# Jak pracují špičkové grafické karty

**NOVÉ GRAFICKÉ ČIPY** od ATI a nVidie posouvají standardy 3D akcelerace. Chip vám přináší podrobný pohled na design grafických čipů a odhaluje rozdíly mezi nimi.

THOMAS LITTSCHWAGER

Grafické karty toho dnes zvládnou mnohem víc než jen zobrazovat počítačové hry – GPU mají na starosti plynulé přehrávání videa ve vysokém rozlišení, umí rychle počítat složité matematické kalkulace a lze s nimi v reálném čase vytvářet realistické 3D scény. Výkon současných generací grafických karet od nVidie i ATI, potažmo AMD, je ohromující. Nový GPU od firmy nVidia nese název GT200 a můžeme jej najít v kartách s modelovým označením GeForce GTX260 a GeForce GTX280. V případě ATI jsou nejvýkonnější grafické čipy pojmenovány RV770 a můžeme se s nimi setkat v kartách Radeon HD4850 a Radeon HD4870. Měření, která dokazují, jak moc jsou v porovnání s ostatními GPU nové čipy výkonné, naleznete v našem pravidelném přehledu CPU a GPU.

V tomto článku se od základu podíváme na architekturu grafických čipů, vysvětlíme, co se skrývá za názvy shader, ROP či texturovací jednotka, a ukážeme, při jakých výpočtech dokáže grafická karta deklasovat jakýkoliv CPU v osobním počítači.

## Základy: Struktura grafické karty

Moderní grafická karta se v zásadě skládá z pěti komponent: systémového rozhraní, grafické paměti, grafického procesoru (GPU), frame bufferu a RAMDAC (Random Access Memory Digital/Analog Converter) převodníku. Systémové rozhraní je umístěno nejblíže k základní desce počítače a v současné době se nejčastěji jedná o PCI Express. Data přicházející z počítače jsou ukládána do grafické paměti, která slouží i jako paměť pro ukládání objektů a textur a která má v současnosti většinou kapacitu 256 až 1 024 MB. V paměti uložená data následně putují do grafického procesoru (GPU), který vypočítá všechny pozice, po-

## INFO

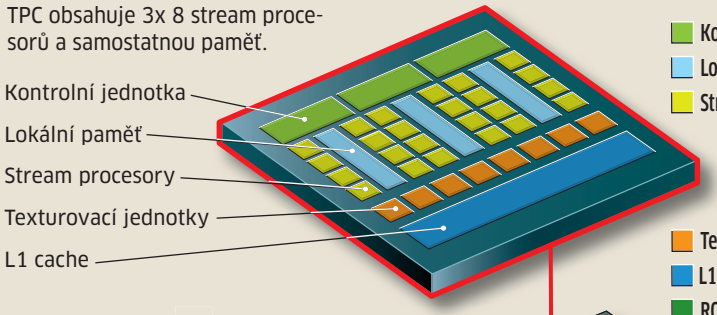
### Grafický procesor nVidia GT200

Čip GT200 je v současnosti vlajkovou lodí společnosti nVidia. nVidia chce samozřejmě disponovat nejrychlejším grafickým procesorem, kterým právě GT200 momentálně je. Setkáme se s ním v nejvýkonnějších high-endových kartách nVidia.

Nové generace GPU od ATI i nVidie se skládají z množství komponent. Uvnitř grafického čipu najdeme hlavně unifikované shadery (stream procesory, SP). GPU GT200 od nVidie obsahuje celkem 240 streamovacích procesorů. Celkový počet je rozdělen do bloků po osmi SP a tři takovéto bloky jsou klastrovány do výsledného Thread Processing Clusteru. Čip s touto architekturou nemusí být použit pouze pro grafické výpočty, ale i pro paralelní výpočty.

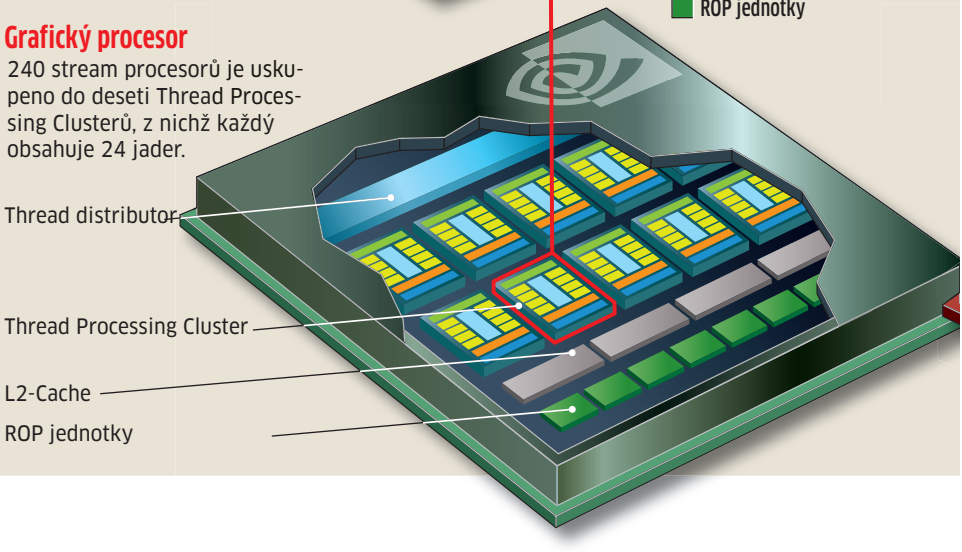
**Thread Processing Cluster (TPC)**

TPC obsahuje 3x 8 stream procesorů a samostatnou paměť.




**Grafický procesor**

240 stream procesorů je uskupeno do deseti Thread Processing Clusterů, z nichž každý obsahuje 24 jader.




ILUSTRACE: HARALD FUCHSLOCH

hyby a rozhraní objektů 3D scény a vytvoří z nich obraz.

Hotový obraz je dále předán do frame bufferu, odkud putuje do RAMDAC převodníku. Ten převádí všechny digitální informace do analogové podoby, v níž je zpracuje analogový VGA monitor. RAMDAC rovněž ovládá digitální výstupy DVI, HDMI či DisplayPort.

## Grafická pipeline: Dálnice pro obraz

Většina částí grafické karty nehraje tak důležitou roli, opravdu důležitý je pouze základ, tedy grafický čip GPU. Aby z hrubých dat, která do grafiky směřují z PC, vyšel na výstupu obraz, je třeba provést řadu složitých operací.

Grafická pipeline, tedy cesta, kudy po grafické kartě putují data, je téměř identická u všech současných modelů. Celý průběh dat od načtení do grafické paměti přes zpracování až po výstup z frame bufferu je třeba pro každý obrázek opakovat. Pro iluzi plynulého a netrhaného obrazu potřebujeme, aby grafika vyprodukovala minimálně 25 obrázků za sekundu. V případě graficky náročných her s řadou efektů musí karta pro dosažení opravdu realistického obrazu vyprodukovat cca 60 snímků za sekundu. GPU má tedy opravdu dost práce.

První zastávkou na pouti po pipeline poté, co jsou data z rozhraní nahrána do GPU, je přípravný procesor (Setup Engine nebo Input Assembler), který předkalkuluje a převádí data. Rozliší data podle typu, určí, zda představují vektory, obraz či kód programu, a zpracuje hrubá data tak, aby se dostala do správného výpočetního modulu. Zde se rozhoduje, zda budou data poslána do vertex shaderu, geometrického shaderu, pixel shaderu nebo jednotky pro zpracování textur.

Každý trojrozměrný obraz se skládá z množství trojúhelníků (viz ilustrace Grafická pipeline). Do vertex shaderů (vertexy jsou vrcholy neboli rohové body polygonů) směřují souřadnice, které ve 3D modelu budou značit rohové body trojúhelníků, jejich zarovnání, měřítko nebo i zkreslení podle úhlu pohledu virtuálního pozorovatele. Předpokládaná viditelná plocha se nazývá „frustum“, což je vlastně komolý jehlan určující pohledový objem. Když je scéna hotová, proběhne kontrola, zda bude objekt vůbec zobrazen (tedy je-li z daného úhlu vidět), zda je správně umístěn do frusta a zda není úplně nebo částečně zakryt jinými objekty.

Neviditelné části obrazu jsou ze scény odstraněny, aby se ušetřil výpočetní výkon grafického procesoru, potažmo aby se zrychlil výpočet viditelných komponent. Tomuto procesu se říká „frustum culling“, pohledové ořezávání či „odstřel podle tělesa záběru“. Pokud se objekt nachází příliš daleko (tedy za hranicí viditelnosti), příliš blízko či za virtuálním pozorovatelem, proběhne obdobný proces, který se nazývá „clipping“ neboli ořez.

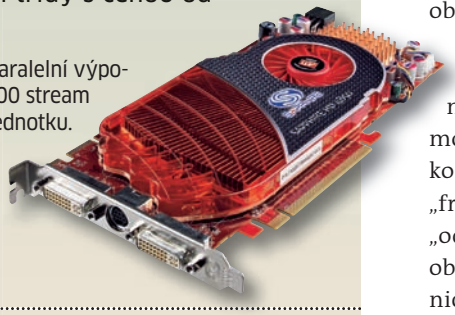
Poslední úlohou vertex shaderu je nasvícení scény. Při tomto procesu je 3D scéna osvětlena světelnými zdroji umístěnými někde v místnosti. Bez tohoto kroku by 3D scéna byla naprosto tmavá. Vertex shader dokáže objekty pouze manipulovat, ale neumí vytvářet nové geometrické prvky, jako jsou body, linky či trojúhelníky. Za tímto účelem byl od listopadu 2006 (jako součást DirectX 10) vytvořen geometrický shader, který dokáže vytvářet nové geometrické formy, jako jsou například rostoucí stromy. Po vytvoření 3D scény je tedy zapnut geometrický shader.

Jakmile dostane požadovaný obraz podobu mřížkového modelu osvětleného světelnými zdroji, je třeba vytvořit něco jako fotografii scény, která má podobu dvojrozměrného obrazu, který by se měl výsledně zobrazit na monitoru. Tomuto procesu se říká screening nebo rendering. Všechny body trojrozměrného objektu, který byl zatím uložen ve →

## Grafický procesor ATI RV770

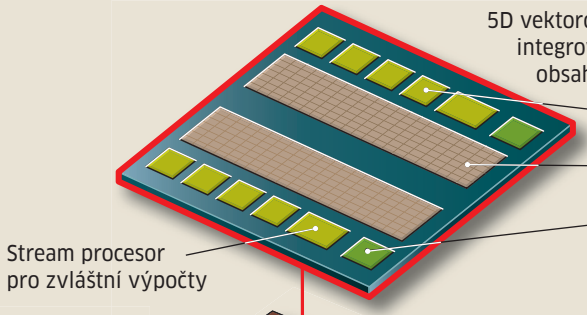
Firma AMD má s procesorem RV770 velké plány. Po několikaleté bitvě má k dispozici procesor, který přiláká od nVidie množství zákazníků požadujících grafickou kartu střední třídy s cenou od 3 500 do 7 000 Kč.

I společnost ATI pracuje na využití grafických karet pro paralelní výpočty. S tímto záměrem vybavila procesor RV770 celkem 800 stream procesory. Po pěticích tvoří tyto SP tzv. 5D vektorovou jednotku. 5D vektorové jednotky jsou poskládány po dvojicích. 16 takovýchto dvojic pak tvoří jedno SIMD jádro. Díky této architektuře dosáhne procesor ATI RV770 výkonu 1 200 GFLOPS, což je o poznání více než 933 GFLOPS naměřených u procesoru nVidia GT200.



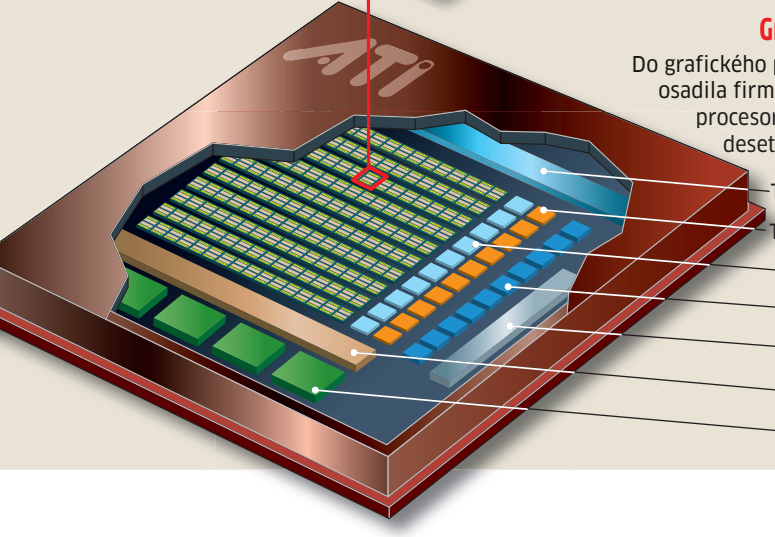
### 5D vektorová jednotka

5D vektorová jednotka obsahuje dvojici integrovaných vektorových jednotek obsahujících pět stream procesorů.

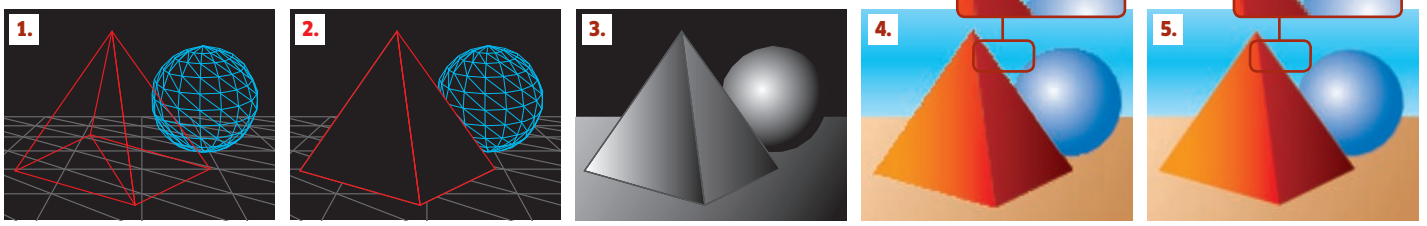



### Grafický procesor

Do grafického procesoru RV770 osadila firma ATI 800 stream procesorů sdružených po deseti do 80 jednotek.



## GRAFICKÁ PIPELINE: JAK SE TVOŘÍ Z ČÍSEL OBRAZ




1. Vertex shader vytvoří ze souřadnic a vektorů 3D scénu.
2. Neviditelné prvky jsou odstraněny ze scény.
3. Scéna je nasvícena jedním nebo více světelnými zdroji.
4. Pixel shader a texturovací jednotky obarví jednotlivé objekty scény.
5. Anti-aliasingem jsou vyrovnány hrubé a ostré hrany objektů.

vektorové podobě, jsou převedeny na pixely. Nyní bude následovat výpočetně nejnáročnější krok – stínování neboli shading v jednotce zvané pixel shader. Ten vypočítává barvu, a je-li to potřebné, i další atributy, jako je průhlednost, odrazivost či struktura každého pixelu.

Z toho se odvíjí barevnost 3D objektů. Jednotlivé kroky shaderů si můžete názorně prohlédnout v informační grafice na této straně.

V podstatě je nyní obraz hotový a zbývají jen finální úpravy – například aplikace různých filtrů pro zvýšení realističnosti scény. Na 3D objekt jsou naneseny textury (tedy kompletní bitmapy či obrázky). Tomuto kroku se říká texture mapping. Tímto způsobem lze snadno vytvořit fotorealistické obrázky, které však postrádají flexibilitu opravdových 3D objektů. Proto například stromy vytvořené pomocí textur vypadají skvěle při pohledu zepředu, ale při pohledu z boku jsou naprosto ploché.

Anizotropické filtrování, k němuž dochází rovněž v texturovacích jednotkách, má na starosti perspektivní zkreslování textur. Po této transformaci se objekty zdají ostré i při pohledu z dálky.

Po dokončení mapování textur se snímek posune do ROP jednotky (Raster Operation Processor, u ATI nazýván Element Render Back-End). Rasterizace obrazu má za následek zhrubnutí jeho hran. Ty jsou převedeny do pixelů a objeví se na nich jakoby „schodový“ efekt. Tento efekt může odstranit anti-aliasing, který podobné problémové hrany vyhlazuje (viz grafika úplně vpravo nahoře).

Hotový obrázek je nyní uložen ve frame bufferu, kde pipeline grafické karty končí. V další části tohoto článku vám ukážeme, čím se technicky odlišují grafické karty obou nejvýznamnějších výrobců.

### EVGA GEFORCE GTX 280

| TECHNICKÁ DATA | |
|---|---|
| GPU | nVidia GT-200 |
| Počet tranzistorů | cca 1,4 miliardy |
| Frekvence čipu | 602 MHz |
| Frekvence shaderů | 1 296 MHz |
| Paměť | 1 024 MB GDDR3 |
| Frekvence paměti | 1 107 MHz |
| Výpočetní výkon | cca 933 GFLOPS |
| **DATA A MĚŘENÍ** | |
| 3DMark Vantage | 10 577 bodů |
| Crysis | 44 fps |
| Half Life 2 Lost Coast | 202,4 fps |

### MSI R 4870 T2D 512

| TECHNICKÁ DATA | |
|---|---|
| GPU | ATI RV770 |
| Počet tranzistorů | cca 965 milionů |
| Frekvence čipu | 750 MHz |
| Frekvence shaderů | 750 MHz |
| Paměť | 512 MB GDDR5 |
| Frekvence paměti | 1 800 MHz |
| Výpočetní výkon | cca 1 200 GFLOPS |
| **DATA A MĚŘENÍ** | |
| 3DMark Vantage | 9 085 bodů |
| Crysis | 39 fps |
| Half Life 2 Lost Coast | 172,9 fps |

## GPU: Srdce grafické karty

Architektura nových GPU je maximálním způsobem určena rozhraním DirectX 10 (sada programovacích rozhraní pro tvorbu her pod Windows Vista). Design jednotlivých čipů se zatím (jak jsme právě popsali) dělil na aritmetické logické jednotky ALU (Aritmetic Logical Unit), které pracují jako pixel shadery nebo vertex shadery.

Rozhraní DirectX 10 přineslo nový model práce se shadery – tzv. unifikované shadery. V praxi to vypadá tak, že v rámci GPU mohou všechny ALU podle momentálních požadavků pracovat jako pixel, vertex nebo geometrické shadery. ATI stejně jako nVidia dnes vyrábí téměř výhradně GPU určené pro rozhraní DirectX 10. S novými čipy lze dosáhnout mnohem vyšší efektivity unifikovaných shaderů. Ke správnému roztřídění hrubých dat do volných ALU slouží tzv. Thread Scheduler neboli plánovač výpočetních vláken. Ten analyzuje a paralelizuje datové toky zpracované Setup Enginem a přiřazuje potřebné úkoly volným ALU.

Rozdíl mezi grafickými procesory ATI a nVidia je v podstatě dán jejich vnitřní strukturou, a to hlavně co se týče architektury unifikovaných shaderů, kterým se rovněž říká stream procesory (SP). nVidia jich na čip umístila 240 a firma ATI jich použila 800. Hlavní rozdíl je ale v tom, co jednotlivé stream procesory zvládnou udělat. SP nVidie jsou jednorozměrné, lze je rovněž využít v každém cyklu jako skalární jednotky (pro výpočet jedné z komponentních hodnot – červená, zelená, modrá či alfa) a pracují v každém cyklu zároveň s MADD instrukcemi (Multiply Add, násobení a sčítání), stejně jako s MUL instrukcemi (Multiply, násobení), určenými pro další výpočty.

Stream procesory v čipech ATI však dokáží v každém kroku provést jen jednu MADD operaci. Pokud bychom chtěli do-

sáhnout srovnatelného výkonu se stream procesory nVidie, museli bychom uvažovat o přiřazení pěti SP procesorů pro každou 5D vektorovou jednotku.

Oba oponenty neodlišují pouze schopnosti samostatných stream procesorů, ale liší se i uskupením a rozložením jednotlivých částí čipu (viz infografika na str. 64-5). ATI integruje do jednoho shaderového jádra pět SP procesorů, což v případě 800 SP procesorů představuje 160 těchto jader. Ze 16 shaderových jader se u modelové řady ATI Radeon HD4800 skládá jedno SIMD jádro. SIMD je zkratka z anglického názvu Single Instruction, Multiple Data. Jedno SIMD jádro dokáže vykonat stejnou aritmetickou operaci simultánně na větším množství dat, ale nedokáže provádět odlišné úkoly. Každému SIMD jádru je přiřazeno 16 KB místní paměti pro rychlý přenos dat mezi jednotlivými stream procesory. Součástí každého jádra jsou dále texturové klastry skládající se ze čtyř texturových procesorů, dekompresní a adresační jednotky, sampleru a filtrovací jednotky a L1 cache paměti pro okamžité uložení textur. Jednotlivá jádra mezi sebou komunikují prostřednictvím sběrnice Data Request Bus, jež zabírá 16 KB celkové cache paměti. Navíc jsou zde čtyři bloky L2 cache, které jsou přímo spojeny s hlavní pamětí a vyměňují si data se SIMD jádry prostřednictvím crossbar. Procesor UTDP (Ultra-Threaded Dispatch Processor) rozděluje datové toky zpracované Setup Enginem podle náležitosti do vertexových, pixelových a geometrických programů a tak zajišťuje optimální výkonnost aritmetických jednotek.

nVidia rozděluje stream procesory v kartách řad GT-200 odlišným způsobem. ALU jsou seskupeny po osmi do tzv. streamovacích multiprocesorů (SM), které disponují stejně jako v případě ATI 16 KB cache. Z celkových 30 SM tvoří tři tzv. Texture Processing Cluster, tedy část dedikovanou pro zpracování textur, k níž navíc patří osm texturovacích jednotek a L1 cache. Tyto klastry tedy mezi sebou spolupracují jako MIMD (Multiple Instructions, Multiple Data). Architektura uvnitř klastru je v případě nVidie popisována jako SIMT (Single Instruction, Multiple Thread) a je to v podstatě modifikace SIMD, kterou používá ATI.

Zatímco ATI si získala reputaci ohledně počtu stream procesorů, nVidia vítězí množstvím ROP procesorů (Raster Operation Processors). GT200 obsahuje 32 ROP jednotek, zatímco GPU ATI RV770 jich nabízí pouze 16. nVidia má náskok i co se týče rozhraní paměti a kapacity integrované hlavní paměti. 1 024 MB paměti čipu GT200 je spojeno s GPU prostřednictvím 512bitového rozhraní, zatímco ATI využívá spojení široké jen 256 bitů a celková kapacita videopaměti dosahuje pouze 512 MB. ATI využívá ve svých kartách Radeon 4870 paměť typu GDDR5, která má prakticky dvojnásobnou propustnost (4 Gb/s namísto 2 Gb/s jako u GDDR3) a pouze čtvrtinovou spotřebu v porovnání s GDDR3, které se drží nVidia.

## GPU: Vyšší výpočetní výkon než CPU

Oblast 3D grafiky není jedinou věcí, nad kterou oba konkurenti přemýšlí. Princip ALU jednotek spolupracujících v rámci GPU vede k silnému paralelnímu výpočetnímu výkonu, což je velký úspěch. Grafické karty dokáží v určitých případech (například v případech simulací či finančních analýz) počítat až 150x rychleji než CPU. Klíčem k úspěchu jsou zde flexibilní, nezávisle programovatelné shadery moderních GPU, které z nich činí „General Purpose GPU“ (GP GPU, víceúčelové GPU).

Rozdíl ve výkonu je v tomto případě úžasný. GT200 dosahuje výkonu až 933 GFLOPS (FLOPS je zkratka z anglického Floating Point Operation Per Second, tedy výkon s operacemi v plovoucí čárce), RV770 však dosahuje výkonu téměř 1 200 GFLOPS. Pro srovnání – čtyřjádrový procesor Intel Core 2 Quad Q6600 ve stejném testu vykazuje 21,4 GFLOPS, takže zdaleka nedosahuje výkonu GPU. Všechny programy však nejdou paralelizovat, takže CPU je v případě každodenních aplikací lepší volbou. Pravda ale je, že v případě specializovaných aplikací, jakými jsou například simulace, je GPU mnohem rychlejší než CPU.

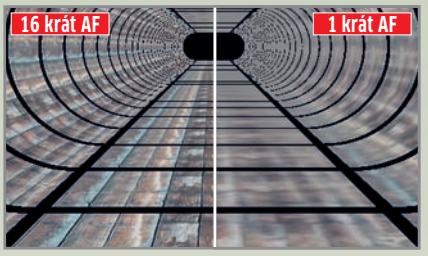
Implementace takových programů byla doposud celkem složitá. nVidia zajišťuje programovací prostředí CUDA, které pracuje s jazyky C a C++. Díky němu lze pro GPU vyvíjet nástroje umožňující mnohem lepší paralelnost zpracování dat. ATI pracuje s podobným projektem CTM (Close to the Metal), který však nenabízí takové ovládací pohodlí jako rozhraní C++.

Největší nevýhodou GPU v porovnání s CPU je to, že hodnoty s plovoucí čárkou lze počítat pouze 32bitově. Složité kalkulace velkého množství dat vyžadují vyšší přesnost a 64bitové zpracování. ATI i nVidia reagují na tyto požadavky úpravou současných GPU, která jim zajistí dvojnásobnou přesnost při výpočtech s plovoucí čárkou. Vypadá to, že výrobci klasických CPU budou mít možná dost brzy nepříjemnou konkurenci.

AUTOR@CHIP.CZ

## SLOVNÍK

**Anizotropické filtrování** Anizotropické filtrování zajišťuje, aby se i vzdálenější textury zdály ostré. V podstatě se sníží jejich rozlišení, takže vzdálené textury nevyžadují tolik výpočetního výkonu.




**Anti-aliasing** Převodem souřadnic a vektorů na pixely vznikají ostré a jakoby schodovité hrany, které anti-aliasing rozpozná a vyhlazuje.

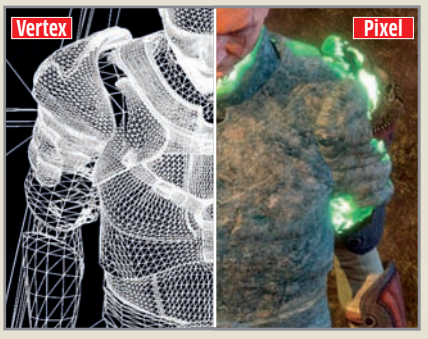
**Frame buffer** Frame buffer ukládá během tvorby výsledné obrázky. Abychom neviděli obraz před dokončením, používají se dva nebo i tři frame buffery (dvojité/trojité bufferování).

**RAMDAC** Převodník RAMDAC získává obrázky z frame bufferu a zpracovává je pro analogový či digitální výstup. Z RAMDAC převodníku směřuje obraz přímo na výstupní konektory grafické karty.

**Raster Operation Processor (ROP)** Uvnitř ROP procesorů jsou umístěny filtry pro anti-aliasing a anizotropické filtrování.

**Pixel shader** Při průchodu pixel shaderem získává každý pixel informace o barvě.

**Geometrický a vertex shader** Ve vertex shaderu vzniká mřížkový model 3D objektu, geometrický shader přidává další objekty.




**Textury** Pro zvýšení realističnosti scény jsou rozhraní 3D objektů pokryta obrázky (texturami).

**Z-buffer** Ukládá informace o hloubce objektu a k němu náležícího pixelu a poté rozhodne, zda je, či není daný pixel skrytý a zda bude, nebo nebude zobrazen.