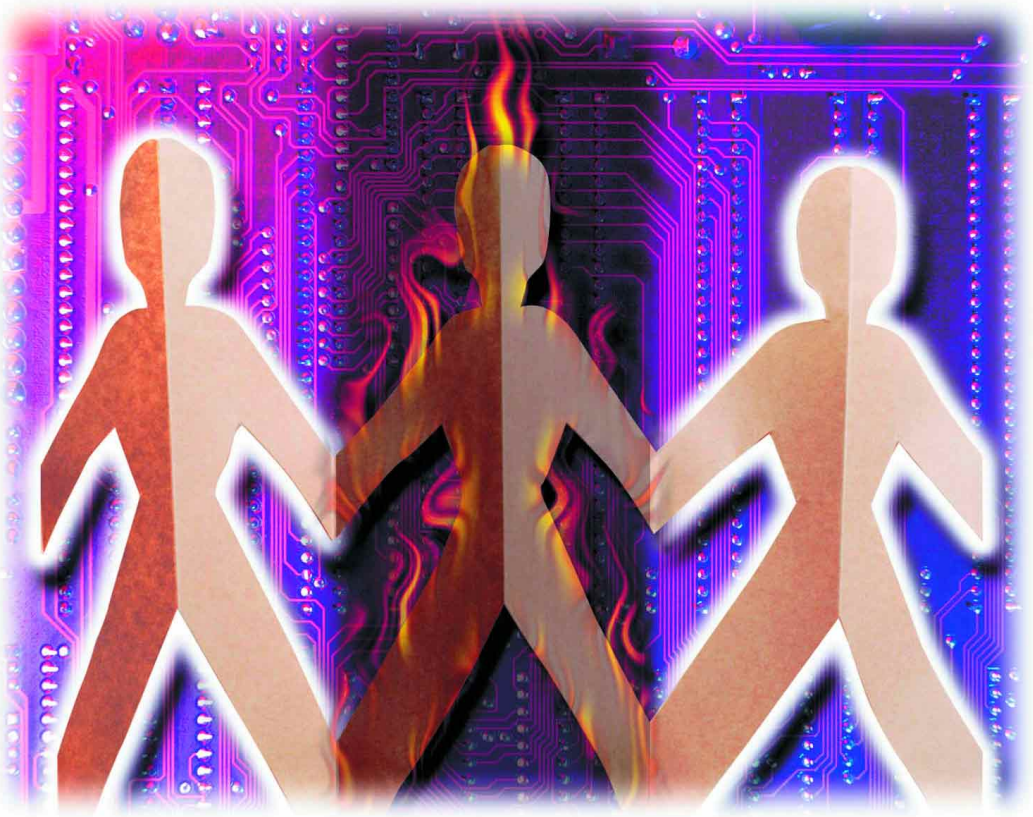
Redundant Array of Inexpensive Disks

REDUNDANCY IS GOOD!

BERNHARD KUHN

When a single hard disk isn't fast enough, or its storage capacity is insufficient, one solution is to connect several drives together. As an added benefit, this can be done in a way that increases reliability by allowing individual drives to fail without losing data.



Three scientists at the University of Berkeley first hit on the idea more than 13 years ago of making a resilient and high performance storage medium out of separate hard disks: They defined five variants of this design and called it a "Redundant Array of Inexpensive Drives", RAID for short. This acronym is often also said to stand for "Redundant Array of Independent Disks". In RAID levels 1 to 5 one drive can fail without the system having to stop working. Later, two more configurations were added: RAID 0 with no error tolerance and RAID 6 with additional fault tolerance.

Hard or soft?

Big corporations are the main users of RAID technology. This isn't surprising: the hardware isn't cheap since apart from the bus controllers (PCI/SCSI) it must include a complete processor unit and a few megabytes of buffer memory (see Fig.1). A RAID controller acts just like an ordinary hard disk controller, although special drivers are often needed by the operating system. For information about specific controllers see the test report on hardware RAID controllers with Linux support in this issue on page18.

As the performance of processors and the complexity of operating systems has increased, it has also become possible to implement error correction using redundant disks in the server itself. This variant, known as "Software RAID" (or "SoftRAID" for short) is enjoying ever-increasing popularity, especially with the home user who is looking for a useful and cheap way to use any old hard disks that may be lying around. (Software RAID is also dealt with in more detail in another article in this issue on page 62).

At the other extreme, an external SCSI-to-RAID bridge can be used without the need for any special

device drivers. From the point of view of the SCSI adapter in the server this behaves like an ordinary SCSI drive. Figure 2 shows a RAID array with integrated SCSI converter.

A RAID array owes its fault tolerance to the fact that it contains at least one extra hard disk which, by a variety of methods, allows the data on a failed drive to be recovered. If a drive fails it should nevertheless be replaced as soon as possible since if a second drive fails all the data will probably be lost.

Fail safe

According to the laws of probability a redundant disk array, when used correctly, should only be out of action for a brief period about once every twenty thousand years. However, leaving aside for a moment the symptoms of ageing of the other components, it's possible for a defective hard disk to cripple the whole (SCSI or IDE) bus (for example, turning it into a "babbling idiot!") so that other drives are also temporarily unable to function. If this happens it will cause the entire system to stop working.

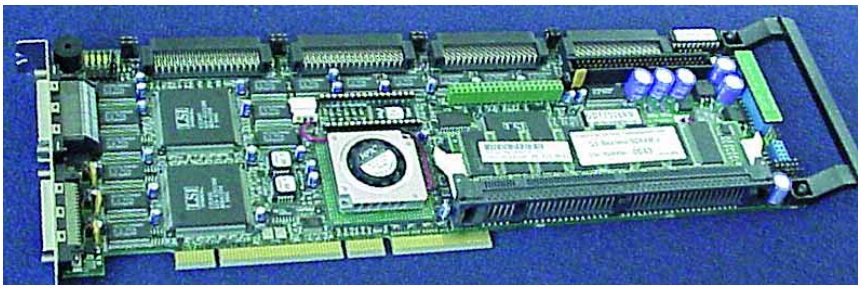
It's true that SCSI hard disks usually die quietly: they just fall silent. But to play it completely safe, it's

Fig. 3: Special cartridges are used to allow drives to be hot-swapped

best to devote a separate channel to each hard disk. This will also avoid any bottlenecks in slower bus systems, but the improvement obviously comes at greater cost.

In order to be able to exchange faulty media during operation (a process known as "hot swapping") hard disks are mounted into special cartridges, which slot into a cage. These cartridges ensure that destructive electrical potentials are discharged on insertion and that the power supply to the drive starts cleanly on insertion and is cut off before removal. The RAID controller software must also be able to correct any transfer errors that might occur due to signal interference during the swap procedure, for example by repeating the read or write cycles affected.

When a defective drive is replaced, reconstruction of the data or error correction codes is performed. Because this can involve examining every bit of data in the RAID system the process can take several hours. During this time, use of the server may be subject to a few restrictions on performance, although the reconstruction should only run when no data read or write operations are pending.



[above]
Fig. 1: a multi-channel RAID controller

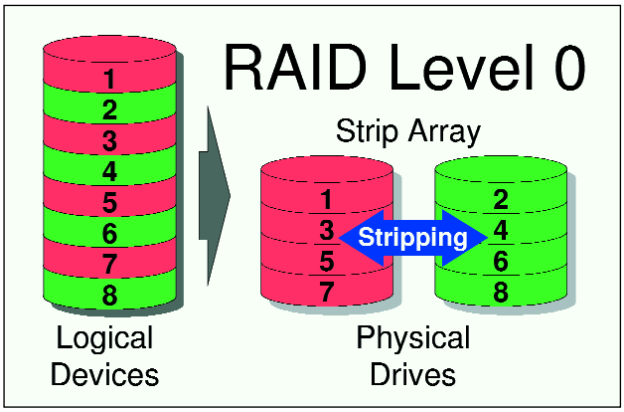
[above]
Fig. 2: SCSI-to-RAID bridge based on BSD: configuration is done via a serial interface

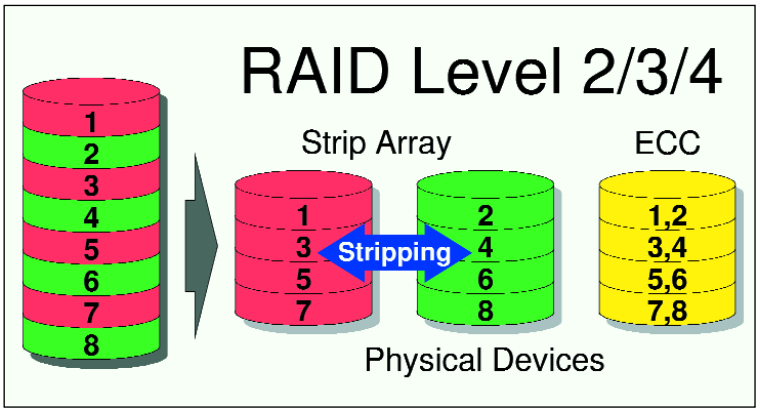
If a disk fails on a Saturday, which is the administrator's day off but a day when the system's users are very busy, the weekend can be saved for everyone by using a "hot spare" hard disk. With this, if a drive fails the data reconstruction on to the spare drive starts automatically. Replacing the defective medium is then not quite so urgent. The price to pay for this is that the capacity of the spare disk remains unused during normal operation. For this reason, this solution is only deployed in mission-critical applications.

In all there are more than a dozen different RAID levels, each involving descendants or combi-

Table 1:RAID Level for servers at a glance

Level	0	1	2-4	5	6	10
minimum hard disks	2	2	3	3	4	4
data hard disk+ error code carrier	n+0	1+1	n+1	n+1	n+2	n+n
Reading performance in normal operation (Factor)	n	1 to 2n	n	n	n	n to 2*n
Ideal reading performance in case of disk failure	0	1	n	n	n	n to 1.5*n
Write performance	n	1	n	n	n	n
Fail-safe	--	++	+	+	+++	++
Performance/Price ratio	++	0	-	+	--	0





[left] Fig. 4: Not really a RAID: RAID increases transfer speed at the cost of reliability

[right] Fig. 5: Redundancy and high transfer performance are achieved by combining RAID with an error correction process.

nations of the basic forms. An administrator should spend some time thinking about precisely which level is best suited to the needs of the applications that will use it. The overview in Table 1 should be taken with a pinch of salt: depending on the application, it could look completely different.

'Striptease' with RAID 0

At the lowest RAID level data is stored without any redundancy. There is therefore no resilience or fault tolerance. Data is written in blocks or "chunks": the first block to the first drive in the array, the second block to the second drive and so on. For this reason, RAID 0 is often referred to as "data striping".

The benefit of RAID 0 is not automatic error recovery but improved performance. It is possible to achieve almost *n* times the performance of a single hard disk, where *n* is the number of drives in the array. This is achieved because *n* read or write operations can take place simultaneously instead of sequentially. However, the probability of failure also increases *n*-fold.

Since a RAID 0 subsystem has no redundancy, if there is a fault the data is normally lost. Files of a size smaller than the block size – depending on the file system used – do have a certain chance of survival, but restoring them manually is tiresome and time-consuming. RAID 0 is thus certainly not a *Redundant* Array of Inexpensive Disks and is suitable only for applications in which large amounts of data must be recorded very quickly only to be discarded after a short processing period, such as in compressionless non-linear video editing.

Mirror on the wall

RAID Level 1 is the simplest form of RAID, and is also known as "disk mirroring." It creates redundancy very simply by writing all data twice: once to each of two disks. If a hard disk goes down, the data is still there, intact, on the second drive.

Since each block of data is synchronously duplicated on the two disks there is no performance increase (or decrease) compared to using a single hard disk. Reading small files also isn't faster, but big files can be read from the two disks in parallel (if the bandwidths of the busses allow such a thing). ie. Chunks 1, 3 and 5 from disk 1 can be read along with chunks 2, 4 and 6 from the other disk. However, the blocks have to be re-interleaved.

RAID 1 can be useful in applications like web servers, file servers or news servers, where some fault tolerance is needed and data tends to be read more often than it is written. However, the disadvantage of it is that you are giving away half your dearly bought storage capacity.

RAID 2/3/4: One more dosen't hurt

If a striping array (RAID 0 with *n* drives) is provided with an additional drive that is used to store error correction and checking (ECC) codes, higher transfer rates and a lower risk of unrecoverable errors are combined. If one disk from the stripe array goes down, the lost data can be completely restored from the contents of the remaining drives plus the error correction information. The transfer rate during write operations (and the speed of restoring) is a function of the processing power of the ECC calculation unit.

RAID levels 2 and 3 both use an algorithm developed in 1950 by R W Hamming to calculate the ECC codes; they differ only in the chunk size that is used. RAID 2 uses a chunk size of just one bit: its benefits are more theoretical than anything else and you won't find any RAID 2 arrays in real life. There are commercial implementations of RAID 3 (with small chunk sizes) but they are seldom used. Higher RAID levels are preferred.

RAID Level 4 uses considerably larger chunks than its predecessors, (usually 4 to 128KB) and uses a simple exclusive-OR operation to generate the error correction codes and to restore data. Figure 5 shows an example with a chunk size of four bits.

The compromise

If data and error codes are distributed equally over the N+1 hard drives according to Fig 6, then they can read n+1 data blocks at once. For example to get the first six data blocks, the RAID-solution reads the boocks 1 and 6 from the first, 2 and 3 from the

Info

D. A. Patterson, G. Gibson, and R. H. Katz, „A Case for Redundant Arrays of Inexpensive Disks (RAID)", Report No. UCB/CSD 87/391, University of California, Berkeley, CA 1987.

Nick Sabine, "An Introduction to RAID": http://www-student.furman.edu/users/n/nsabine/cs25/

Storage Technology Corporation: http://www.stortek.com/StorageTek/hardware/disk/raid/raid.html

second and 4 and 5 from the third drive (two block operations per drive). With RAID 2/3/4, the blocks 1, 3 and 5 would be read from the first drive and 2, 4 and 6 from the second (three blcok operations per drive being necessary). The redundancy information is not used for read operations in normal situations. The amount of space used for error correction purposes is the same as for RAID 4 so, given the benefits, it is hardly surprising that RAID 5 is the preferred level used in practical applications.

No worries!

In especially critical applications provision must be made for the simultaneous loss of two disks. RAID 5 isn't up to this and so to meet this requirement we have RAID 6. RAID 6 calculates two different error correction values from *n* data chunks and, as in RAID 5, distributes these evenly on to all hard disks. The Reed-Solomon error correction code is frequently used. Calculating this requires considerable computing power: consequently RAID 6 systems are not exactly cheap.

Other configurations

A duplicated disk stripe with at least four media as shown in Fig. 8 is also often referred to as RAID 10 (0+1). The hardware RAID controllers needed to implement this are relatively cheap, which helps to offset the cost of providing twice the storage capacity that would otherwise be needed. This solution is usually implemented using ordinary disk controllers with the operating system taking over the RAID function, so in fact it is really a cleverly-disguised software RAID solution.

Other RAID derivatives are RAID 30 or 50. In RAID 50, for example, three RAID 0 arrays are used as data storage for a RAID 5 configuration.

Other RAID levels are also defined, though they are rarely used in practice.

RAID 7 works in a similar way to level four, but requires a microcontroller which processes all I/O activities asynchronously, sorts them appropriately and buffers data. All current RAID controllers (and software solutions) have this ability built into them anyway, so this RAID level is obsolete. However it is still sometimes used in marketing to make a product appear to have something special.

The term "RAID 100" refers to parallel accesses to a RAID 1 system. This is also only possible with the aid of a dedicated microcontroller and is now rarely used.

Software RAID 0 evenly distributes data chunks over all the available hard disks. The same effect can be achieved using the Logical Volume Manager by specifying the "strip" parameter. The Linux LVM, incidentally, is planned to include support for equivalents of RAID 1 and RAID 5.

Conclusion

A RAID for all seasons does not exist! Each RAID level has its own advantages and disadvantages. There is usually a price to be paid for high performance and fail-safe features and so the final decision will often be subject to budgetary constraints.

RAID Level 5 is an outstanding compromise and for this reason it is widely used. Depending on the application, however, adequate protection for the data on a server can be economically obtained using the "poor man's RAID" – RAID 1. ■

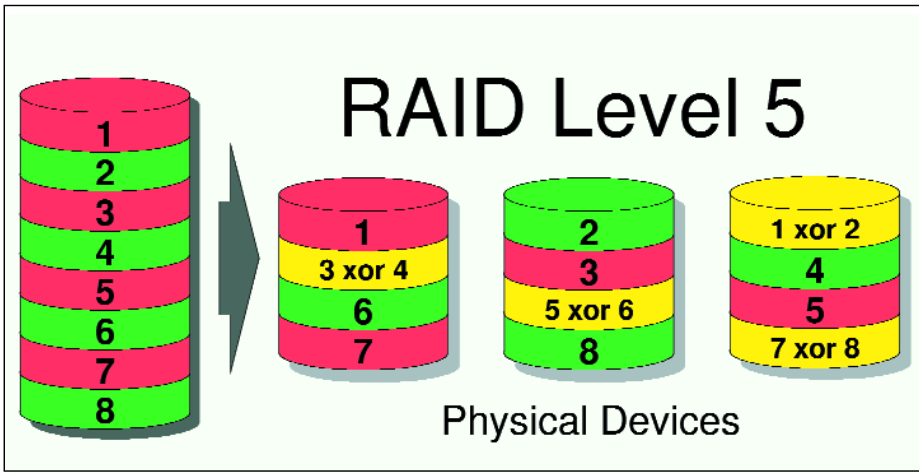
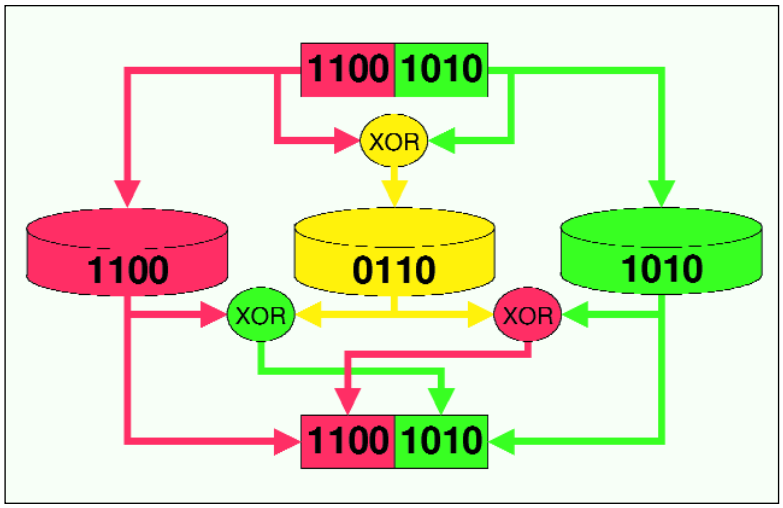
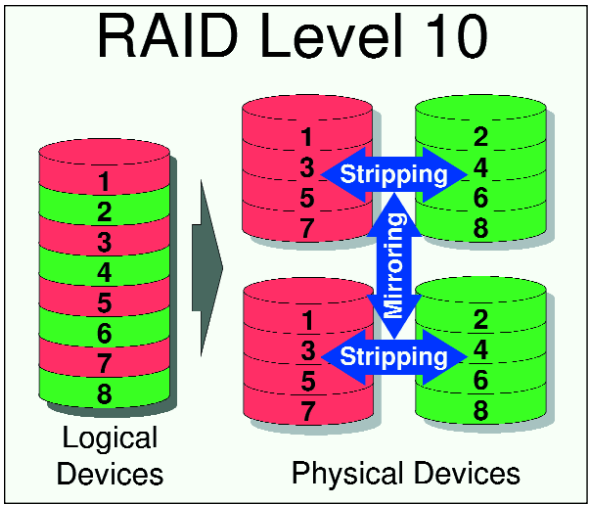


Fig. 6:RAID5 is now state of the art in industry**



**[left]
Fig. 7:dataflow under RAID 5 in normal operation and reconstruction.**



**[right]
Fig. 8: RAID 0+1: Parallel accesses as with RAID**1 using low-cost controllers**