Tung D. Le

Research Scientist
IBM Research - Tokyo
http://www.ibm.biz/leductung

research interest

My main interest lies in the intersection of parallel programming, compiler, and deep learning. In particular, I would love to propose systematic optimizations for parallel programming and AI systems.

academic activities

- **ACM member since July 2016**
- **Associate editor for IEICE transactions on Information and Systems (June 2020 - June 2024)**
- PC members: ScalCom (**2018**, **2019**, 2020, 2021)

professional experience

- 1/2017 - now: research staff member at IBM Research - Tokyo, Japan.
- 4/2016 - 12/2016: postdoctoral researcher at IBM Research - Tokyo, Japan.
- 4/2013 - 3/2016: Ph.D. student at National Informatics Institue, Japan.
    - Topic: A systematic approach to regular-expression-based queries on big graphs
    - Supervisor: Prof. **Zhenjiang Hu**.
- 10/2012 - 3/2013: internship at National Informatics Institue, Japan.
    - Topic: Systematic parallel programming with MapReduce programming model.
    - Supervisor: Prof. **Zhenjiang Hu**.
- 4/2010 - 9/2010: internship at National Informatics Institue, Japan.
    - Topic: Skeleton parallel programming for Hierarchically Tiled Arrays.
    - Supervisor: Prof. **Zhenjiang Hu**.
- 10/2008 - 9/2010: M.Sc. student at **HUST**.
    - Topic: Skeleton parallel programming for Hierarchically Tiled Arrays.
    - Supervisors: **Dr. Huu-Duc Nguyen**
- 7/2007: two-week internship at **The San Diego Supercomputer Center, UCSD**.
- 6/2007 - 9/2012: Researcher at The High Peformance Computing Center, **HUST**.

education

When	What (in computer science) @ Where	With whom
2013 - 2016	Ph.D. @ **SOKENDAI/NII, Japan**	**Prof. Zhenjiang Hu**
2008 - 2010	M.Sc. @ **HUST, Vietnam**	**Dr. Huu-Duc Nguyen**
2002 - 2007	B.S. @ **HUST, Vietnam**	**Prof. Thanh-Thuy Nguyen**

awards

- The IBM Outstanding Technical Achievement Award (OTAA) in 2018, 2019, and 2020.
- The IBM Open Source Recognition Award in 2018 (leader level), 2019 (contributor level), and 2020 (contributor level).
- **Hitachi Scholarship for Ph.D. courses** (October 2012 - April 2016). I am the only person from **HUST, Vietnam** ever who gets the scholarship.
- PLMW Scholarship 2015 to attend **PLMW2015** and **POPL2015**.
- **ICDE2015 Student Travel Award**.

projects

- 2019 - Now: **ONNX Models in MLIR Compiler Infrastructure**
- 2016 - 2019: **IBM PowerAI**
  - Optimized deep learning frameworks Caffe, Chainer on POWER machines.
  - Proposed large model support for tensorflow (TFLMS). Open source was published at **Graph-based Large Model Support for TensorFlow**

patents

granted patents

1. **Tung D. Le**, Imai Haruki, and Yasushi Negishi. *Efficient parallel training of a network model on multiple graphics processing units*. (Mar. 2021). US Patent No. US 10949746 B2, Filed Feb. 3, 2017, Granted Mar. 16, 2021.

2. **Tung D. Le**, Haruki Imai, Yasushi Negishi, and Kiyokuni Kawachiya. *Graph rewriting for large model support using categorized topological sort*. (Jan. 2021) US Patent No. 10884755 B1, Filed Jul . 31, 2019, Granted Jan. 5, 2021.

3. Taro Sekiyama, Kiyokuni Kawachiya, **Tung D. Le**, Yasushi Negishi. *Real-time resource usage reduction in artificial neural networks*. (Feb. 2020). US Patent No. 10558914 B2, Filed Apr. 16, 2019, Granted Feb. 11, 2020.

filed patents

1. Haruki Imai, **Tung D. Le**, Yasushi Negishi, Kiyokuni Kawachiya. *Data swapping for neural network memory conservation*. US Patent App. 17/089245, Filed Nov. 4, 2020.

2. Yasushi Negishi, **Tung D. Le**, Haruki Imai, Kiyokuni Kawachiya. *ReLU compression to reduce gpu memory*. US Patent App. 17/085196, Filed Oct. 30, 2020.

3. **Tung D. Le**. *Neural programmer interpreters with modeled primitives*. US Patent App. 16/514528, Filed Jul. 17, 2019.

4. Gradus Janssen, Vladimir Zolotov, and **Tung D. Le**. *Neural network training using a data flow graph and dynamic memory management*. US Patent App. 16/704240, Filed Dec. 5, 2019.

5. Yasushi Negishi, Haruki Imai, Taro Sekiyama, **Tung D. Le**, Kiyokuni Kawachiya. *Mechanism for choosing execution mode for large neural network*. US Patent App. 16/018680, Filed Jun. 26, 2018.

6. Taro Sekiyama, **Tung D. Le**, Kun Zhao. *Optimizing tree-based convolutional neural networks*. US Patent App. 15/617737, Filed Jun. 8, 2017.

7. Kiyokuni Kawachiya, **Tung D. Le**, Yasushi Negishi. *Balancing memory consumption of multiple graphics processing units in deep learning*. US Patent App. 15/604542, Filed May. 24, 2017.

8. **Tung D. Le**, Haruki Imai, Taro Sekiyama, Yasushi Negishi. *Multi-gpu deep learning using cpus*. US Patent App. 15/843244, Filed Dec. 15, 2017.

9. Taro Sekiyama, **Tung D. Le**. *Localizing tree-based convolutional neural networks*. US Patent App. 15/815771, Filed Nov. 17, 2017.

conference/journal papers

2020

1. **Tung D. Le**, Gheorghe-Teodor Bercea, Tong Chen, Alexandre E Eichenberger, Haruki Imai, Tian Jin, Kiyokuni Kawachiya, Yasushi Negishi, Kevin O'Brien. 2020. *Compiling ONNX Neural Network Models Using MLIR*. arXiv:2008.08272, Retrieved from https://arxiv.org/abs/2008.08272v1

2. Haruki Imai, **Tung D. Le**, Yasushi Negishi, and Kiyokuni Kawachiya. 2020. *Acceleration of large deep learning training with hybrid GPU memory management of swapping and re-computing*. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), December 10-13, 2020, Atlanta, GA, USA. IEEE, 1111-1116.

2019

1. Haruki Imai, Samuel Matzek, **Tung D. Le**, Yasushi Negishi, Kiyokuni Kawachiya. 2019. *High Resolution Medical Image Segmentation Using Data-Swapping Method*. In: Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, Vol. 11766. Springer, Cham.

2. **Tung D. Le**, Haruki Imai, Yasushi Negishi, Kiyokuni Kawachiya. 2019. *Automatic GPU Memory Management for Large Neural Models in TensorFlow*. In Proceedings of the 2019 ACM SIGPLAN International Symposium on Memory Management (ISMM 2019), June 2019, Phoenix, Arizona, USA. Association for Computing Machinery, New York, NY, USA, 1-13.

3. G. Janssen, V. Zolotov and **Tung D. Le**. 2019. *Large Data Flow Graphs in Limited GPU Memory*. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), December 3-12, 2019, Los Angeles, CA, USA. IEEE, 1821-1830.

4. Yuki Ito, Haruki Imai, **Tung Le Duc**, Yasushi Negishi, Kiyokuni Kawachiya, Ryo Matsumiya, and Toshio Endo. 2019. *Profiling based out-of-core hybrid method for large neural networks.* In Proceedings of the 24th Symposium on Principles and Practice of Parallel Programming (PPoPP '19), February 16-20, 2019, Washington, DC, USA. Association for Computing Machinery, New York, NY, USA, 399—400.

2018

1. Minsik Cho, **Tung D. Le**, Ulrich A Finkler, Haruki Imai, Yasushi Negishi, Taro Sekiyama, Saritha Vinod, Vladimir Zolotov, Kiyokuni Kawachiya, David S. Kung, Hillery C. Hunter. 2018. *Large Model Support for Deep Learning in Caffe and Chainer*. 2018 SysML conference.

2. **Tung D. Le**, Haruki Imai, Yasushi Negishi, Kiyokuni Kawachiya. 2018. *TFLMS: Large Model Support in TensorFlow by Graph Rewriting.* arXiv:1807.02037. Retrieved from https://arxiv.org/abs/1807.02037

3. **Tung D. Le**, Taro Sekiyama, Yasushi Negishi, Haruki Imai, Kiyokuni Kawachiya. 2018. *Involving CPUs into Multi-GPU deep learning*. In Proceddings of the 2018 ACM/SPEC International Conference on Performance Engineering (ICPE '18), March 2018, Berlin, Germany. Association for Computing Machinery, New York, NY, USA, 56—67.

2017

1. **Tung D. Le**, Taro Sekiyama, Yasushi Negishi, Haruki Imai, Kiyokuni Kawachiya. 2017. *Accelerating Multi-GPU Deep Learning by Collecting and Accumulating Gradients on CPUs*. SIG Technical Reports 2017-HPC-159(8), 1-8.

2016

1. **Le-Duc Tung**, Zhenjiang Hu. *Towards Systematic Parallelization of Graph Transformations over Pregel*. International Journal of Parallel Program. 45, 2 (April 2017), 320—339.

2. Chong Li, **Le-Duc Tung**, Xiaodong Meng, Zhenjiang Hu. 2016. *Derivation of parallel-efficient structural recursive functions from declarative graph queries*. In Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC ’16), April 2016, Pisa, Italy. Association for Computing Machinery, New York, NY, USA, 1922–1925.

2015

1. **Le-Duc Tung**, Zhenjiang Hu. 2015. *Towards Systematic Parallelization of Graph Trans- formations over Pregel*. In Proceedings of the 8th International Symposium on High-level Parallel Programming and Applications (HLPP 2015), July 2-3, 2015, Pisa, Italy.

2. **Le-Duc Tung**, Zhenjiang Hu. *Pregel meets UnCAL: a Systematic Framework for Transforming Big Graphs*. In Proceedings of the 2015 31st International Conference on Data Engineering Workshops (ICDE2015), April 13-17, 2015, Seoul, South Korea. IEEE, 250-254.

2013

1. **Le-Duc Tung**, Nguyen-Van Quyet, Zhenjiang Hu. 2013. *Efficient Query Evaluation on Distributed Graphs with Hadoop Environment*. In Proceedings of the Fourth International Symposium on Information and Communication Technology (SoICT ’13), December 5-6, 2013, Da Nang, Vietnam. Association for Computing Machinery, New York, NY, USA, 311–319,

2. Nguyen-Van Quyet, **Le-Duc Tung**, Zhenjiang Hu. *Minimizing Data Transfers for Regular Reachability Queries on Distributed Graphs*. In Proceedings of the Fourth International Symposium on Information and Communication Technology (SoICT ’13), December 5-6, 2013, Da Nang, Vietnam. Association for Computing Machinery, New York, NY, USA, 325-334.

2012

1. **D. T. Le**, H. D. Nguyen, T. A. Pham, H. H. Ngo and M. T. Nguyen. 2012. *An Intermediate Library for Multi-GPUs Computing Skeletons*. In Proceedings of the 2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF ’12), March 2012, Ho Chi Minh, Vietnam. IEEE, 1-6.

contact

Office address: **10th floor, tower side, IBM Japan headquarters building**.
Telephone number: +81-(3)-3808-5228
E-mail: tung@jp.ibm.com