APPENDIX

A DNN MODELS

In Table 1, we present the model characteristics of 10 deep learning models used in our evaluation. The number of parameters, total size of all parameters, number of computational operations in inference mode and training mode, and the standard batch size are given below.

Neural Network Model	#Par	Total Par Size (MiB)	#Ops Inference/ Training	Batch Size
AlexNet v2 (Krizhevsky, 2014)	16	191.89	235/483	512
Inception v1 (Szegedy et al., 2014)	116	25.24	1114/2246	128
Inception v2 (Ioffe & Szegedy, 2015)	141	42.64	1369/2706	128
Inception v3 (Szegedy et al., 2015)	196	103.54	1904/3672	32
ResNet-50 v1 (He et al., 2015)	108	97.39	1114/2096	32
ResNet-101 v1 (He et al., 2015)	210	169.74	2083/3898	64
ResNet-50 v2 (He et al., 2016)	125	97.45	1423/2813	64
ResNet-101 v2 (He et al., 2016)	244	169.86	2749/5380	32
VGG-16 (Simonyan & Zisserman, 2014)	32	527.79	388/758	32
VGG-19 (Simonyan & Zisserman, 2014)	38	548.05	442/857	32

Table 1: DNN model characteristics

B TAO VS TIO

In Figure 13, we plot the increase in throughput achieved with scheduling in env_C with and without the scheduling schemes (TIC and TAC). We observe that both TIC and TAC offer significant speedup compared to the baseline (no scheduling). Performance of TIC is comparable to that of TAC indicating that we can achieve improved performance without relying on runtime statistics in current models.

Due to the simplicity of TIC algorithm, we use it as the representative algorithm for scheduling in the cloud GPU environment (env_G).

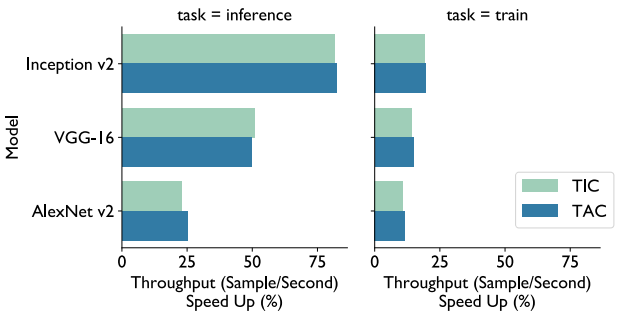

Figure 13: Increase in throughput with the scheduling schemes (TIC and TAC) compared to the baseline (no scheduling). Measured on env_C (CPU-Only).