

Seonmyeong Bak

2860 County Hwy G4, Santa Clara, CA 95051, psn016@gmail.com

TECHNICAL INTERESTS

Deep Learning, High Performance Computing, Parallel Computing, Task-level Parallelism, Parallel Programming Models, Runtime Systems, Resource Management, Computer Architecture

EDUCATION

Georgia Institute of Technology, GA, USA May. 2018 - Dec 2020

Ph.D in Computer Science (Advisor: [Prof. Vivek Sarkar](#))

- Ph.D. Thesis: Runtime Approaches to Improve the Efficiency of Hybrid and Irregular Applications

University of Illinois at Urbana-Champaign, IL, USA (Transferred to Georgia Tech) Sep. 2015 - May. 2018

Ph.D Student in Computer Science (Transferred to Georgia Tech, Advisor: [Prof. Laxmikant Kale](#))

- Worked on the integration of Charm++/AMPI and OpenMP runtime for intra-node load balancing on multicore nodes

Seoul National University, Seoul, Korea Sep. 2012 - Aug. 2014

Master of Science in Electrical Engineering and Computer Science (Advisor : [Prof. Jaejin Lee](#))

- Masters Thesis: Lightweight and block-level concurrent sweeping for JavaScript garbage collection

SungKyunKwan University, Seoul, Korea Mar. 2006 – Aug. 2012

Bachelor of Science in Computer Engineering / Bachelor of Economics (Dual Degree)

- Total Credit: 148, GPA : 4.01 / 4.5 (upper 4.16 / 4.5)
- Major Credit of Computer Science: 52 , GPA 4.13 / 4.5

SKILLS & EXPERTISE

Programming Languages: C, C++, Python, OpenMP, MPI, Charm++, OpenCL, CUDA

Open Source Projects Used for Research: WebKit, GNU/LLVM OpenMP, Charm++/AMPI, Boost, HCLib

HPC Tools: CrayPat, Vtune, NVProf, PAPI, OProf (Profilers), LSF, PBS, SLURM (Job schedulers)

Supercomputers Used: Selene, Pre-EOS, BlueWaters(CrayXE/XK), Stampede1/2, Cori / Theta(CrayXC), Summit, Cori-GPU

Tools: Git, GDB, LLDB, Bash, Mercurial, Redmine, Gerrit

EXPERIENCES

Senior System Software Engineer, NVIDIA Corporation Feb. 2021 – present

- Checkpointing improvement in Megatron-Core with asynchronous, fully-parallel approach
- Working on optimization of ML software stack and applications at NVIDIA for supercomputers
- Suggest or make changes for Megatron and HugeCTR for better performance and scaling
- Performance analysis and debugging of ML applications on DGX SuperPod bring-up machines such as Selene(A100), EOS(H100) and next-gen clusters

Teaching Assistant, Georgia Institute of Technology Jan. 2020 – May. 2020

- CS 4240 Compilers & Interpreters, Spring 2020

[ASTRO](#) Intern, CSR Group, Oak Ridge National Laboratory Aug. 2019 - Dec. 2019

- Extension of OpenMP for SLATE library requirements on Summit/Frontier

Visiting Student, MCS Division, Argonne National Laboratory May. 2018 - Aug. 2018

- Design and Implement user defined scheduling API for OpenMP runtime

Internship in OpenMP Team, Software Service Group, Intel Corporation May. 2017 – Aug. 2017

- Design and Implement the composability of **Intel OpenMP** runtime library
- Optimized OpenMP so that multiple external instances call libraries written in OpenMP runs efficiently without contention on resources
- Evaluated the composability with commercial python frameworks: **Tensorflow**, **DASK**

Internship in ICT Institute, SK Telecom

Jun. 2014 – Aug. 2014

- Developing web framework and applications based on WebRTC

Internship in Business of Mobile communication, Samsung Electronics

Jun. 2011 – Aug. 2011

- Developing a tool to analyze logs in Bada OS (variation of RTOS)

Sergeant, Military Service in the Army of the Republic of Korea

Mar. 2008 – Jan. 2010

- Mandatory Service for Korean males

RESEARCH ACTIVITIES

Task Scheduling in Task Graphs for Improved Communication/Synchronization

Aug 2019 - Dec 2020

- Improved scheduling of tasks with internal communication operations
- Hybrid scheduling of **gang-scheduling** and **work-stealing** and optimized victim selection in work-stealing for improved synchronization and computation/communication overlap, with performance improvements demonstrated for CPU/GPU versions of LU, QR and Cholesky factorization kernels in **SLATE**, the successor of **ScaLAPACK**. This work is under submission

MPI + Asynchronous Many Task (AMT) Programming Model to Improve Resiliency

Sep 2018 - Aug 2019

- Using AMT to improve **resiliency** in parallel applications. Especially using HClb for AMT which is a C++ library and enables task parallelism on applications written in C/C++. Published in **Euro-Par '19**
- Extended version which is interoperable with MPI is accepted to **ExaMPI '20** (to appear)

User-defined scheduling API on OpenMP constructs

May 2018 - May 2019

- Propose a set of APIs to extend the specification of parallel loops in OpenMP by user-provided functions
- Handling **performance variance** and **load imbalance** incurred by input datasets on irregular iterative applications (Graph, Scientific applications using sparse matrix)
- Implemented in LLVM OpenMP runtime and published in **ICPP '19**

Integration of OpenMP into Charm++ / Adaptive MPI for Node-Level Parallelism

Sep 2015 – May 2018

- Comparison of Charm++ and common task-level parallel programming paradigms for node parallelism
- Integrated OpenMP into Charm++/AMPI for improved node-level performance on many-core nodes
 - This work started on GNU OpenMP and has migrated into LLVM OpenMP
 - This feature has been available to the public since Charm++, 6.8.0
 - Published in **ACM ESPM '17** workshop (co-located with SC'17) for the first version
 - Published in **CCGrid '18** with a more optimized version and detailed analysis

Porting of Rubik, Topology-aware Mapping Framework for Cray Machines

Sep 2015 – Aug 2016

- Ported Rubik, a Python framework for topology aware mapping to work on Blue Waters (CrayXE/XK hybrid)
- Developed a new compaction scheme to map a logical complete cuboid into a physical torus network, which is not cuboid. Poster presented at **SC 16 ACM SRC**

TIZEN Memory Management Optimization in Cooperation with Samsung Electronics

Oct 2012 – Aug 2013

- Analysis of memory usage of WebKit engine. Memory Management Optimization for a JavaScript engine in WebKit (JavaScriptCore). Evaluated on a ARM development device (Tizen development kit)
- Published in **LCTES '14** (co-located with PLDI '14).

Mini Projects for SnuCL, a distributed OpenCL framework

July 2012 – Oct 2012

- Porting of AES algorithm from a single-threaded C version to OpenCL

- Comparison of AES block cipher modes and implemented more optimized version than the AMD reference implementation of AES in OpenCL
- Porting of SnuCL, a distributed OpenCL framework to ARM and Verification of the ported version with OpenCL conformance test on a **ARM** development board (Beagleboard)

PUBLICATIONS

[ICS 21] Seonmyeong Bak, Oscar Hernandez, Mark Gates, Piotr Luszczek, Vivek Sarkar. [Task-Graph Scheduling Extensions for Efficient Synchronization and Communication](#) In **ICS '21: 35th ACM International Conference on Supercomputing**, June 15-18, Worldwide online event. (39/157, acceptance rate 24.8%)

[Euro-Par 19] Sri Raj Paul, Akihiro Hayashi, Nicole Slattengren, Hemanth Kolla, Matthew Whitlock, **Seonmyeong Bak**, Keita Teranishi, Jackson Mayo and Vivek Sarkar. [Enabling Resilience in Asynchronous Many-Task Programming Models](#). In **Euro-Par '19: 25th International European Conference on Parallel and Distributed Computing**, August 26-30, Göttingen, Germany. (38/144, acceptance rate 26.4%)

[ICPP 19] Seonmyeong Bak, Yanfei Guo, Pavan Balaji, and Vivek Sarkar. [Optimized Execution of Parallel Loops via User-Defined Scheduling Policies](#). In **ICPP '19: 48th International Conference on Parallel Processing**, August 5–8, 2019, Kyoto, Japan. ACM, New York, NY, USA, 10 pages. (106/405, acceptance rate 26.2%)

[CCGrid 18] Seonmyeong Bak, Harshitha Menon, Sam White, Matthias Diener, and Laxmikant Kale. [Multi-level Load Balancing with an Integrated Runtime Approach](#). In **CCGrid '18: Eighteenth IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing**, May 1-4, 2018, Washington DC, USA (52/250, acceptance rate 20.8%)

[ESPM2 17] Seonmyeong Bak, Harshitha Menon, Sam White, Matthias Diener, and Laxmikant Kale. [Integrating OpenMP into the Charm++ Programming Model](#). In **ESPM2'17: Third International Workshop on Extreme Scale Programming Models and Middleware**, November 12–17, 2017, Denver, CO, USA

[SC 16 SRC] Seonmyeong Bak, Nikhil Jain, Laxmikant Kale [Mapping applications on Irregular allocations](#). Poster presented in **SC '16 ACM SRC: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis**, Salt Lake City, United States, November 2016.

[LCTES 14] Honjune Kim, **Seonmyeong Bak**, Jaejin Lee. [Lightweight and block-level concurrent sweeping for JavaScript garbage collection](#). In **LCTES '14: Proceedings of the 2014 SIGPLAN/SIGBED conference on Languages, Compilers and Tools for Embedded Systems**, pp. 155-164, Edinburgh, United Kingdom, June 2014. (16/51, acceptance 31.3%)

HONORS & AWARDS

Academic Excellence Scholarship

- SungKyunKwan University, Korea (2006 Fall, 2011 Spring, 2011 Fall, 2012 Spring)

Student Research Competition Poster Grant in SC '16

- Grant for hotel, transportation (ground, flight), registration and meals. 26 posters accepted out of 63 submissions

Student Volunteer in SC' 17

- Grant for hotel accommodation and registration, and opportunities to talk with mentors with HPC expertise

IEEE CCGrid '18 Travel Grant Award

- Grant for hotel and airfare. 15 awardees selected