基于预测的 DC 间流量增量调度研究

王然，张宇超，王文东
(wangranse@bupt.edu.cn, yczhang@bupt.edu.cn, wdwang@bupt.edu.cn)
通讯作者：张宇超

摘要

目前已经有许多关于数据中心间网络性能优化的研究[1][2][3][4][5]，但这些已有研究要么未涉及在线流量和离线流量共用一个数据中心传输网络的问题，要么在线流量和离线流量在链路中的分配模式是固定不变的，因此它们未能充分利用链路中的剩余流量。而如果对带宽利用情况进行预测，并且在预测时根据网络变化情况动态地权衡预测灵敏程度和更新频率，从而实时地调整在线流量和离线流量的带宽占比，进行动态的带宽分离，能够提高数据中心网络的利用率。

关键词 数据中心 流量工程 预测算法

一、引言

如今，越来越多的大型企业在世界各地构建起了自己的数据中心以及跨数据中心域的数据传输平台。但是，连接每个数据中心对的长途链路十分昂贵，因此提高数据中心间链路的利用率毫无疑问能够为企业带来巨大的效益，尤其是随着 5G 的到来，传输数据量会急剧膨胀，数据中心间链路利用率的提升将更为紧迫，其带来的效益也将更为显著。

传统的数据中心间的数据复制方法具有三个局限：

第一，现有的非集中式的调度方式容易得到局部最优的决策。在非集中式的调度方式下，每个服务器仅能看到部分可用传输数据源，因此不能利用所有可用的覆盖路径来最大化链路利用率。

第二，高计算开销。为了维护全局视图，以保证全局最优的调度决策，现有的集中式调度方式的计算开销一般都较大，很难应用到实际当中。

第三，固定的带宽分离模式。在这种模式中，链路总带宽被以固定的比例划分给在线流量和离线流量，这会导致在线流量很少的时候，预分配给它的那部分空闲带宽无法被数量众多的离线流量利用。

根据以上分析，本文提出了一种集中控制的，基于带宽使用情况预测的实时增量调度方法，它能够解决传统数据复制传输方式存在的以上不足，让在线流量和离线流量共享带宽，使数据中心网络的链路利用率接近 100%。

二、研究背景及现状

为了提高数据中心间 WAN 的链路利用率，Google 先后提出了 B4 [6]以及它的改良版本 B4 and After[7]，此外还有支持其中 TE 组件的 BwE 组件[2]。

B4 整体采用 SDN 的结构来实现，主要通过一个集中式的流量工程（TE）算法来为应用分配带宽，实现最大最小公平(max-min fairness)，这一方式能够使链路利用率达到接近 100%。

针对带宽需求和可靠性要求的增长，Google 随后提出了 B4 and After，它改进了 B4 原有扩展方案，将每一个站点设计为两层拓扑抽象，同时引入了 sidelinks 和 supernode 级别的 TE 解决容量不对称问题

以 Google 为例的集中式调度对于提升链路利用率的效果立竿见影，但 Google 的场景跟国内大多数企业的场景不同，它具有两套独立的 WAN，除了上述连接众多数据中心的 B4 之外还有一套是链接数据中心和用户的 B2。

而国内的数据中心网络则将在线流量和离线流量混合部署，需要进一步考虑带宽的分离模式。

百度在面临以上问题时，先后提出了 PieBridge[8]和 BDS[9]。PieBridge 通过在残存网络上进行传输，最大化了通信链路的带宽使用，同时通过使用调度器选择数据传输源，显著减少了数据同步的完成时间。在 PieBridge 的理论基础上，BDS 具体通过充分利用 DC 间的覆盖路径来减少数据同步的完成时间，同时给组播流量（离线流量）和延迟敏感型流量（在线流量）固定地分配带宽占比来实现链路带宽的共享。

然而，固定的带宽分离模式仍不能充分地在在线流量和离线流量之间共享带宽，尤其是当在线流量较少，固定分配给它的带宽剩余较多时，会导致严重的带宽浪费。

三、动态分离的传输机制

3.1 实时增量调度

本系统的调度和路由决策由集中控制器产生，并且当控制器出现故障或无法访问时，系统会退回到传统的非集中的控制模式。集中式的调度系统能够产生近似最优的决策，但同时也会带来较小的更新延迟，因此在设计调度方案的时候要对两者进行权衡。由于决策空间十分庞大，使用标准路由和线性规划的方法很难得到近似最优的决策，所以本系统将集中控制解耦为调度和路由两部分，调度过程选出候选数据块，缩小了路由阶段的决策空间，因此能几乎实时地更新。

集中控制器周期性地更新路由和调度决策，周期一般是几秒，每个周期的工作流程如下：

1) 由运行在每个服务器本地的代理确认本地的状态，包括数据块的传输状态（哪些块到来，哪些块未完成），服务器的可用性，磁盘故障，等等。

2) 然后这些数据被包装成一个控制消息，通过代理监视器的高效的消息传递层发送给集中化的控制器。

3) 控制器也从网络监视器接收网络层的数据（延迟敏感型流量的带宽消耗，以及每条数据中心间链路上的利用率），并调用预测模块，根据历史在线流量的信息预测出当前周期各链路上的在线流量使用量，作为调度模块的输入。

4) 一旦接收到来自所有代理和网络监视器的更新信息，控制器就运行集中式的决策制定算法，以得到新的调度和路由决策，并且通过代理监视器的消息传递层将新旧决策间的差异发送给每个服务器本地的代理。

5) 最后，代理为每个数据传输分配带宽，根据控制器的路由和调度决策来进行实际的数据传输。

3.2 在线流量预测

在传统的固定带宽分离的模式中，由于分配给延迟敏感的在线流量以及对延迟不那么敏感的离线流量的带宽是固定的，所以即使在线流量很少时，离线流量也不能利用分配给在线流量但目前空闲的带宽，这将导致很低的链路利用率。所以本系统采用了动态带宽分离模式，根据不同的网络情况来自动调整调度结果：当在线流量到达其峰值时，本系统缩减离线流量

拥有的带宽以避免拥堵，当在线流量到达其谷值时，本系统让离线流量使用更多的带宽以充分使用剩余带宽。图 1 为本系统动态带宽分离的逻辑图。网络监视器读取由代理观察到的 traffic 值，然后执行一个由 EWMA（Exponentially Weighted Moving-Average）[10]和拐点检测算法[11]组成的 Sliding-k 模块，EWMA 负责计算当前周期的流量预测值，拐点检测负责观察历史数据然后判断当前是否有突变出现，使代理监视器平稳而且敏感。

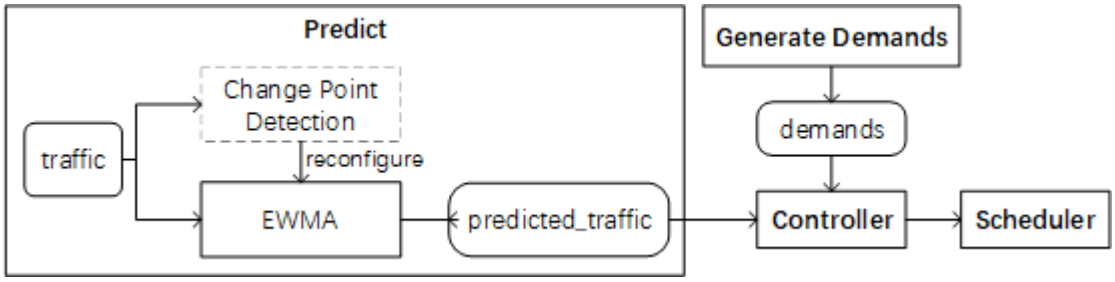

图 1 动态带宽分离的逻辑图

目前存在一些基本方法可以检测在线流量变化并动态调整调度的配置，如 EWMA、k-Sigma 等[12][13]。但这些方法有时候会连续地重配置，甚至在网络很稳定的时候也会如此，这是没有必要的。因此，在预测可用带宽的时候面临着一个权衡：当我们在预测时更偏向于参考最近的数据，预测值将会出现明显的震荡，这将引起不必要的连续重调度；而当我们在预测时更偏向于参考历史数据，预测值受近期检测到的拐点的影响就较小，这使得系统对于网络变化不那么敏感。

为解决以上问题，本系统将 EWMA[10]和拐点检测算法[11]结合，并且设计了定制的 Sliding-k 算法。具体来说，Sliding-k 为 EWMA 算法设置了一个上界 K，若当前没有拐点，k 将被设置为 K，而当一个拐点被检测到时，k 将被设为 0，然后逐渐地增长到 K。

在一个调度周期内，网络变化监视器不断地获取服务器吞吐量的一系列代理观测值，这些值在下一个调度周期内被用来预测可用带宽。

3.2.1 EWMA

EWMA 即指数加权移动平均法，这一方法可以根据历史观测值来估计当前的值，预测时给观测值的权值随时间呈指数递减，离当前时间越近的数据权重越大。

EWMA 的表达式如下：

$$Z_0 = \mu_0, \tag{1}$$

$$Z_i = (1-r)Z_{i-1} + r\bar{X}_i \ , \quad j > 0 \tag{2}$$

- $\bar{X}_i$为 i 时刻的实际值
- 系数 r 表示加权下降的速度，其值越大表示下降地越快，也就是给最近观测值的权重更大
- Z_i为 i 时刻 EWMA 预测的值

将 EWMA 的表达式归纳后可写成如下形式：

$$Z_i = (1-r)^i\mu_0 + r(1-r)^{i-1}\bar{X}_1 + r(1-r)^{i-2}\bar{X}_2 + \\ \ldots + r(1-r)\bar{X}_{i-1} + r\bar{X}_i \ , \quad j \geq 0 \tag{3}$$

从上式可以看出，观测值的权值随着时间呈指数式下降。给近期观测值较大的权重是因

为它对预测值有较大的影响，更能反应近期变化的趋势。

3.2.2 贝叶斯在线拐点检测算法

为了同时保证稳定性以及对网络变化的敏感程度，预测模块引入贝叶斯拐点检测算法。该算法使用消息传递算法计算当前“运行”长度或自上一个变化点以来的时间的概率分布。

贝叶斯拐点检测算法在单变量时间序列上以在线方式执行贝叶斯拐点检测。核心思想是在每个新数据点到达时递归计算“运行长度”的后验概率。运行长度定义为自上次更改点发生以来的时间。

3.2.3 Sliding-k

在对序列进行预测时，若 EWMA 的参数α设置得很小，意味着给更“旧”的观测值的权重更大，预测结果会更加平稳，但对于突发的抖动不够灵敏，若α很大，意味着给较“新”观测值的权重越大，预测结果就会越灵敏，但是预测值会出现频繁的波动，引起连续的重配置(如果预测结果跟上一周期的预测值相差不大,则无需重新配置状态信息,减小更新压力)。所以本系统采用了 Sliding-k 的方式，即当前无拐点时就给更“旧”的观测值更大的权重，以保证预测结果的平滑，而一旦检测出拐点，就给较“新”观测值更大的权重，保证预测结果的灵敏、准确。

其主要步骤如下：

1） 贝叶斯拐点检测算法计算当前周期为拐点的概率。
2） 通过将上一步得出的概率值与设定好的阈值对比，判断当前周期是否为拐点。
3） 如果当前周期为拐点，则 EWMA 的输入序列的窗口大小为 1，也就是 EWMA 输入序列仅包含上一周期的观测值，且参数α设置为 1，即预测时仅参考上一周期的观测值，之后每经过一个周期α值减小 0.1，直到α值降低到指定的稳定值（如 0.6）。
4） 如果当前周期非拐点，则 EWMA 的输入序列为上一个拐点出现时到上一周期这段时间内的一系列观测值。
5） EWMA 根据输入的序列值和α值预测当前周期的值，输出到调度模块。

Sliding-k 根据拐点检测的结果动态地调整预测方式，使得在无拐点时能够保证预测结果的平滑，而有突发拐点时又能灵敏地响应，做出准确的预测。

四、实验评估

我们在实验中对比了固定带宽分离与本系统动态带宽分离的方案。主要是在系统中固定在线流量和离线流量的比值来模拟静态带宽分离的效果。而本系统则动态地预测当前周期的在线流量需求的大小，然后将剩下的空闲带宽都分配给离线流量。

如图 2 所示，实验结果中，蓝线代表了在线流量占总带宽的百分比，黑线代表固定带宽分离方案下总的带宽利用率（在线流量和离线流量占总带宽的百分比），红线代表本系统方案下的总的带宽利用率。从图中可以看到,为在线流量与离线流量设置固定的比值 2:3 时，离线流量最多只能使用完属于自己的总带宽的 60%，尽管在线流量远未使用完分配给它的总带宽的 40%，离线流量也不能去借用这部分空闲流量，这就造成带宽的浪费。其中，图中黑线与蓝线形状相似是因为离线流量在每周期内都使用了 60%的带宽（其上限），也就是说黑

线是蓝线往上平移了 60%，这与我们的设想是一致的。

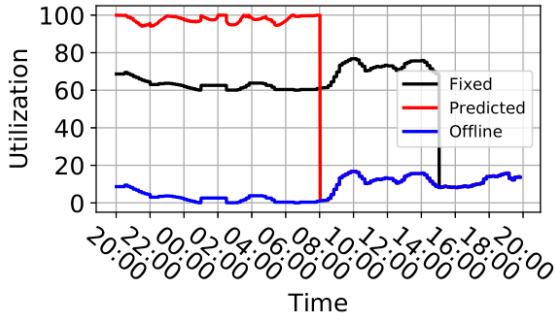

图 2 某条链路上的带宽使用情况

为了验证 Sliding-k 方法的预测效果，我们随机生成了分段的带“毛刺”的序列，首先使用 EWMA 对在线流量进行预测，当α设置为 0.3、0.5、0.9 时预测效果分别如图 3 图 4 图 5 所示，其中黑线代表真实值，红线代表预测值，可以看到α值为 0.3 时预测结果比较平滑但不够准确，特别是在拐点出现的时间点上，而α为 0.9 时预测值十分接近真实值，但这样的预测效果比较抖动，会引起频繁的重配置，增加更新压力，而α为 0.5 时，虽然拐点处的预测值不十分贴近真实值，但是能够兼顾预测平滑和预测灵敏这两点需求。

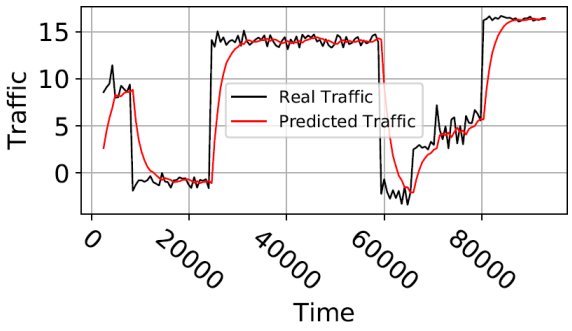

图 3 EWMA 参数 α 为 0.3 时的预测效果

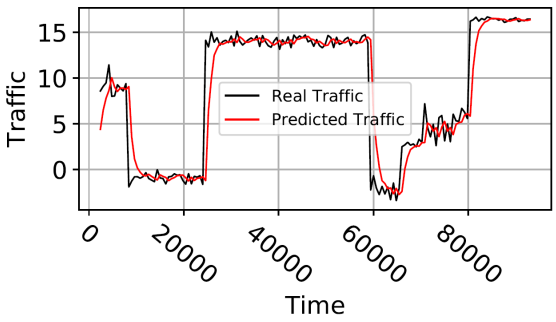

图 4 EWMA 参数 α 为 0.5 时的预测效果

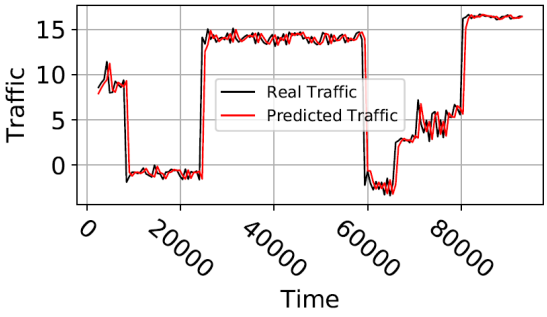

图 5 EWMA 参数 α 为 0.9 时的预测效果

由于想要实现预测值足够平滑，且在拐点处的预测足够灵敏准确，因此要使用本文提出的 Sliding-k 方法。当 EWMA 的参数值设置为 0.3 时，使用 Sliding-k 方法得到的预测效果如图 6，对比α同样为 0.3 但仅使用 EWMA 方法的预测效果（图 3）可以发现，Sliding-k 能够弥补 EWMA 的不足，使拐点出现时的预测效果更加准确，而当α设置为 0.5 时，Sliding-k 预测效果更好，如预测效果在无拐点时平滑，而在有拐点时敏感且准确，达到了预期的预测效果。

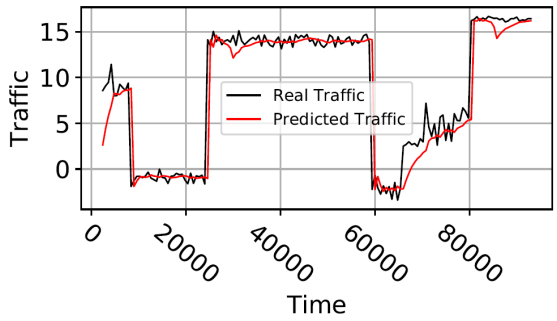

图 6 Sliding-k 方法且 α 为 0.3 时的预测效果

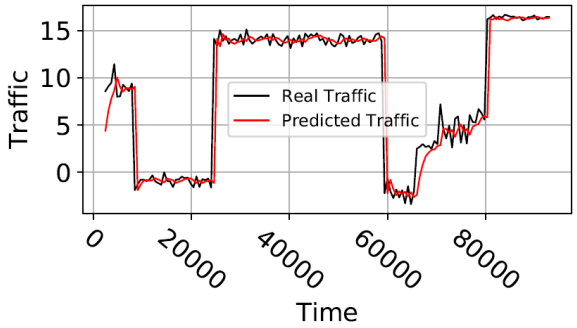

图 7 Sliding-k 方法且 α 为 0.5 时的预测效果

五、结论

传统固定带宽分离的方式会造成这样一种现象，当在线流量处于谷值的时候，那些预留给它的带宽由于已经被固定地分配给了在线流量而不能被大量离线流量充分利用，从而造成数据中心间链路带宽的浪费。而本文使用了动态带宽分离的方式，使用将 EWMA 和拐点检测算法结合的 Sliding-k 算法预测在线流量的带宽占用情况，在保证延迟敏感型的在线流量不受影响的条件下将剩余可用的链路带宽容量分配给离线流量。Sliding-k 算法在预测时先判定当前是否出现拐点，然后再决定预测时“回看”多远，这样即能够保证预测的灵敏，又能够在网络变化不大的时候避免频繁的重配置。这种基于预测的增量调度方式可以使链路利用率接近 100%，充分地利用了链路带宽。未来可以考虑使用 EWMA 之外的更好的时间序列预测算法来提高预测效果，使链路利用率更高而且更加稳定。

参考文献

[1] Hong C Y, Kandula S, Mahajan R, et al. Achieving high utilization with software-driven WAN[C]//ACM SIGCOMM Computer Communication Review. ACM, 2013,

43(4): 15-26.

[2] Kumar A, Jain S, Naik U, et al. BwE: Flexible, hierarchical bandwidth allocation for WAN distributed computing[J]. ACM SIGCOMM Computer Communication Review, 2015, 45(4): 1-14.

[3] Savage S, Collins A, Hoffman E, et al. The end-to-end effects of Internet path selection[C]//ACM SIGCOMM Computer Communication Review. ACM, 1999, 29(4): 289-299.

[4] Zhang H, Chen K, Bai W, et al. Guaranteeing deadlines for inter-data center transfers[J]. IEEE/ACM Transactions on Networking (TON), 2017, 25(1): 579-595.

[5] Zhang Y, Xu K, Wang H, et al. Going fast and fair: Latency optimization for cloud-based service chains[J]. IEEE Network, 2017, 32(2): 138-143.

[6] Jain S, Kumar A, Mandal S, et al. B4: Experience with a globally-deployed software defined WAN[C]//ACM SIGCOMM Computer Communication Review. ACM, 2013, 43(4): 3-14.

[7] Hong C Y, Mandal S, Al-Fares M, et al. B4 and after: managing hierarchy, partitioning, and asymmetry for availability and scale in google's software-defined WAN[C]//Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication. ACM, 2018: 74-87.

[8] Zhang Y, Xu K, Yao G, et al. Piebridge: A cross-dr scale large data transmission scheduling system[C]//Proceedings of the 2016 ACM SIGCOMM Conference. ACM, 2016: 553-554.

[9] Zhang Y, Jiang J, Xu K, et al. BDS: a centralized near-optimal overlay network for inter-datacenter data replication[C]//Proceedings of the Thirteenth EuroSys Conference. ACM, 2018: 10.

[10]Lucas J M, Saccucci M S. Exponentially weighted moving average control schemes: properties and enhancements[J]. Technometrics, 1990, 32(1): 1-12.

[11]Adams R P, MacKay D J C. Bayesian online changepoint detection[J]. arXiv preprint arXiv:0710.3742, 2007.

[12]Roberts S W. Control chart tests based on geometric moving averages[J].

Technometrics, 1959, 1(3): 239-250.

[13]Lucas J M, Saccucci M S. Exponentially weighted moving average control schemes: properties and enhancements[J]. Technometrics, 1990, 32(1): 1-12.

作者信息

题目：基于预测的 DC 间流量增量调度研究

主题：14.数据中心网络

姓名：王然

学位：硕士研究生

单位：北京邮电大学

地址：北京市西土城路 10 号（100876）

手机：13683391866

Email： wangranse@bupt.edu.cn

姓名：张宇超

职称：助理教授

学位：博士

单位：北京邮电大学

地址：北京市西土城路 10 号（100876）

手机：18600281690

Email：yczhang@bupt.edu.cn

姓名：王文东

职称：教授

单位：北京邮电大学

地址：北京市西土城路 10 号（100876）

手机：13901212701

Email：wdwang@bupt.edu.cn