Topic Preview: Routing

Marco Chiesa
KTH Royal Institute of Technology
Routing: selecting paths for network traffic
Routing: selecting paths for network traffic - forwarding devices

Routers, switches, ...
Routing: selecting paths for **network traffic**

- forwarding devices
- links

Optical fibers, copper wires
Routing: selecting paths for network traffic

- forwarding devices
- links
- end-hosts

User devices, IoT devices, ...

Datacenter servers, ...
Routing: selecting paths for network traffic
Routing: selecting paths for network traffic
Routing: selecting paths for network traffic
Routing: *selecting paths* for network traffic

how do we compute these paths?

destination 1

destination 2
Inter-domain routing: selecting paths across independent domains
Intra-domain routing: selecting paths within a single domain
Routing: **selecting paths** for network traffic

<table>
<thead>
<tr>
<th>2:10 pm - 3:50 pm</th>
<th><strong>Main-Conference Session 2: Routing</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Session Chair:</strong> Nate Foster (<em>Cornell, USA</em>)</td>
<td></td>
</tr>
<tr>
<td><strong>Location:</strong> Vigadó, 2nd-Floor Ceremonial Hall</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:10 pm - 2:35 pm</td>
<td><em>Internet Anycast: Performance, Problems and Potential</em> <em>Zhihao Li, Dave Levin, Neil Spring, Bobby Bhattacharjee (UMD, USA)</em></td>
</tr>
<tr>
<td>2:35 pm - 3:00 pm</td>
<td><em>B4 and After: Managing Hierarchy, Partitioning, and Asymmetry for Availability and Scale in Google’s Software-Defined WAN</em> <em>Chi-Yao Hong, Subhasree Mandal, Mohammad Al-Fares, Min Zhu, Richard Alimi, Kondapa Naidu B., Chandan Bhagat, Sourabh Jain, Jay Kimal, Shiyu Liang, Kirill Mendelev, Steve Padgett, Faro Rabe, Saikat Ray, Malveeka Tewari, Matt Tierney, Monika Zahn, Jonathan Zolla, Joon Ong, Amin Vahdat (Google, USA)</em></td>
</tr>
<tr>
<td>3:00 pm - 3:25 pm</td>
<td><em>On Low-Latency-Capable Topologies, and Their Impact on the Design of Intra-Domain Routing</em> <em>Nikola Gvozdiev, Stefano Vissicchio, Brad Karp, Mark Handley (UCL, UK)</em></td>
</tr>
<tr>
<td>3:25 pm - 3:50 pm</td>
<td><em>Asynchronous Convergence of Policy-Rich Distributed Bellman-Ford Routing Protocols</em> <em>Matthew L. Daggitt (Cambridge, UK), Alexander J. T. Gurney (Comcast, USA), Timothy Griffin (Cambridge, UK)</em></td>
</tr>
</tbody>
</table>

Two papers on **inter-domain** routing
Routing: selecting paths for network traffic

Two papers on intra-domain routing specifically, routing in Wide Area Networks (WANs)
Inter-domain routing: selecting paths across independent domains
Inter-domain routing: selecting paths across independent domains
Inter-domain routing: selecting paths across independent domains
The Border Gateway Protocol (BGP): a policy-based path-vector protocol
Path-vector, distance-vector: the Distributed Bellman-Ford routing family

Each node performs the following operations:

- **import**: learn routes from neighbors
- **ranking**: select a best route
- **export**: announce the best route to neighbors
Distributed Bellman-Ford: shortest-path path-vector protocol

Each node performs the following operations:

- **import**: learn routes from neighbors
  - accept all routes but filter loops
- **ranking**: select a best route
  - prefer shortest route
- **export**: export the best route
  - announce best route to everyone
Distributed Bellman-Ford: an example of shortest-path path-vector protocol
Distributed Bellman-Ford: an example of shortest-path path-vector protocol

I have a route with cost 0
Distributed Bellman-Ford: an example of shortest-path path-vector protocol
Distributed Bellman-Ford: an example of shortest-path path-vector protocol
Distributed Bellman-Ford: an example of shortest-path path-vector protocol
Distributed Bellman-Ford: an example of shortest-path path-vector protocol
Distributed Bellman-Ford: an example of shortest-path path-vector protocol
Distributed Bellman-Ford: an example of shortest-path path-vector protocol
Distributed Bellman-Ford: an example of shortest-path path-vector protocol
Distributed Bellman-Ford: an example of shortest-path path-vector protocol
Distributed Bellman-Ford: an example of shortest-path path-vector protocol
Distributed Bellman-Ford: an example of shortest-path path-vector protocol

our desired path is not a shortest-path!
The Border Gateway Protocol: a policy-based path-vector protocol

our desired path is not a shortest-path!
**BGP policy-based path-vector**

Each node performs the following operations:

- **import**: learn routes from neighbors
  - filter routes based on regular expressions (e.g. filter routes through network X)
  - filter routing loops!
- **ranking**: select a best route
  - rank routes based on BGP metrics (e.g., prefer routes through X)
  - break ties based on number of traversed domains
- **export**: export the best route
  - announce routes based on regular expressions (e.g., do not announce a route to X)
The Border Gateway Protocol: a policy-based distance-vector protocol

BGP ranking policy: I prefer to route to South America regardless of path length
The Border Gateway Protocol: a policy-based distance-vector protocol

BGP ranking policy: I prefer to route to South America regardless of path length
Distributed Bellman Ford routing: convergence vs expressiveness
Distributed Bellman Ford routing: convergence vs expressiveness

<table>
<thead>
<tr>
<th>Routing expressiveness</th>
<th>shortest-path path-vector</th>
<th>policy-based path-vector (BGP)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td>Shortest-path path-vector</td>
<td>Policy-based path-vector (BGP)</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Routing expressiveness</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>Guaranteed convergence</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>
Distributed Bellman Ford routing: convergence vs expressiveness

<table>
<thead>
<tr>
<th>Routing Expressiveness</th>
<th>Shortest-Path Path-Vector</th>
<th>Policy-Based Path-Vector (BGP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guaranteed Convergence</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>
The Border Gateway Protocol: routing inconsistencies

I prefer to route through BLUE

I prefer to route through YELLOW
Paper #4:
Asynchronous Convergence of Policy-Rich Distributed Bellman-Ford Routing Protocols

• shortest-path is not expressive for implementing economic goals…

• … but conflicting BGP policies may lead to routing instabilities

Known results for BGP instabilities:
• two stables states > risk of routing instabilities
• so-called Gao-Rexford routing policies are guaranteed to converge to a stable routing

Tradeoff between routing expressiveness and convergence:

• shortest-path is not expressive for implementing economic goals...
• ... but conflicting BGP policies may lead to routing instabilities
Paper #4:  
Asynchronous Convergence of Policy-Rich Distributed Bellman-Ford Routing Protocols

Tradeoff between routing expressiveness and convergence:
- shortest-path is not expressive for implementing economic goals...
- ... but conflicting BGP policies may lead to routing instabilities

In this paper:

“What classes of routing policies (i.e., import, ranking, and export policies) are guaranteed to converge to a stable state when messages can be lost, reordered, and indefinitely delayed?”

- Studies both distance-vector (RIP-like) and path-vector (BGP-like) routing
Recommended readings for paper #4

• L. Gao and J. Rexford. "Stable internet routing without global coordination". In Transactions on Networking 2001

• T. Griffin et al. "The stable paths problem and interdomain routing". In Transactions on Networking 2002

• T. Griffin and J. L. Sobrinho. "Metarouting". In SIGCOMM 2005

• R. Sami et al, "Searching for Stability in Interdomain Routing". In INFOCOM 2009

• M. Chiesa et al, "Using routers to build logic circuits: How powerful is BGP?". In ICNP 2013
Internet Load-balancing

My service IP = 140.0.0.1
Internet Load-balancing:
BGP determines Internet routing paths

My service IP = 140.0.0.1
Internet Load-balancing: BGP determines Internet routing paths

This service is unusable!

My service IP = 140.0.0.1
Reducing user latency:
Add service replicas closer to the users

My service
IP = 140.0.0.1
Reducing user latency: How to reach the "closest" replica?

My service
IP = 140.0.0.1
One approach is **anycast routing**: 
Announce the same IP prefix from different locations
One approach is **anycast routing**:

BGP determines the closest replica!
Notorious problems with BGP

BGP selects the best route based on:

• explicit routing policies (e.g., prefers routes through X over Y)
• number of traversed domains

BGP does not care about:

• physical properties of the route (e.g., geographical distance -> latency)

BGP latency-oblivious routing affects anycast effectiveness!

Paper #1:  
Internet Anycast: Performance, Problems, & Potential

Prior studies:

“[In anycast routing,] clients are often routed to replicas that are hundreds of kilometers away from their closest replicas”
Paper #1:
Internet Anycast: Performance, Problems, & Potential

Prior studies:

“[In anycast routing,] clients are often routed to replicas that are hundreds of kilometers away from their closest replicas”

In this paper:

1. A deep investigation of why anycast fails
2. A technique to fix anycast (spoiler: include geographical hints in BGP)
Recommended readings for paper #4

Anycast routing:


Demand-aware BGP improvements:

• K. Yap et al. "Taking the Edge off with Espresso: Scale, Reliability and Programmability for Global Internet Peering". In SIGCOMM 2017

• B. Schlinker et al. "Engineering Egress with Edge Fabric". In SIGCOMM 2017
**Intra-domain routing:** selecting paths **within a single domain**
Intra-domain routing: selecting paths within a single domain
Intra-domain routing: selecting paths within a single domain

what are the best paths?
Intra-domain routing: selecting paths within a single domain

Objectives, e.g.,:
- min load on links
- min latency
Intra-domain routing: 
**selecting paths within a single domain**

Objectives, e.g.,:
- min load on links
- min latency

Constraints, e.g.,:
- routing expressiveness
Intra-domain routing: selecting paths within a single domain

Objectives, e.g.,:
- min load on links
- min latency

Constraints, e.g.,:
- routing expressiveness

Uncertainty, e.g.,:
- node/link failures
- traffic demands
Intra-domain routing: selecting paths within a single domain

Objectives, e.g.,:
- min load on links
- min latency

Constraints, e.g.,:
- routing expressiveness

Uncertainty, e.g.,:
- node/link failures
- traffic demands

all internal link capacities are 1
Intra-domain routing: selecting paths within a single domain

Objectives, e.g.,:
- min load on links
- min latency

Constraints, e.g.,:
- routing expressiveness

Uncertainty, e.g.,:
- node/link failures
- traffic demands

all internal link capacities are 1 unit
Intra-domain routing:
selecting paths within a single domain

Objectives, e.g.,:
- min load on links
- min latency

Constraints, e.g.,:
- routing expressiveness

Uncertainty, e.g.,:
- node/link failures
- traffic demands

all internal link capacities are 1
Intra-domain routing: selecting paths within a single domain

Objectives, e.g.,:
- min load on links
- min latency

Constraints, e.g.,:
- routing expressiveness

Uncertainty, e.g.,:
- node/link failures
- traffic demands

2 units
all internal link capacities are 1
Intra-domain routing: selecting paths within a single domain

Objectives, e.g.,:
- min load on links
- min latency

Constraints, e.g.,:
- routing expressiveness

Uncertainty, e.g.,:
- node/link failures
- traffic demands

all internal link capacities are 1

high latency
Intra-domain routing: selecting paths within a single domain

Objectives, e.g.,:
- min load on links
- min latency

Constraints, e.g.,:
- routing expressiveness

Uncertainty, e.g.,:
- node/link failures
- traffic demands

2 units

all internal link capacities are 1
Intra-domain routing: selecting paths within a single domain

Objectives, e.g.,:
- min load on links
- min latency

Constraints, e.g.,:
- routing expressiveness

Uncertainty, e.g.,:
- node/link failures
- traffic demands

all internal link capacities are 1
Paper #3:
On low-latency-capable topologies, and their impact on the design of intra-domain routing

Goal: understanding the interplay between network topology and latency
Paper #3: 
On low-latency-capable topologies, and their impact on the design of intra-domain routing

Goal: understanding the interplay between network topology and latency

Fundamental questions investigated in this paper:

1) ”Are there topologies that are more suitable to accommodate latency-sensitive, dynamic traffic demands?”

2) ”What type of routing schemes perform well on such topologies?”
Paper #3: On low-latency-capable topologies, and their impact on the design of intra-domain routing

Goal: understanding the interplay between network topology and latency

Fundamental questions investigated in this paper:

1) “Are there topologies that are more suitable to accommodate latency-sensitive, dynamic traffic demands?”
2) “What type of routing schemes perform well on such topologies?”

State of the art improvements:

• outperforms existing routing schemes on achieving low latency traffic delivery
Paper #2:
B4 and After: Managing Hierarchy, Partitioning, and Asymmetry for Availability and Scale in Google’s Software-Defined WAN

A unique look into Google’s SDN Wide Area Network Routing

Main routing challenges:
• performance
• scalability
• availability
Recommended readings for paper #2 and #3

• Wide Area Network Traffic-Engineering:
  • C. Hong et al. "Achieving high utilization with software-driven WAN”. In SIGCOMM 2013
  • S. Jain et al. "B4: experience with a globally-deployed software defined wan". In SIGCOMM 2013
  • C. Hong et al. "B4 and After: Managing Hierarchy, Partitioning, and Asymmetry for Availability and Scale in Google’s SD-WAN”. In SIGCOMM 2018

• Traffic oblivious Routing:
  • H. Räcke ”Optimal hierarchical decompositions for congestion minimization in networks”. In STOC 2008
  • D. Applegate, E. Cohen ”Making intra-domain routing robust to changing and uncertain traffic demands: understanding fundamental tradeoffs”. In SIGCOMM 2003
  • M. Chiesa et al. ”Oblivious Routing in IP Networks”. In Transactions on Networking 2018

• Semi-oblivious routing:
  • M. Hajiaghayi et al, "Semi-oblivious routing: lower bounds". In SODA 2007
  • P. Kumar et al. ”Semi-Oblivious Traffic Engineering: The Road Not Taken”. In NSDI 2018
Recommended readings for paper #2 and #3

Scalability of the control-plane:
• T. Koponen et al. "Onix: A Distributed Control Platform for Large-scale Production Networks". In OSDI 2010

Distributed routing:
• R. Gallager ”A Minimum Delay Routing Algorithm Using Distributed Computation”. In Transactions on Communications 1977

Hash-based forwarding:
• Z. Cao et al. "Performance of Hashing-Based Schemes for Internet Load Balancing". In INFOCOM 2000
<table>
<thead>
<tr>
<th>Time</th>
<th>Session Title</th>
<th>Speakers</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:10 pm - 2:35 pm</td>
<td>Internet Anycast: Performance, Problems and Potential</td>
<td>Zhihao Li, Dave Levin, Neil Spring, Bobby Bhattacharjee (UMD, USA)</td>
<td>2nd-Floor Ceremonial Hall</td>
</tr>
<tr>
<td>2:35 pm - 3:00 pm</td>
<td>B4 and After: Managing Hierarchy, Partitioning, and Asymmetry for Availability and Scale in Google’s Software-Defined WAN</td>
<td>Chi-Yao Hong, Subhasree Mandal, Mohammad Al-Fares, Min Zhu, Richard Alimi, Kondapa Naidu B., Chandan Bhagat, Sourabh Jain, Jay Kaimal, Shiyu Liang, Kirill Mendelev, Steve Padgett, Faro Rabe, Saikat Ray, Malveeka Tewari, Matt Tierney, Monika Zahn, Jonathan Zolla, Joon Ong, Amin Vahdat (Google, USA)</td>
<td>2nd-Floor Ceremonial Hall</td>
</tr>
<tr>
<td>3:00 pm - 3:25 pm</td>
<td>On Low-Latency-Capable Topologies, and Their Impact on the Design of Intra-Domain Routing</td>
<td>Nikola Gvozdiev, Stefano Vissicchio, Brad Karp, Mark Handley (UCL, UK)</td>
<td>2nd-Floor Ceremonial Hall</td>
</tr>
<tr>
<td>3:25 pm - 3:50 pm</td>
<td>Asynchronous Convergence of Policy-Rich Distributed Bellman-Ford Routing Protocols</td>
<td>Matthew L. Daggitt (Cambridge, UK), Alexander J. T. Gurney (Comcast, USA), Timothy Griffin (Cambridge, UK)</td>
<td>2nd-Floor Ceremonial Hall</td>
</tr>
</tbody>
</table>