Cornell University

Small routing tables

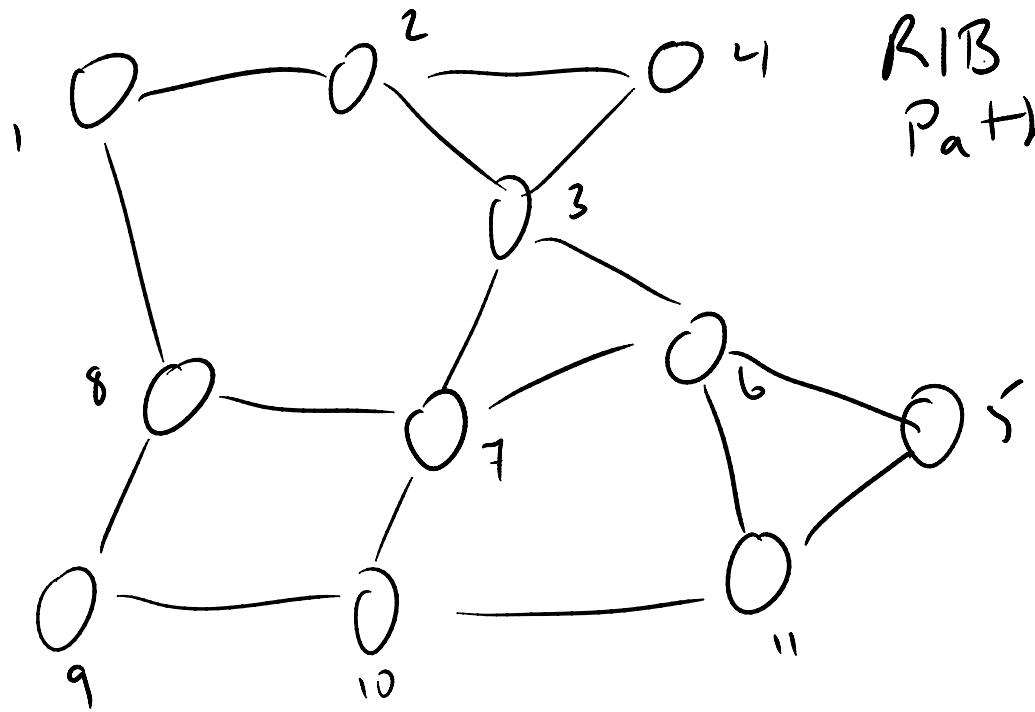
Paul Francis

Outline

- We have a trick for making routing tables very small
  - For hierarchical addresses
  - Global IP, VPNs
  - Called “CRIO” (Core-Router Integrated Overlay)
- And some speculation as to why this might be a good thing

1977

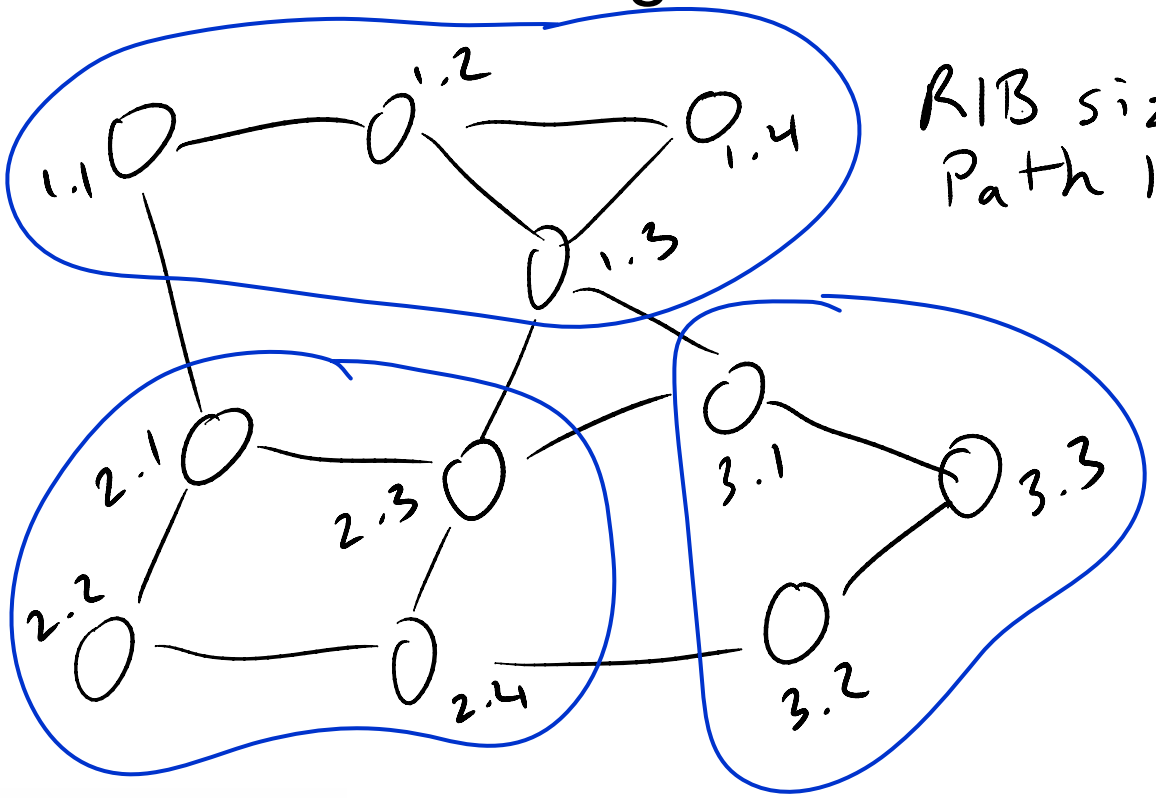
- Folks were looking at the basic trade-off between routing table size and path length



RIB size = 11
Path length = shortest

1977

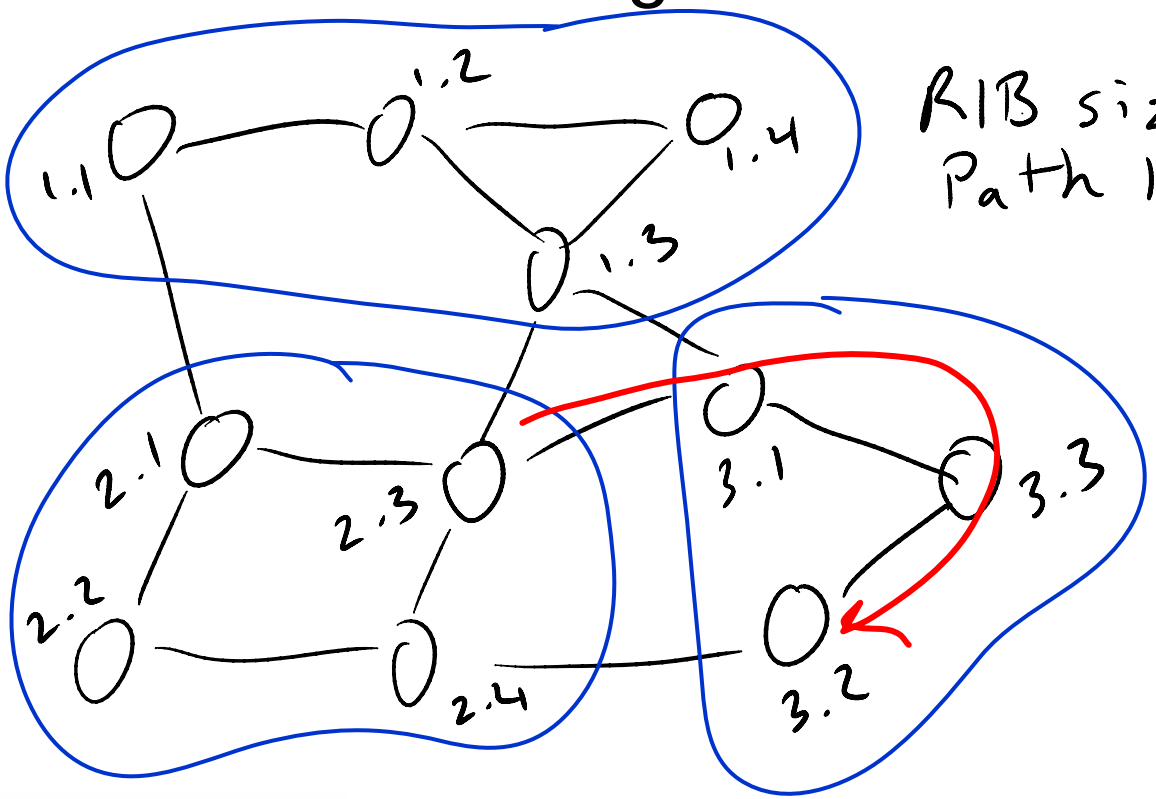
- Folks were looking at the basic trade-off between routing table size and path length



RIB size ≈ 4 or 5
Path length = longer

1977

- Folks were looking at the basic trade-off between routing table size and path length



Path-length / Table size trade-off

- A nice trade-off to have
- This trade-off doesn't exist today
  - Hierarchical nature of internet "forces" an ISP-centric address assignment model
  - Because of multi-homing, sites don't fit neatly into a single "cloud"

CRIO has two parts

- Mapping/tunneling part
  - Can operate stand-alone
- Virtual prefix part
  - Requires mapping/tunneling

Mapping/tunneling part

- BGP keeps routes to major POPs only
  - 1000 – 2000 of these
  - One prefix per POP
- Separate mapping table binds customer prefixes to POPs
- Forwarding is two-step:
  - Map address to POP
  - Tunnel packet to POP address
- Not a new idea
  - Deering's Map-N-Encap, Kim Claffy et. al.

Mapping doesn't shrink FIB per se

- Shifts work from *distributed route computation* problem to *data distribution* problem
  - I would argue that the latter problem is easier
- Data distribution could be done by:
  - OSPF-like flooding
  - ICMP-like notification
    - (Note that with data distribution, not all routers needs to know about a topology change)

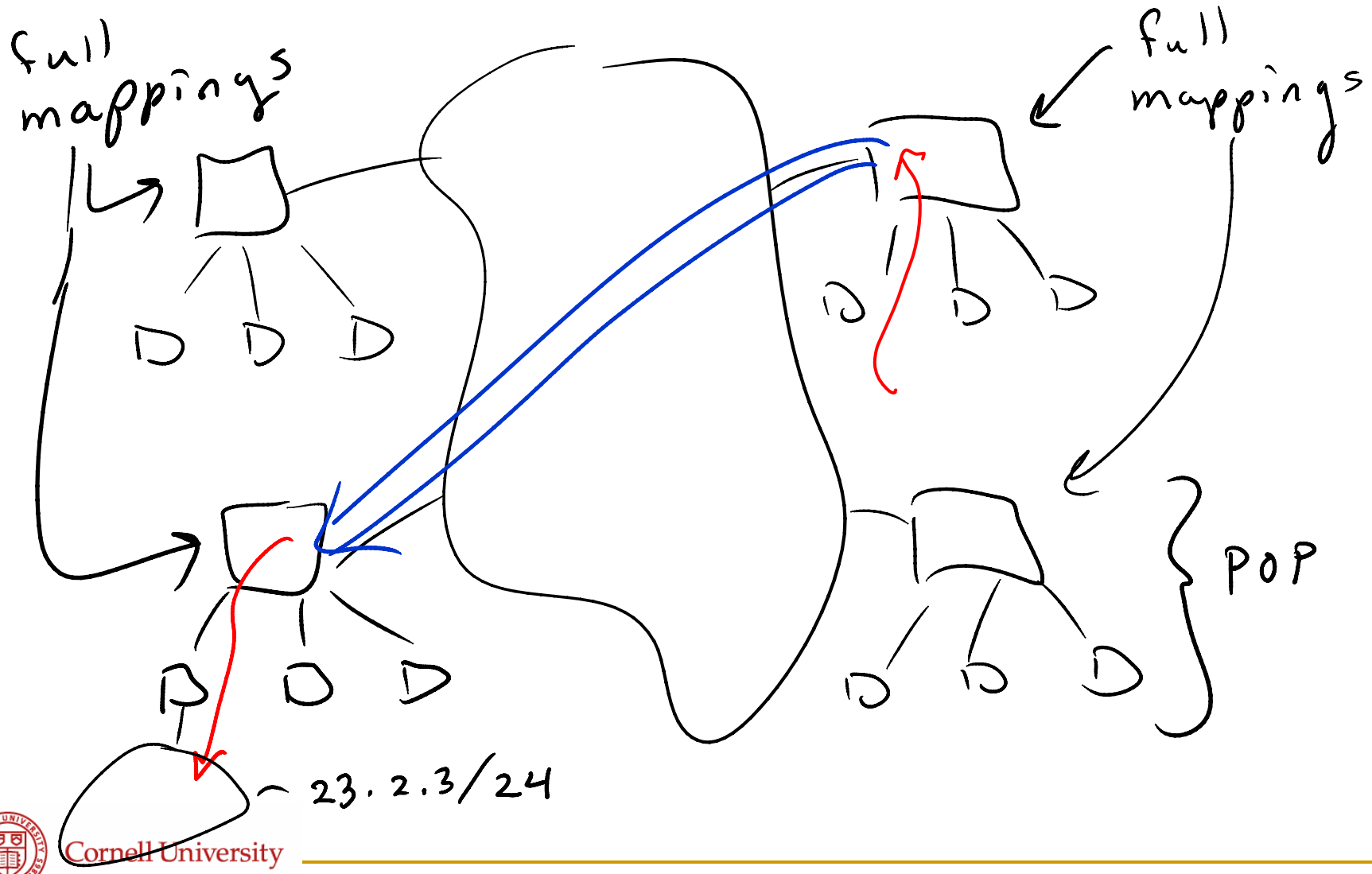
Data distribution easier than route computation

- Streamlined BGP can converge faster
  - A small number of very stable prefixes
  - Operators could crank down the timers
- Easier to debug
  - Mapping table is the same everywhere, BGP RIBs are not
- Easier to secure
  - Secure mapping only, not entire path

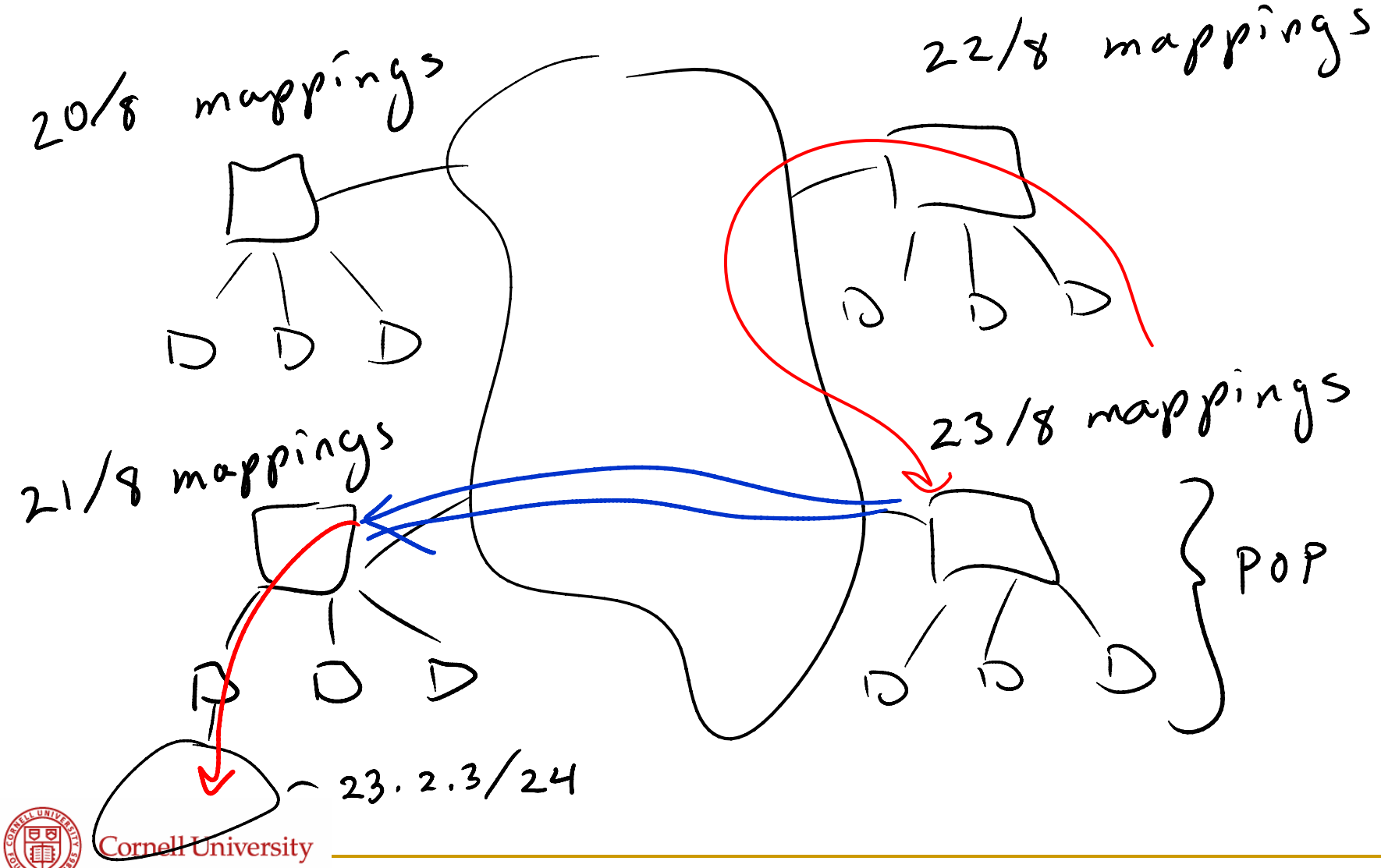
Other mapping characteristics

- Provides a new policy hook
  - For multi-homed nodes, mapping can indicate access preference
- Detunnelling is costly
  - Though it could be implemented lightweight (one-ended tunnels)
- Tunnels introduces new security problems
  - Deflection DoS attack
  - Mitigate by using MPLS or a new protocol field for outer IP header

Mappings without virtual prefixes



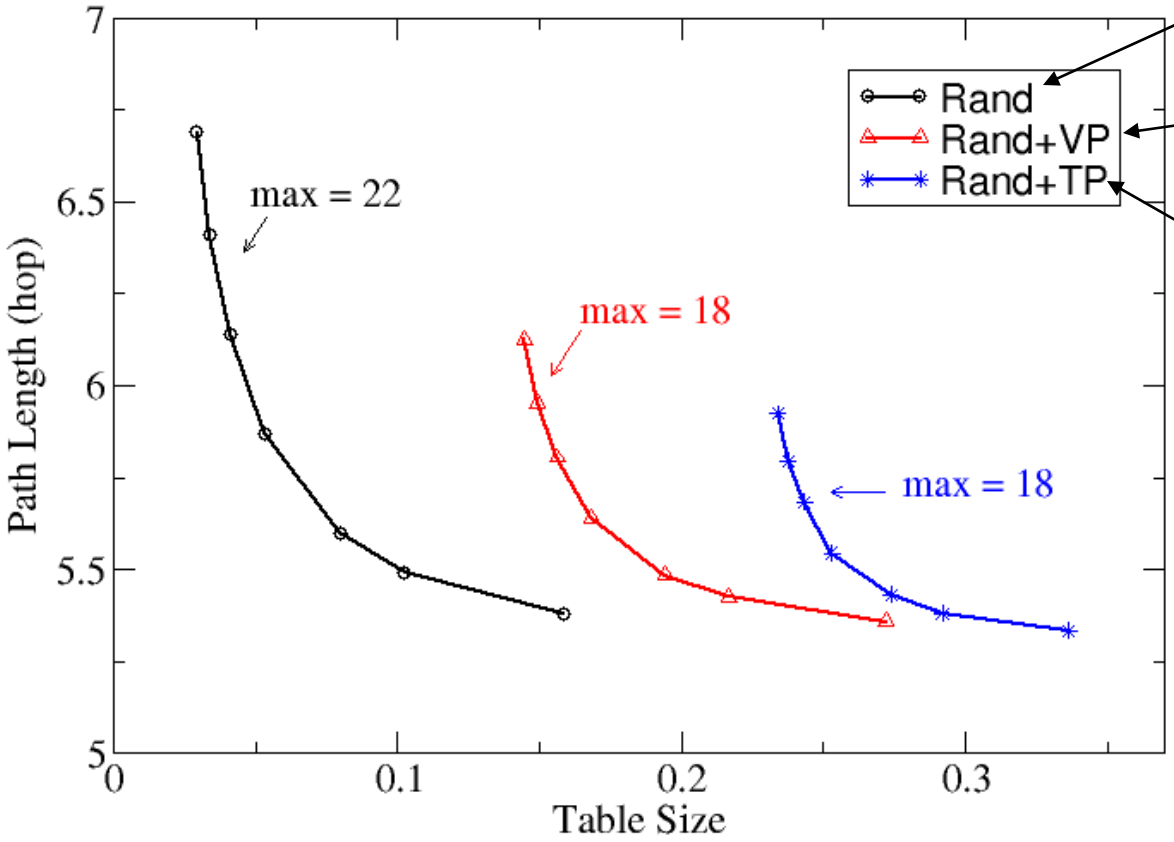
Mappings with virtual prefixes



Virtual Prefixes

- Mappings for a given virtual *super*-prefix are stored only at selected routers
- These routers advertise the virtual prefix into BGP
- Mapping tables and FIBs are smaller, paths are longer
- Completely flexibility as to where individual mappings go
  - Fine-tune size/path-length tradeoff

Path length versus FIB size (for global IP routing)



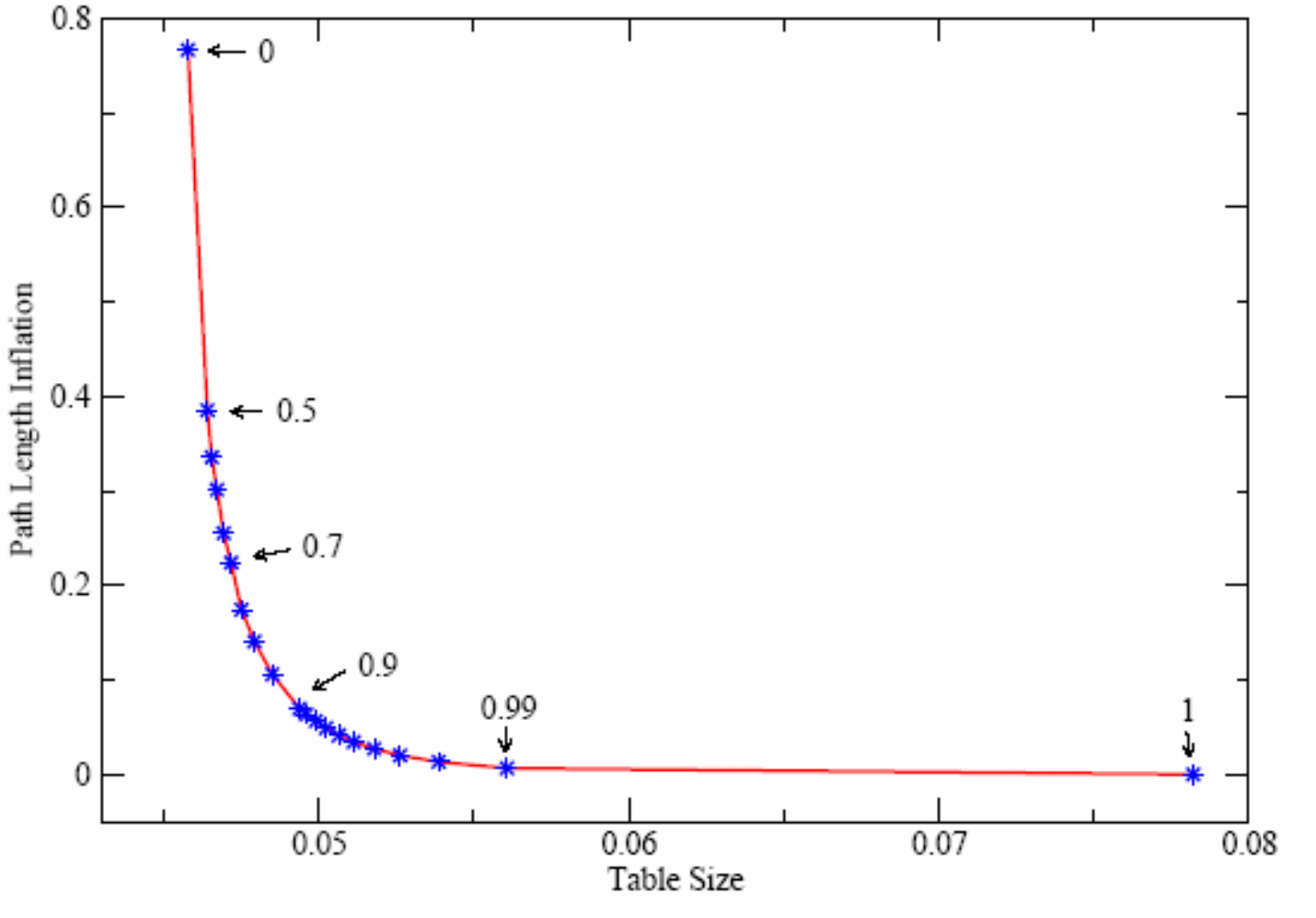
Random across all ISPs

Each ISP has all prefixes

All intra-ISP routes are shortest path

(RIB has around 2000 prefixes)

Path length versus FIB size for VPN routing



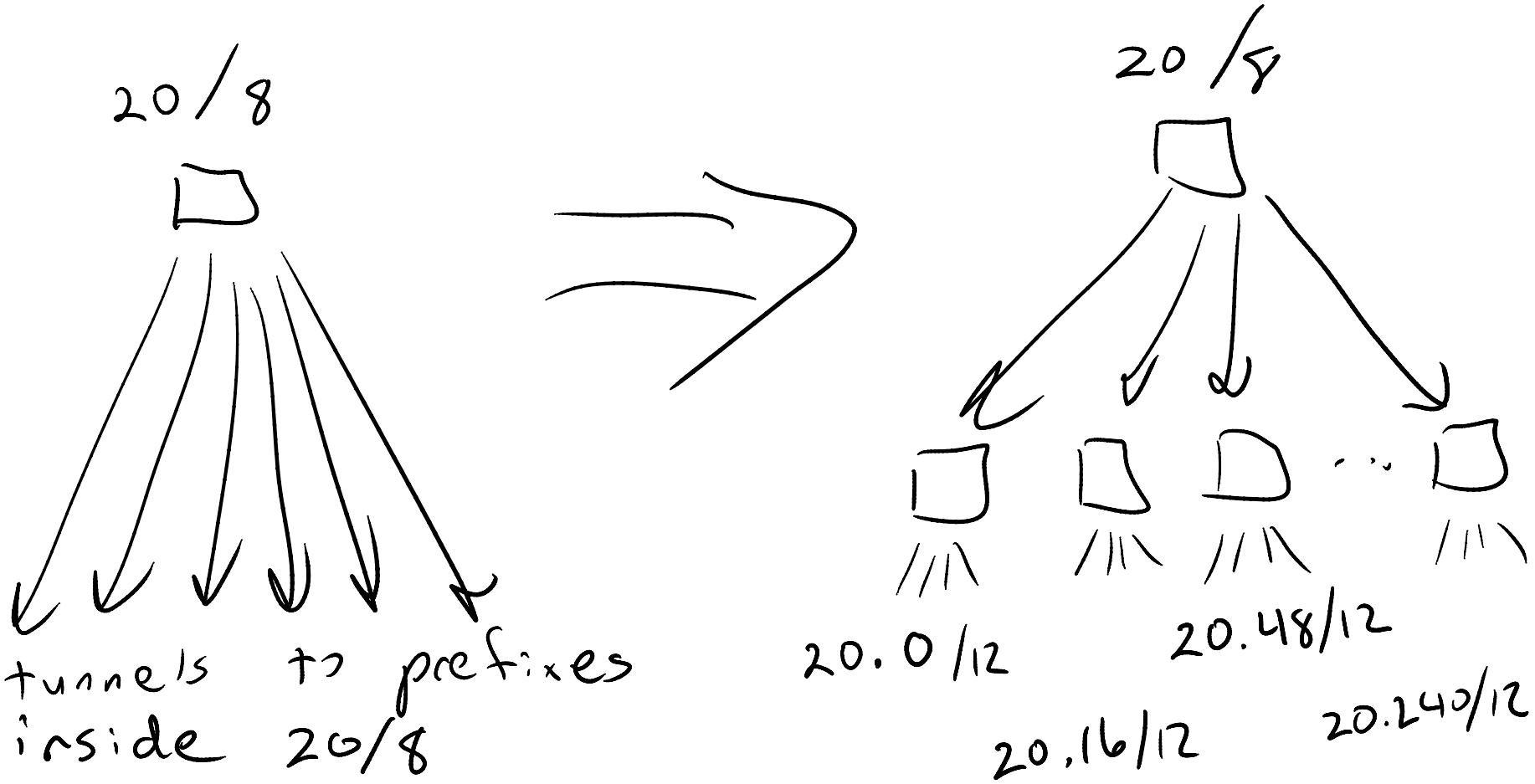
A thought

- Does CRIO allow single-chip forwarding engines?
  - FIB and all processing on a single chip
  - May be possible because ISP can control FIB size
  - On other hand, not all of the table is for hierarchical destination address lookup
    - ACLs, source addressing filtering, etc.
- If so, is there a big advantage to single-chip forwarding engines?
  - After all, much of switch/router memory is due to packet buffering

Really small FIBs

- Can probably shrink the “BGP” FIB component to a few hundred prefixes
  - Using Deering’s metro addressing…all POPs in a metro area have the same prefix
- Can shrink the “mapping” FIB component almost arbitrarily
  - By chaining tunnels (even within a single POP or router)

Chained tunnels



Conclusion

- CRIO gives us back the path-length / table-size trade-off
  - We have shown this for global IP and VPNs
- Interesting, but not clear how valuable this is
  - Faster and simpler BGP (or get rid of BGP altogether)?
  - Better multi-homed traffic engineering?
  - Single-chip forwarding engines?