FastForward Storage and I/O
7.3 – End-to-End Epoch Recovery
DAOS Server Failure Recovery

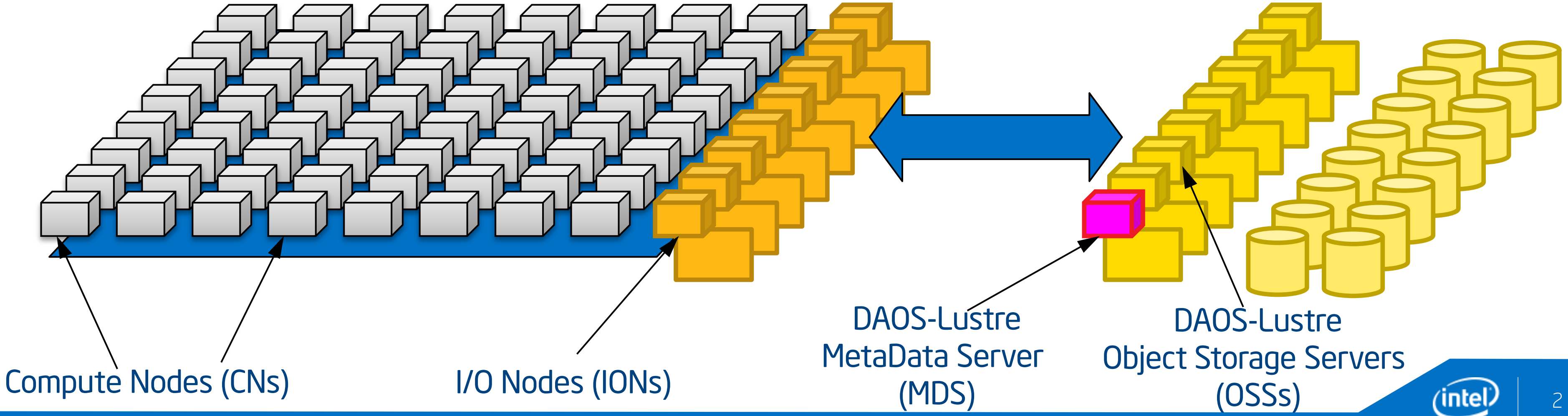
DAOS team, High Performance Data Division, Intel

March 31th, 2014

NOTICE: THIS MANUSCRIPT HAS BEEN AUTHORED BY INTEL UNDER THE SUBCONTRACT WITH LAWRENCE LIVERMORE NATIONAL SECURITY, LLC WHO IS THE OPERATOR AND MANAGER OF LAWRENCE LIVERMORE NATIONAL LABORATORY UNDER CONTRACT NO. DE-AC52-07NA27344 WITH THE U.S. DEPARTMENT OF ENERGY. THE UNITED STATES GOVERNMENT RETAINS AND THE PUBLISHER, BY ACCEPTING THE ARTICLE OF PUBLICATION, ACKNOWLEDGES THAT THE UNITED STATES GOVERNMENT RETAINS A NON-EXCLUSIVE, PAID-UP, IRREVOCABLE, WORLD-WIDE LICENSE TO PUBLISH OR REPRODUCE THE PUBLISHED FORM OF THIS MANUSCRIPT, OR ALLOW OTHERS TO DO SO, FOR UNITED STATES GOVERNMENT PURPOSES. THE VIEWS AND OPINIONS OF AUTHORS EXPRESSED HEREIN DO NOT NECESSARILY REFLECT THOSE OF THE UNITED STATES GOVERNMENT OR LAWRENCE LIVERMORE NATIONAL SECURITY, LLC.

Demonstration Goals

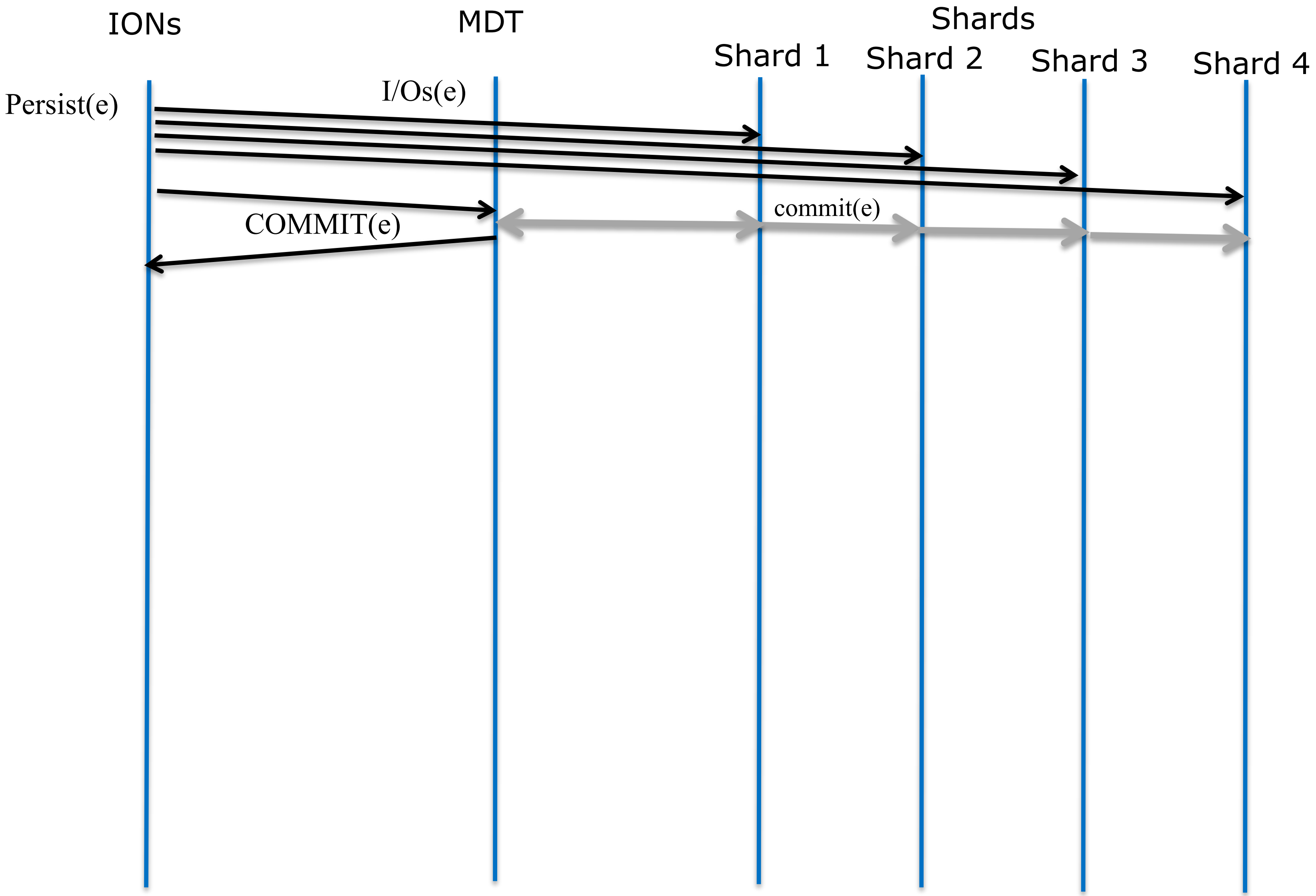
- Demonstrate a fully integrated stack
- Demonstrate resilience of the stack to failures
    - CN & ION failure recovery already demonstrated
    - Simulate transient failure of the DAOS MDS
    - Simulate transient failure of a DAOS OSS
    - Simulate transient failure of multiple DAOS OSSs
    - Simulate permanent failure of a DAOS OSS


Compute Nodes (CNs)
I/O Nodes (IONs)
DAOS-Lustre MetaData Server (MDS)
DAOS-Lustre Object Storage Servers (OSSs)

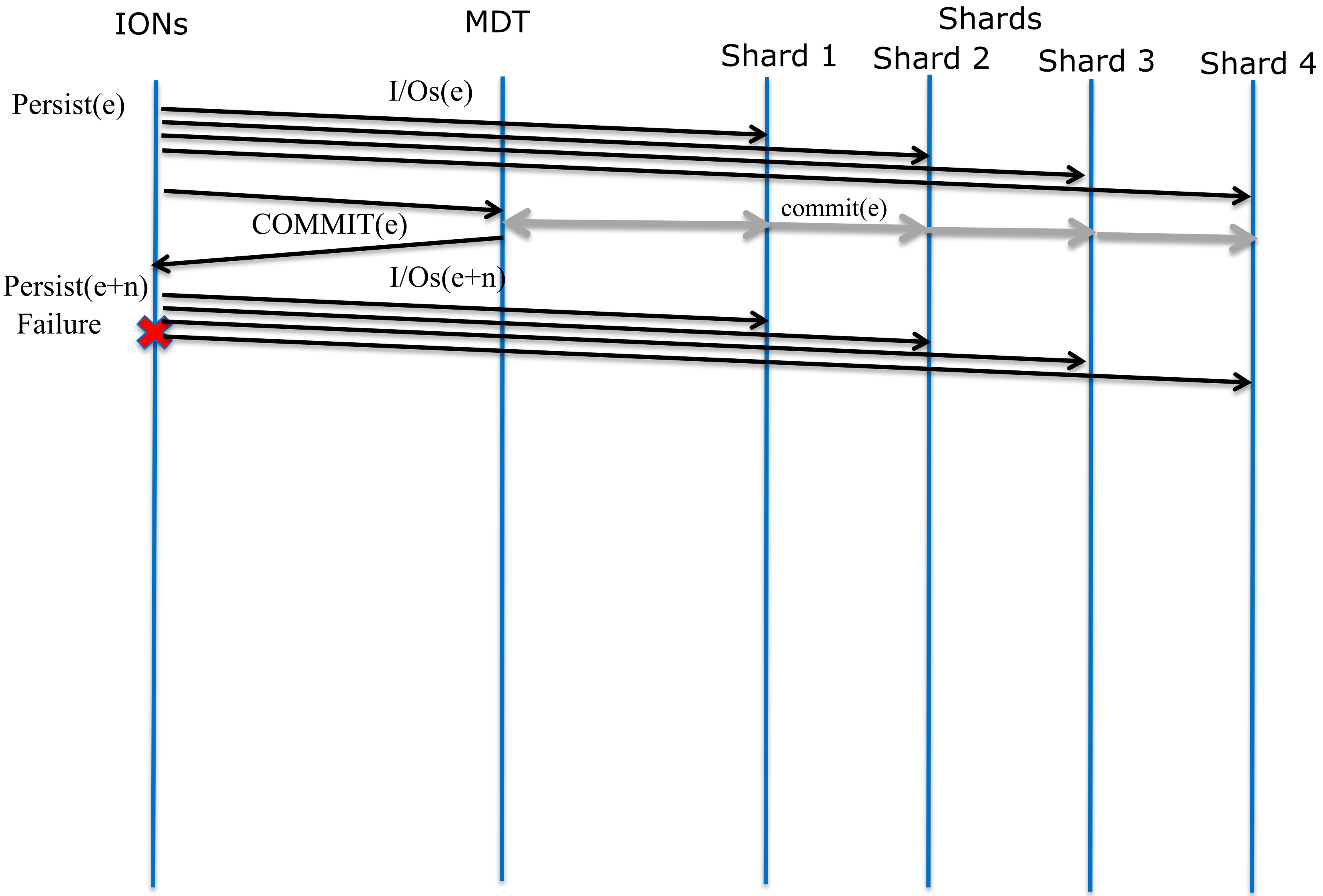
Test Environment

- VPIC runs on 4 CNs: lola-[20,25-27]
- 4 IONs: lola-[12-15] using flash & Lustre cross-mounts to share data stored in burst buffers
- 4 OSSs: lola-[16-19] using flash as well
- 1 MDS: lola-2 stored on a JBOD disk
- Zero copy disabled in VOSD
  - Major problems found in ZFS patch implementing block migration from intent log to DAOS object
  - New less intrusive approach is being implemented
- Failures are simulated by rebooting nodes via IPMI
- "Screen" session with a shell on each node

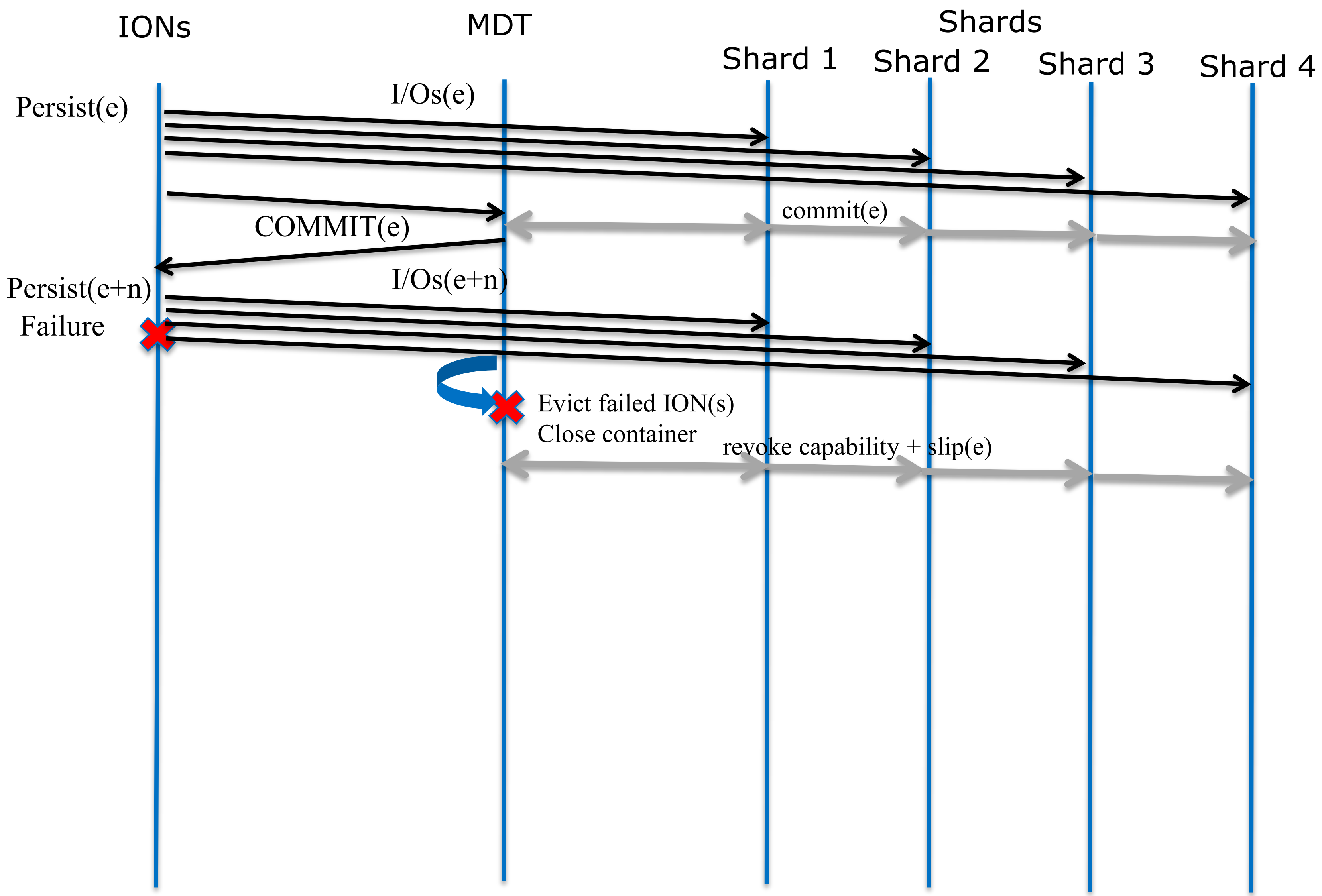
ION failure: what happened at DAOS level …



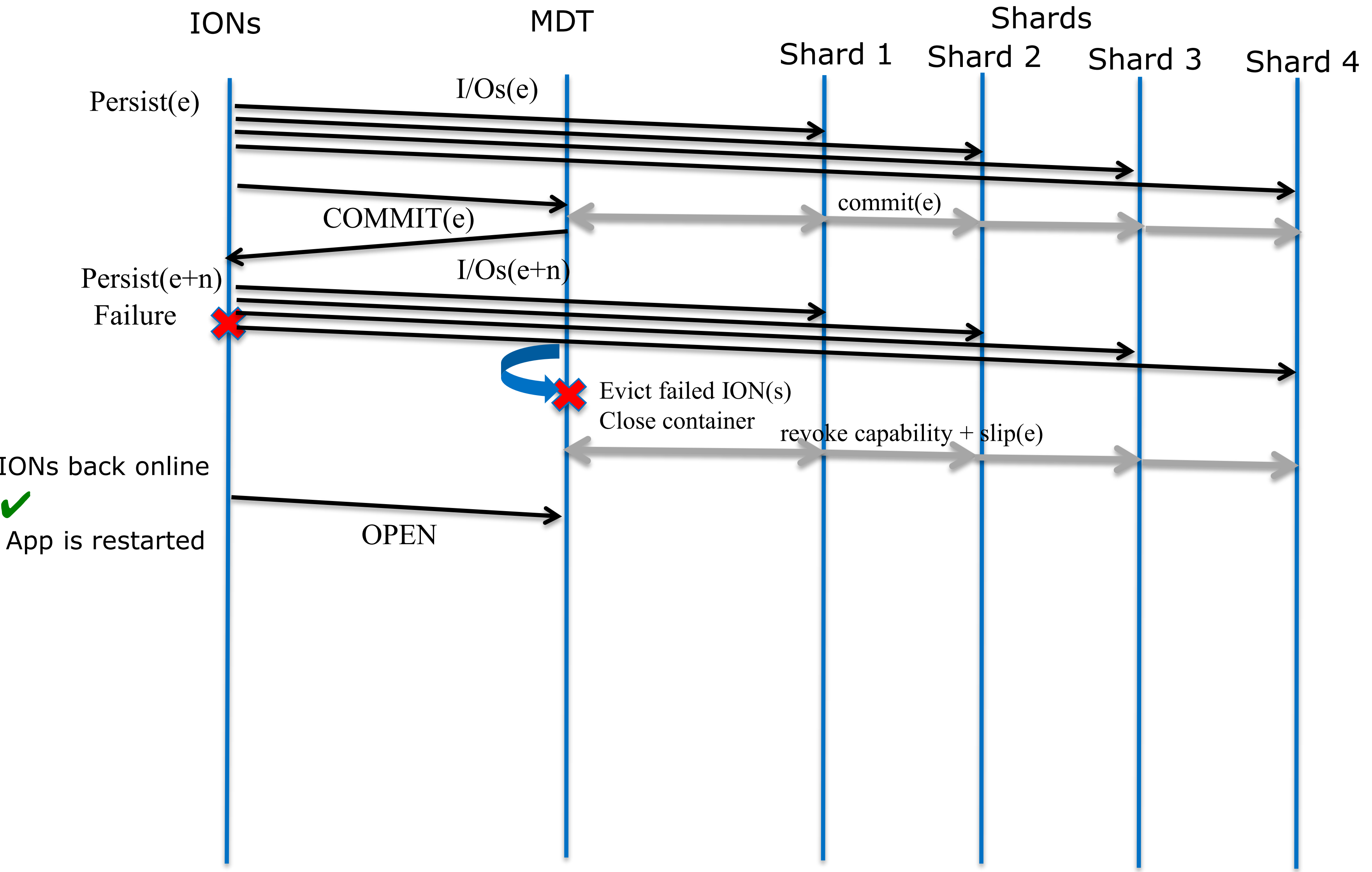
ION failure: what happened at DAOS level …



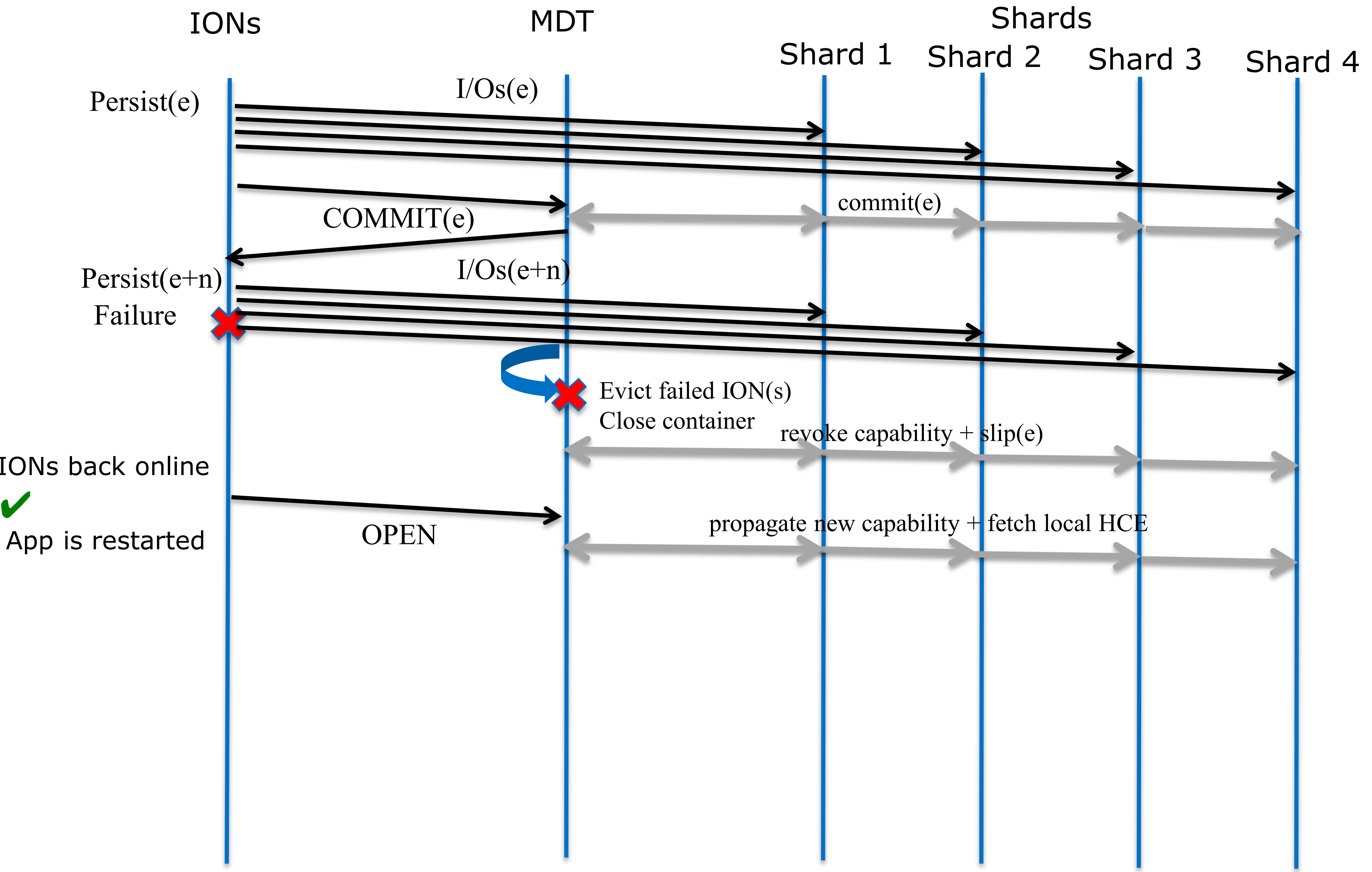
ION failure: what happened at DAOS level …



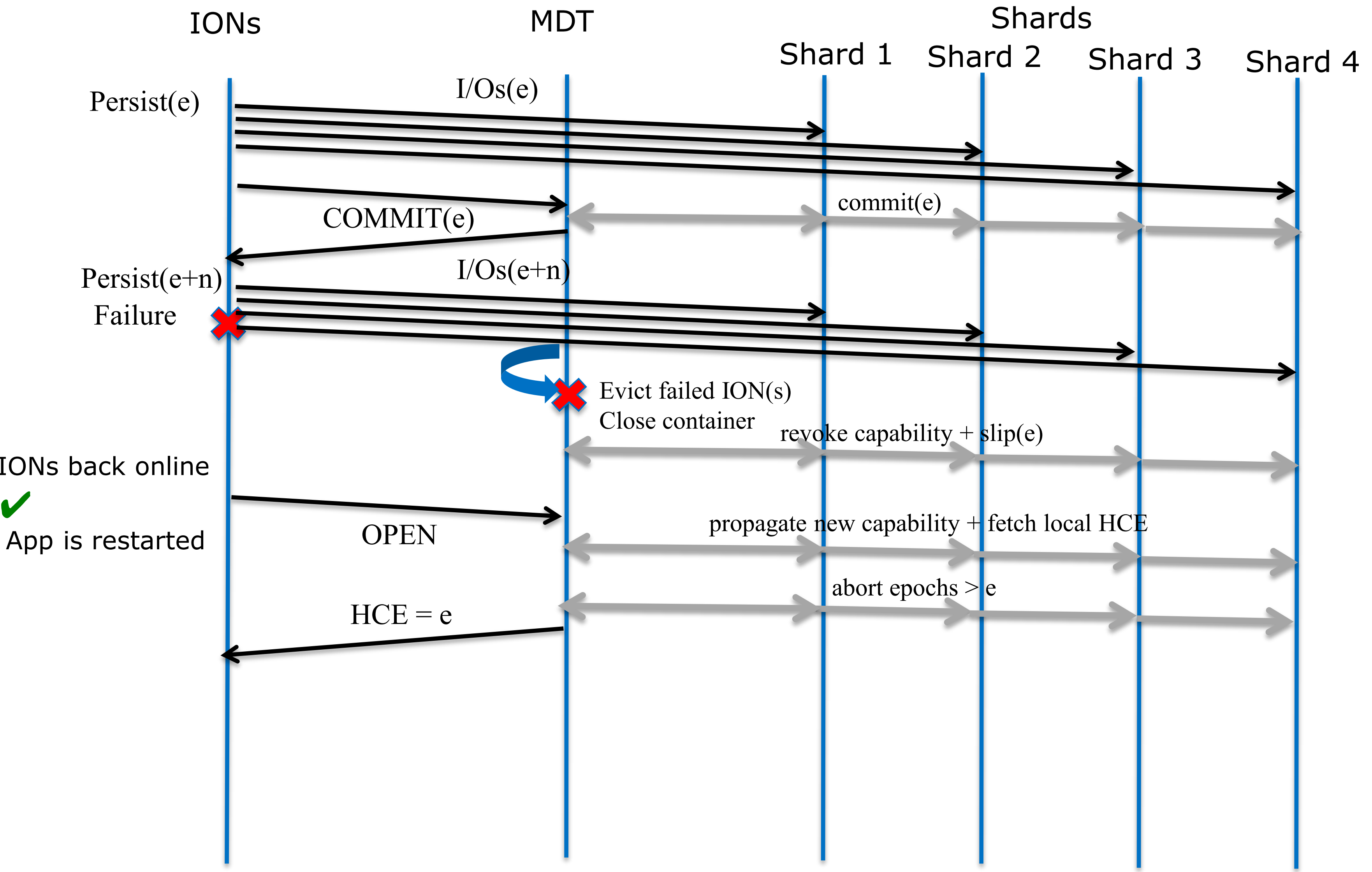
ION failure: what happened at DAOS level …



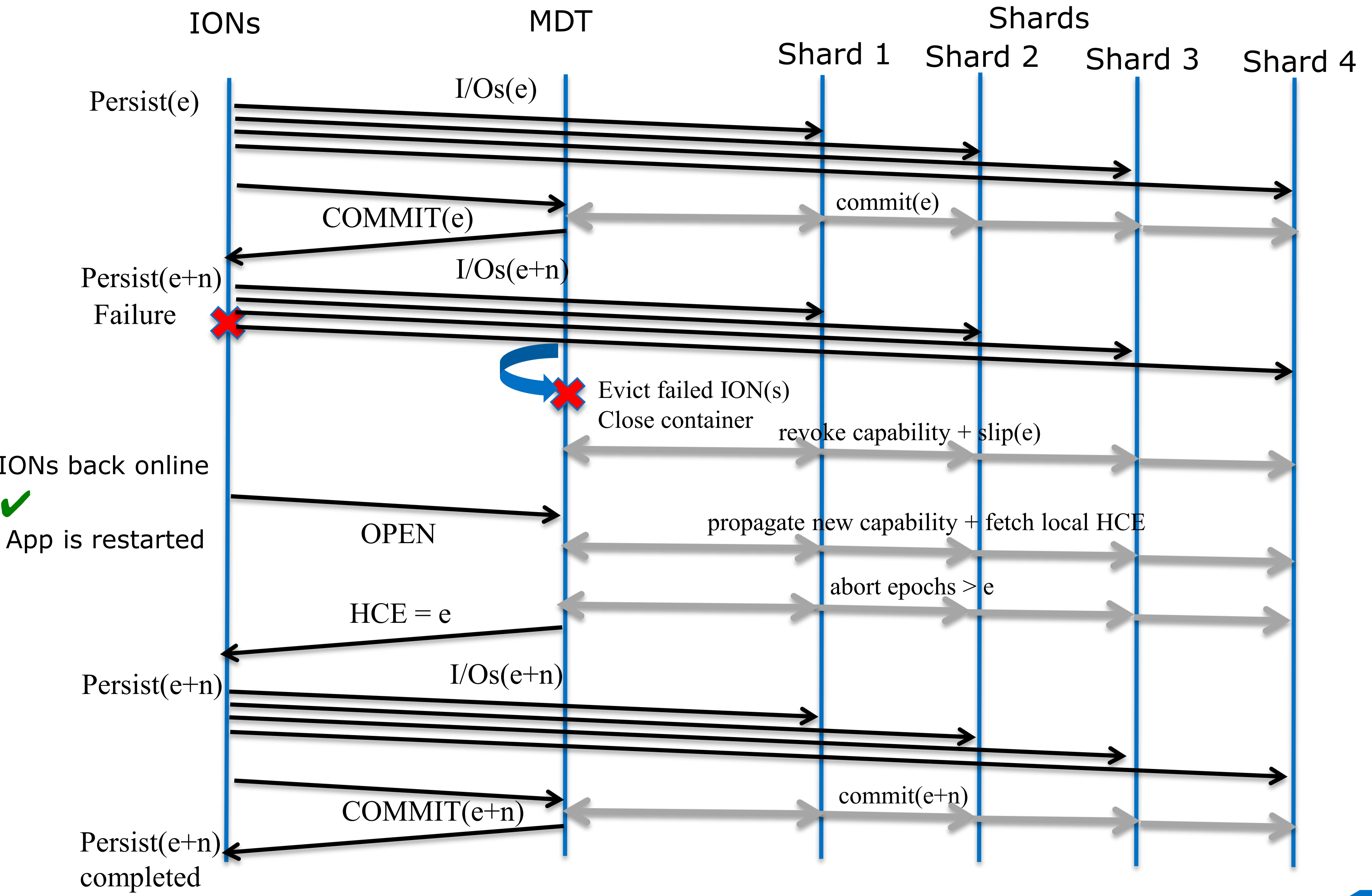
ION failure: what happened at DAOS level …



ION failure: what happened at DAOS level …



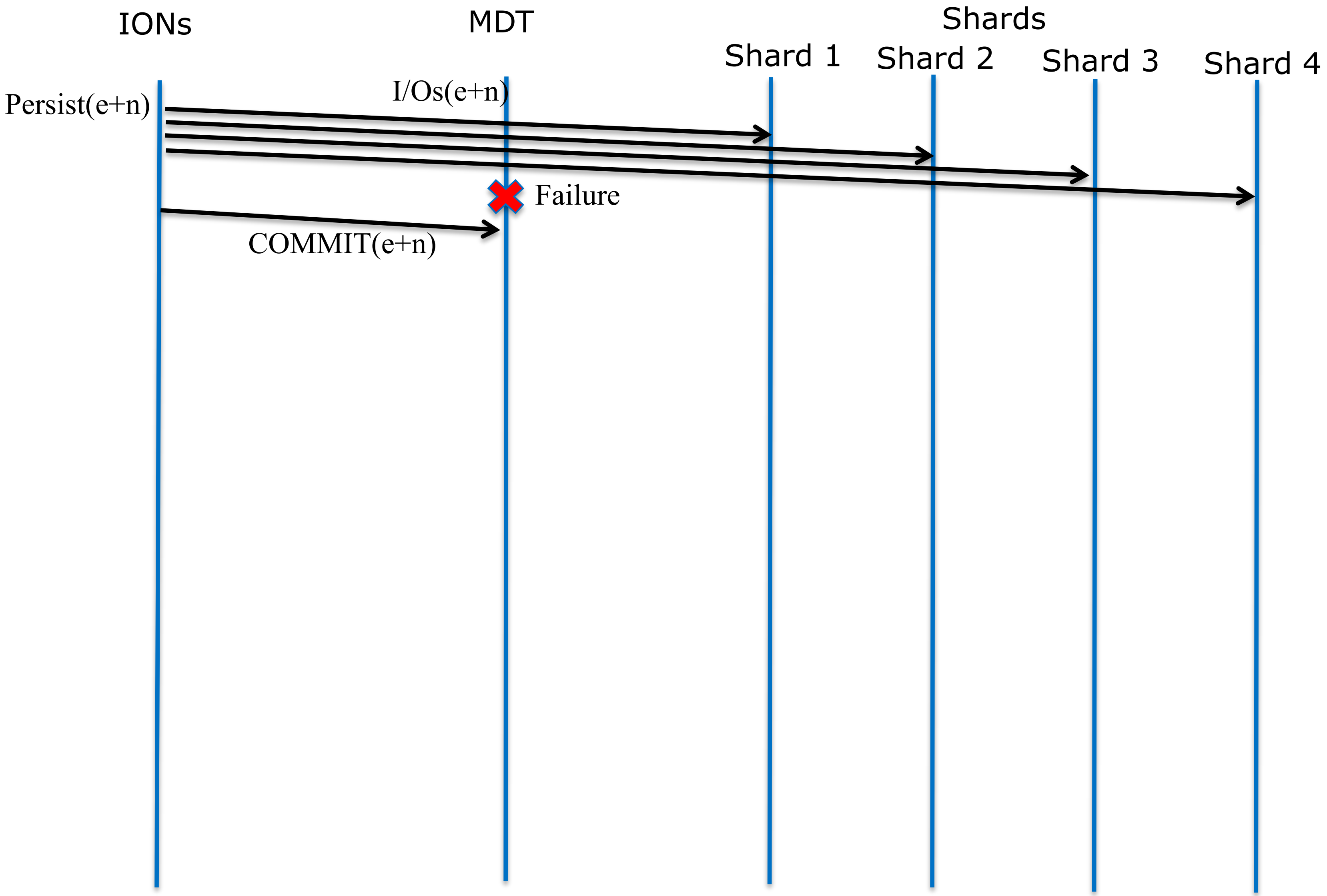
ION failure: what happened at DAOS level …



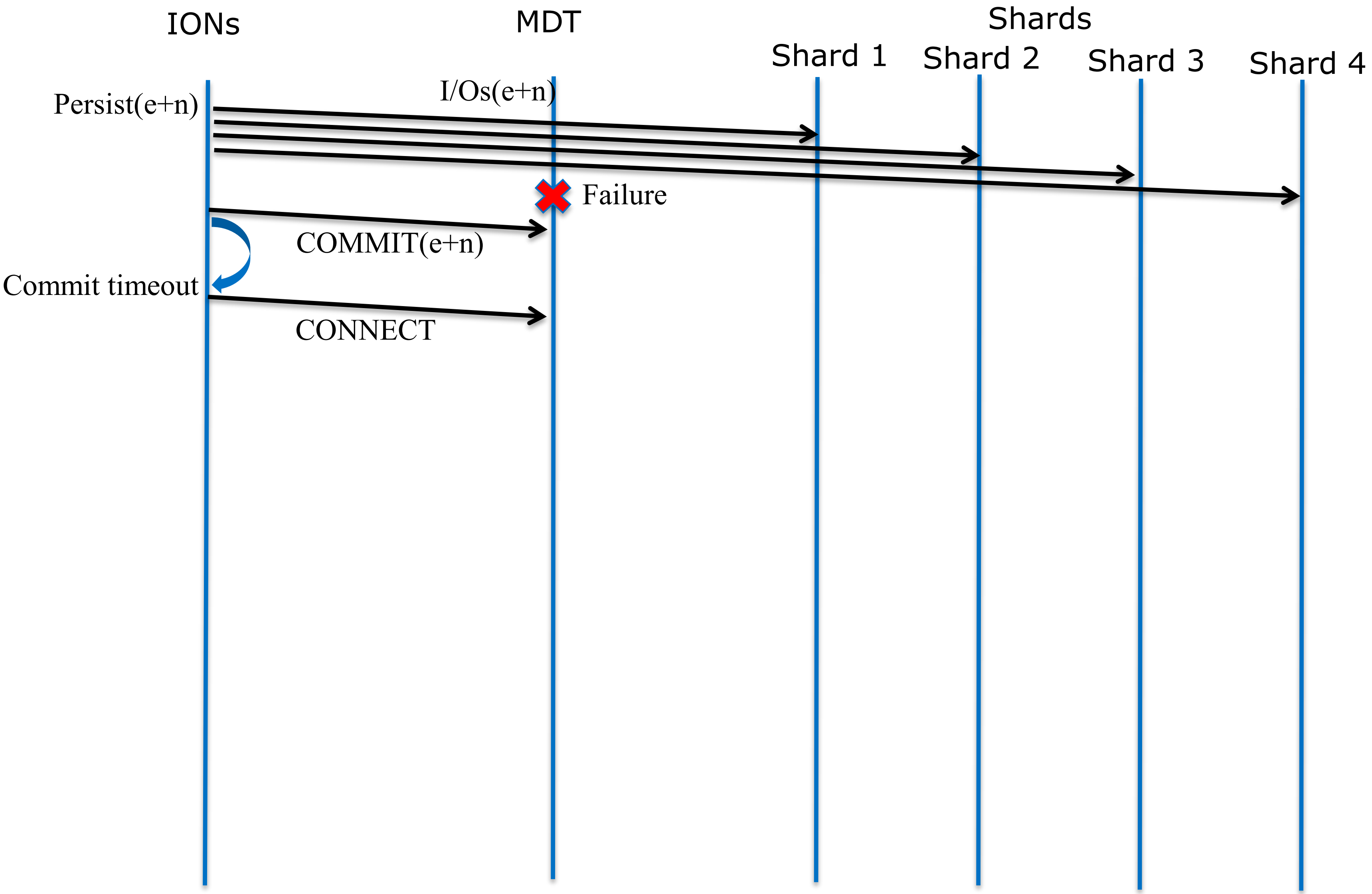
Transient MDS Failure

- Let’s start the demo
  - Run VPIC Application on CNs & IONs
  - Power cycle the MDS (lola-2)
  - Remount the MDT once the MDS is back online
  - Wait for VPIC to complete
  - Clear all burst buffers
  - Run VPIC again in verify mode to check consistency of DAOS data
- Expected result
  - Application continues running until it needs to communicate with the MDT (epoch_query/commit/slip)
  - Application waits for MDT to be up
  - Application resumes and completes successfully
  - Second VPIC run completes successfully with no corruption found

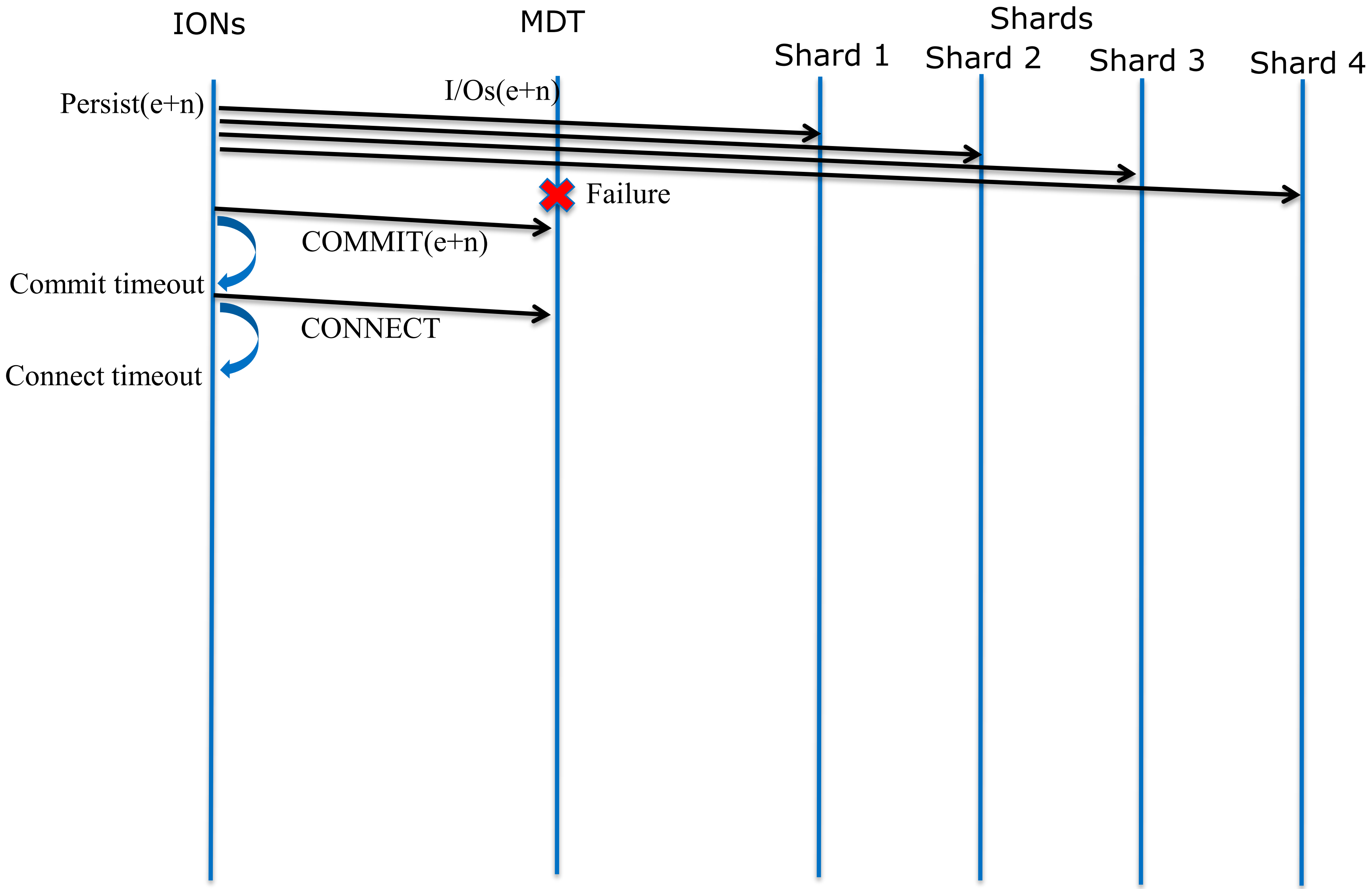
MDS failure: what is going on?



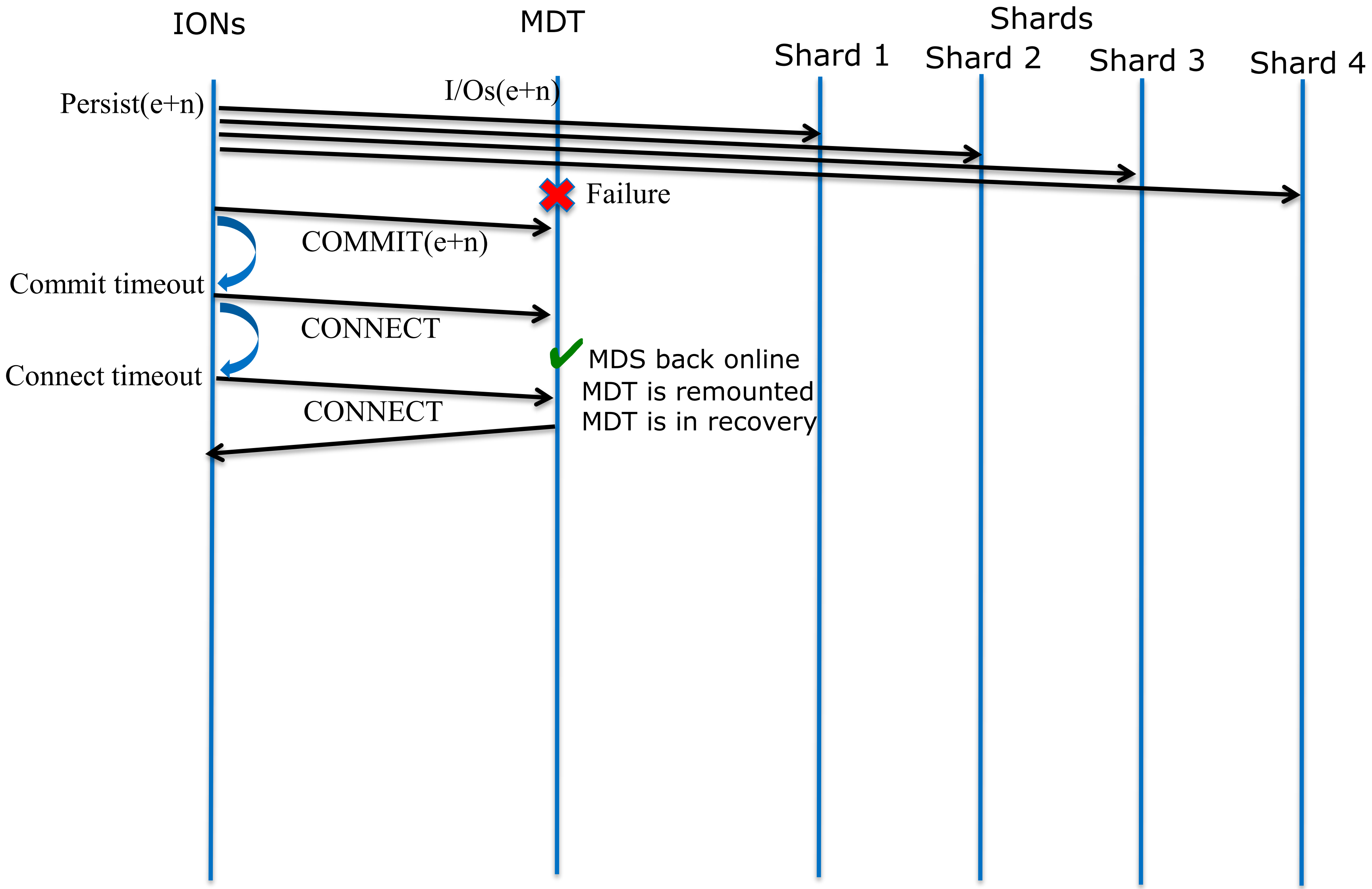
MDS failure: what is going on?



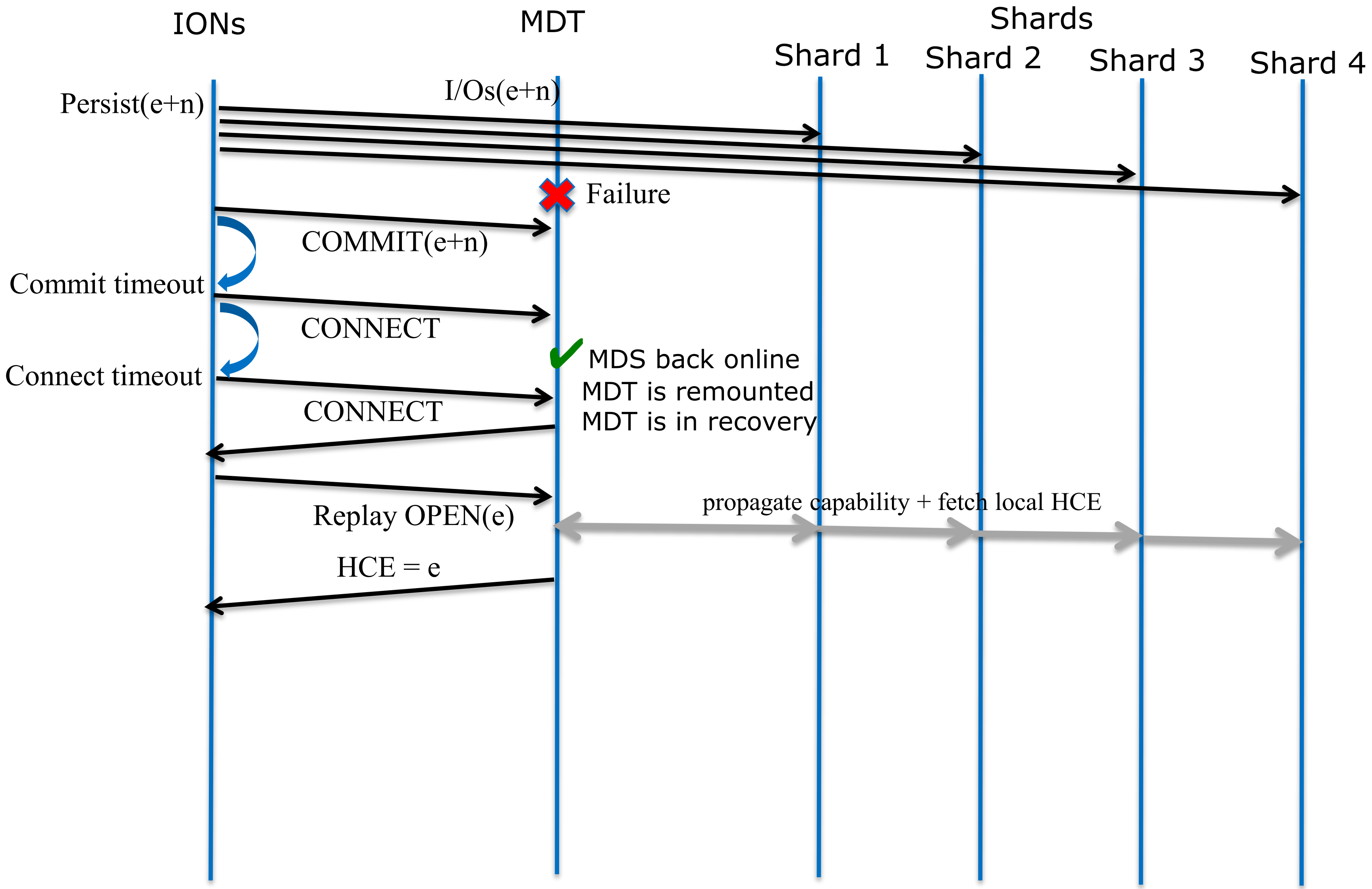
MDS failure: what is going on?



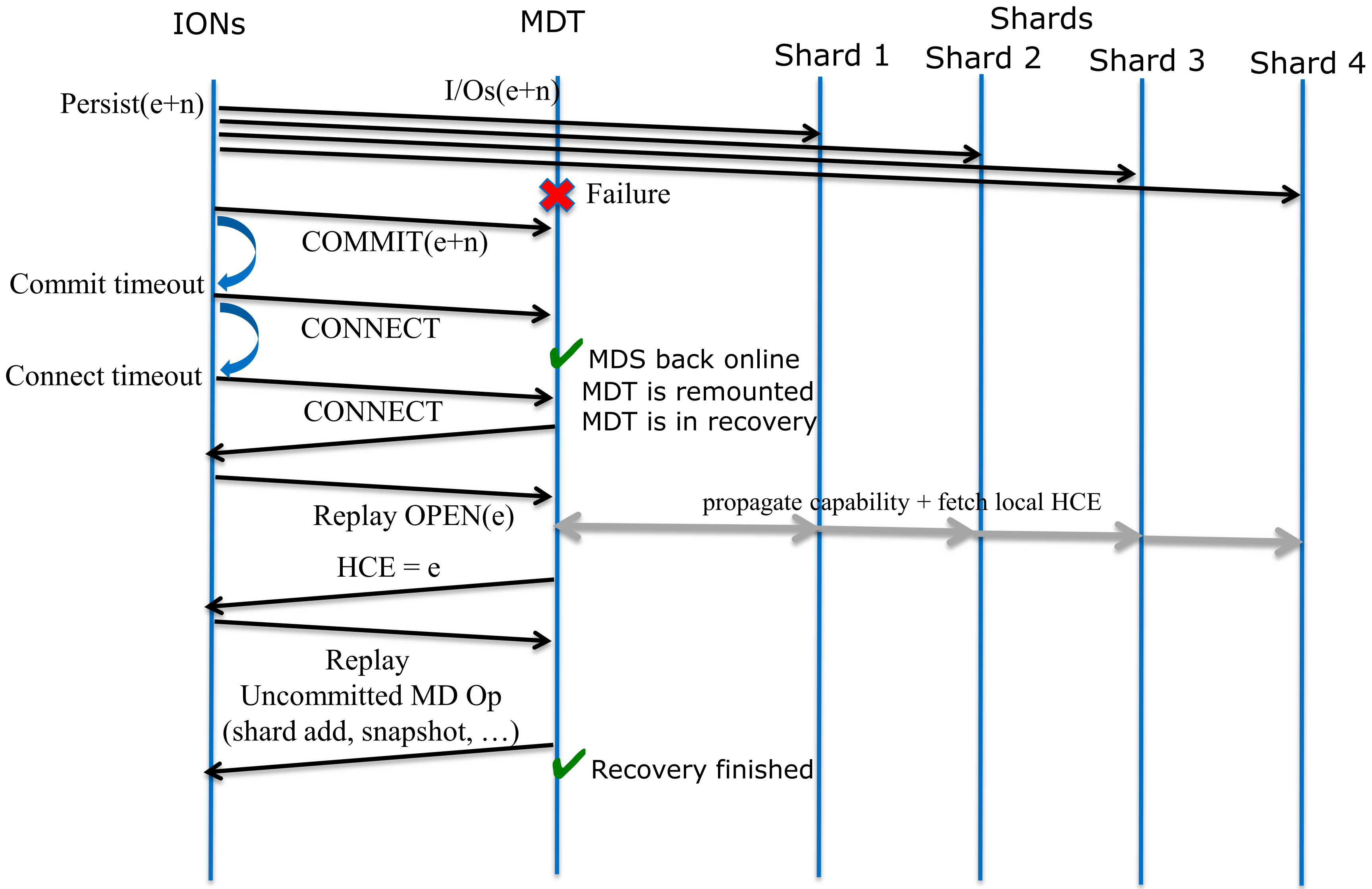
MDS failure: what is going on?



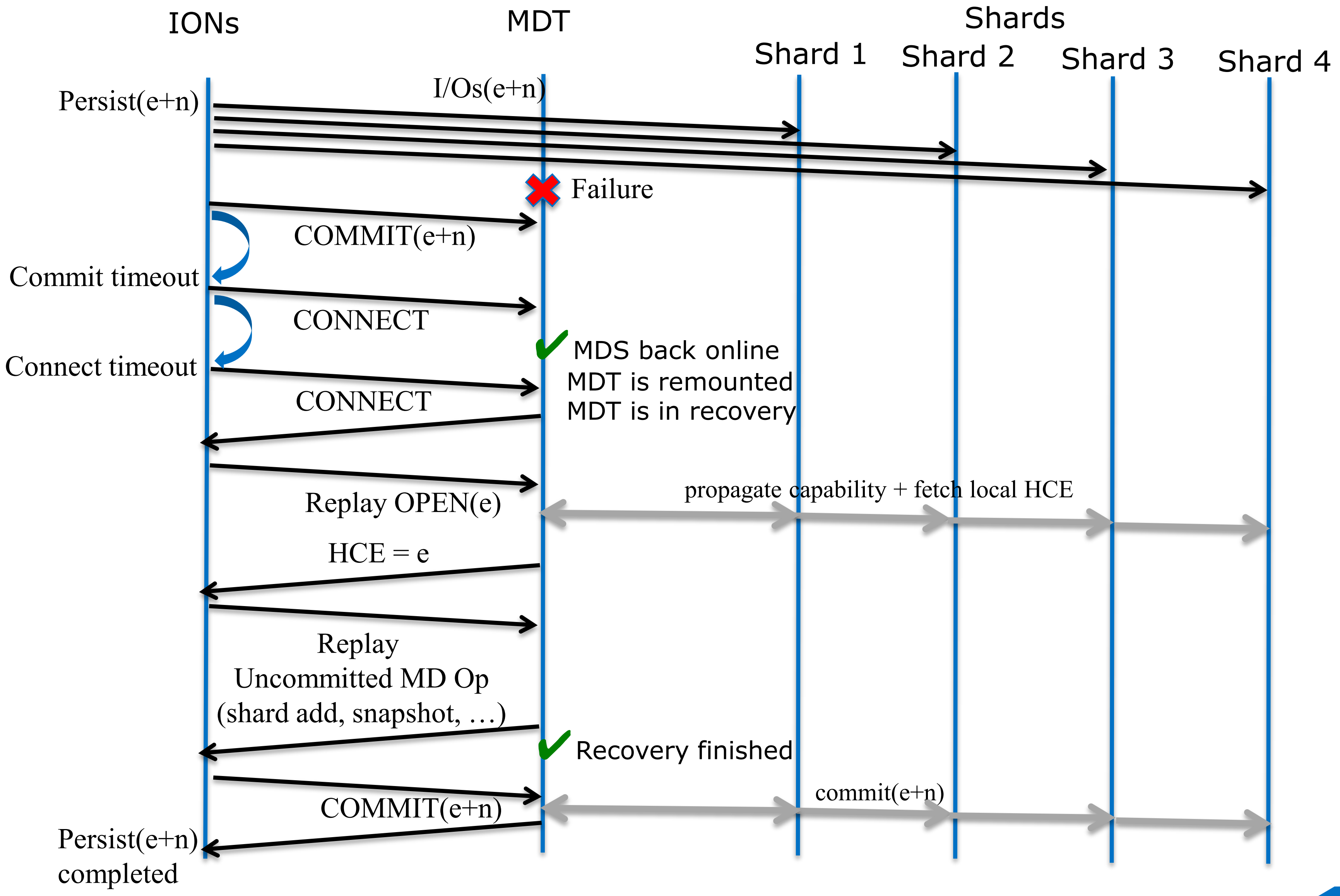
MDS failure: what is going on?



MDS failure: what is going on?



MDS failure: what is going on?



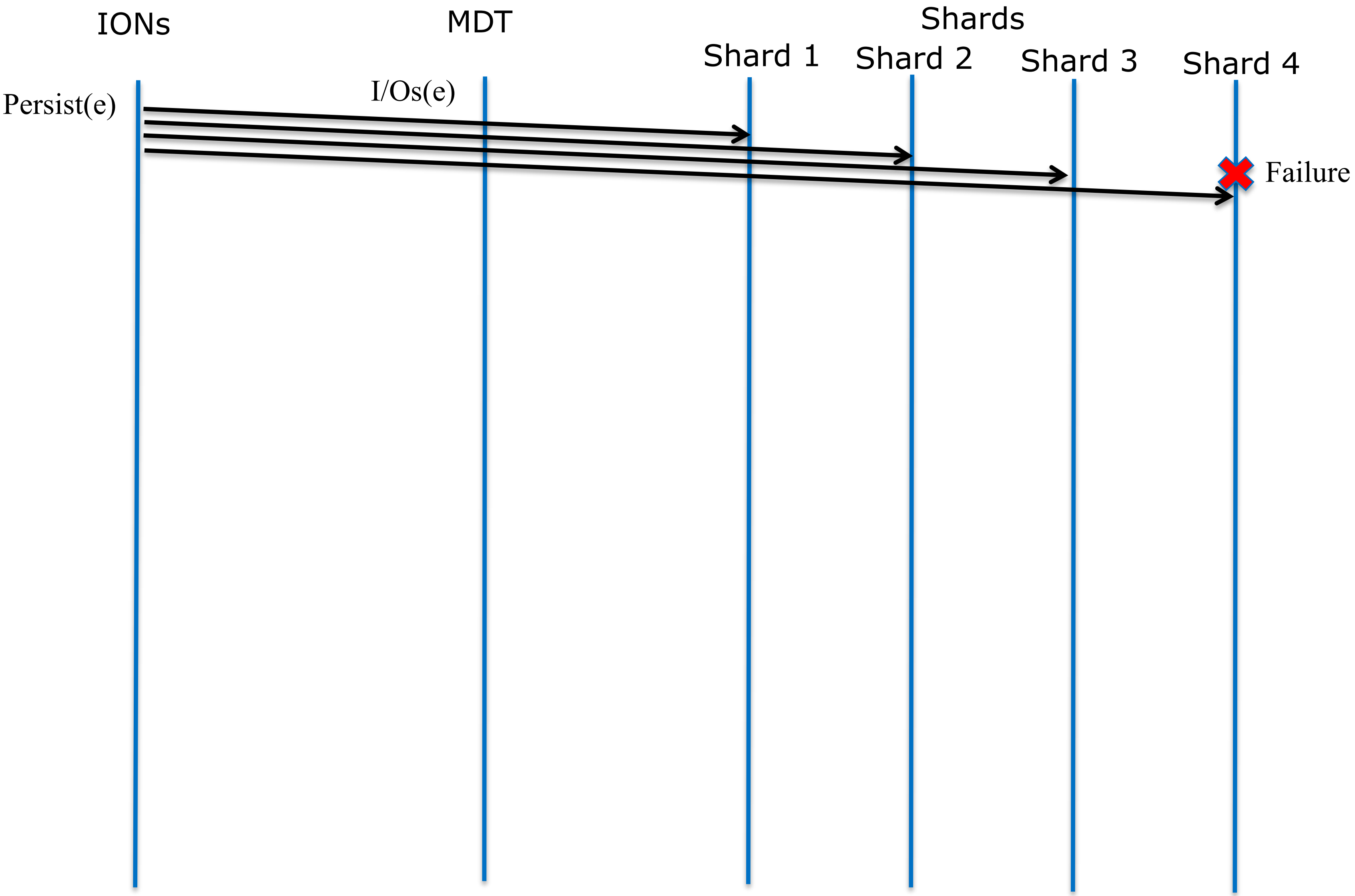
Transient OSS Failure (1/2)

- Let's start the demo
  - Run VPIC Application on CNs & IONs
  - Power cycle an OSS (lola-16)
  - Remount the OSTs once the OSS is back online
  - Wait for VPIC to complete
  - Clear all burst buffers
  - Run VPIC again in verify mode to check consistency of DAOS data

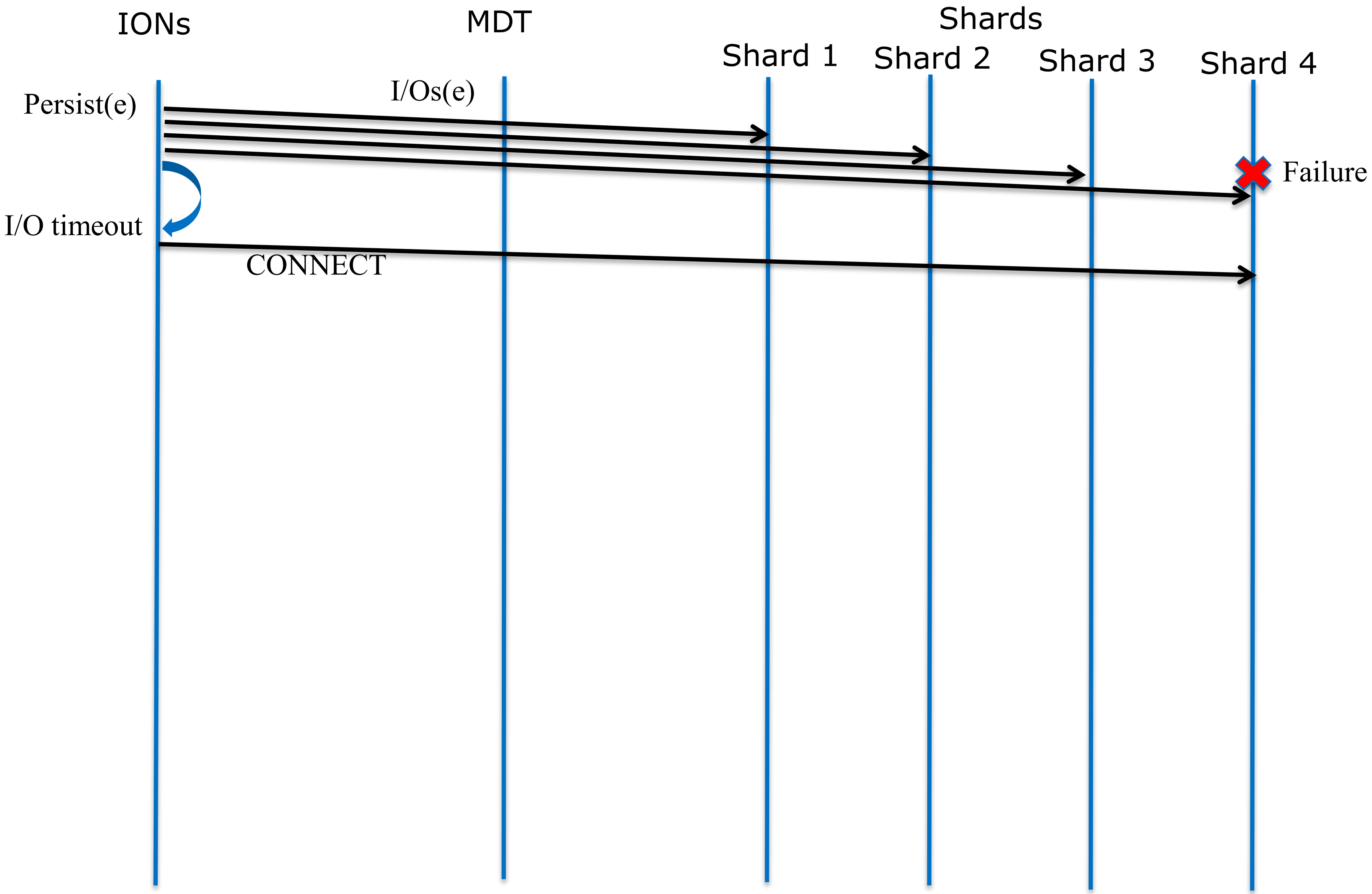
Transient OSS Failure (2/2)

- Expected result
  - Application continues running until it needs to communicate with an OST which is down
    - Might happen through a server collective issued by the MDT
  - Application waits for the OST to be back online
  - Persist() call might fail if OST has lost I/Os on disk
    - HDF retries persist in this case
  - Application resumes and completes successfully
  - Second VPIC run completes successfully with no corruption found

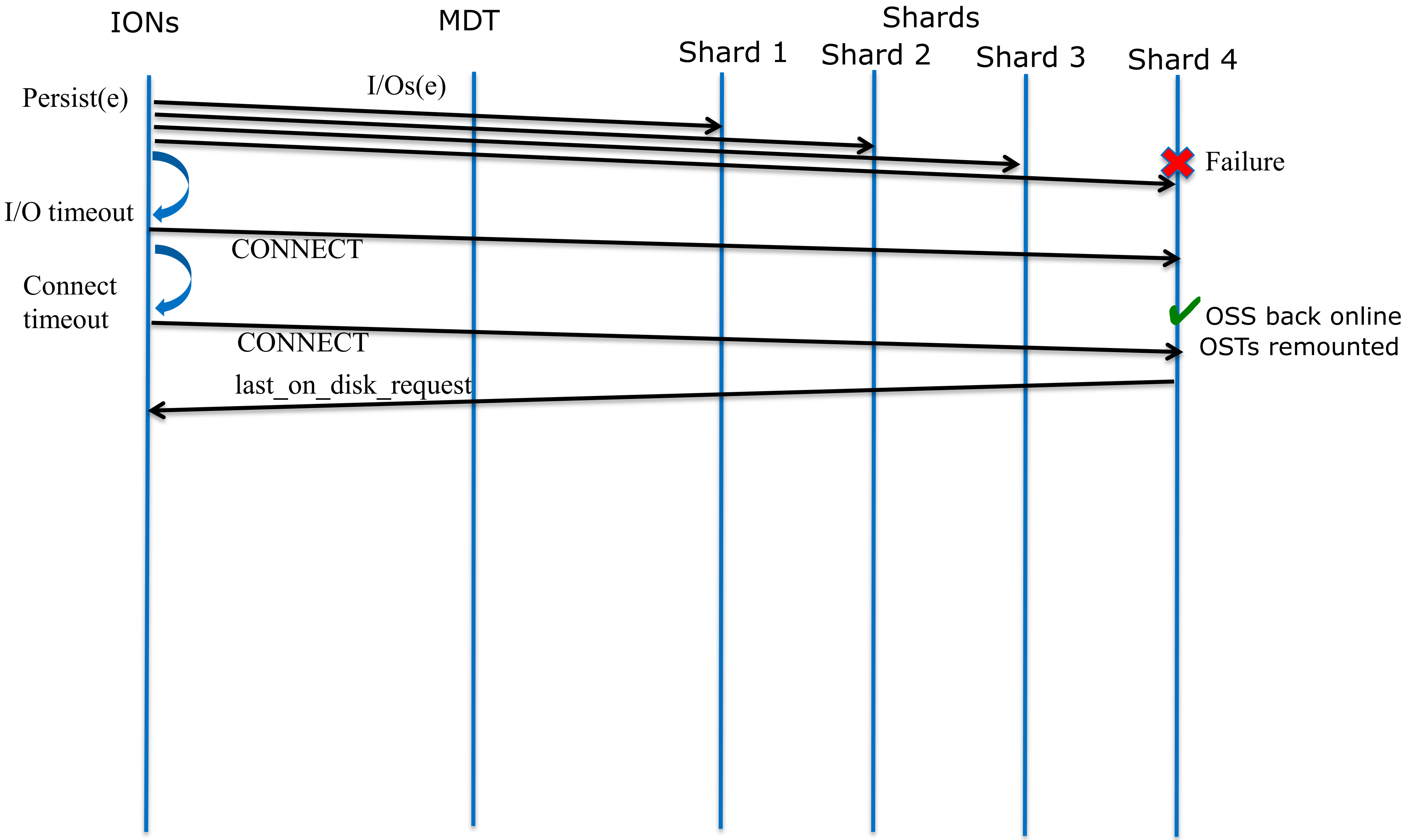
OSS failure: what is going on?



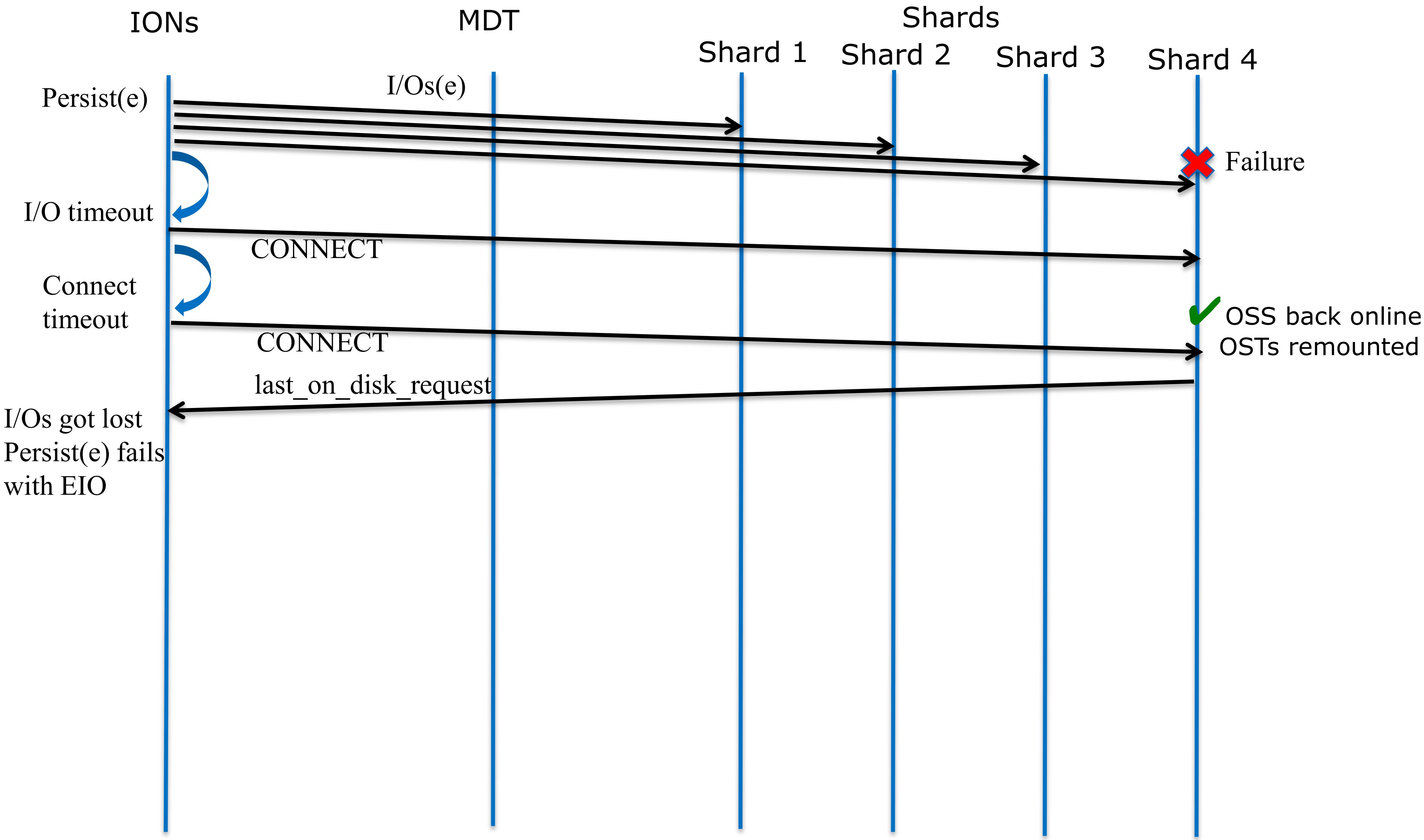
OSS failure: what is going on?



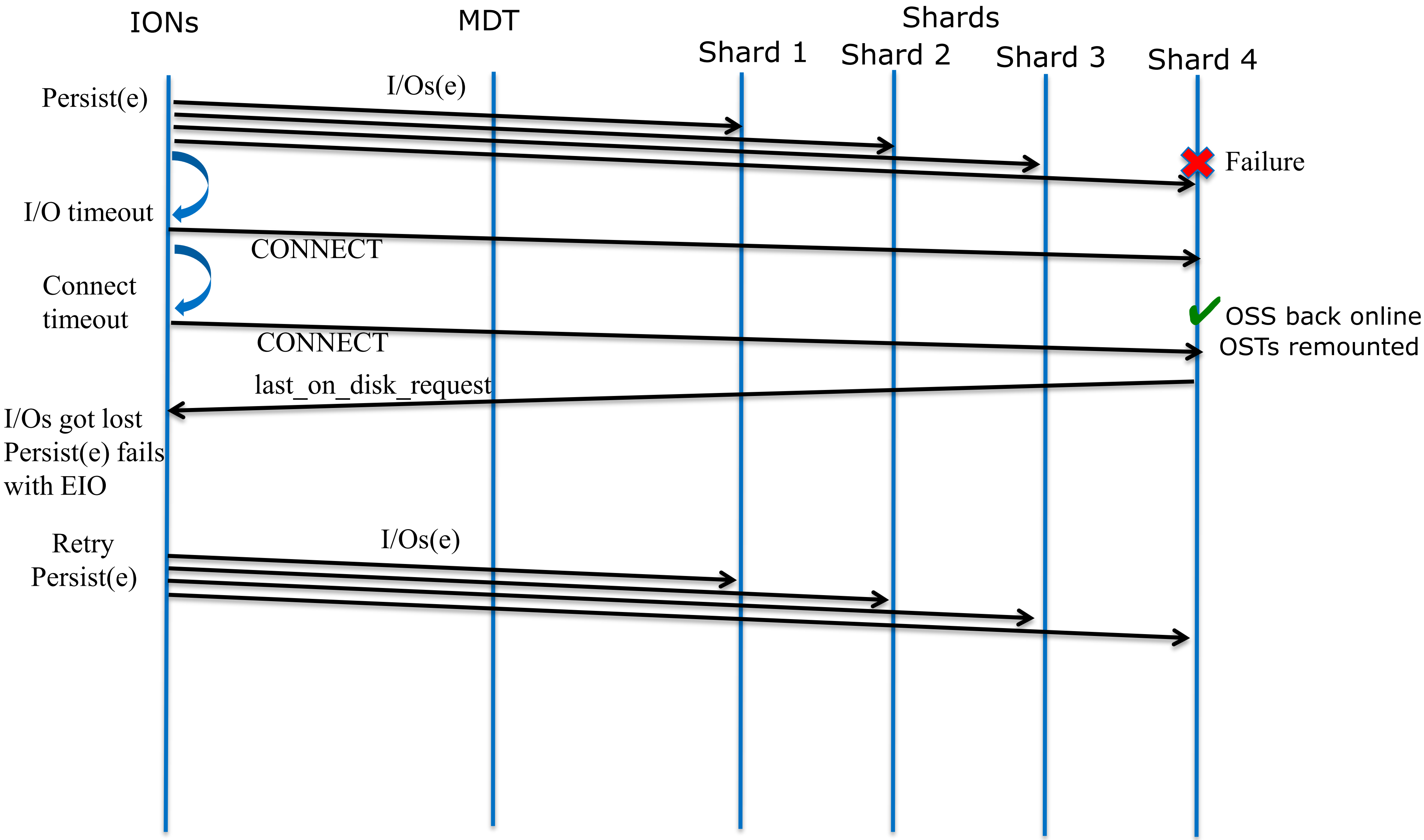
OSS failure: what is going on?



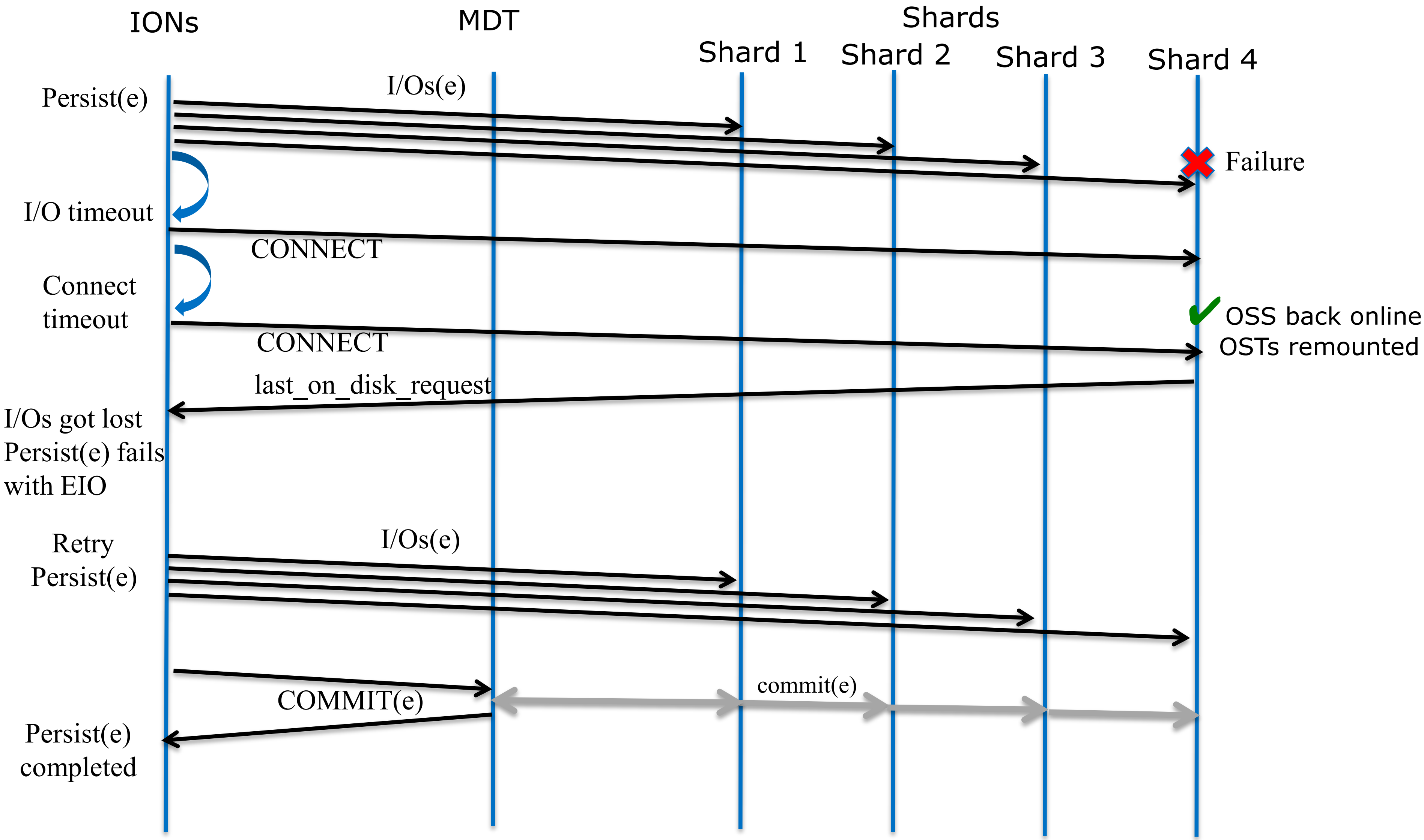
OSS failure: what is going on?



OSS failure: what is going on?



OSS failure: what is going on?



Failure of Multiple OSSs

- Run same demo, but power cycle all OSSs this time
- Same mechanisms involved and same result expected

Permanent OST Failure

- Let’s run the demo
  - Run VPIC Application on CNs & IONs
  - Mark an OST as permanently deactivated with conf_param
  - Any attempt to communicate with the deactivated OST fails with ESHUTDOWN (108)
- Expected result
  - VPIC should fail since some shards are not accessible any more
  - DAOS-HA required to handle such permanent failures

Demonstration Materials

- Source code still available on git.whamcloud.com
  - repository ff/daos_lustre, tag v1.2_DAOS
  - http://git.whamcloud.com/?p=ff/daos_lustre.git
- Built as usual Lustre tree:
  - autogen.sh
  - configure –with-linux=… –with-spl=… –with-zfs=…
  - make & make install (or make rpms)
- This slide deck & scripts are uploaded to the wiki
  - https://wiki.hpdd.intel.com/display/FF/Project+Quarter+7
  - “7.3 DAOS Demo Slides” & “7.3 DAOS Scripts”