Flash Attention, matrix size $m \times n \times n$, $m = 8192/n$

- basic CUDA f16
- basic CUDA f32
- CUDA my backend w/ TC