Future Internet Architecture:
Routing challenges, alternatives and perspectives

1st Japan-EU Symposium
Brussels, June 9-10, 2008

Dimitri Papadimitriou

Alcatel-Lucent BELL NV

dimitri.papadimitriou@alcatel-lucent.be

Outline

- Introduction
- Fundamental causes of Internet routing scalability problems
  - Challenges
- Alternatives
- Perspectives

Introduction

RFC 1287: Towards the Future Internet Architecture (Dec.1991)

Five most important areas for architectural evolution:

1) **Routing and Addressing**: most urgent architectural problem, as it is directly involved in the ability of the Internet to continue to grow successfully
2) Multi-Protocol Architecture
3) **Security Architecture**: experience has shown that it is difficult to add security to a protocol suite unless it is built into the architecture from the beginning
4) **Traffic Control and State**: the Internet should be extended to support "real-time" applications like voice and video -> “traffic control” mechanisms
5) Advanced Applications

RFC 1380: IESG Deliberation on Routing and Addressing (Nov. 1992)

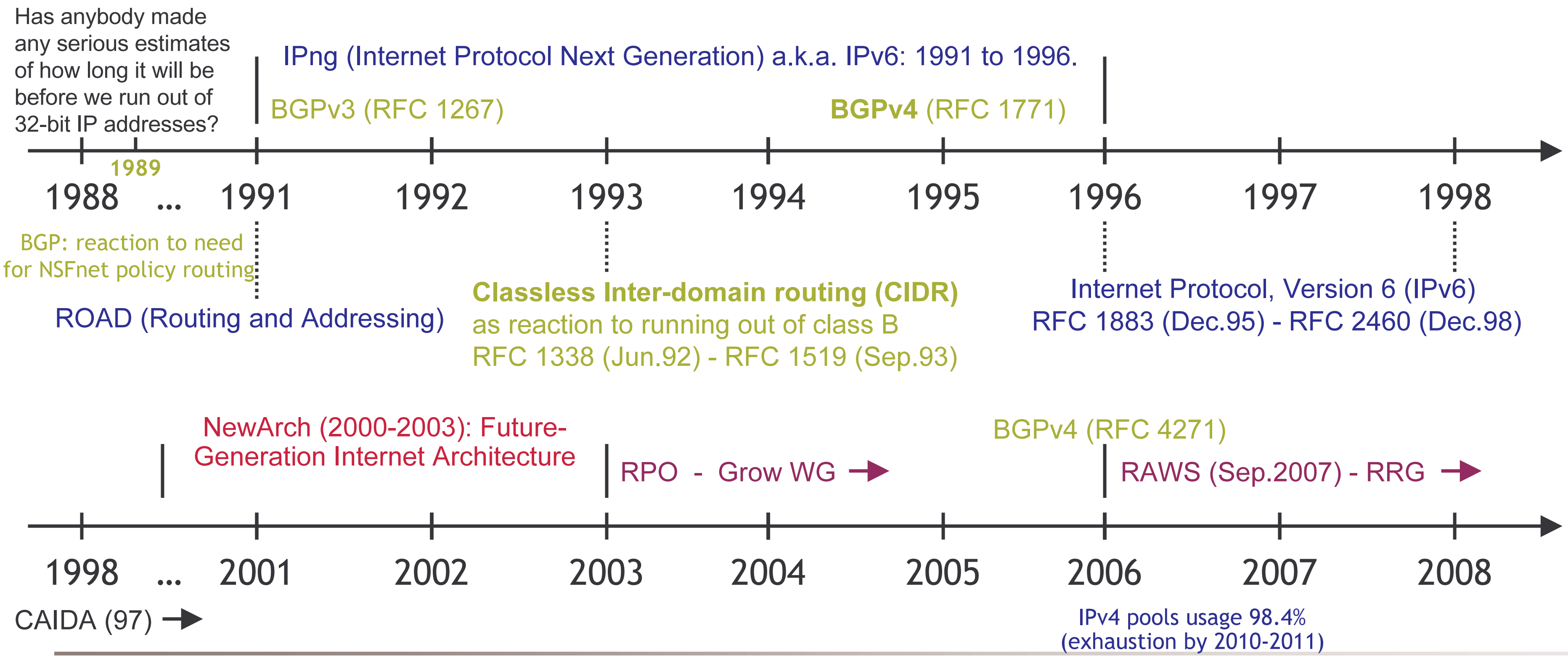
- Summarizes issues surrounding the **routing and addressing scaling problems in the IP architecture**
- Provides a brief background of the ROAD group and related activities in the IETF
- Reports on preliminary Internet Engineering Steering Group (IESG) deliberations on how these routing and addressing issues should be pursued in the Internet Architecture Board (IAB)/IETF

Introduction

RFC 4984: Report from the IAB Workshop on Routing and Addressing (Sep.2007)

Reports outcome of Routing and Addressing IAB Workshop held on Oct., 2006, in Amsterdam

- Goal: develop a shared understanding of the problems that the large backbone operators are facing regarding the **scalability of today's Internet routing system**
- Findings: analysis of the major factors that are driving routing table growth, constraints in router technology, and the limitations of today's Internet addressing architecture



In practice…

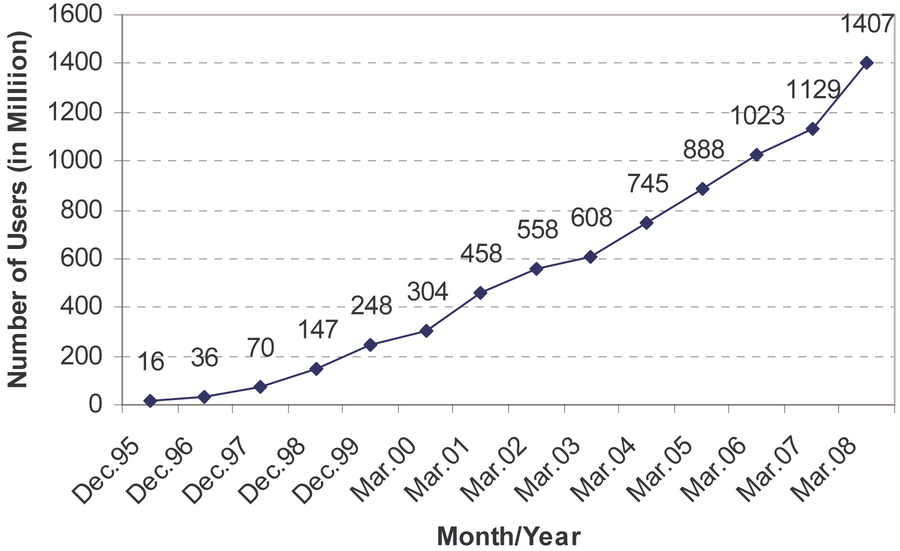
Routing system scalability is a major technological challenge of the Future Internet

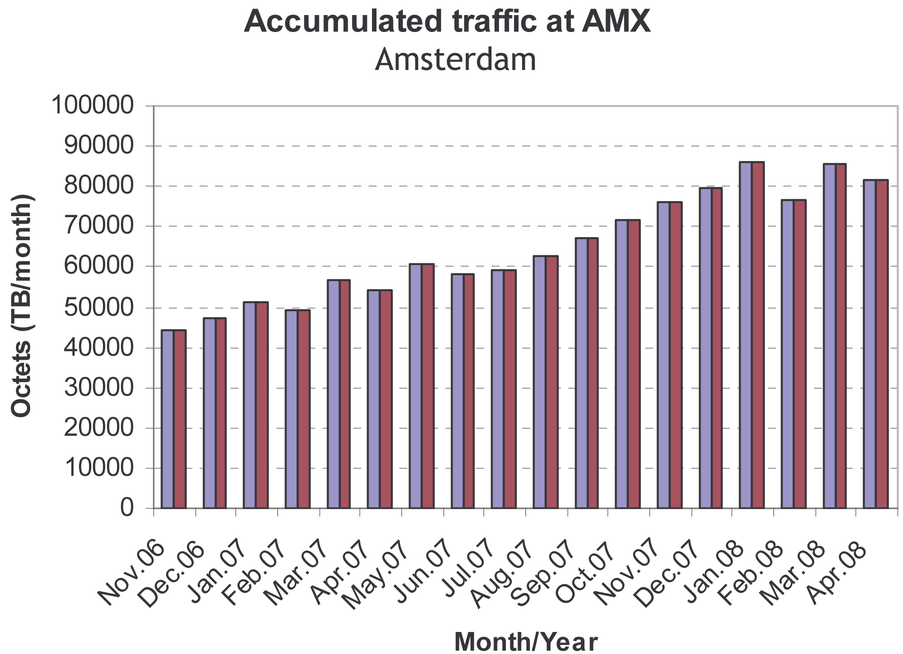
↑ number of routing table entries (traffic engineering/de-aggregation)

↑ number of sites x multi-homing

↑ number of AS's with increasing meshedness but *steady average AS path length*

↑ routing system dynamics (impact on robustness/stability and convergence properties)





Max yearly growth: Nov06->07: 70%

Min yearly growth: Apr07->08: 50%

Fundamental causes of Internet routing scalability problems (1)

Cause 1: Topology vs aggregation

- Host addresses assignment based on topological location
- Conditions to achieve efficient address aggregation and relatively small routing tables (tradeoff routing information aggregation vs routing information granularity)
  - Tree-like graph structure
  - Address assignment that follows topological structure
- Deterioration causes
  - MN mobility (Mobile IP)
  - Site multi-homing (~25% of sites)
  - Traffic engineering (de-aggregation of address prefix): cost vs performance

→ Super-linear growth of routing and forwarding table even if the network itself would not be growing

⇒ Routing protocol must not only scale with increasing network size !

Fundamental causes of Internet routing scalability problems (2)

Cause 2: BGP inter-domain routing system

1. Protocol specifics/implementation: may be circumvented

2. Protocol architecture: BGP is a path-vector protocol (eliminates DV count-to-infinity problem)

→ Path exploration (withdraw/announcement): routers may explore O(N!) (-> computational states) alternate AS paths, N = number of AS, in a complete graph of AS

⇒ Convergence time: upper bound ~ O(N!) and lower bound = Ω[(N-3) x MRAI timer]

Mitigation (examples):

- Root cause analysis/notification (pin location/cause of updates ?): comes with side effects such as complexity and inaccuracy
- Multi AS-path: Backup AS-path (routing diversity): comes with side effect on number of RIB states

→ Exponentially exacerbates the number of possible routing table oscillations

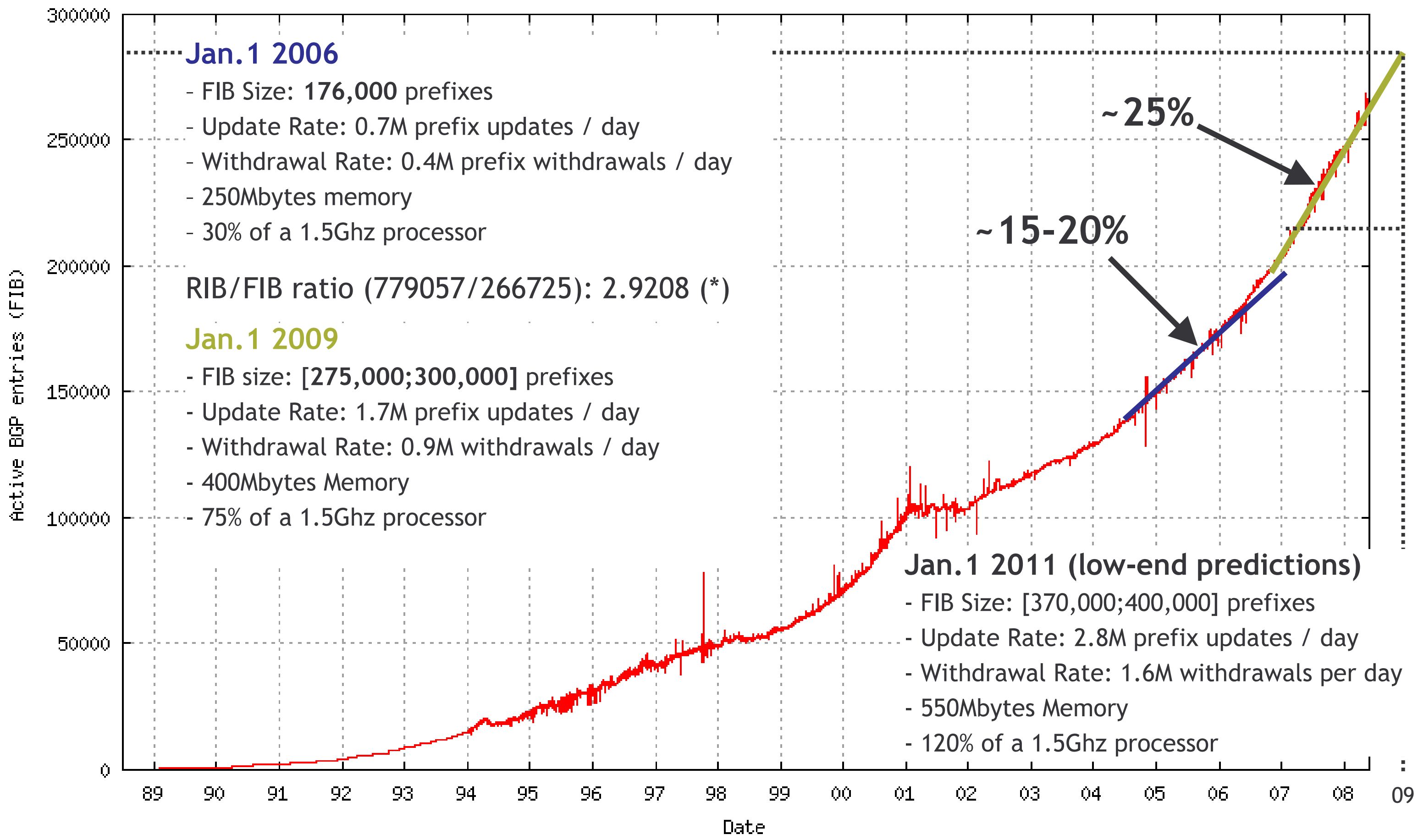
3. Protocol usage: policy-based routing (- no policy distribution)

→ inter-AS oscillations (policy conflicts: local preferences over shortest path selection)

→ intra-AS oscillations (MED-induced oscillations*)

(*) can be eliminated by ensuring cross-AS monotonic ranking

Growth of Active BGP Entries in FIB (from Jan'89 to Mar'08)



(*) - RIB/FIB ratio can vary from ~3 to 30 (function of the number of BGP peering sessions at sample point)

Source: BGP Routing Table Analysis Reports on AS65000 - http://bgp.potaroo.net

Expansion of Internet between 2005 and 2006

Prefixes: 173,800 – 203,800 (+17%)

AS Numbers: 21,200 – 24,000 (+13%)

Addresses: 87.6 – 98.4 (/8) (+12%)

Average advertisement size: smaller (8,450 – 8,100)

Average prefixes per update: smaller (2.1 - 1.95)

Average address origination per AS: smaller (69,600 – 69,150)

Average AS Path length: steady (3.4)

AS transit interconnection degree: growing (2.56 – 2.60)

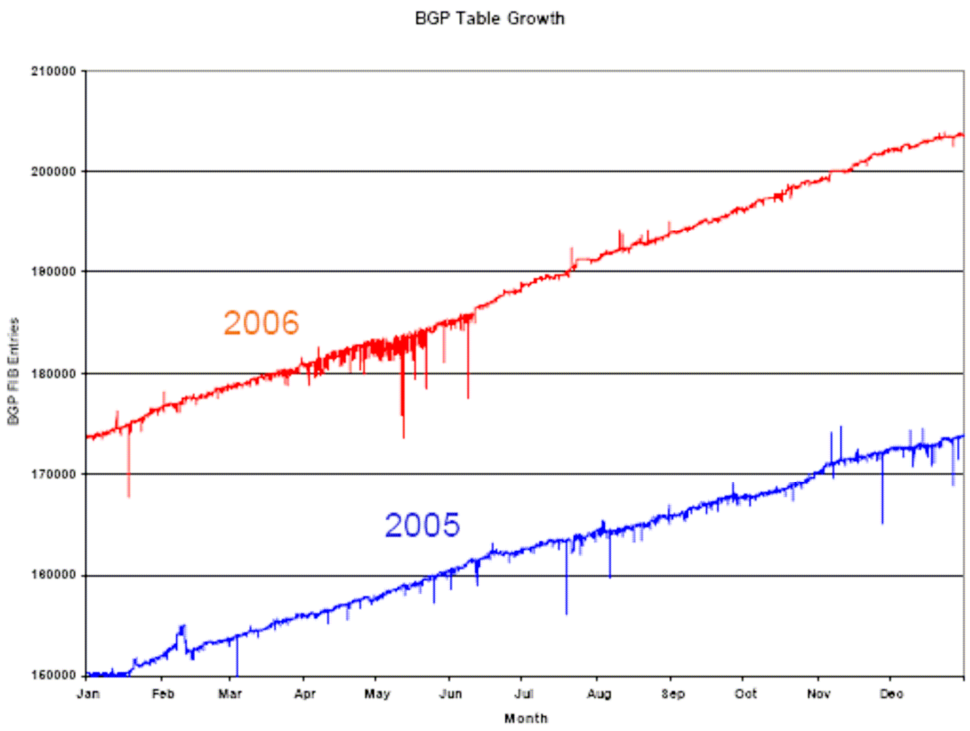
⇒ IPv4 network becomes

- denser (more interconnections)
- finer levels of advertisement granularity (more specific advertisements)

⇒ Higher levels of path exploration before stabilization on best available paths

IPv4 in 2006

Total BGP FIB Entries



Source: IEPG, <http://www.potaroo.net>

Internet routing system - BGP scalability impact

Scaling of routing algorithm (RT size growth rate > linear)

1. Routing engine / system resource consumption -> cost growth rate ~ 1.2-1.3/2years

- Routing space size

  ↑ #routing table entries ⇒ ↑ memory

  ↑ #routing table entries ⇒ ↑ processing and searching (lookup)

- Number of peering adjacencies between routers

  ↑ #peering adjacencies ⇒ ↑ memory (due to dynamics associated with routing information exchanges)

2. Exacerbates BGP convergence time

- BGP convergence time is limited by access speeds of DRAM (used for RIB storage)
  - DRAM capacity growth rate: ~4x every 3.3 years (faster than Moore's law)
  - DRAM access speed growth rate: ~1.2x every 2 years
- BGP convergence time degradation rate (estimation):

$$\frac{\text{routing table growth rate } [1.25\text{-}1.3]}{\text{DRAM access speed growth rate } [1.1]} \sim 10\% \text{ per year}$$

Note: speed limitations can absorbed using parallelism

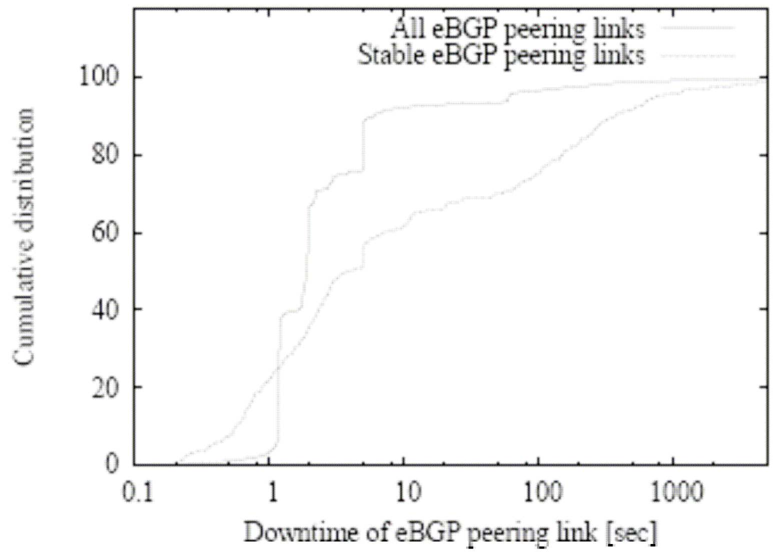
Internet routing system - BGP instability causes

BGP peering link failures

- Common events (~70% of instability) that occurs everywhere but mostly at edge networks and within ASes
- Failure duration: usually transient events with duration ~O(1s)-O(10s)
  - 82% of eBGP peering link failures last less than 180s
  - 22% of eBGP peering link failures lasted less than 1s
- Small number of links are responsible for large fraction of failures (flapping links)

BGP operational instability

Instability	Examples
BGP Session availability	Session establishment/teardown/reset
BGP Session filters	Filter and/or BGP attribute changes usually imply session (soft-)reset or graceful restart
IGP costs changes	IGP metric changes
IP address changes	Renumbering
Originator changes route	Addition/deletion of network prefixes



Source: "Achieving Sub50 Milliseconds Recovery Upon BGP Peering Link Failures", O.Bonaventure et al, ACM Co-Next 2005.

Internet routing system - BGP dynamics impact

Dynamics of routing information exchanges between routers

- Network topology updates (dynamic reaction to topological structure changes due to e.g. link/node failures)
- Routing information updates (impacts number of inter-domain routing messages that exchanged among BGP routers)

→ BGP slow convergence due to uninformed path exploration

Routing convergence: delay between an event and the instant when all routers have correctly reacted to this event

→ Trade-off

- Increase AS-path route diversity >< BGP best route selection (BGP decision process)
- Shorten adv. interval with RCN (leading to more BGP updates) to fasten convergence if dampening parameters not aggressive >< rate limit on sending routing updates (used to effectively dampening some of the oscillations inherent in vectoring approach)

(Some known) Alternatives

Solution Space

Internet evolution results in a multi-dimensional equation with multiple tradeoffs:

[**Functionality** x **Performance** x **Complexity** x **Cost**]

→ Solution Space

1. Either circumvent technological and operational limits of existing network layer in particular shortcomings of IP layer routing (in terms of scalability, stability, convergence but also sub-optimal user performance)

2. Or build an (infrastructure-based) overlay on top of existing IP network layer = add an additional layer of indirection and/or virtualization with benefits (such as customization → genericity, evolvability, & scalability ?) but also side effects

- Change properties in one or more areas of underlying network
- Horizontal and vertical cross-layer interactions (-> impact on overall network performance ?)

Overview - Routing Alternatives

BGP improvements

- **Multi-path**
- **Fast re-routing**
- As-path limit (diameter)
- Route cause notification

Beyond SPF

- **Compact routing**

  Name dependent: TZ scheme, BC scheme

  Name independent: Abraham scheme

Hybrid routing protocols

- **Combination of LS/PV: Hybrid Link-state Path-vector (HLP)**
- Combination of LS/DV: LVA

Others

- **Loc/ID separation** (host-based: SHIM6, HIP - **router-based: LISP,** GSE)
- User-controlled path routing
- Geographical routing
- Hierarchical routing

… Please do not forget the deployability requirement

BGP Improvements

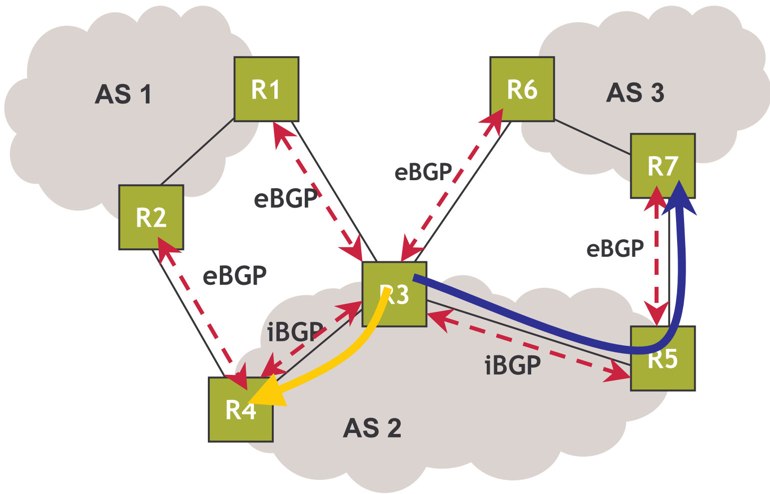
Approach: recover traffic against link failure (local) or AS-path reachability (network-wide)

Alternative_1 (reactive)

- Upon peering link failure, local recovery faster than complete BGP routing convergence
- BGP Fast Re-Route
  - BGP still advertises single best path but propagates peering link information (iBGP)
  - Multi-connected ASs -> backup link between AS pairs (=> reachability maintenance for affected prefixes)
  - In case of long period failure, deprecate the prefix reachability over failed link (instead of advertising failure)

Principles:

- BGP speaker prepared to quickly handle failure by pre-locating alternate next-hop for each BGP peering links
- When BGP peering link fails, detecting router updates its FIB to send packets to alternate next-hop (tunneling)
- Alternate next-hop then send packets to destination without using the failed link



Alternative_2 (proactive)

- BGP advertises set of alternate paths
- Solves a larger problem but requires efficient BGP route selection process
- Note: during past years, lot's of work dedicated to defection routing

(Some additional) BGP Challenges

Ultimate objective: inter-domain routing protocol that is scalable, stable (robust), fast-convergence and yet reroutes traffic extremely fast upon failure

BGP scalability → routing information aggregation

- Pro's: aggregation is beneficial for reducing BGP table size (⇒ reduce processing and hide disruption of sub-prefixes)
- Con's: however aggregation hides much topology information (granularity)

BGP scalability → routing information filtering (BGP decision process)

- Today: linear increase in terms of number of path ⇒ linear increase in number of states/updates
- Goal: super-linear increase in terms of number of path ⇒ supra-linear increase in number of states/updates

Additional constraints:

- Fast convergence: routing diversity (exploit diversity of underlying network graph) ⇒ decrease time performance on inter-domain routing system convergence
- Stability: interaction between BGP and network dynamics and how they mutually influence each other (-> robustness)

Hybrid Link-state Path-vector (HLP)

PV | HLP | LS

Policy

No policy information distribution (but inferable)
->oscillations and uninformed path exploration

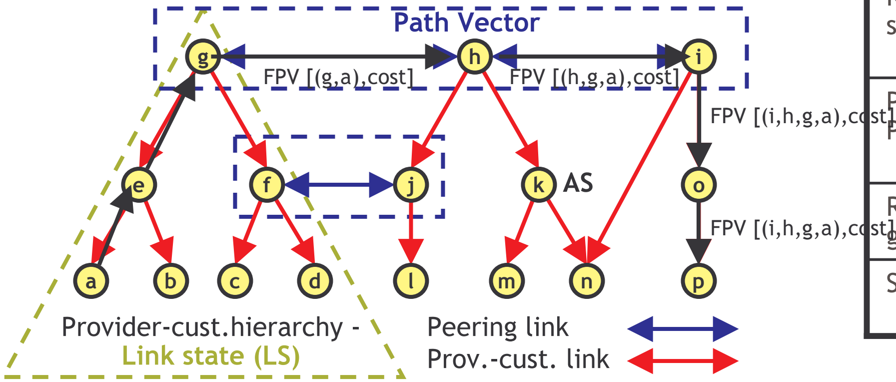
Expose common cases of policies
that BGP can not hide

Routing

Reachability and full path information distribution

Hide some routing information across hierarchy
->limit global visibility (isolation) & routing events effects

Based on hierarchical structure in AS topology, HLP combines LS routing within a provider-customer hierarchy and PV routing across peering hierarchies



Design issue	BGP	HLP
Routing structure	Flat	Hierarchical: avoids error propagation by hiding some path information using hierarchical routing structures
Policy structure Policy distr.	Generic policies No policy distr.	Optimize for common policies (export & route pref. rules) Exposure of common policies
Routing granularity	Prefix- based	AS based: each AS maintains LSDB in its local hierarchy
Style of routing	Path vector (PV)	Hybrid: LS within a given hierarchy and PV across hierarchies

- HLP performs better than BGP in isolation (number of AS's that can potentially be affected by a routing events) and churn reduction (total number of updates generated by an event)
- Convergence and security properties still require further analysis

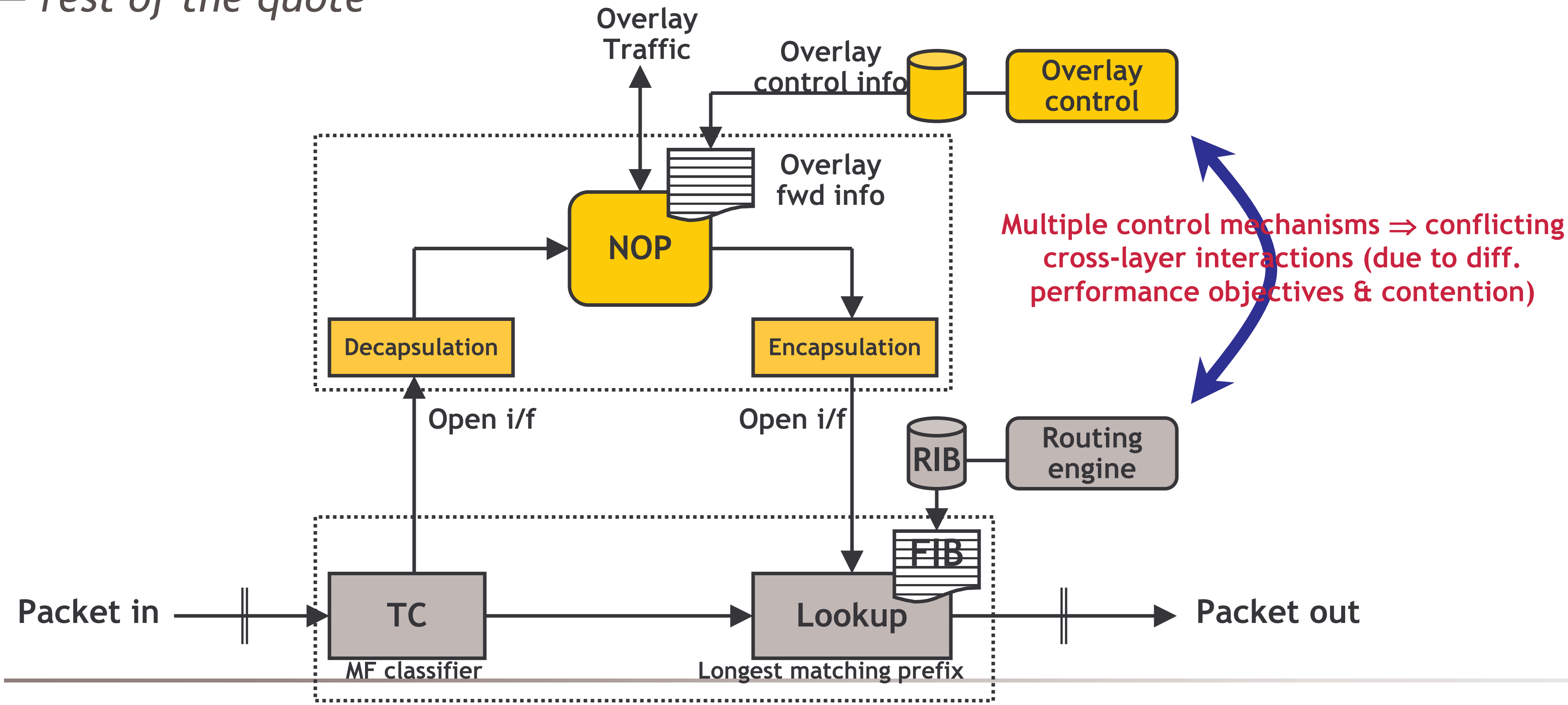
Observation …

"Any problem in computer science can be solved with another layer of indirection." -- here indirection = infrastructure-based overlay routing

– *David Wheeler*

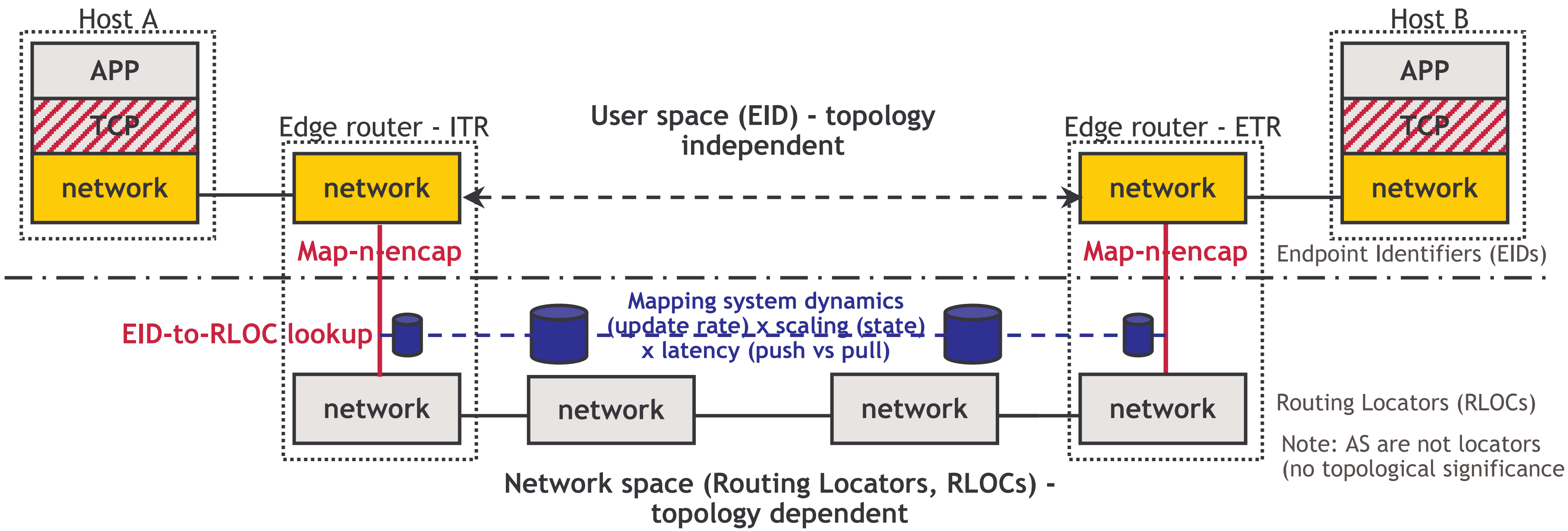
… "But that usually will create another problem."

– *rest of the quote*



Locator/Identifier Separation (Router-based: LISP)

- **Segmentation** between topology independent endpoint identifier (= user address space) and topology dependent locator (= network address space)
  + **Resolution** via distributed database incl. information necessary to translate hosts' topology independent addresses (identifiers) to topology dependent addresses (locators)



- Basic idea: **Loc/ID split** using different numbering spaces for EIDs (allocated per organization) and RLOCs (topology congruent and aggregatable)
- **LISP** = protocol implementing Loc/Id split using **map-n-encap**
  Take advantages of indirection level - Loc/Id split (-> improved routing system scalability via RLOC aggregation while minimizing core routing system changes)

Compact Routing

Stretch = ratio between length of routing path and length of shortest available path from source (s) to destination node (d) - stretch(s,d) = length(path) / dist(s,d)

Routing algorithm stretch = max.ratio over all (s,d) pairs in all graphs

→ *intuitively*: worst-case path-length increase factor relative to shortest paths

Principles

- Build routing algorithms such as, given network topology full view, trade-off between RT sizes and stretch is efficiently balanced

  → Compact routing algorithms make RT sizes compact by omitting some network topology details (in an efficient way) such that resulting path length increase stays small

Stretch	Scaling (mem. size)	Example
Stretch-1	n log n	Shortest-path first all deployed LS-, DV-, or PV-based routing protocols
Stretch-3	$n^{1/2} \log^{1/2} n$	TZ-scheme (average stretch ~ 1.1, ~70% shortest path) Topology-dependent node names and static

Stretch 3 -> need to allow for at least 3-time path length increase to route with sublinear($n^{1/2}$) routing table sizes

Assumptions

- Scale-free Internet topology -> do allow for extremely efficient **static** compact routing
- Routing to not always follow shortest paths
- **… but having <u>full view</u> of network graph (static routing)**

Forwarding vs Routing Scaling: two-dimensional nature of core scaling (1)

In large-scale packet networks: two-dimensional nature of core scaling

- if routing traffic is aggregated, then it is aggregated on the same platform that aggregates data traffic (forwarding)
- Cons.: routers must include state-of-the-art capabilities for both dimensions

⇒ System must scale in terms of capacity and throughput + routing protocol messaging and processing

How to address/reduce impact of two-dimensional nature of core scaling ?

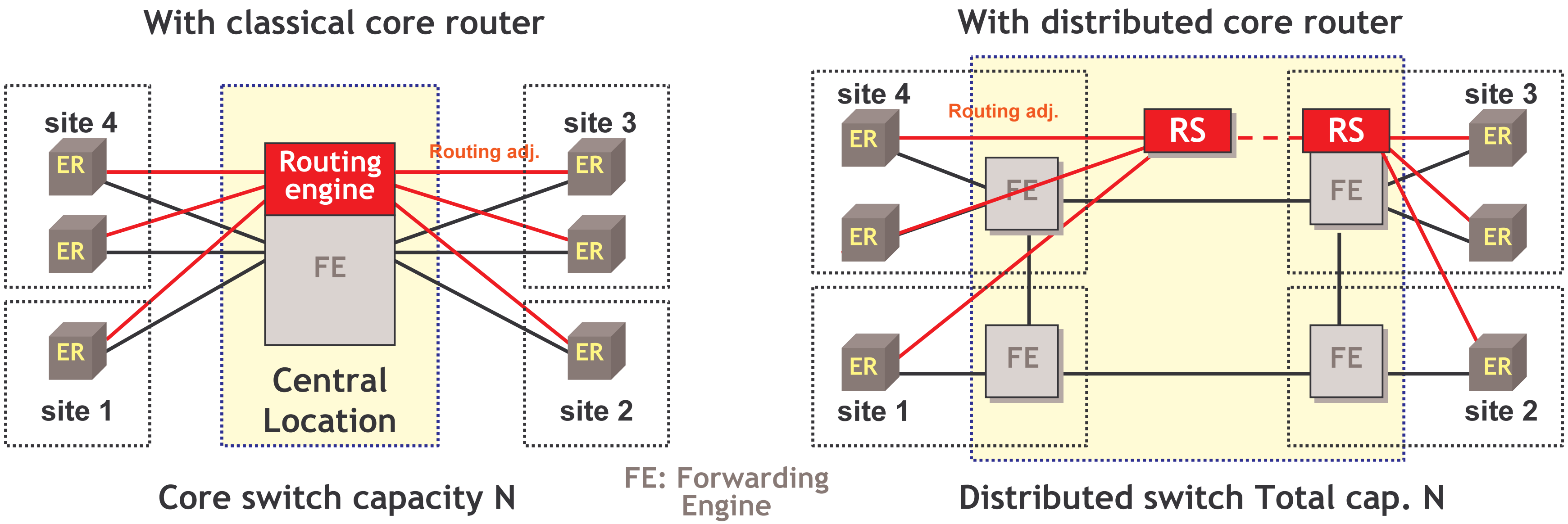
- Remove dependency to distinct expansion rates
  - Internet traffic growth: ~ 50-70% per year
  - Routing table growth: ~ 20-25% per year
- Solve aggregation problem separately with specific (rather than generalized platforms) by decoupling routing from forwarding plane aggregation
  - As traffic increase vs #routing entries
  - As number of AS increases (at periphery)
  - As paths remain sensibly identical (length)

Transit AS needs to accommodate more traffic with less increasing #edges/routes

Forwarding vs Routing Scaling: two-dimensional nature of core scaling (2)

Route server (RS) acting as routing information "re-director": routing information exchanged via established adjacencies with peering routers (routing plane level)

→ Forwarding capacity vs routing capacity differences in expansion rates in both logical and physical spaces are no longer dependent



Core routing without core router for larger scale IP networks that maintains

- Distributed traffic aggregation (no hyper-node aggregation)
- Robustness and resiliency against both node and link failure

Perspectives

Scaling dependency on Topology

Internet topological properties characterized by

- Node degree distribution: approx. long tail power law distr. $P(k) \sim k^{-\gamma}$, $\gamma = 2.254$

  ⇒ Average AS-path length ~constant (avg. 3,4) >< hierarchical routing (performs well for graphs with large distances between nodes)

- Node degree correlation: negative correlation between a node's degree k and its nearest-neighbors average degree (disassortative mixing)

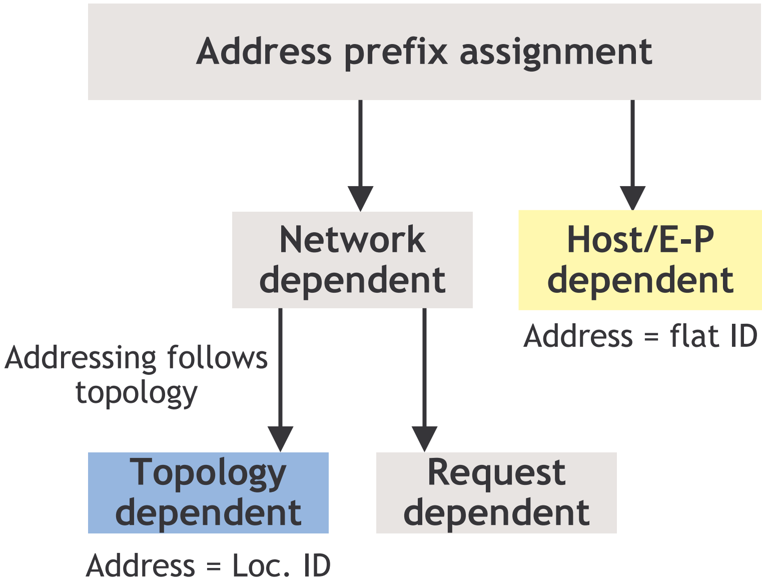
  ⇒ lower-degree nodes tend to connect with higher-degree nodes

- "Clustering": large numbers of triangular subgraphs (3-cycle) >< regular tree structures
- Rich-club connectivity: small number of nodes with a high-degree (fully interconnected -> forming a rich-club) and large number of nodes with a low-degree

Consequence: aggressive aggregation of topology-dependent locators is impossible

⇒ Routing protocols relying on aggregation can not improve RIB scaling on Internet topology

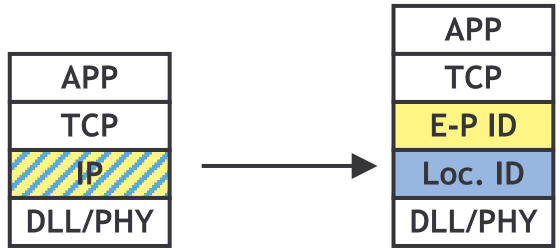
Scaling dependency on Addressing



Topology-dependent: locator address structure designed specifically to enable “topological aggregation” to scale with routing system

>< Addressing space usage as flat ID to prevent topological changes (TCP impact) and renumbering impact

⇒ routing on topology-independent end-point identifier (flat ID) that requires **some form of Loc/ID split**



Only static and topology-dependent tree-based routing exhibit logarithmic scaling on Internet topologies

Dynamic routing on topology-independent flat identifiers is a requirement on Internet topologies

⇒ routing table size cannot scale better than

Stretch	Topology dependent	Topology independent
1 | < 1,4	n log n	- | n log n
3 | < 3	$n^{1/2} \log^{1/2} n$	$n^{1/2} \log^{1/2} n$ | n

Note: same worst-case scaling of name-dependent and name-independent routing but name-independent scaling is worse on average

EU Projects - FP6 & FP7

RiNG (Routing in Next Generation networks) - FP6 CA (http://www.ist-ring.eu/)

- Coordination, study and analysis of Internet routing protocols
- Focus on new approaches to routing / changes to existing routing protocols that may support future Internet growth

  → Developing research & innovation strategies for inter-domain routing evolution

TRILOGY - FP7 IP (http://www.trilogy-project.org/)

- Redesign key elements of Internet architecture incl. inter-domain routing, locator/identifier separation and multiple path-routing
- Enhance routing infrastructure, as well as dissociate routing, TE and congestion control, to improve Internet scalability

  → Prototypes for experimental validation

ECODE - FP7 STREP - FIRE experimental research (http://www.ecode-project.eu)

- Combines networking with machine learning (semi-supervised, on-line, and distributed) to experiment cognitive routing system meeting Internet challenges
- Improve "scalability" of Internet routing system by revisiting its dynamics: e.g. enabling events detections (bogus, topological, etc.) to predict and prevent major instabilities (oscillations, uninformed path explorations) by anticipative actions

Conclusion - Network layer Routing

Difficult to predict future but… some common & base characteristics:

1. Two-part identifier

- End-point identifier e.g. crypto ID or IP address (that remains unchanged if end-host moves or is attached multi-homed to different networks)
- One or several locator identifiers e.g. IP address (that identifies attachment points to network)

2. At routing locator level

- Alt.1: BGP re-considered (is it possible ?) or new candidate such as HLP - but no improvement possible on scale of RT size from aggregation
- Alt.2: Topology-dependent **compact routing** on locators - but still lot’s of room for improvement

3. End-point ID-to-locator mapping information using (distributed) database

- Distribute entries and maintain tables for ID-to-locator name resolution
- End-point identifier ⇒ dynamically update info on where end-point ID is currently located
- Topology-dependent locators ⇒ dynamically update ID-to-locator mapping (network dynamics)

Or move directly to topology-independent compact routing (same worst case)

In any case

- Routing requires coherent full-view (network graph topology or distance to dest) & support of network dynamics ⇒ timely routing updates
- Messaging & processing cost cannot grow slower than linearly on Internet

Conclusion - Impact of Overlay Routing (on top of network layer routing)

Performing dynamic routing at both overlay and native IP layers leads to conflicting cross-layer interactions due to

- Functional overlap (unintended interactions/interferences)
- Vertical: mismatch/conflict in (re-)routing objectives
- Horizontal: contention for limited physical resources (race conditions & load oscillations)

Complex cross-layer interaction amplified by

- Selfish routing where individual user/overlay controls routing of infinitesimal amount of traffic to optimize its own performance without considering system-wide criteria
- Lack of information about other layer(s) ⇒ uninformed optimizations leading to loose-loose situation

⇒ Need to overcome degradation of overall network performance

In addition to many challenges (additional layer does not remove complexity)

- Scalability (state maintenance -> impact on reliability)
- Stability and robustness (coupling effects)
- Security

‘Pour être plus il faut s’unir,

pour s’unir il faut partager,

pour partager il faut avoir une vision.’

(Pierre Teilhard de Chardin)