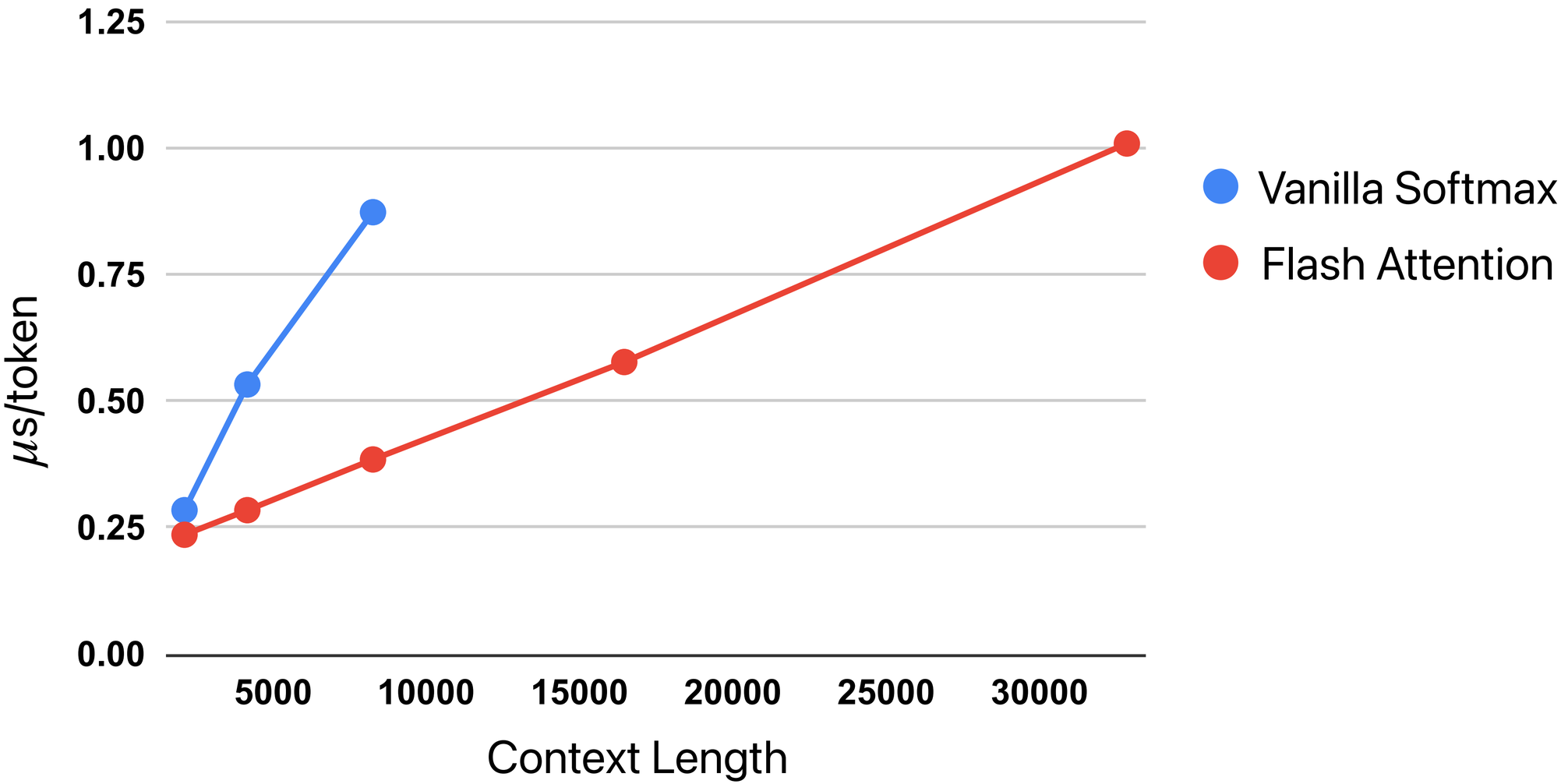
Train Step Latency Per Token



Each token in the training examples looks back at the whole context

Sub-quadratic Time Algorithms?

- Unfortunately seems to be no!
    - Exact or entrywise approximation
- Alman and Song show output of attention is hard to approximate to high precision under reasonable complexity assumptions