SNAP: Stateful Network-Wide Abstractions for Packet Processing

Mina Tahmasbi Arashloo, Yaron Koral, Michael Greenberg, Jennifer Rexford, and David Walker

ABSTRACT
Early programming languages for software-defined networking (SDN) were built on top of the simple match-action paradigm offered by OpenFlow 1.0. However, emerging hardware and software switches offer much more sophisticated support for persistent state in the data plane, without involving a central controller. Nevertheless, managing stateful, distributed systems efficiently and correctly is known to be one of the most challenging programming problems. To simplify this new SDN problem, we introduce SNAP.

SNAP offers a simpler “centralized” stateful programming model, by allowing programmers to develop programs on top of one big switch rather than many. These programs may contain reads and writes to global, persistent arrays, and as a result, programmers can implement a broad range of applications, from stateful firewalls to fine-grained traffic monitoring. The SNAP compiler relieves programmers of having to worry about how to distribute, place, and optimize access to these stateful arrays by doing it all for them. More specifically, the compiler translates one-big-switch programs into an efficient internal representation based on a novel variant of binary decision diagrams. This internal representation discovers read-write dependencies, and constructs a mixed-integer linear program, which jointly optimizes the placement of state and the routing of traffic across the underlying physical topology. We have implemented a prototype compiler and applied it to about 20 SNAP programs over various topologies to demonstrate our techniques’ scalability.

1. INTRODUCTION
The first generation of programming languages for software-defined networks (SDNs) was built on top of OpenFlow 1.0, which offered simple match-action processing of packets. As a result, these systems were partitioned into (1) a stateless packet-processing part that could be analyzed statically, compiled, and installed on OpenFlow switches, and (2) a general stateful component that ran on the controller.

This “two-tiered” programming model can support any network functionality by running the stateful portions of the program on the controller and modifying the stateless packet-processing rules accordingly. However, simple stateful programs—such as detecting SYN floods or DNS amplification attacks—cannot be implemented efficiently because packets must go back-and-forth to the controller, incurring significant delay. Thus, in practice, stateful controller programs are limited to those that do not require per-packet stateful processing.

Today, however, SDN technology has advanced considerably: there is a raft of new proposals for switch interfaces that expose persistent state on the data plane, including those in P4 [6], OpenState [4], POF [35], and Open vSwitch [24]. Stateful programmable data planes enable us to offload programs that require per-packet stateful processing onto switches, subsuming a variety of functionality normally relegated to middleboxes. However, the mere existence of these stateful mechanisms does not make networks of these devices easy to program. In fact, programming distributed collections of stateful devices is typically one of the most difficult kinds of programming problems. We need new languages and abstractions to help us manage the complexity and optimize resource utilization effectively.

For these reasons, we have developed SNAP, a new language that allows programmers to mix primitive stateful operations with pure packet processing. However, rather than ask programmers to program a large, distributed collection of independent, stateful devices manually, we provide the abstraction that the network is one big switch (OBS). Programmers can allocate persistent arrays on that OBS, and do not have to worry about where or how such arrays are stored in the physical network. Such arrays can be indexed by fields in incoming packets and modified over time as network conditions change. Moreover, if multiple arrays must be updated simultaneously, we provide a form of network transaction to ensure such updates occur atomically. As a result, it is easy to write SNAP programs that learn about the network environment and record its state, store per-flow information or statistics, or implement a variety of stateful mechanisms.

While it simplifies programming, the OBS model, together with the stateful primitives, generates implementation challenges. In particular, multiple flows...
may depend upon the same stateful components. To process these flows correctly and efficiently, the compiler must simultaneously determine which flows depend upon which components, how to route those flows, and where to place the components. Hence, to map OBS programs to concrete topologies, the SNAP compiler discovers read-write dependencies between statements. It then translates the program into an xFDD, a variant of forwarding decision diagrams (FDDs) extended to incorporate stateful operations. Next, the compiler generates a system of integer-linear equations that jointly optimizes array placement and traffic routing. Finally, the compiler generates the switch-level configurations from the xFDD and the optimization results. We assume that the switches chosen for array placement support persistent state; other switches can still play a role in routing flows efficiently through the state variables. Our main contributions are:

- A stateful and compositional SDN programming language with persistent global arrays, a one-big-switch programming model, and network transactions. (See §2 for an overview and §4 for more technical details).

- Algorithms for compiling SNAP programs into low-level switch mechanisms (§3). (i) an algorithm for compiling SNAP programs into an intermediate representation that detects program errors, such as race conditions introduced by parallel access to stateful components using our extended forwarding decision diagrams (xFDD) and (ii) an algorithm to generate a mixed integer-linear program, based on the xFDD, which jointly decides array placement and routing while minimizing network congestion and satisfying the constraints necessary for network transactions.

- An implementation and evaluation of our language and compiler using about 20 applications. (§5 §6).

We discuss various data-plane implementations for SNAP, and how it relates to middleboxes in §7 discuss related work in §8 and conclude in §9.

2. SNAP SYSTEM OVERVIEW

This section overviews the key concepts in our language and compilation process using example programs.

2.1 Writing SNAP Programs

DNS tunnel detection. The DNS protocol is designed to resolve information about domain names. However, it is possible to use DNS messages as a tunnel to leak information since they are not typically intended for general data transfer. Detecting DNS tunnels is one of many real-world scenarios that require state to track the properties of network flows §5. The following steps can be used to detect DNS tunneling §5:

1. For each client, keep track of the IP addresses resolved by DNS responses.
2. For each DNS response, increment a counter. This counter tracks the number of resolved IP addresses that a client does not use.
3. When a client sends a packet to a resolved IP address, decrement the counter for the client.
4. Report tunneling for clients that exceed a threshold for resolved, but unused IP addresses.

Figure 1 shows a SNAP implementation of the above steps that detects DNS tunnels in the CS department subnet 10.0.6.0/24 (see Figure 2). Intuitively, a SNAP program can be thought of as a function that takes in a packet plus the current state of the network and produces a set of transformed packets as well as updated state. The incoming packet is read and written by referring to its fields (such as dstip and dns.rdata). The “state” of the network is read and written by referring to user-defined, array-based, global variables (such as orphan or susp-client). Before explaining the program in detail, note that it does not refer to specific network device(s) on which it is implemented. SNAP programs are expressed as if the network was one-big-switch (OBS) connecting edge ports directly to each other. Our compiler automatically distributes the program across network devices, freeing programmers from such details and making SNAP programs portable across topologies.

The DNS-tunnel-detect program examines two kinds of packets: incoming DNS responses (which may lead to possible DNS tunnels) and outgoing packets to resolved IP addresses. Line 1 checks whether the input packet is a DNS response to the CS department. The condition in the if statement is an example of a simple test. Such tests can involve any boolean combination of packet fields. §4 If the test succeeds, the packet could potentially belong to a DNS tunnel, and will go through the detection steps (Lines 2–6). Lines 2–6 use three global variables to keep track of DNS queries. Each

```
1  if dstip = 10.0.6.0/24 & srcport = 53 then
2    orphan(dstip)[dns.rdata] <- True;
3    susp-client(dstip)++; 
4  if susp-client(dstip) = threshold then
5    blacklist(dstip) <- True
6  else id
7  else
8    if srcip = 10.0.6.0/24 & orphan[srcip][dstip]
9      then orphan[srcip][dstip] <- False;
10     susp-client[srcip]--
11   else id
```

Figure 1: SNAP implementation of DNS-tunnel-detect.
variable is a mapping between keys and values, persistent across multiple packets. The orphan variable, for example, maps each pair of IP addresses to a boolean value. If orphan[c][s] is True then c has received a DNS response for IP address s. The variable susp-client maps the client’s IP to the number of DNS responses it has received but not accessed yet. If the packet is not a DNS response, a different test is performed, which includes a stateful test over orphan (Lines 8). If the test succeeds, the program updates orphan[arcip][dstip] to False and decrements susp-client[arcip] (Lines 10–11). This step changes the global state and thus, affects the processing of future packets. Otherwise, the packet is left unmodified — id (Line 12) is a no-op.

Routing. DNS-tunnel-detect cannot stand on its own—it does not explain where to forward packets. In SNAP, we can easily compose it with a forwarding policy. Suppose our target network is the simplified campus topology depicted in Figure 2. Here, I1 and I2 are connections to the Internet, and D1–D4 represent edge switches in the departments, with D3 connected to the CS building. C1–C6 are core routers connecting the edges. External ports (marked in red) are numbered 1–6 and IP subnet 10.0.1.0/24 is attached to port i. The assign-egress program assigns outports to packets based on their destination IP address:

```
assign-egress = if dstip = 10.0.1.0/24
then outport <- 1
else if dstip = 10.0.2.0/24 then outport <- 2
else ...
else if dstip = 10.0.6.0/24 then outport <- 6
else drop
```

Note that the policy is independent of the internal network structure, and recompilation is needed only if the topology changes. By combining DNS-tunnel-detect with assign-egress, we have implemented a useful end-to-end program: DNS-tunnel-detect; assign-egress

Monitoring. Suppose the operator wants to monitor packets entering the network at each ingress port (ports 1-6). She might use an array indexed by import and increment the corresponding element on packet arrival: count[import]++. Monitoring should take place alongside the rest of the program; thus, she might combine it using parallel composition (+): (DNS-tunnel-detect + count[import]++) ; assign-egress. Intuitively, p + q makes a copy of the incoming packet and executes both p and q on it simultaneously.

Note that it is not always legal to compose two programs in parallel. For instance, if one writes to the same global variable that the other reads, there is a race condition, which leads to ambiguous state in the final program. Our compiler detects such race conditions and rejects ambiguous programs.

Network Transactions. Suppose that an operator sets up a honeypot at port 3 with IP subnet 10.0.3.0/25. The following program records, per import, the IP and dstport of the last packet destined to the honeypot:

```
if dstip = 10.0.3.0/25
then hon-ip[import] <- srcip;
   hon-dstport[import] <- dstport
else id
```

Since this program processes many packets simultaneously, it has an implicit race condition: if packets p1 and p2, both destined to the honeypot, enter the network from port 1 and get reordered, each may visit hon-ip and hon-dstport in a different order (if the variables reside in different locations). Therefore, it is possible that hon-ip[1] contains the source IP of p1 and hon-dstport[1] the destination IP of p2 while the operator’s intention was that both variables refer to the same packet. To establish such properties for a collection of state variables, programmers can use network transactions by simply enclosing a series of statements in an atomic block. Atomic blocks co-locate their enclosed state variables so that a series of updates can be made to appear atomic.

2.2 Realizing Programs on the Data Plane

Consider the program DNS-tunnel-detect; assign-egress. To distribute this program, across network devices, the SNAP compiler should decide (i) where to place state variables (orphan, susp-client, and blacklist), and (ii) how packets should be routed across the physical network. These decisions should be made in such a way that each packet passes through devices storing every state variable it needs, in the correct order. Therefore, the compiler needs information about which packets need which state variables. In our example program, for instance, packets with dstip = 10.0.6.0/24 and srcport = 53 need to pass all three state variables, with blacklist accessed after the other two.

Program analysis. To extract the above information, we transform the program to an intermediate representation called extended forwarding decision diagram (xFDD) (see Figure 3). FDDs were originally introduced in an earlier work [32]. We extended FDDs in SNAP to support stateful packet processing. An xFDD is like a binary decision diagram: each intermediate node is a test on either packet fields or state variables. The leaf nodes are sets of action sequences, rather than merely ‘true’ and ‘false’ as in a BDD [1]. Each interior node has two successors: true (solid line), which
determines the rest of the forwarding decision process for inputs passing the test, and false (dashed line) for failed cases. xFDDs are constructed compositionally; the xFDDs for different parts of the program are combined to construct the final xFDD. Composition is particularly more involved with stateful operations: the same state variable may be referenced in two xFDDs particularly more involved with stateful operations: the same state variable may be referenced in two xFDDs.

Figure 3: The equivalent xFDD for DNS-tunnel-detect; assign-egress

In our example program, the MILP places all state variables on D4, which is the optimal location as all packets to and from the protected subnet must flow through D4. Note that this is not obvious from the DNS-tunnel-detect code alone, but rather from its combination with assign-egress. This highlights the fact that in SNAP, program components can be written in a modular way, while the compiler makes globally optimal decisions using information from all parts. The optimizer also decides forwarding paths between external ports. For instance, traffic from I1 and D1 will go through C1 and C5 to reach D4. The path from I2 and D2 to D3 goes through C2 and C6, and D3 uses C5 to reach D4. The paths between the rest of the ports are also determined by the MILP in a way that minimizes link utilization. The compiler takes state placement and routing results from the MILP, partitions the program’s intermediate representation (xFDD) among switches, and generates rules the controller should push to all stateless and stateful switches in the network.

Reacting to network events. The above phases only run if the operator changes the OBS program. Once the program compiles, and to respond to network events such as failures or traffic shifts, we use a simpler and much faster version of the MILP that given the current state placement, only re-optimizes for routing. Moreover, with state on the data plane, policy changes become considerably less frequent because the policy, and consequently switch configurations, do not change upon changes to state. In DNS-tunnel-detect, for instance, attack detection and mitigation are both captured in the program itself, happen on the data plane, and therefore react rapidly to malicious activities in the network. This is in contrast to the case where all the state is on the controller. There, the policy needs to change and recompile multiple times both during detection and on mitigation, to reflect the state changes on the controller in the rules on the data plane.

3. SNAP

SNAP is a high-level language with two key features: programs are stateful and are written in terms of an abstract network topology comprising a one-big-switch (OBS). It has an algebraic structure patterned on the NetCore/NetKAT family of languages [20, 2], with each program comprising one or more predicates and policies. SNAP’s syntax is in Figure 4. Its semantics is defined through an evaluation function “eval.” eval determines, in mathematical notation, how an input packet should be processed by a SNAP program. Note that this is part of the specification of the language, not the implementation. Any implementation of SNAP, including ours, should ensure that packets are processed as defined by

2 State can be spread out across the network. It just happens that in this case, one location turns out to be optimal.
the `eval` function: when we talk about “running” a program on a packet, we mean calling `eval` on that program and packet. We discuss `eval`’s most interesting cases here; see the extended version for a full definition \[39\].

`eval` takes the SNAP term of interest, a packet, and a starting state and yields a set of packets and an output state. To properly define the semantics of multiple updates to state when programs are composed, we need to know the reads and writes to state variables performed by each program while evaluating the packet. Thus, `eval` also returns a log containing this information. It adds “\(Rs\)” to the log whenever a read from state variable \(s\) occurs, and “\(Ws\)” on writes. Note that these logs are part of our formalism, but not our implementation. We express the program state as a dictionary that maps state variables to their contents. The contents of each state variable is itself a mapping from values to values. Values are defined as packet-related fields (IP address, TCP ports, MAC addresses, DNS domains) along with integers, booleans and vectors of such values.

**Predicates.** Predicates have a constrained semantics: they never update the state (but may read from it), and either return the empty set or the singleton set containing the input packet. That is, they either pass or drop the input packet. `id`, passes the packet and `drop`, drops it. The test \(f = v\) passes a packet `pkt` if the field \(f\) of `pkt` is \(v\). Both predicates yield empty logs.

The novel predicate in SNAP is the `state test`, written \(s[e_1] = e_2\) and read “state variable (array) \(s\) at index \(e_1\) equals \(e_2\)”. Here \(e_1\) and \(e_2\) are expressions, where an expression is either a value \(v\) (like an IP address or TCP port), a field \(f\), or a vector of them \(c\). For \(s[e_1] = e_2\), function `eval` evaluates \(e_1\) and \(e_2\) on the input packet to yield two values \(v_1\) and \(v_2\). The packet can pass if state variable \(s\) indexed at \(v_1\) is equal to \(v_2\), and is dropped otherwise. The returned log will include \(Rs\), to record that the predicate reads from the state variable \(s\).

We evaluate negation `¬` by running `eval` on \(x\) and then complementing the result, propagating whatever log \(x\) produces. `\(x \vee y\)` (disjunction) unites the results of running \(x\) and \(y\) individually, doing the reads of both \(x\) and \(y\). `\(x \& y\)` (conjunction) intersects the results of running \(x\) and \(y\) while doing the reads of \(x\) and then \(y\).

**Policies.** Policies can modify packets and the store. Every predicate is a policy—they simply make no modifications. Field modification \(f \leftarrow v\) takes an input packet `pkt` and yields a new packet, `pkt'`, such that `pkt'.f = v` but otherwise `pkt'` is the same as `pkt`. State update \(s[e_1] \leftarrow e_2\) passes the input packet through while (i) updating the store so that \(s\) at `eval(e1)` is set to `eval(e2)`, and (ii) adding \(W\) to the log. The `s[e]++` (resp. `--`) operators increment (decrement) the value of \(s[e]\) and add \(W\) to the log.

Parallel composition `p + q` runs `p` and `q` in parallel and tries to merge the results. If the logs indicate a state read/write or write/write conflict for `p` and `q` then there is no consistent semantics we can provide, and we leave the semantics undefined. Take for example \((s[0] \leftarrow 1) + (s'[0] \leftarrow 2)\). There is no conflict if \(s \neq s'\). However, the state updates conflict if \(s = s'\). There is no good choice here, so we leave the semantics undefined and raise compile error in the implementation.

Sequential composition `p; q` runs `p` and then runs `q` on each packet that `p` returned, merging the final results. We must ensure the runs of `q` are pairwise consistent, or else we will have a read/write or write/write conflict. For example, let `p` be \((f \leftarrow 1 + f \leftarrow 2\), and `pkt[f \mapsto v]` denote “update `pkt`’s `f` field to `v`”. Given a packet `pkt`, the policy `p` produces two packets: `pkt_1 = pkt[f \mapsto 1]` and `pkt_2 = pkt[f \mapsto 2]`. Let `q` be `s[0] \leftarrow f`, running `p; q` fails because running `q` on `pkt_1` and `pkt_2` updates `s[0]` differently. However, `p; q` runs fine for `q = g \leftarrow 3`.

We have an explicit conditional “if \(a\) then \(p\) else \(q\),” which indicates either `p` or `q` are executed. Hence, both `p` and `q` can perform reads and writes to the same state. We have a notation for atomic blocks, written `atomic(p)`. As described in \[32\] there is a risk of inconsistency between state variables residing on different switches on a real network, when many packets are in flight concurrently. When compiling `atomic(p)`, our compiler ensures that all the state in `p` is updated atomically \((\cite{4})\).

**4. Compilation.**

To implement a SNAP program specified on one big switch, we must fill in two critical details: traffic routing and state placement. The physical topology may offer many paths between edge ports, and many possible locations for placing state.\[4\] The routing and placement problems interact: if two flows (with different input and output OBS ports) both need some state variable \(s\), we should select routes for the two flows that pass through a common location where we place \(s\). Further complicating the situation, the OBS program may specify that certain flows read or write multiple state variables

\[\text{In this work, we assume each state variable resides in one place, though it is conceivable to distribute it (see \cite{4,4,4})}\]
in a particular order. The routing and placement on the physical topology must respect that order. In DNS-tunnel-detect, for instance, routing must ensure that packets reach wherever orphan is placed before susp-client. In some cases, two different flows may depend on the same state variables, but in different orders.

We have designed a compiler that translates OBS programs into forwarding rules and state placements for a given topology. As shown in Figure 5, the two key phases are (i) translation to extended forwarding decision diagrams (xFDDs)—used as the intermediate representation of the program and to calculate which flows need which state variables—and (ii) optimization via mixed integer linear program (MILP)—used to decide routing and state placement. In the rest of this section, we present the compilation process in phases, first discussing the analysis of state dependencies, followed by the translation to xFDDs and the packet-state mapping, then the optimization problems, and finally the generation of rules sent to the switches.

### 4.1 State Dependency Analysis

Given a program, the compiler first performs state dependency analysis to determine the ordering constraints on its state variables. A state variable \( t \) depends on a state variable \( s \) if the program writes to \( t \) after reading from \( s \). Any realization of the program on a concrete network must ensure that \( t \) does not come before \( s \). Parallel composition \( p + q \), introduces no dependencies: if \( p \) reads or writes state, then \( q \) can run independently of that. Sequential composition \( pq \), on the other hand, introduces dependencies: whatever reads are in \( p \) must happen before writes in \( q \). In explicit conditionals “if \( q \) then \( p \) else \( q \)”, the writes in \( p \) and \( q \) depend on the condition \( a \). Finally, atomic sections \( \text{atomic}(p) \) say that all state in \( p \) is inter-dependent. In DNS-tunnel-detect, for instance, \texttt{blacklist} is dependent on susp-client, itself dependent on orphan. This information is encoded as a dependency graph on state variables and is used to order the xFDD structure (§4.2), and in the MILP (§4.4) to drive state placement.

### 4.2 Extended Forwarding Decision Diagrams

We use an extended forwarding decision diagrams (xFDDs) as our internal representation for the OBS program. Formally (see Figure 6), an xFDD is either a branch \((t ? d_1: d_2)\), where \( t \) is a test and \( d_1 \) and \( d_2 \) are xFDDs, or a set of action sequences \([a_1, \ldots, a_n]\).

Each branch can be thought of as a conditional: if the test \( t \) holds on a given packet \( pkt \), then the xFDD continues processing \( pkt \) using \( d_1 \); if not, processes \( pkt \) using \( d_2 \). There are three kinds of tests. The field-value test \( f = v \) holds when \( pkt.f \) is equal to \( v \). The field-field test \( f_1 = f_2 \) holds when the values in \( pkt.f_1 \) and \( pkt.f_2 \) are equal. Finally, the state test \( s.e = e \) holds when the state variable \( s \) at index \( e \) is equal to \( e \). The last two tests are our extensions to FDDs. The state tests support our stateful primitives, and as we show later in this section, the field-field tests are required for correct compilation. Each leaf in an xFDD is a set of action sequences, with each action being either the identity; drop; field-update \( f \leftarrow v \); or state update \( s.e \leftarrow e \), which is another extension to the original FDD.

A key property of xFDDs is that the order of their tests (\( \sqsubset \)) must be defined in advance. This ordering ensures that each test is present at most once on any path in the final tree when merging xFDDs. In our xFDDs, we ensure that all field-value tests precede all field-field tests, themselves preceding all state tests. Field-value tests themselves are ordered by fixing an arbitrary order on fields and values. Field-field tests are ordered similarly. For state tests, we first define a total order on state variables by looking at the dependency graph from §4.1. We break it into strongly connected components (SCCs) and fix an arbitrary order on state variables within each SCC. Then for every edge from one SCC to another—i.e., where some state variable in the first SCC depends on some state variable in the second—\( s_1 \) precedes \( s_2 \) in the order, where \( s_2 \) is the minimal element in the second SCC and \( s_1 \) is the maximal element in the first SCC. The state tests are then ordered based on the order of state variables.

We translate a program to an xFDD using the \texttt{to-xfdd} function (Figure 7), which translates small parts of a program directly to xFDDs. Composite pro-
grams get recursively translated and then composed using a corresponding normalization operator: we use $\oplus$ for $p + q$, $\odot$ for $p \cdot q$, and $\ominus$ for $-p$. Figure 7 gives a high-level definition of the semantics of these operators. For example, $d_1 \oplus d_2$ tries to merge similar test nodes recursively by merging their true branches together and false ones together. If the two tests are not the same and $d_1$’s test comes first in the total order, both of its subtrees are merged recursively with $d_2$. The other case is similar. $d_1 \odot d_2$ for leaf nodes is the union of their action sets.

The hardest case is surely for $\ominus$, where we try to add in an action sequence as to an xFDD ($t ? d_1 : d_2$). Suppose we want to compose $f \leftarrow v_1$ with ($f = v_2 ? d_1 : d_2$). The result of this xFDD composition should behave as if we first do the update and then the condition on $f$. If $v_1 = v_2$, the composition should continue only on $d_1$, and if not, only on $d_2$. Now let’s look at a similar example including state, composing $s[srcip] \leftarrow e_1$ with ($s[dstip] = e_2 ? d_1 : d_2$). If $srcip$ and $dstip$ are equal (rare but not impossible) and $e_1$ and $e_2$ always evaluate to the same value, then the whole composition reduces to just $d_1$. The field-field tests are introduced to let us answer these equality questions, and that is why they always precede state tests in the tree. The trickiness in the algorithm comes from generating proper field-field tests, by keeping track of the information in the xFDD, to clear out these cases. The full algorithm is given the extended version 33.

Inconsistent use of state variables is prohibited by the language semantics when composing programs (see 43). We enforce the semantics by looking for these violations while merging the xFDSDs of composed programs and raising a compile error if the final xFDD contains a leaf with parallel updates to the same state variable.

4.3 Packet-State Mapping

For a given program $p$, the corresponding xFDD $d$ offers an explicit and complete specification of the way $p$ handles packets. We analyze $d$, using an algorithm called packet-state mapping, to determine which $flows$ use which states. This information is further used in the optimization problem (14) to decide the correct routing for each flow. Our default definition of a flow is those packets that travel between any given pair of ingress/egress ports in the OBS, though we can use other notions of flow (see 14). Traversing from $d$’s root down to the action sets at $d$’s leaves, we can gather information associating each flow with the set of state variables read or written. We omit the full algorithm due to space constraints.

Furthermore, the operators can give hints to the compiler by specifying their network assumptions in a separate policy:

- `assumption = (srcip = 10.0.1.0/24 & inport = 1)`
- `+ (srcip = 10.0.6.0/24 & inport = 6)`
- `

We require the assumption policy to be a predicate over packet header fields, only passing the packets that match the operator’s assumptions. assumption is then sequentially composed with the rest of the program, enforcing the assumption by dropping packets that do not match the assumption. Such assumptions benefit the packet-state mapping. Consider our example xFDD in Figure 3. Following the xFDD’s tree structure, we can infer that all the packets going to port 6 need all the three state variables in DNS-tunnel-detect. We can also infer that all the packets coming from the 10.0.6.0/24 subnet need orphan and susp-client. However, there is nothing in the program to tell the compiler that these packets can only enter the network from port 6. Thus, the above assumption policy can help the compiler to identify this relation and place state more efficiently.

4.4 State Placement and Routing

To decide state placement and routing, we generate an optimization problem that takes the form of a mixed-integer linear program (MILP), and is an extension of the multi-commodity flow problem. The MILP has three key inputs: the concrete network topology, the state dependency graph $G$, and the packet-state mapping, and two key outputs: routing and state placement (Table 4). The objective is to minimize the sum of link utilization (or a convex function of it) in the network.

**Inputs**. The topology is defined in terms of the following inputs to the MILP: (i) the nodes, some distinguished as edges (ports in OBS), (ii) expected traffic $d_{uv}$ for every pair of edge nodes $u$ and $v$, and (iii) link capacities $c_{ij}$ for every pair of nodes $i$ and $j$. State dependencies in $G$ are translated into input sets $dep$ and $tied$. $tied$ contains pairs of state variables which are
in the same SCC in G, and must be co-located. \textit{dep} identifies state variables with dependencies that do not need to be co-located; in particular, \((s, t) \in \text{dep}\) when \(s\) precedes \(t\) in variable ordering, and they are not in the same SCC in \(G\). The packet-state mapping is used as the input variables \(S_{uv}\), identifying the set of state variables needed on flows between nodes \(u\) and \(v\).

**Outputs and Constraints.** The routing outputs are variables \(R_{uvij}\), indicating which fraction of the flow from edge node \(u\) to \(v\) should traverse the link between nodes \(i\) and \(j\). The constraints on \(R_{uvij}\) (left side of Table 2) follow the multi-commodity flow problem closely, with standard link capacity and flow conservation constraints, and edge nodes distinguished as sources and sinks of traffic.

State placement is determined by the variables \(P_{sn}\), which indicate whether the state variable \(s\) should be placed on the physical switch \(n\). Our constraints here are more unique to our setting. First, every state variable \(s\) can be placed on exactly one switch. A choice we discuss at the end of this section. Second, we must ensure that flows that need a given state variable \(s\) traverse that switch. Third, we must ensure that each flow traverses states in the order specified by the \(\text{dep}\) relation; this is what the variables \(P_{svij}\) are for. We require that \(P_{svij} = R_{uvij}\) when the traffic from \(u\) to \(v\) that goes over the link \((i, j)\) has already passed the switch with the state variable \(s\), and zero otherwise. If \(\text{dep}\) requires that \(s\) should come before some other state variable \(t\)—and if the \((u, v)\) flow needs both \(s\) and \(t\)—we can use \(P_{svij}\) to make sure that the \((u, v)\) flow traverses the switch with \(t\) only after it has traversed the switch with \(s\) (the last state constraint in Table 2). Finally, we must make sure that state variables \((s, t) \in \text{tied}\) are located on the same switch.

The MILP can be configured to decide paths for more fine-grained notions of flows. Suppose packet-state mapping finds that only packets with \(\text{srcip} = x\) need state variable \(s\). We refine the MILP input to have two edge nodes per port, one for traffic with \(\text{srcip} = x\) and one for the rest, so the MILP can choose different paths for them. Moreover, the MILP assigns each state variable \(s\) to one physical switch, to avoid the overhead of synchronizing multiple instances of the same variable. Besides, most network topologies are hierarchical and have core switches that can serve as good candidates for placing state variables common across multiple flows. Still, distributing a state variable remains a valid option. For instance, the compiler can partition \(s[\text{import}]\) into \(k\) disjoint state variables, each storing \(s\) for one port. The MILP can decide placement and routing as before, this time with the option of distributing partitions of \(s\) with no concern for synchronization.

### 4.5 Generating Data-Plane Rules

Rule generation happens in two phases, combining information from the xFDD and MILP to configure the network switches. In generating the configurations, we assume each packet is augmented with a SNAP-header upon entering the network, which contains its original OBS import and future outport, and the id of the last processed xFDD node, the purpose of which will be explained shortly. This header is stripped off when the packet exits the network. We use DNS-tunnel-detect from [2] as a running example, with its xFDD in Figure 3. For exemplary clarity, we assume that all state variables are stored on \(C_6\) instead of \(D_i\).

In the first phase, we break the xFDD down into ‘per switch’ xFDDs, since not every switch needs the entire xFDD to process packets. Splitting the xFDD is straightforward, given placement information: stateless tests and actions can happen anywhere, but reads and writes of states have to happen on switches storing them. For example, edge switches \((I_1\) and \(I_2\), and \(D_1\) to \(D_6\) only need to process packets up to the state tests, e.g., tests 3 and 8, and write the test number in the packet’s SNAP-header showing how far into the xFDD they made progress. Then, they send the packets to \(C_6\), which has the corresponding state variables, orphan and susp-client. \(C_6\), on the other hand, does not need the top part of the xFDD. It just needs the subtrees containing its state variables to continue processing the packets sent to it from the edges. The per-switch xFDDs are then translated to switch-level configurations, by a straightforward traversal of the xFDD (See [4]).

In the second phase, we generate a set of match-action rules that take packets through the paths decided by the MILP. These paths comply with the state ordering used in the xFDD, thus they get packets to switches with the right states in the right order. Note that packets contain the path identifier (the OBS import and outport, \((u, v)\) pair in this case) and the “routing” match-action rules are generated in terms of this identifier to forward
them on the correct path. Additionally, note that it may not always be possible to decide the egress port \( v \) for a packet upon entry if its outpost depends on state. We observe that in that case, all the paths for possible outports of the packet pass the state variables it needs. We pick one of these paths in proportion to their capacity and show, in the extended version \([39]\), that traffic on these paths remains in their capacity limit.

As an example of packets being handled by generated rules, consider a DNS response with source IP 10.0.1.1 and destination IP 10.0.6.6, entering the network from port 1. The rules on \( I_1 \) process the packet up to test 8 in the xFDD, tag the packet with the path identifier (1, 6) and number 8. The packet is then sent to \( C_6 \). There, \( C_6 \) will process the packet from test 8, update state variables accordingly, and send the packet to \( D_4 \) to exit the network from port 6.

5. IMPLEMENTATION

The compiler is mostly implemented in Python, except for the state placement and routing phase \([4.4]\) in which uses the Gurobi Optimizer \([15]\) to solve the MILP. The compiler’s output for each switch is a set of switch-level instructions in a low-level language called NetASM \([31]\), which comes with a software switch capable of executing those instructions. NetASM is an assembly language for programmable data planes designed to serve as the “narrow waist” between high-level languages such as SNAP, and NetCore \([20]\), and programmable switching architectures such as RMT \([7]\). FPGAs, network processors and Open vSwitch.

As described in \([4.5]\), each switch processes the packet by its customized per-switch xFDD, and then forwards it based on the fields of SNAP-header using a match-action table. To translate the switch’s xFDD to NetASM instructions, we traverse the xFDD and generate a branch instruction for each test node, which jumps to the instruction of either the true or false branch based on the test’s result. Moreover, we generate instructions to create two tables for each state variable, one for the indices and one for the values. In the case of a state test in the xFDD, we first retrieve the value corresponding to the index that matches the packet, and then perform the branch. For xFDD leaf nodes, we generate store instructions that modify the packet fields and state tables accordingly. Finally, we use NetASM support for atomic execution of multiple instructions to guarantee that operations on state tables happen atomically.

While NetASM was useful for testing our compiler, any programmable device that supports match-action tables, branch instructions, and stateful operations can be a SNAP target. The prioritized rules in match-action tables, for instance, are effectively branch instructions. Thus, one can use multiple match-action tables to implement xFDD in the data plane, generating a separate rule for each path in the xFDD. Several emerging switch interfaces support stateful operations \([6, 4, 35, 24]\). We discuss possible software and hardware implementations for SNAP stateful operations in \([7]\).

6. EVALUATION

This section evaluates SNAP in terms of language expressiveness and compiler performance.

6.1 Language Expressiveness

We created several stateful network functions (Table 3) that are typically delegated to middleboxes. Examples were taken from the Chimera \([9]\), FAST \([21]\), and Bohatei \([8]\) systems. The code can be found in the extended version \([39]\). Most examples use protocol-related fields in fixed packet-offset locations, which are parsable by emerging programmable parsers. Some fields require session reassembly. However, this is orthogonal to the language expressiveness. As long as these fields are available to the switch, they can be used in SNAP programs. To make them available, one could extract these fields by placing a “preprocessor” before the switch pipeline, similar to middleboxes. For instance, Snort \([33]\) uses preprocessors prior to the detection engine, which can use the fields extracted by them.

6.2 Compiler Performance

The compiler has several phases upon the system’s cold start, yet most events require only part of them. Table 4 summarizes these phases and their sensitivity to network and policy changes.

Cold Start. When the very first program is compiled, the compiler goes through all phases, including MILP model creation, which happens only once in the lifetime of the network. Once created, the model supports incremental additions and modifications of variables and constraints in a few milliseconds.

Policy Changes. Compiling a new program goes through the three program analysis phases and rule generation as well as both state placement and rout-

<table>
<thead>
<tr>
<th>Application</th>
<th># domains sharing the same IP address</th>
<th># distinct IP addresses under the same domain</th>
<th>DNS TTL change tracking</th>
<th>DNS tunnel detection</th>
<th>Sidejacking detection</th>
<th>Phishing/spam detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chimera</td>
<td>[5]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAST</td>
<td>[21]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bohatei</td>
<td>[8]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>[5]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Applications written in SNAP.
Table 4: Compiler phases. For each scenario, phases that get executed are checkmarked.

<table>
<thead>
<tr>
<th>ID</th>
<th>Phase</th>
<th>Topic/TM Change</th>
<th>Policy Change</th>
<th>Cold Start</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>State dependency</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>P2</td>
<td>xFDD generation</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>P3</td>
<td>Packet-state map</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>P4</td>
<td>MILP creation</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>P5</td>
<td>MILP solving</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>P6</td>
<td>Rule generation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 5: Statistics of evaluated enterprise/ISP topologies.

<table>
<thead>
<tr>
<th>Topology</th>
<th># Switches</th>
<th># Edges</th>
<th># Demands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>26</td>
<td>92</td>
<td>207,898</td>
</tr>
<tr>
<td>Berkeley</td>
<td>25</td>
<td>96</td>
<td>342,256</td>
</tr>
<tr>
<td>Purdue</td>
<td>98</td>
<td>232</td>
<td>243,506</td>
</tr>
<tr>
<td>AS 1755</td>
<td>87</td>
<td>322</td>
<td>360,000</td>
</tr>
<tr>
<td>AS 1221</td>
<td>104</td>
<td>302</td>
<td>518,400</td>
</tr>
<tr>
<td>AS 6461</td>
<td>138</td>
<td>744</td>
<td>921,600</td>
</tr>
<tr>
<td>AS 3257</td>
<td>161</td>
<td>656</td>
<td>1,255,440</td>
</tr>
</tbody>
</table>

Table 6: Runtime of compiler phases when compiling DNS-tunnel-detect with routing on enterprise/ISP topologies.

<table>
<thead>
<tr>
<th>Topology/TM Changes</th>
<th>Stanford</th>
<th>Berkeley</th>
<th>Purdue</th>
<th>AS 1755</th>
<th>AS 1221</th>
<th>AS 6461</th>
<th>AS 3257</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>1.1</td>
<td>1.5</td>
<td>1.2</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>P1-P2-P3 (s)</td>
<td>28</td>
<td>47</td>
<td>67</td>
<td>19</td>
<td>21</td>
<td>116</td>
<td>142</td>
</tr>
<tr>
<td>P5 (s)</td>
<td>10</td>
<td>18</td>
<td>27</td>
<td>6</td>
<td>7</td>
<td>47</td>
<td>74</td>
</tr>
<tr>
<td>P6 (s)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.04</td>
<td>0.04</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>P4 (s)</td>
<td>75</td>
<td>150</td>
<td>169</td>
<td>22</td>
<td>32</td>
<td>120</td>
<td>163</td>
</tr>
</tbody>
</table>

Scaling with topology size. We synthesize networks with 10–180 switches using IGen [27]. In each network, 70% of the switches with the lowest degrees are chosen as edges and the DNS tunnel policy is compiled with that network as a target. Figure 9 shows the compilation time for different scenarios, combining the runtimes of phases relevant for each. Note that by increasing the topology size, the policy size also increases in the assign-egress and assumption parts.

Scaling with number of policies. We use the network from the previous experiment with 50 switches and gradually increase the number of policies by parallel composing those from Table 3 each destined to a separate egress port. The policies are mostly similar in complexity to the DNS tunnel example (12). Figure 10 depicts the compilation time as a function of the number of policies. The 10-second jump from 18 to 19 takes place when the TCP state machine policy is added, which is considerably more complex than others.

6.2.2 Analysis of Experimental Results

Creating the MILP takes longer than solving it, in most cases, and much longer than other phases. Fortunately, this is a one-time cost. After creating the MILP instance, incrementally adding or removing variables and constraints (as the topology or state requirement changes) takes just a few milliseconds.

Solving the ST MILP unsurprisingly takes longer as compared to the rest of the phases when topology grows. It takes around 2.5 minutes for the biggest synthesized topology and around 2.3 minutes for the biggest RocketFuel topology. The curve is close to exponential as the problem is inherently computationally hard. However, this phase takes place only in cold start or upon a policy change, which are infrequent and planned in advance.

Re-optimizing routing with fixed state placement is much faster. In response to network events (e.g., link failures), the TE MILP can recompute paths in around a minute across all our experiments, which is the timescale we expected for this phase. Moreover, it can be used even on policy changes, if the user settles for a sub-optimal state placement using heuristics rather than ST MILP. We plan to explore such heuristics.

Given the kinds of events that require complete (policy change) or partial (network events) recompilation, we believe that our compilation techniques meet the requirements of enterprise networks and medium-size
7. DISCUSSION

This section discusses data-plane implementation strategies for SNAP’s stateful operations as well as how SNAP relates to middleboxes.

7.1 Stateful Operations in the Data Plane

A state variable (array) in SNAP is a key-value mapping, or a dictionary, on header fields, persistent across multiple packets. When the key (index) range is small, it is feasible to pre-allocate all the memory the dictionary needs and implement it using an array. A large but sparse dictionary can be implemented using a reactively-populated table, similar to a MAC learner table. It contains a single default entry in the beginning, and as packets fly by and change the state variable, it reactively adds/updates the corresponding entries.

In software, there are efficient techniques to implement a dictionary in either approach, and some software switches already support similar reactive “learning” operations, either atomically [31] or with small periods of inconsistency [24]. The options for current hardware are: (i) arrays of registers, which are already supported in emerging switch interfaces [6]. They can be used to implement small dictionaries, as well as hash tables and Bloom Filters as sparse dictionaries. In the latter case, it is possible for two different keys to hash to the same dictionary entry. However, there are applications such as load balancing and flow-size-based sampling that can tolerate such collisions [21]. (ii) Content Addressable Memories (CAMs) are typically present in today’s hardware switches and can be modified by a software agent running on the switch. Since CAM updates triggered by a packet are not immediately available to the following packets, it may be used for applications that tolerate small periods of state inconsistency, such as a MAC learner, DNS tunnel detection, and others from Table 3. Our NetASM implementation (§5) takes the CAM-based approach. NetASM’s software switch supports atomic update to the tables in the data plane and therefore can perform consistent stateful operations.

Overall, any software/hardware switch with similar capabilities as described above, can implement all or a useful subset of SNAP’s stateful operations and therefore can fit as SNAP’s data-plane target for switches that carry state.

7.2 SNAP and Middleboxes

Networks traditionally rely on middleboxes for advanced packet processing, including stateful functionalities. However, advances in switch technology enables stateful packet processing in the data plane, which naturally makes the switches capable of subsuming a subset of middlebox functionality. SNAP provides a high-level programming framework to exploit this ability, hence, it is able to express a wide range of stateful programs that are typically relegated to middleboxes (see Table 3 for examples). This helps the programmer to think about a single, explicit network policy, as opposed to a disaggregated, implicit network policy using middleboxes, and therefore, get more control and customization over a variety of simpler stateful functionalities.

This also makes SNAP subject to similar challenges as managing stateful middleboxes. For example, many network functions must observe all traffic pertaining to a connection in both directions. In SNAP, if traffic in both directions uses a shared stateful variable, the MILP optimizer forces traffic in both directions through the same node. Moreover, previous work such as Split/Merge [28] and OpenNF [13] show how to migrate internal state from one network function to another, and Gember-Jacobson et al. [12] manage to migrate state without buffering packets at the controller. SNAP currently focuses on static state placement. However, since SNAP’s state variables are explicitly declared as part of the policy, rather than hid-
den inside blackbox software, SNAP is well situated to adopt these algorithms to support smooth transitions of state variables in dynamic state placement. Additionally, the SNAP compiler can easily analyze a program to determine whether a switch modifies packet fields to ensure correct traffic steering—something that is challenging today with blackbox middleboxes [17, 9].

While SNAP goes a step beyond previous high-level languages to incorporate stateful programming into SDN, we neither claim it is as expressive as all stateful middleboxes, nor that it can replace them. To interact with middleboxes, SNAP may adopt techniques such as FlowTags [17] or SIMPLE [9] to direct traffic through middleboxes chains by tagging packets to mark their progress. We focus on programming networks but if verification is of interest in future work, one might adopt techniques such as RONO [23] to verify isolation properties in presence of stateful middleboxes. In summary, interacting with existing middleboxes is no harder or easier in SNAP than it is in other global SDN languages such as NetKAT [2].

8. RELATED WORK

Stateful languages. Stateful NetKAT [19], developed concurrently with SNAP, is a stateful language for “event-driven” network programming, which guarantees consistent update when transitioning between configurations in response to events. SNAP source language is richer and exponentially more compact than stateful NetKAT as it contains multiple arrays (as opposed to one) that can be indexed and updated by contents of packet headers (as opposed to constant integers only). Moreover, they place multiple copies of state at the edge, proactively generate rules for all configurations, and optimize for rule space, while we distribute state and optimize for congestion. Kinetic [18] provides a per-flow state machine abstraction, and NetEgg [42] synthesizes stateful program from user’s examples. However, they both keep the state at the controller.

Compositional languages. NetCore [20], and other similar languages [29, 10, 2], have primitives for tests and modifications on packet fields as well as composition operators to combine programs. SNAP builds on these languages by adding primitives for stateful programming [8]. To capture the joint intent of two policies, sometimes the programmer needs to decompose them into their constituent pieces, and then reassemble them using ; and +. PGA [25] allows programmers to specify access control and service chain policies using graphs as the basic building block, and tackles this challenge by defining a new type of composition. However, PGA does not have linguistic primitives for stateful programming, such as those that read and write the contents of global arrays. Thus, we view SNAP and PGA as complementary research projects, with each treating different aspects of language design space.

Stateful switch-level mechanisms. FAST [21] and OpenState [4] propose flow-level state machines as a primitive for a single switch. SNAP offers a network-wide OBS programming model, with a compiler to distribute the programs across the network. Thus, although SNAP is exponentially more compact than a state machine in cases where the state stores contents of packet header, both FAST and OpenState can be used as a target for a subset of SNAP’s programs.

Optimizing placement and routing. Several projects have explored optimizing placement of middleboxes and/or routing traffic through them. These projects and SNAP share the mathematical problem of placement and routing on a graph. Merlin programs specify service chains as well as optimization objectives [36], and the compiler uses an MILP to choose paths for traffic with respect to specification. However, it does not decide the placement of service boxes itself. Rather, it chooses the paths to pass the existing instances of the services in the physical network. Stratos [11] explores middlebox placement and distributing flows amongst them to minimize inter-rack traffic, and Slick [3] breaks middleboxes into fine-grained elements and distributes them across the network while minimizing congestion. However, they both have a separate algorithm for placement. In Stratos, placement results is used in an ILP to decide distribution of flows. Slick uses a virtual topology on the placed elements with heuristic link weights, and finds shortest paths between traffic endpoints.

9. CONCLUSION AND FUTURE WORK

In this paper, we introduced a stateful SDN programming model with a one-big-switch abstraction, persistent global arrays, and network transactions. We developed algorithms for analyzing and compiling programs, and distributing their state across the network. Based on these ideas, we prototyped and evaluated the SNAP language and compiler on numerous sample programs.

As future work, we plan to extend the SNAP language and compiler to support a wider range of stateful programs. First, we will allow a field to be assigned the value of a state variable at a specific index (i.e., $f \leftarrow a[e]$). This feature would allow SNAP to express applications like NATs and proxies that store connection mappings in state variables and modify packets as they go by. Second, we will add an extra field called content, containing the packet’s payload, where tests on the content field can match on regular expressions to represent DPI capabilities. Third, we will allow fields spread over multiple packets, requiring connection reassembly (a “one big packet” abstraction). Each of these extensions introduce new and interesting research problems to precisely define the language semantics, and extend our compilation algorithms and prototype.
10. REFERENCES


