MetaEase: Heuristic Analysis from Source Code via Symbolic-Guided Optimization

Abstract

Large-scale systems rely on heuristics to tackle NP-hard problems such as traffic engineering, virtual machine placement, and packet scheduling. While these heuristics are efficient, they can exhibit severe performance gaps under certain workloads, leading to outages or costly over-provisioning. This risk has motivated tools that attempt to identify inputs causing worst-case underperformance. However, using these tools in practice often requires rewriting heuristics as formal mathematical models—a process that is time-consuming, error-prone, and excludes many real-world algorithms.

We introduce MetaEase, a practical general-domain analyzer that works by directly analyzing a heuristic's source code, eliminating the need for formal modeling. MetaEase combines code-aware input generation with guided search to uncover worst-case scenarios efficiently, even for heuristics with randomness (e.g., various traffic engineering schemes) or non-convex behavior (e.g., bin packing for virtual machine placement).

Across five problem domains and eight heuristics, MetaEase matched or exceeded MetaOpt, a state-of-the-art optimization-based heuristic analyzer, in most cases; in the remainder, it achieved at least 85–98% of its performance and often ran faster. Against black-box optimization baselines, it won in 88% of settings and ranked in the top two otherwise. MetaEase analyzed Arrow [85], a widely studied networking heuristic, which cannot be analyzed by any of the state-of-the-art heuristic analyzers. We revealed previously unknown performance gaps in Arrow.

1 Introduction

Network operators routinely face challenging optimization problems where it is infeasible to compute an exact solution (NP-hard or otherwise computationally expensive) [3, 5, 9, 24, 33, 47, 48, 52, 53, 64, 67, 72, 78, 81–84]. For practical reasons, most systems rely on heuristics—non-optimal algorithms that trade accuracy for speed and scalability. Examples include "best-fit" or "first-fit" algorithms for VM placement [15], demand pinning for traffic engineering [52], and approximate algorithms for packet scheduling [7–9, 80]. However, heuristics can have severe failure modes that degrade performance and impact user experience [6, 9, 12, 52]. For instance, Microsoft previously employed a traffic engineering heuristic in its wide-area network that, for certain inputs, underperforms the optimal solution by 30% [52]. To mitigate this impact, they must either over-provision the network by 30%, drop 30% of the demand, or *detect such cases and switch to an alternative heuristic* [52].

Recent performance analyzers [6, 9, 12, 52] compare a heuristic's performance against an (often optimal) benchmark, and enable operators to anticipate and mitigate risks by identifying inputs that produce the largest performance gap. These tools reveal problematic inputs, guide heuristic refinement, and support input-aware switching among heuristics with complementary failure modes. They solve:

$$I^* = \arg\max_{I \in \mathbb{R}^n} \text{Benchmark}(I) - \text{Heuristic}(I) \qquad (1)$$

to find the input(s) I^* that maximize the performance gap between the heuristic and the benchmark for a specific instance of the problem.

For example, in traffic engineering, the input I represents traffic demand, while Benchmark(I) and Heuristic(I) denote the total flow routed by the optimal solution and the heuristic, respectively. The objective is to identify the demand I^* that minimizes the traffic routed by the heuristic relative to the optimal. Throughout this discussion, note that I (e.g., demand in traffic engineering) is an unspecified variable in this problem.

Existing performance analyzers that compute I^* have two main limitations. First, they are difficult to use. To apply them, we need to express the heuristic in specialized mathematical forms, such as a convex optimization model [52], feasibility or satisfiability constraints [6, 52], or a network flow model [40]. These tools demand substantial expertise in formal methods or optimization theory; it is hard to even model a simple traffic engineering heuristic (see §2). Any change to the heuristic also requires re-modeling.

Second, existing performance analyzers support a limited range of heuristics. MetaOpt [52] only analyzes heuristics that are expressed as convex optimization or a set of feasibility constraints. FPerf [9], CCAC [12], and Virelay [32] are custom-designed for specific domains (queue management, congestion control, and scheduling, respectively). Many common classes of heuristics, such as ML-based heuristics, fall outside their scope.

While most operators have little experience in formal analysis to model heuristics mathematically, we observe that *implementations* of heuristics are often readily available – after all, one must execute the heuristic to use it! Similarly, benchmark models are usually available and rarely pose a problem[1]. Motivated by this, our goal is to directly find the inputs that make a heuristic *implementation* underperform its benchmark as much as possible, with only access to its source code.

One approach to solving the problem is to treat the heuristic implementation as a black box. This approach is supported

[1] For any domain, we only need to model the benchmark once and can reuse it across heuristics.

Tool	No Heuristic Optimization/Modeling	Heuristic-guided Search	General Domain
MetaEase	✓	✓	✓
Black-box Search	✓	✗	✓
MetaOpt [52]	✗	✓	✓
Virelay [32]	✗	✓	✗
XPlain [40]	✗	✓	✓
FPerf [9]	✗	✓	✗

Table 1: We compare MetaEase with prior analyzers. Unlike black-box methods, MetaEase leverages heuristic structure to guide the search, and unlike model-based approaches, it does not require a mathematical model of the heuristic.

by off-the-shelf techniques such as sample-based gradient ascent [66], Bayesian optimization [18], hill climbing [26], and simulated annealing [41]. These methods sample inputs $I \in \mathbb{R}^n$, compute the performance gap for each (i.e., Benchmark(I)−Heuristic(I)), and iteratively move toward worse-performing regions. However, such a search quickly becomes impractical: for each sample, we must solve the benchmark — often an NP-hard or otherwise costly optimization — to evaluate the outcome and compute the next search direction. Worse, the search frequently gets stuck in local optima and requires multiple runs from different starting points. This approach scales poorly, especially on large problem instances (see Fig. 13 and §5.2).

We present MetaEase, a performance analyzer that enables operators to directly evaluate a heuristic *implementation* and overcomes the limitations of black-box search. MetaEase takes as input (1) the heuristic implementation (written in C) and (2) a benchmark, specified as an optimization model whose solution represents the best achievable performance for the problem. Operators may trade off speed and provide an implementation of the benchmark instead. MetaEase finds the input I^* that maximizes the performance gap between the benchmark and the heuristic. Operators can optionally impose constraints to restrict the search (e.g., to consider only likely inputs). Tab. 1 shows how MetaEase improves upon the state of the art. Unlike prior analyzers [6, 9, 40, 52], MetaEase does not require a mathematical model of the heuristic. It relies solely on a benchmark formulation—typically well studied and easy to specify for most problems. Crucially, users define the benchmark once per domain (e.g., multi-commodity flow for traffic engineering), while heuristics change frequently and vary widely. This separation enables MetaEase to analyze diverse heuristics without repeated modeling effort.

MetaEase addresses black-box search limitations: on the high-level MetaEase runs gradient ascent [2] search on Eq. 1, but unlike black-box approaches, it uses the heuristic's structure to guide its search and has a mechanism to reduce expensive benchmark calls. It exploits the linearity of differentiation in Eq. 1 and expresses the ascent direction as ∇Benchmark(I)−∇Heuristic(I), which allows it to compute the two terms separately. This gradient-based approach provides search progress for Eq. 1 without repeatedly solving the benchmark (§6.2).

For the heuristic term, we fit a smooth surrogate (a Gaussian process [70]) to Heuristic(I) within a small neighborhood of input I and differentiate it to estimate ∇Heuristic(I). As the search progresses, MetaEase updates the surrogate to approximate Heuristic(I) locally. We also show this surrogate-guided direction is more efficient than black-box approaches (Fig. 13).

MetaEase's second key idea is to use symbolic execution tools for static analysis of the heuristic's code to select initial seed points for the search. Black-box methods often start from random inputs, many of which fall into the same "equivalence class" where the heuristic behaves identically. Instead, MetaEase uses KLEE [19], a robust and well-known symbolic execution tool [14, 20], to extract inputs that exercise different code paths. We then run gradient ascent from each seed in parallel and report the maximum performance gap across all searches. Our results show that MetaEase finds gaps up to 44× larger than those from random initialization (§6.1).

To our knowledge, MetaEase is the first system that lets operators analyze their *existing implementations (as code)* of deployed heuristics and quantify when and by how much they underperform relative to a benchmark. MetaEase can analyze any heuristic implementation, and we show that it supports non-convex and probabilistic heuristics, which prior work in this space cannot analyze. We used MetaEase to analyze heuristics from five common problem domains in networking and distributed systems (vector bin packing, WAN traffic engineering, Arrow [85], knapsack, and maximum weight matching). MetaEase *matched or exceeded* the performance gap the state-of-the-art found (which in most cases solves the problem optimally) in 60% of our experiments, was within 2−15% for 30% of cases, and within ∼35% in the rest. We plan to open-source MetaEase upon publication for community use.

2 Motivation and MetaEase Overview

Our goal is to find inputs that maximize the performance gap between the heuristic and the benchmark (Eq. 1). Several tools already address this problem. We next explain how existing heuristic analyzers operate and illustrate their limitations through an example. In this discussion, we focus on tools that support a broad range of heuristics and exclude those limited to specific domains [9, 10, 12, 32] (we describe these works in detail in the related work).

2.1 How existing methods solve the problem

Heuristic analyzers. We know of three general-purpose heuristic analyzers: MetaOpt [52], Vireley [6], and XPlain [40]. MetaOpt solves a bi-level optimization problem and takes as input a mathematical model of the heuristic (and the benchmark), provided either as an optimization problem or as a set of constraints — it supports convex or feasibility heuristics. XPlain [40] lets users model heuristics in a network-flow-based domain-specific language and compiles the model into an

[2] Gradient ascent iteratively updates $x \leftarrow x + \eta \nabla f(x)$ to maximize f.

```
Demands SelectTop20Percent(Demands D) {
    sort_desc(D);
    int k = max(1, 0.2 * size(D));
    return top_k(D, k);
}
```

(a) Expressing top-20% selection in code

Input: $\mathcal{D} = \{d_k\}$

1: $\tau \leftarrow \lfloor 0.2|\mathcal{D}| \rfloor$

2: $r_k \leftarrow \texttt{Rank}\left(d_k, [d_i]_{i \in \mathcal{D}}\right)$

3: $c_k \leftarrow \texttt{IsLeq}(r_k, \tau) \quad (c_k \in \{0,1\})$

4: $\sum_k c_k = \tau$

5: $d_k^{\text{crit}} \leftarrow \texttt{Multiplication}(c_k, d_k)$

6: $d_k^{\text{non}} \leftarrow \texttt{Multiplication}(1 - c_k, d_k)$

(b) Expressing top-20% selection in MetaOpt using its helper functions

Figure 1: How we can express the "top-20% of demands" as (a) code, and (b) in MetaOpt. MetaOpt's encoding requires much more expertise and optimization background.

optimization problem that MetaOpt supports. Vireley [6] uses formal methods: it encodes the heuristic's performance as a set of satisfiability constraints (a generalization of feasibility constraints) and uses an SMT-solver (Z3 [27]) to find instances where the heuristic underperforms.

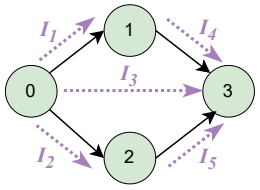
Black-box methods. Black-box methods (e.g., multi-start hill climbing [26], simulated annealing [41], sample-based gradient [17]) treat both benchmark and heuristic as black-box. They do not require a mathematical model but are often sample-inefficient, seed-sensitive, and often stall in local optima. We discuss the challenges of these approaches in §2.3.

2.2 Why existing heuristic analyzers are hard to use

Consider a simple traffic engineering heuristic that operates as follows (Code 2): (1) route the largest 20% of demands optimally, ignoring the rest; (2) freeze these allocations and construct a residual graph by adjusting capacities based on the previous step; and (3) allocate the remaining demands optimally on this residual graph. The goal of this heuristic is to reduce the number of variables in the traffic engineering optimization problem, which enables faster solutions.

One might expect this heuristic is straightforward to analyze using existing tools, since MetaOpt and Xplain support the underlying convex traffic engineering optimization problem. The heuristic essentially solves two optimization instances: first for the largest 20% of demands, then for the remainder. Indeed, MetaOpt and XPlain model a similar heuristic called "demand pinning," which fixes small demands to the shortest paths before it routes the rest.

It is trivial to select the top 20% of demands in code (Fig. 1a). In contrast, it is hard to do so in MetaOpt because demands are variables within the verifier (the input vector $\boldsymbol{I}$ in Eq. 1), which



(a) Simple topology.

Seed	1	2	3	4	5
Initial	$\begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{bmatrix}$ $\begin{bmatrix} 70 \\ 270 \\ 130 \\ 180 \\ 240 \end{bmatrix}$	$\begin{bmatrix} 20 \\ 280 \\ 340 \\ 4 \\ 60 \end{bmatrix}$	$\begin{bmatrix} 170 \\ 340 \\ 60 \\ 360 \\ 140 \end{bmatrix}$	$\begin{bmatrix} 230 \\ 380 \\ 280 \\ 250 \\ 320 \end{bmatrix}$	$* \begin{bmatrix} 1 \\ 1 \\ 200 \\ 1 \\ 1 \end{bmatrix}$
Final Gap	0.0	0.0	0.0	0.0	**398.2**
Time (min)	20	20	20	20	**5**

(b) Hill-Climbing results.

Figure 2: Black-box approaches struggle to find large performance gaps. Hill-Climbing stalls under random initialization, yielding small gaps and long runtimes, and succeeds only with carefully selected inputs (*).

makes it impossible to order them explicitly during problem formulation. To find the top 20%, we must encode the sorting process itself—a task that is difficult to model. We show how to achieve this in MetaOpt (Fig. 1b). Encoding sorting is inherently complex (the exact details are not critical here). A further complication is that the two optimizations must execute sequentially, which is difficult to express (see App. A for details). These challenges are not unique to our setting; similar issues arise in max–min fair traffic engineering methods [53], which also require explicit modeling of sorting-based allocations.

This example shows why it is hard to model even a simple heuristic as an optimization. The MetaOpt authors employ big-M techniques and methods to "convexify" optimizations for such cases. These techniques are difficult for non-experts to devise and apply. Similar challenges occur when one expresses the heuristic as SMT constraints in Virelay.

2.3 Why Black-Box search algorithms do not work

A naive approach is to treat the problem as a black-box search: we run both the benchmark and the heuristic on selected input instances and navigate the input space to find larger performance gaps. But the benchmark/optimal solution for most problems we study is slow—heuristics often approximate NP-hard problems—so each instance is expensive to evaluate. This approach also overlooks information in the heuristic's source code and the benchmark's mathematical representation.

Consider a highly simplified traffic engineering (TE) example (Fig. 2). We ran a multi-start hill-climbing [26] search. With random demand initializations, the search stalls (Fig. 2) and fails to uncover a meaningful performance gap. Only after we manually inject a specially crafted input (marked as * in the table) does it discover a meaningful gap, and even then each instance takes minutes to evaluate.

These issues are common in black-box algorithms: they often converge to poor local optima, are highly sensitive to initialization, and run slowly (see also Fig. 7 and §5).

2.4 MetaEase overview

MetaEase strikes a balance between black-box methods and model-based analyzers. Similar to black-box methods, it does

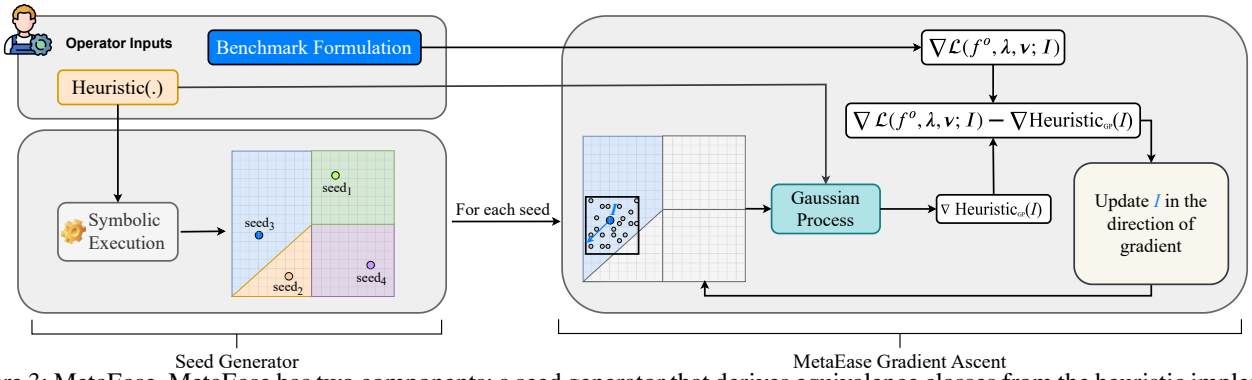


Figure 3: MetaEase. MetaEase has two components: a seed generator that derives equivalence classes from the heuristic implementation through symbolic execution, and a gradient ascent process that uses the benchmark formulation to compute the gradient of the Lagrangian and the heuristic code for an approximate gradient to guide the search.

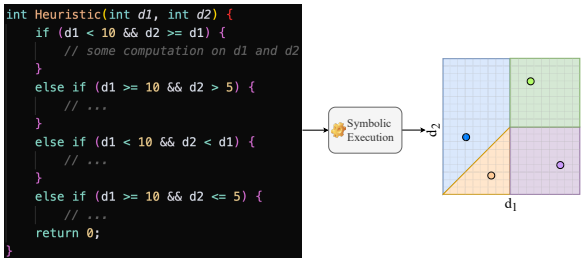


Figure 4: KLEE [19] symbolically explores the heuristic code and generates inputs for distinct *code paths*. Each path represents an equivalence class of inputs sharing the same control flow; here, KLEE produces one input per branch condition.

not require a mathematical formulation of the heuristic, but it uses the heuristic's code structure to guide the search. We show MetaEase's high-level design in Fig. 3.

MetaEase enables operators to analyze any heuristic directly from its *code*. It requires a mathematical model only for the benchmark. This burden is minimal because we need to model each problem domain once and MetaEase already models many. These benchmarks are often fixed, canonical, reusable, and well-optimized. For example, Arrow [85] models the optimal cross-layer Optical–IP problem and MetaOpt [52] models the optimal vector bin-packing and packet scheduling solutions. Users may provide a benchmark implementation, though this may reduce efficiency.

MetaEase finds inputs that cause the heuristic to underperform relative to the benchmark. It (1) uses the linear separability of the pefromance gap objective (Eq. 1) to do gradient ascent, and (2) smartly generates seeds from static analysis of the heuristic's code to improve search efficiency.

(1) Gradient ascent: The Gradient Ascent Module (Fig. 3) exploits the separability of $\text{Gap}(I) = \text{Benchmark}(I) - \text{Heuristic}(I)$ to compute gradients for the benchmark and heuristic independently. When the benchmark is provided as an optimization formulation, we derive the Lagrangian (§3.1) and compute its gradient directly, avoiding repeated (costly) optimal runs. This approach scales better and outperforms black-box methods (see §5.2 and §6.2). If the optimization formulation is unavailable, MetaEase approximates the benchmark gradient using the same mechanism as for the heuristic.

To estimate the heuristic's gradient, we adopt a surrogate-based approach [23, 29, 31, 37, 65], suited for unknown or non-differentiable objectives. At each step, MetaEase fits a Gaussian process to a local region and uses it to compute the heuristic's gradient (Fig. 5, §3.2). When gradient ascent moves beyond this region, we retrain the process. Gaussian processes approximate any function with sufficient data [39, 49] and provide closed-form gradients [49].

(2) Seed generation: The seed generator module (Fig. 3) mitigates poor local optima. Unlike black-box methods that rely on random restarts, we symbolically analyze the heuristic's code using KLEE [19]. KLEE explores branch structures (e.g., if/else conditions) and returns concrete inputs that traverse distinct execution paths (Fig. 4). Each path defines an input region where the heuristic follows the same control flow, forming *equivalence classes* of distinct behaviors.

MetaEase performs "path-aware" gradient ascent within each equivalence class and reports the maximum performance gap across classes. By treating path conditions as boundaries, this approach avoids crossing non-differentiable regions of the heuristic's code (§3.4).

3 MetaEase Approach

MetaEase leverages the heuristic implementation to guide the search, which allows it to handle non-convex and non-differentiable heuristics—cases that prior methods cannot address. In this section, we show how MetaEase overcomes the key challenges outlined earlier. MetaEase:

- does not solve the costly benchmark repeatedly when it maximizes the performance gap (§3.1);
- uses a local surrogate model to estimate the gradients of arbitrary heuristics (§3.2);
- selects informed seeds for gradient-based search to prevent poor local optima (§3.3); and
- ensures stable progress even in the presence of non-differentiabilities (§3.4).

Some of our ideas are inherently novel (e.g., symbolic execution and path-based gradients to identify equivalence classes and help avoid issues with non-differentiable regions in the input space). Others are well-known in the optimization literature [17, 70], but we integrate them into MetaEase in way that allows it to function as a general-purpose heuristic analyzer. We aim to maximize Gap(I) without repeatedly solving the costly benchmark. On the high-level, our idea is to use gradient ascent. Since gradients are linear, we decompose $\nabla\big(\text{Benchmark}(I) - \text{Heuristic}(I)\big)$ as $\nabla\text{Benchmark}(I) - \nabla\text{Heuristic}(I)$, and estimate each term separately. We start with the benchmark.

3.1 How MetaEase estimates ∇Benchmark(I)

We use an established Lagrangian-based method to estimate the gradient [17, 25, 50]. We describe the technique below.

Benchmark as an optimization. The benchmark is itself, in general, an optimization:

$$\text{Benchmark}(I) = \max_{x^o} f_{\text{B}}(x^o; I) \tag{2}$$
$$\text{s.t. } g_i(x^o; I) \geq 0,\ i = 1,\ldots,m, \quad h_j(x^o; I) = 0,\ j = 1,\ldots,p.$$

Here, x^o are the benchmark's decision variables (e.g., how much flow to allocate on each path in traffic engineering), g_i are the m inequality constraints (e.g., the capacity constraints), and h_j are the p equality constraints (e.g., flow conservation). The *variables* I are inputs to the benchmark problem (e.g., the demand matrix). Intuitively, the benchmark defines the optimal outcome for a given input I—for example, in traffic engineering, it computes the maximum total flow that we can route.

Step 1. Lagrangian. The gradient ∇Benchmark(I) captures the change in the benchmark's *optimal value* under small perturbations of I. We would have to re-solve the benchmark for each perturbation of I to directly compute it, which is costly. Instead, we use the benchmark's *Lagrangian*, which combines the objective and the constraints:

$$\mathcal{L}(x^o, \lambda, \nu; I) \triangleq f_{\text{B}}(x^o; I) + \sum_{i=1}^{m} \lambda_i g_i(x^o; I) + \sum_{j=1}^{p} \nu_j h_j(x^o; I) \tag{3}$$

Here, $\lambda_i \geq 0$ are dual variables for the m inequality constraints g_i, and ν_j are dual variables for the p equality constraints h_j. We can approximate how the benchmark's optimal objective changes with the change of I when we analyze the gradients of $\mathcal{L}$ with respect to these variables. This allows us to estimate ∇Benchmark(I) without repeatedly solving the full optimization.

Step 2. Duality. The Lagrangian approximates the benchmark's optimal value. By *weak duality* [17], maximizing the Lagrangian over the benchmark's decision variables x^o yields an upper bound on the true benchmark value for input I:

$$\text{Benchmark}(I) \leq \sup_{x^o} \mathcal{L}(x^o, \lambda, \nu; I), \quad \lambda \geq 0.$$

Here sup (supremum) simply means the largest value of $\mathcal{L}$ over all possible choices of x^o.

When the benchmark is convex, *strong duality* [17] tells us something stronger: we can exactly recover the benchmark value if we optimize in two stages— we first take the maximum over x^o and then the minimum over the dual variables λ, ν:

$$\text{Benchmark}(I) = \min_{\nu, \lambda \geq 0} \sup_{x^o} \mathcal{L}(x^o, \lambda, \nu; I).$$

This "max–min" structure is called a *saddle-point*: one set of variables (x^o) aims to maximize the objective, while the dual variables aim to minimize it. For clarity, we denote ascent variables in green and descent variables in red.

Step 3. Compute the gradient. Our goal is to estimate ∇Benchmark(I)—the sensitivity of the benchmark's optimal value to changes in I. The saddle-point form is useful because the Lagrangian already encodes how both the objective and the constraints respond to I. Empirically, even for non-convex benchmarks, following the gradients of the saddle-point expression provides reliable directions for ∇Benchmark(I).

We treat the saddle-point as a joint optimization and update each variable according to its role:

- maximize over variables x^o (the benchmark's decisions),
- minimize over the dual variables λ, ν (the penalties on constraints), and
- ascend in the input I (the parameter we care about).

In practice, we take ascend for x^o and I, and descend for λ and ν. We approximate ∇Benchmark(I) through these coordinated updates, which allows us to not re-solve the benchmark optimization in each step.

When a benchmark's optimization model is unavailable, operators can provide its implementation. In this case, to compute the benchmark's gradient, we use the same surrogate-based method as for the heuristic. We describe how we estimate ∇Heuristic(I) next.

3.2 How MetaEase estimates ∇Heuristic(I)

Now that we can estimate ∇Benchmark(I), we also need ∇Heuristic(I). Unlike the benchmark, we do not have a mathematical formulation of the heuristic, only its implementation. We therefore use a surrogate-based approach [23, 29, 31, 37, 65]. The idea is to build a *local surrogate* around the current input $I \in \mathbb{R}^n$ and use the surrogate's gradient as an estimate of the heuristic's gradient (Fig. 5).

We define a hypercube (the Δ-neighborhood) centered at I with side length $\Delta > 0$. We draw N samples from this hypercube, run the heuristic implementation on each sample, and fit a Gaussian process (GP) model to the results. We then use the GP's closed-form gradient to approximate ∇Heuristic(I) and update I along that direction. As the search progresses, MetaEase refits the GP whenever I leaves its current hypercube to ensure the surrogate reflects local behavior.

Gaussian processes are a natural fit because they provide smooth, differentiable surrogates with analytically computable gradients [70] (App. B). Intuitively, the GP fits a smooth curve through the heuristic's local outputs, which lets us differentiate

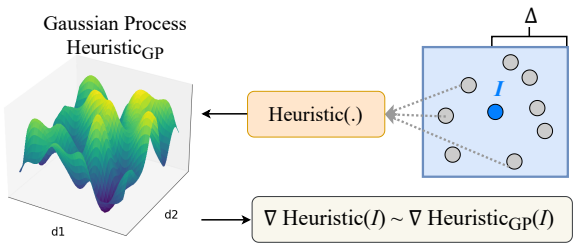


Figure 5: MetaEase fits a local Gaussian-process (GP) surrogate to heuristic evaluations near the current input and uses GP's analytical gradient to estimate gradient of heuristic.

the surrogate rather than the heuristic itself. This yields efficient gradient estimates and reduces runtime. Compared with finite-difference methods, which require evaluating the heuristic along all n coordinate directions[3], MetaEase requires far fewer heuristic evaluations (§6.3).

3.3 How MetaEase avoids bad local optima

When an optimization problem is nonconvex, gradient-based methods can get trapped in local optima [17]. Tools such as MetaOpt largely avoid this issue by using mixed-integer solvers to solve the optimization directly. Consequently, MetaOpt supports only convex heuristics or those expressible as feasibility problems. In contrast, MetaEase accepts *any* heuristic implementation, which may be nonconvex or otherwise ill-behaved. Therefore, we need a strategy for cases where gradient-based search converges to a poor local optimum.

A common mitigation is random multi-start, which restarts the search from multiple random initial seeds [41]. Although this approach improves coverage of the input space, it does not guarantee that we find an optimal or near-optimal solution. As problem size increases, the number of starting seeds we need grows substantially [41].

MetaEase uses the heuristic code to find search seeds. To increase the likelihood that MetaEase finds larger gaps than random seeding, we use the heuristic's *implementation*. We identify *equivalence classes* of inputs that trigger the same behavior in the heuristic and use a representative from each class as seeds. We then run the gradient search we described earlier within each class and retain the best gap across classes.

MetaEase uses KLEE to find equivalence classes. We use KLEE [19], which is a mature, robust, and well-known symbolic execution tool [2, 14, 19, 20], to symbolically execute the heuristic and produce representative inputs for distinct code-paths. [4] Symbolic execution is a well-studied concept in formal methods that is used to automatically derive test cases from a target program's source code. For MetaEase, the test inputs serve as the seed values for its search. In most cases, each code-path corresponds to a set of distinct "choices" the heuristic makes on a given input. This means the heuristic treats inputs that follow the same code-path in the same way — the property we seek in an equivalence class. KLEE returns a set of input seeds $I \in$ KLEE(Heuristic) where each input has a different *code-path signature*, h. We run gradient ascent in parallel across multiple equivalence classes starting from KLEE points. In some cases, KLEE discovers only a single code-path for a heuristic. This indicates that symbolic execution has not exposed any meaningful branching behavior; in such cases, we revert to random multi-start.

3.4 How MetaEase avoids heuristic non-differentiability

Many heuristics contain conditional logic. For example, the demand pinning heuristic studied in MetaOpt pins small demands before optimally routing the remainder; the heuristic in §2 sorts demands and routes the top 20% first; and the first-fit decreasing heuristic for bin-packing sorts items and then places them from smallest to largest. Such branching decisions create highly non-convex and non-differentiable regions in the gap function. In these regions, the gradient can assume extreme values, causing gradient ascent to repeatedly overshoot or undershoot the true optimum (§6.2).

To mitigate this, MetaEase restricts its search *within* an equivalence class of inputs, i.e., inputs that follow the same code path in the heuristic. Empirically, we find that within each equivalence class the heuristic is much smoother, making gradient ascent more stable and effective. While this is not guaranteed to always find the global optimum, in practice it consistently leads to higher gaps.

Fig. 6 and Algo. 1 illustrate how MetaEase performs a *path-aware* gradient step. After computing the benchmark gradient (§3.1) and the heuristic gradient using the GP surrogate (§3.2), MetaEase first attempts a regular gradient ascent step. If the step remains in the same code path, it is taken directly. If instead it crosses into a new path, MetaEase projects the update back into the current equivalence class: it selects from the GP training samples the point whose direction is most aligned with the gradient and moves there instead.

This path-aware technique keeps the search within a single equivalence class, avoiding large jumps across non-differentiable regions. In practice, this allows MetaEase to explore both sides of non-differentiable boundaries separately and reliably uncover higher performance gaps (§6.2). For further details, see App. C.

4 Optional Optimizations

MetaEase can analyze any heuristic implementation that operates on numerical inputs. In this section, we discuss optional optimizations that improve MetaEase's runtime and memory efficiency for larger problem instances. Users can disable these optimizations to run the vanilla MetaEase implementation, which still analyzes the heuristic and computes the worst-case gap, though with higher time and memory costs. We evaluate

[3] Finite differences estimate $\nabla f(I)$ by computing $f(I)$ and $f(I+\delta e_j)$ for each dimension $j = 1,\ldots,n$.

[4] For clarity, we use the terms code-path and equivalence class interchangeably in the remainder of this paper.

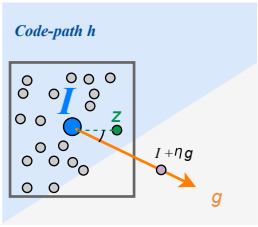


Figure 6: One step of gradient ascent in MetaEase. We move along the estimated gap gradient ($\eta \cdot g$); if the step remains within the same code path, we take it; otherwise, we project to a nearby point z that preserves the path. See the alg. below.

Algorithm 1: Path-aware gradient step in MetaEase

Input: Block size Δ, samples N, step size η

$h \leftarrow \text{path}(I)$ // current code path
$\mathbf{X} \leftarrow \{u \in \text{SAMPLEBOX}(I, \Delta) \mid \text{path}(u) = h\}$
$\mathbf{Y} \leftarrow [\text{Heuristic}(x) \mid x \in \mathbf{X}]$
$\text{Heuristic}_{GP} \leftarrow \text{FITGP}(\mathbf{X}, \mathbf{Y})$
Compute $g_b \leftarrow \nabla\text{Benchmark}(I)$
Compute $g_h \leftarrow \nabla\text{Heuristic}_{GP}(I)$
$g \leftarrow g_b - g_h$
$\tilde{I} \leftarrow I + \eta \cdot g$
if $\text{path}(\tilde{I}) = h$ **then**
  $I \leftarrow \tilde{I}$ // stay in same path
else
  Pick $z \in \mathbf{X}$ most aligned with g
  $I \leftarrow z$ // project back into path

the benefits of these optimizations in §6.

4.1 Projected gradients

When we study large problem instances, we have to look at higher dimensional inputs. This increases the runtime of two MetaEase components: (1) KLEE [19] may take longer to run; (2) we may need more samples to train the Gaussian Process.

To maintain low runtime despite these increases, we use the technique of *projected gradients* [21]. We select a subset of K input dimensions and *freeze* the remaining dimensions. We then apply MetaEase in the subspace defined by the selected K dimensions. Afterward, we unfreeze these values and choose a new subset of dimensions to continue the search. This approach mitigates the slowdowns in (1) and (2) but may affect the quality of the resulting solution.

Operators can use this feature to quickly estimate the heuristic's worst-case performance. We discuss runtime optimizations combined with this approach in App. E.

4.2 Domain customization

As in most verification problems, domain knowledge improves runtime and, in our case, the quality of the identified performance gap. MetaEase allows operators to input additional *hints* to MetaEase based on their expertise.

Hints are lightweight constraints or predicates that MetaEase passes to KLEE, which uses them to bias the selection of representative inputs for each equivalence class. These hints leave the MetaEase pipeline unchanged—they do not affect the objective function or the search algorithm—but they influence the starting points for the search process.

Operators can specify two categories of hints:
Range constraints: intervals that restrict the input domain;
Hardness predicates: conditions over the input space that operators believe cause the heuristic to underperform.

Consider our earlier example: operators know that demands stressing certain edge capacities in the graph are critical—if no demand overloads an edge, we can satisfy all demands without difficulty. Operators can encode this condition as a predicate in MetaEase (e.g., Code 1).

Code 1: Example of domain-specific assumption in TE

```
for (int k = 0; k < K; ++k) {
  klee_assume(0 <= demand[k] && demand[k] <= D_MAX);
} // Demand range constraints
for (int e = 0; e < num_edges; ++e) {
  int edge_load
     [e] = sum_demands_on_edge(demand, edges[e]);
  klee_assume(edge_load[e] > edges[e].Capacity);
} // Hardness predicates
```

5 Evaluation

MetaEase applies across domains such as wide-area and optical-IP cross-layer traffic engineering, knapsack problems, vector bin packing, and maximum weight matching. We evaluate MetaEase on deterministic (greedy, sort-based, and conditional), randomized, and DNN-based heuristics.

Our results show that MetaEase consistently ranks among the top two approaches. In 12 experiments, it matches the worst-case gap of MetaOpt, which computes the exact optimum when run to completion. In six additional experiments, it achieves at least 85–98% of MetaOpt's performance and outperforms all black-box methods in 22 out of 25 experiments.

5.1 Experiment setup

Implementation. We use *KLEE* [19] to compute heuristic equivalence classes and *OR-tools* [59] to solve benchmark optimization problems. We used the problem's optimal solution as a benchmark. To train Gaussian process surrogates for a given input, we uniformly sample $N = 100$ points from a hypercube centered on that input. We set the hypercube size to $\Delta = 1\%$ of the input space (§3.2)[5].

Baselines. We evaluate the performance gap identified by MetaEase relative to MetaOpt, the only open-source heuristic analyzer, for several heuristics (see Tab. 2). For heuristics that are difficult or impossible to model in MetaOpt, we compare against black-box methods, including random sampling, hill climbing [26], simulated annealing [41], and sample-based gradient ascent. We execute each black-box baseline 10 times with different random seeds under the same time budget as MetaEase and report the maximum gap observed across runs.

[5]We measured the variance of all problems and found $N = 100$ balances training time and Gaussian Process accuracy. See §6.3 for an example.

	Benchmark	Heuristic
TE	MaxFlow [44, 52]	DP [43] (conditional) POP [56] (random) Alg in §2 (sorting) DOTE [60] (DNN)
VBP	Optimal [77]	FFD [58]
Knapsack	Optimal [77]	Sort-based greedy [30]
Optical–IP TE	Optimal [85]	Arrow [85]
MWM	Optimal [77]	Greedy [77]

Table 2: Overview of the domains and heuristics we explored in this work. (TE: Traffic Engineering, DP: Demand Pinning, VBP: Vector Bin-packing, MWM: Maximum Weight Matching). Arrow is randomized and optimization-based.

Topology	**Δ Time (h)** (MetaOpt – MetaEase)	**Gap Ratio (%)** (MetaEase / MetaOpt)
Abilene	0.60	91.23%
B4	0.61	100.0%
Swan	0.46	100.0%
Uninett2010	2.23	63.4%
Cogentco	8.00	90.8%

Table 3: For Demand Pinning, MetaEase finds performance gaps comparable to MetaOpt, but in less time in most cases.

Metrics. We define the *normalized gap* as the performance difference between the benchmark and the heuristic, divided by the maximum value attainable by either on any input. For instance, in the case of traffic engineering, the normalized gap corresponds to the difference in total flow, divided by the sum of link capacities in the network. For each method, we report the *maximum normalized gap* they discovered. We also track runtime. All experiments run on an Intel Xeon E5-2630 v4 CPU @ 2.20 GHz with 24 cores and 23 GB of memory, utilizing all available threads. See App. F for detailed MetaEase runtimes.

5.2 MetaEase can analyze *any* numerical heuristic

We demonstrate the generality of MetaEase by comparing it against MetaOpt across several types of traffic engineering heuristics (Tab. 2): Demand Pinning, which is conditional; POP, which involves randomness; and the heuristic in §2, which relies on sorting. We also evaluate DOTE, a DNN-based heuristic that MetaOpt does not support. In this case, we compare MetaEase against the DNN analyzer in [54].

For traffic engineering, we use the total demand routed by the algorithm as our performance metric. Following MetaOpt, we normalize the performance gap by the sum of all link capacities. We evaluate heuristics on two large topologies (Cogentco, Uninett2010) from [1] and three public production topologies: B4 [36], Abilene [74], and SWAN [34] [6]. The

[6] See Tab. 10 for the details of these topologies.

Method	**Normalized Max Gap**
MetaEase	73.13%
DNN Analyzer in [54]	71.84%
Random	63.02%
Simulated Annealing	61.39%
Hill Climbing	58.78%
Sample-based Gradient	58.78%

Table 4: For DOTE, MetaEase finds the largest performance gap on the Abilene topology.

Topology	**Δ Time (h)** (MetaOpt – MetaEase)	**Gap Ratio (%)** (MetaEase/ MetaOpt)
B4	-6.53	88.91%
Abilene	-1.07	98.46%
Swan	0.02	91.19%
Uninett2010	-55.6	65.2%
Cogentco	-12.69	87.8%

Table 5: For POP, MetaEase finds performance gaps within 15% of MetaOpt for most topologies (within 35% for Uninett2010). It needs more time on larger topologies.

average link capacity in each topology is 200 Gbps. For Cogentco and Uninett2010, we restrict the routing to the 4-shortest paths. For other topologies, we use all the available paths. We use projected gradients with $K = 16$ and allow demands to vary within 2× the average link capacity.

For Demand Pinning [43], MetaEase discovers performance gaps within 90% to 100% of what MetaOpt discovers but does so in significantly less time (up to 8 hours faster) across most topologies (Tab. 3). The only exception is on the Uninett2010 topology where MetaEase's gap is 65% of MetaOpt's. Compared to black-box baselines, MetaEase identifies up to 7.3× larger performance gaps under the same time budget (Fig. 8). Fig. 7 illustrates the progress of each algorithm over time for publicly available topologies. Similar to MetaOpt, we set the demand pinning threshold to 5% of average link capacities.

For DOTE [60], MetaEase identifies the largest performance gap among all the baselines (Tab. 4). We used the open-source DOTE implementation [61] and trained it on the Abilene topology until achieving 99.95% test accuracy. We compare against the solution from Namyar et al. [54] (obtained from the authors) as well as black-box search methods.

For POP [56], we compare MetaEase against other baselines in Tab. 5 and Fig. 9. POP divides node pairs and their demands uniformly at random into partitions and assigns each partition an equal share of link capacities. It then solves the original optimization problem (e.g., max-flow) independently for each partition. We report the expected gap similar to MetaOpt. MetaEase's results are within 15% of MetaOpt on most topologies and within 35% for Uninett2010. Analyzing POP using MetaEase is slower because POP runs separate optimizations for each partition and lacks distinct code paths.

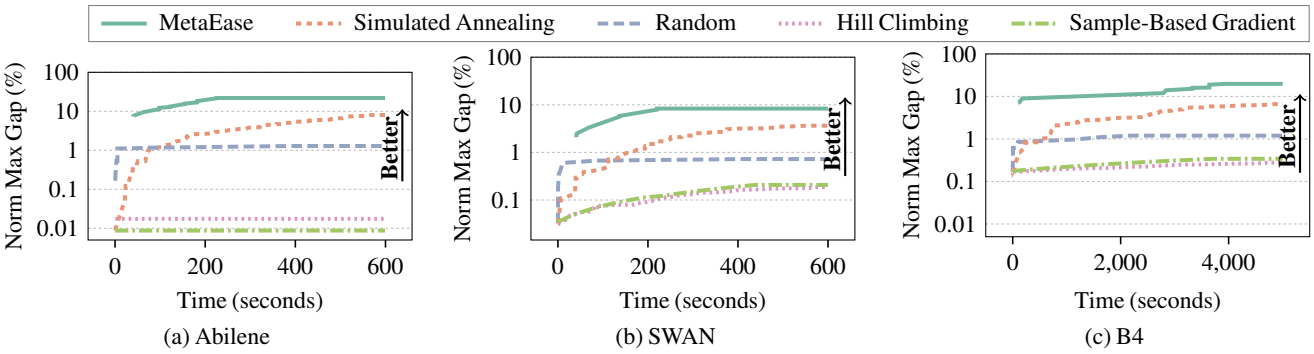


Figure 7: MetaEase finds larger performance gaps faster compared to black-box approaches. For a fair comparison, we allow all methods to fully utilize all threads available on the machine.

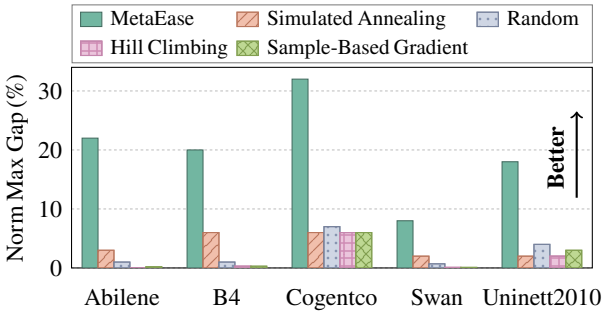


Figure 8: For Demand Pinning, MetaEase finds larger gaps compared to all the black-box baselines for the same time budget. We normalize the gap by the total capacity in each topology.

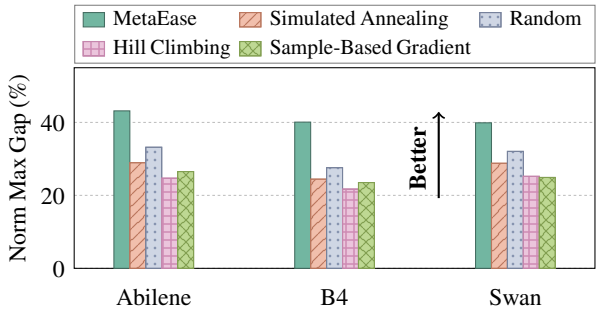


Figure 9: MetaEase finds larger performance gap compared to black-box baselines for POP across three topologies (Abilene, B4, SWAN) under the same time budget.

Thus, KLEE produces no output, impacting convergence.

For the example heuristic in §2, we observe an unexpected result: the heuristic is optimal on the SWAN topology, and neither MetaEase nor MetaOpt found a scenario where it underperforms. On larger topologies, MetaEase finds scenarios with up to 30% larger performance gaps compared to MetaOpt within the same time frame. We set MetaOpt's timeout to 4× that of MetaEase and applied its partitioning technique for efficiency. Despite these adjustments, MetaOpt still could not find scenarios where the heuristic underperforms on most topologies. MetaEase also outperforms black-box algorithms (Fig. 19).

Topology	Δ Time (h) (MetaOpt – MetaEase)	MetaEase Gap (%)	MetaOpt Gap (%)
Abilene	0.50	30%	0.02%
B4	0.88	25.47%	25.47%
Uninett2010	11.6	12.86%	0%
Cogentco	10.4	4.36%	0%

Table 6: For the example heuristic in §2, MetaEase finds the same or a larger performance gap faster than MetaOpt on all topologies.

5.3 MetaEase can analyze diverse problem domains

Domains where the optimal benchmark solution is convex, such as traffic engineering, are well-suited for MetaEase. We show MetaEase also performs well on other (non-convex) problems using four examples (Tab. 2):

For Vector bin-packing (VBP), MetaEase finds equal or better performance gaps than MetaOpt in 5 of 6 settings (Tab. 7). VBP is commonly used in Virtual Machine Placement [58], where the goal is to pack multi-dimensional items into bins of fixed capacity. Performance in VBP depends on the number of used bins, with fewer being better. The optimal form of VBP is APX-hard [79], so many practitioners use first fit decreasing (FFD), a greedy and iterative heuristic [58].

For FFD, MetaEase nearly always matches MetaOpt. In one large-scale setting, MetaEase's performance gap is 5× than that of MetaOpt – MetaOpt failed to improve its objective even after 6 hours. Although MetaEase takes longer (due to KLEE), it continues to progress to a higher gap.

For Knapsack, MetaEase again shows larger performance gaps relative to all black-box baselines (Fig. 10). Given a set of items with associated weights and values, the knapsack problem determines which subset of items maximizes the total value while ensuring the total weight does not exceed a specified limit. It is NP-complete [77] and is used to solve problems such as LTE downlink scheduling [30]. We analyze the commonly used greedy heuristic [30], which computes

Setting (#Balls, #Dims)	Δ Time (h) MetaOpt –MetaEase	Gap Ratio (%) MetaEase / MetaOpt
(10, 1)	2.21	100.00%
(10, 2)	-55.68	100.00%
(15, 1)	1.37	100.00%
(15, 2)	2.20	75.00%
(20, 1)	-16.79	100.00%
(20, 2)	-74.95	500.00%

Table 7: MetaEase finds comparable gaps to MetaOpt across different problem sizes for FFD.

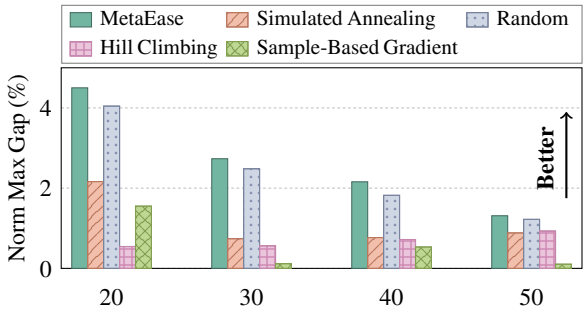


Figure 10: MetaEase outperforms all the baselines when it analyzes knapsack problems for 20, 30, 40, and 50 items. We normalized the gaps by number of items multiplied by the upper bound on each item's weight.

(1) each item's value-to-weight ratio, (2) sorts items by this ratio, and (3) picks items in order until capacity is exhausted.

Black-box approaches also perform well here: when item values and weights vary widely, the greedy heuristic is more likely to underperform, and random sampling is more likely to encounter these adverse cases.

For Arrow [85], MetaEase outperforms black-box baselines (Fig. 11). To our knowledge, no other heuristic analyzer, including MetaOpt, can analyze Arrow since it solves a non-convex optimization with an objective. Arrow jointly assigns optical resources (wavelengths and fibers) after a fiber cut. It introduces a randomized heuristic through its LotteryTickets concept and models the heuristics as a mixed-integer optimization. As a result, the heuristic has a single code path, and although we lose the advantages of using Klee, MetaEase is still able to find competitive performance gaps.

Random search also performs well in this problem because it is very lightweight and explores more samples than other methods. In contrast, MetaEase incurs overhead to run KLEE and train a Gaussian Process; the former provides minimal benefit.

We use the B4 topology (12 nodes, 39 optical links) and the IBM topology (17 nodes, 154 optical links) in our experiments [85]. We run Arrow with 3 LotteryTickets and consider 3 single-fiber cut scenarios. We repeat each setting 10 times and verified that this sample size captures Arrow's variance.

For Maximum Weight Matching, MetaEase shows the largest performance gap on the Abilene, Cogentco, and Uninett2010 topologies and consistently ranks among the top

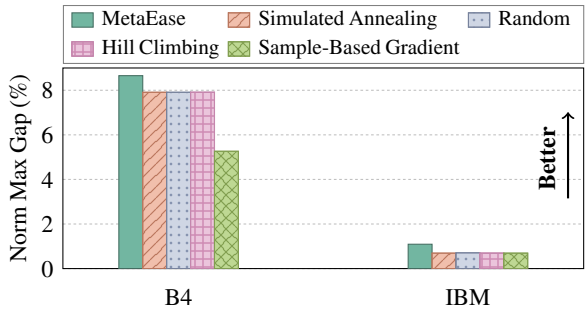


Figure 11: For Arrow, MetaEase finds larger gaps in the same time budget. We show the maximum gaps normalized by the total capacity of the healthy network.

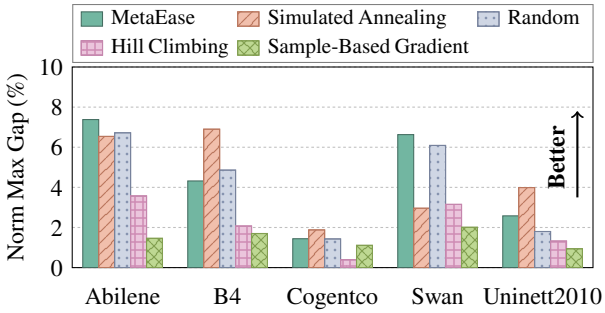


Figure 12: MetaEase finds the largest performance gap across most topologies when we use it to analyze a greedy heuristic that solves the maximum weight matching problem.

two approaches (Fig. 12). Given a graph with weighted edges, maximum weight matching selects a set of non-overlapping edges to maximize the sum of their weights. Prior work used it to understand the throughput limits of datacenter topologies [38, 55]. We analyze a greedy heuristic from [77] and omit the details for brevity.

6 Evaluating MetaEase's Design Choices

We now demonstrate how each module in MetaEase contributes to its improved performance by analyzing the demand pinning heuristic using public topologies.

6.1 Seed Generation Module

MetaEase uses KLEE to partition inputs into equivalence classes and seeds the gradient-based search with samples from each class. Operators may optionally provide domain knowledge to select better starting points for each class (see §4.2).

We compare MetaEase w/ hints with three alternatives in Fig. 14: (i) MetaEase without hints; (ii) using 30 random starting points; and (iii) replacing KLEE with an LLM that analyzes the code and generates equivalence classes. For the latter, we prompted GPT-5 in reasoning mode to analyze the heuristic's code along with the topology and generate diverse demand matrices exposing different heuristic behaviors (see Fig. 20).

MetaEase with hints consistently finds gaps that are larger or comparable to vanilla MetaEase, and it is on average 1.6× faster.

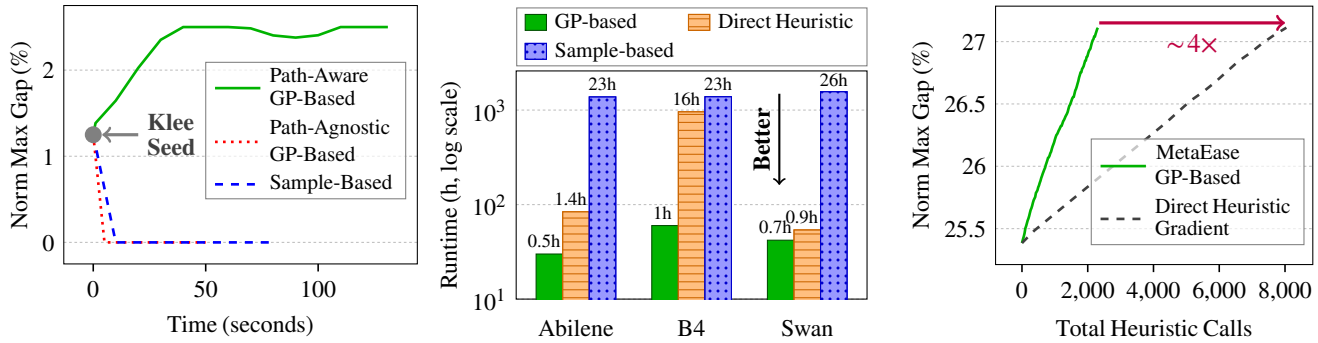


Figure 13: "Path-based" and "Surrogate-based" methods are effective: (left) MetaEase's gradient ascent improves upon the initial point and handles the discontinuities; (middle) surrogate-based is faster than sample-based variants, including vanilla sample-based and one where it only estimates heuristic's gradient; (right) MetaEase reduces the number of heuristic calls.

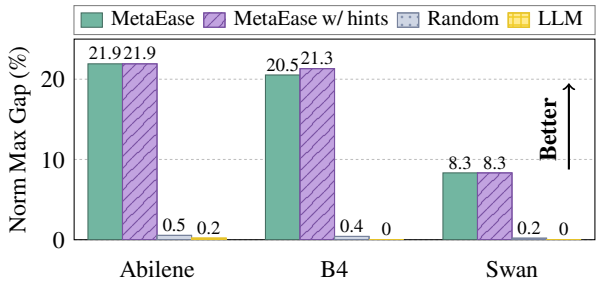


Figure 14: MetaEase with hints consistently produces larger performance gaps over the baselines, followed closely by vanilla MetaEase. All MetaEase variants outperform the baselines.

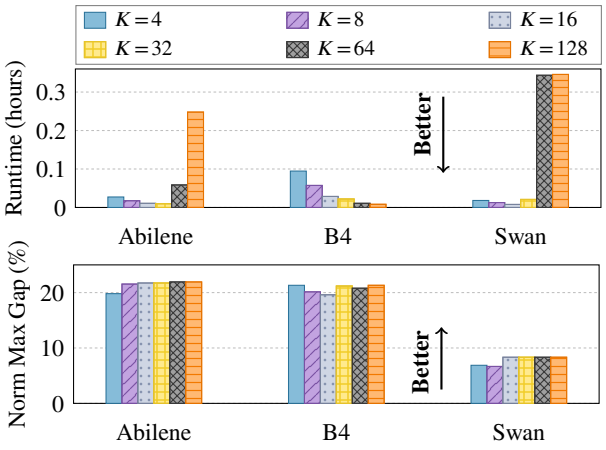


Figure 15: The number of dimensions in projected gradients (K) introduces a tradeoff between the quality of the discovered performance gap and the runtime.

MetaEase also discovers substantially larger performance gaps than other methods. Surprisingly, the LLM-based equivalence classes did not improve the gradient-based search — this may mean we need to engineer a better prompt, but it may also indicate that LLMs are not well suited to this problem.

6.2 Gradient Ascent Module

Path-based gradients help with non-differentiability. We analyze the Demand Pinning heuristic on the Abilene topology. In this case, KLEE returned a starting point for an equivalence class that fell exactly at the demand pinning threshold. The gradient-based search then identified a larger performance gap than this initial point. Using a path-based gradient allowed us to find an even greater gap compared to a search that did not remain within the same equivalence class (Fig. 13–left).

Effectiveness of Gaussian Process-Based Search. We compared gradient ascent using Gaussian Process surrogates with two sample-based variants: (1) directly estimating heuristic's gradient and running gradient ascent for each code path, and (2) the traditional sample-based gradient ascent search. Both sample-based approaches were slower (Fig. 13–middle).

Sample-based gradient ascent is less efficient than the Gaussian Process approach (Fig. 13–right), requiring 4× more heuristic calls than MetaEase. For d variables, it requires $d+1$ heuristic calls per step, whereas MetaEase makes only N calls, regardless of the number of variables. This gap widens on larger topologies: MetaEase is 16.6× faster for Demand Pinning on B4 and 2.8× faster on Abilene (Fig. 13–middle).

6.3 The impact of hyperparameters

MetaEase has three main hyper-parameters: the number of dimensions for projected gradients (K), the block size for training the Gaussian Process (Δ), and the number of samples per block (N). K introduces a trade-off between input space exploration and the time required to run KLEE and train the Gaussian Process (Fig. 15). Block size introduces another trade-off: larger blocks reduce the accuracy of the Gaussian Process surrogate and the gradient estimate, while very small blocks increase MetaEase's runtime (Fig. 16). The number of samples per block, N, directly affects the accuracy of the Gaussian Process (Fig. 17). We evaluate the Gaussian Process using 200 held-out samples. We observe that increasing N improves surrogate quality but also increases runtime.

7 Related Work

General domain heuristic analyzers. To our knowledge, MetaEase is the first general heuristic analyzer that operates

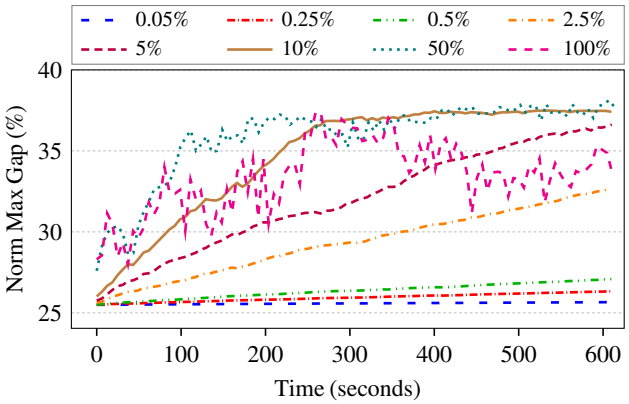


Figure 16: The block size Δ introduces a tradeoff: a larger Δ reduces the quality of the approximate gradient (which causes oscillations in MetaEase), while a smaller Δ increases runtime.

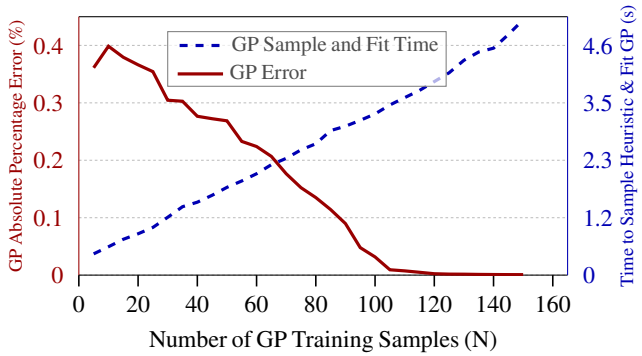


Figure 17: Training the Gaussian Process with more samples improves accuracy but increases runtime.

without a mathematical model of the heuristic, instead exploiting its structure to guide search. Prior approaches either encode heuristics into analytic abstractions—such as XPlain's network-flow model [40] or MetaOpt's optimization formulation [52]—or treat them as black boxes, using simulated annealing, hill climbing [26, 41], sample-based gradient methods [66], Bayesian/surrogate optimization [18, 29, 31, 37, 39, 49, 65, 70], or derivative-free global search [23]. Symbolic execution tools likewise generate targeted adversarial test cases [19, 22, 57].

Domain-specific analysis and anomaly detection. Several systems focus on specific domains: Virelay verifies scheduling heuristics with SMT [32], and FPerf synthesizes adversarial packet schedules to expose throughput and fairness issues [9]. For congestion control, CCAC symbolically explores adversarial traces [12], constraint-guided templates yield provable guarantees [6], and prior work identifies starvation pathologies [11]. Analytic models reveal stability and fairness problems, from classic TCP fluid models [51] to modern analyses of high-bandwidth flow control [28] and CCAs like BBR [68]. HotCocoa adds NIC-oriented abstractions for implementing and analyzing CCAs [10]. Other works motivate solver- or test-based analyzers: Buffy offers a solver-agnostic language for performance analysis [71]; DNS studies find combined performance and security bugs [46]; Dote exposes adversarial samples in learning-enabled systems [54]; Raha detects WAN degradation [13]; and BOLT targets network functions [35]. Control-plane verification has advanced significantly [4, 16, 62, 63, 69, 73, 75, 76]. Collie extends these efforts to RDMA subsystems [42].

8 Discussion

We highlight several additional observations from our development of MetaEase:

MetaEase extends beyond heuristic analysis. Raha [13] demonstrates that MetaOpt can identify failures that trigger worst-case degradation in a network. MetaEase is applicable to such problems as well. Both MetaEase and MetaOpt address a variant of the *Stackelberg game*, where a leader maximizes its payoff by selecting inputs that influence followers, who in turn affect the leader's payoff [52]. We believe the techniques developed in MetaEase can generalize to other Stackelberg variants.

Alternative definitions of equivalence classes may be better. Our goal was to identify regions in the input space where the heuristic makes the same decision, which motivated our use of KLEE. However, other mechanisms may perform better for certain heuristics, and we leave this for future work.

We can also find equivalence classes in the benchmark, too. MetaEase allows operators to provide a benchmark implementation instead of an optimization model. In such cases, we can apply KLEE to the benchmark and define equivalence classes as the Cartesian product of the classes returned for the benchmark and the heuristic.

Random black-box methods may suffice in some cases. For some heuristics (e.g., Arrow), the performance curve is nearly flat. In such scenarios, random black-box techniques can explore the search space quickly and achieve a performance gap close (though still smaller) to that of MetaEase. Detecting when a heuristic falls into this category could enable switching to these faster methods when appropriate.

We can always make MetaEase faster. Since we run MetaEase's gradient ascent from different seeds in parallel, we can always improve speed without an inherent upper limit by using more number of threads.

9 Conclusion

To our knowledge, MetaEase is the first general-domain heuristic analyzer that analyzes a heuristic's *implementation* and enables operators to quantify the performance risk it imposes on their networks. Unlike prior approaches, MetaEase requires no mathematical formulation of a heuristic, making it more accessible to operators. MetaEase combines symbolic execution and gradient-based search to find the maximum gap between heuristics and benchmarks. Our evaluation across diverse heuristics and domains demonstrates its effectiveness. We plan to open-source MetaEase to support operators in assessing heuristic risks.

References

[1] Internet topology zoo. http://www.topology-zoo.org/. Accessed: 2025-09-11.

[2] Klee documentation (v2.1). https://klee-se.org/releases/docs/v2.1/docs/. Accessed: 2025-09-18.

[3] Soheil Abbasloo, Chen-Yu Yen, and H. Jonathan Chao. Classic meets modern: a pragmatic learning-based congestion control for the internet. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '20, page 632–647, New York, NY, USA, 2020. Association for Computing Machinery.

[4] Anubhavnidhi Abhashkumar, Aaron Gember-Jacobson, and Aditya Akella. Tiramisu: Fast and general network verification. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pages 201–219, 2020.

[5] Vamsi Addanki, Maria Apostolaki, Manya Ghobadi, Stefan Schmid, and Laurent Vanbever. Abm: Active buffer management in datacenters. In *Proceedings of the ACM SIGCOMM 2022 Conference*, pages 36–52, New York, NY, USA, 2022. Association for Computing Machinery.

[6] Anup Agarwal, Venkat Arun, Devdeep Ray, Ruben Martins, and Srinivasan Seshan. Towards provably performant congestion control. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 951–978, 2024.

[7] Albert Gran Alcoz, Alexander Dietmüller, and Laurent Vanbever. {SP-PIFO}: Approximating {Push-In}{First-Out} behaviors using {Strict-Priority} queues. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, pages 59–76, 2020.

[8] Albert Gran Alcoz, Balázs Vass, Pooria Namyar, Behnaz Arzani, Gábor Rétvári, and Laurent Vanbever. Everything matters in programmable packet scheduling. In *22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25)*, pages 1467–1485, 2025.

[9] Mina Tahmasbi Arashloo, Ryan Beckett, and Rachit Agarwal. Formal methods for network performance analysis. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, 2023.

[10] Mina Tahmasbi Arashloo, Monia Ghobadi, Jennifer Rexford, and David Walker. Hotcocoa: Hardware congestion control abstractions. In *Proceedings of the 16th ACM Workshop on Hot Topics in Networks*, HotNets '17, page 108–114, New York, NY, USA, 2017. Association for Computing Machinery.

[11] Venkat Arun, Mohammad Alizadeh, and Hari Balakrishnan. Starvation in end-to-end congestion control. In *Proceedings of the ACM SIGCOMM 2022 Conference*, pages 177–192, 2022.

[12] Venkat Arun, Mina Tahmasbi Arashloo, Ahmed Saeed, Mohammad Alizadeh, and Hari Balakrishnan. Toward formally verifying congestion control behavior. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, SIGCOMM '21, 2021.

[13] Behnaz Arzani, Sina Taheri, Pooria Namyar, Ryan Beckett, Siva Kakarla, and Elnaz Jallilipour. Raha: A general tool to analyze wan degradation. In *Proceedings of the ACM SIGCOMM 2025 conference*, 2025.

[14] Roberto Baldoni, Emilio Coppa, Daniele Cono D'Elia, Camil Demetrescu, and Irene Finocchi. A survey of symbolic execution techniques. *ACM Computing Surveys*, 51(3), 2018.

[15] Hugo Barbalho, Patricia Kovaleski, Beibin Li, Luke Marshall, Marco Molinaro, Abhisek Pan, Eli Cortez, Matheus Leao, Harsh Patwari, Zuzu Tang, et al. Virtual machine allocation with lifetime predictions. *Proceedings of Machine Learning and Systems*, 5:232–253, 2023.

[16] Ryan Beckett, Ratul Mahajan, Jitendra Padhye, and David Walker. A general approach to network configuration verification (minesweeper). In *ACM SIGCOMM*, pages 321–334, 2017.

[17] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[18] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[19] Cristian Cadar, Daniel Dunbar, Dawson R Engler, et al. Klee: Unassisted and automatic generation of high-coverage tests for complex systems programs. In *OSDI*, volume 8, pages 209–224, 2008.

[20] Cristian Cadar and Koushik Sen. Symbolic execution for software testing: Three decades later. *Communications of the ACM*, 56(2):82–90, 2013.

[21] Paolo H. Calamai and Jorge J. More. Projected gradient methods for linearly constrained problems. In *Nonlinear Programming 3*, pages 71–116. Springer, 1987.

[22] Koen Claessen and John Hughes. Quickcheck: a lightweight tool for random testing of haskell programs. In *Proceedings of the fifth ACM SIGPLAN international conference on Functional programming*, pages 268–279, 2000.

[23] Andrew R. Conn, Katya Scheinberg, and Luis N. Vicente. *Introduction to Derivative-Free Optimization*. SIAM, 2009.

[24] Emilie Danna, Subhasree Mandal, and Arjun Singh. A practical algorithm for balancing the max-min fairness and throughput objectives in traffic engineering. In *2012 Proceedings IEEE INFOCOM*. IEEE, 2012.

[25] John M. Danskin. The theory of max–min, with applications. *SIAM Journal on Control*, 5(1):64–104, 1967.

[26] Lawrence Davis. Bit-climbing, representational bias, and test suite design. In *Proc. 4th Int'l Conf. on Genetic Algorithms*, pages 18–23. Morgan Kaufmann, 1991.

[27] Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient smt solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer, 2008.

[28] Nandita Dukkipati and Nick McKeown. Flow control in fast long-distance networks. In *IEEE INFOCOM*, pages 1079–1088, 2006.

[29] David Eriksson, Michael Pearce, Jacob Gardner, Ryan D. Turner, and Matthias Poloczek. Scalable global optimization via local bayesian optimization. In *NeurIPS*, 2019.

[30] Nasim Ferdosian, Mohamed Othman, Borhanuddin Mohd Ali, and Kweh Yeah Lun. Greedy–knapsack algorithm for optimal downlink resource allocation in lte networks. *Wireless Networks*, 22(5):1427–1440, 2016.

[31] Alexander Forrester, András Sobester, and Andy Keane. *Engineering Design via Surrogate Modelling: A Practical Guide*. Wiley, 2008.

[32] Saksham Goel, Benjamin Mikek, Jehad Aly, Venkat Arun, Ahmed Saeed, and Aditya Akella. Quantitative verification of scheduling heuristics, 2023.

[33] Chi-Yao Hong, Srikanth Kandula, Ratul Mahajan, Ming Zhang, Vijay Gill, Mohan Nanduri, and Roger Wattenhofer. Achieving high utilization with software-driven wan. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, pages 15–26, 2013.

[34] Chi-Yao Hong, Srikanth Kandula, Ratul Mahajan, Ming Zhang, Vijay Gill, Mohan Nanduri, and Roger Wattenhofer. Achieving high utilization with software-driven wan. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, pages 15–26, 2013.

[35] Rishabh Iyer, Luis Pedrosa, Arseniy Zaostrovnykh, Solal Pirelli, Katerina Argyraki, and George Candea. Performance contracts for software network functions. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*, pages 517–530, 2019.

[36] Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, et al. B4: Experience with a globally-deployed software defined wan. *ACM SIGCOMM Computer Communication Review*, 43(4):3–14, 2013.

[37] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[38] Sangeetha Abdu Jyothi, Ankit Singla, P. Brighten Godfrey, and Alexandra Kolla. Measuring and understanding throughput of network topologies. In *SC '16: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 761–772, 2016.

[39] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018. Posterior mean of GP regression equals kernel ridge regression.

[40] Pantea Karimi, Solal Pirelli, Siva Kesava Reddy Kakarla, Ryan Beckett, Santiago Segarra, Beibin Li, Pooria Namyar, and Behnaz Arzani. Towards safer heuristics with xplain. In *Proceedings of the 23rd ACM Workshop on Hot Topics in Networks*, pages 68–76, 2024.

[41] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[42] Xinhao Kong, Yibo Zhu, Huaping Zhou, Zhuo Jiang, Jianxi Ye, Chuanxiong Guo, and Danyang Zhuo. Collie: Finding performance anomalies in {RDMA} subsystems. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pages 287–305, 2022.

[43] Umesh Krishnaswamy, Rachee Singh, Nikolaj Bjørner, and Himanshu Raj. Decentralized cloud wide-area network traffic engineering with {BLASTSHIELD}. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pages 325–338, 2022.

[44] Umesh Krishnaswamy, Rachee Singh, Paul Mattes, Paul-Andre C Bissonnette, Nikolaj Bjørner, Zahira Nasrin, Sonal Kothari, Prabhakar Reddy, John Abeln, Srikanth Kandula, et al. {OneWAN} is better than two: Unifying a split {WAN} architecture. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 515–529, 2023.

[45] Averill M. Law. *Simulation Modeling and Analysis*. McGraw-Hill Education, New York, 5 edition, 2015.

[46] Si Liu, Huayi Duan, Lukas Heimes, Marco Bearzi, Jodok Vieli, David Basin, and Adrian Perrig. A formal framework for end-to-end dns resolution. In *Proceedings of the ACM SIGCOMM 2023 Conference*, pages 932–949, 2023.

[47] Kshiteej Mahajan, Arjun Balasubramanian, Arjun Singhvi, Shivaram Venkataraman, Aditya Akella, Amar Phanishayee, and Shuchi Chawla. Themis: Fair and efficient {GPU} cluster scheduling. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, pages 289–304, 2020.

[48] Nick McKeown. The islip scheduling algorithm for input-queued switches. *IEEE/ACM transactions on networking*, 2002.

[49] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

[50] Paul Milgrom and Ilya Segal. Envelope theorems for differential and nondifferentiable optimization. *Econometrica*, 70(2):583–601, 2002.

[51] Vishal Misra, Wei-Bo Gong, and Don Towsley. Fluid-based analysis of a network of aqm routers supporting tcp flows with an application to red. In *ACM SIGCOMM*, pages 151–160, 2000.

[52] Pooria Namyar, Behnaz Arzani, Ryan Beckett, Santiago Segarra, Himanshu Raj, Umesh Krishnaswamy, Ramesh Govindan, and Srikanth Kandula. Finding adversarial inputs for heuristics using multi-level optimization. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 927–949, 2024.

[53] Pooria Namyar, Behnaz Arzani, Srikanth Kandula, Santiago Segarra, Daniel Crankshaw, Umesh Krishnaswamy, Ramesh Govindan, and Himanshu Raj. Solving {Max-Min} fair resource allocations quickly on large graphs. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, 2024.

[54] Pooria Namyar, Michael Schapira, Ramesh Govindan, Santiago Segarra, Ryan Beckett, Siva Kesava Reddy Kakarla, and Behnaz Arzani. End-to-end performance analysis of learning-enabled systems. In *Proceedings of the 23rd ACM Workshop on Hot Topics in Networks*, HotNets '24, page 86–94, New York, NY, USA, 2024. Association for Computing Machinery.

[55] Pooria Namyar, Sucha Supittayapornpong, Mingyang Zhang, Minlan Yu, and Ramesh Govindan. A throughput-centric view of the performance of datacenter topologies. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, SIGCOMM '21, page 349–369, New York, NY, USA, 2021. Association for Computing Machinery.

[56] Deepak Narayanan, Fiodar Kazhamiaka, Firas Abuzaid, Peter Kraft, Akshay Agrawal, Srikanth Kandula, Stephen Boyd, and Matei Zaharia. Solving large-scale granular resource allocation problems efficiently with pop. In *Proceedings of the ACM SIGOPS 28th Symposium on Operating Systems Principles*, pages 521–537, 2021.

[57] Carlos Pacheco, Shuvendu K Lahiri, Michael D Ernst, and Thomas Ball. Feedback-directed random test generation. In *29th International Conference on Software Engineering (ICSE'07)*, pages 75–84. IEEE, 2007.

[58] Rina Panigrahy, Kunal Talwar, Lincoln Uyeda, and Udi Wieder. Heuristics for vector bin packing. January 2011.

[59] Laurent Perron and Vincent Furnon. Or-tools.

[60] Yarin Perry, Felipe Vieira Frujeri, Chaim Hoch, Srikanth Kandula, Ishai Menache, Michael Schapira, and Aviv Tamar. {DOTE}: Rethinking (predictive){WAN} traffic engineering. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 1557–1581, 2023.

[61] PredWanTE. Predwante/dote: Implementations for dote traffic engineering. https://github.com/PredWanTE/DOTE, 2023. Accessed: 2025-09-03.

[62] Divya Raghunathan, Ryan Beckett, Aarti Gupta, and David Walker. Acorn: Network control plane abstraction using route nondeterminism. *arXiv preprint arXiv:2206.02100*, 2022.

[63] Divya Raghunathan et al. Abstract interpretation of distributed network control planes (shapeshifter). In *ACM Symposium on Principles of Programming Languages (POPL)*, pages 213–227, 2020.

[64] Sudarsanan Rajasekaran, Manya Ghobadi, and Aditya Akella. {CASSINI}:{Network-Aware} job scheduling in machine learning clusters. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 1403–1420, 2024.

[65] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[66] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[67] Ahmed Saeed, Nandita Dukkipati, Vytautas Valancius, Vinh The Lam, Carlo Contavalli, and Amin Vahdat. Carousel: Scalable traffic shaping at end hosts. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, pages 404–417, 2017.

[68] Simon Scherrer, Markus Legner, Adrian Perrig, and Stefan Schmid. Model-based insights on the performance, fairness, and stability of bbr. In *Proceedings of the 22nd ACM Internet Measurement Conference*, pages 519–537, 2022.

[69] Johann Schlamp, Matthias Wählisch, Thomas C. Schmidt, Georg Carle, and Ernst W. Biersack. Cair: Using formal languages to study routing, leaking, and interception in bgp. *arXiv preprint arXiv:1605.00618*, 2016.

[70] Matthias Seeger. Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106, 2004.

[71] Amir Seyhani, Junyi Zhao, Aarti Gupta, David Walker, and Mina Tahmasbi Arashloo. Buffy: A formal language-based framework for network performance analysis. In *Proceedings of the 23rd ACM Workshop on Hot Topics in Networks*, HotNets '24, page 95–102, New York, NY, USA, 2024. Association for Computing Machinery.

[72] Aashaka Shah, Vijay Chidambaram, Meghan Cowan, Saeed Maleki, Madan Musuvathi, Todd Mytkowicz, Jacob Nelson, Olli Saarikivi, and Rachee Singh. {TACCL}: Guiding collective algorithm synthesis using communication sketches. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 593–612, 2023.

[73] Xiaozhe Shao, Zibin Chen, Daniel Holcomb, and Lixin Gao. Accelerating bgp configuration verification through reducing cycles in smt constraints (binode). *IEEE/ACM Transactions on Networking*, 2021.

[74] Stanford University IT. Abilene core topology. Technical web page, 2015. Accessed: 2025-09-11.

[75] S. Steffen et al. Probabilistic verification of network configurations (netdice). In *ACM SIGCOMM*, 2020.

[76] Alan Tang, Ryan Beckett, Steven Benaloh, Karthick Jayaraman, Tejas Patil, Todd Millstein, and George Varghese. Lightyear: Using modularity to scale bgp control plane verification. In *Proceedings of the ACM SIGCOMM 2023 Conference*, ACM SIGCOMM '23, page 94–107, New York, NY, USA, 2023. Association for Computing Machinery.

[77] V.V. Vazirani. *Approximation Algorithms*. Springer Berlin Heidelberg, 2013.

[78] Guohui Wang, David G. Andersen, Michael Kaminsky, Konstantina Papagiannaki, T.S. Eugene Ng, Michael Kozuch, and Michael Ryan. c-through: part-time optics in data centers. In *Proceedings of the ACM SIGCOMM 2010 Conference*, SIGCOMM '10, page 327–338, New York, NY, USA, 2010. Association for Computing Machinery.

[79] Gerhard J. Woeginger. There is no asymptotic ptas for two-dimensional vector packing. *Information Processing Letters*, 64(6):293–297, 1997.

[80] Zhuolong Yu, Chuheng Hu, Jingfeng Wu, Xiao Sun, Vladimir Braverman, Mosharaf Chowdhury, Zhenhua Liu, and Xin Jin. Programmable packet scheduling with a single queue. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, pages 179–193, 2021.

[81] Haoyu Zhang, Logan Stafman, Andrew Or, and Michael J Freedman. Slaq: quality-driven scheduling for distributed machine learning. In *Proceedings of the 2017 Symposium on Cloud Computing*, pages 390–404, 2017.

[82] Yiwen Zhang, Xumiao Zhang, Ganesh Ananthanarayanan, Anand Iyer, Yuanchao Shu, Victor Bahl, Z Morley Mao, and Mosharaf Chowdhury. Vulcan: Automatic query planning for live {ML} analytics. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 1385–1402, 2024.

[83] Liangyu Zhao, Siddharth Pal, Tapan Chugh, Weiyang Wang, Jason Fantl, Prithwish Basu, Joud Khoury, and Arvind Krishnamurthy. Efficient {Direct-Connect} topologies for collective communications. In *22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25)*, pages 705–737, 2025.

[84] Liangyu Zhao, Siddharth Pal, Tapan Chugh, Weiyang Wang, Jason Fantl, Prithwish Basu, Joud Khoury, and Arvind Krishnamurthy. Efficient {Direct-Connect} topologies for collective communications. In *22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25)*, pages 705–737, 2025.

[85] Zhizhen Zhong, Manya Ghobadi, Alaa Khaddaj, Jonathan Leach, Yiting Xia, and Ying Zhang. Arrow: restoration-aware traffic engineering. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, pages 560–579, 2021.

A Encoding a new TE heuristic in MetaOpt

We wanted to analyze a new sort-based heuristic in MetaOpt. This heuristic routes the demands and selects the top 20% as critical demands, route those demands on the network optimally in stage 1, and then route the rest of the non-critical demands on the residual capacities optimally in stage 2 (see Code 2).

Code 2: New Sort-based Traffic Engineering Heuristic

```
double SortBasedHeuristic(Graph *G) {
   // collect and sort demands
   Demands D = positive_demands(G);
   sort_desc(D);

   // split top 20% as critical
   int k = max(1, 0.2 * size(D));
   Demands Critical = top_k(D, k);
   Demands Remaining = rest(D, Critical);

   // solve optimal on critical demands first
   Routing
     criticalRoutes = SolveOptimal(G, Critical);
   // Adjust the capacities
   Graph G = AdjustCapacities(G, criticalRoutes);
   // solve optimal on non critical demands
   Routing
     nonCriticalRoutes = SolveOptimal(G, Remaining);

   return G->total_met_demands;
}
```

We formulated this heuristic in MetaOpt (see Fig. 18). This formulation is convoluted despite the helper functions MetaOpt provides. The user has to find a way to extract the non-convex parts of the optimization into the outer problem (Raha [13] describes a similar challenge).

To see why this is hard in MetaOpt, let us go through how we can sort the top 20% of demands as part of the optimization. First, we need to introduce binary variables c_i which indicate whether a demand is in the top 20% or not. We face two challenges: (1) since binary variables are non-convex we need to make sure that we do not use them directly to model the heuristic; (2) we need to set the values such that c_i is 1 if the demand falls in the top 20% and to 0 otherwise (since we do not know the values of the demands before we solve the optimization problem we cannot set these values ahead of time). Constraints 10-15 in Fig. 18 allows us to solve the second problem and we also have to apply tricks to move the values c_i out of the heuristic model to avoid the non-convexity they introduce.

Note that optimization formulation is error-prone and it requires careful considerations and testing. Consequently, it's not straightforward to just ask LLMs to write them. There needs to be human interventions to make sure the formulations are correct.

B Gaussian Process surrogate gradient

A Gaussian Process (GP) is a function of the form:

$$\text{Heuristic}_{GP}(I) = m(I) + Cov(I,\mathbf{X})\mathbf{K}^{-1}\big(\mathbf{y} - m(\mathbf{X})\big)$$

which we need to train in order to find the mean function $m(\cdot)$, and the covariance kernel $Cov(\cdot)$. To train it, we use data $(\mathbf{X},\mathbf{y})$, where $\mathbf{X} = \{x_1,\ldots,x_N\}$ are samples from the input region Δ^n and they all are in the same code-path, and $\mathbf{y} = \big(\text{Heuristic}(x_1),\ldots,\text{Heuristic}(x_N)\big)^\top$ are the outputs when we run the heuristic on the input samples.

The analytic gradient of the Gaussian process at I is

$$\nabla\text{Heuristic}_{GP}(I) = \nabla m(I) + \nabla Cov(I,\mathbf{X})\mathbf{K}^{-1}\big(\mathbf{y} - m(\mathbf{X})\big),$$

where

$$\nabla Cov(I,\mathbf{X}) = \big[\nabla_I Cov(I,x_1),\ldots,\nabla_I Cov(I,x_N)\big].$$

The gradient of the GP is closed-form and we leverage that in estimating the gradient of the heuristic in region Δ^n.

C How MetaEase restricts the search within one code path

For each seed I returned by KLEE [19], we run Heuristic(I) and record its code-path signature $h = \text{path}(I)$. This signature defines the equivalence class $Class_h$.

At iteration t of updates, given the current point I_t ($\text{path}(I_t) = h$), we draw a batch of samples in a small box around I_t (side length 2Δ). We evaluate the heuristic on these candidates and retain those with signature h, forming a training set $\mathbf{X}_t$. From $\mathbf{X}_t$, we select N points to fit a Gaussian Process surrogate, $\text{Heuristic}_{GP}(x)$. We then compute the objective gradient

$$g_t = \nabla\text{Benchmark}(I_t) - \nabla\text{Heuristic}_{GP}(I_t),$$

according to §2.4, and propose the next step

$$I_{t+1} = I_t + \eta g_t.$$

If $\text{path}(I_{t+1}) = h$, we accept the step. Otherwise, we project the move back into the same class: specifically, we choose the point $z \in \mathbf{X}_t$ whose direction from I_t has the smallest acute angle with g_t, i.e.,

$$z = \operatorname*{argmax}_{u\in\mathbf{X}_t} \frac{g_t^\top(u - I_t)}{\|u - I_t\|} \quad \text{s.t.} \quad \frac{g_t^\top(u - I_t)}{\|u - I_t\|} > 0.$$

We then set $I_{t+1} = z$. This ensures progress while staying inside $Class_h$.

The process repeats until convergence, budget exhaustion, or until all candidate angles are obtuse ($g_t^\top(u - I_t) < 0$), at which point the ascent terminates.

Tool	Heuristic format as input	Guarantee worst-case type	Heuristic Type
MetaEase	C implementation (Code)	Empirical best found (symbolic-guided)	Domain agnostic. Any heuristic (non-convex, DNN-based, randomized, etc)
MetaOpt [52]	Optimization form	Formal lower bound	Domain agnostic. Only convex or feasibility heuristic
Virelay [32]	Model + assumptions (SMT)	Formal (model-level)	General Domain (focuses on scheduling). Any heuristic expressible in SMT
XPlain [40]	Network-flow Abstraction	Formal lower bound	Domain agnostic. Only convex or feasibility heuristic
FPerf [9]	Queueing model + Query	Witness (existence)	Queue management. Discrete-event, rule-based, non-convex (e.g., FIFO)
Black-box Search	Executable	Empirical best found (blind search)	Domain agnostic. Any heuristic

Table 8: MetaEase compared to prior performance analyzers and black-box baselines. Unlike prior tools, MetaEase operates directly on heuristic code and leverages symbolic-guided search to uncover large performance gaps of heuristic compared to a benchmark.

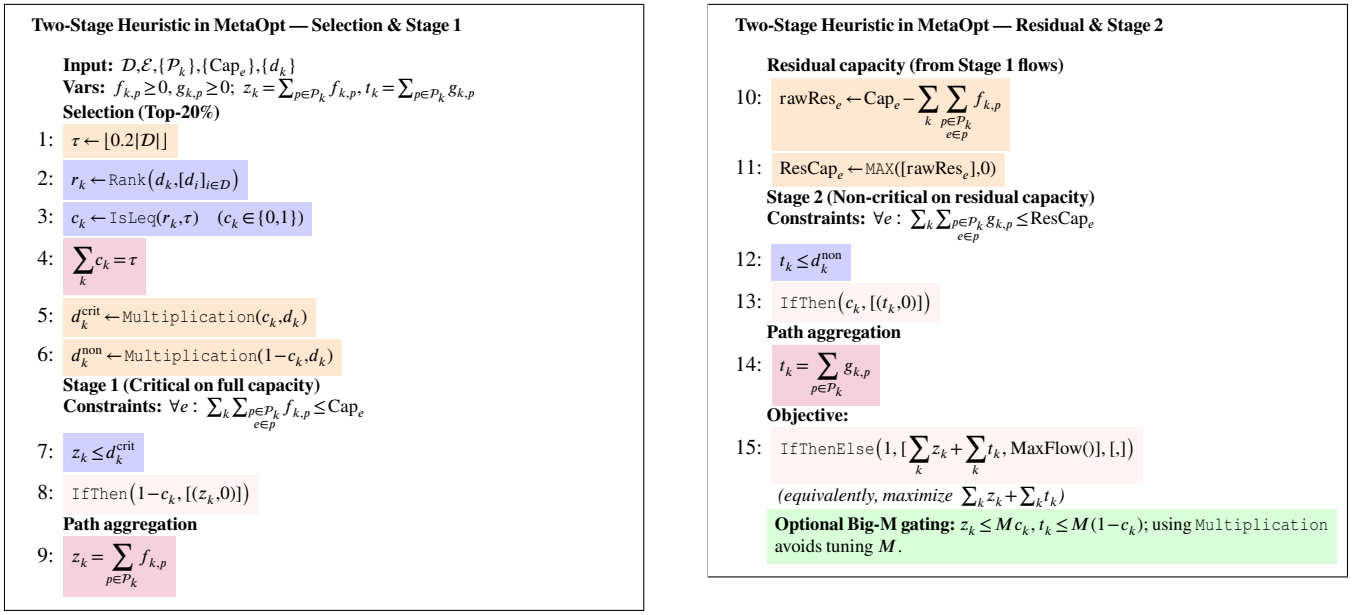


Figure 18: Encoding of Traffic Engineering heuristic in MetaOpt. This heuristic selects the top-20% of demands as critical (1) route critical set optimally, (2) then route remaining non-critical demands on residual capacities. This required a lot of attention and LLMs can't do it without human supervision to check for correctness.

Symbol	Meaning
Δ	Block size for GP sampling around a point §3.2
N	Number of samples for fitting the GP §3.2
η	Learning rate for gradient ascent §3.4
K	Number of variables in projected gradient technique §4.1

Table 9: Parameters in MetaEase's gradient ascent

Topology	#Nodes	#Edges
Cogentco	197	486
Uninett2010	74	202
Abilene	10	26
B4	12	38
Swan	8	24

Table 10: Detail of topologies used for evaluation. We used similar topologies to MetaOpt [52].

D How MetaEase handles randomness

In many cases, operators may be dealing with randomized heuristics, or they may want to measure the performance gap over different scenarios. For example, in studying Arrow's heuristic in §5.3, we look at the performance gap across different standards of fiber cut. In these cases, MetaEase maximizes the *expected* performance gap. Let $I \in \mathcal{I}$ denote the input, ξ the exogenous "scenario" randomness (different scenarios of fiber cut), and τ the heuristic's internal randomness (e.g., number of tickets in Arrow (§5.3), random partitions in POP (§5.2).

Objective. We denote the expected performance gap as

$$\max_{I \in \mathcal{I}} \mathbb{E}_{\tau,\xi}\left[\text{Benchmark}(I;\xi) - \text{Heuristic}(I;\tau,\xi)\right] \tag{4}$$

By linearity of expectation,

$$\max_{I \in \mathcal{I}} \mathbb{E}_{\xi}\left[\text{Benchmark}(I;\xi)\right] - \mathbb{E}_{\tau,\xi}\left[\text{Heuristic}(I;\tau,\xi)\right]. \tag{5}$$

Gradients. Under mild regularity (differentiation under the expectation),

$$\nabla_I \mathbb{E}_\xi\big[\text{Benchmark}(I;\xi)\big] = \mathbb{E}_\xi\big[\nabla_I \text{Benchmark}(I;\xi)\big], \quad (6)$$

$$\nabla_I \mathbb{E}_{\tau,\xi}\big[\text{Heuristic}(I;\tau,\xi)\big] = \mathbb{E}_{\tau,\xi}\big[\nabla_I \text{Heuristic}(I;\tau,\xi)\big]. \quad (7)$$

We keep the same decomposition used in the deterministic case:

$$\nabla \text{Gap}(I) = \underbrace{\nabla \text{Benchmark}(I)}_{\text{via dual/Lagrangian}} - \underbrace{\nabla \text{Heuristic}(I)}_{\text{via local GP surrogate}},$$

now interpreted in expectation per (6)–(7).

Averaging over Samples. We approximate these expectations by averaging over random samples, reusing the same random draws across both benchmark and heuristic to reduce variance [45]. At each ascent step, we approximate the expectations by averages over n_ξ scenarios and n_τ heuristic draws (often $n_\tau = 1$ if the heuristic's randomness is embedded in the scenario). Draw paired samples

$$\{\xi_i\}_{i=1}^{n_\xi}, \quad \{(\tau_{i,j},\xi_i)\}_{j=1}^{n_\tau} \text{ for each } i,$$

reusing the same ξ_i across the benchmark and heuristic to reduce variance [45].

Benchmark term. When the benchmark admits the Lagrangian in Eq. 3, for each scenario ξ_i we form

$$\phi_i(\lambda_i,\nu_i;I,\xi_i) \triangleq \sup_{x^o} \mathcal{L}\left(x^o,\lambda_i,\nu_i;I,\xi_i\right),$$

and compute its I-gradient (no solver call). The MC gradient estimator is

$$\widehat{\nabla}\text{Benchmark}(I) = \frac{1}{n_\xi}\sum_{i=1}^{n_\xi} \nabla_I \phi_i(\lambda_i,\nu_i;I,\xi_i). \quad (8)$$

Heuristic term. For stochastic heuristics, we directly roll out the C implementation $I \mapsto \text{Heuristic}(I;\tau,\xi)$ across n_τ random seeds τ and n_ξ random scenarios ξ. We report the *average performance* over these runs:

$$\overline{\text{Heuristic}}(I) = \frac{1}{n_\xi n_\tau}\sum_{i=1}^{n_\xi}\sum_{j=1}^{n_\tau} \text{Heuristic}(I;\tau_{i,j},\xi_i).$$

We then fit a local Gaussian Process surrogate to $\overline{\text{Heuristic}}(I)$ in the current neighborhood, and take its analytic gradient:

$$\widehat{\nabla}\text{Heuristic}(I) = \nabla_I \text{Heuristic}^{\text{GP}}(I). \quad (9)$$

Notes. We use this rolled out C implementation to find KLEE points as well. This way, the randomness also takes effect on the seed generation. If only the heuristic is randomized, we set n_ξ=1 for the benchmark and average over τ (POP in §5.2); if both the scenarios and heuristic are randomized(Arrow in §5.3), we couple the same ξ_i across both terms.

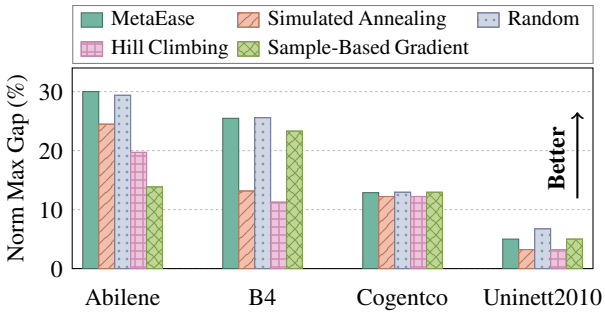


Figure 19: MetaEase delivers top-2 gaps compared to black-box searches when we analyze the example heuristic in §2.

E Other MetaEase runtime optimizations

Operators may want to do quick initial tests to make sure the heuristic they designed performs well in the majority of the input space. MetaEase provides options that allow users to control whether it should analyze the heuristic comprehensively. These options:

Prune equivalence classes. MetaEase can first evaluate the representative points KLEE returns for each equivalence class and remove those with no performance gap (it assumes those regions are probably areas where the heuristic performs well).

Prioritize equivalence classes. MetaEase normally runs gradient ascent in each equivalence class in parallel; but it also has an option where it orders the equivalence classes before it runs gradient ascent based on the performance gap of the representative point for that class — if we assume the representative point is a good estimate of how hard that region is for the heuristic, then this would allow MetaEase to prioritize search over regions that are likely to produce bad outcomes.

Early stop (timeouts). MetaEase also has a timeout feature that limits how long it searches each equivalence class.

F Evaluation Details

Fig. 19 shows the details of comparison of MetaEase with black-box approaches for the example heuristic in §2. MetaEase delivers top-2 gaps.

We also report the raw end-to-end runtime for MetaEase in Tab. 11 across all the problems and all the heuristics in the paper.

G Seed Generation using LLM

To complement symbolic execution seeds, we also explored using large language models (LLMs) to automatically generate diverse seed inputs. The idea is to frame seed generation as a prompt-driven task: given the heuristic code and network topology, the LLM is asked to propose synthetic demand samples that exercise qualitatively different behaviors of the heuristic. Fig. 20 shows the prompt template we use for the `Demand Pinning` heuristic. Based on the results, naively prompting the LLM did not work. The LLM often produced trivial or redundant traffic demands. More work is needed to

make this approach work.

Problem	Instance	Time (s)
TE: Demand Pinning §5.2	Uninett2010	7714
	Cogentco	1558
TE: POP §5.2	Abilene	4883
	B4	24528
	Swan	898
	Uninett2010	202320
	Cogentco	51480
TE: Heuristic in §2	Abilene	100
	B4	250
	Swan	500
	Uninett2010	1640
	Cogentco	5971
DOTE §5.2	Abilene	2478
VBP:FFD §5.3 (# items, # Dim)	10-1	2839
	10-2	215066
	15-1	5857
	15-2	8449
	20-1	71277
	20-2	280748
Knapsack §5.3 (# items)	20	8673
	30	10676
	40	12931
	50	15518
Arrow §5.3	B4	143098
	IBM	6907
Maximum Weight Matching §5.3	Abilene	108
	B4	1628
	Swan	200
	Uninett2010	1751
	Cogentco	2548

Table 11: MetaEase end-to-end execution times for different problems and heuristics.

Prompt Template for Seed Generation

You are an expert in analyzing heuristics. You are given a WAN Traffic Engineering heuristic named `Demand Pinning`, which operates on the given topology (`TOPOLOGY`). This heuristic has a pinning threshold: for each demand, if the demand value is less than the threshold, it is routed through its first shortest path. For the rest of the demands, they are routed on residual capacities optimally.
Your task is to generate a diverse set of synthetic demand samples that explore the range of behaviors of this heuristic.

```
Demand Pinning Code:

// Insert Demand Pinning implementation here

TOPOLOGY:

// Insert JSON topology description here
```

Requirements:

1. **Diversity of Behaviors:** Each sample must be designed so that it triggers a different decision path or behavior. For example:
   - Highly skewed demand (one dominant demand).
   - Uniform low demands across all pairs.
   - Bottleneck saturation forcing rerouting/load balancing.
   - Edge cases: minimal vs. extremely high demand.
   - Sensitivity cases: adding/removing one demand flips the solution.
   - ...
2. **Number of Samples:** Generate at least 8–10 samples, each highlighting a distinct heuristic behavior.
3. **Topology Constraint:** Ensure that demands only use valid source/destination pairs from `TOPOLOGY`.

Output Formatting: Follow exactly the format below. Each sample should be a JSON object containing a list of source-destination pairs with their demand values.

```
// Example output format:
[
 "Sample_1": {
   {"src": "A", "dst": "B", "demand": 12},
   {"src": "B", "dst": "C", "demand": 5},
   ...
},
 ...
]
```

Figure 20: Prompt given to the LLM to generate seed demands for Demand Pinning.