

**CHRIST**(DEEMED TO BE UNIVERSITY)
BANGALORE | DELHI NCR | PUNE

Notice for the PhD Viva Voce Examination

Ms Josephine Eskaline Joyce J, Registration Number: 2071305, PhD Scholar at the Department of Computer Science, School of Sciences, CHRIST (Deemed to be University), will defend her PhD thesis at the public viva-voce examination on Wednesday, 09 September 2026 at 11.00 am in Room No. 044, Ground Floor, R&D Block, CHRIST (Deemed to be University), Bengaluru - 560029, Karnataka, India.

Title of the Thesis	:	Proactive Auto Scaling Using Custom Metrics for Large Language Model-Based Applications in a Kubernetes Environment
Discipline	:	Computer Science
External Examiner - I	:	Dr Modi Chirag N Associate Professor Department of Computer Science and Engineering National Institute of Technology Goa Cuncolim Municipal Area, Salcete Taluka Goa - 403703
External Examiner - II	:	Dr C Shoba Bindu Professor Department of Computer Science and Engineering JNTUA College of Engineering, Ananthapuramu Sir Mokshagundam Vishveshwariah Road, Upparapalli Ananthapuramu - 515002 Andhra Pradesh
Supervisor	:	Dr Shoney Sebastian Professor Department of Computer Science School of Sciences CHRIST (Deemed to be University) Bengaluru - 560029 Karnataka

The members of the Research Advisory Committee of the Scholar, the faculty members of the Department and the School, interested experts and research scholars of all the branches of research are cordially invited to attend this open viva-voce examination.

Registrar (Academics)

Place: Bengaluru
Date: 27 August 2026

ABSTRACT

The accelerated adoption of large language models (LLMs) on cloud-native platforms introduces new autoscaling challenges due to inference-heavy, latency-sensitive, and token-driven workloads. Traditional Kubernetes Horizontal Pod Autoscaler (HPA), based on CPU and memory metrics, fails to capture LLM runtime behavior, leading to slow scaling, cold starts, and degraded performance. This study proposes a proactive autoscaling framework using inference-aware custom metrics and machine learning-based workload prediction. It incorporates metrics such as inference time and tokens processed and evaluates a dual-metric HPA policy that outperforms single-metric approaches, reducing mean latency by up to 50%, peak latency by 18–20%, and improving scaling stability. To overcome reactive limitations, time-series models are used for workload forecasting, with LSTM and GRU outperforming ARIMA and Prophet in accuracy. System-level results demonstrate that prediction-driven autoscaling reduces cold starts by 80%, stabilizes latency, and enhances response time, enabling efficient, workload-aware resource management for latency-critical LLM applications.

Keywords: *LLMs, Kubernetes, HPA, Inference-aware scaling, Custom metrics, Predictive scaling, ML forecasting, Cloud-native AI systems*

Publications:

1. **J. E. Joyce** and S. Sebastian, "Inference-Time-Driven Autoscaling for Inference Workloads: A Comparative Study of Latency-Variant Models in Kubernetes", "Technologies" (Scopus Q1, Accepted)
2. **J. E. Joyce** and S. Sebastian, "Reinforcement Learning based Autoscaling for Kafka-centric Microservices in Kubernetes,"2022 IEEE 4th PhD Colloquium on Emerging Domain Innovation and Technology for Society (PhD EDITS), Bangalore, India, 2022, pp. 1-2.
3. **J. E. Joyce** and S. Sebastian, "Enhancing Kubernetes Auto-Scaling: Leveraging Metrics for Improved Workload Performance,"2023 Global Conference on Information Technologies and Communications (GCITC)", Bangalore, India, 2023, pp. 1-7, doi: 10.1109/GCITC60406.2023.10426170.
4. **Joyce, J.E.**, Sebastian, S. (2025). "Maximizing Efficiency: Unveiling the Potential of Kubernetes Metrics", "Fifth International Conference on Computing and Network Communications" CoCoNet 2023, "Lecture Notes in Electrical Engineering", vol 1219. Springer, Singapore. https://doi.org/10.1007/978-981-97-4540-1_3.
5. **J. E. Joyce**, A. K. Nagaraj, S. Soni, U. A. N. S. Shaik, V. Patil and A. Tripathi, "Autonomous Infrastructure Healing: A Multi-Agent Kubernetes Recovery Framework", "2025 Supercomputing India (SCI)", Bangalore, India, 2025, pp. 1-8, doi: 10.1109/SCI68648.2025.11333843.
6. **J. E. Joyce**, S. Maheshwari and D. Kothamasu, "Secure by Design: Strategic Approaches to Infrastructure as Code", "2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON)", BENGALURU, India, 2025, pp. 1-6, doi: 10.1109/NMITCON65824.2025.11187640.
7. **J. E. Joyce**, P. Bhat and S. Murali, "Designing a Multi-Layered Rate Limiting Framework for Resilient Cloud-Native Systems", "2025 Fourth International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)", Villupuram, India, 2025, pp. 1-6, doi: 10.1109/ICSTSN67075.2025.11397959.
8. **J. E. Joyce** and S. Maheshwari, "A Design-Driven Taxonomy of AI Agentic Patterns", "2025 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)", Bengaluru, India, 2025, pp. 1-6, doi: 10.1109/CONECCT65861.2025.11306519.