

# Kan vi få for mange filsystemer?

**Er det behov for Oracle Cluster File System?  
Hva skal vi med direct NFS?**

**Ingemar Jansson Haverstad**

# Om foredragsholderen



**Ingemar Jansson Haverstad**  
dbWatch Services

- Arbeidet med Oracle siden 1985
- Flere og flere «hull» i kompetansen...
- Konsulent, instruktør og foredragsholder

ingemar@oraklet.no  
[www.oraklet.no/foredrag](http://www.oraklet.no/foredrag)

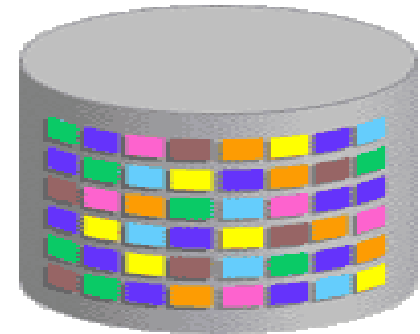
**ORACLE**  
Certified Professional

# Agenda

- Filsystemer generelt
- Oracle Cluster Filsystem - OCFS2
- Direct NFS - dnfs

# Hva kjennetegner et filsystem? <sup>4</sup>

- Generelt
- Begrensninger
- Metadata
- Egenskaper
- Fordelingsprinsipper
- Operativsystem støtte



Kilde - Wiki:

[http://en.wikipedia.org/wiki/Comparison\\_of\\_file\\_systems](http://en.wikipedia.org/wiki/Comparison_of_file_systems)

# Oracle Cluster File System

<http://oss.oracle.com/projects/ocfs2/>

- POSIX-compliant.
- Cluster fil system.
- Real Application Cluster, Web-server og fil-server.
- Utviklet av Oracle.
- OCFS er lisensiert under GPL, versjon 2.
- OCFS er inkludert i 2.6.16 versjonen av Linux-kjernen.
- Tilgjengelig for Linux og Windows.

# Btrfs

<http://oss.oracle.com/>

- «Copy on write» filsystem for Linux.
- Inneholder mange avanserte egenskaper.
- Fokus på sikkerhet feiltoleranse og administrasjon.
- Utviklet av Oracle.
- *Btrfs* er lisensiert under GPL, versjon 2.
- *Btrfs* er inkludert i 2.6.29 versjonen av Linux-kjernen.

Btrfs is under heavy development, and is not suitable for any uses other than benchmarking and review.

# Generelt

<b>Fil system</b>	<b>Opprettet av</b>	<b>Introduksjonsår</b>	<b>Original OS</b>
FAT16	Microsoft	1987	MS-DOS
NTFS	Microsoft	1993	Windows NT
UFS1	Kirk Mc Kussick	1994	
ext3	Stephen Tweedie	1999	Linux
JFS	IBM	1999	OS/2
ZFS	Sun	2004	Solaris
OSCF2	Oracle	2005	Linux
ext4	various	2006	Linux
Btrfs	Oracle	2007	Linux

# Bytes...

**Wiki:** The **tebibyte** is closely related to the **terabyte**, which can be either a synonym for tebibyte, or refer to  $10^{12}$  bytes = 1,000,000,000,000 bytes, depending on context.

Prefiks for <i>bit</i> og <i>byte</i>				
Desimal		Binær		
Verdi	SI	Verdi	IEC	JEDEC
1000	k kilo-	1024	Ki kibi-	K Kilo-
1000 <sup>2</sup>	M mega-	1024 <sup>2</sup>	Mi mebi-	M mega-
1000 <sup>3</sup>	G giga-	1024 <sup>3</sup>	Gi gibi-	G giga-
1000 <sup>4</sup>	T tera-	1024 <sup>4</sup>	Ti tebi-	
1000 <sup>5</sup>	P peta-	1024 <sup>5</sup>	Pi pebi-	
1000 <sup>6</sup>	E exa-	1024 <sup>6</sup>	Ei exbi-	
1000 <sup>7</sup>	Z zetta-	1024 <sup>7</sup>	Zi zebi-	
1000 <sup>8</sup>	Y yotta	1024 <sup>8</sup>	Yi yobi-	

# Begrensninger

Fil system	Maks filnavn	Tegn	Katalog lengde	Fil størrelse	Volum størrelse
FAT16	8.3	Unicode	260(*)	2 GB	2 GB
NTFS	255 tegn	Unicode	255	16 EB	16 EB
UFS1	255 tegn	-	-	32 PB	256 TB
ext3	255 tegn	-	-	2 TB	32 TB
JFS	255 tegn	Unicode	-	4 PB	32 PB
ZFS	255 tegn	-	-	16 EB	ukjent
OSCF2	255 tegn	-	-	8 TB	8 TB
ext4	255 tegn	-	-	16 TB	32 TB
Btrfs	255 tegn	-	ukjent	16 EB	16 EB

# Metadata

	OCFS2	ext4	ZFS	Btrfs
<b>Fil eier</b>	Ja	Ja	Ja	Ja
<b>POSIX rettigheter</b>	Ja	Ja	Ja	Ja
<b>Opprettet dato</b>	Nei	Ja	Ja	Ja
<b>Aksess dato</b>	Ja	Ja	Ja	Ja
<b>Sist endret dato</b>	Ukjent	Ukjent	Ukjent	Ukjent
<b>Dato kopi opprettet</b>	Ukjent	Ukjent	Ukjent	Ukjent
<b>Metadata endret</b>	Ja	Ja	Ja	Ukjent
<b>Arkiv dato</b>	Nei	Nei	Ja	Ukjent
<b>ACL</b>	Nei	Ja	Ja	Ja
<b>MAC</b>	Nei	Ja	(Nei)	Ukjent
<b>Utvidete attributer</b>	Nei	Ja	Ja	Ukjent
<b>Sjekk sum / ECC</b>	Nei	Ja	Ja	Ja

# Egenskaper

	OCFS2	ext4	ZFS	Btrfs
<b>Harde linker</b>	Ja	Ja	Ja	Ja
<b>Symbolske linker</b>	Ja	Ja	Ja	Ja
<b>Blokk journal</b>	Ja	Ja	Ja	Ukjent
<b>Kun metadata journal</b>	Ja	Ja	(Nei)	Ja
<b>Store og små tegn</b>	Ja	Ja	Ja	Ja
<b>Beholder gemena</b>	Ja	Ja	Ja	Ja
<b>Logg for filendringer</b>	Nei	Nei	Nei	Ukjent
<b>Snapshots</b>	Nei	Nei	Ja	Ja
<b>XIP – Execute in Place</b>	Nei	Ja	Ja	Ukjent
<b>Kryptering</b>	Nei	Mei	(Ja)	No
<b>COW – Copy on Write</b>	Ukjent	Ukjent	Ja	Ja
<b>Integrert volumnhåndtering</b>	Ukjent	Nei	Ja	Ja

# Fordelingsprinsipper

	<b>OCFS2</b>	<b>ext4</b>	<b>ZFS</b>	<b>Btrfs</b>
<b>Fordeling av blokk rest</b>	Nei	Nei	Ukjent	Ukjent
<b>Tail packing</b>	Nei	Nei	Delvis	Ukjent
<b>Variabel blokk størrelse</b>	Nei	Nei	Ja	Ukjent
<b>Ekstenter</b>	Ja	Ja	Nei	Ukjent
<b>Allocate-on-flush</b>	Nei	Ja	Ja	Ukjent
<b>Sparse file</b>	Ja	Ja	Ja	Ukjent
<b>Transparent compression</b>	Nei	Nei	Ja	Ukjent

# OCFS2 v1.4

## Nye egenskaper

- 1. Ordered Journal Mode**
- 2. File Attribute Support**
- 3. Performance Enhancements**
- 4. Splice I/O**

Adds support for the new splice(2) system call. This allows for efficient copying between file descriptors by moving the data in kernel.

- 5. Access Time Updates**
- 6. Flexible Allocation**
- 7. Shared Writeable mmap(2)**
- 8. Inline Data**
- 9. Online File system Resize**
- 10. Clustered flock(2)**

The flock(2) system call is now cluster-aware. File locks taken on one node from user-space will interact with those taken on other nodes. All flock(2) options are supported, including the kernel's ability to cancel a lock request when an appropriate kill signal is received. (Note: Support for clustered POSIX file locks, also known as lockf(3) or fcntl(2), has not yet been added. We hope to have that available in the near term.)

# OCFS2 ?

## RAC:

- Lagring av Voting og OCR-diskene.
- Lagring av spfile.
- Alternativ til ASM for «Flash recovery area».



# OCFS2 - Installasjon

```
[root@OEL1 ~]# uname -r  
2.6.18-128.1.10.0.1.el5
```

```
[root@OEL1 ~]# rpm -ivh \  
ocfs2-2.6.18-128.1.10.0.1.el5-1.2.9-1.el5.i686.rpm \  
ocfs2console-1.4.1-1.el5.i386.rpm \  
ocfs2-tools-1.4.1-1.el5.i386.rpm
```

```
Preparing...
```

```
##### [100%]  
 1:ocfs2-tools  
##### [ 33%]  
 2:ocfs2-2.6.18-128.1.10.0  
##### [ 67%]  
 3:ocfs2console  
##### [100%]
```

# OCFS2 - Konfigurering

```
[root@OEL1 ~]# cd /etc/ocfs2
```

```
[root@OEL1 ~]# cat cluster.conf
```

```
cluster:
```

```
node_count = 2
```

```
name = ocfs2
```

```
node:
```

```
ip_port = 7777
```

```
ip_address = 192.168.73.1
```

```
number = 0
```

```
name = xps
```

```
cluster = ocfs2
```

```
node:
```

```
ip_port = 7777
```

```
ip_address = 192.168.73.130
```

```
number = 1
```

```
name = OEL1
```

```
cluster = ocfs2
```

# OCFS2 - Status

```
[root@OEL1 ~]# service o2cb status
Module "configfs": Loaded
Filesystem "configfs": Mounted
Module "ocfs2_nodemanager": Loaded
Module "ocfs2_dlm": Loaded
Module "ocfs2_dlmfs": Loaded
Module "ocfs2_stackglue": Loaded
Filesystem "ocfs2_dlmfs": Mounted
Checking O2CB cluster ocfs2: Online
Heartbeat dead threshold = 7
  Network idle timeout: 10000
  Network keepalive delay: 5000
  Network reconnect delay: 2000
Checking O2CB heartbeat: Active
```

# OCFS2 - Formatering

```
[root@OEL1 ~]# /sbin/mkfs.ocfs2 -L "oradata" /dev/sdb1
mkfs.ocfs2 1.4.1
Cluster stack: classic o2cb
Filesystem label=oradata
Block size=2048 (bits=11)
Cluster size=4096 (bits=12)
Volume size=1069252608 (261048 clusters) (522096 blocks)
17 cluster groups (tail covers 7096 clusters, rest cover 15872
clusters)
Journal size=33554432
Initial number of node slots: 2
Creating bitmaps: done
Initializing superblock: done
Writing system files: done
Writing superblock: done
Writing backup superblock: 0 block(s)
Formatting Journals: done
Formatting slot map: done
Writing lost+found: done
mkfs.ocfs2 successful
```

# OCFS2 - Monitoring

```
[root@OEL1 ~]# mount /dev/sdb1 /oradata
[root@OEL1 ~]# cat /etc/fstab
...
/dev/sdb1 /oradata ocfs2 _netdev,defaults 0 0
```

The screenshot shows a graphical interface for monitoring OCFS2 clusters. At the top, there are menu items: File, Cluster, Tasks, and Help. Below the menu are three buttons: Mount, Unmount, and Refresh, each with a small blue square icon. To the right of these buttons is a text input field labeled "Filter:". Below the buttons is a table with two columns: "Device" and "Mountpoint". The table contains one entry: "/dev/sdb1" under "Device" and "/oradata" under "Mountpoint". Below the table, there are two tabs: "General" and "File Listing". The "General" tab is selected and displays the following information:

- Version: 0.90
- Label: oradata
- UUID: a38f44ea-3b86-4c91-a5e3-b81376d07349
- Maximum Nodes: 2
- Cluster Size: 4 K
- Block Size: 2 K
- Free Space: 950.7 MB (996904960b)
- Total Space: 1019.7 MB (1069252608b)

# Direkte NFS – Server

## Metalink notat 762374.1

```
# id oracle
uid=1005(oracle) gid=1001(oinstall)
groups=1001(oinstall),1002(dba),1003(oper),1004(asm)

# mkdir /dnfs
# chown 1005:1001 /dnfs

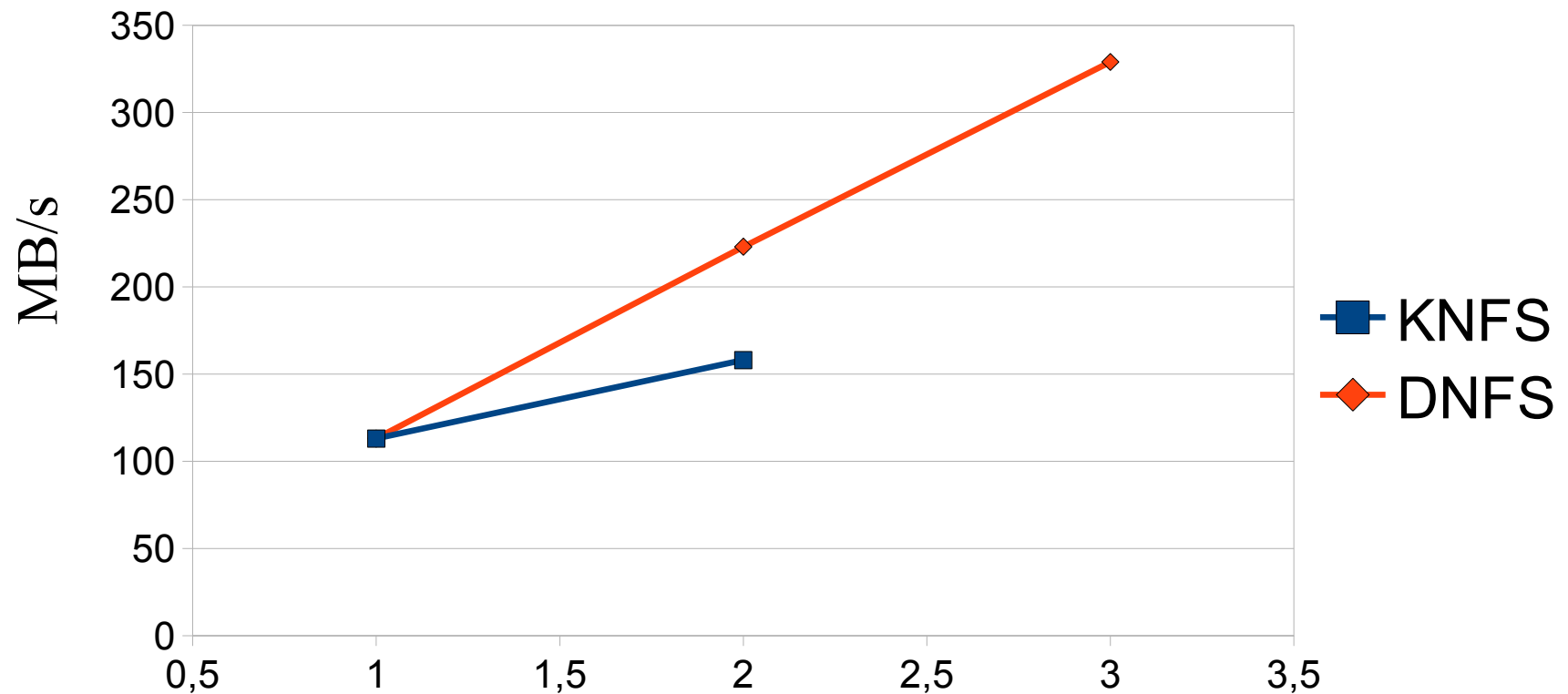
# vi /etc/exports
/oraclenfs * (rw, sync, all_squash, anonuid=500, anongid=500)
# exportfs -a
# exportfs -v

# chkconfig --level 345 nfs on

# service nfs start
```

# Direkte NFS

- Samtidig I/O
- Asynkron I/O
- 4 paralelle nettverkskort



Full table scan throughput Direct NFS vs OS Kernel NFS with Bonding

# Direkt NFS

- Tilgjengelig i *Oracle 11g*.
- Svært enkelt å administrere.
- Forenkler Real Application Cluster oppsett.
- Alternativ til ASM for «Flash recovery area».



# Direkt NFS - Klient

```
# mkdir /oradata1

# vi /etc/fstab
xps:/dnfs /oradata1 nfs \
rw,bg,hard,nointr,rsize=32768,wsiz=32768,tcp,actimeo=0,vers=3,timeo
=600 0 0

# mount /oradata1
# mount
xps:/dnfs on /oradata1 type nfs
(rw,bg,hard,intr,rsize=32768,wsiz=32768,tcp,noac,nfsvers=3,
timeo=600,addr=10.177.52.158)
```

# Direkt NFS - Konfigurierung

```
$ vi $ORACLE_HOME/dbs/oranfstab
server:xps
path: 10.177.52.158
local: 10.177.52.151
path: 10.177.52.159
local: 10.177.52.151
export: /dnfs mount: /oradata1
```

```
$ cd $ORACLE_HOME/lib
$ mv libodm11.so libodm11.so_bak
$ ln -s libnfsodm11.so libodm11.so
```

```
$ sqlplus / as sysdba
SQL> SHUTDOWN IMMEDIATE
SQL> STARTUP
```

```
$ cd /oracle/admin/TEST/bdump
$ tail alertTEST.log
```

```
Oracle instance running with ODM: Oracle Direct NFS Library V 2.0
```

# Sammenfatning

- Mange interessante alternativer.
- OCFS2 fremdeles aktuell.
- Direct NFS gir god ytelse og er enkel å administrere.
- Mulighet for synergier – Btrfs og ZFS.
- Alternativer:
  - ASM til ordinære databasefiler
  - OCFS2 eller Direct NFS til «Flash Recovery Area»