Ani Balasubramaniam

ani@anionline.me

+1 (206) 651-4761

github.com/aniongithub

linkedin.com/in/anibalasubramaniam

6168 158th Ave SE Bellevue WA 98006

www.anionline.me

RESEARCH & TECHNICAL LEADERSHIP PROFILE

Senior technical leader and researcher with **25+ years of experience** designing and scaling **AI infrastructure, cloud systems, simulation platforms, and GPU-accelerated architectures** across industry and applied research environments. Proven track record of **bridging AI research with deployable hardware–software systems**, spanning cloud infrastructure, real-time systems, heterogeneous compute, and AI-driven simulation.

Expert in **co-designing AI workloads with infrastructure constraints**, optimizing for **performance, cost, reliability, and scale**. Experienced in leading **ambitious research agendas**, mentoring senior engineers and researchers, and transitioning **research prototypes into production-grade systems** deployed at global scale.

CORE RESEARCH & TECHNICAL AREAS

- AI Infrastructure Architecture (Cloud, Distributed Systems, Edge)
- Hardware–Software Co-Design for heterogeneous AI Workloads
- GPU / Heterogeneous Compute (CUDA, OpenCL, SIMD, GPGPU)
- AI-Driven Simulation, Differentiable Systems, Digital Twins
- Real-Time Systems, Low-Latency Streaming, Robotics & Autonomy
- Cloud-Native Systems (Kubernetes, OpenStack, GCP, Azure, AWS)
- Performance Modeling, Cost Optimization, Reliability Engineering

---

PROFESSIONAL EXPERIENCE

Senior Director of AI — Nucleus (Fibernetics subsidiary)

Apr 2025 – Present

- Leading R&D for **speech-driven AI agents**, designing **distributed cloud and AI infrastructure** optimized for real-time interaction.
- Defined infrastructure strategy across **training, inference, scaling, and observability**, with strong emphasis on cost-performance tradeoffs.
- Collaborated with product, research, and engineering teams to align **AI system capabilities with infrastructure constraints**.

Founder & Chief Scientist — Text2Motion Inc.

Jan 2024 – Present

- Founded and led a GenAI startup delivering **text-to-3D skeletal animation** for **25,000+ users**, operating high-volume inference pipelines on cost-optimized cloud infrastructure.
- Designed **custom AI models and inference architectures** optimized for **CPU-first execution**, reducing GPU dependency and operating costs by **300%** compared to **GPU/TPU** deployments.

- Architected scalable, low-latency systems on **Google Cloud**, balancing throughput, reliability, and operational efficiency.
- Led full **research-to-production lifecycle**: model design, training, systems architecture, deployment, and observability.

Principal Research Engineer — Microsoft OCTO / Azure IoT AI

Mar 2020 – Jan 2024

- Led research on **differentiable simulation graphs**, distributed AI/ML workloads, and infrastructure optimizations for multi-modal training.
- Designed **AI-driven simulation systems** for autonomous drones, robotics, and industrial environments, integrating compute, sensing, and environmental context.
- Architected and built **embedded hardware testing platforms** enabling sim-to-real reinforcement learning convergence.
- Partnered across Azure, IoT, and Research teams to transition **research prototypes into deployable systems**.

Senior Research Engineer — Amazon (AWS, Twitch, Game Studios)

Nov 2013 – Mar 2020

- Designed and prototyped **ultra-low-latency streaming architectures** using **WebRTC, HLS, and GPU-accelerated encoding**.
- Architected **instance-side infrastructure agents** for AWS AppStream and WorkSpaces, managing KVM, capture, encode, and session orchestration at scale.
- Led R&D for **cloud-based game streaming**, integrating hardware acceleration and real-time networking.
- Built experimental ML-based device prototypes combining **computer vision, hardware interfacing, and cloud services**.

EARLIER RESEARCH & SYSTEMS EXPERIENCE (SELECTED)

- **AIR Worldwide** — Built GPU-accelerated stochastic modeling platforms achieving **order-of-magnitude speedups** over CPU-based systems.
- **Starkey Laboratories** — Redesigned acoustic simulation models for embedded hardware, optimizing SIMD performance across platforms.
- **Vital Images / Eigen / MSL Technologies** — Advanced GPU rendering, medical imaging, and real-time visualization systems.

TECHNICAL SKILLS

Languages: C, C++, Python, C#, TypeScript
AI/ML: PyTorch, PyTorch Lightning, LLMs, Generative AI
Infrastructure: Kubernetes, Docker, OpenStack, GCP, AWS, Azure
Compute: CUDA, OpenCL, SIMD, GPGPU
Systems: Distributed Systems, Real-Time Systems, WebRTC, HPC
Simulation & Graphics: Omniverse, Unreal Engine, OpenCV

EDUCATION

B.E. in Computer Science (Equivalent to B.S in CS)

S.C.S.V.M.V. University, India