Ano VII, v.1 2026 | submissão: 02/06/2026 | aceito: 05/06/2026 | publicação: 06/08/2026

Técnicas de redução de latência em pipelines de processamento em tempo real com computação de borda e computação sem servidor.

Técnicas de redução de latência em pipelines de processamento em tempo real com edge computing e serverless

Técnicas de redução de latência em pipelines de processamento em tempo real com computação borda e sem servidor

Lucas Mohallem Ferraz

Resumo: Este artigo examina técnicas de redução de latência em pipelines de processamento em tempo real, com ênfase na convergência de arquiteturas de Edge Computing e Serverless. A discussão adota uma abordagem histórico-evolutiva, traçando o caminho desde a consolidação da computação em nuvem — definida pelo NIST como acesso sob demanda a recursos computacionais compartilhados, elásticos e mensuráveis — até a descentralização promovida pelas arquiteturas de edge (Publicações Técnicas do NIST). Argumenta-se que a redução de latência depende não apenas de maior capacidade computacional, mas também de decisões arquitetônicas sobre alocação de processamento, granularidade funcional, comunicação assíncrona, particionamento de dados, observabilidade e inicialização a frio.

mitigação em plataformas sem servidor.
Palavras-chave: Computação de borda. Sem servidor. Latência. Processamento em tempo real. Sistemas distribuídos.

1 Introdução

A evolução dos sistemas distribuídos deslocou progressivamente o eixo computacional de centros de dados centralizados para arquiteturas híbridas, elásticas e geograficamente distribuídas.
Inicialmente, os modelos computacionais dependiam fortemente de grandes centros de dados centralizados que armazenavam, processavam e distribuíam informações para clientes e aplicativos. No entanto, o crescimento exponencial do volume de dados e do número de dispositivos conectados tornou a necessidade de arquiteturas mais flexíveis e escaláveis cada vez mais evidente. Nesse contexto, a computação em nuvem consolidou conceitos como elasticidade, autosserviço e pagamento conforme o uso, possibilitando maior eficiência operacional e redução dos custos de infraestrutura.
Apesar das vantagens oferecidas pela computação em nuvem, as aplicações modernas sensíveis ao tempo começaram a enfrentar desafios relacionados à latência e à variabilidade da rede. Em ambientes onde as respostas do sistema devem ocorrer em milissegundos, a distância física entre o usuário, a fonte de dados e os centros de dados pode comprometer significativamente o desempenho percebido. Esse problema tornou-se ainda mais evidente em aplicações críticas, como sistemas de monitoramento de saúde, veículos autônomos, realidade aumentada e ambientes industriais conectados, onde mesmo atrasos mínimos podem causar impactos operacionais significativos ou riscos à segurança.

Nesse contexto, emergiram paradigmas computacionais focados na descentralização do processamento, principalmente a Computação de Borda (Edge Computing) e a Computação Sem Servidor (Serverless Computing). A Computação de Borda aproxima o processamento e o armazenamento das fontes que geram eventos, reduzindo o tempo necessário para transmitir dados para a nuvem central. O modelo Sem Servidor, por sua vez, introduz uma abstração operacional baseada na execução orientada a eventos e na escalabilidade automática, reduzindo a necessidade de gerenciamento direto da infraestrutura. A combinação dessas abordagens representa uma importante evolução arquitetural para sistemas distribuídos modernos, especialmente em cenários de processamento em tempo real.

Ano VII, v.1 2026 | submissão: 02/06/2026 | aceito: 05/06/2026 | publicação: 06/08/2026

2. Evolução histórica: da nuvem centralizada à borda programável

Historicamente, a computação distribuída passou por diversas transformações estruturais impulsionadas pelo crescimento das redes e pela demanda por maior eficiência computacional. A primeira grande mudança ocorreu com a virtualização de infraestrutura, uma tecnologia que possibilitou a consolidação de múltiplas máquinas virtuais em um único host físico. Essa inovação melhorou a utilização de recursos e reduziu os custos operacionais, preparando o terreno para o surgimento da computação em nuvem. A virtualização também viabilizou a expansão de modelos de provisionamento de recursos escaláveis, permitindo que as organizações consumissem infraestrutura de forma flexível e sob demanda.

A segunda transformação significativa foi a consolidação da computação em nuvem por meio dos modelos IaaS, PaaS e SaaS. Esses modelos redefiniram a forma como os aplicativos eram desenvolvidos, implantados e consumidos, possibilitando elasticidade, alta disponibilidade e abstração da infraestrutura.

No entanto, à medida que as aplicações se tornaram cada vez mais dependentes de respostas em tempo real, as limitações das arquiteturas centralizadas tornaram-se mais evidentes. O aumento do tráfego de dados, a necessidade de comunicação contínua entre dispositivos inteligentes e a expansão da Internet das Coisas intensificaram os problemas relacionados à latência, à largura de banda e à dependência de centros de dados remotos.

Em resposta a essas limitações, surgiu o paradigma da Computação de Borda (Edge Computing), caracterizado pelo processamento computacional descentralizado. Em vez de rotear todos os dados para servidores centrais, parte do processamento ocorre próximo à origem dos eventos, reduzindo atrasos e melhorando a eficiência operacional.

Simultaneamente, o paradigma Serverless (Sem Servidor) emergiu como uma evolução da computação em nuvem, simplificando a execução de aplicações por meio da abstração de servidores e da escalabilidade automática orientada a eventos. Como resultado, a Computação de Borda e o Serverless passaram a representar pilares complementares no projeto de arquiteturas modernas para aplicações distribuídas sensíveis à latência.

3. Latência em Pipelines de Tempo Real

A latência em pipelines de processamento em tempo real refere-se ao intervalo de tempo necessário para que um evento seja capturado, processado e entregue ao consumidor final. Em arquiteturas distribuídas modernas, esse processo envolve múltiplas etapas, incluindo captura de eventos, serialização de dados, transmissão pela rede, enfileiramento, processamento, persistência e consumo. Cada uma dessas fases introduz pequenos atrasos que, quando acumulados, podem comprometer o desempenho geral do sistema. Em aplicações críticas, como sistemas financeiros ou monitoramento industrial, mesmo diferenças de alguns milissegundos podem afetar diretamente a experiência do usuário e a confiabilidade operacional.

Em sistemas distribuídos, a análise de latência não pode se limitar à média aritmética dos tempos de resposta. Métricas estatísticas como p95, p99 e p99,9 tornaram-se essenciais para avaliar com precisão a qualidade do serviço, pois representam os piores cenários que os usuários vivenciam. Em muitos ambientes de missão crítica, uma pequena fração de requisições extremamente lentas pode ser suficiente para violar os acordos de nível de serviço (SLAs) e causar falhas operacionais. Portanto, as arquiteturas modernas devem ser projetadas não apenas com foco no desempenho médio, mas também com estabilidade, previsibilidade e resiliência diante de flutuações de carga.

As principais fontes de atraso em pipelines distribuídos incluem a distância geográfica entre os componentes, congestionamento de rede, serialização ineficiente, chamadas síncronas encadeadas, contenção de filas e inicialização tardia de funções serverless. Além disso, problemas relacionados ao armazenamento remoto, replicação robusta de dados e a ausência de mecanismos de controle de fluxo podem aumentar significativamente os tempos de resposta. Portanto, lidar com a latência exige uma abordagem sistêmica que leve em consideração tanto os aspectos físicos da infraestrutura quanto as decisões arquitetônicas que regem o fluxo de dados e a organização dos serviços distribuídos.

Ano VII, v.1 2026 | submissão: 02/06/2026 | aceito: 05/06/2026 | publicação: 06/08/2026

4 Técnicas de Redução de Latência Arquitetural

Uma das principais técnicas de redução de latência envolve o processamento na borda da rede, conhecido como Computação de Borda. Nesse modelo, operações como filtragem, agregação, validação e inferência leve são realizadas próximas à fonte de dados, reduzindo o volume de informações transmitidas para a nuvem central. Essa abordagem diminui significativamente o tempo de ida e volta das solicitações, além de reduzir os custos de comunicação e o consumo de largura de banda. Em aplicações distribuídas de grande escala, a computação de borda também melhora a disponibilidade do sistema, permitindo que certas funcionalidades continuem operando mesmo quando a conectividade com as regiões centrais é interrompida.

Outra técnica relevante é o particionamento funcional de pipelines, no qual diferentes estágios de processamento são distribuídos de acordo com sua sensibilidade temporal. Operações que exigem respostas imediatas permanecem próximas à fonte de dados, enquanto tarefas analíticas, históricas ou de menor prioridade são descarregadas para ambientes centrais baseados em nuvem. Complementando isso, arquiteturas orientadas a eventos utilizam filas, brokers de streaming e logs distribuídos para desacoplar produtores de consumidores, reduzindo as dependências síncronas entre os serviços. Esse desacoplamento aumenta a escalabilidade do sistema e elimina gargalos associados ao processamento concorrente.

Além disso, técnicas específicas do paradigma serverless tornaram-se essenciais para a redução da latência. Estas incluem a mitigação do cold start, o uso de concorrência provisionada, inicialização preguiçosa e a adoção de runtimes mais leves. Caches hierárquicos, replicação seletiva de dados e processamento incremental baseado em janelas e agregações parciais também desempenham um papel fundamental. Em vez de recalcular estados completos, os sistemas mantêm informações resumidas continuamente atualizadas, reduzindo assim o custo computacional das operações em tempo real. A combinação de Edge Computing, Serverless e otimizações arquiteturais avançadas permite a construção de pipelines altamente responsivos e escaláveis.

5. Modelo de Referência Proposto

Um pipeline moderno de baixa latência baseado em Edge Computing e Serverless pode ser estruturado em múltiplas camadas especializadas. A primeira camada corresponde aos dispositivos de origem de dados, incluindo sensores, sistemas embarcados, dispositivos móveis e navegadores. Esses elementos atuam como produtores contínuos de eventos e representam a principal fonte de geração de dados em aplicações distribuídas contemporâneas. O crescimento da Internet das Coisas aumentou significativamente o número de dispositivos conectados, exigindo arquiteturas capazes de processar grandes volumes de dados em
Prazos mínimos.

A segunda camada consiste em nós de borda responsáveis pelas operações iniciais de processamento. Nesta etapa, ocorrem filtragem, normalização, autenticação, inferência leve e mecanismos de cache. Os dados podem então ser encaminhados para uma camada sem servidor regional, composta por funções orientadas a eventos responsáveis pelo enriquecimento, roteamento e persistência. O uso de funções sem servidor permite o escalonamento automático sob demanda, reduzindo o desperdício de recursos computacionais e simplificando o gerenciamento da infraestrutura.

Por fim, o pipeline incorpora uma camada de streaming responsável por filas, logs distribuídos e mecanismos de contrapressão, bem como uma camada analítica para armazenamento histórico, auditoria e treinamento de modelos de aprendizado de máquina. A principal decisão arquitetural neste modelo é determinar onde cada operação será executada, levando em consideração a sensibilidade ao tempo, os custos de movimentação de dados, os requisitos de consistência e as restrições regulatórias. O modelo de referência proposto busca, portanto, equilibrar desempenho, escalabilidade e eficiência operacional em ambientes distribuídos em tempo real.

Ano VII, v.1 2026 | submissão: 02/06/2026 | aceito: 05/06/2026 | publicação: 06/08/2026

6. Discussão Crítica

A convergência de Edge Computing e Serverless representa um avanço significativo no design de sistemas distribuídos modernos, mas não elimina a complexidade inerente a esses ambientes. Na prática, a descentralização do processamento reduz a latência, mas introduz novos desafios relacionados à implantação distribuída, sincronização de dados, segurança e observabilidade.

Quanto mais dispersos geograficamente estiverem os nós de borda, mais difícil se torna gerenciar as operações e manter a consistência entre os componentes do sistema.

O paradigma Serverless, por sua vez, reduz significativamente a carga de gerenciamento de infraestrutura, permitindo que as equipes concentrem seus esforços no desenvolvimento da funcionalidade do aplicativo. No entanto, as plataformas Serverless podem impor limitações quanto ao tempo máximo de execução, variabilidade de desempenho e dependência de fornecedor. Problemas como inicializações a frio e restrições de conectividade também podem afetar aplicativos altamente sensíveis ao tempo, principalmente em cenários com padrões de carga imprevisíveis ou picos repentinos de solicitações.

Portanto, o projeto arquitetônico de pipelines em tempo real deve levar em conta múltiplos fatores além da simples redução de latência. Aspectos como custo operacional, portabilidade, governança, resiliência e manutenibilidade devem ser avaliados de forma holística. Para aplicações críticas, o uso de métricas avançadas de observabilidade, rastreamento distribuído, simulação de falhas regionais e monitoramento contínuo de percentis de latência é fortemente recomendado. Somente uma abordagem integrada que combine engenharia de software e arquitetura distribuída pode garantir um equilíbrio adequado entre esses fatores.

Desempenho, confiabilidade e sustentabilidade operacional a longo prazo.

Conclusão

A redução da latência em pipelines de processamento em tempo real tornou-se um dos principais desafios da computação distribuída contemporânea. O crescimento da Internet das Coisas (IoT), de aplicações móveis e de sistemas inteligentes ampliou a demanda por respostas rápidas, estáveis e escaláveis. Nesse contexto, a Computação de Borda (Edge Computing) destaca-se como uma estratégia fundamental para aproximar o processamento e o armazenamento das fontes que geram eventos, reduzindo os atrasos causados pela distância entre os dispositivos e os data centers centrais.

Em paralelo, o paradigma Serverless introduziu um modelo operacional mais flexível e automatizado, baseado na execução sob demanda e na escalabilidade automática. Essa abordagem reduziu significativamente o esforço necessário para a administração da infraestrutura, permitindo maior agilidade no desenvolvimento.
aplicações distribuídas. No entanto, desafios como inicializações a frio, variabilidade de desempenho e dependência de fornecedores demonstram que a adoção dessas tecnologias requer um planejamento arquitetônico cuidadoso e mecanismos de otimização adequados.

Conclui-se que a combinação de Edge Computing e Serverless representa um estágio maduro na evolução dos sistemas distribuídos modernos. No entanto, a construção de pipelines de baixa latência requer uma abordagem sistêmica que englobe particionamento funcional, comunicação assíncrona, observabilidade avançada, processamento incremental e o uso inteligente de caches hierárquicos.

Assim, as organizações que buscam operar aplicações críticas em tempo real devem adotar arquiteturas que equilibrem desempenho, escalabilidade, confiabilidade e eficiência operacional em ambientes computacionais altamente distribuídos.

Referências

AMAZON WEB SERVICES. O que é AWS Lambda? Disponível em: Documentação oficial do AWS Lambda. Acesso em: 7 de maio de 2026.

Ano VII, v.1 2026 | submissão: 02/06/2026 | aceito: 05/06/2026 | publicação: 06/08/2026

AMAZON WEB SERVICES. Melhorando o desempenho de startups com o Lambda SnapStart. Disponível em: Documentação oficial do AWS Lambda. Acesso em: 7 de maio de 2026.

JONAS, E. et al. Programação em Nuvem Simplificada: Uma Visão de Berkeley sobre Computação Sem Servidor. Berkeley: Universidade da Califórnia, 2019.

MELL, P.; GRANCE, T. A definição de computação em nuvem do NIST. Publicação Especial 800-145 do NIST. Gaithersburg: Instituto Nacional de Padrões e Tecnologia, 2011.

SHI, W. et al. Computação de borda: visão e desafios. IEEE Internet of Things Journal, 2016.

SHI, W.; DUSTDAR, S. A promessa da computação de borda. Computer, IEEE, 2016.