Evaluación del Trabajo Práctico Número 9

Arquitectura de Computadoras 25/06/2024

Instrucciones

- La presente evaluación es individual y podrá utilizar todo el material escrito del que disponga.
- La misma consta de 2 ejercicios en total y la duración será de 80 minutos.
- En cada ejercicio se indica el tiempo estimado para resolverlo y el puntaje que otorga.
- A pesar de que la suma de puntaje de todas las preguntas es 120 puntos, como ayuda adicional para aprobar, en ningún caso el puntaje máximo podrá superar los 100 puntos.
- La evaluación se desarrollará en las hojas provistas por la cátedra a tal fin.
- La buena presentación, ortografía y redacción serán tenidas en cuenta en la calificación.

Ejercicios para resolver

1) **35 Minutos - 55 Puntos (30, 25)**
Un programa realiza el procesamiento de un conjunto de datos. Cada conjunto de datos está compuesto por 16 palabras, y se tienen 512 conjuntos distintos, independientes y ubicados contiguamente en memoria. El procesamiento consiste en sumar las primeras 15 palabras de cada conjunto, y guardar el resultado en la decimosexta palabra.
Este programa se ejecutará en un procesador de dos núcleos, con un CPI ideal igual a 1. Cada núcleo demora 30 ciclos en realizar el procesamiento de un conjunto de datos sin considerar los fallos de caché.
Los núcleos cuentan con un caché de datos asociativo puro con bloques de 32 palabras, lo suficientemente grande para no tener fallos por capacidad, y en el que cada fallo tiene una penalización de 12 ciclos. Estos caches son de tipo snooping, e implementan el protocolo de coherencia MSI visto en clase.
   a) Si la tarea se divide de manera tal que un núcleo procesa todos los conjuntos impares y el otro todos los conjuntos pares, calcule cuánto demora el procesamiento de todos los conjuntos de datos. Recuerde que el procesamiento del conjunto incluye una escritura, y suponga que el primer núcleo arranca un ciclo antes que el segundo.
   b) Repita el apartado anterior, pero suponiendo ahora que la tarea se divide de manera tal que un núcleo procesa la primera mitad de los conjuntos y el otro la segunda mitad.

2) **45 Minutos - 65 Puntos (10, 20, 15, 20**
El entrenamiento de una red neuronal requiere efectuar el producto matricial gran cantidad de veces. Se dispone de una rutina optimizada que completa el cálculo del producto y acumulación de dos elementos de una matriz en 8T. Para completar un producto matricial se requiere, además, el almacenamiento del resultado, que demora 4T, y el cambio de fila y/o columna del siguiente resultado que demora 6T cada uno. De esta forma, el cálculo del producto matricial de una matriz de m filas y n columnas con una matriz de n filas y p columnas demora:

$$\Big[\big(8T \cdot n + 10T\big) \cdot m + 6T\Big] \cdot p$$

El entrenamiento se efectuará utilizando una computadora con 64 núcleos que operan a 4 Ghz y que ejecuta el programa de calculo con un IPC de 1,8.
   a) Si se ejecuta utilizando solo un núcleo del procesador ¿cuál es el tiempo, en segundos, que requeriría completar el producto matricial de una imagen full HD de 1920 x 1080 con un filtro de 1080 x 1080?.
   b) Para aprovechar todos los núcleos se propone dividir la matriz resultado en 64 partes independientes. Si el proceso de división y sincronización consume 1000T ¿cual es el tiempo, ahora en milisengundos, que se demora al efectuar el mismo calculo anterior? ¿Cuál es factor de aceleración?
   c) Si para completar el entrenamiento de la red se requiere repetir el calculo anterior 32 veces para un lote de 1.000.000 imágenes, ¿cuanto tiempo demora, en horas, el entrenamiento usando 64 núcleos?
   d) Si combinar los resultados de dos entrenamiento separados demora lo mismo que el producto de una imagen con un filtro, ¿cuanto tiempo se requiere, en segundo, para completar el entrenamiento anterior si se utiliza un cluster de 256 computadoras de 64 núcleos? ¿Cuál es factor de aceleración?

Soluciones propuestas

1) a) El procesamiento de todo el conjunto de datos demora 10788 ciclos.

   b) En este caso, el procesamiento de todo el conjunto de datos demora 9216 ciclos.

2)
   a) Demora 2,49s.
   b) Demora 38,92ms y se obtiene un factor de aceleración de 63,99.
   c) Demora 346h.
   d) Demora 4865,65s y se obtiene un factor de aceleración de 254,95.