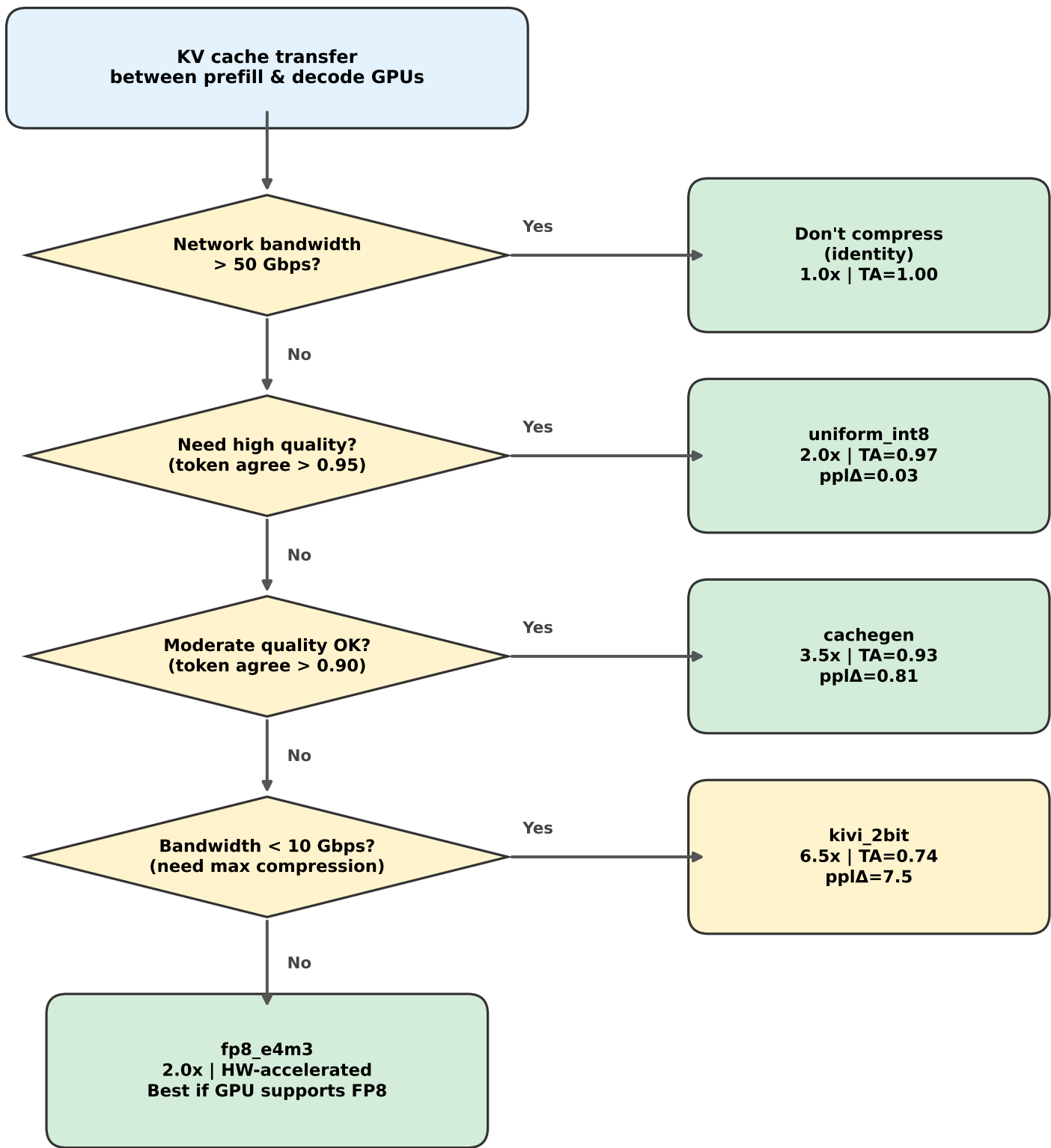


KVShuttle Decision Flowchart

Which KV cache compressor should I use for disaggregated serving?



Avoid for
generation tasks:

uniform_int4
TA=0.59 | pplΔ=13k

cascade
TA=0.33 | pplΔ=18k

Legend

TA = Token Agreement (greedy decode match rate)

pplΔ = Perplexity delta vs uncompressed

N.Nx = Compression ratio

Break-even: GPU pipelined, Tesla T4

Quality: FP16, 5 models, 50 WikiText prompts