Dr. Stefan Reimbold
IBM Research & Development
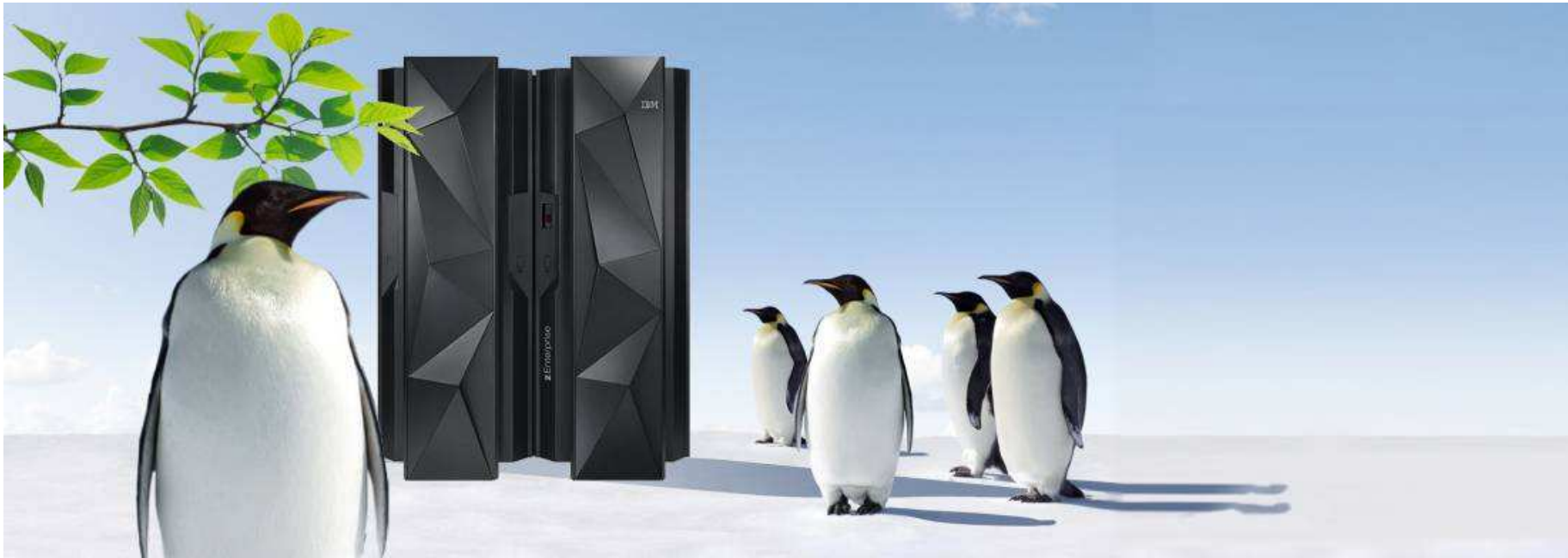
Linux on System z - Filesystems

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at Copyright and trademark information at www.ibm.com/legal/copytrade.shtml.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply. All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions. This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area. All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Agenda

- Linux Filesystems
- btrfs
- Local Filesystems
- Remote Filesystems

Linux Filesystems

	File Size	Volume Size	Journaling
btrfs	16 EB	16 EB	✔
ext2	2 TB	32 TB	
ext3	2 TB	32 TB	✔
ext4	16 TB	1 EB	✔
xfs	8 EB	8 EB	✔

btrfs - Introduction

- Name
  - B-Tree FS
  - Butter FS
  - Better FS

- Features
  - Copy-On-Write Filesystem
  - Based on binary trees
  - Better FS
  - Default FS for SLES 12
  - Improved scalability

btrfs - Features

- Volume size 2^{64} Bytes

- Online volume growth and shrinking

- Online check and defragmentation

- Snapshots (fast backup and rollback)

- Multi block device support

- Integrated RAID capability

- Subvolumes

- Compression

- In-place conversion from ext3/4

- Efficient with small files and directories

- Data deduplication

btrfs - Examples

Create

```
root> mkfs.btrfs /dev/sda1 -f
```

```
root> mkdir /mnt/disk
root> mount /dev/sda1 /mnt/disk
```

```
root> mount | grep sda1
/dev/sda1 on /mnt/disk type btrfs (rw,relatime,seclabel,
space_cache)
```

btrfs - Examples

Show information

```
root> btrfs filesystem show /dev/sda1
Label: none   uuid: 95f0d6e8-42ef-427a-be1e-e76e37084f2e
        Total devices 1 FS bytes used 192.00KiB
        devid    1 size 10.00GiB used 1.03GiB path /dev/sda1

Btrfs v3.12
```

btrfs - Examples

Convert ext4 to btrfs

```
root> btrfs-convert /dev/mapper/data-disk1
/dev/mapper/data-disk1 is mounted
```

```
root> umount /dev/mapper/data-disk1
root> btrfs-convert /dev/mapper/data-disk1
block size is too small
conversion aborted
```

Block size needs to be 4096

```
root> umount /dev/mapper/data-disk1
root> fsck.ext4 /dev/mapper/data-disk1
root> btrfs-convert /dev/mapper/data-disk1
creating btrfs metadata.
creating ext2fs image file.
cleaning up system chunk.
conversion complete.
root> mount /dev/mapper/data-disk1
```

btrfs - Examples

Convert back to ext4

```
root> btrfs-convert -r /dev/mapper/data-disk1
rollback complete.
```

btrfs - Examples

```
root> btrfs filesystem resize -200M /data/disk1
Resize '/mnt/test_btrfs' of '-200M'
ERROR: unable to resize '/mnt/test_btrfs'
```

```
root> btrfs filesystem resize -100M /data/disk1
Resize  /data/disk1  of  -100 M
root> btrfs filesystem resize -100M /data/disk1
Resize '/mnt/test_btrfs' of '-100M'
ERROR: unable to resize '/mnt/test_btrfs'
```

ext3/4 - Introduction

- Has been there from the beginning

- Evolutionary development to ext3/ext4

- Journaling

- ext3 and ext4 are compatible

- ext4

    - Unlimited number fo sub directories
    - Online defragmentation

- Scaling is regarded as a problem

ext3/4 - Example

For ext3 change check interval and max mount count

```
root> mke2fs -j /dev/mapper/data-disk1
root> tune2fs -c 0 -i /dev/mapper/data-disk1
```

xfs - Introduction

- Online defragmentation
- Online growing
- Online dump
- Default file system for Red Hat

SMB - Introduction

- SMB - Server Message Block
  - Originally designed at IBM
  - aka CIFS
- DOS/Windows compatible
- Used for lots of NAS servers

SMB - Issues

- Latency has impact on WAN performance
- Some security vulnerabilities

NFS - Introduction

- NFS v3
    - Already some TCP implementations
    - CacheFS in some implementations

- NFS v4
    - pNFS - parallel client access
    - Service side copy
   - TCP based

NFS - Issues

- Heavy performance loss with small packet loss when using UDP

- Stale file handle problems in case of connectivity interrupts
  - Issue in HA-Environments
  - Depending on implementation

- Security issues
  - Configuration specific
    - root squash
  - Depending on implementation

AFS - Introduction

- Andrew File System
  - Not a product anymore
  - Available as open source at http://www.openafs.org
- Single name space
- Write caching using callbacks
- Kerberos access control
- Online file server migration
- Scalability
  - One installation
    - 25000 clients
    - 50 sites
    - 6 continents

source: http://www-conf.slac.stanford.edu/AFSBestPractices/Slides/MorganStanley.pdf
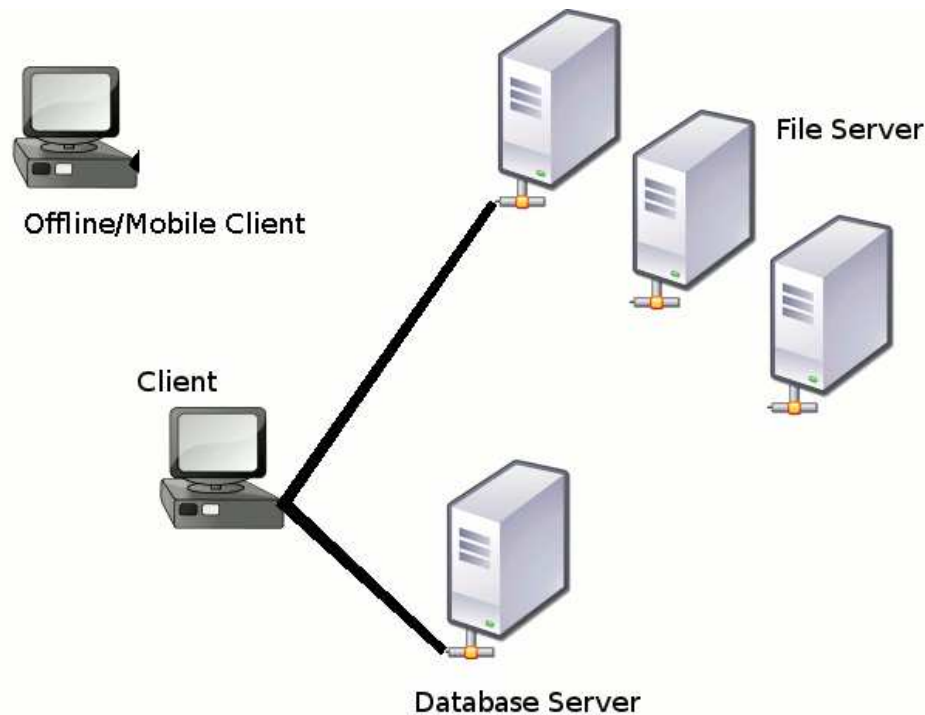
AFS - Volumes

- can be
    - replicated
    - moved between file servers
- Quota per Volume
- Create read-only volume for backup
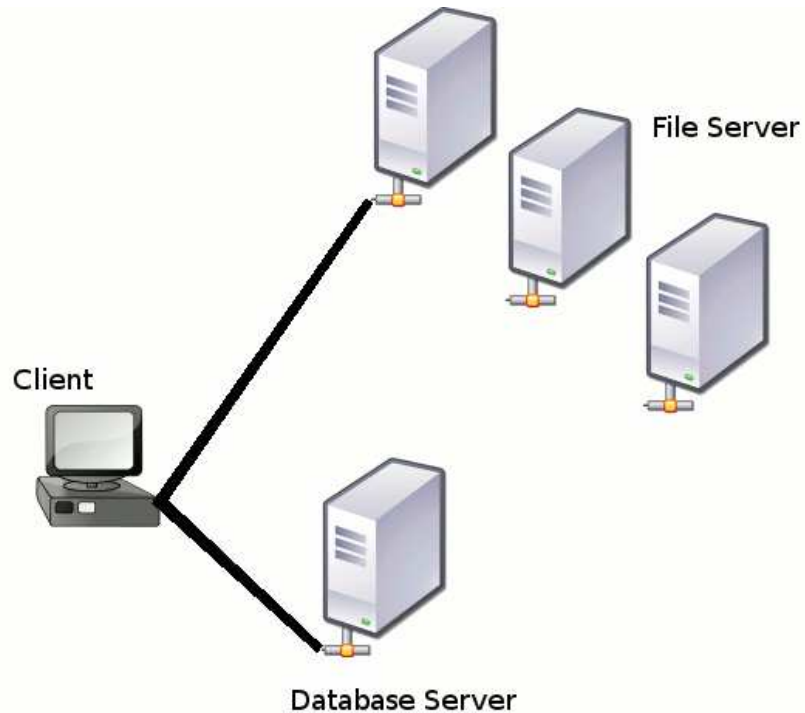
Coda - Introduction

- Developed out of AFS

- Features
  - Disconnected operation - hoarding
  - Continued operation during partial network failures

- Still experimental

Offline/Mobile Client

File Server

Client

Database Server

Links

- developerWorks

  `http://www.ibm.com/developerworks/library/l-linux-filesystem`

- Resources for Linux on System z

  `http://www-03.ibm.com/systems/z/os/linux/resources/index.html`

- IBM Redbooks

  `http://www.redbooks.ibm.com`

Questions ?

Dr. Stefan Reimbold
Diplom-Physiker

Linux on System z Service

Schoenaicher Strasse 220
D-71032 Boeblingen
Mail: Postfach 1380
D-71003 Boeblingen

Phone +49-7031-16-2368
Stefan.Reimbold@de.ibm.com