
from flickr/purplemattfish, Broken hard drive?

revision 5

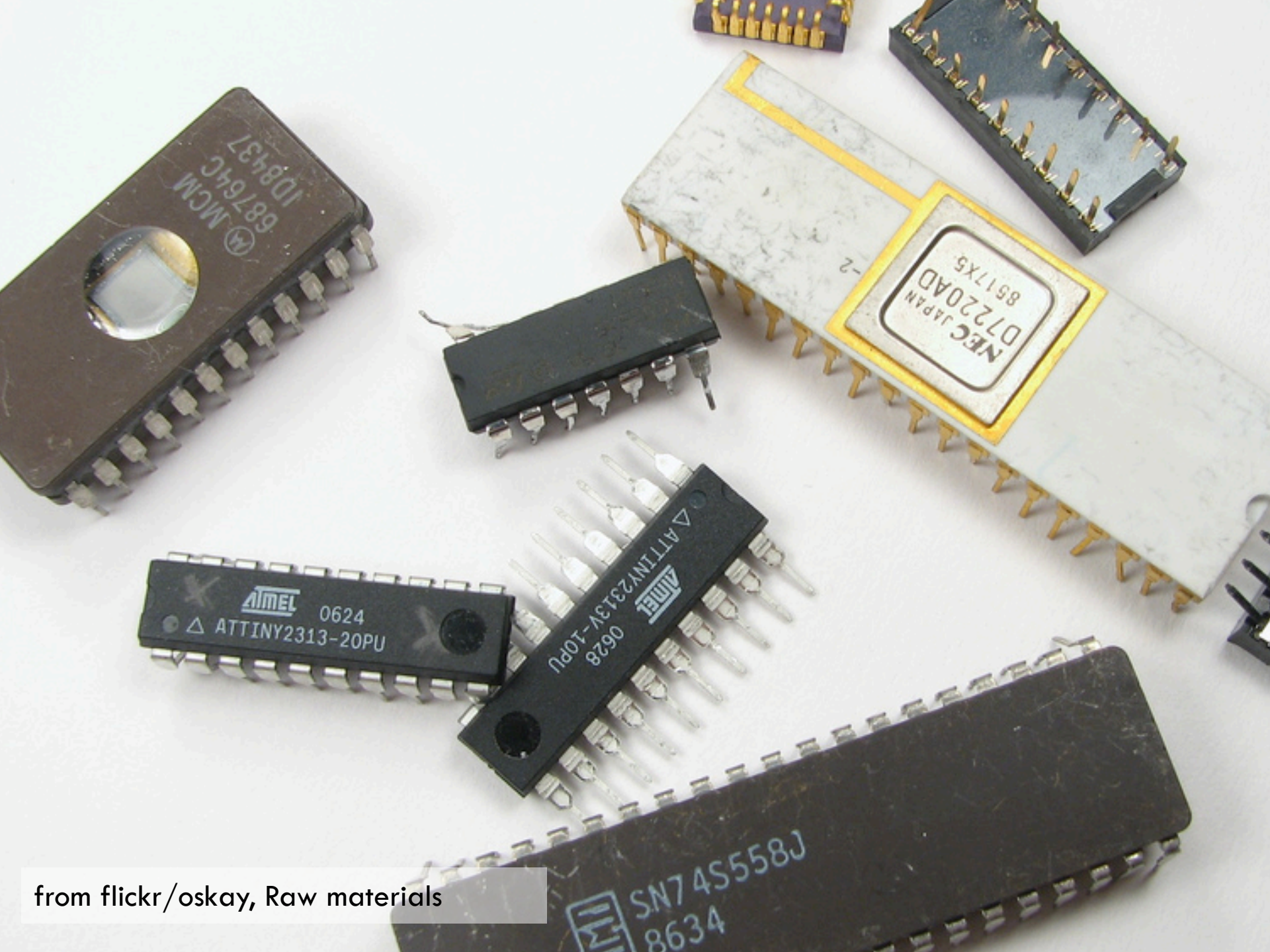
When Bad Things Happen to Good Disks

Erik Riedel, EMC

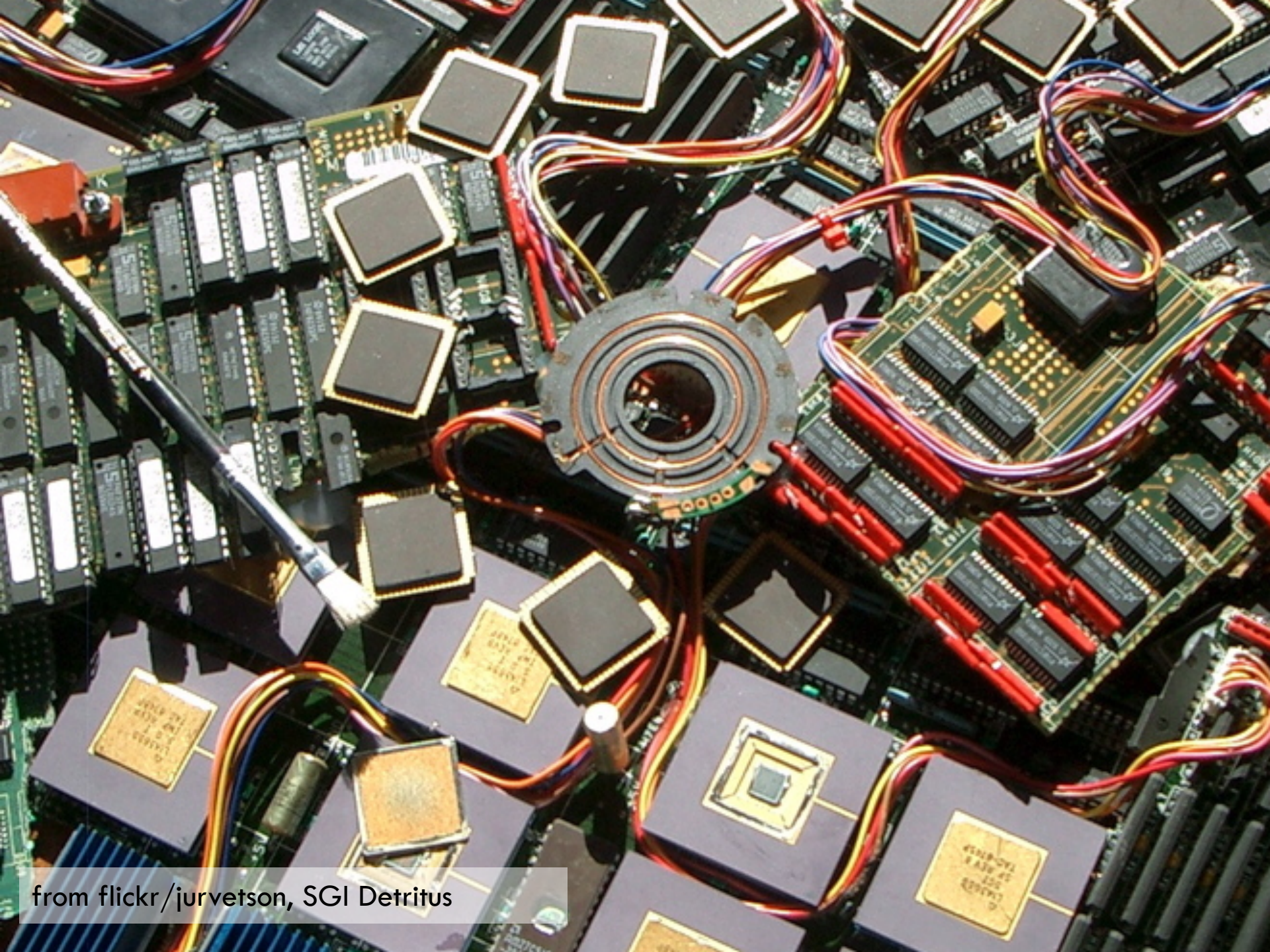
August 2015

aka Disks Don't Have File Descriptors

from flickr/Blude, floppy disks for breakfast



from flickr/oskay, Raw materials

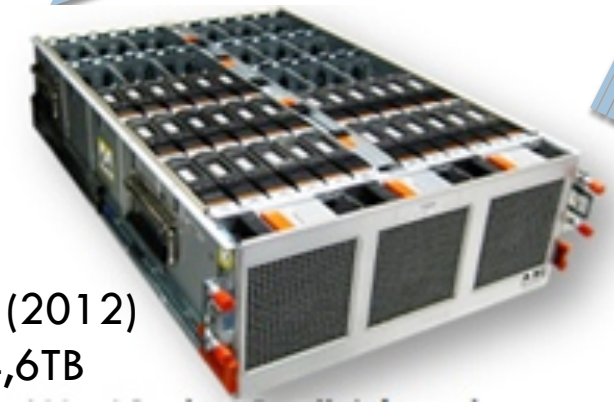


from flickr/jurvetson, SGI Detritus

from flickr/erinhillaw, Math/Physics Bike Rack
and flickr/csavage31, Bike Racks

Gen1 (2008)
1TB

Gen2 (2010)
2,3TB



Gen3 (2012)
3,4,6TB

Gen4 (2014)
6TB

Gen5 (2015)
8TB

- high capacity drives (as many as possible)
- x86 servers/controllers (as few as possible)
- SAS backplanes/cables (not too many, not too few)

Scale Out

RU	NILE DENSE
40	GbE
39	10 GbE
38	10 GbE
37	Empty
36	
35	Rinjin 4 Blade
34	
33	Blank
32	
31	
30	
29	Blank
28	
27	
26	
25	Blank
24	
23	
22	
21	Blank
20	
19	
18	
17	Voyager 15 Disk
16	
15	
14	
13	Voyager 15 Disk
12	
11	
10	
9	Voyager 15 Disk
8	
7	
6	
5	Voyager 15Disk
4	
3	
2	
1	Not Used

480 TB/4n60d
U400

RU	NILE DENSE
40	GbE
39	10 GbE
38	10 GbE
37	Empty
36	
35	Rinjin 4 Blade
34	
33	Blank
32	
31	
30	
29	Blank
28	
27	
26	
25	Blank
24	
23	
22	
21	Blank
20	
19	
18	
17	Voyager 30 Disk
16	
15	
14	
13	Voyager 30 Disk
12	
11	
10	
9	Voyager 30 Disk
8	
7	
6	
5	Voyager 30 Disk
4	
3	
2	
1	Not Used

960 TB/4n120d
U900

RU	NILE DENSE
40	GbE
39	10 GbE
38	10 GbE
37	Empty
36	
35	Rinjin 4 Blade
34	
33	Blank
32	
31	
30	
29	Blank
28	
27	
26	
25	Blank
24	
23	
22	
21	Blank
20	
19	
18	
17	Voyager 60 Disk
16	
15	
14	
13	Voyager 60 Disk
12	
11	
10	
9	Voyager 60 Disk
8	
7	
6	
5	Voyager 60 Disk
4	
3	
2	
1	Not Used

1.9 PB/4n240d
U1900

RU	NILE DENSE
40	GbE
39	10 GbE
38	10 GbE
37	Rinjin 4 Blade
36	
35	Rinjin 4 Blade
34	
33	Voyager 60 Disk
32	
31	
30	
29	Voyager 60 Disk
28	
27	
26	
25	Voyager 60 Disk
24	
23	
22	
21	Voyager 60 Disk
20	
19	
18	
17	Voyager 60 Disk
16	
15	
14	
13	Voyager 60 Disk
12	
11	
10	
9	Voyager 60 Disk
8	
7	
6	
5	Voyager 60 Disk
4	
3	
2	
1	Not Used

3.8PB/8n480d
U4000

17PB/48n3840d

Problem Overview

- set up a collection of 50-node to 500-node Linux clusters at 100s of sites worldwide
- deployed, managed, monitored, serviced by a diverse group of Ops + Service folks (our "customers")
- when something goes (really) wrong, they call your (cell) phone
- approach: keep it simple; make it easy; be proactive; turn off your (cell) phone

It's 4am, the clock is ticking, you have 52* minutes to solve a problem, can you debug it?

*52 minutes is the allowed yearly downtime at "4x 9s" availability

Support calls you at 4am, how many minutes will it take for you to explain what the system is supposed to do, before they can begin to debug and fix it. If it takes 20 minutes to explain the design, you're down to 30 minutes left to fix whatever is wrong. And then nothing else can go wrong until next year.

Marvin Theimer, Amazon (2009 LADDIS workshop talk)

Disks

```
logan-pink:/var/tmp # cs_hal list disks
Disks(s):
SCSI Device Block Device Enclosure   Slot Serial Number       SMART
---------- ------------ ---------- ---- ------------------- ------
/dev/sg6    /dev/sdf     /dev/sg1    C01  WMC130020055        GOOD
/dev/sg7    /dev/sdg     /dev/sg1    A03  Z1F21NXT            FAILED
/dev/sg8    /dev/sdh     /dev/sg1    A00  1EV0XEJB            FAILED
/dev/sg12   /dev/sdl     /dev/sg1    D01  PAGHLBYV            GOOD
/dev/sg13   /dev/sdm     /dev/sg1    D00  PAGHL9TV            GOOD
/dev/sg19   /dev/sds     /dev/sg1    E02  YHGPHUVC            FAILED
/dev/sg20   /dev/sdt     /dev/sg1    B03  S1UYJ1LZ109607      GOOD
/dev/sg123  /dev/sddn    /dev/sg1    E06  AR31021EG513RC      GOOD
/dev/sg33   /dev/sdaf    /dev/sg1    A06  Z1F1CHC8            GOOD
/dev/sg34   /dev/sdag    /dev/sg1    A07  Z1F1CJ5X            GOOD
/dev/sg41   /dev/sdan    /dev/sg1    B06  9XW0BALH            FAILED
/dev/sg42   /dev/sdao    /dev/sg1    A11  9XW094H2            GOOD
/dev/sg43   /dev/sdap    /dev/sg1    B10  PK133WPAG03L9J      FAILED
/dev/sg44   /dev/sdaq    /dev/sg1    B11  WD-WMC190009962     GOOD
/dev/sg51   /dev/sdax    /dev/sg1    E11  Z4D01HKQ            GOOD
/dev/sg89   /dev/sdch    /dev/sg62   E04  9WM2DQ0K            FAILED
/dev/sg64   /dev/sdbi    /dev/sg62   A01  YHH6EJ8A            FAILED
/dev/sg65   /dev/sdbj    /dev/sg62   A02  YHGVVE8A            SUSPECT
/dev/sg68   /dev/sdbm    /dev/sg62   A03  YHH5N6MA            FAILED
/dev/sg69   /dev/sdbn    /dev/sg62   A00  YHH5GWMA            FAILED
/dev/sg71   /dev/sdbp    /dev/sg62   A05  YHH6EWJA            FAILED
/dev/sg72   /dev/sdbq    /dev/sg62   A04  YHH682RA            SUSPECT
/dev/sg73   /dev/sdbr    /dev/sg62   D01  Z1F1CGH7            SUSPECT

 RAID array: 1
    external: 116

total disks: 117
```

Note that /dev/sd* is essentially useless

Where are all my disks? Original smartd config:

/dev/sd[a-z]

shows only 26 drives. What is this, Windows?

A: **/dev/sd[a-z]+**

Disks(s): ONE NODE

SCSI Device	Block Device	Enclosure	Slot	Serial Number	SMART Status
n/a	/dev/md126	RAID vol	n/a	not supported	n/a
/dev/sg0	/dev/sda	intl/sys	0	PWHHBZ7F	GOOD
/dev/sg1	/dev/sdb	intl/sys	1	PWHGVT6F	GOOD
/dev/sg3	/dev/sdc	/dev/sg2	C00	YVHSKHWA	GOOD
/dev/sg4	/dev/sdd	/dev/sg2	A01	YVHRUYEA	GOOD
/dev/sg5	/dev/sde	/dev/sg2	A02	YVHSSHXA	GOOD
/dev/sg6	**/dev/sdf**	**/dev/sg2**	**B00**	**YVHRL21A**	**GOOD**
/dev/sg7	/dev/sdg	/dev/sg2	C01	YVHSB98A	GOOD
/dev/sg8	/dev/sdh	/dev/sg2	A03	YVHSJRRA	GOOD
/dev/sg9	/dev/sdi	/dev/sg2	A00	YVHSMK7A	GOOD
/dev/sg10	/dev/sdj	/dev/sg2	B01	YVHLVEND	GOOD
. . . . .					
/dev/sg63	/dev/sdbj	/dev/sg2	E07	YVHSB4BA	GOOD

Disks(s): ANOTHER NODE

SCSI Device	Block Device	Enclosure	Slot	Serial Number	SMART Status
n/a	/dev/md126	RAID vol	n/a	not supported	n/a
/dev/sg0	/dev/sda	intl/sys	0	PWJMRV8D	GOOD
/dev/sg1	/dev/sdb	intl/sys	1	PWJLVH2F	GOOD
/dev/sg4	/dev/sdu	/dev/sg3	C00	YVK2EWWA	GOOD
/dev/sg5	/dev/sdx	/dev/sg3	A01	YVJWLP3D	GOOD
/dev/sg6	/dev/sdbk	/dev/sg3	A02	YVK078ED	GOOD
/dev/sg7	**/dev/sdbl**	**/dev/sg3**	**B00**	**YVK2V6SA**	**GOOD**
/dev/sg8	/dev/sde	/dev/sg3	C01	YVJWB5KD	GOOD
/dev/sg9	/dev/sdbm	/dev/sg3	A03	YVK2V9BA	GOOD
/dev/sg10	/dev/sdbn	/dev/sg3	A00	YVK1S2RA	GOOD
/dev/sg11	/dev/sdbo	/dev/sg3	B01	YVK2V68A	GOOD
. . . . .					
/dev/sg66	**/dev/sddl**	/dev/sg3	E07	YVK3487A	GOOD

"Need to replace a failed disk on the 3rd node in the cluster, the bad disk is *sdf*".

Density

2012	Disks (raw) @ 3TB	Disks (protected)	Racks @ 480 disks
5 PB	1,700 disks	2,700 disks	6 racks
20 PB	6,700 disks	11,000 disks	23 racks
50 PB	17,000 disks	27,000 disks	56 racks

Density

2012	Disks (raw) @ 3TB	Disks (protected)	Racks @ 480 disks
5 PB	1,700 disks	2,700 disks	6 racks
20 PB	6,700 disks	11,000 disks	23 racks
50 PB	17,000 disks	27,000 disks	56 racks

2014	Disks (raw) @ 6TB	Disks (protected)	Racks @ 480 disks
5 PB	830 disks	1,300 disks	3 racks
20 PB	3,300 disks	5,300 disks	12 racks
50 PB	8,300 disks	13,000 disks	28 racks

Density

Updated from "Long-Term Storage", presented at Library of Congress Workshop in September 2012

2012	Disks (raw) @ 3TB	Disks (protected)	Racks @ 480 disks
5 PB	1,700 disks	2,700 disks	6 racks
20 PB	6,700 disks	11,000 disks	23 racks
50 PB	17,000 disks	27,000 disks	56 racks

2014	Disks (raw) @ 6TB	Disks (protected)	Racks @ 480 disks
5 PB	830 disks	1,300 disks	3 racks
20 PB	3,300 disks	5,300 disks	12 racks
50 PB	8,300 disks	13,000 disks	28 racks

2016	Disks (raw) @ 10TB	Disks (protected)	Racks @ 780 disks
5 PB	500 disks	700 disks	1 rack
20 PB	2,000 disks	2,800 disks	4 racks
50 PB	5,000 disks	7,000 disks	9 racks

from flickr/purplemattfish, Broken hard drive?

```
dino-black:~ % cs_hal list disks
Disks(s):
SCSI Device Block Device Enclosure   Slot Serial Number      SMART
---------- ------------ ---------- ---- ------------------ -----
n/a        /dev/sda     RAID vol   n/a  not supported      n/a
/dev/sg0   n/a          RAID array 0    9QE801ME           GOOD
/dev/sg1   n/a          RAID array 1    9QE834TG           GOOD
/dev/sg3   /dev/sdb     /dev/sg18  0    9WM0R49P           GOOD
/dev/sg4   /dev/sdc     /dev/sg18  1    9WM0R48T           GOOD
/dev/sg5   /dev/sdd     /dev/sg18  2    9WM0R3Z4           GOOD
/dev/sg6   /dev/sde     /dev/sg18  3    9WM0R4VK         SUSPECT:
                                           Reallocated(5)=19
/dev/sg7   /dev/sdf     /dev/sg18  4    9WM0RF21           GOOD
/dev/sg8   /dev/sdg     /dev/sg18  5    9WM0R44B           GOOD
/dev/sg9   /dev/sdh     /dev/sg18  6    9WM0R3E0           GOOD
/dev/sg10  /dev/sdi     /dev/sg18  7    9WM0RF2X           GOOD
/dev/sg11  /dev/sdj     /dev/sg18  8    9WM0R4TX           GOOD
/dev/sg12  /dev/sdk     /dev/sg18  9    9WM0REHK           GOOD
/dev/sg13  /dev/sdl     /dev/sg18  10   9WM0R3EW           GOOD
/dev/sg14  /dev/sdm     /dev/sg18  11   9WM0R4GY           GOOD
/dev/sg15  /dev/sdn     /dev/sg18  12   9WM0R4NZ           GOOD
/dev/sg16  /dev/sdo     /dev/sg18  13   9WM0RF42           GOOD
/dev/sg17  /dev/sdp     /dev/sg18  14   9WM0R3AS           GOOD

 RAID array: 2
   external: 15

total disks: 17
```

disks

Basic cross-layer mapping of disks, and RAIDs, and enclosures.

```
dino-black:~ % cs_hal list fs
Volume(s):
SCSI Device Block Device FS UUID                                Type      Slot  Label SMART              Mount Point
----------- ------------ ------------------------------------- --------- ----- ----- ------             ------------
/dev/sg2    /dev/sda     0ddb9635-ff27-4cd3-8c2f-58a6f5226d30  ext3            BOOT  GOOD /boot
/dev/sg2    /dev/sda     2192b3ef-2a44-4450-9b04-327c00215454  xfs                   GOOD /root2
/dev/sg2    /dev/sda     ffa9607a-4b6f-4218-9266-c083fb1989a1  xfs                   GOOD /var
/dev/sg2    /dev/sda     746b09d4-f07a-49dc-8b40-86220dfc7edc  xfs                   GOOD /
/dev/sg2    /dev/sda     f7c37c92-4bc5-4abf-95a5-efa51c46f6bc  swap v1               GOOD -
/dev/sg3    /dev/sdb     90a52650-e0f3-49e4-810b-a505cdcadb51  xfs       0           GOOD /data-disks/ss-90a52650-e0f3-49e4-810b-a505cdcadb51
/dev/sg4    /dev/sdc     173aef8b-80e9-4be2-a510-3b88d3343f8a  xfs       1           GOOD /data-disks/ss-173aef8b-80e9-4be2-a510-3b88d3343f8a
/dev/sg5    /dev/sdd     bcfb1897-152b-482b-bde6-de9665ad7c51  xfs       2           GOOD /data-disks/ss-bcfb1897-152b-482b-bde6-de9665ad7c51
/dev/sg6    /dev/sde     bc6946ae-770f-4621-9ea5-f2d1e5ec0f28  xfs       3         SUSPECT /data-disks/ss-bc6946ae-770f-4621-9ea5-f2d1e5ec0f28
/dev/sg7    /dev/sdf     52446742-a566-4036-8b0c-5cd7901474f0  xfs       4           GOOD /data-disks/ss-52446742-a566-4036-8b0c-5cd7901474f0
/dev/sg8    /dev/sdg     c9ee0971-d8dc-4621-8958-d79890d0f590  xfs       5           GOOD /data-disks/ss-c9ee0971-d8dc-4621-8958-d79890d0f590
/dev/sg9    /dev/sdh     294bcd25-ab19-40ee-8c03-cd71e94e9e06  xfs       6           GOOD /meta/294bcd25-ab19-40ee-8c03-cd71e94e9e06
/dev/sg10   /dev/sdi     cb5cac6c-1cdf-49ec-8754-a475db3d4afd  xfs       7           GOOD /data-disks/ss-cb5cac6c-1cdf-49ec-8754-a475db3d4afd
/dev/sg11   /dev/sdj     91739495-2a46-47d2-8676-d8b4b3f8fd76  xfs       8           GOOD /data-disks/ss-91739495-2a46-47d2-8676-d8b4b3f8fd76
/dev/sg12   /dev/sdk     9f2a0ae1-d97b-4fb1-873e-6a9bfb2c3254  xfs       9           GOOD /data-disks/ss-9f2a0ae1-d97b-4fb1-873e-6a9bfb2c3254
/dev/sg13   /dev/sdl     404a8c5a-19c0-4949-bd33-edd83ca4ee8f  xfs       10          GOOD /meta/404a8c5a-19c0-4949-bd33-edd83ca4ee8f
/dev/sg14   /dev/sdm     da36046f-41f7-46d4-bcaa-af183002b792  xfs       11          GOOD /data-disks/ss-da36046f-41f7-46d4-bcaa-af183002b792
/dev/sg15   /dev/sdn     a71b6937-8ae5-4a37-96d0-78feeb0e62c4  xfs       12          GOOD /data-disks/ss-a71b6937-8ae5-4a37-96d0-78feeb0e62c4
/dev/sg16   /dev/sdo     34d6f5c5-1f5d-4cea-af5a-af157324aee8  xfs       13          GOOD /meta/34d6f5c5-1f5d-4cea-af5a-af157324aee8
/dev/sg17   /dev/sdp     9cc59415-cab5-4456-881f-a0c533e1823d  xfs       14          GOOD /data-disks/ss-9cc59415-cab5-4456-881f-a0c533e1823d
```

```
dino-black:~ % cs_hal list fs
Volume(s):
SCSI Device Block Device FS UUID                                Type    Slot Label SMART   Mount
----------- ------------ ------------------------------------- ------- ---- ----- ------  ------
/dev/sg2    /dev/sda     0ddb9635-ff27-4cd3-8c2f-58a6f5226d30  ext3         BOOT  GOOD    /boot
/dev/sg2    /dev/sda     2192b3ef-2a44-4450-9b04-327c00215454  xfs                GOOD    /root2
/dev/sg2    /dev/sda     ffa9607a-4b6f-4218-9266-c083fb1989a1  xfs                GOOD    /var
/dev/sg2    /dev/sda     746b09d4-f07a-49dc-8b40-86220dfc7edc  xfs                GOOD    /
/dev/sg2    /dev/sda     f7c37c92-4bc5-4abf-95a5-efa51c46f6bc  swap v1            GOOD    -
```

volumes (file systems)

```
dino-black:~ % cs_hal list disks
Disks(s):
SCSI Device Block Device Enclosure   Slot Serial Number       SMART
----------- ------------ ----------- ---- ------------------ -----
n/a         /dev/sda     RAID vol    n/a  not supported      n/a
/dev/sg0    n/a          RAID array  0    9QE801ME           GOOD
/dev/sg1    n/a          RAID array  1    9QE834TG           GOOD
/dev/sg3    /dev/sdb     /dev/sg18   0    9WM0R49P           GOOD
```

HAL – details

abstracted object model embodied as a library

```
silver-is1-004:~ % cs_hal list disks
Disks(s):
SCSI Device Block Device Enclosure   Slot Serial Number       SMART Status
----------- ------------ ----------- ---- ------------------- ------------
n/a         /dev/md126   RAID vol    n/a  not supported       n/a
/dev/sg1    n/a          RAID array  1    KLH6DNZJ            GOOD
/dev/sg0    n/a          RAID array  0    KLH6DL7J            GOOD
/dev/sg27   /dev/sdy     /dev/sg2    B04  Z1Z0EVBF            GOOD
/dev/sg28   /dev/sdz     /dev/sg2    C04  Z1Z0EKFZ            GOOD
/dev/sg29   /dev/sdaa    /dev/sg2    D04  Z1Z0ETMY            GOOD
/dev/sg30   /dev/sdab    /dev/sg2    E05  Z1Z0EVLG            GOOD
/dev/sg31   /dev/sdac    /dev/sg2    E04  Z1Z0EVH9            GOOD
. . .
. . .
. . .
/dev/sg47   /dev/sdas    /dev/sg2    C11  Z1Z0ETTT            GOOD
/dev/sg48   /dev/sdat    /dev/sg2    D11  Z1Z0EVAM            GOOD
/dev/sg49   /dev/sdau    /dev/sg2    C10  Z1Z0ETFN            GOOD
/dev/sg50   /dev/sdav    /dev/sg2    D10  Z1Z0EVC4            GOOD
/dev/sg51   /dev/sdaw    /dev/sg2    C09  Z1Z0EVCR            GOOD
/dev/sg52   /dev/sdax    /dev/sg2    D09  Z1Z0ETEP            GOOD
/dev/sg53   /dev/sday    /dev/sg2    E11  Z1Z0EKG3            GOOD
/dev/sg54   /dev/sdaz    /dev/sg2    E10  Z1Z0ETLV            GOOD
/dev/sg55   /dev/sdba    /dev/sg2    E09  Z1Z0EV1A            GOOD
/dev/sg56   /dev/sdbb    /dev/sg2    C08  Z1Z0EV90            GOOD
```

```
 RAID array: 2
   external: 60

total disks: 62
```

```
silver-is1-004:~ % cs_hal info sg2
SCSI enclosure   : /dev/sg2
bsg              : /dev/bsg/expander-1:0
id               : 50060480e01b09be
S/N              : 50060480e01b09be
expander count   : 2
zoned            : no
zoning supported : yes
zone saving      : yes
disk slot count  : 60
disk count       : 60
LED              : OFF
vendor           : EMC
model            : ESES Enclosure
firmware         : 0001
SCSI id          : 1:0:0:0
SAS address      : 50060480e01b09be
state            : awake and running
HBA              : 0000:02:00.0
```

```
silver-is1-004:~ % cs_hal info sg27
SCSI disk        : /dev/sg27
block device     : /dev/sdy
size (via SCSI)  : 3726.02 GB
size (via blk)   : 3726.02 GB
vendor           : ATA
model            : ST4000NM0033-9ZM
firmware         : GT00
SCSI id          : 1:0:25:0
S/N              : Z1Z0EVBF
SAS address      : 50060480e832bc16
state            : awake and running
RAID             : no
internal         : no
system disk      : no
type             : rotational
volume count     : 1
volume           : /dev/sdy1
volume size      : 3726.02 GB
filesystem       : 285b59d3-xxxx (xfs; mounted)
slot name        : B04
parent enc       : sg2
parent exp       : sg3
parent HBA       : 0000:02:00.0
LED              : OFF
SMART            : GOOD
```

HAL – blinks

```
silver-is1-004:~ % cs_hal list disks
Disks(s):
SCSI Device Block Device Enclosure   Slot Serial Number       SMART Status
----------- ------------ ----------- ---- ------------------- ------------
n/a         /dev/md126   RAID vol    n/a  not supported       n/a
/dev/sg1    n/a          RAID array  1    KLH6DNZJ            GOOD
/dev/sg0    n/a          RAID array  0    KLH6DL7J            GOOD
/dev/sg27   /dev/sdy     /dev/sg2    B04  Z1Z0EVBF            GOOD
/dev/sg28   /dev/sdz     /dev/sg2    C04  Z1Z0EKFZ            GOOD
/dev/sg29   /dev/sdaa    /dev/sg2    D04  Z1Z0ETMY            GOOD
/dev/sg30   /dev/sdab    /dev/sg2    E05  Z1Z0EVLG            GOOD
/dev/sg31   /dev/sdac    /dev/sg2    E04  Z1Z0EVH9            GOOD
. . .
. . .
. . .
/dev/sg47   /dev/sdas    /dev/sg2    C1
/dev/sg48   /dev/sdat    /dev/sg2    D1
/dev/sg49   /dev/sdau    /dev/sg2    C1(
/dev/sg50   /dev/sdav    /dev/sg2    D1(
/dev/sg51   /dev/sdaw    /dev/sg2    C0
/dev/sg52   /dev/sdax    /dev/sg2    D0
/dev/sg53   /dev/sday    /dev/sg2    E1
/dev/sg54   /dev/sdaz    /dev/sg2    E1(
/dev/sg55   /dev/sdba    /dev/sg2    E0
/dev/sg56   /dev/sdbb    /dev/sg2    C0
```

```
silver-is1-004:~ % cs_hal led sg2 blink
cs_hal: setting LED state of enclosure sg2 from 'OFF' to 'BLINK'

silver-is1-004:~ % cs_hal led sg27 blink
cs_hal: setting LED state of disk sg27 from 'OFF' to 'BLINK'

silver-is1-004:~ % cs_hal led Z1Z0EVBF blink
cs_hal: setting LED state of disk Z1Z0EVBF from 'OFF' to 'BLINK'

silver-is1-004:~ % cs_hal led node on
cs_hal: setting LED state of node to 'ON'
```

```
 RAID array: 2
   external: 60

total disks: 62
```


```
silver-is1-004:~ % cs_hal led node off
cs_hal: setting LED state of node to 'OFF'

silver-is1-004:~ % cs_hal led sg27 off
cs_hal: setting LED state of disk sg27 from 'BLINK' to 'OFF'

silver-is1-004:~ % cs_hal led sg2 off
cs_hal: setting LED state of enclosure sg2 from 'BLINK' to 'OFF'
```

"There are three bad disks in the cluster, they are all out of file descriptors."

"Try a reboot, it might fix those disks [without sending out spares]"

Serial Number:	K85ETA1266AF	Serial Number:	K85ETA1266AF
Serial Number:	K85ET9B260PB	Serial Number:	K85ET9B260PB
FEBRUARY		**OCTOBER**	
Serial Number:	WCAVY2775469	Serial Number:	WCAVY2775469
Serial Number:	WCAVY2775477	Serial Number:	WCAVY2775477
Serial Number:	WCAVY2766210	Serial Number:	WCAVY2766210
Serial Number:	WCAVY2570279	Serial Number:	WCAVY2570279
Serial Number:	**WCAVY2568648**	**Serial Number:**	**WCAVY2568648**
Serial Number:	**WCAVY2568648**	Serial Number:	WCAVY2768844
Serial Number:	WCAVY2768780	Serial Number:	WCAVY2768780
Serial Number:	WCAVY2768776	Serial Number:	WCAVY2768776
Serial Number:	WCAVY2631983	Serial Number:	WCAVY2631983
Serial Number:	WCAVY2764736	Serial Number:	WCAVY2764736
Serial Number:	WCAVY2766146	Serial Number:	WCAVY2766146
Serial Number:	WCAVY2768581	Serial Number:	WCAVY2768581
Serial Number:	WCAVY2775457	Serial Number:	WCAVY2775457
Serial Number:	WCAVY2775980	Serial Number:	WCAVY2775980
Serial Number:	WCAVY2567467	Serial Number:	WCAVY2567467

```
smartctl version 5.37 [x86_64-unknown-linux-gnu] Copyright (C) 2002-6 Bruce Allen
=== START OF INFORMATION SECTION ===
Device Model:     WD2002FYPS-12
Serial Number:    WCAVY2568648
Firmware Version: 02.S0500
User Capacity:    2,000,398,934,016 bytes
ID# ATTRIBUTE_NAME          FLAG     VALUE WORST THRESH TYPE      UPDATED  WHEN_FAILED RAW_VALUE
4 Start_Stop_Count        0x0032   100   100   000    Old_age   Always       -       73
5 Reallocated_Sector_Ct   0x0033   200   200   140    Pre-fail  Always       -       0
9 Power_On_Hours          0x0032   086   086   000    Old_age   Always       -       10762

smartctl version 5.37 [x86_64-unknown-linux-gnu] Copyright (C) 2002-6 Bruce Allen
=== START OF INFORMATION SECTION ===
Device Model:     WD2002FYPS-12
Serial Number:    WCAVY2568648
Firmware Version: 02.S0500
User Capacity:    2,000,398,934,016 bytes
ID# ATTRIBUTE_NAME          FLAG     VALUE WORST THRESH TYPE      UPDATED  WHEN_FAILED RAW_VALUE
4 Start_Stop_Count        0x0032   100   100   000    Old_age   Always       -       73
5 Reallocated_Sector_Ct   0x0033   200   200   140    Pre-fail  Always       -       0
9 Power_On_Hours          0x0032   086   086   000    Old_age   Always       -       10762
```

"The disk with the duplicate serial number needs to be replaced."

Serial Number:	K85ETA1266AF	Serial Number:	K85ETA1266AF
Serial Number:	K85ET9B260PB	Serial Number:	K85ET9B260PB
FEBRUARY		**OCTOBER**	
Serial Number:	WCAVY2775469	Serial Number:	WCAVY2775469
Serial Number:	WCAVY2775477	Serial Number:	WCAVY2775477
Serial Number:	WCAVY2766210	Serial Number:	WCAVY2766210
Serial Number:	WCAVY2570279	Serial Number:	WCAVY2570279
Serial Number:	***WCAVY2568648***	***Serial Number:***	***WCAVY2568648***
Serial Number:	***WCAVY2568648***	**Serial Number:**	**WCAVY2768844**
Serial Number:	WCAVY2768780	Serial Number:	WCAVY2768780
Serial Number:	WCAVY2768776	Serial Number:	WCAVY2768776
Serial Number:	WCAVY2631983	Serial Number:	WCAVY2631983
Serial Number:	WCAVY2764736	Serial Number:	WCAVY2764736
Serial Number:	WCAVY2766146	Serial Number:	WCAVY2766146
Serial Number:	WCAVY2768581	Serial Number:	WCAVY2768581
Serial Number:	WCAVY2775457	Serial Number:	WCAVY2775457
Serial Number:	WCAVY2775980	Serial Number:	WCAVY2775980
Serial Number:	WCAVY2567467	Serial Number:	WCAVY2567467

```
smartctl version 5.37 [x86_64-unknown-linux-gnu] Copyright (C) 2002-6 Bruce Allen
=== START OF INFORMATION SECTION ===
Device Model:     WD2002FYPS-12
Serial Number:    WCAVY2568648
Firmware Version: 02.S0500
User Capacity:    2,000,398,934,016 bytes
ID# ATTRIBUTE_NAME          FLAG     VALUE WORST THRESH TYPE      UPDATED  WHEN_FAILED RAW_VALUE
4 Start_Stop_Count        0x0032   100   100   000    Old_age   Always       -       73
5 Reallocated_Sector_Ct   0x0033   200   200   140    Pre-fail  Always       -       0
9 Power_On_Hours          0x0032   086   086   000    Old_age   Always       -       10762

smartctl version 5.37 [x86_64-unknown-linux-gnu] Copyright (C) 2002-6 Bruce Allen
=== START OF INFORMATION SECTION ===
Device Model:     WD2002FYPS-12
Serial Number:    WCAVY2568648
Firmware Version: 02.S0500
User Capacity:    2,000,398,934,016 bytes
ID# ATTRIBUTE_NAME          FLAG     VALUE WORST THRESH TYPE      UPDATED  WHEN_FAILED RAW_VALUE
4 Start_Stop_Count        0x0032   100   100   000    Old_age   Always       -       73
5 Reallocated_Sector_Ct   0x0033   200   200   140    Pre-fail  Always       -       0
9 Power_On_Hours          0x0032   086   086   000    Old_age   Always       -       10762
```

"The disk with the duplicate serial number needs to be replaced."

```
logan-pink:/var/tmp # cs_hal list disks
Disks(s):
SCSI Device Block Device Enclosure   Slot Serial Number       SMART
---------- ------------ ---------- ---- ------------------ -----------
/dev/sg6    /dev/sdf     /dev/sg1    C01  WMC130020055       GOOD
/dev/sg7    /dev/sdg     /dev/sg1    A03  Z1F21NXT           FAILED:
                                      Reallocated_Sector_Count(5)=1288
/dev/sg8    /dev/sdh     /dev/sg1    A00  1EV0XEJB           FAILED:
                                      self-test fail; read element;
/dev/sg12   /dev/sdl     /dev/sg1    D01  PAGHLBYV           GOOD
/dev/sg13   /dev/sdm     /dev/sg1    D00  PAGHL9TV           GOOD
/dev/sg19   /dev/sds     /dev/sg1    E02  YHGPHUVC           FAILED:
                                      Reallocated_Sector_Count(5)=1814
/dev/sg20   /dev/sdt     /dev/sg1    B03  S1UYJ1LZ109607     GOOD:
                                      self-test in progress; 40% done;
/dev/sg123  /dev/sddn    /dev/sg1    E06  AR31021EG513RC     GOOD
/dev/sg33   /dev/sdaf    /dev/sg1    A06  Z1F1CHC8           GOOD
/dev/sg34   /dev/sdag    /dev/sg1    A07  Z1F1CJ5X           GOOD
/dev/sg41   /dev/sdan    /dev/sg1    B06  9XW0BALH           FAILED:
                                      Reallocated_Sector_Count(5)=168
/dev/sg42   /dev/sdao    /dev/sg1    A11  9XW094H2           GOOD
/dev/sg43   /dev/sdap    /dev/sg1    B10  PK133WPAG03L9J     FAILED:
                                      Offline_Uncorrectable(198)=202
/dev/sg44   /dev/sdaq    /dev/sg1    B11  WD-WMC190009962    GOOD
/dev/sg51   /dev/sdax    /dev/sg1    E11  Z4D01HKQ           GOOD
/dev/sg89   /dev/sdch    /dev/sg62   E04  9WM2DQ0K           FAILED:
                                      Offline_Uncorrectable(198)=435
/dev/sg64   /dev/sdbi    /dev/sg62   A01  YHH6EJ8A           FAILED:
                                      Reallocated_Sector_Count(5)=1552
/dev/sg65   /dev/sdbj    /dev/sg62   A02  YHGVVE8A           SUSPECT:
                                      Reallocated_Sector_Count(5)=17
/dev/sg68   /dev/sdbm    /dev/sg62   A03  YHH5N6MA           FAILED:
                                      Reallocated_Sector_Count(5)=2005
/dev/sg69   /dev/sdbn    /dev/sg62   A00  YHH5GWMA           FAILED:
                                      Reallocated_Sector_Count(5)=797
/dev/sg71   /dev/sdbp    /dev/sg62   A05  YHH6EWJA           FAILED:
                                      Reallocated_Sector_Count(5)=906
/dev/sg72   /dev/sdbq    /dev/sg62   A04  YHH682RA           SUSPECT:
                                      Reallocated_Sector_Count(5)=15
/dev/sg73   /dev/sdbr    /dev/sg62   D01  Z1F1CGH7           SUSPECT:
                                      Reallocated_Sector_Count(5)=16
```

SMART is not *that* smart, but you can work with it.

Example – Proactive Smarts

```
erik-riedels-macbook-pro:logs er1p$ cat 2014-*/halreport | grep SUSP
/dev/sg4    /dev/sdc     /dev/sg3   C00  YVJZ8XRK  SUSPECT: Reallocated(5)=99
/dev/sg49   /dev/sdav    /dev/sg2   D10  YVK6378A  SUSPECT: Reallocated(5)=35
/dev/sg45   /dev/sdaq    /dev/sg3   B10  YVJZW8EA  SUSPECT: Reallocated(5)=19
/dev/sg6    /dev/sde     /dev/sg3   A02  YVK4UJ5A  SUSPECT: Reallocated(5)=10
/dev/sg21   /dev/sdt     /dev/sg3   E02  YVJG6X4D  SUSPECT: Reallocated(5)=66
/dev/sg32   /dev/sdae    /dev/sg3   C05  YVK25MKA  SUSPECT: Reallocated(5)=78
/dev/sg35   /dev/sdag    /dev/sg3   A06  YVJYBDSA  SUSPECT: Reallocated(5)=43
/dev/sg15   /dev/sdn     /dev/sg3   D00  YVJB5TAA  SUSPECT: Reallocated(5)=42
/dev/sg58   /dev/sdbd    /dev/sg3   C07  YVJYRKYA  SUSPECT: Reallocated(5)=59

erik-riedels-macbook-pro:logs er1p$ cat 2014-*/halreport | grep FAIL
/dev/sg12   /dev/sdl     /dev/sg2   A04  YVJZMN3K  FAILED:  Reallocated(5)=110
/dev/sg60   /dev/sdbk    /dev/sg3   E08  YVK2GNRA  FAILED:  Reallocated(5)=1577
/dev/sg37   /dev/sdai    /dev/sg2   B09  YVJYR8KA  FAILED:  Reallocated(5)=101
/dev/sg41   /dev/sdam    /dev/sg3   B08  YVJEZT7A  FAILED:  Reallocated(5)=682

erik-riedels-macbook-pro:logs er1p$ cat 2014-*/halreport | grep GOOD | wc -l
   12227
```

GOOD	12,227
SUSPECT	9
BAD	4

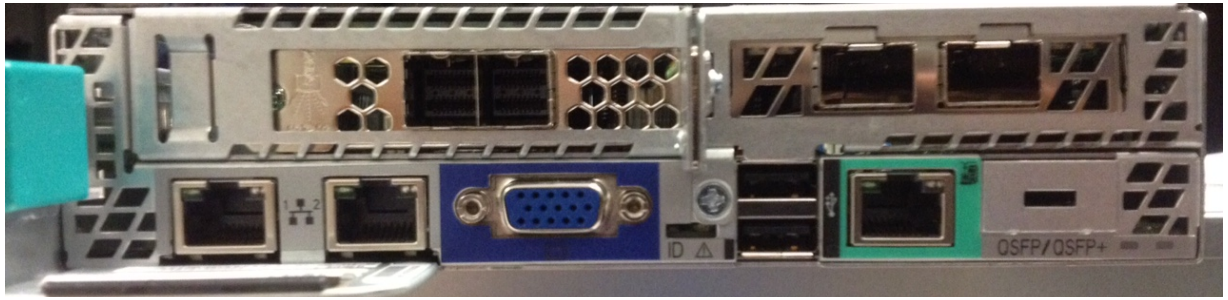
```
smartctl 5.40 2010-10-16 r3189 [x86_64-unknown-linux-gnu] (local build)
=== START OF INFORMATION SECTION ===
Model Family:     Hitachi Ultrastar 7K1000
Device Model:     HUA721010KLA330
Serial Number:    PBHBL6AF
User Capacity:    1,000,204,886,016 bytes
=== START OF READ SMART DATA SECTION ===
SMART overall-health self-assessment test result: FAILED!
Drive failure expected in less than 24 hours. SAVE ALL DATA.
Vendor Specific SMART Attributes with Thresholds:
ID# ATTRIBUTE_NAME          FLAG     TYPE      UPDATED  WHEN_FAILED
RAW_VALUE
5 Reallocated_Sector_Ct   0x0033     Pre-fail  Always   FAILING_NOW 9
9 Power_On_Hours          0x0012     Old_age   Always       -       13073
197 Current_Pending_Sector  0x0022   Old_age   Always       -       1890
198 Offline_Uncorrectable   0x0008   Old_age   Offline      -       9390
```

```
 FAILED:   Offline_Uncorrectable(198)=435
 FAILED:   Offline_Uncorrectable(198)=796
 FAILED:   Offline_Uncorrectable(198)=1385
 FAILED:   Offline_Uncorrectable(198)=2336
 FAILED:   Offline_Uncorrectable(198)=80961
 FAILED:   Reallocated_Sector_Count(5)=52472
 FAILED:   Reallocated_Sector_Count(5)=797
 FAILED:   Reallocated_Sector_Count(5)=906
 FAILED:   Reallocated_Sector_Count(5)=1552
 FAILED:   Reallocated_Sector_Count(5)=1814
 FAILED:   Reallocated_Sector_Count(5)=1818
 FAILED:   Reallocated_Sector_Count(5)=1886
 FAILED:   Reallocated_Sector_Count(5)=1944
 FAILED:   Reallocated_Sector_Count(5)=1999
 FAILED:   Reallocated_Sector_Count(5)=2005
 FAILED:   Reallocated_Sector_Count(5)=3270
 FAILED:   Reallocated_Sector_Count(5)=3809
 FAILED:   Reallocated_Sector_Count(5)=4094
 FAILED:   Reallocated_Sector_Count(5)=4095
 FAILED:   Reallocated_Sector_Count(5)=4281
 FAILED:   Reallocated_Sector_Count(5)=5119
```

keep on running on

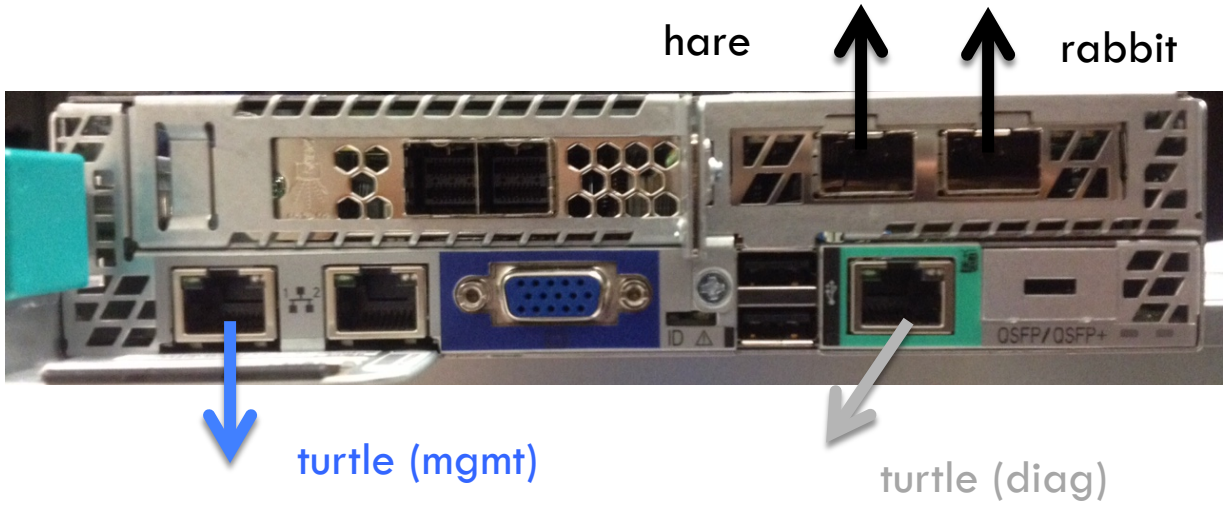
Even from this “very bad” disk with over 9,000 sector errors; we were able to recover >99% of the data with *ddrescue;* 9.5 MB out of 1 TB of data was permanently lost, with some difficulty reconstructing directories.

Networks



“We moved the cable from eth2 to eth3.”

```
provo-vegas:~ # ip add ls
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
2: slave-0: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc mq master public state UP
    link/ether 00:1e:67:9f:1b:3e
3: private: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9000 qdisc mq state UP group default qlen 1000
    link/ether 00:1e:67:69:6d:06
    inet 192.168.219.1/24 brd 192.168.219.255 scope global private
    inet 192.168.219.254/24 brd 192.168.219.255 scope global secondary private
4: unused-0: <BROADCmtu 1500 qdisc mq state DOWN group default qlen 1000
    link/ether 00:1e:67:69:6d:07
5: slave-1: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc mq master public state UP
    link/ether 00:1e:67:9f:1b:3e
6: public: <BROADCAST,MULTICAST,MASTER,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP group default
    link/ether 00:1e:67:9f:1b:3e
    inet 10.249.250.131/21 brd 10.249.255.255 scope global public
7: private.4@private: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP group
    link/ether 00:1e:67:69:6d:06
    inet 169.254.19.1/16 brd 169.254.255.255 scope global private.4
8: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state DOWN group default
    link/ether 56:84:7a:fe:97:99
    inet 172.17.42.1/16 scope global docker0
```



"We moved the cable from eth2 to eth3."

```
provo-vegas:~ # ip add ls
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
2: slave-0: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc mq master public state UP
    link/ether 00:1e:67:9f:1b:3e
3: private: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9000 qdisc mq state UP group default qlen 1000
    link/ether 00:1e:67:69:6d:06
    inet 192.168.219.1/24 brd 192.168.219.255 scope global private
    inet 192.168.219.254/24 brd 192.168.219.255 scope global secondary private
4: unused-0: <BROADCmtu 1500 qdisc mq state DOWN group default qlen 1000
    link/ether 00:1e:67:69:6d:07
5: slave-1: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc mq master public state UP
    link/ether 00:1e:67:9f:1b:3e
6: public: <BROADCAST,MULTICAST,MASTER,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP group default
    link/ether 00:1e:67:9f:1b:3e
    inet 10.249.250.131/21 brd 10.249.255.255 scope global public
7: private.4@private: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP group
    link/ether 00:1e:67:69:6d:06
    inet 169.254.19.1/16 brd 169.254.255.255 scope global private.4
8: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state DOWN group default
    link/ether 56:84:7a:fe:97:99
    inet 172.17.42.1/16 scope global docker0
```

```
provo-vegas:~ # ifconfig
docker0   Link encap:Ethernet  HWaddr 56:84:7A:FE:97:99
          inet addr:172.17.42.1  Bcast:0.0.0.0
          UP BROADCAST MULTICAST  MTU:1500  Metric:1

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:65536  Metric:1

private   Link encap:Ethernet  HWaddr 00:1E:67:69:6D:06
          inet addr:192.168.219.1  Bcast:192.168.219.255  Mask:255.255.255.0
          inet6 addr: fe80::21e:67ff:fe69:6d06/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:9000  Metric:1

private.4 Link encap:Ethernet  HWaddr 00:1E:67:69:6D:06
          inet addr:169.254.19.1  Bcast:169.254.255.255  Mask:255.255.0.0
          inet6 addr: fe80::21e:67ff:fe69:6d06/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1

public    Link encap:Ethernet  HWaddr 00:1E:67:9F:1B:3E
          inet addr:10.249.250.131  Bcast:10.249.255.255  Mask:255.255.248.0
          inet6 addr: fe80::21e:67ff:fe9f:1b3e/64 Scope:Link
          UP BROADCAST RUNNING MASTER MULTICAST  MTU:1500  Metric:1

slave-0   Link encap:Ethernet  HWaddr 00:1E:67:9F:1B:3E
          UP BROADCAST RUNNING SLAVE MULTICAST  MTU:1500  Metric:1

slave-1   Link encap:Ethernet  HWaddr 00:1E:67:9F:1B:3E
          UP BROADCAST RUNNING SLAVE MULTICAST  MTU:1500  Metric:1
```

“We moved the cable from eth2 to eth3.”

```
provo-vegas:~ # lsnet
1 lo 00:00:00:00:00:00 link unknown loopback  virtual
2 slave-0 00:1e:67:9f:1b:3e link up ether ixgbe
3 private 00:1e:67:69:6d:06 link up ether igb
4 unused-0 00:1e:67:69:6d:07 nolink down ether igb
5 slave-1 00:1e:67:9f:1b:3e link up ether ixgbe
6 public 00:1e:67:9f:1b:3e link up ether  virtual
7 private.4 00:1e:67:69:6d:06 link up ether  virtual
8 docker0 56:84:7a:fe:97:99 nolink down ether  virtual
```

```
provo-vegas:~ # ethtool -P slave-0
Permanent address: 00:1e:67:9f:1b:3e
provo-vegas:~ # ethtool -P slave-1
Permanent address: 00:1e:67:9f:1b:3f
```

```
provo-vegas:~ # ipmitool lan print
IP Address Source       : Static
IP Address              : 0.0.0.0
MAC Address             : 00:1e:67:69:6d:08
provo-vegas:~ # ipmitool lan print 2
IP Address Source       : Static
IP Address              : 0.0.0.0
MAC Address             : 00:1e:67:69:6d:09
provo-vegas:~ # ipmitool lan print 3
IP Address Source       : DHCP Address
IP Address              : 10.249.250.121
MAC Address             : 00:1e:67:69:6d:0a
SNMP Community String   : public
```

"We moved the cable from eth2 to eth3."

Topology Map – getrackinfo

```
provo-apricot:/home/emc # getrackinfo -a
```

Node private Ip Address	Node Id	Status	Public Mac	Ip Address	RMM Mac	Ip Address	Node Name
192.168.219.1	N/A	P	N/A	N/A	N/A	N/A	N/A
192.168.219.2	2	MA	00:1e:67:b5:af:84	10.249.249.72	00:1e:67:a3:ad:a4	10.249.249.62	sandy-apricot
192.168.219.3	3	SA	00:1e:67:b5:9f:78	10.249.249.73	00:1e:67:a3:af:43	10.249.249.63	orem-apricot
192.168.219.4	N/A	P	N/A	N/A	N/A	N/A	N/A
192.168.219.5	5	SA	00:1e:67:b5:9e:80	10.249.249.75	00:1e:67:6a:1c:e1	10.249.249.65	layton-apricot
192.168.219.6	6	SA	00:1e:67:b5:ad:40	10.249.249.76	00:1e:67:6a:23:fd	10.249.249.66	logan-apricot
192.168.219.7	N/A	P	N/A	N/A	N/A	N/A	N/A
192.168.219.8	8	SA	00:1e:67:b5:ad:6c	10.249.249.78	00:1e:67:6a:1a:a7	10.249.249.68	murray-apricot
192.168.219.9	9	SA	00:1e:67:b5:b1:e0	10.249.249.91	00:1e:67:6a:24:3e	10.249.249.81	boston-apricot
192.168.219.10	10	SA	00:1e:67:b5:b2:a8	10.249.249.92	00:1e:67:6a:1c:28	10.249.249.82	chicago-apricot
192.168.219.11	11	SA	00:1e:67:b5:b9:48	10.249.249.93	00:1e:67:6a:18:b3	10.249.249.83	houston-apricot
192.168.219.12	12	SA	00:1e:67:b5:b2:d0	10.249.249.94	00:1e:67:6a:1b:b5	10.249.249.84	phoenix-apricot
192.168.219.13	13	SA	00:1e:67:b5:b1:34	10.249.249.95	00:1e:67:6a:04:a4	10.249.249.85	dallas-apricot
192.168.219.14	14	SA	00:1e:67:b5:b1:44	10.249.249.96	00:1e:67:6a:0e:6d	10.249.249.86	detroit-apricot
192.168.219.15	15	SA	00:1e:67:b5:b1:30	10.249.249.97	00:1e:67:a3:b5:3d	10.249.249.87	columbus-apricot
192.168.219.16	16	SA	00:1e:67:b5:a1:b8	10.249.249.98	00:1e:67:a3:ae:7b	10.249.249.88	austin-apricot
192.168.219.17	17	SA	00:1e:67:b5:b5:a8	10.249.249.111	00:1e:67:a3:b1:05	10.249.249.101	memphis-apricot
192.168.219.18	18	SA	00:1e:67:b5:a1:e8	10.249.249.112	00:1e:67:a3:ba:0b	10.249.249.102	seattle-apricot
192.168.219.19	N/A	P	N/A	N/A	N/A	N/A	N/A
192.168.219.20	N/A	P	N/A	N/A	N/A	N/A	N/A
192.168.219.21	N/A	P	N/A	N/A	N/A	N/A	N/A
192.168.219.22	22	SA	00:1e:67:b5:ac:90	10.249.249.116	00:1e:67:69:ed:b1	10.249.249.106	atlanta-apricot
192.168.219.23	23	SA	00:1e:67:b5:9f:e4	10.249.249.117	00:1e:67:69:ef:cd	10.249.249.107	fresno-apricot
192.168.219.24	24	SA	00:1e:67:b5:a2:80	10.249.249.118	00:1e:67:6a:20:2e	10.249.249.108	mesa-apricot

```
Status:
M - Master, S - Slave
E - Epoxy
I - Initializing, U - Updating, A - Active
P - On, O - Off
```

Topology Map – getrackinfo – details

```
provo-vanilla:~ # getrackinfo -a
Node private      Node              Public                                RMM
Ip Address        Id       Status   Mac                 Ip Address        Mac                 Ip Address        Node Name
===============   ======   ======   =================   ================  =================   ================  =========
192.168.219.1     1        MA       00:1e:67:9f:01:96   10.249.248.111    00:1e:67:69:29:8f   10.249.248.101    provo-vanilla
192.168.219.2     2        SA       00:1e:67:9f:01:a2   10.249.248.112    00:1e:67:69:28:72   10.249.248.102    sandy-vanilla
192.168.219.3     3        SA       00:1e:67:9e:ff:9e   10.249.248.113    00:1e:67:69:29:99   10.249.248.103    orem-vanilla
192.168.219.4     N/A      noLink   N/A                 N/A               N/A                 N/A               N/A
192.168.219.5     N/A      noLink   N/A                 N/A               N/A                 N/A               N/A
192.168.219.6     N/A      noLink   N/A
192.168.219.7     N/A      noLink   N/A
192.168.219.8     N/A      noLink   N/A
192.168.219.9     N/A      O        N/A
192.168.219.10    N/A      O        N/A
192.168.219.11    N/A      O        N/A
192.168.219.12    N/A      O        N/A
192.168.219.13    N/A      O        N/A
192.168.219.14    N/A      O        N/A
192.168.219.15    N/A      O        N/A
192.168.219.16    N/A      O        N/A
192.168.219.17    N/A      noLink   N/A
192.168.219.18    N/A      noLink   N/A
192.168.219.19    N/A      noLink   N/A
192.168.219.20    N/A      noLink   N/A
192.168.219.21    N/A      noLink   N/A
192.168.219.22    N/A      noLink   N/A
192.168.219.23    N/A      noLink   N/A
192.168.219.24    N/A      noLink   N/A
Status:
M - Master, S - Slave
E - Epoxy
I - Initializing, U - Updating, A - Active
P - On, O - Off
```

```
provo-vanilla:~ # getrackinfo -v
=====================================================
NodeName              : provo-vanilla
Node Id               : 1

Interfaces(MAC & IP)
--------------------
public                : 00:1e:67:9f:01:96   10.249.248.111/21
private               : 00:1e:67:69:29:8b   192.168.219.1/24
private ipmi          : 00:1e:67:69:29:8d   192.168.219.101/24
private.4(NAN)        : 00:1e:67:69:29:8b   169.254.186.1/16
remote ipmi           : 00:1e:67:69:29:8f   10.249.248.101/21

Network Services
----------------
NTP Configuration:
     server:    10.254.140.21 10.254.140.22
DNS Configuration:
     domain:
     search:    sea.lab.emc.com corp.emc.com emc.com
     server:    192.168.219.254 10.6.149.11
```

But Wait, There's More

```
silver-is1-004:~ % cs_hal info node
Node                  : silver-is1-004
BIOS date             : 06/20/2012
BIOS version          : SE5C600.86B.01.03.0002.062020121504
Board model           : S2600JF
Board S/N             : QSJP23007313
Board vendor          : Intel Corporation
Board version         : G28033-506
Chassis S/N           : FC6ND131900019
Chassis vendor        : ..............................
Chassis model         : S2600JF
System S/N            : FC6AT131900005
Processor count       : 8
Total memory          : 23.0433GB
Available memory      : 17.7322GB
Total swap            : 2GB
Available swap        : 2GB
Shared memory         : 0GB
Host adapter count    : 2
Net interface count   : 4
Enclosure count       : 1
External disk count   : 60
```

But wait, there's MORE – BIOS, BIOS settings, BMC firmware, BMC settings, power supply firmware, fan firmware, HBA firmware, HBA NVDATA, enclosure firmware, enclosure power supply firmware, enclosure fan firmware, …

Release Notes: "This BIOS update fixes a problem where BIOS update fails 20%-40% of the time."

Summary

What We Did

- kept it simple, took control
  - no ~~hardware RAID~~; no ~~databases~~; no ~~events~~ (poll)
  - sg, sd, ~~md, dm,~~ lvm, fs (~~ext3, ext4~~, xfs, ~~btrfs~~)
- built a library – HAL – hardware abstraction layer
  - common library for our app-level services to use
- built some tools – *cs-hal* (for Support to use)
  - `cs-hal list disks`
  - `cs-hal list fs`
  - `cs-hal info sg27`
  - `cs-hal led Z1Z0EVBF blink`
  - `cs-hal led sg27 blink`

Biggest Take-Aways

- when you design a solution for a single machine …
- … think about the poor sap who has to
  - diagnose 200 nodes/machines – real machines
  - … 12,000 drives – real drives
  - … 12,000 file systems (or more) – w/ real customer data
  - … from 5,000 miles away – real miles
  - … in the middle of the night – really dark
  - … all week long
  - … especially on Friday nights

Other Conclusions

- most Linux + tools developers don't have 50+ disks on their systems
- where did `/dev/sddh` come from?
  - device briefly offline => new dev!!
  - hardware comes & goes, the software stays the same
- disks don't have file descriptors
  - sg, sd, md, dm, lvm, fs (ext3, ext4, xfs, btrfs)
- SATA disks are big & cheap and all, but can be a bit "unruly"... temporary disconnects
- hardware RAID is yucky
- databases are often stale
  - trust, but verify => don't trust, skip right to verify

Containers

- All magic comes with a price

http://www.spreadshirt.com/ouat-quote-all-magic-comes-with-a-price

Q & A

erik.riedel@emc.com

from flickr/purplemattfish, Broken hard drive?

Build on 20 Years of Storage Research

- APIs vs. mount points – "no slashes required"
  - blocks vs. files vs. objects vs. "APIs"
- App-driven and policy-automated
  - self-configuring, self-organizing, self-tuning, self-*
- Built in data services
  - self-healing
- Unlimited namespace, dynamic
  - billions and billions of objects, large and small
- Native multi-tenancy
  - security/auth, monitoring, resource isolation

MORE EXAMPLES

Example – DAE reconnects

```
Jul  1 21:37:37 localhost kernel: mptbase  ioc0  LogInfo(0x31130000) Code={IO Not Yet Executed}, SubCode(0x0000)
Jul  1 23:50:06 localhost kernel: mptbase  ioc1  LogInfo(0x31112000) Code={Reset}, SubCode(0x2000)
Jul  1 23:50:09 localhost kernel: mptbase  ioc1  LogInfo(0x31112000) Code={Reset}, SubCode(0x2000)
Jul  1 23:50:12 20xx : WARNING : Disk Event : Disk is moved to DAE: Slot ID: 0 : Serial NO: WCAVY4897042
Jul  1 23:50:12 20xx : WARNING : Disk Event : Disk is moved to DAE: Slot ID: 0 : Serial NO: WCAVY5192630
Jul  1 23:50:13 20xx : WARNING : Disk Event : Disk is moved to DAE: Slot ID: 0 : Serial NO: WCAVY5186052
Jul  1 23:50:14 20xx : WARNING : Disk Event : Disk is moved to DAE: Slot ID: 0 : Serial NO: WCAVY3550485
Jul  1 23:50:14 20xx : WARNING : Disk Event : Disk is moved to DAE: Slot ID: 0 : Serial NO: WCAVY360702
(…all 60 disks…)
Jul  1 23:50:15 20xx : ERROR : DAE Event : DAE (device path: /dev/sg66) lost. : Serial NO: , Device path: /dev/sg66, Device ID:
     5000097a780747be
Jul  1 23:50:15 20xx : WARNING : Disk Event : Disk is moved to DAE: Slot ID: 0 : Serial NO: WCAVY5349410
Jul  1 23:51:14 20xx : INFO : DAE Event : New DAE (device path: /dev/sg66) is added. : Serial NO: , Device path: /dev/sg66, Device
     ID: 5000097a780747be
Jul  1 23:51:14 20xx : WARNING : Disk Event : Disk is moved to DAE: 5f4ad992-724e-48af-8cac-a68b7d859593 Slot ID: 11 : Serial NO:
     WCAVY5182031 , Device path: /dev/sdaq, Slot ID:
Jul  1 23:51:14 20xx : WARNING : Disk Event : Disk is moved to DAE: 5f4ad992-724e-48af-8cac-a68b7d859593 Slot ID: 13 : Serial NO:
     WCAVY5186052 , Device path: /dev/sdas, Slot ID:
(…all 60 disks…)
Jul  1 23:51:16 20xx : WARNING : Disk Event : Disk is moved to DAE: e70905ad-5736-48d9-8a1b-a15a2d116825 Slot ID: 4 : Serial NO:
     WCAVY5349410 , Device path: /dev/sday, Slot ID:
(outage ends, log ends)
```

Reset on the SAS/SATA bus (expander), enclosure identifiers re-assigned to “<NULL>”; enclosure returns after 68 seconds, disks are assigned back where they belong. Entire episode lasts 70 seconds. BUT system management database remembers this “event” for weeks.

HAL – disk view (15 drive node)

```
dino-black:~ % cs_hal list disks
Disks(s):
SCSI Device Block Device Enclosure   Slot Serial Number       SMART Status
----------- ------------ ----------- ---- ------------------- ------------
n/a         /dev/sda     RAID vol    n/a  not supported       n/a
/dev/sg0    n/a          RAID array  0    9QE801ME            GOOD
/dev/sg1    n/a          RAID array  1    9QE834TG            GOOD
/dev/sg3    /dev/sdb     /dev/sg18   0    9WM0R49P            GOOD
/dev/sg4    /dev/sdc     /dev/sg18   1    9WM0R48T            GOOD
/dev/sg5    /dev/sdd     /dev/sg18   2    9WM0R3Z4            GOOD
/dev/sg6    /dev/sde     /dev/sg18   3    9WM0R4VK            SUSPECT:          Reallocated(5)=19
/dev/sg7    /dev/sdf     /dev/sg18   4    9WM0RF21            GOOD
/dev/sg8    /dev/sdg     /dev/sg18   5    9WM0R44B            GOOD
/dev/sg9    /dev/sdh     /dev/sg18   6    9WM0R3E0            GOOD
/dev/sg10   /dev/sdi     /dev/sg18   7    9WM0RF2X            GOOD
/dev/sg11   /dev/sdj     /dev/sg18   8    9WM0R4TX            GOOD
/dev/sg12   /dev/sdk     /dev/sg18   9    9WM0REHK            GOOD
/dev/sg13   /dev/sdl     /dev/sg18   10   9WM0R3EW            GOOD
/dev/sg14   /dev/sdm     /dev/sg18   11   9WM0R4GY            GOOD
/dev/sg15   /dev/sdn     /dev/sg18   12   9WM0R4NZ            GOOD
/dev/sg16   /dev/sdo     /dev/sg18   13   9WM0RF42            GOOD
/dev/sg17   /dev/sdp     /dev/sg18   14   9WM0R3AS            GOOD

 RAID array: 2
   external: 15

total disks: 17
```

HAL – filesystem view (15 drive node)

```
dino-black:~ % cs_hal list fs
Volume(s):
SCSI Device Block Device FS UUID                                Type      Slot  Label SMART               Mount Point
----------- ------------ ------------------------------------- --------- ----- ----- ------              ------------
/dev/sg2    /dev/sda     0ddb9635-ff27-4cd3-8c2f-58a6f5226d30  ext3            BOOT  GOOD /boot
/dev/sg2    /dev/sda     2192b3ef-2a44-4450-9b04-327c00215454  xfs                   GOOD /root2
/dev/sg2    /dev/sda     ffa9607a-4b6f-4218-9266-c083fb1989a1  xfs                   GOOD /var
/dev/sg2    /dev/sda     746b09d4-f07a-49dc-8b40-86220dfc7edc  xfs                   GOOD /
/dev/sg2    /dev/sda     f7c37c92-4bc5-4abf-95a5-efa51c46f6bc  swap v1               GOOD -
/dev/sg3    /dev/sdb     90a52650-e0f3-49e4-810b-a505cdcadb51  xfs       0           GOOD /data-disks/ss-90a52650-e0f3-49e4-810b-a505cdcadb51
/dev/sg4    /dev/sdc     173aef8b-80e9-4be2-a510-3b88d3343f8a  xfs       1           GOOD /data-disks/ss-173aef8b-80e9-4be2-a510-3b88d3343f8a
/dev/sg5    /dev/sdd     bcfb1897-152b-482b-bde6-de9665ad7c51  xfs       2           GOOD /data-disks/ss-bcfb1897-152b-482b-bde6-de9665ad7c51
/dev/sg6    /dev/sde     bc6946ae-770f-4621-9ea5-f2d1e5ec0f28  xfs       3         SUSPECT /data-disks/ss-bc6946ae-770f-4621-9ea5-f2d1e5ec0f28
/dev/sg7    /dev/sdf     52446742-a566-4036-8b0c-5cd7901474f0  xfs       4           GOOD /data-disks/ss-52446742-a566-4036-8b0c-5cd7901474f0
/dev/sg8    /dev/sdg     c9ee0971-d8dc-4621-8958-d79890d0f590  xfs       5           GOOD /data-disks/ss-c9ee0971-d8dc-4621-8958-d79890d0f590
/dev/sg9    /dev/sdh     294bcd25-ab19-40ee-8c03-cd71e94e9e06  xfs       6           GOOD /meta/294bcd25-ab19-40ee-8c03-cd71e94e9e06
/dev/sg10   /dev/sdi     cb5cac6c-1cdf-49ec-8754-a475db3d4afd  xfs       7           GOOD /data-disks/ss-cb5cac6c-1cdf-49ec-8754-a475db3d4afd
/dev/sg11   /dev/sdj     91739495-2a46-47d2-8676-d8b4b3f8fd76  xfs       8           GOOD /data-disks/ss-91739495-2a46-47d2-8676-d8b4b3f8fd76
/dev/sg12   /dev/sdk     9f2a0ae1-d97b-4fb1-873e-6a9bfb2c3254  xfs       9           GOOD /data-disks/ss-9f2a0ae1-d97b-4fb1-873e-6a9bfb2c3254
/dev/sg13   /dev/sdl     404a8c5a-19c0-4949-bd33-edd83ca4ee8f  xfs       10          GOOD /meta/404a8c5a-19c0-4949-bd33-edd83ca4ee8f
/dev/sg14   /dev/sdm     da36046f-41f7-46d4-bcaa-af183002b792  xfs       11          GOOD /data-disks/ss-da36046f-41f7-46d4-bcaa-af183002b792
/dev/sg15   /dev/sdn     a71b6937-8ae5-4a37-96d0-78feeb0e62c4  xfs       12          GOOD /data-disks/ss-a71b6937-8ae5-4a37-96d0-78feeb0e62c4
/dev/sg16   /dev/sdo     34d6f5c5-1f5d-4cea-af5a-af157324aee8  xfs       13          GOOD /meta/34d6f5c5-1f5d-4cea-af5a-af157324aee8
/dev/sg17   /dev/sdp     9cc59415-cab5-4456-881f-a0c533e1823d  xfs       14          GOOD /data-disks/ss-9cc59415-cab5-4456-881f-a0c533e1823d

total: 21
```

```
layton-copper:~ % cs_hal list disks
Disks(s):
SCSI Device Block Device Enclosure   Slot Serial Number       SMART Status
---------- ------------ ---------- ---- ------------------- ------------
n/a        /dev/md126   RAID vol   n/a  not supported       n/a
/dev/sg1   n/a          RAID array 1    PQKJGZNB            GOOD
/dev/sg0   n/a          RAID array 0    PQKHYT9B            GOOD
/dev/sg26  /dev/sdz     /dev/sg2   C04  WMAW30330711        GOOD
/dev/sg27  /dev/sdaa    /dev/sg2   D04  WMAW30130282        GOOD
/dev/sg28  /dev/sdab    /dev/sg2   E05  WMAW30331465        GOOD
/dev/sg29  /dev/sdac    /dev/sg2   E04  WMAW30400512        GOOD
/dev/sg30  /dev/sdad    /dev/sg2   B05  WMAW30330840        GOOD
/dev/sg31  /dev/sdae    /dev/sg2   C05  WMAW30283365        GOOD
/dev/sg32  /dev/sdaf    /dev/sg2   D05  WMAW30331280        GOOD
/dev/sg3   /dev/sdc     /dev/sg2   C00  WMAW30330725        GOOD
/dev/sg4   /dev/sdd     /dev/sg2   A01  WMAW30330535        GOOD
/dev/sg5   /dev/sde     /dev/sg2   A02  WMAW30330800        GOOD
/dev/sg6   /dev/sdf     /dev/sg2   B00  WMAW30331330        GOOD
/dev/sg7   /dev/sdg     /dev/sg2   C01  WMAW30128826        GOOD
/dev/sg8   /dev/sdh     /dev/sg2   A03  WMAW30199450        GOOD
/dev/sg9   /dev/sdi     /dev/sg2   A00  WMAW30103257        GOOD
/dev/sg10  /dev/sdj     /dev/sg2   B01  WMAW30331487        GOOD
/dev/sg11  /dev/sdk     /dev/sg2   A05  WMAW30327185        GOOD
/dev/sg12  /dev/sdl     /dev/sg2   A04  WMAW30327102        GOOD
/dev/sg13  /dev/sdm     /dev/sg2   D01  WMAW30330859        GOOD
/dev/sg14  /dev/sdn     /dev/sg2   D00  WMAW30331130        GOOD
/dev/sg15  /dev/sdo     /dev/sg2   C02  WMAW30331192        GOOD
/dev/sg16  /dev/sdp     /dev/sg2   D02  WMAW30307529        GOOD
/dev/sg17  /dev/sdq     /dev/sg2   E00  WMAW30196937        GOOD
/dev/sg18  /dev/sdr     /dev/sg2   B02  WMAW30331240        GOOD
/dev/sg19  /dev/sds     /dev/sg2   E01  WCAW32612222        GOOD
/dev/sg20  /dev/sdt     /dev/sg2   E02  WMAW30331427        GOOD
```

HAL – disk view
(60 drive node)

```
layton-copper:~ % cs_hal list fs
Volume(s):
```

SCSI Device	Block Device	FS UUID	Type	Slot	Label	SMART	Mount Point
/dev/sg0	/dev/sda	6cf8c9cb-c0c9-498c-ab3f-28140dd66f09	ext3	0	BOOT	GOOD	-
/dev/sg1	/dev/sdb	6cf8c9cb-c0c9-498c-ab3f-28140dd66f09	ext3	1	BOOT	GOOD	-
/dev/sg26	/dev/sdz	c198e38d-41a1-4263-b46a-39bbdc8ed89c	xfs	C04		GOOD	/data-disks/ss-c198e38d-41a1-4263-b46a-39bbdc8ed89c
/dev/sg27	/dev/sdaa	3429b68b-f599-4679-991a-5b98549b2431	xfs	D04		GOOD	/meta/3429b68b-f599-4679-991a-5b98549b2431
/dev/sg28	/dev/sdab	1fccea68-439f-4a8e-be55-a81fd17774bf	xfs	E05		GOOD	/meta/1fccea68-439f-4a8e-be55-a81fd17774bf
/dev/sg29	/dev/sdac	e520b436-35ef-40d1-bd3b-d6d42957bc41	xfs	E04		GOOD	/data-disks/ss-e520b436-35ef-40d1-bd3b-d6d42957bc4
/dev/sg30	/dev/sdad	12c13240-2957-4b7b-b628-df870a6fbd3b	xfs	B05		GOOD	/meta/12c13240-2957-4b7b-b628-df870a6fbd3b
/dev/sg31	/dev/sdae	7e00293c-1069-45c0-bc4e-2f7c7cd52a7b	xfs	C05		GOOD	/meta/7e00293c-1069-45c0-bc4e-2f7c7cd52a7b
/dev/sg32	/dev/sdaf	7dec91ad-4985-4ce5-898c-fe491d5818af	xfs	D05		GOOD	/meta/7dec91ad-4985-4ce5-898c-fe491d5818af
/dev/sg3	/dev/sdc	05705250-0a35-4618-95da-64d0632395fc	xfs	C00		GOOD	/data-disks/ss-05705250-0a35-4618-95da-64d0632395fc
/dev/sg4	/dev/sdd	05b98c0c-c77e-4a90-bcec-e5874cf89988	xfs	A01		GOOD	/data-disks/ss-05b98c0c-c77e-4a90-bcec-e5874cf89988
/dev/sg5	/dev/sde	42d87a05-4f8e-4375-8547-909f597fdaf5	xfs	A02		GOOD	/data-disks/ss-42d87a05-4f8e-4375-8547-909f597fdaf5
/dev/sg6	/dev/sdf	eb8657cc-b681-4698-805c-86fbd82fbccc	xfs	B00		GOOD	/data-disks/ss-eb8657cc-b681-4698-805c-86fbd82fbccc
/dev/sg7	/dev/sdg	1c15a217-418e-48e6-85a2-cb058c63a26f	xfs	C01		GOOD	/data-disks/ss-1c15a217-418e-48e6-85a2-cb058c63a26f
/dev/sg8	/dev/sdh	cd762f32-19c6-46f0-919d-bdde85261d98	xfs	A03		GOOD	/data-disks/ss-cd762f32-19c6-46f0-919d-bdde85261d98
/dev/sg9	/dev/sdi	f29d89c8-c0c7-4ec3-9645-de1d58b2a1cd	xfs	A00		GOOD	/data-disks/ss-f29d89c8-c0c7-4ec3-9645-de1d58b2a1cd
/dev/sg10	/dev/sdj	bc18fc92-9676-48e4-817c-47b10df3ee7a	xfs	B01		GOOD	/data-disks/ss-bc18fc92-9676-48e4-817c-47b10df3ee7a
/dev/sg11	/dev/sdk	d6f8f279-fc48-466c-9db0-ec41064e0b9e	xfs	A05		GOOD	/data-disks/ss-d6f8f279-fc48-466c-9db0-ec41064e0b9e
/dev/sg12	/dev/sdl	8a38f4b7-bf8c-47fe-a99c-d31fe53b6d1e	xfs	A04		GOOD	/data-disks/ss-8a38f4b7-bf8c-47fe-a99c-d31fe53b6d1e
/dev/sg13	/dev/sdm	55ceca7a-8df1-4eb5-a5b3-003a4fa68c36	xfs	D01		GOOD	/data-disks/ss-55ceca7a-8df1-4eb5-a5b3-003a4fa68c36
/dev/sg14	/dev/sdn	40d95e6d-b410-4f3b-bbcb-15f163b63486	xfs	D00		GOOD	/data-disks/ss-40d95e6d-b410-4f3b-bbcb-15f163b63486
/dev/sg15	/dev/sdo	a865b961-4406-4bd8-91ab-4be9d446712e	xfs	C02		GOOD	/data-disks/ss-a865b961-4406-4bd8-91ab-4be9d446712e
/dev/sg16	/dev/sdp	04e94a2a-c01a-4e06-bbe9-41da0ef1a293	xfs	D02		GOOD	/data-disks/ss-04e94a2a-c01a-4e06-bbe9-41da0ef1a293
/dev/sg17	/dev/sdq	1d9051a7-fe09-4b98-bae1-4385bb1ee08c	xfs	E00		GOOD	/data-disks/ss-1d9051a7-fe09-4b98-bae1-4385bb1ee08c
/dev/sg18	/dev/sdr	9a9f43d7-920b-4197-b388-e9a85b953f4b	xfs	B02		GOOD	/data-disks/ss-9a9f43d7-920b-4197-b388-e9a85b953f4b
/dev/sg19	/dev/sds	4b00c0fb-5bb7-4bfe-af6d-c4fba1721db6	xfs	E01		GOOD	/data-disks/ss-4b00c0fb-5bb7-4bfe-af6d-c4fba1721db6
/dev/sg20	/dev/sdt	ff2d72f8-49aa-4983-a666-b8702fee6916	xfs	E02		GOOD	/data-disks/ss-ff2d72f8-49aa-4983-a666-b8702fee6916
/dev/sg21	/dev/sdu	e04bf3c3-cff8-4316-af77-d1e49a0b26cd	xfs	B03		GOOD	/data-disks/ss-e04bf3c3-cff8-4316-af77-d1e49a0b26cd
/dev/sg22	/dev/sdv	d92bca38-296b-45c1-8291-256eebe2b764	xfs	D03		GOOD	/data-disks/ss-d92bca38-296b-45c1-8291-256eebe2b764
/dev/sg23	/dev/sdw	852bf5d8-a06a-4df8-804e-635364abb7d9	xfs	C03		GOOD	/data-disks/ss-852bf5d8-a06a-4df8-804e-635364abb7d9
/dev/sg24	/dev/sdx	c19c43d2-f084-4d65-8a63-ec40c90f6e54	xfs	E03		GOOD	/data-disks/ss-c19c43d2-f084-4d65-8a63-ec40c90f6e54
/dev/sg25	/dev/sdy	4af383d9-71a6-4324-84d6-d2e854900a71	xfs	B04		GOOD	/data-disks/ss-4af383d9-71a6-4324-84d6-d2e854900a71
/dev/sg57	/dev/sdbd	c8343213-f695-4e9b-92c0-106787ea0f40	xfs	C07		GOOD	/data-disks/ss-c8343213-f695-4e9b-92c0-106787ea0f40
/dev/sg58	/dev/sdbe	afc73d9c-1a89-4a62-8536-4410899818ec	xfs	E06		GOOD	/data-disks/ss-afc73d9c-1a89-4a62-8536-4410899818ec
/dev/sg59	/dev/sdbf	99fb488c-7689-4adc-aa13-7af8d5cd91ba	xfs	E08		GOOD	/data-disks/ss-99fb488c-7689-4adc-aa13-7af8d5cd91ba
/dev/sg60	/dev/sdbg	27b3025b-c3f2-4016-8094-c7eeb355f7d4	xfs	D06		GOOD	/data-disks/ss-27b3025b-c3f2-4016-8094-c7eeb355f7d4
/dev/sg61	/dev/sdbh	6660e770-c8fb-46fd-a628-6c485e20ebc0	xfs	C06		GOOD	/data-disks/ss-6660e770-c8fb-46fd-a628-6c485e20ebc0
/dev/sg62	/dev/sdbi	80ddb764-8337-4ef1-9a0d-e6f66405537f	xfs	D07		GOOD	/data-disks/ss-80ddb764-8337-4ef1-9a0d-e6f66405537f
/dev/sg63	/dev/sdbj	e0614cdd-0662-4845-9c31-ebd93121117e	xfs	E07		GOOD	/data-disks/ss-e0614cdd-0662-4845-9c31-ebd93121117e
/dev/sg34	/dev/sdag	c45cf761-4630-4076-99f5-fe5bbc1eb664	xfs	A06		GOOD	/data-disks/ss-c45cf761-4630-4076-99f5-fe5bbc1eb664
/dev/sg35	/dev/sdah	ad9157f0-6382-46fa-899c-5439d84ac64d	xfs	A07		GOOD	/data-disks/ss-ad9157f0-6382-46fa-899c-5439d84ac64d
/dev/sg36	/dev/sdai	5b1d8019-afae-4cdc-9d6c-ccc66c764cc8	xfs	B09		GOOD	/meta/5b1d8019-afae-4cdc-9d6c-ccc66c764cc8
/dev/sg37	/dev/sdaj	0a73ec0d-087d-413f-9cfd-adaf952467a8	xfs	A08		GOOD	/meta/0a73ec0d-087d-413f-9cfd-adaf952467a8
/dev/sg38	/dev/sdak	abb4d427-f891-4af4-a79a-5795a5c2f1d1	xfs	A09		GOOD	/data-disks/ss-abb4d427-f891-4af4-a79a-5795a5c2f1d1
/dev/sg39	/dev/sdal	ff4a6afd-12f2-42cd-8efb-e49d691c0b9d	xfs	A10		GOOD	/meta/ff4a6afd-12f2-42cd-8efb-e49d691c0b9d
/dev/sg40	/dev/sdam	69a19693-609e-4d5e-8482-6de57fa5946e	xfs	B08		GOOD	/meta/69a19693-609e-4d5e-8482-6de57fa5946e
/dev/sg41	/dev/sdan	442e5f89-c528-46fe-8b5a-6a6b01ccf359	xfs	B07		GOOD	/meta/442e5f89-c528-46fe-8b5a-6a6b01ccf359
/dev/sg42	/dev/sdao	8d9052ab-0d4c-4fc5-92ea-e128318d0c21	xfs	B06		GOOD	/meta/8d9052ab-0d4c-4fc5-92ea-e128318d0c21
/dev/sg43	/dev/sdap	04bed093-5748-44d6-a9a0-6e9efee05dac	xfs	A11		GOOD	/data-disks/ss-04bed093-5748-44d6-a9a0-6e9efee05dac
/dev/sg44	/dev/sdaq	a3554dea-8043-43cc-804d-4460860a69f7	xfs	B10		GOOD	/data-disks/ss-a3554dea-8043-43cc-804d-4460860a69f7
/dev/sg45	/dev/sdar	a5eab0f3-4780-46fb-a0e2-f363f0f842f3	xfs	B11		GOOD	/data-disks/ss-a5eab0f3-4780-46fb-a0e2-f363f0f842f3
/dev/sg46	/dev/sdas	af4815c2-3ae8-4787-bb90-abc9a8cac8a9	xfs	C11		GOOD	/data-disks/ss-af4815c2-3ae8-4787-bb90-abc9a8cac8a9
/dev/sg47	/dev/sdat	2ffee3bd-866d-432e-ae7d-d7e4b264fea7	xfs	D11		GOOD	/data-disks/ss-2ffee3bd-866d-432e-ae7d-d7e4b264fea7
/dev/sg48	/dev/sdau	87beea7d-0d01-4418-b120-0b83b6edac81	xfs	C10		GOOD	/data-disks/ss-87beea7d-0d01-4418-b120-0b83b6edac81
/dev/sg49	/dev/sdav	5614a615-fca3-4ab8-8e1f-7e7ddfa9fe0a	xfs	D10		GOOD	/data-disks/ss-5614a615-fca3-4ab8-8e1f-7e7ddfa9fe0a
/dev/sg50	/dev/sdaw	f1148778-f1bd-45c0-9dd1-bafd6c5ffcad	xfs	C09		GOOD	/data-disks/ss-f1148778-f1bd-45c0-9dd1-bafd6c5ffcad
/dev/sg51	/dev/sdax	31aa9f31-c6af-4370-be8c-4726b31341ac	xfs	D09		GOOD	/meta/31aa9f31-c6af-4370-be8c-4726b31341ac
/dev/sg52	/dev/sday	555804d0-4a2a-488e-a92f-be55aa61da37	xfs	E11		GOOD	/data-disks/ss-555804d0-4a2a-488e-a92f-be55aa61da37
/dev/sg53	/dev/sdaz	9d1fe14a-9b03-4918-ab80-febbc960cf9e	xfs	E10		GOOD	/data-disks/ss-9d1fe14a-9b03-4918-ab80-febbc960cf9e

HAL – filesystem view (60 drive node)

HAL – node

```
silver-is1-004:~ % cs_hal list disks
Disks(s):
SCSI Device Block Device Enclosure   Slot Serial Number       SMART Status
----------- ------------ ----------- ---- ------------------- ------------
n/a         /dev/md126   RAID vol    n/a  not supported       n/a
/dev/sg1    n/a          RAID array  1    KLH6DNZJ            GOOD
/dev/sg0    n/a          RAID array  0    KLH6DL7J            GOOD
/dev/sg27   /dev/sdy     /dev/sg2    B04  Z1Z0EVBF            GOOD
/dev/sg28   /dev/sdz     /dev/sg2    C04  Z1Z0EKFZ            GOOD
/dev/sg29   /dev/sdaa    /dev/sg2    D04  Z1Z0ETMY            GOOD
/dev/sg30   /dev/sdab    /dev/sg2    E05  Z1Z0EVLG            GOOD
/dev/sg31   /dev/sdac    /dev/sg2    E04  Z1Z0EVH9            GOOD
. . .
. . .
. . .
/dev/sg47   /dev/sdas    /dev/sg2    C11
/dev/sg48   /dev/sdat    /dev/sg2    D11
/dev/sg49   /dev/sdau    /dev/sg2    C10
/dev/sg50   /dev/sdav    /dev/sg2    D10
/dev/sg51   /dev/sdaw    /dev/sg2    C09
/dev/sg52   /dev/sdax    /dev/sg2    D09
/dev/sg53   /dev/sday    /dev/sg2    E11
/dev/sg54   /dev/sdaz    /dev/sg2    E10
/dev/sg55   /dev/sdba    /dev/sg2    E09
/dev/sg56   /dev/sdbb    /dev/sg2    C08

 RAID array: 2
   external: 60

total disks: 62
```

```
silver-is1-004:~ % cs_hal info node
Node                  : silver-is1-004
BIOS date             : 06/20/2012
BIOS version          : SE5C600.86B.01.03.0002.062020121504
Board model           : S2600JF
Board S/N             : QSJP23007313
Board vendor          : Intel Corporation
Board version         : G28033-506
Chassis S/N           : FC6ND131900019
Chassis vendor        : ..............................
Chassis model         : S2600JF
System S/N            : FC6AT131900005
Processor count       : 8
Total memory          : 23.0433GB
Availble memory       : 17.7322GB
Total swap            : 2GB
Available swap        : 2GB
Shared memory         : 0GB
Host adapter count    : 2
Net interface count   : 4
Enclosure count       : 1
External disk count   : 60
```

HAL – sensors

```
silver-is1-004:~ % cs_hal sensors all
Entity           Type                Label              Status  Info
-----            -----               -----              -----   -----
Power Dist       Power Unit          Pwr Unit Status    OK      OK; extra info unimplemented; actual: [c0 00 00]
Power Dist       Power Unit          Pwr Unit Redund    OK      fully redundant;
System Chassis   Chassis Intrusion   Physical Scrty     OK
System Board     SEL Disabled        System Event Log   OK      OK; extra info unimplemented; actual: [c0 04 00]
System Board     System Event        System Event       OK      OK; extra info unimplemented; actual: [c0 00 00]
System Board     Button/Switch       Button             OK      OK; extra info unimplemented; actual: [c0 00 00]
I/O Module       Module/Board        IO Mod Presence    OK      OK; extra info unimplemented; actual: [c0 02 00]
System Board     Mgmt Subsys Health  BMC Health         OK      OK; extra info unimplemented; actual: [c0 00 00]
System Chassis   Other Units-based   System Airflow     OK      12 CFM
System Board     Temperature         BB Inlet Temp      OK      33 Degrees Celsius
System Board     Temperature         SSB Temp           OK      63 Degrees Celsius
System Board     Temperature         BB BMC Temp        OK      53 Degrees Celsius
System Board     Temperature         P1 VR Temp         OK      39 Degrees Celsius
System Board     Temperature         IB QDR Temp        OK      48 Degrees Celsius
System Board     Temperature         Exit Air Temp      OK      53 Degrees Celsius
Front Panel      Temperature         IOM Temp           OK      40 Degrees Celsius
Drive Backplane  Temperature         HSBP PSOC          OK      40 Degrees Celsius
Front Panel      Temperature         LAN NIC Temp       OK      67 Degrees Celsius
Cooling Device   Fan                 Sys Fan 1A         OK      7387 RPM
Cooling Device   Fan                 Sys Fan 1B         OK      7482 RPM
Cooling Device   Fan                 Sys Fan 2A         OK      7387 RPM
Cooling Device   Fan                 Sys Fan 2B         OK      7654 RPM
Cooling Device   Fan                 Sys Fan 3A         OK      7387 RPM
Cooling Device   Fan                 Sys Fan 3B         OK      7396 RPM
Power Supply     PSU                 PS1 Status         OK
Power Supply     PSU                 PS2 Status         OK
Power Supply     Other Units-based   PS1 Input Power    OK      224 Watts
Power Supply     Other Units-based   PS2 Input Power    OK      196 Watts
Power Supply     Current             PS1 Curr Out %     OK      17 Unspecified
Power Supply     Current             PS2 Curr Out %     OK      14 Unspecified
Power Supply     Temperature         PS1 Temperature    OK      35 Degrees Celsius
Power Supply     Temperature         PS2 Temperature    OK      36 Degrees Celsius
Processor        Processor           P1 Status          OK      OK; extra info unimplemented; actual: [c0 80 00]
Processor        Processor           P2 Status          OK      OK; extra info unimplemented; actual: [c0 80 00]
```

REFERENCES

References – Failures

- "Are Disks the Dominant Contributor for Storage Failures?"
  - System-level failures http://www.usenix.org/events/fast08/tech/jiang.html
  - Weihang Jiang, Chongfeng Hu, Yuanyuan Zhou (UIUC), Arkady Kenevsky (NetApp)
  - Additional related studies
    - http://www.usenix.org/events/fast08/tech/bairavasundaram.html
    - http://www.usenix.org/events/fast08/tech/krioukov.html
- Google & CMU field reliability studies
  - http://www.usenix.org/events/fast07/tech/pinheiro.html
  - http://www.usenix.org/event /fast07/tech/schroeder/schroeder.pdf

References – Designing for Failure @ Scale

- Advice (LADIS 2009 workshop)
  - advice from Amazon - http://bit.ly/iDebZX
  - experience sharing from Google - http://bit.ly/mcvppe
  - from Microsoft - http://bit.ly/ixCh8i - and a number of others - http://bit.ly/jJ2VgW
  - The key take-away from Marvin's Amazon talk was the call for simplicity:
    - "It's 4AM, the clock is ticking, you have 52 minutes to solve problem, can you debug it?"
    - (52 minutes is the allowed yearly downtime at "4 9s" availability – Support calls you at 4am, how many minutes will it take for you to explain what the system is supposed to do, before they can begin to debug and fix it. If it takes 20 minutes to explain the design, you're down to 30 minutes left to fix what's wrong. And then nothing else can go wrong until next year.)