Tony Lee

ML Inference Performance Engineering • GPU Serving Systems • CUDA/Triton Optimization

tungsheng@gmail.com
linkedin.com/in/tonyslee8

PROFESSIONAL SUMMARY

ML inference performance engineer optimizing GPU-backed LLM serving from CUDA/Triton kernels to vLLM runtime behavior and Kubernetes capacity controls, with recent work on KV-cache limits, latency/cost tradeoffs, autoscaling/admission control, H200 Tensor Core matmul tuning, RMSNorm bandwidth profiling, and decode-step replay.

SKILLS

ML Infrastructure / Inference
vLLM • LLM Serving • OpenAI-Compatible APIs • Autoscaling • Admission Control • KV Cache / Context Length • Quantization Evaluation • CUDA / Triton Benchmarking • Kernel Fusion • GPU Profiling • GPU Capacity Planning • GPU Scheduling (NVIDIA Device Plugin) • Inference Load Testing (k6)

Leadership
Architecture Design • Cross-Functional Leadership • Technical Mentorship

Infrastructure / Cloud
Terraform • Kubernetes (EKS) • AWS (VPC, IAM, EC2, ALB) • Karpenter • CI/CD Automation • Prometheus / Grafana • DCGM Exporter

Languages
Go • Python • TypeScript • SQL

Backend
Node.js • GraphQL • REST

EDUCATION

M. E. Electrical Engineering
University of California, San Diego
Sep 2006 - Jun 2008

B. S. Electrical Engineering
University of Illinois at Urbana-Champaign
Sep 2003 - Dec 2005

SELECTED PROJECTS

GPU Inference Decision Lab (AWS EKS + Karpenter + vLLM)

- Built an AWS EKS/vLLM inference performance lab that turns KV-cache/context-length behavior, autoscaling, admission control, scheduler/quantization probes, useful-work cost, and failure-mitigation drills into supported, rejected, or pending architecture evidence.
- Measured burst and 8192/300 long-context behavior: bounded queues preserved 100% delivery near 2s p95 for bursts, 1.20 req/s long-context traffic repeated 62.66-63.40s p95 latency, and FP8 KV was rejected after delivery fell to 47.58-69.12%.

CUDA/Triton GPU Kernel Lab (PyTorch baselines + A10G/H200)

- Built a reproducible CUDA/Triton benchmarking lab for LLM-shaped GPU primitives, comparing custom Triton kernels against PyTorch/cuBLAS baselines with correctness checks, latency percentiles, bandwidth, TFLOP/s, roofline analysis, and Nsight evidence.
- Separated supported wins from caveats: RMSNorm fp16 reached 5.901x with 90.91% DRAM throughput, while same-stream dynamic decode replay reached about 0.156 ms p50 / 0.230 ms p95 as a synthetic resident-KV upper bound.
- Extended the H200 matmul track with standard and persistent-wave Triton schedules; the focused 512x11008x4096 bf16 standard row reached 471.4 TFLOP/s (89.41% of PyTorch/cuBLAS), while persistent waves remained below the standard schedule.

WORK EXPERIENCE

Staff Software Engineer • DTEX Systems
Sep 2024 - Present

- Designed distributed platform services across UI, API, data, and infrastructure boundaries, with emphasis on reliability, deployment safety, and cross-service coordination.
- Improved release stability by containerizing UI services, decoupling shared dependencies, and isolating high-risk deployment surfaces.
- Led a launch-critical OpenSearch Dashboards migration, resolving deployment and environment issues with AWS/OpenSearch engineers.
- Built local development and validation automation that cut dev/test iteration latency by 20% and improved onboarding.

Senior Technical Lead • Cisco Systems, Inc.
Aug 2019 - Apr 2024

- Led architecture for the Onboarding Experience platform, helping enterprise customers complete onboarding in under 30 minutes.
- Drove system design across frontend, backend, APIs, and integrations, resolving tradeoffs across US, Europe, and India engineering groups.
- Led AngularJS-to-React modernization to remove security risk and improve maintainability; introduced Cypress end-to-end testing that cut test creation time by 50%.
- Built CI/CD and test automation improvements that reduced regression risk in high-visibility customer onboarding flows.

Senior Software Engineer • Tico Co., Ltd.
Jan 2019 - Aug 2019

- Improved message loading performance by 60% through frontend and API optimization and reduced codebase size by 30% through modular refactoring.

Co-founder / CTO • Popup Technology Co., Ltd.
Apr 2016 - Oct 2018

- Built the core platform from scratch, led a 5-person engineering team, and shipped features that doubled revenue within 12 months.