RAID pod Linuxem

Stručný přehled

Autor: František Ryšánek <rysanek@fccps.cz>

FCC Průmyslové Systémy s.r.o.

Obsah

Úvodem

Pod Linuxem lze RAID zařídit několika různými způsoby.

Především lze použít pravý hardwarový RAID, ať už v podobě PCI RAID adaptéru nebo v podobě externího SCSI-to-SCSI řadiče. Dále lze použít „softwarový RAID adaptér“, pokud je pod Linuxem podporován. A nakonec je k dispozici také čistě softwarové řešení, které je standardní součástí novějších jader 2.4.

Informace uvedené v tomto dokumentu se týkají jader řady 2.4.

Hardware RAID adaptéry

Solidní PCI HW RAID je pro Linux transparentní – v běžném provozu se tváří jako obyčejný SCSI řadič.

Podmínkou použití hardwarového PCI RAID adaptéru je tedy především dostupnost ovladače pro Linux. Ovladač v operačním systému zařizuje v zásadě pouze prostý blokový přístup k emulovaným diskovým zařízením – o RAIDové funkce se nestará.

Vedlejší součástí ovladače bývá „znakové“ zařízení pro správu polí. Znaková třída zařízení je ovšem spíše zástěrkou, hlavním komunikačním prostředkem tohoto pomocného zařízení jsou typicky ioctl() volání. Jejich obsah je předáván přímo autonomnímu procesoru RAID řadiče.

Toto vše se odehrává v kernelu. Správce systému se ale pohybuje v „user space“ a dává přednost jisté míře uživatelské přítulnosti. Komplementární součástkou správcovského subsystému je proto obvykle user-space utilita, která v nejjednodušším případě vytváří příkazový řádek (též CLI – Command Line Interface) a převádí textové příkazy administrátora na ioctl() volání.

Vyskytují se i složitější formy správcovského user-space softwaru – weboví klienti, dohledoví démoni poskytující SNMP funkce a rozesílání upozornění mailem apod.

V konfiguračním systému jádra („make menuconfig“) se ovladače pro PCI RAID řadiče nacházejí typicky ve větvi „SCSI Support -> SCSI low-level drivers”, tj. mezi ovladači pro obyčejné SCSI řadiče. Výjimkou je ovladač pro řadiče kompatibilní s Mylex DAC960 (tj. z novějších např. AcceleRaid 160 a 170), který se zatoulal do menu „Block devices“.

Tradiční SCSI-to-SCSI RAID řadiče bývají řídící jednotkou externího pole a k počítači/serveru se připojují jako disk – tj. prostřednictvím SCSI kabelu a „obyčejného“ SCSI adaptéru. Vůči počítači se taková externí disková jednotka tváří jako jediný disk, případně jako několik disků (jedno SCSI ID, několik čísel LUN). Potažmo operační systém nepotřebuje pro externí pole speciální ovladač.

Správa polí na externí jednotce se odehrává buď ručně přes vlastní LCD panel externí jednotky, nebo po sériové lince v terminálovém režimu – v tom případě stačí propojit konzolový port externího pole s volným RS232 portem hostitelského počítače a následně je možno použít libovolný terminálový emulátor (komunikační program) jako cu nebo minicom. V tomto případě tedy sám RAIDový řadič vytváří příkazovou řádku nebo systém menu (viz. např. řadiče firmy InforTrend).

Software RAID adaptéry

V linuxu existuje podpora pro nejznámější „hardwarově závislé“ softwarové RAIDy – modul „ataraid“. Tento modul se skládá z univerzální části kódu a dále z volitelných hardwarově závislých „spodních polovin“.

Konkrétně v jádrech 2.4 kolem 2.4.22 jsou podporovány řadiče Promise PDC202xx, HighPoint HPT a Silicon Image SiI. Pokud se hardwaru týče, jedná se v zásadě o prachobyčejné IDE/UATA řadiče, ke kterým výrobce dodává proprietární RAIDové ovladače do Windows. Obvykle je k dispozici také proprietární closed-source port těchto ovladačů do Linuxu, obvykle nepříliš kvalitní – viz např. samostatný dokument o Promise FastTrak TX2000.

Hardwarově závislá vrstva modulu ATARAID se stará především o počáteční detekci polí na diskových jednotkách. Pozor, všechny tyto řadiče mají samostatné hardwarově závislé IDE ovladače a kromě toho jsou tyto řadiče obvykle i bez zakompilované hardwarové podpory detekovány jako obecné IDE řadiče. Modul ATARAID se nestará o low-level přístup k hardwaru a běží v zásadě nad standardními IDE zařízeními.

Můžeme tedy shrnout, že v Linuxu se provozu „hardwarově závislého softwarového RAIDu“ účastní tyto vrstvy:

Vrstva	**Funkce vrstvy**
Společný kód modulu ataraid	Prezentace polí do systému v podobě blokových zařízení
HW specifický kód modulu ataraid	Detekce polí na dostupných IDE discích, některé RAIDové operace
HW specifický IDE ovladač	Standardní IDE funkce, včetně HW specifických low-level funkcí

Pokud vynecháme obě vrstvy modulu ATARAID, i nadále můžeme tyto řadiče používat jako prostý IDE řadič s hardwarově specifickým ovladačem. A naopak, teoreticky nám nic nebrání provozovat modul ATARAID včetně jeho hardwarově specifických volitelných součástek nad libovolným jiným IDE řadičem – ovšem za předpokladu, že na IDE discích nějakým způsobem vytvoříme příslušný proprietární RAID superblok.

Tato možnost je ovšem zajímavá spíše teoreticky a pouze na první pohled – modul ATARAID totiž výrazně zaostává za hardwarově nezávislým RAIDem, který je v Linuxu také k dispozici. Více o tom níže.

V nástroji „make menuconfig“ naleznete ATARAID a všechny relevantní low-level IDE ovladače ve větvi „ATA/IDE/MFM/RLL“. Nízkoúrovňové IDE ovladače jsou v záložce „PCI IDE chipset support“, modul ATARAID se skrývá úplně na konci v záložce „Support for IDE RAID Adapters“.

Velké zklamání představuje ovšem praktická (ne)použitelnost modulu ATARAID. Tento modul nezvládá obnovu pole na pozadí, ba dokonce se zdá, že funkce pro obnovu pole zcela postrádá.

Neexistuje user-space utilita pro správu či alespoň sledování stavu pole. ATARAID běží nad obecnými IDE zařízeními, nemá přímé háčky na hardware – výpadek disku způsobí vodopád I/O chyb z low-level IDE vrstvy, se kterými se ATARAID nedokáže vyrovnat, takže výpadek disku v redundantním poli znamená nefunkčnost celého pole. Pokud je degradováno pole, ze kterého systém bootuje, neodvratným důsledkem je pád systému. Z degradovaného pole ovšem dokonce nejde ani nabootovat. Teprve po obnově pole v přiloženém BIOSu nebo v nějakém jiném operačním systému je možné pole dále používat pod Linuxem.

Modul ATARAID se tedy hodí spíše pro nouzový přístup z Linuxu na pole a oddíly vytvořené v systémech Windows. V tomto smyslu najde uplatnění v dual-boot systémech a na „vyprošťovacích“ disketách či CD-ROM.

Obecné informace o softwarových IDE RAIDech najdete v příslušných kapitolách dokumentu „obecně o RAIDu“. Adaptec HostRAID není v Linuxu podporován.

Čistě softwarový RAID

Toto řešení má v porovnání s jinými typy RAIDu některé specifické vlastnosti.

Především běží nad prakticky jakýmikoli blokovými zařízeními, včetně např. „network block devices“. Typickým podkladovým zařízením je partition na fyzickém disku.

Méně imponující je, že RAIDem prezentované blokové zařízení nelze rozdělit na oddíly (existuje patch MdPart, který toto umožňuje, ale obvykle není k dispozici pro nejnovější jádra, a v modulu md se toho pravidelně poměrně dost mění).

Standardní způsob použití osvětlí jednoduchý příklad. Linux se typicky instaluje na jediný fyzický disk, který se rozdělí na tři oddíly: boot, swap a root. Řekněme, že chceme toto schéma obohatit o RAID typu mirror. Vezmeme tedy dva disky, které fdiskem rozdělíme identickým způsobem opět na zmíněné tři oddíly. Následně vytvoříme tři pole typu mirror - vždy ze dvou oddílů, které si vzájemně odpovídají na dvou použitých discích. Tato tři pole pak montujeme namísto původních tří diskových oddílů – samozřejmě poté, co na nich vytvoříme souborové systémy apod.

Z linuxového čistě softwarového RAIDu lze bootovat – v tom případě je vhodné povolit, aby si ovladač ukládal na fyzické disky informaci o vytvořených polích.

Linuxový RAID provádí obnovu pole automaticky na pozadí a zvládá také průběžnou kontrolu konzistence.

Pro RAIDové operace se využívají volné systémové zdroje, tj. volný čas procesoru a volná kapacita diskových sběrnic.

Linuxový RAID bohužel za současného stavu věcí neustojí degradaci pole za běhu. Jedná se patrně opět o problém s použitím low-level IDE (a SCSI) zařízení, která na výpadek disku reagují totálním kolapsem. RAIDová vrstva nepoužívá přímý přístup k hardwaru, který by umožnil vypořádat se s výpadkem disku po svém, a neumí využít ani hardwarovou podporu pro SCSI/ATA/SATA hot-swap. Linuxové jádro 2.4 jednoduše nepodporuje runtime re-detekci diskových zařízení a oddílů a také nepodporuje hot-swap backplane procesory (SAF-TE/SES na SCSI SCA, ani SAF-TE přes I2C u Serial ATA).

Degradace pole za běhu tedy znamená zatuhnutí pole – v případě, že se z pole bootuje, dojde k pádu celého systému. Po resetu se ovšem pole rozběhne i v degradovaném režimu a lze z něj nabootovat. Pokud mezitím připojíme náhradní disk, lze za běhu spustit obnovu degradovaných polí (příkazem raidhotadd). Dojde tedy pouze k minimálnímu výpadku systému (pokud máme možnost jej resetovat).

Pokud přerušíme obnovu pole restartem systému, po dalším bootu se obnova pole rozběhne znova od začátku.

Utility pro správu a dohled jsou dostatečné.

V nástroji „make menuconfig“ najdete podporu pro čistě softwarový RAID ve větvi „Multi-device support (RAID and LVM)“.

Podrobnější informace a návody ke konfiguraci získáte ve výborném „Linux Software RAID HOWTO“.

Shrnutí

Všechny výše popsané typy RAIDu za normálních podmínek v Linuxu bez problémů fungují. Velký rozdíl je ovšem v chování jednotlivých řešení *v kritických situacích* – tj. při degradaci pole. Je třeba si uvědomit, že zejména kvůli kritickým situacím se pole staví – proto řešení, která v kritických situacích selhávají, nemají valný smysl. Holé uchování dat při havárii disku, doprovázené ovšem nefunkčností operačního systému, je poměrně slabý výsledek. Vedle celkové stability systému v kritických situacích je u typických použití UNIXu důležitá především možnost ošetřit pole přímo z běžícího operačního systému.

HW RAID ovladače, ATARAID a čistě softwarový RAID jsou tři různé oblasti jádra, které jsou navzájem naprosto nezávislé. Pokud konfigurujete HW RAID, není třeba zapínat podporu pro ATARAID ani podporu pro čistě softwarový RAID – totéž platí pro zbylé dvě možnosti.

„Softwarových řadičů“ firmy Promise se plně týká kapitola o „software RAID adaptérech“.

Firma Promise ovšem vyrábí také vícekanálové řadiče, které sice postrádají autonomní procesor, ale obsahují čip pro akceleraci RAID operací (RAID5 XOR, mirroring) a onboard cache – tyto řadiče nejsou obsouženy ATARAIDem a spadají tedy spíše do kapitoly o Hardware RAIDu. Firma Promise nedávno uvolnila zdrojový kód ovladačů pro svou řadu Serial-ATA RAID řadičů, které byly do té doby k dispozici výhradně v closed-source podobě. Těžko říci, kterých řadičů se toto týká. Closed-source ovladače firmy Promise donedávna za mnoho nestály (a management software do Linuxu neexistuje) – je ale možné, že zveřejněním zdrojového kódu ovladačů se situace změní k lepšímu. Zdá se, že na vyčištěném portu pro jádro 2.6 se pracuje. Software firmy Promise pro Windows patří rozhodně k tomu lepšímu, co lze na trhu potkat.

Odkazy

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

Autorova oprava chyby v ovladači pdcraid.c (součást modulu ataraid).
http://sweb.cz/Frantisek.Rysanek/pdcraid/pdcraid.html