A Formal Specification of Tensor Cores via Satisfiability Modulo Theories

BENJAMIN VALPEY, University of Rochester, USA

Because of the complexities of today's processors, users require suitable documentation in order to use them correctly and effectively. While standard microprocessors may have suitable documentation, others, such as GPUs – the focus of this work – fall short in this regard. Without suitable documentation, users resort to manual tests to understand the units. This leads to a splintered and sometimes contradictory knowledge base. Vendors can remedy this issue by providing a formal specification of their hardware implementation. We argue that SMTLIB is a language well-suited for such a task. SMTLIB's advantage over C, the current de-facto specification language other than plain English, is that it is easy to reason about; SMT solvers can be used to prove properties about the hardware specification. However, vendors may not always release such a specification, so we propose a method to help automate the formalization of hardware implementations.

We showcase our method by using it to create a formal specification of the Tensor Cores for NVIDIA's Volta Architecture. Tensor cores are specialized units for matrix multiplication that vary in behavior between different machines. However, their implementation details and numerical behavior are not standardized. Often their exact behavior is not fully documented either, which can make it challenging to write consistent software. Past work has identified properties and developed informal theories of the functioning of tensor cores on NVIDIA's GPUs. Such work resorts to manually-concocted tests which must be repeated as new architectures modify the behavior of tensor cores. In this study, we describe a formal model of the tensor cores in NVIDIA's Volta architecture using the Satisfiability Modulo Theories Library (SMT-LIB). By querying this model to identify discriminating inputs, we determine various properties of the units, including their rounding mode, precision, and accumulation order. Using this approach, we confirm many of the findings from previous studies on tensor cores, but we also identify some discrepancies. Specifically, we find that the hardware does not use IEEE-754 round-to-zero for accumulation. Instead, We also confirm the result of a previous study that showed that tensor cores do not normalize intermediate sums. This previous study asserted that due to the lack of normalization, 3 carry bits are needed for proper accumulation, but were only able to provide inputs that proved that 2 bits were used. Using our model, we find inputs that prove the 5-term accumulator requires 3 extra carry-out bits for accurate accumulation.

Another advantage of our approach is that it can be reused to analyze programs that utilize the hardware it formalizes. To demonstrate this, we use our model to analyze two algorithms designed for tensor cores. The two algorithms use tensor cores to accelerate single-precision multiplication by using tensor cores The second method leveraged the knowledge of tensor cores behavior to create a new method that offers an increase in accuracy at the cost of a slight performance hit. Crucially, their ability to improvement the earlier method relied on knowledge about the hardware implementation. Using our model, we determine whether this second method always produces a result with less error than the original, or if it is possible for its result to be less accurate. We find that the second technique can actually yield less accurate results than the original method for certain inputs.

Our model is provided as an implementation of tensor cores in a parametric Python script that generates SMT-LIB. We expect that our detailed model can not only be used to replicate the behavior of tensor cores in tools such as emulators, but also form the basis of formal "tensor core theories" that can be used to describe and maintain consistency across differing tensor core implementations.

SC '23, November 14–16, 2023, Denver, CO

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

ACM Reference Format:
Benjamin Valpey. 2023. A Formal Specification of Tensor Cores via Satisfiability Modulo Theories. In *SC '23: Student Research Competition Graduate Poster Session, November 14–16, 2023, Denver, CO*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn