Beyond the Bridge: Contention-Based Covert and Side Channel Attacks on Multi-GPU Interconnect

Yicheng Zhang[1], Ravan Nazaraliyev[1], Sankha Baran Dutta[2], Nael Abu-Ghazaleh[1], Andres Marquez[2], Kevin Barker[2]
[1]University of California, Riverside
[2]Pacific Northwest National Laboratory

2024 International Symposium on Secure and Private Execution Environment Design (SEED) | 979-8-3315-0565-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/SEED61283.2024.00014

***Abstract*—High-speed interconnects, such as NVLink, are integral to modern multi-GPU systems, acting as a vital link between CPUs and GPUs. This study highlights the vulnerability of multi-GPU systems to covert and side channel attacks due to congestion on interconnects. An adversary can infer private information about a victim's activities by monitoring NVLink congestion without needing special permissions. Leveraging this insight, we develop a covert channel attack across two GPUs with a bandwidth of 45.5 kbps and a low error rate, and introduce a side channel attack enabling attackers to fingerprint applications through the shared NVLink interconnect.**

I. Introduction

Graphics Processing Units (GPUs) are crucial for accelerating various applications, including computer vision [1], Extended Reality (XR) [2], language models [3], health care [4], and more. As dataset sizes grow, experiments on a single GPU may take days, prompting the use of multi-GPU systems to speed up these applications. Security and privacy concerns within the realm of both CPU and GPU have garnered increasing attention. Conventional I/O interconnects such as PCIe have been demonstrated to be susceptible to side channel attacks [5]–[7]. This study highlights that contemporary high-speed multi-GPU interconnects are also vulnerable to covert and side channel attacks stemming from congestion on these links. By monitoring the data transfer rates of the shared NVLink protocol, adversaries can glean information about other users' activities. Specifically, we devise a covert channel attack across two GPUs, achieving a notable bandwidth. Additionally, we introduce a side-channel attack that allows attackers to profile applications running on multiple GPUs.

In this wild and emerging ideas (WEI) paper, we present the initial outcomes of our attacks on Nvidia multi-GPU systems. This paper introduces the **first** attack on multi-GPU interconnects using congestion timing leakages.

II. Background and Threat Model

NVLink, developed by Nvidia [8], is a high-speed interconnect designed to enable rapid data exchange between CPUs and GPUs. It supports efficient read and write operations on the host and device memory. Each bidirectional link consists of two sublinks, ensuring high-bandwidth communication. Our experiments focus on the second version of NVLink, supported by Tesla V100 GPUs [9].

Threat model. Previous microarchitectural attacks [7], [10], [11] necessitated the co-location of victim and spy users on a single GPU. In contrast, our approach eliminates this requirement. Spy and victim users can share the same NVLink without being located on the same GPU or needing specialized system support. Additionally, the attacker does not require any special privileges or specialized system support.

III. Cross GPU Covert Channel Attack

We set up a scenario involving two Tesla V100 GPUs connected via NVLink. The sender and receiver programs are situated on GPU0 and GPU1, respectively. The sender begins by allocating two 1.25 MB memory buffers: one on the remote GPU1 and another on the local GPU0. To transmit a '1', the sender executes a *cudaMemcpyPeer()* operation, transferring data from GPU1 to GPU0; to indicate a '0', it remains idle. Likewise, the receiver sets up two 256-byte memory buffers on the remote GPU0 and the local GPU1. It then initiates a *cudaMemcpyPeer()* operation to prompt data transfer from GPU0 to GPU1, simultaneously tracking the operation's execution time. The congestion within the GPU interconnects leads to differentiable profiling times: 28,356 clock cycles for '0' and 68,368 for '1'. A threshold of 55,000 clock cycles is employed to distinguish between bit '0' and '1'.

Fig.1 (a) depicts the covert message transmission process. Communication begins upon the receiver's detection of four consecutive "1" bits, at which point the text-to-binary encoded message "Hello,NVLink!" is covertly transmitted. We evaluated the communication bandwidth and error rate by executing over 5 runs of 10000-bit message transmission. The measured bandwidth was recorded as 45.5 kbps, accompanied by an error rate of 3.22%.

IV. Cross GPU Side Channel Attack

In this attack, a background spy application continuously monitors NVLink data transfer latency caused by other users' activities. The NVLink traffic traces are then analyzed to correlate with the victim's actions, allowing us to infer the applications the victim runs.

We demonstrate that an attacker can identify the specific HPC application being executed by eavesdropping on the NVLink timing side channel leakages. To illustrate this, we utilized four applications from the OpenMM [12] benchmark as the victim applications, as listed below. We conducted experiments on two Nvidia Telsa V100s with an NVLink

979-8-3315-0565-3/24/$31.00 ©2024 IEEE
DOI 10.1109/SEED61283.2024.00014

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on November 19,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.

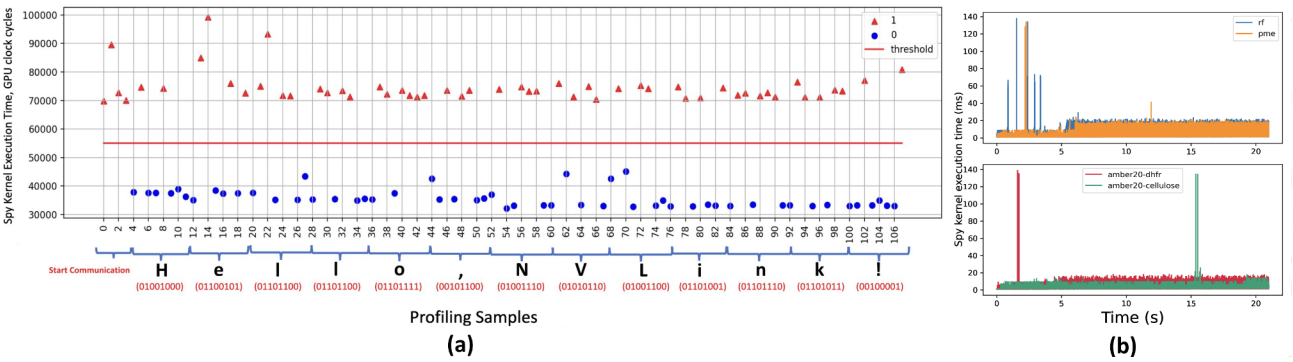

Fig. 1. Two attacks in this work: (a) Cross-GPU covert message; (b) Application fingerprinting via NVLink congestion side channel leakage.

connection. The background spy program performs the following operations: it first allocates two chunks of memory on a remote GPU1 and local GPU0, then executes the vector addition kernel to force sending data from GPU1 to GPU0. Afterward, we use the timer function to record the execution time of the spy kernel and use it to detect the congestion influenced by victim programs.

OpenMM benchmark. OpenMM is a high-performance toolkit tailored for molecular dynamics simulations, supporting multi-GPU systems. We focus on four benchmark applications: rf, pme, amber20-dhfr, and amber20-cellulose.

Observing benchmark distinguishability. Fig. 1 (b) showcases traces for four benchmarks: rf (blue), pme (orange), amber20-dhfr (red), and amber20-cellulose (green), highlighting their distinguishable characteristics. For future endeavors, we intend to employ standard machine learning classifiers, as done in prior studies [13], [14], to further distinguish between these applications.

V. Conclusion

This paper explores congestion-based covert and side channel attacks on multi-GPU systems. We advocate for heightened awareness within our community regarding implementing multi-GPU interconnects with enhanced security measures.

Acknowledgment

We thank our anonymous reviewers for their valuable comments. This work was supported by the U.S. DOE Office of Science, Office of Advanced Scientific Computing Research, under awards 66150: "CENATE - Center for Advanced Architecture Evaluation" and 76125: "AMAIS - Advanced Memory to support Artificial Intelligence for Science." The Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under contract DE-AC05-76RL01830. The work was also partially supported by US National Science Foundation grants CNS-1955650 and CNS-2053383.

References

[1] Y. Zhang, D. Pandey, D. Wu, T. Kundu, R. Li, and T. Shu, "Accuracy-Constrained Efficiency Optimization and GPU Profiling of CNN Inference for Detecting Drainage Crossing Locations," *Workshops of ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC-W)*, 2023.

[2] Y. Zhang, C. Slocum, J. Chen, and N. Abu-Ghazaleh, "It's all in your head(set): Side-channel attacks on AR/VR systems," in *USENIX Security*, 2023.

[3] Y. Li, Z. Li, W. Yang, and C. Liu, "Rt-lm: Uncertainty-aware resource management for real-time inference of language models," *arXiv preprint arXiv:2309.06619*, 2023.

[4] Z. Lai, J. Wu, S. Chen, Y. Zhou, A. Hovakimyan, and N. Hovakimyan, "Language models are free boosters for biomedical imaging tasks," *arXiv preprint arXiv:2403.17343*, 2024.

[5] M. Tan, J. Wan, Z. Zhou, and Z. Li, "Invisible probe: Timing attacks with pcie congestion side-channel," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 322–338.

[6] S. B. Dutta, H. Naghibijouybari, N. Abu-Ghazaleh, A. Marquez, and K. Barker, "Leaky buddies: Cross-component covert channels on integrated cpu-gpu systems," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 972–984.

[7] M. Side, F. Yao, and Z. Zhang, "Lockeddown: Exploiting contention on host-gpu pcie bus for fun and profit," in *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2022, pp. 270–285.

[8] Nvidia, "NVLink and NVSwitch," https://www.nvidia.com/en-us/data-center/nvlink/, 2014.

[9] J. Choquette, O. Giroux, and D. Foley, "Volta: Performance and programmability," *Ieee Micro*, vol. 38, no. 2, pp. 42–52, 2018.

[10] H. Naghibijouybari, A. Neupane, Z. Qian, and N. Abu-Ghazaleh, "Rendered insecure: Gpu side channel attacks are practical," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 2139–2153.

[11] J. Wei, Y. Zhang, Z. Zhou, Z. Li, and M. A. Al Faruque, "Leaky dnn: Stealing deep-learning model secret with gpu context-switching side-channel," in *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2020, pp. 125–137.

[12] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern *et al.*, "Openmm 7: Rapid development of high performance algorithms for molecular dynamics," *PLoS computational biology*, vol. 13, no. 7, p. e1005659, 2017.

[13] S. B. Dutta, H. Naghibijouybari, A. Gupta, N. Abu-Ghazaleh, A. Marquez, and K. Barker, "Spy in the gpu-box: Covert and side channel attacks on multi-gpu systems," in *Proceedings of the 50th Annual International Symposium on Computer Architecture*, 2023, pp. 1–13.

[14] Y. Zhang, R. Yasaei, H. Chen, Z. Li, and M. A. Al Faruque, "Stealing neural network structure through remote fpga side-channel analysis," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4377–4388, 2021.

Authorized licensed use limited to: Univ of Calif Riverside. Downloaded on November 19,2024 at 08:06:56 UTC from IEEE Xplore. Restrictions apply.