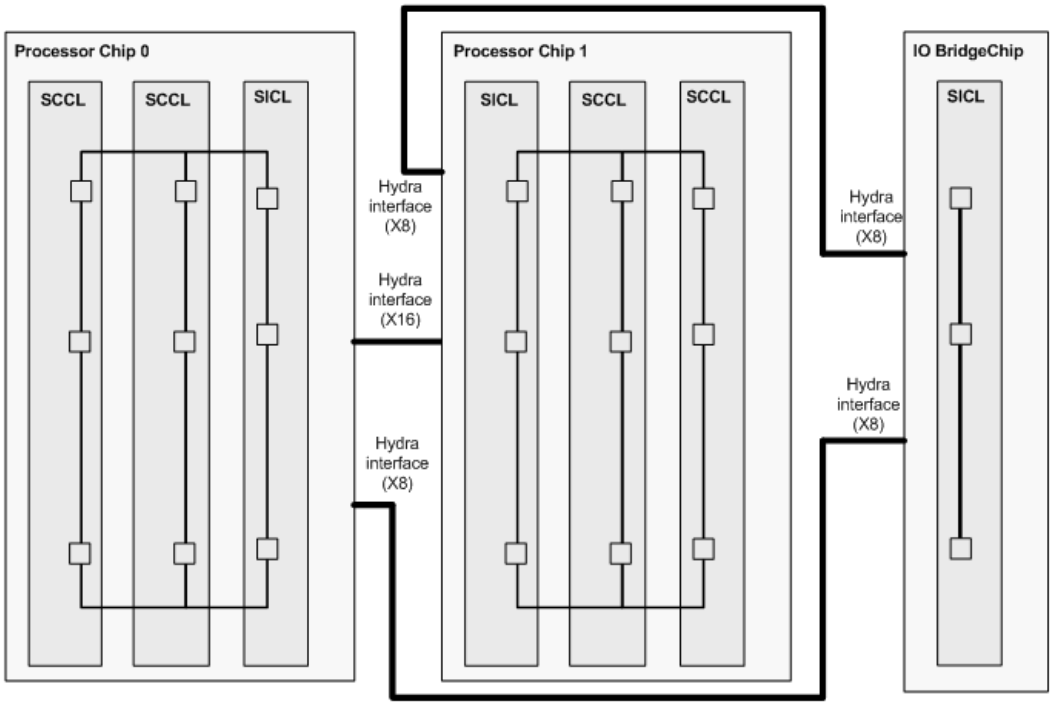
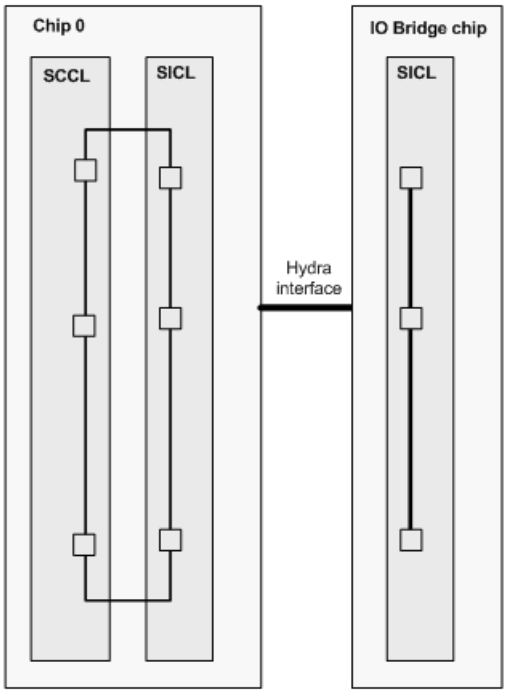


图2-5提供了另一种多片互联方案。本系统有两颗芯片，一颗是华为鲲鹏9205220/3210芯片（包含一个SCCL和一个SICL），另一颗是华为鲲鹏920的7265/7260/5255/5250/5245/5240/5235/5230 I/O桥片（用于I/O扩展），两颗芯片之间通过Hydra接口连接（在本方案中，Hydra接口只用于CPU和I/O桥连接）。本方案多用于存储系统。如果没有PCIe插槽扩展需求，I/O桥片可以省去。

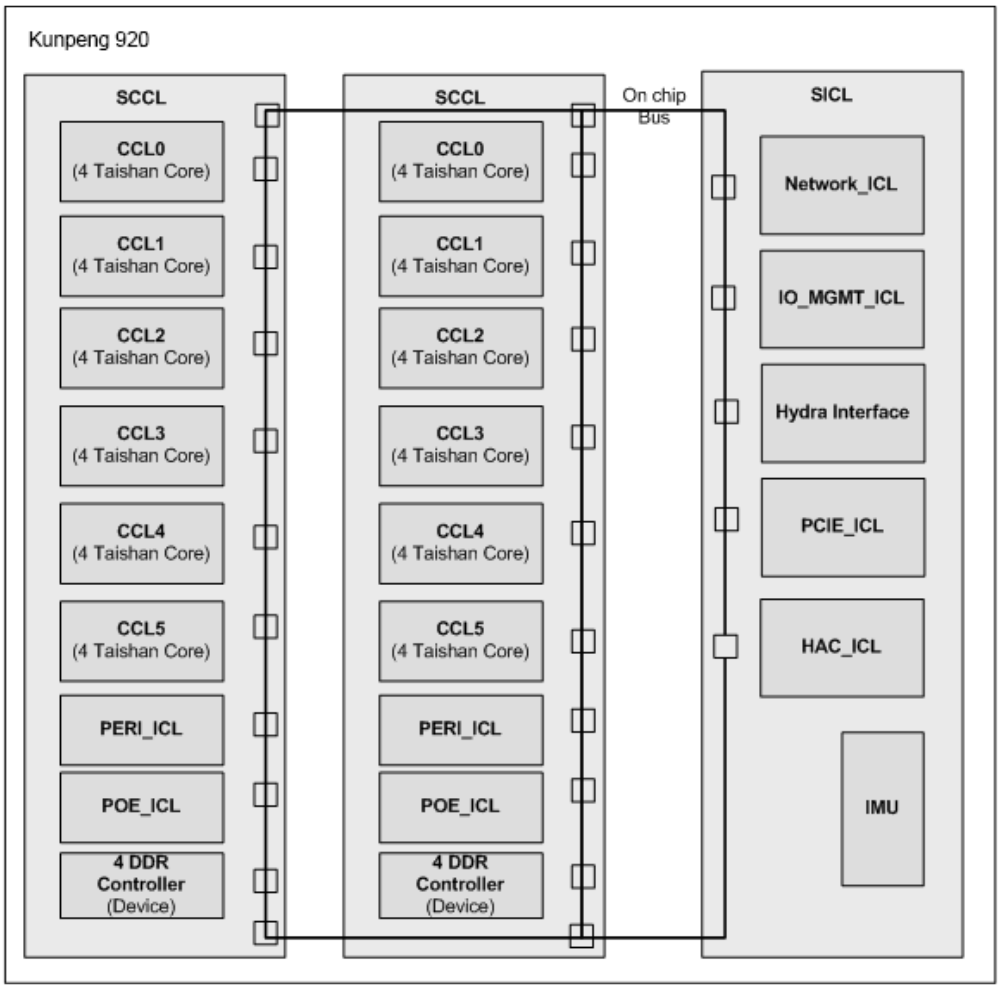
图 2-5 存储场景多片互联方案示例



2.2 芯片组件

华为鲲鹏920 7265/7260/5255/5250/5245/5240/5235/5230芯片组件示意图如图2-6所示。

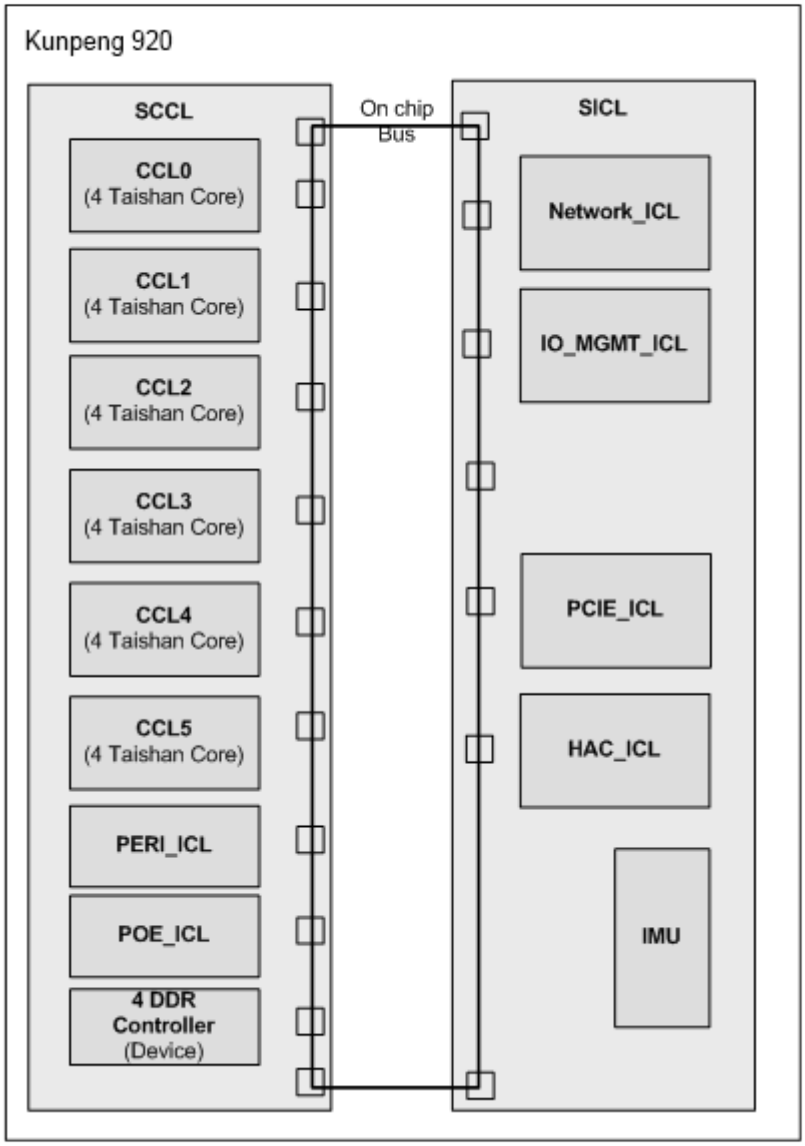
图 2-6 华为鲲鹏 920 7265/7260/5255/5250/5245/5240/5235/5230 组件示意图



由于Hydra以及其他高速I/O接口（如PCIe、GE和XGE）都集成在SICL中，因此每个处理器都包括一个SICL。

华为鲲鹏920 5220/3210芯片组件示意图如图2-7所示。

图 2-7 华为鲲鹏 920 5220/3210 芯片组件示意图



2.3 组件单元

2.3.1 片上总线

Cache一致性总线连接华为鲲鹏920处理器芯片内部的各个组件，为每个内核、设备、集群、以及其他可寻址组件提供对系统内存地址空间的一致访问。（可通过华为自研Cache一致性片间总线Hydra接口级联多颗芯片来实现SMP系统扩展）

通过Cache一致性总线连接支持不同功能或不同版本的集群，从而组合成为不同型号的芯片。

各个内核、设备和集群通过该片上总线来访问内存和其他设备寄存器中的数据。（详见章节**4 内存子系统**）。各个设备通过该片上总线向处理器内核发送中断信号。

2.3.2 CCL

华为鲲鹏920的每个内核集群（CCL）都由4个内核和专用L2 Cache组成。

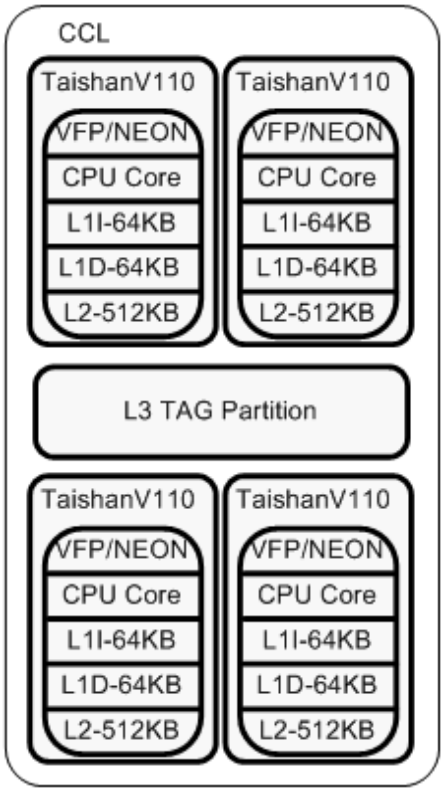
华为鲲鹏920兼容Armv8.2-A架构平台的所有特性。

每个华为鲲鹏920内核都有一个专用L2 Cache。华为鲲鹏920的CCL之间支持完全一致性。总线上的其他功能单元可以一致性地访问每个CCL的缓存中的最新数据。

CCL内部共用一个系统总线接口。

CCL组件如图2-8所示。

图 2-8 华为鲲鹏 920 的 CCL 组件示意图



2.3.3 ICL

I/O集群（ICL）指物理上距离相近且共享公共资源（如共用系统总线接口和内部接口）或逻辑关联紧密的设备组。一个典型的ICL包括以下组成部分：

- 多个设备（如图2-9所示）；
- 0个或1个SMMU（System Memory Management System），为设备提供地址转换和访问保护功能；
- 1个系统总线接口；
- 1个Sysctrl或Subctrl，用于固件初始化和公共配置；
- 1个Dispatch，为访问设备寄存器空间提供物理地址（PA）译码；
- 0个或多个Scheduler，当设备数量较多时，Scheduler可以合并各设备的内存访问流量。

不是所有的ICL都支持DMA能力。但是所有设备都有一个可通过编程访问的寄存器。

ICL的内部结构如图2-9所示。

支持DMA能力的设备主动对系统地址空间发起读写操作，Scheduler将请求进行合并。该请求由SMMU进行处理，通过物理地址（PA）访问总线。

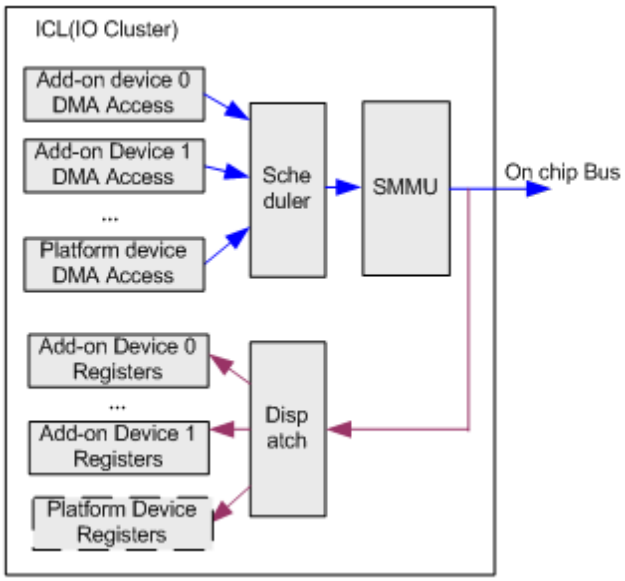
从设备经由Scheduler再到SMMU的路径上的地址，可以是虚拟地址（VA）、中间物理地址(IPA)或物理地址（PA）。如果DMA引擎上配置的内存访问地址不是物理地址（PA），设备通过StreamID识别合适的地址空间。SMMU将地址转换为最终的物理地址（PA）。

StreamID和DeviceID的配置和生成方法，请参见5.5 StreamID与DeviceID对应关系。

用来访问设备寄存器空间的地址是从片上总线经由Dispatch传送给设备的物理地址（PA），如图2-9所示。

各类设备的行为，请参见5.1 设备分类。

图 2-9 ICL 内部结构



2.3.4 SCCL

华为鲲鹏920处理器提供超级内核集群（SCCL）概念。SCCL由物理上距离相近且共享其他资源的多个集群组成。华为鲲鹏920的每个SCCL包括6个CCL、2个ICL和4个DDR控制器。DDR控制器也可以看做一个设备。SCCL内部结构如图2-10所示。

L3 Cache在物理上分为两部分：L3 Cache标签和L3 Cache数据。为了降低snoop时延，CCL集成了L3 Cache标签。L3数据块直接通过片上总线传输。

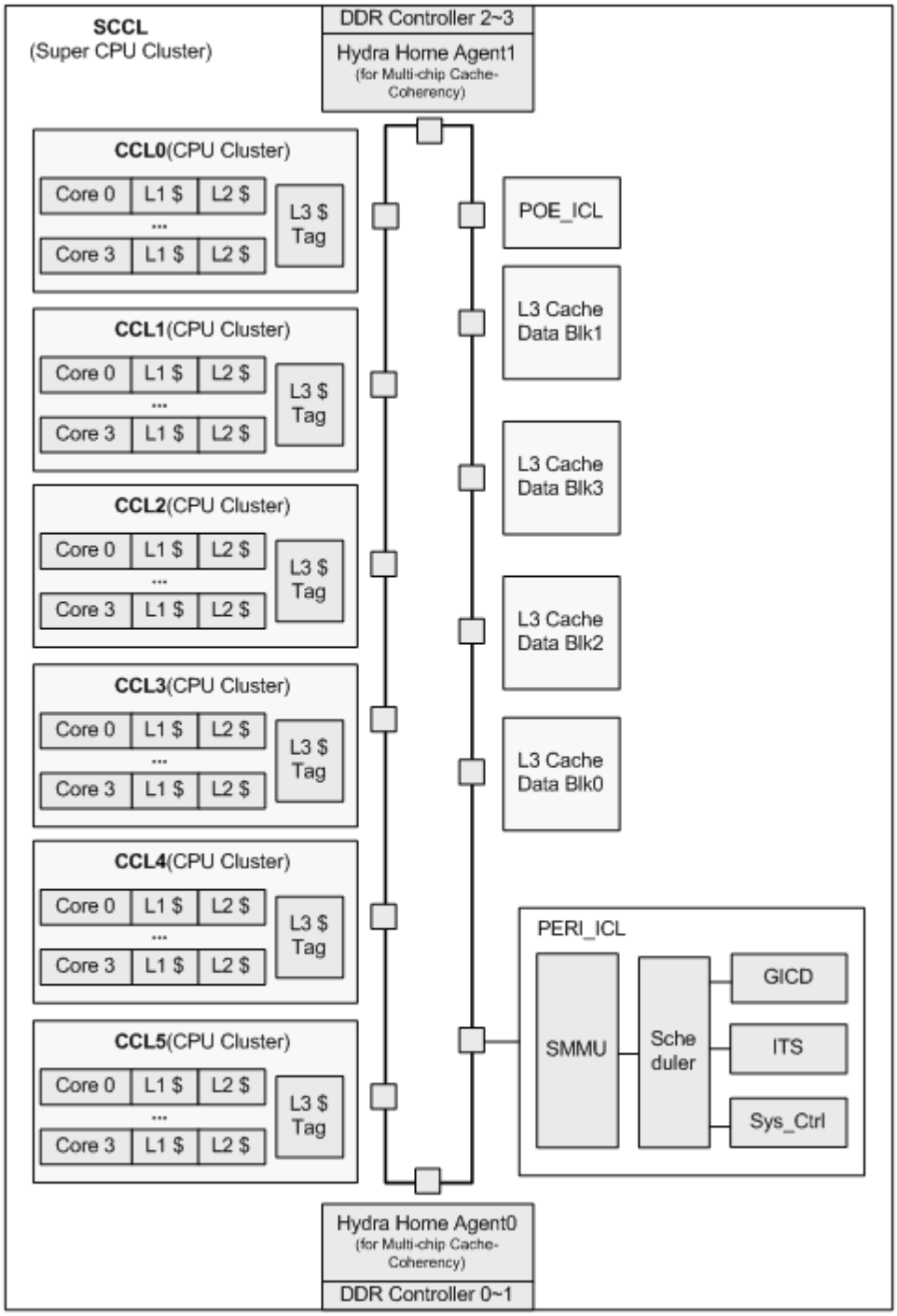
每个SCCL在物理上都具有一个通用中断分发器（Generic Interrupt Controller Distributor，GICD）模块，兼容GICv4规范。当单片或多片系统中有多个SCCL时，只有一个GICD对系统软件可见。

Hydra Home Agent负责处理多片系统的Cache一致性协议。

POE_ICL是系统中的硬件加速器。使用场景如下：

- 报文保序；
- 消息排队；
- 分发消息、任务到指定内核。

图 2-10 SCCL 内部结构

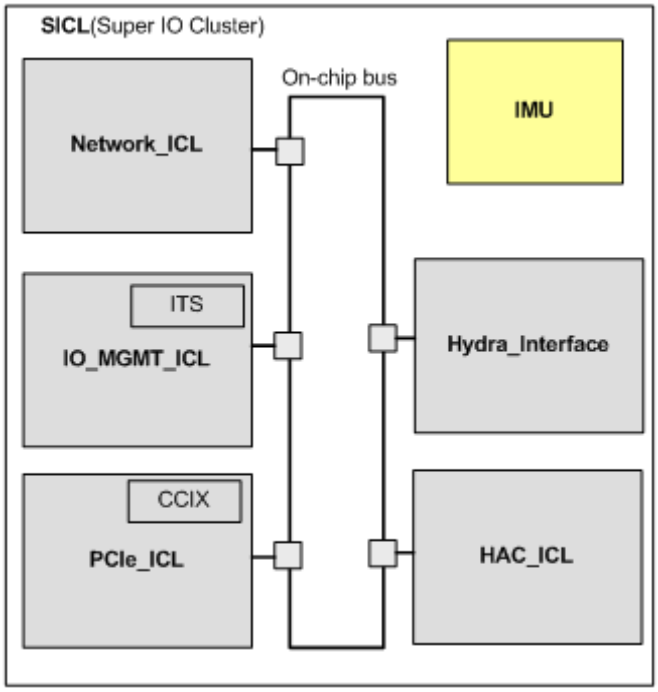


2.3.5 SICL

此外，华为鲲鹏920处理器系列芯片还提供超级I/O集群（SICL）概念。每个SICL由4个ICL、1个Hydra接口和1个独立的IMU组成。每个SICL也可能包含一个内核集群。

其一级内部结构如图2-11所示。

图 2-11 SICL 内部结构



SICL提供SerDes。PCIe、NIC、存储接口控制器和Hydra接口均要用到SerDes功能，它们对应的ICL也包含在SICL中。

Network ICL提供片内以太网络接口控制（Network Interface Control，NIC），包括100G、50G、40G、25G、10G、1G速率接口，支持RoCE和RoCEv2功能。

PCIe_ICL提供根复合体（Root Complex，RC）功能。支持PCIe 4.0规范，最大根端口号为20。PCIe_ICL还支持Cache一致性互联协议（Cache Coherent Interconnect for X，CCIX）。

Hydra接口是保证片间互联Cache一致性的高带宽接口。最多支持4个处理器芯片互联。

HAC_ICL和IO_MGMT_ICL是存储Host控制器、USB Host控制器、硬件加速器和各种I/O的组合。均可在系统设备树中找到。

IMU是芯片的管理单元，完全独立于芯片的其他部分。

说明

在多片互联场景下，SMP系统中存在多个SICL。

1) PCIe、网络、HAC和IO_MGMT的I/O集群如同示例副本一样并行工作。

2) 多片互联场景下，只有Chip 0的IMU对Cache一致性SMP系统可见，其他芯片的IMU对Cache一致性SMP系统不可见。单片系统和多片系统使用不同的IMU配置。

3 CPU 核

华为鲲鹏920处理器集成了Arm TaiShan处理器内核。该处理器基于Armv8.2-A架构平台，满足高性能、低功耗需求，兼容Armv8-A平台所有特性，支持Armv8.1和Armv8.2扩展。

华为鲲鹏920处理器芯片系列各型号的内核配置如表3-1所示。内核与其他组件的逻辑关系请参见2 芯片架构。

表 3-1 各芯片型号的内核配置

芯片型号	SCCL数量	单个SCCL中CCL数量	单个CCL中内核数量	内核总数	Arm架构版本
华为鲲鹏920 3210	1	6	4	24	Armv8.2
华为鲲鹏920 5220	1	8	4	32	Armv8.2
华为鲲鹏920 5255/5250	2	6	4	48	Armv8.2

CCL的配置细节，请参见2.3.2 CCL。

4 内存子系统

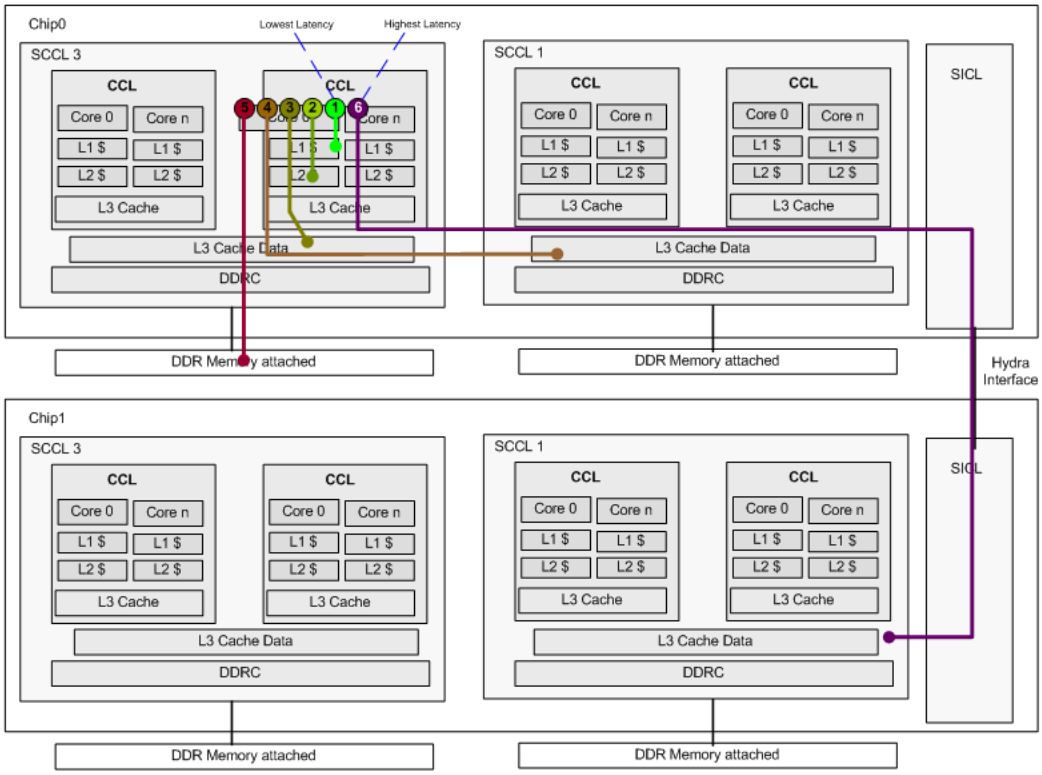
本节介绍单片或多片系统中物理地址空间涉及的内存行为。

当某个CCL或ICL访问物理内存空间时，内存空间在页表、内存属性、订单行为等方面的配置和行为均遵循Armv8架构。CCL访问的内存空间的属性由MMU（ memory management unit，内存管理单元 ）中的页表控制。ICL访问的内存空间的属性由SMMU中的页表或源设备控制。SMMU中的页表可以通过硬件自动同步到CCL中的MMU中，也可以单独配置。

华为鲲鹏920系列处理器支持跨芯片的硬件Cache一致性，整个系统相当于一个SMP系统。系统的内存行为和内存序模型遵循Arm架构。由于物理限制，内存访问延迟受数据所在位置（ 例如不同级别的缓存中或DDR内存中 ）影响。如果目标数据位置在物理上接近内存访问发起者，则时延较低。

物理内存的层次关系如图4-1所示。对典型内存访问路径从1到6进行顺序编号，访问时延从低到高。也就是说，路径1的时延最低，路径6的时延最高。

图 4-1 华为鲲鹏 920 内存层次关系



说明

- 华为鲲鹏920的内核不共享L2 Cache，每个内核都有一个单独的L2 Cache。
- 对L3 Cache按照CCL进行区域划分，每个区域靠近一个CCL，以降低L3 Cache访问时延。换句话说，本地SCCL中的所有内核共享一个L3 Cache，但每个内核的访问时延可能有细微差异。

5 设备及设备拓扑

5.1 设备分类

华为鲲鹏920采用SoC芯片架构，除了内核和内存子系统外，芯片还集成了大量设备。这些片内设备分为三类：

- 固件设备

  用于配置芯片的功能模式及启动流程，如DDR初始化、系统地址译码、SerDes初始化等。

  这些设备在系统中有固定的内存映射地址，还可能已经预先分配了SPI中断。对应的寄存器不能由操作系统或驱动程序直接访问，而只能由BIOS或运行时UEFI进行读写。

  固件设备的工作不依赖于平台设备或外设。

- 平台设备

  作为一款基于Arm的服务器处理器，华为鲲鹏920遵循Arm生态系统的各项规范，包括SBSA规范和RAS相关规范。平台设备包括SMMU、通用中断控制器（Generic Interrupt Controller，GIC）、UART、Watchdog等，是平台的必要组成部分。平台设备提供统一的可编程寄存器。

  所有平台设备都应该采用标准驱动程序，以保证跨版本和跨厂商兼容性。

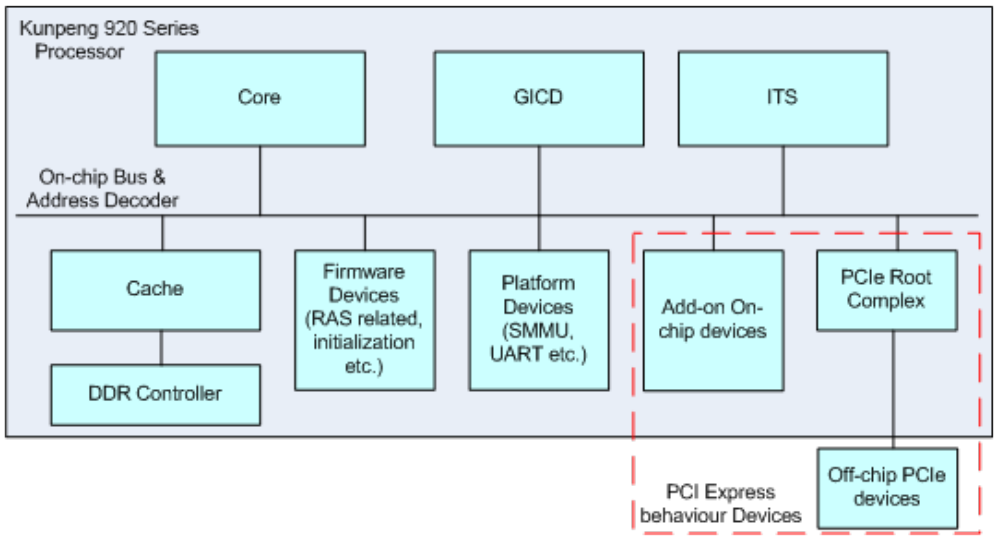
- 片内外设

  芯片内部的部分功能可以用类似功能替代，如NIC、USB控制器等，这些功能统称为“外设”。这些外设使用的驱动程序由专门的厂商提供。

  从软件角度来看，大多数片内外设都是标准PCI设备。标准PCI设备包括如下行为：

- 可被固件或操作系统枚举，如片外PCI设备。
- 可以在PCIe框架中分配内存地址和中断。
- 采用与PCI设备相同的功能复位和电源管理机制。
- 虚拟化特性完全遵循PCIe SR-IOV规范。

图 5-1 华为鲲鹏 920 设备分类



5.2 叠加设备拓扑

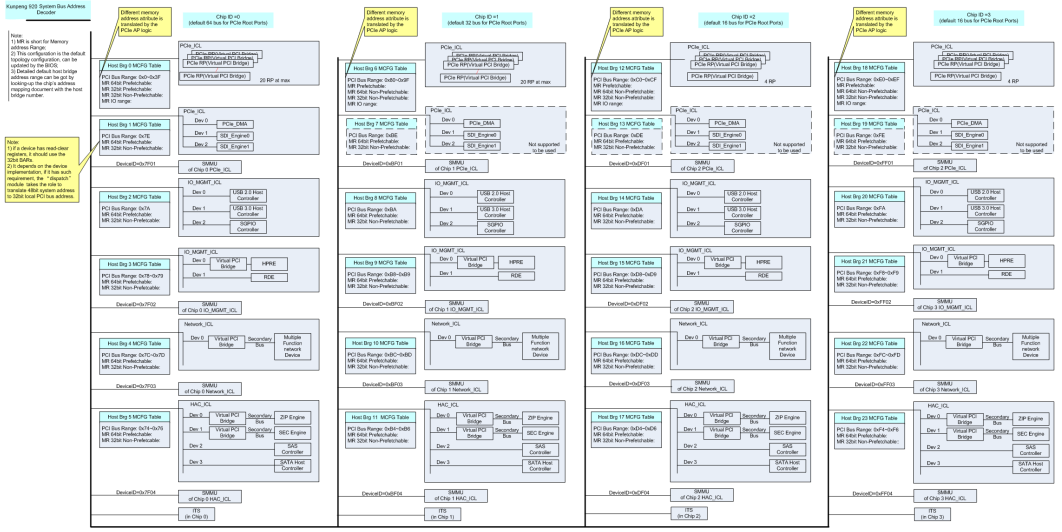
这些外设不是Arm SBSA（Server Base System Architecture，服务器基础架构）规范定义的通用平台设备。外设包括片内外设和片外外设，片外外设通过PCIe与华为鲲鹏920连接。这两种外设都采用PCI拓扑方式组织，遵循PCIe框架。

启动默认外设的拓扑结构如图5-2所示。整个拓扑完全遵循PCIe固件和软件架构。用户可以对照2.3 组件单元描述的集群内部结构，以对系统全面有更清晰的了解。

注意以下几点：

- 不管是单片架构还是多达4片互联的多片架构，整个SMP系统共享一个ECAM空间。提供PCI总线（包含256条子总线）用于挂接设备。整个系统共享一个PCI总线域。
- 受芯片物理设计和系统译码机制的限制，每个SCL都包含多个Host桥。Host桥应由ACPI（Advanced Configuration and Power Interface，高级配置与电源接口）表来描述。不包含在最终产品里的外设不能在ACPI表中体现。以Network ICL为例，如果PCB板只用到了Chip 1的网口控制器，Chip 0的Network ICL就不能体现在ACPI表中。那么，Chip 0的网络功能对软件不可见。
- 不支持虚拟化外设作为集成端点设备（Integrated Endpoint）与Host桥连接。支持虚拟化外设通过虚拟PCI-PCI桥与Host桥连接，可以获取更多PCI功能。对于不在产品PCB板上使用的片内外设，可以在固件里将这些设备的PCI头的厂商ID配置为0xffff，隐藏这些冗余逻辑。
- 片上总线功能通过Chip ID和SCL ID对默认地址译码逻辑进行调整。Chip 0有一套默认的总线号和地址范围分配配置，Chip 1有一套默认的总线号和地址范围分配配置，以此类推。以Chip 0为例，Chip 0的默认总线范围在单片和4片互联场景下相同。用户可以根据需要在固件中对这些配置进行更改。

图 5-2 外设拓扑



5.3 平台设备拓扑

表5-1对所有平台设备进行了总结。

所有平台设备都提供与Arm生态系统兼容的标准编程接口。固件向操作系统提供平台设备的描述，平台设备的驱动由操作系统通过通用软件框架驱动完成。

SMMU模块采用消息中断（Message Signaled Interrupt，MSI）的中断形式，SMMU驱动以编程方式将MSI ID和地址写入SMMU寄存器中。

表 5-1 平台设备拓扑

平台设备	功能	兼容架构版本	数量
通用UART	UART接口	Arm® Server Base System Architecture 3.0	每颗芯片都有一个UART接口。 固件描述决定哪些UART接口软件可见。
通用Watchdog	Watchdog	Arm® Server Base System Architecture 3.0	SMP系统共用一个Watchdog。
GICD	通用中断分发器	Arm® Generic Interrupt Controller Architecture Specification version 4.0	整个系统共用一个GICD。

平台设备	功能	兼容架构版本	数量
IMU消息通道	应用处理器与IMU的信息交互	Arm® Compute Subsystem SCP Version: 1.2 Message Interface Protocols	1组IMU消息通道。
ITS	Interrupt Translation Service（ITS）将MSI/MSI-X设备中断转换成系统的LPI中断，并发送给不同的通用中断控制寄存器（General Interrupt Control Register，GICR）。	Arm® Generic Interrupt Controller Architecture Specification version 4.0	每颗芯片都有一个ITS组件，包含在SICL中。
SMMU	1) 地址转换和内存访问保护； 2) PCIe_ICL中的SMMU也支持地址转换服务（Address Translation Services，ATS）和页面请求接口（Page Request Interface，PRI）虚拟化特性。	Arm® System Memory Management Unit Architecture Specification SMMU architecture version 3.0	每个ICL都有一个专用SMMU组件。

5.4 固件设备

固件设备可以被固件访问。

5.5 StreamID 与 DeviceID 对应关系

SMMU模块借助StreamID对地址转换表进行索引。ITS模块借助DeviceID对引中断转换表进行索引。在PCIe系统架构中，还有两种ID：Requester ID和PASID。Requester ID与PCI{总线，设备，功能}号组合，共同标识PCI层次结构中的功能。PASID是Requester的进程地址空间ID。在典型应用场景中，可能与内核的进程地址空间有关。关于这些ID定义的更多信息,请参见SMMU、GIC和PCIe规范。

根据章节**5 设备及设备拓扑**的描述，华为鲲鹏920采用PCIe系统拓扑进行基础设备管理。与SMMU或ITS产生工作交互的设备都按PCIe的要求分配了唯一的Requester ID：

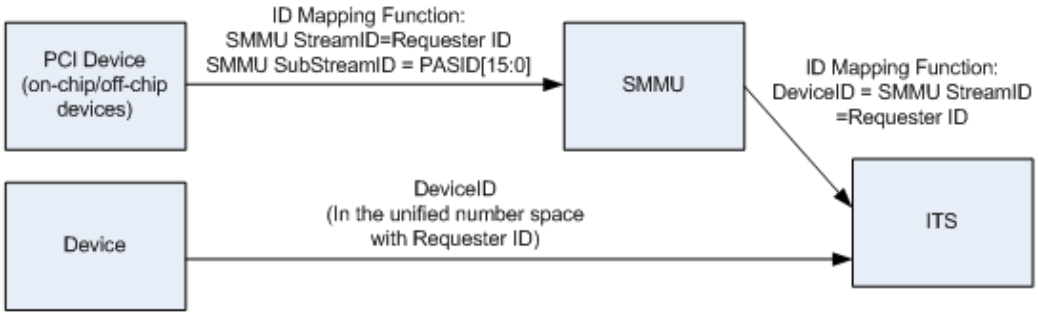
- 设备一般都具有标准的PCIe头。设备Requester ID由PCI枚举软件按标准流程分配。

- 没有标准PCI头的设备在系统启动时直接分配Requester ID。此类Requester ID用的是芯片预留的PCI总线。SMMU DeviceID示例如图5-2。
- 系统（包括多片系统）内的Requester ID均为16位长度的编号。

  这些ID的映射功能如图5-3所示。在整个路径中，一个设备的RequesterID、SMMU StreamID和ITS DeviceID都相同且唯一。设备的PASID映射到SMMU Substream ID，对Stage1转换表进行索引。

  华为鲲鹏920处理器对以上所有ID的映射都是1:1映射。

图 5-3 StreamID 和 DeviceID 的对应关系



5.6 设备内存序

设备内存具有重排序（Re-ordering）的属性。对于标记为“非重排序（nR）”的设备内存访问，其对相同块大小的访问应该按程序中定义的顺序在目标设备中体现。

对华为鲲鹏920的设备内存来说，每个PCIe根端口为块大小。也就是说，对同一个PCIe根端口的非重排序访问，与按照程序中定义的顺序访问效果一致；而对不同PCIe根端口的访问，除非借助内存屏障机制，否则访问顺序是无法保证的。

华为鲲鹏920的片内设备都是按顺序排列的独立的设备内存块，也就是说，对同一个设备的非重排序访问，与按照程序中定义的顺序访问效果一致；而对不同设备的访问顺序，除非借助内存屏障机制，否则访问顺序是无法保证的。

6 PCIe 子系统

华为鲲鹏920支持16GT/s数据传输，兼容PCIe规范V4.0。

6.1 PCIe 软件视图

华为鲲鹏920处理器的每个PCIe_ICL都是一个Host桥。华为鲲鹏920最多支持4片SMP配置，4个PCIe根端口对应4个Host桥。所有这些Host桥都位于统一的ECAM空间和统一的PCI总线域内。

华为鲲鹏920支持在PCIe根端口间进行Peer2Peer传输。当启用Peer2Peer功能时，任何PCIe设备都可以在软件逻辑视图中连接到一条共享PCI总线；PCIe设备共享一个公共地址空间，并且可以彼此发起数据传输。Peer2Peer传输通过Host桥配置进行路由，RC感知系统Host桥内存和总线范围信息，且这些信息可以在BIOS阶段得到同步。

中断同样对软件可见。PCIe设备的MSI/MSI-X中断路径，请参见**5.5 StreamID与DeviceID对应关系**和**5.5 StreamID与DeviceID对应关系**。系统把所有PCIe INTx中断整合为4个SPI中断。注意，这是针对整个PCIe总线域来说的，对应统一的ECAM空间，并非每个根端口都有4个SPI中断。

在Armv8架构下，PCIe设备的配置空间、内存地址空间、I/O地址空间被映射到全局内存地址。配置空间访问指的是对ECAM空间的内存访问。Host桥利用地址转换单元（Address Translation Unit，ATU）机制，将系统访问转换为PCIe内存和I/O请求。华为鲲鹏920不支持对外部的原子操作。转换机制如**图2-6**所示。

表 6-1 PCIe 地址空间和 Arm 内存类型配置

PCIe目标地址空间	PCIe请求	推荐Arm内存类型	PCIe控制器转换机制
PCIe配置空间	配置读； 配置写。	Device-nGnRnE型设备内存	1) 系统基内存地址，用于标识ECAM访问。 2) PCI总线范围，用于过滤配置访问。

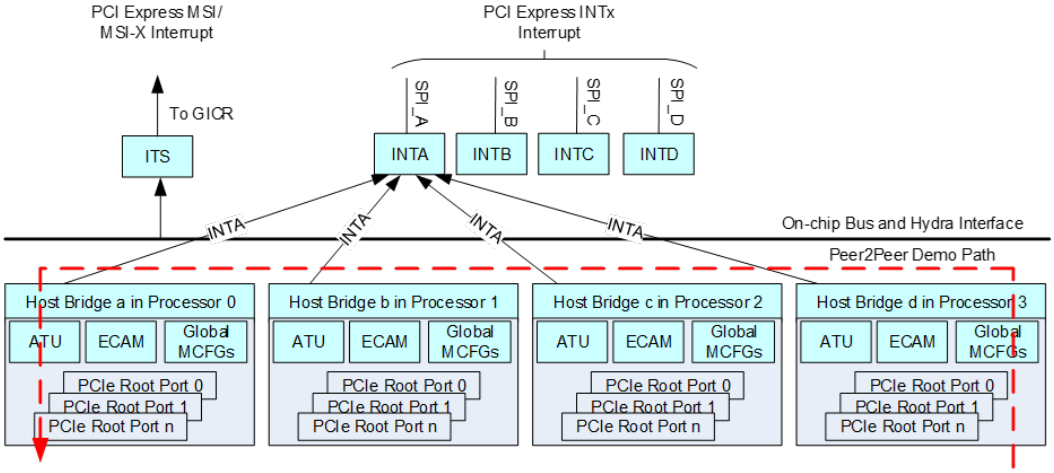
PCIe目标地址空间	PCIe请求	推荐Arm内存类型	PCIe控制器转换机制
I/O空间	I/O读； I/O写。	Device-nGnRnE型设备内存	ATU表项
64bit可预取内存地址空间	供寄存器使用。 内存读； 内存写。	Device-nGnRE型设备内存	ATU表项
	用于内存空间扩展。 内存读； 内存写。	Device-nGRE型设备内存； Device-GRE型设备内存； 不可缓存型普通内存	
32bit可预取内存地址空间	供寄存器使用。 内存读； 内存写。	Device-nRE型设备内存	ATU表项将32bit的PCIe地址映射到Arm内存中大于4G地址空间。
	用于内存空间扩展。 内存读； 内存写。	Device-nGRE型设备内存； Device-GRE型设备内存； 不可缓存型普通内存	
32bit不可预取内存地址空间	内存读； 内存写。	Device-nGnRE型设备内存	ATU表项

说明

1. 强烈建议将PCIe内存映射的寄存器空间设为Device-nGnRE型设备内存，这样华为鲲鹏920才能够保证读写访问的执行顺序。编程时可以通过读请求来Flush在PCIe系统中的滞后写数据（Posted Write Data）。
2. 当PCIe内存空间为Device-R型设备内存或不可缓存型普通内存时，编程时应借助屏障机制来保证访问的执行顺序。
3. 当PCIe配置和I/O空间为Device-E型设备内存时，则无法保证PCIe配置写和I/O写操作能正常生效。因为虽然Arm内存序对配置写和I/O写进行了提前应答，但PCIe无法对配置写和I/O写执行Flush。

这些特性如图6-1所示。图中为4个处理器级联场景示意图。若SMP系统中处理器少于4个，可以直接忽略对应处理器的Host桥。

图 6-1 华为鲲鹏 920 PCIe RC 软件视图



6.2 华为鲲鹏 920 PCIe 硬件视图介绍

华为鲲鹏920的PCIe控制器具有多个根端口的电路逻辑，根端口作为虚拟PCI-PCI桥呈现在软件视图中。如果板端的实际物理端口数小于芯片最大端口数，可以通过BIOS对控制器进行初始化，屏蔽不存在的根端口。例如，x16 PCIe控制器可以配置为单个根端口，即系统可见单桥；也可以配置为4个x4根端口，即系统可见4桥。

华为鲲鹏920支持PCIe热插拔。芯片的I2C总线可以获取PCIe插槽的硬件状态，并将该状态信息同步给根端口指示器。软件用户界面遵循PCIe标准。I2C总线状态收集协议采用华为鲲鹏920专用时序。

6.3 PCIe 系统特点

除了物理、数据链路和事务层之外，还有许多与芯片实现相关的系统特性。华为鲲鹏920并没有全部实现所有特性，详见表6-2。

华为鲲鹏920已实现的系统特性主要满足以下三大编程接口规范：

- PCIe Base规范
- PCI固件规范
- Arm生态规范

表 6-2 PCIe 系统特性移植

系统架构功能	华为鲲鹏920是否支持	备注
中断/功耗管理事件（Power Management Event，PEM）	支持	SMP系统把所有INTx中断转换为4个SPI本地中断。
错误信令和日志	支持	-
虚拟通道(VC)	原生PCIe不支持 仅适用于CCIX	-

系统架构功能	华为鲲鹏920是否支持	备注
设备同步	支持	-
锁定事务	不支持	-
PCIe复位	支持	-
PCIe热插拔	支持	-
功率预算能力	不支持	-
槽位功率限制	不支持	-
RC拓扑发现	支持	ECAM机制
链路速率管理	支持	-
访问控制器服务（Access Controller Services，ACS）	支持	RC/SR-IOV/Peer2Peer
替换路由ID协议（Alternative Routing-ID Interpretation，ARI）	支持	-
多播	不支持	-
原子操作	仅支持对内部的原子操作	不支持对外部的原子操作，即华为鲲鹏920不能通过CPU原子命令向PCIe设备外部发起原子操作。
动态功率分配（Dynamic Power Allocation，DPA）	不支持	-
事务层包处理提示（Transaction Layer Packet Process Hints，TPH）	支持	-
延迟容限报告（LTR，Latency Tolerance Reporting）	不支持	-
Optimized Buffer Flush/Fill（OBFF）	不支持	-
PASID	支持	与SMMU共同工作ID的宽度为16bit，而不是20bit。
轻量级通知（Lightweight Notification，LN）	不支持	-

系统架构功能	华为鲲鹏920是否支持	备注
精确时间测量（Precision Time Measurement，PTM）	不支持	-
就绪通知（Readiness Notifications，RN）	支持	-
增强分配（Enhanced Allocation，EA）	不支持	-

7 网络子系统

华为鲲鹏920通过外接网络设备，提供高带宽以太网络能力。Network ICL控制以太网帧的接收和发送，支持典型的业务处理机制，支持基于RoCEV1和RoCEV2规范的RDMA。华为鲲鹏920的最大线速为100Gbps，RDMA目标时延低于600ns。提供通用网卡，性能调优支持项完全兼容Linux ethtool，也支持Microsoft网络堆栈。

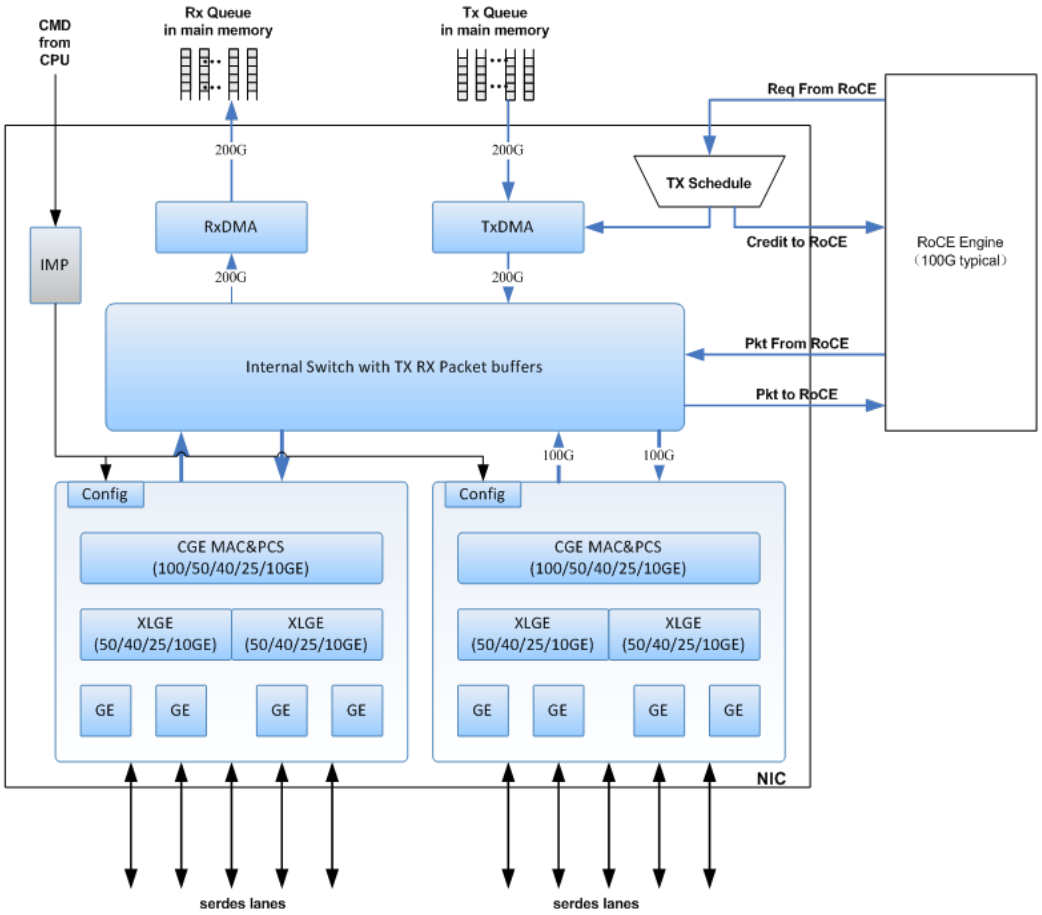
编程方面，Network ICL兼容PCIe系统架构。提供标准的PCIe配置空间，供PCI ECAM机制访问，支持BIOS枚举。虚拟化符合SR-IOV规范。支持256项PCI功能，包括虚拟功能（Virtual Function，VF）和物理功能（Physical Function，PF）。每个PCI功能都有独立的任务和配置空间，便于业务数据处理和硬件配置。

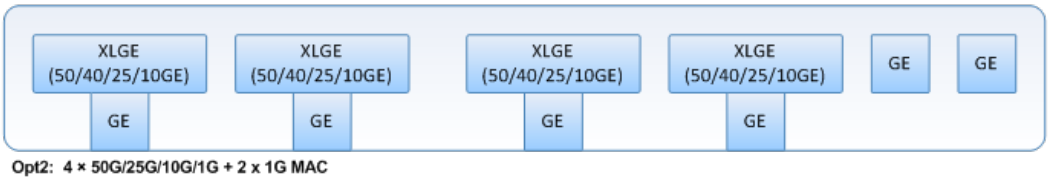
华为鲲鹏920网络子系统基于共享内存架构，如**图7-1**所示。存储交换单元提供收发包缓存，完成包的缓存、存储和交换。TxDMA和RxDMA用于将报文放入报文队列或从CPU内存中载入报文队列。

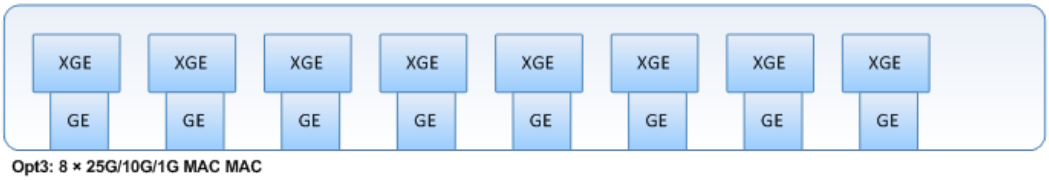
集成管理处理器（Integrated Management Processor，IMP）是一种嵌入式的管理处理器，主要完成命令通道的管理和网络控制器边带接口（Network Controller Sideband Interface，NC-SI）功能。IMP负责解析读写命令，配置硬件和获取硬件配置。IMP还负责响应主板管理控制器（Baseboard Management Controller，BMC）的配置请求。

每个50GE/40GE MAC控制器都可以配置为1个XLGE MAC控制器或者4个XGE MAC控制器。因此，也可以进行如下配置组合：2个100G/50G/40G/25G/10G/1Gbps MAC控制器+2个千兆MAC控制器，或者4个50G/25G/10G/1Gbps+2个千兆MAC控制器，或者8个25G/10G/1Gbps MAC控制器。

图 7-1 华为鲲鹏 920 网络子系统框架示意图







我们把这些功能都视做PCIe框架的功能外设，并且与通过PCIe外接的NIC或HBA网卡在同一个框架中集中管理。通用网络逻辑和RDMA逻辑放在同一个PCIe功能中，不占用不同的PCI功能。

华为鲲鹏920典型网络接口配置如表7-1所示。BIOS初始化Network ICL的逻辑，根据单板的物理连接告知系统这些功能外设的设备ID。

表 7-1 华为鲲鹏 920 网络可用设备配置

PCI厂商ID（华为）	PCI设备ID	功能描述
0x19E5	0xA220	华为鲲鹏920 GE片内NIC。
	0xA221	华为鲲鹏920 GE/XGE/25GE片内NIC。
	0xA222	华为鲲鹏920 GE/XGE/25GE片内NIC，支持RDMA、数据中心桥接（Data Center Bridging，DCB）及优先级流量控制（Priority Flow Control，PFC）。
	0xA224	华为鲲鹏920 GE/XGE/25GE/50G片内NIC，支持RDMA、DCB及PFC。
	0xA226	华为鲲鹏920 GE/XGE/25GE/40G/50GE/100GE片内NIC，支持RDMA、DCB及PFC。
	0xA22E	华为鲲鹏920片上NICVF虚拟功能，不支持RDMA、DCB及PFC。
	0xA22F	华为鲲鹏920片上NICVF虚拟功能，支持RDMA、DCB及PFC。

8 管理子系统

智能管理单元（Intelligent Management Unit，IMU）是华为鲲鹏920的管理子系统，配备64位Armv8内核。IMU在芯片上完全独立。这意味着无论其他子系统发生什么情况，IMU的代码空间、数据空间、外设、时钟、复位功能都可以正常工作。

其主要组件有：

- 64位Armv8内核的指令架构集（Instruction Set Architecture，ISA）子系统，及其外设和中断控制器。外设兼容Arm SBSA。中断控制器符合GICv3标准。
- 片上内存，为IMU内核提供专用的基础管理固件代码空间和数据空间。在其他子系统发生异常时，确保IMU子系统正常工作。该片上专用内存不能被应用处理器访问。
- 系统隔离墙，为IMU子系统外部环境提供安全通道，确保IMU RISC不会因系统异常而阻塞。
- I2C接口，连接其他处理器以及BMC，当多片互联场景下芯片间的高速接口发生异常时，I2C接口可以充当带外通道，供IMU访问其他处理器。

IMU的主要功能如下：

- 通过智能平台管理总线（Intelligent Platform Management Bus，IPMB）协议与BMC等外部管理芯片进行信息交互。
- 通过系统控制管理接口（System Control and Management Interface，SCMI）与应用处理器通信，实现对应用处理器的全面管理，包括电源管理、热管理、状态和事件管理；通过应用处理器序列号（Application Processor Equipment Identity，APEI）与应用处理器共享内存空间和数据。
- 安全启动，启动代码验证，包括统一可扩展固件接口（Unified Extensible Firmware Interface，UEFI）固件和信任固件。安全启动功能可配置。

图 8-1 华为鲲鹏 920 IMU 框架示意图