Vamsi Sripathi

Portland, Oregon
Mobile: 919-599-1040
Email: admin@vamsis.com
Web: http://www.vamsis.com

Objective

To pursue technical leadership positions where I can apply my R&D experience to deliver optimal solutions to emerging computing applications and paradigms.

Core Skillset

- 13 years of experience in applying x86 code optimizations to mathematical libraries, deep learning frameworks and scientific applications on Intel Processors.
    - Directly contributed to several Intel Silicon design wins valued at $M's by delivering targeted code optimizations.
    - Contributed to Deep Learning workloads performance analysis on future Intel CPU ISA extensions. Formulated and delivered software test plan for successful bring-up of Intel AI Accelerator for top hyper-scalers/CSP's.
    - Enabled key external customers and collaborated with cross organizational teams to enhance Intel architectures positioning in deep learning space.
    - Open source contributions: code optimizations to TensorFlow, Caffe, Eigen frameworks.
    - 6 years of experience in optimization of Basic Linear Algebra Subroutines (BLAS) in Intel Math Kernel Library (MKL) for Intel CPUs. Robust product development experience covering 5 major and several minor releases.
- 6 years of leadership experience in driving internal/external technical engagements with cross-org team of junior/mid-level staff engineers.
- More details at http://www.vamsis.com

Work Experience

[May 2018 – Current] Senior HPC Application Engineer – Intel Corporation

Optimization of HPC workloads on Intel CPU and GPU architectures.

- Improved the performance of HotQCD (Lattice Quantum Chromo Dynamics) framework by 1.4x on Intel Sapphire Rapids Processors with High Bandwidth Memory (HBM) by optimizing the lattice memory layout and software prefetching (technical article)
- Improved the performance of MPAS-A (Model for Prediction Across Scales - Atmosphere) by 1.25x on Intel Sapphire Rapids Processors with High Bandwidth Memory (HBM) by applying code tuning techniques.
- Optimization of STREAM benchmark kernels with Intel Data Streaming Accelerator (DSA)
- Delivered performance improvements (1.3x) to a weather application (from European Center for Medium Weather Forecasts [ECMWF]) on Intel Icelake Processors through explicit AVX512 vectorization of prefix-sum/scan operations (talk, slides)
- Analyzed memory bandwidth performance for various memory traffic patterns across Intel, AMD and ARM architectures.

[March 2016 – May 2018]

Senior Machine Learning Engineer – Intel Corporation

Optimization of machine learning/deep learning algorithms and frameworks on current and future generation of Intel architectures.

- Directly contributed to Intel Silicon design win (valued at $M's) with a leading content recommendation provider by delivering targeted code optimizations in a deadline driven engagement. (White-paper). This win

was highlighted in quarterly earnings call by Intel CEO and covered by top media outlets (Reuters, Venturebeat).

- Formulated and delivered software test plan for successful bring-up of Intel Neural Network Processor (NNP, next-gen AI Accelerator) for top hyper-scalers/CSP's, enabled key external customers and collaborated with Intel cross organizational teams to enhance Intel architectures positioning in deep learning space.
- Implemented 8-bit integer matrix-matrix multiplication kernels using Vector Neural Network Instructions (VNNI) targeted for Intel next-gen Cascade Lake processors. Improved the performance of a leading language translation model (Transformer-LT) by accelerating the performance of integer GEMM in TensorFlow.
- 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.
- Proposed and developed run-time profiling capabilities to Intel MKL-DNN library. This feature has been upstreamed and widely used by user community.
- Contributed to Deep Learning workloads performance analysis on future ISA for evaluating next-gen Intel Silicon.
- Ported MKL FFT's to TensorFlow C++ backend which delivered performance gains of 6x and addressed competitive threat for Intel Silicon for a leading medical imaging provider.
- Implemented AVX512 Compiler vector intrinsics and OpenMP parallelism to improve performance of Eigen framework on Intel Xeon Phi (KNL).
- Developed Singularity based containerized solutions for TensorFlow and Caffe frameworks for deployment on large scale distributed clusters.
- Open-Source work/contributions:
  - TACKLE (Thread Affinity Checker, Advisor, Launcher) - https://github.com/vamsi-sripathi/tackle
  - Optimized AVX512 SIMD implementations of prefix-sum, argmax - https://github.com/vamsi-sripathi/simd
  - Improved the performance of N-dimensional tensor broadcast operations in Tensorflow using SIMD techniques in Eigen framework by 4-30x.
  - Improved the performance of RNN workloads by enabling TensorFlow to use Intel MKL as the backend engine for acceleration of linear algebra functions, ported Intel MKL matrix-matrix multiplication API's (GEMM, batched GEMM) to TensorFlow framework.
  - Accelerated the performance of Deep Learning Inference workloads by enabling TensorFlow Serving framework to use Intel MKL.
  - Delivered performance gains of 1.75x (AlexNet topology) for Intel Caffe framework on Intel Xeon Phi architecture (KNL) through SIMD optimization of data transformation operations.
  - Improved the performance of k-means clustering algorithm employed in a geospatial data application by 2.7x on Intel Xeon Phi (KNL) architecture by developing OpenMP parallelization and vectorization techniques.

[SEPTEMBER 2010 – MARCH 2016]

SOFTWARE DEVELOPMENT ENGINEER FOR INTEL MATH KERNEL LIBRARY – INTEL CORPORATION

Design, develop and optimize Basic Linear Algebra Subroutines (BLAS) in MKL for Intel Xeon and Xeon Phi architectures. Robust product development experience covering 5 major and several minor releases.

- Applied a broad set of code tuning techniques to optimize floating point and parallel efficiency of MKL BLAS on Intel architectures.
- Optimized matrix-matrix, matrix-vector and vector-vector operations spanning multiple generations of Intel CPU microarchitectures (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.
- Designed and developed compiler vector intrinsics framework for MKL BLAS optimizations.
- Played a key role in implementing 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), arbitrary number of inputs and memory alignment offsets.

- Proposed and implemented a complete re-design of libmatmul, an Intel Compiler component, to support AVX instructions.

[JUNE - AUGUST 2008]

SUMMER INTERN - OAK RIDGE NATIONAL LABORATORY

Performance analysis of PFLOTRAN to quantify compute load imbalance on application scalability at 16,000 processor cores of Cray XT. Benchmarked application I/O performance (HDF5, MPI-IO) on Lustre file system of Cray XT.

Mentor: Dr. Richard Mills

Group: Computational Earth Sciences, Computer Science and Mathematics division.

[AUGUST 2007 – AUGUST 2010]

GRADUATE RESEARCH ASSISTANT - NORTH CAROLINA STATE UNIVERSITY

Performance analysis and optimization of large-scale scientific applications on supercomputing platforms (Cray XT4/5 and IBM BlueGene/P).

Research Advisor: Dr. Mahinthakumar.

Funding Agency: US Department of Energy (DOE) - Scientific Discovery through Advanced Computing (SciDAC) initiative.

- Performance analysis and optimization of PFLOTRAN, a highly scalable groundwater simulation code, which uses MPI for inter-process communications, PETSc framework for solving numerical equations, HDF5 for parallel I/O on Cray XT and IBM Blue Gene/P (BGP).
- 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.
- Characterized the compute, communication and I/O sub-system differences between Cray XT and IBM BGP. Improved the performance of MPI All_reduce() collective operation on Cray XT5 with hybrid MPI-OpenMP implementation.
- Selective instrumentation of PFLOTRAN and PETSc functions with Tuning and Analysis Utilities (TAU) and Cray Performance Analysis Tools (PAT) on Cray XT and IBM BGP.

Technical Skills

- Programming Languages: C, C++, Fortran 90.
- Parallel Programming: OpenMP, MPI, OpenCL, SYCL/DPC++, Intel TBB.
- Low-Level Code Optimizations: x86 ASM, Compiler SIMD Vector Intrinsics (SSE, AVX1,2,512, VNNI).
- Scripting Languages: BASH, Python.
- Libraries: Intel MKL, Eigen, Intel DML, HDF5, PETSc.
- Deep Learning Frameworks: TensorFlow, Caffe.
- HPC Tools: Intel SW tools (Compilers, VTune, SDE), Cray PAT, TAU, IBM HPCToolkit.
- Dev Tools: Vim, GNU Binutils, Subversion, Mercurial, Git, Singularity, Docker, Valgrind, Total View, Doxygen.
- Databases: MongoDB, MongoEngine, SQLite.
- Web Designing: Flask, JavaScript, CSS, HTML.
- Operating Systems: GNU/Linux (Fedora/RHEL, OpenSUSE/SLES).
- Intel Architectures: Xeon (Nehalem/Westmere, Sandy Bridge/Ivy Bridge, Haswell/Broadwell, Skylake, Cascade Lake, IceLake, Sapphire Rapids) and Xeon Phi (Knights Corner, Knights Landing), Intel Xe GPUs, Neural Network Processors (NNP).
- Supercomputing Platforms: Cray XT, IBM BlueGene/P.

Education

[AUGUST 2007 – AUGUST 2010]: M.S IN 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 **(**Presentation**)**
[JUNE 2003 - APRIL 2007]: BACHELOR OF TECHNOLOGY IN INFORMATION TECHNOLOGY - V. R. SIDDHARTHA ENGINEERING COLLEGE (NAGARJUNA UNIVERSITY), INDIA.

Publications

- **V. Sripathi**, *Optimize QCD Performance on Intel Processors with HBM*. Published in Parallel Universe Magazine – Issue 53, July 2023
- **V. Sripathi**, *Optimizing the Maxloc Operation Using Intel AVX512 Instructions*. Published in Parallel Universe Magazine – Issue 46, October 2021
- **V. Sripathi**, Ruchira Sasanka, *Optimization of Scan Operations Using Explicit Vectorization*. Published in Parallel Universe Magazine - Issue 44, April 2021
- R. T. Mills, **V. Sripathi**, J. Kumar, S. Sreepathi, F. Hoffman, W. Hargrove, *Parallel k-means Clustering of Geospatial Data Sets Using Manycore CPU Architectures*. Eighth Workshop on Data Mining in Earth System Sciences (DMESS 2018), Proceedings of the IEEE International Conference on Data Mining (ICDM 2018) Workshops.
- Sarat Sreepathi, Jitendra Kumar, Forrest M. Hoffman, Richard T. Mills, **Vamsi Sripathi**, William W. Hargrove, *Parallel Multivariate Spatio-Temporal Clustering of Large Ecological Datasets on Hybrid Supercomputers*. IEEE Cluster 2017.
- Murat Guney, Sarah Knepper, Kazushige Goto, **Vamsi Sripathi**, Greg Henry, and Shane Story. *Batched Matrix-Matrix Multiplication Operations for Intel Xeon Processor and Intel Xeon Phi Co-Processor*, 2015 SIAM Conference on Applied Linear Algebra.
- Sarat Sreepathi, **Vamsi Sripathi**, Richard Mills, Glenn Hammond, G. Kumar Mahinthakumar, *SCORPIO: A Scalable Two-Phase Parallel I/O Library With Application to a Large Scale Subsurface Simulator*, IEEE Conference on High Performance Computing (HiPC) 2013, Bengaluru, India.
- G. (Kumar) Mahinthakumar, **Vamsi Sripathi**, Sarat Sreepathi, ***Comparison of parallel solvers for large-scale groundwater contaminant transport simulations***. XIX International Conference on Computational Methods in Water Resources (CMWR 2012).
- R. T. Mills, G. E. Hammond, P. C. Lichtner, **V. Sripathi**, G. Mahinthakumar and B. F. Smith, *Modeling subsurface reactive flows using leadership-class computing***,** Journal of Physics, 2009.
- Jacqueline Chame, Chun Chen, Mary Hall, Jeffrey K. Hollingsworth, Kumar Mahinthakumar, Gabriel Marin, Shreyas Ramalingam, Sarat Sreepathi, **Vamsi Sripathi**, Ananta Tiwari, ***PERI Autotuning of PFLOTRAN***, In Journal of Physics, Proceedings of SciDAC July 2011.
- R. T. Mills, **V. Sripathi**, G. Mahinthakumar, G. E. Hammond, P. C. Lichtner, B. F. Smith, ***Engineering PFLOTRAN for Scalable Performance on Cray XT and IBM BlueGene Architectures***, Proceedings of SciDAC 2010, July 11-15, 2010, Chattanooga, TN, USA. Invited paper.
- **V. Sripathi**, G. E. Hammond, G. Mahinthakumar, R. T. Mills, P. H. Worley and P. C. Lichtner. *Performance Analysis and Optimization of Parallel I/O in a large scale groundwater application on the Cray XT5*, Poster Presentation, *Supercomputing* 2009, Portland, Oregon, Nov 12-16.
- R. T. Mills, **V. Sripathi**, G. Mahinthakumar, G. E. Hammond, P. C. Lichtner, B. F. Smith, ***Experiences and Challenges Scaling PFLOTRAN, a PETSc-based Code for Subsurface Reactive Flow Simulations, Towards the Petascale on Cray XT Systems***, Cray Users Group Meeting, May 2009, Atlanta, GA.
- **Tech Articles:**
    - Optimizing the Maxloc Operation Using Intel AVX512 Instructions. Published in Parallel Universe Magazine - Issue 46, October 2021
    - Optimization of Scan Operations Using Explicit Vectorization. Published in Parallel Universe Magazine - Issue 44, April 2021
    - Optimization of TensorFlow-Serving Application on Intel Xeon Scalable Processors

- Best Practices for Scaling Deep Learning Training and Inference with TensorFlow On Intel Xeon Processor-Based HPC Infrastructures
- Boosting Deep Learning Training & Inference Performance on Intel Xeon and Intel Xeon Phi Processors
- TensorFlow Optimizations on Modern Intel® Architecture
- Recipe: Optimized Caffe for Deep Learning on Intel® Xeon Phi™ processor x200

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***. Invited talk at Seventh Workshop on Data Mining in Earth System Science, part of IEEE International Conference on Data Mining 2017. (slides)

Professional Affiliations/Activities

- Participated in several Intel microarchitecture and external customer hackathons.
- Program committee member for the Seventh Workshop on Data Mining in Earth System Science, in conjunction with 2017 IEEE International Conference on Data Mining (ICDM 2017).
- Mentored Intel MKL summer interns.
- Represented Intel MKL at SuperComputing'11, 12, 13 conferences.
- Student Volunteer for SuperComputing'08, 09 conferences.
- Member of ACM.

Recognitions

- Received several Intel org-level and department-level awards:
  - Artificial Intelligence Products Group
  - Data Center Group
  - Software and Services Group
  - Developer Products Division
  - Developer Relations Division

Patents

- Automated resource usage configurations for deep learning neural network workloads on multi-generational computing architectures