NI Big Physics Summit 2016
11-12 February 2016
CERN, Switzerland

A high throughput data acquisition and processing model for applications based on GPUs

J. Nieto[a], G. de Arcas[a] , M.Ruiz[a] ,R. Castro[b], J.Vega[b], P. Guillen[a]

[a]*Instrumentation and Applied Acoustic Research Group, Technical University of Madrid (UPM), Madrid, Spain. Email: jnieto@sec.upm.es*

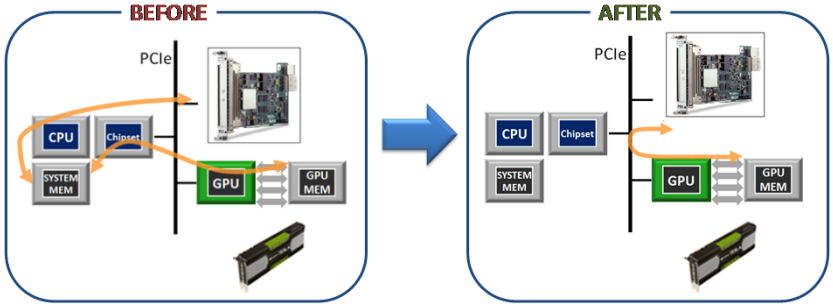
[b]*Data acquisition Group EURATOM/CIEMAT Association for Fusion, Madrid, Spain*

ABSTRACT

There is an increasing interest in the use of GPU technologies for real time analysis in fusion devices. The availability of high bandwidth interfaces has made them a very cost effective alternative not only for high volume data analysis or simulation, and commercial products are available for some interest areas. However from the point of view of their application in real time scenarios, there are still some issues under analysis, such as the possibility to improve the data throughput inside a discrete system consisting of data acquisition devices (DAQ) and GPUs. This work addresses the possibility of using peer to peer data communication between DAQ devices and GPUs sharing the same PCIexpress bus to implement continuous real time acquisition and processing systems where data transfers require minimum CPU intervention. This technology eliminates unnecessary system memory copies and lowers CPU overhead, avoiding bottleneck when the system uses the main system memory.

OBJECTIVES

- To use peer to peer data communication among NI-FlexRIO devices and GPUs sharing the same PCIexpress bus to implement continuous real time DAQ&Processing systems with minimum CPU intervention.
- PCIe Write Through using RDMA for GPUdirect.
- To develop standardized methodologies to integrate these technologies in scientific research environments like EPICS.



RESULTS

	CPU + GPU		GPU		
Bloq Size	Med. (MB/s)	Std. Dev	Med. (MB/s)	Std. Dev	Gain(%)
64KiB	323.11	44.43	710.78	31.67	119.98
128KiB	412.89	42.12	797.09	36.21	93.05
256KiB	488.71	23.99	825.74	13.77	68.96
512KiB	542.87	22.85	858.50	7.61	58.14
1MiB	550.01	11.28	868.70	4.33	57.94

* Obtained executing a bechmark 100 hundred.

CONCLUSSIONS

- We have obtained a direct communication path between the FlexRIO FPGA card and a NVIDIA Tesla k20, improving data transfer rates. In this improvement is important to note the lower standard deviation when data is transferred to the GPU.
- In this new model, a copy of data to the CPU memory is avoided, leaving this only for control tasks. There is an improvement of 57 % when the data block moved is bigger.

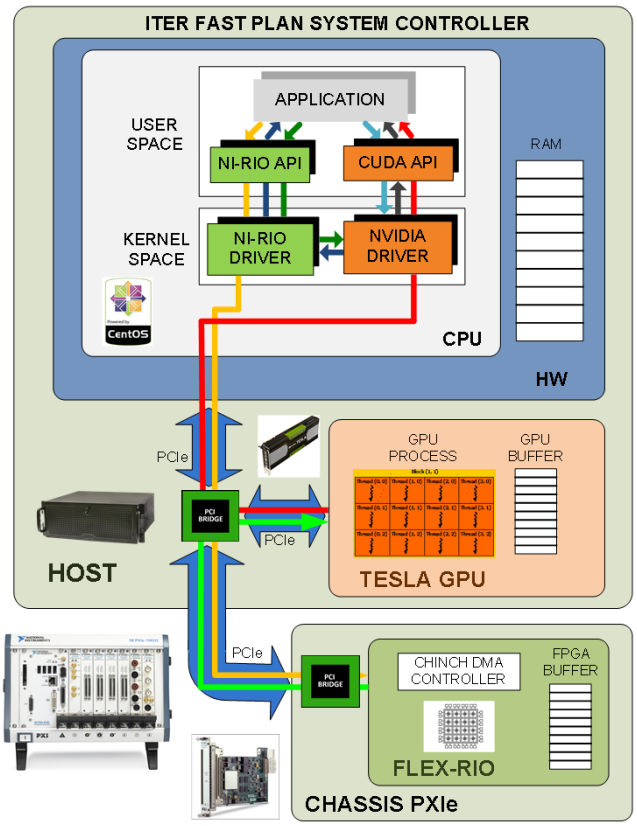
ACKNOWLEDGEMENTS

This work was supported in part by the by the Spanish Ministry of Economy and Competitiveness under project ENE2012-38970-C04-04

TEST SYSTEM

- Hardware: Scenario close to ITER FPSC:
  - Workstation that houses a NVIDIA Tesla k20 and a NI Pcie 8372 cards. Rack NI PXIe-1062Q that houses a FlexRIO 7966R and NI 8370 cards (4 lines PCIe 1.0)
- Software:
  - OS Linux Centos 6.4 64 bits, NI FlexRIO Open Linux Driver, Linux C Api FlexRIO FPGA, NVIDIA Driver, CUDA 5.5 SDK.

METHODOLOGY



Workflow steps:

- From user space DMA FIFO configuration is requested, using functions of NIRIO and CUDA. In kernel space, both drivers reserve GPU memory.
- In user application, data movement between GPU card and FlexRIO FPGA card starts.
- Data is sent to GPU, and processed on the GPU.
- Finally GPU send results to the user application.