BLUE WATERS

SUSTAINED PETASCALE COMPUTING

High Availability on Blue Waters File Systems
&
User Experiences

Kalyana Chadalavada
Jing Li, Sharif Islam
SEAS

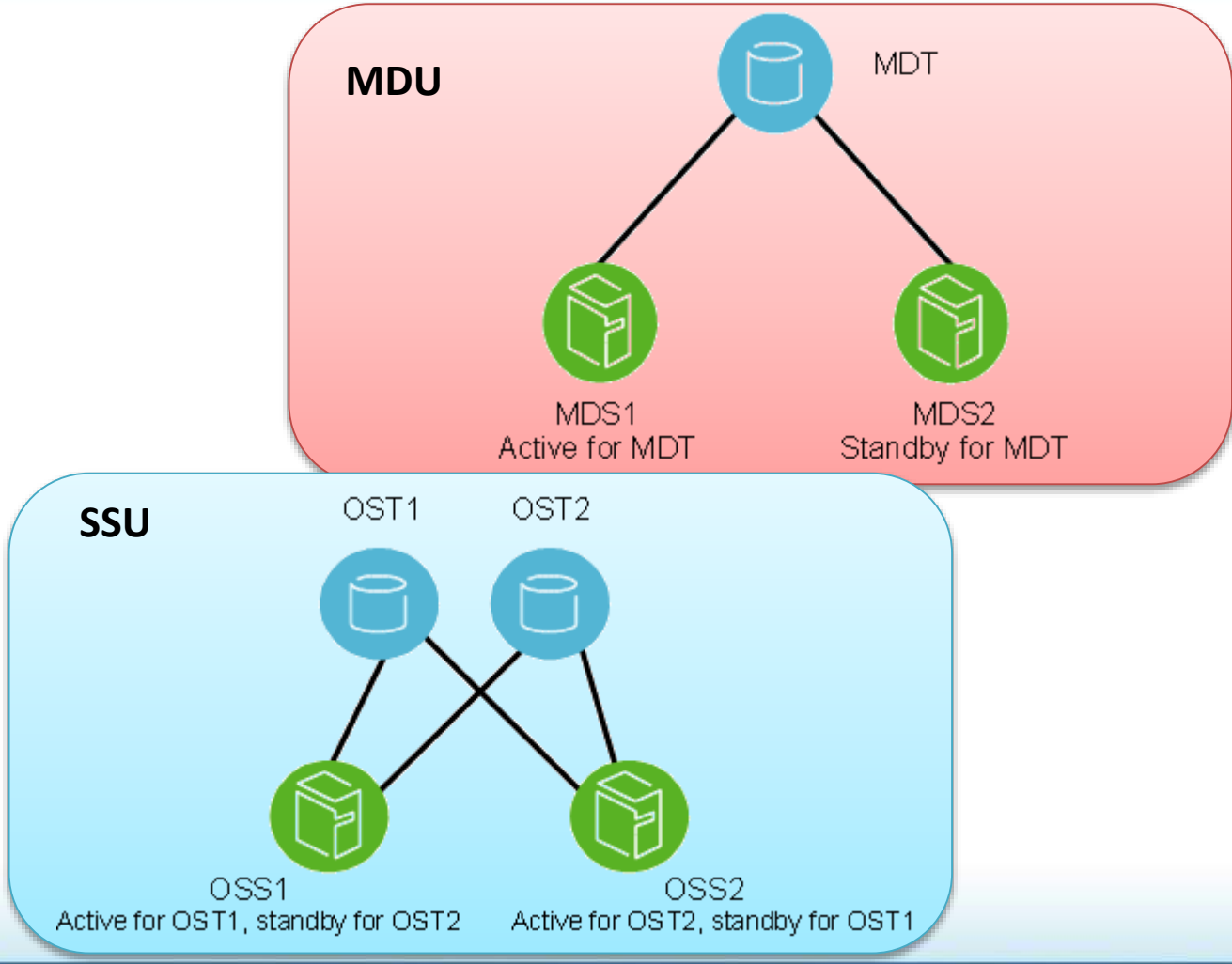


GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

CRAY

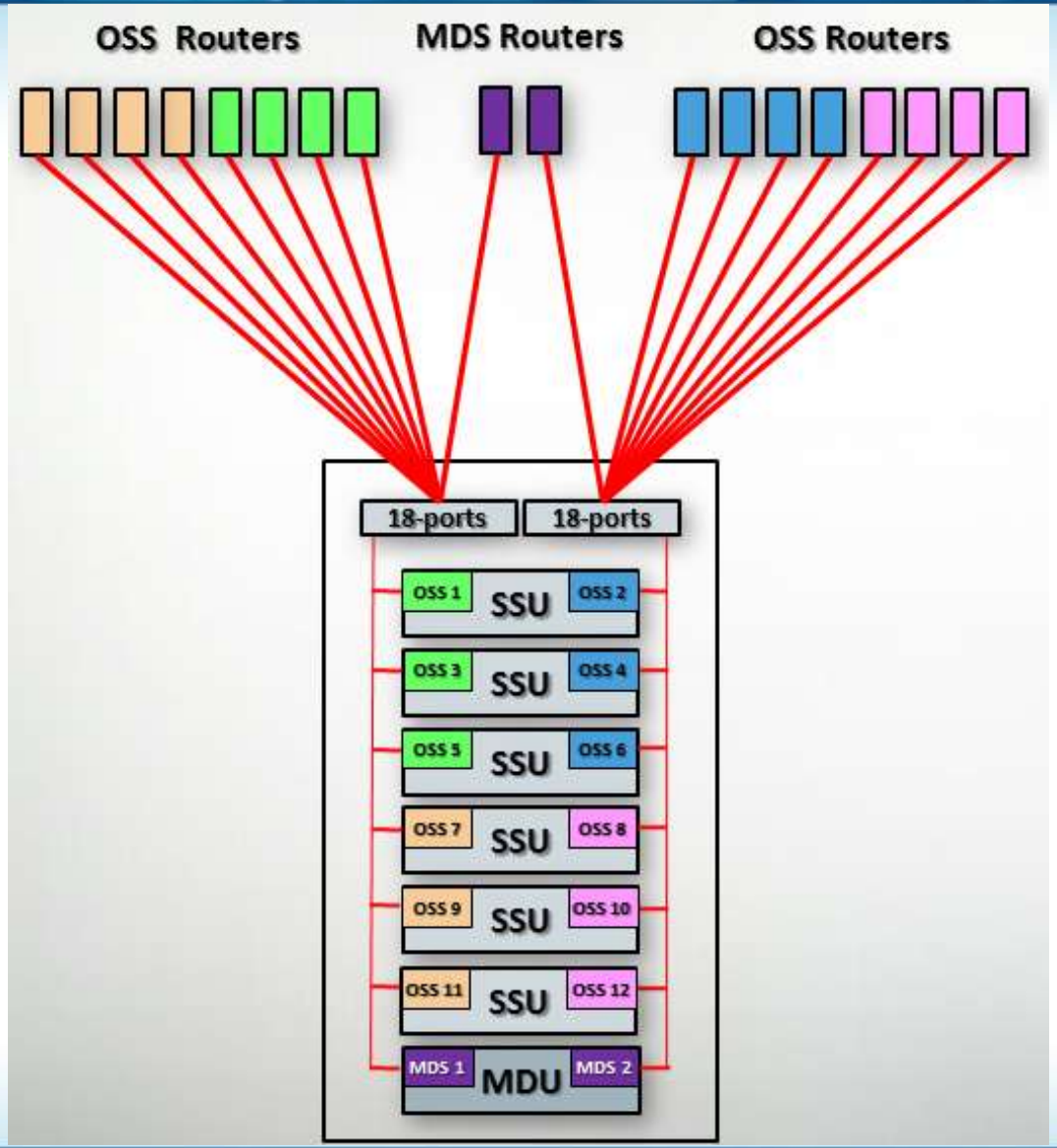
High Availability on BW

- Object Storage Target (OST)
  - RAID 6 – up to 2 disk failures
- Object Storage Server (OSS)
  - Active:Active fail over pair
- Metadata Target (MDT)
  - RAID 6 – up to 2 disk failures
- Metadata Server (MDS)
  - Active:Passive fail over pair



High Availability on BW

- Lustre Networking (LNET)
  - 4:4 Active:Active fail over pair



Possible Failures & Recovery Paths

- HSN (Gemini) fail
  - No recovery
- HSN Quiescence
  - No failure unless timeout triggered, client evicted, error, reconnect, replay
- OSS Failure, MDS failure
  - OST, MDS Failover, requests will wait
- OST Failure, MDT Failure
  - Extremely low probability of more than two disks failing
- LNET Failure
  - Errors possible on eviction, Reroute

Possible failures & Recovery paths

- MDS Failure
  - Detection: Timeout, no ping response
  - Response: Connect to standby MDS, replay metadata transactions
  - No errors seen by the application. Metadata operations take longer
- OST Failure
  - Detection: Communication problems
  - Response: OSC enters recovery, blocks IO to that OST, recovery
  - No errors seen by the application. IO to that OST take longer

Possibly Fatal Failures

- Client Eviction – failure to communicate in a timely manner
  - Client no longer connected to the target, locks & cache are flushed
  - Client cannot detect eviction & reconnect until the next ping or IO operation
    - In progress operations will fail with EIO or ESHUTDOWN
      - Unsubmitted changes must be discarded
  - Possible triggers (to cause timeout)
    - Network quiescence
    - Warm swaps
    - Lustre bugs

Possibly Fatal Failures

- LNET Failures
  - Can cause RPC timeouts leading to evictions
  - Applications will see the error
  - Reroute to next weight class routers
- Other cases
  - Bugs
  - Reproduce and wait for a fix

USER EXPERIENCE

User applications on Blue Waters

- Many users do not implement IO error handling
  - FORTRAN
    - Very visible – application crashes
    - No unexpected tickets later
  - C/C++
    - Silent problems – application continues
    - Users claim file corruption, which is a much bigger concern (*if it were true*)

Experiments on JYC with PSDNS

- OSS Failure
  - Application continued, IO took longer
- LNET Failure
  - Turned off three out of four available LNETS
  - Client evicted
    - "`Cannot send after transport endpoint shutdown`"
  - Repeated tests with ONE failed LNET
    - The job survived using other routers in one instance
    - Other instances, job received error

Encourage Error Handling and Defensive Programing

- FORTRAN – mandatory for resilience
  - IOSTAT – runtime error message.
  - Use in conjunction with ERR branch specifier
- C / C++ - mandatory to avoid problems later
  - Check return value from the IO call
    - User writing to a full disk, claimed file system corruption later
- Recommendation to users
  - Handle specific errors user is interested in
  - For others, retry a few times with sleep after each failed call
  - Make decisions intelligently after seeing an error.

THANK YOU