Contact Information

Name: John Doe
Email: john.doe@example.com
Phone: +1 (555) 123-4567
LinkedIn: linkedin.com/in/johndoeaiinfra
GitHub: github.com/johndoe-ai
Location: San Jose, CA

Education

Master of Science in Computer Science 2012
Stanford University Stanford, CA
GPA: 3.9/4.0

Bachelor of Engineering in Computer Engineering 2010
University of California, Berkeley Berkeley, CA
GPA: 3.8/4.0

Professional Experience

Lead AI Infrastructure Engineer Jan 2019 – Present
Tech Innovations Inc. San Francisco, CA

- Spearheaded the design and implementation of **next-generation AI infrastructure** for high-performance machine learning workloads, utilizing **C++** and **Golang** for core components.
- Developed and optimized **Kubernetes Operators** and **Custom Resource Definitions (CRDs)** to automate the deployment and lifecycle management of **AI infrastructure** on large-scale **Kubernetes clusters**, improving deployment efficiency by **40%**.
- Engineered **eBPF**-based **telemetry collection systems** for **Linux** nodes and **Nvidia GPUs**, providing real-time performance insights and reducing monitoring overhead by **15%** across **1000+ GPU-enabled servers**.
- Implemented advanced **GPU programming** and **memory management** strategies for **CUDA kernels** using **Nvidia MIG** and **Nvidia MPS** concepts, achieving **30%** throughput increase for critical **AI workloads**.

Senior Systems Software Engineer Aug 2012 – Dec 2018
Global Cloud Solutions Seattle, WA

- Designed and developed **distributed system fundamentals** for a cloud-native platform using **C++** and **Python**, ensuring **scalability**, **resilience**, and **reliability** for critical services.
- Optimized data transfer paths by integrating **RDMA** and **UCX** for high-speed communication between compute nodes, resulting in a **20%** reduction in inter-node latency for data-intensive applications.
- Contributed to **Linux kernel development** and wrote **device drivers** for custom hardware accelerators, enhancing system performance and enabling new capabilities for virtualized environments.
- Managed **Linux user space development**, including robust **software packaging**, **system logging**, and **lifecycle management of processes** for core infrastructure components, improving system stability by **25%**.

Projects

AI GPU Orchestration Framework 2023

Developed an open-source framework for dynamic GPU resource allocation in AI clusters.

- Designed and implemented a **Kubernetes Operator** in **Golang** to manage **Nvidia MIG** and **Nvidia MPS** configurations dynamically, improving GPU utilization by **35%** for mixed **AI workloads**.
- Integrated with **Nvidia GPU operators** and **Nvidia container toolkit** to provide seamless GPU access and optimized runtime for **Docker**-based **AI/ML** containers.
- Utilized **CUPTI** and **Nsight** for detailed **performance analysis** and **optimization** of **CUDA kernels**, reducing execution time by **18%** on various deep learning models.

High-Performance Network Fabric for AI 2021

Engineered a low-latency, high-throughput network fabric for AI training clusters.

- Developed a custom **Linux kernel module** and **device driver** in **C** to enable direct memory access for **RDMA** over Ethernet, achieving **95Gbps** throughput.
- Integrated **UCX** to abstract communication protocols, providing a unified interface for **GPU-to-GPU** and **CPU-to-GPU** data transfers, crucial for distributed **Artificial Intelligence** training.
- Conducted extensive **performance benchmarking** and **optimization** using custom tools, demonstrating **2x** speedup for collective communication operations compared to standard TCP/IP.

Technical Skills

Programming Languages: C/C++, **Golang**, **Python**
AI/ML Infrastructure: Artificial Intelligence (AI), **Machine Learning (ML)**, **GPU Programming**, **CUDA** (kernels, general), **UCX**, **RDMA**, **Nvidia GPU operators**, **Nvidia container toolkit**, **Nsight**, **CUPTI**, **Nvidia MIG** concepts, **Nvidia MPS** concepts, **AI Workloads**, **Next-Generation AI**
Operating Systems & Kernel: Linux (user space, kernel-level components), **Linux kernel development/expertise**, **Device driver development/expertise**, **Linux user space development**, **eBPF**
Containerization & Orchestration: Kubernetes (K8s), **Docker**, **Custom Resource Definitions (CRDs)**, **Kubernetes Operators**
Distributed Systems & Networking: Distributed system fundamentals, **High-speed data transfer technologies**
Performance & Optimization: Performance benchmarking, **Performance analysis**, **Performance optimization** (AI infrastructure, CUDA kernels, memory management for GPUs), **Memory management** (for GPUs), **Efficiency**, **High-Performance**
System Operations & Management: Software packaging, **System logging**, **System telemetry**, **Lifecycle management of processes**, **Telemetry collection systems**, **Software component configuration (config)**, **Software upgrade architecture** (seamless, to minimize downtime), **Software installation**, **Software deployment** (AI infrastructure on Kubernetes clusters), **System-level issues** (debugging, problem-solving), **Reliability**, **Scalability**, **Resilience**
Tools & Methodologies: Problem-solving, **Debugging**, **Collaboration**, **Innovation**, **Agility**, Fast-paced environments, Experimentation-rich environments