

Decoding Kernels Runtime Breakdown with Utilisation Curve (One Layer of DeepSeek-671B, Batch/Chip=16k)

