ICT-MobileSummit 2009 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2009
ISBN: 978-1-905824-12-0

Poster Paper

Forwarding Challenges and Solutions for a Publish/Subscribe Network

Dmitrij LAGUTIN[1], Sasu TARKOMA[1]
[1]*Helsinki Institute for Information Technology HIIT, Helsinki University of Technology*
TKK, PO Box 9800, Espoo, 02015 TKK, Finland
Tel: +358 9 451, Fax: +358 9 694 9768, Email: firstname.lastname@hiit.fi

Abstract: The shortcomings of the current Internet have motivated a number of publish/subscribe based networking protocols and solutions, which aim to challenge a host-centric nature of the current Internet. While most of these proposals are based on existing Internet Protocol, some are more radical aiming to build a publish/subscribe based network from scratch. In this paper, we investigate forwarding approaches for a clean state publish/subscribe network. A forwarding should be efficient, flexible, compliant with the current valley-free model of the Internet, and utilize the data-oriented nature of the network as much as possible. We divide our problem into inter-domain and intra-domain cases and present alternatives for both, including a novel up-graph based approach for inter-domain forwarding.

Keywords: Future Internet technologies, publish-subscribe, forwarding, inter-networking.

1. Introduction

For almost 30 years, the Internet has been coping with ever increasing traffic and new applications, including voice and video, while retaining its original architecture drafted almost 40 years ago. The current dominant inter-networking solution, the Internet Protocol suite, works reasonably well for most existing demands but suffers from a number of limitations. The most notable of the design aspects that have turned detrimental is the imbalance of powers in favour of the sender of information, who is overly trusted. The network accepts anything that the sender wants to send and will make a best effort to deliver it to the receiver. This has led to increasing problems with unsolicited traffic, e.g. e-mail SPAM, and distributed denial of service (DDoS) attacks, forcing companies and users to conceal their e-mail addresses and place their systems behind firewalls. Further challenges include those related to efficient support for mobility, efficient global multicast, and multi-homing. In addition, reconciliation of end-to-end reachability with other networking requirements, that arise from the scarcity of IP addresses and an untrustworthy environment, using firewalls, network address translation (NAT), and other middleboxing techniques is a much studied, albeit hard to solve problem.

In the Publish-Subscribe Internet Routing Paradigm (PSIRP) [1] project, which is an EU FP7-funded project with a 30 month lifetime, we see the main reason for the shortcomings of current IP-based inter-networking being deeper embedded in its underlying communication paradigm than in its operational shortcomings. The endpoint-centric communication paradigm that underpins the current Internet, and its predecessor, the telephony network, places rather arbitrary topological constraints on the delivery of information. With the observed increase of information-centric services, such as the World-Wide Web or newer contemporary applications such as sensor networks, the inflexibility of endpoint-oriented topologies increasingly places a burden on solution developers that needs circumvention by virtue of ever increasing number of overlays. This leads to a lack of flexibility and increasing rigidity.

In this paper we analyze the forwarding problem within the scope of publish/subscribe networks and propose a novel inter-domain forwarding solution based on up-graph information. This paper is organized as follows, Section 2 contains background discussion. Section 3 explores the construction of the forwarding identifiers and the forwarding process in the scope of PSIRP. Our approach is analyzed in Section 4 which also discusses future work. Finally, Section 5 contains conclusions.

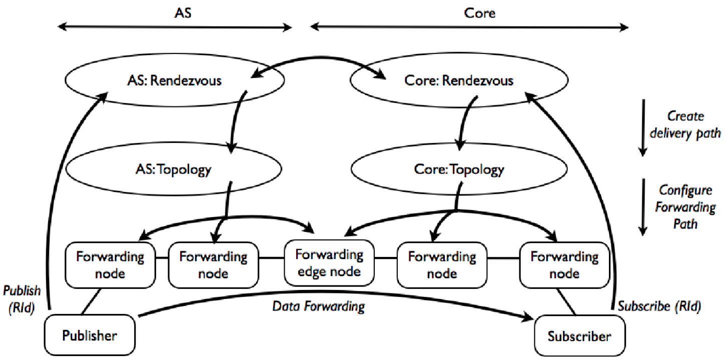
2. Background

Traditionally, the Internet architecture has been host and connection oriented system where users establish connections to specific hosts. However, the situation is changing and nowadays users are more interested in the actual data content than its location within the network. Similarly, the trust-to-trust principle [2] is also an emerging trend, which argues that users are interested to contact entities in which they trust, instead of just contacting arbitrary hosts.

Publish/subscribe networks aim to transform the current host-centric Internet architecture into an information-centric one. Users should be able to retrieve relevant data without having the information about its topological location within the network. If multiple users request the same piece of data, it should be delivered using an efficient multicast technique instead of multiple point-to-point connections. There have been numerous publish/subscribe [3] and data-oriented network approaches, including data-oriented network architecture (DONA) [4], routing on flat labels (ROFL) [5], Triad [6], and Internet indirection infrastructure (i3) [7]. Most of these approaches still utilize the IP protocol for the actual data traffic.

2.1 PSIRP

PSIRP aims to implement a publish/subscribe-based network from scratch without relying on existing technologies like IP. The architecture of PSIRP consists of four main parts: forwarding, rendezvous, caching and routing. The aim of the rendezvous process is to create a forwarding path between a publisher and subscribers, while the forwarding process is responsible for actual data delivery. The overall architecture of PSIRP is described in Figure 1.



Fi

gure 1: The architecture of PSIRP

PSIRP utilizes several kinds of identifiers on a different layers. A rendezvous identifier (Rid) is associated with policy-compliant data dissemination graphs for publication delivery, both in the local domain (intra-domain) and between domains (inter-domain). The rendezvous identifiers are chosen from within a large enough set to provide a probabilistic guarantee of uniqueness without a central allocation authority.

Applications may resolve application identifiers, which are contained within published data, into rendezvous identifiers. It is then the responsibility of the rendezvous functions, with the help of the topology information, to find suitable data transit and delivery paths in the network and denote them with forwarding identifiers (Fids). This resolution from a rendezvous identifier to a set of forwarding identifiers is based upon the rendezvous identifier in conjunction with scoping (as identified through the scope identifier, Sid) and policy mechanisms. The breadth of reference of Fids is variable, potentially limited to single hops or dynamically expandable to encompass full multicast trees.

In this paper we concentrate on the forwarding problem in PSIRP; how to effectively express and construct forwarding identifiers in a clean slate approach, and how to forward traffic from publishers to subscribers based on them.

2.2 Forwarding on the Internet

The current Internet hierarchy can be divided in to three tiers. Tier-1 is an IP network which connects to the entire Internet using settlement free peering. There are a small number of tier-1 networks that typically seek to protect their tier-1 status. A tier-2 network is a network that peers with some networks, but relies on tier-1 for some connectivity for which it pays settlements. A tier-3 network is a network that only purchases transit from other networks.

Autonomous systems (AS) on the Internet can be categorized as follows [8]: customer-to-provider (c2p), peer-to-peer (p2p), and sibling-to-sibling (s2s). In the first, a customer AS pays a provider AS for any traffic sent between the two. In the second p2p category, two domains can freely exchange traffic between themselves and their customers, but do not exchange traffic from or to their providers or other peers. In the third s2s category, two domains are part of the same organization and can freely exchange traffic between their providers, customers, peers, or other siblings. According to [9], every border gateway protocol (BGP) path must comply with the following hierarchical pattern: an uphill segment of zero or more c2p or s2s links, followed by zero or one p2p links, followed by a downhill segment of zero or more p2c or s2s links. Paths with this hierarchical structure are *valley-free* or valid. Paths that do not follow this hierarchical structure are called invalid and may result from BGP misconfigurations or from BGP policies that are more complex and do not distinctly fall into the c2p/p2p/s2s classification. While PSIRP aims to create new publish/subscribe-based network from scratch, the system should be deployable on the current Internet, and therefore should conform to the valley-free model. Effects of the valley-free model on data-oriented networks have been studied in [10], and they must be also taken into account when designing a forwarding solution for PSIRP.

NIRA (A New Inter-Domain Routing Architecture) [11] empowers users the ability to choose a provider and domain level end-to-end path. The motivation for this is that only users know whether a path is valid or not. Such a model creates competition between paths that different ISPs offer because users can choose the most suitable transit paths. NIRA emphasizes policy-based routing and utilizes the valley-free model.

3. Forwarding approaches for a clean slate publish-subscribe network

Our goal is to design a forwarding solution for the publish/subscribe network. The solution should work well with multicast traffic, be scalable, efficient, and deployable on the current Internet.

In the scope of PSIRP, forwarding on the router level is a process that accepts an incoming publication and sends it to zero or more output ports based on information contained in the packet and state maintained in the nodes. The state needed to perform this forwarding decision needs to be managed.

3.1 Constructing the forwarding identifier

The number of forwarding tree identifiers is crucial for scalability of the system. We divide this problem into two parts, namely intra-domain and inter-domain identifiers. These two are distinct cases and this division is useful in reducing the router state.

For the intra-domain case, publications need to be delivered between subscribers and publishers in the local area. Although the number of active data identifiers is huge, on the order of 10^{15}, only a small subset is active in a given domain. Table 1 illustrates the possible choices for the intra-domain forwarding identifiers (Fids), namely using the content identifiers (e.g., Sid+Rid), a Bloom filter [12] based identifier over the content identifiers, B(Sid + Rid) where B is a Bloom filter, a static Fid assigned by the rendezvous system for each publication type separately, and a Bloom filter over the link identifier. Bloom filter is a probabilistic data structure where a simple *AND* operation is used to test whether the element is present in a set, therefore it offers a high matching performance. While the Bloom filter does not produce false negatives, false positives are possible.

Table 1: Comparison of intra-domain forwarding methods

Intra-domain Fid type	**Aggregation**	**Mapping**	**Notes**
Content identifier (Sid+Rid)	No	One-to-one	Offers fine-grained forwarding, but requires excessive state in routers. In addition, the state must be set up for each scope separately.
B(content identifier)	Yes	Almost one-to-one	May result in false positives due to the probabilistic nature of the solution.
Static Fid	No	Many-to-one	Requires setup, a change to Fid requires interaction with the rendezvous system. Possible loss of precision, because scopes +Rid needs to be mapped to the Fid.
B(link identifier)	Yes	One-to-Many	A form of a probabilistic source routing.

The combination of Sid and Rid offers fine-grained forwarding, but requires excessive state in routers. In addition, the state must be set up for each scope separately. Aggregation is not possible and each combination can be seen as its own circuit.

The second alternative offers fine-grained forwarding which may result in false positives due to the probabilistic nature of the solution. Aggregation is possible in routers through Bloom filter union.

Using a static Fid assigned by the rendezvous system, the forwarding tree must be setup, and a change to Fid requires interaction with the rendezvous system. Each Fid can be seen as its own circuit. There is a possible loss of precision, because scopes+Rid needs to be mapped to the Fid.

Bloom filter over link identifiers approach is a form of probabilistic source routing. The rendezvous system issues the Bloom filter, which contains zero or more secure router identifiers. These identifiers correspond to output links through which the packet should pass. This approach has good scalability properties since minimal state is needed in routers; however, it assumes that the rendezvous system knows all the routers and can determine the proper Fid. It may be difficult to change (rewrite) the Fid due to security issues.

Based on this analysis, static Fids are not appealing because of their inflexible nature and the interaction needed with the rendezvous system. On the other hand, static Fids allow very fast forwarding decisions, because of their circuit-switched nature.

A Bloom filter based approach looks to be the most promising one, we can either use content or link identifiers for construction of the Bloom filter, or a combination of both. Basically, this is a question whether the content identifiers (Rid + Sid) are handled only in a rendezvous system or also in routers. Additional study is required to determine the optimal structure of the forwarding identifier.

Inter-domain identifiers require that the domain level structure is taken into account. Here we are concerned to which domains a publication must be sent. This needs to take into account the domain-level paths, and also the active subscriptions in the domains. In the global case, it is not possible to store all the identifiers needed or advertised by other domains. Therefore, a different strategy must be employed for inter-domain Fids. Table 2 presents possible candidate solutions.

Table 2: Comparison of inter-domain forwarding methods

Inter-domain Fid type	Aggregation	Mapping	Notes
Domain + content identifiers	No	One-to-one	Requires excessive state and updates.
Domain identifier + B(content identifier)	Yes	Almost one-to-one	Offers possibility to aggregate Rids.
Static Fid	No	Many-to-one	Requires setup, a change to Fid requires interaction with the rendezvous system. Possible loss of precision, because scopes+Rid needs to be mapped to the Fid.
B(domain identifier)	Yes	One-to-many	A form of a probabilistic source routing.
Identifier derived from up-graphs	Yes	Many-to-one	Rendezvous system is responsible for combining publisher's and subscriber's up-graphs.

The first approach does not lend itself well to inter-domain operation due to the excessive state and updates needed in routers. The second approach may be useful, because of the possibility to aggregate Rids.

With a static Fid assigned by the rendezvous system, the forwarding tree must be setup, and a change to Fid requires interaction with the rendezvous system. Each Fid can be seen as its own circuit. There is a possible loss of precision, because Sid+Rid needs to be mapped to the Fid.

It is also possible to use a Bloom filter over the domain identifier, which is a form of a probabilistic source routing similar to intra-domain case.

The final approach utilizes up-graph information in a similar manner as NIRA. The scale-free nature of the Internet makes the most of the data travel through a very small hub (tier-1). Because of this, in many cases all forwarding trees from one source have the same path towards the center of the network. P2P traffic has brought changes in supply and demand in the network. Increasing amount of data is sent from the very edge of the network. This leads us to our hypothesis that the up-graphs are similar for many edge nodes. Therefore, the up-graph based solution is the most promising one.

3.2 Setting up the forwarding state

There several alternatives where to store the forwarding state. It can be placed in packets, or in the forwarding tables of routers, or in both. In the following, we focus on strategies which require that routers maintain per-distribution tree state. This requires that the per-distribution tree state is built, which is similar to virtual circuits. Ultimately, there is a distribution tree rooted at each publisher; however, many parts of the trees are shared. Since there is no data traffic without publisher, it is reasonable to assume that publishers require activation of a delivery tree (or forest).

We summarize the expected behaviour of the network as follows using an abstract rendezvous service:

1. Publisher appears on network.
2. Publisher tells the rendezvous system its intent to publish.
3. Publisher sends its up-graph (up to tier-1) to the rendezvous system.
4. Rendezvous system knows how to connect publisher to other domains with a policy compliant shortest path.
5. Rendezvous system can run a process to determine a representative set of distribution trees between domains. This set is evaluated in terms of key metrics.
6. Rendezvous system establishes the distribution trees across domains.

No traffic is delivered unless there are subscribers to the content. If forwarding paths are set up and used before any subscribers appear, packets are delivered throughout the network in vain. Therefore the publisher is a subscriber for relevant forwarding identifiers that should be used from the source.

3.3 Forwarding process

The forwarding process within routers would work as follows. Each anycast packet maps to zero or one output ports. Each multicast packet maps to zero or more output ports. Since modular design is one of the basic requirements, we separate these two forwarding processes. Of the two, anycast is clearly simpler and requires the maintenance of one-to-one correspondence between incoming data labels and output ports. This can be achieved using hash tables or probabilistic structures such as Bloom filters.

The multicast forwarding problem is more complicated, because any given packet and its associated label may map to a number of outgoing ports. The two main rules are as follows:

- The input port is never used
- Map to zero or more output ports

Given that there are n ports in the system, a simple strategy that associates a probabilistic structure, such as a counting Bloom filter, with each port takes linear time for finding the proper ports. A port specific update is a constant time operation with this strategy.

4. Analysis and future work

We have divided the forwarding problem into two separate cases, forwarding within the domain and forwarding between domains. The first case is relatively simple to handle. The amount of forwarding states and ids is limited, and therefore we can use a Bloom filter based solution.

The second case requires a solution which is scalable on a global Internet level. Using a forwarding identifiers derived from up-graphs for inter-domain forwarding looks to be a promising solution. However, this presents some challenges. In order to be able to derive delivery trees for subscribers and publishers, the rendezvous system needs to know about the policy-compliant end-to-end paths. This information can be given to the rendezvous system by the subscribers and publishers by sending their up-graphs to the rendezvous entities in question. The up-graphs can then be combined to determine the subset of inter-domain topology that is relevant for the distribution of the information. The rendezvous system must take possible confidentiality issues into account when using this up-graph information. The advantage of this solution is that the rendezvous system does not need to know the global network topology beforehand, it gets the relevant information from publishers and subscribers during the rendezvous process. In addition, this gives end users the possibility to influence the path along which the data will be delivered, creating more competition between network providers.

4.1 Future work

Deployment of the system is an issue which must be carefully considered in the future. The current Internet infrastructure is optimized to IP traffic and our solution should be gradually deployable. In this respect, up-graph based solution is a promising one, since it is compatible with the current tier-based hierarchy of the Internet. We also aim to better take into account the data-oriented nature of traffic in the exact structure of forwarding identifiers. In addition, testing, evaluation and security are important issues.

A low-level Bloom filter based forwarding implementation has been studied in the PSIRP project [13, 14]. Such a solution utilizes link id tags (LIT) which are created by hashing the link identifiers. Therefore, a single link id can have several associated LITs and the Bloom filter used for forwarding is constructed from LITs. Using LITs allows optimization the system according to various parameters, like reducing the amount of false positives[1]. Using such a scheme, a NetFPGA [15] based implementation of the Bloom filter forwarding approach achieved a good performance with a latency overhead of only 1-4 µs compared to a loopback interface. The PSIRP prototype built on FreeBSD platform is briefly described in Figure 2. A detailed description of the prototype is out of scope of this paper and is explained in [14]. We aim to continue to work on the Bloom filter approach and refine it to be more suitable with our proposed forwarding solution.

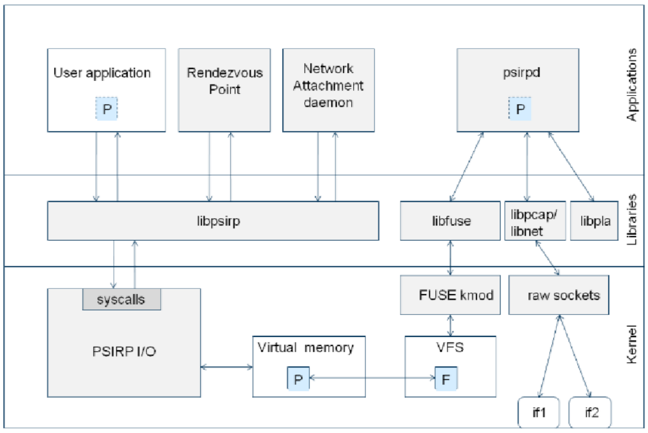


Figure 2: The architecture of the PSIRP prototype

4.2 The security of forwarding approaches

Security is an important issue when designing a new network architecture, mistakes made with the original Internet design where the security was mostly ignored should not be repeated [16].

Bloom filters over link identifiers have some interesting inherent security properties. Denial-of-service attacks are difficult to launch since the link identifiers are not globally known, and it is impossible to derive them from a complete Bloom filter. Therefore, the attacker is unable to construct a valid Bloom filter from themselves to the specific destination without knowing a global network topology and link identifiers.

However, “security thought obscurity” based solutions are usually not bullet-proof. We also consider additional security solutions, like the packet level authentication (PLA) [17], which aims to secure the network by using strong, per packet cryptographic signatures.

[1] Bloom filters always have a risk of false positives. In our case the false positive means that the packet is delivered to the undesired destination.

Such a method would offer a good protection even if the attacker possesses the information about the global topology and link identifiers.

5. Conclusions

This paper explores challenges and potential solutions for a publish/subscribe-oriented scalable forwarding. Such forwarding solutions have not been widely studied yet.

We consider intra- and inter-domain forwarding as two separate cases. Our inter-domain forwarding solution, where the rendezvous process creates a forwarding path between publish and subscriber based on their up-graph, looks to be promising. It offers a good flexibility and is compatible with the valley-free nature of the current Internet.

The PSIRP project is an ongoing work and in the future we will concentrate on deployment, testing and evaluation of our system. This is a challenging task, since the system should be a gradually deployable on the current Internet while being based on a completely different paradigm than current IP networks. There are also additional issues like the security which must be carefully taken into account.

References

[1] M. Ain, et al, Conceptual Architecture of PSIRP Including Subcomponent Descriptions, technical report [online], August 2008, available at: http://psirp.hiit.fi/files/Deliverables/FP7-INFSO-ICT-216173-PSIRP-D2.2_ConceptualArchitecture_v1.1.pdf [Accessed 10th January 2009].

[2] D. D. Clark and M. S. Blumenthal, End-to-end Arguments in Application Design: The Role of Trust, Proc. of TPRC, 2007.

[3] P. Eugster, P. Felber, R.Guerraoui, and A-M. Kermarrec. The many faces of publish/subscribe. ACM Computer survey, Volume 35, Issue 2, pp. 114-131, 2003.

[4] T. Koponen et al., A Data-oriented (and beyond) network architecture, Proc. of ACM SIGCOMM 2007, pp. 181-192, Kyoto, Japan, August 2007.

[5] M. Caesar, T. Condie, J. Kannan, K. Lakshminarayanan, and I. Stoica, ROFL: Routing on Flat Labels, Proc. of ACM SIGCOMM 2006, pp. 363-374, September 2006.

[6] D. R. Cheriton and M. Gritter, TRIAD: A New Next-Generation Internet Architecture [online], July 2000, available at: http://www-dsg.stanford.edu/triad/ [Accessed 2th January 2009].

[7] I. Stoica, D. Adkins, S. Zhuang, S. Shenker, and S. Surana, Internet Indirection Infrastructure, Proc. of ACM SIGCOMM, August, 2002.

[8] X. Dimitropoulos et al., AS Relationships: Inference and Validation, ACM SIGCOMM Computer Communication Review, Volume 37, Issue 1, pp. 29-40, 2007.

[9] L. Gao, On Inferring Autonomous System Relationships in the Internet, In IEEE/ACM Trans. Networking, December 2001.

[10] J. Rajahalme, M. Särelä, P. Nikander, and S. Tarkoma, Incentive-Compatible Caching and Peering in Data-Oriented Networks, Re-Arch'08, 2008.

[11] X. Yang, D. Clark, A. Berger, NIRA: a New Inter-Domain Routing Architecture, IEEE/ACM Transactions on Networking, Volume 15, Issue 4, pp. 775-788, August 2007.

[12] B. H. Bloom, Space/time trade-offs in hash coding with allowable errors, In ACM Communications, Volume 13, Issue 7, pp. 422-426, 1970.

[13] C. Esteve, F. Verdi, M. Magalhaes, Towards a new generation of information-oriented Internetworking architectures, Re-Arch'08, 2008.

[14] P. Jokela e al., LIPSIN: Line Speed Publish/Subscribe Inter-Networking, Technical report TR09-0001 [online], January 2009, available at: http://www.psirp.org/files/Deliverables/PSIRP-TR09-0001_LIPSIN.pdf [Accessed 13th January 2009]

[15] J. W. Lockwood et al., NetFPGA-An Open Platform for Gigabit-rate Network Switching and Routing, In MSE'07: Proceedings of the 2007 IEEE International Conference on Microelectronic Systems Education, pp. 160-161, 2007.

[16] D. Clark, The design philosophy of the DARPA internet protocols, ACM SIGCOMM Computer Communication Review, Symposium proceedings on Communications architectures and protocols SIGCOMM '88, vol. 18, issue 4. 1988.

[17] D. Lagutin, Redesigning Internet - The packet level authentication architecture, licentiate's thesis, Helsinki University of Technology, Finland, June 2008.