

RELATÓRIO TÉCNICO – Testes de Desempenho da API

Aluno: David Leong

Resumo Executivo:

Os testes de desempenho realizados nos dois principais fluxos da API demonstraram diferenças claras de capacidade entre o cenário de I/O (/checkout/simple) e o cenário CPU-bound (/checkout/crypto). A rota de I/O suportou de forma estável até aproximadamente 300 usuários simultâneos, conforme evidenciado no spike test, que registrou 18.036 requisições sem falhas e latência p95 de aproximadamente 295 ms. Já a rota CPU-bound apresentou limitações significativas: durante o teste de estresse, a aplicação começou a falhar entre 200 e 300 usuários virtuais, culminando em 94% de falhas quando atingiu 1000 VUs. Assim, a capacidade máxima estimada é de cerca de 300 usuários no cenário de I/O e aproximadamente 200 a 300 usuários no cenário de CPU.

Evidências:

Os seguintes resultados sintetizam o comportamento observado:

- Smoke Test: 30 requisições com 0% de falhas e p95 » 2,6 ms.
- Load Test (/checkout/simple): 6864 requisições, 0% de falhas e p95 » 298 ms.
- Spike Test (/checkout/simple): 18.036 requisições, 0% de falhas e p95 » 295 ms.
- Stress Test (/checkout/crypto): 68.363 requisições, 94% de falhas e p95 acima de 55 s.

Análise de Estresse:

No endpoint /checkout/crypto, o teste de estresse revelou o ponto de ruptura da aplicação. As primeiras falhas surgiram quando a carga ultrapassou aproximadamente 200 usuários virtuais. A partir desse momento, foram observadas recusas de conexão e encerramentos abruptos de sessão. À medida que o teste avançou para cargas mais altas, especialmente acima de 500 VUs, a latência cresceu para dezenas de segundos, caracterizando saturação severa. O colapso completo ocorreu ao se atingir 1000 VUs, momento em que a taxa de falhas alcançou 94%, confirmando que o limite operacional seguro desse endpoint se encontra entre 200 e 300 usuários simultâneos.

Conclusão:

A API apresenta excelente desempenho no cenário de I/O e capacidade satisfatória para lidar com picos repentinos. Contudo, no cenário de CPU, o sistema atinge rapidamente seu limite e passa a falhar de forma crítica. Os resultados sugerem que operações intensivas em CPU devem ser otimizadas ou isoladas para garantir escalabilidade adequada.