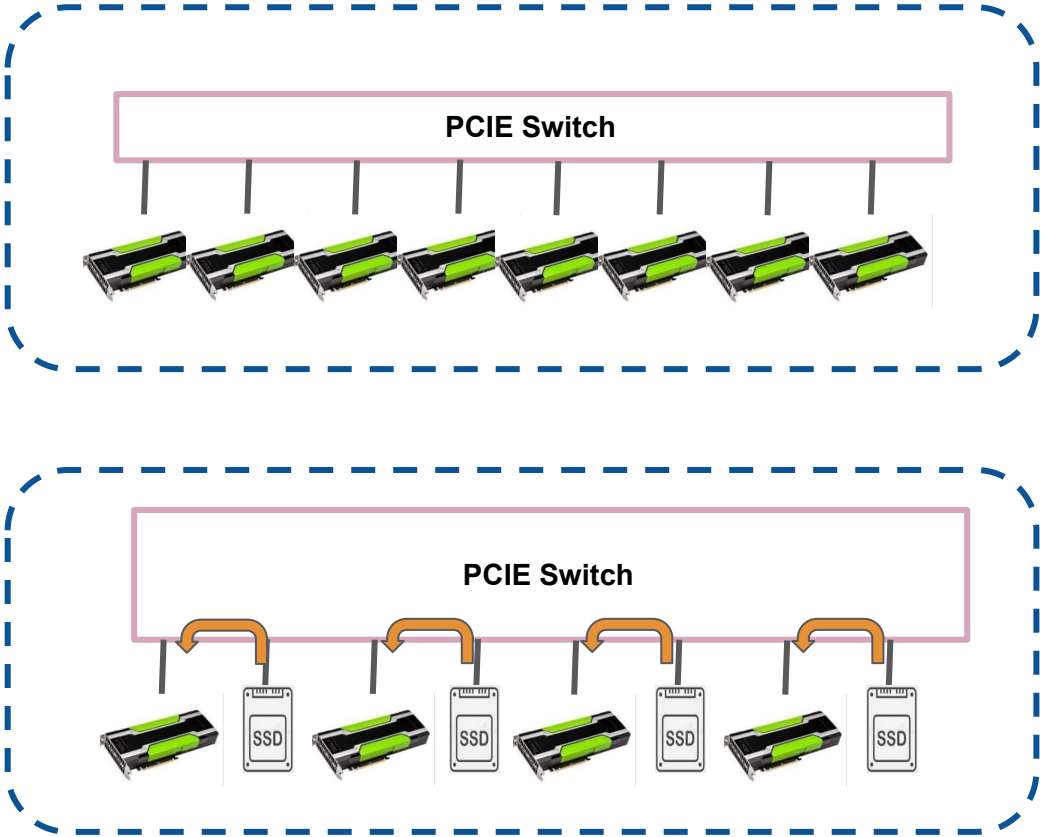
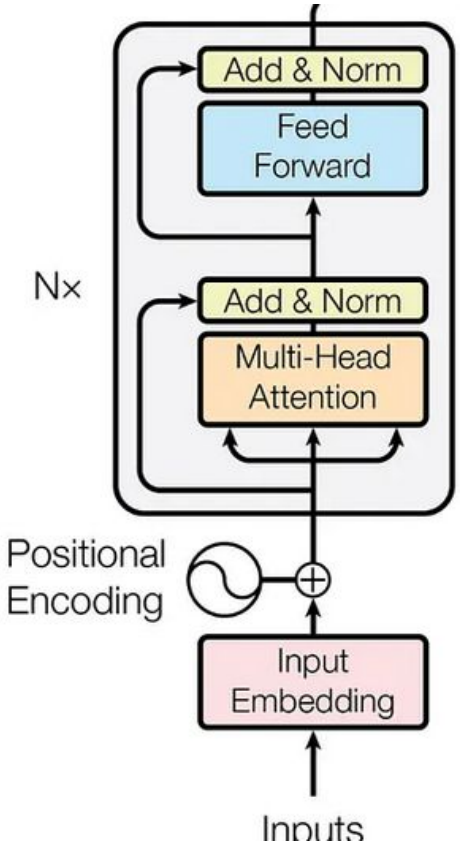
Beyond KV-Cache

GPU Server Restructured



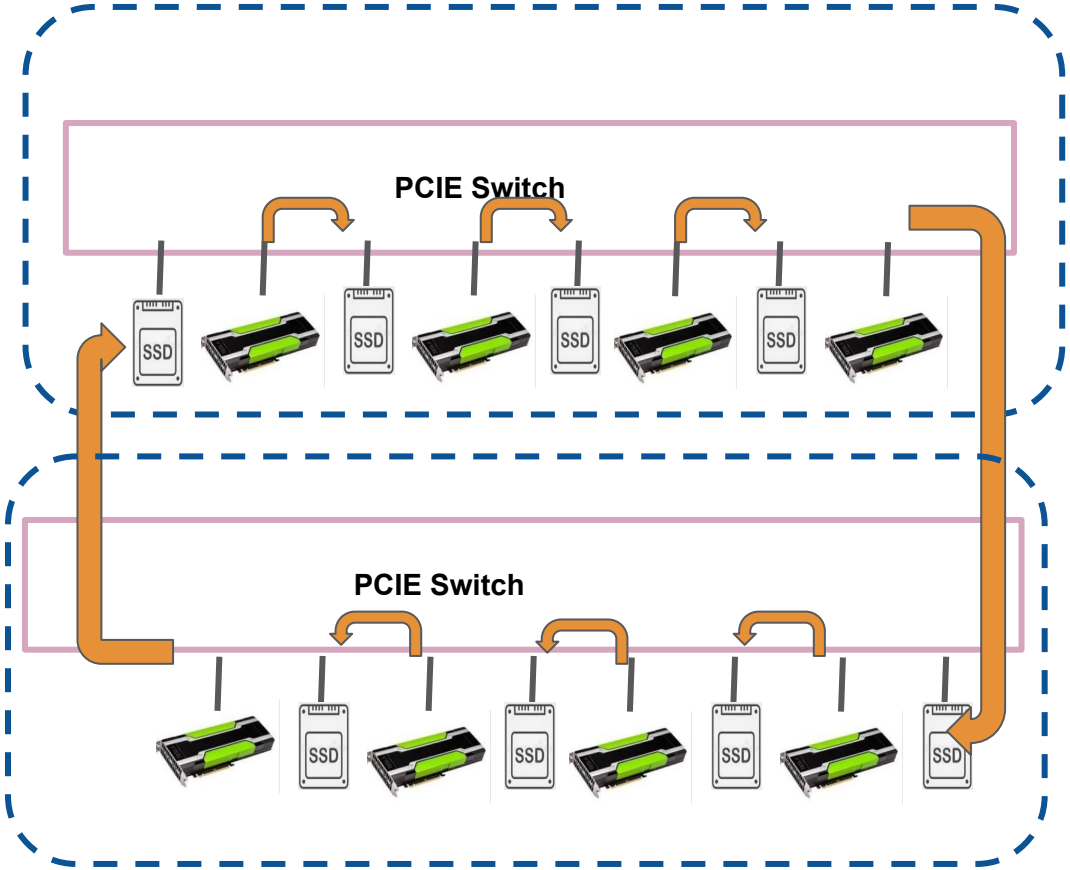
- GPU paired with SSD
- GPU now have dedicated computing scratch pad
- SSD will hold all model parameters
- GPU exchange data through neighbour SSD

SSD LLM inference Streaming Computing



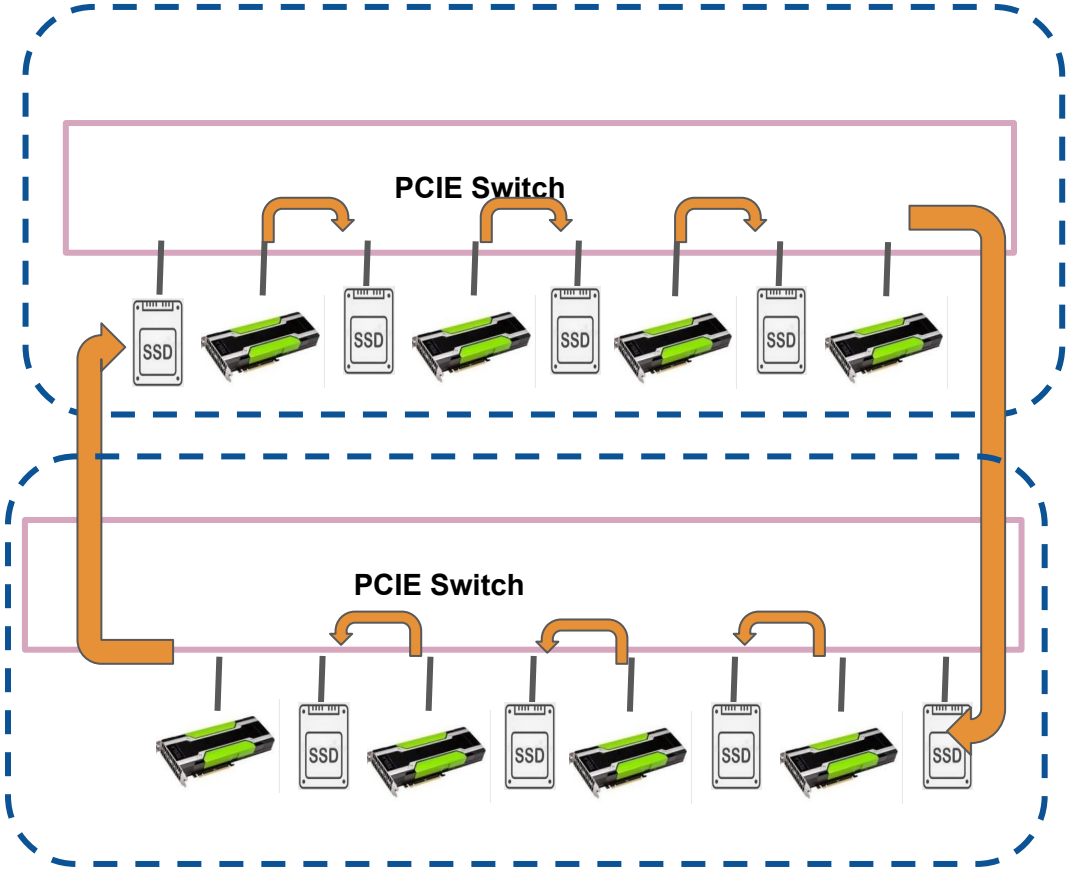
- GPT-3 175B, 96 layer
- Each layer ~1.8B, 3.6G Bytes
- Gen5x16 SSD bandwidth ~ 50GB/s
- Model loading time for one layer: 3.6/50 ~ **70ms**
- Computing: 100K x 1.8B / 100 T ~ **1.8s**

GPU Server Ring Extension



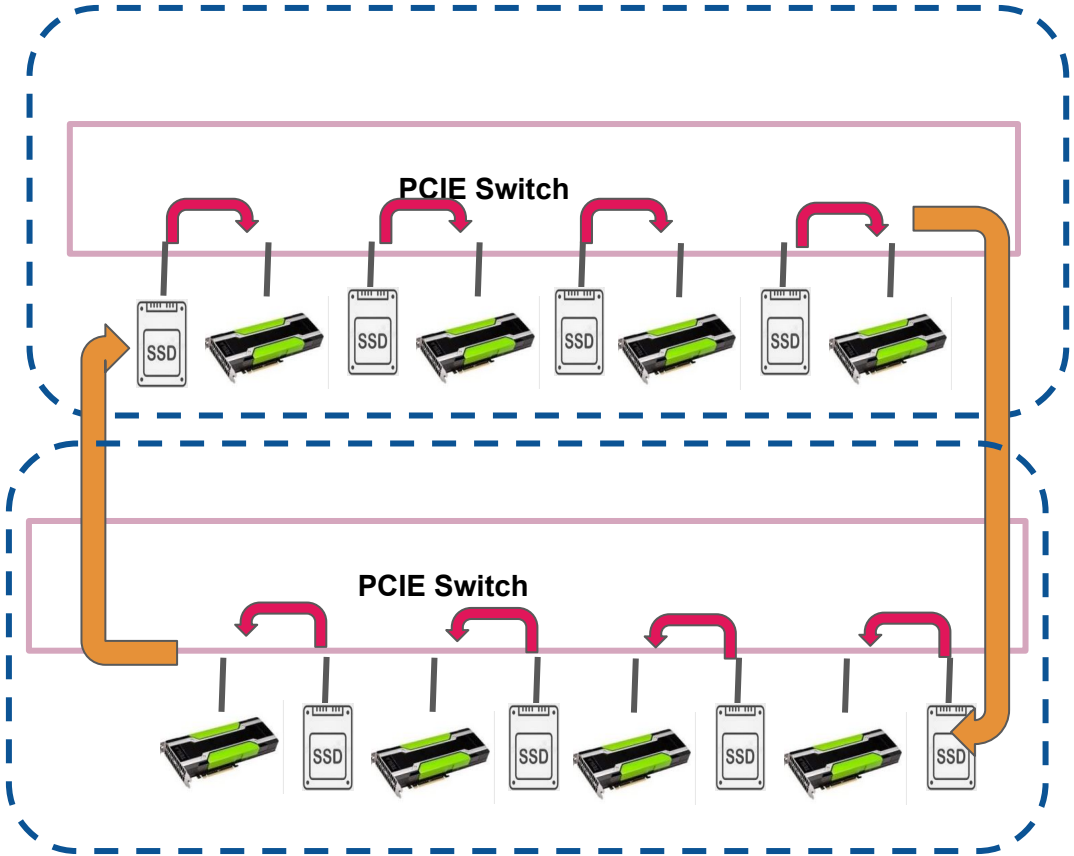
- Commodity GPU, less HBM dependency
- HBM could be fully used for KV-Cache and computing
- Better parallelism could be achieved

SSD-PCIE Ring Reduction - Write Data



- Efficient All - All communication mechanism for partitioned matrix on different GPU
- GPU write intermediate results to neighbour SSD

SSD-PCIE Ring Reduction - Read Data



- GPU read data from the SSD written by the neighbour GPU
- Repeat for n times to complete Ring Reduction

```
warmup - generate
benchmark - generate
100%|
7it/s]#layers: 50
#batches prefill:  50
#batches decoding: 1550
load_weight            (per-layer): 0.017412 s
load_cache_prefill     (per-batch): 0.000009 s
store_cache_prefill    (per-batch): 0.000066 s
compute_layer_prefill  (per-batch): 0.000684 s
load_cache_decoding    (per-batch): 0.000045 s
store_cache_decoding   (per-batch): 0.000052 s
compute_layer_decoding (per-batch): 0.003682 s
100%|
8it/s]
/home/u9/work/FlexLLMGen/lib/python3.12/site-packages/torch/__init__.py:1136: FutureWarning: `torch.distribute
 use `torch.distributed.ReduceOp` instead
  return isinstance(obj, torch.Tensor)
Outputs:
----------------------------------------------------------------------
0: Paris is the capital city of France
----------------------------------------------------------------------
0: Paris is the capital city of France
----------------------------------------------------------------------

TorchDevice: cuda:0
  cur_mem: 0.0089 GB,  peak_mem: 1.8652 GB
TorchDevice: cpu
  cur_mem: 0.0000 GB,  peak_mem: 0.0000 GB
model size: 2.443 GB    cache size: 0.100 GB    hidden size (p): 0.002 GB
peak gpu mem: 1.865 GB  projected: True
prefill latency: 1.314 s        prefill throughput: 389.756 token/s
decode latency: 24.159 s        decode throughput: 1.283 token/s
total latency: 25.473 s total throughput: 1.256 token/s
(FlexLLMGen) u9@u9-System-Product-Name:~/work/FlexLLMGen$
```