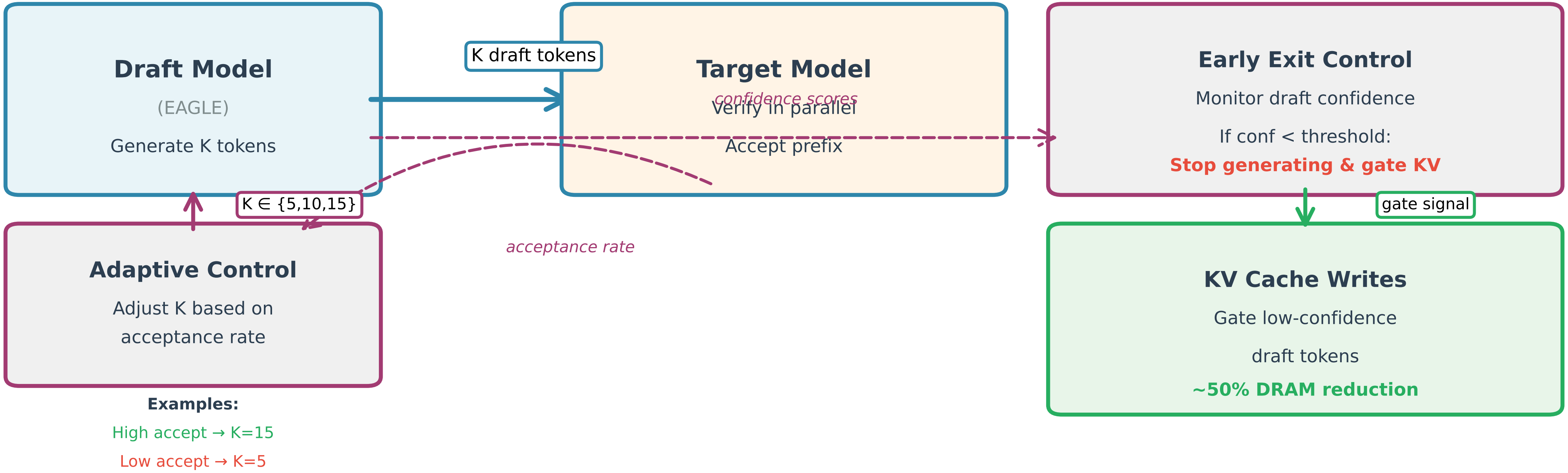
Elastic Speculation Overview

Adaptive Draft Length + Confidence-Based Early Exit (KV Gating)



Draft Model

(EAGLE)

Generate K tokens

Target Model

Verify in parallel

Accept prefix

Early Exit Control

Monitor draft confidence

If conf < threshold:

Stop generating & gate KV

Adaptive Control

Adjust K based on acceptance rate

KV Cache Writes

Gate low-confidence

draft tokens

~50% DRAM reduction

Examples:

High accept → K=15

Low accept → K=5

□ Adaptive Draft Length

~30–50% lower latency vs. fixed-length

Dynamically adjusts speculation depth

□ Confidence-Based Early Exit

~50% less KV-cache DRAM traffic

Minimal latency impact (~1–3%)