Ian F.V.G. Hunter (Ireland, EMEA)

ianfhunter | ianfhunter | ianhunter.ie | ianfhunter@gmail.com | On Request Only

Summary

With over 10 years of experience, mostly in optimizing the performance of next-generation AI chips, with broad secondary skills, and history in both technical and management roles, I am able to apply myself effectively to most positions. I am only open to **fully remote** options due to my residence (though I am open to periodic office visits and/or international travel). I have been working fully remotely with both local and global teams since 2019. I am particularly interested in roles which involve research/prototyping work and facilitate publication / conference attendance.

Work Experience

AMD (Advanced Micro Devices) — SMTS [NPU Architecture, Shift Left] 2024 - Present

My work at AMD focussed on validating and optimizing performance of our next generation AI chips at the presilicon stage, enabling greater confidence in performance projections, providing early feedback to many teams including the compiler, hardware and production-grade software groups. (XDNA1 through 3),

- Prototyping of full network inferences on hardware simulation/emulators/physical devices.
- Upgraded manual processes with more scalable & performant solutions
  - Memory/Stream Allocators, Graph abstractions, Code generation
  - Early error detection, Interoperability with other team's data
  - Visualization tool for Memory lifetime analysis, stream allocations
- Assisted other teams for critical deadlines with performance as a deliverable
- Achieved industry-leading performance on UL Procyon AI Computer Vision Benchmark.
- Technologies: C++, Python, Git, Jenkins, Make, GTKWave, PULP, Numpy

Intel/Movidius — Various Titles [NPU Presilicon Power & Performance] 2015-2024

- Deep Learning Engineering Manager [Intel] (2021-2023)
- Senior Deep Learning Engineer [Intel] (2020-2021)
- Deep Learning Engineer [Intel] (2016-2020)
- Embedded Vision Engineer [Movidius] (2015-2016)

I created the first Neural Network compiler for Movidius's range of embedded processors ('VPUs'/NPUs) in 2015. This became the prime focus of Movidius - we unveiled the Fathom Neural Compute Stick at NIPS and were acquired by Intel the next year to be their NPU offering.

Since then, I improved the compiler and corresponding embedded runtime over several VPU generations - all descendants of that once-prototype. I managed a team of 4-8 employees all dedicated to maximizing the chips' performance during the pre-silicon stage.

Recieved Divisional Award in Q2 2021 & Q1 2022 for key innovations in NPU performance, published one paper in Intel DTTC and filed 2 patents on NPU Compiler technologies.

- Demonstrating performance capabilities of future NPUs for various KPI Networks
- Host-side development of a Python-based AI Compiler
  - Network graph algorithmics using NetworkX, Numpy and other libraries (Dijkstra's Algorithm, Partitioned Boolean Quadratic Programming, etc)
  - Interfacing with various AI frameworks such as ONNX, TensorFlow, Caffe, PyTorch.
  - Cutting-Edge features such as INT4 and other sub-byte type support.
- C++ Device Applications
  - HW Drivers & Libraries (e.g. Matrix Multiplication)

  - Runtime Control for AI VLIW Processors & Hardware Units.
- Web-based Dev Tooling
  - Data Collection / Training System for NN-Based Cost Model (See Publication)
  - Visualization of Tensors + Device Workloads
- Robust Test & Continuous Integration (Jenkins)
- Management
  - Coaching, Guidance, Hiring & Firing, Raises/Promotions, Technical Roadmapping, Charter Definition, Cross-Team Collaboration, JIRA Kanban/Scrum
- Also: Linux, Power Measurement, Direct customer interaction, Technical assistance/Coaching, Priority support, Conference Attendance / Booth hosting, Paper Publication & Patents as below.

Wonga (DevOps, Build Systems, Legacy Software) 2014-2015

FullStack (Web, Android) 2013

GetBulb (Web, Data Visualization) 2012

Winners of Irish Times Digital Innovation Award

Projects

GNOLL Link to Repository

Some of my hobbies are boardgames and tabletop role-playing games. Both of these often use a syntax for describing dice rolls which is non-trivial, widely used and totally organically defined. As there were little-to-no topics on the matter, I took it upon myself to analyse hundreds of rpg systems, concretely define a grammar and publish a easily installed library for anyone to use. My work has been published in the Journal of Open Source Software. (C, Yacc/Lex, Bison/Flex, FFIs to 14 other programming languages)

Education

2017 - 2019 M.S.(Research) Computer Science at **Trinity College Dublin** (Grade: N/A (PASS))

2010 - 2014 B.A.(Mod) Computer Science at **Trinity College Dublin** (Grade: I (GPA 4.0 Equivalent))

Patents & Publications

Patents

"Neural Network Based Power and Performance Model for Versatile Processing Units" (2022a). Patent Pending.

"Graph Neural Network Model for Neural Network Scheduling Decisions" (2023b). Patent Pending.

Journal Papers

GNOLL: Efficient Software for Real-World Dice Notation and Extensions (2023a). *Journal of Open Source Software.*

Kyōgi Karuta Overseas: Analysing how Phonetic Variation in the Kimari-Ji of The Hyakunin Isshu affects Gameplay in International Adaptations (2024a). *Electronic Journal of Contemporary Japanese Studies.*

Conference Papers

Towards Optimal VPU Compiler Cost Modeling by using Neural Networks to Infer Hardware Performances (2022b). *Intel Design & Test Technology Conference.*

Pre-silicon performance benchmarking and actionable architecture feedback (2024b). *AMD EMEA Tech Summit.*

Unlocking Industry-Leading Inference on XDNA2 NPUs (2025b). *AMD Global Technical Authors Conference.* Winner of Best Poster - SW-AI Track.

Posters and Demos

Fathom: Myriad2 Neural Compute Stick (2016a). Embedded Vision Summit.

Movidius Fathom: Deep Learning in a USB Stick (2016b). Conference on Neural Information Processing Systems (NIPS).

Pre-silicon performance benchmarking and actionable architecture feedback (2024c). AMD EMEA Tech Summit.

Industry-Leading Inference on XDNA2 NPUs (2025a). AMD EMEA Tech Summit. Recieved many votes for Best Poster.

Theses

Optimizing Web Content Download in Low-Performance Networks (2014). BA(Mod) Final Year Project. 1st Place for Best Poster, Notable Mention for Presentation. Trinity College Dublin.

Effective Index-Mapping of Quantized Values for Low-Precision Neural Networks on Power-Efficient Embedded Devices (2020). Master's Thesis. Trinity College Dublin.

Skills

Programming Languages	C++, Python, Make, Bash, Javascript, TCL, Cmake
Libraries & Frameworks	Numpy, NetworkX, Pandas, Docker-Compose, Jenkins, Gerrit, Caffe, Tensorflow, ONNX, PyTorch
Working Methodologies Management	Scrum, Performance Management, OKRs
Technologies	Neural Networks, Artificial Intelligence, Inference, Quantization, Optimization, Refactoring Legacy Software, Git, Jenkins, VLIW Processors, Parallelism, Linux, Embedded Systems