Universidad de Costa Rica
Escuela de Ciencias de la Computación e Informática
Programación Paralela y Concurrente
Prof. Esteban Rodríguez Betancourt

Proyecto 2: Page Rank

Para el segundo proyecto del curso, ustedes deberán programar el algoritmo PageRank. Este es un algoritmo propuesto por Larry Page y Sergey Brin para puntuar la importancia de un sitio web. Este algoritmo define la importancia de un sitio web en términos de cuantos links recibe. La fórmula es la siguiente:

$$PR(A) = (1-d) + d \cdot \sum \frac{PR(T_i)}{C(T_i)}$$

Donde:

- La página A es apuntada por las páginas de T_i
- d es *damping factor*. Típicamente se usa 0.85
- $C(A)$ es la cantidad de links que salen de la página A

Existen múltiples formas de calcular esta fórmula. Para este proyecto, vamos a utilizar la forma iterativa, la cual se define así:

$$PR(p_i; t) = \begin{cases} \frac{1}{N}, & t = 0 \\ \frac{1-d}{N} + d \cdot \sum_{p_j \in M(p_i)} \frac{PR(p_j; t-1)}{C(p_j)}, & t > 0 \end{cases}$$

En este caso, tenemos que, para la iteración inicial, todas las páginas web tendrán el mismo PageRank, mientras que en las iteraciones siguientes este valor va a ser refinado. Este proceso se repite hasta que los valores converjan, es decir, la diferencia entre dos pasos de tiempo es menor que algún épsilon. $M(p)$ es el conjunto de páginas que tienen un hipervínculo hacia p.
Para este proyecto deben elaborar un programa que calcule PageRank a partir de un archivo de texto. Este archivo va a tener el formato:
URL origen -> URL destino

El cálculo de esta fórmula debe ser realizado en paralelo, usando las técnicas que hemos visto en clase. El programa deberá escribir en un archivo una lista con las URL origen ordenadas alfabéticamente, seguidas por un espacio, dos puntos, un espacio y el valor de PageRank resultante.

Requisitos

El archivo de entrada será especificado mediante el flag -src
El archivo de salida será especificado mediante el flag -dst
El archivo de salida debe estar ordenado alfabéticamente por URL

El programa deberá hacer uso de paralelismo en cada iteración para acelerar el cómputo
Deben entregar documentación que explique el funcionamiento de su proyecto, cómo compilarlo, cómo correrlo, límites y problemas conocidos.

Evaluación

La evaluación se divide de la siguiente forma:

1. (75%) Implementación
    a. Puede leer correctamente el archivo de entrada
    b. Puede escribir correctamente el archivo de salida
    c. La salida es correcta
    d. El programa puede paralelizar correctamente la ejecución. Se espera que pueda escalar linealmente dado un incremento en la cantidad de núcleos
    e. No hay fugas de memoria ni accesos inválidos detectados por Valgrind
2. (25%) Documentación
    a. Explicación detallada del funcionamiento de su proyecto
    b. Explicación detallada de cómo compilar el proyecto
    c. Explicación detallada de cómo correrlo
    d. Explicación detallada de los límites de su proyecto
    e. Explicación detallada de los problemas conocidos con su proyecto

Avances

Aparte de la entrega principal del proyecto, deberán realizar avances de este. Los avances van a contar como quices en el rubro de ejercicios. Las fechas de los avances serán acordadas en clase.

Otros aspectos

1. El trabajo deberá ser realizado usando un sistema de control de versiones como git.
2. Podrá ser realizado en grupos de a lo sumo 3 personas.
3. Se espera que cada persona tenga una contribución equitativa en el proyecto. En caso de evidentes desbalances la nota individual será proporcional a la contribución de cada persona.
4. La entrega será el domingo 30 de junio a media noche.