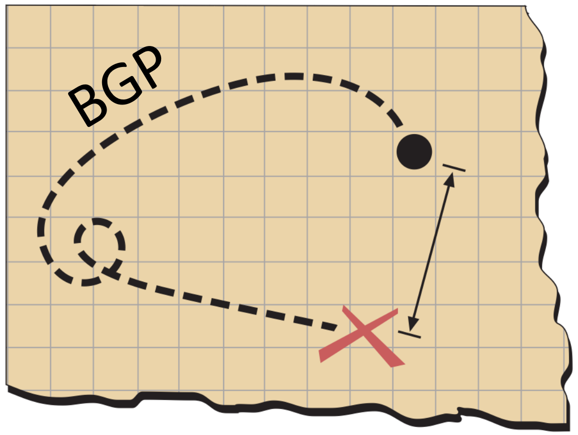
Topic Preview: **Routing**

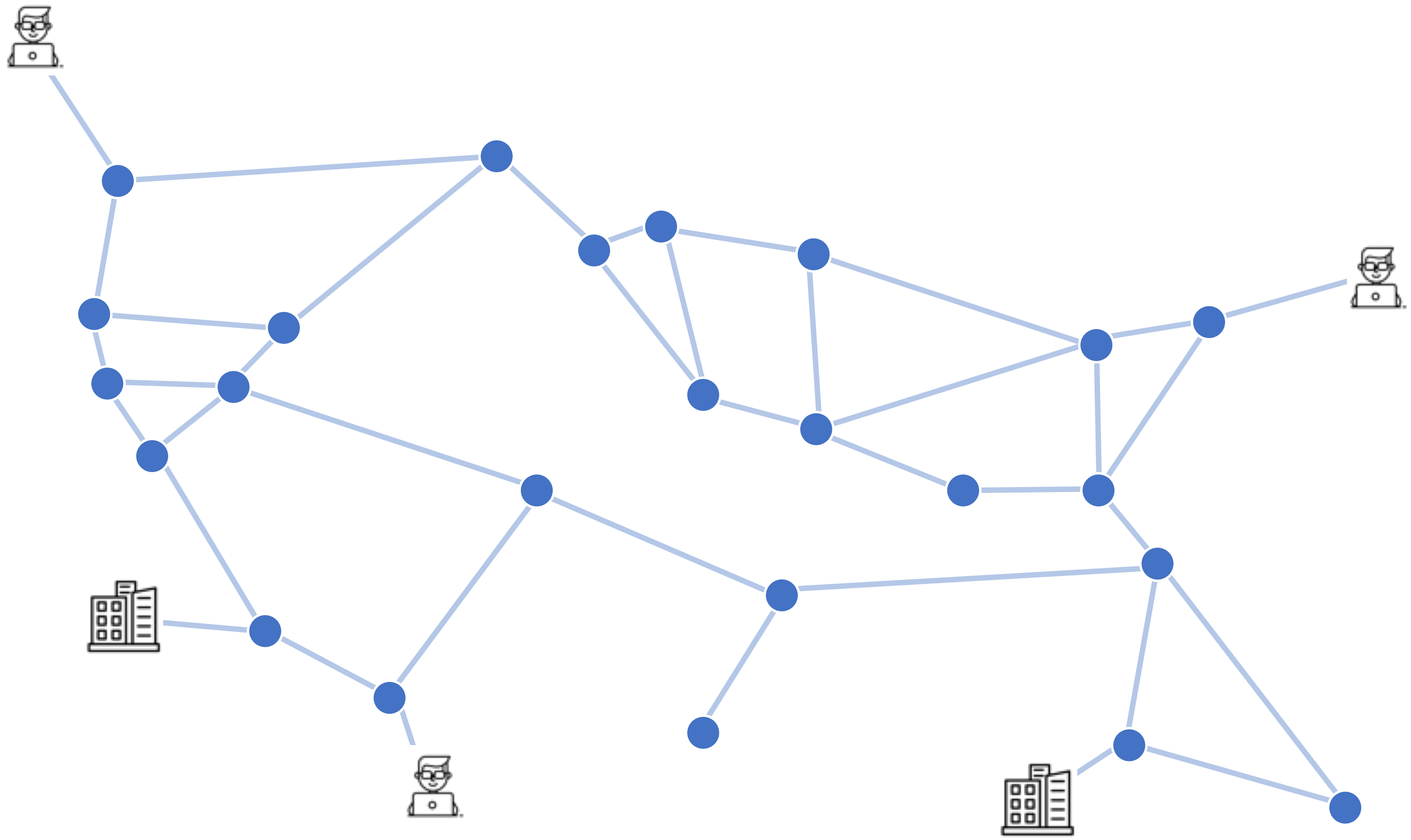


Marco Chiesa

KTH Royal Institute of Technology

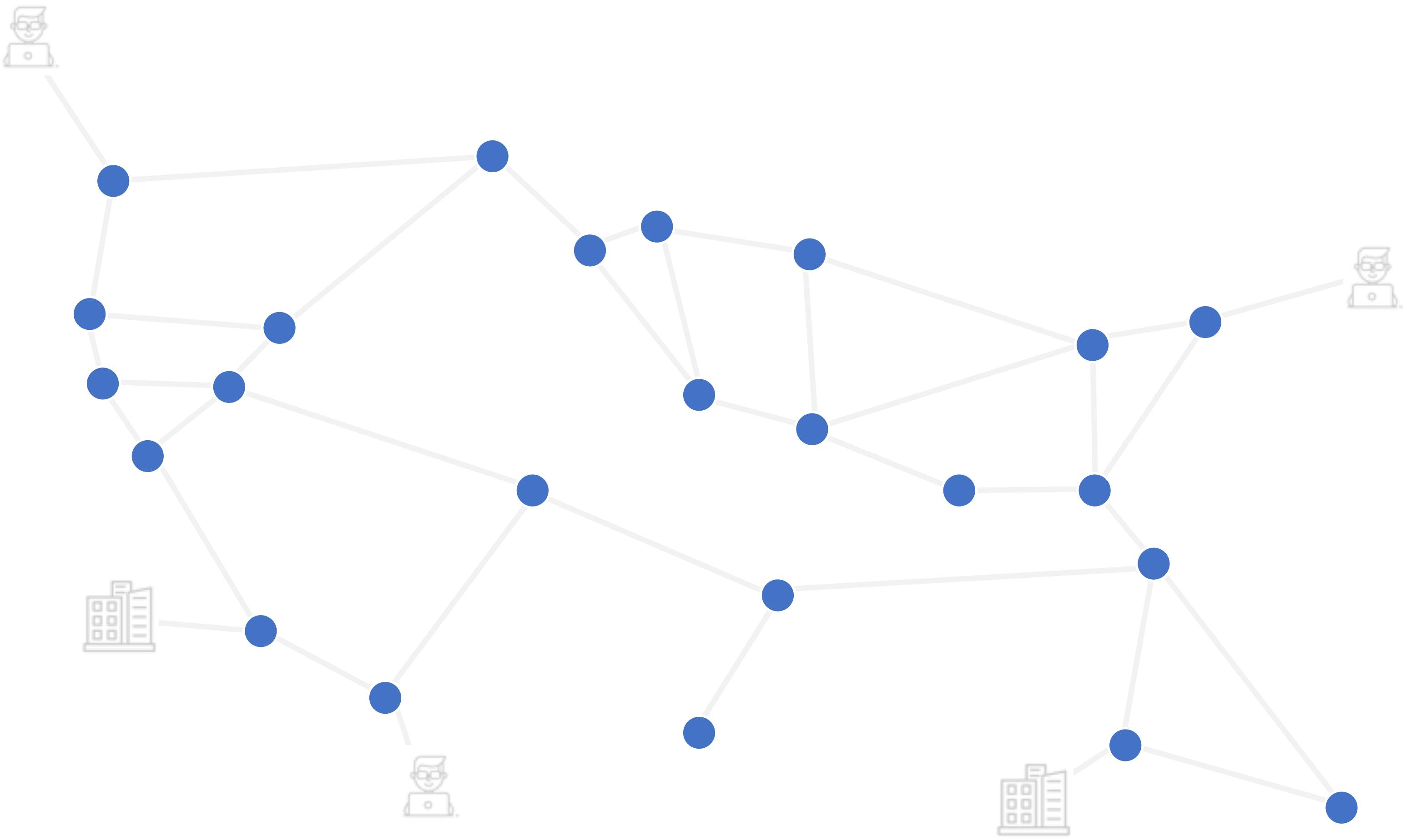
Image credits: The Noun Project: Switch By Opher Aloni, Detour 2 By Pierre-Luc Auclair, CA, Right arrow By heri sugianto, ID, Fast clock By Piero Borgo, IT, Small table By IconMark, PH, Brain By Symbolon, IT Right arrow 2 By Roy, NL, Redo By Bluetip Design, US, World Map by Avnihan Kırışık, datacenter by ProSymbols, sensor by Tami Nova, treasures by Vladimir Belochkin, Pirate by b farias, Cross by Xela Ub, Computer by Ralf Schmitzer, Letter by Daria Moskvina, Time by Xiaoying Huang

Routing: selecting paths for network traffic



Routing: selecting paths for network traffic

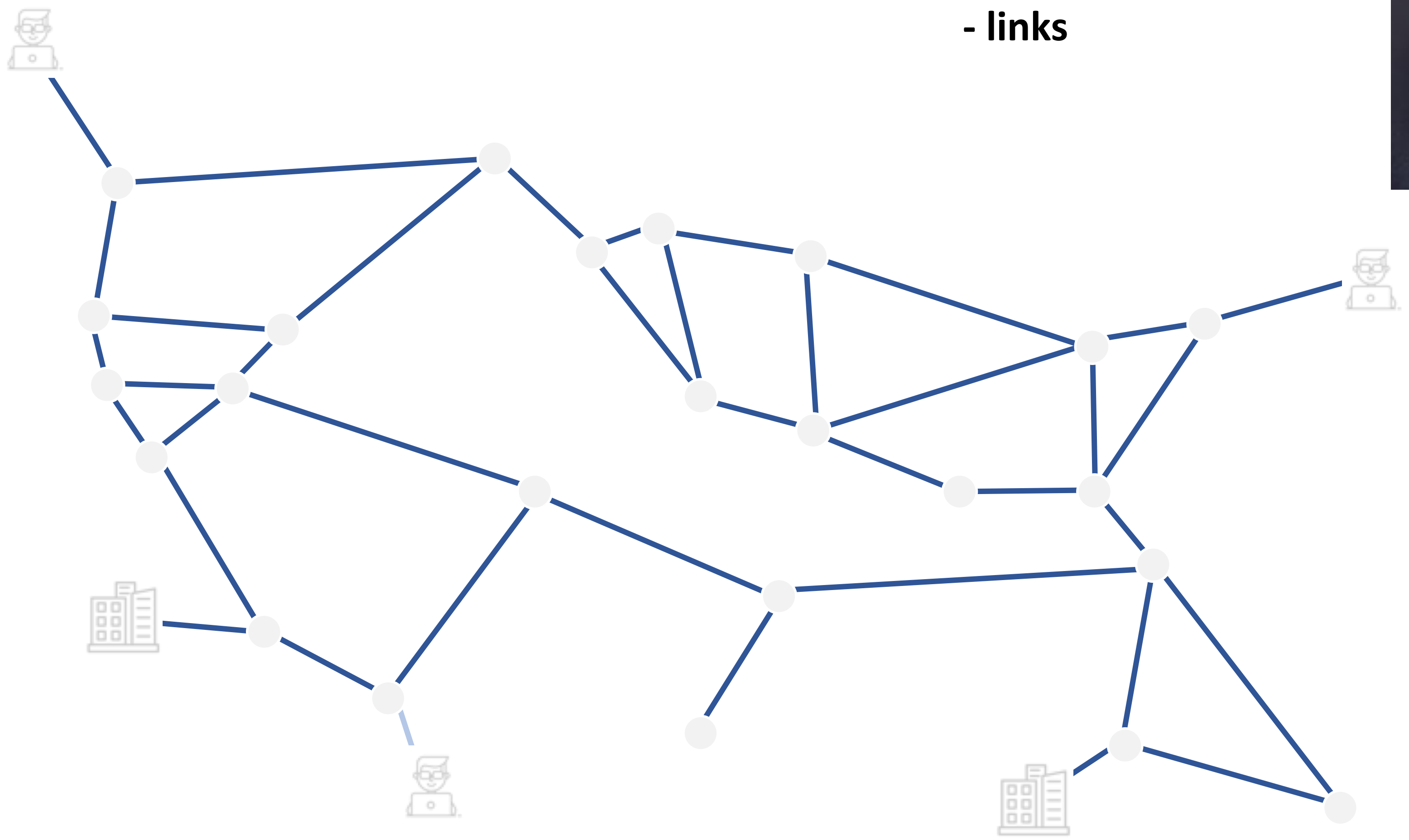
- forwarding devices



Routers, switches, …

Routing: selecting paths for network traffic

- forwarding devices
- **links**



Optical fibers, copper wires

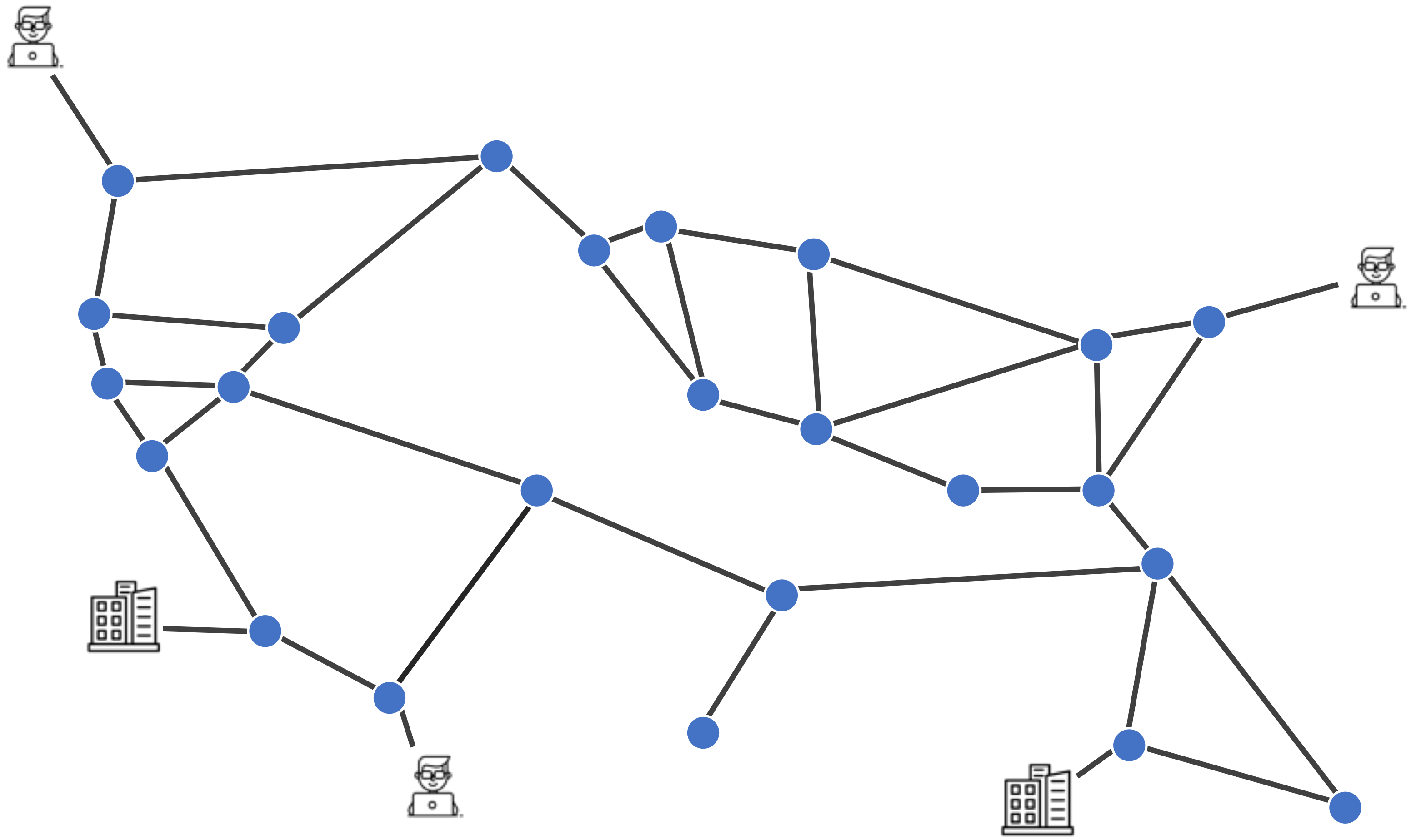
Routing: selecting paths for network traffic

- forwarding devices
- links
- **end-hosts**

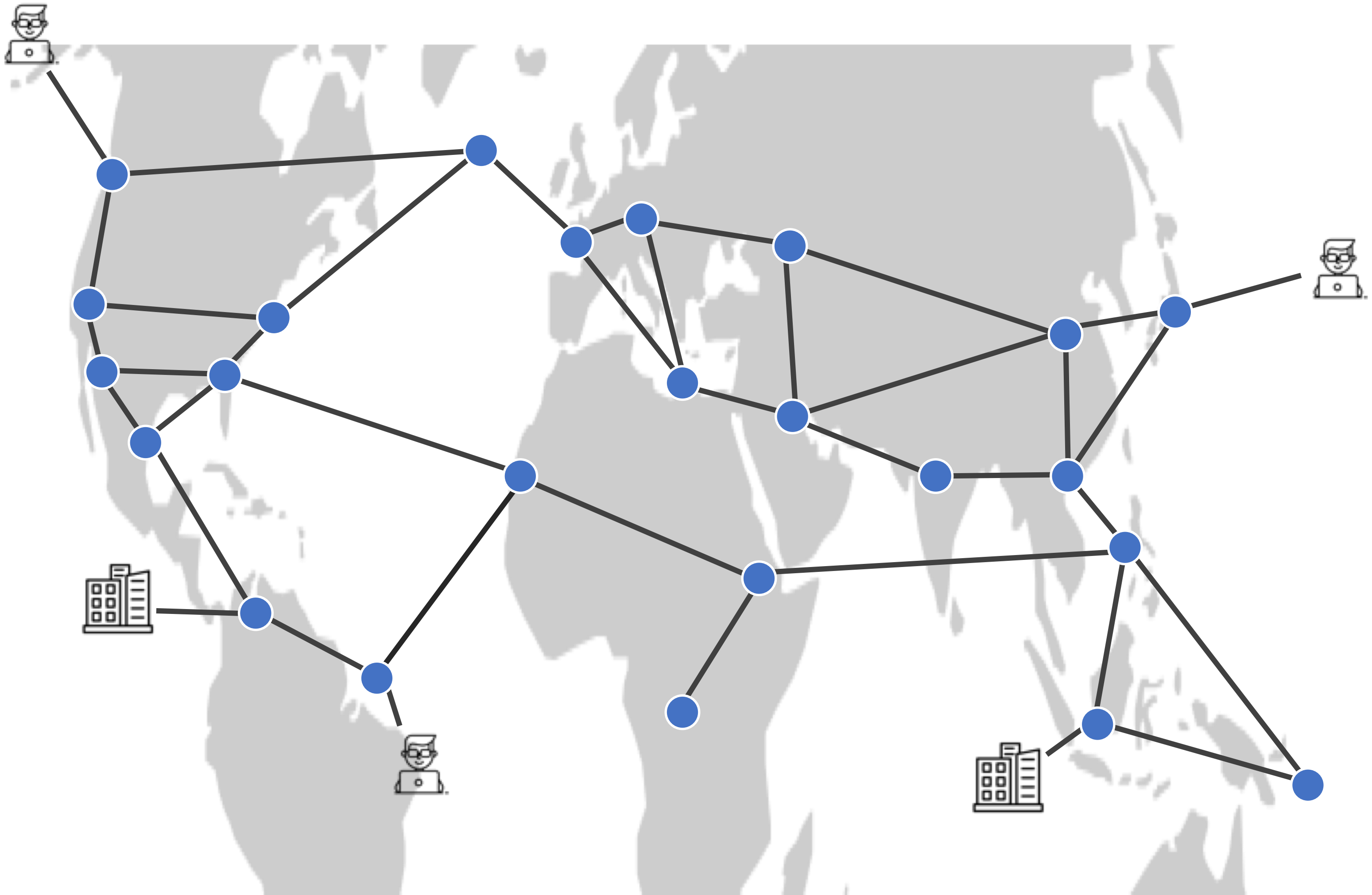
User devices,
IoT devices, …

Datacenter
servers, …

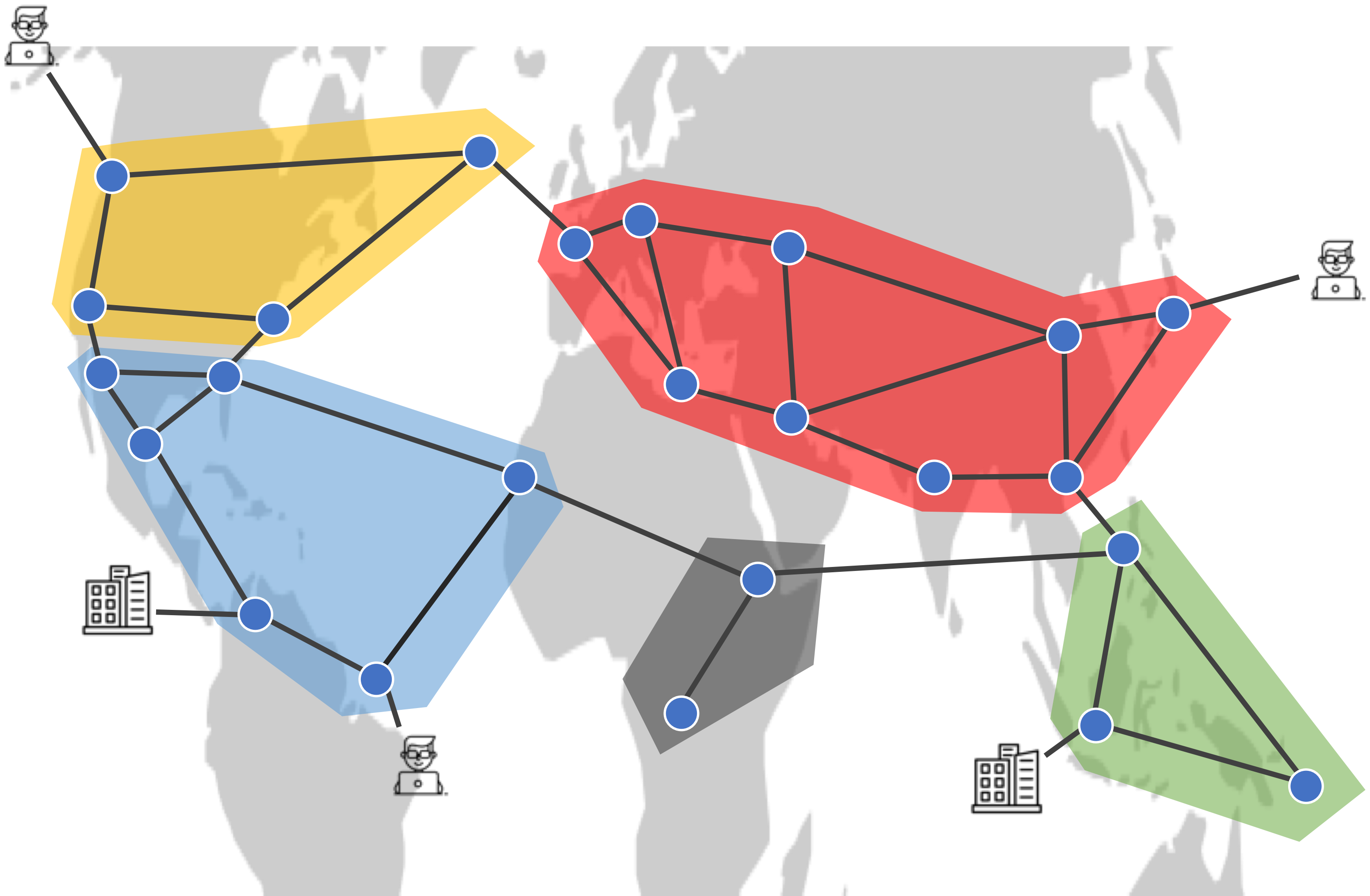
Routing: selecting paths for network traffic



Routing: selecting paths for network traffic

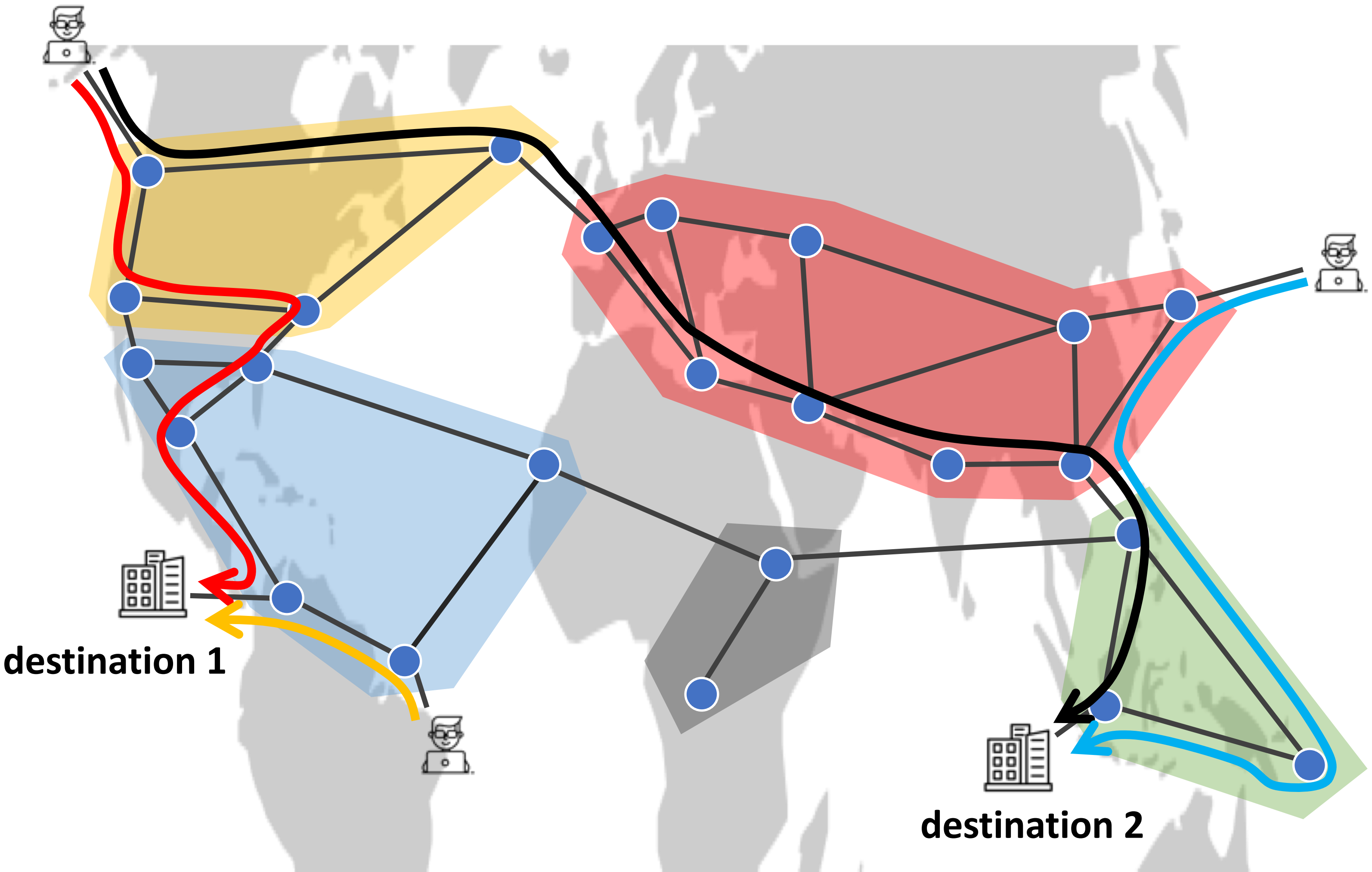


Routing: selecting paths for network traffic

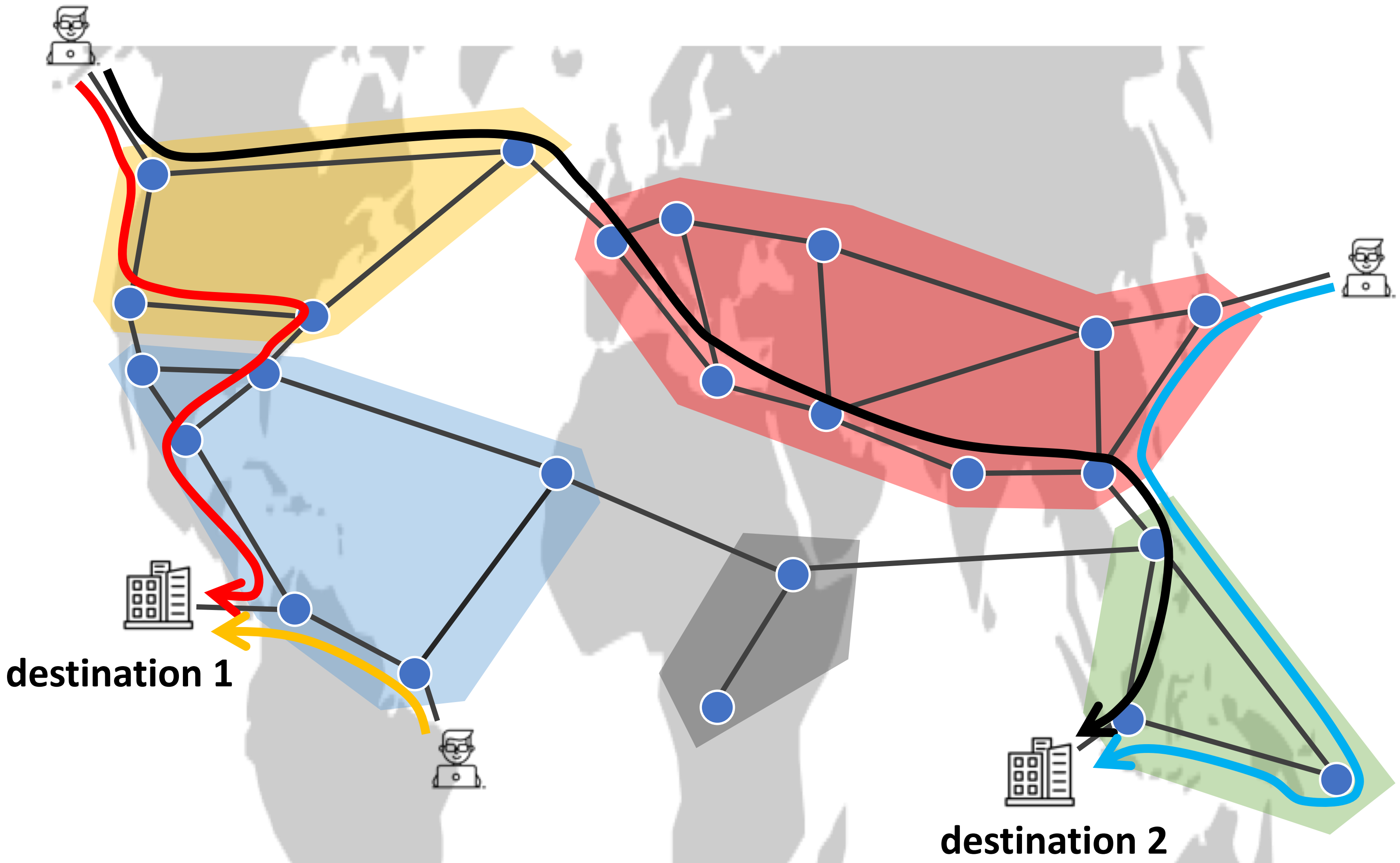


Routing: selecting paths for network traffic

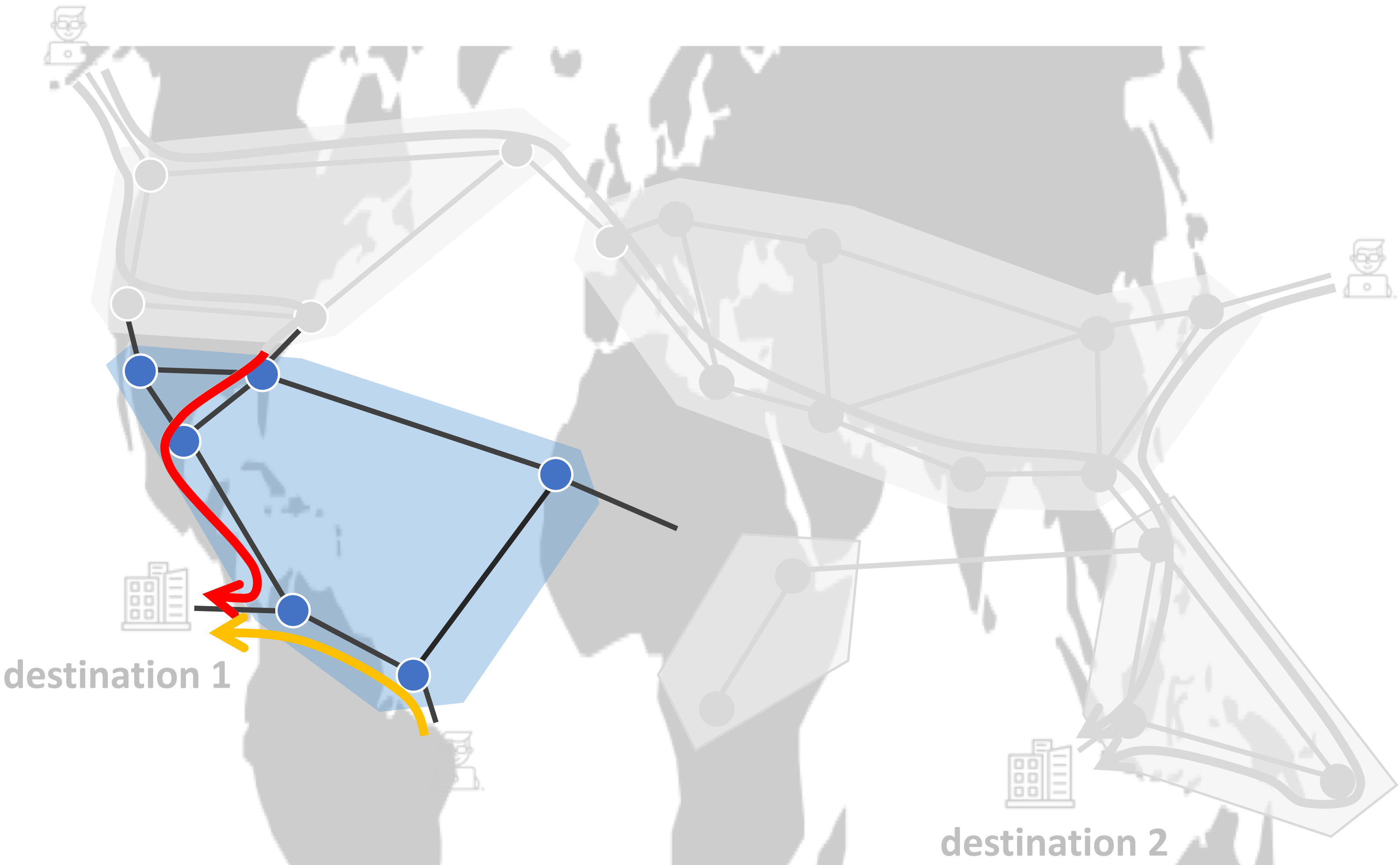
how do we compute these paths?



Inter-domain routing: selecting paths across independent domains



Intra-domain routing: selecting paths within a single domain



Routing: selecting paths for network traffic

2:10 pm - 3:50 pm Main-Conference Session 2: Routing

Session Chair: Nate Foster *(Cornell, USA)*
Location: Vigadó, 2nd-Floor Ceremonial Hall

2:10 pm - 2:35 pm	**Internet Anycast: Performance, Problems and Potential** Zhihao Li, Dave Levin, Neil Spring, Bobby Bhattacharjee *(UMD, USA)*
2:35 pm - 3:00 pm	**B4 and After: Managing Hierarchy, Partitioning, and Asymmetry for Availability and Scale in Google's Software-Defined WAN** 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)*
3:00 pm - 3:25 pm	**On Low-Latency-Capable Topologies, and Their Impact on the Design of Intra-Domain Routing** Nikola Gvozdiev, Stefano Vissicchio, Brad Karp, Mark Handley *(UCL, UK)*
3:25 pm - 3:50 pm	**Asynchronous Convergence of Policy-Rich Distributed Bellman-Ford Routing Protocols** Matthew L. Daggitt *(Cambridge, UK)*, Alexander J. T. Gurney *(Comcast, USA)*, Timothy Griffin *(Cambridge, UK)*

Two papers on **inter-domain** routing

Routing: selecting paths for network traffic

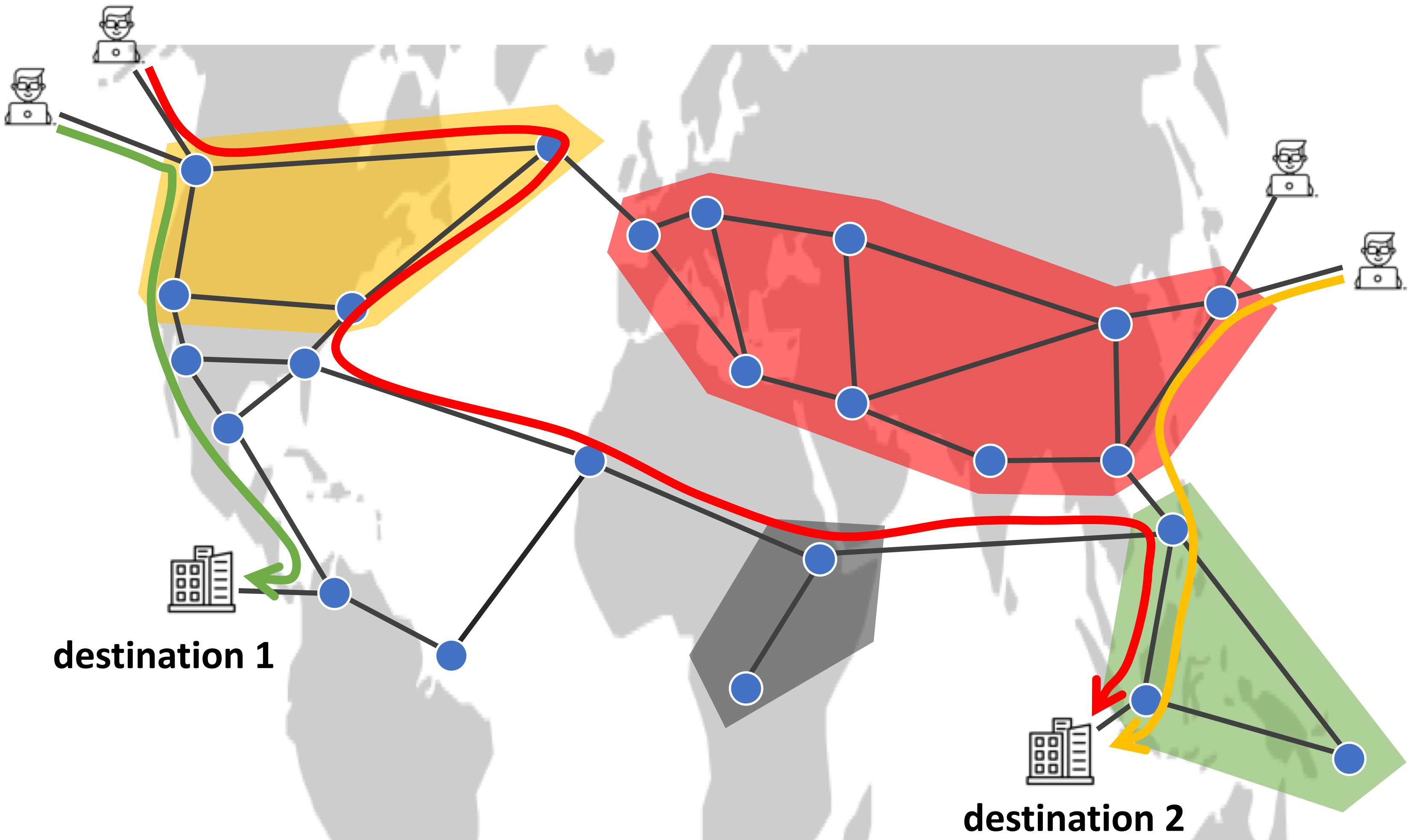
2:10 pm - 3:50 pm Main-Conference Session 2: Routing

Session Chair: Nate Foster *(Cornell, USA)*
Location: Vigadó, 2nd-Floor Ceremonial Hall

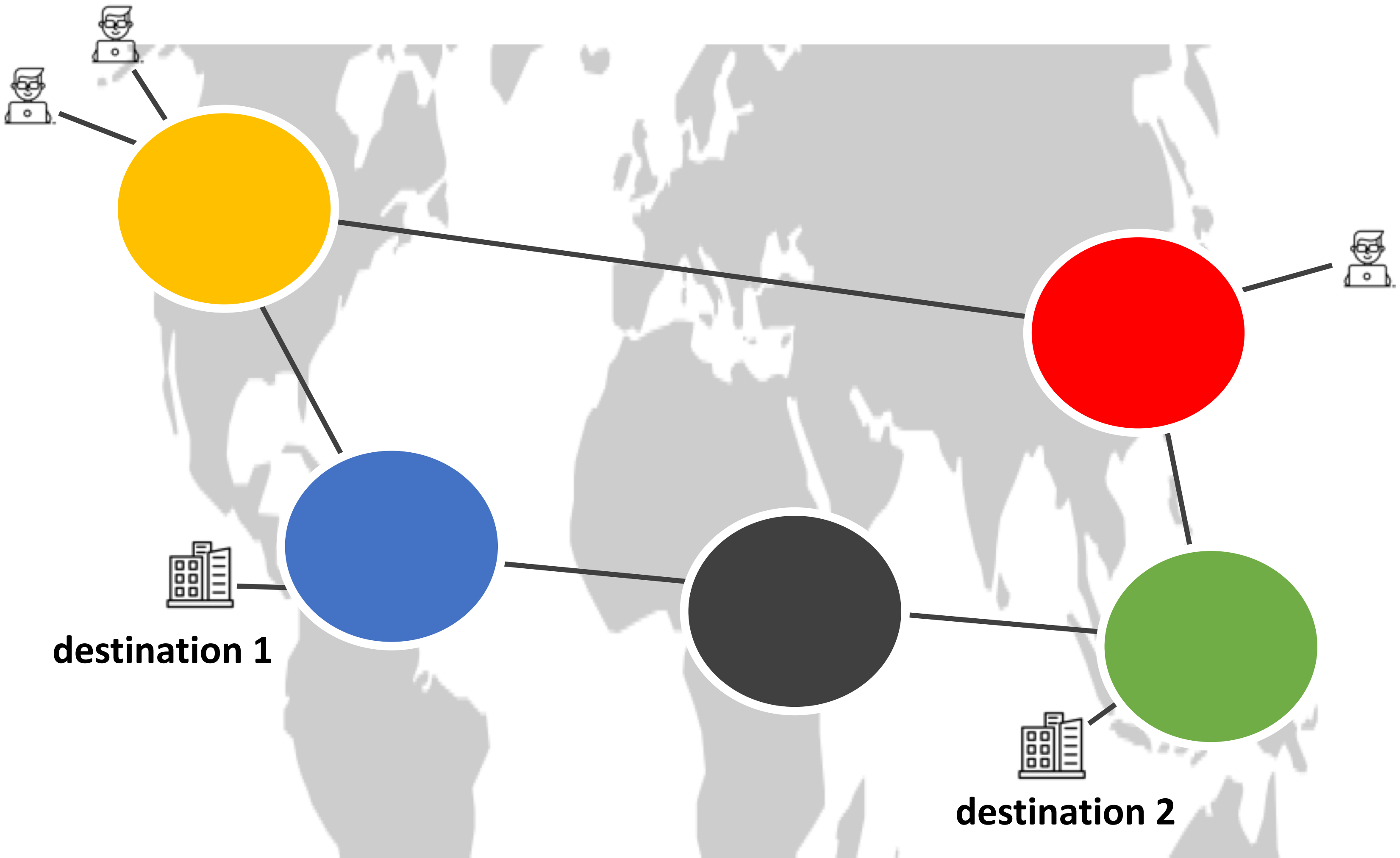
2:10 pm - 2:35 pm	**Internet Anycast: Performance, Problems and Potential** Zhihao Li, Dave Levin, Neil Spring, Bobby Bhattacharjee *(UMD, USA)*
2:35 pm - 3:00 pm	**B4 and After: Managing Hierarchy, Partitioning, and Asymmetry for Availability and Scale in Google's Software-Defined WAN** Chi-Yao Hong, 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)*
3:00 pm - 3:25 pm	**On Low-Latency-Capable Topologies, and Their Impact on the Design of Intra-Domain Routing** Nikola Gvozdiev, Stefano Vissicchio, Brad Karp, Mark Handley *(UCL, UK)*
3:25 pm - 3:50 pm	**Asynchronous Convergence of Policy-Rich Distributed Bellman-Ford Routing Protocols** Matthew L. Daggitt *(Cambridge, UK)*, Alexander J. T. Gurney *(Comcast, USA)*, Timothy Griffin *(Cambridge, UK)*

Two papers on **intra-domain** routing
specifically, routing in **Wide Area Networks (WANs)**

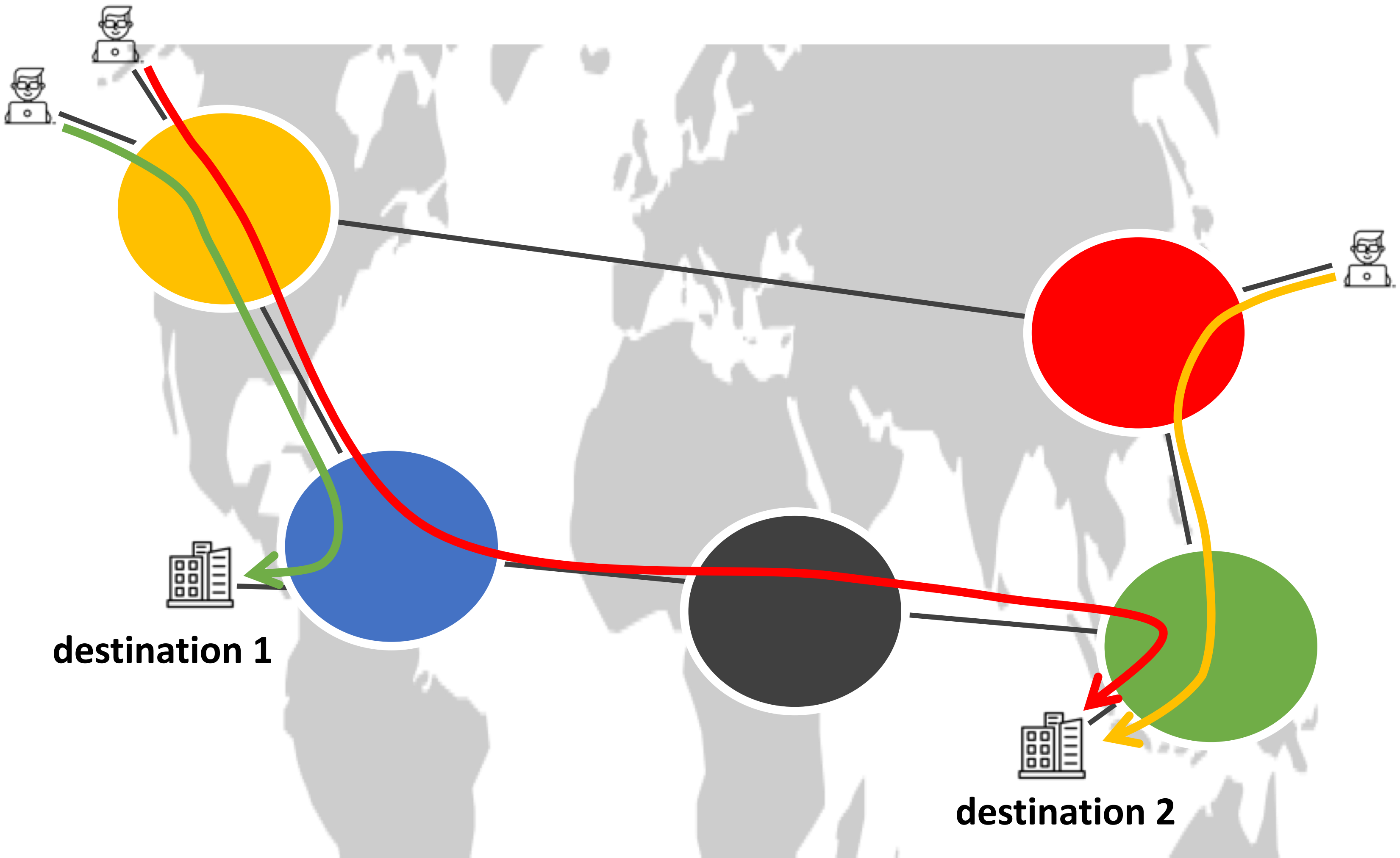
Inter-domain routing:
selecting paths across independent domains



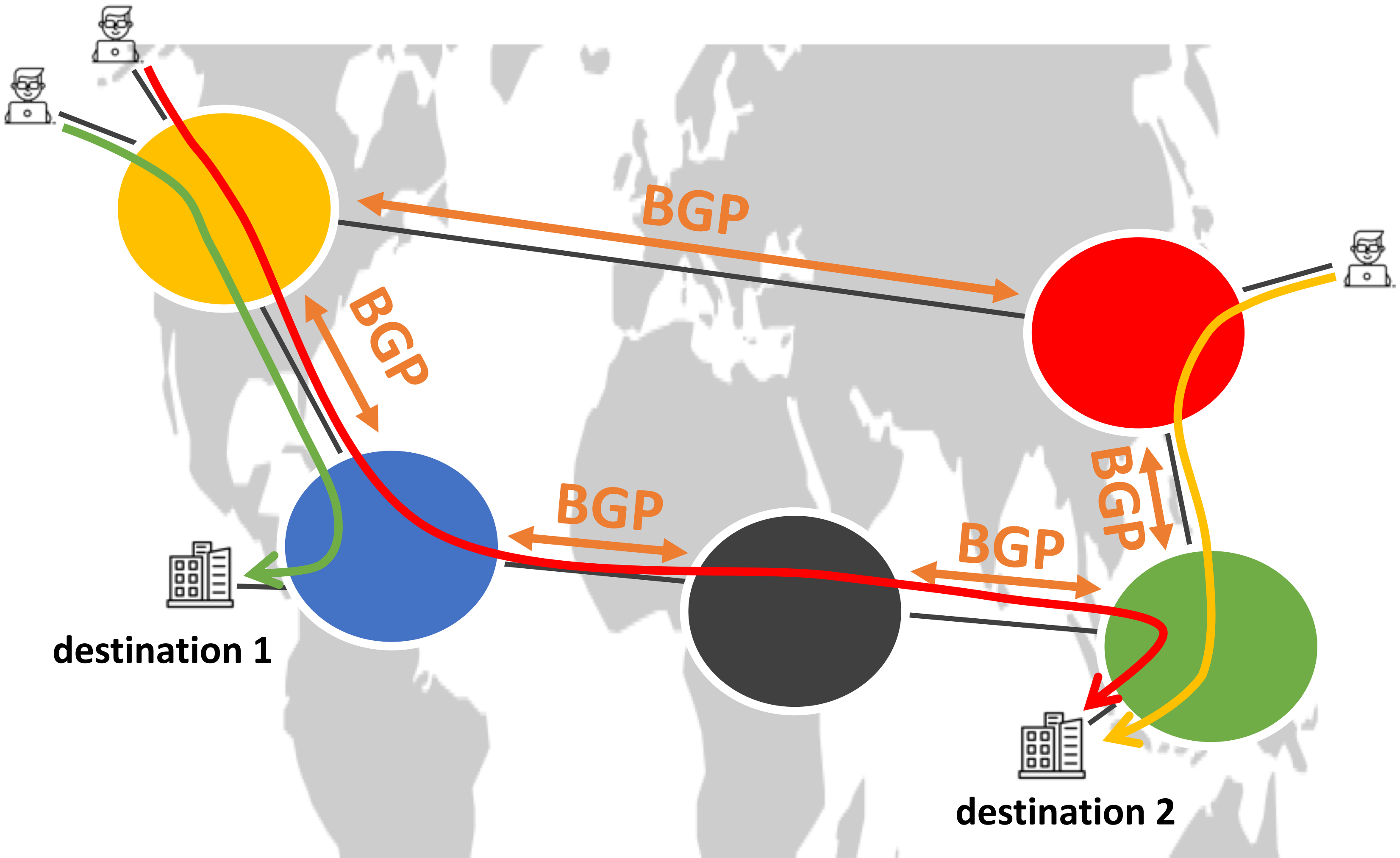
Inter-domain routing: selecting paths across independent domains



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:

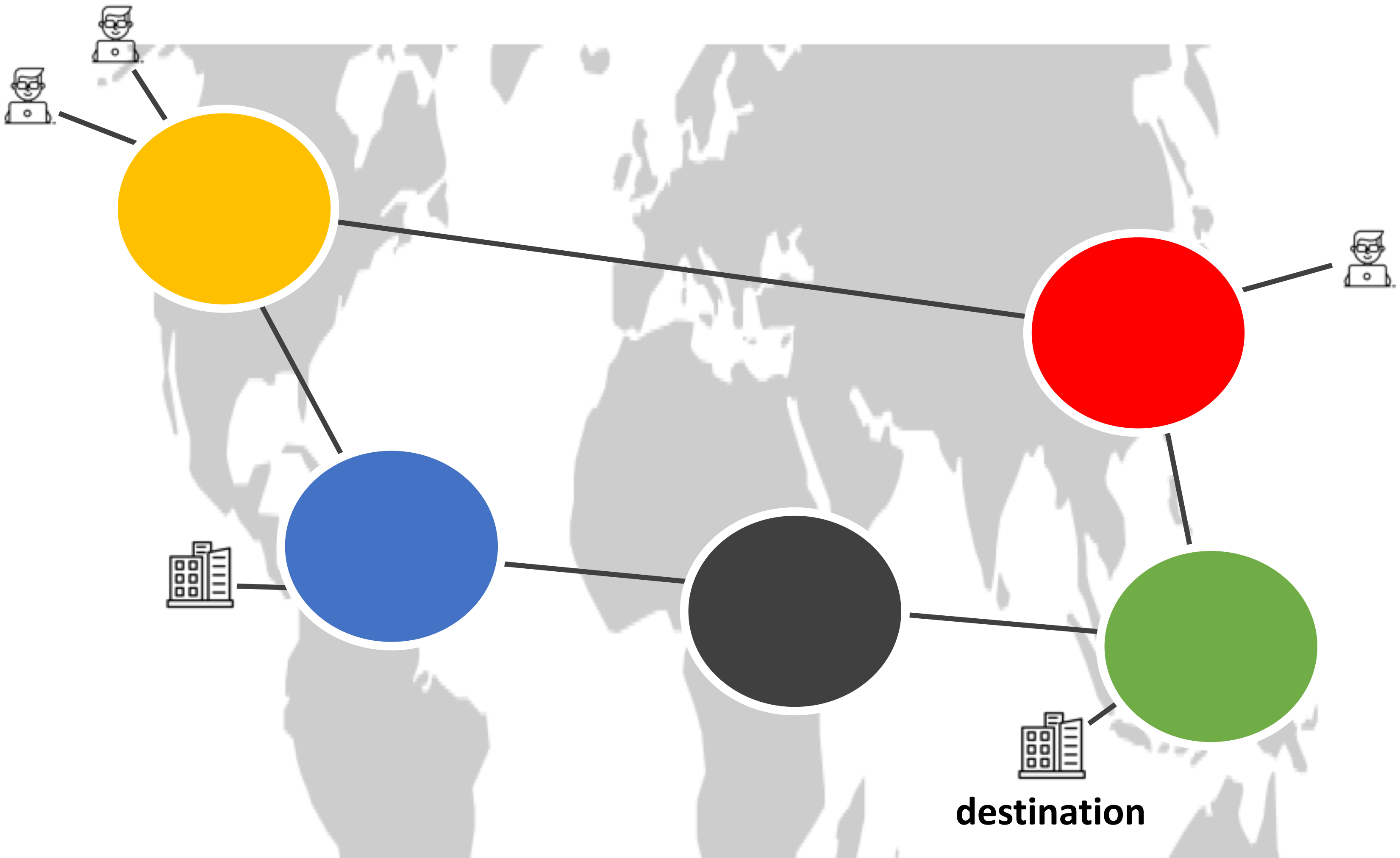
- import: learn routes from neighbors
- ranking: select a best route
- export: announce the best route to neighbors

Distributed Bellman-Ford: shortest-path path-vector protocol

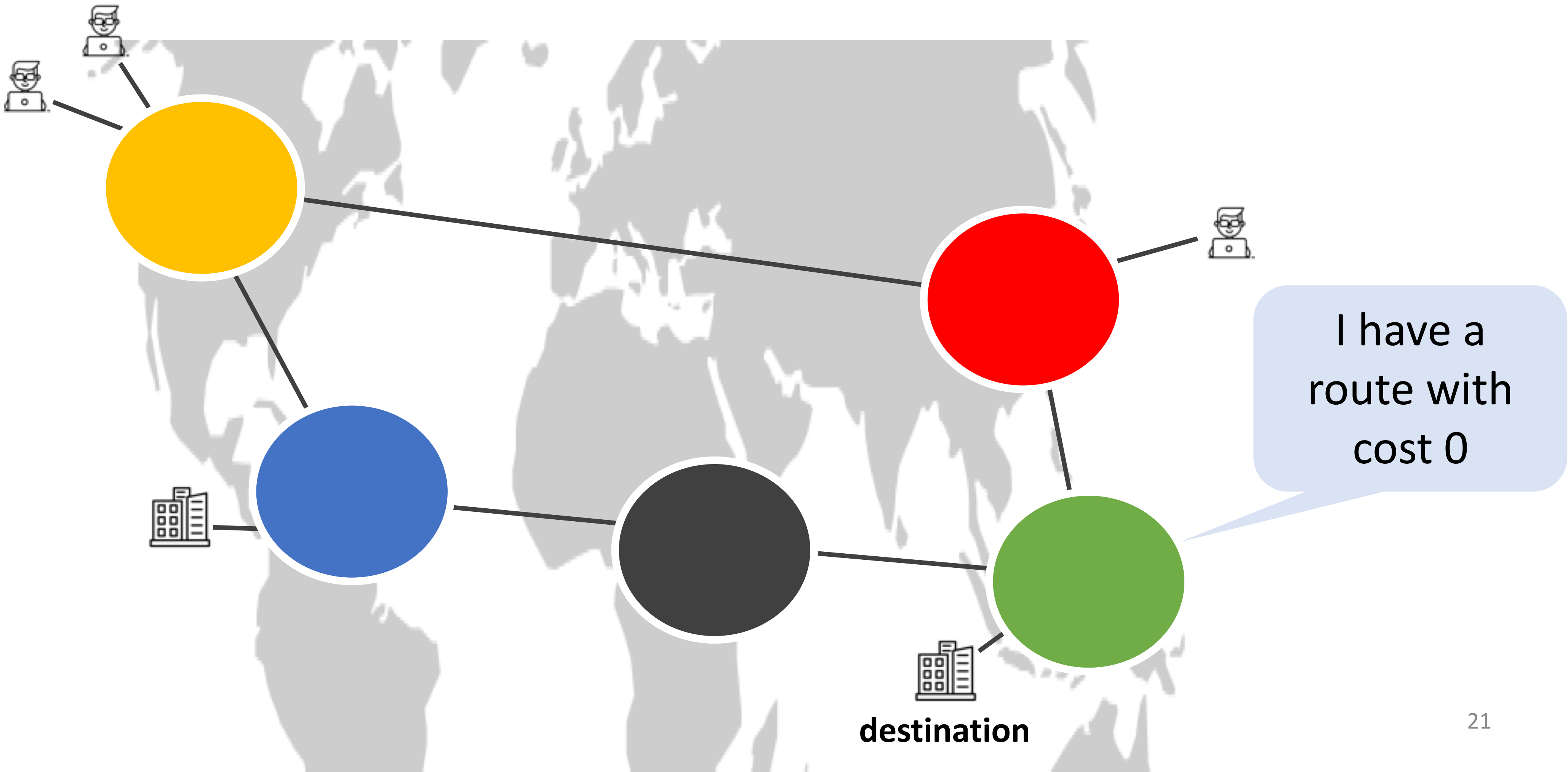
Each node performs the following operations:

- **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

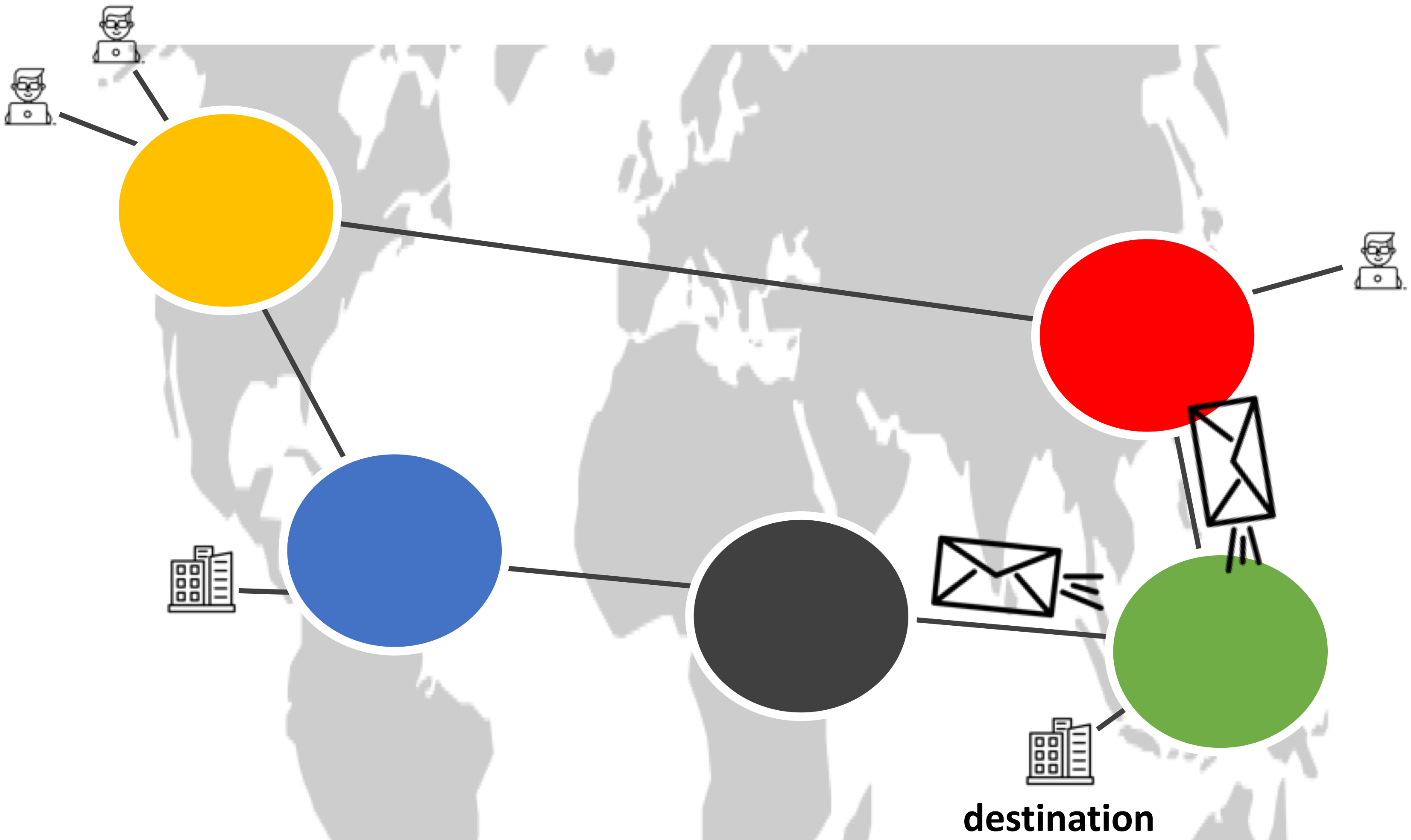
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



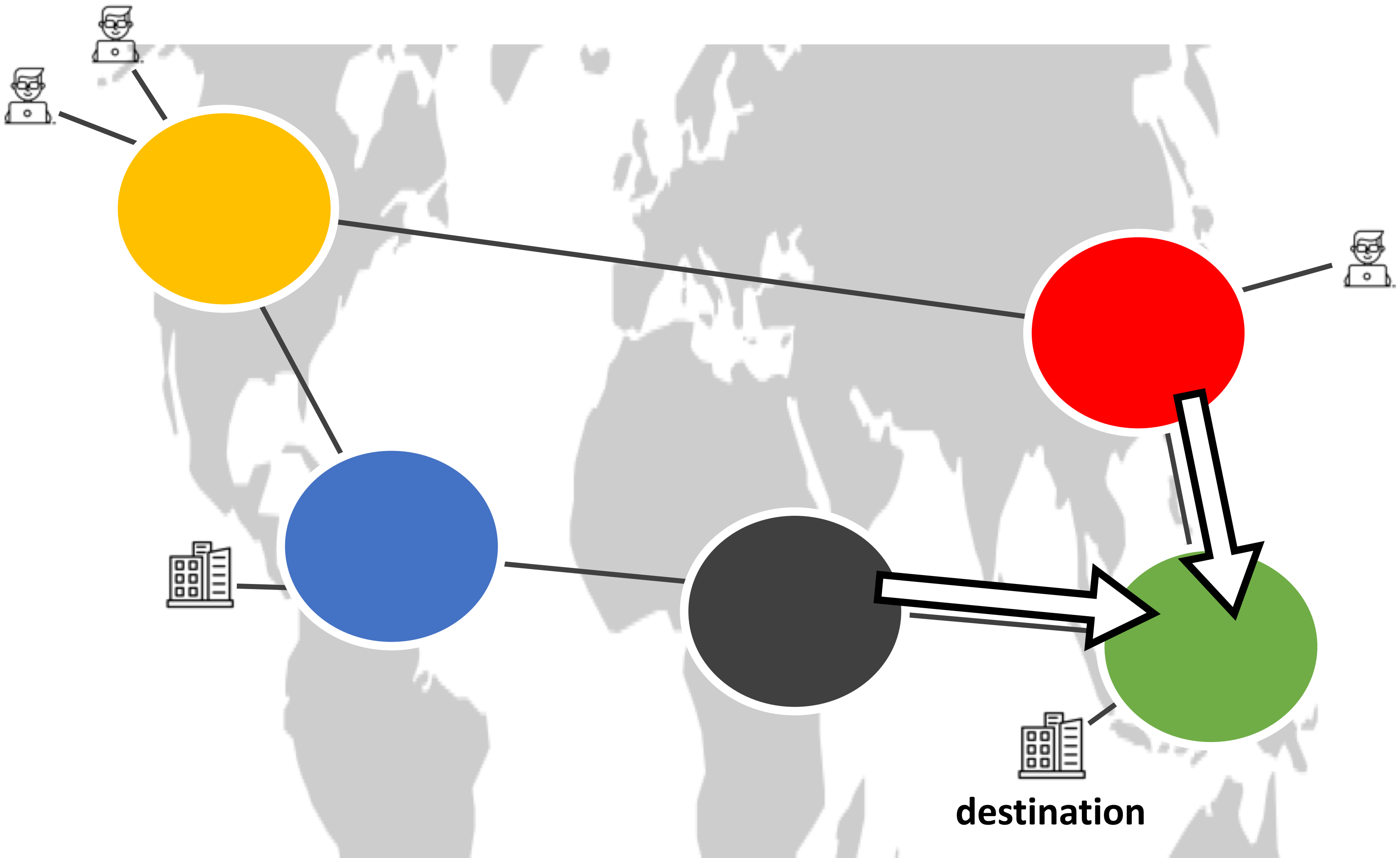
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



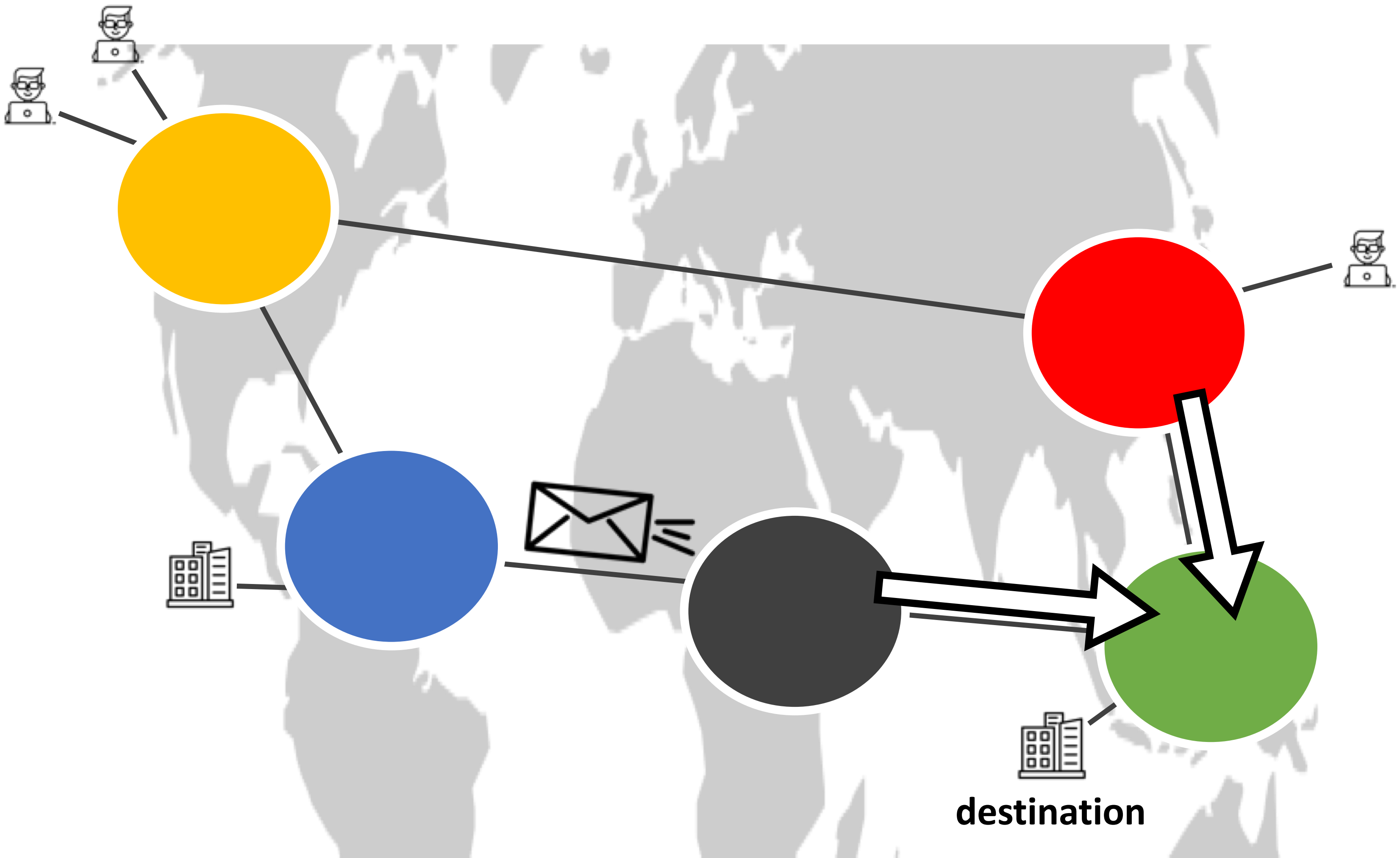
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



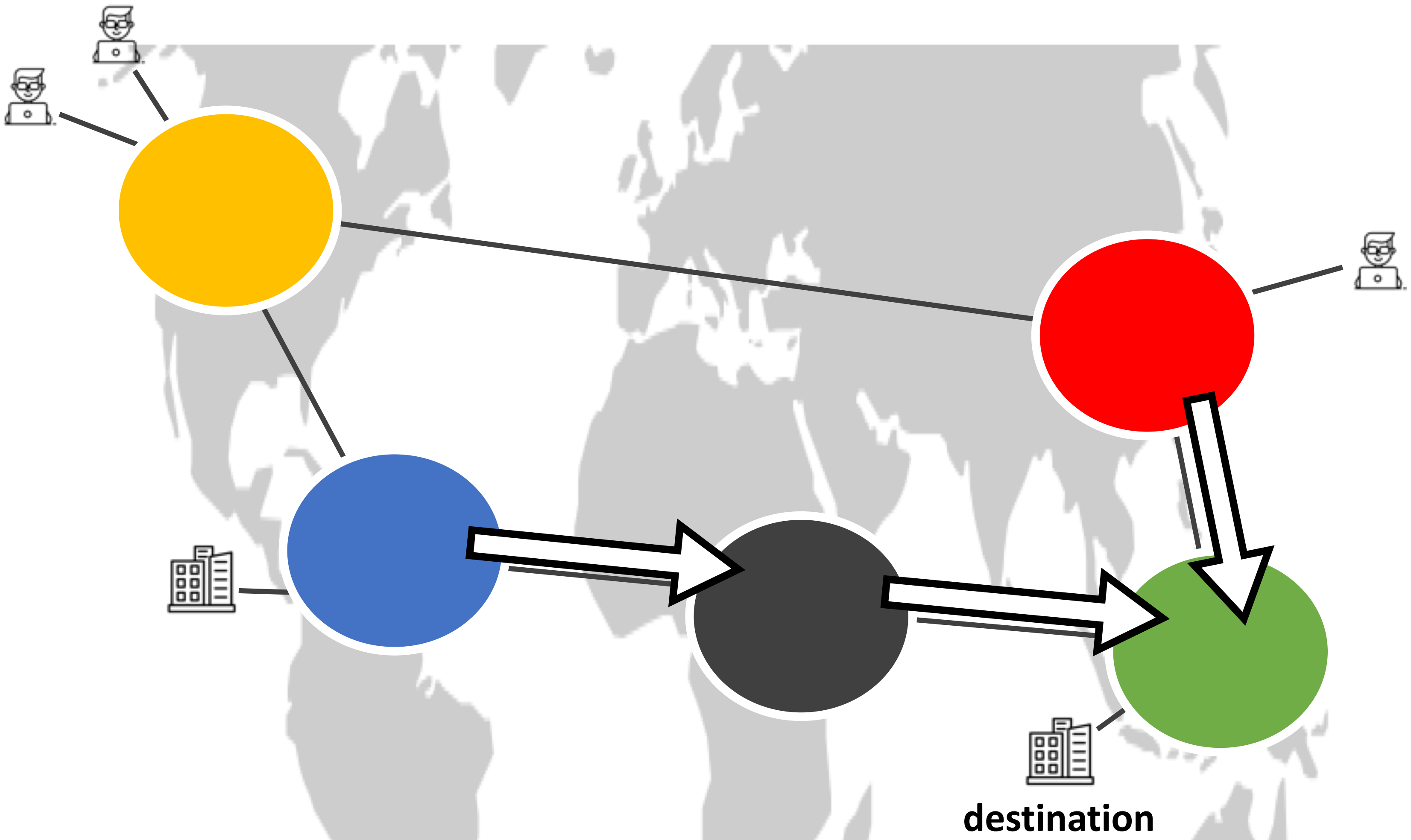
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



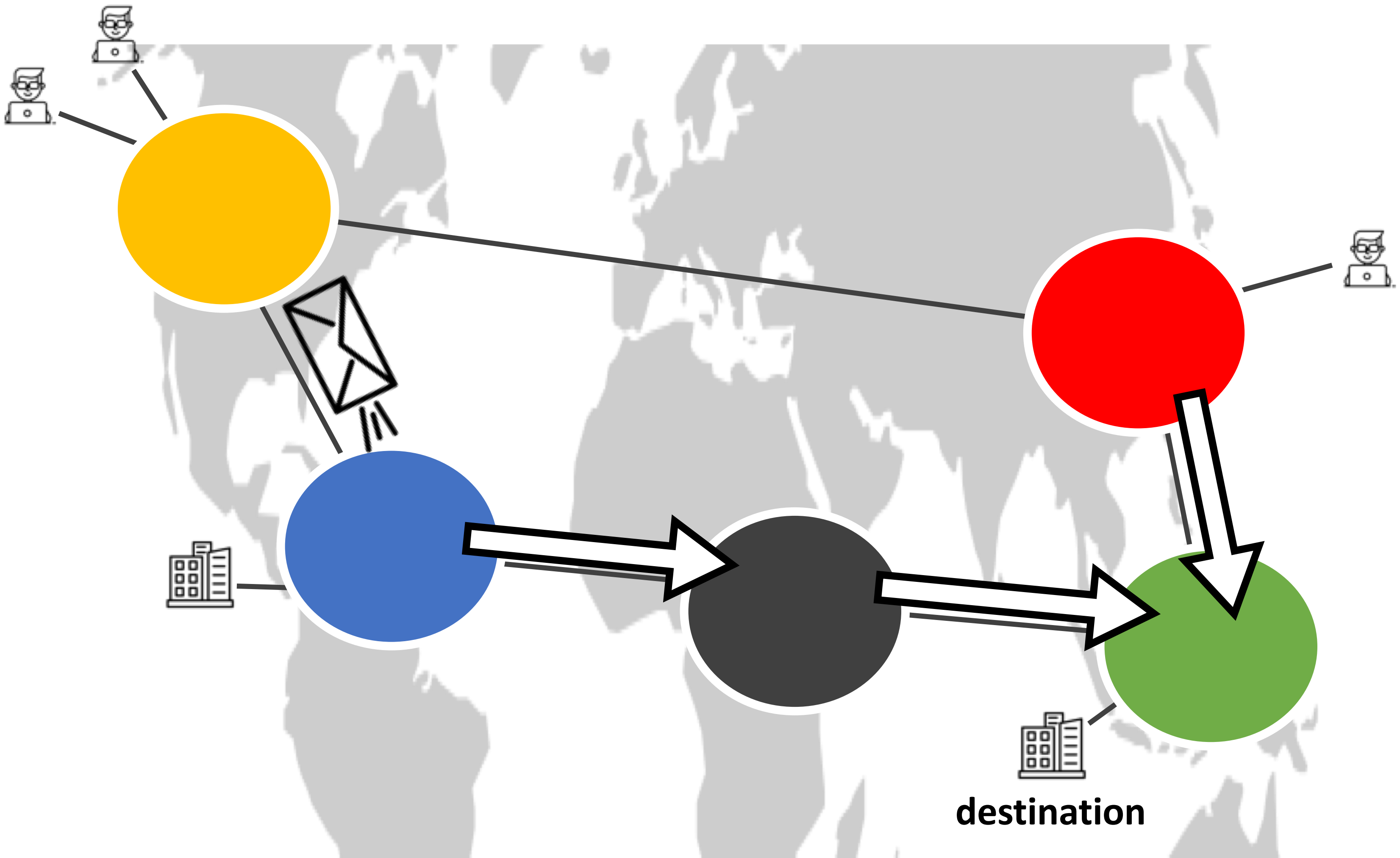
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



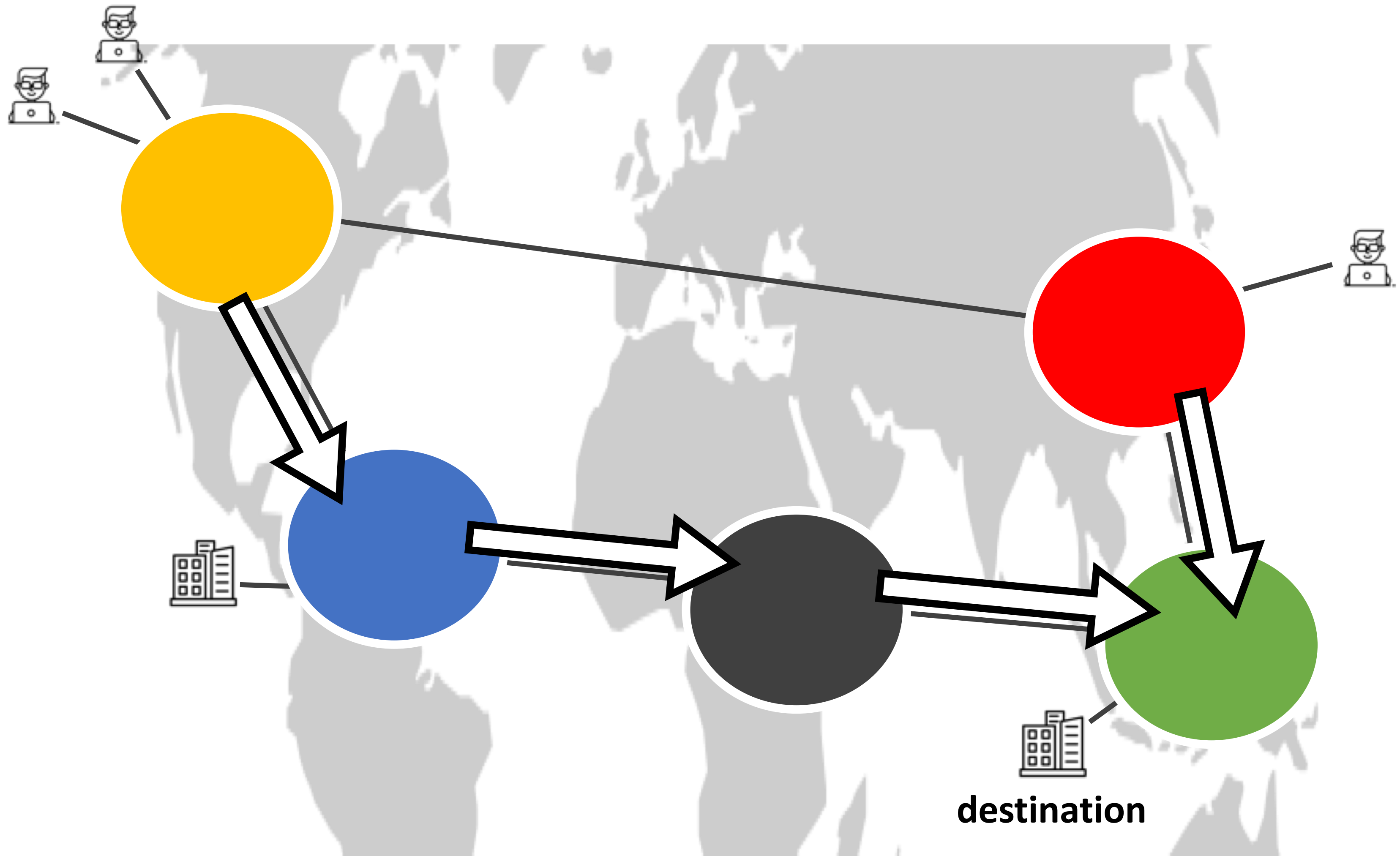
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



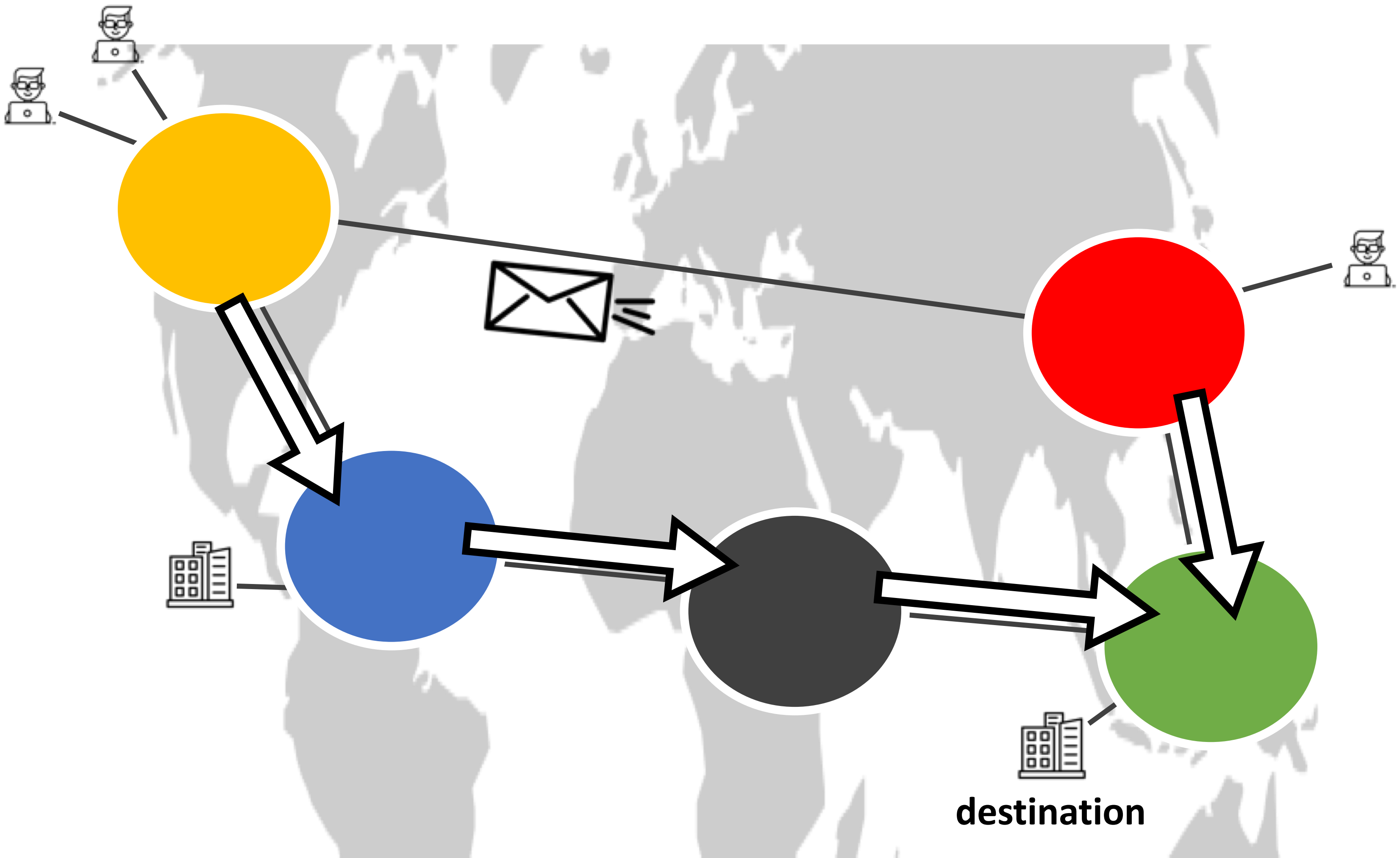
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



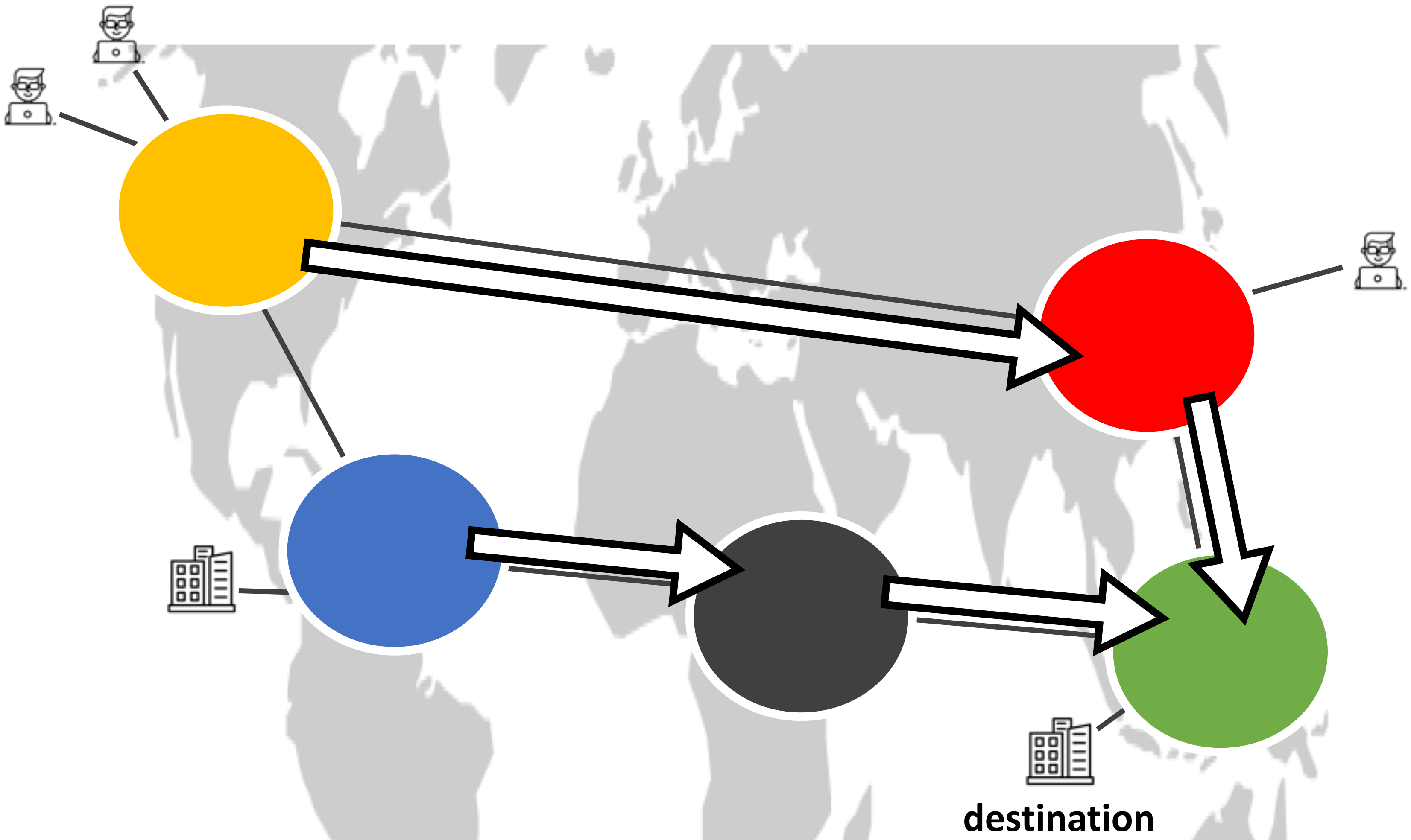
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



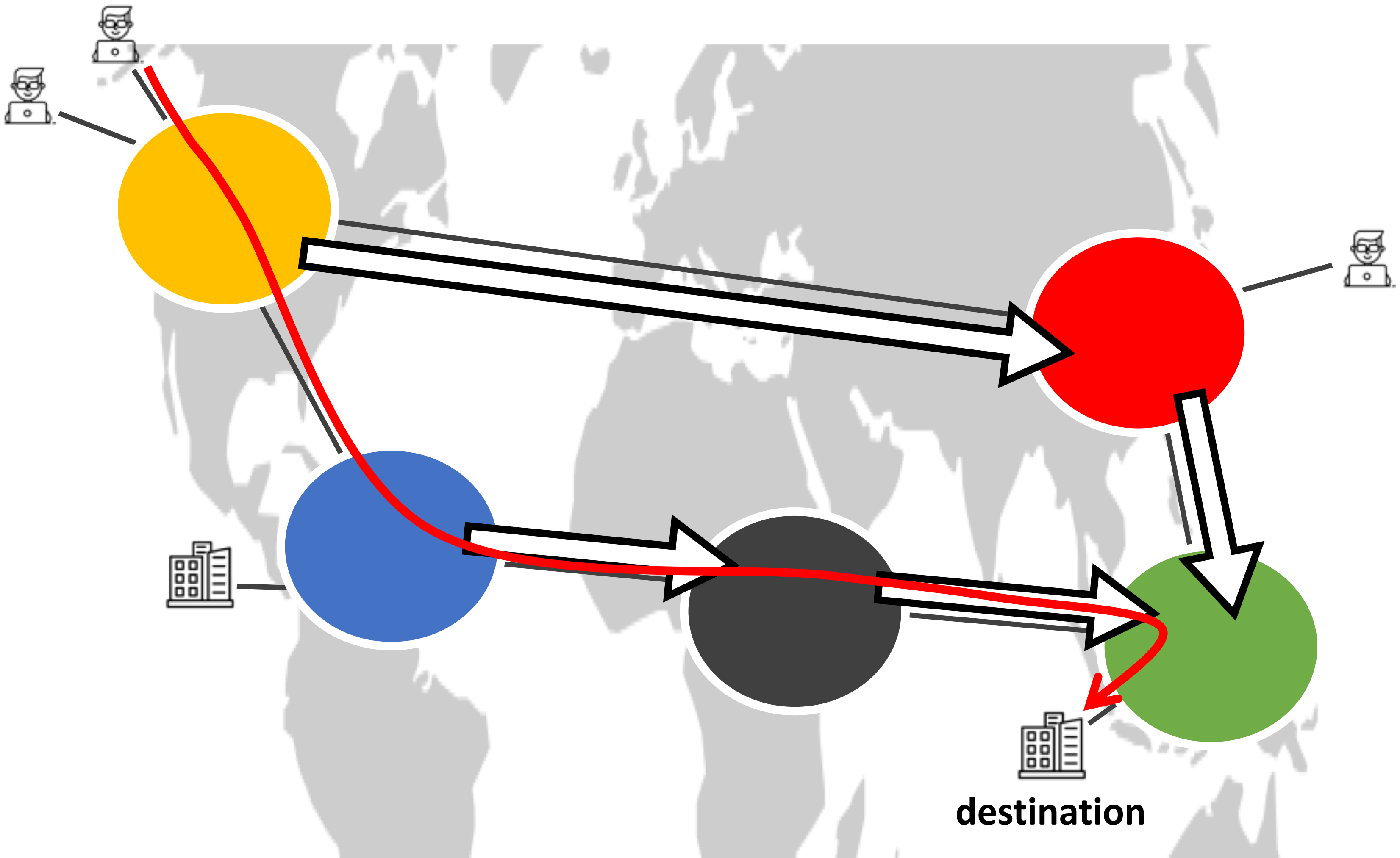
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



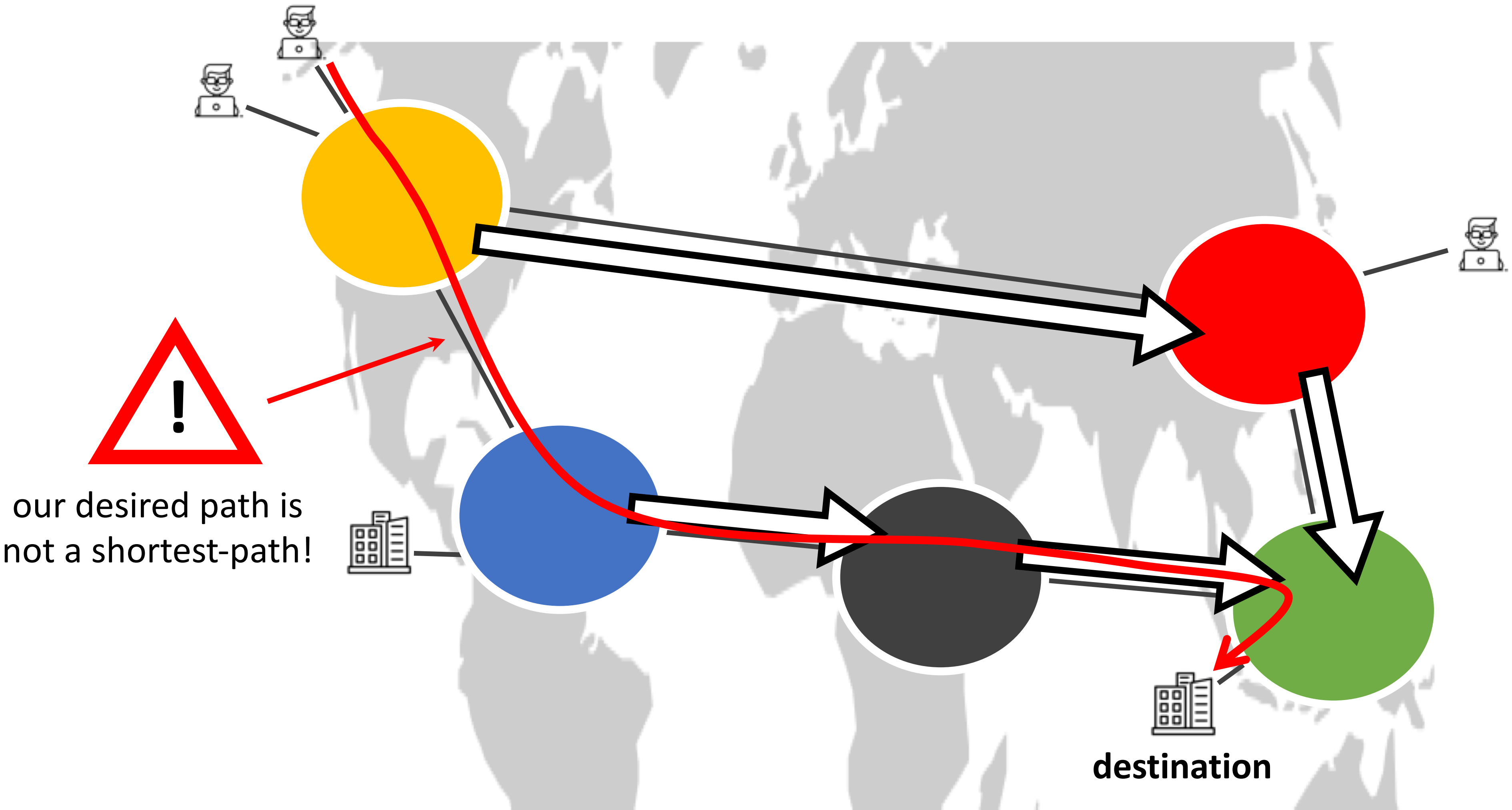
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



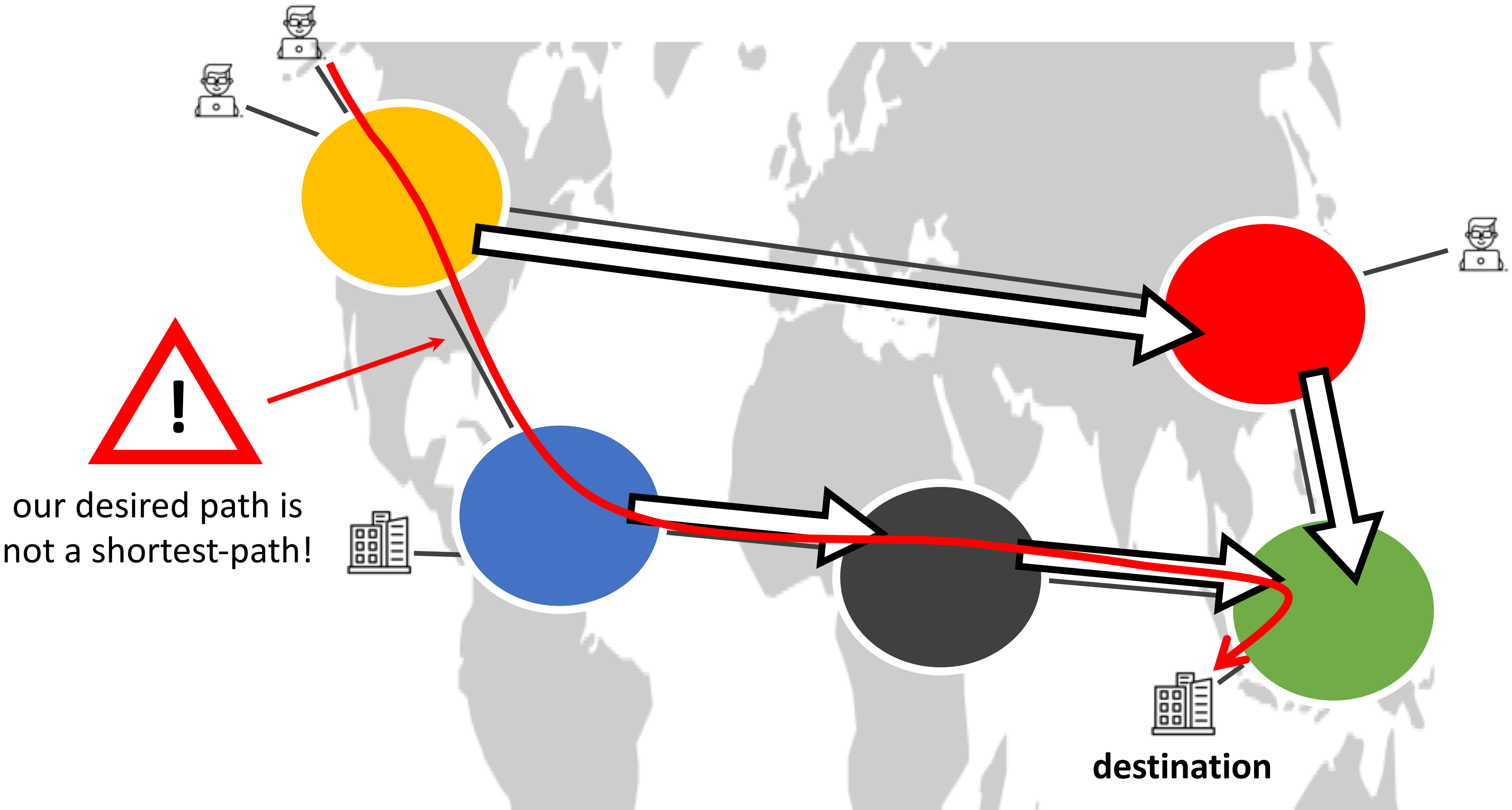
Distributed Bellman-Ford: an example of shortest-path path-vector protocol



Distributed Bellman-Ford: an example of shortest-path path-vector protocol



The Border Gateway Protocol:
a policy-based path-vector protocol



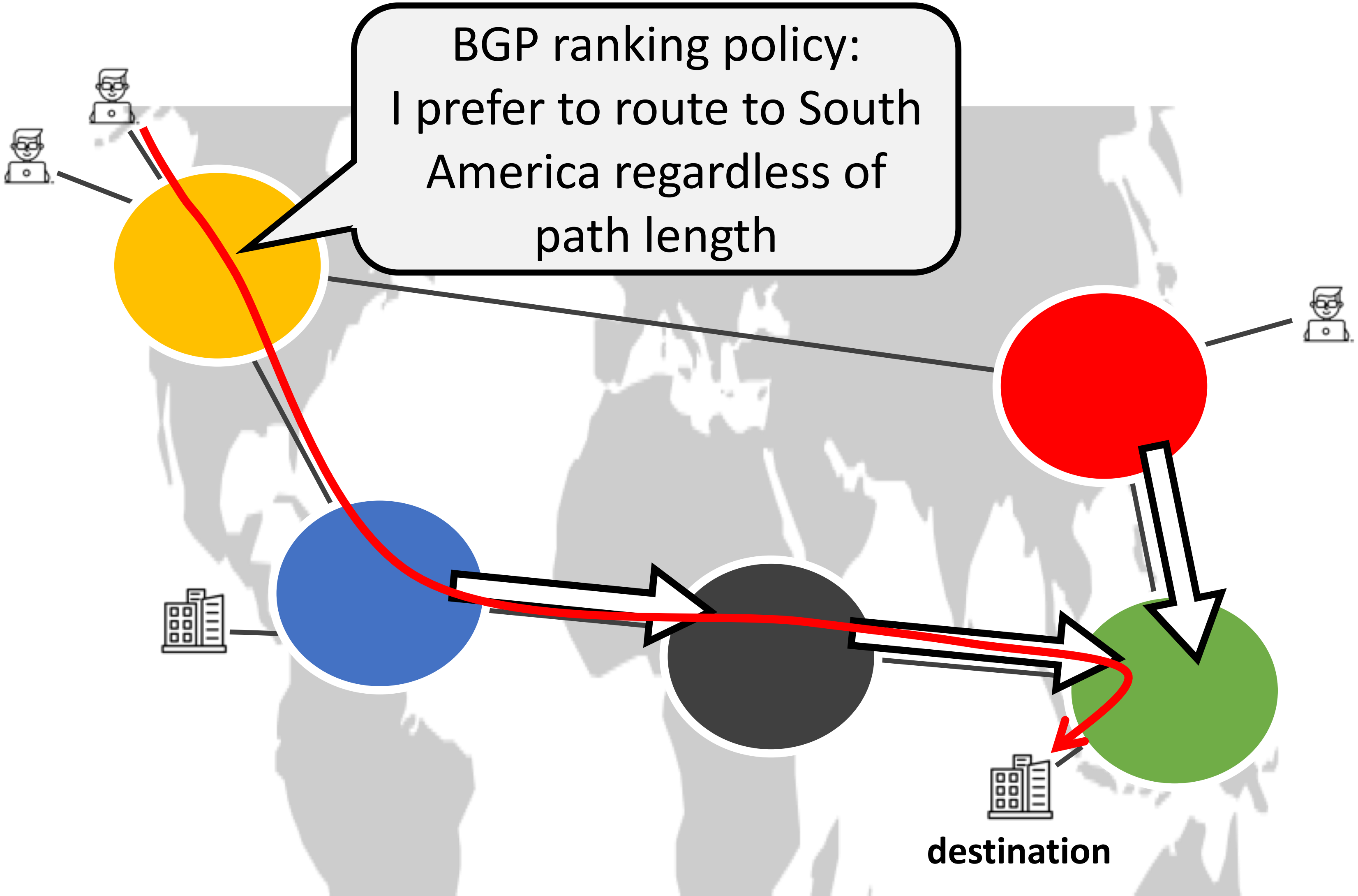
BGP policy-based path-vector

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)

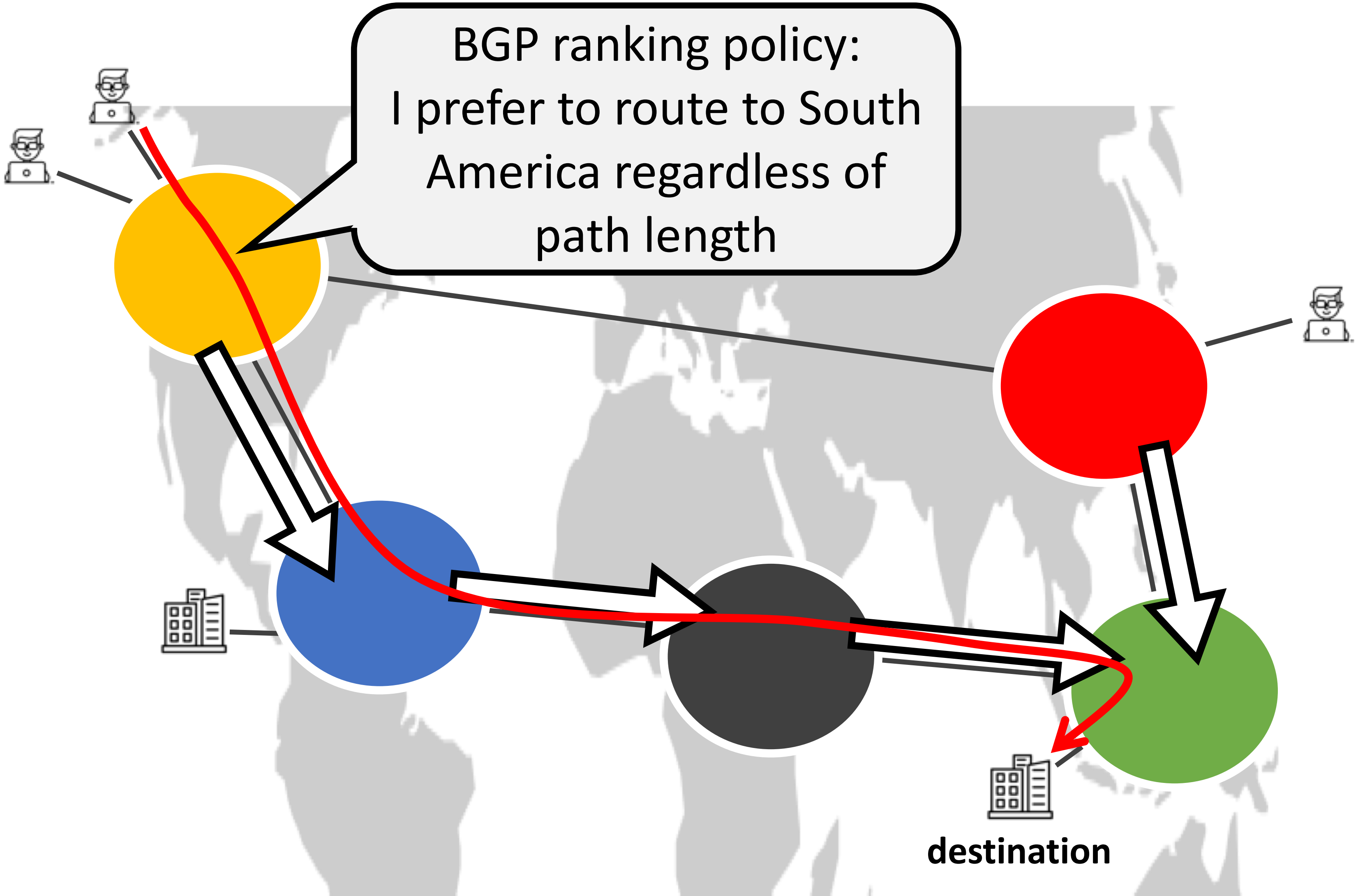
The Border Gateway Protocol:

a policy-based distance-vector protocol



The Border Gateway Protocol:

a policy-based distance-vector protocol



Distributed Bellman Ford routing: convergence vs expressiveness

	shortest-path path-vector
routing expressiveness	low

Distributed Bellman Ford routing: convergence vs expressiveness

	shortest-path path-vector	policy-based path-vector (BGP)
routing expressiveness	low	high

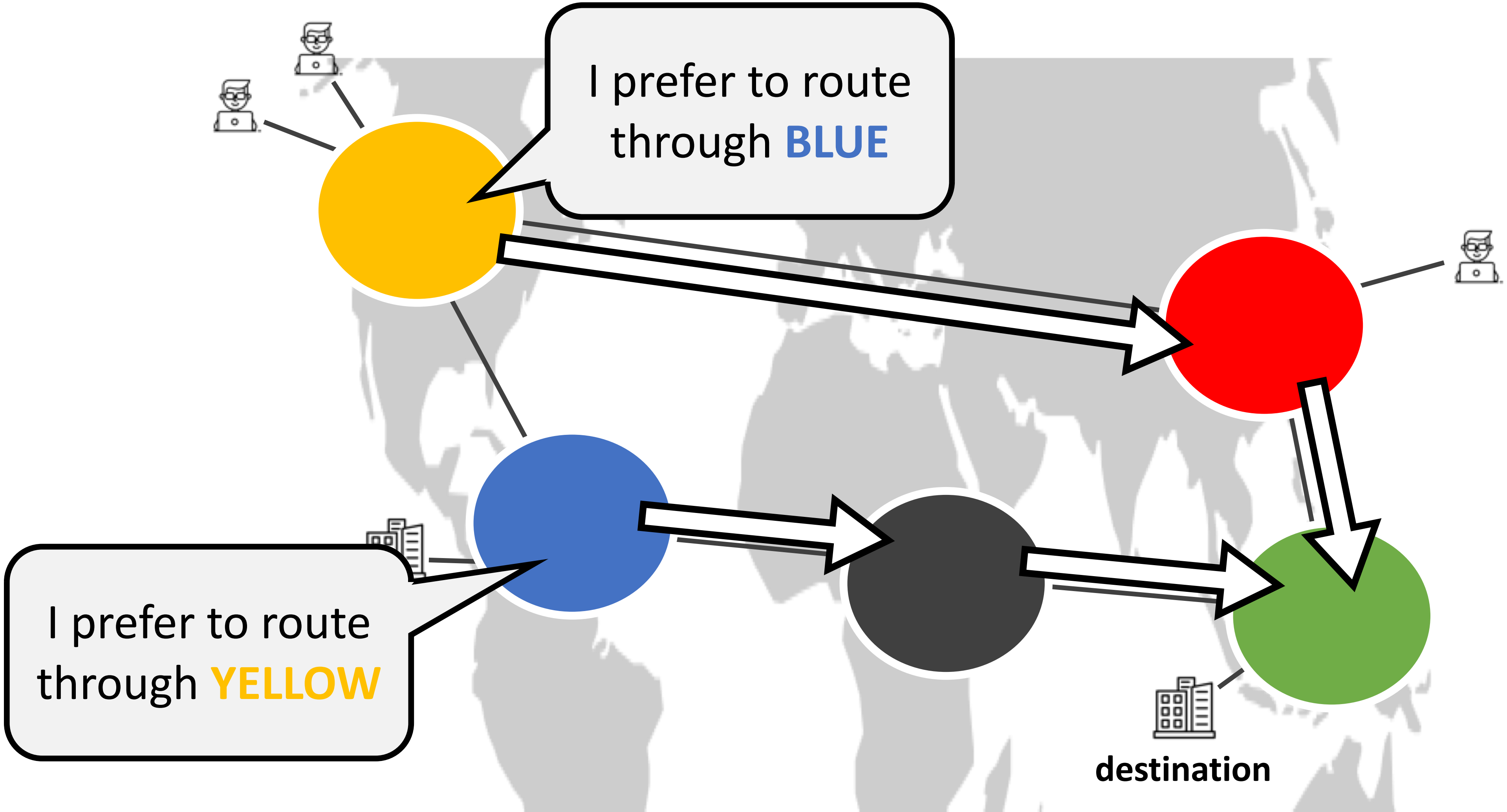
Distributed Bellman Ford routing: convergence vs expressiveness

	shortest-path path-vector	policy-based path-vector (BGP)
routing expressiveness	low	high
guaranteed convergence	yes	

Distributed Bellman Ford routing: convergence vs expressiveness

	shortest-path path-vector	policy-based path-vector (BGP)
routing expressiveness	low	high
guaranteed convergence	yes	no

The Border Gateway Protocol: routing inconsistencies



Paper #4:
Asynchronous Convergence of Policy-Rich Distributed Bellman-Ford Routing Protocols

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

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:

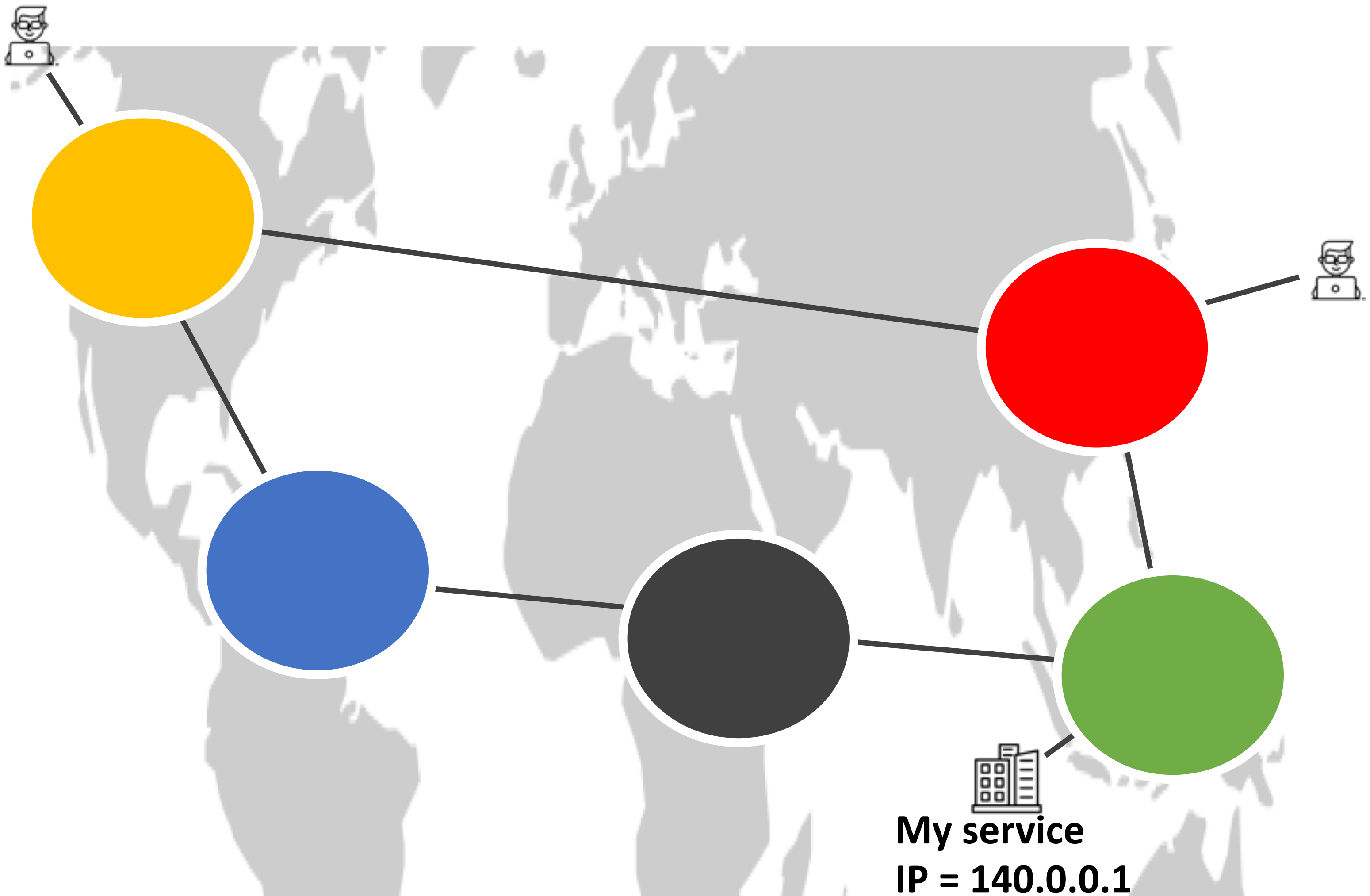
”What __classes of routing policies__ (i.e., import, ranking, and export policies) are guaranteed to converge to a stable state __when messages can be lost, reordered, and indefinitely delayed__?”

- Studies both distance-vector (RIP-like) and path-vector (BGP-like) routing

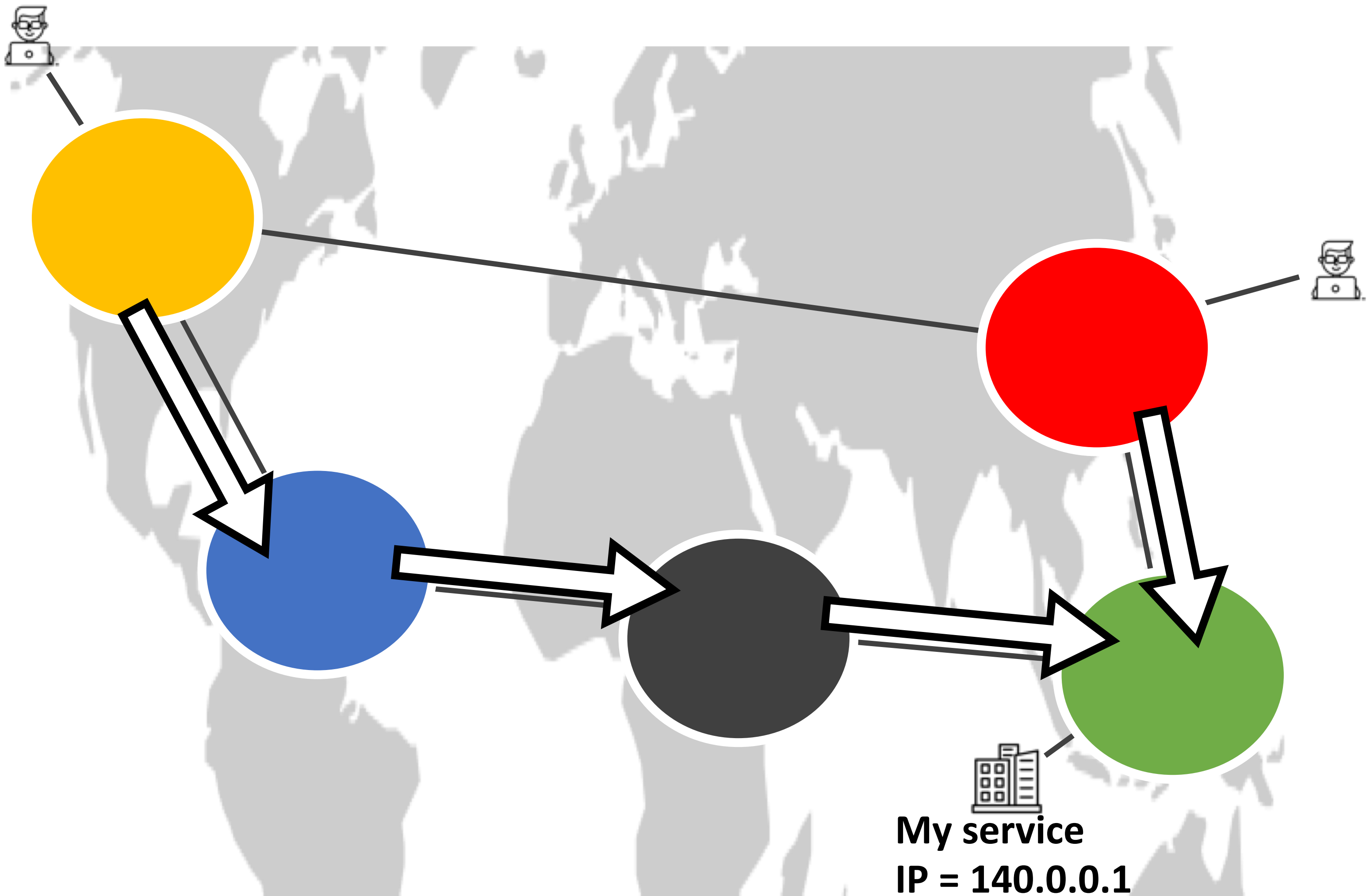
Recommended readings for paper #4

- 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

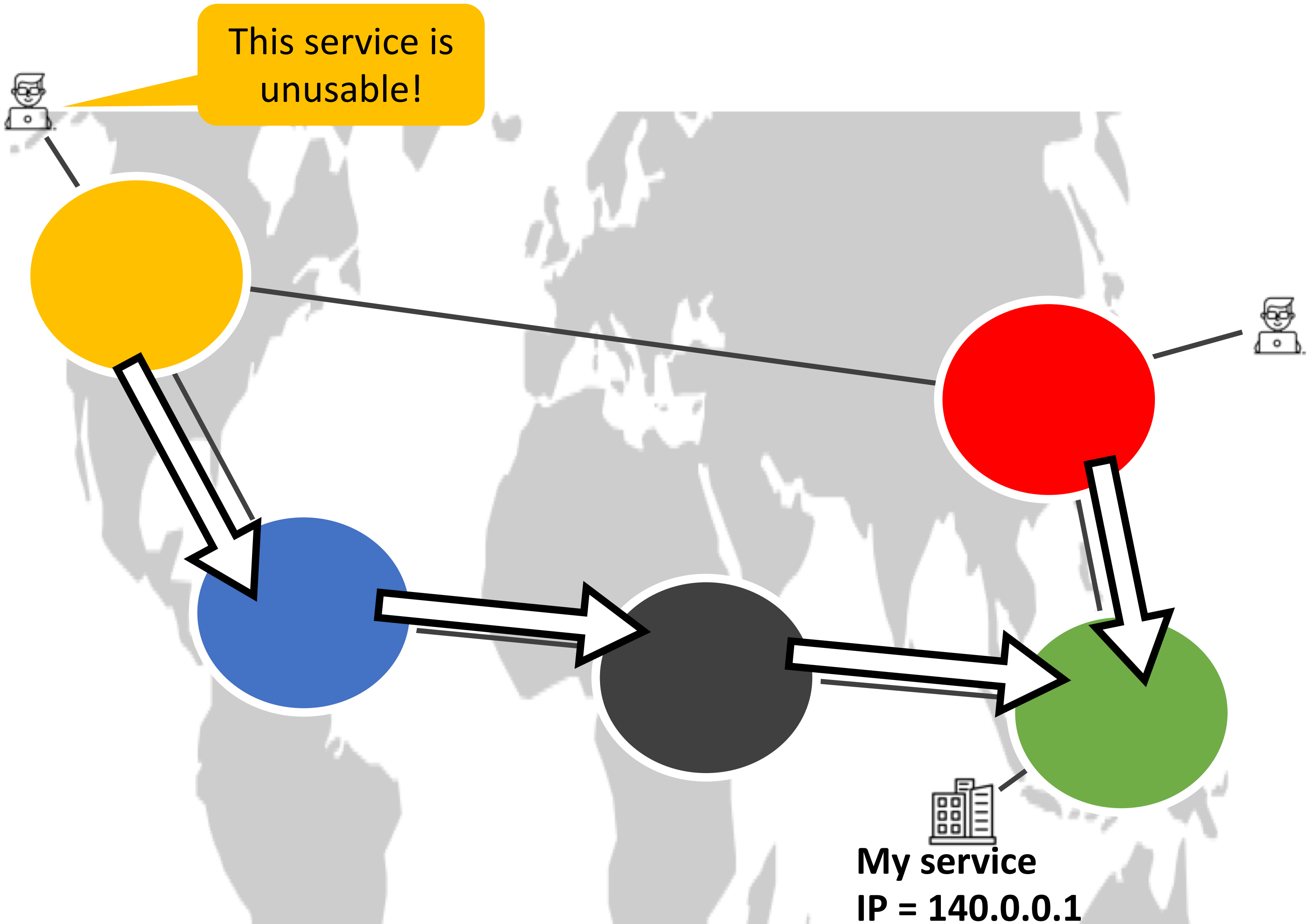
Internet Load-balancing



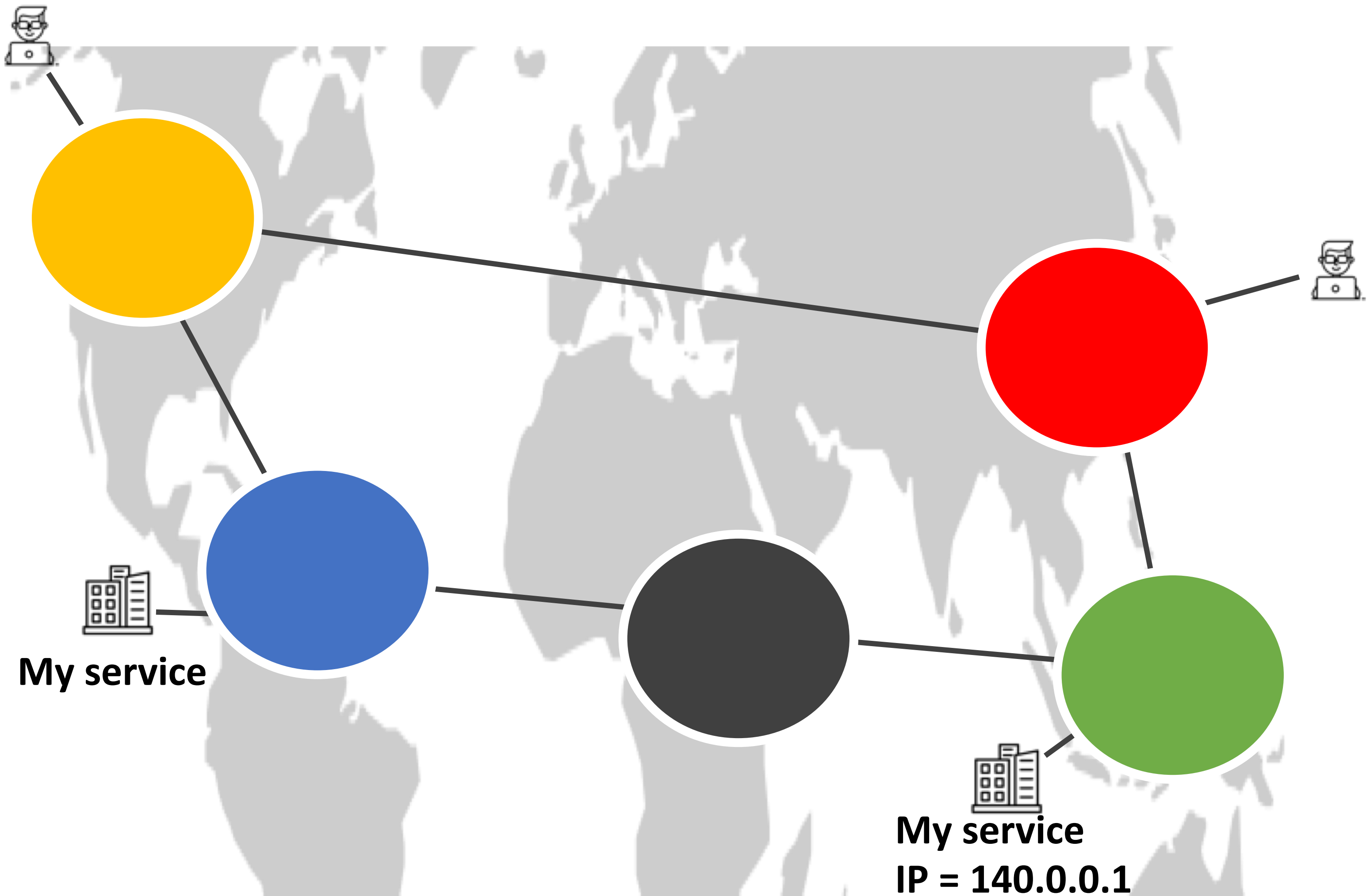
Internet Load-balancing:
BGP determines Internet routing paths



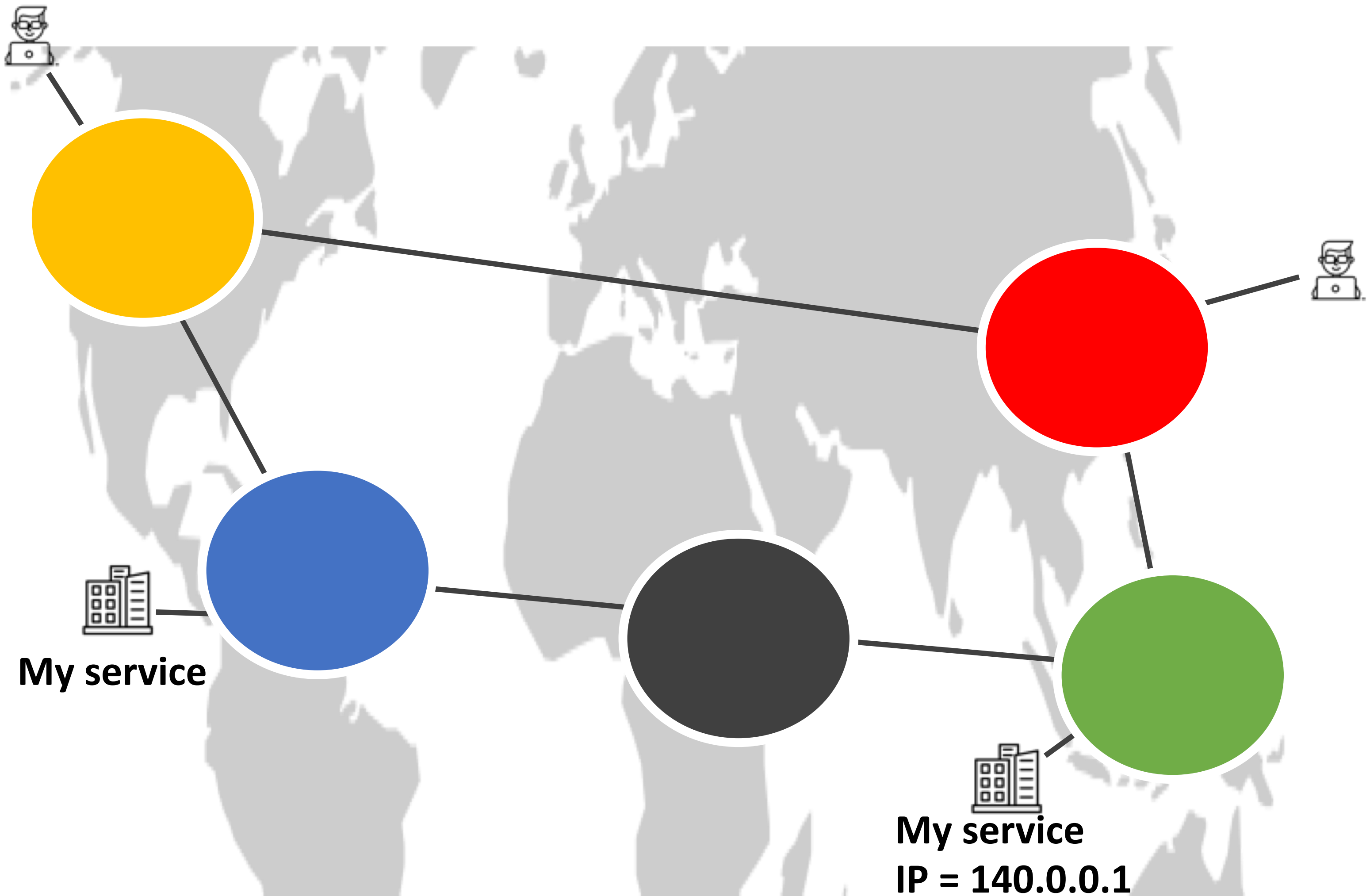
Internet Load-balancing:
BGP determines Internet routing paths



Reducing user latency:
Add service replicas closer to the users

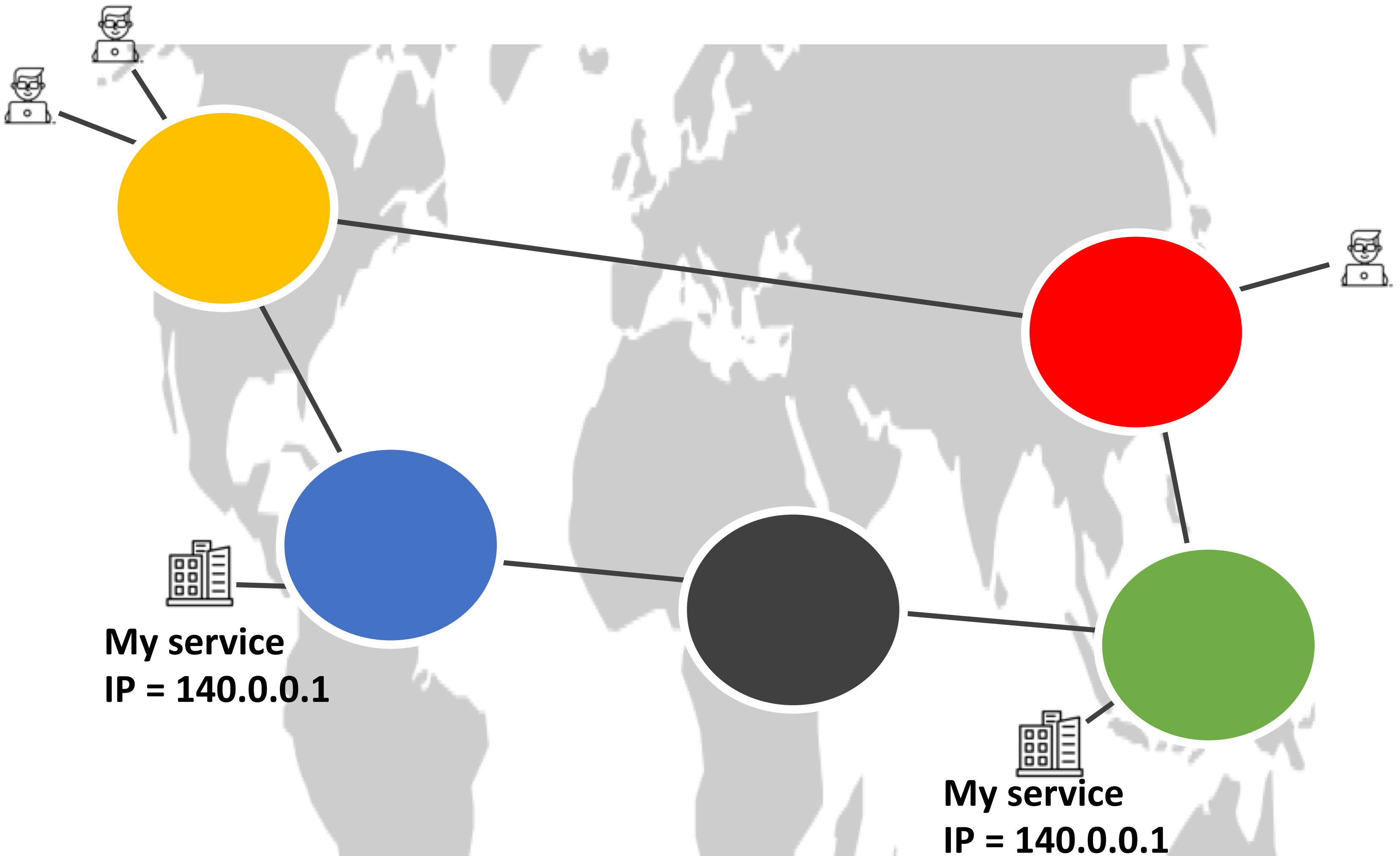


Reducing user latency: How to reach the "closest" replica?



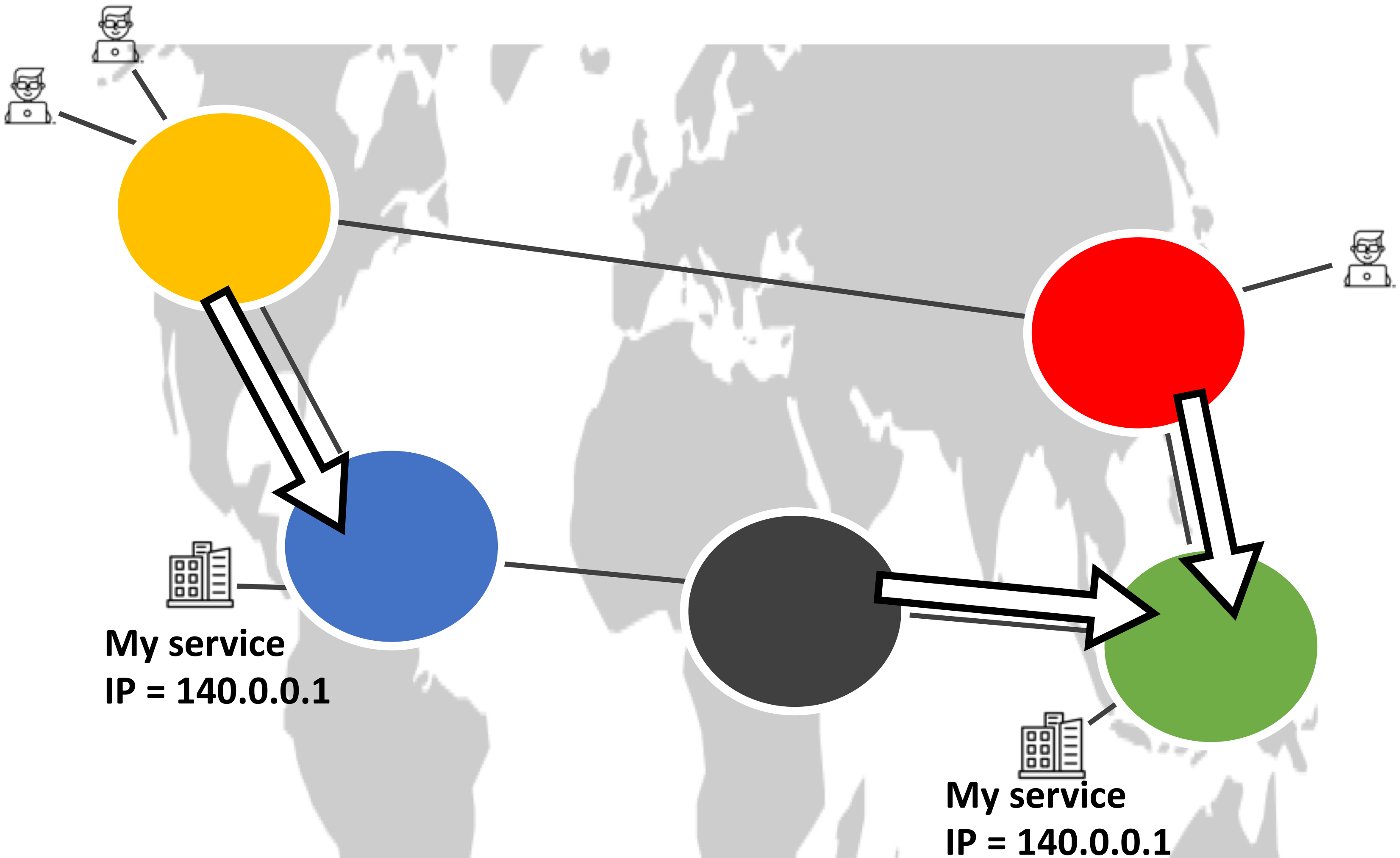
One approach is anycast routing:

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!

[1] Internet Anycast: Performance, Problems, & Potential. In SIGCOMM 2018

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

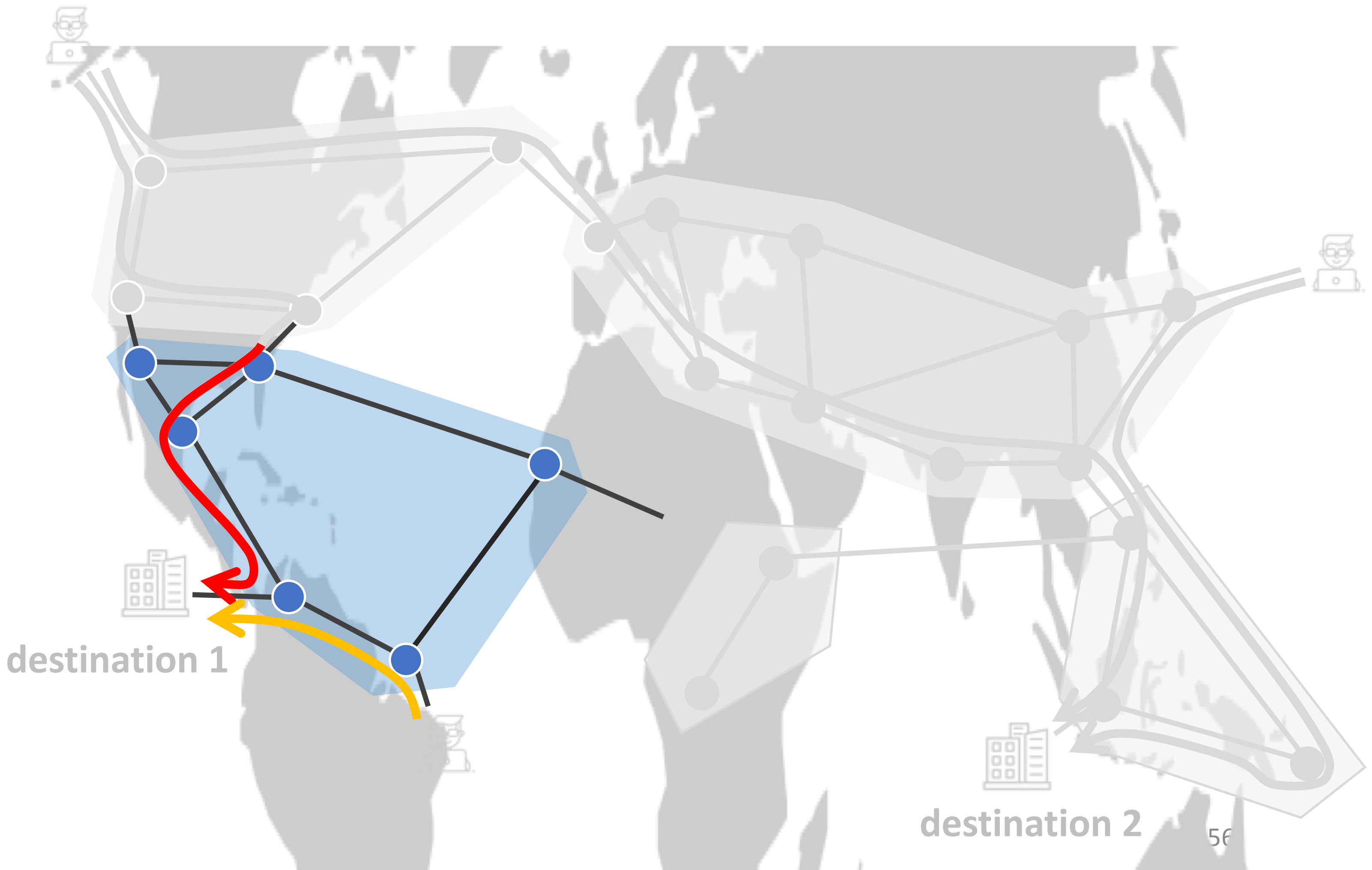
Anycast routing:

- H. Ballani and P. Francis. ”*Towards a global IP anycast service*”. In ACM SIGCOMM, 2005.

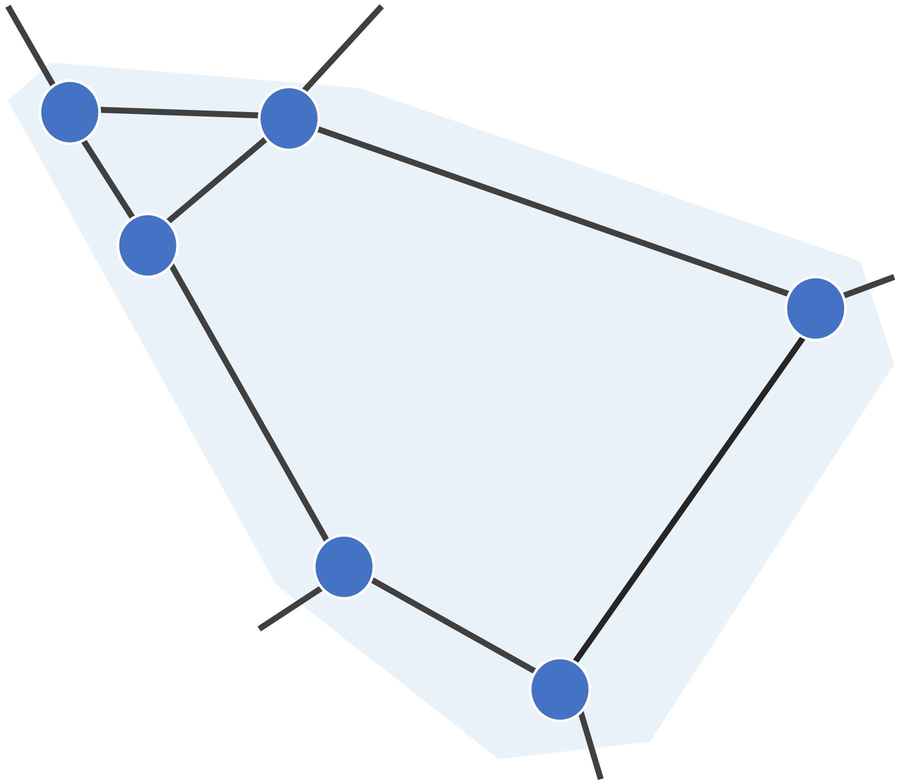
Demand-aware BGP improvements:

- 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

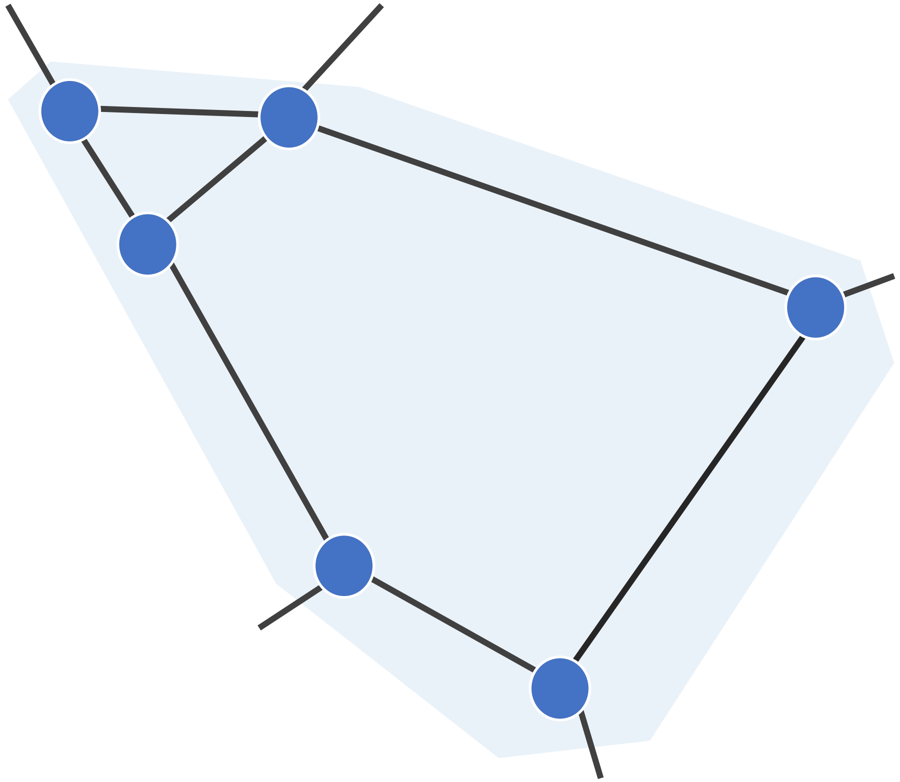


Intra-domain routing: selecting paths within a single domain



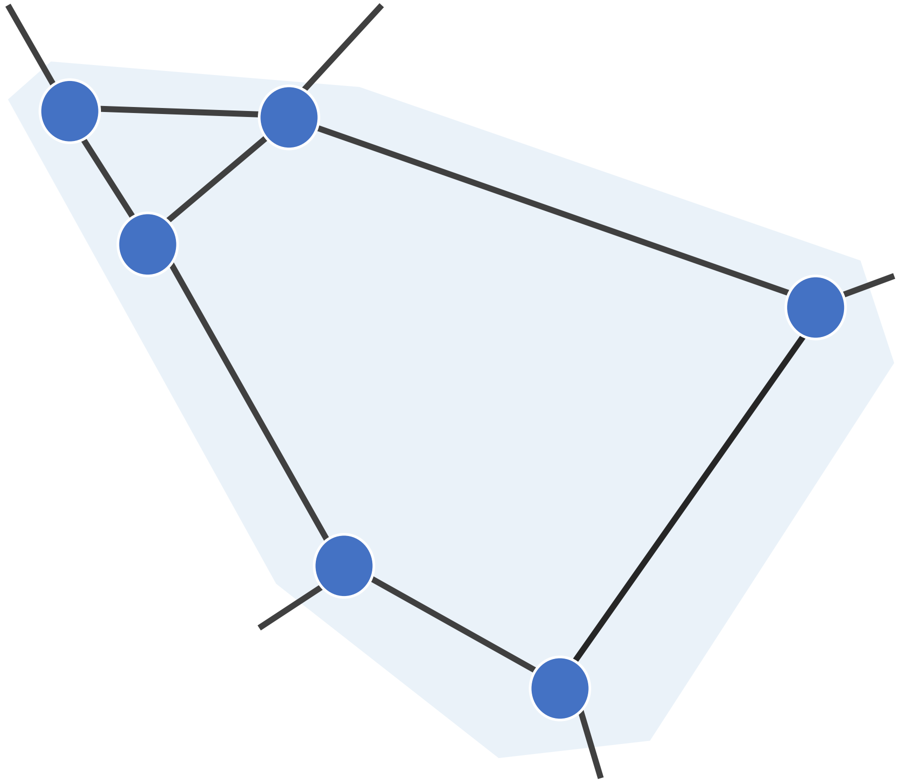
Intra-domain routing: selecting paths within a single domain

what are the best paths?



Intra-domain routing:
selecting paths within a single domain

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



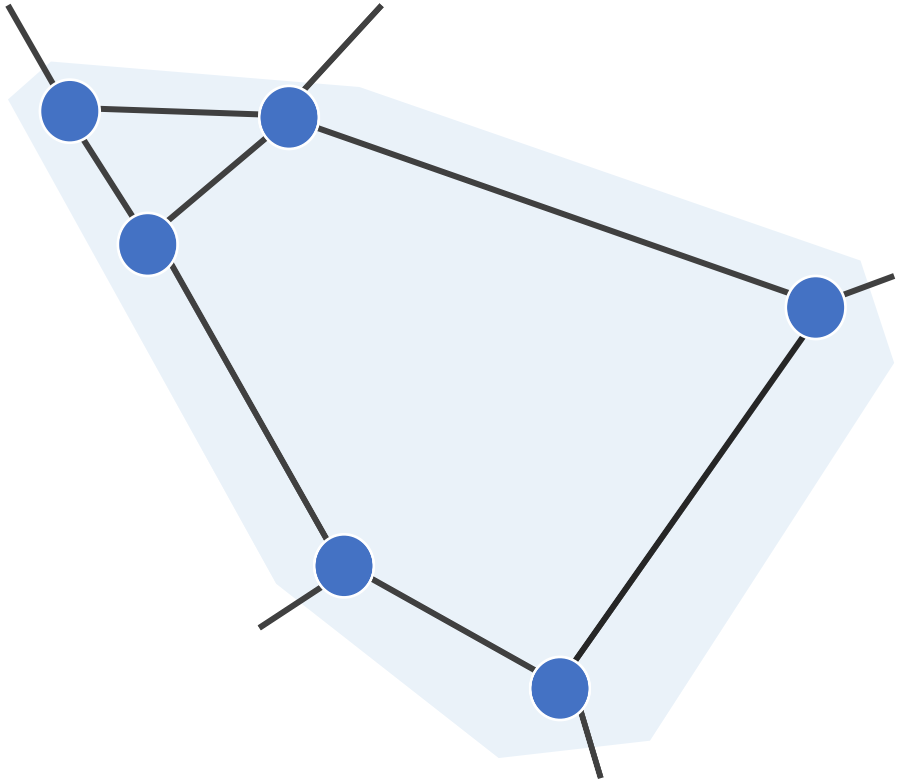
Intra-domain routing: selecting paths within a single domain

Objectives, e.g.,:

- min load on links
- min latency

Constraints, e.g.,:

- routing expressiveness

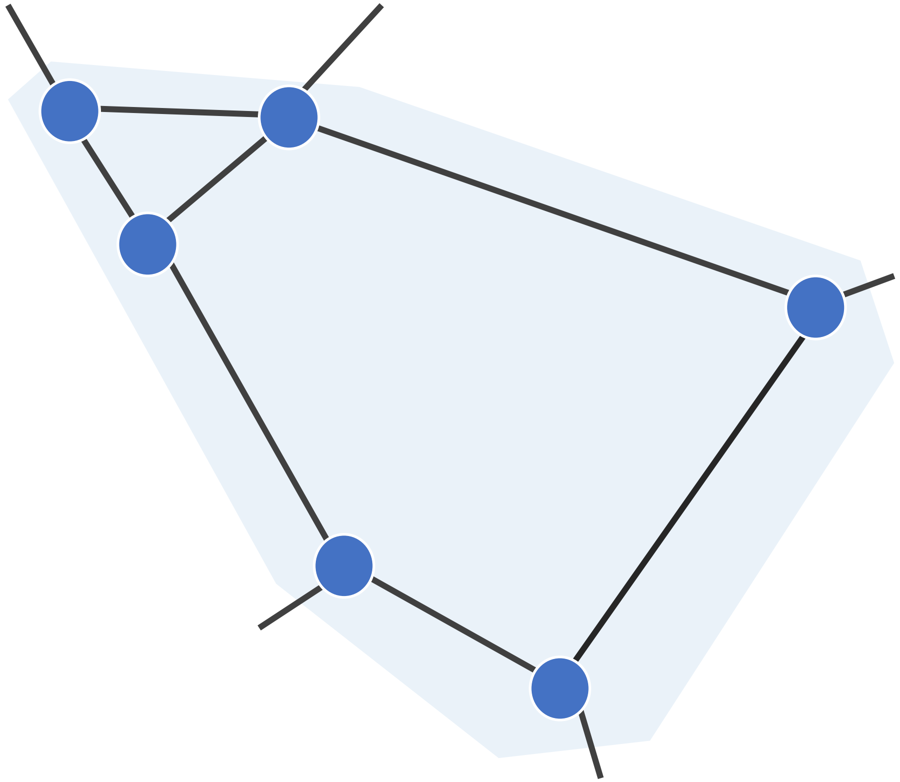


Intra-domain routing: selecting paths within a single domain

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

Constraints, e.g.,:
- routing expressiveness

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

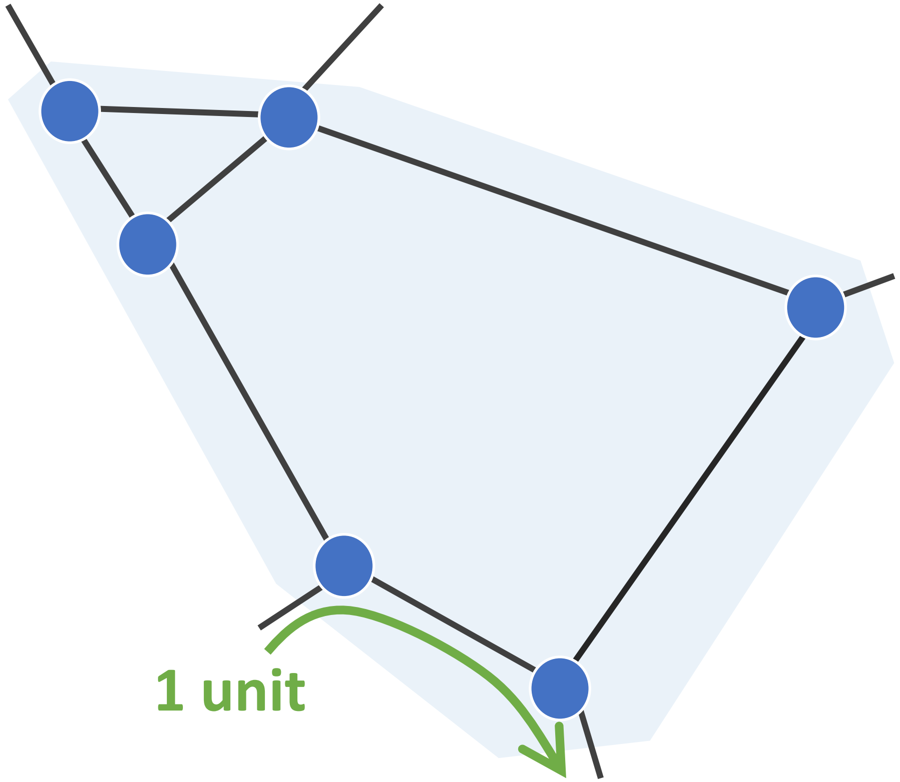


Intra-domain routing: selecting paths within a single domain

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

Constraints, e.g.,:
- routing expressiveness

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



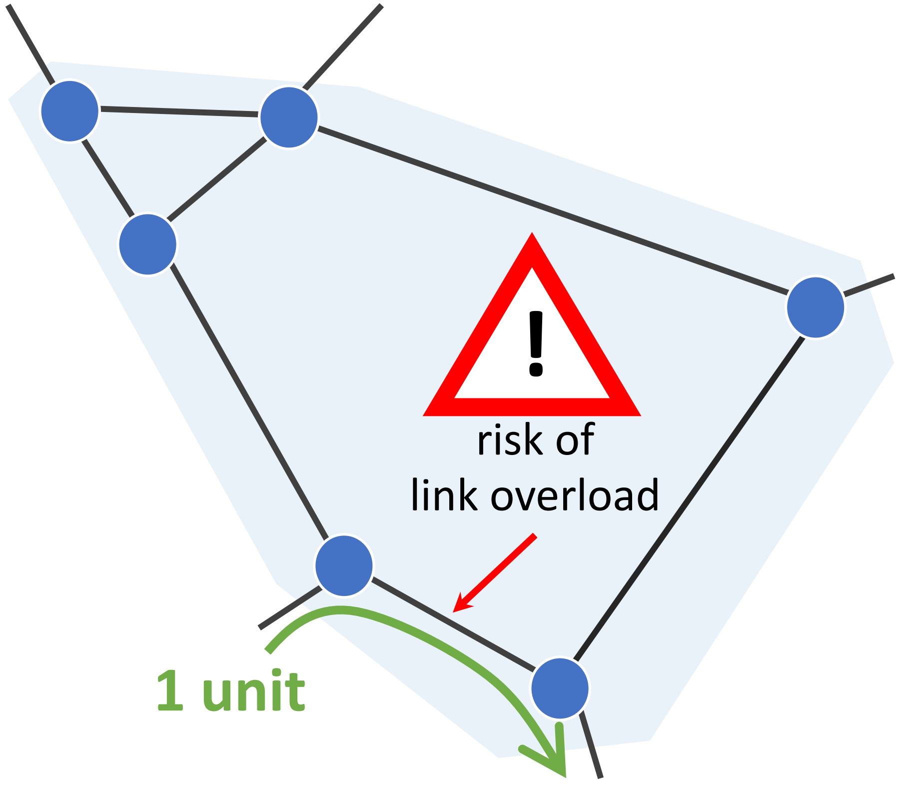
all internal link capacities are 1

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



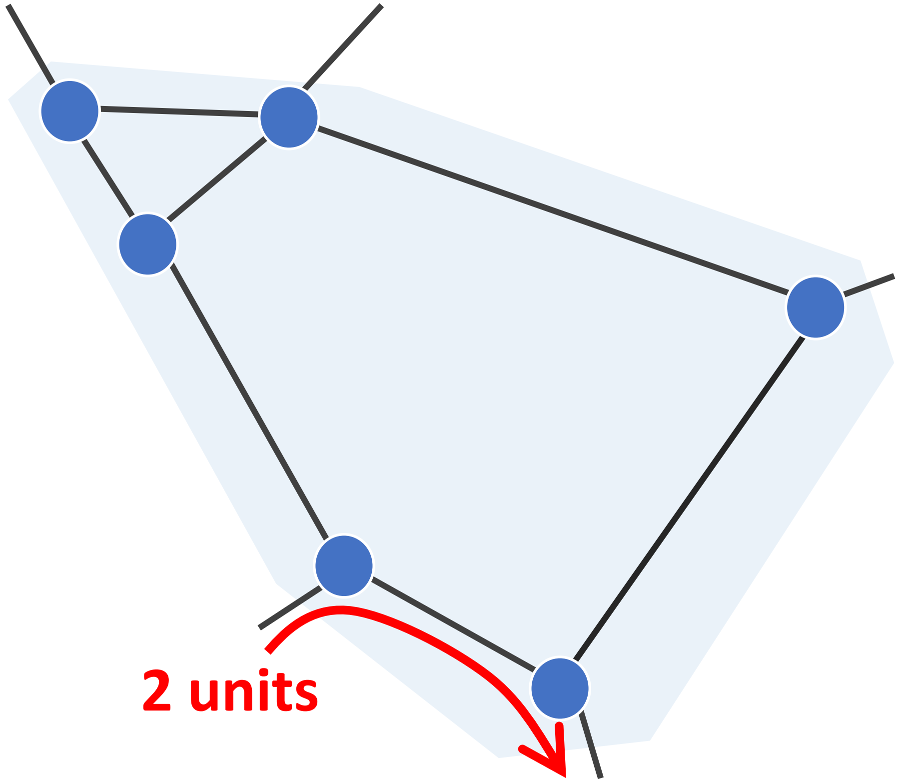
all internal link capacities are 1

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



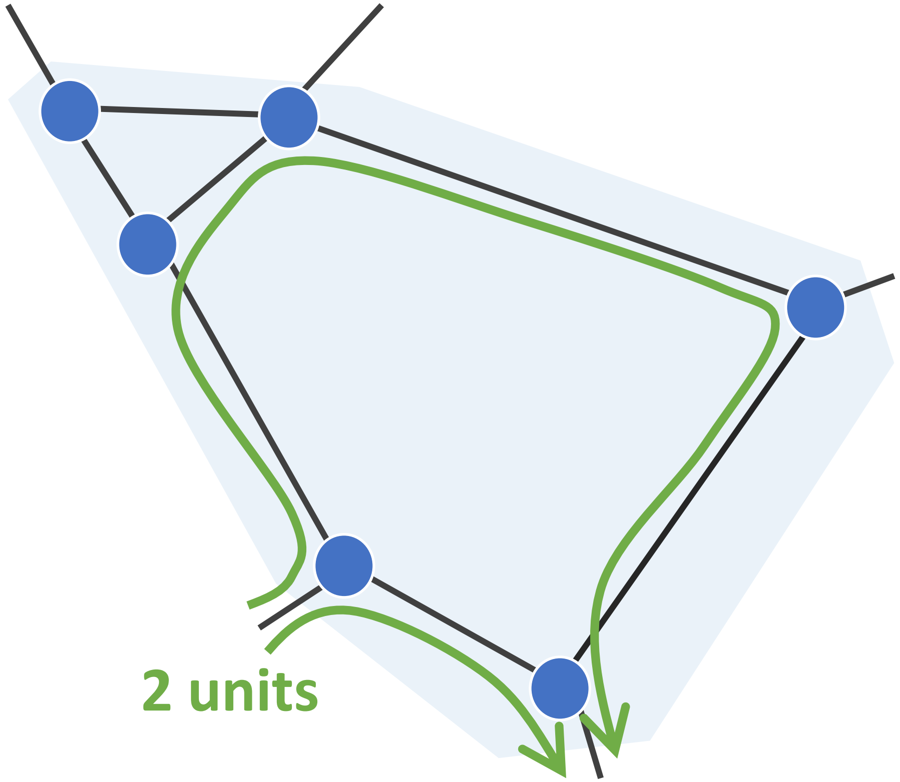
all internal link capacities are 1

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



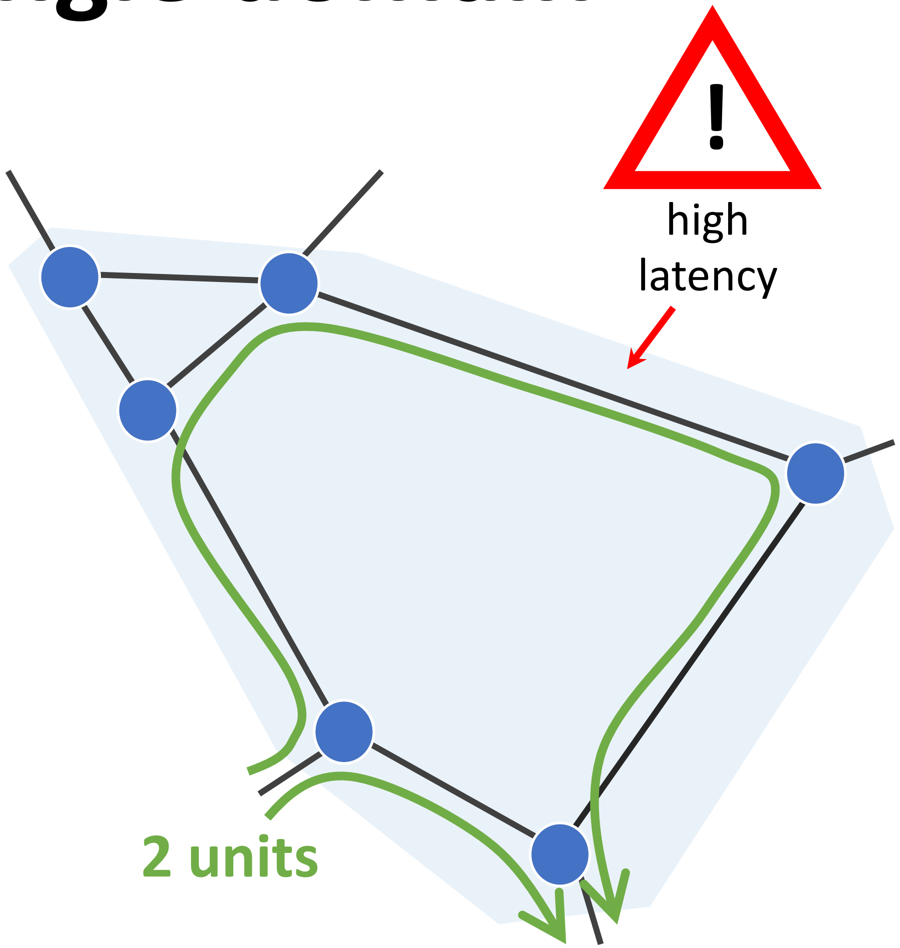
all internal link capacities are 1

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



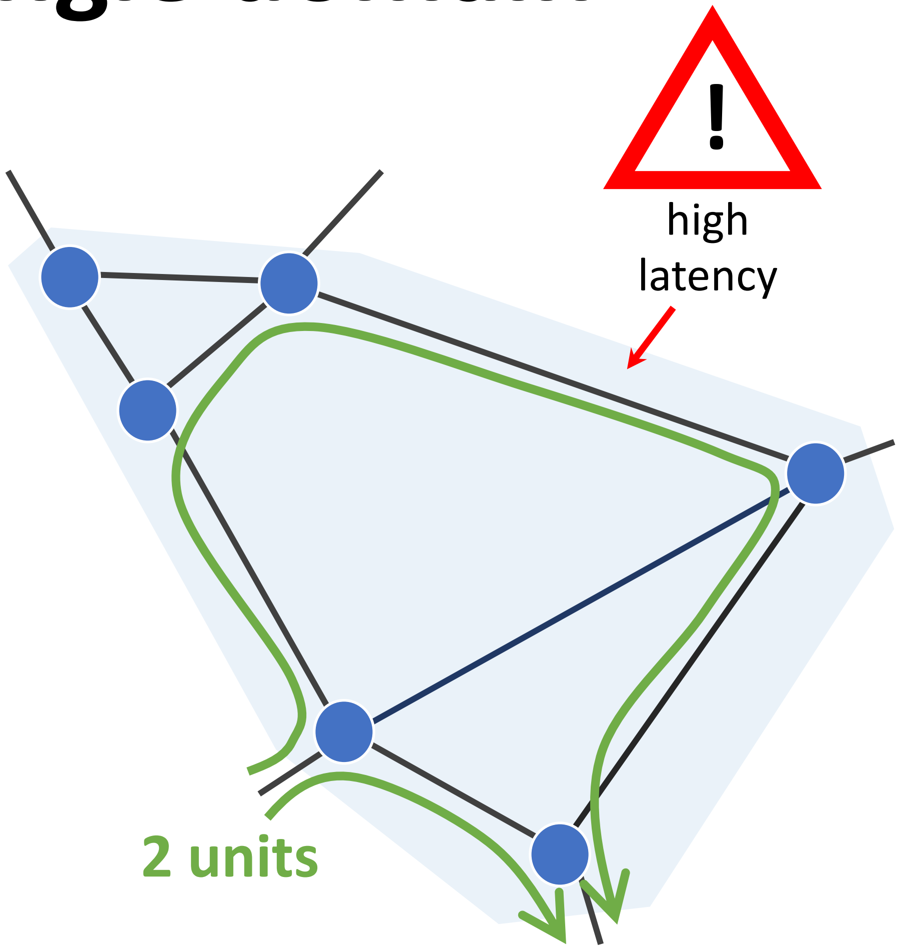
all internal link capacities are 1

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



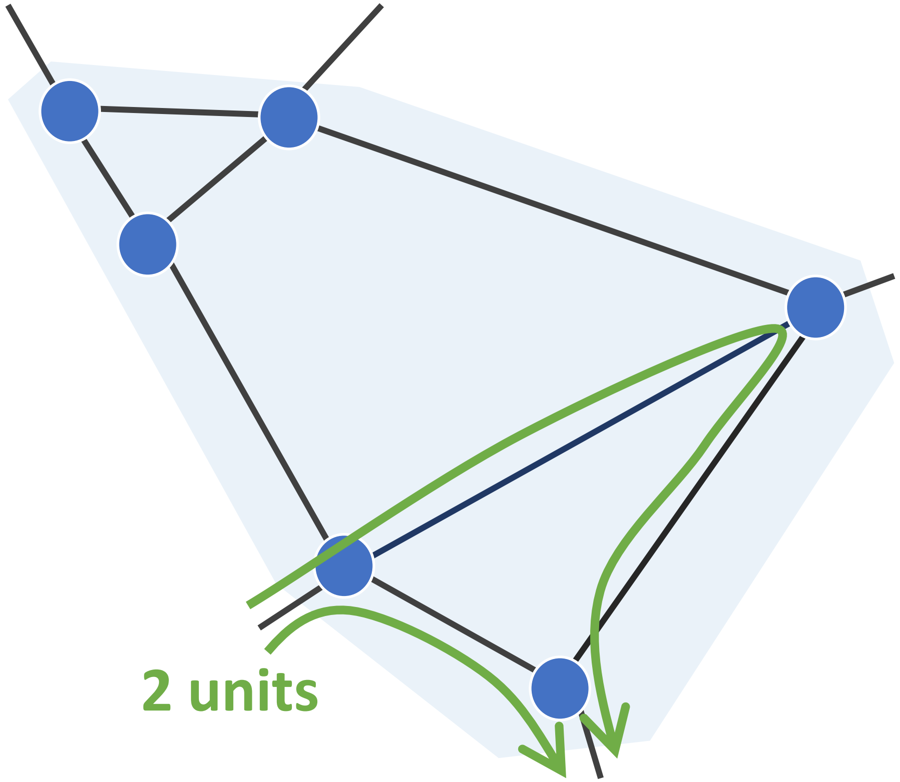
all internal link capacities are 1

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 accomodate 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 accomodate 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
- A. Curtis et al. ”*DevoFlow: scaling flow management for high-performance networks*”. In SIGCOMM 2011

Distributed routing:

- R. Gallager ”*A Minimum Delay Routing Algorithm Using Distributed Computation*”. In Transactions on Communications 1977
- N. Michael et al. "*HALO: Hop-by-Hop Adaptive Link-State Optimal Routing*". In ICNP 2013

Hash-based forwarding:

- Z. Cao et al. "*Performance of Hashing-Based Schemes for Internet Load Balancing*". In INFOCOM 2000

Topic Preview: **Routing**

2:10 pm - 3:50 pm Main-Conference Session 2: Routing

Session Chair: Nate Foster *(Cornell, USA)*
Location: Vigadó, 2nd-Floor Ceremonial Hall

Today, after lunch!

2:10 pm - 2:35 pm	**Internet Anycast: Performance, Problems and Potential** Zhihao Li, Dave Levin, Neil Spring, Bobby Bhattacharjee *(UMD, USA)*
2:35 pm - 3:00 pm	**B4 and After: Managing Hierarchy, Partitioning, and Asymmetry for Availability and Scale in Google's Software-Defined WAN** 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)*
3:00 pm - 3:25 pm	**On Low-Latency-Capable Topologies, and Their Impact on the Design of Intra-Domain Routing** Nikola Gvozdiev, Stefano Vissicchio, Brad Karp, Mark Handley *(UCL, UK)*
3:25 pm - 3:50 pm	**Asynchronous Convergence of Policy-Rich Distributed Bellman-Ford Routing Protocols** Matthew L. Daggitt *(Cambridge, UK)*, Alexander J. T. Gurney *(Comcast, USA)*, Timothy Griffin *(Cambridge, UK)*