Errata

Trinity: Three-dimensional Tensor Program Optimization via Tile-level Equality Saturation

Jaehyeong Park, Youngchan Kim, Haechan An, Gieun Jeong, Jeehoon Kang, Dongsu Han | ASPLOS 2026

https://dl.acm.org/doi/10.1145/3779212.3790240

April 2026

We have identified two inaccuracies in our description of FlashInfer, one of the baselines evaluated in our paper. These errors do not affect our experimental results, conclusions, or any of the paper's core contributions. They are strictly limited to the textual description of the baseline.

Correction 1

Current text (appearing in multiple places):

> - "FlashInfer is manually engineered kernels" (page 2080, last paragraph in Introduction)
> - "FlashInfer is a manually optimized implementation" (page 2087, the paragraph immediately preceding the Evaluation section)
> - "FlashInfer is a hand-tuned kernel library" (page 2088, first paragraph)

Corrected text:

> "FlashInfer is an attention kernel library based on the manually engineered FlashAttention algorithm."

Reason for correction:

We compared against FlashInfer to demonstrate that our system outperforms solutions built on top of manually engineered attention algorithms such as FlashAttention. However, describing FlashInfer as a "manually optimized" or "hand-tuned" kernel library is misleading, as FlashInfer also employs JIT compilation to automate certain low-level optimizations, including tile size selection, loop unrolling, and tensor core fragment mapping.

We believe the corrected description more accurately characterizes FlashInfer. Importantly, our core claim remains valid: FlashInfer's underlying algorithm (FlashAttention) is manually engineered, and our system demonstrably outperforms it. Moreover, the fact that our system outperforms FlashInfer — which extends FlashAttention with additional JIT-based optimizations — further highlights the strength of our approach, rather than diminishing it.

Correction 2

Current text (page 2088, second paragraph of Section 6.2):

> "FlashInfer, despite being a state-of-the-art hand-tuned attention kernel, cannot handle any architecture beyond Vanilla transformer."

Corrected text:

> "FlashInfer, despite being a state-of-the-art attention kernel library, is specialized for core attention computation and does not support operators outside of attention, such as Pre-Norm and SwiGLU-FFN. Furthermore, supporting additional attention variants in FlashInfer requires manually implementing optimized algorithms for each new variant."

Reason for correction:

Our original statement was intended to convey that FlashInfer does not support the specific model architecture variants used in our experiments. However, the phrasing "cannot handle any architecture beyond the Vanilla Transformer" was inaccurate and potentially misleading, as FlashInfer in fact supports a range of attention variants beyond vanilla attention — including MLA, RoPE, and sliding window attention — through its customizable attention template.

The corrected text more accurately characterizes FlashInfer's limitations in two respects. First, FlashInfer is specialized for core attention computation and does not support operators outside of attention, such as Pre-Norm and SwiGLU-FFN, which are used in our experiments. Second, supporting additional attention variants still requires manually implementing optimized algorithms for each new variant. Our core claim therefore remains valid — these limitations make it difficult to cover the growing diversity of model architectures, highlighting the need for automatic optimization.

Jaehyeong Park
on behalf of the authors of Trinity