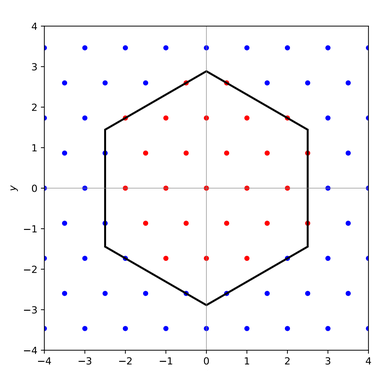
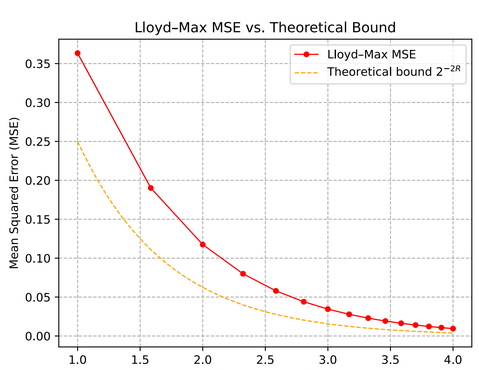
Hardware Costs of Vector Quantization

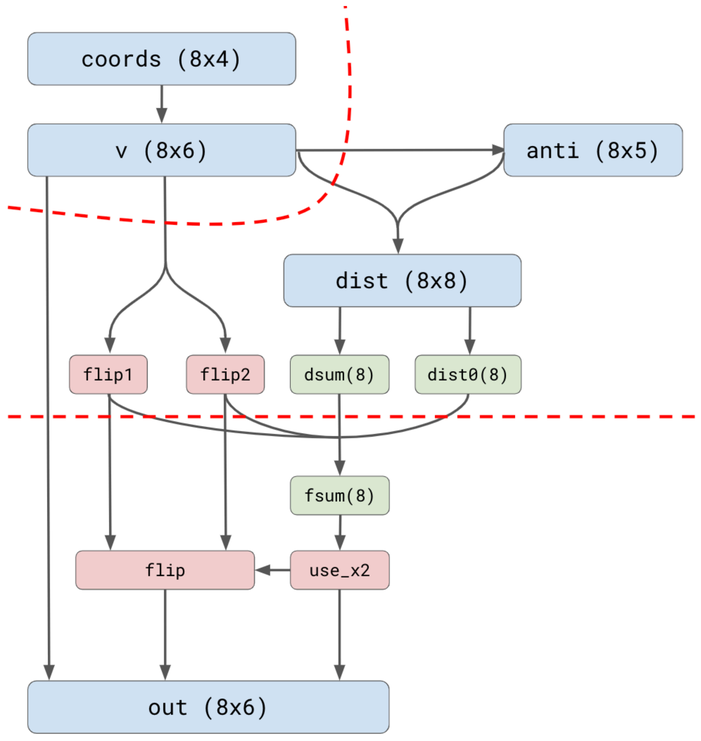
Semyon Savkin, Jose Ricardo Ramos

During inference—especially with reasoning LLMs—the most time and energy intensive operations are memory-bound matrix–vector multiplies, where fetching the weight matrix from DRAM dwarfs the arithmetic cost.. Quantization is a technique of reducing the precision of the data (in our case, model weights) to decrease the storage and reduce the memory-bandwidth bottleneck.

NestQuant is a lattice-based vector quantization method. We fix a lattice Λ in dimension d, which will be the dimension of our VQ. When quantizing a point $x \in \mathbb{R}^d$, we first find the closest (by Euclidean distance) lattice element $y \in \Lambda$, denoted $y = V_\Lambda(x)$. To assign bitstrings, we use a nested-lattice approach: let $q = 2^r$, then y is assigned a bitstring iff $V_{q\Lambda}(y) = 0$. In NestQuant, we choose Λ to be the E_8 lattice, enabling a compact representation with high accuracy.



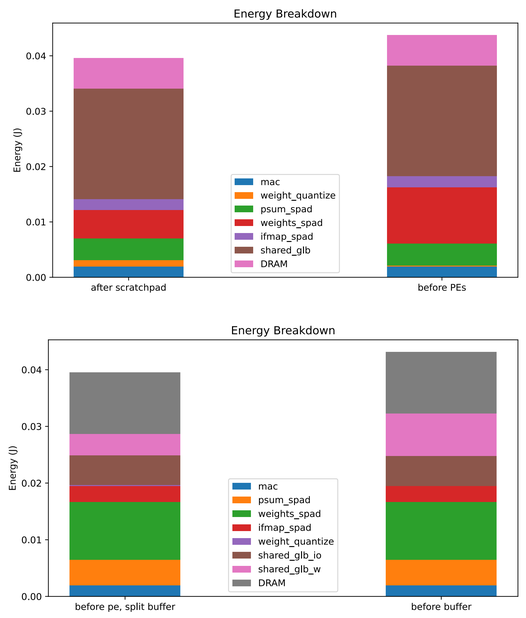


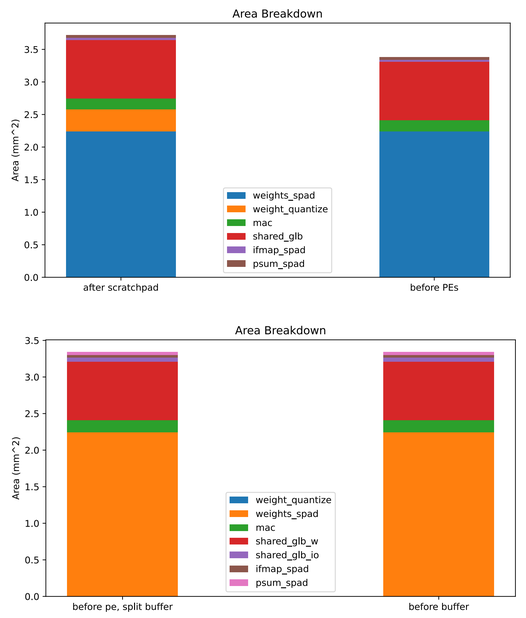


Dequantization consists of two stages: (1) multiply the quantized coordinates by the generating matrix G; (2) subtract the closest point in the scaled lattice $q\Lambda$. For 4-bit quantization with q=16, we implement 2G via bit-shifts, keep only the 6 least significant bits, and resolve rounding by a single 'flip' of the first component. Our pipelined Minispec design supports both 2-stage and 3-stage variants.

We designed a fully-pipelined de-quantization circuit in Minispec. First, the 4-bit codeword is expanded by bit-shifts to approximate multiplication by the E_8 generator matrix G. Next, a lightweight "flip" unit corrects rounding on the MSB. Finally, the scaled lattice offset ($q\Lambda$) is subtracted to recover the real-valued vector. We synthesized both a 2-stage variant (lower area, longer critical path) and a 3-stage variant (higher area, shorter delay) to explore the latency/area trade-off.

Using CiMLoop at 65 nm and a three-stage pipeline, we estimate our quantization circuit's area (scaled by the square of the technology ratio) at ~3 000 μm^2 and its energy (using $\alpha \cdot C_{gate} \cdot V^2$ with C_{gate} = 2 fF, V = 1 V) at ~1 pJ. We integrate this into an Eyeriss-style accelerator (DRAM, shared global buffer, 14×8 PEs with 8-bit dequantized versus 4-bit quantized data) and model the dequantizer as a 32-bit buffer attachment. By comparing placements—adjacent to each PE (before the MAC) versus after the weight buffer—we find that moving the dequantizer closer to the MAC multiplies its instances and invocations, raising both area and energy, whereas locating it after the weight buffer doubles buffer bitwidth savings and reduces memory-access energy. Since dequantizer energy per eight elements is bounded by a single MAC's cost and is negligible relative to buffer energy, the best trade-off is to place the dequantizer after the weight buffer, maximizing net area and energy savings.





	Workload	Config	DRAM (area)	DRAM (energy)	ifmap_spad (area)	ifmap_spad (energy)	mac (area)	mac (energy)	psum_spad (area)	psum_spad (energy)	shared_glb (area)	shared_glb (energy)	shared_glb_io (area)	shared_glb_io (energy)	shared_glb_w (area)	shared_glb_w (energy)	weight_quantize (area)	weight_quantize (energy)	weights_spad (area)	weights_spad (energy)	Total (area)	Total (energy)
0	GEMM	w0	nan	5.50e-03 (13.89%)	3.40e-08 (0.91%)	1.97e-03 (4.97%)	1.68e-07 (4.52%)	1.97e-03 (4.99%)	4.17e-08 (1.12%)	3.98e-03 (10.05%)	8.97e-07 (24.12%)	2.00e-02 (50.44%)	nan	nan	nan	nan	3.38e-07 (9.09%)	1.10e-03 (2.78%)	2.24e-06 (60.24%)	5.10e-03 (12.89%)	3.72e-06 (100.00%)	3.96e-02 (100.00%)
1	GEMM	w3	nan	5.50e-03 (12.57%)	3.40e-08 (1.00%)	1.97e-03 (4.50%)	1.68e-07 (4.97%)	1.97e-03 (4.52%)	4.17e-08 (1.23%)	3.98e-03 (9.10%)	8.97e-07 (26.50%)	2.00e-02 (45.66%)	nan	nan	nan	nan	3.02e-09 (0.09%)	1.38e-04 (0.31%)	2.24e-06 (66.20%)	1.02e-02 (23.33%)	3.38e-06 (100.00%)	4.37e-02 (100.00%)
2	GEMM	w3s	nan	1.09e-02 (27.49%)	3.40e-08 (1.02%)	2.81e-03 (7.12%)	1.68e-07 (5.03%)	1.97e-03 (5.00%)	4.17e-08 (1.25%)	4.51e-03 (11.41%)	nan	nan	5.51e-08 (1.65%)	5.28e-03 (13.36%)	7.99e-07 (23.92%)	3.74e-03 (9.46%)	3.02e-09 (0.09%)	1.38e-04 (0.35%)	2.24e-06 (67.04%)	1.02e-02 (25.82%)	3.34e-06 (100.00%)	3.95e-02 (100.00%)
3	GEMM	w4	nan	1.09e-02 (25.19%)	3.40e-08 (1.02%)	2.81e-03 (6.53%)	1.68e-07 (5.03%)	1.97e-03 (4.58%)	4.17e-08 (1.25%)	4.51e-03 (10.45%)	nan	nan	5.51e-08 (1.65%)	5.28e-03 (12.24%)	7.99e-07 (23.92%)	7.48e-03 (17.34%)	3.02e-09 (0.09%)	5.37e-07 (0.00%)	2.24e-06 (67.04%)	1.02e-02 (23.66%)	3.34e-06 (100.00%)	4.31e-02 (100.00%)
4	DNN	w0	nan	1.35e-03 (18.13%)	3.40e-08 (0.91%)	3.54e-04 (4.74%)	1.68e-07 (4.52%)	4.17e-04 (5.59%)	4.17e-08 (1.12%)	8.19e-04 (10.98%)	8.97e-07 (24.12%)	3.22e-03 (43.14%)	nan	nan	nan	nan	3.38e-07 (9.09%)	2.32e-04 (3.11%)	2.24e-06 (60.24%)	1.07e-03 (14.31%)	3.72e-06 (100.00%)	7.46e-03 (100.00%)
5	DNN	w3	nan	1.34e-03 (16.29%)	3.40e-08 (1.00%)	3.45e-04 (4.18%)	1.68e-07 (4.97%)	4.17e-04 (5.05%)	4.17e-08 (1.23%)	8.19e-04 (9.93%)	8.97e-07 (26.50%)	3.16e-03 (38.33%)	nan	nan	nan	nan	3.02e-09 (0.09%)	2.74e-05 (0.33%)	2.24e-06 (66.20%)	2.14e-03 (25.89%)	3.38e-06 (100.00%)	8.25e-03 (100.00%)
6	DNN	w3s	nan	3.19e-03 (30.27%)	3.40e-08 (1.02%)	4.37e-04 (4.15%)	1.68e-07 (5.03%)	4.17e-04 (3.96%)	4.17e-08 (1.25%)	9.84e-04 (9.35%)	nan	nan	5.51e-08 (1.65%)	1.41e-03 (13.40%)	7.99e-07 (23.92%)	1.56e-03 (14.84%)	3.02e-09 (0.09%)	5.62e-05 (0.53%)	2.24e-06 (67.04%)	2.47e-03 (23.50%)	3.34e-06 (100.00%)	1.05e-02 (100.00%)
7	DNN	w4	nan	3.06e-03 (27.41%)	3.40e-08 (1.02%)	4.37e-04 (3.91%)	1.68e-07 (5.03%)	4.17e-04 (3.73%)	4.17e-08 (1.25%)	9.91e-04 (8.87%)	nan	nan	5.51e-08 (1.65%)	1.10e-03 (9.84%)	7.99e-07 (23.92%)	2.77e-03 (24.78%)	3.02e-09 (0.09%)	1.72e-06 (0.02%)	2.24e-06 (67.04%)	2.40e-03 (21.45%)	3.34e-06 (100.00%)	1.12e-02 (100.00%)

Embedding a pipelined NestQuant dequantizer adds only 3 000 um^2 and 1 pJ per 8-word decode—negligible compared to DRAM/SRAM energy when we halve buffer bit-widths. Placing the dequantizer just after the weight buffer maximizes these savings without multiplying hardware instances. This shows that vector-quantization modules can deliver the accuracy of 4-bit VQ at minimal area/energy cost, unlocking more aggressive low-bit schemes in DNN accelerators.