Tentamen Computerarchitectuur

Woensdag 17 januari 2018, 10:00 - 13:00 uur

Examinator: dr. K. F. D. Rietveld

- Het tentamen is **gesloten boek**, dus het is niet toegestaan om het tekstboek, slides of eigengemaakte aantekeningen te gebruiken.
- Alleen rekenmachines waarin geen teksten kunnen worden opgeslagen zijn toegestaan. Het gebruik van grafische rekenmachines is **niet** toegestaan.
- De vragen mogen worden beantwoord in het Nederlands en Engels.
- Beargumenteer al uw antwoorden. Aan antwoorden zonder uitleg of uitwerking worden geen punten toegekend.

- Bij het inleveren van het gemaakte werk zal u worden verzocht uw naam, studentnummer en aantal ingeleverde bladen te noteren op de presentielijst.

- Het aantal opgaven is 5 met een totaal van 22 onderdelen. Het tentamen bestaat uit 5 pagina's.
- Bij elk onderdeel staat tussen vierkante haken het aantal te behalen punten aangegeven. Het totaal aantal te behalen punten is 100.

In dit tentamen zullen we in verschillende opgaven aspecten bespreken van de Hitachi SuperH SH-4 processor. Deze processor is vooral bekend van het gebruik in de Sega Dreamcast gaming console, welke eind jaren negentig op de markt verscheen. Op de datasheet van deze processor komen we de volgende eigenschappen tegen:

- 200 MHz, 360 MIPS (Dhrystone).
- 1.4 peak GFLOPS.
- 16-bit instructielengte.
- Two-issue superscalar RISC pipeline.
- On-chip 8 KB instruction en 16 KB data cache.
- 16 general purpose registers van 32 bits.
- 32 single-precision floating point registers, opgedeeld in twee banken van 16 registers. Een instructie kan maar instructies uit één bank tegelijkertijd aanspreken.
- 4 single-precision vector registers van 128 bits.
- Speciale 3-D graphics / SIMD instructies die opereren op vectoren van 128 bits (dus vectoren bestaande uit 4 single-precision elementen).
- 64-bit memory bus.
- Totaal van 234 ondersteunde instructies.

Opgave I – Performance [15 punten]

a. [3 punten] Gezien de klokfrequentie en de gegeven MIPS-rating (Millions of Instructions Per Second) van de SH-4 processor, wat zijn de CPI en IPC die gemiddeld zouden worden behaald?

b. [4 punten] Bij nadere bestudering van de datasheet blijkt dat de MIPS-rating is bepaald middels de zogenaamde Dhrystone benchmark. Dit is een synthetische benchmark ontwikkeld door Reinhold Weicker. Bespreek de nadelen van een dergelijke benchmarkmethode.

Stel nu, we hebben een andere RISC processor die opereert op een klokfrequentie van 400 MHz. Deze processor heeft instructietimings zoals te vinden in de tabel hieronder. De fabrikant heeft met een zelf ontworpen benchmark de verdeling van instructies, of instructiemix, gekarakteriseerd zoals is te zien in de rechterkolom.

Operatie	CPI	Aantal in miljoenen
Data transfer	3	280
Arithmetic	1	320
Shift	2	40
Branches	3	96
FP ALU	4	64

c. [4 punten] Wat is de overall CPI die wordt behaald op deze machine met deze instructiemix? En wat is dan de MIPS-rating?

d. [4 punten] Is het eerlijk om de gegeven MIPS-rating van de SH-4 en die van de machine zoals bepaald onder **c** direct met elkaar te vergelijken om een oordeel te vellen over het verschil in performance tussen deze processoren? Waarom wel/niet?

Opgave II - Processor organisatie [16 punten]

In deze opgave bekijken we wederom een aantal aspecten van de SH-4 processor.

a. [4 punten] Leg in uw eigen woorden duidelijk uit wat een "two-issue superscalar pipeline" inhoudt.

b. [4 punten] De instructies van deze architectuur hebben een grootte van 16 bits. Beredeneer of het hier om 2-operand of 3-operand instructies gaat.

c. [4 punten] Hebben we hier te maken met een in-order of out-of-order micro-architectuur? Uit welk gegeven, of het ontbreken van welk gegeven, blijkt dit?

d. [4 punten] In marketingmateriaal werd de Sega Dreamcast gepresenteerd als de eerste 128-bit gaming console. Waarom was dit zo, gezien de specificaties? Vindt u dit terecht? Waarom wel/niet?

Opgave III - Caching [24 punten]

De SH-4 beschikt over een instruction cache van 8 KB en een data cache van 16 KB. Beide caches zijn direct-mapped en hebben cache lines (block size) ter grootte van 32 bytes. De caches zijn verbonden met een gedeeld RAM geheugen.

a. [4 punten] Is er in dit geval sprake van een Von Neumann of Harvard-architectuur? Leg uit.

b. [5 punten] Bereken voor beide caches het aantal sets in de cache en het aantal benodigde bits om die cache te kunnen indexeren.

Hoewel een direct-mapped cache op eenvoudige wijze kan worden geïmplementeerd, blijft de performance van dit type cache achter voor bepaalde patronen van geheugenreferenties.

c. [5 punten] Bepaal een reeks van geheugenreferenties gegeven in blokadressen ("block addresses") voor de data cache, welke voor de direct-mapped cache een beduidend hogere miss rate laat zien dan voor een 4-way set associative cache. In beide gevallen zijn cache lines 32 bytes groot. Ga ervan uit de cache in het begin leeg is. De gebruikte replacement-strategie in de set associative cache is LRU. Motiveer uw antwoord.

Stel dat we overwegen set associativity te implementeren in de instruction cache. Voor de huidige direct-mapped instruction cache is de miss rate 7% en de miss penalty 8 cycles. Wanneer we 2-way set associativity implementeren, wordt de miss rate 5.1% met een miss penalty van 11 cycles en voor 4-way set associativity een miss rate van 4.5% met een miss penalty van 12 cycles. We gaan uit van een workload van 860 miljoen instructies. Er vindt voor iedere instructie één memory access plaats (de instructie fetch).

d. [5 punten] Geef een advies welke cache configuratie het meest geschikt zou zijn voor deze workload op basis van een berekening van het aantal memory stall cycles veroorzaakt door instructie fetches.

e. [5 punten] De SH-4 voert hardware-gebaseerde prefetching uit voor instructies, maar niet voor data. Er is echter wel een instructie aanwezig om software-gebaseerde prefetching te faciliteren. Bespreek twee redenen waarom de ontwerpers besloten zouden kunnen hebben hardware-gebaseerde prefetching voor data niet in de processor op te nemen.

Opgave IV - Instruction-Level Parallelism [26 punten]

In deze opgave bekijken we een 5-stage micro-architectuur welke gelijkenis toont met de SH-4. De gebruikelijke stages zijn: Instruction Fetch (IF), Decode and Register Fetch (ID), Execute (EX), Memory (MEM), Write back (WB)). Voor (single-precision) floating-point instructies zijn de stages echter: IF, ID, F1, F2, FS; waarbij aan het einde van FS ook de register write-back plaatsvindt. Elke stage kost 1 klokperiode. De tabel hieronder laat van elke instructie betekenis, instructiegroep en latency zien. In dit geval geeft latency het aantal klokperioden aan dat moet volgen na de ID stage van een instructie voordat het resultaat bekend is en het dus mogelijk wordt om dit resultaat middels forwarding naar een andere stage te sturen.

Instruction	Betekenis	Groep	Latency
`LD`	Laad waarde uit geheugen	LS	2
`SW`	Sla waarde op in geheugen	LS	1
`ADD`	Integer optelling	EX	1
`CMP`	Integer vergelijking, slaat resultaat op in compare flag	MT	1
`BF`	Branch als compare flag False	BR	2
`FADD`	Floating-point optelling	FE	3
`FMUL`	Floating-point vermenigvuldiging	FE	3

NB: Alleen `CMP` schrijft naar de compare flag en dit resultaat wordt niet overschreven door andere (niet-`CMP`) instructies.

We bekijken nu de volgende pseudo assemblycode (NB. dit zijn geen SH-4 instructies):

```
Loop:   LD    F2,0(R1)        ; F2 <- Mem[R1+0]
        LD    F3,16(R3)       ; F3 <- Mem[R3+16]
        FMUL  F2,F2,F1        ; F2 <- F2 * F1
        FADD  F2,F2,F3        ; F2 <- F2 + F3
        SW    F2,16(R3)       ; Mem[R3 + 16] <- F2
        ADD   R1,R1,4         ; R1 <- R1 + 4
        ADD   R3,R3,4         ; R3 <- R3 + 4
        CMP   R1,R2           ; R1 == R2?
        BF    Loop            ; Branch if false.
```

a. [6 punten] Laat voor één iteratie van de loop zien hoe de pipeline per kloktik wordt gevuld. Indien er een stall optreedt, benoem dan ook de reden waarom.

b. [5 punten] Pas instruction scheduling toe om een verbeterde versie van de assemblycode op te stellen die in minder klokperioden kan worden uitgevoerd.

(Opgave IV gaat verder op de volgende pagina.)

De SH-4 heeft de mogelijkheid om twee op elkaar volgende instructies in parallel uit te voeren wanneer deze instructies geen dependence op elkaar hebben die dit onmogelijk maakt én dit is toegestaan volgens de volgende tabel wanneer we kijken naar de groepen van deze twee instructies:

		2nd Instruction					
		MT	EX	BR	LS	FE	CO
1st Instruction	MT	O	O	O	O	O	X
	EX	O	X	O	O	O	X
	BR	O	O	X	O	O	X
	LS	O	O	O	X	O	X
	FE	O	O	O	O	X	X
	CO	X	X	X	X	X	X

O: Can be executed in parallel

X: Cannot be executed in parallel

Bron: SuperH (SH) 32-Bit RISC MCU/MPU Series SH7750 Programming Manual

c. [4 punten] Wat valt er in deze tabel op over groepen van instructies die tegelijkertijd kunnen worden uitgevoerd? Wat kunt u hieruit afleiden wat betreft het aantal en de typen functional units in een SH-4 processor?

d. [6 punten] Geef een nieuwe versie van de assemblycode met een instructievolgorde zodanig dat het programma correct is en er zoveel mogelijk paren van instructies in parallel zouden worden uitgevoerd.

e. [5 punten] Leg uit wat er wordt verstaan onder de begrippen *branch prediction* en *speculative execution* en licht de relatie tussen deze twee begrippen kort toe.

Opgave V - Parallelism [19 punten]

a. [5 punten] We zien in een Graphics Processing Unit (GPU) verschillende vormen van data-parallel rekenen samen komen: multi-processors, multi-threading en SIMD. Laat voor elk van deze drie vormen zien hoe deze terugkomt in de algemene architectuur van een GPU.

b. [4 punten] We overwegen een programma te optimaliseren dat op dit moment een executietijd heeft van 32 minuten. Het blijkt dat 12 minuten hiervan worden gespendeerd in een rekenkernel welke kan worden vervangen door een GPU-versie, waarmee voor deze kernel een speedup van een factor 50 wordt gerealiseerd. Hier staat tegenover dat het overige deel van het programma 5% langzamer wordt vanwege bus transfer overhead. Wat is de overall speedup die met deze optimalisatie kan worden behaald?

De SH-4 heeft een pieksnelheid van 1.4 GFLOPS. Deze snelheid wordt gehaald via een speciale SIMD instructie die een single-precision 4×4-matrix kan vermenigvuldigen met een 4-vector. De single-precision elementen zijn 4 bytes groot.

c. [5 punten] (a) Bereken het aantal floating-point operaties dat de SH-4 per klokperiode moet kunnen uitvoeren om deze pieksnelheid te kunnen halen.
(b) Wat is de arithmetic intensity van deze matrix-vector vermenigvuldiging?

d. [5 punten] Wat betekenen de afkortingen UMA en NUMA? En leg uit waarom in moderne serversystemen die bestaan uit meerdere multi-core CPUs de stap van UMA naar NUMA is gemaakt.