Providing Bandwidth Guarantees, Work Conservation and Low Latency Simultaneously in the Cloud

Shuihai Hu, Wei Bai, Kai Chen, Chen Tian, Ying Zhang, and Haitao Wu

Abstract—Today’s cloud is shared among multiple tenants running different applications, and a desirable multi-tenant datacenter network infrastructure should provide bandwidth guarantees for throughput-intensive applications, low latency for latency-sensitive short messages, as well as work conservation to fully utilize the network bandwidth. Despite significant efforts in recent years, none of them can achieve these three properties simultaneously. In this paper, we identify the key deficiency of prior solutions and use this insight to motivate our design of Trinity—a simple, practical yet effective solution that achieves bandwidth guarantees, work conservation and low latency simultaneously in the cloud. We implement Trinity using existing commodity hardwares and demonstrate its superior performance over prior solutions using testbed experiments.

Index Terms—Data center network, public cloud, bandwidth guarantee, work conservation, low latency

1 INTRODUCTION

In today’s clouds, the network resource, unlike the compute and storage resources, is shared in an uncoordinated best-effort manner among multiple tenants. For this reason, the tenants may experience varied network performance which can adversely affect their application performance and increase their cost. For example, recent studies on several major cloud infrastructures have revealed that bandwidth and packet latency can vary significantly by an order of magnitude [1], [2], [3], [4], [5]. The lack of predictable performance has prevented users and enterprises from migrating their applications into the cloud, especially for delay sensitive applications such as web search, retail, advertising, recommendation systems, etc.

A natural way to provide predictable network performance is to let the users specify the amount of bandwidth they need and allocate dedicated bandwidth to them, i.e., providing bandwidth guarantees to the tenants. However, such strict bandwidth allocation may result in bandwidth waste if the tenant cannot fully utilize his share. Thus, the cloud network should also provide work conservation to enable the multiplexing economic benefits for the cloud provider. At the same time, it should provide low latency to short flows for small response time. As a result, a good cloud network design should be able to meet these three objectives simultaneously.

While significant efforts [1], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] have been made toward sharing the cloud and obtaining predictable network performance, none of them achieves all the three goals simultaneously. For example, SecondNet [9] and Oktopus [1] provide bandwidth guarantees, but they are not work-conserving. ElasticSwitch [6] aims at work-conserving bandwidth guarantees, however it cannot ensure low latency for short flows, and more importantly, its work conservation is sacrificed due to a fundamental tradeoff between accurately providing bandwidth guarantees and being work-conserving (see details in Section 2).

We identify that a key deficiency of prior solutions such as ElasticSwitch [6] is that they heavily rely on end-to-end rate control while neglecting important support from network. The reason why ElasticSwitch has to sacrifice work conservation for bandwidth guarantees is that: it injects rate control while neglecting important support from network. Consequently, work-conserving traffic of one tenant, if too aggressive, can adversely affect bandwidth guarantee traffic of other tenants and hurt the latency of their short flows.

This directly motivates our design of Trinity in this work. By Trinity, we show that simple network support can be explored to solve the problem. Observing that today’s commodity switches already support 4–8 priority queues [16], [17], [18], our key idea in Trinity is that by simply differentiating the two types of traffic at the end and prioritizing them in the network, we can readily achieve all the triple goals simultaneously.

---

111909/YCC.2018.2890252
Basically, Trinity decouples providing bandwidth guarantees from being work-conserving by segregating these two types of traffic at the end, and leveraging commodity switches to enforce priority queueing in the network. The traffic of bandwidth guarantees is prioritized over that of work conservation. With such prioritization, work conservation can be designed aggressively without affecting bandwidth guarantees. Furthermore, such prioritization also makes it easier for Trinity to achieve low latency for short flows: it only needs to classify packets of short flows as bandwidth guarantee traffic and let them receive priority in the network (see details in Section 3).

Despite being conceptually simple, there are still a few concrete issues we need to address before making Trinity truly effective. First, how to design an aggressive rate control algorithm so that the work-conserving traffic can fully utilize spare bandwidth in the network, while not causing a large number of packet drops at switches. Second, how to handle packet trapping (or starvation) of the work-conserving traffic in the lower priority queue. Third, how to deal with possible packet re-ordering which might occur when a long flow promotes from the lower priority queue to the higher one. In Section 3.3, we introduce how Trinity addresses each of them.

We have implemented a Trinity prototype with commodity servers and switches (Section 4). On the end host, our Trinity kernel module is located as a shim layer over the physical NIC (Network Interface Card) driver in hypervisor. It does not introduce any modification to network stacks or applications of tenants. In the switch, Trinity only requires strict priority queueing and Explicit Congestion Notification (ECN) which are both built-in functions for existing commodity switches.

To evaluate Trinity, we build a testbed with 2 Pronto-3295 Gigabit Ethernet switches and 16 Dell servers. Our experimental results show that:

- Trinity provides accurate bandwidth guarantees while achieving good work conservation. For example, Trinity outperforms ElasticSwitch by 20.88–53.06 percent in terms of the average throughput under different settings;
- Trinity delivers low latency for short flows and improves their flow completion time (FCT) significantly. For example, as for 1 KB short flows, compared to ElasticSwitch, Trinity reduces their FCT by 22-33 percent on average and by 68-71 percent at the 99th percentile;
- Trinity improves the performance of both throughput-sensitive applications and latency-sensitive applications. For instance, in latency-sensitive Memcached application, Trinity outperforms ElasticSwitch by 83 percent in terms of the query completion time at the 99th percentile;
- Trinity is scalable to be applied to large-scale data center networks. For example, Trinity only introduces 9 percent CPU overhead in the worst case when compared to no protection scheme.

The rest of the paper is organized as follows. Section 2 introduces the background and discusses some related works. Section 3 introduces the Trinity design in detail. Sections 4 and 5 describe the Trinity implementation and testbed experiments. Section 6 concludes this paper.

2 BACKGROUND

2.1 Scope and Non-Goals

In today’s clouds, it is easy to share the compute and storage resources effectively among multiple tenants. In contrast, sharing network resource with bandwidth guarantees is believed to be complex and challenging [1], [6], [8], [9], which requires the following three key technologies:

- Tenant abstraction. Tenant abstractions are used to model bandwidth guarantees needed by tenants. Many of these abstractions [1], [9], [10], [15], [19], [20], [21] offer a virtual network, which allows tenants to have an illusion of having their own dedicated networks. These abstractions usually consist of VMs, virtual switches and virtual links associated with bandwidth guarantees. For example, Fig. 1 shows an example of Hose model, where each of the VMs is connected with a virtual switch via a virtual link. Each virtual link is associated with a specified bandwidth to be guaranteed.
- VM placement algorithm. Given a tenant abstraction, VM placement algorithm [1], [9], [10], [15] is needed to determine whether this abstraction can be deployed on the physical datacenter network without violating the guarantees for the existing tenants. For example, Fig. 1 also shows an placement of the two hose models on the physical network.
- Runtime enforcement. Runtime enforcement [6], [8], [14] is used to enforce the bandwidth guarantees specified in each placed tenant abstraction. Note that by only placing VMs on the physical networks, bandwidth guarantees are not necessarily satisfied even if enough bandwidth is provisioned. For example, in Fig. 1, if VM X sends traffic at a rate much larger than $B_X$, the bandwidth guarantee of VM E may be violated.

The focus of this paper is to design a scalable and efficient runtime enforcement mechanism, such that bandwidth guarantees, work conservation and low latency can be simultaneously achieved. We rely on prior solutions for designing tenant abstractions and doing VM placement.
Trinity can support both pipe mode [9], [19], [20] and hose model [1], [21]. Trinity can also work with abstractions based on the hose model, such as TIVC abstraction [10] and TAG abstraction [15].

Trinity is agnostic to the VM placement algorithm. For schemes that support online reconfiguration of bandwidth guarantees [22], [23], Trinity can still be applied as long as the network manager updates the tenant abstractions and their placement after each reconfiguration.

2.2 Problem and Related Work

Given tenant abstractions and a valid placement, the cloud provider needs to have a runtime mechanism to enforce the bandwidth guarantees and utilize unused bandwidth efficiently. We argue that a good mechanism should satisfy three properties in the following.

- Providing bandwidth guarantees means that each VM can share a minimum guaranteed bandwidth to send and receive traffic whenever needed. This is crucial for the predictable application performance, especially for data-intensive applications [24], [25] whose completion time mainly rely on the available network bandwidth.

- Being work-conserving requires that the bottleneck link should be always fully utilized as long as there are sufficient demands. This means that a tenant should be able to dynamically grab free bandwidth, which are either unallocated, or allocated but are not currently used by other tenants. Work conservation benefits both tenants and the provider because tenants can finish their jobs faster and the provider can achieve high resource utilization.

- Delivering low latency for short flows is crucial for many online data-intensive (OLDI) applications such as web services. For better user experience, many OLDI applications operate under soft real-time constraints that requires short flows to be completed before deadlines [26].

To the best of our knowledge, prior solutions do not achieve the three goals simultaneously. Table 1 summarizes some related work according to the objectives they meet and the assumptions they have. Specifically, SecondNet [9], Oktopus [1] and TIVC [10] provide bandwidth guarantees but not work-conserving, while Seawall [11] does the opposite. EyeQ [8] and GateKeeper [14] are work-conserving, but they require the network core to be congestion-free which is not the case for production datacenters [6]. Similarly, FairCloud PS-P [13] is also work-conserving, but at the cost of expensive switch hardware support especially per-VM queues. Furthermore, all these solutions do not consider low latency. On the other hand, Silo [7] considers guaranteed bandwidth and packet latency, but it does not achieve work conservation.

We also note that there exist some traditional solutions that tackles similar problems in the broader context of the Internet. For example, weight fair queuing (WFQ) [27] can be borrowed to achieve bandwidth guarantees and work conservation by using per-tenant dedicated queues. However, today’s commodity switches have a limited number of queues (e.g., 4-8), which is far from enough for clouds with many tenants. For some other advanced schemes like [28], [29], their algorithms are too complicated to be implemented in commodity switches. Deep Dice. The work closest to Trinity is ElasticSwitch [6]. However, there is a fundamental tradeoff between accurately providing bandwidth guarantees and being work-conserving in ElasticSwitch. In order for a tenant to detect the spare bandwidth not being used by other tenants, ElasticSwitch needs to probe the available bandwidth by increasing the flow rates. However, probing too conservatively (i.e., increase gradually but drop dramatically) may under-utilize the available bandwidth and is not sufficiently work-conserving; while probing too aggressively (i.e., increase dramatically but drop gradually) may affect bandwidth guarantees of other tenants when their traffic come back to network.

We show this dilemma using testbed experiments in Fig. 2. As shown in Fig. 2a, there are four VMs of two tenants A and B sharing a same bottleneck link, and VM A1 and B1 send traffic to A2 and B2, respectively. We measure the throughput at A2 and B2 every 5ms. In the first experiment, we assume both tenants have 150 Mbps guarantees and use conservative probe. In this case, the ideal work conservation result should be that both tenants stay

<table>
<thead>
<tr>
<th>Related Work</th>
<th>Design objectives</th>
<th>System requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BW guarantee</td>
<td>Work conservation</td>
</tr>
<tr>
<td>Oktopus [1], TIVC [10]</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>SecondNet [9]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GateKeeper [14]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EyeQ [8]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seawall [11], NetShare [12]</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>FairCloud PS-P [13]</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Silo [7]</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>ElasticSwitch [6]</td>
<td>Yes, tradeoff with work conservation</td>
<td>Yes, tradeoff with BW guarantee</td>
</tr>
<tr>
<td>Trinity</td>
<td>Yes, without tradeoff</td>
<td>Yes, without tradeoff</td>
</tr>
</tbody>
</table>

**TABLE 1**

Summary of Previous Approaches and Comparison to Trinity

---

Authorized licensed use limited to: Nanjing University. Downloaded on June 06,2021 at 13:30:41 UTC from IEEE Xplore. Restrictions apply.
around 500 Mbps. However, in Fig. 2b, we can see that the scheme is not fully work-conserving, because it probes available spare bandwidth too conservatively by increasing rates slowly at the beginning, but dropping too dramatically once it senses congestion. Spare bandwidth in the valleys is wasted.

In the second experiment, we assume tenant A and B have 150 Mbps and 750 Mbps guarantees respectively and use aggressive probe. We let A take more spare bandwidth first, and later on more traffic from B arrives. However, in Fig. 2c, we can see that, under such aggressive probe, B even cannot get back its guaranteed bandwidth for a long while. The reason is that the work-conserving traffic of A adversely throttles the bandwidth guarantee traffic of B. Because A drops gradually upon congestion (but increases dramatically when seeing spare bandwidth), it makes B unable to grab its minimum guarantee of 750 Mbps in a short time (although eventually it will).

We note that, in ElasticSwitch [6], they proposed solutions such as 10 percent headroom, hold-increase and rate-caution which essentially trade work-conservation for bandwidth guarantees, but do not completely solve the problem.

3 THE TRINITY DESIGN

3.1 Design Overview

To solve the above dilemma, we seek network support instead of sticking to pure end-to-end solution. By simply differentiating the two types of traffic at the end and prioritizing them in the network, we break the impasse.

Specifically, Trinity decouples providing bandwidth guarantees from being work-conserving by differentiating traffic of bandwidth guarantees from that of work conservation with two colors at the end (i.e., green indicates bandwidth guarantee traffic, and red indicates work-conserving traffic), and leveraging commodity switch capability to enforce strict priority queuing in the network. That is, the traffic of bandwidth guarantees is always prioritized over that of work conservation in the network. With such prioritization enforced, work conservation can now be designed more aggressively without causing any interference to bandwidth guarantees. This effectively enables Trinity to achieve absolute bandwidth guarantees and work conservation without any tradeoff.

Meanwhile, such prioritization of bandwidth guarantee traffic over work conservation traffic also enables Trinity to optimize and ensure low latency for short flows: it only needs to make sure that the packets of short flows are colored as bandwidth guarantee packets. The reason is as follows. Since tenant’s bandwidth guarantee requirement has already been met by the provider based on the network capacity in the tenant admission control phase, pure bandwidth guarantee traffic can be accommodated by the network without congestion and the packets will experience little, if any, queuing delay. In the case of mixed bandwidth guarantee and work-conserving traffic, as long as the prioritization is in place, bandwidth guarantee packets will not be blocked by work-conserving traffic, and thus still be able to see low latency.

3.2 System Framework

The system framework of Trinity is shown in Fig. 3. Trinity software component runs in the hypervisor layer, and mainly consists of two parts, a Rate Controller (RC) module and a number of VM-to-VM channels. For each source-destination pair of VMs, there is one corresponding VM-to-VM channel.

Rate Controller Module. RC is responsible for determining the traffic rates for bandwidth guarantees and work conservation for each VM pair. At the receiver side, RC periodically measures latest channel rates and ECN signals, and feeds back to the corresponding senders. At the sender side, RC calculates bandwidth guarantee rate $R_G$ and work...
As introduced, for each VM-to-VM channel, RC should be the bandwidth guarantee of VM X. If Trinity is similar.

The basic idea is as follows. When there is no congestion feedback from network, we allow a VM-to-VM channel to send work-conserving traffic aggressively without any limits as long as the NIC allows; When there is congestion feedback from network, we employ aggressive rate control algorithm for work conservation without affecting bandwidth guarantees. Then the question is: what kind of rate control we should employ?

Solution. As introduced, for each VM-to-VM channel, RC needs to compute a bandwidth guarantee rate \( R_G \) and a work-conserving rate \( R_W \).

At first, we describe the calculation of \( R_G \). For pipe models where bandwidth guarantees between pairs of VMs are directly specified, \( R_G \) is the input. For hose models where only per-VM aggregated bandwidth guarantee is specified, Trinity borrows the approach of ElasticSwitch [6] to transform a hose model into a set of minimum bandwidth guarantees between VM pairs:

For a channel \( X \to Y \), RC sets its bandwidth guarantee rate as:

\[
R_G^{X\to Y} = \min(B_X^{X\to Y}, B_Y^{X\to Y}),
\]

where \( B_X^{X\to Y} \) is the guaranteed bandwidth assigned by X's hypervisor for the traffic to Y, and \( B_Y^{X\to Y} \) is the guaranteed bandwidth assigned by Y's hypervisor for the traffic receiving from X. Let \( B_X \) be the bandwidth guarantee of VM X. If X is sending traffic to N destination VMs with unbounded bandwidth demand, we have \( B_X^{X\to Y} = B_X/N \). The computation for \( B_Y^{X\to Y} \) is similar.

To fully utilize the spare bandwidth, Trinity adopts an aggressive algorithm for the calculation of \( R_W \). The basic idea is as follows. When there is no congestion feedback from network, we allow a VM-to-VM channel to send work-conserving traffic aggressively without any limits as long as the NIC allows; When there is congestion feedback from network, we employ aggressive rate control algorithm for work conservation without affecting bandwidth guarantees. Then the question is: what kind of rate control we should employ?

Solution. As introduced, for each VM-to-VM channel, RC needs to compute a bandwidth guarantee rate \( R_G \) and a work-conserving rate \( R_W \).

At first, we describe the calculation of \( R_G \). For pipe models where bandwidth guarantees between pairs of VMs are directly specified, \( R_G \) is the input. For hose models where only per-VM aggregated bandwidth guarantee is specified, Trinity borrows the approach of ElasticSwitch [6] to transform a hose model into a set of minimum bandwidth guarantees between VM pairs:

For a channel \( X \to Y \), RC sets its bandwidth guarantee rate as:

\[
R_G^{X\to Y} = \min(B_X^{X\to Y}, B_Y^{X\to Y}),
\]

where \( B_X^{X\to Y} \) is the guaranteed bandwidth assigned by X's hypervisor for the traffic to Y, and \( B_Y^{X\to Y} \) is the guaranteed bandwidth assigned by Y's hypervisor for the traffic receiving from X. Let \( B_X \) be the bandwidth guarantee of VM X. If X is sending traffic to N destination VMs with unbounded bandwidth demand, we have \( B_X^{X\to Y} = B_X/N \). The computation for \( B_Y^{X\to Y} \) is similar.

To fully utilize the spare bandwidth, Trinity adopts an aggressive algorithm for the calculation of \( R_W \). The basic idea is as follows. When there is no congestion feedback from network, we allow a VM-to-VM channel to send work-conserving traffic aggressively without any limits as long as the NIC allows; When there is congestion feedback from network, we employ aggressive rate control algorithm for work conservation without affecting bandwidth guarantees. Then the question is: what kind of rate control we should employ?
network, we reduce $R_W$ in proportion to an estimation of network congestion.

Formally, let $S$ be the set of VMs hosted on a server, and $\forall X \in S$, we use $B_X$ to denote the bandwidth guarantee of $X$. Assume the capacity of the NIC is $C$, then the spare capacity:

$$C_W = C - \sum_{X \in S} B_X. \quad (2)$$

To apply the idea mentioned above, the hypervisor divides all the active VM-to-VM channels into two sets $P$ and $Q$, and computes work conserving rates for channels in different sets using different schemes. Here $P$ is the set of congestion-free channels, while $Q$ is the set of congestion-caution channels. Initially, we put all active channels in $P$.

Let $C_P$ be the total spare capacity belonging to the congestion-free channels. It is easy to know that:

$$C_P = C_W - \sum_{u \in Q} R_W^{u\rightarrow v}. \quad (3)$$

Here $R_W^{u\rightarrow v}$ is the work-conserving rate of channel $u \rightarrow v$. For a channel $X \rightarrow Y$ in $P$, its work conserving rate is:

$$R_W^{X\rightarrow Y} = C_P \times \frac{R_X^{Y\rightarrow X}}{\sum_{u \in P} R_G^{Y\rightarrow v}}. \quad (4)$$

It means that all channels in $P$ share spare capacity $C_P$ in a weighted fair sharing fashion, where the weight is set as the bandwidth guarantee.

Although our Trinity ensures that work-conserving traffic will not affect bandwidth guarantee traffic, sending too much work-conserving traffic may cause a large number of packet losses in the low priority queues on switches, which will result in TCP timeout and thus hurt the throughput of TCP flows.

To address this, we enable ECN in the low priority switch queues. On the end hosts, we let hypervisors monitor the congestion feedback of ECN marking. Specifically, hypervisors will maintain an estimation of the fraction of red packets that are marked with ECN (denoted as $\beta$) in each period for all channels. If a congestion-free channel is detected to be congested, it will be moved from $P$ to $Q$. For any congestion-caution channel $X \rightarrow Y$ in $Q$, in each period, if $\beta$ is non-zero, we reduce its work conserving rate in proportion to $\beta$ in a manner similar to DCTCP [30], i.e.,

$$R_W^{X\rightarrow Y} = (1 - \beta/2) \times R_W^{X\rightarrow Y}. \quad (5)$$

If $\beta$ is zero, it means that there is no congestion in the network. We then increase its work conserving rate as follows:

$$R_W^{X\rightarrow Y} = \min(R_P, (1 + \alpha) \times R_W^{X\rightarrow Y}). \quad (6)$$

Here $R_P$ is the work-conserving rate this channel will be allocated if it is a congestion-free channel. A congestion-caution channel should get no more allocation than its share as a congestion-free channel. $\alpha$ is a factor used to control the aggressiveness of rate increase. If $R_W^{X\rightarrow Y} = R_P$ after updating rate, the hypervisor will move this channel back to $P$.

In a public cloud, we cannot assume all tenants support ECN in their transport layer protocols. To make our ECN-based solution practical: at the sender side, for all out-going packets, the hypervisor sets the ECN-capable bits in IP header to be true; at the receiver side, the hypervisor estimates the fraction of ECN marked incoming red packets for every VM-to-VM channel, and sends this estimation back to the corresponding hypervisor at the sender side periodically.

In addition, the hypervisor should also record whether a connection supports ECN. For a TCP connection, the hypervisor can know whether it supports ECN in 3-way handshakes. For those flows that disable ECN, to avoid disturbing the function of their transport layer protocols, the hypervisor will clear the ECN bits when delivering packets to upper layer.

**Problem #2: Packet Trapping.** There exist scenarios that red packets can get trapped (starved) in the lower priority queue of a bottleneck switch. For example, initially the switch has spare bandwidth (by other tenants’ bandwidth guarantees but currently not being used) for work-conserving traffic and thus some red packets get in the lower priority queue. Suddenly, the bandwidth guarantee packets of other tenants come back and occupy the bandwidth for a long duration. Then, the work-conserving red packets get trapped due to lower priority. As a consequence, the TCP sender of those red packets responds by retransmitting the packets repeatedly, and these retransmitted packets may get dropped persistently since the bottleneck queue is already full.

**Solution.** Reserving sufficient bandwidth headroom for work-conserving traffic can potentially address this problem, however it is a waste of bandwidth and thus not desirable.

We introduce a simple solution to this problem without bandwidth headroom. As mentioned above in rate control, the hypervisor at the receiver side will estimate the fraction of ECN marked incoming red packets for every VM-to-VM channel periodically. Then if a hypervisor does not receive any red packets for a VM-to-VM channel in the last period, it can send a message to inform the corresponding hypervisor at the sender side of the possible packet trapping. The hypervisor at the sender then checks how many red packets the source VM has sent out for this VM-to-VM channel in the last period. If the source VM does send out some red packets, it indicates packet trapping in the network. The hypervisor then sets the work-conserving rate $R_W$ to a small value (e.g., 10 Kbps), and marks this channel as congestion-caution channel.

**Problem #3: Packet Re-Ordering.** For a short flow, all of its packets are colored as green, there is no out-of-order problem. While for a long flow, due to instantaneous token availability in Token Bucket 1, the packets can alternate between green and red. It is possible that in the same long flow, some packets with smaller sequence numbers are colored as red as tokens run out in Token Bucket 1, while subsequent packets with larger sequence numbers are colored as green because new tokens are being generated. In such case, packet re-ordering could arise because red packets may experience longer queueing delay in the network and reach the destination later than green ones. This is detrimental to TCP throughput, by triggering window collapse and unnecessary retransmissions.

**Solution.** The solution to this problem is twofold. At the sender, we minimize the case that packets of a flow alternate from red back to green. Specifically, we introduce a color
transition delay parameter \( t \): when there is a need to change the colors of packets from red to green, we defer the change by \( t \) seconds. There are two benefits for this delay. First, it increases the chance that some other flows may come up and consume the tokens in Token Bucket 1 without packet re-ordering. Second, it decreases the chance that packet re-ordering happens with the flow itself because this has already reserved some additional time for the red packets to transmit. At the receiver, we adopt a re-sequencing buffer [31], [32], [33] to absorb possible out-of-order packets as shown in Fig. 3. More specifically, if a green packet \( p \) is received and some packets \( p_i \), prior to it have not been received yet, Trinity puts \( p \) into the re-sequencing buffer and a timer is initiated. If all \( p_i \)s are received at a time \( t \) before timeout, they are submitted to TCP receiver together with \( p \) immediately; Otherwise, the whole buffer is submitted when timeout.

4 IMPLEMENTATION AND TESTBED SETUP

4.1 Trinity Implementation

Trinity consists of two components on end-hosts: receiver RX processing and sender TX processing. As a prototype, we have implemented TX and RX processing as a Linux kernel module. We also implemented a ElasticSwitch-like kernel module following the description of [6]. The kernel module is located as a shim layer above the physical NIC driver in hypervisor, without touching network stacks and applications of tenant’s VMs. We also developed an application to configure Trinity kernel module in user space. The application communicates with kernel module using ioctl [34]. Our implementation code in total consists of about 1000 lines of C code and about 900 lines of header files. We now describe each component in detail.

**Sender Trinity Module.** The sender module consists of a hash based flow table and multiple TX contexts. The flow table is used for tracking per-flow state and packet classification. Its operations are as follows: 1) All of the outgoing packets are intercepted by Netfilter hook at LOCAL_OUT and directed to the flow table [35]. 2) Each flow in the flow table is identified by the 5-tuple: source/destination IPs, source/destination ports and protocol. When a packet comes in, we identify its corresponding flow entry (or create a new entry) and update the amount of bytes sent. 3) Based on the bytes sent information, we classify packets and direct them to FIFO queues of corresponding TX contexts.

We allocate a TX context for each VM-to-VM pair. The TX context maintains basic TX information and a rate limiter. Unlike traditional token bucket rate limiter, our rate limiter has two associated FIFO queues, a timer, two rates (\( R_H \) and \( R_C \)) and corresponding two kinds of tokens. The packets of short-flow class and long-flow class are segregated by two FIFO queues. To enforce accurate rates over short timescales and avoid long delay to short-flow packets, we use Linux high-resolution kernel timer, HRTIMER [36], for our rate limiters. Once the timer fires, we update two kinds of tokens and begin packet scheduling. The packets from short-class FIFO queue has the high priority to be dequeued but they can only consume tokens for bandwidth guarantee traffic. After scheduling short-flow class packets, the packets from long-flow class can be dequeued and they can consume both of two kinds of tokens. The dequeued packets consuming different kinds of tokens will be marked with different Different Service Code Point (DSCP) values and enqueued to different priority queues in network switches.

**Receiver Trinity Module.** The receiver modules consists of multiple RX contexts and a control packet generator. We pre-allocate a RX context for each VM-to-VM communication pair. The RX context tracks the VM-to-VM pair’s receive traffic and measures incoming throughput. In each control interval, the RX context calculates the fraction of ECN marking packets and delivers this to source VMs using special feedback packets. Similar to EyeQ [8], our feedback packet is a special minimum sized IP packet (64 bytes) with a special unused IP protocol number (143 in our implementation). We encode the ECN fraction in the IP identification field. Since we only generate a packet for each VM-to-VM pair every control interval, the feedback traffic consumes limited network bandwidth. Considering a VM concurrently receiving traffic from 100 VMs, the feedback traffic only consumes ~50 Mbps throughput over the control interval of 1 ms. Furthermore, we can also piggyback the feedback information on packets back to the source VM. To achieve low latency for control messages, the feedback packets will be marked with DSCP of bandwidth guarantee traffic and sent out without going through rate limiters. To not disturb tenant’s network stacks, the RX context also clears any possible ECT and ECN marks in incoming packets when a tenant disables ECN function.

4.2 Testbed Setup

To evaluate Trinity, we build a dumbbell testbed with 16 servers connected to 2 Pronto-3295 48-port Gigabit switches as shown in Fig. 5. We configure strict priority queuing and per-queue ECN marking on switches. The shared buffer is enabled on our switches by default. With per-queue ECN marking, each queue has its own marking threshold and performs ECN marking independently to other queues. Packets are classified into different priority queues based on

**Fig. 5. Trinity testbed.**
their DSCP values. Each server is a Dell PowerEdge R320 with a 4-core Intel E5-1410 2.8 GHz CPU, 8G memory, a 500 GB hard disk, and a Broadcom BCM5719 NetXtreme Gigabit Ethernet NIC. Each server runs Debian 6.0-64 bit with Linux 2.6.38.3 kernel. Due to the limited number of CPU cores in our physical servers, we emulate multiple VMs by creating multiple virtual network interfaces with different IP addresses to avoid virtualization overheads. In our experiments, each tenant has its own virtual subnets.

5 Evaluation

We evaluate Trinity using testbed experiments. Our evaluation centers around four key questions:

- **Does Trinity have any tradeoff between bandwidth guarantee and work conservation?** By comparing to three other schemes: no protection, static reservation and ElasticSwitch, we show that Trinity can accurately provide minimum bandwidth guarantees while at the same time enabling VMs with large bandwidth demand to fully utilize spare link capacity. Specifically, Trinity outperforms ElasticSwitch by 20.88–53.06 percent in terms of average throughput under different settings.

- **Can Trinity deliver low latency for short flows and benefit their flow completion time?** We evaluate the scenarios where short flows coexist with long flows. Our results show that, compared to ElasticSwitch, Trinity improves the FCT by 22-33 percent on average and 68-71 percent at the 99th percentile for 1 KB short flows; furthermore, it reduces the FCT by 21-38 percent on average and 62-70 percent at the 99th percentile for 20KB short flows.

- **Can Trinity improve the performance of both throughput-sensitive applications and latency-sensitive applications?** As for FileStore application, which is throughput-sensitive, Trinity achieves good work conservation and outperforms ElasticSwitch by 5.4 percent in terms of data shuffle completion time. As for Memcached application, which is latency-sensitive, Trinity ensures low latency and outperforms ElasticSwitch by 83 percent in terms of the query completion time at the 99th percentile.

- **Is Trinity scalable to be applied to large-scale data center networks?** We evaluate the CPU overhead of Trinity under various scenarios. Our results show that, compared to no protection scheme, Trinity only introduces 5 percent additional CPU usage in the most stressed scenario.

In addition, we also evaluate the convergence time of Trinity. We find that under Trinity, rates of VMs can quickly adapt to varying network condition and converge within a short time.

**Schemes Compared:** We mainly compare Trinity against ElasticSwitch [6], static reservation (Oktopus-like [1]) and no reservation in our testbed. Among them ElasticSwitch is our closest work to compare. Qualitative analysis of other schemes like Gatekeeper [14] and EyeQ [8] shows that those approaches cannot provide guarantees when the network core is congested, so we exclude them in our testbed experiments.

**Parameters:** The rate control interval is set to 5 ms. We set ECN marking threshold to be 30 KB as DCTCP [30] recommends. For the rate control algorithm of ElasticSwitch [6], we also use its recommended algorithm.

5.1 Bandwidth Guarantees and Work Conservation

We show that Trinity can provide bandwidth guarantee while achieving good work conservation when multiple tenants are competing for the same bottleneck link.

**Many Connections versus One Connection.** In our experiments, there are four VMs (A1, A2, B1 and B2) of two tenants A and B sharing a same bottleneck link. VM A1 on server S1 sends traffic to VM A2 on server S9 using one TCP connection, while VM B1 on server S2 sends traffic to VM B2 on server S10 using different numbers of TCP connections.

We measure the throughput at VM A2 under four schemes: no protection, static reservation [1], [9], ElasticSwitch, and Trinity. In Fig. 6a, both tenants are provisioned with 100 Mbps guarantees. In Fig. 6b, both tenants are provisioned with 300 Mbps guarantees.
From the results, we make the following two observations: 1) No protection does not provide any bandwidth guarantee as link capacity is shared among all TCP connections. Static reservation provides minimum bandwidth guarantee, but does not utilize any spare bandwidth. ElasticSwitch provides bandwidth guarantees and utilizes part of the spare bandwidth. In contrast, Trinity not only provides bandwidth guarantee but also fully utilizes all the spare bandwidth. In terms of the average throughput, Trinity outperforms ElasticSwitch by 20.88 to 53.06 percent in different bandwidth guarantee settings. 2) ElasticSwitch wastes around 50 percent of the spare bandwidth. For instance, in Fig. 6a, when reserving 20 percent of the link capacity on the bottleneck link as bandwidth guarantees, ideally, VM A2 should achieve around 500 Mbps throughput on average. However, under ElasticSwitch, the average throughput of VM A2 is only about 230 Mbps.

We further look into the reason behind it by measuring the throughput of VM A2 every 5ms (i.e., rate control interval). In Fig. 6c, we show the result of the case where there are 10 TCP connections between VM B1 and B2. As illustrated, under ElasticSwitch, the throughput of VM A2 drops back to minimum guarantee as long as it senses congestion. Due to this conservative rate control, ElasticSwitch can only utilize about half of the spare bandwidth on average. On the other hand, our Trinity achieves nearly ideal throughput at the granularity of millisecond, this is because Trinity adjusts rate for each active VM pair based on a fine-grained estimation of the network congestion as introduced in Section 3.3.

A follow-up question may arise: can ElasticSwitch provide bandwidth guarantee and achieve good work-conservation by using the rate control algorithm of Trinity? We answer this question in the following experiment.

**Tradeoff between Bandwidth Guarantees and Work Conservation.** We denote ElasticSwitch with original rate control as conservative ElasticSwitch (C-ElasticSwitch), and with Trinity’s rate control as aggressive ElasticSwitch (A-ElasticSwitch). In this experiment, we use the same scenario as above, and measure the average throughput of VM A2 and B2 under C-ElasticSwitch, A-ElasticSwitch and Trinity. The number of TCP connections between VM B1 and B2 is set to 10.

The results in Fig. 7 show: 1) C-ElasticSwitch provides bandwidth guarantees but cannot fully utilize spare bandwidth as shown in Fig. 7a; 2) A-ElasticSwitch achieves good work conservation, but fails to provide bandwidth guarantees as shown in Fig. 7b; 3) Trinity provides accurate bandwidth guarantees while achieving good work conservation in both cases.

The takeaway of this experiment is that: 1) There is a tradeoff between bandwidth guarantee and work-conservation. Pure end-to-end solutions are difficult to achieve both goals simultaneously. 2) In-network prioritization with priority queueing is key to eliminating this tradeoff.

### 5.2 Low Latency for Short Flows

We show that Trinity can deliver low latency for short flows when short flows coexist with long flows.

**Tradeoff between Low Latency and Work Conservation.** It has been shown that, when most of the link capacity are reserved as guarantees, ElasticSwitch is work-conserving. However, we will show that there is actually a tradeoff between low latency and work-conservation. In this experiment, we have 6 VMs A1, A2, B1, B2, C1 and C2 of three tenants A, B and C. They are hosted on servers $S_1$, $S_2$, $S_3$, $S_4$, $S_5$, and $S_6$, respectively.

In this experiment, VM A1 sends 1 KB or 20 KB short flows to A2 periodically, and in the meantime, VM B1 and C1 send long flows to VM B2 and C2, respectively. To explore the tradeoff between low latency and work-conservation, we study two cases: 1. Three tenants are all provisioned with 200 Mbps guarantees on the bottleneck link, and thus we have 400 Mbps spare bandwidth; 2. Tenant A is provisioned with 200 Mbps guarantee on the bottleneck link. Tenants B and C are both provisioned with 400 Mbps guarantees on the bottleneck link. Hence no spare bandwidth is left in this case.

For case 1, the results are shown in Table 2. For case 2, the results are shown in Table 3. From the results, we observe that: 1) Compared to ElasticSwitch, Trinity reduces the FCT by 22-33 percent on average and by 68-71 percent at the 99th percentile for 1 KB short flows; furthermore, it reduces the FCT by 21-38 percent on average and by 62-70 percent at the 99th percentile for 20 KB short flows. 2) Although ElasticSwitch is work-conserving when 100 percent of link capacity is reserved as guarantees, it is at the cost of sacrificing latency of short flows. By comparing the results in Table 2 with that in Table 3, we can see that, under ElasticSwitch, the FCT
increases by 17-21 percent on average, and by 30-39 percent at the 99th percentile. In contrast, under Trinity, we do not observe any significant increase on the FCT.

The takeaway of this experiment is two-fold: 1) There is a tradeoff between low latency and work conservation. Pure end-host based solutions are difficult to achieve both goals simultaneously. 2) By letting packets of short flows receive high priority in the network, we can well address this tradeoff and improve the FCT of short flows significantly.

**Short Flows Mixed with Long Flows on End-Host.** If a VM is sending both long flows and short flows to a remote VM, then the congestion on end-host cannot be simply ignored anymore. Recall that in the design of Trinity, packets of short flows have higher priority to consume tokens in Token Bucket 1 over packets of long flows. We show that this mechanism can reduce end-host delay of short flows when short flows are mixed with long flows on end-host.

In this experiment, we change the scenario above by letting VM A1 send both short flows and long flows with unbounded demand to VM A2. In Table 4, we show the results of the case when only 60 percent of link capacity is reserved as guarantees.

From the results, we can find that: compared to ElasticSwitch, Trinity reduces the FCT by 82 percent on average and by 86 percent at the 99th percentile for 1KB short flows; Furthermore, it reduces the FCT by 78 percent on average and by 79 percent at the 99th percentile for 20 KB short flows. This implies that Trinity can reduce both in-network delay and end-host delay.

| Flow Completion Time of Short Flows (60 Percent of Link Capacity is Reserved as Guarantees) |
|-----------------------------------|-----------|-----------|-----------|-----------|
| Flow size | Trinity | ElasticSwitch | Flow size | Trinity | ElasticSwitch |
| Flow size | 1 KB | 20 KB | 1 KB | 20 KB |
| Average FCT(us) | 212 | 857 | 272 | 1083 |
| 99th percentile FCT(us) | 274 | 1104 | 857 | 2878 |

| Flow Completion Time of Short Flows (100 Percent of Link Capacity is Reserved as Guarantees) |
|-----------------------------------|-----------|-----------|-----------|-----------|
| Flow size | Trinity | ElasticSwitch | Flow size | Trinity | ElasticSwitch |
| Flow size | 1 KB | 20 KB | 1 KB | 20 KB |
| Average FCT(us) | 219 | 878 | 328 | 1413 |
| 99th percentile FCT(us) | 291 | 1218 | 1002 | 3997 |

### 5.3 Experiments with Applications

We show that: 1) For throughput-sensitive applications such as FileStore and MapReduce, Trinity can not only provide predictable network performance, but also reduce their shuffle completion times; 2) For latency-sensitive applications like Memcached, Trinity can improve its performance by reducing the query completion times.

**Throughput-Sensitive Applications.** Both FileStore and MapReduce are throughput-sensitive, and data shuffle is a potential bottleneck. In this experiment, we emulate the data shuffle phase of both jobs, and measure the shuffle completion times.

As for FileStore, on each server of \( S_1 \)-\( S_8 \), there is a 300 MB file to be copied to 4 random servers from \( S_9 \)-\( S_{16} \) to create some redundancies. In total, we have 32 communication pairs, and each pair is provisioned with 5Mbps guarantee. We create 8 parallel TCP connections between each pair.

As for MapReduce, we runs 12 VMs, one on each server of \( S_1 \)-\( S_{12} \). We let the 8 VMs on servers \( S_1 \)-\( S_8 \) (act as mappers) send 256 MB traffic to each of the 4 VMs on \( S_9 \)-\( S_{12} \) (act as reducers) to emulate data shuffle. It has 32 communication pairs in total and each pair has 15 Mbps guarantee. We create 2 parallel TCP connections between each pair.

**Fig. 8a** shows that under no protection, FileStore gets more bandwidth by creating more TCP connections and thus completes faster. This result implies that under no protection, tenants can cheat by creating more parallel TCP connections. **Fig. 8b** shows that under static reservation, both jobs have minimum performance guarantee, but spend much more time to finish (more than 500s).

**Fig. 8c** shows that compared to static reservation, ElasticSwitch reduces the shuffle completion time of FileStore and MapReduce jobs by 32.4 and 70.9 percent, respectively, because it can utilize part of the spare bandwidth. **Fig. 8d** shows that compared to ElasticSwitch, Trinity further reduces the shuffle completion time of FileStore by 5.4 percent (159s to 150s) by better utilizing the spare bandwidth on the bottleneck link. Note that 64 percent of
the link capacity of the bottleneck link is reserved as the guaranteed bandwidth in this experiment. Therefore, we can only observe a relatively small improvement of shuffle completion time under Trinity compared to ElasticSwitch.

Latency-Sensitive Applications. To evaluate how Trinity improves the performance of latency-sensitive applications, we run a memcached tenant: 8 instances hosted on servers $S_1$-$S_8$ and 8 clients hosted on servers $S_9$-$S_{16}$. We pre-populate instances with 4B-key, 1 KB-value pairs. The client generates a GET query to all 8 instances and each instance responds with a 1KB value. A query is completed only when the client receives all the responses from instances. The base query completion time is around 730us in our testbed.

In this experiment, we create an adversarial UDP tenant which runs many long-lived UDP flows from $S_9$-$S_{16}$ to $S_1$-$S_8$ and compare Trinity with ElasticSwitch. We plot the CDF of query completion times in Fig. 9. From the figure, we find Trinity outperforms ElasticSwitch by 83 percent in terms of the query completion time at the 99th percentile. The results imply that Trinity can improve the performance of latency-sensitive applications by letting short query messages receive higher priority in the network.

5.4 CPU Overhead

In this part, we show that Trinity is scalable as it introduces low CPU overhead to servers.

CPU Overhead under Various Numbers of Active Flows. In this experiment, we have four VMs A1, A2, A3 and A4 hosted on servers $S_1$, $S_9$, $S_{10}$ and $S_{11}$, respectively. We create two scenarios to evaluate the CPU overhead of both Trinity sender module and Trinity receiver module under various numbers of active flows.

In the first scenario, VM A1 send traffic to VMs A2, A3 and A4 using various numbers of TCP connections. We measure the CPU utilization of VM A1 to study the CPU overhead introduced by Trinity sender module. In the second scenario, VMs A2, A3 and A4 send traffic to VM A1 using various numbers of TCP connections. We measure the CPU utilization of VM A1 to study the CPU overhead introduced by Trinity receiver module. For both scenarios, we record CPU utilization every 1 second using the `sysstat` [42] tool, and calculate the average utilization over 100 seconds.

For both scenarios, we also measure the CPU utilization of no protection scheme. Under no protection scheme, no flow classification, packet coloring or rate limiting is done at end-host hypervisors. So its CPU utilization is mainly the cost of transmitting network traffic, and can be viewed as the base utilization.

The average CPU utilization under both scenarios are plotted in Fig. 10a and 10b, respectively. From the results, we can make two observations. First, the average CPU utilization increases as we reduce the rate control interval of Trinity from 2 ms to 0.5 ms, but the increase is not significant. As CPU overhead of Trinity partly depends on the frequency of updating work conserving rate, it is no wonder that smaller rate control interval leads to higher CPU utilization. As shown in both figures, compared with no protection scheme, the additional CPU utilization introduced by Trinity modules is only about 5 percent when the control interval is as small as 0.5ms. This indicates that Trinity is capable to do fine-grained rate control with small additional CPU utilization.

Second, in both scenarios, the CPU utilization almost stays the same as we increase the number of active flows between each VM pair. This is mainly because total network load is bounded by the link capacity, and adding more
TCP connections does not generate more packets to be processed at end-hosts. This also indicates that performing flow classification in Trinity introduces little CPU overhead.

**CPU Overhead under Varying Network Load.** In this experiment, we build a small testbed consisting of 4 servers connected to a 10 Gbps Mellanox SN2000 switch. Each server is equipped with an Intel 82599EB 10 GbE NIC. We vary the link capacity by configuring per port hardware traffic shaper at the switch. There are 4 VMs A1-A4, one on each server. Similar to last experiment, we use two scenarios to evaluate the CPU overhead of Trinity modules under varying network load.

In the first scenario, VM A1 works as a source to send traffic to VMs A2-A4. In the second scenario, VM A1 works as a destination to receive traffic from VMs A2-A4. Let L be the configured link capacity. For both scenarios, each VM pair has L/8 bandwidth guarantee. We vary the value of L from 1Gbps to 10Gbps, and measure the corresponding average CPU utilization of VM A1, as shown in Fig. 11 (We use no protection scheme as the base).

From the results, we find that the extra CPU utilization introduced by Trinity increases as we increase the network load, but the increase is not significant. At 10 Gbps, Trinity introduces only ~8 percent extra CPU utilization at the sender module, and ~9 percent extra CPU utilization at the receiver module. Note that our current implementation is single-core based, which can hardly scale up to higher link speed, e.g., 40 Gbps or 100 Gbps. The implementation of a multi-core aware Trinity, which can balance the network load across multiple CPU cores, is left as our future work.

**5.5 Convergence Time**

In this part, we show that under Trinity, rates of VMs can quickly adapt to varying network condition and converge within a short time (i.e., a few milliseconds).

To evaluate the convergence time of Trinity, in this experiment, we let four VMs A1, A2, B1 and B2, belonging to tenants A and B, compete a bottleneck link. Both tenants are provisioned with 200 Mbps bandwidth guarantees. And VMs A1, A2, B1 and B2 are hosted by servers S1, S2, S3 and S4, respectively. Note that we set the rate control interval as 1ms in this experiment.

Initially, only VM A1 are sending TCP traffic to VM A2. Some time later, tenant B becomes active, and VM B1 starts to send unbounded UDP traffic to VM B2.

As shown in Fig. 12, we sample the sending rate of VM A1 and VM B1 every 5 ms, and observe how long it takes for both VMs to get converged. As we can see, when VM B1 starts to send UDP traffic to VM B2, rates of both VMs converge to about 500 Mbps within only 5ms. The results indicate that, by adopting ECN-based rate control, Trinity can converge within a short time under mixed traffic load and varying network condition.

**6 CONCLUSION**

This paper presented Trinity, a simple yet effective solution that provides triple properties: bandwidth guarantees, work conservation and low latency simultaneously in the cloud. By differentiating traffic at the end and enforcing prioritization in the network, Trinity eliminates the tradeoff between providing bandwidth guarantees and being work-conserving, while achieving low latency for short flows. We have implemented Trinity using commodity switches and servers, and demonstrated its performance with testbed experiments.

**ACKNOWLEDGMENTS**

This work is supported by the Hong Kong RGC 26200014, 16203715, 613113, CRF C7036-15G and the National Basic Research Program of China (973) under Grant No.2014CB340303.
References


Wei Bai received the BE degree in information security from Shanghai Jiao Tong University, in 2013 and the PhD degree from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology in 2017. He is an associate researcher with Microsoft Research Asia. He was also a recipient of Microsoft Research Asia Fellowship (2015). He is broadly interested in computer networking with a special focus on data center networking. His research work has been published in many top conferences and journals, such as SIGCOMM, NSDI, CoNext and the IEEE/ACM Transactions on Networking.

Kai Chen received the PhD degree in computer science from Northwestern University, Evanston, Illinois, in 2012. He is an associate professor with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong. His research interest includes networked systems design and implementation, data center networks, and cloud computing.

Chen Tian received the BS, MS, and PhD degrees from the Department of Electronics and Information Engineering from Huazhong University of Science and Technology, China, in 2000, 2003, and 2008. He is an associate professor with State Key Laboratory for Novel Software Technology, Nanjing University, China. He was previously an associate professor with the School of Electronics Information and Communications, Huazhong University of Science and Technology, China. From 2012 to 2013, he was a postdoctoral researcher with the Department of Computer Science, Yale University. His research interests include data center networks, network function virtualization, distributed systems, Internet streaming, and urban computing.

Ying Zhang received the PhD degree from the EECS Department, University of Michigan. She has been a software engineer with the Facebook network infrastructure team since 2017. Prior to that, she is a senior research scientist with the Networking and Mobility Lab, Hewlett Packard Labs from 2014. Prior to that, she is a senior researcher with Ericsson Research Silicon Valley Lab from 2009. Her research interest is networking and systems, including network function virtualization, software-defined networking, cloud and data centers, internet routing and measurement, and network security.

Haitao Wu received the bachelor’s degree in telecomm. engineering and the PhD degree in telecomm. and information systems from Beijing University of Post and Telecommunications (BUPT), in 1998 and 2003 respectively. He joined the Wireless and Networking Group, Microsoft Research Asia (MSRA), in 2003. He was transferred to Microsoft Azure product group on data center networking in 2014. He joined Google Network Infrastructure team in 2017. His research interests include data center networks, QoS, TCP/IP, P2P, and wireless networks. He was a member of the IEEE.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.