# 探测器：数据中心网络的拓扑感知监控系统

今天，数据中心已经成为企业信息化、行业信息化和城市信息化的核心所在。随着国家的经济发展和社会管理对信息系统依赖程度的提高，特别是在众多重要行业的数据大集中后，信息系统的安全性、可靠性、可用性日益突出

1. **问题背景**

在大型数据中心网络中，如何排除网络性能问题是一项艰巨的任务。今天的大型数据中心，都会托管各种服务，例如搜索引擎，社交网络和文件共享。 为了高质量的支持这些服务，经过精心设计的数据中心网络（DCN），可以有效地将数千个网络设备连接在一起。例如64位Fattree 结构的DCN拥有超过6万台服务器和5000台交换机。 然而，由于网络规模庞大，频繁升级和管理复杂性，在DCN中，像路由误差、链路故障等问题逐渐成为了常态。为了保证提供给用户高质量的服务，要首先检测并消除会给用户带来直观感受的故障，如丢包和高延迟问题，才不会引起用户的投诉。

为了检测并清除故障，需要在数据中心使用一套高效的网络监控系统。

1. **现有的系统**

过去几年提出了一些方法：被动的监控方法，通过SNMP查询设备计数器或在用户感知到某些问题时通过设备CLI检索信息；使用主动监控方法的系统有Pingmesh ，NetNORAD 。

**2.1被动监控方法**

被动监控方法存储定期从SNMP或CLI轮询得到的交换机计数器上的数据包统计信息。 如果某段链路坏掉，则交换机计数器将显示大量的分组丢失。 但是网络中会出现设备未检测到或忽视的故障，或者设备发生故障，而不是链接失效。如果出现了这种灰色故障，则可能不会被系统检测到。 例如，由于各种原因导致无声数据包丢失时，交换机不显示任何数据包丢弃提示，因此SNMP数据不是完全可信赖的。 此外，交换机计数器可能是嘈杂的，因此通过此方法识别的问题可能会导致也可能不会导致用户感知到端到端延迟或丢失。

**2.2主动监控系统**

为了弥补这种缺陷，一些系统通过发送端到端的探测器来检测灰色故障。

GUO 等人提出的Pingmesh和LAPUKHOV 等人提出的NetNORAD采用端到端探测方法来测量网络延迟和数据包丢失。 Pingmesh通过在DCN内构建两个完整的图来选择探测路径：一个包括同一ToR交换机下的所有服务器，另一个则涵盖了所有的ToR交换机。 NetNORAD类似于Pingmesh，但是将其放在几个pod（交付单元）中，而不是所有的服务器。 他们的方法简化了设计，但带来了非常显著的开销。 虽然可以捕获灰色故障，但是在DCN中采用ECMP时，很难检测到导致链路速率降低的故障：一对服务器之间有很多路径，在特定的链路发生低速率损耗可能不会影响两台服务器之间的整体端到端丢失率。

由于ECMP导致一对服务器间有很多路径，Pingmesh或NetNORAD不知道探测器采用了端到端的哪些路径，所以无法使用他们确定故障的具体位置。 因此需要借助Netbouncer 和fbtracert 等其他工具，通过发送额外的探测器来回溯故障。 这些工具可能无法确定由短暂的位错误，非原子规则更新或网络升级引起的故障导致的暂时错误。 为了确定这种故障，需要把检测和定位紧密耦合起来，就可以只根据检测数据来定位故障，而不是在发现故障后增加额外的探测器。 这种耦合可以进一步实现近实时故障定位。

**2.3 现有系统的局限性**

第一、现有的监控系统很难覆盖到全部的故障检测范围。 传统的被动监控方法，可以检测到链路故障和线路卡故障等故障并清除，被动监控方法面对这种灰色故障就显得力不从心了。由于数据中心中使用了等价路由技术，主动监控系统也很难捕获到链路速度降低的故障。

第二、主动探测系统在每对服务器之间发射探测器，增加了带宽的开销。此外，为了覆盖到所有可能的路径，也需要加入大量的探测器。

第三、现有的主动探测系统在获知了网络中发生的故障后，无法准确定位故障发生的具体位置，仅仅会告知给管理员可疑的源-目的服务器对。管理员需要通过借助额外的工具来找出故障发生的具体位置。由于故障转瞬即逝，管理者很难快速找出故障并解决。 因此，这种检测与故障分离的方式可能会花上很长时间去解决故障。

因此DCN中期望的监控系统应满足三个目标：全面覆盖的故障检测，较低的带宽开销和实时故障定位。

1. **相关工作**

探测器设计：许多现有的工作如Dhamdhere等人提出的Netdiagnoser利用多播协议、Castro等人提出的Network tomography利用交换机上的日志、Sharma等人提出利用网络编码进行网络探测。Chen等人提出了代数方法进行网络监控、Guilbaud等人提出的大规模丢包定位方法都是针对所有链接的失效率进行探究，而不是发生故障的链路。 Zeng等人和Nicolas等人提出了由于可扩展性而不适用于DCN的骨干网络的监控解决方案，主要区别在于探测矩阵设计。

故障定位：我们准确识别错误链接的目标正是二进制网络层析成像的领域。 由于它们的问题规模和损耗特性，诸如Sherlock ，Tomo ，GREEDY ，SCORE 和OMP 等故障定位算法对于DCN来说并不奏效。

DCN监控： Pingmesh 和NetNORAD 等先进监控系统在故障探测的设计上采用了故障检测和定位分离的方案，而为了以较低的开销实现实时排除故障，必须把检测和定位紧密耦合起来。针对这一目的，Liu等人提出的LossRadar 是基于交换机的解决方案，但它需要可编程的交换机。 Sigelman等人提出的Dapper方案和开源的Zipkin 是基于分布式跟踪系统，通过收集时序数据来分析故障源。

1. **主要挑战**

**4.1 探测路径**

现有监控系统的主要局限性在于无法选择最优的探测路径，导致无法收集足够的有用信息，需要借助额外的探测器来重现故障并定位。 因此我们面对的问题是如何仔细选择探测路径以克服这种限制。

给定DCN路由矩阵R，如何选择一组路径构建探测矩阵P，使得P同时满足

（1）最小化路径数

最小化探测路径的数量有利于最小化网络带宽消耗和分析开销，这样我们有可能在短短的几分钟内就可以完成探测并诊断整个DCN。 在相同的探测带宽预算下，它允许每个探测器更频繁地探测同一组路径。

（2）α-覆盖

α-覆盖要求每个链路至少被探测矩阵中的α条路径覆盖。因为链路有可能被来自多个探测器的探测路径覆盖，多次覆盖链路可以提高故障检测的统计准确性，以及更好的探测器故障恢复能力。

（3）β-识别能力。

β-识别能力表明DCN中任何β条链路同时发生故障可以被正确定位。更好的识别能力有助于更高的损失定位精度。

我们发现该问题对于一般的DCN是NP-hard，因为最小集合覆盖问题是该问题的一个特殊情况。

**4.2 故障定位**

故障定位的问题是：给定端到端分组丢失观察结果，找出导致这个结果的错误链路的最小集合。 这个问题有着两个巨大的挑战：

（1）更大的问题规模

我们的研究重点是大规模DCN网络，不同于在较小的网络中的损失定位。 在我们的问题规模上，现有的算法不够快（需要几十秒甚至几分钟）来进行实时丢失定位。

（2）不同的损失模式。

网络故障主要表现为两种分组丢失：完全分组丢失和部分分组丢失，意味着遍历链路的全部或部分分组丢失。 现有的定位技术假设如果路径上的所有链路都很好，那么路径很好。 在数据中心的部分分组丢失的情况下，这是不正确的，例如，分组黑洞可能仅导致使用该链路的路径子集导致链路上的损失。