

EOS Open Storage

Using EOS with Erasure Encoding

EOS workshop

4-5 February 2019

CERN

Europe/Zurich time zone

There is a live webcast for this event.



<http://eos.cern.ch>

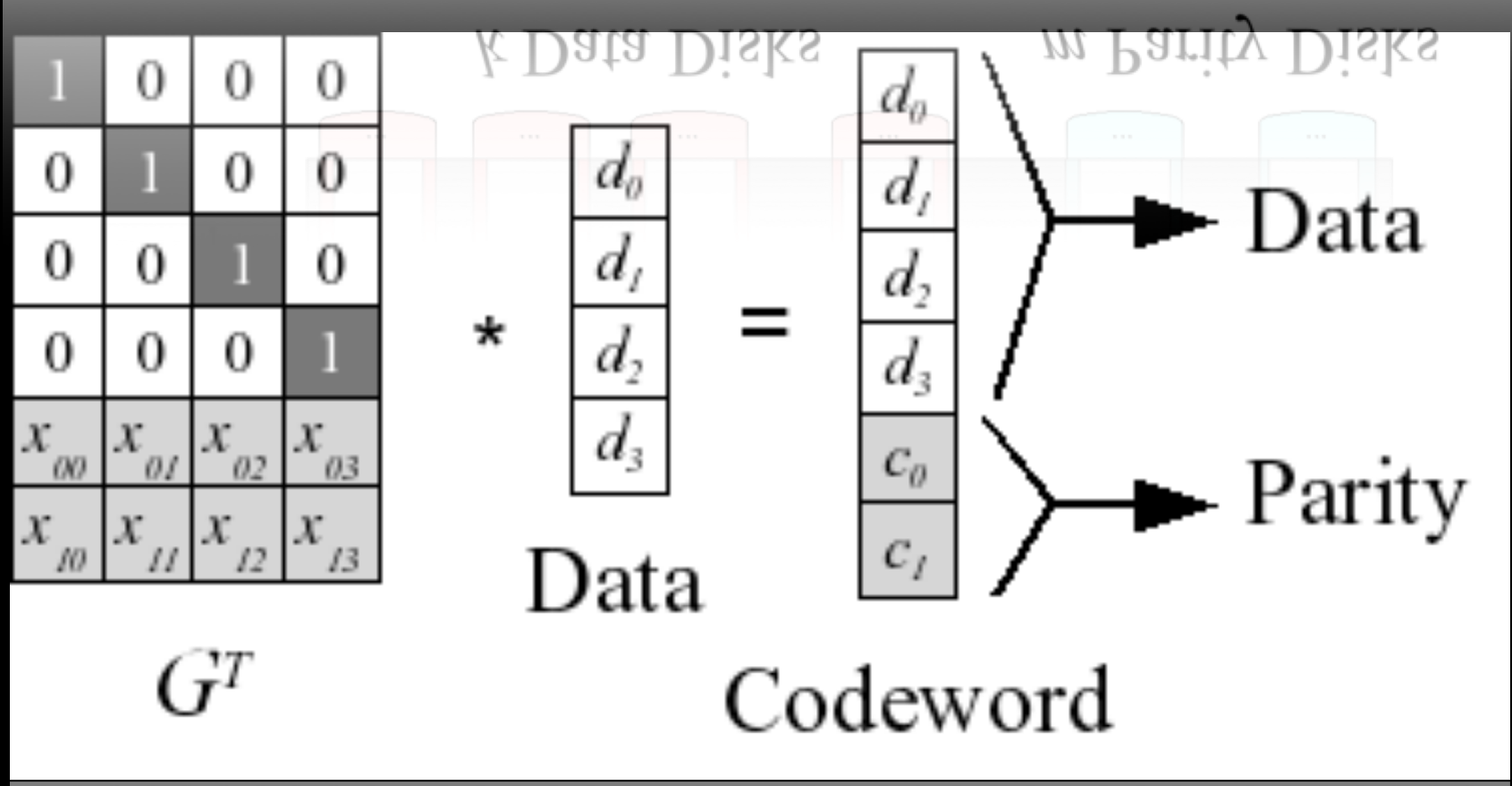
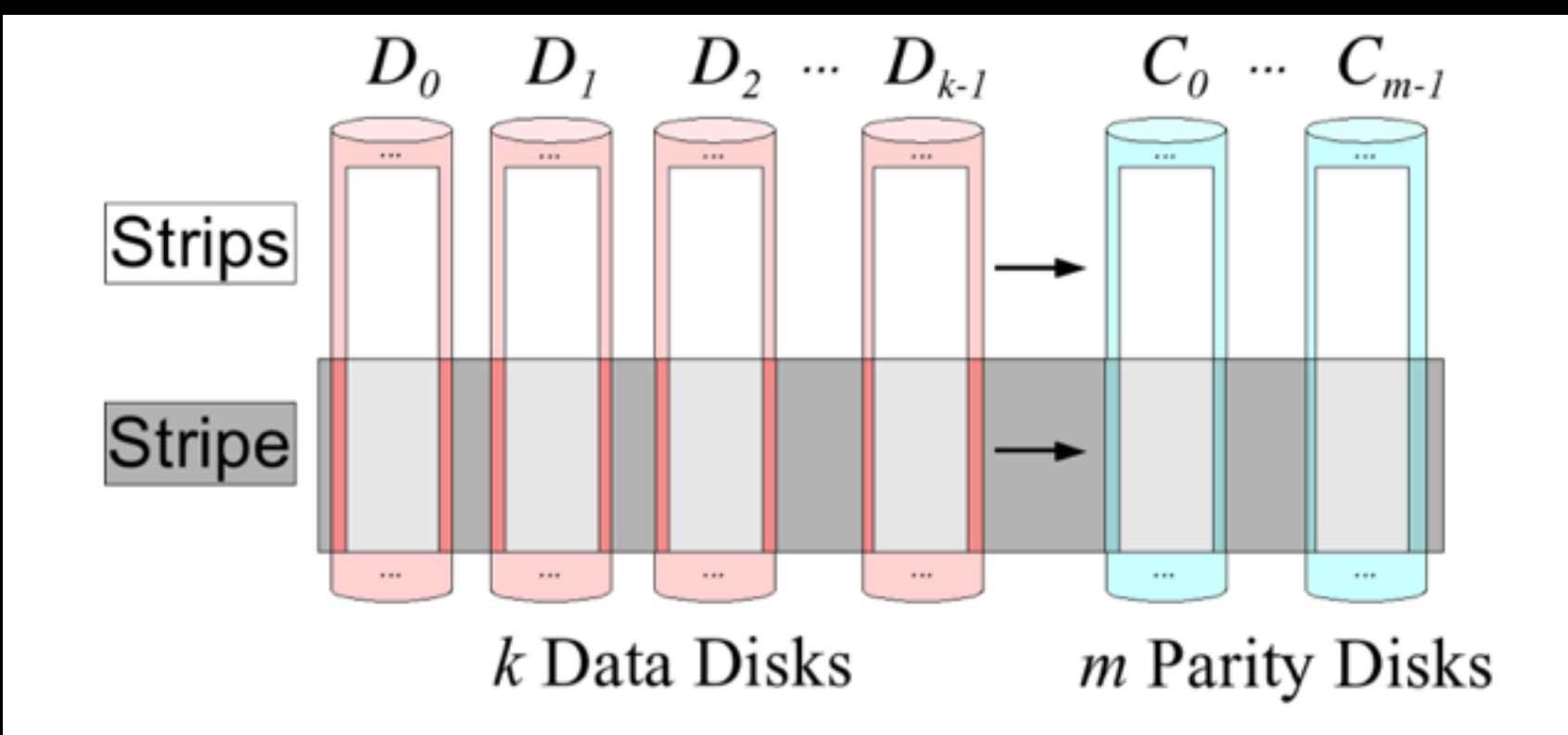
Andreas-Joachim Peters

CERN IT-ST



Erasure Encoding

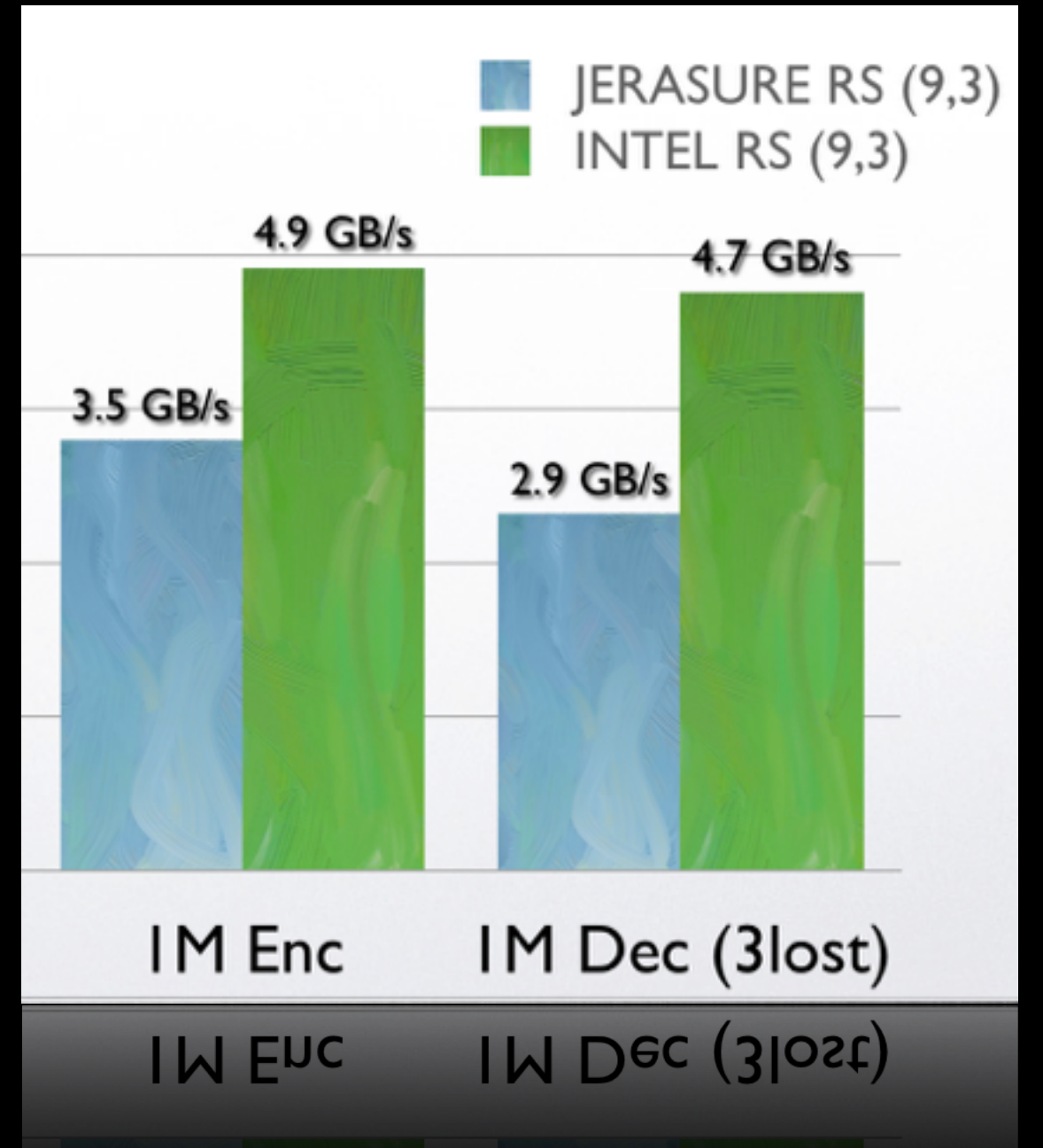
Reed Solomon Encoding - 1960



- developed 1960 at MIT
- widely used in digital storage (RAID/RAIN)
- bar codes
- data transmission
- based on galois-field arithmetic
 - CODEC today fast (GB/s per core) thanks to hardware acceleration of GF multiplications or XOR based codes

Open Source libraries

- Jerasure
- ISA-L (Intel)
- liberasurecode (wrapper)





Erasure Coding in EOS

- uses Reed-Solomon algorithm from JERASURE
 - cauchy matrix & GF-complete library to implement Galois-Field transformations using SSE processor
- Library supports in theory any number of data chunks and parity chunks
- we provide only two default **RAIN** layouts
 - **RAID6** - two parity chunks - redundancy equal to 3 replicas
 - **ARCHIVE** - three parity chunks - redundancy equal to 4 replicas



Erasure Coding in EOS

- EOS Erasure Encoding is **file based**
 - each file is split into blocks (default 1M, maximum 16M)
 - each block gets encoded into M data and K parity chunks
 - blocks are distributed within the same scheduling group
 - each blocks integrity is checked using hardware accelerated 4k crc32c block checksums



Erasure Coding in EOS

- RAID6 - RS(4,2)

```
apeters — root@diamondns:~ — ssh — 84x10
EOS Console [root://localhost] |/eos/diamond/rain/> mkdir -p raid6
EOS Console [root://localhost] |/eos/diamond/rain/> attr set default=raid6 raid6
EOS Console [root://localhost] |/eos/diamond/rain/> attr ls raid6
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="1M"
sys.forced.checksum="adler"
sys.forced.layout="raid6"
sys.forced.nstripses="6"
sys.forced.space="default"
```

- ARCHIVE - RS(5,3)

```
apeters — root@diamondns:~ — ssh — 84x10
EOS Console [root://localhost] |/eos/diamond/rain/> mkdir -p archive
EOS Console [root://localhost] |/eos/diamond/rain/> attr set default=archive archive
EOS Console [root://localhost] |/eos/diamond/rain/> attr ls archive
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="1M"
sys.forced.checksum="adler"
sys.forced.layout="archive"
sys.forced.nstripses="8"
sys.forced.space="default"
```

for RS(M,K): `sys.forced.nstripses = M+K`





Erasure Coding in EOS

- Example: **ARCHIVE** layout in a **6 node storage system**: use RS(3,3)

```
apeters — root@diamonDNS:~ — ssh — 84x10
EOS Console [root://localhost] |/eos/diamond/rain/> mkdir -p raid6
EOS Console [root://localhost] |/eos/diamond/rain/> attr set default=raid6 raid6
EOS Console [root://localhost] |/eos/diamond/rain/> attr ls raid6
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="1M"
sys.forced.checksum="adler"
sys.forced.layout="raid6"
sys.forced.nstripes="6"
sys.forced.space="default"
```

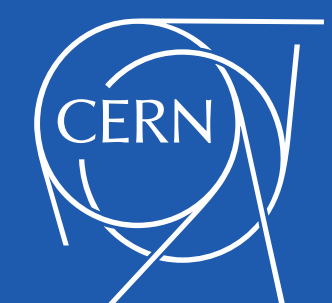


Erasure Coding in EOS

```
apeters — root@diamondns:~ — ssh — 130x35
[root@diamondns ~]# xrdcp /var/tmp/128MB root://localhost//eos/diamond/rain/archive/128MB
[128MB/128MB] [100%] [=====] [5.818MB/s]
```

```
[root@diamondns ~]# eos file info /eos/diamond/rain/archive/128MB
File: '/eos/diamond/rain/archive/128MB'  Flags: 0644
Size: 134217728
Modify: Mon Jan 28 11:14:08 2019 Timestamp: 1548670448.927826000
Change: Mon Jan 28 11:13:46 2019 Timestamp: 1548670426.899151255
CUid: 3 CGid: 4 Fxid: 000019d1 Fid: 6609 Pid: 381 Pxid: 0000017d
XStype: adler XS: 99 e0 65 00 ETAGs: "1774089928704:99e06500"
archive Stripes: 6 Blocksize: 1M LayoutId: 30640522
#Rep: 6
```

no.	fs-id	host	schedgroup	path	boot	configstatus	drainstatus	active	geotag
0	2	diamondns.cern.ch	default.0	/zfsns/fst2	booted	rw	nodrain	online	local
1	7	diamondns.cern.ch	default.0	/zfsns/fst7	booted	rw	nodrain	online	local
2	6	diamondns.cern.ch	default.0	/zfsns/fst6	booted	rw	nodrain	online	local
3	9	diamondns.cern.ch	default.0	/zfsns/fst9	booted	rw	nodrain	online	local
4	3	diamondns.cern.ch	default.0	/zfsns/fst3	booted	rw	nodrain	online	local
5	10	diamondns.cern.ch	default.0	/zfsns/fst10	booted	rw	nodrain	online	local





Erasure Coding in EOS

- **Scheduling Group Layout**

- to use RS(M,K) you need at least (M+K) filesystems per group
- if you have only (M+K), you have a problem to run drain operations because there is no unused disk in a group, which does not have already a stripe



Erasure Coding in EOS

- Example of a 10 FST cluster
 - each node provides a single disk
 - all FSTs live on the same node on different ports (in production you really want to put FSTs on separate nodes for HW redundancy)

EOS Console [root://localhost] |/eos/diamond/rain/> node ls

type	hostport	geotag	status	status	txgw	gw-queued	gw-ntx	gw-rate	heartbeatdelta	nofs
nodesview	diamondns.cern.ch:2001	local	online	on	off	0	10	120	2	1
nodesview	diamondns.cern.ch:2002	local	online	on	off	0	10	120	2	1
nodesview	diamondns.cern.ch:2003	local	online	on	off	0	10	120	2	1
nodesview	diamondns.cern.ch:2004	local	online	on	off	0	10	120	2	1
nodesview	diamondns.cern.ch:2005	local	online	on	off	0	10	120	2	1
nodesview	diamondns.cern.ch:2006	local	online	on	off	0	10	120	2	1
nodesview	diamondns.cern.ch:2007	local	online	on	off	0	10	120	2	1
nodesview	diamondns.cern.ch:2008	local	online	on	off	0	10	120	2	1
nodesview	diamondns.cern.ch:2009	local	online	on	off	0	10	120	2	1
nodesview	diamondns.cern.ch:2010	local	online	on	off	0	10	120	2	1

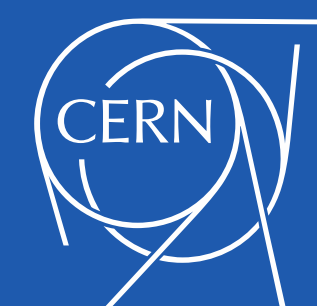


Erasure Encoding EOS

EOS Console [root://localhost] |/eos/diamond/rain/> group ls -l

type	name	status	N(fs)
groupview	default.0	on	10

host	active	port	id	health	uuid	path	schedgroup	headroom	boot	configstatus
drainstatus	scaninterval	statuscomment								
diamondns.cern.ch nodrain	online	2001 604800	1	7d68b00c-86a4-46e4-868d-e4e9e0d2439a N/A		/zfsns/fst1	default.0	0.00	booted	rw
diamondns.cern.ch nodrain	online	2002 604800	2	2ee272c7-22a7-4b48-b235-8afde4c7b234 N/A		/zfsns/fst2	default.0	0.00	booted	rw
diamondns.cern.ch nodrain	online	2003 604800	3	7e8d747c-6dde-4ea2-b30c-7fc1a65dd352 N/A		/zfsns/fst3	default.0	0.00	booted	rw
diamondns.cern.ch nodrain	online	2004 604800	4	00b174a9-a3f6-46c4-8394-b0ac202bc810 N/A		/zfsns/fst4	default.0	0.00	booted	rw
diamondns.cern.ch nodrain	online	2005 604800	5	48d65ae8-f242-466f-9ab3-1811c8ad9242 N/A		/zfsns/fst5	default.0	0.00	booted	rw
diamondns.cern.ch nodrain	online	2006 604800	6	73527d5e-1fa4-4ffd-9c62-e503e0f2b547 N/A		/zfsns/fst6	default.0	0.00	booted	rw
diamondns.cern.ch nodrain	online	2007	7	3ba39b36-840b-4632-aa fb-f3602dcb0a89 N/A		/zfsns/fst7	default.0	0.00	booted	rw
diamondns.cern.ch nodrain	online	2008	8	40565ef6-f23d-4a37-9730-60b019b616c1 N/A		/zfsns/fst8	default.0	0.00	booted	rw
diamondns.cern.ch nodrain	online	2009	9	f7167d44-0e5d-4e44-9a47-3b2eb6779e99 N/A		/zfsns/fst9	default.0	0.00	booted	rw
diamondns.cern.ch nodrain	online	2010	10	9a96a4da-4567-4830-8943-9b4c55ce4c9d N/A		/zfsns/fst10	default.0	0.00	booted	rw





Erasure Encoding EOS

- If you have 10 nodes and 4 disks per node you have to create four scheduling groups each with 1 disk from each node

default.0 (10 filesystems)

default.1 (10 filesystems)

default.2 (10 filesystems)

default.3 (10 filesystems)

- The EOS scheduler selects round-robin each group and then geotree-based within a group



Erasure Encoding in EOS

How to convert from one layout into another

upload a file ...

```
[root@diamondns ~]# xrdcp /var/tmp/128MB root://localhost//eos/diamond/rain/archive/128MB -f  
[128MB/128MB] [100%] [=====] [64MB/s]
```

inspect placement ...

```
EOS Console [root://localhost] |/eos/diamond/rain/archive/> file info 128MB  
File: '/eos/diamond/rain/archive/128MB' Flags: 0644  
Size: 134217728  
Modify: Mon Jan 28 11:30:13 2019 Timestamp: 1548671413.151720000  
Change: Mon Jan 28 11:30:11 2019 Timestamp: 1548671411.668022084  
CUID: 3 CGid: 4 Fxid: 000019d4 Fid: 6612 Pid: 381 Pxid: 0000017d  
XStype: adler XS: 99 e0 65 00 ETAGs: "1774895235072:99e06500"  
archive Stripes: 6 Blocksize: 1M LayoutId: 30640522  
#Rep: 6
```

no.	fs-id	host	schedgroup	path	boot	configstatus	drainstatus	active	geotag
0	1	diamondns.cern.ch	default.0	/zfsns/fst1	booted	rw	nodrain	online	local
1	2	diamondns.cern.ch	default.0	/zfsns/fst2	booted	rw	nodrain	online	local
2	4	diamondns.cern.ch	default.0	/zfsns/fst4	booted	rw	nodrain	online	local
3	8	diamondns.cern.ch	default.0	/zfsns/fst8	booted	rw	nodrain	online	local
4	5	diamondns.cern.ch	default.0	/zfsns/fst5	booted	rw	nodrain	online	local
5	9	diamondns.cern.ch	default.0	/zfsns/fst9	booted	rw	nodrain	online	local



Erasure Encoding in EOS

Converting archive (3,3) to dual replica layout

trigger file conversion manually ...

```
EOS Console [root://localhost] |/eos/diamond/rain/archive/> file convert 128MB replica:2
info: conversion based layout+stripe arguments
success: created conversion job '/eos/diamond/proc/conversion/00000000000019d4:default#03650112'
```

inspect placement ...

```
EOS Console [root://localhost] |/eos/diamond/rain/archive/> file info 128MB
File: '/eos/diamond/rain/archive/128MB'  Flags: 0644
Size: 134217728
Modify: Mon Jan 28 11:30:13 2019 Timestamp: 1548671413.151720000
Change: Mon Jan 28 11:35:23 2019 Timestamp: 1548671723.789229329
CUID: 3 CGid: 4 Fxid: 000019d6 Fid: 6614 Pid: 381 Pxid: 0000017d
XStype: adler XS: 99 e0 65 00 ETAGs: "1775432105984:99e06500"
replica Stripes: 2 blocksize: 4M LayoutId: 00150112
#Rep: 2
```

no.	fs-id	host	schedgroup	path	boot	configstatus	drainstatus	active	geotag
0	9	diamondns.cern.ch	default.0	/zfsns/fst9	booted	rw	nodrain	online	local
1	5	diamondns.cern.ch	default.0	/zfsns/fst5	booted	rw	nodrain	online	local



Erasure Encoding in EOS

Enable the CONVERTER

- Conversion requires to have the CONVERTER thread enabled

```
EOS Console [root://localhost] |/eos/diamond/rain/archive/> space status default
# -----
# Space Variables
# .....
autorepair           := off
balancer             := off
balancer.node.ntx    := 2
balancer.node.rate   := 25
balancer.threshold   := 20
converter            := on
converter.ntx        := 2
```



enable CONVERTER: eos space config default space.converter=on
configure e.g. 2 parallel CONVERSION jobs: eos space config default space.converter.ntx =2



Erasure Encoding in EOS

Converting dual replica to RAID6 - RS (6,2) layout

trigger file conversion manually ...

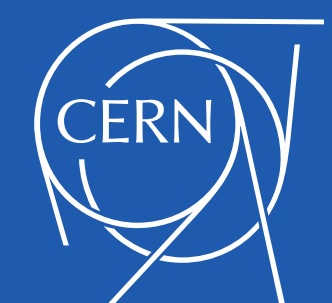
```
EOS Console [root://localhost] |/eos/diamond/rain/archive/> file convert 128MB raid6:8
info: conversion based layout+stripe arguments
success: created conversion job '/eos/diamond/proc/conversion/00000000000019d6:default#20650742'
```

```
EOS Console [root://localhost] |/eos/diamond/rain/archive/> file info 128MB
File: '/eos/diamond/rain/archive/128MB'  Flags: 0644
Size: 134217728
Modify: Mon Jan 28 11:30:13 2019 Timestamp: 1548671413.151720000
Change: Mon Jan 28 11:41:15 2019 Timestamp: 1548672075.454486711
CUid: 3 CGid: 4 Fxid: 000019d8 Fid: 6616 Pid: 381 Pxid: 0000017d
XStype: adler XS: 99 e0 65 00 ETAGs: "1775968976896:99e06500"
raid6 Stripes: 8 Blocksize: 4M LayoutId: 20650742
```

inspect placement ...

#Rep: 8

no.	fs-id	host	schedgroup	path	boot	configstatus	drainstatus	active	geotag
0	1	diamondns.cern.ch	default.0	/zfsns/fst1	booted	rw	nodrain	online	local
1	5	diamondns.cern.ch	default.0	/zfsns/fst5	booted	rw	nodrain	online	local
2	6	diamondns.cern.ch	default.0	/zfsns/fst6	booted	rw	nodrain	online	local
3	7	diamondns.cern.ch	default.0	/zfsns/fst7	booted	rw	nodrain	online	local
4	8	diamondns.cern.ch	default.0	/zfsns/fst8	booted	rw	nodrain	online	local
5	3	diamondns.cern.ch	default.0	/zfsns/fst3	booted	rw	nodrain	online	local
6	9	diamondns.cern.ch	default.0	/zfsns/fst9	booted	rw	nodrain	online	local
7	4	diamondns.cern.ch	default.0	/zfsns/fst4	booted	rw	nodrain	online	local





Erasure Encoding in EOS

Conversion Characteristics

- File **conversion** is asynchronous and queued
 - parallelism depends on your space setting `space.converter.ntx`
 - files change the EOS file id (inode number) after a conversion
 - **warning**: this will trigger a download from CERNBOX clients
- File conversion additionally can specify a *geoplacement* policy (see `eos file -help`)



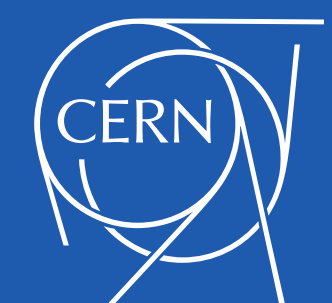
Bulk Layout Conversion

initial layout 2 replica...

```
EOS Console [root://localhost] |/eos/diamond/rain/> attr ls conversion
sys.forced.blocksize="4k"
sys.forced.checksum="adler"
sys.forced.layout="replica"
sys.forced.nstripes="2"
sys.forced.space="default"
```

```
[root@diamonDNS ~]# xrdcp /var/tmp/128MB root://localhost//eos/diamond/rain/conversion/128MB -f
[128MB/128MB] [100%] [=====] [128MB/s]
```

no.	fs-id	host	schedgroup	path	boot	configstatus	drainstatus	active	geotag
0	1	diamonDNS.cern.ch	default.0	/zfsns/fst1	booted	rw	nodrain	online	local
1	8	diamonDNS.cern.ch	default.0	/zfsns/fst8	booted	rw	nodrain	online	local





Bulk Layout Conversion

change new default layout to RAID6 ...

```
EOS Console [root://localhost] |/eos/diamond/rain/> attr -r set default=raid6 conversion
```

verify new default layout ...

```
EOS Console [root://localhost] |/eos/diamond/rain/> attr ls conversion
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="1M"
sys.forced.checksum="adler"
sys.forced.layout="raid6"
sys.forced.nstripes="6"
sys.forced.space="default"
```

convert all existing files to new default layout:

```
[root@diamonDNS ~]# for name in `eos find -f /eos/diamond/rain/conversion/`; do eos file convert $name raid6:6; done
info: conversion based layout+stripe arguments
success: created conversion job '/eos/diamond/proc/conversion/00000000000019e1:default#22650542'
```



Layout Conversion

Improvements

- **Future enhancements**

- create a bulk command to convert and verify layouts of whole tree stub (not existing):

```
file convert -tree /eos/mytree -layout raid6:6
```

```
file convert -verify -tree /eos/mytree -layout raid6:6
```

- avoid rescheduling of an already existing conversion jobs until all existing ones have finished
- implement **conversions like drain jobs**
 - show progress, estimate, failures
- tag layout **policies on spaces**, directories select only the space, then the space policy is applied
 - show risk assessment per group based on the default policy
 - e.g. a group with RS(4,2) is in high danger if two disks are already down



Draining of Erasure Encoded Files

- Filesystem draining is done with the new central drainer:

```
/etc/xrd.cf.mgm:    mgmofs.centraldrain true
```

```
fs config 1 configstatus=drain
```

- Draining of erasure encoded files **amplifies the network bandwidth** required for reconstruction:
 - for a single disk failure $M \cdot \text{vol}(\text{disk})$ has to be read e.g. reconstruct 4 TB = 16 TB of network traffic in a RS(4,2) configuration
 - if several disks are broken, each of them requires $M \cdot \text{vol}(\text{disk})$ to be read



Outlook: EC for CERN

- With decommissioning of Wigner site, we can now move from RS(1,1) to RS(N,2) and benefit from additional space
- >99% of volume is write-once read-many
- e.g. RS(6,2) : 1 TB logical spaces reduces from 2 TB to 1.25 TB
 - CERN capacity of 125 PB of usable space moves to 200 PB of usable space
- Evaluation/Testing/Validation will be done this year during CC migration

THANK YOU

QUESTIONS ?



EOS workshop

4-5 February 2019

CERN

Europe/Zurich                

There is a live webcast for this event.