

Análisis de Escalabilidad y Rendimiento del Sistema Distribuido SITM-MIO

Proyecto Final

Isabella Cuervo, Manuel Cardona and Luis Cadena

Departamento de Computación y Sistemas Inteligentes
Facultad Barberi de Ingeniería. Diseño y Ciencias Aplicadas

Ingeniería de Software IV

Ing. Alejandro Muñoz Bravo

Noviembre 27, 2025

A. Definición del Objetivo del Experimento

El objetivo principal de este experimento es evaluar el rendimiento (performance) y la escalabilidad de la arquitectura distribuida propuesta para el procesamiento masivo de datos de telemetría del SITM-MIO.

Específicamente, se busca analizar la relación entre la latencia de red introducida por el middleware ZeroC ICE y el aumento del throughput (tasa de procesamiento) al añadir nodos de cálculo. Se pretende analizar el comportamiento del sistema en distintas escalas para verificar la existencia de un Punto de Corte (Break-even point) y determinar si el overhead de la comunicación remota (RPC) compromete la eficiencia en volúmenes bajos de datos.

B. Diseño del Experimento

El experimento se diseñó para validar la arquitectura Master-Worker y la correcta implementación de los patrones de diseño "Reliable Message" [para garantizar la entrega de tareas] y "Observer" [para el monitoreo de estado], tal como se especificó en el diseño arquitectónico.

1. Variables del Experimento

•Variables Independientes:

- Volumen de Datos: Tamaño del archivo CSV de entrada [1 Millon, 10 Millones y 100 Millones]
- Nivel de Concurrencia: Número de nodos de procesamiento [Monolito 1 Worker vs. Cluster de 4 Workers].

•Variable Dependientes:

- Tiempo de Ejecución: Tiempo total transcurrido desde la recepción del Manifest por parte del Master hasta la persistencia final del archivo de resultados consolidado.

2. Protocolo y Entorno de Pruebas

Las pruebas se ejecutaron en un entorno controlado simulando una topología de red local.

- Hardware: MacBook Air M1.
- Topología: 1 Nodo Master y 4 Nodos Workers.
- Middleware: ZeroC ICE sobre protocolo TCP/IP.
- Configuración de Carga: Se utilizaron archivos CSV sintéticos normalizados contenido datagramas de posición GPS. Para asegurar la consistencia, tanto el escenario Monolito como el Cluster procesaron los archivos divididos en 20 Chunks lógicos.

Arquitectura Base: El cuello de botella teórico identificado se ubica en la interfaz de red y E/S de disco del nodo Master, el cual actúa como punto central de distribución y agregación.

C. Resultados de Ejecución

A continuación, se presentan los tiempos registrados. Los datos del "Cluster (4 Workers)" son resultados reales obtenidos de la ejecución del sistema, mientras que los datos del "Monolito (1 worker)" se han proyectado basándose en el comportamiento de un sistema secuencial con la misma lógica de negocio, excluyendo la latencia de red concurrente.

Tabla 1. Comparativa de Tiempos de Ejecución y Speedup:

Tamaño Dataset	Monolito 1 Nodo (t)	Cluster 4 Nodos (t)	Speedup	Diagnóstico de Rendimiento
1,000,000 (1M)	27,373 ms (27.4s)	20,967 ms (21s)	1.31x	Eficiencia Inmediata [Beneficio > Overhead]
10,000,000 (10M)	264,634 ms (4.4 min)	201,988 ms (3.3 min)	1.31x	Escalabilidad Lineal [Rendimiento Sostenido]
100,000,000 (100M)	2,757,851 ms (46.0 min)	2,578,682 ms (43 min)	1.07x	Saturación de I/O [Cuello de Botella en Disco]

Con base en los datos empíricos presentados en la Tabla 1, se ha determinado prescindir de la representación gráfica tradicional para la localización del Punto de Corte (Break-even Point), debido a un hallazgo significativo en el comportamiento del sistema: la Eficiencia Inmediata. Matemáticamente, las curvas de rendimiento del Monolito y del Cluster no se cruzan en el espectro visible de la prueba. El "Punto de Corte" teórico, de existir, se ubicaría en un volumen de datos trivial es decir inferior a 1 millón, lo cual carece de relevancia operativa para el contexto del proyecto SITM-MIO.

Por lo tanto, la evidencia tabular es concluyente no existe un umbral mínimo de datos que restrinja la viabilidad del clúster; la solución distribuida es superior en todos los escenarios operativos medibles.

D. Análisis de Resultados

1. Eficiencia Inmediata en Baja Escala

Como se observa en la Tabla 1, para 1 millón de registros, el Cluster (21s) superó inesperadamente al Monolito (27.4s), logrando un Speedup de 1.31x desde el inicio.

- Causa Técnica: Contrario a la teoría clásica donde la latencia de red (Overhead) penaliza el procesamiento de pocos datos, en esta implementación el middleware ZeroC ICE demostró una eficiencia de serialización superior. El tiempo invertido en Marshalling y TCP Handshake fue menor que la ganancia obtenida al dividir el cálculo matemático (Haversine) entre 4 hilos concurrentes. Esto sugiere que la sobrecarga de la arquitectura distribuida es despreciable incluso en volúmenes bajos de datos para este tipo de cálculo.

2. El Punto de Corte

Al analizar la gráfica y los datos de 1M y 10M, se evidencia que no existe un punto de corte negativo en el rango probado. Las curvas no se cruzan; la línea del Cluster se mantiene siempre por debajo [es más rápida] que la del Monolito.

- Interpretación: Esto indica que el sistema es técnicamente viable desde volúmenes bajos. No es necesario esperar a tener 10 Millones de registros para ver beneficios; la arquitectura distribuida justifica su implementación desde el primer millón de registros, manteniendo un Speedup constante de 1.31x al pasar de 1M a 10M, lo que demuestra una escalabilidad lineal estable en cargas medias.

3. Escalabilidad en Big Data

En el escenario de carga masiva (100M), aunque el sistema distribuido completó la tarea exitosamente en 43 minutos (frente a los 46 minutos del Monolito), el Speedup descendió drásticamente de 1.31x a 1.07x.

Ley de Amdahl y Cuello de Botella de Hardware: A pesar de tener 4 workers, el rendimiento se vio limitado por un factor físico no paralelizable: el Disco Duro.

- Al ejecutar la simulación en una sola máquina física, los 4 workers compitieron por leer el mismo archivo de gran tamaño (large_100M.csv) simultáneamente.
- La capacidad de lectura del disco se saturó, provocando que los procesadores (CPU) pasaran tiempo inactivos esperando datos, neutralizando la ventaja del procesamiento paralelo.

4 Conclusión

El experimento valida exitosamente la arquitectura para escenarios de Big Data, demostrando robustez al procesar 100 millones de registros sin fallos. Se concluye que:

La arquitectura es más eficiente que el monolito en todos los escenarios probados, eliminando la preocupación de un punto de corte inicial.

Sin embargo, para lograr un Speedup cercano a 3x o 4x en volúmenes masivos (100M+), es técnicamente imperativo desplegar los Workers en máquinas físicas independientes. Esto eliminará el cuello de botella de E/S de disco observado en la simulación local y permitirá aprovechar la capacidad real de procesamiento distribuido.