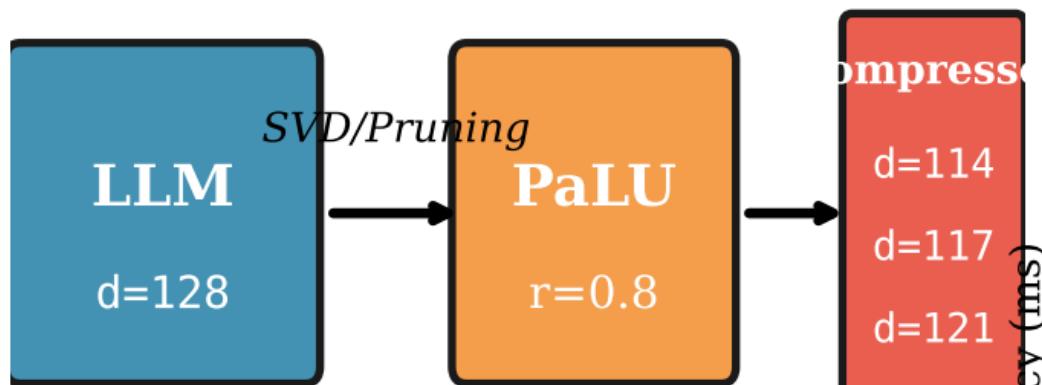


(a) Compression Produces Irregular Dimensions



96.9% misaligned

GPU Alignment Requirements

FlashAttention: $d \% 8 = 0$ | Tensor Core: $K \% 16 = 0$ | float4: $K \% 4 = 0$

(b) SDPA Latency Cliff Effect

