Routing/Addressing Problem Solution Space

John Scudder
jgs@juniper.net

ARIN XX, October 17, 2007

Recap: What is the problem?

- **Problem: The routing table is growing**

- **I'll present current understanding of how to address this**
  - Probably incomplete
  - Certainly lacking in detail
  - Trying to identify tradeoffs
  - Focusing on near-term prospects
  - All IMHO

Overview: Options

- **Stay the course**
  - PI and hole punching for multihoming
  - Bigger hardware
  - Routing protocol evolution
- **Locator/ID split**
  - Network-based — e.g. LISP, 8+8/GSE
  - Host-based — e.g. Shim6, Six/One
- **Other options**
  - Different aggregation/deployment — e.g. geographic
  - Forbid PI, forbid multihoming
  - Clean slate

Stay the course — FIB size

- **Build bigger FIBs!**
- **Some hardware supports 1M+ routes now**
  - … and can be expected to scale up (~10M) within a few years if demand exists
- **But: wide deployment of “legacy” hardware with smaller FIBs**
  - … and big-FIB not available across all product segments
- **5+ year amortization cycles**

Stay the course — Control plane

- **Build bigger route engines!**
  - Similar issues as with FIB
- **Incrementally improve BGP**
  - Various proposals to improve stability, performance
  - Modest (~2-3x) improvements in update rate seem possible
  - No “magic bullet”, fundamental scaling properties stay the same
- **How does BGP degrade?**
  - Performance-wise: Gracefully (just slows down)
  - Memory-wise, ungracefully (falls over)

Stay the course evaluation

- **Pros:**
  - Same old, same old — well understood
  - Low short term risk — “get a bigger one” is a simple algorithm
- **Cons:**
  - Same old, same old — warts and all
  - Doesn’t enable new features and capabilities
  - Cost
  - Risk if hardware not shipping when needed
  - Long term risk difficult to quantify — because predicting the future is difficult
    - Sharp uptick in table growth rate would be a problem

Locator/ID split

Locator/ID split

“Any problem in computer science can be solved with another layer of indirection.”
—David Wheeler

Locator/ID split

"Any problem in computer science can be solved with another layer of indirection."
—David Wheeler

"But that usually will create another problem."
—rest of the quote

Locator/ID split [2]

- **Many proposals**
  - Too many to cover in detail
  - Representative examples in this talk
  - Example, not specific endorsement or criticism!
  - See Routing Research Group home page for much, much more
- **Network based (e.g., LISP)**
  - Premise: too hard to change hosts
- **Host based (e.g., Six/One)**
  - Premise: changing hosts can be done, now is the time (for v6), transition is easier

Locator/ID split [3]

- **Identifier**
  - Endpoint of a communication (a host)
  - Basically, a PI address
- **Locator**
  - Where in the topology the host is at the moment
  - Basically, a PA address
- **Currently, IP address is used both ways at the same time**
- **Why would splitting locator and ID help?**
  - Routers in the core use locators — which act like PA addresses
  - Pushes PI problem into a different component ("mapping service")

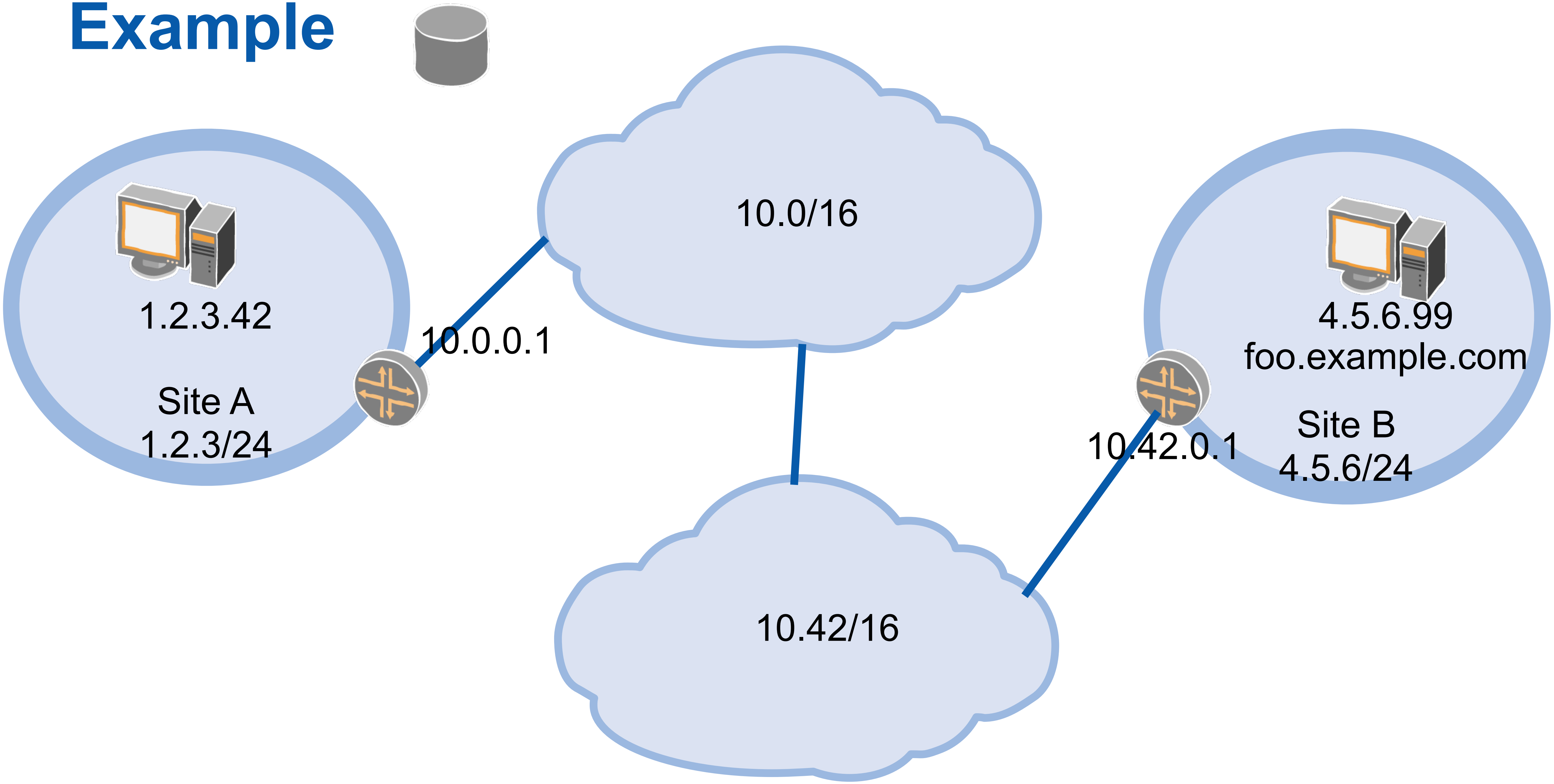
Carrying Identifiers and Locators

- **Hosts want to see identifiers**
- **Routers want to see locators**
- **So, need some way to have both in packets**
- **Map-n-encap (e.g. LISP)**
  - Host sends packet with IP header. IP address in header is an "identifier"
  - Edge router ("Ingress Tunnel Router" or ITR) adds a header with a "locator"
- **Map and rewrite (e.g. 8+8/GSE)**
  - Host sends IPv6 packet with identifier in lower 8 bytes
  - Router writes locator into upper 8 bytes
  - Hosts have to ignore content of upper 8 bytes as it may be changed by routers

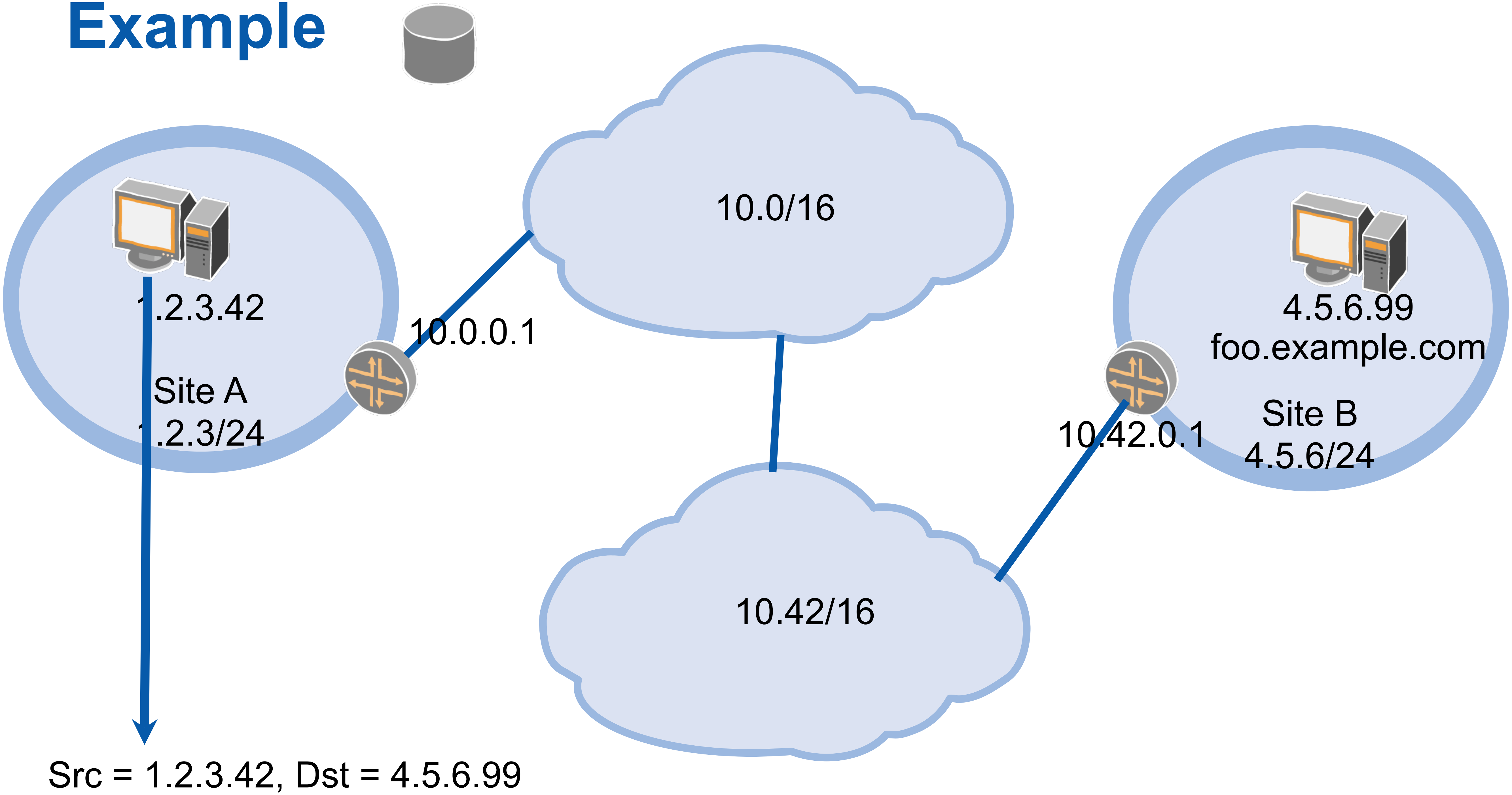
Getting Mappings

- **Ingress Tunnel Routers receive packets with identifier addresses**
  - need to associate with locator addresses
- **Do this by looking up identifier in a “mapping service”**
- **Details of the mapping service are**
  - Contentious
  - Under development, many proposals
  - Not well understood yet
  - Crucially important

Example

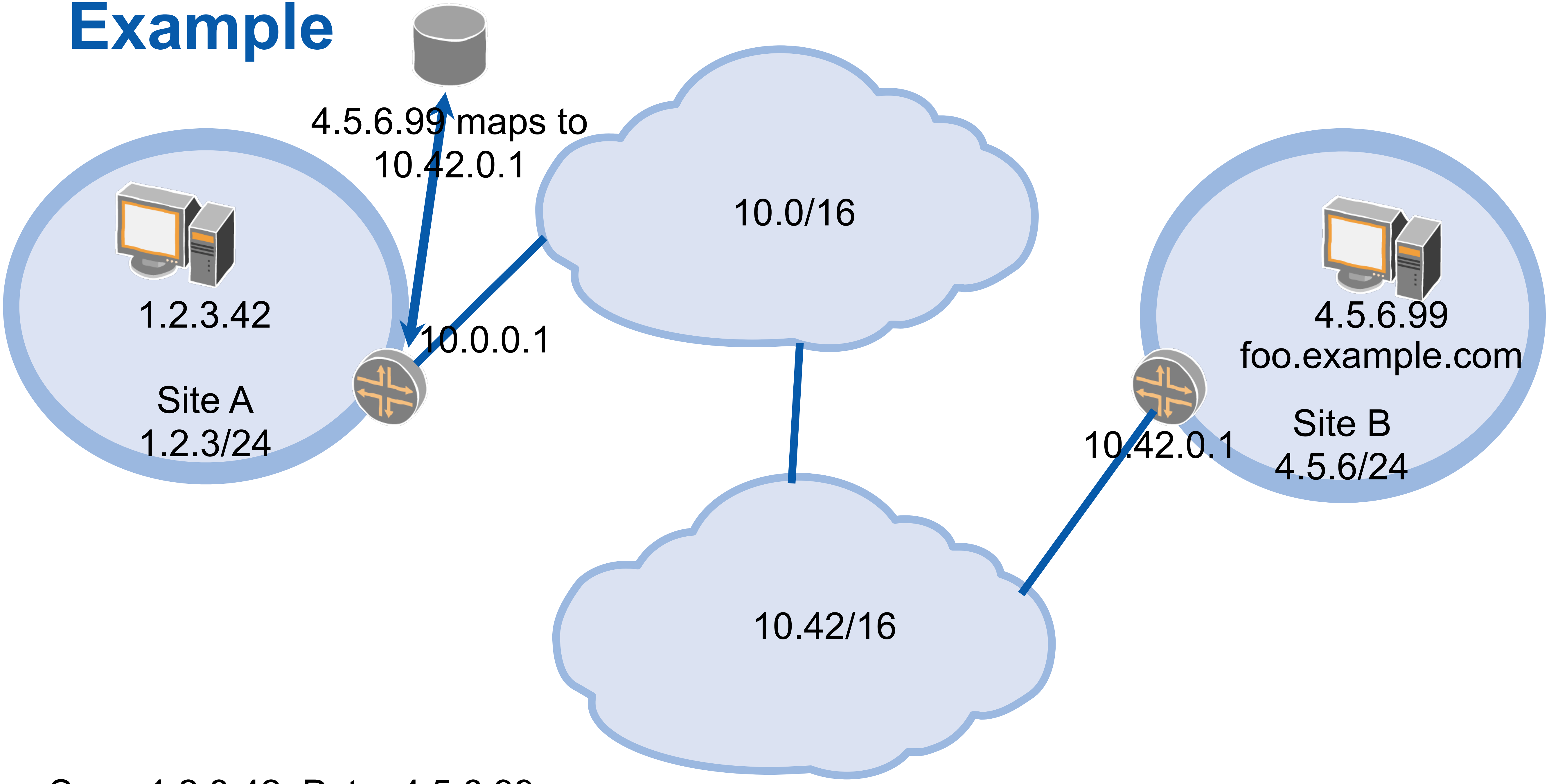

10.0/16
1.2.3.42
Site A
1.2.3/24
10.0.0.1
4.5.6.99
foo.example.com
Site B
4.5.6/24
10.42.0.1
10.42/16

Example



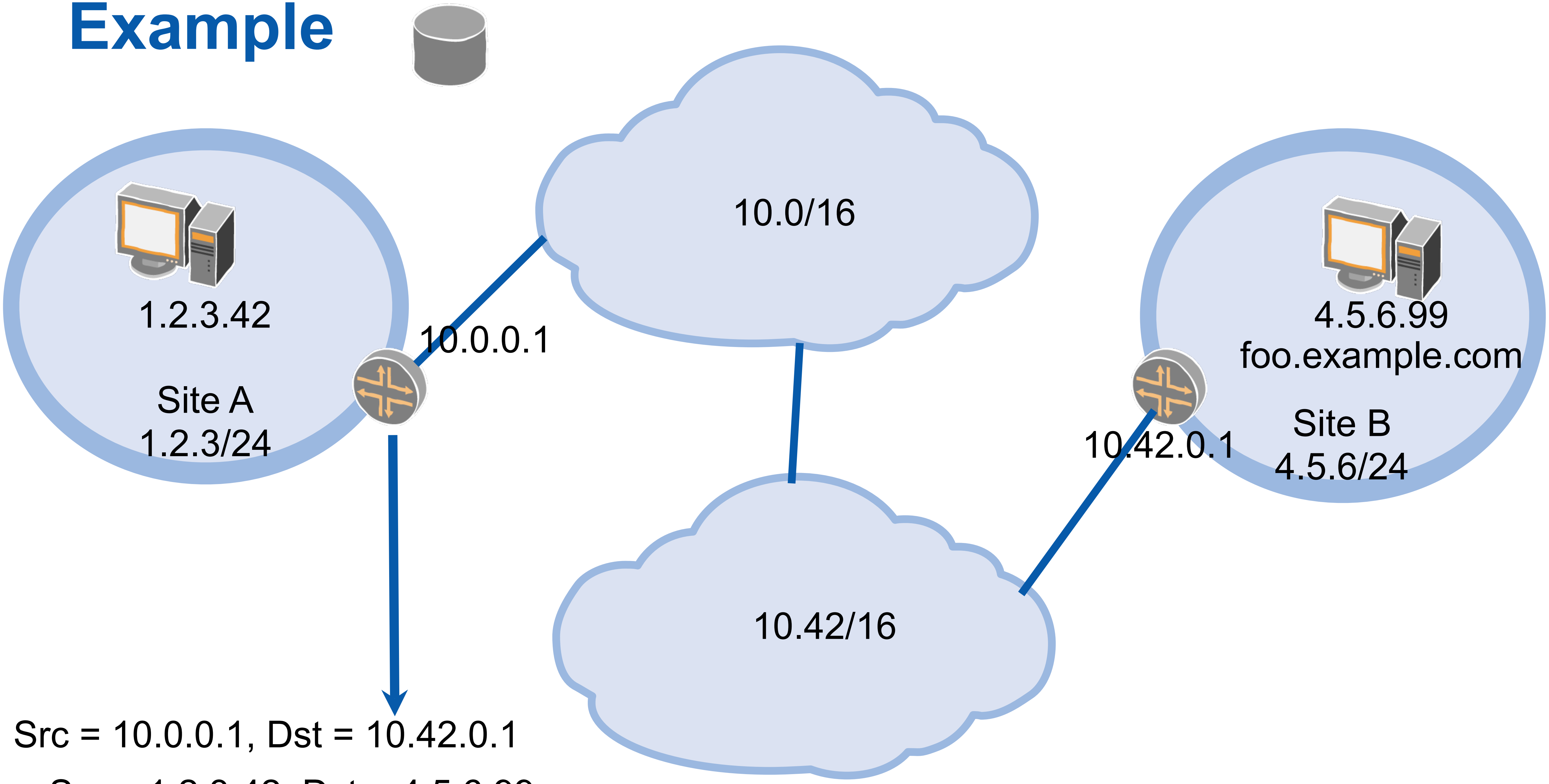
Src = 1.2.3.42, Dst = 4.5.6.99

Example



Src = 1.2.3.42, Dst = 4.5.6.99

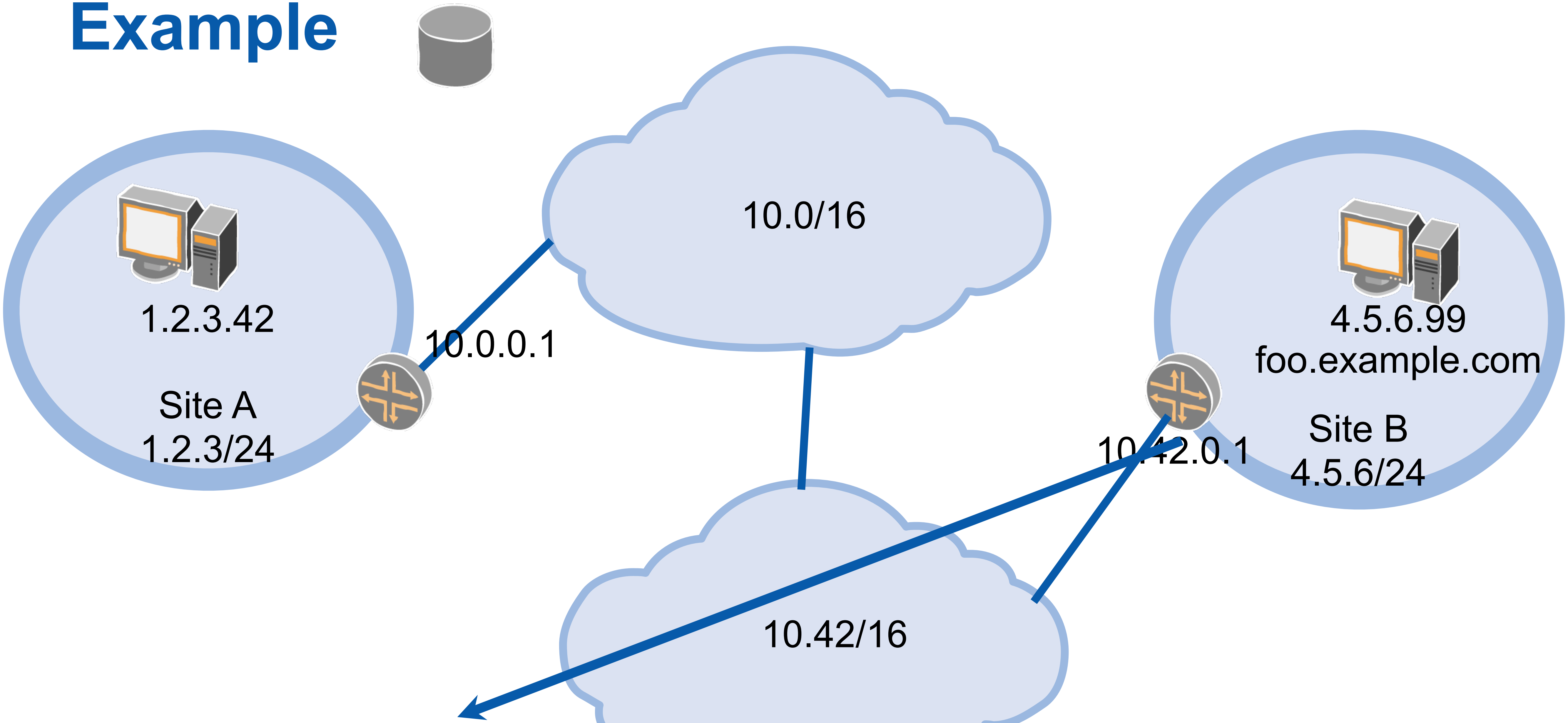
Example



Src = 10.0.0.1, Dst = 10.42.0.1

Src = 1.2.3.42, Dst = 4.5.6.99

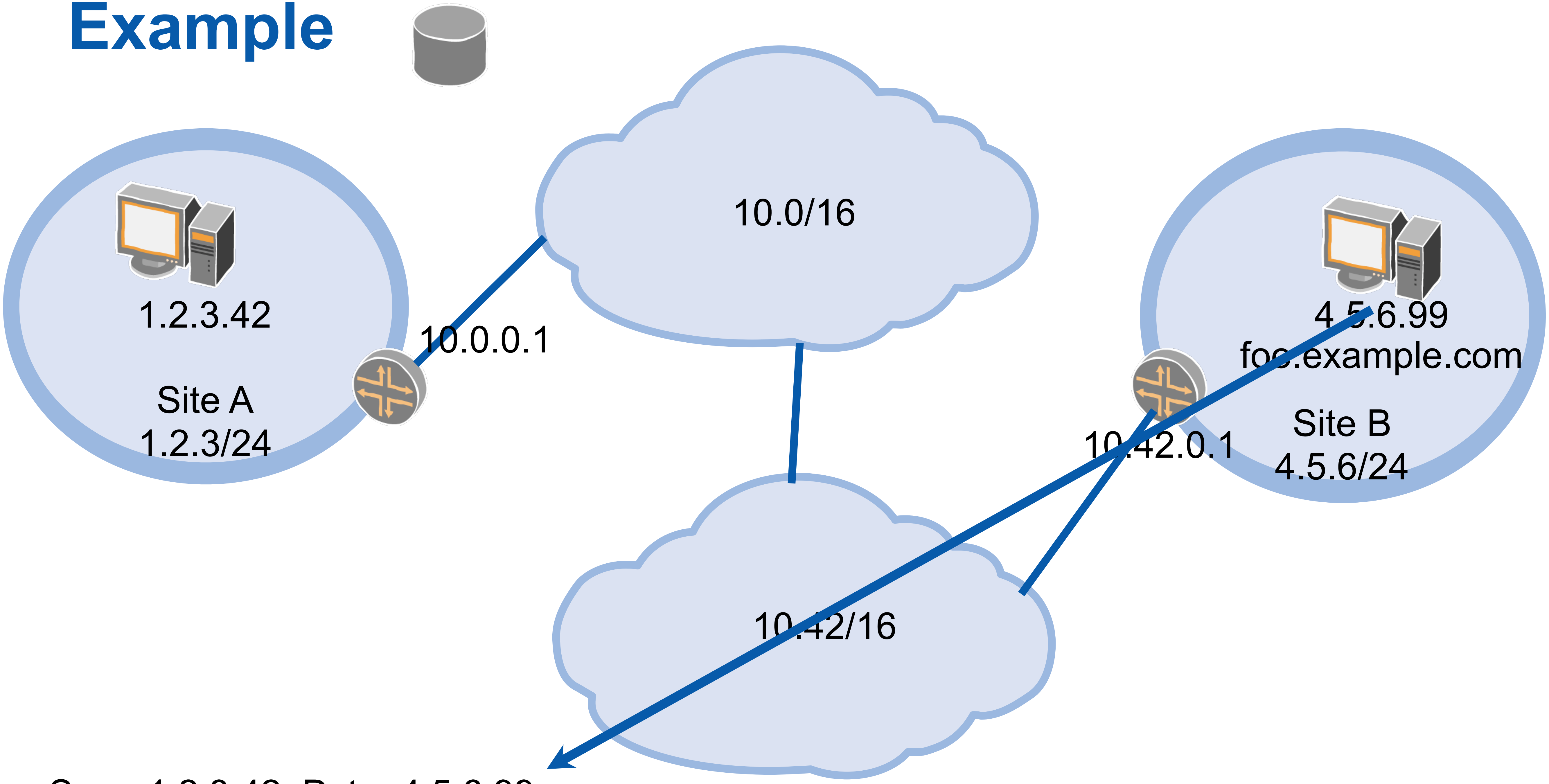
Example



Src = 10.0.0.1, Dst = 10.42.0.1

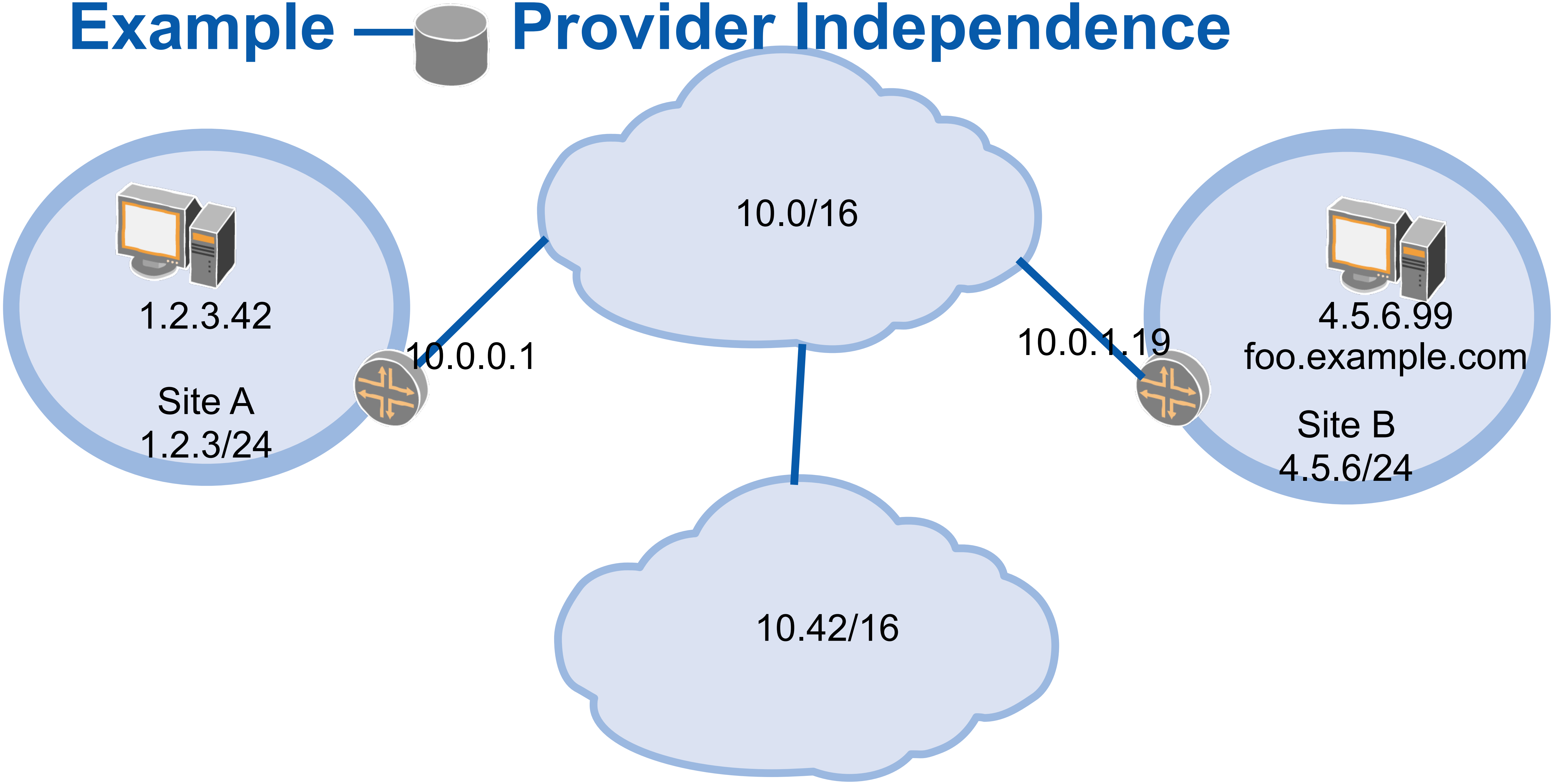
Src = 1.2.3.42, Dst = 4.5.6.99

Example

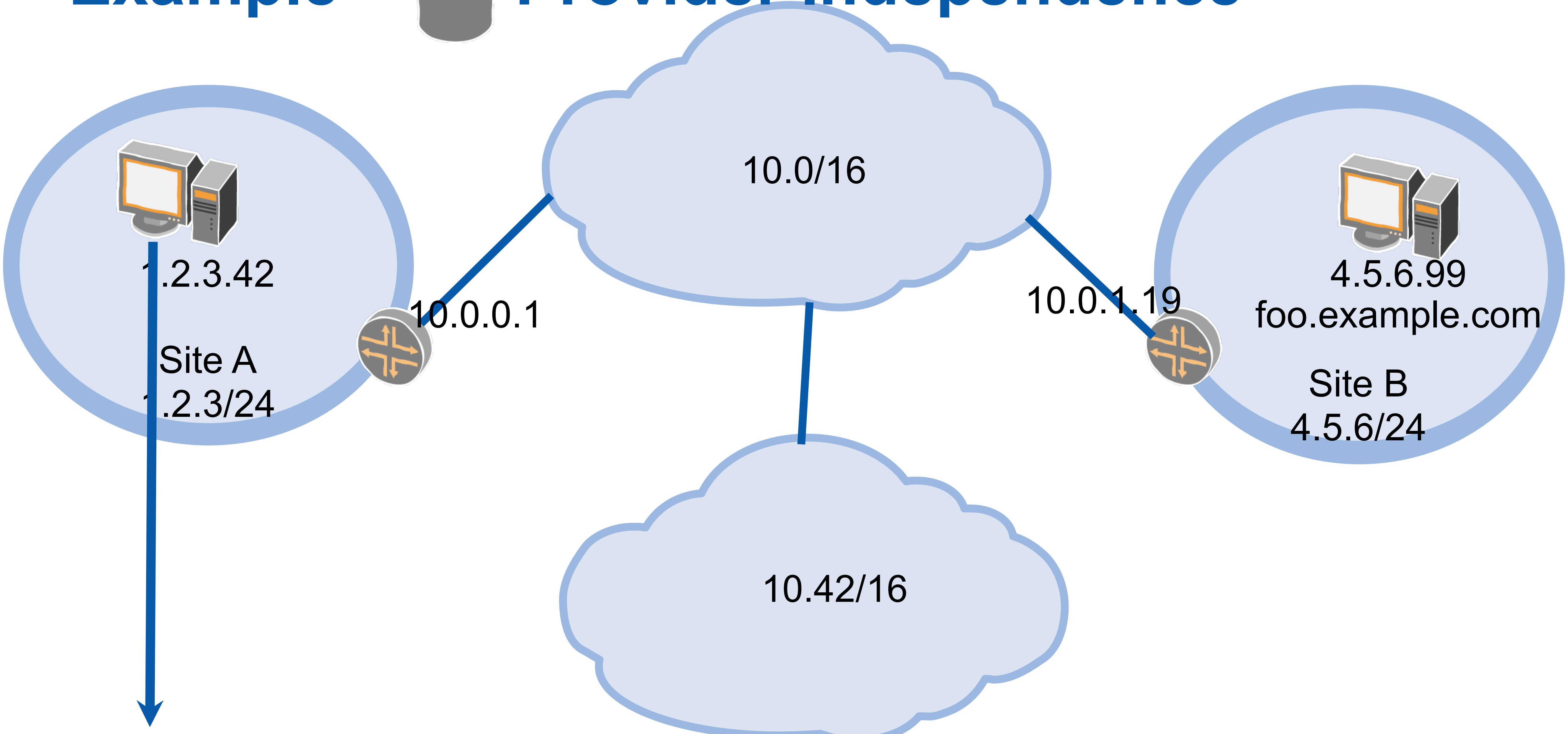


Src = 1.2.3.42, Dst = 4.5.6.99

Example — Provider Independence

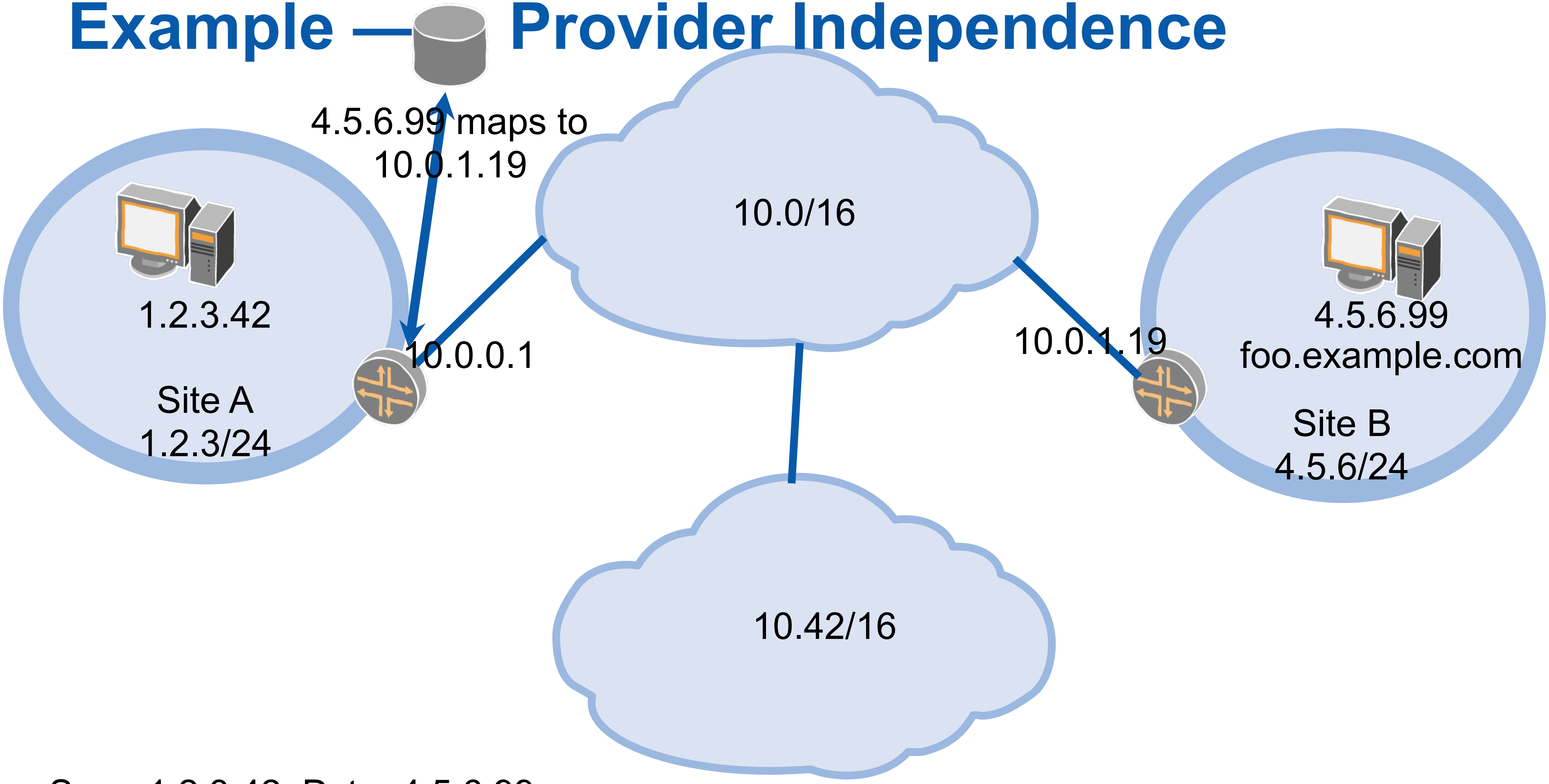


Example — Provider Independence



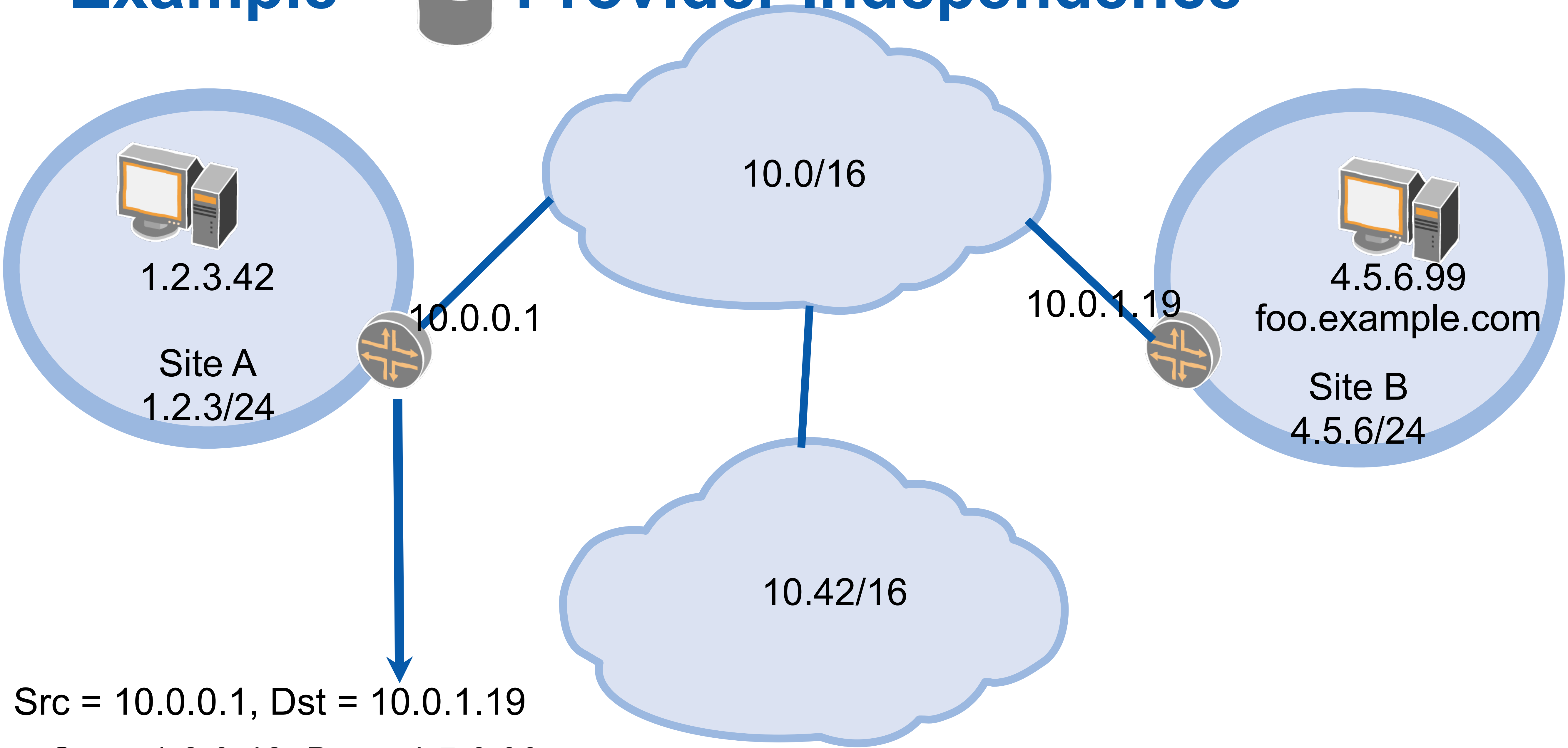
Src = 1.2.3.42, Dst = 4.5.6.99

Example — Provider Independence



Src = 1.2.3.42, Dst = 4.5.6.99

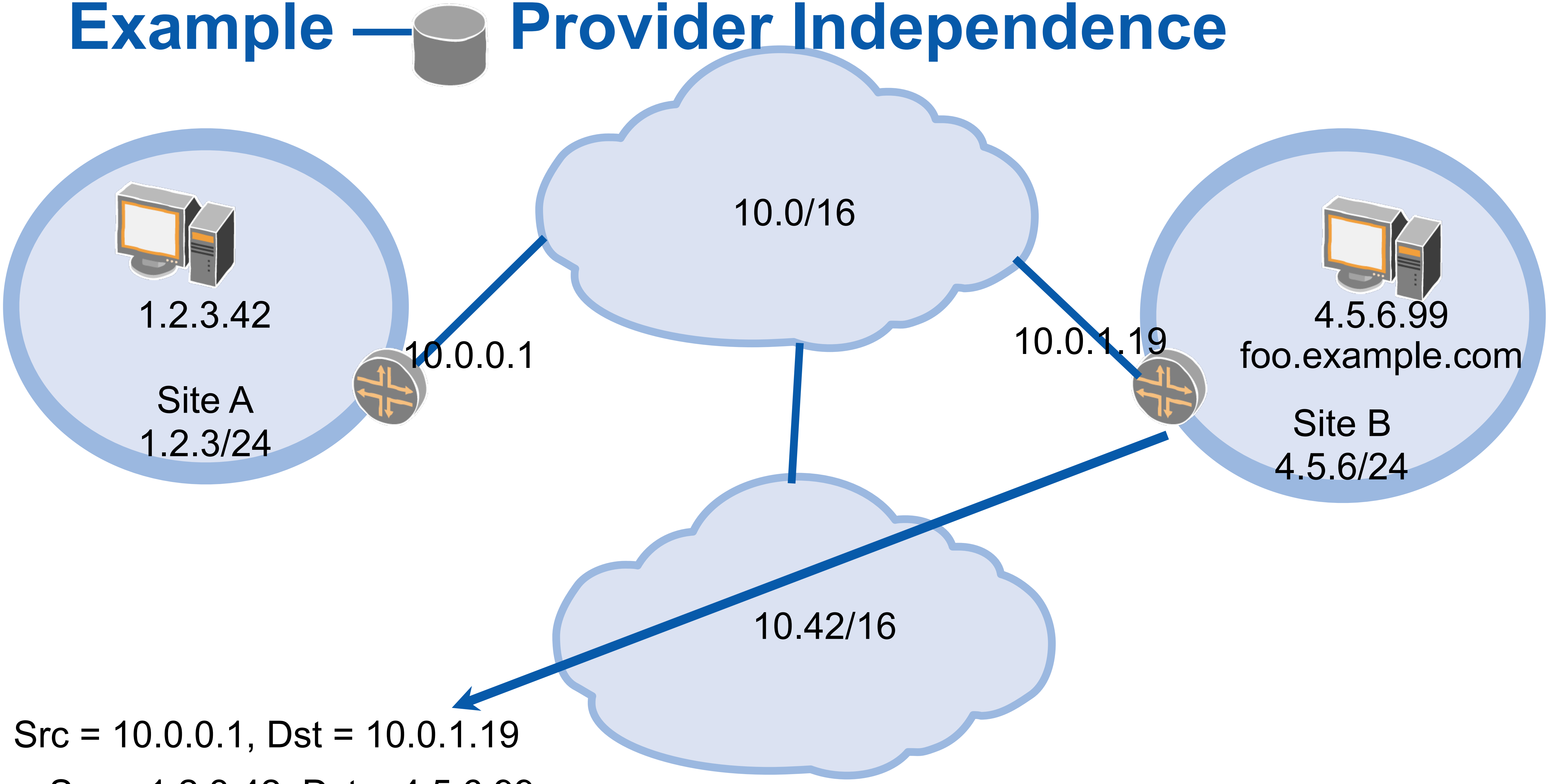
Example — Provider Independence



Src = 10.0.0.1, Dst = 10.0.1.19

Src = 1.2.3.42, Dst = 4.5.6.99

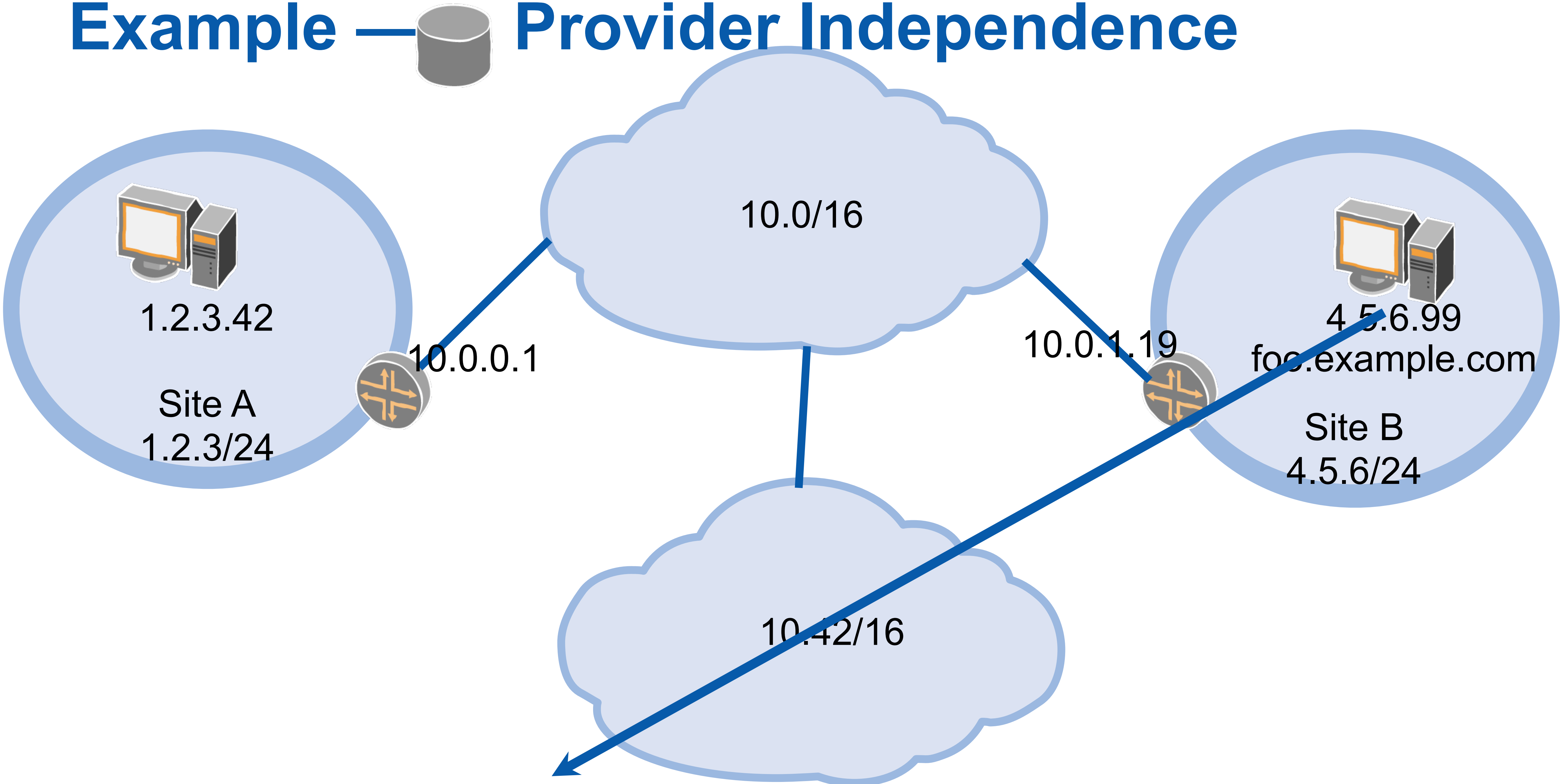
Example — Provider Independence



Src = 10.0.0.1, Dst = 10.0.1.19

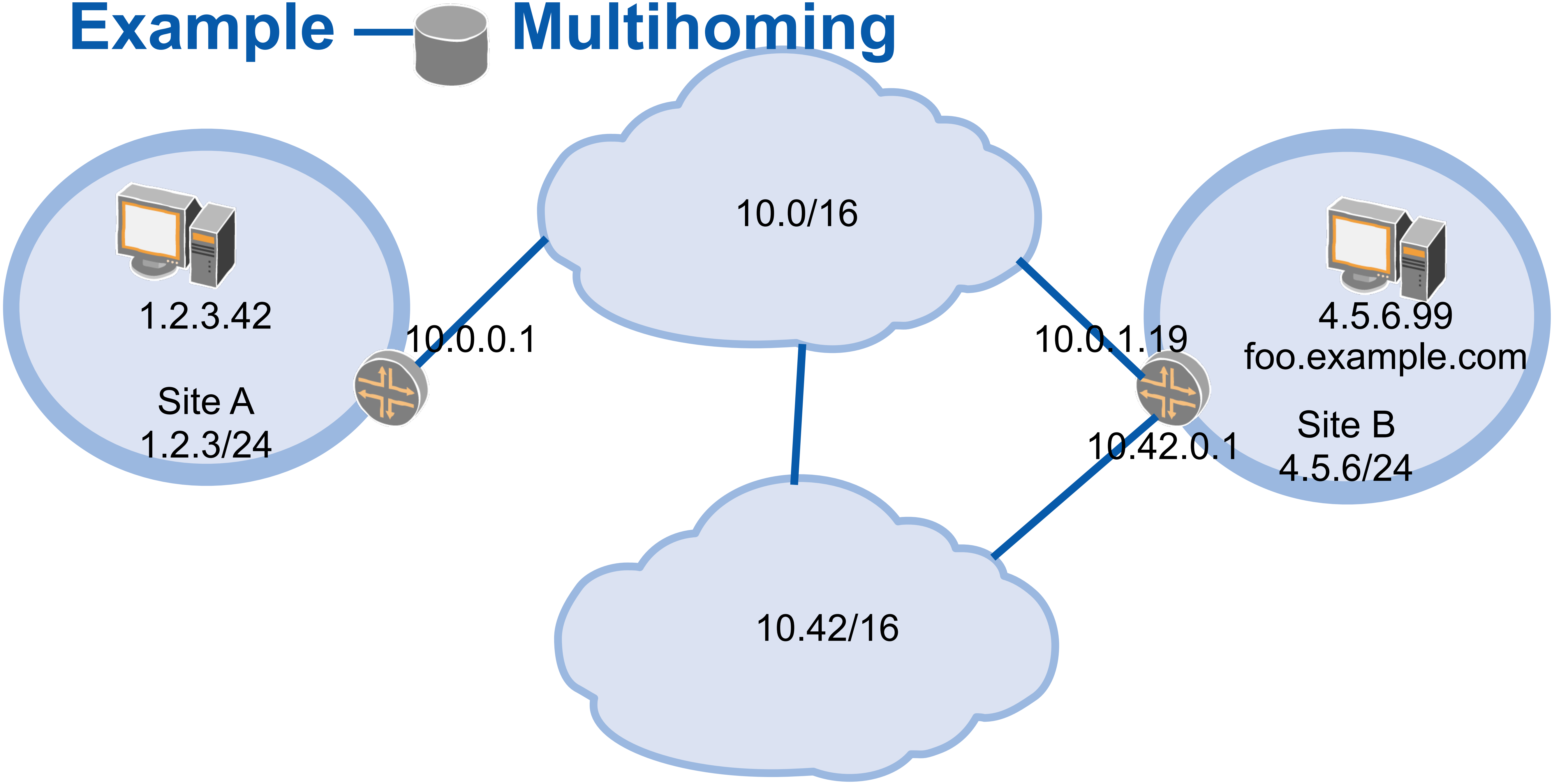
Src = 1.2.3.42, Dst = 4.5.6.99

Example — Provider Independence

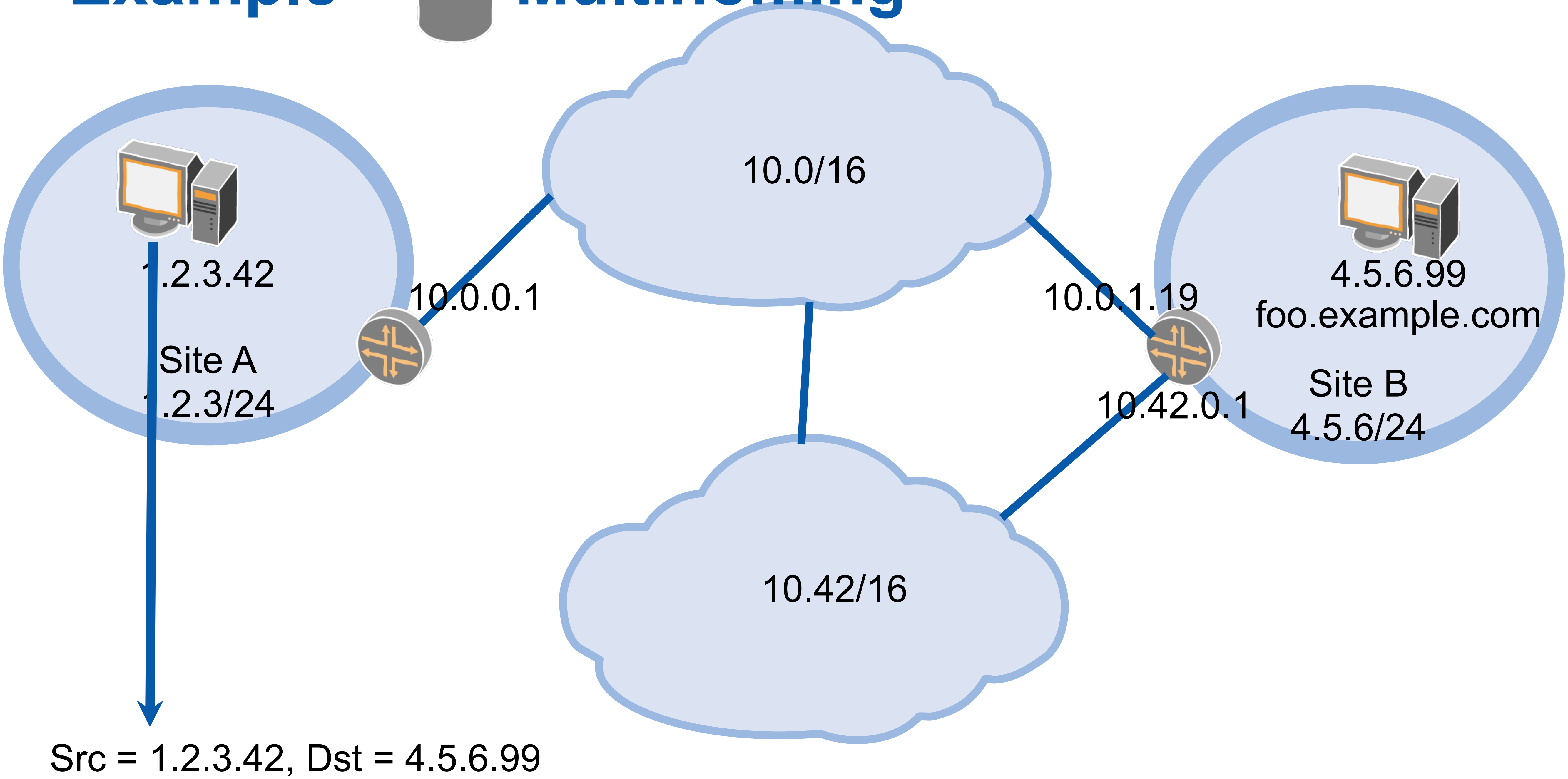


Src = 1.2.3.42, Dst = 4.5.6.99

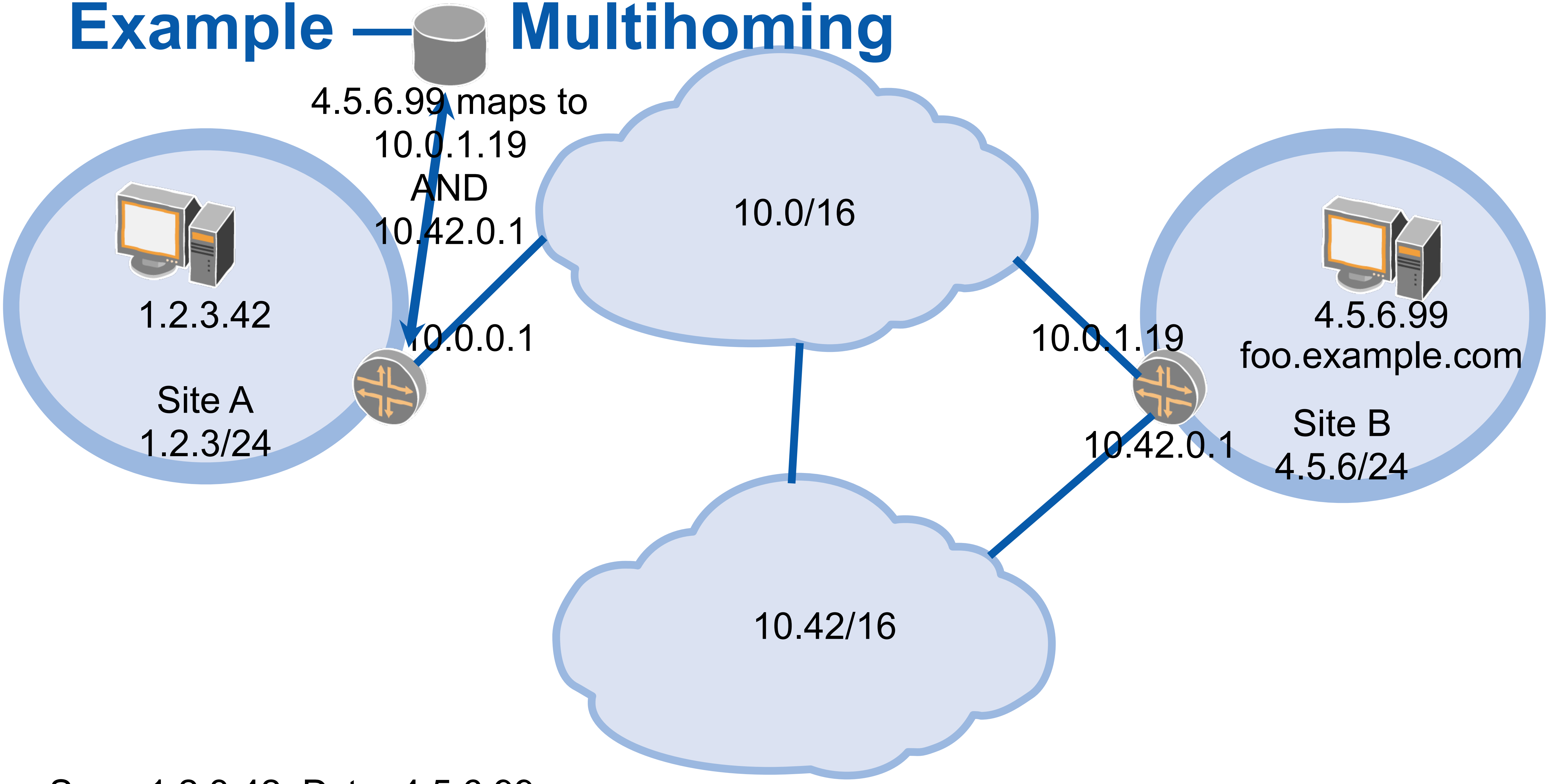
Example — Multihoming



Example — Multihoming

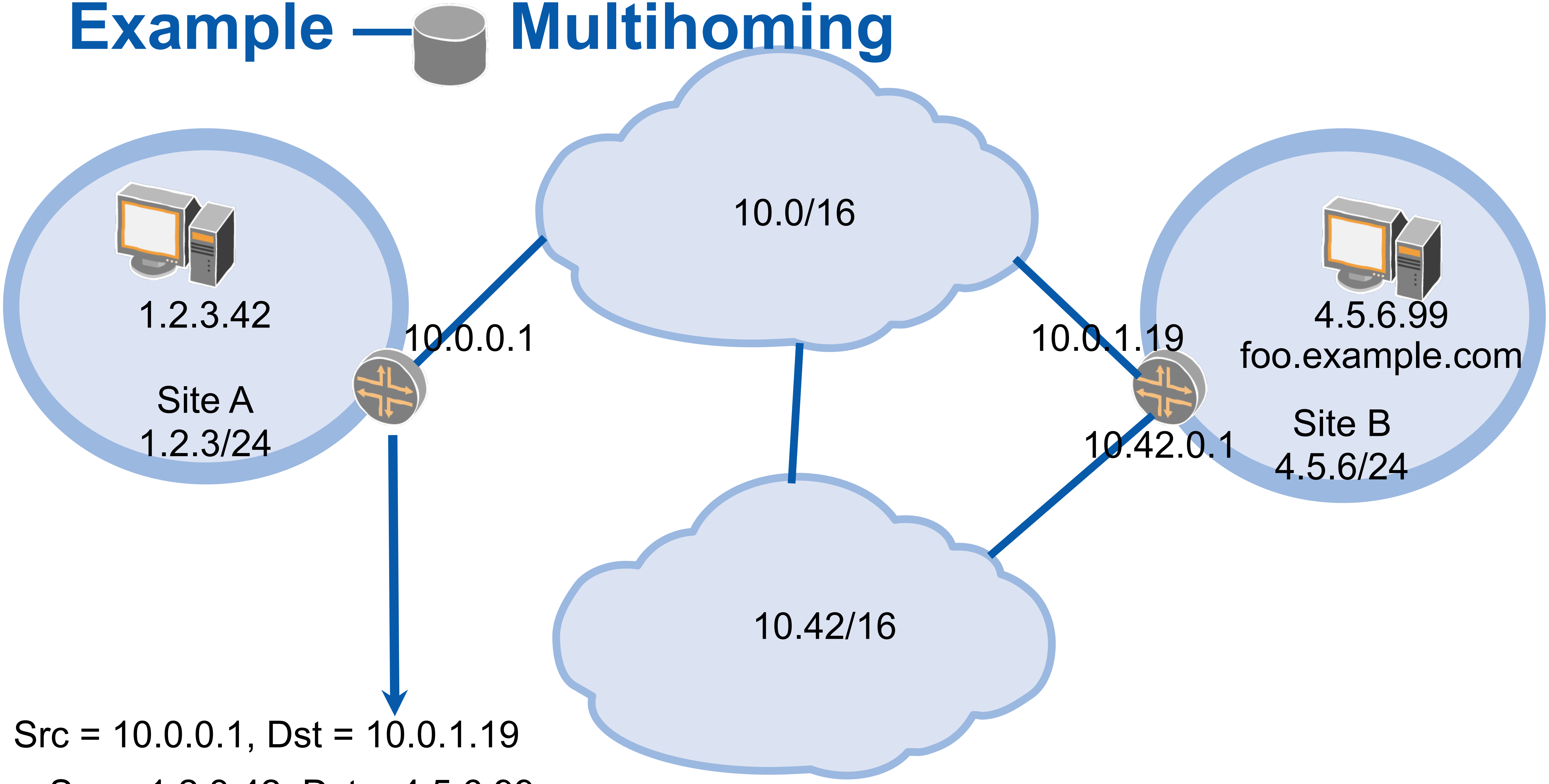


Example — Multihoming



Src = 1.2.3.42, Dst = 4.5.6.99

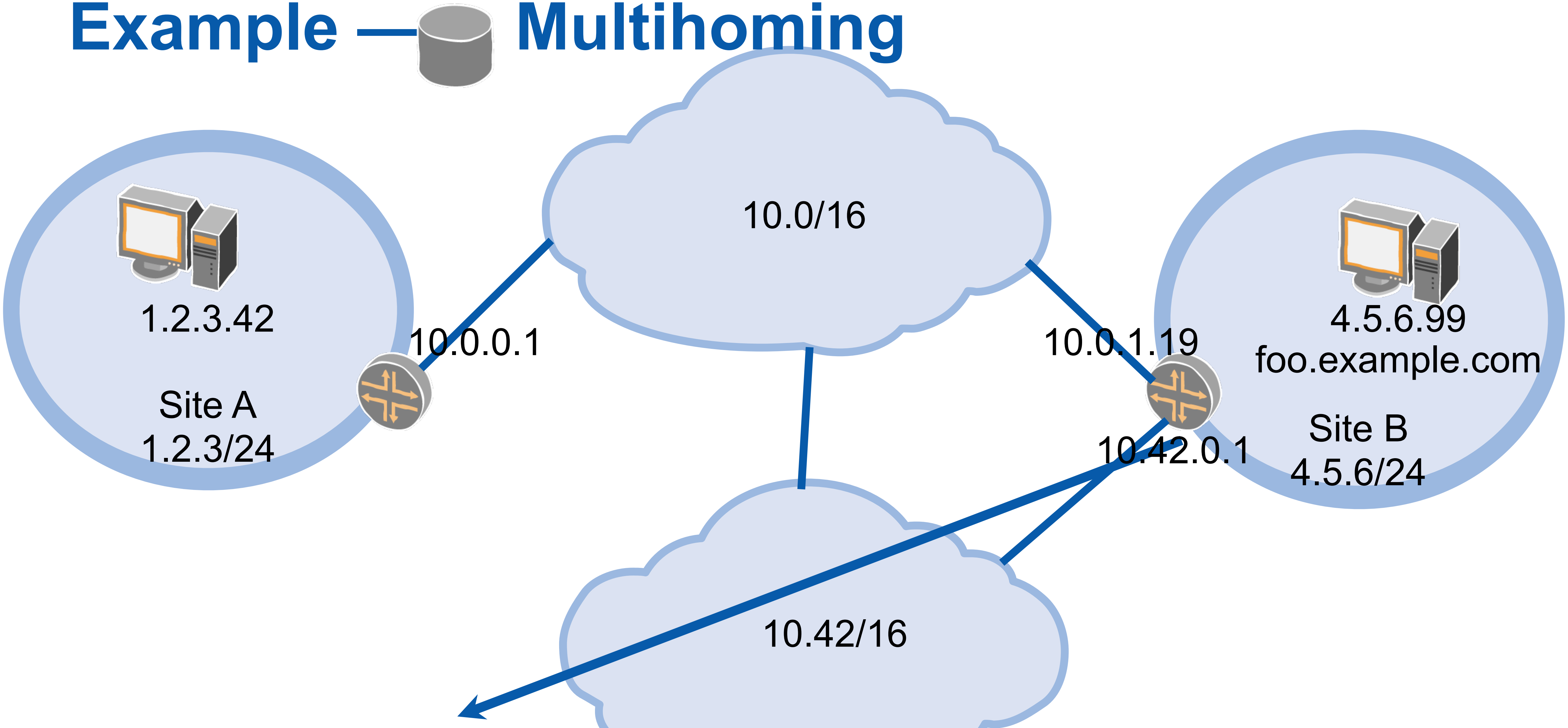
Example — Multihoming



Src = 10.0.0.1, Dst = 10.0.1.19

Src = 1.2.3.42, Dst = 4.5.6.99

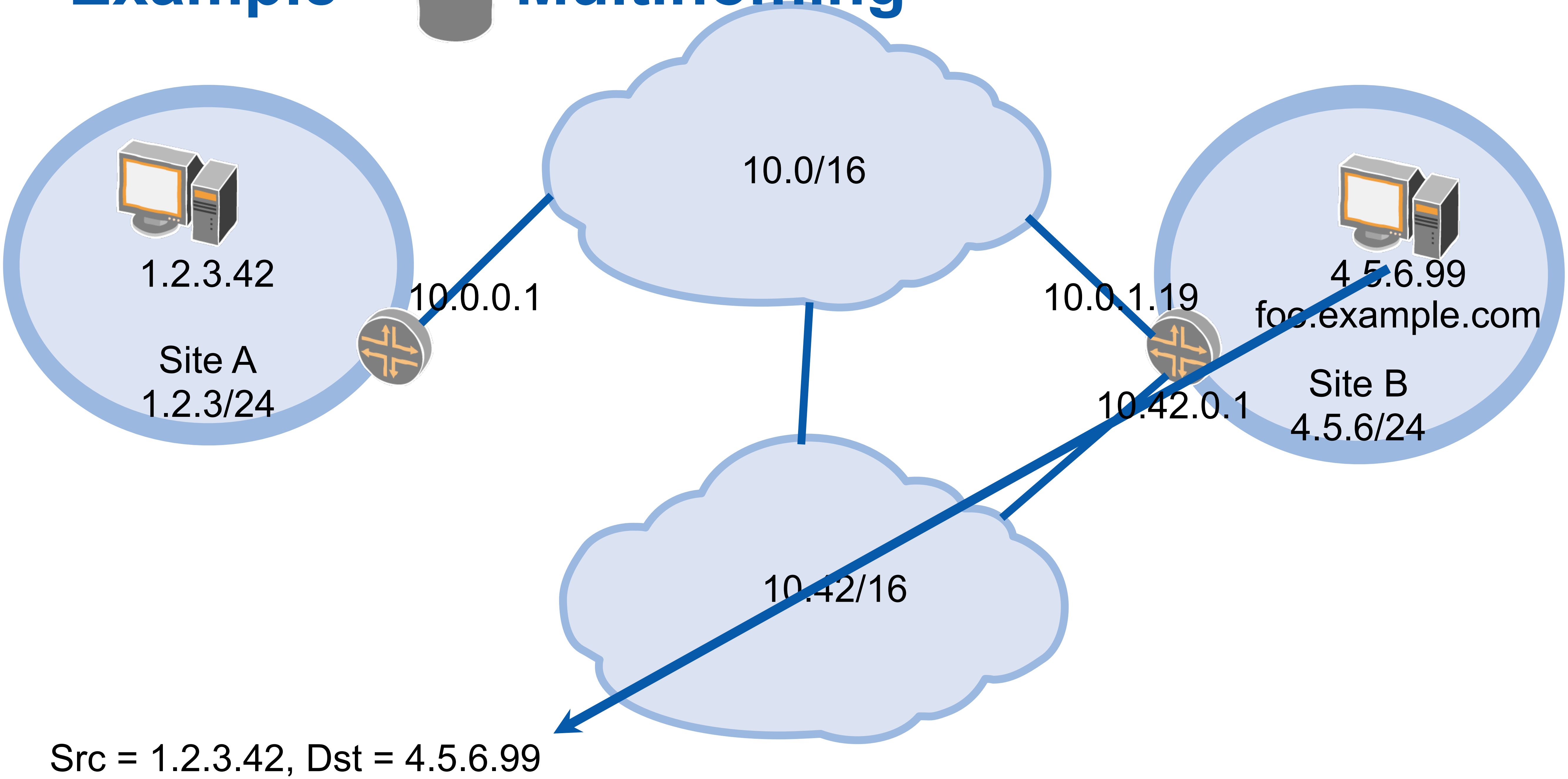
Example — Multihoming



Src = 10.0.0.1, Dst = 10.0.1.19

Src = 1.2.3.42, Dst = 4.5.6.99

Example — Multihoming



Src = 1.2.3.42, Dst = 4.5.6.99

Traffic Engineering

- **Compared to current BGP based multihoming/TE:**
- **Destination site has about the same capabilities**
  - “Prefer to reach me this way”
  - “Load share across both attachments”
- **Source site gains more capabilities**
  - Can override destination site policy
- **ISP loses out**
  - Since destination identity isn’t exposed to ISP network

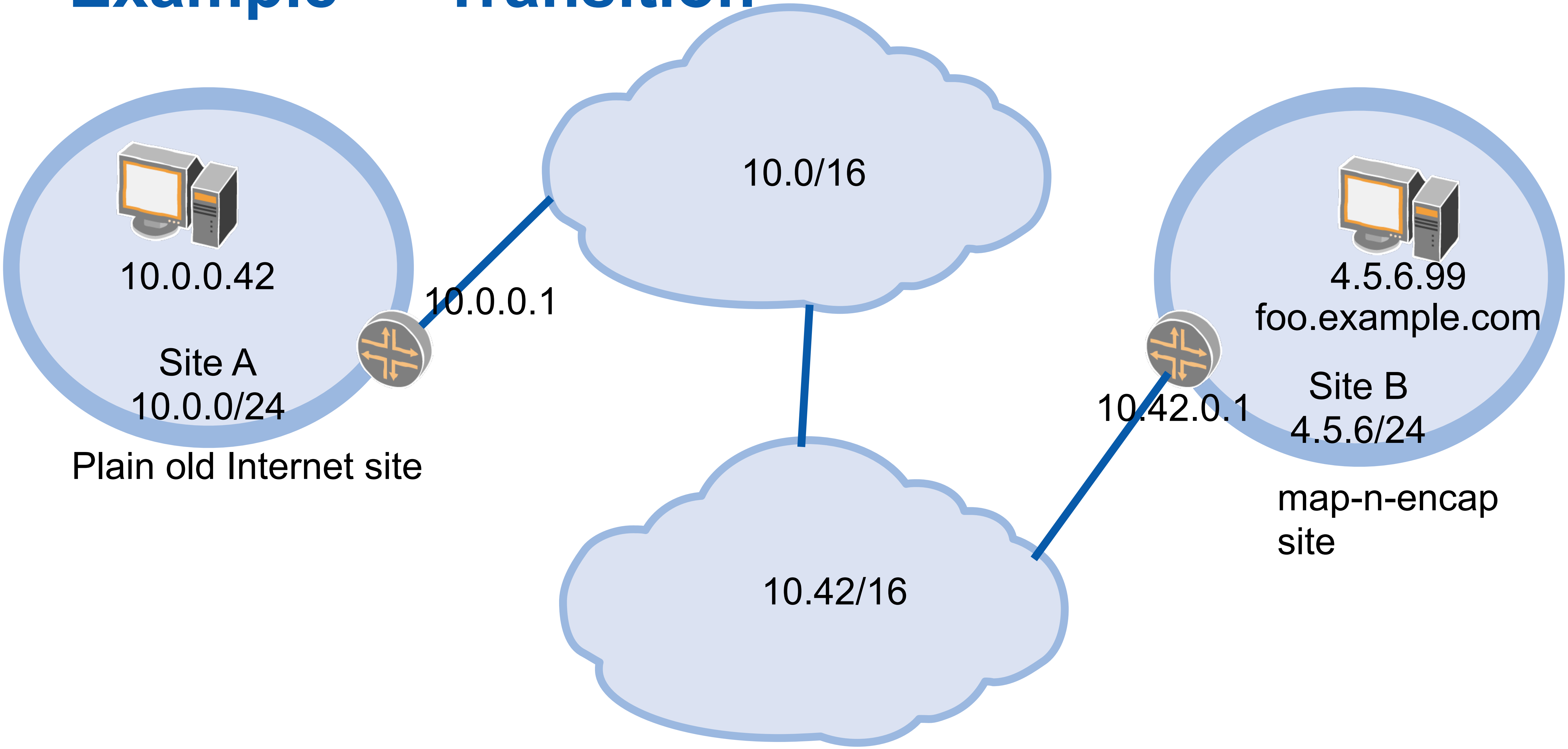
Detecting Failures

- **Currently: control plane signals failures**
  - Multihomed network loses attachment
  - Route is withdrawn from BGP
  - So nobody tries to send packets that way
- **Locator/ID: no failure signaling in control plane**
  - Multihomed network loses attachment
  - Packets are sent that way anyway
  - Rely on ICMP or similar to learn about failure
- **Control-driven vs. data-driven**
- **Implications not well understood**

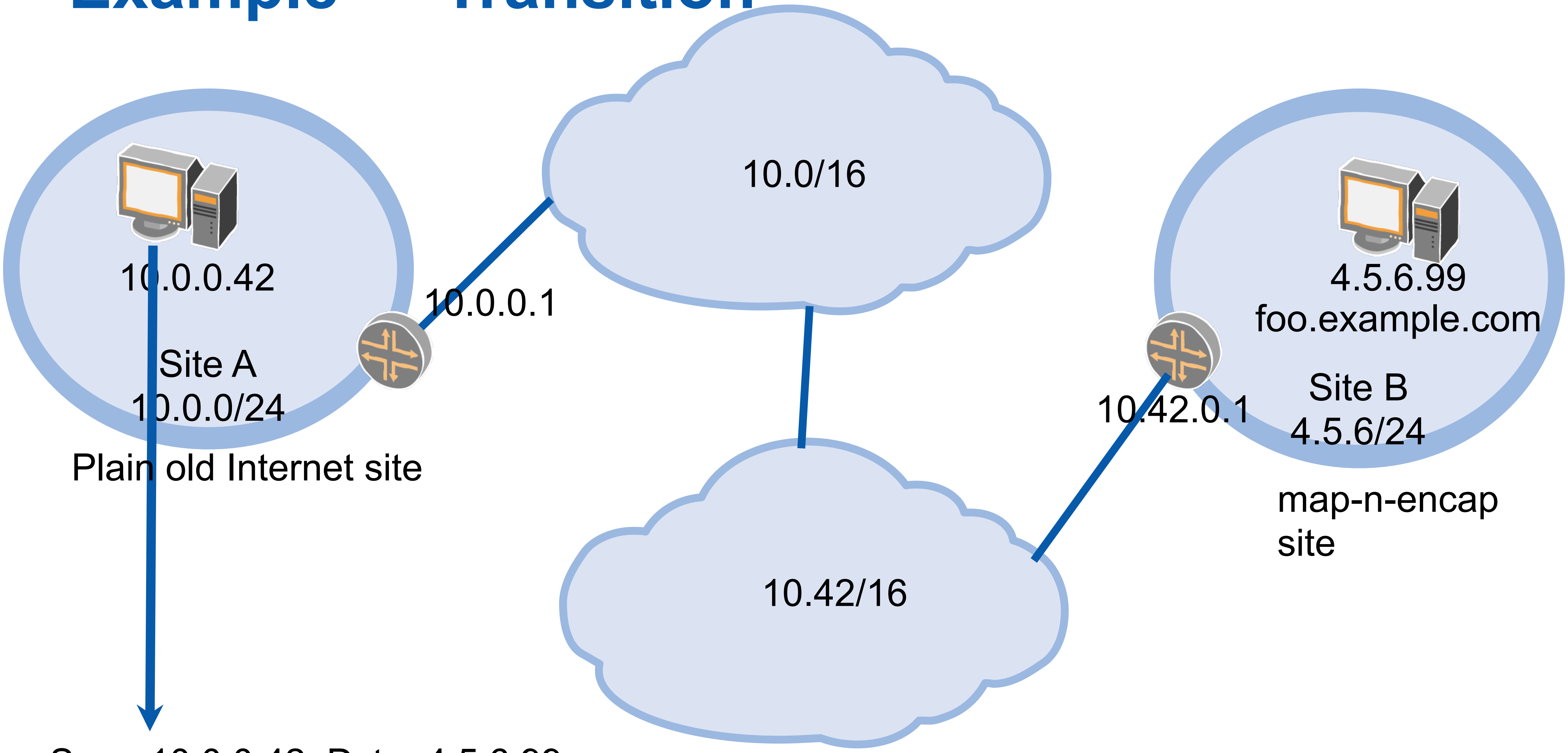
Mapping Database

- **Pull model**
  - Ingress routers query external mapping servers and cache results
  - Reduces state on ingress routers
  - Adds latency, reduces performance
- **Push model**
  - Full mapping database replicated on every ingress router
  - But mapping database likely much larger than current routing table!
  - Did we gain anything?
- **Hybrid approaches possible (e.g. LISP-CONS)**

Example — Transition

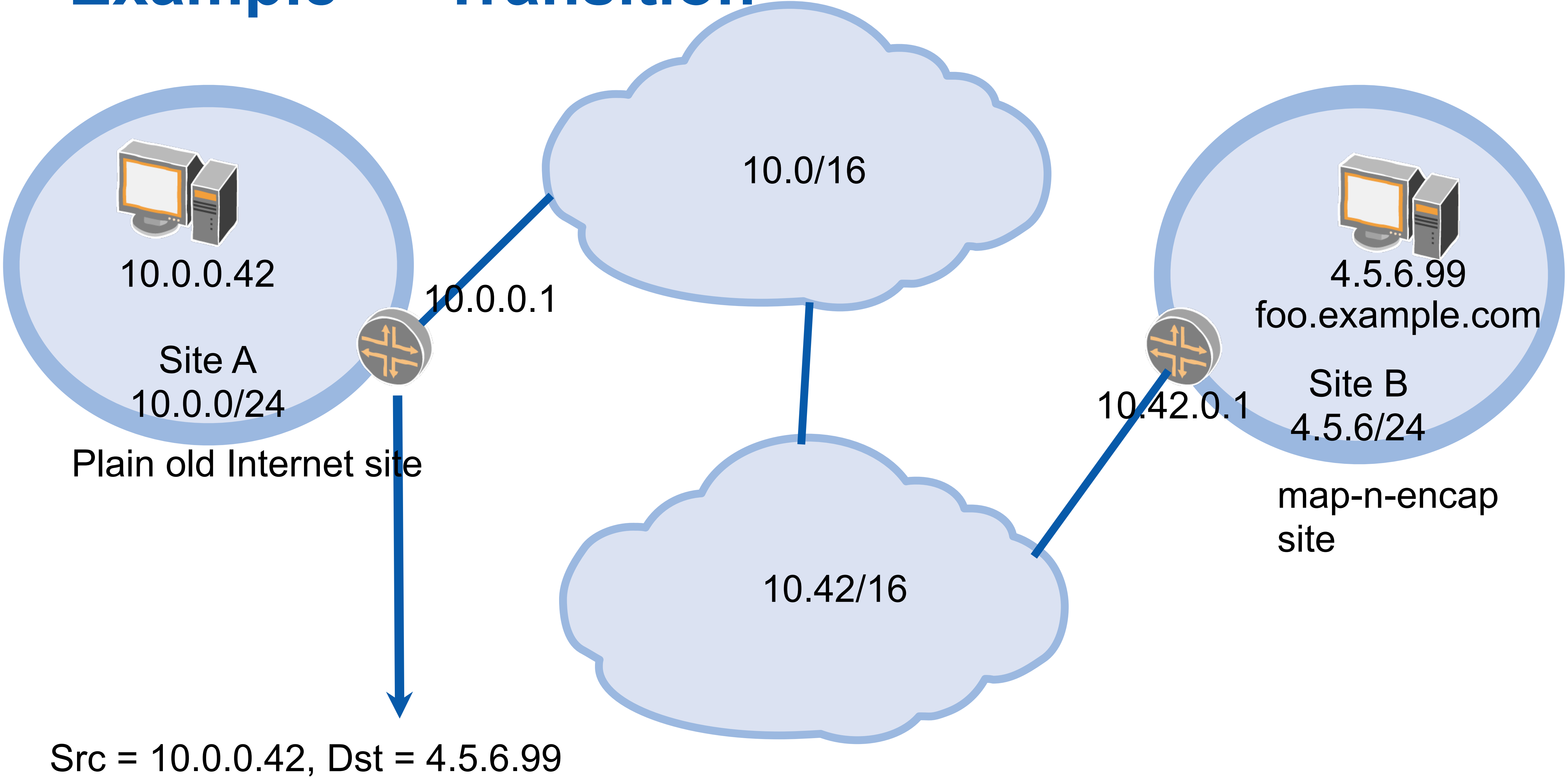


Example — Transition

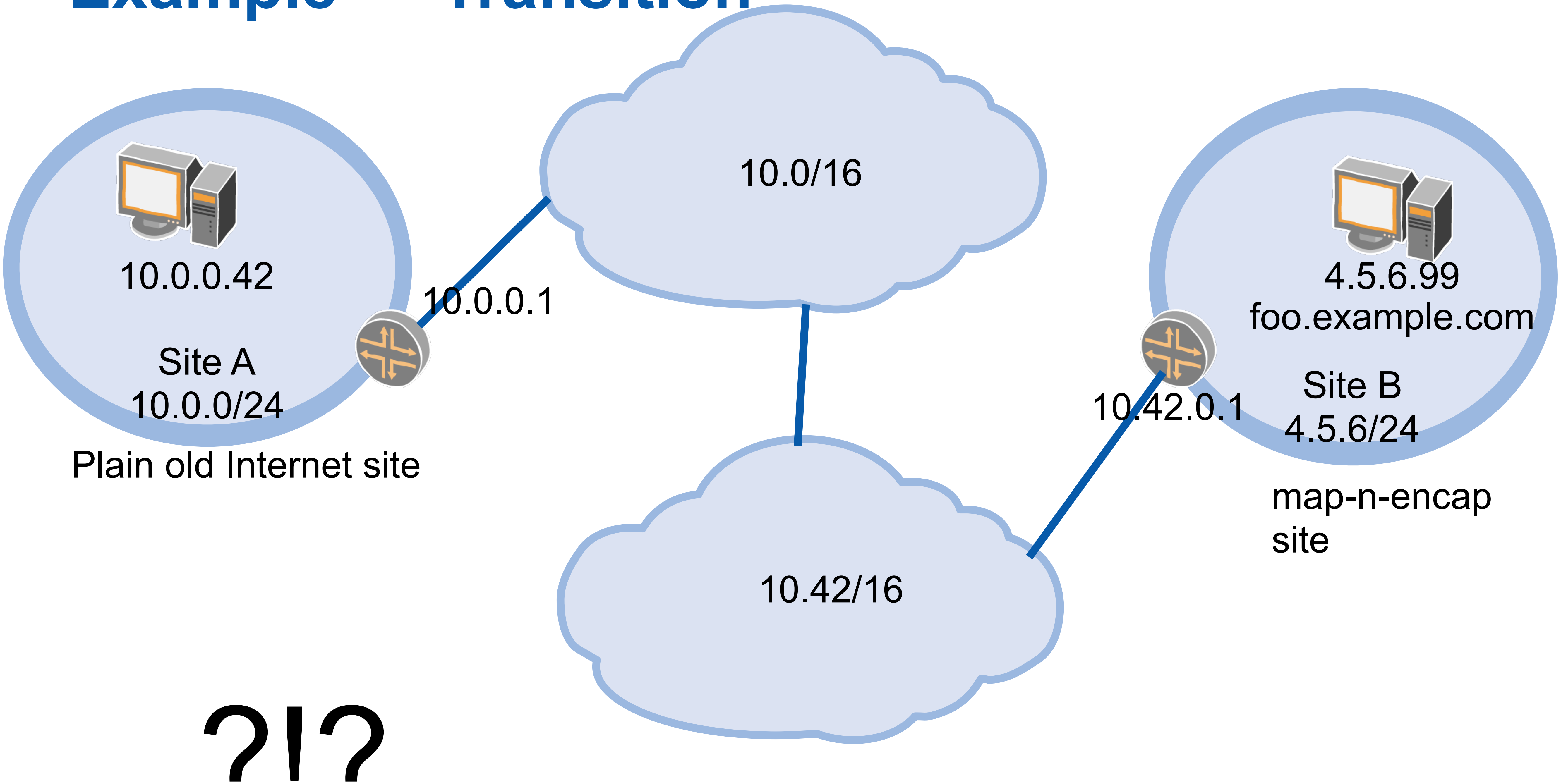


Src = 10.0.0.42, Dst = 4.5.6.99

Example — Transition



Example — Transition



Plain old Internet site

map-n-encap
site

?!?

Network-Based Locator/ID evaluation

- **Pros:**
  - Core routing scales very well
  - Enables increased use of multihoming
  - More flexible traffic engineering
  - May enable denser address space utilization
    - Pushing out IPv4 depletion

Network-Based Locator/ID evaluation [2]

- **Cons**
  - Ingress routers might scale not-so-well if using “push”
  - … or suffer performance problems if using “pull”
  - Potential performance issues — “pull” mapping, tunneling (MTU issues, tunnel overhead), data-driven failure detection, etc
  - Security not well understood
  - Mapping service not well understood, scaling unknown
  - Providers lose TE capabilities
  - No satisfactory transition plan
  - Still in research phase
  - Cost

Host-Based Locator/ID

- **Example: Shim6**
- **Host stack has concept of locator and identifier**
  - By dividing address into low/high bytes a la 8+8/GSE
  - Or by some kind of encapsulation (or “shim”)
- **Network addressing is all PA**
  - Host selects source address (“locator”)
  - Host selects destination address (“locator”)
  - Locators can change during communication
- **Doesn’t address renumbering**
  - Which is one motivation for PI
- **Host makes all traffic engineering decisions**
  - No network control — could be fixed in principle

Host-Based Locator/ID [2]

- **Network based rewriting, e.g. Six/One**
  - Like in 8+8/GSE
- **Fixes some problems**
  - Network can make traffic engineering decisions
  - Renumbering can be supported
- **Incremental transition**
  - If both hosts support host-based locator/ID, use it
  - Otherwise, fall back to regular IP communication
  - But, if not supported, multihoming and TE functionality are degraded

Host-Based Locator/ID evaluation

- **Pros**
  - Core routing scales very well
  - Enables increased use of multihoming
  - More flexible traffic engineering
  - Some hope of incremental transition
- **Cons**
  - Current proposals just IPv6
  - Requires host changes
  - Providers lose TE capabilities
  - Really provide enough benefit to stamp out PI?
  - Still in research phase
  - Cost

Other Options [1]

- **Geographical Addressing/Aggregation**
  - Providers in an area address customers from a geography (e.g., metro) specific block
    - Not from a provider-specific block like today
  - Providers only announce aggregate outside geographical area
  - Providers peer with each other within geographical area to exchange traffic
- **Many proposals for this over the years**
  - Never seem to go anywhere

Geographical Addressing evaluation

- **Pros**
  - Aggregates well, allows PI and multihoming within area
  - No new router hardware or software needed
  - Can be complimentary to other solutions
    - Not one-size-fits-all

Geographical Addressing evaluation [2]

- **Cons**
  - Business model different from current, substantial new coordination and business processes needed
    - Participating providers must structure networks according to geographical scheme
    - Participating providers must peer in each metro
  - Traffic engineering doesn’t work so well
    - Because current TE involves advertising more-specific
  - Not attractive for customers spanning multiple geographies
  - Works best for customers who don’t need PI anyway

Other Options [2]

- **Clean Slate**
  - Catch-all for “anything not covered here”
  - Especially, anything not incrementally deployable
  - Pros: “anything is possible”
  - Cons: but you can’t deploy it
- **Forbid PI, forbid multihoming**
  - No PI, no multihoming… no route table scaling issues!
    - Because perfect aggregation possible
  - Pros: never upgrade your routers again (sort of)
  - Cons: appears unacceptable to customers

Summary

- **Stay the course — scale up hardware, protocols**
  - Development, deployment cycles relatively short
  - Capex high, opex low (relatively speaking)
  - Miracles unlikely
- **Locator/ID or other architectural magic**
  - Development, deployment cycles long (my guess: 5+ years, best case)
  - Capex low (maybe), opex high (maybe)
  - Key issues still unsolved
- **Other approaches exist**
  - But require tradeoffs on PI, multihoming, TE

Conclusion

- **Current architecture will be with us for a while**
  - Upgrade cycles, like it or not
  - Continued planning required
  - Continued management of routing table growth rate required
- **Locator/ID research is promising**
  - But many open questions remain
  - Contributions very welcome
  - Routing Research Group meeting at Vancouver IETF
  - Mailing list: rrg-request@psg.com
  - http://www.irtf.org/charter?gtype=rg&group=rrg
- **Did I mention this is all IMHO?**

Juniper your Net™