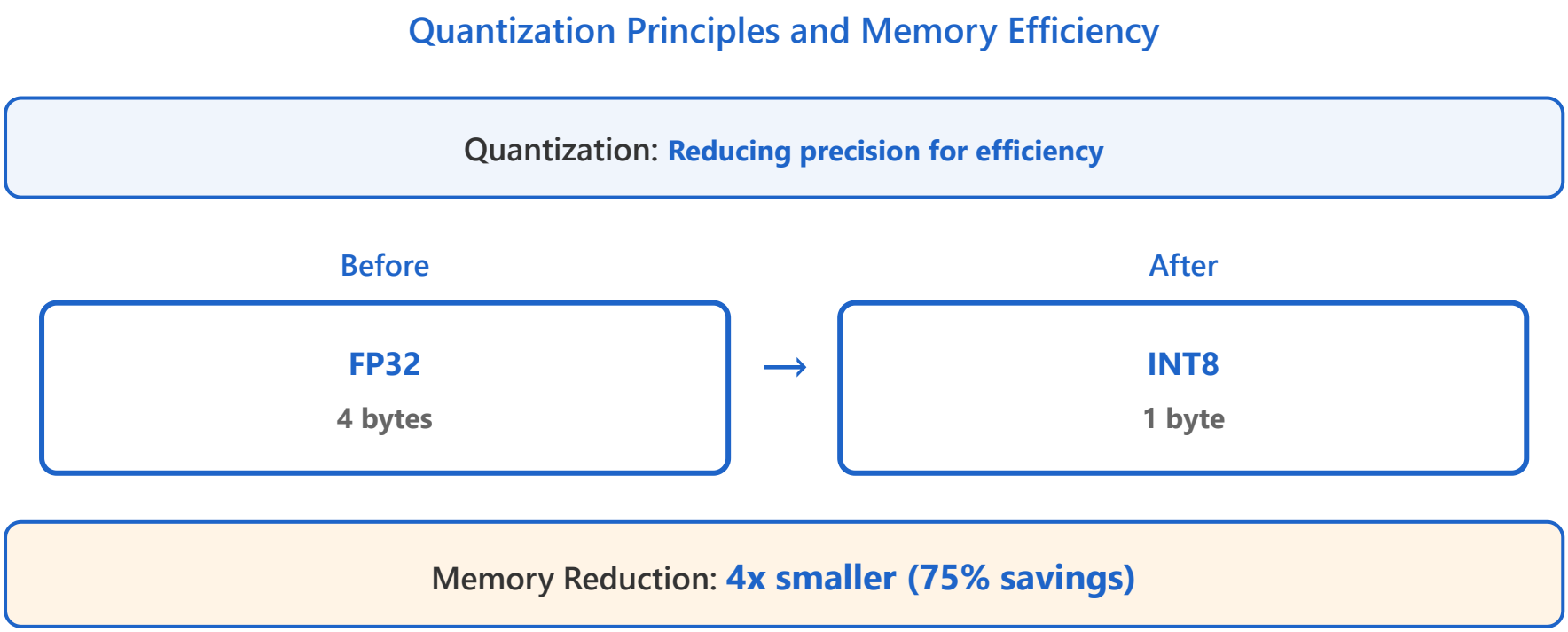
Quantization Principles and Memory Efficiency
Quantization: Reducing precision for efficiency
Before
FP32
4 bytes
After
INT8
1 byte
Memory Reduction: 4x smaller (75% savings)

Quantization Methods

- Post-training quantization
- Quantization-aware training

Accuracy Impact

- Minimal accuracy loss
- Typically <1-2% degradation
- With proper quantization

Real-World Applications

Mobile Deployment

Edge Devices

Faster Inference