PUNEETH N NAIK
📍 Hyderabad, India | ✉️ [puneethnaik60@gmail.com] | 🔗 Github

EDUCATION

Indian Institute of Science, Bangalore (IISc) — *M.Tech, Computer Science & Automation*
2021–2023 | GPA: 7.8/10

- **Thesis:** *Workload Characterization of Transformer Text Generation Inference*
    - Optimized inference for IndicBART (summarization) & mBART (translation) by offloading beam search & logit processing to GPUs with **custom CUDA kernels**.
    - Achieved **32.4% speedup (IndicBART)** and **19% speedup (mBART)** over baseline.
    - Reduced device-to-host transfers by **66.8% (IndicBART)** and **99.1% (mBART)**.
    - Explored **DVFS-based energy optimization**: achieved **15% lower energy** at only 5% latency overhead.

TECHNICAL SKILLS

- **Deep Learning:** PyTorch, TensorFlow, HuggingFace Transformers
- **Inference Optimization:** CUDA, GPU programming, multi-threaded CPU optimization, DVFS
- **Quantization & Deployment:** (add ONNX Runtime, TensorRT if you do a project)
- **Languages:** Python, C++, CUDA, Go, Java
- **Tools:** Docker, Kubernetes, Git, Linux Kernel internals

RELEVANT PROJECTS

Performance Optimization of Transformer Inference — *CUDA, PyTorch*

- Implemented GPU-accelerated **beam search & logit processing** kernels.
- Benchmarked across summarization & MT workloads → **up to 32% throughput gain**.
- Evaluated system-level tradeoffs with **DVFS energy-performance tuning**.

Checkered Matrix Multiplication Optimization — *C++, CUDA*

- Achieved **190% IPC gain (single-thread)** and **289% (multi-thread)** over baseline.
- CUDA implementation achieved **GPU IPC 432.9**.

Directory Cache Coherence Simulator — *Python*

- Implemented MSI directory-based coherence protocol from scratch.

IsoSurface Visualizer — *C++, OpenGL*

- Implemented Phong shading for realistic isosurface rendering.

EXPERIENCE

Salesforce — Software Engineer *(2023–Present)*

- Designed & built distributed logging/indexing systems (**Go, Java, Lucene**).
- Hackathon-winning project: leveraged **control flow graph analysis** to predict log volume growth.
- Built **Tatzelwurm DFS** (inspired by GFS): supports async replication, WAL-based recovery, auto-chunkserver detection.
- Developed **Seshat** indexing layer on DFS with Lucene integration.

Salesforce FutureForce Intern (2022)

- Led POC for migration from **EC2 → Kubernetes (EKS)**, reducing ops overhead by 90%.
- Implemented **service observability (SLOs, dashboards, alerts)** across multi-AZ deployments.
- Extended Splunk operator with custom metrics + node affinity features.

OPEN SOURCE CONTRIBUTIONS

- **Xterm.js** (12.2k ★): APIs for scrollbar control, selection color customization.
- **AwaitWhat** (40 ★): Python async visualization tool — added APIs for task tracing.
