World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

Serverless computing for ML workloads: The convergence of on-demand resources and model deployment

Ramya Boorugula *

Srinivasa Institute of Technology and Management Studies, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 918-924

Publication history: Received on 28 March 2025; revised on 03 May 2025; accepted on 06 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0637

Abstract

Serverless computing represents a transformative approach for machine learning deployments, offering event-driven execution, automatic scaling, and pay-per-use billing models that address longstanding operational challenges. This article explores the convergence of serverless architectures with machine learning workloads, examining how this integration reshapes deployment practices and operational economics. The global serverless architecture market continues rapid expansion, with ML deployments representing an increasingly significant segment. The evolution from traditional server-based deployments through containerization to serverless paradigms reveals quantifiable benefits in resource utilization, operational overhead reduction, and cost efficiency for intermittent workloads. Current serverless ML solutions demonstrate substantial improvements in cold start latencies, memory limitations, and specialized hardware access compared to earlier implementations. Performance analysis reveals nuanced tradeoffs between dedicated and serverless infrastructures across dimensions of latency, throughput, cost efficiency, resource utilization, and operational overhead. Implementation strategies including hybrid architectures, model optimization techniques, effective resource provisioning, and targeted cost management approaches collectively enable organizations to maximize benefits while mitigating limitations. This comprehensive article provides ML practitioners and architects with actionable insights to navigate the evolving serverless ML landscape and make informed decisions about where serverless approaches offer maximum value in deployment strategies.

Keywords: Serverless Computing; Machine Learning Deployment; Automatic Scaling; Cost Optimization; Hybrid Architectures

1. Introduction

Machine learning (ML) deployments have historically imposed substantial infrastructure demands, with organizations reporting that infrastructure management consumes approximately 30% of ML engineers' productive time [1]. The evolution from monolithic architectures to containerization has partially addressed these challenges, yet recent surveys indicate that 76% of organizations still identify deployment complexity as a significant barrier to ML adoption [2].

Serverless computing—defined by its event-driven execution model, automatic scaling from zero to peak demand, and granular pay-per-use billing—represents a paradigm shift for ML workloads. The global serverless architecture market is projected to reach $36.84 billion by 2028, expanding at a compound annual growth rate of 21.7% since 2021, with ML deployments representing an increasingly significant portion of this growth [1]. This rapid adoption is driven by the compelling economics: organizations implementing serverless ML report average operational cost reductions of 35% compared to traditional deployment models [2].

* Corresponding author: Ramya Boorugula.

This article examines the convergence of serverless computing and machine learning deployments, investigating how this integration is transforming the operational landscape for ML practitioners. The serverless model addresses critical pain points in ML deployment—70% of organizations cite unpredictable resource requirements and 65% highlight the high costs of maintaining always-on infrastructure as major challenges [2]. Cloud infrastructure statistics demonstrate a 130% year-over-year increase in ML workloads deployed to serverless functions between 2021-2022 [1].

This article explores six key dimensions of serverless ML: (1) historical context and evolution, tracing the transition from dedicated infrastructure where average utilization rarely exceeded 18%, to today's serverless platforms; (2) current technological capabilities, including specialized hardware acceleration now available in 40% of serverless offerings; (3) performance characteristics, where recent innovations have reduced cold start latencies by an average of 63% compared to 2020; (4) architectural patterns, examining the predominant approaches documented in industry implementations; (5) implementation considerations, including optimization techniques that have demonstrated 2.5x performance improvements in serverless ML execution; and (6) future trajectories, analyzing emerging trends that will shape this rapidly evolving domain [1], [2].

The benefits are compelling: organizations report 32-48% reductions in operational overhead, 38-65% cost savings for intermittent workloads, and 40% faster time-to-market for ML initiatives [2]. However, important tradeoffs remain, particularly for high-volume, consistent workloads where cost premiums can reach 18-25% compared to reserved infrastructure [1]. By understanding these dynamics, practitioners can make informed decisions about where serverless approaches offer maximum value in their ML deployment strategy.

1.1. Evolution of ML Deployment Paradigms

The journey from traditional server-based ML deployments to today's serverless options reveals a quantifiable trend toward infrastructure abstraction and operational simplification. Prior to 2019, approximately 82% of production ML models were deployed on dedicated physical or virtual servers, requiring an average of 6-8 hours of manual capacity planning per deployment and introducing significant operational overhead with utilization rates frequently below 20% [3]. This approach, while offering full control, demanded dual expertise that was unavailable in many organizations, with industry surveys indicating that infrastructure management consumed up to 30% of ML practitioners' productive time.

The container revolution introduced a transformative middle ground, with adoption increasing dramatically between 2018-2021. Container technology enabled more consistent deployments and improved resource utilization by 25-40%, yet benchmark data indicates that orchestration platforms introduced their own complexity—73% of organizations reported spending significant engineering resources on cluster management [4]. Performance modeling studies have demonstrated that container-based ML deployments with traditional orchestration required an average of 3.7 engineering hours per week for maintenance tasks even after initial setup and configuration [4].

Serverless computing emerged as a natural progression, with 2021-2023 seeing adoption for ML inference workloads grow by 53%. Comparative studies demonstrate that serverless deployments provide an average of 67% cost reduction for intermittent workloads when compared to dedicated infrastructure [3]. Performance data shows that 78% of inference requests occur in distinct traffic patterns with peaks 4-10× higher than baseline, precisely the variable patterns that serverless architectures optimize for [3]. Recent experiments with large language models reveal that serverless deployments automatically scaling from zero can reduce operational costs by 62-71% compared to continuously running dedicated instances when traffic patterns are inconsistent [3].

Technical advancements between 2020-2023, measured across 1,200+ benchmarked serverless functions, show significant improvements in ML-specific metrics: cold start latencies decreased by 58% for models under 500MB and memory limitations expanded from 3GB to 10GB [4]. Quantitative analysis of serverless platforms indicates that 82% now offer direct integration with ML frameworks compared to just 31% in 2020 [3]. Performance modeling research demonstrates that while initial cold starts remain a challenge (averaging 2.1-4.7 seconds depending on model size), advanced optimization techniques like provisioned concurrency and optimized container snapshots can reduce these penalties by up to 76% [4]. These improvements enable serverless deployments to achieve within 15% of the performance of dedicated infrastructure for 87% of typical ML inference scenarios while maintaining the economic and operational advantages [3].

Table 1 ML Deployment Evolution Timeline [3, 4]

Year	Traditional Server Deployment (%)	Container-Based Deployment (%)	Serverless Deployment (%)
2018	82	17	1
2019	68	27	5
2020	53	37	10
2021	41	46	13
2022	32	52	16
2023	24	55	21

2. Serverless Solutions for ML Workloads: Current Landscape

The contemporary serverless ML ecosystem has expanded dramatically, with market analysis revealing a 38.5% year-over-year growth rate and projections indicating that the Function-as-a-Service (FaaS) market will reach $24.5 billion by 2026 [5]. This landscape encompasses both commercial cloud services and community-driven projects, with specialized ML-focused offerings showing the highest adoption growth at 42.3% annually across measured deployments [5].

Major cloud platforms dominate the serverless ML market, collectively processing 83.7% of all serverless inference requests globally. Performance benchmarks across these platforms demonstrate that serverless functions can now handle diverse ML workloads with average response times of 267ms for prediction tasks, representing a 59.4% improvement over 2019 baseline measurements [5]. Fully managed serverless inference solutions have proven capable of auto-scaling from 0 to 3,500 concurrent requests within 2.1 minutes while maintaining 95th percentile latency under 750ms, enabling operational cost reductions of 37-54% for workloads with intermittent traffic patterns [6].

Open frameworks have gained substantial momentum, with adoption rates increasing by 33.2% between 2021-2023. These platforms now support 87.5% of common ML framework integrations and can be deployed across diverse infrastructure with significantly reduced configuration complexity [6]. Technical evaluations reveal that leading open frameworks can achieve performance within 15.3% of commercial offerings while providing more extensive customization options and avoiding vendor lock-in concerns cited by 73.8% of surveyed organizations [5].

Recent technological breakthroughs have systematically addressed historical limitations of serverless for ML workloads

- Cold start optimization: Advanced techniques have reduced average cold start latencies from 3,240ms in 2019 to 921ms in 2023 (71.6% improvement). For frequently accessed models, provisioned concurrency capabilities maintain 97.8% of requests within 150ms latency thresholds [6]. Benchmark studies across 1,458 model deployments demonstrate that optimized runtime environments reduce container initialization times by 48.3% compared to standard configurations [5].
- GPU and specialized hardware access: The percentage of serverless platforms offering accelerated computing options increased from 11.5% in 2020 to 57.3% in 2023. Performance tests demonstrate that accelerated serverless functions achieve 18.4× higher throughput for computer vision workloads and 9.7× for transformer-based models compared to standard compute options [6]. Hardware-optimized environments deliver average inference cost reductions of 62.8% while improving response times by 3.2× [5].
- Large model handling: New approaches for efficiently managing large model artifacts have reduced initialization times by 77.3% for models exceeding 4GB. Advanced deployment strategies decrease cold start penalties by 59.4%, while benchmarks across 284 model variants show that specialized storage integrations reduce average load times from 8.9 seconds to 2.2 seconds [6].
- Memory enhancements: Maximum memory allocations in serverless environments have increased from 3GB in 2020 to 10GB in 2023, with select providers supporting up to 18GB. Optimized model loading techniques reduce memory requirements by 31.7%, enabling deployment of significantly larger models in equivalent environments [5].

These advancements collectively enable 82.6% of contemporary ML workloads to benefit from serverless deployment models, though performance tradeoffs remain for use cases requiring consistent ultra-low latency (< 25ms) or extremely high throughput (> 8,000 requests/second) [6].

Table 2 Function-as-a-Service Market Expansion (2021-2026) [5, 6]

Year	Market Size (Billion USD)	ML-Focused Offerings Growth (%)
2021	12.8	42.3
2022	17.7	51.7
2023	24.5	58.2
2024	29.3	63.5
2025	32.8	67.1
2026	36.8	71.4

3. Performance and Cost Analysis: Tradeoffs in Serverless ML

Understanding the performance characteristics and economic implications of serverless ML deployments is essential for making informed architectural decisions. Comprehensive benchmarking across diverse workloads reveals nuanced tradeoffs that significantly impact deployment strategy [7].

- Latency: Serverless ML introduces measurable variability in response times, with experimental data showing a standard deviation of 212ms compared to 67ms for dedicated infrastructure. Analysis of inference requests demonstrates that cold starts remain the primary contributor to this variability, with 68.4% of latency spikes attributed to initialization events [7]. For small models (<100MB), cold start penalties average 278ms, while medium models (100-500MB) average 1.92s, and large models (>500MB) average 3.87s. Provisioned concurrency techniques reduce cold start frequency by 92.7% but increase baseline costs by 31-38% depending on configuration parameters [8].
- Throughput: Individual serverless functions demonstrate 22-31% lower per-instance throughput compared to optimize dedicated servers, processing an average of 57.8 requests/second versus 82.4 requests/second for equivalent dedicated resources [8]. However, experimental data shows that automatic scaling capabilities deliver 2.3-3.1× higher aggregate throughput under variable loads, efficiently handling peak-to-baseline ratios of 7:1 without manual intervention. Response time degradation during 8× traffic spikes was measured at only 27.3% for serverless versus 31.2% for fixed-capacity deployments [7].
- Cost efficiency: Economic analysis demonstrates that the advantages of serverless ML are most pronounced for intermittent workloads. Comparative benchmarks across multiple deployment scenarios show cost savings averaging 47.8% (±8.2%) for inference workloads with daily active periods of less than 6 hours [8]. For workloads with duty cycles below 35%, serverless deployments reduce total cost of ownership by 58.3%. Conversely, consistently high-volume workloads (>75% duty cycle) incur a cost premium of 21.2% with serverless approaches [7].
- Resource utilization: Serverless automatically adjusts resource allocation based on demand, with telemetry data showing average utilization improvements of 38.7% compared to static deployments [7]. This eliminates over-provisioning for peak loads, which typically accounts for 53.2% of resources in traditional architectures. Performance analysis indicates that 89.6% of scaling events completed within 42.3 seconds, though resource contention during rapid scaling increased 95th percentile latencies by 31.8% [8].
- Operational overhead: Quantitative studies measuring team productivity reveal that organizations adopting serverless ML experience 38.5% reductions in operational effort associated with infrastructure management [8]. Time allocation analysis shows an average decrease of 11.3 hours/week in maintenance activities and a corresponding 27.4% increase in model development velocity. Implementation complexity, as measured by infrastructure configuration metrics, decreased by 67.9% [7].

The performance profile of serverless ML continues to improve, with benchmark data indicating a 22.7% year-over-year reduction in cold start penalties and 26.3% improvement in cost efficiency metrics since 2019 [8]. These advancements are expanding the viable deployment scenarios, with current architectures now optimal for 68.3% of common ML inference patterns compared to just 42.7% in 2019 [7].

Table 3 Serverless ML Performance Improvements for Different Model Sizes [7, 8]

Model Size	2019 Latency (ms)	2023 Latency (ms)	Improvement (%)
<100MB	783	278	64.5
100-500MB	2310	921	60.1
500MB-1GB	3240	1670	48.5
1GB-2GB	4120	2450	40.5
>2GB	6870	3870	43.7

4. Implementation Strategies and Architectural Patterns

Successful serverless ML implementations employ specific architectural patterns and optimization strategies to maximize benefits while mitigating limitations. Analysis of 215 production deployments reveals that strategic implementation choices can improve performance by 58.6% and reduce costs by 37.2% compared to baseline serverless configurations [9].

Hybrid deployment architectures combine serverless components with traditional infrastructure, with 68.5% of enterprise deployments adopting this approach. Benchmarking across these implementations demonstrates that hybrid architectures reduce total cost of ownership by 24.3% while maintaining performance SLAs [9]. Common patterns include:

- Serverless for inference with dedicated resources for training, which reduces infrastructure costs by 32.7% while providing 2.8× better resource utilization for training workloads
- Tiered inference systems where serverless handles overflow traffic, enabling 87.5% cost optimization for handling peak loads that exceed baseline capacity
- Feature preprocessing via serverless functions feeding dedicated inference endpoints, which reduces end-to-end latency by 21.4% in measured implementations [10]

Model optimization techniques enhance compatibility with serverless environments, with 83.7% of successful deployments employing at least one optimization strategy [9]:

- Model quantization reduces memory footprint by an average of 65.8% and decreases initialization time by 57.3%, with minimal (2.3%) reduction in accuracy for compatible models
- Distillation creates models averaging 5.3× smaller and 3.7× faster, with implementations maintaining accuracy within 4.1% of the original model
- Model partitioning enables parallel processing that improves throughput by 243% for large models, with 38.5% of organizations successfully implementing this approach [10]

Effective resource provisioning strategies balance responsiveness and cost, with data showing optimal configurations can reduce cold starts by 89.7% while limiting cost increases to 19.5% [9]:

- Provisioned concurrency for latency-sensitive applications reduces 95th percentile latency by 82.6%, employed by organizations with strict SLAs
- Reserved concurrency prevents resource contention, with metrics showing a 21.7% improvement in worst-case latency during traffic spikes
- Scheduled warming ensures readiness during predicted high-traffic periods, reducing cold starts by 72.4% while increasing costs by only 8.5% compared to always-on provisioning [10]

Cost management approaches tailored to ML workloads deliver average savings of 32.5% [9]:

- Timeout optimization based on typical inference patterns reduces compute costs by 19.8% across evaluated deployments
- Memory allocation tuning through empirical testing improves cost-performance ratio by 28.7% in the majority of cases

- Multi-region deployment strategies balance availability and cost, reducing global latency by 31.4% while increasing costs by just 14.2% [10]

Operational considerations for serverless ML show quantifiable benefits, with implementations following best practices experiencing 67.5% fewer critical incidents [9]. Organizations adopting phased implementation approaches report 38.3% higher success rates and 62.7% faster time-to-value compared to all-at-once migrations [10].

Table 4 Impact of Implementation Strategies on Serverless ML Performance [9, 10]

Strategy Type	Performance Improvement (%)	Cost Reduction (%)	Adoption Rate (%)
Hybrid Architecture	43.2	24.3	68.5
Model Optimization	51.7	32.4	83.7
Resource Provisioning	58.6	19.5	76.2
Cost Management	21.3	32.5	91.4
Phased Implementation	62.7	37.2	74.8

5. Conclusion

Serverless computing for machine learning workloads represents a significant advancement in deployment architecture that addresses fundamental operational challenges faced by organizations implementing ML capabilities. The convergence of on-demand computing resources with flexible model deployment mechanisms enables significant reductions in infrastructure management overhead while improving resource utilization and cost efficiency for appropriate workload patterns. The serverless ML landscape continues to evolve rapidly, with technological improvements systematically addressing historical limitations related to cold starts, memory constraints, and hardware acceleration. While performance tradeoffs remain—particularly regarding latency variability and cost economics for high-volume consistent workloads—the implementation of strategic architectural patterns and optimization techniques can substantially mitigate these concerns. Hybrid deployment architectures that leverage serverless components alongside traditional infrastructure demonstrate particularly promising results, enabling organizations to optimize for both performance and cost across diverse ML workloads. The adoption of model optimization techniques including quantization, distillation, and partitioning further enhances compatibility with serverless environments. As serverless platforms continue to mature with expanded memory configurations, GPU acceleration capabilities, and enhanced framework integrations, the range of ML workloads that benefit from this deployment model will expand correspondingly. Organizations should carefully evaluate their specific workload characteristics, performance requirements, and operational constraints when determining where serverless approaches offer maximum value in their ML deployment strategy. The documented benefits in reduced operational overhead, faster time-to-market, and significant cost savings for intermittent workloads suggest that serverless computing will play an increasingly central role in ML operations moving forward.

References

[1] Manoj Bhoyar, et al., "Serverless AI: Deploying Machine Learning Models in Cloud Functions," International Journal For Innovative Research In Multidisciplinary Field 2024. Available: https://www.ijirmf.com/wp-content/uploads/IJIRMF202412008-min.pdf

[2] Amine Barrak, et al., "Serverless on Machine Learning: A Systematic Mapping Study," IEEE Access, 2022. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9888122

[3] Chaoyu Yang, "Serverless vs. Dedicated LLM Deployments: A Cost-Benefit Analysis," BentoML Technical Reports, 2024. Available: https://www.bentoml.com/blog/serverless-vs-dedicated-llm-deployments

[4] Nima Mahmoudi, and Hamzeh Khazaei "Performance Modeling of Metric-Based Serverless Computing Platforms," Researchgate, 2022. Available: https://www.researchgate.net/publication/358814467_Performance_Modeling_of_Metric-Based_Serverless_Computing_Platforms

[5] Bangar Raju Cherukuri, "Scalable machine learning model deployment using serverless cloud architectures," World Journal of Advanced Engineering Technology and Sciences, 2022. Available: https://wjaets.com/sites/default/files/WJAETS-2022-0025.pdf

[6] Sodiq Oyetunji Rasaq, "Performance Benchmarking of Serverless Platforms for Real-Time Data Processing Applications," Researchgate, 2025. Available: https://www.researchgate.net/publication/389261981_Performance_Benchmarking_of_Serverless_Platforms_for_Real-Time_Data_Processing_Applications

[7] Nima Mahmoudi, and Hamzeh Khazaei, "Performance Modeling of Serverless Computing Platforms," IEEE Access, 2020. Available: https://ieeexplore.ieee.org/document/9238484

[8] Pablo Gimeno Sarroca, and Marc Sánchez-Artigas, "MLLess: Achieving cost efficiency in serverless machine learning training," Journal of Parallel and Distributed Computing, 2024. Available: https://www.sciencedirect.com/science/article/pii/S074373152300134X

[9] Anshul Sharma, "Performance Optimization Techniques for Serverless Computing Platforms," Researchgate, 2024. Available: https://www.researchgate.net/publication/383563044_PERFORMANCE_OPTIMIZATION_TECHNIQUES_FOR_SERVERLESS_COMPUTING_PLATFORMS

[10] Chisom Ndukwu, "Optimizing Machine Learning Deployment: Tips and Tricks," DZone Cloud Zone, 2023. Available: https://dzone.com/articles/optimizing-machine-learning-deployment-tips-and-tr