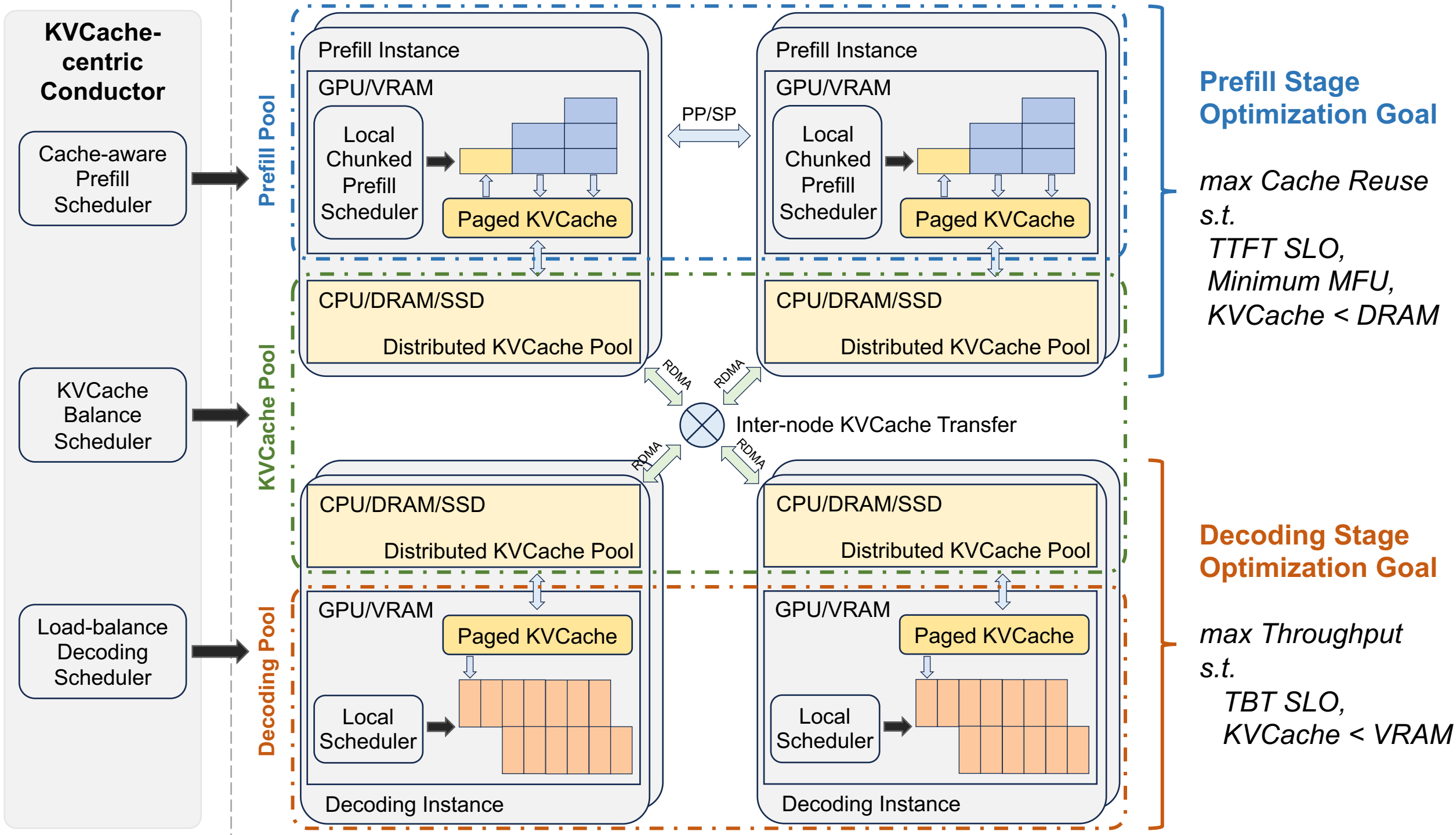
KVCache-centric Conductor
Cache-aware Prefill Scheduler
KVCache Balance Scheduler
Load-balance Decoding Scheduler
Prefill Pool
KVCache Pool
Decoding Pool
Prefill Instance
GPU/VRAM
Local Chunked Prefill Scheduler
Paged KVCache
PP/SP
CPU/DRAM/SSD
Distributed KVCache Pool
RDMA
Inter-node KVCache Transfer
Local Scheduler
Decoding Instance
Prefill Stage Optimization Goal
max Cache Reuse
s.t.
TTFT SLO,
Minimum MFU,
KVCache < DRAM
Decoding Stage Optimization Goal
max Throughput
s.t.
TBT SLO,
KVCache < VRAM