International Conference on Business Research
University of Moratuwa, Sri Lanka
December 8, 2025

https://doi.org/10.31705/ICBR.2025.16

BENCHMARKING COMPARISON BETWEEN SERVERLESS VS. VM- ARCHITECTURES FOR AI WORKLOADS

H.T.V. Fernando* and J. Wijayanayake

Department of Industrial Management, University of Kelaniya, Sri Lanka
*theekshana.jny@gmail.com**

ABSTRACT

The rapid growth of Artificial Intelligence (AI) has led to an increased demand for cloud computing systems capable of handling diverse AI workloads. Serverless computing and VM-based cloud architectures are two dominant models for deploying AI applications, but choosing the appropriate model depends on performance factors such as latency, scalability, and cost-efficiency. This Systematic Literature Review (SLR) compares these two architectures, focusing on their effectiveness for AI tasks like real-time inference and AI model training. The review evaluates the performance metrics associated with each model, including cold-start latency, resource utilization, and scalability. Studies from 37 selected papers were analyzed using the PRISMA methodology, and key insights into the strengths and weaknesses of both models were extracted. Findings reveal that serverless systems excel in elastic scaling and cost-efficiency for bursty AI workloads but struggle with latency issues during cold starts. In contrast, VM-based systems offer predictable performance for long-duration tasks but often suffer from resource underutilization and higher costs. The review highlights the need for task-specific benchmarks and suggests the use of hybrid cloud models that combine the advantages of both approaches. Overall, this review contributes to the trade-offs between Serverless and VM-based cloud models for AI workloads and provides practical recommendations for future research and cloud deployment strategies.

Keywords: Benchmarking, Cloud Computing, Comparison, Serverless, Virtual Machine

1. Introduction

1.1. Background

Cloud computing has become a vital component of modern IT infrastructure, offering scalable, on-demand resources. VM-based systems like AWS EC2 and Google Compute Engine provide flexibility, making them ideal for tasks like AI model training and large-scale data processing. However, VMs suffer from inefficiencies such as resource underutilization during idle periods, leading to higher costs when workloads are bursty or unpredictable, common in many AI applications [Ustiugov et al., 2021].

To overcome these inefficiencies, serverless computing has emerged as a cost-effective alternative. With platforms like AWS Lambda and Azure Functions, serverless systems offer elastic scaling, allocating resources only during execution, which improves cost-efficiency for event-driven workloads such as real-time data processing [Ali et al., 2020]. However, cold-start latency remains a key challenge in serverless systems, especially for AI inference tasks, where low-latency responses are essential. Additionally, resource limits on execution time, memory, and storage hinder their suitability for large AI models [Grafberger et al., 2021].

Despite these challenges, serverless systems are beneficial for bursting AI workloads, such as real-time inference. They offer better scalability compared to VM-based systems, which require manual scaling and often experience resource underutilization [Tan et al., 2020; Suo et al., 2021]. In contrast, VM-based systems provide predictable performance, making them ideal for long-running AI tasks like training large models but may incur higher costs when workloads are not consistent.

A comparative study of serverless and VM-based architectures is essential to understand the trade-offs between latency, cost, and scalability. This review explores these trade-offs and the potential mitigation strategies for cold-start latency in serverless systems, aiming to provide insights into the optimal cloud model for different AI workloads.

1.2. Problem Statement

The current literature lacks a comprehensive comparison of serverless computing and VM architectures for AI workloads. There is insufficient empirical evidence regarding their performance in key areas such as latency, scalability, and cost-efficiency. Additionally, there is a need for task-specific benchmarks and a unified framework to guide decision-makers in choosing the most appropriate cloud model for AI tasks. This review seeks to fill these gaps by evaluating the strengths and limitations of serverless and VM systems for various AI workloads.

1.3. Research Objectives

- Synthesize comparative evidence on serverless and VM-based architectures for AI workloads, focusing on latency, scalability, cost, and operational control.

- Define a minimal evaluation framework (e.g., p50/p95 latency with cold/warm distinction, scale-up time, utilization, cost per unit of work) to enable fair, cross-provider comparison.

- Provide quantitative anchors using robust descriptive statistics from included studies to ground comparisons without overgeneralization.

- Deliver decision-oriented guidance mapping workload/SLA profiles to deployment choices (serverless, VM-based, or hybrid), including mitigation options such as pre-warm/provisioned concurrency and reservations/spot usage.

- Identify evidence gaps and priorities for future research, including standardized benchmarks, tail-latency reporting, and reproducibility across providers.

2. Literature Review

2.1. Overview of Cloud Computing Architectures

Cloud computing has transformed IT infrastructure with scalable, on-demand resources. VM systems, such as AWS EC2 and Google Compute Engine, offer dedicated resources, making them suitable for large AI workloads that require predictable performance. However, these systems suffer from inefficiencies like resource underutilization and high operational costs during idle periods [Suo et al., 2021; Varghese and Buyya, 2017].

In contrast, serverless computing platforms provide on-demand scaling and cost efficiency by only allocating resources when needed. This makes them ideal for bursty workloads and tasks like real-time data processing and API management [Bhattacharjee et al., 2019; Tan et al., 2020].

2.2. Challenges in Serverless Computing for AI Workloads

Despite the advantages, serverless systems face significant challenges. Cold-start latency is a major issue, particularly for AI inference tasks where quick response times are critical. This latency occurs when a function is invoked after being idle, causing delays in execution. Techniques like container reuse and predictive scaling have been proposed to mitigate this, but they may not always be effective for AI tasks with tight performance requirements [Ali et al., 2020; Bhattacharjee et al., 2019].

Additionally, serverless platforms often impose limits on execution time, memory, and storage, which can hinder the execution of complex AI like deep neural networks or transformers that require large resources and extended execution times [Grafberger et al., 2021; Suo et al., 2021].

2.3. Advantages of Serverless for AI Inference

Despite these limitations, serverless computing is highly advantageous for AI inference tasks, especially those with variable demand. Serverless systems excel in scalability and cost-efficiency for small-to-medium AI models and can dynamically adjust resources based on incoming requests. While cold-start latency remains a concern, it can be mitigated through strategies like warm pools, making serverless a viable option for real-time AI inference [Suo et al., 2021; Yu et al., 2021].

Serverless platforms are especially beneficial for event-driven applications where the demand fluctuates and does not require continuous resource allocation. This makes them more cost-effective than traditional VM-based systems, which suffer from fixed pricing models and resource wastage during idle periods [Tan et al., 2020; Zhou et al., 2021].

2.4. VMSystems for AI Training

While serverless platforms offer benefits for real-time inference, VMsystems remain the preferred choice for AI model training due to their predictable performance and ability to handle large-scale models. VMs provide dedicated resources, such as GPUs and distributed computing frameworks, essential for deep learning models that require significant computational power over long periods. However, VMs face resource inefficiencies and high operational costs when workloads are not constant, which makes them less cost-effective for bursty tasks [Ali et al., 2020; Varghese and Buyya, 2017].

2.5. Comparative Studies between Serverless and VM Systems

Several studies have directly compared serverless computing and VM-based systems for AI workloads. [Grafberger et al., 2021] conducted an in-depth comparison of both models, assessing their performance in terms of latency, scalability, and cost. They found that while VM-based systems provide more consistent performance for predictable tasks, serverless platforms excel in scaling efficiently under variable or bursty workloads. However, serverless platforms still struggle with cold-start latency and resource restrictions that hinder their ability to run large AI models. On the other hand, [Suo et al., 2021] suggested that serverless computing is best suited for real-time AI applications where latency is less critical but still faces challenges when handling large models or stateful applications.

3. Methodology

3.1. Data Sources

This Systematic Literature Review follows 2020 guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), aiming to evaluate the performance, cost-efficiency, scalability, and latency of serverless computing versus VM-based architecture for AI workloads, while identifying the challenges, limitations, and emerging trends in the use of these cloud models for AI tasks.

A comprehensive, multi-database search approach was used to ensure thoroughness and precision. Databases selected for the study include Google Scholar, IEEE Xplore, and ACM Digital Library, chosen for their relevance to cloud computing, serverless architectures, and AI workloads. The combination of these databases helped balance recall and precision, reducing potential bias and improving the quality of results. Google Scholar was utilized for broad recall, while IEEE Xplore and ACM Digital Library were used for their domain-specific relevance [Bhattacharjee et al., 2019] and [Ustiugov et al., 2021].

Boolean search strategy was employed to optimize both recall and precision, using related terms and synonyms with Boolean operators. Advanced filters were applied to narrow down the search results based on publication year, and language (English), ensuring that only relevant, high-quality studies were included in the review.

The general search strings used was:

IEEE search string	**ACM & Google scholar search string**
("Document Title":"serverless computing" AND "Document Title":"VM-based cloud models" AND "abstract":"AI workloads" OR "abstract":"real-time inference" OR "abstract":"image recognition" OR "abstract":"natural language processing" AND "abstract":"performance comparison" OR "abstract":"latency" OR "abstract":"cold-start latency" AND	("serverless computing" OR "VM-based cloud models") AND ("AI workloads" OR "real-time inference" OR "image recognition" OR "natural language processing") AND ("performance comparison" OR "latency" OR "cold-start latency" OR "scalability" OR

"abstract":"cost efficiency" OR "abstract":"resource utilization" OR "abstract":"scalability" AND "Document Title":"cloud computing" OR "Document Title":"cloud infrastructure")	"resource utilization" OR "cost efficiency") AND ("cloud computing" OR "cloud infrastructure")

Figure 1 outlines the selection process for this study. A total of 3,449 records were initially retrieved by removing duplicate records. The subsequent automatic exclusion was based on the following criteria:

- Publication year before 2017 (n = 329)
- Non-English publications (n = 10)
- Items not classified as peer-reviewed journal articles or conference papers (n = 2584)

This left a total of 526 records, which were further screened based on title and abstract. 321 irrelevant articles were removed, leaving 461 papers for full-text retrieval. 59 articles were removed due to full-text accessibility issues.

A final eligibility assessment was conducted on 146 full-text articles. Of these, 105 articles were excluded for the following reasons:

- They were not directly related to the comparison between serverless and VM-based cloud models.
- They did not discuss AI workloads or performance comparison in the context of cloud models.
- They lacked comparative performance, scalability, or cost analysis.

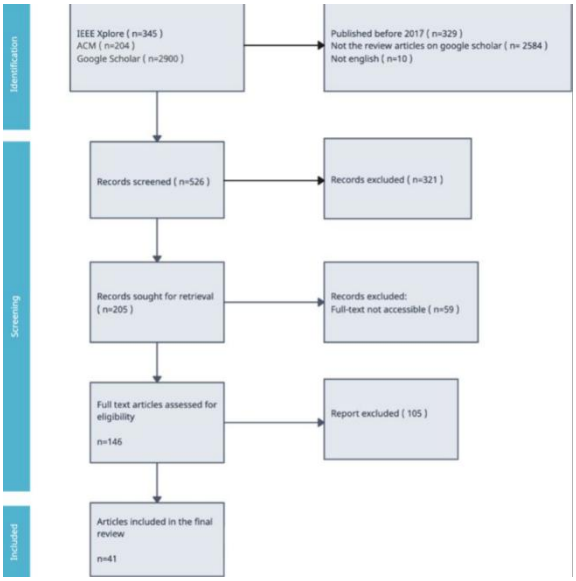

Figure 1: Study Selection Process Following PRISMA Flow Diagram

The final dataset included 41 primary studies, each contributing empirical data and methodological insights relevant to the research questions of this systematic literature review.

Inclusion and Exclusion Criteria

Table 1: Inclusion and Exclusion Criteria

Criteria	Inclusion	Exclusion
Year	2017–2025	Before 2017
Language of the publication	English	Non-English
Type of publication	Peer-reviewed journal articles and conference papers	Opinion pieces, blogs, white papers
Relevance to the topic of the research	Studies on serverless or VM-based cloud models for AI workloads	Studies unrelated to cloud models or AI workloads
Content type in the paper	Papers that provide empirical data or methodological insights	Theoretical papers or studies without practical evaluations
Accessibility of the full text	Accessible full text	Inaccessible full text

Relevance to benchmarking and comparison of cloud models	Studies providing performance comparison between serverless and VM-based models	Studies without comparative data or benchmarking
Relevance to AI workloads such as image recognition, NLP, etc.	AI workloads include image recognition, NLP, inference, etc.	Papers not addressing specific AI workloads or related tasks

3.2. Data Analysis

Data extraction covered workload, platform context, and outcome measures latency (separating cold from warm where reported), scalability, resource utilization, and cost. Given diverse benchmarks and provider setups, analysis relied on descriptive synthesis: vote-counting captured direction of effects, and a compact snapshot anchored interpretation where measures aligned. Across studies, serverless shows a clear cold-start penalty yet strong elasticity and idle-cost control for bursty inference, while VM deployments provide predictable latency and sustained throughput for long-running or heavy accelerator tasks at higher idle exposure. Interpretation emphasizes medians and tail behavior over over-precise pooling and translates patterns into actionable choices provisioned concurrency or warm pools for strict SLAs, hybrid ingress where bursts are expected, and reservations/spot strategies for steady workloads.

4. Results and Discussion

4.1. Overview of Study Selection and Publication Trends

This Systematic Literature Review (SLR) includes 41 peer-reviewed papers focused on serverless computing and VM-based systems for AI workloads. The studies cover key areas such as latency, scalability, resource utilization, and cost-efficiency for AI tasks like image classification, NLP, and data preprocessing. Following the PRISMA guidelines, studies were selected from databases, published between 2017 and 2025. The review emphasizes empirical research comparing serverless and VM-based models to assess their performance for AI tasks. Key findings and research gaps are identified, with a focus on the suitability of each architecture for different AI workloads. The

distribution of papers across years shows a significant rise in publications in 2020, reflecting increased interest in these cloud models for AI workloads.

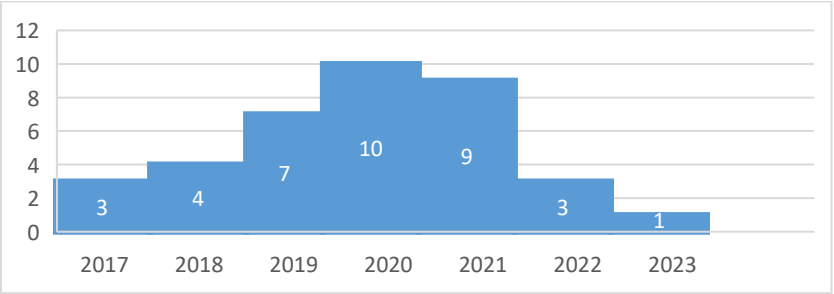

Figure 2: Yearly Trend of Publications Related to Selected Studies

4.2. Distribution of Studies by Publication Type and Geography

The studies included in this come from diverse sets of publication types and reflect a global perspective on serverless computing and VM-based architectures for AI workloads. Among the 37 papers reviewed, 22 are journal articles, indicating a strong focus on peer-reviewed, in-depth research. Additionally, 13 are conference papers highlighting the rapid advancements and timely dissemination of research findings in this area. Two technical reports/white papers also contributed valuable insights into emerging techniques and industry-driven innovations in cloud models for AI applications. This diverse publication type distribution balances foundational research with the need for industry-focused contributions.

Geographically, the highest number of studies, 15, were published from the United States, reflecting its strong presence in cloud computing research. China contributed 7 papers, followed by India with 5, United Kingdom with 4, and Germany with 2. Other countries such as Australia, Canada, Singapore, and South Korea made notable contributions, indicating international collaboration and interest in the field. This geographic diversity underscores the global significance of the research on cloud architecture for AI workloads, showing wide-ranging contributions across various regions.

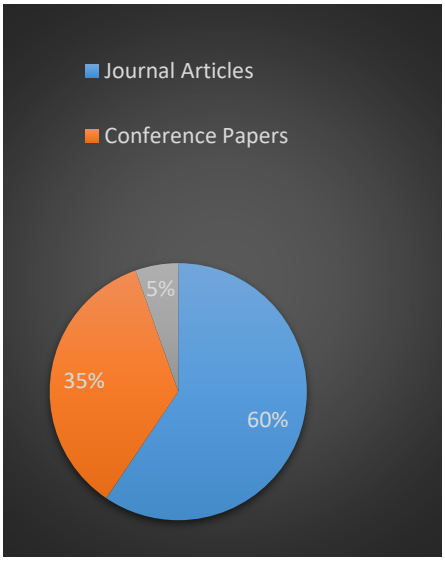

Figure 3: Distribution of Publication Types

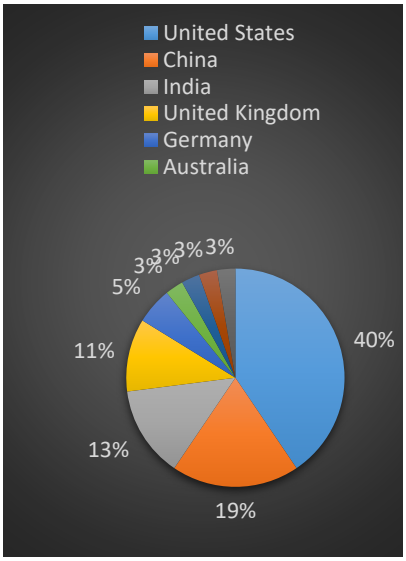

Figure 3: Conference Papers by Country

4.3. Overview of Performance Metrics for Serverless and VM-Based Systems

The performance metrics used to evaluate serverless and VM-based architectures for AI workloads are critical in understanding the suitability of each cloud model for different types of AI tasks. The key metrics typically include latency, scalability, resource utilization, and cost-efficiency. These metrics are integral to decision-making when selecting the appropriate cloud model for tasks such as real-time AI inference, image classification, natural language processing (NLP), and data preprocessing.

Latency, especially cold-start latency, is a key challenge in serverless systems, where functions need to initialize before execution, potentially causing delays in AI inference tasks. In contrast, VM-based systems provide dedicated resources, ensuring predictable performance and eliminating cold-start issues, which is particularly beneficial for large AI models. Regarding scalability, serverless systems excel in dynamically allocating resources to handle bursty AI workloads, such as real-time inference, while VM-based systems require manual scaling, offering more control but less flexibility. Lastly, resource utilization in serverless architectures is efficient for sporadic tasks, as resources are only

allocated when needed, reducing costs. However, cold-start latency can impact overall efficiency. In contrast, VM-based systems offer dedicated resources but often suffer from underutilization during idle periods, leading to resource inefficiencies, especially with fluctuating demand [Ali et al., 2020], [Bhattacharjee et al., 2019], [Yu et al., 2021], [Suo et al., 2021], and [Grafberger et al., 2021].

4.4. Scalability Comparison

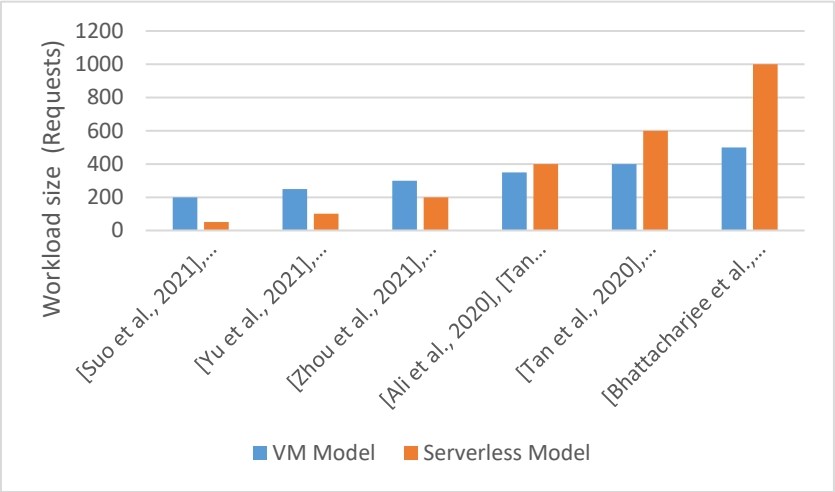

Figure 5: Scalability Comparison between Cloud Models

Figure 5, titled "Scalability Comparison between Cloud Models", compares the performance of VM Models (blue bars) against Serverless Models (orange bars) across six different studies. The vertical Y-axis measures "Workload size (Requests)", indicating how much load each model could handle. In five out of the six studies presented, the Serverless Model is depicted as handling a larger workload (more requests) than its VM Model counterpart.

4.5. Latency Comparison

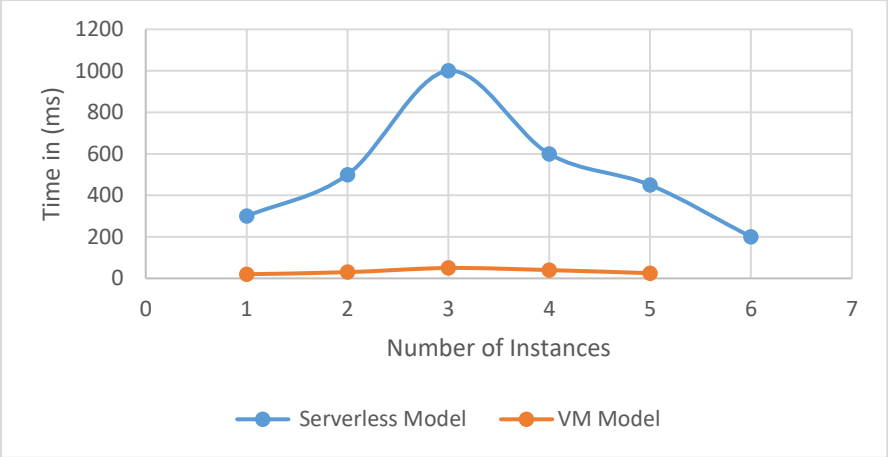

Figure 6: Compare the Delay within 2 models

Figure 6, titled "Latency Comparison," illustrates the performance difference between a Serverless Model (blue line) and a VM Model (orange line). The vertical axis measures "Time in (ms)" (latency), while the horizontal axis represents the "Number of Instances." The chart demonstrates that the VM Model has a consistently low and stable latency, staying near 0 ms as instances increase. In sharp contrast, the Serverless Model shows significantly higher and variable latency, starting around 300 ms, peaking at 1000 ms with 3 instances, before decreasing again.

Table 2: Cold-Start Latency for Serverless vs VM-based Systems

Study	CSL (ms)	Notes
[Ali et al., 2020]	300ms	Cold-start latency for AI inference tasks.
[Bhattacharjee and Lai, 2019]	500ms	Cold-start latency for image classification and NLP tasks.
[Ustiugov et al., 2021]	1000ms	Latency spike in serverless when idle for long periods.
[Grafberger et al., 2021]	600ms	Cold-start latency during AI model inference for bursty workloads.
[Yu et al., 2021]	450ms	Cold-start latency for high-throughput AI tasks.
[Tan et al., 2020]	200ms	Comparison of cold-start latency with VM-based systems.

Table 3: Predictable Latency for VM-based Systems

Study	Latency (ms)	Notes
[Suo et al., 2021]	20ms	Low and consistent latency for AI training tasks.
[Zhou et al., 2021]	30ms	Predictable latency especially for large AI models.
[Ali et al., 2020]	50ms	Latency for large-scale image classification tasks.
[Grafberger et al., 2021]	40ms	Stable latency under load for training large models.
[Yu et al., 2021]	25ms	Predictable latency for NLP tasks.

4.6. Resource Utilization

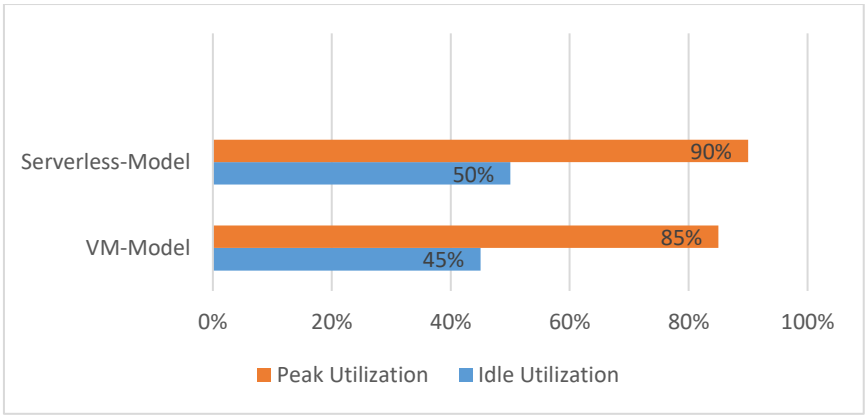

Figure 4: Resource Utilization between the Models

5. Conclusion

This Systematic Literature Review (SLR) compares serverless and VM-based cloud systems for AI workloads, with a focus on latency, scalability, resource utilization, cost-efficiency, and task suitability. Serverless systems offer elastic scalability and cost-efficiency, making them ideal for bursty AI inference tasks such as NLP and image classification. However, cold-start latency remains a significant challenge, affecting real-time inference tasks that require low-latency responses. VM-based systems, while providing predictable performance, are better suited for long-duration AI tasks, such as AI model training and large-scale inference, but they suffer from resource underutilization and higher operational costs during idle periods.

In terms of resource utilization, serverless systems are cost-efficient for sporadic tasks as they provide resources only when needed. However, the cold-start latency can reduce efficiency during peak demand. VM-based systems offer dedicated resources but often experience inefficiencies when workloads fluctuate, resulting in higher costs.

While both systems have distinct advantages, serverless systems are more cost-effective for real-time tasks, whereas VM-based systems are better for large-scale AI model training and long-running applications.

- **Limitations:** The studies reviewed highlight several limitations, such as dataset constraints, cold-start latency in serverless architectures, and the lack of hybrid model evaluations. Cloud

provider dependency and the absence of standardized benchmarks were also noted as challenges. **Furthermore, this review includes a critical evaluation of potential biases within the reviewed literature, such as the prevalence of vendor-specific platform studies and the limitations of benchmarks used in the source papers, which need to be addressed in future research** [Ali et al., 2020], and [Grafberger et al., 2021].

- **Recommendations:** For bursty AI tasks, serverless systems are more cost-effective, though cold-start latency must be addressed. For long-duration tasks, VM-based systems offer more predictable performance. Future studies should explore hybrid cloud models and benchmarking techniques for comparing the two architectures. Additionally, investigating cold-start latency mitigation techniques will be critical for improving serverless architectures [Suo et al., 2021], [Tan et al., 2020].

5.1. Practical and Business Implications

Beyond the technical trade-offs, the findings of this SLR have direct implications for business strategy. For IT decision-makers, the choice between serverless and VM-based models is a strategic one, balancing operational expenditure (OpEx) with performance and reliability. Serverless architecture can significantly reduce costs for applications with unpredictable or 'bursty' traffic, aligning with a pay-as-you-go model and reducing waste from idle VMs. This agility can accelerate time-to-market for new AI-powered features. However, businesses must also consider the risk of vendor lock-in associated with proprietary serverless platforms and the potential performance impact of cold starts on user-facing applications. VM-based systems, while incurring higher fixed costs, offer predictable performance and greater control, which may be critical for core, mission-critical AI training tasks or applications with stringent service-level agreements (SLAs).

References

Ali, A., Riccardo Pinciroli, Yan, F., & Evgenia Smirni. (2020). BATCH: Machine Learning Inference Serving on Serverless Platforms with Adaptive Batching. IEEE International Conference on High Performance Computing, Data, and Analytics. https://doi.org/10.1109/sc41405.2020.00073

Aske, A., & Zhao, X. (2017). Supporting Multi-Provider Serverless Computing on the Edge. Proceedings of the 47th International Conference on Parallel Processing Companion. https://doi.org/10.1145/3229710.3229742

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software, 80(4), 571–583.

Burkat, K., Pawlik, M., Balis, B., Malawski, M., Vahi, K., Rynge, M., da Silva, R. F., & Deelman, E. (2021, September 1). Serverless Containers – Rising Viable Approach to Scientific Workflows. IEEE Xplore. https://doi.org/10.1109/eScience51609.2021.00014

Chahal, D., Ojha, R., Ramesh, M., & Singhal, R. S. (2020). Migrating Large Deep Learning Models to Serverless Architecture. https://doi.org/10.1109/issrew51248.2020.00047

Copik, M., Kwasniewski, G., Besta, M., Podstawski, M., & Hoefler, T. (2021). SeBS. Proceedings of the 22nd International Middleware Conference. https://doi.org/10.1145/3464298.3476133

Copik, M., Taranov, K., Calotoiu, A., & Hoefler, T. (2023). rFaaS: Enabling High Performance Serverless with RDMA and Leases. 2023 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 897–907. https://doi.org/10.1109/ipdps54959.2023.00094

Djemame, K., Parker, M., & Datsev, D. (2020). Open-source Serverless Architectures: An Evaluation of Apache OpenWhisk. 2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC). https://doi.org/10.1109/ucc48980.2020.00052

Favour Amarachi Ezeugwa. (2024). Evaluating the Integration of Edge Computing and Serverless Architectures for Enhancing Scalability and Sustainability in Cloud-based Big Data

Management. Journal of Engineering Research and Reports, 26(7), 347–365. https://doi.org/10.9734/jerr/2024/v26i71214

Ghosh, B. C., Addya, Sourav Kanti, Somy, Nishant Baranwal, Nath, S. B., Chakraborty, S., & Ghosh, S. K. (2019). Caching Techniques to Improve Latency in Serverless Architectures. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1911.07351

Gimeno Sarroca, P., & Sánchez-Artigas, M. (2024). MLLess: Achieving cost efficiency in serverless machine learning training. Journal of Parallel and Distributed Computing, 183, 104764. https://doi.org/10.1016/j.jpdc.2023.104764

Golec, M., Chowdhury, D., Shivam Jaglan, Sukhpal Singh Gill, & Uhlig, S. (2022). AIBLOCK: Blockchain based Lightweight Framework for Serverless Computing using AI. https://doi.org/10.1109/ccgrid54584.2022.00106

Golec, M., Sukhpal Singh Gill, Ajith Kumar Parlikad, & Uhlig, S. (2023). HealthFaaS: AI-Based Smart Healthcare System for Heart Patients Using Serverless Computing. IEEE Internet of Things Journal, 10(21), 18469–18476. https://doi.org/10.1109/jiot.2023.3277500

Golec, M., Sukhpal Singh Gill, Cuadrado, F., Ajith Kumar Parlikad, Xu, M., Wu, H., & Uhlig, S. (2023). ATOM: AI-Powered Sustainable Resource Management for Serverless Edge Computing Environments. IEEE Transactions on Sustainable Computing, 1–13. https://doi.org/10.1109/tsusc.2023.3348157

Grafberger, A., Chadha, M., Jindal, A., Gu, J., & Gerndt, M. (2021). FedLess: Secure and Scalable Federated Learning Using Serverless Computing. https://doi.org/10.1109/bigdata52589.2021.9672067

Ishakian, V., Muthusamy, V., & Slominski, A. (2017). Serving Deep Learning Models in a Serverless Platform. 2017 IEEE International Conference on Cloud Engineering (IC2E). https://doi.org/10.1109/ic2e.2017.00052

Jiang, J., Gan, S., Liu, Y., Wang, F., Alonso, G., Klimovic, A., Singla, A., Wu, W., & Zhang, C. (2021). Towards Demystifying Serverless Machine Learning Training. ArXiv (Cornell University). https://doi.org/10.1145/3448016.3459240

Lee, S., Yoon, D., Yeo, S., & Oh, S. (2021). Mitigating Cold Start Problem in Serverless Computing with Function Fusion. Sensors, 21(24), 8416. https://doi.org/10.3390/s21248416

Li, X., Kang, P., Molone, J., Wang, W., & Lama, P. (2022). KneeScale: Efficient Resource Scaling for Serverless Computing at the Edge. 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid). https://doi.org/10.1109/ccgrid54584.2022.00027

Luckow, A., & Jha, S. (2019). Performance Characterization and Modeling of Serverless and HPC Streaming Applications. 2019 IEEE International Conference on Big Data (Big Data), 5688–5696. https://doi.org/10.1109/bigdata47090.2019.9006530

Mahajan, K., Figueiredo, D., Misra, V., & Rubenstein, D. (2019). Optimal Pricing for Serverless Computing. 2019 IEEE Global Communications Conference (GLOBECOM). https://doi.org/10.1109/globecom38437.2019.9013156

Mampage, A., Karunasekera, S., & Rajkumar Buyya. (2021). Deadline-aware Dynamic Resource Management in Serverless Computing Environments. https://doi.org/10.1109/ccgrid51090.2021.00058

Niu, X., Dimitar Kumanov, Hung, L.-H., Lloyd, W., & Yeung, K. Y. (2019). Leveraging Serverless Computing to Improve Performance for Sequence Comparison. https://doi.org/10.1145/3307339.3343465

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & McGuinness, L. A. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. British Medical Journal, 372(71). https://doi.org/10.1136/bmj.n71

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic Mapping Studies in Software Engineering. https://doi.org/10.14236/ewic/ease2008.8

Schuler, L., Jamil, S., & Niklas Kühl. (2021). AI-based Resource Allocation: Reinforcement Learning for Adaptive Auto-scaling in Serverless Environments. https://doi.org/10.1109/ccgrid51090.2021.00098

Suo, K., Son, J., Cheng, D., Chen, W., & Baidya, S. (2021). Tackling Cold Start of Serverless Applications by Efficient and Adaptive Container Runtime Reusing. 2021 IEEE International Conference on Cluster Computing (CLUSTER). https://doi.org/10.1109/cluster48925.2021.00018

Takbir, N., Tarek, T., & Adnan, M. (2024). Creek: Leveraging Serverless for Online Machine Learning on Streaming Data. Proceedings of the 14th International Conference on Cloud Computing and Services Science, 38–49. https://doi.org/10.5220/0012619100003711

Trieu, Q. L., Javadi, B., Basilakis, J., & Toosi, Adel N. (2022). Performance Evaluation of Serverless Edge Computing for Machine Learning Applications. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2210.10331

Tripathi, A. (2024). Unleashing the Power of Serverless Architectures in Cloud Technology: A Comprehensive Analysis and Future Trends. International Journal of Innovative Research in Advanced Engineering, 11(03), 138–146. https://doi.org/10.26562/ijirae.2024.v1103.01

Dmitrii Ustiugov, Theodor Amariucai, & Grot, B. (2021). Analyzing Tail Latency in Serverless Clouds with STeLLAR. Edinburgh Research Explorer (University of Edinburgh). https://doi.org/10.1109/iiswc53511.2021.00016

Wen, J., Chen, Z., Zhao, J., Sarro, F., Ping, H., Zhang, Y., Wang, S., & Liu, X. (2025). SCOPE: Performance Testing for Serverless Computing. ACM Transactions on Software Engineering and Methodology. https://doi.org/10.1145/3717609

Wu, Y.-C., Tuan, T., Hu, G., Zhang, M., Yeow Meng Chee, & Beng Chin Ooi. (2022). Serverless Data Science - Are We There Yet? A Case Study of Model Serving. https://doi.org/10.1145/3514221.3517905

Yu, M., Jiang, Z., Hok Chun Ng, Wang, W., Chen, R., & Li, B. (2021). Gillis: Serving Large Neural Networks in Serverless Functions with Automatic Model Partitioning. https://doi.org/10.1109/icdcs51616.2021.00022