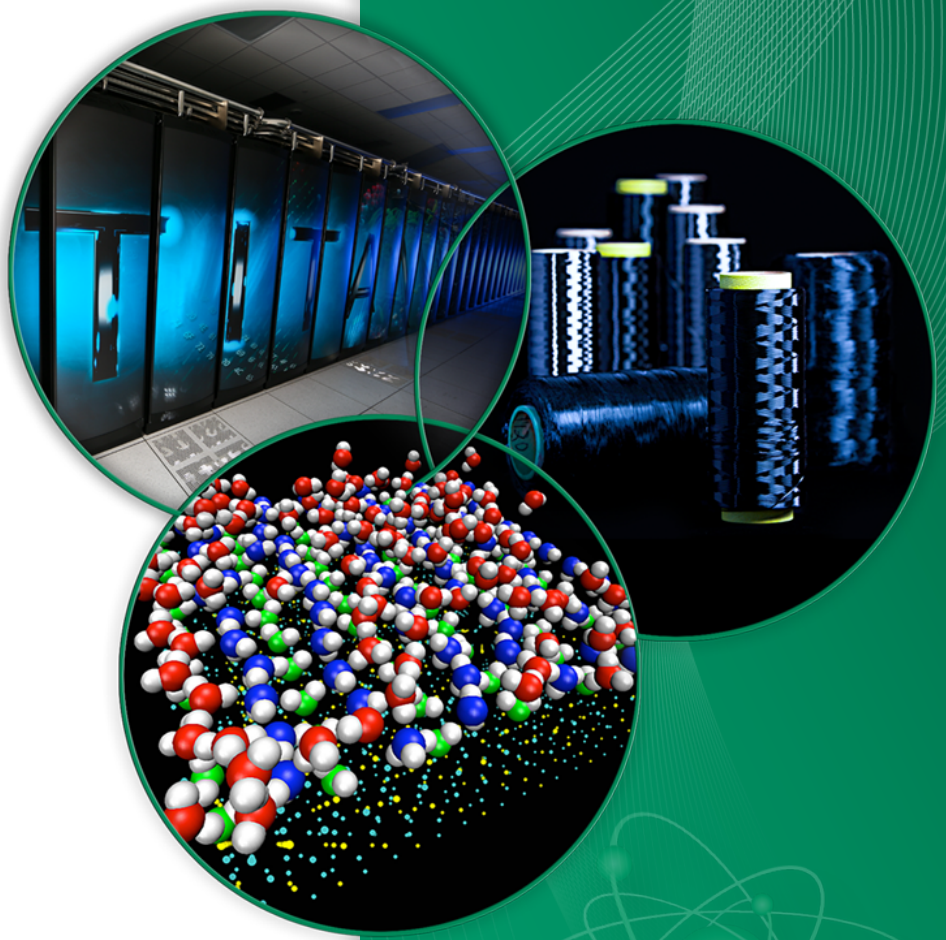
VERBS Error Codes discussion

Pavel Shamis



Typical MPI post send flow

- Using a code-snapshot from the OMPI codebase


```
        /* check for a send wqe */
        if (qp_get_wqe(ep, qp) < 0) {
            qp_put_wqe(ep, qp);
            OPAL_THREAD_LOCK(&ep->endpoint_lock);
            opal_list_append(&ep->pending_put_frags, (opal_list_item_t*)frag);
            OPAL_THREAD_UNLOCK(&ep->endpoint_lock);
            return OPAL_SUCCESS;
        }
.
.
.
         if(ibv_post_send(ep->qps[qp].qp->lcl_qp, &frag->sr_desc, &bad_wr))
            return OPAL_ERROR;
```

Verbs MLX5 provider

- QP.c

```
int mlx5_post_send(struct ibv_qp *ibqp, struct ibv_send_wr *wr,
                                   struct ibv_send_wr **bad_wr)
.
.
if (unlikely(mlx5_wq_overflow(&qp->sq, nreq,to_mcq(qp->ibv_qp.send_cq)))) {
                                  mlx5_dbg(fp, MLX5_DBG_QP_SEND, "work queue overflow\n");
                                   errno = ENOMEM;
                                   err = -1;
                                   *bad_wr = wr;
                                   goto out;
                        }
.
.
.
```

Keeping track of outstanding wqes (again !)

How the code should look

```
.
.
.
        if(ibv_post_send(ep->qps[qp].qp->lcl_qp, &frag->sr_desc, &bad_wr))
           if (errno == NOMEM) {
                OPAL_THREAD_LOCK(&ep->endpoint_lock);
                opal_list_append(&ep->pending_put_frags, (opal_list_item_t*)frag);
                OPAL_THREAD_UNLOCK(&ep->endpoint_lock);
                return OPAL_SUCCESS;
            } else {
                return OPAL_ERROR;

            }
```

But …

- Error codes are not documented in man pages
- It is not clear if the codes are well defined across providers
- Can we do better job ?