CPU-GPU Heterogeneous Computing for AI Inference: Technical Landscape and Patent Analysis

The era of CPU-assisted AI inference has arrived, but patent protection has largely expired. CaraServe's pioneering CPU-assisted cold start technique and CPU embedding offload approaches are now documented in academic literature, with **major patent filing deadlines already passed**. Production deployments at Meta, Alibaba, and ByteDance demonstrate commercial viability, while the patent landscape shows significant existing IP from Google, Intel, and Apple covering foundational heterogeneous inference concepts.

Part 1: CaraServe and CPU-assisted cold start innovations

CaraServe technical architecture delivers 1.4× latency speedup

The CaraServe paper (arXiv:2401.11240, January 20, 2024) from HKUST, CUHK-Shenzhen, Shanghai AI Lab, and Huawei Cloud presents a CPU-assisted LoRA serving system that addresses cold start latency for adapter-based LLM inference. The core innovation is remarkably elegant: **while a LoRA adapter loads from host memory to GPU, the CPU simultaneously begins prefilling using the adapter weights still resident in main memory**.

The technical implementation uses layer-wise synchronization between CPU and GPU processes. At each Transformer attention layer, the base LLM process transfers the input tensor from GPU device memory to host memory. A CPU LoRA process then performs its computation on this tensor and returns results. A custom CUDA operator enables pipelined asynchronous memory copy and signaling, reducing LoRA invocation overhead to **under 1 millisecond** through shared memory that eliminates serialization. The system achieves up to **99% SLO attainment** compared to baselines like vLLM, HuggingFace PEFT, and S-LoRA.

CaraServe's rank-aware scheduling algorithm builds performance models for BGMV (Batched Gather Matrix-Vector) and MBGMV kernels using linear regression with **$R^2 = 0.96$ accuracy**. The scheduler evaluates batch heterogeneity across LoRA adapters with different ranks and routes requests to servers that minimize cost scores. **No public code repository exists** for CaraServe as of December 2025.

ServerlessLLM optimizes storage, not CPU computation

ServerlessLLM (OSDI 2024, UC Berkeley/Edinburgh) takes a fundamentally different approach from CaraServe. Rather than using CPU for computation during loading, it exploits underutilized near-GPU storage —host DRAM, NVMe SSDs, and SATA drives—to achieve **10-200× lower latency** than KServe and Ray Serve. The system uses O_DIRECT I/O for direct disk-to-GPU transfers and a loading-optimized checkpoint format.

Its most innovative contribution is live migration of LLM inference: rather than transferring multi-gigabyte KV-caches over the network, ServerlessLLM migrates only input/generated tokens (kilobytes) and re-computes the

KV-cache at the destination. **Open source code is available** under Apache-2.0 license at github.com/ServerlessLLM with 622+ stars.

PowerInfer pioneered CPU-GPU hybrid neural computation

PowerInfer (SJTU, arXiv December 2023/SOSP 2024) presents perhaps the most direct CPU-GPU computational split for inference. It exploits the power-law distribution in neuron activation: "**hot neurons**" that consistently activate are preloaded on GPU, while "**cold neurons**" activated only for specific inputs are computed on CPU. Both processors independently compute their assigned neurons, combining results on GPU. MLP-based predictors achieve **93%+ accuracy** in identifying which neurons will activate, enabling 13-29 tokens/second on consumer GPUs.

Paper	Venue	CPU Role	Cold Start Approach	Code Available
CaraServe	arXiv 2024	LoRA execution during loading	CPU prefill concurrent with GPU load	No
ServerlessLLM	OSDI 2024	None (storage-centric)	Multi-tier storage + live migration	Yes (Apache-2.0)
PowerInfer	SOSP 2024	Cold neuron computation	Hot/cold neuron splitting	Yes
dLoRA	OSDI 2024	None	Dynamic adapter merge/unmerge	No
S-LoRA	MLSys 2024	None	Unified paging + BGMV kernels	Yes (Apache-2.0)

Part 2: CPU embedding lookup research distinguishes storage from computation

The solved versus unsolved problem

A critical distinction exists in CPU-GPU embedding architectures that research often conflates: **CPU storage with GPU lookup** (solved and widely deployed) versus **CPU lookup with GPU MLP** (actively researched but challenging). In the first paradigm, embedding tables reside in CPU DRAM while the GPU fetches and processes them. In the second, the CPU actively performs embedding lookup operations—index mapping, sparse access, and pooling—then sends dense vectors to the GPU for MLP computation.

Prism represents production-scale CPU lookup deployment

Prism (NSDI 2025, Alibaba) is the most mature production system matching the CPU lookup paradigm. Running on **10,000+ GPUs and 800,000+ CPUs** at Alibaba since late 2022, Prism disaggregates recommendation inference between CPU nodes (128 cores, 1TB memory, 200 Gbps RDMA) and GPU nodes

(128 cores, 8×A100s, 4×200 Gbps RNICs). CPU nodes execute embedding-intensive operations including lookups; GPU nodes handle MLP scoring. RDMA transfers dense embedding vectors between tiers.

The system reduces CPU fragmentation by **53%** and GPU fragmentation by **27%** while achieving 5-9× throughput improvement over baseline. During seasonal traffic events, Prism enables **90%+ GPU savings** through capacity loaning. The implementation adds 2,000 lines of Python and 3,500 lines of C++ to TensorFlow v1.12.

FleetRec pioneered heterogeneous embedding clusters

FleetRec (ETH Zurich/Alibaba, KDD 2021) introduced GPU-FPGA-CPU heterogeneous clusters for recommendation inference. FPGAs with HBM (Xilinx Alveo U280, 8GB HBM) handle embedding lookups for most tables, while **CPU servers perform lookups for tables exceeding 16GB**— [Kai-zeng] the largest tested was 200M entries at 51.2GB. GPUs (Titan RTX) handle exclusively MLP computation. Performance benchmarks show **15.5-49× throughput speedup** over CPU baselines with 21-92.5% latency reduction. [Kai-zeng]

Open source code exists at github.com/fpgasystems/GPU-FPGA-Recommendation-System, [GitHub] [github] though with limited adoption (17 stars).

GDRec takes the opposite approach

GDRec (Chinese Academy of Sciences, Journal of Computer Research and Development 2023/2024) **eliminates CPU from the embedding access path entirely**. Using CUDA Unified Virtual Addressing for DRAM and a lightweight GPU-side NVMe driver for SSD, the GPU directly accesses storage resources with zero-copy. [Ict] This falls under the "CPU storage + GPU lookup = solved" category and achieves 1.9× throughput over NVIDIA HugeCTR.

Research status summary for CPU embedding lookup

Paper	CPU Does Lookup?	Production Deployed?	Key Differentiator
Prism	Yes	Yes (10k+ GPUs at Alibaba)	RDMA disaggregation
FleetRec	Yes (large tables)	Research	FPGA primary, CPU secondary
FlexEMR	Yes	Research/prototype	Multi-threaded RDMA engine
Hotline	Yes (non-popular)	Research	Hardware accelerator required
GDRec	No (bypasses CPU)	Research	GPU direct storage access

Part 3: Industry has deployed tiered heterogeneous architectures

Meta operates 50TB+ embeddings on CPU memory with GPU compute

Meta's ZionEX platform runs **50TB+ DLRM models** with 4-8 socket CPUs providing 6TB memory per node alongside 8 GPUs. Terabyte-scale embedding tables reside in CPU main memory (DDR4/DDR5) while MLPs and dense computations execute on GPUs. Meta's RecMG paper (HPCA 2025) describes how caching/prefetch models execute on CPU to save GPU cycles across their 856+ embedding tables with 500M+ production accesses. This represents the industry's largest-scale deployment of heterogeneous embedding inference.

Alibaba's PAI platform uses PS/Worker architecture

Beyond Prism, Alibaba's PAI (Platform for AI) handles thousands of daily training jobs using parameter server/worker architecture. The **PEARL strategy** (Partitioned Embedding And Replicated) distributes large sparse variables across CPU memory on multiple PS nodes while GPU workers handle computation. Their EasyRec Processor supports deployment on both GPU instances (T4, A10, GU30) and CPU instances (Intel g6/g7/g8), with an Item Feature Cache module that caches features in CPU memory.

ByteDance processes billions of requests with heterogeneous clusters

ByteDance operates tens of thousands of GPUs (V100, P4, T4) processing **billions of inference requests daily**. Their open-source Monolith framework provides collisionless embedding tables built on TensorFlow supporting batch/real-time training and serving. BytePS optimizes distributed training for heterogeneous GPU/CPU clusters. For batch inference processing 200TB+ data, ByteDance explicitly designs pipelines where "costlier GPUs are reserved for inference while CPUs handle data preprocessing."

Google chose hardware-level acceleration over CPU offload

Rather than CPU-based embedding operations, Google invested in **TPU SparseCores**—specialized dataflow processors within TPU v4+ that accelerate embedding lookups by **5-7× using only 5% of die area and power**. Each TPU v4 chip contains 4 SparseCore processors with 2.5MB scratchpad memory each, scaling to multiple terabytes across pod slices. Snap achieved ~3× better throughput and 30% lower cost using TPU v3-32 versus 4×A100 for ad ranking by leveraging this hardware-level embedding acceleration.

NVIDIA HPS provides commercial tiered caching

NVIDIA's Merlin Hierarchical Parameter Server offers production-ready tiered caching: GPU memory (embedding cache for hot keys) → CPU memory (HashMap local or Redis distributed) → SSD storage (RocksDB). This exploits the power-law observation that **0.16% of categories are referenced by 95.9% of samples** in Criteo's 1TB dataset. HPS achieves up to **60× speedup** over CPU-only solutions nvidia and integrates with Triton Inference Server for production deployment.

Part 4: Patent landscape reveals foundational IP but gaps in cold start

Google holds core embedding-MLP pipeline patents

Google's US9141916B1 patent family (2015-2020) covers the foundational pattern of processing input features using embedding functions to generate numeric values, then processing through deep networks for classification. These establish broad claims over the **embedding → MLP pipeline architecture** central to recommendation systems. Any implementation of CPU embedding lookup followed by GPU MLP computation touches this patent space.

Intel dominates CPU-GPU routing and load balancing

Intel's patent portfolio extensively covers heterogeneous inference routing:

- **US20140052965A1** (2012): Dynamic CPU-GPU load balancing using power thresholds
- **US10304154B2**: Pre-selecting hardware portions (CPU/GPU) for inference tasks at training time
- **US20230025245A1**: OpenVINO-based heterogeneous execution with layer-wise device assignment

These patents establish claims over dynamic routing of work between CPU and GPU based on resource availability and workload characteristics—directly relevant to CaraServe-style approaches.

Apple's per-layer allocation patent is highly relevant

Apple's **EP4283526A3** describes assigning annotations indicating whether neural network operations should execute on CPU or GPU based on hardware capabilities, with per-layer optimization for heterogeneous execution. This overlaps significantly with CaraServe's layer-wise synchronization approach.

Huawei's Chinese patent explicitly covers lookup-to-CPU mapping

Huawei's **CN111723935A** describes mapping neural network operators to computing devices based on characteristics, explicitly stating that **index table look-up operators map to CPU while multiplication-heavy operations map to GPU/NPU**. This Chinese filing directly addresses the CPU embedding lookup paradigm.

Potential white space exists in cold-start specific approaches

The patent search found **no patents specifically covering CPU-assisted serving during model loading/cold start**. This represents potential white space, though it may be covered by broader claims in existing heterogeneous inference patents. Novel patentable subject matter might include:

- Cold-start specific CPU-GPU coordination mechanisms
- Model loading latency reduction through heterogeneous compute
- SLA-aware dynamic routing between CPU embedding and GPU inference during warmup

---

Part 5: Patentability analysis shows expired opportunities

CaraServe's patent window has closed

CaraServe was published on arXiv on **January 20, 2024**. Under 35 U.S.C. § 102(b)(1)(A), inventors have a 12-month grace period to file after their own disclosure. The US filing deadline of **January 20, 2025 has expired**. No USPTO filings were found for CaraServe or the specific CPU-assisted LoRA serving approach. The core innovations—CPU prefilling during GPU loading, layer-wise synchronization, rank-aware scheduling—are now **prior art** and likely unpatentable.

European patent protection was never possible: the EPO offers only a 6-month grace period limited to "evident abuse" or recognized international exhibitions—**not for voluntary academic publications**. The moment CaraServe appeared on arXiv, European patent rights were destroyed.

All major 2024 cold-start papers face expired deadlines

Paper	Publication Date	US Filing Deadline	Status
CaraServe	Jan 20, 2024	Jan 20, 2025	**EXPIRED**
ServerlessLLM	Jan 25, 2024	Jan 25, 2025	**EXPIRED**
FleetRec	Aug 2021	Aug 2022	**EXPIRED**
S-LoRA	~2023	~2024	Likely expired

Academic publication definitively constitutes prior art

Under US patent law (post-America Invents Act), prior art includes anything "described in a printed publication" before the effective filing date. ArXiv preprints are publicly accessible from their posting date and **constitute prior art as of that date**. A reference must be "disseminated or otherwise made available to the extent that persons interested and ordinarily skilled in the subject matter can locate it"—arXiv definitively meets this standard.

The 12-month grace period applies only to the **inventors' own publications**. If third parties publish similar work during that window, those disclosures remain prior art against the original inventors. International jurisdictions offer weaker or no protection:

- **European Patent Office**: No grace period for voluntary publications
- **China**: 6-month grace period, limited to specific circumstances
- **Japan/Korea**: 12 months but require formal declarations within 30 days of filing

What remains potentially patentable

Given extensive prior art, novel claims would need to demonstrate:

- **Specific implementation improvements** not disclosed in published papers
- **Novel combinations** of existing techniques with new applications
- **Hardware-software co-design** for specific configurations
- **Novel scheduling algorithms** with measurable improvements over published methods

Trade secret protection may be the only viable IP strategy for implementations where patent protection has been lost. Future researchers should coordinate with technology transfer offices **before publication** to preserve patent rights—the academic incentive to publish quickly often destroys commercial IP value.

Conclusion

The landscape of CPU-GPU heterogeneous computing for AI inference reveals a maturing field where **production deployments have outpaced patent protection**. Meta, Alibaba, and ByteDance operate at scales of 50TB+ embeddings and billions of daily requests using tiered CPU-GPU architectures, while key academic innovations from 2024 now exist in the public domain with expired patent windows.

CaraServe's elegant CPU-assisted cold start approach and the broader CPU embedding lookup paradigm remain commercially valuable despite their prior art status. The patent landscape shows foundational IP from Google (embedding-MLP pipelines), Intel (heterogeneous routing), and Apple (per-layer allocation) that any new implementation must navigate. However, white space exists in cold-start specific coordination mechanisms and disaggregated embedding server architectures.

For practitioners, NVIDIA's Hierarchical Parameter Server offers the most production-ready commercial solution for heterogeneous embedding inference, NVIDIA Developer while Prism demonstrates that CPU lookup + GPU MLP can scale to 10,000+ GPUs in production. For researchers seeking IP protection, the lesson is clear: **file before you publish**, as academic disclosure destroys patent rights faster than most realize.