

Triton

Kernel Triton

The aim of Triton is to provide an open-source environment to write fast code at higher productivity than CUDA, but also with higher flexibility than other existing DSLs.

<https://github.com/openai/triton>

<https://triton-lang.org/main/index.html>

Diagram illustrating the workflow of Triton AI:

- kernel** (represented by a 1x10 grid) is input to the **Triton compiler** (represented by a 1x10 grid).
- The **Triton compiler** outputs a **Triton** component (represented by a 1x10 grid).
- This **Triton** component is used in a **Triton + Triton** configuration (represented by a 1x10 grid) to produce the final **AI** output (represented by a 1x10 grid).
- The **AI** output is then used in a **Triton** configuration (represented by a 1x10 grid) to produce the final **kernel** output (represented by a 1x10 grid).

Triton [] [] [] [] [] ----Block-wise [] [] Block [] [] [] [] [] Block [] [] [] Triton compiler

[] [] [] [] [] [] [] [] Block [] [] [] [] [] [] [] [] Triton compiler [] [] []

Passes Triton NV GPU kernel
 coalescing pipeline/prefetch
 shared memory bank-conflict swizzling

Revision #1

Created 11 January 2025 09:46:28 by Colin

Updated 12 January 2025 06:35:20 by Colin