



INFORME COMPARATIVO: CONVERSIÓN DE URLS A PDF CON MÚLTIPLES HILOS

Sistemas Distribuidos

Diego Jose Torres Rangel
Universidad Pontificia Bolivariana

Informe Comparativo: Conversión de URLs a PDF con Múltiples Hilos

1. Objetivo del Estudio

Analizar el impacto del paralelismo (número de hilos) en el rendimiento de un programa Java para la conversión de documentos web a PDF utilizando Headless Chrome, con una muestra de 32 URLs.

2. Especificaciones del Hardware

- **Procesador (CPU):**
 - Modelo: AMD Ryzen 5 4600H con gráficos Radeon integrados
 - Núcleos: 6
 - Hilos lógicos: 12
 - Velocidad base: 3.00 GHz (operando a 3.40 GHz)
 - Caché L1: 384 KB
 - Caché L2: 3 MB
 - Caché L3: 8 MB
 - Virtualización: Habilitada
- **Memoria RAM:**
 - 16 GB
- **Almacenamiento:**
 - Unidad de estado sólido (SSD) de 1 TB

3. Metodología

Se desarrolló un programa en Java que realiza las siguientes operaciones:

1. Recibe una lista de 32 URLs como entrada
2. Utiliza un número configurable de hilos para procesar las URLs en paralelo
3. Emplea Headless Chrome para convertir cada página web a formato PDF
4. Mide el tiempo total de ejecución para la conversión de todas las URLs

Las pruebas se ejecutaron variando el número de hilos de 1 a 16, manteniendo constantes las demás variables.

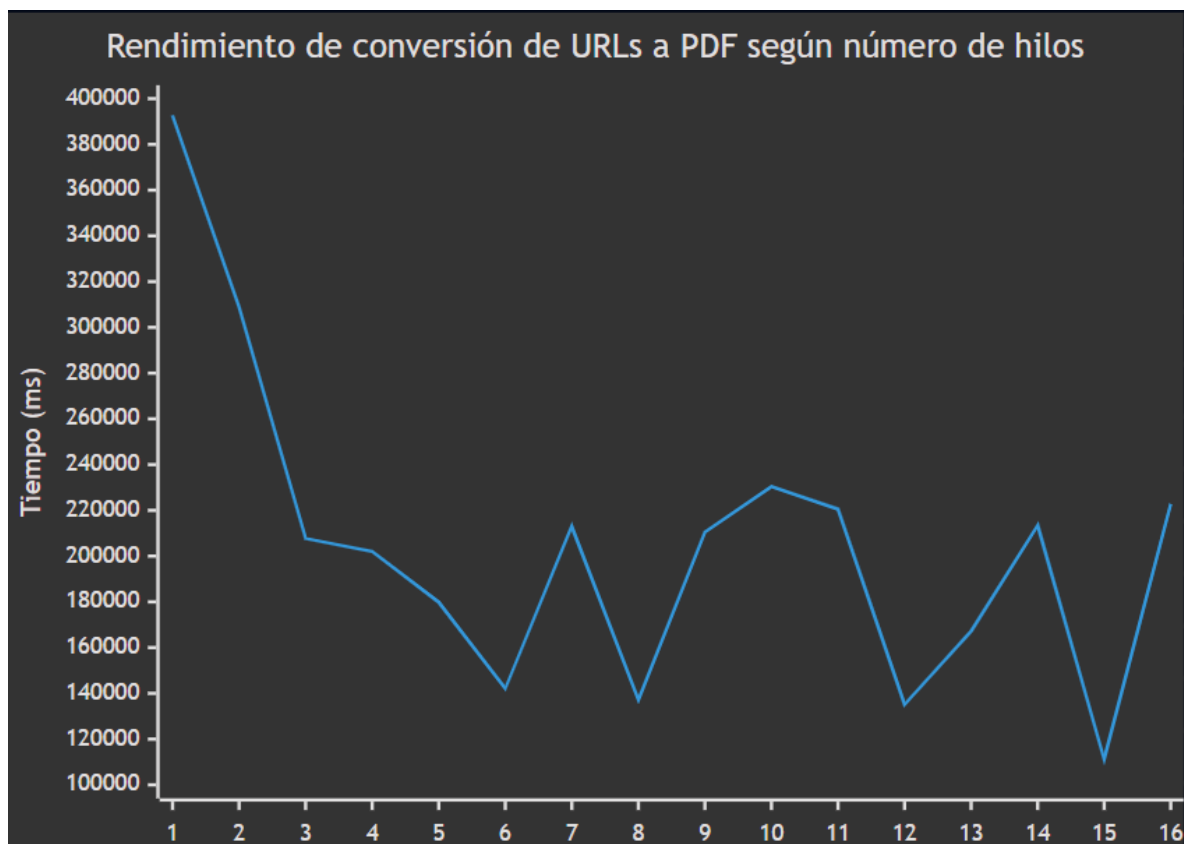
4. Resultados Obtenidos

4.1 Tabla de Resultados

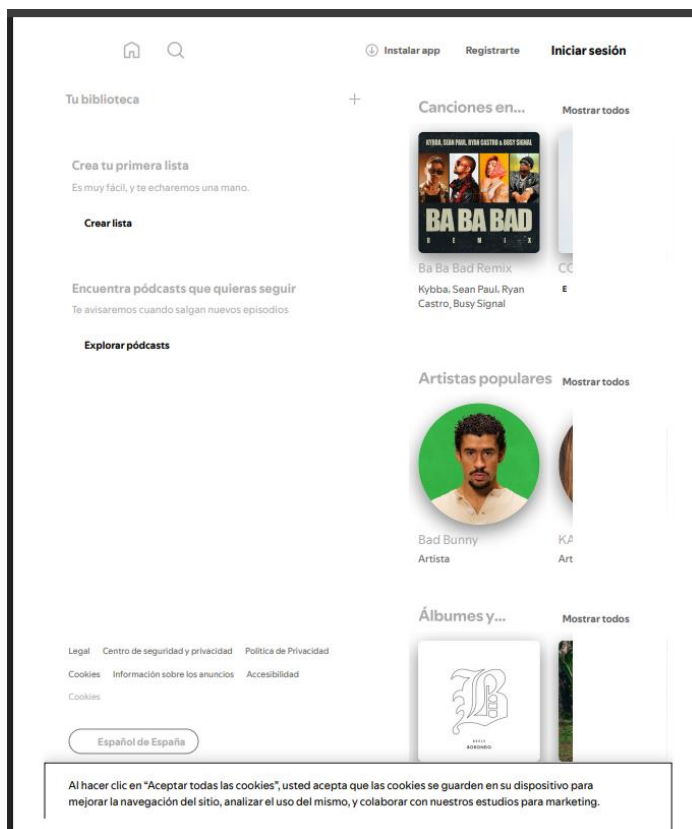
Número de hilos	Tiempo (ms)	Tiempo (s)	PDFs generados	% Completitud
1	392,747	392.7	32	100%
2	309,070	309.1	31	96.9%
3	207,705	207.7	32	100%
4	202,095	202.1	32	100%
5	179,844	179.8	32	100%
6	142,152	142.2	32	100%
7	213,015	213.0	31	96.9%
8	137,204	137.2	32	100%
9	210,512	210.5	31	96.9%
10	230,372	230.4	31	96.9%
11	220,548	220.5	30	93.8%
12	135,130	135.1	31	96.9%
13	167,154	167.2	29	90.6%
14	213,429	213.4	30	93.8%
15	111,212	111.2	31	96.9%
16	222,762	222.8	27	84.4%

Nota: Durante las pruebas con 15 y 16 hilos, se observó que la CPU operaba al 100% de su capacidad y la memoria RAM alcanzó el 93% de utilización.

4.2 Gráfico Comparativo



5. pdf página Spotify



6. Capturas de Resultados por Número de Hilos

```
=== Resultados ===  
Tiempo total: 392747 ms  
Archivos PDF generados: 32  
Hilos utilizados: 1  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 309070 ms  
Archivos PDF generados: 31  
Hilos utilizados: 2  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 207705 ms  
Archivos PDF generados: 32  
Hilos utilizados: 3  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 202095 ms  
Archivos PDF generados: 32  
Hilos utilizados: 4  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 179844 ms  
Archivos PDF generados: 32  
Hilos utilizados: 5  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 142152 ms  
Archivos PDF generados: 32  
Hilos utilizados: 6  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 213015 ms  
Archivos PDF generados: 31  
Hilos utilizados: 7  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 137204 ms  
Archivos PDF generados: 32  
Hilos utilizados: 8  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 210512 ms  
Archivos PDF generados: 31  
Hilos utilizados: 9  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 230372 ms  
Archivos PDF generados: 31  
Hilos utilizados: 10  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 220548 ms  
Archivos PDF generados: 30  
Hilos utilizados: 11  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 135130 ms  
Archivos PDF generados: 31  
Hilos utilizados: 12  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 167154 ms  
Archivos PDF generados: 29  
Hilos utilizados: 13  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 213429 ms  
Archivos PDF generados: 30  
Hilos utilizados: 14  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 111212 ms  
Archivos PDF generados: 31  
Hilos utilizados: 15  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

```
=== Resultados ===  
Tiempo total: 222762 ms  
Archivos PDF generados: 27  
Hilos utilizados: 16  
Directorio de salida: ./pdfs  
Resultados guardados en resultados.csv
```

7. Análisis de Resultados

1. **Tendencia general:** Se observa una reducción significativa en el tiempo de ejecución al aumentar el número de hilos, aunque con importantes fluctuaciones.
2. **Puntos óptimos de rendimiento:**
 - Con 6 hilos: 142,152 ms (100% de completitud)
 - Con 8 hilos: 137,204 ms (100% de completitud)
 - Con 12 hilos: 135,130 ms (96.9% de completitud)
 - Con 15 hilos: 111,212 ms (96.9% de completitud)
3. **Relación con la arquitectura:** Los resultados muestran mejoras de rendimiento significativas hasta aproximadamente 6 hilos, lo que corresponde al número de núcleos físicos del procesador. Se observan ganancias adicionales con 8, 12 y 15 hilos.
4. **Trade-off entre velocidad y completitud:** A medida que aumenta el número de hilos, se evidencia una tendencia a la disminución en el número de PDFs generados correctamente. Con 16 hilos, solo se completaron 27 de los 32 PDFs (84.4%).
5. **Saturación de recursos:** Las pruebas con 15 y 16 hilos llevaron al sistema a operar al 100% de CPU y 93% de memoria, evidenciando la saturación de recursos. Si bien con 15 hilos se logró el mejor tiempo, con 16 hilos el rendimiento se degradó significativamente.
6. **Errores de conexión:** Se observaron errores de tipo "Connection reset" que explican la variabilidad en el número de PDFs generados, especialmente con un alto número de hilos. Estos errores son consecuencia de las limitaciones de red y los mecanismos de protección de servidores web ante múltiples conexiones simultáneas.

8. Conclusiones

1. **Paralelismo óptimo:** Para este caso específico y con el hardware utilizado, el rendimiento óptimo se alcanzó con 15 hilos, logrando completar la tarea en 111,212 ms (aproximadamente 1.85 minutos), casi 4 veces más rápido que con un solo hilo.
2. **Punto de equilibrio:** Si se prioriza tanto el rendimiento como la fiabilidad, el uso de 6 a 8 hilos representa el mejor compromiso, con tiempos de ejecución aproximadamente 2.8 veces más rápidos que con un solo hilo y 100% de completitud en la generación de PDFs.

3. **Ley de rendimientos decrecientes:** Los resultados demuestran que aumentar el paralelismo más allá del número de núcleos físicos (6) y lógicos (12) del procesador no siempre produce mejoras proporcionales y puede incluso degradar el rendimiento.
4. **Limitaciones externas:** El rendimiento del sistema está limitado no solo por los recursos de hardware, sino también por factores externos como la velocidad de conexión a internet, la respuesta de los servidores web y los mecanismos de protección contra múltiples conexiones simultáneas.
5. **Recomendación final:** Para aplicaciones similares que realicen tareas de procesamiento web intensivas, se recomienda establecer el número de hilos cercano pero no superior al número de núcleos lógicos del procesador, considerando también la posible degradación en la fiabilidad al aumentar excesivamente el paralelismo.

9. Consideraciones Adicionales

Para futuros estudios, sería valioso:

- Implementar mecanismos de reintento para conexiones fallidas
- Analizar el impacto de diferentes velocidades de conexión a internet
- Evaluar el rendimiento con diferentes tipos de páginas web (simples vs. complejas)

