CAICT 中国信通院

ACTIC 全国先进计算技术创新大赛

2024A3-基于DPU的SRv6技术赋能算力网络

A3-房银轰

- 队长：伊啸
- 队员：伊啸，李义

- 2024年8月17日

CAICT 中国信通院

ACTIC 全国先进计算技术创新大赛

目录

content

1 团队介绍

2 赛题回顾

3 主要工作

4 总结

CAICT 中国信通院

ACTIC 全国先进计算技术创新大赛

1 团队介绍

团队介绍

伊啸

- 北京大学软件与微电子学院硕士研究生在读;
- 现实习任职于字节跳动抖音电商后端开发;

李义

- 中山大学系统科学与工程学院硕士研究生在读;
- 研究方向为基于大模型的智能运维、无人机仿真平台架构设计

CAICT 中国信通院

ACTIC 全国先进计算技术创新大赛

2 赛题回顾

赛题回顾

1.赛题要求：体现DPU技术及产品在某个业务场景的应用效果

评判标准：

（1）体现DPU面向重点行业、重点领域的应用场景探索；

（2）体现应用DPU的主要技术指标： **高带宽、低时延、大规模ROCE组网**等；

（3）体现DPU在性能、成本、业务创新等方面的技术优势和经济价值；

2.赛题

整个应用方案应涵盖以下主要内容：

（1）当前业务场景面临的主要技术难点和痛点问题分析；

（2）提炼总结出能够利用DPU技术特点和优势（高带宽、低时延、大规模组网、高性能存储）的应用方案设想；

（3）利用**hadoop、spark、pytorch、k8s、openstack**等业界成熟的主流应用软件，结合DPU提供的OVS、DPDK、SPDK、裸金属架构、容器CNI、服务网格、安全、管理等基础设施卸载和加速能力的应用落地方案；

（4）应用落地方案应包括：整体架构、利用的应用软件功能说明、使用的DPU技术特性和能力详细说明、实现的应用效果和经济价值；

（5）业务发展对下一步我国DPU技术、市场、产业生态发展的诉求和展望；

CAICT 中国信通院

ACTIC 全国先进计算技术创新大赛

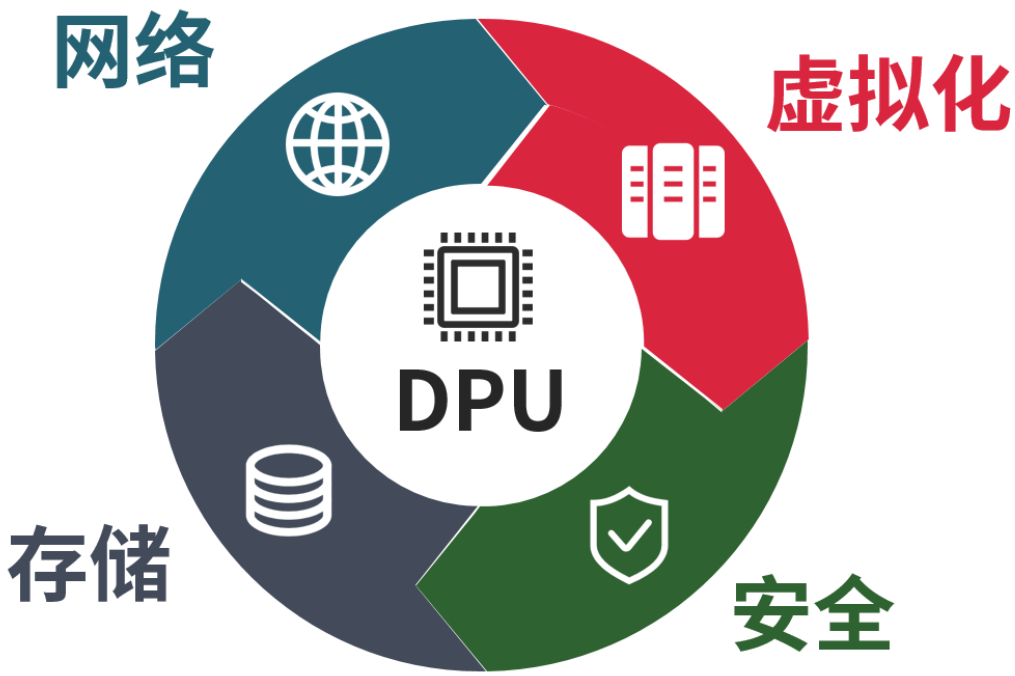
3 主要工作

DPU技术概述

数据中心标配： CPU + DPU + GPU
CPU 通用计算，GPU 加速计算，DPU 数据处理

Datacenter Tax（数据中心税）

- 以网络为例。主机在收发数据时，需要进行海量的网络协议处理。根据传统的计算架构，这些协议处理都是由 CPU 完成的。
- 线速处理 10G 的网络，需要的大约 4 个 Xeon CPU 的核。也就是说，仅仅是进行网络数据包的处理，就要占用一个 8 核高端 CPU 一半的算力。
- 现在数据中心网络不断升级，从 10G 到 40G 、 100G ，甚至 400G 高速网络，这些性能开销**如何承受**?



云计算领域痛点

资源争抢限制

随着资源需求的增多，资源争抢容易造成服务质量不稳定，尤其在大负载、大流量时I/O性能容易出现严重抖动，无法保障稳定的SLA体验。

计算特性损失

一方面，虚拟机相比物理机存在一定的性能损失；另一方面，私有云无法更好地利用公有云弹性云主机资源，限制了云主机的使用场景。

裸金属管理问题

裸金属服务器虽能提升性能，但所有CPU资源专属于用户，无法支持云管理，需外部管理，且与云的弹性、灵活原则不符。此外，对接远程存储时存在安全风险。

高性能网络和存储需求

随着AI应用增多，云上AI任务对网络和存储延迟要求更高。传统网卡架构下的RDMA和NVMe协议难以适应云计算多租户的灵活性需求。

DPU云计算场景全卸载

提升 I/O 性能	安全性增强	算力释放
• DPU 在具备标准网卡能力的同时，利用专用硬件完成网络和存储 I/O，释放主机 CPU 算力资源的同时可显著。	• 业务（主机 CPU）与虚拟化软件（DPU）的硬件载体分离，业务与云平台的隔离性以及主机的安全性进一步提高。	• DPU 在单部件成本上有所增加，但是 DPU 的引入解放了更高成本的主机 CPU 算力，释放了更多可售卖资源

算力网络发展需求

算力网络是以计算为核心，通过网络实现连接，通过感知实现资源匹配与调度的平台

人工智能（AI）和深度学习

- 随着深度学习模型的复杂性不断增加，训练这些模型所需的计算资源也随之增加。
- 深度神经网络中的参数数量和层数的增加使得训练任务更加庞大和耗时

在实时推理场景下

- 部署在实时环境中的 AI 应用（例如自动驾驶、人脸识别和语音识别等），要求快速、低延迟的推理能力
- 为了满足这些实时性要求，同样需要强大的计算资源来执行模型推理。

算力网络发展的挑战

技术挑战

算力网络涉及到异构硬件和芯片、接入和互联网络、数据中心/云计算， 要求支持灵活的可编程能力。

安全和隐私挑战

随着算力网络的广泛应用，如何保证数据的安全和用户的隐私， 也是一个重要的挑战。

感知和度量挑战

算力网络需要增强算力感知，推动算力度量，识别、标识和路由的技术体系，这是一个复杂的任务。

编排和服务挑战

算力网络需要增强算力编排，推动算网一体统一编排和调度的融合发展，推动以算力业务驱动的算力经济发展模式。

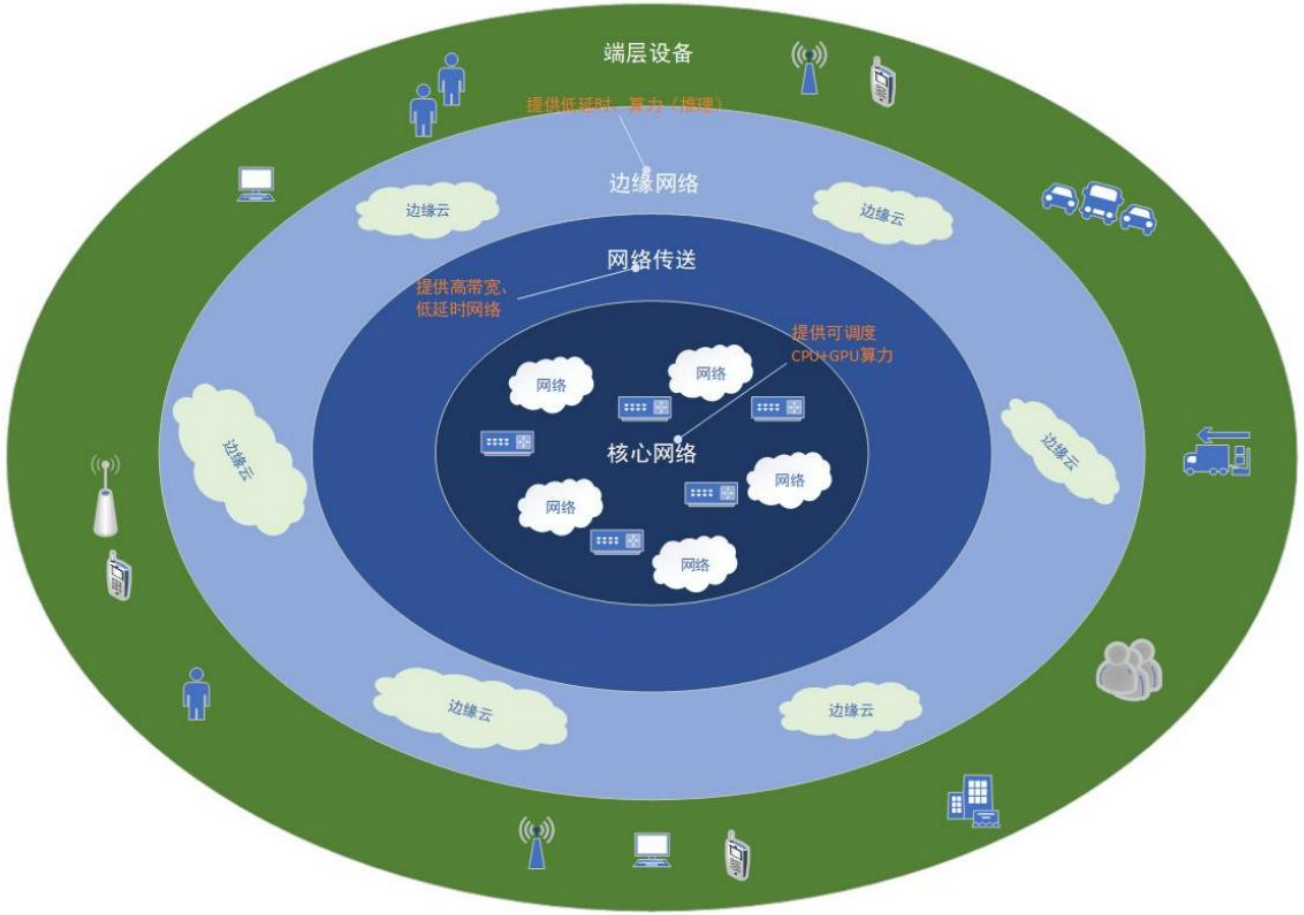
标准和规范挑战

目前，算力网络的标准和规范还不完善，这可能会阻碍算力网络的发展。

算力网络需求分布特点

算力资源匹配不均衡：5G、AI、IOT、 智能汽车、工业 4.0，大语言模型的蓬勃发展

- **网络边缘区域：**
  - 需要低延时，高算力特性；
- **网络传输区域**
  - 需要高带宽，低延时特性；
- **核心网络**
  - 高数据吞吐、巨量通用 CPU 算力和高并行 GPU 算力的特点



数据中心、超算中心、边缘云等“孤岛”网络，各自为战！

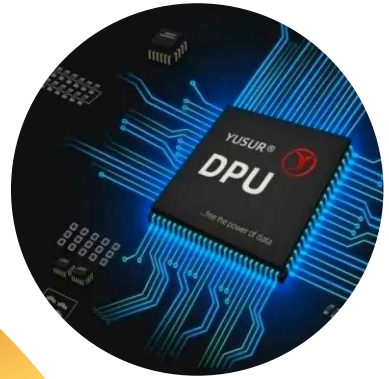
SRv6技术解决了传统广域网的N多难题

IPv6 ＋ **Segment Routing** ＋ **More Function**

大规模终端接入

基于应用的网络分片

安全，工业，5G等可编程需求



SRv6是什么?

定义：简单来讲即SR+IPv6，是新一代IP承载协议。它采用现有的IPv6转发技术，通过灵活的IPv6扩展头，实现网络可编程性。

原理：将报文路径切分为Segment，用IPv6地址形式的SID区分。为此，IPv6报文增加SRH扩展头，含Segment List指定路径。头节点添加SRH后，中间节点依据其信息转发报文。

SRv6

什么是SRv6？：

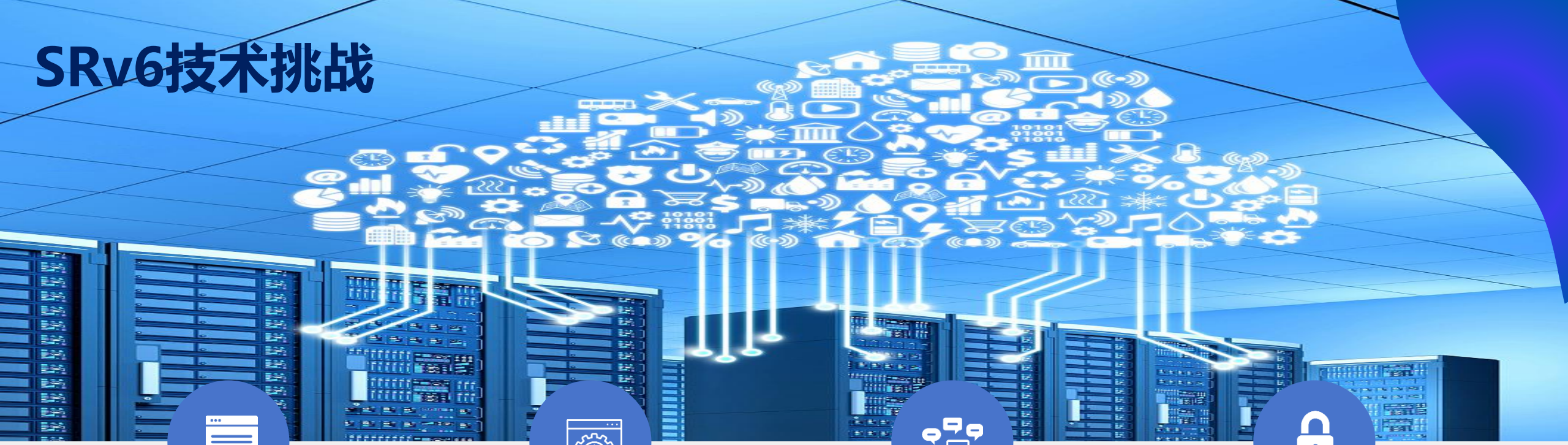
- 是指将 Segment Routing 技术应用于 IPv6 数据平面
- 通过在 IPv6 报文中插入一个路由扩展头 SRH （Segment Routing Header）
  - 在 SRH 中包含了由 IPv6 地址列表表示的 segment list
- 报文的目的地址将逐段的被更新，完成逐段转发

为什么要SRv6？：

- 可减少网络中实施的协议数量，从而降低运营支出（OpEx）
- 分段路由可原生支持网络可编程性，不但可以优化分布式计算场景下的网络性能，也可以无缝支持 NFV 环境；
- SRv6 同时支持 SDN、服务链和隧道，可简化 NFV 实施；

传统的 SRv6 实现方案主要依赖 CPU 以软件形式实现，额外消耗 CPU 算力！

SRv6技术挑战



可编程性

SRv6 技术在 IPv6 数据包中添加一个**可编程的扩展头**，实现网络数据平面的可编程；要实现对 SRv6 报文头的解析并做出相应的报文处理动作，就需要 HOST CPU 和软件的介入。**当前已有处理效率都相对较低！**

性能

SRv6 技术需要在 IPv6 数据包中添加一个扩展头，这会增加数据包的长度，增加额外的报文编辑的开销，从而影响网络的性能。

兼容性

SRv6 技术需要在 IPv6 数据包中添加一个扩展头，导致 SRv6 数据包无法被 IPv4 网络识别

安全性

SRv6 技术需要在 IPv6 数据包中添加一个扩展头，增加网络的攻击面

DPU 支持 SRv6 数据面卸载

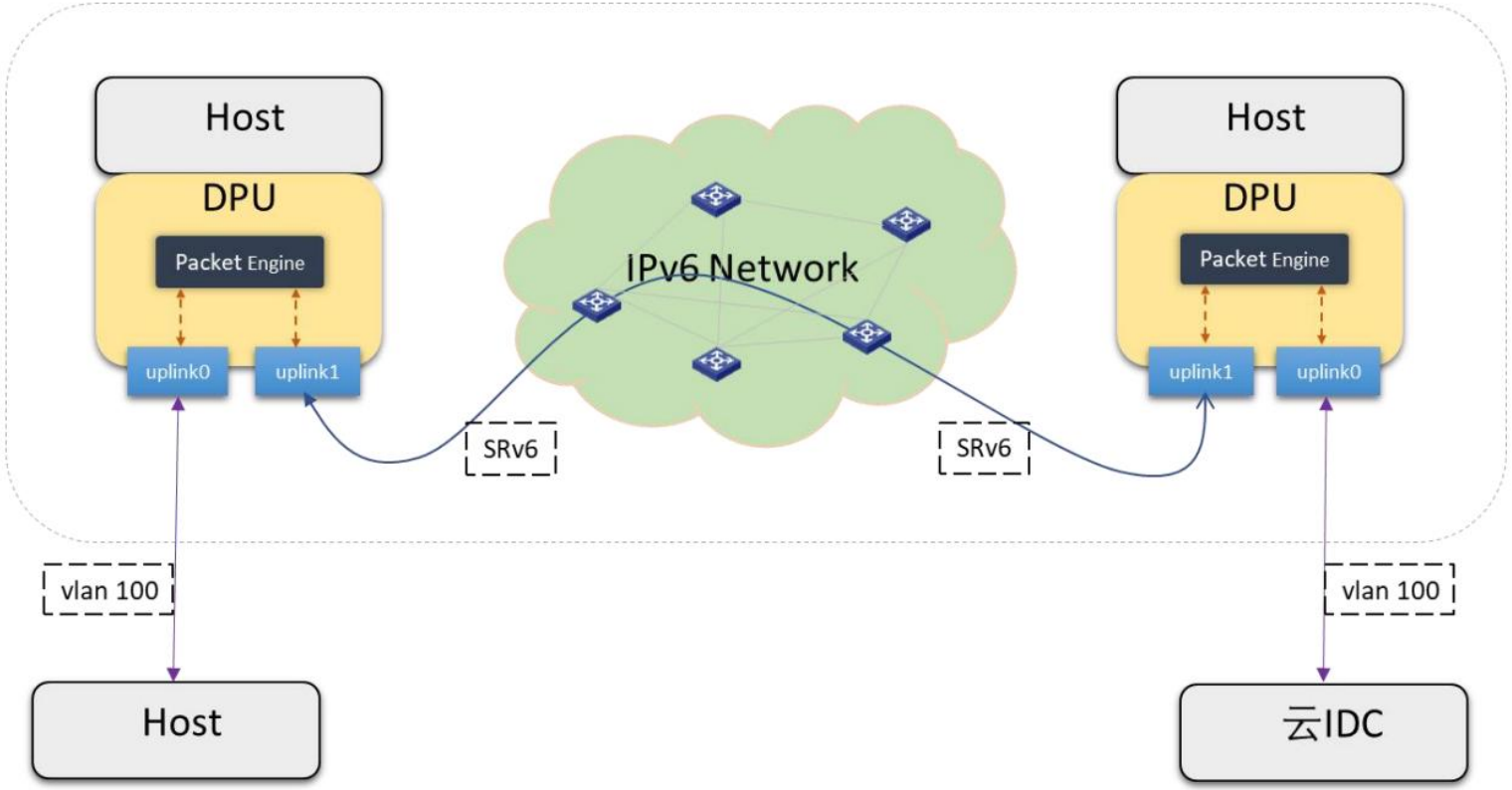
问题：如何高效实现对 SRH 扩展头的处理，以满足 SRv6 协议交互的要求?

回答：DPU!

DPU 内部包含“网络可编程硬件引擎”：

- 提供了灵活、多层次的可编程资源，以满足各种报文协议规范的处理需求
- 具有专用的端口管理表，支持 Port 级流量处理；
- 具有可编程报文解析器，支持用户自定义报文格式的解析；
- 具有模糊匹配以实现包分类，及多种报文匹配算法，例如：Range match、EM、 WC，etc.
- 具有专用的流量镜像、Per-Flow 级 QoS、多播广播数据流处理引擎；

DPU 支持 SRv6 数据面卸载



- SRv6 报文的处理完全被 DPU 中的网络可编程硬件引擎处理，不会引入 HOST 计算资源
- 并且由于 DPU 硬件引擎对报文的编辑、转发完全基于硬件逻辑， 处理效率远远高于依赖 HOST CPU 的通用计算能力。

DPU 支持 SRv6 数据面卸载

基于 DPU 的网络可编程硬件引擎 | SRv6 报文过滤 | Local SID | SRv6 Segment list 处理和编辑 | SRv6 报文封装解封装 | SRv6 转发 | VRF

从数据上来分析

- 通过 HOST 算力来实现 SRv6 报文的处理，通常**以内核形式**来处理 报文，处理效率依赖于 HOST CPU 的性能以及使用的 CPU 核心数量有上下的浮动
- 但是基本上内核的处理效率通常在每秒 10 万报文的处理量级变化，与 DPU 的处理能力相比还是**有较大的劣势。**

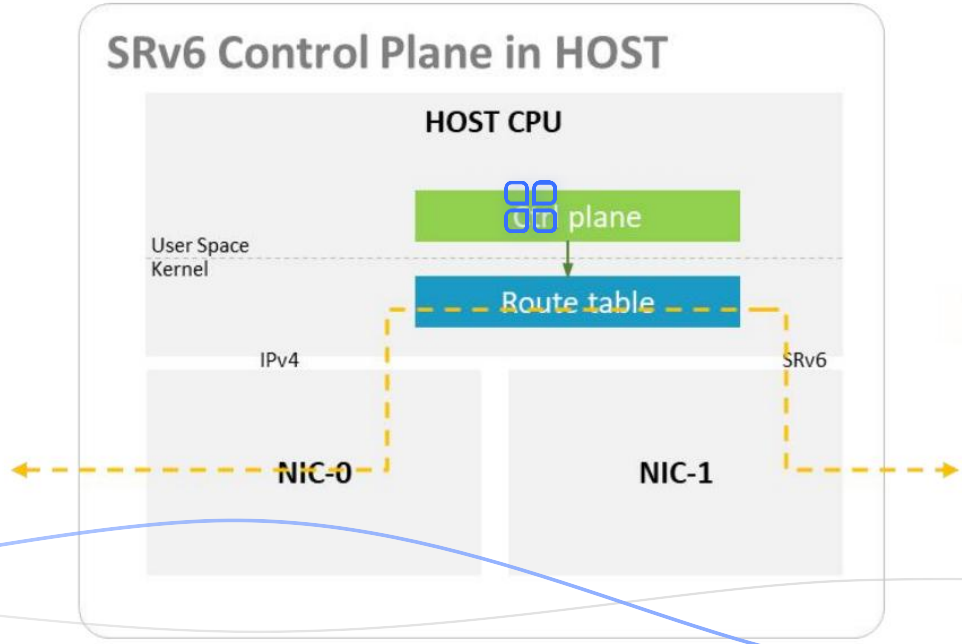
基于 DPU 的SRv6 数据面卸载

- 当前随着 DPU 接口带宽从 25G 向 100G->200G->400G 的快速发展， 报文的转发效率也达到了**每秒百兆报文**的处理量级。
- 综合 报文握手逻辑、协议逻辑处理等的实现，总体上与传统的以 HOST CPU+内核方式处理报文 的效率相比有 **100 倍级**的性能提升。

DPU 支持 SRv6 控制面卸载

传统的数据中心

需要在宿主机中部署、运行、维护大量且复杂的软件系统来完成IaaS的功能；



现在的DPU

可以卸载IaaS功能，实现统一的控制面管理。

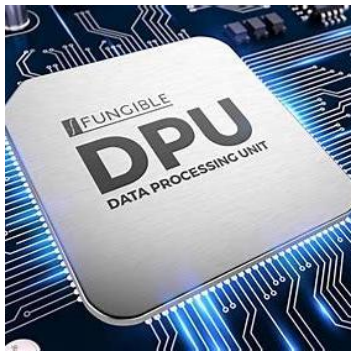
CPU Cluster
提供通用可编程能力

- 在 SRv6 控制面支持上，DPU 中基于通用可编程能力可以体现现有 SRv6 的控制逻辑 平滑的卸载到 DPU 内容，
- 由于 CPU Cluster 支持成熟的 OS 系统（例如：Linux OS、Euler OS 等等）使得软件的迁移不会引入过多的适配成本
- 并且卸载到 DPU 的 OS 上之后可以 充分获得 DPU 原生 SDK 的功能支持，优化对 DPU 底层硬件组件的直接调用和配置能力。

DPU 支持 P4 可编程

DPU的可编程网络处理引擎在硬件层面提供高效的灵活可扩展能力

- 软件层面可以通过P4语言将芯片上资源及硬件的可编程能力开放给终端用户
- DPU芯片通过 P4 可编程能力满足差异化的业务需求，可以真正实现将 SRv6流量在数据面上全卸载至芯片硬件处理
- 也就意味着对SRv6报文的解析，查找匹配报文编辑等过程全流程完全无需软件介入
- 完全消除CPU软件侧的资源开销的同时，可以百倍提升单节点的报文转发性能，并降低转发时延。



≈

单 DPU 节点的报文处理能力可以替代数百台 x86 服务器，真正实现数据中心降本增效！

CAICT 中国信通院

ACTIC 全国先进计算技术创新大赛

4 总结

DPU在算力网络中发展展望

例如需要面对“跨区域算力调度”，“区域内算力调度”的场景，提供电力与算力网络融合，高效合理调度全局算力和网络资源，助力 “双碳战略”的落地实施；

面对 AI 智能算力变化的训练推理场景，提供智能感知的能力将训练和推理任务调度至最优的智算网络中；

面对图像/视频渲染、多媒体闲时转码、切片服务、药物分析、大数据，科学计算、基因测序、金融和交易数据分析等计算类场景，提供闲散算力资源整合与灵活调配；

面对直播场景中，将连麦、转码、 渲染、弹幕、切片等业务的场景，提供基于全局负载均衡、统一算力调度策略的就近或质量最优的服务策略。

DPU在算力网络中发展展望

2023 年 6 月中国信通院首届算力互联互通大会

将 ODPU 作为算网云开源操作系统（CNCOS）项目 1.0 的子项进行了发布

在 DPU 管理、计算卸载、存储卸载、网络卸载、安全卸载和 RDMA 支持等方面提供通用软件开发框架和兼容性接口。

与此同时，中国信息通信研究院、中国通信标准化协会等部门和组织也在积极制定相应的标准

2023 年 10 月 9 日国家工信部联合六大部门发布的**《算力基础设施高质量发展行动计划》**

DPU 在提升算力高效运载能力方面起到重要作用

骨干网、城域网全面支持 IPv6，且 SRv6 等创新技术使用占比达 **40%**

大力支持和推广 DPU 相关应用，将加快推进我国数字经济建设步伐

CAICT 中国信通院

ACTIC 全国先进计算技术创新大赛

感谢观看！

A3-房银轰

- 队长：伊啸
- 队员：伊啸，李义
- 2024年8月17日