Leveraging Hardware Probes and Optimizations for Accelerating Fuzz Testing of Heterogeneous Applications

Anonymous Author(s)

ABSTRACT

There is a growing interest in the computer architecture community to incorporate heterogeneity and specialization to improve performance. Developers can create *heterogeneous applications* that consist of both *host* code and *kernel* code, where compute-intensive kernels can be offloaded from CPU to hardware accelerators. Testing such applications on *real* heterogeneous architectures is extremely challenging as kernels are black boxes, providing no information about the kernels' internal execution to diagnose issues such as silent hangs or unexpected results. Additionally, inputs for heterogeneous applications are often large matrices, leading to a vast search space for identifying bug-revealing inputs.

We propose a novel fuzz testing technique, HFuzz, to enable efficient testing on real heterogeneous architectures. HFuzz aims to increase both the observability of hardware kernels and testing efficiency through a three-pronged approach. First, HFuzz automatically generates test guidance by inserting device-side in-kernel hardware probes in addition to host-side software monitors. Second, it performs rapid input space exploration by offloading compute-intensive input mutations to hardware kernels. Third, HFuzz parallelizes fuzzing and enables fast on-chip memory access, by utilizing four FPGA-level optimizations including loop unrolling, shannonization, data preloading, and dynamic kernel sharing.

We evaluate HFuzz on seven open-source OneAPI subjects from Intel. HFuzz speeds up fuzz testing by 4.7× with HW-accelerated input space exploration. By incorporating HW probes in tandem with SW monitors, HFuzz finds 33 defects within 4 hours and reveals 25 unique, unexpected behavior symptoms that could not be found by SW-based monitoring alone. HFuzz is the first to design hardware optimizations to accelerate fuzz testing.

ACM Reference Format:

Anonymous Author(s). 2023. Leveraging Hardware Probes and Optimizations for Accelerating Fuzz Testing of Heterogeneous Applications. In *Proceedings of The 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2023).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

There has been a growing interest in developing specializable hardware accelerators for domain-specific workloads for various performance and energy benefits [1–3]. As an example, FPGA can be easily customized to accelerate applications across a wide variety of domains [4, 5] at lower power and higher performance than general-purpose CPUs [6–8]. Major hardware vendors are offering or plan to offer packages that include both CPUs and FPGAs [9, 10]. Such hardware packages have also been made into all major clouds to accelerate various analytic and learning tasks.

In recent years, fuzz testing has emerged as an effective test generation technique for large software systems [11]. Most fuzzing techniques, such as AFL [12], start from a seed input, generate new inputs by mutating the previous input, and add new inputs to the queue if they improve a given guidance metric such as branch coverage. In this paper, we focus on *fuzz testing* (i.e. *fuzzing*) of applications on a heterogeneous platform with a CPU host and an FPGA device. Such a *heterogeneous application* consists of *host* code and *kernel* code, and the host code offloads compute-intensive kernels from the CPU to the FPGA to run. Despite the potential benefits of FPGAs and their commercial availability to a broad user base, programming FPGAs is notoriously difficult in practice. Ensuring the correctness of FPGA programs, even seemingly-simple kernels, could take a substantial amount of time in terms of months [13]. As such, FPGA programming can be done by only a small handful of hardware experts [14–16]. Automatic fuzz testing of heterogeneous applications, together with root cause analysis of failures, can greatly simplify FPGA programming, thereby making FPGAs accessible to the masses.

There has been significant effort to ease the development of heterogeneous applications with FPGAs. The most successful effort is *high-level synthesis* (HLS) [17]. HLS raises the level of programming abstraction from hardware description languages (such as Verilog) to C/C++ dialects (such as SYCL/DPC++[18]), enabling C/C++ developers on FPGAs. Even when heterogeneous applications are written in HLS languages, debugging and testing these heterogeneous applications can remain a significant challenge due to the following reasons:

Lack of observability. FPGA is a device of massive parallelism but little debugging support exists to help high-level programmers. Kernels run on an FPGA device as black boxes, and it often confuses programmers, e.g., when the kernels silently deadlock. General-purpose FPGA debugging [19, 20] works at the gate level and even when in-circuit debugging information is available, it is difficult to correlate low-level gate signals with high-level variables in HLS programs.

Consider a scenario where an application multiplies two matrices `A` and `B` to create a new matrix `M`: M=A×B and then applies a reciprocal transformation on each element of `M`. This application has two kernels offloaded to FPGA: (1) `matrix_multiply` and (2) `transformer`. To transfer the intermediate result `M` from the first to the second kernel, a pipe is established to facilitate data transfer. For each element in the matrix `M`, the first kernel writes its computed value to the designated pipe, and the second kernel `transformer`

ESEC/FSE 2023, 11 - 17 November, 2023, San Francisco, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

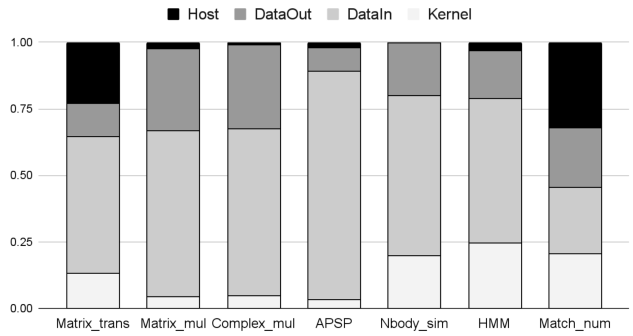

Figure 1: Latency breakdown of running applications on heterogeneous architectures. On average, data transfer into kernels takes 60% of execution time, highlighted in red.

reads it from the pipe, computes the reciprocal, and transfers the final result back to the host. With FPGA emulation, the application works as expected because both kernels run at the same speed. However, when run on an actual FPGA, the speed of the first kernel generating a value can be different from the speed of the second kernel consuming it. The developer should check the size of the pipe, delay writing if it is full or delay reading if it is empty. If such check is not done, the pipe would be saturated or depleted, resulting in data loss and wrong reciprocal outcomes. Currently, due to a lack of observability into the dynamic usage of the pipe, the developer may find it difficult to diagnose the root cause.

Costly transfer of data with high redundancy. Traditional iterative fuzzing techniques often mutate a small part of a seed input to generate new inputs. While this approach works well for many CPU programs, it is extremely ineffective for applications that are run on heterogeneous architectures. Inputs of heterogeneous applications are often large matrices and tensors, leading to significant data access and transfer overheads —the host, which mutates the matrices, must send newly mutated matrices (e.g., with only a few elements modified) to the device. Figure 1 illustrates the latency breakdown of running applications on Intel's heterogeneous architecture. On average, data transfer from CPU to hardware kernels takes 60% of the execution time. For a 100k×100k matrix, a single process of offloading the new generated matrix from the fuzzer to the device would take 2 minutes, prohibiting fast fuzzing on heterogeneous architectures.

Overlooked opportunities for FPGA-level optimizations. Fuzzing heterogeneous applications may be approached in a naïve manner by treating hardware kernel invocations as analogous to software function calls and repeatedly invoking them from an iterative input mutation loop. However, this approach ignores the potential of FPGA optimizations, as the mutations often consist of independent tasks that can be parallelized efficiently when offloaded to the FPGA side. In other words, the nature of fuzzing (i.e., iterative input generation and program invocation) unlocks new micro-architecture level performance optimizations. Indeed, we can treat the domain of heterogeneous applications, not only as a new target domain, but as a new enabler for accelerating automated test generation. When software-style matrix input mutation is offloaded to FPGA and is then combined with subsequent kernel invocation, many micro-architecture level optimizations such as loop unrolling, data preloading, shannonization, and dynamic kernel sharing are now applicable for further performance speed-up.

HFuzz. We developed HFuzz, a novel fuzz testing tool that aims to quickly reveal bugs in heterogeneous applications. Our key insights are elaborated below:

First, to improve error observability during testing, HFuzz injects *hardware probes inside the kernels* in tandem with software monitors inside the host. This is different from prior approaches that consider an FPGA kernel as a black box and inject software monitors only [21]. In HFuzz, both software monitors and hardware probes are designed to effectively detect overflows caused by intermediate variables within the FPGA kernel, as well as pipe saturation errors that may occur during data transfer between different devices. These hardware probes are injected through source-to-source transformation and then synthesized for FPGA. With timely execution feedback from the hardware probes, HFuzz prioritizes inputs that provide a new behavior signal at the FPGA execution level. For example, HFuzz monitors the saturation of a communication pipe between two FPGA kernels and retains the inputs that lead to a new maximum pipe saturation level for further mutations.

Second, HFuzz offloads input mutations into FPGA kernels to reduce unnecessary data transfer. For a `vector-add` example, instead of repeatedly transferring a mutated input vector of size 10^6, HFuzz retains the initial input vector in the FPGA buffer and mutates the elements of the vector within the FPGA kernel. For another example, the host-side mutation of a seed matrix with 10,000 elements for 1,000 times takes 9.1 seconds, in our evaluation, while in-kernel input mutation takes only 2.1 seconds.

Third, HFuzz implements four types of FPGA-level optimizations to speed up fuzzing. For example, one such optimization is *dynamic kernel sharing in parallel fuzzing loops*, which enables a more effective search space exploration when utilizing multiple input generators, each with its own seed queue. HFuzz then invokes the target kernel function using a mutated input selected from one of the seed queues and dynamically increases the probability of choosing that input generator if the input yields new behavior signals at the hardware execution level. The other three micro-architecture level optimizations are *loop unrolling* which enables parallel iteration, *shannonization* which precomputes operations and reduces the latency of critical paths, and *data pre-loading* for fast memory access by moving data from global memory to local memory. HFuzz is the first to directly leverage the performance enhancing power of FPGA for automated testing of heterogeneous applications on an FPGA device.

We evaluate HFuzz's effectiveness on seven programs. These programs are from Intel's OneAPI benchmarks for heterogeneous applications with FPGA kernels [22]. We compare HFuzz against four alternatives: (Alternative 1: AFL-like) an AFL-like grey-box fuzzing tool that uses branch coverage as feedback and runs on the host entirely, (Alternative 2: HeteroFuzz) the state-of-the-art testing tool for heterogeneous applications using software monitors only, (Alternative 3: NoKernelMutation) HFuzz with CPU-side input mutation without offloading it to FPGA, and (Alternative 4: NoHWOptimization) HFuzz without FPGA-level optimizations. It took HFuzz much less time (i.e., 7%, 9.7%, 21.3%, and 29.4% of the time used by the four alternatives) to find the same number of defects. Given the same time budget (4 hours), HFuzz found 11×, 4.13×,

```
1  for(int s = 1; s <= nsteps; ++s) {
2    ...
3  // Kernel: calculate velocity
4      h.parallel_for(n, [=](item<1> i){
5        acc0=0; acc1=0; acc2=0;
6        #pragma unroll factor=2
7        for(int j=0; j<n; j++) {
8          if (j==i) {continue};
9          int8 dx, dy, dz;
10         dx = p[j].pos[0]-p[i].pos[0];
11         dy = p[j].pos[1]-p[i].pos[1];
12         dz = p[j].pos[2]-p[i].pos[2];
13         int8 sqr=dx*dx+dy*dy+dz*dz;
14         acc0+=(kG*p[j].mass/sqr)*dx;
15         acc1+=(kG*p[j].mass/sqr)*dy;
16         acc2+=(kG*p[j].mass/sqr)*dz;}
17         p[i].vel[0]+=acc0*dt;
18         p[i].vel[1]+=acc1*dt;
19         p[i].vel[2]+=acc2*dt;});});
```

Figure 2: Nbody-simulation: a heterogeneous version with DPC++ high-level synthesis.

2.36×, and 1.03× more defects than the four alternatives. We tried longer time (24 hours) but no more defect is found after 4 hours. Per the open science policy, we make HFuzz's artifacts, benchmark programs, and datasets available with this submission (uploaded with this submission).

In summary, this work makes the following contributions:

- To our knowledge, HFuzz is the first fuzz testing technique that uses hardware probes in tandem with software monitors to guide test input generation for heterogeneous applications.
- HFuzz is the first to unlock new micro-architecture level performance optimizations for fuzz testing by mapping both iterative input mutation and kernel invocation to FPGA-side computation. It implements four FPGA-level optimizations and accelerates fuzzing by 3.4×.
- HFuzz accelerates fuzz testing by 4.7× by directly synthesizing input mutations within kernels on FPGA. This also reduces the host-device data transfer overhead by 66%.
- With a 4-hour budget on seven benchmarks, HFuzz was able to discover 33 defects while traditional coverage-guided fuzzing only uncovered 3. Out of these 33 defects, 25 could not have been found without the use of device-side feedback.

2 BACKGROUND

2.1 Heterogeneous applications with FPGA

Driven by performance and energy benefits, heterogeneous computing applications [23] contain code that is executed on different kinds of processors such as CPU, GPU, and FPGA.

FPGAs are field programmable gate arrays. Modern FPGAs include millions of look-up tables (LUTs), thousands of embedded block memories (BRAMs), thousands of digital signal processing blocks (DSPs), and millions of flip-flop registers (FFs) [24]. Intel provides CPU+FPGA multi-chip packages; with its recent acquisition of Altera, such integration is expected to be even tighter in the future. FPGA has made its way into modern data centers, including Microsoft's Azure, Amazon F1, and Intel DevCloud [25–27].

A heterogeneous application typically consists of *host* code executed on the CPU and *kernel* code to be synthesized and executed on FPGA or GPU. Host code initializes the device, allocates the device memory, transfers data to the device, and invokes the compute-intensive kernel on the device side. After the execution, it transfers the kernel output back to the host and deallocates the memory.

To simplify kernel development, high-level-synthesis (HLS) [17, 28] lifts the abstraction of hardware development by automatically generating register-transfer level (RTL) descriptions from code written in C-like dialects. One example of HLS C/C++ dialects is Intel's Data Parallel C++ (DPC++), a cross-platform abstraction layer that enables code to be targeted to different CPUs, GPUs, and FPGAs [29, 30]. With DPC++, users can specify which hardware platform to implement a kernel on. For example, a user may use a compiler flag `-Xsboard=intel_s10sx_pac` to select Intel's FPGA S10. The user can develop a kernel function `f`, calling `h.parallel_for(n,f)` with a job handler `h`. This handler executes `f` with `n` degree parallelism on FPGA S10. Consider the following example.

2.2 An illustrating example: Nbody-simulation

Figure 2 illustrates the simulation of `n` particles moving over a sequence of `nsteps`. Lines 10-12 calculate the distance between particles, while Lines 14-16 calculate the acceleration. In lines 17-19, the program subsequently updates the particles' velocities based on the acceleration. These computations are extracted as compute-intensive kernels and offloaded to an FPGA. To enable parallelism and speed up the velocity calculation, the developer uses `h.parallel_for` and loop unrolling `#pragma unroll factor=2` (highlighted in red) at Line 4 and 6.

When writing a heterogeneous application, a user must conservatively estimate the limit of hardware resources and specify bidwidths for custom types and the size of buffers and pipes because all hardware resources are finite. Due to the need to statically specify hardware resources, a heterogeneous application often contains defects that cannot be detected statically via a compiler analysis. This is a problem that universally exists with all HLS languages. To illustrate, consider the real defects in the Nbody-simulation.

Divide By Zero in Nbody-Simulation. For code in Figure 2, with the input `p.pos=[(1,2,4),...,(1,2,4)]`, the velocity calculation on an FPGA A10 device produces absurdly large numbers `p.vel=[(-214748364,..),..]`. This is because, when the kernel inputs contain two particles with the same position, a divide-by-zero may happen inside the kernel in Lines 14-16 due to `sqr=0` at Line 13.

Overflow in Nbody-Simulation. When the kernel calculates the acceleration of two particles in Figure 2, an in-kernel overflow could occur if two particles are close to each other (i.e., `sqr`≈`0` at Line 13). This is because when `sqr` is close to zero, `acc` becomes large. When the inputs `p.pos=[(81,0,0),(81,1,0),(81,0,1),...]` are sent to the kernel, it produces a small value `sqr=1`, leading to overflow for the variables `acc1`; finally, the wrong result is sent back to the host.

State-of-the-Art. Grey-box fuzzing [21] generates program inputs based on per-iteration execution feedback. Suppose that a user uses grey-box fuzzing to monitor the value range of the inputs and outputs of kernels on the host-side (CPU) code. For the divide-by-zero bug that could occur in Figure 2, because `sqr` is an in-kernel variable and does not appear in the host code, software-side grey-box fuzzing [21] cannot easily reveal defects that originate from the inside of the kernel.

HFuzz addresses the limitations of existing work by utilizing hardware probes to monitor the intermediate states of kernels. HFuzz

identifies the in-kernel local variable `sqr` at Line 13 and inserts hardware probes to track its value range. The input generation process is then optimized by prioritizing inputs that result in new minimum or maximum values of `sqr`. As a result, HFuzz is able to effectively detect overflow when `sqr` reaches the small value `sqr=1` and divide-by-zero defects when `sqr` reaches its minimum value 0.

3 APPROACH

HFuzz aims to find inputs that can trigger both in-kernel errors and host-side errors for heterogeneous applications written in Intel's DPC++ HLS[18]. HFuzz contains three novel components that work in concert: (1) in tandem monitoring of software and hardware feedback by injecting software monitors and in-kernel probes (Section 3.1); (2) offloading input mutations to hardware kernels (Section 3.2), and (3) FPGA-level optimizations to speed up iterative input generation and kernel invocation (Section 3.3). HFuzz's design builds on two key insights. First, hardware-level parallelism can bring notable performance enhancement for iterative fuzzing, which is often characterized by independent task-level parallelism. Second, grey-box fuzzing's effectiveness can be significantly improved by observing signals from both hardware and software.

The Fuzzing Process. The overall workflow of HFuzz is shown in Algorithm 1. HFuzz takes as input a program p written in Intel's DPC++ and produces concrete inputs that trigger defects in p. HFuzz first applies a source-to-source transformation to p to produce an instrumented version $p\prime$, by inserting in-kernel probes and software monitors that can guide fuzz testing. HFuzz selects an input generator G from a set of generator S. It then randomly offloads a random seed input in' from G's seed queue into the kernels. To generate new inputs, HFuzz creates a new mutation kernel job in addition to the original kernel, and utilizes parallelism within FPGAs to mutate the input locally. The target function directly accesses the new input from local memory. In this process of input mutation and target execution, HFuzz incorporated four FPGA level optimizations for performance efficiency. As shown in Algorithm 1 Line 10-15, inputs that advances either software or hardware feedback are saved to the input queue for the next fuzzing iteration.

3.1 Injecting HW Probes in addition to SW Monitors

HFuzz, *for the first time*, directly introduces application-specific observability to hardware kernels by inserting hardware probes. It leverages these kernel probes in tandem with software-level monitors to form effective feedback signals to stretch heterogeneous application behavior.

Hardware Probes. While OS virtualization could provide the appearance of unbounded resources for the code executed on traditional CPUs, kernel functions are physically mapped to *resource-limited* heterogeneous architectures. This distinction leads to unique failures that are often induced by *resource limitations* on the device-side, which are not easily detectable when running software simulators. For example in Figure 2, a local variable `sqr` customizes regular integers to 8-bit integers for resource efficiency. Overflow conditions can occur if the variable's value exceeds its customized bitwidth. As another example, pipe saturation between two consecutive kernel functions can lead to read and write failures. In fact, such incorrect *intermediate computation states* within hardware kernels have been identified as the primary reason for hardware-originated bugs. HFuzz takes advantage of this observation, identifies local variables within kernels that hold intermediate states, and injects hardware probes to expose potential failures in kernel.

Algorithm 1: Fuzzing workflow.

```
   Input: program p, input generator set S, mutation operator set O
1  FuzzingLoop(p, S)
2  begin
3      p' = INSTRUMENT(p);
4      Feedback = ∅;
5      for 1..MAX do
6          G = S.select_input_generator();
7          in' = RANDOM_SELECT(G);
8          F_HW, F_SW =
             p'.host + in_kernel_mutate_execute(in', O);
9          for F ∈ {F_HW ⋃ F_SW} do
10             if F ∉ Feedback then
11                 INCREASE_PROB(S, G);
12                 good_input = REGENERATE(F.m, in');
13                 G = G ⋃ {good_input};
14                 Feedback = Feedback ⋃ {f};
15             end
16         end
17     end
18 end
   Input: kernel_input k_s, mutation_ops_set O
   Output: F_HW is a queue of triples (f, m, out) where f is
           kernel-feedback, m is mutation, and out is kernel output
19 In_Kernel_Mutate_Execute(k_s, O)
20 begin
21     for i = 1..MAX do
22         operator o = SELECT_OP(O);
23         start s = RANDOM_GENERATE();
24         end e = RANDOM_GENERATE();
25         mutation m = {(o, s, e)};
26         In_queue = In_queue ⋃ MUTATE_INPUT(o, s, e, k_s);
27     end
28     foreach in ∈ In_queue do
29         (f, m, out)=EXECUTEONDEVICE(in) ;
30         F_HW = F_HW ⋃ (f, m, out) ;
31     end
32     return F_HW
33 end
```

HFuzz automates the process of hardware probe insertion through source to source transformation, creating an instrumented kernel. From such instrumented kernel, intermediate states in the HW device are sent directly to the host code using dedicated host-kernel communication channels. The channels are implemented as global FIFO buffers and can be accessed from both the host and the kernel. The kernel side writes hardware feedback into the channels, while the host side reads information from the channels. Both read and write operations are non-blocking, in order to minimize any additional overhead to the original kernel logic. To expose intermediate computation states, HFuzz identifies in-kernel local variables and pipe usage via a C/C++ AST analysis [31]. As shown in Figure 3, in-kernel variable `sum` is highlighted in green, and pipe usage is highlighted in red. With a focus on in-kernel local variable and pipe

Table 1: Mutations accelerated by hardware.

Category	Description	SW Mutations	In-kernel Mutations	Average Speedup
M1 Sparsity Mutation	Replace non-zeros with zeros from index *s* to *e*, or do the opposite	`for i in s..e do {vector[i]=0}`	`# pragma unroll for i in s..e{vector[i]=0});`	4.31×
M2 Copy Mutation	Replace each element from index *s* to *e* with element at *s*	`for i in s..e do {vector[i]=vector[s]}`	`# pragma unroll for i in s..e{vector[i]=vector[s]});`	3.98×
M3Addition Mutation	Add constant *a* to each element from index *s* to *e*	`for i in s..e do {vector[i]+=a}`	`# pragma unroll for i in s..e{vector[i]+=a});`	3.21×
M4 Bit Mutation	Mutate an element with binary XOR given a constant *x*	`for i in s..e do {vector[i]^= (1«x)}`	`# pragma unroll for i in s..e {vector[i]^= (1«x)});`	4.42×

```
1  //First kernel...
2    h.parallel_for(range(M, P), [=](auto index) {
3      int sum = 0;
4      #pragma unroll factor=2
5      for (int i = 0; i < num_element; i++) {
6        sum += a[index[0]][i] * b[i][index[1]];
7        if (min_sum>sum) min_sum=sum;
8        if (max_sum<sum) max_sum=sum;}
9      bool flag;
10     KToKPipe::write(sum, flag);
11     KToKPipeSize++;//Pipe usage Probe
12     DeviceToHostKToKPipe::write(KToKPipeSize);
13     DeviceToHostMax_sum::write(max_sum);//sum's Value Range Probe
14     DeviceToHostMin_sum::write(min_sum);});});
15 //Second kernel...
16   h.single_task([=]() {
17   for (size_t i = 0; i < number_element; ++i) {
18     out[i] = KToKPipe::read();
19     KToKPipeSize--;//Pipe Usage Probe
20     DeviceToHostKToKPipe::write(KToKPipeSize);
21     out[i] = reciprocalTransform(output[i]); }});});
22 for(int i=0; i<number_elements; i++) {//SW monitor for kernel output
23   outmin=min(outmin, output[i]);
24   outmax=max(outmax, output[i]);}
```

Figure 3: Matrix transform: inserted value range probes are in the green rectangle. Inserted pipe usage probes are in the red rectangles. Inserted SW monitors are in the orange rectangle.

monitoring, HFuzz aims to uncover the two most commonly seen errors in custom hardware accelerators: overflows resulting from the resource and bitwidth finitization, as well as read/write failures caused by communication pipe saturations.

- *Value Range Probe*: HFuzz creates a value range monitor that checks the maximum and minimum value for each in-kernel variable. In Figure 3, HFuzz inserts probes on the intermediate variable `sum` which saves the cumulative sum of the product `a[index[0]][i]*b[i][index[1]]`. These probes monitor the minimum and maximum value of `sum`. HFuzz also constructs channels `DeviceToHostMax_sum` and `DeviceToHostMin_sum` to send these captured values back to the host at Line 13-14.
- *Pipe Usage Probe*: HFuzz creates a pipe usage monitor for each communication pipe. Consider the same example in Figure 3. HFuzz uses an AST analysis tool [31] to identify the locations of two kernel functions: `matrix_multiply` at Line 1-14 and `transformer` Line 16-21. We identify the variable name, `KToKPipe` used for pipe-based data transfer between the two kernels. By using `KToKPipe::write()` and `KToKPipe::read()`, the first kernel writes its result `sum` at Line 10 and the second kernel reads the value from this pipe at Line 18 in Figure 3. HFuzz applies source to source transformation to inject a counter-based usage monitor for this pipe and update the counter `KToKPipeSize` at Line 11 and Line 19 in Figure 3. Then HFuzz sends this counter value to the host by creating another direct communication channel, called `DeviceToHostKToKPipe` at Line 12 and Line 20.

Software Monitors. In addition to in-kernel probes, HFuzz inserts a set of software monitors on the host side, specialized to the custom FPGA accelerator synthesized on the device. We monitor: (1) the number of loop iterations, because it is related to pipelining and loop unrolling, common optimizations for parallelization implementation on FPGA; (2) the value range of each kernel input and output; (3) the kernel execution time, as hang or unexpectedly slow execution could be an indicator of failures. HFuzz retrieves the time and loop unrolling information from the HLS compilation report generated by DPC++. Besides, to monitor the value range of each kernel input and output, HFuzz inserts a value range monitor before and after each kernel, as shown in Line 22-24 of Figure 3.

3.2 Offloading input mutations to kernels

The traditional fuzzing process involves repeatedly mutating seed inputs and feeding them into a target program. The implicit assumption underlying such mutations is that seed inputs can be mutated and sent to the target program fast. Unfortunately, this assumption does not hold true for heterogeneous applications. Inputs to heterogeneous applications are often large matrices, leading to significant data transfer overheads between CPU and FPGA. We observe that local data transfer—data transfer within FPGAs, consumes less than 89% of the time required for data transfer between the fuzzer and the kernel. Additionally, in the process of fuzzing, a variety of *independent* mutation operations are frequently employed on small segments of the same seeds with the aim of exploring the input space. Thus, we can avoid repetitive data transfer by offloading the seed inputs to hardware kernels and mutating them directly within FPGAs. To achieve this, HFuzz creates a dedicated kernel for mutations *in parallel to* the original kernel, as well as a segment of on-chip memory for the storage of seeds and newly generated inputs. The mutation kernel and the original kernel function are both synthesized to the FPGA hardware concurrently. Table 1 shows four supported mutation operators. Because mutation operators are all order-independent and deterministic, HFuzz modifies all elements in the seed input at once. A resulting input can be re-generated given the seed and a concrete instance of mutation.

Consider Figure 3 as an example. The first kernel code computes the matrix product with two input matrices. We show how HFuzz tracks the feedback and mutates the input step by step in Table 2. With the initial seed input offloaded to the kernel, HFuzz tracks hardware feedback from the in-kernel variable `sum` at Line 2 by the inserted in-kernel probes in the green rectangle (column *Hardware Probes* in Table 2). After we apply the *M3 Addition Mutation* with

```
1  for (int i = s; i < e; i++) {
2    if (A[i]==0) {A[i] = generate_number(seed);}}
```

(a) Original mutation

```
1  int local_A[e-s];
2  #pragma unroll factor=4
3  for (int i = 0; i < e-s; i++) {local_A[i] = A[i+s];}
4  int t = generate_number(seed);
5  for (int i = 0; i < e-s; i++) {
6    if (local_A[i]==0) {
7    local_A[i] = t;
8    t = generate_number(seed);}}
9  #pragma unroll factor=4
10 for (int i = 0; i < e-s; i++) {A[i+s] = local_A[i];}
```

(b) Optimized mutation in kernel

Figure 4: Sparsity mutation: replace the zero elements to non-zero elements from index s to index e.

loop unrolling optimization, from the starting offset s=1 to the ending offset e=4 on array a, a greybox fuzzer that only monitors the value range for the kernel interface variables a and b would discard the input `[-20,5,7,7,9,20]` because it does not achieve a new value spectra at the software level. However, HFuzz saves the corresponding mutation information, since this input registers a new feedback at the hardware level for the in-kernel variable sum.

3.3 FPGA optimizations for fuzzing

Traditional fuzz testing can be naïvely applied to heterogeneous applications by treating hardware kernel invocations as equivalent to software function calls. However, such straightforward application of software-style fuzzing results in severe performance inefficiencies. In heterogeneous applications, there is a distinct *opportunity* to utilize hardware micro-architecture level optimizations to accelerate the traditional fuzzing process. Both iterative matrix mutations and target executions involve independent tasks, enabling task-level parallelism.

HFuzz applies four FPGA optimizations to accelerate iterative matrix mutations and target execution, including loop unrolling, shannonization, local memory access, and dynamic kernel sharing. These optimizations are not specific to HFuzz or Intel's heterogeneous architecture, and thus also are applicable to other applications on other FPGAs. For instance, loop unrolling is a technique that can be used to optimize iterative computations that do not have significant data dependencies between iterations, and it can be applied independently of the specific FPGA platform.

1. Dynamic kernel sharing. In traditional fuzzing, the difficulty of testing often arises from the need to explore deep branches within the program. However, when testing heterogeneous applications, errors tend to occur due to variations in the range of values for in-kernel variables and resource usage. This presents a significant challenge of rapid input space exploration especially when inputs are large matrices.

We propose a dynamic, probabilistic kernel-sharing method to interleave the exploration of input search space originating from multiple seeds in heterogeneous applications. To implement this method, HFuzz employs four input generators that share the same target kernel and each has its own seed queue. These input generators start with different seed inputs and, during each iteration, one generator is chosen based on an activation probability array. The selected generator then picks a seed input from its queue, mutates it within the kernel, and sends the generated input to the target kernel function via on-chip memory on the device. If the generated input results in new feedback, it is saved in the generator's seed queue for use in future fuzzing iterations.

HFuzz utilizes an adaptive approach to input generation by selecting an input generator and its associated seed queue based on an activation probability array. The selection process involves evaluating the performance of each generator and adjusting its probabilities accordingly. For instance, if a new input generated by generator G results in new feedback, it will be considered a favored generator and its activation probability will be increased. Otherwise, it will be labeled as an inactive generator and its activation probability will be decreased. This approach allows for efficient input space exploration and ensures that the test generation is focused on areas that are likely to yield new feedback:

$$P_G = \begin{cases} P_G + \alpha & \text{if } G \text{ is chosen and HFuzz gets new feedback} \\ P_G - \frac{\alpha}{l-1} & \text{if } G \text{ is not chosen and HFuzz gets new feedback} \\ P_G - \alpha & \text{if } G \text{ is chosen and HFuzz gets no new feedback} \\ P_G + \frac{\alpha}{l-1} & \text{if } G \text{ is not chosen and HFuzz gets no new feedback} \end{cases} \quad (1)$$

In our experiment, we set the number of generators l to be 4. The initial activation probability for each generator P_G is set to $1/l = 0.25$. The update factor α is predefined as 0.05. In Table 2, in the second execution (ID 2), inputs generated by generator G increased the hardware monitor range. As a result, HFuzz increases the activation probability of G from 0.25 to $0.25+\alpha = 0.3$.

2. Data preloading [32]. Matrix mutation on large matrices requires a significant amount of data read and write operations. To improve efficiency, it is crucial to minimize memory access time for input vectors or matrices. Many heterogeneous computing systems, such as Intel oneAPI, have both *global memory* that can be accessed by both kernel and host code, and on-chip *local memory* that is only accessible by kernel code. Accessing local memory within the kernel typically has a shorter latency than accessing global memory. We thus apply data preloading to transfer data from global memory to local memory.

In Figure 4b, HFuzz reduces memory access costs (highlighted in red) by transferring data from array A to the local array local_A. This results in a reduction of memory access costs, as seen at Line 6-7 in the optimized code, compared to the original code in Figure 4a at Line 2. This optimization leads to a 1.31x speedup in the mutation process.

3. Shannonization [33]. Sparsity mutation replaces zero elements with non-zero elements. It necessitates the implementation of a null check for each element in the matrix. As shown in Line 2 of Figure 4a, an `if` statement is added to accomplish this. However, this `if` statement induces extra hardware overhead, as it increases the delay in the critical path. Each time the `if` condition is satisfied (i.e. `A[i]==0`), the operation `generate_number` needs to be computed, which can slow down the overall performance.

Table 2: Example execution of input generator G.

ID	Mutation Operator	Kernel Inputs	Variable	Hardware Probes Min	Hardware Probes Max	Software Monitors Min	Software Monitors Max	New Value Range	Over-flow	Save Input	Memorization HW Range	Memorization SW Range	P_G
Seed	N/A		sum	-56	168			N/A	No	N/A	[-56,168]		0.25
		a[][1]=[-20, 2, 4, 4, 6, 20]	a			-20	20	N/A				[-20, 20]	
		b[1][]=[1, -10, -4, -14, 28, 0]	b			-14	28	N/A				[-14, 28]	
1	**M3**		sum	-202	54			**Yes**	No	**Yes**	**[-202, 168]**		0.3
	start s=1	a[][1]=[-20, 5, 7, 7, 9, 20]	a			-20	20	No				[-20, 20]	
	end e=4	b[1][]=[1, -10, -4, -14, 28, 0]	b			-14	28	No				[-14, 28]	
2	**M2**		sum	-70	140			No	No	No	[-202, 168]		0.25
	start s=1	a[][1]=[-20, 5, 5, 5, 5, 20]	a			-20	20	No				[-20, 20]	
	end e=4	b[1][]=[1, -10, -4, -14, 28, 0]	b			-14	28	No				[-14, 28]	
3	**M3**		sum	20	-140			**No**	**Yes**	**Yes**	**[-202, 168]**		0.3
	start s=1	a[][1]=[-20, 8, 10, 10, 12, 20]	a			-20	20	No				[-20, 20]	
	end e=4	b[1][]=[1, -10, -4, -14, 28, 0]	b			-11	28	No				[-14, 28]	

Shannonization improves performance by precomputing operations within a loop and removing them from the critical path. In this example, HFuzz applies shannonization (highlighted in green in Figure 4b) by precomputing the operation `generate_number` at Line 4, and removing it from the critical path inside the branch at Line 6. Then HFuzz precomputes the next value of `t = generate_number` at Line 8 for a later iteration of the loop to use when required (that is, the next time `local_A[i]==0`). This precomputation can be done simultaneously within the loop, allowing for a reduction in the critical path delay and leading to a 1.24x speedup in the sparsity mutation process.

4. Loop unrolling [34]. Software-style mutations on large vectors and matrices are often performed by modifying one or some particular elements. Line 2 in Figure 4a shows an example mutation based on a `for` loop. Such direct application of loops on hardware neglects the potential for hardware parallelism, resulting in inefficient use of hardware resources.

Loop unrolling improves performance by creating multiple copies of the loop body, thus the required number of iterations is reduced. In the example shown in Figure 4b, the `#pragma unroll` directive (highlighted in orange) causes the kernel to unroll the loop by a factor of 4, as specified by the `factor=4` argument. The compiler then expands the pipeline by quadrupling the number of operations and loading three times more data. This results in a 4x speedup of the loop process.

4 EVALUATION

We evaluate the following research questions:

RQ1 How much improvement in defect detection capability is achieved by incorporating both device-side feedback and host-side feedback in HFuzz?

RQ2 How much speed-up is achieved by in-kernel input mutations?

RQ3 How much speed-up is achieved by the FPGA-level optimizations for fuzzing?

RQ4 How much overhead is incurred by injecting hardware probes in HFuzz?

To assess the improvement in defect detection and fuzzing acceleration, we compare HFuzz against four baselines.

(1) Alternative 1 AFL-like: This option uses a branch-coverage guided fuzzing similar to AFL and performs input mutations on CPU side.

(2) Alternative 2 HeteroFuzz: This option is a replication of the state-of-art work HeteroFuzz [21] for Intel DPC++. Compared to HFuzz, it does not have in-kernel probes on FPGA devices and considers only software monitoring feedback.

(3) Alternative 3 NoKernelMutation: This option disables in-kernel mutations and performs input mutations on the CPU.

(4) Alternative 4 NoHWoptimization: This option disables hardware optimizations and only uses one input queue instead.

Benchmarks. We choose seven applications from Intel's OneAPI GitHub repositories [22]: (R1) Matrix-transform. It has two kernels—one for matrix multiplication M=A*B and the other for reciprocal transformation on each element of M; (R2) Matrix-mul: multiplication of two matrices; (R3) Complex-mul: multiplication of two vectors of complex numbers in parallel; (R4) APSP: the Floyd-Warshall algorithm to find the shortest path between the pairs of vertices in a graph; (R5) Nbody-sim: Simulation of a dynamical system of particles under the influence of gravity; (R6) Hidden-Markov-model: a statistical model using a Markov process; (R7) Match-num: reading data from the host and sending the numbers that match a set of pre-defined constants back to the host.

These benchmarks are widely used in hardware acceleration literature [13] and cover a representative set of optimizations used in kernels (e.g., custom bitwidth, loop unrolling, etc.) and exhibit different memory usage patterns (e.g., buffer memory and unified shared memory for kernel input and output, kernel-to-kernel pipe and kernel-to-host pipe, local memory for in-kernel variables, etc.). Testing difficulties for heterogeneous applications do not depend on the code size; rather, it depends on how hardware resources are synthesized (e.g., in-kernel variables, loop unrolling) and the communication channel details between software and hardware and between hardware kernels. These benchmarks' kernels are widely used and their code size is similar to commercial HLS benchmarks. They are complex in both optimizations and memory arrangements and hard to get right.

Experimental Environment. All experiments were conducted on Intel DevCloud A10 nodes [27]. The automated kernel probe insertion was implemented using DPC++ compiler and Pycparser [31]. The refactored programs were synthesized to RTL and targeted to Intel Arria 10 GX FPGA [35]. We also tried HFuzz on other FPGAs like Intel Stratix 10 SoC FPGA [36] and achieved similar results.

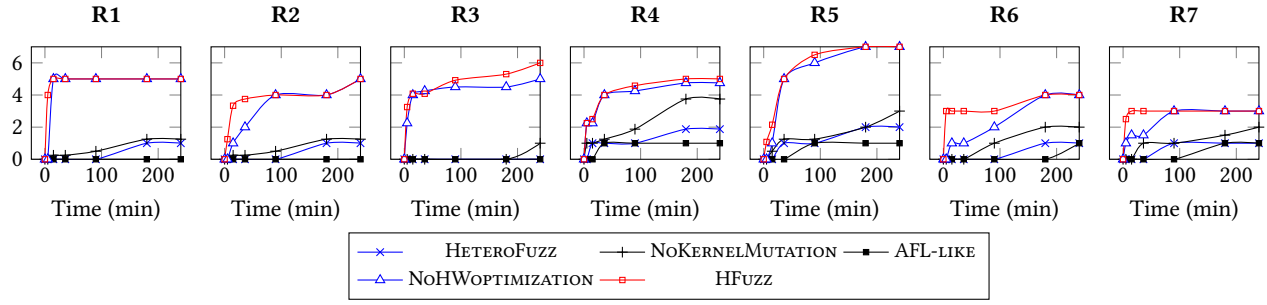

Figure 5: # Number of Defects

Table 3: Example symptoms of kernel defects in R1.

ID	Symptom	Description	HeteroFuzz Find
S1	Kernel Runtime Overflow	The value of intermediate variables `sum` at line 2 of Figure 3 exceeds its bitwidth capacity, leading to a wrong result.	✓
S2	Pipe Write Failure	Pipe write failure happens when FPGA attempts to write into pipe when pipe is full.	×
S3	Pipe Read Hang	Pipe read hang happens when FPGA attempts to read synchronously from an empty pipe.	×
S4	Division by Zero	`sum` in line 5 of Figure 3 equals 0, leading to divide by zero at line 21.	×
S5	Incorrect Loop Unrolling	CPU and FPGA produce different results when the input array size `num_element` is not multiple of 2.	✓

4.1 Defect detection by HW and SW feedback

We assess the effectiveness of HFuzz's feedback guidance by comparing the number of defects detected through combined hardware probes and software monitors to that of HeteroFuzz, which relies solely on software monitors. For each benchmark, we generate test inputs using HFuzz and HeteroFuzz for 4 hours. We tried longer time (24 hours) but no more defect is found after 4 hours. Using the generated inputs, we then perform differential testing between CPU-only executions and CPU+FPGA executions and measure the number of defects (i.e., diverging outcomes) found.

Figure 5 shows the average experimental results from ten runs. HFuzz is able to detect 3.1× more defects than HeteroFuzz. For example, for R5 Nbody-simulation, without monitoring in-kernel variable `sqr`, HeteroFuzz cannot find divide-by-zero error we mentioned in Section 2.2 at Line 16-18 in Figure 2. When using HeteroFuzz, the value range of kernel inputs does not reflect the change in the square of distance between particles `sqr`. HFuzz, instead, directly monitors the value range of in-kernel variable `sqr`, and finds the defects when `sqr` reaches its minimum value 0. In total, HeteroFuzz finds 8 unique defects in 16.5 hours, while HFuzz finds the same defects in 1.6 hours—almost 90% reduction in the testing time.

Table 3 lists five defects found by HFuzz in R1 `Matrix-transform`.

First, S1 shows an overflow occurred in the FPGA execution due to the in-kernel variable `sum` at Line 3 in Figure 3. It happens when the input vector `a` includes a large number such as 2090401586. By monitoring in-kernel variable `sum`'s value range, HFuzz increases the chance of generating a new vector with large numbers.

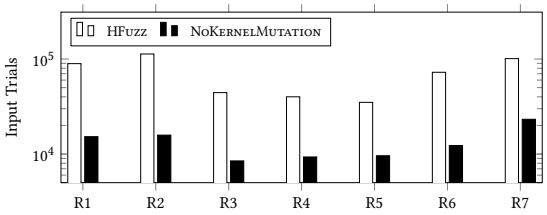

Figure 6: Number of Input Trials

Second, two kernels in R1 use a 128-byte pipe to facilitate direct data transfer. As mentioned in Section 1, when the first kernel produces results faster than the second kernel can consume, the pipe may become saturated. Consequently, a pipe write failure occurs silently and the newly written value is lost, shown as S2 in Table 3. This may further lead to another defect S3: pipe read hang. The second kernel in Figure 3 reads values from the pipe for `number_elements` times. However, if the number of values successfully written to the pipe is less than `number_elements`, the second kernel will hang at this pipe read. Both defects cannot be detected by prior work HeteroFuzz because host-side software monitors cannot detect the saturation of commutation pipes.

Third, S4 depicts a divide-by-zero error caused by the intermediate result `sum` in the second kernel `reciprocalTransform` at Line 21 in Figure 3. It happens when both two input matrices are sparse matrices. On CPU, this execution may raise a division-by-zero exception; however, it silently returns an unexpected number on FPGA instead. By monitoring `sum`'s value range, HFuzz triggers this defect by generating inputs using *Sparsity Mutation*.

Fourth, since R1 makes two copies of the loop body at Line 4 in Figure 3 by using `#pragma unroll factor=2`, a wrong result happens if the number of loop iterations `num_elements` is not a multiple of the unroll factor 2.

> HFuzz achieves 10.3× speed-up and finds 25 new defects compared to HeteroFuzz, demonstrating the combined benefit of hardware probes and software monitors.

4.2 Speed-up from in-kernel input mutations

To assess speed-up enabled by offloading input mutations to FPGA devices, we compare HFuzz with a downgraded version NoKernelMutation. We measure the number of generated inputs and defects found within the same 4-hour budget.

Figure 6 reports the average number of input trials within 4 hours. For example, in R7, NoKernelMutation generates 23225 inputs, while HFuzz generates 100918 inputs (5.3× speed-up) by avoiding redundant data transfer and parallelizing input mutations. In R2,

NoKernelMutation and HFuzz enumerate 15824 and 112940 inputs respectively, leading to 7.1× speed-up. R2 achieves higher speedup than R7 because its performance is more dominated by data transfer as shown in Figure 1.

Figure 5 shows the number of defects found by NoKernelMutation. While NoKernelMutation reports 14 unique defects in 24 hours, HFuzz detects the same defects in 5.1 hours, which translates to 4.7× speed-up in defect detection. These defects are not found by NoKernelMutation, because it wastes time in sequentially mutating inputs in CPU and sending the large data to the kernel.

> HFuzz reduces the need for data transfer by offloading mutations into kernels and speeds up fuzzing by 4.7×.

4.3 Speed-up from FPGA-level optimizations

To evaluate the effectiveness of FPGA-level optimizations for input generation, we created a downgraded version of our tool NoHWoptimization, which disables this feature. We evaluated the time taken to find the same defects. The results are shown in Figure 5. Compared to NoHWoptimization, HFuzz finds the same 33 bugs 3.4x faster, taking only 8.3 hours as opposed to 28 hours.

In R1 (e.g., Figure 3), the detected defects include (1) a divide-by-zero error when the kernel takes as input two sparse matrices and (2) an overflow error when the kernel takes as input two dense matrices with large elements. Because inputs leading to these defects are distinct from each other, traditional mutational fuzzers with a single input queue may be inefficient to find them. In fact, it takes 2 hours to mutate two sparse matrices into dense ones. HFuzz uses one hardware optimization technique, called dynamic kernel sharing, to enable simultaneous exploration of input subspaces originating from different seeds. For that, HFuzz utilizes multiple input generators. One generator A starts with dense matrices and another generator B starts with sparse matrices. HFuzz can detect these two bugs by interleaving generator A and generator B based on runtime feedback. For example, when generator A reaches its maximum value and triggers an overflow, it can no longer provide any new feedback. HFuzz will switch to generator B and detect the divided-by-zero error. HFuzz reduces the detection time to 5 mins.

> HFuzz achieves 3.4× speed-up in the detection of detects by implementing hardware optimizations. Loop unrolling, shannaization, and fast memory access directly speed up the mutation process. Dynamic kernel sharing enables efficient input space exploration.

4.4 Probe Overhead

Inserting hardware probes into the original kernels may cause extra overhead on hardware resources, as reported in Table 4. We measure four types of hardware resource, including ALUT (a lookup table implementing the boolean function), FF (flip flops for storing temporary data), RAM (random access memory blocks), and DSP (a digital signal processing unit for common fixed-point and floating-point arithmetics). In general, the overhead depends on the complexity of the original kernels. In R2, compared to the original kernel with 9592 ALUTs and 14466 FFs, inserted probes used 22% more ALUTs and 33% more FFs. For a relatively complex kernel R4, the overhead is 6% ALUT and 10% FFs. The extra resource usage mainly comes from (1) the probe computation including read and write, and (2) the kernel dispatch logic establishes the communication between kernel and host.

Table 4: Resource overhead from injecting hardware probes.

ID/Program		#LUT	#FF	#RAM	#DSP	Freq /MHz
R1/ Matrix_trans	Orig	15932	25088	137	4.5	247
	Probe	17905	34320	192	4.5	246
R2/ Matrix_mul	Orig	9592	14466	492	16	259
	Probe	12032	19443	492	16	247
R3/ Complex_mul	Orig	11545	18494	106	6	273
	Probe	11203	27117	106	6	253
R4/ APSP	Orig	60468	92249	555	195	221
	Probe	64327	101229	558	195	212
R5/ Nbody_sim	Orig	23642	44352	309	34	270
	Probe	27612	50549	317	34	260
R6/ HMM	Orig	48706	64987	395	67	257
	Probe	56562	87392	491	67	247
R7/ Match_num	Orig	2239	1357	67	12	279
	Probe	3828	2033	73	12	259

Such overhead could be further reduced by manual optimizations. For example, Curreri [20] performs resource sharing by using the same FIFO probe for multiple feedback signals.

> Hardware probe insertion uses 24% extra LUT, 29% extra FF, 8% extra RAM, and reduces frequency by 5% on average. However, it enables an overall 10.3× speed-up in defect detection by providing hardware feedback.

5 RELATED WORK

Fuzz Testing. Traditional fuzzing starts from a seed input, runs the program on the selected input, generates new inputs by mutating the previous input, and adds new inputs to the queue if they improve a given guidance metric such as branch coverage. Instead of using coverage as guidance, several techniques use custom guidance mechanisms. UAFL [37] incorporates typestate properties and information flow analysis to detect the use-after-free vulnerabilities. BigFuzz [38] monitors dataflow operator coverage in tandem with branch coverage for dataflow-based analytics. For example, MemLock [39] employs both coverage and memory consumption metrics. AFLgo [40] extends AFL to direct fuzzing towards user-specified target sites. PerfFuzz [41] uses the execution counts of exercised instructions together with branch coverage to identify inputs revealing pathological performance. HeteroFuzz [21] generates concrete test inputs for heterogeneous applications to perform differential testing between CPU vs. CPU+FPGA. Unlike HFuzz, HeteroFuzz treats the kernels as black boxes and performs software-level monitoring only. All these techniques rely on pure software-level feedback either at the level of code coverage or using custom monitors. None leverages hardware probes in tandem with software monitors to guide test input generation, like HFuzz.

A fuzzing loop consists of multiple invocations of a target program with different inputs in an independent manner; thus, it provides a natural opportunity for parallelism. AFL++ [42] injects a fork server, which tells the target to fork itself to run, and thus realizes parallel fuzzing across multiple CPU cores or across a fleet of systems. For example, P-Fuzz [43] distributes unique seeds to run fuzzing in parallel, and PAFL [44] maintains global and local guiding information for synchronizing parallel fuzzing jobs. While these techniques accelerate fuzz testing via distributed computation on CPU, unlike HFuzz, none accelerates fuzzing by using FPGAs.

HFuzz pushes iterative input mutation directly to an FPGA kernel, and benefits from the massive hardware parallelism intrinsic to FPGA during iterative testing of heterogeneous applications.

Coverage-guided greybox fuzzing adds test cases into the set of seeds if they exercise the new path or new behavior. However, most seeds exercise the same "high-frequency" paths. To explore more paths with the same number of tests, researchers develop strategies to select seeds wisely. AFLFast [45] models coverage-based greybox fuzzing as a Markov chain, and assigns different selection probabilities for different seeds. EcoFuzz [46] improves AFLFast's Markov chain model and presents a variant of the Adversarial Multi-Armed Bandit model. EcoFuzz sets three states of the seeds set and develops a unique adaptive scheduling algorithm. While these techniques select seeds based on probabilities, none of them leverages FPGA-level optimizations to speed up seed selection with dynamic kernel sharing.

High Level Synthesis & In-Circuit Debugging. To ease the development of heterogeneous applications, HLS tools automatically generate RTL descriptions from C/C++ programs. To help debugging HLS-generated circuits, *Inspect* [47] introduces software debugger-like capabilities, including gdb-like breakpoints, step, and data inspection. It tracks file names and line numbers in HLS code, so that HW probes at the level of wires and registers could be linked to specific lines in the HLS code. A user can monitor each variable for its data width and the number of elements in an array. Monson and Hutchings [48] design a debugger for HLS-generated FPGA-based circuits via source instrumentation by connecting C expressions to top-level ports that serve as debug signals. HLScope [49] is a performance debugger that traces the cause of stalls for HLS-generated circuits. Curreri et al. realize in-circuit assertions for timing analysis and stall-relate bugs [20]. While these debuggers and HFuzz leverage a similar mechanism of injecting HW probes, HFuzz's goal is different—it improves the effectiveness of grey-box fuzzing for heterogeneous applications by designing meaningful monitors at both software and hardware levels.

In the hardware design community, *circuit verification*, including formal verification and runtime verification, has been used to validate code written in hardware description languages (Verilog, VHDL, etc.). For example, RFUZZ [50] is a circuit-level input generator for FIRRTL IR (UC Berkeley's RTL variant). RFUZZ invents a notion of *MUX toggle coverage* for circuit testing at the gate level and employs a rapid memory resetting on FPGA for RTL circuit verification. However, their monitors are gate-level and not application-specific. Qin and Mishra present a scalable test generation technique [51] for hardware kernels in Verilog by interleaving concrete and symbolic execution to bridge the gap between model checking and testing. Kourfali and Stroobandt [19] exploit parameterization of LUTs and routing infrastructures in an FPGA to create a virtual debugging overlay network inside circuits. These circuit testing and verification techniques find bugs in kernels at RTL level, while HFuzz targets *end-to-end testing of heterogeneous applications written in HLS*. In other words, it is not feasible to directly compare HFuzz against these in-circuit verification techniques.

FPGA Performance Optimizations. Ma et al. explored various loop optimization techniques, such loop tiling, loop interchange, and loop unrolling to reduce memory consumption and data movement when mapping deep convolutional neural networks [52] to FPGA. Zhang et al. adopt data buffering techniques to hide the memory access latency and interconnects, avoiding data transfer overhead from the global memory to FPGAs on-chip memory [53]. Li et al. [54] use pipeline optimizations when mapping layer-by-layer computation to multiple FPGAs resources. Pipelining can increase hardware utilization and achieve high throughput by preventing the computing engines to become idle due to imbalanced computation speed across layers. Other widely used kernel optimizations include I/O optimization by sharing resources among computation tasks at different time stamps. Another optimization is *retiming*, which moves edge-triggered registers across combinatorial gates or LUTs to improve timing while ensuring identical behavior, etc [55]. Inspired by these FPGA-level performance optimizations, HFuzz designs four unique FPGA-level optimizations to accelerate the combined computation of input generation and kernel invocation: dynamic kernel sharing, shannonization, loop unrolling, and data buffering. HFuzz is a pioneering tool—the first to embody FPGA-level optimizations to enhance fuzzing efficiency and effectiveness for heterogeneous applications.

SNAP [56] leverages the existing CPU pipeline and hardware features to optimize the bitmap update required for coverage-guided testing. As opposed to SNAP that targets fuzzing traditional programs running on a CPU and simply uses existing hardware features as a black box acceleration aid, HFuzzHFuzz designs new FPGA-level optimizations for mapping input generation and kernel invocation to FPGAs and empirically demonstrates significant fuzzing speed-up from these optimizations (3.4×).

6 CONCLUSION

In recent years, performance improvement in CPU has slowed significantly to only a few percent—due to challenges in power supply scaling, heat dissipation, space and cost. This trend necessitates the needs to embrace heterogeneous computer architectures such as GPU and FPGA. In particular, FPGA is a promising, *reprogrammable* alternative for improving performance and energy efficiency. However, due to the lack of observability into FPGA execution and complex interaction between CPU and kernel execution on FPGA, developing and testing heterogeneous applications is extremely inaccessible to regular software engineers.

HFuzz is the first grey-box testing approach leverages the *capability of heterogeneous hardware* for testing *heterogeneous applications*. In particular, HFuzz injects hardware probes in addition to injecting software monitors to better guide input generation and offloads iterative input generation to hardware accelerators. HFuzz speeds up fuzzing by offloading input mutations to FPGAs by 4.7× without sacrificing any defect detection capability. It speeds up testing 10.3× on average by gathering meaningful signals from hardware execution directly by injecting in-kernel probes. This work fits the domain of software testing, as it targets HLS C/C++ dialects and it has the potential to significantly improve correctness in the new era of *heterogeneous computing*, where regular software developers write code in HLS C/C++ to exploit custom hardware acceleration.

7 DATA AVAILABILITY

Per the open science policy, we make HFuzz's artifacts, benchmark programs, and datasets available with this submission (uploaded with this submission).

REFERENCES

[1] A. A. Chien, A. Snavely, and M. Gahagan, "10x10: A general-purpose architectural approach to heterogeneity and energy efficiency," *Procedia Computer Science*, vol. 4, pp. 1987–1996, 2011.

[2] J. Cong, M. A. Ghodrat, M. Gill, B. Grigorian, K. Gururaj, and G. Reinman, "Accelerator-rich architectures: Opportunities and progresses," in *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2014, pp. 1–6.

[3] J. Cong, V. Sarkar, G. Reinman, and A. Bui, "Customizable domain-specific computing," *IEEE Design Test of Computers*, vol. 28, no. 2, pp. 6–15, 2011.

[4] J. Casper and K. Olukotun, "Hardware acceleration of database operations," in *Proceedings of the 2014 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 151–160. [Online]. Available: https://doi.org/10.1145/2554688.2554787

[5] J. Cong, L. Guo, P.-T. Huang, P. Wei, and T. Yu, "Smem++: A pipelined and time-multiplexed smem seeding accelerator for dna sequencing," in *2018 IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2018, pp. 206–206.

[6] A. M. Caulfield, E. S. Chung, A. Putnam, H. Angepat, J. Fowers, M. Haselman, S. Heil, M. Humphrey, P. Kaur, J.-Y. Kim, D. Lo, T. Massengill, K. Ovtcharov, M. Papamichael, L. Woods, S. Lanka, D. Chiou, and D. Burger, "A cloud-scale acceleration architecture," in *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2016, pp. 1–13.

[7] A. Putnam, A. M. Caulfield, E. S. Chung, D. Chiou, K. Constantinides, J. Demme, H. Esmaeilzadeh, J. Fowers, G. P. Gopal, J. Gray, M. Haselman, S. Hauck, S. Heil, A. Hormati, J.-Y. Kim, S. Lanka, J. Larus, E. Peterson, S. Pope, A. Smith, J. Thong, P. Y. Xiao, and D. Burger, "A reconfigurable fabric for accelerating large-scale datacenter services," *Commun. ACM*, vol. 59, no. 11, p. 114–122, Oct. 2016. [Online]. Available: https://doi.org/10.1145/2996868

[8] L. Guo, J. Lau, Z. Ruan, P. Wei, and J. Cong, "Hardware acceleration of long read pairwise overlapping in genome sequencing: A race between fpga and gpu," in *2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2019, pp. 127–135.

[9] I. Cutress, "Intel shows xeon scalable gold 6138p with integrated fpga, shipping to vendors," https://www.anandtech.com/show/12773/intel-shows-xeon-scalable-gold-6138p-with-integrated-fpga-shipping-to-vendors, May 2018.

[10] P. Alcorn, "Amd to fuse fpga ai engines onto epyc processors, arrives in 2023," https://www.tomshardware.com/news/amd-to-fuse-fpga-ai-engines-onto-epyc-processors-arrives-in-2023, May 2022.

[11] V. Manes, H. Han, C. Han, s. cha, M. Egele, E. Schwartz, and M. Woo, "The art, science, and engineering of fuzzing: A survey," *IEEE Transactions on Software Engineering*, vol. PP, pp. 1–1, 10 2019.

[12] M. Zalewski, "American fuzz loop," http://lcamtuf.coredump.cx/afl/, 2021.

[13] H. Rong, "Programmatic control of a compiler for generating high-performance spatial hardware," *CoRR*, vol. abs/1711.07606, 2017. [Online]. Available: http://arxiv.org/abs/1711.07606

[14] Y.-H. Lai, E. Ustun, S. Xiang, Z. Fang, H. Rong, and Z. Zhang, "Programming and synthesis for software-defined fpga acceleration: Status and future prospects," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 14, no. 4, sep 2021. [Online]. Available: https://doi.org/10.1145/3469660

[15] K. Rupnow, Y. Liang, Y. Li, and D. Chen, "A study of high-level synthesis: Promises and challenges," in *2011 9th IEEE International Conference on ASIC*, 2011, pp. 1102–1105.

[16] D. F. Bacon, R. Rabbah, and S. Shukla, "Fpga programming for the masses," *Commun. ACM*, vol. 56, no. 4, p. 56–63, apr 2013. [Online]. Available: https://doi.org/10.1145/2436256.2436271

[17] J. Cong, B. Liu, S. Neuendorffer, J. Noguera, K. Vissers, and Z. Zhang, "High-level synthesis for fpgas: From prototyping to deployment," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 4, pp. 473–491, 2011.

[18] Intel, "Dpc++ reference," https://oneapi-src.github.io/DPCPP_Reference/, 2021.

[19] A. Kourfali and D. Stroobandt, "In-circuit debugging with dynamic reconfiguration of fpga interconnects," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 13, no. 1, jan 2020. [Online]. Available: https://doi.org/10.1145/3375459

[20] J. Curreri, G. Stitt, and A. D. George, "High-level synthesis techniques for in-circuit assertion-based verification," in *2010 IEEE International Symposium on Parallel Distributed Processing, Workshops and Phd Forum (IPDPSW)*, 2010, pp. 1–8.

[21] Q. Zhang, J. Wang, and M. Kim, "Heterofuzz: Fuzz testing to detect platform dependent divergence for heterogeneous applications," in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2021, pp. 242–254.

[22] Intel, "Base: Vector add sample," https://github.com/oneapi-src/oneAPI-samples/tree/master/DirectProgramming/DPC%2B%2B/DenseLinearAlgebra/vector-add, 2021.

[23] A. R. Brodtkorb, C. Dyken, T. R. Hagen, J. M. Hjelmervik, and O. O. Storaasli, "State-of-the-art in heterogeneous computing," *Scientific Programming*, vol. 18, no. 1, pp. 1–33, 2010.

[24] Xilinx, "Ultrascale architecture and product data sheet: Overview," https://www.xilinx.com/support/documentation/data_sheets/ds890-ultrascale-overview.pdf, 2021.

[25] M. Zahran, "Heterogeneous computing: Here to stay," *Communications of the ACM*, vol. 60, no. 3, pp. 42–45, 2017.

[26] Amazon.com, "Amazon ec2 f1 instances: Run custom fpgas in the aws cloud," https://aws.amazon.com/ec2/instance-types/f1, 2021.

[27] Intel, "Devcloud," https://www.intel.com/content/www/us/en/developer/tools/devcloud/overview.html, 2022.

[28] D. D. Gajski, N. D. Dutt, A. C. Wu, and S. Y. Lin, *High—Level Synthesis: Introduction to Chip and System Design*. Springer Science & Business Media, 2012.

[29] R. Reyes and V. Lomüller, "Sycl: Single-source c++ accelerator programming," in *Parallel Computing: On the Road to Exascale*. IOS Press, 2016, pp. 673–682.

[30] J. Reinders, B. Ashbaugh, J. Brodman, M. Kinsner, J. Pennycook, and X. Tian, *Data parallel C++: mastering DPC++ for programming of heterogeneous systems using C++ and SYCL*. Springer Nature, 2021.

[31] E. Bendersky, "Pycparser c parser and ast generator written in python," 2012.

[32] Intel, "Fpga optimization guide for intel® oneapi toolkits - transfer loop-carried dependency to local memory," https://www.intel.com/content/www/us/en/develop/documentation/oneapi-fpga-optimization-guide/top/optimize-your-design/throughput-1/single-work-item-kernels/loops/transfer-loop-carried-dependency-to-local-memory.html, 2022.

[33] ——, "Fpga optimization guide for intel® oneapi toolkits - shannonization to improve fmax/ii," https://www.intel.com/content/www/us/en/develop/documentation/oneapi-fpga-optimization-guide/top/optimize-your-design/throughput-1/single-work-item-kernels/loops/shannonization-to-improve-fmax-ii.html, 2022.

[34] ——, "Fpga optimization guide for intel® oneapi toolkits - unroll loops," https://www.intel.com/content/www/us/en/develop/documentation/oneapi-fpga-optimization-guide/top/optimize-your-design/throughput-1/single-work-item-kernels/loops/unroll-loops.html, 2022.

[35] ——, "Intel® arria® 10 gx fpga overview," https://www.intel.com/content/www/us/en/products/details/fpga/arria/10/gx/products.html, 2022.

[36] ——, "Intel® stratix® 10 gx fpga overview," https://www.intel.com/content/www/us/en/products/details/fpga/stratix/10.html, 2022.

[37] H. Wang, X. Xie, Y. Li, C. Wen, Y. Li, Y. Liu, S. Qin, H. Chen, and Y. Sui, "Typestate-guided fuzzer for discovering use-after-free vulnerabilities," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ser. ICSE '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 999–1010. [Online]. Available: https://doi.org/10.1145/3377811.3380386

[38] Q. Zhang, J. Wang, M. A. Gulzar, R. Padhye, and M. Kim, "Bigfuzz: Efficient fuzz testing for data analytics using framework abstraction," in *The 35th IEEE/ACM International Conference on Automated Software Engineering*, 2020.

[39] C. Wen, H. Wang, Y. Li, S. Qin, Y. Liu, Z. Xu, H. Chen, X. Xie, G. Pu, and T. Liu, "Memlock: Memory usage guided fuzzing," in *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, 2020, pp. 765–777.

[40] M. Böhme, V.-T. Pham, M.-D. Nguyen, and A. Roychoudhury, "Directed greybox fuzzing," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, D. Evans, T. Maklin, and D. Xu, Eds. United States of America: Association for Computing Machinery (ACM), 2017, pp. 2329–2344, aCM Conference on Computer and Communications Security 2017
, CCS 2017 ; Conference date: 30-10-2017 Through 03-11-2017. [Online]. Available: https://ccs2017.sigsac.org/

[41] C. Lemieux, R. Padhye, K. Sen, and D. Song, "Perffuzz: Automatically generating pathological inputs," in *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis*, ser. ISSTA 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 254–265. [Online]. Available: https://doi.org/10.1145/3213846.3213874

[42] A. Fioraldi, D. Maier, H. Eißfeldt, and M. Heuse, *AFL++: Combining Incremental Steps of Fuzzing Research*. USA: USENIX Association, 2020.

[43] C. Song, X. Zhou, Q. Yin, X. He, H. Zhang, and K. Lu, "P-fuzz: A parallel grey-box fuzzing framework," *Applied Sciences*, vol. 9, no. 23, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/23/5100

[44] J. Liang, Y. Jiang, Y. Chen, M. Wang, C. Zhou, and J. Sun, "Pafl: Extend fuzzing optimizations of single mode to industrial parallel mode," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 809–814. [Online]. Available: https://doi.org/10.1145/3236024.3275525

[45] M. Böhme, V.-T. Pham, and A. Roychoudhury, "Coverage-based greybox fuzzing as markov chain," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1032–1043.

[46] T. Yue, P. Wang, Y. Tang, E. Wang, B. Yu, K. Lu, and X. Zhou, "Ecofuzz: Adaptive energy-saving greybox fuzzing as a variant of the adversarial multi-armed bandit," in *Proceedings of the 29th USENIX Conference on Security Symposium*, 2020, pp. 2307–2324.

[47] N. Calagar, S. D. Brown, and J. H. Anderson, "Source-level debugging for fpga high-level synthesis," in *2014 24th International Conference on Field Programmable*

Logic and Applications (FPL), 2014, pp. 1–8.

[48] J. S. Monson and B. Hutchings, "Using source-to-source compilation to instrument circuits for debug with high level synthesis," in *2015 International Conference on Field Programmable Technology (FPT)*, 2015, pp. 48–55.

[49] Y.-K. Choi and J. Cong, "Hlscope: High-level performance debugging for fpga designs," in *2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2017, pp. 125–128.

[50] K. Laeufer, J. Koenig, D. Kim, J. Bachrach, and K. Sen, "Rfuzz: Coverage-directed fuzz testing of rtl on fpgas," in *Proceedings of the International Conference on Computer-Aided Design*, ser. ICCAD '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3240765.3240842

[51] X. Qin and P. Mishra, "Scalable test generation by interleaving concrete and symbolic execution," in *Proceedings of the 2014 27th International Conference on VLSI Design and 2014 13th International Conference on Embedded Systems*, ser. VLSID '14. USA: IEEE Computer Society, 2014, p. 104–109. [Online]. Available: https://doi.org/10.1109/VLSID.2014.25

[52] Y. Ma, Y. Cao, S. Vrudhula, and J.-s. Seo, "Optimizing loop operation and dataflow in fpga acceleration of deep convolutional neural networks," in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, 2017, pp. 45–54.

[53] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong, "Optimizing fpga-based accelerator design for deep convolutional neural networks," in *Proceedings of the 2015 ACM/SIGDA international symposium on field-programmable gate arrays*, 2015, pp. 161–170.

[54] H. Li, X. Fan, L. Jiao, W. Cao, X. Zhou, and L. Wang, "A high performance fpga-based accelerator for large-scale convolutional neural networks," in *2016 26th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, 2016, pp. 1–9.

[55] P. Garrault and B. Philofsky, "Hdl coding practices to accelerate design performance," *Xilinx White Paper*, vol. 231, pp. 1–22, 2006.

[56] R. Ding, Y. Kim, F. Sang, W. Xu, G. Saileshwar, and T. Kim, "Hardware support to improve fuzzing performance and precision," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 2214–2228.