This model version has very low data transition between host and Device to get the most out of execution timing details. All layers in this model depend on input device matrices and output device matrices. Device Outputs are transmitted from one layer to another