随着网络尺度和数据流量的爆发性增长，人们急需适用于大规模网络的管理方案。基于流的测量方法对网络管理至关重要。精确的，及时的和高效的流级别的信息采集能为数据中心网络的安全分析、流量工程和QOS提供帮助。

Despite the growing applicability of data centers in a wide variety of scenarios, there are very few systematic measurement studies [19, 3] of data center usage to guide practical issues in data

center operations

尽管人们越来越关注提升数据流采样精度的方法，但是几乎没有人会关注mice流的影响和采样节点间的协作策略。论文[1]说 mice流在数据中几乎占比80%以上。mice流具有生命周期短，数据包少且分散等特点。因此，mice流往往比较难捕获，严重影响了网络全局可视化。通常，异常攻击流量具备这一特性。出于安全考虑，我们的系统应该具备一种对mice敏感的机制。

另一方面，数据采集的过程中容易发生同一个包被重复采集的情况。这种情况不仅占用收集器带宽，损耗计算资源（lead to a massive cost of the collector's channel bandwidth and computing resource ），甚至可能降低系统的采样精度和采样的有效性。这是数据包传播本身的性质决定的。所以，我们应该找到一种节点间协同的采样策略，避免采集到重复的包。

我们都知道，采样节点数量过多会耗费资源。但数量太少又限制采样精度。然而当前节点的选取方法太过单调[2][3]，甚至会采集到大量的重复的流。这并不能适应在大规模网络下的节点间流量突发性和爆发性（explosive growth）的特点。所以我们需要一种策略能从大规模网络中选取少量最具有代表性的节点。Furthermore，为了在减少性能损耗的同时尽可能的提高采样精度,采样时间的分配应该能够自适应变化。

为了解决上述问题，一个灵敏的，自适应的和协同的，并且适用于大规模网络的采样机制是必不可少的。我们

where a network controller can quickly take action based on the data it provides.

Network monitoring and measurement are crucial in network management to facilitate quality of service routing and performance evaluation.

Inspired by the visibility and central control of SDN, we propose CeMon, a generic low-cost high-accuracy monitoring system that supports various network management tasks.

Software-deﬁned networking (SDN) replaces the distributed, per-switch control planes of traditional networks with a (logically) centralized control plane that programs the forwarding behavior of all the switches in a given network. This centralized control plane, run on a controller, can act as a control loop that (i) gathers trafﬁc and other measurements from the network and (ii) uses the gathered information to compute and install forwarding behaviors in the switches.

随着网络尺度和数据流量的爆发性增长，基于流Flow-level的测量方法对网络管理至关重要。在大规模网络中，采样精度是至关重要的，高精度的Flow-level能为管理者和网络应用提供更丰富的信息支持,例如IDS，TE等，以及应用感知网络等。尤其是网络安全而言，网络中Mice流在所有流中几乎占比80%以上。他们数据包少且分散等特点，分布式攻击的容易隐藏在其中。 These malware propagations via the Internet can be prevented by capturing suspicious packets on the network and inspecting them by using security application solutions such as an intrusion detection system (IDS)，因此越高精度的流采集，网络安全性就越高。

采样的模式可分为the systematic packet sampling (SPS) method captures every packet for a sampling duration from a starting point in time, and probabilistic packet sampling (PPS) is a method to selectively capture packets from traffic flows with a uniform or non-uniform probability p. 它们将网络中的数据包按照一定的路由规则转发至统一的分析器上(e.g. IDS)。而对于流采样精度则意味在整个网络中，实际采集的流的数量和当前的流数量的比值。

传统网络中，由于网络节点的封闭性和隔离性，没有很灵活的方式应对网络流量的变化

和根据实时网络状况。得益于SDN网络的出现， SDN replaces the distributed, per-switch control planes of traditional networks with a (logically) centralized control plane that programs the forwarding behavior of all the switches in a given network. This centralized control plane, run on a controller, can act as a control loop that (i) gathers trafﬁc and other measurements from the network and (ii) uses the gathered information to compute and install forwarding behaviors in the switches. 借助于它集中全网络信息的能力和精准的控制能力，在SDN网络下的流量采集能够有更多的策略。SDN控制器能够通过简单的OpenFlow[2]协议将任意节点的流量，以不同的策略进行收集。这是传统网络中无法比拟的。

在大规模网络中采样，为减少对网络的入侵性，往往选择网络中的某些节点进行采样。为最大化流的采样精度，许多研究使用社交网络中的方法来研究该问题[1,3,4]，文章[1，5]采用节点当前拥有的活跃流数量作为影响力量化节点价值，以此选则采样节点。节点的采样率确定，是影响采样精度的另外一个原因。采样率是网络中采样的一种约束条件,如IDS的处理能力，是有限的采样资源。在给定约束下为每个采样节点确定合理的采样率是至关重要要的。例如[3]使用等比分配的方式确定各节点的采样率。采样点的选择，和节点采样率的确定是影响采样精度的重要两个因素。

在采样中，采样的有效性也是一个不可忽视的重要问题[3]。它表示redundant samples的包数量在整个采样中数据包的比值。当多个交换机在同一时间采样同一条流的同一个包，产生了重复包。重复率越高的采样，会使得效率降低，对整个采样系统而言是资源的浪费，甚至可能会影响到采样精度。

因此，在给定收集器能力约束下，如何确定采样节点和确定采样率，同时如何降低重复率，节省出更多的采样资源，来最大化采样精度是我们的动机。

In this paper, 我们使用SPS采样模式。在量化节点价值时，作为采样点选择和采样时长的基本依据，我们不仅考虑了节点当前的活跃流为节点带来的直接价值，还考虑了节点在采样周期内的潜在价值：包括节点在拓扑中的位置和节点的活跃度影响力来综合量化节点的价值。我们也首次提出了节点协同采样的概念：不仅只考虑了采样节点的选择和采样时长分配两个维度对采样精度的影响，我们考虑了节点在时间维度上的采样次序分布对采样精度的影响----通过节点间采样时间在时间维度上的合理安排，降低重复率，节省出更多的采样资源，来提高采样精度。节点的采样次序应该被安排为一定的次序，满足这些次序能够提升采样精度。因此，我们对最大化流采样精度问题进行了量化并建立了模型，把节点的综合影响力作为评价节点的价值标准，把节点的采样时间在时间维度上的次序安排作为优化采样精度的因素。

基于上述模型，我把它分解成三个子问题，提出了AFCS策略，它对采样点选择，采样时长分配，节点协同采样优化分别进行了求解。它包括了：xxx,xxxx,xxx三个算法；这三个算法是相关联的三个算法，前一个问题的解作为第二个问题的输入。AFCS通过这三步走得到最大化流采样精度问题的近似最优解。

Related Work

There is a rich literature on improving the accuracy of trafﬁc ﬂow sampling and measurements. We will brieﬂy discuss the most relevant work here. In[2]中，用社交网络中的中介中心度度量方式利用节点当前活跃流量化其影响力，以此选择前Top-K个影响力大的节点作为采样点；再基于各采样节点当前速率和收集器约束等比分配采样率。In [3]中，针对数据中心的流量采集和测量，使用多控制器的架构，并构造了一种基于节点实时活跃流的贪心策略选择节点。In[5][8]中提出对网络中的流采集和检测大象流的方式。In [13]中基于SPS采样模式，利用流量路径信息去算得合适的采样率来提高可疑流量的采样精度。

在大部分研究中，在选择采样节点和分配采样率(采样时长)时，对节点价值的量化只考虑了其当前活跃流带给该节点的直接价值，没有考虑到节点的潜在价值，例如某些节点当前活跃流为0，但是在采样周期T内，可能会到来新的流，而这取决于节点的潜在价值。

同时在分配采样率/采样时长时，是从节点的速率来确定采样率或时长，却忽略了不同节点在不同采样率/采样时长下带来的价值是不同的。

我们的贡献：

Another important problem associated with network traffic monitoring is the acquisition of network traffic packets in a less-intrusive manner.