MM-INT: Telemetria em *Switches* Programáveis com Múltiplas Filas usando Roteamento Multicaminhos na Origem

Mateus N. Bragatto[1], João Paulo M. Clevelares[2], Cristina K. Dominicini[3], Rodolfo S. Villaça[2], Fábio L. Verdi[1]

[1]Departamento de Computação (Dcomp) – Universidade Federal de São Carlos (UFSCar)
Sorocaba/SP, Brasil.
mateusbragatto@estudante.ufscar.br,verdi@ufscar.br

[2]Departamento de Informática (DI) – Universidade Federal do Espírito Santo (UFES)
Vitória/ES, Brasil.
joao.clevelares@edu.ufes.br, rodolfo.villaca@ufes.br

[3]Instituto Federal do Espírito Santo (IFES) – Campus Serra
Serra/ES, Brasil.
cristina.dominicini@ifes.edu.br

Abstract. *This article emphasizes the importance of queues associated with the ports of switches in network monitoring. Traditionally, data collection about these queues is done using programmable data planes and telemetry based on INT (In-band Network Telemetry) probes, assuming there is only a single queue per output port. The MM-INT (Multiqueue Multicast - INT) is a solution that utilizes registers to store data from all queues and ports, enabling the efficient collection of monitoring information. The MM-INT avoids probe overload and employs the origin-based routing mechanism and multicast trees for the probes. The results demonstrate significant reductions in the number of probes sent compared to other traditional solutions found in the literature.*

Resumo. *Este artigo destaca a importância das filas associadas às portas dos switches no monitoramento de uma rede. Tradicionalmente, a coleta de dados sobre essas filas é feito com o uso de planos de dados programáveis, e telemetria baseada em sondas INT (In-band Network Telemetry), assumindo que há apenas uma única fila por porta de saída. O MM-INT (Multiqueue Multicast - INT) é uma solução que utiliza registradores para armazenar dados de todas as filas e de todas as portas, permitindo a coleta eficiente de informações de monitoramento. O MM-INT evita a sobrecarga de sondas e emprega o mecanismo de roteamento na origem e árvores multicaminhos para as sondas. Os resultados demonstram reduções significativas na quantidade de sondas enviadas, em comparação com outras soluções tradicionais da literatura.*

1. Introdução

Nos *switches* de uma rede, as filas são associadas às portas do dispositivo e representam uma importante fonte de informação de monitoramento, pois podem indicar o nível de congestionamento dessa rede [Kim et al. 2018]. Além disso, seu nível de ocupação pode afetar outras métricas, tais como tempo de resposta, atraso de transmissão e *jitter*. No entanto, coletar dados sobre as filas, em tempo real, sempre foi uma tarefa de alto

custo devido ao *overhead* gerado na rede e nos dispositivos [Arslan and McKeown 2019]. Entretanto, esse cenário tem mudado nos últimos anos graças à popularização dos planos de dados programáveis, com a linguagem P4, e ao monitoramento com telemetria *inband* baseado em sondas do tipo INT (In-band Network Telemetry).

Neste sentido, instruções de telemetria INT direcionam a coleta de métricas finas nos equipamentos de rede, em baixa granularidade e alta frequência, permitindo que os operadores de rede obtenham uma fotografia mais nítida do estado da rede a partir dos seus dispositivos. Porém, os trabalhos atualmente existentes na literatura concentram-se em coletar as métricas de desempenho das interfaces de rede, assumindo que há apenas uma única fila por porta de saída. Entretanto, em equipamentos programáveis mais recentes, há no mínimo 8 filas associadas à cada porta, sendo que em alguns equipamentos mais modernos este número já alcança até 128 filas. Esta quantidade de filas é extremamente útil, pois diferentes classes de fluxos de rede podem ser utilizadas, cada fila com uma qualidade de serviço diferente.

O maior desafio quando múltiplas filas estão presentes está relacionado ao fato de que, tradicionalmente, uma única sonda é capaz de coletar a telemetria apenas da fila por onde ela passa, associada à porta de saída de encaminhamento da sonda. Sendo assim, se uma porta possui duas filas, para monitorar todas essas filas, duas sondas deveriam ser enviadas. Para monitorar todas as filas, de todas as portas de cada *switch* da rede, seria necessário um acréscimo de sondas proibitivo, aumentando a sobrecarga de monitoramento.

Neste sentido, este artigo apresenta o *MM-INT* (Multiqueue Multicast - INT), uma solução capaz de coletar informações de todas as filas de cada porta física de um *switch* programável por meio da linguagem P4. Em resumo, a solução desenvolvida neste trabalho utiliza registradores para armazenamento dos metadados de cada fila. Estes registradores são alimentados sempre que um pacote de dados passa por uma fila no equipamento. Uma sonda INT, na solução *MM-INT*, é capaz de realizar a coleta dos metadados INT de todas as filas e de todas as portas por meio da leitura desses registradores. A solução proposta no *MM-INT* opõe-se ao mecanismo de monitoramento baseado na telemetria INT tradicional, pois, ao invés de obter os metadados INT diretamente da fila associada à porta de saída da sonda, coleta os dados armazenados nos registradores internos do dispositivo. Em resumo, tais registradores acumulam dados de monitoramento de todas as filas associadas a todas as portas do *switch* monitorado.

Além disso, a solução aqui apresentada faz uso do mecanismo de roteamento na origem e árvores multicaminhos (*multicast*) para diminuir a quantidade de sondas necessárias para cobrir todos os *switches* da rede, evitando sobreposições dessas sondas ao percorrer a topologia. Para isso, adotamos o mecanismo *MPolKA-INT* [de O. Pereira et al. 2023], que usa roteamento baseado na origem, substituindo as tabelas de roteamento tradicionais por um rótulo usado para o encaminhamento dos pacotes na rede. A solução foi implementada usando a linguagem P4 e avaliada em um ambiente Mininet com *switches* BMv2 (Behavioral Model version 2). A avaliação do *MM-INT* mostrou que a solução reduziu em 4x a quantidade de sondas enviadas e em $\approx$2.8x a quantidade de dados transferidos para a coleta da telemetria de rede quando comparada a soluções tradicionais encontradas na literatura.

2. Fundamentação Teórica e Posicionamento do Problema de Pesquisa

Esta seção está dividida em duas partes: i) descrição do funcionamento da coleta de métricas em redes programáveis usando INT; e ii) funcionamento do MPolka-INT.

2.1. Telemetria usando INT

Graças aos avanços recentes na tecnologia, à popularização dos dispositivos de redes programáveis e à linguagem P4, é possível que estes dispositivos informem o estado da rede sem intervenção do plano de controle [Arslan and McKeown 2019]. Nesse caso, os pacotes de dados podem conter cabeçalhos que são instruções de telemetria, os quais podem ser usados para coletar dados de desempenho. As instruções de telemetria estão descritas na especificação de plano de dados INT, a qual define três modos de operação: INT-XD (*eXport Data*), INT-MX (*eMbed Instructions*) e INT-MD (*eMbed Data*).

No modo INT-XD, cada dispositivo exporta os metadados usando instruções INT pré-configuradas nas tabelas de fluxo, atuando diretamente do plano de dados para o sistema de monitoramento. Neste modo, nenhuma modificação nos pacotes do tráfego dos clientes é realizada. No modo INT-MX, o nó de origem cria instruções INT no cabeçalho do pacote, de modo que em cada nó de trânsito as instruções INT são lidas, os respectivos metadados são coletados e transmitidos diretamente para o sistema de monitoramento. Neste modo, pequenas modificações são realizadas nos pacotes de dados, originado pelos clientes, visto que instruções de telemetria precisam ser inseridas no cabeçalho desses pacotes. No modo INT-MD, instruções INT e metadados são inseridos nos pacotes a cada salto na rede. Este é o modo de operação padrão definido pela especificação INT e o escolhido para este trabalho. Para não gerar dúvidas, o modo INT-MD será referenciado como INT ao longo deste artigo.

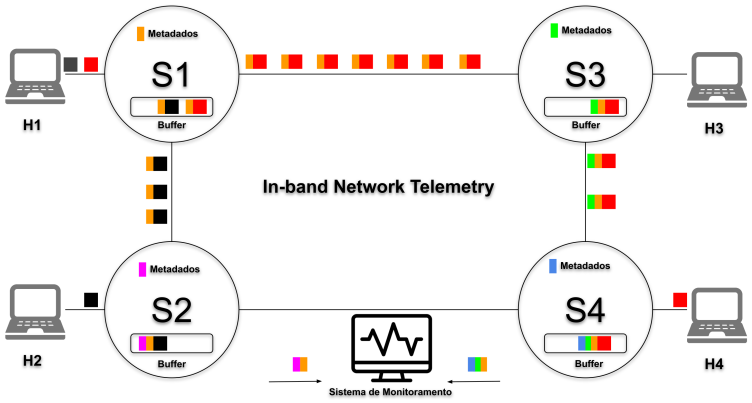

Figura 1. Exemplo de rede com In-band Network Telemetry.

A Fig. 1 exemplifica a operação do INT em uma rede. A rede é formada por quatro sistemas finais (*H1, H2, H3* e *H4*) e quatro *switches* programáveis com suporte à telemetria INT (*S1, S2, S3* e *S4*). Cada *switch* possui um conjunto de metadados, de interesse do sistema de monitoramento, representados por retângulos nas cores laranja (*S1*), magenta (*S2*), verde (*S3*) e azul (*S4*). Na rede existem dois fluxos de dados, um representado pelos pacotes na cor vermelha e o outro na cor preta. O fluxo da cor vermelha deve percorrer o seguinte caminho na rede $f1 = \{H1, S1, S3, S4, H4\}$. Já o fluxo preto deve percorrer o caminho $f2 = \{H1, S1, S2, H2\}$. A cada salto na rede, instruções de telemetria no plano de dados dos dispositivos orientam a coleta e adição dos metadados aos pacotes

que estão atravessando cada *switch*. Esse processo se repete ao longo de todo o caminho, desde o primeiro *switch* após a origem até o último *switch* antes do destino. No *switch* de destino, os metadados são então extraídos do pacote e encaminhados para um sistema de monitoramento da rede. O pacote original é então encaminhado ao seu destino.

Um dos principais atrativos do uso do INT está justamente no nível de granularidade alcançada, uma vez que os próprios pacotes que atravessam a rede carregam informações para o sistema de monitoramento. Além dos modos descritos na especificação INT, é possível utilizar outras abordagens para coleta dos metadados em redes programáveis. Uma delas é a possibilidade de se utilizar um fluxo de pacotes exclusivos para telemetria da rede. Estes pacotes exclusivos são denominados **sondas** e são responsáveis pela coleta dos metadados de telemetria, portanto, não alterando os pacotes de dados. A principal vantagem desta abordagem é evitar que os pacotes de dados sejam fragmentados, já que o acréscimo de metadados de telemetria nesses pacotes pode ultrapassar a MTU (*Maximum Transmission Unit*) da camada de enlace. A principal desvantagem desta abordagem é realizar a cobertura completa da topologia da rede para envio das sondas. Tipicamente, soluções tradicionais geram uma quantidade excessiva de sondas para cobrir toda a topologia da rede. Neste sentido, o uso do *MPolKA-INT*, conforme proposto neste artigo, reduz o número de sondas necessárias para essa tarefa por meio do uso de soluções de roteamento na origem e árvores de *multicast*, que será melhor detalhada na Seção 2.2.

Até aqui foi apresentado como a telemetria INT atua para coleta de metadados nos dispositivos de rede. Entretanto, nada foi dito sobre as múltiplas filas existentes nesses equipamentos. A Fig. 2, extraída de [Sharma et al. 2015], mostra em detalhes o funcionamento interno de uma fila de saída de um dispositivo de rede. Na figura, pode-se observar que os pacotes que entram no equipamento são direcionados para diferentes filas de saída antes de serem encaminhados para o enlace. Cada fila de saída possui sua própria taxa de ocupação, que se altera dinamicamente conforme mais ou menos pacotes chegam e saem da fila. Na figura, a retirada dos pacotes da fila ocorre por meio de um escalonamento baseado em pesos, sendo que uma fila com maior prioridade (mais peso) terá seus pacotes retirados (enviados para o enlace de saída) mais rapidamente.

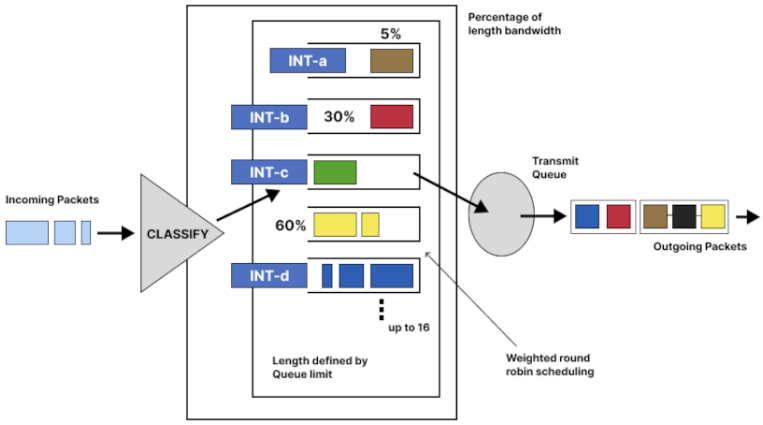

Figura 2. Múltiplas filas em uma porta de saída de um equipamento de rede. Adaptado de [Sharma et al. 2015].

Além disso, a mesma Fig. 2 ilustra essencialmente o comportamento padrão de um dispositivo de rede. Entretanto, considerando um dispositivo programável tradicional,

capaz de ser monitorado via telemetria INT, algumas observações devem ser feitas. Primeiramente, uma sonda INT somente é capaz de coletar métricas (metadados) da fila na qual ela está enfileirada. Ou seja, para cada fila da porta associada ao enlace de saída, uma sonda INT diferente deve ser gerada de forma que as métricas de cada fila ao longo de um caminho sejam coletadas. Pode-se observar, na figura, a existência de múltiplas sondas INT, representadas em azul, uma para cada fila (INT-a, INT-b, etc.). Cada sonda ilustra a coleta dos metadados individualmente em sua fila específica.

Entretanto, conforme já dito, esta solução gera a necessidade de uma grande quantidade de sondas para coleta de dados de monitoramento de todas as filas presentes na porta de saída de um dispositivo programável, inviabilizando a coleta de telemetria de toda a rede. O uso de registradores, disponibilizados pelos dispositivos programáveis, permite o armazenamento de dados de telemetria a serem coletados na passagem de um pacote e que eles sejam mantido para próximos pacotes. Desta forma, ao definir registradores específicos, é possível armazenar informações de cada uma das filas em que irão ser feitas leitura e escrita apenas no registrador o qual representa o conjunto de porta física e fila lógica a qual um determinado pacote tenha passado.

2.2. *MPolKA-INT*

MPolKA-INT [de O. Pereira et al. 2023] propõe um mecanismo para executar a telemetria da rede, explorando múltiplos caminhos, com diminuição da sobrecarga nos planos de controle e de dados, eliminando a redundância das informações de telemetria e substituindo as tabelas de roteamento das sondas pelo roteamento na origem. Para cobertura dos nós de interesse na rede, a proposta usa a solução de roteamento *Multipath Polynomial Key-based Architecture* (*M-PolKA*) [Guimarães et al. 2022], pois é um método de roteamento na origem, com suporte a multicaminhos, em que é possível codificar um rótulo de rota representando um ciclo ou árvore de maneira agnóstica à topologia [Guimarães et al. 2022].

O roteamento M-PolKA é baseado no Sistema de Números Residuais (RNS) e na aritmética polinomial usando campos de Galois de ordem 2 GF(2). Neste esquema, nos nós de núcleo, os estados de transmissão das portas de saída são dados pelo resto da divisão polinomial binária (operação de *mod* polinomial) do identificador de rota do pacote pelo identificador do nó do dispositivo de rede. Sua implementação em switches programáveis explora mecanismos de clonagem de pacotes e a reutilização do hardware de CRC (*Cyclic Redundancy Redundancy Check*), que permite a operação do *mod* polinomial.

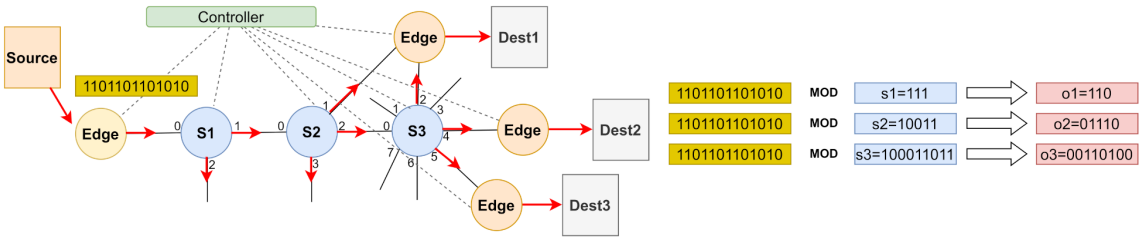

Figura 3. Exemplo de roteamento M-PolKA [Guimarães et al. 2022]

Conforme mostrado no exemplo da Fig. 3, a arquitetura *M-PolKA* é composta por: (i) nós de borda (em amarelo), (ii) nós de núcleo (s_1, s_2 e s_3, em azul) e (iii) controlador logicamente centralizado (em verde). O roteamento depende de três identificadores polinomiais: (i) *nodeID*: um identificador fixo atribuído aos nós de núcleo pelo controlador

em uma fase de configuração da rede; (ii) *t_state*: um identificador atribuído ao estado de transmissão das portas de saída em cada nó de núcleo; e (iii) *routeID*: um identificador de rota multicaminho, calculado pelo controlador usando o RNS e incorporado no pacote pelos nós de borda. Na Fig. 3 para um fluxo de exemplo, após calcular o *routeID* (10101100101100), o controlador instala regras na borda para incorporar este rótulo nos pacotes desse fluxo. Depois, cada nó de núcleo calcula seu *t_state* dividindo este *routeID* pelo seu *nodeID*. Em s_1, o restante do *routeID* (10101100101100) quando dividido pelo seu *nodeID* (111) é 110. Assim, em s_1, as portas 1 e 2 encaminham os pacotes para o próximo salto. Da mesma forma, em s_3 (*nodeID* 100011011), o resultado da operação de *mod* é 00110100 e apenas as portas 2, 4 e 5 encaminham os pacotes para o próximo salto.

Na solução de telemetria do MPolKA-INT, primeiramente, o controlador calcula o *routeID*, usando o roteamento *M-PolKA* descrito anteriormente, para representar a árvore *multicast* para a rota da sonda de monitoramento que irá cobrir todos os nós de interesse. Então, o nó gerador insere o *routeID* calculado, assim como as informações de telemetria do *switch*, no cabeçalho do pacote. Depois, a sonda percorre cada *switch* coletando as informações de telemetria. Em cada salto, o pacote é clonado para todas as portas de saída ativas no vetor de estados de transmissão. Para eliminação da redundância, os metadados de telemetria relativos aquele nó são inseridos apenas na primeira porta de saída.

3. Trabalhos Relacionados

Nesta seção estão relacionados alguns trabalhos relevantes da literatura que abordam a telemetria de redes *inband* usando o INT. Mais especificamente, visando destacar as principais contribuições deste artigo, os trabalhos relacionados estão divididos em 2 categorias: i) uso do INT com soluções de roteamento na origem; ii) uso do INT em dispositivos que possuem múltiplas filas ou em cenários que exigem o monitoramento na granularidade no nível de portas.

Pode-se afirmar que o INT-Path [Pan et al. 2019] foi o trabalho pioneiro que juntou o roteamento na origem e a telemetria INT para permitir a especificação de um caminho, representado como uma lista de nós incluídos no cabeçalho do pacote. O INT-Probe [Pan et al. 2021] resolve o mesmo problema, mas, adicionalmente, propõe um algoritmo de planejamento de caminhos de sondas para cobertura total da rede. O SR-INT [Zheng et al. 2021a] usa um cabeçalho de comprimento fixo e mitiga o problema da sobrecarga de sondas INT explorando o protocolo de roteamento por segmentos (do inglês *Segment Routing*). Apesar de haver uma boa quantidade de trabalhos relevantes que incorporam soluções de roteamento na origem com a telemetria *inband* usando INT, todos os trabalhos citados até o momento fazem uso de sondas *unicast* e focam em reduzir o custo da tarefa de monitoramento por meio da minimização da emissão de sondas INT.

Ainda com o mesmo objetivo, MPINT [Zheng et al. 2021b] inova ao aplicar a telemetria usando INT ao tráfego *multicast*, sem fazer uso de soluções de roteamento na origem. Por sua vez, o MPolKA-INT [de O. Pereira et al. 2023] é o trabalho pioneiro ao aplicar a telemetria *inband* ao tráfego *multicast*, com uso de soluções de roteamento na origem, cujos resultados mostraram a eficiência da solução para ampliar a cobertura da rede com o menor número possível de sondas.

Entretanto, nenhum destes trabalhos relacionados abordam o problema de monitoramento de múltiplas filas em uma mesma porta. Em [Harkous et al. 2021], é

Tabela 1. Sumário de informações sobre os trabalhos relacionados.

Ref.	INT	Roteamento na Origem	*Multipath*	Múltiplas Filas
[Pan et al. 2019]	✓	✓	X	X
[Pan et al. 2021]	✓	✓	X	X
[Zheng et al. 2021a]	✓	✓	X	X
[Zheng et al. 2021b]	✓	X	✓	X
[de O. Pereira et al. 2023]	✓	✓	✓	X
[Harkous et al. 2021]	X	X	X	✓
[Kundel et al. 2021]	X	X	X	✓
[Vogt et al. 2022]	✓	X	X	✓
MM-INT	✓	✓	✓	✓

apresentada uma solução para uso de filas virtuais no pipeline P4, explorando sua aplicação para o gerenciamento de tráfego em diferentes dispositivos programáveis. P4-CoDel [Kundel et al. 2021], por sua vez, destaca desafios e demonstra, usando P4, como implementar algoritmos de Gerenciamento Ativo de Filas (AQM) em hardware programável. Os autores também fazem referência a uma versão para o algoritmo fq_CoDel, que aborda o gerenciamento de múltiplas filas, porém não há maiores detalhes a esse respeito no artigo. Por fim, Vogt et al. [Vogt et al. 2022] apresentam o IPGNET, uma solução de monitoramento baseada no INT, onde os autores destacam a viabilidade de correlacionamento de dados de telemetria, mesmo em múltiplas filas. Entretanto, no IPGNET não se faz referência ao método de coleta de dados dessas múltiplas filas.

Em suma, conforme pode ser observado na Tabela 1, é fácil destacar uma lacuna na literatura com relação à proposição de soluções voltadas para: i) viabilização de coleta de dados de telemetria de múltiplas filas em dispositivos programáveis, e ii) uso de soluções de roteamento na origem, com sondas *multipath*, para minimizar a sobrecarga na cobertura dos nós em uma rede programável P4.

4. Coleta de Metadados INT em Múltiplas Filas

Na Seção 2.1, especificamente na Fig. 2, foi apresentada uma limitação do monitoramento baseado em telemetria INT, em sua forma tradicional, em relação à situações de múltiplas filas nas portas dos dispositivos de rede. A limitação consiste no fato de que uma sonda INT só consegue coletar os metadados de telemetria da fila associada à porta pela qual a sonda está passando. Portanto, a solução mais simples e ingênua seria enviar uma sonda para cada fila existente em todas as portas de cada *switch* da rede. Entretanto, esta solução não se apresenta como viável dada a quantidade de sondas que seriam necessárias para cobrir todas as filas de todas as portas e de todos os *switches* na rede monitorada.

Sendo assim, este artigo apresenta o *MM-INT*, uma solução que adota o uso de registradores para armazenar os metadados de telemetria de todas as filas existentes nos dispositivos programáveis por meio da linguagem P4, de forma que uma sonda INT coleta tais dados diretamente destes registradores. Como os registradores fazem parte da memória compartilhada do dispositivo, uma única sonda é capaz de ler estes registradores e coletar a telemetria de todas as filas e portas deste equipamento.

A Fig. 4 ilustra o mecanismo elaborado para suportar a solução *MM-INT*, proposta neste artigo. A figura apresenta um *switch* programável com dois *pipelines*: ingresso e egresso. Nossa solução é implantada apenas no pipeline de egresso, após o *Traffic Manager*,

onde se encontram as informações de telemetria de todas as portas e filas do equipamento. O exemplo apresentado possui duas filas na porta física, uma verde e outra vermelha. O mecanismo funciona da seguinte maneira: cada pacote de dados, antes de ser removido da fila para ser transmitido para o enlace de saída, invoca as chamadas para obtenção de telemetria INT

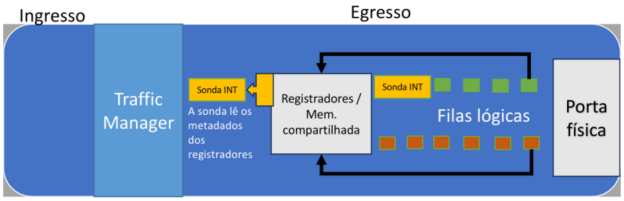

Figura 4. Estrutura interna em cada *switch* para coleta dos metadados de telemetria em múltiplas filas.

O dado da telemetria retornado por essas chamadas é armazenado em registradores, sendo que cada fila possui seu registrador próprio. A sonda INT, representada em amarelo na figura, será responsável por ler dos registradores os dados de telemetria armazenados, adicionando-os em seu cabeçalho. No *MM-INT*, a sonda é capaz de obter os dados de telemetria de todas as portas e filas de um *switch* programável. Após a coleta dos dados de telemetria, a sonda é encaminhada para o próximo *switch*, conforme o caminho definido na origem. Além disso, note que a sonda INT também será encaminhada por alguma fila, porém, na nossa solução, tal sonda não invocará as chamadas INT para obtenção da telemetria, já que tais dados foram obtidos a partir dos registradores. Na figura, a sonda INT foi encaminhada pela fila verde. Importante relembrar que, se os dados de telemetria não estivessem armazenados nos registradores, a sonda INT obteria tais dados quando encaminhada por alguma fila, porém, seria capaz de obter a telemetria apenas daquela fila (na figura, apenas da fila verde).

4.1. Detalhamento e discussão entre as soluções

Para facilitar a compreensão das vantagens associadas ao *MM-INT*, nesta seção são exemplificadas duas soluções tradicionais encontradas na literatura para coleta de informações de telemetria. Tais soluções serão comparadas com o *MM-INT* na Seção 5. Neste sentido, as soluções encontradas na literatura são aquelas que não usam (ou usam pouco) mecanismos para redução da quantidade de sondas INT enviadas na rede. A Fig. 5 será usada para apoiar essa discussão.

Solução $S1$: esta abordagem representa o mecanismo mais simples para a coleta de dados de telemetria na rede, que envia sondas para cobrir cada fila existente em cada porta dos *switches* da topologia. As sondas precisam ser geradas na origem (por algum *host* gerador de sondas pertencente ao sistema de monitoramento) e serem capazes de coletar dados de todas as filas de todas as portas do equipamento. É importante relembrar aqui que as sondas INT tradicionais coletam informações apenas da fila associada à porta de saída por onde passa no equipamento. Na Fig. 5, as sondas INT, representadas por setas coloridas, devem ser enviadas em ambas as direções na árvore de encaminhamento (ida e volta), partindo do nó raiz (*switch* A) para cada nó folha (*switches* D, E, F), e também a partir de cada nó folha para a raiz, para cobrir as portas de entrada nos *switches* no caminho de ida. Em uma árvore qualquer, o número de sondas INT pode ser obtido através da seguinte

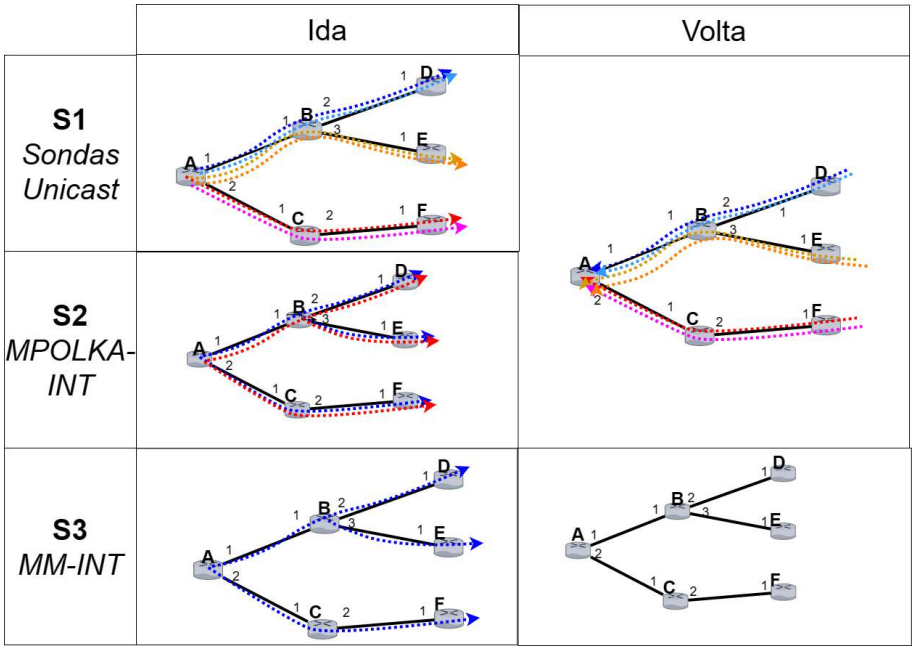

Figura 5. Comparação com as Soluções $S1$ e $S2$. Exemplo de 2 filas por porta.

equação: *nf x nq x 2*, sendo *nf* a quantidade de nós folhas na árvore de distribuição e *nq* a quantidade de filas em cada porta física. Neste exemplo, estamos considerando que a quantidade de filas por porta é o mesmo para todos os *switches* na rede. Na **Solução** $S1$ é importante observar a ocorrência da coleta em duplicidade, representada pelas quatro setas coloridas passando pelos caminhos A-B (ida) e B-A (volta). Estas setas representam as 4 sondas repetidas que passam por este enlace coletando as mesmas métricas.

Solução $S2$: esta solução introduz o MPolKa-INT, ou seja, apenas uma sonda é gerada e enviada em cada direção da árvore de multicast representada na Fig. 5. Seguindo a lógica do funcionamento do MPolka-INT, a sonda deve ser clonada em cada bifurcação para cobrir todos os ramos da árvore. Entretanto, já que há múltiplas filas por porta, a sonda deve ser clonada e recirculada para cada fila, para coletar individualmente a telemetria de cada fila associada à porta de saída da sonda. Para exemplificação, tome como exemplo o funcionamento do *switch* B da Fig. 5. A sonda INT chega no *switch* B pela porta 1. Esta sonda será encaminhada por alguma fila da porta 2. Entretanto, antes de ser encaminhada, a sonda será clonada três vezes, uma para cada fila restante (2 portas, 2 filas por porta). Cada sonda clonada e recirculada seguirá por sua respectiva fila, conforme mostra a figura, coletando então as métricas de telemetria de uma fila (por onde passou) associada à porta de saída de sonda. Porém, pode-se observar que a simples aplicação do MPolka-INT reduz a quantidade de sondas na ida, porém, na volta, quando o envio da sonda é feito a partir dos nós folha, ainda ocorrem algumas duplicidades. Primeiro, permanece a necessidade do envio de 1 sonda por fila. Além disso, no exemplo observa-se a duplicidade de sondas no enlace B-A, evidenciado pela quantidade de setas neste enlace no caminho de volta. Portanto, a Solução $S2$ reduz, mas não elimina, a duplicidade de sondas INT, quando comparada com a Solução $S1$.

Solução $S3$, *MM-INT*: A Fig. 5 também ilustra a Solução $S3$, que faz uso do *MM-INT*, onde apenas uma sonda é gerada na origem, que será clonada apenas quando necessário,

nas bifurcações da topologia (ramos da árvore de encaminhamento multicaminhos). Como pode-se observar na Fig. 5, as sondas são geradas apenas no sentido de ida, já que é possível coletar a telemetria de todas as filas e portas de um mesmo *switch* a partir da leitura dos registradores internos implementados na solução. Elimina-se, portanto, a necessidade de se enviar sondas no sentido contrário (volta), e não há duplicidades de sondas em nenhum enlace. Em resumo, com o *MM-INT* um sistema de monitoramento é capaz de realizar a coleta de dados de telemetria de todas as filas, associadas a todas as portas de um *switch* usando uma árvore de multicaminhos com roteamento na origem, reduzindo drasticamente a quantidade de sondas necessárias para cobrir toda uma topologia, eliminando completamente a duplicidades de sondas na rede.

4.2. Implementação do *MM-INT*

Nos exemplos que seguem, considere que por padrão um pacote será enviado sempre para a fila 0 da porta de saída do *switch* programável. Na etapa de *parsing*, é verificado por meio do campo *etherType* do cabeçalho se um pacote é uma sonda de telemetria (TYPE_SR) ou dados (TYPE_IPV4). Um pacote de dados é encaminhado usando tabelas de roteamento tradicionais baseadas no endereço IP de destino. Caso o campo TOS não esteja configurado com o valor 55 (configuração escolhida para um pacote de telemetria), é um pacote de dados e então é realizada a seleção da fila. Conforme a seleção, os dados de telemetria relativos a essa fila são gravados nos registradores correspondentes.

A sonda INT é encaminhada na rede usando o encaminhamento M-PolKA, descrito na Seção 2.2. Sendo assim, na etapa de *parsing* para uma sonda INT, é obtido o valor do *routeID* como identificador de rota multicaminho a ser percorrido, além de extraída a pilha de cabeçalhos INT. Após a execução do *pipeline*, o pacote é clonado para todas as portas de saída ativas no vetor de estados de transmissão (*bit* 1). Para eliminação da redundância das informações de telemetria, utilizamos o recurso do *resubmit* para alterar o cabeçalho de um pacote clonado e inserir os metadados de telemetria apenas na primeira porta de saída ativa.

5. Experimentação e Avaliação do *MM-INT*

As avaliações foram realizadas em ambiente virtualizado com Mininet. Uma das primeiras contribuições neste sentido foi adaptar o emulador Mininet para suportar múltiplas filas, alterando os arquivos necessários tanto da arquitetura (*v1model*) quanto do compilador (*p4c*). Muito embora a implementação do BMv2 suporta até oito filas lógicas por porta física, não havia uma versão do Mininet que integrasse o BMv2 com este suporte. Essa versão está publicamente disponível no GitHub[1] com uma máquina virtual já configurada.

A topologia utilizada para a avaliação funcional e comparativa do *MM-INT* é mostrada na Fig. 7. Importante destacar que, para fins de demonstração, esta topologia já está em formato de árvore[2]. Para este artigo, configuramos duas filas lógicas por porta (*fila 0 e fila 1*). A implementação foi feita usando a linguagem P4 e todos os artefatos para reprodução do experimento também estão disponíveis no GitHub[3].

[1]https://github.com/dcomp-leris/p4-multiqueue.

[2]O *MM-INT* funciona em qualquer topologia no formato de árvore ou em uma topologia genérica que seja convertida para uma árvore para fins de roteamento (e.g., uma spanning tree). Por questões de espaço, analisaremos apenas uma topologia neste artigo.

[3]https://github.com/dcomp-leris/MM-INT.

5.1. Avaliação Funcional da Implementação do *MM-INT*

O primeiro experimento mostra que a implementação do *MM-INT* é funcional. A maneira mais simples de fazer isso foi gerar tráfego intenso em todas as direções entre *hosts* ligados aos nós folhas da topologia (não mostrados na Fig. 7), de forma que todas filas fiquem ocupadas. Em paralelo, foram geradas sondas INT que seguirão pela árvore conforme multicaminhos definidos na origem pelo MPolka-INT. Uma aplicação desenvolvida em Python foi usada como coletor de sondas INT nos *switches* de borda (`SW1`, `SW3`, `SW4` e `SW7`). A Fig. 6 mostra a ocupação das filas,em número de pacotes, apenas do *switch* `SW1`. Para os outros *switches*, os gráficos são semelhantes e não estão apresentados por falta de espaço. Note que a figura mostra a ocupação das duas filas lógicas em cada porta física do *switch* `SW1`.

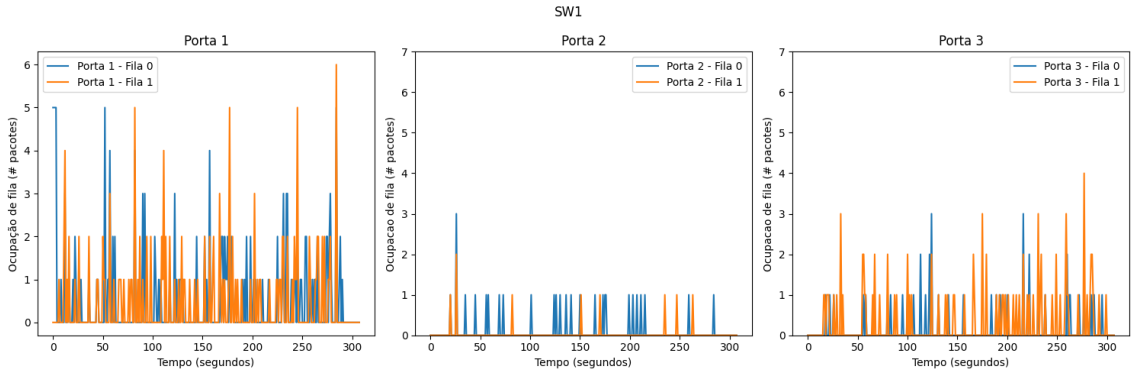

Figura 6. Coleta INT realizada no *switch* `SW1`.

5.2. Avaliação de Desempenho do *MM-INT*

Em seguida, após certificar-se de que a solução implementada está funcionando, vamos analisar a quantidade de sondas enviadas para as soluções $S1$, $S2$ e $S3$ (*MM-INT*). Por questões de espaço, a topologia apresentada na Fig. 7 ilustra o envio das sondas apenas para a Solução $S1$, na ida, no sentido do nó raiz (`SW1`) para os nós folha. Portanto, pode-se observar as flechas (sondas) saindo do switch `SW1` para os *switches* `SW1`, `SW3`, `SW4`.

Nesta topologia, considerando a Solução $S1$ serão necessárias 6 sondas para ida e 6 para a volta. Na ida, temos 2 sondas (1 para cada fila lógica) saindo de `SW1` para `SW3`, 2 sondas saindo para `SW4`, e 2 sondas saindo para `SW7`. Na volta, dos nós folhas para a raiz, o processo é repetido resultando em mais 6 sondas. As duplicidades ocorrem tanto na ida quanto na volta, sendo 2 sondas no enlace `SW1-SW2` e outras 2 no enlace `SW2-SW6`, totalizando 4 sondas duplicadas na ida, e outras 8 sondas duplicadas na volta, totalizando 12 sondas repetidas. É importante destacar que a quantidade de sondas duplicadas depende da topologia. Outras topologias em árvore serão analisadas em trabalhos futuros.

Considerando a Solução $S2$, a quantidade de sondas enviadas é o mesma da Solução $S1$, ou seja, 12 sondas. Uma das grandes vantagens da $S2$, que já usa o MPolka-INT, é que apenas uma sonda é gerada pelo nó gerador de sondas, e as clonagens ocorrem apenas nas bifurcações, como explicado na Seção 4.1. Assim, menos sondas passam pelos enlaces, já que não há duplicidade de sondas na ida, apenas na volta. A quantidade de sondas repetidas na volta é a mesma obtida na Solução $S1$ (8 sondas).

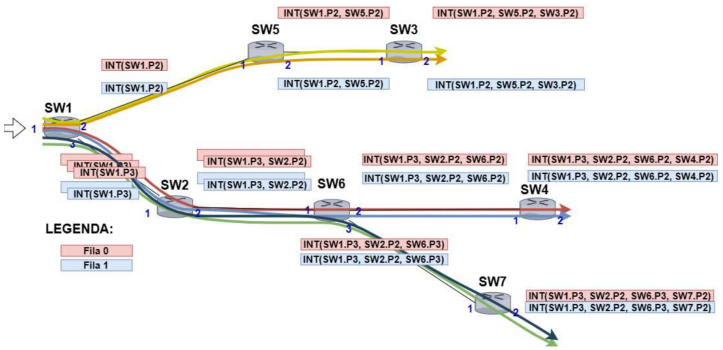

Figura 7. Envio de sondas INT usando a Solução $S1$ (INT original). A figura representa apenas a sonda partindo do `SW1` para os nós folhas (`SW3`, `SW4` e `SW7`).

Além da quantidade de sondas enviadas e as duplicidades, outras métricas devem ser analisadas para verificar as vantagens e desvantagens do *MM-INT*. Para isso, também foram analisadas as seguintes métricas: tamanho da sonda (em *Bytes*), ocupação de memória e quantidade de dados (em *Bytes*) necessários para as sondas INT.

Tabela 2. Resultados da avaliação para a topologia usada neste artigo.

Mecanismo	Tam. pacote (bytes)	# sondas	Mem. (em bytes)	# dup.	Bytes totais
S1	61	12	0	12	2300
S2	75	12	0	8	2174
MM-INT	122 (2 portas) 154 (3 portas)	3	64 (2 portas) 96 (3 portas)	0	814

A Tabela 2 apresenta os resultados obtidos. Como pode-se observar, o *MM-INT* ocupa uma pequena quantidade de memória do *switch*, uma vez que precisa armazenar as informações de telemetria das filas lógicas nos registradores (*statefull*). As Soluções $S1$ e $S2$ não precisam armazenar nenhum estado em registradores, portanto, não ocupam espaço em memória. Entretanto, a quantidade de memória utilizada pelo *MM-INT* é insignificante e, mesmo em um *switch* com 32 portas físicas e 128 portas lógicas por porta física (Tofino2), seriam necessários em torno de 65KB. Considerando que este switch possui 40MB de memória compartilhada, entendemos que a solução é escalável em termos de ocupação de memória. Por outro lado, o tamanho da sonda INT para o *MM-INT* neste caso pode chegar a aproximadamente 65KB, por embutir a informação coletada nesses registradores, o que ultrapassa o limite de MTU (Unidade Máxima de Transmissão) de 1500 *Bytes* em uma rede Ethernet, ocorrendo fragmentação.

O *MM-INT* se destaca quando analisamos a quantidade de sondas enviadas e duplicadas pois envia 4x menos sondas em relação às outras soluções avaliadas, sem geração de sondas duplicadas. Devido ao número menor de sondas, cada pacote no MM-INT é capaz de coletar mais informações de telemetria em um único envio. Portanto, ao analisarmos o tamanho do pacote, o *MM-INT* requer mais *Bytes* para carregar os metadados de telemetria, e esse tamanho depende da quantidade de portas e de filas por porta. A depender do tamanho do pacote, pode-se ultrapassar a MTU da camada de enlace, causando fragmentação. Nesses casos, uma estratégia para limitar o tamanho do pacote de coleta é seletivamente definir o que cada sonda pode coletar. Pode-se definir quais portas e metadados serão coletados por sonda, tais como apresentados em soluções já existentes na literatura [Papadopoulos et al. 2023, Ben Basat et al. 2020, Tang et al. 2020].

Mesmo com o *MM-INT* possuindo o pacote de telemetria maior, quando se analisa a quantidade total de *Bytes* usados para a coleta de dados desta topologia, observa-se que o *MM-INT* requer uma quantidade de *Bytes* muito inferior em relação às outras duas soluções analisadas neste artigo, na ordem de 2.8x menos *Bytes*. Assim, mesmo que o *MM-INT* requeira um cabeçalho maior para a telemetria, no agregado total, utiliza-se menos largura de banda da rede já que envia muito menos sondas (redução de 4x) e não gera duplicidade.

5.3. Limitações do Protótipo

A solução apresentada neste trabalho se mostrou viável e funcional quando comparada às soluções tradicionais, especialmente por se tratar de um protótipo em estágio inicial. Entretanto, há algumas limitações que precisam ser destacadas, que serão alvo de pesquisa e desenvolvimento em trabalhos futuros relativos ao *MM-INT*:

- A solução pressupõe que a quantidade de filas lógicas por porta é o mesmo em todos os equipamentos.
- O tamanho do pacote INT é maior, e pode não escalar caso haja muitas portas físicas e filas lógicas nos equipamentos.
- Há um certo atraso entre a informação de telemetria armazenada nos registradores e o momento em que a sonda passa pelo *switch*.
  A implementação pode sofrer alterações quando aplicada a um dispositivo físico como o *switch* Tofino. Especificamente, a questão das leituras e escritas realizadas nos registradores e o número de estágios de pipeline de execução exigirão adaptações para serem suportados na arquitetura *TNA (Table Type Architecture)*.
- A fragmentação pela utilização do *MM-INT* pode vir à ocorrer devido à limitação imposta pela MTU. Este ponto sugere a realização de uma melhor análise sobre qual seria o limite relacionado a quantidade de portas e filas por *switch* coletado, além da quantidade de *switches* em que uma única sonda poderia fazer a coleta.

6. Conclusões e Trabalhos Futuros

Neste artigo apresentamos uma solução para coleta de telemetria em switches programáveis, o *MM-INT*, que permite a coleta de dados de múltiplas filas, diminuindo a sobrecarga na rede e eliminando redundância de sondas. O *MM-INT* faz uso de uma infraestrutura de roteamento na origem e explora múltiplos caminhos para realizar a cobertura completa da rede monitorada. Como principais resultados, o *MM-INT* foi capaz de minimizar a quantidade de sondas necessárias para coleta dos dados de telemetria das múltiplas filas utilizando INT, além de diminuir o total ocupado em Bytes para cobrir a coleta de todos os switches na topologia avaliada. Além disso, a solução foi incorporada ao Mininet e disponibilizada publicamente.

Como trabalhos futuros, as principais demandas relacionam-se com a continuação da implementação do *MM-INT*, endereçando as limitações descritas na Seção 5.3. Além disso, pretende-se avaliar a viabilidade, escalabilidade de desempenho da proposta em *switches* programáveis reais, do tipo Tofino 2.

Agradecimentos

Os autores agradecem o apoio financeiro da FAPESP, processos $2022/13544-8$ e $2020/05182-3$, e da FAPES, processos $941/2022$, $2023/RWXSZ$ e $2022/ZQX6$.

Referências

Arslan, S. and McKeown, N. (2019). Switches know the exact amount of congestion. In *Proc. of the Workshop on Buffer Sizing*, BS '19, New York, NY, USA. ACM.

Ben Basat, R. et al. (2020). Pint: Probabilistic in-band network telemetry. In *Proceedings of the ACM SIGCOMM*, SIGCOMM '20, page 662–680, New York, NY, USA. ACM.

de O. Pereira, I. et al. (2023). MPolKA-INT: Stateless Multipath Source Routing for In-Band Network Telemetry. In Barolli, L., editor, *Advanced Information Networking and Applications*, pages 513–524, Cham. Springer International Publishing.

Guimarães, R. S. et al. (2022). M-PolKA: Multipath Polynomial Key-Based Source Routing for Reliable Communications. *IEEE Transactions on Network and Service Management*, 19(3):2639–2651.

Harkous, H. et al. (2021). Virtual queues for p4: A poor man's programmable traffic manager. *IEEE Transactions on Network and Service Management*, 18(3):2860–2872.

Kim, Y. et al. (2018). Buffer management of virtualized network slices for quality-of-service satisfaction. In *IEEE Conf on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, number 18725013 in 1, pages 1–4, Verona, Italy. IEEE.

Kundel, R. et al. (2021). P4-codel: Experiences on programmable data plane hardware. In *ICC 2021 - IEEE International Conference on Communications*, pages 1–6.

Pan, T. et al. (2019). Int-path: Towards optimal path planning for in-band network-wide telemetry. In *IEEE INFOCOM 2019*, pages 487–495. IEEE.

Pan, T. et al. (2021). Int-probe: Lightweight in-band network-wide telemetry with stationary probes. In *IEEE 41st ICDCS*, pages 898–909.

Papadopoulos, K., Papadimitriou, P., and Papagianni, C. (2023). Deterministic and probabilistic p4-enabled lightweight in-band network telemetry. *IEEE Transactions on Network and Service Management*, 20(4):4909–4922.

Sharma, R., Sehra, S., and Sehra, S. K. (2015). Review of different queuing disciplines in voip, video conferencing and file transfer. *IJARCCE*, pages 264–267.

Tang, S., Li, D., Niu, B., Peng, J., and Zhu, Z. (2020). Sel-int: A runtime-programmable selective in-band network telemetry system. *IEEE Transactions on Network and Service Management*, 17(2):708–721.

Vogt, F. G. et al. (2022). Innovative network monitoring techniques through in-band inter packet gap telemetry (ipgnet). In *Proc of the 5th Int Workshop on P4 in Europe*, EuroP4 '22, page 53–56, New York, NY, USA. ACM.

Zheng, Q., Tang, S., Chen, B., and Zhu, Z. (2021a). Highly-efficient and adaptive network monitoring: When int meets segment routing. *IEEE Transactions on Network and Service Management*, 18(3):2587–2597.

Zheng, Y., Pan, T., Zhang, Y., Song, E., Huang, T., and Liu, Y. (2021b). Multipath in-band network telemetry. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–2.