

Runtime Profiling: Latencies and Scheduling on Floating-Point and Hybrid Float 6 Accelerators

Tensor Op	CPU (ms)	TP FP32 (ms)	Gain	TP HF6 (ms)	Gain
Conv _A	112.24	12.49	8.9	6.92	16.2
Conv _B	213.13	16.39	13.0		
Conv _C	46.59	3.59	12.9		

