Disaggregated Prefill & Decode Architecture

Decoupling Compute and Memory Walls in High-Throughput LLM Serving Clusters

Author: Lesty Tech Systems Architecture Team | **Date:** November 9, 2026 | **Status:** Validated Engineering Reference | **Target:** Tier-1 Foundry, Silicon & Systems Architecture Audits

EXECUTIVE ARCHITECTURAL ABSTRACT:

Standard production large language model (LLM) serving systems co-locate the prompt processing (prefill) phase and the autoregressive token generation (decode) phase across the same physical GPU pool. This monolithic design induces fundamental hardware inefficiency. Prefill operations are compute-bound, exhibiting high arithmetic intensity (40-128 FLOP/byte) that saturates tensor cores. Decode iterations are strictly memory-bandwidth-bound (< 0.85 FLOP/byte), starving tensor units while waiting on HBM transfers. Co-locating both phases causes sudden prefill bursts to preempt ongoing decodes, blowing up tail latency (p99 ITL) and causing severe head-of-line blocking. This whitepaper evaluates Disaggregated Prefill and Decode (Splitwise / DistServe / Mooncake) architecture, detailing the network transport mechanics, RDMA KV-cache migration fabrics, and cluster-level scheduling algorithms that decouple the compute and memory walls in hyperscale AI inference clusters.

KEY PHYSICAL PARAMETERS & EXPERIMENTAL BENCHMARK TELEMETRY

Operational Metric	Monolithic Colocated	Chunked-Prefill vLLM	Splitwise Disaggregated	DistServe Fabric
Arithmetic Intensity	Mixed / Contended	Interleaved Slices	Dedicated High (Prefill)	Dedicated High (Prefill)
p99 TTFT Tail Latency	1,850 ms (Contended)	820 ms (Chunk Sliced)	285 ms (Isolated GPU)	240 ms (Isolated GPU)
p99 ITL Jitter	85 - 140 ms Spikes	35 - 55 ms Spikes	< 14 ms (Zero Preemption)	< 12 ms (Zero Preemption)
GPU Tensor Core Saturation	38.2% (Interrupted)	52.4% (Interleaved)	88.6% (Prefill Cluster)	91.2% (Prefill Cluster)
KV Migration Overhead	0 ms (Local HBM)	0 ms (Local HBM)	2.8 ms (RoCEv2 800G)	2.2 ms (InfiniBand NDR)

1. The Compute vs. Memory Roofline Duality in Transformer Inference

Transformer model inference consists of two phases with completely divergent computational characteristics. During prefill, the model processes an input prompt of N tokens in parallel. The self-attention and feed-forward layers execute dense General Matrix Multiplications (GEMM), achieving arithmetic intensities exceeding 50 FLOPs per byte, fully saturating modern systolic tensor arrays. During decode, the model processes exactly one new token per sequence. Matrix-vector operations (GEMV) dominate; the GPU must stream gigabytes of model weights and historical Key-Value cache tensors through its memory bus to compute a handful of arithmetic operations, resulting in an arithmetic intensity of under 1 FLOP per byte and leaving 80% of silicon compute transistors idle.

2. Head-of-Line Blocking & Tail Latency Collapse in Monolithic Serving

When a serving engine mixes prefill and decode tasks in the same batch (even with continuous batching), high-priority prefills preempt ongoing decode iterations. A 4,000-token prompt prefill can monopolize GPU execution units for several hundred milliseconds. During this window, all concurrently running streaming decodes are paused, causing severe Inter-Token Latency (ITL) jitter and degrading real-time user experiences. Furthermore, memory allocation becomes severely fragmented: prefill demands vast intermediate activation buffers, while decode demands deep, persistent KV cache storage, causing premature Out-Of-Memory (OOM) evictions.

TECHNICAL ANALYSIS CONTINUED: DISAGGREGATED PREFILL & DECODE ARCHITECTURE

3. Disaggregated Architecture: Dedicated Compute & Memory Pools

Disaggregated serving physically separates the inference cluster into two specialized tiers: Prefill Instances (P-nodes) and Decode Instances (D-nodes). P-nodes are provisioned with high-compute configurations (optimized for dense FP8/FP16 tensor throughput) and execute prompt GEMMs at maximum batch size. Once the prefill completes and the initial output token is generated, the P-node transfers the generated Key-Value cache tensors across an ultra-low-latency network fabric directly into the memory of a designated D-node. D-nodes operate continuously on high-batch decodes with 100% predictable step latencies, completely insulated from prompt arrival bursts.

4. GPUDirect RDMA Network Fabrics & KV Cache Streaming

The viability of disaggregated serving depends on the speed of KV cache migration. Transmitting a 32-layer, 8,000-token KV cache for a 70-billion-parameter model represents approximately 2 gigabytes of tensor data. Utilizing standard TCP/IP networking would introduce dozens of milliseconds of transfer delay, erasing the latency advantages of disaggregation. Production disaggregated clusters utilize 400G/800G RoCEv2 or InfiniBand fabrics with GPUDirect RDMA. Tensors are streamed directly from prefill HBM to decode HBM in under 3 milliseconds, fully pipelined with first-token network transmission.

STRATEGIC ENGINEERING TAKEAWAYS & DEPLOYMENT CHECKLIST:

- Disaggregated serving eliminates the resource conflict between compute-bound prompt prefill and memory-bound token decode.
- Physical cluster separation slashes p99 Inter-Token Latency (ITL) jitter by over 80%, providing consistent sub-15ms streaming.
- High-speed 800G GPUDirect RDMA fabrics enable multi-gigabyte KV cache migrations between nodes with negligible sub-3ms overhead.
- Overall cluster GPU efficiency increases by 40-60%, allowing operators to independently scale prefill and decode hardware resources.