Oracle DB erfolgreich betreiben auf SPARC/LDoms/Solaris/ZFS

Marcel Hofstetter

hofstetter@jomasoft.ch
https://jomasoftmarcel.blogspot.ch

CEO / Enterprise Consultant
JomaSoft GmbH

Oracle ACE „Solaris“

Agenda

- Wer ist JomaSoft?
- SPARC und LDoms Technologie
- ZFS
- Oracle DB Cloning mit ZFS
- INMEMORY / DAX

Wer ist JomaSoft?

ORACLE Gold Partner
Specialized
Oracle Solaris 11

ORACLE Gold Partner
Specialized
SPARC T-Series Servers

- Software Unternehmen gegründet im Juli 2000
- Spezialisiert im Bereich **Solaris,**
  Software Entwicklung & Services/Beratung
- Produkt **VDCF** (Virtual Datacenter Cloud Framework):
  Installation, Management, Betrieb, Monitoring, Security und DR von Solaris 10/11, sowie Virtualisierung mittels LDoms und Solaris Zonen
- VDCF wird seit 2006 produktiv in Europa genutzt

Love Work Play Mobistar

bmvit
Bundesministerium für Verkehr, Innovation und Technologie

Zürcher Kantonalbank

swisscom

NOTENSTEIN PRIVATBANK

LOTERIE ROMANDE

VOLKSBANK LIECHTENSTEIN

Marcel Hofstetter

Informatiker seit 25+ Jahren
Solaris seit 20 Jahren
CEO bei der JomaSoft GmbH seit 18 Jahren

Internationaler Speaker:
Oracle OpenWorld, DOAG, UKOUG, SOUG, AOUG

Oracle ACE „Solaris“

SOUG (Swiss Oracle User Group) – Speaker of the Year 2016

Hobby: Familie, Reisen, Wine & Dine, Kino

https://www.linkedin.com/in/marcelhofstetter

https://twitter.com/marcel_jomasoft

https://jomasoftmarcel.blogspot.ch

Oracle SPARC CPUs

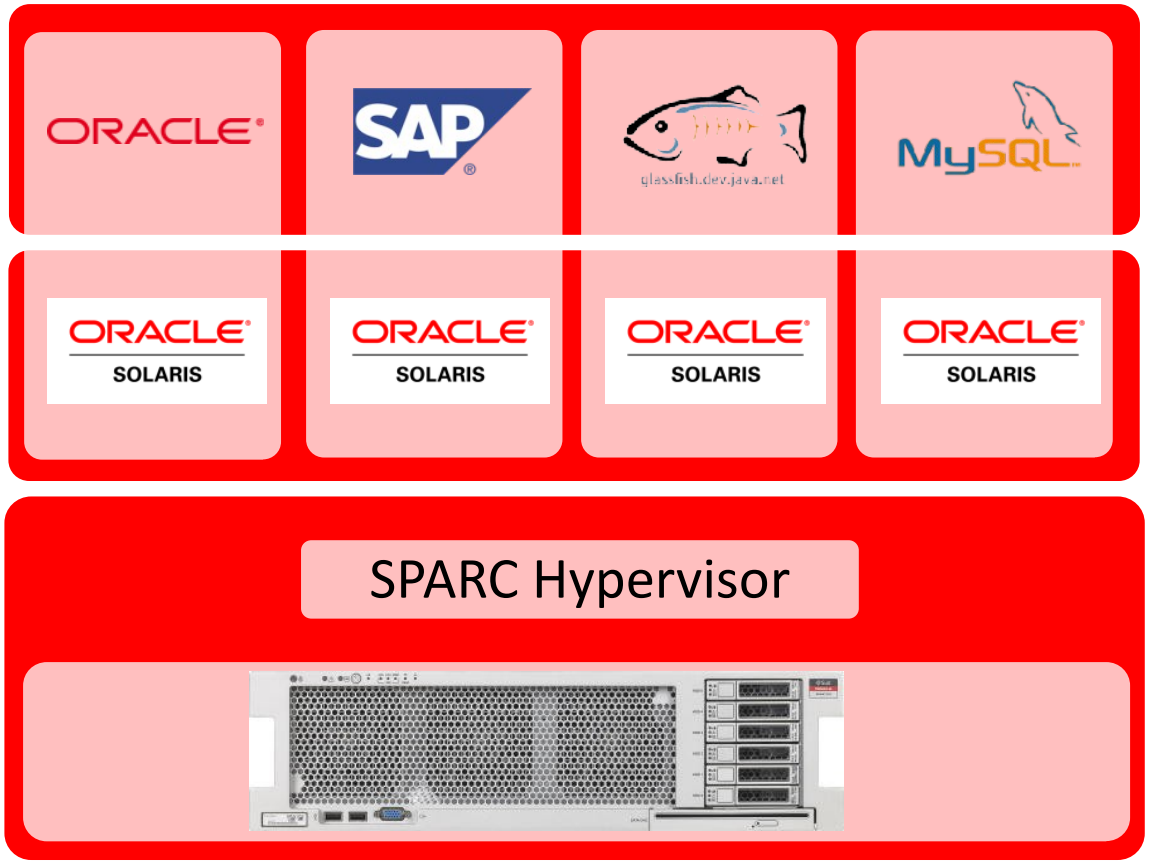
	SPARC M8 (2017)	SPARC S7 (2016)	SPARC M7 (2015)	SPARC T5 (2013)
Processor Cores	32 (5th Gen)	8 (4th Gen)	32 (4th Gen)	16 (3rd Gen)
Cache per Core	2 MB	2 MB	2 MB	0.5 MB
Memory Bandwidth per Core	5.6 GB/sec	6.0 GB/sec	5.3 GB/sec	5.0 GB/sec
Memory Access	127ns	97ns	131ns	163ns
I/O Bandwidth	145 GB/sec	32 GB/sec	145 GB/sec	32 GB/sec
CPU Frequency	5.0 GHz	4.27 GHz	4.13 GHz	3.6 GHz

Oracle VM Server for SPARC (LDoms)

Isoliertes Betriebssystem und Applikationen in jeder Logical Domain

Firmware basierter Hypervisor

Jede Logical Domain läuft mit dediziertem Memory und CPU Threads
→ Zero Overhead

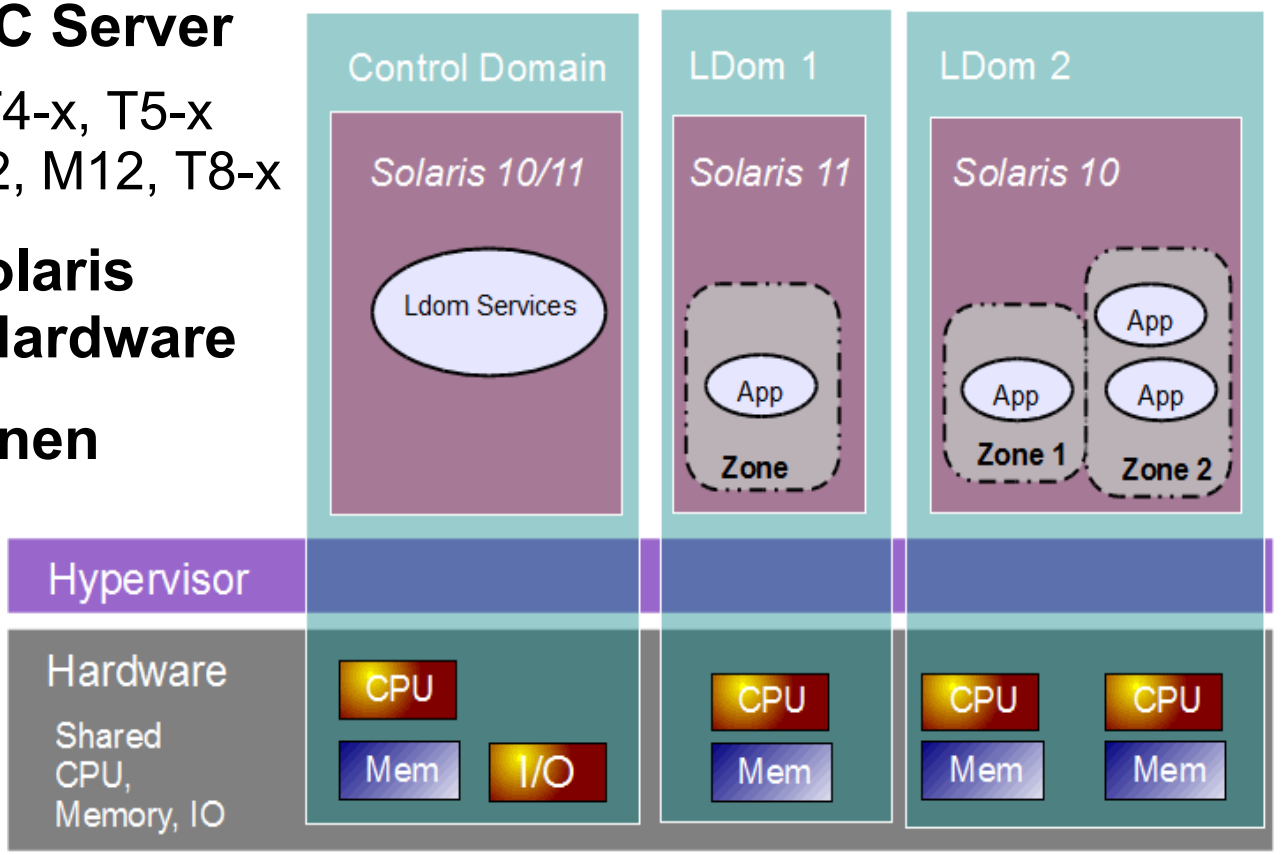


Logical Domains (LDoms)

- **Oracle/Fujitsu SPARC Server**

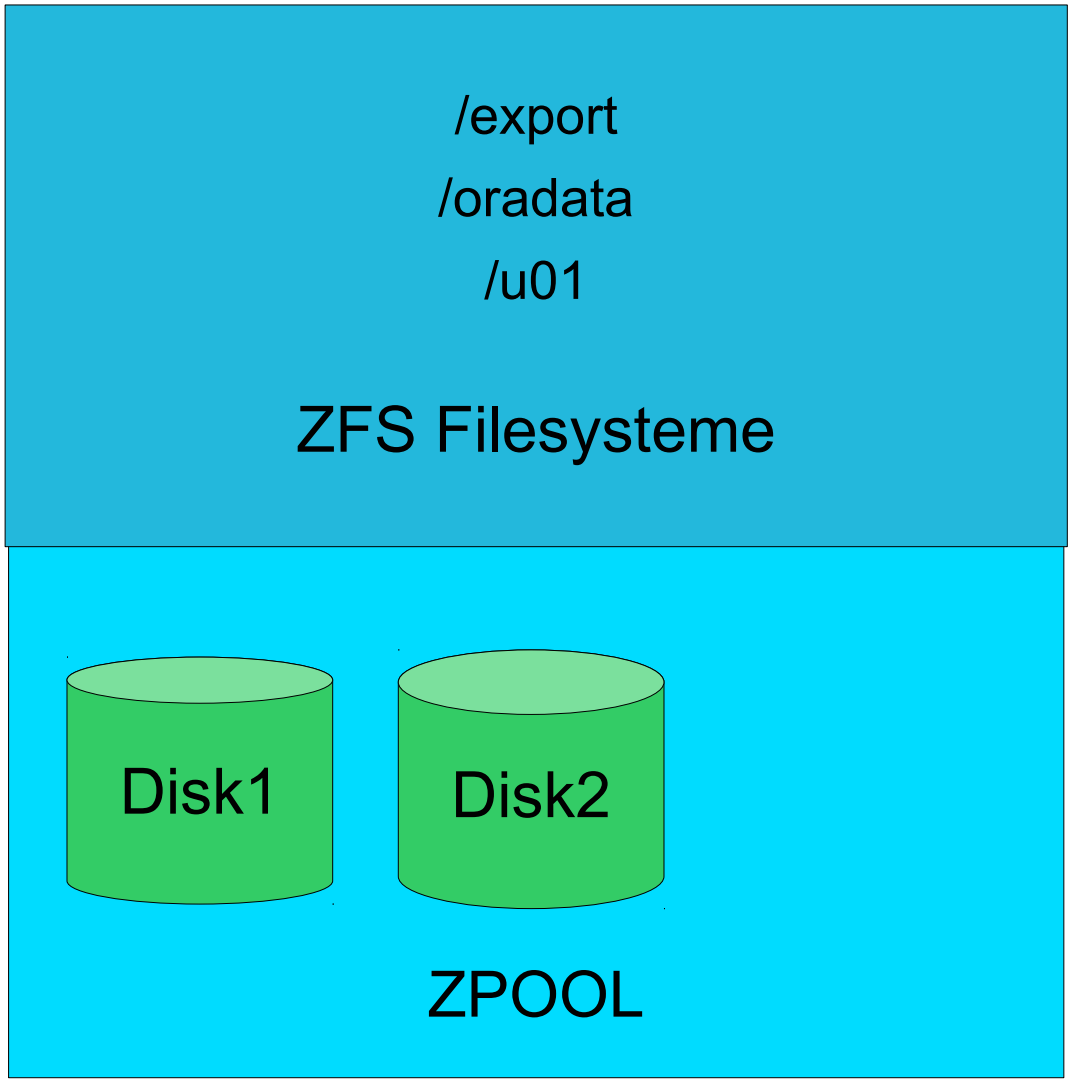
  Systeme: T5xx0, T3-x, T4-x, T5-x
  M5, M6, M10, T7-x, S7-2, M12, T8-x

- **Mehrere, separate Solaris Instanzen auf einer Hardware**
- **Kombinierbar mit Zonen**
- **Live Migration (auf andere Hardware ohne Unterbruch)**



- **LDoms & Zonen sind Hard Partitions für Oracle DB Lizenzierung**

Solaris ZFS



- Einfache Bedienung
- Flexibel
- Filesystem Grösse optional
- Snapshots & Clones
- COW / Kein fscheck

- Stripe, Mirror, RAID
- Disk hinzu → Grösser
- Export & Import

- Seit Solaris 11.4 auch Disk entfernen

NEW

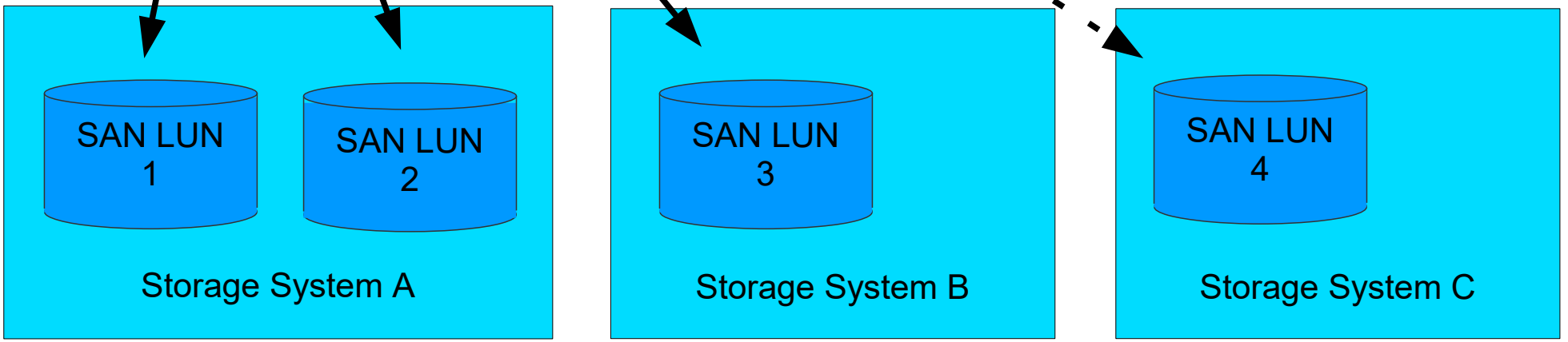
Vorteile von ZFS Host-Based Mirroring

+ Performance-Vorteil (Read)

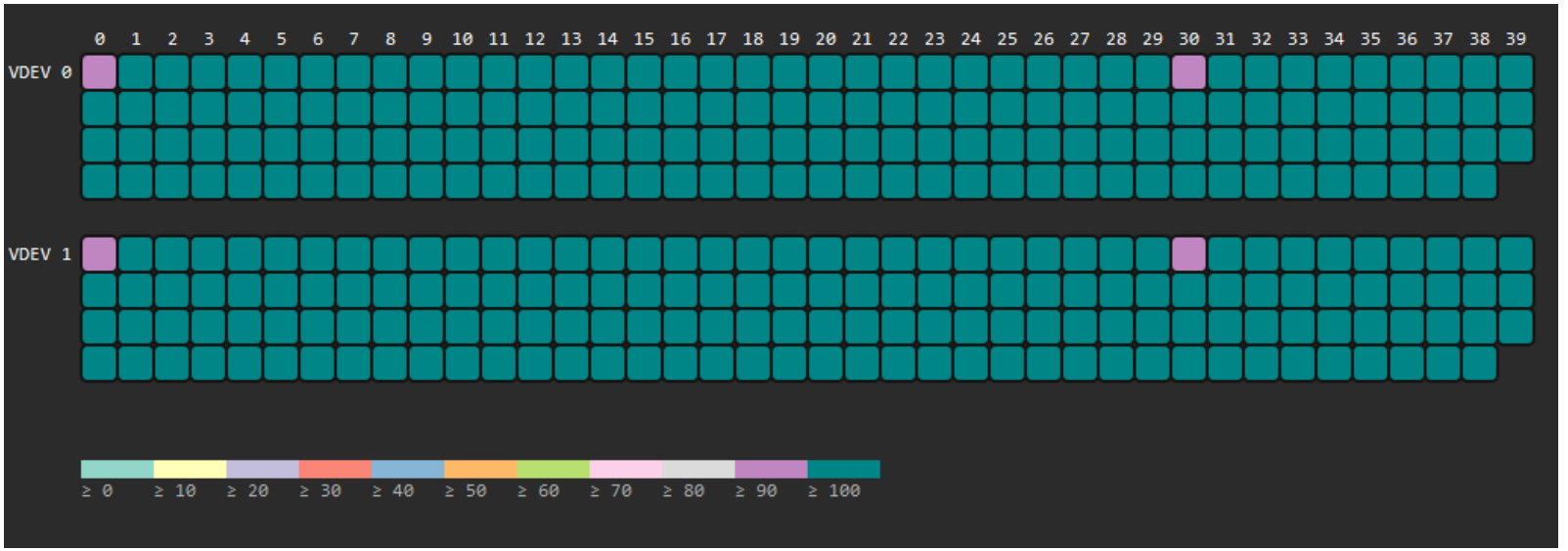
+ ZFS kann inkonsistente Daten korrigieren (Checksums)

+ Migration auf neuen Storage während Betrieb

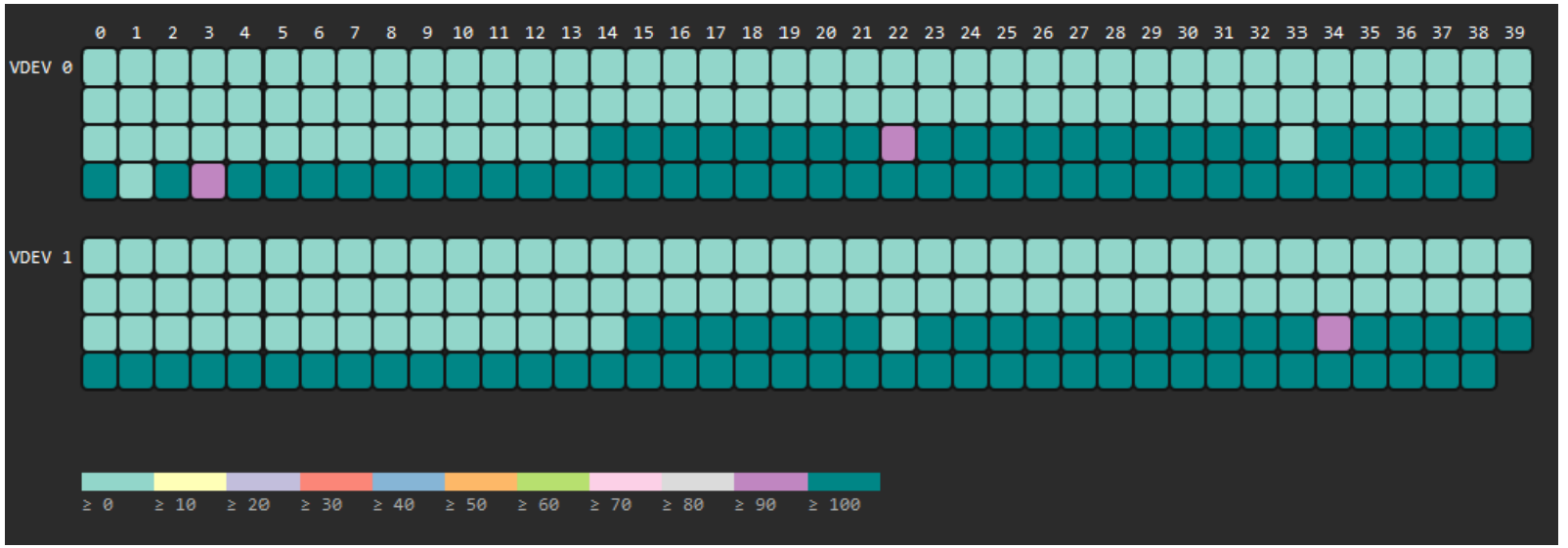
- Benötigt zusätzlichen Diskplatz



ZFS - Fragmentation



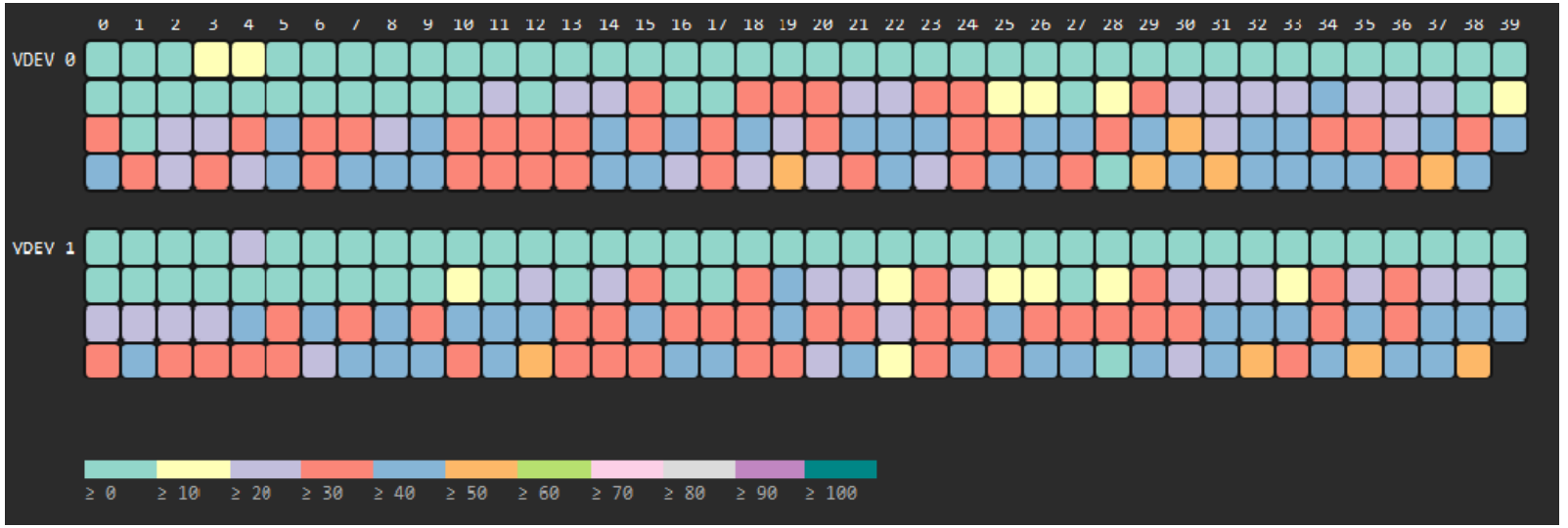
Leerer ZPOOL



Wird gefüllt

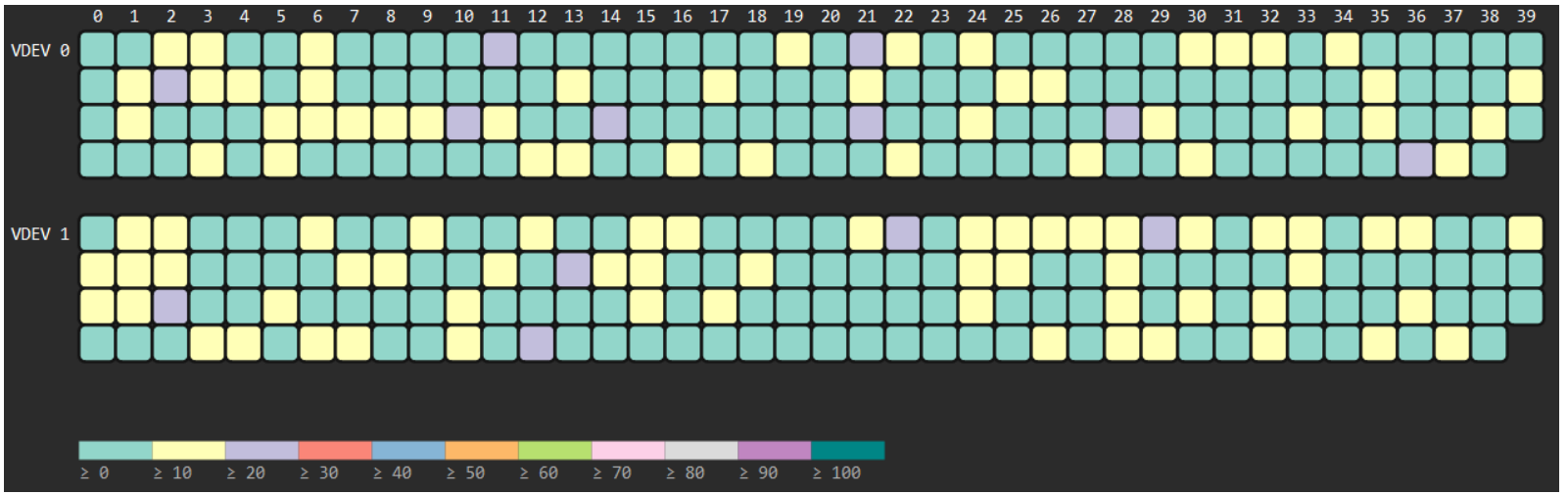
Hier bei 60%

ZFS - Fragmentation



Nach ein paar delete

76% voll



Später

90% voll

fragmentiert

ZFS - Fragmentation

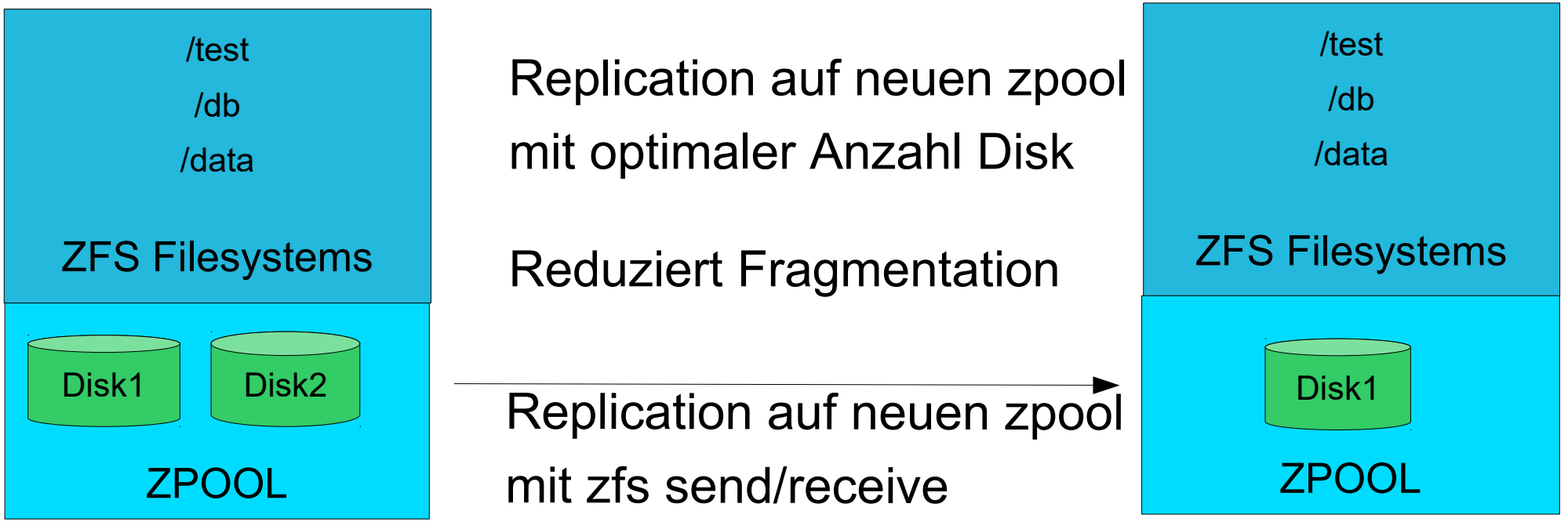
Bei sehr grossen ZPOOLS mit hohem Füllgrad wird write sehr langsam, weil freie Blocks gesucht werden müssen

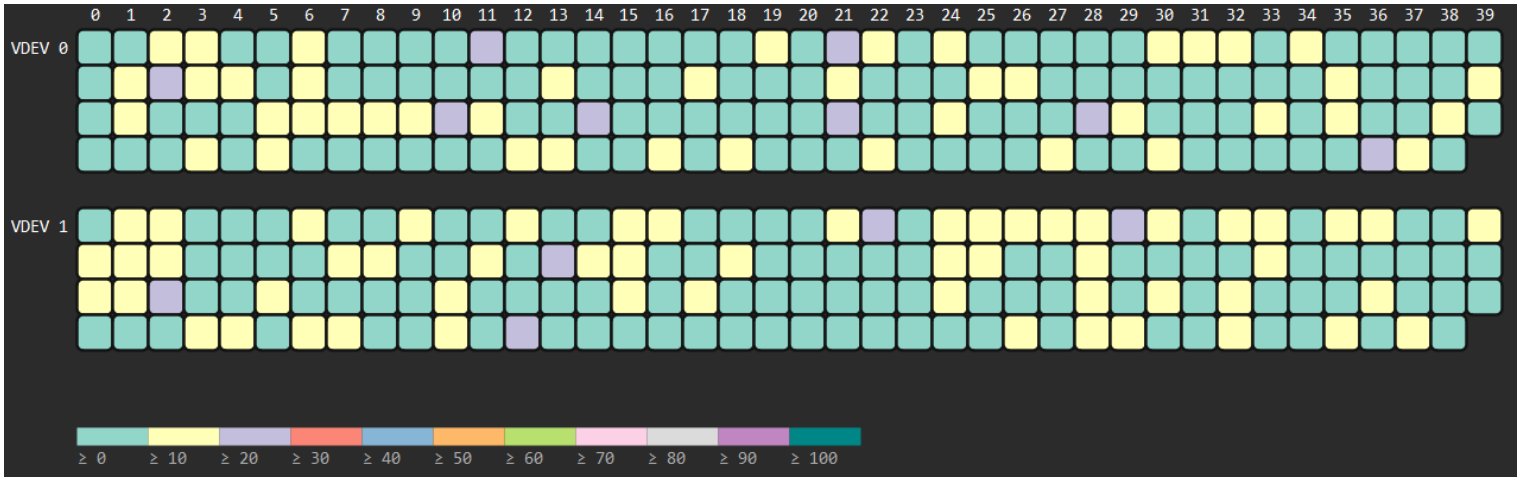
Wie entschärfen?

- Aktuelle Solaris Versionen einsetzen mit besseren Algorithmen
- ZPOOL Log Disk anhängen und logbias=latency
- Wenn möglich Datenbank export
- Daten replizieren auf neuen ZPOOL (nicht spiegeln!)

- (ZPOOL Füllgrad unter 80% halten)

VDCF - ZPOOL Online Replication



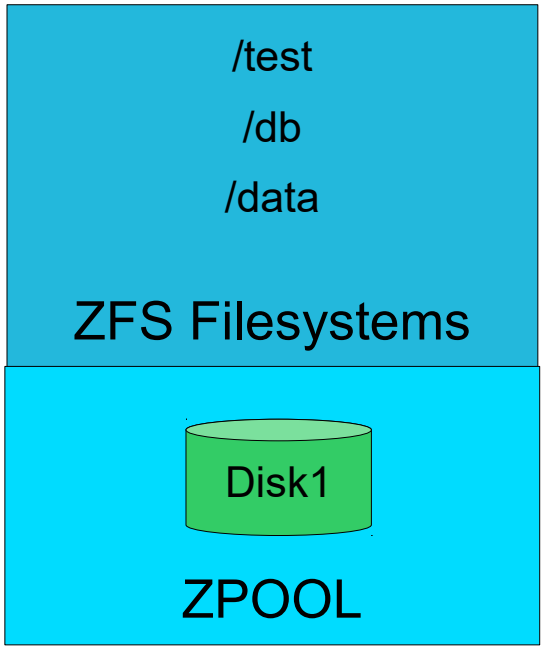


VDCF - ZPOOL Online Replication

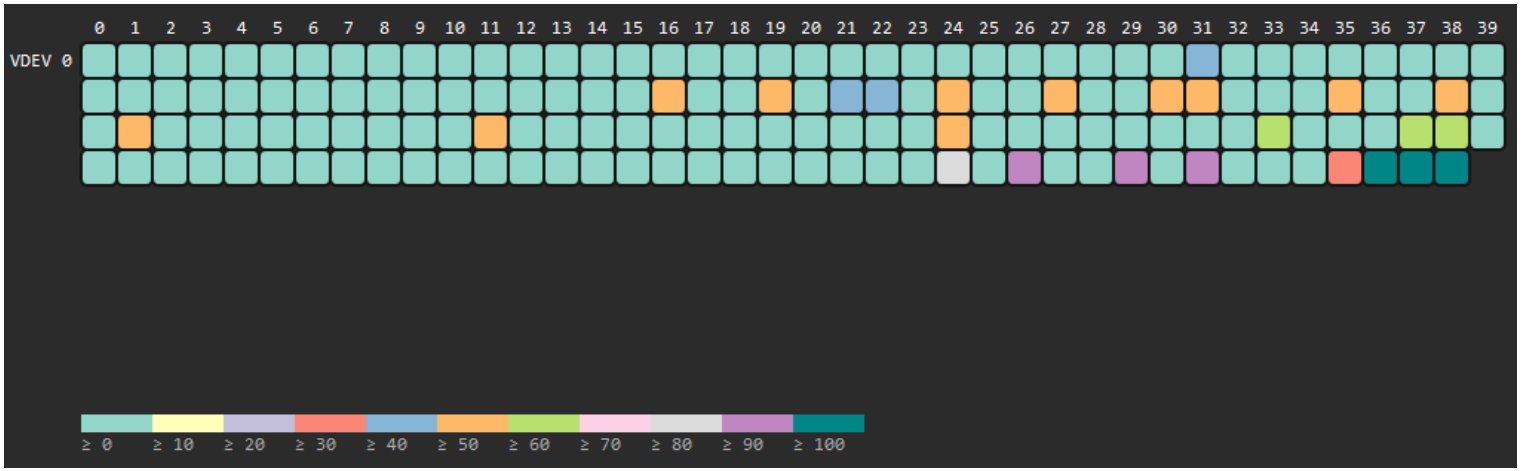
Daten online repliziert

Wiederholung für Delta/Änderungen

Kurze Downtime für Remount



Resultat nach der Online Replication



ZFS - Performance Empfehlungen

Genügend Memory für Applikationen und ZFS Cache (Read)

Genügend CPUs in der Global Zone für I/O

Mehrere Disk/LUN verwenden um vom Striping zu profitieren

Bei ZFS auf SAN mit vielen physischen Disks: zfs_vdev_max_pending (Default 10) erhöhen

Separate, schnelle ZPOOL Log Disk (Write)

Oracle DB Cloning mit ZFS

Ziel: Einsparen von Disk Space

Speziell natürlich bei sehr grossen Datenbanken: X TB

Clone DB erstellen soll schneller gehen
- Kurze Downtime der Quell DB
- In ein paar Minuten automatisiert erledigt

Konkrete Einsparungen bei einem JomaSoft Kunden
10 Test-Datenbanken: Anstatt 10 x 3 TB nur ca 4 TB.

Memory und CPU wird trotzdem pro Clone DB benötigt

Oracle DB Cloning mit ZFS

DBA Tasks

Solaris Admin Task

Ablauf

Vorbereitung

Control File Statements der Source DB in Trace speichern
& Anpassen für Clone DB

Shutdown der App & Source DB

ZFS Filesystem Cloning mit VDCF
1. ZFS Snapshot (ReadOnly)
2. Read/Write ZFS Clone basierend auf ZFS Snapshot
3. ZFS Clone mounten

Start der Source DB & App

Erstellen Control File für Clone DB und Start

Allenfalls noch Anpassungen an den Daten ...

Oracle DB Cloning mit ZFS

Anpassungen im Control File SQL
Set #2. RESETLOGS case

```
STARTUP NOMOUNT
CREATE CONTROLFILE SET DATABASE "SLOB" RESETLOGS NOARCHIVELOG
MAXLOGFILES 16
MAXLOGMEMBERS 2
MAXDATAFILES 1024
MAXINSTANCES 1
MAXLOGHISTORY 292
LOGFILE
GROUP 1 '/data/SLOB/onlinelog/o1_mf_1_fxzwszfw_.log' SIZE 4096M BLOCKSIZE 512,
GROUP 2 '/data/SLOB/onlinelog/o1_mf_2_fxzwtg7k_.log' SIZE 4096M BLOCKSIZE 512,
GROUP 3 '/data/SLOB/onlinelog/o1_mf_3_fxzwtx72_.log' SIZE 4096M BLOCKSIZE 512
-- STANDBY LOGFILE
DATAFILE
'/data/SLOB/datafile/o1_mf_system_fxzwvdct_.dbf',
'/data/SLOB/datafile/o1_mf_sysaux_fxzwvg5p_.dbf',
'/data/SLOB/datafile/o1_mf_sys_undo_fxzwvgsj_.dbf',
'/data/SLOB/datafile/o1_mf_iops_fxzxyzxt_.dbf'
CHARACTER SET US7ASCII

~~RECOVER DATABASE USING BACKUP CONTROLFILE~~
```

Oracle DB Cloning mit ZFS

Automatisiertes Cloning mit VDCF (1/2)

```
-bash-4.4$ ./clone_db
Executing on vServer v0133 (Oracle 18c Demo) on Node g0059
Oracle Corporation      SunOS 5.11      11.4    September 2018
Executing command as oracle: /oracle/prepare_clone_cntl_file SLOB SLOB3
Trace File with controlfile found
/u01/app/oracle/diag/rdbms/slob/SLOB/trace/SLOB_ora_24904.trc
ControlFile: REUSE -> SET
ControlFile: Ignore RECOVER DATABASE
ControlFile Script created for SLOB3: /data/SLOB/SLOB3_newctl.sql

Executing on vServer v0133 (Oracle 18c Demo) on Node g0059
Executing command as oracle: /oracle/shutdown_source SLOB

Cloning filesystem /data/SLOB on vServer v0133
ZFS snapshot created for vServer v0133
Filesystem </data/SLOB3> mounted on vServer v0133
/data/SLOB of vServer v0133 successfully cloned and mounted as /data/SLOB3 on vServer v0133
```

Oracle DB Cloning mit ZFS

Automatisiertes Cloning mit VDCF (2/2)

```
Executing on vServer v0133 (Oracle 18c Demo) on Node g0059
Executing command as oracle: /oracle/start_source SLOB
Executing command as oracle: /oracle/start_clone SLOB3
Copyright (c) 1982, 2018, Oracle.  All rights reserved.
Connected to an idle instance.
ORACLE instance started.
Total System Global Area 3254772544 bytes
Fixed Size                  8643392 bytes
Variable Size            1090519040 bytes
Database Buffers         2147483648 bytes
Redo Buffers                8126464 bytes
Control file created.
Database altered.
Disconnected from Oracle Database 18c Enterprise Edition Release
18.0.0.0.0 - Production
Version 18.3.0.0.0

Source DB was down 35 seconds
Cloning Duration was 136 seconds
```

Oracle DB Cloning mit ZFS

Platzbedarf

Auf dem System

```
-bash-4.4$ df -h | egrep "Filesys|SLOB"
Filesystem             Size   Used  Available Capacity  Mounted on
/data/SLOB              41G    23G         18G    57%    /data/SLOB
/data/SLOB3             41G    23G         18G    57%    /data/SLOB3
```

Effektiv

```
# zfs list -t all | egrep "NAME|SLOB"
NAME                                                  USED  AVAIL  REFER  MOUNTPOINT
SLOB                                                  31.2G  17.6G    31K  legacy
SLOB/data                                             31.2G  17.6G    31K  legacy
SLOB/data/data_SLOB                                   23.2G  17.6G  23.2G  legacy
SLOB/data/data_SLOB@20181106-170922                   14.3M      -  23.2G  -
SLOB/data/data_SLOB3                                  8.06G  17.6G  23.2G  legacy
```

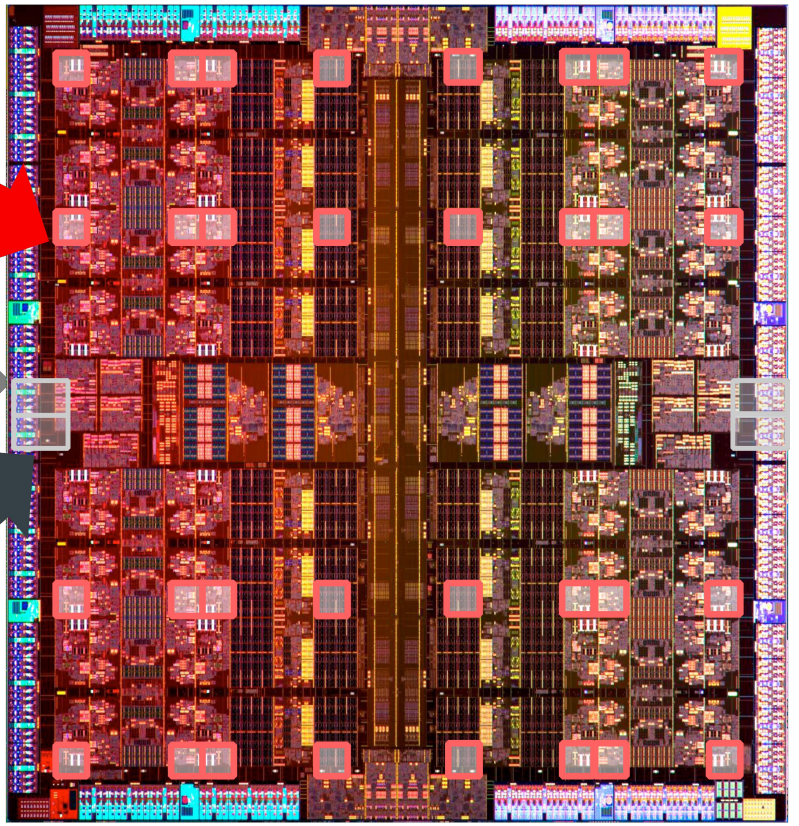
INMEMORY / DAX

Oracle SPARC S7,M7 und M8
DAX → Oracle Data Analytics Accelerator

Security in Silicon:
Silicon Secured Memory
Cryptography Acceleration

SQL in Silicon:
Database In Memory Accelerator Engines

Capacity in Silicon:
Decompression Engines



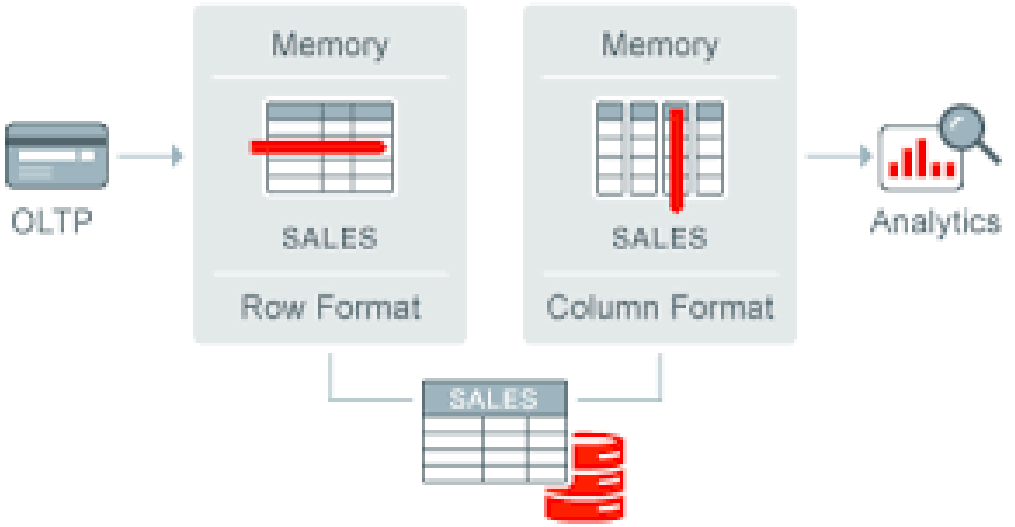
INMEMORY / DAX

Test Setup mit SLOB

```
SQL> show parameter inmemory_size
NAME                                 TYPE        VALUE
------------------------------------ ----------- ------------------------------
inmemory_size                        big integer 1G


SQL> ALTER TABLE USER1.CF1 INMEMORY;
Table altered.


SQL> select count(*) from USER1.CF1;
  COUNT(*)
----------
     10000
```



INMEMORY / DAX

Resultat / 8 Reader / 1 x SPARC S7-core

awr_0w_8r.20181107_165153.txt

DB Name	DB Id	Unique Name	DB Role	Edition	Release	RAC	CDB
SLOB	3718155087	SLOB	PRIMARY	EE	18.0.0.0.0	NO	NO

Host Name	Platform	CPUs	Cores	Sockets	Memory(GB)
v0133	Solaris[tm] OE (64-bit)	**8**	**1**	1	16.00

	Snap Id	Snap Time	Sessions	Curs/Sess
Begin Snap:	105	07-Nov-18 16:46:32	44	1.3
End Snap:	106	07-Nov-18 16:51:51	44	1.3
Elapsed:		5.31 (mins)		
DB Time:		42.12 (mins)		

Load Profile	Per Second	Per Transaction	Per Exec	Per Call
DB Time(s):	7.9	126.4	0.00	8.15
DB CPU(s):	7.9	125.4	0.00	8.09
Background CPU(s):	0.0	0.5	0.00	0.00
Redo size (bytes):	8,690.5	138,454.2		
Logical read (blocks):	**125,562,125.4**	**2,000,411,835.5**		
Block changes:	42.8	681.1		
Physical read (blocks):	0.6	10.1		
Physical write (blocks):	3.2	51.2		
Read IO requests:	0.3	5.2		
Write IO requests:	1.4	22.8		
Read IO (MB):	0.0	0.1		
Write IO (MB):	0.0	0.4		
IM scan rows:	125,536,275.3	2,000,000,000.0		
Session Logical Read IM:	125,536,275.3	2,000,000,000.0		
User calls:	1.0	15.5		

```
-bash-4.4$ grep offload awr_0w_8r.20181107_165153.txt
Statistic                                     Total     per Second     per Trans
IM simd compare HW offload calls          4,000,000       12,553.6     200,000.0
IM simd decode unpack HW offload          4,000,000       12,553.6     200,000.0
```

INMEMORY / DAX

Spannend, dass die 1 core LDOM alle 4 DAX Units des SPARC S7 Socket verwenden kann

```
Host Name         Platform                          CPUs Cores Sockets Memory(GB)
---------------- ---------------------------------- ---- ----- ------- ----------
v0133            Solaris[tm] OE (64-bit)             8     1       1      16.00
```

```
-bash-4.4$ daxstat 10
DAX    commands fallbacks    input     output %busy
  4       63809         0   106.1M      5.4M     0
  5       63810         0   106.1M      5.5M     0
  6       63810         0   106.1M      5.4M     0
  7       63803         0   106.1M      5.5M     0
```

SLOB Performance auf SPARC M8

```
1 CHIP SPARC M8 / 32 Cores / 256 Threads / 5.0 GHz
SLOB Logical Read Results

running Solaris 11.3

-bash-4.4$ more awr_0w_256r.20180227_113756.txt

WORKLOAD REPOSITORY report for

DB Name         DB Id      Instance     Inst Num Startup Time    Release     RAC
------------ ----------- ------------ -------- --------------- ----------- ---
SLOB          3692102215 SLOB                1 27-Feb-18 10:55 12.1.0.2.0  NO

Host Name        Platform                         CPUs Cores Sockets Memory(GB)
---------------- -------------------------------- ---- ----- ------- ----------
ldom2            Solaris[tm] OE (64-bit)           256    32       1      96.00

              Snap Id      Snap Time      Sessions Curs/Sess
            --------- ------------------- -------- ---------
Begin Snap:      1231 27-Feb-18 11:26:54        53        .5
  End Snap:      1232 27-Feb-18 11:37:50        51        .5
   Elapsed:               10.94 (mins)
   DB Time:            2,753.99 (mins)

Top ADDM Findings by Average Active Sessions
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Finding Name                              Avg act Percen Task Name
----------------------------------------- ------- ------ -----------------------
Top SQL Statements                         251.89  98.68 ADDM:3692102215_1_1232
PL/SQL Execution                           251.89   1.95 ADDM:3692102215_1_1232
Load Profile                     Per Second   Per Transaction  Per Exec  Per Call
~~~~~~~~~~~~~~~             ---------------   --------------- --------- ---------
             DB Time(s):              251.8           5,007.3      0.00     30.14
              DB CPU(s):              251.6           5,002.5      0.00     30.11
      Background CPU(s):                0.1               1.4      0.00      0.00
      Redo size (bytes):            8,475.1         168,516.6
  Logical read (blocks):       50,221,277.5     998,583,141.3
          Block changes:               20.5             406.6
 Physical read (blocks):                2.1              41.1
Physical write (blocks):                6.1             121.8
       Read IO requests:                0.4               8.3
```

SPARC M8 vs Intel Price/Performance

Oracle SPARC T8-1 vs HPE Proliant DL560 Gen10

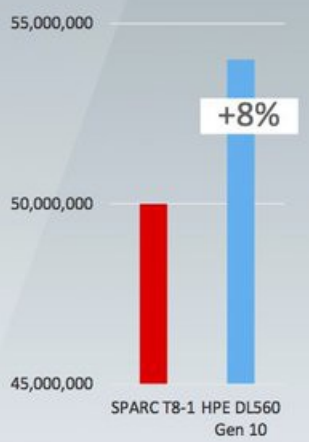
SLOB Benchmark

SPARC T8-1 offers same Price/Performance BUT with Massive License Savings!

SPARC T8-1
1 SPARC M8 Chip
32 cores / 128GB
Solaris 24/7
16 x Oracle Database EE Licenses

$48,320

(HW List Price)



HPE DL560 Gen 10
4 x 3.0GHz Xeon Gold 6154
72 cores / 128GB
RHEL 3Yr 9x5
36 x Oracle Database EE Licenses

$51,697

(HW List Price)

Based on Publicly Available Pricing and Published Performance Results

Quelle: Twitter / @Phil_Dunn1

Links

Solaris 11.4 Download
https://www.oracle.com/technetwork/server-storage/solaris11/downloads/index.html

zfs_msviz: A Tool to Visualise ZFS Metaslab allocations
MOS Doc ID 1624945.1

Hard Partitioning with LDoms

https://www.oracle.com/technetwork/server-storage/vm/ovm-sparc-hard-partitioning-1403135.pdf

Hard Partitioning with Oracle Solaris Zones

https://www.oracle.com/technetwork/server-storage/solaris11/technologies/os-zones-hard-partitioning-2347187.pdf

JomaSoft VDCF - Virtual Datacenter Cloud Framework

https://www.jomasoft.ch/vdcf/

Oracle DB erfolgreich betreiben auf SPARC/LDoms/Solaris/ZFS

Fragen?

Marcel Hofstetter

hofstetter@jomasoft.ch
https://jomasoftmarcel.blogspot.ch

CEO / Enterprise Consultant
JomaSoft GmbH

Oracle ACE „Solaris“

https://www.linkedin.com/in/marcelhofstetter

https://twitter.com/marcel_jomasoft

https://jomasoftmarcel.blogspot.ch

Weitere interessante Vorträge an der #DOAG2018

Di, 20.11. 16:00 Raum Prag	„Praktische Erfahrungen mit SPARC S7-2 Server“ Marcel Hofstetter
Mi, 21.11. 10:00 Raum Prag	„EU-DSGVO und Infrastruktur – ein Fazit nach 6 Monaten“ Jan Brosowski & Ralf Zenses
Do, 22.11. 09:00 Raum Hongkong	„Live Long And Prosper. Solaris 11.4 Vorteile in der Praxis“ Thomas Nau
Do, 22.11. 10:00 Raum Hongkong	„Oracle Solaris 11.4 and Beyond“ Joost Pronk& Jan Brosowski
Do, 22.11. 12:00 Raum Hongkong	„System Monitoring mit Solaris 11.4 DTrace und Analytics“ Thomas Nau
Do, 22.11. 13:00 Raum Hongkong	„SAP und Solaris 11.4 Erste Erfahrungen“ Andris Perkons & Jan Brosowski
Do, 22.11. 14:00 Raum Hongkong	„Was bringt Solaris 11.4“ Marcel Hofstetter