

# Research-Based Application of Composable GPU Solution to Multiple AI Training Usage

## CUSTOMER STORY

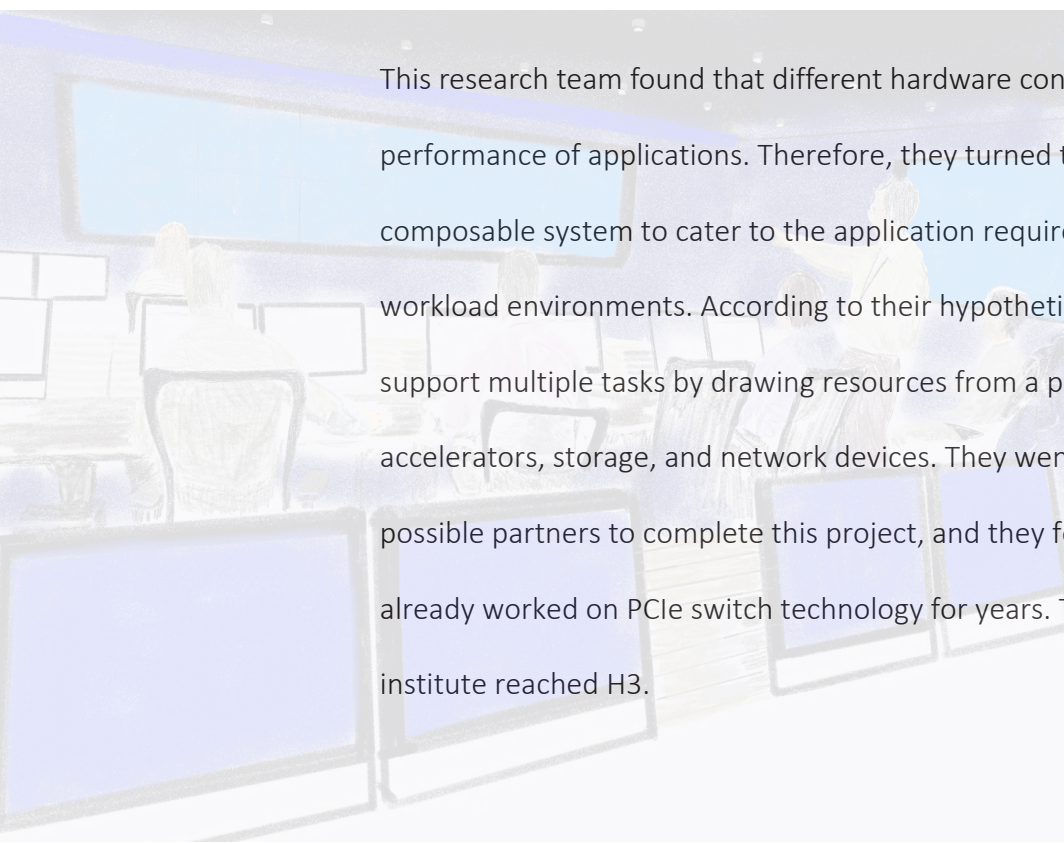
H3 PLATFORM INC.

**H3** 11F.-1, No.79, Sec. 1, Xintai 5th Rd, Xinbei, Taiwan

# BACKGROUND

**A**long with the mega-trends of cloud computing proliferation, AI growth, and Network cloudification, a broader range of emerging workloads exerts pressure on data centers to expand their infrastructure scale.

Based on internal requirements and research motivation, a widely known research institute sought a solution to overcome future possible infrastructure restrictions. They wanted to design a flexible and extensible fundamental architecture that could improve production efficiency without tearing down or building new data centers.



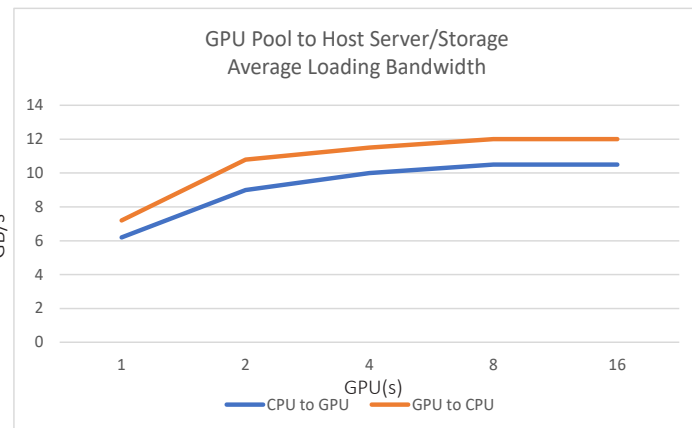
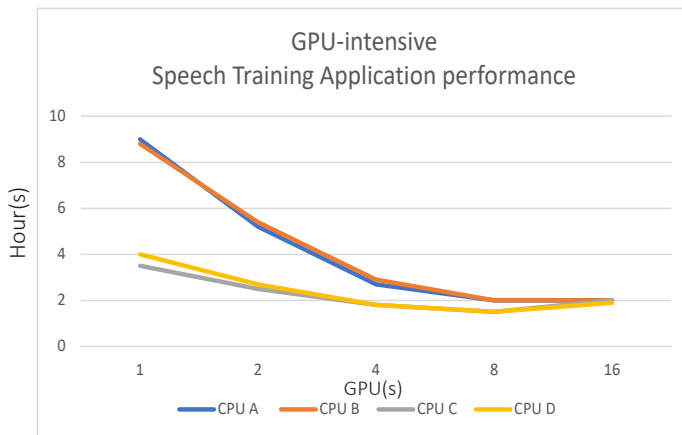
This research team found that different hardware configuration affects the performance of applications. Therefore, they turned to the elasticity of the composable system to cater to the application requirements in different AI workload environments. According to their hypothetical concept, this design has to support multiple tasks by drawing resources from a pool with processors, accelerators, storage, and network devices. They went online to search for possible partners to complete this project, and they found H3. Since 2014, H3 has already worked on PCIe switch technology for years. Therefore, this well-known institute reached H3.



## SOLUTION

Traditionally, a server is used to pair up with fixed accelerators and memory. No matter how much the workload is, the server uses the same package to process. Such an operation leads to over or under-provisioning. H3's PCIe switch technology realizes the avant-garde concept of pooling all the resources in individual chassis at a rack level. Disaggregated devices can be dynamically distributed to upstream compute servers through PCIe interconnects, effectively tackling the resource waste problem.

Standing on the solid experiences of PCIe switch technology development, H3 had a breakthrough in connecting Broadcom PEX9797 switches to maximize the number of lanes (288 PCIe 3.0 lanes) for device and host linkage. Such an infrastructure design outperforms the traditional one by non-blocking sharing independent accelerators, storage, and communications at a rack level. Users can arrange the needed lanes optimally according to application bandwidth requirements. For example, the research team found that in a specific fixed configuration environment, due to the interconnect bandwidth saturation, the GPU performance faces the bottleneck as the bandwidth saturates by its maximum of 16 GB/s (Figure 1&2). In other words, no matter how many GPUs got installed, that hardware system can only offer the performance of about four GPUs. Thanks



Diagrams retrieved from Chung (2018)

to PCIe switch technology, the bandwidth can earn the opportunity to be raised or reduced with dynamic lane arrangement instead of being locked. The number of GPUs turns meaningful. As well, the resources become composable and arrangeable according to the tasks. For example, Figures 3 and 4 show system can allocate different sets of accelerators to every server for their specific workload, widely enhancing agility and flexibility.

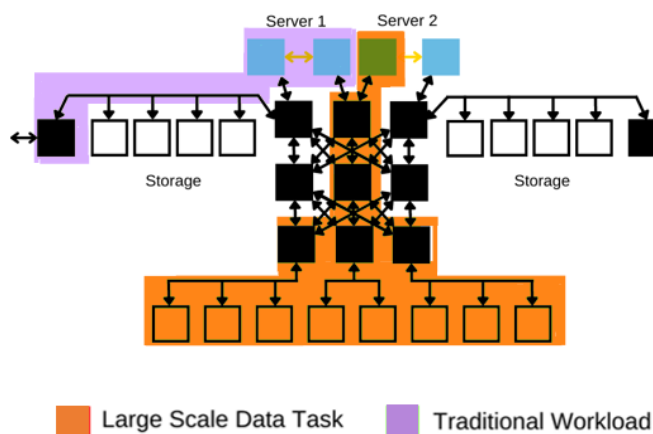


Figure 3

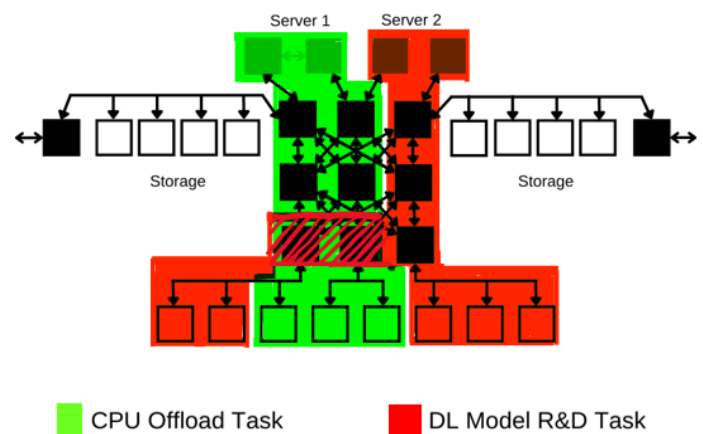


Figure 4

## RESULTS

According to the research report, this research confirms that a composable system can effectively buffer the influence of different component configurations on the overall performance. One excerpt from their research paper that a researcher can conduct AI workloads efficiently with the flexibility of the composable system is a firm acknowledgment of our profession and hard work. Not only having produced the valuable research paper, but this research team also adopted composable solutions in their labs to run different AI workloads since that time. This research team is still our customer, having purchased two generations of PCIe composable systems and management solutions.

