David Bennett

Longmont, CO
https://www.linkedin.com/in/dave-bennett-23a6a743/

PROFESSIONAL SUMMARY

Technical leader and hands-on developer with 40+ years solving complex optimization problems across hardware, software, and IT domains. Combines deep coding expertise with proven ability to guide engineering teams. Transforms cutting-edge research into shipping products, with 23 patents spanning FPGA design, compiler technology, CPU architecture, and cloud security.

TECHNICAL EXPERTISE

- Languages: Python, C/C++, Java, CUDA, HIP, Scala, etc.
- Software Architecture: Overall system design, algorithm acceleration, distributed systems, compiler design
- Technical Leadership: Team scaling, M&A integration, team management, project planning, mentoring
- Information Security: Encryption, data protection, cloud security, hardware security
- Compiler Development: LLVM, MLIR, optimization, automatic parallelization and partitioning
- Cloud Platforms: AWS, Google Cloud, Azure
- Cloud Infrastructure: REST, Docker, Cloud Security, Serverless, High-Reliability/High-Availability, Load-Balancing
- Embedded Systems: Linux kernel, real-time systems, driver development
- Large Language Models: Claude, ChatGPT, LLaMa, Mistral, Deepseek
- LLM APIs: Anthropic, OpenAI, Hugging Face, LLaMa, Mistral, Deepseek
- AI Accelerators: NVIDIA/AMD GPUs, AWS Neuron, Google TPU, Groq, FPGA
- AI/ML Stack: Model optimization, training and fine-tuning, PyTorch, JAX, XLA, LoRA, Amazon Neuron Kernel Interface (NKI)
- Databases: DynamoDB, MongoDB, Redis
- CPU Architecture: Memory hierarchy, cache design, hardware/software codesign
- HDL Languages: Verilog, Chisel
- Verification tools: Verilator, Icarus Verilog, Firesim, Chipyard

PROFESSIONAL EXPERIENCE

MyTorch AI **Principal** **Longmont, CO** **1/2025-present**

MyTorch is a development and deployment platform for PyTorch and other AI platforms that lets users run workloads on cloud or enterprise GPUs with zero code changes.

- Designed, co-developed, and deployed a serverless cloud platform that redirects PyTorch workloads to GPU resources without any code modification, eliminating queue wait times and reducing customer GPU costs by 50% or more.
- Developed client API and server-side implementation over gRPC, providing a secure, low-latency message-passing and remote execution system.
- Eliminated unnecessary client-server round trips through an asynchronous execution model that defers data transfer until explicitly required, enabling non-blocking training loops with near-zero network overhead.
- Reduced network bandwidth through a tensor proxy architecture that substitutes lightweight UUID handles for tensor data, with lazy attribute synchronization and a hash-based partial update protocol for future further minimization.
- Created a prompt pruning algorithm that significantly reduces time-to-first-token (TTFT) and per-token generation latency while preserving semantic fidelity more accurately than comparable published methods.
- Conceived and implemented a novel bilinear attention variant that replaces standard multi-head attention, reducing memory by 50% and memory bandwidth by an order of magnitude, with results competitive with MHA on standard benchmarks.

Johnson Lumber Ace Hardware **Consultant** **Remote** **10/2025-11/2025**

Johnson Lumber is one of the largest building materials providers in the Bay Area.

- Designed and delivered an AI-assisted ordering system that analyzes handwritten and typewritten customer materials lists, matches them to part databases and loads them into the company's point-of-sale system.
- Evaluated multiple LLM platforms (OpenAI, Anthropic, DeepSeek), performing comparative testing and prompt engineering to select the best-performing model for production use.
- Developed a hybrid matching approach combining LLM output with deterministic rules, significantly improving part-match accuracy and reducing manual intervention.
- Built desktop (Qt) and web-based user interfaces backed by the REST API, enabling sales staff to review, modify and extend the generated parts lists; the system supports flexible deployment including fully local, client-server desktop, or web-based operation.
- Designed a REST-based backend service to centralize business logic, orchestrating LLM inference and rules-based part matching.
- Increased order processing throughput from ~2 orders per hour to 15+ orders per hour, delivering significant productivity improvement.

Silicon Compiler Corp. **Consultant** **Remote** **3/2025-7/2025**

Silicon Compiler Corporation licensed CacheQ's technology to develop ASIC accelerators to lower the cost of frequently used micro-services.

- Built a high-performance computing toolchain to automatically transform unmodified C code into a RISC-V ASIC with CPU-intensive code implemented directly in hardware.
- Designed and implemented hardware/software interfaces for RISC-V accelerators using the Chipyard framework.
- Developed custom RISC-V cores in Chisel/Scala — including custom opcodes and memory extensions — integrated via the ROCC interface using the CacheQ toolchain to accelerate compute-intensive workloads.
- Integrated comprehensive system verification into the toolchain flow using Verilator and FireSim, ensuring any C code submitted to the pipeline generates automatically verifiable ASIC designs with FPGA-based real-time testing.
- Delivered a fully functional ASIC design pipeline capable of generating production-ready designs with minimal manual intervention, demonstrating a complete end-to-end solution to automate ASIC development.

8090 Solutions, Inc. **VP of Engineering** **Longmont, Colorado** **8/2024-1/2025**

8090 Solutions is a startup company using AI to automate business operations.

- Worked on agentic development platform built around a centralized knowledge graph, specialized AI agents, and multiplayer human-AI collaboration.
- Defined strategy for Transpiler, a tool that would target machine learning kernels to multiple target architectures.
- Created benchmarks to compare the performance of AWS Neuron, Google TPU, AMD and NVIDIA GPUs.
- Preliminary LLM fine-tuning using LoRA.
- Modified kernels to improve performance of LLaMa 3 on AWS Neuron devices using XLA and AWS's Neuron Kernel Interface (NKI).

Groq, Inc **Contractor** **Remote** **10/2023-6/2024**

Groq designs and manufactures custom processors for accelerating artificial intelligence (AI) inference workloads.

- Worked on support for memory spilling, which allowed larger models to be executed in a smaller number of chips. Primarily worked on test and build scripts to test the implementation of spilling.
- Did extensive profiling and performance analysis of Groq's cloud deployment and produced detailed

reports and offered recommendations on ways to improve the performance of their inference platform.
- Optimized compiler performance through LLVM/MLIR transformations targeting Groq's architecture-specific features, improving generated code size and execution efficiency.
- Developed a high-performance lookup data structure for KV Cache optimization, enabling prompt prefix skipping to eliminate redundant processing of system prompts and fixed content.

CacheQ Systems, Inc. **Co-founder and CTO** **Longmont, Colorado** **1/2018-8/2024**

CacheQ was a venture capital-funded startup that developed high-level language compilers targeting a variety of heterogeneous hardware platforms, including GPUs, FPGAs and custom accelerators.

- Cofounded, raised venture capital and defined product-market fit through collaboration with early adopters and enterprise partners.
- Defined development roadmap, resource allocation and timelines.
- Built and led a team of 10 developers in creating innovative compiler technology, achieving performance on unmodified code that surpassed hand-coded results using vendor-specific tools such as CUDA, HIP and FPGA HLS tools.
- Designed and implemented compiler backend for automatic parallelization and optimization targeting GPUs, FPGAs, and multicore CPUs – requiring no code annotations or pragmas, preserving clean CPU-first development.
- Designed distributed shared memory system enabling seamless memory access across heterogeneous computational elements including GPUs, FPGAs and other custom hardware.
- Enabled transparent GPU and FPGA acceleration by generating optimized kernel code and managing data movement across heterogeneous targets.
- Partnered with AI hardware vendors to co-develop compilers to target next-generation accelerators.

IONU Security, Inc. **Cofounder and CTO** **Longmont, Colorado** **2/2013-1/2018**

IONU was a startup providing a software security system focused on prevention of data loss even if the network is penetrated. A secondary product applied the same technology to the IoT market.

- Played a key role in corporate strategy, market strategy and business planning.
- Responsible for overall design, implementation and deployment of IONU's highly secure enterprise data security system.
- Hired and managed a development team of approximately ten software developers.
- Managed IT department and SaaS deployment with High Reliability and High Availability with load balancing and automated failover.
- Developed applications for Windows, OS-X, iOS and Android to provide secure access to enterprise data without leaking.
- Source code was licensed to Dropbox.
- Developed low-inertia highly secure registration for IoT devices and developed demonstrations with Arduino and Renesas boards.

CPU Technology, Inc. **V.P. Software Development** **Longmont, Colorado** **10/2008-2/2013**

CPU Tech developed ultra-secure processors for products containing intellectual property in the compiled software that must be protected under extreme circumstances (mostly military applications). The company was sold to Boeing.

- Managed a team of about fifteen software engineers in Pleasanton, CA and Longmont, CO, producing software and firmware for extremely secure microprocessors used in DoD applications requiring security against foreign government intrusion.
- Ported Linux Operating System to run on the CPU Tech's secure processors.
- Designed and developed a secure C compiler that blocked rogue code from being inserted by the compiler.
- Worked with the IC development team to create one of the earliest PUF (physical unclonable function) implementations to make it into a product. This was required to meet the standards dictated by the DoD for the export of classified technology.

- Designed and implemented a consumer-oriented secure computing environment, which was spun out of CPU Tech into IONU.

Xilinx, Inc. **Distinguished Engineer** **Longmont, Colorado** **6/2004-10/2008**

Xilinx was a Fortune 500 company that invented FPGAs (Field Programmable Gate Arrays), used in electronics ranging from automobiles to high-end networking equipment to SETI. They were acquired by AMD in 2022.

- Led team of 4-5 researchers exploring next-generation FPGA architecture and tools.
- Pioneered innovative C-to-FPGA compiler enabling software developers to leverage FPGA acceleration without hardware expertise. Won a company-wide prize for innovation, beating out dozens of competitors.
- Developed novel design partitioning algorithms that maintained quality while improving scalability.
- Worked closely with lead FPGA architect to design a new asynchronous FPGA.
- Managed relationship with UC Berkeley working on a multi-university project to explore next generation parallel computing architectures.

Xilinx, Inc. **V.P. Software** **Longmont, Colorado** **4/1995-6/2004**

- Managed all software product development at Xilinx after the acquisition of NeoCAD, reporting initially to the CEO and later to the Executive VP of product development.
- Grew department from under 100 to over 300 engineers with facilities in Boulder and Longmont Colorado, San Jose California, Dublin Ireland, Grenoble France and Toronto Ontario.
- Responsibilities included defining product direction, development strategy, development process, division of labor, schedules and personnel.
- Responsible for an annual budget of approximately $50 million.
- During this period, code base grew from 500,000 lines of code to over 20 million.
- Oversaw acquisition of three companies, not including NeoCAD, and integration of those engineering groups into Xilinx.
- During this period Xilinx won numerous awards for having the best EDA software (not just the best software by an FPGA company) from Electronic Design News (EDN), one of the premier periodicals for electronics engineers.
- Numerous completely new Xilinx products were developed or initiated within my group during my tenure. These included:
  - *Xilinx Core Generator*
  - *Matlab HDL Compiler*
  - *Chipscope*
  - *DSP toolkit*
  - *Embedded Development Kit*
  - *6-Input LUT logic block*
  - *System Generator*
  - *MicroBlaze CPU*

NeoCAD Inc. **V.P. of Engineering** **Boulder, Colorado** **11/1990-4/1995**

NeoCAD developed the only complete design solution with support for multiple FPGA vendors. NeoCAD was acquired in April of 1995 by Xilinx, Inc.

- Co-founded this venture capital-funded company and managed all engineering activities, both software and hardware. This included concept, design, implementation and productization.
- Build from scratch a team of approximately 40 Engineers, including Software Development, Software QA, Hardware Development and Technical Writers.
- Managed all of IT including acquisition of equipment and architecture of network and server infrastructure.
- Reverse engineered low-level architectural details of Xilinx and Actel FPGAs so that our tools were able to place, route, compute timing and program them. Both vendors believed this was impossible until we did it.
- With VP of Marketing, negotiated multimillion-dollar licensing contracts with Motorola and AT&T to resell NeoCAD's software. Worked with those companies to define and develop support for their FPGAs.

Daisy/Cadnetix Inc. **Director of CAD/CAM Technology** **Boulder, Colorado** **6/1983-11/1990**

Cadnetix, Inc. was the leading developer of computer-aided design tools for printed circuit boards. It was acquired by

Daisy Systems, Inc. in 1988 and the merged company was the world's largest Electronic Design Automation tool supplier at that time.

- Supervised development of all of Cadnetix PCB CAD products, which accounted for 80% of the company's revenue. Responsibilities included hiring, scheduling, personnel development, and budgets.
- Managed over 30 software engineers and other managers developing applications for electronic design.
- Designed, managed and developed the Cadnetix Route Engine, the company's flagship product and recognized industry leader in fully automatic PCB routing, which was built around a bit-slice processor.
- After the Daisy/Cadnetix merger, was given additional responsibility for the activities of remote development site in Israel with over 20 software engineers.