Aligning Market and Index Vectors

An implementation of the vector virtual machine for financial applications

Sonia Kolasinska

November 26, 2025

Abstract

This document summarizes several implementation strategies for core vector alignment within the WebAssembly (WASM) execution environment. The objective was to select an algorithm that ensures stable, predictable transaction costs while efficiently handling the atomic **Join-Filter (JFLT)**, **Join-Update (JUPD)**, and **Join-Add (JADD)** operations. This review incorporates all empirical gas data, including a new finding that confirms the unpredictable and excessive cost of the riskiest solution on mixed data. This report validates the selection of the **Hybrid Linear/Logarithmic Solution (d)**, which provides optimal flexibility and cost control across various data scenarios.

1 Problem: Vector Alignment for Market Delta Calculation

The primary requirement is to synchronize two sorted vectors: a larger Market State Vector ($\mathbf{M}$) and a smaller Index Order Vector ($\mathbf{N}$). The process involves performing **Join-Add (JADD), Join-Update (JUPD), and Join-Filter (JFLT)** operations to generate the market size delta.

Specifically:

1. **Join-Filter:** Project elements from $\mathbf{M}$ onto $\mathbf{N}$.

2. **Join-Update:** Replace selected elements in $\mathbf{M}$ with values from corresponding data in $\mathbf{N}$.

3. **Join-Add:** Add to selected elements in $\mathbf{M}$ the values from corresponding data in $\mathbf{N}$.

Due to the nature of dynamic array operations in WASM, in-place removal of elements from the beginning of the large vector $\mathbf{M}$ can lead to excessive data shifting, introducing non-deterministic execution costs.

2 Solutions Proposed

Four primary implementation strategies were considered and tested to validate their theoretical complexity against practical WASM gas cost. The size of the market vector is $\mathbf{M}$ and the size of the user vector is $\mathbf{N}$.

3 Measurements: Simple Test Observations

Initial testing utilized a simple, low-rigor contiguous dataset (Test Set A) to establish baseline costs. This was followed by a more relevant test case (Test Set B) featuring a mix of sparse and contiguous data, representative of real market orders. The tests used a fixed dataset size: **50 asset index (N) and 150 asset market (M)**.

Table 1: Overview of Proposed Vector Alignment Solutions

Version	Implementation Method	Theoretical Complexity	Core Mechanism	Risk Note
(a)	Linear Scan with In-Place Filtering	$O(N+M) \rightarrow \mathbf{O(M^2)}$	In-place mutation (removal)	Risk of cost spike due to array shifting.
(b)	Linear Scan with Append to a New Vector	$\mathbf{O(N+M)}$	Linear read, selective copy to new vector.	Safe and highly efficient on simple data.
(c)	Pure Binary Search and Append	$O(N \cdot \log M)$	Binary search for every element in $\mathbf{N}$.	Safe, but higher inherent search overhead.
(d)	**Hybrid Linear/Logarithmic**	$\mathbf{O(N + J \cdot \log M)}$	Linear scan for contiguous blocks, binary search jump (J) for gaps.	Optimal flexibility and cost management.

Table 2: Comparative Empirical Gas Costs (Test Sets A and B)

Version	Implementation Method	Gas Cost (Test Set A)	Gas Cost (Test Set B - Mixed Data)
(a)	Linear scan, filtering **in-place**	$3,380,000$	$\mathbf{3,993,165}$
(b)	Linear scan, filtering and **appending**	$\mathbf{3,380,000}$	$3,990,747$
(c)	Pure Binary Search and Append	$3,970,000$	$3,991,967$
(d)	**Hybrid Linear/Logarithmic**	$3,420,000$	$\mathbf{3,990,453}$

Test Set B utilized asset labels with multiple contiguous sections and sparse sections (e.g., $1, 2, 5, 9$ followed by a block $31..41$). The cost observed for Solution (a) in Test Set B ($3,993,165$) was the highest measured across all efficient solutions, validating the risk associated with its $O(M^2)$ worst-case complexity.

4 Conclusions

The Test Set B confirmed that Solution (a)'s inherent risk of $\mathbf{O(M^2)}$ cost spikes makes it unacceptable. Solutions (b), (c), and (d) all offer deterministic safety.

The critical findings from the mixed-data Test Set B are the empirical validation of both the risk and the theoretical choice:

1. **Deterministic Safety:** All final candidates are safe.

2. **Risk Validation:** Solution (a) incurred the highest gas cost ($3,993,165$) under mixed-data conditions, proving its cost is volatile and unacceptable.

3. **Comparative Efficiency:** The three deterministic solutions performed closely: Solution (d) at $3,990,453$ gas, Solution (b) at $3,990,747$ gas, and Solution (c) at $3,991,967$ gas. This confirms that the hybrid approach (**d**) is the most gas-efficient under these mixed-data conditions, validating the strategic decision to prioritize its adaptive logic.

Mandate: The **Hybrid Solution (d)** is selected for implementation. It offers the required deterministic safety and has been empirically proven to be the most gas-efficient solution for handling the varied, complex data distributions of the market vectors.