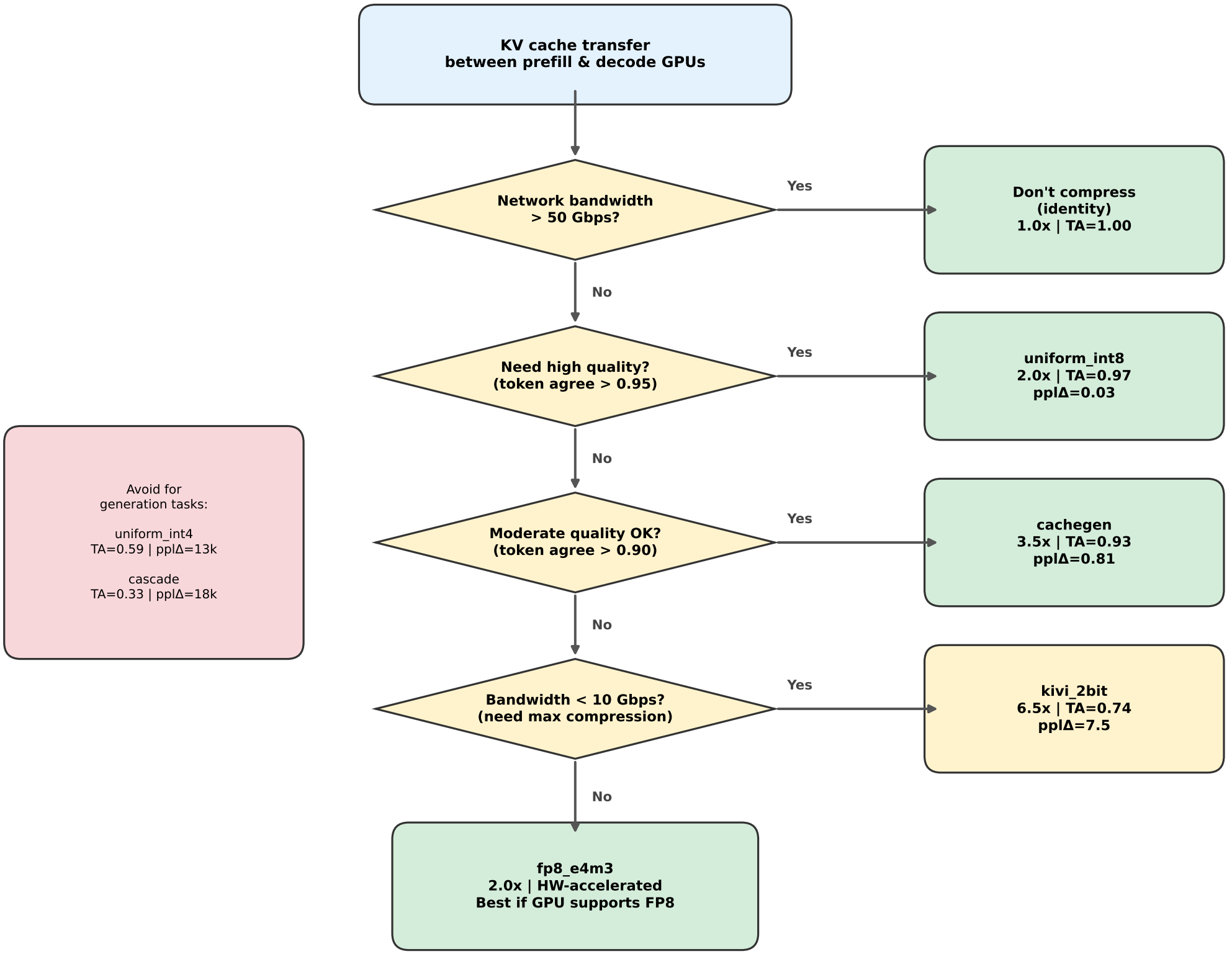
KVShuttle Decision Flowchart

Which KV cache compressor should I use for disaggregated serving?



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