

(a) Tokens/Second Data for 4 Bit Quantization

