Vishal Pramod Kasliwal

Last update on January 16, 2023

vishal.kasliwal@gmail.com • +1.267.206.9287 • https://github.com/AstroVPK
6289 Mahan Dr. • San Jose, CA 95123. • USA
US Permanent Resident. • Indian Citizen
Fluent in English & Hindi

Experience

Luminous Computing - Office of the CTO — Santa Clara, CA

I work for the CTO at Luminous Computing. My role focuses on technical product defintion & development.

Technical Product Definition & Performance Architecture — *April '22 – present*

I work in the CTO's office within Luminous Computing. My role lies at the intersection of technical product definition, performance architecture & hardware/software co-development. In the Office of the CTO, I'm tasked with determining & defining what Luminous will build. My role consists of

1. engaging with external customers to understand the evolving needs of the marketplace, expectations of system & software behavior
2. identifing & collating key AI workloads of interest
3. analyzing workloads with the goal of understanding workload characteristics
4. mapping the workload into Luminous Computing's software & hardware to determine expected system performance, discover bottlenecks, & problems
5. architecting hardware & software features to improve workload performance
6. working with the engineering team to productize the improvements

I interact and advise the engineering team on the architecture as it develops. I also work on on future iterations of Luminous' product and help the CTO define the forward roadmap.

Responsibilities:

- Build analytical performance models for deep-learning workloads executing on novel hardware system designs to answer key questions about the value proposition of Luminous' product
- Analyze behavioral performance models of AI HW accelerator system components (compute, memory, & interconnects) within Luminous' internal system-level architecture modeling and simulation framework
- Participate in identifying system- and component-level bottlenecks that exist within and/or across the network, memory, and compute subsystem boundaries
- Advise the engineering team on hardware and software features that can improve runtime of ML models of interest on current and future architectures
- Assure that what the engineers are building - cutting across the software stack, the digital architecture, the compiler, and the hardware - is consistent with the intended goals of the product requirements and product value proposition
- Use architecture & modeling to identify key system figures-of-merit (FOM) & bottlenecks. Utilize this knowledge to help the CTO define a forward technology roadmap for future products that address these bottlenecks.

Intel Corporation - Senior GPU Software Development Engineer — Santa Clara, CA

I work in the Software Architecture group within AXG. My role focuses on architecting Intel's Level Zero GPU driver.

Level Zero Architecture — *August '21 – April '22*

My responsibilities included

- defining the software architecture of the Level Zero GPU driver.
- defining new architectural features for enhancing Intel's GPU architecture for HPC & Deep Learning workloads.
- driving adoption of best practices, i.e. architecture-specific know-how, in software products.

Intel Corporation - Senior Deep Learning Software Engineer — Santa Clara, CA

I work in the Machine Learning Performance (MLP) organization in the Machine Learning Distributed Compute (MLDC) group on accelerating Deep Learning workloads on Intel's Xe-HPC discrete accelerator cards for Deep Learning & High-Performance Computing. My role focuses on pre-Si performance optimization via hardware-software co-design. I develop & test computation- & communication-kernels using hardware simulators.

Hardware-Software Co-Design — *April '19 – August '21*

My responsibilities included

- defining architectural features for enhancing the architecture for Deep Learning workloads.
- driving adoption of best practices, i.e. architecture-specific know-how, in software products.
- evaluating & projecting pre-Si Deep Learning workload performance.
- creating the overall strategy for distributing Deep Learning workloads across multiple-cards & -nodes i.e. scale-up/scale-out strategy.

oneCCL Development *October '20 – August '21*

I was responsible for

- researching & implementing superior algorithms for hierarchical & non-hierarchical collective communication kernels.
- developing collective communication kernels for low-precision datatypes (bfloat16, fp16, etc...).

post-Si Performance Validation *October '20 – August '21*

My contributions included

- developing performance validation tests for scale-up & scale-out.
- working with the performance validation team to understand the observed performance behavior of new hardware.

Wave Computing - Senior Staff Research & Development Software Engineer Campbell, CA

Wave Computing is developing the next-generation of solutions for speeding up Deep Learning applications using Dataflow Processing Units (DPUs), which contain thousands of interconnected dataflow Processing Elements (PEs). DPUs power Wave Computing's custom appliance for developing, testing, and deploying Deep Learning models. I use to develop the compute- and data movement- software kernels which are executed by Wave Computing's Dataflow Processing Units (DPUs) for Deep Learning acceleration.

Deep Learning Kernel- & Library-Development *Dec '17 – April '19*

I have worked extensively on compute- & communicate-kernels, tools, and the supporting library for enabling Deep Learning workloads on Wave's DPUs.

- Developed compute kernels for various Deep Learning layers such as Average Pooling, Convolutions, Feed-Forward, Activations, Concatenation & fork, etc...
- Developed a tool for visualizing the place & route performed by JitPR.
- Owner of the libary of routines for performing IEEE 754 rounding.
- Owner of the library of routines for changing precision.
- Authored various block Matrix Multiplication operations.
- Authored various multi-byte addition operations.
- Authored various multi-byte shift operations.

Tools Development *Nov '18 – Apr '19*

I have developed tools for

- Simulating (functional & performance) Deep Learning kernels.
- Visualizing the placement & routing performed by JitPR.

Management Role *March '18 – April '19*

I assisted my superior with the management of the kernel team consisting of seven engineers. My duties included

- maintaining the schedule of work being done by the members of the team.
- assisting in the planning of new work items.
- identifying & interviewing new compute-kernel team candidates.
- report on the progress of the compute-kernel team to the VP of product engineering.

Major accomplishments include

- I developed a workflow for creating new compute-kernels.
- I integrated the schedule of work into a JIRA managed project.

The outcome of my efforts is better project management leading to a significant reduction in the time taken to develop new compute-kernels.

Colfax International - High Performance Computing (HPC) Research Engineer Sunnyvale, CA

HPC Computational Fluid Dynamics (CFD) Application Development *Mar '17 – Dec '17*

HPC consulting project. I parallelized a CFD simulation code written in C for a client in the oil & gas sector resulting in a $\sim$ 10X speedup. I have also refactored the client's codebase to enable independent time-evolution in different regions of the simulation via domain-decomposition methods.

C/C++ Compiler Analysis *Nov '17*

HPC research project. I investigated suitability of C++ compilers for HPC applications. I developed optimized scientific computational kernels and investigated the performance obtained by each compiler from each kernel. I analyzed compiled binary code to determine reason for differences in performance. A technical report of my findings can be obtained at Colfax Research.

Intel Advisor Lecture *May '17 – Jul '17*

I presented a lecture on Intel Advisor for the Stanford University course ME 344: Introduction to High Performance Computing on July 20th, 2017.

Large Synoptic Survey Telescope (LSST) Data Management (Princeton University) Princeton, NJ

Postdoctoral Research Associate *Sept '15 – Feb '17*

LSST Data Management is building a C++ & Python software stack to analyze raw imaging data from LSST. I worked on the software stack to add functionality, documentation, & tests. I developed & implemented algorithm to propagate covariance when stacking images and worked on techniques for optimal image stacking & differential chromatic refraction. I worked on a machine-learning based star-galaxy classifier and on converting the LSST stack to use py.test.

Department of Physics & Astronomy (University of Pennsylvania) Philadelphia, PA

Postdoctoral Researcher *Sep '15 – Feb '17*

I developed and implemented a parallelized Bayesian algorithm to estimate orbital parameters from stochastic light curves of binary supermassive black holes. I also developed and implemented Python framework to automatically wrangle astronomical time-series data from a variety of sources including web-servers, SQL servers, data servers and local data files.

Principle Developer *Sept '15 – Feb '17*

I architected and implemented kālī, an open-source high performance library to model stochastic time-series data in a Bayesian framework. kālī is capable of modeling time-series data as variants of C-ARMA processes (a type of Gaussian random process). Written primarily in C++and exposed to Python using Cython, kālī uses scikit-learn for machine learning, Intel MKL for fast linear algebra, Intel Bull Mountain technology for hardware random number generation, & OpenMP 4.0 for vectorization & parallelization. kālī is being used to study astronomical time-series data by multiple research groups at Caltech, UPenn, & Drexel.

Department of Physics (Drexel University) Philadelphia, PA

AGN Variability Analysis *June '09 – Aug '15*

I developed C++sofwtare for Intel Xeon Phi accelerator cards to model AGN variability. I developed vectorized & parallelized the C++ pipeline to forward-model and fit data to model using MLE of 2^{nd}-order statistics.

LSST Photo-z Analysis *Sept '08 – May '09*

I used MLE & machine learning (neural networks) to establish optimal y-band filter for LSST galaxy photo-z distance estimation.

Department of Physics (Virginia Commonwealth University) Richmond, VA

Adjunct Instructor *Jun '07 – Aug '08*

I taught *Introduction to Astronomy* course.

AFM Image Analysis *Aug '05 – May '07*

I implemented an IDL pipeline to analyze AFM images of silicon surfaces etched using oxygen.

Department of Physics (University of Richmond) Richmond, VA

Cosmic Microwave Background Analysis *May '03 – May '05*

I used IDL to perform statistical tests of the utility of the bispectrum for detection of non-Gaussianity in the CMB.

Education

Drexel University Philadelphia, PA

PhD. in Physics *2008 – 2015*

Probing AGN Accretion Physics through AGN Variability: Insights from *Kepler*

Virginia Commonwealth University Richmond, VA

M.S in Physics & Applied Physics *2005 – 2007*

CAFM Studies of Epitaxial Lateral Overgrowth GaN Films

University of Richmond Richmond, VA

B.S. in Mathematics & Physics *2001 – 2005*

The Bispectrum as a Quantifier of non-Gaussianity in the Cosmic Microwave Background

Certifications

deeplearning.ai — COURSERA.ORG

Custom Models, Layers, and Loss Functions with TensorFlow
Certificate earned on December 17, 2020.

Deep Learning Specialization
Certificate earned on June 24, 2018.

Sequence Models
Certificate earned on June 24, 2018.

Convolutional Neural Networks
Certificate earned on January 22, 2018.

Structuring Machine Learning Projects
Certificate earned on December 17, 2017.

Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization
Certificate earned on November 26, 2017.

Neural Networks and Deep Learning
Certificate earned on November 4, 2017.

Skills

Technical expertise: My expertise lies in the research, design & implementation of high-performance scientific/numerical software on exotic, highly-parallel, and distributed hardware. Current interests include hardware-software co-design, particularly in the area of AI accelerators. I earned my doctorate in Astrophysics by applying mathematial & statistical analysis and machine learning to complex time-domain data. While I am familiar with developing Deep Learning applications using TensorFlow & PyTorch, my speciality lies in knowledge of the innards of these frameworks & their execution on various accelarators suh as GPUs, DPUs, etc.... Used to working with(in) teams, I am fond of using Agile methodologies (Scrum) and continuous integration (Jenkins) to manage & deliver on-time & bug-free software. I enjoy writing C-For-Metal, OpenCL, FlowGraph, C, C++, & Python, and am learning Julia/CUDA/Go/Io/Prolog. I have excellent knowledge of parallelization technologies: OpenMP, MPI, C++11 threads, & POSIX threads, & hardware architecture. Over the years, I have gained extensive experience programming on novel computing platforms such Intel's Xe-HPC accelerators , Wave Computing's Dataflow Processing Units (DPU) & Intel's Xeon Phi (Knight's Corner & Knight's Landing). I have extensive knowledge & experience with various programming toolchains including the Intel toolchain, GNU toolchain, LLVM toolchain, PGI toolchain, AOCC toolchain, Valgrind, gdb, make, SCons etc.... My preferred development platform is Linux ($\geq$ 15 years of development experience) although I have also developed on Windows (1 year of dev-ex), & Mac OSX (7 years of de-ex). I am very comfortable with writing in LaTeX ($\geq$ 15 years of experience) and have a good understanding of the UNIX programming environment and tools (memory-management, process spawning, etc...)

Public speaking: With years of experience delivering highly technical talks to both expert & general audiences, I am comfortable with public speaking & outreach.

Natural languages: English (*native language*) and Hindi (*native language*).

Service

Particles Member of the Reviewer Board of the journal Particles published by MDPI.

The National Science Foundation Served on a grant review panel for the Division of Astronomical Sciences.

The Astrophysical Journal Peer reviewed publications.

The Astronomical Journal Peer reviewed publications.

Monthly Notices of the Royal Astronomical Society Peer reviewed publications.

Publications

A Performance-Based Comparison of C/C++ Compilers Colfax Research, 2017

Science-driven Optimization of the LSST Observing Strategy arXiV, 2017

Large Synoptic Survey Telescope Galaxies Science Roadmap arXiV, 2017

Extracting Information from AGN Variability MNRAS, 470, 3, 3027-3048, 2017

The LSST Data Management System Proceedings of ADASS XXV, 2015

Do the Kepler AGN light curves need reprocessing? MNRAS, 453, 2075, 2015

Are the variability properties of the Kepler AGN light curves consistent with a damped random walk? MNRAS, 451, 4328, 2015

Thirty Meter Telescope Detailed Science Case: 2015 http://arxiv.org/abs/1505.01195, 2015

AFM and CAFM studies of ELO GaN films Proc. SPIE 6473, 647308, 2007

Local electronic and optical behaviors of a-plane GaN grown via epitaxial lateral overgrowrth Appl. Phys. Lett., 90, 011913, 2007

Grants

Kepler Guest Observer Program Co-Investigator on Kepler Guest Observer Program accepted proposals K2 GO16088, K2 GO14088, K2 GO12013, K2 GO8052, & K2 GO10052

NASA Grant NNX14AL56G Helped write proposal for awarded NASA Grant NNX14AL56G. Grant was used to fund my Ph.D. research.

Presentations

Webinar on Intel & A Day in the Life of a Deep Learning Software Engineer HiCounselor, September 24th, 2019, Sunnyvale, CA

Applications of High Performance Computing in Artificial Intelligence Rajasthan Student Startup Exposure Program 2018 (RSSEP2018), September 8th, 2018, San Jose, CA

Intel Advisor Stanford University ME344: Introduction to High Performance Computing, July 20th, 2017, Stanford, CA

Optical Variability Signatures from Massive Black Hole Binaries 229th Meeting of the American Astronomical Society, 2017, Grapevine, TX

Extracting Information From AGN Variability: an LSST AGN Collaboration Proposal 2017 LSST AGN Science Collaboration Roadmap Development Meeting, 2017, Grapevine, TX

Extracting Information from AGN Variability 2016 KARL LSST Workshop, November 2016, Louisville, KY.

Surveying the Dynamic Sky with the LSST 2016 KARL LSST Workshop, November 2016, Louisville, KY.

AGN Variability: Insights from Kepler 2016 Hotwiring the Transient Universe V Meeting, October 2016, Villanova, PA.

Probing Accretion Processes through Variability 2016 TMT Science Forum 'International Partnership for Global Astronomy', May 2016, Kyoto, Japan.

AGN Variability: Insights from Kepler Princeton HSC Science Discussion Series, March 2016, Princeton, NJ.

AGN Variability on Short Timescales: What does Kepler tell us about AGN Variability? 2015 TMT Science Forum 'Maximizing Transformative Science with TMT', June 2015, Washington, DC.

What can Kepler tell us about AGN variability? 225th Meeting of the American Astronomical Society, January 2015, Seattle, WA.

Do Kepler AGN Light Curves Exhibit a Damped Random Walk? 24th Meeting of the American Astronomical Society, June 2014, Boston, MA.

The Bispectrum of Galactic Dust: Implications for Microwave Background non-Gaussianity 204th Meeting of the American Astronomical Society, May 2004, Denver, CO.

Conferences

HotChips 2020 on behalf of Intel Corporation August 16th - 18th, 2020 San Jose, CA

The Next AI Platform 2020 on behalf of Intel Corporation June 25th, 2020 San Jose, CA

The Next FPGA Platform 2020 on behalf of Intel Corporation Jan 22nd, 2020 San Jose, CA

The Next AI Platform 2019 on behalf of Intel Corporation May 9th, 2019 San Jose, CA

Super Computing 17 (SC17) on behalf of Colfax International November 12th - 17th, 2017 Denver, CO