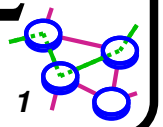
CS551
Delayed Internet Routing Convergence
[Labovitz00]

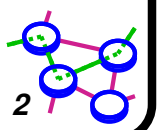
Bill Cheng

http://merlot.usc.edu/cs551-f12



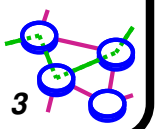
Context

- BGP widely deployed in the Internet
  - but poorly understood
  - ISP's don't tell you what they are doing
- BGP Problems: *Delayed Convergence*
- Question:
  - How long does it take for a route to *fail-over*?
- How to answer this question:
  - experimental methodology
  - explanation of observation using simple model



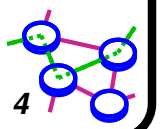
Key Idea

- Convergence time takes longer than we expected
- Observes 2-3 minute convergence times (6x longer than expected), BGP timer goes off every 30 seconds
- Study and understand BGP convergence time
  - simulation
  - measurement
- Suggests bounds of O(n!) worst case for BGP convergence, O((n-3)*30s), where n is the number of AS's



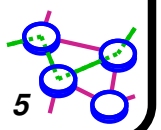
Why Is Convergence Important?

- **Robustness**
  - **PSTN (telephone) fail-over times are in milliseconds**
  - **Internet fail-over times are in 10s of seconds**
    - **open problem: how can Internet do *much* better?**

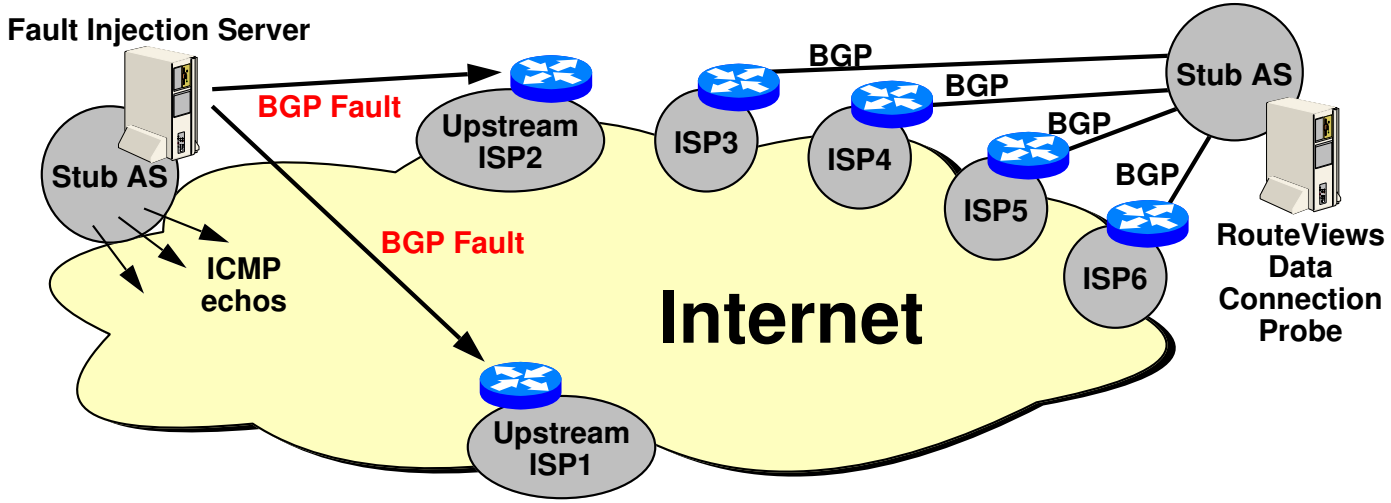


Methodology

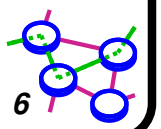
- Introduce artificial faults across Internet
  - but for only *their* AS, of course!
  - failures, repairs, fail-over
- Simluation to study worst case behavior
- Analysis: helps understand worst case bounds
- Two years of traces
- Measure:
  - *Tup:* time for good news to propagate
  - *Tdown:* time for bad news to propagate
  - *Tshort:* time to switch from a longer route to a shorter one
  - *Tlong:* time to switch from a shorter route to a longer one
- In general, want bad news to travel fast, good news to travel slowly



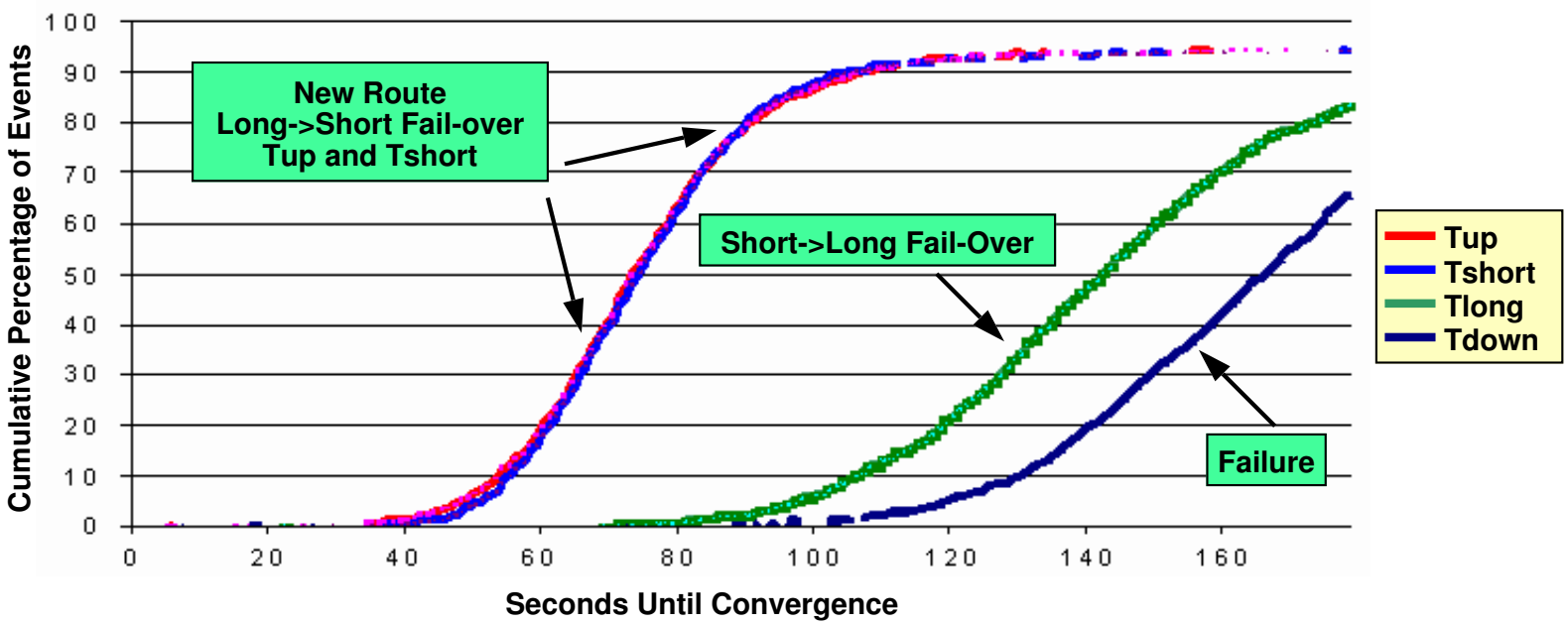
Methodology



- Internet-scale experimentation
- What kind of complexities/errors can arise?
- How do you deal with these errors on *real* routes?



Observed Convergence Latency

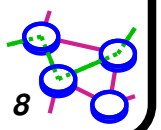


- Less than half of Tdown events converge within two minutes
- Long tailed distribution (up to 15 minutes)

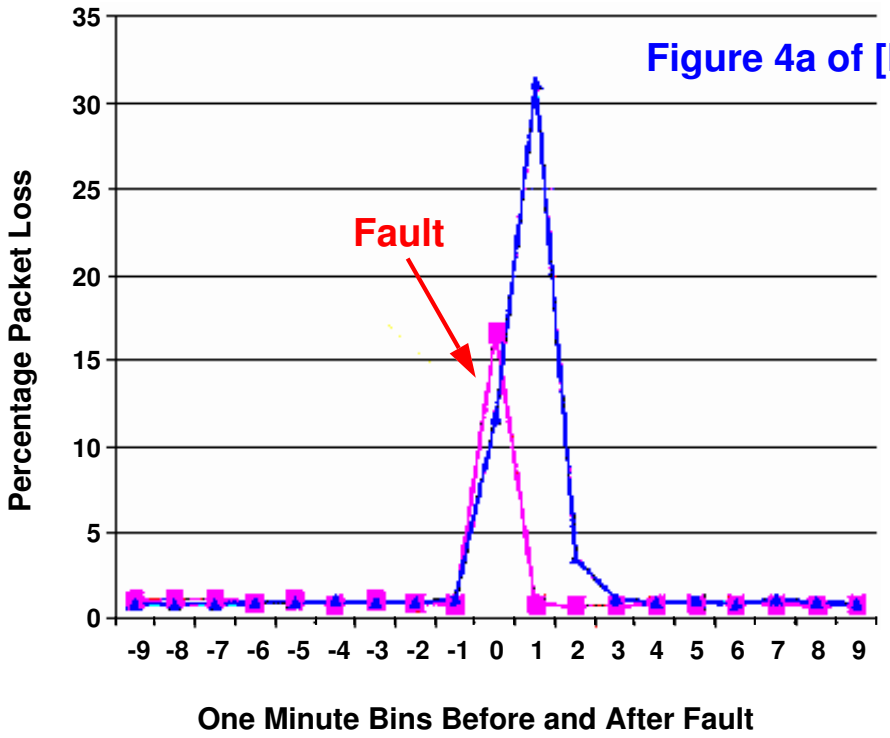
⇨ In general, want bad news to travel fast and good news to travel slowly

Other Observations

- No correlation between network distance (latency, router, or AS hops) and convergence times
- Why is long convergence bad?

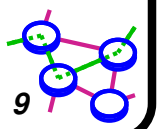


Impact on Traffic



Why does loss go up?

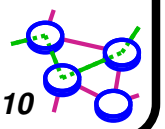
- Because there are route loops in the net causing packet drops.
- Some people use old paths, result in routing loops



How To Tell What's Going On?

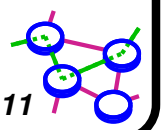
- **Simulate BGP**
  - model one router per AS
  - assume full routing mesh
  - ignore latency
  - synchronous processing via global queue

    - ***simple model that captures key details***

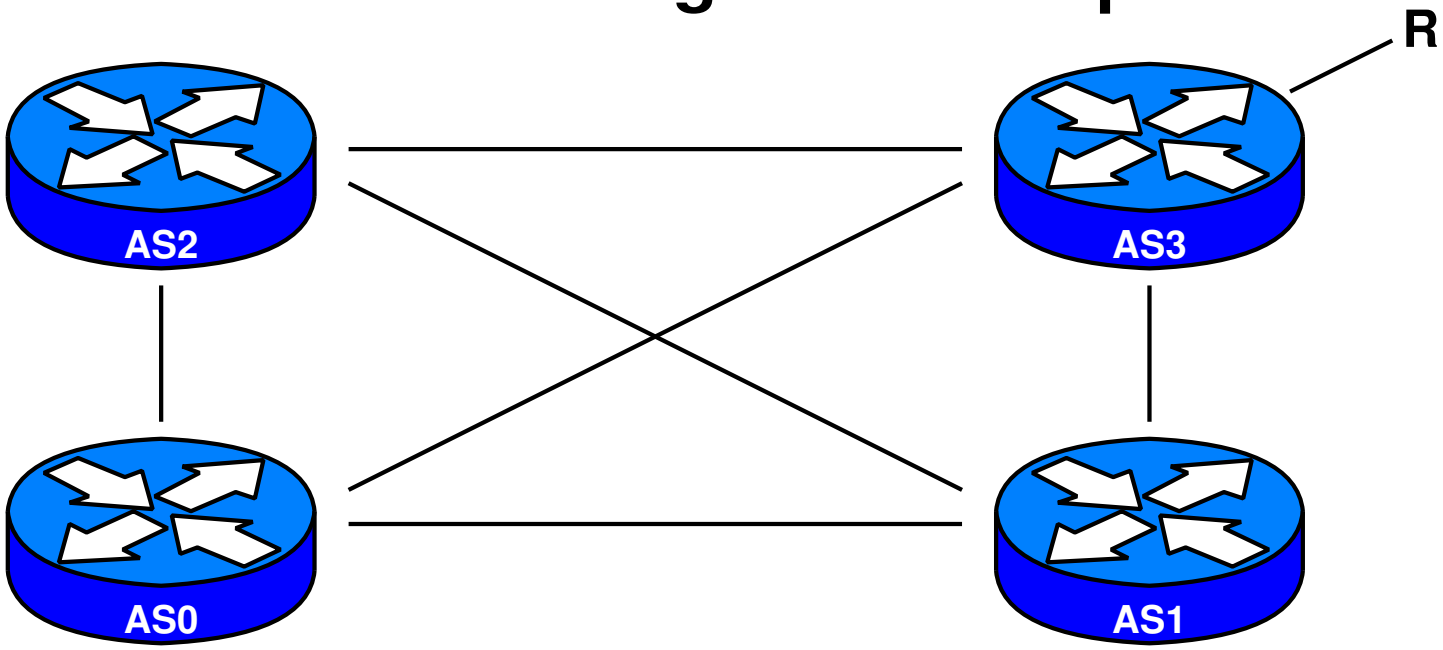


What's Going On?

- There are many possible routes (indirect through other AS's) and it takes a long time with BGP to figoure out that none work
  - BGP can try all paths of length 2, then 3, then 4
    - $\Rightarrow$ O(n!) steps
  - even with MinRouteAdver timers it still can take O(n) steps (13 steps vs. 48 steps originally)

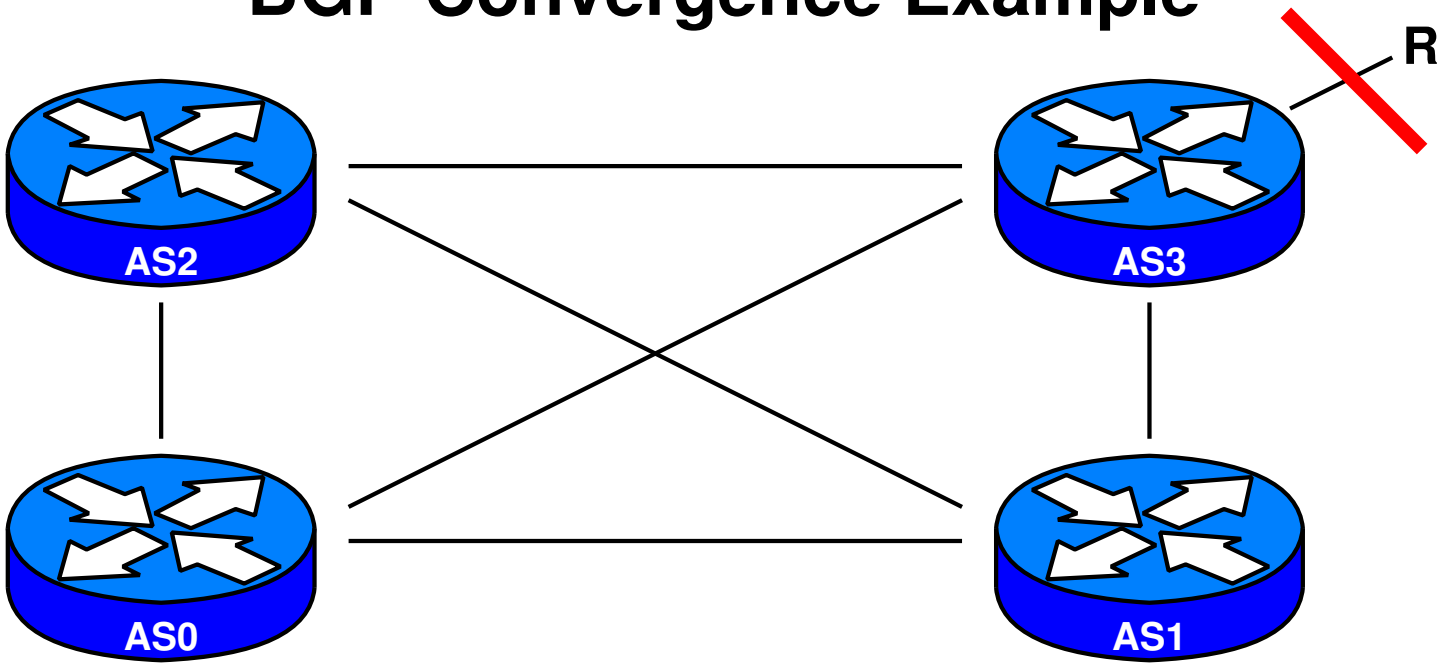


BGP Convergence Example



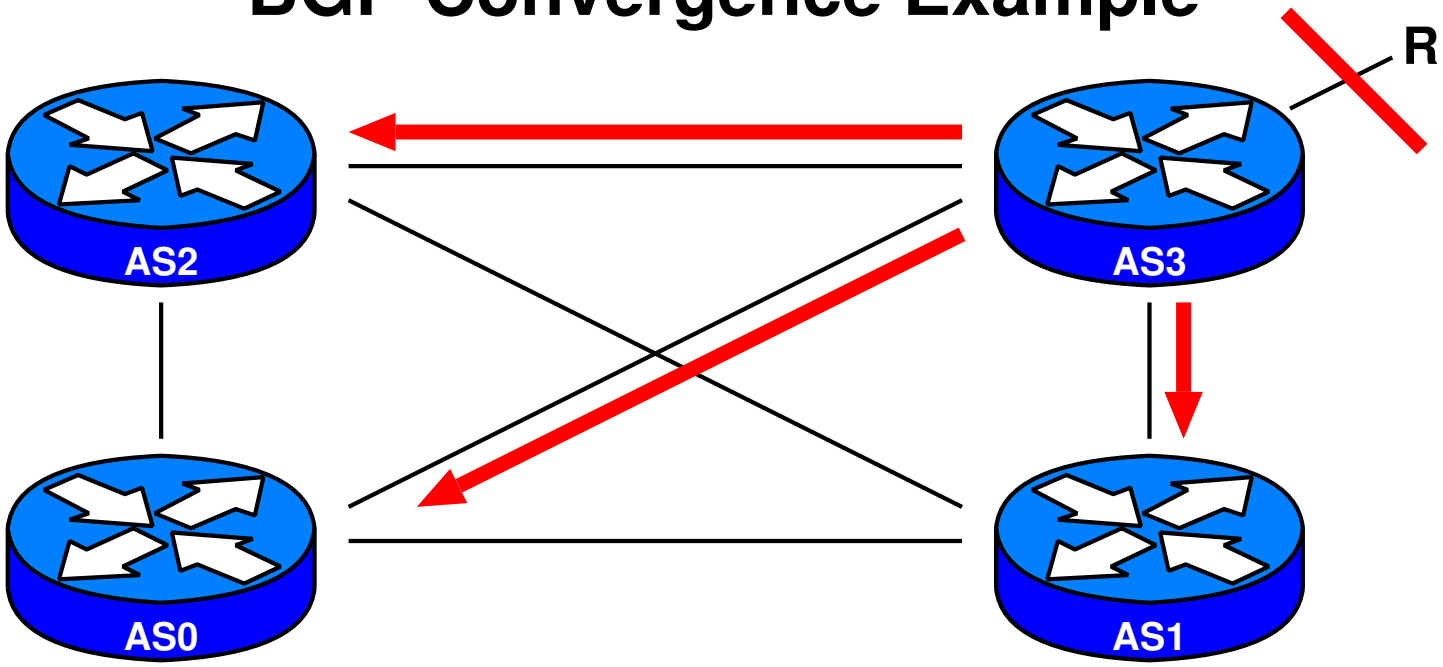
AS0	AS1	AS2
*B R via 3	*B R via 3	*B R via 3
B R via 1 3	B R via 0 3	B R via 0 3
B R via 2 3	B R via 2 3	B R via 1 3

BGP Convergence Example



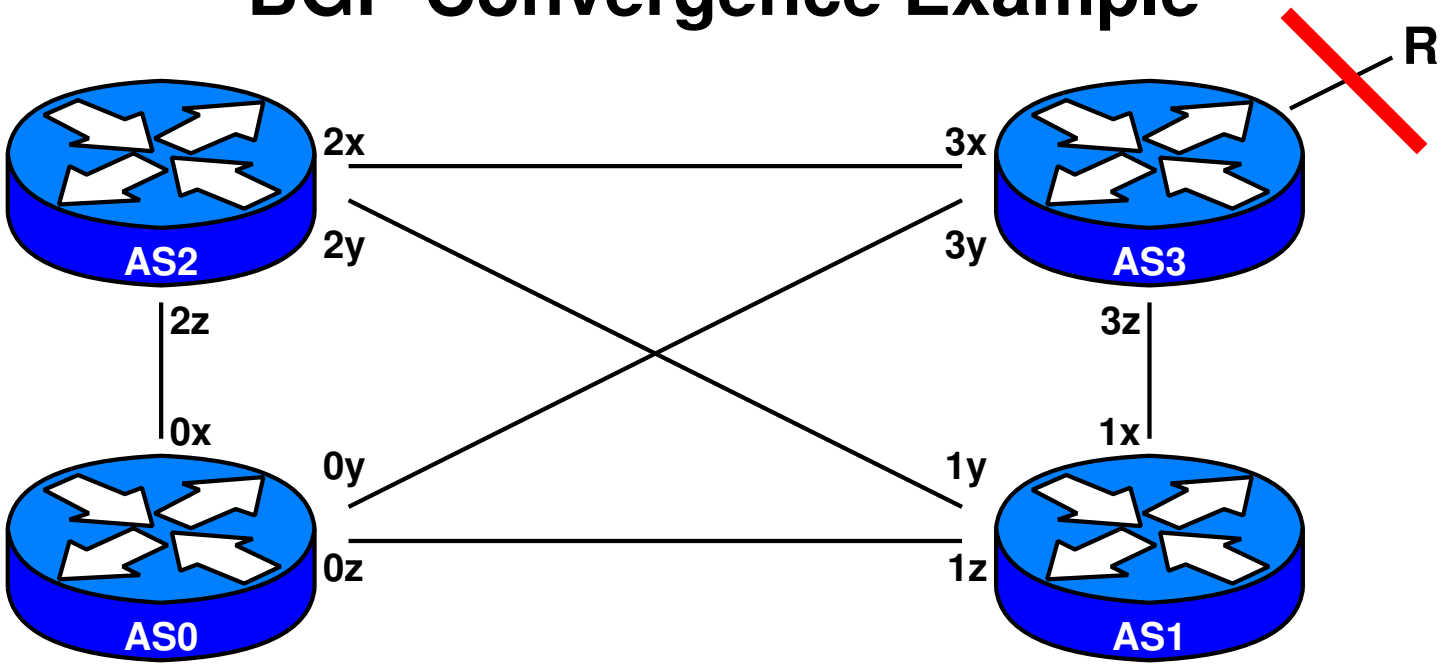
*B R via 3 B R via 1 3 B R via 2 3	*B R via 3 B R via 0 3 B R via 2 3	*B R via 3 B R via 0 3 B R via 1 3
AS0	AS1	AS2

BGP Convergence Example



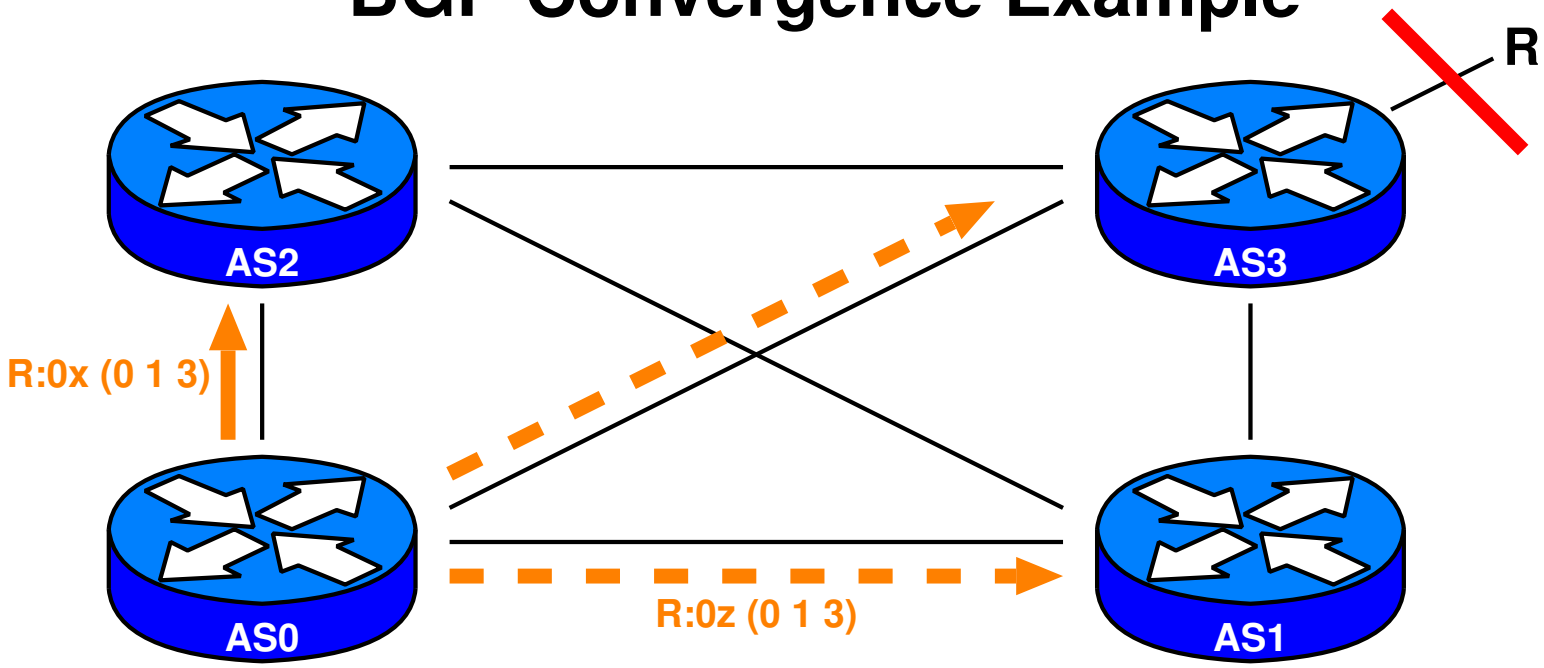
*B R via 3 B R via 1 3 B R via 2 3	*B R via 3 B R via 0 3 B R via 2 3	*B R via 3 B R via 0 3 B R via 1 3
AS0	AS1	AS2

BGP Convergence Example



AS0	AS1	AS2
~~*B R via 3~~	~~*B R via 3~~	~~*B R via 3~~
*B R via 1 3	*B R via 0 3	*B R via 0 3
B R via 2 3	B R via 2 3	B R via 1 3

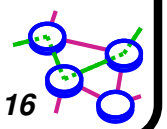
BGP Convergence Example



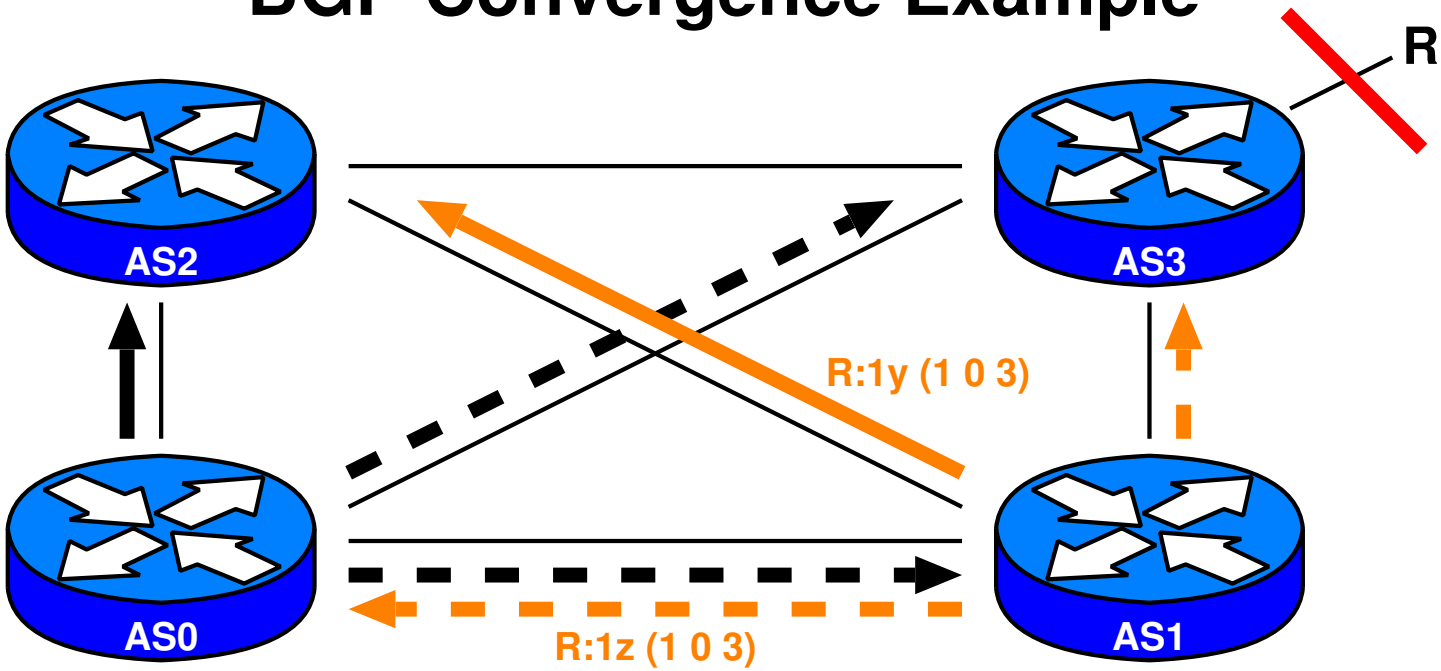
AS0	AS1	AS2
~~*B R via 3~~	~~*B R via 3~~	~~*B R via 3~~
*B R via 1 3	*B R via 0 3	*B R via 0 3
B R via 2 3	B R via 2 3	B R via 1 3

Withdraw

Update



BGP Convergence Example

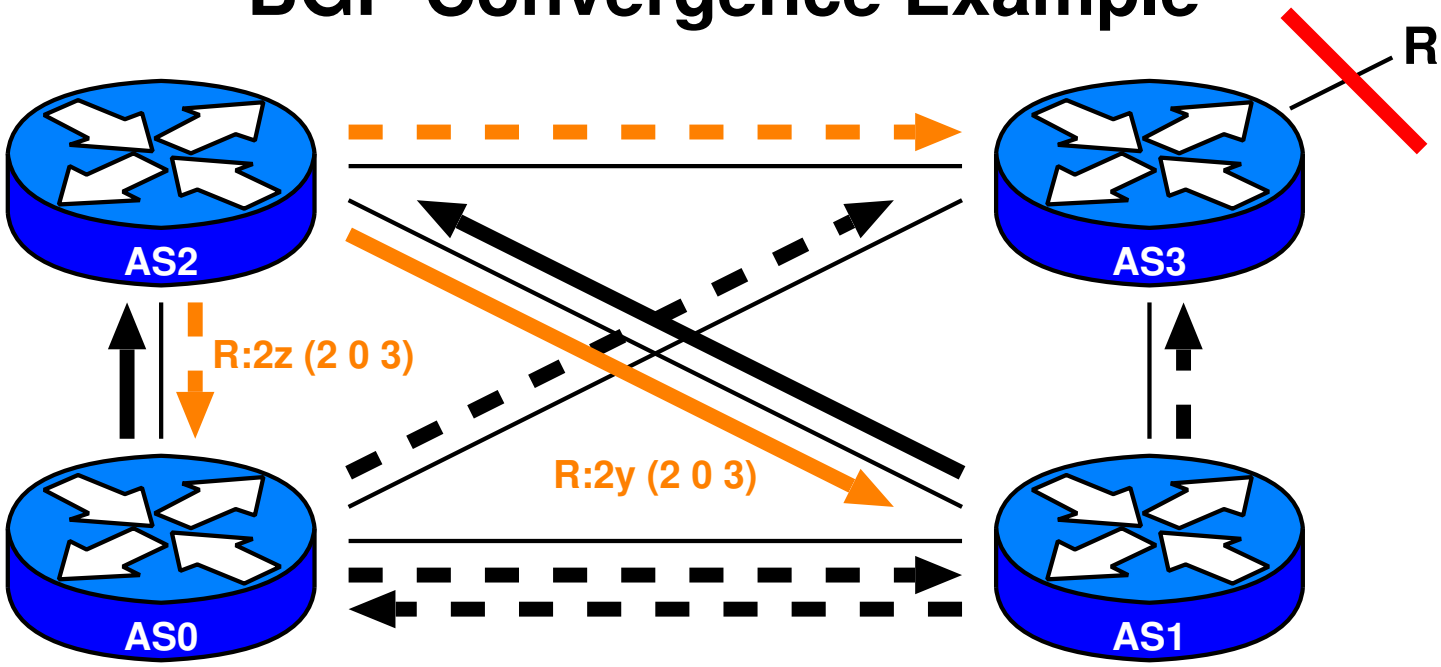


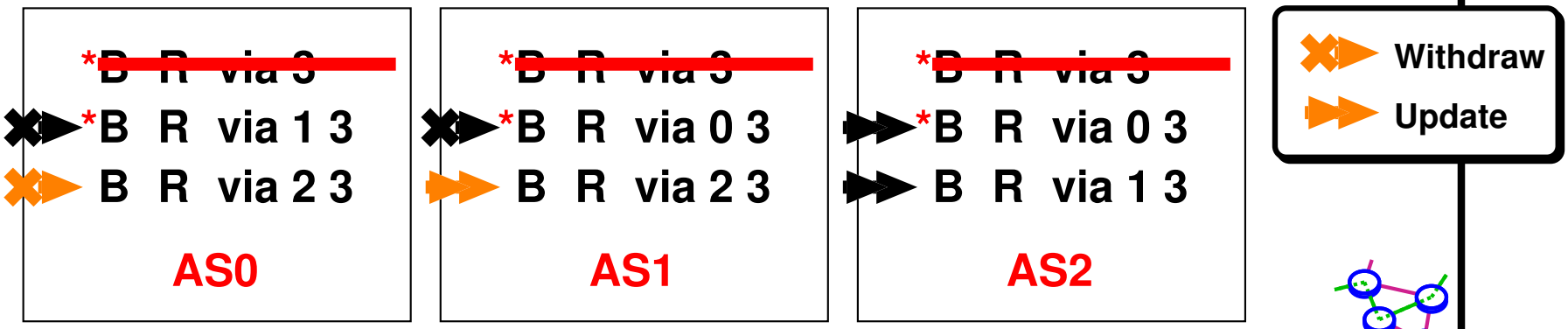
AS0	AS1	AS2
~~*B R via 3~~	~~*B R via 3~~	~~*B R via 3~~
*B R via 1 3	*B R via 0 3	*B R via 0 3
B R via 2 3	B R via 2 3	B R via 1 3

Withdraw

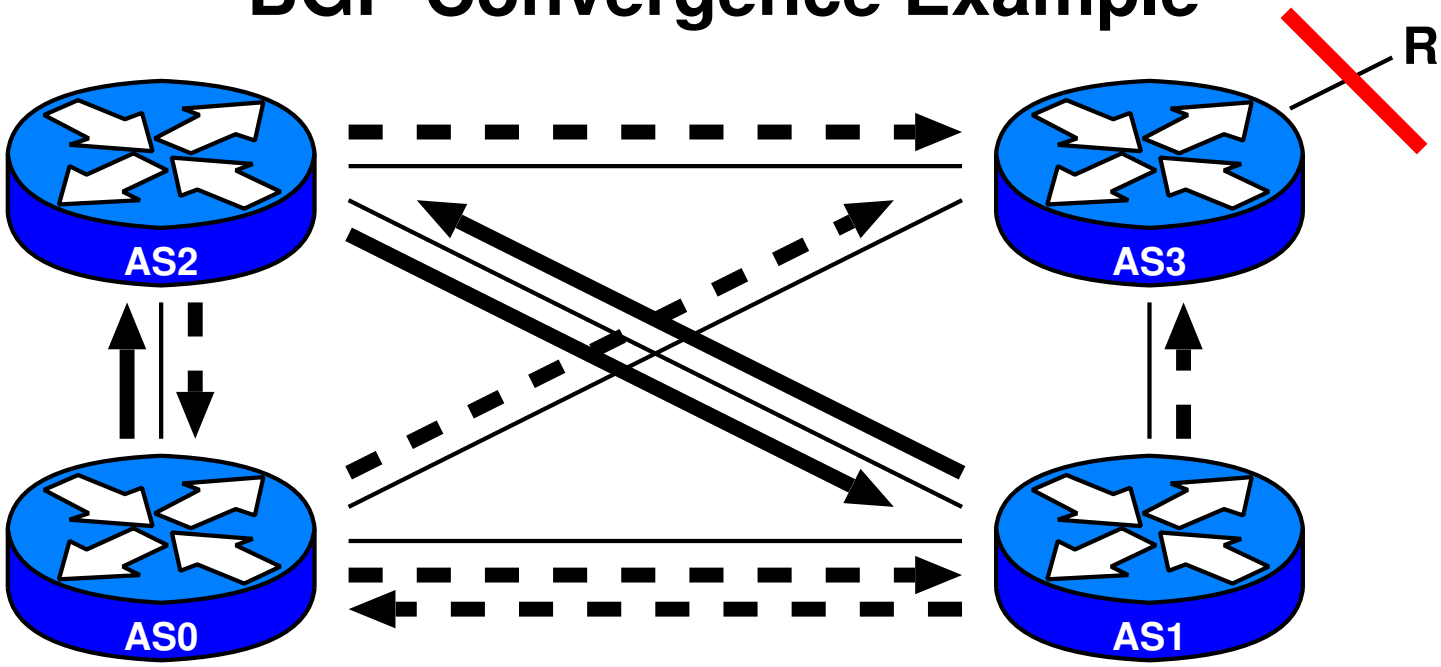
Update

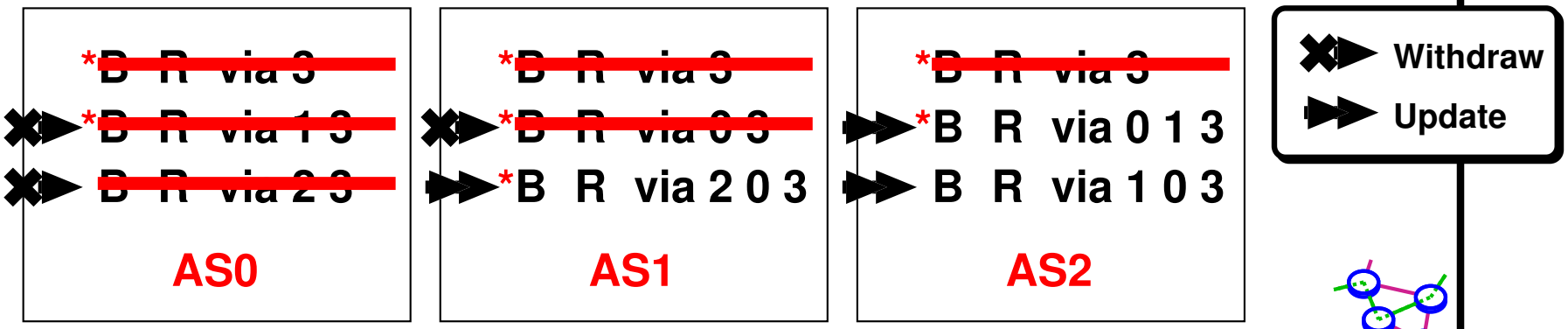
BGP Convergence Example



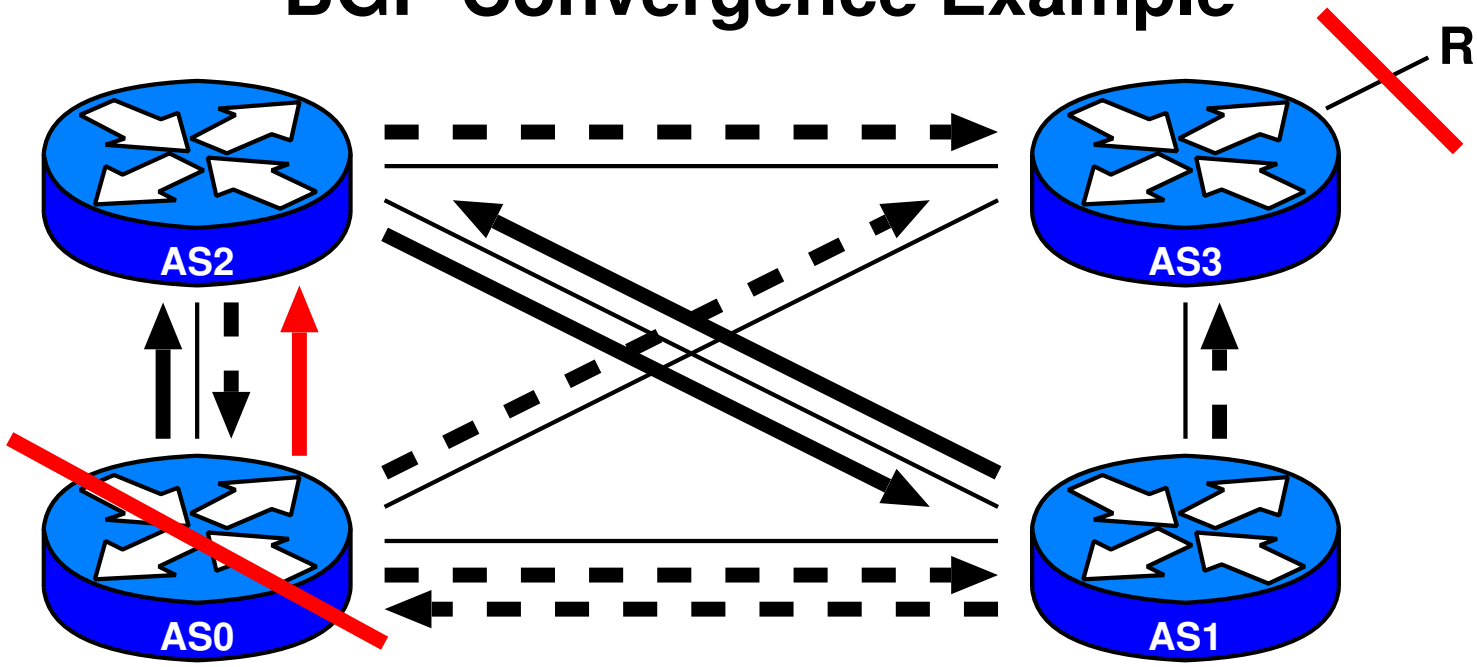


BGP Convergence Example





BGP Convergence Example

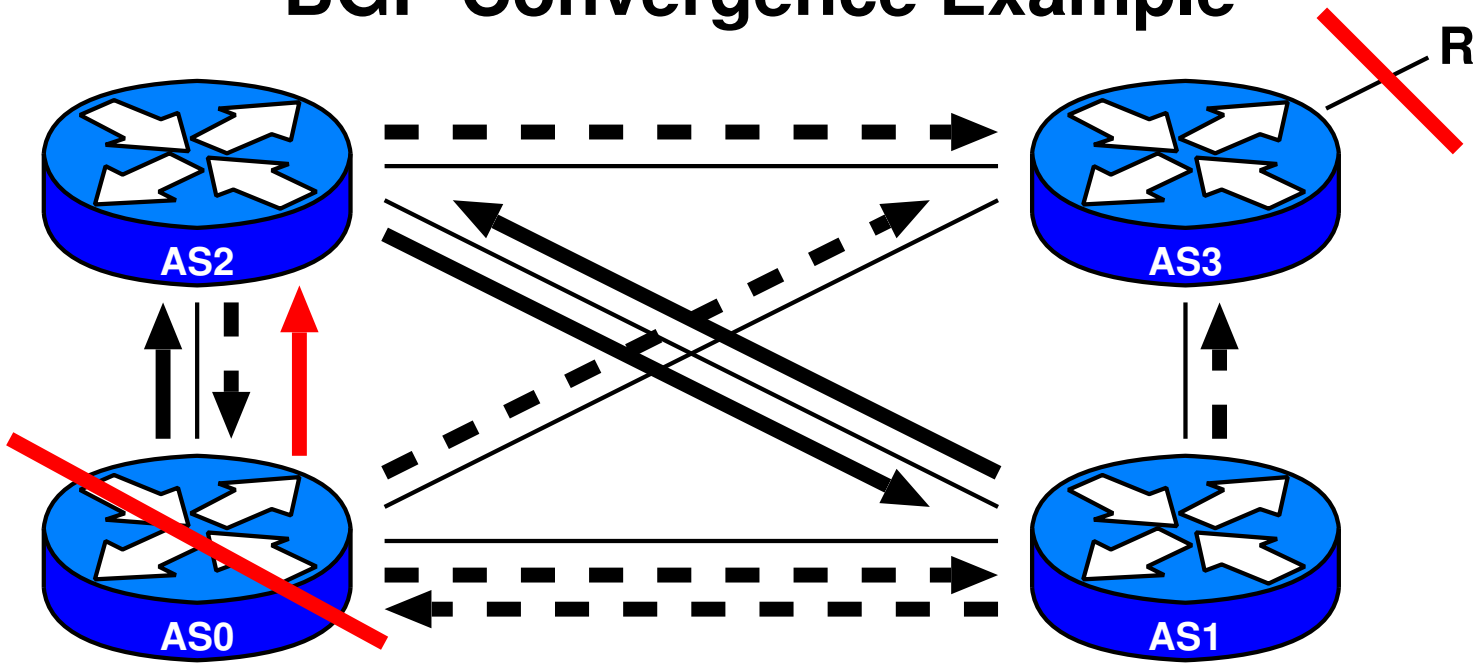


AS0	AS1	AS2
~~*B R via 3~~	~~*B R via 3~~	~~*B R via 3~~
~~*B R via 1 3~~	~~*B R via 0 3~~	*B R via 0 1 3
~~B R via 2 3~~	*B R via 2 0 3	B R via 1 0 3

Withdraw

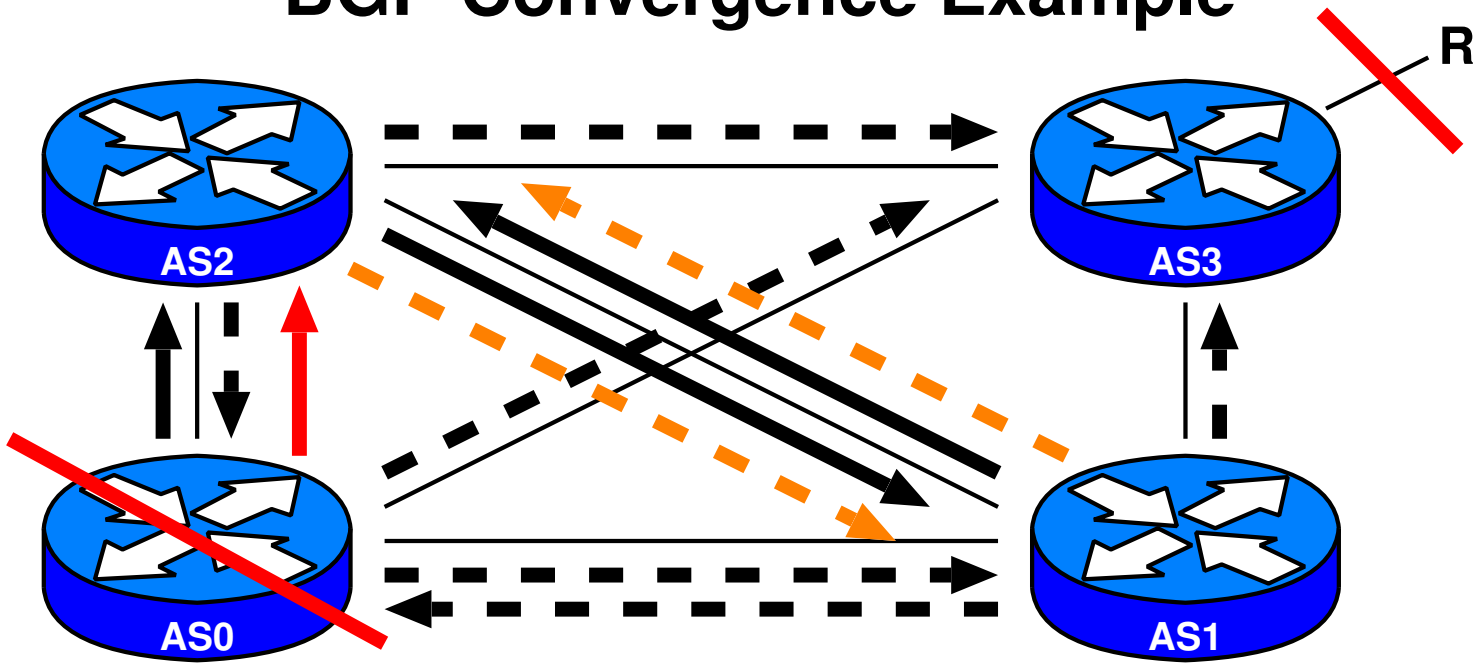
Update

BGP Convergence Example



AS0	AS1	AS2
~~*B R via 3~~	~~*B R via 3~~	~~*B R via 3~~
~~*B R via 1 3~~	~~*B R via 2 3~~	~~*B R via 2 1 3~~
~~B R via 2 3~~	*B R via 2 0 3	*B R via 1 0 3

BGP Convergence Example

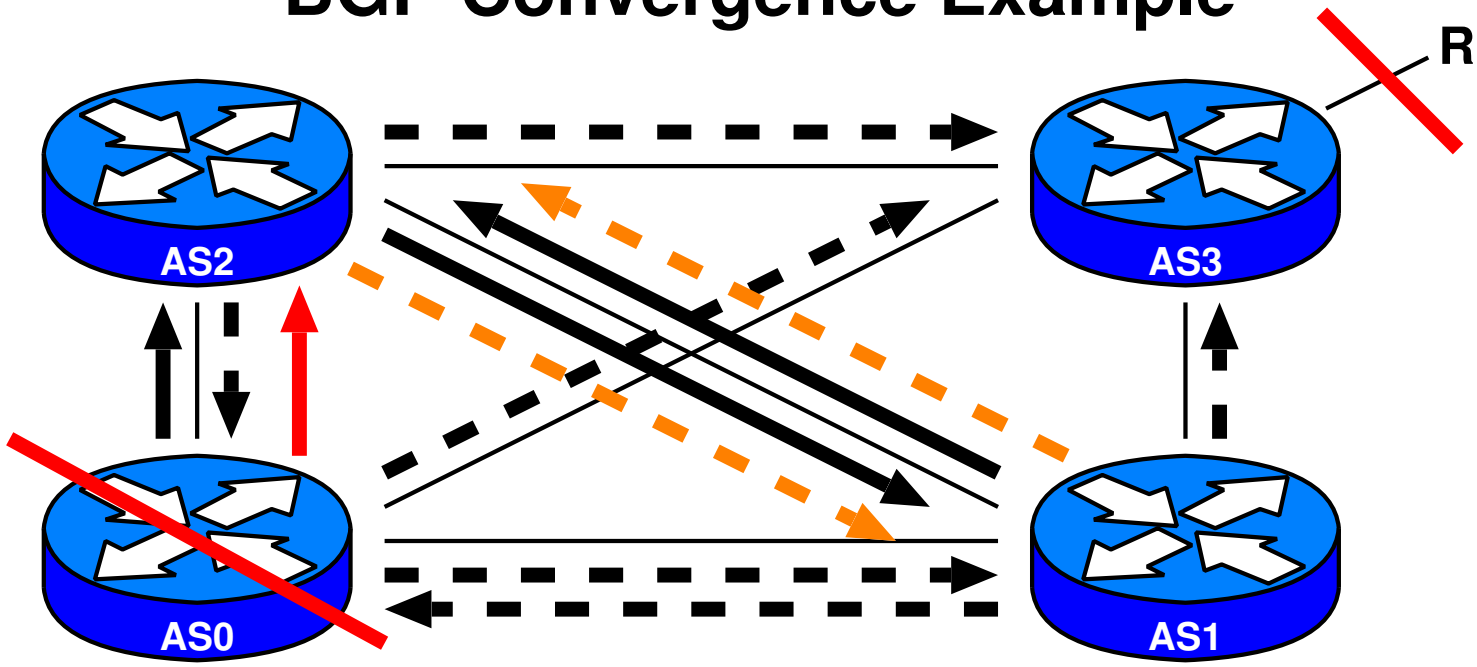


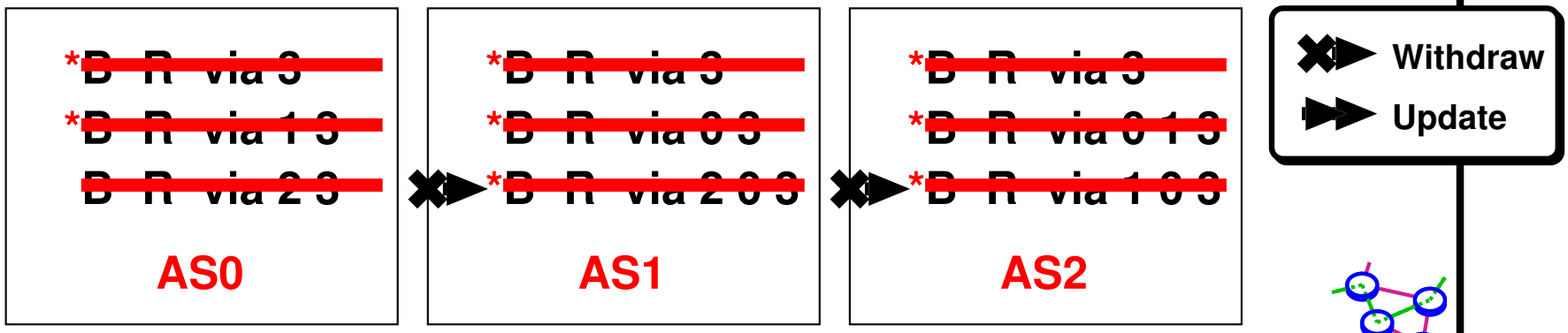
AS0	AS1	AS2
~~*B R via 3~~	~~*B R via 3~~	~~*B R via 3~~
~~*B R via 1 3~~	~~*B R via 2 3~~	~~*B R via 2 1 3~~
~~B R via 2 3~~	*B R via 2 0 3	*B R via 1 0 3

Withdraw

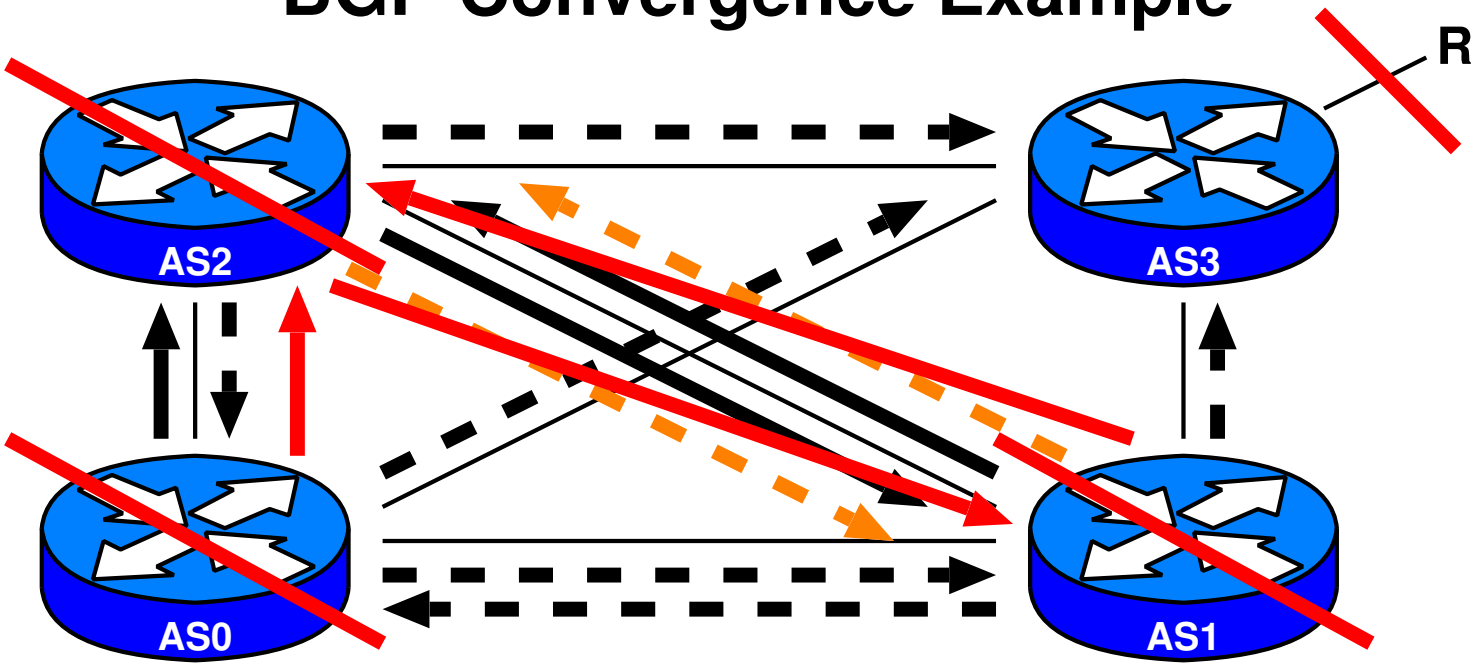
Update

BGP Convergence Example





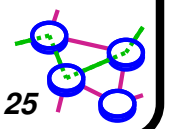
BGP Convergence Example



AS0	AS1	AS2
*B R via 3	*B R via 3	*B R via 3
*B R via 1 3	*B R via 0 3	*B R via 0 1 3
B R via 2 3	*B R via 2 0 3	*B R via 1 0 3

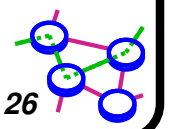
Why Does This Happen?

- In BGP, the theoretical worst case occurs when all possible alternate paths are explored
  - O(n!) such paths
  - explains pathological convergence time



What About MinRouteAdver?

- BGP has minimum advertisement interval timers
  - designed to limit updates
  - and to encourage aggregation
- How does it affect convergence?
  - by delaying announcements, routers figure out the pain sooner
  - see section 5.2
- n-3 rounds of MinRouteAdver (rather than n!)

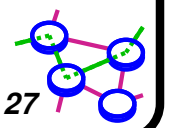


Does This Explain Measurements?

- Tup/Tshort converge quickly because they shorten path length and therefore are quickly accepted
- Tdown/Tlong converge slowly because BGP tries hard to find all alternatives
  - Tlong actually *sometimes* goes quicker if it's "not long enough" and can preempt some of the thrashing

Other Observations

- Could do loop detection at *sender* side and not just receiver side



Discussion

- Context
  - written when the Internet was a large infrastructure
  - some problems were known in BGP, but until then the problems were only hypothetical
- Impact
  - shook the faith of a lot of people (operators and academics alike) in the wisdom of BGP design
- Pros
  - real experimentation (from actual data)
  - relatively simple result
- Cons
  - still a debate about whether operators care about convergence delays