TCPTokens: Introducing Currency into Datacenter Congestion Control

Anand Jayarajan
University of British Columbia
anandj@cs.ubc.ca

Michael Przystupa
University of British Columbia
michael.przystupa@gmail.com

Robert Reiss
University of British Columbia
rreiss@cs.ubc.ca

Fabian Ruffy
University of British Columbia
fruffy@cs.ubc.ca

ABSTRACT

Datacenter networks have become a hotbed for research in the recent past. Many works have been published on improving bisection bandwidth, cost, and latency. As a core contribution of this research, centralized scheduling has entered the stage as a dominant strategy to manage flows. Benefitting from a global view and full control over the network, centralized schedulers are able to enforce fine-grained traffic control, frequently achieving near-optimal bandwidth optimization. However, modern congestion control algorithms and schedulers are still fundamentally reactive. Many algorithms are designed to only respond to packet loss or significant increase in latency, not to actively prevent it. In this project, we plan to explore the opportunities of employing a preemptive and tightly controlling central network scheduler. Using tokens as a global tool to enforce traffic limits and admission control, the scheduler is able to proactively steer the flow of traffic. The notion of "knowledge" and predictive analysis in networks is a growing trend in research, which we intend to leverage in this system.

We investigate the potential in such a new form of interactive congestion control and analyze it against state-of-the-art solutions. Our tool of choice to model our system is Mininet, a rapid prototyping emulator for datacenter networks. We will compare our "Iroko" system against established TCP-congestion systems such as Hedera and DCTCP. Based on the findings and measurement results, we will reassess the feasibility and prospects of a predictive congestion control algorithm.

KEYWORDS

TCP, Congestion Control, SDN, Datacenter

1 INTRODUCTION / BACKGROUND

Initially, research in network congestion was dominated by the assumption of a decentralized, autonomous network - end-hosts only have control over the amount of traffic they send, and are unaware of the intentions or traffic rates of their peers. Similarly, switches and routers are unaware of global traffic patterns and only forward based on their local notion of optimality.

In line with these assumptions, TCP was designed to optimize traffic globally on a simple fairness principle: nodes react to packet loss and assume that others will behave accordingly. An equilibrium is reached when all end-hosts achieve a traffic rate that, in sum, conforms to the maximum available bandwidth of the most congested link. TCP works well in scenarios where many different distrustful participants compete for limited bandwidth. Still, TCP is a *reactive protocol*; the fact that packet loss and latency increases occur in the network already indicates a problem. Packet loss is primarily due to overflowing queues in forwarding elements or mismatched hardware capabilities, implying that traffic has not been optimally distributed. Ideally, a network should always be "zero-queue", i.e., latency will merely be induced by propagation, and not queuing delay.

Queueing has commonly not been a dominating issue in wide-area and enterprise networks, as traffic is sufficiently distributed and diverse, with only few "hot" target hosts. [1, 4] Traffic optimization is a substantial challenge; network operators have no control over the individual network elements nor its participants. Under these conditions, TCP and its extensions can be considered a best-effort solution to globally maximize utilization.

Developments in the past decade have changed the general networking environment. Datacenters have emerged as an exciting new research frontier, posing novel design challenges and opportunities. Driven by minimization of costs and maximization of compute power, data centers must run at maximum utilization to achieve an optimal compute/cost ratio. Inefficient routing can quickly lead to bufferbloat [7] and the eventual collapse of a high-load network, requiring more sophisticated approaches to solve congestion control.

On the other hand, operators now have the ability to freely control and adapt their network architecture, leading to highly customized systems and fine-grained optimization. In such an environment, Software-Defined Networking (SDN) has emerged as a new networking paradigm. [6]

Moving away from the principle of distributed communication and routing, SDN introduces the notion of "centralized management". A single controller with global knowledge is able to automatically modify and adapt the forwarding tables of all switches in the network, while notifying end hosts of changes in the network. These two new trends in system design; full architectural control and centralized management, facilitated new opportunities in the space of TCP congestion research. Traffic can now be managed in a *centralized* fashion based on *global knowledge* of the entire topology and traffic patterns.

A new line of centralized schedulers has emerged that can achieve close to optimal bandwidth utilization. [1, 4, 8, 9, 12] However, these schedulers are still *reactive* in nature. The central controller responds to changes in the network or requests by applications,

which may cost valuable round-trip latency. Often, short-term flows or bursts are unaccounted for, which causes undesirable packet loss and back-propagating congestion.

The idea of admission control and service guarantees in networks is not new. [11, 13]. However, such designs traditionally aim to assure quality and bandwidth guarantees in a contentious, decentralized, and untrusted environments such as the internet. In a datacenter, these conditions do not apply. End-hosts are generally considered reliable and restricted in behavior, which allows for great simplification of enforcement and prioritization policies.

A desirable solution is a global, centralized arbiter which is able to predict and fairly distribute flows in the network before bursts or congestion even occur. By treating the network's compute and forwarding power as a single finite resource, a controller acts akin to a OS scheduler distributing CPU time slices to processes. This design approach follows SDN's aspiration of introducing operating systems abstractions to the networking domain space.

In this project, we plan to explore the possibilities of a centralized, proactive flow scheduler. We ask ourselves the following research questions:

(1) What are the requirements for such a centralized, predictive scheduler to succeed?
(2) Is it possible to preemptively regulate a network by analyzing global traffic patterns?
(3) What latency, packet loss and utilization are we able to achieve?

In the scope of this course, we attempt to answer these questions and design a simple predictive scheduler in Mininet. If successful, we will benchmark our results and evaluate the level of utilization compared to contemporary scheduling systems.

2 DESIGN / PROPOSED APPROACH

2.1 Overview and Goals

In our initial simple system, "Iroko", a centralized controller regulates all node traffic by rate-limiting end-hosts. We have opted for a centralized manager in favor of a distributed protocol to leverage the advantages of a global network view. [1] By polling each switch individually for port and utilization statistics we are able to infer a global traffic matrix, which we can amalgamate with static routing and topology information.

Iroko guarantees minimal congestion and low average access latency. Reducing network-global packet loss and jitter is the priority and objective function of the arbiter, which will enforce these goals by restricting host bandwidth. In an ideal Iroko system, packet-loss will only rarely, if ever, occur. In respect to the free lunch theorem we aim to trade off maximum bandwidth utilization for optimal system stability and reliable latency.

It is important to note that routing decisions are not in the current scope of Iroko. While it may certainly beneficial to include dynamic and adaptive routing decisions in a predictive scheduler, we do not include these features in the current system due to time and complexity limitations. Iroko operates on static flow routes defined and generated by ECMP. It is assumed that these routes will not vary substantially and that the computed ECMP forwarding hash is consistent.

[1] For now, we do not expect our design to scale up to thousands of switches. Iroko is intended for small to mid-tier size data centers, which may benefit from a simplistic, centralized scheduling model.

2.2 Controlling traffic flow

Rate-limiting is performed on the granularity of the IP protocol, allowing us to restrict traffic to specific hot regions and to package flows into groups. End-hosts are guaranteed a limited amount of bandwidth which they may send to each destination address. The traffic window size for any flow is calculated based on the host's current traffic availability. Total bandwidth of a group of flows may never exceed the imposed bandwidth limitation, but nodes are free to distribute the allocated resources on singular subflows.

The central arbiter enforces these bandwidth and access limitations by assigning "tokens" to nodes. These tokens are plain information packets specifying the affected destination address flow, its bandwidth guarantee, and the expiration time. Tokens act as the rate-limiter of the system and form a queue each end-host will cycle through.

When a node opens a flow, it will look up the current bandwidth restriction for the destination IP in its token database and calculate the maximum congestion window possible for the particular TCP connection. In our simple design, flows are always assigned a fair fraction.

Once a token has expired for a particular flow-group, the restrictions of the next token in the queue will be applied. This mechanism is the underlying basis for a predictive access control algorithm and does not require any application level modification. On end-hosts, only the transport layer services will have to be modified.

2.3 Adapting traffic flow

Initially, the controller will compute optimal route configuration based on the topology and link bandwidth using a simple heuristic bin-packing approach. End-hosts will be initialized with a fixed low-to-medium bandwidth guarantee that will be adjusted over time. This bandwidth guarantee will be below the host's proportional bi-section bandwidth that would be used to reach any other host. That is, if every end host transfers using the full bandwidth initially allocated to it, there would be no dropped packets due to congestion.

Of course, this would vastly underutilize network resources, thus the controller will dynamically adjust these allocations to ensure that hosts that require the bandwidth are allocated it, while host that are not currently utilizing all their allocated bandwidth have their allocated bandwidth reduced. To avoid starving hosts, some small amount of bandwidth must always be allocated to a given host even if it not utilizing network resources. Thus, the network, by design, will never reach 100% utilization.

Conversely however one should rarely see dropped packets due to network congestion. Furthermore, the controller must be able to react quickly to changing bandwidth needs for hosts. Ideally, using a predictive method.

2.4 Statistical System

One potential implementation for our arbiter is to use reinforcement learning. In reinforcement learning, an agent (in our case the arbiter) performs actions in the environment and receives a reward signal from the environment to adjust its decisions in the next epoch [14].

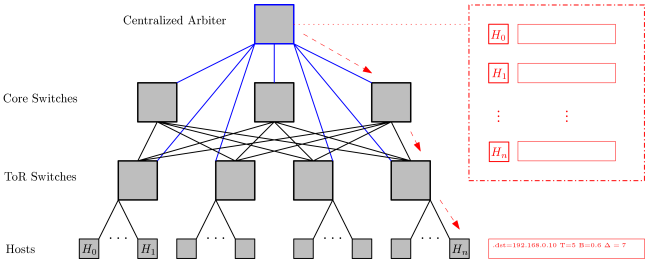

Figure 1: Simple sketch of the Iroko architecture. A central scheduler maintains a consistent view of the current network activity and computes the optimal end-host bandwidth distribution. Bandwidth is enforced at end-hosts on a per-IP basis.

The reward is generally used to represent the goal the agent is trying to achieve. The agent can freely choose whatever actions are necessary to achieve this goal [14].

The environment in this case is defined as the statistics collected from each switch. This decision is based on the fact that this data is easy to collect and provides a current snapshot of the performance of the network. We wish to optimize over the packet loss, and define the reward to be good if packet loss is decreasing (+1 reward); bad if it is increasing (-1 reward); and acceptable if packet loss remains the same within a threshold (0 reward). Using this representation, we take full advantage of all the information provided in our data center; This representation is replicable and can be applied beyond the scope of this project.

Given the data center assumption, the number of end-hosts is known in the topology. We can define our arbiter's actions as a discretized representation among an n-dimensional vector where each cell is one of three values: increase decrease, or maintain the allocated bandwidth. This representation offers sufficient complexity. There are 3^n possible actions to perform and to achieve optimal results the agent must explore the environment extensively, precluding a significantly larger action space.

A potential improvement to our representation is a hierarchical learning agent. In this case, our arbiter would manage agents for each host and these sub-agents would manage their respective host whether to increase, decrease, or maintain bandwidth for their particular host. This problem can be referred to as hierarchical reinforcement learning [5], but may prove to be beyond the scope of our current proposed solution, although has the added benefit of solving a set of sub-problems to address the global problem of packet loss and simplifies the representation spaces in the sub-problems.

Thus we frame our problem as one which can now be solved using techniques in reinforcement learning. A classic algorithm to consider is the use of SARSA [14], which is an online temporal difference algorithm used in classical reinforcement learning problems. One advantage of SARSA is its online nature [14] which promotes more conservative decisions. Given full representation of our environment and actions, we can store information in an artificial neural network which is another popular choice in reinforcement learning literature. This model would store the reward signal information that is used to make future action choices, and is a theoretically proven function approximator which has made intractable environmental representations manageable [14].

3 IMPLEMENTATION

We intend to emulate our system in Mininet [10] to observe traffic patterns and infer a suitable token algorithm. Mininet has proven itself to be a viable tool to model new congestion control algorithms [15], and will help us prototype our concept efficiently. We will build a custom SDN controller that interacts with traditional OpenFlow software switches as well as end-hosts. End-hosts will run a custom real-world traffic generation script which adjusts based on information packets sent by the controller. If time permits, we may expand our implementation to MaxiNet, which can emulate large-scale network stress tests on multiple physical hosts.

We initially considered a second implementation alternative in C/C++ based on the FastPass [12] source code. This would provide us with a fully deployable system which we could fork our implementation from. A major advantage of this approach is the ability to test scenarios and traffic algorithms using real software code. However, several concerns made us favour a Mininet emulation instead. Firstly, FastPass relies on DPDK integration, which requires actual hardware interfaces. The central arbiter in the FastPass design would need to run on a dedicated machine, which increases prototyping and development complexity substantially.

Secondly, the FastPass code is highly specialized and optimized research code with only little available documentation. Modifying and evolving the source code will require thorough understanding of kernel and networking development, a significant time-sink. For a class project, these may be major initial hurdles, taking away from the research aspect of the design concept. Consequently, we have decided to pursue an approach which allows us to quickly develop an understanding of the problem without being obstructed by engineering work.

4 EVALUATION

In the absence of a commercial data center, the implementation of our network design is going to be done on top of mininet with simulated FatTree topologies of varying size. We stress test with iPerf and simulate data center traffic using tcpreplay and packet traces provided by [3] .

To evaluate the general effectiveness of our system, we plan to measure against existing centralized as well as decentralized solutions. The centralized design will be based on Hedera [1], a common and influential datacenter scheduler. The decentralized congestion control mechanism will be DCTCP [2], a state-of-the-art TCP congestion algorithm.

Since we are primarily concerned about reducing the latency and packet drops while keeping utilisation at maximum, the measurements to get a good insight into how well the design can perform are as follows:

(1) Latency: We are aiming for a low latency network which means that 99th percentile latency in the network across all flows should as low as possible. Latency should be measured when all hosts are sending packets at the maximum limit and also with random traffic patterns. There should be

low latency even during sudden bursts as the transmission rate is limited by a base value.

(2) Packet drop rate: Ideally this metric will approach zero, as the objective of Iroko is to minimize packet loss.

(3) Fairness: Fairness can be tested by introducing a new host to a completely saturated network and increasing the transmission rate to see if all the hosts gets fair share of the total bandwidth. We are assuming all flows should be at equal priority. Differential priority is out of scope.

(4) Responsiveness: This metric depends on the predictive power and efficiency of Iroko. It may be interesting to analyze the response time of the central scheduler compared to decentralized DCTCP and TCP. In addition, it is valuable to identify at what data center and flow size the central computing node of the network may become a bottleneck.

(5) Network utilization: We will measure the used bandwidth and the total load on the hosts. These metrics can be measured by varying the load from each host. Load to utilization ratio should be 1 until the saturation point and after that utilization should stay stable at the maximum bisection bandwidth. It is also important to check that there is no starvation happening in any host. This can be measured by simulating random traffic patterns and plotting residual bandwidth and excess load. We expect overall utilization to be lower compared to Hedera and DCTCP. Although Iroko will maximize goodput, we do not measure it on grounds of simplicity.

As baseline we will also compare to the default TCP congestion algorithm to estimate how our scheduler fares against the default case. As our initial simple design is very constrained in its ability to route and control traffic we expect to achieve less overall optimality against DCTCP and Hedera, but still gain a substantial advantage over common TCP.

5 TIMELINE

- October 15th: Finalize draft.
- October 22nd: Iterate over literature and thoroughly analyze related work. Set up an initial prototyping environment in Mininet on which each team member can work on.
- October 30th: Finalize the initial simple design and explore naive congestion modelling in Mininet. Integrate existing Mininet emulations of Hedera and DCTCP.
- November 12th: Develop initial naive Iroko prototype involving controller and end-host design.
- November 19th: Develop intelligent centralized congestion control heuristic.
- November 26th: Run evaluations and microbenchmarks.
- December 3rd: Iterate and improve the scheduler, explore topology and bandwidth constraints. Test at scale.
- December 10th: Write report and prepare presentation.

REFERENCES

[1] Mohammad Al-Fares, Sivasankar Radhakrishnan, Barath Raghavan, Nelson Huang, and Amin Vahdat. 2010. Hedera: Dynamic Flow Scheduling for Data Center Networks.. In *NSDI*, Vol. 10. 19–19.

[2] Mohammad Alizadeh, Albert Greenberg, David A Maltz, Jitendra Padhye, Parveen Patel, Balaji Prabhakar, Sudipta Sengupta, and Murari Sridharan. 2010. Data center tcp (dctcp). In *ACM SIGCOMM computer communication review*, Vol. 40. ACM, 63–74.

[3] Theophilus Benson, Aditya Akella, and David A. Maltz. 2010. Network Traffic Characteristics of Data Centers in the Wild. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC '10)*. ACM, New York, NY, USA, 267–280. https://doi.org/10.1145/1879141.1879175

[4] Theophilus Benson, Ashok Anand, Aditya Akella, and Ming Zhang. 2011. MicroTE: Fine grained traffic engineering for data centers. In *Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies*. ACM, 8.

[5] Magnus Borga. 1993. *Hierarchical Reinforcement Learning*. Springer London, London, 513–513. https://doi.org/10.1007/978-1-4471-2063-6_139

[6] Nick Feamster, Jennifer Rexford, and Ellen Zegura. 2013. The road to SDN. *Queue* 11, 12 (2013), 20.

[7] Jim Gettys and Kathleen Nichols. 2011. Bufferbloat: Dark buffers in the internet. *Queue* 9, 11 (2011), 40.

[8] Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, et al. 2013. B4: Experience with a globally-deployed software defined WAN. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 3–14.

[9] Xin Jin, Hongqiang Harry Liu, Rohan Gandhi, Srikanth Kandula, Ratul Mahajan, Ming Zhang, Jennifer Rexford, and Roger Wattenhofer. 2014. Dynamic scheduling of network updates. In *ACM SIGCOMM Computer Communication Review*, Vol. 44. ACM, 539–550.

[10] Bob Lantz, Brandon Heller, and Nick McKeown. 2010. A network in a laptop: rapid prototyping for software-defined networks. In *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*. ACM, 19.

[11] Sridhar Machiraju, Mukund Seshadri, and Ion Stoica. 2002. A scalable and robust solution for bandwidth allocation. In *Quality of Service, 2002. Tenth IEEE International Workshop on*. IEEE, 148–157.

[12] Jonathan Perry, Amy Ousterhout, Hari Balakrishnan, Devavrat Shah, and Hans Fugal. 2015. Fastpass: A centralized zero-queue datacenter network. *ACM SIGCOMM Computer Communication Review* 44, 4 (2015), 307–318.

[13] Ion Stoica and Hui Zhang. 1999. *Providing guaranteed services without per flow management*. Vol. 29. ACM.

[14] Richard S. Sutton and Andrew G. Barto. 1998. *Introduction to Reinforcement Learning* (1st ed.). MIT Press, Cambridge, MA, USA.

[15] Lisa Yan and Nick McKeown. 2017. Learning Networking by Reproducing Research Results. *ACM SIGCOMM Computer Communication Review* 47, 2 (2017), 19–26.