Jonghyun Bae

1220 N Mathilda Ave, Sunnyvale, CA, USA 94089

jonghbae90@gmail.com | jonghyunbae.github.io

Current Position

Google — ***Sunnyvale, CA, USA***

Visiting faculty researcher — *April 2024 - Present*

- Evaluation, analysis, design, and optimization of Google's vector DB
- **Supervisor:** Jichuan Chang

Education

Seoul National University — ***Seoul, Korea***

Ph.D. in Computer Science and Engineering — *Sep 2017 - Feb 2022*

- **Dissertation:** A Large-Batch, High-Throughput Training System for Deep Neural Networks
- **Advisor:** Professor Jae W. Lee

Sungkyunkwan University — ***Suwon, Korea***

M.S. in Electrical and Computer Engineering — *Mar 2015 - Aug 2017*

- **Dissertation:** Jointly Optimizing Task Granularity and Concurrency for In-Memory MapReduce Frameworks
- **Advisors:** Professors Jae W. Lee and Jaehyuk Choi

Sungkyunkwan University — ***Suwon, Korea***

B.S. in Semiconductor Systems Engineering — *Mar 2009 - Feb 2015*

Professional Experience

Lawrence Berkeley National Laboratory — ***Berkeley, CA, USA***

Postdoctoral researcher — *Oct 2022 - March 2024*

- Analyzing and optimizing performance of scientific applications on HPC environment
- **Supervisor:** Leonid Oliker, Khaled Ibrahim

Artificial Intelligence Institute of Seoul National University (AIIS) — ***Seoul, Korea***

Postdoctoral researcher — *April 2022 - Aug 2022*

- Leveraging high-performance storage system for ML/AI performance optimization
- **Supervisor:** Professor Jae W. Lee

NAVER Clova AI Research — ***Seongnam, Korea***

Research Intern — *May 2020 - Aug 2020*

- Research on automatic search for effective augmentation policies
- **Advisor**: Ji-Hoon Kim

Google Summer of Code — ***Virtual***

Student Project — *Mar 2014 - Aug 2014*

- Improving the R-interactive-Graphics-via-HTml (RIGHT) Package
- **Advisor**: Junghoon Lee and Jae W. Lee

Research Experience

Optimizing In-Memory Data Storage for Scalable Atomic Simulation [MLG-HPCE'23][IPDPSW'24]

- Analyzed how the performance of each communication method varies with training environments
- Implemented high-performance data reader for the compressed data and an online decompression manager leveraging GPU
- Presented an in-memory distributed data store designed for GNN training on large-scale graph data
- Optimized in-memory data loader for distributed DNNs that can switch between one-sided and collective methods

Accelerating Data Preparation for DNN Training [ECCV'22] [APSys'23]

- Analyzed resource utilization and performance of data preparation pipeline on DNN training
- Developed accelerator-friendly lossless image format for high-throughput DNN training
- Leveraged multiple image formats to balance the loading and decoding step in data preparation pipeline

High-Performance ML/AI Training on Modern NVMe SSDs [USENIX FAST'21a] [USENIX FAST'21b]

- Exploiting SSDs to provide large memory for NN training running on GPUs
- Designed and implemented FlashNeuron, a prototype SSD-based NN training system, on PyTorch and NVIDIA CUDA
- Suggested faster and smarter AutoAugment, a dynamic data augmentation method improving accuracy and robustness

High-Throughput Big Data Analytics on Modern Low-latency SSDs [BigData'17] [IEEE Micro'19] [USENIX ATC'19]

- Analyzed performance and energy efficiency of native and virtualized Apache Spark clusters from memory and storage perspectives
- Minimized spill/garbage collection (GC) overhead of Apache Spark in native and virtualized environments
- Developed SSDStreamer, a high-performance SSD-based object caching system for Apache Spark targeting ML workloads
- Proposed a NAND block erase suspension technique for reducing tail latency and improving QoS

Scalable Visualization on SparkR [SparkSummit'16]

- Expanded `ggplot2`, the most widely used data visualization package in R, to handle big data (SparkR DataFrame) on distributed nodes
- Implemented APIs of `dplyr`, `reshape2`, and `lubridate` packages to enable big data processing.

Improving the R-interactive-Graphics-via-HTml (RIGHT) Package [UseR'14]

- Offloaded R code for analysis on a server to overlay the results on the plot and update them interactively, supporting `ggplot2`-like R API
- Selected for Google Summer of Code (GSoC) in 2014

Publications

Conferences

[SAC'25] SkipLSM: Fast Retrieval of Hot Key-Value Pairs on LSM Tree

Jongsung Lee, Sam Son, **Jonghyun Bae**, Yunho Jin, Tae Jun Ham and Jae W. Lee
ACM/SIGAPP Symposium On Applied Computing (SAC), Sicily, Italy, March 2025 (To appear)

[ECCV'22] L3: Accelerator-Friendly Lossless Image Format for High-Resolution, High-Throughput DNN Training

Jonghyun Bae, Woohyeon Baek, Tae Jun Ham, and Jae W. Lee
European Conference on Computer Vision (ECCV), Tel-Aviv, Israel, October 2022

[USENIX ATC'21] ASAP: Fast Mobile Application Switch via Adaptive Prepaging

Sam Son, Seung Yul Lee, Yunho Jin, **Jonghyun Bae**, Jinkyu Jeong, Tae Jun Ham, Jae W. Lee, and Hongil Yoon
USENIX Annual Technical Conference (ATC), Virtual, July 2021

[USENIX FAST'21a] FlashNeuron: SSD-Enabled Large-Batch Training of Very Deep Neural Networks

Jonghyun Bae, Jongsung Lee, Yunho Jin, Sam Son, Shine Kim, Hakbeom Jang, Tae Jun Ham, and Jae W. Lee
USENIX Conference on File and Storage (USENIX FAST), Virtual, February 2021

[USENIX FAST'21b] Behemoth: A Flash-centric Training Accelerator for Extreme-scale DNNs

Shine Kim*, Yunho Jin*, Gina Sohn, **Jonghyun Bae**, Tae Jun Ham, and Jae W. Lee
(* Equally contributed)
USENIX Conference on File and Storage (USENIX FAST), Virtual, February 2021

[HPCA'21] Layerweaver: Maximizing Resource Utilization of Neural Processing Units via Layer-Wise Scheduling

Young H. Oh, Seonghak Kim, Yunho Jin, Sam Son, **Jonghyun Bae**, Jongsung Lee, Yeonhong Park, Dong Uk Kim, Tae Jun Ham, and Jae W. Lee
IEEE International Symposium on High-Performance Computer Architecture (HPCA), Virtual, February 2021

[ISCA'20] A Case for Hardware-Based Demand Paging

Gyusun Lee*, Wenjing Jin*, Wonsuk Song, Jeonghun Gong, **Jonghyun Bae**, Tae Jun Ham, Jae W. Lee, and Jinkyu Jeong
(* Equally contributed)
International Symposium of Computer Architecture (ISCA), Virtual, May 2020

[USENIX ATC'19] Practical Erase Suspension for Modern Low-latency SSD

Shine Kim, **Jonghyun Bae**, Hakbeom Jang, Wenjing Jin, Jeonghun Gong, Seungyeon Lee, Tae Jun Ham, and Jae W. Lee
USENIX Annual Technical Conference (USENIX ATC), Renton, WA, July 2019

[BigData'17] Jointly Optimizing Task Granularity and Concurrency for In-Memory MapReduce Frameworks

Jonghyun Bae*, Hakbeom Jang*, Wenjing Jin, Jun Heo, Jaeyoung Jang, Joo-Young Hwang, Sangyeun Cho, and Jae W. Lee
(* Equally contributed)
IEEE International Conference on Big Data (BigData), Boston, MA, December 2017

Journals

[IEEE Micro'19] SSDStreamer: Specializing I/O Stack for Large-Scale Machine Learning

Jonghyun Bae, Hakbeom Jang, Jeonghun Gong, Wenjing Jin, Shine Kim, Jaeyoung Jang, Tae Jun Ham, Jinkyu Jeong, and Jae W. Lee
IEEE Micro, vol. 39, Sep-Oct, 2019

[IEICE TIS'18] Eager Memory Managerment for In-Memory Data Analytics

Hakbeom Jang, **Jonghyun Bae**, Tae Jun Ham, and Jae W. Lee
IEICE Transactions on Information and Systems, March 2018

Workshops

[MLG-HPCE'24] MDLoader: A Hybrid Model-Driven Data Loader for Distributed Graph Neural Network Training

Jonghyun Bae, Jong Youl Choi, Massimiliano Lupo Pasini, Kshitij Mehta, Pei Zhang, and Khaled Ibrahim
Workshop on Machine Learning with Graphs in High Performance Computing Environments (Be held in conjunction with SC '24), Atlanta, GA, Nov 2024.

[IPDPSW'24] MDLoader: A Hybrid Model-driven Data Loader for Distributed Deep Neural Networks Training

Jonghyun Bae, Jong Youl Choi, Massimiliano Lupo Pasini, Kshitij Mehta, and Khaled Ibrahim
IEEE International Parallel and Distributed Processing Symposium Workshops, San Francisco, CA, May 2024 (To appear)

[MLG-HPCE'23] DDStore: Distributed Data Store for Scalable Training of Graph Neural Networks on Large Atomistic Modeling Datasets

Jong Youl Choi, Massimiliano Lupo Pasini, Pei Zhang, Kshitij Mehta, Frank Liu, **Jonghyun Bae**, and Khaled Ibrahim
Workshop on Machine Learning with Graphs in High Performance Computing Environments (Be held in conjunction with SC '23), Denver, CO, Nov 2023

[APSys'23] Liquid: Mix-and-Match Multiple Image Formats to Balance DNN Training Pipeline

Woohyeon Baek*, **Jonghyun Bae***, Donghyun Lee, Hyunwoong Bae, Yeonhong Park, and Jae W. Lee
(* Equally contributed)
The 14th ACM SIGOPS Asia-Pacific Workshop on Systems 2023, Seoul, Korea, August 2023

[SparkSummit'16] ggplot2.SparkR: Rebooting ggplot2 for Scalable Big Data Visualization

Jonghyun Bae, Sangoh Jeong, Wenjing Jin, and Jae W. Lee
Spark Summit East, New York, MA, December 2016

[UseR'14] RIGHT: An HTML Canvas and JavaScript-based Interactive Data Visualization Package for Linked Graphics

ChungHa Sung, **Jonghyun Bae**, SangGi Hong, TaeJoon Song, Jae W. Lee, and Junghoon Lee
The R User Conference (UseR!-Poster), Los Angeles, CA, July 2014

Skills

Languages C, C++, Java, NVIDIA CUDA, JavaScript, Python, R, Scala
Frameworks PyTorch, Caffe, Apache Spark, Apache Hadoop, Nvprof

Teaching

Big data engineering 3 (Parallel and distributed databases) ***Seoul National University, Korea***
Teaching assistant *Mar 2017 - Aug 2017*
- Training course for Apache Spark running on a distributed cluster system

Data structures and algorithms ***Sungkyunkwan Univesity, Korea***
Teaching assistant *Sep 2016 - Dec 2016*
- Basic algorithmic techniques for computational problems arising frequently in applications

Logic design laboratory ***Sungkyunkwan Univesity, Korea***
Teaching assistant *Mar 2014 - June 2014, Mar 2015 - June 2015, Mar 2016 - June 2016*
- Hand-on experiments of digital logic design and implementation through FPGA using VHDL

Activities

Korea Electric Power Corporation (KEPCO) Data Science Lab *Seoul, Korea*

Invited talk *May 2022*

- A Large-Batch, High-Throughput Training System for Deep Neural Networks

Korea Computer Congress 2021 *Jeju, Korea*

Invited session: Top Conference session *June 2021*

- FlashNeuron: SSD-Enabled Large-Batch Training of Very Deep Neural Networks