(12) PATENT APPLICATION PUBLICATION
(21) Application No.202541070309 A

(19) INDIA

(22) Date of filing of Application :23/07/2025
(43) Publication Date : 01/08/2025

(54) Title of the invention : "Next-Generation Cloud Architectures for Large-Scale Generative AI Systems"

(51) International classification	:G06F0009500000, G06F0021620000, G06N0003063000, G06N0003045000, G06F0009455000
(86) International Application No	:NA
Filing Date	:NA
(87) International Publication No	: NA
(61) Patent of Addition to Application Number	:NA
Filing Date	:NA
(62) Divisional to Application Number	:NA
Filing Date	:NA

(71)**Name of Applicant :**
1)MURUGALAKSHMI k
Address of Applicant :Assistant Professor, Department of Artificial Intelligence and Data Science, VSB College of Engineering Technical Campus, Kinathukadavu. ----------- -----------
2)Dr. R. Murugadoss
3)Mrs. D.S. Jayakumari
4)Ms. D.Jeevitha
5)Mrs. B. Jebaranjani
6)Mr. S. Soundhar
7)Mr. F. Theophilus
8)Mrs. G. Nithya
9)Ms. G. Yogapriya
Name of Applicant : NA
Address of Applicant : NA
(72)**Name of Inventor :**
1)MURUGALAKSHMI k
Address of Applicant :Assistant Professor, Department of Artificial Intelligence and Data Science, VSB College of Engineering Technical Campus, Kinathukadavu. ----------- -----------
2)Dr. R. Murugadoss
Address of Applicant :Professor, Department of AI & DS, V.S.B College of Engineering and Technical Campus, Coimbatore Coimbatore ----------- -----------
3)Mrs. D.S. Jayakumari
Address of Applicant :Assistant Professor, Department of AI & DS, V.S.B College of Engineering and Technical Campus, Coimbatore coimbatore ----------- -----------
4)Ms. D.Jeevitha
Address of Applicant :Assistant Professor, Department of AI & DS, V.S.B College of Engineering and Technical Campus, Coimbatore coimbatore ----------- -----------
5)Mrs. B. Jebaranjani
Address of Applicant :Assistant Professor, Department of AI & DS, V.S.B College of Engineering and Technical Campus, Coimbatore coimbatore ----------- -----------
6)Mr. S. Soundhar
Address of Applicant :Assistant Professor, Department of AI & DS, V.S.B College of Engineering and Technical Campus, Coimbatore coimbatore ----------- -----------
7)Mr. F. Theophilus
Address of Applicant :Assistant Professor, Department of AI & DS, V.S.B College of Engineering and Technical Campus, Coimbatore coimbatore ----------- -----------
8)Mrs. G. Nithya
Address of Applicant :Assistant Professor, Department of AI & DS, V.S.B College of Engineering and Technical Campus, Coimbatore coimbatore ----------- -----------
9)Ms. G. Yogapriya
Address of Applicant :Assistant Professor, Department of AI & DS, V.S.B College of Engineering and Technical Campus, Coimbatore coimbatore ----------- -----------

(57) Abstract :

Abstract The rapid evolution of generative AI systems—such as large language models (LLMs), diffusion models, and multi-modal transformers—demands a fundamental rethinking of cloud architectures to meet the scale, latency, security, and efficiency requirements of modern workloads. This paper explores next-generation cloud architectures designed to support large-scale generative AI systems. It presents a layered framework that integrates heterogeneous computing (GPUs, TPUs, FPGAs), high-throughput networking, advanced memory hierarchies, and disaggregated storage. The architecture leverages innovations in distributed training, model parallelism, and elastic inference serving, orchestrated via AI-native platforms such as Kubernetes with custom scheduling and autoscaling policies. Moreover, it addresses critical challenges in energy efficiency, cost optimization, and compliance with data privacy regulations through federated and edge-cloud hybrid deployments. We also examine the role of AI accelerators, service meshes, serverless AI pipelines, and zero-trust security models in enabling reliable, real-time AI model delivery. The proposed architectures aim to provide scalable, resilient, and sustainable infrastructure foundations for the future of generative AI in industries ranging from healthcare and finance to autonomous systems and creative content generation.

No. of Pages : 9 No. of Claims : 10