

Training  
Dense

Inference  
Dense

Low Speed  
High Power



Training  
Dense

Fine-tune  
Sparse

Inference  
Sparse

High Speed  
Low Power



GPU

GPU ?

GPU/FPGA/ASIC