ICS-2018: The 32nd ACM/SIGARCH International Conference on Supercomputing

PFault: A General Framework for Analyzing the Reliability of High-Performance Parallel File Systems

Jinrui Cao, Om Rameshwar Gatla, Mai Zheng

New Mexico State University, USA

Dong Dai, Vidya Eswarappa, Yan Mu, Yong Chen

Texas Tech University, USA

Parallel File Systems (PFSes) are Important

- A crucial component of any HPC Systems
  - enable file management & global namespace across nodes

lustre® File System

PVFS

OrangeFS™

ceph

Increasing Scale & Complexity Causes Reliability Challenge

- More and more difficult to get right
  - many traditional verification methods (e.g., model checking) are not scalable

- Failure handling is particularly under-studied
  - Failure events (“faults”) can happen at any time after deployment, but may never happen during development

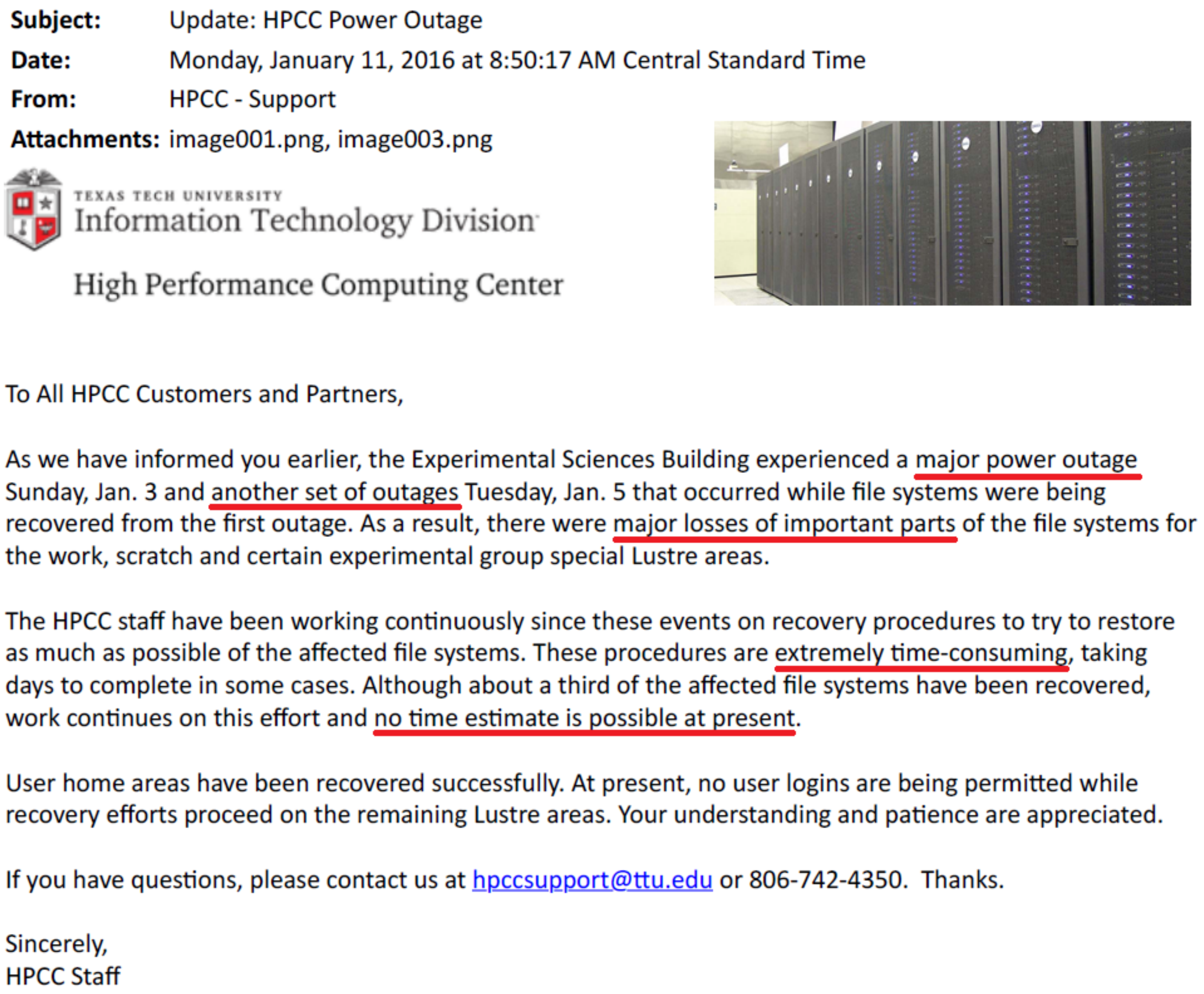
- Has been exposed in other large-scale systems

“Why Does the Cloud Stop Computing?: Lessons from Hundreds of Service Outages” [SoCC’16]

“Failure recovery: When the cure is worse than the disease” [HotOS’13]

THE UNIVERSITY OF CHICAGO

Microsoft



Subject: Update: HPCC Power Outage
Date: Monday, January 11, 2016 at 8:50:17 AM Central Standard Time
From: HPCC - Support
Attachments: image001.png, image003.png

TEXAS TECH UNIVERSITY
Information Technology Division

High Performance Computing Center

To All HPCC Customers and Partners,

As we have informed you earlier, the Experimental Sciences Building experienced a major power outage Sunday, Jan. 3 and another set of outages Tuesday, Jan. 5 that occurred while file systems were being recovered from the first outage. As a result, there were major losses of important parts of the file systems for the work, scratch and certain experimental group special Lustre areas.

The HPCC staff have been working continuously since these events on recovery procedures to try to restore as much as possible of the affected file systems. These procedures are extremely time-consuming, taking days to complete in some cases. Although about a third of the affected file systems have been recovered, work continues on this effort and no time estimate is possible at present.

User home areas have been recovered successfully. At present, no user logins are being permitted while recovery efforts proceed on the remaining Lustre areas. Your understanding and patience are appreciated.

If you have questions, please contact us at hpccsupport@ttu.edu or 806-742-4350. Thanks.

Sincerely,
HPCC Staff

Subject: Update: HPCC Power Outage

Date: Monday, January 11, 2016 at 8:50:17 AM Central Standard Time

From: [illegible]

Attachments: image001.png, image003.png

Information Technology Division

High Performance Computing Center

To All HPCC Customers and Partners,

As we have informed you earlier, the Experimental Sciences Building experienced a major power outage [illegible] being [illegible] from the first outage. As a result, there were major losses of important parts of the file systems for [illegible] to try to restore as much as possible of the affected file systems. These procedures are extremely time-consuming, taking days to [illegible] in some cases. Although about a third of the affected file systems have been recovered, work con[illegible] on this effort and no time estimate is possible at present. [illegible] recovery efforts proceed on the remaining Lustre areas. Your understanding and patience are appreciated.

If you have questions, please contact us at hpccsupport@ttu.edu or 806-742-4350. Thanks.

Sincerely,
HPCC Staff

2018:"...**power outage** one reason behind **AWS cloud disruption**"

2018:"...**lost power** in the largest consumer electronics show"

2017:"... **massive AWS outage** ... caused by human error"

2017:"Red Hat Suffers **Massive** Data Center Network **Outage**"

2017:"**data center outage** ... cancellation of over 400 flights"

2016:"Verizon **data center failure** ... air travel **delays**"

2016:"**Data Center Power Outage** Brings **Down** GitHub"

2016:"Delta: **Data Center Outage** Cost Us **$150M**"

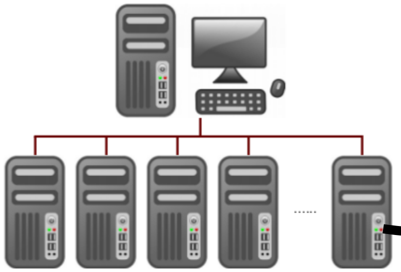
2015:"**Data center outage disrupts** Fujitsu cloud"

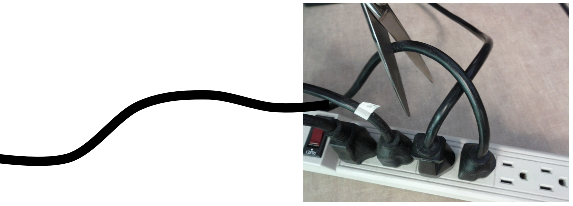
2015:"Lightning strikes and old disks cause Google **Data Loss**"

2014:"... **Data Center Outage** Takes **Down** StackExchange..."

Our Contributions

- PFault
  - A general fault-injection framework for analyzing the failure handling of PFSes
    - transparent to PFSes
    - easy to deploy in practice





- Case Study on Lustre
  - Uncover a number of unexpected recovery issues, including a resource leak problem
  - Build a tool (LeakCK) to mitigate the resource leak problem

Outline

I. Motivation

II. Background of PFS

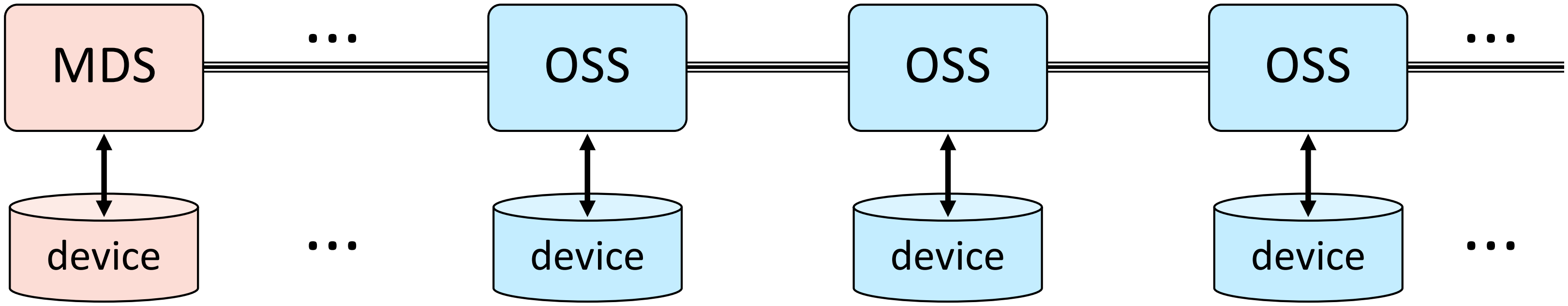
III. Design of PFault

IV. Case Study: Lustre

V. Conclusion & Future Work

PFS 101

- File management & global namespace across nodes in a cluster
  - crucial for any HPC and Big Data systems
  - two types of nodes
    - Metadata Server/**MDS**: metadata of PFS
    - Object Storage Server/**OSS**: user data



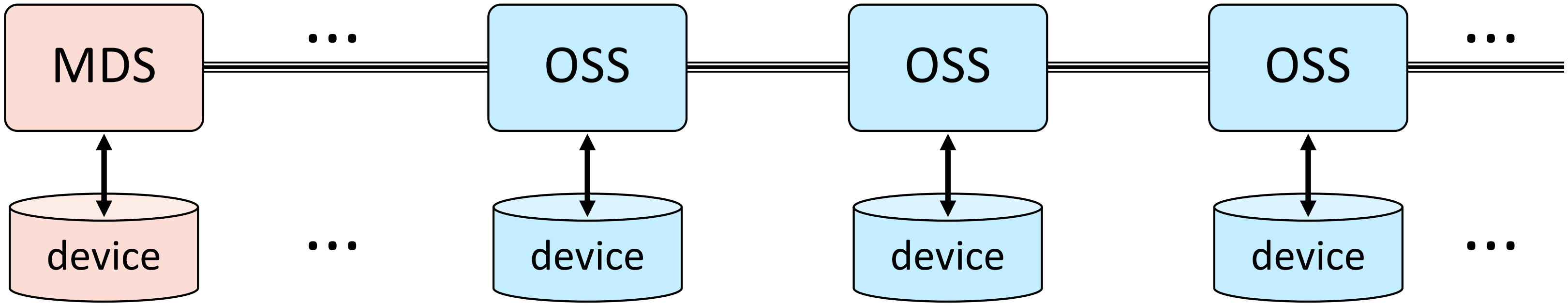
lustre File System   PVFS   OrangeFS   ceph

PFS 101

- Include sophisticated redundancy to handle failure events
  - e.g. stand-by servers
  - but may not work as expected

"Redundancy Does Not Imply Fault Tolerance: Analysis of Distributed Storage Reactions to Single Errors and Corruptions" [FAST'17]

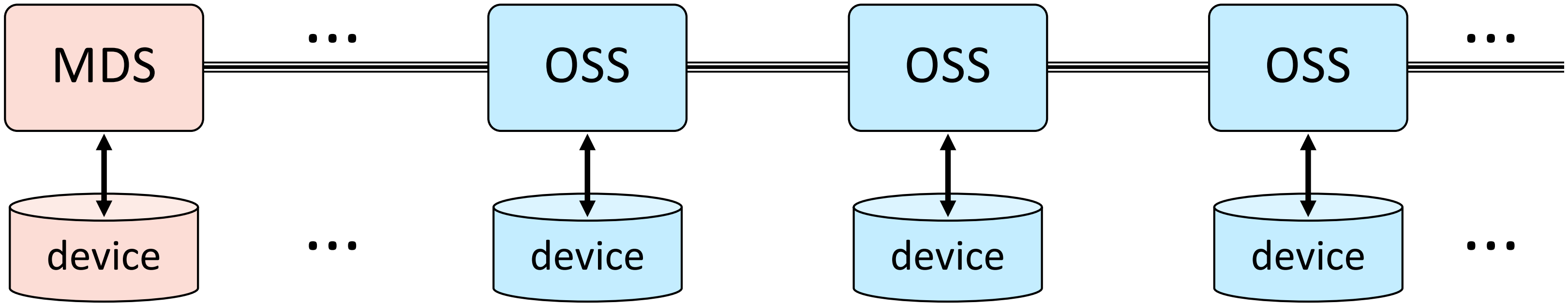
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON



lustre File System PVFS OrangeFS ceph

PFS 101

- PFS checker ("global checker")
  - "the last line of defense" to recover a broken PFS
    - e.g., LFSCK (Lustre), cephfs-fsck (Ceph)
  - detect /repair global inconsistencies/corruptions across nodes
  - depend on local file system checker ("local checker")
    - e.g., e2fsck (Ext4)



lustre File System   PVFS   OrangeFS   ceph

Outline

I. Motivation

II. Background of PFS

III. Design of PFault

IV. Case Study: Lustre

V. Conclusion & Future Work

What can happen to PFS in practice?

- "A Behind-the-Scenes Tour", Jeff Dean@Google
  - typical first year for a new cluster:
    - ~20 rack failures (40-80 machines instantly disappear)
    - ~1 PDU failure (~500-1000 machines suddenly disappear)
    - ~0.5 overheating (power down most machines in <5 mins)
    - ~5 racks go wonky (40-80 machines see 50% packet loss)
    - ~8 network maintenances (4 might cause connectivity losses)
    - ~1000 individual machine failures
    - ~thousands of hard drive failures
    - ~slow disks, bad memory, etc.

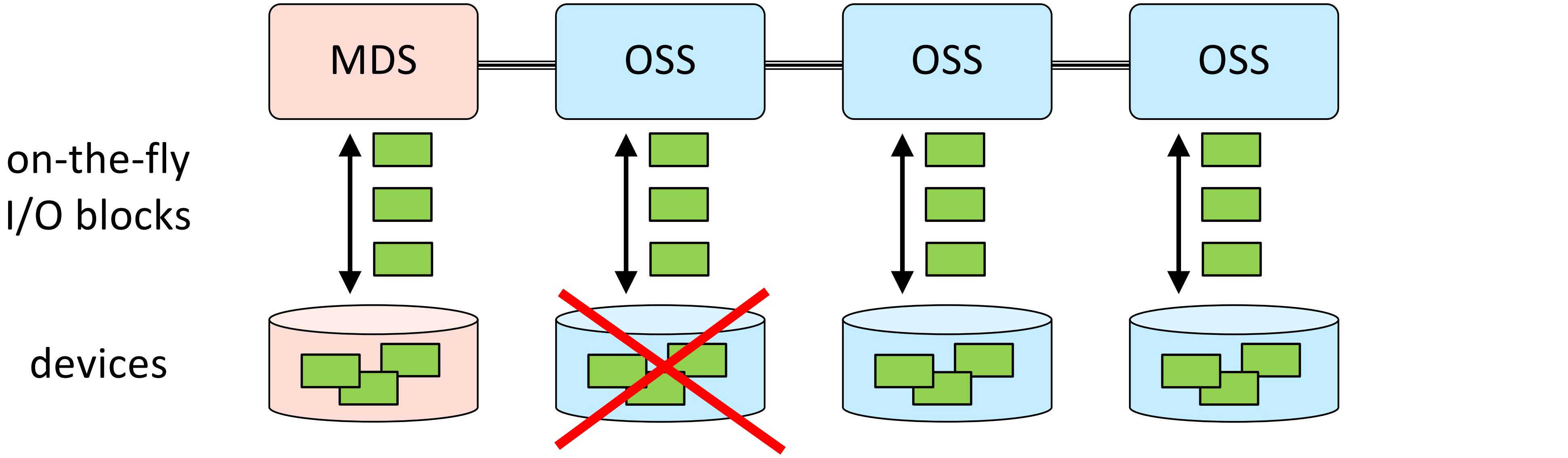
- And many others
  - Gunawi*[FAST'18]*, Cano*[SoCC'16]*, Xia*[FAST'15]*, Huang*[SoCC'15]*, Zheng*[FAST'13]*, Dinu*[HPDC'12]*, Clement*[NSDI'09]*, Bairavasundaram*[FAST'08]*, Schroeder*[FAST'07]*, Yang*[OSDI'06]*, ...

THE UNIVERSITY OF CHICAGO McGill Georgia Tech TEXAS The University of Texas at Austin RICE Unconventional Wisdom

THE OHIO STATE UNIVERSITY WISCONSIN UNIVERSITY OF WISCONSIN–MADISON UNIVERSITY OF TORONTO Stanford University UNIVERSITY of WASHINGTON

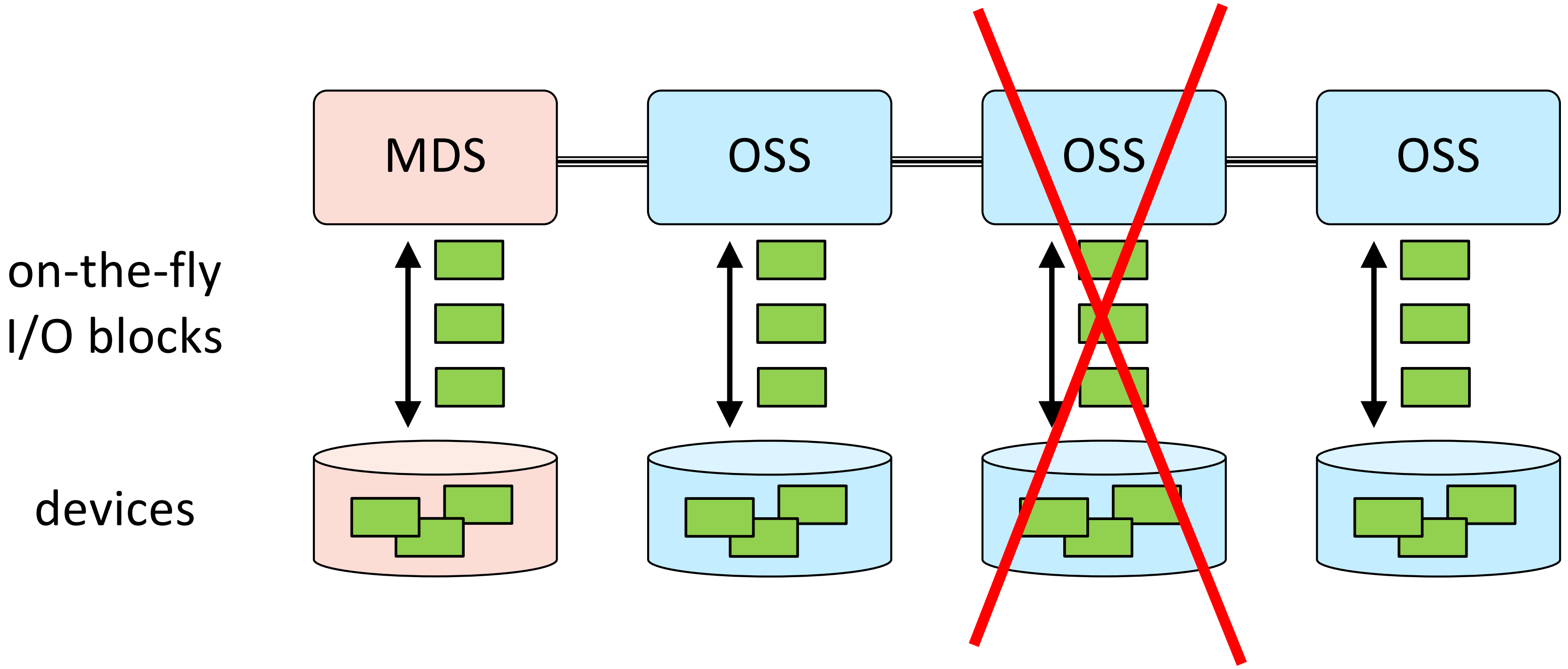
Three Representative Fault Models

- #1: Whole Device Failure
  - lose connection to one or more devices entirely
    - may be caused by controller failure, accumulation of sector errors, etc.
    - may happen on any subset of nodes



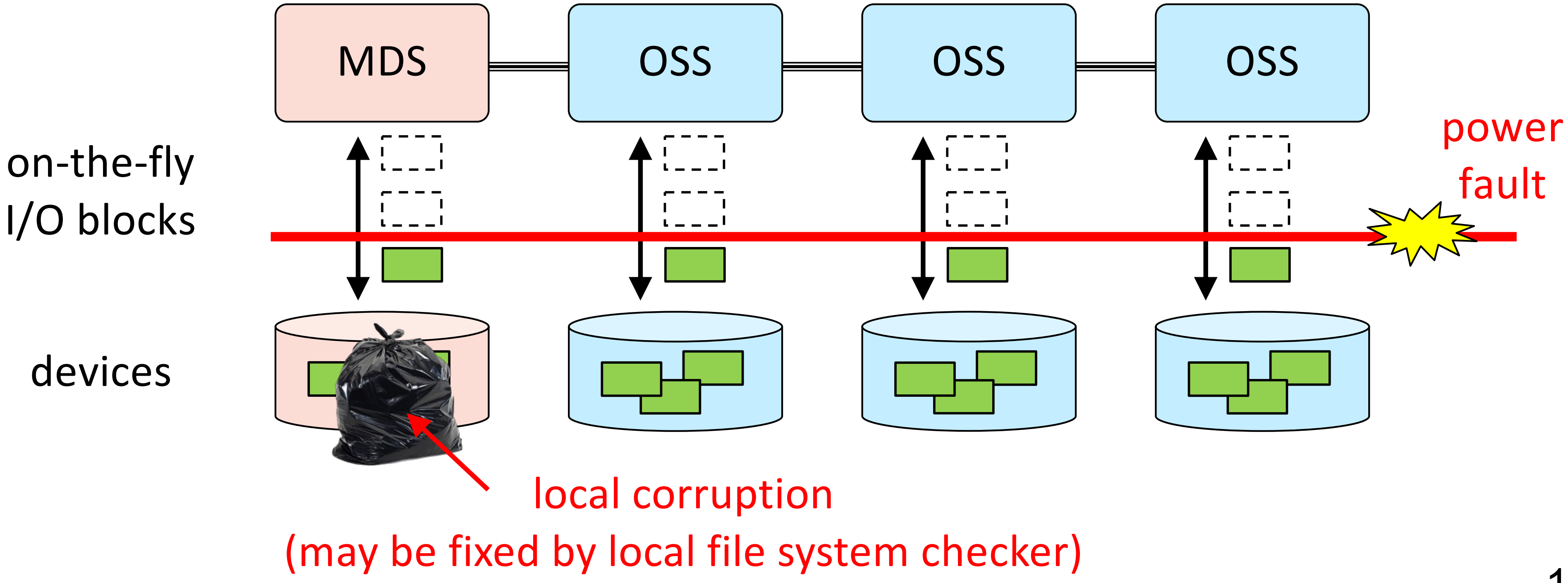
Three Representative Fault Models

- #2: Network Partitioning
  - lose connection to one or more nodes entirely
    - may be caused by malfunction of network interface cards and switches, etc.
    - may happen on any subset of nodes

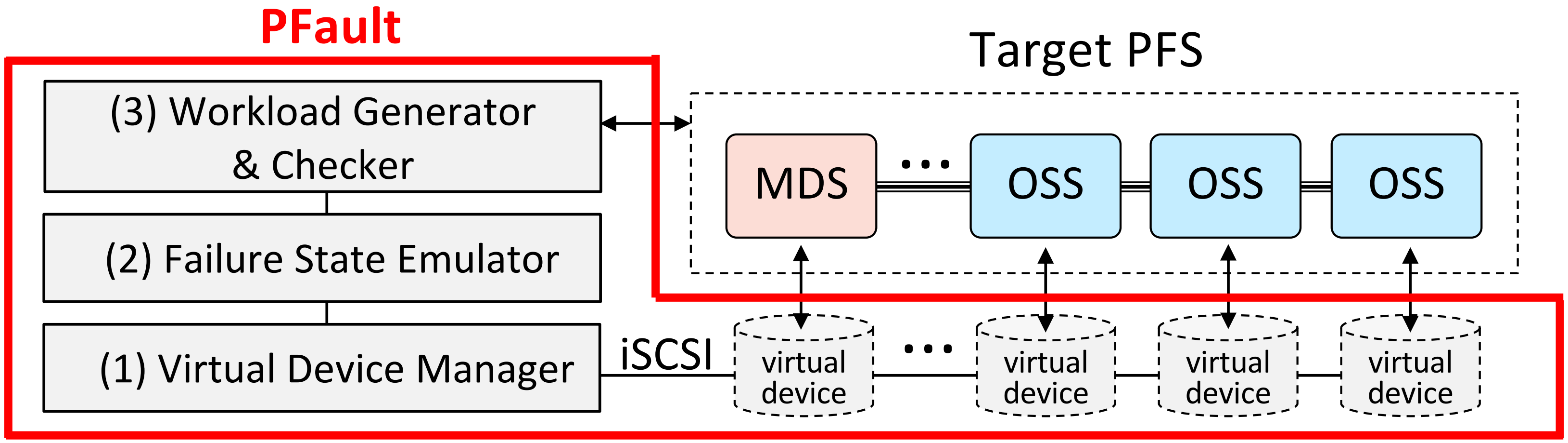


Three Representative Fault Models

- #3: Global Inconsistency
  - all devices and nodes are still accessible
  - all local file systems on individual nodes are consistent (locally)
    - e.g., local file systems are corrupted (due to power outages, latent sector errors, etc.), and then repaired by the local checker
  - the global state across nodes is inconsistent

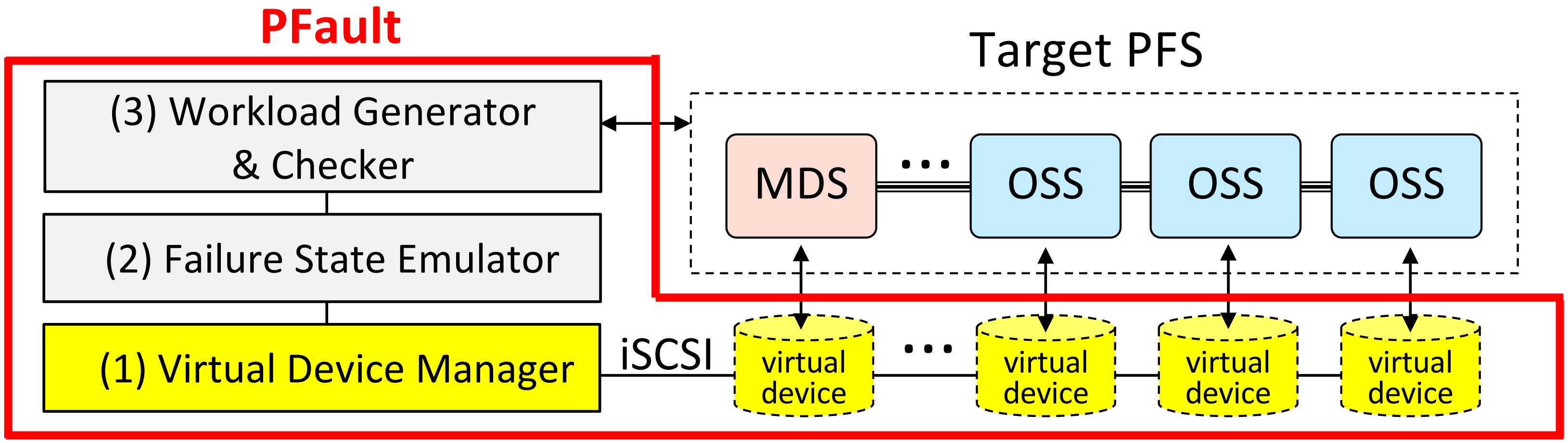


Design Overview



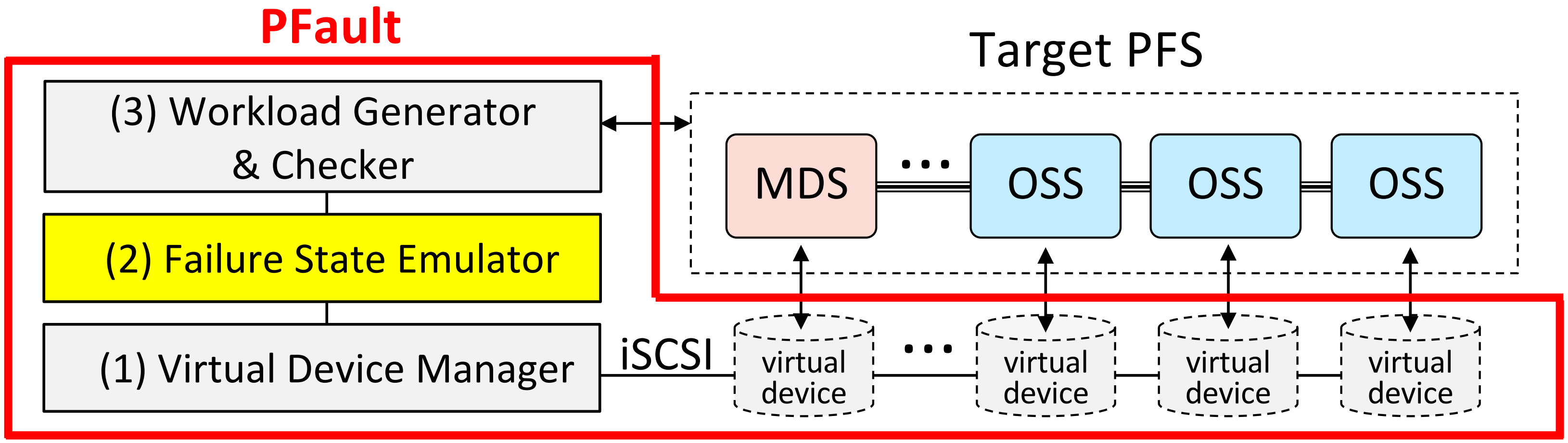
- Three main components
  - (1) Virtual Device Manager:
    - manages the persistent state of the target PFS
  - (2) Failure State Emulator:
    - inject faults based on fault models
  - (3) Workload Generator & Checker:
    - generate I/O operations & check correctness of recovery

Design Overview



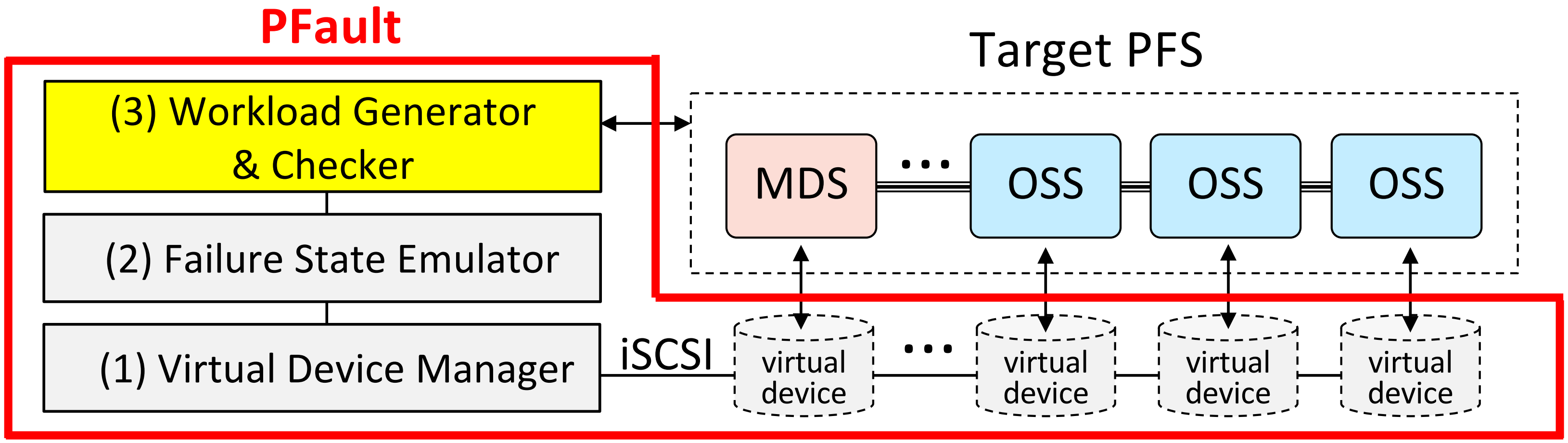
- Three main components
  - (1) Virtual Device Manager:
    - manages the persistent state of the target PFS
  - (2) Failure State Emulator:
    - inject faults based on fault models
  - (3) Workload Generator & Checker:
    - generate I/O operations & check correctness of recovery

Design Overview



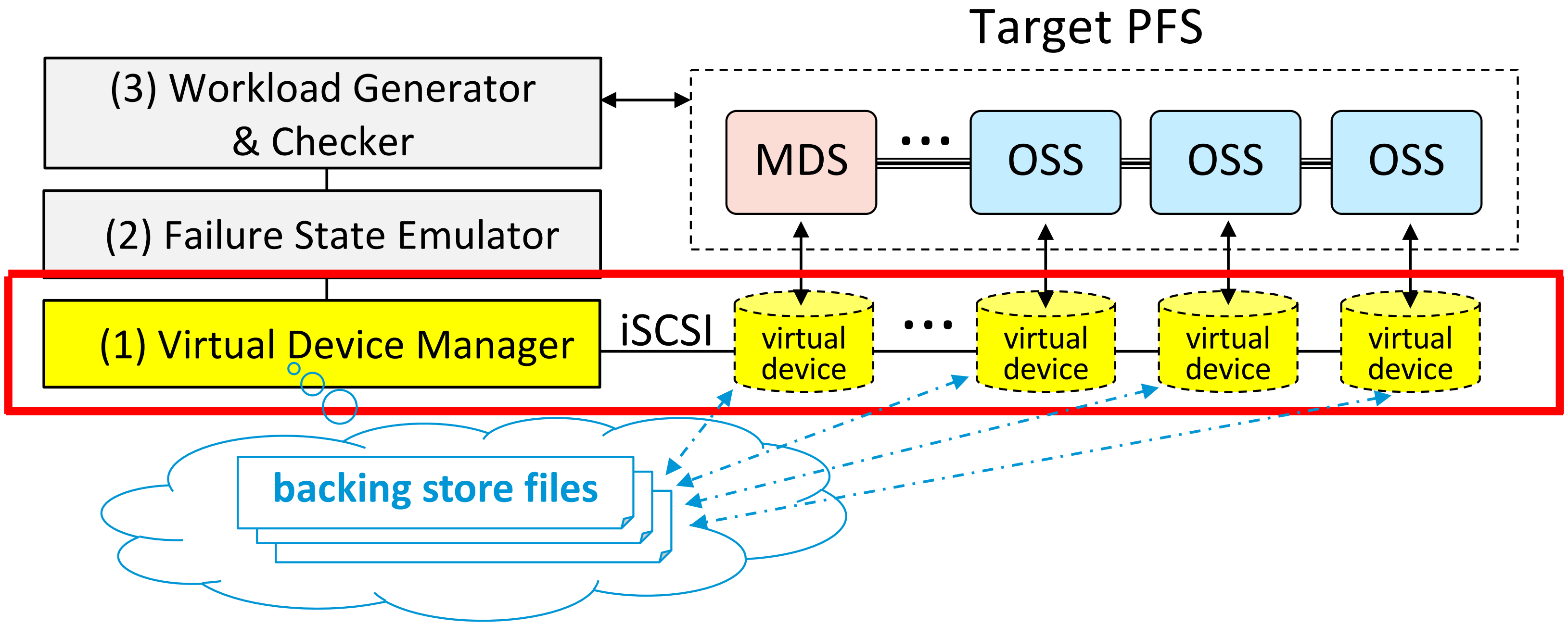
- Three main components
  - (1) Virtual Device Manager:
    - manages the persistent state of the target PFS
  - (2) Failure State Emulator:
    - inject faults based on fault models
  - (3) Workload Generator & Checker:
    - generate I/O operations & check correctness of recovery

Design Overview



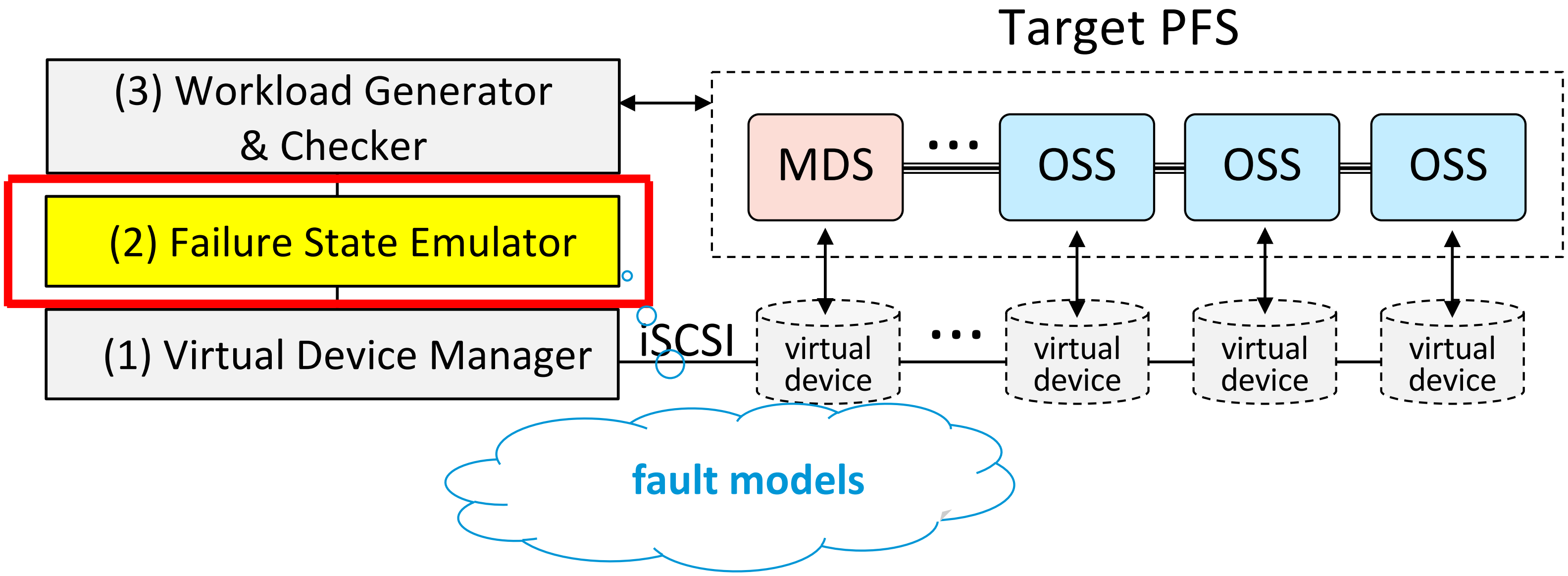
- Three main components
  - (1) Virtual Device Manager:
    - manages the persistent state of the target PFS
  - (2) Failure State Emulator:
    - inject faults based on fault models
  - (3) Workload Generator & Checker:
    - generate I/O operations & check correctness of recovery

(1) Virtual Device Manager



- Manages the persistent state of the target PFS
  - decouple PFault from the PFS via iSCSI (remote storage protocol)
  - mount iSCSI virtual devices on storage nodes
    - transparent to PFS
  - collect I/O commands via virtual devices
    - maintaining backing store files to represent individual device states

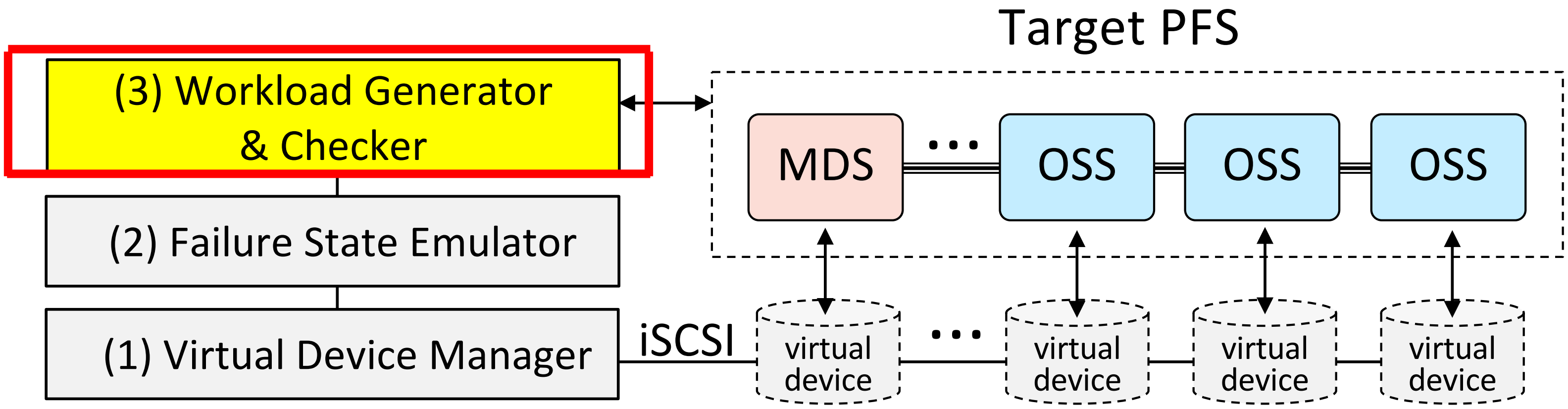
(2) Failure State Emulator



- Emulate failure states based on fault models

Fault Models	Emulating Methods
Whole Device Failure	use logout command in iSCSI to disconnect device
Network Partitioning	disable network cards on selected nodes
Global Inconsistency	corrupt selected local file systems using file system utilities (e.g., debugfs); repair affected local file systems using local checker (e.g., e2fsck)

(3) Workload Generator & Checker



- Generate I/O operations & check correctness of recovery
  - Apply two types of workloads before fault injections

Workloads Examples	Descriptions	Purposes
file manipulation	create, write, delete files	age PFS
Montage	astronomical image mosaics	age PFS
WikiW-init	write a set of Wikipedia files w/ known MD5 checksum	generate verifiable data

  - Check recovery after fault injections
    - Run PFS checker and examine the response and logs
    - Examine the correctness of verifiable workloads (e.g., verify checksum)

Outline

I. Motivation

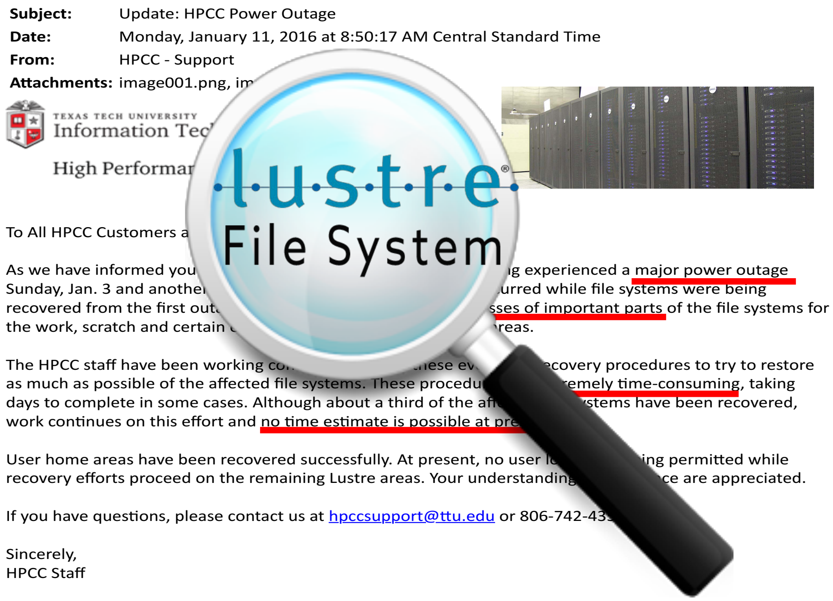
II. Background of PFS

III. Design of PFault

IV. Case Study: Lustre

V. Conclusion & Future Work

Result Overview


lustre®
File System
TEXAS TECH UNIVERSITY
Information Tec
High Performar

Subject: Update: HPCC Power Outage
Date: Monday, January 11, 2016 at 8:50:17 AM Central Standard Time
From: HPCC - Support
Attachments: image001.png, im

To All HPCC Customers a

As we have informed you ig experienced a major power outage
Sunday, Jan. 3 and anothei urred while file systems were being
recovered from the first out sses of important parts of the file systems for
the work, scratch and certain reas.

The HPCC staff have been working co hese ev ecovery procedures to try to restore
as much as possible of the affected file systems. These procedu emely time-consuming, taking
days to complete in some cases. Although about a third of the af stems have been recovered,
work continues on this effort and no time estimate is possible at pre

User home areas have been recovered successfully. At present, no user l ing permitted while
recovery efforts proceed on the remaining Lustre areas. Your understanding ce are appreciated.

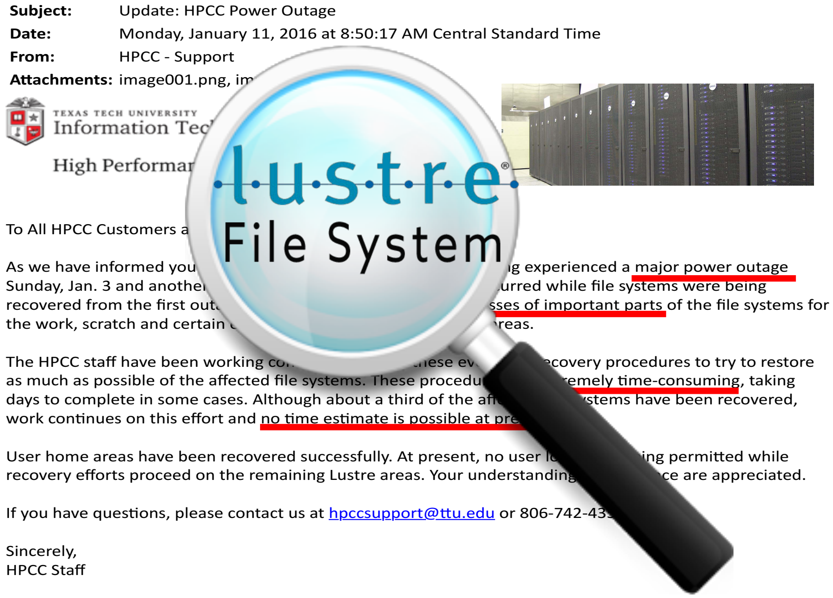
If you have questions, please contact us at hpccsupport@ttu.edu or 806-742-43

Sincerely,
HPCC Staff

- Target PFS: Lustre
  - dominate core HPC market
  - suffered from data loss in HPCC
- Luster checker (LFSCK) itself may behave abnormally
  - crash, hang, etc.

Node Affected	Fault Models	Desired Behavior of LFSCK	Actual Behavior
MDS	Whole Device Failure	report device error	crash (with an I/O error)
	Network Partitioning	report network error	hang (> 1hour)
	Global Inconsistency	report & fix inconsistency	finish w/o any report; resource leak
OSS	Whole Device Failure	report device error	finish w/o any report
	Network Partitioning	report network error	hang (> 1hour)
	Global Inconsistency	report & fix inconsistency	reboot OSS node

Result Overview



Subject: Update: HPCC Power Outage
Date: Monday, January 11, 2016 at 8:50:17 AM Central Standard Time
From: HPCC - Support
Attachments: image001.png, im

TEXAS TECH UNIVERSITY
Information Tec

High Performar

To All HPCC Customers a

As we have informed you g experienced a major power outage Sunday, Jan. 3 and anothe urred while file systems were being recovered from the first out sses of important parts of the file systems for the work, scratch and certain reas.

The HPCC staff have been working co these ev ecovery procedures to try to restore as much as possible of the affected file systems. These procedu remely time-consuming, taking days to complete in some cases. Although about a third of the aff ystems have been recovered, work continues on this effort and no time estimate is possible at pre

User home areas have been recovered successfully. At present, no user lo ing permitted while recovery efforts proceed on the remaining Lustre areas. Your understanding ce are appreciated.

If you have questions, please contact us at hpccsupport@ttu.edu or 806-742-43

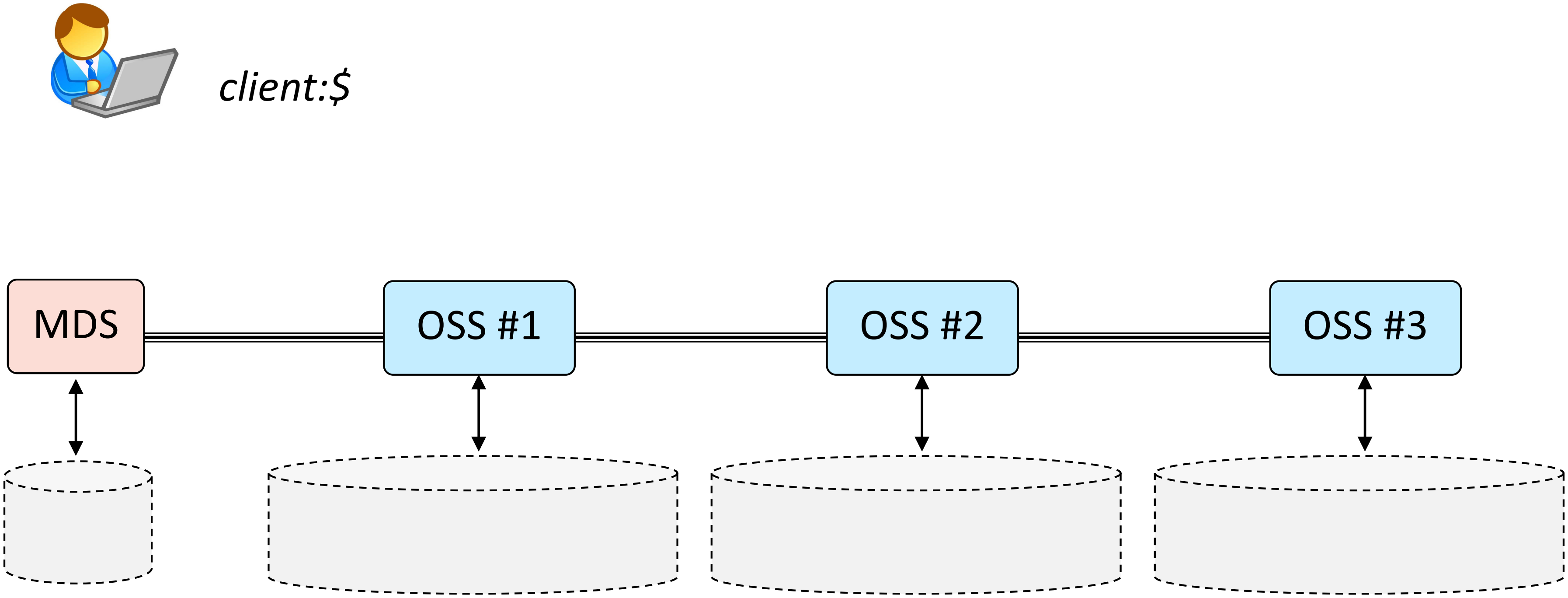
Sincerely,
HPCC Staff

- Target PFS: Lustre
  - dominate core HPC market
  - suffered from data loss in HPCC
- Luster checker (LFSCK) itself may behave abnormally
  - crash, hang, etc.

Node Affected	Fault Models	Desired Behavior of LFSCK	Actual Behavior
MDS	Whole Device Failure	report device error	crash (with an I/O error)
	Network Partitioning	report network error	hang (> 1hour)
	Global Inconsistency	report & fix inconsistency	finish w/o any report; resource leak
OSS	Whole Device Failure	report device error	finish w/o any report
	Network Partitioning	report network error	hang (> 1hour)
	Global Inconsistency	report & fix inconsistency	reboot OSS node

The Resource Leak Problem

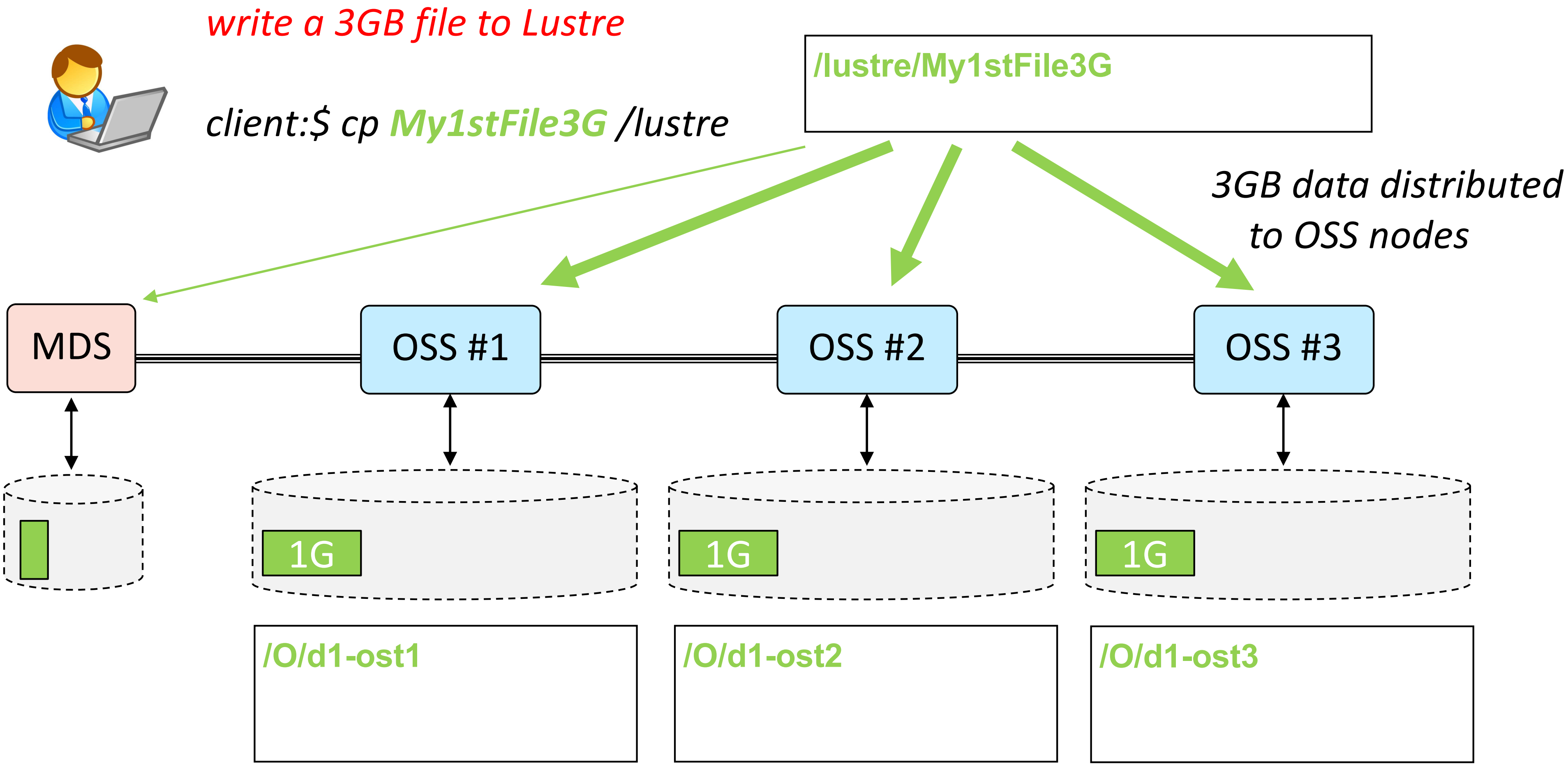
- LFSCK may fail to detect/recycle orphan objects



client:$

The Resource Leak Problem

- LFSCK may fail to detect/recycle orphan objects



write a 3GB file to Lustre

*client:$ cp **My1stFile3G** /lustre*

/lustre/My1stFile3G

3GB data distributed to OSS nodes

MDS

OSS #1

OSS #2

OSS #3

1G

1G

1G

/O/d1-ost1

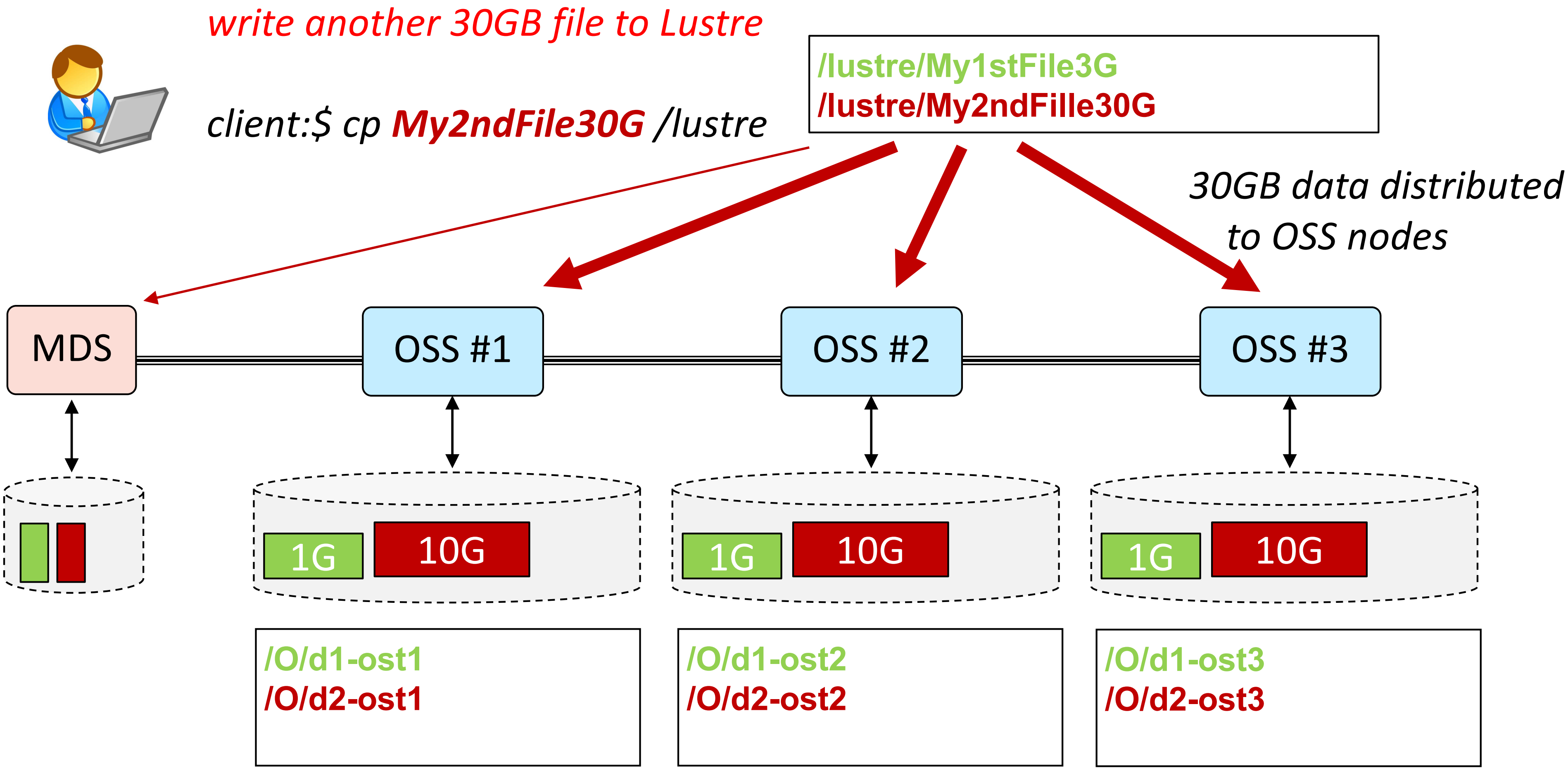
/O/d1-ost2

/O/d1-ost3

a 1GB object file is generated on each OSS node

The Resource Leak Problem

- LFSCK may fail to detect/recycle orphan objects



write another 30GB file to Lustre

client:$ cp ***My2ndFile30G*** */lustre*

/lustre/My1stFile3G
/lustre/My2ndFille30G

30GB data distributed to OSS nodes

/O/d1-ost1
/O/d2-ost1

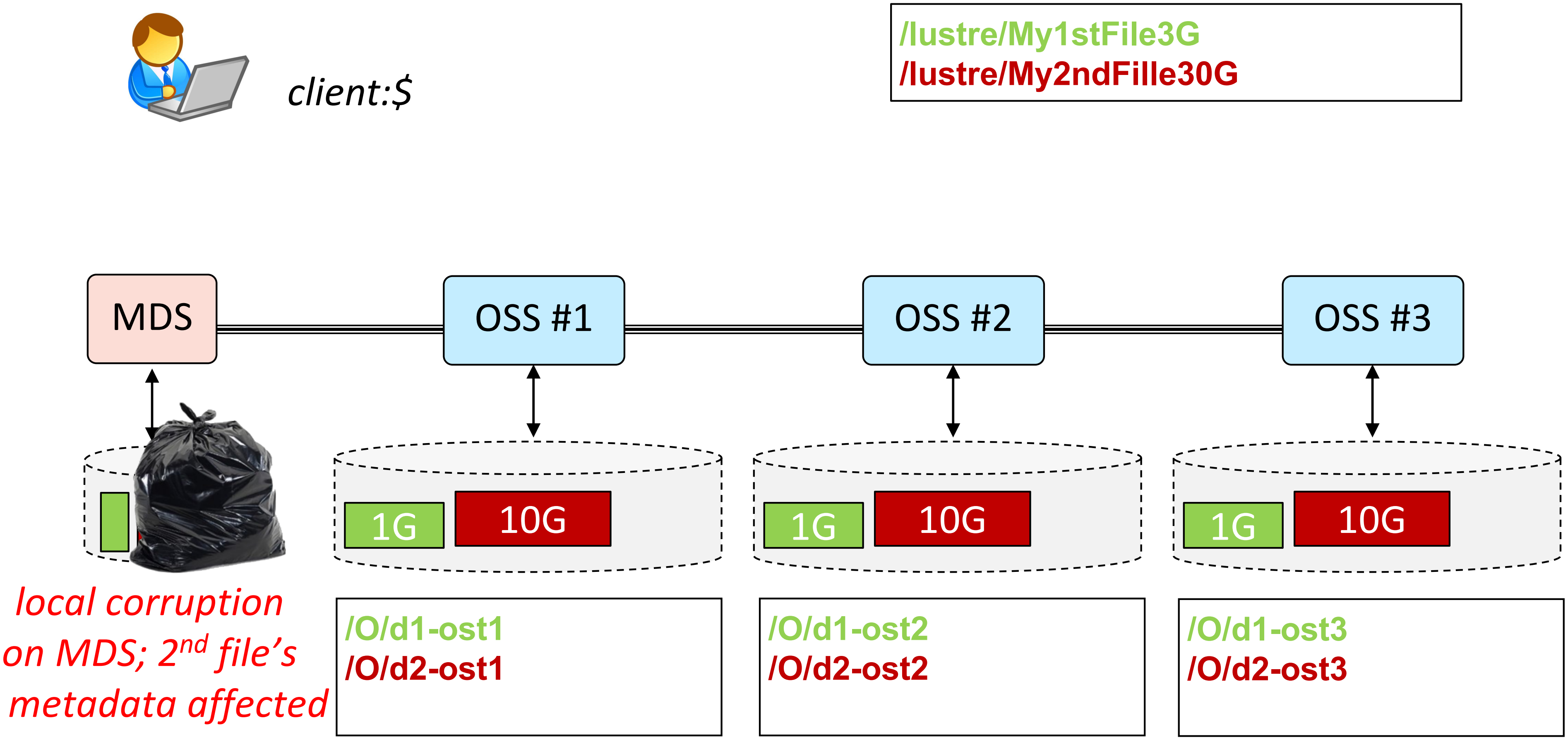
/O/d1-ost2
/O/d2-ost2

/O/d1-ost3
/O/d2-ost3

a 10GB object file is generated on each OSS node

The Resource Leak Problem

- LFSCK may fail to detect/recycle orphan objects



client:$

/lustre/My1stFile3G
/lustre/My2ndFille30G

MDS

OSS #1

OSS #2

OSS #3

1G

10G

1G

10G

1G

10G

/O/d1-ost1
/O/d2-ost1

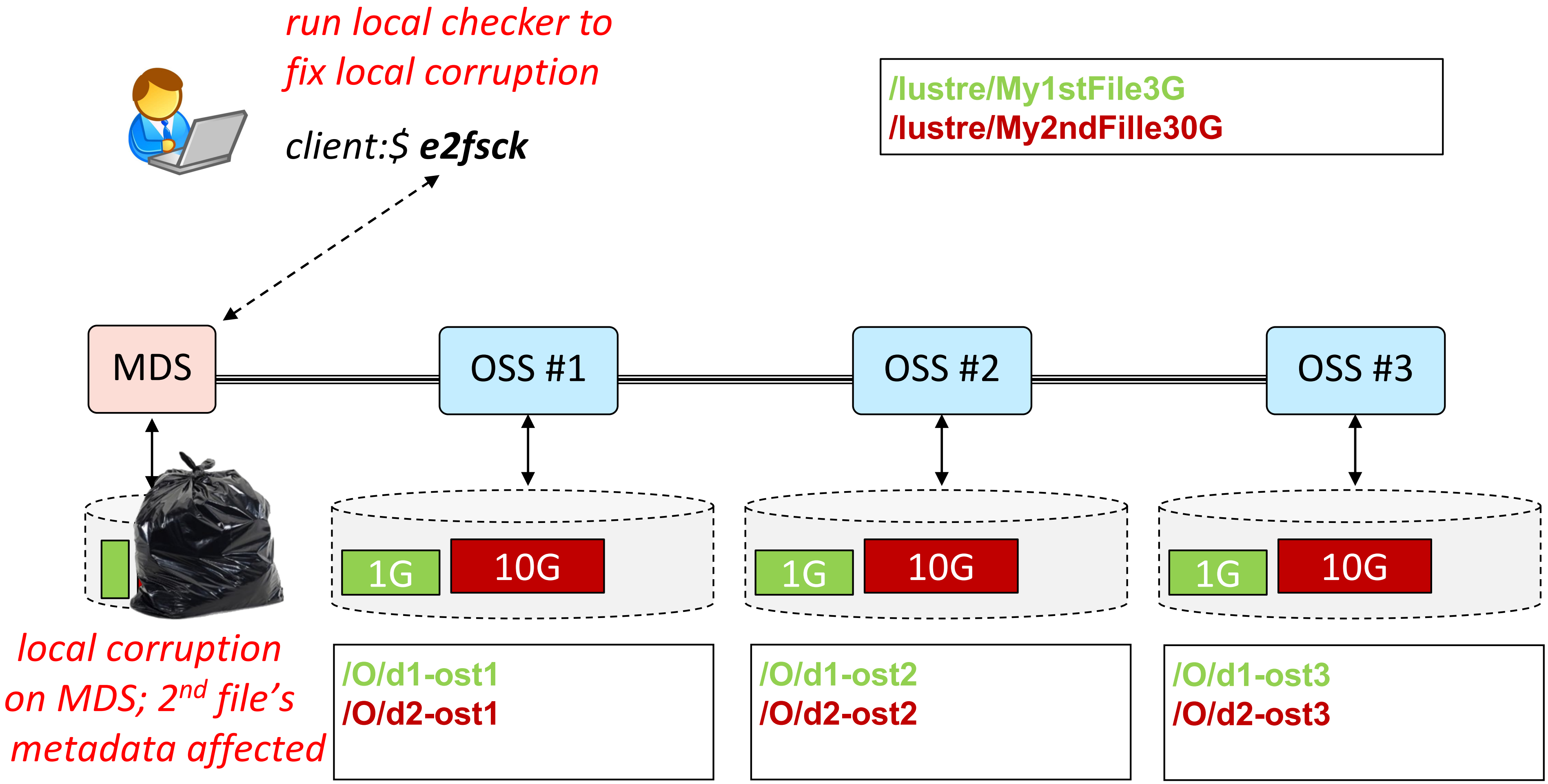
/O/d1-ost2
/O/d2-ost2

/O/d1-ost3
/O/d2-ost3

local corruption on MDS; 2nd file's metadata affected

The Resource Leak Problem

- LFSCK may fail to detect/recycle orphan objects



run local checker to fix local corruption

*client:$ **e2fsck***

/lustre/My1stFile3G

/lustre/My2ndFille30G

MDS

OSS #1

OSS #2

OSS #3

1G 10G

1G 10G

1G 10G

local corruption on MDS; 2nd file's metadata affected

/O/d1-ost1

/O/d2-ost1

/O/d1-ost2

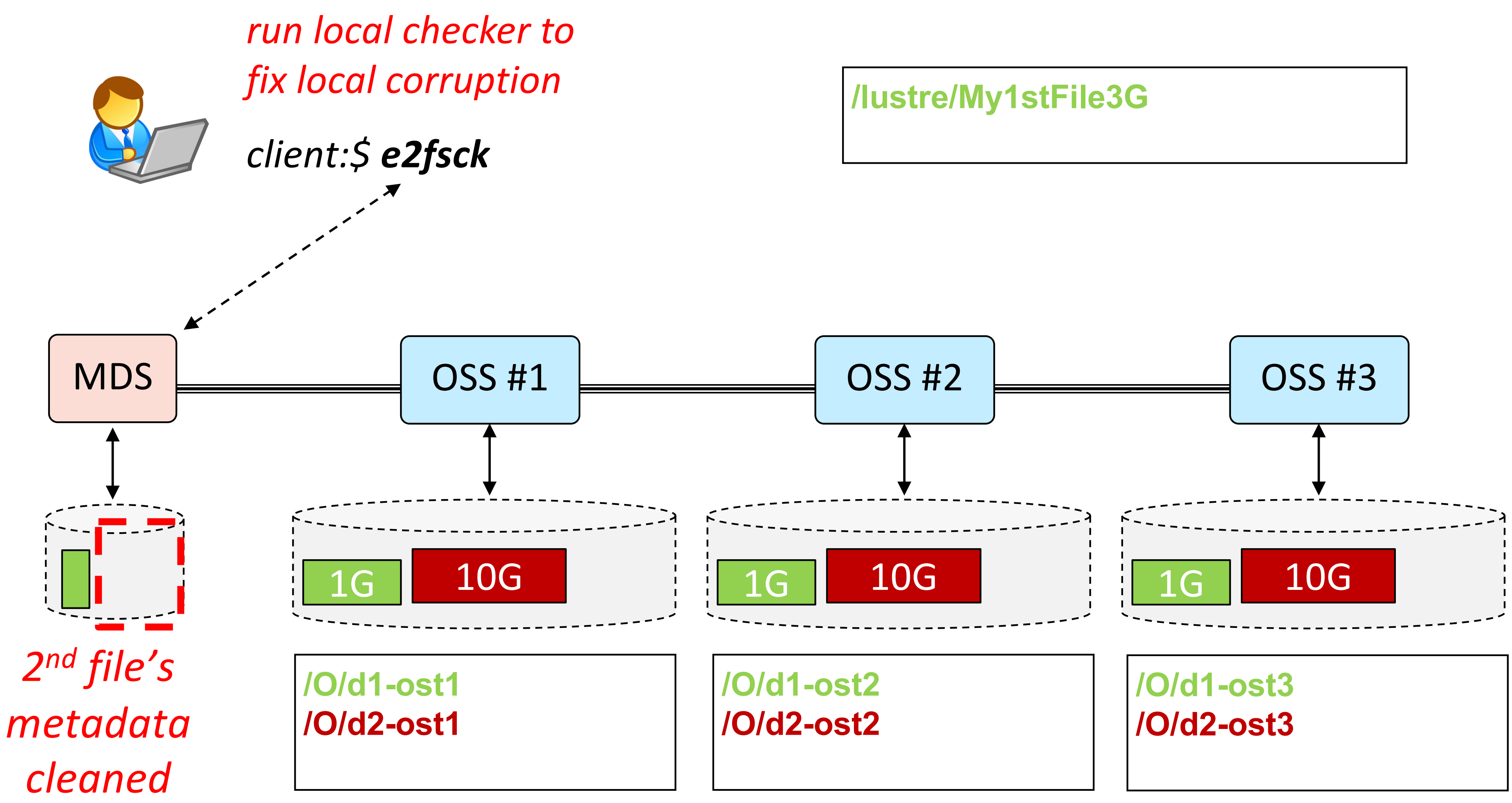
/O/d2-ost2

/O/d1-ost3

/O/d2-ost3

The Resource Leak Problem

- LFSCK may fail to detect/recycle orphan objects



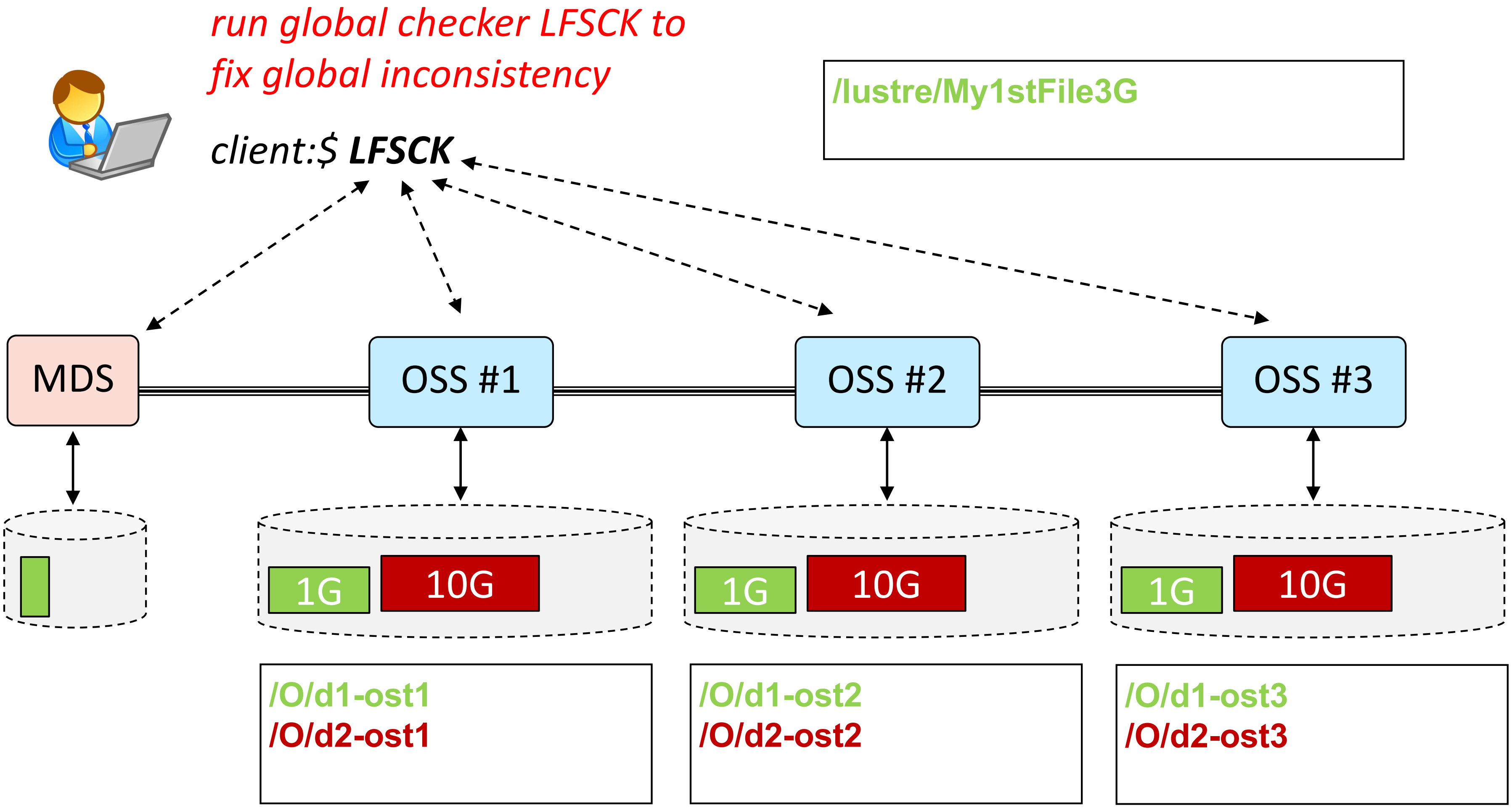
run local checker to fix local corruption

*client:$ **e2fsck***

2nd file's metadata cleaned

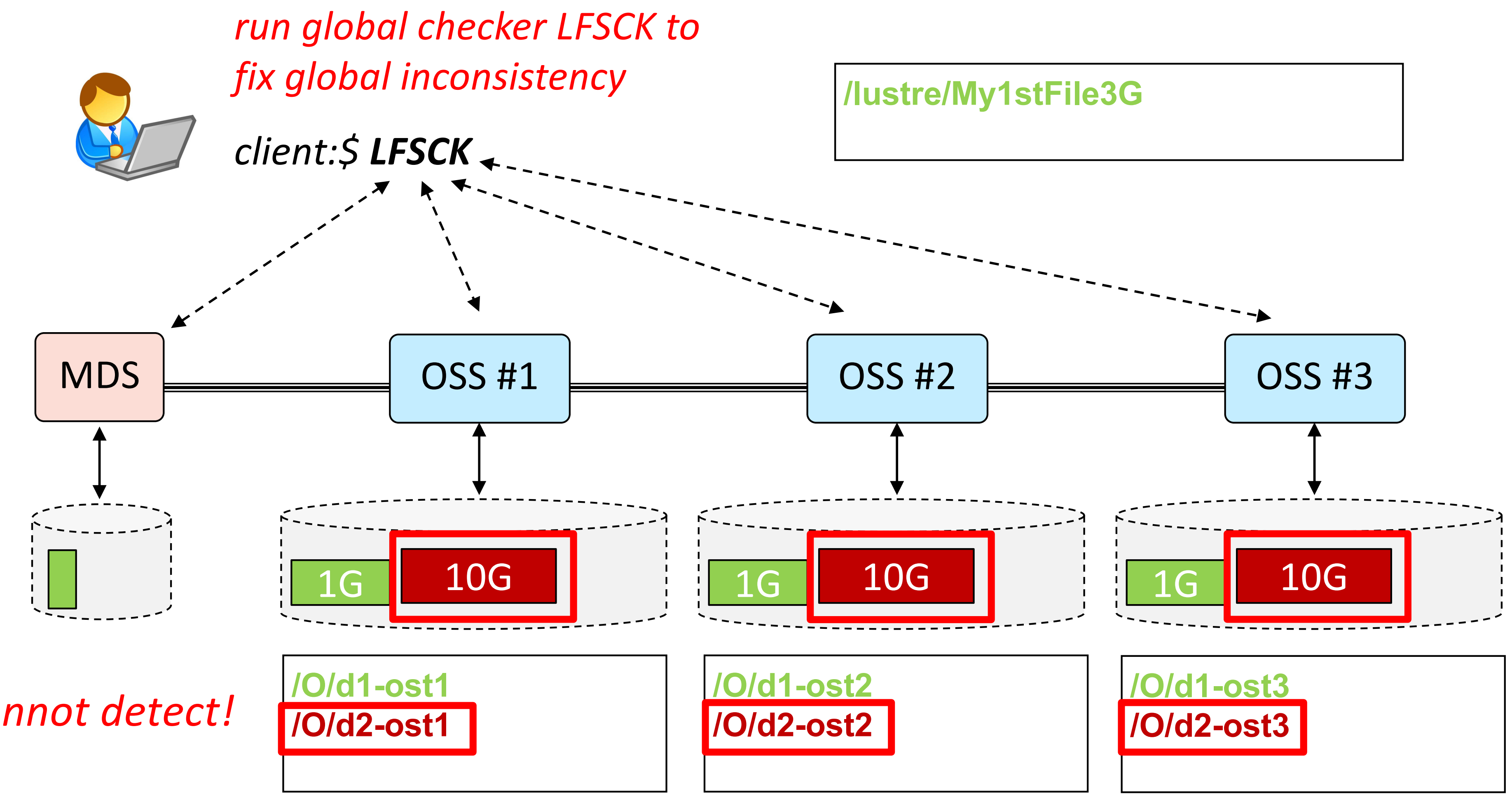
The Resource Leak Problem

- LFSCK may fail to detect/recycle orphan objects



The Resource Leak Problem

- LFSCK may fail to detect/recycle orphan objects



run global checker LFSCK to fix global inconsistency

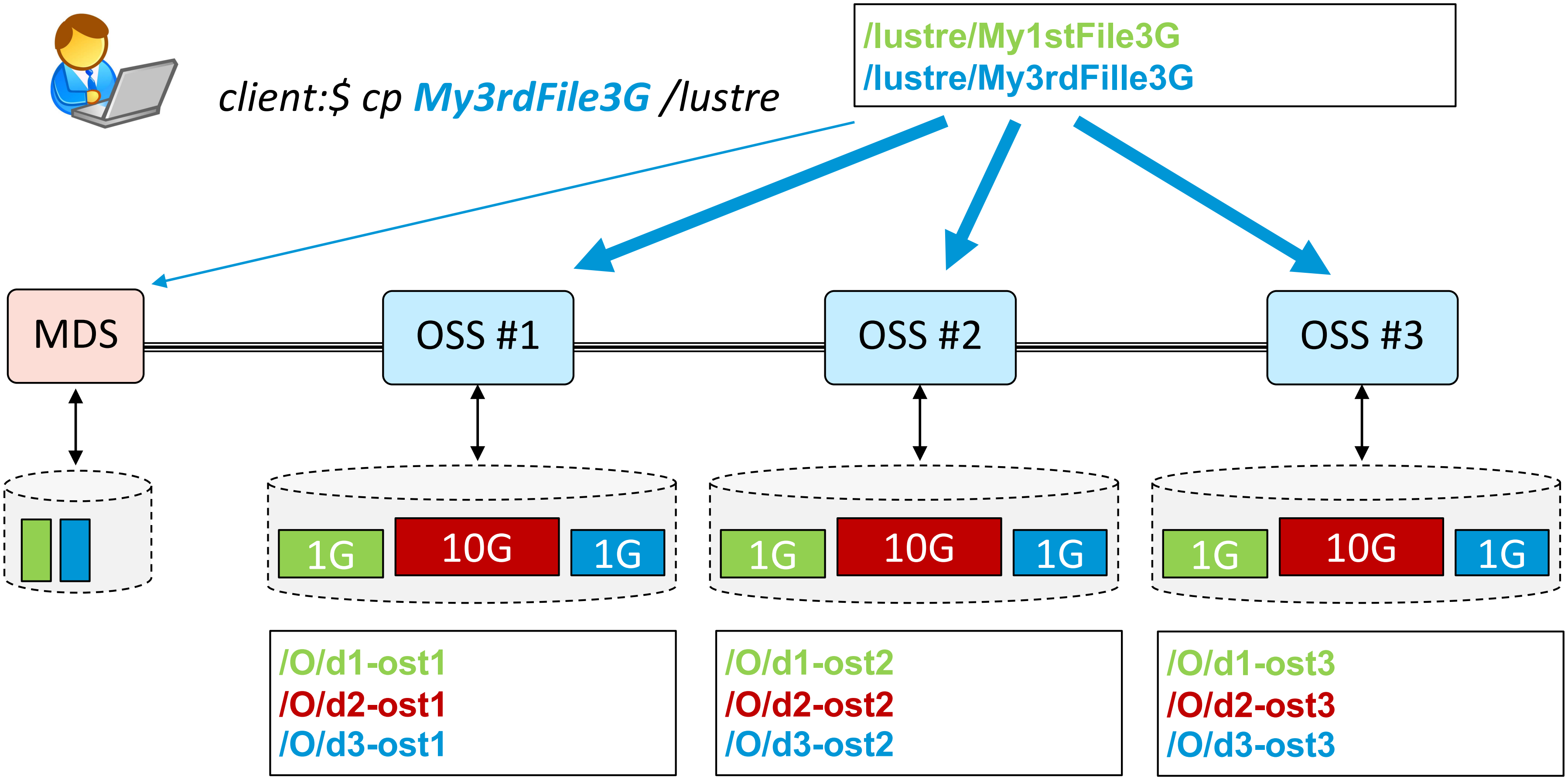
client:$ ***LFSCK***

cannot detect!

The Resource Leak Problem

- LFSCK may fail to detect/recycle orphan objects

keep writing until Lustre reports full



*client:$ cp **My3rdFile3G** /lustre*

/lustre/My1stFile3G
/lustre/My3rdFille3G

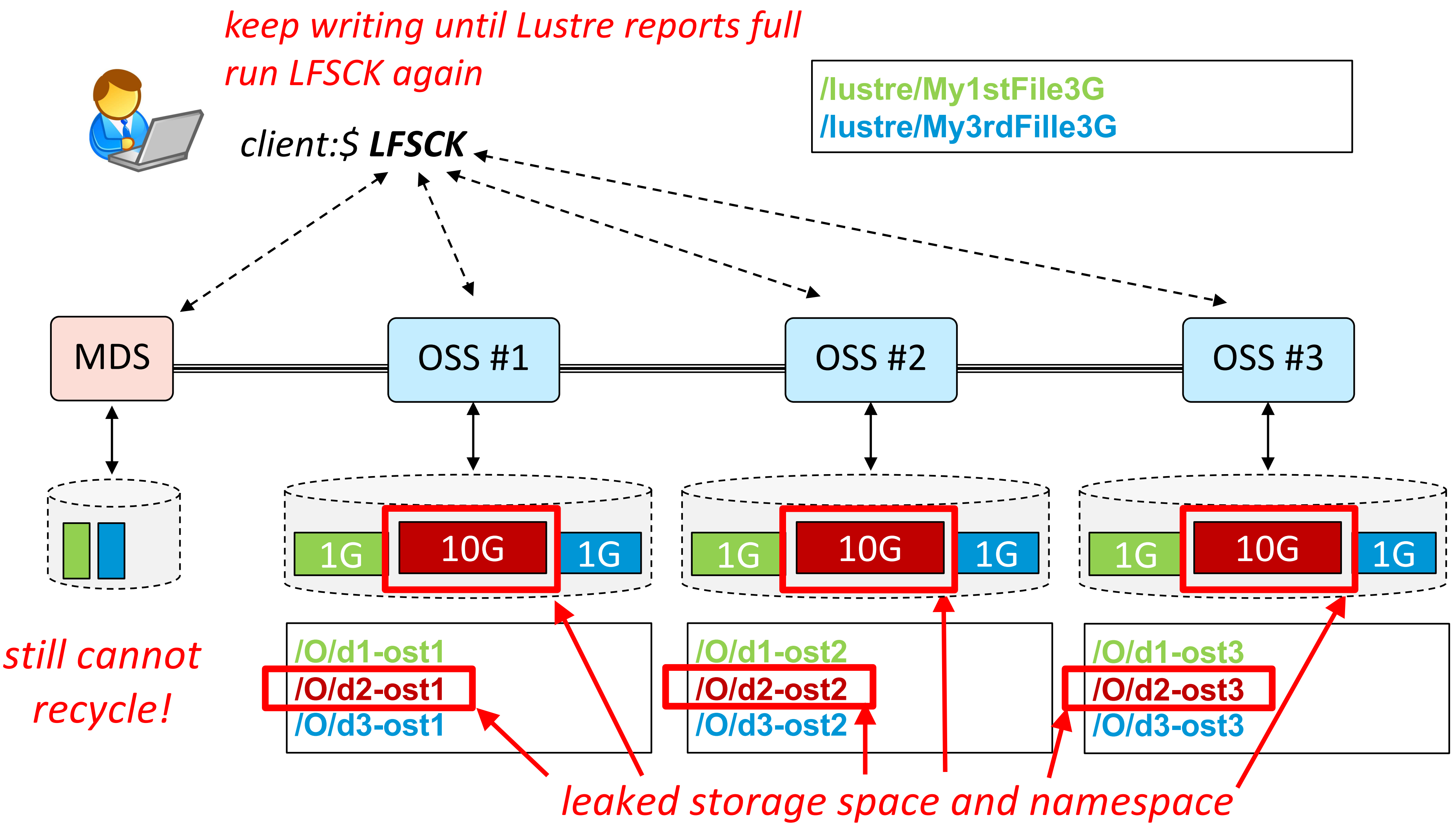
/O/d1-ost1
/O/d2-ost1
/O/d3-ost1

/O/d1-ost2
/O/d2-ost2
/O/d3-ost2

/O/d1-ost3
/O/d2-ost3
/O/d3-ost3

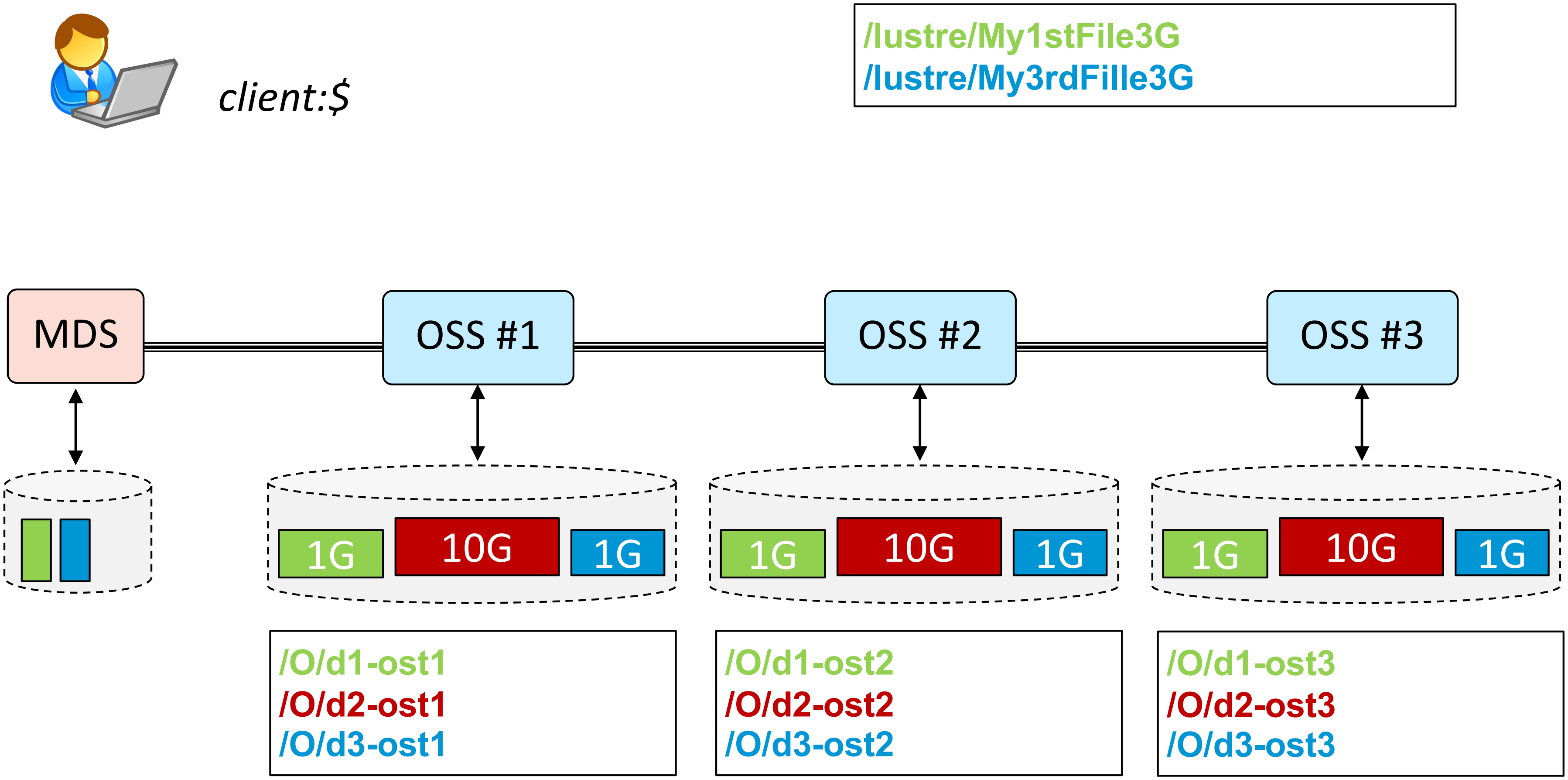
The Resource Leak Problem

- LFSCK may fail to detect/recycle orphan objects



Our Patch: *leak-ck*

- Detect orphan objects based on access time (atime)



client:$

/lustre/My1stFile3G
/lustre/My3rdFille3G

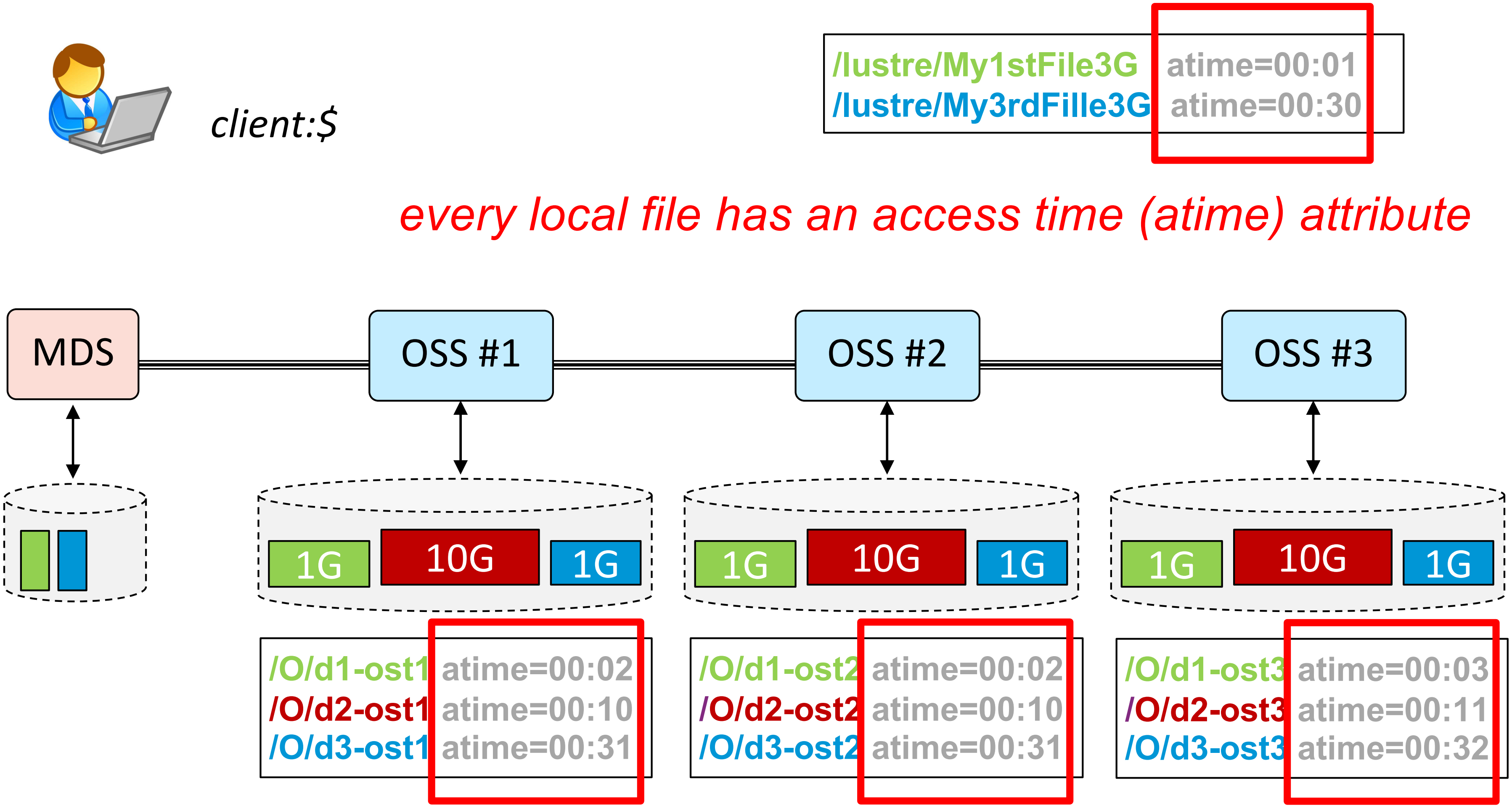
/O/d1-ost1
/O/d2-ost1
/O/d3-ost1

/O/d1-ost2
/O/d2-ost2
/O/d3-ost2

/O/d1-ost3
/O/d2-ost3
/O/d3-ost3

Our Patch: *leak-ck*

- Detect orphan objects based on access time (atime)



client:$

/lustre/My1stFile3G atime=00:01
/lustre/My3rdFille3G atime=00:30

every local file has an access time (atime) attribute

/O/d1-ost1 atime=00:02
/O/d2-ost1 atime=00:10
/O/d3-ost1 atime=00:31

/O/d1-ost2 atime=00:02
/O/d2-ost2 atime=00:10
/O/d3-ost2 atime=00:31

/O/d1-ost3 atime=00:03
/O/d2-ost3 atime=00:11
/O/d3-ost3 atime=00:32

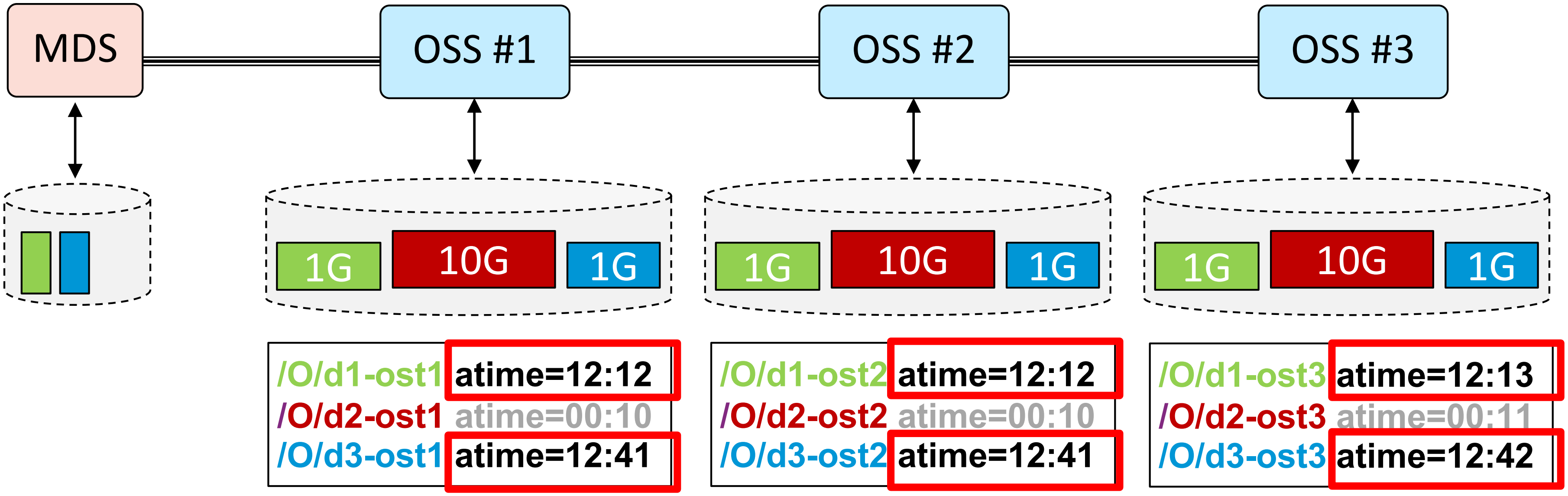
Our Patch: *leak-ck*

- Detect orphan objects based on access time (atime)

client:$ ***leak-ck*** */lustre*

/lustre/My1stFile3G	atime=12:11
/lustre/My3rdFille3G	atime=12:40

touching user files leads to propagated atime updates



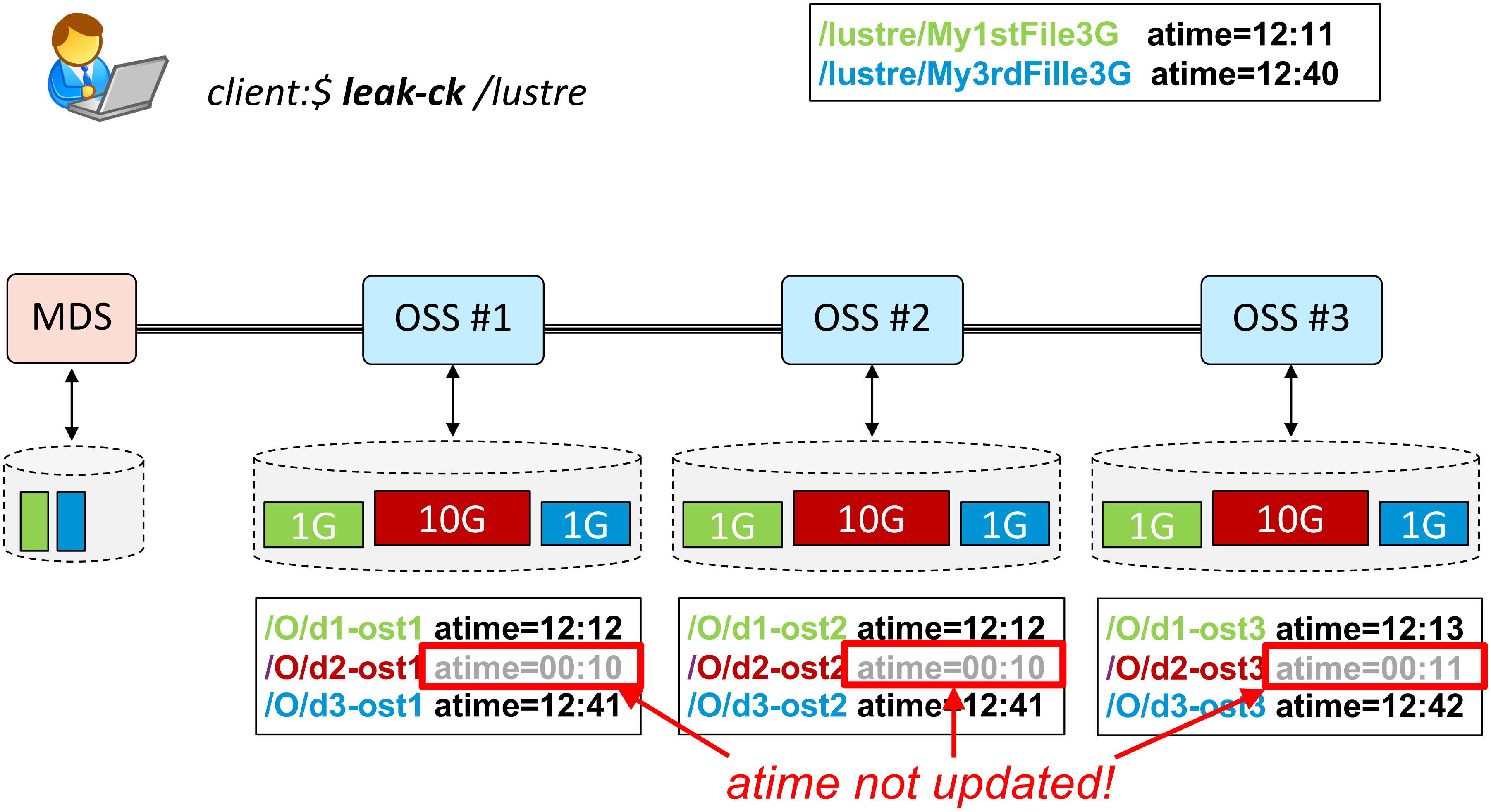
/O/d1-ost1	atime=12:12
/O/d2-ost1	atime=00:10
/O/d3-ost1	atime=12:41

/O/d1-ost2	atime=12:12
/O/d2-ost2	atime=00:10
/O/d3-ost2	atime=12:41

/O/d1-ost3	atime=12:13
/O/d2-ost3	atime=00:11
/O/d3-ost3	atime=12:42

Our Patch: *leak-ck*

- Detect orphan objects based on access time (atime)



*client:$ **leak-ck** /lustre*

/lustre/My1stFile3G atime=12:11
/lustre/My3rdFille3G atime=12:40

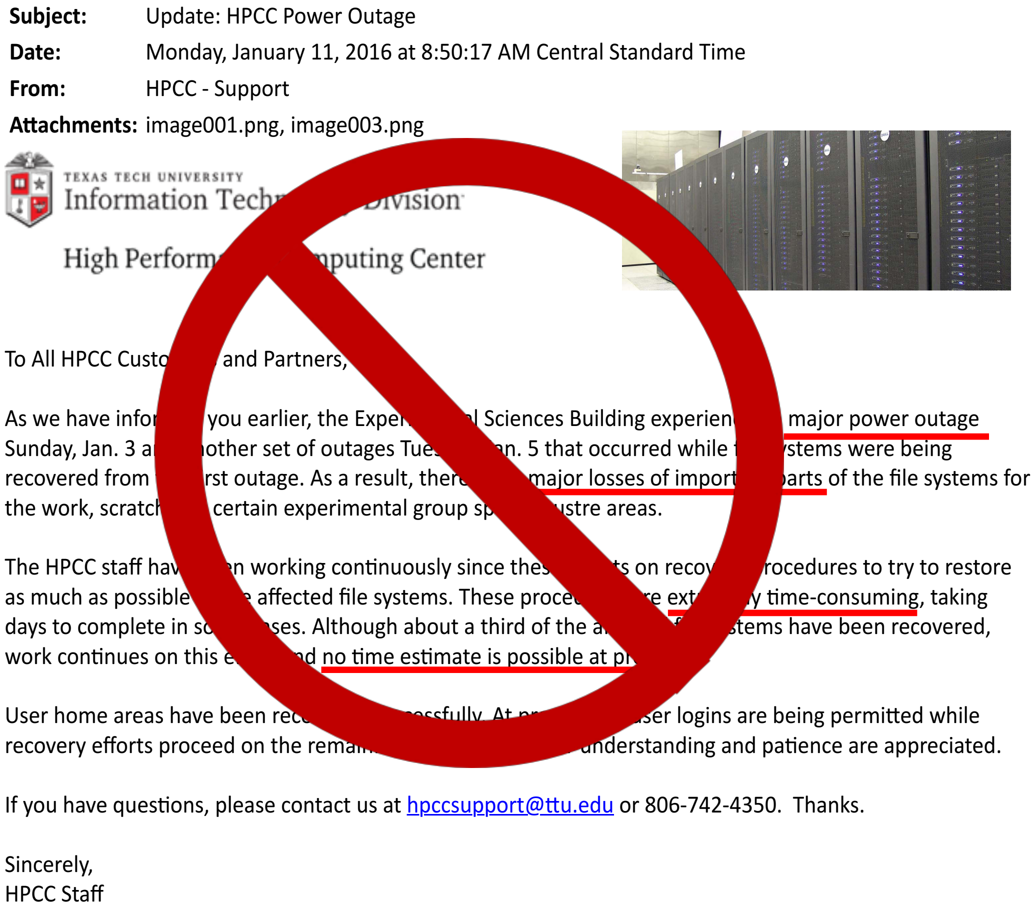
/O/d1-ost1 atime=12:12
/O/d2-ost1 atime=00:10
/O/d3-ost1 atime=12:41

/O/d1-ost2 atime=12:12
/O/d2-ost2 atime=00:10
/O/d3-ost2 atime=12:41

/O/d1-ost3 atime=12:13
/O/d2-ost3 atime=00:11
/O/d3-ost3 atime=12:42

atime not updated!

The Downtime & Data Loss at HPCC Could Have Been Prevented



Subject: Update: HPCC Power Outage
Date: Monday, January 11, 2016 at 8:50:17 AM Central Standard Time
From: HPCC - Support
Attachments: image001.png, image003.png

TEXAS TECH UNIVERSITY
Information Tech Division
High Perform puting Center

To All HPCC Custo and Partners,

As we have infor you earlier, the Exper l Sciences Building experien major power outage Sunday, Jan. 3 a other set of outages Tue n. 5 that occurred while t ystems were being recovered from rst outage. As a result, there major losses of import parts of the file systems for the work, scratch certain experimental group sp ustre areas.

The HPCC staff hav n working continuously since thes ts on recov rocedures to try to restore as much as possible e affected file systems. These proce re ext y time-consuming, taking days to complete in so ses. Although about a third of the a stems have been recovered, work continues on this e d no time estimate is possible at pr

User home areas have been rec ssfully. At pr ser logins are being permitted while recovery efforts proceed on the remai understanding and patience are appreciated.

If you have questions, please contact us at hpccsupport@ttu.edu or 806-742-4350. Thanks.

Sincerely,
HPCC Staff

- Many recovery issues (e.g., crash, hang) can be deterministically exposed by PFault

- Will release the prototype soon

GitHub

Outline

I. Motivation

II. Background of PFS

III. Design of PFault

IV. Case Study: Lustre

V. Conclusion & Future Work

Conclusion & Future Work

- PFault framework + Lustre study
- A wake-up call

  - there are vulnerabilities in widely-used PFSes which may lead to downtime and/or data loss
  - consistent with other studies on large-scale systems
  - will likely become more challenging as the scale & complexity of HPC systems keep increasing
- Future directions
  - understand root causes (crash, hang, resource leak)
  - automate the diagnosis
  - other large-scale systems

Conclusion & Future Work

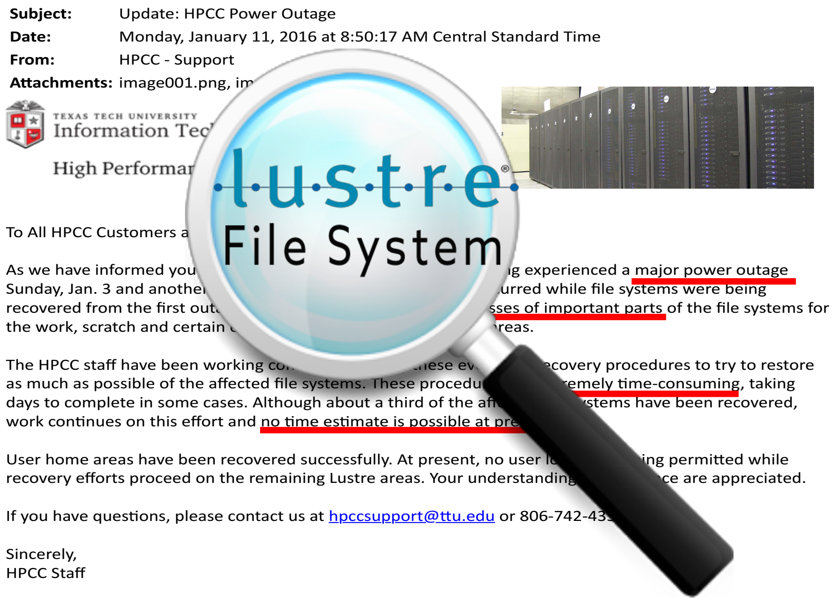
- PFault framework + Lustre study
- A wake-up call
  - there are vulnerabilities in widely-used PFSes which may lead to downtime and/or data loss
  - consistent with other studies on large-scale systems
  - will likely become more challenging as the scale & complexity of HPC systems keep increasing

- Future directions
  - understand root causes (crash, hang, resource leak)
  - automate the diagnosis
  - other large-scale systems

Thank You!

Backup

Result Overview


lustre
File System

Subject: Update: HPCC Power Outage
Date: Monday, January 11, 2016 at 8:50:17 AM Central Standard Time
From: HPCC - Support
Attachments: image001.png, im

TEXAS TECH UNIVERSITY
Information Tec
High Performar

To All HPCC Customers a

As we have informed you g experienced a major power outage Sunday, Jan. 3 and anothe urred while file systems were being recovered from the first out sses of important parts of the file systems for the work, scratch and certain reas.

The HPCC staff have been working co hese ev ecovery procedures to try to restore as much as possible of the affected file systems. These procedu emely time-consuming, taking days to complete in some cases. Although about a third of the af ystems have been recovered, work continues on this effort and no time estimate is possible at pre

User home areas have been recovered successfully. At present, no user lo ing permitted while recovery efforts proceed on the remaining Lustre areas. Your understanding ce are appreciated.

If you have questions, please contact us at hpccsupport@ttu.edu or 806-742-43

Sincerely,
HPCC Staff

- Target PFS: Lustre
  - dominate core HPC market
  - suffered from data loss in HPCC
- Luster checker (LFSCK) itself may behave abnormally
  - crash, hang, etc.

Node Affected	Fault Models	Desired Behavior of LFSCK	Actual Behavior
MDS	Whole Device Failure	report device error	crash (with an I/O error)
	Network Partitioning	report network error	hang (> 1hour)
	Global Inconsistency	report & fix inconsistency	finish w/o any report; resource leak
OSS	Whole Device Failure	report device error	finish w/o any report
	Network Partitioning	report network error	hang (> 1hour)
	Global Inconsistency	report & fix inconsistency	reboot OSS node

More Details of Crash

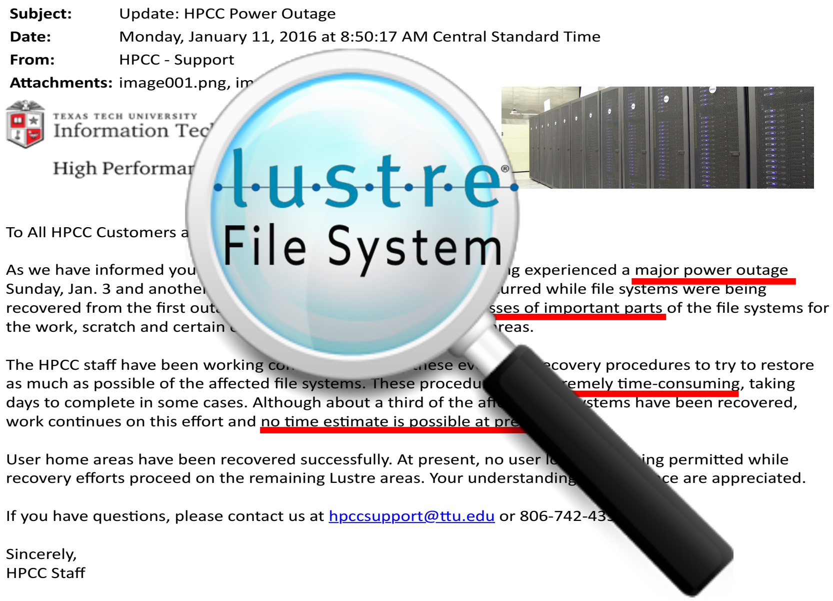
Node Affected	Fault Models	Desired Behavior of LFSCK	Actual Behavior
MDS	Whole Device Failure	report device error	crash (with an I/O error)

- Logs of Lustre and LFSCK

	Logs on MGS	Logs on MDS	Logs on OSSes
Logs of Lustre	*y1*	*y1, y7*	*y1, y3*
Logs of LFSCK	--	*no log*	*initial state*

Message Type	Meaning	Example
y1	*Disconnection*	*...genops.c:1244:class_disconnect() disconnect: cookie 0x923a4db81e68...*
y3	*MDS Recovery failed*	*...ptlrpc_connect_interpret() recovery of lustre-MDT0000_UUID...failed...*
y7	*Failing over MDT*	*...obd_config.c:652:class_cleanup() Failing over lustre-MDT0000...*

Result Overview



Subject: Update: HPCC Power Outage
Date: Monday, January 11, 2016 at 8:50:17 AM Central Standard Time
From: HPCC - Support
Attachments: image001.png, im

TEXAS TECH UNIVERSITY
Information Tec
High Performar

To All HPCC Customers a

As we have informed you ng experienced a major power outage
Sunday, Jan. 3 and anothe urred while file systems were being
recovered from the first out sses of important parts of the file systems for
the work, scratch and certain reas.

The HPCC staff have been working co these ev ecovery procedures to try to restore
as much as possible of the affected file systems. These procedu remely time-consuming, taking
days to complete in some cases. Although about a third of the af stems have been recovered,
work continues on this effort and no time estimate is possible at pre

User home areas have been recovered successfully. At present, no user lo ing permitted while
recovery efforts proceed on the remaining Lustre areas. Your understanding ce are appreciated.

If you have questions, please contact us at hpccsupport@ttu.edu or 806-742-43

Sincerely,
HPCC Staff

- Target PFS: Lustre
  - dominate core HPC market
  - suffered from data loss in HPCC
- Luster checker (LFSCK) itself may behave abnormally
  - crash, hang, etc.

Node Affected	Fault Models	Desired Behavior of LFSCK	Actual Behavior
MDS	Whole Device Failure	report device error	crash (with an I/O error)
	Network Partitioning	report network error	hang (> 1hour)
	Global Inconsistency	report & fix inconsistency	finish w/o any report; resource leak
OSS	Whole Device Failure	report device error	finish w/o any report
	Network Partitioning	report network error	hang (> 1hour)
	Global Inconsistency	report & fix inconsistency	reboot OSS node

More Details of Hang

Node Affected	Fault Models	Desired Behavior of LFSCK	Actual Behavior
MDS/OSS	Network Partitioning	report network error	hang (> 1hour)

- Logs of Lustre and LFSCK

	Logs on MGS	Logs on MDS	Logs on OSSes
Logs of Lustre	*no log*	*y2, y4*	*y3*
Logs of LFSCK	--	*initial state*	*initial state*

Message Type	Meaning	Example
y2	*MGS Recovery failed*	*...ptlrpc_connect_interpret() recovery of MGS on MGC 192.x.x.x...failed...*
y3	*MDS Recovery failed*	*...ptlrpc_connect_interpret() recovery of lustre-MDT0000_UUID...failed...*
y4	*OSS Recovery failed*	*...ptlrpc_connect_interpret() recovery of lustre-OST0001_UUID...failed...*