

Modelos Matemáticos para Orquestração de Telemetria *In-Band*

**Viviane Cátia Köhler, Rogério Correa Turchetti,
Guilherme Dhein, Olinto César Bassi de Araújo**

viviane@redes.ufsm.br , turchetti@redes.ufsm.br
gdhein@redes.ufsm.br, olinto@ctism.ufsm.br
Universidade Federal de Santa Maria

RESUMO

Telemetria in-band estabelece que sinais para monitoramento de rede de computadores podem ser enviados através de pacotes de dados de usuário. Como o espaço nesses pacotes é limitado e as aplicações de monitoramento são heterogêneas em relação à característica monitorada, exigindo diferentes informações para o correto funcionamento, o problema combinatório é decidir qual o conjunto de itens de telemetrias que otimiza o gerenciamento geral da rede. O Problema de Orquestração de Telemetria In-Band apresenta variantes focadas na redução da sobrecarga causada pela inclusão dos dados de telemetria nos fluxos. Neste trabalho são estudadas duas variantes do problema focadas nas aplicações de monitoramento e os respectivos modelos matemáticos. Os modelos propostos permitem resolver na otimalidade de instâncias com 50 e 100 dispositivos de rede e 50, 100 e 150 fluxos. São apresentados resultados computacionais comparativos com um modelo que considera como critério de otimização a maximização da coleta de telemetrias.

PALAVRAS CHAVE. Telemetria In-Band, Monitoramento de Rede, Otimização, Programação Linear Inteira Mista

Tópicos. Outras aplicações de PO, Programação Matemática.

ABSTRACT

In-band telemetry establishes that computer networks monitoring signals can be sent via user data packets. As the space in these packets is limited and the monitoring applications are heterogeneous about the monitored characteristic, requiring different information for correct operation, the combinatorial problem is to decide which set of telemetry items optimizes the overall management of the network. The In-Band Telemetry Orchestration Problem has variants focused on reducing the overhead caused by including telemetry data in flows. This paper studies two problem variants focused on monitoring applications and their respective mathematical models. The proposed models allow the optimal solution of instances with 50 and 100 network devices and 50, 100, and 150 flows. Comparative computational results are presented with a model that considers maximizing telemetry collection as an optimization criterion.

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

Paper topics. Other applications in OR, Mathematical Programming.

1. Introdução

O monitoramento é fundamental para gerenciamento e controle eficientes de redes de computadores, e passa pelo acesso a informações sobre o funcionamento corrente da infraestrutura. A telemetria de rede *in-band* consiste na coleta de estatísticas de dados sobre o comportamento da rede diretamente do plano de dados e, em algum ponto da infraestrutura, as informações são extraídas e transmitidas para uma aplicação de monitoramento para processamento e reação aos eventos observados [Dallanora, 2022]. Este tratamento das informações coletadas é realizado por uma aplicação fora do plano de dados/controlado da rede, podendo afetar o próprio desempenho das aplicações de monitoramento [Hohemberger et al., 2019]. A telemetria *in-band* se coloca como um paradigma para tratar o envio dessas informações, e se caracteriza pelo transporte de indicadores de *status* através de sua incorporação em pacotes de dados de usuário, e não através de pacotes gerados com propósito específico. Desta forma, informações sobre o *status* de uma rede são transmitidas pela inserção de metadados nos pacotes de dados de usuários.

Neste contexto, podemos utilizar uma Rede Definida por Software (SDN – *Software-Defined Network*) para implementar Funções Virtualizadas de Rede (NFV – *Network Function Virtualization*) com habilidades de realizar a telemetria mais próximo do plano de dados, reduzindo atrasos de monitoramento. Para ilustrar melhor este cenário, considere que nas Redes SDN o plano de controle é separado do plano de dados [Kreutz et al., 2015]. Concretamente, isto significa que a programação dos *switches* é realizada por um controlador que é responsável pelo gerenciamento no plano de controle. Apesar de simples, esta abordagem possibilita solucionar problemas e limitações conhecidas da arquitetura de rede atual [Farris et al., 2019; Chiras et al., 2024]. Neste plano de controle é possível adicionar NFVs para tratar diversos problemas da rede com a vantagem de estar próximo ao plano de dados, trazendo agilidade à coleta e monitoramento dos dados [Turchetti e Duarte, 2017].

Em geral, as NFVs exploram tecnologias de virtualização para transformar equipamentos de redes em entidades virtuais [ETSI, 2015]. Estas entidades podem ser implementadas em conjunto com outras tecnologias emergentes [Han et al., 2015], este é o caso do cenário ilustrado anteriormente que considera a execução em redes SDN. Neste contexto, as NFVs são utilizadas para prover soluções no plano de controle como também no plano de dados da rede. Indústria, academia e organismos de padronização têm mostrado interesse em projetos para NFV, pois podem prover novos serviços aos usuários em um rápido ciclo de desenvolvimento, reduzindo custos de aquisição de equipamentos físicos e o custo global necessário para o gerenciamento de um grande número de dispositivos [Xilouris et al., 2014; Haleplidis et al., 2014].

A tecnologia de virtualização permite a definição do problema da orquestração da telemetria *in-band* como um problema de otimização, que consiste em elaborar um plano de coleta de dados da rede e planejar seu envio através dos pacotes. O objetivo é garantir a eficácia de aplicações de monitoramento, cada uma dependente de um conjunto específico de informações para que possa realizar inferências e apresentar diagnósticos sobre o estado corrente da rede. A complexidade do problema deriva do fato de que o espaço disponível em pacotes de usuário não é um recurso ilimitado. Portanto, é preciso planejar o envio dos diversos itens de telemetria através do espaço disponível limitado nos fluxos da rede, de modo a atender da melhor forma possível as necessidades das aplicações de monitoramento.

Conforme [Tan et al., 2021], os fundamentos da telemetria *in-band* remontam aos primeiros anos do corrente século. Mas a atenção à orquestração é mais recente, geralmente com o foco de atender às necessidades de monitoramento reduzindo a sobrecarga na rede induzida pela inclusão de nova informação. Este é o principal objetivo da forma de orquestração proposta em [Bhamare et al., 2022], trabalho no qual os autores propõem um *framework* para a especificação de um con-

junto ideal de fluxos a serem usados para o envio de itens de telemetria pela rede. Os autores de [Marques et al., 2019] propõem modelos para duas variantes de orquestração de telemetria *in-band*: na primeira, o objetivo é minimizar o número de fluxos ativos necessários, enquanto na segunda o objetivo é balancear as demandas de telemetria entre os fluxos. Nos dois casos, a intenção é evitar saturação em pontos da rede e, de forma subjacente, minimizar a sobrecarga. São apresentadas ainda soluções heurísticas para as duas variantes.

A abordagem proposta em [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. Os autores apresentam 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.

Já em [Hohemberger et al., 2019], o problema da orquestração é abordado com foco na quantidade e qualidade dos itens de telemetria coletados na rede e recebidos pelas aplicações de monitoramento. 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.

No presente trabalho são apresentadas duas variantes do problema de orquestração, que se originam dos modelos de orquestração focados na maximização do número de itens de telemetria propostos em [Hohemberger et al., 2019] e [N'Diaye, 2024]. São apresentados modelos matemáticos para as duas novas variantes, os quais destacam-se pelas funções objetivo que utilizam como critérios de otimização as informações de telemetria necessárias para o funcionamento das aplicações de monitoramento. Dessa forma, fica incorporada aos próprios modelos propostos a definição sobre o melhor conjunto de itens de telemetria que devem ser coletado: aquele conjunto que melhor atende à demanda de informação das aplicações. As diferenças entre as variantes serão apresentadas e discutidas no decorrer do trabalho.

O restante do trabalho está organizado da seguinte forma. Na Seção 2 é apresentada uma breve descrição do problema. A apresentação dos modelos matemáticos é feita na Seção 3. Os modelos foram utilizados para a resolução de instâncias do problema propostas na literatura, e os resultados computacionais são apresentados e discutidos na Seção 4. O artigo é encerrado com as considerações finais, na Seção 5.

2. Definição do problema

Considere um grafo $G = (D, L)$ no qual $D = \{1, 2, \dots, d_l\}$ é o conjunto de d_l vértices que representam os dispositivos de encaminhamento programáveis, e $L = \{l_{ij} \mid i, j \in D\}$ é o conjunto de arestas que representam as conexões existentes entre pares de dispositivos na infraestrutura da rede. Um conjunto $M = \{1, 2, \dots, m_l\}$ estabelece as m_l aplicações de monitoramento em execução. A localização física dessas aplicações é irrelevante para o problema, podendo ser considerada a possibilidade de um servidor à parte. Um conjunto $V = \{1, 2, \dots, v_l\}$ define os itens de telemetria que podem ser coletados na rede. Cada dispositivo $d \in D$ é capaz de incorporar um subconjunto $V_d \subseteq V$ desses itens a um fluxo, e cada item de telemetria $v \in V$ tem associado a si um parâmetro b_v que define seu tamanho em bytes.

Caminhos pré-definidos no grafo G estabelecem fluxos ativos na rede, cada um dos caminhos iniciado em um vértice de origem $d_o \in D$, a partir do qual itens de telemetria podem ser incorporados aos pacotes, e terminado em outro vértice de $d_e \in D$, no qual os itens coletados ao longo dos fluxos que ali convergem são agrupados e enviados para as respectivas aplicações de M . É em pacotes transferidos através desses fluxos que cada um dos dados de telemetria pode ser incorporado por um dispositivo $d \in D$. O conjunto de f_l fluxos ativos é dado por $F = \{1, 2, \dots, f_l\}$.

Os fluxos não são necessariamente disjuntos, de forma que mais de um fluxo podem passar por um dispositivo ou uma aresta/conexão. Associado a cada fluxo $f \in F$ há um parâmetro C_f que determina o tamanho do espaço disponível no fluxo para a incorporação de itens de telemetria. É esse parâmetro que estabelece um limite para os itens que podem ser transmitidos às aplicações, pois o somatório dos tamanhos dos itens de telemetria incorporados em um fluxo não pode exceder o espaço disponível.

O monitoramento realizado por uma das aplicações depende de um conjunto particular de itens de telemetria. Alguns itens, porém, apresentam dependência espacial, o que significa que devem ser coletados no mesmo dispositivo $d \in D$ para que sejam relevantes para o monitoramento da rede. Para formalizar o conceito de dependência espacial, é preciso considerar o conjunto $\mathcal{P}(V)$, ou seja, o conjunto de partições, em que cada partição representa um subconjunto de itens de telemetria de V . O conjunto de itens de telemetria que devem ser recebidos pela aplicação $m \in M$, já com as restrições de dependência espacial definidas, é o conjunto dado por $R_m \subseteq \mathcal{P}(V)$.

O conjunto R_m é, portanto, um subconjunto do conjunto de partições de V . Cada uma destas partições será referida individualmente no restante do texto por um índice $p \in P_m$, sendo P_m o conjunto dos índices dos componentes do conjunto R_m .

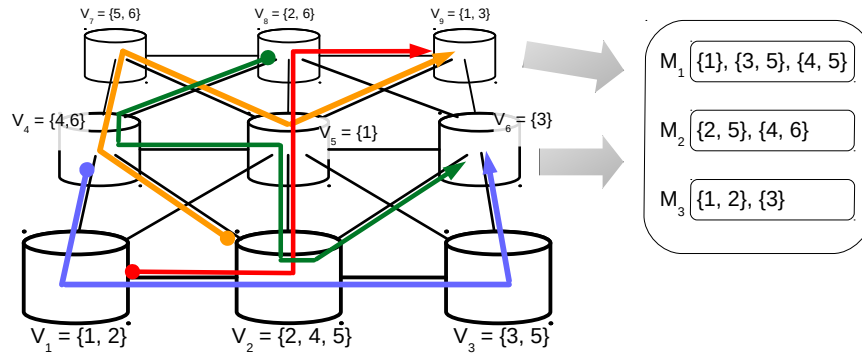


Figura 1: À esquerda, representação da estrutura física da rede, com quatro fluxos ativos identificados por cores distintas. Junto aos dispositivos estão os conjuntos de itens de telemetria que podem ser coletados por um fluxo. Três aplicações são representadas à direita, juntamente com os respectivos conjuntos R_m . Os itens recebidos pelos dispositivos ao final dos fluxos devem ser repassados para essas aplicações.

A Figura 1 ilustra a estrutura descrita até aqui. No lado esquerdo estão representados os dispositivos e suas conexões, que formam o grafo $G = (D, L)$. A cada um dos nove dispositivos representados está associado o respectivo conjunto V_d com os itens de telemetria que podem ser coletados naquele ponto do fluxo. A união desses conjuntos indica que o conjunto de todos os itens de telemetria que podem ser coletados na rede do exemplo é dado por $V = \{1, 2, 3, 4, 5, 6\}$. Os fluxos ativos são representados por quatro caminhos coloridos, um par terminando no dispositivo 6 e outro no dispositivo 9.

Os itens coletados ao longo dos fluxos são enviados por esses dois dispositivos para as aplicações, representadas mais a direita. O conjunto de partições de V necessária para o funcio-

namento de cada uma das aplicações está listado junto ao identificador da aplicação. O requisito para o funcionamento da aplicação M_1 , por exemplo, é dado por $R_{M_1} = \{\{1\}, \{3, 5\}, \{4, 5\}\}$. Essa configuração indica que a aplicação M_1 precisa receber o item de telemetria 1 sem qualquer dependência espacial, os itens 3 e 5 desde que tenham sido coletados no mesmo dispositivo, e os itens 4 e 5 também oriundos de um mesmo dispositivo. Além disso, como são três partições de V no conjunto R_{M_1} , o conjunto de índices é dado por $P_{M_1} = \{1, 2, 3\}$, em que a partição $\{4, 5\}$, por exemplo, é identificada pelo índice $p = 3$.

Faz-se referência, ao longo do presente texto, ao fato de que aplicações necessitam de itens de telemetria para seu funcionamento. A partir da definição do conjunto de partições R_m , é possível perceber que trata-se de uma simplificação, haja vista que a aplicação depende, na verdade, de elementos de $\mathcal{P}(V)$ (ou partições do conjunto V). A simplificação é adotada por fins de clareza no texto, mas registre-se que os itens sempre são organizados conforme as partições para garantir a satisfação da dependência espacial. A referência, portanto, ao recebimento de itens não diz respeito a itens individuais, mas sim a itens que combinados completam uma das partições necessárias.

É importante salientar que a dependência espacial não define um dispositivo específico para coleta dos itens de telemetria. Além disso, não há restrição sobre o fluxo pelo qual os itens devem atingir o dispositivo final, ou seja, itens com dependência espacial devem ser coletados em um mesmo dispositivo, mas podem ser enviados por diferentes fluxos.

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 objetivos relacionados ao número de aplicações atendidas. Na primeira variante proposta, especificamente, o objetivo é maximizar o número de aplicações atendidas. Vale lembrar que uma aplicação é considerada atendida quando todos os itens que ela necessita são recebidos, respeitadas as restrições espaciais. De modo mais formal, o objetivo é maximizar o número de conjuntos R_m recebidos, considerando todas as aplicações $m \in M$.

Perceba que maximizar as aplicações atendidas não é o mesmo que maximizar o número de partições do conjunto V (com restrição de dependência espacial) demandadas pelas aplicações. Uma orquestração que maximiza o número de tais partições recebidas, mas não exerce controle sobre a diversidade dessas partições para atender os requisitos definidos por R_m , pode resultar na inutilidade de grande parte dos itens de telemetria coletados e transmitidos.

Por outro lado, e por motivo semelhante, o objetivo da primeira variante proposta não garante o maior número de aplicações distintas sendo atendidas, possibilitando que uma aplicação tenha seu conjunto de partições atendido várias vezes, em detrimento ao atendimento de outra. A segunda proposta visa tratar exatamente essa situação. A coleta de itens de telemetria é planejada de forma a maximizar o número de atendimentos da aplicação menos atendida, denotado por C_{min} . O principal objetivo é evitar que uma ou mais aplicações sejam negligenciadas. Ao maximizar C_{min} , o número de aplicações não atendidas fica reduzido, pois o número de atendimentos da aplicação menos atendida tende a ser não nulo. Além disso, assume-se que a redundância é interessante para a melhoria no monitoramento por qualquer aplicação. Por isso, havendo espaço disponível para coleta de itens além dos necessários para maximizar as aplicações atendidas, este espaço será ocupado por itens redundantes. Observe que este modelo favorece, de certa forma, uma distribuição balanceada entre as aplicações atendidas em função da necessidade de elevar os atendimentos da aplicação menos atendida.

Por outro lado, é importante salientar que informações incompletas (nem todos elementos de R_m sendo recebidos) impedem o correto monitoramento por parte de uma aplicação. Se os requisitos de uma aplicação não puderem ser satisfeitos na totalidade, é irrelevante que ela receba qualquer item de telemetria, sendo interessante que o espaço que seria ocupado seja destinado para

a redundância nas informações de outras aplicações ou não seja utilizado para não sobrecarregar desnecessariamente o fluxo da rede.

3. Modelos Matemáticos

Neste seção são apresentados dois modelos para o Problema de Orquestração de Telemetria *In-Band*. O primeiro modelo é denominado INT-L, e considera a maximização do número de aplicações atendidas. O segundo modelo é denominado INT-C, e considerada, como objetivo principal, a maximização de C_{min} , ou seja, a maximização do número de vezes que a aplicação menos atendida tem seus requisitos satisfeitos. Como objetivo secundário, tem-se a maximização do número de aplicações atendidas.

O modelo INT-L é formalizado a seguir.

Conjuntos

D : conjunto de vértices do grafo, que representam os dispositivos a partir dos quais os itens de telemetria são enviados pelos fluxos;

F : conjunto de fluxos ativos na rede;

M : conjunto de aplicações de monitoramento;

V_d : conjunto de itens de telemetria que podem ser coletados no dispositivo $d \in D$;

P_m : conjunto de índices das partições que fazem parte do conjunto R_m , ou seja, conjunto de índices das partições de V que fazem parte dos requisitos para o funcionamento da aplicação $m \in M$;

D_{mp} : conjunto que contém os elementos $d \in D$ capazes de fornecer os itens de telemetria da partição p dentre as partições de R_m ;

Parâmetros

b_v : tamanho em bytes do item de telemetria $v \in V$;

C_f : capacidade de um fluxo $f \in F$, ou seja, a quantidade de bytes que podem ser enviados pelo fluxo f ;

Variáveis

y_{dvf} : variável binária que determina se o item de telemetria $v \in V$ é coletado no dispositivos $d \in D$ e enviado pelo fluxo $f \in F$;

s_{mpd} : variável binária que determina se a partição de índice p dentre as partições de R_m é atendida pelo dispositivo d ;

$\hat{s}_m$: variável inteira que determina quantas vezes o conjunto R_m de partições de itens de telemetria da aplicação m é satisfeito;

Modelo INT-L

$$\text{Max} \quad \sum_{m \in M} \hat{s}_m \quad (1)$$

s.a:

$$\sum_{d \in D} \sum_{v \in V_d} b_v y_{dvf} \leq C_f \quad \forall f \in F \quad (2)$$

$$\sum_{f \in F} y_{dvf} \leq 1 \quad \forall d \in D, \forall v \in V_d \quad (3)$$

$$s_{mpd} \leq \sum_{f \in F} y_{dvf} \quad \forall m \in M, \forall p \in P_m, \forall d \in D_{mp}, \forall v \in V_d \quad (4)$$

$$\hat{s}_m \leq \sum_{d \in D_{mp}} s_{mpd} \quad \forall m \in M, \forall p \in P_m \quad (5)$$

$$y_{dvf} \in \{0, 1\} \quad \forall d \in D, \forall v \in V_d, \forall f \in F \quad (6)$$

$$s_{mpd} \in \{0, 1\} \quad \forall m \in M, \forall p \in P_m, \forall d \in D_{mp} \quad (7)$$

$$\hat{s}_m \in \mathbb{N} \quad \forall m \in M \quad (8)$$

A função objetivo (1) maximiza o número de vezes que as aplicações têm os respectivos requisitos atendidos. A restrição (2) determina que o somatório dos tamanhos dos itens de telemetria enviados em cada fluxo $f \in F$ não pode exceder a capacidade do respectivo fluxo. A restrição (3) define que um item de telemetria $v \in V$ só pode ser enviado por um único fluxo $f \in F$ a partir do dispositivo $d \in D$.

Conforme referido anteriormente, os itens de telemetria necessários para o funcionamento de uma aplicação são organizados de acordo com a restrição de dependência espacial, de forma que os conjuntos R_m de requisitos de cada aplicação são conjuntos de partições do conjunto V . A restrição seguinte (4) determina se uma das partições do conjunto R_m de requisitos da aplicação $m \in M$, partição essa referida pelo índice p , é atendida por itens coletados no dispositivo d . Determinados os requisitos recebidos, é possível verificar quais aplicações tem os respectivos conjuntos R_m atendidos na totalidade. Isso é determinado na restrição (5), que estabelece o número de vezes que os requisitos da aplicação $m \in M$ foram atendidos. As restrições (6 a 8) determinam o domínio das variáveis.

Vale lembrar que a dependência espacial não define um dispositivo específico para o envio dos itens de telemetria, determinando apenas que os itens que fazem parte de um determinado componente de R_m sejam coletados no mesmo dispositivo. Perceba também que a modelagem é concebida pela consideração de um ciclo de monitoramento, não sendo possível que um único dispositivo colabore com a contagem da restrição (5) mais de uma vez. Mas havendo espaço para redundância nos fluxos, diferentes dispositivos podem enviar o mesmo conjunto de itens.

O segundo modelo tem como objetivo a maximização de C_{min} . Ou seja, o objetivo é modelado considerando um critério *maxmin*, no qual se procura maximizar a aplicação com o menor número de requisitos atendidos. Com isso, é esperado que nenhuma aplicação seja negligenciada. Além disso, havendo espaço para redundância, busca-se como um objetivo secundário maximizar o número de vezes que as aplicações têm os requisitos satisfeitos.

O modelo INT-C é composto pelas mesmas restrições do modelo anterior, e acrescenta duas novas. Por simplificação, a apresentação a seguir desse modelo não repete as restrições idên-

ticas, resumindo-se a apresentar a nova função objetivo e as restrições adicionais. São utilizados ainda um novo parâmetro e uma nova variável.

Variável

C_{min} : variável inteira não negativa que indica a quantidade de conjuntos de partição de itens R_m recebidos pela aplicação $m \in M$ com menor atendimento;

Parâmetros

ϵ : número suficientemente pequeno;

Modelo INT-C

$$\text{Max } C_{min} + \epsilon \sum_{m \in M} \hat{s}_m \quad (9)$$

s.a:

$$(2) - (8)$$

$$C_{min} \leq \hat{s}_m \quad \forall m \in M \quad (10)$$

$$C_{min} \geq 0 \quad (11)$$

A função objetivo (9) maximiza o número de conjuntos de partições de telemetrias coletados da aplicação menos atendida (C_{min}) e o número de vezes que as aplicações têm os requisitos atendidos. O parâmetro ϵ é utilizado para definir uma hierarquia entre os objetivos. Escolhendo um número ϵ suficientemente pequeno, o primeiro objetivo (maximizar C_{min}) torna-se prioritário, enquanto o segundo objetivo se torna secundário. A restrição (10) determina o valor de C_{min} e a restrição (11) determina o domínio desta variável.

4. Resultados

Todos os modelos utilizados nos testes computacionais foram implementados na linguagem Python 3.8 e resolvidos com o solver IBM ILOG CPLEX 20.1.0, com configuração padrão e limite de 8 *threads*, em uma máquina Intel Quad-Core Xeon E5-2697 v2 2.70GHz, com 64GB de memória RAM. O tempo limite para cada execução foi limitado a 600 segundos.

O conjunto de instâncias utilizado foi proposto por [N'Diaye, 2024]. As instâncias foram geradas aleatoriamente para redes com 50 e 100 nós. O tamanho dos itens de telemetria foi definido considerando um intervalo de valores discretos de $[10, 20]$, e a capacidade dos fluxos foi definida considerando um intervalo discreto de $[10, 40]$, ambos com distribuição uniforme. Para a rede com 50 nós, foram geradas instâncias com 50, 100 e 150 fluxos; 8 e 16 itens de telemetria; e 2, 4, 6 e 8 aplicações. Para a rede com 100 nós, foram geradas instâncias com 100 e 200 fluxos, todas com 16 itens de telemetria e 2, 4, 6 e 8 aplicações. O conjunto de testes totaliza 800 instâncias.

Na comparação entre os modelos, o Coeficiente de Gini (12) é utilizado para avaliar a equidade no atendimento dos requisitos das aplicações. Embora esta métrica não seja diretamente tratada nos modelos, o modelo INT-C, ao maximizar C_{min} , promove uma melhor equidade no atendimento das aplicações. O Coeficiente de Gini é largamente utilizado em economia para mensurar a concentração de renda. O índice varia no intervalo $[0, 1]$ e é de fácil interpretação: valor 0 representa a situação de equidade estrita, enquanto valor 1 corresponde ao extremo oposto, isto é, os recursos são monopolizados por apenas uma entidade. Na prática o limitante superior do coeficiente depende do número de dados considerados e é, desta forma, menor do que 1 para um conjunto

finito de dados. No caso presente, se o Coeficiente de Gini for igual a zero, significa que todas as aplicações foram atendidas igualmente, e valores maiores indicam que uma ou mais aplicações foram atendidas em detrimento de outras.

$$G = \frac{\sum_{m \in M} \sum_{m' \in M} |\hat{s}_m - \hat{s}_{m'}|}{2|M| \cdot \sum_{m \in M} \hat{s}_m} \quad (12)$$

As Tabelas 1 e 2 resumizam os resultados obtidos para instâncias que representam redes com 50 e 100 nós/dispositivos, respectivamente. A estruturação das tabelas é semelhante. Os resultados estão separados de acordo com o número de aplicações (|M|), número de fluxos (|F|) e quantidade de itens de telemetria (|V|), e a configuração das instâncias equivalentes a cada linha pode ser verificada nas três primeiras colunas de cada tabela. Cada resultado apresentado nas linhas corresponde à média aritmética dos valores obtidos a partir da execução dos modelos em 30 instâncias para as redes com 50 nós e em 10 instâncias para as redes com 100 nós.

O modelo referido nas tabelas 1 e 2 como Modelo INT-T é o modelo da literatura proposto em [N'Diaye, 2024] baseado em [Hohemberger et al., 2019], que objetiva recolher o maior número de partições de V que fazem parte dos requisitos de aplicações. Os dois modelos seguintes são os modelos originais do presente trabalho. Para cada modelo e formato de instância são exibidas colunas indicando os valores médios obtidos para o tempo de execução, a quantidade de atendimentos à aplicações ($\sum \hat{s}_m$), quantidade de atendimentos à aplicação menos atendida (C_{min}), quantidade de partições de V que fazem parte dos requisitos de aplicações recebidas ($\sum s_{mpd}$) e Coeficiente de Gini para ilustrar a equidade na distribuição dos atendimentos a aplicações.

Os primeiros resultados a serem destacados são os que dão conta da correteza dos modelos. Tanto na Tabela 1 como na Tabela 2, o Modelo INT-T, da literatura, sempre supera os demais na quantidade de partições recebidas, enquanto os modelos INT-L e INT-C nunca tem seus valores superados no que se refere à quantidade de aplicações atendidas e quantidade de atendimentos à aplicação menos atendida, respectivamente. Entre os dois últimos, percebe-se algumas vezes a igualdade de valores, indicando que o número máximo de atendimentos possível também é o número de atendimento ótimo quando se objetiva evitar que uma solução seja negligenciada.

Um resultado que chama a atenção é relativo à quantidade de partições do conjunto V (relevantes para alguma aplicação) recebidas. Percebe-se claramente que a satisfação dos conjuntos de requisitos das aplicações não está relacionado diretamente à maximização dos itens de telemetria coletadas ao longo da rede. Isso ocorre porque as partições, isoladas, não tem utilidade, já que uma aplicação só é atendida se receber a totalidade das partições de seus requisitos.

O Modelo INT-I, conforme esperado e observado na coluna $\sum s_{mpd}$, maximiza o recebimento desses pacotes. Mas se essa coluna é observada associada à coluna que apresenta o número de aplicações atendidas, percebe-se que maior número de atendimentos pode ser obtido mesmo com um conjunto substancialmente menor de partições recebidas. No caso de maior diferença no número de pacotes recebidos, na linha $|M| = 8$, $|F| = 50$, $|V| = 16$ da Tabela 1, tanto o Modelo INT-L quanto o Modelo INT-C precisaram de menos do que 9,5% das partições recebidas no Modelo INT-T, atendendo em média 8,33% mais aplicações. Este é um resultado muito importante, pois indica uma redução na coleta de itens de telemetria ao longo da rede, o que está de acordo com um objetivo muito buscado em outros trabalhos da literatura que visam reduzir a sobrecarga causada pela telemetria *in-band*.

O Coeficiente de Gini é exibido nas tabelas para ilustrar o fato de que o Modelo INT-C tende a gerar soluções com melhor distribuição dos atendimentos às aplicações. Em nenhum dos conjuntos de soluções exibidos o valor do Coeficiente de Gini do Modelo INT-C foi superior aos valores obtidos com os demais modelos. Isso evidencia que o critério de otimização deste modelo traz consigo uma característica de equidade que também é desejável.

Nas tabelas, como cada linha representa a média de execuções para instâncias do mesmo tipo, não ocorrem situações em que o Coeficiente de Gini resultou menor para o Modelo INT-L do que para o Modelo INT-C. Mas em situações específicas, para algumas instâncias individuais, essa situação ocorreu, indicando maior equidade na solução do primeiro modelo em relação ao segundo. Cabe reafirmar que a equidade é um efeito (desejável) que ocorre na maioria dos casos, mas não está explicitamente representada na função objetivo. No Modelo INT-C, uma vez que esteja maximizado o número de atendimentos da aplicação menos atendida, a distribuição dos atendimentos redundantes pode resultar desigual. Por isso, embora o objetivo de evitar que aplicações sejam negligenciadas leve a soluções em geral melhor distribuídas, eventualmente isso pode não ocorrer.

Tabela 1: Rede com 50 nós

Instance			Modelo INT-T					Modelo INT-L					Modelo INT-C				
IMI	IFI	IVI	Tempo	$\sum \hat{s}_m$	C_{min}	$\sum s_{mpd}$	Gini	Tempo	$\sum \hat{s}_m$	C_{min}	$\sum s_{mpd}$	Gini	Tempo	$\sum \hat{s}_m$	C_{min}	$\sum s_{mpd}$	Gini
2	50	8	0,5	16,8	6,5	126,1	0,132	5,4	19,6	8,4	56,2	0,083	3,6	19,5	8,8	56,2	0,060
		16	1,7	7,0	2,2	114,2	0,203	1,6	10,5	4,2	45,9	0,107	4,5	10,4	4,7	46,7	0,059
	100	8	4,8	28,9	12,5	210,7	0,069	9,5	29,8	12,9	90,0	0,068	12,7	29,8	12,9	90,0	0,068
		16	8,4	16,0	6,3	225,3	0,120	7,0	19,3	8,2	93,5	0,078	40,5	19,1	8,5	94,0	0,064
	150	8	0,2	31,1	14,0	220,8	0,052	0,1	31,1	14,0	92,1	0,052	0,1	31,1	14,0	92,1	0,052
		16	46,1	24,2	10,0	312,8	0,093	23,2	25,9	11,0	127,8	0,079	23,8	25,9	11,0	128,1	0,076
	Média		10,3	20,7	8,6	201,7	0,112	7,8	22,7	9,8	84,2	0,078	14,2	22,6	10,0	84,5	0,063
4	50	8	1,4	29,5	5,8	413,9	0,095	0,8	31,3	6,3	92,0	0,080	0,9	31,1	6,7	92,0	0,068
		16	7,0	12,0	1,9	310,9	0,210	6,5	14,3	2,4	58,2	0,205	9,8	14,2	2,8	58,6	0,163
	100	8	4,6	52,3	10,1	719,5	0,094	2,5	54,0	10,4	155,9	0,093	2,1	53,9	10,5	155,9	0,089
		16	26,3	29,1	6,0	707,3	0,089	19,7	32,8	6,8	152,2	0,081	26,9	32,7	7,2	151,8	0,070
	150	8	1,6	65,4	12,9	913,3	0,093	0,3	65,6	12,9	192,2	0,094	0,3	65,6	12,9	192,2	0,094
		16	132,9	41,4	7,6	1022,3	0,119	76,7	44,9	8,7	211,6	0,109	68,5	44,8	9,0	212,4	0,101
	Média		28,9	38,3	7,4	681,2	0,117	17,7	40,5	7,9	143,7	0,110	18,1	40,4	8,2	143,8	0,097
6	50	16	8,3	14,7	1,5	692,8	0,209	8,0	17,0	1,9	82,6	0,164	12,1	16,9	2,0	82,4	0,159
		8	2,1	43,1	5,5	877,4	0,097	1,5	45,5	5,9	130,3	0,088	1,7	45,2	6,3	129,9	0,078
	100	16	81,9	41,0	5,0	1447,9	0,116	73,9	44,8	5,8	205,4	0,099	125,1	44,6	6,1	205,5	0,089
		8	18,6	75,4	9,7	1645,3	0,094	16,9	78,2	10,0	239,4	0,094	24,4	78,1	10,2	239,0	0,091
	150	16	194,9	58,1	7,3	2257,6	0,100	101,8	62,9	7,9	309,9	0,099	160,0	62,6	8,2	308,4	0,091
		8	20,6	96,7	11,7	2051,8	0,102	5,7	97,5	11,7	294,0	0,101	3,6	97,5	11,7	293,8	0,100
	Média		50,8	55,8	6,9	1494,2	0,119	32,4	58,6	7,3	210,4	0,107	50,8	58,4	7,5	210,0	0,101
8	50	16	13,9	21,6	1,8	1194,1	0,159	10,7	23,4	1,9	112,5	0,157	21,0	23,4	2,1	112,5	0,149
		8	3,1	54,9	5,2	1525,6	0,092	2,5	57,7	5,6	171,4	0,087	2,3	57,4	5,9	170,5	0,076
	100	16	111,5	51,1	4,7	2523,4	0,102	57,7	54,5	5,1	264,4	0,095	37,4	54,4	5,4	263,4	0,089
		8	4,9	97,8	9,2	2748,5	0,107	4,2	101,6	9,3	303,3	0,103	24,2	101,3	9,6	302,2	0,098
	150	16	252,6	75,1	6,7	3864,9	0,109	141,9	80,9	7,4	401,3	0,108	164,3	80,3	7,8	400,0	0,097
		8	14,8	126,1	12,0	3635,4	0,092	21,6	127,5	12,1	390,6	0,090	22,1	127,5	12,1	390,7	0,090
	Média		61,7	72,5	6,7	2580,7	0,110	37,2	75,6	7,0	274,2	0,106	42,7	75,4	7,3	273,5	0,100

Tabela 2: Rede com 100 nós

Instance			Modelo INT-T					Modelo INT-L					Modelo INT-C				
IMI	IFI	IVI	Tempo	$\sum \hat{s}_m$	C_{min}	$\sum s_{mpd}$	Gini	Tempo	$\sum \hat{s}_m$	C_{min}	$\sum s_{mpd}$	Gini	Tempo	$\sum \hat{s}_m$	C_{min}	$\sum s_{mpd}$	Gini
2	100	16	14,9	15,3	4,9	235,8	0,184	17,5	24,4	9,2	104,9	0,106	46,9	23,8	11,4	103,8	0,021
	200	16	114,2	31,1	10,9	512,4	0,165	33,7	44,8	18,4	226,6	0,092	192,1	44,6	19,1	228,6	0,076
	Média		61,7	72,5	6,7	2580,7	0,110	37,2	75,6	7,0	274,2	0,106	42,7	75,4	7,3	273,5	0,100
4	100	16	47,0	24,7	4,5	747,2	0,147	38,0	30,5	5,3	152,6	0,135	41,7	30,4	5,9	153,9	0,114
	200	16	241,0	59,5	11,0	1515,6	0,102	275,6	69,6	15,2	341,2	0,064	232,2	68,7	15,9	339,6	0,043
	Média		144,0	42,1	7,8	1131,4	0,124	156,8	50,1	10,3	246,9	0,099	136,9	49,6	10,9	246,8	0,079
6	100	16	72,3	31,0	3,2	1363,8	0,153	61,0	38,1	4,1	183,2	0,141	63,0	37,5	4,7	182,1	0,118
	200	16	355,0	87,6	11,5	3129,6	0,073	405,9	95,1	13,0	462,7	0,073	405,6	94,1	13,7	460,8	0,058
	Média		213,6	59,3	7,4	2246,7	0,113	233,4	66,6	8,6	323,0	0,107	234,3	65,8	9,2	321,5	0,088
8	100	16	131,9	47,8	4,0	2523,2	0,121	127,8	53,3	4,5	258,1	0,123	167,8	52,3	5,2	255,4	0,098
	200	16	432,1	108,6	10,0	5384,0	0,088	448,6	115,1	11,2	588,8	0,077	488,1	113,9	12,3	583,9	0,063
	Média		282,0	78,2	7,0	3953,6	0,104	288,2	84,2	7,9	423,5	0,100	327,9	83,1	8,8	419,7	0,081

5. Considerações finais

Neste trabalho, foram apresentados modelos para duas variantes do Problema da Orquestração de Telemetria *In-Band*. O foco destas variantes está na satisfação dos requisitos das aplicações de monitoramento, buscando na primeira versão maximizar o número de atendimentos às aplicações, e acrescentando um componente de equidade dos atendimentos na segunda versão. Os modelos propostos permitem a solução na otimalidade de instâncias do problema com 50 e 100 nós e 50, 100 e 150 fluxos. As soluções possibilitam observar características interessantes que resultam da nova interpretação para problemas de orquestração e das novas formulações.

Em primeiro lugar, os resultados permitem afirmar que os modelos satisfazem os objetivos desejados, principalmente quando comparados com resultados obtidos por modelo que busca maximizar as partições de telemetria recebidas. Fica evidente, pelos resultados, que a satisfação de aplicações não está relacionada diretamente ao recebimento do máximo de itens de telemetria, mas sim do seu direcionamento conforme a necessidade das aplicações. Mesmo recebendo substancialmente mais partições de itens de telemetria, o Modelo INT-T, focado em maximizar este recebimento, atende menos aplicações em comparação com os outros dois modelos.

O Modelo INT-C, focado em maximizar o número de atendimentos à aplicação menos atendida, também cumpriu seu papel. Os resultados mostram que esse modelo, em muitos casos, resulta no mesmo número de atendimentos a aplicações que no Modelo INT-L. Mesmo quando constrói soluções em que menos conjuntos completos de partições de telemetria são coletados na rede, o Modelo INT-C evita que uma aplicação seja negligenciada enquanto outras são atendidas numerosas vezes. A avaliação através do Coeficiente de Gini evidencia também outra característica interessante das soluções do Modelo INT-C: a melhora no balanceamento da distribuição.

Por questão de justiça, é importante destacar que o primeiro modelo focado na maximização do recebimento de partições do conjunto de itens de telemetria não foi proposto originalmente na literatura como uma solução isolada. Ele é parte de uma solução composta ainda por uma etapa baseada em aprendizagem de máquina, responsável por melhor definir quais itens são importantes para coleta e envio pela rede. Mas cabe salientar que as alternativas propostas no presente trabalho incorporam essa definição ao modelo, dispensando o auxílio de outra ferramenta. Ao elaborar funções objetivo focadas na satisfação dos requisitos das aplicações, a importância aos itens de telemetria resta definida de forma subjacente, pela sua própria existência dentre as informações necessárias em pelo menos uma aplicação.

Por fim, vale observar que um dos objetivos mais comuns em trabalhos de orquestração de telemetria *in-band* é a diminuição da sobrecarga na rede. Torna-se evidente, pelos resultados, que focar na maximização do atendimento das aplicações reduz drasticamente o número de informações de telemetria necessárias, o que significa uma redução, na mesma proporção, do número de informações coletadas e transmitidas. Portanto, os novos modelos colaboram, indireta porém substancialmente, para a redução da sobrecarga na rede causada pela telemetria *in-band*.

Como trabalhos futuros, existe a possibilidade de formular o problema de modo a calcular os fluxos com um limite para o número de dispositivos que integram o caminho e apenas o ponto de coleta de telemetrias determinado.

Referências

- Bhamare, D., Kessler, 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.
- 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.

- Chiras, A., Peratikou, A., e Stavrou, S. (2024). Increasing security of containerized blockchain using sdn. In *2024 Panhellenic Conference on Electronics Telecommunications (PACET)*, p. 1–5.
- Dallanora, L. M. (2022). Orquestração dinâmica da telemetria de rede in-band em planos de dados programáveis. Master's thesis, Universidade Federal do Pampa.
- ETSI (2015). ETSI gs nfv 002: Architectural Framework. <http://www.etsi.org/technologies-clusters/technologies/nfv>. Acessado em 08/06/2024.
- Farris, I., Taleb, T., Khettab, Y., e Song, J. (2019). A survey on emerging sdn and nfv security mechanisms for iot systems. *IEEE Communications Surveys Tutorials*, 21(1):812–837.
- Haleplidis, E., Denazis, S., Koufopavlou, O., Lopez, D., Joachimpillai, D., Martin, J., Salim, J., e Pentikousis, K. (2014). ForCES Applicability to SDN-Enhanced NFV. In *Third European Workshop on Software Defined Networks (EWSDN)*.
- Han, B., Gopalakrishnan, V., Ji, L., e Lee, S. (2015). Network function virtualization: Challenges and opportunities for innovations. *IEEE Communications Magazine*, 53(2).
- 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.
- Kreutz, D., Ramos, F. M. V., Veríssimo, P. E., Rothenberg, C. E., Azodolmolky, S., e Uhlig, S. (2015). Software-defined networking: A comprehensive survey. *Proceedings of the IEEE*, 103(1):14–76.
- 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.
- 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. ISSN 1389-1286. URL <https://www.sciencedirect.com/science/article/pii/S1389128620313396>.
- Turchetti, R. C. e Duarte, E. P. (2017). NFV-FD: implementation of a failure detector using network virtualization technology. *International Journal of Network Management*, 27(6).
- Xilouris, G., Trouva, E., Lobillo, F., Soares, J., Carapinha, J., McGrath, M., Gardikis, G., Paglierani, P., Pallis, E., Zuccaro, L., Rebahi, Y., e Kourtis, A. (2014). T-NOVA: A marketplace for virtualized network functions. In *European Conference on Networks and Communications (EuCNC)*.