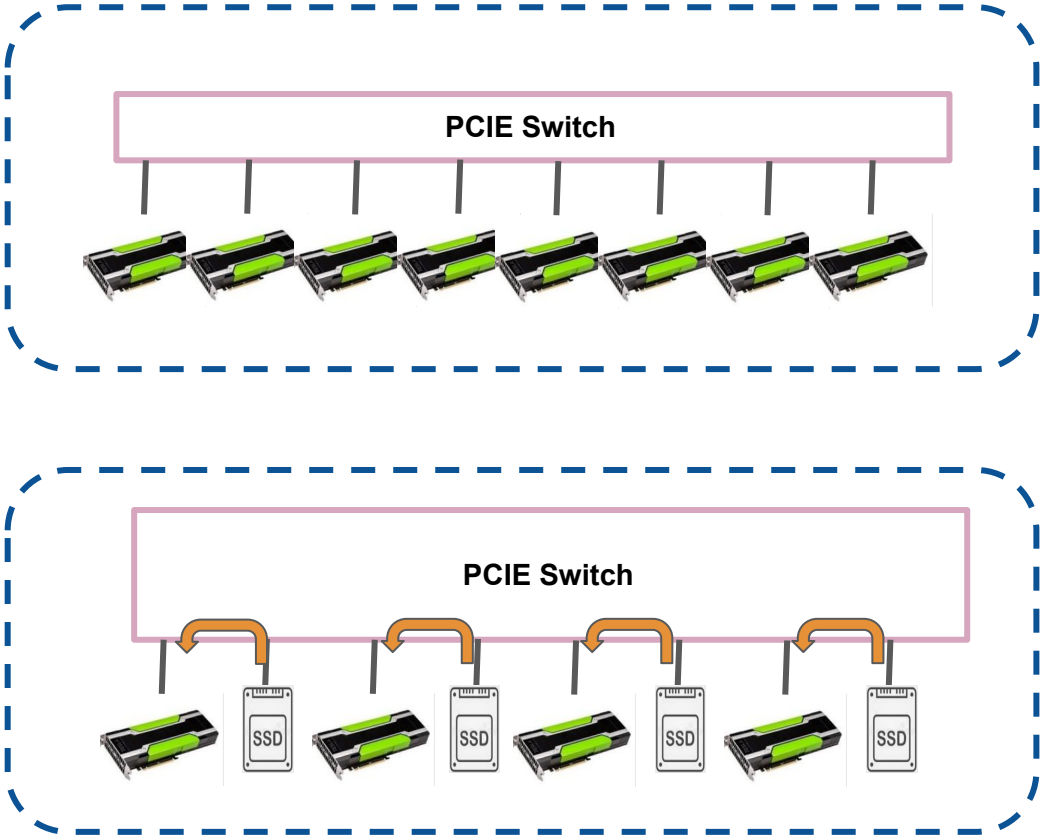
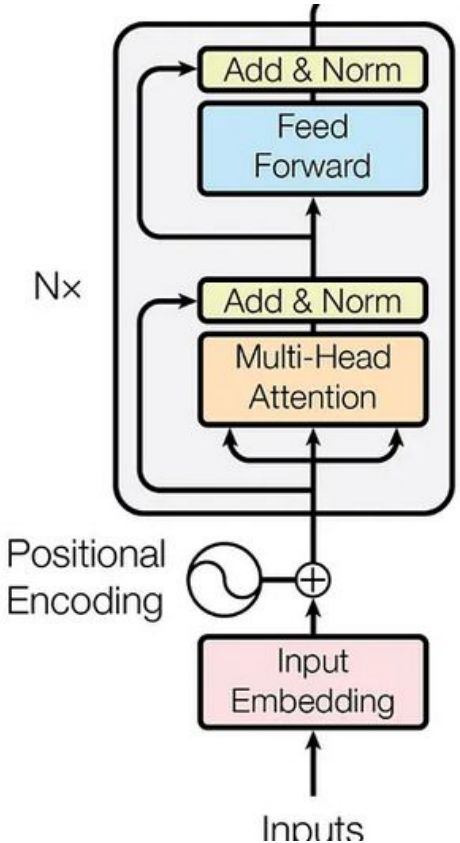
Beyond KV-Cache

GPU Server Restructured



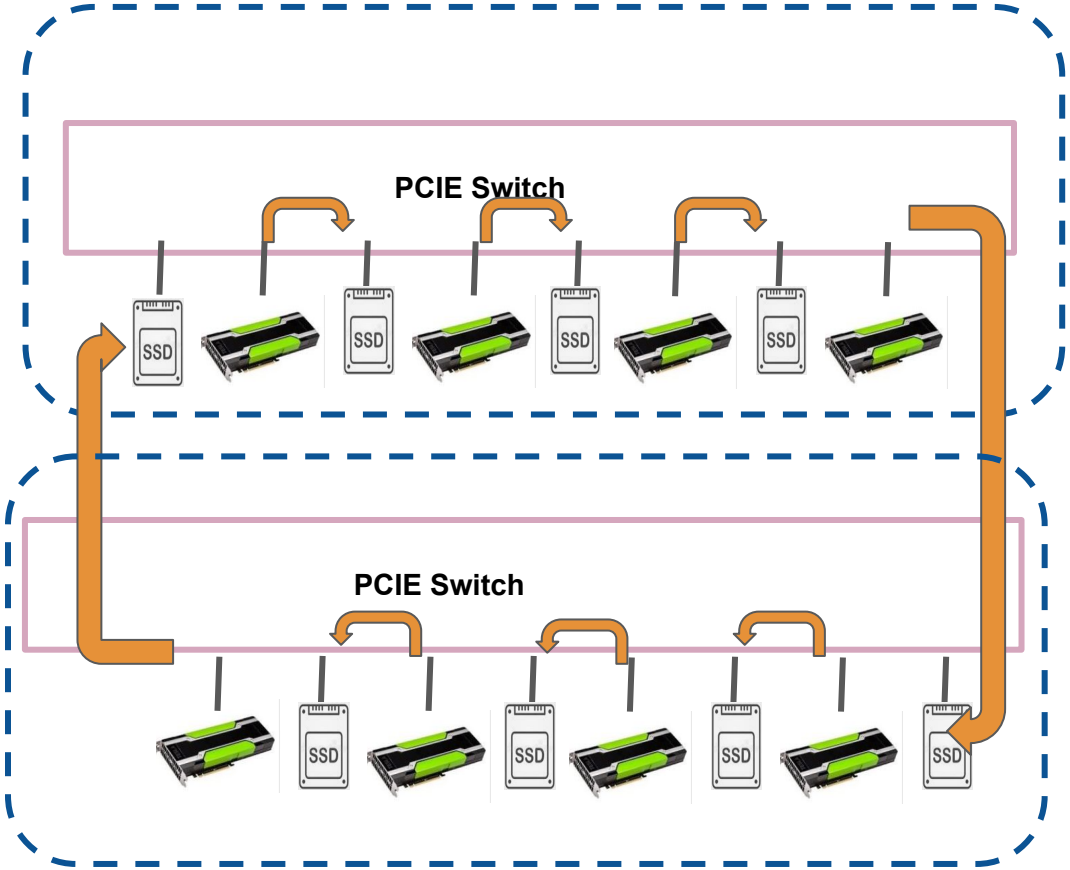
- GPU paired with SSD
- GPU now have dedicated computing scratch pad
- SSD will hold all model parameters
- GPU exchange data through neighbour SSD

SSD LLM inference Streaming Computing



- GPT-3 175B, 96 layer
- Each layer ~1.8B, 3.6G Bytes
- Gen5x16 SSD bandwidth ~ 50GB/s
- Model loading time for one layer: 3.6/50 ~ **70ms**
- Computing: 100K x 1.8B / 100 T ~ **1.8s**

GPU Server Ring Extension



- Commodity GPU, less HBM dependency
- HBM could be fully used for KV-Cache and computing
- Better parallelism could be achieved