Domain specific run time optimization for software data planes

Original
Domain specific run time optimization for software data planes / Miano, Sebastiano; Sanaee, Alireza; Risso, FULVIO GIOVANNI OTTAVIO; Rétvári, Gábor; Antichi, Gianni. - STAMPA. - (2022), pp. 1148-1164. (Intervento presentato al convegno 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems tenutosi a Lausanne (Switzerland) nel February 2022) [10.1145/3503222.3507769].

Availability:
This version is available at: 11583/2972687 since: 2022-10-29T10:44:39Z

Publisher:
IEEE

Published
DOI:10.1145/3503222.3507769

Terms of use:
This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
ACM postprint/Author's Accepted Manuscript, con Copyr. autore

(Article begins on next page)
Domain Specific Run Time Optimization for Software Data Planes

Sebastiano Miano  
Queen Mary University London  
London, United Kingdom  
s.miano@qmul.ac.uk

Aliereza Sanaee  
Queen Mary University London  
London, United Kingdom  
a.sanaee@qmul.ac.uk

Fulvio Risso  
Politecnico di Torino  
Turin, Italy  
ffulvio.risso@polito.it

Gábor Rétvári  
Budapest University of Technology and Economics & Ericsson Research  
Budapest, Hungary  
g.antichi@qmul.ac.uk

ABSTRACT
State-of-the-art approaches to design, develop and optimize software packet-processing programs are based on static compilation: the compiler’s input is a description of the forwarding plane semantics and the output is a binary that can accommodate any control plane configuration or input traffic.

In this paper, we demonstrate that tracking control plane actions and packet-level traffic dynamics at run time opens up new opportunities for code specialization. We present Morpheus, a system working alongside static compilers that continuously optimizes the targeted networking code. We introduce a number of new techniques, from static code analysis to adaptive code instrumentation, and we implement a toolbox of domain specific optimizations that are not restricted to a specific data plane framework or programming language. We apply Morpheus to several eBPF and DPDK programs including Katran, Facebook’s production-grade load balancer. We compare Morpheus against state-of-the-art optimization frameworks and show that it can bring up to 2x throughput improvement, while halving the 99th percentile latency.

CCS CONCEPTS
• Networks → Middle boxes / network appliances; Data center networks; End nodes; • Software and its engineering → Dynamic compilers; Just-in-time compilers.

KEYWORDS
Data Plane Compilation, LLVM, eBPF, XDP, DPDK

1 INTRODUCTION
Software Data Planes, packet processing programs implemented on commodity servers, are widely adopted in real deployments [9, 37, 44, 53, 75, 86, 93, 94]. Since data plane programs tend to be performance-critical, the code is usually transformed through a sequence of offline optimization steps (e.g., inlining, loop unrolling, branch elimination, or vectorization [6, 59]) during the compilation process [39, 57]. These are mainly static transformations, independent of the actual input the code will process in operation, as this is unknown until then [10, 28]. Thus, the resulting code is generic, as it contains logic for protocols and features that may never be triggered in a deployment, performs costly memory loads to access values that are only known at run time, and takes difficult-to-predict branches conditioned on variable data.

Dynamic compilation, in contrast, enables program optimization based on invariant data computed at run time and produces code that is specialized to the input the program is processing [7, 28, 38]. The idea is to continuously collect run time data about program execution and then re-compile it to improve performance. This is a well-known practice adopted by generic programming languages (e.g., Java [28], JavaScript [38], and C/C++ [7]) and often produces orders of magnitude more efficient code as shown for data-caching services [74], data mining [21] and databases [55, 97]. Unfortunately, this is not the case for packet-processing programs [6, 29, 68], since their performance critically depends on highly dynamic domain-specific knowledge, such as traffic patterns, match-action table content, and network configuration (§2). Obtaining and tracking this knowledge efficiently is extremely challenging: lightweight online tracing tools (e.g., Linux perf [25]) are restricted to CPU performance counters, whereas capturing all domain-specific information requires tracking packet-level and instruction-level logs which is prohibitively costly. As an example, GCC FDO instrumentation, when applied in this context, may easily incur ~900% mean overhead [54]. Therefore, existing solutions tailored for the networking domain (Table 2) resort to offline profiling, which requires operators to collect representative samples of data-plane configuration and match-action tables from production deployments and still completely miss out on dynamic traffic-level insights. The main challenge we tackle in this paper is unsupervised dynamic compilation for network code, which captures just enough domain-specific knowledge to enable efficient dynamic performance optimization, but inexpensive enough to be run online, inside the data plane pipeline.
Table 1: A comparison of some popular dynamic optimization frameworks and Morpheus.

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain specific</th>
<th>Unsupervised adaptation to control plane actions</th>
<th>Unsupervised adaptation to data plane traffic</th>
<th>Data plane agnostic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bolt [74]</td>
<td>✗</td>
<td>-</td>
<td>✗</td>
<td>✗</td>
<td>Offline profile-guided optimizer for generic software code.</td>
</tr>
<tr>
<td>eSwitch [68]</td>
<td>✓</td>
<td>-</td>
<td>✗</td>
<td>✗</td>
<td>Policy-driven optimizer for DPDK-based OpenFlow software switches.</td>
</tr>
<tr>
<td>PacketMill [34]</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>Packet metadata management optimizer for DPDK-based software data planes.</td>
</tr>
<tr>
<td>NFReducer [29]</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>Policy-driven optimizer for network function virtualization.</td>
</tr>
<tr>
<td>Morpheus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Run time compiler and optimizer framework for arbitrary networking code.</td>
</tr>
</tbody>
</table>

We present Morpheus, a system to optimize network code at run time using *domain-specific dynamic optimization techniques*. Morpheus operates in unsupervised mode: it does not require any *a priori* knowledge about control plane configuration or data plane traffic patterns. We discuss its design challenges (§3), such as automatically tracking highly variable input (e.g., inbound traffic) that may change tens, or hundreds of millions times per second. We show that the required profiling and tracing facilities, if implemented carelessly, can easily nullify the performance benefit of code specialization. We introduce several novel techniques; we leverage *static code analysis* to build an understanding of the program *offline* and propose a low-overhead *adaptive instrumentation* mechanism to minimize the amount of data collected *online*. Then, we invoke several dynamic *optimization passes* (e.g., dead code elimination, data-structure specialization, just-in-time compilation, and branch injection) to specialize the code against control plane actions and data plane traffic patterns. Finally, we protect the consistency of the specialized code against changes to input that is considered invariant by injecting *guards* (§4).

Our implementation exploits the LLVM JIT compiler toolchain to apply the above ideas at the LLVM Intermediate Representation (IR) level in a generic fashion and separates data plane specific code to several backend plugins to minimize the effort in porting Morpheus to a new architecture (§5). The code currently contains an eBPF and a DPDK/C plugin. We apply Morpheus to a number of packet processing programs, including the production-grade L4 load balancer Katran from Facebook, using synthetic and real-world traffic traces (§6). Our results show that Morpheus can improve the performance of the unoptimized (statically compiled) eBPF application up to 94%, while reducing packet-processing latency by up to 123% at the 99th percentile. Applying Morpheus to a DPDK application, we increase performance by up to 469%. Finally, we measured Morpheus against state-of-the-art network code optimization frameworks such as ESwitch [68] and PacketMill [34]: we show that Morpheus boosts the throughput by up to 80% and 294%, respectively, compared to existing work.

**Contributions.** In this paper, we:
- demonstrate that tracking packet-level dynamics opens up new opportunities for network code specialization;
- design and implement Morpheus, a system working with standard compilers to optimize network code at run time;
- extensively evaluated Morpheus by applying it to two different I/O technologies (i.e., DPDK and eBPF), and a number of programs including production-grade software;
- share the code in open source to foster reproducibility [63].

## 2 THE CASE FOR DOMAIN-SPECIFIC OPTIMIZATIONS

State-of-the-art profile-guided optimization tools (PGO), such as Google’s AutoFDO [21, 46] and Facebook’s Bolt [74], can dynamically rewrite the targeted code using execution profiles recorded offline; e.g., by simplifying load instructions or reordering basic blocks to speed up the most frequently executed code paths. Fig.1a shows the single-core throughput obtained when applying AutoFDO and Bolt combined (PGO) to a sample DPDK firewall application [31], which performs basic L2/L3/L4 packet processing followed by a lookup into an access control list (ACL), over a stream of 64-byte packets at 40Gb line rate (see §6 for the details of the configuration). In line with the expectations [21, 54, 74], we managed to improve the performance of the targeted code by a mere 4.2%.

The behavior of packet-processing code can, however, be deeply influenced by specific metrics (e.g., match-action table access patterns, table sizes and content) that cannot be tracked with generic profiling mechanisms (i.e., Linux perf) used by standard PGO tools. Lacking such *domain-specific* insight, meaningful only in the packet processing context, generic purpose PGO tools cannot be fully exploited for dynamically optimizing network code [6, 29, 34, 68, 79].

To understand the potential of *domain-specific* optimizations, we present a series of preliminary benchmarks using real network code. We consider two applications: the DPDK sample *firewall* discussed above and Katran [44], Facebook’s open-source L4 eBPF/XDP *load balancer*.

**The promise of policy-driven optimizations.** Most data-plane programs are developed as a single monolithic block containing various features that might be activated depending on the specific network configuration in use at any instance of time. For example, many large-scale cloud deployments still run on pure IPv4 and so the hypervisor switches would never have to process IPv6 packets [48] or adopt a single virtualization technology (VLAN/VxLAN/GRE/Genève/GTP) and so switches would never see other encapsulations in operation [53, 73]. This implies that, depending on dynamic input that is unknown at compile time, a huge body of...
Removing unused code based on run time configuration can have a profound effect on software performance. To show this, we configured our firewall as a TCP signature-based Intrusion Detection System (IDS), with only TCP wildcard rules generated with Class-Bench [90]. This opens up an opportunity for optimization: all non-TCP packets can bypass the ACL table, avoiding a wasteful lookup. Fig. 1b shows the run time benefit of this optimization (under the Run time configuration bar) for a synthetic input traffic trace where only about 10% of the packets are UDP. Although around 90% of the traffic still has to undergo an ACL lookup, just avoiding this costly operation for a small percentage of traffic increases performance with about 4.7%, without changing the semantic in any way. In many practical scenarios, like DDoS blocking, security groups [45, 71] or whitelist-based access control, most firewall rules are fully-specified; for instance, in the official Stanford ruleset [52] on average ~45% of the rules are purely exact-matching. This opens up another dynamic optimization opportunity: add in front of the ACL an exact-matching lookup table to sidestep the costly wildcard lookup. The result in Fig. 1b (under the Table specialization bar) shows a further ~8% performance improvement with this simple modification.

A similar effect is visible with the load-balancer (Fig. 1c): configuring Katran as an HTTP load balancer [14, 70] allows to dynamically remove all the branches and code unrelated to IPv4/TCPProcessing, which reduces the number of instructions by ~58% (as reported by the Linux perf tool), yielding ~17.1% decrease in the number of L1 instruction cache-load misses. Better cache locality then translates into ~12% performance improvement (from 4.09 Mpps to 4.69 Mpps).

**Takeaway #1:** Specializing networking code for slowly changing input, like flow-rules, ACLs and control plane policies, substantially improves the performance of software data planes.

**The need for tracking packet-level dynamics.** The potential to optimize code for specific network configurations has been explored in prior work, for OpenFlow [68], P4 software [79, 91] and hardware targets [5], network functions [29], and programmable switches [34] (see Table 1). In order to maximize performance, however, we need to go beyond specializing the code for relatively stable run time configuration and apply optimizations at the packet level.

Consider the DPDK firewall application. We installed 1000 wild-card rules and generated highly skewed traffic, so that from the thousand active unique 5-tuple flows only 5% accounts for 95% of the traffic. This opens up the opportunity to inline the match-action logic for the recurring rules. As the results show (Fig. 1b, under the Fast Path bar), we obtain ~42% performance improvement with this simple traffic-dependent optimization. With the eBPF load balancer the effect is also visible: configuring 10 Virtual IPs (VIP) (both TCP and UDP), each with hundred different back-end servers, a similarly skewed input traffic trace presents the same opportunity to inline code, yielding ~24% performance edge (Fig. 1c).

**Takeaway #2:** For maximum performance, networking code must be specialized with respect to inbound traffic patterns, despite the potentially daunting packet-level dynamics.

**3 CHALLENGES**

Static compilation performs optimizations that depend only on compile-time constants: it does not optimize variables whose value is invariant during the execution of the program but remain unknown until then. Dynamic compilation, in contrast, enables specializing the code with respect to invariant run time data [7]. This opens up a broad toolbox of optimization opportunities, to propagate, fold and inline constants, remove branches and eliminate code never triggered in operation, or even to completely sidestep costly match-action table processing. The unsupervised optimization of networking code, however, presents a number of unique challenges:

**Challenge #1:** Low-overhead run time instrumentation. Unsupervised dynamic optimization rests on the assumption that program variables remaining invariant for an extended period of time are promptly detected. This is the job of profiling tools. Although low-overhead solutions exist [25, 26, 40, 58], they track high-level code behavior information such as cache events, branch misses or memory accesses which is not enough for packet-processing code (§2). Less lightweight and more accurate tools [8, 19, 27], instead, are not practical to be used at data-plane time scales: recording at run time instruction-level logs for code that processes potentially tens of millions of packets per second can introduce an overhead that makes the subsequent optimization pointless. For example, GCC FDO instrumentation can easily incur ~900% mean overhead [54]. We tackle this challenge in Morpheus by using static code analysis to understand the structure of the program offline (§4.1) and leveraging an adaptive instrumentation mechanism to minimize the amount of data that is collected online (§4.2).

**Challenge #2:** Dynamic code generation. Once run time profiling information is available, the dynamic compiler applies domain-specific optimizations to specialize the code for that profile. Here, code generation must integrate seamlessly into the compiler toolchain, to avoid interference with the built-in optimizations. Furthermore, a toolbox of domain-specific optimization passes must be identified, which, when applied to networking code, promise significant speedup (§4.3).

**Challenge #3:** Consistency. The dynamically optimized data plane is contingent on the assumption that the data considered
We designed Morpheus with an ambitious goal: to build a portable 
We tackle this challenge by injecting 
At each invocation, the compiler performs an extensive offline code 
Finally, the IP address of the backend associated with 
of the possible inputs it may receive during run time. Networking 
4.1 Code Analysis

Listing 1: Simplified Katran main loop

```c
1 int process_packet (packet pkt) {
2    u32 backend_idx;
3    parse_l3_headers(pkt);
4    parse_l4_headers(pkt);
5    vip.vip = pkt.dstIP;
6    vip.port = pkt.dstPort;
7    vip.proto = pkt.proto;
8    vip_info = vip_map.lookup(vip);
9    if (vip_info->flags & F_QUIC_VIP) {
10       backend_idx = assign_to_backend(pkt);
11       goto send;
12    }
13    if (!backend_idx) {
14       backend_idx = conn_table.lookup(pkt);
15       conn_table.update(pkt, backend_idx);
16    }
17    send:
18    backend = backend_pool.lookup(backend_idx);
19    encapsulate_pkt(backend->ip);
20    return XDP_TX;
21 }
```
Table 2: Dynamic optimizations in Morpheus. Applicability of each optimization depends on the map size, access profile (RO/RW), and availability of instrumentation information. Note that optimizations marked as “traffic-dependent” can also be applied, at least partially, without packet-level information (e.g., small RO maps can always just-in-time compiled). For full efficiency, these passes rely on timely instrumentation information (e.g., to JIT heavy hitters from a large map as a fast-path).

4.2 Instrumentation

In the second pass, Morpheus profiles the dynamics of the input traffic by generating heatmaps of the maps’ access patterns, so that the collected statistics can then be used to drive the subsequent optimization passes. Specifically, Morpheus uses a sketch to keep track of map accesses, by storing instrumentation data in a LRU (least-recently-used) cache alongside each map and adapting the sampling rate along several dimensions to control the run time cost of profiling. The dimensions of adaptation are as follows. (1) Size: small maps are unconditionally inlined into the code and hence instrumentation is disabled for these maps. (2) Dynamics: Morpheus does not record each map access, but rather it samples just enough information to reliably detect heavy hitters [33]. (3) Locality: instrumentation caches are per-CPU and hence track the local traffic conditions at each execution thread separately, i.e., specific to the RSS context. This improves per-core heavy hitter detection in presence of highly asymmetric traffic. (4) Scope: after identifying heavy hitters in the CPU context, local instrumentation caches are run together to identify global heavy hitters. (5) Context: if a map is accessed from multiple call sites then each one is instrumented separately, so that profiling information is specific to the calling context. (6) Application-specific insight: the operator can manually disable instrumentation for a map if it is clear from operational context that access patterns prohibit any traffic-dependent optimization (see Table 2). Traffic-independent optimizations are still applied by Morpheus in such cases.

**Running example.** Consider the vip_map in our sample program, identified as an RO map in the first pass. In addition, suppose that there are hundreds of VIPs associated with TCP services stored in the vip_map and only a single one is running QUIC, but the QUIC service receives the vast majority of run time hits. Then, instrumentation will identify the QUIC VIP as a heavy hitter and Morpheus will seize the opportunity to specialize the subsequent QUIC call-path explicitly. Note that this comes without direct traffic monitoring, only using indirect traffic-specific instrumentation information.

4.3 Optimization Passes

The third step of the compilation pipeline is where all online code transformations are applied. Before deploying any code transformation, Morpheus has to protect the consistency of the new specialized code against changes to the invariants the optimizations depend on. To do this, Morpheus uses guards, a standard mechanism used by dynamic compilers to guarantee code consistency by injecting simple run time version checks at specific points in the code [43]. When the control flow reaches a guard, it atomically checks if the version of the guard is the same as the version of the optimized code; if yes, execution jumps to the optimized version, otherwise it falls back to the original code (“deoptimization”). Below, we describe all the various run time optimizations currently applied by Morpheus; see Table 2 for a summary.

4.3.1 Just-in-time compilation (JIT). Empirical evidence (see §2) suggests that table lookup is a particularly taxing operation for software data planes. This is because certain match-action table types (e.g., LPM or wildcard), that are relatively simple in hardware, are notoriously expensive to implement in software [35]. Therefore, Morpheus specializes tables at run time with respect to their content and dynamic access patterns, as learned in the instrumentation pass. Specifically, empty maps are completely removed, small maps are unconditionally just-in-time (JIT) compiled into equivalent code, and larger maps are preceded by a similar JIT compiled fast-path cache, which is in charge of handling the heavy hitters. Note that the consistency of the the fast-path cache must be carefully protected against potential changes made to the specialized map entries; Morpheus places guards into the code to ensure this (see later).

**Running example.** Consider again Listing 1 and suppose that there are only two VIPs configured in the vip_map. Being an exact-matching hash it is trivial to compile the vip_map into an “if-then-else” statement, representing each distinct map key as a separate branch. To do so, Morpheus uses the insights from the code analysis phase to discover that relevant fields in the lookup are the destination address (pkt. dstIP), port (pkt. dstPort) and the IP protocol (pkt. proto). Then, for each entry in the map, it builds a separate “if” conditional to compare the entry’s fields against the relevant packet header fields and chains these with “else” blocks. Since the instrumentation and the just-in-time compiled map are specific to unique combinations of destination address/port/protocol, the lookup semantics is correctly preserved even for longest prefix matching (LPM) caches and wildcard lookup.

4.3.2 Constant propagation. Specializing a table does not only benefit the performance of the lookup process but has far reaching consequences for the rest of the code. This is because a specialized table contains all the constants (keys and values) inlined, which
makes it possible to propagate these constants to the surrounding code so to inline memory accesses. In Morpheus, constant propagation opportunistically extends to larger maps that cannot be wholly just-in-time compiled: if a certain table field is found to be constant across all entries, then it is also inlined into the surrounding code. This optimization is thereby two-faceted: it can be used to specialize the code with respect to the inbound traffic (traffic-dependent, former case) but can also be applied without packet-level information (traffic-independent, the latter case). Morpheus does not implement constant propagation itself; rather, it relies on the underlying compiler toolchain to perform this pass.

Running example. Suppose there are only two backends in the backend_pool. Here, the map lookup (line 24) is rewritten into an "if-then-else" statement, with two branches for each backend. Correspondingly, in each branch the value of the backend variable is constant, which allows to save the costly memory dereference backend->ip (line 25) by inlining the backend IP address right into the specialized code.

4.3.3 Dead code elimination. Depending on the specific configuration, a large portion of code may sit unused in memory at any point in time. Such "dead code" can be found using a combination of static code analysis and the instrumentation information obtained from the previous pass. Upon detection, Morpheus removes all dead code on the optimized code path. As previously, this operation is outsourced to the compiler.

Running example. Consider the vip_map lookup site (line 10) and suppose that there are no QUIC services configured. As a consequence, the vip_info->flags is identical across all the entries in the vip_map and the constant propagation pass inlines this constant into the subsequent conditional (line 10). Thus, the condition vip_info->flags & F_QUIC_VIP always evaluates to false and the subsequent branch can be safely removed.

4.3.4 Data Structure Specialization. Morpheus adapts the layout, size and lookup algorithm of a table against its content at run time. For example, if all entries share the same prefix length in an LPM map, then a much faster exact-matching cache [68] can be used. This is done by first associating a backend-specific cost function with each applicable representation (this can be automatically inferred using static analysis and symbolic execution [74, 76]).

expected cost of each candidate, and finally implement the table that minimizes the cost.

4.3.5 Branch Injection. This pass applies to the cases when certain fields take only few possible values in a table, which makes it possible to eliminate subsequent code that handles the rest of the values. This optimization was used in §2 to sidestep the ACL lookup for UDP packets in the firewall use case: if we observe that the "IP protocol" field can have only a single value in the ACL (e.g., TCP), then we can inject a conditional statement before the ACL lookup to check if the IP protocol field in a packet is TCP, use symbolic execution to track the use of this value throughout the resultant branch, and invoke dead code elimination to remove the useless ACL lookup on the non-TCP "else" branch.

4.3.6 Guard elision. As discussed before, Morpheus uses guards to protect the consistency of the optimized code. Since each packet may need to pass multiple checks while traversing the datapath, guards may introduce non trivial run time overhead [92]. To mitigate this, Morpheus heuristically eliminates as many guards as possible; this is achieved by using different schemes depending if changes to the code are made from the control plane or from the data-plane itself, as in the case when a program implements a stateful network function.

Handling control plane updates. Theoretically, each table should be protected by a guard when the contents are modified from the control plane. This would require packets to perform one costly guard check for each table. To reduce this overhead, Morpheus collapses all table-specific guards against control plane updates into a single program-level guard, injected at the program entry point. Once an RO map gets updated by the control plane, the program-level guard directs all incoming packets to the original (unoptimized) datapath until the next compilation cycle kicks in to re-optimize the code with respect to the new table content.

Handling updates within the data plane. The optimized datapath must be protected from data-plane updates as well, which requires an explicit guard at all access sites for RW maps. If the guard tests valid then a query is made into the just-in-time compiled fast-path map cache and, on cache hit, the result is used in the subsequent code. Once a modification is made to the map from inside the data plane, the guard is invalidated and map lookup falls back to the original map.
Fig. 3 presents a breakdown of the strategies Morpheus uses to protect the consistency of optimized code. For RW maps (Fig. 3a), first an instrumentation cache is inlined at the access sites, followed by a guard that protects the just-in-time compiled fast-path against data-plane updates. Note that the constant propagation and dead code elimination passes are suppressed, since these passes may modify the code 

Running example. Once static code analysis confirms that the vip_map and backend_pool maps are RO, Morpheus opportunistically eliminates the corresponding guards at the call site. This then implies that, as long as the VIPs and the backend pool are invariant, the optimized code elides the guard. Since the conn_table map is RW, it is protected with a specific guard at the call site (line 17). Thus, the specialized map is used only as long as the connection tracking module’s state remains constant; once a new flow is introduced into conn_table (line 20) the specialized code is immediately invalidated by bumping the data-plane version. This does not invalidate all optimizations: as long as the rest of the (RO) maps are not updated by the control plane, the program-level guard remains valid and the corresponding RO map specializations still apply.

4.4 Update

Upon invocation, Morpheus executes the above passes to create the optimized datapath and uses the native compiler toolchain to transform the optimized code to target native code. Meanwhile, control plane updates are temporarily queued without being processed. This allows the “old” code to process packets without any disruption while the optimization takes place. Once compilation is finished, the optimized code is injected into the data path, the program-level guard is updated [41] and the outstanding table updates are executed.

5 IMPLEMENTATION

Morpheus is implemented in about 6000 lines of C++ code and it is openly available at https://github.com/Morpheus-compiler/Morpheus, with the artifacts archived on Zenodo [63]. The code is separated into a data plane independent portable core, containing the compiler passes, and technology-specific plugins to interact with the underlying technology (i.e., eBPF, DPDK).

The Morpheus core extends the LLVM [57] compiler toolchain (v10.0.1) for code manipulation and run time code generation. We opted to implement Morpheus at the intermediate representation (IR) level as it allows to reason about the running code using a relatively high-level language framework without compromising on code generation time. Moreover, this also makes the Morpheus core portable across different data plane frameworks and programming languages [85].

The data plane plugins are abstracted via a backend API. This API exports a set of functions for the core to identify match-action table access sites based on data-plane specific call signatures; compute cost functions for data structure specialization; rewrite data plane dependent code using templates; and provide an interface to inspect guards. Additionally, the backend can channel instrumentation data from the data plane to the compiler core, implement the data plane dependent parts of the pipeline update mechanism, and provide a mechanism for the Morpheus core to intercept, inspect, and queue any update made by the control plane. The latter allows the compilation pipeline to be triggered when Morpheus intercepts a control plane event, e.g., an update to a table. Currently, only eBPF (fully) and DPDK (partially) are supported, but the architecture is generic enough to be extended to essentially any I/O framework, like netmap [81] or AF_XDP [2].

5.1 The eBPF Plugin

Morpheus leverages the Polycube [65] framework as an eBPF backend to manage chains of in-kernel packet processing programs. Polycube readily delivers almost all the needed components for an eBPF backend. We added a mechanism for updating the data plane program on-the-fly and defined templates to inject guards. We discuss these components next.

Pipeline update. Once the optimized program is built, Morpheus calls the eBPF LLVM backend to generate the final eBPF native code, loads the new program into the kernel using the bpf() system call, and directs execution to the new code. In Polycube, a generic data plane program is usually realized as a chain of small eBPF programs connected via the eBPF tail-call mechanism, using a BPF_PROG_ARRAY map to get the address of the entry point of the next eBPF program to execute. Thus, injecting a new version of an eBPF program boils down to atomically update the BPF_PROG_ARRAY map entry pointing to it with the address of the new code.

Guards. Morpheus relies on guards to protect the specialized code against map updates. The program-level guard is implemented as a simple run time version check [41]. For stateful processing, Morpheus installs a guard at each map lookup site and injects a guard update pre-handler at the instruction address corresponding to the map update eBPF function (map_update_e1em). This handler will then safely invalidate the guard before executing the map update.

5.2 The DPDK Plugin

Morpheus leverages FastClick [13], a framework to manage packet-processing applications based on DPDK. FastClick makes implementing most components of the backend API trivial; below we report only on pipeline updates and guards.

Pipeline update. A FastClick program is assembled from primitive network functions, called elements, connected into a dataflow graph. Every FastClick element holds a pointer to the next element along the processing chain. To switch between different element implementations at run time, Morpheus adds a level of indirection to the FastClick pipeline: every time an element would pass execution to the next one, the corresponding function call is conveyed through a trampoline, which stores the real address of the next element to be called. Then, atomic pipeline update simplifies into
rewriting the corresponding trampoline to the address of the new code. In contrast to eBPF, which explicitly externalizes into separate maps all program data intended to survive a single packet’s context, a FastClick element can hold non-trivial internal state, which would need to be tediously copied into the new element. As a workaround, our DPDK plugin disables dynamic optimizations for stateful FastClick elements.

Guards. Since stateful FastClick elements are never optimized in Morpheus and RO elements always elide the guard, our DPDK plugin currently does not implement guards, except a program-level version check at the entry point.

6 EVALUATION

Our testbed includes two servers connected back-to-back with a dual-port Intel X710 40Gbps NIC. The first, a 2x10-core Intel Xeon Silver 4114 CPU @2.20GHz with support for Intel’s Data Direct I/O (DDIO) [1] and 37.5 MB of L3 cache, runs the various applications under consideration. The second, a 2x10 Intel Xeon Silver 4114 CPU @2.20GHz with 13.75MB of L3 cache, is used as packet generator. Both servers are installed with Ubuntu 20.04.2, with the former running kernel 5.10.9 and the latter kernel 4.15.0-112. We also configured the NIC Receive-Side Scaling (RSS) to redirect all flows to a single receive queue, forcing the applications to be executed on a single CPU core, while Morpheus was pinned to another CPU core on the device-under-test (DUT).

In our tests, we used pktgen [30] with DPDK v20.11.0 to generate traffic and report the throughput results, and the DPDK burst replay tool [80] to replay the different packet traces. Unless otherwise stated, we report the average single-core throughput across five different runs of each benchmark, measured at the minimum packet size (64-bytes). For latency tests, we used Moongen [32] to estimate the round-trip-time of a packet from the generator to the DUT and back. Finally, we used perf v5.10 to characterize the micro-architectural metrics of the DUT (e.g., cache misses, cycles, number of instructions).

In order to benchmark Morpheus on real applications, we chose four eBPF/XDP-based packet processing programs from the open-source eBPF/XDP reference network function virtualization framework Polycube [67], plus Facebook’s Katran load-balancer used earlier as a running example [44].

The L2 switch, the Router and the NAT applications were taken from Polycube [67]. The L2 switch use case is a functional Ethernet switch supporting 802.1Q VLAN and STP, with STP and flooding delegated to the control plane while learning and forwarding implemented entirely in eBPF, using an exact-matching MAC table supporting up to 4K entries. The Router use case implements a standard IP router, with RFC-1812 header checks, next-hop processing and checksum rewriting, configured with an LPM table taken from the Stanford routing tables [52]. The NAT is an eBPF re-implementation of the corresponding Linux Netfilter application, configured with a single two-way SNAT/masquerading rule: the source IP of every packet is replaced with the IP of the outgoing NAT port and a separate L4 source port is allocated for each new flow. BPF-iptables is an eBPF/XDP clone [64] of the well-known Linux iptables framework, configured with 5-tuple rules generated by Classbench [90]. We used the Classbench trace generator [89] to generate packets matching the created rule set using a Pareto cumulative density function to control the locality of reference. We used the same default parameters suggested by the ClassBench paper [90] to generate traces of varying locality, in particular the no-locality trace uses $\alpha = 1, \beta = 0$ as Pareto parameters, the low locality uses $\alpha = 1, \beta = 0.0001$, and the high locality uses $\alpha = 1, \beta = 1$. Finally, Katran [44] was configured as a web-fronted, with 10 TCP services/VIPs and 100 backend servers for each VIP.

For each benchmark, we generated 3 traffic traces with varying locality, to demonstrate the ability of Morpheus to track packet-level dynamics and optimize the programs accordingly. In particular, we created a high-locality traffic trace, where few flows account for most of the traffic, a no-locality trace with flows generated at random by a uniform distribution and a low-locality trace that sits in the middle between the two previous cases.

6.1 Benefits of Optimizations

We first show the impact of Morpheus on the mentioned programs, when attached to the XDP hook of the ingress interface.

Morpheus improves packet-processing throughput. In Fig. 4, we show the impact of Morpheus under different traffic conditions. Throughput is defined as the maximum packet-rate sustained by
Figure 6: 99th percentile (P99) latency with Morpheus. The graph shows both the latency for the optimized and non-optimized code paths, under small load (10pps) and heavy load (highest rate without packet drop).

a program without experiencing packet loss. When a small subset of flows sends the majority of traffic (high-locality), Morpheus consistently provides more than 50% throughput improvement over the baseline, with a 2x speed-up for the Router. This is because it can track heavy flows and optimize the code accordingly. To confirm the benefit of packet-level optimizations in Morpheus, we compared it to a faithful eBPF/XDP re-implementation of ESswitch, a dynamic compiler that does not consider traffic dynamics [68]. The results (Fig 4) show that Morpheus delivers 5–10x the improvement compared to ESswitch for high-locality traces, while it falls back to ESswitch for uniform traffic.

Morpheus benefits at the micro-architectural scale. In Fig. 5, we show that Morpheus reduces the last-level CPU cache misses by up to 96% and halves the instructions and branches executed per packet. At low or no traffic locality, the effects of packet-level optimizations diminish, but Morpheus can still bring considerable performance improvement: we see ~30% margin for BPF-iptables even for the no-locality trace. This is because the optimization passes in Morpheus are carefully selected to be applicable independently from packet-level dynamics (see Table 2).

Morpheus reduces packet-processing latency. In Fig. 6, we compared the 99th percentile baseline latency for each application against the one obtained with Morpheus, both in a best-case scenario when all packets travel through the optimized code path (e.g., the right branch in Fig. 3a), and a worst-case scenario with all packets falling back to the default branch instead of taking the fast-patch cache for each map (the left branch in Fig. 3a). The left panel in Fig. 6 shows the latency measured at low packet rate (10pps) so to avoid queuing effects [16], whereas the right panel shows latency under the maximum sustained load without packet drops [17]. First, we observe that Morpheus never increases latency, despite the considerable additional logic it injects dynamically into the code (guards, instrumentation; see below); in fact, it generally reduces it even in the worst case scenario. Notably, it reduces Katran’s packet-processing latency by about 123%.

6.2 What is the cost of code instrumentation?

Clearly, the price for performance improvements is the additional logic, most prominently, instrumentation, injected by Morpheus into the fast packet-processing path. To understand this price, we compared our adaptive instrumentation scheme (§4.2) against a naive approach where all map lookups are explicitly recorded. Fig. 7 shows that instrumentation involves visible overhead: the instrumented code performs worse than the baseline. The naive approach imposes a hefty 14–23% overhead, but adaptive instrumentation reduces this to just 0.9%–9%. Most importantly, this reduction does not come at a prohibitive cost: adaptive instrumentation provides enough insight to Morpheus to make up for the performance penalty imposed by it and still attain a considerable throughput improvement on top (see the green stacked barplots). In contrast, the performance tax of naive instrumentation may very well nullify optimization benefits, even despite full visibility into run time dynamics (e.g., for the L2 switch or Katran).

We also studied the impact of packet sampling rate on instrumentation. Indeed, Morpheus collects information on packet-level dynamics only on a subset of input traffic in order to minimize the overhead. Fig. 8 highlights that Morpheus can strike a balance between overhead and efficiency by adapting the sampling rate. At a low sampling rate (e.g., recording every 100th packet) Morpheus does not have enough visibility into dynamics, which renders traffic-dependent optimizations less effective (but the traffic-invariant optimizations still apply). Higher sampling rates provide better visibility but also impose higher overhead. At the extreme (BPF-iptables, 100% instrumentation rate), optimization is just enough to offset the price of instrumentation. In conclusion, we found that setting the sampling rate at 5%–25% represents the best compromise.
### 6.3 How fast is the compilation?

In Table 3, we indicate with $t_1$ the time to analyze, instrument and optimize the LLVM IR code, and with $t_2$ the time to generate the final eBPF code, starting from the LLVM IR. Note that $t_1$ is highly dependent on table size: the bigger the table, the more time needed to read and analyze it. We show the results for high-locality and no-locality traffic. The former is the best case since Morpheus needs to track fewer flows, thus requiring lighter instrumentation tables that are faster to analyze. The latter is the worst case. Generally, table read time (i.e., $t_1$) dominates over compilation time, consistently staying below 100ms and reaching only for Katran in the worst-case scenario almost 600ms. This is because Katran uses huge static maps with tens of thousands of entries to implement consistent hashing. Recent advances in the Linux kernel allow to read maps in batches, which would cut down this time by as much as 80% [95], reducing recompilation time for Katran below 100ms. Furthermore, the time needed to inject the optimized datapath into the kernel depends on the complexity of the program, since all eBPF code must pass the in-kernel verifier for a safety check before being activated. This also ensures that a mistaken Morpheus optimization pass will never break the data plane. In our tests, injection time varies between 0.5 to 3.4ms in the best case and at most 6.1ms in the worst case. Finally, in all the tests we run the Morpheus compilation pipeline on a separate core with respect to the data plane application, and we noticed that in most of the case it consumes, on average, ~5% of CPU, with a peak of 15% for applications that consist of a pipeline of multiple programs chained together (e.g., BPF-iptables) and with lot of table entries to be analyzed (e.g., Katran).

### 6.4 Morpheus in action

To test the ability of Morpheus to track highly dynamic inputs, we fed the Router application with time-varying traffic and observed the throughput over the time (Fig. 9a). Recompilation period was conservatively set to 1 second. In the first 5 seconds we generate uniform traffic; here, the traffic-independent optimizations applied by Morpheus yield roughly 15% performance improvement over the baseline. At the 5th second, the traffic changes to a high-locality profile: after a quick learning period Morpheus specializes the code, essentially doubling the throughput. We see the same effect from the 10th second, when we switch to another high-locality trace with a new set of heavy-hitters: after a brief training period Morpheus dynamically adapts the optimized datapath to the new profile and attains 60–100% performance improvement. We also repeated the same test using a real-world traffic trace (CAIDA 2019 dataset, equinix-nyc [20]), counting 30M packets with an average size of 910B. The trace experiences also a low degree of traffic locality, with the most hit entry matched around 0.4% overall. In Figure 9b, we show how Morpheus consistently improves the throughput of the router by factor of ~10%.

Finally, in Figure 10, we report the multi-core scaling of Morpheus. Here, we still use the router when processing input traffic characterized by low-locality. The constant performance increase is enabled mainly by our adaptive instrumentation mechanism, which is able to track the flow states across the different cores, i.e., specific to the RSS context and, depending on their distribution optimize the code accordingly.

### 6.5 What can go wrong?

The flip side of dynamic optimization is the potential for a misguided run time code transformation to harm performance. With generic languages this can happen when the dynamic compiler steals CPU cycles from the running code [28, 87]; in such cases careful manual compiler parameter tuning and deep application-specific knowledge is needed to make up for the lost performance [72]. Similar issues may arise with dynamically optimizing network code, as we show below on the NAT use case [67]. The NAT is organized as a single large connection tracking table, updated from within the data plane on each new flow. This represents a worst-case scenario for Morpheus: fully stateful code, so that guards cannot be opportunistically elided, coupled with potentially high traffic dynamics. Yet, since traffic-independent optimizations can still be applied (Table 2) Morpheus can improve throughput by around 5% (from 4.36 to 4.58...
6.6 Morpheus with DPDK programs

We applied Morpheus to a DPDK program, the FastClick [13] version of the eBPF Router application, the same one used in PacketMill’s paper [34]. We configured it with either 20 or 500 rules from the Stanford routing tables [52] and generated traffic with different levels of locality. We compared the throughput and the latency of the baseline code, the Morpheus optimized and its version transformed with PacketMill, state-of-the-art DPDK packet-processing optimizer. In our tests PacketMill uses the following optimizations: removing virtual function calls, inlining variables, and allocating/defining the elements’ objects in the source code.

Fig. 11a reports the average throughput results. For only 20 prefix rules and with low locality traffic, PacketMill outperforms Morpheus by about 9%, whereas for high-locality traffic and larger forwarding tables Morpheus produces a whopping 469% improvement over PacketMill. The reason for the large performance drop from 20 rules to 500 rules is that LPM lookup is particularly expensive in FastClick (linear search), but Morpheus can largely avoid this costly lookup by inlining heavy hitters. The 99th percentile latency results (Fig. 11b) confirm this finding, with Morpheus decreasing latency 5-fold compared to PacketMill with high-locality traffic.

On the other hand, the reason for the lower performance of Morpheus in the low/no locality case (20 rules) can be found in the main difference between the two systems. First, Morpheus requires instrumentation to track table access patterns, which produces some run time overhead, while PacketMill does not apply online optimizations and so it does not need instrumentation at all. Second, PacketMill implements some optimizations that Morpheus does not (although nothing prevents us from implementing them), but in most cases the effect of these additional optimizations is masked by the speedup brought by the Morpheus traffic-level optimizations.

7 DISCUSSIONS

Add other optimizations to Morpheus compilation pipeline. Morpheus is orthogonal to most optimizations proposed in recent literature and can be extended to support them. For example, PacketMill optimizations [34], such as the reordering of metadata fields, could be easily integrated into Morpheus with the added benefit that, having access to the number of accesses to a given variable thanks to Morpheus’s instrumentation, we could obtain a more accurate reordering compared to PacketMill, which only estimates access patterns. Finally, traditional PGO optimizations can be used with Morpheus too, allowing the compiler to optimize the code on-the-fly, as opposed to traditional PGO approaches where the profile is collected offline.

Extend Morpheus to other data plane technologies. Morpheus comprises a data plane-independent core implementing the bulk of the optimization passes, and separate data plane-specific plugins. This makes it relatively easy to port Morpheus to new dataplanes that use well defined APIs (e.g. for table lookups or packet reception). This is the case of eBPF/XDP and DPDK FastClick, but any data plane with a clear API can be potentially used with Morpheus. To do this, the developer would need to write a new data plane plugin, providing the function signatures for the relevant API calls as well as some simple operators for Morpheus to read match-action table content and manipulate the underlying code (e.g., inject a guard).

The choice of working at the IR level. In Morpheus, all the optimizations are directly applied at the LLVM IR level. A major drawback of this approach is that by doing so we lose direct access to the low-level machine code, making certain optimizations impossible: peephole, vectorization/SIMD, or other micro-architectural optimizations [12, 22, 49, 88, 96]. Nevertheless, this choice provides also a series of benefits: (i) IR code is in the Static Single Assignment (SSA) form and SSA simplifies the use of different compiler optimization algorithms; (ii) Morpheus optimization passes can exploit flow information performed in the compiler itself to gather information about the code under consideration: for instance, we use LLVM MemorySSA analysis to retrieve information about variable and load/store dependencies; (iii) working at the IR level allows Morpheus to re-use part of the other optimizations already available in the compiler suite; (iv) finally, it allows to keep the optimization passes as generic as possible with respect to the language in which the data plane is written.

It is not all about table lookups. Morpheus heavily optimizes also the code surrounding the table lookups using the insights it
obtains during code analysis. It separates table lookup code into a fast-path, with the lookup results specialized for the heavy hitters, and a generic slow-path. It uses the table lookup code to also gather information about the table’s content, and how it is used in the rest of the code. This allows for optimizing the entire fast-path code after the table lookup without affecting the slow path in any way. For example, constants are folded from the JITted lookup code (effectively a set of nested if-then-else statements) into the surrounding code (i.e., each branch of the if-then-else is specific to a certain value of the conditional), unreach code is removed or tables are specialized depending on their run time entries.

Because of this, caching/JITting table lookups is just one, albeit very important, optimization that Morpheus performs. As shown in Figure 4, the results with the “no locality” trace demonstrate the combined benefits of all the optimizations that are independent from the input traffic, such as dead code elimination, constant propagation and data structure specialization, while the rest of the cases (“low/high-locality”) show the additional effect of traffic-dependent optimizations. Note that some optimizations cannot be directly measured since they are the results of a combination of other passes; e.g., the contribution of dead code elimination is dependent on constant propagation.

Morpheus dependence on compilation periodicity. The performance of Morpheus depend on how fast Morpheus can recompile the targeted code (see Table 3). In the presence of traffic changes that are faster than that, then traffic-dependent optimizations become less effective. Nevertheless, Morpheus can still speedup the original function, since other traffic-independent optimizations are still valid (Figure 4). Potentially, we could disable traffic-level optimizations when Morpheus discovers highly variable traffic that goes under the recompilation period (Section 6.5); this would reduce the impact of guards and instrumentation and increase the benefits of traffic-independent optimizations. Examining these techniques remains as future work.

8 RELATED WORK

Generic code optimization has a long-standing stream of research and prototypes [12, 21, 46, 49, 69, 74, 77, 82]. In the context of networking, domain-specific data-plane optimization has also gained substantial interest lately.

Static optimization of data-plane programs. Several packet I/O frameworks present specific APIs for developers to optimize network code [23, 24, 36, 42, 75], or implement different paradigms to efficiently execute packet-processing programs sequentially or in parallel [11, 50, 60, 62, 83, 88, 96]. Other proposals aim to remove redundant logic or merge different elements together [18, 51, 84]. These works provide static optimizations; Morpheus, on top of these, also considers run time insight to specialize generic network code.

Dynamic optimization of packet-processing programs. ESwit- ch [68, 79] was the first functional framework for the unsupervised dynamic optimization of software data planes with respect to the packet-processing program, specified in OpenFlow, being executed. PacketMill [34] and NFReducer [29] leverage the LLVM toolchain [57] instead of OpenFlow: PacketMill targets the FastClick datapath by exploiting the DPDK packet I/O framework and NFReducer aims to eliminate redundant logic from generic packet processing programs using symbolic execution. Morpheus is strictly complementary to these works: (1) it applies some of the same optimizations but it also introduces a toolbox of new ones (e.g., branch injection or constant propagation for stable table entries); (2) Morpheus can detect packet-level dynamics and apply more aggressive optimizations depending on the specific traffic patterns; and (3) Morpheus is data-plane agnostic, in that it performs the optimizations at the IR-level using a portable compiler core and relies on the built-in compiler toolchain to generate machine code and a data-plane plugin to inject it into the datapath.

Profile-guided optimization for packet-processing hardware. P2GO [91] and P5 [5] apply several profile-driven optimizations to improve the resource utilization of programmable P4 hardware targets. Some of the ideas presented in this work can also be used with programmable P4 hardware, provided it is possible to re-synthesize the packet processing pipeline without traffic disruption, with a notable difference: P2GO and P5 require a priori knowledge (i.e., the profiles) while Morpheus aims at unsupervised dynamic optimization.

9 CONCLUSIONS & FUTURE WORK

We presented Morpheus, a run time compiler and optimizer framework for arbitrary networking code. We demonstrated the importance of tracking packet-level dynamics and how they open up opportunities for a number of domain-specific optimizations. We proposed a solution, Morpheus, capable of applying them without any a priori information on the running program and implemented on top of the LLVM JIT compiler toolchain at the IR level. This allows to decouple our system from the specific framework used by the underlying data plane as much as possible. Finally, we demonstrated the effectiveness of Morpheus on a number of programs written in eBPF and DPDK and released the code in open-source to foster reproducibility of our results.

We consider Morpheus only as a first step towards more intelligent systems that can adapt to network conditions. As future work, we intend to integrate a run time performance prediction model [15, 47, 61, 76, 78] into Morpheus, enabling the compiler to reason about the effect of each different dynamic optimization pass. This would allow for selecting the most efficient subset of optimizations and adapt the recompilation timescales to the current network conditions.

ACKNOWLEDGMENTS

We would like to thank our shepherd, Edouard Bugnion, for his insightful suggestions; the anonymous ASPLOS reviewers for their enlightening comments, the artifact evaluation committee for their effort in reviewing our artifacts, and Srinivas Narayana for his feedback on the first draft of the paper. This work is partially supported by the UK’s EPSRC under the projects NEAT (EP/T007206/1) and by Facebook as part of the 2020 Networking Systems request for proposals (RFP) award. Gábor Révész was funded by the NKFIH/OTKA Project #135606, the MTA-BME Information Systems Research Group and the MTA-BME Network Softwarization Research Group.


