57º SBPO
SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
Gramado • RS
DE 5 A 9 DE OUTUBRO DE 2025

Programação Matemática Aplicada à Orquestração de Telemetrias *In-Band* para o Monitoramento de Redes de Computadores

Thierno Bocar N'Diaye[1], Viviane Cátia Köhler[2], Célio Trois[2], Guilherme Dhein[2], Olinto César Bassi de Araújo[2]
thierno.ndiaye@imsp-uac.org, viviane@redes.ufsm.br, trois@inf.ufsm.br, gdhein@redes.ufsm.br, olinto@ctism.ufsm.br
[1]University of Nouakchott Al Aasriya - Mauritânia
[2]Universidade Federal de Santa Maria - Brasil

RESUMO

A Orquestração de Telemetria *In-Band* é um problema relacionado a redes de computadores, consistindo no planejamento do envio de itens de telemetria através de cabeçalhos de pacotes. Tais itens são importantes para o monitoramento da rede e, consequentemente, para sua disponibilidade e seu bom desempenho. No presente trabalho, é apresentado um novo modelo matemático para o problema. A primeira característica original do novo modelo é a incorporação de restrição relativa à disponibilidade de memória nos dispositivos da rede, que limita o número de pacotes que podem ser manipulados concomitantemente em cada um deles. Passou a ser considerada também a dependência temporal, que determina um intervalo máximo entre os instantes de recebimento de alguns itens de telemetria pelas aplicações de monitoramento. Resultados obtidos com testes computacionais atestam a coerência do modelo. Além disso, demonstram que as novas características incorporadas, enquanto aproximam o modelo de situações práticas, dificultam a resolução das instâncias.

PALAVRAS CHAVE. Telemetria *In-Band*, Monitoramento de Rede, Dependência Temporal, Programação Linear Inteira Mista

Tópicos. Outras aplicações de PO, Programação Matemática.

ABSTRACT

In-Band Telemetry Orchestration is a problem related to computer networks, involving the planning of telemetry item transmission through packet headers. These items are important for monitoring the network and, consequently, for its availability and performance. This paper presents a new mathematical model for the problem. The first original feature of the new model is the incorporation of a restriction on the availability of memory in the network devices, which limits the number of packets that can be handled concurrently in each of them. Temporal dependency is also taken into account, which determines a maximum interval between the times when some telemetry items are received by the monitoring applications. Computational results confirm the consistency of the model. They also show that the new features incorporated, while bringing the model closer to practical situations, make it more difficult to solve the instances.

KEYWORDS. In-Band Network Telemetry, Network Monitoring, Temporal Dependency, Mixed Integer Linear Programming.

Paper topics. Other applications in OR, Mathematical Programming.

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

1. Introdução

O monitoramento das redes de computadores é uma atividade essencial para garantir o desempenho, a disponibilidade e a segurança das infraestruturas de comunicação. Tradicionalmente, esse processo é realizado por meio de protocolos específicos como SNMP (Simple Network Management Protocol) [Stallings, 1998] ou NetFlow [Claise, 2004], que coletam dados em intervalos definidos, muitas vezes com atrasos significativos em relação aos eventos de rede [Tian e Gao, 2023]. Esse gerenciamento requer monitoramento da infraestrutura, que se torna mais complexo com o crescimento do número de dispositivos e da interconexão.

O surgimento das Redes Definidas por Software (SDN - *software-defined networking*) e da Programação do Plano de Dados (PDP - *programmable data plane*) ampliou significativamente as possibilidades de personalização e controle das redes, reconfigurando os planos de controle e de dados sem comprometer a infraestrutura existente [Tan et al., 2021]. Essas tecnologias abriram caminho para arquiteturas de monitoramento mais flexíveis, permitindo a integração de novos paradigmas, como a telemetria *in-band*, ao mesmo tempo em que mantêm compatibilidade com técnicas tradicionais de coleta e amostragem de fluxos.

Conforme [Dallanora, 2022], telemetria de rede *in-band* consiste na coleta de dados sobre o estado corrente da infraestrutura diretamente do plano de dados. Os dados coletados, posteriormente agrupados, são enviados à aplicação de monitoramento responsável pelo seu processamento, com o objetivo de detectar eventos importantes e, potencialmente, reagir a eles. Os fundamentos da telemetria *in-band* têm origem no início do século corrente, mas a atenção à necessidade de um planejamento otimizado para a coleta e envio dos itens é mais recente [Borges et al., 2024].

É importante salientar que as aplicações de monitoramento são heterogêneas. Cada uma dependente de um conjunto diferente de indicadores (referidos na sequência como itens de telemetria ou apenas itens) para realizar as inferências que identificam o funcionamento correto ou não da rede em relação a alguma métrica específica. O planejamento de uma estratégia que estabelece uma forma de coleta e envio dos itens de telemetria, visando o atendimento das aplicações de monitoramento, configura um problema de otimização combinatória conhecido como o Problema da *Orquestração* da Telemetria *In-band*, usualmente referido apenas como Orquestração. A possibilidade de que cada item de telemetria se origine e seja coletado em diferentes dispositivos da rede, a existência de múltiplas alternativas de pacotes nos quais os dados podem ser incorporados, e restrições relativas à possibilidade de incorporação de cada item de telemetria a cada pacote são exemplos de características que contribuem para a complexidade do problema.

A literatura sobre Orquestração de Telemetrias apresenta diferentes interpretações quanto ao objetivo a ser otimizado, sendo o mais comum a redução da sobrecarga gerada pela inserção de metadados (itens de telemetria) em pacotes de usuário. Esse é o foco de [Bhamare et al., 2022], que propõem um framework para definir um conjunto ideal de fluxos para envio dos pacotes. Já [Marques et al., 2019] propõem modelos matemáticos para duas variantes do problema, ambas visando minimizar a sobrecarga e evitar a saturação da rede: uma busca reduzir o número de fluxos ativos; a outra, balancear a demanda de telemetria entre os fluxos. Também são apresentadas soluções heurísticas para ambas.

A proposta de [Polverini et al., 2024] está focada em buscar correlações espaciais impostas pela topologia da rede e que podem ser exploradas na interpretação dos dados de telemetria. É apresentado um algoritmo que, a partir do conhecimento da topologia da rede e de um conjunto restrito de sinais de telemetria, consegue inferir outros valores não recebidos explicitamente, assim reduzindo a quantidade de itens efetivamente enviados. Em [Chen et al., 2023], os autores procuram integrar a telemetria *in-band* à técnica de roteamento por segmentos. São apresentados um modelo de programação linear inteira mista e soluções baseadas em heurística gulosa e geração de colunas.

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

Em [Hohemberger et al., 2019], o problema da orquestração é abordado com foco na quantidade e qualidade dos itens de telemetria coletados na rede. É apresentado um modelo matemático cujo objetivo é maximizar a quantidade de itens de telemetria recebidos pelas aplicações de monitoramento. Mas a definição sobre quais itens são realmente importantes para serem coletados é feita por um processo a parte, baseado em aprendizagem de máquina. O trabalho de [N'Diaye, 2024] estende o modelo de maximização de itens recebidos, dispensando a pré-definição dos fluxos, que são calculados em tempo de execução como parte da solução do modelo matemático.

A orquestração realizada com foco na maximização do número de itens de telemetria coletados, conforme apresentado em [Hohemberger et al., 2019] e [N'Diaye, 2024], foi a principal motivação para a proposta de duas novas variantes em [Köhler et al., 2024]. Nessas variantes, o foco do modelo matemático deixa de ser a maximização do número de itens de telemetria coletados, passando a ser o número de atendimentos a aplicações. Na primeira variante, maximiza-se o número de aplicações de monitoramento com os requisitos atendidos. A segunda visa promover equidade, maximizando os atendimentos à aplicação menos atendida. Nos dois casos, fica incorporada aos próprios modelos propostos a definição sobre o melhor conjunto de itens de telemetria que devem ser coletados: aquele conjunto que melhor atende à demanda de informação das aplicações.

O presente trabalho apresenta uma modelagem original para o problema da orquestração que se alicerça no trabalho descrito no parágrafo anterior. Mantém-se a função objetivo focada na equidade apresentada em Köhler et al. [2024]. As principais diferenças em relação àquele trabalho e a outros da literatura se encontram em dois novos fatores modelados com a intenção de tornar a representação da rede mais acurada em relação a situações reais. O primeiro fator diz respeito ao limite de capacidade dos dispositivos para o tratamento dos pacotes. É constantemente referida na literatura a sobrecarga (*overhead*) causada pela necessidade de parar um pacote para nele escrever uma telemetria. Mas essa sobrecarga pode se tornar muito mais sensível se o número de pacotes superar uma capacidade do dispositivo relacionada à sua estrutura de memória interna [Trois et al., 2016]. Existe, portanto, um limiar no número de pacotes que podem ser tratados em um dispositivo de forma que a sobrecarga siga moderada, e essa característica foi incorporada ao modelo proposto.

O segundo fator está relacionado ao momento da chegada dos itens de telemetria às aplicações de monitoramento. Mesmo dois itens coletados concomitantemente em um dispositivo, por seguirem por fluxos diferentes, podem chegar às aplicações em instantes separados por um intervalo considerável, a ponto de inviabilizar sua utilização. O novo modelo incorpora uma dependência temporal entre telemetrias que são inter-relacionadas.

O restante do trabalho está organizado da seguinte forma. A Seção 2 traz a descrição do problema. O novo modelo matemático é apresentado na Seção 3. Os resultados computacionais obtidos a partir do modelo são descritos e discutidos na Seção 4 e o artigo se encerra na Seção 5, com as considerações finais.

2. Definição do problema

Considere que a supervisão do funcionamento de uma rede ocorre pela execução de l_m aplicações de monitoramento, cujos índices formam o conjunto $M = \{1, 2, \ldots, l_m\}$. Para seu correto funcionamento, cada aplicação depende do recebimento de um conjunto particular de itens de telemetria coletados potencialmente em diferentes dispositivos. Os índices de todos os l_v itens de telemetria necessários para o funcionamento de todas as aplicações de monitoramento são agrupados no conjunto $V = \{1, 2, \ldots, l_v\}$. Existe a possibilidade de dependência espacial entre itens coletados, que se traduz na obrigatoriedade de que itens específicos sejam obtidos no mesmo dispositivo da rede para que sejam úteis para a aplicação monitoramento. O dispositivo no qual as aplicações de monitoramento são executadas é irrelevante para o problema de orquestração das telemetrias.

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

Na Figura 1 é representado um conjunto simplificado de $l_m = 3$ aplicações de monitoramento. A aplicação M_2, por exemplo, necessita, para seu funcionamento, dos itens de telemetria de índices 1, 2, 4, 5 e 6. Mas a representação dos índices 1 e 4 agrupados indica que os respectivos itens de telemetria apresentam dependência espacial e, portanto, devem ser coletados em um mesmo dispositivo da rede. Outros exemplos são os itens de índices 2 e 6 da mesma aplicação M_2 ou os itens de índices 1, 3 e 4 da aplicação M_1.

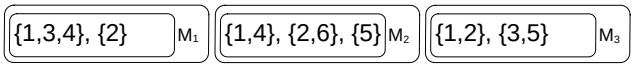


Figura 1: Um exemplo simplificado de 3 aplicações de monitoramento com um total de 6 itens de telemetria

Devido a dependência espacial, as demandas de cada aplicação de monitoramento são estabelecida, portanto, a partir de partições do conjunto V. Para a formalização desse conceito, é preciso considerar o conjunto das partições do conjunto de itens de telemetria, dado por $\mathcal{P}(V)$. A necessidades da aplicação de monitoramento de índice $m \in M$, representada por R_m, corresponde a um subconjunto desse conjunto de partições, ou seja, $R_m \subseteq \mathcal{P}(V)$.

Um conjunto chamado P_m agrupa os índices das partições de R_m, ou seja, cada partição em R_m pode ser referenciado individualmente através de um índice $p \in P_m$. A composição de cada partição em particular é dada por V_{mp}, de modo que cada conjunto V_{mp} define uma coleção de índices de itens de telemetria que precisam ser coletadas de um mesmo dispositivo como requisito parcial para a ativação da aplicação $m \in M$.

Outro conjunto, $P_m^s \subseteq P_m$, armazena os índices das partições da aplicação de índice m que são compostas por mais de uma telemetria. Esse conjunto restringe os índices de P_m àqueles que indicam dependência espacial, pois tal dependência não existe para partições com apenas uma telemetria, sendo exclusiva de partições em que $|V_{mp}| > 1$. Voltando à Figura 1, é possível observar, para fins de exemplo, que $R_2 = \{\{1,4\},\{2,6\},\{5\}\}$, $P_2 = \{1,2,3\}$, $V_{21} = \{1,4\}$ e $P_2^s = \{1,2\}$.

Ao longo do texto, usa-se uma descrição genérica que descreve as aplicações como dependentes de itens de telemetria para o seu funcionamento. Essa afirmação, embora verdadeira em essência, é incompleta, já que o atendimento à aplicações de monitoramento não se dá apenas pelo recebimento de itens de telemetria separados, mas também de itens que precisam estar combinados de modo a atender restrições de dependência. O uso da afirmação mais genérica se dá por motivo de simplicidade, mas é preciso salientar que os referidos itens que devem ser recebidos são sempre organizados em partições de V.

Para o melhor atendimento das demandas de aplicações, é necessário um planejamento que otimize a coleta dos itens de telemetria nos dispositivos da rede. O conjunto de dispositivos é representado por seus índices em $D = \{1, 2, \ldots, l_d\}$ e as conexões existentes entre os dispositivos são representadas no conjunto $L = \{l_{ij} \mid i, j \in D\}$. A rede é representada pelo grafo $G = (D, L)$, no qual os l_d dispositivos estabelecem os vértices e as conexões em L estabelecem as arestas.

O conjunto V de todas as telemetrias necessárias para o monitoramento também pode ser interpretado como o conjunto dos itens que podem ser coletados em dispositivos da rede. Os itens que podem ser providos pelo dispositivo de índice $d \in D$ formam o conjunto $V_d \subseteq V$. Cada item de telemetria tem um tamanho dado por b_v, que estabelece o espaço em bytes que este item ocupa em um pacote caso seja incorporado a ele para envio.

Os pacotes transitam pela rede através de um conjunto de l_f fluxos ativos dado por $F = \{1, 2, \ldots, l_f\}$. O objetivo do problema é planejar, portanto, quais itens devem ser coletados pelo conjunto de pacotes que começam a trafegar pelos fluxos em um determinado instante, um pacote por fluxo. Obviamente, em uma aplicação real, esses instantes se sucedem. Em um caso real, é

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

possível aplicar a mesma configuração planejada em sucessivas iterações, ou então replanejar em instantes em que a configuração da rede tenha sofrido alteração.

Cada fluxo é um caminho sobre o grafo G, estabelecido *a priori*, com pontos inicial e final em dispositivos de D. Os fluxos não são necessariamente disjuntos, podendo compartilhar arestas ou apresentar dispositivos em comum, tanto como pontos de partida, chegada ou ao longo do caminho. O ponto final de um fluxo representa o dispositivo a partir do qual os dados de telemetria, coletados nos dispositivos dispostos ao longo do caminho, são enviados às aplicações de monitoramento. Cada fluxo têm associado a si um parâmetro de capacidade C_f, que representa o espaço em bytes disponível no pacote que por ali trafega para inclusão de itens de telemetria. É esse parâmetro que restringe o número de itens de telemetria que podem ser coletados em cada fluxo.

Perceba que um mesmo item de telemetria pode ser gerado e potencialmente coletado, através de escrita em um pacote, em diferentes dispositivos. Se acrescentarmos a isso a possibilidade de que um mesmo item seja enviado por diferentes fluxos que passam pelo dispositivo no qual é coletado, fica evidente a grande diversidade disponível para a orquestração da coleta das telemetrias.

A Figura 2 ilustra a estrutura da rede descrita até aqui. O grafo é composto por $l_d = 7$ vértices que representam dispositivos, o conjunto de itens de telemetria associados a cada vértice de índice $d \in D$ é representado pelo conjunto de índices V_d e existem $l_f = 4$ fluxos identificados por cores distintas. O fluxo azul, por exemplo, se origina no dispositivo de índice 3 e se completa no dispositivo de índice 7, após passar pelos arcos $(3, 1)$, $(1, 2)$, $(2, 5)$ e $(5, 7)$. Observe que todos os fluxos terminam nos dispositivos 2 ou 7, de onde são repassados para as aplicações. Linhas tracejadas indicam conexões possíveis, mas não incluídas em qualquer fluxo.

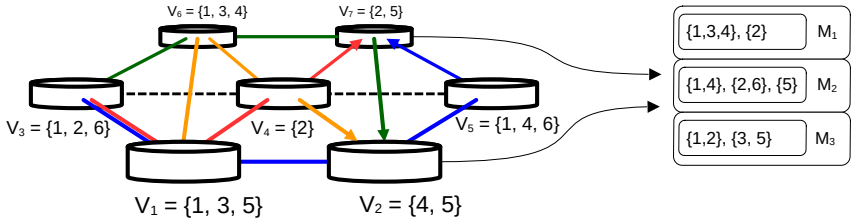


Figura 2: A estrutura do grafo para uma rede simplificada, com 7 dispositivos e as 3 aplicações da Figura 1

A estrutura descrita até esse ponto repete a considerada no artigo Köhler et al. [2024] e é a estrutura básica para diversos trabalhos voltados à orquestração de telemetrias *in-band*. As novidades do presente trabalho, apontadas na Seção 1, demandam mais algumas definições.

O limite no número de pacotes que podem ser atendidos em um dispositivo é dado por L_d. Esse limite é importante devido à possibilidade de que um dispositivo faça parte de diversos fluxos diferentes, podendo ser colocado como ponto de coleta de telemetria em grande número deles e, consequentemente, se tornando um gargalo. Em situações reais, o desrespeito a esse limite não infactibiliza a solução, provocando apenas um atraso relacionado à hierarquia de memória do dispositivo muito mais significativo que a sobrecarga inerente à escrita das telemetrias nos pacotes. Na modelagem do problema aqui proposta, porém, esse atraso de maior vulto é considerado inaceitável, e soluções que não estão de acordo com esse limite são consideradas infactíveis.

A dependência temporal está relacionada à dependência espacial. Para a sua compreensão, é importante destacar que a dependência espacial, embora torne obrigatória a coleta de todos os itens de qualquer partição V_{mp} em um mesmo dispositivo sempre que $|V_{mp}| \geq 2$, não estabelece qual é esse dispositivo e nem o fluxo de envio. Para fins de exemplo, considere, novamente na Figura 2, a partição $\{3, 5\} \in \mathcal{P}(V)$. Essa partição é necessária para a aplicação M_3 (já que $\{3, 5\} \in R_3^s$) e só pode ser atendida pelo dispositivo 1, pois apenas V_1 contém os dois índices de itens de telemetria.

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

Apesar de apenas um dispositivo ser capaz de atender à dependência temporal, existem diversas possibilidades para o envio dos itens de telemetria. É possível enviar ambos por apenas um entre os fluxos azul, vermelho e laranja, ou enviar cada uma das telemetrias por um desses fluxos.

A configuração em que um item de telemetria segue pelo fluxo azul enquanto outro item segue pelo fluxo vermelho está entre as viáveis. Nesse caso, o item enviado pelo fluxo vermelho passará por apenas um dispositivo intermediário (4) antes de chegar ao dispositivo final (7) do fluxo, enquanto o item enviado pelo fluxo azul passará por dois dispositivos intermediários (2 e 5). Esse caminho mais longo, por si só, pode resultar em um atraso na entrega do item. O custo inerente ao tamanho do caminho é pré-calculado para todos os vértices e todas as possibilidades de fluxos. Ou seja, para cada vértice $d \in D$ e para cada fluxo $f \in F_d$, sendo $F_d \subset F$ o conjunto dos índice dos fluxos que passam pelo dispositivo d, existe um parâmetro l_{df} que representa esse custo de envio.

Além desse custo, a necessidade de escrever em um pacote de um fluxo gera uma sobrecarga. O tempo para efetivamente escrever cada item de telemetria no cabeçalho de um pacote é considerado irrelevante frente ao tempo necessário para preparar o pacote para escrita. Por isso, o tempo de escrita é específico para cada dispositivo, dado por w_d, mas o número de telemetrias que serão escritas pelo dispositivo não é considerado. Como resultado, itens que transitam por fluxos nos quais mais dispositivos incorporam telemetrias aos pacotes também têm atrasos maiores.

Fica claro que, embora o planejamento considere um ciclo no qual os pacotes partem juntos das origens dos fluxos, eles podem chegar aos seus destinos em tempos diversos. A dependência temporal decorre do fato de que alguns itens de telemetria que apresentam dependência espacial também apresentam um limite máximo para a diferença entre os seus tempos de recebimento nas aplicações (ou no final dos fluxos). As partições de uma aplicação $m \in M$ que apresentam dependência temporal são representadas pelo conjunto de índices $P^t_m \subseteq P^s_m$, o que explicita que a essa dependência se aplica a um subconjunto das partições que indicam dependência espacial. Havendo dependência temporal estabelecida para a partição de índice p de P^s_m, ou seja, se o item de índice p também faz parte de P^t_m, a diferença máxima aceitável para os tempos de recebimento dos itens de telemetria dessa partição é representado por T_{mp}.

O objetivo geral da proposta do presente trabalho é orquestrar o envio de itens de telemetria através dos espaços disponíveis em pacotes trafegando pelos fluxos, de modo a otimizar o número de aplicações atendidas conforme a função objetivo proposta no segundo modelo do trabalho de [Köhler et al., 2024]. É importante destacar que o recebimento do maior número possível de partições V_{mp} não implica no atendimento do maior número de aplicações. Uma aplicação m só é atendida se todo seu conjunto de partições R_m é atendido, respeitadas as dependências espacial e temporal. Por isso, a cobertura dos conjuntos R_m é mais importante que o simples número de itens de telemetria ou partições individuais recebidas.

Da mesma forma, atender o maior número de aplicações não implica em um melhor monitoramento da rede supervisionada. Como a redundância é permitida, a busca pelo maior número de ativações de aplicações de monitoramento pode resultar em muitas ativações de algumas (notadamente aquelas que necessitam de poucos itens de telemetria) em detrimento de outras. Por isso, optou-se pela função objetivo focada em maximizar o número de atendimentos à aplicação menos atendida, denotado por C_{min}. Evita-se, assim, que uma ou mais aplicações sejam negligenciadas, pois são os atendimentos à menos atendida que definem o sucesso do planejamento. Este objetivo ainda favorece, de forma indireta, uma distribuição balanceada entre as aplicações atendidas.

A possibilidade de redundância é interessante, por possibilitar uma maior acuidade nas inferências feitas pelas aplicações. Por isso, a função objetivo contempla também a maximização do somatório dos atendimentos aos conjuntos de partições das aplicações, através de um objetivo secundário cujo peso é regulado por uma parâmetro. Por outro lado, informações incompletas (R_m

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

sendo recebidos incompleto) impedem o correto funcionamento de uma aplicação. Se os requisitos de uma aplicação não puderem ser satisfeito na totalidade, é irrelevante que ela receba qualquer item de telemetria, sendo interessante que o espaço ou seja deixado sem utilização, evitando sobrecarga desnecessária na rede, ou seja destinado para a redundância nas informações de outras aplicações.

Por fim, é importante salientar que a orquestração de telemetrias foca exclusivamente no atendimento das demandas das aplicações de monitoramento. A forma como os itens serão utilizados nas aplicações de monitoramento e a funcionalidade de cada aplicação não estão no escopo do problema de orquestração.

3. Modelo Matemático

O modelo proposto nesse trabalho para o Problema de Orquestração de Telemetria *In-Band*, denominado INT-E, é apresentado nessa seção. O modelo tem como objetivo principal a maximização do número de vezes que a aplicação menos atendida tem seus requisitos satisfeitos. A maximização do número de aplicações atendidas é colocada como um objetivo secundário. O modelo INT-E é formalizado a seguir.

Conjuntos

- F: conjunto de fluxos de dados na rede;
- D: conjunto de vértices do grafo que representam os dispositivos da rede;
- V: conjunto de itens de telemetria;
- V_d: conjunto de itens de telemetria que podem ser extraídos do dispositivo $d \in D$;
- F_d: conjunto de fluxos que passam pelo dispositivo $d \in D$ na rede;
- M: conjunto das aplicações voltadas ao monitoramento;
- P_m: conjunto de índices das partições da aplicação $m \in M$;
- P_m^t: índices de partições da aplicação m que possuem dependência temporal;
- V_{mp}: itens da telemetria na partição $p \in P_m$ da aplicação $m \in M$;
- D_{mp}: conjunto de dispositivos que podem enviar itens de telemetria da partição $p \in P_m$ da aplicação $m \in M$;

Variáveis

- y_{dvf}: variável binária que determina se o item de telemetria $v \in V$ é coletado no dispositivo $d \in D$ e enviado pelo fluxo $f \in F$;
- r_{df}: variável binária que indica se o dispositivo $d \in D$ enviou algum item de telemetria pelo fluxo $f \in F_d$;
- s_{mpd}: variável binária que indica se a partição $p \in P_m$ da aplicação $m \in M$ foi atendida pelo dispositivo $d \in D$;
- $\hat{s}_m$: variável inteira que indica o número de vezes em que a aplicação $m \in M$ teve todos os seus requisitos de itens de telemetria satisfeitos;
- C_{min}: variável inteira que representa o menor número de vezes em que uma aplicação $m \in M$ teve todos os seus requisitos atendidos;

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

Modelo INT-E

$$\textbf{Max} \quad C_{min} + \epsilon \sum_{m \in M} \hat{s}_m \tag{1}$$

s.a.:

$$\sum_{d \in D} \sum_{v \in V_d} b_v y_{dvf} \leq C_f \qquad \forall f \in F \tag{2}$$

$$\sum_{f \in F_d} y_{dvf} \leq 1 \qquad \forall d \in D, \forall V \in V_d \tag{3}$$

$$\sum_{v \in V_d} \sum_{f \in F_d} y_{dvf} \leq L_d \qquad \forall d \in D \tag{4}$$

$$t_{dvf} \geq l_{df} + \sum_{d' \in D_{df}} w_d \, r_{d'f} - \mathbb{M}(1 - y_{dvf}) \qquad \forall d \in D, \forall v \in V_d, \forall f \in F_d \tag{5}$$

$$t_{dvf} \leq l_{df} + \sum_{d' \in D_{df}} w_d \, r_{d'f} + \mathbb{M}(1 - y_{dvf}) \qquad \forall d \in D, \forall v \in V_d, \forall f \in F_d \tag{6}$$

$$\sum_{v \in V_d} y_{dvf} \leq |V_d| \, r_{df} \qquad \forall d \in D, \forall f \in F_d \tag{7}$$

$$\hat{t}_{\max_{mpd}} \geq t_{dvf} \qquad \forall m \in M, \forall p \in P^t_m, \forall d \in D_{mp}, \forall v \in V_{mp}, \forall f \in F_d \tag{8}$$

$$\hat{t}_{\min_{mpd}} \leq t_{dvf} \qquad \forall m \in M, \forall p \in P^t_m, \forall d \in D_{mp}, \forall v \in V_{mp}, \forall f \in F_d \tag{9}$$

$$\hat{t}_{\max_{mpd}} - \hat{t}_{\min_{mpd}} - T_{mp}(1 - \tau_{mpd}) \leq T_{mp} \qquad \forall m \in M, \forall p \in P^t_m, \forall d \in D_{mp} \tag{10}$$

$$s_{mpd} \leq \tau_{mpd} \qquad \forall m \in M, \forall p \in P^t_m, \forall d \in D_{mp} \tag{11}$$

$$s_{mpd} \leq \sum_{f \in F} y_{dvf} \qquad \forall m \in M, \forall p \in P_m, \forall d \in D_{mp}, \forall v \in V_{mp} \tag{12}$$

$$\hat{s}_m \leq \sum_{d \in D_{mp}} s_{mpd} \qquad \forall m \in M, \forall p \in P_m \tag{13}$$

$$C_{min} \leq \hat{s}_m \qquad \forall m \in M \tag{14}$$

$$y_{dvf} \in \{0, 1\} \qquad \forall d \in D, \forall v \in V_d, \forall f \in F \tag{15}$$

$$r_{df} \in \{0, 1\} \qquad \forall d \in D, \forall f \in F_d \tag{16}$$

$$t_{dvf} \in \mathbb{N} \qquad \forall d \in D, \forall v \in V_d, \forall f \in F_d \tag{17}$$

$$\hat{t}_{\max_{mdp}} \in \mathbb{N} \qquad \forall m \in M, \forall p \in P_m, \forall d \in D_{mp} \tag{18}$$

$$\hat{t}_{\min_{mdp}} \in \mathbb{N} \qquad \forall m \in M, \forall p \in P_m, \forall d \in D_{mp} \tag{19}$$

$$\tau_{mpd} \in \{0, 1\} \qquad \forall m \in M, \forall p \in P_m, \forall d \in D_{mp} \tag{20}$$

$$s_{mpd} \in \{0, 1\} \qquad \forall m \in M, \forall p \in P_m, \forall d \in D_{mp} \tag{21}$$

$$\hat{s}_m \in \mathbb{N} \qquad \forall m \in M \tag{22}$$

$$C_{min} \in \mathbb{N} \tag{23}$$

A função objetivo (1) considera a maximização de duas parcelas de forma hierárquica: C_{min}, que representa o número de vezes em que os requisitos da aplicação menos atendida são satisfeitos, e o total de vezes em que os requisitos de todas as aplicações são atendidos. O parâmetro ϵ é definido de modo que a maximização da segunda parcela não degrade o valor da primeira. A restrição (2) assegura que a capacidade de cada fluxo $f \in F$ seja atendida, garantindo que a soma dos tamanhos dos itens de telemetria atribuídos a um dado fluxo não exceda a correspondente capacidade. A restrição (3) estabelece que cada item de telemetria $v \in V$ deve ser enviado por apenas um único fluxo $f \in F$, a partir de um dispositivo $d \in D$. O limite de telemetrias que um dispositivo pode enviar, independentemente do fluxo, é determinado pela restrição (4).

As restrições (5–10) modelam a dependência temporal, ou seja, garantem que telemetrias

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

da partição de índice $p \in P_m$ da aplicação $m \in M$, também relacionadas por dependência espacial, cheguem ao nó de destino respeitando a diferença máxima de tempo T_{mp}, mesmo quando enviadas por fluxos distintos. As restrições (5) e (6) determinam o tempo necessário para que um item de telemetria atinja o destino quando enviado pelo dispositivo $d \in D$, por meio do fluxo $f \in F_d$. Esse tempo é composto por um custo fixo de envio, representado por l_{df}, somado ao tempo de escrita w_d, que corresponde ao processamento realizado pelos demais dispositivos ao longo do caminho até o destino. Este último procedimento está em sinergia com a restrição (7), que determina se o dispositivo $d \in D$ enviou algum item de telemetria pelo fluxo $f \in F_d$. É importante observar que, caso o dispositivo d envie o item de telemetria $v \in V_d$ por meio do fluxo f, o tempo de envio t_{dvf} é determinado de forma única. Caso contrário, ou seja, se o envio não ocorrer, essa variável permanece livre para assumir qualquer valor. Por fim, as restrições (8) - (10) determinam se a diferença de tempo de chegada no destino das telemetrias que compõem a partição de índice p da aplicação m satisfaz a diferença de tempo máxima T_{mp}.

As restrições (11) e (12) verificam se todos os requisitos da partição p, associada à aplicação m e composta por itens de telemetria provenientes do dispositivo d, foram atendidos, considerando, respectivamente, a dependência temporal — quando aplicável — e a dependência espacial. Note que uma partição pode ter seus requisitos atendidos por mais de um dispositivo. Consequentemente, uma aplicação pode ser considerada atendida mais de uma vez, desde que todas as suas partições tenham seus respectivos requisitos satisfeitos. A contagem do número de vezes em que uma partição é atendida é determinada pela restrição (13). A restrição (14) determina o valor de $Cmin$, enquanto as restrições (15) - (23) determinam o domínio das variáveis.

4. Resultados computacionais

O modelo matemático foi implementado em Python 3.8 e resolvido com o solver IBM ILOG CPLEX 20.1.0, usando configuração padrão, até 8 *threads* e limite de 1800 segundos por execução. Os testes foram realizados em um processador Intel Xeon E5-2697 v2 (2.70GHz) com 64GB de RAM. As instâncias seguiram as diretrizes de [N'Diaye, 2024], sendo geradas aleatoriamente sobre redes com 50 nós. Para cada combinação entre 50, 100 ou 150 fluxos; 8 ou 16 itens de telemetria; e 2, 4 ou 6 aplicações, foram produzidas 10 instâncias, totalizando 180.

Os tamanhos dos itens de telemetria e as capacidades dos fluxos foram definidos com valores inteiros uniformemente distribuídos nos intervalos $[10, 20]$ e $[10, 40]$, respectivamente. Os parâmetros adicionais foram definidos assim: o intervalo máximo entre chegadas de itens com dependência temporal varia entre 5 e 15; o tempo de escrita por dispositivo está entre 1 e 5 unidades de tempo; e a capacidade de tratamento simultâneo de pacotes por dispositivo d varia entre $[g/2, g]$, em que g é o número de fluxos que passam por d.

A Tabela 1 resume os resultados dos testes, organizados por configuração de instância, variando número de aplicações, fluxos e itens de telemetria (colunas iniciais). Cada linha apresenta a média dos resultados obtidos em 10 instâncias. Também é incluída uma linha com a média geral para configurações com o mesmo número de aplicações.

As seis colunas seguintes da Tabela 1 apresentam métricas de qualidade das soluções: tempo de execução, número de atendimentos completos por aplicação, número de atendimentos à aplicação menos favorecida, número de partições completas coletadas, *gap* médio (devido a soluções não ótimas em 1800s) e coeficiente de Gini sobre os atendimentos. Essas métricas são exibidas para os modelos INT-C e INT-E, o primeiro proposto na literatura e o último neste trabalho. Os resultados referentes ao modelo INT-C foram gerados com a implementação do artigo Köhler et al. [2024], obtida com os autores. Embora compartilhem a função objetivo, os modelos possuem restrições distintas, o que impede uma comparação qualitativa direta. A comparação com INT-C visa medir o impacto das novas restrições introduzidas no INT-E.

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

Tabela 1: Rede com 50 nós

Instâncias			Modelo INT-C						Modelo INT-E					
|M|	|F|	|V|	Tempo	$\sum \hat{s}_m$	C_{min}	$\sum s_{mpd}$	gap_C	*Gini*	Tempo	$\sum \hat{s}_m$	C_{min}	$\sum s_{mpd}$	gap_C	*Gini*
2	50	8	7,7	26,6	12,9	159,3	0,0	0,068	316,3	26,2	12,6	155,0	0,9	0,074
		16	273,2	19,7	9,2	123,9	1,4	0,112	921,9	19,2	9,0	120,4	2,6	0,112
	100	8	0,4	43,2	16,9	172,8	0,0	0,115	45,3	43,2	16,9	172,8	0,0	0,113
		16	75,9	32,3	14,3	235,1	0,0	0,090	746,9	31,3	13,7	222,6	6,7	0,090
	150	8	0,2	40,2	16,6	170,8	0,0	0,029	18,3	40,2	16,6	170,8	0,0	0,028
		16	66,3	33,4	13,0	249,4	0,0	0,017	392,5	33,0	12,9	244,8	0,6	0,021
Média			**70,6**	**32,6**	**13,8**	**185,2**	**0,2**	**0,072**	**406,9**	**32,2**	**13,6**	**181,1**	**1,8**	**0,073**
4	50	8	571,3	45,5	9,9	208,3	0,0	0,038	918,6	44,9	9,5	199,1	2,4	0,074
		16	638,1	25,9	5,8	186,3	0,0	0,153	1612,1	24,3	5,3	173,1	5,6	0,156
	100	8	564,9	73,0	13,1	343,1	0,0	0,148	1240,2	71,3	12,8	331,2	2,5	0,178
		16	972,6	49,4	11,6	350,2	0,0	0,106	1800,0	34,7	7,6	197,3	19,1	0,106
	150	8	0,3	86,9	16,4	365,3	0,0	0,052	138,7	86,9	16,4	365,3	0,0	0,070
		16	1086,7	62,9	10,8	444,2	0,1	0,054	1282,3	58,1	9,1	403,2	27,4	0,081
Média			**639,0**	**57,3**	**11,3**	**316,2**	**0,0**	**0,092**	**1165,4**	**53,4**	**10,1**	**278,2**	**9,5**	**0,113**
6	50	8	201,8	61,3	8,8	283,1	0,0	0,081	1197,2	59,3	8,3	270,8	8,0	0,214
		16	429,8	33,8	5,0	249,9	0,0	0,150	1661,4	31,6	4,2	220,3	16,1	0,169
	100	8	627,0	106,2	12,7	539,6	0,0	0,150	1514,7	99,2	11,1	492,3	17,5	0,394
		16	1474,8	63,7	9,2	518,2	1,1	0,178	1800,0	58,8	6,6	413,3	44,8	0,126
	150	8	2,5	136,0	13,3	610,0	0,0	0,162	587,3	133,6	13,1	591,8	2,0	0,181
		16	1329,6	87,7	10,2	639,6	0,1	0,070	1800,0	56,0	5,6	362,9	31,6	0,235
Média			**677,6**	**81,5**	**9,9**	**473,4**	**0,2**	**0,132**	**1426,9**	**72,9**	**8,1**	**390,1**	**25,2**	**0,172**

A comparação dos modelos revela um aumento significativo na dificuldade de obtenção da solução ótima no INT-E, devido às novas restrições. Isso se reflete tanto no maior tempo de resolução quanto no número de testes encerrados por tempo esgotado. Em 3 das 18 configurações, todas as 10 instâncias atingiram o limite de 1800s, e o tempo médio para instâncias com 4 ou 6 aplicações superou 1000 segundos. Os *gaps* observados reforçam a complexidade do novo problema, indicando a necessidade do uso de metaheurísticas para obter soluções viáveis em contextos práticos.

A ausência de otimalidade em algumas instâncias limita uma avaliação geral da função objetivo. No entanto, nas instâncias com *gap* zero, as soluções do modelo INT-C superam as do INT-E tanto nos atendimentos à aplicação menos favorecida quanto no total de aplicações atendidas. Isso era esperado, pois o modelo INT-C é uma relaxação do INT-E, e qualquer solução viável para o INT-E é também viável para o INT-C — mas não o contrário. Portanto, o INT-E não pode produzir soluções melhores em termos da função objetivo.

Em alguns casos, como nas linhas relativas a instancias com $|M| = 2$, $|F| = 150$ e $|V| = 8$ ou $|M| = 4$, $|F| = 150$ e $|V| = 8$, entre outras, o resultado relativo ao número de atendimentos a aplicações é o mesmo para ambos modelos, o que representa o melhor resultado possível para o modelo INT-E. Nas demais, em que não há igualdade, o resultado relativo ao modelo novo é ligeiramente inferior. As novas restrições representam um acréscimo nos requisitos a serem cumpridos para o atendimento à aplicações, e essa maior dificuldade fica explicita nesses números.

O Coeficiente de Gini (24), comum na economia para medir desigualdade, é aqui aplicado para avaliar a equidade no atendimento às aplicações. Variando entre 0 e 1, valores próximos de 0 indicam distribuição equitativa, enquanto valores maiores refletem concentração. Na prática, o limite superior depende do número de elementos e é inferior a 1 em conjuntos finitos. No contexto analisado, Gini zero indica atendimento igualitário entre as aplicações, e valores mais altos revelam desequilíbrios.

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

$$G = \frac{\sum_{m \in M} \sum_{m' \in M} |\hat{s}_m - \hat{s}_{m'}|}{2|M| \cdot \sum_{m \in M} \hat{s}_m} \tag{24}$$

A equidade tende a melhorar com a maximização dos atendimentos à aplicação menos favorecida, pois isso reduz a disparidade entre aplicações. No entanto, a segunda parte da função objetivo — que busca maximizar o total de atendimentos — pode prejudicar essa equidade, favorecendo aplicações com demandas menores e menos restrições. Embora esse objetivo tenha peso reduzido na parametrização, ele ainda influencia a solução. Assim, o Coeficiente de Gini é usado apenas como métrica auxiliar para avaliar o impacto indireto do modelo na equidade, sem ser um objetivo direto nem critério de validade dos modelos.

A comparação entre os modelos mostra um aumento geral do Coeficiente de Gini no INT-E, indicando perda de equidade com as novas restrições. Como já mencionado, a maximização dos atendimentos tende a favorecer aplicações com demandas menos restritivas. Com a introdução das restrições de dependência temporal no INT-E, há um aumento tanto na complexidade das demandas quanto na disparidade entre elas. Assim, o resultado é coerente com a nova modelagem proposta.

Em duas linhas da tabela, referentes à configurações de instâncias com $|M| = 2$, $|F| = 150$ e $|V| = 8$ ou $|M| = 6$, $|F| = 100$ e $|V| = 16$, o coeficiente Gini é menor para os resultados do modelo INT-E. Uma explicação para esse caso é o fato de que não foram obtidas soluções ótimas para todas as instâncias, e o cálculo é realizado com valores que não representam as soluções finais.

Além da dependência temporal, o modelo INT-E introduz um limite para o número de pacotes processados simultaneamente por dispositivo. Para avaliar o impacto de cada fator, realizaram-se testes com a remoção alternada dessas restrições. Os resultados indicam que a dependência temporal tem influência significativamente maior, afetando tanto o tempo para obter a solução ótima quanto sua configuração em relação ao modelo INT-C.

5. Considerações finais

Este trabalho propôs uma nova modelagem para o Problema da Orquestração de Telemetria *In-Band*, incorporando restrições de dependência temporal e de limite de pacotes por dispositivo ao modelo existente. O novo modelo, denominado INT-E, foi testado com instâncias sintéticas, assim como o modelo da literatura (INT-C). Embora não seja possível uma comparação qualitativa direta, dado que os modelos possuem restrições distintas, os resultados permitem avaliar o impacto das novas restrições introduzidas. De um modo geral, os resultados indicam que o modelo INT-E apresenta soluções cuja obtenção requer mais tempo de processamento e que apresentam resultados ligeiramente piores no que se refere ao número de atendimentos às demandas de aplicações.

Essa dificuldade é inerente às novas características incorporadas ao modelo, e que se originaram de necessidades de aplicações reais de orquestração. Ao invés de apontar para a necessidade de se voltar ao modelo original, o conjunto de resultados demonstra a maior complexidade inerente ao problema de orquestração quando o modelo incorpora novas características reais. E demonstram também a necessidade de se buscar alternativas viáveis para a resolução de instâncias em tempo razoável para uso efetivo das soluções encontradas.

O estudo para a futura modelagem de novas características do problema, notadamente a atribuição de pesos que diferenciem as aplicações de monitoramento, dando-lhes níveis de importância, ou a ponderação do grafo, aumentando a acuidade na definição do tempo de transporte de um pacote por um fluxo, são possíveis caminhos para a sequência da pesquisa. Mas desponta como necessária a busca por métodos baseados em heurísticas, para garantir a obtenção rápida de soluções de qualidade.

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br

Referências

Bhamare, D., Kassler, A., Vestin, J., Khoshkholghi, M. A., Taheri, J., Mahmoodi, T., Öhlén, P., e Curescu, C. (2022). Intopt: In-band network telemetry optimization framework to monitor network slices using p4. *Computer Networks*, 216. ISSN 13891286.

Borges, E. S., Martinello, M., Bonella, V. B., dos Santos, A. J., Gomes, R. L., Dominicini, C. K., Guimarães, R. S., Menegueti, G. T., Barcellos, M., e Ruffini, M. (2024). Pot-polka: Let the edge control the proof-of-transit in path-aware networks. *IEEE Transactions on Network and Service Management*, 21(4):3681–3691.

Chen, B., Chen, F., Tang, S., Zheng, Q., e Zhu, Z. (2023). On orchestration of segment routing and in-band network telemetry. *IEEE Transactions on Network and Service Management*, 20: 4047–4060. ISSN 19324537.

Claise, B. (2004). Cisco systems netflow services export version 9. RFC, RFC Editor. URL `https://www.rfc-editor.org/rfc/rfc3954`.

Dallanora, L. M. (2022). Orquestração dinâmica da telemetria de rede in-band em planos de dados programáveis. Master's thesis, Universidade Federal do Pampa.

Hohemberger, R., Castro, A. G., Vogt, F. G., Mansilha, R. B., Lorenzon, A. F., Rossi, F. D., e Luizelli, M. C. (2019). Orchestrating in-band data plane telemetry with machine learning. *IEEE Communications Letters*, 23:2247–2251. ISSN 15582558.

Köhler, V. C., Dhein, G., Turchetti, R. C., e de Araújo, O. C. B. (2024). Modelos matemáticos para orquestração de telemetria in-band. In *LVI Simpósio Brasileiro de Pesquisa Operacional*, volume 56.

Marques, J. A., Luizelli, M. C., da Costa Filho, R. I. T., e Gaspary, L. P. (2019). An optimization-based approach for efficient network monitoring using in-band network telemetry. *Journal of Internet Services and Applications*, 10. ISSN 18690238.

N'Diaye, T. B. (2024). *A New Proposal for Optimizing the Routing of the In-Band Network Telemetry Problem*. PhD thesis, Institut de Mathématiques et des Sciences Physiques (IMSP-Benin). To be published.

Polverini, M., Sardellitti, S., Barbarossa, S., Cianfrani, A., Lorenzo, P. D., e Listanti, M. (2024). Reducing the in band network telemetry overhead through the spatial sampling: Theory and experimental results. *Computer Networks*, 242. ISSN 13891286.

Stallings, W. (1998). *SNMP, SNMPv2, SNMPv3, and RMON 1 and 2*. Addison-Wesley Longman Publishing Co., Inc.

Tan, L., Su, W., Zhang, W., Lv, J., Zhang, Z., Miao, J., Liu, X., e Li, N. (2021). In-band network telemetry: A survey. *Computer Networks*, 186:107763.

Tian, Y.-C. e Gao, J. (2023). Network management architecture. In *Network Analysis and Architecture*, p. 321–359. Springer.

Trois, C., Bona, L., Fabro, M. D., e Martinello, M. (2016). Carving software-defined networks for scientific applications with spaten. In *2016 IEEE 41st Conference on Local Computer Networks (LCN)*, p. 200–203. IEEE.

DOI: 10.59254/sbpo-2025-212269
https://proceedings.science/p/212269?lang=pt-br