Kelvin K.W. Ng

Email: nkwkelvin@gmail.com
Phone: (267) 290-6113
LinkedIn: kelvin-kai-wing-ng
GitHub: github.com/Kelvin-Ng
Website: kelvin-ng.github.io

Research Focus

ML systems researcher in accelerator–network co-design for distributed training and inference, with experience in asynchronous runtimes, fine-grained scheduling, device-direct networking, collective communication, computation–communication overlap, compiler–runtime co-design, and data-center network simulation.

Education

University of Pennsylvania — Philadelphia, PA, USA
Ph.D. in Computer and Information Science, GPA: 4/4, Advisor: Vincent Liu — Aug 2019–May 2025

- Thesis: Resource Sharing for Machine Learning Serving

The Chinese University of Hong Kong — Hong Kong
B.Sc. in Computer Science, GPA: 3.749/4, Major GPA: 3.817/4 — Sep 2015–Jul 2019

- Thesis: Efficient Similarity Search with Configurable Probabilistic Recall Guarantees

Experience

Amazon Web Services — Cupertino, CA, USA
Applied Scientist — May 2025–Present

- **Asynchronous runtime:** Designed and implemented a graph-based runtime spanning device operations—memory transfers, kernel execution, and collective communication—and host-side CPU tasks; the runtime enforces complex dependencies while automatically overlapping independent work.
- **Fine-grained on-device scheduling:** Designed and implemented a microcode-level cooperative scheduler that interleaves multiple jobs at instruction granularity, filling otherwise-idle accelerator engines when one job stalls.
- **Device-direct Trainium–EFA communication:** Designing a flexible, transparent path for Trainium accelerators to use Elastic Fabric Adapter (EFA) directly for cross-host collective communication, bypassing the host CPU and supporting fine-grained computation–communication co-optimization for expert-parallel workloads such as DeepEP.

University of Pennsylvania — Philadelphia, PA, USA
Graduate Research Assistant — Aug 2019–May 2025

- **ParaFlex (SoCC '25):** Developed a heterogeneous LLM serving system that aligns pipeline-stage execution times across models with different sizes and parallelism configurations, enabling efficient multiplexing on shared GPUs.
- **Paella (SOSP '23):** Designed a cross-layer system spanning the compiler, client, and software-defined GPU scheduler to control kernel execution order, enabling flexible application-level scheduling and efficient GPU sharing.
- **MimicNet (SIGCOMM '21):** Built an ML-based approximate network simulator that retains packet-level visibility where needed while approximating the rest of a data-center network, achieving over two orders of magnitude speedup with low error.
- **Shared-parameter model serving:** Designed a system that jointly generates sharding, placement, and scheduling decisions while accounting for parameters shared across workloads, such as fine-tuned model variants.

ByteDance Bellevue, WA, USA
Research Scientist Intern May–Aug 2024

- Developed efficient in-place gradient compression techniques for collective communication in distributed training.
- Designed topology-aware, adaptive collective communication techniques for distributed training and inference.

The Chinese University of Hong Kong Hong Kong
Undergraduate Research Assistant May 2016–Jul 2019

- **Pyramid (BigData '19):** Developed a distributed, HNSW-based similarity search framework that partitions similar items across nodes and selectively routes each query, achieving high throughput and low latency while remaining robust to node failures and stragglers.
- **HSQ:** Developed vector-quantized gradient compression for communication-efficient federated learning.

Publications

[1] T. Luo, **K. K. W. Ng**, Z. P. Khor, S. Sankhe, B. T. Loo, and V. Liu, "Multiplexed heterogeneous LLM serving via stage-aligned parallelism", in *Proceedings of the 2025 ACM Symposium on Cloud Computing*, 2025.

[2] **K. K. W. Ng**, H. M. Demoulin, and V. Liu, "Paella: Low-latency model serving with software-defined GPU scheduling", in *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023.

[3] Q. Zhang, **K. K. W. Ng**, C. Kazer, S. Yan, J. Sedoc, and V. Liu, "MimicNet: Fast performance estimates for data center networks with machine learning", in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, 2021.

[4] X. Dai, X. Yan, **K. K. Ng**, J. Liu, and J. Cheng, "Norm-explicit quantization: Improving vector quantization for maximum inner product search", in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI'20)*, 2020.

[5] X. Dai, X. Yan, K. Zhou, H. Yang, **K. K. Ng**, J. Cheng, and Y. Fan, *Hyper-Sphere Quantization: Communication-efficient SGD for federated learning*, 2019. arXiv: 1911.04655 [cs.LG].

[6] S. Deng, X. Yan, **K. K. Ng**, C. Jiang, and J. Cheng, "Pyramid: A general framework for distributed similarity search on large-scale datasets", in *Proceedings of the 2019 IEEE International Conference on Big Data (BigData'19)*, 2019.

[7] C. W. Kazer, J. Sedoc, **K. K. Ng**, V. Liu, and L. H. Ungar, "Fast network simulation through approximation or: How blind men can describe elephants", in *Proceedings of the 17th ACM Workshop on Hot Topics in Networks (HotNets'18)*, 2018.

[8] J. Li, X. Yan, J. Zhang, A. Xu, J. Cheng, J. Liu, **K. K. Ng**, and T.-c. Cheng, "A general and efficient querying method for learning to hash", in *Proceedings of the 2018 International Conference on Management of Data (SIGMOD'18)*, 2018.

[9] F. Shang, Y. Liu, K. Zhou, J. Cheng, **K. K. Ng**, and Y. Yoshida, "Guaranteed sufficient decrease for stochastic variance reduced gradient optimization", in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (AAAI'18)*, 2018.

Skills

- **Programming:** C++, CUDA, Python
- **ML systems:** PyTorch, vLLM, TVM, JAX, XLA, TensorFlow
- **Tools:** OMNeT++, Docker, Kubernetes, Flink, Spark, Hadoop, YARN

Professional Service

Technical Program Committee, ENC 2025

Languages

- Cantonese (native); English and Mandarin (fluent)