MoE Transformer Encoder **Transformer MoE Transformer** with device placement Encoder Encoder Encoder Encoder Encoder Encoder output output (shard 1) output (shard E) output Add & Norm Add & Norm Add & Norm Add & Norm Feed Forward Feed Forward Feed Forward Feed Forward FFN FFN FFN FFN Nx Add & Norm Add & Norm Add & Norm Add & Norm Multi-Head Multi-Head Multi-Head Multi-Head Attention Attention Attention Attention (N/2)x(N/2)x(N/2)xInput embeddings + Add & Norm Add & Norm Add & Norm Positional embeddings All-to-All Dispatch FFN₁ FFN_E FFN_1 FFN_E Gating All-to-All Combine Gating Gating Model-parallel Add & Norm Add & Norm Add & Norm MoE Multi-Head Multi-Head Multi-Head Attention Device Attention Attention 1...E Device 1 Device E Input embeddings + Input embeddings + Input embeddings + Positional embeddings Positional embeddings Positional embeddings (shard 1) (shard E)