Stefan Haberland – Linux on System z development

17 July 2014

zdsfs -
Direct Linux access to
z/OS data sets

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

AIX*
BladeCenter*
DataPower*
DB2*
FICON*
GDPS*
HiperSockets
IBM*
IBM eServer
IBM (logo)*
InfiniBand*
Parallel Sysplex*
POWER*
POWER7*
PowerVM
PR/SM
Smarter Planet
System x*
System z*
System z9*
System z10
WebSphere*
z9*
z10 BC
z10 EC
zEnterprise
z/OS*
zSeries*
z/VM*
z/VSE

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries. Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Windows Server and the Windows logo are trademarks of the Microsoft group of countries.
InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective companies.

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.

Why you want to use zdsfs

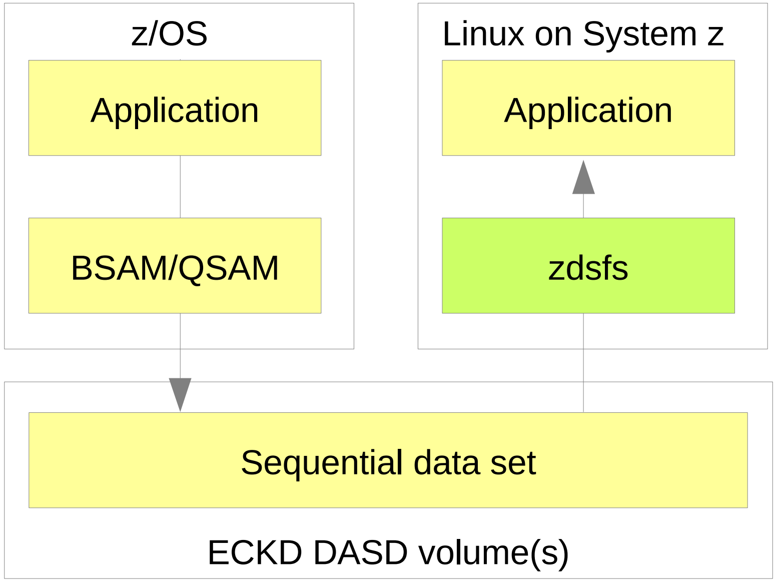
- **Lots of data stored and processed on z/OS**
  - Linux on System z is nearby
- **Significantly improve processing time for batch applications working with data generated in z/OS**
- **Easier to implement new applications and business processes on Linux**
  - larger community
  - a lot more pre build software
- **Reduce z/OS CPU cycles**
- **Extract Transfer Load (ETL) requires a lot CPU cycles**
  - sometimes more than warehousing itself
  - offload to Linux

Overview

- **Goal**
  - Transfer bulk data from z/OS to Linux on System z
  - Faster than networked transfer (e.g. FTP, NFS)
  - Use less CPU cycles than networked transfer

- **NOT intended for CONCURRENT access**
  - Not a cluster file system

- **Approach**
  - Read records from DASD volumes
  - Translate into Linux file system semantics
    - Physical Sequential data set → File
    - Partitioned data set → Directory containing members as files



Linux disk layout

- **Block devices → directly addressable sectors of fixed size**
- **Smallest unit is one sector**
  - size is a power of 2 → usually between 512 and 4096 bytes
  - for current Linux systems the upper limit is the memory page size
- **Applications mostly do not use block devices directly**
  - use of a filesystem

DASD and z/OS

- **Direct Access Storage Device – DASD**
  - matrix of tracks, addressable via cylinder and head number
  - within each track the OS or application can store records of an arbitrary size
  - results in an variable number of records per track
  - each record has count, key and data field – CKD
  - today Extended Count Key Data - ECKD is in use

- **z/OS makes full use of the flexibility of ECKD DASD**
  - Data set consists of one ore more extents
    - extents are areas of consecutive tracks on a DASD
  - each data set can have records of variable size
  - the Volume Table of Content (VTOC) describes each data set, its extents and parameters
  - one data set can be distributed over several DASD devices

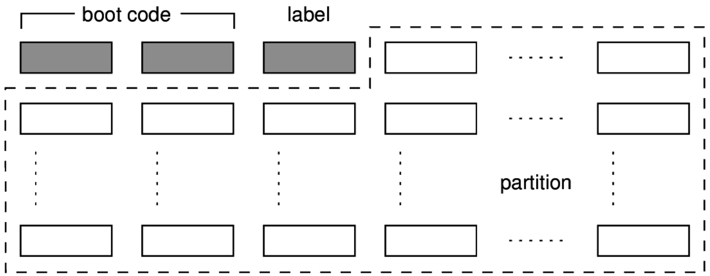
Linux on System z use of ECKD DASD

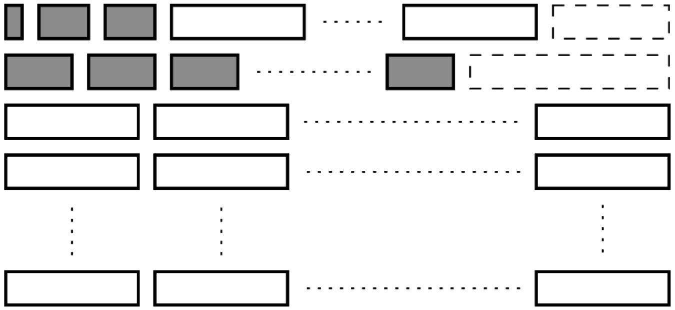
Linux disk Layout – LDL

- Similar to Linux disks
- Fixed block size
- Not usable by other System z OS

Compatible disk Layout – CDL

- Different record size in first tracks
- Fixed block size for data tracks
- Has VTOC and is readable by z/OS
  - – partitions are seen as a dataset





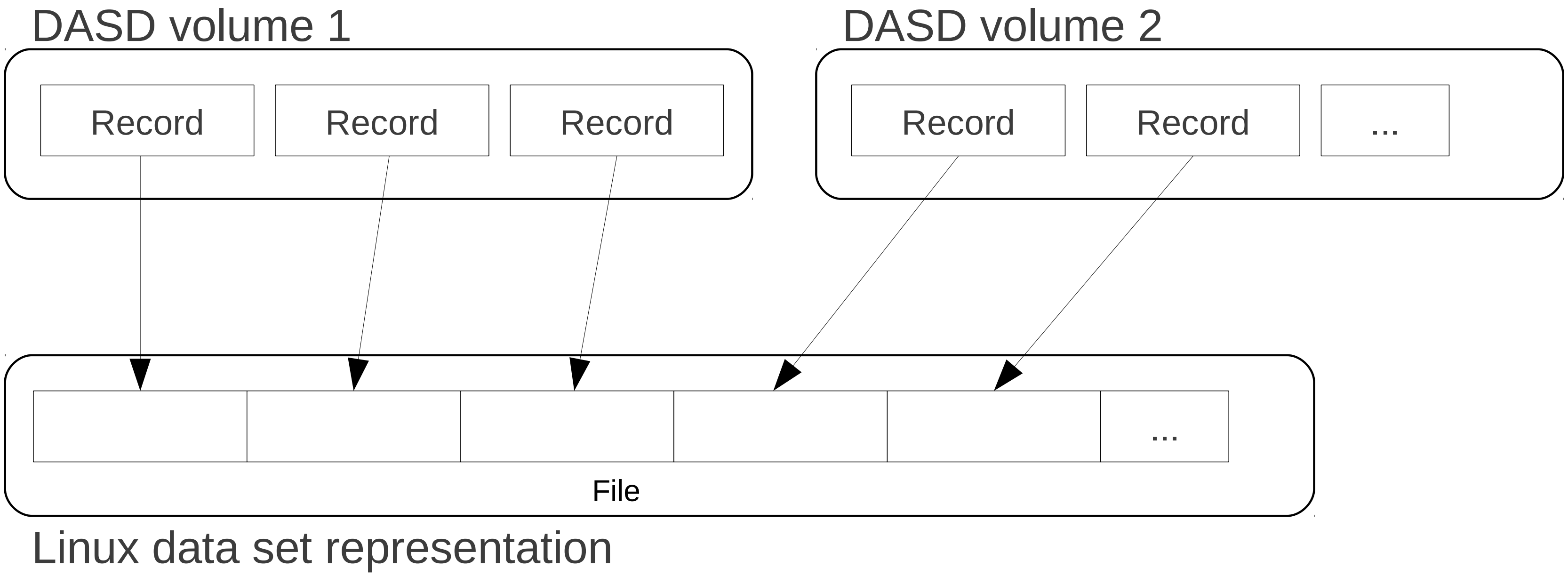
Linux on System z feature - raw track access

- **Read and write full track images including Count and Key values**
- **Track has a well known size → 58786 byte for a DASD of type 3390**
- **Disk is mapped to a sequence of tracks of a fixed size**
- **Page size for Linux on System z is 4096 byte**
  - block size is 4096 byte, too
  - need to map a track to 16 separate 4096 byte blocks
  - but not directly accessible
- **Accessing such a device like a normal block device will most likely end up in I/O errors**
  - need to use DIRECT_IO to bypass most of the block layer optimizations
  - track alignment
- **But this still gives only raw data and no understanding of the low level data formats like:**
  - ECKD track layout
  - VTOC layout
  - z/OS data set layout

zdsfs

- **Filesystem in Userspace (FUSE)**
  - – enables user space filesystems
- **Supported data sets:**
  - – physical sequential data sets (PS)
  - – partitioned data sets (PDS)
- **Other data set formats like VSAM or extended format data sets are not supported**
- **Limited to basic operations:**
  - – readdir
  - – stat
  - – open
  - – read
  - – seek
- **Optimized for sequential read access**
  - – random access possible but with performance impact
- **Option to include record descriptor words in data stream**
- **PS data sets → simple files**
- **PDS data sets → directories with members as files**
- **One or more DASD devices possible**

Record mapping

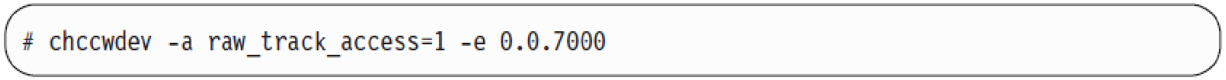


Limitations

- **Data set format restrictions**
  - No VSAM
  - No Extended-Format data sets
- **Access not controlled by z/OS authorization mechanisms**
  - Use of DASDs dedicated to data transfer recommended
  - Linux authorization mechanisms apply
- **Access not logged by z/OS auditing mechanism**
  - Users must consider “z/OS write to DASD = read by Linux”
- **No catalog access**
  - Users need to specify the DASDs on which a data set is located
- **Complex usage**
- **One-way data transfer only**
- **File size (total data set size) only approximated (number and size of extends)**

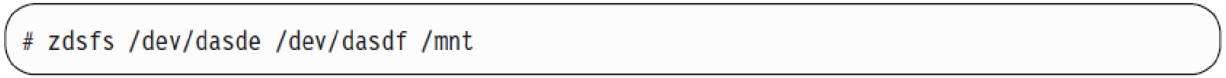
Usage

- z/OS: Write data set to DASDs
- z/OS: Set DASDs offline to ensure consistent on-disk state of data set
- Linux: Set DASD online in raw-track-access-mode



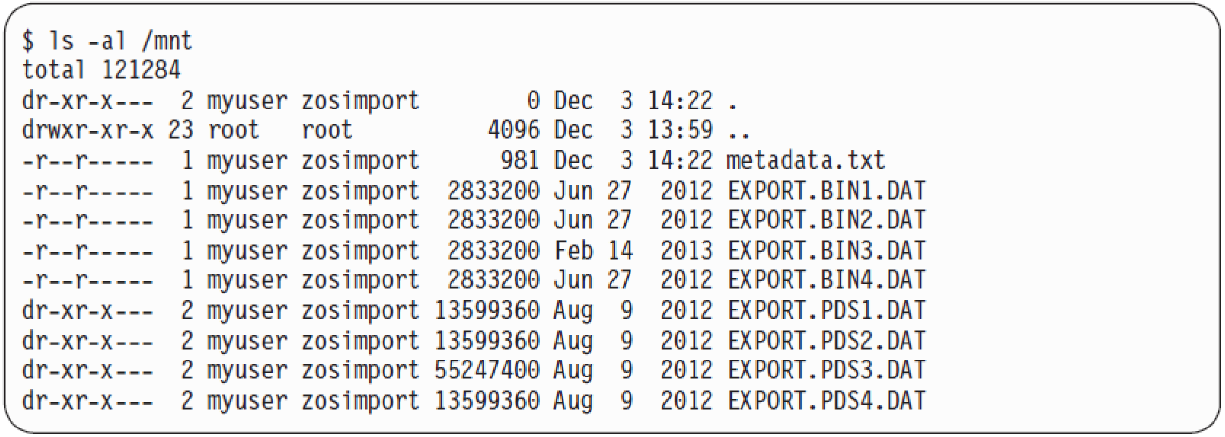
```
# chccwdev -a raw_track_access=1 -e 0.0.7000
```

- Linux: Run zdsfs to “mount” the data set



```
# zdsfs /dev/dasde /dev/dasdf /mnt
```

- Linux: Access data set



```
$ ls -al /mnt
total 121284
dr-xr-x---  2 myuser zosimport        0 Dec  3 14:22 .
drwxr-xr-x 23 root   root          4096 Dec  3 13:59 ..
-r--r-----  1 myuser zosimport      981 Dec  3 14:22 metadata.txt
-r--r-----  1 myuser zosimport  2833200 Jun 27  2012 EXPORT.BIN1.DAT
-r--r-----  1 myuser zosimport  2833200 Jun 27  2012 EXPORT.BIN2.DAT
-r--r-----  1 myuser zosimport  2833200 Feb 14  2013 EXPORT.BIN3.DAT
-r--r-----  1 myuser zosimport  2833200 Jun 27  2012 EXPORT.BIN4.DAT
dr-xr-x---  2 myuser zosimport 13599360 Aug  9  2012 EXPORT.PDS1.DAT
dr-xr-x---  2 myuser zosimport 13599360 Aug  9  2012 EXPORT.PDS2.DAT
dr-xr-x---  2 myuser zosimport 55247400 Aug  9  2012 EXPORT.PDS3.DAT
dr-xr-x---  2 myuser zosimport 13599360 Aug  9  2012 EXPORT.PDS4.DAT
```

zdsfs options

- **-o ignore_incomplete**
  - Represents all complete data sets in the file system, even if there are incomplete data sets
- **-o rdw**
  - Keeps record descriptor words (RDWs) of data sets that are stored by using the z/OS concept of variable record lengths
- **-o tracks=<n>**
  - Specifies the track buffer size in tracks
  - Increasing the track buffer size might improve your system performance
- **-o seekbuffer=<s>**
  - Sets the maximum seek history buffer size in bytes
  - Speed up the performance of a seek operation

Dataset Meta data

Static meta data provided in two ways:

- **File “metadata.txt” in top level of mounted directory:**

```
# cat metadata.txt
dsn=WEIN.TEST2.TXT,recfm=FB,lrecl=80,dsorg=PS
dsn=WEIN.WEIN.DASDECKD.C,recfm=F,lrecl=100,dsorg=PS
dsn=WEIN.DASDECKD.C,recfm=F,lrecl=100,dsorg=PS
dsn=WEIN.DASDECK2.C,recfm=F,lrecl=100,dsorg=PS
```

  – advantage: can be copied along with the data sets

- **Via extended file attributes:**

```
# getfattr -d WEIN.DASDECKD.C
# file: WEIN.DASDECKD.C
user.dsorg="PS"
user.lrecl="100"
user.recfm="F"
```

  – advantage: generic tools and APIs available

Attention

Set devices in z/OS offline before mounting them in Linux.

Through zdsfs file system the whole DASD is accessible in Linux but the access is not controlled by z/OS auditing mechanisms.

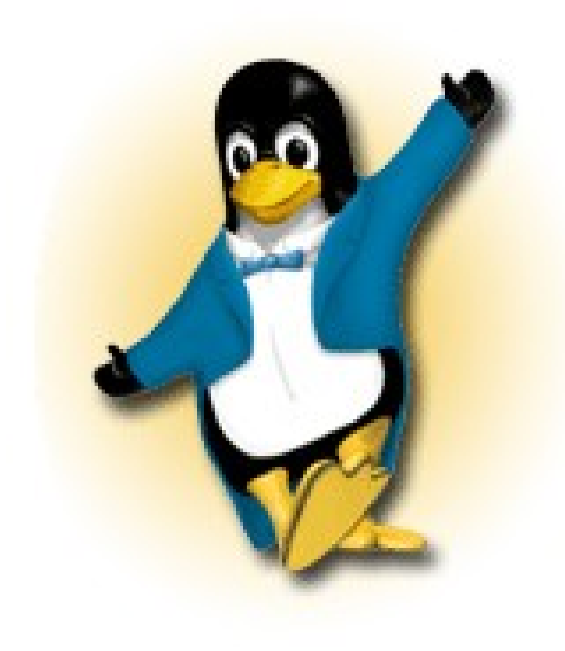
To avoid security problems the disk may be dedicated in z/OS only for providing data to Linux.

Further reading

- **Device Drivers, Features, and Commands (Kernel 3.12) - SC33-8411-23**

  http://public.dhe.ibm.com/software/dw/linux390/docu/l312dd23.pdf

Questions?



Stefan Haberland

Linux on System z
Development

Schönaicher Strasse 220
71032 Böblingen, Germany

Phone +49 (0)7031-16-1760
stefan.haberland@de.ibm.com