数据中心网络拥塞控制方案的分析与优化研究

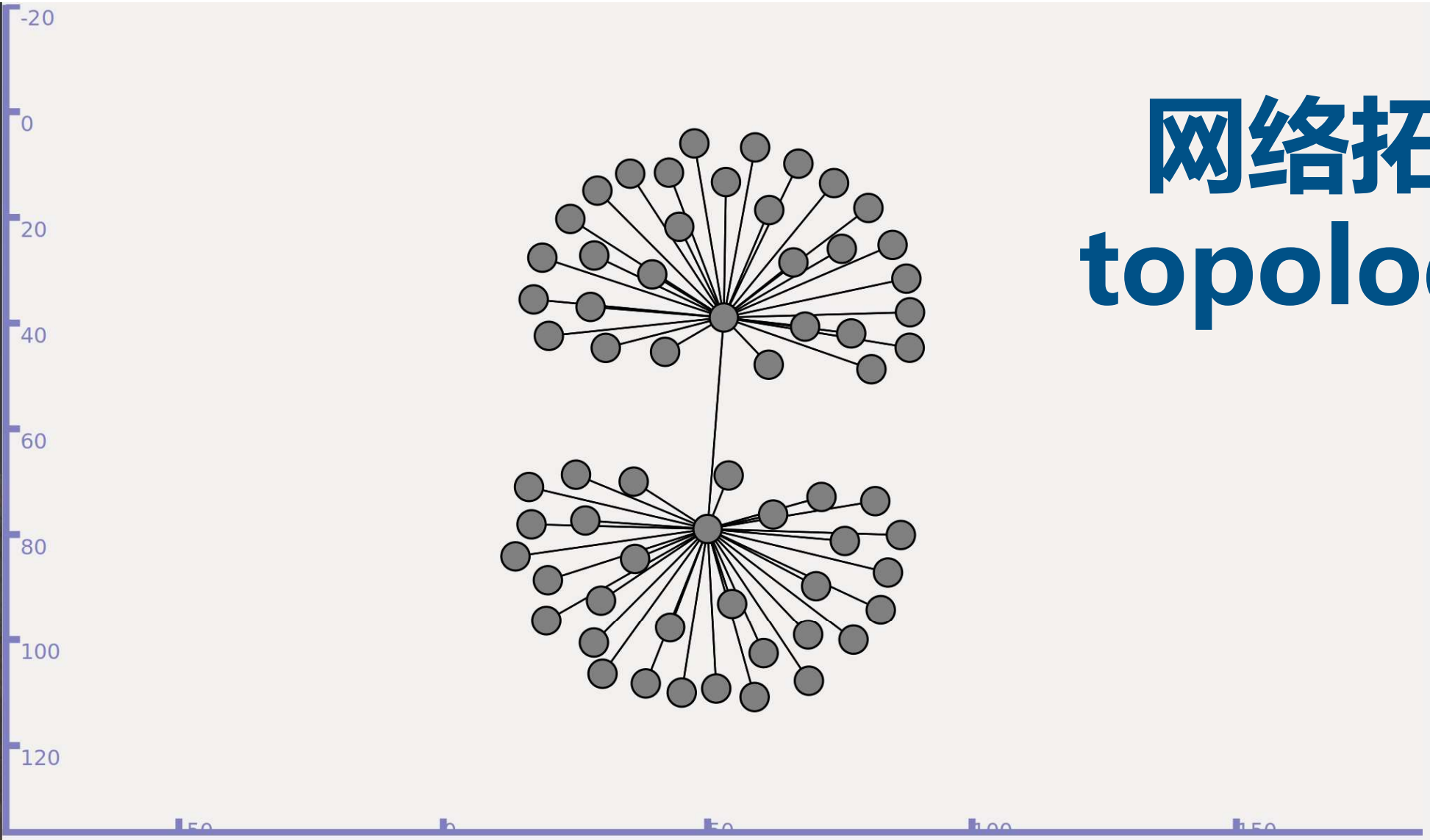
指导教师：李卓　　负责人：武嘉闻

成员：张鑫沂　谢紫昂　张宇彬　张艺彤

1 项目实施进展情况


Activities
Files
四 21:49
Home
ns-allinone-3.33
ns-3.33
Recent
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
Trash
Ubuntu …
Other Locations
bindings
build
contrib
doc
examples
__pycache__
scratch
src
testpy-output
utils
waf-tools
AUTHORS
CHANGES.html
CONTRIBUTING.md
LICENSE
Makefile
README.md
RELEASE_NOTES
test.py
testpy.supp
utils.py
VERSION
waf
waf.bat
wscript
wutils.py
在Linux中
搭建实验环境


bruce@bruce: ~/ns-allinone-3.33/ns-3.33
File Edit View Search Terminal Help
bruce@bruce:~/ns-allinone-3.33/ns-3.33$ sudo ./waf
[sudo] password for bruce:
Waf: Entering directory `/home/bruce/ns-allinone-3.33/ns-3.33/build'
Waf: Leaving directory `/home/bruce/ns-allinone-3.33/ns-3.33/build'
Build commands will be stored in build/compile_commands.json
'build' finished successfully (0.618s)
Modules built:
antenna aodv applications
bridge buildings config-store
core csma csma-layout
dsdv dsr energy
fd-net-device flow-monitor internet
internet-apps lr-wpan lte
mesh mobility netanim
network nix-vector-routing olsr
point-to-point point-to-point-layout propagation
sixlowpan spectrum stats
tap-bridge test (no Python) topology-read
traffic-control uan virtual-net-device
visualizer wave wifi
wimax
Modules not built (see ns-3 tutorial for explanation):
brite click dpdk-net-device
mpi openflow
bruce@bruce:~/ns-allinone-3.33/ns-3.33$
编译 ns-3
网络模拟器



网络拓扑
topology

算法与协议

Data Center TCP (DCTCP)

Mohammad Alizadeh[‡†], Albert Greenberg[†], David A. Maltz[†], Jitendra Padhye[†],
Parveen Patel[†], Balaji Prabhakar[‡], Sudipta Sengupta[†], Murari Sridharan[†]

[†]Microsoft Research [‡]Stanford University
{albert, dmaltz, padhye, parveenp, sudipta, muraris}@microsoft.com
{alizade, balaji}@stanford.edu

ABSTRACT

Cloud data centers host diverse applications, mixing workloads that require small predictable latency with others requiring large sustained throughput. In this environment, today's state-of-the-art TCP protocol falls short. We present measurements of a 6000 server production cluster and reveal impairments that lead to high application latencies, rooted in TCP's demands on the limited buffer space available in data center switches. For example, bandwidth hungry "background" flows build up queues at the switches, and thus impact the performance of latency sensitive "foreground" traffic.

To address these problems, we propose DCTCP, a TCP-like protocol for data center networks. DCTCP leverages Explicit Congestion Notification (ECN) in the network to provide multi-bit feedback to the end hosts. We evaluate DCTCP at 1 and 10Gbps speeds using commodity, shallow buffered switches. We find DCTCP delivers the same or better throughput than TCP, while using 90% less buffer space. Unlike TCP, DCTCP also provides high burst tolerance and low latency for short flows. In handling workloads derived from operational measurements, we found DCTCP enables the applications to handle 10X the current background traffic, without impacting foreground traffic. Further, a 10X increase in foreground traffic does not cause any timeouts, thus largely eliminating incast problems.

Categories and Subject Descriptors: C.2.2 [Computer-Communication Networks]: Network Protocols
General Terms: Measurement, Performance
Keywords: Data center network, ECN, TCP

1. INTRODUCTION

In recent years, data centers have transformed computing, with large scale consolidation of enterprise IT into data center hubs, and with the emergence of cloud computing service providers like Amazon, Microsoft and Google. A consistent theme in data center design has been to build highly available, highly performant computing and storage infrastructure using low cost, commodity components [16]. A corresponding trend has also emerged in data center networks. In particular, low-cost switches are common at the top of the rack, providing up to 48 ports at 1Gbps, at a price point under $2000 — roughly the price of one data center server. Several recent research proposals envision creating economical, easy-to-manage data centers using novel architectures built atop these commodity switches [2, 12, 15].

Is this vision realistic? The answer depends in large part on how well the commodity switches handle the traffic of real data center applications. In this paper, we focus on soft real-time applications, supporting web search, retail, advertising, and recommendation systems that have driven much data center construction. These applications generate a diverse mix of short and long flows, and require three things from the data center network: low latency for short flows, high burst tolerance, and high utilization for long flows.

The first two requirements stem from the *Partition/Aggregate* (described in §2.1) workflow pattern that many of these applications use. The near real-time deadlines for end results translate into latency targets for the individual tasks in the workflow. These targets vary from ~10ms to ~100ms, and tasks not completed before their deadline are cancelled, affecting the final result. Thus, *application requirements for low latency directly impact the quality of the result returned and thus revenue.* Reducing network latency allows application developers to invest more cycles in the algorithms that improve relevance and end user experience.

The third requirement, high utilization for large flows, stems from the need to continuously update internal data structures of these applications, as the freshness of the data also affects the quality of the results. Thus, high throughput for these long flows is as essential as low latency and burst tolerance.

In this paper, we make two major contributions. First, we measure and analyze production traffic (>150TB of compressed data), collected over the course of a month from ~6000 servers (§2), extracting application patterns and needs (in particular, low latency needs), from data centers whose network is comprised of commodity switches. Impairments that hurt performance are identified, and linked to properties of the traffic and the switches.

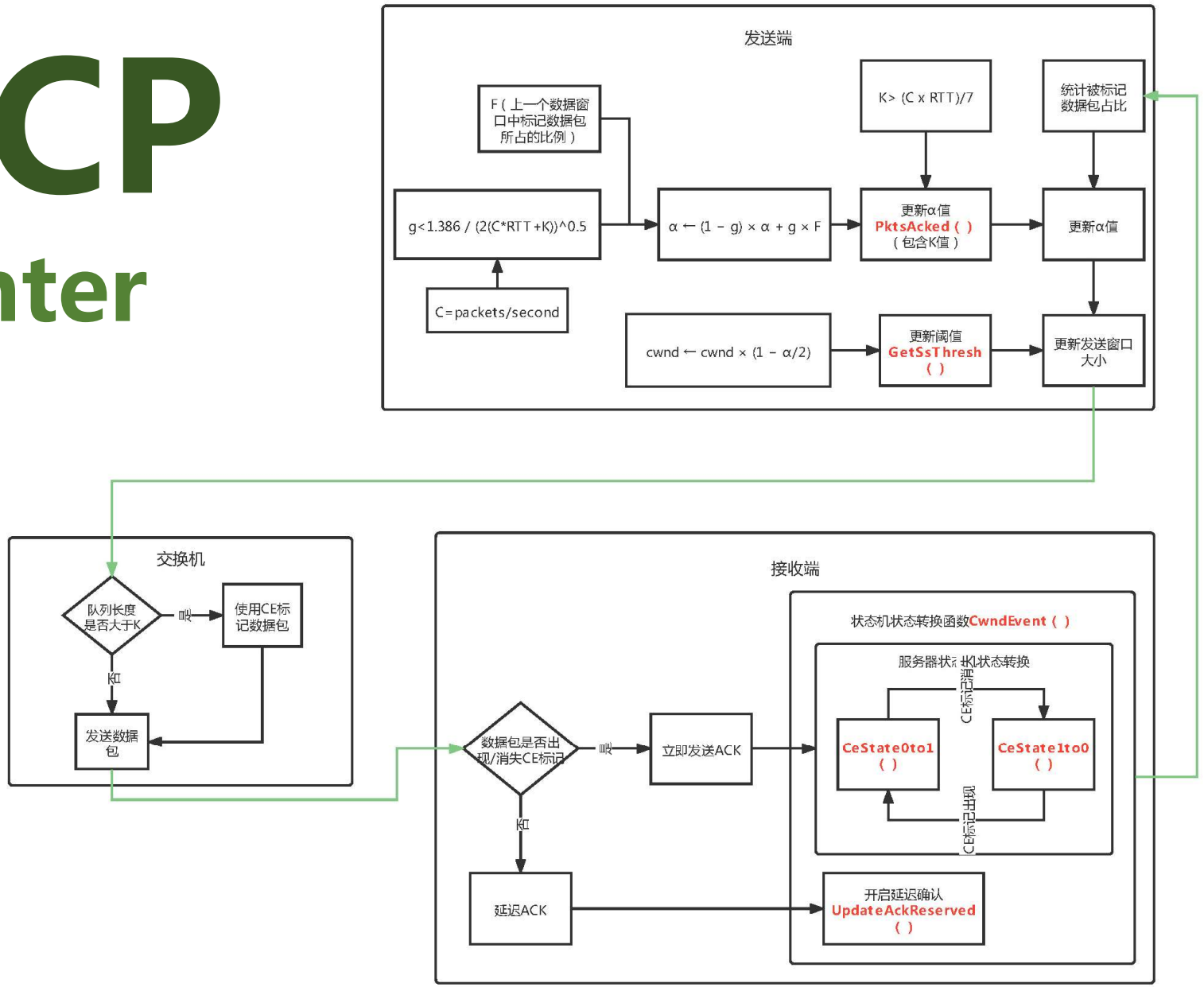
Second, we propose Data Center TCP (DCTCP), which addresses these impairments to meet the needs of applications (§3). DCTCP uses Explicit Congestion Notification (ECN), a feature already available in modern commodity switches. We evaluate DCTCP at 1 and 10Gbps speeds on ECN-capable commodity switches (§4). We find DCTCP successfully supports 10X increases in application foreground and background traffic in our benchmark studies.

The measurements reveal that 99.91% of traffic in our data center is TCP traffic. The traffic consists of query traffic (2KB to 20KB in size), delay sensitive short messages (100KB to 1MB), and throughput sensitive long flows (1MB to 100MB). The query traffic experiences the incast impairment, discussed in [32, 13] in the context of storage networks. However, the data also reveal new impairments unrelated to incast. Query and delay-sensitive short messages experience long latencies due to long flows consuming

SIGCOMM'10, August 30–September 3, 2010, New Delhi, India.

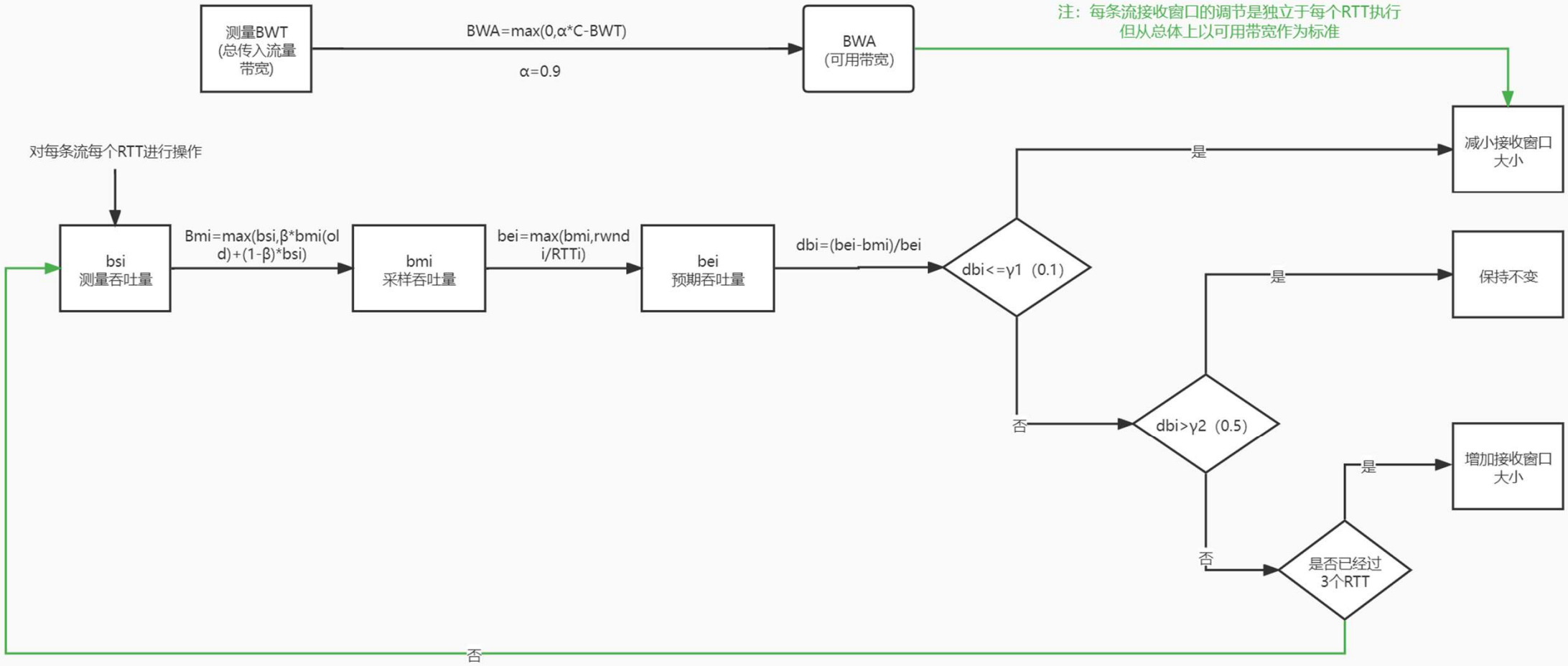
DCTCP

Data Center TCP



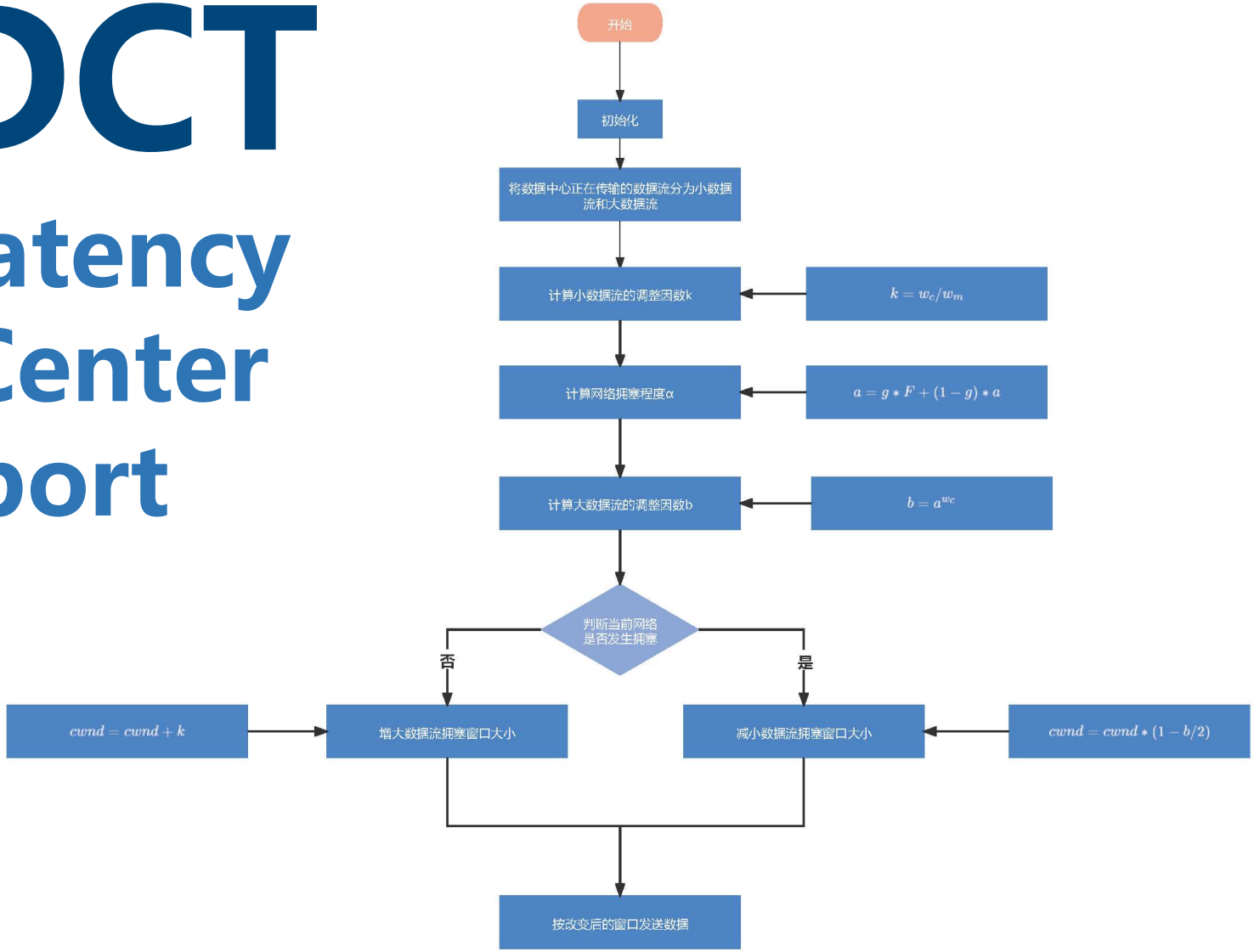
ICTCP

Incast Congestion Control for TCP



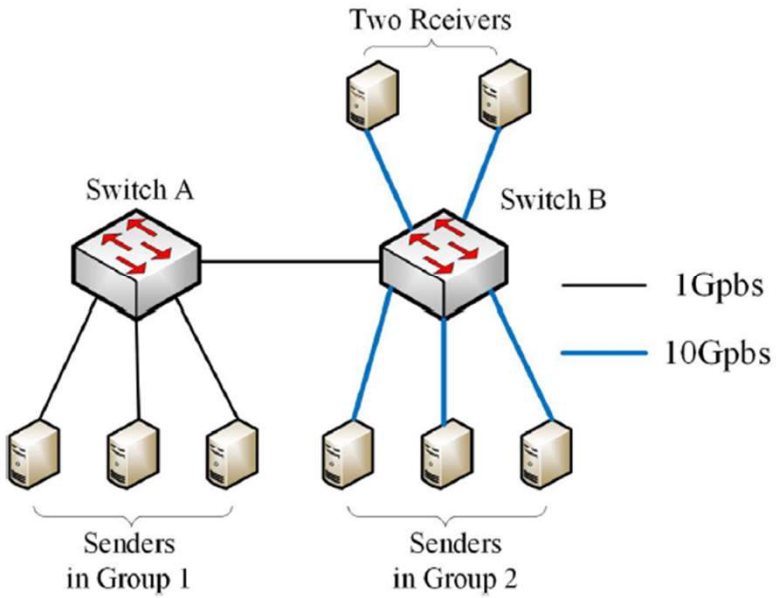
L2DCT

Low Latency Data Center Transport

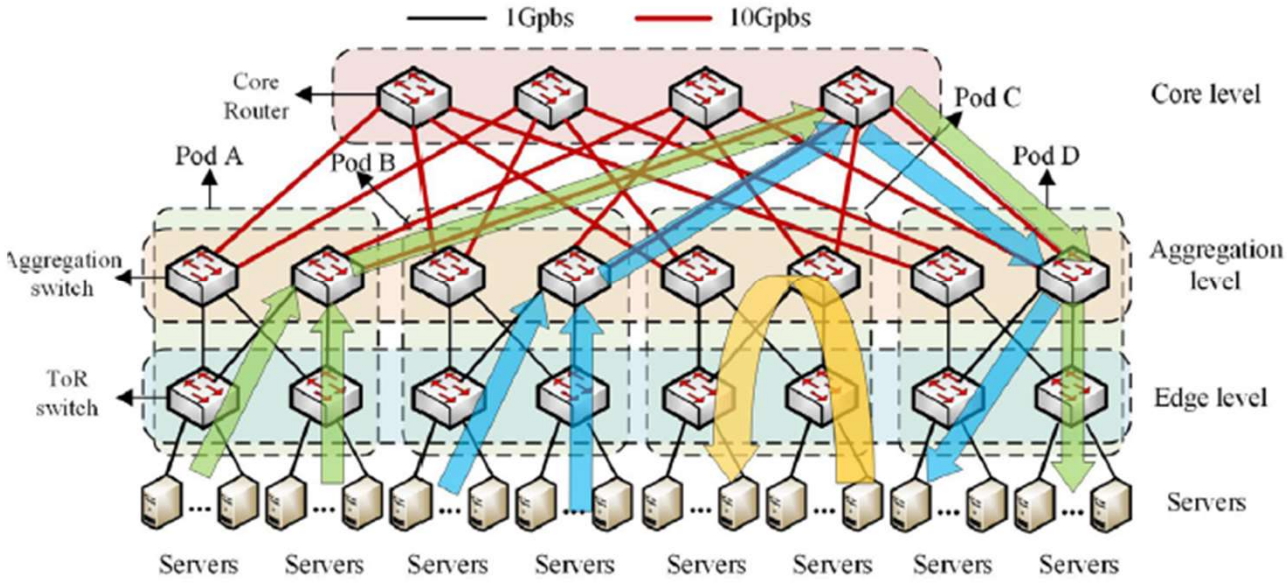


2 已取得的成果

网络拓扑模型

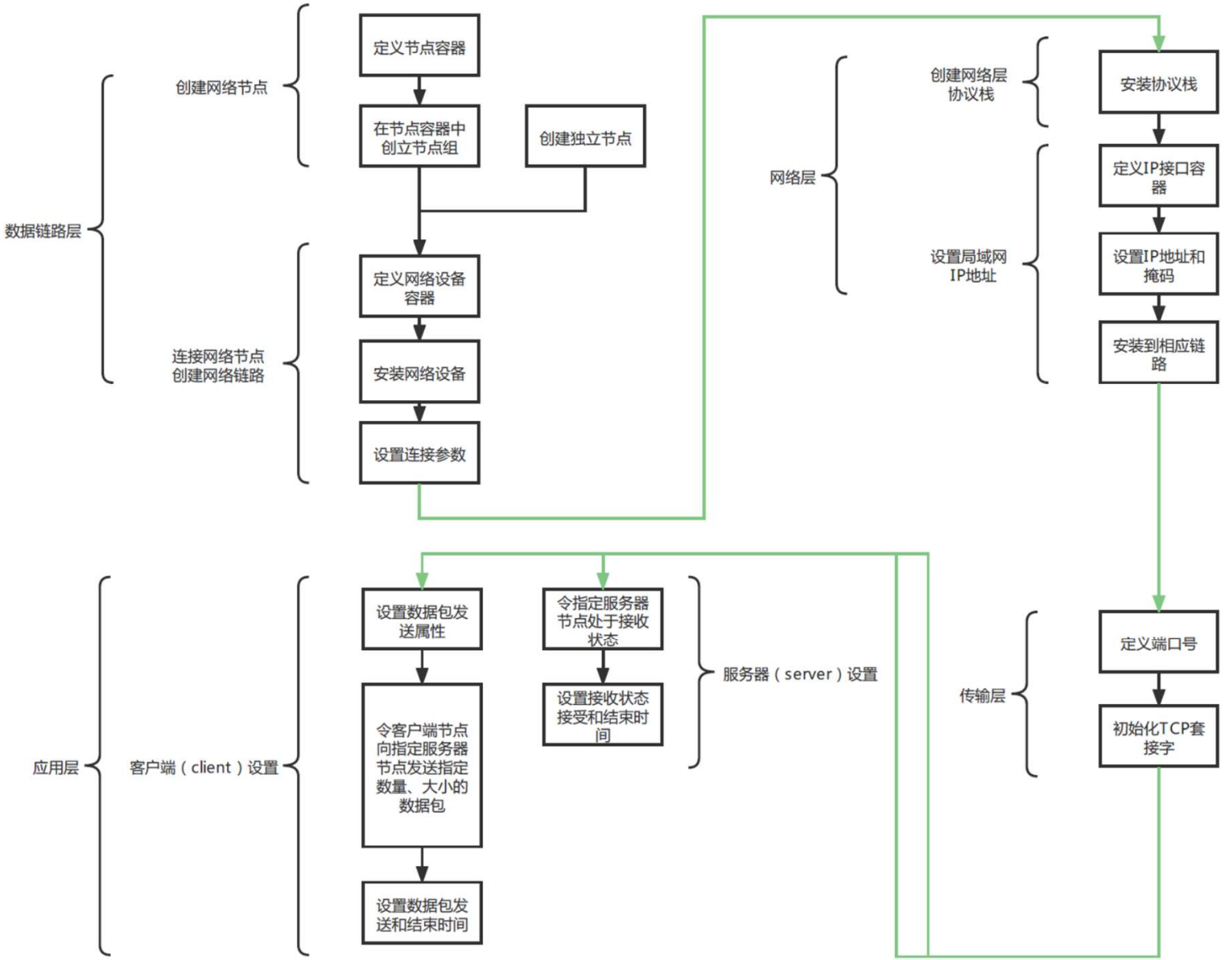


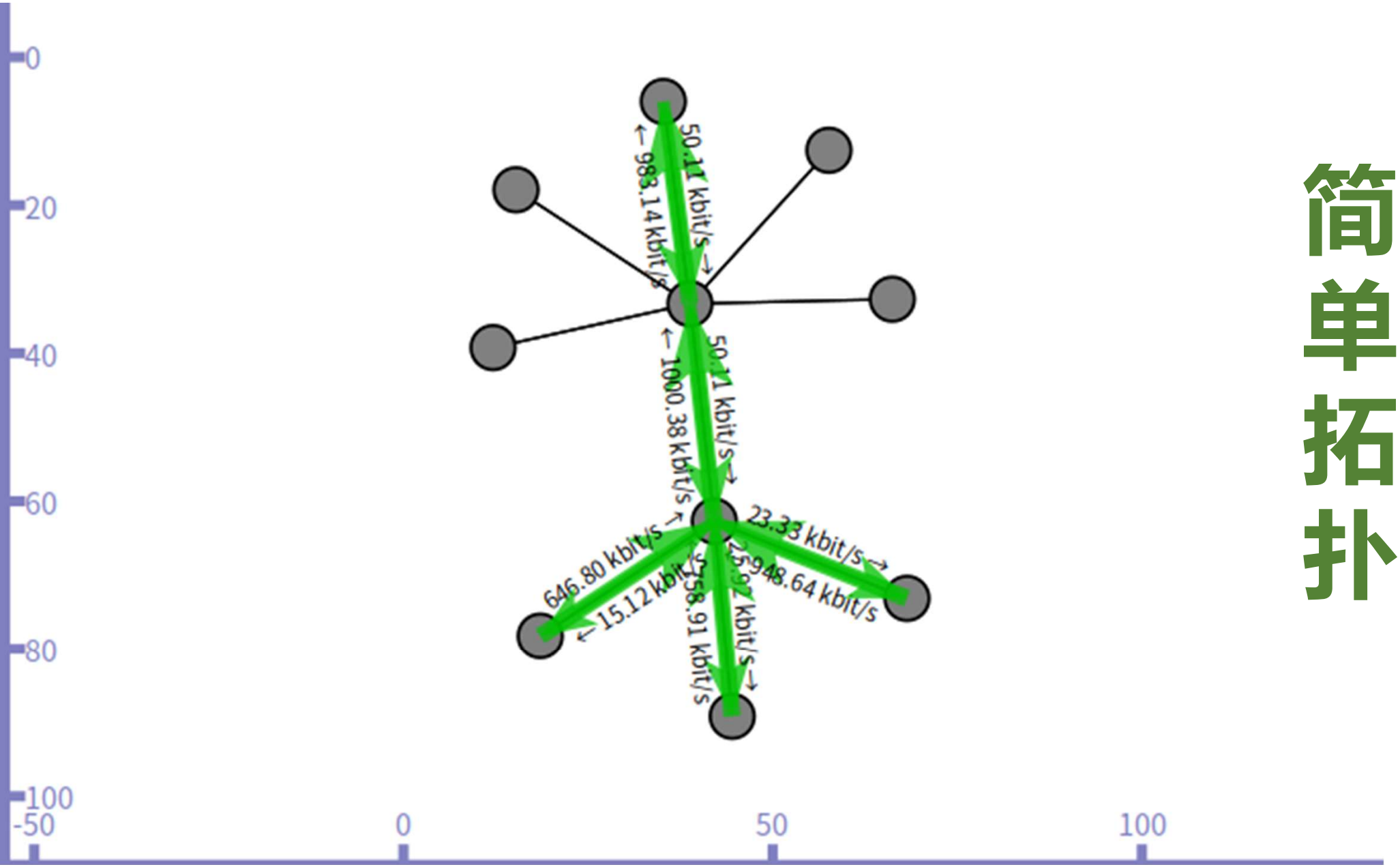
简单拓扑
小网络模型



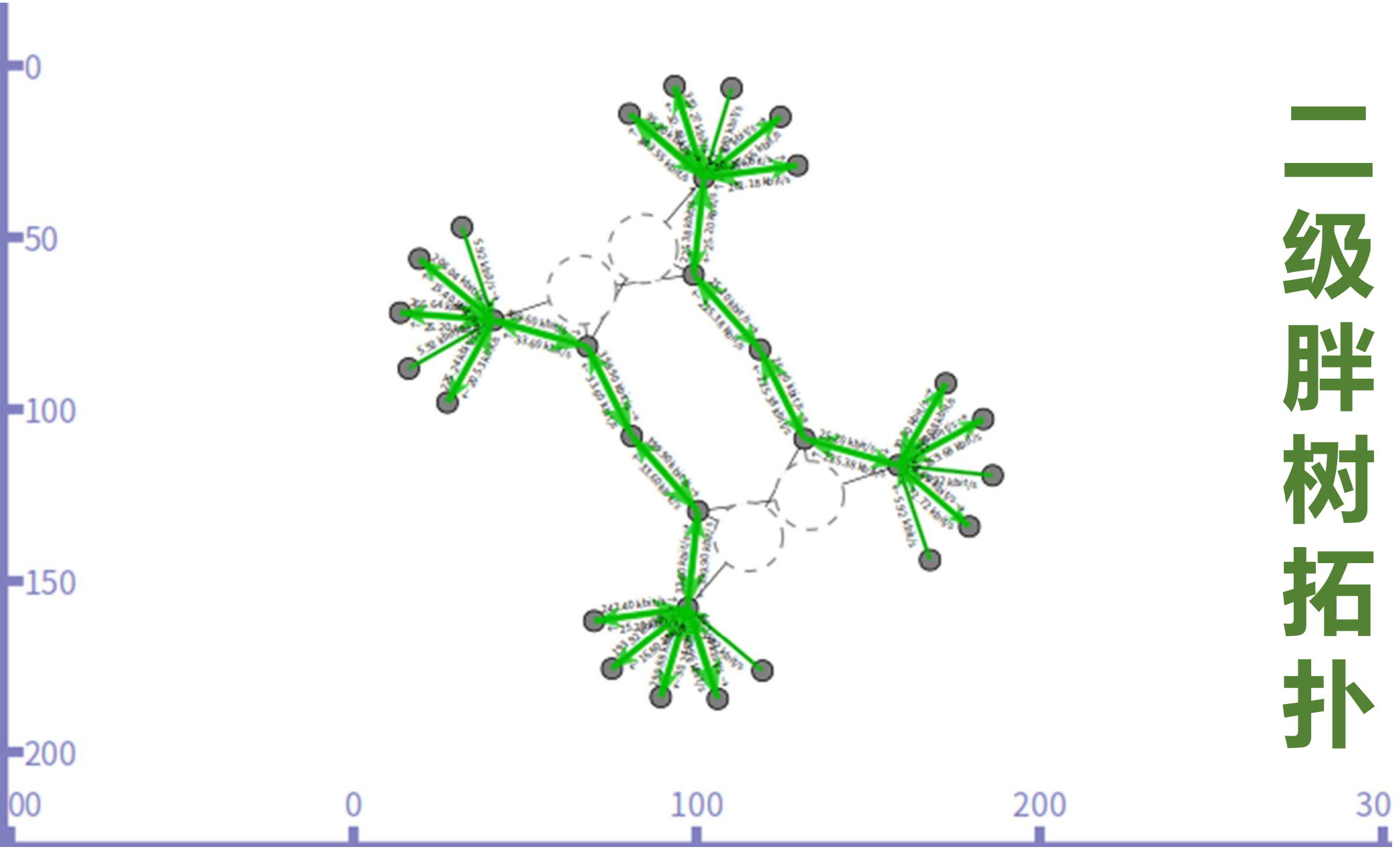
胖树拓扑
大网络模型

拓扑流程

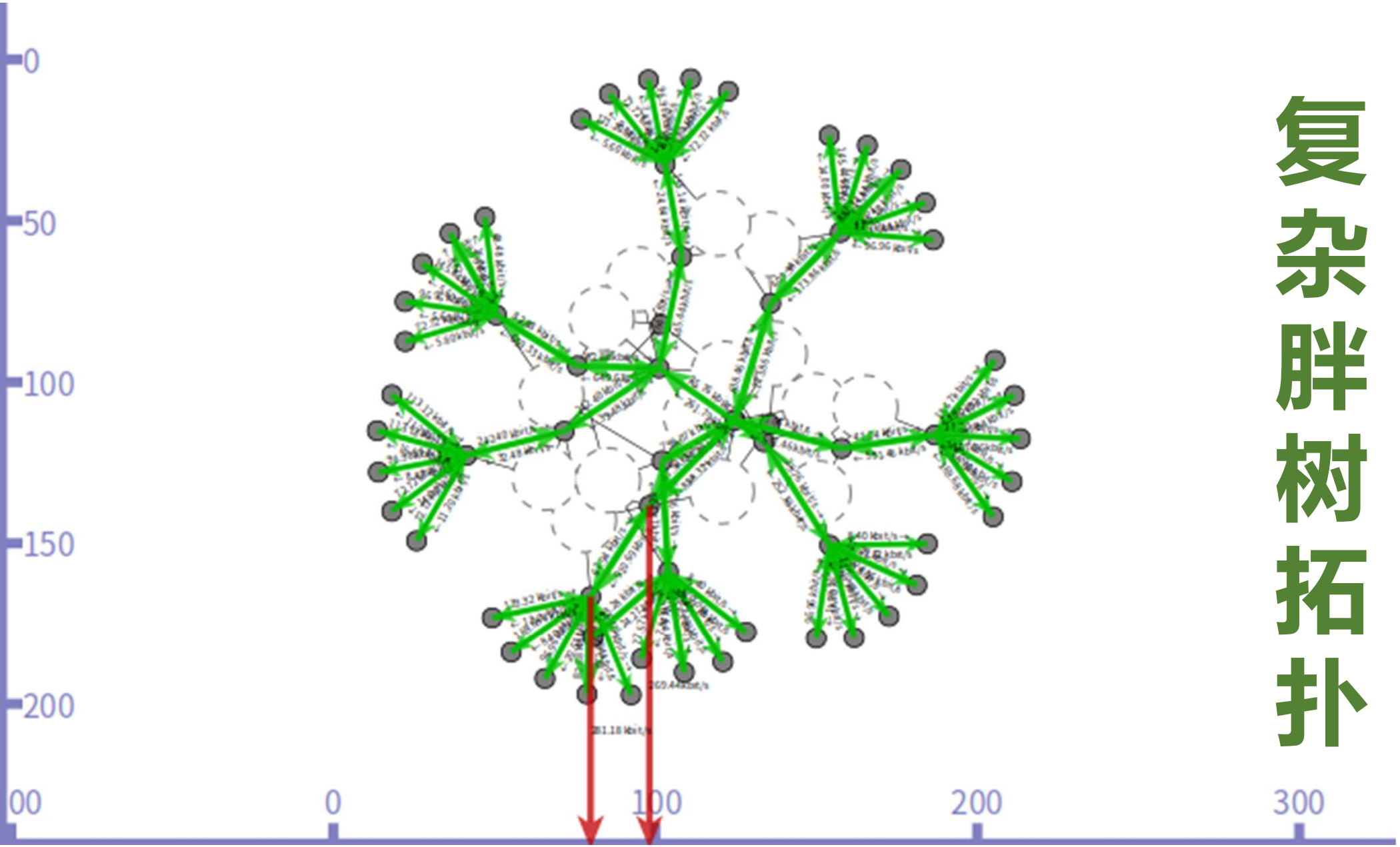

数据链路层
创建网络节点
定义节点容器
在节点容器中创立节点组
创建独立节点
连接网络节点创建网络链路
定义网络设备容器
安装网络设备
设置连接参数
网络层
创建网络层协议栈
安装协议栈
设置局域网IP地址
定义IP接口容器
设置IP地址和掩码
安装到相应链路
应用层
客户端（client）设置
设置数据包发送属性
令客户端节点向指定服务器节点发送指定数量、大小的数据包
设置数据包发送和结束时间
服务器（server）设置
令指定服务器节点处于接收状态
设置接收状态接受和结束时间
传输层
定义端口号
初始化TCP套接字


0
20
40
60
80
100
-50
0
50
100
50.11 kbit/s
983.14 kbit/s
1000.38 kbit/s
50.11 kbit/s
23.33 kbit/s
948.64 kbit/s
646.80 kbit/s
15.12 kbit/s
758.91 kbit/s

简单拓扑



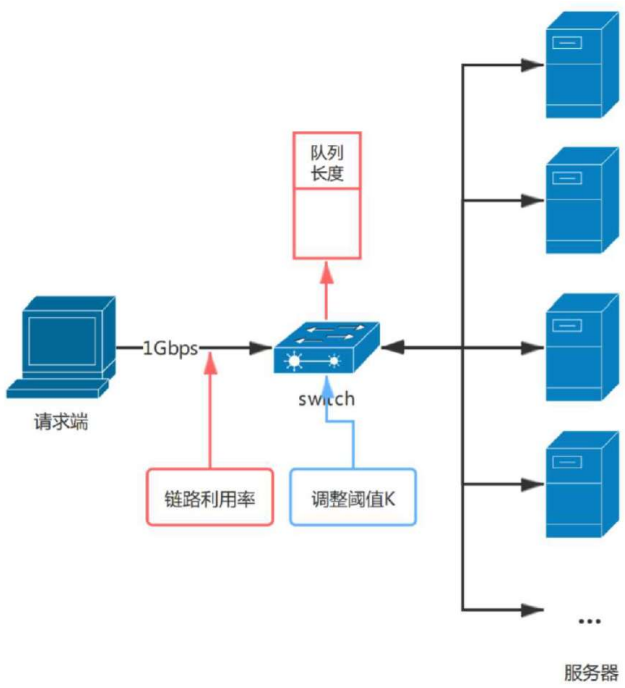
二级胖树拓扑

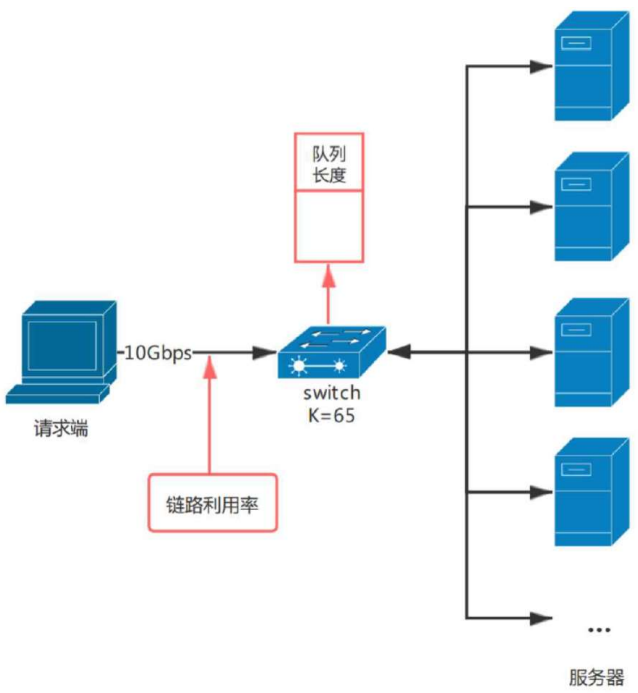


复杂胖树拓扑

吞吐量和队列长度

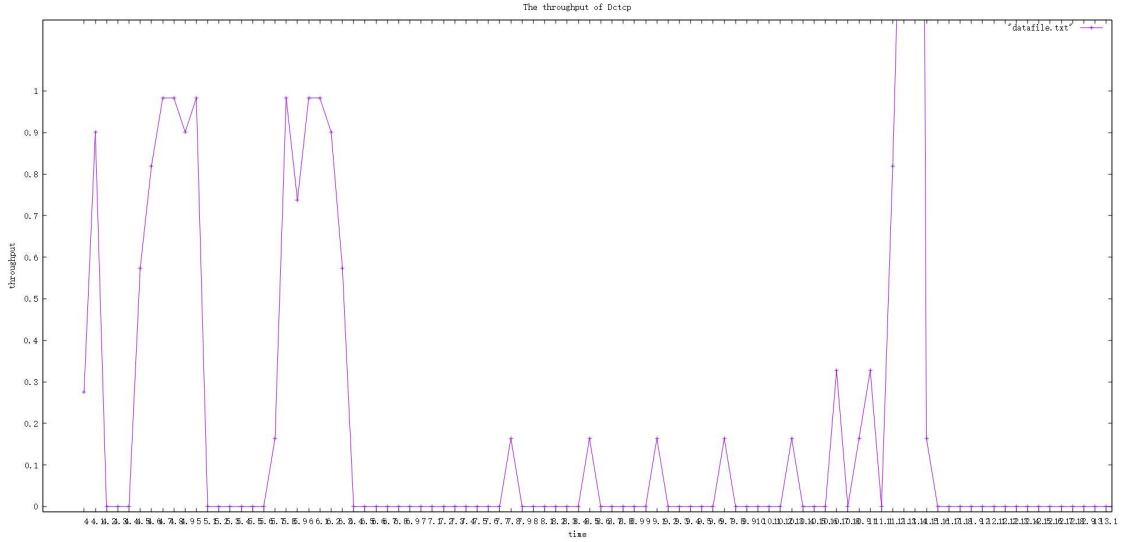
吞吐量和交换机内队列长度是衡量网络拥塞情况的重要指标，也是决定网络性能的重要标准。

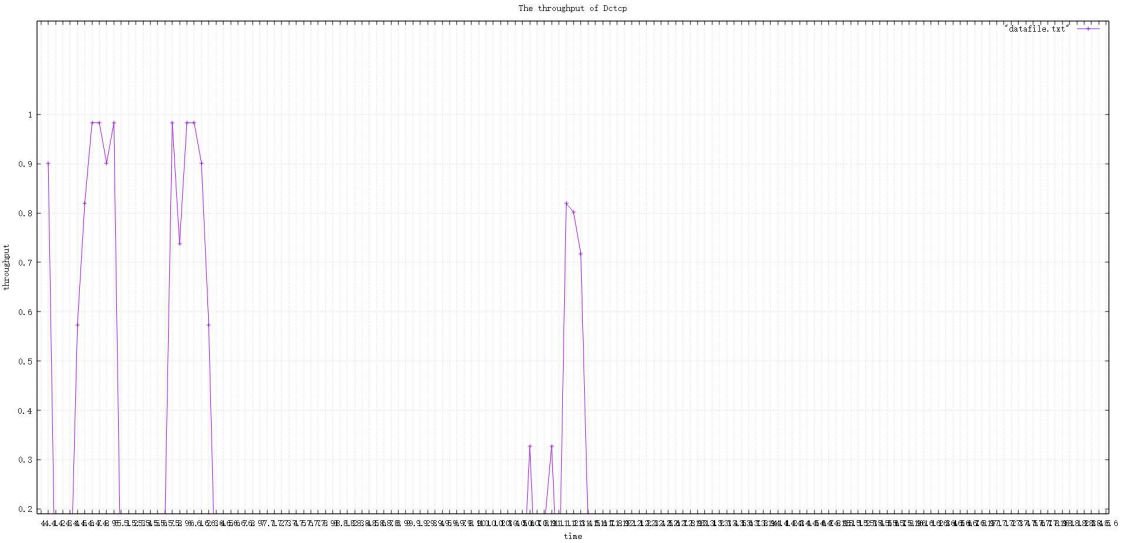


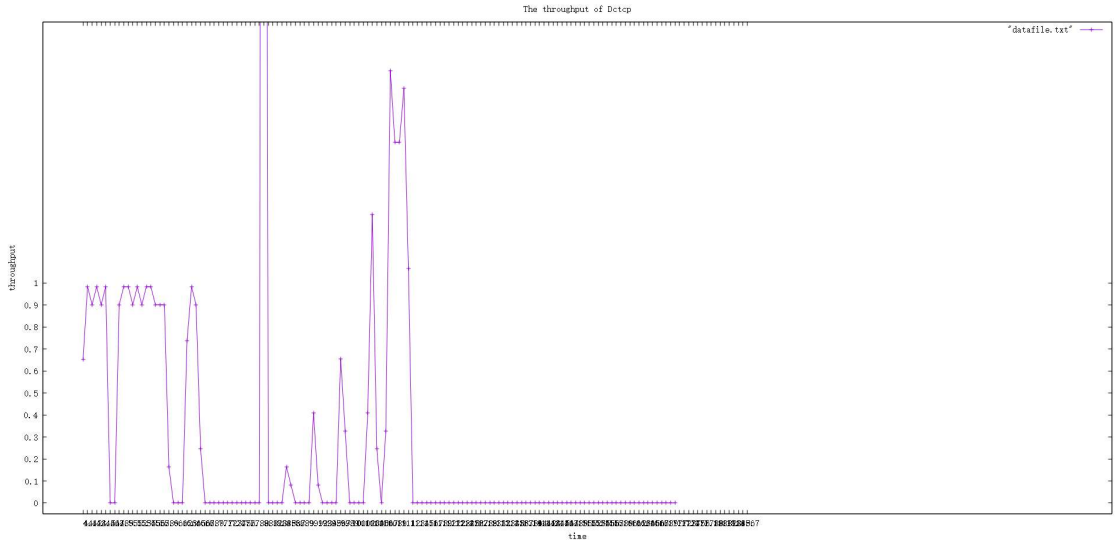


DCTCP协议

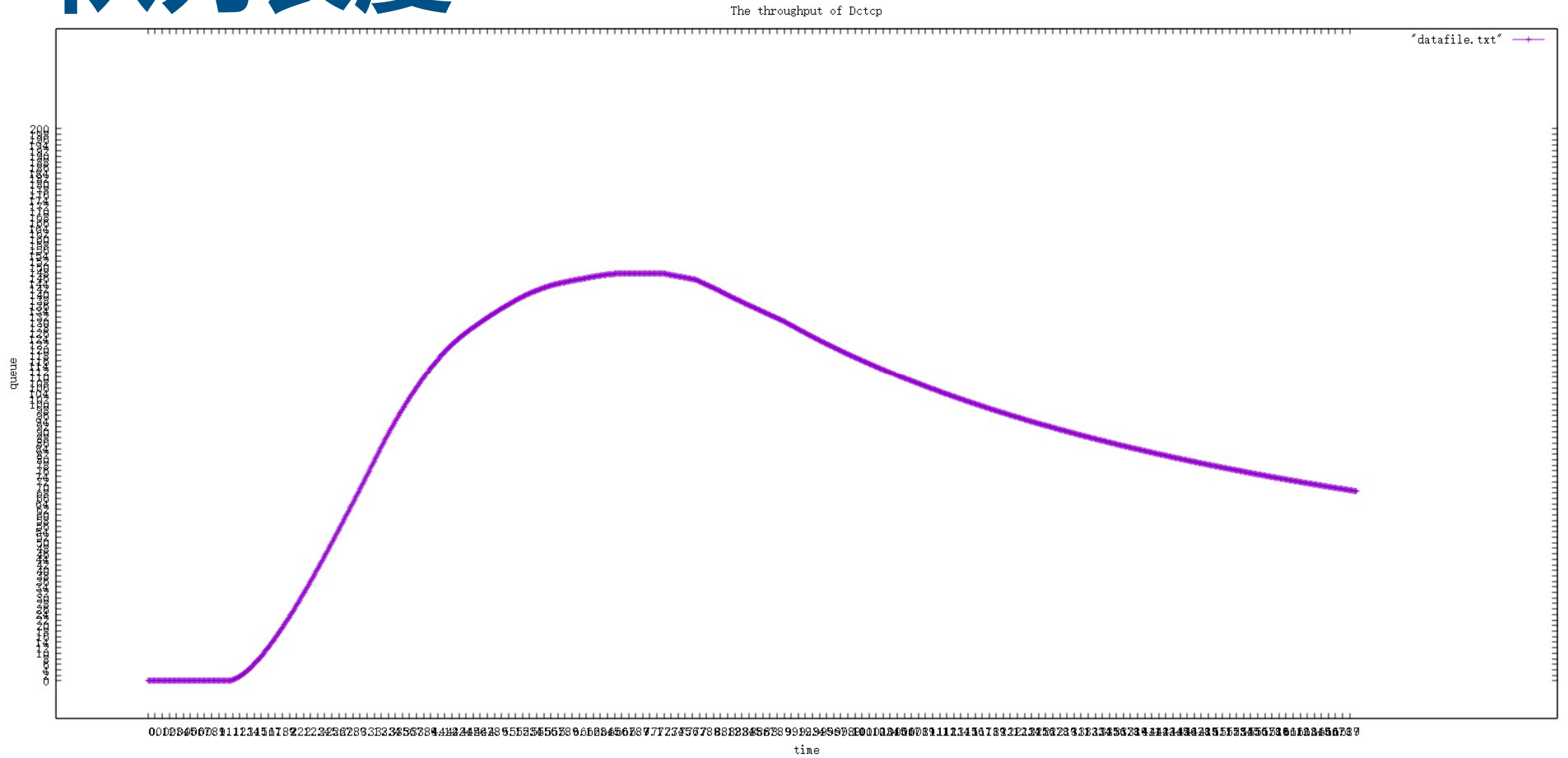
吞吐量


The throughput of Dctcp
"datafile.txt"
throughput
time


The throughput of Dctcp
"datafile.txt"
throughput
time


The throughput of Dctcp
"datafile.txt"
throughput
time

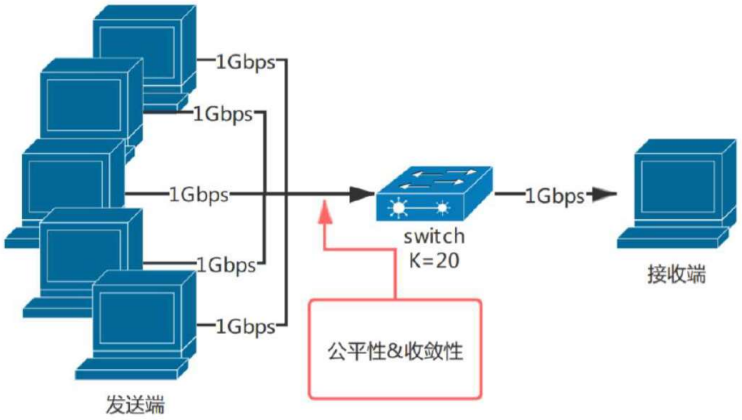
队列长度

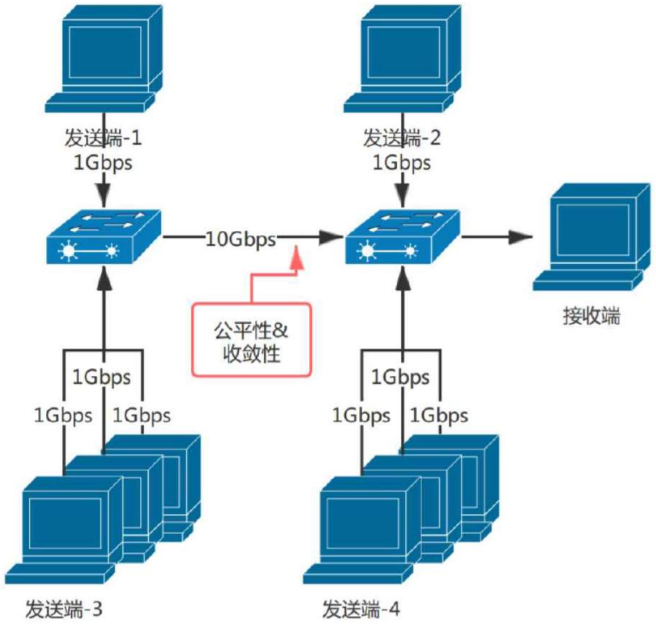


3 下阶段计划及预期成果

公平性和收敛性

公平性和收敛性共同反映了多流同时传输时各流相互抢占资源所导致的部分拥塞情况，同时收敛性可以反映链路拥塞情况在不同网络环境转换时的变化。





预期成果：发表论文一篇

数据中心网络要求具备三个条件：低延迟的短流流量、高突发容限和高利用率的长流流量。支持推动了大量网络搜索、零售、广告和推荐系统的数据中心的建设。

优化 TCP Incast 问题，可以让数据中心使用更多的低成本的交换机，使用低成本的商用组件构建高可用性、高性能的计算和存储基础架构。

4 成员分工情况


现阶段我们根据实验需要
分为了三个小组
张鑫沂 DCTCP协议
张艺彤 张宇彬 ICTCP 协议
武嘉闻 谢紫昂 L2DCT协议

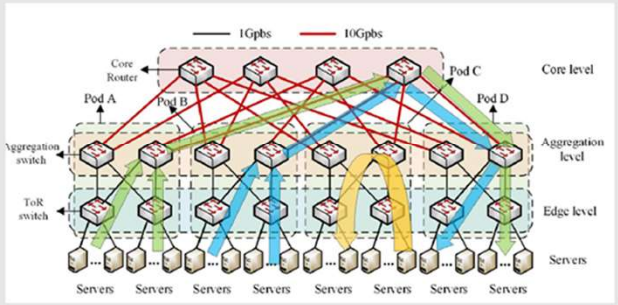
谢 谢

DCTCP

NS-3

ICDCP

L2DCT



指导教师：李 卓

负 责 人：武嘉闻

团队成员：张鑫沂
谢紫昂
张宇彬
张艺彤