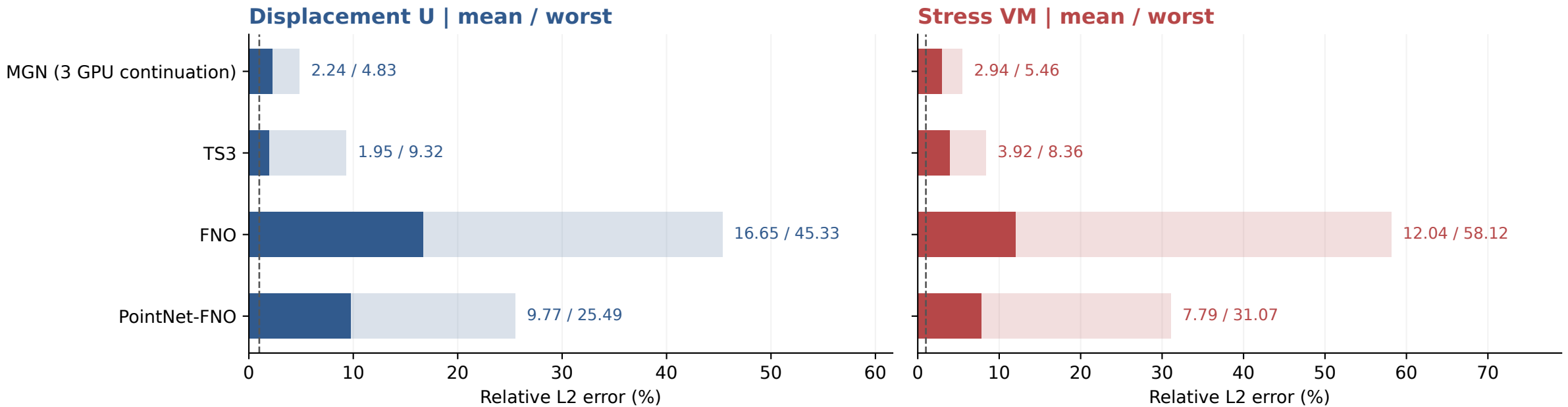
1000-case experiment | FINAL 100-test results

800 train / 100 validation / 100 test; 6000 updates; best selected using validation only



Solid = mean; pale = worst case; dashed = 1%. Negative VM predictions are preserved.