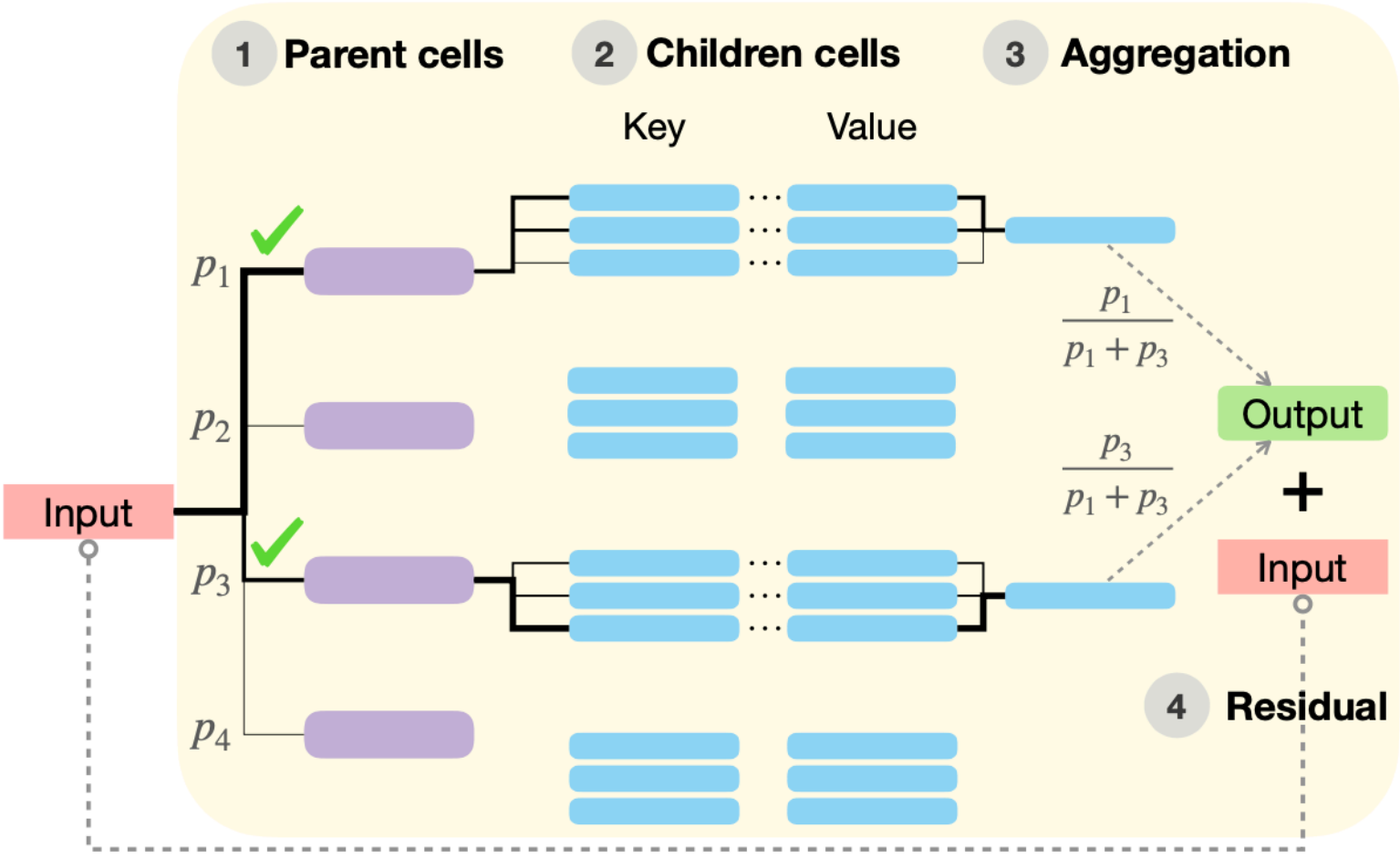

Figure 1: SPARTAN adds a sparse, hierarchically organized memory after each Transformer layer that is shared by all positions in the input sequence. ① The input corresponds to a single position, and chooses a sparse subset of parent cells based on a computed probability distribution (here, top-2). ② The corresponding children cells are used to compute an input-conditioned representation, ③ which is aggregated via a weighted sum based on the probability distribution in step 1 to give the output. ④ It is added to the input through a residual connection which serves as the input to the next Transformer layer. SPARTAN outperforms baselines on GLUE while giving a 90% increase in throughput on resource-constrained devices (§ 5).