Quantized Cache-to-Cache: Communication-Budgeted KV Transfer for Heterogeneous LLMs

Anonymous Authors

Abstract

We study precision-aware communication between heterogeneous LLMs by quantizing KV-cache transfers in Cache-to-Cache (C2C). Our goal is to reduce bandwidth and memory while preserving accuracy. We present post-training quantization, cache-length reduction, and token-level selective transfer with sparse fusion, and show accuracy vs. bytes transmitted curves. This draft contains placeholders for results.

1 Introduction

Large language models (LLMs) often communicate via text, which is slow and lossy. Cache-to-Cache (C2C) communicates via KV-cache projection and fusion, but does not address precision constraints. We ask: *How low can KV precision go before accuracy collapses, and can we recover performance under tight bandwidth budgets?*

2 Background and Motivation

C2C projects sharer KV caches into receiver space and fuses them with learned gates. This retains richer semantics than text relay. However, KV caches are large: they scale with sequence length, heads, and hidden size. Quantization and cache-length reduction can shrink the communication footprint.

3 Related Work

Layer-selective cache communication has been explored in KVComm and Q-KVComm, which emphasize selecting layers and quantizing shared caches (KVComm, Q-KVComm). Token-level KV selection has been studied for single-model inference (e.g., ZipCache and TokenSelect), and value-norm criteria have been shown to outperform attention-only scores for token importance (VATP) (ZipCache, TokenSelect, VATP). Our work is grounded in C2C's cache projection and fusion (C2C) and relates to recent KV-cache alignment and latent communication efforts (KV Cache Alignment, Latent Collaboration). Our contribution is distinct: we apply projector-aware, token-level sparsity within a C2C projector+fuser pipeline for heterogeneous models, and quantify accuracy-per-byte tradeoffs.

4 Method

4.1 Post-Training Quantization (PTQ)

We quantize the KV caches using INT8 or INT4/NF4 with per-head scaling. We evaluate accuracy and latency under fixed precision budgets.

4.2 Cache-Length Reduction

We prune KV tokens using a fixed ratio (e.g., top-50%, 25%, 10%), reducing transmitted bytes further.

4.3 Selective & Compressed Cache Transfer (SparseC2C)

We select a sparse subset of token positions to transfer and fuse. Let $I \subset \{1, \ldots, T\}$ be selected tokens and S_I the gather operator. We fuse only selected tokens and scatter updates back:

$$(\tilde{K}_\ell^R, \tilde{V}_\ell^R) = S_I^\top(K_\ell^R, V_\ell^R), \quad (\tilde{K}_\ell^S, \tilde{V}_\ell^S) = S_I^\top(K_\ell^S, V_\ell^S)$$

$$(\tilde{K}_\ell^{R\prime}, \tilde{V}_\ell^{R\prime}) = \mathcal{F}_\ell\big(\tilde{K}_\ell^R, \tilde{V}_\ell^R, \Pi_\ell^K(\tilde{K}_\ell^S), \Pi_\ell^V(\tilde{V}_\ell^S)\big)$$

We then scatter the update to the full cache. We use *projector-aware* token scoring by computing value norms in receiver space ("proj_vnorm_topk"), which ties selection to the cross-model mapping.

4.4 Communication-Budget Curves

We report accuracy as a function of transmitted bytes, enabling fair comparison under equal communication constraints.

5 Experiments

5.1 Setup

We evaluate on OpenBookQA and ARC-C with a Qwen3-0.6B receiver and Qwen2.5-0.5B sharer. All models are frozen; only the projector is trained when QAT is enabled.

5.2 Main Results

We report preliminary cache-length pruning results from INT8 PTQ runs. Baseline FP16/INT8/INT4 full runs are currently partial (100/169 samples) and are omitted here; we will re-run them to match the full 500/1150 sample counts.

Table 1: OpenBookQA accuracy (%, 500 samples) for cache-length pruning (INT8).

Order mode	75%	50%	25%	10%
Front	44.6	43.0	38.8	38.6
Back	52.2	52.0	50.8	49.2

Table 2: ARC-C accuracy (%, 1150 samples) for cache-length pruning (INT8).

Order mode	75%	50%	25%	10%
Front	40.2	46.3	38.3	40.7
Back	55.7	57.2	56.2	53.7

5.3 Communication-Budget Curve

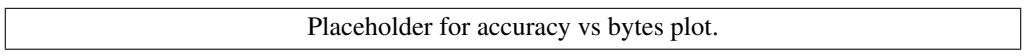

Figure 1: Accuracy vs bytes transmitted (placeholder).

6 Discussion

Quantized C2C provides large bandwidth reductions with limited accuracy drop. Cache pruning further improves the tradeoff, suggesting a practical path to deployable multi-LLM communication.

7 Conclusion

We introduce precision-aware C2C and report accuracy vs bytes curves. This establishes a communication-budget perspective for cross-model KV transfer and opens the door to low-latency, low-bandwidth agent collaboration.

Acknowledgments

Placeholder.