



Grafické a multimediální procesory (GMU)
Příprava na zkoušku

Obsah

1	Popište základní principy návrhu energeticky úsporného procesoru CPU/GPU, jak se vyhodnocuje energetická úspornost.	4
2	Popište vývoj 2D/3D grafického řetězce s různým stupněm zřetězení.	4
3	Popište 1- až n-rozměrné multiprocesorové propojovací struktury (Origin/Onyx) a typy propojovacích sběrnic Silicon Graphics.	5
4	Popište principy výstavby, činnosti a použití GPGPU.	5
5	Popište koncepci a vlastnosti různých typů Streaming Multiprocessoru	6
6	Popište principy komprese dat v systému Pascal (delta-komprese a další).	6
7	Popište architekturu grafických multiprocesorů a nové principy činnosti, jako komprese dat, preempce a její typy.	6
8	Popište princip tensorového jádra Turing, k čemu slouží, jaké formáty dat se používají.	7
9	Popište principy a funkce preempce Pascal.	8
10	Popište vývoj unifikovaného adresového prostoru, principy činnosti, a jeho hardwarovou podporu.	8
11	Popište principy návrhu energeticky úsporného GPU pro mobilní zařízení.	8
12	[TODO] Popište hlavní vývojové etapy a principy grafických systémů Mali pro mobilní zařízení.	9
13	Vysvětlete a zdůvodněte koncepci kachlíček (tiles) 16x16 pixelů v grafice Mali.	9
14	[TODO] Popište formáty dat podporované v grafice Mali, skalární i vektorové.	9
15	Popište a zdůvodněte principy tří základních typů komprese textur - ztrátové, bezztrátové a adaptivní.	9
16	Popište alespoň 3 rozdíly mezi architekturami GPU a CPU.	9
17	Co je to kernel?	10
18	Co je to vlákno?	10
19	Co je to divergence vláken a kdy vzniká?	10
20	Co je to warp/wavefront?	10
21	Co je to multiprocesor (streaming multiprocessor/compute unit) GPU a k čemu slouží.	10
22	Co je to fronta příkazů (command queue)?	11

23 Jakým způsobem je možné synchronizovat úlohy ve frontě příkazů s vykonáváním mimo pořadí (out-of-order execution) nebo mezi frontami v OpenCL?	11
24 Jakým způsobem se eliminuje/zmírňuje dopad aritmetických/paměťových latencí.	11
25 Co je to konflikt banků v lokální paměti a kdy vzniká?	11
26 Co je to zarovnaný přístup do paměti?	11
27 Popište konstantní/uniformní paměť.	11
28 Popište lokální/sdílenou paměť.	11
29 K čemu slouží texturovací jednotky?	12
30 [TODO] Jaké jsou rozdíly mezi bufferem a texturou v OpenCL?	12
31 Jaký je rozdíl mezi normalizovanými a nenormalizovanými texturovacími koordinátami?	12
32 Co je to register spilling?	12
33 Jaké jsou způsoby komunikace mezi vlákny ve skupině a mezi skupinami?	12
34 K čemu slouží atomické instrukce? Popište jejich vlastnosti a příklad algoritmu kde je možné je využít.	13
35 Jaké mohou být příčiny nízké výkonnosti kernelů?	13
36 Co je to obsazenost multiprocesoru (occupancy), na čem je závislá?	13
37 [TODO] Co je to aritmetická a paměťová latence?	14
38 [TODO] Co je to separabilní filtr a jakým způsobem lze přistupovat k jeho paralelizaci?	14
39 K čemu slouží paralelní redukce?	14
40 Co je to dynamický paralelismus?	15
41 [TODO] Jaké jsou možné typy přenosů dat mezi CPU a GPU?	15
42 K čemu slouží P2P přístup a jakým způsobem ho lze využít.	15
43 K čemu slouží operace shuffle?	15
44 Vyjmenujte alespoň 3 api pro GP-GPU.	15
45 Co to je SpirV?	15
46 K čemu slouží operace ballot?	15
47 K čemu lze využít atomicCompSwap?	16

- 48 [TODO] Jaký je vztah mezi NDRange/Grid/Dispatch, pracovní skupinou a vláknem/invokací? 16
- 49 Jaké paměti jsou na GPU a popište jejich vlastnosti (odkud je lze plnit, odkud je lze číst, relativní rychlost, relativní velikost). 16
- 50 Jaký je rozdíl mezi pracovní skupinou a warpem/wavefrontem. 18
- 51 K čemu slouží příkaz barrier? Jakou má souvislost s rozdělením vláken do skupin, warpů? Jaký to má vztah k větvení programu? 18

1 Popište základní principy návrhu energeticky úsporného procesoru CPU/GPU, jak se vyhodnocuje energetická úspornost.

Vyhodnocuje se pomocí poměru MFLOPS/W

- další snižování napájecího napětí není možné
- energie na operaci se snižuje lineárně s rozměry prvku
- současné technologie 3 W pro mobilní zařízení, 150 W stolní pc
- v klasické jednovláknové architektuře se většina energie spotřebuje na režii dodání dat (načtení operandů spotřebuje více E než operace)
- Řešení úpravou výpočetní aritmetiky
 - Snížení přesnosti kódu čísel, např. odseknutím n nejnižších bitů
 - přibližné počítání je samostatný specializovaný obor
- Řešení lokalitou dat
 - energeticky efektivní architektury musí snížit objem dat přenesených na instrukci
 - např. přístup do mimočipové paměti $250\times$ dražší
 - požadavek na zvýšení šířky pásma paměti mimo čip a snížení příkonu komunikace

2 Popište vývoj 2D/3D grafického řetězce s různým stupněm zřetězení.

GTXR = generování → traverzace (sledování paprsku) → transformace → rasterizace → zobrazení

1980+

fixed pipeline

1990+

první pokusy a zřetězení – různé architektury = různé řešení

Např.:

- hluboký pipeline – 15 úrovní zřetězení
- paralelizace 2d a 3d řetězce
- zdvojený řetězec - v každém cyklu 2 pixely

2000+

Převedení některých SW částí na hardware

- transformace, shading
- interpolace snímku
- vybírání, vytváření a vykrývání vrcholů
- zavedení z-bufferu

Současnost

Klasická OpenGL pipeline

- programovatelné části – vertex, geometry, control, evaluation, fragment shader, compute shader
- konfigurovatelné části – teselace
- možné extra buffery + transform feedback buffer, dispatch indirect buffer

3 Popište 1- až n-rozměrné multiprocesorové propojovací struktury (Origin/Onyx) a typy propojovacích sběrnic Silicon Graphics.

Origin

Propojovací struktura architektury koncepce Origin2000 umožňuje vytvářet až 5D krychle. Díky její pružnosti a koncepční pokrokovosti byla převzata i do navazujících architektur SGI, jako například Onyx2, a další.

Architektura Origin2000 je tvořena množstvím výpočtových uzlů, propojených navzájem propojovací sítí CrayLink. Každý výpočtový uzel obsahuje jeden nebo dva procesory s pamětí cache, sdílenou paměť, adresář pro řízení koherence cache s hlavní pamětí a dvě různé jednotky interface — jeden XIO, který připojuje I/O jednotky, a druhý CrayLink, propojující uzly přes jednotku Router.

Onyx2

Každý procesor je spolu se svou pamětí cache a částí operační paměti počítače (o kapacitě od 128 MB do 8 GB) součástí jednoho procesorového uzlu, který obsahuje čtyři procesory. Obousměrný přenos dat mezi procesory v jednom uzlu dosahuje hodnoty 1,6 GB/s, resp. 800 MB/s v každém směru zvlášť. Jednotlivé uzly jsou mezi sebou propojeny pomocí propojovací sítě, jejíž topologie sice vychází z ideální hyperkrychle, ale z důvodu optimalizace datových přenosů jsou mezi uzly vytvořeny i další přídavné datové spoje (již z principu musí jít o diagonály), zejména mezi procesory a zobrazovacími subsystémy.

4 Popište principy výstavby, činnosti a použití GPGPU.

Jde o techniku využití GPU k obecným (negrafickým) výpočtům, které běžně probíhají na CPU. Zatímco dříve bylo zpracování dat na GPU pevně dané (static pipeline), dnes už je převážná část programovatelná, a tudíž nabízí možnost obecných výpočtů. K tomu slouží např. rozhraní OpenCL, CUDA či ATI Stream.

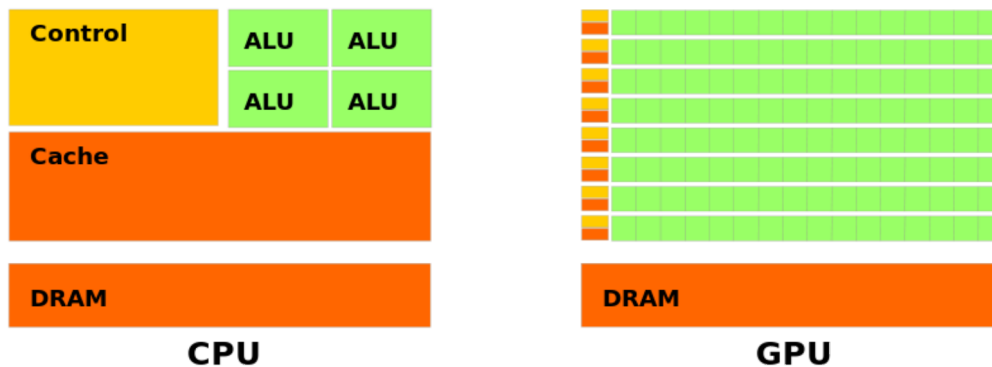
Z povahy grafických čipů vychází také způsob práce s daty, který je silně orientován na paralelní proudové zpracování dat. Data jsou často rozdělena do 2D mřížky, kde nad každou buňkou pracuje jedna výpočetní jednotka. Tyto jednotky sdílí kód programu, nazývaný také kernel. Možnost synchronizace či sdílení dat mezi jednotkami v průběhu výpočtu je omezená a značně snižuje výsledný výkon. (Alespoň v OpenCL možnost sdílení paměti existuje.) Z běžných součástí GPU lze použít například texturovací jednotky jako vstupní rozhraní a framebuffer jako výstupní.

Syntaxe programovacího jazyka může být podmnožinou jazyka C (v OpenCL), ovšem existují omezení plynoucí právě z paralelního proudového zpracování. Není možné používat rekurzi, volání vnořených funkcí má svá omezení, jsou zavedeny speciální funkce a datové typy pro práci s vektory, je žádoucí znovuvyužívat použité proměnné a obecně musí optimalizace probíhat jinak než u běžných (CPU) programů - s ohledem na povahu HW.

Vhodnými úlohami pro zpracování na GPU je například zpracování videa, zvuku či řeči, provádění fyzikálních simulací (kapaliny, osvětlení, počasí) anebo třeba úlohy z kryptografie.

5 Popište koncepci a vlastnosti různých typů Streaming Multiprocessoru

Většinu plochy čipu GPU zabírá velké množství relativně jednoduchých skalárních procesorů, případně jednotek VLIW, které jsou organizovány do větších celků zvaných *streaming multiprocessory*. Vzhledem k tomu, že se jedná o architekturu SIMT, řízení jednotek a plánování instrukcí je jednoduché a spolu s velmi malou vyrovnávací pamětí zabírá malé procento plochy GPU čipu.



Obrázek 1: Porovnání struktury CPU a GPU

6 Popište principy komprese dat v systému Pascal (delta-komprese a další).

Komprese v paměti cache L2

Jsou uplatněny tři principy delta komprese barev.

GPU vypočítává rozdíly mezi barvami pixelů v bloku a ukládá blok jako referenční pixely plus delta hodnoty – rozdíly od referenčních hodnot.

Pokud jsou delta hodnoty malé, pak je zapotřebí na jeden pixel jen malý počet bitů. Standardní kompresní metoda vede na kompresní poměr 2:1.

GeForce GTX 1080 Pascal zavedla tyto kompresní režimy:

- Zdokonalená komprese 2:1
- Komprese 4:1 pro případ velmi malých hodnot delta
- Kompresní režim 8:1 kombinuje kompresi konstantních barev 4:1 pro bloky 2x2 pixelů s kompresí 2:1 hodnot delta mezi těmito bloky

7 Popište architekturu grafických multiprocesorů a nové principy činnosti, jako komprese dat, preempce a její typy.

viz taky 5, 6 a 9

Preempce

Preempce je v informatice přerušování právě vykonávaného procesu (úlohy) bez toho, aby byla vyžadována jeho spolupráce. Přerušovaná úloha je pozastavena, nahrazena jinou úlohou a později je pomocí stejného mechanismu její činnost opět obnovena.

Preempce pracovní zátěže, např. pro časové posunutí (warp) ještě před vykreslením snímku. Používá se:

- Preempce na pixelové úrovni
- Preempce na úrovni vláken
- Preempce na úrovni instrukcí

V zásadě jde o pozastavení vykreslování, uložení kontextu, kde a jak později pokračovat, a rychle přejít na žádost o preempci.

8 Popište princip tensorového jádra Turing, k čemu slouží, jaké formáty dat se používají.

tensor – veličiny, dané např. třemi nebo čtyřmi vektory

Princip činnosti tensorového jádra

- Používané neuronové sítě mají tisíce vrstev a milióny neuronů.
- Pro jejich rychlé trénování a dedukování je zapotřebí vysoký výpočetní výkon.
- Zde jsou základními operacemi násobení matic – matice (Matrix-Matrix multiplication GEMM).
- Používají se k násobení velkých matic vstupních dat a vah v propojených vrstvách sítě.
- Násobení matic se provádí v některých aplikacích v jednoduché přesnosti.
- Pro trénování a dedukci se používá poloviční přesnosti.
- V jádře Volta se používá smíšené přesnosti, násobení matic se vstupy v FP16 a akumulace v FP32.
- Tensorová jádra jsou navržena pro činnost s vysokou energetickou efektivností.

Každé tensorové jádro pracuje s maticemi 4×4 a provádí operaci Matrix Multiply and Accumulate

$$D = A \times B + C$$

kde A , B , C a D jsou matice 4×4 . Vstupy pro maticové násobení A , B jsou v FP16, akumulační matice C , D mohou být matice FP16 nebo FP32.

$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

FP16 or FP32 FP16 FP16 FP16 or FP32

Obrázek 2: Matrix Multiply and Accumulate

9 Popište principy a funkce preempce Pascal.

Preempce je v informatice přerušení právě vykonávaného procesu (úlohy) bez toho, aby byla vyžádána jeho spolupráce. Přerušená úloha je pozastavena, nahrazena jinou úlohou a později je pomocí stejného mechanismu její činnost opět obnovena.

Výpočtová preempce

Brání dlouho běžícím aplikacím ovládnout výlučným způsobem celý systém a znemožnit tak běh dalších aplikací, nebo časově překročit.

Preempce pracovní zátěže

V zásadě jde o pozastavení vykreslování, uložení kontextu, kde a jak později pokračovat, a rychle přejít na žádost o preempci.

10 Popište vývoj unifikovaného adresového prostoru, principy činnosti, a jeho hardwarovou podporu.

- Programátor nebude muset pracovat s více paměťovými prostory a limitující přenosovou šířkou pásma na paměť a systém bude celkově energeticky úspornější
- Společně s koherencí cache zavádí GCN virtuální paměť jako kombinaci hardwarové podpory a driveru
- Virtuální paměť GCN je definovaná jako kompatibilní s x86
- Tím se zjednoduší přenášení dat mezi CPU a GPU
- Dále vytváří možnost bez problémů sdílet jediný adresový prostor pro jednotky CPU i GPU
- Sdílení dat bez jejich kopírování je důležitý faktor výkonnosti i příkonové úspornosti v heterogenních systémech

11 Popište principy návrhu energeticky úsporného GPU pro mobilní zařízení.

- V režimu nízké kvality je průměrná kvalita obrázku 26,89 dB s úsporou energie 33,66 %
- V režimu vysoké kvality se dosahuje v průměru 45,94 dB a úspory energie 31,63 %
- Výslednou kvalitu obrázku může vybrat uživatel
- Výběr přesnosti se provádí v zásadě automaticky a neovlivňuje vlastní program shaderu
- Návrh systému EHPS používá FP se sníženou přesností na 32 bitech a FX na 24 bitech, ALU SIMD a SFU pro úsporu energie
- Základní technikou úspory energie v shaderu je používat pevnou řádovou čárku a snížit přesnost čísel s pohyblivou řádovou čárkou

12 [TODO] Popište hlavní vývojové etapy a principy grafických systémů Mali pro mobilní zařízení.

13 Vysvětlete a zdůvodněte koncepci kachliček (tiles) 16x16 pixelů v grafice Mali.

Tento princip je srovnáván s tradičním bezprostředním režimem (immediate mode), používaným typicky v počítačích PC a v terminálech. Vykreslování s použitím kachliček znamená vykreslovat buffer snímku z menších úseků – z kachliček 16×16 pixelů. Když je kachlička dokončená, vypíše se do paměti.

Bezprostřední režim

Vykreslování v bezprostředním režimu se provádí jako přesná posloupnost příkazů pro každý primitiv v každém volání kreslení, nepoužívá se paralelní zpracování a pipelining.

14 [TODO] Popište formáty dat podporované v grafice Mali, skalární i vektorové.

15 Popište a zdůvodněte principy tří základních typů komprese textur - ztrátové, bezztrátové a adaptivní.

Komprese ASTC

Adaptive Scalable Texture Compression je kompresní metoda ARM. Je to ztrátová metoda pro kompresi a dekompresi textur, ukládaných v bufferech v pamětech DDR – Double Data Rate.

Ukládání geometrických dat vyžaduje běžně přes 64 bytů na vrchol, ve srovnání se 4 bity na texel pro Ericsson Texture Compression 2 (ETC2).

Komprese AFBC

ARM Frame Buffer Compression je bezztrátová kompresní metoda pro kompresi a dekompresi toku videa např. ve formátu H.264. Umožňuje přístup k datům v libovolném místě toku videa.

Metoda AFBC je nyní dostupná pro všechny video stroje Mali, budoucí GPU Mali a jako licencovaný IP pro displejové procesory.

16 Popište alespoň 3 rozdíly mezi architekturami GPU a CPU.

CPU

- programování pomocí běžných jazyků (C, C++, ...)
- menší počet komplexních výkoných výpočetních jednotek
- predikce skoků, spekulativní vykonávání, přejmenování registrů, vykonávání instrukcí mimo pořadí, ...

GPU

- programování/konfigurace jednotek pomocí speciálního API (OpenCL, Vulkan)
- velké množství jednoduchých skalárních procesorů (SIMD jednotek) organizovaných do desítek multiprocesorů

- cca 20× vyšší výkon a 10× paměťová propustnost
- na každý multiprocessor lze naalokovat řádově stovky až tisíce vláken
- přepínání vláken bez penalty
- různé druhy pamětí sloužící různým účelům
- in-order zpracování

17 Co je to kernel?

- program pro GPU, spouští se paralelně na více vláknech zároveň
- sdružuje se do pracovních skupin
- typicky definován v externím kódu (OpenGL, OpenCL), spouští se pomocí API
- lze jej řadit do front

18 Co je to vlákno?

- jedna instance kernelu (viz 17)
- jsou organizovány do pracovních skupin
- každé vlákno má svůj index v rámci skupiny i v rámci kernelu – slouží k indexaci dat, nad kterými má pracovat (až 3D index)

19 Co je to divergence vláken a kdy vzniká?

Když se vlákna jednoho warpu dostanou do různých větví kódu – toto zpomaluje vykonávání programu, protože jednotlivé větve nelze vykonávat najednou (SIMD ne MIMD) a musí se provádět postupně.

Ideálně potřebujeme aby vlákna dělala vždy tu samou instrukci, pokud ne, dochází k divergenci.

20 Co je to warp/wavefront?

- Skupina vláken (podskupina pracovní skupiny) která patří do jedné pracovní skupiny a vykonávají se společně ve stejnou dobu na jednom multiprocessoru.
- Nvidia – *warp*, AMD – *wavefront*
- Typická velikost 32 nebo 64

21 Co je to multiprocessor (streaming multiprocessor/compute unit) GPU a k čemu slouží.

- Jsou to hlavní výpočetní jednotky na GPU které vykonávají kód warpu
- Nvidia – *streaming multiprocessor*, AMD – *compute unit*
- akce běžící na různých MP jsou na sobě nezávislé
- na jednom MP může běžet víc warpů, pokud na to stačí zdroje
- MP má plánovač, který řeší vyhodnocování warpů
- na každý MP lze naalokovat až tisíce vláken a typicky obsahuje až desetitísíce 4 B registrů

22 Co je to fronta příkazů (command queue)?

- fronta, ve které jsou řazeny jednotlivá volání kernelů pro GPU
- OpenGL nemá přímý přístup ke frontě – vytváří se automaticky při vytvoření kontextu
- OpenCL může mít i více front, které manipulují s buffery, spouští kernelů
- Implicitně *in-order* fronta (lze přenastavit)

23 Jakým způsobem je možné synchronizovat úlohy ve frontě příkazů s vykonáváním mimo pořadí (out-of-order execution) nebo mezi frontami v OpenCL?

Pomocí eventů

24 Jakým způsobem se eliminuje/zmírňuje dopad aritmetických/paměťových latencí.

Pokud nějaké vlákno čeká vlivem latence → je místo něj spuštěno jiné vlákno, pokud máme přístup k dostatečnému množství vláken které lze spustit můžeme kompletně vyblokovat efekty latence.

25 Co je to konflikt banků v lokální paměti a kdy vzniká?

Banky jsou malé paměťové elementy (přístup do banky zabere 2 cykly) o velikosti 4 B. Nedojde-li ke konfliktu bank, lze k nim přistupovat současně prostřednictvím *half-warpu*. V rámci celého warpu se jedná o 2 “transakce”, kde jedna je pro dolní a druhá pro horní *half-warpu*. Konflikt bank spočívá v tom, že dvě paměťové adresy v rámci half-warpu odkazují do stejné banky. V takovém případě se přístup do bank serializuje a není již současný.

26 Co je to zarovnaný přístup do paměti?

Přístup k datům postupně jak jsou v paměti – je třeba si je takto přichystat aby se vždy přistupovalo k smysluplným datům – toto umožňuje lépe pracovat s pamětí – zlepšuje to práci cache a plně se využije načítání větších bloků paměti najednou.

Obvykle se doporučuje zarovnávat paměť podle velikosti cache-line, protože jinak může jeden proces zabrat $\frac{2}{3}$ řádku a další po něm následující musí jít přes dvě cache-line a bude pomalejší. Ideálně teda zarovnat na celý řádek i když se mrhá daty. Taký přístup k datům je pak rozumnější (indexy).

27 Popište konstantní/uniformní paměť.

viz 49

28 Popište lokální/sdílenou paměť.

viz 49

29 K čemu slouží texturovací jednotky?

Pod pojmem textura rozumíme dvourozměrný obrázek, který je určen k potahování předmětů a vytváření jejich reálného vzhledu. Vstupem do Texturovací části grafického pochodu je transformovaný a na trojúhelníky rozdělený osvětlení a nastínovaný objekt (zpravidla ve stupních šedi, či v některém základních barevných tónů - RGB).

Textura tudíž slouží jako tapeta, vzor povrchu, kterým se daný předmět potáhne (a může se jednat klidně o poloprůhlednou texturu vody, či kouře a nejen o klasickou představu cihlové zdi).

V 95 % případů nestačí klasická jedna textura jako mapa, reprezentující povrch objektu. Při zobrazení např. stínového přechodu se na texturu umísťují další a další vrstvy poloprůhledných textur (LightMap), které dávají objektu světelně přechodový vzhled (viz. kapitola o stínování). Problematické jsou vždy poloprůhledné a nesouvislé textury jako je kouř, voda, sklo, atd. V jednom cyklu tak vyvstává nutnost nanést až několik textur najednou, např. Textura + Průhledná textura + odrazová textura, atd. Obecně jsou na jednom pixelu navrženy až tři textury, při texturování složitějších částí jako již zmiňovaná voda může počet vrstev 6 i více.

Čím více dokáže grafická karta v jednom cyklu nanést textur na pixel (Multitexturing), tím je větší pravděpodobnost, že celá operace texturování pixelu bude zvládnuta v tom samém cyklu.

30 [TODO] Jaké jsou rozdíly mezi bufferem a texturou v OpenCL?

Textury budou nejspíše využívat texturovací jednotky

31 Jaký je rozdíl mezi normalizovanými a nenormalizovanými texturovacími koordinátami?

Nenormalizované jsou v rozsahu $[0, x]$, kde x je velikost obrazu.

U normalizovaných $[0, 1]$.

32 Co je to register spilling?

Když existuje více proměnných než je možné udržet v registrech musí se některé proměnné ukládat do paměti.

33 Jaké jsou způsoby komunikace mezi vlákny ve skupině a mezi skupinami?

V rámci skupiny

1. zápis do lokální nebo globální paměti
2. synchronizace
3. čtení z lokální nebo globální paměti

Omezení při současném zápisu do stejné proměnné z více vláken – potřeba atomických instrukcí

Jednosměrná komunikace mezi skupinami

- pomocí globální paměti
- není zaručeno pořadí vyhodnocení pracovních skupin
- při současném zápisu do stejné proměnné z více skupin – potřeba atomických instrukcí

34 K čemu slouží atomické instrukce? Popište jejich vlastnosti a příklad algoritmu kde je možné je využít.

K výlučnému přístupu k daným datům a zabráněním WAR, WAW a podobných paměťových chyb.

Použití třeba u paralelní redukce (suma, min, max, ...) kdy všechny vlákna chtějí přičíst svou hodnotu k celkovému výsledku, ale bez výlučného přístupu jeden udělá `sum += y1` a než zapíše, tak jiný udělá své `sum += y2` a změní původní hodnotu a opožděný první mu ji přepíše a tím se znehodnotí výsledek.

35 Jaké mohou být příčiny nízké výkonnosti kernelů?

- slabá obsazenost (occupancy)
- aritmetické latence - RAW konflikty
- divergence vláken - větvení, cykly
- paměťové latence
- paměťová propustnost globální/lokální paměti

36 Co je to obsazenost multiprocesoru (occupancy), na čem je závislá?

Vytíženost multiprocesoru — čím větší tím lepší, závisí na divergenci vláken a na správném rozložení pracovních skupin

Pro maximální rychlost výpočtů má smysl co největší obsazenost, ale když vlákna nemakají (není pro ně práce) je vhodné uvolnit místo pro jiné, ať se nebrzdí výpočet jiných workgroup a podobně. Vhodné je taky uvažovat nad rozmístěním vláken v rámci multiprocesoru.

Obsazenost multiprocesoru (occupancy)

- % naalokovaných vláken z maximálního počtu
- alokovatelných vláken na multiprocesor
- čím větší % tím větší množství aktivních warpu na multiprocesor → lepší pokrytí aritmetických či paměťových latencí
- vlákna se alokují po celých skupinách

Vliv na obsazenost

- slabá granularita alokace skupin
- příliš velké skupiny → velký dopad synchronizačních funkcí
- příliš malé množství vláken/skupinu → limitace maximálním množstvím skupin
- velikost skupiny → na nových GPU lze naalokovat maximálně 16 - 32 skupin
- spotřeba registrů na vlákno → maximální velikost na nových GPU typicky 64 kb na multiprocesor
- spotřeba lokální paměti na skupinu → maximální velikost na nových GPU typicky 64kB – 96kb na multiprocesor

37 [TODO] Co je to aritmetická a paměťová latence?

Doba čekání na dokončení paměťových operací a aritmetických operací – načítání dat zpravidla trvá více než jeden takt, podobně tomu bývá u dělení a jiných složitých aritmetických operací.

Aritmetická latence

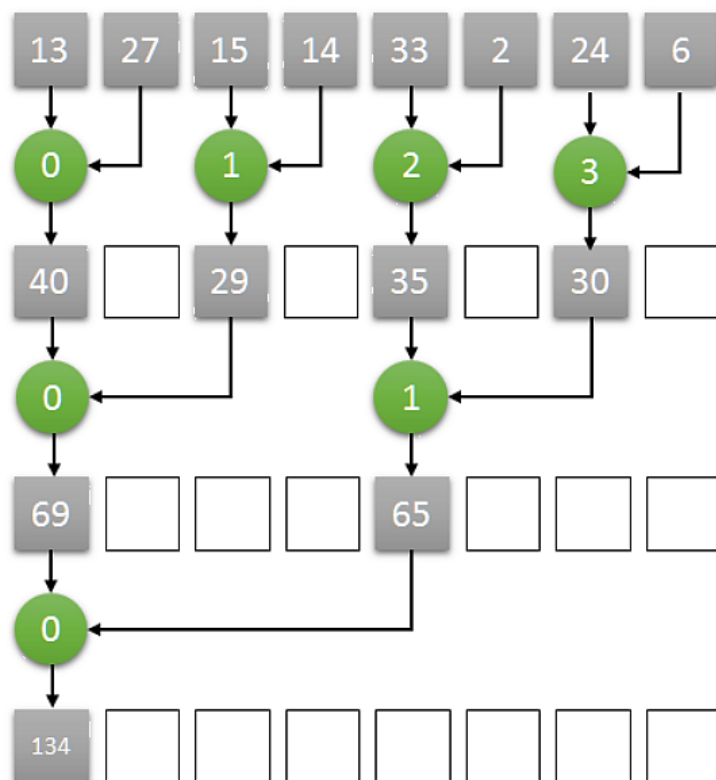
- způsobena čtením právě zapsané hodnoty (RAW - read after write konflikt)
- u moderních GPU typicky 10 – 20 cyklů
- řešení přeskupením instrukcí – automaticky pomocí kompilátoru
- řešení zpracováním více elementů/thread – stoupá spotřeba registrů, klesá potřeba aktivních warpů na multiprocessor

38 [TODO] Co je to separabilní filtr a jakým způsobem lze přistupovat k jeho paralelizaci?

GMU 06, p. 22

39 K čemu slouží paralelní redukce?

Paralelizace cyklu se závislostí na předešlé iteraci při operaci která je asociativní – jde ze složitosti $O(n)$ na $O(\log(n))$



Obrázek 3: Příklad paralelní redukce

40 Co je to dynamický paralelismus?

Počet aktivních vláken se mění během výpočtu (došla práce pro některé vlákna, tak je zruším).
Schopnost vlákna kernelu založit další kernel

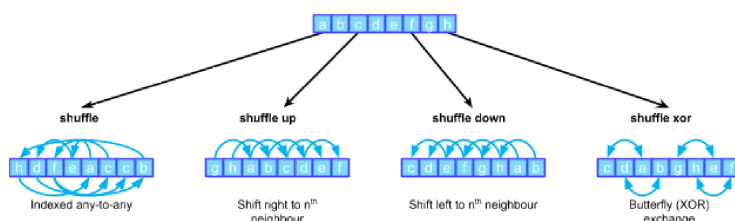
41 [TODO] Jaké jsou možné typy přenosů dat mezi CPU a GPU?

42 K čemu slouží P2P přístup a jakým způsobem ho lze využít.

Přímý přístup k paměti, používá se hlavně ke zvýšení propustnosti paměti.

43 K čemu slouží operace shuffle?

Výměna dat mezi vlákny ve warpu



Obrázek 4: Paralelní redukce s shuffle

44 Vyjmenujte alespoň 3 api pro GP-GPU.

VulkanRT, CUDA, OpenGL, DirectCompute, OpenCL

45 Co to je SpirV?

Softwarový nástroj překládající zdrojové kódy různých platforem (OpenGL, CUDA, ...) na jednotnou nízkourovňovou API. Cílem je zjednodušit vývoj a umožnit použití všeho na všem.

46 K čemu slouží operace ballot?

- Speciální instrukce vrací bitovou masku
- Umožňuje komunikaci mezi vlákny ve warpu/wavefrontu
- Každé vlákno ve warpu vyhodnotí podmínku a zapíše výsledek na bit proměnné jehož číslo odpovídá číslu vlákna ve warpu
- `ballot` instrukce je možná na NVIDIA i AMD
- délka warpu je různá (32, 64 bit)
- `ballot` lze simulovat pomocí lokální paměti, ale je to pomalejší

47 K čemu lze využít atomicCompSwap?

viz 34

Instrukci `atomicCompSwap` lze použít pro implementaci nepodporovaných instrukcí jako jsou například: `atomicMax`, `atomicMin`

`atomicCompSwap(mem, compare, data)` provede atomické porovnání `mem` a `compare`. Pokud se rovnají, pak se obsah `data` zapíše do `mem`, jinak se nedělá nic. Funkce vrací původní obsah `mem` bez ohledu na výsledek porovnání.

Možná použití například v řadicích algoritmech – porovnej dvě položky a když jsou opačně, prohoď. (Atomicky proto, aby soused nepracoval s mojí položkou a já neprohodil nesmysl.)

48 [TODO] Jaký je vztah mezi NDRange/Grid/Dispatch, pracovní skupinou a vláknem/invokací?

Pracovní skupina se skládá z vláken

NDRange (openCL)

- global size – počet vláken celkově v dané dimenzi
- local size - počet vláken v jedné pracovní skupině
- počet pracovních grup je určen z global a local size hodnot

Grid (CUDA) – 3 dimenze – počet pracovních grup?

Dispatch (OpenGL) – funkce která spustí počet pracovních grup (ve 3 dimenzích) definovaný v jejím volání?

49 Jaké paměti jsou na GPU a popište jejich vlastnosti (odkud je lze plnit, odkud je lze číst, relativní rychlost, relativní velikost).

Typ	Umístění	Rychlost	Cache	Velikost	Dostupnost
global	off-chip	0,11 B/tick	1	3 072 MB	vše
local	on-chip	1,33 B/tick	0	0,72 MB	skupina
constant	off-chip	4 B/tick	1	64 kB	vše
register	on-chip	16 B/tick	0	4 MB	vlákno

Globální paměť

- off-chip paměť operační paměť grafického adaptéru
- velikost v řádech GB
- latence v řádu stovek cyklů → nutnost běhu mnoha warpů pro překrytí latence (multithreading)
- rychlost o řád nižší než on-chip paměti (lokální paměti, registry)
- u novějších architektur cachováno pomocí L2 cache, případně i L1 cache → zmírnění dopadu opakovaného načítání stejných dat
- paměť rozdělena po segmentech
- pokud je na jednu adresu ze segmentu přistoupeno z warpu přečte se celý segment
- paměťové segmenty se pohybují mezi 32 B a 128 B
- u starších architektur nezarovnaný přístup - serializace

Konstantní paměť

- lze ji plnit a číst z aplikace
- maximální velikost bufferu dle specifikace 64 kb
- optimalizace na broadcast → přístup do jiných lokací v rámci warpu → serializace
- cachované pomocí konstantní cache
- omezený počet bufferů na kernel – minimálně 8 dle specifikace
- cachování ve speciální cache multiprocesoru
- použití: read-only konstanty používané všemi vlákny

Lokální paměť

- Sdílená v rámci work-group
- Umístěná přímo na čipu – větší propustnost a menší latence než globální paměť (často se před výpočtem plní z globální paměti)
- minimální velikost je pro OpenCL 1.0 16 kB, u novějšího 32 kB
- velikost na předchozích generacích GPU typicky 32 – 48 kB, na aktuální generaci GPU typicky 64 – 96 kB
- u CPU a některých starších GPU emulovaná pomocí globální paměti – pomalé, nepoužívat
- rozdělena do banků s velikostí buňky typicky 4 B
- přístupu do stejného banku více vláken z warpu – serializace
- novější GPU umějí broadcast a multicast – přístup z více vláken z warpu na stejnou adresu nezpůsobuje bank konflikty
- použití: kooperace mezi vlákny, uživatelské cachování

Texturovací paměť

- data umístěna v globální paměti
- 1D až 3D textury
- data prochází přes texturovací jednotky → interpolace, data mimo rozsah, různé adresování
- adresování
 - normalizované → koordináty staženy do rozsahu $< 0,0, 1,0 >$
 - nenormalizované → koordináty v rozsahu $< 0,0, size >$, první texel na pozici 0,5 0,5
- interpolace
 - nejbližší soused → zaokrouhlení na nejbližší texel
 - lineární → lineární přechod mezi sousedními texely
- řešení okrajů
 - opakování obrazu
 - okrajová hodnota
 - zrcadlení
- optimalizace na načítání bloků
- cachování i na starších architekturách, na nových i do L1 cache multiprocesoru

Registry

- nejrychlejší paměť
- přístup do registrů bez latence
- typicky desetitisíce 4 B registrů na multiprocessor
- pro každé vlákno je alokován potřebný počet registrů – při příliš velké spotřebě omezuje obsazenost
- neblahý vliv na spotřebu registrů může mít rozbalování smyček
- příliš mnoho registrů na vlákno → register spilling

50 Jaký je rozdíl mezi pracovní skupinou a warpem/wavefrontem.

Oboje jsou skupiny vláken, které spolu mohou komunikovat – běží na stejné výpočetní jednotce/streaming multiprocessoru

Pracovní skupina

- uživatelem definovaná skupina vláken (work-group)
- může obsahovat více wavefrontů — je jim nadřazená
- různé wavefronty mohou vykonávat různé instrukce – synchronizovat je nutné explicitně

warp/wavefront

- pro všechna vlákna se vykonává stejná instrukce (při divergenci některé ale nemusí pracovat)

51 K čemu slouží příkaz `barrier`? Jakou má souvislost s rozdělením vláken do skupin, warpů? Jaký to má vztah k větvení programu?

- Slouží k synchronizaci vláken
- Vlákna můžeme synchronizovat pouze v rámci skupiny (work-groupy)
- Synchronizuje se pomocí příkazu `barrier`
- Všechna vlákna ve skupině jej musí zavolat, než kterékoliv může pokračovat
- Pokud je `barrier` v podmínce, musí ji navštívit všechna vlákna ve skupině nebo žádné
- V případě, kdyby jedno vlákno vykonávalo jinou *control flow* cestu, ostatní vlákna by zůstaly čekat na `barrier` donekonečna