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	**4.41**	**48.3**
$Conv_C$	46.59	3.59	12.9	**0.99**	**47.0**

CPU + TP (Hybrid-Float6 @ 200 MHz)

Total = 27.9 ms (35.84 IPS)