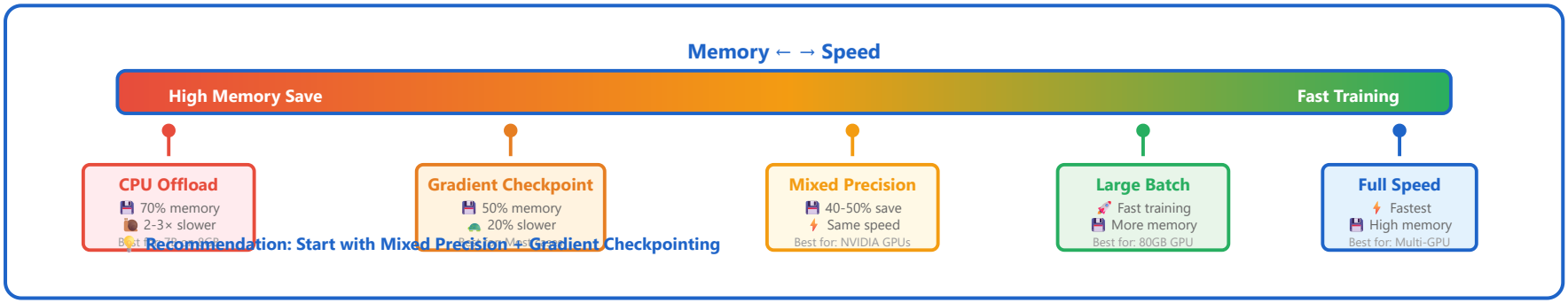
Memory-Compute Trade-offs

GPU Memory vs Training Time
Optimization strategies for limited resources



Memory Optimization Techniques

- **Gradient Checkpointing:** 50% memory, 20% slower
- **Batch Size Reduction:** Linear memory scaling
- **Gradient Accumulation:** Effective batch size increase
- **Mixed Precision:** 40-50% memory reduction
- **CPU Offloading:** 70% memory, 2-3x slower

Recommended Configurations

- **24GB GPU:** 7B models with QLoRA, batch=4
- **40GB GPU:** 13B models with QLoRA, batch=8
- **80GB GPU:** 30B models with LoRA, batch=16

- **Multi-GPU:** 70B models with DeepSpeed ZeRO-3

📈 Resource Planning

- Profile memory usage before full training run
- Leave 20% GPU memory as buffer
- Monitor CUDA OOM errors and adjust