

James Wilson

Senior AI Systems Architect | LLM Infrastructure Expert

Seattle, WA | (206) 555-7821 | james.wilson@example.com

linkedin.com/in/jameswilson-ai | github.com/jwilson-llm-infra

Professional Summary

Strategic AI Systems Architect with 12+ years of experience designing and implementing infrastructure for large-scale AI systems. Specialized in high-performance computing, model serving architectures, and cost-efficient training strategies. Proven track record of building systems supporting trillion-parameter models while optimizing for performance and resource utilization.

Professional Experience

Principal AI Infrastructure Architect

MegaScale AI, Seattle, WA (Apr 2018 - Present)

- Designed distributed training architecture supporting models exceeding 1 trillion parameters. - Reduced training costs by 45% through custom sharding strategies and optimizer improvements. - Architected inference serving platform handling 50,000+ requests per second with 99.9% availability. - Led team of 15 engineers across infrastructure, performance optimization, and reliability workstreams.

Lead ML Infrastructure Engineer

CloudCompute Inc., San Francisco, CA (Feb 2014 - Mar 2018)

- Built cloud-native infrastructure for training and serving ML models at scale. - Implemented custom CUDA kernels improving compute efficiency by 30%. - Developed automated deployment and monitoring systems for ML infrastructure. - Created cost modeling tools for accurately forecasting compute requirements.

High Performance Computing Engineer

SuperComputing Technologies, Austin, TX (Jul 2010 - Jan 2014)

- Designed and implemented distributed computing systems for scientific applications. - Optimized memory hierarchies and communication patterns for parallel computing. - Developed custom MPI implementations for specialized hardware. - Created performance modeling tools for large-scale distributed systems.

Technical Skills

- **Systems Design:** Distributed Systems, Fault Tolerance, High Availability Architecture
 - **Programming Languages:** C++, CUDA, Python, Go, Rust
 - **ML Infrastructure:** PyTorch FSDP, DeepSpeed ZeRO, Megatron-LM, Alpa
 - **Cluster Management:** Kubernetes, Slurm, Ray, Dask
 - **Cloud Platforms:** AWS (EC2, EKS, S3), GCP (GKE, TPU), Azure
 - **Performance Optimization:** Mixed Precision, Kernel Fusion, Memory Optimization
 - **Serving Systems:** TorchServe, Triton, vLLM, Ray Serve
 - **Monitoring & Reliability:** Prometheus, Grafana, Datadog, SLO/SLI frameworks
-

Education

PhD, Computer Engineering

Georgia Institute of Technology, Atlanta, GA (Graduated: Jun 2010)

- Dissertation: "Scalable Architectures for Distributed Computing" - Research focus on high-performance computing systems

Master of Science, Electrical Engineering

University of Texas, Austin, TX (Graduated: May 2006)

- Specialized in computer architecture and systems design

Bachelor of Science, Computer Engineering

Purdue University, West Lafayette, IN (Graduated: Jun 2004)

- Minor in Mathematics - Graduated with Highest Distinction

Patents & Technical Publications

- US Patent 11,856,972: “System and Method for Efficient Model Parallelism in Neural Networks”
 - US Patent 11,542,391: “Architecture for Distributed Training of Large Models”
 - Wilson, J., et al. (2022). “Cost-Efficient Training Strategies for Trillion-Parameter Models.” *MLSys Conference*.
 - Wilson, J., et al. (2020). “Optimizing Memory Hierarchies for Distributed Training.” *SC Conference*.
 - Wilson, J., et al. (2018). “Fault Tolerance in Large-Scale Model Training.” *NeurIPS Systems Workshop*.
-

Industry Leadership

- Technical Advisory Board, ML Systems Consortium (2020-present)
 - Keynote Speaker, SuperComputing Conference 2022: “The Path to Exascale AI”
 - Committee Member, MLSys Conference (2019-present)
-

Professional Certifications

- Google Cloud Certified Professional Cloud Architect (2021)
 - AWS Certified Solutions Architect - Professional (2020)
 - NVIDIA DLI Certified Instructor - Accelerated Computing (2019)
-

Languages: English (native), German (proficient)