A.

Triton and Torch matmult results match.
The compile+autotune+first-run time in seconds: 0.8167842279999604
Average kernel time in ms: 0.034508800506591795

B.

Achieved GFLOPs/s: 62230.028760048976
Machine: NVIDIA GeForce RTX 4090 and its Peak GFLOPs/s: 82600.0
Percent of Peak GFLOPs/s: 76%

C.

Sweep Results 256-4096:

N	1st Kernel call	time ms	GFLOPS	% Peak
256	0.6223	0.025	1365.3	1.65%
512	0.6238	0.024	11073.8	13.41%
768	0.6298	0.025	36530.1	44.23%
1024	0.0001	0.031	69167.3	83.74%
1536	0.6687	0.127	57273.7	69.34%
2048	0.6939	0.215	79830.7	96.65%
3072	0.7530	0.752	77076.0	93.31%
4096	0.8112	1.674	82105.4	99.40%

E.

1. Bar.sync: It is a block level barrier for threads using a shared memory(tile for triton).
2. Ld.shared/st.shared: Multiple threads concurrently read from input matrices, concurrent reads are safe to do. Output writes are independent
3. Cp.async: Asynchronous data movement and pipelining