Accelerating Stencil Computation with Tensor Core Unit on GPU

1 ROOFLINE ANALYSIS

We adopt the roofline model [2] to analyze the effectiveness of stencil computation by Artemis, *TCstencil*, and *TC-w/o tc*. Since Nsight profiler do not supprot collect FLOP in half, so we give the theoretical value. Table 1 shows the the maximum operational intensity $I_{max} = \pi/\beta$ (FLOP/Byte) in FP16 precision of A100 and V100, where π (TFLOPS) is the maximum attainable performance the β (GB/s) is maximum memory bandwidth. We are only considering the float multiply operation andthe access to global memory of the programe. For updating the inner points in 16×16 mesh with 9pt stencil, the Arithmetic Intensity (AI, FLOP/Byte) of Artemis/*TC-w/o tc*/*TCstencil* is 0.50/2.53/5.33. As for 25pt stencil, the AI of Artemis/*TC-w/o tc*/*TCstencil* is 0.50/7.03/6.

Table 1: The performance information of A100 and V100 in FP16 precision [1].

Item	A100 FMA	A100 TC	V100 FMA	V100 TC
π	78	312	31.4	125
β	1555		900	
I_{max}	50.1	200.64	34.88	128.8

REFERENCES

[1] NVIDIA. 2021-1-30. NVIDIA Turing Architecture Whitepaper. online. https://www.nvidia.com/en-us/data-center/a100/ Accessed March 30, 2021.

[2] Samuel Williams, Andrew Waterman, and David Patterson. 2009. Roofline: an insightful visual performance model for multicore architectures. *Commun. ACM* 52, 4 (2009), 65–76.