# Probability of RAID System Crashes

(c) Peter Thoemmes, 2013-01-16 (orig.: 2012-12-15)

The probability to read an unrecoverable bit from a hard disk is given by the URE (unrecoverable error rate) or also called BER (bit error rate). As disks are block devices, meaning one can only read block-wise (a physical block is a chunks of 512 bytes = 4096 bits), the probability of reading an unrecoverable block is 4096 times higher then the probability to read an unrecoverable bit:

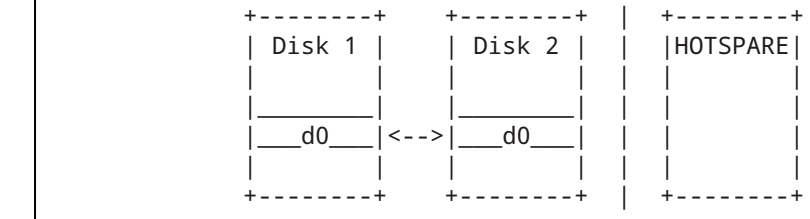
```
p_bad_bit = 1/URE

p_bad_block = 4096 x p_bad_bit

    Example:
        URE = 10^14
        ---> p_bad_bit = 1 x 10^-14
        ---> p_bad_block = 4.1 x 10^-11
```

The total number of blocks on a disk is calculated like this:

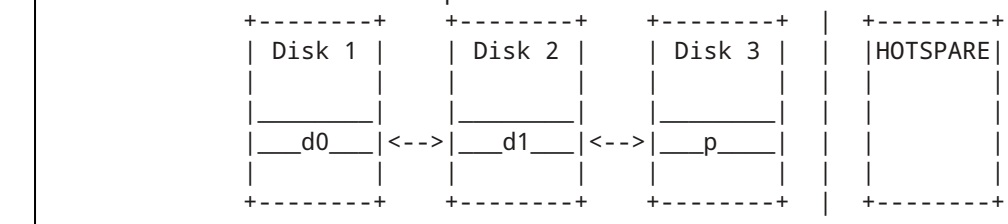
```
num_blocks = disksize/512Byte
           = 2 x disksize/KB
           = 2048 x disksize/MB
           = 2097152 x disksize/GB
           = 2147483648 x disksize/TB

    Example:
        Disk size = 2TB
        ---> num_blocks = 4.29 x 10^9
```

Now for real life usage we are looking at 3 basic RAID setups here:

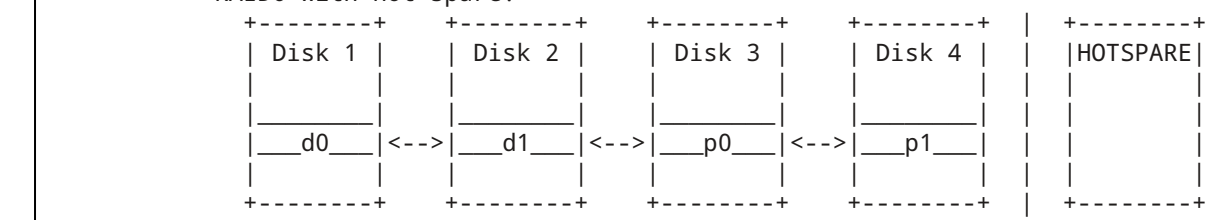
MIRROR with hot-spare:




RAID5 with hot-spare:

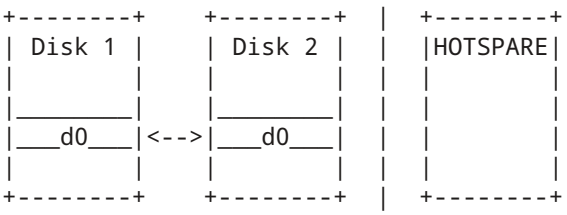



RAID6 with hot-spare:



Using a Mirror with 2 disks, the probability to read a bad block from the first disk, followed by a bad block from the second disk is given by the square of p_bad_block. If a defect disk was hot-swapped with the SPARE, then during building up the degraded RAID by the RAID controller (unaware which file system is on top) there is a higher probability to loose the RAID, as *all* blocks need to be read without error. So for each block read there is the probability p_bad_block to get a bad block.

MIRROR:



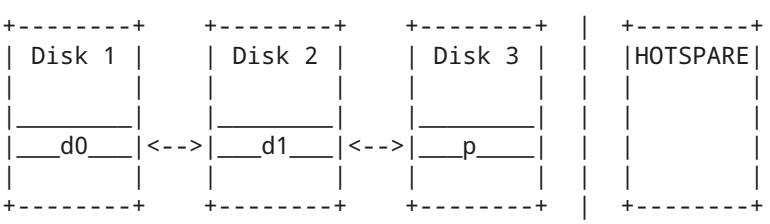

p_bad_read_mirror = p_bad_block^2
p_bad_rebuild_mirror = num_blocks x p_bad_block

Example:
URE = 10^14, Disk size = 2TB
---> p_bad_read_mirror = 16.8 x 10^-22
p_bad_rebuild_mirror = 0.176 = 17.6 %

This is a really good value for reading data in normal mode, but an extremely bad value for the rebuild phase of a degraded mirror.

Using a RAID5 with 3 disks, the information is not mirrored block by block, but striped (including 1 parity bit) over all disks. As RAID5 uses only 1 parity bit, a bad read will happen if 2 of the striped blocks are bad (double-failure). There is 3 possible double-failures here (d0/d1, d0/p and d1/p) and so we get 3 times the square of p_bad_block as the probability for a bad read. In degraded mode we have to read each single block without error and there is 2 possible single failures here (d0 and d1) which we get with the probability p_bad_block:

RAID5:



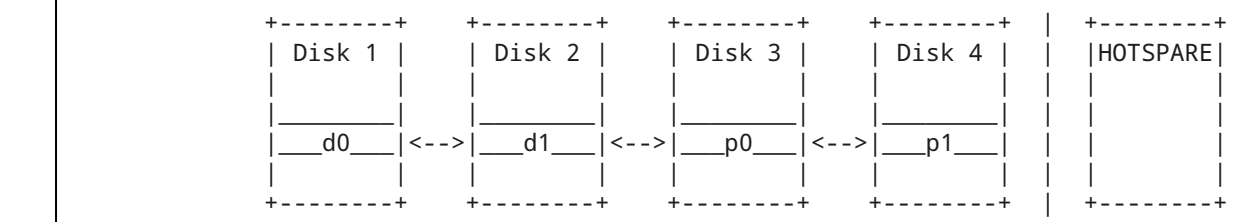

p_bad_read_raid5 = 3 x p_bad_block^2
p_bad_rebuild_raid5 = 2 x num_blocks x p_bad_block

Example:
URE = 10^14, Disk size = 2TB
---> p_bad_read_raid5 = 50.4 x 10^-22
p_bad_rebuild_raid5 = 0.352256 = 35.2 %

As for the mirror, this is a really good value for reading data in normal mode, but an but an extremely bad value for the rebuild phase of a degraded RAID5.

Using a RAID6 with 4 disks, the information is striped (including 2 parity bits) over all disks. As RAID6 uses 2 parity bits, a bad read will happen if 3 of the striped blocks are bad. There is 4 possible triple-failures here (d0/d1/p0, d0/d1/p1, d0/p0/p1 and d1/p0/p1) and so we get 4 times the cubic of p_bad_block as the probability for a bad read. In degraded mode we have to read without double-failure and there is 3 possible double-failures here (d0/d1, d0/p0 and d1/p0) which we get with the probability p_bad_block^2. In case 2 disks went lost, we have to read each single block of the 2 disks left without error and there is 2 possible single failures here (d0 and d1) which we then would get with the probability p_bad_block:

RAID6:




p_bad_read_raid6 = 4 x p_bad_block^3
p_bad_rebuild_raid6 = 3 x num_blocks x p_bad_block^2
p_bad_rebuild_2disks_raid6 = 2 x num_blocks x p_bad_block

Example:
URE = 10^14, Disk size = 2TB
---> p_bad_read_raid6 = 276 x 10^-33
p_bad_rebuild_raid6 = 21.7 x 10^-11
p_bad_rebuild_2disks_raid6 = 0.352 = 35.2 %

RAID6 shows up with perfect values for reading data in normal mode, as well as for the rebuild phase of a degraded RAID6 with one disk lost. Only if 2 disks went lost, it has problems to build up the RAID again.

That's a terrible bad result. It means, that for big disks, like 2TB, RAID1 (Mirror) and RAID5 is no longer safe. Only RAID6 can always hot-swap and rebuild a defect disk safely. Only two failed disks at the same time will usually bring down the RAID6 here.

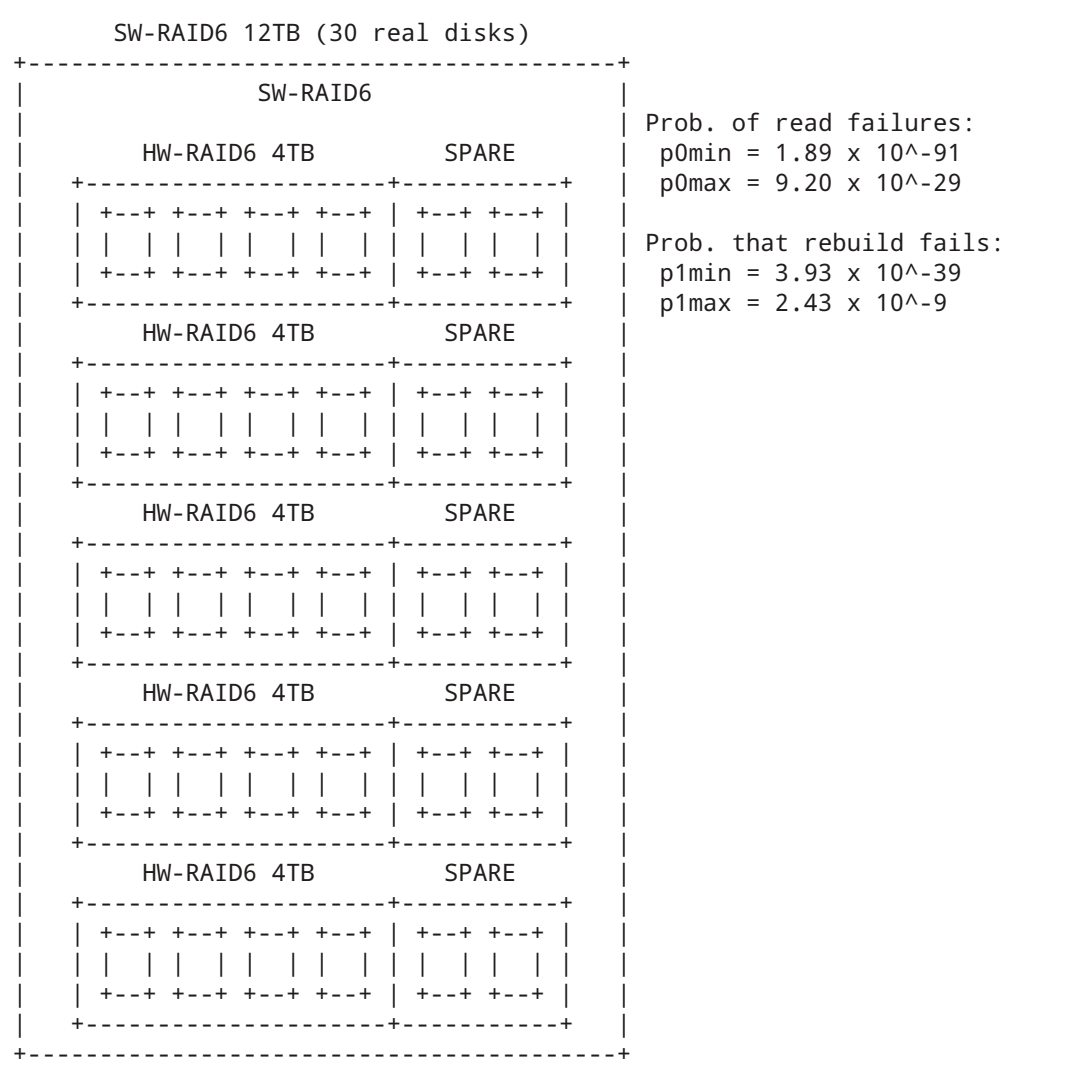
If we want to have more then 4 disks, it is safe to cascade the RAID6 systems. We may group 4 disks + 2 hot-spare disks to a HW-RAID6. Then we use those to build a new RAID6 on top. The new virtual disks would have a virtual p_bad_block' of value p_bad_read_raid6 (276 x 10^-33) in normal mode and p_bad_rebuild_raid6 (21.7 x 10^-11) if the HW-RAID6s all are in degraded mode with 1 disk failed.
Using a RAID6 with 5 such virtual disks, there is 9 possible triple-failures when reading in normal mode. For rebuilding one disk there is 6 possible double-failures and for rebuilding 2 disks there is 3 possible single failures. The number of blocks of one virtual disk (4TB) was 8589934592:

p_bad_read_raid6' = 9 x p_bad_block'^3
p_bad_rebuild_raid6' = 6 x num_blocks x p_bad_block'^2
p_bad_rebuild_2disks_raid6' = 3 x num_blocks x p_bad_block'

Example for disk size = 4TB:

MIN (all HW-RAID6 in normal mode):
p_bad_block' = 276 x 10^-33
p_bad_read_raid6' = 9 x 276^3 x 10^-99 = 1.89 x 10^-91
p_bad_rebuild_raid6' = 6 x 8589934592 x 276^2 x 10^-66 = 3.93 x 10^-39
p_bad_rebuild_2disks_raid6' = 3 x 8589934592 x 276 x 10^-33
= 7.12 x 10^-12

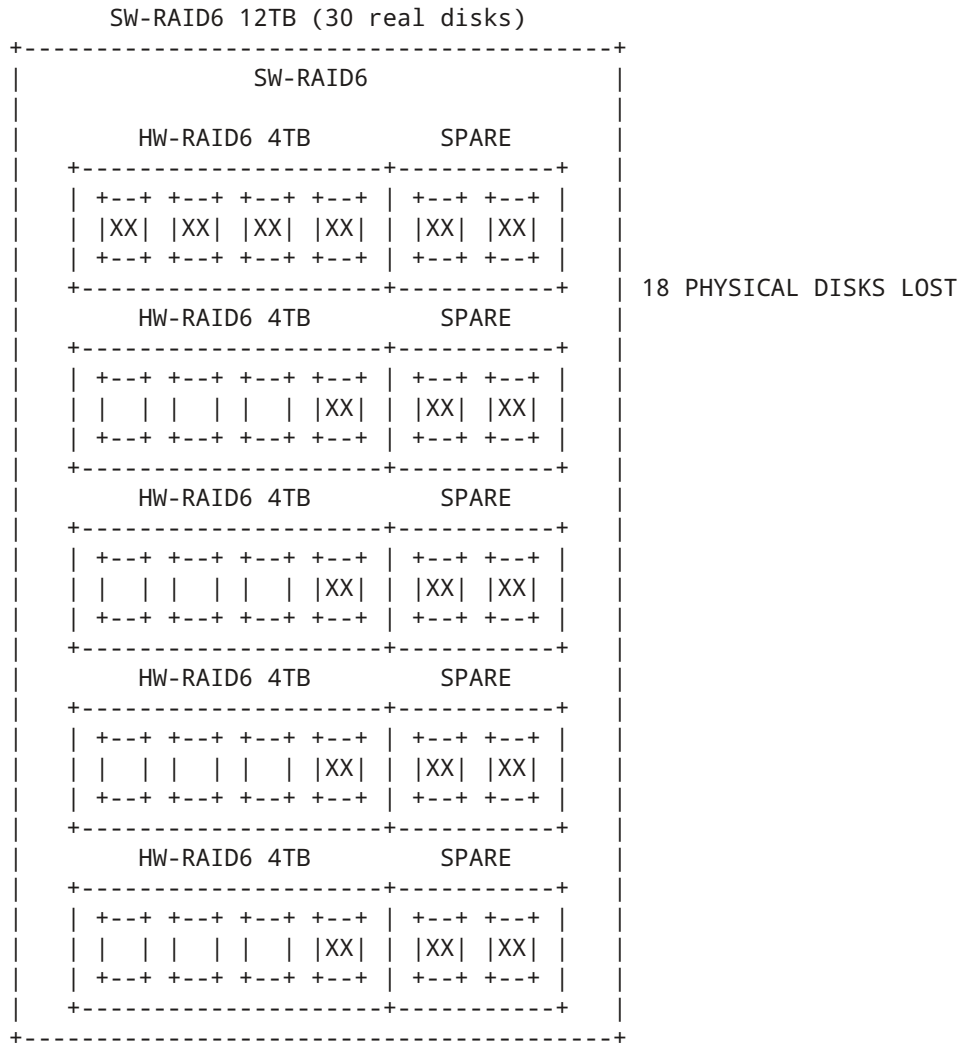
MAX (all HW-RAID6 in degraded mode with 1 disk failed):
p_bad_block' = 21.7 x 10^-11
p_bad_read_raid6' = 9 x 21.7^3 x 10^-33 = 9.2 x 10^-29
p_bad_rebuild_raid6' = 6 x 8589934592 x 21.7^2 x 10^-22 = 2.43 x 10^-9
p_bad_rebuild_2disks_raid6' = 3 x 8589934592 x 21.7 x 10^-11
= 5.59
= 559 % (SURELY COMPLETELY DOWN!!!)




So as a summary we can say, that we only allow one disk to fail in each RAID6 group (HW-RAID6 and SW-RAID6). Then, for 2TB disks with an URE of 10^14 we get following values for the probabilities:

```
HW-RAID6:
   p_bad_read_raid6 = 2.76 x 10^-31
   p_bad_rebuild_raid6 = 2.17 x 10^-10

SW-RAID6:
   p_bad_read_raid6' = 1.89 x 10^-91 ... 9.20 x 10^-29
   p_bad_rebuild_raid6' = 3.93 x 10^-39 ... 2.43 x 10^-9
```

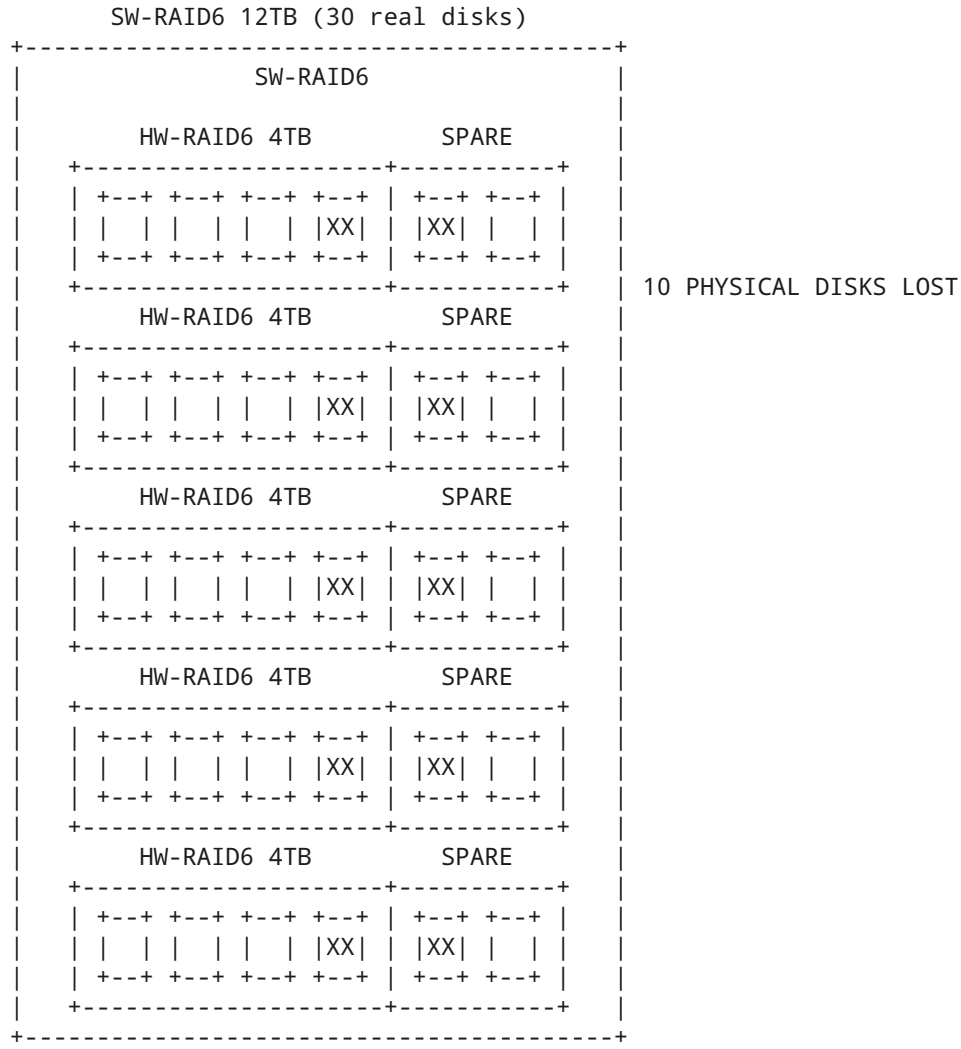
So the WORST CASE ALLOWED, which still would run the system, would be to loose *18* disks:




This means, that 60% of the physical disks would still not destroy the data, but require manual action to safely rebuild the system. All, except one HW-RAID6 may host just one operational defect disk. Only one HW-RAID6 may fail totally.

Even up to 6 more disks may fail without loosing the data, but then the system would run in such a degraded mode, that it could not be safely rebuild again, but it would be worse trying it.

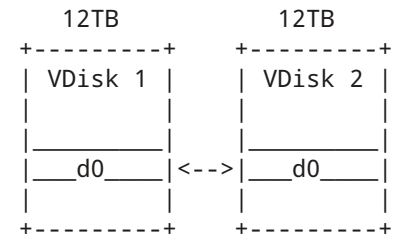
The WORST CASE ALLOWED, which would still fully automatically rebuild the whole system, would be to loose *10* physical disks:



We could compare the above example to a 12TB MIRROR on top of 2 times 3 of the same virtual HW-RAID6 4TB disks. The mirror would then provide 3 times 4TB, meaning 12TB capacity like the system shown above.

The new virtual disks would have a virtual p_bad_block' with the value p_bad_read_raid6 (276 x 10^-33) in normal mode and p_bad_rebuild_raid6 (21.7 x 10^-11) if the HW-RAID6s all are in degraded mode, rebuilding one failed physical disk.

So we would get following:




p_bad_read_mirror' = p_bad_block'^2
p_bad_rebuild_mirror' = num_blocks' x p_bad_block'

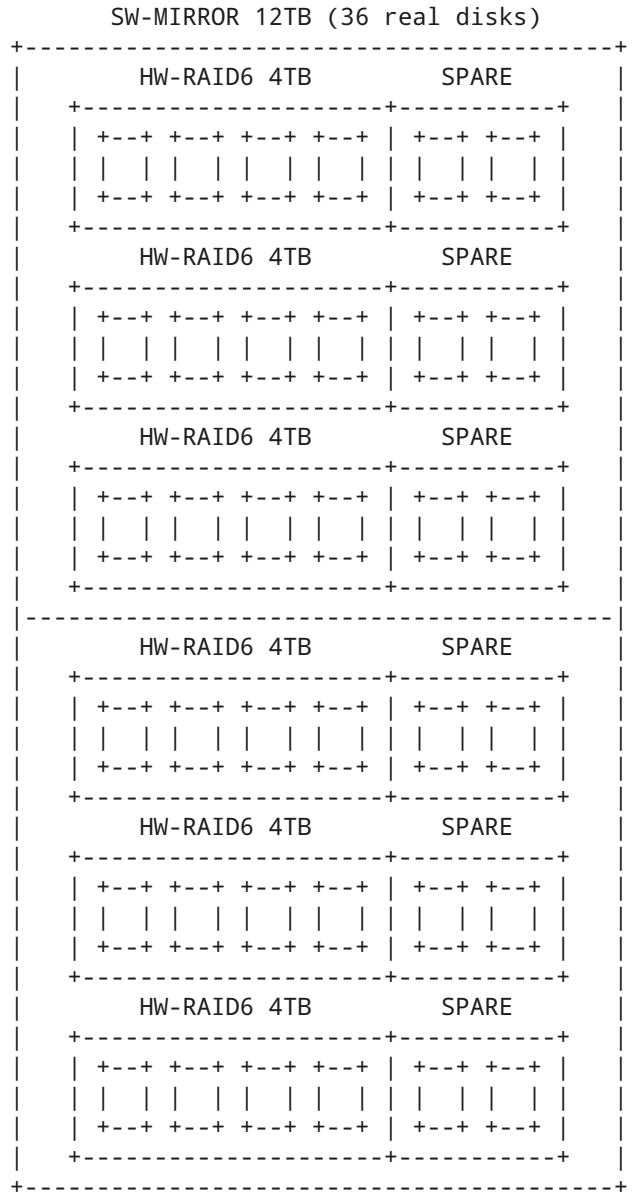
The number of blocks is given by

num_blocks' = 2147483648 x disksize/TB
= 2147483648 x 12
= 25769803776

So our comparable mirror would show up with following probabilities:

MIN (all HW-RAID6 in normal mode):
p_bad_block' = 276 x 10^-33
p_bad_read_mirror' = 276^2 x 10^-66 = 7.62 10^-62
p_bad_rebuild_mirror' = 25769803776 x 276 x 10^-33 = 7.11 x 10-21

MAX (all HW-RAID6 in degraded mode with 1 disk failed):
p_bad_block' = 21.7 x 10^-11
p_bad_read_mirror' = 21.7^2 x 10^-22 = 4.71 x 10^-20
p_bad_rebuild_mirror' = 25769803776 x 21.7 x 10^-11
= 5.59
= 559 % (SURELY COMPLETELY DOWN!!!)



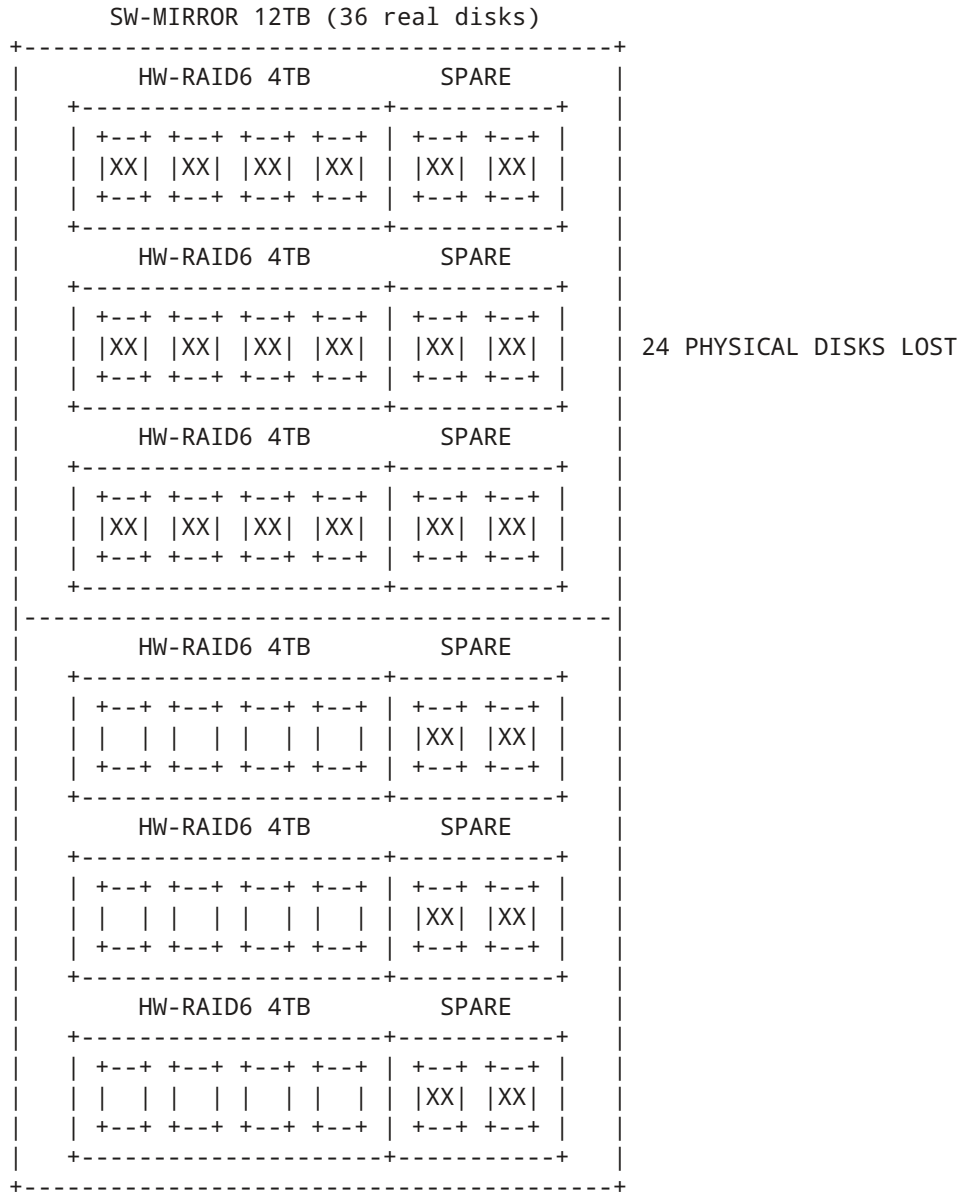

So as a summary we can say, that we only allow one disk to fail in each HW-RAID6 group. Then, for 2TB disks with an URE of 10^14 we get following values for the probabilities:

HW-RAID6:
   p_bad_read_raid6 = 2.76 x 10^-31
   p_bad_rebuild_raid6 = 2.17 x 10^-10

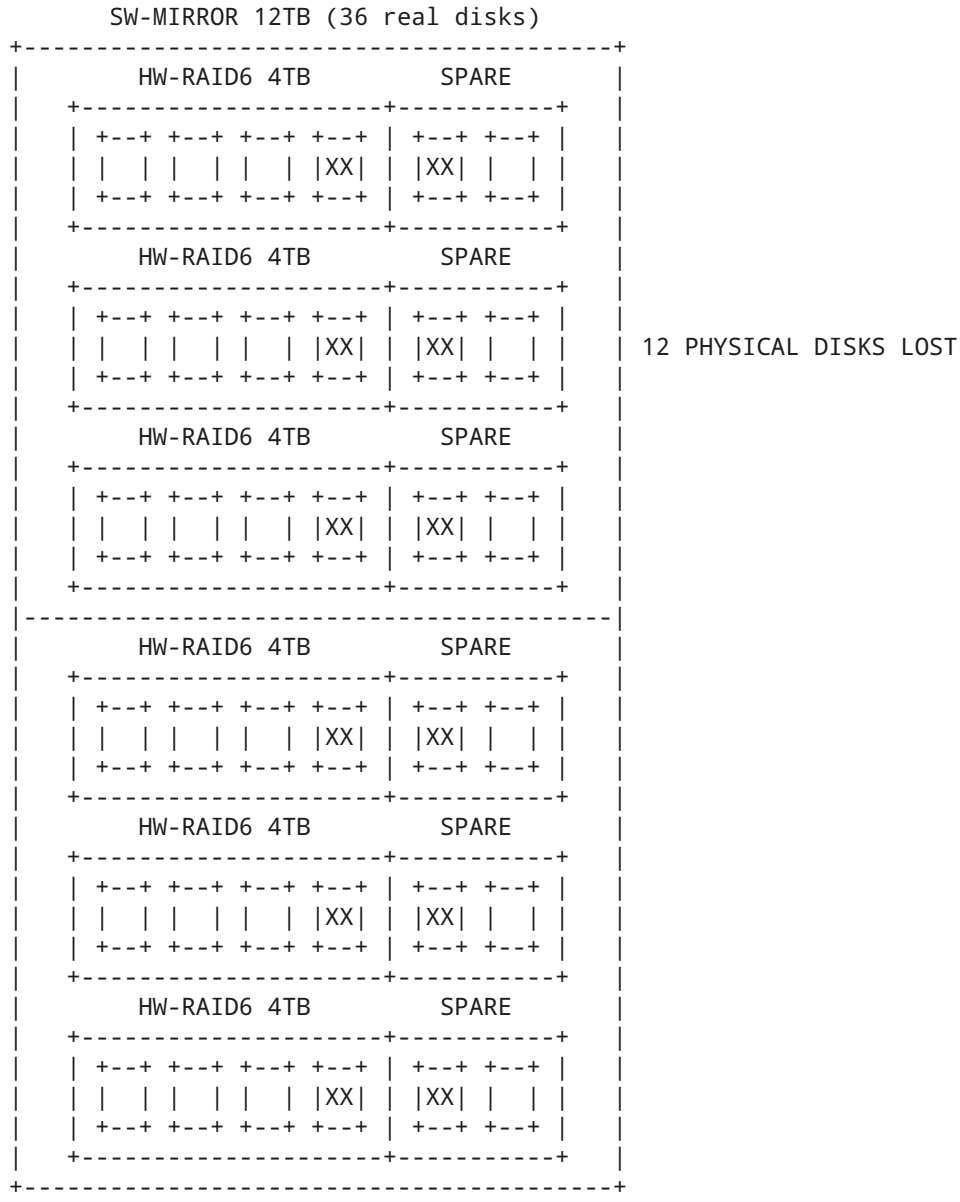
SW-MIRROR:
   p_bad_read_mirror' = 7.62 10^-62 ... 4.71 x 10^-20
   p_bad_rebuild_mirror' = 7.11 x 10-21 ... 5.59

SW-RAID6:
   p_bad_read_raid6' = 1.89 x 10^-91 ... 9.20 x 10^-29
   p_bad_rebuild_raid6' = 3.93 x 10^-39 ... 2.43 x 10^-9

Looking at the maximum number of lost disks without loosing data, we find *24* physical disks:



Looking at the maximum number of lost disks without loosing the rebuild capability, we find 12 physical disks:



```
+-------------------------------------------------------------------------------+
| OVERVIEW COMPARING DIFFERENT SETUPS FOR GETTING A 4TB VIRTUAL DISK:           |
+-------------------------------------------------------------------------------+
| (c) Peter Thoemmes, 2012-12-15                                                |
+-------------------------------------------------------------------------------+

Physical disk parameters:  URE = 10^14
                           p_bad_bit = 1/URE = 1 x 10^-14
                           p_bad_block = 4096 x p_bad_bit = 4.1 x 10^-11
                           num_blocks = 2147483648 x disksize/TB

MIRROR (RAID1) --- 4TB VIRTUAL DISK

         4TB           4TB            4TB
      +--------+    +--------+  |  +--------+
      | Disk 1 |    | Disk 2 |  |  |HOTSPARE|
      |        |    |        |  |  |        |
      |________|    |________|  |  |        |
      |___d0___|<-->|___d0___|  |  |        |
      |        |    |        |  |  |        |
      +--------+    +--------+  |  +--------+
    p_bad_read_mirror = p_bad_block^2
    p_bad_rebuild_mirror = num_blocks x p_bad_block
        num_blocks = 8.58 x 10^9
        p_bad_read_mirror = 16.8 x 10^-22
        p_bad_rebuild_mirror = 0.352 = 35.2 % (REALLY BAD!!!)

RAID5 --- 4TB VIRTUAL DISK

         2TB           2TB           2TB            2TB
      +--------+    +--------+    +--------+  |  +--------+
      | Disk 1 |    | Disk 2 |    | Disk 3 |  |  |HOTSPARE|
      |        |    |        |    |        |  |  |        |
      |________|    |________|    |________|  |  |        |
      |___d0___|<-->|___d1___|<-->|___p____|  |  |        |
      |        |    |        |    |        |  |  |        |
      +--------+    +--------+    +--------+  |  +--------+
    p_bad_read_raid5 = 3 x p_bad_block^2
    p_bad_rebuild_raid5 = 2 x num_blocks x p_bad_block
        num_blocks = 4.29 x 10^9
        p_bad_read_raid5 = 50.4 x 10^-22
        p_bad_rebuild_raid5 = 0.352 = 35.2 % (REALLY BAD!!!) ---> like Mirror

RAID6 --- 4TB VIRTUAL DISK

         2TB           2TB           2TB           2TB            2TB
      +--------+    +--------+    +--------+    +--------+  |  +--------+
      | Disk 1 |    | Disk 2 |    | Disk 3 |    | Disk 4 |  |  |HOTSPARE|
      |        |    |        |    |        |    |        |  |  |        |
      |________|    |________|    |________|    |________|  |  |        |
      |___d0___|<-->|___d1___|<-->|___p0___|<-->|___p1___|  |  |        |
      |        |    |        |    |        |    |        |  |  |        |
      +--------+    +--------+    +--------+    +--------+  |  +--------+
    p_bad_read_raid6 = 4 x p_bad_block^3
    p_bad_rebuild_raid6 = 3 x num_blocks x p_bad_block^2
    p_bad_rebuild_2disks_raid6 = 2 x num_blocks x p_bad_block
        num_blocks = 4.29 x 10^9
        p_bad_read_raid6 = 276 x 10^-33
        p_bad_rebuild_raid6 = 21.7088 x 10^-11
        p_bad_rebuild_2disks_raid6 = 0.352 = 35.2 % (REALLY BAD!!!)
```

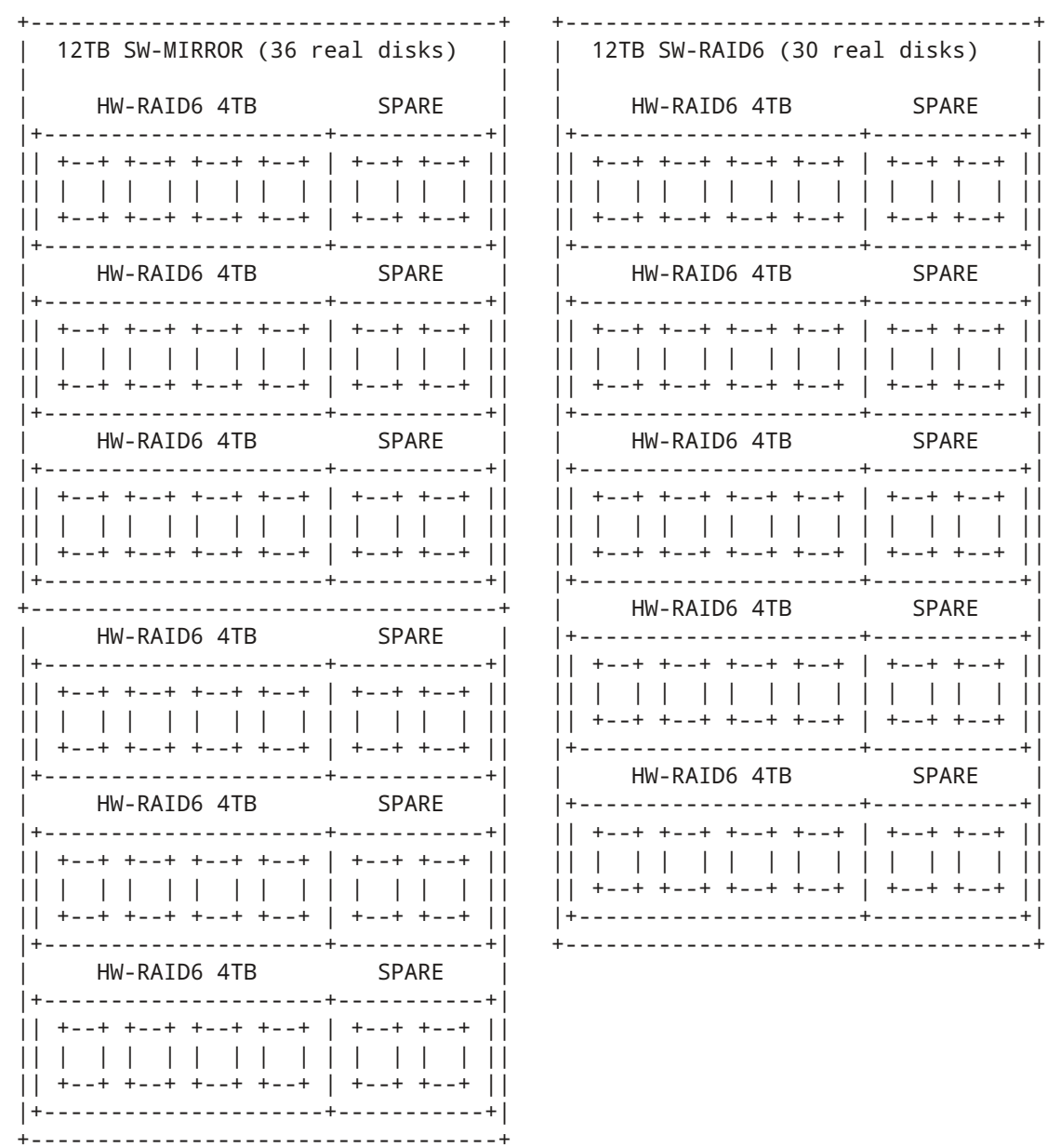
# OVERVIEW COMPARING DIFFERENT SETUPS FOR GETTING A 12TB VIRTUAL DISK:

(c) Peter Thoemmes, 2013-01-16

This overview compares two setups, both providing 12TB data storage and using 4TB HW-RAID6 virtual disks underneath. The HW-RAID6 group 4 disks + 2 hot-spare disks to one virtual disk of 4 TB size. The physical size is 2TB and the URE 10^14, which results in following probabilities for failing:

```
     HW-RAID6:
        p_bad_read_raid6 = 2.76 x 10^-31
        p_bad_rebuild_raid6 = 2.17 x 10^-10   (1 failed disk is rebuild)
```

The two SW-RAID setups are MIRROR and RAID6 are compared by looking at 2 case studies: BEST CASE is the one with all HW-RAID6 running in normal mode, WORST CASE is the one with all HW-RAID6 running in DEGRADED mode with 1 disk failed and currently rebuild (we allow maximum 1 disk to fail in each HW-RAID6 group):



In any case the SW-RAID6 is doing MUCH better. In the worst case, the SW-MIRROR will most probably (559%) not build up again!

| | | SW-MIRROR FAILURE | SW-RAID6 FAILURE |
|---|---|---|---|
| READ | BEST CASE | 2.76 x 10^-31 | 1.89 x 10^-91 |
| | WORST CASE | 4.71 x 10^-20 | 9.20 x 10^-29 |
| REBUILD | BEST CASE | 7.11 x 10-21 | 3.93 x 10^-39 |
| | WORST CASE | 5.59 | 2.43 x 10^-9 |

It is also useful to compare the maximum number of failed physical disks. The SW-RAID6 is the winner in this case as well:

Maximum number of failed disks without data loss:

SW-MIRROR .....: 18 (50%)
SW-RAID6 ......: 18 (60%)

Maximum number of failed disks with a still successful rebuild:

SW-MIRROR .....: 12 (33%)
SW-RAID6 ......: 10 (33%)

Beside this the SW-RAID6 system requires only 30 disks, compared to the 36 used by the SW-MIRROR.