Iridescent: A Framework Enabling Online System Implementation Specialization

Vaastav Anand, Deepak Garg, Antoine Kaufmann
Max Planck Institute for Software Systems

Abstract

Specializing systems to specifics of the workload they serve and platform they are running on often significantly improves performance. However, specializing systems is difficult in practice because of compounding challenges: i) complexity for the developers to determine and implement optimal specialization; ii) inherent loss of generality of the resulting implementation, and iii) difficulty in identifying and implementing a single optimal specialized configuration for the messy reality of modern systems.

To address this, we introduce Iridescent, a framework for automated online system specialization guided by observed overall system performance. Iridescent lets developers specify a space of possible specialization choices, and then at runtime generates and runs different specialization choices through JIT compilation as the system runs. By using overall system performance metrics to guide this search, developers can use Iridescent to find optimal system specializations for the hardware and workload conditions at a given time. We demonstrate feasibility, effectivity, and ease of use.

1 Introduction

Specializing system implementations to workload characteristics and hardware can significantly improve performance and efficiency [4, 8, 16, 18, 22, 24, 27, 30, 31, 39, 41, 53, 55, 60, 70, 72, 76–78, 82, 86]. Achieving these benefits requires manual modification of the system implementations and recompilation. Part of the performance benefit arises from cascading compiler optimizations, e.g. by removing a feature enabling the compiler to eliminate dead code, in turn enabling further optimizations [27, 61, 83]. Today, system *specialization* for performance is manual, developer-driven, and iterative.

Building optimal systems in practice is a challenge because of three compounding factors: First, specialization is difficult for developers to implement, as it requires trial-and-error and maintaining multiple versions of key code paths. Next, specialization fundamentally comes at the cost of generality — a specialized system either performs poorly outside its regime, or completely fails. Finally, predicting performance implications of specialization choices for different workloads and hardware is almost impossible. To make matters worse, workload and platform conditions are dynamic and optimal specialization choices depend on them.

We propose a fundamentally different approach: *automated online system specialization guided by observed overall system performance.* Our goal is to specialize performance critical system systems at runtime to the exact momentary workload and hardware conditions, *without the human developer on the critical path.* To enable this, we re-distribute tasks between ahead of time development and runtime operation. Ahead of deployment, rather than choosing a concrete set of specializations, *developers specify the space of possible specializations*, along with a search strategy. After deployment, our runtime iteratively *generates and measures the performance of different specialized code* versions, choosing the currently optimal specialization based on observed overall system performance. This allows compile-time specializations at runtime, suitable for search-based auto-tuning [75, 78, 82].

To enable this, we introduce Iridescent, a framework for practical development of fast and efficient systems with online specialization. Iridescent enables both the incremental specialization of existing code-bases with minimal modifications, and in-depth design for specialization for maximizing performance. Iridescent builds on LLVM and targets systems written in typical (non-managed) languages such as C(++) or Rust that are architected around an event-loop. Concretely, Iridescent requires developers to separate the code-base into a specialized, performance critical handler core, and the remaining outer event loop code. Iridescent then provides the developer with a specialization API to annotate possible specializations in the code thereby specifying the space of possible specializations. Additionally, Iridescent provides hooks to optionally customize JIT code generation in depth. Finally, Iridescent provides a runtime API to control the exploration of different specialized implementations as the system runs.

We integrate Iridescent with multiple systems and show that with Iridescent enabled specializations, systems can gain a performance boost of upto 30×. Moreover, with Iridescent, developers can easily configure the system to automatically explore the specialization space to adaptively find the best performing specialization for dynamic workloads at runtime.

Machine / Workload	N=1024	N=256	N=64
IceLake	32	32	32
IvyBridge	16	16	4
CoffeeLake	32	4	4
AlderLake-p	32	4	2
AlderLake-e	64	4	4

Table 1: Optimal value of block size, B, for our block matrix multiply, across 5 hardware platforms and 3 workloads.

2 Motivation & Challenges

We illustrate the benefits and challenges of specialization with an example system: a server executing square matrix multiplications for clients. To optimize cache locality we use a blocked matrix multiplication algorithm [11, 84, 85]. We identify one key workload parameter, N, the matrix size, and a key configuration parameter, B the block size.

Code specialization improves performance. Both parameters offer a substantial opportunity for optimization through specialization. Unsurprisingly, we find fixing B as a compile-time constant instead of leaving it a variable improves throughput by up to 6.5×, by enabling the compiler to unroll and vectorize the inner loops. Note that B is an internal parameter that may affect performance, but any valid choice results in correct behavior with every workload. Similarly, assumptions about the workload, N, can also simplify the algorithm, e.g. by assuming that N is a multiple of B we avoid the need for copying and 0-padding partial blocks.

Optimal configuration depends on workload and HW. We now compare different block sizes for different workloads (matrix sizes) on 5 different processors. As Table 1 shows, different block sizes yield optimal performance for different workload and processor combinations. Picking a fixed B ahead of a practical deployment in a dynamic environment will not yield optimal performance. Moreover, even with single single-size workloads it is difficult to predict what block size will be ideal for a concrete processor.

2.1 Specialization is Effective but Complex

There is a long line of prior work that has established the performance and efficiency benefits of specializing code, and systems code more specifically. For example, interpreted languages may generate specialized instances of functions with constant parameters and optimize them accordingly [21, 49]. This is a generic optimization and typically done automatically and transparently by the runtime. Other optimizations are specific to individual systems and concrete concerns, such as inlining small table lookups in software network functions as if statements [53]. Like most compiler optimizations, these approaches are typically guided by simple heuristics around local metrics (e.g. frequency of the same parameter value, or table size) based on developer intuition.

Specialization effects often cascade. A key aspect of specialization techniques is that the effects combine. For example, PacketMill [27] first de-virtualizes function pointers [43] for the Click modular router [44], and then based on this further eliminates dead code and data structure fields. The analysis for the latter optimization is impossible before de-virtualization as the virtual calls prevent the compiler from analyzing the complete packet handling code.

2.2 Specialization is Challenging in Practice

(C1) Developer Complexity. A key barrier to specialization is the increased complexity for developers. Developers now need to reason about what assumptions will definitely hold and can help specialize the system for better performance. Next the developer needs to implement the required specialization and evaluate it. In practice, this frequently results in having to maintain multiple code versions. Unsurprisingly, this is not only complex but also laborious and frustrating, since it is often an iterative hit-and-miss process.

(C2) Loss of Generality. Specializations may hurt performance or break the system when the underlying assumptions are violated. For example, a network function may process 99.999% of packets if the processing loop and data structures are specialized in a specific way. However, specializing the system to this class of packets, and thereby simplifying code and data structures may break handling of the rare 0.001% cases. The need to gracefully handle rare or unexpected cases forces developers to specialize conservatively.

(C3) Optimal Specialization Choices. Modern system performance is a complex product of the emergent behaviors across all system components, it is also not feasible for developers to distill this into simple local cost models. Optimally specializing systems statically is fundamentally impossible since concrete workload and hardware parameters affect these choices, and both are dynamic in practice.

3 Approach

We enable practical specialization of systems for improved performance and efficiency. Our key idea to enable this given the challenges above, is to *automatically explore a space of possible specializations provided by the developer at runtime based on observed overall system performance.*

Simple specialization with simple annotations. We first observe that, while developers struggle with determining *optimal* specializations and implementing a system based on them, it is much easier for developers to suggest *possible* specialization assumptions. We thus enable developers to annotate system code with simple annotations to this end.

Specialization	Effect
Value	Convert runtime variables to constants; triggers const. prop., unrolling, vectorization, ...
Assumption	Add boolean assumptions; triggers dead code elimination, branch elimination, ...
Fast-Path	Add if-else based early exits with previously computed/cached results.
Custom Code	Generate custom code at runtime. e.g. dynamic LPM matching via nested branches.

Table 2: Iridescent specialization types and their effects.

Figure 2a shows an example of the annotations for the matrix multiplication server above and how developers can configure specializations at runtime. Additionally, developers can also provide LLVM transformations to perform intrusive app-specific specializations, and thereby generate further candidate configurations in the overall specialization space. Crucially, we do not rely on the developer to filter or rank specializations. Table 2 shows the broad class of specializations supported by Iridescent and their corresponding effects.

Guarded specialization with a fall-back. Later, when automatically instantiating specialized system versions through JIT compilation, we inject guard conditions into the code to use the specialized version of the code when applicable. If the specialized version is not applicable, the guard conditions ensure that we fall back to the unspecialized version of the system code for that invocation. While triggering this guard incurs overhead, it does ensure correctness. If the guard triggers sufficiently rarely, in combination with the benefit for the common case, the specialization is still a net win.

Online exploration guided by developer policy. Finally, we completely forgo cost models and heuristics for predicting which choice in the specialization space is optimal, and instead explore different points online as the system runs. Our guards ensure that the system always behaves correctly in this process. We rely on the developer to provide a policy for guiding this exploration. The developers control when to explore which points, manually, through existing autotuning solutions or by leveraging simple search strategies from our library. A key task of this policy is comparing the (application specific) overall system performance metrics, e.g. request throughput, when running different specialization points. Overall system performance metrics also implicitly factor in any overheads, e.g. triggering specialization guards for some calls. This ensures that the system converges to the optimal specialization point for the concrete combination of dynamic system, workload, and hardware conditions.

4 Iridescent Design

We now present the design of our approach in Iridescent.

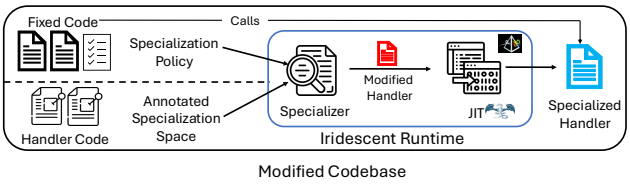

Figure 1: Architecture of a Iridescent-specialized system.

Iridescent Specialization API	
`spec_enum(lbl, x, ...)`	Enumeration spec. point (x∈...)
`spec_range(lbl, x, l, h)`	Range spec. point (l≤x≤h)
`spec_generic(lbl, x)`	Policy-controlled spec. point
`spec_assume(lbl, cond)`	Specialization assumption
`spec_custom_*(lbl, ...)`	Custom spec. point
Iridescent Policy API	
`Iridescent(handler_ir)`	Initialize runtime
`.handler(h)`	Get specialized handler
`.spec_space()`	Obtain code spec. space
`.specialize(c)`	Choose specialization; `c` maps spec. point labels ↦ values
`.add_custom_spec(n, gen)`	Add app specialization
`.customize_opts(passes)`	Modify codegen optimizations

Table 3: Iridescent API overview.

4.1 Anatomy of a Iridescent System

System Requirements for Iridescent. Iridescent targets performance critical systems not implemented in managed languages. We observe that typical performance critical systems code is almost exclusively architected with an event loop structure – an outer loop handling a sequence of events, such as processing requests or data elements, and an inner handler that processes the event. While Iridescent is designed to specialize these systems in varying dynamic environments, for these systems to fully gain the benefits of Iridescent, these systems should have stability in the workload and operating conditions for a sufficient period of time. Conditions should at least remain stable for the time Iridescent takes to recompile a new specialized variant (typically 10–400 ms, see §6.4).

Isolating performance critical code for specialization. Since Iridescent targets performance critical systems not implemented in managed languages, modifying code of the running system is non-trivial in general. We leverage the event-loop structure for pragmatic specialization, by requiring developers to separate the system code into two parts (Figure 1): (i) the *fixed code*, including initialization and the outer loop; (ii) the performance-critical event *handler code*. The developer compiles the fixed code as before, but integrates calls to the Iridescent API (Table 3). In the handler

```
void matmul(u64 *L, u64 *R, u64 *O, int N, int B) {
  B = spec_enum("B",B,2,4,8,16,32,64);
  N = spec_general("N",N);
  for (int kk = 0; kk < N; kk += B)
    for (int jj = 0; jj < N; jj += B)
      for (int i = 0; i < N; i++)
        for (int jjj = jj; jjj < jj + B; jjj++)
          for (int kkk = kk; kkk < kk + B; kkk++)
            M3[i*N+jjj] += L[i*N+kkk] * R[kkk*N+jjj];
}
```

(a) Handler Code

```
Iridescent rt("handler_code.ll");
void spec_policy() {
  Conf N_cs[2] ={{}, {{"N", 256}}};
  do { for (c : cartesian(rt->spec_space(), N_cs)) {
         rt.specialize(c); tput_counter = 0;
         sleep(...); // Sleep for some time to monitor
         if (tput_counter > best) {
           best = tput_counter; best_c = c; } }
    rt.specialize(best_c);
  } while(every 1min);
}
int main() {
  auto *matmul = rt.handler("matmul");
  rt.reg_opt_pipeline(opt_passes);
  launch_thread(spec_policy);
  while(true) { // Main processing loop
    Req *req = get_req(); Resp *res = new Resp();
    matmul(req.A, req.B, res.C, 2, req.N);
    tput_counter++;
  }
}
```

(b) Fixed Code

Figure 2: MMulBlockBench microbenchmark

code, the developer adds lightweight Iridescent specialization API calls and compiles to LLVM intermediate representation. In the fixed code, the developer obtains function pointers for the handlers through the Iridescent policy API, and calls these as before. Figure 2 illustrates this with our matrix multiplication server running example.

Handling Global State. During dynamic code generation, Iridescent will link handlers against symbols in the fixed code, to enable calls and accesses to global state. Since Iridescent repeatedly rebuilds handlers, *Iridescent requires that the handler code does not include definitions of global variables that must be preserved.* The developer has to move all global state to the fixed code, and reference it from the handlers.

4.2 Defining Specialization Space

The developer defines the specialization space through *specialization point* annotations in the handler code. At runtime, Iridescent replaces these annotations with specialized code, instrumentation, or disables them, guided by the policy. Iridescent offers four types of specialization points:

Value Specialization Point. A value specialization point signals to Iridescent that the handler code could be specialized to a constant for the wrapped expression. `spec_enum` indicates the value could be one of the specified values; `spec_range` instead specifies a range of integers; `spec_generic` leaves it to the policy to determine possible values.

Assumption Specialization Point. This provides a *possible specialization assumption* to Iridescent. The condition is a boolean predicate that Iridescent can provide to compiler optimizations to further optimize, through LLVM's builtin `llvm.assume`. Crucially, the assumption need not always hold. Unlike for the LLVM intrinsic where incorrect behavior results, Iridescent catches this with its specialization guards.

Fast-Path Specialization Point. This is a generic version of the hot key specialization of Morpheus [53]. It works in two phases: (i) the instrumentation phase: Iridescent inserts instrumentation code that samples (sampling rate selected by the developer) invocations of the specialized function to find the most popular inputs to the function along with their calculated outputs; (ii) the specialization phase: Iridescent updates the specialized function to have a specialized if-else chain where the top-N inputs along with their outputs are converted into a series of if-else checks to avoid paying the cost of re-computation for the heavy-hitting inputs. The value of N may be developer-provided or could be configured as a runtime constant that can be further specialized. If the input does not match any of the branches in the if-else chain, the computation simply defaults to the generic version.

Custom User-Defined Specialization Point. These specialization points, annotated through `spec_custom_*`, invoke developer-defined specializations registered in the fixed code through `add_custom_spec`. An example is generating an if-else chain as a special-case fast-path for an expensive data structure lookup or computation (§5.2).

4.3 Defining Specialization Policies

The developer implements the specialization policy in the system's fixed code using the Iridescent policy API. The policy decides what, when, and how Iridescent should specialize.

Figure 2b illustrates how developers can control the exploration and selection of specializations. In line 4, the `spec_space` API returns the specialization space generated by the annotations in the handler code. We combine this space with other specializations through a Cartesian product to obtain a complete set of specializations. In lines 4–8, the fixed code automatically tries out the different combinations in the set of specializations and chooses the best performing combination. To choose the best, the specialization policy uses the target end-to-end performance metric, in this case the throughput. We trigger re-exploration at a fixed-time interval to adapt to workload changes.

The developer can also optionally configure custom optimization passes through the `customize_opts` API to interpose on and customize code generation. Developers may also choose to enable instrumentation for select specialization points to dynamically identify opportunities for specialization. For example, the policy for the matrix multiplication server could detect the most frequent matrix size N, and only decide to specialize if $\geq$70% of requests have this value.

Exploring the specialization space. Iridescent provides a periodic exhaustive search strategy for simple cases as a library routine, with a simple call-back for providing the strategy with the system performance metric. For more complex cases, Iridescent integrates with OPPerTune [75] to leverage its state-of-the-art search strategy for configuration tuning of web services. Additionally, Iridescent also provides a thin interface for developing custom search strategies. We expect most systems will either utilize existing auto-tuning or implement custom strategies maximizing overall system performance, using the Iridescent building blocks.

4.4 Specialization Runtime

The specialization runtime has two key components: (i) the specializer generating code by applying selected specializations; (ii) the JIT compiling and optimizing the modified specialized code and making it available to the fixed code.

4.4.1 Specializer

The specializer generates the specialized version of the handler code. It operates at function granularity, and generates versions of functions with specialization points replaced depending on the chosen specialization configuration.

For each specialization point, Iridescent performs the following actions, depending on the configuration the policy supplied for this point. If the policy marks a point as disabled, the Iridescent specializer simply removes the annotation and replaces it with the original value, or skips over the point if it is an assumption point. To specialize an enabled value specialization point, Iridescent replaces the specialization point annotation in the handler code with the constant value supplied by the policy. By default, the specializer will also insert a specialization guard, which the developers may explicitly disable. For a custom specialization point, the specializer replaces the specialization point with the custom source code provided by the previously registered handler for that point type. The recompilation process then enables standard compiler optimization to take advantage of the newly inserted constants or code.

Instrumentation. Some specialization policies benefit from collecting runtime data to generate possible specialization configurations. To support such data collection, the specializer can optionally enable instrumentation for each specialization point. For a specialization point with instrumentation enabled, Iridescent additionally generates code for collecting and storing the observed actual values. The policy retrieves this information included in the result of the `spec_space` call.

4.4.2 JIT

Compiling the specialized code. The specialization runtime adds the specialized code generated by the specializer to the JIT. The JIT compiles the code and runs the developer-specified pipeline of optimization passes, including any custom developer-provided optimization passes, to generate an optimized version of the specialized code. By default, the JIT applies the default `O3` optimization pipeline if the developer has not provided a specific pipeline. In addition, the JIT also keeps a copy of the original, non-specialized version of the specialized code. This generic code serves as fallback for situations where the specialized code is not applicable.

Using the specialized code . To use the specialized handlers, the fixed code obtains function pointers to specialized code from the specialization runtime. The fixed code then uses these pointers to invoke the specialized handlers. If no specialization has been enabled the handlers default to the non-specialized versions of the functions. For simplicity, the fixed code only does this once at start of execution. The JIT creates a trampoline function which calls the most recent specialized version of the function. The trampoline function is stored at a fixed address and does not change across specialization iterations.

4.4.3 Correctness through Specialization Guards

Iridescent inserts a runtime check called a specialization guard for each enabled specialization point. The policy can disable this guard for individual points. The specialization guard ensures that the specialization conditions for generated code holds during execution. For this, the specializer inserts an early exit from the specialized version of the code on guard failure by throwing an exception. The Iridescent trampoline function catches these exceptions and re-routes control to the original non-specialized version of the function transparently without exposing the exception to the fixed code.

Restoring state and side effects. To ensure correct restoration of the state, the specializer will call a developer-defined cleanup function for reversing any side-effects before throwing the exception. It is critical to note here that not all side-effects are reversible (e.g. sending a packet to a neighbor), so Iridescent leaves the developer with the necessary control.

4.5 Prototype Implementation

We have implemented Iridescent in 5.5K lines of C++. Our implementation uses the LLVM IR and JIT [45] to generate the specialized code at runtime. The specializer is implemented as a set of LLVM transformation passes that operate on the LLVM IR of the instrumented handler code.

5 Case Studies

We evaluate Iridescent with four open-source systems and libraries and the MMulBlockBench microbenchmark. We explain each system along with the changes we made for integrating Iridescent below.

5.1 Tiling Benchmarks

MMulBlockBench. We implement a blocked/tiled version of matrix multiplication algorithm [11, 84, 85]. To integrate this benchmark with Iridescent, we convert the tile size, B, as a value specialization point.

CloverLeaf. We use a C-version of the CloverLeaf application from the SPEC ACCEL2023 benchmark [19]. In this application, we implement tiled-versions of the various mathematical kernels, and use value specialization points for the tile sizes, T_i and t_J. In total, this application has a 24 unique specialization points, each of which that can take 7 unique valeus resulting in a total of 7^{24} possible configurations.

5.2 LibLPM

LibLPM [66] is an open-source Longest Prefix Match (LPM) library written in C with built-in support for IPv4 and IPv6 addresses. LPM is a key operation in packet routing for finding the closest matching route in routing tables for incoming packets. To integrate LibLPM with Iridescent, we create two different specializations:

LibLPM-FP specializes the lookup function to add a fast-path specialization point. This specialization adds an exact match if-else fast-path for the top-N input IP addresses.

LibLPM-NI creates a code-generation specialization point in the LPM library. The code-generation specialization point generates a new version of the lookup function which generates a nested-if-else tree of checks consisting of prefix match checks for the incoming address for each lpm entry. The nested-if-else tree starts at the least specific matching rule (based on prefix length) and progressively checks for the most specific matching rule until it can't find one anymore. With this code generation specialization, we embed the prefix rules directly into the codebase, forgoing expensive dynamic datastructure traversals and enabling further optimizations for individual branches.

LibLPM-NI-FP combines the previous two specializations by adding a fast-path specialization point in the generated nested-if specialized lookup function.

5.3 TAS

TAS [42], TCP acceleration as a service, is a lightweight software TCP network fast-path that is optimized for common-case client-server RPCs. TAS executes common-case TCP operations in an isolated fast path that uses DPDK [25], while handling corner cases in a slow path.

To integrate TAS with Iridescent, we convert its `BATCH_SIZE` macro to a value specialization point. In TAS, the `BATCH_SIZE` variable is used in three different scenarios: (i) to determine the number of packets to read from the NIC, (ii) to determine the number of packets to read from the application queues, and (iii) to determine the number of packets to read from the queue manager. For more fine-grained control, we convert each of these usage instances of the `BATCH_SIZE` to three separate runtime constant specialization points - (i) `rx_batch`, (ii) `queues_batch`, and (iii) `qman_batch`.

5.4 FastClick

FastClick [7] is an extended version of the Click Modular Router [44] with Netmap and DPDK support for running the modular router in userspace. A FastClick (or Click) router is assembled from individual packet processing modules called elements. Each element implements simple router functions such as packet classification, scheduling, routing, and interacting with network devices. Users configure the router with different graphs of connected elements.

To integrate FastClick with Iridescent, we modify the `LinearIPLookup` element of FastClick to create a Iridescent-enabled `LinearIPLookup` element. The original element uses a linear search algorithm to find the best matching route for an incoming packet based on the routing table. We modify the packet-processing function of the `LinearIPLookup` element to add a fast-path specialization point. The target of this specialization point is the internal lookup function of the `LinearIPLookup` element.

5.5 Network Functions

Network Functions is a suite of the network functions that include DPDK and ebpf implementations of common network functions such as a NAT, Router, policer, among others. These functions have been developed and used for evaluation by various research projects such as Pix [36] and Vigor [88]. We extract the implementations from the Pix artifact [35].

To integrate the Network Functions with Iridescent, we convert the `BATCH_SIZE` used by the DPDK version of these functions into a runtime constant specialization point. In the

Machine (Processor)	**Constant** (c) (cycles/op)	**Variable** (v) (cycles/op)	**Benefit** (v/c)
IceLake	175297	284944	61%
IvyBridge	250434	661295	246%
CoffeeLake	168817	581130	348%
AlderLake-p	173350	583724	336%
AlderLake-e	206924	557572	269%

Table 4: Impact of turning block size, B, as compile-time constant at runtime for N=64 for different hardwares (Table 1)

network functions, the BATCH_SIZE controls the number of packets that are read from a network port at a given time.

6 Evaluation

In this section, we evaluate how much Iridescent improves performance of systems by enabling online specializations. We show how we can use online specialization with Iridescent for three use-cases: (i) enabling compile-time optimizations at runtime; (ii) enabling incremental specializations at runtime; (iii) design exploration at runtime. We answer the following questions:

- Do Iridescent-enabled optimizations improve performance? (§6.2)
- Can Iridescent find an optimal specialization through exploration at runtime under dynamic conditions? (§6.3)
- What is the overhead of using Iridescent? (§6.4)

6.1 Experimental Setup

Testbed. For our experiments, unless otherwise stated, we configure two identical machines as client and server. They are directly connected with a pair of 100 Gbps Mellanox ConnectX-5 Ethernet adapters. Each machine has two Intel Xeon Gold 6152 processors at 2.1 GHz, each with 22 cores for a total of 44 cores and 187 GB of RAM per machine. We run Linux kernel 5.15 with Debian 11.

6.2 Compiler Optimizations at Runtime

Compile-time optimizations often produce more performant code. These compile time optimizations include constant propagation, loop unrolling, dead-code elimination, and use of vectorization instructions. With Iridescent, we can enable compile time optimizations using runtime data.

Iridescent enables cascading compiler optimizations. First, we show the potential benefits of enabling compiler optimizations at runtime for the MMulBlockBench microbenchmark in Table 1. Specifically, for the workload $N = 64$, we compare the performance of keeping the optimal block size, B, as a runtime parameter to that of converting it into a constant at runtime using Iridescent, thus enabling downstream compiler optimizations. For both configurations, we execute the function a fixed number of times and measure the number of cycles spent per function execution. Table 4 shows the benefit of converting B into a compile-time constant for different machines. For each machines, we get at least 50% reduction in cycles, and greater than 240% reduction in cycles for four of the five operating conditions. This improvement is a direct impact of Iridescent's specialization allowing different compiler optimizations to cascade and combine together to produce a more efficient version of the code. In this specific case, the specialization of B allows the JIT compiler to first easily unroll the loops in the matrix multiplication function and then leverage optimized vector instructions.

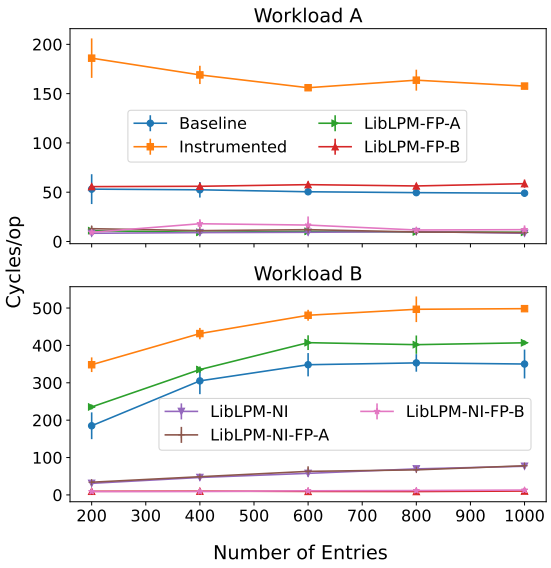

Figure 3: Average cycles required to perform a table lookup with varying table sizes under two different workloads.

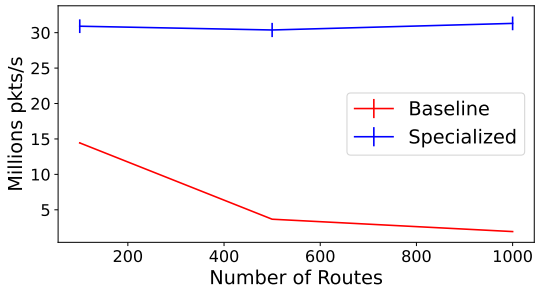

Figure 4: FastPath router throughput with Iridescent specialized LinearIPLookup element for different routing table sizes at 100% fast-path hit rate.

Iridescent incorporates existing specializations. To showcase Iridescent's ability to easily incorporate exist-

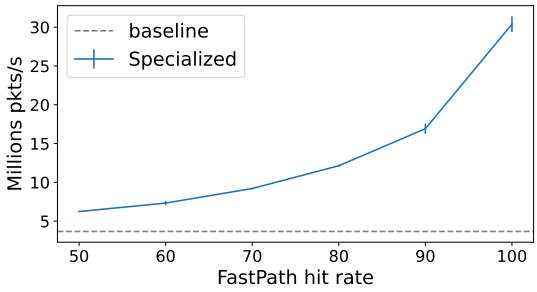

Figure 5: FastPath router throughput with Iridescent specialized LinearIPLookup element for varying fast-path hit rates.

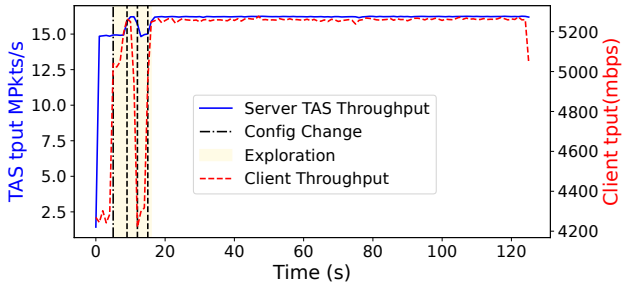

Figure 6: Automatic exploration and specialization of different configurations for TAS.

ing runtime specializations, we implement Morpheus' [53] map hot-keys specialization of as a fast-path specialization point in the packet processing function of FastClick's LinearIPLookup element. We configure the fast-path specialization point with a 0.01% sampling rate for the instrumentation phase. We execute the FastPath router on one machine and treat it as the device under test and execute Pktgen to execute an open-loop workload to generate packets with destination ip addresses from a given set of IP addresses. We repeat this experiment for different sizes of the routing table in the LinearIPLookup as well as different sizes of the fast-path to measure the throughput of the router under different conditions. Figure 4 shows that Iridescent achieves up to 15× improvement in the throughput based on the size of the table at 100% fast-path hit rate. The improvement increases with larger table sizes, as the basline requires longer linear scans here. Figure 5 shows that for a router with a 1000-entry routing table, even with a 50% fast-path hit rate, Iridescent achieves an approximately 5× throughput improvement. The improvement increases with the the fast-path hit rate.

Iridescent enables custom specializations. Next, we show the benefits of custom compile-time optimizations for LibLPM. For this experiment, we set up eleven different configurations resulting from a combination of five different LPM table sizes and two different workloads - Workload A and Workload B. In Workload A, the incoming IP address is an IP address that matches a very specific prefix entry in the routing table. In Workload B, the incoming IP address is an IP address that only matches the LPM prefix entry of prefix length 0. Thus, Workload A and Workload B cover the best and worst cases respectively for the LPM lookup function. We execute these workloads with seven specializations:

(i) Baseline: no Iridescent-enabled specialization.
(ii) Instrumented: Iridescent-enabled specialization capturing the most popular inputs (10% sampling); data used to guide the fast path specializations.
(iii) LibLPM-FP-A: fast-path specialization of size 1 specialized for Workload A.
(iv) LibLPM-FP-B: fast-path specialization of size 1 specialized for Workload B.
(v) LibLPM-NI: nested-if code generation specialization.
(vi) LibLPM-NI-FP-A: Combine LibLPM-NI and -FP-A.
(vii) LibLPM-NI-FP-B: Combine LibLPM-NI and -FP-B.

Figure 3 shows average cycles per LPM lookup lookup function for all different combinations of configurations and specializations. For Workload A, all three specializations, LibLPM-FP-A, LibLPM-NI, LibLPM-NI-FP-A, achieve up to 6× reduction in cycles per execution. However, unlike the LibLPM-FP and LibLPM-NI-FP-A specialization, this LibLPM-NI specialization does not require an additional instrumentation phase to sample executions to find the optimal candidates for the fast path. LibLPM-NI improves performance across the board for all workloads by up to 6×. Despite this performance boost, LibLPM-NI is not the best specialization for Workload B. For Workload B, the best optimizations are the *-FP-B specializations as they improve by up to 30× as all instantiations of Workload B hit the fast-path.

6.3 Runtime Design Exploration

Iridescent enables runtime design exploration. We show runtime design exploration for TAS [42]. We experiment in a typical server-client setting, with the server providing an echo service. The client is multi-threaded and executes an open-loop workload of 64 byte packets. Both the server and client are TAS-enabled. In our experiment, we modify the server-side TAS and measure the throughput of the server side TAS as millions of packets processed per second. With Iridescent we explore different values of the `rx_batch` in the TAS source code on the server side. Figure 6 shows Iridescent automatically determining and then selecting the best-configuration on the server side by exploring the different values of the batch size for different code locations.

Iridescent automatically adapts to changing workloads. To show that Iridescent can automatically adapt to workloads, we first use the MMulBlockBench microbenchmark. We execute a Iridescent-enabled version of the function and a non-Iridescent enabled baseline. In the baseline, the block size B is a runtime parameter. In the Iridescent-enabled version, Iridescent explores different values of B by specializing

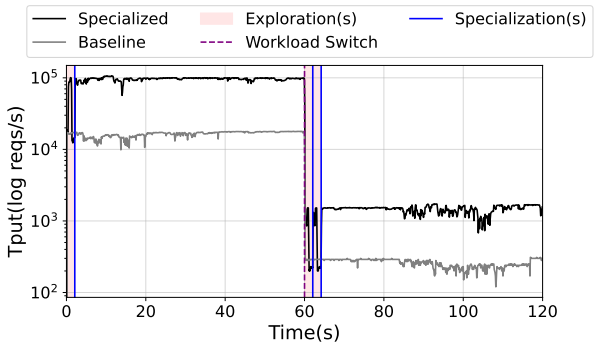

Figure 7: Automatic exploration and specialization of MMulBlockBench for two different workloads.

the value in the code and converting B into a compile-time constant. We execute each version with two different workloads in succession, each lasting one minute. We measure the overall throughput (executions/s) for both configurations. Figure 7 shows how Iridescent explores the different block sizes and finds the most performant configuration compared to the baseline. Moreover, Iridescent can automatically detect the workload change based on the drop in throughput and restart the A/B testing process.

Next, we show runtime design exploration and adaptation for two different network functions, vignat and vigpol, from vigor [88]. We setup our experiment with a packet generator on one machine and a device under test (DUT) on the other machine. We use two ports for the DUT. The first port acts as the internal facing port whereas the second port as the external facing port. The DUT executes the specific network function for every incoming packet generated by the packet generator. The network function implementation has different execution pathways for packets depending on whether they arrive on the internal port or the external port. The network function is enabled with the Iridescent specialization policy to trigger an exploration phase to find the best performing `BATCH_SIZE` whenever there is a change in the workload. We execute the experiment in two phases: in the first phase, packets arrive only on the external facing port and in the second phase, packets arrive only on the internal facing port. Figure 8 shows the results for the NAT and the Policer network functions. Iridescent finds the best performing configuration in each phase with different optimal values.

Iridescent transparently enables exploration-based adaptation with custom specializations. We show runtime design exploration for FastClick LinearIPLookup. We run the FastPath router on one machine, treat that machine as the Device Under Test and run Pktgen to execute an open-loop workload generating packets with destination ip addresses from a given set of IP addresses. At the 1 minute-mark, we completely switch the destination IP address set

System	Time (ms)	Size (KB)
MMulBlockBench	10 ± 1	10
LibLPM-FP	72 ± 9	80
Network Functions	98 ± 5	67
TAS	340 ± 5	360
FastClick	11 ± 1	9.2

Table 5: JIT compilation time and corresponding size

with no overlap with the initial address. The router is configured with a Iridescent specialization policy that triggers an exploration whenever it detects a large change ($\geq 25\%$) in measured throughput. Figure 9 shows how the throughput of the router changes for the different IP sets. At the start, the specialization policy triggers an exploration as it detects a large in-flux of packets, a significant change from 0 incoming packets. First, Iridescent runs a quick instrumentation phase for around 100ms to find the most popular incoming destination IP addresses. After the instrumentation phase, Iridescent switches to an exploration phase to explore the size of the fast-path, i.e., how many addresses should be included in the generated if-else source code. Once the exploration finishes, Iridescent switches to the best performing fast-path size and generates the specialized code which persists until the workload switch at the 1-minute mark. Due to the workload change, the specialization policy kicks in automatically, and Iridescent restarts the instrumentation and exploration phase. As shown in Figure 9, Iridescent can quickly adapt to changes in workload to generate optimal specialized code under dynamic conditions.

Iridescent can explore large specialization spaces. We explore the large specialization space of CloverLeaf with OPPerTune's exploration algorithm integrated with Iridescent. We limit the number of configurations to be explored to 20 out of a total of 7^{24} total configurations. We observe that, Iridescent can find an configuration that reduces the number of cycles spent per iteration of the main CloverLeaf loop by $3\times$ on average. By leveraging state-of-the-art autotuning exploration strategies, Iridescent can efficiently find an optimal configuration from large specialization spaces.

6.4 Iridescent but at what Cost?

Compilation cost. Table 5 shows the JIT compilation time for each of the target systems. Note that, this compilation happens off the critical path so it is not a performance bottleneck. However, this cost does dictate the propagation delay, i.e., the time taken for the specialized version of the code to be available for the execution.

To analyze how the compilation time changes with the increase in the size of the JIT, we consider the code generated by Iridescent for the LibLPM-NI specialization. Figure 10 shows the change in the compilation time for the LibLPM-NI

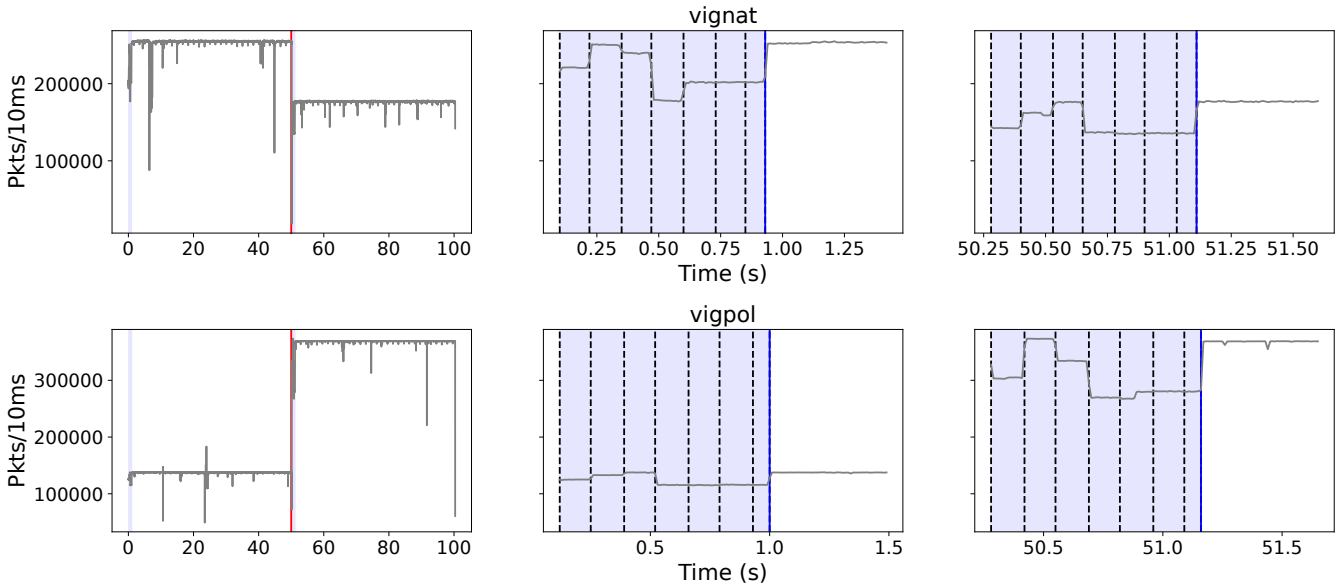

Figure 8: Batch size exploration with Network Functions

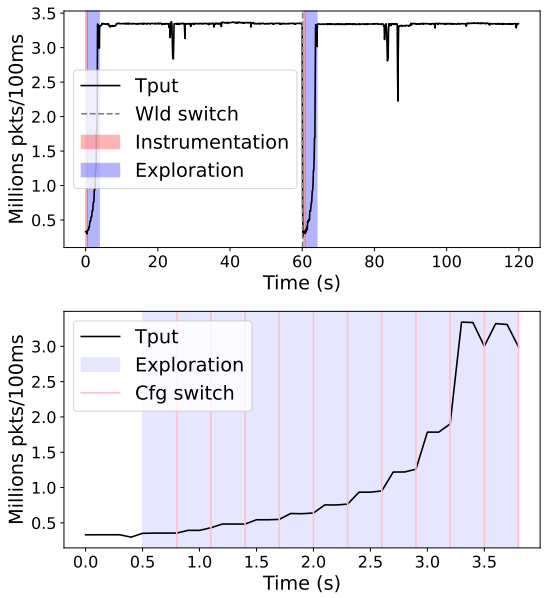

Figure 9: Optimal fast path size exploration with Iridescent.

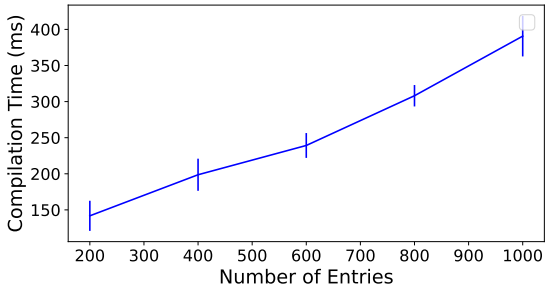

Figure 10: Compilation time of LibLPM-NI (nested-if specialization) as a function of the number of elements in the LPM lookup table.

specialization as a function of the number of the elements in the lpm lookup table. The compilation time increases linearly with increase in the number of lpm entries. This is because Iridescent generates one basic block in the specialized code for each lpm entry. As a result, the code size generated by Iridescent grows linearly with the total number of entries.

Developer cost. Table 6 shows the necessary lines of code changes to integrate Iridescent specializations into the target systems. For all cases, fewer than 100 lines of code change.

Implementing the custom LPM specialization only required 162 lines of code for enabling the custom user-defined LPM specialization.

Instrumentation overhead. To measure the overhead of adding instrumentation, we devise a microbenchmark called SimpleBench, consisting of two very simple functions, `f` and `g`. `f` computes the square of the input and `g` computes the product of its two inputs. For `f`, we add a general specialization point for its input `a`. For `g`, we add a range-based specialization for its second input `b`. For both the specialization points `a` and `b`, we first execute their respective functions, `f` and `g`, in normal mode and then use Iridescent to add instrumentation. We execute each configuration one million times. As a baseline, each function on average takes up to 8 cycles per execution in normal mode. For `a`, which is a general specialization point, adding instrumentation adds an overhead

System	Handler Annotations (LoC)	Spec Policy (LoC)
MMulBlockBench	2	47
LibLPM	4	9
Network Functions	5	43
TAS	8	38
FastClick	5	88

Table 6: Lines of Code change required to integrate Iridescent.

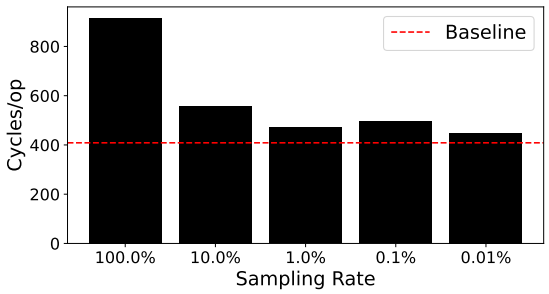

Figure 11: Impact of sampling rate on instrumentation overhead for LibLPM-FP configuration for Workload B.

of around 450 to 500 cycles per execution. For `b`, which is a specialization point designed for more efficient instrumentation for specialization points with values in fixed ranges, adding instrumentation adds an overhead of around 1 extra cycle per operation. When possible, Iridescent users should opt to use the more efficient specialization point. However, as this ma not always be possible and users may be forced to use the general specialization point, Iridescent allows users to specify a sampling rate to avoid incurring the instrumentation cost for low-latency computations. Figure 11 shows how decreasing the sampling rate for the LibLPM-FP Workload B configuration can decrease the instrumentation overhead.

Specialization guards and failures. To measure the cost of specialization guards and failures, we use the SimpleBench microbenchmark. For function `f`, we use Iridescent to specialize the value of `a` and generate two different versions. In the first version, we disable the specialization check at the entrypoint of `f`. We then execute this version of the function with the specialized value as input one million times. Each execution of this version takes 7 cycles. In the second version, we enable the specialization check at the entrypoint of `f`. If the specialization check fails, i.e. the input value of `a` does not match the specialized value of `a`, then this version of `f` throws an exception that is caught by Iridescent-inserted trampoline code which subsequently calls the generic version of the function. We first execute this version of the function with inputs such that the specialization guards always pass. This check adds an extra cycle per execution of the function. Next, we execute this version of the function with inputs such that the specialization guards always fail. This results in exception handling behavior which adds approximately 5000 cycles of overhead per execution. Thus, the raw cost of a specialization check is 1 extra cycle whereas that of a specialization failure is approximately 5000 cycles.

7 Related Work

7.1 Offline Automatic Specializations

We classify offline automatic specialization techniques in four categories. (i) Compile pass selection techniques automatically select the optimal sequence of compiler optimization passes through auto-tuning methods for trying out different combinations [58, 58, 59, 62], (ii) Specialized compiler optimization passes that can generate specialized code for specific hardware features like vectorization [56] or prefetching [73], or for other specific optimizations [2, 47, 52, 69], (iii) Feedback-Driven Optimizations (FDO) [13, 23, 33, 33, 34, 38, 57, 61, 67, 71, 79, 87, 89, 90], also known as PGO (Profile-Guided Optimizations), are compiler optimizations consisting of multiple executions of the program to be optimized by collecting relevant metrics from a benchmarking run and then using these metrics to generate optimized versions. FDO techniques in their current form lack flexibility as they are specialized to work well for only one workload (the workload it was trained on), (iv) Specialized Code Tuning approaches either use trial runs to tune specific parameters in the code [46, 64] or apply specialized optimizations to different pieces of the code [64, 81]. Unlike Iridescent, existing offline automatic optimizations cannot adapt to changes in workload or environmental settings dynamically. Consequently, these techniques are sub-optimal for the conditions they were not trained on.

7.2 Online Automatic Optimizations

We classify existing online automatic optimizations in four categories — (i) runtime code selection, (ii) configuration optimizations, (iii) generic code specializations, (iv) domain-specific specializations.

Runtime code selection. In this optimization technique, developers pre-generate multiple different versions of the code at compile-time and then use measured runtime statistics and offline-trained models to switch between the versions [20, 40, 76, 78]. These techniques are limited as they must anticipate and generate all possible versions a priori. Consequently, in contrast to Iridescent, these techniques can not fully leverage runtime information for specialization.

Config optimizations. Runtime auto-tuning techniques such as OtterTune [82] often perform tuning through

trial sessions to find optimal values for different configuration knobs. Some AutoTuning techniques such as OPPerTune [75] and SoftSKU [77] apply the AutoTuning paradigm of measurement-driven search at runtime without any trials. However, these techniques are critically limited to environmental parameters and knobs as they work transparently with respect to the deployed application. These techniques are currently not applicable to code optimizations or specializations. Iridescent enables the auto-tuning paradigm of measurement-driven search at runtime to find the best combination of code specializations for the system.

Code specializations. Runtime code specialization techniques [5, 17, 18, 22, 32, 65] produce more efficient versions of the code by exploiting values and invariants that only exist at runtime. Today, these specializations manifest in three ways: (i) binary rewriting techniques [3]; (ii) Transparent Dynamic Optimization (TDO) techniques such as Dynamo [6], The Transmeta Code Morphing Software [23], DynamoRIO [9]. These optimize code at runtime without requiring any modifications by capturing and optimizing traces (sequences of instructions) that are commonly executed; (iii) optimizations in JIT compilers for interpreted languages in the form of Type Specialization [14, 15, 57] for specializing the types of data for dynamically typed languages or as Value Specialization for interpreted languages [21, 49] where the parameter values of hot functions are converted into constants. These specializations may be optionally applied depending on the computation context [28, 37, 54, 63]. In contrast to Iridescent, they are generic and typically done automatically by the runtime based on internal cost models which may not be appropriate for specific application contexts. Moreover, these specializations are limited in scope and do not take the system's end-to-end performance into consideration and do not leverage domain- and situation-specific optimizations.

Domain-specific specializations. Recently, value specializations have extended beyond interpreted languages for specific use cases such as value specializations for GPU kernels [29], or through incremental specializations for packet processing frameworks [27, 53, 68]. Iridescent provides a general framework for configuring, applying, exploring, and selecting such specializations at runtime.

8 Discussion

Debugging. Iridescent makes it harder for developers to debug performance issues. This is because the generated specialized code is often different from the code that was originally written by the developer, which makes it harder for the developers to fully understand the execution sequences. This is further exacerbated by the fact that different workloads may lead to different specializations making it difficult to reproduce issues seen in production. Combining Iridescent with monitoring and distributed tracing techniques could offer developers a solution to their debugging problems.

Leveraging scale & ML for exploration. Several ML-based techniques exist for tuning configurations [1, 12, 26, 38, 48, 50, 51, 68, 74, 75, 80, 82]. These ML techniques can be implemented as algorithms with OPPerTune [75], which we have already integrated with Iridescent. Moreover, Iridescent could further incorporate SmartChoices [10] to allow different ML techniques to apply to specific specialization points. As part of future work, Iridescent could potentially also parallelize exploration across a large number of replicas to explore larger subsets of the specialization space.

Limitations. While Iridescent is designed to operate in dynamic and irregular operating conditions, exploration overhead overshadows performance benefits if the operating conditions do not remain stable for long enough to reap the benefits of the specialization optimizations. Ideally, developers should configure the specialization policy to opt for the general version of the handler code in such scenarios.

9 Conclusions

In this work, we presented Iridescent, a framework enabling online workload-driven runtime specialization of systems to improve performance. Iridescent provides a toolkit for developers to implement system-specific specializations that utilize runtime data and invariants to generate more performant systems. Iridescent provides the necessary tooling for developers to configure automatic runtime exploration of the space of potential specializations to find the best performing configuration at each instant. By combining developer insight with automated compiler optimizations, Iridescent enables systems to adapt efficiently to changing workloads and environments with minimal manual effort.

Acknowledgments

We thank Yuchen Qian for his contributions to an early prototype Iridescent implementation.

References

[1] Ibrahim Umit Akgun, Ali Selman Aydin, Andrew Burford, Michael McNeill, Michael Arkhangelskiy, and Erez Zadok. Improving storage systems using machine learning. *ACM Transactions on Storage*, 19(1):1–30, 2023.

[2] Christie L Alappat, Johannes Seiferth, Georg Hager, Matthias Korch, Thomas Rauber, and Gerhard Wellein. Yasksite: stencil optimization techniques applied to explicit ode methods on modern architectures. In *2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 174–186. IEEE, 2021.

[3] AP Arif Ali and Erven Rohou. Dynamic function specialization. In *2017 International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS)*, pages 163–170. IEEE, 2017.

[4] Omid Alipourfard, Hongqiang Harry Liu, Jianshu Chen, Shivaram Venkataraman, Minlan Yu, and Ming Zhang. CherryPick: Adaptively unearthing the best cloud configurations for big data analytics. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pages 469–482, 2017.

[5] Matthew Arnold, Stephen Fink, David Grove, Michael Hind, and Peter F Sweeney. Adaptive optimization in the jalapeno jvm. In *Proceedings of the 15th ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications*, pages 47–65, 2000.

[6] Vasanth Bala, Evelyn Duesterwald, and Sanjeev Banerjia. Dynamo: A transparent dynamic optimization system. In *Proceedings of the ACM SIGPLAN 2000 conference on Programming language design and implementation*, pages 1–12, 2000.

[7] Tom Barbette, Cyril Soldani, and Laurent Mathy. Fast userspace packet processing. In *2015 ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)*, pages 5–16. IEEE, 2015.

[8] Sapan Bhatia, Charles Consel, A-F Le Meur, and Calton Pu. Automatic specialization of protocol stacks in operating system kernels. In *29th Annual IEEE International Conference on Local Computer Networks*, pages 152–159. IEEE, 2004.

[9] Derek Bruening, Timothy Garnett, and Saman Amarasinghe. An infrastructure for adaptive dynamic optimization. In *International Symposium on Code Generation and Optimization, 2003. CGO 2003.*, pages 265–275. IEEE, 2003.

[10] Victor Carbune, Thierry Coppey, Alexander Daryin, Thomas Deselaers, Nikhil Sarda, and Jay Yagnik. Smartchoices: hybridizing programming and machine learning. *arXiv preprint arXiv:1810.00619*, 2018.

[11] Steve Carr and Ken Kennedy. Blocking linear algebra codes for memory hierarchies. In *PPSC*, pages 400–405. Citeseer, 1989.

[12] Stefano Cereda, Stefano Valladares, Paolo Cremonesi, and Stefano Doni. Cgptuner: a contextual gaussian process bandit approach for the automatic tuning of it configurations under varying workload conditions. *Proceedings of the VLDB Endowment*, 14(8):1401–1413, 2021.

[13] Dehao Chen, David Xinliang Li, and Tipp Moseley. Autofdo: Automatic feedback-directed optimization for warehouse-scale applications. In *Proceedings of the 2016 International Symposium on Code Generation and Optimization*, pages 12–23, 2016.

[14] Maxime Chevalier-Boisvert and Marc Feeley. Interprocedural type specialization of javascript programs without type analysis. *arXiv preprint arXiv:1511.02956*, 2015.

[15] Maxime Chevalier-Boisvert, Laurie Hendren, and Clark Verbrugge. Optimizing matlab through just-in-time specialization. In *International Conference on Compiler Construction*, pages 46–65. Springer, 2010.

[16] Inho Choi, Anand Bonde, Jing Liu, Joshua Fried, Irene Zhang, and Jialin Li. Ml-native dataplane operating systems. In *Proceedings of the 16th ACM SIGOPS Asia-Pacific Workshop on Systems*, pages 8–14, 2025.

[17] Charles Consel, Luke Hornof, François Noël, Jacques Noyé, and Nicolae Volanschi. A uniform approach for compile-time and run-time specialization. In *Partial Evaluation: International Seminar Dagstuhl Castle, Germany, February 12–16, 1996 Selected Papers*, pages 54–72. Springer, 2005.

[18] Charles Consel and François Noël. A general approach for run-time specialization and its application to c. In *Proceedings of the 23rd ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 145–156, 1996.

[19] Standard Performance Evaluation Corporation. Spec accel benchmark suite. Available from https://www.spec.org/accel, 2023.

[20] Diego Costa and Artur Andrzejak. Collectionswitch: A framework for efficient and dynamic collection selection. In *Proceedings of the 2018 International Symposium on Code Generation and Optimization*, pages 16–26, 2018.

[21] Igor Rafael de Assis Costa, Henrique Nazaré Santos, Péricles Rafael Alves, and Fernando Magno Quintao Pereira. Just-in-time value specialization. *Computer Languages, Systems & Structures*, 40(2):37–52, 2014.

[22] Jeffrey Dean, Craig Chambers, and David Grove. Selective specialization for object-oriented languages. *ACM SIGPLAN Notices*, 30(6):93–102, 1995.

[23] James C Dehnert, Brian K Grant, John P Banning, Richard Johnson, Thomas Kistler, Alexander Klaiber, and Jim Mattson. The transmeta code morphing/spl trade/software: using speculation, recovery, and adaptive retranslation to address real-life challenges. In

International Symposium on Code Generation and Optimization, 2003. CGO 2003., pages 15–24. IEEE, 2003.

[24] Bangwen Deng, Wenfei Wu, and Linhai Song. Redundant logic elimination in network functions. In *Proceedings of the Symposium on SDN Research*, pages 34–40, 2020.

[25] DPDK Project. Data plane development kit. http://www.dpdk.org/, 2022. Retrieved Feb 2, 2022.

[26] Songyun Duan, Vamsidhar Thummala, and Shivnath Babu. Tuning database configuration parameters with ituned. *Proceedings of the VLDB Endowment*, 2(1):1246–1257, 2009.

[27] Alireza Farshin, Tom Barbette, Amir Roozbeh, Gerald Q Maguire Jr, and Dejan Kostić. Packetmill: toward per-core 100-gbps networking. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 1–17, 2021.

[28] Olivier Flückiger, Guido Chari, Ming-Ho Yee, Jan Ječmen, Jakob Hain, and Jan Vitek. Contextual dispatch for function specialization. *Proceedings of the ACM on Programming Languages*, 4(OOPSLA):1–24, 2020.

[29] Giorgis Georgakoudis, Konstantinos Parasyris, and David Beckingsale. Proteus: Portable runtime optimization of gpu kernel execution with just-in-time compilation. In *Proceedings of the 23rd ACM/IEEE International Symposium on Code Generation and Optimization*, pages 507–522, 2025.

[30] Hamid Ghasemirahni, Tom Barbette, Georgios P Katsikas, Alireza Farshin, Amir Roozbeh, Massimo Girondi, Marco Chiesa, Gerald Q Maguire Jr, and Dejan Kostić. Packet order matters! improving application performance by deliberately delaying packets. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pages 807–827, 2022.

[31] Hamid Ghasemirahni, Alireza Farshin, Dejan Kostic, and Marco Chiesa. Just-in-time packet state prefetching. *arXiv preprint arXiv:2407.04344*, 2024.

[32] Brian Grant, Matthai Philipose, Markus Mock, Craig Chambers, and Susan J Eggers. An evaluation of staged run-time optimizations in dyc. *ACM SIGPLAN Notices*, 34(5):293–304, 1999.

[33] Wenlei He, Hongtao Yu, Lei Wang, and Taewook Oh. Revamping sampling-based pgo with context-sensitivity and pseudo-instrumentation. In *2024 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 322–333. IEEE, 2024.

[34] Urs Hölzle and David Ungar. Optimizing dynamically-dispatched calls with run-time type feedback. In *Proceedings of the ACM SIGPLAN 1994 conference on Programming language design and implementation*, pages 326–336, 1994.

[35] Rishabh Iyer, Katerina Argyraki, and George Candea. Performance interface extractor (pix) artifact. Accessed November 2024 from https://github.com/dslab-epfl/pix, 2022.

[36] Rishabh Iyer, Katerina Argyraki, and George Candea. Performance interfaces for network functions. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pages 567–584, 2022.

[37] Era Jain and Subhajit Roy. Phase directed compiler optimizations. In *2016 IEEE 23rd International Conference on High Performance Computing (HiPC)*, pages 270–279. IEEE, 2016.

[38] Saba Jamilan, Tanvir Ahmed Khan, Grant Ayers, Baris Kasikci, and Heiner Litz. Apt-get: Profile-guided timely software prefetching. In *Proceedings of the Seventeenth European Conference on Computer Systems*, pages 747–764, 2022.

[39] Anuj Kalia, Michael Kaminsky, and David Andersen. Datacenter RPCs can be general and fast. In *16th USENIX Symposium on Networked Systems Design and Implementation*, NSDI, 2019.

[40] Melanie Kambadur and Martha A Kim. Nrg-loops: adjusting power from within applications. In *Proceedings of the 2016 International Symposium on Code Generation and Optimization*, pages 206–215, 2016.

[41] Ajaykrishna Karthikeyan, Nagarajan Natarajan, Gagan Somashekar, Lei Zhao, Ranjita Bhagwan, Rodrigo Fonseca, Tatiana Racheva, and Yogesh Bansal. SelfTune: Tuning cluster managers. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 1097–1114, 2023.

[42] Antoine Kaufmann, Tim Stamler, Simon Peter, Naveen Kr. Sharma, Arvind Krishnamurthy, and Thomas Anderson. TAS: TCP acceleration as an OS service. In *14th ACM European Conference on Computer Systems*, EuroSys, 2019.

[43] Eddie Kohler, Robert Morris, and Benjie Chen. Programming language optimizations for modular router configurations. In *Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS X, pages 251–263, New York, NY, USA, 2002. Association for Computing Machinery.

[44] Eddie Kohler, Robert Morris, Benjie Chen, John Jannotti, and M. Frans Kaashoek. The Click modular router. *ACM Transactions on Computer Systems*, 18(3):263–297, August 2000.

[45] Chris Lattner and Vikram Adve. LLVM: A compilation framework for lifelong program analysis & transformation. In *9th International Symposium on Code Generation and Optimization*, CGO, 2004.

[46] Wen-Chuan Lee, Yingqi Liu, Peng Liu, Shiqing Ma, Hongjun Choi, Xiangyu Zhang, and Rajiv Gupta. White-box program tuning. In *2019 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 122–135. IEEE, 2019.

[47] David Leopoldseder, Lukas Stadler, Thomas Würthinger, Josef Eisl, Doug Simon, and Hanspeter Mössenböck. Dominance-based duplication simulation (dbds): code duplication to enable compiler optimizations. In *Proceedings of the 2018 International Symposium on Code Generation and Optimization*, pages 126–137, 2018.

[48] Zhao Lucis Li, Chieh-Jan Mike Liang, Wenjia He, Lianjie Zhu, Wenjun Dai, Jin Jiang, and Guangzhong Sun. Metis: Robustly tuning tail latencies of cloud systems. In *2018 USENIX Annual Technical Conference (USENIX ATC 18)*, pages 981–992, 2018.

[49] Caio Lima, Junio Cezar, Guilherme Vieira Leobas, Erven Rohou, and Fernando Magno Quintão Pereira. Guided just-in-time specialization. *Science of Computer Programming*, 185:102318, 2020.

[50] Chen Lin, Junqing Zhuang, Jiadong Feng, Hui Li, Xuanhe Zhou, and Guoliang Li. Adaptive code learning for spark configuration tuning. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 1995–2007. IEEE, 2022.

[51] Ashraf Mahgoub, Alexander Michaelson Medoff, Rakesh Kumar, Subrata Mitra, Ana Klimovic, Somali Chaterji, and Saurabh Bagchi. OPTIMUSCLOUD: Heterogeneous configuration optimization for distributed databases in the cloud. In *2020 USENIX Annual Technical Conference (USENIX ATC 20)*, pages 189–203, 2020.

[52] Jason Mccandless and David Gregg. Compiler techniques to improve dynamic branch prediction for indirect jump and call instructions. *ACM Transactions on Architecture and Code Optimization (TACO)*, 8(4):1–20, 2012.

[53] Sebastiano Miano, Alireza Sanaee, Fulvio Risso, Gábor Rétvári, and Gianni Antichi. Domain specific run time optimization for software data planes. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 1148–1164, 2022.

[54] Subrata Mitra, Manish K Gupta, Sasa Misailovic, and Saurabh Bagchi. Phase-aware optimization in approximate computing. In *2017 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 185–196. IEEE, 2017.

[55] László Molnár, Gergely Pongrácz, Gábor Enyedi, Zoltán Lajos Kis, Levente Csikor, Ferenc Juhász, Attila Kőrösi, and Gábor Rétvári. Dataplane specialization for high-performance openflow software switching. In *Proceedings of the 2016 ACM SIGCOMM Conference*, pages 539–552, 2016.

[56] Sourena Naser Moghaddasi, Haris Smajlović, Ariya Shajii, and Ibrahim Numanagić. Vectron: A dynamic programming auto-vectorization framework. In *Proceedings of the 23rd ACM/IEEE International Symposium on Code Generation and Optimization*, pages 644–659, 2025.

[57] Guilherme Ottoni. Hhvm jit: A profile-guided, region-based compiler for php and hack. In *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 151–165, 2018.

[58] Haolin Pan, Jinyuan Dong, Mingjie Xing, and Yanjun Wu. Synergy-guided compiler auto-tuning of nested llvm pass pipelines. *arXiv preprint arXiv:2510.13184*, 2025.

[59] Haolin Pan, Hongbin Zhang, Mingjie Xing, and Yanjun Wu. A hybrid, knowledge-guided evolutionary framework for personalized compiler auto-tuning. *arXiv preprint arXiv:2510.14292*, 2025.

[60] Heng Pan, Peng He, Zhenyu Li, Pan Zhang, Junjie Wan, Yuhao Zhou, XiongChun Duan, Yu Zhang, and Gaogang Xie. Hoda: a high-performance open vswitch dataplane with multiple specialized data paths. In *Proceedings of the Nineteenth European Conference on Computer Systems*, pages 82–98, 2024.

[61] Maksim Panchenko, Rafael Auler, Bill Nell, and Guilherme Ottoni. Bolt: a practical binary optimizer for data centers and beyond. In *2019 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 2–14. IEEE, 2019.

[62] Sunghyun Park, Salar Latifi, Yongjun Park, Armand Behroozi, Byungsoo Jeon, and Scott Mahlke. Srtuner: Effective compiler optimization customization by exposing synergistic relations. In *2022 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 118–130. IEEE, 2022.

[63] Gabriel Poesia and Fernando Magno Quintão Pereira. Dynamic dispatch of context-sensitive optimizations. *Proceedings of the ACM on Programming Languages*, 4(OOPSLA):1–28, 2020.

[64] Mihail Popov, Chadi Akel, Yohan Chatelain, William Jalby, and Pablo de Oliveira Castro. Piecewise holistic autotuning of parallel programs with cere. *Concurrency and Computation: Practice and Experience*, 29(15):e4190, 2017.

[65] Aleksandar Prokopec, Gilles Duboscq, David Leopoldseder, and Thomas Würthinger. An optimization-driven incremental inline substitution algorithm for just-in-time compilers. In *2019 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 164–179. IEEE, 2019.

[66] Mindaugas Rasiuckevicius. Longest prefix match (lpm) library. Accessed June 2025 from https://github.com/rmind/liblpm, 2022.

[67] Probir Roy and Xu Liu. Structslim: A lightweight profiler to guide structure splitting. In *Proceedings of the 2016 International Symposium on Code Generation and Optimization*, pages 36–46, 2016.

[68] Fabian Ruffy, Zhanghan Wang, Gianni Antichi, Aurojit Panda, and Anirudh Sivaraman. Incremental specialization of network programs. In *Proceedings of the 23rd ACM Workshop on Hot Topics in Networks*, pages 264–272, 2024.

[69] Silvius Rus, Lawrence Rauchwerger, and Jay Hoeflinger. Hybrid analysis: static & dynamic memory reference analysis. In *Proceedings of the 16th international conference on Supercomputing*, pages 274–284, 2002.

[70] Krzysztof Rzadca, Pawel Findeisen, Jacek Swiderski, Przemyslaw Zych, Przemyslaw Broniek, Jarek Kusmierek, Pawel Nowak, Beata Strack, Piotr Witusowski, Steven Hand, et al. Autopilot: workload autoscaling at google. In *Proceedings of the Fifteenth European Conference on Computer Systems*, pages 1–16, 2020.

[71] Marija Selakovic, Thomas Glaser, and Michael Pradel. An actionable performance profiler for optimizing the order of evaluations. In *Proceedings of the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 170–180, 2017.

[72] Hashim Sharif, Muhammad Abubakar, Ashish Gehani, and Fareed Zaffar. Trimmer: application specialization for code debloating. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, pages 329–339, 2018.

[73] Savvas Sioutas, Sander Stuijk, Henk Corporaal, Twan Basten, and Lou Somers. Loop transformations leveraging hardware prefetching. In *Proceedings of the 2018 International Symposium on Code Generation and Optimization*, pages 254–264, 2018.

[74] Gagan Somashekar, Amoghavarsha Suresh, Saurabh Tyagi, Vikas Dhyani, K Donkada, Anurag Pradhan, and Anshul Gandhi. Reducing the tail latency of microservices applications via optimal configuration tuning. In *2022 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, pages 111–120. IEEE, 2022.

[75] Gagan Somashekar, Karan Tandon, Anush Kini, Chieh-Chun Chang, Petr Husak, Ranjita Bhagwan, Mayukh Das, Anshul Gandhi, and Nagarajan Natarajan. OPPerTune: Post-deployment configuration tuning of services made easy. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 1101–1120, 2024.

[76] Jithendra Srinivas, Wei Ding, and Mahmut Kandemir. Reactive tiling. In *2015 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 91–102. IEEE, 2015.

[77] Akshitha Sriraman, Abhishek Dhanotia, and Thomas F Wenisch. Softsku: Optimizing server architectures for microservice diversity@ scale. In *Proceedings of the 46th International Symposium on Computer Architecture*, pages 513–526, 2019.

[78] Akshitha Sriraman and Thomas F Wenisch. μTune: Auto-tuned threading for OLDI microservices. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 177–194, 2018.

[79] Mark Stephenson and Ram Rangan. Pgz: automatic zero-value code specialization. In *Proceedings of the 30th ACM SIGPLAN International Conference on Compiler Construction*, pages 36–46, 2021.

[80] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

[81] SFX Thiago Teixeira, Corinne Ancourt, David Padua, and William Gropp. Locus: a system and a language for program optimization. In *2019 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pages 217–228. IEEE, 2019.

[82] Dana Van Aken, Andrew Pavlo, Geoffrey J Gordon, and Bohan Zhang. Automatic database management

system tuning through large-scale machine learning. In *Proceedings of the 2017 ACM international conference on management of data*, pages 1009–1024, 2017.

[83] Patrick Wintermeyer, Maria Apostolaki, Alexander Dietmüller, and Laurent Vanbever. P2go: P4 profile-guided optimizations. In *Proceedings of the 19th ACM Workshop on Hot Topics in Networks*, pages 146–152, 2020.

[84] Michael Wolfe. Iteration space tiling for memory hierarchies. In *Proceedings of the Third SIAM Conference on Parallel Processing for Scientific Computing*, pages 357–361, 1987.

[85] Michael Wolfe. More iteration space tiling. In *Proceedings of the 1989 ACM/IEEE conference on Supercomputing*, pages 655–664, 1989.

[86] Shaofeng Wu, Qiang Su, Zhixiong Niu, and Hong Xu. Tomur: Traffic-aware performance prediction of onnic network functions with multi-resource contention. *arXiv preprint arXiv:2405.05529*, 2024.

[87] Minghui Yang, Gang-Ryung Uh, and David B Whalley. Efficient and effective branch reordering using profile data. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 24(6):667–697, 2002.

[88] Arseniy Zaostrovnykh, Solal Pirelli, Rishabh Iyer, Matteo Rizzo, Luis Pedrosa, Katerina Argyraki, and George Candea. Verifying software network functions with no verification expertise. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, pages 275–290, 2019.

[89] Yuxuan Zhang, Nathan Sobotka, Soyoon Park, Saba Jamilan, Tanvir Ahmed Khan, Baris Kasikci, Gilles A Pokam, Heiner Litz, and Joseph Devietti. Rpg2: Robust profile-guided runtime prefetch generation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 999–1013, 2024.

[90] Peng Zhao and José Nelson Amaral. Feedback-directed switch-case statement optimization. In *2005 International Conference on Parallel Processing Workshops (ICPPW'05)*, pages 295–302. IEEE, 2005.