Abstract. Software solutions on GPU platforms has limits. -> dedicated hardware on FPGA.

CPU, GPU, FPGA, ASIC.

GPU -> easy platforms, fast enough. But power hungry comparatively.

ASIC -> long design cycle.

1. Algorithm optimization, 2. Quantization, 3. Weight Reduction.

Rooflint model??? [28]

CTC(computational communication ratio). (X) -> operation times / external data access time

Peak computing power (Y)

1. Common Characteristics of algorithms.

Reducing parameters -> less accuracy.

2. Parallel.

Task / Data / HW level parallelism.