

CS551

Delayed Internet Routing

[Labovitz00]

Bill Cheng

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

Copyright © William C. Cheng

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

Copyright © William C. Cheng

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

Copyright © William C. Cheng

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?

Copyright © William C. Cheng

Methodology

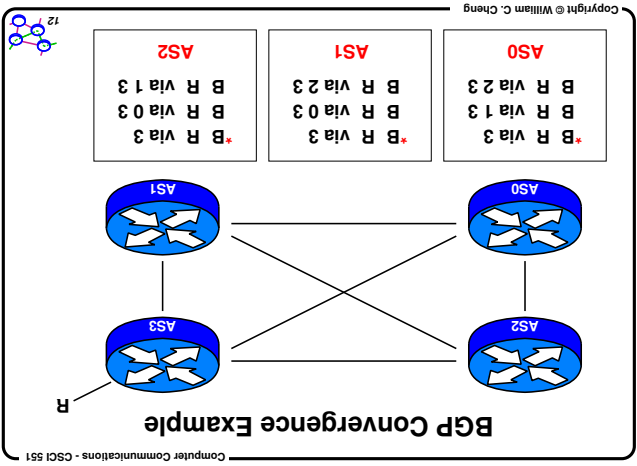
- Introduce artificial faults across Internet
 - but for only *their* AS, of course!
 - failures, repairs, fail-over
- Simulation to study worst case behavior
- Analysis: helps understand worst case bounds
- Two years of traces
- Measure:
 - Tip*: time for good news to propagate
 - Down*: time for bad 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

Copyright © William C. Cheng

Methodology

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

Copyright © William C. Cheng



Copyright © William C. Cheng

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

Computer Communications - CSC1 551

Copyright © William C. Cheng

Other Observations

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

Computer Communications - CSC1 551

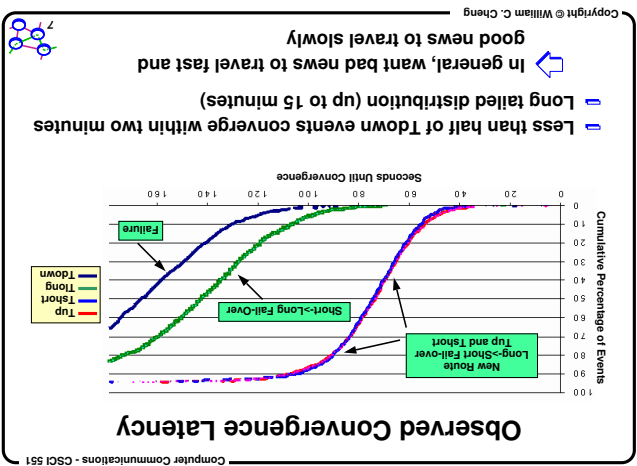
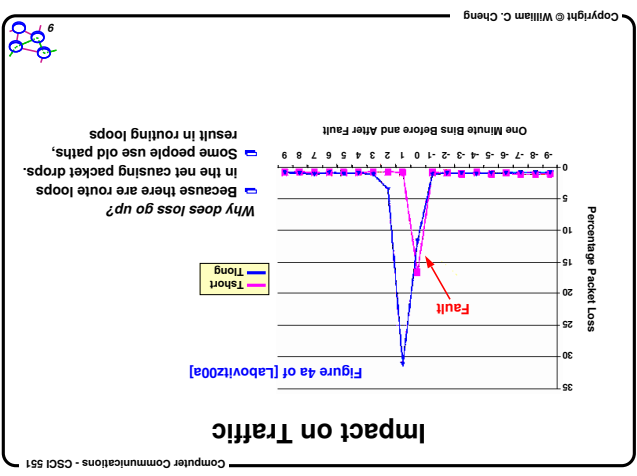
Copyright © William C. Cheng

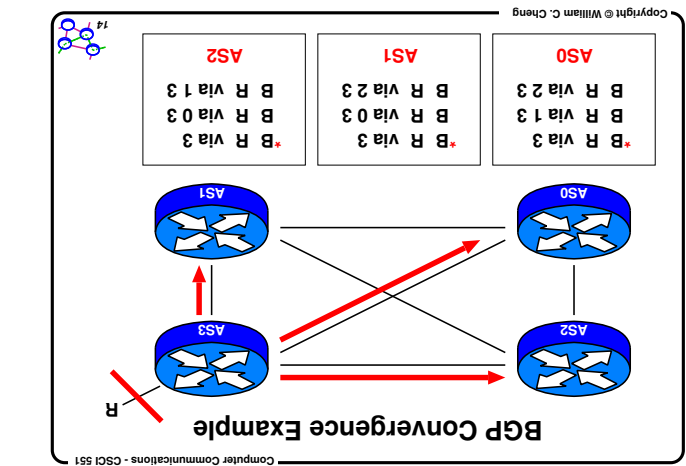
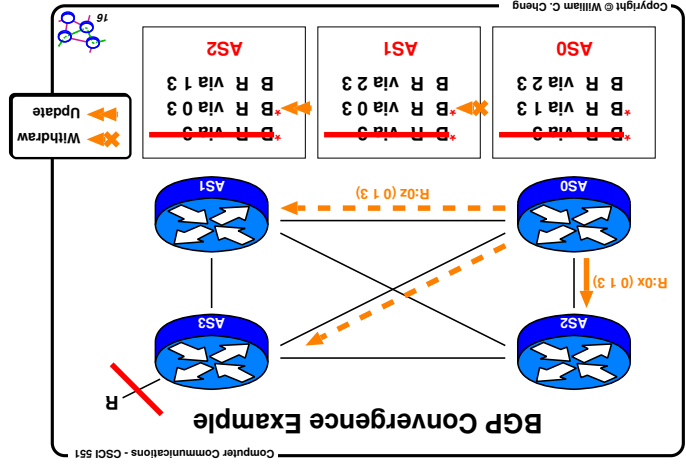
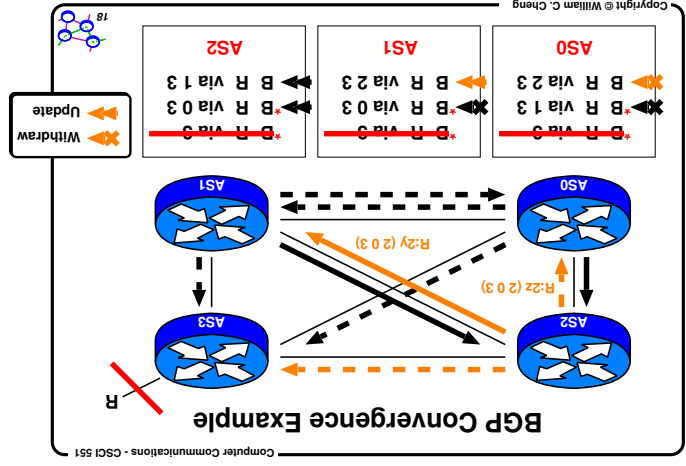
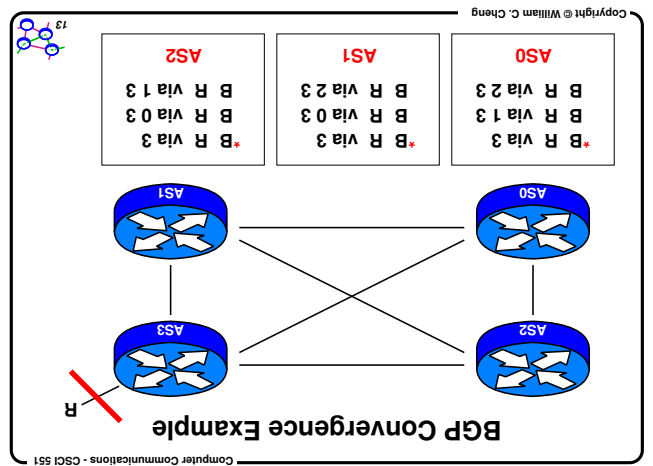
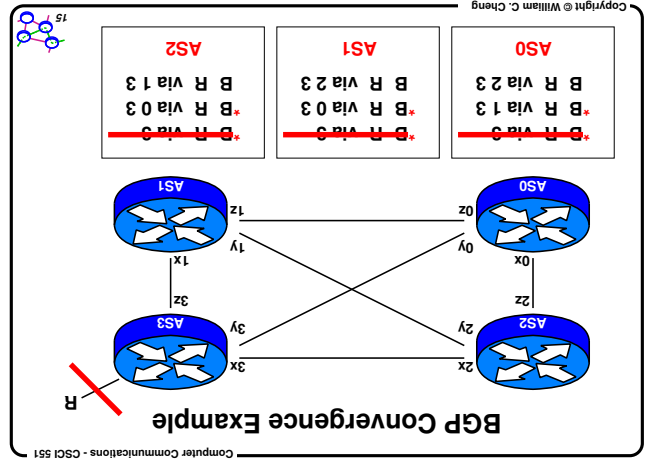
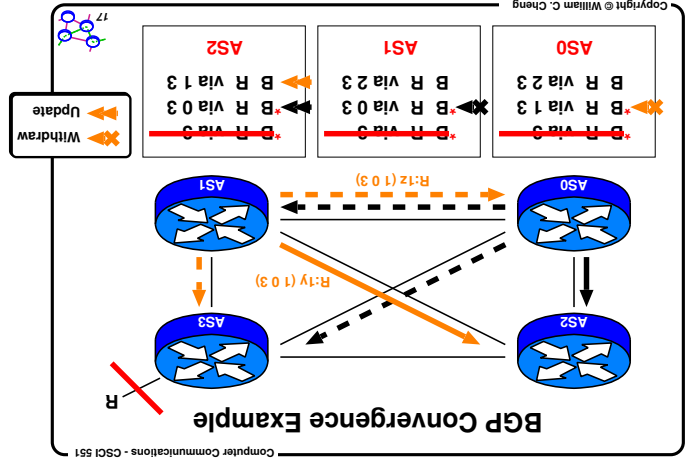
What's Going On?

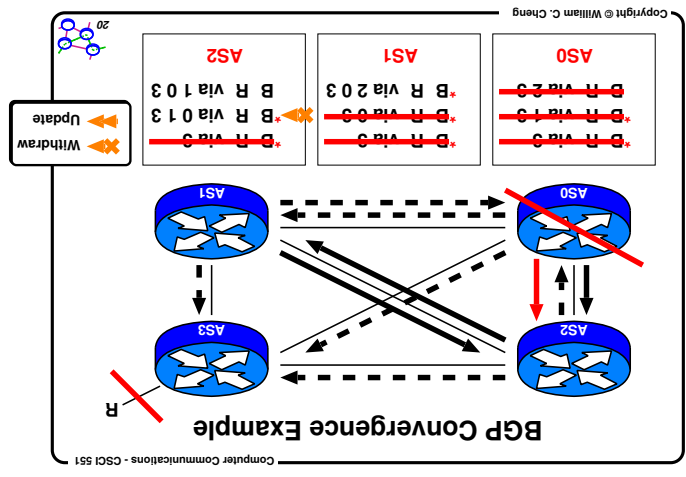
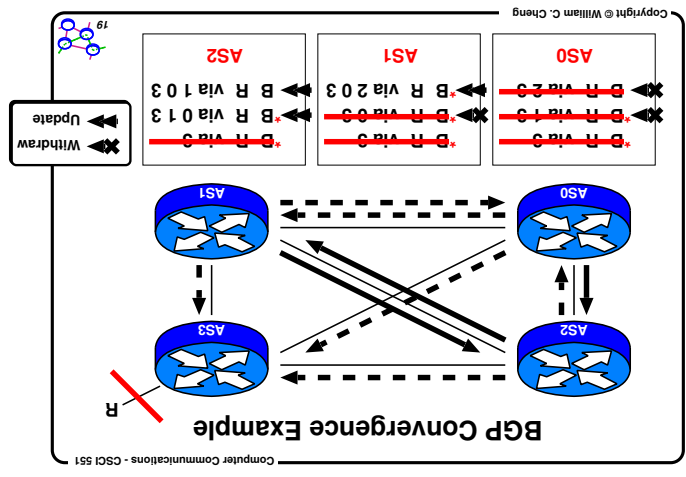
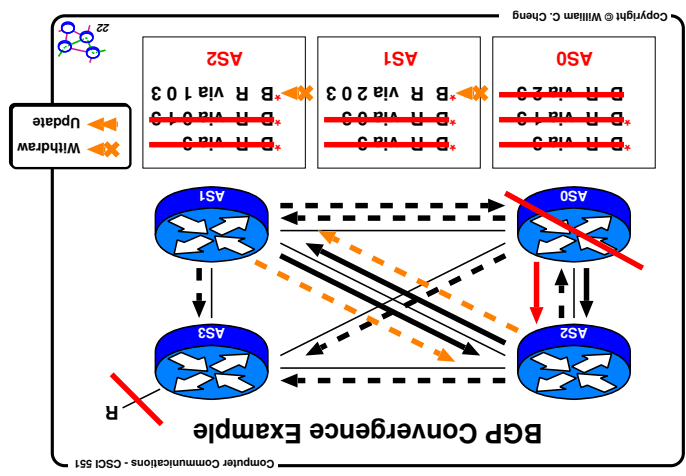
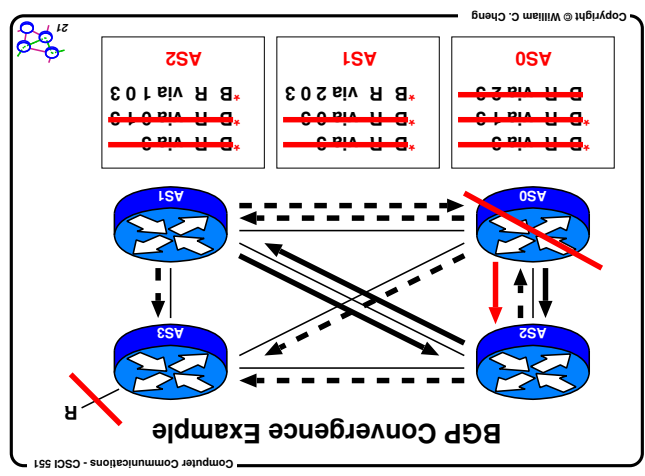
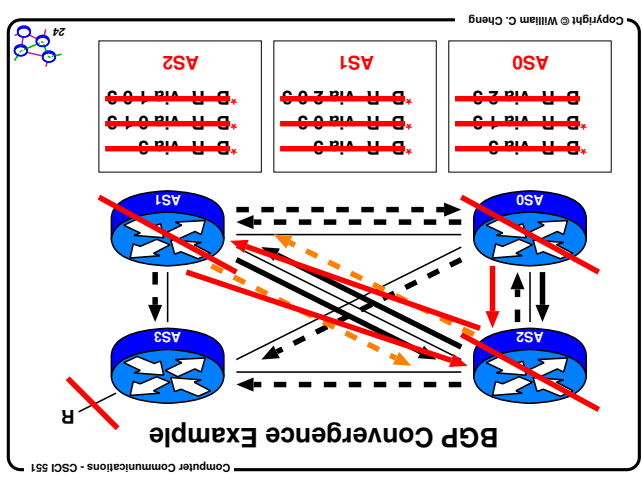
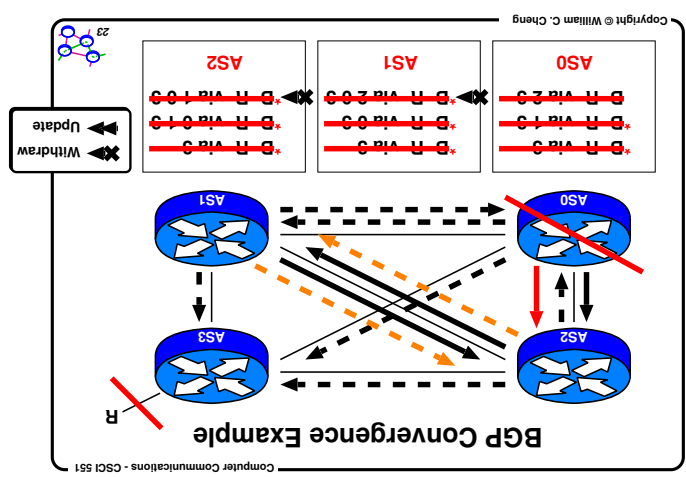
There are many possible routes (indirect through other AS's) and it takes a long time with BGP to figure out that none work

- BGP can try all paths of length 2, then 3, then 4
- $\leq O(n!)$ steps
- even with MinRouteAdver timers it still can take $O(n)$ steps (13 steps vs. 48 steps originally)


Computer Communications - CSC1 551







Copyright © William C. Cheng



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

Computer Communications - CSC1 551

Copyright © William C. Cheng



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!)

Computer Communications - CSC1 551

Copyright © William C. Cheng



Does This Explain Measurements?


- Up/short converge quickly because they shorten path length and therefore are quickly accepted
- Down/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

Computer Communications - CSC1 551

Copyright © William C. Cheng



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

Computer Communications - CSC1 551