

Bireswar Kundu

Software Engineer

Hardworking, Honest & Innovative Software Engineer with 6 years of experience in designing, developing & deploying enterprise-grade applications. Specializes in bringing Generative AI models to production, ensuring scalability and efficiency with a strong track record of delivering high-impact solutions.

EXPERIENCE

SLB, Pune — Software Engineer

AUGUST 2024 – Current

- ❖ **Gen AI based Solution for Domain Specific Highly Scalable Translation Module**

- ❖ **Environment:** Python, Embedding, Huggingface, Transformers, Azure DevOps, Azure OpenAI, PostgreSQL
- ❖ **Role:** Design & Development

Collaborated with multiple teams to analyze the existing architecture and limitations of the current application infrastructure, leading the development of a large-scale translation feature module in multiple languages for standard work documents by leveraging GenAI offerings from Azure OpenAI. Spearheaded the POC by significantly contributing to architectural design, development, cost estimation, and final delivery.

- ❖ **Enterprise Web App for Obsolescence Management of Electronic Components**

- ❖ **Environment:** Django, Angular, Material UI, Linux, Azure Web Apps, Azure DevOps, PostgreSQL
- ❖ **Role:** Design & Development

Responsible for setting the layout of the project's code structure, creating and managing the repository (branching strategy), DevOps pipelines, deployment, development of the UI and assisting team in development of backend functionalities, like IAM, user management, logging, schedulers, etc. Significantly contributed through the means of code reviews and guidance/assists, actively talking part in knowledge & skill development of myself & team.

- ❖ **Optimization of Enterprise Web App for RFQ Generation**

- ❖ **Environment:** Django, Angular, Material UI, Linux, Azure Cloud, Azure DevOps, PostgreSQL
- ❖ **Role:** Development & Optimization

Responsible for refactoring the existing Django API server, making it robust, scalable, and high-performance version, reducing the execution time by more than 50%. Successfully took over the DevOps (CI/CD - pipeline) responsibility of the project. Contributed to the development of a robust Angular-based front-end converting Figma designs to responsive modern UI Components, by leveraging advanced concepts like observers, interceptors, guards, lazy-loading, etc.

Mobile: **+91 7584075066**

Email: biru1997.bk@gmail.com

LinkedIn:

<https://linkedin.com/in/bireswar-kundu-1643b0160>

SKILLS

Languages: Python, C, JavaScript, Typescript

SDLC: Agile

ML Ops: Azure OpenAI, Google Vertex AI, Transformers, vLLM, ROCm, Huggingface, Langchain

Databases: MySQL, PostgreSQL, CouchDB, InfluxDB, Vector DB

Cloud: Microsoft Azure Cloud, Oracle Cloud, IBM Cloud

IDEs: Pycharm, VS Code, Jupyter

Version Control: GIT, Azure Repos

Frameworks: Node.js, Flask, Django, Angular

DevOps Tools: GitHub Actions, Azure DevOps

IAM: Okta, OAuth, Open ID, SAML, SCIM

Other Tools: Docker, Kubernetes, Dataiku, Graph QL, React.js, Elastic Stack, SOAP API

OPERATING SYSTEM

Windows, Linux

CERTIFICATIONS

Azure Fundamentals

LANGUAGES

English, Hindi

AWARDS

Next 5% Executive Award in AMD'23

Advanced Micro Devices (AMD), Bengaluru — Systems Engineer 2

JUNE 2023 – AUGUST 2024

- ❖ **LLM Performance Benchmarking, Finetuning & Quantization**
- ❖ **Environment:** Python, Pytorch, ROCm, llama.cpp, vLLM, Huggingface
- ❖ **Role:** Research & Development

Contributed to the development of scripts for benchmarking LLMs & creating system requirements for the LLM performance suite. Briefly worked on LLM finetuning to make sure the inferences were useful to the specific areas. Also contributed to benchmark quantized models and different quantization processes of LLM.

- ❖ **Metal to Model (Platform for on-demand AI System Provisioning and LLM hosting)**
- ❖ **Environment:** Python, Oracle Cloud, ROCm, vLLM, Harbor, Rancher, Kubernetes, Pytorch, Transformers, Llamafire, Nginx Ingress, Linux
- ❖ **Role:** Research & Development

Developed generic templates of LLM deployment for inference in an on-premise Kubernetes environment, complete automation from downloading, loading to hosting. Deployed nginx ingress controller with TLS for routing & load balancing through single domain name.

- ❖ **GenAI Initiatives (Internal AI Chatbot, Code Searching, Code Explain & Code Completion)**
- ❖ **Environment:** Python, Flask, React JS, Okta, Langchain, Fast API, Azure Open AI Services, Google Cloud, Huggingface Models, Vector DB, Linux
- ❖ **Role:** Design, Development, Test & Integration, Deployment

Internal Chatbot web application with image gen capabilities, based on different LLM model offerings from cloud services and hosting LLM models on-premise from hugging face org. Responsible for creating & managing the lifecycle of API gateways and UI lifecycle, crunching source codes, and leveraging prompt engineering to achieve Gen AI initiative goals.

- ❖ **ML in Predictive Failure Analysis of Integrated Circuits**
- ❖ **Environment:** Python, Dataiku, Jupyter Notebook
- ❖ **Role:** Research & Development

A system to streamline the identification of processor die based on FAB manufacturing and testing data.

Responsible for feature engineering with available data, developing anomaly detection techniques & testing different combinations of data used as features for training prediction and anomaly detection models.

EDUCATION

National Institute of Technology,
Jamshedpur -

M.tech 2021 - 2023 CGPA - 9.2

Kalinga Institute of Industrial Technology,
Bhubaneswar —

B.Tech 2015 – 2019 CGPA - 8.55

Banerjee Danga High School, Garbeta —
WBCHSE

2013 – 2015 Percentage - 81%

Banerjee Danga High School, Garbeta —
WBBSE

2013 Percentage - 90%

PROJECTS

Novel Approach to secret detection in Enterprises using Machine Learning Algorithms — Google COLab Project for Final Year Thesis in M.Tech

Serverless Computing & OTA Firmware Update to Edge Devices — Using the serverless architecture of cloud environments to bring efficient approaches to OTA firmware updates of edge devices

Advanced Micro Devices (AMD), Hyderabad — *Co-Op/Intern*

JUNE 2022 - June 2023

- ❖ **DevSecOps (Secret Scanning in Application Security)**
- ❖ **Environment:** Python, SAST & SCA tools, GitHub, Azure Functions, PostgreSQL, Azure DevOps, Kubernetes, Elastic Stack
- ❖ **Role:** Design, Development, Test & Integration

A complete secret scanning solution for DevOps pipelines in source code versioning and management systems like GitHub, GitLab, BitBucket, etc.

The secret scanning tool is designed & developed with the capability to scan throughout the whole commit history of git repositories along with the feature to perform continuous-scan through CI/CD pipelines (for new source codes). Governance like management and tracking is implemented using PostgreSQL database and automated notification processes.

- ❖ **Automation (Cisco Jabber Automation & VPN Monitoring Automation)**
- ❖ **Environment:** Python, Cisco Unified Communication Manager, GitHub, InfluxDB, GraphQL, Service Now, Azure Cloud
- ❖ **Role:** Development, Test & Integration

Automation to record different VPN metrics like ping, latency, etc., and analytics-based dashboard creation with automated notification generation on various criteria in metric activity.

Cisco Jabber phone & physical phone provisioning-de-provisioning automation based on service now ticket for user's organization status.

GS Lab, Pune — *Software Engineer*

AUGUST 2019 - DECEMBER 2021

- ❖ **SIEM (Identity & Access Management)**
- ❖ **Environment:** Python, OAuth, Open ID, SAML, SCIM
- ❖ **Role:** Microservice Development, Daily Ops, & Integration

IAM domain of Security Information and Event Management (SIEM) deals with managing the identity of consumers & their access to different technology resources. Actively contributed to integrating IAM protocols with client applications to embed applications with the client's IAM provider Instance.

- ❖ **Internet of Things**
- ❖ **Environment:** Python, Docker, IBM Cloud, CouchDB, Mosquitto MQTT Broker, Postman, Git
- ❖ **Role:** Design, Development & Integration

Gateway-level multi-threaded application that acts as a bridge between cloud and end devices to connect and control them remotely via the MQTT protocol. Implemented for industrial protocols like Modbus, Canbus, OPC, and RFID devices.

GS Lab, Pune — *Intern*

SEPTEMBER 2018 - JULY 2019

- ❖ **SIEM (Identity & Access Management)**
- ❖ **Environment:** Node.js, Express.js, React.js, IBM Watson Services, IBM Cloudant NoSQL
- ❖ **Role:** Learning & Development

Automate health monitoring of patients. Ease communication between doctors and patients, analyze patient records, detect symptoms, and provide possible cures. Track the patient's heart rate and automate calling the nearest ambulance. A vision to get healthcare to the patient, not the other way around. Inefficiency in the public healthcare system is one of the biggest barriers and reasons for the unavailability of essential life-saving services