DEPLOYING SHADOW BUFFER IN CONTEXT OF CLOCK-SYNCHRONIZED EDGE-BASED NETWORK FUNCTIONS

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 63/320,160, filed March 15, 2022, which is hereby incorporated by reference herein in its entirety.

TECHNICAL FIELD

[0002] This disclosure relates generally to network transmissions and coordinated control of network traffic within data flows.

DESCRIPTION OF THE RELATED ART

[0003] Modern internet infrastructure typically includes large data centers that generate huge amounts of network traffic. When demand is high, data center output may be constrained (e.g., by a capacity of switches, gateways, and the like) and may have to meter network traffic. Such transient congestion scenarios cause bottlenecks and may cause dropped packets. To ensure that packet transmissions have succeeded in the face of such situations, systems have been developed to transmit acknowledgments from receiving nodes to sending nodes as packets are received. However, these acknowledgments are inefficient, in that they contribute to yet further network traffic. Moreover, these acknowledgments are limited to functioning in single-sender to single-receiver scenarios. Yet further, where acknowledgments are not received, packets are simply re-transmitted ad-hoc, potentially running into a same congested switch and achieving a same dropped result, resulting in scenarios where packets are perpetually delayed or even never received by their destination. Additionally, these scenarios are rooted in congestion having already occurred, and are not sufficient to prevent congestion from occurring in the first instance.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004] Figure (FIG.) 1 is an exemplary system environment for implementing netcam and priority functions, according to an embodiment of the disclosure.

[0005] FIG. 2 is a network traffic diagram showing multiple sender hosts sending multiple data flows to a single receiver host, according to an embodiment of the disclosure.

[0006] FIG. 3 is a network traffic diagram showing a timestamping operation at both a sender and receiver side of a data transmission, according to an embodiment of the disclosure.

[0007] FIG. 4 is a data flow diagram showing netcam activities during normal operation and where an anomaly is detected, according to an embodiment of the disclosure.

[0008] FIG. 5 is a network traffic diagram showing a receiver host receiving both high and low priority traffic from sender hosts, according to an embodiment of the disclosure.

[0009] FIG. 6 is a data flow diagram showing netcam activities where priorities are accounted for in determining netcam activity, according to an embodiment of the disclosure.

[0010] FIG. 7 is a flowchart that illustrates an exemplary process for performing netcam activities, according to an embodiment of the disclosure.

[0011] FIG. 8 is a flowchart that illustrates an exemplary process for performing netcam activities in a multiple priority scenario, according to an embodiment of the disclosure.

[0012] FIG. 9 is a data flow diagram showing netcam activities where shadow buffer considerations are depicted, according to an embodiment of the disclosure.

[0013] FIG. 10 is a flowchart that illustrates an exemplary process for performing netcam activities in coordination with shadow buffer considerations, according to an embodiment of the disclosure.

DETAILED DESCRIPTION

[0014] The figures and the following description relate to preferred embodiments by way of illustration only. It should be noted that from the following discussion, alternative embodiments of the structures and methods disclosed herein will be readily recognized as viable alternatives that may be employed without departing from the principles of what is claimed.

[0015] Systems and methods are disclosed herein for coordinating control of data flows in the face of transient congestion. A "netcam" monitors network traffic between clock-synchronized sender and receiver hosts that are part of a data flow. The term "netcam" as used herein, is a term that is short for "network camera," and is a module that tracks network traffic and ensures remedial action is taken where traffic of a data flow in clock-synchronized systems lags beyond tolerable limits. The netcam instructs sender and receiver hosts to

buffer copies of network traffic according to some parameter (e.g., buffer a certain number of packets, buffer packets for a rolling window of time, etc.). Buffers may be overwritten on a rolling basis where the parameter is achieved (e.g., overwrite oldest packet when new packet is transmitted or received and when buffer is full). The netcam may have all sender and receiver hosts write buffer data where an anomaly is detected, and may have the sender hosts re-transmit the written packets. The re-transmission may be subject to jitter (e.g., a time delay between packet transmissions of the data flow), such that where transmission delay or failure occurred due to a given sequence of packet transmission, the jitter causes enough change to nonetheless have the re-transmission attempt succeed. The netcam may determine a need to write and re-transmit packets differently depending on a priority of a data flow. The netcam may instruct shadow buffers at receiver hosts to monitor path usage and capacity, where high usage and/or low capacity may cause the netcam to predict an upcoming anomaly and take remedial action similar to that taken where a buffer is full.

[0016] Advantageously, the netcam implementations disclosed herein enable both improved network transmissions and forensic analysis. The improved network transmissions occur in that writing latest packet transmission attempts to buffers across all machines in a data flow enable re-transmission of an exact set of packets from many machines without reliance on acknowledgment packets that may get lost or dropped across a complex web of machines. Moreover, virtual machines may have bugs that are difficult to detect or isolate. Writing packet sequences associated with an anomaly enables failure analysis, which may enable identification of a faulty virtual machine. Yet further, using shadow buffers to predict anomalies may prevent a scenario where traffic becomes over-congested, enabling remedial action to occur while some capacity remains on a path and without pausing traffic. Additional advantages and improvements are apparent from the disclosure below.

[0017] FIG. 1 is an exemplary system environment for implementing netcam and priority functions, according to an embodiment of the disclosure. As depicted in FIG. 1, netcam environment 100 includes sender host 110, network 120, receiver host 130, and clock synchronization system 140. While only one of each of sender host 110 and receiver host 130 is depicted, this is merely for convenience and ease of depiction, and any number of sender hosts and receiver hosts may be part of netcam environment 100.

[0018] Sender host 110 includes buffer 111, Network Interface Card (NIC) 112, and netcam module 113. Buffer 111 stores a copy of outbound data transmissions until one or more criteria for overwriting or discarding packets from the buffer is met. For example, the

buffer may store data packets until it is at capacity, at which time the oldest buffered data packet may be discarded or overwritten. Other criteria may include a time lapse (e.g., discard packets after predetermined amount of time has elapsed from its transmission timestamp), an amount of packets buffered (e.g., after a predetermined amount of packets are buffered, begin to discard or overwrite oldest packet as new packets are transmitted), and the like.

[0019] In an embodiment, buffer 111 stores information relating to given outbound transmissions, rather than entire packets. For example, a byte stamp may be stored rather than the packet itself, the byte stamp indicating an identifier of the packet and/or flow identifier and a time stamp at which the packet (or aggregate data flow) was sent. In such an embodiment, the stored information need not be overwritten, and may be stored to persistent memory of sender host 110 and/or clock synchronization system 140. This embodiment is not mutually exclusive to buffer 111 storing copies of packets, and they may be employed in combination.

[0020] NIC 112 may be any kind of network interface card, such as a smart NIC. NIC 112 interfaces sender host 110 and network 120.

[0021] Netcam module 113 monitors data flow for certain conditions, and triggers functionality based on the monitored data. As an example, netcam module 113 may, responsive to detecting network congestion, instruct all hosts that are part of a data flow to perform one or more of various activities, such as pausing transmissions, taking a snapshot of buffered data transmissions (that is, writing buffered data packets to persistent memory), and performing other coordinated activity. As used herein, the term data flow may refer to a collection of data transmissions between two or more hosts that are associated with one another. Further details of netcam module 113 are described in further detail with respect to FIGS. 2-8 below. Netcam module 113 may be implemented in any component of sender host 110. In an embodiment, netcam module 113 may be implemented within NIC 112. In another embodiment, netcam module 113 may be implemented within a kernel of sender host 110.

[0022] Network 120 may be any network, such as a wide area network, a local area network, the Internet, or any other conduit of data transmission between sender host 110 and receiver host 130. In some embodiments, network 120 may be within a data center housing both sender host 110 and receiver host 130. In other embodiments, network 120 may facilitate cross-data center transmissions over any distance. The mention of data centers is

merely exemplary, and sender host 110 and receiver host 130 may be implemented in any medium including those that are not data centers.

[0023] Receiver host 130 includes netcam buffer 131, NIC 132, netcam module 133, and shadow buffer 134. Netcam buffer 131, NIC 132, and netcam module 133 operate in similar manners to the analog components described above with respect to sender host 110. Buffer 131 may be a same size or a different size from buffer 111, and may additionally or alternatively store byte stamps for received packets. Any further distinctions between these components as implemented in sender versus receiver host will be apparent based on the disclosure of FIGS. 2-8 below.

[0024] Shadow buffer 134 may be used for tracking data traffic in a manner that enables an early warning of when congestion is likely to come. For example, as data traffic is buffered, congestion may occur when the buffer is full, the congestion preventing further data traffic from flowing until the congestion is cleared. A shadow buffer may increment a counter more quickly than regular buffer (e.g., increment by 1.1 where 1 unit of data is received at a regular buffer), and/or may decrement the counter more slowly than a regular buffer (e.g., decrement by 0.9 or 0.95 where 1 unit of data is cleared at the regular buffer). The term regular buffer, as used herein, may refer to activity of buffers 111 and/or buffer 131 and/or other buffers disclosed herein having similar functionality to that of buffers 111 and/or 131. While only one shadow buffer 134 is depicted in FIG. 1, multiple shadow buffers may be employed at receiver hosts, and each shadow buffer may be allocated to a different subset of data flows, such as data flows each corresponding to a same application. The shadow buffers may increment/decrement at different rates (e.g., to show more congestion for lower priority applications, and to show less congestion for higher priority applications). Alternatively, the shadow buffers may increment/decrement at same rates, but different thresholding may be applied for different applications as to when a data flow should be considered to be facing congestion. Data buffered in a regular buffer includes data traffic (e.g., network packets) received by a receiver; the data is removed from the regular buffer as the data is processed and/or routed to a next destination. Activity described herein of netcam module 113 and/or netcam system 140 taking action with respect to conditions being met with respect to regular buffers may equally be performed where shadow buffer 134 indicates congestion.

[0025] Netcam system 140 includes clock synchronization system 141. Netcam system 140 may monitor data observed by the netcam modules implemented in hosts, such as netcam

module 131 and 133. Netcam system 140 may detect conditions that require action by the netcam modules and may transmit instructions to affected netcam modules to take coordinated action for a given data flow. Clock synchronization system 141 synchronizes one or more components of each host, such as the NIC, the kernel, or any other component within which the netcam modules act. Details of clock-synchronization are described in commonly-owned US Pat. No. 10,623,173, issued April 14, 2020, the disclosure of which is hereby incorporated by reference herein in its entirety. Each host is synchronized to an extremely precise degree to a same reference clock, enabling precise timestamping across hosts regardless of host location, bandwidth conditions of the host, jitter, and the like. Further details of netcam system 140 are disclosed below with reference to FIGS. 2-8. Netcam system 140 is an optional component of netcam environment 100, and the netcam modules of the sender and/or receiver hosts can operate netcam modules without reliance on a centralized system, other than reliance on a reference clock with which to synchronize.

[0026] There are many advantages of netcam environment 100. The netcam modules are edge-based, given that they can run in the kernel or in NICs (e.g., smart NICs) of a host (e.g., physical host, virtual machine, or any other form of host). In an embodiment, the netcam functionality may run as an underlay, meaning that it may run, e.g., as a shim, on a layer of the OSI system under a congestion control layer (e.g., layer 3 of the OSI system). The netcam modules and/or netcam system 140 may instruct hosts to perform activity upon detection of a condition (e.g., a congestion signal is detected using a shadow buffer), such as pausing transmission of a data flow across affected hosts, taking a snapshot (that is, writing some or all of the buffered data, such as the last N bytes transmitted and/or the bytes transmitted in the last S seconds, where N or S may be default values or defined by an administrator), and any other activity disclosed herein. Further advantages and functionality are described below with respect to FIGS. 2-8.

[0027] FIG. 2 is a network traffic diagram showing multiple sender hosts sending multiple data flows to a single receiver host, according to an embodiment of the disclosure. As depicted in FIG. 2, sender host 1 is sending data flow 211 to receiver host 200, sender host 220 is sending data flow 221 to receiver host 200, and, represented by sender host 230, any number of additional hosts may be transmitting respective data flows (represented by data flow 231) to receiver host 200. As depicted in FIG. 2, each data flow sent by each sender host is different; however, this is merely for convenience two or more sender hosts may transmit data from the same data flow. Moreover, a single sender host may send two or

more different data flows to receiver host 200. While only one receiver host is depicted, sender hosts may transmit data flows to any number of receiver hosts.

[0028] We turn to the moment now to FIG. 3 to discuss operation of netcam modules at sender and receiver hosts. FIG. 3 is a network traffic diagram showing a timestamping operation at both a sender and receiver side of a data transmission, according to an embodiment of the disclosure. As depicted in FIG. 3, when sender host 310 transmits a packet to receiver host 320, netcam module 113 of receiver host 320 records sender timestamp 311. Similarly, when receiver host 320 receives the packet, netcam module 133 of receiver host 320 applies receiver timestamp 321. The timestamp reflects a time at which the data packet was sent or received by the relevant component on which the netcam module is installed (e.g., NIC, kernel, etc.). Sender timestamps may be stored in buffers 111 and 131, appended to packets, transmitted for storage in netcam system 140, or any combination thereof.

[0029] Because sender host 310 is synchronized to a same reference clock as receiver host 320, the elapsed time between the time of sender timestamp 311 and receiver timestamp 321 reflects a one-way delay for a given packet. In an embodiment, upon receiving a given packet, receiver host 320 transmits an acknowledgment packet to sender host 310 that indicates receiver timestamp 321, by which netcam module 113 can calculate the one-way delay by subtracting the sender timestamp 311 from the receiver timestamp 321. Other means of calculating the one-way delay are within the scope of this disclosure. For example, the sender timestamp 311 may be appended to the data transmission, and receiver host 320 may thereby calculate the one-way delay without a need for an acknowledgment packet. As yet another example, the netcam modules of sender hosts and receiver hosts may transmit, either in batches or individually, timestamps to netcam system 140, which may calculate one-way delay therefrom. For the sake of convenience and brevity, the scenario where sender host 110 calculates one-way delay based on an acknowledgment packet will be the focus of the following disclosure, though one of ordinary skill in the art would recognize that any of these means of calculation equally apply.

[0030] In an embodiment, the netcam system then determines whether the one-way delay exceeds a threshold. For example, after calculating one-way delay, sender host 110 may compare the one-way delay to the threshold. The threshold may be predetermined or dynamically determined. Predetermined thresholds may be set by default or may be set by an administrator. As will be described further below, different thresholds may apply to different

data flows depending on one or more attributes of the data flows, such as their priority. The threshold may be dynamically determined depending any number of factors, such as dynamically increasing the threshold as congestion lowers, and decreasing the threshold as congestion rises (e.g., because delay is more likely to be indicative of a problem where congestion is not a cause or is a minor cause). In one embodiment, thresholds may be set on a per-host basis, as they may depend on a distance between a sender host and a receiver host. In such an embodiment, the threshold may be a predefined multiple of a minimum one way delay between a sender and a receiver host. That is, the minimum amount of time by which a packet would need to travel from a sender host to a receiver host would be a minimum one-way delay. The multiple is typically 1.5x-3x the minimum, but may be any multiplier defined by an administrator of the netcam. The threshold is equal to the multiple times the minimum one-way delay. Responsive to determining that the one-way delay exceeds the threshold, netcam module 113 may instruct sender host 110 to take one or more actions.

[0031] In an additional or alternative embodiment, determining whether to take one or more actions may be performed using a separate measure of a status of a shadow buffer (e.g., shadow buffer 134). In short (further detail will be described below), during a given data flow, and in parallel with buffering data using a regular buffer, netcam module 133 may instruct shadow buffer 134 be incremented for each unit of data traffic received by receiver host 320. Netcam module 133 may define a dynamic drain rate, which is a rate at which netcam module 133 instructs shadow buffer 134 be decremented. The dynamic drain rate may be determined by netcam module 133 based on a number of units of data removed from buffer 131 per unit of time (e.g., multiplied by a factor that causes drain to occur more slowly in shadow buffer 134 than it occurs in buffer 131). Netcam module 133 may calculate a dwell time as a function of the counter of shadow buffer 134 and the dynamic drain rate (e.g., the dwell time may be calculated by a value of the counter of the shadow buffer divided by the dynamic drain rate). From here, netcam module 133 may determine a one-way delay of the shadow buffer to be the actual one-way delay (determined from the sender and receiver timestamps, described above) as aggregated with the dwell time. The one-way delay of the shadow buffer may be used for comparison against the threshold (in addition to, or instead of, the one-way delay of the regular buffer) to determine whether to take one or more actions.

[0032] Whether driven by the regular buffer or the shadow buffer one-way delay, these one or more actions may include pausing transmission from that sender host when one-way delay is high, which reduces congestion and thereby reduces packet drops on network 120 in

general. The pause may be for a predetermined amount of time, or may be dynamically determined proportionally to the magnitude of the one-way delay. In an embodiment, the pause may be equal to the one-way delay or may be determined by applying an administrator-defined multiplier to the one-way delay. In an embodiment, the netcam determines whether a prior pause is being enforced, and if so, may reduce the pause time based on a prior amount of pause time that has already elapsed from previously acknowledged packets. Moreover, a given data flow may not be the only data flow contributing to congestion, and thus its pause duration may be smaller than the one-way delay or the one-way delay threshold.

[0033] Another action that may be taken is to write some or all buffered data packets (e.g., from either or both of the sender host and receiver host) to persistent memory responsive to the one-way delay exceeding the threshold. Diagnosis may then be performed on the buffered data packets (e.g., to identify network problems). Further actions are described with respect to FIGS. 4-8 in further detail below.

[0034] In some embodiments, data flows may be associated with different priorities. Netcam modules may determine priority of data flows either based on an explicit identifier (e.g., an identifier of a tier of traffic within a data packet header), or based on inference (e.g., based on heuristics where rules are applied to packet header and/or payload to determine priority type). Priority, as used herein, refers to a precedence scheme for which types of data packets should be allowed to be transmitted, and which should be paused, during times of congestion. The priorities disclosed herein avoid a need for underutilizing a link or making explicit allocations of bandwidth, and instead are considered in the context of choosing what packets to transmit during network congestion.

[0035] In order to prioritize high priority packets, a high one-way threshold may be assigned to high priority traffic, and a low, relative to the high one-way threshold, may be assigned to the low priority traffic. These thresholds may be used for comparison against either, or both of, a shadow buffer one-way delay and/or a regular buffer one-way dealy. In this manner, low priority packets will have anomalies detected more frequently than high priority packets, because a lower one-way delay is required to be detected for a low priority packet for an anomaly to be detected by a netcam module, whereas high priority packets will have anomalies detected only when a higher one-way delay threshold has been breached. Following from the above discussion of determining the one-way threshold for a given host, different one-way thresholds may be applied to different data packets that are sent by or received by a same host depending on priority. In priority embodiments, the one-way

threshold may be determined in the manner described above (e.g., by applying a predetermined multiplier to the threshold), where the determination is additionally influenced by applying a priority multiplier. The priority multiplier may be set by an administrator for any given type of priority, but will be higher for higher priorities, and lower for lower priorities. Priority need not be binary – any number of priority tiers may be established, each corresponding to a different type or types of data traffic, and each having a different multiplier. Priorities and their associated multipliers may change over time for given data flows (e.g., where a data flow begins transmitting a different type of data packet that does not require high latency transmission, priority may be reduced).

[0036] Additionally or alternatively to using a priority multiplier on one-way delay thresholds and differentiating one-way delay thresholds based on priority of a given packet or data flow within which a packet is transmitted, the netcam modules may manipulate the pause time of paused traffic during a pause operation differently depending on priority. A low pause time may be assigned to higher priority traffic, and a relatively high pause time may be assigned to lower priority traffic, ensuring that lower priority traffic is paused more often than high priority traffic during times of congestion, and thereby ensuring that higher priority traffic has more bandwidth available while the lower priority traffic is paused. The pause times may be determined in the same manner as described above, but with the additional step of applying an additional pause multiplier to the pause times, with lower pause multipliers (e.g., multipliers that are less than 1, such as 0.7x) for high priority traffic, and higher pause multipliers (e.g., multipliers that are more than 1) for lower priority traffic.

[0037] Priority may be allocated in any number of ways. In an embodiment, one or more "carpool lanes" may be allocated that can be used by data flows having qualifying priorities. For example, a "carpool lane" may be a bandwidth allocation that does not guarantee a minimum bandwidth for a given data communication, but that can only be accessed by data flows satisfying requisite parameters. Exemplary parameters may include one or more priorities that qualify to use the reserved bandwidth of a given "carpool lane." As an example, a carpool lane may require that a data flow has at least a medium priority, and thus both medium and high priorities qualify in a 3-priority system having low, medium, and high priorities. As another example, multiple carpool lanes may exist (e.g., a carpool lane that can only be accessed by high priority traffic in addition to a carpool lane that can be accessed by both medium and high priority traffic).

[0038] In an embodiment, guaranteed bandwidth may be allocated to a given priority. For example, a high priority data flow may be allocated a minimum bandwidth, such as 70mbps. In such an embodiment, excess unused bandwidth from what is guaranteed may be allocated to lower priority data flows until such a time that the bandwidth is demanded by a data flow that qualifies for the guarantee. Guaranteed bandwidth may be absolute or relative. Relative guarantees guarantee that a given priority data flow will receive at least a certain relative amount more bandwidth than a low priority data flow. For example, a high priority data flow may be guaranteed 3x the bandwidth of a low priority data flow, and a medium priority data flow may be guaranteed 2x the bandwidth of a low priority data flow.

[0039] Returning to FIG. 2, where two or more sender hosts transmit data from a same data flow, those nodes, in tandem, and in addition to any receiver hosts that are receiving the data from the data flow, may be referred to as a "cluster." In an embodiment, a data flow may be identified by a collection of identifiers that, if all detected, represent that a data packet is part of a data flow. For example, a netcam module of any host may determine a flow identifier that identifies a data flow to which a packet belongs based on a combination of source address, destination address, source port number, destination port number, and protocol port number. Other combinations of identifiers may be used to identify a data flow to which a packet is a part. As stated before, the hosts of the cluster are all clock-synchronized against a same reference clock, no matter their form (e.g., server, virtual machine, smart NIC, etc.).

[0040] In a scenario where data flows 211 and 221 are a same data flow, sender host 210, sender host 220, and receiver host 200 form a cluster. Following this example, buffering of data packets (across both regular buffers and shadow buffers) may occur on a per-flow level across a cluster of hosts. That is, one or more netcam modules and/or netcam system 140 may record within buffers of hosts of a data flow all packets transmitted or received within whatever parameter the buffer uses to record and then overwrite data (e.g., most recently transmitted packets, packets transmitted/received within a given amount of time, etc.). Moreover, a receiver node receiving packets of a data flow from multiple sender hosts (e.g., receiver host 200 receiving packets from sender hosts 210 and 220) may maintain a single shadow buffer for the data flow, or may maintain separate shadow buffers, one for each of sender host 210 and sender host 220. In an embodiment, indicia of a timed sequence, relative to the reference clock, is stored with the buffered data (e.g., sender timestamp 311 and/or receiver timestamp 321 is stored with a buffered data packet). Thus, sender host 210

and sender host 220 may store in their buffers 111 data packets that share a given flow ID, and receiver host 200 may store received packets within buffer 131. Alternatively or additionally, transmitted and/or received packets may be transmitted to netcam system 140, which may buffer received data.

[0041] From this vantage point of buffering a certain amount of data at each host of a cluster, different functionality of host netcam modules is possible responsive to detection of an anomaly (e.g., the aforementioned conditions mentioned with respect to FIG. 2 above). FIG. 4 is a data flow diagram showing netcam activities during normal operation and where an anomaly is detected, according to an embodiment of the disclosure. Data flow 400 reflects host activities and netcam activities (e.g., activities taken by netcam modules of sender/receiver hosts or netcam system 140) during normal function, and during an "anomaly function" (that is, action taken where an anomaly is detected). Data flow 400 first shows normal function, where hosts send or receive 402 data flows, and the netcam module or system (referred to generally in this figure as "netcam") determines 404 whether an anomaly is detected (e.g., based on one-way delay, as discussed above). Where no anomaly is detected, on the assumption that the buffer is full from prior storage of data packets, the host(s) (e.g., of a cluster) overwrite 406 their buffer(s) (e.g., meaning overwrite oldest packet or follow some other overwrite heuristic as described above). Of course, where buffers are not full, overwriting is not necessary, and storing to a free memory of the buffer occurs. Normal function repeats unless an anomaly is detected.

[0042] Anomaly function occurs where an anomaly is detected. Different anomaly functions are disclosed herein, and data flow 400 focuses on illustrating a particular anomaly function of re-transmitting buffered data. Where sending/receiving 408 information of a data flow by hosts (e.g., of a cluster), the netcam may detect 410 an anomaly. As mentioned above, anomalies are detected where one-way delay (e.g., of a shadow buffer and/or of a regular buffer) exceeds a threshold. Recall that for a cluster, the threshold may vary between hosts of the cluster depending on distance between sender and receiver hosts. Responsive to detecting the anomaly, the netcam instructs 412 the buffered data to be stored at all hosts of the cluster. That is, where an anomaly occurs on even one host of a cluster, data from all nodes of the cluster is stored. This may occur by instructing the hosts to store the buffered data (or the portion thereof relating to the data flow) to persistent memory, or by keeping the buffered data within the buffer and pausing data transmissions, or a combination thereof with different instructions for different hosts. Note that where pause is used, pause time may vary

across the different nodes of the cluster, as mentioned above. Regardless of how the data is stored, the netcam may jitter 414 retransmission timing. Recall that the timed sequence of packet transmissions and receptions is reflected in the stored data packets. The netcam may jitter 414 the retransmission timing by altering the timed sequence (e.g., creating longer lag between a previous time gap between transmissions, transmitting the packets in a different order, etc.). The jitter may occur according to a heuristic, or may be random. Jitter is applied in case the prior attempted timed sequence was the cause of the failure (e.g., because the prior attempted timed sequence itself may cause too much transient congestion), and thus the jitter may in such a scenario result in a success where re-transmission without jitter would fail. The netcam then re-transmits 416 the buffered data (or portion thereof). Note that it may be more expedient and computationally efficient to re-transmit the entire buffer, including data unrelated to the data flow or the anomaly, rather than isolating the packets of the data flow that relate to the anomaly. Normal function then resumes until another anomaly is detected.

[0043] Re-transmission with jitter is only one example of anomaly function, and any number of functions may occur responsive to detection of an anomaly. For example, additionally or alternatively to the anomaly function depicted in data flow 400, the buffered data may be written to persistent memory and stored for forensic analysis. In such a scenario, responsive to detecting an anomaly, the netcam may transmit an alert to an administrator and/or may generate an event log indicative of the anomaly. Any other aforementioned anomaly function is equally applicable. As an example of forensic analysis, a known type of attack on a system such as a data center is a timing attack. Timing attacks may have "signatures," in that an inter-packet spacing of traffic can be learned (e.g., by training a machine learning model using timing patterns as labeled by whether the timing pattern was a timing attack, by using pattern recognition, etc.). Forensic analysis may be performed to determine whether the data was a timing attack. Timing attacks may be blocked (e.g., by dropping data packets from a buffer upon netcam module 113 determining that the buffered data represents a timing attack).

[0044] As mentioned above, buffered data may include byte stamps (as opposed to, or in addition to, buffered packets). Byte stamps may be used in analyzing an anomaly (e.g., in forensic analysis, network debugging, security analysis, etc.). An advantage of using byte stamps, rather than buffered data packets, is that storage space is saved, and byte stamps are computationally less expensive to process. Byte stamps for an amount of time corresponding to an anomaly may be analyzed to determine a cause of the anomaly. The trade off in using

byte stamps, rather than buffered packets, is that buffered packet data is more robust and may provide further insights into an anomaly.

[0045] FIG. 5 is a network traffic diagram showing a receiver host receiving both high and low priority traffic from sender hosts, according to an embodiment of the disclosure. As depicted in FIG. 5, sender host 510 transmits high priority data flow 511 to receiver host 500, and sender host 530 transmits low priority data flow 531 to receiver host 500. Where network congestion occurs and an anomaly is detected, the sender hosts may treat the high and low priority traffic differently. In an embodiment, sender host 530 detects network congestion sooner than sender host 510 because low priority data flow 531 is associated with a lower one-way delay threshold than high priority data flow 511. Therefore, sender host 530 may perform remedial action, such as pausing network transmissions of low priority data flow 531, for a pause time, while high priority data flow 511 continues to transmit because its higher one-way delay threshold has not yet been reached. Where high priority data flow 511 does reach its higher one-way delay threshold, and a pause action is responsively taken, that pause time may be lower than the pause time for low priority data flow 531, thus ensuring that high priority data flow 511 resumes sooner and during a time of less congestion than it would face if low priority data flow 531 were not paused for extra time while high priority data flow 511 continued.

[0046] Similarly, with respect to shadow buffer operation, a high priority shadow buffer may be separately maintained by receiver host 500 for high priority data flow 511, and a low priority shadow buffer may be separately maintained by receiver host 500 for low priority data flow 531. The drain rate may be weighted differently on the basis of priority. For example, the high priority shadow buffer may have a higher drain rate relative to a drain rate used for the low priority shadow buffer, thus resulting in the high priority shadow buffer being less likely to cause a detection of an anomaly than the low priority shadow buffer.

[0047] While depicted as two separate sender hosts, sender hosts 510 and 530 may be a same host, where one sender host transmits both high and low priority traffic to receiver host 500. Thus, a same sender host may take remedial action (e.g., pause) responsive to detecting an anomaly of low priority data flow 531 while continuing to transmit high priority data flow 511 as normal. Sender hosts may have multiple buffers 111, each buffer corresponding to a different priority of data.

[0048] FIG. 6 is a data flow diagram showing netcam activities where priorities are accounted for in determining netcam activity, according to an embodiment of the disclosure.

Data flow 600 begins with one or more sender hosts (e.g., sender host 110) sending 602 a data flow and applying sender timestamps (e.g., sender timestamp 311). A receiver host (e.g., receiver host 130) receives 604 the data flow and applies receiver timestamps (e.g., receiver timestamp 321). Netcam activity then occurs. As described above, the netcam activity may occur at the sender host(s) (e.g., by receiving ACK packets indicating receiver timestamps and using netcam modules to compute one-way delay), at receiver hosts (e.g., where sender timestamps are included in the data flow and netcam modules compute one-way delay therefrom), at netcam system 140, or some combination thereof.

[0049] The netcam determines 606 one-way delay of data packets in data flows. As explained above, the one-way delay computation may depend on a priority of the data flow, and thus different data flows may have different one-way delay thresholds ("priority thresholds"). One-way delay may be determined from packets generally, and/or may be aggregated with dwell time to form a shadow buffer one-way delay. The netcam compares 608 the determined one-way delay (or delays, in the case where shadow buffer one-way delay is used) to the respective priority threshold. Responsive to determining 610 that the one-way delay is greater than the threshold for a given priority data flow, anomaly function is initiated. As depicted in FIG. 6, some anomaly function may include one or more of pausing 612 transmission of the data flow associated with the given priority and/or storing 614 the buffered data flow associated with the given priority (e.g., for forensic analysis). As described above, the pause time may vary depending on the priority level of the paused data flow.

[0050] FIG. 7 is a flowchart that illustrates an exemplary process for performing netcam activities, according to an embodiment of the disclosure. Process 700 may be executed by one or more processors (e.g., based on computer-readable instructions to perform the operations stored in a non-transitory computer-readable memory). For example, netcam modules 113, 133, and/or netcam system 140 may execute some or all of the instructions to perform process 700. Process 700 is described with respect to netcam module 113 for convenience, but may be executed by any other netcam module and/or system.

[0051] Process 700 begins with, for a data flow transmitted between a sender host (e.g., sender host 110) and a receiver host (e.g., receiver host 130), recording 702, on a first rolling basis, by the sender host, a first pre-defined amount of sent network traffic of the data flow (e.g., recording to buffer 111) and recording 704, on a second rolling basis, by the receiver host, a second pre-defined amount of received network traffic of the data flow (e.g., recording

to buffer 131), wherein the sender host and the receiver host are clock-synchronized (e.g., using a reference clock of clock synchronization system 141.

[0052] Netcam module 113 monitors 706 for an anomaly in the data flow based on time stamps of data packets in the network traffic (e.g., by subtracting sender timestamp 311 from receiver timestamp 321 and comparing the result to a one-way delay threshold). Netcam module 113 determines 708 whether an anomaly is detected during the monitoring (e.g., based on whether the comparison shows the one-way delay to be greater than the threshold). Responsive to determining that no anomaly is detected during the monitoring, netcam module 133 may passively allow an overwriting 710 of the recorded sent network traffic and the recorded received network traffic with newly sent network traffic and newly received network traffic, respectively (e.g., recording the latest network traffic over the oldest recorded data packet(s) and going on to repeat elements 702-708). Responsive to determining that an anomaly is detected during the monitoring, netcam module 113 pauses 712 the data flow, causes the sender host to store the recorded sent network traffic to a first buffer, and causes the receiver host to store the recorded received network traffic to a second buffer.

[0053] FIG. 8 is a flowchart that illustrates an exemplary process for performing netcam activities in a multiple priority scenario, according to an embodiment of the disclosure. Process 800 may be executed by one or more processors (e.g., based on computer-readable instructions to perform the operations stored in a non-transitory computer-readable memory). For example, netcam modules 113, 133, and/or netcam system 140 may execute some or all of the instructions to perform process 800. Process 800 is described with respect to netcam module 113 for convenience, but may be executed by any other netcam module and/or system.

[0054] Process 800 begins with netcam module 113 identifying 802 a first data flow between a first sender host (e.g., sender host 110) and a receiver host (e.g., receiver host 130), the first data flow having a high priority (e.g., high priority data flow 511), the sender host and the receiver host synchronized using a common reference clock. Netcam module 113 (e.g., of a different sender host or a same sender host as sender host 110) identifies 804 a second data flow between a second sender host and the receiver host (e.g., low priority data flow 531), the second data flow having a low priority, where the second sender host may be the same or a different host as the first sender host.

[0055] Netcam module 113 assigns 806 a first delay threshold to the first data flow based on the high priority and a second delay threshold to the second data flow based on the low

priority, the first delay threshold exceeding the second delay threshold. Netcam module 113 monitors 808 first one-way delay of data packets of the first data flow relative to the first delay threshold, and monitors 810 second one-way delay of data packets of the second data flow relative to the second delay threshold. Responsive to determining that the first one-way delay of data packets of the first data flow exceed the first delay threshold, netcam module 113 pauses 812 transmission of data packets of the first data flow from the first sender host to the receiver host for a first amount of time. Responsive to determining that the second one-way delay of data packets of the first data flow exceed the second delay threshold, netcam module 113 pauses 814 transmission of data packets of the second data flow from the second sender host to the receiver host for a second amount of time that exceeds the first amount of time.

[0056] FIG. 9 is a data flow diagram showing netcam activities where shadow buffer considerations are depicted, according to an embodiment of the disclosure. Data flow 900 begins with a sender host sending 902 a data flow and applying sender timestamps, and a receiver host receiving 904 the data flow and applying receiver timestamps. These activities are performed in the manner described above with respect to elements 602 and 604 of FIG. 6. As mentioned with respect to FIG. 1, in an embodiment, the receiver host maintains both one or more regular buffers and one or more shadow buffers, where a regular buffer stores data packets as they are received, and a shadow buffer maintains a counter that ticks up as data packets are received and drains according to a dynamic drain rate (that is, decrements according to the dynamic drain rate over each unit of time). Different shadow buffers may be used for different data flows on a same receiver host, and the different data flows may have different priorities.

[0057] A shadow buffer may be in an idle state or an active state. Netcam module 133 of receiver host 130 may determine a shadow buffer to be in an active state responsive to receiving traffic of a data flow (that is, a shadow buffer for that data flow transitions from an idle state to an active state). Netcam module 133 may determine a shadow buffer to be in an idle state responsive to determining that the traffic is no longer received. For example, traffic may be deemed to be no longer received for a data flow where at least a threshold amount of time has passed since a last packet of the data flow was received. As another example, where traffic is consistently received for a data flow on a packet-by-packet basis over each unit of time, and a unit of time passes where a packet is not received for the data flow, netcam module 133 may determine that the traffic is no longer received. Thus, netcam module 133

may continue toggling a state of a shadow buffer for a data flow from idle to active and back depending on whether traffic is received for a data flow. As will be described further below, the state of the shadow buffer is used by netcam module 133 to determine other attributes relating to the shadow buffer, such as drain rate.

[0058] Assuming that the shadow buffer was idle, responsive to receiving a first packet of the data flow in 904, netcam module 133 transitions 905a the shadow buffer from an idle state to an active state, and increments 905b a counter of the shadow buffer that indicates a unit of data traffic received. Where the shadow buffer is already in an active state, 905a is not performed, but 905b continues as each unit of traffic (e.g., packet) is received. In an embodiment, netcam module 133 increments the counter by multiplying the unit of data traffic received by a factor. For example, for every packet received, the counter may be incremented by multiplying the unit by a number greater than 1 (e.g., 1.01, or 1.1). As a particular example where there are multiple priorities, if a packet is received, the shadow buffer may be multiplied by 1.01 if it is a high priority flow, or by 1.1 if it is a low priority flow. The higher the factor, the more quickly the shadow buffer counter will have a number that exceeds a threshold reflecting an anomaly (e.g., a scenario that merits pausing traffic and/or performing remedial measures).

[0059] The netcam (that is, either netcam system 140 or netcam module 133, or some distributed processing) performs the netcam activity depicted in the right-most column of FIG. 9. For convenience, the activity will be referenced as performed netcam module 133, but distributed or entire processing by netcam system 140 is equally possible.

[0060] Netcam module 133 determines 906 a one-way delay of data packets for each data flow, and determines 908 a dynamic drain rate for each shadow buffer corresponding to each respective data flow. While 906 and 908 are depicted sequentially in FIG. 9, these may be performed in parallel with one another or in an opposite order from what is depicted. Element 906 may occur at any point between where it is depicted in FIG. 9 up until the occurrence of 914. Element 906 may be performed in the same manner described above with respect to 606 of FIG. 6.

[0061] Netcam module 133 may determine the dynamic drain dynamic drain rate based on a number of units of the data removed from the regular buffer per unit of time while the shadow buffer is in the active state. That is, if three bytes are removed from the regular buffer for transmission to a next node in a data flow per microsecond, then the rate of 3 per microsecond is a basis from which the dynamic drain rate is determined, multiplied by a

factor less than 1 (e.g., 0.9 or 0.95) such that drain from the shadow buffer occurs more slowly than drain from the regular buffer. The reason to decrement the shadow buffer at a slower rate than the regular buffer is, again, to ensure that where an anomaly might occur on the regular buffer, it is first detected using the shadow buffer. Netcam module 133 may select a factor to multiply by the drain rate based on priority of data flow, where high priority data flows have higher drain rates (e.g., 0.95-0.99), where medium and low priority data flows have lower drain rates (e.g., 0.9-0.94 for medium and 0.85-0.89 for low).

[0062] Netcam module 133 may determine the dynamic drain rate on any cadence, such as each time a data packet is received by receiver host 130, or on a slower cadence, such as for every Nth data packet received in a given data flow. Netcam module 133 may limit performance of determining 908 the dynamic drain rate to scenarios where the shadow buffer is in an active state. Where the shadow buffer is in an idle state, netcam module 133 may render a last determined dynamic drain rate as a static drain rate to use over time to decrement the shadow buffer until such a time that the shadow buffer re-enters an active state, whereafter netcam module 133 may recalculate a new dynamic drain rate.

[0063] The dynamic drain rate is used by netcam module 133 for two purposes. First, the dynamic drain rate is used to decrement the shadow buffer counter over time. Second, the dynamic drain rate is used to calculate a "dwell time." The term dwell time, as used herein, refers to a value that may be aggregated with the actual one-way delay of packets on a data flow as a congestion signal for determining whether there is an anomaly in the data flow that requires remedial measures to be taken.

[0064] Netcam module 133 determines 910 the dwell time as a function of the counter of the shadow buffer (e.g., which is a proxy of a length of the regular buffer with some added length based on the incremental and drain multiplier factors) and the dynamic drain rate. In an embodiment, netcam module 133 calculates the dwell time by dividing a value of the counter of the shadow buffer by the dynamic drain rate.

[0065] Netcam module 133 determines 912 a congestion signal for the data flow based on the dwell time. In an embodiment, netcam module 133 determines the congestion signal by mathematically aggregating a one way delay between the sender host and the receiver host with the dwell time. Similar to calculating the dynamic drain rate and incrementing the counter, netcam module 133 may weight the dwell time by a factor. For example, the dwell time may be weighted depending on priority of a data flow, where a larger multiplier may be used for lower priority data flows, and a smaller multiplier may be used for higher priority

data flows (e.g., 1.01-1.05 for a high priority data flow; 1.06-1.14 for a medium priority data flow; 1.15-1.30 for a low priority data flow). This, again, will cause higher priority data flows to be impacted less frequently than lower priority data flows that will more quickly have their congestion signal reach a threshold that triggers remedial action.

[0066] In a similar manner to FIG. 6's discussion of elements 610-614, netcam module 133 may determine 914 that the congestion signal exceeds a threshold (e.g., a priority-specific threshold, similar to that used for regular buffers), and may take remedial action. The remedial action may include storing 916 data or indications of data for the associated data flow, and/or pausing 918 transmission of the associated data flow.

[0067] FIG. 10 is a flowchart that illustrates an exemplary process for performing netcam activities in coordination with shadow buffer considerations, according to an embodiment of the disclosure. Process 1000 may be executed by one or more processors (e.g., based on computer-readable instructions to perform the operations stored in a non-transitory computer-readable memory). For example, netcam modules 113, 133, and/or netcam system 140 may execute some or all of the instructions to perform process 1000. Process 1000 is described with respect to netcam module 133 for convenience, but may be executed by any other netcam module and/or system.

[0068] Process 1000 begins with netcam module 133 maintaining 1002 a plurality of buffers at a receiver host, the plurality of buffers comprising a regular buffer and a shadow buffer (e.g., buffer 131 and shadow buffer 134). Netcam module 133, responsive to receiving a data flow from a sender host that is clock-synchronized with the receiver host using a common reference clock, performs 1004: storing a first indication of data of the data flow to the regular buffer (e.g., storing a data packet or metadata corresponding to the data packet to buffer 131), transitioning the shadow buffer from an idle state to an active state (e.g., where this is the beginning of traffic in the data flow since a last break in traffic), and incrementing a counter of the shadow buffer that indicates a unit of data traffic received (e.g., counter of shadow buffer 134 that corresponds to the data flow).

[0069] Netcam module 133 determines 1006 a dynamic drain rate based on a number of units of the data removed from the regular buffer per unit of time while the shadow buffer is in the active state, where the shadow buffer reverts to an idle state responsive to a break in the receiver host receiving the data flow. Netcam module 133 calculates 1008 a dwell time as a function of the counter of the shadow buffer and the dynamic drain rate, and determines 1010 a congestion signal for the data flow based on the dwell time (e.g., the congestion signal

used to detect an anomaly in the same manner described with respect to 708 of FIG. 7).

WHAT IS CLAIMED IS:

1. A computer-implemented method comprising:

maintaining a plurality of buffers at a receiver host, the plurality of buffers comprising a regular buffer and a shadow buffer;

responsive to receiving a data flow from a sender host that is clock-synchronized with the receiver host using a common reference clock:

storing a first indication of data of the data flow to the regular buffer,

transitioning the shadow buffer from an idle state to an active state, and

incrementing a counter of the shadow buffer that indicates a unit of data traffic received;

determining a dynamic drain rate based on a number of units of the data removed from the regular buffer per unit of time while the shadow buffer is in the active state, wherein the shadow buffer reverts to an idle state responsive to a break in the receiver host receiving the data flow;

calculating a dwell time as a function of the counter of the shadow buffer and the dynamic drain rate; and

determining a congestion signal for the data flow based on the dwell time.

2. The computer-implemented method of claim 1, wherein the plurality of buffers comprise a plurality of shadow buffers, the plurality of shadow buffers comprising the shadow buffer.

3. The computer-implemented method of claim 2, wherein each of the plurality of shadow buffers acts on a different subset of a plurality of data flows received by the receiver host.

4. The computer-implemented method of claim 3, wherein each different subset is determined to include data flows common to a different given application to which the data flows correspond, each different application having a different priority, the drain rate weighted on a basis of the different priority.

5. The computer-implemented method of claim 3, wherein each different subset is determined to include data flows common to a given set of senders.

6. The computer-implemented method of claim 1, where incrementing the counter comprises multiplying the unit of data traffic received by a factor.

7. The computer-implemented method of claim 1, further comprising, while the shadow buffer is in the active state, decrementing the shadow buffer at a slower rate than the data is removed from the regular buffer.

8. The computer-implemented method of claim 7, wherein the slower rate is determined by multiplying the rate at which data is removed from the regular buffer by a factor greater than 0.9 but less than 1.

9. The computer-implemented method of claim 1, wherein the dwell time is calculated by a value of the counter of the shadow buffer divided by the dynamic drain rate.

10. The computer-implemented method of claim 1, wherein the dynamic drain rate is held responsive to the shadow buffer transitioning to an idle state, and wherein the dynamic drain rate is re-determined responsive to the shadow buffer transitioning back to an active state.

11. The computer-implemented method of claim 1, wherein determining the congestion signal comprises mathematically aggregating a one way delay between the sender host and the receiver host with the dwell time.

12. The computer-implemented method of claim 11, wherein mathematically aggregating the one way delay comprises weighting the dwell time by a factor.

13. A non-transitory computer-readable medium comprising memory with instructions encoded thereon, the instructions, when executed, causing one or more processors to perform operations, the instructions comprising instructions to:

maintain a plurality of buffers at a receiver host, the plurality of buffers comprising a regular buffer and a shadow buffer;

responsive to receiving a data flow from a sender host that is clock-synchronized with the receiver host using a common reference clock:

store a first indication of data of the data flow to the regular buffer,

transition the shadow buffer from an idle state to an active state, and

increment a counter of the shadow buffer that indicates a unit of data traffic received;

determine a dynamic drain rate based on a number of units of the data removed from the regular buffer per unit of time while the shadow buffer is in the active state, wherein the shadow buffer reverts to an idle state responsive to a break in the receiver host receiving the data flow;

calculate a dwell time as a function of the counter of the shadow buffer and the dynamic drain rate; and

determine a congestion signal for the data flow based on the dwell time.

14. The non-transitory computer-readable medium of claim 13, wherein the plurality of buffers comprise a plurality of shadow buffers, the plurality of shadow buffers comprising the shadow buffer.

15. The non-transitory computer-readable medium of claim 14, wherein each of the plurality of shadow buffers acts on a different subset of a plurality of data flows received by the receiver host.

16. The non-transitory computer-readable medium of claim 15, wherein each different subset is determined to include data flows common to a different given application to which the data flows correspond, each different application having a different priority, the drain rate weighted on a basis of the different priority.

17. The non-transitory computer-readable medium of claim 15, wherein each different subset is determined to include data flows common to a given set of senders.

18. The non-transitory computer-readable medium of claim 13, where the instructions to increment the counter comprise instructions to multiply the unit of data traffic received by a factor.

19. The non-transitory computer-readable medium of claim 13, the instructions further comprising instructions to, while the shadow buffer is in the active state, decrement the shadow buffer at a slower rate than the data is removed from the regular buffer.

20. The non-transitory computer-readable medium of claim 19, wherein the slower rate is determined by multiplying the rate at which data is removed from the regular buffer by a factor greater than 0.9 but less than 1.