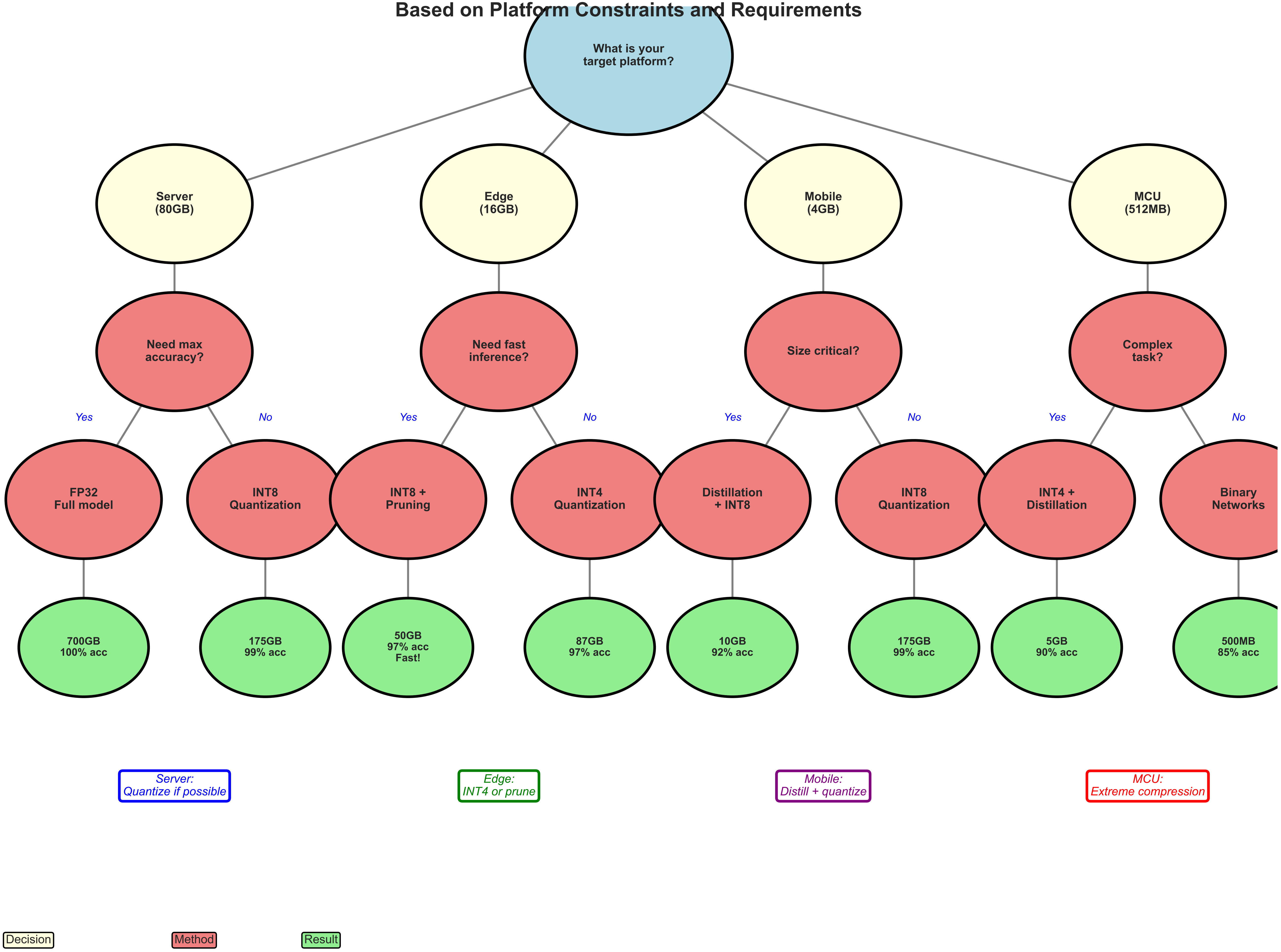
Compression Decision Tree: Choose Your Strategy Based on Platform Constraints and Requirements


What is your target platform?
Server (80GB)
Edge (16GB)
Mobile (4GB)
MCU (512MB)
Need max accuracy?
Need fast inference?
Size critical?
Complex task?
Yes
No
Yes
No
Yes
No
Yes
No
FP32 Full model
INT8 Quantization
INT8 + Pruning
INT4 Quantization
Distillation + INT8
INT8 Quantization
INT4 + Distillation
Binary Networks
700GB 100% acc
175GB 99% acc
50GB 97% acc Fast!
87GB 97% acc
10GB 92% acc
175GB 99% acc
5GB 90% acc
500MB 85% acc
Server: Quantize if possible
Edge: INT4 or prune
Mobile: Distill + quantize
MCU: Extreme compression
Decision
Method
Result