**Optimizer Wall Clock Time vs. Model Size (CUDA)** Step Time (ms) with 95% CI Average ! 10<sup>0</sup> 10<sup>5</sup>  $10^{6}$ 10<sup>7</sup> **Number of Model Parameters** 

Optimizer
--- adagrad
--- adam
--- adamp
--- adan

amsgraddiffgradlion

madgradpulsegradrmsprop