

AnchorTP: Resilient LLM Inference with State-Preserving Elastic Tensor Parallelism

Wendong Xu^{1,*}, Chujie Chen², He Xiao¹, Kuan Li³, Jing Xiong¹, Chen Zhang¹,
Wenyong Zhou¹, Chaofan Tao¹, Yang Bai⁴, Bei Yu⁴, Ngai Wong¹

¹Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong

²Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

³Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong

⁴Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong

*Corresponding author. Email: wdxu@connect.hku.hk

Abstract—Large Language Model (LLM) inference services demand exceptionally high availability and low latency, yet multi-GPU Tensor Parallelism (TP) makes them vulnerable to single-GPU failures. We present AnchorTP, a state-preserving elastic TP framework for fast recovery. It (i) enables Elastic Tensor Parallelism (ETP) with unequal-width partitioning over any number of GPUs and compatibility with Mixture-of-Experts (MoE), and (ii) preserves model parameters and KV caches in GPU memory via a daemon decoupled from the inference process. To minimize downtime, we propose a bandwidth-aware planner based on a Continuous Minimal Migration (CMM) algorithm that minimizes reload bytes under a byte-cost dominance assumption, and an execution scheduler that pipelines P2P transfers with reloads. These components jointly restore service quickly with minimal data movement and without changing service interfaces. In typical failure scenarios, AnchorTP reduces Time to First Success (TFS) by up to 11× and Time to Peak (TTP) by up to 59% versus restart-and-reload.

I. INTRODUCTION

Online large language models (LLMs) serving systems have become critical infrastructure [1], [2]. To meet compute and memory capacity demands, inference commonly adopts tensor parallelism (TP) across multiple GPUs [3], [4], [5]. The tight coupling in TP makes systems highly sensitive to single-GPU failures or link degradation [6], [7]: once a communication group breaks, service is interrupted and the time to first success (TFS) can reach tens of minutes [8], [9]. This work aims to reduce TFS (time from fault to first successful token; often called TTFT) to seconds while preserving throughput and latency, and to shorten TTP (time from first success to stabilized peak throughput).

However, existing TP implementations hardwire divisibility constraints on key dimensions, fix per-GPU tensor shapes, and assume a static scale of collective communications. When topology changes, shapes and collectives globally mismatch, leading to high reconfiguration costs.

To achieve fast recovery, the industry typically employs two main approaches: restart-and-reload, which simplifies the process but yields long TFS/TTP [10], [11], [12]; and static redundancy, which relies on standby nodes or replicas, sacrificing flexibility and cost efficiency. Practice reveals three challenges [13], [14], [15]: (i) topology-aware migration costs under bandwidth

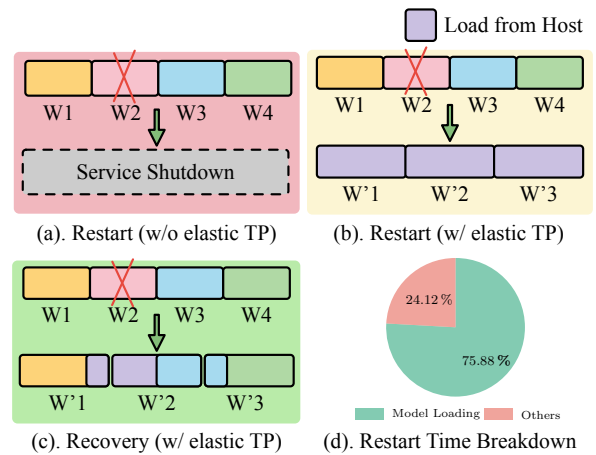


Fig. 1. Recovery strategies when one GPU fails in a four-GPU deployment. (a) Without elastic TP, service cannot resume. (b) With elastic TP but no state preservation, service restarts with three GPUs but fully reloads parameters from host. (c) With state-preserving elastic TP, parameters/KVs on surviving GPUs are reused via planned P2P transfers with minimal reload. (d) Time breakdown for a typical restart-and-reload on Qwen3-14B [16]; host-to-GPU reload dominates.

tiering and limited reachability, (ii) compute–communication rebalancing after TP-scale changes, and (iii) runtime state persistence where KVs, memory pools, and communication groups are tightly bound to the initial topology. These constraints shrink the feasible design space and motivate maximizing in-place reuse while minimizing necessary migrations (see Figure 1).

As illustrated in Figure 1(d), restart-and-reload is bottlenecked by host-to-GPU model reload time. We observe that significant memory redundancy in production deployments creates an opportunity for low-cost state-preserving recovery.

The widespread adoption of Mixture-of-Experts (MoE) architectures [17], [18], [19] has made sparse activation and expert routing the norm, further tightening the coupling between state and communication and raising the requirements for KV-state consistency and topology reconfiguration. Expert-parallel load balancing (EPLB) [20], [21] primarily focuses on optimizing routing popularity and expert utilization, which is essentially dynamic load balancing. EPLB strategies rely on the recency

and temporal coherence of request traffic; once a service restart and reload occurs, this coherence is disrupted, which can bias estimates of traffic distribution and hot experts [22]. As a result, larger scheduling and routing adjustments may be triggered, which in the short term exacerbate expert load imbalance and degrade performance.

To address these challenges, we propose AnchorTP, a disaster-resilient elastic inference framework. AnchorTP decouples state from control: the state plane persistently holds GPU memory ownership of model parameters and KVs, while the control plane drives elastic TP reconfiguration and minimal-migration recovery after failures. Without changing the service interface, we convert costly reloads into a small number of bandwidth-aware migrations (Figure 1b,c). We focus on multi-GPU inference under device-offline and link-degradation scenarios, targeting low TFS and reduced TTP by maximizing in-place reuse and minimizing necessary migrations. While our evaluation concentrates on single-node deployments, the core principles extend naturally to multi-node environments where the primary difference lies in communication overhead rather than algorithmic complexity.

We summarize our contributions as follows:

- We decouple a state plane and a control plane. The state plane anchors GPU memory for parameters and KVs via a daemon and IPC handles, enabling in-place fast recovery.
- We implement Elastic Tensor Parallelism (ETP) with unequal-width sharding for attention and linear layers. ETP breaks divisibility constraints, supports arbitrary TP-scale reconfiguration, and remains compatible with MoE.
- We design a Continuous Minimal Migration (CMM) planner with bandwidth-aware execution. It minimizes reload bytes by interval mapping under the cost model and overlaps P2P with reloads to reduce wall-clock time.

II. PRELIMINARIES

A. Parallel Topologies

Distributed inference for LLMs typically employs a hybrid parallel strategy combining Tensor Parallelism (TP) and Expert Parallelism (EP) in MoE models [23], [24]. TP addresses single-card memory bottlenecks by partitioning linear layer computations across multiple GPUs, while EP distributes different experts across devices to achieve sparse activation of computation. Together, these two forms of parallelism constitute a tightly-coupled parallel topology. Within this topology, all participating GPUs must operate collaboratively at a predefined, fixed scale.

B. Communication and Migration Costs

Fault recovery strategies generally fall into two categories: restart-and-reload, which incurs a minute-scale RTO due to I/O, and in-place recovery, which converts the RTO into an inter-device communication problem.

After a failure, the core of system recovery is the reorganization of surviving GPU resources [8]. This process inevitably involves data migration, including model parameters and the KV cache. The efficiency of this migration directly

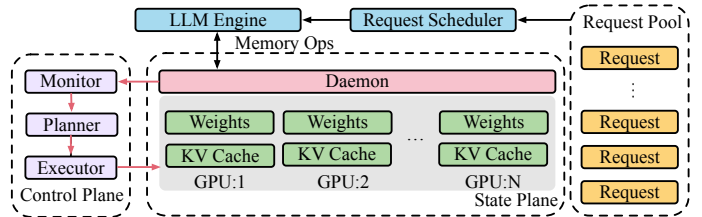


Fig. 2. AnchorTP overview with two planes. The state plane runs daemons that pin GPU memory for model parameters and the KV cache. The control plane monitors failures, plans recovery via our Continuous Minimal Migration (CMM) algorithm, and the executor coordinates data migration and system reinitialization.

determines recovery time, with communication bandwidth as the main bottleneck. In a typical multi-GPU server, inter-device bandwidth varies significantly: Peer-to-Peer (P2P) direct access over high-speed interconnects (e.g., IF/XGMI) is much faster than host-mediated PCIe paths relayed through CPU memory [25]. Therefore, an efficient plan must be topology-aware and maximize high-bandwidth paths to minimize recovery costs.

C. Recovery objectives and practical metrics

For online fault tolerance, we ground recovery objectives in measurable quantities. We use Time to First Success (TFS) as the operational form of the recovery objective: the time from fault injection to the first successful response. We report Time to Peak (TTP) as the time from TFS until throughput stabilizes at its peak. Consistency and freshness are reflected by state-retention coverage and reload fraction rather than a standalone RPO metric. In particular, we report (i) data reuse ratio and reload fraction for parameters and KVs, and (ii) the P2P/Host/Reload migration breakdown and overlap ratio during execution.

III. METHODOLOGY

We propose AnchorTP, an architecture for state-preserving recovery and elastic tensor parallelism shown in Figure 2. Its core idea is to decouple long-lived state management from dynamic topology orchestration with two planes: a state plane that pins parameters/KV, and a control plane that plans with CMM and executes recovery. The architecture is designed to work seamlessly across both single-node and multi-node deployments, with the fundamental recovery mechanisms remaining identical regardless of the physical topology.

Under normal operation, daemons in the state plane hold references to the model weights and the key-value cache (KV cache). If a GPU fails, the inference process may crash; however, because the daemons persist, the GPU driver does not reclaim this critical state. The control plane then activates: it detects the failure through heartbeat checks and performance monitoring, assesses surviving GPU resources, and leverages elastic tensor parallelism to derive a viable new topology. Using our continuous minimal migration algorithm, it computes an optimal data-migration plan that maximizes data reuse while minimizing host reloads. The control plane then coordinates the actual migration, prioritizing P2P transfers and pre-allocating

destination buffers to minimize recovery time. Finally, it launches a new inference system instance to resume service.

A. State Management

Traditional restart-and-reload is the most straightforward response to failures, but it yields a long TFS. We decompose it into multiple stages: init, load, and warmup. Among these, loading model state from disk into GPU memory is the primary bottleneck. For a 30B model with MoE in half precision roughly needs 60 GB of weights and typically takes tens of seconds even on fast hardware. Eliminating disk reload is therefore key to achieving second-level TFS.

To address this, we decouple a long-lived daemon from the inference system. Our design leverages the common observation that production LLM services often have significant memory redundancy, as VRAM is provisioned for worst-case KV cache usage. In our prototype, we provide a state plane and a device-memory pool interface that hold (or register) critical GPU memory regions. The daemon’s role is simple and stable: at service startup it allocates or registers the required GPU memory and remains the owner of these allocations. This pinning of state has minimal overhead as it occupies otherwise under-utilized VRAM. The daemon itself does not participate in complex computation or communication, maximizing its stability.

The inference system acts as the user of these memories. Via inter-process communication (IPC), it acquires handles from the daemon and performs reads and writes.

Furthermore, the state plane maintains backup timestamps and recent usage information for each session’s KV cache. During the recovery phase, if the calculated KV demand exceeds the memory budget, the system will gradually release cold KVs using a least recently used (LRU) eviction policy to prioritize the availability of hot sessions and new requests. When necessary, only the minimum set of hotspots that can cover steady-state throughput will be retained, and on-demand replay will be triggered for evicted sessions to rebuild the KV.

As a result, model parameters and KV-cache state on surviving GPUs are preserved in place. When the control-plane orchestrator launches new inference processes, they can immediately obtain handles to the surviving memory from the daemon and execute the recovery plan.

B. Elastic Tensor Parallelism

Elastic tensor parallelism is a core capability that enables AnchorTP to achieve fault recovery. Traditional tensor parallelism assumes the tensor dimension is divisible by the parallel degree. Our framework removes this constraint by allowing partitioning across any number of GPUs, even if shards have unequal sizes.

Concretely, during partitioning we allow some GPUs to hold shard size $\lfloor S/g_t \rfloor$ while others hold $\lceil S/g_t \rceil$, where S is the tensor dimension and g_t is the new TP size. This flexibility underpins elastic recovery: regardless of how many GPUs survive after a failure, the system can always find a valid sharding plan without being constrained by divisibility.

To make the above partitioning reusable during recovery, we establish several layout invariants on the state plane: parameters

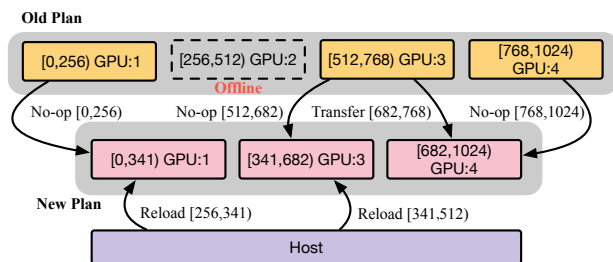


Fig. 3. Example (4→3 GPUs). 1024 rows (modeled as a 1D byte interval) are split across 4 GPUs. After GPU:2 fails, the target plan is [0, 341], [341, 682], [682, 1024]. GPU:1 keeps [0, 256] and reloads [256, 341]; GPU:3 reloads [341, 512] and keeps [512, 682]; GPU:4 receives [682, 768] via P2P from GPU:3 and keeps [768, 1024]. Only 256 rows are reloaded; the rest use P2P.

reside as contiguous blocks in a daemon-managed device-memory pool and are exposed via versioned handles that provide a unified address view; the KV cache is localized by attention head-groups per token block (e.g., paged KV blocks) and kept consistent with the parallel mapping. On top of these invariants, we provide two minimal interface primitives: a parameter re-assignment/reloading primitive, and KV-cache operators that support variable-length shards. Both primitives are decoupled from communication and memory management, enabling in-place remapping within a freeze window so that mapping switches complete with minimal migration.

C. Topology-aware recovery planning

Building on state preservation, we design a pragmatic two-stage recovery procedure. The first stage, logical migration planning, is topology-agnostic. It operates on a unified one-dimensional byte layout and emits a minimal-reload plan by maximally reusing surviving data intervals. The second stage, physical execution scheduling, is topology-aware. It takes the logical plan and orchestrates data movement to minimize wall-clock recovery time, respecting hardware constraints like peer reachability and bandwidth hierarchy (e.g., prioritizing fast XGMI/NVLink over slower PCIe paths). This decoupled design keeps the planning algorithm simple and provably optimal in terms of reload bytes, while the scheduler handles the complexities of hardware performance.

As illustrated in Figure 3, the CMM algorithm first constructs the old layout view in a unified one-dimensional byte space. It identifies the total size $H = \max(e_i)$ and evenly partitions the interval $[0, H)$ by the target GPU count M to derive the target layout. Then, for each target interval, it computes intersections with surviving source intervals. Non-empty intersections are mapped to P2P Transfer tasks. Any remaining gaps in the target interval that are not covered by surviving data are mapped to Reload tasks from host memory. This greedy, interval-based procedure is simple, deterministic, and guarantees a minimal total reload volume, as formally argued in Equation (1). The overall procedure is summarized in Algorithm 1.

We give a brief argument that the planner achieves minimal reload. Let the global space be $[0, H)$, the surviving source layout be disjoint intervals $\{[s_i, e_i]\}_{i \in S}$, and the target layout be disjoint intervals $\{[u_j, v_j]\}_{j=1}^M$ with $\bigcup_j [u_j, v_j] = [0, H)$. For each target $[u_j, v_j)$, the maximum content reusable without

Algorithm 1 Continuous Minimal Migration Algorithm

Require:

Current row plan: $\{(gpu_i, s_i, e_i, alive_i)\}_{i=1}^N$ from daemon,
target GPU count M

Ensure: Migration plan: $\{(gpu_j, sources_j)\}_{j=1}^M$

- 1: $H \leftarrow \max(e_i)$ for all i
 - 2: **for** $j = 1$ to M **do**
 - 3: $s_j \leftarrow \lfloor (j-1) \times H/M \rfloor$
 - 4: $e_j \leftarrow \lfloor j \times H/M \rfloor$
 - 5: **end for**
 - 6: **for** $j = 1$ to M **do**
 - 7: $tar_range \leftarrow [s_j, e_j]$
 - 8: **for** $i = 1$ to N **do**
 - 9: **if** $alive_i = \text{true}$ **then**
 - 10: $inter \leftarrow tar_range \cap [s_i, e_i]$
 - 11: **if** $inter \neq \emptyset$ **then**
 - 12: $AddTransferPlan(inter, gpu_i, gpu_j)$
 - 13: $tar_range \leftarrow tar_range \setminus inter$
 - 14: **end if**
 - 15: **end if**
 - 16: **end for**
 - 17: **if** $tar_range \neq \emptyset$ **then**
 - 18: $AddReloadPlan(tar_range, gpu_j)$
 - 19: **end if**
 - 20: **end for**
 - 21: **return** migration plan
-

reload equals the total measure of $\bigcup_{i \in S} ([s_i, e_i] \cap [u_j, v_j])$. Hence the minimum reload over all plans is

$$\begin{aligned}
 \text{Reload}^* &= \sum_{j=1}^M \left(|[u_j, v_j]| - \sum_{i \in S} |[s_i, e_i] \cap [u_j, v_j]| \right) \\
 &= H - \sum_{i \in S} \sum_{j=1}^M |[s_i, e_i] \cap [u_j, v_j]|.
 \end{aligned} \tag{1}$$

The CMM planner enumerates these intersections and assigns each overlap directly, thereby attaining the upper bound on reuse for every target GPU. Its reload volume thus equals the theoretical minimum derived in Equation (1). This optimality holds under the assumption that reload cost per byte is strictly higher than any P2P transfer cost.

Next, the execution scheduler optimizes the wall-clock time for the migration plan generated by CMM.

Before launching any data movement, the scheduler first scans the entire migration plan and pre-allocates the required destination buffers on all target GPUs in a single pass. This shifts allocator overhead off the critical recovery path and avoids latency spikes caused by on-the-fly allocations during migration.

The scheduler then pipelines migration tasks to maximize hardware utilization, informed by the underlying topology. It maintains a cost model for different communication paths and prioritizes scheduling transfers over the highest-bandwidth available links first. Crucially, it overlaps high-latency reloads (host-to-GPU transfers via PCIe) with lower-latency P2P transfers (GPU-to-GPU copies, often via XGMI), effectively hiding the

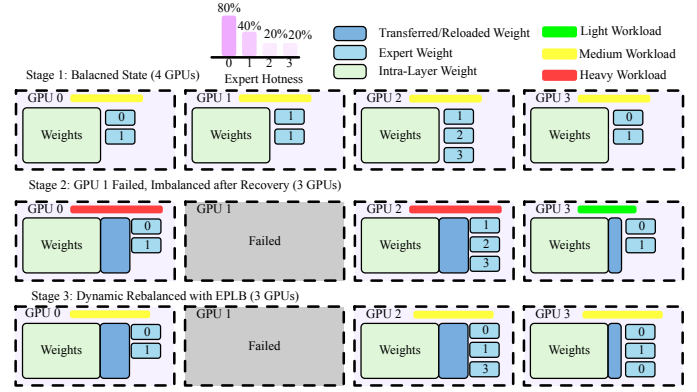


Fig. 4. Example of EPLB rebalancing after a failure. (a) Initially, 4 GPUs are perfectly balanced. (b) After GPU 1 fails, AnchorTP re-shards the parameters, leaving GPU 3 with a smaller shard and thus more free compute resources. (c) EPLB, aware of this, intelligently places the recovered Expert 1 and new replicas of the hotspot Expert 0 onto the most idle GPU (GPU 3), achieving a new, performance-optimal state that is not arithmetically balanced but maximizes system throughput.

P2P communication latency under the longer reload time and thus shortening the overall recovery duration. In addition, the runtime KV cache and parameters share the same planning and task generation process. On KV cache misses, we rebuild via on-demand recomputation (token replay) without blocking the recovery pipeline.

Beyond failure recovery, this two-stage planning mechanism also underpins dynamic rescheduling during normal operation. When the system detects hardware performance drift, resource bottlenecks, or load hotspots, it proactively triggers rescheduling and reuses the same planning and optimization pipeline. To address potential MoE expert imbalance after elastic adjustments, the system applies an EPLB-like [20] dynamic load balancing strategy. This strategy restores balance by replicating hot-spot experts and adjusting routing to optimally utilize available resources. For instance, after an $8 \rightarrow 7$ GPU recovery, the new parameter sharding might leave one GPU with a smaller shard and thus more idle compute capacity. As illustrated in Figure 4, an intelligent EPLB planner detects this heterogeneity and schedules more expert replicas to the under-utilized GPU. This results in a new state that, while not arithmetically balanced in terms of request count, is performance-optimal as it maximizes the throughput of the entire system.

IV. IMPLEMENTATION

We implement a lightweight inference framework based on nano-llm [26] and realize the aforementioned fault-tolerance design in practice. On the state plane, a daemon manages a device-memory pool and IPC handles to anchor model parameters and the KV cache. The migration planner operates over a unified one-dimensional byte space and generates Transfer/Reload tasks on interval boundaries. The orchestrator distributes versioned plans, while the executor pre-allocates destination buffers and overlaps Reloads with P2P transfers. ETP in AnchorTP enables unequal-width sharding for attention and linear layers, making arbitrary TP degrees feasible and allowing reuse of surviving shards. The implementation follows minimal-interface and decoupling principles: communication, memory, and operator changes are

independent; stale shards are promptly released after remapping takes effect to curb fragmentation, which eases integration into existing services. For model reload, our loader reads safetensors by offsets: it parses per-tensor offset/length from the header and issues range reads to fetch only required byte intervals.

V. EVALUATION

This chapter systematically evaluates AnchorTP’s fault-recovery capability and elastic scaling efficiency on a single-node, multi-GPU setup, covering recovery latency, resource utilization, and end-to-end performance.

We evaluate recovery in 4→2 and 8→6 TP-degradation scenarios, and additionally test 8→7 for ablations.

A. Experimental Setup

The evaluation is conducted on a single-node, multi-GPU ROCm 6.3 platform to first establish the foundational performance of our recovery primitives in a controlled environment with high-speed interconnects. The setup consists of 8× AMD Instinct MI210 with 64GB memory and dual NUMA (GPUs 0–3 on node0 and GPUs 4–7 on node1). Intra-group interconnect is via Infinity Fabric (IF/XGMI), inter-group connectivity is over PCIe, and RCCL is used as the communication library. The software stack is Python 3.12 and PyTorch 2.8.0.

We use the following metrics: TFS, defined as the time from fault injection to the first successful response; TTP, the time from TFS until throughput stabilizes at its peak; and the total run time of the benchmark. During the recovery window (from failure detection to TFS), incoming requests are queued by the serving frontend and processed once the service is restored, ensuring no requests are dropped.

Models and Workload. We use Qwen3-30B-A3B [16], Mixtral-8×22B (141B) [27], and Qwen3-8B/14B [16]. The workload is a fixed set of 1,000 ShareGPT requests [28] replayed with a fixed arrival pattern.

Baselines. Static Parallelism (SP) keeps the original TP size without online rescheduling; upon failure, service halts (we report its no-failure runtime as a baseline reference). Elastic TP (restart-only) restarts with arbitrary TP sizes (supports non-divisible TP) but performs no state-preserving recovery (full reload at restart).

B. End-to-End Performance

We evaluate end-to-end performance under injected failures at 25% and 50% of the request stream.

For Qwen3-30B-A3B, we simulate one GPU failure at 25% and another at 50% of the request stream. For Qwen3-30B-A3B, all GPUs are in the same IF group. For Mixtral-8×22B (8→6), failures are induced symmetrically with one GPU terminated in each NUMA node’s IF group to ensure balanced degradation.

SP cannot recover from failure; ‘/’ indicates TFS/TTP are undefined. Even when KV-cache misses occur and trigger on-demand replay, AnchorTP degrades gracefully and continues serving, which is strictly preferable to service interruption in SP. TFS of AnchorTP is about 10.8× (25%) and 9.9× (50%) faster for Qwen3-30B-A3B, and about 10.5× (25%) and 10.2× (50%) for Mixtral-8×22B, compared to the Elastic

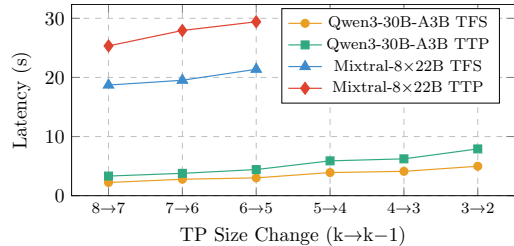


Fig. 5. Per-switch TFS and TTP as TP decreases ($k \rightarrow k-1$) for Qwen3-30B-A3B and Mixtral-8×22B. Lower is better.

TP baseline, as it significantly reduces the costly full model reload from host memory. This efficiency translates directly to a significantly lower total runtime overhead, which is reduced by factors of 5.6× and 4.7× for Qwen3-30B-A3B and Mixtral-8×22B, respectively. As our analysis, the TFS for the restart-only baseline is dictated by the total model size, with Mixtral-8×22B taking roughly four times longer to load than Qwen3-30B-A3B, consistent with its larger parameter count. In contrast, AnchorTP’s TFS is determined by minimal data migration, resulting in consistently low recovery times.

C. Ablation Study

To independently validate the contributions of AnchorTP’s core components, a series of ablation studies were conducted to isolate the impact of each component. First, the effectiveness of daemon-based state management is assessed by comparing the full system against the “Elastic TP (restart-only)” baseline. As shown in Table I, removing state preservation forces a full reload, significantly increasing TFS and TTP and underscoring the daemon’s pivotal role. Second, the “Static Parallelism (SP)” baseline in Table I illustrates the consequence of lacking elastic reconfiguration: upon failure, the service cannot adapt and is irrecoverably interrupted, establishing elastic TP as a foundational prerequisite for fault tolerance.

The superiority of our Continuous Minimal Migration algorithm is evaluated against alternative strategies. The results in Table II demonstrate its efficiency in minimizing reload and P2P transfer times. The evaluated methods include CM (ours), a Greedy algorithm with local optimizations, and a Full Reload approach without data reuse. For Greedy in our experiments, missing bytes are reloaded from host memory on the target GPU and we do not schedule cross-device transfers for those gaps. In numbers, CM achieves the lowest reload time 17.6s while incurring only small P2P (1.9s); Greedy increases reload time to 26.2s because gaps are not filled via P2P; Full Reload is dominated by host reload (197s) and has no P2P component, corroborating the advantage of reuse-aware planning.

To assess AnchorTP’s performance stability under continuous degradation and its adaptability to different model scales, we progressively decrease the TP degree for two models: Qwen3-30B-A3B (from TP=8 to 2) and Mixtral-8×22B (from TP=8 to 6). For each change (TP=k → k-1), we report per-switch TFS and TTP, as shown in Figure 5.

As depicted in Figure 5, both TFS and TTP exhibit a consistent upward trend as the Tensor Parallelism (TP) degree is progressively reduced. This trend is attributed to the increased

TABLE I

END-TO-END RECOVERY PERFORMANCE COMPARISON. WE REPORT TFS AND TTP (IN SECONDS) AFTER FAILURES INJECTED AT 25% AND 50% OF THE REQUEST STREAM. ‘OVERHEAD’ IS THE TOTAL RUNTIME INCREASE OVER THE NO-FAILURE STATIC PARALLELISM (SP) BASELINE; LOWER IS BETTER.

| Model | TP Δ | Method | TFS@25 (s) | TFS@50 (s) | TTP@25 (s) | TTP@50 (s) | Total (s) | Overhead (s) \downarrow |
|------------------------|-------------------|---------------------------|-------------------|-------------------|------------------|------------------|--------------------|---------------------------|
| Qwen3-30B-A3B | 4 \rightarrow 2 | Static Parallelism (SP) | / | / | / | / | 107.84 \pm 5.68 | 0 (Baseline) |
| | | AnchorTP (ours) | 4.48 \pm 0.35 | 4.98 \pm 0.43 | 6.23 \pm 0.82 | 7.91 \pm 1.07 | 139.74 \pm 6.76 | +31.90 |
| | | Elastic TP (restart-only) | 48.43 \pm 2.95 | 49.21 \pm 3.17 | 14.15 \pm 1.74 | 19.33 \pm 2.11 | 285.82 \pm 9.81 | +177.98 |
| Mixtral-8 \times 22B | 8 \rightarrow 6 | Static Parallelism (SP) | / | / | / | / | 273.65 \pm 8.42 | 0 (Baseline) |
| | | AnchorTP (ours) | 18.71 \pm 1.44 | 19.52 \pm 1.12 | 25.33 \pm 2.34 | 28.94 \pm 2.88 | 345.12 \pm 9.67 | +71.47 |
| | | Elastic TP (restart-only) | 195.82 \pm 5.07 | 198.49 \pm 6.53 | 35.69 \pm 3.53 | 41.27 \pm 3.71 | 610.79 \pm 17.61 | +337.14 |

TABLE II

PLANNER COMPARISON ON RELOAD AND P2P TIME (MIXTRAL-8 \times 22B, 8 \rightarrow 7 GPU). LOWER IS BETTER.

| Method | Reload Time (s) \downarrow | P2P Time (s) \downarrow |
|-------------|------------------------------|---------------------------|
| CM (ours) | 17.56 \pm 1.39 | 1.87 \pm 0.33 |
| Greedy | 26.15 \pm 1.73 | / |
| Full Reload | 197.31 \pm 7.98 | / |

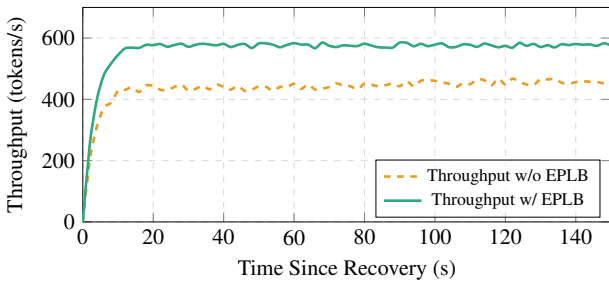


Fig. 6. Impact of EPLB on system throughput. With EPLB enabled, the system not only reaches a higher peak throughput but also stabilizes much faster, as indicated by the shorter TTP window. This demonstrates that rebalancing accelerates the convergence to a new, optimal steady-state.

complexity of rescheduling and the higher load placed on each surviving GPU. Notably, TFS shows a moderate, near-linear increase, demonstrating the efficiency of our minimal migration algorithm in handling state reconstruction across various scales. In contrast, TTP rises more steeply, underscoring the inherent performance challenge as fewer devices must handle the same workload, prolonging the time required to stabilize at a new peak throughput. The significantly higher absolute latencies for Mixtral-8 \times 22B compared to Qwen3-30B-A3B further highlight the pronounced impact of model scale on recovery overhead. Overall, the results validate AnchorTP’s robust scalability, maintaining efficient recovery even under significant degradation.

When serving MoE models like Mixtral-8 \times 22B under full load, elastic recovery (8 \rightarrow 7 GPUs) induces load imbalance from stale expert mappings. A lightweight EPLB strategy [20] restores balance: without EPLB the system stabilizes at 436.61 tokens/s, while with EPLB it reaches 562.32 tokens/s (+29%) and shortens TTP (Figure 6).

VI. RELATED WORK

a) LLM Distributed Systems: LLM training and inference rely on advanced parallelism including Data Parallelism (DP), Tensor Parallelism (TP), Pipeline Parallelism (PP), and Fully Sharded Data Parallel (FSDP). Inference stacks such as vLLM, TensorRT-LLM, and Megatron variants provide high-throughput kernels and KV management. DreamDDP[29] accelerates DP under low-bandwidth settings by partially localizing SGD and overlapping synchronization; AsymGroup[30] dynamically forms asymmetric 2D DP groups to manage heterogeneity.

b) Online Disaster Recovery and Elasticity for Serving: As LLM serving scales, failures become common and elasticity is required. Adaptive fault tolerance leverages prediction and dynamic resource allocation[31]. For training, Nonuniform Tensor Parallelism (NTP) mitigates single-GPU failures by reconfiguring TP within a DP replica via gradient resharding[32]. Resource managers like FaPES[33] borrow GPUs across inference and training to improve utilization. Unlike these, AnchorTP targets inference-time recovery under TP-scale changes by preserving state in-place and performing minimal migration.

c) Parameter/KV Hot Loading for Serving: Full-model reload dominates switch latency in large models. Parameter hot loading decouples model architecture from parameters to accelerate switching (e.g., FastPTM[34]). KV-centric serving (paged KV) further reduces memory churn in vLLM-like systems. AnchorTP complements these by pinning parameters and KVs in GPU memory via a daemon and by enabling unequal-width TP so that surviving shards are reused rather than reloaded.

VII. DISCUSSION

a) Assumptions and Limitations: Our design targets single-node, multi-GPU inference with peer reachability within the node. The reload-minimality of CMM assumes per-byte costs where reload is strictly more expensive than transfer/no-op and ignores link reachability during planning; execution scheduling then respects hardware topology. Extending CMM to multi-node settings requires incorporating link reachability and cost weighting (e.g., PCIe, NVLink/XGMI, and RDMA tiers) into planning.

b) Load Rebalancing and MoE Dynamics: Post-recovery, topology changes can perturb expert load. We integrate expert-parallel load balancing (EPLB) to re-stabilize routing weights

with lightweight regulation, prioritizing intra-node balance before inter-GPU adjustments. This complements CMM by restoring steady-state performance without invalidating preserved KVs.

VIII. CONCLUSION

This paper introduced AnchorTP, a disaster recovery framework designed to address the high-availability challenges in Tensor Parallelism-based LLM inference services. Traditional approaches suffer from service interruptions due to fixed device scales and long downtimes from state reloading. AnchorTP overcomes these limitations through two key innovations: Elastic Tensor Parallelism, which enables flexible service reconfiguration on surviving GPUs, and State-Preserving Recovery, which uses a decoupled daemon to preserve critical states in GPU memory, thus eliminating costly reloads. A topology-aware migration planner further accelerates recovery by minimizing data movement.

Experiments demonstrate that AnchorTP is highly effective, reducing recovery time by over $10\times$ and time-to-peak-performance by up to 59% compared to restart methods. This work significantly enhances the resilience of LLM inference services. Future work will extend CMM to multi-node topologies with link reachability and cost weighting, and integrate predictive fault handling for preemptive warm migration.

REFERENCES

- [1] L. Zheng, L. Yin, Z. Xie, C. L. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. E. Gonzalez *et al.*, “Sglang: Efficient execution of structured language model programs,” *Advances in neural information processing systems*, vol. 37, pp. 62 557–62 583, 2024.
- [2] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, “Efficient memory management for large language model serving with pagedattention,” in *Proceedings of the 29th symposium on operating systems principles*, 2023, pp. 611–626.
- [3] B. Wang, Q. Xu, Z. Bian, and Y. You, “Tesseract: Parallelize the tensor parallelism efficiently,” in *Proceedings of the 51st International Conference on Parallel Processing*, 2022, pp. 1–11.
- [4] F. Brakel, U. Odyurt, and A.-L. Varbanescu, “Model parallelism on distributed infrastructure: A literature review from theory to llm case-studies,” *arXiv preprint arXiv:2403.03699*, 2024.
- [5] B. Wu, S. Liu, Y. Zhong, P. Sun, X. Liu, and X. Jin, “Loongserve: Efficiently serving long-context large language models with elastic sequence parallelism,” in *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles*, 2024, pp. 640–654.
- [6] D. Arfeen, D. Mudigere, A. More, B. Gopireddy, A. Inci, and G. R. Ganger, “Nonuniform-tensor-parallelism: Mitigating gpu failure impact for scaled-up llm training,” *arXiv preprint arXiv:2504.06095*, 2025.
- [7] S. Cui, A. Patke, H. Nguyen, A. Ranjan, Z. Chen, P. Cao, B. Bode, G. Bauer, C. Di Martino, S. Jha *et al.*, “Characterizing gpu resilience and impact on ai/hpc systems,” *arXiv preprint arXiv:2503.11901*, 2025.
- [8] N. Blagojev, O. Ersoy, and L. Y. Chen, “All is not lost: Llm recovery without checkpoints,” *arXiv preprint arXiv:2506.15461*, 2025.
- [9] J. Xia, M. Zhang, J. Huang, Y. Liu, X. Hu, X. Liu, and C. Hu, “Mnemosyne: Lightweight and fast error recovery for llm training in a just-in-time manner,” in *Proceedings of the 9th Asia-Pacific Workshop on Networking*, 2025, pp. 157–163.
- [10] F. Strati, S. Mcallister, A. Phanishayee, J. Tarnawski, and A. Klimovic, “D\`ej\`avu: Kv-cache streaming for fast, fault-tolerant generative llm serving,” *arXiv preprint arXiv:2403.01876*, 2024.
- [11] B. Wu, L. Xia, Q. Li, K. Li, X. Chen, Y. Guo, T. Xiang, Y. Chen, and S. Li, “Transom: An efficient fault-tolerant system for training llms,” *arXiv preprint arXiv:2310.10046*, 2023.
- [12] T. Xie, J. Zhao, Z. Wan, Z. Zhang, Y. Wang, R. Wang, R. Huang, and M. Li, “Realm: Reliable and efficient large language model inference with statistical algorithm-based fault tolerance,” *arXiv preprint arXiv:2503.24053*, 2025.
- [13] Y. Cheng, “A scalable approach to distributed large language model inference,” *knowledge.uchicago.edu*, 2025.
- [14] M. Xu, J. Liao, J. Wu, Y. He, K. Ye, and C. Xu, “Cloud native system for llm inference serving,” *arXiv preprint arXiv:2507.18007*, 2025.
- [15] M. Zhang, X. Shen, J. Cao, Z. Cui, and S. Jiang, “Edgeshard: Efficient llm inference via collaborative edge computing,” *IEEE Internet of Things Journal*, 2024.
- [16] Q. Team, “Qwen3 technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2505.09388>
- [17] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.
- [18] S. Masoudnia and R. Ebrahimpour, “Mixture of experts: a literature survey,” *Artificial Intelligence Review*, vol. 42, no. 2, pp. 275–293, 2014.
- [19] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, “Deepseek-v3 technical report,” *arXiv preprint arXiv:2412.19437*, 2024.
- [20] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025.
- [21] D. Wisdom, “Load balancing and memory optimizations for expert parallel training of large language models,” Ph.D. dissertation, Massachusetts Institute of Technology, 2024.
- [22] Y. Zeng, C. Huang, Y. Mei, L. Zhang, T. Su, W. Ye, W. Shi, and S. Wang, “Efficientmoe: Optimizing mixture-of-experts model training with adaptive load balance,” *IEEE Transactions on Parallel and Distributed Systems*, 2025.
- [23] S. Singh, O. Ruwase, A. A. Awan, S. Rajbhandari, Y. He, and A. Bhatle, “A hybrid tensor-expert-data parallelism approach to optimize mixture-of-experts training,” in *Proceedings of the 37th International Conference on Supercomputing*, 2023, pp. 203–214.
- [24] R. Zhu, Z. Jiang, C. Jin, P. Wu, C. A. Stuardo, D. Wang, X. Zhang, H. Zhou, H. Wei, Y. Cheng *et al.*, “Megascall-infer: Efficient mixture-of-experts model serving with disaggregated expert parallelism,” in *Proceedings of the ACM SIGCOMM 2025 Conference*, 2025, pp. 592–608.
- [25] R. Nakamura, Y. Kuga, and K. Akashi, “How beneficial is peer-to-peer dma?” in *Proceedings of the 11th ACM SIGOPS Asia-Pacific Workshop on Systems*, 2020, pp. 25–32.
- [26] X. Yu, “nano-vllm,” 2025. [Online]. Available: <https://github.com/GeeekExplorer/nano-vllm>
- [27] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand *et al.*, “Mixtral of experts,” *arXiv preprint arXiv:2401.04088*, 2024.
- [28] ShareGPT, “Sharegpt,” <https://sharegpt.com/>, 2023.
- [29] Z. Tang, Z. Tang, J. Huang, X. Pan, R. Yan, Y. Wang, A. C. Zhou, S. Shi, X. Chu, and B. Li, “Dreamddp: Accelerating data parallel distributed llm training with layer-wise scheduled partial synchronization,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.11058>
- [30] K. Tae Kim, S.-J. Im, and E.-Y. Chung, “Asymgroup: Asymmetric grouping and communication optimization for 2d tensor parallelism in llm inference,” *IEEE Access*, vol. 13, pp. 120 591–120 602, 2025.
- [31] Y. Jin, Z. Yang, X. Xu, Y. Zhang, and S. Ji, “Adaptive fault tolerance mechanisms of large language models in cloud computing environments,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.12228>
- [32] D. Arfeen, D. Mudigere, A. More, B. Gopireddy, A. Inci, and G. R. Ganger, “Nonuniform-tensor-parallelism: Mitigating gpu failure impact for scaled-up llm training,” 2025. [Online]. Available: <https://arxiv.org/abs/2504.06095>
- [33] X. Zhao, S. Yang, J. Wang, L. Diao, L. Qu, and C. Wu, “Fapes: Enabling efficient elastic scaling for serverless machine learning platforms,” in *Proceedings of the 2024 ACM Symposium on Cloud Computing*, ser. SoCC ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 443–459. [Online]. Available: <https://doi.org/10.1145/3698038.3698548>
- [34] F. Cai, D. Yuan, Z. Yang, Y. Xu, W. He, W. Guo, and L. Cui, “Fastptm: Fast weights loading of pre-trained models for parallel inference service provisioning,” *Parallel Computing*, vol. 122, p. 103114, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167819124000528>