Vamsi Sripathi

Technical Lead - HPC Software Optimization, Intel

admin@vamsis.com 919.599.1040
github.com/vamsi-sripathi linkedin.com/in/vamsisripathi vamsis.com

Summary

- 15 years of experience in applying x86 code optimizations to mathematical libraries, deep learning frameworks and scientific applications on multiple generations of Intel Processors.
- 8 years of leadership experience in driving HPC/AI technical engagements with cross-org team of junior/mid-level staff engineers.
- 6 years of software product development experience in Intel Math Kernel Library.
- Demonstrated track-record of identifying performance bottlenecks and optimizations for compute and memory bound workloads through robust understanding of CPU micro-architecture.

Key Accomplishments

- Directly contributed to several Intel Silicon design wins valued at $M's by delivering targeted code optimizations.
- Upstreamed code optimizations (AVX512 vectorization, cache blocking, prefetching, NUMA) to deep learning frameworks (TensorFlow, Caffe, Eigen) and HPC domains (Climate/Weather, Quantum Chromo Dynamics, Geospatial). Optimized AVX512 instruction sequence for prefix-sum, argmax that beats Intel, GCC, Clang Compiler performance.
- Enabled key external customers and collaborated with cross-organizational teams to enhance the positioning of Intel platforms in the HPC/AI markets. Worked with a wide spectrum of customers - US/Europe National Labs (ORNL, ANL, LANL, NCAR, ECMWF, TUDA), CSPs (Meta, Amazon), Taboola, Lenovo, HPE, GE, Siemens, Ford, General Motors.
- Proposed and collaborated with lead architects in development and evaluation of future Intel hardware features and ISA extensions. Synthesized complex application workloads to representative kernels used in performance debug (emulation, post-Silicon) of Intel CPUs and GPUs.
- Implemented AVX, AVX2, AVX512 optimizations to Basic Linear Algebra Subroutines (BLAS)/matrix operations in Intel Math Kernel Library (MKL) for Intel CPUs. Designed and developed compiler SIMD vector intrinsics framework for MKL BLAS optimizations. Robust product development experience spanning 5 major releases of MKL.
- Optimized MPI I/O performance on Lustre file-system at 100K processes of Cray XT petascale system.
- Senior Member of ACM and IEEE. Publications in IEEE and ACM conferences.

Employment History

September 2010 - Current	Technical Lead - HPC Software Optimization, **Intel**. Senior Software Optimization Engineer (HPC, AI), **Intel**. Software Development Engineer (Math Kernel Library, 2010 - 2016), **Intel**.
August 2007 – August 2010	Graduate Research Assistant, **North Carolina State University**.
Summer 2008	Intern (Computational Earth Sciences), **Oak Ridge National Laboratory**.

Education

2007 – 2010 **M.S., Computer Science** North Carolina State University, USA.
Thesis: *Performance Analysis and Optimization of Parallel I/O in a Large Scale Groundwater Application on Petascale Architectures*

2003 – 2007 **B.Tech., Information Technology** V. R. Siddhartha Engineering College / Acharya Nagarjuna University, India.

Skills

Programming Languages	C, C++, Fortran 90
Parallel Programming	OpenMP, MPI, CUDA, OpenCL, SYCL/DPC++, Intel TBB
Low-level Code Optimizations	x86 ASM, Compiler SIMD Vector Intrinsics (AVX+)
Scripting Languages	BASH, Python
Intel Architectures	Xeon (AVX, AVX2, AVX512), Xeon + HBM, Xe GPUs, Xeon Phi
Development Tools	VTune, Intel EMON, SDE, GNU Binutils, GDB, Screen, Vim, Git
Libraries	Intel MKL, Eigen, Intel DML, HDF5, PETSc
Deep Learning Frameworks	TensorFlow, Caffe
Misc	Singularity, SQLite, MongoDB, HTML/CSS
Operating Systems	GNU/Linux
Supercomputing Platforms	Cray XT, IBM BlueGene/P

Patents

- Automated Resource Usage Configurations for Deep Learning Neural Network Workloads on Multi-generational Computing Architectures
- Method, Computer Program, and Computer System for Prefetching Data During Execution of an Application Program (Pending)

Invited Talks

- *Optimization of ACRANEB2 Radiation Kernel on Intel Xeon Processors*. Delivered at ECMWF's 19th Workshop on High Performance Computing in Meteorology Fall 2021. (*video*, *slides*)
- *Optimization of TensorFlow-Serving Application on Intel Xeon Scalable Processors*. Delivered at Intel Extreme Performance Users Group (IXPUG) Fall 2018. (*slides*)
- *TensorFlow Performance Optimizations on Intel Architectures*. Delivered at Argonne National Laboratory Leadership Facility (ALCF) Developer Session, July 2018. (*slides*)
- *Scalable Algorithms for Clustering Large Geospatiotemporal Data Sets on Manycore Architectures*. Delivered at Seventh Workshop on Data Mining in Earth System Science, part of IEEE International Conference on Data Mining 2017. (*slides*)

Technical Projects

- High Performance Computing
  - Improved HotQCD (Lattice Quantum Chromo Dynamics) performance by 1.4x on Intel Sapphire Rapids CPUs with HBM by optimizing the lattice memory layout and software prefetching. (*slides*, *article*)
  - Improved the performance of MPAS-A (Model for Prediction Across Scales - Atmosphere) by 1.25x on Intel Sapphire Rapids CPUs with HBM by applying code tuning techniques. (*slides*)
  - Delivered 1.3x performance improvements to a weather application (from European Center for Medium Weather Forecasts [ECMWF]) on Intel Icelake Processors through explicit AVX512 vectorization of prefix-sum operations and cache blocking. (*slides*, *article*)
  - Optimized STREAM benchmark by employing hybrid Intel Data Streaming Accelerator (DSA) and CPU software pipelining methodology. (*article*)
  - Root-caused HBM performance anomalies on Intel Sapphire Rapids Processors and developed software mitigations (data alignment, prefetching).
  - Ported SPECFEM3D-Cartesian (computational seismology) from CUDA to SYCL for Intel GPUs.
  - Developed kernels used in performance debug (emulation, post-Silicon) of Intel CPUs and GPUs.
- Deep Learning/AI
  - Directly contributed to $M's Intel Silicon design win with a leading content recommendation provider by delivering targeted AVX512 SIMD code optimizations in a deadline driven engagement (*white-paper*, *slides*). This win was highlighted in quarterly earnings call by Intel CEO and covered by media outlets.
  - Implemented 8-bit integer matrix-matrix multiplication kernels using Vector Neural Network Instructions (VNNI) on Intel Cascade Lake Processors and improved the performance of Transformer-LT (language translation model) with TensorFlow. (*paper*)
  - Formulated and delivered software test plan for successful bring-up of Intel AI Accelerator for top hyperscalers/CSP's.
  - Proposed and developed run-time profiling capabilities to Intel MKL-DNN library.
  - Ported MKL FFT's to TensorFlow C++ backend which delivered performance gains of 6x and addressed competitive threat for Intel Silicon.
  - Improved performance of Eigen on Intel Xeon Phi (KNL) with AVX512 intrinsics and OpenMP.
- Intel Math Kernel Library (MKL)
  - Designed, developed and optimized Basic Linear Algebra Subroutines (BLAS) in MKL for Intel Xeon and Xeon Phi architectures.
  - Applied a broad set of code tuning techniques to optimize floating point and parallel efficiency of BLAS. Optimized matrix-matrix and matrix-vector operations spanning multiple generations of Intel CPU micro-architectures (AVX - 256bit SIMD, AVX2 - 256bit SIMD + FMA, AVX512 - 512bit SIMD + FMA).
  - Developed Intel Thread Building Blocks (TBB) parallelism for MKL BLAS and enhanced BLAS OpenMP performance on high core count architectures.
  - Implemented bitwise numerical reproducibility of floating-point operations in MKL BLAS under variable data alignment conditions. Developed a test suite to validate bitwise accurate results of all BLAS with multiple precisions (FP32, FP64, COMPLEX64, COMPLEX128) and memory alignment offsets.
  - Proposed and implemented a complete re-design of Intel Compiler's libmatmul to support AVX ISA.

- Open-Source Work - *https://github.com/vamsi-sripathi?tab=repositories*
  - Optimized AVX512 SIMD implementations of *prefix-sum*, *argmax* that beats Intel, Clang and GCC Compiler performance by 5-20x.
  - Improved the performance of N-dimensional tensor broadcast operations in Tensorflow by 4-30x using SIMD techniques in Eigen framework.
  - Ported Intel MKL matrix API's (GEMM, batched GEMM) to TensorFlow framework and improved the performance of RNN workloads.
  - Delivered performance gains of 1.75x (AlexNet topology) for Intel Caffe framework on Intel Xeon Phi architecture (KNL) through SIMD optimization of data transformation layers.
  - Improved the performance of k-means clustering algorithm by 2.7x on Intel Xeon Phi (KNL) architecture by developing OpenMP parallelization and vectorization techniques.
  - Accelerated the performance of Deep Learning Inference workloads by enabling TensorFlow Serving framework to use Intel MKL.
  - Developed a tool called TACKLE (Thread Affinity Advisor, Checker, and Launcher) that recommends the ideal thread affinity, NUMA binding and taskset settings for OpenMP based applications. (*github*)
- Graduate Research
  - Performance analysis and optimization of PFLOTRAN, a highly scalable groundwater simulation code, which uses MPI for inter-process communications, PETSc for solving numerical equations, HDF5 for parallel I/O on Cray XT and IBM Blue Gene/P (BGP) supercomputing platforms. (*slides*)
  - Optimized the performance of HDF5 parallel read and write I/O in PFLOTRAN by 40x and 3x respectively at 100,000 processor cores of Cray XT5 on Lustre file system. (*paper*, *poster*)
  - Characterized compute, communication and I/O system differences between Cray XT and IBM BGP. Improved performance of MPI All_reduce() on Cray XT5 with hybrid MPI-OpenMP implementation.

Selected Publications

For full list of papers/technical articles, see my *Google Scholar profile*

- **Sripathi**, **V.** (2024). Accelerating Memory-Bandwidth-Bound Kernels Using the Intel Data Streaming Accelerator. *"Parallel Universe Magazine", Issue-57.*
  https://www.intel.com/content/www/us/en/developer/articles/news/parallel-universe-magazine/issue-57-july-2024.html
- **Sripathi**, **V.** (2023). Optimize QCD Performance on Intel Processors with HBM. *"Parallel Universe Magazine", Issue-53.*
  https://www.intel.com/content/www/us/en/developer/articles/news/parallel-universe-magazine/issue-53-july-2023.html
- **Sripathi**, **V.** (2021a). Optimization of Scan Operations Using Explicit Vectorization. *"Parallel Universe Magazine", Issue-44.*
  https://www.intel.com/content/www/us/en/developer/articles/technical/optimize-scan-operations-explicit-vectorization.html
- **Sripathi**, **V.** (2021b). Optimizing the Maxloc Operation Using Intel AVX512 Instructions. *"Parallel Universe Magazine", Issue-46.*
  https://www.intel.com/content/www/us/en/developer/articles/news/parallel-universe-magazine/issue-46-october-2021.html

- Bhandare, A., **Sripathi**, **V.**, Karkada, D., Menon, V., Choi, S., Datta, K., & Saletore, V. (2019). Efficient 8-bit quantization of transformer neural machine language translation model. https://arxiv.org/abs/1906.00532
- Mills, R. T., **Sripathi**, **V.**, Kumar, J., Sreepathi, S., Hoffman, F., & Hargrove, W. (2018). Parallel k-means clustering of geospatial data sets using manycore cpu architectures. *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, 787–794. https://doi.org/10.1109/ICDMW.2018.00118
- Sreepathi, S., Kumar, J., Mills, R. T., Hoffman, F. M., **Sripathi**, **V.**, & Hargrove, W. W. (2017). Parallel multivariate spatio-temporal clustering of large ecological datasets on hybrid supercomputers. *2017 IEEE International Conference on Cluster Computing (CLUSTER)*, 267–277. https://doi.org/10.1109/CLUSTER.2017.88
- Sreepathi, S., **Sripathi**, **V.**, Mills, R., Hammondz, G., & Mahinthakumar, G. K. (2013). Scorpio: A scalable two-phase parallel i/o library with application to a large scale subsurface simulator. *20th Annual International Conference on High Performance Computing*, 443–451. https://doi.org/10.1109/HiPC.2013.6799128
- Mills, R. T., Hammond, G. E., Lichtner, P. C., **Sripathi**, **V.**, Mahinthakumar, G., & Smith, B. F. (2009). Modeling subsurface reactive flows using leadership-class computing. *Journal of Physics: Conference Series*, *180*(1), 012062. https://doi.org/10.1088/1742-6596/180/1/012062