Routing přerušení a kolize prostředků na platformě x86 aneb sedm generací PC AT

Autor: František Ryšánek <rysanek@fccps.cz>

FCC Průmyslové systémy s.r.o.

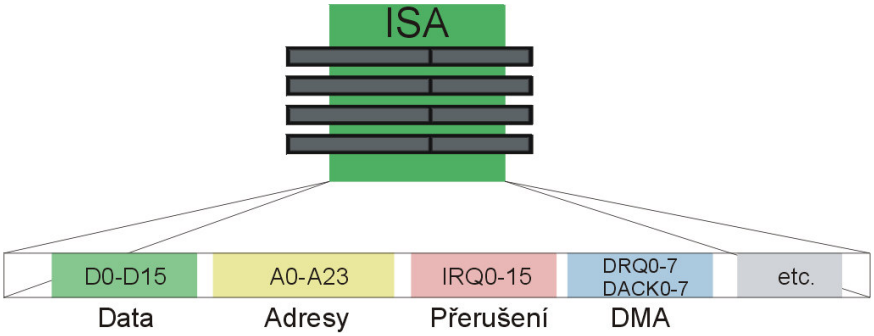
Obsah

Základní charakteristiky sběrnic

Na architektuře PC se vyskytoval a vyskytuje větší počet druhů periferních sběrnic. Na úvod tohoto pojednání o přerušeních zmíníme pouze dva: ISA a PCI. Hrají ve vývoji architektury asi nejvýznamnější roli.

ISA

ISA je šestnáctibitová sběrnice. Má oddělené adresní a datové vodiče (A0-A23, D0-D15). Zápis nebo čtení probíhá tak, že procesor současně nastaví adresu a datové slovo a na další hraně hodinového signálu transakce proběhne. K zápisu nebo přečtení jednoho slova tedy stačí jeden takt sběrnice.

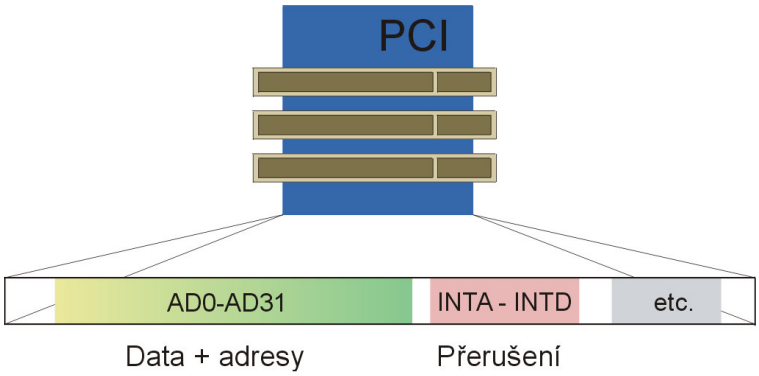


Transakce se pochopitelně účastní několik dalších řídících signálů (vodičů). Sběrnice také umí DMA bez účasti procesoru - k tomu účelu obsahuje osm DRQ signálů (a odpovídajících osmkrát DACK).

Pro signalizaci přerušení je k dispozici šestnáct IRQ signálů (vodičů), známých jako IRQ0 až IRQ15.

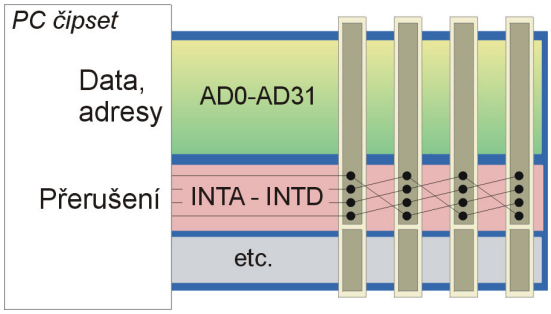
PCI

PCI je dvaatřicetibitová sběrnice. Tytéž vodiče se používají pro adresy i pro data (AD0-AD31). Zápisové a čtecí transakce tedy využívají jednoduchý multiplex – a vyžadují více než jeden sběrnicový takt. Při blokových přenosech se ovšem adresa přenáší pouze na začátku transakce a po ní následuje větší sekvenční blok dat – takže vliv multiplexu na latenci transakcí a efektivní kapacitu sběrnice není velký.



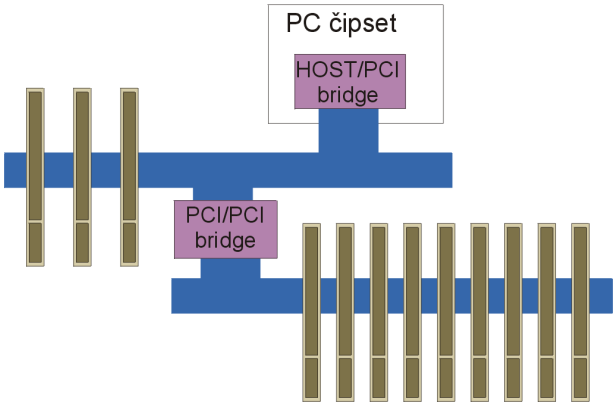
Inicializace zařízení a provozních transakcí se opět účastní několik dalších řídících signálů (vodičů). Mechanismus přímého přístupu do paměti na platformě PC (“PCI DMA”) je realizován pomocí bus masteringu.

Pro signalizaci přerušení jsou k dispozici čtyři signály (vodiče) INTA až INTD. Tyto signály jsou mezi rozšiřujícími sloty “zadrátovány” zvláštním způsobem: na každém dalším slotu se o jeden signál “rotují”.

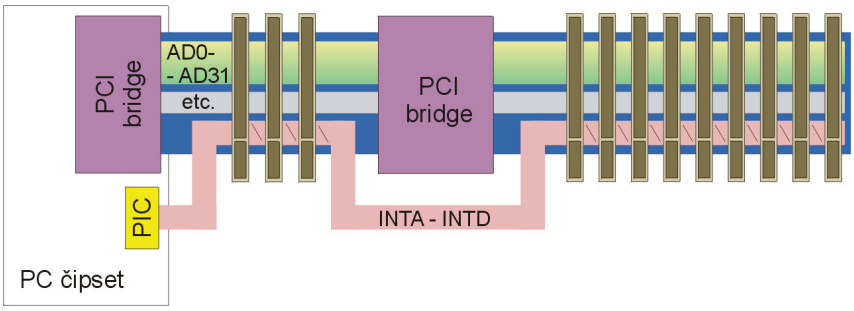


Platí konvence, že deska ve slotu obsazuje jako první vždy signál INTA (a teprve v případě potřeby ostatní) – takže je přímo hardwarově zařízena samovolná minimalizace sdílení IRQ. V případě, že jsou na jedné sběrnici více než čtyři sloty, sdílení se pochopitelně nevyhneme[1].

Sběrnice PCI má kvůli vyšší rychlosti poměrně striktně vymezené elektrické parametry. Jde především o impedance vedení, délky vodičů, parazitní kapacity, vstupní impedance připojených zařízení. V důsledku toho lze na jeden elektrický segment připojit pouze poměrně malý počet zařízení (slotů). Integrované PCI bridge PC čipsetů obvykle zvládají tři až pět externích PCI slotů. Elektrické segmenty lze ovšem řetězit pomocí PCI bridgů (PCI-to-PCI) - tak lze dosáhnout většího počtu připojených zařízení. PCI bridge v podobě samostatných součástek obslouží např. až devět PCI slotů. Zůstává pouze potenciální problém s délkou a sdílením IRQ signálů, protože o tyto se PCI bridge nestará.



PCI IRQ jsou vlastně na zbytku sběrnicových protokolů PCI do značné míry nezávislá – nejsou obsluhována PCI bridgi, na straně PC čipsetu jsou přivedena přímo na vstupy programovatelného řadiče přerušení. Hardwarově specifická tabulka mapování PCI IRQ na jednotlivé sloty je součástí PC BIOSu a s činností sběrnice PCI vlastně přímo nesouvisí. PCI bridge se stará o své transakce pro přenos dat, se kterými IRQ vodiče mají pramálo společného – pokud pomineme nápadnou časovou korelaci jejich aktivity, která je ovšem způsobena nepřímo, oklikou přes procesor nebo přes periferní zařízení. Pokud je sběrnice PCI "prodloužena" bridgem, signály INTA-INTD musejí PCI bridge obcházet - jakoby ani nebyly integrální součástí PCI sběrnice, ale spíše jakýmsi appendixem prosazeným platformou x86 PC. Přitom oboje pochází z dílny Intelu.



Sběrnice PCI není závislá na platformě x86 a jejích adresačních pravidlech – PCI má svůj vlastní adresní prostor (32bitový nebo 64bitový), který si hostitelský operační systém prostřednictvím host bridge mapuje / překládá na svou nativní adresaci. Připomeňme, že PC procesory 386 a vyšší samy nativně používají pro virtualizaci a ochranu paměti několik paralelních adresních prostorů (fyzická adresace, prostor jádra, prostory jednotlivých běžících user-space procesů).

[1] Pakliže ovšem mezi sloty není rozdistribuováno více čtveřic IRQ signálů – zdá se, že i takovou konfiguraci některé PC čipsety podporují.

Trendy vývoje sběrnic

PCI je jednou z prvních sběrnic, u kterých je znát obecný trend směrem k “paketizaci” transakcí.

Nové periferní sběrnice mají typicky omezený počet řídících i datových/adresových signálů a transakce mají podobu zpráv nebo rámců, které obsahují adresu, data a další řídící informace. Kromě “datových” transakcí se vyskytují i čistě režijní transakce, které přebírají význam původních samostatných signálů (vodičů) – takovou režijní transakcí může být např. přerušení.

Systémoví architekti se snaží s ohledem na výrobní složitost (a tedy cenu) snižovat či alespoň omezit počet vývodů procesorů a zejména čipsetů. V čipsetech pro PC je integrováno stále více periferií, což samo o sobě rychle zvyšuje počet vývodů. Moderní polovodičová technologie naopak umožňuje výrobu budičů pro krátké sběrnice taktované na stovkách MHz až jednotkách GHz. Proto se výrazně omezují počty vodičů zejména v pomalejších periferních sběrnicích – čím pomalejší sběrnice, tím menší počet signálů je zapotřebí. Viz např. proprietární sběrnice Intel HubLink a LPC nebo AMD HyperTransport, nebo “otevřené” JEDEC Flash či Intel PCI-Express. Tento trend často vede až k převodu sběrnic na klasický sériový provoz – viz např. Serial ATA a jeho příbuzný SAS (Serial-Attached SCSI) nebo externí sběrnice jako USB 1.1/2.0, FireWire nebo FiberChannel. Některé novinky na tomto poli zavádějí prvky známé dosud spíše z datových sítí a telekomunikací.

Tendence k paketizaci transakcí se do jisté míry projevuje i u paměťových sběrnic a FSB – zde jsou ovšem požadavky na kapacitu natolik vysoké, že se příliš neuplatňuje efekt “zužování” sběrnic. Dnešní FSB a paměťové cesty mají šířku typicky 64 nebo 128 bitů, což odpovídá “dimenzi” procesorové platformy nebo jejímu dvojnásobku. A konkrétně široká paměťová sběrnice DDR DRAM zřejmě zvítězila nad “zúženou” sběrnicí Rambus DRAM – roli hrála vyšší cena RDRAM způsobená licenční politikou a také vyšší latence transakcí. Zatím se tedy příliš nedaří omezovat trvalý růst počtu vývodů procesorů s každou novou generací.

Velice zajímavě vypadá příští generace sběrnice PCI. Je založená na sériovém přenosu dat point-to-point spoji a switchování provozu. Topologie tedy není nepodobná např. 100BaseT Ethernetu nebo ATM – ovšem rychlosti jsou řádově vyšší a počty portů na switchujících elementech řádově nižší. V systému je díky tomu na daný počet koncových periferních zařízení relativně velký počet switchů. Odhlédneme-li však od nové topologie, je do značné míry dodržena sběrnicová sémantika PCI. Zajímavou novinkou je možnost konstruovat systém s několika host bridgi – což vypadá jako obdoba SCSI sběrnice sdílené v redundantní konfigurací mezi více počítači.

Druhy sběrnicových prostředků

Tato kapitola probere sběrnicové prostředky známé na platformě PC – tak jak je ukazují operační systémy. A také jednu kategorii prostředků, které v operačním systému vidět *nejsou*.

I/O porty

Architektura x86 striktně rozlišuje přístup na vstupně/výstupní porty od přístupu do paměti. Porty mají svou vlastní adresaci, která není zaměnitelná s paměťovým adresovým prostorem – přestože na sběrnici ISA oba druhy přístupu používají tytéž datové a adresní vodiče. Rozlišení přístupu do paměti od přístupu na porty se na sběrnici ISA děje několika jednoúčelovými signály, na sběrnici PCI pomocí "PCI příkazu" (opět čtyři samostatné vodiče/signály).

Konkrétní zařízení na sběrnici může používat jeden či více I/O portů – tzn. reaguje na zápis a čtení příslušných portových adres. Jedná se obvykle o blok fixního počtu portových adres (např. 16) od určité bázové adresy, kterou lze konfigurovat.

Pro přístup na porty se používají **instrukce `in` a `out`** – cílem resp. zdrojem mohou být pouze vybrané registry procesoru (nikoli paměťová pozice). Jedna instrukce přesune vždy jedno slovo (byte, word, dword).

IO porty se používají z principu typicky pro pomalejší vstupně/výstupní periferie (PIO režim). Pomalejší proto, že tento styl vstupně výstupních operací probíhá "s osobní účastí procesoru" a tedy spotřebovává hodně procesorového času (výkonu). To platí zejména v dnešní době, kdy jsou procesory výrazně rychlejší než periferní sběrnice (takže se na procesoru vkládá spousta čekacích cyklů).

Paměťový rozsah

Jak na sběrnici ISA, tak na sběrnici PCI existují zařízení typu "paměť" – tj. zařízení, která reagují na zápis či čtení určitého rozmezí paměťových adres. Velikost paměti je opět typicky fixní, konfiguruje se základní (nultá) adresa.

Pro přístup do paměti se používá **instrukce `mov`** (a některé další). Kopírovat data je možné pouze z paměti do registrů procesoru a naopak (a také mezi registry procesoru navzájem) - ačkoli některé assemblery podporují i makro mov z paměti do paměti. Jedna instrukce mov přesune vždy jedno slovo (byte, word, dword, qword) – existují řetězcové instrukce, které slouží pro efektivní kopírování větších bloků dat.

Mezi paměťově mapovaná zařízení patří typicky ROM paměti rozšiřujících karet s vlastním BIOSem (programovým kódem), případně "vstupně-výstupní zařízení mapovaná do paměti" (MMIO).

DMA "kanály"

Zkratka DMA znamená "přímý přístup do paměti". Pomocí tohoto mechanismu může periferní zařízení číst či zapisovat z/do operační paměti počítače bez účasti procesoru.

Jedná se o osm signálů/vodičů ve sběrnici ISA, které jsou obsluhovány DMA řadičem. Konkrétní kanál může být využit nanejvýš jedním zařízením – lze nakonfigurovat, které zařízení bude používat který kanál. Před použitím DMA konkrétním zařízením je třeba řadič naprogramovat (je vidět jako

několik portů – tj. programuje se pomocí instrukcí in/out). Nakonec se přenos spustí – obvykle asynchronně, tj. bez přímého pokynu procesoru, na pokyn periferního zařízení (zatahá za DMA signál). Ukončení přenosu je typicky signalizováno procesoru signálem IRQ.

Jedná se o prostředek specifický pro sběrnici ISA (a její odvozeniny). Je používán typicky u zařízení, která přenášejí velké objemy dat a nemají nadměrně zatěžovat procesor (SCSI řadiče[2]), případně u zařízení, která potřebují zajistit nepřerušovaný isochronní tok dat nezávisle na zátěži procesoru (typicky zvukové karty).

Tzv. "PCI DMA" vypadá z hlediska operačního systému obdobně, je však prováděno pomocí bus-master transakcí sběrnice PCI.

IRQ

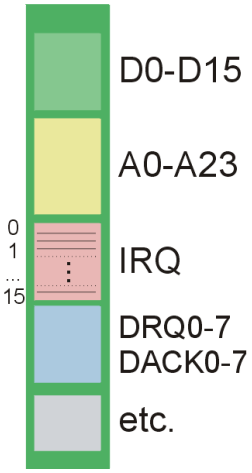
Přerušení je obecně vstupní signál procesoru, který používají periferie v situaci, kdy "chtějí mluvit s procesorem". Jinak řečeno, jde o základní hardwarový prostředek pro asynchronní signalizaci vnějších událostí. Z pohledu softwaru se přerušení jeví jako svévolné či implicitní volání funkce, která byla předem registrována jako *obsluha* konkrétního přerušení (anglicky ISR – Interrupt Service Routine).

Zkratka IRQ znamená Interrupt Request – žádost o přerušení. Jedná se o signál (vodič) ve sběrnici – vyskytuje se jak na ISA, tak na PCI.

Ve standardním PC AT je jich šestnáct (IRQ0 až IRQ15) – tj. šestnáct jednotlivých vodičů ve sběrnici ISA. Ne všechny jsou vyvedeny do ISA *slotů* - některé jsou využívány zabudovanými systémovými zařízeními (klávesnice, časovač, matematický koprocesor).

Sběrnice PCI obsahuje čtyři "IRQ" signály, značené obvykle INTA až INTD. Tyto jsou na platformě IBM PC konfigurovatelně mapovány na původní ISA IRQ.

Na novějším PC hardwaru se v oblasti přerušení objevilo mnoho novinek – nepozorovaně a bez humbuku. Více o tom v dalších kapitolách.



Novinky na PCI - GNT a spol.

Na sběrnici PCI existuje několik režijních signálů, které jsou individuální pro každý slot, resp. PCI zařízení – nejdůležitější jsou asi tyto:

– **IDSEL** – výstup PCI bridge – používá se při konfiguraci zařízení na sběrnici, k jednoznačné identifikaci slotu před přidělením prostředků. Signály IDSEL dnes existují jako samostatné piny vlastně jen v konektorech PCI slotů. PCI bridge typicky budí piny IDSEL0-15 adresně/datovými signály z rozmezí AD16-31 – nebo-li, IDSEL PIN je přímo u svého PCI slotu připojen na příslušný AD signál. Tento multiplex je součástí standardu PCI.

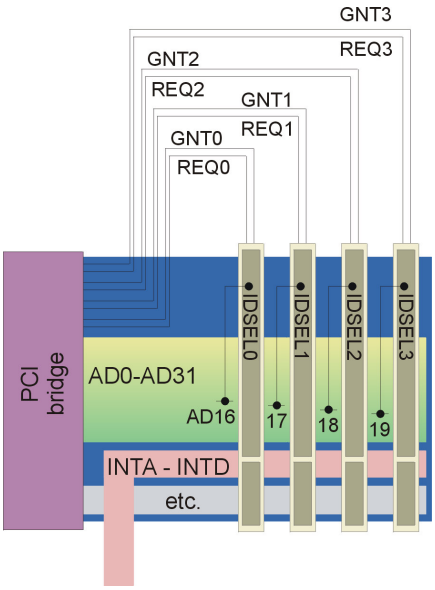
[2] původní standard IDE s využitím DMA nepočítá – zřejmě z cenových důvodů.

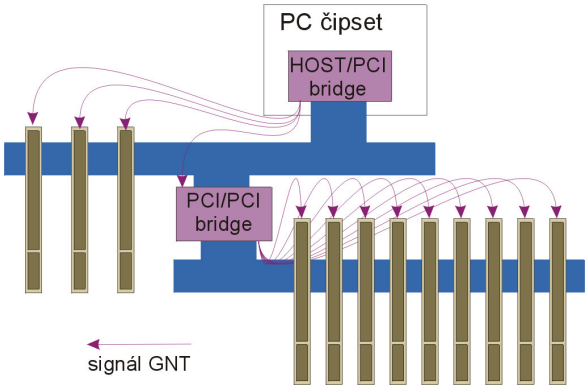
– **REQ** – vstup PCI bridge – žádost o přidělení sběrnice (bus request). Použije ho zařízení, které chce provést bus-master transakci.

– **GNT** – výstup PCI bridge – potvrzení o přidělení sběrnice (bus grant). Zařízení, které žádalo o přidělení sběrnice, může nyní provést bus-master transakci.

Bus-master transakce slouží podobnému účelu jako ISA DMA – tj. k přenosům větších objemů dat po sběrnici bez přímé účasti procesoru. Lze je používat i mezi zařízeními navzájem – nejen z periferního zařízení do operační paměti hostitelského systému. Většina dnešních periferních PCI karet na bus-masteringu vysloveně závisí – jako příklady lze uvést grafické karty, síťové karty, zvukové karty nebo diskové řadiče všeho druhu.

PCI-to-PCI bridge má dva porty – "master" port a "slave" port. Na "slave" portu, tj. na nadřízené sběrnici, se chová do značné míry jako obyčejné zařízení – obsazuje jedinou instanci signálů IDSEL, REQ a GNT. Na "master" portu, tj. na podřízené sběrnici, naopak poskytuje N instancí signálů IDSEL, REQ a GNT svým "slave" zařízením. Na obrázku jsou zakresleny signály GNT – signály IDSEL jsou vedeny analogicky, signály REQ analogicky opačným směrem.





Signál IDSEL je nezbytně nutný ke zprovoznění PCI slotu/zařízení – a díky multiplexu se signály AD o signály IDSEL není nouze. Tento signál tedy nechybí u žádného slotu či onboard zařízení. Lze ovšem uvažovat, že k některému slotu nepovedou signály REQ a GNT. Takový slot pak nebude schopen bus-masteringu.

Tato skupina signálů bohužel není zcela bezproblémová. **Problémy s bus-masteringem patří mezi nejzapeklitější, protože nejsou přímo viditelné v operačním systému** – více k tomu níže.

Způsoby konfigurace sběrnicových prostředků

ISA

Využití sběrnicových prostředků (jednotlivých IRQ, portů apod.) různými kartami v ISA slotech a zařízeními integrovanými na základní desce se konfiguruje přímo na těchto zařízeních a způsobem specifickým pro konkrétní zařízení: pomocí jumperů, utilitou která se spouští z DOSu, pomocí jednotného ISA PnP, volbami v BIOS SETUPu a snad i jinak. Ne všechna zařízení umí využívat např. kterýkoli IRQ signál či volně volitelný port – často "umí" jen několik málo prostředků od každého typu. Takže už v původním PC AT může být konfigurace IRQ lehčím hlavolamem.

PCI

O konfiguraci prostředků na sběrnici PCI se stará mechanismus PnP – lze ho nanejvýš do jisté míry usměrnit. Na sběrnici PCI se prostředky konfigurují až na výjimky jednotně, nezávisle na konkrétním typu, výrobci a modelu PCI zařízení, ať už jde o rozšiřující PCI kartu či zařízení integrované on-board nebo on-chip. Jednotnou konfiguraci prostředků provádí standardizovaný PCI PnP software, který je součástí BIOSů a operačních systémů, s vydatnou podporou standardního PCI hardwaru (bridgů a zařízení). Specifikace sběrnice PCI obsahuje povinné konfigurační registry PCI zařízení a povinně obsluhované signály, jakož i postup inicializace sběrnice a zařízení PnP softwarem – výsledkem je, že všechna zařízení mají přiděleny nekonfliktní zdroje ještě předtím, než se začnou vůbec zavádět ovladače.

Speciality PCI

Na rozdíl od čtyř "tradičních" sběrnicových prostředků **signály REQ/GNT nepatří mezi zdroje spravované pomocí PnP či vůbec nějak softwarově viditelné, a proto nelze z operačního systému zjistit jejich přiřazení ani je jakkoli konfigurovat.** Na "kancelářských" PC motherboardech neexistuje ani hardwarová možnost změny konfigurace. Na některých průmyslových modulárních platformách založených na IBM PC lze přiřazení signálů GNT/REQ a IDSEL nastavit pomocí jumperů.

Sdílení prostředků několika zařízeními a hardwarové konflikty

Obecná pravidla sdílení prostředků

Obecně platí, že prostředky výše uvedených typů by se sdílet neměly, protože to vede ke konfliktům, tj. k nefunkčnosti jednotlivých postižených zařízení, nebo i celého počítače.

Z tohoto pravidla existují jisté výjimky.

Lze si představit, že několik zařízení na sběrnici ISA sdílí nějaké rozmezí portů nebo blok paměti v režimu pro zápis. Při čtení by nastaly komplikace (budiče do zkratu = konflikt).

Sdílení DMA kanálu je snad teoreticky za určitých okolností myslitelné, prakticky se však nepoužívá – zařízení, která používají ISA DMA, se v PC ostatně těžko sejdou v dostatečně hojném počtu, aby byla o DMA kanály nouze. A pokud by se i sešla, sdílení DMA kanálu by patrně v takové situaci nepředstavovalo schůdnou variantu.

Pokud se týče IRQ, tradičně na platformě PC platívalo, že konkrétní IRQ signál smí být využíván nanejvýš jedním zařízením. V MS-DOSu zařizovaly obsluhu konkrétních přerušení ovladače hardwaru v podobě rezidentních programů, případně přímo koncové aplikace. Obsluha přerušení se instalovala přímo přepsáním příslušného vektoru v tabulce přerušení – v zásadě se nepočítalo s tím, že by IRQ signál mohl být vyvolán jedním z několika zařízení a tedy se také nepočítalo s tím, že by se volalo postupně několik obslužných rutin, než by jedna "zabrala". Toto mohlo fungovat u víceportového zařízení jednoho výrobce, ale ne mezi různými výrobci hardwaru a příslušných ovladačů.

Modernější operační systémy, jako jsou Windows nebo open-source UNIXy, již sdílení IRQ umožňují – především na sběrnici PCI pracuje sdílení poměrně bez problémů. Funguje to tak, že obslužná rutina se již nezavěšuje natvrdo přímo do tabulky vektorů přerušení, ale je při inicializaci ovladače "registrována" do systému pomocí jednotného API. Zároveň musí odpovídat určitému volacímu "prototypu" a musí vracet návratovou hodnotu, která sdělí, zda tato obslužná rutina "zabrala" - tj. zda bylo nalezeno zařízení, které přerušení poslalo. Na jedno IRQ lze registrovat několik obslužných rutin – do tabulky vektorů přerušení je zavěšena systémová "superobsluha", která při každém přerušení spouští sekvenčně všechny registrované obsluhy, než jedna "zabere". Řadič přerušení zajistí, že pokud je konkrétní přerušení vyvoláno naráz několika zařízeními, bude "superobsluha" volána opakovaně. Tak to alespoň funguje pod Windows – ve volně šiřitelných UNIXech to bude podobné.

Kdy sdílení funguje - příklady hardwaru

Multiportové sériové karty

Klasickým případem sdílení přerušení jsou multiportové sériové karty, například od výrobce Moxa. Karta obsahuje několik sériových UARTů, např. klasických 16550. Každý UART je na jiném I/O portu, sdílejí ovšem společné přerušení. Na kartě je dále společný "stavový" registr (další I/O port), ze kterého si obsluha IRQ přečte, které porty mají k dispozici data.

Dialogic BLT ISA

Patrně na samé hranici možností pracuje se sběrnicí ISA firma Dialogic (nyní Intel).

Jedna z několika řad ISA karet od firmy Dialogic používá tzv. technologii BLT (Board Locator Technology). Několik karet sdílí 32kB blok paměti v režimu čtení+zápis a také jedno konkrétní IRQ. Sdílejí jedinou instanci těchto dvou prostředků. Patrně se používá několik paměťových míst ve zmíněném bloku pro sdílený zápis jako adresační "registr", který pomáhá multiplexovat čtecí operace mezi několik karet. Obsluha přerušení je rovněž připravena obsloužit několik karet tohoto typu.

Každá BLT karta má šestnáctipolohový "ciferníkový" přepínač, kterým se nastaví její unikátní pořadové číslo v systému (pořadí na TDM sběrnici SCbus, kterou jsou karty vzájemně propojeny prostřednictvím plochého kabelu). Tím hardwarové nastavení končí. Karty BLT nepoužívají standardní ISA PnP, ani jumpery, ani utilitu pro konfiguraci IRQ a paměti. Oba tyto zdroje jsou konfigurovány za běhu ovladačem, což nastoluje zajímavý problém typu slepice/vejce: jak může ovladač nastavit IRQ a rozsah paměti dřív, než může tyto zdroje použít, aby se s kartou domluvil? Sama karta ve sběrnici ISA ze své pozice nemá šanci zjistit, které zdroje jsou v hostitelském systému volné a deterministicky si nějaké vybrat. Je to záhada. Nutno ovšem podotknout, že BLT funguje v naprosté většině případů bez problémů. Autor se domnívá, že ovladač používá pro počáteční komunikaci nějaký nepublikovaný neobsazený I/O port. Ovladač nastavuje IRQ a rozsah adres mimo rámec systémového PnP – pouze si ověří, zda jsou dané zdroje volné. Dokonce se doporučuje zablokovat použité IRQ v BIOSu pro "legacy ISA" – tím se toto IRQ spolehlivě znepřístupní PnP enginu jak v BIOSu, tak v systému Windows.

Obecné sdílení IRQ na PCI

V moderních operačních systémech je běžné, že několik PCI zařízení sdílí jedno přerušení. Na moderním PC hardwaru, nabitém periferiemi všeho druhu, to vlastně ani jinak nejde (resp. jde, pokud je k dispozici APIC – více o tom níže).

Vybrané typy hardwarových zádrhelů

Dědictví sběrnice ISA, programátorské chyby, nepořádné implementace PnP, špatná dokumentace hardwaru a další neduhy způsobují, že správa systémových prostředků na platformě PC není vždy natolik bezproblémová, jak by měla být.

Na sběrnici PCI vlastně ani neexistuje možnost nějaké prostředky nastavit "natvrdo" – natolik je sběrnice PCI provázaná s PnP softwarem. Na PCI hardwaru nelze nastavit, které prostředky má zabírat, a nelze říci softwaru, kde má hardware hledat. Pokud je chyba v softwaru účastnícím se PnP konfigurace konkrétního zařízení, holýma rukama se s tím dá těžko něco udělat.

Historicky také existovalo a existuje dost železa pro sběrnici ISA, které nepodporuje ISA PnP. V BIOSu lze konfigurovat prostředky využívané on-chip periferiemi (typicky sériové a paralelní porty a dva IDE kanály). Pro přídavná ISA zařízení je v BIOSu typicky mnohem méně.možností – obvykle lze pro "legacy ISA" vyhradit konkrétní IRQ, ale tím možnosti BIOSu končí. Pod Windows lze ovladačům určit pro ne-PnP zařízení také ostatní prostředky.

Pomineme banální konflikty klasických prostředků u ne-PnP zařízení a neřešitelné programátorské chyby v PnP – k těm není mnoho co dodat. Uvedeme si zde však některé zajímavé konkrétní problémy, kterých se lze vyvarovat.

Vypnuté onboard zařízení dál blokuje IRQ

Zejména v dobách 486 se stávalo, že onboard zařízení blokovalo IRQ signál i v případě, kdy byl tento prostředek u příslušného zařízení vypnut v BIOSu. Zařízení pouze přestalo signál aktivovat – budič tohoto signálu jej ovšem držel v neaktivním stavu, takže nebylo možné IRQ použít pro jiné zařízení. (Správně by měl budič při vypnutí IRQ přejít do stavu vysoké impedance, aby signál neblokoval.) Pokud máte podezření, že s "uvolněným" IRQ není něco v pořádku, zkuste použít jiné IRQ.

Priority IRQ

Některá zařízení mají vyšší nároky na odezvu systému na přerušení. Jedná se typicky o zařízení zpracovávající rychlé isochronní toky dat, např. video, zvuk nebo telefonní hovory. Podobně vysoké nároky můžou mít průmyslové periferie pro sběr dat nebo vyhodnocování událostí. Tato zařízení by měla být konfigurována tak, aby používala pokud možno přerušení s co nejvyšší prioritou. Při běhu na nízké prioritě se může dostavit trhané přehrávání videa nebo zkreslení či slyšitelný praskot ve zvykovém výstupu.

Díky kaskádovému řazení dvou integrovaných obvodů Intel 8259, které dohromady tvoří řadič přerušení (PIC) standardního PC AT, jsou standardní priority přerušení "promíchané" – seřazeno od nejvyšší priority po nejnižší vypadá pořadí IRQ takto: 0, 1, 2, 8, 9, 10, 11, 12, 13, 14, 15, 3, 4, 5, 6, 7.

Z toho někteří autoři vyvozují doporučení, že zařízení vyžadující rychlou odezvu by měla být umístěna na přerušení 9, 10, 11 apod. (přerušení 0,1,2 a 8 jsou využívána systémovými zařízeními).

Pořadí hardwarových priorit se ovšem dá softwarově nastavit a těžko říci, jak s prioritami zacházejí různé verze Windows. Kromě toho v moderních operačních systémech, které podporují multitasking a multithreading i v jádře, probíhá zpracování přerušení o něco složitějším způsobem: různé obsluhy běží v preemptivním multitaskingu s různou *softwarovou* prioritou, obsluha IRQ v rámci ovladače bývá rozdělena na několik částí, z nichž některé běží ve vlastním režimu přerušení, jiné jako odložená "mini-úloha", či jako trvale běžící "jaderné vlákno" apod. Situací také bezesporu zamíchá univerzálně programovatelný APIC (viz níže).

Z informace o standardních hardwarových prioritách tedy nelze vyvozovat obecně platné závěry.

Nesprávný HAL

Operační systém Windows od verze 2000 používá pro práci s klíčovým hardwarem tzv. "hardwarovou abstrakční vrstvu" – zkratka HAL pochází z anglického Hardware Abstraction Layer. To zní tajemně a složitě, možná dokonce pokročile - je ale jednodušší a mnohem přiléhavější představit si pod pojmem HAL konkrétní ikonku/objekt ve "správci zařízení".

Jedná se o jediný objekt ve složce "počítač", který většině správců patrně přijde komický a zbytečný. Jmenuje se pokaždé trochu jinak, protože existuje několik základních typů: "standardní PC", "jednoprocesorový osobní počítač s rozhraním ACPI, "multiprocesorový osobní počítač s rozhraním ACPI" apod.

Kámen úrazu je v tom, že na rozdíl od všeho ostatního hardwaru ve správci zařízení, tento objekt nelze odinstalovat.

Pokud do systému Windows přidáváme rozšiřující karty a jiné periferie a zase je odebíráme, není problém průběžně přidávat a odebírat ovladače. Dokonce je možné přesypat hotovou instalaci na jiný harddisk na jiném diskovém řadiči (systém se bude bránit, protože používá persistentní přiřazení

písmenek diskům - ale lze ho přesvědčit ručním zásahem do boot.ini a do přiřazení písmenek ve větvi MountedDevices v registrech). Takže jde radikálně změnit i bootovací disk.

Jedna věc ale nejde změnit: nelze změnit HAL. Obecně proto nejde přestěhovat Windows na jiný *čipset*. Což by se někdy potenciálně velmi hodilo při upgradu hardwaru.

Pokud si Windows při instalaci zvolily HAL typu "Standardní PC", není toto pravidlo až tak striktní. Pokud na novém hardwaru Windows najdou disky podle původních hardwarových identifikátorů na svém místě, pak přinejmenším nastartují. Takže funguje přinejmenším přestěhování windows na novější čipset téhož výrobce.

Funguje to ovšem jen do jisté míry. Pokročilejší schopnosti nového hardwaru mohou zůstat v lepším případě nevyužity. Například pokud nový čipset a BIOS podporuje ACPI a obsahuje APIC (viz níže), neexistuje způsob, jak Windows přinutit, aby toho začaly využívat. A co je horší, může se stát, že starý HAL nebude nový hardware úplně správně inicializovat, takže mohou nastat neřešitelné konflikty sběrnicových prostředků nebo obecně nefunkčnost některých hardwarových zařízení.

Pokud si Windows při instalaci zvolily ACPI HAL, čipset již naprosto nelze změnit. ACPI HAL se při instalaci "ušije na míru" konkrétnímu čipsetu a při přenosu na jiný ACPI-kompatibilní čipset nelze očekávat úspěšný start systému.

Tato kapitola vychází z praktických zkušeností s pokusem o migraci Windows 2000 z čipsetu Intel 440BX (PIII, bez ACPI) na čipset Intel 845 (P4, podpora ACPI) – a z následné debaty v newsech (autor ke své smůle nevlastní žádný MS certifikát). Windows na novém hardwaru nastartovaly a tvářily se spokojeně, ale některé druhy zásuvných karet nefungovaly správně, podle všeho kvůli neodhaleným hardwarovým konfliktům.

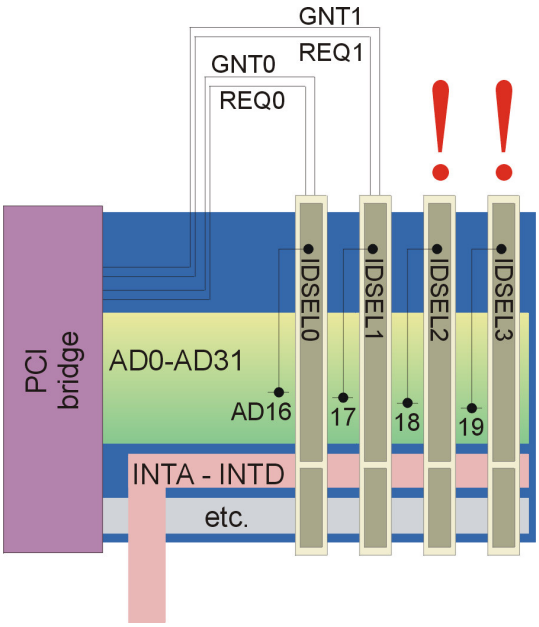
Poznámka na okraj: pod Linuxem tyto problémy s HALem nejsou. Linux si základní vlastnosti systému detekuje a konfiguruje při každém startu. Použije se vše, pro co je podpora v právě startujícím jádře. Při větších rozdílech mezi starým a novým hardwarem není problém startovat alternativní jádra na tomtéž kořenovém svazku. Případně je možné vybranému jádru upravit boot/root device – ať už z původního systému běžícího na starém hardwaru, nebo ze záchranného CD na novém hardwaru. Upgrade hardwaru pod nainstalovaným Linuxem tedy principiálně není problém.

Nefunkční bus-mastering v některých slotech

Sloty neschopné busmasteringu se bohužel na PC motherboardech běžně vyskytovaly a patrně dosud vyskytují. Důvodem jsou obvykle chybějící signály REQ/GNT. Zejména první motherboardy třídy 486 se sběrnicí PCI obvykle umožňovaly bus mastering pouze v jednom PCI slotu (typicky č.1).

Na novějších motherboardech (Pentium a výš) se situace postupně zlepšovala, na dnešních deskách s větším počtem PCI slotů (a zcela bez ISA slotů) již nejde o běžný problém.

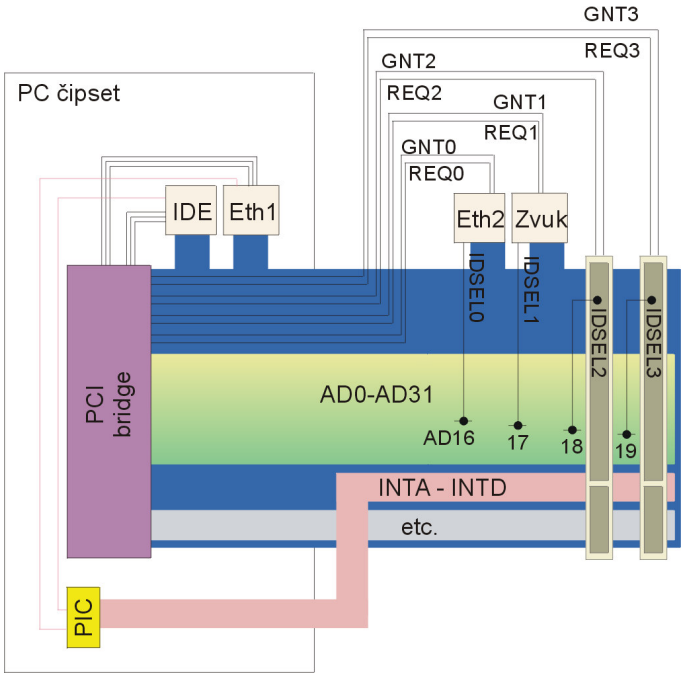
Schopnost resp. *neschopnost* busmasteringu bohužel nebývá dostatečně důrazně zdokumentována – tato informace bývá v manuálech spíše dovedně utajena.



Ještě horší je, že softwarově nelze zjistit, zda je konkrétní PCI slot schopen či neschopen bus-masteringu, kolik GNT signálů má konkrétní PCI bridge, který GNT signál obsluhuje který slot, či zda dokonce není některý z těchto signálů sdílen více sloty/zařízeními.

Typickými příznaky "zmrzačeného" slotu je, že zařízení je při startu nalezeno, jsou mu přiděleny nekonfliktní prostředky, operační systém se tváří třeba i úplně spokojeně – až na to, že zařízení nefunguje, tj. nepřenáší data. Např. síťová karta nepřijímá a neodesílá pakety, v Linuxu zůstává nula na softwarových počítadlech v utilitě ifconfig.

Toto chování lze vysvětlit tak, že konfigurace prostředků proběhne, protože v této fázi se používají portové a paměťové operace (nikoli PCI DMA neboli bus-mastering). Selže teprve "užitečný provoz", který běží přes "PCI DMA" (bus-master transakce).



Sdílený GNT signál

South Bridge používané na moderních motherboardech obsahují obvykle několik interních PCI zařízení a dále poskytují signály pro obsluhu několika rozšiřujících PCI slotů – vedle hlavních datově/adresních signálů se jedná mj. o výše zmíněné signály IDSEL, GNT a REQ.

Zařízení integrovaná on-chip v čipsetu neobsazují externí GNT/REQ páry. Zařízení integrovaná on-board na motherboardu (jako samostatné čipy) tyto signály naopak obsazují.

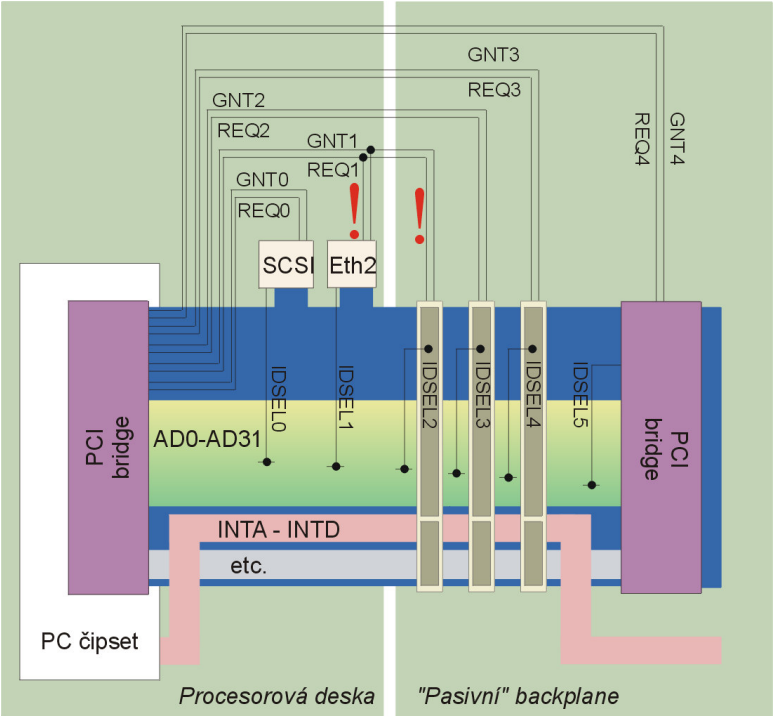
Externích signálů REQ/GNT má konkrétní typ bridge fixní počet – tím je také dán maximální počet připojených bus-master zařízení. V určitých konfiguracích modulárního PC hardwaru se může stát, že dva sloty sdílí jediný GNT signál[3]. Přesněji řečeno, obvykle je v těchto případech GNT signál pro rozšiřující PCI slot přiveden také na některé "onboard" zařízení. Signály IDSEL přitom zůstávají oddělené. Vzniknou tak svého druhu siamská dvojčata (viz obrázek na další straně).

Tak se stane, že jsou obě zařízení správně detekována operačním systémem, jsou jim oběma přiděleny nekonfliktní zdroje, operační systém se tváří spokojeně (alespoň zpočátku), ale ani jedno z obou zařízení nefunguje – nepřenáší užitečná data. Případně celý systém při pokusu o přenos dat zatuhne. Příčinou je pochopitelně zmiňovaný konflikt signálů REQ/GNT na sběrnici PCI.

V takových případech je řešením ponechat postižený rozšiřující slot neobsazený – sběrnice PCI přiděluje prostředky zařízením, nikoli prázdným PCI slotům. Případně pomůže vypnout postižené onboard zařízení, pokud to jde.

[3] Z logiky věci vyplývá, že na závadu je především sdílení signálu REQ (dochází ke kolizi budičů) – dokumentace však obvykle hovoří pouze o signálu GNT.

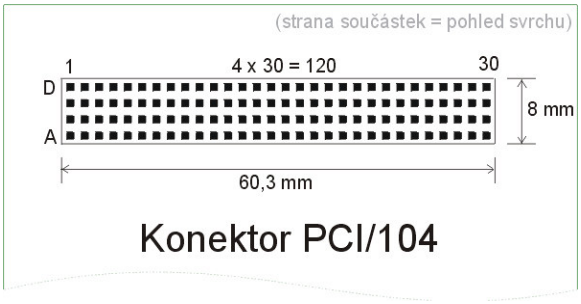
Seriózní výrobci průmyslového hardwaru poslední dobou obvykle publikují v manuálech k procesorovým deskám tzv. “mapu sběrnice PCI”. Jedná se vlastně o tabulku přiřazení IRQ a GNT signálů jednotlivým PCI slotům a onboard zařízením. Z této tabulky lze jednoznačně vyčíst, zda je tentýž GNT signál použit pro více zařízení. Kromě toho bývá v případě konfliktu pod tabulkou výrazné explicitní varování.



Přesto se občas vyskytne nedokumentovaný konflikt GNT signálů. Pokud máte podezření na tento typ konfliktu, pokuste se jej eliminovat uvolněním konkrétního PCI slotu (přestěhováním rozšiřujících karet). Také se pokuste získat katalogový list a nejlépe i aplikační poznámku k velkým bridgům na motherboardu (north bridge a south bridge) – pokud jde o novější součástky od firmy Intel, máte v tomto směru štěstí. Z této dokumentace lze zjistit, kolik vlastně daný PCI bridge podporuje externích PCI zařízení, resp. kolik má REQ a GNT pinů. Každé bus-master zařízení, ať již ve slotu či on-board (ovšem nikoli on-chip), potřebuje vlastní GNT signál – pozor, pokud máte “pasivní” backplane s velkým počtem PCI slotů, dceřinný PCI bridge se na nadřízené sběrnici počítá také mezi bus-master zařízení (zabere jednu sadu signálů, “představuje další PCI slot”).

Specifika sběrnice PC/104+ (PCI/104)

Původní sběrnice PC/104 je průmyslovou variantou sběrnice ISA – používá stohovatelný konektor, který obsahuje elektrické signály odpovídající “stolní” sběrnici s ISA. Novější varianta PC/104+ obsahuje navíc další stohovatelný konektor se sběrnicí PCI. Samotný tento PCI konektor se někdy nově označuje zkratkou PCI/104.



PCI sběrnice v konektoru PCI/104 ovšem neodpovídá klasickému PCI slotu. Klasický PCI slot podporuje na rozšiřující kartě nanejvýš jedno PCI zařízení. Konektor PCI/104 je stohovatelný, a proto obsahuje signály pro obsluhu několika PCI zařízení – nebo chcete-li pro emulaci několika konvenčních PCI slotů.

Konektor PCI/104 konkrétně obsahuje klasické čtyři IRQ signály (INTA-INTD), čtyři signály IDSEL, tři signály REQ a tri signály GNT. Jeden konektor PCI/104 tedy dokáže adresovat čtyři PCI sloty (nebo ekvivalentní zařízení) a každé může dostat vlastní IRQ, ovšem pouze tři z nich mohou

používat bus mastering. A to ještě za předpokladu, že žádný ze signálů IDSEL, REQ a GNT není sdílen s nějakým onboard zařízením na PC/104+ “motherboardu”.

Kromě nativních zařízení do slotu PCI/104 existují také redukce pro připojení klasické PCI karty do slotu PCI/104. V obou případech platí, že pokud má být možno PCI/104 zařízení stohovat, případně se vyhnout konfliktu s onboard hardwarem, je třeba aby rozšiřující PCI/104 zařízení bylo vybaveno mechanismem pro změnu konfigurace – chcete-li, aby se umělo tvářit jako kterýkoli ze čtyř PCI slotů. To lze zajistit větším počtem jumperů, nebo malým počtem jumperů ve spojení s polovodičovými multiplexery (čímž se ovšem konfigurovatelnost omezí na několik málo kombinací). V ideálním případě lze konfigurovat jednotlivé signály nezávisle.

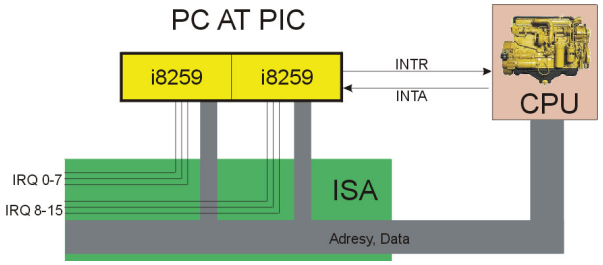
Zpracování přerušení v podání různých generací x86 PC/AT

Metody reprezentace a přenosu přerušení na sběrnicích se v průběhu několika generací PC/AT změnily. Navenek vypadá všechno takřka stejně, ale uvnitř jsou změny možná až překvapivě radikální.

Následující výčet generací procesorů je v zájmu přehlednosti zjednodušený. Jednotlivé vlastnosti se například u různých výrobců čipsetů (a procesorů) objevily o generaci později, případně naopak i v posledních modelech generace předchozí: cache, FPU, MMX, TSC, APIC, novější sběrnicové technologie, členění integrovaných periferií do různých pouzder apod. Tato kapitola se soustředí především na zacházení s IRQ, na které zmíněné vlastnosti povětšinou nemají podstatný vliv.

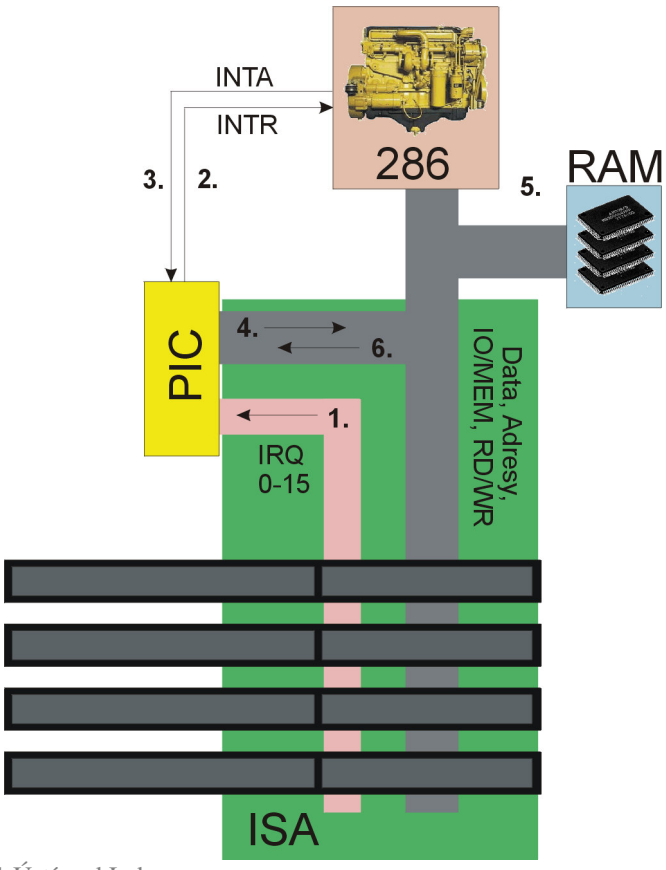
286

Signály ISA IRQ0-15 nevedou přímo do procesoru – na platformě PC AT se používá externí řadič přerušení (PIC – Programmable Interrupt Controller). Pro zjednodušení nebudeme dále pitvat fakt, že se jedná o dva osmivstupové čipy Intel 8259. Řadič přerušení multiplexuje šestnáct ISA přerušení na jediný signál INTR, který je předán procesoru – zároveň řadič procesoru na datových pinech sběrnice sdělí, který typ přerušení má být vyvolán.



Celý děj vypadá přesně takto:

1) periferie žádá o přerušení příslušným elektrickým signálem (ISA IRQ)
2) řadič vyšle procesoru signál INTR
3) procesor potvrdí příjem signálem INTA (INTerrupt Acknowledged) – po signálu INTA pošle dva pulzy
4) při druhém pulzu řadič přerušení vystaví na datových vodičích sběrnice ISA (na dolních osmi bitech) typ přerušení, které bylo vyvoláno. Pozor, typ přerušení neznamená číslo IRQ signálu, ale číslo položky v tabulce vektorů přerušení (je jich 256) – viz níže.
5) procesor provede volání obsluhy přerušení – tj. spustí se procedura, na kterou ukazuje příslušný vektor přerušení. (Implicitní instrukce `int` chová podobně jako explicitní instrukce `call` – tj. uloží do zásobníku návratovou adresu a provede skok na první instrukci obsluhy.) Obsluha



se baví s periferií, která přerušení vyvolala – podle druhu použitých prostředků buť přes I/O porty (PIO) instrukcemi `in`/`out` nebo přes paměťový přístup (MMIO) instrukcí `mov` a příbuznými

6) těsně než obsluha skončí, potvrdí řadiči přerušení (zápisem na jeho konkrétní port – instrukcí `out`), že přerušení bylo obslouženo.

"Vektor přerušení" je adresa jeho obslužné rutiny (adresa první instrukce rutiny v operační paměti). Standardní obslužné rutiny jsou součástí BIOSu, ale další software může obsahovat / instalovat své vlastní obsluhy. To může dělat operační systém, ovladače, v DOSu také přímo aplikace.

Vektory přerušení jsou uloženy v tabulce vektorů. Tabulka vektorů (v reálném režimu 8086) je uložena v prvním 1 kB operační paměti procesoru – tedy v bloku adres 0x0000 až 0x03FF.

Typ přerušení	Adresa	byte 0	byte 1	byte 2	byte 3
int 0x00	0x0000	vek	tor	č.	0
int 0x01	0x0004	vek	tor	č.	1
int 0x02	0x0008	vek	tor	č.	2
...	...	...	...	...	...
int 0xFF	0x03FC	vek	tor	č.	255

Jak již výše uvedeno, řadič přerušení neposílá procesoru číslo IRQ signálu, ale typ přerušení, neboli pořadové číslo 32bitové pozice v tabulce vektorů. Mapování šestnácti čísel IRQ na typy přerušení (int) tedy udržuje ve svých registrech řadič přerušení. Standardní „mapa" vypadá takto:

IRQ	INT	Zařízení
0	0x08	Časovač
1	0x09	Klávesnice
2	(0x0A)	Kaskáda na druhý čip řadiče
3	0x0B	COM2/COM4
4	0x0C	COM1/COM3
5	0x0D	Různé (XT pevný disk ?)
6	0x0E	Disketová jednotka
7	0x0F	Různé (Tiskárna)
8	0x70	Hodiny reálného času
9	0x71	Přesměrované IRQ2 (různé)
10	0x72	Různé
11	0x73	Různé
12	0x74	Různé (PS/2 myš)
13	0x75	Mat.koprocesor
14	0x76	Různé (IDE kanál 1)
15	0x77	Různé (IDE kanál 2)

Je zřejmé, že tabulka vektorů má mnohem více pozic (256), než je počet IRQ ve standardním PC (16, u XT 8). To znamená, že v tabulce vektorů jsou "díry". K čemu to?

Na platformě PC se totiž vyskytují i tzv. "softwarová přerušení", dostupná pomocí explicitní instrukce `int`. Tato explicitní softwarová "přerušení" obsazují volné vektory. "Softwarová přerušení" jsou velice nízkoúrovňovým prostředkem (na úrovni strojových instrukcí procesoru), jak umožnit registraci různých systémových volání na standardní a dobře známé pozice, které se kromě notoricky známého čísla vyznačují také stabilní sadou argumentů a stabilním způsobem jejich předávání (registry, zásobník).

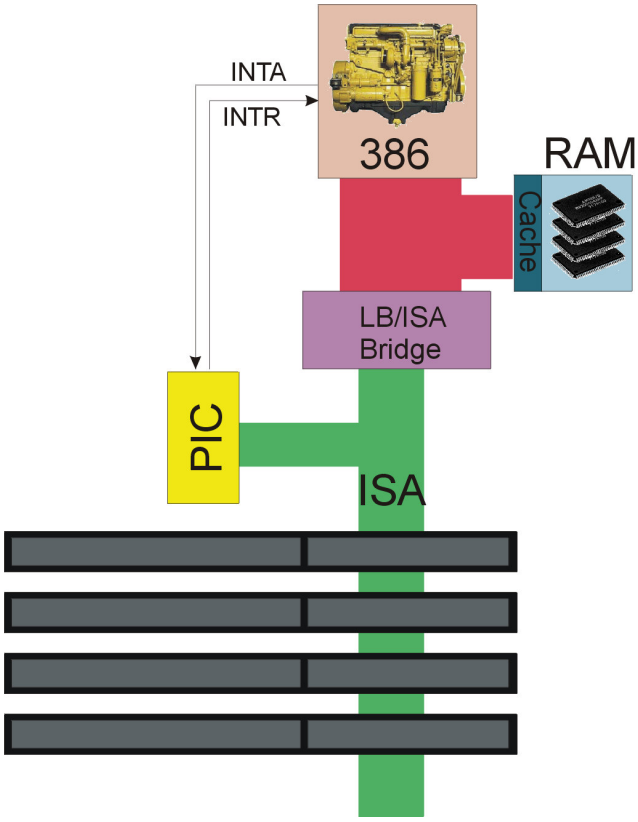
Slůvko "přerušení" je zde tedy matoucí – vlastně jde o prachobyčejné volání systémového API, nikoli o obsluhu externě generované události. Rozdíl mezi instrukcemi `call` a `int` je v tom, že argumentem instrukce `call` je adresa obsluhy, kdežto argumentem instrukce `int` typ přerušení (pozice v tabulce vektorů). Tento nepřímý způsob volání šetří objem kódu – není třeba v programovém kódu zapisovat explicitní "dereference".

Protože lze explicitně volat čistě softwarová přerušení, lze také explicitně volat obsluhy hardwarových přerušení. To lze využít např. k řetězení (rekurzi) obsluh – funkčnost původní obsluhy zůstane zachována, nově vložená obsluha si přidá nějaký svůj kód. Takto se v MS-DOSu řešilo např. transkódování klávesnice nebo použití časovače.

Pro úplnost dodejme, že v chráněném režimu procesoru x86 má tabulka vektorů přerušení (IDT) lehce jinou podobu, může začínat kdekoli v paměti, ale princip zůstává stejný.

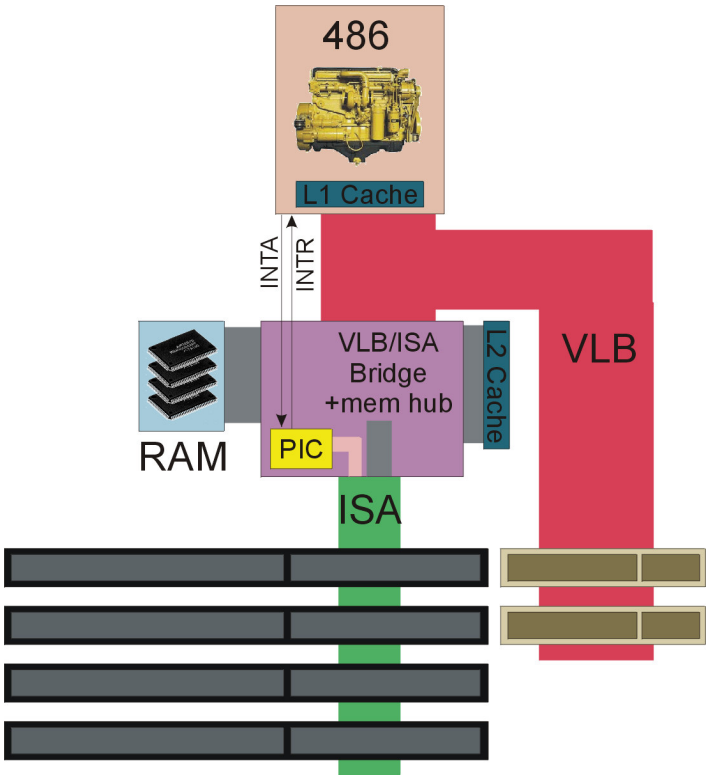
386

Pokud se týče zpracování přerušení, procesor 386 a jeho doprovodné "podhoubí" nepřineslo mnoho nového. Procesorová lokální sběrnice je kvůli rychlejšímu přístupu do paměti již rychlejší než ISA, proto jsou tyto dvě sběrnice vzájemně odděleny bridgem. Objevuje se paměťová cache. Na základním mechanismu zpracování přerušení PICem se ovšem nic nemění.



486

Pokud se týče zpracování přerušení, procesor 486 se opět nechová o mnoho jinak než původní 286. Změna oproti platformě 386 spočívá patrně především v tom, lokální procesorová sběrnice byla standardizována do podoby VESA Local Bus a lze na ni připojovat omezený počet rychlých periferií. Vesa Local Bus je ke konkrétní kartě přivedena společným slotem se sběrnicí ISA a při práci s periferiemi se používají klasická "ISA" IRQ.



Bridge okolo procesoru se začínají komplikovat, objevuje se druhá úroveň cache (první úroveň se stěhuje na procesor). Objevují se první implementace sběrnice PCI – host bridge je typicky připojen až za sběrnici VLB, PCI zde tedy vystupuje v roli čistokrevné periferní sběrnice. V samém závěru éry 486 se vyskytly i čipsety, které VLB úplně vypustily.

Pentium

Pokud pomineme nesmělé první vlaštovky na sklonku éry 486, sběrnice PCI se naplno prosadila právě ve spojení s procesory Pentium. Sběrnice ISA již není připojena přímo na "memory hub" (nyní north bridge), ale je zapojena až za sběrnici PCI, přes další bridge. Jinak řečeno, sběrnice PCI je nyní spojovacím článkem mezi north bridgem a south bridem – a zároveň obsluhuje zařízení v PCI slotech.

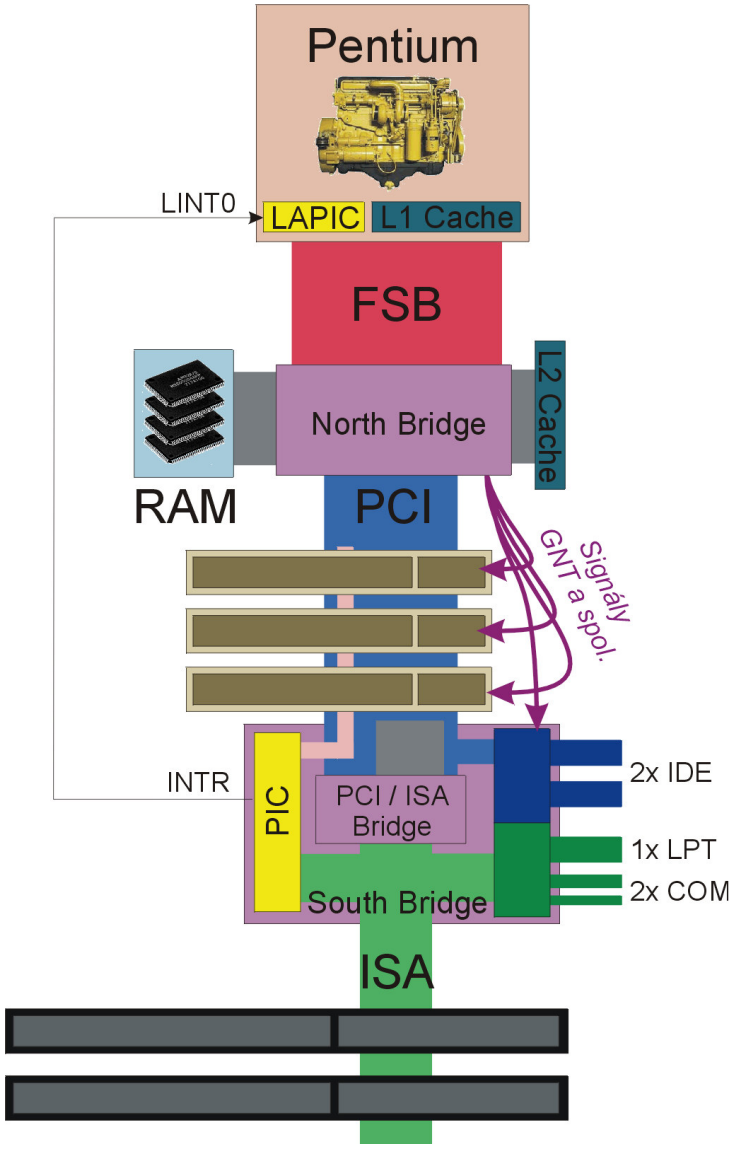
Pokračuje integrace standardních systémových součástek. Pomalé systémové periferie, které byly donedávna částečně v samostatných obvodech na motherboardu a částečně na rozšiřující kartě ve VLbus slotu, jsou nyní v některých případech sdruženy do south bridge. Jsou totiž připojeny na sběrnici ISA, a PCI/ISA bridge je beztak klíčovou součástkou south bridge. IDE řadič má nyní kromě původního ISA portu také PCI port – kvůli rychlejšímu přístupu.

Sběrnice PCI není pupeční šňůrou vázána na architekturu x86 PC a nebyla součástí původního PC AT – aby se zachovala zpětná kompatibilita, jsou signály INTA-INTD na platformě PC mapovány dodatečným obvodem (maticovým přepínačem) na původní ISA přerušení. Schéma mapování se konfiguruje mechanismem PnP – počáteční nastavení obstará BIOS, případně přiloží ruku k dílu i operační systém. Po pravdě řečeno, signály INTA-INTD se vlastně sběrnicových protokolů PCI přímo neúčastní.- patrně v duchu svého prapůvodního účelu, kterým je signalizace asynchronních událostí z periferií na procesor.

Pitoreskní je, že zatímco “master” bridge na sběrnici, tj. host bridge, je součástí north bridge, tradiční řadič přerušení je nadále dostupný přes sběrnici ISA – a ta je dostupná přes south bridge. Takže procesor s řadičem PIC komunikuje přes sběrnice FSB a PCI a příslušné bridge (přes north bridge a přes PCI/ISA bridge v south bridgi).

Tak jako jiné systémové periferie, PIC je nyní typicky součástí south bridge.

Zdá se, že v signalizaci přerušení zde nastává první zásadnější změna. Signál INTR zůstává alespoň u Intelských součástek nadále propojen, ale signál INTA (Interrupt Acknowledge) zřejmě zmizel v propadlišti dějin. Některé zdroje uvádějí, že u této generace čipsetů se již signály INTR a INTA přenášejí po sběrnici PCI ve formě “INTR transakcí” – patrně to bude pravda u jiných výrobců, než je Intel. Ostatně podle specifikace PCI 2.2 smějí “INTR transakce” používat i přímo koncová zařízení na sběrnici PCI.



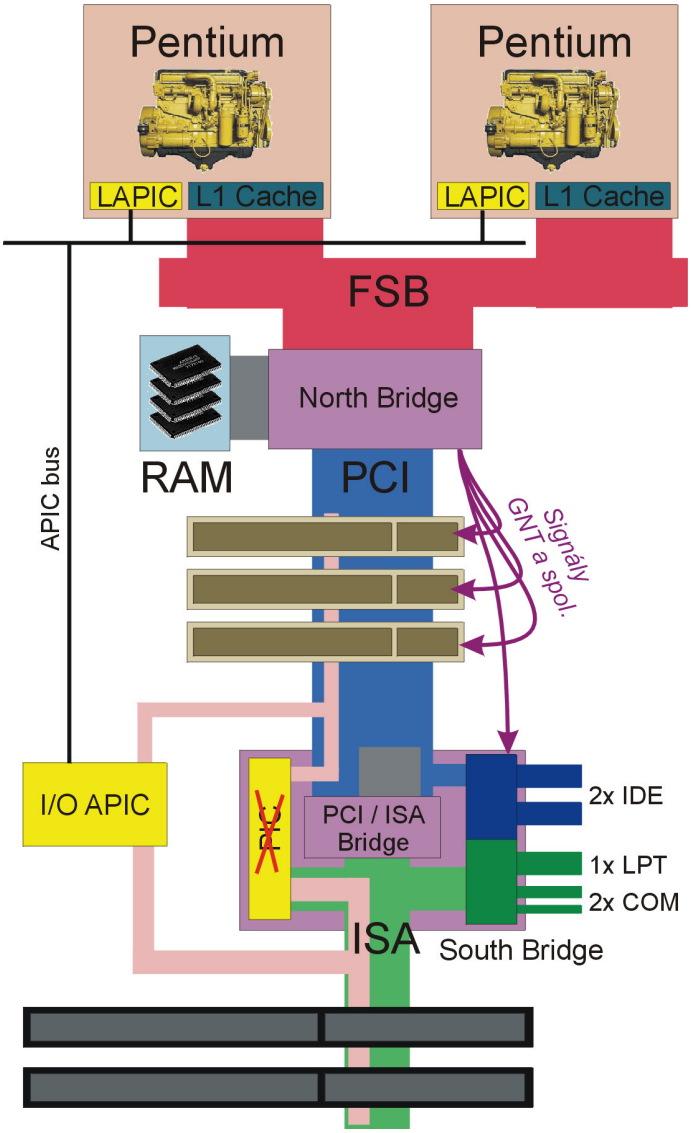
South bridge se vůči integrovanému PICu tváří jako procesor - přijme signál INTR a zkonvertuje jej na PCI transakci (“zprávu”) s tímtéž významem. Tuto zprávu odešle na north bridge, kde je zkonvertována do sběrnicového protokolu FSB a odeslána procesoru. Procesor pošle ACK stejným způsobem – ACK transakce proběhne přes FSB a PCI na south bridge, který ji zkonvertuje na vstupní signál PICu. Zbytek ošetření přerušení probíhá stejně - je tedy zakončen explicitním zápisem na ISA port PICu, který znamená, že přerušení je obslouženo.

Intermezzo – na scénu přichází APIC

Co je to APIC

Zkratka APIC znamená “Advanced Programmable Interrupt Controller” – jde o novou vývojovou generaci řadiče přerušení na platformě PC. Jeho nejviditelnějším přínosem je to, že ruší dosavadní omezení platformy PC na šestnáct IRQ signálů. Nejběžnější APICy od firmy Intel, ať už samostatné nebo integrované v čipsetech, rozšiřují počet IRQ signálů na 24 a totéž platí o integrovaných APICech v čipsetech VIA KT333 a vyšších. Specifikace APIC nicméně žádné takové omezení neobsahuje – potažmo na ně nespoléhá ani implementace softwarové obsluhy APICů v operačních systémech. Je možné zkonstruovat systém se stovkami IRQ linek.

Praktickým důsledkem nasazení APICu je, že není třeba sdílet IRQ signály – což zrychluje odezvu systému na jednotlivá IRQ, snižuje zátěž procesoru a odstraňuje možný zdroj nestability systému. Ke snížení zátěže rovněž přispívá vylepšená signalizace - APICu již není třeba potvrzovat přijetí přerušení a ukončení jeho obsluhy.



Podrobnosti o architektuře

Ve skutečnosti se nejedná o jeden obvod, ale přinejmenším o dva, které vzájemně spolupracují:

- I/O APIC – právě on je náhradou původního PICu. Tento obvod sdružuje jednotlivé IRQ linky a informaci i vyvolaném přerušení předává jednotným způsobem na procesor.
- Local APIC (LAPIC) – je integrální součástí procesoru. U firmy Intel je lokální APIC standardní součástí všech procesorů již od prvních generací rodiny Pentium (přesně od P24T, což byl overdrive do patice 486).

Obě komplementární části spolu komunikují po speciální sériové sběrnici zvané překvapivě “APIC bus”. I/O APIC zařídí “špinavou práci” se sběrem přerušení z jednotlivých elektrických signálů, LAPIC zařídí obsluhu přerušení na procesoru. Událost zvaná přerušení tedy již není přenášena jednotlivými elektrickými signály, ale jako zpráva na “APIC bus”.

Architektura APIC ovšem umožňuje dokonce i vysloveně distribuované konfigurace – s několika I/O APICy a několika LAPICy.

APIC je např. nezbytnou součástkou multiprocesorových strojů s procesory x86 – viz ilustrace. Z logiky věci má smysl obsluhovat každé přerušení pouze jedním procesorem – a jedině pomocí APICů lze směrovat přerušení z různých IRQ linek různým procesorům, což je jediný možný přístup s ohledem na rozkládání zátěže.

Těžko říci, zda APIC z multiprocesorové oblasti přímo pochází, nebo zda jde o shodu okolností – historicky se ovšem APICy objevovaly nejprve v multiprocesorových strojích, na uniprocesorových stolních motherboardech jde o relativní novinku.

Větší počet I/O APICů v systému naopak umožňuje obsluhovat v systému větší počet periferních zařízení bez potřeby sdílet IRQ signály. Pokud nestačí jeden I/O APIC čip s 24 nebo 32 vstupními linkami, použije se jich víc – vedlejším důsledkem je, že mohou být blíže svým periferiím, takže se šetří délkou IRQ signálů a plochou desky plošných spojů. Větší počet I/O APICů v systému nepředstavuje problém – jsou propojeny smíněnou sériovou sběrnicí APIC bus, na které mají každý svůj unikátní identifikátor.

Kvůli zpětné kompatibilitě dodnes stroje vybavené APICem obsahují klasický PIC a startují s jeho pomocí, tj. s přerušeními obsluhovanými konvenčním způsobem a mapovanými s omezením na původní sadu 16 IRQ signálů[4]. Použití APICu je třeba povolit v BIOSu a jeho vlastní inicializaci zařizuje operační systém. V DOSu se tedy APIC neuplatní, k jeho využití je zapotřebí relativně čerstvá verze Windows, Linuxu, FreeBSD apod.

V APIC režimu se sníží výskyt sdílených přerušení, což zvyšuje rychlost obsluhy přerušení a zamezuje plýtvání procesorovým časem na "plané obsluhy". Zároveň v případě APICu odpadají dva ACKy - zápisové cykly po sběrnici PCI (jeden z nich dokonce až po ISA), které jsou nezbytné u původního PICu. Obsluha přerušení pomocí APICu je po všech stránkách čistší, rychlejší a efektivnější než tradičním způsobem přes PIC.

APIC vs. ACPI

Popis zapojení IRQ signálů od jednotlivých PCI a jiných zařízení na různé vstupní piny jednotlivých I/O APICů je na konkrétní základní desce uložen v jejím ACPI BIOSu. ACPI je totiž něco víc, než jen vylepšený power-management (nástupce APM). ACPI mluví i do PnP, konkrétně do směrování přerušení. Od ACPI BIOSu se operační systém před inicializací APICu dozví, jak jsou přerušení na základní desce zapojena. Proto lze tvrdit, že vlastnosti APIC a ACPI sice nejsou totožné (neplést!), ale prakticky spolu úzce souvisejí.

[4] multiprocesorová PCčka z téhož důvodu startují v uniprocesorovém režimu.

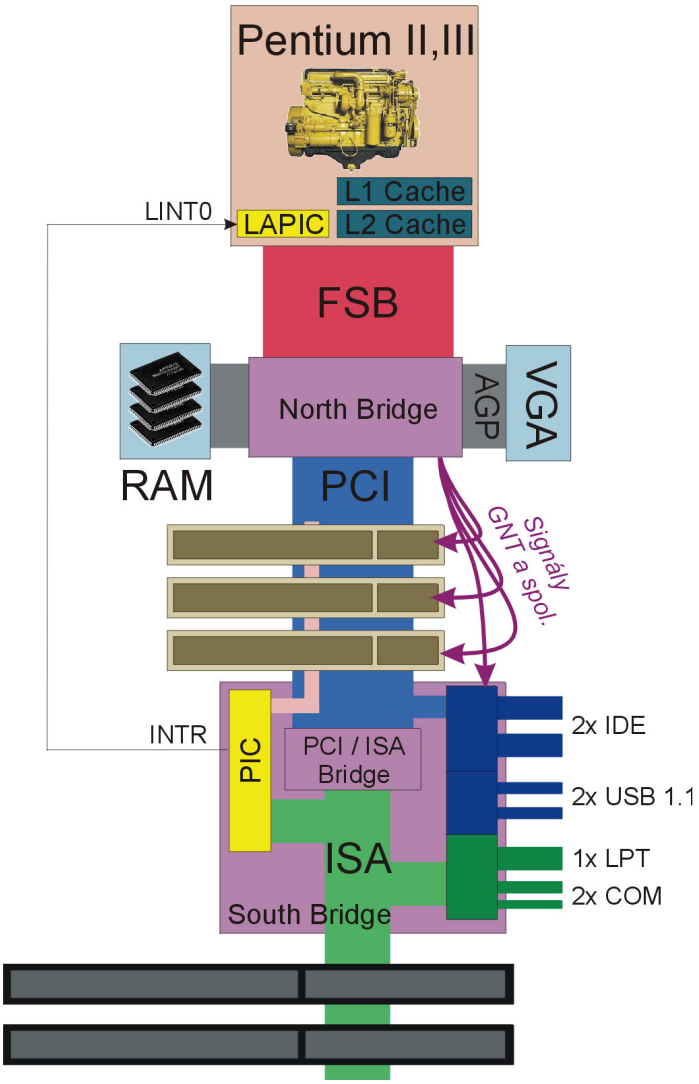
Pentium II, Pentium III

V oblasti směrování IRQ a zpracování přerušení s touto generací procesorů nepřišlo mnoho nového. PCI zůstává spojnicí mezi north bridgem a south bridgem. V případě multiprocesorových strojů se APIC používá analogicky jako u procesorů Pentium.

Obecně je se na této platformě objevilo několik málo novinek. Nejvýznamnější je asi přestěhování L2 cache na procesor. Cache se vyskytuje buď v podobě druhého čipu ve společném pouzdře (viz Pentium Pro) nebo nověji přímo na čipu s procesorem.

Na north bridgi přibyla podpora AGP – rychlé sběrnice pro grafický adaptér. V operačním systému se tváří částečně jako PCI zařízení, hardwarově je ale odlišná a umí navíc některé rychlé operace s pamětí – např. sdílení systémové RAM v roli videopaměti a DMA přesuny velkých bloků dat (textury) mezi systémovou operační pamětí a videoRAM.

Na South Bridgi přibyla podpora USB (zařízení na sběrnici PCI). Pomalé porty (sériové, paralelní, floppy, řadič klávesnice) se naopak pomalu stěhují do externích "SuperIO" čipů na sběrnici ISA.

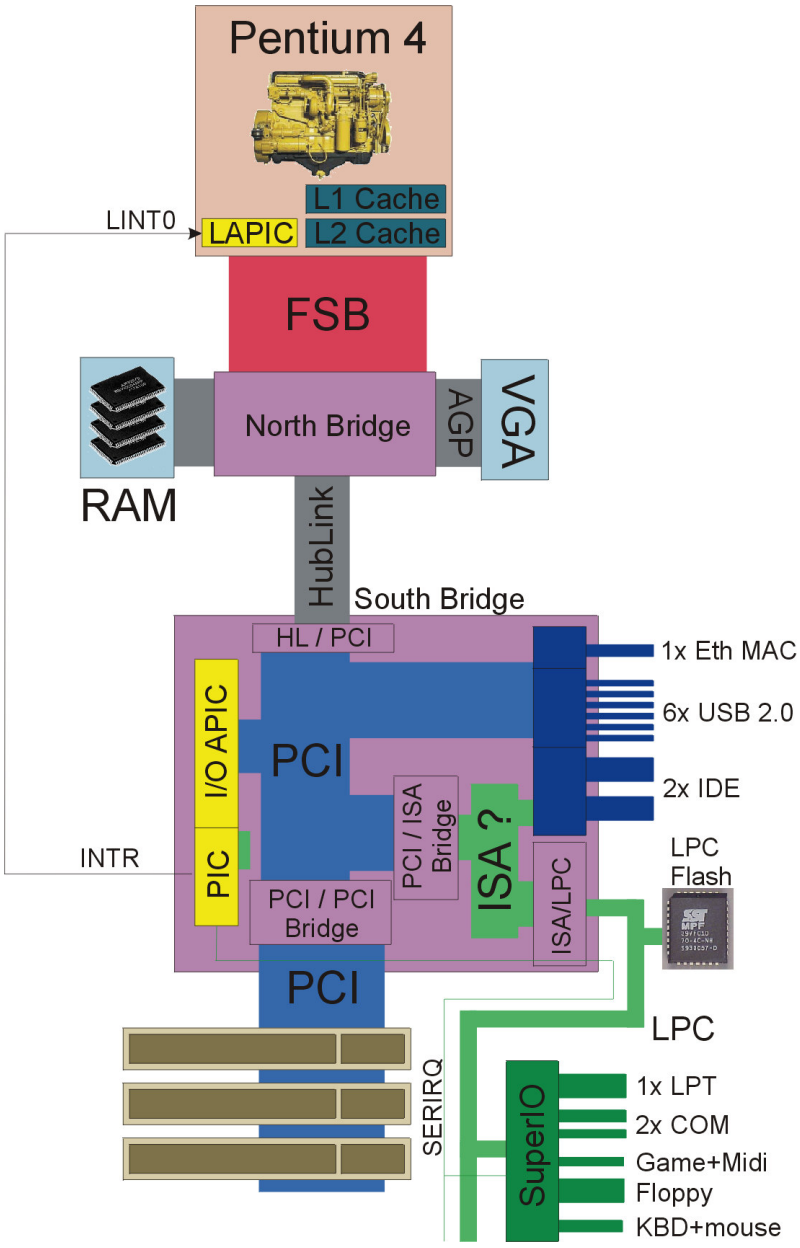


Pentium 4

Spolu s P4 přišla alespoň u Intelu další rošáda ve sběrnicích.

Jako spojnice mezi north bridgem a south bridgem již neslouží PCI, ale zúžená sběrnice zvaná HubLink (8 resp. 12 bitů) s kapacitou 266 MBps, tj. asi jako PCI32/66MHz.

Sběrnice PCI nyní opět plní funkci skutečně periferní sběrnice – vyskytuje se pouze "jižně" od south bridge.



Vlastně je to trochu složitější. South bridge (ICH) obsahuje interní PCI-to-PCI bridge. Externí PCI zařízení jsou vidět logicky až za tímto PCI bridgem, tj. na "sekundární" PCI sběrnici. Interní PCI zařízení integrovaná v south bridgi jsou ovšem vidět na "primární" PCI sběrnici, hned za host bridgem a *před* PCI bridgem. Logicky tedy vyvstává otázka, kde se vlastně nachází host bridge – tj. můstek mezi FSB a touto primární PCI sběrnicí Možné odpovědi jsou dvě: buď jde o bridge mezi interní primární PCI a sběrnicí HubLink (a tedy se nachází v south bridgi), nebo jde o bridge mezi HubLinkem a FSB (a tedy se nachází v north bridgi). Jinými slovy, je otázkou, zda HubLink emuluje PCI, nebo zda jde o prodloužení interní sběrnice north bridge. Intel ve svých katalogových listech interní architekturu čipsetů nijak nerozvádí, takže toto lze těžko říci. Prakticky na tom beztak málo záleží.

Pokud se týče integrovaných periferií, USB dospěla do verze 2.0 a razantně se zvedl počet USB portů, pomalé periferie se definitivně odstěhovaly mimo south bridge do SuperIO skanzenů.

Zajímavá je sběrnice LPC, přes kterou jsou pomalé periferie připojeny. LPC = Low Pin Count. Sběrnice LPC osahuje čtyřbitovou datovou cestu plus několik řídících signálů, používá víceslovné "transakce" – na tomto základě emuluje sběrnici ISA. IRQ signály se nepřenášejí transakcemi, ale jediným signálem SERIRQ, který je přiveden přímo na odpovídající kompatibilní vstup řadiče přerušení.

Přes LPC je připojen i tzv. “firmware hub”, tj. paměť Flash obsahující BIOS základní desky.

Nativní ISA se na south bridgi již nevyskytuje – přinejmenším není dostupná na vnějších vývodech tohoto čipu. Je otázkou, zda se signály sběrnice ISA ještě vyskytují někde uvnitř south bridge a v jakém rozsahu. V systému stále existují zařízení s klasickými ISA adresami – většina je jich připojena pomocí LPC (takže by se dalo uvažovat o PCI-to-LPC bridgi maskovaném jako PCI-to-ISA bridge), ale vedle nich mají ISA port také IDE kanály a klasický řadič přerušení (PIC), systémový časovač apod. Tyto nejsou vidět jako samostatná PCI zařízení resp. samostatné ISA bridge, takže se lze domnívat, že v nějaké formě ISA uvnitř south bridge dosud žije. Těžko říci jak moc je úplná, na jakém taktu vlastně běží apod.

Na poli zpracování přerušení došlo k další drobné revoluci: south bridge ICH4/ICH5, vyskytující se jako doprovod north bridgů i845 - i875, obsahují nyní integrovaný I/O APIC. Zmizela ovšem samostatná sběrnice APIC bus[5], přerušení se nyní směrují k Local APICu jako transakce na sběrnicích HubLink a FSB. Za těchto okolností je mírně překvapivé, že dosud existuje signál INTR, který vede ze south-bridge (z PICu) rovnou na procesor – patrně jde o relikt nezbytný pro “legacy PIC” režim.

Zmíněné south bridge mají nyní osm PCI IRQ vstupů – první čtyři jsou standardní INTA-INTD, další čtyři jsou konfigurovatelné jako INTE-INTH nebo jako vstupy pro všeobecné použití (GPIO).

Jak je tomu u jiných výrobců procesorů a chipsetů

Ostatní výrobci jdou podobnou cestou. Také používají APIC podle specifikací firmy Intel, také mají v novějších čipsetech APIC integrovaný, také doručují přerušení jako transakce na PCI i na proprietárních rychlých sběrnicích. Tito ostatní výrobci mají své vlastní rychlé sběrnice – např. otevřenou/AMD HyperTransport, VIA VLink, SiS MuTIOL apod.

Zajímavou anomálií je integrace paměťového řadiče na procesoru AMD Opteron/Athlon64. To způsobuje lehce jinou topologii “velkých systémových bridgů” a také specifické chování multiprocesorových systémů (na jazyk se dere zkratka NUMA).

[5] přesněji řečeno, na south bridgi jsou její signály dosud dostupné – zřejmě aby se umožnilo použití rozšiřujících externích I/O APICů.

Shrnutí – nový pohled na přerušení

Během několika generací PC AT se přerušení odpoutalo od “drátěných” IRQ a více se přiblížilo svému archetypu, kterým je “asynchronní událost”. Počet IRQ signálů do budoucna již není omezujícím faktorem. Na nejnovějších strojích je přerušení předáváno jako transakce (zpráva, frame) na “zúžených” či “částečně serializovaných” sběrnicích.

Na sběrnici LPC se emulují ISA IRQ signály pomocí sériového signálu SERIRQ. “Paralelní” PCI signály INTA-INTD nadále fungují, různé segmenty PCI lze obsluhovat několika sadami IRQ. Již v éře procesorů Pentium se signály INTR/INTA z PICu na CPU emulovaly transakcemi na sběrnici PCI. Po pravdě řečeno, podle specifikace PCI 2.2 mají i koncová zařízení legitimní možnost posílat IRQ rovnou jako sběrnicové transakce (což se patrně příliš neujalo). Jednoúčelová sběrnice APIC Bus zřejmě patří minulosti, alespoň na uniprocesorových systémech – nahradily ji transakce na rychlých sběrnicích (PCI, HubLink, FSB).

Ve světle tohoto posledního vývoje je lehce úsměvné, že operační systémy dodnes tvrdošíjně pojmenovávají typy přerušení i v režimu APIC jejich původními čísly IRQ – přestože reálně na sběrnici PCI se tento koncept dávno vyprázdnil a se zmizením sběrnice ISA postrádá jakýkoli reálný základ. Kdyby nebylo APICů, které přinesly IRQ 16-24 (na strojích ServerWorks se dvěma procesory až IRQ31, se čtyřmi procesory až IRQ63), uživatel by si ani nevšiml, že je něco jinak – když mu klávesnice stále visí na IRQ1, disketová jednotka na IRQ6 a koprocesor na IRQ13. Výrobci operačních systémů jsou si asi vědomi, že uživatel je rád balamucen, že balamucení je v zájmu zachování uživatelova duševního zdraví, smyslu života, vyrovnaného fungování fyziologických pochodů apod.

Rejstřík zkratek

IRQ	Interrupt Request
DMA	Direct Memory Access, přímý přístup periferie do paměti bez účasti procesoru
PIO	Port I/O, vstup a výstup přes I/O porty (instrukcemi in, out)
MMIO	Memory-Mapped I/O, periferní vstup a výstup přes paměťově mapovanou oblast (instrukcí mov).
REQ	PCI bus Request (při bus-master operacích)
GNT	PCI bus Grant (při bus-master operacích)
FSB	Front Side Bus, lokální sběrnice mezi procesorem a north bridgem
ISR	Interrupt Service Routine, obslužná rutina přerušení
GMCH	Graphics and Memory Controller Hub - north bridge s podporou AGP (Intel P4)
MTXC	System Controller - north bridge (Intel Pentium/PII/PIII)
ICH	I/O Controller Hub – south bridge (Intel P4)
PIIX	PCI-to-ISA/IDE Xcelerator – south bridge (Intel Pentium/PII/PIII)
PIC	Programmable Interrupt Controller, klasický řadič přerušení na platformě IBM PC
APIC	Advanced PIC, rozšířený řadič přerušení (zkratka se často používá ve smyslu I/O APIC)
I/O APIC	Input/Output APIC, jedna ze dvou komplementárních částí architektury APIC. Sdružuje IRQ signály a konvertuje je na zprávy zasílané procesoru po sériové sběrnici APIC Bus.
LAPIC	Local APIC, integrální součást moderních procesorů – “přijímač” přerušení na sběrnici APIC Bus na straně procesoru. Zajišťuje spouštění obsluh přerušení na základě přijatých IRQ událostí.
ACPI	Advanced Configuration and Power Interface

Literatura

Intel 875 (P4) North Bridge (82875 GMCH)
ftp://download.intel.com/design/chipsets/datashts/25252502.pdf

Intel 875 (P4) South Bridge (82801EB ICH5)
ftp://download.intel.com/design/chipsets/datashts/25251601.pdf

Intel 845G (P4) North Bridge (82845 GMCH)
ftp://download.intel.com/design/chipsets/datashts/29074602.pdf

Intel 845G (P4) South Bridge (82801DB ICH4)
ftp://download.intel.com/design/chipsets/datashts/29074401.pdf

Intel 440BX (PII-PIII) North Bridge (82443 AGPset)
http://www.intel.com/design/chipsets/datashts/29063301.pdf

Intel 430TX (Pentium) North Bridge (82439TX MTXC)
ftp://download.intel.com/design/chipsets/datashts/29055901.pdf

Intel 430TX/440BX South Bridge (82371EB PIIX4)
ftp://download.intel.com/design/intarch/datashts/29056201.pdf

Obchodně laděné informace o čipsetu VIA KT333
http://www.via.com.tw/en/apollo/KT333.jsp
http://www.via.com.tw/en/images/KT333/WP2501002KT333RIB.PDF

Intel 21154 PCI-to-PCI bridge (v samostatném pozdře)
http://www.intel.com/design/bridge/docs/21154_documentation.htm

Různé PCI-to-ISA bridge

Winbond W83628F+W83629D:
http://www.winbond.com/e-winbondhtm/partner/PDFresult.asp?Pname=631

Intel 82380AB (Part of the Intel 380 mobile chipset):
http://www.intel.com/design/chipsets/datashts/29056302.pdf

National Semiconductor PC87200:
http://www.national.com/ds.cgi/PC/PC87200.pdf

Specifikace sběrnice Intel LPC 1.1
http://www.intel.com/design/chipsets/industry/25128901.pdf

Specifikace PCI/104
http://www.pc104.org/technology/PDF/PCI-104%20v1_0.pdf

Intel 82093 IO APIC (v samostatném pouzdře)
ftp://download.intel.com/design/chipsets/datashts/29056601.pdf

Windbond W83627HF SuperIO čip s rozhraním LPC
http://www.winbond.com/PDF/sheet/w83627hf.pdf

Zapojení vývodů procesorů 286, 386, 486
http://www.hardware-bastelkiste.de/cpu_286.html
http://www.hardware-bastelkiste.de/cpu_386.html
http://www.hardware-bastelkiste.de/cpu_486.html

Přerušení na platformě 80x86 (tradiční PIC)
http://www.csc.uvic.ca/~mcheng/360/notes/INT86.html

Sběrnice v PC – přehled od firmy Rambus
http://www.rambus.com/rdf/rdf2002/pdf/2FeibusIntro.pdf

Úvod do PCI, sponzorovaný firmou Xilinx
http://www.cs.uml.edu/~bill/cs592/PCI_slides.pdf

Úvod do PCI Express, od firmy Intel
ftp://download.intel.com/intelpress/pciexpresscomplete/PCIEC_Tutorial.pdf

Vynikající poznámka o I/O APICech, od fy. Microsoft
http://www.microsoft.com/whdc/hwdev/platform/proc/IO-APIC.mspx

Aplikační poznámka AMD - lehce irelevantní, okrajově o PIC/APIC/LAPIC
http://www.amd.com/us-en/assets/content_type/white_papers_and_tech_docs/24919.pdf

Poznámky k nastavení interruptů pro karty Intel Dialogic (obsahuje přehled priorit IRQ)
http://resource.intel.com/telecom/support/tnotes/gentnote/dl_hard/tn169.htm

USENET příspěvek o zpracování přerušení v čipsetech ALI Alladin

Autor: Geoffrey Levand
Diskusní skupina: comp.lang.asm.x86
Datum: 18.5.1999

When an ISA device interrupts, the 8259 expects to be connected to the CPU. The signaling between CPU and 8259 is very archaic. The 8259 asserts the CPU's INTR. Then the CPU pulses the INTA# of the 8259, and on the second pulse, the 8259 puts a byte on the data bus that indicates the highest priority interrupt that is pending.

When the system has a north bridge (host to PCI) and a south bridge (PCI to ISA), the 8259 is isolated from the CPU across the PCI bus. There are no INTR and INTA# signals on the PCI bus, so the PCI designers cooked up the PCI interrupt ack to handle this.

The 8259 is inside the south bridge, so the south bridge can pretend it is the CPU when the 8259 asserts the INTR. The south bridge forwards the interrupt on to the north bridge using a PCI interrupt line (I'm familliar with the Alladin V, and it has a dedicated line).

The north bridge then fakes-out the CPU by asserting INTR. While the CPU is doing the double pulse, the north bridge does an interrupt ack. The south bridge gets this and fiddles with the 8259 to get the byte of data and puts it on the PCI bus. The north bridge gets the data byte, and forwards it on to the host bus.