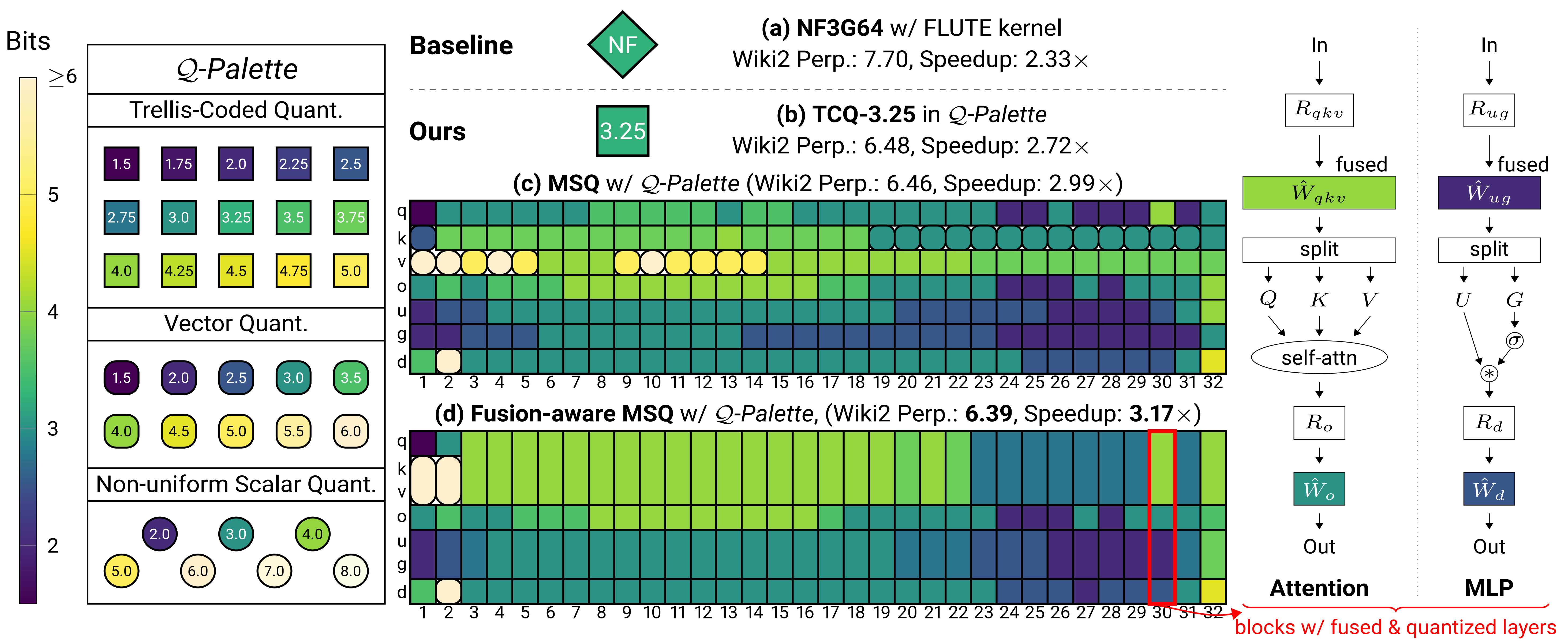
Q-Palette: Fractional-Bit Quantizers Toward Optimal Bit Allocation for Efficient LLM Deployment

Deokjae Lee, Hyun Oh Song

Seoul National University

Overview



Contributions

1. We develop **Q-Palette**, a set of versatile rotation-based quantizers with efficient inference CUDA kernels and wide fractional-bit support.

2. Built on Q-Palette, we propose a novel mixed-scheme quantization framework that **jointly optimizes quantizer selection and layer fusion**.

Results

More results are in our paper!

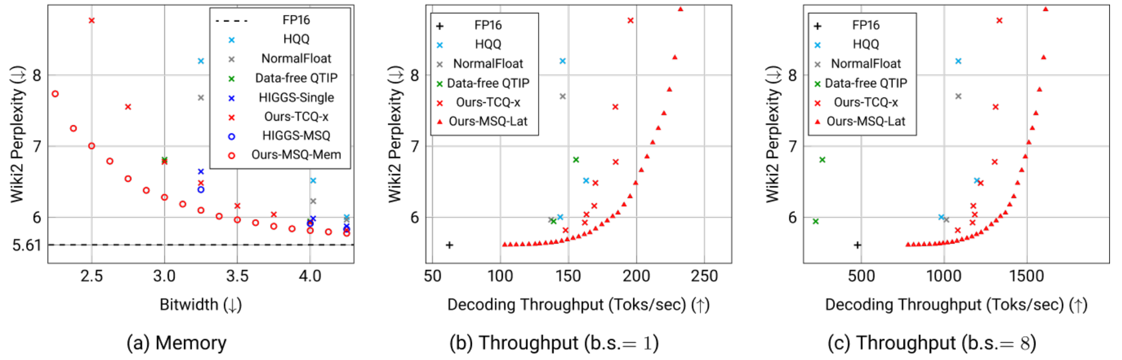


Figure 1. Performance trade-offs under different constraints (LLaMA 3.1-8B, RTX4090 GPU).

Motivation

- Recent quantization methods utilize **rotation** to Gaussianize weights, reducing outliers.
- Based on HIGGS's linearized surrogate loss

$$\mathcal{L}(\{Q(W_l)\}_{l=1}^{L}) - \mathcal{L}(\{W_l\}_{l=1}^{L}) \approx \sum_{l=1}^{L} a_l \underbrace{\|Q(W_l) - W_l\|^2/\|W_l\|^2}_{=:\,\mathrm{err}(Q;W_l)},$$

the memory-constrained mixed scheme quantization (MSQ) can be formally written by

$$\underset{P_{lq}\in\{0,1\}}{\text{minimize}} \quad \sum_{l=1}^{L}\sum_{q=1}^{|\mathcal{Q}|} P_{lq}\cdot \underbrace{(a_l\cdot \mathrm{err}(Q_q;W_l))}_{\text{loss term } \ell_{lq}}$$

$$\text{subject to} \quad \underbrace{\sum_{q=1}^{|\mathcal{Q}|} P_{lq} = 1 \quad \forall 1\le l\le L,}_{\text{Exclusive assignment}} \quad \underbrace{\sum_{l=1}^{L}\sum_{q=1}^{|\mathcal{Q}|} P_{lq}\cdot \mathrm{bit}(Q_q;W_l)\, d_l^{\mathrm{in}} d_l^{\mathrm{out}} \le M.}_{\text{Memory constratint}}$$

- Here, for Gaussianized weights, we have a natural lower bound on distortion term $\mathbb{E}[\mathrm{err}(Q)] \ge 2^{-2\mathrm{bit}(Q)}$, given by Gaussian distortion-rate theorem.
- Assuming ideal Gaussian quantizers that achieve the optimal distortion bound, the optimal bit allocation is given by

$$b_l^* = \max\left\{0, \frac{1}{2\ln(2)}\left(\ln\frac{a_l}{d_l^{\mathrm{in}} d_l^{\mathrm{out}}}\right) + C\right\} \quad \forall 1\le l\le L, \quad \text{for } C \text{ s.t. } \sum_{l=1}^{L} b_l^* d_l^{\mathrm{in}} d_l^{\mathrm{out}} = M.$$

- However, in practice, quantization is performed with a set of non-ideal quantizers, and the performance gap is given by 1) how closely each quantizer approaches the distortion bound, and 2) how finely the available bitwidths approximate the optimal bit allocations, motivating the design of **Q-Palette** (Sec 3.2).

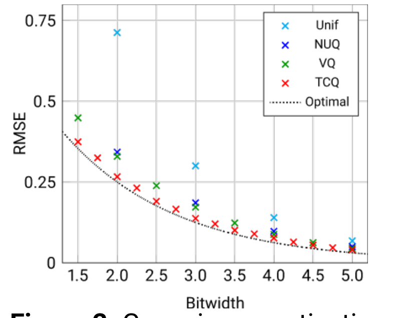


Figure 2. Gaussian quantization error of Q-Palette quantizers.

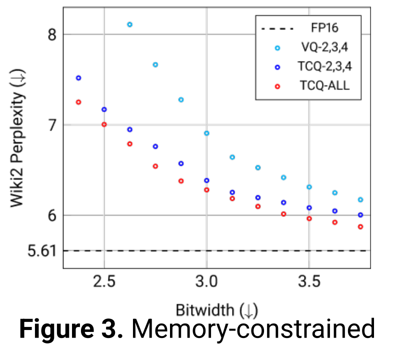


Figure 3. Memory-constrained MSQ results for different quantizer sets (LLaMA 3.1-8B).

Fusion-aware MSQ

- We propose **fusion-aware MSQ**, a novel MSQ framework that jointly optimizes quantization with the additional design dimension of **layer fusion**. (see Overview)
- Fusion-aware MSQ simultaneously determines 1) how to group layers for fusion and 2) which quantizer to assign to each fused group, encoded by P_{gq}.
- The fusion-aware MSQ problem is formulated as

$$\underset{P_{gq}\in\{0,1\}}{\text{minimize}} \quad \sum_{g\in\mathcal{G}}\sum_{q=1}^{|\mathcal{Q}|} P_{gq}\cdot \sum_{l\in g}\ell_{lq}$$

$$\text{subject to} \quad \underbrace{\sum_{g\in\mathcal{G}:l\in g}\sum_{q=1}^{|\mathcal{Q}|} P_{gq} = 1 \quad \forall 1\le l\le L,}_{\text{Exclusive assignment}} \quad \underbrace{\sum_{g\in\mathcal{G}}\sum_{q=1}^{|\mathcal{Q}|} P_{gq}\cdot c_{gq} \le C.}_{\text{Resource constraint}}$$

- Here, $\mathcal{G}$ denotes the set of all fusible layer groups, where each group consists of linear layers sharing the same input (*e.g.*, Q, K, V layers in the same Transformer block).
- The problem above is ILP and can be solved by solvers such as SCIP solver.