Storage Integration with Host-based Write-back Caching

Flash Memory Summit 2017

Andy Banta @andybanta

NetApp SolidFire

Agenda

- Patented information
- How virtual machines use storage
- Caching methods
  - And who can and needs to use them
- Single-host write-back caching integration
- High Availability write-back caching integration

Patented information

- Authors: Andy Banta and Erik Cota-Robles
- Wholly owned by VMware
- Discovery is fully described by the patent
  - Not giving away any secrets here

(12) **United States Patent**
Banta et al.

(10) **Patent No.: US 9,519,581 B2**
(45) **Date of Patent: Dec. 13, 2016**

(54) **STORAGE INTEGRATION FOR HOST-BASED WRITE-BACK CACHING**

(71) Applicant: **VMware, Inc.**, Palo Alto, CA (US)

(72) Inventors: **Andrew Banta**, Reno, NV (US); **Erik Cota-Robles**, Mountain View, CA (US)

(73) Assignee: **VMware, Inc.**, Palo Alto, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 162 days.

(21) Appl. No.: **14/028,101**

(22) Filed: **Sep. 16, 2013**

(65) **Prior Publication Data**

US 2015/0081979 A1 Mar. 19, 2015

(51) **Int. Cl.**
G06F 12/08 (2016.01)

(52) **U.S. Cl.**
CPC ***G06F 12/0804*** (2013.01); *G06F 12/0866* (2013.01)

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,549,959 B1 * 4/2003 Yates et al. 710/22
8,549,230 B1 * 10/2013 Chatterjee et al. 711/135
8,904,117 B1 * 12/2014 Kalekar G06F 12/0804 711/119
2002/0166031 A1 * 11/2002 Chen et al. 711/141
2014/0068197 A1 * 3/2014 Joshi G06F 12/0866 711/135
2014/0143506 A1 * 5/2014 Gole G06F 12/0842 711/143

OTHER PUBLICATIONS

Fusion-IO, "Fusion-io: A New Standard for Enterprise-class Reliability," 2011, available at http://www.fusionio.com/hitepapers/fusion-io-a-new-standard-for-enterprise-classreliability/.
IBM, "Disk Cache Write-Back versus Write-Through" in "Tuning IBM System x Servers for Performance" (section 11.6.7 p. 281), Aug. 2009, available at http://www.redbooks.ibm.com/redbooks/pdfs/sg245287.pdf.
Pure Storage, "FlashArray, The All-Flash Array for Every Enterprise", available at http://www.purestorage.com/flasharray (viewed Sep. 16, 2013).
The Register, "Dell spills its hot cache Fluid, hopes to beat off rivals", Jun. 18, 2012, available at http://www.theregister.co.uk/2012/06/18/dell_hermes/.
Ricardo Koller, Leonardo Marmol, Raju Rangaswami, Swaminathan Sundararaman,Nisha Talagala, Ming Zhao "Write Policies for Host-side Flash Caches," Feb. 2013, http://users.cis.fiu.edu/~raju/WWW/publications/fast2013/paper.pdf.
Violin Memory, Inc., Storage at the Speed of Memory, available at http://www.violin-memory.com/products/ (viewed Sep. 16, 2013).

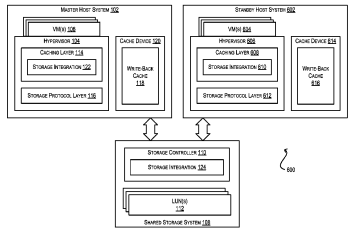
* cited by examiner

Primary Examiner — Midys Rojas
Assistant Examiner — Jane Wei

(57) **ABSTRACT**

Techniques for enabling integration between a storage system and a host system that performs write-back caching are provided. In one embodiment, the host system can transmit to the storage system a command indicating that the host system intends to cache, in a write-back cache, writes directed to a range of logical block addresses (LBAs). The host system can further receive from the storage system a response indicating whether the command is accepted or rejected. If the command is accepted, the host system can initiate the caching of writes in the write-back cache.

18 Claims, 7 Drawing Sheets

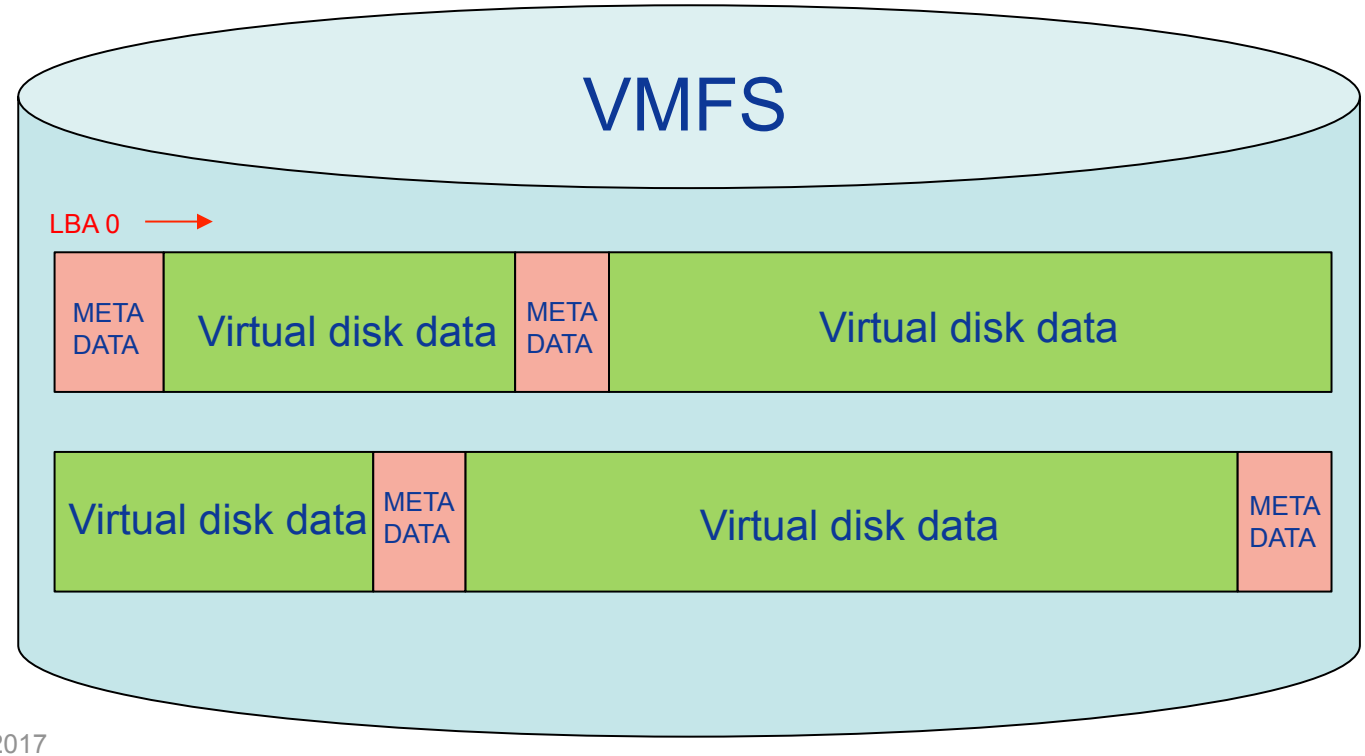


How VMware uses storage

- Shared filesystem on block storage
  - Others exist, not as interesting
- One or more LUs make up one filesystem
- Virtual disks intermixed with metadata
- Various locks to prevent or coordinate concurrent access to VMs and metadata.


Flash Memory Summit

VMFS layout


VMFS
LBA 0
META DATA
Virtual disk data
META DATA
Virtual disk data
Virtual disk data
META DATA
Virtual disk data
META DATA

Caching with VMFS

- Hosts can’t cache metadata
  - Heartbeat operations assure storage health
  - Used to coordinate access among multiple hosts
  - Information about usage and layout
- Hosts can cache virtual disk data
  - Assuming single host access
  - Assuming no underlying storage operations

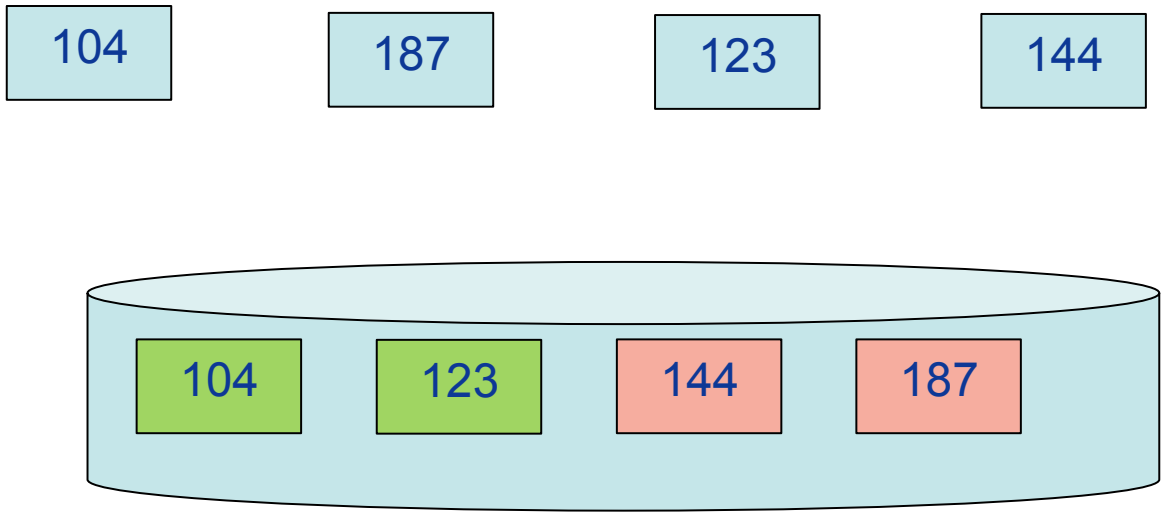
Write-back vs. Write-through

- Write-through
  - Write operations complete to backing storage
    - Therefore no write speed up
  - Provides transactional consistency
    - Required in many cases
- Write-back
  - Speeds up write operations
  - Might cause data corruption or loss

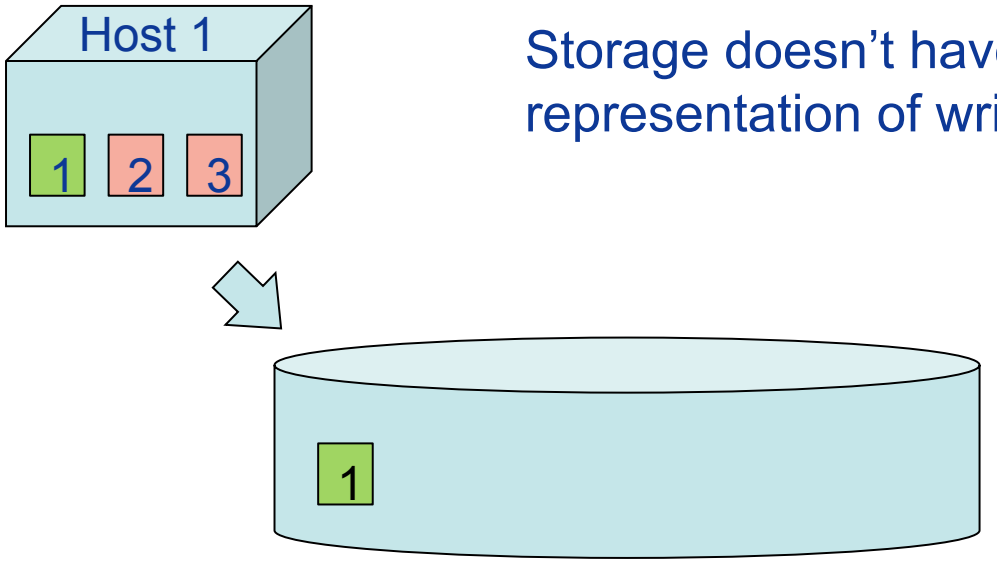
Write-back options

- Write ordering is important
  - Flushing should be in write order
  - Out-of-order flush can cause corruption
- Not for all workloads
  - Desktops, web servers are OK
  - Databases not so much
- Can be used for crash consistency

Write ordering corruption

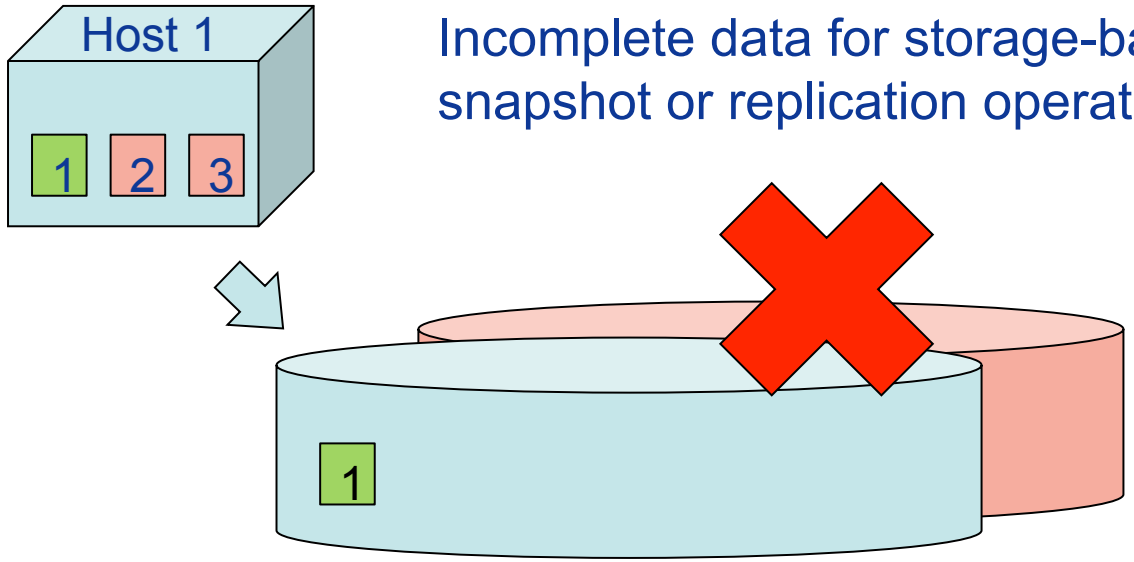


Write-back pitfalls



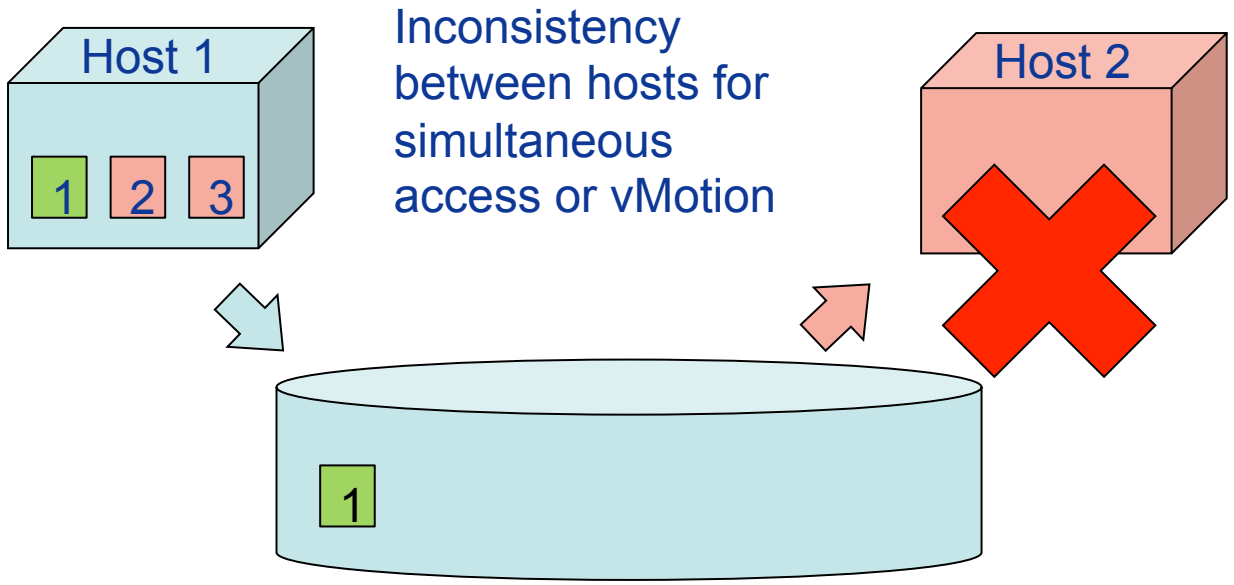
Storage doesn't have an up-to-date representation of written data

Write-back pitfalls

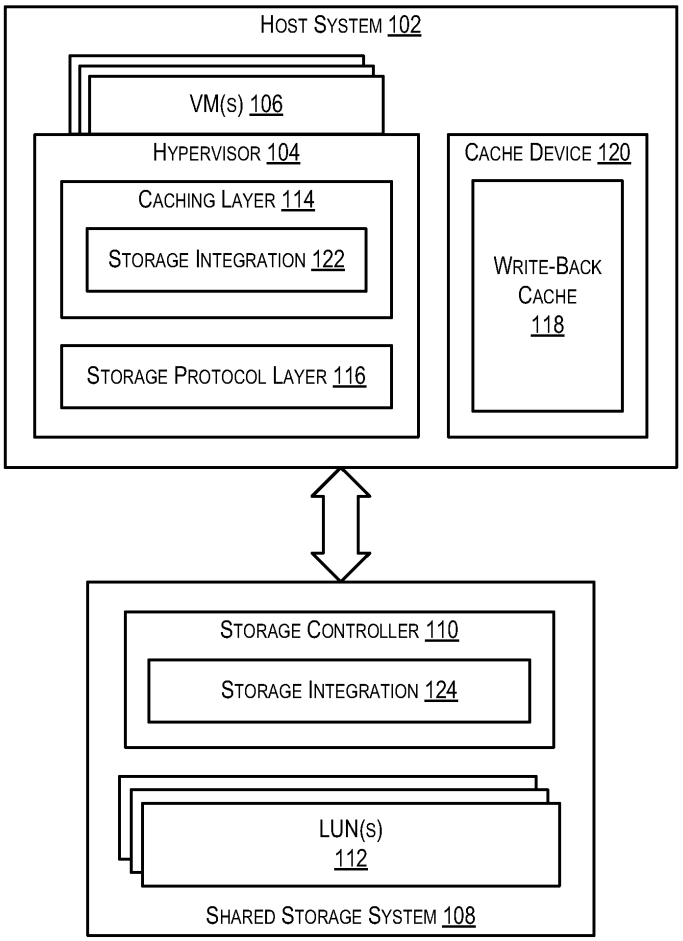


Incomplete data for storage-based snapshot or replication operations

Write-back pitfalls



Storage Integration with host-based write-back caching



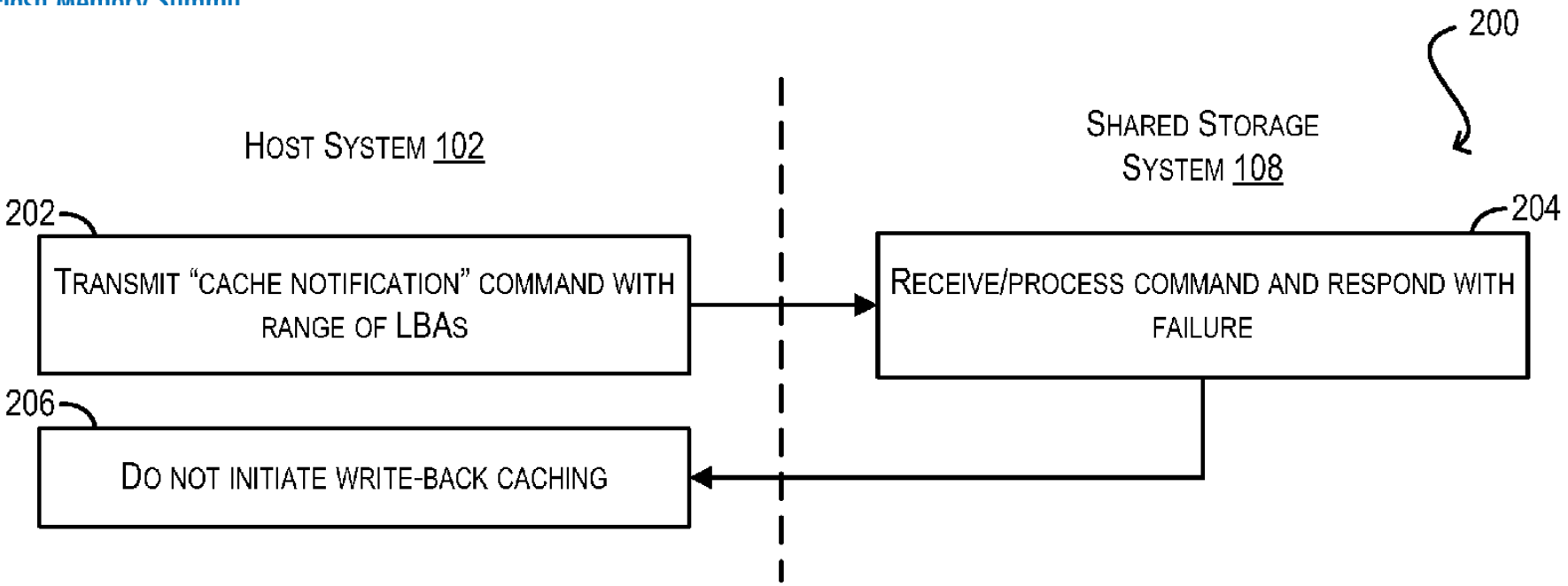
New commands and responses

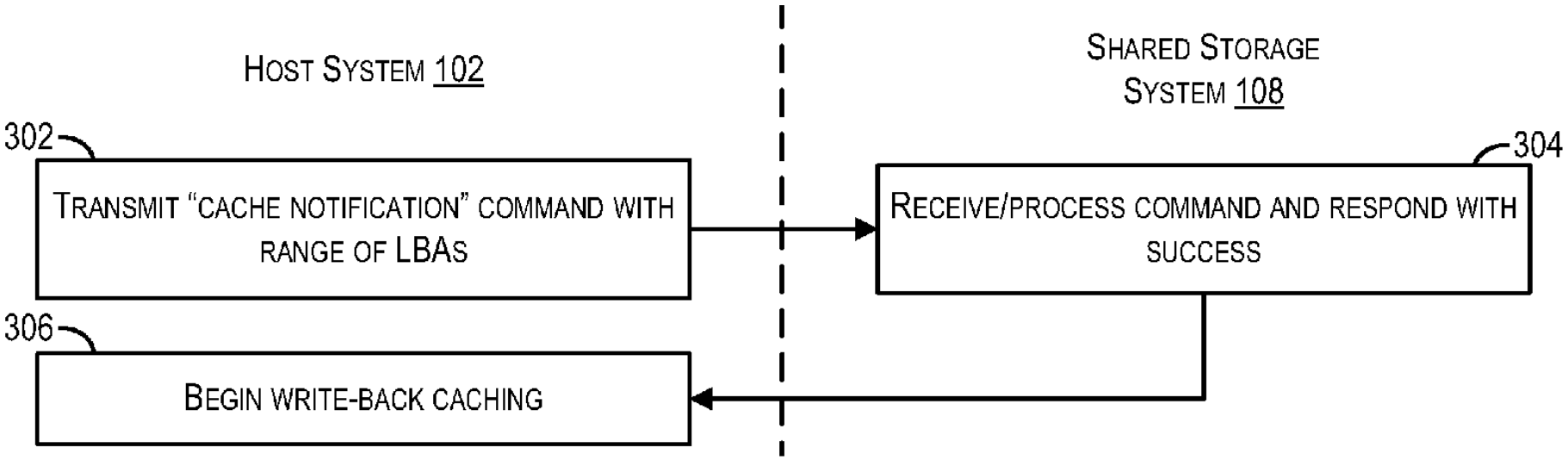
- Simple expansion of SCSI T10 standard
- `Cache Notification` Command
  - Capability discoverable at Inquiry
  - Host informs storage of intent to do write-back caching
  - Provides a range of blocks (LBAs)
- `Cache Flushed` Command
  - Flag to indicate if caching is complete

New commands and responses

- Simple expansion of SCSI T10 standard
- `Flush Required` Unit Attention
  - Recoverable, Not ignorable
  - Synchronized with `Flush Completed` command
  - Flag for `one time` or `permanent`

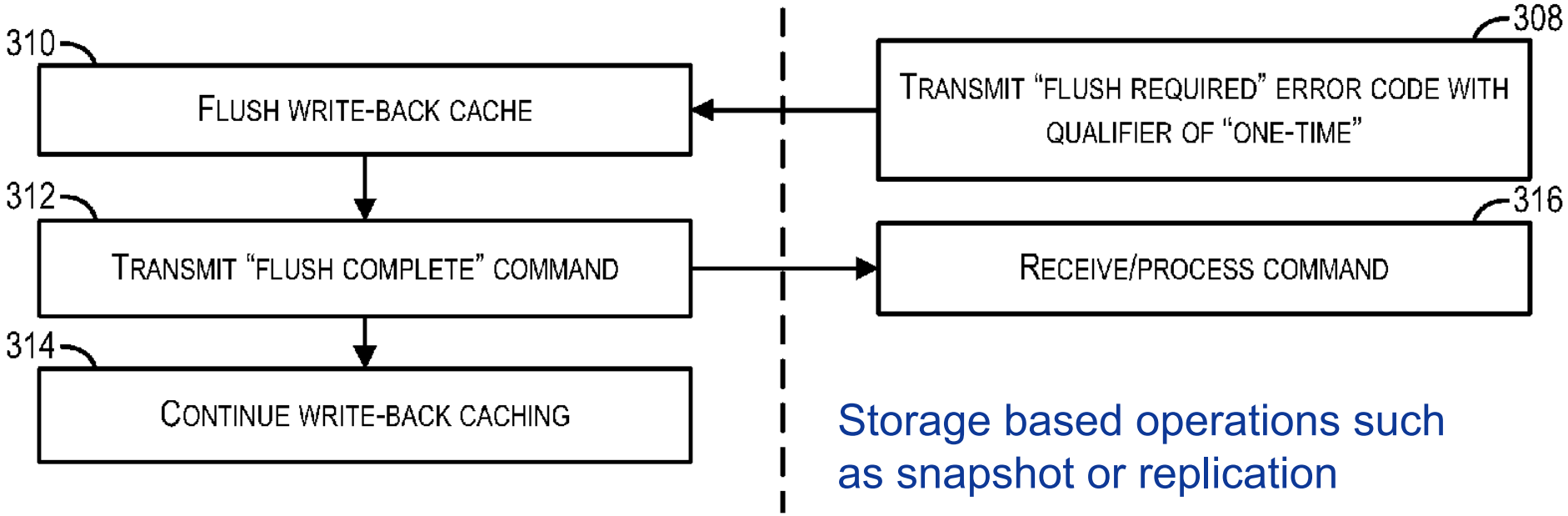
No caching scenario





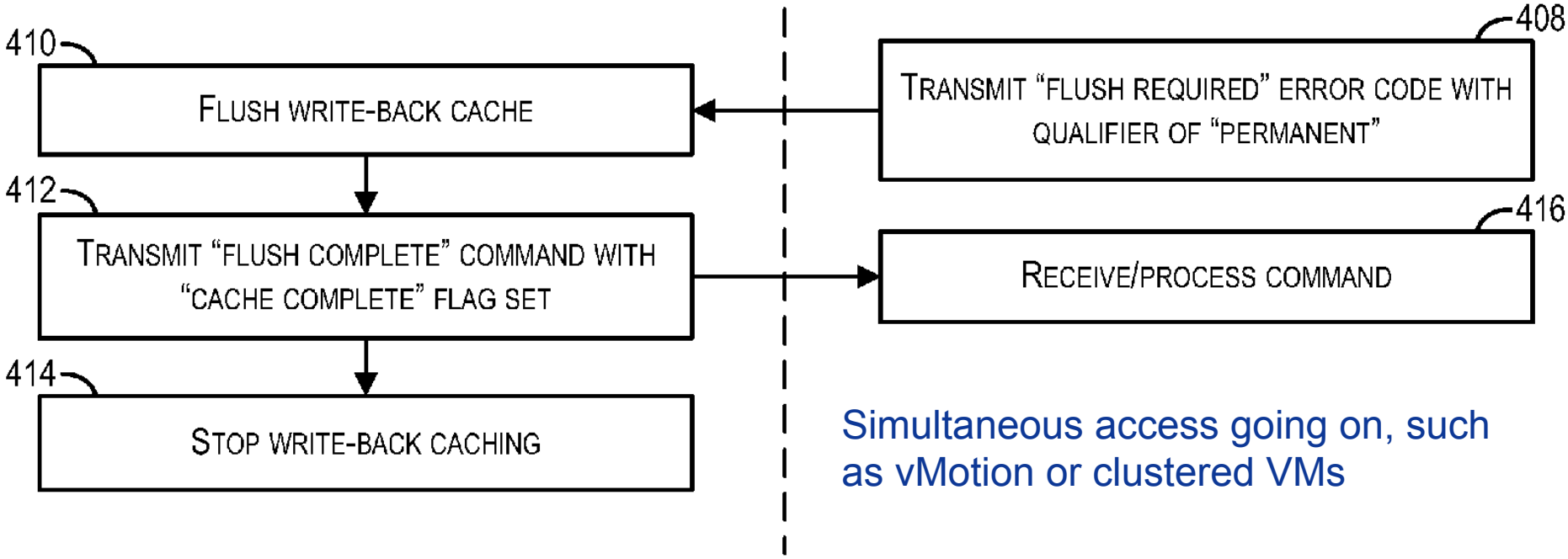
Start caching example

One time flush

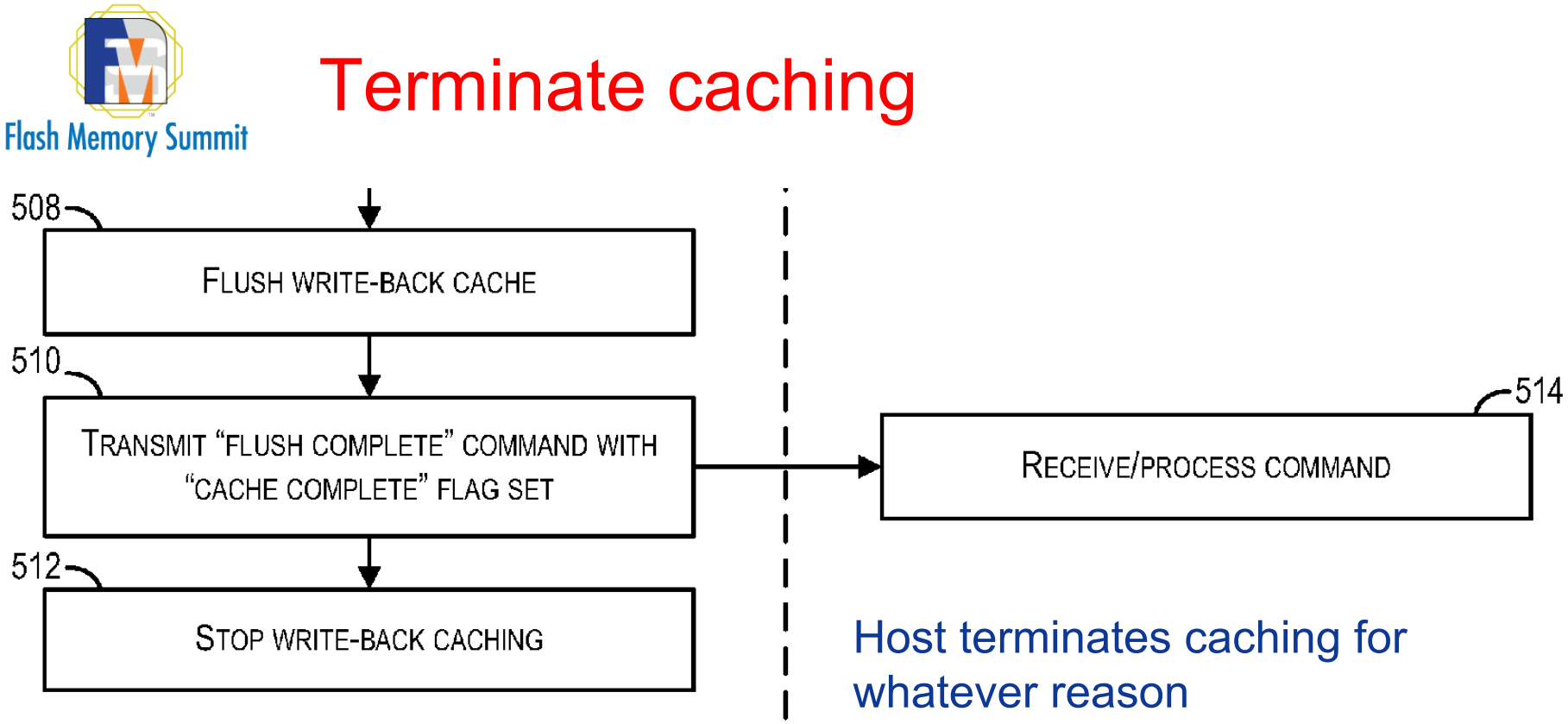


Storage based operations such as snapshot or replication

Disable caching



Simultaneous access going on, such as vMotion or clustered VMs



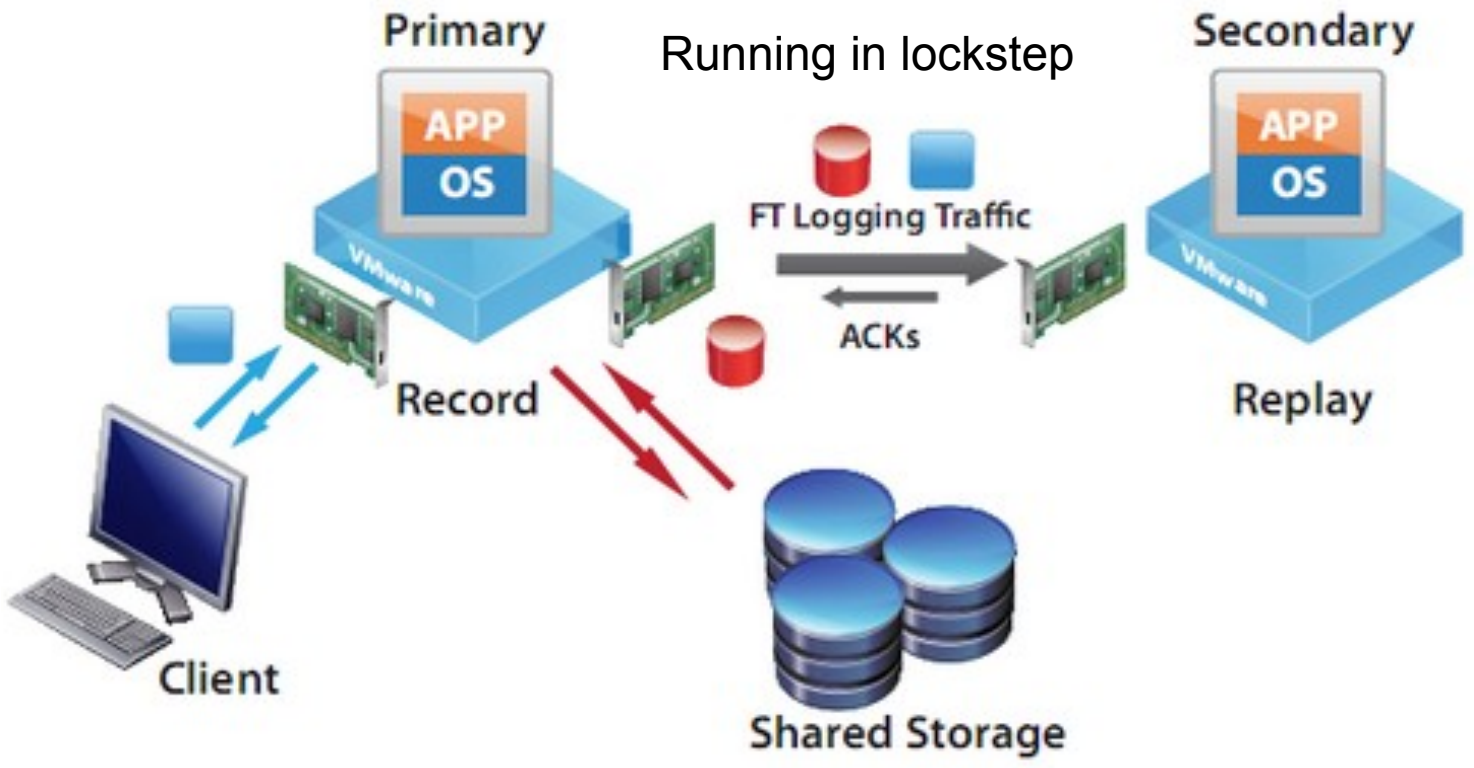
Terminate caching

Host terminates caching for whatever reason

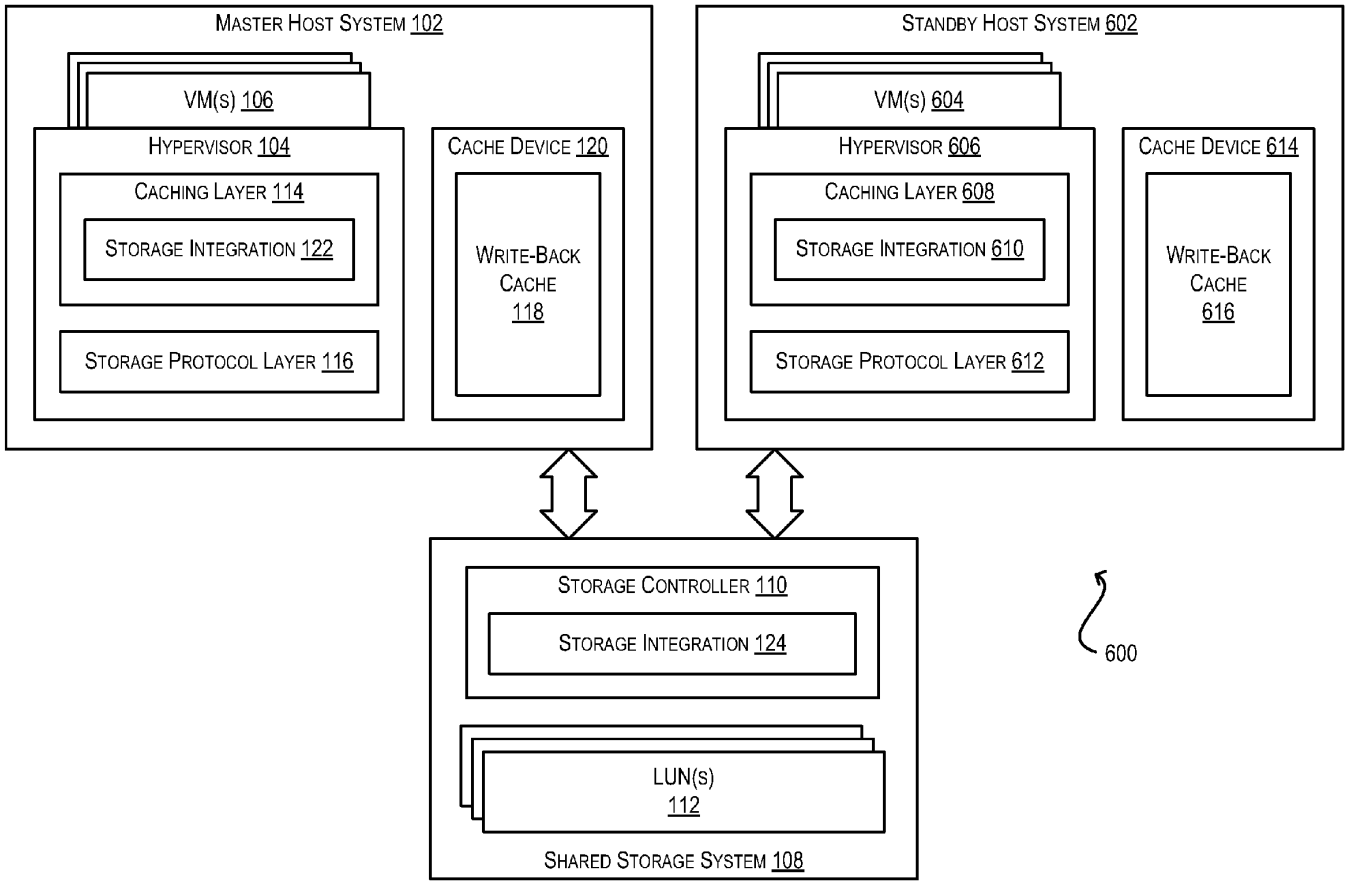
Cache coordination

Action	Unit Attention	Host Command
Storage operation	Flush Required – one time	Flush complete - No Cache complete
Secondary host operation	Flush Required - permanent	Flush complete – Cache complete
Primary host operation	N/A	Flush complete – Cache complete

VMware Fault Tolerance



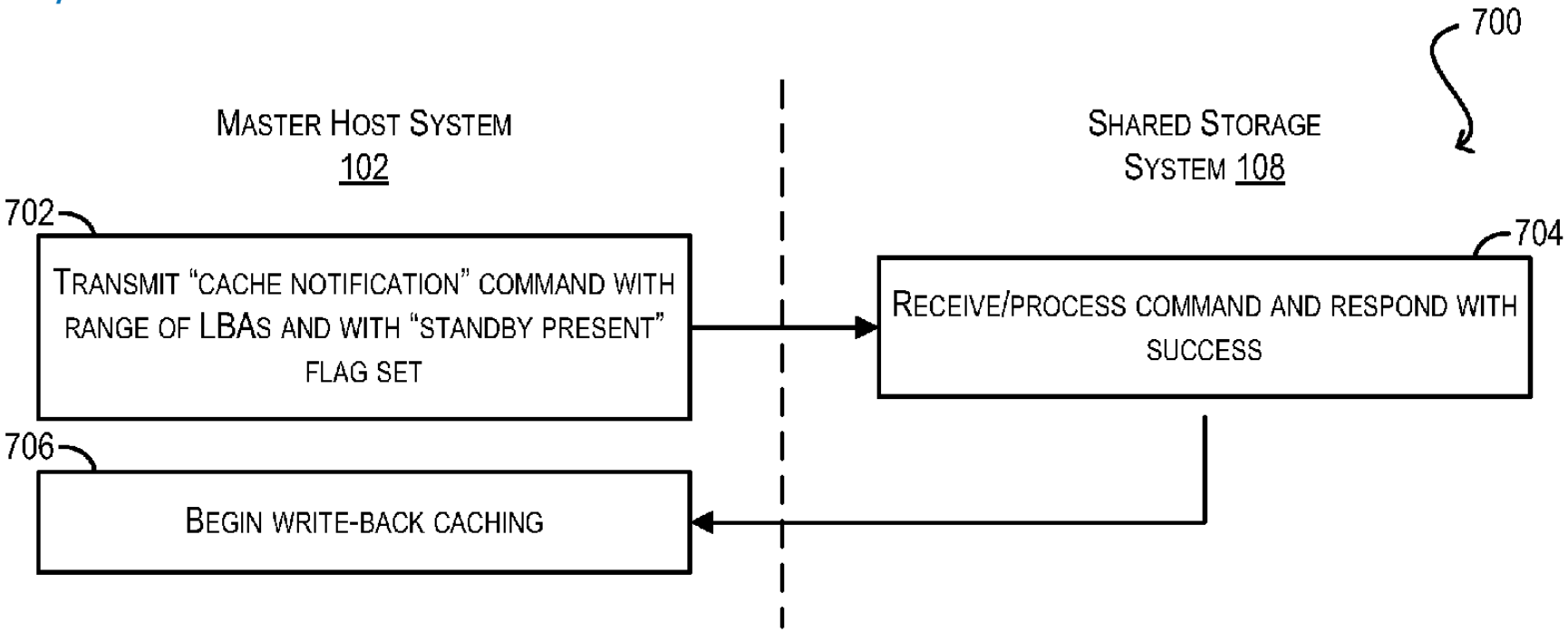
Fault Tolerance with write-back caching



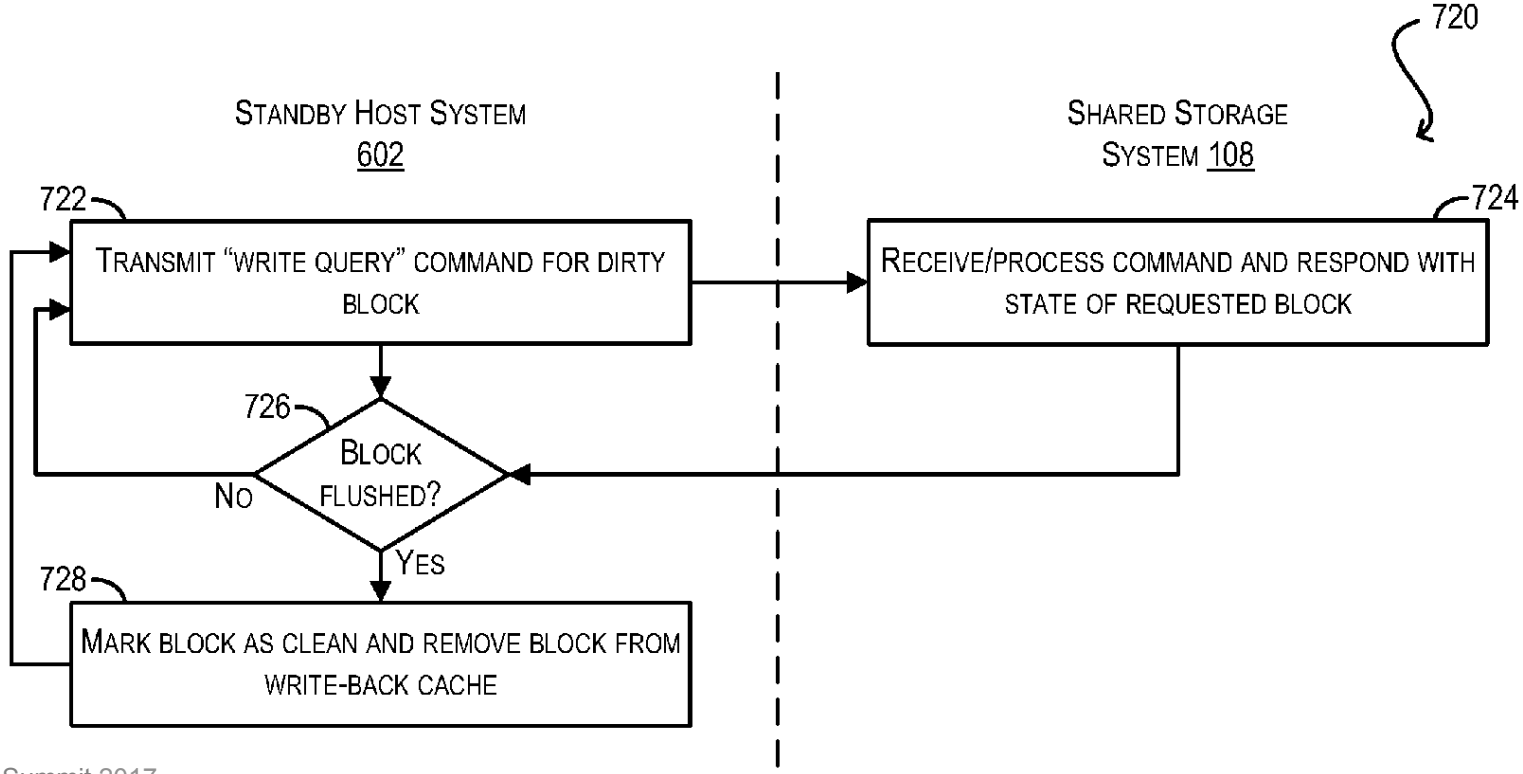
New commands and responses

- Continued expansion of SCSI T10 standard
- `Cache Notification` Command
  - `Standby Present` indicator
  - `Assuming Control` indicator
- `Write Query` Command
  - Checks if particular LBAs have been written
  - Used to prune the standby write-back cache

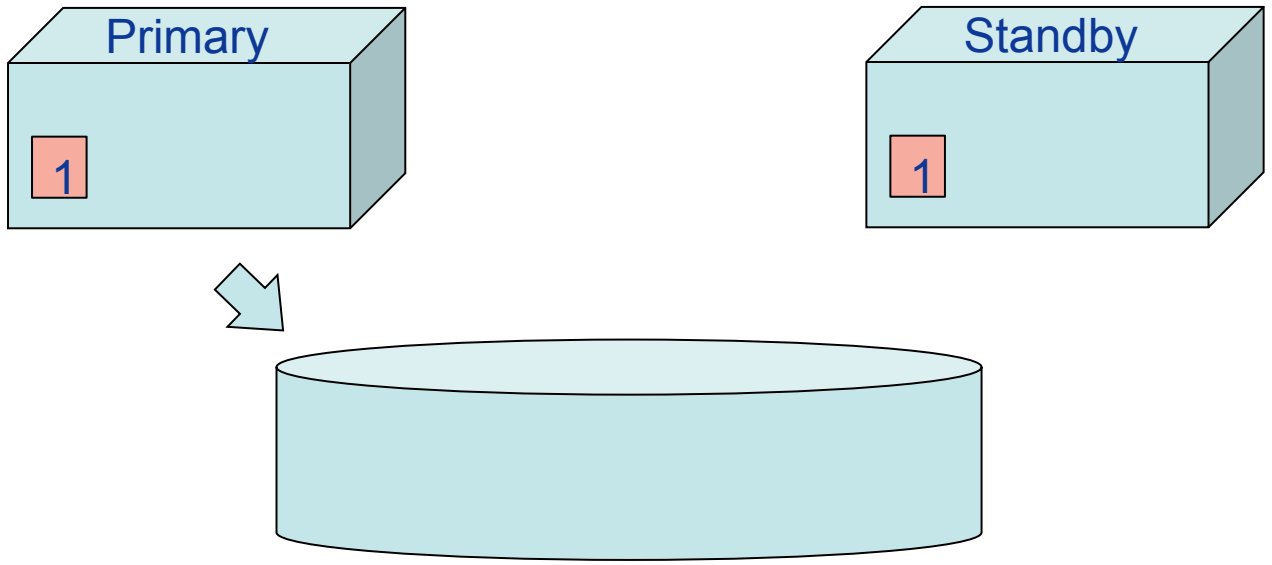
Initiate write-back caching with standby



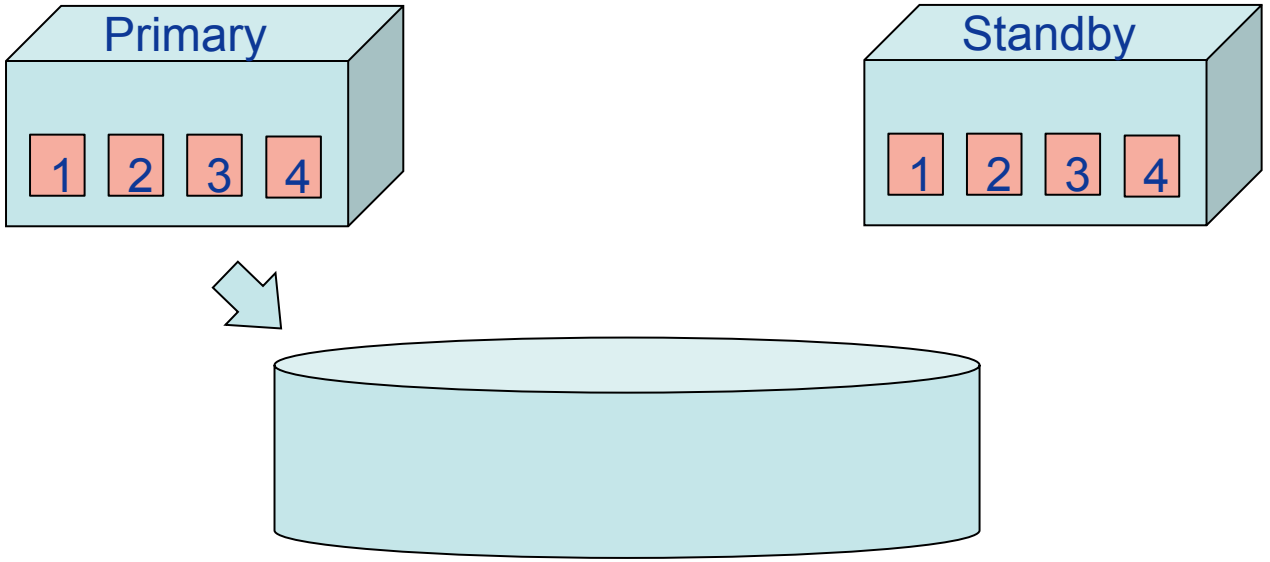
Standby checking cache status



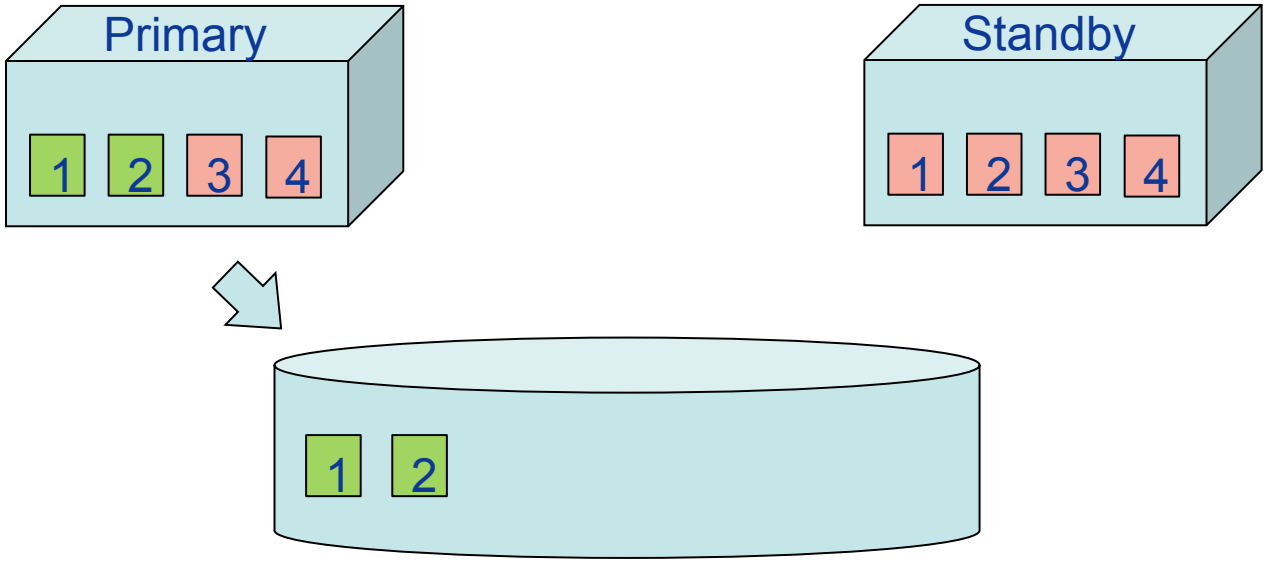
Fault tolerance example



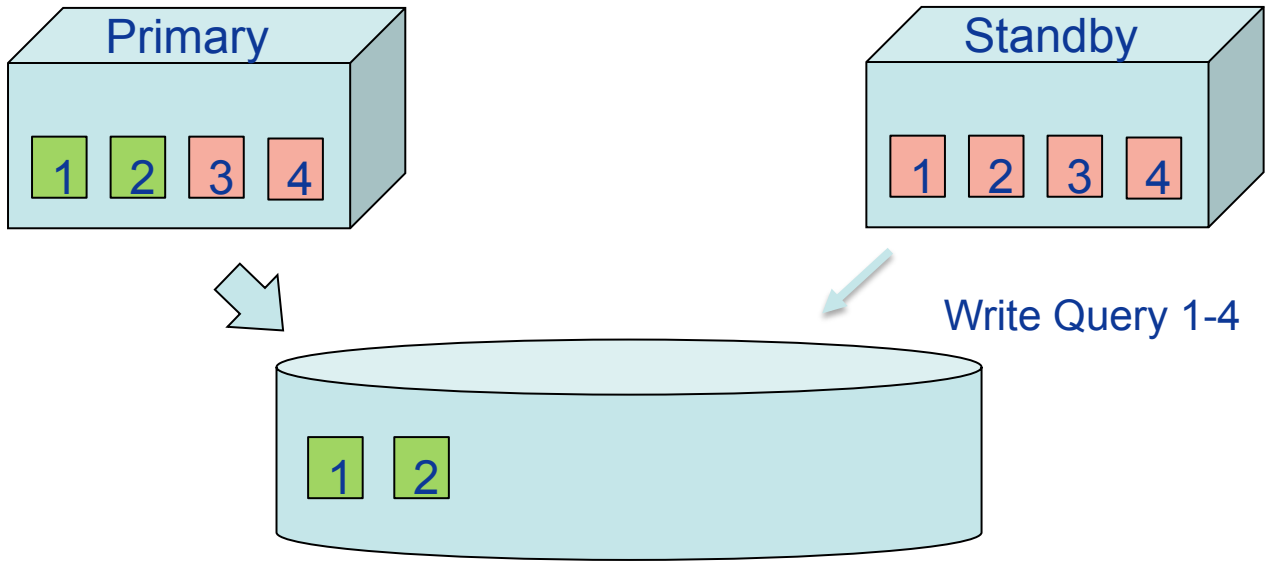
Fault tolerance example



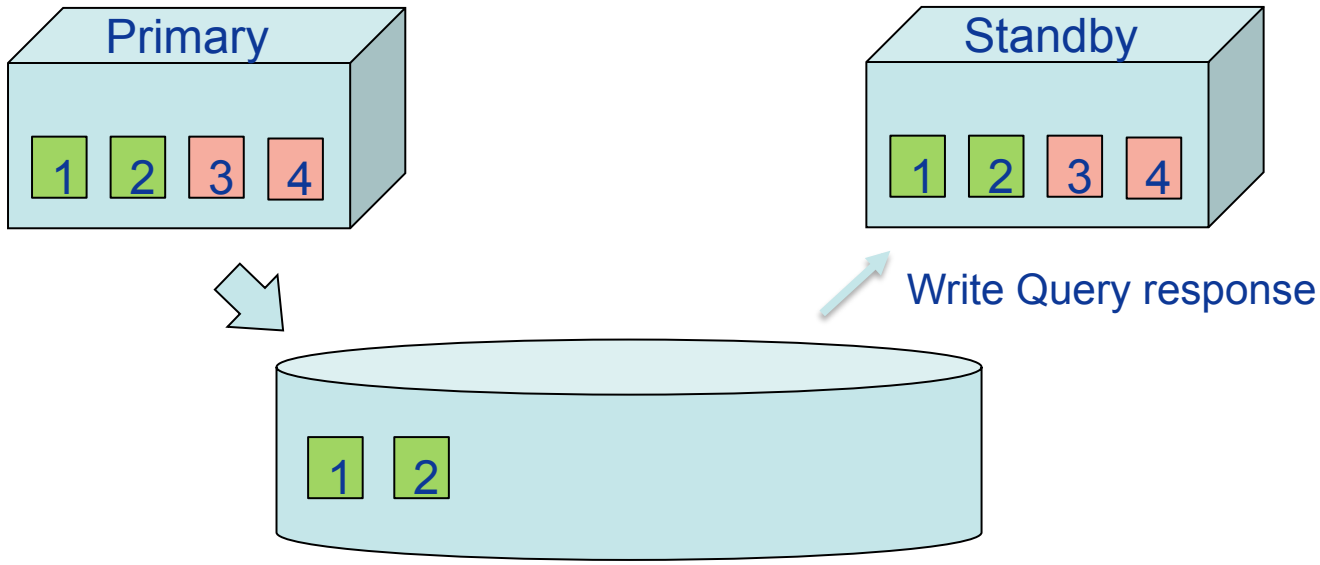
Fault tolerance example



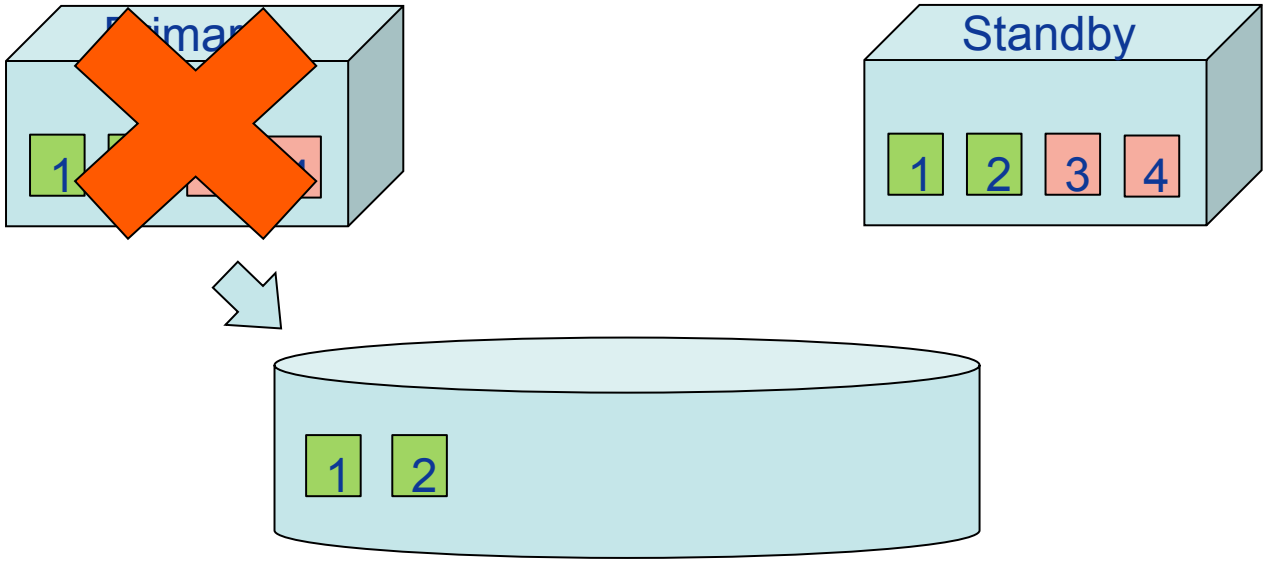
Fault tolerance example



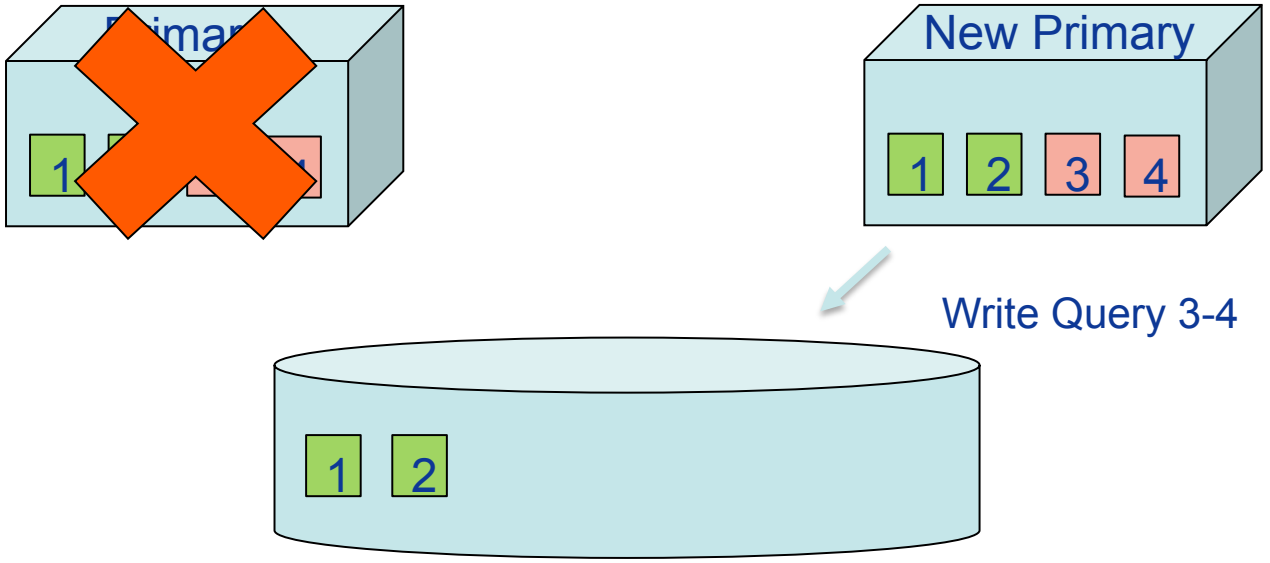
Fault tolerance example



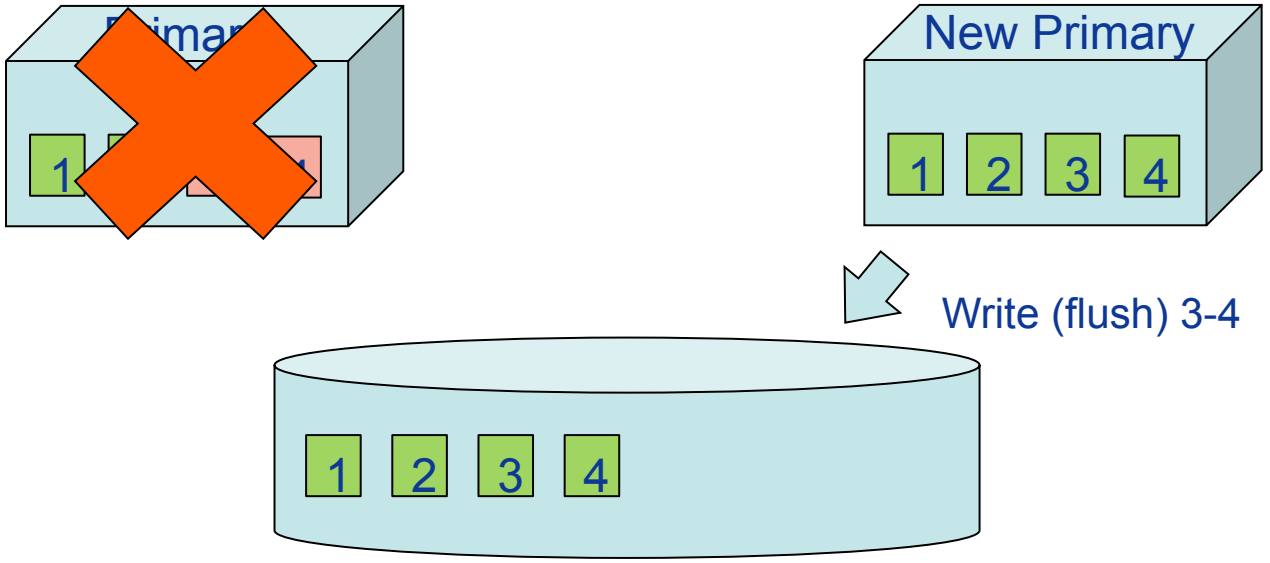
Fault tolerance example



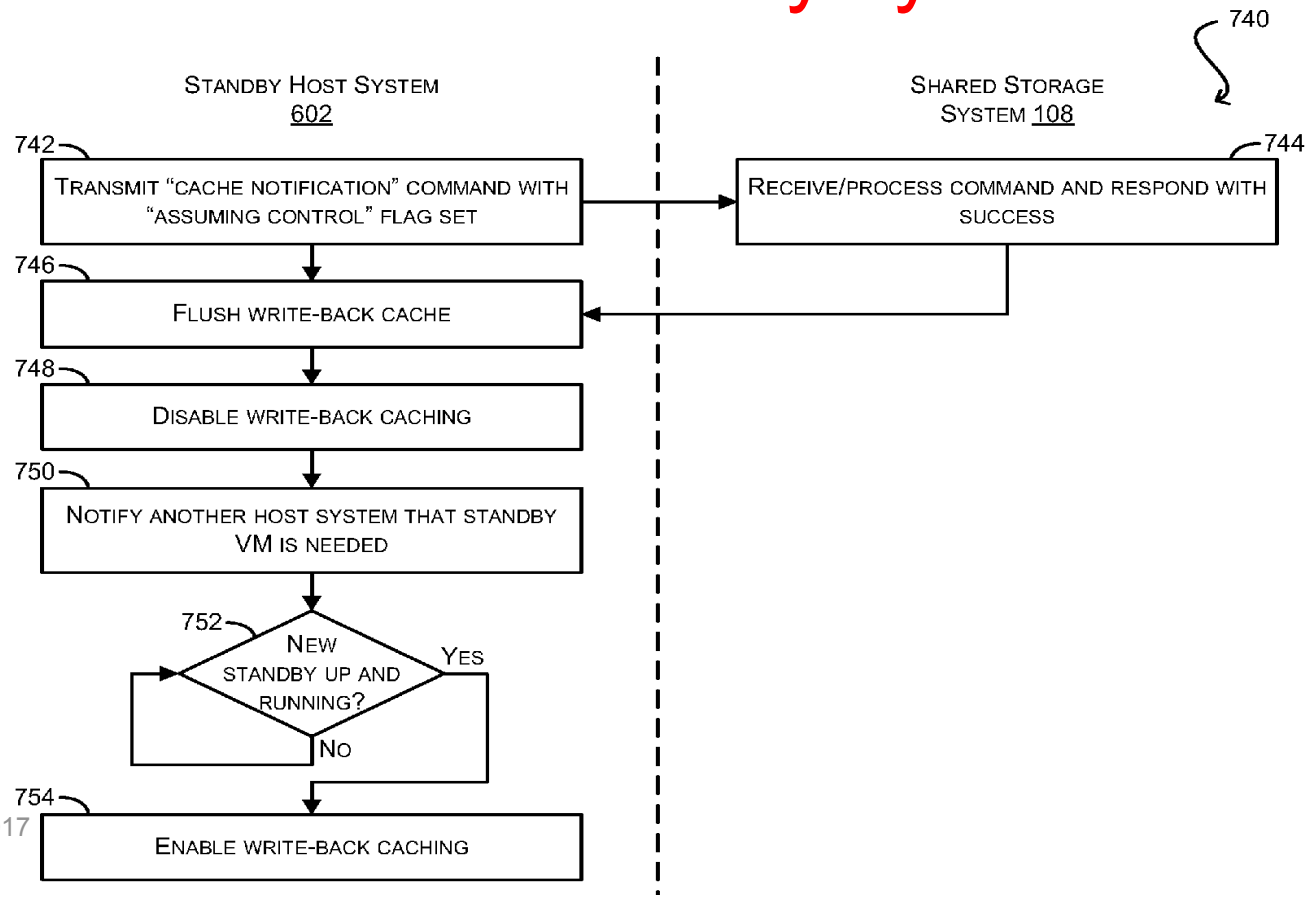
Fault tolerance example



Fault tolerance example



Failover to secondary system



Thank You

- Questions?

Free-format slide title