Simulation Modeling of Zoom Network Traffic

Mehdi Karamollahi  Carey Williamson
University of Calgary
Calgary, AB, Canada
{mehdi.karamollahi,cwill}@ucalgary.ca

Abstract—In this paper, we develop a synthetic workload model for the Zoom network application based on empirical Zoom traffic measurements on a campus network. We then use this model in a simulation study of Zoom network traffic at the campus scale. The simulation results show that hybrid learning places a substantial load on the campus network, especially on the wireless network. Additional simulation experiments investigate the potential benefits of locally-hosted Zoom infrastructure, improved load balancing strategies for Zoom servers, and multicast delivery models for Zoom network traffic. The results show that the multicast approach offers the greatest potential benefit.

Index Terms—Zoom videoconferencing, traffic measurement, workload modeling, network simulation, capacity planning

I. INTRODUCTION

Zoom is a highly popular network application for remote teaching and learning. It is one of several videoconferencing applications that have gained prominence during the COVID-19 pandemic, with Meet, Teams, and Webex being others [2].

Our university adopted Zoom as its solution of choice for remote work and learning during the pandemic. At our institution, remote teaching started in March 2020, and continued for about two years. A small number of students (about 10%) were allowed back on campus in Fall 2020, with in-person learning offered for some small upper-year courses, while all large courses remained online. In Fall 2021, about 50% of students transitioned back to in-person learning on campus, though many larger classes remained online.

Zoom worked quite well at our university during the first year of the pandemic. However, we encountered several network performance problems during the second year of the pandemic, with a hybrid mix of in-person and online offerings. The underlying reason was the higher load on the campus network: many students on campus for in-person courses were also using Zoom for their online courses, which might take place just before or just after their in-person classes, making it difficult to commute home for these.

This hybrid of learning modalities created a “perfect storm” that exacerbated Zoom performance problems on the campus network. There are several reasons for this. First, there were lots of students on campus, with many using Zoom for their (large) online classes. Second, many of these students were on the wireless network, using mobile devices like laptops and smartphones. Third, most classes used the campus-licensed version of Zoom, which generates higher video bit rates than the free version (e.g., 3 Mbps vs 1 Mbps). Fourth, Zoom is a bandwidth-intensive application that does not always share the network fairly with other applications [11], [17]. Furthermore, Zoom triggers dynamic FEC (Forward Error Correction) when packet losses and delays occur [11], leading to extra load on congested networks. Last but not least, our campus Zoom setup was configured to prefer using regional Zoom servers in Canada, rather than the larger pool of Zoom servers in the US and around the world. This configuration setting produces fairly high loads on these Zoom servers [5].

The goals of this paper are to provide a better understanding of Zoom performance on a campus network, and to explore possible solutions for improving Zoom performance. Our prior works have provided a macro-level view of overall campus network traffic during the pandemic [8], as well as a micro-level analysis of Zoom and its performance problems [3]. While these prior works have proposed several potential improvements to Zoom, the performance benefits of these solutions have not been evaluated quantitatively.

In this paper, we develop a synthetic workload model for Zoom, and use it in simulation studies investigating several different Zoom scenarios on our campus network. Specifically, we consider the potential benefits of locally-hosted Zoom servers, improved load balancing strategies for Zoom servers, and the use of multicast for Zoom content delivery.

The main contributions of our paper are: (1) the design and implementation of a synthetic workload model for campus-level Zoom traffic; (2) a simulation study exploring the sensitivity of Zoom performance to different workload parameters and configuration settings; and (3) practical recommendations to improve Zoom performance on a campus network.

The main insights that emerge from our work are:

- Zoom traffic has sharp spikes in its meeting and session arrivals based on course scheduling, making load balancing a challenge;
- hybrid learning modalities can overload the campus wireless network when too many students are using Zoom;
- improved load balancing strategies can spread load more robustly across Zoom servers; and
- multicast is a promising solution for Zoom delivery.

The remainder of this paper is organized as follows. Section [II] provides background information on Zoom and discusses prior related work. Section [III] introduces our empirical dataset. Section [IV] describes our synthetic workload model and its validation. Section [V] presents simulation results investigating different Zoom configuration scenarios, Zoom server load balancing, and multicast delivery models. Finally, Section [VI] concludes the paper.
II. BACKGROUND AND RELATED WORK

A. Zoom

Zoom is a videoconferencing network application that has been widely used for remote work and learning since the COVID-19 pandemic emerged. Zoom meetings with only two participants operate in direct peer-to-peer mode [3]. Larger meetings operate in client-server mode, using a Zoom Multi-Media Router (MMR) server in the cloud (e.g., AWS).

A Zoom meeting consists of multiple Zoom sessions, with one session for each participating user. A Zoom session involves one TCP connection for the control and management of the session, and three UDP connections: one for transmitting audio, one for video, and one for screensharing data. The Zoom application is highly resilient, and can adapt video bitrates to the available bandwidth on a network. It can also dynamically restore sessions that experience failure at the UDP or TCP connection level [3].

A free Zoom account has some limitations, such as a 40-minute limit on meetings, and limited video bitrates. A license is required to remove the limitations. Therefore, many enterprises purchase Zoom licenses for their members to use Zoom without limitations.

There are multiple ways to deploy and use Zoom within enterprise networks. The first and most straightforward is using Zoom cloud servers to manage meetings. Those servers are maintained and administered by Zoom or Amazon on their data centers. Another is the on-premise solution for organizations to host a Zoom Meeting Zone that contains servers to manage Zoom meetings. In this paper, we explore the performance differences between these deployments for a campus network.

B. Related Work

The impacts of the COVID-19 pandemic on the Internet were significant. Many research works have tried to identify, measure, and characterize the changes caused by sudden global lockdowns in various contexts [1, 2, 5, 6, 7, 8, 10].

Initial reports showed a substantial increase in Internet traffic, with weekday peaks 45% higher than pre-lockdown levels, and weekend/evening peaks 20%-40% higher [9]. Despite the rise in traffic and service demand, the Internet showed high resilience and adaptability with limited reports of problems [7].

The changes, however, varied across network types. ISPs and IXPs experienced larger downstream increases, since more people accessed the Internet from their residential networks [5]. Mobile networks saw reduced mobility, specifically in crowded municipal areas [10]. Educational organizations such as university campus networks, on the other hand, experienced decreases in downstream traffic, since fewer people remained on campus after the lockdowns. In contrast, these networks experienced growth in upstream traffic, from remote users accessing internal infrastructure [5, 8].

Several studies have reported the rise of videoconferencing applications and online learning platforms [15, 16]. MacMillan et al. [11] compared Zoom, Meet, and Teams using an experimental testbed. They investigated how these applications perform under different network conditions and showed that the answers heavily rely on the application. For example, Zoom can consume more than 75% of the available bandwidth when competing with Meet and Teams.

Chang et al. [2] developed a cloud-based framework to evaluate Zoom, Webex, and Meet in terms of QoE metrics. They used 48 hours of emulated videoconferencing sessions, and compared the applications based on geographically-distributed clients, media bitrates, and other evaluation metrics.

Our prior work [8] took a longitudinal view of Zoom, Teams, and Meet as three prominent videoconferencing applications. We provided evidence of issues with Zoom TCP connections and session management on our campus network when the network is congested, and too many users share a relatively small subset of Zoom servers during peak hours.

In [5], we developed tools to analyze Zoom connections to identify Zoom sessions and meetings, providing an in-depth analysis of Zoom traffic. We showed that this traffic could stress the campus network due to many concurrent meetings, temporally correlated arrivals, high video bit rates, and long-lasting sessions. Observed anomalies included congested WiFi networks, excessive FEC traffic, disrupted TCP connections, and sluggish TLS handshakes at Zoom servers. A key insight is that unstable WiFi on a home network disrupts only a single user, while on a campus network it disrupts hundreds of users, all of whom need to reconnect to Zoom at the same time.

In this paper, we introduce strategies that could mitigate these issues by reducing the load on the client-side network as well as the Zoom servers. We discuss how multicast could be a promising solution to ameliorate Zoom performance issues on our campus network.

III. EMPIRICAL MEASUREMENT DATA

In this paper, we study Zoom usage on the University of Calgary network, as an example of a typical campus edge network. Our network is used by 32,000 undergraduate/graduate students and 3,000 faculty/staff.

Our data was collected from a mirrored stream of all Internet traffic passing through the edge routers on our campus network. This traffic was processed into connection-level summaries using Zeek (formerly known as Bro [14]). Each connection summary represents communication between a sending host (originator) and a receiving host (responder), where one of the hosts is on the campus network, and the other elsewhere on the Internet. Relevant fields from the connection-level data were then loaded into Vertica [2] which is a big data analytics platform that we used for traffic analysis.

We have over two years of Zoom traffic data that we also used in our prior works [3, 8], but for the purposes of this paper we focus on a small subset of the data from the Fall

1 Even though UDP is a connectionless protocol, we use the term connection to refer to the sustained bidirectional flow of streaming media traffic on each of the three UDP ports used by clients.

2 https://www.vertica.com/
2021 semester. Figure 1 illustrates Zoom traffic during the week of September 19-25, 2021. Hybrid teaching and learning started in September 2021, and we picked this week to ensure that all online classes were well underway. Figure 1(a) shows the hourly count for TCP and UDP connections to Zoom for the chosen week, while Figure 1(b) shows the corresponding average hourly data rate in Gbps. The diurnal pattern on working days is evident in these graphs, with two major peaks, one in the morning and the other in the early afternoon, and a small peak in the late evening, especially on the data rate plot. The peak bandwidth consumption for Zoom is almost 1 Gbps, which places a substantial load on our (already busy) 6 Gbps external link for commercial Internet traffic. These plots show that Zoom usage is quite consistent on weekdays, and we subsequently chose the Wednesday of this week for detailed fine-grain analysis. We tested our methodology on other days to ensure that the chosen day (Sept 22, 2021) is representative of the daily Zoom traffic on our network.

Figure 2 shows three important characteristics of empirically-observed Zoom network traffic. First, Zoom usage has a diurnal pattern, with distinct spikes in Zoom meetings initiated on an hourly basis due to the course scheduling for lectures. Figure 2(a) shows these pronounced peaks, along with smaller secondary peaks that occur at half-hour intervals for other meetings. Second, Zoom meetings have widely varying durations, ranging from a few minutes to a few hours. Figure 2(b) shows the empirical distribution, with a median duration near one hour. Third, Zoom meetings have wide-ranging numbers of participants. Most meetings are small, with only a few participants, but meetings with several hundred participants also occur. Figure 2(c) shows the empirical distribution, which exhibits a power-law structure.

IV. WORKLOAD MODELING

This section discusses our approach to modeling Zoom traffic based on our empirical data. We focus on modeling a single day of Zoom traffic, and build a synthetic workload model for this purpose. Our model is implemented with about 200 lines of C code. Table I summarizes the key parameters in the model, which we explain and justify next.

<table>
<thead>
<tr>
<th>Item</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Zoom meetings</td>
<td>3500</td>
</tr>
<tr>
<td>Meeting arrival process</td>
<td>Poisson, time-varying rate</td>
</tr>
<tr>
<td>P2P Zoom probability</td>
<td>0.2</td>
</tr>
<tr>
<td>Meeting participants</td>
<td>2 + Geometric(10)</td>
</tr>
<tr>
<td>Meeting duration</td>
<td>Exponential(1800)</td>
</tr>
<tr>
<td>Lecture duration</td>
<td>3600 + Normal(0,300)</td>
</tr>
<tr>
<td>Session join latency</td>
<td>Normal(0,120)</td>
</tr>
<tr>
<td>Prob(lecture)</td>
<td>0.9</td>
</tr>
<tr>
<td>Prob(home)</td>
<td>0.5</td>
</tr>
<tr>
<td>Prob(WiFi)</td>
<td>0.8</td>
</tr>
<tr>
<td>Server load balancing</td>
<td>RAND</td>
</tr>
</tbody>
</table>

Our synthetic workload model starts at the meeting level, which we subsequently extend to the session level and the
connection level. At the meeting level, we use a Poisson arrival process, but with time-varying rates. The background level of Zoom meetings is very low in the early morning hours (midnight to 8:00am), higher during the working day (8:00am to 5:00pm), and low in the evening (6:00pm to midnight). On top of this background load, we induce impulses in the meeting arrival process during the 10-minute interval prior to each hourly lecture slot. We also add softer impulses at the half-hour intervals.

Figure 3 shows the arrival count process for the synthetic Zoom traffic generated using our model. This traffic pattern is structurally similar to the empirical traffic shown in Figure 2(a). Most notably, the overall meeting counts are similar, and the synthetic workload captures the diurnal patterns, including the primary and secondary spikes.

Additional details of the workload model are as follows. Meeting durations are generated conditionally based on whether the meeting is for a lecture class or not. Lectures have durations near one hour, while other meetings have durations drawn from an exponential distribution. Session arrivals to a meeting are staggered randomly using a Gaussian distribution, to reflect students arriving early or late for class (as observed in the empirical data). About 20% of the meetings are peer-to-peer meetings with exactly two participants. All other meetings have a number of participants drawn from a shifted geometric distribution. We have parameters to represent the probability of a Zoom user being on their home residential network rather than on campus, and a parameter to represent the probability of being on WiFi when on the campus network.

In terms of model validation, one important property is the number of concurrent Zoom meetings that occur, since this reflects the interactions between the arrival process, meeting duration, number of participants, and network resource usage. Figure 4 provides a comparison between the synthetic and empirical traffic, showing that the primary structure and the overall volume of Zoom activity are well represented. The main limitation in the model is the inability of the exponential distribution to model heavy-tailed meeting durations. This weakness is most evident in the evening traffic, since we do not explicitly model the longer lectures in evening classes, nor lengthy personal Zoom calls that take place in off-peak hours. Nonetheless, we find the workload model suitable for our simulation purposes.

V. SIMULATION RESULTS

This section describes the simulation experiments conducted using our Zoom traffic model.

A. Simulation Scenario

We have built our own discrete-event simulation model for Zoom traffic studies. The simulator is also written in C, and involves about 700 lines of code, including routing, load balancing, and instrumentation.

The simulator operates at the flow level (i.e., Zoom meetings, sessions, or connections) and not the packet level. The primary events in the simulator are the arrivals and departures of Zoom meetings, each with randomly generated participants and durations. The simulation model tracks the location and type of each Zoom user, as well as the network resource usage for each user session during the Zoom meeting. We model a single day of Zoom traffic, with about 3500 Zoom meetings.

Figure 5 shows the network model used in our simulations. The three main entities are the campus network, the ISP network, and the Zoom cloud infrastructure. A router (R) connects the university network (U) to the Zoom network (Z) and the ISP network. Within the campus network, we distinguish between mobile users on the WiFi network, and desktop users on Ethernet LANs. We also represent possible P2P Zoom users using P1 and P2. Inside the ISP network, we model home (H) users (i.e., faculty, staff, or students) using Zoom, including possible P2P users P3 and P4. For the Zoom infrastructure, we represent the regional data centers in Toronto and Vancouver, which handle 90% of our campus Zoom traffic.

B. User Location

Our first simulation experiment is a one-factor experiment focusing on the sensitivity to the location of Zoom users. We use the home probability parameter for this purpose, which
determines the likelihood of a Zoom participant being on their home ISP network rather than on campus. The default setting for this parameter is 50%, but we consider settings with either more or fewer users on their home networks.

Figure 6 shows the results of this simulation experiment. The graph shows the number of concurrent Zoom meetings traversing the network link between U and R in Figure 5 with one line for each of our three parameter settings. In the default setting with 50% of the users at home, there are about 130 concurrent Zoom meetings during the busy part of the day. When 75% of the users are at home, this Zoom traffic drops by about half, since it routes directly between the ISP and the Zoom data centers. When only 25% of the users are at home, more Zoom traffic traverses the campus network. These results are as expected, and help explain why the Zoom performance problems on our campus network were much worse in Fall 2021 than in Fall 2020, since many more students were on campus in Fall 2021, and 85% of them were using WiFi.

C. Local Zoom Meeting Zone

Rather than relying on Zoom infrastructure in the cloud, one possible configuration option with Zoom is to deploy local Zoom infrastructure (i.e., controller, data center) on campus. This option is intuitively appealing, since it reduces the network round-trip time to the Zoom servers, and should reduce the volume of external Zoom traffic generated. However, it might be an expensive solution to deploy, so we use simulation to investigate its performance benefits.

Figure 7 presents the simulation results for this locally-hosted Zoom solution (shown as “Local” in Figure 5). The results show that the effectiveness of this solution is highly dependent on the location of Zoom users. In the middle graph (Figure 7(b)), which shows the default setting with 50% of the users at home, there is really no benefit. That is, since half of the users are on campus, and half are at home, the external Zoom traffic volume remains the same, regardless of where the Zoom servers are located.

When more users are on campus (see Figure 7(a)), the benefits of local Zoom hosting are clear. However, one could also argue that in-person teaching might make more sense in this case, rather than Zoom, so the cost-benefit tradeoff may not be worthwhile. Furthermore, when more users are at home (see Figure 7(c)), the traffic trends reverse, with even greater demand on the campus network infrastructure. In short, there does not seem to be much rationale for investing in locally-hosted Zoom infrastructure (a counter-intuitive result).

D. Load Balancing

Our next simulation experiment takes a closer look at load balancing strategies for Zoom servers, as suggested in prior work [3]. The crux of the issue here is that the selection of a Zoom MMR server must be made dynamically when a meeting is first initiated, before knowing how many participants it will have, or how long it will last. Furthermore, this decision is often made during the sharp impulses in the workloads, when load levels are rapidly changing. As such, it is not uncommon for Zoom servers to have very different loads, in terms of meetings or users. The latter scenario occurs often in our own campus Zoom usage, since preference is given to the limited pool of regional Zoom servers in Canada.

We consider three possible load balancing strategies. First, we consider the RAND policy, which chooses Zoom servers uniformly at random. We do not know what policy Zoom controllers actually use to select MMR servers, but the RAND policy serves as a simple baseline point to approximate their empirically-observed behaviours [3]. We then consider two load-aware policies: one based on the number of Zoom meetings (MTG) currently hosted, and one based on the number of Zoom participants (USERS) currently hosted. We believe that such policies should be implementable in practice, particularly for recurring Zoom meetings (the common case), which can use past history as a hint about participants and duration.

For this simulation, we focus solely on the Zoom traffic to the Vancouver data center (about 1400 meetings). For these experiments, we artificially constrain the number of Zoom
servers to $N = 10$ to highlight the key trends in our results. Note that each Zoom meeting is an atomic unit from a load balancing point of view; it is not possible to split a large Zoom meeting across multiple servers.

Figure 8 shows the results from this simulation experiment. Each column of graphs presents a different load balancing policy (RAND, MTG, USERS), while the rows represent two different load balancing metrics: concurrent Zoom meeting count on the top row, and concurrent Zoom users supported on the bottom row. On each graph, the upper line shows the busiest Zoom server (out of $N = 10$), while the lower line shows the least busy Zoom server. A vertical gap between the two lines indicates load imbalance.

Figure 8 shows (as expected) that load-aware policies are better than load-oblivious ones. For RAND, the number of concurrent Zoom meetings hosted by different Zoom servers can differ by up to a factor of two on this workload, and the number of users by 3-4x.

The two load-based policies keep the number of meetings and/or users much more consistent across the Zoom servers. The MTG policy assigns each new Zoom meeting to the least busy server (in terms of meetings), keeping these counts nicely balanced, though small differences can occur when meetings terminate, since no rebalancing is done then. User counts can still differ by a factor of 2x, since meetings have highly heterogeneous sizes. The USERS policy improves upon the latter metric, by assigning each new Zoom meeting to the least busy server (in terms of users). Load balancing is never perfect, since meetings have different sizes, and terminate at different times. However, load-aware policies do improve greatly upon the load-oblivious policies.

We also considered Round Robin (RR) as another load-unaware policy, which assigns Zoom meetings across the available Zoom servers in a cyclic fashion. Although using
RR improves upon RAND, particularly in meeting counts, it still differs a lot in user counts since meetings have widely varying sizes. For space reasons, we do not present the RR results here.

In summary, there is always an inherent tradeoff between meetings and users, since balancing one of these often unbalances the other. Such tradeoffs are not uncommon in network-based systems with heterogeneous workloads [4], [18]. In our Zoom context, load-aware policies greatly reduce the overall load on the busiest servers, and spread the load more evenly across all servers. Improved load distribution also helps mitigate risk if any particular server or meeting should experience a network disruption.

E. Multicast Model

The final simulation experiment explores the possible use of multicast for Zoom content delivery. The key insight here is that for cloud-hosted Zoom meetings, there is no need to send separate copies of the same audio, video, and data streams from the Zoom data center to each collocated user on the campus network. Rather, a single multicast stream can traverse the Internet to reach the campus network, at which point it is replicated for local delivery to the individual users at different locations on campus. Whether the multicast support is provided at the application layer (in Zoom or a proxy), or natively at the network layer (i.e., IPv4, IPv6), is irrelevant in our simulation model. We simply assume its existence, and focus on the potential savings in the campus network traffic.

Figure 9 shows the results from this multicast Zoom model. Note that the results in these graphs are for Zoom session counts, not Zoom meeting counts as in the earlier graphs. The results confirm the intuitively obvious, namely that the benefits of multicast reduce the Zoom session count dramatically.

The benefits of multicast support for Zoom increase when more users are on campus (see top row of graphs in Figure 9). Furthermore, the performance advantages of multicast support for Zoom are quite robust across this range of Zoom workload parameter settings.

The benefits of multicast Zoom would decrease if small P2P Zoom calls were more prevalent, as illustrated in the bottom row of graphs in Figure 9. P2P Zoom calls do not benefit from multicast, since they are point-to-point. Their prevalence in the workload also reduces the number of lecture-based Zoom meetings generated, and hence the session counts observed. Multicast is still useful for the larger meetings with multiple collocated participants, but the overall benefits are smaller.

VI. CONCLUSIONS

In this paper, we provide a modeling methodology for the simulation and analysis of Zoom network traffic on a campus network. Our approach relies upon a synthetic workload model for campus-level Zoom traffic developed from empirical Zoom measurement data. However, the same methodology could be generalized to other types of networks when empirical data is available to calibrate and validate the model.

Note that 25% of P2P Zoom calls in our model remain internal to the campus network (P1 to P2), 25% remain internal to the ISP network (P3 to P4), so only 50% of the P2P calls traverse the U-R-ISP links (P1/P2 to P3/P4).
The main conclusions from our work are as follows. First, the cost-benefit tradeoff of the on-premise solution to deploy local Zoom Meeting Zones at the campus networks may not be worthwhile. While there are workload parameter settings where this solution makes sense, other settings can actually increase the load on the campus network when a large proportion of users are off-campus. In an era of transition to hybrid work models, managers of enterprise networks similar to our campus network might need to rethink on-premise network solutions, such as the Zoom infrastructure considered in this paper. Second, despite the tradeoff between balancing the number of meetings or users on Zoom servers, the load-aware policies better distribute the load across the servers, reducing the high loads on the busy servers. We showed how these methods could reduce the risk of disruptions that users experience during the meetings due to highly loaded servers and could greatly mitigate the performance issues with Zoom connections that users experience on our network. Finally, with hybrid learning modalities, many collocated students on campus may connect to the same Zoom meetings. Multicast support for Zoom is a promising solution for enterprise networks like our campus. It dramatically lowers the Zoom session count, reducing the incoming load on the campus network.

Our future work will study the dynamics of Zoom’s video bit rate adaptation, and its FEC strategies, on congested and lossy WiFi networks.

ACKNOWLEDGEMENTS

The authors thank the MASCOTS 2022 reviewers for their constructive suggestions that helped to improve the final version of our paper. We also thank Albert Choi and Kiana Gardner for their collection and analysis of Wireshark traces from Zoom test sessions. The authors are grateful to University of Calgary Information Technologies (UCIT) and the Conjoint Faculties Research Ethics Board (CFREB) for enabling the collection and analysis of our Zoom measurement data. Financial support for this work was provided by Canada’s Natural Sciences and Engineering Research Council (NSERC). The software tools described in this paper are available from http://pages.cpsc.ucalgary.ca/~carey/software.htm.

REFERENCES


