TECH BRIEF

ATTO
DIRECT
2
GPU

ATTO Direct2GPU™ Technology

SUPERCHARGE DATA MOVEMENT BETWEEN BLOCK STORAGE AND **NVIDIA GPU**s

THE BENEFITS

Ultra-low Latency
Delivers efficient data transfers while ensuring the lowest latency

Superior Performance
Application performance can be improved up to 50% (I/O dependent)

Increase overall system performance
Reduce CPU utilization by up to 40% allowing you to run more applications on expensive CPUs

Reduce PCIe bus traffic
PCIe bus traffic can be reduced by up to 50%, opening paths for additional data transfers and increasing efficiency

OVERVIEW

With ATTO Direct2GPU™ technology, data transfers between ATTO storage adapters and NVIDIA GPUs are accelerated, eliminating bottlenecks caused by unnecessary system memory copies, resulting in increased overall performance and efficiency while freeing up bandwidth for additional applications and ensuring that the GPU isn't waiting for data. Direct2GPU guarantees that GPUs are able to operate at peak efficiency, that CPU bandwidth isn't unnecessarily burdened and that the PCIe bus doesn't needlessly become clogged with traffic. Direct2GPU technology is enabled on ATTO Celerity™ Fibre Channel and ATTO ExpressSAS® SAS/SATA Host Bus Adapters (HBAs) and provides performance benefits for a wide range of GPU-accelerated applications from analytic databases, high-performance computing and machine learning to digital content creation including rendering, compositing and image processing.
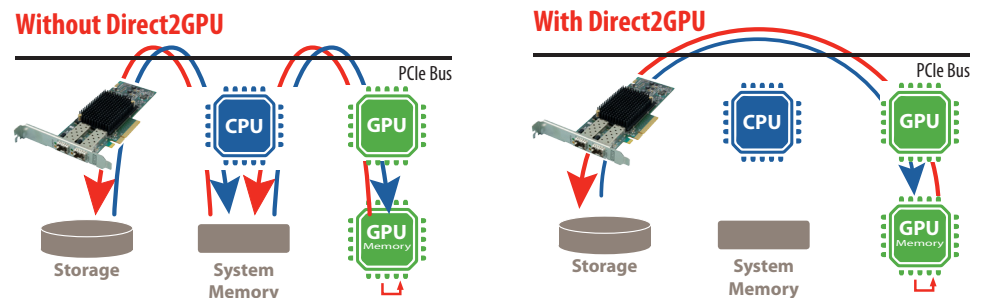
HOW DOES ATTO DIRECT2GPU™ TECHNOLOGY WORK?

A typical "write" operation to load data into the GPU memory for processing requires allocation of both GPU and System memory. Data is moved from storage, across the PCIe bus and into the System memory. From there, it is moved across the PCIe bus again and into the GPU memory for processing. When the GPU processing is complete, the data is written back to the storage, following the same steps, but in reverse order. The operation consumes unnecessary CPU and PCIe bandwidth.

Using Direct2GPU technology, data transfers are simplified, reducing PCIe bus traffic and CPU utilization. ATTO Direct2GPU moves data directly from storage into GPU memory via direct memory access (DMA) with minimal CPU and system memory utilization.

WHAT APPLICATIONS BENEFIT FROM ATTO DIRECT2GPU™ TECHNOLOGY?

Applications where the GPU is regularly starved, waiting for data transfers to be completed from local and other block storage devices, are prime candidates for implementing Direct2GPU technology.

Without Direct2GPU — PCIe Bus — CPU — GPU — Storage — System Memory — GPU Memory

With Direct2GPU — PCIe Bus — CPU — GPU — Storage — System Memory — GPU Memory

TECH BRIEF

ATTO

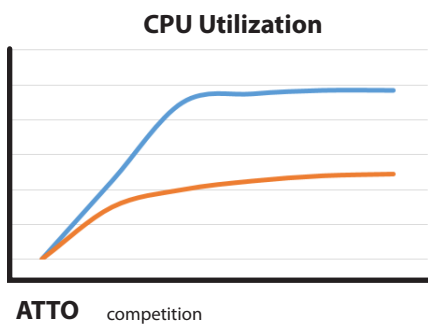## MEDIA AND ENTERTAINMENT APPLICATIONS

- Applications using GPUs for data transformation such as color correction, finishing, post-production and data manipulation
- Frame grabbers for vision systems.
- Video stream transcoding and encoding working in parallel with high-performance GPUs

HIGH PERFORMANCE COMPUTING (HPC)

- Large databases and business analytics
- Scientific computing and modeling such as seismic simulation, computational fluid dynamics and other dynamic simulations

MACHINE AND DEEP LEARNING

- High-frequency trading applications
- Cloud computing and analytics

CPU Utilization

ATTO   competition

ATTO Direct2GPU™ Technology

SUPERCHARGE DATA MOVEMENT BETWEEN BLOCK STORAGE AND NVIDIA GPUS

PARTNERS IN DEVELOPMENT

ATTO partners with application developers and system builders to bring the benefits of Direct2GPU™ technology to NVIDIA user communities. Our work begins with an application environment analysis to evaluate whether your particular work flow could benefit from direct data transfers. We'll then provide a developer's kit that includes everything you will need to easily and quickly integrate Direct2GPU technology into your application. This kit includes all of the code libraries, sample code, hardware drivers and documentation needed to get the job done. Also, we have a trained development team at the ready to answer any questions you might have. Our goal is to make the process hassle-free so that your end users can experience the benefits of ATTO Direct2GPU technology.

Interested in getting started? Contact ATTO Technology at direct2gpu@atto.com or +1 716.691.1999 and ask us how to get your developer's kit.

ATTO CELERITY™ AND EXPRESSSAS® HBAS—BRINGING DIRECT2GPU TECHNOLOGY TO LIFE

ATTO has over 30 years of experience developing and delivering reliable, first-to-market storage connectivity solutions to our customers. Our 32Gb/s and 16Gb/s ATTO Celerity™ Gen 6 Fibre Channel HBAs are consistently the highest-performing adapters for cloud computing, database and virtualized deployments. ATTOExpressSAS® 12Gb/s SAS/SATA HBAs provide dense, efficient connection from servers and workstations to local storage in data center, scientific computing, video editing and post-production, as well as other high-performance application environments.

ATTO Advanced Data Streaming (ADS™) Technology provides controlled acceleration of data to deliver the most consistent performance and reliable data transfers. ATTO Celerity and ATTO ExpressSAS HBA solutions are the ideal solution for users looking to achieve the highest I/O and data throughput for advanced video and enterprise-class IT applications.

	Celerity™ FC-324G	Celerity™ FC-164G	Celerity™ FC-162G	ExpressSAS® G120F	ExpressSAS® G12F0
Ports	4	4	2	16 Internal	16 External
PCIe Bus	X16 PCIe 3.0	X8 PCIe 3.0	X8 PCIe 3.0	X8 PCIe 3.0	X8 PCIe 3.0
Form Factor	Full Height	Low Profile	Low Profile	Low Profile	Low Profile
Max. Transfer Rate	12.8 GB/s	6400 MB/s	3200 MB/s	8000 MB/s	8000 MB/s
SKU	CTFC-324E-G00	CTFC-164P-G00	CTFC-162P-G00	ESAH-120F-G00	ESAH-12F0-G00