GPGPU: 200x rychleji než CPU

S využitím výkonu grafických karet je možné video, obrázky a programy zpracovat mnohem rychleji než kterýmkoliv procesorem – díky GPGPU. *Thomas Littschwager, autor@chip.cz*

Lidské jednání je – alespoň z velké části – sekvenční. Znamená to, že člověk celkově sleduje jeden cíl za druhým a mezi ně neustále vřazuje menší, důležité úkoly. A podobnou cestou se ubíral vývoj hlavních procesorů pro počítačové systémy. CPU postupně zpracovává množství nejrůznějších úloh, je velmi flexibilní a přizpůsobivá.

Díky pokrokům v průmyslu polovodičů se dnes v procesorech o závratné pracovní frekvenci a s ohromujícím výkonem starají až čtyři jádra. Přesto však průběh úloh zůstává sériový, ačkoliv by při čtyřech jádrech mohl být odpovídající počet úloh zpracován paralelně.

Pro většinu výpočtů je pochopitelně základem sériový postup, neboť nejprve musí být k dispozici výsledky jednoho kroku, aby bylo možno zahájit další. Za jistých okolností ovšem zase mnohé jiné výpočty na sobě nijak nezávisí a bylo by je možné bez problémů nechat proběhnout současně, tedy paralelně. To je případ například dotazů do databanky, vědeckých simulací nebo také určitých prvků překódování videa. Pro takové úlohy už současné procesory nejsou úplně vhodné.

Optimální náhrada však přitom vězí prakticky v každém počítači: grafická karta. Její procesor, GPU, byl vlastně vyvinut výhradně pro paralelní výpočty velkého množství obrazových bodů – přístup k nim je však možný jenom prostřednictvím jednoho z grafických programových rozhraní (API, Application Programming Interface) DirectX a OpenGL. Důmyslná koncepce navržená Stanfordovou univerzitou však nyní umožňuje pomocí speciálně vyvinutých programů využít výkon grafického procesoru nejen pro jeho základní poslání. Řeč je o GPGPU (General-Purpose Computing on Graphics Processing Units, tj. víceúčelové nasazení grafických procesorů).

Programy založené na GPGPU slibují úžasné zvýšení výkonu: videa lze překódovávat až dvacetkrát rychleji a plug-iny do Photoshopu počítají znatelně svižněji než při použití normálních CPU. Přinášíme vám teď přehled technických základů a možností, které takové „přeškolení“ GPU přináší.

Potenciál GPU: 3D karta jako lepší procesor

Technické předpoklady pro GPGPU přineslo zavedení grafických karet pro DirectX 8 – toto rozhraní totiž vyžaduje volně programovatelné shadery (výpočetní jednotky uvnitř GPU). Pak je totiž možné – přinejmenším teoreticky – nechat pixelovými shadery zpracovat libovolný programový kód. Teoretická výkonnost je pozoruhodná: vezmeme-li v úvahu výkonnostní pokroky výpočtů v pohyblivé čárce u CPU ve srovnání s GPU, uvidíme, že první s druhou zdaleka nedokáže držet krok. Jestliže počet GFlops/s (miliard operací v pohyblivé čárce za sekundu) u CPU v posledních čtyřech letech vzrostl z 5 GFlops/s (Pentium 4) na 20 GFlops/s (řada Core 2 Duo), grafické karty se v tomto směru za stejnou dobu zlepšily z 15 GFlops/s (ATI Radeon 9700) na neuvěřitelných 520 GFlops/s (řada nVidia GeForce 8800).

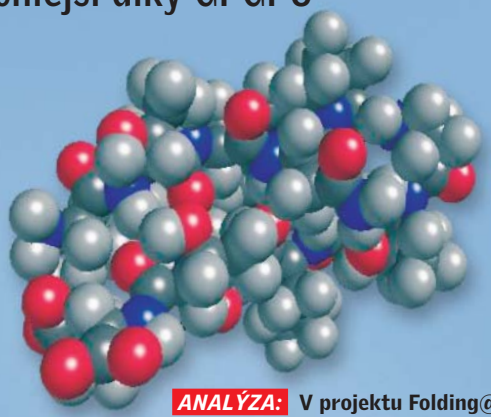
Za tímto nárůstem rychlosti stojí stále se zvyšující počet shaderů uvnitř grafického čipu, které všechny pracují paralelně. Z tohoto souběhu pak profituje vhodný software: může-li program současně využívat každý jednotlivý shader, je možné – třeba jako u ATI Radeon X1950 XTX – současně zpracovávat 48 datových proudů. Rychlé Quad Core procesory od Intelu přitom zvládnou čtyři, jinak jsou dnes normální dva.

Čím se liší GPU a CPU

Při těchto výsledcích se samozřejmě vnucuje otázka, jsou-li už tedy klasické procesory zastaralé. Budeme napříště potřebovat jen výkonné grafické čipy? Ne tak docela. Výhoda GPU oproti CPU je totiž víceméně omezená: nejen že datové pakety zpracovávané GPU musí být na sobě nezávislé, ale →

Výzkum nemocí 20krát rychlejší díky GPGPU

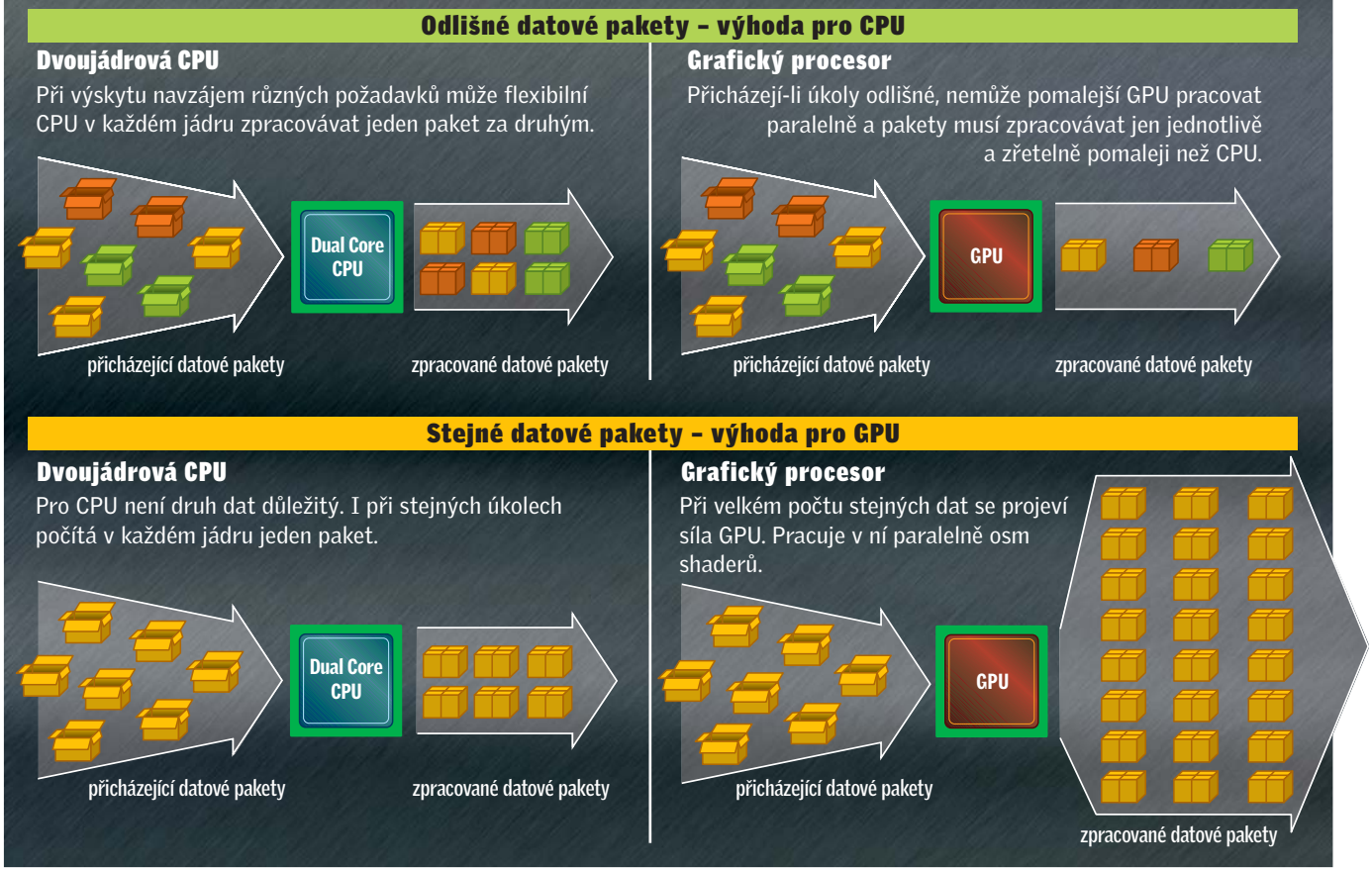
Jako jednu z prvních aplikací nového konceptu pro uživatele portovala Stanfordova univerzita pro GPGPU (momentálně jen pro GPU řady ATI X1000) program Folding@home (F@h). Aplikace je určena k výpočtům sbalování (folding) proteinů, při nichž se využívá výpočetní výkon všech zúčastněných uživatelů podobně jako u Seti@home. Výsledky distribuovaných výpočtů slouží při výzkumu příčin onemocnění, jako jsou Alzheimerova nebo Parkinsonova choroba. Nasazení GPU propůjčuje projektu mimořádné výhody: například Radeon X1950 zde počítá až 20krát rychleji než Core 2 Duo E6600.



ANALÝZA: V projektu Folding@home se propočítává sbalování bílkovin – buď s využitím CPU, nebo GPU.

Srovnání hlavního procesoru a grafického čipu

Podle druhu výpočtu může být výhodnější použít buď klasický procesor (CPU), nebo grafický čip. GPU vítězí, vzdor své výrazně nižší frekvenci, nad flexibilními CPU především při řešení mnoha téměř identických úloh.



navíc musí být výpočetní operace potřebné k jejich zpracování prakticky identické. Grafické procesory jsou totiž poměrně úzce specializovány jen na určité operace, zatímco CPU jsou schopny v každém taktu a v každém jádře počítat s naprosto odlišnými daty a podle úplně odlišných programů. Obecně lze tedy říci, že grafický procesor je vhodnější na úkoly týkající se hlavně manipulace s daty, hlavní procesor je zase lepší na čisté výpočty a úlohy s převažujícími logickými operacemi.

Vývoj: V čele nVidia a AMD

Důvodem, proč se GPGPU vzdor mohutnému potenciálu při vhodných úlohách stále ještě neprosadilo, je především obtížné programování potřebného softwaru: až do konce roku 2006 vůbec neexistovalo odpovídající vývojové prostředí (SDK, Software Development Kit), v němž by bylo možno GPGPU software vytvářet pomocí známých programovacích jazyků, jako je Java, C nebo C++. Programovat se proto muselo prostřednictvím obou grafických programových rozhraní, tedy DirectX a OpenGL, což ovšem podle okolností vyžaduje přizpůsobení nejen při implementaci, ale především také ve struktuře programu. Grafické API totiž některé výpočetní operace vůbec neumožňuje, například určení minima a maxima v seznamu položek. Takto omezená nabídka příkazů znemožnila zejména snadný a nekomplikovaný převod stávajících programů do podoby schopné využít GPGPU. Vlastnosti různých GPU jsou navíc zpravidla natolik odlišné, že nástroj vyvinutý pro jeden grafický čip většinou neběží na konkurenčním produktu – což ovšem celou věc pro masový trh zcela diskvalifikuje.

Je však nasnadě, že právě velcí výrobci grafických čipů AMD a nVidia vidí v GPGPU nadějnou budoucnost – a hlavně mohutný finanční potenciál. Již více než před rokem oznámila společnost ATI, dnešní „dcera“ AMD, zahájení vývoje GPGPU videotranskodéru, který má být 20krát rychlejší než každá CPU. Dnes už jak AMD, tak nVidia disponují vlastním SDK na bázi GPGPU, určeným programátorům softwaru: „Close to the Metal“ (CTM) pochází od AMD, „Cuda“ od nVidie.

Obě vývojová prostředí pracují s kompilátorem oblíbeného programovacího jazyka „C“, který program překládá tak, aby s ním mohla pracovat GPU. Kromě toho společnost AMD v září 2006 představila kartu s GPU na bázi FireGL (viz obrázek), která je určena výhradně pro GPGPU nástroje – první servery s klastrem GPU se již používají v praxi.

Dvěstěkrát rychleji: GPGPU slibuje mimořádný pokrok

Dnešní kombinace hardwaru a softwaru konečně programátorům otevírá dveře k potenciálu grafické karty, aniž by při vývoji museli kompletně měnit své návyky. Mimo oblast vědeckých zařízení pro GPGPU je prozatím k dispozici jen velmi málo aplikací, ale už u onoho mála existujících nástrojů se ukazuje, jaké rychlostní výhody lze v budoucnu očekávat. Momentálně nejznámějšími aplikacemi, které využívají GPGPU, jsou výpočty fyzikálních efektů a také Folding@home (viz rámeček str. 28). Při simulaci burzovních kurzů dokázala nVidia podle vlastních údajů urychlit výpočty až 200krát.

Bude ovšem trvat ještě měsíce, ne-li roky, než bude z GPGPU profitovat také uživatel běžných programů. V plánu jsou antivirové nástroje nezatěžující CPU nebo webové servery zvládající vysoký výskyt databázových dotazů. Každopádně byl ale v podobě CTM a Cuda položen základní kámen ke snadnějšímu programování.

ODKAZY

www.gpgpu.org
www.nvidia.com/cuda
www.ati.com/developer
http://folding.stanford.edu