封面

摘要（中文）关键词

摘要（英文）关键词

目录：

1. 绪论

1.1研究背景

1.2国内外研究现状及趋势

1.3论文研究意义及创新

1.4论文结构安排

2、研究基础

2.1数据中心网络模型基础

2.2拥塞控制算法及机制基础

2.2.1流量调度

2.2.2负载均衡

2.2.3发送速率控制

2.3神经网络算法基础

2.4NS3仿真平台基础

3.基于神经网络的拥塞控制算法研究

3.1算法整体设计及模块划分

3.2数据预处理模块设计

3.2.1算法设计

3.2.2算法分析验证

3.3RTT预测模块设计

3.3.1神经网络模型设计

3.3.2数据集、损失函数、优化器等参数设计

3.3.3算法分析验证

3.4PID速率控制模块设计

3.4.1网络模型搭建

3.4.2参数设计

3.4.3模型分析

3.5本章小结

4、神经网络模型训练仿真及算法比较

4.1 NS3仿真平台环境搭建及设计

4.1.1环境配置

4.1.2网络拓扑仿真设计

4.2神经网络模块化训练

4.2.1数据集获取及展示

4.2.2RTT预测模块训练效果展示

4.2.3PID速率控制模块训练效果展示

4.3算法拥塞控制效果展示

4.3.1RTT时延

4.3.2发送速率

4.4.算法分析比较

4.4.1分析比较指标

4.4.2算法比较

4.4.2.1TIMELY

4.4.2.2HPCC

4.4.2.3DCTCP

4.4.3比较结果分析总结

4.4.3.1算法优势及缺点

4.4.3.2算法创新点

5、全文总结及期望

5.1全文总结

5.2后续工作展望

致谢

参考文献

摘要:数据中心网络传输旨在实现低延时和无损行为。其中，RTT作为包往返时间，被证明和交换机队列长度有着很大的关联程度，是网络终端节点能够获取的为数不多的拥塞信号。目前随着网络硬件的进步，RTT的测量已经能够达到微妙级的精度。然后，本论文设计了一种神经网络拥塞控制算法，充分利用了数据包RTT时延，来对发送速率进行调整，保持低延时和高带宽。并在基于NS3搭建的仿真平台上部署了该算法。设计容易发生拥塞的INCAST通信模式，通过实验表明，它提供了很好的拥塞控制效果，平均时延控制在10us以下，在大规模数据注入时也没有发生拥塞。同时该算法不依赖于网络硬件的修改，在数据中心等网络大规模部署是十分容易的。

关键词：神经网络、拥塞控制、数据中心网络、INCAST、RTT

Abstract: Data center network transmissions are designed for low latency and lossless behavior. Among them, RTT, as the packet round-trip time, proves to have a great correlation with the length of the switch queue, and is one of the few congestion signals that network endpoints can obtain. At present, with the advancement of network hardware, RTT measurements have been able to achieve subtle levels of accuracy. Then, we design a neural network congestion control algorithm that makes full use of the packet RTT delay to adjust the transmission rate to keep the latency low and high bandwidth. We deployed the algorithm on a simulation platform based on NS3. The INCAST communication mode is designed to be prone to congestion, and experiments show that it provides good congestion control effect, the average delay is controlled below 10us, and no congestion occurs during large-scale data injection. At the same time, the algorithm does not depend on the modification of network hardware, and it is very easy to deploy it on a large scale in networks such as data centers.

Keywords: neural networks, congestion control, data center networks, INCAST, RTT

1. 绪论

1.1研究背景

根据2021年颁布的《数据中心智能无损网络白皮书》[1]，数据中心支持了很多类型的业务，包括大数据机器学习等；随着人工智能在数据中心的发展和应用，数据中心中传输的数据呈指数级的增长，海量数据的传送使网络承受更大的传输压力，数据发送的瓶颈逐渐从终端转移到网络设备上，低时延和无损行为成为了数据中心网络新的必要需求。

云时代的数据中心专注于应用转型和服务的快速部署。在 AI 时代，数据中 心提供了实现数字化生活所需的信息和算法。高速存储和人工智能分布式计算的 结合，将大数据转化为快速数据，供人、机、物访问。高性能、大规模、无丢包 的数据中心网络对数字转换的顺利进行至关重要。

人工智能、网络性能等高性能应用的关键指标包括吞吐量、时延和拥塞。吞 吐量是指快速传输大量数据的网络总容量。时延是指跨数据中心网络事务的总延迟。当流量超过网络容量时，会发生拥塞。丢包是严重影响吞吐量和时延的因素。

为应对这种变化趋势，数据中心使用了新远程直接内存访问（Remote Direct Memory Access, RDMA） 协议消除了数据副本，释放了 CPU 资源，能够完成路径选择和取出顺序计算，但是RDMA 效率的提高给网络带来了更大的压力，将瓶颈转移到数据中心网络基础设施上，这使得网络比以往更容易发生拥塞，而传统拥塞控制算法已经无法适应目前的数据中心数据传输，因此针对数据中心网络设计新的拥塞控制算法和机制就非常重要。

1.2研究现状及趋势

数据中心无损网络的拥塞控制，目前已提出了多种解决思路和方法，大体方向一共有三种，分别是仅通过端节点的拥塞控制算法、需要网络中交换机或控制器配合的流调度和负载均衡；通过端节点的拥塞控制算法又包含了基于丢包的TCP Cubic算法、基于ECN丢包的DCQCN[2]和DCTCP[3]算法、基于RTT信息的TIMELY[4]算法等，流调度包括了基于截止时间的D2TCP[5]、基于流大小的pFabric[6]等，负载均衡包括了基础简单算法ECMP和复杂算法Hedera[7].大部分拥塞控制算法由于需要特殊的硬件或者报文支持，如HPCC[8]算法需要INT报文字段，阿里提出的vFabric[9]需要能够发送特殊探针的边界路由器。

拥塞控制研究发展有以下几个方向，一是基于SDN的中心化控制，集中式控制可以对网络全局信息进行掌控，更能精确的规划控制流量，但同时面临着可部署规模的问题；二是人工智能与流量控制技术结合的研究方向，人工智能具有强大的自适应力和自学能力，能为拥塞控制提供一套有效的决策工具[10]，例如华为在2021年提出的ACC[11]中提出了强化学习的拥塞控制算法等。

1.3论文研究意义及创新

由于数据中心网络的数据传送的突发性、以及流种类的复杂性，目前的拥塞控制算法和机制主要有两个特点（缺点）：一是具有针对性，对于特定情况或特定流的拥塞处理较好，但是对于网络中的突发流量导致的拥塞情况效果较差；二是大部分算法都需要专门的硬件配合来获取所需的信息，但数据中心网络规模非常大，大部分未实现商用。

本论文基于数据中心提出了一种新的拥塞控制算法，结合了AI人工智能，设计了一个新的LSTM+PIDNN神经网络框架。解决了 上述两个缺点，首先神经网络具有很强的学习能力，能够根据网络状态变化调整参数，解决了大部分算法的“针对性”；同时神经网络的实现以及相应数据（RTT）的获取不需要通过专门的硬件实现，而只需要相应的软件修改；因此更容易大规模部署；

1.4论文结构安排

本文的章节安排如下：

第一章阐述了数据中心网络的背景及拥塞控制算法概况。对当前数据中心中拥塞控制算法进行了总结了分析，然后介绍了本论文的算法设计思想。

第二章讲述了数据中心网络的拓扑及通信模式的基本概念，以及不同拥塞控制算法的基本概念。包括对于CLOS拓扑和INCAST通信模式，以及DCTCP、TIMELY、HCPP、ECMP等。

第三章介绍了本论文设计的基于神经网络的数据中心拥塞控制算法，包括了模块划分，模型设计搭建及分析验证，以及数据集设计和网络训练方式。

第四章介绍了算法仿真效果。包括网络拓扑搭建及模型训练，效果分析。

第五章对算法效果和其他拥塞控制算法进行比较，主要指标包括平均RTT、发送速率等。

第六章对全文进行了总结。

2、数据中心拥塞控制基础

2.1数据中心网络模型基础

2.1.1数据中心网络拓扑

新一代数据中心网络架构大多数都为CLOS网络架构，CLOS是一个对称的“接入-核心-接入”网络架构，具有以下特点

* ·多级交换，典型为三级交换架构。
* 在每一级的每个单元都与下一级的设备全连接。
* 到指定目的地，在第1级交换单元存在多条路由，而后续交换单元都只存在唯一的一条路由。
* 严格意义上的无阻塞。
* 支持递归，可无限扩展。



目前，基于CLOS架构搭建的数据中心网络拓扑有Fat-Tree和Spine-Leaf

2.1.2数据中心网络协议

高带宽、低延迟是目前数据中心应用的基本需求。RDMA 通过 Memory Region 机制使得网卡能够直接读写用户态的内存数据，避免了数据拷贝和上下文切换；并将网络协议栈从软件实现 offload 到网卡硬件实现，极大降低了 CPU 开销。随着 RoCEv2（RDMA over Converged Ethernet v2）技术的成熟，RDMA 可以部署在数据中心已有的网络设施上，RDMA 成为数据中心高速网络通信的主流方案。



使用RDMA协议有以下优点：

* 零拷贝(Zero-copy)
* 内核旁路(Kernel bypass)
* 不需要CPU干预(No CPU involvement消息基于事务(Message based transactions)
* 支持分散/聚合条目(Scatter/gather entries support)

2.1.3数据中心网络对于拥塞控制协议的要求

由于RDMA的重传机制，丢包对 RDMA 造成大幅性能下降，因此 RDMA 要求 L2 层或者 L3 层网络是无损的。



丢包对 RDMA 吞吐量的影响

由上图可知，RDMA吞吐量对于丢包率的增长是十分敏感的，当网络丢包率到1%,时，RDMA的读写速率下降到几乎为1零，只有当丢包率保持在一个非常低的比率时，才能够完整发挥RDMA的性能；

为了实现不丢包，数据中心网络RoCEv2引入了基于优先级的流量控制PFC协议，基于优先级的流量控制（PFC）功能是当接收设备输入缓存区的占用超过设 定阈值时，暂停上游发送设备，防止因缓存区溢出造成丢包。虽然提供了 RoCE 所必需的无损环境，但 PFC 的大规模使用也存在一些问题，包括 PFC 死锁的可能性。

因此，单是上述的PFC协议是无法避免网络拥塞的，数据中心网络还需在网络上的各个节点采取不同的拥塞控制协议或者机制，多者之间联合调控。

2.2拥塞控制算法及机制

根据算法执行的位置进行分类，拥塞控制算法主要分为在终端节点控制的拥塞控制协议和在交换机等交换设备上实现的流量工程；两者虽然实现在CDN网络中不同的位置节点，但是都是为了合理调度网络资源以减少网络拥塞情况的发生。下面将具体进行介绍

2.2.1拥塞控制

在终端节点是拥塞控制主要就是根据获取的网络信息来控制TCP层发送速率，网络信息大多数从接收到的数据包中某个标志字段分析得出，大致分为三类：基于丢包率（TCP cubic）、基于ECN标志字段(DCTCP,HULL,DCQCN)和RTT(TIMELY);

2.2.2流量工程

流量工程主要是实现在网络中交换机等交换设备上，主要分为流调度和负载均衡两种；流调度以流为调度单位，主要在端口队列实现。通过某种机制或算法控制流的发送数量或顺序，有基于截止时间的D2TCP、基于流大小的pFabric和基于链路价格的FCP等；而复杂均衡则是控制流或数据包走不同的网络链路，避免大量数据占用同一条链路资源，使得不同物理路径之间负载平衡，例如ECMP算法；

除了拥塞控制和流量工程之外，有些拥塞控制算法结合了二者；例如阿里巴巴提出uFAB算法，需要终端和边缘交换机联合调控，边缘交换机发送探测数据包收集网络信息，后再将统计信息发送给连接的终端节点，终端节点根据网络状态控制流发送速率；同时交换机还可以根据网络状态控制流发送路径。



2.3算法评价分析

目前数据中心拥塞控制算法种类繁多；其中大部分对于特定拥塞场景或者对于特定东西向流（视频流、语音流、AI流等）有着很好的控制效果，但是不能够作为处理拥塞的普适算法；其中一些算法需要对网络协议中报文字段进行增加或者对设备硬件进行修改，这些算法在拥塞处理上结果很理想，能够达到很好的预期，但是往往不能够大规模部署，停留在实验阶段。很多拥塞算法参数固定，对网络状态变化适应能力差，速率控制比较机械，不够灵活，不能够适应智能化网络发展趋势。

2.4本章小结

本章先介绍了数据中心网络协议模型，再根据种类介绍了多种拥塞控制算法，并对他们的优缺点进行分析总结。

3.基于神经网络的拥塞控制算法研究

3.1算法设计思路及模块划分

基于神经网络的拥塞控制算法是主要使用神经网络来控制终端节点数据发送速率以达到控制拥塞的效果，通过ACK控制报文的收集，统计出网络数据包的往返时延RTT，作为算法的输入，再经算法计算得出结果，输出作为速率调节因子。

使用神经网络实现拥塞控制算法主要为了利用神经网络的预测和学习功能，数据中心流量大，当检测拥塞发生时，到调用控制算法进行调节时，很有可能已经错过了拥塞控制的最佳时机，控制策略具有一定的滞后性，一是检测拥塞到拥塞信号传递给终端之间需要一定的时间，二是拥塞控制算法开始到作用到网络交换机拥塞点需要一定的传播时间；这两者是造成滞后性的两个方面，尤其是第一部分最为重要，往往越拥塞，拥塞信号传递给终端的时间就越长；为此，使用神经网络对网络拥塞情况进行预测，提前预知拥塞的发生，从而提前作出反应。

传统拥塞控制算法都没有对网络的将来状态进行预测，一是考虑到数据中心网络状态复杂、包括有网络拓扑、发送速率、丢包率、ECN标记等，二是预测需要比较复杂的数学模型进行学习，而传统算法绝大很多参数已经固定，无法准确预测，而预测的准确度又对控制效果起着决定性作用。采用神经网络则克服了第二个问题，对于一中复杂的状态，论文【】中指出网络RTT和队列长度之间有着很高的相关性，而队列长度决定了网络拥塞情况，因此本论文将网络的拥塞状态用RTT表示，将本来的多维度预测变为一维预测，同时从结果上表明了仅通过准确的RTT预测，就能够使控制算法更好、更准确的控制速率。

由于数据中心网络中RTT短时间内变化剧烈，有着很高频率的抖动，因此不能够对原始RTT进行预测，数据预处理模块算法对原始数据进行一定的处理，使得由RTT数据变为能够训练和测试网络的数据集。

当实现预测之后，还需设计一个速率控制，终端对于拥塞的控制主要就是通过调节发送速率来实现的，这里使用神经网络搭建PID速率控制器，利用神经网络的学习功能去优化控制参数，尽可能实现更好的控制效果；相较于其他控制算法，使用PID算法对于RTT变化反应灵敏，且不需要很多参数，同时也能够控制发送速率使得真实RTT向目标收敛。

基于上述分析，·神经网络拥塞控制算法从功能划分上主要分为四个模块：

* 网络信息采集模块
* 数据预处理模块
* RTT时间序列预测模块
* PIDNN速率控制模块



每个模块实现各自功能，按照顺序相互配合，及算法按照“网络信息采集模块->数据预处理模->RTT时间序列预测模块->PIDNN速率控制模块”的流程运行，后两个模块需要其提前将网络参数进行训练才能够使用。

3.2网络信息收集模块设计

DCN网络数据量庞大，单位时间发送数据包数据巨大，因此对每个数据包RTT都执行算法是不现实的，会浪费大量的CPU计算资源，因此设计了RTT的采集策略，降低了速率调节的频率。

**RTT采集策略**：每一个RTT完成时间内只统计第一个数据包：



图X、RTT统计方式

例如图X所示，发送方一共连续发送了5个数据包（D1 D2 D3 D4 D5）并收到了数据包D1的确认ACK1，完成时间定义为数据包从发送到收到ACK确认之间的时间，即T0到T3的时间，而这个时间段内发送方还发送了D2 D3 D4三个数据包，根据统计策略，在这个时间段内，只会统计D1的往返延时RTT1，而不会统计D2 D3 D4这三个数据包的往返延时，T3之后则会统计D5的往返延时，同时不会统计从收到D5确认报文ACK5时间段内发送的其他报文RTT.采取这样的策略主要考虑以下几点优点：

* 由于CDN网络发送速率大，报文数量多，速率控制次数和频率是由统计的RTT次数和频率决定的，即每统计一次RTT，就需要根据算法进行一次数据调节。因此如果每一个报文都去进行统计，则速率越大，算法运行频率越快，这对于终端计算负载很大。因此采取此策略降低算法运行频率，以缓解计算压力。
* 降低了对于算法计算时间限制，设运行一次拥塞控制算法所需的时间为T,如果采取逐包预测，则该时间T则需要小于连续接收到的两个ACK时间间隔，否则上一数据包算法还没有计算完，下一个数据包的RTT已经能够计算出来，预测模块便失去了作用，且整体算法跟不上ACK的频率。如果采取每一个RTT完成时间内只统计第一个数据包，只需要满足时间T小于第一个数据包的完成时间；实际情况中远大于；这放宽了对与算法时间上的限制，因此可以采用更为复杂的算法去实现控制，提高了控制的准确度。
* 降低了算法运行频率同时也降低了发送速率变化频率，这使得发送窗口大小更加稳定。
* 前后连续几个报文的RTT几乎相等，该策略降低了相同数据的重复运算的次数。
* 避免了不同RTT之间统计数据的交叉储存。

但该采集策略也有以下缺点：

* 采集频率不固定，收到数据包完成时间的影响，完成时间越大，采集频率越低，完成时间越小，采集频率越高；而完成时间受网络状态影响最大，因此需要尽可能保证网络的稳定性，需要依靠PID算法控制速率，使得RTT收敛到一个较低的范围。
* 当网络拥塞时，RTT时延迅速增加，完成时间也会大幅增加，这样会导致采集频率降低，速率控制频率降低；该情况可以通过拥塞预测解决，因此需要很低的预测偏差（MAPE在10%以下）。

**RTT计算方法：**

根据图X中统计数据可知，

根据数据数据包发送时间统计T0，在仿真平台中，由于发送的并非真实数据包，因此接收端在收到发送方数据包时不需处理直接返回相应的ACK报文，因此，当发送方根据报文序列号确认收到ACK为被采集的报文，统计出T3，得出采集RTT值；采集完一个值后，下一个发送的数据包则作为下一个被采集的报文，更新T0，以此流程反复，直到整个流完全发送完毕。

3.3数据预处理模块设计

3.3.1算法设计

数据预处理对于后续的预测准确度影响很大，选择一个合适的处理算法，能够避免网络训练发生过拟合和欠拟合的情况。

数据预处理模块包括了单节点数据包时延平滑处理、时序特征计算和采样；

**平滑处理：**

数据中心RTT前后变化幅度很大，一般是以微妙us为单位，为了解决RTT的剧烈变化，因此显示对原始数据进行平滑处理，处理公式如下：

是信息采集模块采集的第个RTT, 是处理后的光滑数据，是光滑因子，一般取0.2；

这里的t并非真正意义上的时间，而是时间序列的顺序，即;

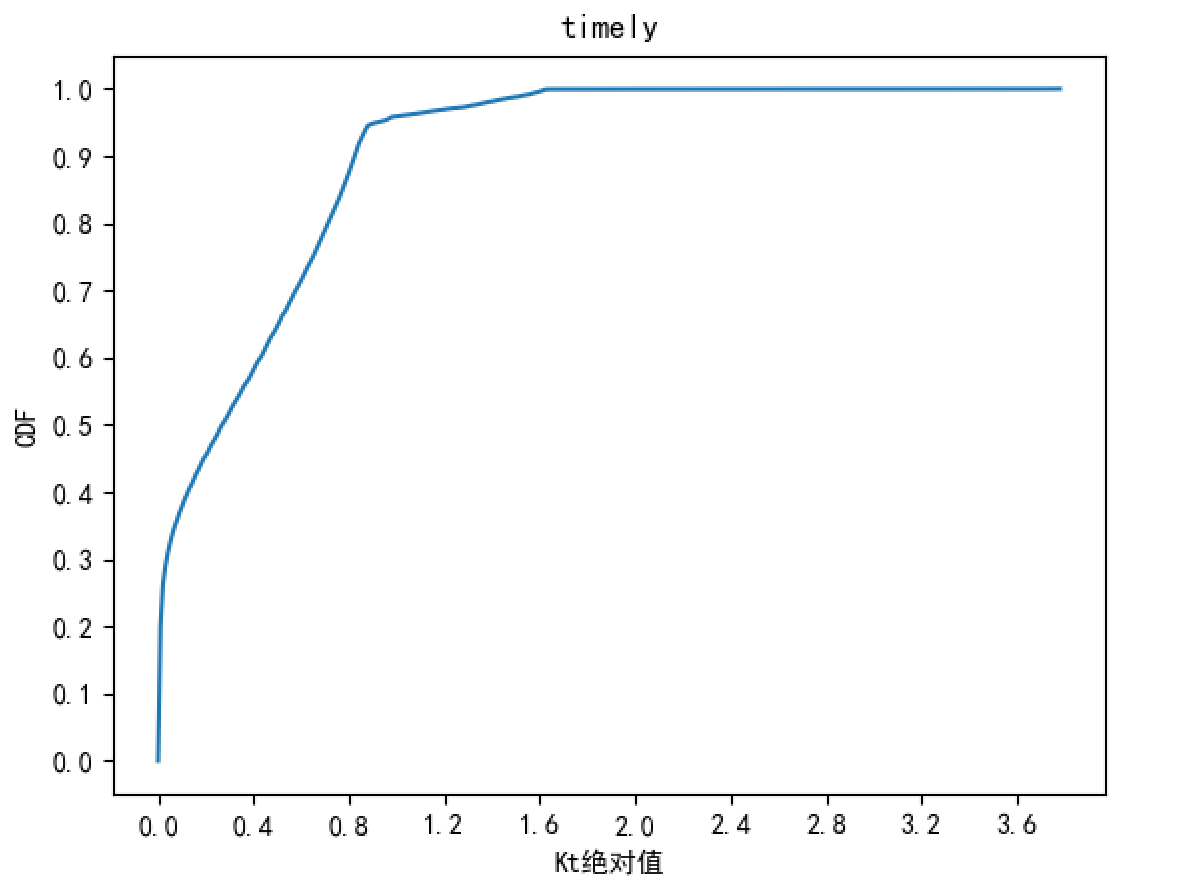
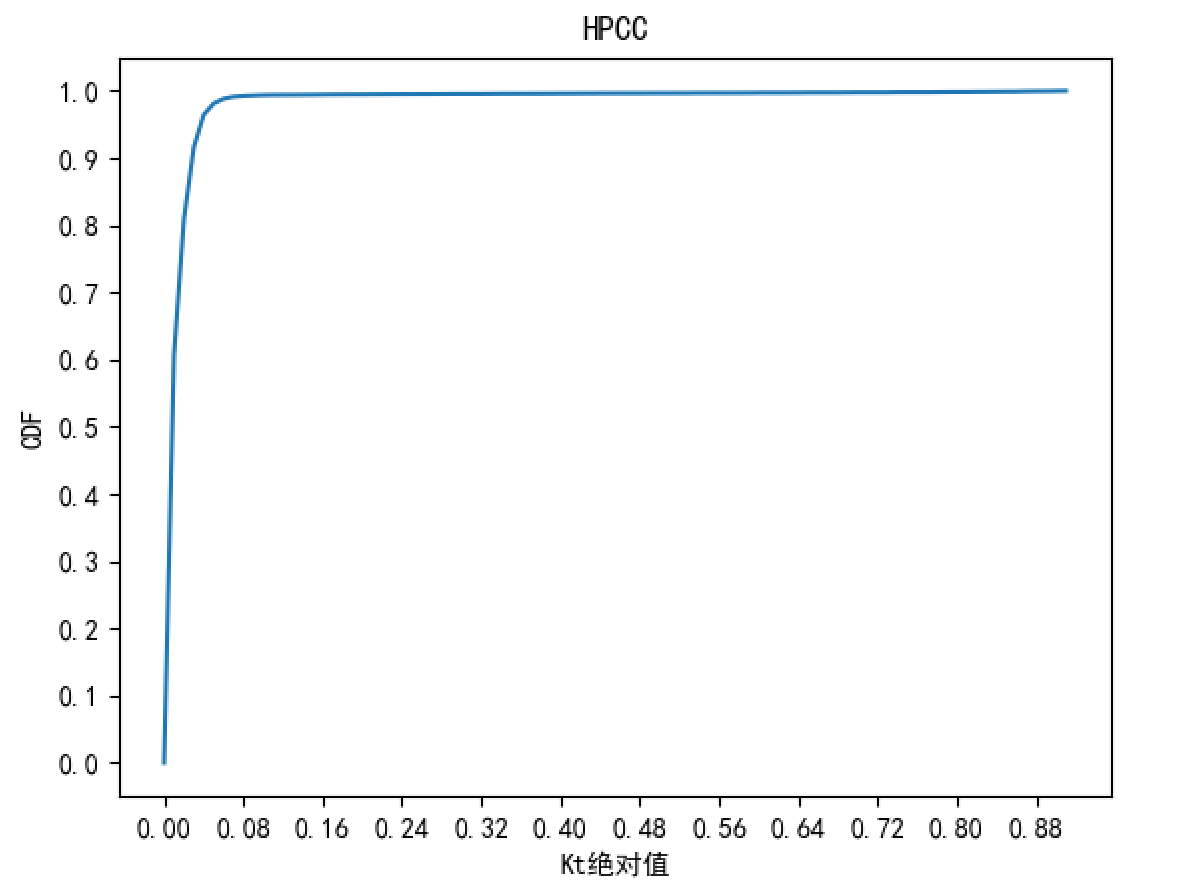
**时序特征计算：**

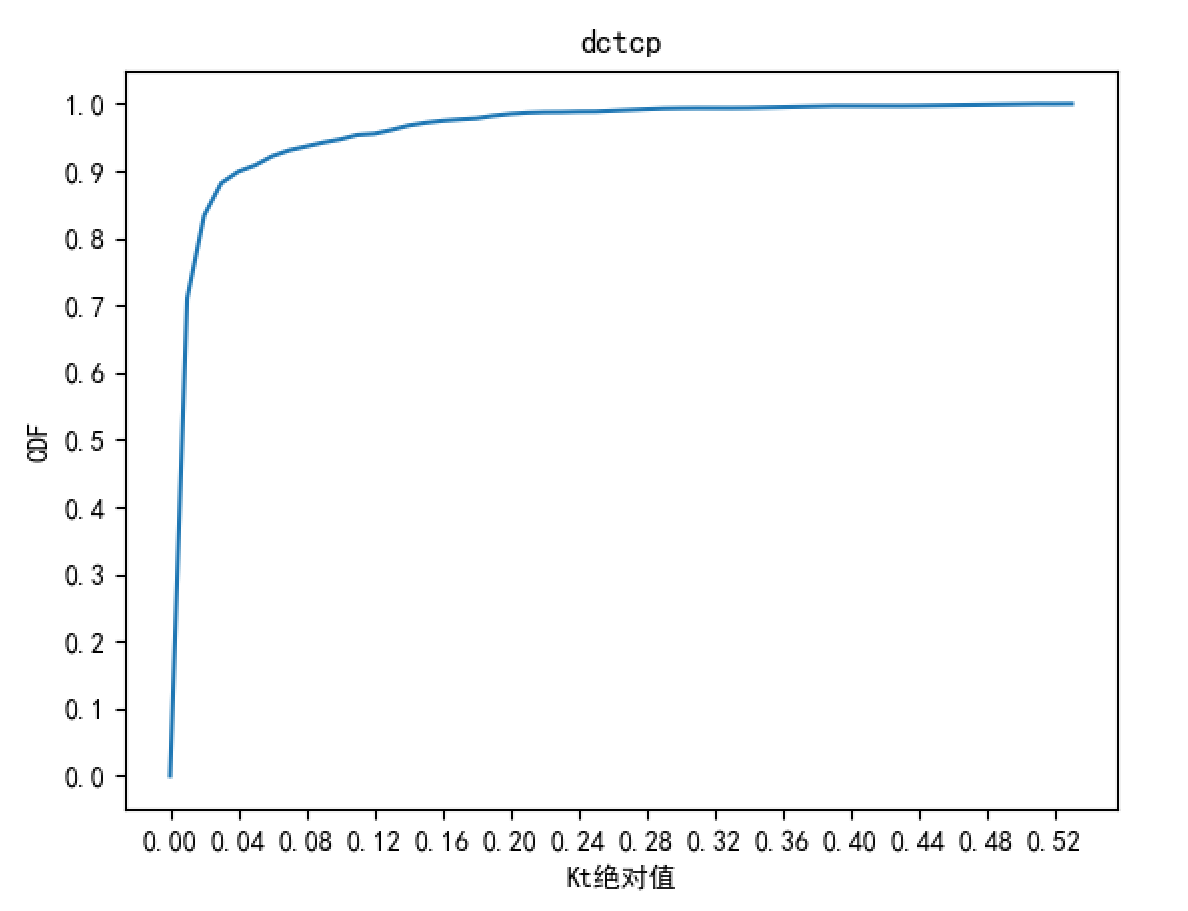
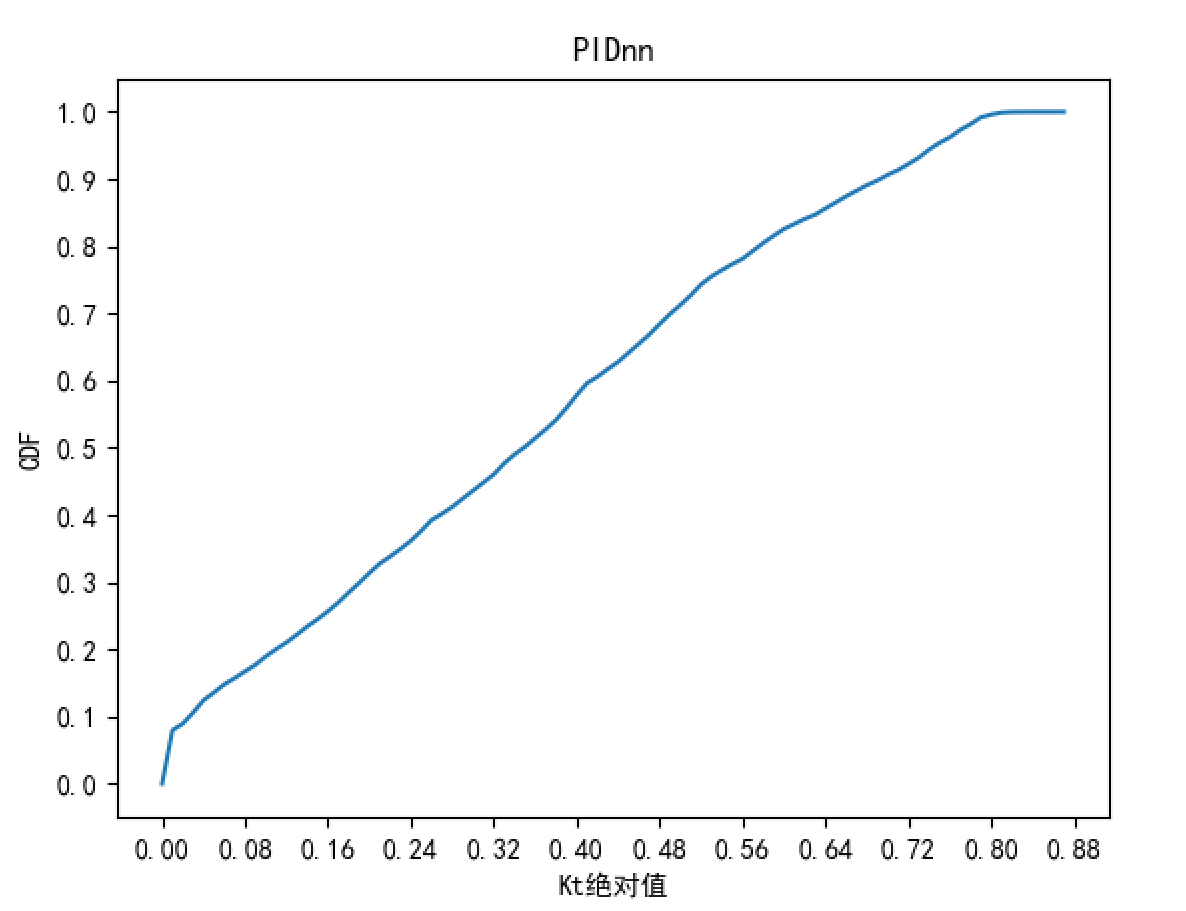
由于平滑因子σ接近于0，平滑效果很好，但是处理前后RTT损失了很多时序特征，如果不考虑损失的时序特征，会导致神经网络预测效果大幅度降低。因此根据平滑前后值，设计了一种算法，提取其特征值，计算方法如下：

即为采集模块采集的第个RTT与对应时刻平滑值的变化幅度特征,保留了平滑处理前后损失的时序抖动，因此通过平滑值及其特征就能够得出真实值；而为采集模块采集的第个RTT与对应t时刻平滑值的变化幅度特征,是对未来变化的一种预测特征，即在真实过程中的大小是未知的，通过对的预测，再根据当前时刻的平滑值，得到对未来RTT真实值的一个预测。在实际RTT时间序列预测过程中，就是通过连续多个已知 对进行预测，从而实现对下一时刻RTT的预测。

**阈值划分及特征采样：**

并不是越多的数据都用于训练预测效果越好，观察多个拥塞控制算法采集后的变化幅度特征发现，大部分的变化幅度特征值的大小都聚集到一个很小范围，造成这种情况的原因是不同的拥塞控制算法对于RTT控制各有特点（例如HCPP算法目的是为了尽量让交换机队列保持空状态，因此RTT大多数情况很低，变化幅度特征值大多数情况趋近0，而TIMELY算法根据前后的RTT变化梯度进行速率控制，这样就会导致RTT有很强的抖动，因此变化幅度特征值的绝对值大多数时间保持在一个很高的值），但无论哪儿个拥塞控制算法都会使变化幅度特征值的大小分布不均匀，这样往往会导致过拟合的结果（即训练集预测指标很高，而测试集指标远远低于训练集）。观察多个算法(HPCC,TIMELY,DCTCP以及未优化参数的PIDnn)运行过程中特征值的大小并统计，CDF（累计分布函数cumulative distribution function）统计结果如下：



这四个图能不能画到一起

观察统计结果可知;HPCC算法90%的特征绝对值和DCTCP算法 87%都处在一个较低的范围（0.，0.04），而TIMELY算法特征绝对值主要分布在（0.08，1）范围内，但在（1，3）范围内仍有分布；未优化参数的PIDnn分布相对均匀。如果将所有的数据都作为数据集训练网络，则大量的低绝对值会造成网络的过拟合；为了避免该情况的发生，根据特征值绝对值的大小划分为四个范围，为[0,0.02）、[0.02，0.08）、[0.08，0.15）、[0.15，inf）,并根据从每个范围中采样出近似数量的数据作为最终用于训练和测试神经网络的数据集。

3.4RTT预测模块设计

3.4.1神经网络模型设计

**神经网络模型选择：**

预测模块受到以下两方面的约束：

* 数据中心RTT的量级在微妙，因此预测模块需要很快的前向传播时间，在下一个RTT到达之前预测出结果，因此网络模型不能够过于复杂或庞大，否则预测也就失去了意义
* 速率控制要求预测模块预测出RTT变化趋势，因此需要较低的预测偏差，这对拥塞控制的有着决定性影响，因此当网络模型过于简单造成准确度不足时，预测反而对速率控制有着负作用。

通过比较LSTM、GRU、Simple RNN三种神经网络：

* Simple RNN模型最为简单，预测效果最差，多次训练后LOSS值很大
* GRU模型较为简单，同等规模下预测效果较好，但效果不如LSTM，没有达到预期
* LSTM模型复杂，预测效果最好

最终实现的预测神经网络总共有两层隐藏层

第一层为三LSTM神经元（根据LSTM传播特性可知，LSTM前向传播时间与该层神经元数量相关，考虑到LSTM内部复杂性、约束限制和前后特征相关性，只使用三个神经元），隐藏层维度为16；输入为三个连续的。

第二层为linear层，负责将多维转化为一维时序特征，并作为预测输出。



图X、时间序列预测网络

其中LSTM神经元有三个输出门，分别是遗忘门、输入门和输出们；神经元的输入包括时间序列,上一个神经元输出结果和上一个细胞元状态；

遗忘门该门会读取和，输出一个在 0到 1之间的数值给每个在细胞状态中的数字。1 表示“完全保留”，0 表示“完全舍弃”;，计算公式如下,其中为遗忘门输出，：

输入门负责下一步决定让多少新的信息加入到 cell 状态中来。实现这个需要包括两个步骤：首先，一个叫做“input gate layer ”的 sigmoid 层决定哪些信息需要更新；一个 tanh 层生成一个向量，也就是备选的用来更新的内容， 。再下一步，把这两部分联合起来，对 cell 的状态进行一个更新，现在是更新旧细胞状态的时间了，更新为。前面的步骤已经决定了将会做什么，现在就是实际去完成。把旧状态与相乘，丢弃掉我们确定需要丢弃的信息。接着加上。这就是新的候选值，根据我们决定更新每个状态的程度进行变化，计算公式如下：

输出门需要确定输出什么值。这个输出将会基于我们的细胞状态，但是也是一个过滤后的版本。首先，我们运行一个 sigmoid 层来确定细胞状态的哪个部分将输出出去。接着，我们把细胞状态通过 tanh 进行处理（得到一个在 -1 到 1 之间的值）并将它和 sigmoid 门的输出相乘，最终我们仅仅会输出我们确定输出的那部分,计算公式如下：

上述公式中，、、、是神经元中的权值参数，、、、是神经元中的偏置参数，预测神经网络的训练目的就是对这两种参数的优化更新。而和是两个激活函数，前者保证输出值在0到1之间，后者保证输出值在-1到1之间，两者都是用来加入非线性因素的，提高神经网络对模型的表达能力，具体计算公式如下：

LSTM神经元结构图如下：



图X、LSTMCell内部结构

3.4.2数据集、损失函数、学习率等参数设计

结合数据预处理和预测神经网络，设计的参数如下：

数据集：

从数据预处理模块和RTT时间序列预测模块，用到的网络信息只有采集到的数据包往返时延，使用不同拥塞控制算法，数据发送速率不同，所造成的数据包RTT也不同，但是从神经网络的角度分析，它对与RTT的预测不考虑用了什么拥塞控制算法，只关心网络的模型、网络的输入输出和最终的预测效果，RTT数据的来源算法是不重要的，只需保证此RTT是真实仿真环境下的统计输出，因此在获取数据时，采集出的RTT数据来自于了论文后续优化PIDNN参数时的仿真数据和使用TIMELY算法的仿真数据，总共有19058对数据。

Input:输入为三个连续的特征值、

Label:标签设置为当前时刻的预测特征

Train:每轮从数据集中随机选择800对数据进行训练

Test: 每轮从数据集中随机选择200对数据进行训练

Epoch:由于数据集大小限制，设置训练19轮

损失函数：

学习率：0.001

根据预测神经网络的输出out和公式X可知，(t+1)时刻RTT的预测结果如下:

3.4.3算法分析验证

在19轮数据集训练过程中，训练集损失loss变化如下：



图X，训练集loss变化图

上图为数据集训练过程中整体loss变化图（做了一个时间平均处理），横坐标为训练次数，纵坐标为训练loss,观察上图可知，随着训练次数的增加，训练loss逐渐减小，且下降速度由快到慢，后趋于平稳，说明训练模型参数逐渐向最优值收敛。

使用[MAPE(Mean Absolute Percentage Error,平均绝对百分比误差)](http://www.baidu.com/link?url=5hBZCJkzkfqi1vOMeVt5LfKSLbuF5iTbJ7Gg_7bsT3zcyThn-isTSjC25gAIwdbCnUvMyv8EzsTYaBYxeDMStw3fA-1lH37Tg9rYoivaAqpB87FroyukbdPcVRMyP9VF" \t "_blank)评价指标评估预测效果

MAPE值越小，说明预测结果和真实结果越接近，预测越准确，每一轮预测统计MAPE值如

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MAPE/轮数 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 训练集 | 0.113 | 0.089 | 0.083 | 0.064 | 0.061 | 0.050 | 0.052 | 0.055 | 0.051 |
| 测试集 | 0.081 | 0.079 | 0.059 | 0.083 | 0.058 | 0.067 | 0.053 | 0.045 | 0.044 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MAPE/轮数 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| 训练集 | 0.046 | 0.044 | 0.042 | 0.039 | 0.040 | 0.042 | 0.040 | 0.040 | 0.036 | 0.035 |
| 测试集 | 0.042 | 0.045 | 0.039 | 0.044 | 0.037 | 0.043 | 0.035 | 0.038 | 0.042 | 0.036 |

观察上表可以看出，随着训练轮式的增加，整体MAPE指标在逐渐降低，说明预测神经网络预测效果越来越好，同时可以看到初始几轮的MAPE值“较低”，但真实效果却并非“较好”，原因如下：

预测神经网络并非对未来时刻RTT的直接预测，而是对特征值的预测再通过公式X计算出对RTT的预测值，因此根据公式X可以看出，其MAPE指标在计算过程中经过了一定的转化，因此看到初始几轮的MAPE值“较低”而非真正的较低，从数值上知识最终几轮MAPE指标的3倍左右，但实际预测效果差距非常大

3.5PID速率控制模块设计

在一些自动化设备控制中，由于被控变量和控制变量之间没有准确的数学映射关系，因此常常采用PID模糊控制算法，经过经验修改PID算法参数，使得该模糊控制算法通常能够发挥很好的控制效果；而且算法结构简单、稳定性好、工作可靠、调整方便。

本论文将PID算法中各个控制模块作为神经网络的神经元，将该算法实现为有一种神经网络，通过反向传播学习的方法替代人工经验调整参数，使其能够适应网络状态变化，在预测模块的配合下，能够对未来拥塞情况或趋势进行提前的预知，提前调节速率进行防范，从而达到良好的拥塞控制效果。

3.5.1网络模型搭建

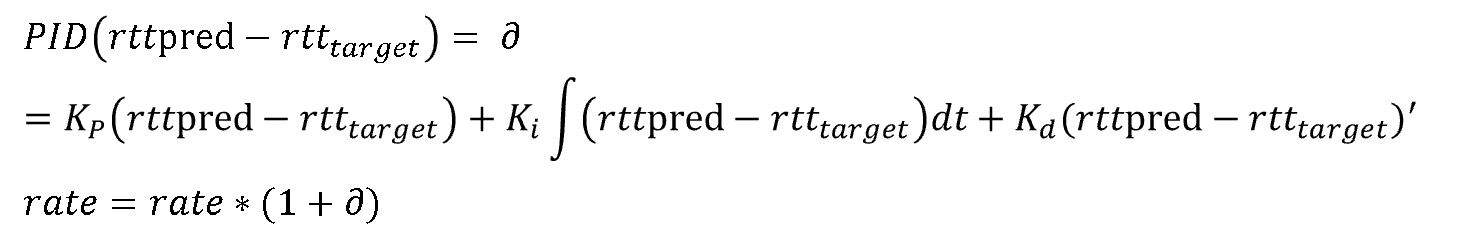
速率控制模块使用PID算法进行速率调节，上一个预测模块的RTT预测值作为该模块的输入，输出为速率调节因子。

PID控制算法总共包括了三个小模块，差分模块、积分模块和为分模块；

差分模块：最重要的调节因子，根据实际输入和目标值之差的特定比例作为调节因子

积分模块：对实际输入和目标值之差进行一定时间的累计求和，作为长时间调节因子

差分模块：实际输入和目标值之差进行时间求导，作为瞬时调节因子。



公式需要改，rttpred 改为rttnew



PID神经网络虽然隐藏层只有一层且只有三个神经元，但是每个神经元实现的功能不同。

3.5.2参数设计

PID算法不能够之间拿来使用，算法中四个参数KP、KI、KD、RTTtarget需要调整到合适的范围才能够对拥塞进行控制，其中前三个参数可以通过仿真平台进行实时自动修改调整，而后一个参数则需要人工经验手动修改。

根据误差公式：

通过反向求导对参数KP、KI、KD进行优化

由公式xxx可知，根据误差对三个控制参数求导：

其中，根据公式XX可得到：

但是无法精确求出，由代替，而仿真环境中，由于rtt和rate抖动较大，因此为了参数的快速收敛，将该值替换为一个接近1的常数

参数更新：

其中为参数更新的学习率，可以设置为一个很小的常数；当参数收敛困难时，也可以设置为一个随时间减小的变量;本论文中为了加快参数的收敛，更新率设置如下：

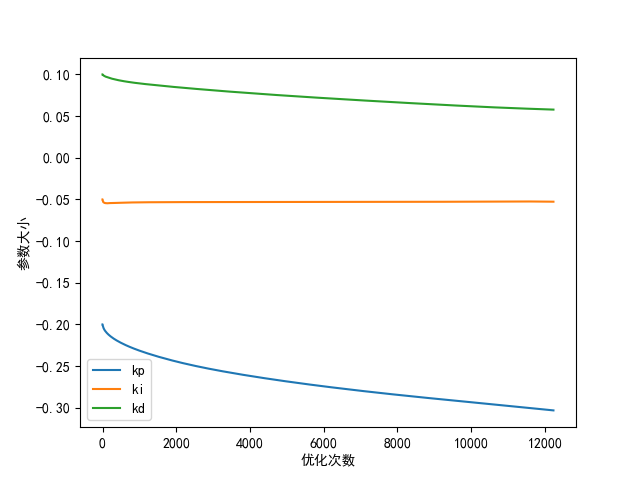
当训练时间较短时，参数更新快，而随着时间增加，参数逐渐趋于稳定，不再更新。

观察网络环境可知，当t较小时，网络数据包发送处于起步状态，网络状态波动较大，因此RTT震动幅度加大，对于参数算影响较大，同时此时刻学习率较大，这种情况容易导致参数更新发生突变，这对于PID参数的学习更新会造成严重的错误，可能导致后续参数不断负优化，为了避免这种情况发生，设置求导梯度范围为在[-0.1，0.1]之间.

3.5.3参数优化

PID速率控制模块训练主要基于在线的方式，因为每次反向传播中的误差计算需要下一时刻的rtt,而下一时刻的RTT又收到速率控制模块的影响，因此不能够通过其他的拥塞控制算法（TIMELY）获取的下一时刻RTT数据进行训练，只能使用PID算法；因此在训练时必须使用在线的方式，在仿真平台中部署还未训练的PIDNN，然后使用该速率控制模块调节发送速率，并根据该算法调节下的下一时刻RTT通过反向传播参数梯度求导的方式，是用公式X对参数进行优化，当多轮的训练后PIDNN参数收敛，在加入数据预处理模块和训练好的时间序列预测模块，形成整体算法。

PID参数的初始化对于参数后续优化影响很大，根据经验判断参数的大致范围，以为列,设置初始值,优化过程中参数变化如下：



图X，PID算法参数优化曲线

根据上图可知，随着学习次数的增加，三个参数逐渐趋于稳定，最终优化结果为；

3.6本章小结

本章将设计的基于神经的数据中心网络拥塞控制算法进行模块划分，介绍各模块的意义、功能、算法实现；其中包括了基于神经网络实现的RTT数据包往返时延预测模块以及PIDNN速率控制模块，使用这两个模块需要提前对网络参数进行训练优化，并根据网络特征采取不同的优化方法；最终完成各个模块的搭建。

下一步需要将各模块先后部署到NS3仿真平台进行功能等相关测试，验证通过后合并所有模块实现整体算法部署到NS3仿真平台进行算法验证。

4、神经网络模型训练仿真及算法比较

4.1 NS3仿真平台环境搭建及设计

4.1.1环境配置

NS3仿真平台基于LINUX系统中的C++实现，系统复杂，要求配置特定的环境变量，在实现算法仿真前需要将相应的编译环境及环境变量完成配置（仿真平台环境配置不包括神经网络训练所使用到的环境）。

具体配置如下：

虚拟机平台：VMware Workstation 16 pro

镜像：ubuntu-22.04-desktop-amd64.iso

C++版本：gcc version 4.8.5

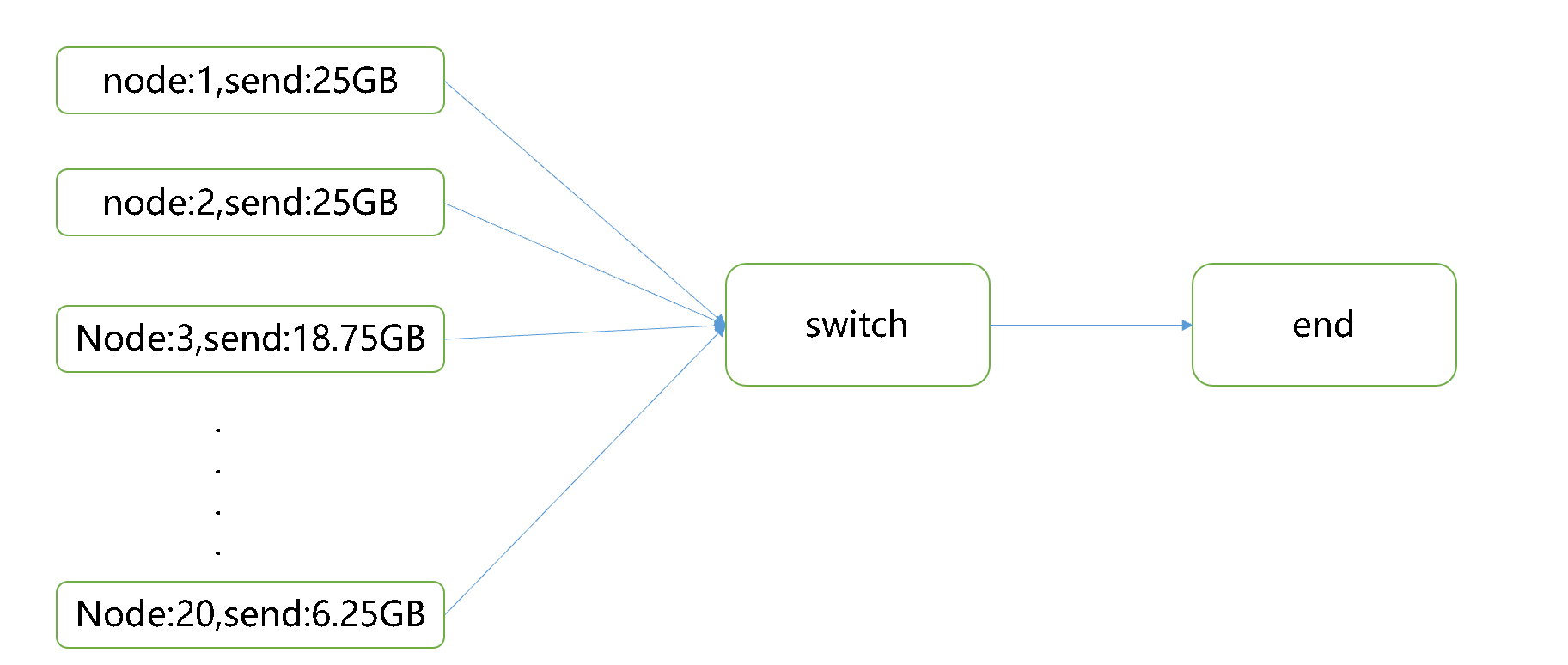
Python版本：Python 2.7.18

仿真平台：基于NS3 3.17的 [HPCC: High Precision Congestion Control (SIGCOMM' 2019)](https://rmiao.github.io/publications/hpcc-li.pdf)

仿真代码编写软件：Visual Studio Code 1.74.3

4.1.2网络拓扑仿真设计

为了尽可能地模拟数据中心拥塞场景，在设计仿真拓扑时选择了最容易发生拥塞的INCAST通信场景，其中设计20个终端节点像一个目的发送大量数据包，每个终端节点连续发送同一条流，并使用一个交换机将该20条流进行合并，使用一条链路传递给终端，该通信模式十分考研发送端对于发送速率的控制（因为交换机的流合并处理方式，因此发送速率的控制决定了合并出的拥塞程度，避免了交换机上相关的负载均衡协议或者机制对模拟的影响），具体网络拓扑和流大小如下图所示：



总共20个节点，每个节点负责发一条大数据流，其中前三个节点发送流的大小分别是25GB、25GB、18.75GB,其余17条流的大小都为6.25GB，网络发送速率限制为1Mbps-100Mbps;初始发送速率为1Mbps.

4.2神经网络模块化效果展示

将整体算法分为三个部分分别展示效果，即数据采集、RTT预处理预测和PID速率控制，一方面为了验证各模块在仿真平台上是否满足设计预期功能，另一方面同时观察各自效果，便于模块修正，最后再将三个部分同时部署到仿真平台中。

4.2.1数据集获取及展示

RTT时间序列预测模块需要根据设计的采集策略获取原始RTT值，基于NS3平台搭建好仿真模型，先后部署TIMELY/ DCTCP拥塞控制算法，并搭建设计的网络拓扑，分别对20个节点进行RTT采集，由于流数量较多，这里仅展示节点1采集数据。



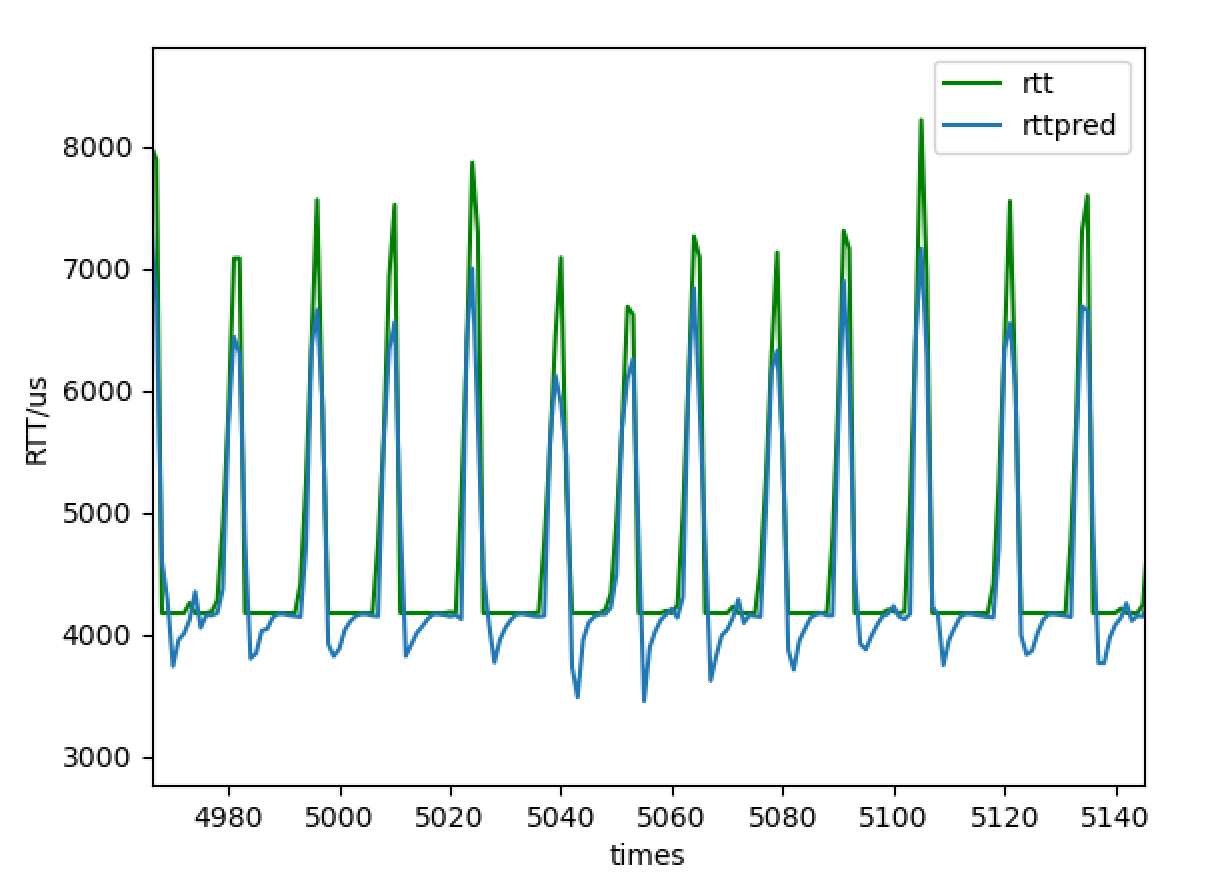
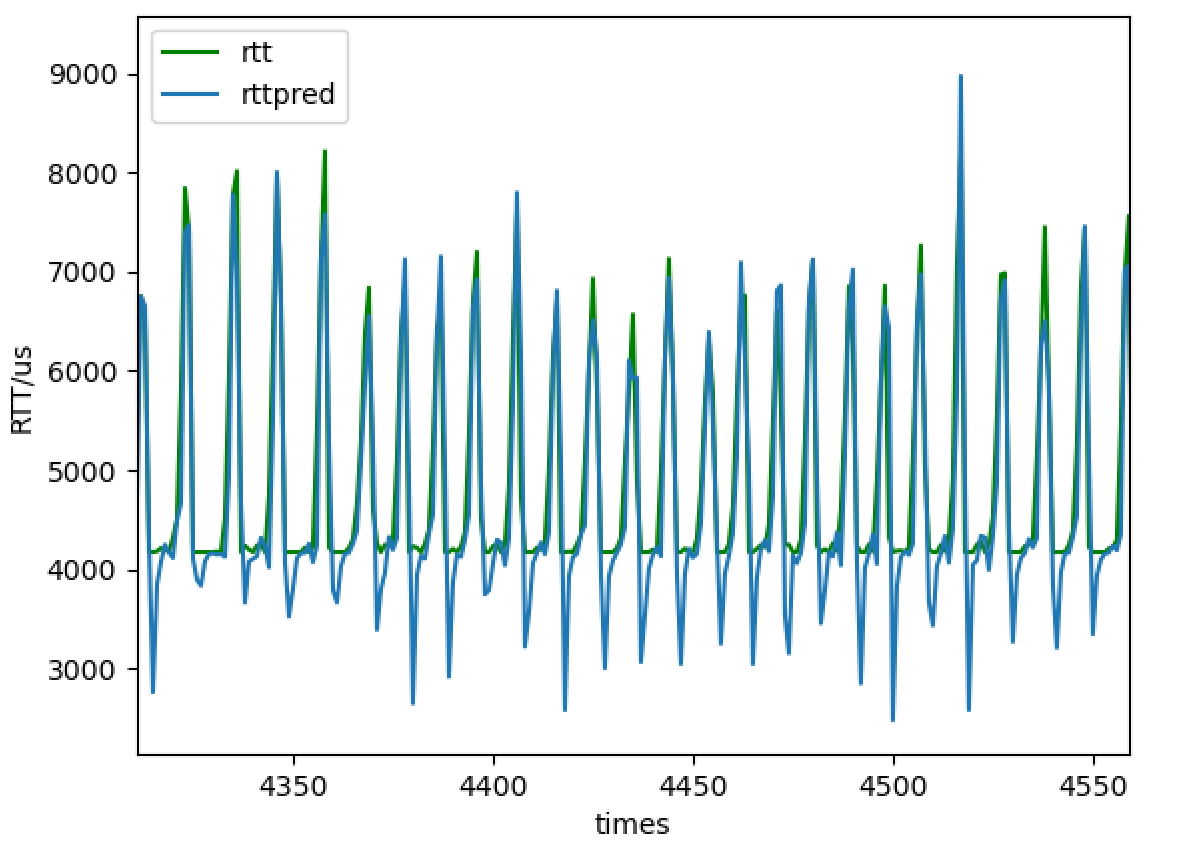
图X,DCTCP算法RTT采集数据



图X、TIMELY算法RTT采集数据

4.2.2RTT预测模块效果展示

通过论文第X部分的数据集划分及训练方法，对上述数据集进行训练，根据19轮训练后的网络模型，使用PYTORCH模型中的torch.save函数将训练后网络保存为pt文件。基于公式X在仿真平台使用C++复现该神经网络算法，并使用torch.load读取训练后的参数人工加入到C++代码中，运行仿真平台测试效果，这里选取其中两个节点RTT预测结果，由于数据量过大，这里仅展示部分时刻的RTT预测结果，如下图所示：



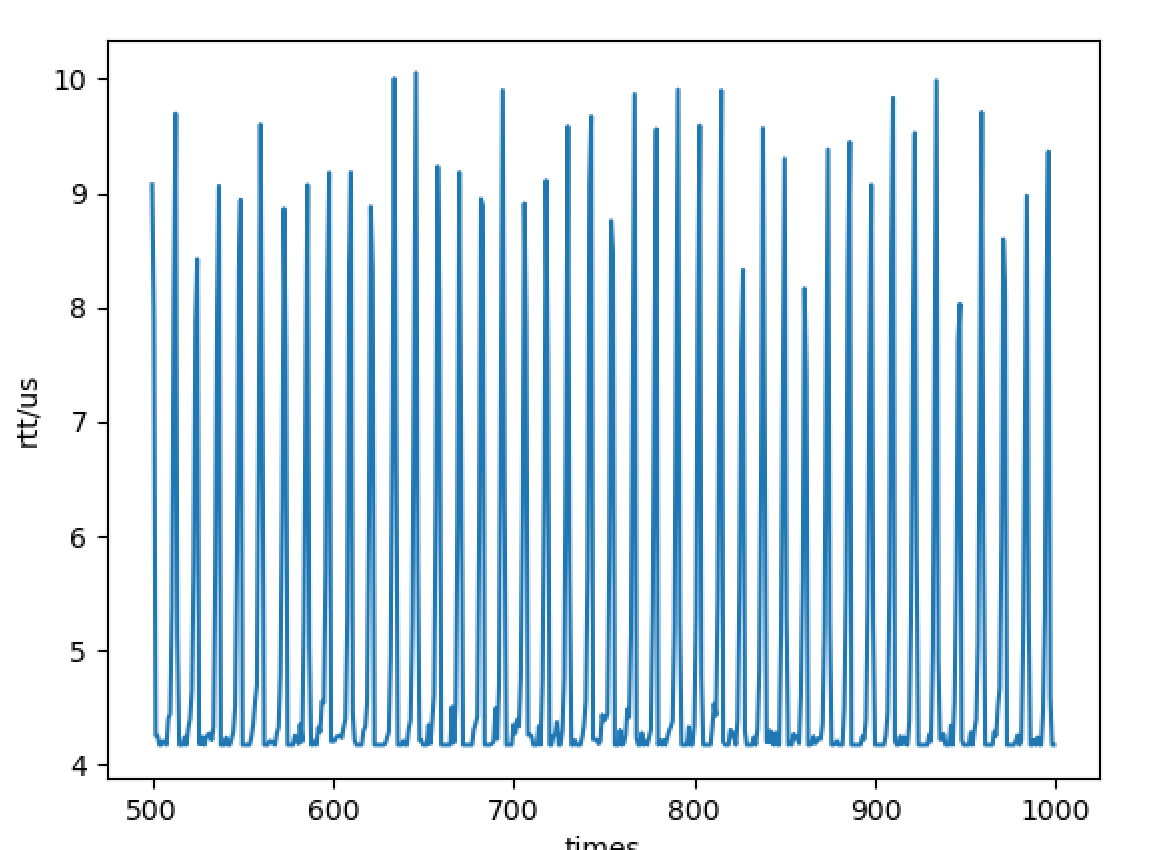
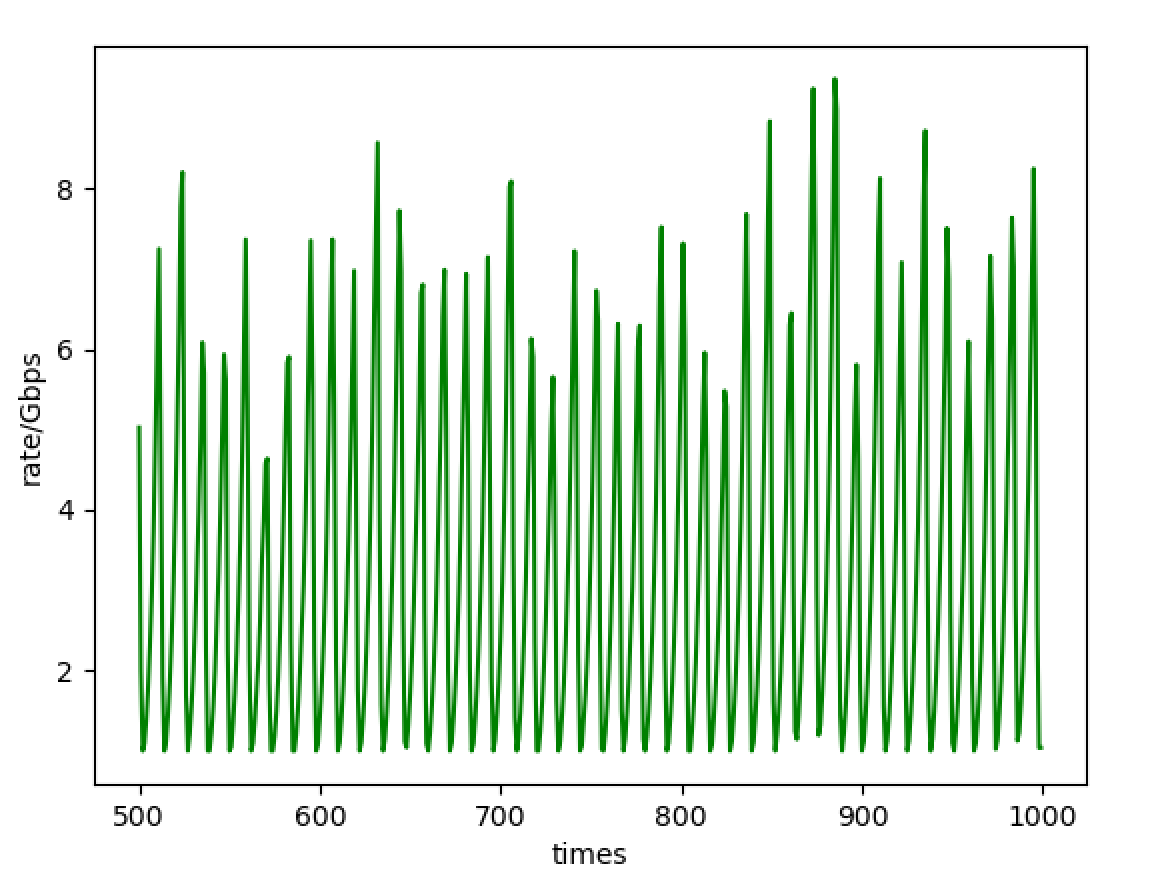
图x,仿真平台部分RTT预测图

通过上图可以观察出，RTT时间序列预测模块整体预测效果很好，预测曲线几乎与真实曲线重合，预测效果分析：

* 能够预测出RTT的迅速增加和降低，即能够预测出网络拥塞程度的变化。
* 在RTT峰值和谷值出预测效果值与真实值重合度低，并非预测不准确，是由于速率控制模块控制拥塞导致的，当RTT迅速增加时或降低时，预测模块能够准确判断该趋势并给出预测结果，但控制模块为了抑制RTT过高增加（拥塞加剧）和RTT过低（队列空，链路带宽利用率低，资源浪费）的情况会调整速率，进而影响后几个时刻的RTT。

4.2.3PID速率控制模块效果展示

为了尽量保持发送链路交换机端口队列接近空队列，设置,使用上一章已经调整好的参数，观察该PID控制算法下节点10第500-第1000个采集数据包的RTT时延和RATE速率变化：

图X

观察上图发现，节点10第500-第1000个采集数据包的RTT时延和RATE速率变化有着相同的规律，即不断上下震动，这是基于RTT这一类拥塞控制算法共有的特点（基于RTT控制拥塞，当RTT增加，发送速率降低，随着速率降低，RTT随着降低，这时算法观测到RTT降低又重新增加发送速率，而RTT又随之增加，如此往复）；左图最低值大概在4us左右，这是仿真拓扑中固定的链路传播延时，有图最小值为1Gbps,这是仿真平台控制的最小发送速率。统计所有节点采集数据，其中平均RTT为5.0272微妙，平均发送速率16.0872Gbps, 99%RTT为10.105微妙，最大RTT为13.009微妙。

4.3算法拥塞控制效果展示

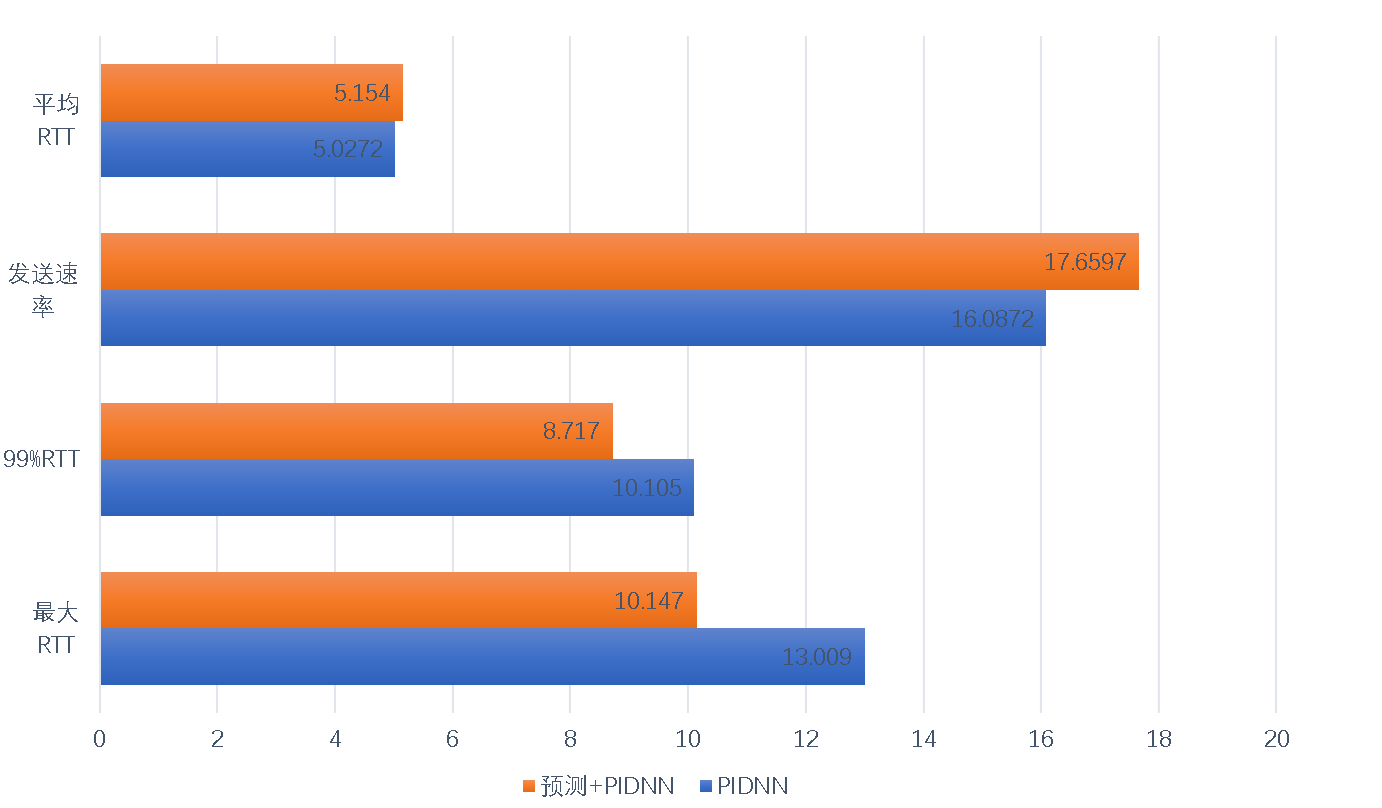
通过上一节可知，各模块在仿真平台上分别部署实现各自功能；该节将各个模块进行合并，同时部署到仿真平台中进行测试，观察算法整体效果以及各模块之间配合情况。

在仿真平台中部署整个基于神经网络的数据中心拥塞控制算法，每当选择不同的，具体控制效果就不同，观察其对所有发送节点采集报文平均时延及其发送速率的影响,其中平均RTT单位为微妙us,平均发送速率单位为Gbps：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| (us) | 5 | 6 | 7 | 9 |
| 平均RTT | 5.1540 | 6.8418 | 9.4241 |  |
| 平均发送速率 | 17.6597 | 25.3849 | 23.9605 |  |
| 99%RTT | 8.717 | 15.785 | 33.686 |  |
| 最大RTT | 10.147 | 18.46 | 35.65 |  |

随着增加，相关指标RTT分析

观察上标中那一列，与之前4.2.3 单一PID速率控制模块一样，前者加入了RTT预测功能，将公式X中rtt（t+1）变为rttperd,效果展示结果比较：



图X，PID控制算法加入预测模块对比

其除了平均RTT指标较后者较差，其余三个指标都有一定成都的提升，尤其是最大RTT和99%RTT,这说明加入了RTT预测模块后对于拥塞变化反应更快，速率调整更及时；加入了RTT预测模块后，当预测RTT大于，速率便会提前下降，对于拥塞具有更快的反应,使得最终效果有了一定的提升。

在实际应用过程中，可以根据不同的需求调整，例如当发送低延时要求的控制报文时，可以设置以保持链路中交换机端口队列接近空队列；当要求数据尽可能快速发送，而对时延要求较低时，可以设置以提高数据发送速率，降低发送总时间.

-------------------------------------------------------------------------------

===========================================================

4.4.算法分析比较

4.4.1分析比较指标

算法分析数据基于以下内容：

* 统计的RTT时延和发送速率
* 基于仿真平台提供的FCT分析和跟踪工具。

4.4.2算法比较

4.4.2.1TIMELY

4.4.2.2HPCC

4.4.2.3DCTCP

4.4.3比较结果分析总结

4.4.3.1算法优势及缺点

4.4.3.2算法创新点

4.5本章小结