Academic supervisor: Asst/P Goh Weihan
Industry supervisor: Arthur Loo
Members: Foo Chang Jie, Muhammad Azfar Bin Adam, Nicholas Koh Wei Xuan, Wong Chong Peng
[BEng (Hons) Information and Communications Technology (Information Security)]

Design and Performance Analysis of a Portable Multi-GPU Unit

Motivation

The main objective of the project is to explore the possibilities of a multi-GPU setup, with the focus on analyzing the impact of these setups on the performance of Artificial Intelligence, Machine Learning as well as Information Security.

With the rise of GPU usage for these tasks, it is critical for us to understand and find out which setup would affect the performance of these tasks. This project aims to analyze the different types of multi-GPU setups alongside a baseline to compare and find the optimal setup, with reasonable thesis and reasoning to the results obtained, in hope of assisting the community in this field.

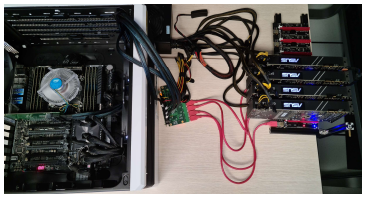
Benchmarks chosen

IS benchmark	Floating-point Computing Power benchmark	AI/ ML benchmark
hashcat	**Highly Parallel LINPACK**	**MLPerf**
• Well recognized • Wide range of hashes • Widely used	• Used to populate TOP500 • Runs reliably • Generates a single, easily comparable result	• Industry recognized • Consistently updated • Runs reliably • Widely used

Our Setups

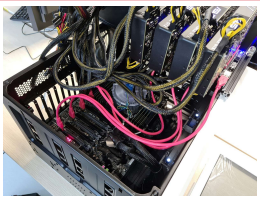
Setup A

All components directly connected to motherboard. Individual cards are connected to the motherboard electrically via PCIe 3.0 x16.



Setup B

GPUs are connected via the PCIe switch. The switch is connected electrically to the motherboard via PCIe 3.0 x4 and the individual cards PCIe 3.0 x1.



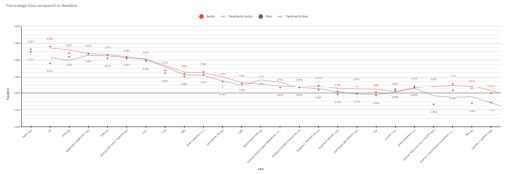
Setup C

GPUs are connected via riser cards. Individual cards are connected to the motherboard electrically via PCIe 3.0 x1.

Item	Details
Motherboard	Asus X99-E WS/USB 3.1
CPU	Intel Xeon(R) E5-1650 v4 @3.60GHz, 6 cores 12 threads
Memory	131072 MB DDR4 2400 MHz (8x Crucial 16GB 2400 MHz)
GPU	4x Asus TURBO-GTX1080TI-11G
PCIe switch	PLX Technology PEX 8614 PCIe
PSU	2 x 1600W

Benchmark Results

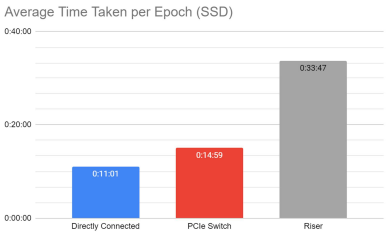
hashcat



Setup B sees an average performance gain of 2.08% across the board, and up to 5.59% improvement. Overall, Setup B performed better in 96% of the hashes.

Setup C sees an average performance gain of 1.61% across the board, and up to 4.97% improvement. Overall Setup C performed better in 76% of the hashes.

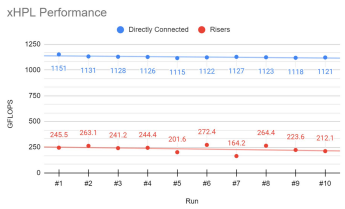
MLPerf (Single Stage Detector)



The average time per epoch increased from 11:01 to 14:59 in Setup B, which is a 36.01% difference with a S.D. of 3 seconds (0.51%).

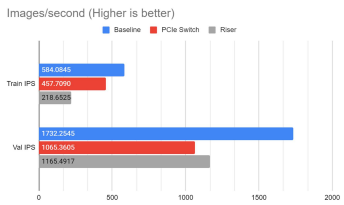
The average time per epoch is 40:18 in Setup C, with a S.D. of 11.97mins (29.68%). As the deviation is large, we compared it using the median of 33:47, which is a 125.52% difference.

NVIDIA CUDA accelerated HPLinpack



Setup B was unable to run HPL with the same configuration due to it being unable to handle high-bandwidth-burst workloads, thus we will not be using it in the comparison.
Setup C decreased in performance by an average of 892.95 GFLOPS, which is a 79.29% decrease in performance, with a S.D. of 33.40 GFLOPS (14.32%).

MLPerf (ResNet50)



Setup B average training IPS is 457.7, a decrease in performance of 22% with an S.D. of 1.16 (0.26%) and validation IPS is 1065.4, a decrease of 38% with an S.D. of 14.38 (1.35%).

Setup C average training IPS is 218.7, a decrease in performance of 62.5% with an S.D. of 3 (1.38%) and validation IPS is 1165.5, a decrease of 32.7% with an S.D. of 22.4 (1.92%).

Analysis & Findings

We found that Setup A outperformed the other setups in all benchmarks other than hashcat. After experimenting with different variables, we concluded that Setup A was bottlenecked by temperature. Temperature plays a huge factor in performance due to NVIDIA's GPU Boost 3.0, where the cards drops boost bin once it exceeds the target temperature.
Even though both Setup B & C are effectively connected via PCIe 3.0 x1 per card, there is a significant difference in performance. It is likely due to the speed difference of the task distribution in the chipset and PCIe switch.

Conclusion

Low bandwidth workloads like hash cracking can perform on externally connected GPUs with negligible performance impact. However, the bandwidth limitation that comes with not directly connecting the GPUs via riser cards and PCIe switches affects the performance of high bandwidth requirement tasks like HPL and AI/ML by a huge amount. This results in mining-rig-esque setups not being ideal for such tasks. The type of switch used will also affect the performance of task distribution.
If the GPUs are connected differently, like with NVIDIA's NVSwitch technology, externally connected GPUs can perform as well as directly connected ones.