Lingyun Yang 杨凌云

✉ lyangbk@cse.ust.hk | 🌐 lingyunyang.com | 📖 Reading Notes | ☎ (+86) 135-0284-6103 | 📍 Hong Kong

Education

Hong Kong University of Science and Technology **Hong Kong, China**
Ph.D. in Computer Science and Engineering, Advisor: **Wei Wang** Sep. 2020 – Present

- Research Domain: Cluster Management, Model Serving Systems, Resource Scheduling

South China University of Technology **Guangzhou, China**
B.Eng. in Computer Science and Technology, Elite Class (GPA 3.82/4) Sep. 2016 – Jul. 2020

- National Scholarship, The First Prize Scholarship

Research and Selected Projects

I have a broad interest in resource management for large-scale data centers / AI infrastructure. Specifically, my research focuses on: (a) improving resource efficiency for AI/GPU clusters; (b) building efficient and low-cost AI model serving systems.

SwiftDiffusion: Efficient Diffusion Model Serving with Add-on Modules (review, 2407.02031)
***Lingyun Yang**, Suyi Li, Xiaoxiao Jiang, Hanfeng Lu, Zhipeng Di, Weiyi Lu, Jiawei Chen, Kan Liu, Yinghao Yu, Tao Lan, Guodong Yang, Lin Qu, Liping Zhang, Wei Wang*

- Built an efficient text-to-image system that generates images with diffusion models and add-on modules (i.e., ControlNets and LoRAs).
- Incorporated system designs, e.g., ControlNet-as-a-Service, async bounded LoRA loading, latent parallelelism for CFG computation.
- Achieved up to **7.8×** in serving latency and **1.6×** in throughput without compromising image quality.

Prism: GPU-Disaggregated Serving for Deep Learning Recommendation Models at Scale (review)
***Lingyun Yang**, Yongchen Wang, Yinghao Yu, Qizhen Weng, Jianbo Dong, Kan Liu, Chi Zhang, Yanyi Zi, Hao Li, Zechao Zhang, Nan Wang, Yu Dong, Menglei Zheng, Lanlan Xi, Xiaowei Lu, Liang Ye, Guodong Yang, Binzhang Fu, Tao Lan, Liping Zhang, Lin Qu, Wei Wang*

- Proposed a GPU-disaggregated DLRM serving system to eliminate resource mismatch and meet elastic demand; leveraged RDMA network to offload computation on separate GPU and CPU nodes; resource-aware graph partitioning; topology-aware scheduling.
- In daily scenarios (e.g., a crowded GPU cluster with > 90% allocation rate), reduced CPU fragments by **53%** and GPU fragments by **27%**; In the Double 11 Shopping Festival, saved up to **90%** of GPUs when loaning GPU servers from training clusters.

Beware of Fragmentation: Scheduling GPU-Sharing Workloads with Fragmentation Gradient Descent (ATC'23)
***Lingyun Yang**, Qizhen Weng, Yinghao Yu, Wei Wang, Xiaochuan Tang, Guodong Yang, Liping Zhang* hkust-adsl/kubernetes-scheduler-simulator

- Formally quantified **statistical** GPU resource fragments in shared GPU clusters.
- Proposed the fragmentation gradient descent (FGD) scheduling algorithm to reduce resource fragmentation.
- Reduced unallocated GPUs by up to **49%** compared to state-of-the-art scheduling policies.

Large-Scale GPU Sharing and Overcommitment in Production

- Enabled large-scale GPU sharing in production clusters, with over **10k** shared GPU containers running daily.
- Support the co-location of GPU tasks with different priorities (e.g., latency-sensitive, best-effort).
- Designed and implemented the node-level agent and the cluster-level controller.

Morphling: Fast, Near-Optimal Auto-Configuration for Cloud-Native Model Serving (SoCC'21)
***Lingyun Yang**, Luping Wang, Yinghao Yu, Wei Wang, Bo Li, Xianchao Sun, Jian He, and Liping Zhang* kubedl-io/morphling

- An auto-configuration framework for AI serving on Kubernetes; part of Alibaba's open-sourced KubeDL, a CNCF sandbox project.
- Combined meta-learning and bayesian optimization to quickly find the **optimal** resource configuration (e.g., CPU cores, GPU memory, GPU timeshare, GPU type) and runtime parameters (i.e., batch size).

Work Experience

Alibaba Group **Hangzhou, China**
Research Intern, Cluster Management Group, Mentor: **Yinghao Yu** Dec. 2020 – Present

- Conducted several research projects including Morphling, GPU sharing, FGD, Prism, SwiftDiffusion (details as mentioned above).

Microsoft Research Asia **Beijing, China**
Research Intern, Innovation Engineering Group Jul. 2019 – Jun. 2020

- Conducted research on model robustness of face recognition. Star of Tomorrow Internship Award of Excellence.

Technical Skills

- **Programming:** Golang, C++, Python, JavaScript, asynchronous, multithread, multiprocess, distributed, RDMA
- **Machine Learning:** PyTorch, TensorFlow, Numpy, Matplotlib, HuggingFace
- **Full Stack:** Web Frontend, Backend, SQL, Grafana, Docker, Kubernetes, Git, CI/CD