

Transparent link failover in NIXL/UCX communication backend

Evgeny Leksikov · Zihao Zhao · William Gallagher · Leonid Genkin · Yossi Itigin · Gal Shalom

1 The problem

- A single NIC or link failure aborts every rank in the job.
- In-flight transfers are lost.
- Recovery means restarting from the last checkpoint and repeating the work since it was written.
- Failure probability rises with cluster size and run length, so the largest jobs are hit most often.

The job keeps running through the fault: no restarts, no checkpoint rollback, transfers are retransmitted transparently

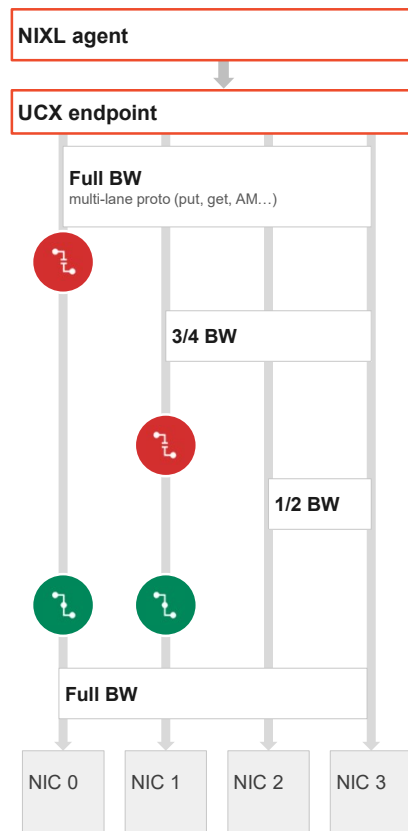
2 What we built

- A new level of error handling in UCX communication backend, built on UCX multi-lane support.
- Multiple candidate paths are kept between every pair of ranks, across NICs and transports.
- Reliability of transports detects the link and NIC failure with first completion with error. That does not add overhead for no failure cases.
- Failed RDMA operations migrate transparently to an alternate route.
- Failover and recovery is coordinated inside the UCX layer and independent for each endpoint.
- Currently supported TX operations: RDMA SEND, READ, WRITE. Other operations falls back to UCX EP failure.

3 Controlled fallback

- When the failed link returns, traffic shifts back to the preferred high-performance path.
- A controlled recovery protocol preserves message ordering and completion semantics across both transitions.
- UCX API change: new endpoint parameter `err_handling_mode=failover`.

4 Architecture



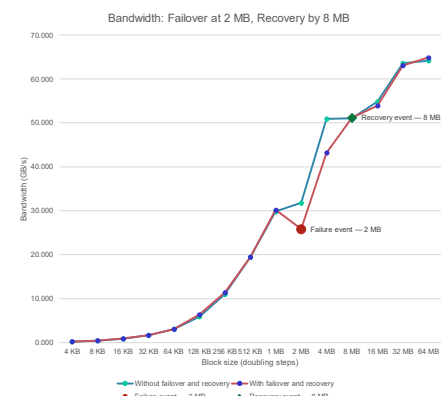
5 Failover sequence

- 1 Detect**
Detects a failed link or NIC on the first CQE with error, while transfers are still in flight.
- 2 Reconfigure**
Outstanding and pending operations are re-posted to the next best available path.
- 3 Continue**
Operations complete normally; endpoint observes no error and the job does not require restart.
- 4 Probing**
Keep alive mechanism periodically probing failed routes.
- 5 Fail back**
When the preferred path returns, the endpoint restores its original optimal configuration and traffic shifts back.

6 Evaluation setup

- Microbenchmarks with `ucx_perftest` on multi-NIC clusters.
- Bandwidth and latency measured during normal operation, failover and recovery.
- Faults injected at both the link and the switch level.
- Fault-free runs of the same configuration used as the baseline.

7 Results



- nixlbench WRITE DRAM-DRAM UCX
- 2 Phys NICs ConnectX-7 (2 + 1 virt ports)
- 1 phys link failure and recovery

8 Takeaways

- Link and NIC failures no longer end a distributed job for multi-NIC clusters.
- In-flight multi-lane data transfers survive the failures until at least one data path is available to continue.
- High bandwidth and low latency are maintained through failover and recovery, at modest overhead against fault-free runs.
- Fallback returns traffic to the preferred path(s) without breaking ordering or consistency.