HCOLL / IBM Power System S822LC (8335-GTB) SUPPORT

README

Rev 1.0

NOTE:
THIS HARDWARE, SOFTWARE OR TEST SUITE PRODUCT (“PRODUCT(S)”) AND ITS RELATED DOCUMENTATION ARE PROVIDED BY MELLANOX TECHNOLOGIES “AS-IS” WITH ALL FAULTS OF ANY KIND AND SOLELY FOR THE PURPOSE OF AIDING THE CUSTOMER IN TESTING APPLICATIONS THAT USE THE PRODUCTS IN DESIGNATED SOLUTIONS. THE CUSTOMER'S MANUFACTURING TEST ENVIRONMENT HAS NOT MET THE STANDARDS SET BY MELLANOX TECHNOLOGIES TO FULLY QUALIFY THE PRODUCT(S) AND/OR THE SYSTEM USING IT. THEREFORE, MELLANOX TECHNOLOGIES CANNOT AND DOES NOT GUARANTEE OR WARRANT THAT THE PRODUCTS WILL OPERATE WITH THE HIGHEST QUALITY. ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT ARE DISCLAIMED. IN NO EVENT SHALL MELLANOX BE LIABLE TO CUSTOMER OR ANY THIRD PARTIES FOR ANY DIRECT, INDIRECT, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES OF ANY KIND (INCLUDING, BUT NOT LIMITED TO, PAYMENT FOR PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY FROM THE USE OF THE PRODUCT(S) AND RELATED DOCUMENTATION EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Mellanox Technologies
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Overview

This README applies specifically to the HCOLL 3.6 support for the IBM Power System S822LC (8335-GTB) server released in December 2016. The scope is to define restrictions which may be applicable to end-user applications looking to exploit/utilize the fabric collective accelerator.

HCOLL Restrictions

The following HCOLL restrictions exist for the HCOLL 3.6 / IBM 8335-GTB release:

- HCOLL, as of the 3.6 release, does not fully support mixed MPI datatypes.

  In this context, mixed datatypes refers to collective operations where the datatype layout of input and output buffers may be different on different ranks, i.e.:

  For an arbitrary MPI collective operation,

  ```
  MPI_Collective_op( input, count1, datatype-in_i, output, count2,
  datatype-out_i, communicator)
  ```

  Where i = 0,...,(number_of_mpi_processes - 1)

  Mixed mode means when i is not equal to j, (datatype-in_i, datatype-out_i) is not necessarily equal to (datatype-in_j, datatype-out_j).

  Mixed MPI datatypes, in general, can prevent protocol consensus inside HCOLL, resulting in hangs. However, because HCOLL contains a datatype engine with packing and unpacking flows built into the collective algorithms, mixed MPI datatypes will work under the following scenarios:

  - If the packed length of the data (a value all ranks must agree upon regardless of datatype) can fit inside a single HCOLL buffer (the default is (64Kbytes – header_space)), then mixed datatypes will work.
  - If the packed length of count*datatype is bigger than an internal HCOLL buffer, then HCOLL will need to fragment the message. If the datatypes and counts are defined on each rank so that all ranks agree on the number of fragments needed to complete the operation, then mixed datatypes will work. Our datatype engine cannot split across primitive types and padding, this may result in non-agreement on the number of fragments required to process the operation. When this happens, HCOLL will hang with some ranks expecting incoming fragments and other believing the operation is complete.

- The MPI_THREAD_MULTIPLE threading mode is not currently supported for Spectrum MPI applications exploiting HCOLL.