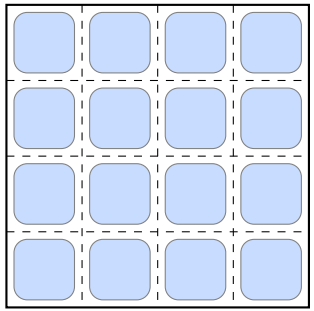
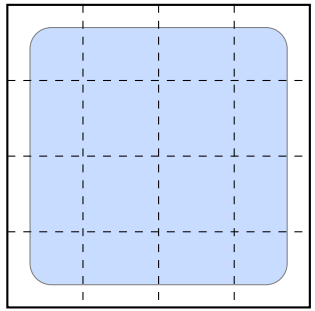
How the *model weights* are split over cores

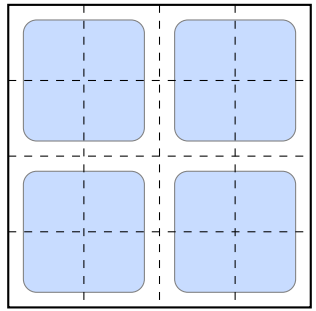
Data Parallelism



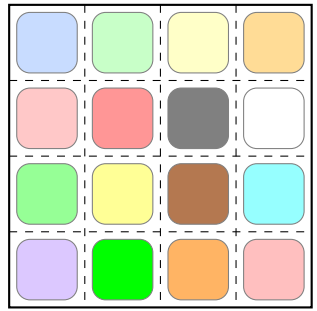
Model Parallelism



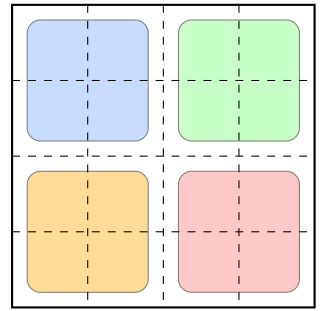
Model and Data Parallelism



Expert and Data Parallelism

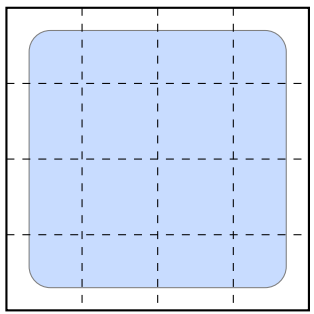


Expert, Model and Data Parallelism

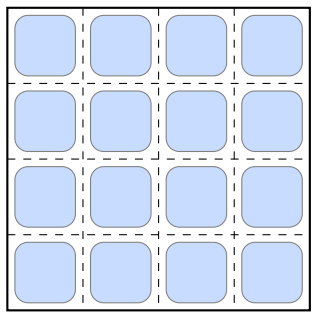


How the *data* is split over cores

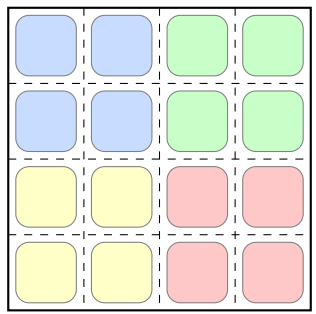
Data Parallelism



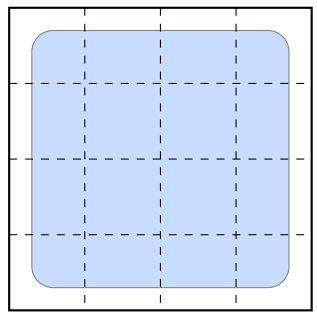
Model Parallelism



Model and Data Parallelism



Expert and Data Parallelism



Expert, Model and Data Parallelism