Poznámky o RAIDu

Autor: František Ryšánek <rysanek@fccps.cz>

FCC Průmyslové systémy s.r.o.

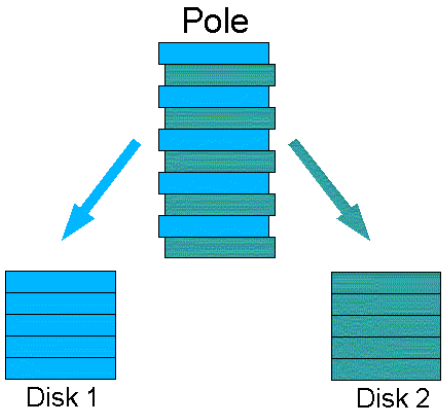
RAID = ”Redundant Array of Inexpensive Disks” - virtuální disk skládající se z více disků.

Obvykle se ovšem nejedná o právě lacinou technologii.

Malá pole se stavějí kvůli spolehlivosti, v případě rozsáhlejších konfigurací jde spíše o jediný způsob, jak při daném stavu technologií pořídit na trhu větší kapacitu, než jaká je k dispozici v rámci jediné diskové jednotky.

Typy polí - "RAID levels"

RAID 0 (striping) = maximální výkon

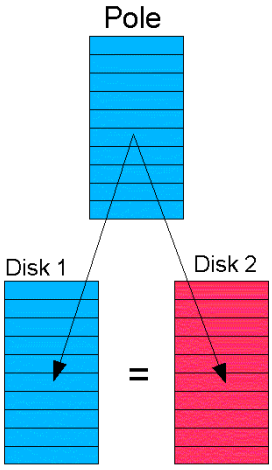


Vytvoří virtuální disk o kapacitě rovnající se téměř součtu kapacit obou (všech) fyzických disků. Virtuální disk je "nakrájen na proužky" o velikosti řádově 64 kB, které se pak rovnoměrně rozdělí mezi oba disky.

Při čtení i zápisu má potom toto pole takřka dvojnásobný (n-násobný) trvalý datový tok oproti jednotlivému fyzickému disku.

Pozor, toto pole je nebezpečné! Pokud selže kterýkoli zúčastněný fyzický disk, pole se nenávratně rozpadne! Nelze zachránit ani část dat, protože na každém disku je pouze část - každý druhý 64k proužek. Jako kdyby data prošla skartovačkou.

RAID 1 (mirroring) = maximální spolehlivost



Vytvoří virtuální disk o kapacitě rovnající se kapacitě jednotlivého zúčastněného disku (nebo polovině součtu kapacit všech disků - lze využít pouze sudý počet disků).

Data z virtuálního disku se ukládají dvakrát - na každý z obou disků jedna kompletní kopie. Nemá smysl hovořit o ”užitečných” datech a paritě – obě kopie jsou rovnocenné. Při výpadku kteréhokoli jednotlivého fyzického disku pole ”degraduje” – virtuální disk nadále funguje, nedojde ke ztrátě dat. Pole lze pole opravit (= obnovit redundanci) - pochopitelně až po výměně vadného disku.

Toto pole má výkon na úrovni jednotlivého disku.

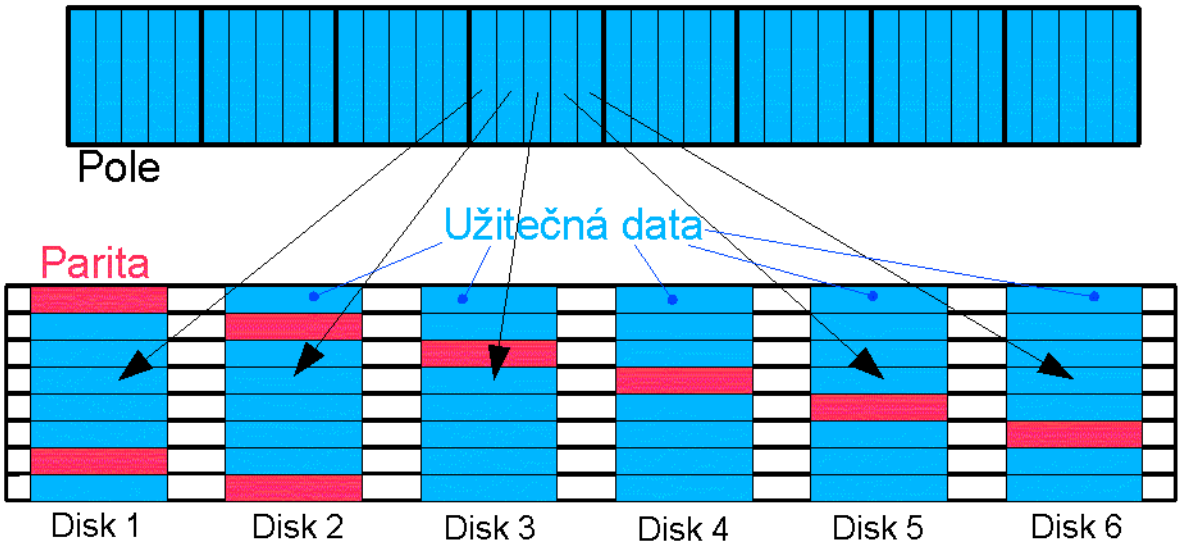
RAID 0+1 (striping+mirroring)

Má stejnou spolehlivost a kapacitu jako mirror, ale vyšší výkon, a lze ho použít přes více disků. Proto některé moderní RAID řadiče údajně ve skutečnosti ani neumí prostý RAID 1 (mirroring).

RAID 5 (redundancy) = výkon, kapacita, spolehlivost

Toto pole lze vytvořit na třech a více fyzických discích (N). Vytvoří se virtuální disk o kapacitě rovnající se přibližně kapacitě N-1 krát kapacita jednotlivého disku.

Data se rozpočítají mezi více disků. Zjednodušeně řečeno: virtuální svazek se nakrájí na ”proužky”, které se rozdělí mezi všechny disky, a na redundantní disk (disky) se ukládá paritní informace. Vtip je v tom, že pro každou sadu proužků se na paritu použije jiný disk, takže nedochází k nadprůměrnému vytěžování jediného disku (což by snižovalo celkovou průchodnost systému).



Takovéto pole má o něco menší výkon a kapacitu než je součet všech použitých fyzických disků. Optimálního výkonu pole dosahuje při ukládání a čtení velkých spojitých bloků dat. Je třeba si uvědomit jednu jedovatou vlastnost: speciálně při zápisu malých bloků dat musí pole před uložením načíst všechny proužky v sadě a dopočítávat a ukládat novou hodnotu paritního proužku – což může razantně snižovat výkon. Dá se proti tomu bojovat velkou cache na RAIDovém adaptéru.

Pokud jeden disk vypadne, řadič díky redundantnímu ukládání dopočítá chybějící data. Pokud vypadne víc než jeden disk, pole nenávratně zhavaruje.

RAID 6 = Výkon, kapacita, dvojnásobná spolehlivost

Toto pole lze vytvořit na čtyřech a více fyzických discích (N). Vytvoří se virtuální disk o kapacitě rovnající se přibližně kapacitě N-2 krát kapacita jednotlivého disku.

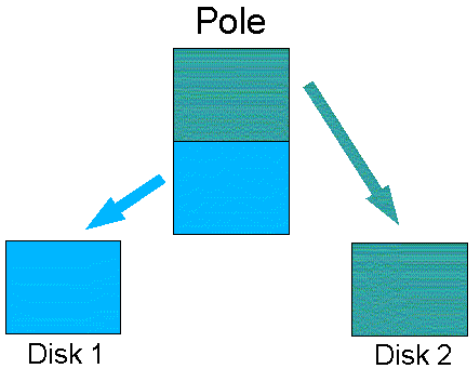
Podobně jako u RAIDu 5 se data stříhají na „proužky“ a rozkládají se cyklicky mezi disky.

Podobně jako u RAIDu 5 je mezi disky cyklicky rozložena také parita, ovšem na rozdíl od RAIDu 5 jsou na paritu obětovány v každé sadě *dva* proužky (zvané P a Q) – tj. globálně dohromady dva disky. Každý z obou paritních proužků je počítán jiným algoritmem tak, aby se data dala složit dohromady při současném výpadku *kterýchkoli dvou disků* – povšimněte si, že tuto úroveň bezpečnosti

nevykazují ani pole úrovně 10 či 50, u kterých pole přežije havárii příznivé kombinace i dvou a více disků, ale nepřežije havárii *libovolných* dvou disků (lze najít kombinace dvou disků, které způsobí havárii pole – tj. dva disky v jedné RAID5 či RAID1 sadě).

Podobně jako RAID 5, i RAID 6 vykazuje „jedovatost“ při zápisu malých bloků dat. RAID 6 má ještě o něco menší kapacitní a výkonovou efektivitu než RAID 5, ale kapacitní efektivita je pořád o mnoho lepší než u RAID 10.

Linear (spanning) = maximální kapacita



Vytvoří virtuální disk o kapacitě rovnající se součtu kapacit obou (všech) fyzických disků. Data se ukládají pouze jednou. Disky se prostě "spojí za sebe".

Výkon pole odpovídá výkonu jednotlivého disku. **Pokud selže kterýkoli zúčastněný fyzický disk, pole se nenávratně rozpadne!**

Lze spekulovat o tom, že by se data z jednotlivého disku dala ručně zachránit. Například jednotlivý soubor, pokud poznáme z obsahu sektorů jeho začátek a konec. Prakticky by to asi bylo možné za předpokladu, že souborový systém na virtuálním disku bude mít nulovou fragmentaci a zachraňovaný soubor nebude překračovat hranice disků.

JBOD

Zkratka znamená „just a bunch of drives“ – jde o přímý přístup k jednotlivým diskům, které nejsou součástí pole. Tento způsob přístupu k fyzickému disku lze použít např. v situaci, kdy v počítači máme pouze RAIDový řadič a potřebujeme zkopírovat data ze samostatného disku, který byl původně připojen přes obyčejný (non-RAID) řadič.

Tímto termínem se také někdy označují externí expanzní „storage“ skříně, které neobsahují vlastní SCSI-to-SCSI RAID řadič.

Další typy polí

Existují další, méně používané „číslované“ typy polí: RAID-3, 4, 7.

Spare drive = náhradní disk

Většina RAID řadičů umožňuje označit další disk v systému jako "spare", tedy náhraní. Na tento disk přejde automaticky (bez zásahu obsluhy) provoz při výpadku některého aktivního disku v běžícím poli RAID1 nebo RAID5. Na obsluhu pak zbyde úkol vyměnit havarovaný fyzický disk a případně nový disk označit jako spare.

Obnova redundance po degradaci pole

Degradovaný RAID 1 (mirror) a RAID 5 lze opravit – obnovit redundanci pole.

Než započneme s obnovou, je třeba vyměnit havarovaný disk nebo odstranit poruchu – pokud hardware umožňuje hot swap a firmware a software řadiče nekladou překážky, obejde se výměna disku i bez restartu počítače.

Některé řadiče při obnově pole vyžadují, aby byl náhradní disk napřed explicitně inicializován a označen jako spare, a teprve poté ho pole přijme a začne s dopočítáváním dat.

Jiné řadiče se spokojí s fyzickou výměnou disku – automatická obnova zafunguje, pokud je detekován původní disk (kterému například jen vypadlo napájení) nebo čistý nový disk. Automatická obnova nenastartuje, pokud nově zasunutý disk již patřil do nějakého jiného pole či obsahuje platnou standardní tabulku rozdělení (řadič se bojí, aby nedošlo ke ztrátě dat na nově zasunutém disku).

Pokročilejší konfigurace polí

Flexibilnější svazky

U dražších řadičů lze vytvářet "virtuální svazky" menší než je maximální dostupná kapacita - lze tak vytvořit například dvě pole RAID5 napříč čtyřmi disky tak, že jedno pole zabírá dolních 30% kapacity každého disku a druhé pole zabírá zbylých 70%.

Konkrétní fyzické disky lze ponechat mimo pole a přistupovat k nim přímo apod.

Hierarchická pole

Některé řadiče umožňují vytvářet hierarchická pole, například dvouúrovňová: na nižší úrovni se několik fyzických disků složí do pole zvaného "virtuální disk" a na vyšší úrovni se několik virtuálních disků složí do "virtuálního svazku".

Na každé úrovni může být použit jiný algoritmus (typ/úroveň pole - 0,1,5).

V systému lze potom vyhrazovat lokální a globální náhradní disky apod.

Pozor, v praktických implementacích obvykle na horní vrstvě hierarchie nelze použít jiný „RAID level“ než 0 (údajně kvůli výkonu) a možnost definovat více než 2 vrstvy se reálně nevyskytuje.

SAN = sítě pro ukládání dat

Zkratka SAN znamená Storage Area Network. Síť se skládá z inteligentních jednotek polí, nízkoúrovňových síťových prvků zvaných "SAN switch" a optické kabeláže.

Klíčovým standardem v této oblasti je FiberChannel.

Sítě SAN umožňují geografické zálohování "diskových" kapacit, pružné přidělování prostoru různým uživatelským strojům a právnickým subjektům, rozšiřování fyzické kapacity polí za běhu apod.

IDE/ATA/SATA vs. SCSI vs. FiberChannel

Jedná se o různé sběrnice pro připojení disků, případně jiných zařízení.

IDE	MBps
IDE (ATA) PIO 0	3,33
IDE (ATA) PIO 1	5,22
IDE (ATA) PIO 2	8,33
IDE Multiword DMA 0	4,16
IDE Multiword DMA 1	13,33
IDE Multiword DMA 2	16,66
EIDE (Fast ATA-2) PIO 3	11,11
EIDE (Fast ATA-2) PIO 4	16,66
Ultra-DMA/0 (ATA/16)	16,66
Ultra-DMA/1 (ATA/25)	25,0
Ultra-DMA/2 (ATA/33)	33,33
Ultra-DMA/3 (ATA/44)	44,4
Ultra-DMA/4 (Ultra-ATA/66)	66,66
Ultra-DMA/5 (Ultra-ATA/100)	100
Ultra-DMA/6 (Ultra-ATA/133)	133
Serial-ATA/150	150
Serial-ATA II / 300	300
(Serial-ATA/600 ?)	(600)

SCSI	Bitů	MBps
Asynchronní	8	5
Fast	8	10
Fast Wide	16	20
Ultra	8	20
Ultra Wide	16	40
Ultra2	8	40
Ultra2 Wide	16	80
Ultra160 (Ultra3)	16	160
Ultra320	16	320
FibreChannel 1Gb	2 opt	100
FibreChannel 2Gb	2 opt	200
FibreChannel 4Gb	2 opt	400
(FibreChannel 10Gb)	(2 opt)	(1000)
SAS/300	2	300
(SAS/600 ?)	(2)	(600)
(SAS/1200 ??)	(2)	(1200)

SCSI je patrně nejstarší a nejsolidnější – dodnes jde o lehce dražší technologii, která se montuje hlavně do serverů. V průběhu času prošla několika vývojovými generacemi – odstupujícím králem je Ultra320 SCSI (320 MB/s half duplex). V oblasti diskových polí existuje související standard SCA, který standardizuje konektory, elektroniku a rozšíření sběrnicových protokolů SCSI pro hot-swap funkce (standardní rozhraní SES a SAF-TE). SCA se vyskytuje v kombinaci s Ultra2-Wide (80 MBps), U160 a U320 SCSI.

Jako nástupník SCSI se dere na výsluní standard Serial-Attached-SCSI (SAS) s počáteční rychlostí 300 MBps full duplex. Nejde již o paralelní sběrnici – ve skutečnosti se jedná o siamské dvojče standardu SATA-II/300. Mezi SAS a SATA-II/300 existuje jistá omezená kompatibilita. SAS ovšem na rozdíl od SATA-II podporuje podobné multi-lane a „switched“ topologie jako PCI Express. Struktura adresy SAS zařízení (disku) je vzdáleně podobná adrese na SCSI nebo FibreChannelu, ale poněkud odlišně strukturovaná (disk:enclosure:slot oproti channel:ID:LUN resp. NAI:AL_PA).

FibreChannel je optická technologie, částečně nástupce SCSI, spíš ale standard pro sítě SAN. Vyskytuje se v kapacitách 1, 2 a 4 Gbps (giga bity, nikoli bajty) – tj. cca 100, 200 a 400 MBps full duplex. Ve vývojové fázi (a snad v roli „stacking portů“ firmy Qlogic) je standard FC 10 Gbps.

IDE/EIDE/ATA/ATAPI je levná disková technologie do stolních a kancelářských počítačů. Také prošla několika vývojovými generacemi, nejnovější jsou UltraATA/133 a SerialATA-II/300 (133 a 300 MB/s). Standard SerialATA a zejm. SATA-II již zahrnuje jistou podporu pro hot-swap – výrobci jí často říkají SAF-TE, ale zdá se, že řadič komunikuje s hot-swap backplanem „out of band“ po I2C, tj. nikoli po samotné diskové sběrnici tak jako u SCSI.

První generace standardu SATA (SATA150) byla v zásadě sériovou náhražkou paralelního IDE – výhodou sériové sběrnice jsou užší kabely, jednodušší konektory, menší počet potřebných budičů sběrnice -> menší spotřeba. Již první generace SATA standardu interně používala v zásadě SCSI příkazy (podobně jako nadstavba paralelního IDE zvaná ATAPI), ovšem bez mechanismu „Tagged Command Queueing“ známého z pravého SCSI, takže vlastně nebylo možno využít full-duplexní povahu sběrnice SATA.

„Frontování požadavků“ nazvané v tomto případě NCQ (Native Command Queueing) přidal až standard SATA-II, obvykle spojovaný s rychlostí 300 MBps.

V průběhu vývoje specifikace nastalo lehké zmatení pojmů: SATA-II bez přídomku 300 může znamenat v zásadě SATA150 s podporou NCQ... proto pozor při studiu ceníků.

Taky pozor na skutečnost, že disky Hitachi SATA-II/300 (aktuální generace, tj. po 500 GB včetně) se prodávají nastavené na maximální rychlost 150 MBps. Chcete-li využít rychlost 300 MBps, je třeba disk přepnout pomocí softwarové utility od firmy Hitachi – která ovšem funguje pouze na SATA řadičích zabudovaných v čipsetu, které ještě zvládají emulovat klasické I/O porty IDE. Momentálně toto platí o integrovaných čipsetech Intel a VIA. Naopak tento požadavek nesplňují přídavné řadiče Promise nebo Marvell SATA-II (které se instalují jako SCSI řadič a vyžadují specifický ovladač), stejně tak jako všechny možné RAIDy pro disky SATA-II, ať už v podobě PCI karty nebo externího pole připojeného přes SCSI. Proto pokud chcete Hitachi disk provozovat na 300 MBps v RAIDu, musíte ho napřed přes nějaký motherboard překonfigurovat na vyšší rychlost. Nebo se spokojit s rychlostí 150 MBps + NCQ, což taky není k zahození, vzhledem k tomu, že SATA disky vč. 500GB modelů nepřekračují trvalou rychlost 80 MBps.

Klasická (velká a drahá) externí disková pole obsahují řadič typu "SCSI-to-SCSI“. K hostitelskému počítači se taková pole připojují přes SCSI sběrnici a hostitelský SCSI řadič (odtud správný název "host bus adapter"). Nověji též přes FibreChannel.

RAID řadiče do PC jsou relativně novější záležitost – zasouvají se přímo do sběrnice PCI. K řadiči se pak pomocí sběrnice IDE nebo SCSI připojují přímo fyzické disky.

Zajímavým hybridem mezi ”drahými” a ”levnými” poli jsou externí pole typu IDE-to-SCSI nebo SATA-to-SCSI: k počítačům se připojují pomocí SCSI, ale uvnitř používají IDE/SATA disky.

Způsob adresace na SCSI sběrnici velí, že na jednu sběrnici lze připojit až 16 zařízení. Jedním z těchto zařízení je samotný SCSI řadič (host adapter), druhým může být řídící čip hot-swapové mechaniky (SCA backplane processor s rozhraním SES nebo SAF-TE). Čili teoreticky lze připojit přinejmenším 14 disků. Na druhou stranu je třeba podotknout, že čtyři dnešní SCSI disky snadno vyčerpají kapacitu sběrnice U320, takže omezujícím faktorem počtu zařízení na jednu sběrnici bude typicky celková průchodnost.

Na jeden kanál IDE řadiče lze tradičně připojit dva disky, existují experimenty se čtyřmi disky na kanál. V případě použití v RAIDu je ovšem třeba si uvědomit, že jediný disk může v případě poruchy v elektronice nebo při špatném kontaktu na konektoru snadno vyřadit z provozu oba disky na daném kanálu, což znamená nevyhnutelnou havárii jakéhokoli redundantního pole (přechodnou a možná i trvalou). Především však způsob zapojení master/slave výrazně snižuje celkovou průchodnost IDE kanálu. Proto se doporučuje v IDE RAIDech připojovat každý disk samostatně na jeho vlastní IDE kanál a ostatně IDE RAID řadiče tento způsob připojení většinou přímo vyžadují.

Software vs. Hardware RAID (řadiče do PC)

Hardwarový RAID

Pole je vytvářeno hardwarem řadiče. Typicky jde o specializovaný RISC procesor, případně vybavený HW urychlovačem operací pro dopočítávání parity (XOR čip). HW RAID řadiče bývají vybaveny velkou diskovou cache – řádově několik desítek až stovek MB RAM.
Takový řadič umí fungovat nezávisle na stavu operačního systému a aplikací hostitelského počítače - například obnovuje pole během restartu počítače, kromě krátké pauzy během hardwarového resetu. Zvládá dokonce zpracovávat několik úloh najednou - například obnovu několika polí paralelně.

Do hostitelského systému se hardwarový RAID řadič tváří obvykle jako obyčejný SCSI řadič a nakonfigurovaná pole prezentuje jako SCSI disky. Skutečné, fyzické disky ovšem nejsou v hostitelském operačním systému vidět. Takto se chovají jak SCSI RAID řadiče, tak IDE RAID řadiče – tj. řadiče, které vytvářejí emulované SCSI svazky nad fyzickými IDE disky. (Výjimkou je okrajová konfigurační volba zvaná ”pass-through”, kterou umožňují některé RAID řadiče.)

Historicky zdaleka nejoblíbenějším procesorem, kolem kterého výrobci navrhují své RAIDové řadiče, je rodina Intel IOP. Starší procesory z této rodiny, v éře U160 SCSI, byly vybaveny jádrem i960. Novější procesory Intel z rodiny IOP300 disponují jádrem Xscale. Tato změna se neobešla bez určitého ustrnutí na trhu RAIDů a poněkud zamíchala kartami.

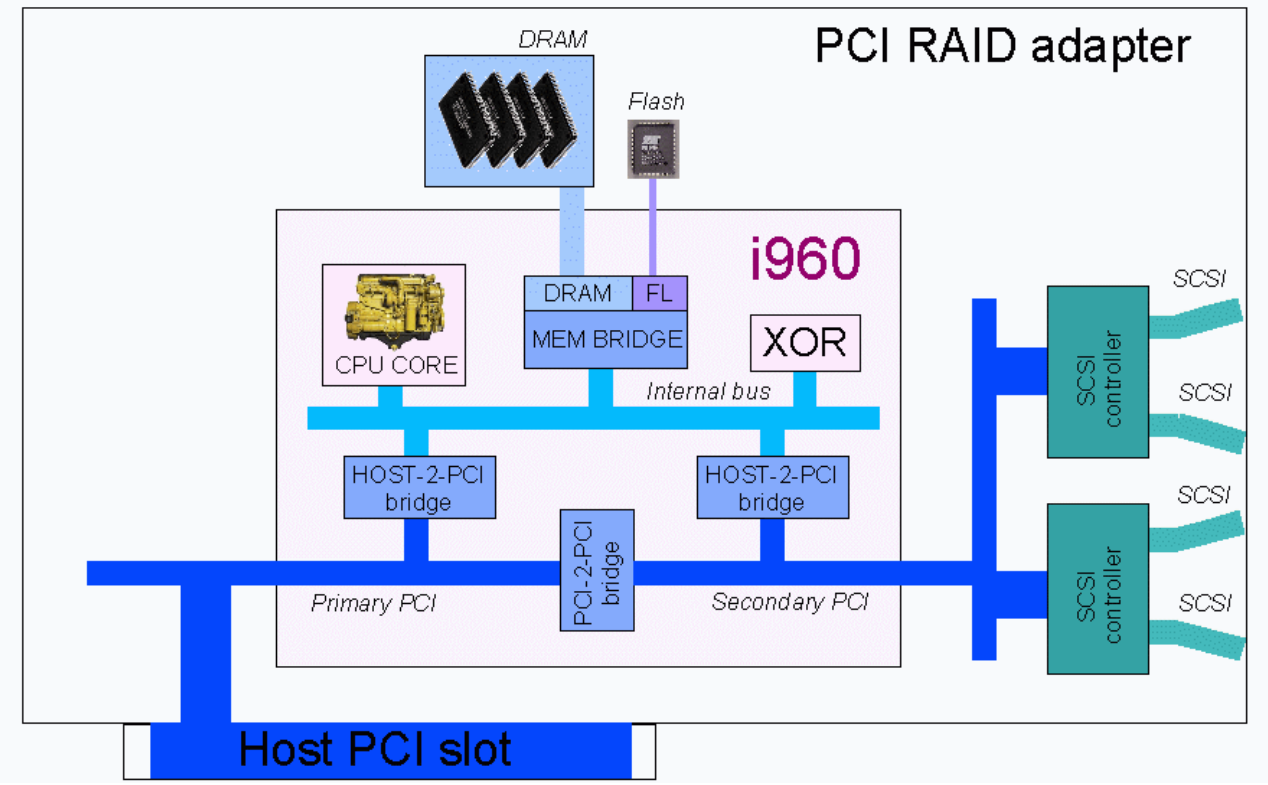
Někdy v roce 2005 se začaly objevovat analogicky uspořádané jednoúčelové čipy od dalších výrobců.

Firma AMCC (která tou dobou koupila 3ware) uvedla čip PowerPC 440SP/SPe.

Firma Adaptec exkluzivně využívá čipy RoC (Raid-on-Chip), na které bez velkého humbuku přešla se zvými FSA=AAC RAIDy od i960, a zřejmě se tak vyhnula procesorům Intel IOP3xx s jádrem Xscale. Zdá se, že dosud existují už dvě generace čipů RoC, pracovně je nazývejme takto:
- **RoC1: Adaptec 2130S, 2230S** – PCI-X, jádro snad MIPS nebo možná „přetaktované“ i960, on-chip SCSI řadič, dodávány s CLI nástrojem AACCLI
- **RoC2: Adaptec 2420SA, 2820SA, 4800SAS, 4805SAS** – PCI-X nebo PCI-e x8, jádro MIPS, RAID6, on-chip SATA2 nebo SAS řadič (4-8 kanálů), dodávány s CLI nástrojem ARCCONF

Těžko říci, kdo pro Adaptec čipy RoC vyrábí – možná Broadcom nebo Vitesse, možná každou generaci někdo jiný. Důležité je, že výkonově poměrně přesně odpovídají rodině Intel IOP3xx okolo IOP321/331.

Výše uvedené procesory firmy Intel i konkurenční se používají v hardwarových RAIDech pro SCSI i IDE/SATA disky. Typické uspořádání RAIDového řadiče s výše zmíněnými procesory vypadá takto:



Toto typické zapojení obsahuje několik vtipných řešení:

- XOR akcelerátor integrovaný na čipu RAID procesoru (novější procesory umí nejen XOR pro RAID5, ale taky XOR + Reed Solomon pro RAID 6)

- jako budiče diskových sběrnic jsou použity v zásadě obyčejné SCSI (resp. IDE, SATA) řadiče pro sběrnici PCI, které se vyskytují i samostatně na PCI kartách. Zde jsou použity jako podřízená zařízení procesoru i960/IOP, který s nimi komunikuje po sekundární PCI sběrnici. Výhody jsou zřejmé: RAIDové řadiče se dají vyvíjet jako skládačka, jejíž díly se mezi sebou baví dobře definovanými protokoly PCI. K dispozici je spousta dílů, které se dají použít jako podřízená zařízení.

- datové přenosy jdou do hlavní paměti řadiče, která funguje jako cache a ve které operuje XOR akcelerátor.

- procesor obsahuje PCI-to-PCI bridge mezi primární a sekundární sběrnicí, takže teoreticky by operační systém hostitelského počítače mohl mluvit přímo se zařízeními na sekundární sběrnici. Prakticky je toto obvykle znemožněno. Vestavěný PCI bridge procesoru i960/IOP umožňuje definovat “privátní” zařízení na sekundární sběrnici (takže operační systém nenajde prosté diskové řadiče), ba dokonce může zakázat bridgi hlásit se jako bridge. Navenek se pak karta tváří jako jediné PCI zařízení. Na kartách od různých výrobců vystupuje i960/IOP pod různými PCI VendorID a DeviceID.

Známe-li u konkrétního řadiče typ použitého procesoru z rodiny i960 či IOP resp. konkurenčních čipů obdobné koncepce (obvykle lze zjistit na webu), můžeme na základě níže uvedené tabulky spočítat teoretickou průchodnost – či alespoň odhadnout relativní vzájemné pozice různých řadičů na startovním poli.

Procesor (popř. IO chip + CPU core chip)	takt PCI [MHz]	šířka PCI [bity]	int.sběr.* [MHz]	takt jádra [MHz]	typ RAM*	kapacita RAM
Adaptec RoC2 [MIPS] (+6)	buď 133 nebo x8	buď 64 PCI-X nebo PCI-e x8	?	?	DDR	?
Adaptec RoC1 [MIPS? Xscale?]	133	64	?	?	DDR	?
AMCC PowerPC 440SPe+ (+6)	266 x8 x4	64 PCI-X 2.0 PCI-e x8 2x PCI-e x4	2x 166 128bit	až 667	DDR333 DDR2/667	16 GB 8 GB
AMCC PowerPC 440SP+ (+6)	266	3x64 PCI-X 2.0	2x 166 128bit	až 667	DDR333 DDR2/667	4 GB
IOP333 (i80333) (+6)	133 x8	2x 64 PCI-X 1x PCIe x8	333 64b	až 800	DDR333 DDR2/400	2 GB 1 GB
IOP332 (i80332)	133 x8	2x 64 PCI-X 1x PCIe x8	266 64b	až 800	DDR333 DDR2/400	2 GB 1 GB
IOP331 (i80331) (+6)	133	2x 64	266 64b	až 800	DDR333 DDR2/400	2 GB 1 GB
IOP219 (i80219**)	133	64	200 64b	až 600	DDR200	1 GB
IOP321 (i80321)	133	64	200 64b	až 600	DDR200	1 GB
IOP315 (i80314** + 2x i80200)	133	2x 64	133 64b	733	2xDDR200	12 GB!
IOP310 (i80312 + i80200)	66	2x 64	100 64b	733	SDRAM100	512 MB
IOP303 (i80303 =jádro i960)	66	2x 64	100 64b	100	SDRAM100	512 MB
IOP302 (i80302 =jádro i960)	66	2x 64	66 64b	100	SDRAM66	128 MB
i960RN (i80960RN)	33	2x 64	66 64b	100	SDRAM66	128 MB
i960RM,RS (i80960RM,RS**)	33	2x 32	66 64b	100	SDRAM66	128 MB
i960RD (i80960RD**)	33	2x 32	66 32b	66	SDRAM66	128 MB

**interní sběrnice Intel IOP je 64bitová, paměť je 64bitová nebo u starších 66 MHz variant volitelně 32bitová*
*** procesory i960 RS, RD a starší postrádají „Application Accelerator“, tj. hardware XOR.*
Totéž platí také pro IOP219, který vznikl odvozením z IOP321.
Také dvou-čipové řešení IOP315 nemá XOR akcelerátor, protože není určeno pro „storage“ aplikace.
+6 : procesory IOP333, novější revize IOP331, PowerPC 440 SP/SPe a RoC2 mají HW podporu RAID6

Architektura Intel Xscale použitá v nových čipech IOP310 a vyšších má instrukční sadu kompatibilní s jádrem ARM/Xscale – a je tedy zpětně nekompatibilní s vysloužilým jádrem i960. Aby to nebylo tak jednoduché, čipy 80302 a 80303 nesou obchodní označení IOP302 resp. IOP303, přestože mají jádro i960JT.

Čipsety IOP310 a IOP315 jsou dvoučipová řešení – v zásadě north bridge + samostatné Xscale jádro.

IOP219 je „ořezaná“ verze IOP321, určená pro jiné než „storage“ nasazení. Přesto např. Areca tento čip kombinuje s externím onboard XOR akcelerátorem a výsledek je překvapivě velmi výkonný.

Orientační empirické hodnoty sekvenční průchodnosti RAID5 jsou zhruba tyto:
IOP302/303 = 50-60 MBps (LSI, Adaptec), IOP321 = 100-120 MBps (LSI 320-2X, Accusys), Adaptec RoC = 120 MBps (2130S), IOP331 = 150 MBps, Areca IOP219 + ext. XOR = 200 MBps

Zdá se, že ukončení dalšího rozvoje architektury i960 kdesi na úrovni 100 Mips postavilo mnohé výrobce RAID řadičů před nelehký úděl. Potřeba vývoje nových, výkonnějších řadičů je nutila do přechodu na novou platformu, což znamená přeportovat velkou část firmwaru. Zdálo se, že IOP303 stačí právě tak na dva kanály U320 a výš se nikomu nechtělo. Tato situace se odráží v níže uvedené tabulce – snad dva roky jediné nové řešení, založené na IOP321, byl LSI MegaRAID 320-2X, prodávaný též pod nálepkou Intelu jako SRCU42X. Intel na své stránce věnované rodině IOP prezentoval řešení firmy LSI Logic jako zásadní success story.

Je tu ještě další důvod, proč procesory IOP s jádrem Xscale měly zpočátku těžký život. IOP310 znamenal dva čipy namísto jediného (IOP303). Naproti tomu IOP321, první rozumně použitelný IOP s jádrem Xscale, má pouze jeden port PCI-X. Proto řadiče s IOP321 jsou typicky externí řešení, která se k hostitelskému počítači připojují přes SCSI (Accusys, Areca). U této koncepce nevadí, že se procesor baví s diskovými řadiči i s „hostitelským“ SCSI/FC řadičem po jediném segmentu PCI. Patrně jediná existující implementace PCI RAIDu s procesorem IOP321, tj. LSI MegaRAID 320-2X, má proto navíc ještě externí PCI-to-PCI Bridge (velký stříbřitý čip s logem „Tundra“), který za sebou schovává samotný IOP321 a jemu podřízený dvoukanálový řadič sběrnice SCSI (patrně nějaký LSI). Víc čipů = víc nožiček = větší cena. A taky větší topný výkon...

Pro srovnání MegaRAID SATA300-8x je již založen na čipu IOP331 a proto bychom na něm externí PCI bridge hledali marně.

Příklady RAID karet

Výrobce a model	Procesor	Diskové kanály
LSI MegaRAID SAS 8480E	IOP333	8x SAS/300 ext.
LSI MegaRAID SAS 8408E	IOP333	8x SAS/300 int.
Adaptec 4800SAS (Marauder-X) Adaptec 4805SAS (Marauder-E)	RoC2 (PCI-X) RoC2 (PCI-e)	8x SAS/300 int. + 8x SAS/300 ext.
Adaptec ASR2230S (Lancer)	RoC1	2x U320 SCSI
Adaptec ASR2130S (Lancer)	RoC1	1x U320 SCSI
LSI Logic MegaRAID SCSI 320-4X	IOP321	4x U320 SCSI
Intel SCRU42X = MegaRAID 320-2X	IOP321	2x U320 SCSI
LSI Logic MegaRAID SCSI 320-2X	IOP321	2x U320 SCSI
LSI Logic MegaRAID SCSI 320-2	IOP303	2x U320 SCSI
LSI Logic MegaRAID SCSI 320-1	IOP302	1x U320 SCSI
ICP Vortex GDT8x24RZ	IOP303	2x U320 SCSI
Adaptec ASR2200S (Vulcan)	IOP303	2x U320 SCSI
Adaptec ASR2120S (Crusader)	IOP302	1x U320 SCSI
Adaptec 5400S (Mustang)	StrongARM@233	4x U160 SCSI
Adaptec ASR3410S	IOP303	4x U160 SCSI
Mylex AcceleRAID 352	i960RM	2x U160 SCSI
Mylex AcceleRAID 170	i960RM	1x U160 SCSI
Mylex AcceleRAID 160 == AcceleRAID 170LP	i960RS	1x U160 SCSI
Adaptec 2820SA (Intruder)	RoC2	8x SATA2/300
Adaptec 2420SA (Intruder)	RoC2	4x SATA2/300
Areca ARC1210 Areca ARC1220, 1230, 1260, 1270	IOP332 (PCI-e) IOP333 (PCI-e)	4x 8x, 12x, 16x, 24x SATA2/300
Areca ARC1110, 1120, 1130, 1160, 1170	IOP331	4x, 8x, 12x, 16x, 24x SATA2/300
LSI MegaRAID SATA300-8X	IOP331	8x SATA2/300 (8 disků)
Adaptec 21610SA	IOP303	16x SATA150 (16 disků)
Adaptec 2810SA	IOP303	8x SATA150 (8 disků)
Adaptec 2410SA	IOP302	4x SATA150 (4 disky)
Intel SRCS14L = ICP Vortex GDT8546RZ	IOP303	4x SATA150 (4 disky), 64 MB
ICP Vortex GDT8546RZ	IOP303	4x SATA150 (4 disky), 128 MB
LSI Logic MegaRAID SATA-8*	IOP302	8x SATA150 (8 disků)
LSI Logic MegaRAID SATA-6	IOP302	6x SATA150 (6 disků)
LSI Logic MegaRAID SATA-4	IOP302	4x SATA150 (4 disky)
Promise FastTrak SX6000	i960RM	6x UATA100 (6 disků)
LSI Logic MegaRAID I4	i960RS	4x UATA100 (4 disky)

**tento produkt se vyskytoval pouze v marketingové brožuře, bez obrázku – jinde na webu není*

Zajímavou a mladou značkou je Taiwanská firma Areca, která si rychle buduje jméno mezi odbornou veřejností výkonnými a komfortními RAIDy na bázi Intel IOP / Xscale. Firma odstartovala v oboru externích RAIDů, její první Xscale-based produkty byly založeny na čipu IOP321. Na základě zkušeností s externími RAIDy firma vyvinula také rodinu interních SATA2 RAIDů do sběrnic PCI-X a PCI Express, s čipy Intel IOP331 a IOP332/333.

Specialitou firmy je externí onboard XOR akcelerátor s podporou RAID6. Díky němu měla Areca podporu RAID6 mnohem dříve, než zbytek trhu – který RAID6 v potu tváře doplňuje teprve poté, co Intel a další výrobci procesorů tuto novinku narychlo implementovali ve svém hardwaru. Jakmile byl firmou Intel uveden čip IOP219, firma Areca jej použila ve svých externích RAIDech – v kombinaci s firemním externím XOR akcelerátorem je tento čip překvapivě výkonnější, než konkurenční produkty založené na IOP321 nebo IOP331.

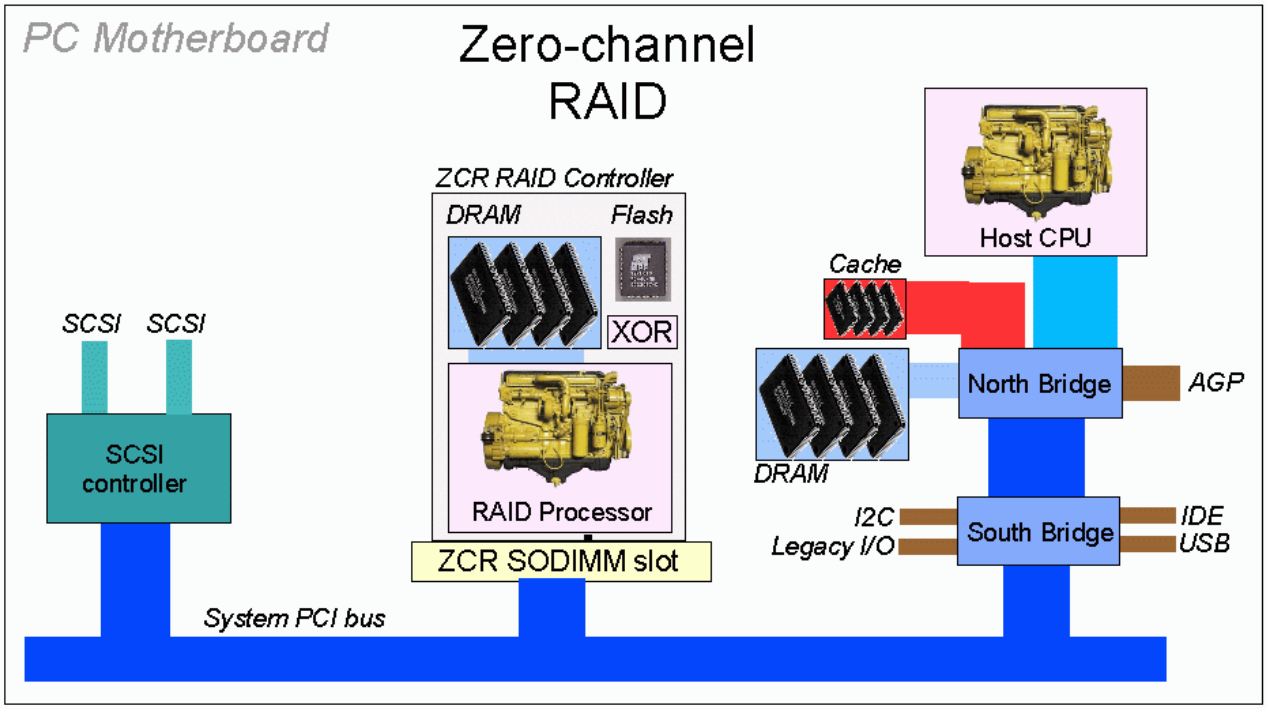
Mezi několik málo produktových řad, které nevyužívají procesory Intel IOP, patří externí RAID řadiče renomované firmy InforTrend, založené na čipech s architekturou PowerPC. Toto bylo svého času náležitě okomentováno v jedné z poznámek na firemním webu, zejména ve vztahu k výše popsanému předělu v architektuře Intelovských IOP procesorů. Firma Infortrend používá „vanilkové“ procesory PowerPC 750 od firmy IBM, které kombinuje se svým vlastním externím „ASICem“ (dvouportový PCI bridge + řadič RAM + XOR akcelerátor) – dlouho se prodával ASIC133 s pamětí SDRAM, přechod na ASIC266 s pamětí DDR proběhl zhruba v době, kdy se začaly prakticky více uplatňovat procesory Intel IOP s jádrem Xscale. Pozdější revize ASIC266 zvládají RAID6.

Na procesorech Intel také tradičně nejsou založeny IDE a SATA RAID řadiče firmy 3ware – jsou mezi nimi některé velmi zajímavé kousky, jako 8- a 12- a 16-kanálové UATA/133 / SATA / SATA2 řadiče. Klíčovou součástkou, kterou firma 3ware zmiňuje ve svých materiálech, je hardwarová „přepínací matice“ pro multiplexování datových toků k jednotlivým diskům (StorSwitch) – zdá se, že hlavní výhodou má být přímý přenos dat z akcelerovaných ATA/SATA kanálů na rychlou interní sběrnici řadiče a do jeho pracovní paměti (tj. žádná privátní PCI).

Svého času se nikde neuvádělo, kolik mají vlastně řadiče 3ware k dispozici RAM – není to oficiálně známo o řadách 7000 a 8500. Při bližším zkoumání fyzických exemplářů lze zjistit, že tyto řady měly cca 500 kB až 2 MB DRAM v roli cache pro data (plus malou SRAM pro řídící procesor). Jako řídící procesor se používal 16bitový mikrokontrolér značky SGS Thomson – tolik pro srovnání s 64bitovými RISC jádry Intel aj.
Řada 3ware 9500 (SATA) má zřejmě dosud slabý procesor, nicméně již disponuje 128 MB ECC SDRAM v modulu SO-DIMM, rozšiřitelnou na 256 MB.
Teprve řada 3ware 9550, vzniklá již pod křídly AMCC, má „dospělý“ procesor PowerPC a standardně 128 nebo 256 MB RAM. Nutno ovšem podotknout, že základním kamenem řadiče je nadále StorSwitch (nové generace) a použitý procesor PowerPC 405 je nějaké 32bitové univerzální embedded jádro od IBM, tj. nejedná se o storage procesor AMCC v pravém slova smyslu, jakým je např. „dělo“ AMCC PowerPC 440SPe+.

Značce 3ware nikdo nevezme jednu velikou zásluhu a v rámci daného tržního segmentu kvalitní řešení: legendární dvoukanálové IDE a SATA RAIDy 3ware 7000-2 a 8006-2 byly a jsou patrně jedinými *hardwarovými* IDE/SATA RAIDy na trhu, a potažmo ve své kategorii kralují kvalitou.

Mezi hardwarové RAIDy patří také jedno cenově výhodné řešení – tzv. řadiče Zero-Channel RAID. Jde o zásuvnou kartu do motherboardu do speciální ZCR patice (SODIMM). Karta obsahuje pouze RAID procesor a jeho pracovní paměť (velkou část zabírá cache RAIDu), ale žádné low-level řadiče/budiče SCSI/SATA sběrnic. ZCR totiž využívá obyčejné SCSI/SATA kanály, které jsou integrovány na hostitelské základní desce. Základní desky a počítače se SODIMM (mini-PCI) paticí pro ZCR se pak prodávají typicky s integrovaným U320 SCSI nebo SATA řadičem, u kterého je zmíněna možnost upgradu na RAID.



ZCR SODIMM slot totiž obsahuje sběrnici PCI. V momentě, kdy do základní desky zamontujeme ZCR kartičku, kartička vyřadí BIOS onboard SCSI kanálů a naopak zařadí do POST řetězce svůj vlastní BIOS. Takže při startu počítače se objeví RAID řadič namísto původního obyčejného SCSI. RAID procesor na ZCR modulu pak přenáší data ze SCSI kanálů do své paměti přes hostitelskou PCI sběrnici – a přes tutéž sběrnici je po zpracování předává hostitelskému systému. Největším problémem tohoto řešení je proto poměrně nízký výkon – úzkým místem je segment systémové PCI sběrnice, který ZCR využívá pro komunikaci dovnitř i ven. Na nízkém výkonu se ovšem podílí také celková ”lacinost” řešení, která diktuje použití malé cache, pomalého RAID procesoru apod.

Stručný přehled pohnuté historie HW RAIDů

Trh hardwarových RAIDů je neustále v pohybu, vedle technologického vývoje s ním cvičí především obchodní aliance a akvizice.

Svého času velice slavná samostatná značka Mylex, používající procesory Intel i960, začala koncem devadesátých let upadat – napřed ji v r.1999 koupila firma IBM, v r.2002 proběhl prodej „divize Mylex“ firmě LSI Logic. Pod křídly LSI Logic byl nakonec ukončen vývoj produktů z rodin Mylex AcceleRAID a ExtremeRAID.

Firma LSI Logic mezitím v roce 2001 převzala jinou slavnou značku: RAIDovou divizi firmy AMI, tj. především produktovou řadu MegaRAID spolu s veškerým know-how a 200 zaměstnanci. Tuto rodinu firma dále rozvíjí – v úzké spolupráci s firmou Intel proběhl u rodiny MegaRAID pod křídly LSI přechod od původních procesorů i960 a IOP302/303 k novějším IOP321/331/333 (v letech 2003-2006). S tím souvisí, že RAIDy LSI jsou ještě v roce 2006 prodávány také pod značkou Intel.

Další samostatná a kvalitní značka RAIDů, původně soukromá německá firma ICP Vortex, byla v roce 2001 koupena firmou Intel. Portfolio firmy ustrnulo u řešení založených na IOP302/303 – nenechte se zmást irelevantní zmínkou o IOP310 v tiskové zprávě o převzetí. Jediným praktickým důsledkem převzetí bylo, že Intel lepil svoje logo na některé starší RAIDy této značky.

Tato Intelská epizoda skončila v roce 2003 prodejem divize ICP Vortex firmě Adaptec. Adaptec ještě nějakou dobu nechal značku ICP Vortex samostatně naživu, potažmo firma Intel ještě začátkem roku 2005 dodávala některé starší řadiče ICP Vortex se svým logem. Samostatná existence značky ICP Vortex patrně prakticky skončila v průběhu roku 2005, a to skladovou nedostupností jak originálních řadičů ICP Vortex, tak jejich ekvivalentů pod značkou Intel. Zdá se, že Adaptec pod značkou ICP Vortex nadále nabízí své vlastní produkty založené na IOP302/303 a RoC, ale pravděpodobně se jedná o labutí píseň této tradiční značky.

Svého času samostatná firma 3ware byla na jaře 2004 převzata firmou AMCC. Firma AMCC se na přelomu století vypracovala ve velmi aktivního dodavatele čipových řešení (jednostranně zaměřených aplikačních procesorů), původně především pro oblast telekomunikací a počítačových sítí, nově ovšem také pro oblast storage technologií. Vedle řadičů značky 3ware má v portfoliu výkonné IO procesory s jádrem PowerPC, které přímo konkurují Intelu – zdá se, že v roce 2006 by se mohly reálně prosadit na trhu koncových produktů. Jádro PowerPC naznačuje úzké vztahy s IBM.

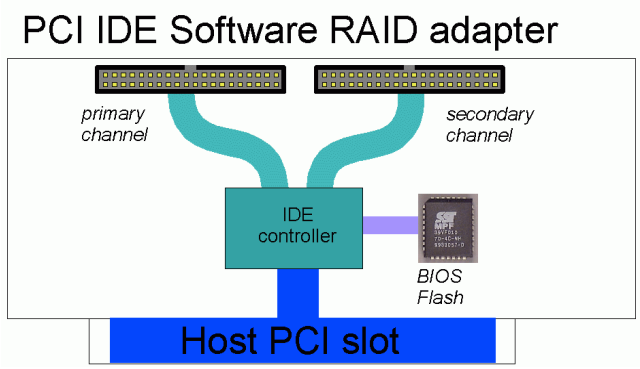
Akvizice značky 3ware firmou AMCC se projevila v prvé řadě přechodným výpadkem dodávek řadičů 3ware v krátkém období (cca 1-2 měsíce), kdy firma AMCC sjednotila celosvětovou distribuční síť. V letech 2005 a 2006 značka 3ware nadále existuje a uvádí na trh nové rodiny výrobků.

Poněkud staršího data je akvizice firmy DPT firmou Adaptec (koncem r.1999). Tato fůze přinesla do portfolia Adaptecu rodinu RAIDů DPT (freebsd device asr) a bezpochyby také určité know-how. Rodina DPT ještě nejakou dobu přežívala v podobě zero-channel RAIDů, jako je Adaptec ZCR2005 a 2015. V porovnání s koncepčně mladšími RAIDy rodiny FSA (freebsd device aac) měla ovšem rodina DPT podstatně slabší interní inteligenci a slabší podporu v obslužném softwaru, nemluvě o historicky daném nižším výkonu.

Rodina PCI RAID řadičů Adaptec FSA/AAC, charakterizovaná především svou společnou skladebností a logikou RAIDu a jednotným ovládacím softwarem, je zjevně jakožto perspektivní vývojová větev nadále rozvíjena a obohacována o nové přírůstky. V průběhu vývoje došlo k úspěšnému přechodu z jader s architekturou i960 na MIPS (RoC) – překvapivě nikoli na Intel ARM/Xscale, přestože praotec rodiny, ASR5400S, měl procesor StrongARM. Pozor, starší verze AAC firmwaru nejsou kompatibilní s novými verzemi ovládacích utilit (CLI, ale hlavně storage manager) a naopak – patrně jde o přirozený důsledek dlouholetého vývoje.

Softwarový RAID

Pole je udržováno softwarem (ovladači), který běží na hostitelském CPU, s určitou minimální asistencí ze strany hardware, nebo spíše bez ní. Pole typicky při mirroringu zapisuje na oba disky paralelně a čte jenom z jednoho. Paralelní zápis obvykle není transparentně ošetřen hardwarem, data musí skutečně dvakrát protéct přes relevantní systémové sběrnice (paměťová, northbridge-southbridge, PCI).

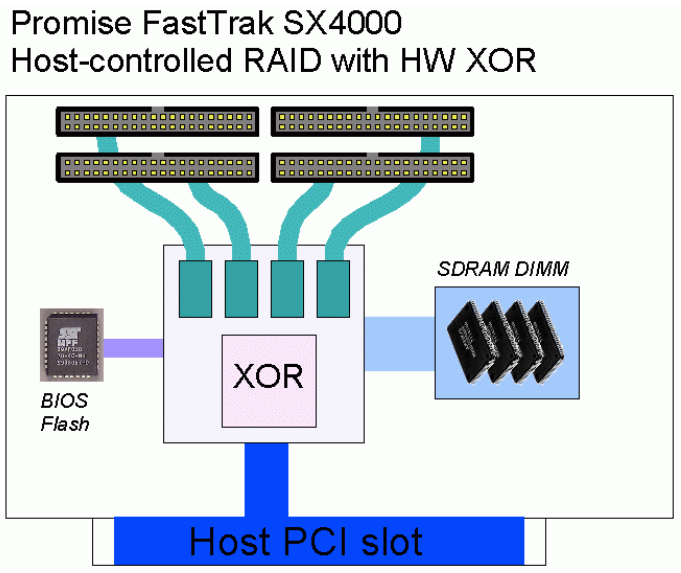


Hardware takového "RAID" řadiče a jeho BIOS se dále nejjednodušším možným způsobem postarají o rozlišení "čistých" disků a disků účastnících se v nějakém poli při bootu, aby operační systém bootoval stejně, ať už je mirrorové pole v pořádku, nebo má jeden či druhý disk vadný. U obzvlášť nekvalitních (a levných) řadičů je v případě havárie a výměny disku zapotřebí vyvolat obnovu pole ručně v BIOSu - automaticky se nespustí.

BIOS a ovladače jsou pochopitelně napsané tak, aby pole nešlo vytvořit na jiném IDE řadiči než na tom, se kterým se dodávají – ačkoli v diskusních skupinách probleskly zprávy, že například ovladače firmy Promise se dají hacknout, aby běžely na jakémkoli obyčejném IDE řadiči.

"Softwarové" RAID řadiče se prakticky vyskytují výhradně v provedení IDE. Pokud uvažujeme o SCSI řadiči, není tedy třeba se obávat, že by zatěžoval procesor hostitelského počítače. Přesněji řečeno, Adaptec u svých *obyčejných* U320 řadičů zavedl jako konfigurační volbu v BIOSu tzv. ”HostRaid”, ale prodává je nadále jako obyčejné SCSI řadiče.

Softwarový RAID typicky podporuje pouze RAID0, RAID1 a JBOD, tj. typy polí, kde většinu práce při softwarové emulaci zařídí DMA. Při provozu pole RAID5 je totiž třeba chroupat bajt po bajtu celý objem dat operací XOR. Tato operace je při daném objemu dat výpočetně velice náročná, takže by hodně zatěžovala hostitelský procesor – takovou ostudu si patrně žádný výrobce SW RAIDu nedovolí. Proto jsou vícekanálové IDE RAID řadiče prakticky výhradně hardwarové – protože u více než dvou disků připadá v úvahu RAID5.



Někde na půl cesty mezi HW a SW RAIDem stojí některé vícekanálové řadiče firmy Promise (např. FastTrak S150 SX4 nebo SX4000), které nemají autonomní procesor, ale obsahují v samostatném pouzdře ASIC pro akceleraci XOR operací. Nejde tedy o plnohodnotný HW RAID. Budiž tomuto řešení polehčující okolností, že má na kartě s řadiči a XOR akcelerátorem ještě svou vlastní RAM (až 256 MB), která slouží jako disková cache a především jako buffer pro provádění XOR operací – takže celý objem čtených či zapisovaných dat neputuje po systémových sběrnicích několikrát tam

a zpátky, ale je přenesen pouze jednou. Hardwarový akcelerátor také provádí mirroring v režimu RAID1, takže i při zápisu data putují po systémových sběrnicích pouze jednou. Přesto na procesor a systém přerušení hostitelského počítače zbývá poměrně dost práce s mapováním virtuálního svazku (pole) na fyzické sektory jednotlivých disků a s úkolováním XOR ASICu.

Existují také čistě softwarová řešení, z nichž nejznámější je asi nativní podpora SW RAIDu v Linuxovém jádře. Toto řešení funguje nad jakýmkoli fyzickým diskem, připojeným přes jakýkoli SCSI nebo IDE řadič – ba dokonce nad jakýmkoli jiným "blokovým zařízením". Podobně ve Windows NT Server (také 2000,XP) existuje již v základní výbavě podpora mirroringu disků.

Jaký pořídit RAID? Hardwarový nebo softwarový?

Někteří autoři argumentují, že softwarový RAID na dnešních gigahertzových procesorech je leckdy výkonnější než hardwarový RAID, a při srovnatelném výkonu o hodně lacinější. Má to ale některé již zmíněné háčky:

1) softwarově emulovat lze reálně pouze RAID0, 1 a JBOD. RAID5 v komerčních SW RAIDech není implementován, open-source implementace ukazují proč – žere desítky procent CPU.
2) softwarová emulace spotřebovává při diskových operacích procesorový výkon a kapacitu sběrnic a zvyšuje počet obsluhovaných přerušení, takže
3) zmiňovaného vysokého výkonu dosáhne pouze v případě, že systém je kromě diskových operací nezatížený (což se stává prakticky jen při jednostranném benchmarkování diskového systému)
4) softwarová emulace RAIDu se nehodí pro systémy vytížené jinou činností (servery), kde se výkonnostní výhoda oproti HW RAIDu ztrácí (HW RAID není přímo ovlivněn vytížením hostitelského CPU)
5) havárie operačního systému může u softwarového RAIDu poškodit integritu pole. Naproti tomu hardware RAID je zapouzdřený – integritu pole si hlídá nezávisle na stavu operačního systému na hostitelském počítači.
6) konkrétní praktické implementace softwarového RAIDu jakožto laciná řešení mívají horší chování při výpadku – menší komfort pro obsluhu, správcovský software je v kritických oblastech "řešení výpadku" méně přehledný a více chybový, proklamovaná obnova na pozadí se v kritické situaci nemusí povést apod. Je třeba hledat a předem testovat.

Softwarový RAID je tedy spíše levné řešení, které se hodí pro stolní systémy a malé servery, jejichž průměrné zatížení je spíše malé a kde nevadí kratší výpadek potřebný k případné offline obnově pole. Při dnešní ceně velikých IDE disků může softwarový mirror za mizivou cenu výrazně zlepšit zabezpečení takového malého serveru proti ztrátě dat a výpadku síťových služeb. (Pozor, jak upozorňuje většina autorů, tohle ještě neznamená, že můžeme přestat zálohovat!)

Naopak kvalitní a drahý hardwarový RAID je jedinou správnou volbou u systémů, které pracují pod vyšší průměrnou zátěží, jejich systémové sběrnice mají na práci i jiné věci než jen komunikaci s diskovým systémem, čeká se od nich vysoká průchodnost při všech úrovních zátěže a každá minuta výpadku stojí peníze, takže možnost obnovy pole na pozadí za plného provozu výrazně přispívá ke klidnému spánku správce, jeho nadřízených a uživatelů.

Pokud už se rozhodneme pro softwarový RAID, je tu další paradoxní skutečnost: nativní softwarová řešení, dodávaná zdarma v ceně dnešních operačních systémů, jsou díky přirozenému sladění s operačním systémem (tj. díky lepšímu vývojovému zázemí výrobce operačního systému) leckdy v běžném provozu stabilnější, komfortnější a výkonnější než softwarové RAIDy třetích stran, dodávané spolu s přídavným pseudo-RAIDovým hardwarem. Toto platí zejména v Linuxu – viz samostatný dokument na toto téma.

Jinými slovy, v dnešní době je třeba se před koupí "softwarového RAIDového řadiče" velmi zamyslet, zda nevyhazujeme peníze oknem. Dnešní UltraATA řadiče integrované na základních deskách nijak nezaostávají za hardwarem, který se prodává jako součást balíkových softwarových IDE RAIDů. Dokonce pokud je "RAIDový" IDE řadič integrovaný na Vaší základní desce, možná bude lepší používat ho jako obecný IDE řadič a spustit nad ním hardwarově nezávislý softwarový RAID dodávaný s operačním systémem. Definitivní volba je ovšem u různých typů a výrobců hardwaru velmi individuální – je třeba testovat.

Jedinou oblastí, kde proprietární softwarový IDE RAID dosud vede, je podpora hot-swap rámečků. Podpora pro hot-swap mechaniky SES a SAF-TE totiž není v operačních systémech běžná, s podporou přidávání a odebírání diskových zařízení za běhu je to ještě horší.

Dá se nějak na první pohled rozeznat SW a HW RAID řadič?

Ano, dá - podle počtu a velikosti pouzder integrovaných obvodů na kartě. Výrobci mají dokonce fotografie běžně vystavené na webu, takže není třeba kupovat zajíce v pytli.

"Softwarový RAID řadič" v provedení IDE má na desce typicky pouze dva větší integrované obvody:

1) konvenční dvoukanálový IDE/UltraATA řadič a
2) paměť Flash/EEPROM, která obsahuje BIOS.

Na serveru www.anandtech.com toto bylo okomentováno doslova tak, že RAIDový adaptér tohoto typu "se skládá z obyčejného IDE řadiče a EPROMky".

Naproti tomu hardwarový RAID řadič má na desce přinejmenším tyto velké součástky:

1) RAID procesor, simulující navenek SCSI zařízení a spravující pole – typicky s jádrem i960
2) přinejmenším jeden obyčejný diskový řadič, jedno- či dvoukanálový, IDE nebo SCSI. Na desce vícekanálového RAID adaptéru je typicky N/2 dvoukanálových diskových řadičů. Na desce je totiž privátní PCI sběrnice mezi RAID procesorem a diskovými řadiči.
3) paměti DRAM pro firmware a data RAID procesoru (několik pouzder SDRAM nebo DDR) – největší objem paměti za provozu zabírá disková cache.
4) Flash/EEPROM pro firmware a BIOS řadiče (pokud není integrována on-chip na RAID procesoru)
5) jednoúčelový čip pro akceleraci XOR operací, na kterých je založeno dopočítávání parity v polích RAID4/RAID5 (může být integrován on-chip na RAID procesoru, což je i případ i960)

Poznámka k výměnám a mazání disků

Většina RAIDů si konfiguraci pole ukládá na disky (a snad také někam do eprom/nvram na desce).

Takže na disku, který byl jednou přiřazen do pole, je od té chvíle napsáno, že se účastnil pole, jakého typu, jaké disky se ho účastnily (SCSI kanál/Device_ID nebo IDE kanál/pozice) a na jaké pozici byl umístěn aktuální disk.

Toto ukládání konfigurace na disk má usnadnit obnovu pole po havárii řadiče.

Pokud zhavaruje řadič, je třeba po jeho výměně připojit disky na původní pozice – porušení původního číslování disků může řadiči pořádně zamotat hlavu.

Některé řadiče se s tím vyrovnají, pouze BIOS řadiče při bootu upozorní, že si disky v RAIDu vyměnily místa a zeptá se, zda si správce přeje změnu konfigurace přijmout.

Také se ale může stát, že řadič po proházení disků vidí dvě degradovaná pole namísto jednoho, které je v pořádku.

Problémy mohou nastat především ve chvíli, kdy se snažíme obnovit redundanci pole po degradaci (tj. po havárii disku) - vložíme náhradní disk, a ejhle, objeví se další neúplné pole a řadič odmítá použít nově vložený disk pro obnovu degradovaného pole.

Náhradní disk totiž nebyl čistý, už na něm kdysi nějaké pole bylo, patrně pochází z rozebraného stoje nebo jsme ho použili při "laboratorních" experimentech s konfigurací pole.

Pokud má řadič rozumný software, lze takového ”ducha” bývalého pole jednoduše zrušit a disk inicializovat. Pokud toto z nějakého důvodu nezabere, je třeba takový "RAIDem poznamenaný" disk smazat – přepsat nulovými daty, aby se choval jako čerstvě vybalený z krabice. Tuto mazací operaci je třeba provést na jakémkoli jiném, obyčejném řadiči (IDE nebo SCSI) – tj. na řadiči bez RAIDových schopností.

Smazání disku lze provést nejsnáze v unixu příkazem

```
cp /dev/zero /dev/<blokové zařízení disku
```

Alternativně lze použít program dd:

```
dd if=/dev/zero of=/dev/<blokové zařízení disku>
```

Ten se ale pro přepis celého disku příliš nehodí – dělá sync() po každém bloku, takže mu zápis trvá poměrně dlouho. Zvětšení bloku zajisté pomůže, ale přesto je vhodnější použít cp nebo cat – na konci disku nezůstane nesmazaný ”nedělitelný zbytek” a zápis bude beztak rychlejší, protože cp a cat využijí write-back buffering jádra.

Blokové zařízení disku se např. pod Linuxem bude jmenovat /dev/hda až hdf nebo sda až sdf (hd = IDE disk, sd = SCSI disk, následuje pořadové písmeno – při větším počtu kanálů a disků může být poslední písmeno i vyšší než ”F”).

Patrně by se našla i nějaká DOSová utilita, která umí totéž. Někteří čtenáři doporučují Norton Wipe-info v režimu „celý disk“ a přímou editaci sektorů na disku umí třeba Norton Disk Editor.

Většina řadičů při konfiguraci prostého mirroru ukládá data na disky tak, že když disk vymontujeme z pole a připojíme ho na obyčejný (non-RAID) řadič, lze na něm obvykle najít korektní partition table a dokonce se z něj pokusí nastartovat operační systém. Toto funguje zejména u jednodušších řadičů, které mirrorují celý disk – tj. neumí vytvořit napříč dvěma fyzickými disky více ”zrcadlených logických svazků” (virtuálních disků).

Zdá se tedy, že jednodušší RAIDové řadiče v případě mirroringu ponechají na disku MBR s tabulkou rozdělení, která je v zásadě na první pohled korektní. Režijní blok dat týkající se pole uloží řadič kamsi do "neviditelného prostoru", který si vyrobí tak, že část prostoru fyzického disku nepoužije pro provoz pole.

Pokud takový "RAIDem poznamenaný" disk chceme použít samostatně na obyčejném řadiči, bude zřejmě také vhodné ho vyčistit. RAIDem poznamenaný MBR může obsahovat nesprávné informace o začátku a konci použitelného místa apod. - disk by při přímém použití operačním systémem nemusel fungovat správně. To platí i v případě prostého zrušení a znovuvytvoření oddílů FDISKem bez zrušení celé partition table.

Pro přímé použití operačním systémem stačí smazat pár prvních sektorů - např. příkazem dd s následujícími parametry:

dd if=/dev/zero of=<blokové zařízení disku> bs=1k count=1000

Pokud po této operaci spustíme fdisk, nenajde partition table a bude se tvářit, že je disk čistý jako z výroby (a vytvoří novou tabulku rozdělení).

Bohužel ne vždy zafunguje smazat pouze MBR a pár prvních sektorů – zejména pro opětovné použití v poli je vhodné smazat také režijní blok RAIDu, aby řadič nedetekoval dávno zapomenuté pole, které nebylo před odpojením a odložením disku korektně zrušeno. Pak nezbývá než přepsat celý disk. Trvá to déle, ale výsledek je stoprocentní.

Na okraj ještě poznámka pro zvídavé čtenáře: zmíněný ”neviditelný” režijní blok je na fyzickém disku uložen typicky na začátku disku hned za MBR, nebo na samém konci disku (na posledním cylindru, stopě apod.). Jeho přesnou polohu lze zjistit například tak, že založíme prázdné pole na ”vynulovaném” disku a následně tento disk prohledáme sektor po sektoru na obyčejném (non-RAID) řadiči. Znalost pozice řežijního bloku u konkrétního modelu RAIDového řadiče lze použít pro rychlejší hromadné čištění RAIDem poznamenaných disků.

Lze tedy mravoučně shrnout, že je jedině vhodné každý disk, který vyřadíme z pole a uschováme pro pozdější potřebu, pro jistotu před odložením smazat, aby později nemohl způsobit problémy.

Ostatní poznámky

Lavinový efekt

Některá literatura uvádí, že při degradaci a obnově redundance na polích typu mirror a RAID5 může dojít k zajímavému (a fatálnímu) lavinovému jevu, kdy po pádu prvního disku následují v rychlém sledu další.

Hlavní příčinou je skutečnost, že pole obvykle průběžně nekontroluje bezvadnost celé plochy všech disků. Takže nepoužívané soubory například po delší dobu vůbec nejsou čteny ani na jednom disku. Kromě toho například pole typu mirror typicky optimalizuje rychlost přístupu tak, že čte pouze z jednoho z obou disků – z toho, jehož hlava je blíž požadovanému sektoru. A nezapomínejme na volné místo na virtuálním disku, které pochopitelně také nikdo nečte.

To vše znamená, že pozornosti RAIDového řadiče mohou delší dobu unikat vadné sektory na různých místech několika disků. Když pak pole objeví na jednom z disků vadný sektor, nahlásí chybu, správce vymění vadný disk a spustí na degradovaném poli obnovu redundance. Obnova redundance spočívá v tom, že řadič prochází celý prostor funkčních disků a dopočítává obsah vyměněného disku. Při tomto důkladném procházení všech disků může narazit na další vadné sektory na nepoužívaném místě jiného disku, které dosud unikaly pozornosti. Což pak vypadá, jako že ”odešly dva disky najednou”. To může být do jisté míry pravda - výrazně kazové série disků se v minulosti u různých výrobců několikrát vyskytly. Není však příliš reálné, že by odešly dva disky v rozmezí několika hodin či minut.

Většinou by tedy pomohla průběžná kontrola povrchu všech disků v poli – takže by byl čas na nově vzniklé vadné sektory reagovat výměnou vadného disku. Průběžnou kontrolu disků lze u některých řadičů zapnout – pochopitelně do jisté míry na úkor průchodnosti pole.

Nouzový provoz havarovaného pole

Tradiční pravidlo zní, že když v RAIDu 1 nebo 5 vypadne víc než jeden disk, pole nenávratně zhavaruje. Řadič prohlásí oba disky za havarované a začne tvrdošíjně odmítat jakýkoli přístup k poli, i když se třeba jedná jen o pár sektorů na každém z obou disků. I když vlastně žádný z obou poškozených disků nezhavaroval fatálně, oba reagují na příkazy a nepoškozené sektory z nich lze přečíst.

Některé kvalitní řadiče však zvládají ”nouzový havarijní provoz” na takto poškozeném poli. Pokud má více disků v poli vadné sektory, které se však nesejdou pohromadě v konkrétní vnitřně redundantní ”sadě proužků”, řadič poskládá data z funkčních sektorů. I pokud jsou již některé ”sady proužků” nenávratně ztraceny, řadič prostě ohlásí problém na tomto konkrétním místě, ale okolní sektory virtuálního disku jsou nadále dostupné.

Je třeba podotknout že se jedná opravdu o krajní prostředek, jak zachránit alespoň část dat ze sektorů, které jsou ještě v pořádku. Problém je přinejmenším v tom, že fyzický disk se o přečtení vadného sektoru pokouší obvykle poměrně dlouho (jednotky až desítky sekund), než konečně vrátí chybu – takže i pokud řadič informaci o vadných sektorech cachuje, může v nouzovém havarijním režimu výrazně poklesnout výkon pole.

Kterak simulovat výpadek disku

Manuály k některým řadičům říkají, že jediným korektním způsobem, jak simulovat výpadek disku, je zvolit v menu XY položku ”fail drive”. A že vytrhnout disk z hot-swap rámečku není úplně korektní způsob simulace výpadku.

Linuxové ”software-RAID howto” ovšem prohlašuje bez obalu, že ”opravdový” výpadek disku dost dobře nasimulovat nelze. A to zřejmě nikoli proto, že se zmiňovaný materiál týká shodou okolností čistě softwarového řešení. Existuje totiž několik základních skupin typických poruch, z nichž některé nelze nasimulovat, aniž bychom disk doopravdy zničili.

Mezi reálné poruchy, které mohou nastat, patří cca tyto:

1) ”odcházení” sektorů. Disk jako celek je živý, ale není schopen číst data z některých sektorů. Porucha je většinou způsobena proniknutím nečistot do prostoru rotujících ploten, nebo mechanickou poruchou ložisek, ať už u ploten nebo u hlav – proto se tato porucha po ploše disku obvykle rychle šíří.

2) porucha elektroniky. Odejde řídící elektronika disku, takže přestane reagovat na příkazy. Pokud porucha elektroniky zasáhne i budiče IDE sběrnice, může zablokovat celý IDE kanál.

3) špatný kontakt v konektorech nebo porucha kabelu. Tato porucha způsobí bitové chyby a může způsobit špatné ukládání dat (což pole obvykle nezjistí!) nebo i detekované chyby parity (a degradaci pole). Disk není poškozen – po opravě špatného kontaktu jej lze dále používat.

4) úplné odpojení datového kabelu. Disk není poškozen – po nápravě poruchy jej lze dále používat.

5) porucha napájení disku (rozpojený napájecí konektor). Disk není poškozen – po nápravě poruchy jej lze dále používat.

Fyzicky lze simulovat v zásadě jedině odpojení napájecího či datového kabelu. Pokud nejsou k dispozici hot-swapové rámečky, jedná se v případě provedení za provozu o simulaci lehce rizikovou a pro někoho možná drastickou. Pokud provedeme odpojení disku při vypnutém napájení celého systému, je simulace z bezpečnostního pohledu košer, ale zase neotestujeme chování řadiče a jeho ovladačů při výpadku disku za provozu – a nutno podotknout, že v této situaci lze u konkrétních řadičů testováním odhalit nejeden zádrhel.

RAID není odolný vůči bitovým chybám

Bitové chyby mohou vznikat kvůli vadné kabeláži nebo obzvlášť zákeřnou poruchou disku.

Ani redundantní varianty RAIDu neprovádějí on-the-fly kontrolu parity uložených dat. Redundantní informace se vytváří pouze jako ochrana proti výpadku. Pokud řadič nedetekuje výpadek disku, rozhodně nekontroluje, zda si data v redundantní sadě navzájem odpovídají. Při běžném čtení z pole se redundantní informace vůbec nenačítá.

RAID není náhražkou za zálohování!

Nasazením redundantního RAIDu (RAID1/mirror nebo RAID5) správci rozhodně neodpadá povinnost pravidelně zálohovat data.

Různé řadiče mají různou úroveň schopností

Hardware, firmware, BIOS, ovladače a obslužný software různých řadičů mají velmi různou úroveň a nelze říci, že by existoval nějaký řadič, který by byl ve všech operačních systémech po všech stránkách bezvadný. Pokud v určitý moment konkrétní výrobce zapracuje a zazáří jako hvězda na jasném letním nebi, není vůbec jisté, že ho technický pokrok v oblasti hardwaru a překotné vylepšování (a "vylepšování") operačních systémů nezmění během pár měsíců v nebezpečnou morální trosku.

Existují solidní hardwarové řadiče, které díky špatnému obslužnému softwaru nezvládají obnovu pole na pozadí za běhu operačního systému nebo nezvládají odstínit výpadek disku od operačního systému, a existují laciné výrobky, které všechny důležité funkce víceméně potají emulují v ovladačích, a jsou v tom tak dobré, že v některých operačních systémech takřka dosahují schopností kvalitních hardwarových řešení.

Neustále vycházejí nové verze Windows i OpenSource operačních systémů – každá nová verze nese riziko hraničící s jistotou, že výrobci (a dobrovolní nadšenci) nestihnou řádně otestovat a opravit ovladače v nové verzi operačního systému.

Neustále vycházejí nové modely hardwaru – zejména v OpenSource OS každá nová revize hardwaru (a což teprve model) znamená riziko (jistotu), že staré ovladače nebudou fungovat s novým železem, přestože osoby blízké výrobci ujišťují, že "se v API nic nezměnilo"...

Z toho plyne ponaučení: nové železo je třeba před nasazením nebo hromadným nákupem otestovat pod plánovaným cílovým operačním systémem. Nebo najít někoho, kdo to udělal (udělá) za nás.

Těžko na cvičišti, lehko na bojišti

V kritických aplikacích (což je u RAIDů poměrně běžný způsob použití) je nanejvýš vhodné, aby si správce systému vyzkoušel nasimulovat výpadek a pole opět opravit. Zejména pro UNIXové operační systémy bývá ovládací software a dokumentace poměrně slabá, což se projeví typicky při havárii pole. Není nad to, mít postup obnovy pole alespoň v mlhavé paměti a v poznámkách.

Odkazy

Disková sekce Anandtechu
http://www.anandtech.com/storage/index.html

Starší recenze několika ATA RAIDů od Anandtechu
http://www.anandtech.com/storage/showdoc.html?i=1491

Procesory Intel IOP
http://www.intel.com/design/iio/

Intel IOP315 – stránka obsahuje porovnání několika IOP čipů
http://www.intel.com/design/iio/iop315.htm

Manuály k i960 a IOP

IOP321 (Xscale)	ftp://download.intel.com/design/iio/manuals/27351702.zip
IOP303 (i960 vč. XOR)	ftp://download.intel.com/design/iio/manuals/27335301.zip
i960RD (bez XOR)	ftp://download.intel.com/design/iio/manuals/27273602.pdf

Sběrnice a bridge v PC
http://www.rambus.com/rdf/rdf2002/pdf/2FeibusIntro.pdf

RAID řadiče Adaptec
http://www.adaptec.com/worldwide/product/prodtechindex.html?sess=no&language=English+US&cat=%2fTechnology%2fRAID

RAID řadiče Intel
http://www.intel.com/design/servers/buildingblocks/raid.htm?iid=ipp_home+server_raid_cont&

RAID řadiče LSI Logic (Mylex a MegaRAID)
http://www.lsilogic.com/products/stor_prod/raid/index.html

RAID řadiče ICP Vortex
http://www.icp-vortex.com/english/product/prod_e.htm

Domovská stránka firmy 3ware a marketingově laděný popis jejich StorSwitch architektury
http://www.3ware.com
http://www.fcenter.ru/articles.shtml?hdd/4243

RAID řadiče Promise S150 SX4 a SX4000
http://www.promise.com/product/product_detail_eng.asp?productId=112&familyId=2
http://www.promise.com/product/product_detail_eng.asp?productId=94&familyId=2#

Externí RAID řadiče InforTrend
http://www.infortrend.com

Linux Software RAID HOWTO
http://www.ibiblio.org/pub/Linux/docs/HOWTO/other-formats/html_single/Software-RAID-HOWTO.html

Linux Documentation Project (index HOWTO dokumentů a jiné dokumentace)
http://www.tldp.org

Příloha – program pro hledání nenulových sektorů

```
#include <stdio.h>
#include <sys/types.h>
#include <sys/stat.h>
/* #include <fcntl.h> */
#include <asm/fcntl.h>
#include <errno.h>

void usage()
{
   printf("Usage examples:\n");
   printf(" dscan /dev/sdc\n");
   printf(" dscan /dev/hda\n");
   printf(" dscan ./my_ordinary_file\n");
}

int main(int argc, const char** argv, const char** env)
{
   int fd, i;
   unsigned int sector=0, a_hundred_megs=0;
   char data[512];

   if (argc < 2)
   {
      usage();
      exit(1);
   }

   fd = open(argv[1],O_RDONLY|O_LARGEFILE);
   if (fd < 0)
   {
      perror("Error opening input");
      usage();
      exit(1);
   }

   while ( read(fd,data,512) == 512 )
   {
      for (i=0; i<128; i++)
      {
         if ( ((unsigned int*)data)[i] != 0 )
         {
            printf("Non-zero data found at sector %u\n", sector);
            break;
         }
      }

      sector++;
      a_hundred_megs++;
      if (a_hundred_megs == 200000)
      {
         fprintf(stderr, ".\n");
         a_hundred_megs = 0;
      }
   }

   perror("Seems like we're finished");
   close(fd);

   exit(0);
}
```