Puneeth N Naik

puneethnaik60@gmail.com | +916362960529
Github:**Puneethnaik**, **puntertram**

EXPERIENCE

SALESFORCE | Member of Technical Staff

Aug 2023 - Present | Hyderabad, India

- **Splunk on Alibaba/Kubernetes**: Enabled production rollout with zero regressions, **establishing** the standard onboarding model for new environments.
- **Operator Debugging and Stabilization**: **Resolved** critical Splunk operator failures blocking production and **mentored peers** on debugging techniques, improving team incident resolution efficiency.
- **Patch Pipelines and DIY Releases**: **Built** and onboarded pipelines to DIY model, reducing release times by **95%** (days -> hours) and **standardizing processes** across teams.
- **Puppet Upgrade and Runtime Migration**: **Upgraded** Puppet v5->v7 and migrated workloads to RHEL9 across Splunk-on-EC2 (**1500+ hosts**) and Splunk-on-Kubernetes
- **Auto Index Management**: **Automated** Splunk index lifecycle for new Salesforce cells, cutting provisioning time by **95% (days -> hours)**
- **Log Flow Visualization**: **Architected, implemented** end-to-end log flow mapping from source to the sink, reducing operational toil by **83%**, accelerating on-call troubleshooting.
- **Long-Term Retention**: **end-to-end ownership** of S3 freeze component and **implemented** STS token auto-refresh, eliminating manual intervention and enhancing system reliability and uptime.

SALESFORCE | Intern

May 2022 – Jul 2022 (FutureForce) | Hyderabad, India

- **POC for EC2 -> EKS Migration**: **Built** a near production-ready POC to migrate log storage from EC2 to EKS, benchmarking performance under simulated loads. **Reduced** patching and Day-2 ops by **90%**
- **Enhanced** service reliability by defining SLOs with full observability, **authoring** runbooks to reduce incident resolution time, and expanding deployments from single-AZ to multi-AZ
- **Splunk Operator Enhancements**: **Contributed** to Splunk operator by enabling additional metrics export and supporting custom node affinity policies, easing operations.

PROJECTS

EFFICIENT TRANSFORMER INFERENCE | Thesis Project

- Characterized LLM inference workloads for **summarization** and **machine translation**.
- Implemented **GPU kernels** for LogitProcess and BeamSearchProcess, achieving up to **32% speedup** (indicBART) and **19%** (mBART).
- Reduced device-to-host transfers by **66–99%** through multi-threaded GPU optimization.
- Applied **DVFS-based energy optimization**, lowering energy consumption by **15%** with minimal latency impact.
- **Maintained** one-shot deployment of experiments with fully automated scripts, optimized CUDA kernels, and SOPs for execution on HPC clusters at USC, IISc.

EDUCATION

INDIAN INSTITUTE OF SCIENCE

MTech - Computer Science And Automation
07 2021 - 07 2023
GPA: 7.80 / 10.0

RAMAIAH INSTITUTE OF TECHNOLOGY

BE - Computer Science And Engineering
07 2015 - 07 2019
GPA: 9.19 / 10.0

SKILLS

Programming: Go, C/C++, Python
Machine Learning / AI: PyTorch, Huggingface Transformers, Energy-aware LLM inference
DevOps / Infrastructure: Docker, Kubernetes, Slurm, GPUs, CUDA, Linux, Linux Kernel internals
Tools / Build: Git, Jenkins, CMake, Makefile, Pybind

OPEN SOURCE

Xterm.js (TypeScript Terminal Emulator, *19k)

- Created API to turn off scrollbar.
- Implemented escape-sequence support to disable scrollbar and clear scroll buffer.
- Developed API to make opaque selection color translucent, improving text readability.

AwaitWhat (Asyncio Dependency Visualizer, *40)

- Built API to display the task awaited by `asyncio.wait` and remaining timeout.