经过本次对学术相关会议及期刊的调研，我们总结出了三种传统数据中心业务Web、Cache、和Hadoop的流量模型以及他们的数据传输要求，除此之外，还分析了最新关于手机云存储的流量分布和数据传输特征。最后，还介绍了一种将光交换技术应用于数据中心网络的新型架构，并分析其优点和性能。

# 数据中心网络的流量模型

## Web、Cache和Hadoop业务的流量分析

在对Web、Cache、Hadoop三种业务测试的过程中，数据中心使用的是传统的接入层、汇聚层和核心层的三层结构，每一个服务器都通过10Gbps的以太网链路与其机架上方的Tor-switch相连接，每一个Tor-switch相应地又与其上方的汇聚层交换机（”fabric”）通过40Gbps或100Gbps的链路相连接，然后，汇聚层交换机再与核心层交换机（”spines”）相连接，由此构成树状结构。本实验一共测量了30台机架上Tor-switch层在24小时内的流量数据，每个业务各占其中10台机架，并且每个机架上都单独运行一个业务，以确保采集到的数据能够代表相应业务的流量特征。实验取样的是所有机架上的Tor-switch连接的任意一个服务器端口的数据，每个小时任取其中的两分钟的数据进行记录，最终得到数百TB的数据。

### 1.1 关于流量分布特征的研究

我们先对测量的数据以25µs的时间间隔进行采样，如果在采样间隔里链路的利用率超过了50%则我们称其在该时间间隔为Hot状态，如果一个链路在连续的采样间隔中出现了不间断的Hot状态则我们称其为一次burst。以下是关于Web、Cache、Hadoop三种业务burst的持续时间长度的CDF曲线，如Figure 1所示：

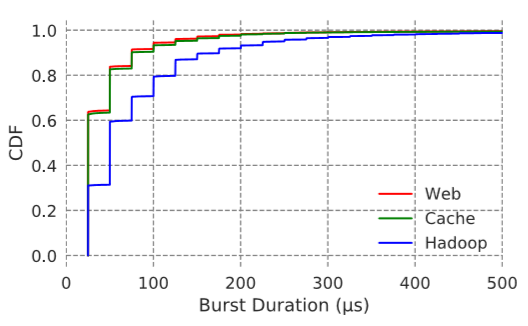


Figure 1

由持续时间的CDF图线可以发现，三种业务超过90%的burst的持续时间长度都小于200µs，而且Web业务的机架的测得的burst的持续时间大概90%都要小于50µs，由此可以得出结论：链路处在高利用率的持续时间长度一般都很短，大概在200µs以内。

继续以25µs为时间间隔，再考虑相邻两个时间间隔是否出现burst是不是会相互影响。

令随机变量来表示在第t个时间间隔内是否出现Hot状态，表示第t个时间间隔内出现了Hot状态，表示第t个时间间隔内没有出现Hot状态。于是根据测得的结果得到以下数据，如Figure 2所示：

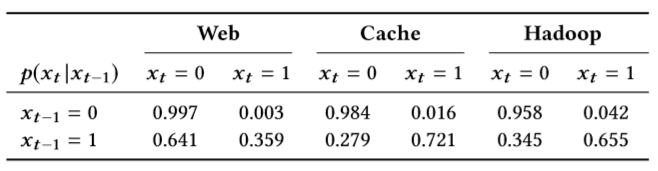


Figure 2

通过表格数据我们可以计算参数r=，如果则说明相邻两个间隔是否为burst状态没有相互影响，若或者则说明相邻两个间隔有联系。计算上述数据可得：

因为三种业务的，所以三种业务下，相邻时间间隔出现高利用率burst状态是相互影响的。

最后研究的是相邻两次burst之间的时间间隔大小规律，根据测得的数据可以得到25µs采样数据下相邻两次burst的时间间隔大小的CDF曲线，如Figure 3所示：

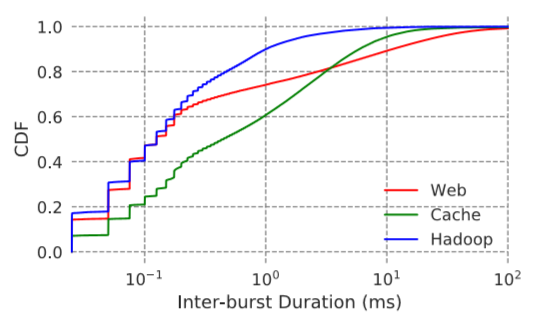


Figure 3

根据曲线可以发现，相邻两次burst之间的时间间隔大多时候都是很小的，特别是对于Web业务和Cache业务来说，它们的间隔大小超过40%都是在0.1ms以下，但是之后的时间间隔大小的分布非常广，且随着时间间隔的增大其概率密度也不断变小。

### 1.2 关于传输数据包大小的研究

根据所测得的数据，得到在burst中和burst以外Web、Cache、Hadoop三种业务的数据包大小情况柱状图，如Figure 4所示：

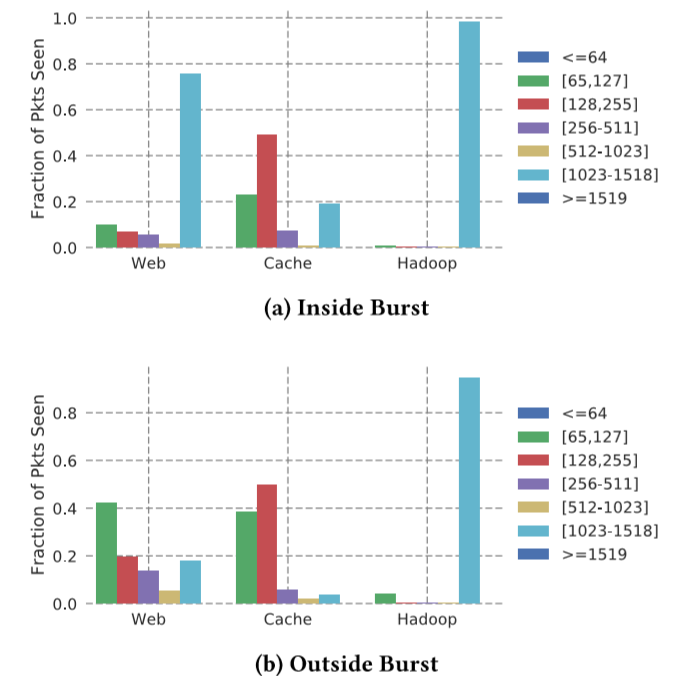


Figure 4

对比两张柱状图可以发现，大体上，burst以外的数据包大小要小于burst内的数据包，而在Hadoop业务下无论是burst内还是burst外主要都是大数据包（大概这与Hadoop业务本身的特点相关），对于Web业务来说在burst以内大数据包占主导，超过了数据包总量的75%，而在burst以内由中小尺寸的数据包主导，但对于Cache业务来说无论是burst以外还是burst内都是小数据包占大多数。

### 1.3 传输链路间负载均衡情况

在本测量环境中，每个Tor-switch都分别通过四个上和下行的链路与汇聚层交换机相连，Tor-switch采用Equal-Cost MultiPath（ECMP）的策略将负载分配至每一个链路上，原则上，这样的策略可以保证比较好的均衡效果，但是实际测试效果如下，用平均绝对离差（Mean Absolute Deviation）的CDF曲线表示，实线表示采样间隔为25µs的情况，虚线表示采样间隔为1s的情况，如Figure 5所示：

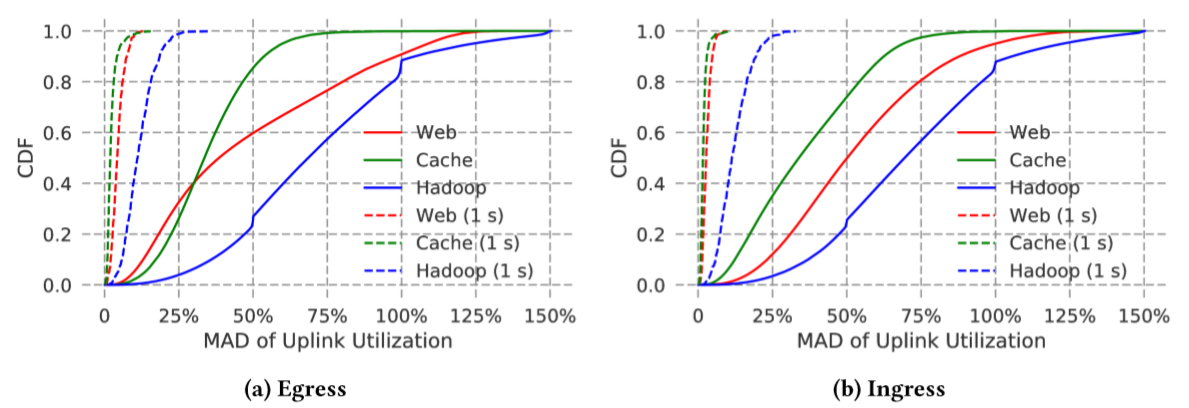
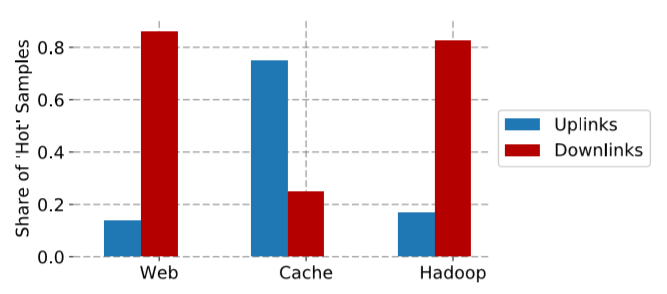


Figure 5

有图可知在采样间隔为1s时，三种业务的负载相差并不大，基本做到了负载均衡，但是当用25µs的采样间隔进行采样时我们却发现，三种业务的负载都相差较大，特别是Hadoop业务，不同链路的负载情况甚至出现倍数关系，基本上可以判定从极短时间上看ECMP的负载均衡效果比较糟糕，其原因可能有两个：（1）ECMP的均衡对象是流的大小而非数据包；（2）它使用的是一致性Hash，不能使均衡效果做到最优。

### 1.4 上行和下行链路Hot状态对比

使用300µs采样周期进行测试，得到上行和下行链路在测试时间间隔中出现Hot状态的频数，并绘制以下上行和下行Hot状态比例柱状图：



根据图中信息，Web和Hadoop业务的Hot状态多出现于下行链路中，上行链路中出现的Hot状态的比例均不足20%。而Cache业务与之相反，在上行链路中更易出现Hot状态，而下行链路中出现Hot状态的比例不足30%。

### 1.5 Burst对Buffer占用率的影响

使用50ms的抽样周期对测量的流量数据抽样，测量发生burst的端口的比例与buffer占用率的关系，这里对buffer占用率根据其最大值进行归一化处理，得到以burst端口比例为横坐标、buffer归一化处理后的占用率为纵坐标的箱型图，如Figure 6所示：

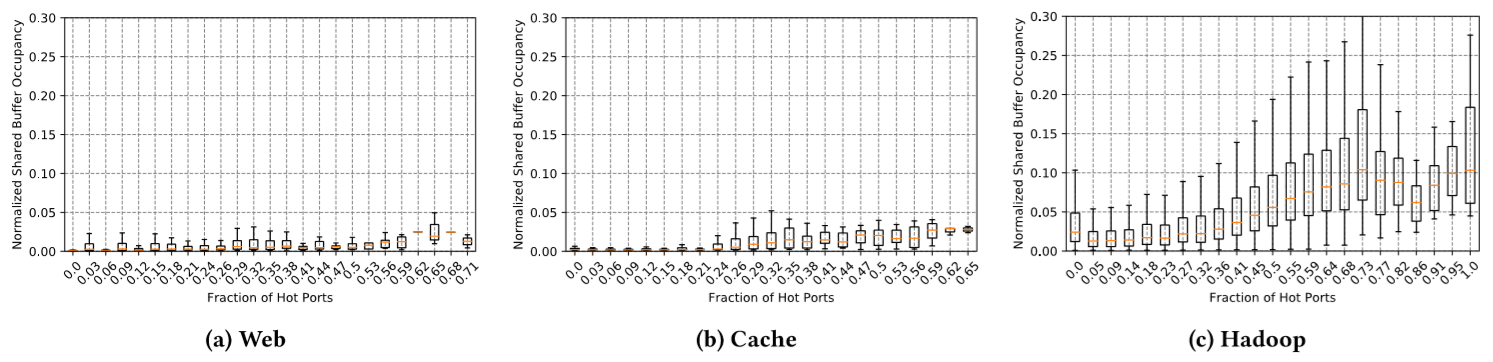


Figure 6

由图可知，Hadoop的burst的端口的比例最大值最高，而Web和Cache业务处于高利用率的端口比例在归一化条件下最高分别为71%和65%。而且相比于其他两种业务，Hadoop对buffer的占用率更高，而且占用率的分布范围也更广。同时，在burst端口比例较高时，三种业务的buffer占用率都保持持平趋势。

## 手机云存储业务

手机云存储不同于传统的桌面云存储业务，它有许多新的特点：

1. 手机云存储的传输流量特征遵循包含两个component的混合高斯模型；
2. 上传至云端的文件数目要远高于下载的文件数；

本实验的数据采集自日活跃用户超过百万的一款中国的云存储服务商，用户可以通过其电脑或手机进行上传、下载、删除和分享操作，并且支持多文件的批量操作，上传文件时本地文件也不会丢失。其中数据传输使用的是HTTP服务，每个HTTP的报文都是固定大小512KB（除了每次HTTP传输的最后一个报文），如果超出了固定大小限制的文件则需要被拆分成多个符合大小要求的报文进行传输。对于一个文件，HTTP请求可以使用多个TCP连接进行传送，同时，一个TCP连接也可以传送多个文件的HTTP请求，这时需要按一定的序列逐个发送，并且只有在收到上一个报文的应答后下一个报文才会发送。

无论是上传和下载，用户设备首先会与metadata server联系。上传时，用户先发送一个包含文件名和文件MD5值的metadata至metadata server，metadata server先检查该文件是否已存在于云存储中，如果已存在，则直接将文件加入到该用户的云存储空间并通知用户设备无需上传，如果不存在，则发送给用户设备离其最近的front-end server的身份信息，通知用户设备与之建立连接。用户设备收到front-end server的身份信息后便向其发送一个file storage operation request，里面包含了文件名、文件大小、文件MD5值、待传的chunk数量、每一个单独的chunk对应的MD5值等信息，然后用chunk storage request初始化存储过程。下载时，用户设备先请求预设的metadata server下载文件对应的MD5值，然后使用这个值通过file retrieval operation request去从storage server请求关于该文件的信息，之后再通过chunk retrieval request请求一个一个的数据报文。

这里定义从一个file operation request的开始到下一次file operation request的开始的区间算作一个file operating interval，用T表示；给定时间参数τ，从某个的file operating interval的结束开始到下一个的file operating interval的最后一个报文的末尾作为结束，称其为一个session。

本实验采集的数据是2015年8月里该云存储业务的所有移动端用户的一周的HTTP请求日志，一共从1148640个活跃用户使用的1396494个移动设备中获得了349092451条日志记录，其中一个用户可能在使用多个移动设备。

### 2.1 流量大小和文件传输数量的时间分布情况

首先将得到的数据绘制成数据量随时间变化的曲线，如Figure 7所示：

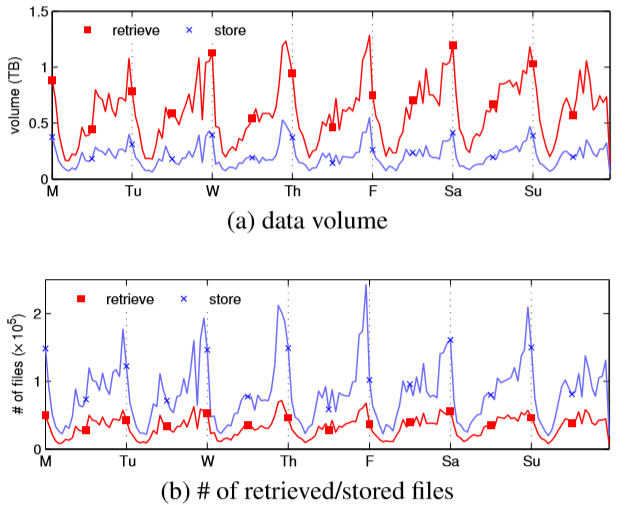


Figure 7

由图可以发现用户的数据流量大小和文件数目整体随时间呈现大致的周期性变化，在每天的23:00左右呈现出一个较高的尖峰，随后在3:00左右为一个低谷，这与手机用户的日常使用习惯相符。其次，对于流量大小，用户下载的流量大小始终都明显高于上传的大小，而对于文件数目，上传的文件数却又明显高于下载的文件数，由此可以推断，手机用户喜爱上传大量容量较小的文件，而下载时都偏向下载大容量文件，且上传的文件数目远高于下载的文件数。

根据分析结果可知，无论是storage server还是metadata server，为了能够应付数据传输的峰值情况，在大多数时候的实际资源利用率都很低，这时就需要使用可以弹性调整其资源大小的服务来减少这种浪费，其次，大容量数据传输对该云存储存储空间和带宽需求也是一个巨大的挑战。

### 2.2 File storage/retrieval operating interval的大小分布情况

通过采集到的数据机绘制出file operating interval的大小分布情况，如Figure 8所示：

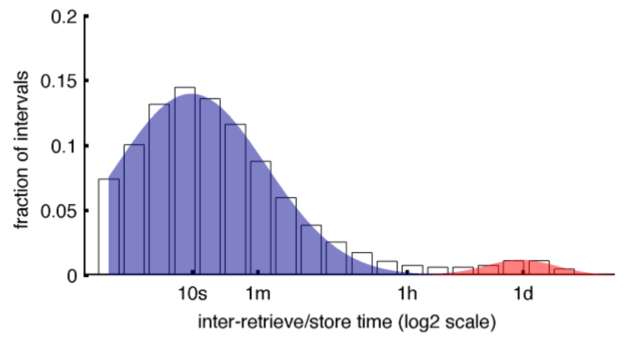


Figure 8

由图可见，file operating intervals的分布是一条双峰曲线，其中一个波峰对应的interval大小为10s，另一个波峰对应的interval大小为1天，而中间频率几乎为0的波谷对应的interval大小为1h，由之前关于session的定义，这里就推荐令，这样，当file operating interval小于1h时，它与它之后紧邻的下一个file operating interval就同处在一个session内，否则，它传输的最后一个报文的结束就标志着其所在session的结束。具体关系见Figure 9：

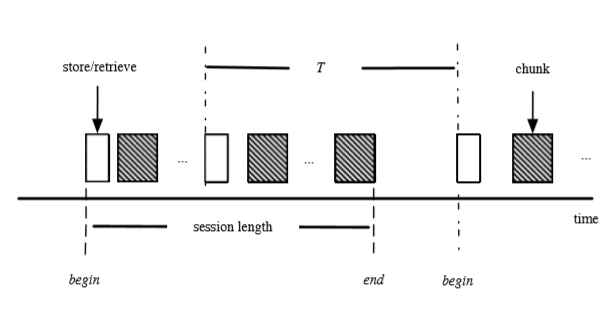


Figure 9

如此一来，手机传输的流量变化特征就服从包含两个component的混合高斯模型，可以通过expectation maximization(EM)算法求得其各个参数，其中一个component相当于横坐标左边的部分，其均值约为10s，另一个compenent相当于右侧的部分，其均值约为1天。

### 2.3 Session的大小变化规律

Figure 10中图(a)绘制的是单个session所包含的file operating interval的数量的CDF曲线，图(b)和(c)描述的是上传和下载两种情况下session的数据量大小随着其包含的file operating interval数量的变化情况：

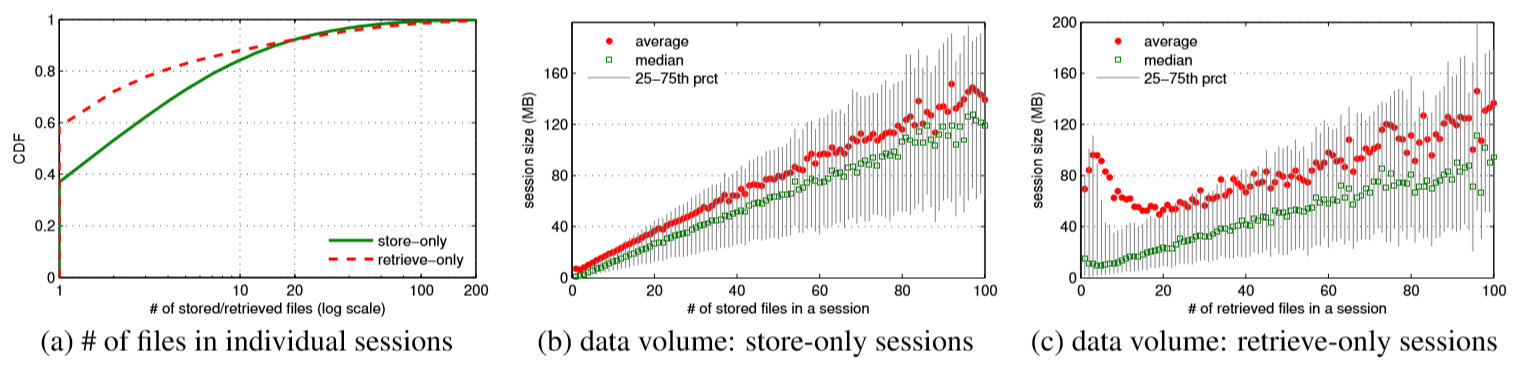


Figure 10

由图(a)可以观察到，差不多半数的session都只包含单个file operating interval，还有大约10%的session包含20个以上的interval，并且下载过程中的session包含的file operating interval的数量总体上要小于上传时的Interval数量。

根据图(b)可以发现，在上传过程中单个session的数据量大小和其所包含的file operating interval的数量呈线性关系，可以推断上传的每个文件数据量大小基本一致。而在下载过程中情况有所不同，单个session包含少量file operating interval时，其中的文件大小平均值较大，而且随着包含的interval数量的增多，session数据量大小波动也非常大，说明下载过程中传输的文件的大小不尽相同。

# 基于光交换的新型数据中心网络架构

近些年，光交换技术在数据中心网络中的应用越来越受到学术界的关注，并取得了一定的研究成果：光网路可以在很大范围的多层、多个域网络内提供动态路由；光网络的连接很有希望实现数据中心之间高触发、高带宽的连接需求；光网络可以通过网络虚拟化来提高网络设备的效率等等。有一些数据中心已经用到了光交换技术，并且都依赖于一个中心的控制器，这里我们研究的也是基于一个中心控制器的基于光交换的数据中心网络。

## 网络架构简介

我们介绍的这个架构是基于光突发技术（Optical Burst Switching）。首先，光突发交换技术需要将网络分为两层——控制平面和数据平面，数据平面拓扑又包括了服务器机架（rack）上的Tor-switch和快速光交换机（Fast Optical Switch），分别属于edge层和core层。常见的快速光交换机有arrayed waveguide grating routers（AWGRs）、semiconductor optical amplifiers（SOAs）和1×N photonic switches，他们的转发数据包的处理时间都在纳秒（ns）量级，可以实现非常快速的转发。

每一个服务器机架中的单个服务器都与其顶层的Tor-switch通过电路连接，而Tor-switch又与其之外的快速光交换机连接，并通过management network与控制平面的控制器连接，控制器可以分别传送控制信息给网络中的Tor-switch和快速光交换机，其具体结构如Figure 11中图(a)所示：

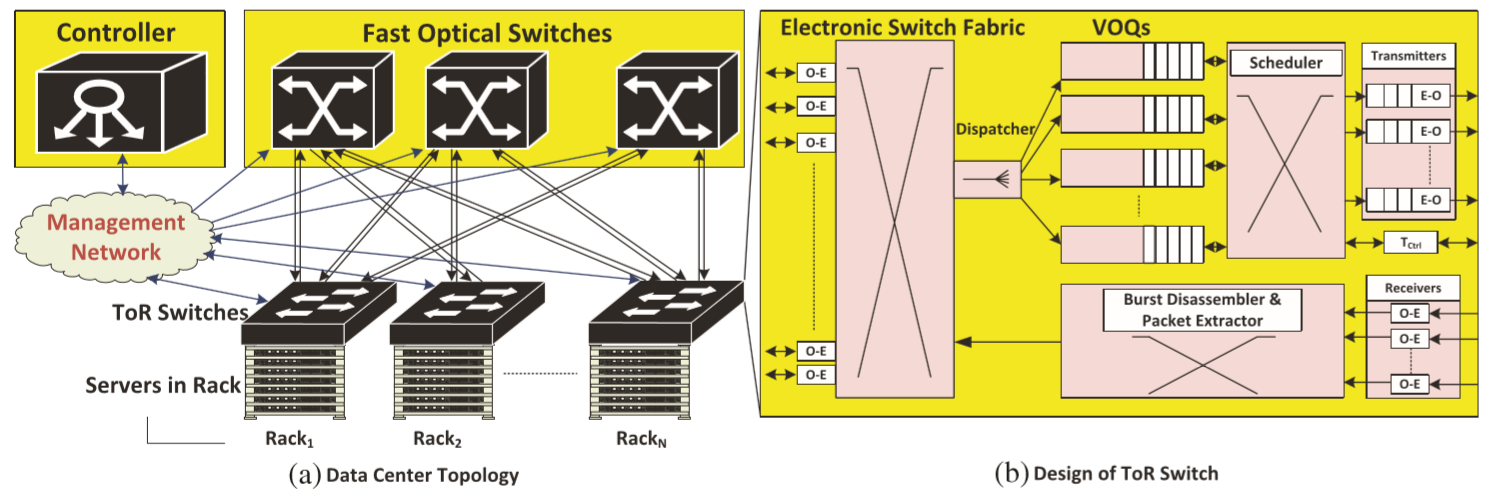


Figure 11

这里推荐使用的快速光交换机是带有1024个接口的SOAs，但是也可以使用带有512个接口的AWGRs。接口数量越多，其连接的Tor-switch数量也就相应的越多，因此负责的服务器数量也就越多，但是，这样的扩展性也取决于控制平面控制器的性能。

Tor-switch内部结构如Figure 11图(b)所示，每一个Tor-switch都通过电路交换网与rack下层的服务器相连接，用以负责rack内的数据交换。同时，Tor-switch也负责rack之间的数据交换，这里采用的是在Tor-switch内设置个虚拟的输出队列（Virtual Output Queues），N是数据中心网络中所有Tor-switch的数量，每一个VOQ都对应一个特定的目的Tor-switch，用于收集想要发送到目的Tor-switch的所有数据。每一个Tor-switch都维护着一个VOQ表来记录每个VOQ对应的VOQ编号和目的rack网络地址。Tor-switch中的dispatcher模块根据VOQ表将从rack下层服务器收到的数据包匹配至对应的目的网络地址的VOQ中。Scheduler则负责根据控制器发挥的控制信息来将数据转移至对应的发送接口发出，其中，Tor-switch还通过一个transceiver与控制器通信。

控制平面主要由中心控制器组成，负责路由、调度（scheduling）和交换处理功能。它接收来自数据中心网络中所有Tor-switch的连接建立请求，寻找对应的俗陋方案、给每个连接请求分配时隙并根据时隙分配调度网络资源实现光交换。控制层维护着一个连接状态记录数据，记录着所有光交换机在对应时隙上的状态，数据平面的光交换机就是根据控制器预先分配的光路来进行数据包的转发。控制平面与数据平面通过一个管理网络（Management Network）相连接。

## 数据转发过程

### burst的产生与发送

burst的形成可以是基于时间的，也可以是基于长度的。当数据包到达一个空的VOQ中时，计时器启动，所有到达该VOQ的数据包开始聚集，当计时器超时或VOQ中聚集的数据长度超过了阈值时，就产生一个控制数据包（control packet），使用transceiver发送给控制平面的控制器。这个控制数据包包含了数据的长度、源和目的IP地址和源和目的Tor-switch ID值等信息，具体结构见Figure 12。在数据中心网络中每一个Tor-switch都被指定了一个ID编号，从0至，N是网络中rack的数量，这些ID编号被控制器用于做路由和调度算法。控制器处理了控制数据包后，分配一个开始时间（start time）和Tor-switch接口号，数据之后会从接口号对应的Tor-switch接口发出，并将这些信息加入到控制数据包中发回给对应的源Tor-switch。当源Tor-switch接收到返回回来的控制数据包后，就根据其中的信息抽取VOQ中的数据产生一个burst，并在指定的时隙从对应的Tor-switch接口发出。尽管VOQ非空，scheduler模块依然会在控制数据包发出后初始化一个新的计时器，因为在控制数据包发出的RTT时间间隔内会有新的数据包达到该VOQ。

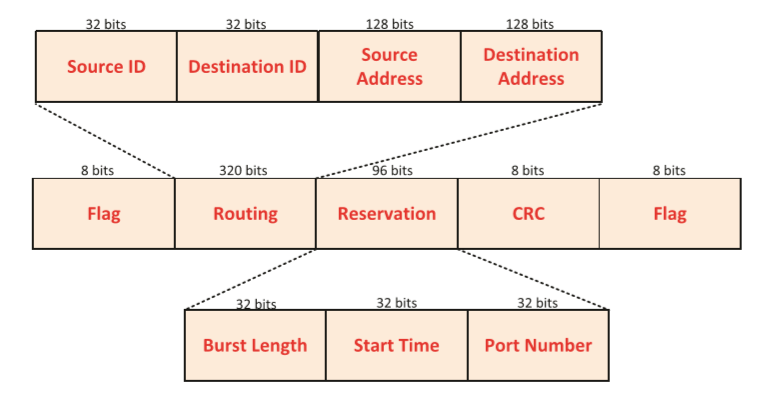


Figure 12

### burst接收过程

当burst到达目的Tor-switch后，Tor-switch中的burst disassembler和packet extractor模块便会将数据包从burst中抽取出来并通过电路交换发送给其所在rack下的各个服务器。

### 控制平面的处理过程

控制器维护着所有光交换机的连接情况信息，并实现路由、调度和操作交换机的处理。其资源分配机制用horizon scheduling来表示，horizon指的是最近的可以使用的空余传输资源的时隙，如Figure 13所示。其中图(a)展示的是在控制数据包到达前的各传输通路的状态，horizon scheduling寻找其中有可用通路资源的最近时间来安排本次burst的发送。图(b)展示了分配时隙之后的状态。

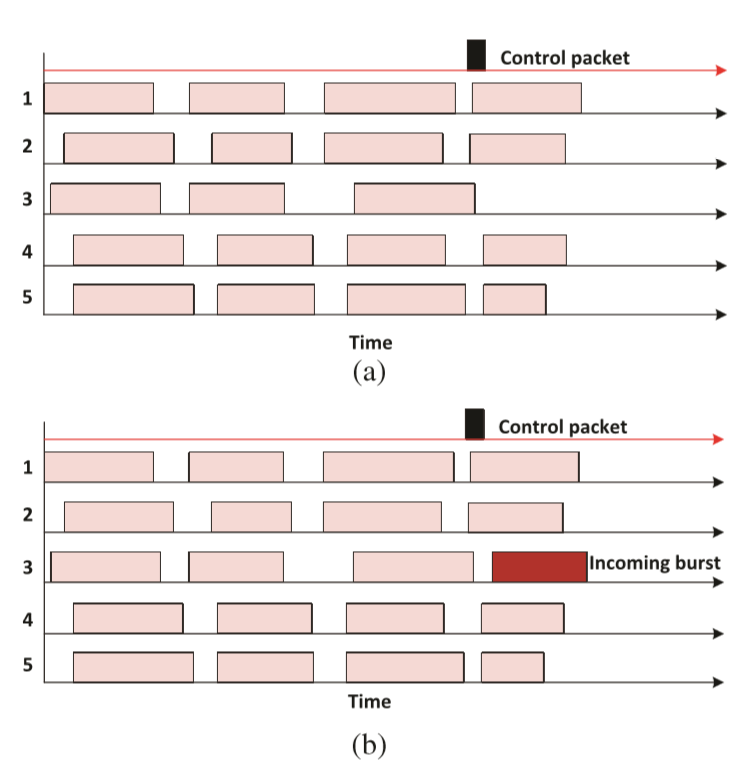


Figure 13

在控制数据包发送后，控制器会产生一个处理信息（configuration message），控制器在其中设置了输入接口、输出接口和执行时间等信息，同时也将其源IP设置成控制器的IP并把目的IP设置为光交换机的IP地址，最后将该处理信息发送至对应的光交换机，光交换机的控制器会根据处理信息中的说明进行操作。

## 性能评估

传统的光突发交换技术相比于这里的改进架构，缺少了对控制数据包control packet的回传，于是burst的产生和发送时间需要Tor-switch设置一个时间偏置来决定，及在Tor-switch将control packet发出后便开始计时，达到时间偏置后就发送burst，控制器则根据收到的control packet直接为其分配网络资源。这样的发送方式导致的结果就是可能有多个burst争抢同样的网络资源导致冲突，从而造成较高的丢包率。而这里的改进算法资源分配是由控制器统一分配管理的，在一定的负载下可以实现0冲突。

这里就对传统的光突发交换的数据中心网络和改进架构的数据中心网络进行对比，来衡量他们的时延、吞吐量和丢包率，同时也与一个典型leaf-spine架构的两层电交换数据中心网络进行了比较，点交换网络如Figure 14所示。

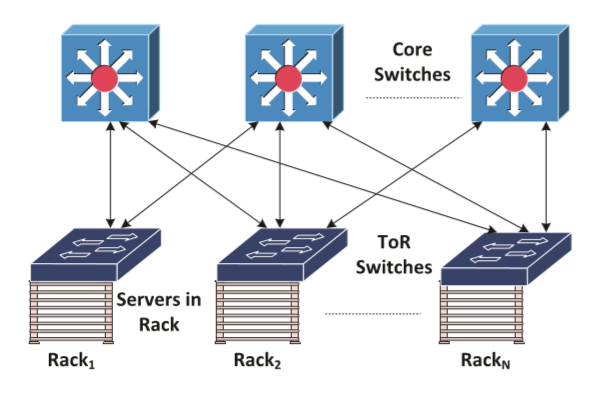


Figure 14

为了测试的结果尽量具备广泛性，该实验测量了10Gbps和40Gbps两种数据产生速率，包含了Tor-switch同时向1台、10台、20台Tor-switch发送数据包的情况，即TDC={1，10，20}，并且还对比了oversubscription ratios为1:1和2:1的情况，即源rack需要发送的数据的目的地址全在其他rack上和50%在其他rack上还有50%在源rack上两种情况。测得的数据如下。

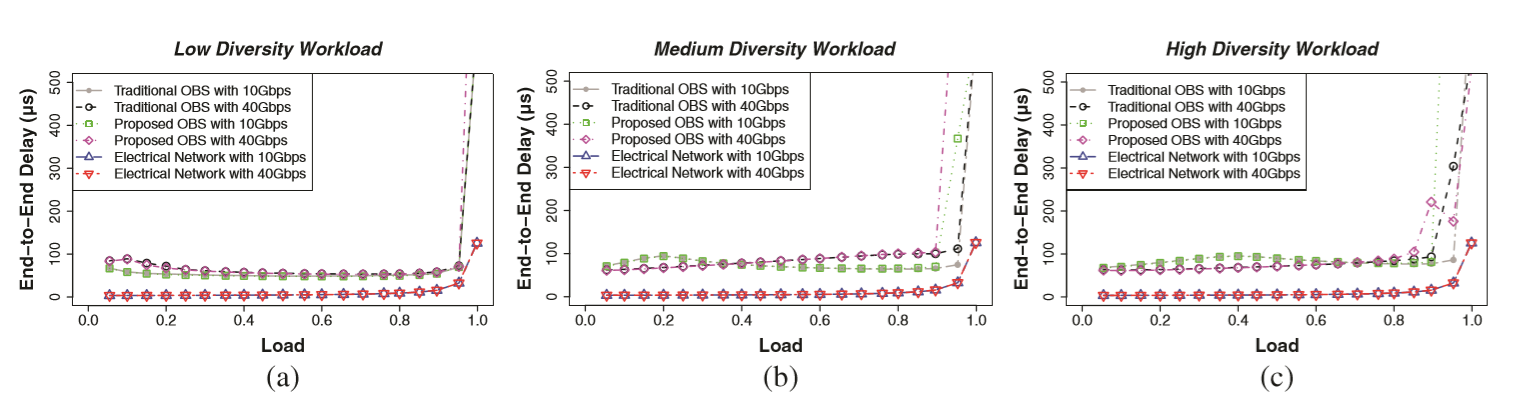


Figure 15

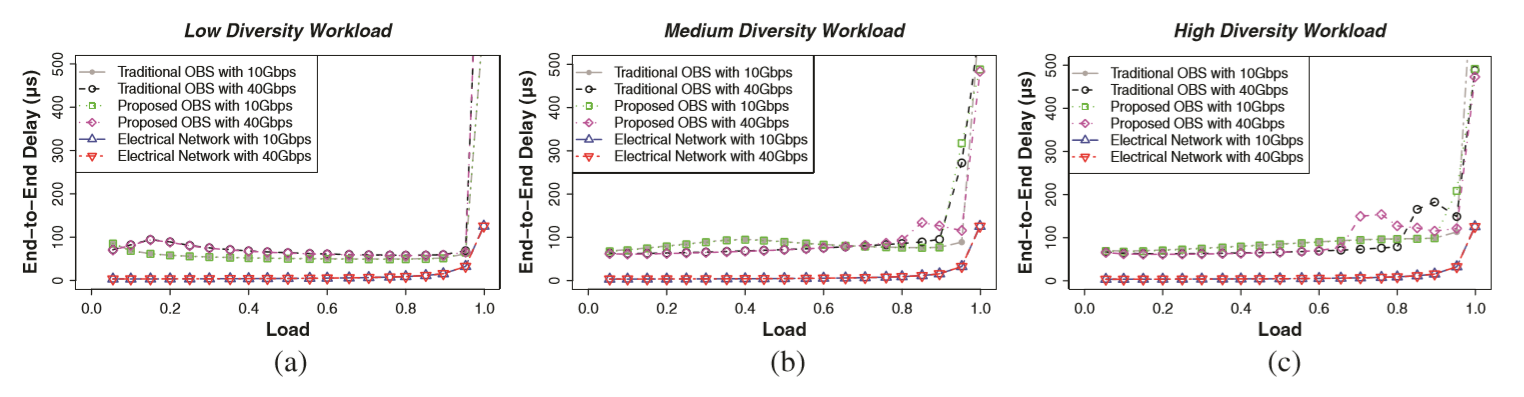


Figure 16

Figure 15和Figure 16分别是oversubscription ratios为1:1和2:1两种情况下的端到端时延情况对比，他们的图(a)、(b)、(c)对应的分别是TDC={1，10，20}三种情况，由对比结果可以发现改进前后时延差别不大，且都比电交换的要差一些，这说明在数据中心这样集中的网络环境下，增加回传机制导致的control packet的RRT延时对总时延影响并不大。

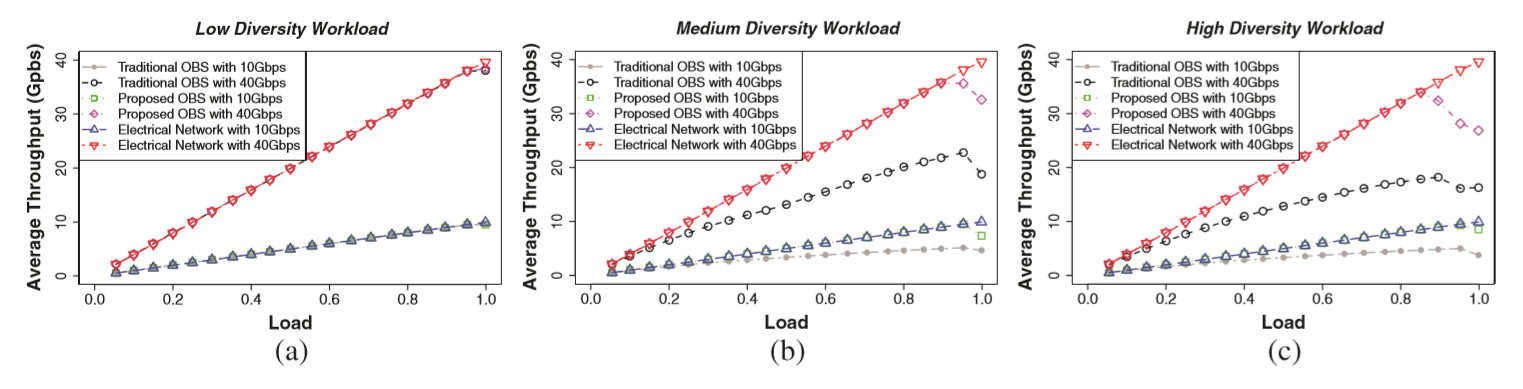


Figure 17

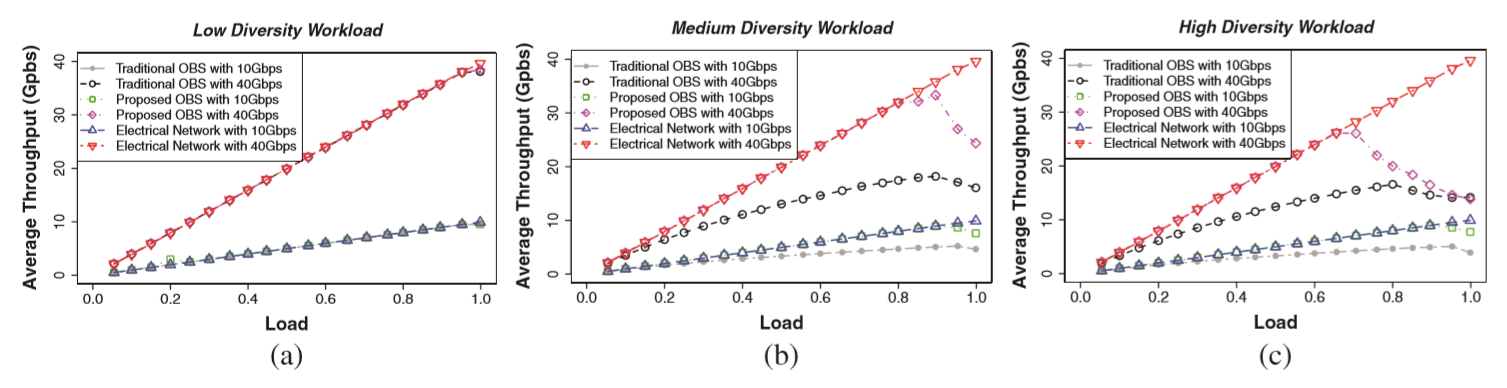


Figure 18

Figure 17和Figure 18分别描述的是oversubscription ratios为1:1和2:1两种情况下的平均吞吐量情况，他们的图(a)、(b)、(c)对应的分别是TDC={1，10，20}三种情况。对比发现改进后的基于光突发交换的DCN架构的吞吐量要比改进前的大，但是改进后的架构在高负载情况下会小于电交换数据中心网络架构。

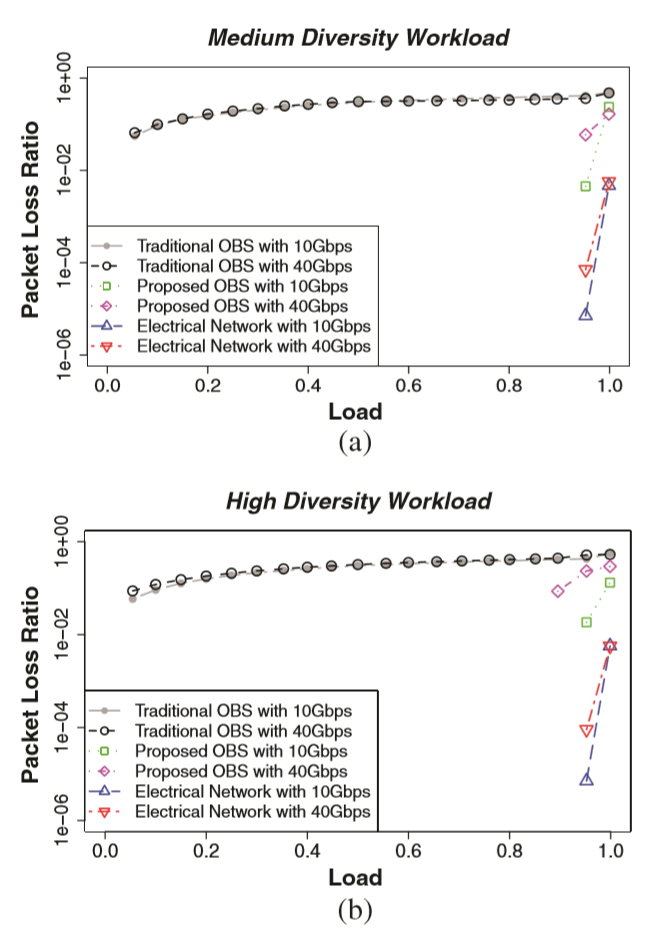


Figure 19

Figure 19中展示了oversubscription ratios为1:1情况下丢包率的大小对比，图(a)对应着TDC为10的情况，图(b)对应TDC为20的情况。由图可知改进的架构和电交换网络的丢包率只有在很高的负载下才会变得较高，而传统的基于光突发的数据中心网络始终保持着较高的丢包率。

由上面的仿真测试数据可知，这里提到的改进的基于光突发的数据中心网络在时延基本不变的情况下提高了吞吐量并降低了丢包率，并且与电交换数据中心网络的差距较小。

## 使用机器学习对光网络的改进

除了以上介绍的基于光突发交换技术的数据中心网络外，还有使用光分组交换的数据中心网络，相比于光突发的传输单元burst，光分组传输的是光分组，并且整体架构依然是由控制平面的控制器和数据平面的Tor-switch和光交换机组成。

在数据包的处理和转发过程上，电数据包先进入到Tor-switch的与目的rack对应的VOQ中等待被转为光分组发送出去，而每一个Tor-switch在一个时隙上只能转换一个VOQ的光分组并发送，因此需要控制平面的控制器进行决策并给予某个VOQ发送权限。当VOQ收到发送权限后就会在下一个时隙发送一个光分组，光分组经过光交换机到达目的Tor-switch。

决定下一个时隙发送哪个VOQ的光分组一般是通过轮询调度（Round-Robin Scheduling）来决定，这既容易实现，保证了处理速度，同时也绝对的公正，每个待发送数据的VOQ机会均等。

但是，轮询调度的方式却忽视了数据中心网络中的数据特征。

根据之前的研究可以发现，数据中发送的流的大小不是均匀分布的，而是呈现出两极化的特点，如Figure 20的图(a)和图(b)所示，数据中心的流量主要由大流和小流组成，且大流和小流的数据长度、持续时间、时延敏感度都有很大差异，因此采用完全公正的方式来决定发送顺序在很多情况下都会不够合理，例如发送了持续时间较长的大流，而使得对时延敏感的小流一直等待。

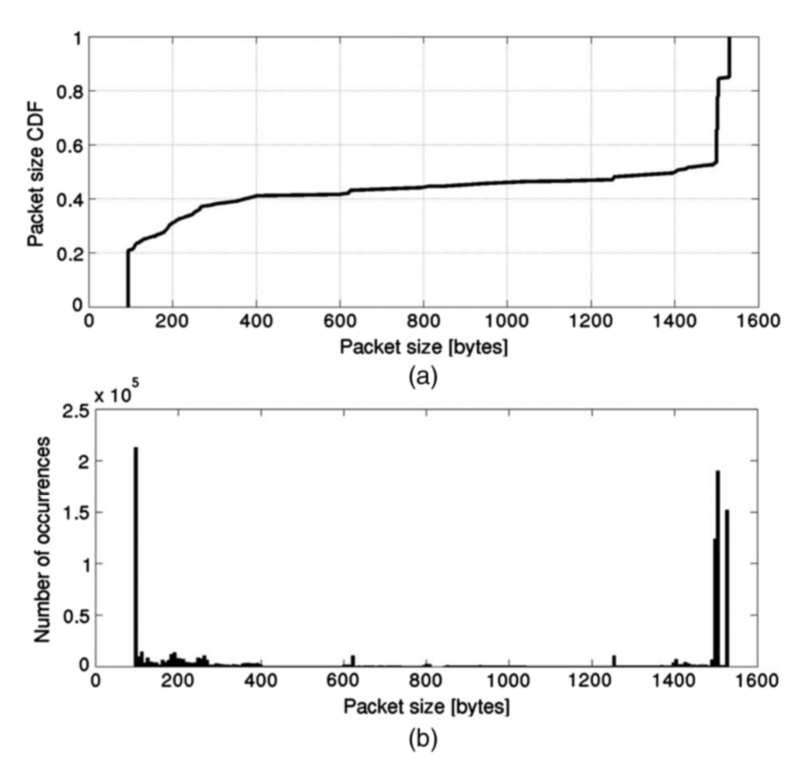


Figure 20

因此可以通过识别VOQ中的大流和小流等特征来赋予每个VOQ一个优先级大小，每次决策都发送优先级最大的。这里，改进的方案考虑优先级的因素有四点：VOQ队列越长优先级越大、数据包数目越多优先级越大、等待时间越长的数据优先级越大、转发过程不需要调整光交换机的优先级越大。通过四个参数（weight factor）将这些指标量化，分别是、、、。优先级计算公式为。

这时调度方案就是根据优先级调度（Priority-Aware Scheduling）。算法具体过程如Figure21所示。

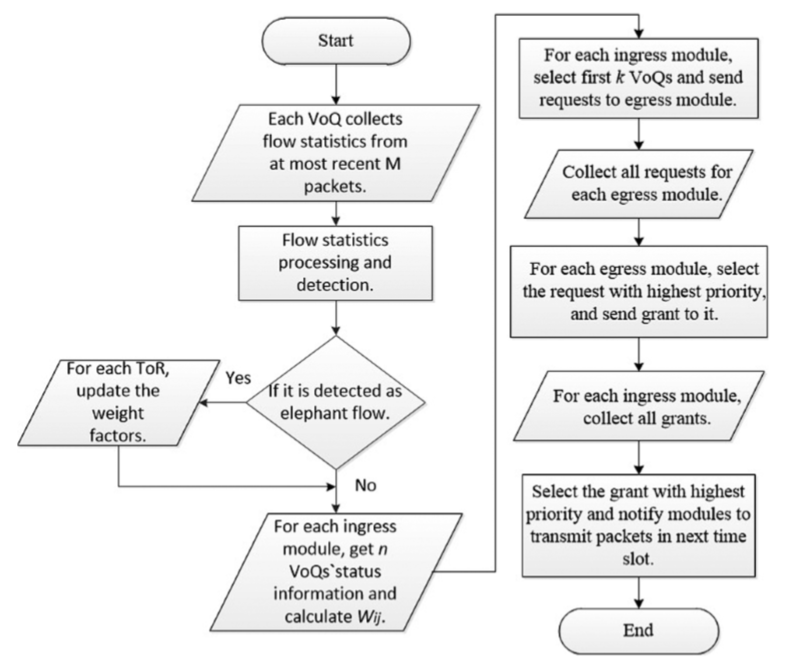


Figure 21

其中，算法中判断最大流的方式使用机器学习的分类算法，这里分别使用C4.5决策树和Naïve-Bayes discretization（NBD），当分类结果发现是大流时，则升高并降低，若分类检查出是小流时则降低并升高，最后计算各个VOQ的权重决定发送顺序。