POROVNÁNÍ NÁVRHŮ NA ZMĚNU INTERNETOVÉ ARCHITEKTURY

COMPARISON OF PROPOSALS SUGGESTING INTERNET ARCHITECTURE CHANGE

Vladimír Veselý, Miroslav Švéda

Abstrakt

Čím dál tím více organizací se uchyluje k multihoming připojení, využívá tvarování provozu, požaduje snadnou mobilitu zařízení anebo se trápí přečíslováním adres při přechodu k jinému poskytovateli. Od roku 2006 je čím dál tím více patrnější krize směrovacího systému páteře Internetu, na který všechny předešlé vlastnosti kladou čím dál tím větší nároky. TCP/IP, na kterém je Internet postaven již skoro třicet let, pod světlem nových požadavků ukazuje, že není řešením na věčné časy. Tato práce si klade za cíl provést porovnání vlastností existujících relevantních návrhů na změnu architektury Internetu. Zaměřuje se na hledání podobností, diskutuje rozdíly, poskytuje bázi znalostí pro udělání si představy, jaké návrhy mají největší potenciál.

Klíčová slova: *směrovací systém, Core-Edge Separation, Core-Edge Elimination*

Abstract

More organizations tend to deploy multihoming, traffic engineering, mobility, deal with address renumbering or hopes for more independence when changing providers. It is more and more apparent that default free zone routing system is heading towards crisis influenced by all previously mentioned network capabilities. Internet is using TCP/IP for nearly thirty years. However, current TCP/IP stack is not flexible enough to accommodate new needs. This paper aims to provide confrontation between relevant proposals suggesting Internet architecture change. Thesis focuses on finding similarities, discussing invariances and providing knowledge base for drawing decision, which suggestions have the best potential.

Key words: *routing system, Core-Edge Separation, Core-Edge Elimination*

1 INTRODUCTION

Nowadays Internet routing and addressing architecture is facing variety of challenges that were not so apparent in early days of the TCP/IP stack. Among those challenges, there are multihoming, mobility, traffic engineering, renumbering, device localization and identification. All of them stress routing scalability of **default-free zone (DFZ)** and lead to growth of the global routing tables.

The main goal of this paper is to provide overview on existing proposals that have capability to upgrade current Internet architecture.

Paper is divided as follows. In this section is provided motivation, brief description of problems and properties of ideal solution. Section 2 provides basic theory behind decoupling identification and localization. Relevant proposals are their properties are mentioned in Section 3. Comparison of properties is provided in Section 4. Section 5 draws conclusion.

1.1 Motivation and Problems

The growing amount of transferred data comes hand to hand with increasing number of users. Paths between nodes in the Internet are becoming shorter, faster, more redundant and more reliable. More existing IPv4 addresses are used as **Provider Independent (PI)** rather than **Provider Aggregatable (PA)** addresses of Internet Service Provider (ISP). Free IPv4 address space is depleted and IPv6 is still fighting to reach at least 2% of overall traffic despite the fact that it has been more than 16 years since its standardization.

The most severe and apparent issues are listed down below as subchapters. Some of the problems are based on review from RFC 6227 (1), RFC 4984 (2), some of them from mutual community observations of current trends.

Routing Scalability

The most affected nodes struggling with the situation are DFZ routers. Every year the **Forwarding Information Base (FIB)** size of of those routers increases. The rate, at which number of prefixes is growing in the FIB, is object of discussions, but it is definitely faster than linea. On the following graphs (Fig. 1, Fig. 12) from (3) we can see historical progress in the size of Border Gateway Protocol (BGP) for IPv4 and also IPv6 – on the x-axis is year, on the y-axis is the number of prefixes:

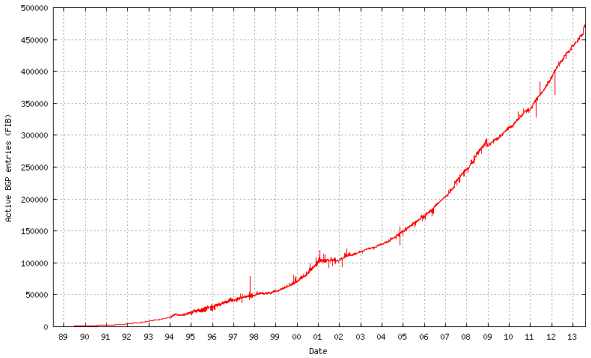

Fig. 1: IPv4 FIB size

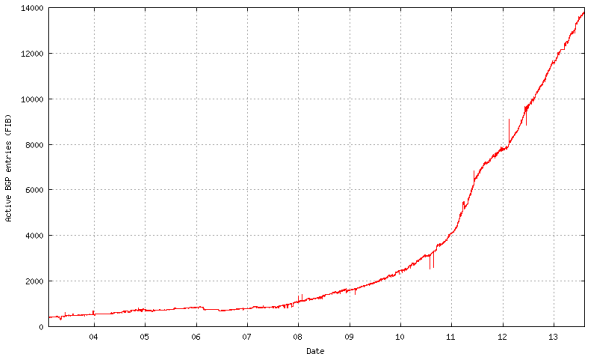

Fig. 2: IPv6 FIB size

Each prefix must be processed which adds to the **control plane** load, consumes more of router's CPU performance and memory and last but not least increases the size and potential number of exchanged routing updates.

Decoupling Identification and Location

IP address serves multiple roles nowadays:

- Identification – **Identifier** is a bit string which is used during the communication's lifetime. It identifies communicating parties in a way that IP address verifies the source of packets.
- Localization – **Locator** is a bit string which specifies packet destination where it should be delivered. It locates the place in the Internet topology where a device is attached. Routing protocols interpret IP address as locator and build up routing tables based on situation which routers route traffic towards a destination. Locator is also known as **Point of Attachment (PoA)**.

Identifiers and locators have different requirements on uniqueness and lifetime. Identifiers must be unique with respect to each set of communicating parties, while locators must be unique within one or more routing domains. Identifiers must be valid at least during the

maximum lifetime of a communication between given devices. Locators must be valid as long as routing system within a routing domain needs them.

Traditionally, the IP address is used both as identifier and locator. However, what if any node has more than one IP address, which one identifies it? Topologically device is situated at one place, although PoA addresses express the networks to which device is connected. Moreover, PoA could have completely different location from the perspective of DFZ.

Multihoming

Multihoming stands for situation when the customer is using two or more ISPs for transit services as it is defined by RFC 4116 (4). The goal of customer is to achieve one or more of following: a) redundancy of Internet connection; b) load-balancing; c) transport layer survivability against outages.

Mandatory prerequisite for multihoming is that every customer is uniquely identified as **autonomous system (AS)** with own **autonomous system number (ASN**). Generally multihoming is nowadays accomplished with the help of BGP which informs others about path to customer's network via two or more ISP transit systems.

Trouble with multihoming is closely connected with IP address semantics problem described in previous subchapter. Assume one router connected with two interfaces (two PoAs) to different ISPs for the sake of requested connection redundancy. If one PoA goes down then it does not imply that whole router and networks behind it are unavailable.

Traffic Engineering

Directing of traffic to use other paths than those precomputed by IGP/EGP is called **traffic engineering (TE)**. We differentiate between two types according to direction of traffic flow: a) outbound TE; b) inbound TE.

TE is performed by tuning BGP attributes of certain router, thus increasing RIB size and introducing additional load to control plane.

1.2 Ideal Solution

One of the major goals for any upcoming change of the Internet architecture is to make routing system scaling independent on the growing number of prefixes, users and interconnections between autonomous systems. It is expected that a solution decouples identifier address namespace from location address namespace in a manner that identifiers would be location-independent, while locators location-dependent. More scalable solution for multihoming is strongly desired to allow organizations multihome without adding pressure to DFZ routing tables. Traffic engineering is necessity to network operation of any organization. However, solution for inbound traffic engineering should pose no burden to DFZ routing tables.

2 THEORY

RFC 6115 clearly states that IETF has rough consensus that: a) separating identity and location of devices as one of the major goals for new architecture; b) multihoming and traffic engineering issues need to be solved in scalable manner. However, there is no consensus on how to do it properly.

Theoretically there are two ways how to decouple identity and locality:

- **Map-and-encap** – It evolves from ENCAPS protocol (5). When a source sends packet towards destination outside of source network, packet must traverse through border router between two address spaces (locator space and identifier space). Here at first border router performs mapping of identifier to appropriate locator ("map" phase). Then packet is encapsulated using returned locator address ("encap" phase). Hence, map-and-encap principle wraps a new header (called *outer header*) using locator addresses around original header (called *inner header*) with identifier addresses. When encapsulated packet reaches destination network, the border router strips off outer header and send original packet towards receiver. Map-and-encap usually does not require changes to hosts or to the core routing infrastructure (that is DFZ). Unfortunately, with additional overlay encapsulation comes size overhead;
- **Rewriting** – Originally this principle comes from papers about 8+8 (6) and later GSE (7). It utilizes IPv6 so that in upper part of IPv6 address PCI's fields is stored locator and in lower part identifier. If source sends packet outside its domain, border router takes addresses containing only identifiers and fills upper bits with appropriate locators. Then locators are removed from addresses upon reception by destination border router. Rewriting schemes may differ whether they perform either destination or both destination and source addresses rewrites.

According to (8), possible solutions could be categorized into two classes which are not in opposite. Over the years following terms were established to describe them:

- **Core-Edge Separation (CES)** – A subset address space (*edge*) corresponding to endsite addresses is separated from the transit DFZ (*core*). This "edge" address space is than handled differently for routing. Subsequently DFZ routing table increases its site only a new ISP transit network instead of a new edge network. Some kind of mapping system is needed to glue core and edge address spaces. CES is depicted schematically on Fig. 3 where it shows communication between *PC-A* and *PC-B* using (green) identifiers and (red) locators;

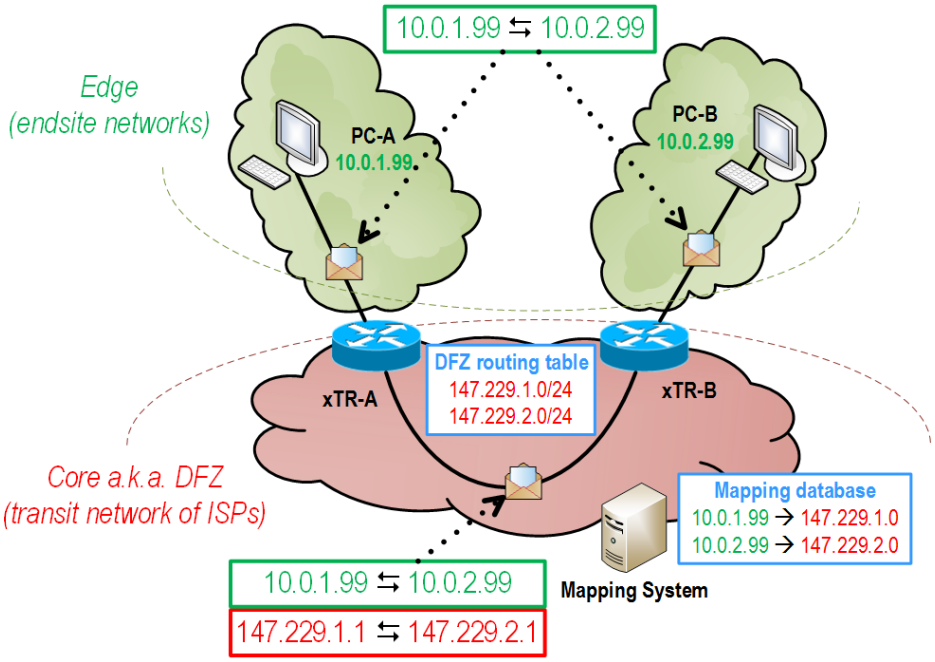

Fig. 3: Core-Edge Separation solution

- **Core-Edge Elimination (CEE)** – The goal of CEE is to eliminate all PI and de-aggregated PA prefixes from the core. Hosts then use either PA addresses provided by ISPs or usually something different (not in IP address name space) as identifier. Some changes in host network behavior are necessary to deploy CEE. Illustrated on Fig. 4.

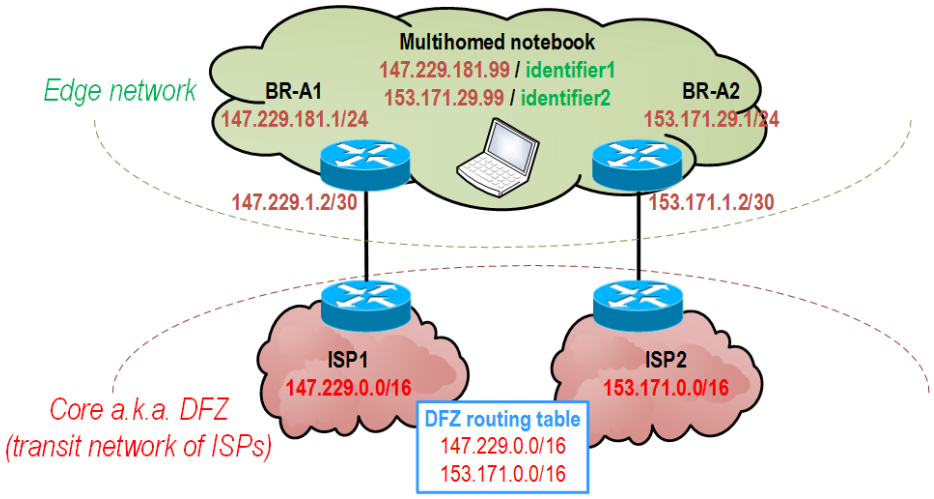

Fig. 4: Core-Edge Elimination solution

3 EXISTING PROPOSALS

Down below is Tab. 1 that summarizes solution candidates. It is outside the scope of this paper to describe them in depth. Hence, astute reader is advised to follow bibliography links.

Locator/Id Split Protocol (LISP)
LISP focuses on separation of locators and identifiers into two distinct address spaces using mapping and encapsulation on routers residing on the borders between those two spaces. Only locators are present in DFZ, thus are possible subject of topological aggregation. With separation of identifiers comes ability to renumber cost effectively. LISP contains by design traffic engineering techniques so that more-specific prefixes could be removed from global routing table. With LISP there is no need to change hosts or DFZ routers. LISP utilizes robust mapping system based on pull model, where queries are data driven. However, it may introduce delay or even packet losses, when ID-to-loc mapping is being discovered. (9)
Host Identifier Protocol (HIP)
Network layer employs IP address as locator, transport and application layer uses identifier in form of cryptographic private-public key pair. Each host is responsible for generating this kind of pair. HIP makes use of DNS or distributed hash table (DHT) to obtain identifier. (10)
Level 3 Multihoming Shim Protocol for IPv6 (Shim6)
Shim6 splits locator/id in a manner that IPv6 address field contains locator and extension header contains identifier. Shim6 employs initial 4-way handshake with DNS lookup during which locator sets are exchanged. Keepalive mechanism tracks locator's reachability. (11)
Routing Architecture for the Next Generation Internet (RANGI)
RANGI append one new layer between network and transport layer just as HIP. Hence, flows and connection are bound to host identifier instead of IP address that now serves as a locator. Unlike to HIP, RANGI host identifiers are hierarchical with organized structure. (12)
Internet Vastly Improved Plumbing (Ivip)
Ivip works with map-and-encap principle as LISP. However, Ivip uses global mapping system instead of hierarchical pull model. It maps only single locator to a given identifier and mappings are updated in real-time. Ivip employs direct IP-in-IP encapsulation. (13)

Hierarchical IPv4 Framework (hIPv4)
hIPv4 introduces additional hierarchy of IPv4 address space by dividing it into area and endpoint locators. Both of them are inserted as optional fields into new shim header between network and transport layer. hIPv4 utilizes DNS for locator distribution. (14)
Name Overlay Service for Scalable Internet Routing (NOL)
NOL utilizes session layer and introduces new devices performing translation between public PA and private PI address namespace which prevent PI to enter DFZ. NOL leverages DNS to store name as a new kind of record. (15)
Global Locator, Local Locator, and Identifier Split (GLI-Split)
GLI-Split decouples addresses into global/local locators and static identifiers. It encodes two different namespaces (each one 64 bits or less) onto single IPv6 address. The communication with legacy Internet is without any proxies or stateful NAT. (16)
Tunneled Inter-Domain Routing (TIDR)
Loc/ID split is performed on BGP level as a new attribute. When a packet to identifier prefix is being routed, it is encapsulated into tunnel. (17)
Identifier-Locator Network Protocol (ILNP)
ILNP decouples identity and locality inside IPv6 address field. Multiple locators might be used by a device simultaneously, whereas applications bind to single identifier. ILNP needs DNS for backward/forward resolution of locators/identifiers to domain name. (18)
Name-Based Sockets (NBS)
NBS are a new alternative for socket-based communication. Unlike nowadays BSD sockets that are bind to IP addresses, NBS are bind to domain names. Applications communicate using domain names where appropriate IP address selection is leaved on TCP/IP stack. (19)
A Practical Transit-Mapping Service (APT)
APT is a copy of LISP with operational restrictions that helps to more clear Loc/ID split design. APT uses periodical synchronization of mapping system. Identifier to locator mappings are carried using new BGP attribute. (20)
Internet Routing Overlay Network with Routing and Addressing in Networks with Global Enterprise Recursion (IRON-RANGER)
IRON-RANGER utilizes own tunneling and path MTU discovery protocol called SEAL which redefines semantics of some ICMP messages. IRON-RANGER is architecturally derived from ISATAP. (21)
Tunneling Route Reduction Protocol (TRRP)
TRRP interconnects border routers between core and edge using GRE. DNS lookup (above overloaded TXT resource record) helps to find tunnel endpoint. TRRP does not support multicast. (22)
Six/One Router (Six/One)
Six/One rewrites edge's local and core's remote addresses at the borders. Six/One takes advantage of special IPv6 extension header. (23)

Tab. 1: Brief description of existing proposals

4 COMPARISON

The following table Tab. 2 summarizes properties of each proposal above. Abbreviations used as columns names means:

- *mrd* – Whether proposal employs map-and-encap ("M") or rewrite principle ("R") or it is something inherently different ("diff");
- *CE* – Whether proposal is Core-Edge Separation ("CES"), Core-Edge Elimination ("CEE") or generally different ("diff") solutions;

- *IPv = Internet Protocol version* – Which IP version does proposal supports ("v4/v6/v4v6");
- *RS = Routing Scalability* – Whether proposal is a relief to DFZ ("yes/no");
- *DIL = Decoupling of Identification and Localization* – Whether proposal performs ("yes") locator/identifier split or not ("no");
- *MH = Multihoming* – Whether proposal supports better multihoming or not ("yes/no") or it is supported conditionally together with multi-path protocol ("cond");
- *Mob = Mobility* – Whether proposal supports seamless mobility or not ("yes/no") or it is supported conditionally together with utilization of multi-path protocol ("cond");
- *TE = Traffic Engineering* – Whether proposal contains TE by design or not ("yes/no") or it is supported conditionally with utilization of multi-path transport protocol ("cond");
- *Ren = Renumbering* – *Whe*ther proposal supports easier renumbering ("yes/no");
- *Dep = Deployability* – Whether proposal allows communication between upgraded and non-upgraded devices ("yes/no") or whether it is not applicable ("n/a").

Name	mrd	CE	IPv	RS	DIL	MH	Mob	TE	Ren	Dep
LISP	M	CES	v4v6	yes	yes	yes	yes	yes	yes	yes
HIP	R	CEE	v6	yes	yes	yes	yes	no	yes	no
SHIM6	R	CEE	v6	no	yes	yes	no	no	no	yes
RANGI	R	CEE	v6	yes	yes	yes	yes	yes	yes	yes
Ivip	M	CES	v4v6	yes	yes	yes	yes	yes	yes	yes
hIPv4	diff	diff	v4	yes	yes	cond	cond	cond	yes	no
NOL	R	diff	v4v6	yes	yes	yes	yes	yes	no	no
GLI-Split	R	CEE	v6	yes	yes	yes	yes	yes	yes	yes
TIDR	M	CES	v4v6	no	yes	yes	no	yes	yes	yes
ILNP	R	CEE	v6	yes	yes	yes	yes	yes	yes	yes
NBS	diff	CEE	v4v6	yes	yes	cond	cond	cond	no	no
APT	M	CES	v4v6	yes	yes	yes	yes	yes	yes	yes
IRON-RANGER	M	CES	v4v6	yes	yes	yes	yes	yes	yes	yes
TRRP	M	CES	v4v6	yes	no	yes	no	yes	no	yes
Six/One	R	CES	v6	yes	yes	yes	no	no	yes	yes

Tab. 2: Properties comparison of existing proposals

Let us focus on comparing CES and CEE solutions because they are majority of proposals. CES are believed to be superior to CEE and subsequent paragraphs provide some overview about pros and cons of both.

CES resulting features:

- Locator/Identifier split is generally performed as depicted on the Fig. 5:
- Edge networks are separated from DFZ routing tables or are at least highly aggregated. Routing scalability is visible in direct proportion to how widely is CES solution adopted;
- CES benefits are available immediately to adopters – multihoming, inbound TE and if possible also mobility;

- Deployment of CES does not affects DFZ routers, but new devices on the border between core and edge are needed to interconnect this two address spaces together with mapping system;
- CES solutions do not require host stack, API or application changes;
- Tunneling and overlaying imposes additional size overhead on fragments, thus introducing MTU concerns when employing CES.

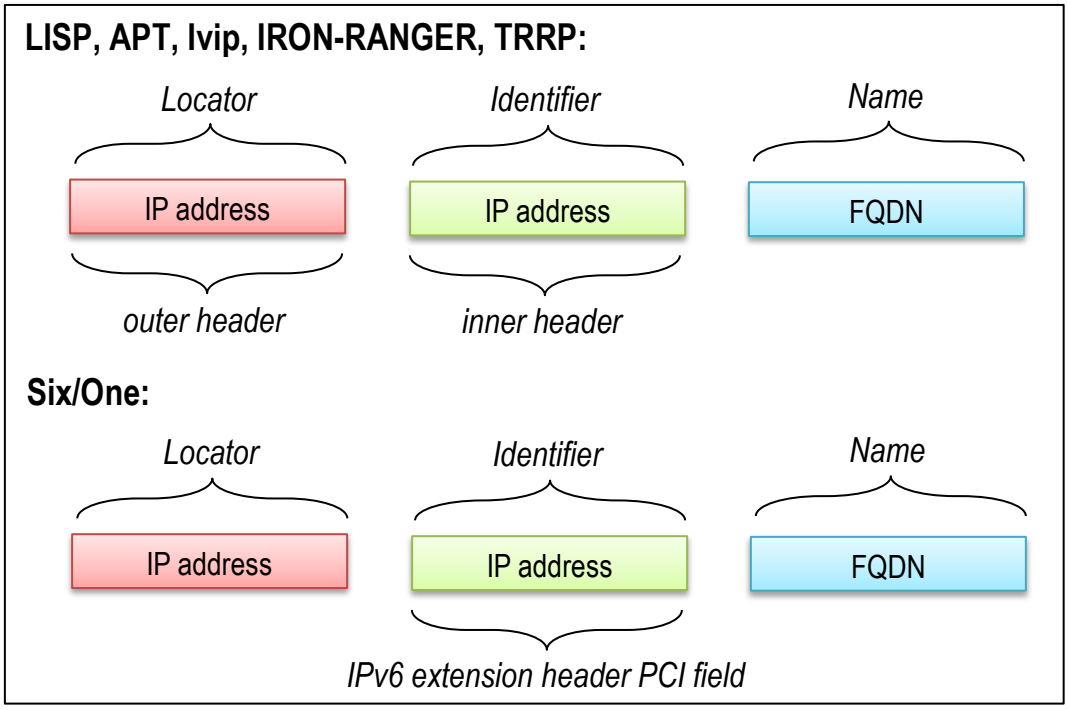

Fig. 5: Kinds of CES locator/identity split

CEE resulting features:

- The most of CEE solutions separates locator/id in a way that there both of them are completely different namespaces. Some of them are depicted on Fig. 6;
- CEE benefits are visible and widely available to adopters only after majority of network migrate;
- Routing scalability is attained in a way that applications are no longer dependent on stable PI (or de-aggregated PA) addresses. Hence, PA addresses could be easily preferred and administratively more available than PI addresses.
- CEE host stack must determine which locator should use. Besides that, potential set of locators could be retrieved, thus implying resolving multihoming, inbound TE issues and ideally mobility issues;
- DFZ routers are not affected and no additional tunneling devices are needed, however a new infrastructure (or at least upgrade of current one, i.e. DNS) must be present to provide mapping between identifiers and locators;
- CEE solutions needs host stack changes and applications augmentations;
- The most of CEE solutions do not support IPv4 and have some troubles with NAT so additionally clutches are needed.

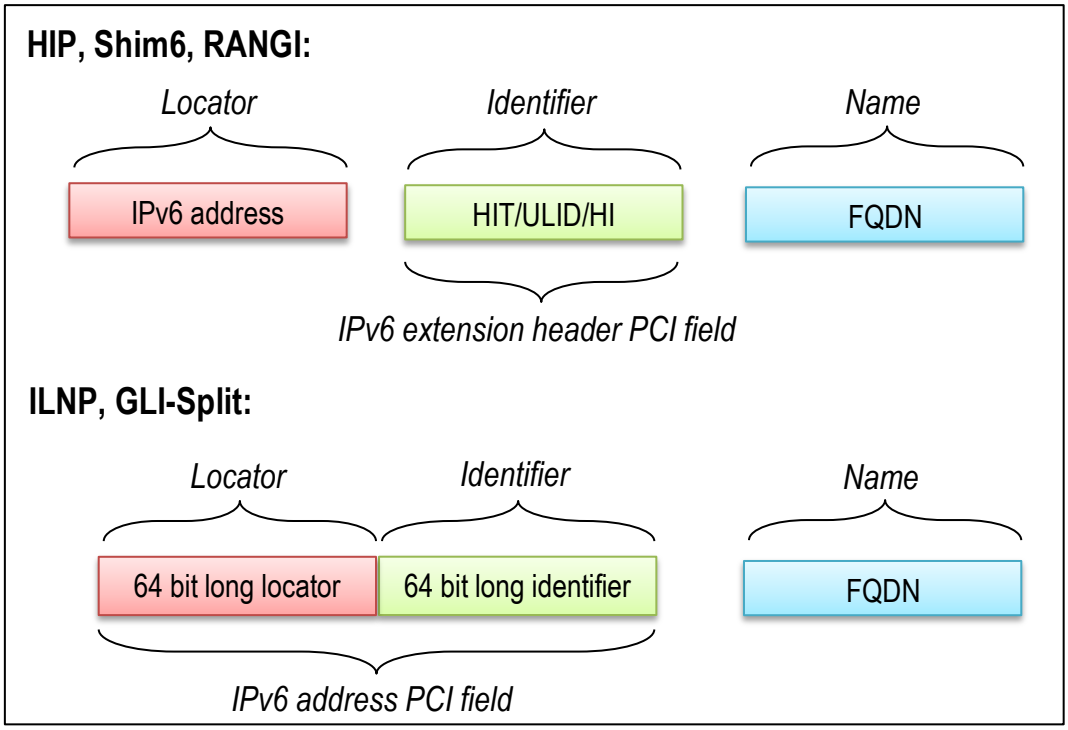

Fig. 6: Kinds of CES locator/identity split

5 CONCLUSION

It is assumed that CES are easier for voluntarily adoption rather than CEE. On the one hand, purpose of routing system is to serve hosts, thus goal is to make routing system more scalable with the help of CES solution that targets network not hosts. On the other hand, CEE solutions are believed to lead to better final shape of the Internet, because of: a) routing should be as simple as possible without unnecessary tunneling clutches; b) utilization of IP address as identifier is a fundamentally wrong concept. CES is "network-centric" and CEE is "host-centric". Unfortunately, synthesis between CES and CEE does not exist.

Both of them need a scalable mapping system. Nevertheless, CES mapping system is arguably more efficient because: a) CES lookups are needed only during initial communication towards a host inside edge network in opposite to CEE lookups that must be performed by senders and receivers for any newly established communications; b) CES mapping system is better designed for caching to alleviate unnecessary resolutions; c) it is unlikely that organizations already using PI addresses would down-grade for PA addresses.

Development of ILNP is pursued further in IETF. However, from our perspective the most promising is LISP because implementation already exist and is supported by vendors like Cisco; also LISP can coexist in both IPv4 and IPv6 world and provides benefits since day one of deployment.

This work was supported by the Brno University of Technology and by the research grant IT4Innovation ED1.1.00/02.0070 by Czech Ministry of Education Youth and Sports.

Bibliography

1. **Li, T.** RFC 6227: Design Goals for Scalable Internet Routing. [Online] May 2011. http://tools.ietf.org/html/rfc6227.
2. **Meyer, D., Zhang, L. and Fall, K.** RFC 4984: Report from the IAB Workshop on Routing and Addressing. [Online] September 2007. http://tools.ietf.org/html/rfc4984.

3. **Huston, G.** BGP Reports - BGP Table Data. [Online] August 7, 2013. http://bgp.potaroo.net/index-bgp.html.
4. **Abley, J., et al.** IPv4 Multihoming Practices and Limitations. [Online] http://tools.ietf.org/html/rfc4116. RFC 4116.
5. **Hinden, R.** RFC 1955: New Scheme for Internet Routing and Addressing (ENCAPS) for IPng. [Online] June 1996. http://tools.ietf.org/html/rfc1955.
6. **Smart, R. and Clark, D.** [RRG] GSE History. [Online] January 1995. http://www.ietf.org/mail-archive/web/rrg/current/msg02455.html.
7. **O'Dell, M.** GSE: The Alternative Addressing Architecture for IPv6. [Online] February 1997. http://tools.ietf.org/html/draft-ietf-ipngwg-gseaddr-00.
8. *Towards A New Internet Routing Architecture: Arguments for Separating Edges from Transit Core.* **Jen, D., et al.** Los Angeles : UCLA, 2008. DOI:10.1.1.156.147.
9. **Farrinaci, D, et al.** RFC 6830: The Locator/ID Split Protocol (LISP). [Online] January 2013. http://tools.ietf.org/html/rfc6830.
10. **Moskowitz, R. and Nikander, P.** RFC 4423: Host Identity Protocol (HIP) Architecture. [Online] May 2006. http://tools.ietf.org/html/rfc4423.
11. **Nordmark, E. and Bagnulo, M.** RFC 5533: Shim6: Level 3 Multihoming Shim Protocol for IPv6. [Online] June 2009. http://tools.ietf.org/html/rfc5533.
12. **Xu, X.** Routing Architecture for the Next Generation Internet (RANGI). [Online] August 2010. http://tools.ietf.org/html/draft-xu-rangi-04.
13. **Whittle, R.** Ivip (Internet Vastly Improved Plumbing) Architecture. [Online] March 2010. http://tools.ietf.org/html/draft-whittle-ivip-arch-04.
14. **Frejborg, P.** RFC 6306: Hierarchical IPv4 Framework. [Online] July 2011. http://tools.ietf.org/html/rfc6306.
15. *Name overlay (NOL) Service for Improving Internet Routing Scalability.* **Wang, Y., Zhang, W. and Bi, J.** Venice, Italy : Second International Conference on Advances in Future Internet (AFIN), July, 2010. pp. 17-21. 978-1-4244-7528-5.
16. *Global Locator, Local Locator, and Identifier Split (GLI-Split).* **Menth, M., Hartmann, M. and Klein, D.** 1, Basel, Switzerland : MDPI AG, January, 2013, Future Internet 2013, Vol. V, pp. 67-94. ISSN 1999-5903.
17. **Adan, J.** Tunneled Inter-domain Routing (TIDR). [Online] November 2006. http://tools.ietf.org/html/draft-adan-idr-tidr-01.
18. **Atkinson, R. and Bhatti, S.** RFC 6740: Identifier-Locator Network Protocol (ILNP) Architectural Description. [Online] November 2012. http://tools.ietf.org/html/rfc6740.
19. **Ubillos, J., et al.** Name-Based Sockets Architecture. [Online] September 2010. http://tools.ietf.org/html/draft-ubillos-name-based-sockets-03.
20. **Jen, D., et al.** APT: A Practical Transit Mapping Service. [Online] November 2007. http://tools.ietf.org/html/draft-jen-apt-01.
21. **Templin, F.** RFC 5720: Routing and Addressing in Networks with Global Enterprise Recursion (RANGER). [Online] February 2010. http://tools.ietf.org/html/rfc5720.
22. **Herrin, W.** Tunneling Route Reduction Protocol (TRRP). [Online] http://bill.herrin.us/network/trrp.html.
23. *Six/One Router: A Scalable and Backwards Compatible Solution for Provider-Independent Addressing.* **Vogt, C.** Seattle, USA : ACM, 2008. 978-1-60558-178-1/08/08.

Kontaktní údaje

Ing. Vladimír Veselý
Vysoké učení technické v Brně, Fakulta informačních technologií
Božetěchova 2, 612 66 Brno

Tel: 541 14 1321
email: ivesely@fit.vutbr.cz

Prof. Ing. Miroslav Švéda, CSc.
Vysoké učení technické v Brně, Fakulta informačních technologií
Božetěchova 2, 612 66 Brno
Tel: 541 14 1288
email: sveda@fit.vutbr.cz