車載インフォテイメント(IVI)向けのLinuxファイルシステム (Linux File System Analysis for IVI Systems)

October 22, 2014
Yasushi Asano <yasushi.asano@jp.fujitsu.com>
Fujitsu Computer Technologies, Ltd

1279ka02

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Agenda

- Background
- File System Comparison for IVI
- Evaluation of File System Requirements
  - Robustness
  - Boot Up Time
  - Performance
- Conclusions

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Who am I?

- Embedded Software Engineer at Fujitsu Computer Technologies
  - Embedded Linux Distribution and Driver Development (In-House Use), Linux Porting, Technical Support, Training

- Our Distribution is used for Fujitsu's Products
  - Server System Controller, Network Equipment, Printer, **IVI**, and many other systems

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

What is File System?

- One of the Features of Operating Systems
  to Store and Organize **Data** as **Files**
  on **Media**, such as HDD, CD/DVD/BD, Flash Drive

- Linux Supports many kinds of File Systems
  - Disk File Systems : Ext2/3/4, XFS, ReiserFS ZFS, Btrfs, ...
  - Flash File Systems : JFFS/JFFS2/YAFFS, UBIFS, LogFS, F2FS, ...
  - Network File Systems : NFS, Samba, AFS, ...

- Stored Data as Files in IVI Systems
  - 2D/3D Maps, Videos
  - Sounds of Voice, Music, Buzzer, ...
  - Information about Traffic, Shops, Disaster, ...
  - Sensing Data
  - System Logs

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Motivation

- **AGL Requirements Version 1.0** is Planned to be Released Soon
  - **Robust File System**
  - References to **Btrfs**, Ext2/3/4, Vfat, UBIFS

- Fujitsu has been Contributing to **Btrfs** from an Earlier Time
  - No.1 Contributor

- How Suitable are Btrfs and other File Systems for IVI?
  - Functional Requirements?
  - Non-Functional Requirements?

[git://git.kernel.org/pub/scm/linux/kernel/git/stable/linux-stable.git]

```
$ git log
Author: Linus Torvalds <torvalds@linux-foundation.org>
Date:   Sun Oct 5 12:23:04 2014 -0700

    Linux 3.17

$ git log fs/btrfs/ | gitdm
Top changeset contributors by employer
None                       632 (28.1%)
Fujitsu                    518 (23.1%)
Fusion-io                  294 (13.1%)
Oracle                     224 (10.0%)
Red Hat                    194 (8.6%)
Novell                     180 (8.0%)
Facebook                    51 (2.3%)
Intel                       14 (0.6%)
IBM                          9 (0.4%)
Google                       9 (0.4%)
```

→ **FS Suitability Analysis for IVI**

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

File System Comparison for IVI

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

AGL Requirements (Architecture version 0.8.2)

- Robust File System
  - Data stored in a file system must not be corrupted even in an immediately power shutdown.
  - File system must start to work immediately after the system is boot-up.

- Quick boot
  - The system must be ready to operate quickly (such as 5sec).

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Functional Comparison

Type of Storage Device	Name	Btrfs		Ext2/3/4						FAT	UBIFS
		Btrfs	Btrfsck	**Ext2**	E2defrag	**Ext3**	**Ext4**	E4defrag	E2fsck	**Vfat**	**UBIFS**
Internal Managed (**SSD, eMMC,** etc.)	File Systems	✓	N/A	✓	N/A	✓	✓	N/A	N/A	✓	N/A
	Robust File System for managed internal storage	✓		✓		✓	✓				
	Power Failure Tolerance	✓				✓	✓				
	Quick Recovery after power loss	✓				✓	✓				
	Multi-threaded I/O	N/A		N/A		N/A	N/A			N/A	
	On-demand integrity checker	✓		✓							
	Read-only mode	✓	N/A	✓	N/A	✓	✓	N/A	N/A	✓	
	Non-blocking unmounting *	✓		✓		✓	✓			✓	
		7		**5**	N/A	**7**	**7**	N/A	N/A	**3**	
Internal Non-managed (**raw NOR and NAND FLASH memory**)	File System for non-managed internal storage	N/A									✓
	All P1 requirements from FS.1.1.x list										N/A
	Wear leveling										✓
	Error detection /correction										✓
	Tolerance to flipping bits										
	Read/write disturb awareness										
	Bad block management										✓
											4
removable managed (**USB stick, SD card**)	File Systems for removable storage	✓	N/A	✓	N/A	✓	✓	N/A	N/A	✓	N/A
	Restricted functionality from security point of view	✓		✓		✓	✓			✓	
	Automount/autounmount **	✓		✓		✓	✓			✓	
		3		3	N/A	3	3	N/A	N/A	3	

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Functional Comparison for P1 (contd.)

- Btrfs and Ext3/4 are the Most Suitable Candidates for Internal Managed Storage Devices (eMMC, SSD, ...)
- Btrfs and Ext3/4 are also Available for Removable Managed Storage Devices (USB Stick, SD card, ...)
- Ext4 is the Successor to Ext3

→ We Focused on Btrfs and Ext4 as Target of Evaluation

→ In this time, We added F2FS as a file system to be verified.

- All AGL Requirements are Functional ✓

→ We started to Evaluate “Power Failure Tolerance” as the one of Most Important Requirements of IVI

Other Requirements of File Systems

- Short Boot Time ✓
  - Time to Show Splash Screen, Home Screen, and Play Startup Sounds
    (within a few seconds in most cases)
- Performance ✓
  - I/O Throughput
  - Application QoS (Quality of Service) :
    Constant Performance under High Load
    Not to Keep HMI Applications Waiting for a Long Time
- Security
  - Permission Control, Encryption, …
- Scalability

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Overview of Ext4

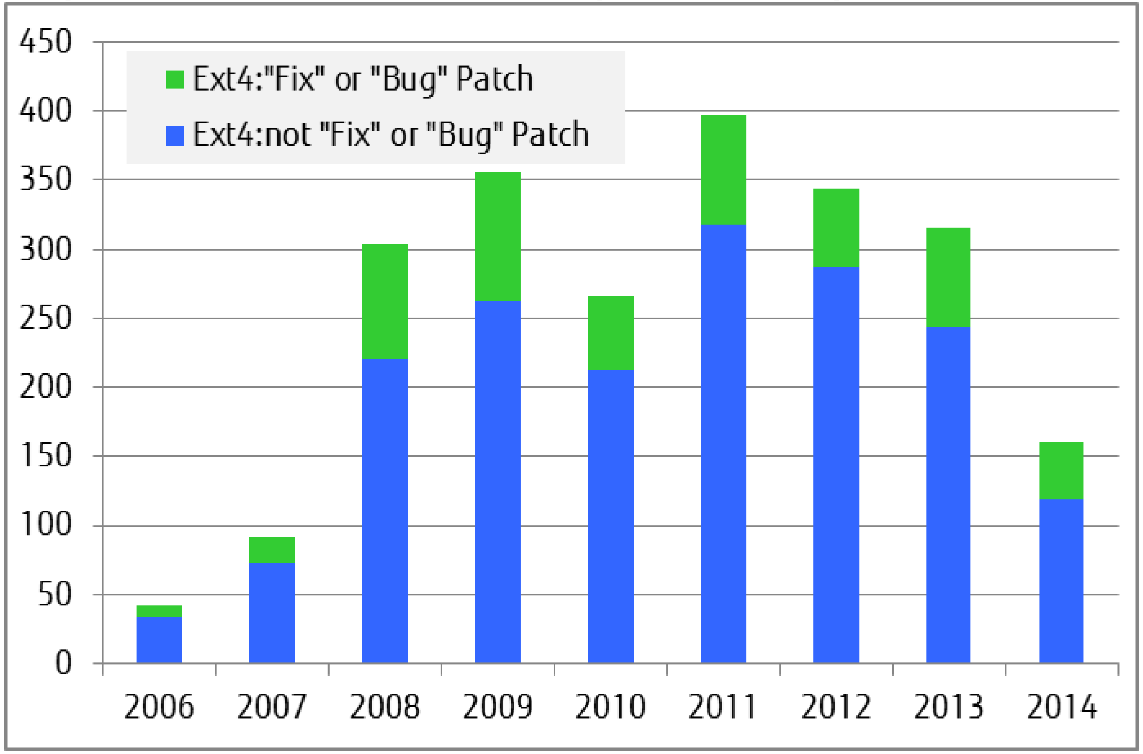
- Journaling File System
  Developed as the Successor to Ext3
- Merged in Mainline **Kernel 2.6.19 in Nov 2006**
- Key Features
  - Large Volume and File Size
  - **Journaling** and Journal Checksum
  - Persistent pre-allocation, ...
- **Standard File System** for Many Major Linux Distros
  - Fedora 11+
  - RHEL 5.6+
  - Ubuntu 9.10+
  - Debian 6.0+

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Overview of Ext4 (contd.)

Development Status

- Mature Enough for Production Use
- Principal Developer of the ext3/4, Theodore Ts'o, [from Wikipedia] stated that although ext4 has improved features,
  it is not a major advance, **it uses old technology**, and **is a stop-gap**.
  Ts'o believes that **Btrfs is the better direction** because
  "it offers improvements in scalability, reliability, and ease of management".



Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Overview of Btrfs

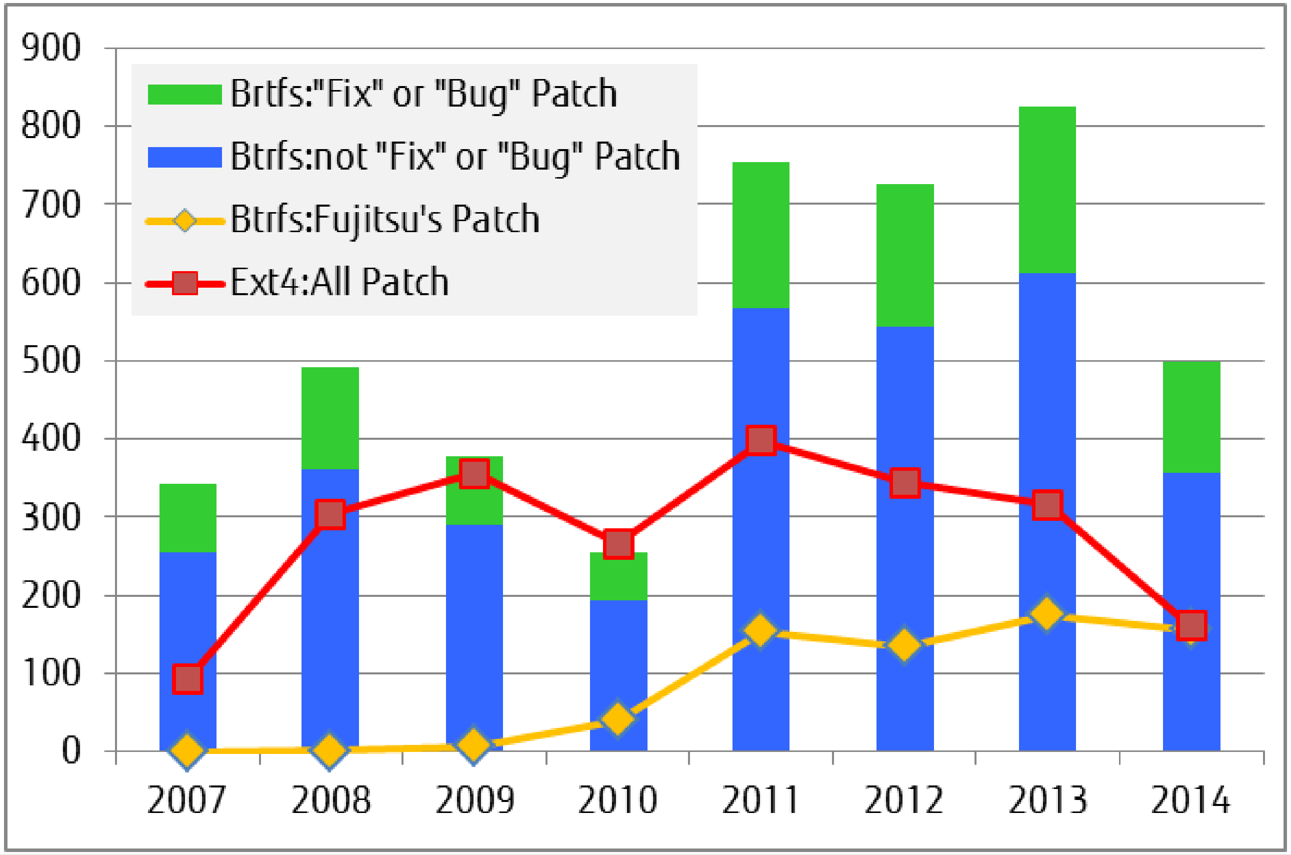
- File System aimed at implementing Advanced Features while focusing on Fault Tolerance, Repair and Easy Administration
- Development began at Oracle in 2007, Merged in Mainline **Kernel 2.6.29 in Jan 2009**
- Key Features
  - Btree Data Structures, Copy on Write (**CoW**) Logging All Data and Metadata (→ Data Consistency and Easy Snapshots)
  - Writable and Read-only **Snapshots**, **Transparent Compression** ,RAID, ...
- Supporting Distributions
  - MeeGo as Standard File System since 2010
  - OpenSUSE 13.2 using Btrfs by **Default** will be released in Nov 2014
  - Oracle Linux since 2012
  - RHEL 7 as a Tech Preview → Btrfs may be supported by Next Version of RHEL
- Facebook
  - Uses Btrfs on their Web Servers

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Overview of Btrfs (contd.)

Development Status

- Some Features are Under Development
- Development has been More Active in the Last Few Years (Twice as Many Patches as Ext4)



Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Overview of F2FS

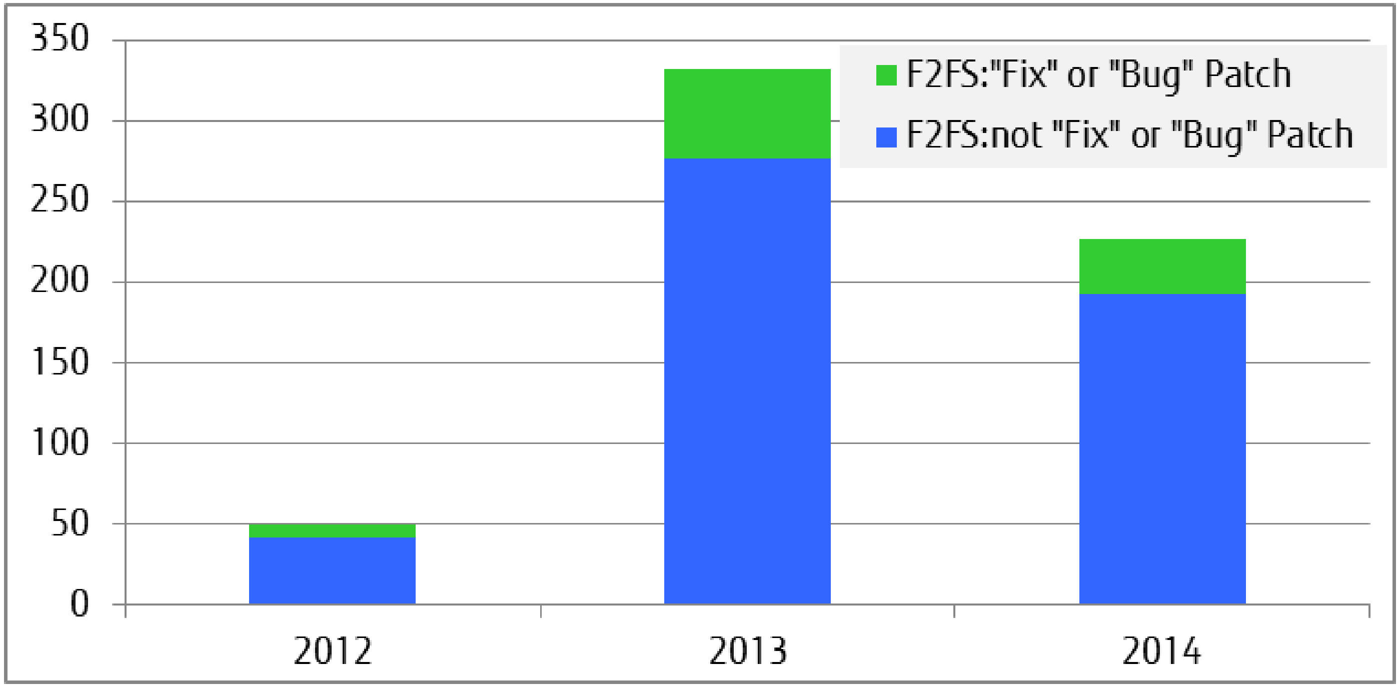
- file system exploiting NAND flash memory-based storage devices, which based on Log-structured File System (LFS).
- Merged in Mainline Kernel 3.8
- designed for delivering maximum file-system performance on flash-based storage devices.
- focused on addressing the issues in LFS
- Key Features
  - Flash Awareness
  - Wandering Tree Problem
  - Cleaning Overhead

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Overview of F2FS (contd.)

Development Status

- Some features are still planned
  - Better direct I/O
  - Transparent compression
  - Data deduplication
  - Removable device support



Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Evaluation of File System Requirements

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Evaluation Plan

- Evaluated Characteristic Requirements of IVI
- Target File System : Btrfs ,Ext4 and F2FS

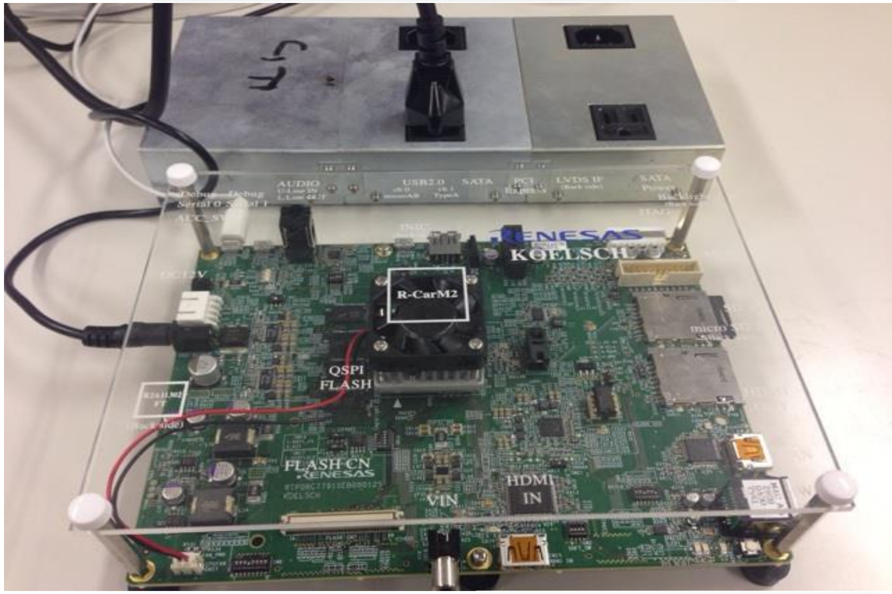
- Eval 1 : Robustness ✓
  - Power Failure Tolerance
- Eval 2 : Boot Up Time ✓
  - FS Mount Time
- Eval 3 : Performance ✓
  - Basic File I/O Throughput
  - File I/O Throughput under High Load

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Eval 1 : Robustness

- Tolerance to Unexpected Power Failure while Writing to Files
- Eval Environment
  - Board
    - Processor : ARMv7
    - Storage : 16GB Micro SD Card
  - Software
    - Yocto based Fujitsu In-House Distro with Kernel 3.17-rc1
- Tools
  - Power Supply Control Unit
    - Periodically Turns On and Off DC Power Supply every Minute
  - File Writing Application
    - Continuously Creates 4KB Files and Writes to it

Power Supply Control Unit



Evaluation Board

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Eval 1 : Robustness (contd.)

Analysis of Results

	Number of Power Failure	Results
Btrfs	1,000+	No Abnormal Situation Occurred
Ext4	1,000+	No Abnormal Situation Occurred
F2FS	1,000+	No Abnormal Situation Occurred

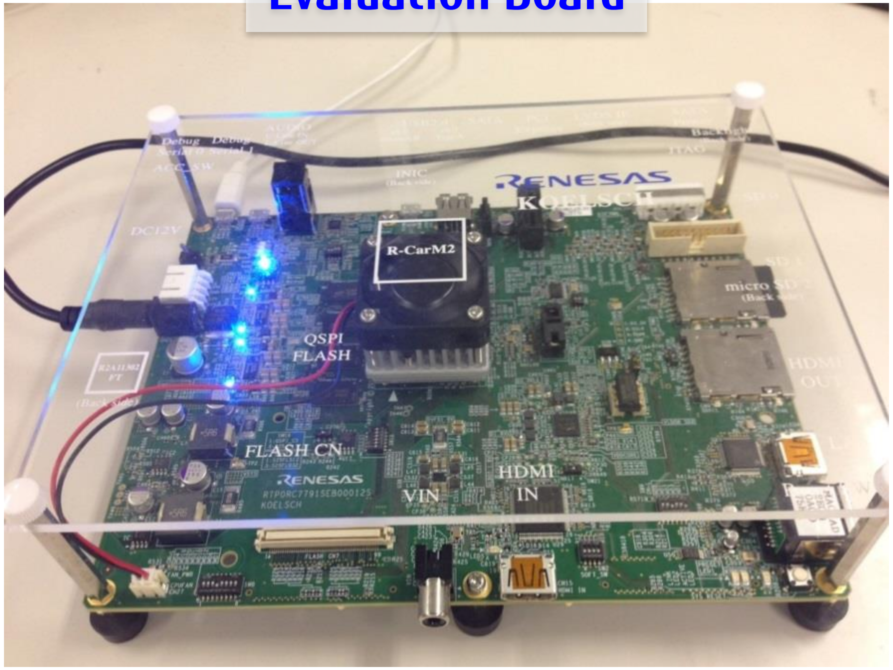
- All File systems showed Very Strong Power Failure Tolerance

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Eval 2 : Boot Up Time

- Time Length of File System Mount
  - Measured after Boot Sequence was Completed in order to measure Real Mount Time
- Eval Environment
  - Board : Renesas R-Car M2 Evaluation Board
  - Software : Fujitsu In-House Distro with Kernel 3.17-rc1 (Simon's tree)
- Tools
  - time command
    - Report how long it took for a command to execute
- Conditions
  - Number of Files : about 4000, 70000, 1000000

Evaluation Board

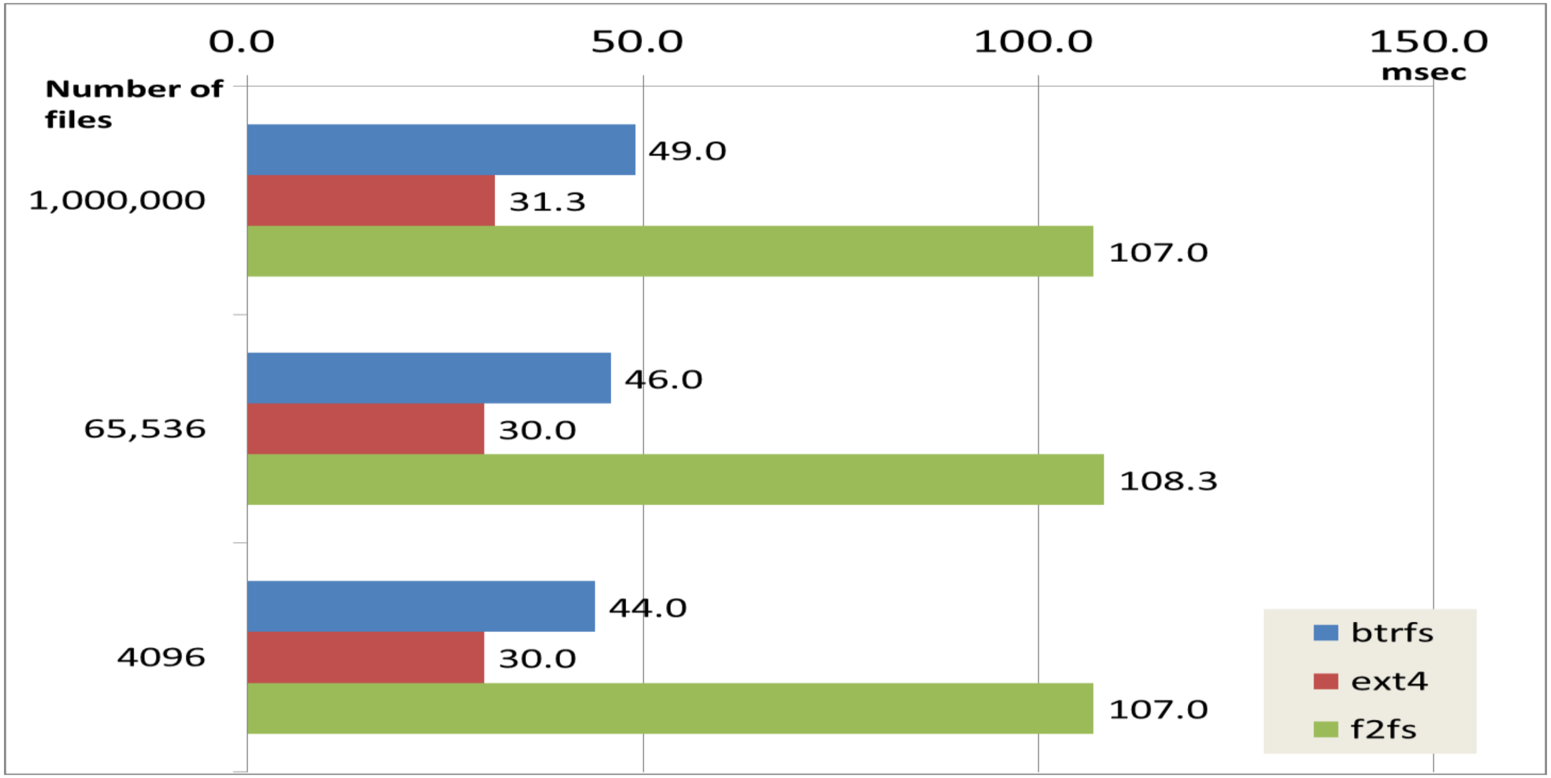


Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Eval 2 : Boot Up Time (contd.)

Analysis of Results

- Btrfs options : rw,relatime,ssd,space_cache
- Ext4 options : rw,relatime,data=ordered
- F2FS options:rw,relatime,background_gc=on,user_xattr,acl,active_logs=6
- Average of 3 Attempts



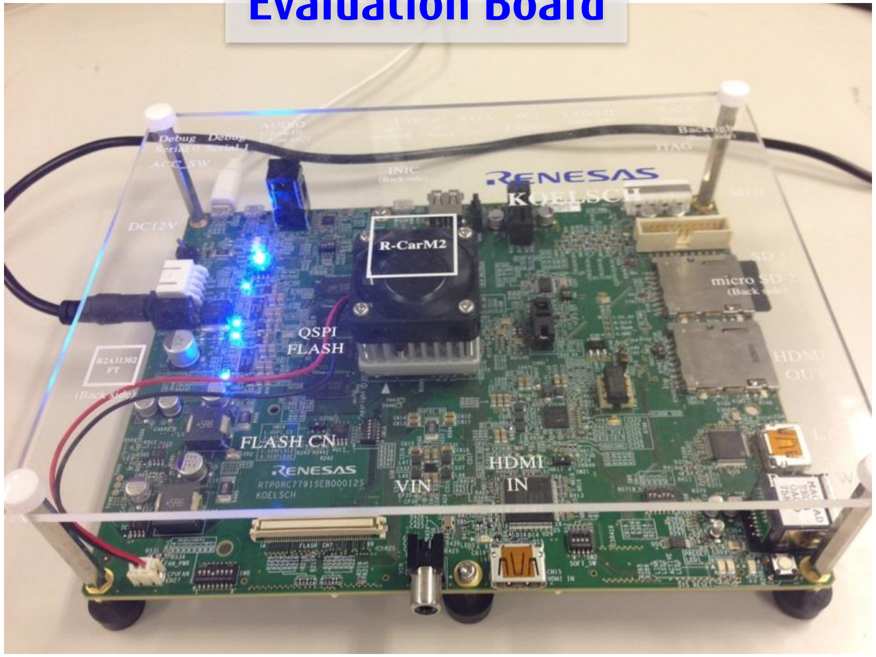
- Ext4 was mounted 3 times or 4 times Faster than F2FS
- Mount Time < 100 msec may have Tiny Impact on a Few Sec Boot Time Reqs

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Eval 3 : Performance

- Basic File I/O Throughput and Throughput under High Load
- Eval Environment
  - Board
    - Renesas R-Car M2 Evaluation Board
    - Processor : R-Car M2
      1.5GHz Cortex-A15 Dual
      780MHz SH-4A (not used)
    - Memory : 1GB DDR3
    - Storage : 32GB Intel X25-E e-SATA SSD
  - Software : Fujitsu In-House Distro
    with Kernel 3.17-rc1(Simon's tree)

Evaluation Board



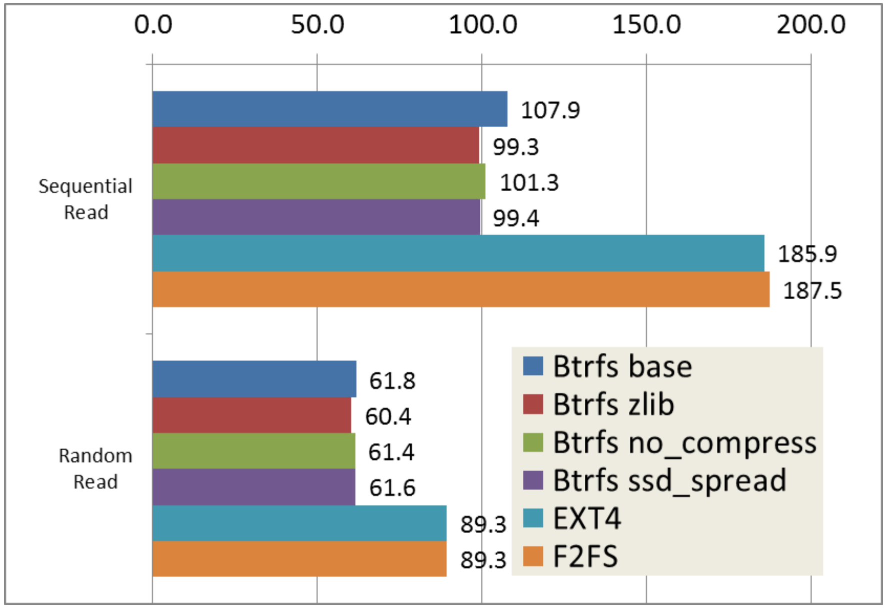
- Tools
  - FIO : to Benchmark and to Make High Load
    (with "yes >> /dev/null" for Userspace Load)
- Conditions
  - Single (for Basic) and Multiple (for High Load) FIO Running
  - FIO makes One Large File (R:2GB, W:1GB) and Reads from/Writes to the Same File with Small Block Size (Seq:64KB, Rand:4KB) (to Simulate DB like Behavior)
  - Some Combinations of Throughput-Related Mount Options

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

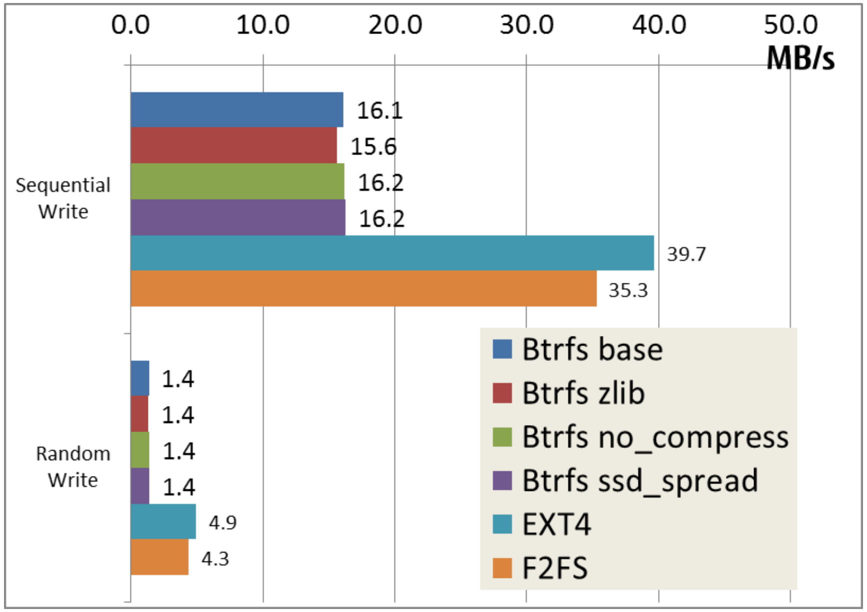
Eval 3 : Performance (contd.)

Analysis of Results : **I/O Throughput** with Single FIO

Read



Write



- Read/Write : Ext4 and F2FS were almost Twice Faster than Btrfs
- Every FS has Advantages and Disadvantages,
  We could see the Other Results on Other Use Cases
- phoronix.com's Benchmark Results
  show Btrfs was the Overall Winner

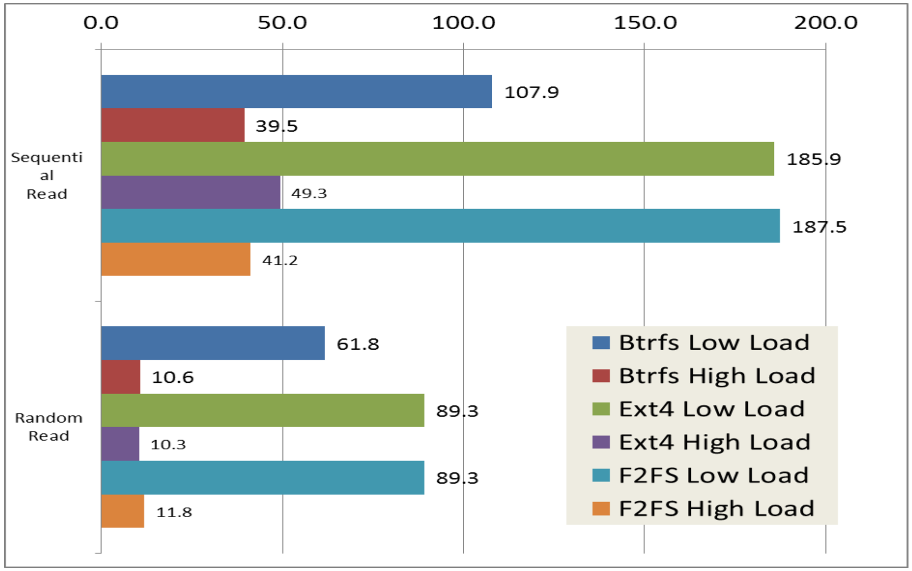
- Btrfs base options : rw,noatime,compress=lzo,ssd,discard,space_cache,autodefrag,inode_cache
- Ext4 options : rw,noatime,discard
- F2FS options : rw,noatime,discard
- File Open with O_SYNC flag, Block Size : Seq 64KB, Rand 4KB, I/O Scheduler : noop
- Average of 3 Attempts

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

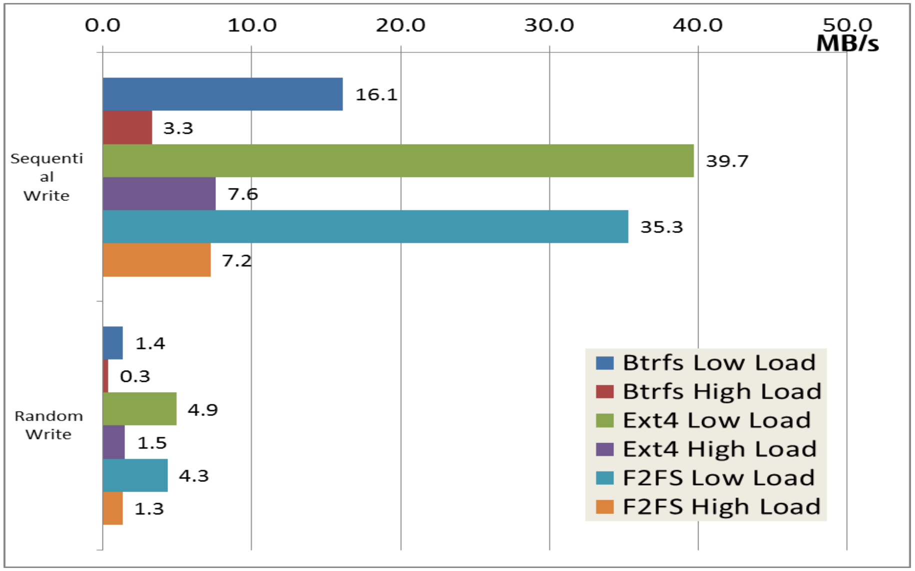
Eval 3 : Performance (contd.)

Analysis of Results : **I/O Throughput** under **High Load**

Read



Write



- Ext4 : Every I/O Throughput Decreased Significantly under High Load

- Btrfs options : rw,noatime,compress=lzo,ssd,discard,space_cache,autodefrag,inode_cache
- Ext4 options : rw,noatime,discard
- F2FS options: rw,noatime,discard
- File Open with O_SYNC flag, Block Size : Seq 64KB, Rand 4KB, I/O Scheduler : noop
- Average of 3 Attempts
- to Make High Load : FIO Seq Read x 2 + Rand Read x 2 + "yes >> /dev/null"

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Conclusions

- **Suitability for IVI Requirements**
  - **Ext4** and **Btrfs** are the most Suitable FS from Functional Aspects
  - Other FS (XFS, NILFS2, ...) may Need to be Evaluated
- **Evaluation Results**

  under Some Specific Environments and Conditions (Like This Study)

	Power Failure Tolerance	Mount Time	I/O Throughput
Btrfs	5	4	Read:3, Write:2, HighLoad:3
Ext4	5	5	Read:5, Write:5, HighLoad:4
F2FS	5	2	Read:5, Write:4, HighLoad:4

Values: 5=Excellent, 4=Very Good, 3=Good, 2=Fair, 1=Poor

  - Effective Mount Options of Btrfs
    - Base : rw,noatime,compress=lzo,ssd,discard,space_cache,autodefrag,inode_cache
    - for Throughput compress : lzo > no compression > zlib

      SSD awareness : ssd_spread > ssd
- **More Evaluations will be Needed for IVI,**

  like phoronix.com's Great Work

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

For the Future

Forecasting the Future …

- HW Specs will become more Rich

  → Priority of Requirements may Change
  such as CPU/Memory Usage, Compression, Boot Up Time, …

- Development of EVs/FCVs may Cause a Change for Requirements of File Systems

  → Power Failures may Almost Never Occur on EVs/FCVs?

- We have to Adapt File Systems to Those Changes Flexibly and Rapidly

- Fujitsu will Continue to Improve Btrfs
- Why don't we Use and Evaluate Btrfs with Various Requirements, Environments, and Conditions to Make Btrfs more Suitable for IVI

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

One-stop Service for Constructing Embedded Linux-based Systems

FUJITSU

We provide technical support for the development of new devices with over 15 years of experience in the development of embedded Linux-based systems

http://www.fujitsu.com/jp/group/fct/index.html

Category	Menu	Service overview
Technical support	Consulting	Validity determination in your devices adopting Linux and required extraction of hardware design support.
	Q&A	Detailed information about constructing a Linux-based system.
		Advice on implementation and operation of your system.
		Troubleshooting for handling problems on devices.
	Training	Techniques and know-how on building an embedded Linux system.
	Provision of information	Provision of vulnerability notes and bug reports.
Development support	Porting Linux	Linux porting and device driver development service to support customer platforms.
	Bug Fix Support	Fixing security vulnerabilities and Linux bugs.

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Thank you for your Attention

References

Btrfs

- https://btrfs.wiki.kernel.org/

Ext4

- http://en.wikipedia.org/wiki/Ext4

Benchmarking

- http://freecode.com/projects/fio
- http://www.phoronix.com/

AGL Requirements

- Keynote of Automotive Linux Summit, Oct 24th, 2013

  Kenichi Murata, Toyota Motor Corporation

The names of products are the product names, trademarks or registered trademarks of the respective companies.
Trademark notices ((R),TM) are not necessarily displayed on system names and product names in this material.

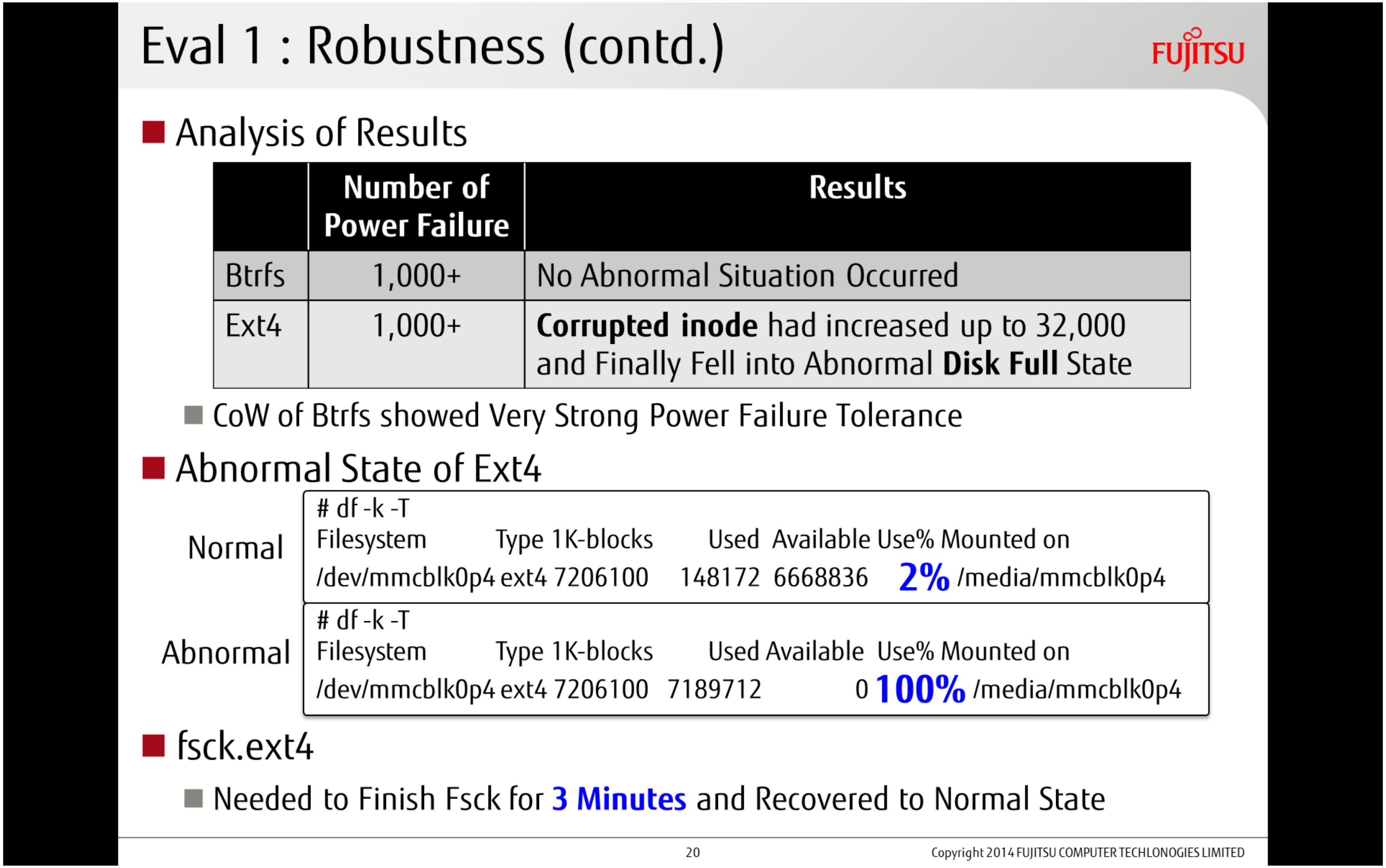
Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

FUJITSU

shaping tomorrow with you

Eval 1 : Robustness (Additional Resources)

- We found a big problem for boot time requirements in EXT4 with kernel v3.15



Eval 1 : Robustness (contd.)

- Analysis of Results

	Number of Power Failure	Results
Btrfs	1,000+	No Abnormal Situation Occurred
Ext4	1,000+	**Corrupted inode** had increased up to 32,000 and Finally Fell into Abnormal **Disk Full** State

  - CoW of Btrfs showed Very Strong Power Failure Tolerance

- Abnormal State of Ext4

Normal

```
# df -k -T
Filesystem        Type 1K-blocks      Used  Available Use% Mounted on
/dev/mmcblk0p4 ext4 7206100    148172  6668836    2% /media/mmcblk0p4
```

Abnormal

```
# df -k -T
Filesystem        Type 1K-blocks      Used Available  Use% Mounted on
/dev/mmcblk0p4 ext4 7206100   7189712            0 100% /media/mmcblk0p4
```

- fsck.ext4
  - Needed to Finish Fsck for **3 Minutes** and Recovered to Normal State

20

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED

Eval 1 : Robustness (Additional Resources)

v3.15 -> v3.16   It was fixed

```
From f9ae9cf5d72b3926ca48ea60e15bdbb840f42372 Mon Sep 17 00:00:00 2001
From: Theodore Ts'o <tytso@mit.edu>
Date: Fri, 11 Jul 2014 13:55:40 -0400
Subject: [PATCH 44/46] ext4: revert commit which was causing fs corruption after journal replays

Commit 007649375f6af2 ("ext4: initialize multi-block allocator before
checking block descriptors") causes the block group descriptor's count
of the number of free blocks to become inconsistent with the number of
free blocks in the allocation bitmap.  This is a harmless form of fs
corruption, but it causes the kernel to potentially remount the file
system read-only, or to panic, depending on the file systems's error
behavior.

Thanks to Eric Whitney for his tireless work to reproduce and to find
the guilty commit.

Fixes: 007649375f6af2 ("ext4: initialize multi-block allocator before checking block descriptors"Cc: stable@vger.kernel.org
# 3.15
Reported-by: David Jander <david@protonic.nl>
Reported-by: Matteo Croce <technoboy85@gmail.com>
Tested-by: Eric Whitney <enwlinux@gmail.com>
Suggested-by: Eric Whitney <enwlinux@gmail.com>
Signed-off-by: Theodore Ts'o <tytso@mit.edu>
---
 fs/ext4/super.c |   51 ++++++++++++++++++++++++---------------------------
 1 files changed, 24 insertions(+), 27 deletions(-)

diff --git a/fs/ext4/super.c b/fs/ext4/super.c
index 6297c07..6df7bc6 100644
--- a/fs/ext4/super.c
+++ b/fs/ext4/super.c
@@ -3879,38 +3879,19 @@ static int ext4_fill_super(struct super_block *sb, void *data, int silent)
                                            goto failed_mount2;
                             }
               }
   :
   :
```

Copyright 2014 FUJITSU COMPUTER TECHLONOGIES LIMITED