

DRAM Latency Model

PIM Optimizer

December 23, 2025

1 DRAM Latency Model

DRAM latency consists of two components: RowBuffer data transfer and Row Activation. Since RowBuffer is part of DRAM, these two latencies should be summed together.

1.1 Per-Datatype DRAM Latency

For each datatype $t \in \{\text{input}, \text{weight}, \text{output}\}$:

$$\text{DRAM_latency}_t = \underbrace{\frac{\text{mem_reads}_t}{\text{rowbuffer_bandwidth}}}_{\text{RowBuffer Data Transfer}} + \underbrace{\text{row_acts}_t \times \text{activation_latency}}_{\text{Row Activation}} \quad (1)$$

Where:

- mem_reads_t — Number of bytes read from RowBuffer for datatype t
- $\text{rowbuffer_bandwidth}$ — RowBuffer read bandwidth (bytes/cycle)
- row_acts_t — Number of row activations for datatype t
- $\text{activation_latency}$ — Latency per row activation (cycles), typically 25 cycles

1.2 Total DRAM Latency

The total DRAM latency is the maximum of all three datatype latencies:

$$\text{DRAM_latency}_{\text{total}} = \max(\text{DRAM_latency}_{\text{input}}, \text{DRAM_latency}_{\text{weight}}, \text{DRAM_latency}_{\text{output}}) \quad (2)$$

1.3 Expanded Form

Combining the above equations:

$$\text{DRAM_latency}_{\text{total}} = \max_{t \in \{I, W, O\}} \left(\frac{\text{mem_reads}_t}{\text{BW}_{\text{rowbuf}}} + \text{row_acts}_t \cdot T_{\text{act}} \right) \quad (3)$$

Where:

- $I = \text{input}$, $W = \text{weight}$, $O = \text{output}$
- $\text{BW}_{\text{rowbuf}} = \text{RowBuffer bandwidth}$
- $T_{\text{act}} = \text{Row activation latency}$

2 Overall Latency Model

The overall latency is determined by the bottleneck among compute and memory:

$$\text{Latency} = \max(\text{Compute_cycles}, \text{DRAM_latency}_{\text{total}}, \text{GlobalBuffer_latency}, \dots) \quad (4)$$