# Performance Migration to Open Solutions: OFED* for Embedded Fabrics

Kenneth Cain
Mercury Computer Systems, Inc.
kcain@mc.com

## Introduction

This presentation details a new, infrastructure-based approach that more effectively enables open standard middlewares and protocols for switched fabric based computing in HPEC systems. The Open Fabrics Enterprise Distribution (OFED) software by the OpenFabrics Alliance (OFA) [1] is considered in the context of a Mercury Computer Systems implementation for Power architecture and serial RapidIO [2] based multi-computers. We will assess challenges with the traditional HPEC middleware model, industry transitions underway that motivate a new approach moving forward, demonstrate performance and "price of portability" of OFED and Open MPI [3], and consider future expansion to more middlewares and architectures.

## Problem Statement

The HPEC community has a long history of pursuing the adoption, specification and implementation of open standard middlewares [4]. The goals and benefits are well understood, including reduced investment for application development and migration/portability among multiple platforms and deployment environments (e.g., airborne, ground). Examples of open standards explored by the community include MPI [5], MPI/RT [6], DRI [7], CORBA [8], DDS [8] and VSIPL/VSIPL++ [9]. The efforts have each been focused in one or more *scaling domains* such as *scale-in* for single-node SIMD, multicore and heterogeneous/accelerator based computing; and (the topic here) *scale-out* for communications among multiple nodes tightly connected by a high speed, switched fabric.

The approach taken to this point has been to posit the suitability of a specific open standard and to promote its development and adoption in the HPEC community. Occasionally this process includes complementary roles defined for other middlewares/APIs. In most cases it is assumed that middleware provides the (final) "insulation layer" between the application and the platform. However, the realities of developing complex systems suggest a need for additional consideration at the underlying platform infrastructure layer – that is vendor-specific and not an open standard.

Some problems that have arisen from the traditional approach include:

- A limited ecosystem of open standard software – resulting from complexity and costs of implementations for various platforms, in addition to the price of portability (performance of open standards compared to native platform software).
- For COTS system vendors, ISVs and open source software, repeated effort to map middleware to specific platforms – on a per-middleware basis and again with shifts in hardware/software or overall system technology.

Some current industry trends and transitions will influence a new approach moving forward:

- Increasing adoption and maturity of Remote Direct Memory Access (RDMA) techniques in a diversity of application areas such as High Performance Computing (HPC), real-time market data, Electronic Design Automation (EDA), and (emerging) cloud/data center networking. This has resulted in closer alignment with HPEC that has relied on RDMA for the same benefits of performance, predictability and efficiency.
- Transitioning toward higher speed switched fabric technology (10/20/40/100 Gbps) [10].
- Transitioning beyond open middlewares and APIs to also include protocols – facilitating communication between the real-time computing resource and enterprise networks and grids, and increasing the ability to converge multiple network accessible services onto the embedded resource.
- An economic need to leverage mass-market driven hardware and software, and a transition from hand-tuned applications to "offload" and optimized implementations of standard interfaces.

## Approach

Rather than propose a single middleware solution, instead we seek to identify and leverage a *pivot point* – a significant concentration of industry investment, and an infrastructure that satisfies requirements of multiple middlewares and protocols "above" (in a typical software layering diagram) and maps to multiple hardware technologies "below".

In the case of open standards based fabric computing for HPEC, we claim that OFED can act as a very effective pivot point. This is already the case in multiple market segments, particularly for HPC based on blade / cluster architectures whose requirements overlap with HPEC in key areas such as the use of RDMA, and in which parallel applications are often programmed using MPI. OFED serves multiple middlewares (e.g., MPI, sockets, messaging APIs), storage protocols, filesystems and databases – and also provides high performance, low overhead fabric transfer support for InfiniBand [11] and RDMA over 10 Gigabit Ethernet (via the Internet Wide Area RDMA Protocol – iWARP [12] and the newly released RDMA over Converged Ethernet – RoCE specification [11]). Among the middlewares served by OFED, MPI is a relevant example because it must deliver high fabric transfer performance due to its central role in the *data plane*

* OFED: OpenFabrics Enterprise Distribution, a software distribution by the OpenFabrics Alliance (OFA) [1]

of many applications. For this reason multiple commercial and open source MPI library suppliers have invested in mapping to the high performance OFED RDMA based primitives and have studied and implemented various aspects of performance and scalability in that environment.

Mercury has developed an OFED *device provider* software module for the Freescale MPC8641D System on Chip DMA engine and the serial RapidIO (sRIO) switched fabric. This software plugs into the overall OFED environment and provides a mapping to sRIO of the fabric transfer services (queue pairs, RDMA read/write and send/receive operations) used by upper layer software packages such as MPI.

Mercury has leveraged the existing HPC industry investment in this middleware/OFED mapping to demonstrate open source MPI libraries Open MPI and MVAPICH2 [13] running in the embedded multicomputer environment with very good performance. Performance enablers (contiguous memory allocators) have also been incorporated into the Mercury Open MPI port to allow MPI applications to achieve data rates approaching the limits of what low level software can achieve over sRIO. These enhancements do not require applications to call vendor-specific APIs, only MPI standard interfaces. Finally, a mechanism of interoperation between the open source library Open MPI and Mercury's vendor-specific middleware, ICS/DX has been demonstrated. We will also present a broader migration strategy toward open solutions that offers an opportunity for a smooth transition – with consideration for different migration starting points reflecting differences in application/deployment environments (ground/laboratory/airborne) and different software dependencies.

The Mercury port of Open MPI has also been optimized to provide better intra-node (scale-in) message passing performance by using an AltiVec accelerated vector copy routine. Also for scale-in, the Mercury OFED sRIO provider has a "loopback" transfer capability, using either the processor or the DMA engine to copy data among processes on the same node.

## Results and Data

The following performance and analysis will be presented:

- Low-level OFED RDMA primitives performing nearly as well as Mercury optimized software offerings (ICS/DX and IOVEC) in the Mercury embedded / sRIO environment, and performing comparably to OFED RDMA in InfiniBand and iWARP fabrics.
- MPI over OFED achieving good initial performance in the Mercury embedded / sRIO environment with a similar price of portability as measured in InfiniBand and iWARP fabrics.
- Open MPI performance over OFED/sRIO outperforming Open MPI over TCP/IP on both 1 Gigabit Ethernet and sRIO.
- Analysis of fundamental differences in message passing (embodied by MPI) and RDMA (embodied by vendor-specific and OFED infrastructure) – accounting for MPI's price of portability.
- MPI intra-node message passing performance, including optimizations added to Open MPI and OFED/sRIO.

Figure 1 provides a summary of performance data already measured.

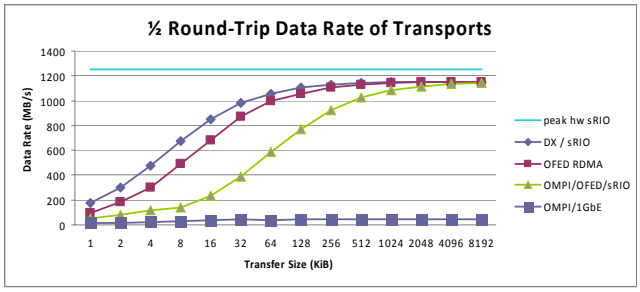


**Figure 1: MPI, OFED, and ICS/DX Bandwidth Results**

## Futures

Plans for the Mercury OFED implementation include expansion to cover additional middlewares of significance to the HPEC community – based on current initiatives in progress in the customer / DoD prime contractor community. Additionally, the OFED infrastructure will be enabled in upcoming Mercury embedded systems reflecting new in processor architecture / switched fabric / DMA engine implementations.

## References


[1] OpenFabrics Alliance, http://www.openfabrics.org

[2] RapidIO Trade Association, http://www.rapidio.org

[3] Open MPI: Open Source High Performance Computing, http://www.open-mpi.org

[4] A. Skjellum, "The State and Future of Middleware for HPEC", H*PEC Workshop Proceedings*, 2009.

[5] Message Passing Interface (MPI) Forum Home Page, http://www.mpi-forum.org

[6] MPI/RT: Real-Time MPI Standard and Forum, http://www.mpirt.org/

[7] Data Reorganization, http://www.data-re.org

[8] OMG Specifications, http://www.omg.org/technology/documents/spec_catalog.htm

[9] Vector Signal Image Processing Library, http://www.vsipl.org/

[10] R. Blau, "Using Layer 2 Ethernet for High-Throughput, Real-Time Applications", *HPEC Workshop Proceedings*, 2008

[11] InfiniBand Trade Association, http://www.infinibandta.org

[12] RDMA Consortium, http://www.rdmaconsortium.org

[13] MVAPICH: MPI over InfiniBand, 10GigE/iWARP and RoCE, http://mvapich.cse.ohio-state.edu/


* OFED: OpenFabrics Enterprise Distribution, a software distribution by the OpenFabrics Alliance (OFA) [1]