1 各位老师上午好，我是张晗，我的导师是尹霞老师，我的博士论文题目是**数据中心网络中的应用传输速率和延迟优化，**

2 我的工作主要从以下6个方面展开，首先是介绍研究背景和主要工作，然后介绍一下相关研究综述和分析，接着是流级别传输的理论模型及优化方法，随后介绍任务级传输的理论模型及优化方法，然后是传输调度系统设计、实现及其性能评估 ，最后介绍一下结论与展望

3 首先介绍一下第一部分研究背景和主要工作

4 **数据中心已经成为支撑互联网在线服务的主要平台之一，著名企业比如Facebook，Youtube,移动，电信等都有自己的数据中心，数据中国心主要进行密集计算，存储以及向用户提供在线服务。**

**评价数据中心网络对应用服务质量的标准主要是延迟和带宽。**

**数据中心的应用对带宽和延迟的需求不同，有的应用必入PTP 同步，Memcached，Naiad是低延迟优先策略，**

**大数据备份，Hadoop是高带宽策略。**

**然而数据中心的资源有限，网络难以满足所有应用的需求，因此需要进行合理的资源调度。**

5 基于此，我的工作首先构建基于应用的数据中心传输调度框架，使应用提供更好的**用户体验**

然后对数据中心流和任务进行调度和传输优化，提高**网络资源利用率**。

**为了达到以上的目的**

本文进行流级别传输优化：根据应用重要程度给数据中心流设置期限，优化流传输延迟

然后 进行任务级别传输优化：给数据中心任务分配优先级，根据网络拥塞和应用优先级进行调度

6 这是本文的框架，论文总共分为10章第一章是引言部分，第二章是国内外研究现状，后面根据两条技术路线展开我们的工作，首先是流级别传输，然后是任务级别优化，对于流级别优化，第三张是基于ECN标记的流模型，随后假设流大小已知，进行第四章，介绍基于负载自适应原则的传输优化方案，和第5章基于流持续时间的传输优化方案。随后，介绍的是任务级别传输，进行第6章传输任务调度模型，随后进行第7章，基于流信息的任务级传输优化方案，和第8章，基于重要性的任务级传输调度方案。 第9章是调度与传输优化系统，最后一章是总结与展望

7 下面介绍一下相关研究综述和分析

8 这部分，对应论文的第2章，介绍的是国内外研究现状

9 数据中心使用传统传输方法面临一下的问题Map-reduce等高并发引起大量丢包TCP使得缓冲区很大，延迟高带宽公平分配，难以满足所有应需求网络资源利用率低，应用未获得应得的网络资源

10 根据传统传输方式面临的问题，数据中心传输方案可以分成以下4种方案。根据传输调度的技术，可以分成分布式和集中式的方法，

根据传输调度的粒度，可以分成流级别和任务级别的方案。

根据优化标准，可以分成显示期限调度和金星流完成时间优化。

根据网络模型可以分成考虑路径的方法以及非阻塞结构的方法。

11 数据中国中心主要的传输优化方案有DCTCP，D3，PDQ，D2TCP，pFabric，PASE，FASTPASS，CODA，Karuna等

12 从我们提出的研究路线，对已有的方案进行分类，在2011前的方案主要是公平分配带宽的方案。 后来分成了流级别优化方案和任务级别优化方案。

流级别优化方案，主要有XXX 任务级别优化方案主要有XXX

13 此表格是对主要方案从调度技术，期限，应用完成时间，优化级别，考虑期限与性嫩关系，任务重要性等方面对方案进行的总结

14 总结一下上面的表格，可以得到以下的结论：

无论当前的流级别还是任务级别的优化，都没有考虑拥塞程度和应用性能的关系，导致方案在重度拥塞时失效 当前流级别传输方案，都侧重单独解决期限问题或者流完成时间优化，而并没有同时对二者优化 任务级别的传输调度，大多优化任务完成时间，并没有结合任务重要性进行调度，导致网络资源并没有合理的分配 **因此 无论是流级别还是任务级别，可优化空间大**

15 基于以上各方案的问题，本文提出的研究目标是构建基于应用的数据中心传输调度框架，使应用提供更好的**用户体验**对数据中心流和任务进行调度和传输优化，提高**网络资源利用率**。**然后本文的研究路线是**进行**流级别传输优化**：根据应用重要程度给数据中心流设置期限，优化流传输延迟，进行**任务级别传输优化**：给数据中心任务分配优先级，根据网络拥塞和应用优先级进行调度

16 总结一下本章的内容，首先论述了传统传输方式面临的问题，然后讨论了传输优化方案面临的问题 那么针对一系列的问题，下面介绍本文的解决方案

17 首先介绍一下流级别的传输方法

18 流级别传输方法主要是3，4，5章 第3章介绍的是基于ECN标记的流模型

19 下面是基于ECN标记的流模型的过程

20 基于ECN标记的流传输过程，交换机设定标记数据包的条件，满足条件数据包被标记 如果数据包被标记，那么ACK中有ECN，在发送端上统计ECN的比例，进而估算拥塞程度 根据拥塞程度根据拥塞程度进行窗口计算如下所示，和是关于拥塞程度的函数 下面的三个式子是基于ECN标记的流模型

相关工作发表在Computer Networks上

21 下面是以DCTCP为例进行模型的实例化验证，下面的第一个式子是表示滑动窗口，第二个式子表示的是队列长度的描述，第三个式子是拥塞程度的式子描述，最后一个式子是队列的描述。图展示的是基于ECN标记的流模型和ns-2仿真模型下对拥塞程度，队列长度，窗口大小的描述的对比。 可以发现，基于ECN标记的流模型和仿真结果相符

22 总结一下本章介绍的内容：本章主要介绍了4部分内容，首先介绍的是基于ECN标记的流模型，然后介绍了基于ECN标记进行带宽计算拥塞窗口的方法，第四部分介绍了基于ECN标记的流模型的描述形式，最后用DCTCP为例对模型进行验证。

23 下面介绍一下基于负载自适应原则的传输优化方案

24 数据中心存在大量数据，传输存在期限。

数据中心的传输模型Partition-aggregate如图所示，此模型是树状结构，在此模型中，根节点会发送请求，然后叶节点会回复数据。

因为有期限限制，因此如果传输错失期限，那么应用的性能会受到影响，从而影响用户体验。 针对此问题，业界提出的解决方案是D2TCP，但是此方案存在的问题是**D2TCP在网络拥塞严重时失效。** 因此，本设计网络拥塞严重时依然有效的速率控制机制。 本章的相关工作发表在INFOCOM 15

25 为了验证我们的结论，在ns-2上，启动4条流，设置联路带宽是1Gbps，RTT是100ms，可以发现D2TCP拥塞程度相同的情形下，期限因子不，4条流的速率相近，D2TCP和DCTCP的性能相似

26 下面分析出现上面现象的原因，D2TCP滑动窗口的计算如下所示，其中α: 和DCTCP中相同，是拥塞控制因子(0 ≤ α ≤ 1)d: 期限因子

如图所示是D2TCP的gamma函数值。

分析gamma函数，定义，

pd的物理含义如图，是相同alpha下，不同d值的差异。

可以发现如图所示，是d1和d2不同时，二者的关系。右图展示得是二者的差的曲线。可以发现期限不同**，**D2TCP的惩罚函数随拥塞先增大后减小。**因此，拥塞程度严重时，滑动窗口变化相同，从而速率相近。**

27 基于D2TCP重载情形下失效的问题，本文提出PI和PD来设计区分函数，其中PI表示期限期限近的流要比期限远的流得到更多带宽，PD表示的是拥塞程度和带宽区分的关系。

LPD策略如图所示。两条曲线表示不同d下pd的值

可以发现随着拥塞程度增加，pd增大

因此期限不同**，**LPD的惩罚函数随拥塞先增大而增大

28 LPD有以下的特性流达到的稳定速率与之对应的期限成反比 ，证明主要是通过LPD不同期限的滑动窗口的关系进行的。

29 根据之前基于ECN标记的流模型，建立LPD的流模型，如下所示。图显示的是LPD流模型和ns-2仿真效果基本类似。

根据LPD流模型得到LPD不动点，根据不动点分析tmax的取值，可以得到tmax可以根据要调整的最长期限的流进行设置

30 下面介绍一下LPD的关键参数分析，LPD中参数K可以根据带宽C和R进行设置。K的值约设定为联路利用率为100%

拓展LPD，可以得到LPD－e。此外，剩余流大小，剩余期限等因子也可以考虑

31 此实验验证LPD解决研究动机提出的问题，可以看到，使用LPD带宽得到区分，拥塞窗口差很大，因子有差异，拥塞程度基本相同

因此，网络拥塞严重时，LPD依然可以根据期限区分数据流带宽

32 本文同时对LPD进行仿真实验验证

仿真实验的拓扑，使用simple和复杂拓扑，如图所示，变化变换fan-in程度，代表不同拥塞程度，同时变换spine-leaf超额认购系数x，40Gbps/x进行测试

数据中心流量使用data mining和web search两种流量对于期限的测试，使用紧急期限(20ms)，中等期限(30ms)，松弛期限(40ms)评价指标使用错失期限数据流比例，流完成时间

33 仿真实验的主要结果如图：首先是错失期限对比，然后是流完成时间对比。

可以看到LPD性能比D2TCP,L2DCT提高50%右边是tmax设置，

34 同时在真实环境下的实验，此实验是基于linux内核3.2.61 拓扑如右所示，以下是主要实验结果。 可以发现LPD可以根据期限进行带宽调整，同时保持交换机队列短

35 本章主要，分析了数据中心流级别优化方案D2TCP在网络负载严重时失效原因提出数据中心“越拥塞，越区分”的拥塞控制机制设计原则

根据“越拥塞，越区分”设计并且实现LPD建立LPD的流模型和简单模型，并对LPD参数进行分析在ns-2和真实环境下评估LPD性能

36 下面介绍一下论文的第5章，基于流持续时间的传输优化方案

37 数据中心应用的侧重不同Hadoop，spark等应用侧重流完成时间

OLDI，搜索等应用关注截止期限 数据中心优化策略存在的问题

基于期限的策略(eg..D2TCP)，只优化期限 优化流完成时间策略(eg..L2DCT)，只优化FCT

38 为了验证我们的假设，在ns-2中启动10个发送端和1个接收端

10 个发送端，1个接收端，K＝25，2 个发送端不断发送背景流，4个发有期限的流，4个发短流 可以发现使用L2DCT流平均完成时间减少，D2TCP，有期限的流错失期限比例少。 因此 L2DCT和D2TCP难以同时优化期限和流完成时间

39 目标：设计合理的流量控制机制，使的网络资源得到更合理的分配，同时优化期限和流完成时间

方法改进TCP滑动窗口机制引入流持续时间到窗口变化函数中，同时考虑网络拥塞程度和带宽的关系

优势，只改进滑动窗口机制，简单，易于部署。引入流持续时间在窗口的变化过程中，同时可以优化有期限的流和流完成时间

40 基于以上的问题，我们提出基于流持续时间的速率控制（Flow Duration Rate Control，简称FDRC）首先计算拥塞控制因子， 然后计算滑动窗口

41 因为FDRC也是基于ECN标记的策略，根据基于ECN标记的流模型方法，实例化得到FDRC的流模型 根据fluid流模型得到不动点，如图 根据不动点可以推算出FDRC重要参数DMIN和DMAX的取值，可以看到DMIN和DMAX需要进行合理设置

42 下面进行决绝实验动机提出的问题的验证。

相同实验参数下可以看到期限错失和D2TCP基本相同，FCT优化能力和L2DCT基本相同

43 使用数据中心真实流量Web search, 数据挖掘，期限设置使用紧急，一般，松散期限，可以发现FDRC的性能优于D2TCP，L2DCT。

44 真实环境下的FDRC的性能如图所示，可以看到FDRC可以同时优化有期限的流和流平均完成时间

45 本章做的工作如下：

发现并且分析了了数据中心流级别优化方案L2DCT和D2TCP只优化流完成时间或者期限，并不能同时优化两个目标

把流持续时间引入传输层速率控制机制中

设计并且实现FDRC

建立FDRC的流模型和简单模型，并对FDRC参数进行分析

在ns-2和Linux内核评估FDRC性能

46 以上介绍了流级别的传输优化方案，下面开始介绍任务级传输的理论模型及优化方法

47 首先介绍第6章，传输任务调度模型

48 数据中心的非阻塞模型如图所示，在此模型中，不考虑中间路径的拥塞

假设拥塞只发生在第一跳和最后一跳

51 在数据中心非阻塞模型下，我们定义流组，流组顾名思义，就是一组流表示流大小，没有从i到j的流，那么＝0

根据流组的定义，理想化权重流组完成时间优化问题 （简称，IWCCTM问题)

其中第一个式子是优化目标

第二个式子是

52 如果只允许非抢先调度,那么 IWCCTM 问题和开放商店中任务的权重完成时间之和的问题是等价的,问题的复杂度是 NP-hard

证明的基本思路是，把数据中心的流组拆分成2m个传输子任务，然后每个子任务和开放商店中的任务等价，因为开放并行商店中的最小化任务权重完成时间问题是NP-hard，所以对流组的调度也是NP-hard的

53 解决IWCCTM问题的离线近似算法如图所示，算法的思路是首先计算每个端口的负载，然后找出负载最重的端口，然后找出最小权重负载比的流组，随后进行下一轮迭代

54 下面证明离线算法的2－近似算法

首先得到IWCCTM的另一种表述方式，然后得到此方式的对偶函数，然后引入一个引理，此引理是任务处理时间的关系

55 最后，我们使用引理通过放缩对离线算法进行验证

56 下面对本章进行总结，介绍了数据中心非阻塞结构模型，提出流组传输的IWCCTM问题，证明IWCCTM的NP-hardness，提出IWCCTM的2－近似离线算法，并证明其近似度

57 后面我们介绍第7章基于流信息的任务级传输优化方案

58 著名的公司比如facebook等都采用纠删码存储系统进行文件存储。

(n,k) MDS纠删码用来对文件进行编码，文件被分成n个块，存储在不同的机器上，任何的k个模块都能恢复此文件

（1）数据块存储在不同的机器上，客户端收到用户的请求后选择适当的server［选择哪些数据源］

（2）所有的数据块到达客户端以后才能对数据进行重新组合，得到用户需要的文件［如何优化传输］

纠删码存储系统经常采取随机选源和流级别传输，但是，随机选源和流级别传输常常使分布式系统性能受阻。

59 为了验证之前的问题，假设在纠删码存储系统中有两个文件A和B，两个文件被编码在两个文件被编码在3台机器上，如图所示，文件A被编码在S1，S2，S3，文件B被编码在S2，S3，S4上，文件A的选取有（S1，S2），（S1，S3），（S2，S3），文件B的选取有（S2，S3），（S3，S4），（S2，S4）

如果请求文件A和文件B都在t=0到达，A选择(S1,S2)，B选择(S3,S4)，AFAT =1，A选择(S2,S3)，B选择(S3,S4)，AFAT＝2

如果请求A在t=0到达，请求B在t=0.1，A选择(S1,S2)，B选择(S2,S3)，使用TCP作为传输协议AFAT =1.95，A选择(S1,S2)，B选择(S2,S3)， 如果A的优先级更高AFAT =1.5

60 基于此，本文提出**文件访问最小化问题（IFATM）问题。**

**首先优化目标是**

**然后限制条件，其中第一个是非抢占是调度情形下的入端口访问限制**

**第二个式子是出端口上的限制**

假设源选择固定，得到理想化文件访问最小化问题（IWCCTM）

**此理想情况下，源选择是全选择，即所有文件存储在所有机器上，文件的数据块需要从文件所有节点中取出，而不需要进行源选择**

**因为**简单的IWCCTM是NP-hard的，因此IFATM是NP-hard的

61 SIFATM问题的思路是先找出负载最重的端口，然后找出最小权重负载比的流组，随后进行下一轮迭代

事实上，只解决理想化的IWCCT问题，在实际文件系统中难以部署

62 针对现实中的问题，本文提出在线调度策略，在线调度首先进行选源，

然后计算每个服务器的负载，随后根据服务器的负载进行优先级调度

对于源的选择，是选取负载最小的前K个节点进行

63 下面是纠删码调度系统的设计，本系统采取集中式处理机制，

中央控制节点计算带宽，并且把计算的结果发送给每个节点

算法6 是客户端管理器的过程，带宽计算过程如算法7所示

64 评价指标：平均文件访问时间(Average File Access Time, 简称 AFAT) ，使用相对于TCP提高幅度进行衡量

数据，60个机架，720台服务器，30天的请求，AT&T纠删码存储系统

开源，https://github.com/zhanghan1990/D-Target

65 主要的结果如图所示，可以发现D-Target性能比任务级别传输方案高1倍

在线策略有20%的性能损失

66 介绍了纠删码存储系统使用随机源选择和TCP存在的性能问题

提出了文件访问最小化问题（IFATM）

证明IFATM的NP-hardness

提出解决纠删码存储系统优化问题的在线方法和离线方法

设计并且实现D-Target系统，并使用AT&T就删码存储系统数据对之进行性能评估

67 下面介绍一下第8章基于重要性的任务级传输调度方案

68 数据中心是应用共享资源类型的，不同应用对网络资源的需求不同

如图所示web,vRouter等应用共享数据中心资源

当前的大多数策略把应用同等看待，事实上，同等看待应用是不够的，使的网络资源得不到合理分配

69 为了验证以上数据中心存在问题，我们测量，AT&T 数据中心中，应用和他们对应的紧急程度 (60 机架, 720 机器,1个月的流量数据)，如表所示，红色的表示紧急，黄色是重要，绿色是正常，蓝色是不重要，黑色是不紧急

70 然后对以上的测量数据使用TCP和Varys进行传输，然后测量每个应用的传输任务完成时间，

结果如图所示，可以发现，从调度结果看，使用Varys，event，vRouter等紧急应用，Varys的调度效果反而更差

71 下面分析出现这个问题的原因，如图所示，假设有两条流组，第一个任务有3条流，第二个流组2条流，两条流组的重要性不同，如果使用TCP传输，结果如图，s你用Varys，结果如图，可以发现，

Vayrs只是根据网络特性调度，而没有考虑应用性

72 针对此，我们解决此问题的方法，刚才提到的IWCCTM测量可以解决此

简单离线算法如图所示，此算法和IWCCTM相比数据中心存在负载均衡机制

忽略端口能力的差异性，优先调度

针对此思想，提出流大小未知的在线调度策略

73 在线调度策略实现了最短流组优先策略

76 测量AT&T数据中心流量，发现并且分析Varys存在的问题

提出使用权重代表任务的紧急程度，并且划分AT&T传输任务5个优先级

提出调度权重任务完成时间的简单离线策略和在线策略

使用Facebook流量和AT&T流量评估在线策略

77