Performance Benchmarks on Ironwood TPU

November 13, 2025

Methodology

- The performance is measured for both the forward (FWD) and backward (BWD) passes of the Splash Attention kernel across various input shapes and hyperparameters. The benchmarks were run with a fixed configuration of `batch_size = 1`, `q_heads = 32`, and `kv_heads = 8` with `bfloat16 (BF16)` numerical precision.
- **Theoretical Limit (ms):** This is the minimum possible execution time for a purely compute-bound kernel, often referred to as the "Roofline Latency, computed as Required compute FLOPs / HW Peak FLOPS
    - Required FLOPs FWD: `0.5 (assumed sparsity for causal attn) * batch_size * q_heads * q_seq_len * kv_seq_len * (2 * qk_head_dim + 2 * v_head_dim)`
    - Required FLOPs BWD: `2.5 * Required FLOPs FWD`
    - Ironwood HW Peak FLOPS: **1.15 x 10^15**
- **Measured Time (ms):** The actual execution time of the kernel, captured during the benchmark run.
- **MFU (Model FLOPs Utilization):** This metric quantifies how efficiently the kernel utilizes the tensor cores of the TPU. It is expressed as a percentage: `MFU (%) = (Theoretical Limit (ms) / Measured Time (ms)) * 100`
- To achieve the best possible performance, various kernel hyperparameters were exhaustively tuned for each configuration of interest. These include parameters that control how data is tiled and processed within the TPU's memory hierarchy, such as block sizes (`block_q`, `block_kv`, etc.) and memory layouts (`q_layout`, `k_layout`, etc.).

Forward Pass (FWD) Benchmarks

Theoretical Limit (ms) / Measured Time (ms) = MFU%

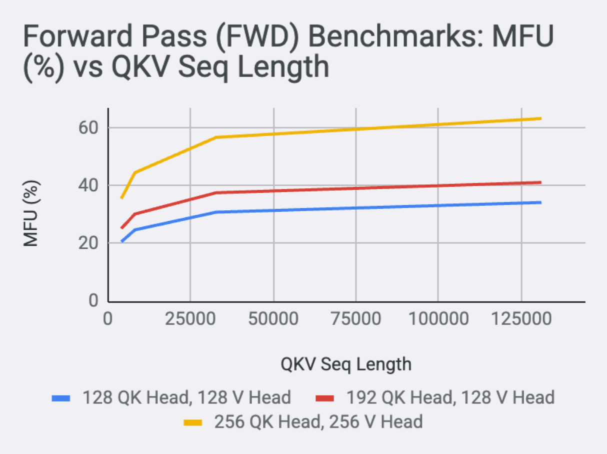
Head Dimension ⟶ **Sequence Length (for Q and KV) ↓**	**QK=128, V=128**	**QK=192, V=128**	**QK=256, V=256**
4096	0.12ms / 0.58 ms = 20.62% MFU	0.15 ms / 0.59 ms = 25.42% MFU	0.24 ms / 0.67 ms = 35.82% MFU
8192	0.48 ms / 1.93 ms = 24.87% MFU	0.60 ms / 1.98 ms = 30.30% MFU	0.96 ms / 2.14 ms = 44.86% MFU
32768	7.65 ms / 24.75 ms = 30.91% MFU	9.56 ms / 25.41 ms = 37.62% MFU	15.30 ms / 26.93 ms = 56.81% MFU
131072	122.38 ms / 357.20 ms = 34.26% MFU	152.98 ms / 371.30 ms = 41.20% MFU	244.76 ms / 386.39 ms = 63.35% MFU

Backward Pass (BWD) Benchmarks

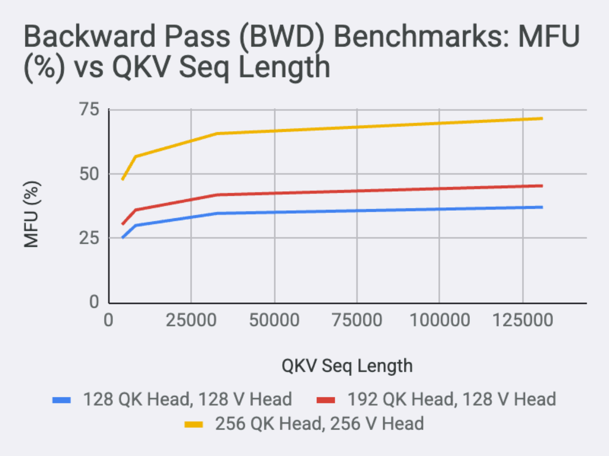
Theoretical Limit (ms) / Measured Time (ms) = MFU%

Head Dimension ⟶ **Sequence Length (for Q and KV) ↓**	**QK=128, V=128**	**QK=192, V=128**	**QK=256, V=256**
4096	0.30 ms / 1.19 ms = 25.21%	0.37 ms / 1.23 ms = 30.08%	0.60 ms / 1.25 ms = 48.00%
8192	1.20 ms / 3.98 ms = 30.15%	1.49 ms / 4.14 ms = 35.99%	2.39 ms / 4.20 ms = 56.90%

32768	19.12 ms / 55.05 ms = 34.73%	23.90 ms / 56.97 ms = 41.95%	38.24 ms / 58.11 ms = 65.81%
131072	305.95 ms / 823.43 ms = 37.16%	382.44 ms / 840.74 ms = 45.49%	611.90 ms / 853.37 ms = 71.71%



Forward Pass (FWD) Benchmarks: MFU (%) vs QKV Seq Length



Backward Pass (BWD) Benchmarks: MFU (%) vs QKV Seq Length