DESEN SUN

Email: desen.sun@uwaterloo.ca | Tel: 6479657428 | Address: Waterloo, Canada

Education

University of Waterloo *Waterloo, Sept. 2024-Jun. 2028*

Beihang University *Beijing, Sept. 2020-Jun. 2023*

University of Electronic Science and Technology of China *Chengdu, Sept. 2016-Jun. 2020*

PUBLICATIONS

Proceeding Papers

1. (**Usenix Security'26**) Breaking Diffusion with Cache: Exploiting Approximate Caches in Diffusion Models.
   Desen Sun, Shuncheng Jie, Sihang Liu.
   USENIX Security Symposium, 2026

2. **(PPoPP'26)** MixFusion: A Patch-Level Parallel Serving System for Mixed-Resolution Diffusion Models.
   Desen Sun, Zepeng zhao, Yuke Wang.
   ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, 2026

3. (**Sigmetrics'26**) EmbAdvisor: Adaptive Cache Management for Sustainable LLM Serving.
   Yuyang Tian, **Desen Sun,** Yi Ding, Sihang Liu.
   ACM SIGMETRICS, 2026

4. (**PPoPP'23**) PiPAD: Pipelined and Parallel Dynamic GNN Training on GPUs.
   Chunyang Wang*, **Desen Sun*,** Yuebin Bai. (***: Co-first author**)
   ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, 2023

5. (**TPDS'23**) CD-MSA: Cooperative and Deadline-Aware Scheduling for Efficient Multi-Tenancy on DNN Accelerators.
   Chunyang Wang, Yuebin Bai, **Desen Sun**.
   IEEE Transactions on Parallel and Distributed Systems, 2023

Preprint Papers

6. FlexCache: Flexible approximate cache system for video diffusion.
   Desen Sun, Yuyang Tian, Sihang Liu.
   Under submission

RESEARCH EXPERIENCE

Breaking Diffusion with Cache: Exploiting Approximate Caches in Diffusion Models (Published on Usenix Sec'26)

1. **Background.** To reduce the excessive computation overhead from Text to Image diffusion models, some companies have leveraged approximate cache to achieve higher throughput. Although this technique brings sufficient performance benefits, it leaks some extra channels that may be exploited by attackers.
2. **Vulnerabilities**. I propose three potential risks: Covert Channel, Prompt Stealing and Poison Attack. The attacker may exploit the latency difference and output similarities to determine whether one prompt hit one specific cache. And an attack can reconstruct one prompt by probing multiple prompts to the cache. The attack can also inject caches with specific logos and the logos will be shown up in normal users' output.
3. **Achievement.** Our evaluation shows that we succeed in all three attacks, demonstrating the new security concern of the approximate cache technique.

EmbAdvisor: Adaptive Cache Management for Sustainable LLM Serving (Published on Sigmetrics'26)

1. **Background**. The carbon emission has become the key concern in the wide use of AI services. Caching further exacerbates significantly embodied carbon emissions. Simply reducing the storage may violate the SLO constraints.
2. **My task**. I Propose that the system should set the SLO as the constraint, and QPS is another metric that affects the SLO, influencing the storage decision. Moreover, I suggest to add the timeline of carbon emission experiment. Finally, I write the first draft of the paper and revise it multiple times until it gets accepted.
3. **Achievement**. The evaluation shows that our system reduces 15.1% carbon emissions with the SLO constraints.

FlexCache: Flexible approximate cache system for video diffusion

1. **Background.** The approximate cache technique has proven to improve the throughput of image generation serving system, but simply adopting it to video generation works bad due to heavy storage overhead and low hit rate
2. **Solution.** I first propose two-level compression method that reduces the storage footprint to occupy more caches. I also design a background-object decouple scheme to improve the hit rate and computation savings. A cache replacement policy is adopted to be aware of the diverse storage cost.
3. **Achievement.** Our evaluation shows that FlexCache can achieve **1.26×** higher throughput and **25%** lower cost compared to the state-of-the-art approximate cache system.

MixFusion: A Patch-Level Parallel Serving System for Mixed-Resolution Diffusion Models (Published on PPoPP'26)

1. **Background.** Diffusion model has been a mature solution to generate images. However, users tend to generate images with different resolutions. Since such requests with different resolution also differ in input and output, they are hard to be combined into one batch, which result in the waste of the GPU's resource..
2. **Solution.** I separate the images into patches to enable batching scheme among different resolutions of images. A CUDA kernel is developed to exchange the boundaries of patches to maintain quality. I also design the patch-level cache reuse mechanism to save more computations. An SLO-aware scheduler takes charge of maximizing service quality.
3. **Achievement**: We can achieve **30%** higher SLO compared to the SOTA diffusion serving system while not hurting the quality.

PiPAD: Pipelined and Parallel Dynamic GNN Training on GPUs (Published on PPoPP'23)

1. **Background**. Although current researchers have made the training of basic GNN efficiently, the temporal GNN still faces the lack of load-balance among thread blocks since it's hard to reorder every discrete graph. It costs too much time. On the other hand, they failed to utilize the common part within the adjacent two graphs.
2. **Solution**. I extract the common part of two adjacent graphs to minimize redundant computations. I also develop a CUDA kernel that ensure the load balance of the GPU threads. The kernel also effectively exploits the GPU memory bandwidth and the computation locality.
3. **Achievement**: Our evaluation shows that PiPAD can achieve **1.22 – 9.57**× speedup over the state-of-the-art DGNN frameworks on three representative models.

CD-MSA: Cooperative and Deadline-Aware Scheduling for Efficient Multi-Tenancy on DNN Accelerators (Published on TPDS'23)

1. **Background.** With the development of deep neural network (DNN), DNN accelerator keeps updating continuously, whereas the current accelerator still has the problem of low processing efficiency. Specially, assigning all systolic arrays to a single request degrades SLO satisfaction by delaying requests.
2. **My Task.** I take charge of constructing the accurate and fine-grained DNN implementation analysis model to simulate the performance of existing TPU-like accelerators. Furthermore, I also complete the programming task of the schedulability analysis.
3. **Achievement:** CD-MSA is able to increase the latency-bounded throughput, SLA satisfaction rate and weighted system throughput for 61%, 63% and 34%, respectively.

Work Experience

Alibaba, Beijing, Supervised by Huimin Yi *Jul. 2023-Aug. 2024*

I have worked at the large model training group in Alibaba(the biggest E-commerce company in China, kind of like Amazon) as an engineer for my gap year.

During my last job, I mainly focused on optimizing the XLA compiler to decrease the memory consumption. I reduced the memory footprint by **20%** with the XLA compiler, which achieves **13%** performance improvements. I also help build the efficient training framework for recommendation models.

Professional Services

Artifact Evaluation Committee for PPoPP'26

Mentor for UR2PhD program