Ph.D. Jinwoong Kim

CONTACT INFORMATION

Mobile: +82-10-5415-1511
E-mail: aragnom@gmail.com

RESEARCH INTERESTS

Machine Learning Platforms
NLP Inference Optimization
Large Scale Cluster Scheduling and Monitoring Systems
Parallel Multi-dimensional Indexing on the GPU
Distributed and Parallel Systems
Database Systems
Non-Volatile Memory based Logging

EXPERIENCE

Machine Learning Research Engineer at Naver Corp. *Sep 2017 - Present*
Full time (Search CIC)
Main Project 1. Build a system to serve Hundreds Billion scale Transformer-based Language Model (such as GPT3) for Real Services
- Built an ML pipeline to train and deploy GPT3 models
- Evaluated the comprehensive performance of GPT3 models and summarized the most important performance factors
- Profiling GPT3 models and deep dive into FasterTransformer (developed by Nvidia) to improve the inference performance
- Applied sparse training GPT3 models to improve the inference performance
- Deployed GPT3 model servers for real services

Main Project 2. Build kubeflow (KF) based Machine Learning Platform - AiSuite
- BuilT a KF-based platform on top of Kubernetes (k8s)
- Built a system component to serve GPT3 model servers and a pipeline to fine tune GPT3 models

Main Project 3. Develop and operate Yarn-based Machine Learning Platform - C3DL
- C3DL is Yarn-scheduled base machine learning platform. We added external components to support GPU scheduling
- Developed a monitoring system that monitors the thousands scale GPU cluster using Nvidia DCGM (Data Center GPU Managers)
- To maximize the GPU utilization, developed a GPU sharing services using Nvidia MPS so that multi-tenants can run their jobs together on the same GPU

Machine Learning Research Engineer at Naver Corp. *Sep 2017 - June 2019*
Full Time, (Clova CIC)
Main Project 1. Develop Machine Learning Platform - NSML
Our team developed NSML from the scratch, which is machine learning framework for a multi-tenant GPU cluster like k8s
NSML deploys user's TensorFlow or pyTorch jobs on GPUs
I implemented the entire scheduler and monitoring systems
Main Project 2. Build a hyper-parameter tuning systems - AutoML
- I developed AutoML framework that tunes hyper-parameter automatically on NSML
- It automatically finds idle GPU serves and runs automl jobs

Publish and Present
- Published a paper and demonstrated at Sysml2019
- Presented AutoML at Deview 2018

EDUCATION

Ms-Ph.D Program in Computer Science and Engineering *Mar 2011 - Aug 2017*
Ulsan National Institute of Science and Technology (UNIST)
Thesis Title : "Exploiting Graphics Processing Units for Massively Parallel Multi-Dimensional Indexing"
Advisor : Prof. Beomseok Nam

- Designed and implemented multi-dimensional indexing structures like R-Trees for Multi-GPU and Multi-Node Systems using CUDA language from-the-scratch.

B.S. in Computer Engineering *Mar 2005 - Feb 2011*
Chungbuk National University, South Korea

JOURNAL REVIEWING ACTIVITES

1. IEEE Transactions on Computers (TC), 2020
2. ACM Transactions on Storage (TOS), 2019
3. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2018

ACADEMIC/ RESEARCH EXPERIENCE

Research Intern
Carnegie Mellon University, Pittsburgh, PA, USA(Prof. Andrew Pavlo) *May - Aug, 2015*
Designed and developed the in-memory database system, *Peloton*
- Implemented the Catalog, DDL, DML, Bootstrap, Logging, etc.
- Designed the non-volatile memory based logging, *Write-Behind Logging*, VLDB, 2016.

Visiting PhD Student
University of California Berkeley, Berkeley, CA, USA(Prof. Ikhlaq Sidhu) *Nov 2015 - Jan 2016*

Research Assistant
Data Intensive Computing Lab, UNIST *Spring 2011 - Fall 2017*

- Designed and implemented parallel indexing schemes for multi-dimensional range query processing on the GPU
- Implemented a distributed semantic caching framework for MapReduce
- Worked on multi-dimensional query processing with distributed cache infrastructure in cloud environments
- Implemented non-volatile memory based *Heap manager* to improve logging performance for SQLite

Korea Institute of Science and Technology Information (KISTI) *May 2015 - Nov 2015*

- Designed and implemented the GPU-based multi-dimensional indexing for *GLOVE*

LG Electronics (CTO) *Jan 2015 - Jun 2015*

- Worked on multi-thread query processing for SQLite.

Teaching Assistant

- TA for Prof. Beomseok Nam, Object-Oriented Programming *Spring 2014*
- TA for Prof. Beomseok Nam, Advanced Programming *Fall 2013*
- TA for Prof. Beomseok Nam, Introduction to Database Systems *Winter 2012*
- TA for Prof. Tsz-Chiu Au, Engineering Programming *Fall 2012*

- TA for Prof. Young-ri Choi, Engineering Programming *Spring 2012*
- TA for Prof. Beomseok Nam, Engineering Programming *Fall 2011*

HONORS AND AWARDS

Naver WoW Project 2nd Prize, Naver Corp. *2019*
Naver PhD Fellowship, Naver Corp. *2016*
Prof. Ram Kumar Fellowships at ICDE, Ramkumar Foundation. *2015*
Merit-based Scholarship, Chungbuk National University *Spring 2009*
Merit-based Scholarship, Chungbuk National University *Fall 2008*

PUBLICATIONS AND PREPRINTS

16 Heungseok Park, **Jinwoong Kim**, Minkyu Kim, Ji-Hoon Kim, Jaegul Choo, Jung-Woo Ha, and Nako Sung "VisualHyperTuner: Visual Analytics for User-Driven Hyperparameter Tuning of Deep Neural Networks", 2nd MLSys Conference **(MLSys)**, Demo, Stanford, CA, USA, Apr. 2019.

15 **Jinwoong Kim**, Minkyu Kim, Heungseok Park, Ernar Kusdavletov, Dongjun Lee, Adrian Kim, Ji-Hoon Kim, Jung-Woo Ha, and Nako Sung "CHOPT : Automated Hyperparameter Optimization Framework for Cloud-Based Machine Learning Platforms", arXiv:1810.03527, 2018

14 Hanjoo Kim, Minkyu Kim, Dongjoo Seo, **Jinwoong Kim**, Heungseok Park, Soeun Park, Hyunwoo Jo, KyungHyun Kim, Youngil Yang, Youngkwan Kim, Nako Sung, and Jung-Woo Ha "NSML: Meet the MLaaS platform with a real-world case study", arXiv:1810.09957, 2018

13 **Jinwoong Kim** and Beomseok Nam, "Co-Processing Heterogeneous Parallel Index for Multi-Dimensional Datasets", Journal of Parallel and Distributed Computing**(JPDC)**, Vol. 113, pp 195-203, Mar. 2018.

12 Nako Sung, Minkyu Kim, HyunWoo Jo, Youngil Yang, **Jinwoong Kim**, Leonard Lausen, Youngkwan Kim, Gayoung Lee, Donghyun Kwak, Jung-Woo Ha, and Sung Kim, "NSML: A Machine Learning Platform That Enables You to Focus on Your Models", ML System Workshop at NIPS, 2017

11 Wook-Hee Kim, Jihye Seo, **Jinwoong Kim**, and Beomseok Nam, "clfB-tree: Cacheline Friendly Persistent B-tree for NVRAM", To appear in ACM Transactions on Storage**(TOS)**, Special issue on NVM and Storage, 2017.

10 Moohyeon Nam, **Jinwoong Kim**, Beomseok Nam "Parallel Tree Traversal for Nearest Neighbor Query on the GPU" 45th International Conference on Parallel Processing **(ICPP)**, Philadelphia, PA, USA, Aug. 2016.

9 Wookhee Kim, **Jinwoong Kim**, Woongki Baek, Beomseok Nam and Youjip Won "NVWAL: Exploiting NVRAM in Write-Ahead-Logging " 21st International Conference on Architectural Support for Programming Languages and Operating Systems **(ASPLOS)**, Atlanta, GA, USA, Apr. 2016

8 **Jinwoong Kim**, Sehoon Lee, Joong-Youn Lee, Beomseok Nam, Min Ah Kim "GLOVE: An Interactive Visualization Service Framework with Multi-Dimensional Indexing on the GPU " 27th International Conference for High Performance Computing, Networking, Storage and Analysis**(SC)**, Poster, Austin, TX, Nov. 2015.

7 **Jinwoong Kim**, Won-Ki Jeong, and Beomseok Nam "Exploiting Massive Parallelism for Indexing Scientific Datasets on the GPU" IEEE Transactions on Parallel and Distributed Systems**(TPDS)**, Vol. 26, No. 8, pp 2258-2271, Aug. 2015. **(Selected as the featured paper of Aug. 2015 issue)**

6 Youngmoon Eom, **Jinwoong Kim**, and Beomseok Nam "Multi-dimensional Multiple Query Scheduling with Distributed Semantic Caching Framework" Cluster Computing, Vol. 18, No. 3, pp 1141-1156, Springer, Jun. 2015.

5 Youngmoon Eom, **Jinwoong Kim**, Deukyeon Hwang, Jaewon Kwak, Minho Shin, Beomseok Nam "Improving Multi-dimensional Query Processing with Data Migration in Distributed Cache Infrastructure", 21st IEEE International Conference on High Performance Computing (HiPC 2014)(23% a/r). Goa, India, Dec. 2014.

4 Youngmoon Eom, Jonghwan Moon, **Jinwoong Kim**, Beomseok Nam, "Collaborative Multi-dimensional Dataset Processing with Distributed Cache Infrastructure in the Cloud", 2nd International Workshop on Autonomic Management of Grid and Cloud Computing (AMGCC'14) (in conjunction with IEEE CAC 2014), London, UK, Sep. 2014.

3 **Jinwoong Kim**, Sul-Gi Kim, Beomseok Nam, "Parallel Multi-dimensional Range Query Processing with R-Trees on GPU", Journal of Parallel and Distributed Computing **(JPDC)**, Vol. 73, Issue 8, 1195-1207, Elsevier, Aug, 2013.

2 Beomseok Nam, Deukyeon Hwang, **Jinwoong Kim**, and Minho Shin, "High-Throughput Distributed Query Scheduling with EMA-based Statistical Prediction", Special Issue on Data Intensive eScience, Distributed and Parallel Databases, Vol. 30, issue 5–6, pp 401-414, Springer, Jun. 2012.

1 **Jinwoong Kim**, Sumin Hong, and Beomseok Nam "A Performance Study of Traversing Spatial Indexing Structures in Parallel on GPU", 3rd International Workshop on Frontier of GPU Computing (FGC), in conjunction with HPCC 2012, Liverpool, UK, Jun. 2012.

INVITED TALKS

6 "Experience of serving large-scale natural language processing models", DEVIEW, South Korea, Nov. 2021.

5 "Struggles to Serve GPT3 without Knowledge of NLP", Naver Europe Workshop, Apr. 2021.

4 Naver AI Techtalk, UNIST, South Korea, Dec. 2019.

3 "NSML: Machine learning as a platform & Automize Model Tuning" DEVIEW, South Korea, Oct. 2018.

2 "NSML: A Machine Learning Platform That Enables You to Focus on Your Models" The 9th Asian Conference on Machine Learning **(ACML)**, Poster, South Korea, Nov. 2017.

1 "Exploiting Massive Parallelism for Indexing Multi-dimensional Datasets on the GPU" Parallel Data Lab **(PDL)** at CMU, Pittsburgh, PA, USA, Aug. 2015.

SKILLS

- Programming Languages :
  C/C++/C#, Python, CUDA, Go

- Libraries, and Knowledge:
  SQL, MPI, Docker, Nvidia MPS, DCGM, GPU Profiling with Nsight, Ncompute
  Linux, Git, LaTex, ELK, Yarn, Ansible, Zookeeper, Kubernetes

REFERENCES

- **Beomseok Nam**
  Associate professor
  College of Software
  Sungkyunkwan University, Suwon, South Korea
  E-mail: bnam@skku.edu

- **Won-Ki Jeong**
  Professor
  Department of Computer Science and Engineering
  Korea University, Seoul, South Korea

E-mail: wkjeong@korea.ac.kr

- **Woongki Baek**
  Associate Professor
  Department of Computer Science and Engineering
  and the Graduate School of Artificial Intelligence
  at Ulsan National Institute of Science and Technology (UNIST)
  Ulsan, South Korea
  E-mail: wbaek@unist.ac.kr

- **Andrew Pavlo**
  Associate Professor
  Department of Computer Science
  Carnegie Mellon University, Pittsburgh, USA
  E-mail: pavlo@cs.cmu.edu

Last updated on **December 9, 2021**