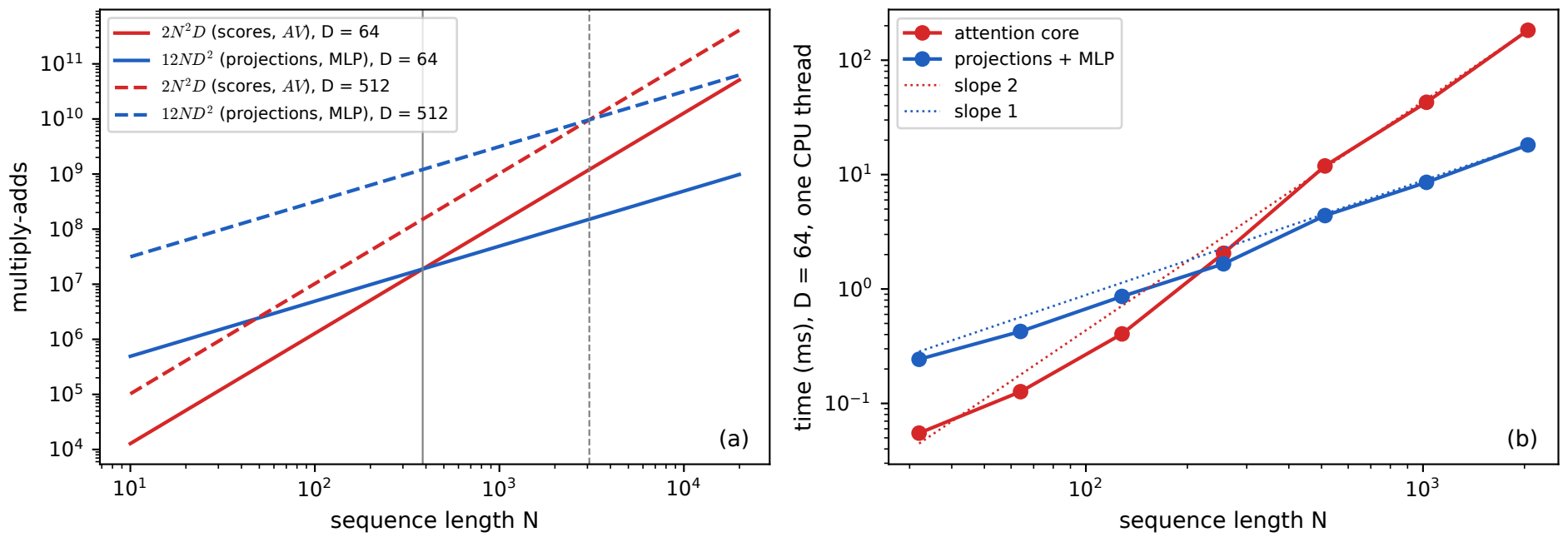

$2N^2D$ (scores, AV), D = 64
$12ND^2$ (projections, MLP), D = 64
$2N^2D$ (scores, AV), D = 512
$12ND^2$ (projections, MLP), D = 512
multiply-adds
sequence length N
(a)
attention core
projections + MLP
slope 2
slope 1
time (ms), D = 64, one CPU thread
sequence length N
(b)