

(a) Tokens/Second Data for F16 Quantization

