Maximální výpočetní výkon

O výkon osobních počítačů se až doposud staral hlavně procesor CPU. To už ale dnes neplatí. Použijte výpočetní výkon grafické karty a váš počítač může být až **DESETKRÁT RYCHLEJŠÍ**.

THOMAS LITTSCHWAGER

Pokud často pracujete s náročnými aplikacemi, potřebujete co nejvýkonnější počítač. Dlouho platilo, že pokud potřebujete vysoký výpočetní výkon, musíte si pořídit co nejvýkonnější procesor CPU.

Situace se však mění. Dnes obsahuje počítač více procesorů a všechny disponují ohromným výpočetním výkonem. Jeden z nich však vyčnívá. Je to GPU (Graphics Processing Unit), tedy procesor, kterým je osazena grafická karta. Moderní hry vyžadují, aby grafický procesor zvládl v reálném čase zpracovat řadu simultánních složitých výpočtů, a ve většině případů je GPU výkonnější než hlavní procesor počítače CPU. Čipy grafických karet pracují o frekvenci až 1 000 MHz a využívají superrychlou paměť s kapacitou až 2 GB. Těžko byste hledali rychlejší systémový koprocesor.

To je důvod, proč se výrobci grafických karet a řada dalších programátorů rozhodli využít výpočetní kapacitu grafických karet pro jiné účely, jako například pro úpravu videa a fotografií, pro simulace přírodních jevů a pro výpočty predikce vývoje finančních trhů. Společnost nVidia vytvořila za tímto účelem před třemi lety architekturu CUDA (Compute Unified Device Architecture), což je programovací rozhraní se syntaxí podobnou programovacímu jazyku C, pomocí kterého lze zpracovávat procesy programů určených pro zpracování na CPU v grafickém čipu. Rozhraní CUDA je však určeno pouze pro grafické karty od generace GeForce 8000 výše. Konkurenční karty ATI od společnosti AMD podporují všeobecný standard OpenCL od Khronos Group (OpenCL dnes podporují i karty nVidia), který umožňuje provozování programů na procesorech kompatibilních s OpenCL (tedy jak CPU, tak GPU). Do hry nedávno vstoupil i Microsoft, který do nového rozhraní DirectX 11 zahrnul patřičnou instrukční sadu, díky které lze spouštět procesy programů na grafickém procesoru GPU.

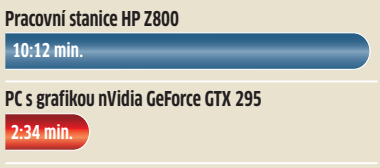
V tomto článku vám vysvětlíme, jak výpočty na grafickém procesoru fungují, které programy dokáží výpočetní síly GPU využít a jakým způsobem je možné maximalizovat výkon svého počítače.

Technologie: GPU podporuje CPU

O softwarovou stránku věci je díky výše zmíněným rozhraním postaráno. Výrobci grafických karet slibují, že díky využití výpočetního výkonu GPU lze dosáhnout v určitých situacích zrychlení systému až o několik tisíc procent. Otázkou zůstává, proč bylo dodnes tak málo programů, které dokáží využít potenciálu grafického čipu, a jak je to vůbec možné, že jsou grafické čipy tak rychlé. Obě otázky mají společnou odpověď: základní problém spočívá v tom, že CPU a GPU pracují odlišným způsobem. Moderní CPU obsahují čtyři jádra a k některým z nich lze pomocí technologie HyperThreading připočítat stejný počet jader virtuálních, takže

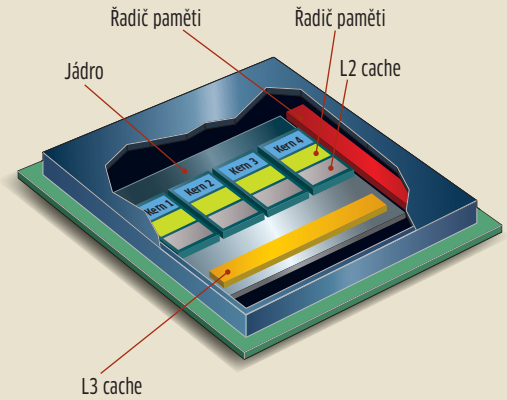
Přímý souboj dvou výkonných sestav

Uvedený graf ukazuje, jak dokáže zapojení GPU urychlit činnost počítače. První sestavou je výkonná pracovní stanice HP Z800, která disponuje dvěma procesory Xeon celkem s 16 jádry, a druhou je běžné PC s grafickou kartou nVidia GeForce GTX 295 se 480 unifikovanými jednotkami.
Úkol: Zpracování videa pomocí kodeku H.264 a programu CyberLink PowerDirector 8.
Výsledek: Pracovní stanici Z800 zabral převod videa 10 minut a 12 sekund, zatímco sestava s grafikou GeForce GTX 295 jej zvládla za čtvrtinovou dobu.



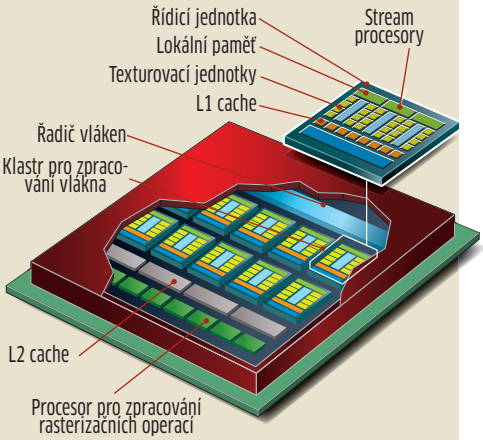
Intel Core i7

Nové procesory Intelu jsou osazeny až čtyřmi fyzickými jádry se samostatnou vyrovnávací pamětí na úrovni L2. Každé fyzické jádro disponuje jedním virtuálním, takže celkem mají Core i7 až osm jader. Výměnu dat zajišťuje společná vyrovnávací paměť L3.



Grafický procesor

GPU (na obrázku uvádíme schéma čipu nVidia GT 200) obsahuje 40× více výpočetních jednotek. 240 stream procesorů je uskupeno do deseti klastrů schopných samostatně zpracovávat jednotlivá a na sobě nezávislá vlákna programu.



CPU navenek vykazuje celkem osm jader. V těchto procesorech lze tedy zároveň provozovat až osm programů (nebo, lépe řečeno, až osm programových vláken). Jádra byla vyvinuta pro všeobecné použití, takže jsou velmi flexibilní a v každém kroku se každé jádro dokáže vyrovnat s odlišnými typy úloh.

Výkon GPU: 240 jader proti 8

Grafické čipy dnes mohou nabídnout až 240 jader, tedy čtyřicetkrát víc než nejmodernější CPU. Říká se navíc, že očekávaný grafický čip ATI Radeon 5000 má obsahovat až 1 600 jader. Tato jádra (kterým se v případě grafických čipů přezdívá streamovací jednotky nebo stream procesory) však nejsou úplně samostatná, jako je tomu v případě jader CPU, ale jsou sdružena do klastrů, které v každém stream procesoru zpracovávají jedno vlákno. V rámci jednoho klastru však dokáže GPU v daném vláknu využít pouze jednu zpracovávanou operaci, a nehodí se tak pro zpracování složitých úkolů. Řadě programů ale stačí, aby procesor GPU vykonával pouze jeden úkol, a právě v tomto případě dokáže GPU plně ukázat brutální výpočetní výkon nakumulovaných jader. Uvedeme praktický příklad. Procesor má za úkol spočítat, kolikrát se v dané knize objevuje určité slovo. CPU tuto úlohu zvládne tak, že začne na první stránce, prozkoumá celý text a skončí na stránce poslední. GPU však rozdělí knihu na malé části, které rozdělí ke zpracování mezi všechna dostupná jádra, a spočítá tak výskyt daného slova ve zlomku času, který tato úloha zabere klasickému CPU. V praktickém využití je tedy GPU nejvýhodnější pro výpočty vědeckých úloh a pro zpracování videa. Nepracuje se tu s množstvím stran jedné knihy, ale se sčítáním a násobením čísel s plovoucí desetinnou čárkou ve velkých maticích. Jde tedy o opakované vykonávání stejného úkonu.

Infografika na straně 58 ukazuje technologický rozdíl mezi zpracováním úlohy v rámci CPU a GPU. Je-li třeba, aby program vykonával řadu různých procesů, kvůli nižší taktovací frekvenci a kvůli omezením jednotlivých kroků zpracování dané úlohy se s tím GPU nedokáže tak dobře vyrovnat. Flexibilní a univerzálnější jádra CPU mají v tomto ohledu výhodu. Pokud jsou ovšem zpracovávaná vlákna a zpracovávané datové pakety podobné, má velké množství paralelních jader GPU velkou výpočetní výhodu. Na obrázku kvůli zjednodušení uvádíme pouze osm jader GPU, moderní GPU však disponují až 240 jádry.

Technická omezení GPU však při tvorbě a optimalizaci softwaru nepředstavují jedi-

ný problém. Největší problém představuje masivní paralelismus jader GPU. Aby bylo možné využít všech 240 jader GPU, je nutné rozdělit program do 240 částí (nebo vláken). Aby bylo možné jednotlivé části kódu zpracovávat paralelně, musí být tyto části na sobě nezávislé. Zpracování jednotlivého procesu nesmí být závislé na výsledku procesu jiného, protože ten je zpracováván současně, a ne před začátkem procesu závislého. Ve výsledku pak nezáleží na tom, který proces doběhne dříve, jelikož výsledek jeho výpočtu nehraje roli při ostatních paralelně zpracovaných výsledcích.

Současné osmijádrové CPU procesory se také potýkají s podobnými problémy, ale při optimalizaci kódu musejí programátoři počítat pouze s osmi jádry, a ne s 240. Mnoho programů má dodnes problémy s využitím osmi jader, málokterý program pak vlastně dokáže plně využít všech možností CPU.

Důvod spočívá v tom, že řadu programů nelze paralelizovat, nebo je to přinejmenším velmi obtížné. Kompilátory, tedy softwarové nástroje, které programátoři využívají k optimalizaci kódu, mají v podstatě za úkol vyhledat části kódu, které mohou být zpracovávány paralelně. Tento automatický mechanismus však často selhává. Pokud vstupní data jedné zpracovávané části záleží na výsledku jiného bloku a pokud se liší pořadí, ve kterém tyto procesy probíhají, automatická optimalizace často nefunguje.

Programátor má možnost sám rozhodnout, zda je paralelní zpracování určitých bloků vůbec možné, ale aby to zjistil, musí se ručně probírat stovkami a tisícovkami řádků programového kódu. Ani manuálně nelze do několika vláken rozdělit každý program. Pokud každý krok zpracování kódu závisí na výsledku kroku předchozího, můžeme na paralelní zpracování zapomenout a běh programového vlákna musí chtě nechtě zůstat sériový. V tom případě je daný kód nevhodný pro zpracování prostřednictvím GPU.

I přes všechny výše zmíněné obtíže se za poslední rok objevilo množství programů, které dokáží alespoň částečně využít výhod zpracování pomocí GPU. Ty nejdůležitější jsme otestovali a prozradíme vám, jak jsou výkonné. Za tímto účelem jsme si pořídili počítač vybavený grafickou kartou nVidia GeForce GTX 295 (2x GPU, celkem 480 výpočetních jednotek, cena grafické karty cca 11 000 Kč), který jsme postavili proti špičkové pracovní stanici HP Z800, vybavené dvěma procesory Intel Xeon DP 5560 (8x 2,8 GHz, celkem 16 jader, cena dvou procesorů cca 45 000 Kč).

Vysoký výkon: Až 20× rychlejší

Lepší využití výkonu grafické karty s sebou přináší masivní zrychlení, které s největší pravděpodobností budete moci využít i na svém počítači. Pokud vlastníte například grafickou kartu řady ATI Radeon HD 3000 či 4000 nebo konkurenční modely nVidia GeForce 8000 a novější, dokáže váš počítač využít všech výše uvedených výhod. Pomocí nástroje Cuda-Z můžete zkontrolovat, zda je vaše grafická karta nVidia kompatibilní s rozhra-

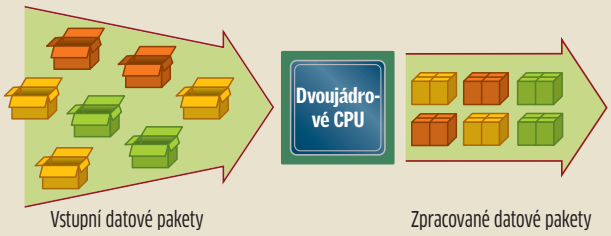
Porovnání CPU a grafického procesoru

Klasické CPU, nebo grafický procesor? Volba záleží na typu zpracovávané úlohy. I přes nižší pracovní frekvenci mohou mít grafické procesory vyšší výkon než CPU, zvláště pak pokud mají za úkol zpracovávat téměř identické datové pakety.

Rozdílné datové pakety: Výhodnější je CPU

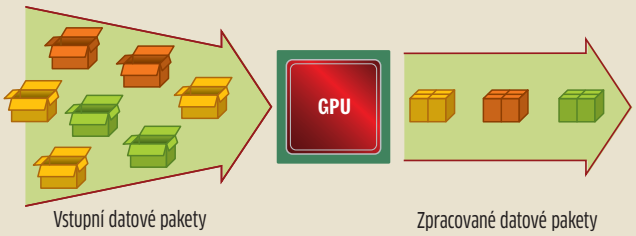
Dvoujádrové CPU

Dvoujádrový procesor CPU dokáže paralelně zpracovávat dvě vlákna s rozdílnými datovými pakety.



Grafický procesor

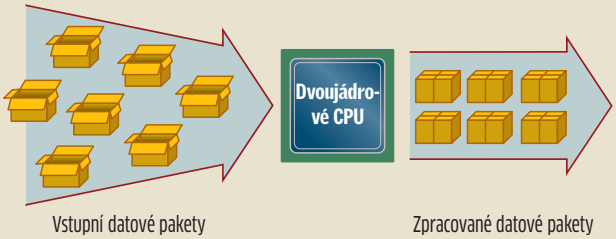
GPU nedokáže paralelně zvládat složité programové procesy a pracuje s nimi individuálně a mnohem pomaleji.



Shodné datové pakety: Výhodnější je GPU

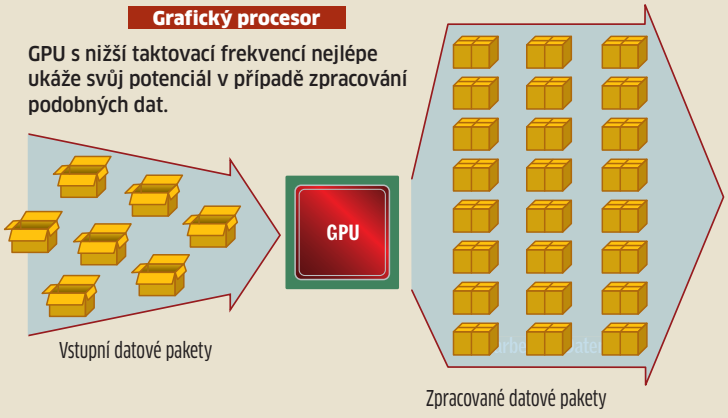
Dvoujádrové CPU

Dvoujádrové CPU zpracovává paralelně i datové pakety dvou shodných úloh.



Grafický procesor

GPU s nižší taktovací frekvencí nejlépe ukáže svůj potenciál v případě zpracování podobných dat.



ním CUDA. Ke stažení jej najdete na stránkách **http://sourceforge.net/projects/cuda-z/**.

Teď už vám schází jen vhodné programy. Zatím neexistuje mnoho freewarových nástrojů, které dokáží k urychlení práce využít výkonový potenciál grafické karty, ale můžeme najít řadu placených aplikací, s jejichž využitím dosáhneme opravdu znatelného zrychlení počítače.

Většina programů, které si dokáží vzít na pomoc grafický procesor, patří do kategorie aplikací určených pro převod a úpravu multimediálních souborů. Právě u videa a obrázků totiž většinou dochází k opakovanému provádění stejných kroků, které jsou aplikovány na velké množství dat, a výsledek jednotlivých kroků nezávisí na tom, jak skončí algoritmus jiného vlákna.

Na využití výpočetní síly GPU se výrazně zaměřila společnost CyberLink, která tuto schopnost využila ve svých nejdůležitějších programech. Výhody pomocného výkonu grafických karet nVidia i AMD podporuje zvláště program PowerDirector od verze 7. Kromě několika filtrů starajících se o barevné převody a zdroje světla používá tento program výkon grafické karty především pro referování filmů ve vysokém rozlišení za pomoci kodeku H.264. Doba převodu je při použití vhodné grafické karty zhruba třetinová. Kodek H.264 spoléhá na pomocný výkon GPU rovněž v případě nástroje CyberLink MediaShow Espresso, který slouží ke snadnému převodu filmů do formátů pro iPod, PlayStation Portable a Xbox. U této utility zabere převod při využití výkonu GPU pouze čtvrtinu času.

Velice výrazného zrychlení lze pomocí GPU dosáhnout i v případě multimediálního přehrávače CyberLink PowerDVD 9, který dokáže převádět DVD filmy s rozlišením 720 × 576 bodů do HD rozlišení. Pokud bychom DVD video převáděli klasicky pouze za pomoci CPU, probíhalo by to tak pomalu, že byste si mohli v klidu prohlédnout každý frame filmu, ale pomocí GPU probíhá převod téměř plynule.

Mezi programy pro zpracování videa zabírá zvláštní roli aplikace Badaboom od firmy Elemental Technology. Tato utilita slouží k převodu filmů do různých formátů a dokáže zapojit výkon GPU tak efektivně, že jeho pomocí je konverze 20krát rychlejší, což v praxi znamená, že převod minutového videa trvá při využití GPU pouze 20 sekund, zatímco pokud aplikace využívá pouze výkonu CPU, tak jí konverze filmu zabere téměř šest minut. Dvacetinásobného zrychlení jsme sice dosáhli za pomoci slabšího procesoru, ale na naší velmi výkonné testovací sestavě jsme i tak zaznamenali desetinásobný nárůst výkonu.

Při úpravě videa dokážeme pomocí přidaného výkonu GPU částečně urychlit práci, ale existují i programy, které díky němu nabízejí úplně nové funkce. Nejlepším příkladem je Adobe Photoshop CS4, ve kterém lze v reálném čase pracovat s obrázky v jejich plném rozlišení. Díky podpoře GPU tak můžete naprosto plynule zvětšovat snímky, na kterých pracujete, a přidávat k nim 3D vrstvy.

Folding@home je vědecká aplikace, která běží na řadě domácích počítačů. Tento program slouží k výzkumu molekul a využívá volného výpočetního výkonu domácích počítačů k tomu, aby se urychlil vývoj boje proti všemožným nemocím. Speciální klient aplikace Folding@home dokáže rovněž zapojit výkon grafické karty, a to dokonce se čtyřicetinásobným zrychlením oproti provozu pouze za pomoci CPU. Tuto aplikaci jsme spouštěli na obou testovaných systémech, a i tak byl výkon akcelerovaný GPU desetkrát vyšší než na sestavě vybavené dvěma osmijádrovými procesory Xeon.

Pokud bychom se na celou věc podívali z hlediska nákladů, pak pro vyrovnání výkonu grafické karty v hodnotě 11 000 Kč bychom museli koupit 20 Xeonů, což by stálo asi 450 000 Kč.

Budoucnost: DirectX využije GPU

Plný výkon grafického procesoru stále není zdaleka využit. Rozhraní CUDA od nVidie zatím ve standardním nastavení podporuje pouze zapojení jedné samostatné grafické karty, takže logicky čekáme na okamžik, kdy bude možné zapřáhnout plný výkon dvou karet zapojených v režimu SLI. Bude to právě rozhraní DirectX, které do budoucna přinese plné zapojení GPU do celkového výpočetního výkonu počítače. Microsoft do svého nejnovějšího API totiž vedle funkcí starajících se o 3D výkon grafické karty zařadil i programovací rozhraní DirectCompute. DirectCompute lze využít k tvorbě programů, které dokáží optimálně využít všech možností nového DirectX 11, ale i staršího a zatím hodně využívaného rozhraní DirectX 10. Unifikované shadery, které bylo již pod DirectX 10 možné použít jako pixel, vertex a geometrické shadery, dostaly nově čtvrtou úlohu jako výpočetní shadery.

Vzhledem k tomu, že rozhraní DirectX se od svého uvedení na trh v roce 1995 stalo nejoblíbenějším rozhraním pro tvorbu her, lze to samé očekávat i od DirectCompute. Jeho základní výhoda spočívá ve faktu, že dokáže bez rozdílu využít grafické karty všech výrobců, a nezáleží tedy na tom, zda počítač používá grafiku nVidia, ATI nebo Intel.

Spolupráce GPU a CPU se výrazně promítne do podoby a rychlosti, kterou budou mít osobní počítače již v blízké budoucnosti. Dlouhou dobu pracuje Intel na projektu Larrabee, který má za cíl vyvinout procesor, jenž bude obsahovat několik všestranných jader, které budou schopné zpracovat jak výpočetní, tak grafické aplikace.

Jak ATI, tak nVidia pracují na přídavných kartách. ATI svůj projekt nazvala FireStream, nVidia používá název Tesla. Tyto karty budou obsahovat grafický procesor, který ale nebude mít za úkol zpracovávat 3D grafiku, a dokonce nemají ani videovýstup. Budou sloužit výhradně jako koprocesory. Hlavním procesorem počítače bude stále CPU, ten ale bude využívat služeb specializovaných koprocesorů, ať již budou součástí přídavných karet, základních desek, nebo dokonce samotných CPU.

AUTOR@CHIP.CZ

INFO

Optimalizované nástroje

Dodatečný výkon grafického procesoru využívají hlavně programy pro práci s multimediálními soubory, ale je ideální i pro vědecké výpočty. Přinášíme seznam nejdůležitějších nástrojů a programů, které zvládají využít výpočetního výkonu grafického procesoru.

VIDEO A VĚDECKÉ APLIKACE

Adobe Photoshop CS4 (více funkcí) Snímání a zvětšování snímků

ArcSoft Total Media Theatre (real time) Převod videa snímaného ve standardním rozlišení do rozlišení HD

CyberLink PowerDVD 9 (real time) Převod videosouborů z rozlišení DVD do HD

Folding@home (40× rychlejší) Celosvětový projekt, který má za úkol molekulární výzkum vývoje nemocí

Loilo Super LoiLoScope (10× rychlejší) Software pro úpravu videa, převod do kodeku H.264

Motion DSP vReveal (5× rychlejší) Automatický převod filmů (pracuje pouze s kartami nVidia)

TMPGEnc 4.0 Xpress (4,5× rychlejší) Videofilm akcelerovaný prostřednictvím GPU

PŘEVOD VIDEA

CyberLink MediaShow Espresso (4× rychlejší) Převod videa prostřednictvím kodeku H.26

CyberLink PowerDirector 7/8 Ultra (3× rychlejší) Převod videa prostřednictvím H.264 a několik realtimových filtrů

Elemental Technologies Badaboom (10× rychlejší) Převod videa prostřednictvím kodeku H.264 (spolupracuje pouze s grafickými čipy nVidia)

MediaCoder (5× rychlejší) Převod videa prostřednictvím kodeku H.264

Nero Move it 1.5 (4× rychlejší) Převod videa prostřednictvím kodeku H.264

Roxio WinOnCD 2010 (5× rychlejší) Převod videa prostřednictvím GPU