

## 以竞赛速度构建 AI 网络：借助多路径可靠连接（MRC）构建协调式 fabric

人工智能（AI）并没有击垮网络，它改变了竞赛的规则。

传统以太网围绕相互独立的流量流设计，很像车辆在高速公路上行驶。AI 工作负载则根本不同：成千上万个图形处理器（GPU）作为一个同步系统运行，其中延迟、拥塞或报文投递上的微小变化，都可能影响整个计算的性能。

因此，AI 网络管理不太像管理高速公路交通，更像是控制一场一级方程式（F1）赛车。胜负不仅仅取决于速度，更取决于协调、时机、态势可见性，以及在条件变化时快速反应的能力。正因如此，多路径组网、拥塞信令、遥测、自动化与智能传输机制，正成为现代 AI fabric 的关键组成要素。

未来的 AI 网络，不会仅由它提供多少带宽来定义，而将由它在整个系统中协调流量的智能程度来定义。

### 真正的约束是协调，而非带宽

一个常见的假设是，AI 网络主要是一个带宽挑战。这并不正确。

现代网络已经能够提供巨大的吞吐量。更大的挑战在于：当成千上万个 GPU 产生高度同步的通信模式时，如何维持可预测的性能。AI 工作负载会产生独特的网络行为，包括：

- Incast 拥塞：多个发送方同时向同一目的地发送；
- 队列堆积：同步流争用共享资源；
- 尾延迟放大：最慢的通信路径拖累整个工作负载；
- 同步停顿：GPU 在等待其他参与者完成后才能继续。

等传统基于流的机制（如等价多路径 ECMP）仍是最重要的基础。然而，大规模 AI 训练的流量会暴露出一些情形：仅依赖流级哈希，可能无法在高度同步的需求下充分利用现有 fabric，或无法维持可预测的性能。

这正是多路径可靠连接（MRC, Multipath Reliable Connection）的动机之一。MRC 是由开放计算项目（OCP, Open Compute Project）交付的一种架构，它在保留 RoCEv2 传输的同时，引入了报文喷洒（packet spraying）与源驱动的拥塞缓解等技术。MRC 通过改进流量在多条网络路径上的分发与协调方式，旨在提升 fabric 利用率，并为大规模 AI 工作负载提供更可预测的性能，同时为现有 AI 网络提供一条实用的迁移路径。

挑战已不再只是更快地搬运报文，而是让系统的所有部分协同前进。在 F1 赛道上，问题不在于单车的速度；当整个车队在同一时刻抵达同一个弯道时，挑战才出现——即便是最快的赛车，也需要协调才能取胜。AI 流量也是如此。

## 报文喷洒：利用整条赛道

报文喷洒代表了超越传统单路径转发的演进。传统网络通常将一个流分配给一条由 ECMP 选定的路径。这对许多应用效果良好，但来自 AI GPU 的流量常常产生同步的流量模式，多个大流争用相同资源。报文喷洒将来自某个工作负载的报文分布到多条可用路径上，从而带来超出传统基于流转发的路径多样性。

报文喷洒的收益包括：

- 通过更均衡的流量分布提升 fabric 利用率；
- 更好地应对同步流量突发；
- 减少拥塞热点；
- 在单条路径状况变化时提升韧性。

回到 F1 类比：设想一个车队代表来自单个 AI GPU 的流。在传统网络中，每辆车都走同一条赛车线；一旦该线拥塞，整个车队都会减速。报文喷洒让车队可以使用赛道上的多条赛车线，通过更有效地利用可用空间，避免局部瓶颈并提升整体性能。

部分实现通过动态负载均衡扩展这一概念，在转发决策中纳入额外的路径信息。然而，用上赛道上的每条车道只是第一步。来自 AI GPU 的流量还需要能够维持可靠性、并随状况变化而自适应调整的机制。

## SRv6：连接多条赛道

SRv6 是 MRC 架构的基础组成，因为它提供了跨多个网络平面的确定性路径引导机制。它是对报文喷洒的补充：使流量能够在多条可用路径上转发，而更上层的流量工程则决定使用哪些路径。

用 F1 类比来说，报文喷洒让赛道的每条车道都被高效使用，而 SRv6 将多条赛道连接成一个协调的赛车综合体——它提供赛道之间的路线，使流量能够作为单一 fabric 跨越多个网络平面。

随着 AI 基础设施持续演进，这些能力可以很自然地超出单一数据中心，用同一套基于策略的转发模型连接分布式的 AI 工厂。但当下，其直接价值在于实现 AI fabric 内部跨多个网络平面的高效、确定性连接。

## 多路径协调：将多条路径变成一个系统

使用多条路径带来了新机会，也引入了新挑战。网络必须确保流量在路径状况变化时，依然保持可靠、高效与自适应。这正是 MRC 的另一个要素变得重要的地方。MRC 通过实现跨多条路径的可靠投递，并允许传输端点在随状况自适应时纳入网络反馈，来增强多路径通信。

其目标是使 AI 工作负载在经历以下情况时仍高效运行：

- 拥塞；
- 报文丢失；
- 瞬时网络事件；
- 变化的路径状况。

用 F1 术语说：报文喷洒打开了整条赛道，MRC 则提供了让每位车手在保持同一场比赛一部分的同时自适应调整的协调能力。目标不是找到一条完美路径，而是把整个 fabric 当作一个协调的资源来使用。

## 熵：保持流量分散

只有当流量保持分散时，多路径才能发挥全部收益。这正是流量熵（entropy）概念的重要之处。更高的流量熵会在流量中制造更多的无序、不确定或复杂，从而带来整个网络更高的利用率。

在赛车中，即便是最快的赛车也会自然地汇聚到同一条最优赛车线上。当整个车队都选择同一条线时，拥塞就会出现——不是因为赛车缺乏能力，而是因为所有人都在争用同一有限空间。AI 流量同理。

高度同步的 GPU 间通信会无意间将流量集中在可用路径的一个子集上，造成热点，而其它资源却未被充分利用。维持有效的熵，可确保流量持续利用整个 fabric。

这需要传输层与网络层通过以下技术协同：

- 传输级变化；
- 流标识符；
- 路径多样性机制；
- 多路径转发。

核心原则很简单：当传输层与网络层协同以维持分布时，多路径的效果最佳。

## 协调式传输：在拥塞发生前做出反应

使用多条路径只解决了一部分挑战。来自 AI GPU 的流量，还需要目的端点和网络在状况演进时做出一致的响应。

在 AI 规模下，成千上万个端点与网元可能同时通过许多路径通信。若无协调反馈，各自独立的反应会导致利用率低下、流量行为不稳定以及不可预测的应用性能。

协调式传输模型允许端点纳入网络反馈，并智能地响应变化中的状况。示例包括：

- ECN（显式拥塞通知）：在报文丢失发生前提供早期拥塞通知；
- SACK 与 NACK：改进丢包检测与恢复；
- 选择性重传：仅恢复缺失的信息；
- 报文修剪（packet trimming）：在拥塞事件中提升恢复效率；
- 附加信令机制：提供对网络状况的感知。

ECN 是最清晰的例子之一。传统拥塞响应往往在报文被丢弃之后才开始；ECN 则在队列开始堆积时通过标记报文提供更早的拥塞指示，使端点能够在拥塞影响性能之前调整。回到 F1 类比：ECN 是黄旗，而非撞车——车手从赛道获得预警，并在事故发生前调整。同理，ECN 可以主动反应，而非等到性能已经下降。

报文修剪遵循同样的主动理念：在拥塞时不丢弃整包，而是转发其中有用的部分，仅要求重传缺失的数据。这减少了恢复流量，并帮助 AI 工作负载在负载下维持吞吐。在 AI 规模下，拥塞是可预期的；目标不是消除它，而是尽早检测、高效传达，并以最小中断恢复。

## 更贴近数据通路的智能

随着 AI 网络持续扩展，另一个新兴行业趋势是将更多智能移近芯片（silicon）。现代 ASIC 已超越简单的报文转发。新的硬件能力能够对瞬时状况做出更快响应，例如：

- 链路临时劣化；
- 短时拥塞事件；
- 对延迟敏感的中断。

用 F1 术语说，这就像把反应时间从维修站墙直接移入赛车内部——因为智能更靠近事件发生处，网络能够更快地响应。这些能力与更高层的传输与控制机制互补，而非取代它们。

方向是明确的：AI 网络正变得更快、更具适应性、更分布式。

## 可见性：维修站墙的优势

没有遥测，就没有 F1 车队能取胜。赛道上的每个决策都由来自赛车的实时信息驱动。AI 网络遵循同样的原则。运营商需要对以下方面保持持续可见：

- 队列行为；
- 拥塞状况；
- 路径利用率；
- 工作负载同步；
- fabric 整体健康度。

可见性不再只是理解“发生了什么”，而是理解“当下正在发生什么”。Nokia 构建了一种架构：源自 Nokia SR Linux 网络操作系统（NOS）的流式遥测，由事件驱动自动化（EDA, Event-Driven Automation）平台处理，将网络可见性转化为运营优势。

遥测提供信息，自动化提供响应。二者共同为 AI 规模网络创建了一个闭环运营模型。

## **MRC 的承诺与协调式 AI fabric 的新架构模型**

Nokia 认为，下一代 AI fabric 将结合：

- 报文喷洒与多路径组网，以最大化可用资源利用率；
- SRv6，以提供跨多条路径的确定性报文路由；
- 传输协调与反馈机制，以提升可靠性与自适应能力；
- 遥测与自动化，以提供可见性与闭环运营；
- 硬件辅助智能，以改善响应时间。

AI 网络关乎这些能力以及其它能力如何协同工作。用 F1 术语说：

- 报文喷洒利用整条赛道；
- SRv6 在单一协调环境中跨多条路径喷洒报文；
- 多路径协调让每辆车协同前进；
- 遥测提供对变化状况的感知；
- 硬件智能改善反应时间。

未来的 AI 网络将更快地转发流量，并协调整个系统。因为在 AI 基础设施中，竞争优势并非来自单一技术，而是来自在一切都以极限运行时，让整个系统保持稳定、可预测且同步。