◎网络与通信

PCIe 交换架构在数据中心网络中的优化研究

何泉初

（中国电子科技集团公司第五十八研究所，江苏　无锡　214000）

摘　要：随着数据中心规模不断扩大，传统以太网架构面临性能瓶颈，PCIe 交换技术以其高带宽、低延时、可扩展等优势，为数据中心网络优化提供了新思路。本文分析了引入 PCIe 交换架构对提升数据中心网络性能的重要意义，探讨了 PCIe 交换在数据中心中诸多应用优势，从拓扑优化、路由调度、容错冗余等方面提出了面向数据中心网络 PCIe 交换架构优化方法，以期为构建高性能数据中心网络提供参考。

关键词：数据中心网络；PCIe 交换技术；性能优化；互连架构

中图分类号：TN402　　**文献标志码：**A　　**文章编号：**1672-4739（2025）01-0185-03

Research on Optimization of PCIe Switching Architecture in Data Center Networks

HE Quanchu

（*China Electronics Technology Group Corporation*, *Wuxi* 214000, *China*）

Abstract: With the continuous expansion of data center scale, the traditional Ethernet architecture is facing performance bottlenecks. PCIe switching technology, with its high bandwidth, low latency, and scalability, provides a new approach for optimizing data center networks. This paper analyzes the significance of introducing PCIe switching architecture in enhancing the performance of data center networks, discusses the various application advantages of PCIe switching in data centers, and proposes optimization methods for PCIe switching architecture in data center networks from aspects of topology optimization, routing scheduling, and fault tolerance redundancy. The aim is to provide references for constructing high-performance data center networks.

Keywords: data center network; PCIe switching technology; performance optimization; interconnection architecture

0　引言

数据中心是支撑云计算、大数据等新兴应用关键基础设施，随着数据中心规模急剧扩大，服务器数量快速增长，东西向流量激增，传统以太网架构逐渐暴露出带宽不足、延迟较高等短板，制约了数据中心进一步发展。PCIe 交换技术具有高速串行点对点通信、全双工传输、硬件 QoS 保证等特性，近年来在高性能计算领域得到广泛应用，将 PCIe 交换引入数据中心，构建扁平化互连架构，有望突破传统以太网瓶颈，大幅提升系统性能。

1　PCIe 交换架构引入数据中心网络意义

PCIe5.0 是高速串行点对点互联技术，单根链路带宽高达 32GB/s，能有效缓解服务器 I/O 瓶颈，引入 PCIe 交换，可构建扁平化数据中心架构，支撑东西向带宽快速增长。相比传统以太网，PCIe 在硬件层面支持 QoS，能更好保证关键业务性能，提升用户体验，PCIe 支持直接 IO 虚拟化，可实现存储、网卡等设备跨服务器共享，提高资源利用率，PCIe 为优化数据中心互连、提升整体性能提供了新思路新方法。

2　PCIe 交换架构在数据中心网络中应用优势

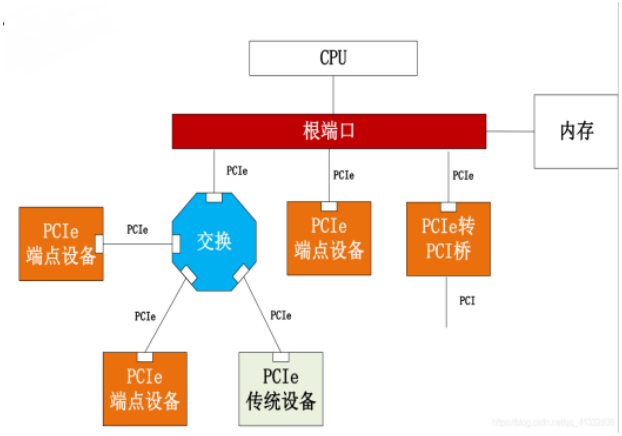


图 1　PCIe 交换架构图

收稿日期：2024-09-19

作者简介：何泉初（1995.11-），男，汉族，四川成都人，硕士，助理工程师，研究方向：PCIe 交换、时间敏感网络交换。

PCIe5.0 交换在数据中心网络中具有诸多应用优势。一是超高带宽，130GB/s 交换容量可满足网络快速扩容需求，单根 PCIe 链路可提供 100Gb/s 以上带宽，远超以太网。二是超低延时，端到端转发延迟低至 200ns 量级，远低于以太网微秒级，满足金融、科研等时延敏感应用。三是灵活互联，PCIe 点对点拓扑结构，使得服务器互联更加灵活便捷，减少级联环节。四是硬件加速，PCIe 原生支持 RDMA，可通过硬件卸载降低 CPU 开销，有效提升通信性能。五是扁平架构，服务器直连 PCIe 交换芯片，省去传统以太网多层交换，网络扁平化程度高。

3 面向数据中心网络 PCIe 交换架构优化方法

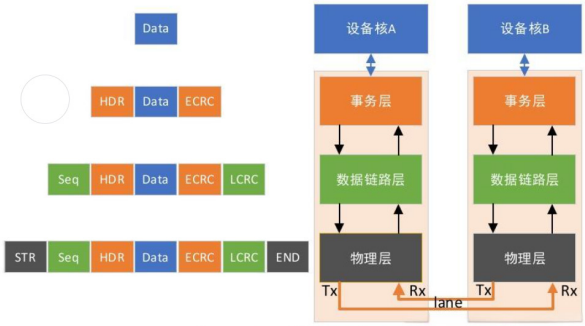


图 2 PCIe 交换架构优化图

3.1 拓扑优化与连接布局

3.1.1 多级交换拓扑结构设计

针对不同规模数据中心，PCIe 交换网络需要采用与之相适应的多级交换拓扑结构，对于支持上万台服务器互联超大规模数据中心，可采用 Clos、Fat tree 等多级交换拓扑，这些拓扑通过分层递归方式，将海量服务器划分为多个 Pod，每个 Pod 内又划分为多个机架，形成了高度模块化、可扩展网络结构。Pod 内所有服务器通过兼具上下行链路 leaf 交换机互联，Pod 之间通过 spine 层交换机实现无阻塞互通，这种分层多级拓扑结构，不仅能够支持大规模网络平滑扩容，而且能有效降低链路阻塞概率，满足云计算等应用对东西向带宽需求，而对于中小规模数据中心，则可采用单级或两级交换拓扑。单级拓扑即每台服务器通过 PCIe 直连到同一台交换机，两级拓扑即增加一级汇聚交换，分担上行链路压力，相比多级拓扑，这种扁平化架构虽然扩展性有限，但成本更低，部署更简单，更适合规模有限中小型数据中心。

3.1.2 服务器到交换机链路聚合

PCIe 技术支持多链路聚合，可充分利用这一特性优化服务器到交换机互联带宽，每台服务器可配置多块 PCIe 网卡，分别与不同交换机端口相连，通过链路聚合技术。将这些物理链路捆绑为一个逻辑链路，从而成倍提升服务器互联带宽，这种多链路互联方式还能避免单点故障，当某条链路或端口出现问题时，流量可自动切换到其他活动链路，从容应对故障。多链路聚合还为实现负载均衡提供了便利，通过均衡调度算法，可将服务器流量动态分摊到各条链路，充分利用链路带宽，降低拥塞概率。而在服务器和交换机之间，也可通过双网卡、双交换机方式实现冗余互联，进一步提升可靠性。

3.1.3 交换机间互联带宽优化

在多级交换架构中，上层交换机往往承载着大量来自下层汇聚流量，容易出现带宽瓶颈，制约网络整体性能，因此在设计拓扑结构时，要注重优化各级交换机，尤其是上层交换机互联带宽。主要可采取以下策略：一是提高上层交换机端口配置带宽，在选型时，要充分考虑峰值流量需求，选择端口带宽足够大交换芯片，如单端口 400G 及以上，避免带宽不足导致排队阻塞。二是采用双层 spine 架构，即增加第二级 spine 交换机，分担流量疏导任务，将部分流量调度职责下沉至第一级 spine，减轻核心层流量转发压力，扩大网络扁平化程度，这种双层互联结构也具有更强容错能力，能应对交换机级故障。

在 PCIe 交换组网时，除了传统树状拓扑外，还可考虑引入环形、格子等拓扑元素，尤其是在 Pod 内或机架内局部组网中，通过在服务器之间，或服务器与交换机之间配置冗余链路，形成闭环或网格状拓扑连接，能显著提升局部网络可靠性和抗毁性。一旦发生链路中断，可依靠环形拓扑提供冗余路径快速恢复通信，而格子状拓扑能在多个方向上提供冗余互联，进一步降低网络阻断概率。

3.2 路由与流量调度优化

3.2.1 基于 PCIe 最短路径路由算法

路由计算是网络高效运转核心，针对 PCIe 交换架构，需要设计专门路由算法，在满足最短路径同时降低计算开销。传统的最短路径算法，如 Dijkstra、Floyd 算法，虽然能够找出两点之间最短路径，但算法复杂度高，难以满足 PCIe 交换网络实时性需求，因此需要针对 PCIe 拓扑特点，设计改进路由算法。一种思路是采用分布式计算思想，打破集中式路由瓶颈，可在各台 PCIe 交换机上分别部署路由模块，各自负责计算一部分最短路径，最后汇总形成全局路由表，这种分布式计算可显著降低单个交换机计算负担，提升路由效率。另一种思路是采用递归、分治策略，充分利用 PCIe 多级树状拓扑层次性，先各自计算 Pod 内叶子交换机到上层 spine 交换机最短路径，再迭代计算 Pod 间最短路径，逐层向上收敛，最终形成端到端最短路径。这种分层计算避免了在一开始就计算整个拓扑全局最短路径，能较好控制计算规模，还可引入路径缓存机制，将反复查询路径结果缓存下来，减少重复计算开销。

3.2.2 负载均衡与拥塞避免策略

均衡调度网络流量、避免拥塞是提升 PCIe 交换网络性能的关键，由于 PCIe 交换芯片普遍集成了 QoS 功能，支持对不同优先级流量进行分类调度，因此我们要充分发掘 PCIe 硬件特性，在芯片内实现精细化流量调度策略。对于不同类型流量，可采用不同调度算法，如对于实时性强小颗粒流量，可采用优先级抢占式调度，保证其时延需求；对于尾长分布大流量，可采用轮询或最小完成时间优先等调度算法，避免带宽被大流量独占。还要在各级交换节点之间协同实现端到端 QoS 保证，通过在入口交换机处为流量打上 QoS 标记，并将该信息在转发过程中持续传递，直至送达服务器，实现全链路服务质量保障。

针对网络拥塞问题，要采取基于虚拟输出队列（VOQ）缓存管理机制，PCIe 交换机普遍支持 VOQ，可针对每个项目端口设置缓存队列。当某个端口发生拥塞时，可将该端口流量缓存在 VOQ 中，而不阻塞其他端口，从而避免无谓 Head-of-Line 阻塞，提高网络整体吞吐量。在拥塞发生时，还可采取基于 VOQ 占用比流量限制措施，当某个 VOQ 占用量超过预设阈值时，触发限流保护机制，通过丢包、间隔发送等措施，动态削减端口流量，直至拥塞解除，这种基于速率拥塞控制，可精准调控瓶颈端口流量，抑制局部拥塞向整网蔓延。

3.2.3 优先级与 QoS 质量保证机制

针对关键业务应用，要通过赋予其较高 QoS 优先级，提供带宽和时延确定性保障，满足差异化服务质量需求，PCIe 交换机支持通过多种方式标识流量优先级。一是基于 VLAN 优先级（PCP）二层 QoS 标识，可在以太网帧头中携带 PCP 字段，将流量映射到不同 QoS 等级。二是基于 DSCP 三层 QoS 标识，可在 IP 包头中携带 DSCP 字段，实现更细粒度 QoS 分类。三是基于链路自定义 QoS 标识，可在 PCIe 包头中自定义保留字段，实现灵活可控优先级分配，支持多种 QoS 标识方式，能更全面地适配不同应用服务质量诉求。在识别优先级后，下一步是进行优先级映射和队列调度，可在交换芯片中预设多个优先级队列，将不同 QoS 等级流量映射到对应队列，队列调度可综合采用 SP、WRR、DRR 等算法，既保证高优先级业务带宽，又兼顾低优先级业务公平性[1]。

3.3 容错冗余与可靠性保障

3.3.1 交换机冗余备份设计

随着数据中心网络规模持续扩张，对业务连续性提出了更高要求，任何网络故障都可能带来巨大经济损失和客户投诉。因此在 PCIe 交换网络设计中，必须重视容错冗余和可靠性保障，尤其是在核心骨干部分，需进行充分交换机冗余备份，可采用主备模式或多机互备模式，部署冗余 PCIe 交换机。在主备模式中，两台交换机采用主从配置，正常情况下由主交换机负责业务转发，备用交换机处于热备状态。一旦检测到主交换机故障，备用交换机可快速接管所有流量，继续提供无中断转发服务，而在多机互备模式中，多台交换机并列部署，分担流量转发，任一交换机发生故障，其余机器可共同分担故障机器流量，整体容量虽有所下降，但不至于完全中断业务。对于规模较大数据中心网络，单一主备冗余很难满足可靠性需求，需要进行分层分区冗余设计，可考虑在每个 Pod 内部署主备或多机互备，同时在核心层再部署第二级冗余备份，形成纵深防御。这种分层冗余，可在不同粒度上实现故障隔离和流量接管，确保在交换机级故障发生时，本 Pod 乃至全网连通性[2]。

3.3.2 链路故障快速检测与恢复

由于 PCIe 交换网络普遍采用分布式 Fetree 拓扑，跨越多级交换端到端链路一旦发生故障，定位和恢复时间可能较长，严重影响网络性能，因此必须缩短链路故障检测和响应时间，通过及时发现、快速收敛，将故障负面影响降到最低。要制定完善链路探测机制，周期性地主动发送探测报文，实时感知链路连通状态，可基于 LinkAwait 协议，由下游节点主动发起探测，上游节点进行回应确认。一旦连续多个探测周期未收到上游回应，即可判定链路中断，立即启动告警和链路切换流程，探测频率要根据网络规模、稳定性等因素灵活设置，过高会加重探测开销，过低又难以及时发现故障，需要慎重权衡[3]。

在故障发生后要通过软硬协同实现快速链路切换，交换硬件需要支持多条等价路径配置，包括等价上联、等价下联等形式，在主用链路 failed 后，可自动将流量切换到等价备用链路，交换芯片可内置保护组表项，将端口隔离到不同广播域，避免故障端口环路风险[4]。交换软件也要及时感知链路状态变化，快速收敛转发表项，通过 Software Define Networ 等技术，将故障通知第一时间同步到控制平面，触发路由重计算、MAC 表刷新等操作，尽可能缩短故障恢复时间，还可引入 Bidirectional Forwarding Detection（BFD）等快速检测协议，通过毫秒级故障检测，进一步提升故障响应速度。

3.3.3 网络全局异常处理机制

数据中心网络中可能发生形式多样的网络异常，包括流量暴增、广播风暴、服务器故障等，这些异常往往难以预期，且极易引发连锁反应，危及整网。需要建立一套全局异常处理机制，对各种突发事件进行实时感知和处置，最大限度消除异常扩散。构建全方位的网络监测体系，实现网络全局的可视化监控，利用 SNMP、sFlow、Netconf 等协议，采集交换机的各项告警信息包括端口 flap、带宽超限、队列溢出等，并汇聚到集中的监控平台。可视化手段实时展示网络健康度，以拓扑图、曲线图等形式直观呈现链路带宽、交换机负载等关键指标，为管理人员实时掌控全局状态，一旦发现流量突增、拓扑频变等异常，平台自动预警并精准定位故障点，为及时止损赢得宝贵时间[5]。

4 结 语

随着云计算、人工智能等数据中心典型应用加速落地，高速互联的需求只会愈发强烈。PCIe 交换技术必将在数据中心的互联变革中扮演关键角色，我们要立足长远，综合考量性能、成本、可靠性、可管理性等多元要素，因地制宜设计最佳互联方案。这不仅需要有一流的技术实力，还需要勇于创新的精神和持之以恒的韧劲。

参考文献：

[1] 伊雯雯，王喜，张书奎 . 递归型数据中心网络上 g- 额外连通度及容错单播算法研究[J]. 计算机应用与软件 ,2024,41(1):261-268.
[2] 潘志勇，程宝雷，樊建席，等 . 数据中心网络 BCDC 上的顶点独立生成树构造算法 [J]. 计算机科学 ,2022,49(7):10.
[3] 刘源 . 低时延数据中心网络流量控制技术研究 [D]. 天津大学 ,2022.
[4] 林凡森，刘鑫，陆晓峰 . 一种 PCIe 交换电路设计与实现 [J]. 电子与封装 , 2021,21(03):52-56.
[5] 王澧，王蕾，张玲 . 一种 PCIe 总线 MAC 模块的设计与验证 [J]. 电子与封装 , 2015,15(10):20-25.