FLOP utilization rates of training runs of given sizes GPU: V100 SXM GPU: A100 SXM GPU: H100 SXM

10²⁹

10³⁰

10³¹

 10^{28}

Training compute (16-bit FLOP)

 10^{32}

8.0

0.6

0.4

0.2

0.0

 10^{25}

 10^{26}

 10^{27}

Best achievable utilization rate