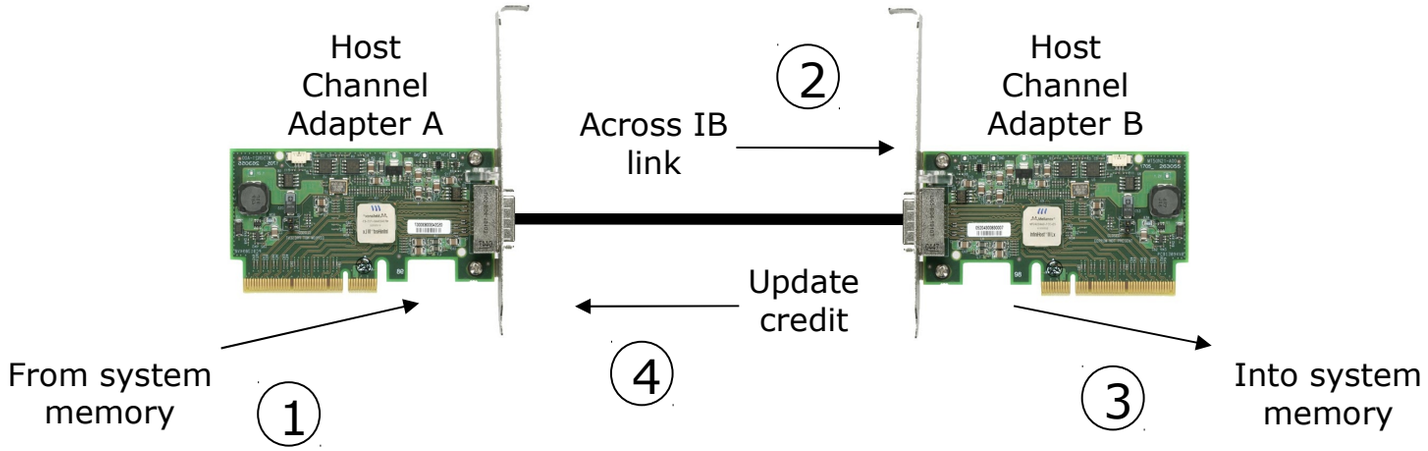
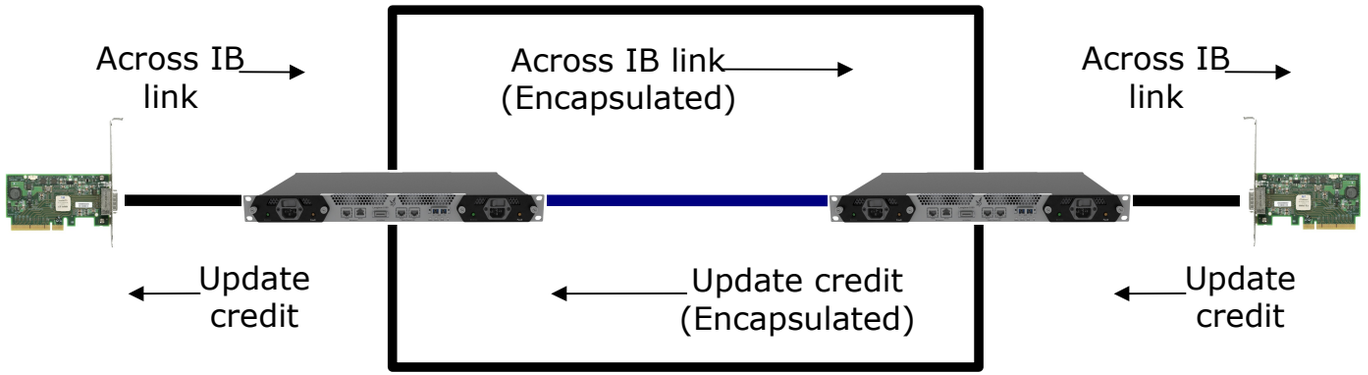
dsync – Fast, Secure & Lossless Wide Area File Migration using InfiniBand

Supercomputing 2010
New Orleans, LA

David Southwell, Ph.D
CTO – Obsidian Strategics Inc.
780.964.3283
dsouthwell@obsidianstrategics.com

Why InfiniBand range-extenders?





Longbow-maintained flow control state machine

Longbow Product Family

X100

WAN – OC-192 / 10GbE / dark fiber
Switch / Router modes
Global reach

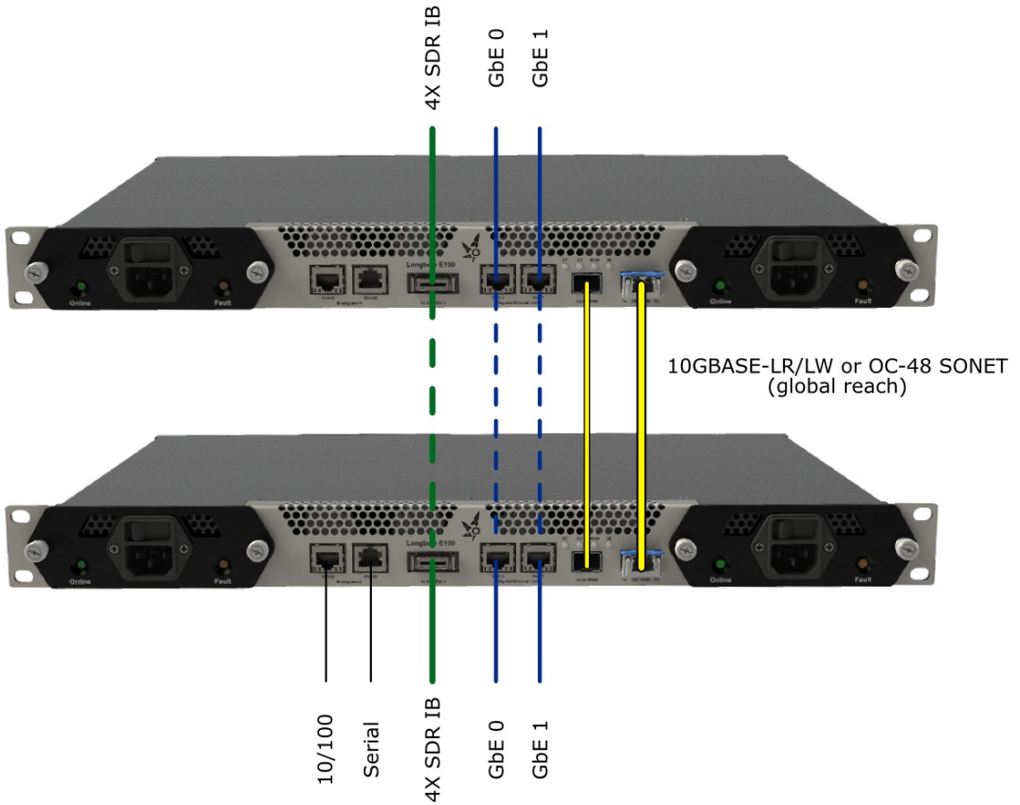
E100

WAN – 10GbE / dark fiber or OC-48 SONET
Switch / Router modes
AES-192-GCM cryptography
Global reach

C100 series

Campus / Metro – dark fiber
Switch mode only
300m - 80km reach

Longbow Operation



- IB and WAN open standards compliant
~1 microsecond port-to-port latency
- Full 4X SDR bandwidth
- Software transparent
- High Availability architectures

Various models support:

• Additional encapsulation of dual Gigabit Ethernet links

• InfiniBand routing modes, which allows each site to keep its own unique subnet (scaling, fault isolation)

• Encryption and authentication to protect data and equipment from unauthorized external access

All models use out-of-band management – https, SNMP, CLI

dsync tool

Several approaches, including the Lustre file system, which supports InfiniBand natively and has received performance refinements for high-latency connections.

A new option is the recently developed directory synchronization tool, dsync:

DSYNC(1) User Commands DSYNC(1)

NAME

dsync − Fast remote directory copy tool

SYNOPSIS

dsync [−**hqaprdltnogDI46**] [−**option**=*config_string*] [−**config**=*config_file*] [−**e**=*path*] [−−**dsync−path**=*path*] [−−**delete**] [−−**numeric−ids**] [−−**include**=*filt*] [−−**exclude**=*filt*] [−−**devices**] [−−**specials**] [−−**force**] [−−**inplace**] [−−**size−only**] [−−**prefer−include**=*filt*] [−−**prefer−exclude**=*filt*] [−−**order** {tree | breadth | depth}] [−−**address**=*IP*] [−−**no−splice**] [−−**direct−io**] {*source*} {*destination*}

DESCRIPTION

dsync is a file copying tool optimized for very high performance networks and storage. It strives to copy files without causing the available CPU resources to become a bottleneck. In particular, the main motivation for this program is to get beyond TCP speeds using 10 and 40gbit IB RDMA networks.

dsync tool

dsync assumes the network is faster than processor/ storage, which opens up a new set of optimizations.

Pipeline directory scanning, data transfer, file I/O and integrity checking.

Heavily multi-threaded, largely lockless internal architecture.

File system and storage hardware independent.

Zero-copy user-space RDMA based transfers.

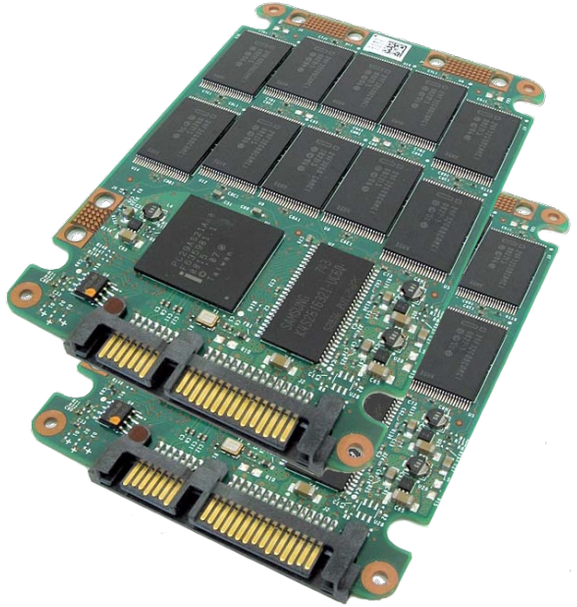
Supports multiple parallel InfiniBand paths – if they are available - for higher aggregate performance and fault tolerance.

Runs on young Linux kernels, and likes Nehalem!

dsync/Longbow E100 performance demo

Hardware set up comprises a pair of servers with two SSDs in a RAID-0 configuration, communicating via Mellanox HCAs connected via Longbow E100s.

Kernel	2.6.35-22-generic #35-Ubuntu
CPU	2xSMP Intel(R) Pentium(R) D CPU 3.00Ghz, 2MB Cache
RAM	2GB (DDR2-533)
IO	Intel Corporation 82801G (ICH7 Family) IDE Controller (non-AHCI SATA)
	Mellanox Technologies MT25204 [InfiniHost III Lx HCA] (SDR IB)
Disc	2xCorsair CSSD-F12 1.1 (120 Gbytes each)
	Linux MD RAID 0, 64k Chunk, XFS



= 483 Mbytes/s

[read/write sustained as a Linux MD RAID 0]

dsync/Longbow E100 performance demo

Two workloads are used for the wide area tests:

1) 73,579 small files for 1.821GBytes in total (standard Ubuntu distribution - /)

[Measures the system's ability to manipulate large numbers of small files, stressing CPU loading and disk sub-system's IOPS performance.]

2) Ten 1.07 Gbyte files

[Measures streaming IO throughput, emphasizing efficiency of encryption mechanisms and high-latency wide-area transport]

rsync is compared with **dsync**, over **TCP/IPoIB** and **RDMA** transports across a 10GbE WAN with emulated latency injected by the Longbow E100s.

Tests are made over **0**, **20ms** and **100ms** WAN RTT latencies

[20ms =~ 2,000km, 100ms =~ 10,000km site separations]

Performance results [1] – 73k files test

Test	Local	2,000km	10,000km	Notes
rsync/ IPSEC/ TCP/IP	**582 s**	-	-	[2KB MTU, IPoIB]
	(3.1 MB/s)			
dsync/ IPSEC/ TCP/IP	**109 s**	**87 s**	**119 s**	[2KB MTU, IPoIB]
	(16.71 MB/s)	**(20.93 MB/s)**	**(15.93 MB/s)**	
dsync/ RDMA	**60 s**	**57 s**	**60 s**	[AIO + O_DIRECT, 1MByte buffer, 128KB blocks]
	(30.35 MB/s)	**(31.95 MB/s)**	**(30.35 MB/s)**	

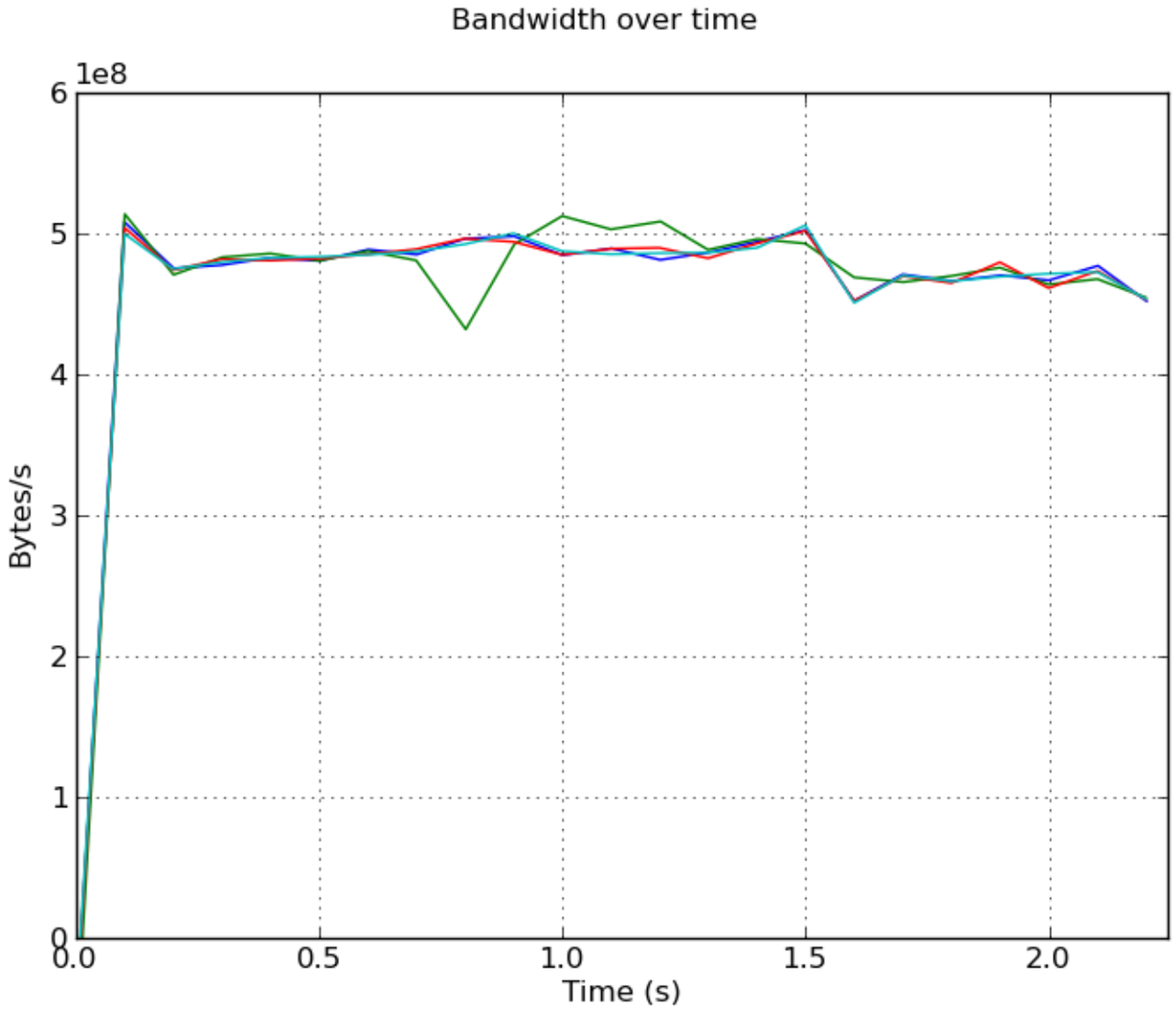
This test is significantly limited by the disk sub-system's meta-data IOP rate, but the spread between dsync+TCP/IP to dsync+RDMA, and dsync to rsync are still apparent.

Performance results [2] – Ten 1.07GByte files test

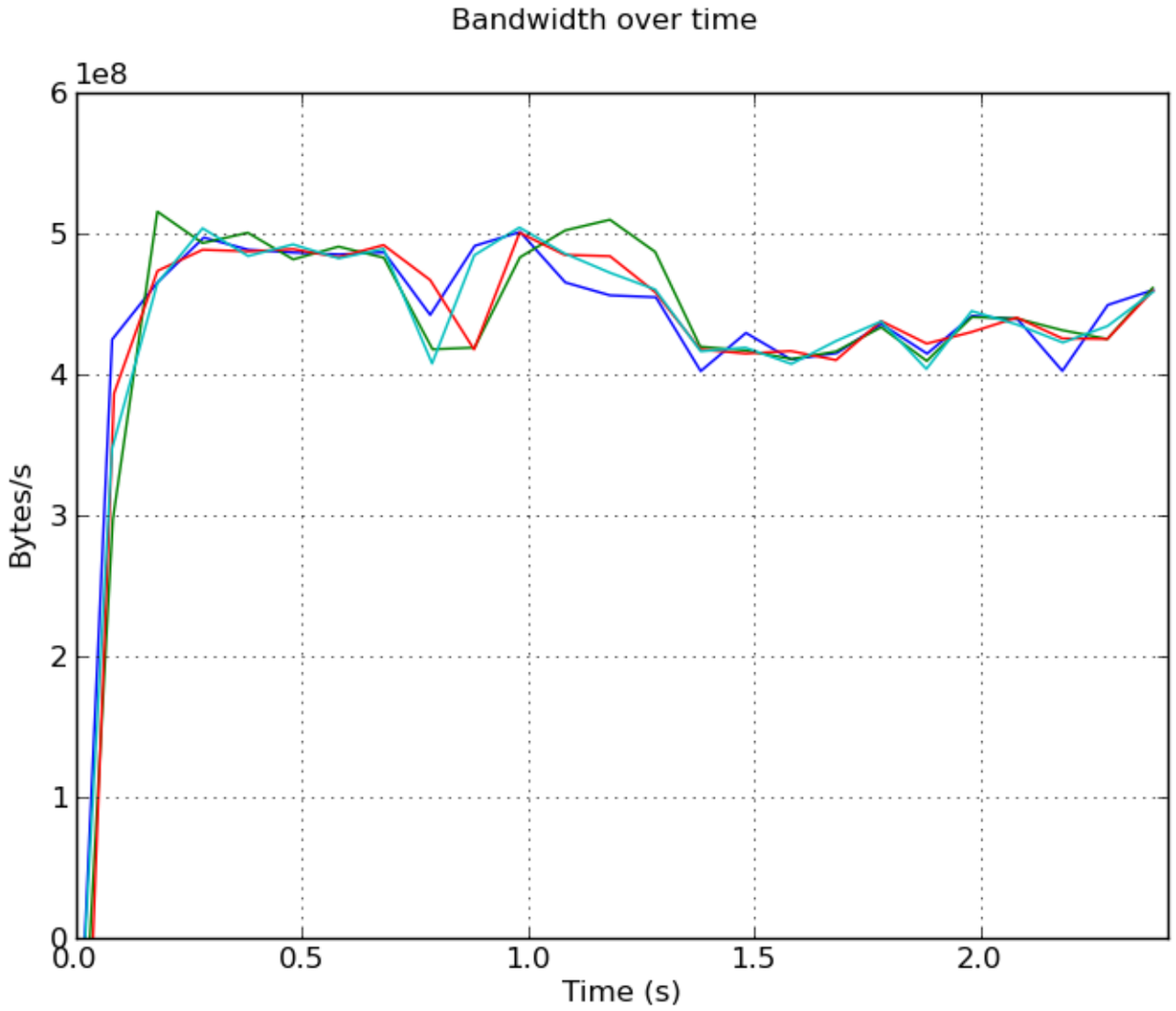
Test	Local	2,000km	10,000km	Notes
rsync/ IPSEC/ TCP/IP	**262 s**	-	-	[2KB MTU, IPoIB]
	(41 MB/s)			
dsync/ IPSEC/ TCP/IP	**48 s**	**232 s**	**488 s**	[2KB MTU, IPoIB]
	(222.8 MB/s)	**(46.12 MB/s)**	**(21.93 MB/s)**	
dsync/ RDMA	**24 s**	**26 s**	**24 s**	[AIO + O_DIRECT, 128MByte buffer, 512KB blocks]
	(445.8 MB/s)	**(411.5 MB/s)**	**(445.8 MB/s)**	

This throughput test highlights rsync vs. dsync at speed, and RDMA over TCP/IP efficiency over high latency connections.

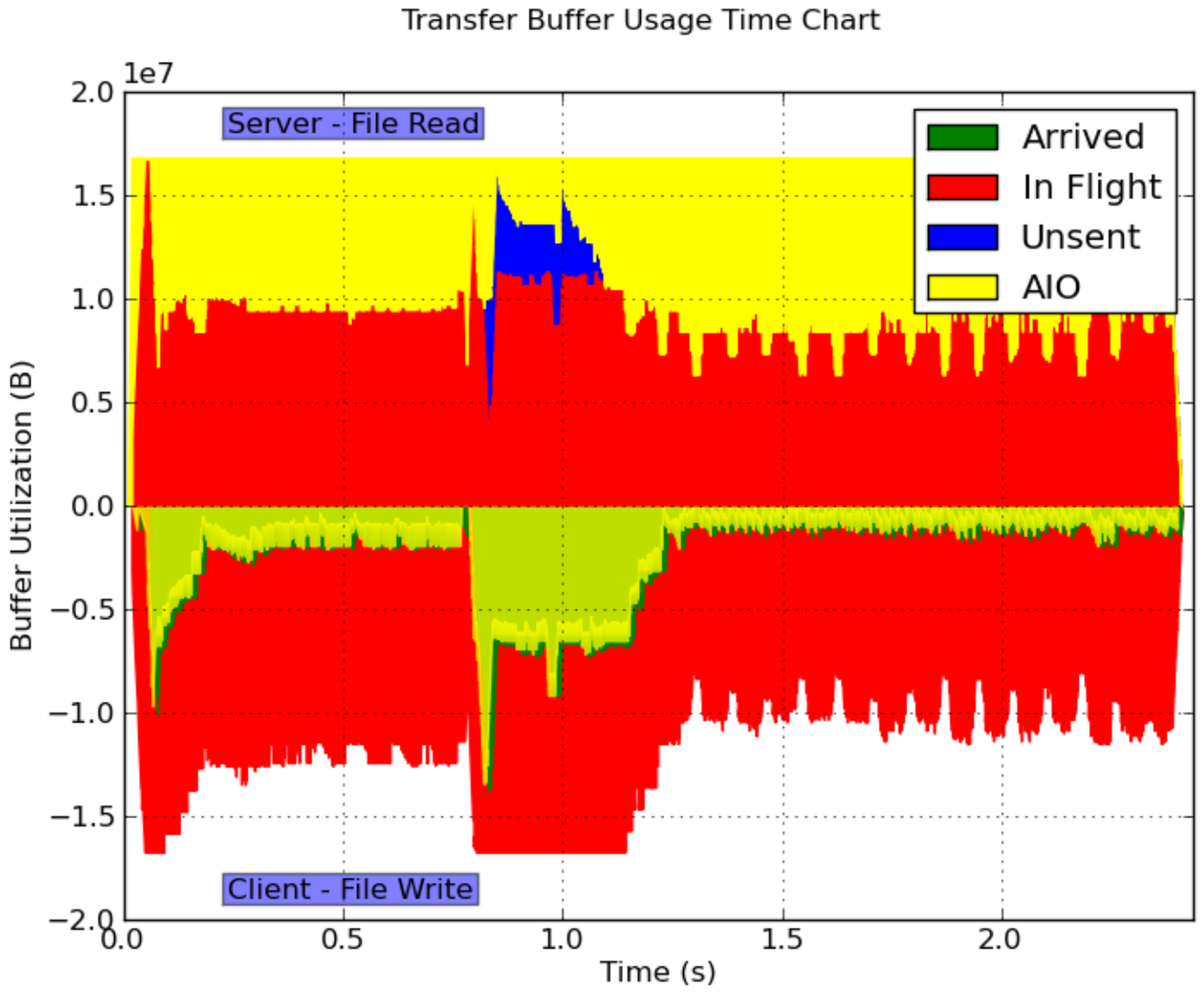
Throughput (0 ms RTT, dsync-Longbow E100)



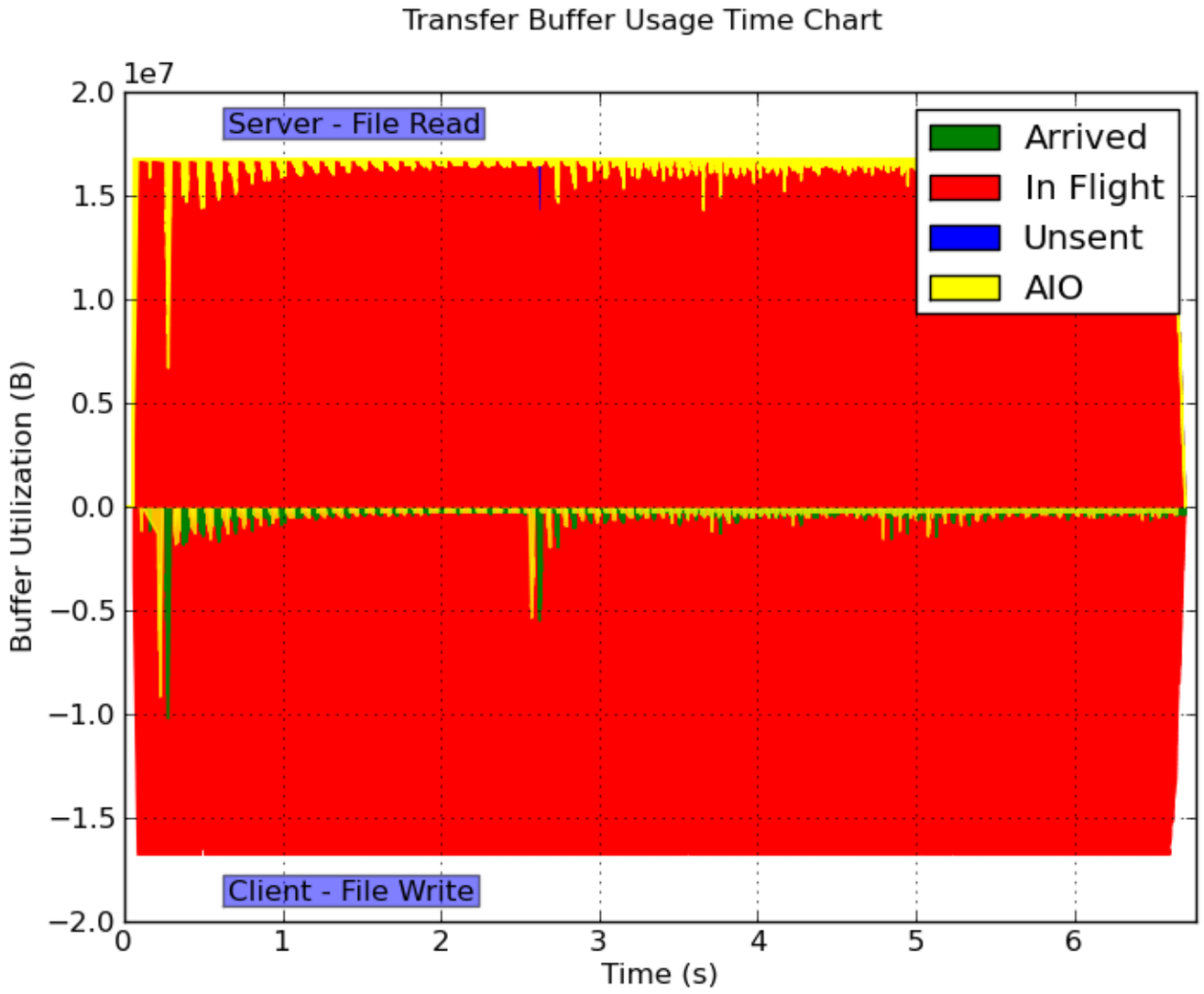
Throughput (2,000km, dsync-Longbow E100)



Buffer trace (2,000km, dsync-Longbow E100)



Buffer trace (10,000km, dsync-Longbow E100)



Wrap up

Over high bit-rate, high latency WAN links:

- TCP/IP is unstable and inefficient
- Lossless InfiniBand is stable and efficient
- File replication protocols need to be redesigned
- Integrated cryptography is much more efficient than software-based types

- High performance storage sub-systems are required
- Small file performance cannot be optimized beyond storage IOPS limit
- AIO and O_DIRECT can significantly enhance IO performance

Thank you!