High Performance Data-Transfers in Grid Environment using GridFTP over InfiniBand

Hari Subramoni[*], Ping Lai[*], Raj Kettimuthu[**], Dhabaleswar. K. (DK) Panda[*]

[*]Computer Science and Engineering Department
The Ohio State University, USA
[**]Mathematics and Computer Science Division
Argonne National Laboratories, USA

Outline

- Introduction & Motivation
- Designing Globus-XIO ADTS Driver
- Experimental Results
- Conclusions & Future Work

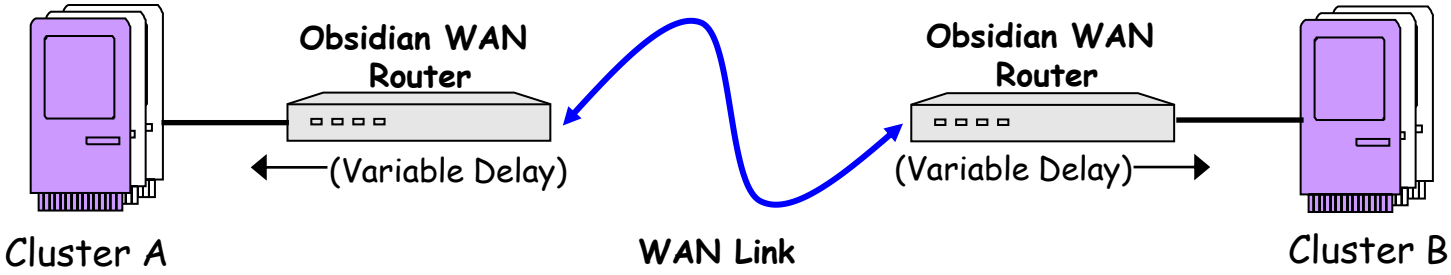
Introduction

- Increasing demands in high end computing has lead to the deployment of compute and storage nodes on a global scale
- Efficient bulk data transfer within and across such clusters is important for
  - Data-set distribution, content replication, remote site backup etc
- GridFTP, designed using the Globus XIO framework, is the most popular mechanism to achieve this in Grids
- But, performance of GridFTP in WAN is limited by relatively low communication bandwidth of existing network protocols

Introduction (cont.)

- System Area Network (SAN) gains momentum
  - InfiniBand, 10Gigabit Ethernet/iWARP etc.
  - High bandwidth, low latency and unique capabilities
- InfiniBand
  - Open Industry Standard based
  - High Performance
    - High Bandwidth (~ 40Gbps)
    - Low Latencies (~1 us)
  - Multiple Transport modes
    - Including RC, UD
  - Two communication semantics
    - Channel semantics: send/recv
    - Memory semantics: RDMA operations
  - WAN capabilities!!
    - Obsidian Longbow routers
    - Bay company products

InfiniBand WAN



- Point-to-point inter-cluster links
- Single Data Rate (SDR)
- Varying delay emulates the WAN distance

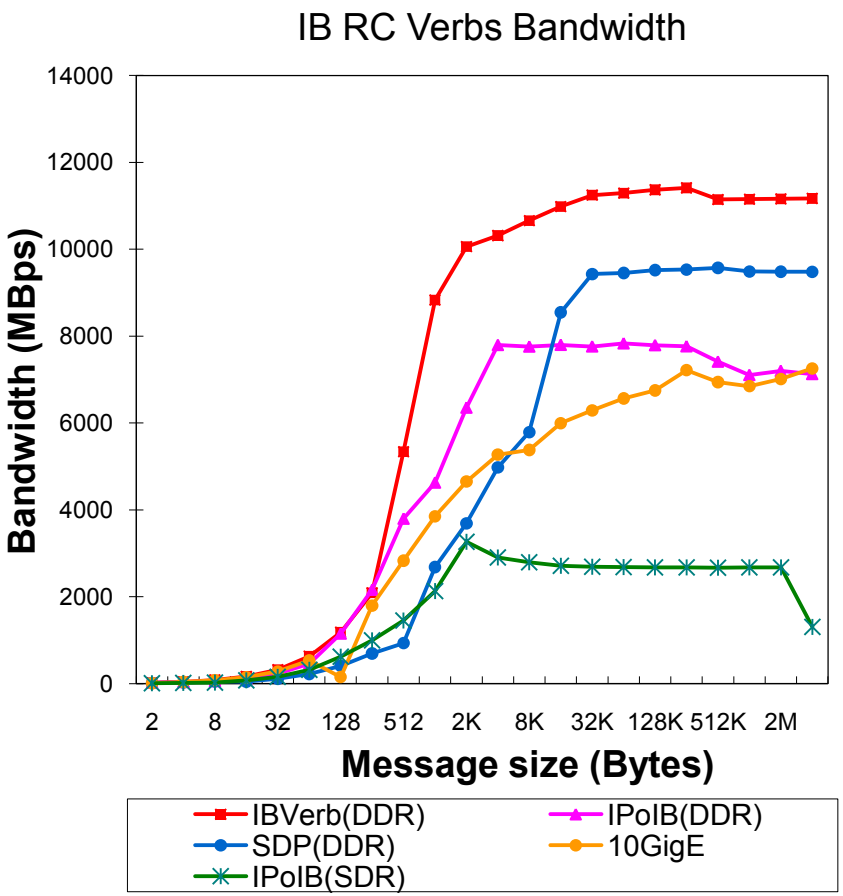
Delay (us)	Distance Emulated(km)
0	0
10	2
100	20
1000	200
10000	2000

Links emulate each *km* of WAN link length with an increase of 5 *us* to each packet latency

Motivation

- Historically, wide-area data transport has been handled mostly by TCP
- Performance achieved using TCP is only a small part of the available bandwidth
- Although multiple efforts have been made to improve the performance of TCP or implement reliable transfer on top of UDP, none of these are able to utilize the complete network bandwidth
- InfiniBand on the other hand is able to saturate the network

- Previous works have shown that IB has better performance over WAN than optimized versions of TCP [1]



[1] N. Rao, S. Poole, P. Newman and, S. Hicks, Wide Area InfiniBand RDMA: Experimental Evaluation, HPiDC'09

Motivation (cont)

- Other interesting developments has been taking place in parallel
- Authors have previously designed a File Transfer Protocol over IB (FTP-ADTS[2]) with features like zero copy and pipelined, data transfers
- But, initial version of FTP-ADTS was not designed for disk based data transfers, a necessity for large volume data transfers used by real world applications such as UltraViz and CCSM

[2] P. Lai, H. Subramoni, S. Narravula, A. Mamidala and D. K. Panda, Designing Efficient FTP Mechanisms for High Performance Data-Transfer over InfiniBand, ICPP '09.

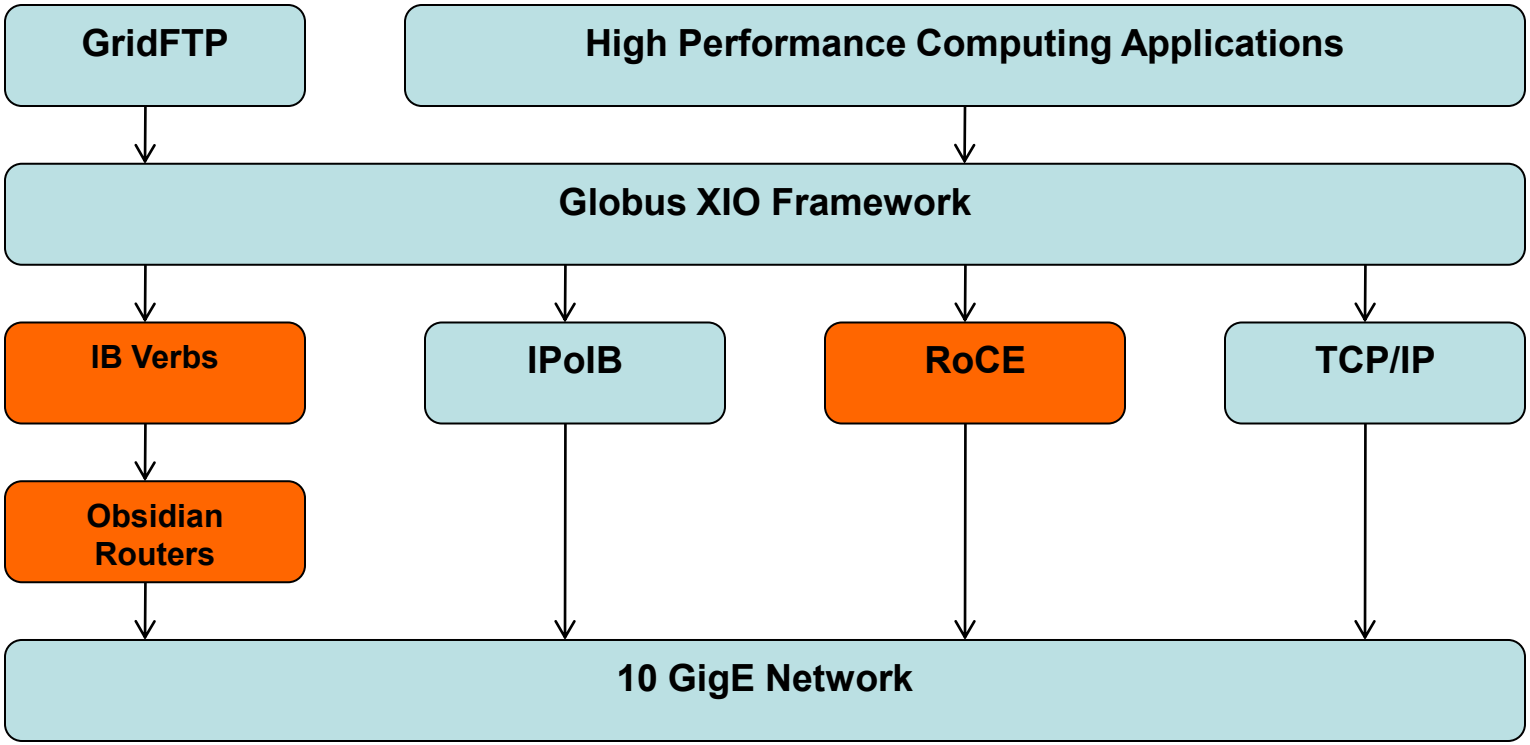
Problem Statement

- Multiple solutions for data transfers in Grid
  - GridFTP
    - well accepted, well defined and easy to use
    - low performance
  - FTP-ADTS
    - high performance
    - basic FTP interface and less support for disk based data transfers
- Hybrid approach, combining ease of use of GridFTP with high performance offered by FTP-ADTS, could be optimum
- *Can we provide native IB support to GridFTP by designing a new Globus-XIO ADTS driver for GridFTP?*
- *Can we enhance the design of FTP-ADTS for disk based data transfers using separate **Disk-IO** threads to stage data from slower speed disks to memory?*

Outline

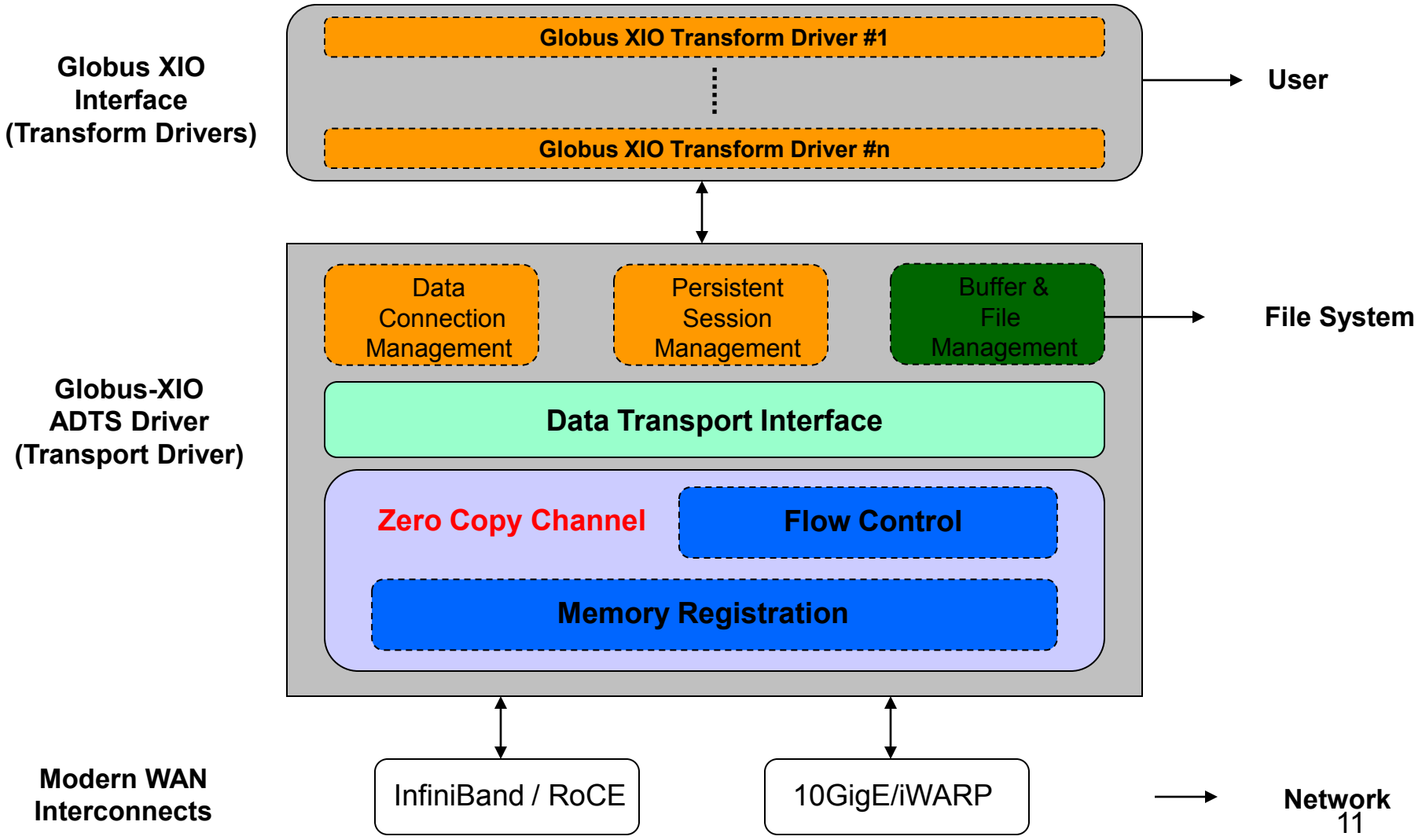
- Introduction & Motivation
- Designing Globus-XIO ADTS Driver
- Experimental Results
- Conclusions & Future Work

Communication Options in Grid



- Multiple options exist to perform data transfer on Grid
- Globus-XIO framework currently does not support IB natively
- We create the Globus-XIO ADTS driver and add native IB support to GridFTP

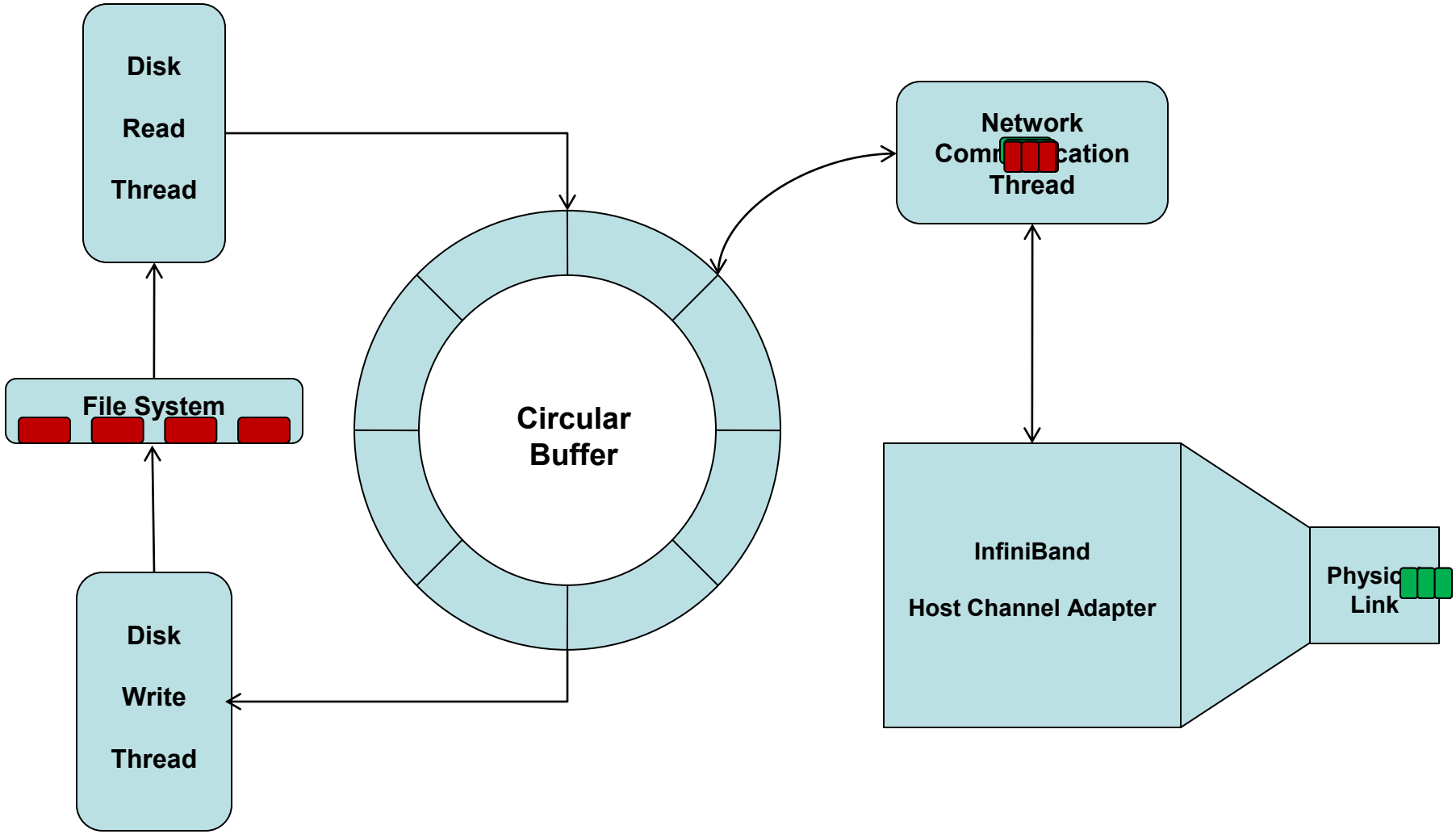
Globus-XIO Framework with ADTS Driver



Buffer and File Management

- Two types of buffers
  - File Buffer
    - Used to stage data from source, by the Network or Disk-IO threads
    - Size varied from 2 MB to 64 MB
  - Network Buffer
    - Used by network thread to inject data into the network
    - Set to 1 MB based on results of previous evaluation
- Disk read thread pre-fetches set of locations from the disk into circular buffer
  - Size of one element of Circular Buffer = Size of File Buffer
- Low and high watermarks defined to limit the minimum and maximum number of file buffers read by disk read thread before yielding to Network Thread
- Write thread follows similar design
- Low and high watermark values as well as the size of the file buffer need to be carefully tuned to achieve best performance

Buffer and File Management (Cont)



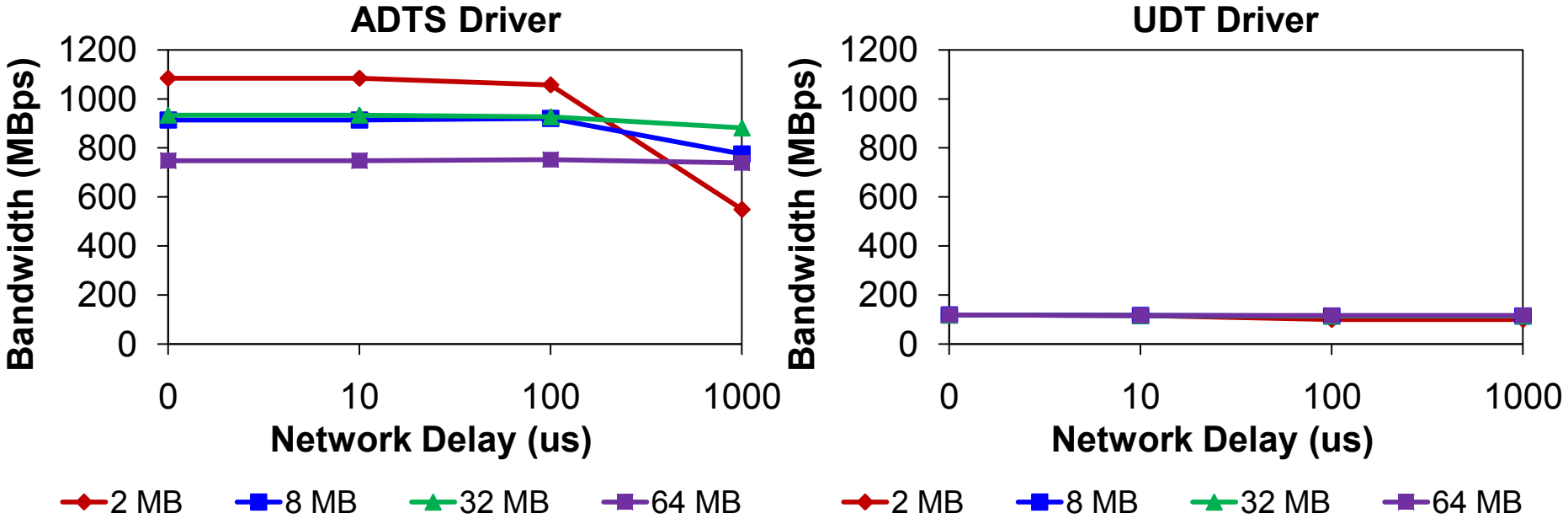
Outline

- Introduction & Motivation
- Designing Globus-XIO ADTS Driver
- Experimental Results
- Conclusions & Future Work

Experimental Setup

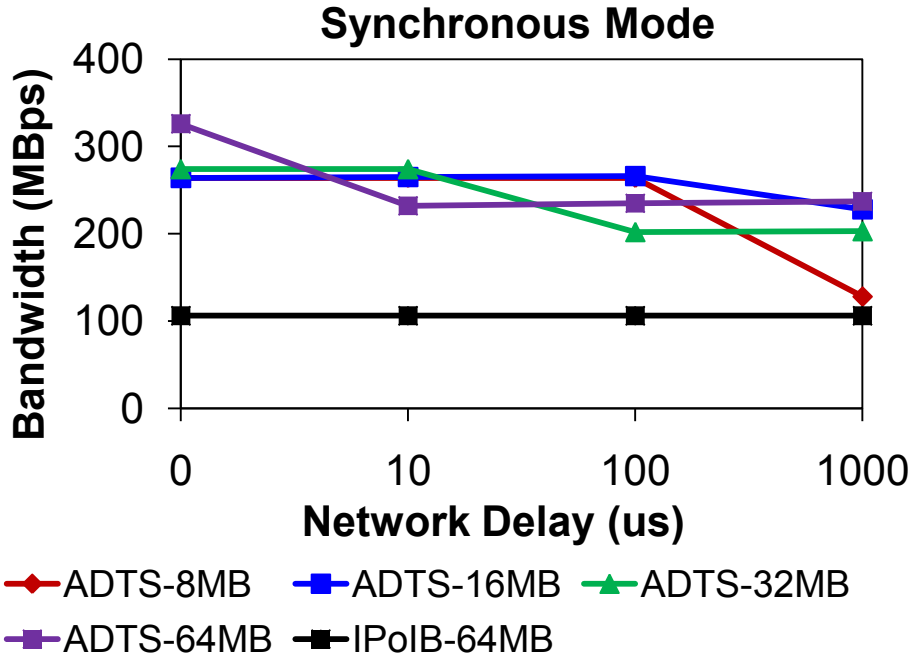
- Experimental Testbed
  - Dual quad-core Xeon Nehalem processors, at 2.40 GHz with 12 GB RAM and PCIe 2.0 interface
  - Red Hat Enterprise Linux Server release 5.3
    - Linux kernel 2.6.18-128
  - InfiniBand (IB) QDR ConnectX HCAs with OFED 1.4.2
  - For all IPoIB (TCP/IP) based tests, auto-tuning of the socket buffers was enabled
    - HTCP congestion control mechanism was used due to bug with the CUBIC congestion control protocol on the 2.6.18 kernels
  - Nodes are connected with Obsidian WAN routers
  - Numbers shown are for *FTP-Get* operation

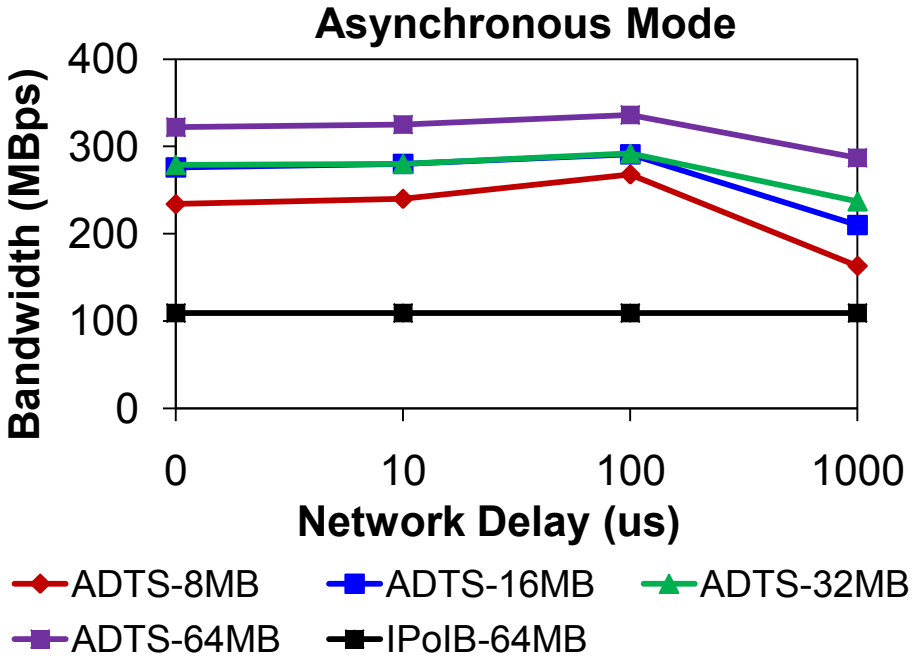
Performance of Memory Based Data Transfer



- Performance numbers obtained while transferring 128 GB of aggregate data in chunks of 256 MB files at application level
- ADTS based implementation is able to saturate the link bandwidth
- Best performance for ADTS obtained when performing data transfer with a file buffer of size 32 MB

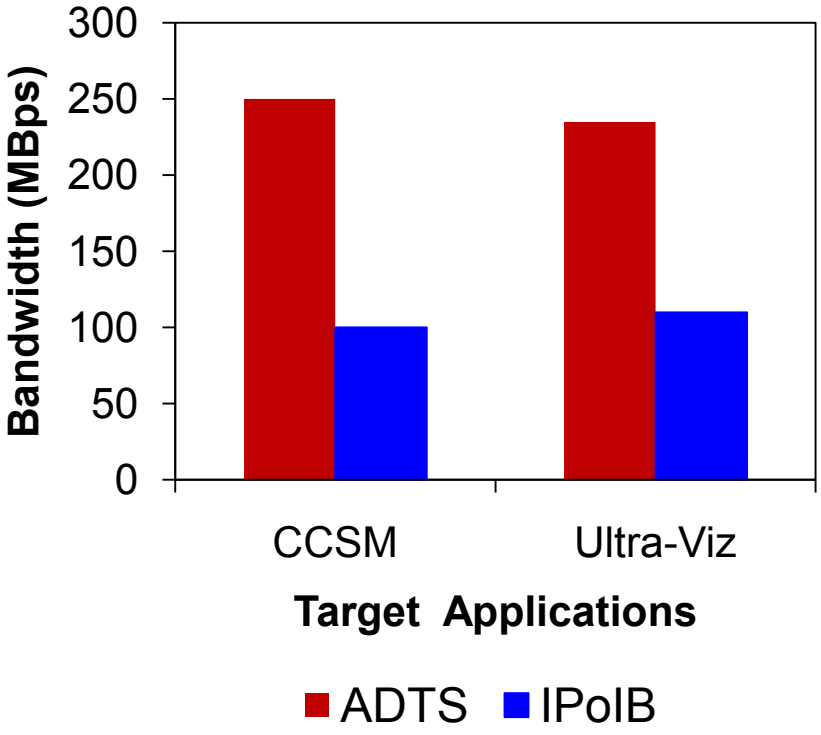
Performance of Disk Based Data Transfer





- Performance numbers obtained while transferring 128 GB of aggregate data in chunks of 256 MB files at application level
- Predictable as well as better performance when Disk-IO threads assist network thread (Asynchronous Mode)
- Best performance for ADTS obtained with a circular buffer with individual buffers (also file buffers) of size 64 MB

Application Level Performance



- Application performance for FTP *get* operation for disk based transfers
- Community Climate System Model (CCSM)
  - Part of Earth System Grid Project
  - Transfers 160 TB of total data in chunks of 256 MB
  - Network latency - 30 ms
- Ultra-Scale Visualization (Ultra-Viz)
  - Transfers files of size 2.6 GB
  - Network latency - 80 ms

The ADTS driver out performs the UDT driver using IPoIB by more than 100%

Outline

- Introduction & Motivation
- Designing Globus-XIO ADTS Driver
- Experimental Results
- Conclusions & Future Work

Conclusions & Future Work

- Design and implement an ADTS based Globus-XIO driver for Grid-FTP, capable of utilizing optimizations done in ADTS including memory registration cache, and pipelined data transfer
- Propose and design a simple data staging mechanism for the ADTS library allowing us to achieve better and predictable file transfer performance
- GridFTP running over the new Globus-XIO-ADTS driver achieves significantly better performance (up to 100% improvement) for both disk and memory based transfers
- Future work
  - Conduct experiments on actual wide area networks and study the impact of cross traffic on performance of the ADTS driver
  - Explore the impact of various transport level features such as re-transmission timeout on the performance of the ADTS driver
  - Investigate whether, InfiniBand as a protocol, is suited for WAN data transfers

Thank you

{subramon, laipi, panda} @cse.ohio-state.edu
{kettimut} @mcs.anl.gov

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

Mathematics and Computer Science Division
http://mcs.anl.gov/

Questions???