中国移动 研究院
China Mobile CMRI

全调度以太网技术架构白皮书

The Technical Framework White Paper of Global Scheduling Ethernet

（2023 年）

中国移动通信研究院

.

前　言

本白皮书面向未来智算中心规模建设和 AI 大模型发展及部署需求，联合产业合作伙伴共同提出全调度以太网（GSE）技术架构，旨在突破智算中心网络性能瓶颈，打造无阻塞、高带宽及超低时延的新型智算中心网络，助力 AIGC 等高性能业务快速发展。

目　录

缩略语列表

缩略语	英文全名	中文解释
AI	Artificial Intelligence	人工智能
AIGC	AI-Generated Content	人工智能生产内容
CPU	Central Processing Unit	中央处理器
DPU	Data Processing Unit	数据处理单元
ECMP	Equal Cost Multi Path	等价多路径路由
ECN	Explicit Congestion Notification	显式拥塞通告
FC	Fibre Channel	光纤通道
GPU	Graphics Processing Unit	图形处理器
GSF	Global Scheduling Fabirc	全调度交换网络
GSOS	Global Scheduling Operating System	全调度操作系统
GSP	Global Scheduling Processor	全调度网络处理节点
HoL	Head-of-line blocking	队首阻塞
JCT	Job Completion Time	任务完成时间
ML	Machine Learning	机器学习
PFC	Priority-based Flow Control	基于优先级的流量控制
PHY	Physical	端口物理层
PKTC	Packet Container	报文容器
RDMA	Remote Direct Memory Access	远程直接内存访问
RoCE	RDMA over Converged Ethernet	融合以太网承载RDMA
VOQ	Virtual Output Queue	虚拟输出队列
DGSQ	Dynamic Global Scheduling Queue	动态全局调度队列

1.背景与需求

目前，AIGC（AI-Generated Content，人工智能生产内容）发展迅猛，迭代速度呈现指数级增长，全球范围内经济价值预计将达到数万亿美元。在中国市场，AIGC 的应用规模有望在 2025 年突破 2000 亿元，这一巨大的潜力吸引着业内领军企业竞相推出千亿、万亿级参数量的大模型，底层 GPU 算力部署规模也达到万卡级别。以 GPT3.5 为例，参数规模达 1750 亿，作为训练数据集的互联网文本量也超过 45TB，其训练过程依赖于微软专门建设的 AI 超算系统，以及由 1 万颗 V100 GPU 组成的高性能网络集群，总计算力消耗约为 3640 PF-days（即每秒一千万亿次计算，运行 3640 天）。

分布式并行计算是实现 AI 大模型训练的关键手段，通常包含数据并行、流水线并行及张量并行等多种并行计算模式。所有并行模式均需要多个计算设备间进行多次集合通信操作。另外，训练过程中通常采用同步模式，多机多卡间完成集合通信操作后才可进行训练的下一轮迭代或计算。智算中心网络作为底层通信连接底座，需要具备高性能、低时延的通信能力。一旦网络性能不佳，就会影响分布式训练的质量和速度。

面向未来智算中心规模建设和 AI 大模型发展及部署需求，中国移动联合多家合作伙伴推出了全调度以太网技术方案（GSE），打造无阻塞、高带宽及超低时延的新型智算中心网络，助力 AIGC 等高性能业务快速发展。

2. GSE 网络架构介绍

2.1 总体设计目标

全调度以太网面向 AI、HPC 等高性能计算场景设计，架构设计遵循以下三大原则：

- ✧ 全调度以太网构建开放透明标准化的技术体系，供所有高性能计算生态涉及到的芯片（GPU、DPU、CPU 等）、设备（服务器、交换机、网卡等）、仪表、操作系统等上下游产业共同使用。
- ✧ 全调度以太网可适应多种高性能计算场景，凡是涉及到无损、高带宽利用率、超低时延需求的业务场景均可通用。
- ✧ 全调度以太网不是重造以太网，而是将高性能计算需求融入以太网，可最大限度地重用以太网物理层，兼容以太网生态链，如光模块、PHY 层芯片等。

2.2 整体架构概况

为打造无阻塞、高带宽、低时延的高性能网络，GSE 架构应运而生，该架构主要包括计算层、网络层和控制层三个层级，包含计算节点、GSP、GSF 及 GSOS 等四类设备。

2.2.1 GSE 整体架构

全调度以太网是具备无阻塞、高吞吐、低时延的新型以太网架构，可更好服务于高性能计算，满足 AI 大模型部署及训推需求。全调度以太网架构自上而下分为三层，分别为控制层、网络层和计算层，其中关键点在于创新的引入一种全新的动态全局队列调度机制。动态全局调度队列（DGSQ）不同于传统的 VOQ，其不是预先基于端口静态分配，而是按需、动态基于数据流目标设备端口创建，为了节省队列资源数量，甚至可以基于目标或途径设备的拥塞反馈按需创建。基于 DGSQ 调度以实现在整个网络层面的高吞吐、低时延、均衡调度。

- ✧ 控制层：包含全局集中式 GSOS，以及 GSP 和 GSF 设备端分布式 NOS。其中，集中式 GSOS 用于提供网络全局信息，实现基于全局信息编址（例如设

备节点 ID 等）、日常运维管理等功能。设备端分布式 NOS 具备独立的控制面和管理面，可运行容器的负载均衡、DGSQ 调度等属于设备自身的网络功能，通过设备分布式管控能力，提升整网可靠性。

- ✧ 网络层：通过 GSP 和 GSF 的分工协作，构建出具备全网流量有序调度、各链路间负载均衡、网络异常精细反压等技术融合的交换网络，是全调度以太网的主要实现层。其中，Fabric 部分可支持二层 GSF 扩展，以满足更大规模的组网需求。
- ✧ 计算层：包含高性能计算卡（GPU 或 CPU）及网卡，为全调度以太网的服务层。初期将计算节点作为全调度以太网边界，仅通过优化交换网络能力提升计算集群训练性能。未来考虑计算与网络深度融合，将 GSP 相关方案延伸到网卡层或者 GPU 直出网卡模块实现，与网络层进行联动形成算网协同的全调度以太网，进一步提升高性能计算性能。

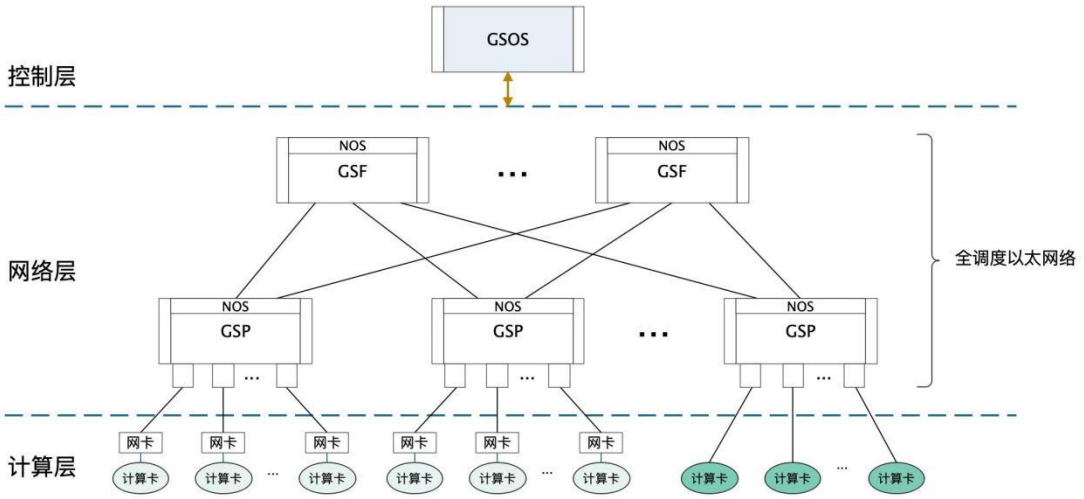

图 2-1 GSE 技术分层架构

2.2.2 GSE 架构设备

GSE 架构包括计算节点、GSP、GSF 及 GSOS 四类设备，各设备间协同工作，分工如下：

- ✧ 计算节点：即服务器侧的计算卡、网卡，提供高性能计算能力。
- ✧ GSP：网络边缘处理节点，用以接入计算流量，并对流量做全局调度；流量上行时，具备动态负载均衡能力。流量下行时具备流量排序能力。
- ✧ GSF：网络核心交换节点，作为 GSP 的上一层级设备，用于灵活扩展网络规

模，具备动态负载均衡能力，以及反压信息发布能力。

✧ GSOS：全调度操作系统，提供整网管控的集中式网络操作系统能力。

2.2.3 GSE 架构特性

考虑到AIGC 等AI/ML应用快速发展以及当前标准以太网规模部署现状，GSE架构应具备灵活可扩展性，并最大限度兼容以太网特性。

GSE架构特性具体如下：

✧ **灵活扩展**：支持万卡高性能计算集群部署，以 GSP +GSF 的两层网络为常用形态，支持横向扩容。当计算节点进一步扩大，两层网络架构不足以支撑时，可灵活扩展成 GSP +GSF+GSF 的三层网络架构，保留扩展到更多层 GSF 组网的能力，以满足业务部署需求。

✧ **生态开放**：秉持生态开放的原则，构建标准开放的技术协议栈，促成多厂家设备间的互联互通，共同构建全调度以太网的网络层，为大规模分布式计算提供高效的网络基础。

✧ **硬件通用**：所有网络节点均支持标准以太网，无需专用的信元处理节点，可与标准以太设备无缝切换。其中，GSP 和 GSF 设备虽然角色分工不同，但均以以太报文交换为基础，转发硬件具有通用性，设备角色可以由软件版本控制，从而支持更灵活的部署和维护。

2.3 关键技术特征

2.3.1 兼容以太网技术

以太网标准是当前普适性最好的通信标准之一，中国移动以通用开放的宗旨联合产业链共同打造 GSE 网络，最大程度兼容现有以太网标准，兼容性主要体现在如下几方面：

✧ 遵循现有以太网 PHY、MAC 层协议：遵循现有 IEEE 802.3 协议对以太网物理层、MAC 层的定义，以兼容现有以太网器件（含光模块、网卡、交换机等），将 GSE 以功能增量的形式融入到现有以太网中，对以太网进行增强。

✧ 完整的以太网业务报文传输：在整个 GSE 网络中，以完整以太网报文形式

进行传输，最大程度保留以太网报文承载内容的完整性，以便后续在 GSE 网络中兼容更多的特性，如在网计算。

- ✧ 遵循现有管控系统与运维习惯：管控系统、运维系统的构建与以太网转发技术一样复杂，且与转控平面的协同体系已成熟。GSE 网络最大程度上沿用现有管控及运维系统，做到架构不变、运维习惯不变，保证现有以太网的管理手段和运维手段的兼容继承。

2.3.2 无阻塞网络

随着网络规模的不断提升，报文交换从单网络节点内单跳到网络节点间多跳实现，各节点间也从松耦合关系变化为联合转发，业界通过 CLOS 架构搭建大规模分布式转发结构来满足日益增长的转发规模需求。该架构下，各节点分布式运行，自我决策转发路径，无法实现最优的整网性能。为使得大规模多节点转发效果和单节点一致，需要解决分布式转发结构内部的阻塞问题。

造成网络阻塞的核心原因是分布式转发结构中各节点无法完全感知全局信息，当一个网络节点发送给另一个网络节点时，无法感知下游节点网络情况，导致流量在下游产生拥塞。例如在基于 ECMP 进行负载均衡的网络中，网络节点仅站在自身视角将流量通过哈希选路发送，最终导致链路拥塞、出端口拥堵、交换网络利用率低等问题。DGSQ 技术是解决这个问题的关键技术，该技术将互不可见的网络节点通过与交换网全局队列映射联合起来，最终达到整网最优的转发效果。

2.3.3 提高有效带宽

基于 DGSQ 技术，可保证分布式交换网络入口节点发往交换网络的流量从出口节点看是最优的。但流量在网络中交换时，传统 ECMP 负载均衡会导致链路负载不均以及哈希极化，特别是在有巨型流存在的情况下，无论巨型流持续时间多长，所到之处均可能引起拥塞和丢包。当前交换网络缺乏有效的带宽控制和优先级管理，丢包将是无差别的，会给应用带来直接的负面影响。基于 Packet 的逐包负载分担技术，将任意流量转化成极短的数据单元传输，彻底消除哈希极化问题，进而提高交换网络的带宽利用率。

2.3.4 优化长尾时延技术

AI 大模型训练存在大量 Map-Reduce 流量模型，任意一轮计算的结束均依赖最后一个结果的返回，降低网络长尾时延可有效提升训练完成时间。交换网络整体转发时延和转发路径上中间节点的拥塞情况正相关，消除中间节点的拥塞就可消除长尾时延。DGSQ 调度和高精度负载均衡技术融合是解决该问题的关键，一方面，通过 DGSQ 的 PUSH+PULL 结合机制控制进入交换网络的报文数据量不会超过整网的转发容量；另一方面，通过高精度负载均衡的加持，双管齐下可以消除交换网络任一节点的拥塞。

3. GSE 网络核心技术

与传统以太网基于流进行负载分担的机制不同，GSE 交换网络采用定长的 PKTC 进行报文转发及动态负载均衡，通过构建基于 PKTC 的 DGSQ 全调度机制、精细的反压机制和无感知自愈机制，实现微突发及故障场景下的精准控制，全面提升网络有效带宽和转发延迟稳定性。

其具体流量转发流程如图所示：

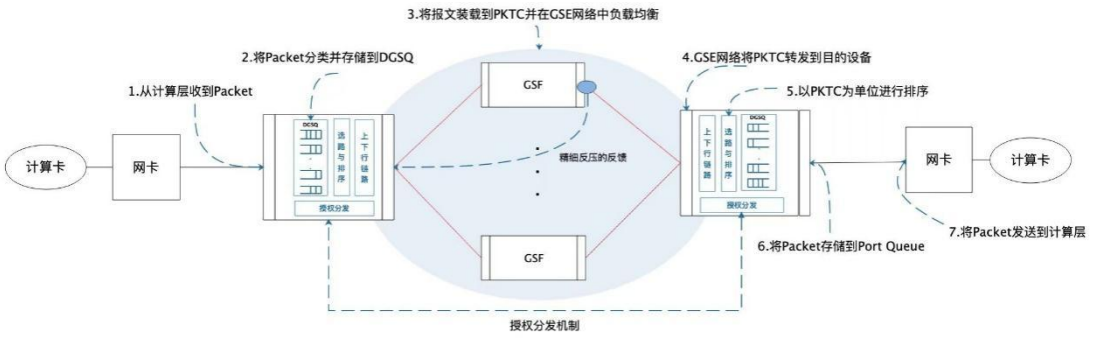

图 3-1 GSE 网络端到端流量转发示意图

(1) 源端 GSP 设备从计算侧收到 Packet 后，通过转发表找到最终出口，并基于最终出口按需将报文分配到对应的 DGSQ 中进行授权调度。

(2) 源端 GSP 设备获得授权后，Packet 将遵循 PKTC 的负载均衡要求，将报文发送到 GSE 网络中。

(3) 当报文到达目的端 GSP 设备后，先进行 PKTC 级别的排序，再通过转发表

将报文存储到物理 Port 的队列，最终通过端口调度将报文发送到计算节点。

3.1 PKTC 机制

PKTC 是区别于 CELL 转发的一种核心转发机制，该机制下以太网报文在逻辑上组成虚拟容器，并以该容器为最小单元在交换网络中传输。本节分将从 PKTC 概念、PKTC 开销和 PKTC 位置三方面进行阐述。

3.1.1 PKTC 概念

基于报文的转发在实现负载均衡时，首先需要克服报文长度随机产生的影响，因此需要对负载均衡的基本转发单元进行归一化处理，建立定长报文容器。报文容器可以容纳报文数量的设定可依据业务报文长度的分布情况进行调整，要求至少能够容纳 1 个最长的业务报文，且总长度在芯片转发能力和解乱序能力允许的情况下尽可能短，以达到精细切分数据流，充分提高瞬间负载均衡度的目的。

为解决上述问题，本方案提出报文容器的概念，设计原理如下图所示：

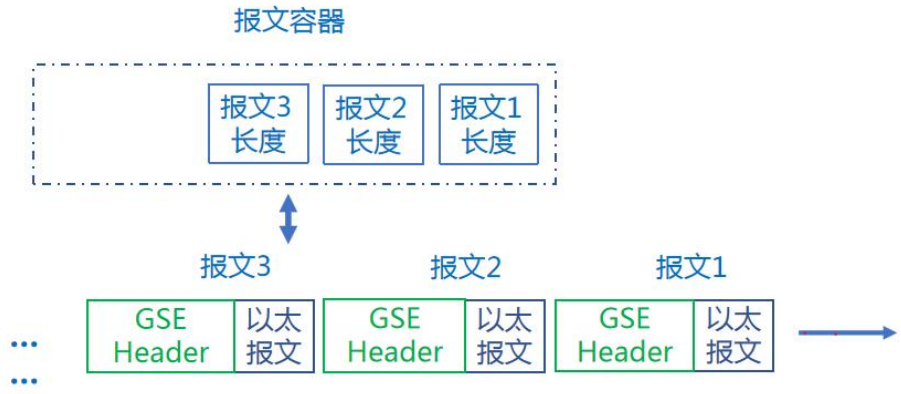

图 3-2 PKTC 转发机制示意图

报文容器的实现是逻辑虚拟的，当一个报文进入 GSP 节点时，GSP 节点将记录其归属的报文容器编号、在该容器中占用的字节数等信息，当报文字节数超过虚拟报文容器设定长度时，将该报文调度并纪录到下一个报文容器中。

GSE 网络各节点均直接转发报文，无需缓存报文构建实际容器。对于归属于相同报文容器内的所有报文，在交换网络中将被负载均衡到唯一路径进行转发，

以保证该报文容器内报文之间不再乱序，以降低出口 GSP 节点解乱序压力。

3.1.2 PKTC 开销

基于逐包的转发机制，需要在数据包中携带相关信息，才能被交换网正确识别处理并发送至目标节点。所以报文在进入 GSP 时需要区分 DGSQ，DGSQ 的标识和系统 DGSQ 建立目标有关。一般情况下，可基于源设备、目标端口以及在该端口下的优先级建立唯一的 DGSQ 标识。当然，也可根据业务需求简化 DGSQ 精细度，例如在一个目标端口下设置 4、2 或 1 个优先级，降低 DGSQ 队列的需求量，降低交换芯片开销。

进入 DGSQ 后的报文，需要经过下行调度授权才能被发送到交换网络中。此时，可将同一个入口 Leaf 节点发往同一个出口 Leaf 节点的报文组成一个解乱序队列，即在每个报文容器内的所有数据包添加相同的序列号（容器的序列）以及源 GSP ID，下行收到这些报文后，可基于源 GSP ID 和序列号进行解乱序处理。

下图以增加标准以太网报文头为例描述，其他内部以太报文构建方式下报文容器的构建和转发原理一致。

图 3-3 PKTC 头构建方式示意图

3.1.3 GSE Header 位置

GSE 网络需要对业务报文添加额外信息以用于全局负载均衡转发以及排序，这些信息有三种携带方式，包括：

✧ 在标准以太帧之外增加标准扩展头：这种携带方式最大的好处是不破坏原始

业务报文，但是在兼容性和传输效率上会有一定损失。如果为了提升以太网的兼容性而选择外加以太网 Tunnel 的方式，传输效率会进一步降低。

图 3-4 标准扩展头方式

✧ 重定义标准以太帧：重新定义报文的 MAC 头，这种携带方式的最大好处是传输效率高，但是兼容以太网能力较差，只有在特定场景下才可使用。

图 3-5 重定义以太帧方式

✧ 在以太网 MAC 或 IP 之后扩充协议头，这种方式的最大好处是平衡了以太网的兼容性和传输效率，但是网络中对 GSE 额外信息的处理会需要深入到报文内部信息，会影响转发时延。

图 3-6 协议头扩充方式

3.2 基于 PKTC 的负载均衡技术

为了减少并消除传统 ECMP 转发模型中出现的哈希极化、负载不均等问题导致的长尾时延或丢包，基于 Packet Container 的技术可以分为负载信息构建、动态路径切换、流量排序机制三个部分。

3.2.1 动态负载信息构建

对出端口负载信息的评估量化后，可随机选出负载较轻的链路之一，为后续流量的 PKTC 路径选择提供依据。如下图所示的转发模型，GSP1 作为接入交换机，当某段 PKTC 通过 GSP1 交换机去往 GSP2 的 A2 口时，需要对上行链路进行负载评估，以决策此段 PKTC 的传输出口。

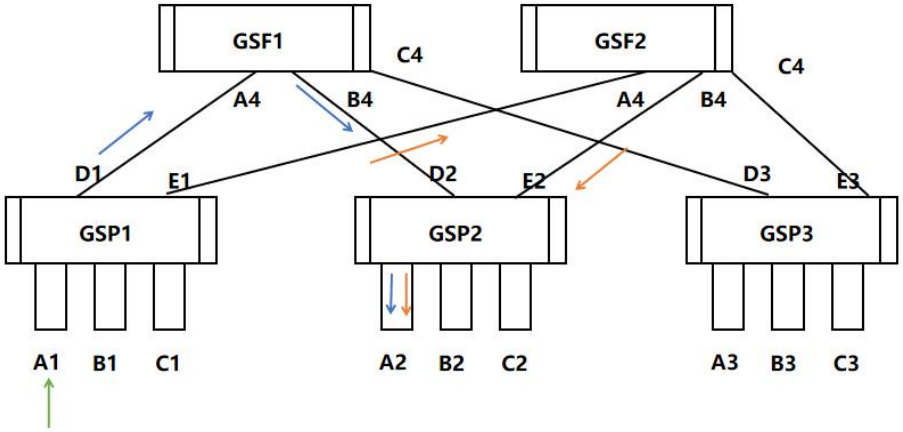

图 3-7 流量转发模型示意图

决策过程可以参考下图所示：在 PKTC 的路径选择上，先进行拥塞 Level 的选择，选择 Level 层级最低的出口集合，再从这些出口集合中随机选择一个出口，防止在多路径选择下存在同步效应。

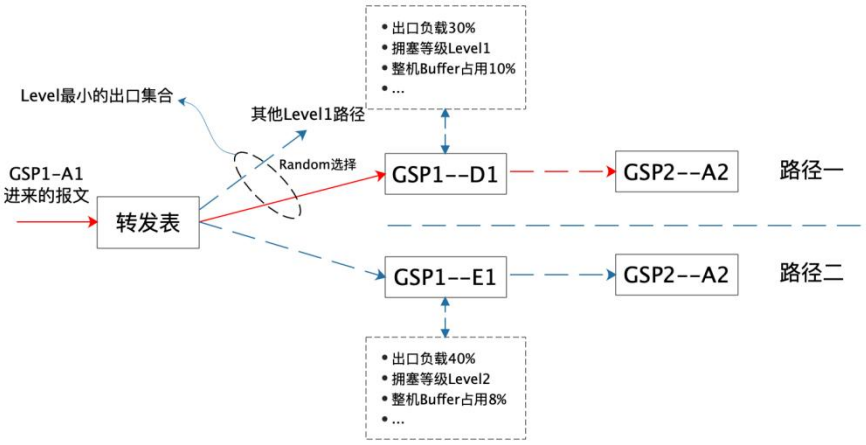

图 3-8 动态负载均衡决策过程

3.2.2 动态路径切换技术

当出口的负载出现动态变化后，每一个 PKTC 都可以按照算路算法进行路径的重新选择，以保证全局的负载均衡效果。在切换过程中，需要保证每个 PKTC 在路径选择上的一致性，否则会增加乱序程度，加大排序压力。路径选择仍按照先选 Level 层级，再随机选择出口的方式进行。

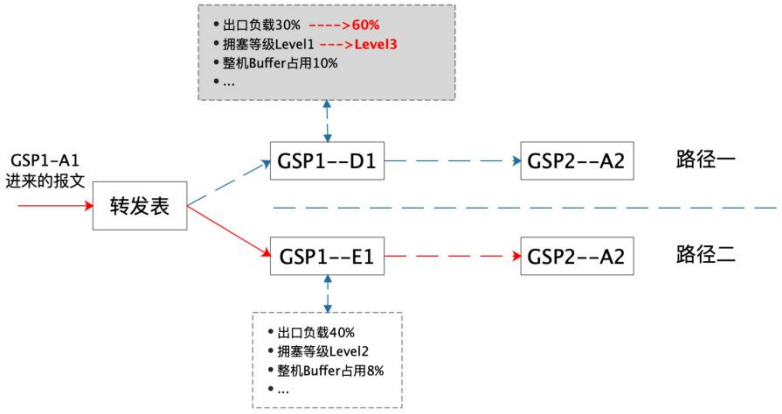

图 3-9 动态路径切换机制

3.2.3 流量排序机制

流量经过负载均衡和动态路径切换后，形成多传输路径。由于不同路径的传输时延存在一定差异，所以当不同路径的流量到达最终出口所在的节点时需要进行重排序处理。

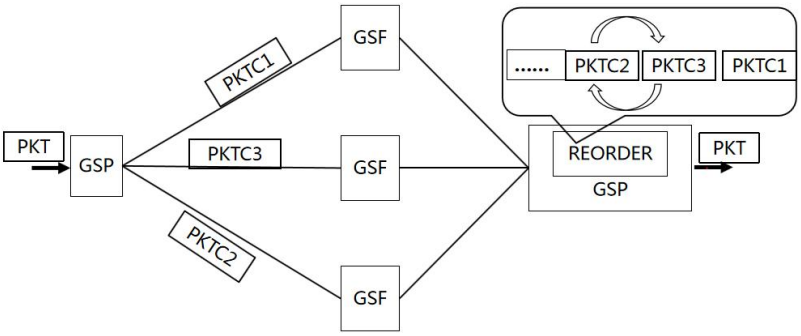

图 3-10　目的端流量排序机制

3.3 基于 PKTC 的 DGSQ 调度技术

网络传输中，常常会出现在某些时刻多个口打一个口的现象。如果这个现象是短暂的，在出口处可以通过一定的 Buffer 进行吸收；如果时间持续过长且多个入口的流量相加远大于出口的线速带宽，为了避免丢包，出口设备需启用反压机制保护流量，而反压一旦出现，网络的转发性能就会大幅度下降。

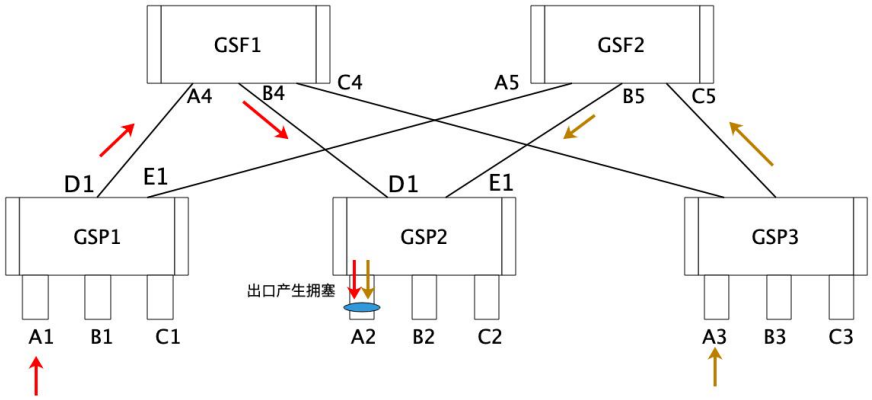

图 3-11 网络 Incast 流量发生场景

如上图所示，GSP1 的 A1 口和 GSP3 的 A3 口同时向 GSP2 的 A2 口发送流量，且流量相加大于 A2 的出口带宽，造成 A2 口出口队列拥塞。针对这种情况，仅通过负载均衡是无法规避的，需全局控制保证送到 A2 的流量不超过其出口带宽才可避免。因此，引入基于全局的转发技术和基于 DGSQ 的调度技术，才可实现全局流量的调度控制。

3.3.1 基于全局视图的转发技术

在传统数据中心以太网转发模型中，转发表以以报文携带的信息为主体，并且根据下一跳连接的出口，编辑报文头信息，如下图所示：

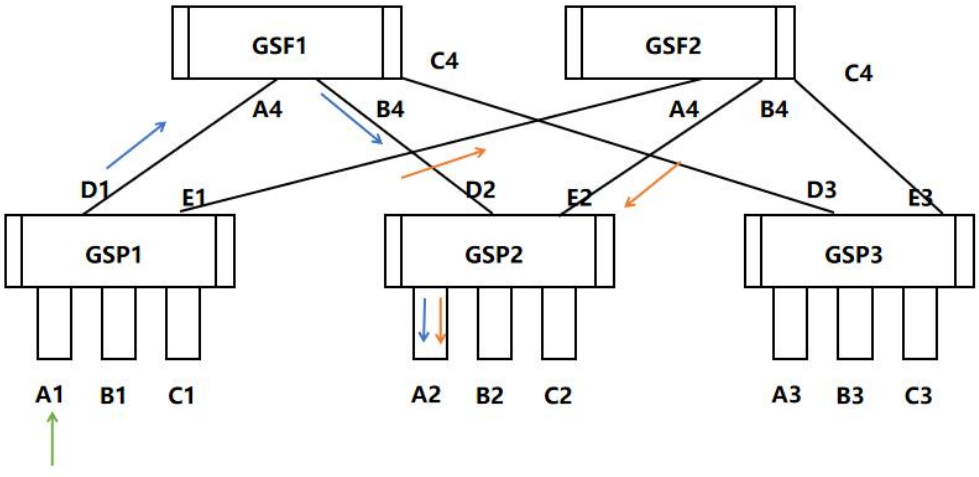

图 3-12 流量转发模型示意图

从 GSP1 任意端口进来的报文转发到 GSP2-A2 口，需要在 GSP1 上形成转发表及对应的出口信息，这些信息是本设备基于自身及相邻设备的状态形成，但对

后续路径上网络设备的状态既不感知也不控制，该方式无法构建无阻塞的全调度以太网。需要构建一种基于全局视野的转发技术，支持在接入交换机的转发表中指明最终目的，并通过端到端路径调度及综合化授权机制，动态形成负载分担信息并形成下一跳出口信息。

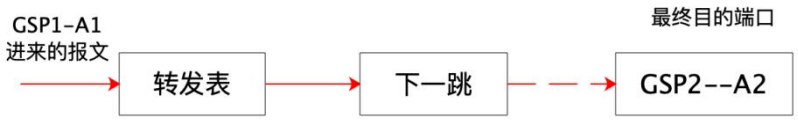

图 3-13 基于全局视图的选路机制

3.3.2 基于 DGSQ 的调度技术

基于 DGSQ 的全局调度技术如下图所示，在 GSP 上建立网络中所有设备出口的虚拟队列，用以模拟本设备到对应端口的流量调度。本设备 DGSQ 的调度带宽依赖授权请求和响应机制，由最终的设备出口、途经的设备统一进行全网端到端授权。由于中间节点的流量压力差异，GSP 去往最终目的端口不再通过 ECMP 路径授权权重选择路径，而是需要基于授予的权重在不同的路径上进行流量调度。通过这种方式，可保证全网去任何一个端口的流量不但不会超过该端口的负载能力，也不会超出中间任一网络节点的转发能力，可降低网络中 Incast 流量产生的概率，减少全网内部反压机制产生。

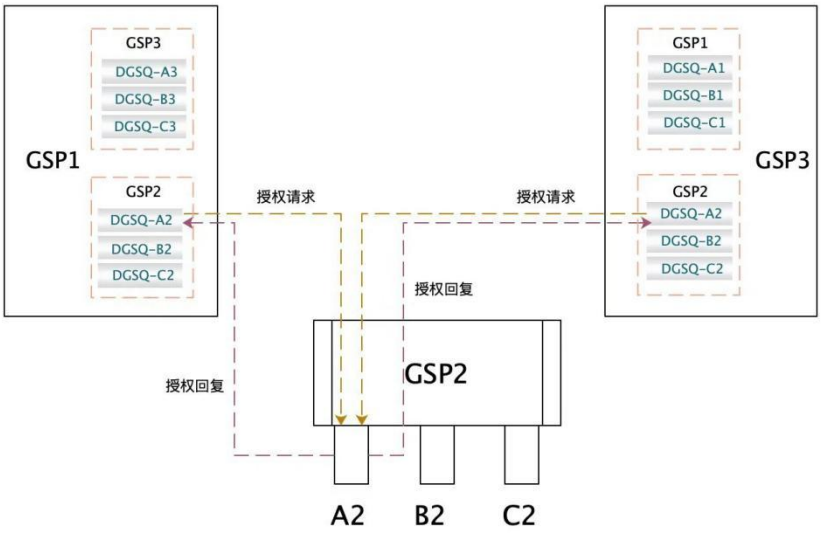

图 3-14 基于 DGSQ 的调度技术

3.4 精细的反压机制

基于 PKTC 的负载均衡技术和 DGSQ 全局调度技术在平稳状态下可很好得进行流量调控与分配，但在微突发、链路故障等异常场景下，短时间内网络还是会产生拥塞，这时仍需要依赖反压机制来抑制源端的流量发送。传统 PFC 或 FC 都是点到点的局部反压技术，一旦触发扩散到整个网络中，引起 HoL、网络风暴等问题。在全调度以太网技术中，需要有精细的反压机制来守护网络的防线，通过最小的反压代价来稳定网络的负载。

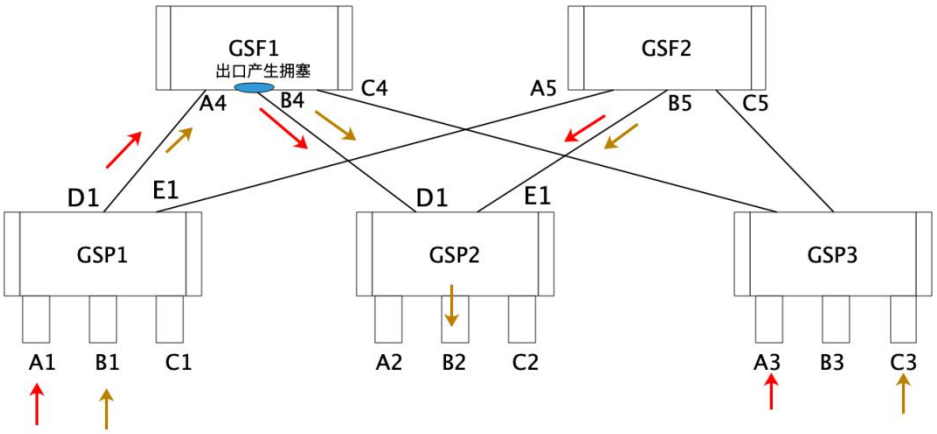

图 3-15 网络拥塞场景示意图

例如，如上图所示，GSF1 的 B4 出口出现拥塞，会降低甚至暂停对此端口的 DGSQ 调度授权。如果还有其他路径选择，将会触发采用动态负载均衡的方式切换到其他链路；如果当前网络中只有这一条链路，或者其他链路也即将处于拥塞状态，则不构成切换条件，此时需要启动反压机制。为了牺牲最小限度的流量保证整网流量的稳定，反压的范围需要控制得足够精确。例如只抑制去往 GSP2 的流量，去往其他设备的流量不受影响。更近一步的精细控制策略是通过 GSF1-B4 去往 GSP2 的流量被抑制，其他设备的流量不受影响。最终的精细程度将在后续的 GSE 标准中制定。

3.5 无感知自愈机制

全调度以太网架构中，通过全调度技术构建了入端口到出端口的虚拟队列路

径，对入端口的转发业务而言无需感知到出端口的每一跳路径，仅需要明确出端口即可。其对 GSF 组成的 Fabric 网络是无感知的，路径的可达性及切换由 Fabric 网络的负载均衡技术保障。

GSF 采用了基于 PKTC 的逐级负载均衡技术。当 Fabric 网络中的某条链路或某台 GSF 发生故障时，与其相连的设备节点能够实时感知到链路状态变化，并自动将相应链路从负载均衡备选列表中移除，回收 DGSQ 涉及此路径的调度授权，从而让 PKTC 分摊到其它可用链路。当设备或链路故障恢复后，相连设备节点同样可以实时感知到链路状态变化，并完成自愈。基于 PKTC 的负载均衡技术在以上链路切换过程中可以保持稳定的均衡性，不会像基于流的负载均衡受哈希结果或链路数量少的影响，可避免某条链路负载突发叠加的情况。

3.6 低时延转发技术

转发面主要通过简化、并行化和旁通转发流程等手段降低设备内转发路径的时延。随着端口速率的不断提升，高速信号完整性的挑战也越来越大，需要不断引入更为强大的 FEC 算法（FEC，forward error correction，前向纠错）。FEC 越强大其编解码复杂度也越高，所增加的时延也越大，100G 以上速率 FEC 所占用的时延已经达到整体转发时延的 20%左右。

FEC 的过程又可以分为检错逻辑和纠错逻辑。在低速的 FEC 处理中往往没有做上述流程的区分，但随着速率提升、检测及纠错逻辑的复杂，细分差异化处理会变为越来越有意义。检错和纠错分离技术可提前校验数据块内是否存在误码。在无错情况下，可旁路 FEC 译码流程，消除无错场景下 FEC 收帧和译码时延，降低无错情况下的接口时延，消除高增益 FEC 码字的时延弊端；有错的情况下，才进一步进行纠错处理。因为发生误码的概率毕竟远小于无误码，此方式可以优化端口的平均转发时延。灵活 FEC（FlexFEC）技术可以根据链路的误码率状态，自动选择合适的 FEC 纠错算法，以便在保持可靠性的同时提供低延迟。

3.7 全调度以太网操作系统

全调度以太网的 GSOS 综合考虑了分布式 NOS、集中式 SDN 控制器的优势，分为全调度控制器、设备侧 NOS 两大部分，同时采用带内的带内管理通路。

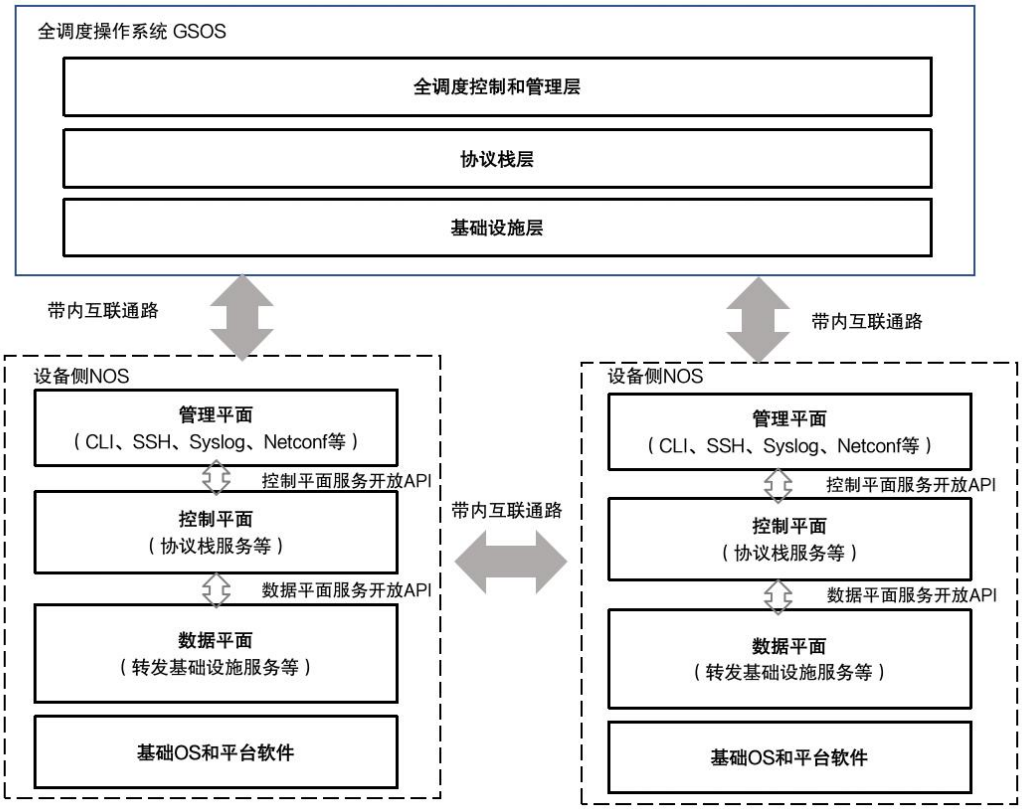

图 3-16 全调度以太网操作系统架构

- ✧ 设备侧 NOS：GSP 和 GSF 的盒式设备支持独立部署 NOS，并构建出分布式网络操作系统。每台 GSP 和 GSF 具备独立的控制面和管理面，可以运行属于设备自身的网络功能，提升系统可靠性，降低部署难度。分布式 NOS 可以将单点设备故障限制在局部范围，避免对整网造成影响。为了提供开放性服务并支持全调度以太网特性，NOS 还将传统一体化的网络功能服务分层解耦成控制平面服务、数据平面服务，并开放服务接口。数据平面服务的开放性为全调度以太网的部署提供了更大的灵活性，例如可以与控制器配合建立全调度 DGSQ 系统、根据网络规模或软件实现情况来选择合适的分布式或集中式发现同步协议来建立 Fabric 互联网络等。
- ✧ 全调度 GSOS：集中式 GSOS 提供了更好的网络全局信息，简化基于全局端口信息的 DGSQ 系统的建立和维护。同时 GSOS 也是整网运维监控的大脑，可协同设备实现对实时路径、历史的记录及呈现以支撑网络运维。
- ✧ NOS 控制管理通路：得益于全调度以太网架构的兼容性原则，网络的 GSF 节点也可以支持以太网报文交换特性。这样可将管理和控制平面统一到数据转发平面，形成带内（In-band）互联通路，并在 Fabric 互联的数据转发平面

中预留内部高优先级通道，以保障控制管理通路的优先级。全调度以太网不再采用带外（Out-band）控制管理通路，而是统一到带内通路，便于运维管理，避免维护两套物理网络。

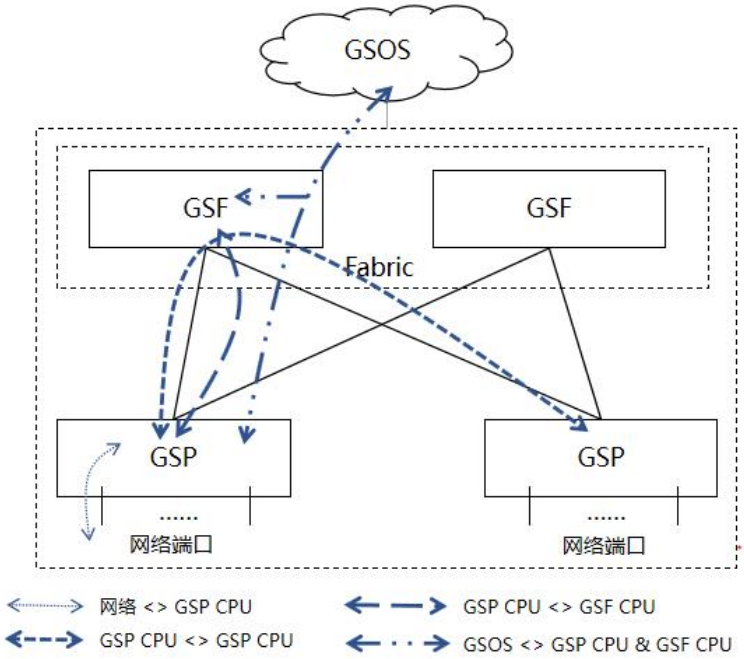

图 3-17：带内模式的控制和管理通路

4. 组网应用展望

GSE 面向无损、高带宽、超低时延等高性能网络需求业务场景，兼容以太网生态链，通过采用全调度转发机制、基于 PKTC 的负载均衡技术、基于 DGSQ 的全调度技术、精细的反压机制、无感知自愈机制、集中管理及分布式控制等技术，实现低时延、无阻塞、高带宽的新型智算中心网络。该技术架构落地时，存在两种方式，一种是仅在网络侧运行该架构，一种是端到端均运行该架构。

✧ 仅在网络侧运行该架构

GSE 本身可以支持网卡侧无感知的组网解决方案，若网卡侧有能力参与协同，则可以更精细化地提供端到端的全调度特性。GSP 设备的 DGSQ 队列可以将其状态反馈给网卡侧，供网卡或业务感知网络状态，从而进行更好端网协同。例如，当 GSP 的某个 DGSQ 队列达到一定水线时，表明其对应计算节点到对端计算节点的网络流量存在拥塞情况，此时 GSP 可以将该信息通过反压机制反馈

给对应网卡，网卡或者业务侧可以根据该信息适当地调整往这个对端计算节点的发包速率，从源头上避免可能的拥塞恶化或丢包情况。

✧ 端到端均运行该架构

将 GSE 的功能在网络组建中重新分工，原有机制不变，网卡或者 GPU 的网卡模块实现授权分发和反压响应，交换机依然集成基于 PKTC 的负载均衡选路、流量排序、精细反压信息产生以及最基础的基于全局的转发控制。这样在原有 GSE 组网模型和功能不变的情况下，利用网卡最接近业务侧的优势，可从业务源头调度流量。

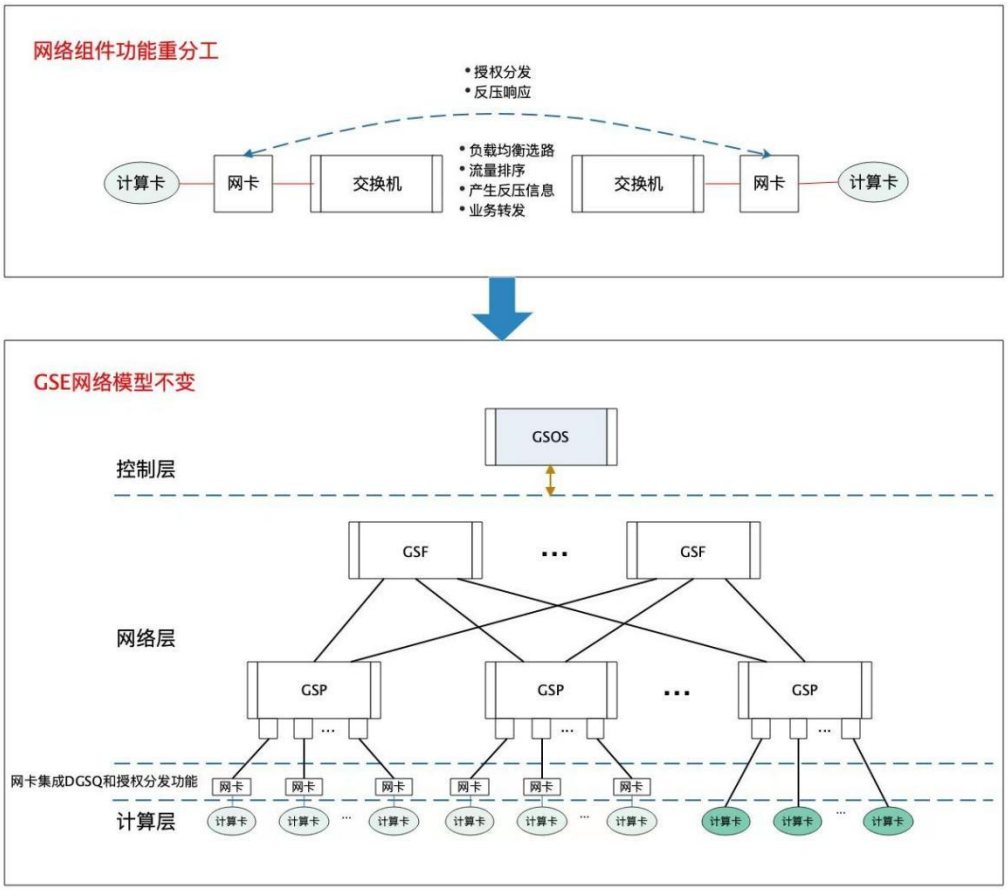

图 4-1 GSE 技术后续演进方向