

# Research Statement

Junxue ZHANG, USTC

Datacenter Networking (DCN) is one of the key components in datacenters<sup>1</sup> and is crucial for the performance and stability of the applications deployed in the datacenters, such as Generative AI, serverless computing, *etc.* As a DCN researcher, I have been answering the question of *How to build high-performance, intelligent and secure DCN* for the past five years.

Answering this question is challenging due to the highly *shared* environment of the datacenter at that time, which results in irregular, unpredictable and distrustful networking communications.

**Irregularity:** The irregularities problem arises from different applications having diverse communication goals (*e.g.*, throughput-sensitive or latency-sensitive) and patterns (*e.g.*, peer-to-peer, incast), requiring different path selection and queue control strategies to achieve *high performance*.

To address this goal, I primarily employ the methodology of active sensing to enable the DCN to be aware of irregularities and take quick reactions to mitigate their performance degradation. Specifically, HERMES<sup>2</sup> is among the first to sense irregularities in congestion information on different paths and actively switch flows to their desired paths to achieve 10.0% better average flow completion time. ECN<sup>#3</sup> senses queue buildups caused by irregularities in round-trip time (RTT) and actively marks ECN to eliminate the queue, thereby achieving 23.4% lower latency.

**Unpredictability:** Accurately predicting networking traffic patterns has long been a critical task in optimizing the DCN, especially for congestion control, flow scheduling, and more. Therefore, the DCN needs to be *intelligent* in handling the unpredictability of networking traffic.

To achieve this goal, I have developed machine learning-driven congestion control algorithms that precisely predict and control networking traffic. These algorithms have been deployed in real-world environments. MOCC<sup>4</sup> leverages a multi-objective reinforcement learning (RL) algorithm to predict the optimal sending rate for different applications with diverse communication goals with  $14.2\times$  convergence time. Additionally, I have designed and opensourced LiteFlow<sup>5</sup>, which is the first userspace and kernel-space hybrid approach<sup>6</sup> to efficiently deploy ML-based congestion control algorithms with 44.4% better throughput.

**Distrust:** In a shared environment, concerns about trust can arise due to the presence of potentially untrusted or malicious entities. These entities may attempt to eavesdrop on communications, compromise data integrity, or introduce security vulnerabilities. Therefore, a *secure* DCN, or even end-to-end data generation and consumption paradigm is needed.

To meet the requirements, I mainly incorporate a new encryption technology called homomorphic encryption<sup>7</sup> to enable an end-to-end data protection mechanism. To make this technology practical, I designed FLASH<sup>8</sup>, a hardware-based accelerator, to efficiently improve the performance of the homomorphic encryption by 2 – 3 orders of magnitudes.

<sup>1</sup> The other two are computation and storage.

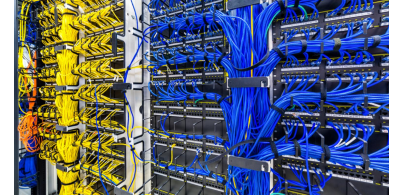


Figure 1: Complex Networking Wiring inside a Datacenter [source].

<sup>2</sup> Hong Zhang, Junxue Zhang, Wei Bai, Kai Chen, and Mosharaf Chowdhury. Resilient Datacenter Load Balancing in the Wild. In *Proc. of SIGCOMM 2017*

<sup>3</sup> Junxue Zhang, Wei Bai, and Kai Chen. Enabling ECN for Datacenter Networks with RTT Variations. In *Proc. of CoNEXT 2019*

<sup>4</sup> Yiqing Ma, Han Tian, Xudong Liao, Junxue Zhang, Weiyan Wang, Kai Chen, and Xin Jin. Multi-objective Congestion Control. In *Proc. of EuroSys 2022*

<sup>5</sup> Junxue Zhang, Chaoliang Zeng, Hong Zhang, Shuihai Hu, and Kai Chen. LiteFlow: Towards High-performance Adaptive Neural Networks for Kernel datapath. In *Proc. of SIGCOMM 2022*

<sup>6</sup> LiteFlow is mainly designed for the Linux OS, but the idea can be extended to other OSes.

<sup>7</sup> Homomorphic encryption allows certain operations, such as addition and multiplication, to be directly performed on ciphertexts without decryption first.

<sup>8</sup> Junxue Zhang, Xiaodian Cheng, Wei Wang, Liu Yang, Jinbin Hu, and Kai Chen. FLASH: Towards a High-performance Hardware Acceleration Architecture for Cross-silo Federated Learning. In *Proc. of NSDI 2023*

### Real-world Impact

My research has garnered recognition from both the academic and industrial sectors. Notably, I have received the Best Paper Award at ICNP 2023<sup>9</sup>, as well as an Honorable Mention for the HKUST CSE Best PhD Dissertation Award, 2021-22<sup>10</sup>. Furthermore, my research findings have been adopted by prominent enterprises such as China Construction Bank, WeBank, SenseTime, and others.

### Networked System for LLM

The advent of large language models (LLMs) has revolutionized the world, driving remarkable advancements across various industries, including programming, personal assistance, and more. Notably, LLM training and inference have emerged as dominant workloads in data centers, creating new demands and opportunities for research in DCN.

Building on my previous work, my current research focuses on addressing the fundamental question of “*How can we co-design DCN and ML systems specifically tailored for LLMs?*” To tackle this question, I have employed the methodology of “*first-principle thinking*” to leverage the unique characteristics of LLM training and inference jobs and develop a networked system for LLMs. The primary objective of this research is to deliver an optimized and efficient system that meets the specific requirements of LLM applications.

### Requirements

LLM training and inference involve the utilization of hardware accelerators such as NVIDIA GPUs and Google TPUs, which work collaboratively to enhance performance. These accelerators are interconnected through high-speed DCNs, such as Remote Direct Memory Access (RDMA)<sup>11</sup>. The efficiency of DCNs significantly impacts the overall performance of end-to-end model training. It is evident that there is an increasing demand for higher DCN bandwidth in order to effectively scale the training and inference of LLMs with a larger number of parameters. Figure 3 quantitatively illustrates the evolution of bandwidth used for LLM applications, indicating that the future requirements for LLMs will soon reach 800Gbps or even higher, such as 1.6Tbps/3.2Tbps.

### Features of LLM Training and Inference

In stark contrast to the aforementioned shared environment, LLM training and inference exhibit a distinct characteristic: *the presence of a single ML job or the use of dedicated, reserved regions*. For instance, in LLM training, the entire GPU cluster (and sometimes even the entire data center) is exclusively allocated to a single LLM training job to maximize performance. In the case of LLM inference, while the entire data center may not be dedicated to a single inference job, a significant portion of machines is reserved<sup>12</sup>. This unique feature has fundamentally transformed the assumptions underlying DCNs, making them highly regular and predictable.

<sup>9</sup> Best Paper: Enabling Load Balancing for Lossless Datacenters.

<sup>10</sup> CSE Best Dissertation Award 2021-22 Recipients Announced.



Figure 2: LLM Revolution [source].

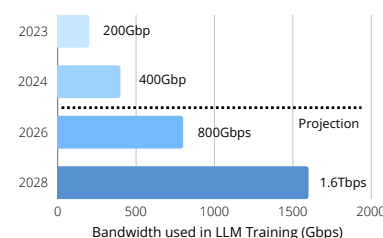


Figure 3: DCN Bandwidth Evolution.

<sup>11</sup> For more information on RDMA, refer to the Wikipedia page: [link].

<sup>12</sup> DeepSeek-AI. DeepSeek-V3 Technical Report, 2024

**Regularity:** As ML jobs solely rely on NCCL<sup>13</sup> for communications, the networking traffic associated with LLM consists primarily of the communication patterns of all-reduce and all-to-all, along with the RDMA WRITE operation involving intermediate numbers<sup>14</sup>. Additionally, due to the exclusive occupation of the entire cluster (or a significant portion) by a single ML job, there is no interference from background applications or traffic, thereby ensuring a high degree of regularity.

**Predictability:** The communication traffic pattern and volume of LLM can be accurately predicted, as it remains unaffected by input data. By analyzing the model and hyper-parameters, the majority of communication patterns and sizes can be predetermined. Although the presence of MoE (Mixture of Experts, see Figure 4) may introduce some unpredictability, this unpredictability is limited to a small region and can be partially determined through probabilistic analysis.

### *AINIC – Streamlined High-performance NIC for LLM Training*

One example of this research outcome is AINIC, a high-performance, specialized network interface card (NIC) specifically designed for LLM training. Due to the demanding requirement for high bandwidth, a software-based solution is not feasible. Therefore, our focus is on a hardware implementation using FPGA initially, with plans for an ASIC implementation in the future. This approach aligns with NVIDIA’s Infiniband or RoCE NICs, which also utilize hardware. However, AINIC sets itself apart by taking a different approach, aiming to streamline the NIC by removing unnecessary components (as depicted in Figure 5). Instead of following a development roadmap that adds more functions, AINIC prioritizes deep optimization of the remaining key components to achieve exceptional levels of performance.

**Connection v.s. Datagram?** As previously mentioned, each server is dedicated to running a single job in the LLM training scenario. Consequently, there is no need for establishing connections to enable fine-grained control between process-to-process pairs. Instead, a simplified NIC design can be achieved by focusing solely on implementing reliable datagrams (RD) with server-to-server control. This approach effectively meets the requirements of LLM training while streamlining the NIC design process.

Furthermore, by exclusively supporting the RD transport, a multiple engines design is employed. Each engine operates as an independent pipeline for sending packets. This approach is feasible because RD eliminates the need for strict ordering of packet delivery, which is unnecessary in LLM training<sup>15</sup>. To further utilize the networking bandwidth, per-packet load balance can be adopted due to the single job feature. Additionally, the rudimentary support for packet reordering serves as an impetus for exploring a packet trimming infrastructure backend.

**Do We Need Congestion Control?** Given that a significant portion of the traffic can be predetermined, AINIC adopts an offline rate calculation method with a hardware rate limiter<sup>16</sup> instead of relying on traditional congestion control algorithms based on AIMD<sup>17</sup>. We will model the traffic pattern and volume adapted by different parallelism strategies, including

<sup>13</sup> NVIDIA Collective Communications Library [link].

<sup>14</sup> An comprehensive discussion on why only use RDMA WRITE can be found here: [link].

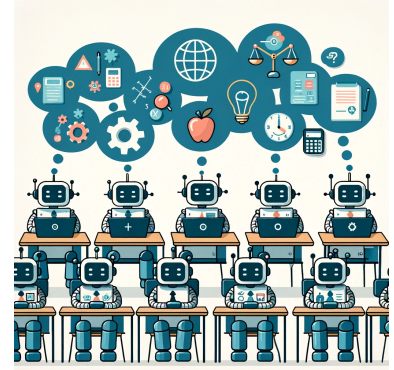


Figure 4: Each Expert Contributes to Different Topics [source].

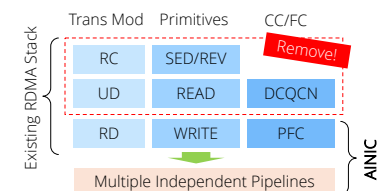


Figure 5: Design Philosophy of AINIC.

<sup>15</sup> Hao Wang, Han Tian, Jingrong Chen, Xinchun Wan, Jiacheng Xia, Gaoxiong Zeng, Wei Bai, Junchen Jiang, Yong Wang, and Kai Chen. Towards Domain-Specific Network Transport for Distributed DNN Training. In *Proc. of NSDI 2024*

<sup>16</sup> Zilong Wang, Xinchun Wan, Luyang Li, Yijun Sun, Peng Xie, Xin Wei, Qingsong Ning, Junxue Zhang, and Kai Chen. Fast, Scalable, and Accurate Rate Limiter for RDMA NICs. In *Proc. of SIGCOMM 2024*

<sup>17</sup> Additive Increase and Multiplicative Decrease: [link].

data parallelism, tensor parallelism, pipeline parallelism, sequence parallelism, and expert parallelism. Based on the analysis results, the offline controller precomputes the communication pattern of the LLM training and leverages a simple rate control interface on the AINIC to schedule flows. Since the dynamic control for MoE is limited to a small region, a centralized congestion controller should suffice for effective management<sup>18</sup>.

A more radical question we are going to explore is, “*Can we completely eliminate congestion control by removing the aforementioned offline and online controllers and let the AINIC send flows at line rate?*” Some of our preliminary results suggest a positive answer to this question, as we have found that congestion is not severe during LLM training. Therefore, even if we completely remove congestion control, there may still be a very low rate of packet loss, which can be effectively recovered using packet trimming. Alternatively, these potential packet losses can be eliminated by enabling PFC<sup>19</sup> with minimal performance degradation.

**Is WRITE all We Need?** Indeed! In LLM Training, the sole communication mechanism employed is NCCL, which exclusively utilizes the RDMA WRITE primitive. Consequently, AINIC is designed to solely implement the RDMA WRITE primitive and the corresponding synchronization operations, such as WRITE with IMM or their equivalent counterparts.

### *Privacy Matters!*

Privacy is crucial for LLM training for several reasons: 1) LLM training often requires large amounts of data, including text from various sources. Privacy ensures that sensitive or confidential information present in the training data, such as personally identifiable information, trade secrets, or proprietary content, is protected from unauthorized access or exposure. 2) Many jurisdictions have privacy regulations in place, such as the General Data Protection Regulation (GDPR)<sup>20</sup> in the European Union. Adhering to these regulations is essential to avoid legal consequences and maintain trust with users and stakeholders.

Therefore, in the design of AINIC, we should consider integrating several potential privacy-preserving methods on the hardware, such as Trust Execution Environment<sup>21</sup> and Fully Homomorphic Encryption, to effectively address the increasingly essential privacy issues.

### **What’s Going On?**

In addition to AINIC, I am committed to exploring other networked systems surrounding LLMs. My ongoing research projects include developing systems for efficient token filtering<sup>22</sup> and optimizing KV cache transfer within a pre-fill/decode disaggregation architecture for LLM inference.

<sup>18</sup> Jonathan Perry, Amy Ousterhout, Hari Balakrishnan, Devavrat Shah, and Hans Fugal. Fastpass: a Centralized “Zero-queue” Datacenter Network. In *Proc. of SIGCOMM 2014*

<sup>19</sup> Priority Flow Control: [\[link\]](#).

<sup>20</sup> <https://gdpr-info.eu>

<sup>21</sup> NVIDIA has introduced confidential computation technology into their GPUs. Interested readers can consult the [\[link\]](#).

<sup>22</sup> Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruo Chen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Not all tokens are what you need for pretraining. In *Proc. of NeurIPS 2024*