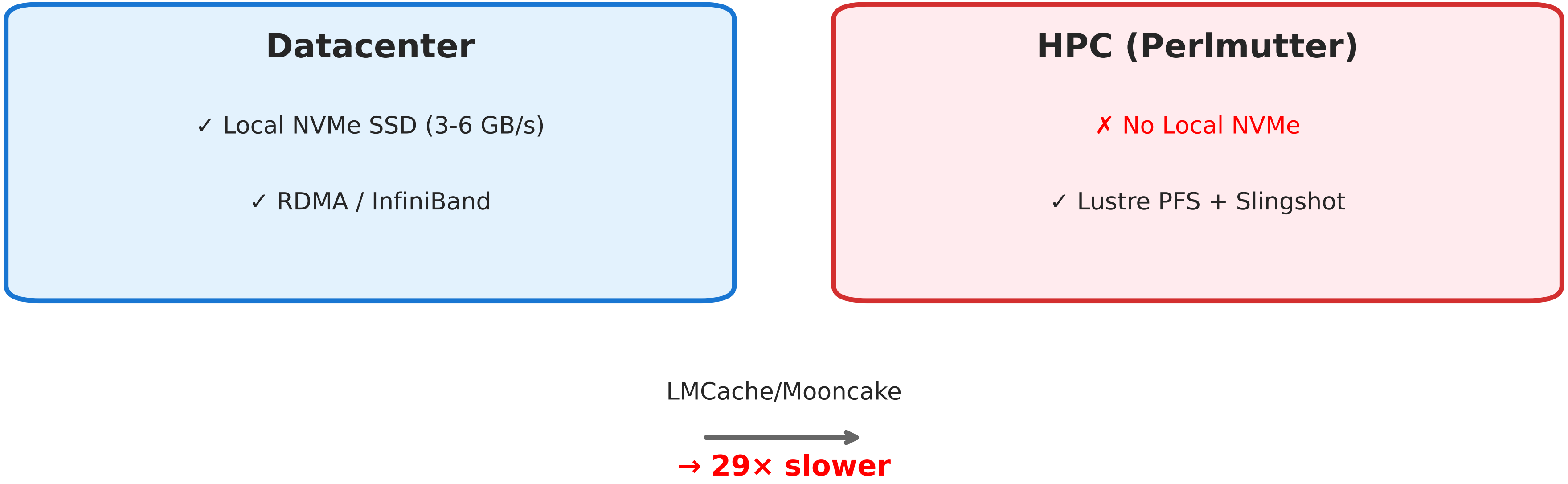
Why Datacenter KV Cache Systems Fail on HPC



Datacenter

✓ Local NVMe SSD (3-6 GB/s)

✓ RDMA / InfiniBand

HPC (Perlmutter)

✗ No Local NVMe

✓ Lustre PFS + Slingshot

LMCache/Mooncake

→ 29× slower