An Open-Source Python-Based Hardware Generation, Simulation, and Verification Framework

Shunning Jiang  Christopher Torng  Christopher Batten
School of Electrical and Computer Engineering, Cornell University, Ithaca, NY
{ sj634, clt67, cbatten }@cornell.edu

ABSTRACT
We present an overview of previously published features and work in progress for PyMTL, an open-source Python-based hardware generation, simulation, and verification framework that brings compelling productivity benefits to hardware design and verification. PyMTL provides a natural environment for multi-level modeling using method-based interfaces, features highly parametrized static elaboration and analysis/transform passes, supports fast simulation and property-based random testing in pure Python environment, and includes seamless SystemVerilog integration.

1 INTRODUCTION
There have been multiple generations of open-source hardware generation frameworks that attempt to mitigate the increasing hardware design and verification complexity. These frameworks use a high-level general-purpose programming language to express a hardware-oriented declarative or procedural description and explicitly generate a low-level HDL implementation. Our previous work [17] has classified these framework into three major categories. Hardware preprocessing frameworks (HPFs) intermingle a high-level language for macro-processing and a low-level HDL for logic modeling. HPFs enable more powerful parametrization but create an abrupt semantic gap in the hardware description [16, 25]. Hardware generation frameworks (HGFs) completely embed parametrization and behavioral modeling in a unified high-level “host” language [4, 5, 8, 19, 21], but still generate a low-level HDL implementation for simulation. This limits the use of available host-language features, requires test benches be written in the low-level HDL, and creates a modeling/simulation language gap that may require the designer to frequently cross language boundaries during iterative development. All these challenges have inspired completely unified hardware generation and simulation frameworks (HGSFs) where parametrization, static elaboration, test benches, behavioral modeling, and a simulation engine are all embedded in a general-purpose high-level language [3, 6, 12, 14, 15, 22].

Our previous work on PyMTL [17, 20] demonstrated the potential for a Python-based HGSF to improve the productivity of hardware development. The Python language provides a flexible dynamic type system, object-oriented programming paradigms, powerful reflection and introspection, lightweight syntax, and rich standard libraries. HGSFs that are built upon these productivity features enable a designer to write more succinct descriptions, to avoid crossing any language boundaries for development, testing, and evaluation, and to use the complete expressive power of the host language for verification, debugging, instrumentation, and profiling. A typical workflow using PyMTL is shown in Figure 1. The designer starts from developing a functional-level (FL) design-under-test (DUT) and test bench (TB) completely in Python. Then the DUT is iteratively refined to the cycle level (CL) and register-transfer level (RTL) along with verification and evaluation using Python-based simulation and the same TB. The designer can then translate a PyMTL RTL model to Verilog and use the same TB for co-simulation. Note that designers can also co-simulate existing SystemVerilog source code with a PyMTL test bench. The ability to simulate/co-simulate the design in the Python runtime environment drastically reduces the iterative development cycle, eliminates any semantic gap, and makes it feasible to adopt verification methodologies emerging in the open-source software community [11, 23]. Finally, the designer can push the translated DUT through an FPGA/ASIC toolflow. Section 2 gives an overview of key PyMTL features that enable this productive workflow.

Section 3 discusses a variety of PyMTL use cases in the computer architecture community. PyMTL has been used by over 400 students for computer architecture course lab assignments. Multiple research papers at top conferences have used PyMTL for productive CL and RTL modeling [9, 10, 18, 26]. PyMTL has also been used in three chip tapeouts: BRGTC1 [29] in IBM 130 nm, Celerity [1, 13] in TSMC 16 nm, and BRGTC2 [28] in TSMC 28 nm.

2 OVERVIEW OF PYMTL FEATURES
In this section, we introduce the following key productivity features of PyMTL: multi-level modeling, method-based interfaces, highly parametrized static elaboration, analysis and transform passes, pure-Python simulation, property-based random testing, Python/SystemVerilog integration, and fast simulation speed.

Multi-Level Modeling – PyMTL provides a unified environment for modeling hardware at the functional level (FL), cycle level (CL), and register-transfer level (RTL) by providing mechanisms that ensure compatible communication at cross-level boundaries. This multi-level modeling approach systematically builds confidence in verifying a single RTL design-under-test (DUT). The designer is encouraged to first create straightforward FL models which can serve
as golden models for CL/RTL modeling, along with test benches (TB) which can also be reused for CL/RTL verification/simulation. The Python language enables rapid algorithmic exploration at the functional level. Then the designer refines the FL model into CL model for cycle-approximate design-space exploration. After the FL and CL models are implemented, verified, and evaluated, the designer can implement the actual hardware in RTL and reuse the same test bench that has validated the FL/CL models. Figure 2 shows an example of the same design implemented at different levels.

Seamless multi-level modeling in PyMTL also shines in composing multiple models at different levels together for faster design-space exploration. Sometimes the designer might be working under a tight time constraint, and wants to implement only the performance-critical DUT in RTL. To reduce the time spent to simulate the very first composition, the designer can implement critical components in RTL, and non-critical ones in CL based on rough performance estimates (such as a cycle-level cache with a single-cycle hit latency). Later, the CL components may be refined to RTL without any change to other RTL components.

Method-Based Interfaces (Work in Progress) – Method-based interfaces provide designers greater semantic meaning for inter-model communication by raising the level of abstraction at the interface. Currently, cross-layer (e.g., CL to FL/RTL) communications are handled by PyMTL adapters that provide FL/CL with methods to call and RTL with valid/ready handshake signals. However, under the hood they are all implemented using RTL signals and just wrapped with methods. These overheads slow down the simulation of FL/CL models, even though FL/CL models are supposed to be simpler and simulate faster than RTL. In addition, these adapters must be instantiated and managed manually by the designer, which adds extra complexity to the design effort. To address these challenges, we take inspiration from Bluespec’s method-based interfaces [24] and SystemC’s transaction-level modeling (TLM) [27]. We are working on true method-based interfaces for FL/CL modeling and automatic interface coercion between different levels as shown in Figure 3. When composing two models at the same level, CL and FL interfaces can be “connected” by passing a method pointer for FL/CL, and RTL interfaces are still connected via signals. When composing two models at different levels, PyMTL automatically inserts an appropriate adapter that tries to preserve as many method calls as possible, instead of resorting to port-based connections.

Highly Parametrized Static Elaboration – Constructing a highly parametrized hardware generator is one of the key motivations behind modern productive hardware modeling frameworks. In PyMTL, the designer can leverage Python’s object-oriented programming and dynamic typing features to intuitively parametrize PyMTL components, as opposed to using low-level HDL’s limited parametrization constructs and static typing. Python’s extensive support for polymorphism allows the designers to pass parameters of different types around and instantiate different models or logic blocks based on value or type. The static elaboration process executes valid Python code and can be inspected step by step.

Analysis and Transform Passes (Work in Progress) – PyMTL provides APIs to query, add, and remove certain components in the model hierarchy where the root node is the top-level model. Inspired by passes over intermediate representations (IR) in the software realm, the designers can write passes that call these APIs to analyze and transform the whole design. Previous work in Chisel [4] and PyRTL [12] advocate for adding another hardware IR level between the host language and a low-level HDL. We argue that PyMTL passes over the module hierarchy at Python level are more intuitive and productive. Analysis passes usually query a list of modules in the hierarchy and accumulate the obtained information for a grand goal. For example, we can query the total number of models that has at least two input ports, a list of ports that starts with a specific name, or even the average number of statements of all logic blocks. Transform passes modify the model hierarchy. Increment-only transform passes add components to instrument the design without invoking APIs to remove components or connections. An example is to add a child module and bring a signal up to the top level by recursively going up the hierarchy to the top. Other passes involve both adding and removing components or connections. An example is to insert a wrapper component between a module and its child module, which requires removing the child module first, instantiating a new wrapper module, instantiating the same child module within the wrapper module, and establishing all the connections. Figure 4 shows code for these example passes.

Pure-Python Simulation – Unlike HGFs which translate the high-level hardware description to low-level HDL and use an HDL simulator, PyMTL’s simulation kernel is built in Python. The designer can track the simulation cycle by cycle and line by line in Python code at runtime instead of relying solely on waveform-based debugging. Note that PyMTL can also dump waveforms.
An Open-Source Python-Based HGSVF

# Analysis pass example:
# Get a list of processors with >=2 input ports

def count_pass( top ):
    ret = []
    for m in top.get_all_modules_filter(
        lambda m: len(m.get_input_ports()) >= 2):
        if isinstance( m, AbstractProcessor ):
            ret.append( m )
    return ret

# Increment-only transform pass example:
# Bring up state variable of every state machine to a top-level
# output port

def debug_port_pass( top ):
    for m in top.get_all_modules():
        if m.get_full_name().startswith("ctrl"):
            signal_type = m.state.get_type()
            port_name = "debug_state" + mangle( m.get_full_name() )
            m.outputport = m.add_output_port( port_name,
                OutPort( signal_type ) )
            m.add_connection( m.outputport, m.state )

while m.has_parent():
    p = m.get_parent()
    p.outputport = p.add_output_port( port_name,
        OutPort( signal_type ) )
    p.add_connection( p.outputport, m.outputport )
    m, m.outputport = p, p.outputport

# Transform pass example:
# Wrap every ctrl with CtrlWrapper

def debug_port_pass( top ):
    for m in top.get_all_modules():
        if m.get_full_name().startswith("ctrl"):
            ctrl = p.delete_component( "ctrl" )
            new_ctrl = w.add_component( "ctrl_wrap", CtrlWrapper() )
            ctrl = w.add_component( "ctrl", m )
            ...< connect ports >

Simulating in a pure Python runtime means the designer can leverage all the existing third-party Python packages for verification. This significantly helps bring the success of open-source software to open-source hardware. For example, machine learning accelerator designers can import packages like PyTorch and TensorFlow to generate input/reference datasets for test benches and reuse the algorithm implementation for FL/CL models. PyMTL also leverages existing open-source software testing/verification facilities. pytest testing framework is used for instantiating numerous tests from a single concise definition, and coverage.py tool is used for line-by-line code coverage.

Property-Based Random Testing (Work in Progress) - Constraint-based hardware verification frameworks such as UVM have been widely adopted in the chip-building industry. However, there is no simulator that supports UVM in the open-source hardware/EDA community. As the first step, PyMTL integrates hypothesis, a sophisticated property-based random testing framework which was originally designed for verifying Python software. Hypothesis automatically generates numerous random test cases according to a given "hypothesis strategy". After one test case fails, hypothesis will try to construct a minimal failed test case by "auto-shrinking". PyMTL will develop specialized strategies such as generating a random-length list of random packets for verifying PyMTL designs as well as to verify the PyMTL framework itself by generating random logic statements.

Python/SystemVerilog Integration - PyMTL supports importing SystemVerilog code for plug-and-play co-simulation and composition with other PyMTL models or imported SystemVerilog models. Combined with the ability to translate PyMTL RTL models into Verilog code, PyMTL becomes a holistic hardware composition and verification framework. Below are the three major advantages.

First, Verilog code generated by PyMTL can be validated by co-simulating with the same PyMTL test bench from FL/CL/RTL development. This helps make sure there are no code generation mismatches and simulation mismatches. PyMTL generates Verilog for tagged components, calls Verilator (an open-source SystemVerilog simulator) to compile each generated Verilog file, compiles this C++ into a shared library and dynamically links it back to PyMTL program using CFFI (C Foreign Function Interface), and replaces the tagged PyMTL model with the Verilog model. Compared to hardware generation frameworks (HGF) where there is usually no way to test if the translated HDL matches the high-level code, PyMTL builds more confidence for RTL developers.

Second, PyMTL can help verification engineers even if they are not writing PyMTL RTL models. They can hand-write SystemVerilog code, write a simple PyMTL wrapper as shown in Figure 5, and build a PyMTL test bench to drive the simulation. At a larger design scale, PyMTL can simulate a design composed of many FL, CL, RTL, and SystemVerilog models based on how detailed the designer models each component. This enables the designer to productively verify SystemVerilog models using supporting FL/CL PyMTL models in an end-to-end testing approach.

Third, PyMTL can act as a glue for composing multiple SystemVerilog models and PyMTL RTL models. The imported code of a child Verilog model is preserved when PyMTL translates the top-level module to Verilog. Structural composition of SystemVerilog models can take advantage of PyMTL's parameterization power to create a larger composition of wrapped SystemVerilog models.

Fast Simulation Speed – Adapted from our recent work [17]. Table 1 shows an apples-to-apples simulation performance comparison of an iterative divider, a single RISC-V RV32IM [2] core hooked up to a two-port test memory, and 32 cores hooked up to a 64-port test-memory. PyMTL offers competitive simulation performance compared to other HGSFs, but commercial HDL simulators can still be orders of magnitude faster than HGSFs. Our work shows that a carefully designed HGSF can close the simulation-performance gap by deeply co-optimizing the HGSF and the underlying general-purpose JIT compiler within the host high-level language. With Mamba techniques, PyMTL’s pure-Python simulation performance matches a commercial Verilog simulator and is 10X-20X faster than the original PyMTL.

Table 1: Simulation Performance Comparison – CPS = simulated cycle per second; CVS = commercial Verilog simulator.

<table>
<thead>
<tr>
<th></th>
<th>PyMTL</th>
<th>MyHDL</th>
<th>PyRTL</th>
<th>Migen</th>
<th>IVerilog</th>
<th>CVS</th>
<th>Mamba</th>
</tr>
</thead>
<tbody>
<tr>
<td>Divider</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>118K CPS</td>
<td>0.8×</td>
<td>2.2×</td>
<td>0.03×</td>
<td>0.6×</td>
<td>9.3×</td>
<td>20×</td>
</tr>
<tr>
<td>1-core</td>
<td>20K CPS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1×</td>
<td>15×</td>
<td>16×</td>
</tr>
<tr>
<td>32-core</td>
<td>360 CPS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.8×</td>
<td>25×</td>
<td>12×</td>
</tr>
</tbody>
</table>

Figure 5: PyMTL Source Code for SystemVerilog Import

```python
class DUT( VerilogModel ):
    def __init__( self ):
        s.in_ = InPort( Bits32 )
        s.out = OutPort( Bits32 )

    # Connect top level ports of DUT
    @property
    def s(self):
        return s

    # Corresponding PyMTL ports
    s.in_ = p_inport( Bits32 )
    s.out = p_outport
      ...

    # Increment-only transform pass example:
    # Bring up state variable of every state machine to a top-level
    # output port
    def debug_port_pass( top ):
        for m in top.get_all_modules():
            if isinstance( m, AbstractProcessor ):
                signal_type = m.state.get_type()
                port_name = "debug_state" + mangle( m.get_full_name() )
                m.outputport = m.add_output_port( port_name,
                    OutPort( signal_type ) )
                m.add_connection( m.outputport, m.state )

    while m.has_parent():
        p = m.get_parent()
        p.outputport = p.add_output_port( port_name,
            OutPort( signal_type ) )
        p.add_connection( p.outputport, m.outputport )
        m, m.outputport = p, p.outputport

    # Transform pass example:
    # Wrap every ctrl with CtrlWrapper
    def debug_port_pass( top ):
        for m in top.get_all_modules():
            if isinstance( m, AbstractProcessor ):
                if m.get_full_name().startswith("ctrl"):
                    signal_type = m.state.get_type()
                    port_name = "debug_state" + mangle( m.get_full_name() )
                    m.outputport = m.add_output_port( port_name,
                        OutPort( signal_type ) )
                    m.add_connection( m.outputport, m.state )

            while m.has_parent():
                p = m.get_parent()
                p.outputport = p.add_output_port( port_name,
                    OutPort( signal_type ) )
                p.add_connection( p.outputport, m.outputport )
                m, m.outputport = p, p.outputport

    # By default PyMTL imports module DUT of DUT.v
    # in the same folder as the python source file.
    class DUT( VerilogModel ):
        def __init__( self ):
            s.in_ = InPort( Bits32 )
            s.out = OutPort( Bits32 )

        # Connect top level ports of DUT
        @property
        def s(self):
            return s

        # Corresponding PyMTL ports
        s.in_ = p_inport( Bits32 )
        s.out = p_outport

        # Increment-only transform pass example:
        # Bring up state variable of every state machine to a top-level
        # output port
        def debug_port_pass( top ):
            for m in top.get_all_modules():
                if isinstance( m, AbstractProcessor ):
                    signal_type = m.state.get_type()
                    port_name = "debug_state" + mangle( m.get_full_name() )
                    m.outputport = m.add_output_port( port_name,
                        OutPort( signal_type ) )
                    m.add_connection( m.outputport, m.state )

            while m.has_parent():
                p = m.get_parent()
                p.outputport = p.add_output_port( port_name,
                    OutPort( signal_type ) )
                p.add_connection( p.outputport, m.outputport )
                m, m.outputport = p, p.outputport
```

Figure 4: Example of Analysis and Transform Passes
3 PyMTL USE CASES

In this section, we discuss how PyMTL has already been employed for teaching in the classroom, for driving computer architecture research, and for building silicon prototypes.

3.1 PyMTL for Teaching

PyMTL has been used by over 400 students across two universities, including in a senior-level undergraduate computer architecture course at Cornell University (ECE 4750), in a similar course at Boston University (EC 513), and in a graduate-level ASIC design course at Cornell University (ECE 5745). The computer architecture courses involved multiple design labs (integer multiplier, simple RISC-V processor, set-associative blocking cache, and bus/ring network), culminating in a final lab composing all previous components to build a multi-core system. Students chose whether to design in PyMTL, in SystemVerilog, or with a mix, but they were required to test their designs using PyMTL. Overall, PyMTL accelerates students’ learning curve. Developing and simulating the design in a pure-Python runtime environment accommodates students with more of a software background. PyMTL’s SystemVerilog integration feature shortens the iterative development cycle for students with more of a hardware background.

3.2 PyMTL for Architecture Research

PyMTL has driven experiments for multiple computer architecture research projects that have been published at top conferences. The LTA [18] and XLOOPS [26] papers both modeled novel accelerators as PyMTL CL models, which were then composited with general-purpose control processors in gem5 [7] using PyMTL/gem5 co-simulation support (co-simulation is implemented using the C Foreign Function Interface, CFFI). Gem5’s popular and well-supported CPU models, memory system, and runtime features are directly reusable, enabling researchers to focus their efforts on exploring the accelerator design-space in PyMTL.

Similarly, DAE [10] and ParallelXL [9] both leveraged PyMTL with particular emphasis on the framework’s highly parametrized static elaboration features for RTL design. These two papers explored the design of architectural templates that efficiently generate tuned accelerators.

3.3 PyMTL for Silicon Prototyping

PyMTL designs have been taped out in advanced nodes including TSMC 16 nm, TSMC 28 nm, and IBM 130 nm. The associated projects have been published: Celerity [1, 15], BRGTC1 [29], and BRGTC2 [28].

Celerity is a 5x5 mm 385M-transistor chip in TSMC 16 nm designed and fabricated by a large team of over 20 students and faculty from UC San Diego, University of Michigan, and Cornell as part of the DARPA Circuit Realization At Faster Timescales (CRAFT) program. PyMTL played a key role in the integration of a complex HLS-generated BNN (i.e., binary neural network) with the broader system’s general-purpose compute tier (i.e., five modified Chisel-generated RISC-V Rocket cores) as well as its massively parallel compute fabric (i.e., 496-core RISC-V tiled manycore processor). The BNN was wrapped with PyMTL-generated parameterized RoCC wrappers and adapters, and these wrappers were generic and were automatically generated using reflection. The wrapped BNN was translated back to Verilog for composition with the rest of the chip. The control blocks for the high-speed links between the BNN and the manycore were also designed and verified in PyMTL.

Figure 6: BRGTC1 and BRGTC2

BRGTC1 (i.e., Batten Research Group Test Chip 1) was implemented nearly entirely using PyMTL. BRGTC1 marked the first exploration of the interaction between PyMTL RTL and HLS-generated Verilog models. The prototype is a small 2x2 mm 1.3M-transistor chip in IBM 130 nm, and it pairs a simple pipelined 32-bit RISC processor developed in PyMTL with an HLS-generated application-specific accelerator. Fabricated BRGTC1 chips have been post-silicon validated for functionality using assembly tests. Our BRGTC1 lab bench setup re-uses the PyMTL test suite to drive signals through a “host” FPGA base board (i.e., Xilinx Zedboard) and out to an FMC-connected daughter card that contains the test chip.

BRGTC2 (i.e., Test Chip 2) was built with much more aggressive use of PyMTL, resulting in a 1x1.25 mm 6.7M-transistor chip in TSMC 28 nm. The chip contains four RISC-V RV32IMAF cores sharing a 32kB instruction cache, a 32kB data cache, and a single-precision floating point unit, along with microarchitectural mechanisms to mitigate the performance impact of resource sharing. The BRGTC2 project stressed PyMTL’s features extensively. For example, multi-level modeling supported debugging efforts by enabling the team to swap in an FL cache to narrow the location of an RTL bug down to other components. A single PyMTL test suite is used to test all modeling abstractions, including the FL, CL, RTL, and even gate-level models. As another success, the floating-point unit was designed using a collection of Synopsys DesignWare components that were each Verilog-imported into PyMTL for composition and simulation. Overall, the PyMTL framework was a tremendous success as a productive hardware modeling framework for designing two non-trivial research test chips.

4 CONCLUSION


ACKNOWLEDGMENTS

This work was supported in part by NSF CRI Award #1512937, NSF SHF Award #1527065, DARPA POSH Award #FA8650-18-2-7852, and a donation from Intel. The authors acknowledge and thank Derek Lockhart for his valuable feedback and his work on the original PyMTL framework. The authors would like to thank Ajay Joshi for using PyMTL for the computer architecture course at Boston University. The author also thank all the students who have provided feedback to PyMTL. U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation theron. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of any funding agency.
REFERENCES