Øving 2
TDT4200

Anette Fossum Morken

Del 1, oppgave 1

a

i)

Nivida Maxwell -tråder -homogenius

ii)

ARM big.LITTLE er en samling av to forskjelige prosessorer. Den ene prosessoren er energikrevende og den andre er lite energikrevende slik at når enkle oppgaver, som tekstbehandling, skal gjøres gjøres det på den mmindre energikrevende prosessoren og når tyngre oppgaver skal gjøres gjøres de på den mer energikrevende prosessoren. Siden den tydleig består av to forskjellige prosessorer har denne heterogenious kjerner

iii)

vilje@NTNU er en superdatamaskin som består av prosessorere i kluster. Hver prosessor igjen bastår av et vist antall tråder.

-numa

iv)

En vanlig CPU i dag består av flere like kjerner hvor hver kjerne består av et vist antall tråder.

b

SIMT passer inn i Flynns taksonomi i SIMD, der det på Nivida er en instruksjon flere tråder vil det i Flynns taksonomi bli en instruksjon multiple data.

c

i)

Nivida Maxwell passer inn i Flynns både som SIMD og MIMD siden den har mulighet til å gjøre plere ting på en gang.

ii)

ARM big.LITTLE blir i dag mest sett på som SIMD, men siden den består av to forkjellige prosessorere kan den i teorien gjøre to forskjellige oppgaver samtidig.

iii)

vilje@NTNU MIMD

iv)

En vanlig CPU SIMD

Del 1, oppgave 2

a

tråder er organisert i blokker, bokker er organisert i gridd.

b

Setter opp ligninger for tiden GPUen vil bruke og CPUen vil bruke:

$$GPU = \frac{2n}{r} + \frac{5n}{r} + 5h_{GPU}n7h_{GPU}log_2(n)$$
$$CPU = 5h_{CPU}n7h_{CPU}log_2(n)$$

der $h_{GPU} = 1$, $h_{CPU} = 10$, r er båndbredden og n er mengde data som overføres. Det lønner seg å bruke GPUen når $GPU > CPU$ det finnes ved å regne ut GPU og CPU for mange forskjellige n og plotte dem og se hvor de krysser hverandre slik at $GPU > CPU$.

c

?????

d

0.0.1 i)

Wrap:

0.0.2 ii)

Occupancy

0.0.3 iii)

Minne fortetting(coalescing) er å slå sammen flere minneoverføringer mellom minet og tråder til en overføring.

0.0.4 iv)

Lokalt minne er minne som en tråd kun har tilgang til skal en tråd ha informasjon fra en annen må dette deles gjennom message passing.

v)

Delt minne er at alle trådene i en blokk har tilgang til informasjonen som ligger her.

Del 2, oppgave 1

c

TIden ble tatt på forskjellige steder i koden, hele programmet, minneoverforingen fra CPUen til GPUen og for minneoverføringen CPU-GPU og tilbake og invertringen av bildet. Hele programmet bruker rundt $0,19$ns, minneoverføringen bruker

antall prosent tiden brukt til minneoverføringen er av hele kjøretiden: antall prosent tiden brukt til minneoverføringen er av selve oppaven til programmet:

DEt første man kan merke seg at I/O tar mye tid, så monneoverføringen er ikke den største tidstyven i dette programmet, men når man ser på CPU-GPU og tilbake og invertringen av bildet ser man at minneoverføring er en prosess som tar tid.