LUZ

MATCHING ENGINE

Matching Engine de Alta Performance

Latência de nanossegundos. Durabilidade de grau institucional. Integração FIX imediata.

Um motor de negociação que entrega latência sub-microssegundo no matching, processa 1 milhão de ordens por segundo por nó, replica cada ordem com consenso Raft — zero perda mesmo em falha de hardware — e conecta-se à sua infraestrutura FIX existente sem reescrita. Construído em Java moderno sobre componentes open-source auditados, operável pela sua própria equipe.

1.000.000 /s
Ordens por segundo, por nó

~6,6 µs
Latência mediana de processamento (nó único)

0
Ordens perdidas — Raft em 3 nós, sustentado a 1M/s

O VALOR, EM UMA PÁGINA

O QUE VOCÊ GANHA	COMO ENTREGAMOS
Execução mais competitiva	Latência mediana de **100 ns no matching**; **~6,6 µs** de processamento ponta a ponta por nó
Capacidade para crescer	**1.000.000 de ordens/s por nó**, sem reescrita ao escalar
Zero perda de ordens	Consenso **Raft** em 3 nós: tolera falha de um nó inteiro mantendo o estado do book
Integração sem dor	**FIX 4.4** (entrada de ordens) + **FIX 5.0 SP2** (dados de mercado) — e acesso direto **SBE/Aeron** para HFT
Conformidade e controle	Risco pré-trade (fat-finger), *self-trade prevention*, multi-instrumento
Custo eficiente	Roda em **AWS Graviton (ARM64)** — melhor desempenho por dólar e por watt
Operável e observável	Métricas OpenTelemetry → Prometheus/Grafana, snapshots, recuperação rápida

POR QUE ISSO IMPORTA PARA O SEU NEGÓCIO

Latência é receita

Para participantes algorítmicos e market makers, cada microssegundo entre o sinal e a execução é arbitragem ganha ou perdida. Mais do que a média, o que importa é a **previsibilidade**: uma cauda de latência estável, sem picos de *garbage collection* ou contenção. Nosso motor processa cada ordem com **alocação zero de memória no caminho crítico** — não há pausas de GC, não há *spikes*. A latência que você mede hoje é a que você terá sob carga amanhã.

Durabilidade não é opcional

Reguladores exigem integridade do estado de negociação mesmo em falha de hardware. Replicação síncrona de banco de dados é lenta demais; não replicar é risco operacional inaceitável. Nossa resposta é o **consenso Raft via Aeron**: cada ordem é replicada e confirmada por maioria de nós **antes** de ser executada — durabilidade real, com latência de microssegundos, não de milissegundos.

Time-to-market

O sistema fala **FIX nativamente** nos dois lados — entrada de ordens e dados de mercado. Suas corretoras e clientes conectam com os mesmos *adapters* que já usam. Para os mais sensíveis a latência, há o caminho direto **SBE binário sobre Aeron**, sem overhead de protocolo.

DESEMPENHO COMPROVADO — EM HARDWARE DE NUVEM REAL

Medições de ponta a ponta em **AWS EC2 Graviton4**, com carga realista e **sem nenhuma otimização de infraestrutura** — números conservadores; hardware dedicado melhora ainda mais.

Um único nó — máxima performance

CARGA	LATÊNCIA P50	LATÊNCIA P99	PERDA
100.000 /s	6,6 µs	8,2 µs	0
500.000 /s	6,9 µs	9,3 µs	0
1.000.000 /s	**8,3 µs**	**14,2 µs**	0

Cluster de 3 nós — durável, tolerante a falhas

CARGA	LATÊNCIA P50	LATÊNCIA P99	PERDA
100.000 /s	330 µs	584 µs	0
500.000 /s	400 µs	622 µs	0
1.000.000 /s	**401 µs**	**613 µs**	0

> Mesmo replicando cada ordem em três nós para durabilidade total, o cluster sustenta **1 milhão de ordens por segundo com latência mediana de ~400 µs e p99 abaixo de 700 µs — sem perder uma única ordem**. A diferença entre o nó único (~6,6 µs) e o cluster (~400 µs) é o custo da durabilidade: o preço, em microssegundos, de garantir que nenhuma ordem se perca diante de uma falha.

PRONTO PARA EVOLUIR COM O HARDWARE

Cada nova geração de chip, mais performance — sem mudar uma linha de código.

O motor foi desenhado para ser **limitado por CPU** no caminho crítico — sem locks, sem alocação, sem espera de I/O. Na prática, cada nova geração de processador se traduz diretamente em mais performance. A migração para a geração mais recente de chips entregou de imediato um terço menos latência e o dobro de capacidade no cluster durável. Seu investimento acompanha a curva do hardware — automaticamente.

CAPACIDADES DE PRODUTO

Tipos de ordem	Limit, market, stop-buy, stop-sell, com prioridade preço-tempo (FIFO).
Risco pré-trade	Limites de quantidade, valor, posição, crédito e *price band* — aplicados antes da execução.
Self-Trade Prevention	Políticas CANCEL_NEWEST, CANCEL_OLDEST, CANCEL_BOTH por participante.
Multi-instrumento	Até 64 books simultâneos, configuráveis sem alteração de código.
Dados de mercado L2	Profundidade de book configurável, via FIX 5.0 SP2 ou SBE/Aeron.
Recuperação rápida	Snapshots periódicos do book e do estado de risco — *restart* em tempo proporcional ao gap, não ao histórico inteiro.
Observabilidade nativa	Throughput, latência (histograma), rejeições e saúde do cluster exportados via OpenTelemetry para Prometheus/Grafana.

CONCLUSÃO

Velocidade, durabilidade e facilidade de integração — sem concessões.

Um motor de matching de grau institucional não precisa escolher entre velocidade, durabilidade e facilidade de integração — este entrega os três. Latência de nanossegundos no núcleo, 1 milhão de ordens por segundo com replicação Raft e zero perda, FIX nativo nos dois lados, e um modelo de custo eficiente em nuvem ARM que melhora a cada geração de hardware.

Solicite uma prova de conceito →

Os números de desempenho foram medidos em AWS EC2 Graviton4, com o suite de carga do produto, sem otimização de infraestrutura. Resultados variam conforme hardware, topologia de rede e perfil de carga. Documentação técnica completa disponível sob solicitação.