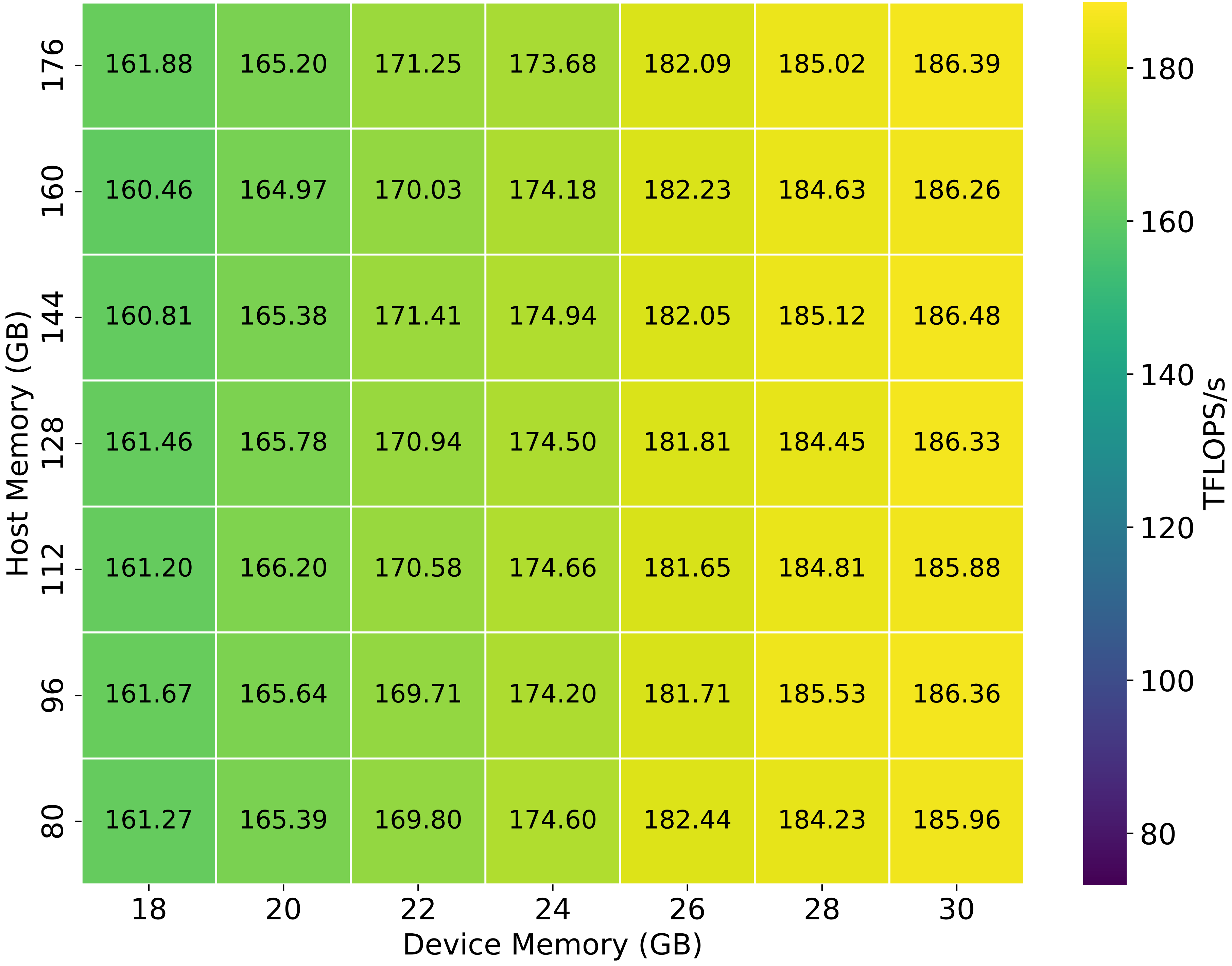
Performance for Model: 8B, Seq Len: 1024 on RTX5090


Host Memory (GB)
Device Memory (GB)
TFLOPS/s