International Journal of Electrical and
Electronics Engineering (IJEEE)
ISSN(P): 2278-9944; ISSN(E): 2278-9952
Vol. 8, Issue 1, Dec - Jan 2019, 1-20
© IASET

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

STREAMING MULTIPROCESSOR ARCHITECTURE, TENSOR CORES AND NVIDIANVLINK OPTIMIZED FOR DEEP LEARNING AND HIGH PERFORMANCE COMPUTING (NVIDIA TESLA V100&NVIDIA VOLTA V100)

Omkar Rajesh Kachare

Research Scholar, Department of Electrical and Electronics Engineering,
California State University, Sacramento, CA, USA

ABSTRACT

AI is not defined by any one industry. It exists in the fields of supercomputing, healthcare, financial services, big data analytics and gaming. It is the future of every industry and market because every enterprise needs intelligence. Modern High Performance Computers (HPC) data centers are the key to solving some of the world's most important scientific and engineering challenges. NVidia Tesla V100 GPU accelerated computing platform powers these data centers, with industry leading applications to accelerate HPC and AI workloads.[8] The features of the V100 GPU are,

- *New Streaming Multiprocessor (SM) Architecture optimized for deep learning.*
- *NVidiaNVLink fabric.*
- *150 Teraflop per second.*
- *640 Tensor cores.*
- *Paired NVidia CUDA and Tensor cores to deliver high performance.*

The main objective of this paper is to study NVidiaNVLink fabric and Tensor cores, which are used for transfer of data from CPU and GPU, and Streaming Multiprocessors Architecture optimization for deep learning.

The rapid growth in deep learning workloads has driven the need for a faster and more scalable interconnect, as PCIe bandwidth increasingly becomes the bottleneck at the multi-GPU system level. NVidiaNVLink technology addresses the interconnect issues by providing higher bandwidth, more links and improvises stability for multi- GPU and multi GPU-CPU system configurations. A single NVidia tesla V100 GPU supports up to 6 NVLink connections with a total bandwidth of 300GB/sec, which is 10x the bandwidth of PCIe gen 3.[11]

KEYWORDS: *GPUs the best choice for computing deep neural network based applications and machine learning applications*

Article History

Received: ***04 Nov 2018*** **|** ***Revised:*** ***14 Nov 2018*** **|** ***Accepted:*** ***27 Nov 2018***