Ejercicio tema 1

Jose Antonio Lorencio Abril

5. Procesadores de propósito general y de propósito específico. En este problema vamos a comparar el rendimiento y el consumo energético de un centro de datos que estuviera equipado con un procesador de propósito general (Haswell E5-2699 v3), una GPU de NVIDIA (la K80), y un ASIC de Google diseñado especialmente para cargas de Deep Learning y denominado TPU (v1). Las características hardware son las siguientes:

Sistema	**Chip**	**TDP**	**Consumo de potencia (reposo)**	**Consumo de potencia (carga)**
Propósito general	Haswell E5-2699 v3	504 W	159W	455W
GP-GPU	NVIDIA K80	1838 W	357W	991W
ASIC	TPU (v1)	861W	290W	384W

Y el rendimiento de cada sistema viene dado a continuación, donde A y B representa cargas de trabajo típicas de ejecución de inferencia en redes neuronales, y C es una carga de trabajo que representa una aplicación más general:

		Productividad (Throughput)		
Sistema	**Chip**	**A**	**B**	**C**
Propósito general	Haswell E5-2699 v3	5482	13194	12000
GP-GPU	NVIDIA K80	13461	36465	15000
ASIC	TPU (v1)	225000	280000	2000

Se pide:

- **Si el centro de datos descrito gasta el 70% de su tiempo ejecutando la carga de trabajo A y el resto en la carga de trabajo B, ¿cuál es la aceleración (speedup) de usar las TPUs sobre las GPUs?**

  Calculamos la mejora para cada carga de trabajo:

  $$m_A = \frac{225000}{13461} = 16.71$$

  $$m_B = \frac{280000}{36465} = 7.68$$

  y usamos Amdahl para obtener la aceleración

  $$acc = \frac{1}{\frac{0.7}{16.71} + \frac{0.3}{7.68}} = 12.35$$

- **Asumo que el consumo de potencia para cada procesador es una función lineal entre su valor en reposo (idle) y su carga (busy). Experimentalmente, hemos obtenido que para las cargas de trabajo del apartado anterior, el procesador de propósito general trabaja al 59.4%, las GPUs al 55.9% y las TPUs al 88.6%. ¿Cuál es el consumo de potencia para cada una de estas tres plataformas?**

  Para el Haswell, tenemos una potencia de $159 + 0.594 \cdot (455 - 159) = 334.824W$.

Para la GPU, es $347 + 0.559\,(991 - 357) = 711.406W$.

Para la TPU, es $290 + 0.886\,(384 - 290) = 373.284W$.

- **Pasado el tiempo, el centro de datos tiene la siguiente utilización: 40% en la carga de trabajo A, 10% en B, y 50% en C. ¿Cuál es la aceleración de usar GPUs o TPUs sobre la plataforma con procesadores de propósito general?**

  Hacemos como en el primer apartado. Primero lo calcularemos para la GPU:

$$m_A = \frac{13461}{5482} = 2.46$$

$$m_B = \frac{36465}{13194} = 2.76$$

$$m_C = \frac{15000}{12000} = 1.25$$

  y, por tanto

$$acc_{GPU} = \frac{1}{\frac{0.4}{2.46} + \frac{0.1}{2.76} + \frac{0.5}{1.25}} = 1.67$$

  Y para la TPU

$$m_A = \frac{225000}{5482} = 41.04$$

$$m_B = \frac{280000}{13194} = 21.22$$

$$m_C = \frac{2000}{12000} = 0.17$$

$$acc_{TPU} = \frac{1}{\frac{0.4}{41.04} + \frac{0.1}{21.22} + \frac{0.5}{0.17}} = 0.34$$