HRBR Layout, Bangalore,
Karnataka, India

ALLEN PHILIP JOHN

+91-8886778140
allenphilip93@gmail.com

SKILLS

- Python | FastAPI | PyTorch | Torch Vision | Torch Audio | TensorRT | ModeOpt | flashattn | xformers | cudnn | triton | Quantization | AOT | OpenCV | ffmpeg | scenedetect | Scikit-learn | Transformers | LDM | VAE | TensorParallelism | FSDP2 | TorchCompile | Torch Lightning
- Adobe Firefly Video | Adobe Firefly Audio | T5 | Upsampler | SSR | Whisper | Azure TTS | Adobe SoundLift | Adobe AutoDub | Adobe's Lipsync
- Pydantic | Kubernetes | Argo | Airflow | Databricks | MLFlow | Timestream | Grafana | Prometheus | Sagemaker | Storm | Zookeeper | S3
- Java | Spring | Splunk | NewRelic | Elasticsearch | Redis | Kafka | MongoDB | SQL | GraphQL | AWS AppSync | Lambda | Terraform | Athena
- Microservices | Distributed Systems | ETL | ML Engineering | Backend | Git | VS Code | Language English – *Professional proficiency*

EMPLOYMENT

Machine Learning Engineer 4 (P4) **Adobe, Bangalore, IN** **Dec 2023 - Present**

Firefly Platform Integration for Video GenAI Models, Core Tech & Products

- Adobe Firefly Video (Text/Image to Video) in Firefly.com
  - Profiled the video diffusion model for identifying possible bottlenecks to improve latency and reduce infra costs
  - Worked on post training FP8 quantization of the model and reduced latency by 35% with minimal quality loss
  - Experimented with attention optimization techniques to reduce latency and improve GPU utilization
  - Wrapped our custom attention implementation using TRT plugins and reduced latency by 25%

- Adobe Audio Generative Extend in Premier Pro
  - Designed and led the team for productionizing the inference pipeline for Firefly Audio Generative Extend in PPro
  - Closely worked with research and services team to improve GPU utilization and providing a scalable infra

- Enhance Speech V2 in Adobe Podcast
  - Designed and led the team for productionizing the inference pipeline to support upto 150 RPM with over 50 A100 GPUs in the cluster
  - Extensively profiled the GPU, CPU, IO utilization to ensure that the infra costs and inference time are optimized
  - Execute multiple improvements to add parallelism, increase TFLOPS, keep IO latencies to the minimum

- Miscellaneous
  - Optimized inference pipeline for Adobe LipSync model to sync lips in video to AI generated dubbed audio.
  - Reduced the cost of operations by 33% using GPU optimization techniques and reduced average latency by 40%.
  - Involved in profiling multiple speech and vision models including Whisper & other models in Adobe.
  - Executed load tests, integration tests, scaling, costing & monitoring strategies for production with the SRE team.

Computer Scientist 2 (P4) **Adobe, Bangalore, IN** **May 2022 – Nov 2023**

Commerce Data Platform, Adobe Business Platform (ABP)

- Enabled near real-time tracking of Adobe commerce workflows using an event streaming platform consuming async events peaking at times over 100,000 RPM.
- Designed a cost-effective data warehousing solution for real-time as well as long-term analytics using a combination of AWS Opensearch, S3, Athena and AWS Glue ETL.
- Engaged with multiple teams across ABP to build requirements and established an audience for the platform.

Anomaly Detection ML Platform, Adobe Business Platform (ABP)

- Conceptualized and productionized a ML platform to intelligently detect anomalies in key commerce KPIs derived from customer engagement data using Spark, Airflow & Sagemaker.
- Collaborated with multiple teams, PdMs and analysts to build requirements and to operationalize the ML platform
- Currently monitoring for over 1000 timeseries metrics through a self-service UI for analysts to self-onboard KPIs, tune the ML models & configure alerting themselves.

Computer Scientist (P3) **Adobe, Bangalore, IN** **May 2021 – May 2022**

AppStore Integration Service (AIS), Adobe Business Platform (ABP)

- Enriched the real-time monitoring of the AIS services which process over 2 million subscription requests every day.

- Re-designed the subscription renewal batch job to make it horizontally scalable.

Senior ML Engineer (SSE) **Blue Yonder, Hyderabad, IN** **Dec 2016 – May 2021**

Promotional Demand Modelling Pipeline

- Led the technical design and implementation of batch processing ML pipelines for promotional demand modelling using GLM with Fourier and Xgboost models for brick-and-mortar retailers like Walmart, Loblaws, Woolworths
- Collaborated with cross-functional teams for the deployment & performance tuning to fit the jobs within SLAs for customers with over 90 million SKUs in production

Data Scientist, Intern **Predible Health, Bangalore, IN** **June 2016 – Oct 2016**

- Implemented a multi-context deep learning algorithm for precise segmentation of digital pathological slides of ovaries for abnormality detection as a consultancy project for Sunpharma Pvt Ltd

Software Engineer **Oracle Pvt Ltd, Hyderabad, IN** **June 2014 – June 2016**

Oracle Cloud Commerce

- Worked on full stack development projects implementing core product features, improving code coverage and drove the PoC of designing B2B user management, multiple catalog, and price list management.

EDUCATION

Tamil Nadu, IN **Indian Institute of Technology, Madras** **August 2010 – April 2014**

- B.Tech in Electrical Engineering
- Graduation project in “Learning Theory using Fourier Analysis of Boolean Functions”
- Relevant Coursework: Data Structures & Algorithms | Machine Learning | Linear Algebra | Calculus

AWARDS & ACCOMPLISHMENTS

- Recipient of the Best Rookie Award in Oracle (2015)
- Recipient of multiple Spot Awards in Blue Yonder (2018-2021)
- Among the 7 recipients of the prestigious ACE award across all BlueYonder in H1-2020
- Recipient of the Kudos Award in Adobe (2022, 2024)