

66.26 Arquitecturas paralelas

# Trabajo Práctico Final

**Integrantes:**

Alumno	padron
Llauró, Manuel Luis	95736
Blanco, Sebastian Ezequiel	98539

**GitHub:**

<https://github.com/BlancoSebastianEzequiel/66.26-TP-Final>

# Índice

<b>1. Objetivo</b>	<b>1</b>
<b>2. Desarrollo teorico</b>	<b>2</b>
2.1. Speed up	2
2.2. Ley de Amdahl	2
2.3. Ley de Gustafson	3
2.4. Map-reduce	3
<b>3. Implementacion</b>	<b>4</b>
3.1. Explicacion del modelo	4
3.2. Multiplicacion de matrices por bloques	4
3.2.0.1. Preprocesamiento	4
3.2.0.2. Mapeo	5
3.2.0.3. Reduccion	5
3.3. multiplicacion de matrices de elemento por fila	5
3.3.0.1. Preprocesamiento	5
3.3.0.2. Mapeo	5
3.3.0.3. Reduccion	5
3.4. Multiplicacion de matrices de columna por fila	6
3.4.0.1. Preprocesamiento	6
3.4.0.2. Mapeo	6
3.4.0.3. Reduccion	6
3.5. Forma de ejecucion	6
3.6. Datos sobre la computadora que se utilizó	6
<b>4. Resultados</b>	<b>8</b>
4.1. Multiplicacion por bloques	8
4.1.0.1. Salida Amdahl	8
4.1.0.2. Salida Gustafson	9
4.2. Multiplicacion elemento por fila	11
4.2.0.1. Salida Amdahl	11
4.2.0.2. Salida Gustafson	12
4.3. Multiplicacion columna por fila	15
4.3.0.1. Salida Amdahl	15
4.3.0.2. Salida Gustafson	16
<b>5. Conclusiones</b>	<b>19</b>

# 1. Objetivo

Se propone la verificación empírica de la ley de amdahl (trabajo constante) versus la ley de Gustafson (tiempo constante) aplicada a un problema de paralelismo utilizando el modelo de programación MapReduce.

Haremos una multiplicación de matrices (ambas de  $N \times N$ ) y se realizarán las mediciones de tiempo variando la cantidad de threads involucrados en el procesamiento. Luego se realizarán las mismas mediciones manteniendo fija la cantidad de threads pero variando la dimensión de las matrices.

## 2. Desarrollo teorico

### 2.1. Speed up

Es la mejora en la velocidad de ejecución de una tarea ejecutada en dos arquitecturas similares con diferentes recursos.

La noción de speedup fue establecida por la ley de Amdahl, que estaba dirigida particularmente a la computación paralela. Sin embargo, la speedup se puede usar más generalmente para mostrar el efecto en el rendimiento después de cualquier mejora en los recursos.

De forma genérica se define como:

$$speed\_up = \frac{Rendimiento\_con\_mejora}{Rendimiento\_sin\_mejora} \quad (1)$$

En el caso de mejoras aplicadas a los tiempo de ejecución de una tarea:

$$speed\_up = \frac{T\_ejecucion\_sin\_mejora}{T\_ejecucion\_con\_mejora} \quad (2)$$

### 2.2. Ley de Amdahl

Utilizada para averiguar la mejora máxima de un sistema de información cuando solo una parte de éste es mejorado.

Establece que la mejora obtenida en el rendimiento de un sistema debido a la alteración de uno de sus componentes está limitada por la fracción de tiempo que se utiliza dicho componente.

Suponiendo que nuestro algoritmo se divide en una parte secuencial **s** u una parte paralelizable **p** y siendo **N** la cantidad de threads, entonces podemos decir que:

$$speed\_up = \frac{s + p}{s + \frac{p}{N}} \quad (3)$$

Amdahl establece un límite superior al speedup que puede obtenerse al introducir una mejora en un determinado algoritmo. Este límite superior está determinado por la porción de la tarea sobre la que se aplique la mejora. Entonces si tomamos la ecuacion anterior y calculamos el limite de la misma con **N** tendiendo a infinito tenemos:

$$speed\_up\_max = 1 + \frac{p}{s} \quad (4)$$

### 2.3. Ley de Gustafson

Establece que cualquier problema suficientemente grande puede ser eficientemente paralelizado. La ley de Gustafson está muy ligada a la ley de Amdahl, que pone límite a la mejora que se puede obtener gracias a la paralelización, dado un conjunto de datos de tamaño fijo, ofreciendo así una visión pesimista del procesamiento paralelo. Por el contrario la ley de Gustafson propone realizar mas trabajo con la misma cantidad de recursos, de esta manera aprovecho la paralelizacion para calcular mas cosas.

Entonces siendo  $s$  el tiempo de la ejecucion de la seccion serie, siendo  $p$  el tiempo de la ejecucion de la seccion paralela y siendo  $N$  la cantidad de procesadores podemos calcular el speed up como:

$$speed\_up = \frac{s + p * N}{s + p} \quad (5)$$

### 2.4. Map-reduce

MapReduce es una técnica de procesamiento y un programa modelo de computación distribuida. El algoritmo MapReduce contiene dos tareas importantes.

Map toma un conjunto de datos y se convierte en otro conjunto de datos, en el que los elementos se dividen en tuplas (pares: clave, valor).

Reduce toma la salida de un mapa como entrada y combina los datos tuplas en un conjunto más pequeño de tuplas.

La principal ventaja de MapReduce es que es fácil de escalar procesamiento de datos en múltiples nodos.

De acuerdo a este modelo, basado en la programación funcional, la tarea del usuario consiste en la definición de una función map y una función reduce y definidas estas funciones, el procesamiento es fácilmente paralelizable, ya sea en una sola máquina o en un cluster.

## 3. Implementacion

### 3.1. Explicacion del modelo

La implementación del MapReduce para resolver el problema esta basado en el siguiente esquema:



Figura 1: Esquema de un map reduce

En nuestro caso creamos una clase llamada `MapReduce` la cual usa una librería de `python` llamada `multiprocessing` en donde usamos el modulo `pool` el cual ofrece un medio conveniente para paralelizar la ejecución de una función a través de múltiples valores de entrada, distribuyendo los datos de entrada a través de procesos (paralelismo de datos).

Entonces lo que hicimos fue instanciar dos `pool`, uno para hacer el map y el otro para el reduce de manera que el primero se le pasa como atributo la cantidad de worker en el cual se quiere paralelizar el problema y el segundo solo se usa uno de manera tal que la fase de reduce se la serie.

### 3.2. Multiplicacion de matrices por bloques

#### 3.2.0.1 Preprocesamiento

Generamos una lista de tuplas donde cada una tiene la posición `(r, c)` de un bloque de la matriz A, tiene el bloque en cuestión `a_block_rc`, y la fila número `c` de bloques de la matriz B, quedando con este formato:

```
(r, c, a_block_rc, b_block_c)
```

### 3.2.0.2 Mapeo

Recibimos la posición  $r$ ,  $c$  del bloque  $a$ , el bloque  $a$  y una lista de bloques  $b$  que es la fila  $c$  de bloques en la matriz B.

Entonces multiplicamos el bloque  $a$  por cada bloque de la lista de bloques  $b$  y guardamos en un vector una tupla con una clave  $r$ ,  $c_b$  donde  $c_b$  es el índice en la lista de bloques  $b$  y como valor guardamos la multiplicación. Por cada multiplicación, agregamos una de estas tuplas al vector de salida para luego devolver este.

### 3.2.0.3 Reduccion

Recibimos la posición de un bloque de salida y una lista de multiplicaciones parciales de bloques. Se suman estas multiplicaciones parciales y se devuelve un vector con los valores resultantes de la multiplicación. Pero por cada valor se calcula la posición de salida del mismo en la matriz resultante y nos deshacemos de la posición de los bloques

## 3.3. multiplicacion de matrices de elemento por fila

### 3.3.0.1 Preprocesamiento

Consiste en generar una lista de tuplas a partir de las dos matrices. Se itera por cada elemento ( $a_{ij}$ ) de la matriz A y se guarda en cada tupla el número de fila del elemento  $a_{ij}$ , el elemento  $a_{ij}$  y la fila  $j$  de la matriz B.

### 3.3.0.2 Mapeo

De esta manera, en la función map, obtenemos partes de esta lista de tuplas y devolvemos un par clave, valor donde la clave es la posición de salida de la matriz resultante  $(i, j)$  y el valor es la multiplicación del elemento  $a_{ij}$  contra cada elemento de la fila  $j$  de la matriz B

### 3.3.0.3 Reduccion

Obtenemos una posición de salida y una lista de valores que resultaron de la multiplicación que se hizo en el map. Entonces se suman las multiplicaciones parciales y se obtiene el valor en la posición de salida de la matriz resultante

## 3.4. Multiplicacion de matrices de columna por fila

### 3.4.0.1 Preprocesamiento

Consiste en generar una lista de tuplas a partir de las dos matrices. Se guarda en cada tupla la columna `i` de la matriz A y la fila `i` de la matriz B

### 3.4.0.2 Mapeo

Recibimos una columna de la matriz A y una fila de la matriz B y por cada elemento de la columna `elem_a` lo multiplicamos por cada elemento de la fila `elem_b` obteniendo una matriz parcial de la multiplicacion. por cada multiplicacion guardamos en un vector una tupla con un par clave valor donde la clave es la posicion de salida de la matriz resultante y el valor es la multiplicacion anteriormente mencionada. Finalmente se devuelve el vector de tuplas.

### 3.4.0.3 Reduccion

Se recibe la posicion de salida de la matriz resultante y una lista de multiplicaciones parciales. Entonces se suman estas y se devuelve la posicion de salida y la suma.

## 3.5. Forma de ejecucion

Para el caso de Amdahl multiplicamos dos matrices de `10x10` con `1`, `2`, `4`, `8`, `16` y `32` threads.

Para el caso de gustafson se usan siempre 4 threads multiplicando dos matrices de `2x2`, `4x4`, `8x8`, `16x16`, `32x32` y `64x64`

Para realizar el calculo se debe ejecutar:

```
$ sh scripts/run.sh.
```

Luego para generar los graficos que vemos en el informe se debe ejecutar:

```
$ sh scripts/generate_output_data.sh
```

## 3.6. Datos sobre la computadora que se utilizó

El equipo sobre el que se realizarán las mediciones es una laptop con un procesador Intel core I7 que posee 4 nucleos a 2.7 Ghz, es decir, soporta hasta 4 threads en paralelo, con 16 Gb de memoria y corriendo sobre un sistema Linux.

Para averiguar estos datos en linux se ejecutaron los siguientes comandos:

- Cantidad de cores: `$ grep -c processor /proc/cpuinfo`



- Velocidad de reloj: `$ lscpu | grep GHz`
- Memoria RAM: `$ free -g`

## 4. Resultados

### 4.1. Multiplicacion por bloques

#### 4.1.0.1 Salida Amdahl

	number_of_threads	parallel_time	serial_time	matrix_dimension
0	1	50205.847025	18839.280605	300
1	2	31137.020111	19088.112116	300
2	3	28101.762295	19261.277676	300
3	4	25487.717390	24137.875080	300
4	8	27197.943211	24288.087368	300
5	16	26089.009523	23751.881361	300
6	32	28919.217348	19207.900763	300

Figura 2: Salida de los tiempos en serie y paralelo

De acuerdo a estos datos podemos calcular el speed up maximo, real y teórico.

	number_of_threads	theoretical_speed_up	real_speed_up	max_speed_up
0	1	1.000000	1.000000	3.664956
1	2	1.571271	1.449222	3.664956
2	3	1.940857	1.654400	3.664956
3	4	2.199538	1.626546	3.664956
4	8	2.749158	1.859518	3.664956
5	16	3.141679	1.963597	3.664956
6	32	3.383203	2.393000	3.664956

Figura 3: Speed up real, teorico y maximo segun la cantidad de threads

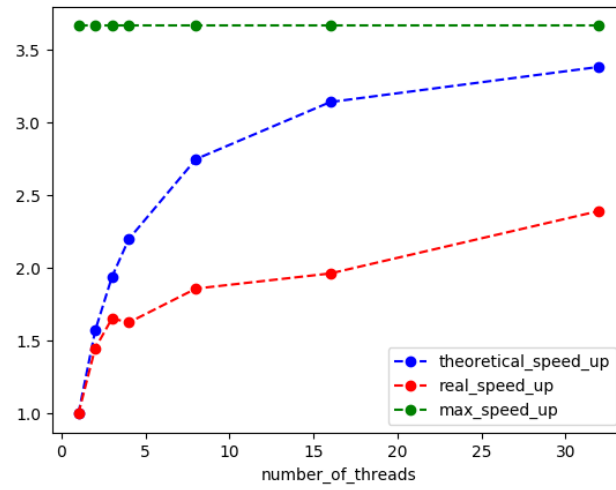


Figura 4: Grafico

Podemos observar que

#### 4.1.0.2 Salida Gustafson

	number_of_threads	parallel_time	serial_time	matrix_dimension
0	4	33.545256	13.228893	2
1	4	39.066076	13.394833	4
2	4	60.608625	13.423681	8
3	4	42.693377	14.817476	16
4	4	64.430237	22.891760	32
5	4	221.577644	91.615438	64
6	4	728.224516	305.559397	100
7	4	27029.505491	19489.433289	300

Figura 5: Salida de los tiempos en serie y paralelo con el error

Podemos ver que

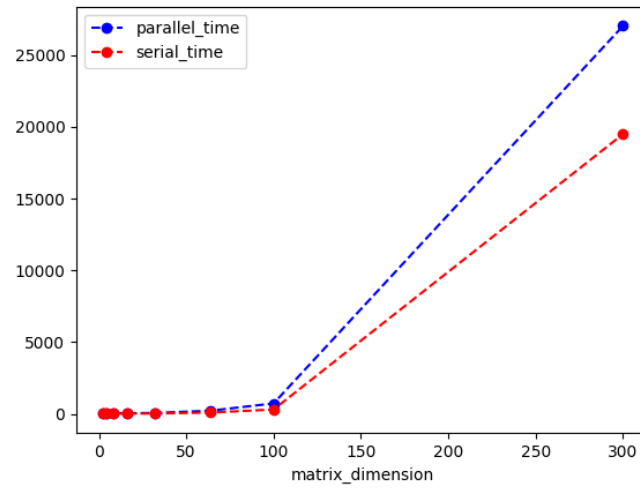


Figura 6: Tiempo paralelo y serie en funcion de la dimension de las matrices de entrada

Luego a partir de estos datos podemos calcular el speed up y obtuvimos lo siguiente:

	matrix_dimension	speed_up
<b>0</b>	2	3.151525
<b>1</b>	4	3.234011
<b>2</b>	8	3.456034
<b>3</b>	16	3.227060
<b>4</b>	32	3.213540
<b>5</b>	64	3.122438
<b>6</b>	100	3.113279
<b>7</b>	300	2.743129

Figura 7: Tabla de valores del speed up

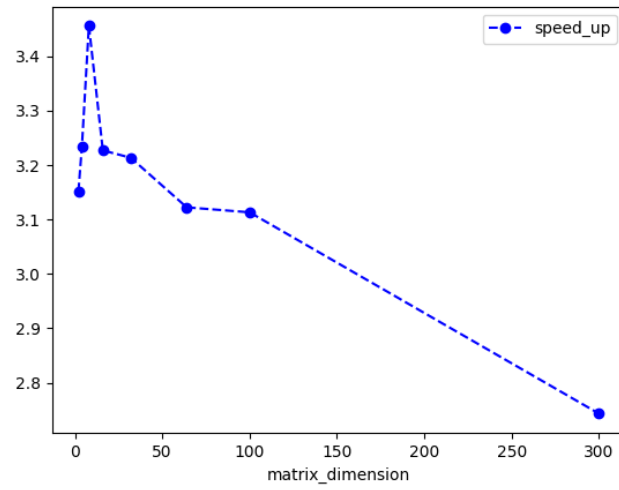


Figura 8: Grafico del speed up

## 4.2. Multiplicacion elemento por fila

### 4.2.0.1 Salida Amdahl

	number_of_threads	parallel_time	serial_time	matrix_dimension
0	1	47537.959337	5190.705538	300
1	2	29165.563345	5323.356152	300
2	3	24212.626219	4589.113474	300
3	4	23993.258238	5441.639662	300
4	8	22454.158545	5337.180376	300
5	16	22864.304304	4607.271910	300
6	32	22783.949137	5616.553307	300

Figura 9: Salida de los tiempos en serie y paralelo

De acuerdo a estos datos podemos calcular el speed up maximo, real y teórico.

	number_of_threads	theoretical_speed_up	real_speed_up	max_speed_up
0	1	1.000000	1.000000	10.158285
1	2	1.820761	1.732577	10.158285
2	3	2.506509	2.275021	10.158285
3	4	3.088027	2.572991	10.158285
4	8	4.736271	3.412513	10.158285
5	16	6.460399	4.551069	10.158285
6	32	7.897927	4.487678	10.158285

Figura 10: Speed up real, teorico y maximo segun la cantidad de threads

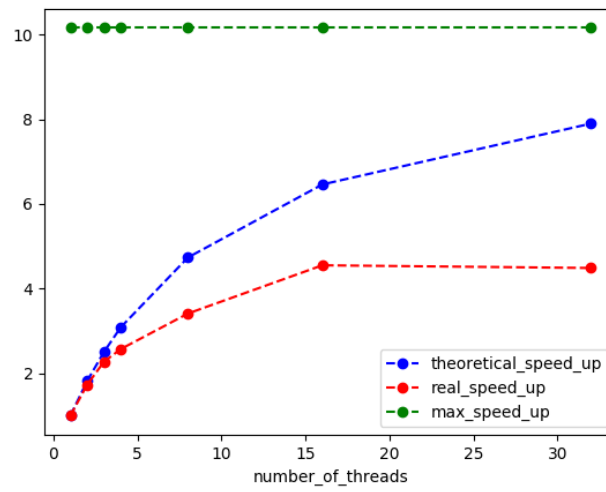


Figura 11: Grafico

Podemos observar que

#### 4.2.0.2 Salida Gustafson

	number_of_threads	parallel_time	serial_time	matrix_dimension
0	4	34.696817	28.255463	2
1	4	38.287878	12.956142	4
2	4	34.617901	25.972605	8
3	4	64.780474	28.531790	16
4	4	59.980869	15.392780	32
5	4	232.058287	47.659159	64
6	4	795.195341	230.800629	100
7	4	23386.375189	5251.749754	300

Figura 12: Salida de los tiempos en serie y paralelo con el error

Podemos ver que

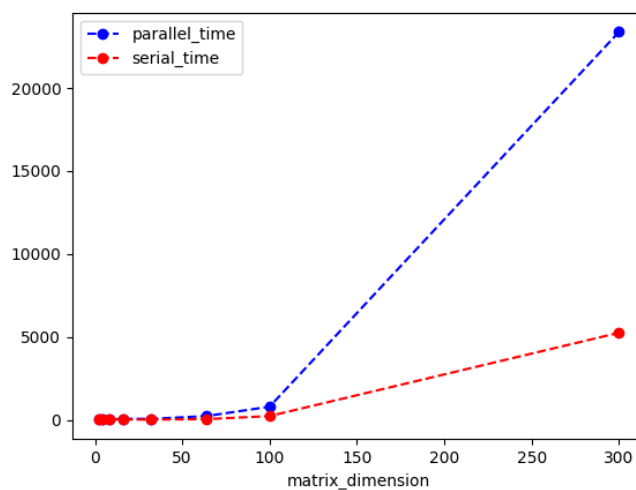


Figura 13: Tiempo paralelo y serie en funcion de la dimension de las matrices de entrada

Luego a partir de estos datos podemos calcular el speed up y obtuvimos lo siguiente:

	<b>matrix_dimension</b>	<b>speed_up</b>
<b>0</b>	2	2.653482
<b>1</b>	4	3.241503
<b>2</b>	8	2.714026
<b>3</b>	16	3.082700
<b>4</b>	32	3.387341
<b>5</b>	64	3.488850
<b>6</b>	100	3.325142
<b>7</b>	300	3.449851

Figura 14: Tabla de valores del speed up

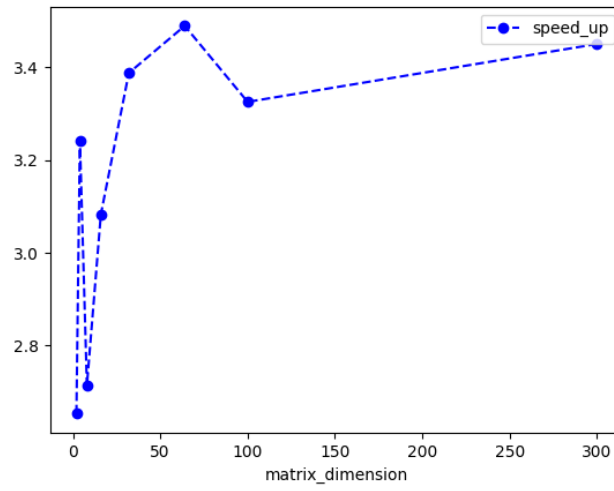


Figura 15: Grafico del speed up



### 4.3. Multiplicacion columna por fila

#### 4.3.0.1 Salida Amdahl

	number_of_threads	parallel_time	serial_time	matrix_dimension
0	1	60482.304811	5493.925095	300
1	2	31412.675858	6107.174158	300
2	3	24666.881800	5500.903606	300
3	4	23425.884247	6119.422197	300
4	8	21565.107822	5516.240358	300
5	16	20271.022797	6139.587879	300
6	32	19509.930372	5528.530836	300

Figura 16: Salida de los tiempos en serie y paralelo

De acuerdo a estos datos podemos calcular el speed up maximo, real y teórico.

	number_of_threads	theoretical_speed_up	real_speed_up	max_speed_up
0	1	1.000000	1.000000	12.008942
1	2	1.846260	1.720028	12.008942
2	3	2.571702	2.198306	12.008942
3	4	3.200477	2.467065	12.008942
4	8	5.054018	3.297826	12.008942
5	16	7.114054	3.565856	12.008942
6	32	8.935029	4.079110	12.008942

Figura 17: Speed up real, teorico y maximo segun la cantidad de threads

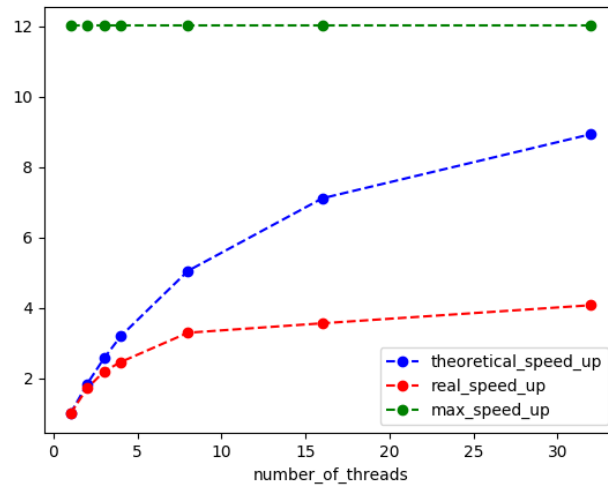


Figura 18: Grafico

Podemos observar que

#### 4.3.0.2 Salida Gustafson

	number_of_threads	parallel_time	serial_time	matrix_dimension
<b>0</b>	4	33.103228	26.368618	2
<b>1</b>	4	38.636208	12.801409	4
<b>2</b>	4	38.946629	26.823282	8
<b>3</b>	4	41.731834	11.095762	16
<b>4</b>	4	91.148853	15.094280	32
<b>5</b>	4	278.873682	50.332308	64
<b>6</b>	4	756.765842	208.662271	100
<b>7</b>	4	23710.624456	6019.118547	300

Figura 19: Salida de los tiempos en serie y paralelo con el error

Podemos ver que

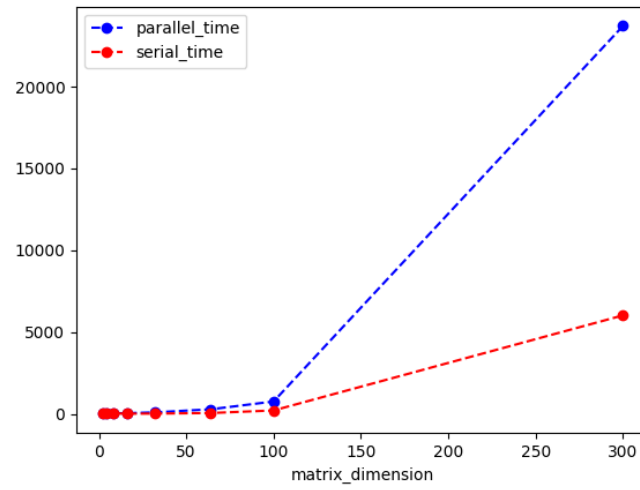


Figura 20: Tiempo paralelo y serie en funcion de la dimension de las matrices de entrada

Luego a partir de estos datos podemos calcular el speed up y obtuvimos lo siguiente:

	<b>matrix_dimension</b>	<b>speed_up</b>
<b>0</b>	2	2.669860
<b>1</b>	4	3.253382
<b>2</b>	8	2.776495
<b>3</b>	16	3.369888
<b>4</b>	32	3.573781
<b>5</b>	64	3.541330
<b>6</b>	100	3.351597
<b>7</b>	300	3.392616

Figura 21: Tabla de valores del speed up

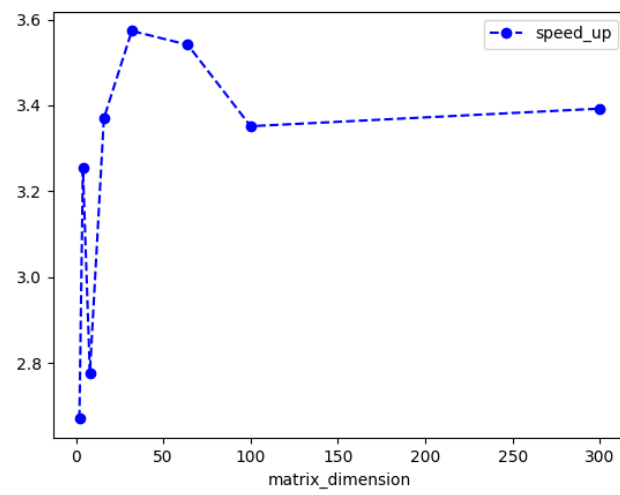


Figura 22: Grafico del speed up

## 5. Conclusiones

Podemos decir que nuestros resultados no son lo que debería ocurrir en un ambiente ideal donde solo nuestro proceso corre. En la parte de amdahl, el speedup sigue aumentando aún cuando la cantidad de threads de procesamiento supera la cantidad de threads del sistema (4). Esto no es lo que se esperaba. Ocurre que al aumentar la cantidad de threads, la sección serie del problema no se mantuvo sino que tuvo una pequeña variación equilibrando la caída de performance de la sección paralela. Esta variación en el tiempo de la parte serie del problema se debe a la forma en que quedan organizados los datos una vez que son procesados por los `map_workers`.