

# Voltaire® Fabric Collective Accelerator™ (FCA™)

## Slashing MPI job runtime



### The MPI Scalability Challenge

Collective functions in the MPI API involve communication between all processes in a process group (which can mean the entire process pool or a program-defined subset). These types of calls are often useful at the beginning or end of a large distributed calculation, where each processor operates on a part of the data and then combines it into a result. Examples of popular collective functions are MPI\_Barrier, MPI\_Broadcast, MPI\_Reduce and MPI\_Allreduce.

The performance of collective communication operations is known to have a significant impact on the scalability of some applications. Indeed, the global, synchronous nature of some collective operations directly implies that they will become the bottleneck when scaling to thousands of ranks (where a rank is an MPI process, typically running on a single core). This fact has led many researchers to try to improve the efficiency of collective operations. Several challenges face today's collective function communication for applications scaling beyond a few dozens of cores. The impact of these challenges significantly increases as multi-core environments continue to dominate HPC, and the number of cores per node continues to grow.

- In large fabrics, servers are typically connected through several tiers of switches, yet all traffic passes through all tiers because the switches are not capable of handling offloading or aggregation. More tiers and longer routes add latency and jitter that rapidly accumulate and significantly impact runtime.
- In many cases a single rank receives data from many ranks at once, significantly increasing the risk of congestion.
- Use of unicast instead of multicast significantly increases latency and congestion when distributing results from a single rank to thousands of ranks.
- The actual physical connectivity topology of the fabric and process placement is not taken into account in processing the collective function.
- Server OS "noise" quickly accumulates on large, synchronous operations, severely impacting overall latency as well as predictability (jitter).

In the era of proprietary interconnects, some vendors implemented their own optimization mechanisms for collective functions, understanding the great impact these function calls have on overall scalability of applications. To date, InfiniBand vendors have not successfully implemented these capabilities, since significantly improving MPI collectives requires intelligence in all parts of the fabric: host (MPI stack), switch and management system.

### The Voltaire Solution

With the aim of scaling out fabrics and improving performance from an application perspective, not just a server/network perspective, Voltaire has designed a unique patent-pending technology called Fabric Collective Accelerator™ (FCA™) software. FCA significantly reduces the runtime of collective operations on any fabric, and is available as an add-on to Voltaire's Unified Fabric Manager™ (UFM™) software.

### Features

- Offload collective function communication & computation from MPI process into Voltaire switches
- Efficient collective communication flow optimized to job and topology
- Monitor performance of collective operations

### Benefits

- Significantly reduce MPI job runtime (up to 40%)
- Improve collective function scalability above and beyond any proprietary interconnect
- Eliminate congestion caused by collective function calls
- No additional hardware to install or manage
- No space/power/cooling penalty
- Seamless integration with MPI job scheduler
- Zero provisioning penalty (parallel to job scheduler initialization)

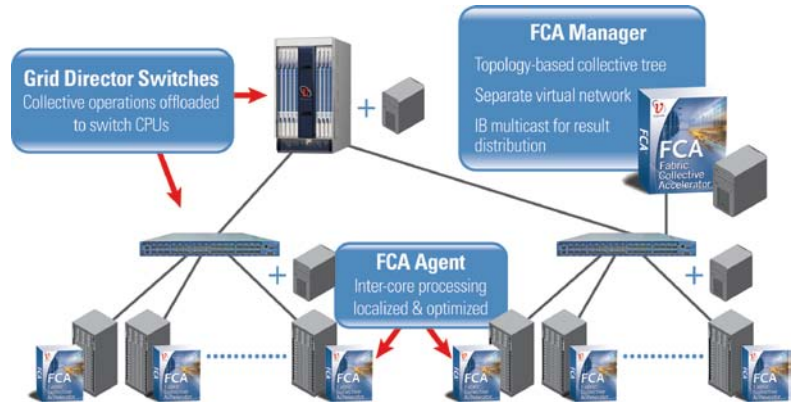
### Supported MPI Collectives

- Allreduce, Reduce, Barrier, BCast, Allgather/AllgatherV
- Unlimited message size
- Open MPI 1.4.1 and up
- Platform MPI 8.0 & up

# Voltaire® Fabric Collective Accelerator™ (FCA™)

The FCA algorithm uses data from the job scheduler and UFM to establish a topology map as it relates to a specific job, as well as the processing power of Voltaire's switches to offload significant parts of the computation.

Using the FCA algorithm with Voltaire switches and UFM ensures a single message per physical wire for any collective function, as opposed to potentially hundreds or thousands of messages per wire using traditional algorithms for collective function handling. This non-blocking collective architecture finally allows InfiniBand to scale collective communication to thousands of nodes better than any interconnect (standard or proprietary) in the market.



Voltaire FCA provides breakthrough performance with no additional hardware

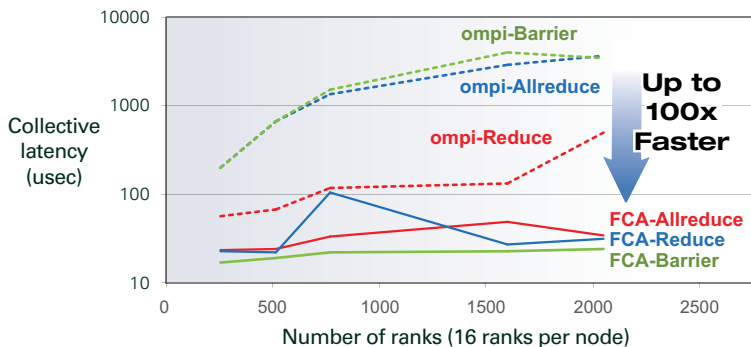
As a result, this well integrated, application-oriented, fabric-wide solution can reduce the runtime of collective operations by more than 90%, resulting in an up to 40% reduction in total MPI job runtime.

Reaching such significant improvements requires new intelligence in several parts of the fabric.

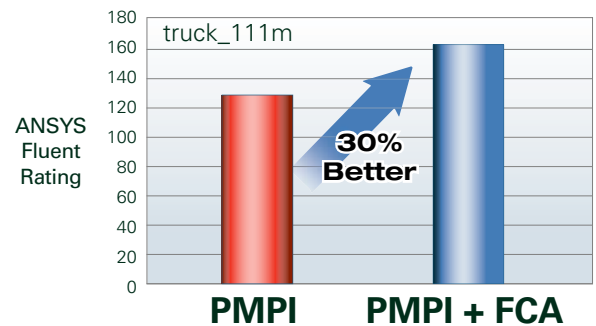
- First, full information regarding topology and job parameters needs to be obtained, processed and translated into a collective map that can then be programmed to all compute entities in the fabric. This is handled by Voltaire UFM.
- Second, each compute entity, be it a server or a CPU on a switch, needs to be able to perform an efficient collective computation on the ranks below it in the tree, and forward the result to the next level.
- Finally, the computation result is distributed to all ranks using InfiniBand (hardware-based) multicast coupled with a reliable message delivery protocol.

Initialization takes no more than 2-3 seconds even on clusters with tens of thousands of ranks, and occurs in parallel with the job scheduler provisioning/resource allocation phase—so there is no penalty on the job start/runtime.

## Benchmark Comparison



IMB (Pallas) benchmarks show 100X faster performance and up to 99.5% runtime reduction using FCA with OpenMPI



ANSYS Fluent benchmarks on 192 cores show up to 30% gain in runtime acceleration using FCA with Platform MPI



Contact Voltaire to Learn More

1.800.865.8247  
info@voltaire.com  
www.voltaire.com

©2010 Voltaire Inc. All rights reserved. Voltaire and the Voltaire logo are registered trademarks of Voltaire Inc. Grid Director is a trademark of Voltaire Inc. Other company, product, or service names are the property of their respective owners. Information in this document is subject to change without notice. Voltaire assumes no responsibility for any errors that appear in this document. All statements regarding Voltaire's future direction and intent are subject to change or withdrawal without notice.