# Vishal Pramod Kasliwal

Last update on May 21, 2024

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

## Summary

*Former astrophysicist with a passion and insaitiable appetite for building highly scalable AI-systems.*

Two AI hardware startups (Luminous Computing & Wave Computing) and two AI chip behemoths (Intel Corp. & AMD Corp.) have taught me that AI systems are incredibly hard to design & build correctly. Having left my roots in Astrophysics behind (2017), I've become an expert on High Performance Computing, AI acceleration, and the process of designing chips and systems to accelerate AI for both training as well as inference. My experiences at chip startups and large established corporations have left me with a healthy respect for teamwork. My previous role at Luminous Computing was focused on technical product definition & performance architecture. It gave me the chance to interact not only with potential customers to better understand their needs & requirements, but also at a deep technical level with the engineering team to create a truly performant AI system for powering the future. My current role at AMD is focused on developing highly performant software libraries for AMD's CDNA architecture GPGPUs.

## Experience - Industry

### Intel Corporation - Senior Staff GPU Compute Architect — Santa Clara, CA

**GPU Compute Architect** — *May '24 – present*

I drive end-to-end GPU compute hardware/software codesign. My responsibilities include

- Building expertise in workloads in AI and HPC to drive compute architecture
- Analyzing performance of these workloads, determining bottlenecks and using these insights to drive compute architecture
- Specifing and building the tools needed for such analysis
- "Hands on" ML kernel/operator optimization for Windows and Linux ecosystem (DirectML, OpenVINO, SYCL)
- Working closely with higher level software stacks (AI frameworks, high level language compilers and runtime) to drive end-to-end performance
- Transitioning the architectural specifications and prototype code to software engineering, oversee development and follow through to product deployment and support
- Interfacing for LevelZero in partnership with other compute architects

### Advanced Micro Devices (AMD) - Machine Learning Libraries — Santa Clara, CA

**ML Performance Optimization** — *Aug '23 – May '24*

I worked on Machine Learning Libraries to accelerate AI performance on AMD's CDNA Data Center GPGPUS (MI100x, MI200x, etc...). My responsibilities include

- optimizing low-precision GEMM kernels, i.e. 8-bit floating-point (E4M3 & E5M2), 8-bit integer, and 16-bit floating point (bf16 & fp16), for multiple generations of AMD's CDNA architecture.
- identifying & eliminating performance bottlenecks via runtime profiling.

I work with a geographically distributed team spread over two continents & use modern software development practices to manage the project.

### Luminous Computing - Sr. Staff Engineer — Santa Clara, CA

**Technical Product Definition & Performance Architecture** — *Apr '22 – May '23*

I worked directly with the CTO at Luminous Computing on determining & defining what Luminous would build. I focused on a combination of technical product definition, performance architecture & hardware/software co-development. My responsibilities included

- engaging with external customers to understand the evolving needs of the marketplace, expectations of system & software behavior.
- identifying & collating key AI workloads of interest.
- analyzing workloads with the goal of understanding workload characteristics.
- mapping the workload into Luminous Computing's software & hardware.
- building analytical performance models to determine expected system performance, discover system- and component-level bottlenecks, & problems.

- co-architecting hardware & software features to improve workload performance.
- working with the engineering team to productize recommended architectural improvements.
- assuring 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.
- identifying key system figures-of-merit (FOM) & bottlenecks and utilizing this knowledge to define a forward technology road-map for future products that address these bottlenecks.

I worked with the engineering team on the chip & system architecture as it developed. I also worked on future iterations of Luminous' product and defined the forward road-map of the company.

### Intel Corporation - Senior GPU Software Architect — Santa Clara, CA

I worked in the Software Architecture group within AXG. My role focused 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 worked 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 focused on pre-Si performance optimization via hardware-software co-design. I developed & tested computation- & communication-kernels using hardware simulators.

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

My responsibilities included

- 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.
- 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.

**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 data-types (bfloat16, fp16, etc...).

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

Wave Computing was 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 were meant to power Wave Computing's custom appliance for developing, testing, and deploying Deep Learning models. I developed the compute- and data movement-software kernels which were meant to be 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 library 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 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.

## Research Experience in Academia (prior to March, 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 $2^{\text{nd}}$-order statistics.

---

## 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

---

## Publications, Grants, & Service

**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

**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.

**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.