FC-DT: Otimização Proativa de Recursos em Ambientes Névoa-Nuvem com Suporte de Gêmeos Digitais

Lucas Silva Lopes[1], José Miqueias[1], Iure Fé[1], Jonas Nunes[1]
Luiz Fernando Bittencourt[2], José Valdemir Junior[1] e Francisco Airton Silva[1]

[1]Universidade Federal do Piauí – PI – Brasil

[2]Universidade de Campinas – SP – Brasil

{lucaslopes092020, jmiqueias, iure.fe}@ufpi.edu.br

{jonas.nunes, valdemirreis, faps}@ufpi.edu.br

bit@ic.unicamp.br

Resumo. *Ambientes de computação em névoa e nuvem são essenciais para aplicações sensíveis à latência, mas variações dinâmicas de carga dificultam o cumprimento de Acordos de Nível de Serviço (SLA). Estratégias de alocação estática ou reativa podem resultar em violações de SLA ou sobreprovisionamento, elevando os custos operacionais. Este trabalho propõe o FC-DT, um Gêmeo Digital (GD) fundamentado em Redes de Petri Estocásticas (SPNs) para o gerenciamento preditivo e dinâmico de recursos na névoa-nuvem. A SPN fornece uma base formal para simular a dinâmica estocástica do sistema. Explorando essa capacidade, o FC-DT incorpora um mecanismo de tomada de decisão proativa que executa simulações de cenários alternativos em tempo de execução. Com base nos resultados das simulações, o FC-DT ajusta a alocação de recursos de forma coordenada entre as camadas de névoa e nuvem, antecipando possíveis violações de SLA e mantendo a configuração mínima necessária para seu cumprimento. Resultados experimentais mostram que a abordagem garante o cumprimento do SLA mesmo sob picos de carga, reduzindo o uso médio de recursos em 29,65% quando comparado a uma configuração estática mínima necessária para atender ao SLA.*

Abstract. *Fog and cloud computing environments are essential for latency-sensitive applications, however, dynamic workload variations hinder the fulfillment of Service Level Agreements (SLAs). Static or reactive allocation strategies may lead to SLA violations or overprovisioning, thereby increasing operational costs. This paper proposes FC-DT, a Digital Twin (DT) grounded in Stochastic Petri Nets (SPNs) for predictive and dynamic resource management in fog-cloud systems. SPNs provide a formal foundation to simulate the stochastic dynamics of the system. Leveraging this capability, FC-DT integrates a proactive decision-making mechanism that performs runtime simulations of alternative scenarios. Based on the simulation outcomes, FC-DT dynamically adjusts resource allocation in a coordinated manner across fog and cloud layers, anticipating potential SLA violations while maintaining the minimum configuration required for compliance. Experimental results demonstrate that the proposed approach ensures SLA fulfillment even under workload peaks, reducing average resource usage by 29.65% compared to the minimum static configuration required to meet the SLA.*

1. Introdução

Ambientes distribuídos de computação em névoa e nuvem são amplamente utilizados por aplicações sensíveis à latência, como sistemas de monitoramento, análise contínua de dados e serviços interativos, nas quais a proximidade do processamento em relação às fontes de dados é essencial para atender aos requisitos de tempo de resposta [Alli and Alam 2020]. Nesse contexto, o paradigma do contínuo névoa–nuvem organiza os recursos de forma hierárquica, aproximando parte do processamento das fontes de dados para reduzir latência e tráfego, enquanto a nuvem fornece suporte elástico para demandas computacionais mais intensivas [Bonomi et al. 2012]. Evidências indicam que esse paradigma pode tornar o processamento de dados até 7,5 vezes mais rápido e gerar uma economia de energia de até 80% ao evitar tráfego desnecessário [Fernando et al. 2025]

Nesse cenário, a alocação de recursos ao longo do contínuo névoa–nuvem ocorre em ambientes caracterizados pela heterogeneidade dos dispositivos e pela variabilidade estocástica das cargas de trabalho. Os nós distribuídos ao longo desse contínuo apresentam diferentes capacidades computacionais, restrições energéticas e condições de conectividade, enquanto as demandas das aplicações podem variar significativamente ao longo do tempo [Mahmud et al. 2018]. Essas características tornam-se especialmente relevantes em aplicações de missão crítica, que impõem limites estritos de tempo de resposta definidos por SLA. Como consequência, a coordenação eficiente dos recursos distribuídos torna-se um fator central para garantir níveis adequados de qualidade de serviço em ambientes névoa–nuvem [Bittencourt et al. 2025].

A eficácia do paradigma névoa–nuvem depende, portanto, de estratégias de alocação de recursos capazes de se adaptar dinamicamente às variações de carga, às restrições de latência e à disponibilidade dos recursos. Abordagens baseadas em planejamento de capacidade apresentam limitações na adaptação a variações de demanda, o que pode resultar em desperdício de recursos ou subprovisionamento [Carvalho et al. 2017]. De forma semelhante, mecanismos convencionais de *autoscaling* operam de maneira predominantemente reativa, estando sujeitos a atrasos na resposta a picos de carga [Lorido-Botran et al. 2014]. Abordagens baseadas em inteligência artificial buscam antecipar essas variações de carga, porém enfrentam desafios relacionados ao elevado custo computacional de treinamento e execução [Boscaro et al. 2025]. Assim, a ausência de mecanismos de orquestração adaptativos e preditivos pode transformar os benefícios da proximidade do processamento em novos gargalos, comprometendo o cumprimento dos requisitos de SLA em ambientes névoa–nuvem [Bittencourt et al. 2017].

Diante dessas limitações, o conceito de GD surge como uma alternativa promissora para apoiar a alocação de recursos na névoa-nuvem. O GD é uma representação virtual do sistema físico continuamente sincronizada por meio de um fluxo de dados bidirecional, no qual informações do ambiente real atualizam a representação virtual, enquanto decisões e políticas avaliadas no GD podem ser retroalimentadas ao sistema físico. Essa sincronização torna viável a avaliação preditiva de políticas de alocação antes de sua implementação no ambiente real [Yao et al. 2023]. Trabalhos na literatura têm explorado o uso de GD como mecanismo de apoio à tomada de decisão em sistemas distribuídos, especialmente por meio da simulação de cenários e da antecipação do comportamento do sistema. Essas abordagens abrangem desde aplicações industriais e de manufatura [Pires et al. 2021], passando por arquiteturas voltadas à segurança,

confiabilidade e armazenamento de dados em ambientes distribuídos [Lv and Lou 2022, Lakhan et al. 2023], até soluções focadas em redução de latência, mobilidade e eficiência energética em computação de borda, redes veiculares, névoa e nuvem [Guo et al. 2023, Van Huynh et al. 2023, Hayawi et al. 2025, Qadir et al. 2025, Yang et al. 2025]. Apesar desses avanços, observa-se que parte das abordagens não integra explicitamente requisitos de SLA ao processo decisório [Guo et al. 2023, Ramesh et al. 2024, Yang et al. 2025, Hayawi et al. 2025], e nenhuma realiza a alocação preditiva de recursos ao longo do contínuo névoa–nuvem. Além disso, muitas soluções permanecem reativas ou restritas a camadas específicas da infraestrutura [Van Huynh et al. 2023, Qadir et al. 2025]

Este trabalho propõe o FC-DT, um GD fundamentado no formalismo de SPN, capaz de antecipar estados futuros do sistema e orientar decisões proativas de alocação de recursos com base em requisitos de SLA, considerando de forma integrada as camadas de névoa e nuvem. A adoção do GD viabiliza um ambiente controlado de experimentação, no qual o desempenho do sistema é avaliado sob diferentes condições, permitindo a seleção e a aplicação automática de configurações de alocação de recursos. Nessa proposta, o sistema físico é representado por um modelo SPN continuamente atualizado com os dados operacionais do sistema físico, o que possibilita calcular métricas de desempenho, realizar simulações de múltiplos cenários, avaliar cada cenário e tomar uma decisão automaticamente. As SPNs estendem as Redes de Petri clássicas ao associar variáveis aleatórias aos tempos de disparo das transições, constituindo um arcabouço matemático adequado para a modelagem de sistemas nos quais tempo e incerteza são fatores críticos [Marsan et al. 1998]. A escolha da SPN em detrimento de outros formalismos justifica-se por sua capacidade de representar explicitamente concorrência, paralelismo e sincronização de eventos, tanto de forma gráfica quanto analítica.

O restante deste artigo está estruturado da seguinte maneira. A Seção 2 discute os trabalhos relacionados. A Seção 3 detalha a arquitetura do FC-DT, enquanto a Seção 4 descreve o modelo SPN desenvolvido para um sistema névoa-nuvem. A validação do modelo é apresentada na Seção 5. Na Seção 6, explora-se um estudo de caso sobre alocação preditiva de recursos orientada ao SLA. Por fim, as conclusões e sugestões para trabalhos futuros são apresentadas na Seção 7.

2. Trabalhos Relacionados

Esta seção apresenta uma visão geral da literatura recente sobre sistemas distribuídos que empregam GD, com ênfase na melhoria do desempenho por meio da simulação de cenários, na qual foram identificados nove trabalhos. A Tabela 1 resume a proposta de cada estudo, destacando ainda se os trabalhos consideram SLA, se abordam os paradigmas de computação em névoa e em nuvem, e se realizam alocação preditiva de recursos. Neste trabalho a alocação preditiva de recursos refere-se ao uso do GD para antecipar o estado operacional futuro (como carga, mobilidade e desempenho) a fim de tomar decisões proativas de alocação de recursos [Guo and Yang 2018].

Proposta. Os trabalhos analisados abrangem desde a otimização energética em sistemas manufatura [Pires et al. 2021], passando por arquiteturas voltadas à segurança e confiabilidade em nuvem híbrida e em sistemas de saúde [Lv and Lou 2022, Lakhan et al. 2023], até soluções direcionadas à redução de latência, justiça na computação em borda, mobilidade e eficiência em redes veiculares, UAVs e arquiteturas multicamadas [Guo et al. 2023, Van Huynh et al. 2023, Hayawi et al. 2025, Qadir et al. 2025]. Também há propostas voltadas à detecção de ataques em sistemas

Tabela 1. Trabalhos Relacionados

Trabalho	Proposta	SLA	Névoa	Nuvem	Alocação Preditiva de Recursos
[Pires et al. 2021]	Uma arquitetura de gêmeo digital voltada à otimização energética em sistemas de manufatura, com foco em eficiência operacional.	✗	✗	✗	✓
[Lv and Lou 2022]	Um modelo de segurança para armazenamento em nuvem híbrida, baseado em gêmeos digitais, destinado à proteção de dados industriais contra intrusões e à melhoria do desempenho.	✗	✓	✓	✗
[Guo et al. 2023]	Uma estratégia de implantação de UAVs assistida por gêmeos digitais, voltada à otimização da alocação de recursos e do desempenho em cenários dinâmicos.	✓	✗	✗	✗
[Van Huynh et al. 2023]	Um *framework* de minimização de latência com justiça em computação em borda, apoiado por gêmeos digitais, para assegurar baixa latência e equidade entre usuários.	✓	✗	✗	✗
[Lakhan et al. 2023]	Um *framework* seguro e tolerante a falhas para IIoHT, baseado em gêmeos digitais, que otimiza dados de sensores de saúde visando menor latência, maior confiabilidade e mitigação de riscos de segurança.	✓	✓	✓	✗
[Ramesh et al. 2024]	Um *framework* de gêmeo digital para simulação e detecção de ataques de segurança em sistemas de irrigação inteligentes.	✗	✗	✓	✗
[Yang et al. 2025]	Um *framework* dinâmico de gerenciamento de carga em arquiteturas de computação em névoa, que integra gêmeos digitais para reduzir o consumo de energia e aumentar a eficiência dos nós.	✗	✓	✗	✓
[Hayawi et al. 2025]	Uma estrutura de delegação de tarefas com gêmeos digitais para a internet dos veículos, projetada para garantir eficiência e baixa latência no processamento.	✗	✓	✗	✓
[Qadir et al. 2025]	Uma rede multicamadas assistida por gêmeos digitais, desenvolvida para assegurar baixa latência e eficiência energética.	✓	✗	✓	✗
Este Trabalho	Um gêmeo digital fundamentado em SPN para sistemas névoa–nuvem, voltado à alocação preditiva de recursos orientada por SLA.	✓	✓	✓	✓

inteligentes de irrigação [Ramesh et al. 2024] e ao gerenciamento dinâmico de carga em arquiteturas de névoa [Yang et al. 2025]. Em contraste, este trabalho propõe um GD fundamentado em SPN para sistemas névoa–nuvem, com foco na alocação preditiva de recursos orientada por SLA.

Acordos de Nível de Serviço. Quatro trabalhos incorporam métricas de SLA como critério para avaliação de desempenho ou suporte à tomada de decisão [Guo et al. 2023, Van Huynh et al. 2023, Lakhan et al. 2023, Qadir et al. 2025]. Os demais tratam desempenho de forma indireta, sem explicitar garantias associadas a SLA, mesmo em cenários sensíveis à latência [Pires et al. 2021, Lv and Lou 2022, Ramesh et al. 2024, Yang et al. 2025, Hayawi et al. 2025]. Neste trabalho, os requisitos de SLA são utilizados como elemento central para orientar decisões automáticas de alocação de recursos.

Névoa. A computação em névoa é considerada em quatro estudos. Algumas propostas adotam arquiteturas hierárquicas que incluem a névoa como camada intermediária [Lv and Lou 2022, Lakhan et al. 2023], enquanto outras exploram cenários em que o processamento próximo à origem dos dados é essencial para lidar com variações de carga ou mobilidade [Yang et al. 2025, Hayawi et al. 2025]. O presente trabalho considera a névoa como parte integrante do sistema.

Nuvem. A computação em nuvem é abordada em quatro trabalhos, no contexto de arquiteturas voltadas à segurança, confiabilidade, armazenamento de dados e desempenho [Lv and Lou 2022, Lakhan et al. 2023, Ramesh et al. 2024, Qadir et al. 2025]. Nos demais estudos, a nuvem não é integrada ao processo de alocação de recursos. Em contraste, este trabalho incorpora a nuvem de forma integrada à névoa, permitindo analisar o comportamento conjunto dessas camadas sob requisitos de desempenho e SLA.

Alocação Preditiva de Recursos. A alocação preditiva de recursos é explorada em três trabalhos [Pires et al. 2021, Yang et al. 2025, Hayawi et al. 2025], com base em previsões de carga, mobilidade e consumo energético. O restante dos trabalhos mantêm caráter predominantemente reativo. Diferentemente, este trabalho utiliza um GD fundamentado em SPN para antecipar estados futuros do sistema e orientar decisões proativas de alocação de recursos baseadas em SLA.

Diferenças do presente trabalho em relação aos trabalhos relacionados. Este trabalho se diferencia da literatura existente por combinar GD com modelagem formal baseada em SPN, possibilitando a análise preditiva do comportamento de sistemas distribuídos no contínuo névoa–nuvem. Além disso, a proposta integra as camadas de névoa e nuvem, incorpora requisitos de SLA como critério central de decisão e utiliza alocação preditiva de recursos. Essa combinação não é observada de forma conjunta nos trabalhos analisados.

3. Arquitetura do FC-DT

Esta seção apresenta a arquitetura proposta, que descreve o FC-DT, um GD voltado ao gerenciamento dinâmico e preditivo de recursos em ambientes névoa–nuvem. O objetivo da arquitetura é manter o cumprimento do SLA, ajustando de forma autônoma a alocação de recursos computacionais diante de variações nas condições operacionais do sistema. Para isso, o FC-DT opera em um ciclo contínuo de monitoramento, sincronização, análise e implantação. A Figura 1 apresenta uma visão geral da arquitetura proposta e seus principais componentes.

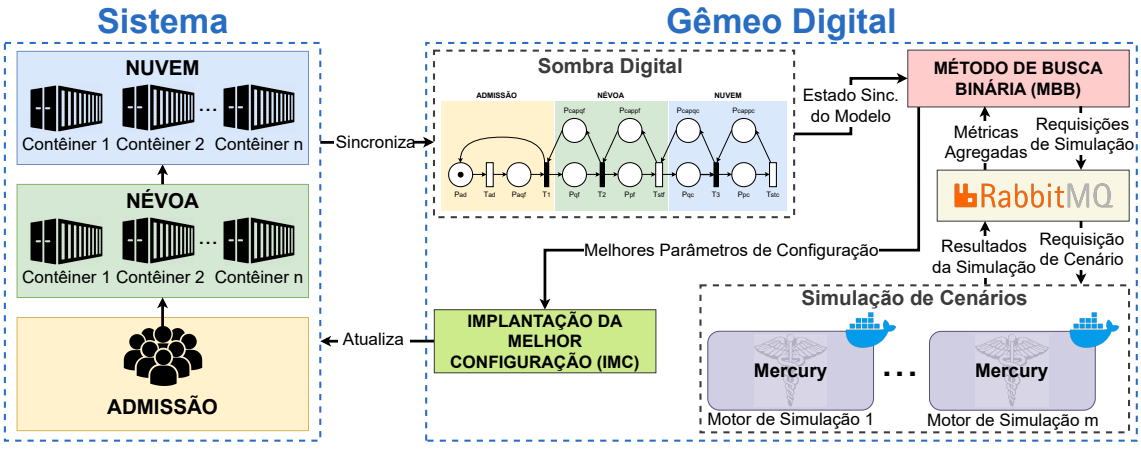

Figura 1. Arquitetura do FC-DT

Verifica-se da Figura 1 que o sistema físico é estruturado em três camadas hierárquicas: admissão, névoa e nuvem. A camada de admissão é responsável pela geração e pelo encaminhamento das requisições ao sistema. As requisições são então processadas na camada de névoa, que concentra recursos de processamento distribuídos

próximos ao usuário final. Após essa etapa, o fluxo de execução prossegue para a camada de nuvem, responsável pelo processamento subsequente das requisições e pelo suporte computacional centralizado. Essas camadas são compostas por contêineres que executam o fluxo de trabalho da aplicação e, em conjunto, constituem o sistema físico monitorado pelo GD composto por quatro módulos principais: a Sombra Digital, o Método de Busca Binária (MBB), o módulo de Simulação de Cenários e o módulo de Implantação da Melhor Configuração (IMC). A Sombra Digital mantém uma réplica virtual continuamente sincronizada com o estado operacional do sistema névoa–nuvem, representada formalmente por modelos SPN. Essa sincronização é realizada a partir de dados coletados do sistema físico, permitindo que o modelo reflita de forma consistente a carga, a utilização de recursos, os tempos de serviço e o estado das filas.

O Estado Sincronizado do Modelo constitui a entrada para o MBB, que atua como o núcleo de decisão da arquitetura. Para mitigar riscos de violação de SLA, o MBB adota um viés pessimista na estimativa da carga, definido como a consideração deliberada de uma demanda superior à média observada, de modo a antecipar picos de carga. A alocação de recursos é tratada como um problema de otimização em um espaço finito e ordenado. Os limites mínimo e máximo de recursos disponíveis são definidos por um operador durante a criação do GD e podem ser atualizados ao longo do tempo, refletindo restrições operacionais ou mudanças na infraestrutura. Esses parâmetros delimitam o intervalo inicial de busca utilizado pelo MBB, que aplica iterativamente o algoritmo de busca binária. Em cada iteração, o método seleciona uma configuração intermediária dentro do intervalo atual e a encaminha ao módulo de Simulação de Cenários. A comunicação entre os módulos é realizada por meio de um *message broker* (RabbitMQ), enquanto as simulações são executadas pela ferramenta Mercury [Maciel et al. 2017], responsável por parametrizar o modelo SPN e retornar métricas de desempenho, como o tempo médio de resposta e a vazão.

Com base nos resultados, o MBB ajusta o intervalo de busca: se a configuração atende ao SLA, restringe-se à metade inferior, buscando reduzir o consumo de recursos; caso contrário, desloca-se para a metade superior, aumentando os recursos. O processo prossegue até a convergência, garantindo que todas as configurações relevantes sejam avaliadas. Ao final, o MBB seleciona a configuração que satisfaz o SLA com o menor consumo possível de recursos. A escolha pela busca binária justifica-se pela eficiência, enquanto uma busca sequencial exigiria N simulações para um sistema com N contêineres, o MBB reduz a complexidade para $O(\log_2 N)$. Em ambientes névoa–nuvem, nos quais o tempo de decisão é crítico, essa redução é essencial para a viabilidade do gerenciamento preditivo de recursos. Por fim, a configuração selecionada é encaminhada ao módulo IMC, responsável por traduzir as decisões do GD em ações operacionais no sistema físico, aplicando as atualizações necessárias nas camadas de contêineres e fechando o ciclo de controle da arquitetura. Ao integrar modelos formais, monitoramento em tempo real do sistema e mecanismos de decisão preditiva, a arquitetura proposta permite antecipar variações de carga e avaliar o impacto de decisões de alocação de recursos antes de sua aplicação no ambiente real. Dessa forma, a arquitetura fornece uma base para o gerenciamento preditivo de recursos em ambientes névoa–nuvem.

4. Modelo SPN para Sistemas Névoa-Nuvem

Esta seção descreve o modelo SPN adotado para representar o comportamento de um sistema névoa–nuvem, capturando a dinâmica de chegada e processamento das requisições

ao longo das camadas de admissão, névoa e nuvem. O modelo constitui a base do GD e permite a avaliação preditiva do desempenho do sistema sob diferentes configurações. A Figura 2 apresenta uma visão geral do modelo SPN, que é composto por lugares, transições e tokens. Os lugares representam estados do sistema e armazenam tokens que indicam requisições em espera ou em processamento. As transições descrevem eventos que alteram o estado do sistema à medida que as requisições percorrem as camadas do sistema. Essa estrutura permite modelar o fluxo contínuo de chegada, processamento e saída das requisições, considerando restrições de capacidade e interdependência entre processos [Marsan et al. 1998].

Os lugares do modelo estão distribuídos entre as camadas de admissão, névoa e nuvem. Na admissão, `Pad` atua como a fila inicial, armazenando as requisições que chegam ao sistema. O lugar `Paqf` mantém as requisições aceitas e prontas para envio à névoa. Na camada de névoa, `Pqf` representa a fila de espera e `Ppf` as requisições em processamento. Os lugares `Pcapqf` e `Pcappf` limitam a capacidade da fila de espera e de processamento da Névoa. Na nuvem, `Pqc` armazena as requisições em espera e `Ppc` reúne as requisições sendo processadas. Os lugares `Pcapqc` e `Pcappc` definem, respectivamente, a capacidade máxima de entrada e o limite de processamento da nuvem. As transições seguem a mesma estrutura e determinam o avanço dos tokens entre os lugares.

As transições são temporizadas quando associadas a tempos de serviço estocásticos e imediatas quando descrevem eventos lógicos sem atraso. As transições `Tad`, `Tstf` e `Tstc` são temporizadas, enquanto as transições `T1`, `T2` e `T3` são imediatas. Na admissão, `Tad` controla o tempo entre chegadas das requisições, enquanto `T1` executa a passagem imediata para a névoa. Na camada de névoa, `T2` representa a passagem da requisição da fila de espera para o processamento, e `Tstf` realiza o processamento e transferência das requisições para a nuvem. Na camada de Nuvem, `T3` envia as requisições da fila de espera para o processamento, e `Tstc` marca a saída definitiva do sistema. Essa configuração garante a coerência temporal e causal do fluxo, refletindo a movimentação real das requisições ao longo das camadas do sistema. A partir do modelo SPN, são extraídas métricas de desempenho como o tempo médio de resposta (MRT) do sistema e a vazão. As métricas alimentam o processo de decisão do FC-DT, permitindo a seleção preditiva de configurações de recursos que garantam os requisitos de desempenho.

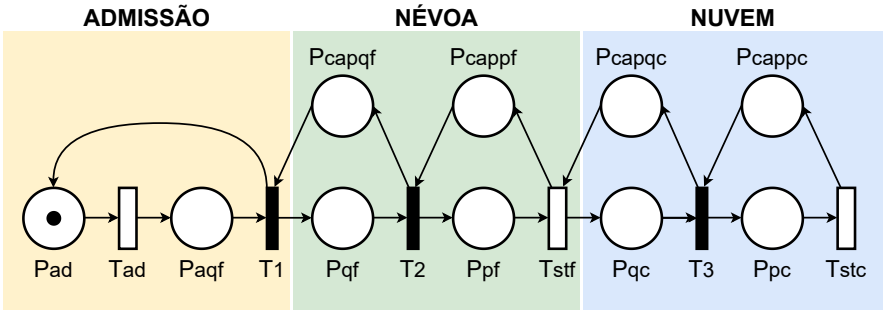

Figura 2. Modelo SPN para Sistemas Névoa-Nuvem.

4.1. Métricas

O MRT é uma das principais métricas adotadas neste trabalho e expressa o tempo médio decorrido entre a chegada de uma requisição ao sistema e a sua conclusão. Com base

na Lei de Little [Little and Graves 2008], o MRT é calculado a partir da razão entre a quantidade média de requisições no sistema e a vazão observada. Nessa formulação, $E\{\cdot\}$ denota o operador valor esperado no regime estacionário, enquanto $\#P$ representa o número médio de tokens presentes no lugar P do modelo SPN, conforme definido na Equação 1.

$$\text{MRT} = \frac{E\{\#Pqf\} + E\{\#Ppf\} + E\{\#Pqc\} + E\{\#Ppc\}}{\text{Vazão}} \tag{1}$$

O numerador da equação 1 corresponde aos lugares associados à espera e ao processamento das requisições nas camadas de névoa e nuvem, uma vez que esses estados representam efetivamente o período em que a requisição ocupa recursos do sistema. A vazão do sistema é definida como a taxa média de conclusão de requisições e é apresentada na Equação 2. Essa métrica é obtida a partir do valor esperado do número de tokens no lugar `Ppc`, dividido pelo tempo de serviço da nuvem (STC), associado à transição temporizada `Tstc`.

$$\text{Vazão} = \frac{E\{\#Ppc\}}{STC} \tag{2}$$

A simulação de cada modelo é conduzida pelo GD em regime estacionário. Esse tipo de simulação possibilita observar a evolução do sistema a partir de condições iniciais até a convergência para um comportamento estável, fornecendo uma estimativa do desempenho do sistema sob as condições consideradas.

5. Validação do Modelo SPN Comparando com um Sistema Real

Nesta seção, é apresentada a validação do modelo proposto, comparando os resultados de desempenho gerados pelo modelo SPN com as métricas do sistema, quando executado sob as mesmas condições. O objetivo é avaliar se o modelo consegue replicar o comportamento do sistema diante das mudanças na taxa de chegada, confirmando, dessa forma, sua capacidade de representar o cenário real e servir como base para simular cenários alternativos. O sistema avaliado foi o Pasid-Validator, emulador que permite a emulação de um ambiente névoa-nuvem e já foi utilizado em estudos para emular diferentes sistemas distribuídos [Silva et al. 2025, Araújo et al. 2025].

A Tabela 2 apresenta os parâmetros utilizados no processo de validação do modelo, incluindo a duração total do experimento, fixada em 1 hora, a fim de garantir a coleta de dados representativos. O tempo entre chegadas de requisições (AD) é variado ao longo do intervalo de 0,045 segundos a 0,3 segundos com o objetivo de gerar diferentes níveis de carga no sistema e avaliar o comportamento do modelo sob condições dinâmicas de operação. Essa variação permite observar o impacto dos picos de chegada sobre as métricas de desempenho analisadas. Os tempos médios de serviço foram configurados em 1,2 segundos para a camada de névoa e 1,8 segundos para a camada de nuvem, refletindo as capacidades de processamento de cada camada. Em termos de recursos disponíveis, foram considerados 32 contêineres tanto na névoa quanto na nuvem, assegurando uma configuração simétrica entre as infraestruturas avaliadas. Adicionalmente, foi adotado um tempo de monitoramento de 15 segundos, durante o qual o sistema é continuamente observado. Ao final desse intervalo, são calculados os valores médios do tempo entre chegadas das requisições, bem como dos tempos de serviço na névoa e na

nuvem. Esses valores médios são então utilizados como entrada do modelo, a partir dos quais as métricas de desempenho são estimadas.

Tabela 2. Parâmetros da Validação.

Parâmetro	Valor	Descrição
Duração do experimento	1 h	Período total de execução utilizado na validação do modelo.
Tempo entre chegadas	[0,045 s, 0,3 s]	Intervalo de variação do tempo entre chegadas de requisições.
Tempo de serviço da Névoa	1,2 s	Tempo médio de processamento da requisição na camada de névoa.
Tempo de serviço da Nuvem	1,8 s	Tempo médio de processamento da requisição na camada de nuvem.
Quantidade de recursos na Névoa	32 contêineres	Quantidade de recursos disponíveis para processamento.
Quantidade de recursos na Nuvem	32 contêineres	Quantidade de recursos disponíveis para processamento.
Tempo de monitoramento	15 s	Intervalo utilizado para coletar dados do sistema e calcular valores médios de entrada do modelo SPN.

A Figura 3 apresenta a comparação do MRT, em segundos, obtido pelo sistema real e pelo modelo. Observa-se que o comportamento dinâmico do modelo acompanha de forma consistente o comportamento do sistema real. Durante os períodos de maior carga, associados aos picos de chegada de requisições, ocorre um aumento temporário no tempo de resposta em ambos os casos. Esse aumento está relacionado ao acúmulo de solicitações nas filas, caracterizando uma saturação momentânea dos servidores. À medida que a carga diminui, o tempo de resposta retorna rapidamente ao seu nível médio, indicando a recuperação da estabilidade do sistema. Esse comportamento evidencia que o modelo representa adequadamente o processo de formação e esvaziamento das filas.

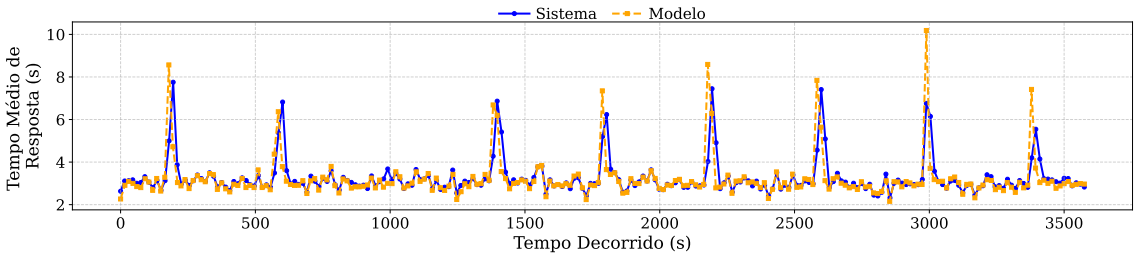

Figura 3. Validação com o MRT.

A Figura 4 apresenta a Vazão utilizada no processo de validação. Assim como observado para o MRT, as curvas do modelo e do sistema real apresentam forte aderência visual ao longo do tempo. As variações na vazão acompanham diretamente as flutuações na carga de trabalho, demonstrando que o modelo é capaz de reproduzir corretamente o impacto dos picos de chegada e dos períodos de menor demanda sobre a capacidade de processamento do sistema.

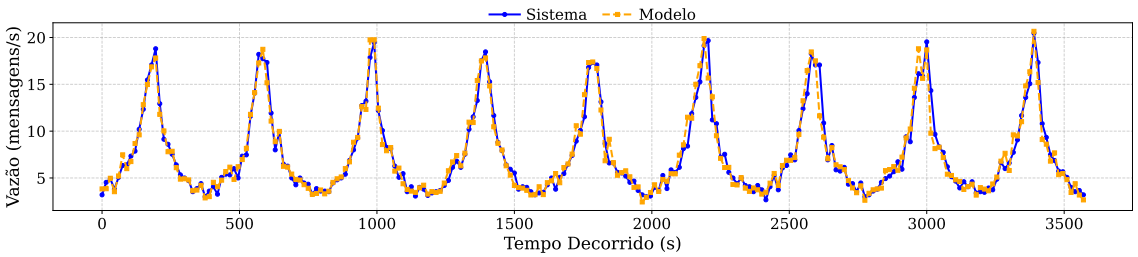

Figura 4. Validação com a Vazão.

Além da correspondência visual entre as curvas, os valores médios das métricas de desempenho apresentaram equivalência estatística, conforme mostrado na Tabela 3. A análise comparativa, conduzida por meio de um teste t pareado com nível de confiança de 95% e baseada em 240 observações, indicou que as diferenças entre as médias do sistema real e do modelo não são estatisticamente significativas. Para o MRT, obteve-se um valor-p de 0,4152, enquanto para a Vazão o valor-p foi de 0,8093. Esses resultados indicam a ausência de discrepâncias relevantes entre os dados observados e os estimados pelo modelo.

Tabela 3. Resultado do Teste t Pareado.

Métrica	Diferença Média	IC 95%	t	Valor-p
MRT (s)	0,0398	[-0,0563, 0,1359]	0,8161	0,4152
Vazão (msg/s)	-0,0151	[-0,1383, 0,1081]	-0,2417	0,8093

Dessa forma, a correspondência entre os resultados experimentais e os obtidos pelo modelo SPN confirma que o modelo proposto representa com precisão o comportamento do sistema real. Com essa validação, o modelo torna-se uma base confiável para a realização de análises preditivas, permitindo avaliar diferentes configurações de recursos e padrões de carga com menor custo e tempo de experimentação em ambiente real.

6. Estudo de Caso

Esta seção avalia a capacidade do FC-DT em antecipar violações de SLA e em reduzir o sobreprovisionamento de recursos, considerando como métrica principal o MRT, cujo limite é definido em 5 segundos. Esse valor representa um compromisso entre requisitos de usabilidade e restrições operacionais de sistemas distribuídos, uma vez que tempos de resposta próximos a 1–2 s estão associados a uma experiência ideal, enquanto latências superiores a 10 s tendem a causar perda de atenção e insatisfação do usuário [Nielsen 1994]. Além disso, avalia-se a capacidade do FC-DT em ajustar dinamicamente a quantidade de contêineres, mantendo o desempenho exigido com o menor consumo possível de recursos.

Para essa análise, foram adotados os mesmos parâmetros experimentais utilizados na validação do modelo, conforme apresentado na Tabela 2, com exceção do AD, que nesta avaliação varia de 0,06 segundos a 0,3 segundos. Após o período de monitoramento, o FC-DT filtra os valores de AD coletados e utiliza, no cálculo da média, apenas aqueles abaixo do percentil 75, incorporando o viés pessimista definido na Seção 3 à estimativa da carga. O desempenho da arquitetura proposta é comparado com três configurações estáticas de referência, resumidas na Tabela 4, que variam a quantidade de contêineres alocados entre as camadas de névoa e nuvem. Em contraste, o FC-DT realiza a alocação de recursos de forma dinâmica, respeitando um limite máximo de 32 contêineres por camada. Para seu funcionamento, o FC-DT utiliza dois contêineres adicionais dedicados às tarefas de simulação e tomada de decisão.

A Figura 5 apresenta o MRT do sistema, considerando uma média móvel de 15 segundos, e tem como objetivo avaliar a capacidade do FC-DT em manter o MRT abaixo do limite de SLA sob variações de carga, em comparação com configurações estáticas. Observa-se que as Configurações 1 e 2 apresentam aumentos de latência em períodos de maior carga, ultrapassando repetidamente o limite de 5 s, especialmente em torno

Tabela 4. Configurações Estáticas de Referência.

Configuração	Contêineres na Névoa	Contêineres na Nuvem
Configuração 1	20	29
Configuração 2	21	28
Configuração 3	21	29

dos instantes 200 s, 800 s, 1400 s e 3200 s. Em contraste, tanto o FC-DT quanto a Configuração 3 mantêm o MRT consistentemente abaixo do limiar estabelecido, contudo, o FC-DT alcança esse resultado por meio de ajustes dinâmicos na alocação de recursos, enquanto a Configuração 3, por se tratar de uma configuração estática, resulta em sobre-provisionamento de recursos durante períodos de baixa carga.

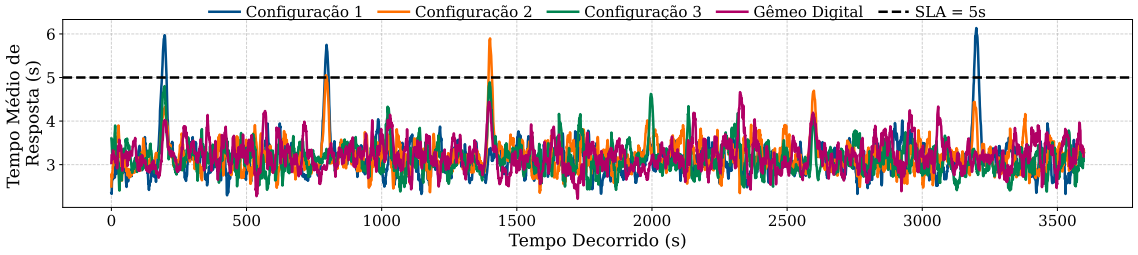

Figura 5. Tempo Médio de Resposta com Média Móvel de 15 Segundos.

A Figura 6 apresenta a vazão do sistema ao longo do tempo, caracterizando a variabilidade da carga imposta durante o experimento. Observam-se períodos de alta variabilidade, com picos que atingem aproximadamente 18 a 20 mensagens por segundo. O gráfico indica ainda que todas as configurações avaliadas, assim como o FC-DT, conseguem atender às requisições submetidas, mantendo níveis de vazão equivalentes ao longo do tempo. Esses picos de carga explicam os aumentos de latência observados nas configurações estáticas que não se adaptam à demanda, evidenciando a necessidade de mecanismos de ajuste proativo de recursos.

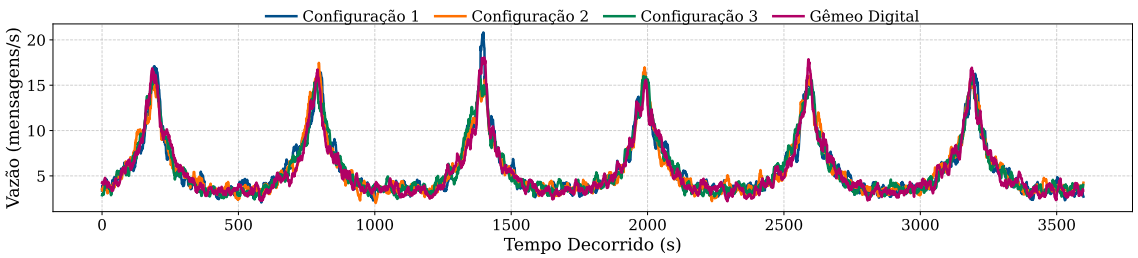

Figura 6. Vazão com Média Móvel de 15 Segundos.

A Figura 7 mostra a adaptação dinâmica da infraestrutura na camada de névoa. Em períodos de baixa demanda, como entre 1000 s e 1300 s, o FC-DT reduz a alocação para aproximadamente 5 contêineres, enquanto as configurações estáticas mantêm um número elevado de contêineres ativos, caracterizando sobreprovisionamento. À medida que a carga aumenta, o FC-DT ajusta progressivamente a alocação, elevando a quantidade de contêineres para sustentar o desempenho do sistema. A adaptação correspondente na camada de nuvem é apresentada na Figura 8. Em períodos de maior demanda, o FC-DT

escala rapidamente os recursos, garantindo a manutenção do MRT dentro do limite de SLA. Em contrapartida, as configurações estáticas mantêm recursos alocados mesmo em períodos de baixa carga, sem ganhos proporcionais de desempenho.

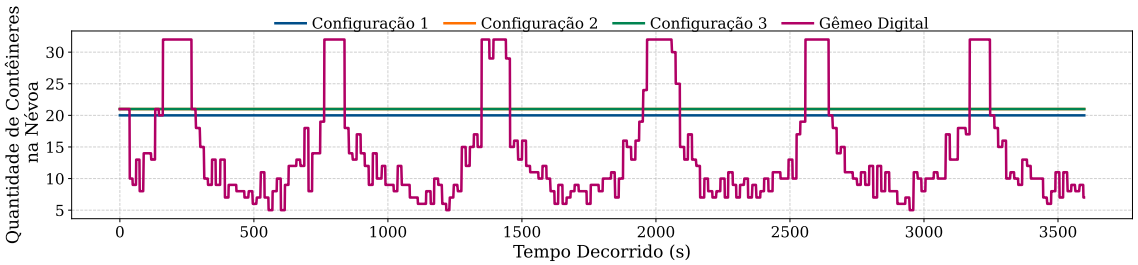

Figura 7. Quantidade de Contêineres na Névoa.

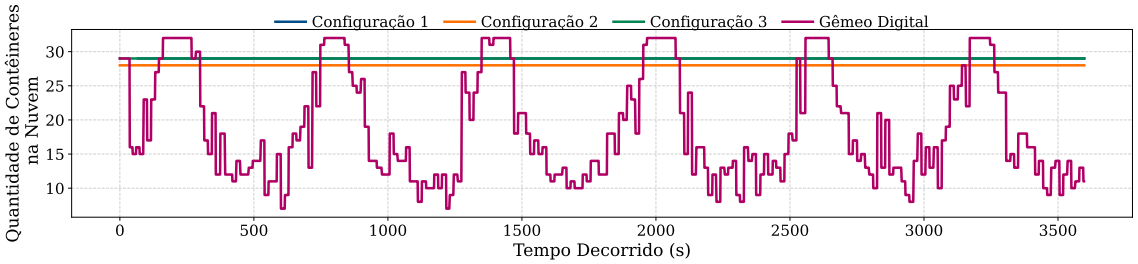

Figura 8. Quantidade de Contêineres na Nuvem.

A análise do uso de recursos evidencia o impacto prático da abordagem proposta. Enquanto a Configuração 3 atende ao SLA com 50 contêineres, somando as camadas de névoa e nuvem, o FC-DT opera com uma média de 35,177 contêineres ao longo do experimento, resultando em uma redução média de 14,823 contêineres, o que corresponde a um ganho de aproximadamente 29,65% no uso dos recursos, sem comprometer o cumprimento do SLA. Esses resultados indicam que variações proativas na alocação de recursos podem ser determinantes para evitar violações de SLA e reduzir o consumo de recursos, com impacto direto na diminuição dos custos operacionais. Ao avaliar cenários de forma preditiva, o FC-DT identifica automaticamente a configuração mínima necessária para sustentar a carga projetada, otimizando o equilíbrio entre desempenho e custo operacional em ambientes distribuídos de névoa–nuvem.

7. Conclusão

Este trabalho propôs o FC-DT, um GD fundamentado em SPN para a alocação preditiva de recursos em ambientes distribuídos de névoa-nuvem, com o objetivo de antecipar violações de SLA e reduzir o consumo de recursos computacionais. A abordagem proposta explora simulações de cenários alternativos para ajustar dinamicamente a alocação de contêineres, evitando a degradação de desempenho e reduzindo o sobreprovisionamento característico de estratégias estáticas. Os resultados de validação confirmaram que o modelo representa adequadamente o sistema real, permitindo análises preditivas confiáveis. Na avaliação experimental, o FC-DT manteve o MRT abaixo do limite de SLA de 5 segundos mesmo sob picos de carga, alcançando uma redução de 29,65% no uso de recursos, correspondente a uma economia de 14,823 contêineres em relação a configuração fixa que atende o SLA. A estratégia de filtragem dos tempos entre chegadas, com a remoção de valores iguais ou superiores ao percentil 75, contribuiu para

decisões proativas, embora possa causar sobreprovisionamento. Como trabalhos futuros, destacam-se a incorporação de algoritmos de predição de carga para evitar o sobreprovisionamento causado pela filtragem do valores de AD, a avaliação sob cargas aleatórias, a comparação entre estratégias de simuação de cenários baseadas em tempo, eventos e abordagens híbridas, bem como a consideração de falhas e o uso de múltiplos modelos SPN para abranger diferentes aspectos do sistema.

Referências

Alli, A. A. and Alam, M. M. (2020). The fog cloud of things: A survey on concepts, architecture, standards, tools, and applications. *Internet of Things*, 9:100177.

Araújo, J. M., Lopes, L. S., Lima, L. N., Barbosa, V., Sabino, A., Feitosa, L., Delicato, F. C., Nguyen, T. A., and Silva, F. A. (2025). Optimizing intelligent camera surveillance in smart buildings: An spn-based edge-fog analysis. In *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*, pages 15–28. SBC.

Bittencourt, L. F., Diaz-Montes, J., Buyya, R., Rana, O. F., and Parashar, M. (2017). Mobility-aware application scheduling in fog computing. *IEEE Cloud Computing*, 4(2):26–35.

Bittencourt, L. F., Rodrigues-Filho, R., Spillner, J., De Turck, F., Santos, J., Fonseca, N. L., Rana, O., and Parashar, M. (2025). The computing continuum: Past, present, and future. *Computer Science Review*, 58.

Bonomi, F., Milito, R., Zhu, J., and Addepalli, S. (2012). Fog computing and its role in the internet of things. In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, pages 13–16.

Boscaro, M., Mason, F., Chiariotti, F., and Zanella, A. (2025). To train or not to train: Balancing efficiency and training cost in deep reinforcement learning for mobile edge computing. In *ICC 2025-IEEE International Conference on Communications*, pages 2352–2357. IEEE.

Carvalho, M., Menascé, D. A., and Brasileiro, F. (2017). Capacity planning for iaas cloud providers offering multiple service classes. *Future Generation Computer Systems*, 77:97–111.

Fernando, N., Shrestha, S., Loke, S. W., and Lee, K. (2025). On edge-fog-cloud collaboration and reaping its benefits: a heterogeneous multi-tier edge computing architecture. *Future Internet*, 17(1):22.

Guo, H., Zhou, X., Wang, J., Liu, J., and Benslimane, A. (2023). Intelligent task offloading and resource allocation in digital twin based aerial computing networks. *IEEE Journal on Selected Areas in Communications*, 41(10):3095–3110.

Guo, J. and Yang, C. (2018). Predictive resource allocation with deep learning. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pages 1–7. IEEE.

Hayawi, K., Sajid, J., Malik, A. W., and Mathew, S. S. (2025). Digital twin assisted task offloading for workload management at fog nodes. *IEEE Internet of Things Journal*.

Lakhan, A., Lateef, A. A. A., Abd Ghani, M. K., Abdulkareem, K. H., Mohammed, M. A., Nedoma, J., Martinek, R., and Garcia-Zapirain, B. (2023). Secure-fault-tolerant efficient industrial internet of healthcare things framework based on digital twin federated

fog-cloud networks. *Journal of King Saud University-Computer and Information Sciences*, 35(9):101747.

Little, J. D. and Graves, S. C. (2008). Little's law. In *Building intuition: insights from basic operations management models and principles*, pages 81–100. Springer.

Lorido-Botran, T., Miguel-Alonso, J., and Lozano, J. A. (2014). A review of auto-scaling techniques for elastic applications in cloud environments. *Journal of grid computing*, 12(4):559–592.

Lv, Z. and Lou, R. (2022). Edge-fog-cloud secure storage with deep-learning-assisted digital twins. *IEEE Internet of Things Magazine*, 5(2):36–40.

Maciel, P., Matos, R., Silva, B., Figueiredo, J., Oliveira, D., Fé, I., Maciel, R., and Dantas, J. (2017). Mercury: Performance and dependability evaluation of systems with exponential, expolynomial, and general distributions. In *2017 IEEE 22nd Pacific Rim international symposium on dependable computing (PRDC)*, pages 50–57. IEEE.

Mahmud, R., Koch, F. L., and Buyya, R. (2018). Cloud-fog interoperability in iot-enabled healthcare solutions. In *Proceedings of the 19th international conference on distributed computing and networking*, pages 1–10.

Marsan, M. A., Balbo, G., Conte, G., Donatelli, S., and Franceschinis, G. (1998). Modelling with generalized stochastic petri nets. *ACM SIGMETRICS performance evaluation review*, 26(2):2.

Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.

Pires, F., Ahmad, B., Moreira, A. P., and Leitão, P. (2021). Digital twin based what-if simulation for energy management. In *2021 4th IEEE international conference on industrial cyber-physical systems (ICPS)*, pages 309–314. IEEE.

Qadir, M. A., Naeem, M., and Ejaz, W. (2025). Digital twin-assisted multi-layer networks for low-latency and energy-efficient communication. *Computer Communications*, page 108219.

Ramesh, K., Sasirekha, G., Rao, M., Bapat, J., and Das, D. (2024). Digital twin based what-if simulation of security attacks in smart irrigation systems. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6. IEEE.

Silva, L. G., Barbosa, V., Cardoso, I., Alves, M., Lopes, L. S., Rego, P. A., and Silva, F. A. (2025). Dynamically adaptive rsus in vanets: An approach focused on sustainability. *International Journal of Communication Systems*, 38(16):e70283.

Van Huynh, D., Nguyen, V.-D., Khosravirad, S. R., Karagiannidis, G. K., and Duong, T. Q. (2023). Distributed communication and computation resource management for digital twin-aided edge computing with short-packet communications. *IEEE Journal on Selected Areas in Communications*, 41(10):3008–3021.

Yang, T.-T., Shen, S.-Y., and Huang, S.-X. (2025). Energy optimization and overheating mitigation using digital twins in fog computing networks. In *IET Conference Proceedings CP929*, volume 2025, pages 561–567. IET.

Yao, J.-F., Yang, Y., Wang, X.-C., and Zhang, X.-P. (2023). Systematic review of digital twin technology and applications. *Visual computing for industry, biomedicine, and art*, 6(1):10.