

Constellation: A High Performance Geo-Distributed Middlebox Framework

Milad Ghaznavi
University of Waterloo
eghaznav@uwaterloo.ca

Ali José Mashtizadeh
University of Waterloo
mashti@uwaterloo.ca

Bernard Wong
University of Waterloo
bernard@uwaterloo.ca

Raouf Boutaba
University of Waterloo
rboutaba@uwaterloo.ca

Abstract

Middleboxes are increasingly deployed across geographically distributed data centers. In these scenarios, the WAN latency between different sites can significantly impact the performance of stateful middleboxes. The deployment of middleboxes across such infrastructures can even become impractical due to the high cost of remote state accesses.

We introduce Constellation, a framework for the geo distributed deployment of middleboxes. Constellation uses asynchronous replication of specialized state objects to achieve high performance and scalability. The evaluation of our implementation shows that, compared with the state-of-the-art [80], Constellation improves the throughput by a factor of 96 in wide area networks.

1 Introduction

Middleboxes, such as firewalls, load balancers, and intrusion detection systems are pervasive in computer networks [20, 32, 64, 68]. The network function virtualization vision enables middleboxes to be flexibly deployed across a network and provisions new instances on demand. Middleboxes may have multiple instances to satisfy traffic demand and share state across instances to cooperatively process traffic.

Although the instances of a middlebox are typically deployed in the same data center, there has been an increasing demand for deploying middlebox instances across wide area networks [3, 21, 34, 58]. This growth stems from the trend towards building multi-data center applications, which necessitates global scale network management.

In many cases, even though the instances are connected by high latency wide area links, it is still necessary for them to share state [2, 22, 54, 61, 81]. Examples include distributed rate limiters that share traffic information to monitor and limit the traffic of multi-data center applications [61], intrusion detection systems with instances across a large ISP network that share statistics to detect attacks [22, 54, 81], and proxies in a content delivery network that actively share their health statuses [78].

Existing middlebox frameworks that support state sharing have focused on optimizing for local area network deployments [38, 47, 63, 80]. Full control over routing allows these frameworks to maintain affinity between traffic flows to middlebox instances. This results in fewer remote accesses since per-flow state remains mostly local to an instance [47, 63, 80]. However, in wide area networks, traffic can span multiple administrative domains, giving a middlebox framework much less control over routing. Asymmetric routing and multipath protocols [57, 74] compound this issue because a single flow may traverse multiple instances, thus requiring state sharing to process such flows. The result is more remote accesses to shared state across wide area links, which increases packet latency and reduces middlebox throughput. These frameworks use synchronous state access for correctness, which is only practical within a local area network as it can add a network roundtrip delay to each packet.

In this paper, we introduce Constellation, a framework for geo-distributed middlebox deployments. Constellation provides a state management system that is highly scalable and performant even when middlebox instances are deployed across wide area networks. It separates the middlebox state from its application logic and abstracts shared state using *convergent state objects*, which can be independently updated yet still converge. Transparent to the middlebox application logic, Constellation asynchronously replicates state objects to other middlebox instances. Replication makes the state local to each instance, and convergence allows a middlebox instance to mutate state with only lightweight coordination.

Asynchronous replication of convergent state enables more flexible load balancing. Replication subsumes the need for flow-instance affinity, and enables any middlebox instance to process any packet with high performance, as the instance already has the required state. Replication also provides seamless dynamic scaling since traffic load can be rebalanced among instances without waiting for state migration.

Standard conflict free replicated data types (CRDTs) [66, 67] are convergent objects that can be used to represent the shared state for a class of common middleboxes. However,

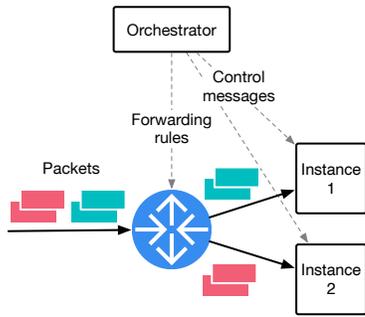


Figure 1: NFV environment: An orchestrator distributes the traffic workload among multiple middlebox instances.

to support middleboxes such as intrusion detection systems and network monitors [25, 31, 35, 52, 53], we also develop new CRDTs including a counting bloom filter and counting sketch. These CRDTs rely on an *ordering* property that is provided by our framework. Moreover, state updates in some middleboxes may necessitate violating CRDT properties. To address this limitation, we introduce *derivative state objects* to support packet processing where the same set of objects are always accessed together. This is commonly used in network address translators and load balancers.

The properties of convergent state objects offer unique opportunities for us to build an efficient and reliable multicast state replication layer. Using the idempotence and commutativity properties of state objects, this layer coalesces state updates for more efficient utilization of wide area network bandwidth. It also provides higher tolerance for straggler instances, as they can receive and apply batched updates to reduce bandwidth and processing resources compared with executing uncoalesced operations.

We implemented Constellation using Click [48], and evaluated our framework by comparing its performance with S6 [80], the state-of-the-art middlebox framework for local area networks. Our results show that Constellation scales linearly with throughput and experiences no overhead due to network end-to-end latency. Over wide area networks, Constellation can process 96× the bandwidth of S6, which was not designed to tolerate latency. In local area networks, Constellation can process up to 11× the bandwidth of S6, which comes from eliminating the heavyweight mechanisms that S6 uses to hide remote state accesses (see § 6). Finally, we show that the complexity of our middleboxes is similar to synchronous approaches when compared to S6.

2 Background and Motivation

Figure 1 shows a typical network function virtualization (NFV) environment, where an orchestrator manages middlebox instances deployed on servers and the network connecting these servers. In response to traffic load, the orchestrator

dynamically adds or removes middlebox instances, and installs forwarding rules in the network to redistribute traffic.

The above operations are sufficient to scale stateless middleboxes, e.g., firewalls that pass or block individual packets based on static rules. However, scaling *stateful* middleboxes becomes challenging, since in addition to the aforementioned operations, the middlebox state must be migrated simultaneously with the new workload distribution [58, 63, 80].

2.1 Middlebox State

Stateful middlebox instances maintain *dynamic state* regarding traffic flows, which changes how they process packets [40, 62, 71]. For example, stateful firewalls filter packets based on information collected about flows [23].

The middlebox state consists of partitionable and shared state. The partitionable state is accessed by a single instance, for example the cache in a web proxy [16]. Shared state can be for a single flow or a collection of flows processed across middlebox instances, and multiple instances query and mutate them. For example, IDS instances read and update port-counts per external host to detect attacks.

2.2 Recent Work: State Management for LAN

Existing frameworks that manage shared state are optimized for local area networks and support synchronous accesses to state [38, 46, 47, 63, 80]. State sharing using this model leads to remote accesses that incur performance cost relative to the network latency.

Two main approaches exist. One approach separates middlebox state into a remote data store [46, 47]. Remote state accesses increase packet latency and can reduce throughput by ~60% [46], because extra CPU and bandwidth resources are consumed for remote I/Os. Another approach [38, 63, 80] distributes state across middlebox instances. An instance must query remote instances for non-local state [80]. Frequent remote accesses can significantly degrade performance. Full control over routing allows the existing frameworks to reduce remote accesses by consistently routing a traffic flow to the same instance so that the flow state remains local to that instance.

Synchronous remote accesses of shared state lead to decreased performance for both approaches. These frameworks introduce several optimizations to maintain performance of their middlebox implementations. The state-of-the-art framework, S6 [80], masks the overhead of remote accesses using concurrency. An instance creates a microthread per packet to enable context switching to other packets while waiting on synchronous requests. However, as we will discuss in § 6.2, our results show that the overhead of using a microthread per packet halves the maximum throughput of the framework.

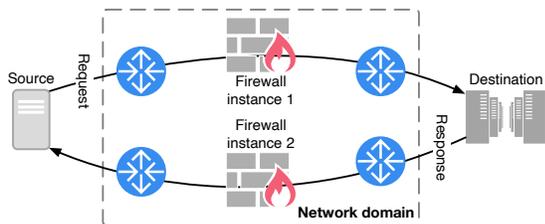


Figure 2: Firewall and asymmetric routing: A source creates a connection with a destination. The request and response streams of this connection take separate paths and pass through two firewall instances.

Another optimization is to trade consistency for performance. An instance caches state and performs reads and writes locally [47, 80]. To avoid permanent divergence, cached state copies must be merged periodically. Doing so naively can result in the consistency anomalies, such as lost updates.

These optimizations complicate the middlebox design. To achieve acceptable levels of performance and scalability, developers may need to use a combination of these mechanisms. Reasoning about their correctness is complicated; an incorrect usage can be a source of subtle bugs in middlebox applications.

2.3 Geo-distributed Middleboxes

Middleboxes are increasingly being deployed across wide area networks, e.g., ISP networks, multiple data centers, and content delivery networks. In such deployments, middlebox instances share state to cooperatively process traffic.

For example, rate limiter instances monitor and limit traffic loads from multiple locations in a content delivery network [61, 78]. They share their state so that they can limit the global traffic load of multi-data center applications [61] and control the impact of traffic spreaders [78].

IDS instances deployed across an ISP network [22] share their local statistics to detect intrusions [22, 54, 81]. Using network-wide statistics collected from different network locations is essential to detect attacks, such as port scanning and denial of service [54, 69, 77, 81]. Moreover, multiple NAT instances share the same flow table to translate network addresses across an ISP network [1].

In the CoDeeN peer-to-peer content delivery network [78], distributed proxies share their health status. To handle a cache miss, a proxy redirects a content request to another peer that is healthy based on this information.

These middlebox deployments face two challenges. The first challenge is due to characteristics of wide area network traffic. Traffic can span multiple administrative domains with less control over routing. This increases shared state hindering scalability and performance. Moreover, asymmetric routing and multipath routing [57, 74, 75] undermine the

State Location	Latency	Throughput
Local Machine	1–10× ns [24]	100 M–1 G
Remote LAN	10–100× μs [36, 59]	10 k–100 k
Remote WAN	10–100× ms [28, 49]	10–100

Table 1: Time to access state in different locations: The throughput of remote state across a wide area network can be as low as 10 to 100 accesses per second.

flow-instance affinity. In asymmetric and multipath routing, a traffic flow, e.g., sub-flows of a MPTCP session [57], may traverse different paths and consequently different middlebox instances. For a correct operation, the instances must share even per-flow state.

Figure 2 shows an example asymmetric routing scenario where the traffic of a connection passes through two firewall instances. A firewall commonly allows connections to initiate only from protected zones, e.g., “Source” in Figure 2. State sharing among firewall instances is essential, since the second instance allows the response stream only if the first instance shares that it has observed the request stream [2, 39].

The second challenge is wide area network latency making state sharing extremely costly. Table 1 lists the access times to shared state when it is stored locally, or remotely over a local and wide area networks. Existing frameworks [46, 47, 80], designed for infrequent state sharing in local area networks, cannot tolerate frequent state sharing over networks with higher latency [26, 28, 49, 55].

3 Design Overview

Table 2 shows a list of popular middleboxes with a selection of their state. This list is not exhaustive but provides a representative set of abstract data types used by middleboxes. To better understand the needs of middleboxes in wide area networks, we start by examining the needs of these popular middleboxes.

3.1 Study of Common Middleboxes

Our study reveals two key observations. First, most middleboxes maintain shared state for collecting traffic statistics, resource ownership, and resource usage. Second, middleboxes mostly operate on relatively small shared state and trade off precision for higher scalability and performance [22]. A corollary is that most operations on shared state are simple even when the middleboxes are not [46, 47].

Purpose of sharing state: Middleboxes collect statistics about traffic for detection and mitigation purposes. As shown in Table 2, an IDS/IPS track statistics of traffic connections and sessions to detect abnormal or malicious communications. The instances of a signature based IDS need to share their

Middlebox	State	Purpose	Shared Abstract data type		Size (B)
IDS/IPS [15, 19, 31, 50, 52]	Session context	Session inspection	✓	Map	$96 \times c$
	Connection context	Connection inspection	✓	Map	$\sim 250 \times c$
	Bloom filter	String rule matching	×	Bloom filter	6250
	Flow size distribution	Traffic summaries	✓	Count-min sketch	20 k–200 k
	Port scanning counter	Port scan detection	✓	Counters	$112 \times n$
Firewall [10, 14, 23, 39, 76]	Flow table	Stateful firewall/dynamic rules	✓	Map	$32 \times c$
	Connection context	Connection inspection	✓	Map	$264 \times c$
Network monitor [41, 43, 44, 69]	Traffic dispersion graph	Anomaly detection	✓	Graph	$256 \times n$
	Heavy hitters	Tracking top elephant flows	✓	Count-min sketch	-
Load balancer [7, 13, 32]	Server pool	Available backend servers	×	-	$20 \times n$
	Server pool usage	Usage of backend servers	✓	Vector	$28 \times n$
	Flow table	Connection/Session persistence	✓	Map	$28 \times c$
NAT [14, 21]	Available address pool	Tracking available addresses	✓	Set	$80 \times c$
	Flow table	Address mapping	✓	Map	$74 \times c$
Web proxy [17, 35]	Stored entries	Metadata of cached contents	×	Map	$104 \times c$
	Cache contents	Caching in memory or storage	×	-	Available storage
	Cache digests	Compact cache summary	✓	Counting bloom filter	$20 \times c$

Table 2: Examples of common middleboxes: A list of common middlebox applications are shown. For “Size (B)”, c and n are respectively the number of connections/sessions and hosts/servers. Note that we provide a representative set of state for each middlebox, and they are not exhaustive. Moreover, for each middlebox application, we list the state of multiple implementations, and a single implementation does not necessarily include all the presented state.

statistics to detect advanced attacks that can exploit multi-path routing in a wide area network. These attacks split their signatures across multiple paths to circumvent traditional signature based detection approaches [54]. A stateful firewall inspects the collected statistics to block malicious connections and maintain dynamic rules for outgoing connections. As mentioned before, the firewall instances may be required to share their state to handle asymmetric flows [2]. A network monitor maintains a traffic dispersion graph that embodies the communications between network nodes. A network wide representation can monitor thousands of hosts to detect large scale attacks [44, 73].

Middleboxes also track resource ownership and usage for resource management purposes. As shown in Table 2, a load balancer manages backend servers, and distributes load among them based on their usage. A NAT manages a set of available public addresses and allocates these addresses among network flows. NAT instances in an ISP network share their state to correctly route asymmetric flows [1].

Shared state implementation: Shared state tends to be small, which reduces communication overheads between instances [54] and per-packet processing costs. For the middleboxes shown in Table 2, to serve millions of flows, the shared state requires only few 100 MB of the memory. For example, the most memory intensive middleboxes are web

proxies that keep a large cache local to each instance [16]. Advanced proxies share a compact summary of their cached contents [35] to allow redirecting content requests to nearby instances. This improves the quality of service in serving content requests in a content delivery network.

Although many middleboxes make complex decisions based on shared state, their operations on shared state are simple. Others have also observed that middleboxes operate on shared state with a simple set of operations [46, 47]. For example, an IDS collects lightweight packet summaries, but performs complex detection operations locally. Rate limiter instances share their observed flow rates [61] and use probabilistic analysis to shape traffic of multiple data centers.

Middleboxes collect approximate statistics when collecting precise statistics leads to high memory or processing overheads. They sometimes use compact and approximated statistics for a faster request serving. As shown in Table 2, IDSes and network monitors often use count-min sketches or bloom filters, which are probabilistic data structures, to track top heavy hitters. Web proxy uses cache digests to quickly check local contents when serving requests.

3.2 Constellation Design Choices

Our observations lead us to design Constellation. Constellation is a geo-replicated middlebox framework that deploy

a cluster of middlebox instances distributed across a wide area network. Constellation provides for the management of shared state across the entire deployment.

Constellation separates the design of middleboxes into middlebox logic and middlebox state to hide the complexity of state sharing from the middlebox logic. Transparent to the logic, Constellation asynchronously replicates updates to shared state from each middlebox instance. Using convergent state, Constellation eliminates most of the complexity of managing asynchrony. Our key design choices are as follows.

Asynchronous state replication: Constellation replicates middlebox state to all instances asynchronously. Each middlebox instance collects and sends its updates of shared state to all other instances in near real time, and they apply these updates to their state locally. All instances access shared state locally without querying remote instances.

Asynchronous access to local state allows middlebox instances to share state over high latency links of a wide area network. Replication also supports flexible load distribution and seamless dynamic scaling by subsuming the need for flow-instance affinity. If a middlebox instance is overwhelmed, traffic can be immediately rerouted to an existing instance that is underutilized. In removing an excess instance, the orchestrator can reassign traffic load from this instance to another without waiting for state migration.

Storing a replica of shared state requires more memory than that of existing frameworks, but the overhead is not substantial. As we observed in § 3.1, many middleboxes operate on lightweight shared state with small memory requirement. Even larger memory usage does not change the cost of running middleboxes in the cloud, where computation to memory ratios are fixed. Middlebox applications already require substantial compute resources that usually goes hand-in-hand with more than enough memory.

Asynchronous state replication also trades the consistency of state across instances for performance. For many cases, our state model framework automatically resolves inconsistencies using *convergent state objects*.

Convergent state objects: Constellation provides a set of state objects to develop the middlebox state. These objects are guaranteed to be *convergent*, i.e., middlebox instances will observe the same local value for a state object after they receive and perform the same set of state updates applied in other instances. Convergence eliminates complexities that may arise due to asynchronous state replication assuring developers about the correctness of shared state.

As we discussed in § 3.1, most middleboxes perform simple operations on shared state. For these middleboxes, Constellation’s builtin convergent state objects can be used to

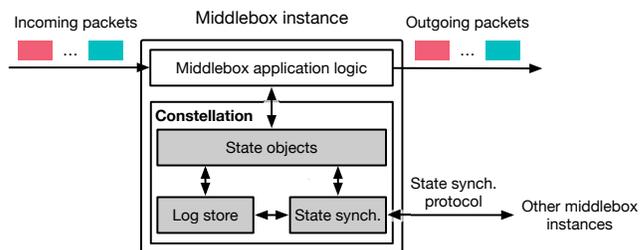


Figure 3: Constellation Middlebox Framework

implement their shared state. For more complex cases, Constellation provides mechanisms for developers to customize state objects to reconcile conflicting state updates.

Asynchronous state replication even when using convergent state objects may introduce some artifacts. We discuss these artifacts and their impact on middleboxes in § 5.2.

Centralized orchestration: A logically centralized orchestrator [5, 11, 12, 27] monitors traffic load, scales the number of instances according to load variations, and distributes the load among the instances. Constellation does not involve the orchestrator in state replication to avoid creating a potential performance bottleneck.

4 Constellation Middlebox Framework

Constellation is designed to allow developers to create geo-distributed middlebox applications with seamless scalability and network latency tolerance. Figure 3 shows a middlebox instance and the main components of the Constellation framework. Our framework works as follows.

The *middlebox logic* registers a set of *state objects* that model the middlebox state. The state objects provide APIs to access and mutate state, which are used by the middlebox logic during its packet processing (discussed in § 4.1).

Constellation internally tracks local state updates by recording them into its *log store*. The *state synchronization* component replicates the recorded state updates to other instances concurrent with packet processing. All other instances will apply state updates received by the state synchronization component to their local state objects (discussed in § 4.2).

During adding or removing of middlebox instances, Constellation adjusts the membership of the middlebox cluster while keeping other instances to process traffic. Constellation selects only one existing instance as a source of the state in bringing a new instance up-to-speed; other instances experience almost no performance disruption by this membership change (discussed in § 4.3).

4.1 State Objects

A state object encapsulates a set of variables and operations to access and update their values. Operations are designed

to guarantee that state objects remain convergent, which simplifies reasoning about asynchronous state replication.

To ensure convergence, Constellation uses data structures based on CRDTs [66, 67] for its state objects. Two general types of CRDTs exist [66]: i) *state based* CRDTs where a state update is performed completely local, and the *entire* state object is disseminated for state synchronization; ii) an operation based CRDTs where operations are propagated. Due to the high bandwidth overhead of state based CRDTs, we opt for operation based CRDT.

For convergence, each instance sends its local update operations to other instances, and each downstream instance applies the received operations. All operations are *idempotent* and *commutative* so instances can safely apply the operations out-of-order and still converge [66].

We create a library of convergent state objects based on our study of various middleboxes shown in Table 2. The library consists of a collection of abstract data types based on operation-based CRDTs, for instance different flavors of map, set, register, and counter objects. These convergent objects can implement the basic state of common middleboxes.

Constellation addresses two limitations of existing CRDTs in implementing the state of middleboxes. First, Constellation provides atomic updates across multiple objects to address the limitation of CRDTs in supporting multi-object updates. Second, Constellation develops a *convergent* counting bloom filter and count-min sketch which are common in intrusion detection systems, network monitors, and proxies (see Table 2). These objects have not been proposed in prior work to the best of our knowledge.

Multi object updates: Middleboxes need to mutate multiple objects simultaneously, but this may violate the properties of CRDTs. This limitation is because convergence is guaranteed for operations on individual CRDT objects [66]. Reasoning about the convergence of general multi-object operations is complicated, because the objects may diverge when mutations on multiple objects do not commute.

In operating on multiple state objects, there are middleboxes that always update the objects together. This is common in managing resources where middleboxes update resource usage or ownership when they allocate or release the resources. For example, a NAT updates an address pool and a flow table object together in processing a new flow, and a load balancer updates the server pool usage and its flow table together in assigning a new flow to a backend server.

Operations on multiple state objects do not violate convergence when two conditions hold. First, such a multi-object operation commutes with other operations defined for the objects. Second, the operation is idempotent.

Constellation supports multi-object operations by defining a derivative object that contains multiple state objects,

and performs operations on its state objects simultaneously. Constellation *atomically* replicates this derivative object in a downstream instance to ensure convergence. We discuss using derivative state objects to develop a NAT in § 5.1.

We also take advantage of the *ordering property* of our system to develop a convergent counting bloom filter object and count-min sketch object. We discuss the ordering property in § 4.2.

Counting bloom filter: A counting bloom filter (CBF) is a memory efficient object for approximate counting. CBF is used in packet classification, deep packet inspection, and network monitoring [25, 31, 35, 53] where keeping accurate statistics with fine granularity does not scale to traffic load.

A CBF represents a large set of n counters using a smaller vector of m counters. It uses k hash functions to update the counters. A CBF exposes count and value operations. The count(x) operation computes k hash values for an operand x (e.g., x can be a five tuple flow identifier). Each hash value provides an index $0 \leq i < m$, and the CBF increments the counter at k computed indices. On value(x) operation, the CBF computes the same set of k hashes and returns the minimum value among the relevant counters. The returned value is an approximation, since the exact value of x 's counter is less than or equal to this value.

For convergence, value and count operations must be commutative and idempotent [66]. As value does not mutate any counter, we focus on count operation. Addition commutes, thus count also commutes; however, addition is not idempotent. Idempotence can be provided using Constellation's ordering feature. This feature prevents applying a duplicate count operation, thus a local count operation performed in a middlebox instance is only applied once at any other instance across the entire middlebox cluster.

Count-min sketch: A count-min sketch (CMS) is a probabilistic data structure similar to counting bloom filter and has application in packet classification and deep packet inspections [52, 65].

A CMS has k arrays of n counters. A CMS uses k hash functions to update the counters, one hash function per array. A CMS provides the same set of operations as a counting bloom filter. To provide convergence, we use the same technique as that of a counting bloom filter. The idempotence of CMS is provided using Constellation's ordering feature.

4.2 Asynchronous State Replication

An efficient replication system for convergent state objects has two requirements. First, we must deliver and apply all operations to all other instances quickly to achieve fast convergence. Convergent state objects require a one-to-all dissemination of updates and allow commutativity of updates. Constellation uses multicast to build a replication layer that

allows unordered delivery of state updates. Multicast uses bandwidth efficiently compared to having $O(n^2)$ point to point transports and reduces latency compared to other topologies, e.g., forwarding state updates through middlebox instances one by one.

Second, we must mitigate straggler instances that can slow down the replication for the entire middlebox cluster. Instances may fall behind because of insufficient bandwidth or processing power. We use the idempotence and commutativity properties of CRDTs to build a congestion control scheme that coalesces state updates to adaptively meet bandwidth and processing constraints of stragglers. This congestion control can bound how far behind any given instance is from others. Increasing coalescing can reduce bandwidth requirements, but state updates fall further behind.

Multicast replication: During a middlebox operation, Constellation records state updates into a log store. For each state object, the log store allocates a queue of *log records* and maintains a sequence number to track these records. A log record tracks an operation by recording the local order at which it is performed. Specifically, a log record denotes that an operation was performed on a set of operands at a particular sequence number. For example, for an IDS’s port counter, log record $(inc, \{22\}, 7)$ shows that a counter was incremented for SSH port 22 at sequence number 7.

The state synchronization component allocates a multicast group per state object. Via this group, middlebox instances share and exchange their local log records of the object. For convergence, all log records must be delivered and applied to other instances. Log records are released once delivered.

Constellation runs two threads at the sender and receiver sides to transmit log records. Send threads use multicast to send local log records to other instances. Receive threads receive and apply log records, send acknowledgements, and prunes log records delivered to all other instances.

For each state object, the send thread continuously retrieves outstanding log records, i.e., log records that have not yet replicated in other instances, from the queue of this object, creates a *state message* from the records, and sends the message to the associated multicast group. The send thread round-robins between queues belonging to different objects.

A state message carries a list of log records, and an *acknowledgement* vector. Each instance maintains the acknowledgement vector to track log records received from other instances. This vector contains the highest sequence numbers seen for each instance.

The receive thread uses the acknowledgement vector to track external log records. Upon receiving a state message, the receive thread applies the operations from the message,

and updates the instance’s acknowledgement vector. The receive thread increments a sequence number of the vector when all log records up to and including this sequence number have been received.

Since operations are idempotent and commutative, state objects converge even when applying duplicate or out of order log records [66]. The receive thread can be also configured to apply log records in order. This *ordering property* provides idempotence for an easier implementation of state objects that are not intrinsically idempotent. We used this property to provide idempotence for count operation of counting bloom filter and count-min sketch in § 4.1.

The receive thread also prunes the log store according to received state messages. For each object, this thread records the last acknowledgement that it has received from other instances. The smallest acknowledgement shows the latest log record that has been replicated in all instances. Accordingly, the receive thread prunes all log records up to the smallest acknowledgement.

We provide reliable multicast by retransmitting lost log records due to packet drops. If a log record has not been acknowledged, the send thread retransmits the record after a timeout based on the maximum round trip time of any instance. If a multicast channel is idle, the send thread periodically transmits keep-alive messages containing the latest acknowledgement vector.

Adaptive bandwidth optimization: There are cases when middlebox instances may fall behind in replication due to transient events, such as temporary congestion in the network. In these cases, Constellation uses *coalesced* log records instead of sending individual log records. Coalescing can significantly reduce the bandwidth usage of state replication.

The idempotence and commutativity properties of state objects allow Constellation to coalesce related log records, i.e., the records of state operations modifying the same object. For example, a series of increments and decrements to a single counter can be represented as adding the sum of the operations. To coalesce related operations, the send thread calls back into to the associated state object.

Constellation detects instances that are falling behind by monitoring the round trip time (RTT) of each instance in the multicast group. The receive thread measures the minimum and average RTTs for each instance using acknowledgements. When the average RTT is higher than the minimum RTT by a set threshold, the instance is marked as congested.

Upon detection, the send thread starts to send coalesced log records using an adaptive *lookahead window* based on RTT. The instance continuously monitors the average RTT to increase or decrease the lookahead window. The larger the lookahead window, the more coalescing opportunity.

Constellation adjusts the lookahead window to trade-off the bandwidth reduction from coalescing and the increased synchronization latency from delaying the transmission of log records. The lookahead window size is set based on the throughput. An instance coalesces log records up to a maximum lookahead or when an acknowledgement is received. If the acknowledgement arrives early, the instance immediately transmits any already coalesced updates, which effectively reduces the lookahead size.

4.3 Dynamic Scaling

A scaling event changes the members of multicast groups and consequently impacts state replication. For a correct replication during and after this membership change, Constellation ensures three properties: i) *unique identification*: with a new set of members, each active instance remains uniquely identifiable; ii) *membership agreement*: instances agree upon active members so that their send and receive threads can work in harmony; and iii) *convergence*: the new set of members still remain convergent for all state objects.

We assume that the orchestrator is fault tolerant. If a failure occurs during a scaling event, the orchestrator reliably detects and notifies all instances of the change.

Scale-out event: A new instance joins replication groups and copies a snapshot of middlebox state from an existing instance before starting to process traffic. Adding a new instance is broken down into four steps.

First, the orchestrator deploys a new instance with unique identifier. To ensure uniqueness, it is sufficient that the orchestrator generates a new identifier or reuses the identifier of a removed instance.

Second, the new instance joins the replication groups and starts to record state messages. It sends a `join` message containing its identifier to the multicast groups to announce that its joining. Existing instances confirm receiving a `join` message by adding the instance to the acknowledgement vector of its state messages. Upon receiving this confirmation, the new instance starts acknowledging state messages received from existing instances. It does so by sending empty state messages with an acknowledgement vector. The new instance retransmits the `join` message until all existing instances confirm receiving the message. This ensures the membership agreement property.

Third, the new instance requests an existing instance for a snapshot of the state and metadata (includes the acknowledgement vector of each state object). The existing instance executes a `fork` system call [6] to duplicate its process to take a state snapshot and transmit it to the new instance. For the snapshot consistency, `fork` is synchronized with packet processing and state synchronization.

Fourth, upon receiving the state snapshot, the new instance applies log records from state messages recorded since

the second step. Lastly, the new instance notifies the orchestrator to redistribute traffic load to it.

Taking snapshots using `fork` is fast, since memory is not immediately copied. The memory is marked as copy-on-write; the operating system copies memory after the original or child process modifies it. Constellation further reduces this overhead by using `madvise` [8]. Since the duplicated process does not process incoming packets, Constellation tells the operating system to exclude memory pages reserved for receiving incoming packets. This significantly reduces the pause time of `fork`.

Scale-in event: Another benefit of Constellation is that it can scale-in with no state loss and virtually zero packet loss. Removing an excess instance takes four steps.

First, the orchestrator reroutes the traffic load of the excess instance to other instances. Due to state replication, other instances have the state necessary to process this load. Second, the orchestrator notifies the excess instance, and this instance waits for some time for remaining inflight traffic to arrive. Then, the instance drains its outstanding log records to ensure convergence. Third, the excess instance sends a leave message to a multicasting group to announce that it is leaving. The instance will retry until all instances acknowledge receiving this message, which ensures membership agreement. Finally, once all other instances have acknowledged the leave message, the excess instance notifies the orchestrator to reclaim all resources.

5 Implementation and Experience

We built Constellation using the Click modular router [48]. It consists of 6141 SLOC for the runtime and 2155 SLOC for the middlebox implementations. We discuss our development experience in using our system compared to existing frameworks that provide synchronous middlebox state management. We dive into the implementation of a flow table and a NAT. Lastly, we discuss the artifacts caused by the use of asynchronous replication.

Convergent flow table: A *flow table* is used to track network flows and has application in several middleboxes as shown in Table 2. A flow table is a mapping keyed on a hashing of packet headers with values that can be network addresses or some attributes regarding the flows.

A flow table supports `add` and `value` operations. The `add(k, v)` operation either inserts flow `k` and value `v`, or updates the value of flow `k` with `v`. The `value(k)` operation returns a value associated with key `k`.

For convergence, we focus on `add` operation, since `value` does not mutate the state. The `add` operation is idempotent but not commutative. Enforcing a deterministic ordering on concurrent `add` operations “artificially” makes `add` commutative. Specifically, a global ordering across the middlebox

cluster determines the winner in a race between two add operations modifying the same key k .

The object exposes a callback that allows developers to customize this ordering. By default, the object uses a numerical comparison, where (k, v) wins against (k, v') only if the binary value of v is greater than that of v' .

5.1 Network Address Translator

A NAT bridges two address spaces [21, 42, 72]. NAT instances share two state objects. Using a flow table object, a NAT instance maps traffic flows coming from one address space to another address space. A NAT instance identifies each flow by a unique port number from an available port pool object [72]. Both flow table and port pool are updated together, thus we can use Constellation’s derivative state object feature to support multi object operations on these two objects.

In rare incidents, due to asynchronous local accesses, NAT instances may concurrently allocate an identical port number for different flows. This violates the NAT’s *unique port assignment* invariant, and flow translations may collide.

Constellation’s convergent flow table enable us to resolve this inconsistency. We use its callback so that among two collided flows, a flow with larger numerical value of its five tuple wins the race enabling all instances to converge.

5.2 Artifacts of Asynchronous Replication

Asynchronous replication may cause middleboxes to experience temporary inconsistencies until instances converge. We study a number of middleboxes including the ones shown in Table 2 for their possible artifacts. There are three categories of artifacts: lag or reduced precision; packet loss; and duplicates and collisions. In practice, these artifacts are non-issues, as they are rare and are already mitigated by existing protocols or end user applications. Our design makes the tradeoff of dealing with small artifacts for substantial performance gains on both local area networks and wide area networks.

Lag or reduced precision: The most common problem for most middleboxes is that asynchronous replication induces a lag in measurement or reduces the measurement precision. For example, an IPS may set a threshold for when it blocks traffic and may lag by approximately the round-trip time between instances. A distributed rate limiter may be imprecise in its ability to set a limit, but for longer flows it can still maintain a tight bounded error.

Packet loss: Packet loss issues can arise for stateful firewalls, NATs and load balancers. This may occur when traffic passes through a different instance while a connection is being established but before state is synchronized between instances. For example in Figure 2, the second firewall instance may receive the response traffic before its state is

synchronized with the first instance. Since most protocols will retry dropped packets, this artifact will only result in a small increase in latency.

Duplicates and collisions: NATs and load balancers may also suffer from collisions or duplicate mappings, if two instances simultaneously generate conflicting mappings. For example, two NAT instances might reuse the same public IP and port for two connections to the same destination IP and port. In this scenario we can terminate one connection and have the client to reconnect. For even very large networks this is exceedingly rare, and disconnects from other issue sources would be orders of magnitude more common.

An alternative approach is to design a NAT with an extra table to allow instances to acquire leases on regions of the public IP and port space. When a connection arrives the instance would allocate out of one of these pools, thus preventing collisions. This would require the middlebox to eagerly reserve new ranges when it is running low on the current pool.

6 Evaluation

We start with a description of our setup and methodology in § 6.1, and then we measure the overhead of Constellation framework in § 6.2. We measure Constellation’s performance during its normal operation and dynamic scaling in § 6.3 and § 6.4, respectively. Then in § 6.6, we measure the impact of Constellation’s artifacts in our IDPS example. Finally, we discuss the implementation complexity in § 6.7.

6.1 Experimental Setup and Methodology

We compare Constellation with S6 [80] and Sharded. S6 is the-state-of-art in elastic scaling of middleboxes. We use the publicly available implementation of S6 [79]. A S6’s middlebox application runs as a process that uses DPDK toolkit [45]. Sharded is a baseline system used to measure the performance upper bound, as middlebox traffic is sharded with no shared state. Moreover, we use two middleboxes, a NAT and an IDPS. Our implementation of IDPS includes only the port-scan detection/mitigation functionality.

We use a server cluster each equipped with a single Intel D-1540 Xeon with 8 cores and 64 GiB of memory. The servers are connected with an Intel Ethernet Connection X557 10 Gbps NIC to a Supermicro SSE-X3348T switch. A separate 10 Gbps Mellanox ConnectX-3 NIC connected to a Mellanox switch is used as the *state channel* for state synchronization. All servers run Ubuntu 18.04.

We use MoonGen [33] to generate traffic and measure performance. Traffic from a generator server is sent through a middlebox instance then back to the generator. We measure latency and total throughput at the traffic generators. The packet size in our experiments is 64 B. MoonGen measures end-to-end latency by sending timestamped 128 B packets

	Toolkit	No Op.	Reference	Read	Write	R+W
S6	11.80	5.96	3.66	2.52	2.08	1.38
Constellation	10.00	9.28	N/A	9.26	9.20	9.20

Table 3: Throughput of a pass-through middlebox in Mpps: “R+W” denotes read and write. S6 runs on DPDK, while Constellation uses DPDK+Click adding overhead to the toolkit baseline. Reference adds the overhead of finding which instance owns a state object. We measure the read and write costs separately and together. The performance difference between Constellation and S6 is the *cost of synchronization*.

while it simultaneously sends load of 64 B packets. We also developed a tool to accurately timestamp received packets at microsecond granularity, which allows us to accurately measure throughput changes. Using this tool, we capture the impact of Constellation’s dynamic scaling in § 6.4.

Unless stated otherwise, we run 5 second experiments and repeat each experiment 10 times. The confidence intervals of our results are all within 5%. As a result, we do not report them in our plots.

6.2 Performance Breakdown

Table 3 shows a performance breakdown for a pass through middlebox operating on a counter object. Using this middlebox, we benchmark S6 and Constellation to breakdown the performance cost of common middlebox operations. We configure the middlebox to either perform no operation, or perform a read, a write, or a read and write per packet.

S6 runs directly on DPDK, while Constellation is built using DPDK+Click which reduces baseline throughput by ~ 15%. The no-operation measurement shows the cost from the S6 and Constellation frameworks. S6 process each packet in a separate microthread, built using Boost coroutines [4] to allow an independent microthread to process a packet while another microthread is blocked on a remote state access. Context switching between microthreads results in a loss of 49% of its performance.

The remaining columns measure the cost of reading and writing shared state. The reference column measures the time required for S6 to discover which instance owns the key of a flow. The read and write costs are measured separately and together. S6 slows down by a further 76% percent over the no-operation column, excluding the cost of microthreads.

6.3 Performance in Normal Operation

We measure the maximum aggregated throughput and the end-to-end latency of NAT and IDPS. For wide area experiments, we deploy our NAT instances in a simulated WAN.

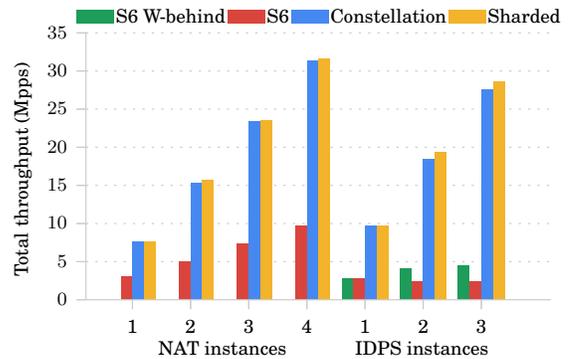


Figure 4: Total throughput of middleboxes in LAN: Compared to linear scaling, Constellation is within 2–4% for NAT and 1–5% for IDPS.

Using tc [18], we configure the servers running instances to artificially add WAN latency [26, 49, 55] to the state channel. For both our LAN and WAN, we use a traffic load where each NAT instance receives 2000 new flows per second.

Throughput in LAN. Figure 4 shows the maximum aggregated throughput of the NAT and IDPS instances deployed in our LAN. For IDPS, we configure S6 in two modes. In the first mode, labeled “S6,” the state updates are immediately synchronized. In the second mode, labeled “S6 W-behind,” the remote counters are updated by a 10 ms delay.

As shown for both middleboxes, Constellation’s throughput scales linearly with increasing the number of instances, within 2–4% of the ideal scaling for NAT and within 1–5% for the IDPS. For S6, per instance throughput of the NAT drops up to 21% due to the overhead of state synchronization. The throughput of IDPS drops for S6 and flattens for “S6 write-behind” going from 2 to 3 instances. In the S6 system, each instance has to query other instances to retrieve the values of their state objects. IDPS instances pay this overhead once every few packets (i.e., 50% and 66% of packets for 2 and 3 IDPS instances), while NAT instances incur this cost once every few flows (i.e., 50% and 66% of flows for 2 and 3 IDPS instances).

Compared to S6 for NAT, Constellation improves throughput by 2.5–3.2× and is within 2–4% of Sharded’s aggregated throughput. For IDPS, Constellation achieves a 3.4–6.3× and 3.4–11.2× higher throughput compared to that of “S6 write-behind” and S6, respectively.

Throughput in WAN. We evaluate the impact of the state channel with WAN delay on the NAT throughput. As shown in Figure 5, the aggregated throughput of Constellation’s NAT is independent of the WAN delay of the state channel. However, S6’s throughput drops significantly. Compared to our LAN measurements, Constellation’s WAN throughput is within 2–3% of its LAN throughput, while S6 becomes 6

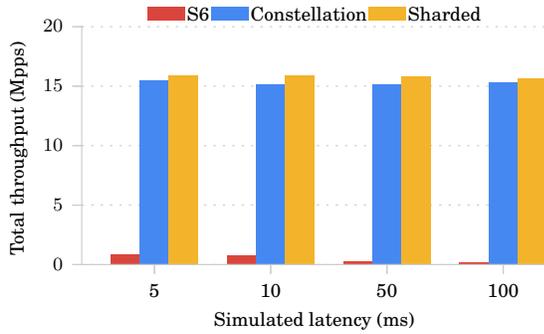


Figure 5: Total throughput of 2 NAT instances in WAN: Constellation’s throughput is largely independent of latency, but synchronous accesses to remote state slow down S6’s throughput by 6 \times to 32 \times going from 5 to 100 ms latency.

	1 instance	2 instances	3 instances
S6	21 \pm 1 μ s	25 \pm 1 μ s	26 \pm 1 μ s
Constellation	31 \pm 1 μ s	44 \pm 3 μ s	46 \pm 2 μ s
Sharded	31 \pm 1 μ s	32 \pm 1 μ s	34 \pm 2 μ s

Table 4: NAT average latency: Constellation’s latency remains constant going from 2 to 3 instances. Its latency increase going from 1 to 2 is due to Click’s scheduling overhead.

to 32 \times slower. Constellation’s throughput is 17 to 96 \times of S6’s and is within 2–5% of Sharded’s.

Constellation accesses the state locally and does not perform immediate state synchronization when a NAT instance writes or queries the state of flows. This asynchrony allows Constellation’s NAT instances to operate at the same throughput level over the WAN as its local area network. On the other hand, S6’s middlebox instances access state stored in a distributed hash table. Due to state distribution in this hash table, an instance owns a half of the state and must remotely query the other instance to operate on the other half. The overhead of this synchronous remote access is the root cause of S6’s performance drop.

Latency in LAN. Table 4 presents the average end-to-end latency of the NAT in our LAN. For a fair comparison, the NAT instances are under S6’s sustainable load of 1 Mpps with 2 k new flows per second. Going from one middlebox instance to two or more instances, both S6 and Constellation enable their state synchronization mechanisms between instances.

As shown in Table 4, going from 2 to 3 instances, Constellation’s latency overhead does not increase. Compared to Sharded, Constellation adds 12 μ s overhead. In our implementation, the receive thread of the state synchronization

and the middlebox logic run on the same processor core, and we use Click’s scheduler [48] to schedule them. The latency increase from 1 to 2 instances is due the overhead of Click scheduler [48]. S6’s latency slightly increases going from 1 to 2 and 3 instances. Its latency is lower than Sharded, since S6’s middleboxes run on DPDK, while Sharded’s middleboxes uses DPDK+Click which adds Click’s overhead to the baseline DPDK (recall from § 6.2).

To investigate the latency overhead in more details we report the latency of the first instance of the 2 NAT instances in Figure 6 under different traffic loads. Figure 6a shows that average latency remains steady for all systems as the traffic load increases until they approach their respective saturation points. Near these points, packets start to be queued, and latency rapidly spikes. The average latency of Constellation and Sharded remain under 209 μ s, while S6’s average latency spikes up to 500 μ s. As shown in Figure 6b, 99-percentile latency has a trend similar to that of the average latency. Constellation’s and Sharded’s peak latency values are 451 μ s and 539 μ s, and S6 exhibits a peak latency of up to 2.1 ms.

6.4 Dynamic Scaling

We use our cluster of two NAT instances to quantify the performance of Constellation during dynamic scaling. Each instance is under a 5 Mpps load with 2 k new flows per second. We use our tool to measure throughput at microsecond scale resolution. We discuss only the performance of the instance that is involved in state transmission to the new instance. This instance resides in the same local area network as the new instance.

As shown in Figure 7, the first instance does not experience notable throughput degradation. Packet drop is also zero. Excluding unnecessary memory pages from copy-on-write protection allows fork to complete in only a fraction of a millisecond. In a separate experiment, not shown here, we measured that fork lasts for 10 ms without this optimization.

Once packet processing resumes, we observe a throughput burst for packets queued during fork pause time. During state transmission, throughput temporarily fluctuates. This is due to state updates in processing the first packets of new flows, since they write into copy-on-write memory pages containing the state and incur memory copying overheads.

6.5 Coalescing Benefits

We measure the bandwidth saving by coalescing state updates of a counting bloom filter and a count-min sketch to evaluate coalescing benefits. We use network traces from a backbone network [9] where we use only valid IP packets.

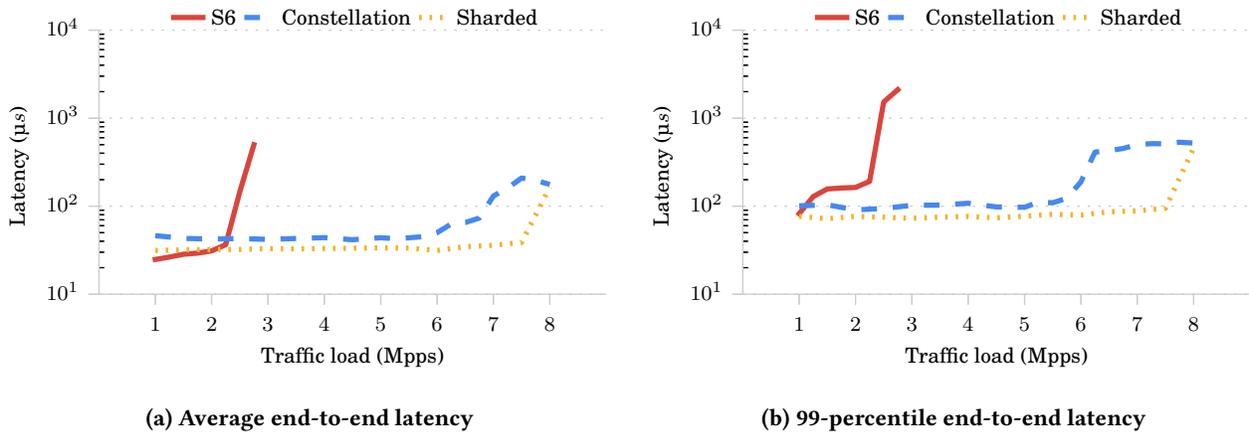


Figure 6: End-to-end Latency of the first instance of 2 NAT instances deployed in our LAN: Constellation’s average and 99-percentile latency remain steady under these traffic loads. When Constellation approaches the saturation point, its latency increases up to 0.2 and 0.5 ms for the average and 99 percentile, respectively.

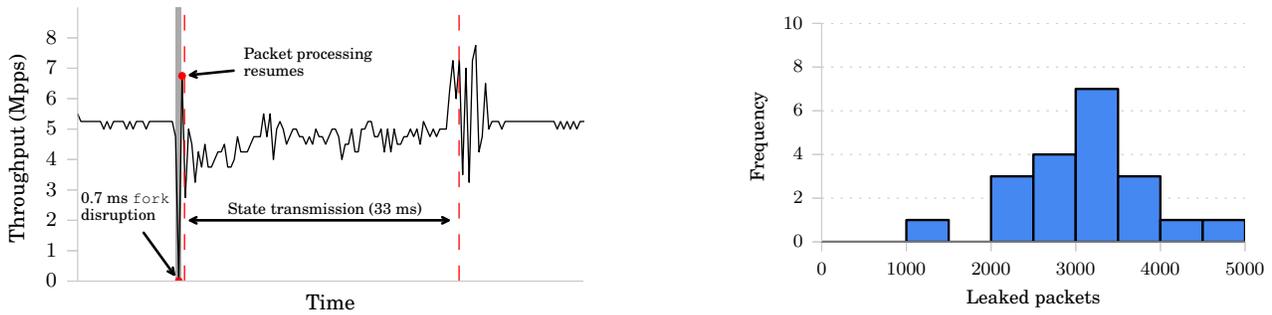


Figure 7: Throughput of the NAT instance transmitting its state in a scale-out event: The instance experiences a sub-millisecond throughput disruption during fork. The throughput becomes unsteady for few milliseconds during state transmission of 20 k concurrent flows.

6.6 Inconsistency Artifacts

We measure the impact of Constellation’s asynchronous replication in mitigating flooding a port number. We deploy two IDPS instances in a simulated WAN with 5 ms delay and configure them with a mitigation policy as follows. An instance blocks traffic destined to a port number if the traffic volume passes the threshold of 1024 Mbits. Two traffic generators flood the instances at 1 Mpps. For each instance, we measure the number of its *leaked packets*, i.e., the number of packets that pass through an instance in the distributed deployment compared with a theoretical centralized IDPS with infinite packet processing that receives the aggregated traffic and filters packets after crossing a given threshold.

Figure 8: IDPS leaking packets due to asynchronous replication: Histogram shows the number of packets leaked beyond the target threshold for Constellation’s IDPS.

Figure 8 shows a histogram of the number of leaked packets. Asynchronous replication delays blocking the attack. Constellation IDPS reacts to the flood within 5 ms and leaks on average 3.2 k packets.

Delaying an IDPS response by a few milliseconds is a good trade-off as it allows the system to keep up with the throughput demands of high speed networks. Previous work has shown that IDSEs unable to keep up with the traffic can be bypassed to successfully launch attacks [29, 60, 70].

6.7 Development Complexity

Comparing the complexity of different middlebox frameworks is extremely difficult. Using the lines of code, we provide a rough estimation of the complexity.

We compared the code of the NAT implemented in S6 to the one implemented in Constellation as described in § 5.

Both NATs are roughly the same size, with 361 lines of code for Constellation and 283 lines for S6. We measured S6's code before the source-to-source translator that provides syntactic sugar to simplify the system implementation. This result illustrates that it is not significantly difficult to build middleboxes with asynchronous replication.

7 Related Work

We have discussed existing frameworks that support state sharing for general middleboxes in § 2.2. As mentioned, they are not optimized for wide area networks. Next, we compare Constellation with two other lines of related work.

Middlebox specific frameworks: Some systems are highly specialized for particular middleboxes. Most of them deploy middlebox instances as shards with no shared state [21, 32, 37, 56]. Unlike others, vNIDS [51], a microservice based network intrusion detection system, and Yoda [37], an application layer load balancer, share their state in a central data store.

Database replication protocols: Two phase commit is a common protocol to replicate transactions in distributed databases. The protocol supports any transaction using synchronous coordination; however, it does not scale for a geo-distributed replication, since a transaction involves multiple rounds of message passing between a coordinator and replicas in different sites.

Multi datacenter consistency (MDCC) [49] optimizes performance by involving a coordinator only when transactions conflict. MDCC also optimizes commutative transactions with domain integrity constraints (e.g., a bank account balance must remain non-negative with concurrent deposits and withdrawals) by involving a coordinator only when concurrent transactions may violate constraints (e.g., the account balance is close to become zero).

Highly available databases relax generality of transactions or consistency guarantees for higher scalability and performance. Some systems split data into shards and restrict updates to only a single shard. Eventually consistent databases [30] allow asynchronous state replication with complex resolution mechanisms to resolve conflicts; however, these mechanisms can cause consistency anomalies.

8 Conclusions

WAN latency can significantly impact the performance of a stateful middlebox whose instances are deployed across a WAN. We introduced Constellation, a framework for the geo-distributed middleboxes. Using asynchronous state replication of convergent state objects, Constellation achieves high performance and scalability. Our results show that Constellation can improve middlebox performance by almost two orders of magnitude compared to the state-of-the-art [80].

References

- [1] [n. d.]. Asymmetric Routing and Firewalls. <http://brbccie.blogspot.com/2013/03/stateful-nat-with-asymmetric-routing.html>. [Online].
- [2] [n. d.]. Asymmetric routing with multiple network paths. <https://docs.microsoft.com/en-us/azure/expressroute/expressroute-asymmetric-routing>. [Online].
- [3] [n. d.]. Barracuda CloudGen Firewall. <https://www.barracuda.com/products/cloudgenfirewall>. [Online].
- [4] [n. d.]. boost.coroutine. <https://github.com/boostorg/coroutine>. [Online].
- [5] [n. d.]. Contrail SD-WAN. <https://www.juniper.net/us/en/products-services/sdn/contrail>. [Online].
- [6] [n. d.]. fork - create a child process. <http://man7.org/linux/man-pages/man2/fork.2.html>. [Online].
- [7] [n. d.]. HAProxy Load Balancer's development branch. <https://github.com/haproxy/haproxy>. [Online].
- [8] [n. d.]. madvise - give advice about use of memory. <http://man7.org/linux/man-pages/man2/madvise.2.html>. [Online].
- [9] [n. d.]. MAWI Working Group Traffic Archive. <https://mawi.wide.ad.jp/mawi/samplepoint-F/2020/202002161400.html>. [Online].
- [10] [n. d.]. NPF: packet filter with stateful inspection, NAT, IP sets, etc. <http://rmind.github.io/npf/>. [Online].
- [11] [n. d.]. ONAP Architecture Overview. https://www.onap.org/wp-content/uploads/sites/20/2018/11/ONAP_CaseSolution_Ar. [Online].
- [12] [n. d.]. ONOS E-CORD Proof of Concept Demonstrates Open Disaggregated ROADM. <https://www.fujitsu.com/us/Images/ONOS-E-CORD-use-case.pdf>. [Online].
- [13] [n. d.]. Open-sourcing Katran, a scalable network load balancer. <https://github.com/facebookincubator/katran>. [Online].
- [14] [n. d.]. pf: packet filter. <https://man.openbsd.org/pf.4>. [Online].
- [15] [n. d.]. Snort++. <https://github.com/snort3/snort3>. [Online].
- [16] [n. d.]. SQUID Frequently Asked Questions - Memory. <http://www.comfsm.fm/computing/squid/FAQ-8.html>. [Online].
- [17] [n. d.]. Squid: Optimising Web Delivery. <http://www.squid-cache.org>. [Online].
- [18] [n. d.]. tc - show / manipulate traffic control settings. <https://linux.die.net/man/8/tc>. [Online].
- [19] [n. d.]. The Zeek Network Security Monitor. <https://github.com/zeek/zeek>. [Online].
- [20] Amazon. [n. d.]. Elastic Load Balancing. <https://aws.amazon.com/elasticloadbalancing/>. [Online].
- [21] Fabien André, Stéphane Gouache, Nicolas Le Scouarnec, and Antoine Monsifrot. 2018. Don'T Share, Don'T Lock: Large-scale Software Connection Tracking with Krononat. In *Proceedings of the 2018 USENIX Conference on Usenix Annual Technical Conference (USENIX ATC '18)*. USENIX Association, Berkeley, CA, USA, 453–465. <http://dl.acm.org/citation.cfm?id=3277355.3277400>
- [22] Azeem Aqil, Karim Khalil, Ahmed O.F. Atya, Evangelos E. Papalexakis, Srikanth V. Krishnamurthy, Trent Jaeger, K. K. Ramakrishnan, Paul Yu, and Ananthram Swami. 2017. Jaal: Towards Network Intrusion Detection at ISP Scale. In *Proceedings of the 13th International Conference on Emerging Networking EXperiments and Technologies (CoNEXT '17)*. ACM, New York, NY, USA, 134–146. <https://doi.org/10.1145/3143361.3143399>
- [23] P Ayuso. 2006. Netfilter's connection tracking system. *login* 31, 3 (2006).
- [24] Bevin Brett. 2016. Memory Performance in a Nutshell. <https://software.intel.com/en-us/articles/memory-performance-in-a-nutshell>.

- [Online].
- [25] Andrei Broder and Michael Mitzenmacher. 2004. Network Applications of Bloom Filters: A Survey. *Internet Mathematics* 1, 4 (2004), 485–509. <https://doi.org/10.1080/15427951.2004.10129096> arXiv:<https://doi.org/10.1080/15427951.2004.10129096>
- [26] Chen Chen, Changbin Liu, Pingkai Liu, Boon Thau Loo, and Ling Ding. 2015. A Scalable Multi-datacenter Layer-2 Network Architecture. In *Proceedings of the 1st ACM SIGCOMM Symposium on Software Defined Networking Research (SOSR '15)*. ACM, New York, NY, USA, Article 8, 12 pages. <https://doi.org/10.1145/2774993.2775008>
- [27] Chen Chen, Changbin Liu, Pingkai Liu, Boon Thau Loo, and Ling Ding. 2015. A Scalable Multi-datacenter Layer-2 Network Architecture. In *Proceedings of the 1st ACM SIGCOMM Symposium on Software Defined Networking Research (SOSR '15)*. ACM, New York, NY, USA, Article 8, 12 pages. <https://doi.org/10.1145/2774993.2775008>
- [28] P. Coelho and F. Pedone. 2018. Geographic State Machine Replication. In *2018 IEEE 37th Symposium on Reliable Distributed Systems (SRDS)*. 221–230. <https://doi.org/10.1109/SRDS.2018.00034>
- [29] Scott A. Crosby and Dan S. Wallach. 2003. Denial of Service via Algorithmic Complexity Attacks. In *Proceedings of the 12th Conference on USENIX Security Symposium - Volume 12 (SSYM'03)*. USENIX Association, Berkeley, CA, USA, 3–3. <http://dl.acm.org/citation.cfm?id=1251353.1251356>
- [30] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. 2007. Dynamo: Amazon’s Highly Available Key-Value Store. *SIGOPS Oper. Syst. Rev.* 41, 6 (Oct. 2007), 205–220. <https://doi.org/10.1145/1323293.1294281>
- [31] S. Dharmapurikar, P. Krishnamurthy, T. Sproull, and J. Lockwood. 2003. Deep packet inspection using parallel Bloom filters. In *11th Symposium on High Performance Interconnects, 2003. Proceedings.* 44–51. <https://doi.org/10.1109/CONNECT.2003.1231477>
- [32] Daniel E. Eisenbud, Cheng Yi, Carlo Contavalli, Cody Smith, Roman Kononov, Eric Mann-Hielscher, Ardas Cilingiroglu, Bin Cheyney, Wentao Shang, and Jinnah Dylan Hosein. 2016. Maglev: A Fast and Reliable Software Network Load Balancer. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*. Santa Clara, CA, 523–535. <https://www.usenix.org/conference/nsdi16/technical-sessions/presentation/eisenbud>
- [33] Paul Emmerich, Sebastian Gullenmüller, Daniel Raumer, Florian Wohlfart, and Georg Carle. 2015. MoonGen: A Scriptable High-Speed Packet Generator. In *Proceedings of the 2015 Internet Measurement Conference (IMC '15)*. ACM, New York, NY, USA, 275–287. <https://doi.org/10.1145/2815675.2815692>
- [34] ETSI. 2017. *NFV Whitepaper*. Technical Report. European Telecommunications Standards Institute. https://portal.etsi.org/NFV/NFV_White_Paper.pdf
- [35] Li Fan, Pei Cao, Jussara Almeida, and Andrei Z. Broder. 2000. Summary Cache: A Scalable Wide-area Web Cache Sharing Protocol. *IEEE/ACM Trans. Netw.* 8, 3 (June 2000), 281–293. <https://doi.org/10.1109/90.851975>
- [36] Mario Flajslik and Mendel Rosenblum. 2013. Network Interface Design for Low Latency Request-Response Protocols. In *Presented as part of the 2013 USENIX Annual Technical Conference (USENIX ATC 13)*. USENIX, San Jose, CA, 333–346. <https://www.usenix.org/conference/atc13/technical-sessions/presentation/flajslik>
- [37] Rohan Gandhi, Y. Charlie Hu, and Ming Zhang. 2016. Yoda: A Highly Available Layer-7 Load Balancer. In *Proceedings of the Eleventh European Conference on Computer Systems (EuroSys '16)*. ACM, New York, NY, USA, Article 21, 16 pages. <https://doi.org/10.1145/2901318.2901352>
- [38] Aaron Gember-Jacobson, Raajay Viswanathan, Chaithan Prakash, Robert Grandl, Junaid Khalid, Sourav Das, and Aditya Akella. 2014. OpenNF: Enabling Innovation in Network Function Control. In *Proceedings of the 2014 ACM Conference on SIGCOMM (SIGCOMM '14)*. ACM, New York, NY, USA, 163–174. <https://doi.org/10.1145/2619239.2626313>
- [39] M. G. Gouda and A. X. Liu. 2005. A model of stateful firewalls and its properties. In *2005 International Conference on Dependable Systems and Networks (DSN'05)*. 128–137. <https://doi.org/10.1109/DSN.2005.9>
- [40] Y. Gu, M. Shore, and S. Sivakumar. 2013. A Framework and Problem Statement for Flow-associated Middlebox State Migration. <https://tools.ietf.org/html/draft-gu-statemigration-framework-03>.
- [41] Guanyao Huang, A. Lall, C. Chuah, and Jun Xu. 2009. Uncovering global icebergs in distributed monitors. In *2009 17th International Workshop on Quality of Service*. 1–9. <https://doi.org/10.1109/TWQoS.2009.5201394>
- [42] T. Hain. 2000. *Architectural Implications of NAT*. RFC 2993. RFC Editor. 1–29 pages. <http://www.rfc-editor.org/rfc/rfc2993.txt>
- [43] Marios Iliofotou, Michalis Faloutsos, and Michael Mitzenmacher. 2009. Exploiting Dynamicity in Graph-based Traffic Analysis: Techniques and Applications. In *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies (CoNEXT '09)*. ACM, New York, NY, USA, 241–252. <https://doi.org/10.1145/1658939.1658967>
- [44] Marios Iliofotou, Prashanth Pappu, Michalis Faloutsos, Michael Mitzenmacher, Sumeet Singh, and George Varghese. 2007. Network Monitoring Using Traffic Dispersion Graphs (Tdgs). In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC '07)*. ACM, New York, NY, USA, 315–320. <https://doi.org/10.1145/1298306.1298349>
- [45] DPK Intel. 2015. Data plane development kit.
- [46] Murad Kablan, Azzam Alsudais, Eric Keller, and Franck Le. 2017. Stateless Network Functions: Breaking the Tight Coupling of State and Processing. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. USENIX Association, Boston, MA, 97–112. <https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/kablan>
- [47] Junaid Khalid and Aditya Akella. 2019. Correctness and Performance for Stateful Chained Network Functions. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*. USENIX Association, Boston, MA, 501–516. <https://www.usenix.org/conference/nsdi19/presentation/khalid>
- [48] Eddie Kohler, Robert Morris, Benjie Chen, John Jannotti, and M. Frans Kaashoek. 2000. The Click Modular Router. *ACM Trans. Comput. Syst.* 18, 3 (Aug. 2000), 263–297. <https://doi.org/10.1145/354871.354874>
- [49] Tim Kraska, Gene Pang, Michael J. Franklin, Samuel Madden, and Alan Fekete. 2013. MDCC: Multi-data Center Consistency. In *Proceedings of the 8th ACM European Conference on Computer Systems (EuroSys '13)*. ACM, New York, NY, USA, 113–126. <https://doi.org/10.1145/2465351.2465363>
- [50] Abhishek Kumar, Minh Sung, Jun (Jim) Xu, and Jia Wang. 2004. Data Streaming Algorithms for Efficient and Accurate Estimation of Flow Size Distribution. In *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '04/Performance '04)*. ACM, New York, NY, USA, 177–188. <https://doi.org/10.1145/1005686.1005709>
- [51] Hongda Li, Hongxin Hu, Guofei Gu, Gail-Joon Ahn, and Fuqiang Zhang. 2018. vNIDS: Towards Elastic Security with Safe and Efficient Virtualization of Network Intrusion Detection Systems. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS '18)*. ACM, New York, NY, USA, 17–34. <https://doi.org/10.1145/3243734.3243862>

- [52] Zaoxing Liu, Ran Ben-Basat, Gil Einziger, Yaron Kassner, Vladimir Braverman, Roy Friedman, and Vyas Sekar. 2019. Nitrosketch: Robust and General Sketch-based Monitoring in Software Switches. In *Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM '19)*. ACM, 334–350. <https://doi.org/10.1145/3341302.3342076>
- [53] Y. Lu, B. Prabhakar, and F. Bonomi. 2006. Perfect Hashing for Network Applications. In *2006 IEEE International Symposium on Information Theory*. 2774–2778. <https://doi.org/10.1109/ISIT.2006.261567>
- [54] J. Ma, F. Le, A. Russo, and J. Lobo. 2015. Detecting distributed signature-based intrusion: The case of multi-path routing attacks. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. 558–566. <https://doi.org/10.1109/INFOCOM.2015.7218423>
- [55] Hatem Mahmoud, Faisal Nawab, Alexander Pucher, Divyakant Agrawal, and Amr El Abbadi. 2013. Low-latency Multi-datacenter Databases Using Replicated Commit. *Proc. VLDB Endow.* 6, 9 (July 2013), 661–672. <https://doi.org/10.14778/2536360.2536366>
- [56] Vladimir Olteanu, Alexandru Agache, Andrei Voinescu, and Costin Raiciu. 2018. Stateless Datacenter Load-balancing with Beamer. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI '18)*. USENIX Association, Renton, WA, 125–139. <https://www.usenix.org/conference/nsdi18/presentation/olteanu>
- [57] Christoph Paasch and Olivier Bonaventure. 2014. Multipath TCP. *Commun. ACM* 57, 4 (April 2014), 51–57. <https://doi.org/10.1145/2578901>
- [58] Shoumik Palkar, Chang Lan, Sangjin Han, Keon Jang, Aurojit Panda, Sylvia Ratnasamy, Luigi Rizzo, and Scott Shenker. 2015. E2: A Framework for NFV Applications. In *Proceedings of the 25th Symposium on Operating Systems Principles (SOSP '15)*. ACM, New York, NY, USA, 121–136. <https://doi.org/10.1145/2815400.2815423>
- [59] Diana Popescu, Noa Zilberman, and Andrew Moore. 2017. Characterizing the impact of network latency on cloud-based applications performance. <https://doi.org/10.17863/CAM.17588>
- [60] Thomas H. Ptacek and Timothy N. Newsham. 1998. *Insertion, evasion, and denial of service: Eluding network intrusion detection*. Technical Report. SECURE NETWORKS INC CALGARY ALBERTA.
- [61] Barath Raghavan, Kashi Vishwanath, Sriram Ramabhadran, Kenneth Yocum, and Alex C. Snoeren. 2007. Cloud Control with Distributed Rate Limiting. *SIGCOMM Comput. Commun. Rev.* 37, 4 (Aug. 2007), 337–348. <https://doi.org/10.1145/1282427.1282419>
- [62] Shriram Rajagopalan, Dan Williams, and Hani Jamjoom. 2013. Pico Replication: A High Availability Framework for Middleboxes. In *Proceedings of the 4th Annual Symposium on Cloud Computing (SOCC '13)*. ACM, New York, NY, USA, Article 1, 15 pages. <https://doi.org/10.1145/2523616.2523635>
- [63] Shriram Rajagopalan, Dan Williams, Hani Jamjoom, and Andrew Warfield. 2013. Split/Merge: System Support for Elastic Execution in Virtual Middleboxes. In *Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI '13)*. USENIX, Lombard, IL, 227–240. <https://www.usenix.org/conference/nsdi13/technical-sessions/presentation/rajagopalan>
- [64] Karthikeyan Ranganathan. [n. d.]. Netflix Shares Cloud Load Balancing And Failover Tool: Eureka! <https://bit.ly/2Xa9PBW>. [Online].
- [65] Robert Schweller, Ashish Gupta, Elliot Parsons, and Yan Chen. 2004. Reversible Sketches for Efficient and Accurate Change Detection over Network Data Streams. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement (IMC '04)*. ACM, New York, NY, USA, 207–212. <https://doi.org/10.1145/1028788.1028814>
- [66] Marc Shapiro, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. 2011. *A comprehensive study of Convergent and Commutative Replicated Data Types*. Research Report RR-7506. Inria – Centre Paris-Rocquencourt ; INRIA. 50 pages. <https://hal.inria.fr/inria-00555588>
- [67] Marc Shapiro, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. 2011. Conflict-Free Replicated Data Types. In *Stabilization, Safety, and Security of Distributed Systems*, Xavier Défago, Franck Petit, and Vincent Villain (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 386–400.
- [68] Justine Sherry, Sylvia Ratnasamy, and Justine Sherry At. 2012. A Survey of Enterprise Middlebox Deployments.
- [69] Vibhaalakshmi Sivaraman, Srinivas Narayana, Ori Rottenstreich, S. Muthukrishnan, and Jennifer Rexford. 2017. Heavy-Hitter Detection Entirely in the Data Plane. In *Proceedings of the Symposium on SDN Research (SOSR '17)*. ACM, New York, NY, USA, 164–176. <https://doi.org/10.1145/3050220.3063772>
- [70] R. Smith, C. Estan, and S. Jha. 2006. Backtracking Algorithmic Complexity Attacks against a NIDS. In *2006 22nd Annual Computer Security Applications Conference (ACSAC'06)*. 89–98. <https://doi.org/10.1109/ACSAC.2006.17>
- [71] Robin Sommer, Matthias Vallentin, Lorenzo De Carli, and Vern Paxson. 2014. HILTI: An Abstract Execution Environment for Deep, Stateful Network Traffic Analysis. In *Proceedings of the 2014 Conference on Internet Measurement Conference (IMC '14)*. ACM, New York, NY, USA, 461–474. <https://doi.org/10.1145/2663716.2663735>
- [72] P. Srisuresh and K. Egevang. 2001. *Traditional IP Network Address Translator (Traditional NAT)*. RFC 3022. RFC Editor. 1–16 pages. <http://www.rfc-editor.org/rfc/rfc3022.txt>
- [73] S. Staniford-Chen, S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, C. Wee, R. Yip, and D. Zerkle. 1996. GrIDS - A Graph-Based Intrusion Detection System for Large Networks. In *In Proceedings of the 19th National Information Systems Security Conference*. 361–370.
- [74] R. Stewart. 2007. *Stream Control Transmission Protocol*. RFC 4960. RFC Editor. 1–152 pages. <http://www.rfc-editor.org/rfc/rfc4960.txt>
- [75] Renata Teixeira, Aman Shaikh, Tim Griffin, and Jennifer Rexford. 2004. Dynamics of Hot-potato Routing in IP Networks. *SIGMETRICS Perform. Eval. Rev.* 32, 1 (June 2004), 307–319. <https://doi.org/10.1145/1012888.1005723>
- [76] Guido VAN ROOIJ. 2000. Real Stateful TCP Packet Filtering in IP Filter. *SANE 2000* (2000). <https://ci.nii.ac.jp/naid/10014926276/en/>
- [77] Shobha Venkataraman, Dawn Song, Phillip B Gibbons, and Avrim Blum. 2004. *New streaming algorithms for fast detection of superspreaders*. Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.
- [78] Limin Wang, Kyoung Soo Park, Ruoming Pang, Vivek Pai, and Larry Peterson. 2004. Reliability and Security in the CoDeeN Content Distribution Network. In *Proceedings of the Annual Conference on USENIX Annual Technical Conference (ATEC '04)*. USENIX Association, USA, 14.
- [79] Shinae Woo. [n. d.]. S6: Elastic Scaling of Stateful Network Functions. <https://github.com/NetSys/S6>. [Online].
- [80] Shinae Woo, Justine Sherry, Sangjin Han, Sue Moon, Sylvia Ratnasamy, and Scott Shenker. 2018. Elastic Scaling of Stateful Network Functions. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI '18)*. USENIX Association, Renton, WA, 299–312. <https://www.usenix.org/conference/nsdi18/presentation/woo>
- [81] Jiarong Xing, Wenqing Wu, and Ang Chen. 2019. Architecting Programmable Data Plane Defenses into the Network with FastFlex. In *Proceedings of the 18th ACM Workshop on Hot Topics in Networks (HotNets '19)*. ACM, New York, NY, USA, 161–169. <https://doi.org/10.1145/3365609.3365860>