One solution for IPv6 addressing and routing

Victor Grishchenko

Abstract

The addressing/routing model presented in this article is one more attempt to go beyond simple ad-hoc in improving IP routing. The model implements a vision of "ocean of connectivity and absolute minimum of support". By realigning basic routing primitives I sketch a routing model that is simple, flexible, universal, self-maintained and self-reconfigurable, serves both classical wired, new wireless and envisioned ubiquitous internet infrastructure. The resulting scheme is applicable both locally and globally. Computational complexity estimates for the proposed model are very promising.

The current state of IPv6 is depressing. Rates of IPv6 adoption resemble more of "IPv6 is technology of the future and always will be" than of the fax effect in action. One of particular obstacles preventing IPv6 adoption is lack of multihoming support. IPv6 route aggregation technology relies heavily on strict addressing hierarchies, which are not multihoming-friendly. The latest result by the multi6 IETF working group is RFC 4219 [25], the standard questionnaire for proposed IPv6 multihoming solutions. The only undoubtable IPv6 advantage is "just more bits". By all other means, including the routing technology, IPv6 is just "IPv4+": RIP, OSPF and BGP for IPv6 are carbon copies of their originals. Because of the aforementioned considerations, I will skip any IPv6 backward compatibility issues (because the current IPv6 legacy just does not worth the price).

The objective of the proposed model is: given a link (2nd) layer, to configure automatically network (3rd) layer, which is able to provide services with some practical QoS to the transport (4th) layer. Thus, 3rd layer has to be self-deployed, self-maintained and self-reconfigured.

Scope of the paper is limited to the 3rd layer of the TCP/IP stack, sometimes addressing changes induced on the 4th layer. No practical migration plan or commercial issues are considered.

Special instruction to the reader: although the most of design patterns used are well-known (CIDR, multiple addresses, IX-rooted address assignments, etc), the reader has to be warned that in this particular architecture use of a pattern may differ from its current use. Thus, a reader has to estimate consequences of this or that solution with caution; despite a solution itself is well-known, a context of its application may be different.

1 General approach

Mission of a network is to deliver data. A good network must also have low CAPEX and OPEX, high reliability, quality of service and flexibility (adaptability to change). Summarizing extensive network research, RFC 3439 [17] finds complexity as a primary danger for the objectives listed. It is a notable fact [12] that human mistakes are the root cause of 80% of outages. Necessity of human intervention is also a natural CAPEX/OPEX driver.

Given that considerations, the proposed TAMARA model has an objective to simplify routing domain as far, as possible and to make human interaction as unnecessary, as possible. (TAMARA stands for Topology-aware Actively auto-Maintained Agile Routing Architecture.)

1.1 Complexity

One feature of the current routing architecture is abundance of pure virtual, but independent entities (things-in-themselves). IP addresses are different from the actual network topology, so a mapping is needed. Autonomous systems contain some (rather random) prefixes, so a mapping is needed. Intranet addresses are different from Internet addresses, so one more mapping is needed. And so on, and so forth.

The main technical motivation behind the "area" kind of entities (OSPF areas, autonomous systems – ASes) is a stereotypical approach to solving the problem of routing computational complexity. Indeed, by dividing a generic graph of size N into $\sim \sqrt{N}$ areas of size $\sim \sqrt{N}$ we may lower routing computational load from $O(N)$ to $O(2\sqrt{N})$ per node, i.e. inter-area routing plus intra-area routing. The inevitable overhead is stretch (e.g. 3 for Thorup-Zwick [13]). Of course, using some good hierarchic assignment we may need just logarithmic amount of routing data, but under the IPv4 policies, this kind of assignment is applicable at the edges only. One of initial IPv6 goals was to assign addresses in tree-like fashion, but this idea conflicts with multihoming.

Thus, the proposed model (Sec 2) makes an attempt to greatly simplify the routing domain by using only one routing entity: a prefix (an IPv6 prefix, like *a:b:c::*, which is also a device locator, i.e. it is used as an address).

Prefixes-addresses are assumed to be just a reflection of the network's topology. Prefixes-addresses have no administrative or business semantics. If the topology changes, addresses also change. This approach is widely known as locator/identifier split. The problem of computational complexity is resolved in another way, avoiding introduction of extra entities.

1.2 Human factor

Necessity of human interaction is a natural CAPEX/OPEX driver. Current dynamic configuration solutions [5, 6] target the trivial case of leaf devices obtaining their addresses from routers. Theoretically, this approach could be extended recursively to a tree of routers. A key limitation for this approach (i.e. why we do not have self-configuring networks) is the fact that a router can't recognize which uplink or path is better, so topology alteration involves human effort in most cases.

Today, on the lower end we have dynamic address configuration for leaf devices (DHCP, IPv6). On the higher end we have routing system configured with deterministic shortest path algorithms (OSPF, BGP). The objective is to melt and blend both approaches into some universally scalable self-configuration solution using some metric that lets a router to distinguish "good" paths and uplinks from "bad". On different levels, different approaches might be *accented*, but the solution must allow the full set of options both for bigger and smaller entities – all in self-similar, fractal fashion.

1.3 Robustness

Contemporary DFZ routing table growth [18] indicates strong demand for multihoming and load balancing among smaller and smaller entities. Thus, the topology of the Internet becomes more rich and average degree of AS connectivity graph grows [19]. From my personal experience, some kinds of even smaller players also have high demand for multihoming solution and, if a separate AS or provider-independent addresses are unavailable, then some NAT solution is usually used. Thus, the proposed model has to allow multihoming at any scale. Ideally, mobile device handoff in heterogeneous environments is supposed to be implemented as a smaller version of multihoming. A device uses several uplinks at once: some new uplinks are gained, some old are lost, but the device remains reliably connected without a need for complete uniform GSM-like coverage.

Thus, robustness is planned to be achieved by the simplest way possible. Namely, by excessive redundancy at all levels of topology – both topological redundancy (multihoming, backup links) and overprovisioning.

2 The TAMARA model

2.1 Reachability

The best way to keep things simple is to minimize primitives. The central point of this paper is that a notion of *ip prefix* may be the only routing entity. Prefixes establish some qusi-hierarchic stucture, thus defining natural way of aggregation. Existence of implicit hierarchy is noted even in AS graph which is formally flat [11]. In scale-free topology, although it is not tree-like, some kind of hierarchy still has place. This phenomenon is described as skeleton [21], whose existence, probably, emerges from the effect of funneling [1] and power-law node degree distribution.

Proposition 1 *IP prefix is owned by some device (either one physical device or a well-connected set of). It is simultaneously an address of the device and an "area descriptor" for all the downlink devices. Any device may redistribute sub-prefixes of its owned prefixes further to downlink devices (complete stateful self-configuration,* $DHCP^n$*).*

World of devices is the *gemba* of internet routing; real routing decisions happen there. Such device-bound IP addresses are tightly dependent on the actual 3rd layer topology and thus are much more meaningful in the context of routing. Also, this point-centric addressing approach makes RTT applicable as a distance metric (inter-AS RTT is meaningless in many cases).

Still, anyone is free to centrally administer a group of devices to build a functional equivalent of today's autonomous system.

The natural roots for this addressing hierarchy are IXes (internet exchange points), as [9] proposes. It is reasonable because IXes are "points" with concrete geographical locations and good connectivity. This allows to practice geo-aggregation and, ideally, to have logarithmic-size routing tables under some conditions [26], see Sec. 3.1.

Now, the proposal is similar to IPv6 clean hierarchy plus the device-owns-addresses principle. And it targets only the case of hierarchical routing, which mostly takes place at the edges these times. Still, this way of routing equals to "skeleton-only" in the terms of [21]. In most cases, skeleton holds $> 50\%$ of graph's betweenness centrality, so hierarchical routing may not be as bad as it probably seems.

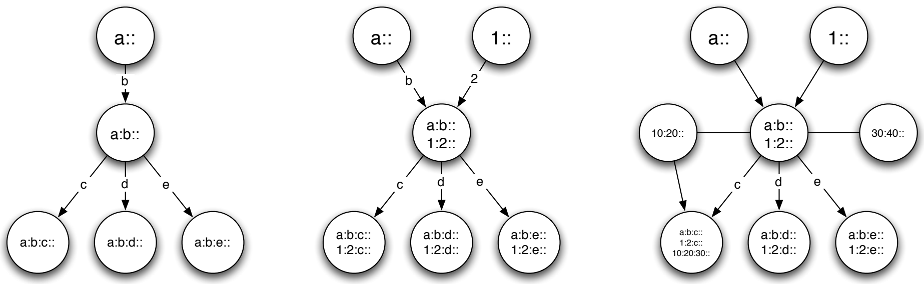

Figure 1: Connectivity patterns. The first picture corresponds to pure hierarchic routing, similar to telephone (PSTN) network or the IPv6 ideal, the case of Sec. 2.1. The second picture corresponds to the case of multiple uplinks, see Sec. 2.2. The third picture includes peering links, see Sec. 2.3. For the sake of simplicity, it is assumed that each device spends 16 bits to number its downlinks (so, 1:2:d::). From the practical point of view, there are only 8 16-bit fragments in an IPv6 address, so smaller chunks have to be used, e.g. 8-bit.

2.2 Reliability and redundance

Hierarchic addressing and routing is obviously not enough.

Proposition 2 *A simple and logical extension is to let a device to own several prefixes to handle the cases of multihoming, IP-level handoff, etc.*

This idea is not original from any point of view. Thus, a device takes subprefixes from several uplinks and each device will finally have a bunch of addresses. Any one address is sufficient to have connectivity. By choosing a particular address a sender may e.g. route packet by a shorter path. Still, each routed packet (at the IP layer) has one destination address. Selection of particular address(es) to use can be transparently implemented on the 4th layer by generalizing e.g. TCP. Required changes include "vectorization" of timing and other calculations of TCP. "Vectorized" TCP can be transparently compatible with today's TCP if using address vector of size 1.

Thus, all addresses of a device will form a directed acyclic graph (toponame), see Fig. 1. Each path in this graph describes some network path from respective root IX to the target device. For estimations on size of address bunch, see Sec. 3.

At this step, TAMARA achieves redundancy which lower many risks of dynamic configuration (now, a device may *try* new uplinks, still being online). It also provides more freedom for end sites, at different aspects. These extended addresses (toponames) also fit the case of scale-free/power-law topology well [10, 24], namely the key role of hubs and logarithmical path length. Toponames enable flexible routing overcoming fragility and restrictions of pure hierarchy. At this stage, a device may send a packet to other device using any of its addresses. A packet will (simplistically) travel to some common grandparent device and then go "down" to the destination. Both ends may choose whichever address is better.

2.3 Optimization and peering

Not all paths can be calculated from toponames. Some shorter paths use "shortcuts". Shortcuts are peering links, i.e. those not involved into uplink-downlink relation. Existence of those links and paths is not reflected in toponames. To exploit those links we have to employ some logic at intermediary devices (routers). To automate routing decisions we have to set distance metric first.

Distance metric used by routers have to effectively minimize need for human intervention. So, a device has to be able to make sane routing decisions. As a reference case, I choose statistical round trip time (effective RTT, eRTT) as a distance metric. It differs from plain RTT in the case of packet loss: both transmission and retransmission are included into the resulting calculation of the eRTT average. So, eRTT discriminates lossy channels. eRTT is good on several reasons, e.g. it favors real-time traffic and enables geo-aggregation [26]. (Note: a device may freely use a different metric.)

Measurement is performed by direct probing. It is a cheap and reliable method. Particular implementation mechanism may involve extra pings added to the bypassing stream or some additional IPv6 header or other methods. Those probes may be directed to the original packet's destination or to respective prefix from the local routing table, or other variant. Assuming the Net to have diameter of ~ 30 hops [14] and number of probes 1 per thousand of bypassing packets, we may have total overhead of probe packets well under 3% at any point.

A protocol is needed to propagate information about shortcuts (route announces). With the exception of the very top layer (inter-IX links), such a protocol provides just optimization hints for routers. So, an average router is free to support a routing table of size it can afford. The protocol is supposed to be a kind of path [27] or distance [7] vector protocol able to work at arbitrary level of detail in spirit of [26]. In fact, more correct definition is "a path or distance *tree* protocol". The tree of prefixes is deep and detailed for the router's vicinity and is more and more coarse for farther entities.

(It is much trickier to enhance link-state [8] protocol this way. Although, it seems possible to do generally following the spirit of [22].)

3 Estimates

TAMARA provides more possibilities than vanilla IPv4 model. Generally, the explained model is able to maintain richer topologies with excessive redundancy. The question is how computational complexity of TAMARA relates to IPv4.

TAMARA involves two methods of route calculation: by name (i.e. by prefix), and by route announces. eRTT stretch is assumed to be bounded by some value, say 1.1. Prefixes and announces are direct substitutes: an additional prefix describes some valuable route and thus makes announce propagation in that direction unnecessary (see [24] for details).

3.1 Routing complexity, globally

At global scale, RTT is well correlated with great circle distance, mostly due to the speed of light, c (see [26] for details and references). Although this rule is not universal and may be wrong because of peering wars, abundance of cheap fiber and other factors, it is still the normal way of things. A theory [26] based on this correlation, geo-friendly address assignment and distance metrics (i.e. the case of TAMARA) predicts routing tables of logarithmical size assuming fixed eRTT stretch (say, 1.1).

This is due to the fact that a device (a point) may use larger routing aggregates (i.e. shorter prefixes) for farther entities. The dependence between distance and aggregate size is linear. As a packet approaches destination, forwarding routers become more and more aware of the destination's routing details. Further, a router may employ routing table maintenance policies described in [26], e.g. to drop entries providing less than 10% in eRTT gains (or less than 10 Mbytes-seconds of total gains per minute). Finally, simple geometric considerations lead to logarithmical routing table size, at any device [26].

If distance and stretch are measured in eRTT, this also leads to some practical QoS level: packet loss and RTT are minimized, while jitter and out-of-order arrival could be traded for RTT using buffers at end points.

3.2 Routing complexity, 100 km vicinity

In a vicinity of 100km the effect of speed-of-light is negligible, so a network is assumed just to follow scale-free (power-law) topology [10]. In my previous work [24] I modeled toponame-enhanced routing on both synthetic and real-world scale-free graphs. On scale-free graphs, shortest-path routing is resolved using sublinear amount of computational resources per node. Advantages of algorithms [24] used include fully symmetric P2P implementation (no central coordinating entities) and the fact that routing information *grows* sublinearly as a graph grows; no pre-computation or complete topology information is needed at any point.

Reasonable estimates for *shortest path* routing performance is toponames of size $O(n^{0.2})$ (loosely corresponds to prefixes per device) and routing tables of size $O(n^{0.6})$, where n is number of routers in ∼100km vicinity. Also, a device may use just some fraction of own addresses.

Both approaches of Sec. 3.1 and 3.2 describe some ideal edge cases. In fact, geo-approach depends on small value of "fatness of a cover" [26], i.e. number of prefixes per device. So, it approximates well on large distances (some thousands of kilometers) where only large aggregates are used. Locality-ignorant approach is also a kind of approximation, because RTT depends on distance in other ways than the speed-of-light limitation. There are also other types of locality e.g. organizational locality, influence of population density patterns and others.

Anyway, TAMARA compares wery well to BGP. BGP needs an order of $\sqrt{N}$ records at each DFZ device imperatively (where N is the total number of IPv4 devices) – without any RTT guarantees. For TAMARA, this order of complexity is kind of the worst case possible (uncompromissed shortest paths, no locality) inside some vicinity (i.e. order of $n^{0.6}$, not $N^{0.5}$).

3.3 Stability

Address stability, namely the frequency of address change, is a major concern for exclusively dynamic address assignment scheme. First, addresses assigned to wired infrastructure have to be changed only on significant topology change. Wired infrastructure close to IXes is unlikely to be ever renumbered. Second, stability of addresses may be ensured by anycast schemes – such as anycast ring, a group of well-connected devices having a shared address. Third, for mobile devices necessity of address change evolves at mechanical timescale ($\gg 1sec$), while network operations happen at faster IP timescale ($\ll 1sec$). Speed of renumbering is not a major concern from the theoretical point of view, because the network is logarithmically deep (i.e. a change in some grand uplink will propagate to any endpoint in logarithmical number of steps).

Routing table (in)stability could be presented as number of record *times* average frequency of record change. It has to be noted that the logic of changes in a maintained routing table is different from the cases of unmaintained tables (e.g. BGP). In fact, both current eRTT of a path, a fact of path withdrawal and a fact of address change could be detected independently by each probing router (i.e. each router maintaining the respective record). On security and stability reasons, announces of new shortcuts are better to propagate slowly. Thus, routing tables can be stable enough.

Load stability is relative rarity of dramatic changes in traffic patterns. (Informally, that channel load does not change from 0% to 100% and back every minute.) The following is the current working hypothesis on stability of traffic flows in the TAMARA model: if frequency of measurement is one order higher than frequency of change and changes occur incrementally, then the network will settle to some "market equilibrium".

For example, if some link goes down, then a device may redistribute at most 10% of the corresponding load to other links every 10 seconds. The remained load destined for the failed link has to be dropped. After each redistribution step, a device has to measure performance of other links once a second at least. This algorithm, hypothetically, avoids "shock wave" of load redistribution or routing microoscillations. Assuming redundant connectivity and overprovisioning, load of the failed link will likely be absorbed by other routes. Also, accidental drops provide feedback both to other routers to align routing and to end points to use other addresses or to use smaller TCP window. Considering the self-similar nature of traffic oscillations [3], eRTT moving average seem to be a convenient control parameter here.

Beyond best educated guesses, some definitive conclusion needs large-scale simulations, involving different attack scenarios, which is problematic under the current level of detail. Thus, it is future work.

4 Discussion

4.1 Parallels

Geo-aggregation is a rather popular approach [9, 15, 16, 4]. I will nickname all the previous proposals involving administrative distribution of locality-dependent addresses as well as IPv6 addresses derived from longtitude and latitude as *naive geo-aggregation* proposals. In the presented model, geographical considerations (e.g. speed of light) define only the RTT lower bound, which is enough to introduce locality and to perform efficient route aggregation (which hop-count models can't do). At the same time, addresses are distributed and routing is performed following the actual network layer topology and not due to some artificial geopolitical considerations.

shim6 [28, 29] is a proposal of upgrading IPv6 stack to use multiple addresses of destination simultaneously. Thus, multihoming is achieved by using multiple addresses at each end point on a multihomed site. Shim6 is primarily targeted to those entities who can't afford larger multihoming tools (ASes, PI addresses) and positioned as a patch to the existing IPv6 architecture. Opinion of many [30] is that it will hardly succeed this way, considering that all IPv6 stacks have to be upgraded and thus complicated, while the gains are much less universal than the expenses. A significant difference from this proposal is that shim6 layer is supposed to transparently function between IP and TCP layers. Thus, it becomes a kind of NAT-inside-the-stack. In my personal opinion, this approach is more than suboptimal for clarity, simplicity and reliability of the architecture.

DHCP [5] is a stateful Dynamic Host Configuration Protocol. It is used to configure leaf devices (end points) in a network. The explained proposal assumes stateful dynamic configuration at all levels of network topology.

NIRA [23] – New Internet Routing Architecture has many common points with the explained model, e.g. path selection by end-points using different pairs of source/destination addresses. Address assignment policies are modeled to maintain hierarchy rooted at Core ASes, not IXes. Address bunch size estimations [23] are not exhausting, but convincing. As far as I understand, computational complexity estimations of [24] also hold for NIRA. Anyway, except for area-centricity vs point-centricity choise, the proposals heavily intersect. (RTT-based metrics are meaningless for area-centric approaches.)

FARA [20] is a very general network architecture proposal. The key concept of FARA, namely the decoupling of end-system named from network addresses is also employed in the explained model.

CIDR [2] and routing prefix aggregation was a temporarily successful IPv4 routing record aggregation approach. CIDR-like aggregatin mechanics is employed in the explained model, with the exception that TAMARA is much more aggregation-friendly than the current IPv4.

Path/distance vector protocols, e.g. RIP, BGP, etc are orthogonal to the key concept of TAMARA. One may interpret TAMARA shortcut announce protocol (a path or distance tree protocol) as an extension to those protocols. Although, BGPv4 [27] already features CIDR aggregation.

5 Conclusion

The main contribution of this work is the concept of actively self-maintained network featuring eRTT as a routing metric and active probing as a base of routing technology. It also summarizes some previous results on computational complexity of routing in scale-free graphs. Other concepts, although well-known, constitute mutually-supportive ensemble in this particular configuration.

The resulting architecture has the following key advantages.

- It is an effort to reverse complexity/robustness spiral effect [17]. Simplicity (less entities) and directly measurable routing metrics make human interaction unnecessary in the topology change – configuration change loop. Absence of human interaction removes the root cause of most of outages and lowers expenditures. Cheaper infrastructure lets to build excessively redundant solutions. Redundancy lowers risks of dynamic configuration, lets simpler solutions and makes human interaction less necessary (at least, lets to act in terms of maintenance shedule, not reaction time). Simpler solutions tend to be cheaper again, etc etc

- This solution is generally applicable both at local and global levels; TAMARA is targeted for digital convergence, masses of self-configurable devices, agile infrastructure and *participatory* model for IP devices.

- TAMARA provides some basic QoS for public IP networks (assuming some reasonable level of over-provisioning).

Simpler technical advantages include smaller routing tables, easy multihoming, IP-level handover, dynamic configuration of network layer, fault-tolerance, place-and-play networks, topology-awareness of all the IP devices and thus, a set of possible innovations regarding mirroring, traffic optimization, network diagnostics etc etc

References

[1] Stanley Milgram: "The Small World Problem", Psychology Today, May 1967. pp 60 - 67.

[2] V. Fuller, T. Li, J. Yu, K. Varadhan: RFC 1519, "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy", 1993

[3] Will E. Leland et al: "On the Self-Similar Nature of Ethernet Traffic", ACM SIGCOMM, 1993

[4] Y. Rekhter, T. Li: RFC1887, "An Architecture for IPv6 Unicast Address Allocation", 1995

[5] R. Droms: RFC2131, Dynamic Host Configuration Protocol, 1997

[6] S. Thomson, T. Narten: RFC2462, IPv6 Stateless Address Autoconfiguration, 1998

[7] G. Malkin: RFC 2453, "RIP Version 2", 1998

[8] J. Moy: RFC 2328, "OSPF Version 2", 1998

[9] R. Hinden, M. O'Dell, S. Deering: RFC 2374 "An IPv6 Aggregatable Global Unicast Address Format", 1998, OBSOLETE

[10] Michalis Faloutsos, Petros Faloutsos, Christos Faloutsos: "On Power-Law Relationships of the Internet Topology". ACM SIGCOMM'99.

[11] L. Gao, "On Inferring Automonous System Relationships in the Internet", IEEE Global Internet, 2000

[12] "Making Smart Investments to Reduce Unplanned Downtime", D. Scott, Tactical Guidelines, TG-07-4033, Gartner Group Research Note, March 1999.

[13] Mikkel Thorup and Uri Zwick: "Compact routing schemes", in proc. of ACM Symposium on Parallel Algorithms and Architectures, 2001

[14] Andre Broido and kc claffy: "Internet topology: connectivity of IP graphs", in Proc. of SPIE ITCom, 2001

[15] T. Hain: "An IPv6 Provider-Independent Global Unicast Address Format", Internet draft, 2001

[16] T. Hain: "Application and Use of the IPv6 Provider Independent Global Unicast Address Format", 2001

[17] RFC 3439 - Some Internet Architectural Guidelines and Philosophy. Editors: R. Bush, D. Meyer, 2002

[18] Tian Bu, Lixin Gao, and Don Towsley: "On Characterizing BGP Routing Table Growth", GlobalInternet 2002

[19] Georgos Siganos, Michalis Faloutsos, Christos Faloutsos: "The Evolution of the Internet: Topology and Routing", Technical Report, 2002

[20] Clark, D., Braden, R., Falk, A., and Pingali, V.: "FARA: Reorganizing the Addressing Architecture". ACM SIGCOMM 2003 FDNA Workshop, Karlsruhe, August 2003.

[21] Kim et al: "Scale-free trees: The skeletons of complex networks", Phys. Rev. E 70, 046126 (2004), available at http://arxiv.org/abs/cond-mat/0403719

[22] Andrew Y. Wu, M. Garland, J. Han: "Mining Scale-free Networks using Geodesic Clustering", KDD 2004

[23] Xiaowei Yang. "NIRA: A New Internet Routing Architecture", PhD thesis (MIT-LCS-TR-967), 2004

[24] Victor Grishchenko: "Computational complexity of one reputation metric", IEEE/CreateNet SECURECOMM'05 SECOVAL, available at http://oc-co.org/articles/computation-rep-rec.final.pdf

[25] E. Lear: RFC4219, "Things Multihoming in IPv6 (MULTI6) Developers Should Think About", 2005

[26] Victor S. Grishchenko: "Geo-aggregation permits low stretch and routing tables of logarithmical size". arXiv:cs.NI/0510028

[27] Y. Rekhter, T. Li, S. Hares: RFC 4271, "A Border Gateway Protocol 4 (BGP-4)", 2006

[28] Site Multihoming by IPv6 Intermediation (shim6) IETF Working Group, charter at http://www.ietf.org/html.charters/shim6-charter.html

[29] E. Nordmark, M. Bagnulo: "Level 3 multihoming shim protocol", 2006

[30] shim6 discussion at NANOG, see http://www.merit.edu/mail.archives/nanog/, February/March 2006