**OFI WG Bi-Weekly telecom – 08/23/2016**

**Agenda:**

* Roll call, agenda bashing
* Intel proposal for Collective Offloads
* Github pull request 2279

**GitHub Issue 2279**

* Fabric and Domain naming conventions; USNic needs to be able to distinguish which is which, since the sense of those has flipped in the USNic design. Proposal is to increment two version numbers – tarball version number changes to 1.4, API revision number as reflected in fabric.h
* Nobody can remember if there is a convention that the tarball version number matches the API revision number, or if they are distinct.
* The proposed change doesn’t change the API, just an indicator to a provider (e.g. USNic, sockets providers) concerning how to interpret the convention of how fabric and domain names are exposed.
* **Since nobody can remember the convention, henceforth, the convention will be that the tarball number will not necessarily reflect the API revision number.**
* Jeff to go bump up the revision number for the sockets provider.
* Jeff to add a comment in fabric.h reflecting the above convention.

**Intel proposal for Collective Offloads -** <http://mug.mvapich.cse.ohio-state.edu/static/media/mug/presentations/2016/Collective-Offload-OFI.pdf>

* Main objective is to improve the performance of collectives by enabling offloading of them.
* Currently there are two types of collectives.
* Four basic implementation techniques
  + Use MPI SEND and RECEIVE, but you don’t get the advantage of off loading.
  + Use high level Collective Library, but tends to obscure the algorithm being used, and doesn’t actually solve the problem.
  + Use Fabric-specific features
  + Generalized Pattern Offload – preferred, the focus of the presentation.
* Proposed changes:
  + Allow for preparation for future commands
  + Allow commands to be arranged in the correct order
  + Call to OFI to create the schedule structures
  + Run the schedule
  + Adds a ‘send atomic’ feature and flags to optimize schedule execution
* This mechanism could conceivably replace triggered ops.
* Concern that this may end up taking libfabric in a different direction. May want to consider this as a layer on top of libfabric, instead of replacing triggered ops. The same thing could be achieved by building mechanisms based on triggered ops, with provider-specific optimizations for performance boosts. Concern is about creating mission creep inside libfabric. It also may complicate testing.
* Current prototype is implemented inside a utility provider and does not appear overly complex, but a code walkthrough would illustrate the level of complexity.
* **Code walkthrough** will be *scheduled for four weeks from now (9/20)* to get a better feel for the complexity introduced.
* As part of the code walkthrough, Howard suggests a post mortem on triggered ops, which turned out to be not as useful as originally thought, but took a lot of work to implement.
* Question – did you look at building this on top of extended triggered ops? A – yes, and the extensions are almost equally as bad.
* The proposal is attempting to avoid vendor specific provider optimizations. But that begs the question, what should vendor specific provider optimizations be used for, if not for something like this?

**Agenda for 9/6:**

* Intel Proposal for RDMA Extensions for Non-Volatile Memory

**Agenda for 9/20**

* Code walkthrough on the collective offload proposal
* Post mortem on triggered ops (and maybe others) which turned out to be not as useful as originally thought.

**Webex link:** <https://cisco.webex.com/ciscosales/j.php?MTID=m9389b0513c9ae643d57e2381e254dcf5>  
Webex password: ofi

**OFIWG Download Site:** [www.openfabrics.org/downloads/OFIWG](http://www.openfabrics.org/downloads/OFIWG)

**Github:** <https://github.com/ofiwg/libfabric>

**OFI Software Download Site:** [www.openfabrics.org/downloads/OFI](http://www.openfabrics.org/downloads/OFIWG)

**Link to WebEx Recording** –

|  |  |
| --- | --- |
| [**Play recording**](https://cisco.webex.com/ciscosales/lsr.php?RCID=67e107907c6d430f8345dbf7dd941ed5) (39 min 25 sec) | |
| Recording password: Qi8MyFht |  |

|  |
| --- |
|  |

**Next regular telecon**

Next meeting: Tuesday, 9/6/16

9am – 10am Pacific daylight time