

Performance Comparisons (TF 2.0 on DGX Station)							
Task	GPUs	AMP/XLA	Val Acc (Avg)	Time-to-result	Peak Throughput	E2E Speed-up	Throughput Speed-up
CoLA	1 x V100 16GB		0.824	451	59	1.0	1.0
CoLA	1 x V100 16GB	AMP	0.820	284	98	1.6	1.7
CoLA	1 x V100 16GB	AMP+XLA	0.823	234	157	1.9	2.7
CoLA	4 x V100 16GB		0.812	207	218	2.2	3.7
CoLA	4 x V100 16GB	AMP	0.828	184	365	2.5	6.2
CoLA	4 x V100 16GB	AMP+XLA	0.828	251	505	1.8	8.6
MRPC	1 x V100 16GB		0.835	204	58	1.0	1.0
MRPC	1 x V100 16GB	AMP	0.823	134	97	1.5	1.7
MRPC	1 x V100 16GB	AMP+XLA	0.827	141	158	1.4	2.7
MRPC	4 x V100 16GB		0.827	150	218	1.4	3.8
MRPC	4 x V100 16GB	AMP	0.817	143	365	1.4	6.3
MRPC	4 x V100 16GB	AMP+XLA	0.818	222	529	0.9	9.1
SST2	1 x V100 16GB		0.922	3436	59	1.0	1.0
SST2	1 x V100 16GB	AMP	0.918	2088	97	1.6	1.6
SST2	1 x V100 16GB	AMP+XLA	0.922	1352	158	2.5	2.7
SST2	4 x V100 16GB		0.929	1025	216	3.4	3.7
SST2	4 x V100 16GB	AMP	0.926	672	364	5.1	6.2
SST2	4 x V100 16GB	AMP+XLA	0.925	601	508	5.7	8.6