Forwarding and Routing With Packet Subscriptions

Theo Jepsen, Ali Fattaholmanan, Masoud Moshref, Nate Foster, Antonio Carzaniga, and Robert Soulé

***Abstract*— In this paper, we explore how programmable data planes can naturally provide a higher-level of service to user applications via a new abstraction called packet subscriptions. Packet subscriptions generalize forwarding rules, and can be used to express both traditional routing and more esoteric, content-based approaches. We present strategies for routing with packet subscriptions in which a centralized controller has a global view of the network, and the network topology has a hierarchical or general structure. We also describe a compiler for packet subscriptions that uses a novel BDD-based algorithm to efficiently translate predicates into P4 tables that can support O(100K) expressions. Using our system, we have built eight diverse applications. We show that these applications can be deployed in brownfield networks while performing line-rate message processing, using the full switch bandwidth of 6.5Tbps.**

***Index Terms*— Publish/subscribe, information-centric networking, network programmability (SDN/NFV/In-network computing).**

I. Introduction

THE advent of programmable data planes [1]–[3] is having a profound impact on networking, with clear benefits to network operators (e.g., increased visibility via fine-grained network telemetry) and to switch vendors (e.g., software development is faster and less expensive than hardware development). However, the benefits to users are still relatively unexplored, in the sense that today's programmable data planes offer the same forwarding abstractions that fixed-function devices have always provided—e.g., match on IP address, decrement TTL, and send to the next hop.

While the Internet is based on a well-motivated design [4], classic protocols such as TCP/IP provide a lower level of abstraction than modern distributed applications expect, especially in networks managed by a single entity, such as data centers. As a case in point, today it is common to deploy services in lightweight containers. Address-based routing for containerized services is difficult, because containers deployed on the same host may share an address, and because containers may move, causing its address to change. To cope with these networking challenges, operators are deploying identifier-based routing, such as Identifier Locator Addressing (ILA) [5]. These schemes require that name resolution be performed as an intermediate step. Another example is load balancing: to improve application performance and reduce server load, data centers rely on complex software systems to map incoming IP packets to one of a set of possible service end-points. Today, this service layer is largely provided by dedicated middleboxes. Examples include Google's Maglev [6] and Facebook's Katran [7]. A third example occurs in big data processing systems, which typically rely on message-oriented middleware, such as TIBCO Rendezvous [8], Apache Kafka [9], or IBM's MQ [10]. This middleware allows for a greater decoupling of distributed components, which in turn helps with fault tolerance and elastic scaling of services [11].

Although the current approach provides the necessary functionality—the middleboxes and middleware abstracts away the address-based communication fabric from the application—the impedance mismatch between the abstraction that networks offer and the abstraction that applications need adds complexity to the network infrastructure. Using middleboxes to implement this higher-level of network service limits performance, in terms of throughput and latency, as servers process traffic at gigabits per second, while ASICs can process traffic at terabits per second. Moreover, middleboxes increase operational costs and are a frequent source of network failures [12], [13]. *Given the existence of programmable devices, can't we do better?*

In this paper, we propose a new network abstraction called *packet subscriptions*. A packet subscription is a stateful predicate that, when evaluated on an input packet, determines a forwarding decision. Packet subscriptions generalize traditional forwarding rules; they are more expressive than basic equality or prefix matching and they can be written on arbitrary, user-defined packet formats. A packet subscription compiler generates both the data plane configuration and the control plane rules, providing a uniform interface for programming the network. Packet subscriptions easily express a range of higher-level network services, including pub/sub [11], in-network caching [14], [15], and identifier-based routing [5].

In some respects, packet subscriptions share a similar motivation to prior work on content-centric networking [16]–[19]. However, in contrast to this prior work, we are *not* proposing a complete re-design of the Internet [19], [20]. Instead, we argue that higher-level network abstractions are already used extensively by distributed applications, and this functionality can be naturally provided by the network data plane. Moreover, packet subscriptions can be implemented efficiently

Manuscript received December 23, 2020; revised January 21, 2022; accepted April 26, 2022; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor S. Subramaniam. This work was supported by the Swiss National Science Foundation (SNF) under Project 200021_166132. *(Corresponding author: Theo Jepsen.)*

Theo Jepsen is with the Department of Computer Science, Stanford University, Stanford, CA 94305 USA (e-mail: jepsen@stanford.edu).

Ali Fattaholmanan and Antonio Carzaniga are with the Department of Informatics, Università della Svizzera italiana, 6904 Lugano, Switzerland.

Masoud Moshref is with Barefoot Networks/Intel, Santa Clara, CA 24863 USA.

Nate Foster is with Barefoot Networks/Intel, Santa Clara, CA 24863 USA, and also with the Department of Computer Science, Cornell University, Ithaca, NY 14853 USA.

Robert Soulé is with Barefoot Networks/Intel, Santa Clara, CA 24863 USA, and also with the Department of Computer Science, Yale University, New Haven, CT 06520 USA.

Digital Object Identifier 10.1109/TNET.2022.3172066

in controlled, data center deployments, in which the entire network is in a single administrative domain, and operators have the freedom to directly tailor the network to the needs of the applications. Packet subscriptions interoperate with other routing schemes (e.g. IP), so they are also suitable for brownfield deployments with heterogeneous network devices.

Furthermore, although security is an important concern for any network design, the controlled data center setting mitigates some of the most pressing security issues. Moreover, it is reasonable to assume that for applications that utilize a publish/subscribe style of communication, the data is public.

Supporting packet subscriptions as a network-level service requires addressing a series of challenges. At the network-wide level, the challenge is routing. Analogous to IP, routing on packet subscriptions amounts to placing and possibly combining rules throughout the network so as to induce the right flows of packets from publishers to subscribers. At the switch-local level, the main problem is forwarding, meaning efficiently matching packets against a set of local rules. Also at the switch-local level, there are technical challenges in efficiently parsing structured packets and allocating switch memory.

To address these challenges, we have designed a new network architecture, named Camus. Applications provide Camus with filters written in a packet subscription language. Camus provides a controller component that determines a global routing policy based on the subscriptions, and a compiler component that generates the control and data plane configurations for the local forwarding decisions that collectively realize the routing strategy.

At the routing level, Camus offers two different routing strategies for hierarchical topologies. One strategy reduces the number of forwarding rules stored in switches at the expense of routing all traffic through the network core. The second strategy makes the opposite trade-off; it avoids sending traffic through the core, at the expense of greater storage requirements. Both strategies assume a data-center network deployment, in which a centralized controller has a global view of the network, and the network topology is organized as a hierarchical structure, such as a Fat Tree or Clos architecture. For general topologies, Camus constructs and routes on a spanning tree of the topology.

With respect to forwarding, naïvely translating packet subscriptions into FIB entries would require significant amounts of TCAM and SRAM memory, which is a scarce resource on network hardware. Instead, Camus uses an algorithm based on Binary Decision Diagrams (BDDs) [21], [22]. The Camus compiler translates logical predicates into P4 tables that can support O(100k) filter expressions within the limited resources of a programmable switch ASIC. Moreover, Camus provides functionality for parsing application-specific message formats, which requires reading deeply into the packet, and processing messages that have been batched together into a single network packet.

We have used Camus to provide communication for eight diverse applications. We performed an in-depth performance evaluation for three of them: a financial application for filtering market feeds (i.e., the ITCH protocol provided by NASDAQ); video streaming services using Cisco's hybrid ICN (hICN) [23]; and in-band network telemetry (INT) event detection. This diversity of applications demonstrates the flexibility and expressiveness of Camus. Moreover, our prototype demonstrates substantial improvements in throughput over software based alternatives, while processing messages at line-rate.

This paper extends our workshop [24] and conference [25] papers by demonstrating the utility and expressiveness of packet subscriptions with additional example applications, and expanding the discussion on architecture practicality. Overall, this paper makes the following contributions:

- It introduces a high-level design of a packet subscription language targeting programmable ASICs (§II).
- It demonstrates a strategy to route via packet subscriptions in hierarchical and general network topologies.(§IV).
- It presents an algorithm to efficiently compile packet subscription to P4 tables and control plane rules (§V).
- It describes techniques for parsing batches of application-level messages deep inside a packet (§VI).
- It experimentally evaluates an implementation of in-network pub/sub using packet subscriptions against software based alternatives (§VIII).

II. Packet Subscriptions

A packet subscription is a filter that determines whether a packet is of interest, and therefore whether it should be forwarded to an application. So, when end-points submit a packet subscription to the global controller, they are effectively saying "send me the packets that match this filter". The following is an example of a stateless filter:

```
ip.dst == 192.168.0.1
```

It indicates that packets with the IP destination address `192.168.0.1` should be forwarded to the end-point that submitted this filter.

One can interpret this filter the traditional way: each host is assigned an IP address, and the switches forward packets toward their destinations. However, in this traditional interpretation, the network is responsible for assigning IP addresses to end-points. Instead, with packet subscriptions it is the application that assigns IP addresses. In other words, packet subscriptions empower applications with the ability to define the routing structure for the network.

Another interpretation is that the subscription is equivalent to joining a multicast group with a given IP address. However, with packet subscriptions, the IP address has no particular global meaning, and instead it is just another attribute of the packet. Applications can use other attributes for routing, and in particular they can express their interests by combining multiple conditions on one or more attributes.

For example, suppose that a trading application is interested in ITCH messages about Google stock. The following filter

$h \in$ *Packet headers*	$n \in$ *Numbers*	
$f \in$ *Header fields*	$s \in$ *Strings*	
$v \in$ *State variables* (e.g., `my_counter`, see Figure 4)		
$g \in$ *State aggregation functions* (e.g., `avg`)		
$c ::= c_1 \wedge c_2 \mid c_1 \vee c_2 \mid !\, c \mid \mathrm{e}$		Filter: logical expression
$e ::= a > n \mid a < n \mid a == n \mid \ldots$		Numeric constraint
$\mid a$ *prefix* $s \mid a == s \mid \ldots$		String constraint
$a ::= h.f \mid v \mid g(v_0 \ldots v_n)$		Attributes

Fig. 1. Packet subscription language abstract syntax.

matches ITCH messages where the `stock` field is the constant `GOOGL` and the `price` field is greater than 50:

```
stock == GOOGL  ∧ price > 50
```

Packet subscriptions may also be stateful—i.e., their behavior may depend on previously processed data packets. To specify a stateful packet subscription, we can use operations on variables in the switch data plane, such as computing a moving average over the value contained in a header field:

```
stock == GOOGL ∧ avg(price) > 60
```

In addition to checking the `stock` field, this filter requires that the moving average of the `price` field exceeds the threshold value 60. The macro `avg` stores the current average, which is updated whenever the rest of the filter matches.

In general, a packet subscription is a logical expression of constraints on individual attributes of a packet or on state variables (see Figure 1). Each constraint compares the value of an attribute or a state variable (or an aggregate thereof) with a constant, using a specified relation. The Camus subscription language supports basic relations over numbers (e.g., equality and ordering) and over strings (e.g., equality and prefix).

The packet subscription language is designed to be expressive while also allowing for an efficient realization in the network [26]. In particular, the simple structure and semantics is easy to understand, since it corresponds to a very basic query language such as a subset of the WHERE clause of an SQL expression. The structure and semantics is also versatile and expressive, since it can represent non-trivial conditions over application-defined data within packets. And, crucially, subscriptions can be aggregated, using exact or approximate reductions, and then compiled into appropriate table structures for fast evaluation in network switches. As we will see later in Section V, the aggregation and reduction algorithms exploit the simple structure of subscriptions, as well as the semantics of the numeric and string relations.

The language also supports stateful predicates, to a limited extent. First, it can only evaluate predicates that reason about local state. It cannot filter on global state (e.g., the sum of values at more than one device). Second, re-evaluating stateful predicates on multiple devices can lead to unexpected results (e.g., the average of the average of the average). Therefore, it only evaluates stateful functions at the last hop switch before a subscriber. And, third, due to the underlying hardware constraints, the types of computations it can perform is limited. For these reasons, the stateful functions that it supports are restricted to basic aggregations over tumbling windows, including count, sum, and average. This is similar to systems such as Linear Road [27] and Sonata [28].

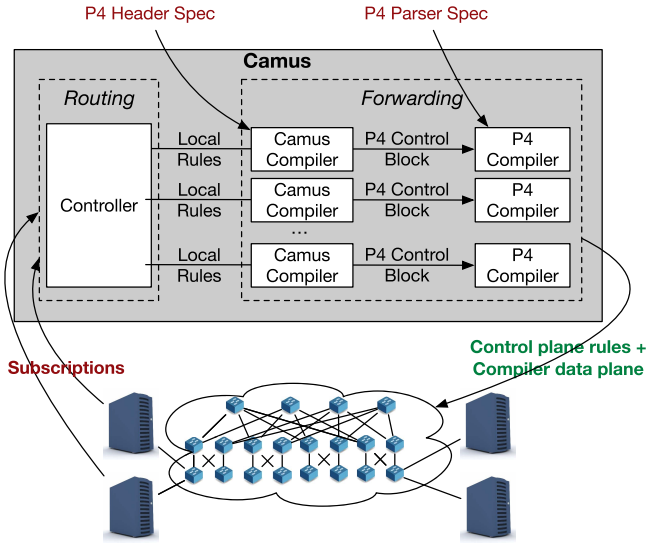


Fig. 2. Overview of camus.

III. Network Architecture

Adopting packet subscriptions as a new network abstraction requires that we re-think the network architecture. Figure 2 illustrates the architecture design. Subscribers express interest in messages, and publishers send messages. The switches running the Camus pipeline process the messages and forward them to interested subscribers.

Camus assumes a logically centralized controller with an omniscient view of the network (i.e., the current topology and device state). One could imagine a decentralized version of Camus, whereby each switch control plane runs the Camus compiler, and subscription information is disseminated through the network (à la conventional routing protocols). However, we leave such a design for future work.

Applications provide the controller with a set of filters written in the Camus subscription language. The application domain is characterized by a set of headers and corresponding packet formats. Camus requires that headers and packet formats be specified through user-provided P4 code.

The Camus controller combines the end-point subscriptions and computes a global routing policy. Although Camus supports general topologies, we focus on static, hierarchical network topologies, such as the Fat Tree architecture. This architecture is common in data-center networks [29], which is our expected deployment for packet subscriptions. It also simplifies the job of the controller, as the topology naturally forms tree-structures by simply distinguishing links that go up or down the hierarchy. We discuss how routing is handled by Camus in Section IV.

To implement the routing policy, the controller emits a set of local rules that are compiled to run on the individual switches in the network. These rules determine the runtime control plane configuration of the switch, whereas the static

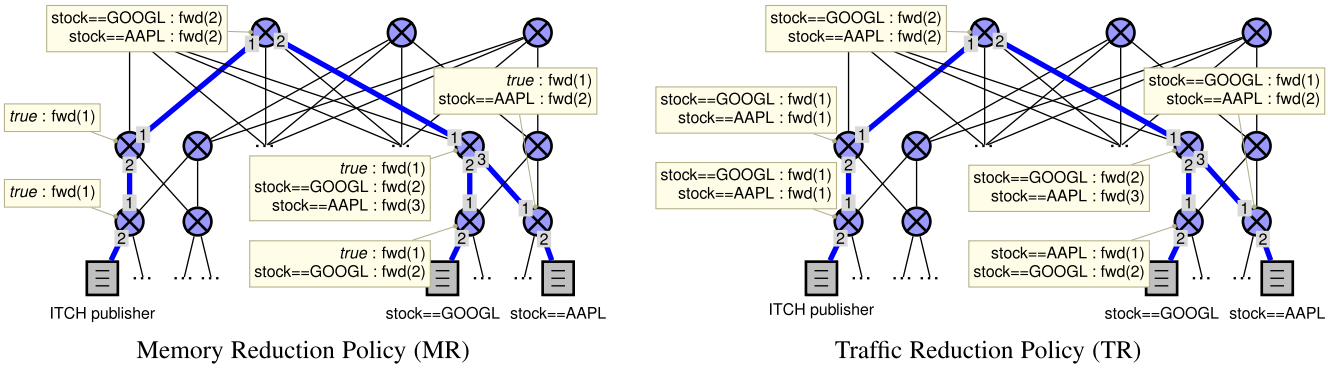


Fig. 3. Routing on subscriptions in the fat tree topology: routing policies.

data plane is configured once using the user-provided packet format specification. Camus relies on programmable switch hardware [1], [2] to realize an application-specific packet processing pipeline. Our prototype implementation uses the P4 compiler to program the switch.

More specifically, the Camus compiler takes the generated rules together with the P4 packet header specification, and generates two outputs: (i) a P4 control block that specifies the control-flow and match-action tables in the pipeline, and (ii) a set of control-plane rules to populate the tables. The P4 compiler then takes the P4 parser specification (packet format) and the control block generated by the Camus compiler to generate the switch image for the packet processing pipeline. We discuss the details of compiling and forwarding in Section V.

Forwarding with packet subscriptions requires that the generated pipeline can parse application-specific message formats, which are often deep in the packet header. In Section VI, we present techniques to read deep into the packet, and process messages that have been batched together into a single network packet.

IV. Routing on Subscriptions

Routing, in general, is a complex issue that raises a number of challenges. We start with hierarchical topologies that are typical in data centers, which are the primary focus of this paper. Then we also show that similar routing schemes can also work on general network topologies.

A. Hierarchical Topologies

We evaluate two approaches to routing with packet subscriptions on an arbitrary hierarchical datacenter topology. These two strategies explore trade-offs between memory and traffic. Neither of these schemes is "better", in the sense that the choice between them depends on the needs and resources of the specific network.

B. Expected Datacenter Deployment

We assume a static, hierarchical network such as a Fat Tree [29] topology. This removes some of the complexity of routing since the topology already enforces tree-like structures in which all simple paths are shortest paths. Figure 3 illustrates an example with three levels in the hierarchy: a top-of-rack layer (ToR), an aggregate layer, and a core layer.

Considering the chosen context (datacenters), Camus currently relies on a centralized controller with a global view of the network. However, there is nothing inherent in the design of packet subscriptions that prevents the use of distributed routing protocols, like BGP or OSPF.

C. Routing Policies

The main task of the controller is to convert the subscriptions into a global routing policy, and then generate local rules for every switch in the network to realize the policy.

A routing policy associates each port p in a switch s with a set of filters F_p^s. Switch s forwards an incoming packet to all the ports p, other than the ingress port, such that the packet matches at least one filter in F_p^s. Figure 3 shows these associations for two policies that we discuss below. The diagrams focus on the router along a particular set of paths (a tree) taken by messages originating from an ITCH publisher on the left-hand side of the network, and going to two subscribers on the right-hand side. For each switch s along those paths, the diagrams show the local forwarding rules derived from the corresponding sets F_p^s.

Since we focus on specific paths, in the diagrams we refer to each specific port numbers. However, Camus treats the upward ports of a switch—those that link to higher-layer switches—as a single logical *up* port. For example, ports 1 in the ToR and aggregate layers are all *up* ports. When forwarding a packet to the *up* port, Camus actually chooses one of the corresponding physical ports, at random or round-robin (ECMP could be used for flow-based protocols). Also, a packet received on one of the upward ports is never forwarded to the *up* port.

As a general correctness condition, F_p^s must match a *superset* of the packets identified by the subscriptions of the hosts reachable from switch s through port p (completeness). And when port p leads directly to a host h, F_p^s must match the *exact* set of packets to which h has subscribed (soundness). Different correct policies may then differ in how precisely each set F_p^s approximates the exact set of packets that must be forwarded from s through p. Intuitively, a loose F_p^s would require fewer rules and therefore less switch memory, but would also generate unnecessary traffic.

Algorithm 1: Routing in a Fat Tree Network

Input: $Network = (Hosts, Switches, access, up, down)$
Input: $access : Hosts \rightarrow (Ports \times Switches)$
Input: $up : Switches \rightarrow (Ports \times Switches)^*$
Input: $down : Switches \rightarrow (Ports)^*$
Input: $Subscriptions : Hosts \rightarrow Filter^*$
Output: A set of sets of subscriptions $F = \{F_p^s : s \in S, p \in P\}$

foreach *switch* s *and port* p **do**
  $F_p^s \leftarrow \emptyset$
foreach $h \in Hosts$ **do**
  $s, p \leftarrow access(h)$ ▷ h connects to s on port p
  $F_p^s \leftarrow F_p^s \cup Subscriptions(h)$
foreach $src \in Switches$, *bottom up* **do**
  $F^{src} \leftarrow \emptyset$ ▷ local, temporary set
  foreach $p \in down(src)$ **do**
    ▷ bottom up, so F_p^{src} already computed
    $F^{src} \leftarrow F^{src} \cup F_p^{src}$
  foreach $dst, q \in up(src)$ **do** src connects to dst on dst's local port q
    $F_q^{dst} \leftarrow F_q^{dst} \cup F^{src}$
if *memory policy* **then** Memory Reduction
  foreach $src \in Switches$ **do**
    $F_{up}^{src} \leftarrow \{true\}$
else if *traffic policy* **then** Traffic Reduction
  foreach $src \in Switches$ **do**
    $dst, q \leftarrow$ first up link $\in up(src)$
    $F_{up}^{src} \leftarrow \emptyset$
    foreach $p \in down(dst)$ **do**
      if $p \neq q$ **then**
        $F_{up}^{src} \leftarrow F_{up}^{src} \cup F_q^{dst}$

Camus implements two policies: one that favors memory (MR) and one that favors traffic (TR), illustrated in Figure 3. In the first policy, every downward port d is associated with a set F_d^s that matches the *exact* (minimal) set of packets that are of interest to hosts reachable through d, while F_{up}^s is the *true* filter that matches every packet. In the second policy, F_{up}^s also matches the exact and therefore minimal set of packets that are of interest to hosts reachable through (one of) the up port. The controller uses Algorithm 1 to compute the filter sets F_p^s for all switches and ports.

D. Filter Approximation Scheme

In addition to the somewhat crude approximation used in the memory reduction policy that simply replaces all the filters associated with an *up* port with a single *true* filter, we also develop a more refined approximation based on filter rewriting. This rewriting is specifically designed to control the amount of false positives, and at the same time to favor *aggregation* of filters. This aggregation is particularly beneficial in combination with the optimizations performed by the Camus compiler for each local switch.

The general idea of this approximation scheme is to rewrite individual constraints so as to reduce the number of *unique* constraints. One way to do that, is to discretize and therefore cluster the comparison constants used in the constraints. In particular, Camus rewrites all numeric constants as multiples of a chosen discretization unit α. For example, choosing $\alpha = 10$, Camus would rewrite constraints `price > 53` and `price > 57` as `price > 50`, and correspondingly constraints `price < 53` and `price < 57` as `price < 60`. As we show experimentally in Section VIII, this simple scheme leads to significant improvements in the compilation time and also in the aggregation of filters, at the expense of only a modest increase in traffic.

Once the sets of filters are computed for each link, the Camus controller turns these sets of subscriptions into an intermediate representation, which is then compiled and installed onto the switch. The intermediate representation appends a forwarding directive to the subscription filter.

Returning to the running example, if the trading application running on a server connected to port 1 of a switch is interested in ITCH messages about Google stock, then the localized rule at the last hop switch would be:

```
stock == GOOGL: fwd(1)
```

The rule asserts if the field `stock` is equal to the constant `GOOGL`, then the message should be forwarded to port 1. A forwarding action may be unicast or multicast:

```
stock == GOOGL: fwd(1,2,3)
```

In this case, messages are forwarded to ports 1, 2, and 3.

The routing schemes we describe require that predicates be re-evaluated at every switch that packets pass through. We considered, but did not implement, an alternative design in which paths through the network would be enumerated and each associated with a "tag". Predicates evaluated at the edge would attach a tag indicating the path that the packet should take. This design seemed like an appealing way to reduce the work done by each switch. However, since multiple predicates can match on a given packet, one would need to attach multiple path tags to each packet. Then the switch would need to check for equality on all of these tags. So, there would not really be an advantage; we would just need to generate an equivalent rule that matches on tags instead of application header fields. Nevertheless, we expect that there should be methods to optimize the storage of rules, since there could be overlap in rules at higher levels in the hierarchy.

E. General Topologies

Routing in general topologies presents challenges not present in hierarchical topologies. Hierarchical topologies are loop-free by construction (in combination with extremely simple forwarding rules) even for multicast flows such as the ones induced by subscriptions. With general topologies, the routing scheme must be designed to avoid loops. A general

way to do that is to route on trees as suggested in previous work [30].

At a basic level, the control plane builds one or more spanning trees, each with its associated FIB entries computed from the subscriptions. As in the TR routing policy described in Section IV-A, each edge (u, v) in a tree defines a partition of the subscriptions of the whole network into two sets: those on the u side (including those of u, if any), and those on the v side (including v). So, the FIB on u must contain rules that represent all the subscriptions on the v side, and must assign those rules to the port that connects u to v. And vice-versa, the FIB on v must contain rules that represent all the subscriptions on the u side, assigned to the port that connects v to u. With these FIBs in place, every packet is initially assigned to and therefore routed within one tree.

This basic scheme can be realized in various ways with many generic and specific optimizations. There are various optimization criteria when constructing the trees. For example, minimizing the distances between nodes, on average or for high-intensity publishers, or reducing the memory (FIB entries) required on all or specific switches. It is also possible to aggregate the FIBs associated with different trees [30] so that the overall memory complexity is sublinear in the number of trees. There are also various strategies for assigning packets to trees.

As stated earlier, our primary focus in this paper is on data-center deployments, so we present a routing scheme for general topologies only at a high level, and we limit our analysis only to memory usage.

In particular, we present two tree-construction algorithms that are intended to work well with the BDDs built by Camus. Both algorithms are based on minimal spanning trees (MST) computed using Prim's algorithm. The first algorithm, which we simply refer to as *MST*, construct a minimum spanning tree considering all edges having equal weight $w(u, v) = 1$. With uniform weights, this algorithm does not select any particular MST and therefore serves as a general baseline.

The second algorithm, which we call *MST++*, uses a heuristic edge-weight function $w(u, v) = deg(u)deg(v)$, where $deg(v)$ is the degree of vertex v. This leads to low-degree spanning trees. (Finding an MST of *minimal* degree is NP-hard.) Using a low-degree tree means that, in each switch, all the subscriptions are partitioned into a few groups each associated with a local port. This in turn allows the Camus compiler to achieve higher rates of compression for its BDD. The experimental evaluation presented in Section VIII-V shows that the *MST++* algorithm is indeed effective in reducing memory usage (FIB entries).

V. Forwarding With Subscriptions

With local subscriptions assigned by the controller to a switch, Camus sets up the forwarding structures on that switch. The key challenge is to compile subscriptions to forwarding structures that are memory efficient and run at line-rate.

Camus compiles the rules into two steps: *static* and *dynamic*. The *static* step is performed once per application, and generates the packet processing pipeline (i.e., packet parsers and a sequence of match-action tables) deployed on the switch. The *dynamic* compilation step is performed whenever the subscription rules are updated, and generates the control-plane entries that populate the tables in the pipeline.

```
1  header itch_order {
2      bit<16> stock_locate;
3      ...
4      bit<32> shares;
5      bit<64> stock;
6      bit<32> price;
7  }
8  @pragma query_field(itch_order.shares)
9  @pragma query_field(itch_order.price)
10 @pragma query_field_exact(itch_order.stock)
11 @pragma query_counter(my_counter, 100, 1024)
```

Fig. 4. Specification for ITCH message format.

This compilation strategy assumes long-running, mostly stable filters. Supporting highly dynamic filters would require an incremental algorithm. Prior work has demonstrated that such incremental algorithms are feasible. BDDs—our primary internal data structure—can leverage memoization [31], and state updates can benefit from table entry re-use [32].

A. Compiling the Static Pipeline

In general, a packet processing pipeline includes a packet parsing stage followed by a sequence of match-action tables. The compiler installs different pipelines for each application, as different applications require different protocol headers, packet parsers, and tables to match on header fields.

To generate the static plane, users provide a message format specification, based on data packets structured as a set of named attributes. Each attribute has a typed atomic value. For example, a particular ITCH data packet representing a financial trade would have a string attribute called `stock`, and two numeric attributes called `shares` and `price`.

Figure 4 shows the specification for the ITCH application. The message format specification extends a P4 header specification with annotations that indicate state variables and fields that will be used by the filters. In the figure, lines 8–11 contain annotations indicating that the fields `shares`, `price`, and `stock` from the `itch_order` header will be used in subscriptions. Thus, the compiler should generate P4 code that matches on those fields. As an optimization, users may specify the match type. The annotation on line 10 specifies that the match should be exact by appending the suffix `_exact`. The annotation on line 11 declares a counter state variable. The first argument is the name of the counter (`my_counter`) and the second is its window size (100μs).

To support state variables, the compiler statically pre-allocates a block of registers that are then assigned to specific variables dynamically. The compiler also outputs the necessary code to update state variables in response to subscription actions at periodic intervals—e.g., to implement the tumbling window used on line 11 in Figure 4. Notice that the use (read/write) of state variable is determined by subscription rules, which are not known statically. Therefore, the static compiler outputs generic code for various update functions,

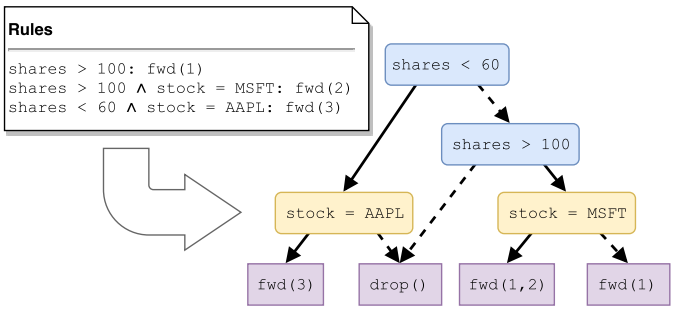


Fig. 5. BDD for three rules. Solid and dashed arrows represent true and false branches, respectively.

and the dynamic compiler effectively links subscription actions to that code. In particular, the dynamic compiler links an update action of the general form $var \leftarrow fun(args)$ with a subscription action by associating that action to what amounts to pointers to var, fun, and $args$. However, the dynamic compiler in our current prototype only supports actions without arguments.

B. Compiling Dynamic Filters

A naïve approach for representing subscription rules would use one big match-action table containing all the rules—each rule would be encoded using a single table entry. However, this approach would be inefficient because the table would require a wide TCAM covering all headers but containing only a few unique entries per header. Furthermore, programmable switch ASICs only support matching a single entry in a table, but a packet might satisfy multiple rules. Hence, we would require a table entry for every possible *combination* of rules, resulting in an exponential number of entries in the worst case.

Instead, our compiler generates a pipeline with multiple tables to effectively compress the large but sparse logical table used by the program. To do this, the compiler represents the subscription rules using a binary decision diagram (BDD) [21], [22]. BDDs are often used to obtain compact representations of functions on a wide input domain for which a single table would be too large. A BDD is a rooted acyclic graph in which non-terminal nodes encode conditions on the input and terminal nodes encode the results (see Figure 5).

The evaluation of the overall function of the BDD that encodes all subscription rules starts at the root node and recursively evaluates the conditions (if) at each node, proceeding to the true (then) or false (else) branch as appropriate. Evaluation terminates when it reaches a terminal node (actions).

We now briefly describe the algorithm for building a BDD out of subscriptions rules. What is important for our purposes is to define the structure of the BDD, so we can implement the BDD evaluation as a sequence of table lookups.

C. Representing Rules with a BDD

The subscription rules are first normalized into disjunctive form, yielding a set of independent rules in which the condition in each rule consists of a conjunction of atomic predicates. An atomic predicate is defined by an equals, greater-than, or less-than relationship between a field and a constant. For example, the rules in Figure 5 are in disjunctive normal form. The compiler then builds the BDD incrementally by evaluating the condition at each node using the Shannon expansion and adding nodes for the predicates in the condition as needed.

The compiler reduces the BDD using a combination of standard and domain-specific transformations. (i) If two nodes are isomorphic, one is deleted. The incoming edges of the deleted node are updated to point to the remaining copy. (ii) If both outgoing edges of a node point to the same successor, then that node is deleted. The incoming edges of the deleted node are updated to point to the successor. (iii) If any ancestors n' of a new node n implies that n is always true or always false (based on domain-specific knowledge of the filter), then n is not added; instead, it reduces to a direct connection to its true or false branch, respectively. The overall effect is to share common structure and remove redundant nodes and unsatisfiable paths [33].

As is standard in ordered BDDs, the conditions in the BDD are arranged in a fixed order. For example, every path in the BDD of Figure 5 consists of a sequence of atomic predicates such that the conditions on field `shares` precede the conditions on field `stock`. This is essential for the representation and evaluation of the BDD as a sequence of table lookups, as we discuss next. The choice of an order can significantly impact the size of a BDD. Determining an optimal field order is NP-hard, but simple heuristics often work well in practice.

D. BDDs to Tables

The BDD can be seen as a state machine, where each state corresponds to a predicate, and the transition function is the evaluation of the predicate on the input packet. However, this naïve evaluation would require an excessively long sequence of evaluation steps. We instead implement BDD evaluation using a fixed-length pipeline.

Since every path in the BDD traverses predicates that consider fields in order, and that order is the same for every path, we use that ordering to effectively slice the BDD into a fixed number of field-specific components. Each component is a subgraph of the BDD that contains all and only those nodes that predicate on a particular field. By extension, we also consider the set of terminal nodes as a component. For example, the BDD in Figure 5 has three components consisting of the blue, yellow, and red nodes, corresponding to the `shares` and `stock` fields, and to actions, respectively.

We can now consider the evaluation of the BDD as a state-machine at the level of the field-specific components. Thus the transition function out of the component of field f depends on the value of field f in the packet. However, since the component of field f is a macro-state corresponding to potentially many states of the BDD, the transition function must also depend on the BDD state in which we enter the component. This entry BDD state and the value of field f are necessary and sufficient to determine the path through the component of field f and therefore the transition function

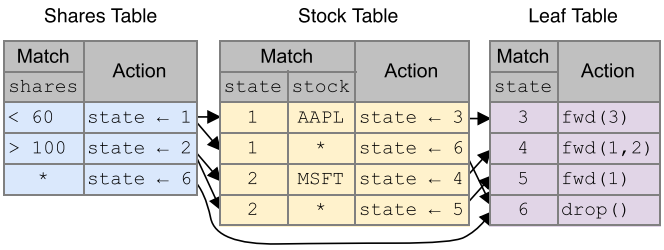


Fig. 6. Table representation of the BDD in Figure 5.

Algorithm 2: Translating BDD to Tables

Input: The BDD graph, G
Output: A set of tables $T_f : state \times dom(f) \rightarrow state$

```
foreach field f do
    C_f ← subgraph of G predicating on field f
    In ← {n ∈ C_f with in-edges from outside C_f}
    Out ← {n ∉ C_f with in-edges from C_f}
    foreach path p = (u ∈ In, ..., v ∈ Out) in C_f do
        range ← ⊤        ▷ all allowable values for field f
        foreach node n ∈ p do
            range ← range ∩ predicate(n)
        T_f ← T_f ∪ {(u, range) ↦ v}
```

for that component. We represent this transition function as a match-action table where we match on the entry state and on the value of field f, and where the action points to the next component and BDD state.

Figure 6 shows all the component-specific match-action tables corresponding to the transition functions for the BDD of Figure 5. The three tables also define the three-stage processing pipeline. The evaluation through the pipeline stores the current BDD state in metadata. The initial state is set to 0 and can be omitted entirely from the first table. The actions define the entry state for the next stage, except for the *Leaf* table where the action corresponds to the overall evaluation of the BDD. For example, the rightmost path through the BDD in Figure 5 corresponds to the path through the 2nd, 4th, and 3rd entries of the *Shares*, *Stock*, and *Leaf* tables in Figure 6.

Notice that it is possible for multiple rules to match the same packet. For example, in Figure 5, the first two rules could match the same packet, so the actions `fwd(1)` and `fwd(2)` are merged into the single action `fwd(1,2)`. The compiler translates this to forwarding to a multicast group that comprises ports 1 and 2.

We compute the transition tables with Algorithm 2. In essence, for each field-specific component C_f in the BDD, Algorithm 2 identifies a set of *In* nodes within C_f that are the destinations of all the edges that enter C_f from components of preceding fields, and a set of *Out* nodes outside C_f that are the destinations of all the edges that exit from C_f to components of succeeding fields. Then Algorithm 2 computes the transition table by iterating over all the paths that connect *In* and *Out* nodes. In general, a BDD could have an exponential number of such paths. However, the domain-specific optimizations we use guarantee that there is at most one path between any pair of *In* and *Out* nodes, which in turn guarantees that the number of paths is at most quadratic.

E. Resource Optimizations

One of the scarce resources in switching ASICs are TCAM memories that allow matching on a subset of bits in headers but consume large area of die and high power. The compiler uses three techniques to reduce TCAM usage. First, by default the compiler generates P4 code that implements range matches, which usually require an expensive TCAM lookup. However, the user can guide the compiler by specifying a matching type for each field that may not require a TCAM lookup. Second, matching on a range in TCAM is not scalable to hundreds of thousands of ranges as each range-match requires multiple TCAM entries ($O(\#bits)$). To cope with this, the compiler uses exact matches instead of range when possible, allowing it to leverage SRAM while saving TCAM. Third, some fields, like `shares`, will probably have only a few unique range predicates. The compiler can map values for that field and the corresponding range predicates onto a lower-resolution domain (e.g., 8-bits).

VI. Efficient Packet Parsing

There are two major challenges that Camus must address with respect to packet parsing: (1) generating multiple copies of a packet with different subsets of messages, and (2) parsing deep inside the packet to handle all application messages.

There are three key observations about the functionality of the hardware that we leverage for efficient parsing. First, a switch may advance arbitrarily deep in a packet in the parsing stage if the packet bytes are not written to the Packet Header Vector (PHV) that is sent through the programmable pipeline. Second, packets may only be replicated in the cross-bar between the ingress and egress pipelines. Third, a number of ports may be dedicated as recirculation or loopback ports, which re-send packets back through the pipeline. Recirculating a packet effectively increases the depth of the processing pipeline, allowing for additional processing at the expense of slightly increased latency and reduced overall throughput.

A. Per-Subscriber Subsets of Messages

To send different subsets of a message to different destinations, Camus uses the following strategy. First, the ingress pipeline creates a port mask that indicates which messages should be sent to which port on a switch. Next, in the crossbar, Camus replicates the packet, creating a separate copy for each output port. Finally, at egress, Camus uses the port mask to prune different messages from each of the replicas, filtering appropriately for each port. To avoid sending the port mask from ingress to egress naïvely, which would add bandwidth overhead, Camus uses a domain-specific optimization. It stores the mask in an unused packet header field (e.g. `ethernet.srcAddr`, which will be overwritten at the end of the pipeline anyway).

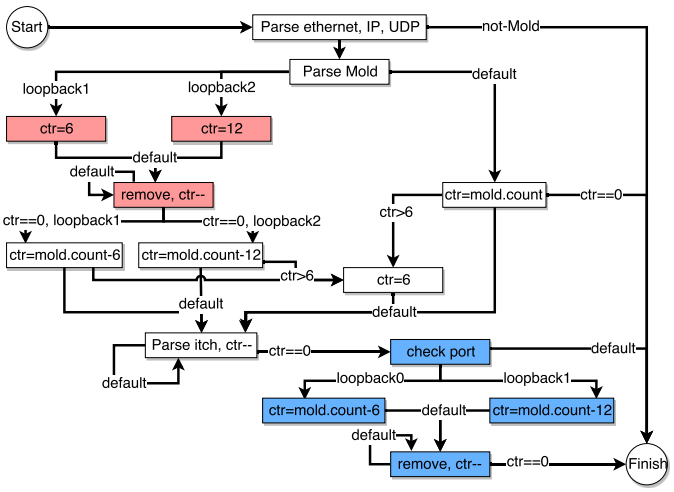


Fig. 7. ITCH ingress parser for 3 loopback ports.

B. Parsing Deep

Since hardware switches have limited memory for carrying packet data through the programmable pipeline, it may not be possible to parse all application-level messages in a single pass. Figure 7 shows the state machine for parsing deep into the packet. Camus processes the messages in multiple parallel passes after the first pass. In the first pass, it multicasts packets on recirculation ports to make multiple copies. When the copies return to ingress, each recirculation port starts parsing at different offsets of each copy. This technique is implemented in a parser loop: the red boxes in Figure 7 show that in each iteration the parser matches on a counter (ctr); updates the counter; and shifts the parser buffer without extracting any headers. Then, the parser reads messages and finally truncates the outgoing packets using another loop to remove the messages that cannot be parsed (blue boxes). This design is also extensible to multiple recirculation passes.

VII. Discussion

The main goals of this work are to: (i) provide applications with a more expressive interface to the network, while (ii) providing high throughput and low latency communication. Although these two goals are not inherently incompatible, there is a tradeoff between the two. To realize a practical solution, we had to consider a few delicate aspects:

A. Fixed-Width Headers

To support packet processing at line rate, Camus only supports fixed-width headers. This is not an issue for many applications that already use fixed-width headers (e.g., ITCH, ILA, INT). Other applications can use libraries like Flatbuffers [34] which use binary serialization formats.

B. Streams

As the name implies, packet subscriptions operate on individual packets (or datagrams). The design in this paper already supports subscribing to a stream if each packet in the stream contains the same header that is used by the filter. Subscribing to streams where the header is only present in the first packet would require the switch to store the matching rule of the first packet, and apply it to subsequent packets in the stream.

C. Multicast Groups

When multiple filters overlap, i.e. they match the same packet, the packet may have to be forwarded out multiple ports. Camus handles this by associating a multicast group with the set of overlapping filters. Multicast groups are a limited resource on the switch, so Camus cannot support an unbounded number of overlapping subscriptions. In practice, subscriptions are selective, so the number of multicast groups on the switch is not a limiting factor (see our evaluation in Section VIII-IV).

VIII. Evaluation

The central thesis of this paper is that the network can and should provide higher-level abstractions than those currently offered. To evaluate this claim, the first part of our evaluation focuses on cases studies. We relate our experiences building eight diverse applications to demonstrate qualitatively that packet subscriptions are expressive and beneficial. The second part of our evaluation focuses on operational aspects, demonstrating that the abstraction can be practically and efficiently implemented and deployed. Overall, the evaluation is structured along five main research questions:

- Q1: Are packet subscriptions useful and expressive enough to implement a diversity of applications?
- Q2: Does the Camus architecture have a reasonable and practical design?
- Q3: What is the impact of packet subscriptions on application performance?
- Q4: Is forwarding with packet subscriptions efficient, in terms of performance and memory?
- Q5: Is routing with packet subscriptions efficient, in terms of traffic load and memory (FIB)?

A. Implementation

The evaluation uses our prototype compiler implementation, which was written in OCaml, and is publicly available.[1] The compiler parses the application specifications written in $P4_{16}$ using the p4v library [35], patched to support our custom annotations. We use our own implementation of a multi-terminal BDD library with reduction optimizations.

B. Experimental Setup

We ran Camus in our cluster with three 32-port Barefoot Tofino switches and four servers, with varying topologies. Each server was running Ubuntu 16.04 on 12 cores (dual-socket Intel Xeon E5-2603 CPUs @ 1.6GHz), 16GB of 1600MHz DDR4 memory, and Intel 82599ES 10Gb or Mellanox ConnectX-5 100Gb Ethernet controllers.

[1] https://github.com/usi-systems/camus-compiler

C. Q1: Expressiveness

To evaluate the utility and expressiveness of packet subscriptions, we implemented eight different applications. We chose these applications because they: (i) represent a diverse domain space, (ii) exercise packet subscriptions in different ways, and (iii) are inspired by real-world applications and use-cases.

1) Market Data Filter: Financial exchanges, such as the Nasdaq stock market and the Chicago Mercantile Exchange, publish price and trade-related information in market data feeds. Different exchanges use different message formats. Nasdaq publishes data in the ITCH format.

ITCH data is delivered to subscribers as a stream of IP multicast packets, each containing a UDP datagram. Inside each UDP datagram is a MoldUDP header containing a sequence number, a session ID, and a count of the number of ITCH messages inside the packet. There are several ITCH message types. An `order` message indicates a new order that has been accepted by Nasdaq. It includes the stock symbol, number of shares, price, message length and a buy/sell indicator. In the descriptions below, packet subscriptions can refer to fields in the traditional header stack, or in the application-specific message format.

This application is an implementation of the Nasdaq ITCH Market data feed filter and router. The feed is delivered as a stream of IP multicast packets. The switch splits ITCH packets into multiple messages, and forwards them to back-end servers based on the subscription.

2) Network Telemetry Analytics: One recent approach to network monitoring is to collect fine grained statistics on every packet that passes through a switch [36], [37]. Once the data is collected, it is sent to an analytics system for processing, such as Barefoot Networks Deep Insight [38] or Broadcom's BroadView Analytics [39]. These analytics systems are usually built following the Lambda architecture design [40] (e.g., Spark [41] for anomaly detection, Cassandra [42] for storage, and Kafka [9] for communication), which requires each of the components to scale out to cope with the input load. We used packet subscriptions to filter and route interesting (i.e., anomalous) events from the stream of in-band network telemetry (INT) data. For example, the subscription can select events indicating flows that experience high latency. In this application, packet subscriptions perform the work normally done by Kafka and Spark.

3) Identifier Based Routing: As mentioned in Section I, to address networking challenges exposed by cloud computing deployments with containerized services, several web-service companies have deployed identifier-based routing, such as ILA at Facebook [5]. ILA is attractive for brownfield deployments, as it does not require a custom packet header; the service identifier is stored in the destination address field of the IPv6 header. By decoupling the service identifier from the service locator, services can migrate without requiring changes (or notifying) clients that wish to use the service. This can be useful for load balancing, where the client does not need to contact a specific machine, but any machine providing the service.

We used packet subscriptions to implement an ILA identifier based routing scheme that enables communication with a web service that migrates between servers. Instead of contacting the service using a physical address, the client makes a request by specifying a service identifier. The server running the web service subscribes to packets matching the identifier. When the service migrates to another server, the subscription is updated.

4) Video Streaming: Video streaming is a powerful use case for pub/sub communication, since there is a single publisher and an unknown number of subscribers. Because many subscribers want the same content, network bandwidth can be reduced by caching copies in the network.

Prior work on Information Centric Networking (ICN) made this same observation, and several systems have implemented some combination of pub/sub communication and in-network caching [43], [44]. Notably, Cisco has recently developed an ICN-style network architecture to address the problem of streaming to clients in unknown locations in 5G networks [45]. Their system, named hybrid ICN (hICN) [46] embeds a content identifier in an IP address, allowing content to be routed in a heterogeneous (hybrid) network of standard L2 hardware switches and software-based hICN forwarders. The hICN forwarders serve as content caches. While the caches can reduce latency and improve bandwidth utilization, they are only effective for "hot" content that is likely to reside in the cache. If there are many cache misses, then the software-based forwarder can become a bottleneck.

The design of the system could be improved if packets were only sent to a forwarder if a cache hit were likely. We used Camus to implement this improved design. We wrote Camus subscription filters that refer to meter state and content identifier. The filters only route packets to the software forwarders if the meter rate exceeds a threshold (i.e., a cache hit is likely). Otherwise, packets are sent upstream through other hardware switches, on the path to the original producer of the content.

5) DNS Resolver: We implemented a DNS resolver that maps a host name to an IP address. This application differs from the others, in that we extended the subscriptions with a custom action. In Section I, we discussed the `fwd` action that is built into the compiler. For the resolver, we added a new action, `answerDNS(ip)`, that resolves a query to an IP address; crafts a DNS response; and sends it back to the source. A DNS entry can be added with a single subscription rule. For example,

```
name == h105: answerDNS(10.0.0.105)
```

would reply with a DNS AA (Authoritative Answer) message containing `10.0.0.105` to DNS queries for host 105 (`h105`). If the switch does not have a subscription for a DNS entry, the default action is to forward the request to a DNS server. In this application, packet subscriptions are essentially used to implement a caching layer to reduce response latency and alleviate load on DNS servers.

6) Motor Highway Monitoring: Partially inspired by the classic stream processing benchmark, Linear Road [47], we implemented an Internet of Things (IoT) Motor Highway Monitoring application. The application is used to detect cars that exceed the speed limit for a given highway. We assume

that each car is equipped with a mote that emits 10 position report packets per second. The position report packet contains a fixed-width header indicating the car's identifier, current coordinates and speed. These packets are forwarded through a switch which is connected to a monitoring server. The application detects cars travelling at over 55 mph within areas defined by lat/long coordinate ranges. For example, the rule

```
x > 10 ∧ x < 20 ∧ y > 30 ∧ y < 40
        ∧ spd > 55: fwd(1);
```

would forward a position report to a monitor server on port 1 if the car is speeding in the region bounded by the lat/long range of 30-40 and 10-20, respectively. Although the subscription predicates on many fields, it can be evaluated in a single pipeline pass, without recirculation.

7) Publish/Subscribe: Many application-level middleware messaging services provide pub/sub communication, such as Kafka [9], ActiveMQ [48], and Siena [49]. We implemented an API-compatible replacement for the Apache Kafka message queue [9]. Instead of sending a message to the Kafka server, the publisher sends the message to our switch, which routes the message to the subscribers. Although our Kafka shim does not provide the same features as Kafka (e.g., persistence), it can support higher throughputs, reducing the number of Kafka servers that would otherwise be necessary. We note that our implementation is limited to message sizes of 512 bytes. Sending messages larger than the MTU would require splitting the message between packets. However, 512 bytes is the typical size of a JSON message [50], and fits within the MTU.

8) Traditional IP: To demonstrate that packet subscriptions generalize traditional forwarding rules, we implemented IP. Each host is assigned an IP addresses, and the switches forward packets toward their destinations.

9) Limitations and Scope: The applications we implemented are intended to demonstrate the expressiveness of packet subscriptions. Later we also compare their performance with equivalent server-based applications, as in the case of Apache Kafka. However, this comparison is slightly unfair, since the applications written with packet subscriptions do not provide all the features of software-based pub/sub systems (e.g., replication, fault-tolerant delivery and logging). Still, the comparison is appropriate for many application scenarios in which providing timely data is more important than offering availability of old data. For example, it is common for real-time data analytic systems to disable replication for fault-tolerance to increase performance.

D. Q2: Architecture Practicality

To evaluate the practicality of the Camus architecture, we focused on the following questions: (i) Can multiple applications using packet subscriptions co-exist? (ii) Can packet subscriptions co-exist with traditional IP traffic? (iii) Can packet subscriptions generalize traditional IP traffic? For all the experiments in this section, the main result is "it works"—unexciting, but exactly what we'd hope to see.

1) Multiple Applications: Packet subscriptions for multiple applications can co-exist on the same switch. Assuming that their rules fit within the switch memory and their aggregate throughput does not exceed the bandwidth supported by the switch, there is no additional latency to that typically experienced in the network (e.g., because of queueing).

To demonstrate that applications can co-exist, we deployed both the ITCH and INT applications on the same switch. We used this switch in a topology with three servers: the first server published packets for both applications, and the other two processed either ITCH or INT packets. Both applications ran simultaneously without problems in the same network.

2) Co-Existence with IP: To demonstrate the feasibility of a brownfield deployment, we implemented a basic L2/L3 IP switch, and used it as ToR switch for servers communicating with Kafka. Then, we extended the switch pipelines for two packet subscription applications, ITCH and INT. We measured the Kafka publisher throughput for both switch configurations. In the Camus configuration, we introduced some INT and ITCH traffic and observed that the Kafka traffic, using traditional IPv4 forwarding, was not impacted.

3) Generalizing IP: To demonstrate that packet subscriptions generalize traditional forwarding rules, we used them to implement traditional IP forwarding. We compiled Camus rules to forward IP packets. We deployed this on a switch connecting a cluster of four servers running an unmodified Kafka application. The application worked as expected.

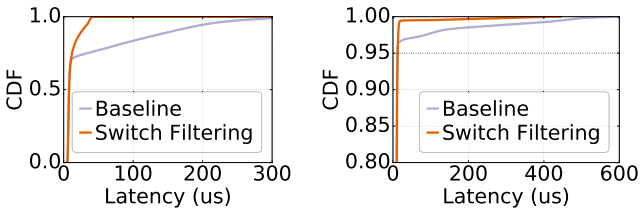
E. Q3: Application Performance

To evaluate the impact on application performance, we compared the in-network applications to traditional software based alternatives in terms of both latency and throughput.

1) Market Data Filter: We evaluated latency in the context of processing ITCH messages. In the experiment, a publisher sends a feed of ITCH messages to a subscriber that filters the feed for *add-order* messages with stock symbol `GOOGL`. We measured the end-to-end latency, between publication and delivery.

The publisher and subscriber (both using DPDK) are collocated on a server for accurate timestamping, and communicate using separate NICs connected via the switch. We ran the experiment in two configurations: in the baseline configuration, the subscriber performs the filtering; in the second configuration, the filtering is done on the switch with Camus. We used two workloads: a Nasdaq trace from August 30th 2017 and a synthetic feed with multiple ITCH messages per packet (Zipf distribution). The number of messages of interest (i.e. for `GOOGL`) is 0.5% of the Nasdaq trace, and 5% of the synthetic feed. We sent the feeds at 8.25 Mpps, which is 90% of the maximum filtering throughput of the subscriber. To be clear, Camus can run at a much higher throughput, but we slowed down the feed rate for the subscriber.

Figure 8 shows the latency CDF for both workloads. For the Nasdaq trace, all messages arrived within 50us with Camus, compared to 300us for the baseline. For the synthetic workload, 99.5% of the messages arrived within 20us with Camus, compared to 96.5% with the baseline. The synthetic feed has multiple messages per packet, which requires packet replication within the switch to split the packet; nevertheless,



(a) Nasdaq trace, 1 msgs/pkt (b) 5% GOOGL, 1-12 msgs/pkt

Fig. 8. ITCH experiments with two different workloads.

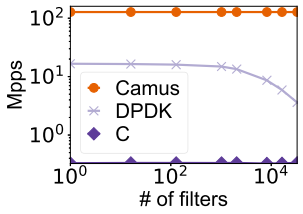


Fig. 9. Filtering INT packets from a 100G link.

the latency is still significantly lower than that of the baseline. Overall, filtering messages on the switch with Camus reduces the tail latency, allowing applications to meet their latency requirements under high throughput.

2) Network Telemetry Analytics: For the INT filtering application, we compared with a C program running in userspace and a C program using DPDK, both running on a commodity server. To generate a high-bandwidth packet stream, as one would expect when collecting data from a realistic datacenter, we used a switch to generate a stream of INT packets on a 100G link. We used filters that match less than 1% of the packets. The filters check that the INT packet passed through a switch with a latency above a threshold, for example:

```
int.switch_id = 2 and int.hop_latency > 100
```

Figure 9 shows the results. DPDK has better performance than the plain C program, but is fundamentally limited by the CPU clock speed: at 1.6GHz, spending about 100 instructions per packet, DPDK can process 16 Mpps. Camus, on the other hand, processes the whole 100G stream at line rate. Moreover, we found that the software based filtering does not scale with the number of filters: the latency for DPDK drastically increases after 10K filters. Camus installs the filters in hardware memory, so it has low latency, regardless of the number of filters.

3) Streaming Video: There are two benefits of our approach: it reduces the load on the hICN forwarder and reduces the latency for "cold" content by bypassing the cache. To evaluate this claim, we deployed the VPP/DPDK [51] implementation of hICN with the topology in Figure 10. On two clients we ran the hICN performance measurement tool, `hiperf`, to stream content for the same identifier, while in parallel another client pulled content for many different identifiers, which are unlikely to be cached. First we ran a baseline IPv6 setup where all the requests from the clients pass through the hICN

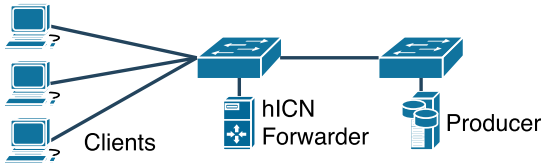


Fig. 10. Topology for hICN video streaming experiment.

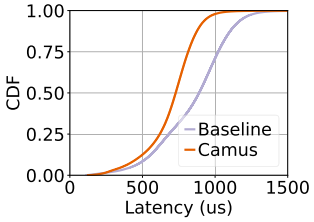
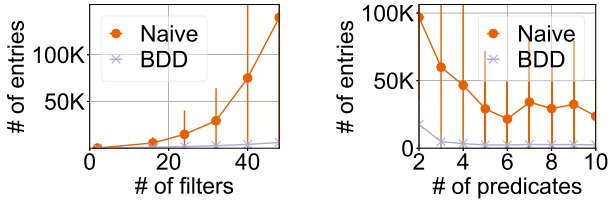


Fig. 11. Lower tail latency for uncached content in hICN.



(a) Number of subscriptions (b) Filter selectivity

Fig. 12. Compiler BDD memory efficiency.

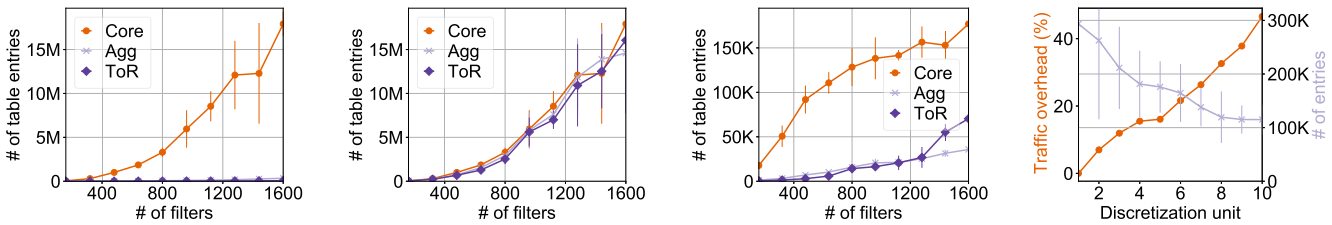
forwarder, which processes packets at about 3.5 Gbps. Then, we ran with the Camus stateful predicates, which only sends "hot" requests to the forwarder. Figure 11 shows the latency for receiving data that is unlikely to be cached. Camus reduces the 95th percentile latency by 21%, because it detects requests for uncached data, and thus avoids the latency of the forwarder. Furthermore, the reduced load on the hICN forwarder allows it to stream to the other clients at a throughput 3% higher than the baseline.

F. Q4: Efficiency of Forwarding

We evaluate the efficiency of forwarding with Camus in terms of performance and memory usage.

1) Performance: Camus can support the full switch bandwidth of 6.5Tbps. The latency of the pipeline, which depends on the application, is less than 1μs.

2) Memory: We compare the memory usage of our compiler to a baseline of naïvely representing all the filters in one big table (see Section V). One big table has entries for all permutations of overlapping queries; workloads with similar queries cause an explosion in the table size. We chose this baseline because it is an intuitive representation of query forwarding logic. We generated workloads using the Siena Synthetic Benchmark Generator [52], which has been used to evaluate prior work in pub/sub systems [49]. Figure 12 shows the total size of the tables (number of entries) as we



(a) MR Policy (b) TR Policy (c) TR Policy + Approx. (d) Approx. Granularity

Fig. 13. Switch memory usage for two hierarchical topology routing policies and approximation.

TABLE I
SWITCH RESOURCE USAGE FOR THREE APPLICATIONS

	SRAM	TCAM	MCast Groups
ITCH (100K filters)	3.7%	4.17%	46%
INT (100K filters)	0.73%	3.45%	0%
hICN (1M filters)	74.9%	0.69%	0%

vary: (a) the number of subscriptions; and (b) the selectiveness of subscriptions (number of predicates).

Given the low growth rate of table entries as workloads become more complex, the experiments show that Camus uses available space effectively, especially compared to the baseline. More selective subscription conditions (i.e. more predicates per filter) require fewer entries, because they result in fewer paths in the BDD.

To understand the memory implications for applications, we measured the switch resource usage for three applications with different filter sizes. For ITCH we generated filters of the form "`stock == S ∧ price > P: fwd(H)`", where `S` is one of a 100 stock symbols, `P` is in the range (0, 1000) and `H` is one of 200 end-hosts. For INT we generated the filters described in Section VIII-III with 100 switches and 1000 `hop_latency` ranges. For hICN (described in Section VIII-III) we used 1M unique content identifiers. Table I shows that these applications are well within the limits of the switch resources. In fact, these applications can be deployed together, or with other standard network functionality. ITCH is the only application that makes heavy use of multicast groups, which is because many end-hosts have overlapping filters.

G. Q5: Efficiency of Routing

To compare the routing efficiency for different design decisions, we measured both memory usage (in hierarchical and general topologies) and compile time.

1) Memory Usage in Hierarchical Topologies: We compare the switch memory usage at various layers of a Fat Tree topology for different routing policies and with the approximation described in Section IV. We used Mininet [53] to emulate the topology depicted in Figure 3 with 20 switches and 16 hosts that publish and subscribe to an ITCH feed. The filters were generated with the synthetic benchmark generator described in the previous section.

Figures 13a and 13b show the memory usage for the two policies with an increasing number of filters, where each filter checks three variables. With the MR policy, the Agg and ToR layer only need to store southbound filters, which uses less memory. On the other hand, the TR policy requires storing the filters from the whole network. The TR policy uses more memory but can utilize the bandwidth more efficiently which results in less congested links in the network.

Figures 13a and 13c show that memory usage is reduced by discretizing the filters while aggregating filters at different layers in the network. Although this approximation reduces both compile time and memory usage, it causes extra traffic in the network. Figure 13d illustrates the correlation between the discretization unit (α) and the percentage of extra traffic forwarded in the core layer of the network.

2) Memory Usage in General Topologies: We compare the memory usage for the two tree-construction algorithms (MST and MST++) described in Section IV-II. We use two network graphs from the SNAP dataset [54]: *CAIDA* is an AS-level graph derived in 2007 and consisting of 26475 nodes and 106762 edges; *AS-733* is another, smaller AS-level graph derived in 2000 and consisting of 6474 nodes and 13233 edges. For each graph, we generate two spanning trees, one using the MST algorithm, and one using the MST++ algorithm, and we carry out a series of experiments on those four trees.

We assign packet subscription rules to some randomly selected nodes. We select increasing numbers of nodes to obtain an increasing total number of subscriptions. Each selected node gets either one or ten rules depending on the experiment. This is intended to test scenarios with more distributed or centralized applications, respectively. Rules are based on two variables in all the experiments. After assigning subscription rules to switches, and referring to one of the four trees, we compute and assign forwarding rules to all the switches in the network as described in Section IV-II. Finally, we run the Camus compiler for each switch, and record the maximal number of actual table entries produced by the compiler for any switch. For each network graph and tree, we run repeated trials for this experiment (more than ten) for each X-axis value, and plot the median of the maxima obtained in each trial.

Figure 15 shows the results of this evaluation. The results for the MST algorithm serve as a baseline and already demonstrate that routing on general topologies is feasible even without specific optimizations. The results for the MST++ algorithm show that Camus admits to specific and effective optimizations.

3) Dynamic Reconfiguration: When subscriptions change, or if there is a change in network topology (e.g., caused by

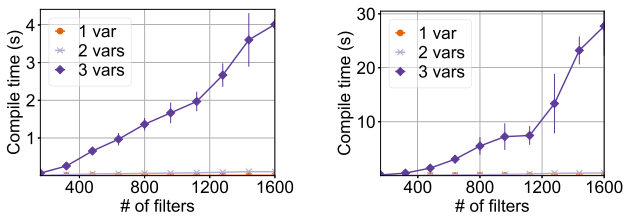


Fig. 14. Compile time for two routing policies.

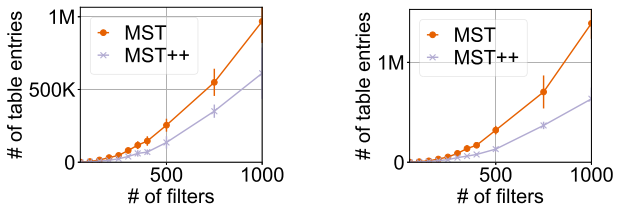


Fig. 15. Comparison of switch memory usage for tree selection algorithms in two general topologies from the SNAP dataset.

a link or switch failure), the runtime table entries need to be recompiled. We evaluate how the compile time of runtime table entries is affected by the number of subscriptions and the number of variables each subscription contains. Figure 14a shows the time to recompile the BDD tables for the MR policy with $\alpha = 10$. These results are similar for TR (Figure 14b) and are two orders of magnitude faster than without approximation (i.e. $\alpha = 1$). The bottleneck is compiling the ToR layer, since it stores all the original (i.e. unapproximated) subscriptions. Whereas the MR policy only needs to recompile entries for 8 out of the 20 switches, the TR policy must recompile entries for all switches. Moreover, the compile time is negligible for 1 to 2 variables, which is a reasonable number of variables for applications like INT and ITCH.

IX. Related Work

Packet subscriptions can be seen as a domain-specific language for networking. Another perspective is to see packet subscriptions as a network-level implementation of publish/subscribe messaging system.

A. Network Programming Languages

Several languages support the configuration of networks of programmable switches, including Frenetic [55], Pyretic [56], Merlin [57], and NetKAT [58]. Packet Subscriptions differs from this work in that it is more than a control-plane language for network designers or administrators. In particular, it provides stateful filtering rules that realize a form of in-network processing, and therefore amount to data-plane programs. The Marple [59] language also evaluates queries in-network, but for the domain-specific application of network telemetry. The packet subscription compilation algorithm is similar to the FDD compilation algorithm by Smolka *et al.* [31].

B. Publish/Subscribe Messaging Systems

Packet subscriptions are comparable to application-level middleware messaging services, such as Kafka, ActiveMQ [48], and Siena [49]. Eugster *et al.* provide a comprehensive survey of pub/sub systems [11]. Packet subscriptions are also comparable to the large body of prior work on information-centric networking (ICN) [60]–[62]. ICN is founded on the idea of addressing data packets using symbolic names rather than network addresses. Prior work reports throughput limits that are well below those of packet subscriptions. Also, notice that these systems implement a stateless prefix matching, which is a problem that is significantly simpler than the content-based and stateful filtering of packet subscriptions.

X. Conclusion

Today, networks provide a lower level of abstraction than what is expected by modern distributed applications. This paper argues that the emergence of programmable data planes has created an opportunity to resolve this incongruity, by allowing the network to offer a more expressive interface.

The core technical contributions of this paper include the design of an expressive filter language that generalizes traditional forwarding rules; a set of algorithms for routing with packet subscriptions; and techniques for compiling complex filters to reconfigurable network hardware using BDDs.

These techniques are widely applicable to a range of network services. As a demonstration, we have used our prototype controller and compiler to build a diverse set of applications, including a financial application for filtering market feeds; detecting network events using INT; and stateful forwarding of hICN streams. These applications demonstrate predictable, low-latency packet processing using the full switch bandwidth.

References

[1] (2019). *XPliant Ethernet Switch Product Family*. [Online]. Available: https://www.cavium.com/XPliant-Ethernet-Switch-Product-Family.html

[2] P. Bosshart *et al.*, "Forwarding metamorphosis: Fast programmable match-action processing in hardware for SDN," in *Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, Aug. 2013, pp. 1–12.

[3] P. Bosshart *et al.*, "P4: Programming protocol-independent packet processors," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 3, pp. 87–95, Jul. 2014.

[4] D. Clark, "The design philosophy of the DARPA internet protocols," in *Proc. Symp. Proc. Commun. Archit. Protocols (SIGCOMM)*, 1988, pp. 106–114.

[5] P. Lapukhov. (2016). *Internet-Scale Virtual Networking Using Identifier-Locator Addressing*. [Online]. Available: https://www.nanog.org/sites/default/files/20161018_Lapukhov_Internet-Scale_Virtual_Networking_v1.pdf

[6] D. E. Eisenbud *et al.*, "Maglev: A fast and reliable software network load balancer," in *Proc. USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Mar. 2016, pp. 523–535.

[7] N. Shirokov and R. Dasineni. (May 2018). *Open-Sourcing Katran, a Scalable Network Load Balancer—Facebook Engineering*. [Online]. Available: https://code.fb.com/open-source/open-sourcing-katran-a-scalable-network-load-balancer/

[8] (2019). *Tibco Rendezvous*. [Online]. Available: https://www.tibco.com/products/tibco-rendezvous

[9] J. Kreps, N. Narkhede, and J. Rao, "Kafka: A distributed messaging system for log processing," in *Proc. 6th Int. Workshop Netw. Meets Databases (NetDB)*, Jun. 2011, pp. 1–7. [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/09/Kafka.pdf

[10] (2019). *IBM MQ*. [Online]. Available: https://www-03.ibm.com/software/products/en/ibm-mq

[11] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec, "The many faces of publish/subscribe," *ACM Comput. Surveys*, vol. 35, no. 2, pp. 114–131, Jun. 2003.

[12] J. Sherry, S. Hasan, C. Scott, A. Krishnamurthy, S. Ratnasamy, and V. Sekar, "Making middleboxes someone else's problem: Network processing as a cloud service," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, Aug. 2012, pp. 1–12.

[13] R. Potharaju and N. Jain, "Demystifying the dark side of the middle: A field study of middlebox failures in datacenters," in *Proc. Conf. Internet Meas. Conf. (IMC)*, Oct. 2013, pp. 9–22.

[14] X. Li, R. Sethi, M. Kaminsky, D. G. Andersen, and M. J. Freedman, "Be fast, cheap and in control with SwitchKV," in *Proc. USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Mar. 2016, pp. 31–44.

[15] X. Jin *et al.*, "NetCache: Balancing key-value stores with fast in-network caching," in *Proc. ACM Symp. Operating Syst. Princ. (SOSP)*, Oct. 2017, pp. 121–136.

[16] A. Carzaniga and A. L. Wolf, "Content-based networking: A new communication infrastructure," in *Proc. NSF Workshop Infrastruture Mobile Wireless Syst. (IMWS)*, Oct. 2001, pp. 59–68.

[17] T. Koponen *et al.*, "A data-oriented (and beyond) network architecture," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, 2007, pp. 181–192.

[18] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proc. ACM Int. Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, Dec. 2009, pp. 1–12.

[19] E. Nordström *et al.*, "Serval: An end-host stack for service-centric networking," in *Proc. USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Apr. 2012, pp. 85–98.

[20] D. Fisher, "A look behind the future internet architectures efforts," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 3, pp. 45–49, Jul. 2014.

[21] S. B. Akers, "Binary decision diagrams," *IEEE Trans. Comput.*, vol. -27, no. 6, pp. 509–516, Jun. 1978.

[22] Bryant, "Graph-based algorithms for Boolean function manipulation," *IEEE Trans. Comput.*, vol. C-35, no. 8, pp. 677–691, Aug. 1986.

[23] (2017). *Mobile Video Delivery With Hybrid ICN*. Cisco Systems. [Online]. Available: https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/ultra-services-platform/mwc17-hicn-video-wp.pdf

[24] T. Jepsen, M. Moshref, A. Carzaniga, N. Foster, and R. Soulé, "Packet subscriptions for programmable ASICs," in *Proc. 17th ACM Workshop Hot Topics Netw.*, Nov. 2018, pp. 176–183.

[25] T. Jepsen, A. Fattaholmanan, M. Moshref, N. Foster, A. Carzaniga, and R. Soulé, "Forwarding and routing with packet subscriptions," in *Proc. 16th Int. Conf. Emerg. Netw. Exp. Technol.*, Nov. 2020, pp. 282–294.

[26] A. Carzaniga, D. S. Rosenblum, and A. L. Wolf, "Achieving scalability and expressiveness in an internet-scale event notification service," in *Proc. 19th Annu. ACM Symp. Princ. Distrib. Comput. (PODC)*, Portland, OR, USA, 2000, pp. 219–227.

[27] T. Jepsen, M. Moshref, A. Carzaniga, N. Foster, and R. Soulé, "Life in the fast lane: A line-rate linear road," in *Proc. ACM SIGCOMM Symp. SDN Res. (SOSR)*, Mar. 2018, pp. 1–7.

[28] A. Gupta, R. Harrison, M. Canini, N. Feamster, J. Rexford, and W. Willinger, "Sonata: Query-driven streaming network telemetry," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, Aug. 2018, pp. 357–371.

[29] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proc. ACM SIGCOMM Conf. Data Commun. (SIGCOMM)*, 2008, pp. 1–12.

[30] M. Papalini, A. Carzaniga, K. Khazaei, and A. L. Wolf, "Scalable routing for tag-based information-centric networking," in *Proc. 1st Int. Conf. Inf.-Centric Netw. (INC)*, 2014, pp. 17–26.

[31] S. Smolka, S. Eliopoulos, N. Foster, and A. Guha, "A fast compiler for NetKAT," in *Proc. 20th ACM SIGPLAN Int. Conf. Funct. Program.*, Aug. 2015, pp. 328–341.

[32] X. Jin, J. Gossels, J. Rexford, and D. Walker, "Covisor: A compositional hypervisor for software-defined networks," in *Proc. USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Oakland, CA, USA, May 2015, pp. 87–101.

[33] W. Chan, R. Anderson, P. Beame, and D. Notkin, "Combining constraint solving and symbolic model checking for a class of systems with non-linear constraints," in *Int. Conf. Comput. Aided Verification (CAV)*, Jun. 1997, pp. 316–327.

[34] (2021). *Flatbuffers*. [Online]. Available: https://google.github.io/flatbuffers/

[35] J. Liu *et al.*, "P4V: Practical verification for programmable data planes," in *Proc. Conf. ACM Special Interest Group Data Commun. (SIGCOMM)*, Aug. 2018, pp. 490–503.

[36] (2017). *Inband Network Telemetry (INT)*. [Online]. Available: https://github.com/p4lang/p4factory/tree/master/apps/int

[37] N. Handigol, B. Heller, V. Jeyakumar, D. Mazières, and N. McKeown, "I know what your packet did last hop: Using packet histories to troubleshoot networks," in *Proc. USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Apr. 2014, pp. 71–85.

[38] (2019). *Barefoot Networks Deep Insight*. [Online]. Available: https://www.barefootnetworks.com/products/brief-deep-insight/

[39] (2019). *Broadcom BroadView Analytics*. [Online]. Available: https://www.broadcom.com/products/ethernet-connectivity/software/broadview-analytics

[40] N. Marz, *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. Sebastopol, CA, USA: O'Reilly Media, 2013. [Online]. Available: http://www.amazon.de/Big-Data-Principles-Practices-Scalable/dp/1617290343

[41] M. Zaharia *et al.*, "Apache spark: A unified engine for big data processing," *Commun. ACM*, vol. 59, no. 11, pp. 56–65, Oct. 2016.

[42] A. Lakshman and P. Malik, "Cassandra: A decentralized structured storage system," *ACM SIGOPS Operating Syst. Rev.*, vol. 44, no. 2, pp. 35–40, Apr. 2010.

[43] P. Jokela, A. Zahemszky, C. Esteve Rothenberg, S. Arianfar, and P. Nikander, "LIPSIN: Line speed publish/subscribe inter-networking," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, Aug. 2009, pp. 1–12.

[44] W. K. Chai *et al.*, "Cache 'less for more' in information-centric networks," in *Proc. Int. Conf. Res. Netw.* Berlin, Germany, 2012. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-30045-5_3

[45] M. Cooney. (Mar. 2018). *Cisco, Verizon Take Information-Centric Networking for a Real-World Spin*. Network World. [Online]. Available: https://www.networkworld.com/article/3264650/cisco-verizon-take-information-centric-networking-for-a-real-world-spin.html

[46] L. Muscariello, G. Carofiglio, J. Auge, and M. Papalini. (May 2020). *Hybrid Information-Centric Networking*. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-muscariello-intarea-hicn-04

[47] A. Arasu *et al.*, "Linear road: A stream data management benchmark," in *Proc. 30th Int. Conf. Very Large Data Bases*, Aug. 2004, pp. 1–12.

[48] Apache. (2019). *Apache Activemq*. [Online]. Available: http://activemq.apache.org/

[49] A. Carzaniga and A. L. Wolf, "Forwarding in a content-based network," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, Aug. 2003, pp. 163–174.

[50] (2017). *Benchmarking Kafka Performance Part 1: Write Throughput*. [Online]. Available: https://hackernoon.com/benchmarking-kafka-performance-part-1-write-throughput-7c7a76ab7db1

[51] (2019). *Vector Packet Processing (VPP)*. [Online]. Available: https://fd.io/

[52] (2019). *Siena Synthetic Benchmark Generator*. [Online]. Available: https://www.inf.usi.ch/carzaniga/cbn/forwarding/

[53] (2019). *Mininet*. [Online]. Available: http://mininet.org

[54] J. Leskovec and R. Sosič, "SNAP: A general-purpose network analysis and graph-mining library," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 1, p. 1, 2016.

[55] N. Foster *et al.*, "Frenetic: A network programming language," in *Proc. Int. Conf. Funct. Program. (ICFP)*, Sep. 2011, pp. 1–13.

[56] C. Monsanto *et al.*, "Composing software defined networks," in *Proc. USENIX Symp. Netw. Syst. Design Implement. (NSDI)*, Apr. 2013, pp. 1–13.

[57] R. Soulé *et al.*, "Merlin: A language for provisioning network resources," in *Proc. ACM Int. Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, Dec. 2014, pp. 213–226.

[58] C. J. Anderson *et al.*, "NetKAT: Semantic foundations for networks," in *Proc. Symp. Princ. Program. Lang. (POPL)*, Jan. 2014, pp. 1–14.

[59] S. Narayana *et al.*, "Language-directed hardware design for network performance monitoring," in *Proc. Conf. ACM Special Interest Group Data Commun.*, Aug. 2017, pp. 85–98.

[60] D. Perino, M. Varvello, L. Linguaglossa, R. Laufer, and R. Boislaigue, "Caesar: A content router for high-speed forwarding on content names," in *Proc. 10th ACM/IEEE Symp. Archit. Netw. Commun. Syst. (ANCS)*, Oct. 2014, pp. 137–148.

[61] Y. Wang *et al.*, "Fast name lookup for named data networking," in *Proc. IEEE 22nd Int. Symp. Quality Service (IWQoS)*, May 2014, pp. 198–207.

[62] H. Yuan and P. Crowley, "Reliably scalable name prefix lookup," in *Proc. ACM/IEEE Symp. Archit. Netw. Commun. Syst. (ANCS)*, May 2015, pp. 111–121.

Theo Jepsen received the B.A. degree in computer science from Cornell University, the M.S. degree from the University of Wisconsin–Madison, and the Ph.D. degree from the Università della Svizzera italiana (USI). He is currently a Post-Doctoral Researcher with Stanford University. His research interests include distributed systems, databases, and software-defined networking.

Ali Fattaholmanan received the B.S. degree in software engineering and the M.S. degree in information technology from the Sharif University of Technology, Tehran, Iran. He is currently pursuing the Ph.D. degree in informatics with the Università della Svizzera italiana (USI), Lugano, Switzerland. His research interests include programmable networks devices and congestion control schemes for cloud networks.

Masoud Moshref received the Ph.D. degree from the University of Southern California. He is currently a Software Engineer with Google. Before joining Google, he was a Software Engineer with Barefoot Networks. His research interests include congestion control, networks monitoring, and programmable networking devices.

Nate Foster received the B.A. degree in computer science from the Williams College, the M.Phil. degree in history and philosophy of science from the University of Cambridge, and the Ph.D. degree in computer and information science from the University of Pennsylvania. He is currently a Professor of computer science with Cornell University and a Platform Architect at Intel. His research interests include programming language design, semantics, and implementation; and software-defined networks.

Antonio Carzaniga received the Ph.D. degree from the Politecnico di Milano, Italy. He is currently a Full Professor with the Università della Svizzera italiana (USI), Switzerland. Prior to joining USI, he was an Assistant Research Professor with the University of Colorado Boulder.

Robert Soulé received the B.A. degree from Brown University and the Ph.D. degree from NYU. After his Ph.D. degree, he was a Post-Doctoral Researcher with Cornell University. He is an Assistant Professor with Yale University and a Research Scientist with Barefoot Networks, an Intel Company. Prior to joining Yale, he was an Associate Professor with the Università della Svizzera italiana, Lugano, Switzerland.