Scaling FIBs with Virtual Aggregation:
How Much Stretch?
How Much FIB savings?

An Evaluation

By

Dan Jen
jenster@cs.ucla.edu

Disclaimer

- Much of the work in this presentation did not make it into the draft.
  - Recently approved work
- I will update the draft soon to reflect the recent work.

Outline

- Motivation
- VA Primer
- Evaluation Setup
- Evaluation Results
- Concluding Remarks

Why Should We Care about VA?

- Some believe that VA can scale FIBs indefinitely, a major RRG goal.
  - VA distributes the DFZ FIB entries over many routers. ISPs can choose how much to distribute the storage. A tuning knob.
  - “If DFZ doesn't fit amongst 4 routers, store it amongst 8 routers!”
- *Of course, FIB size is not the only scaling dimension of the RRG. Others include RIB size, churn rate, and processing requirements. This will be touched upon later in the presentation.*

Why Should We Care about VA?

- Relatively Low Deployment Barriers
  - Independently deployable by ISPs
    - No 3rd-party infrastructures
  - ISPs immediately get full scalability benefits upon deployment
    - Don't need to wait for universal deployment before full scaling benefits realized.
  - Seamless Interworking with current Internet.
    - All changes are internal and transparent to outside.

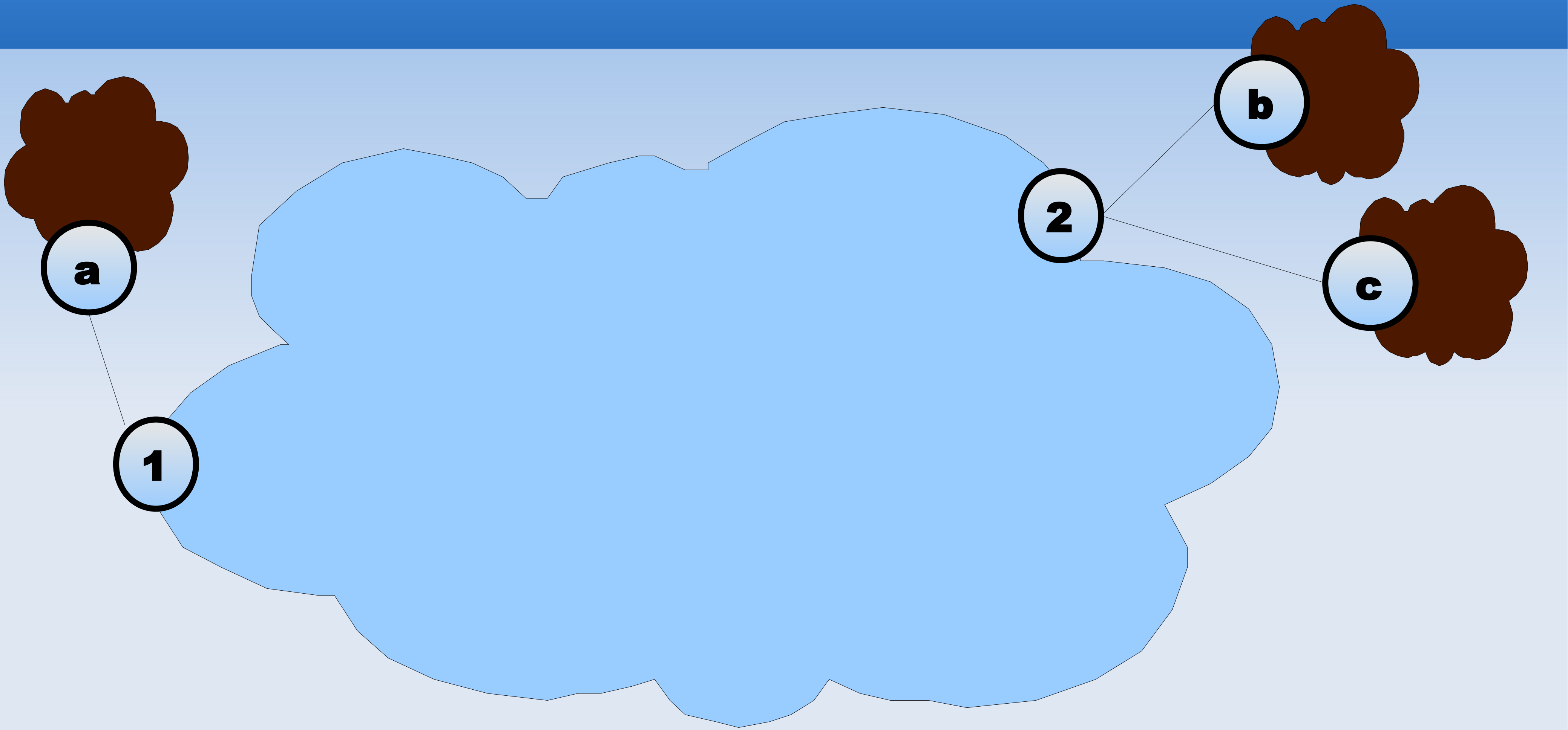
However, How Good is VA?

- VA gets FIB savings, but has drawbacks
  - suboptimal paths (“stretch”)
  - load on networks
- My evaluation focuses on the stretch/savings tradeoff.
- **If an ISP just deploys VA in a simple, intuitive manner, how much stretch and how much FIB savings would a real ISP experience?**

Outline

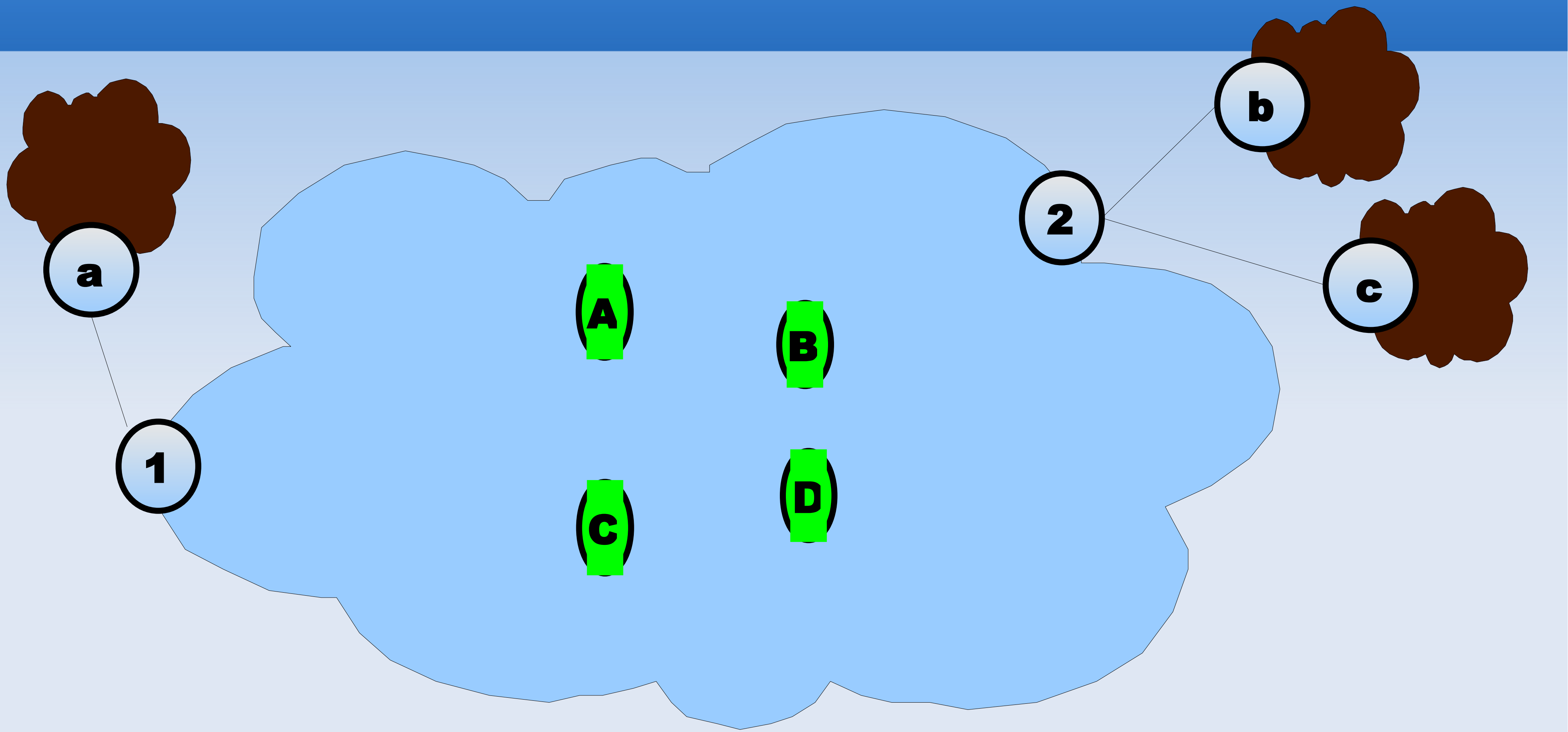
- Motivation
- VA Primer
- Evaluation Setup
- Evaluation Results
- Concluding Remarks

VA Primer



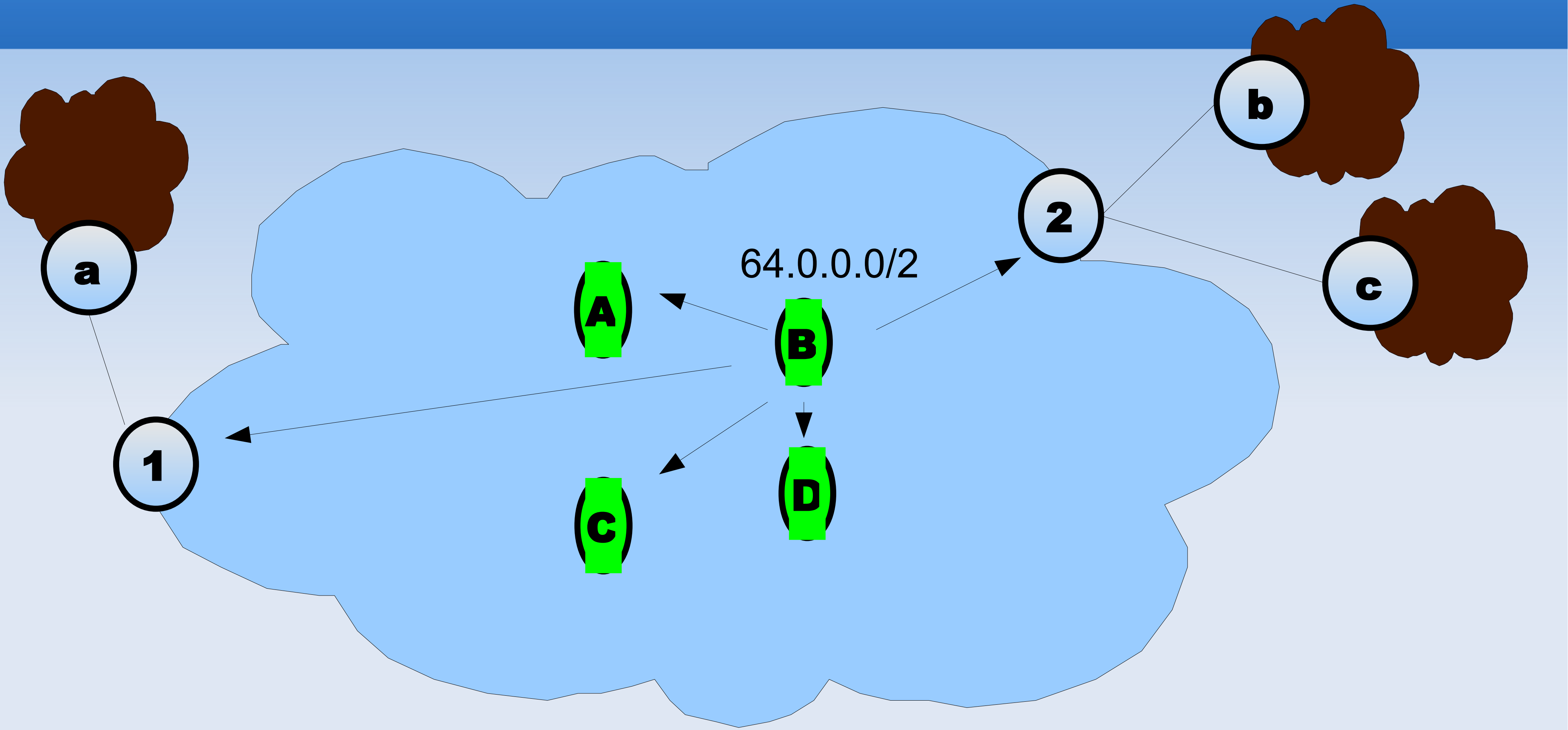
- Brown ISPs represent external peerings (customers, peers)
- External Peers represent possible egress points out of ISP

VA Primer



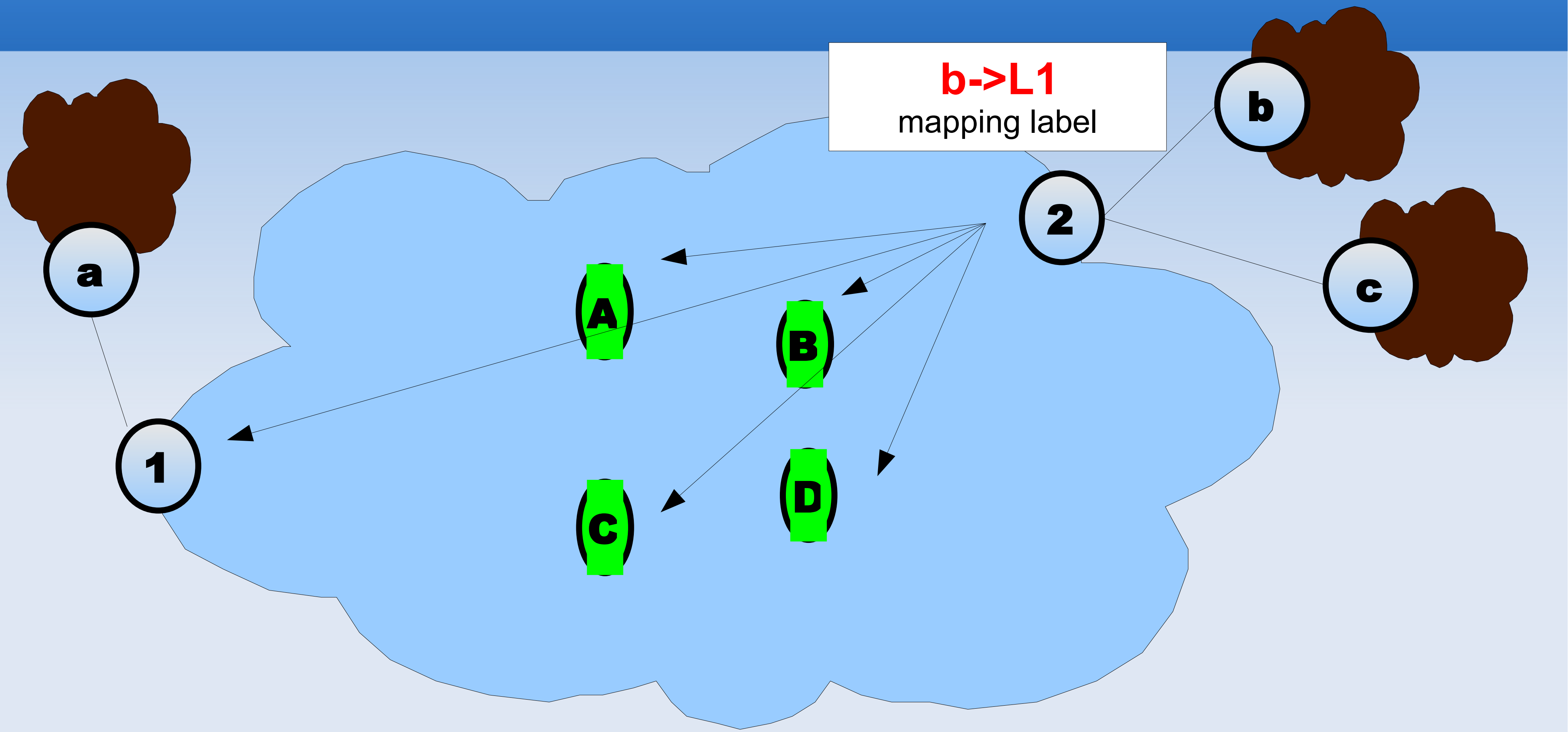
- 2 kinds of routers in a VA:
    - Directory (D) and non-directory(ND) routers
    - A thru D are directory routers, 1,2 are ND routers
- FIB distributed among directory routers. ND routers needn't store FIB.

VA Primer



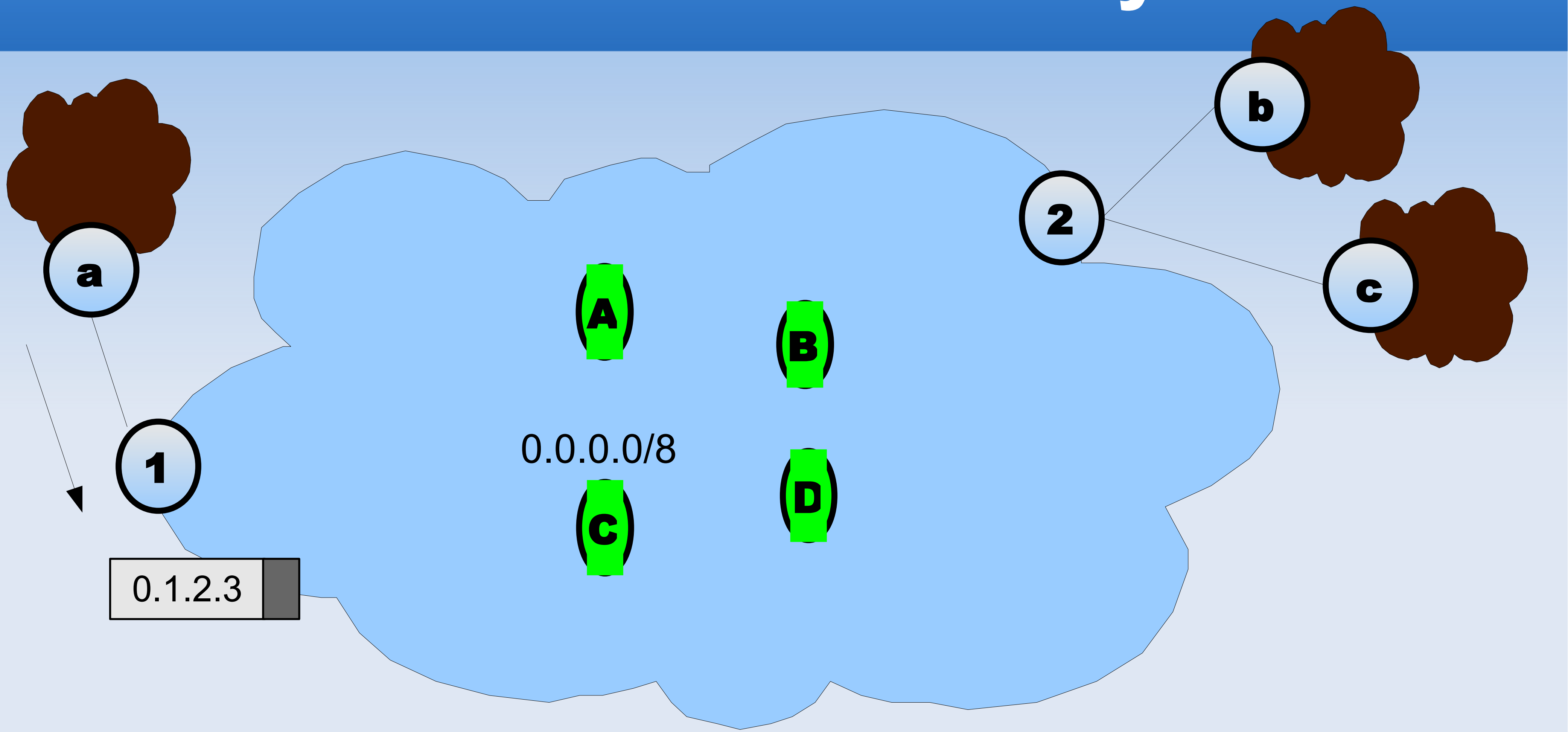
- D routers announce Virtual Prefixes, representing the range of addresses for which it has more specific information.

VA Primer



- For each external peer (a,b,c), a mapping is created between the external peer and a label (could be any tunneling).
- Both D and ND routers store these mappings in FIB.

VA Packet Delivery

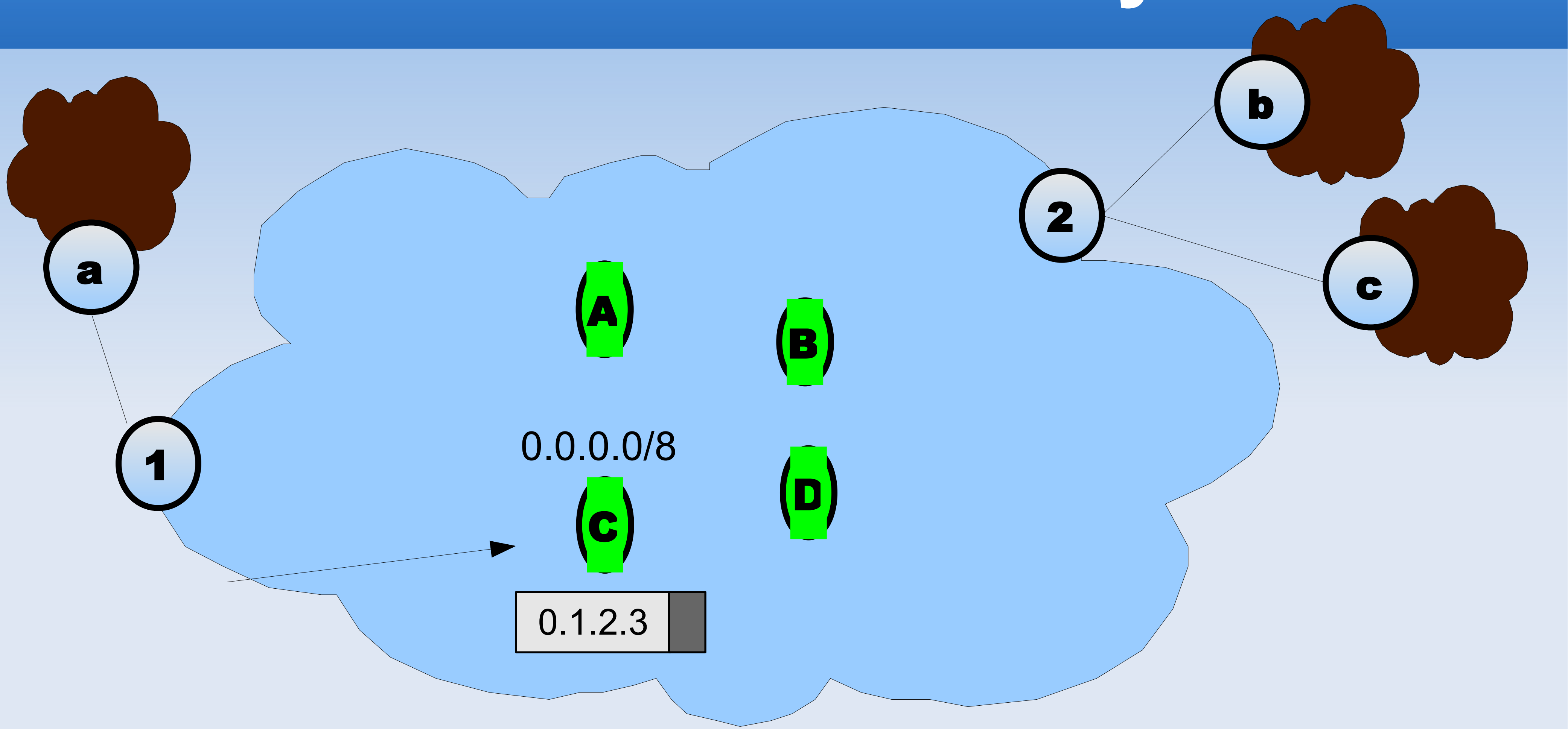


Assume:

- destination 0.1.2.3 is supposed egress ISP out of external peer 'b'

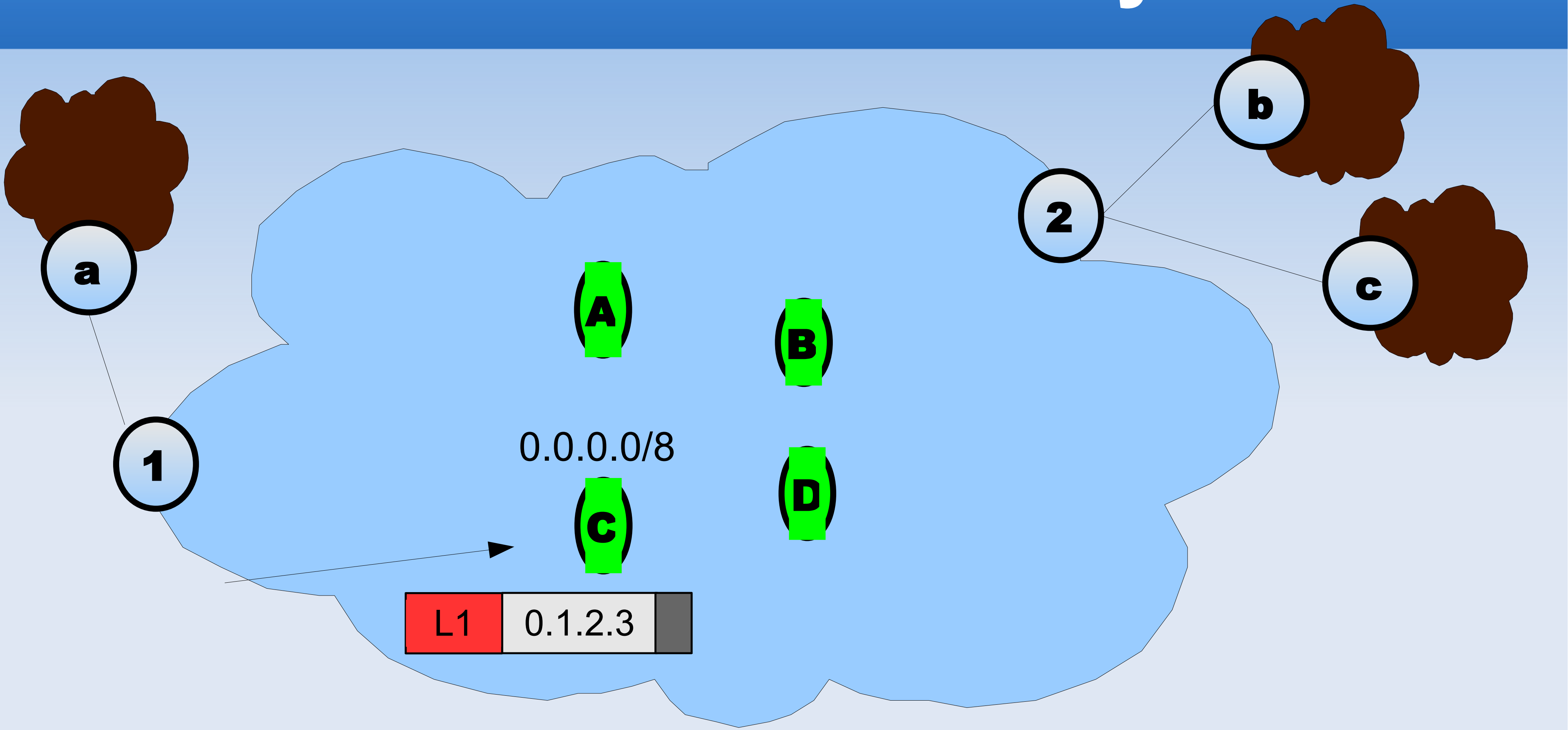
- directory router 'C' is carrying all FIB entries under 0/8.

VA Packet Delivery



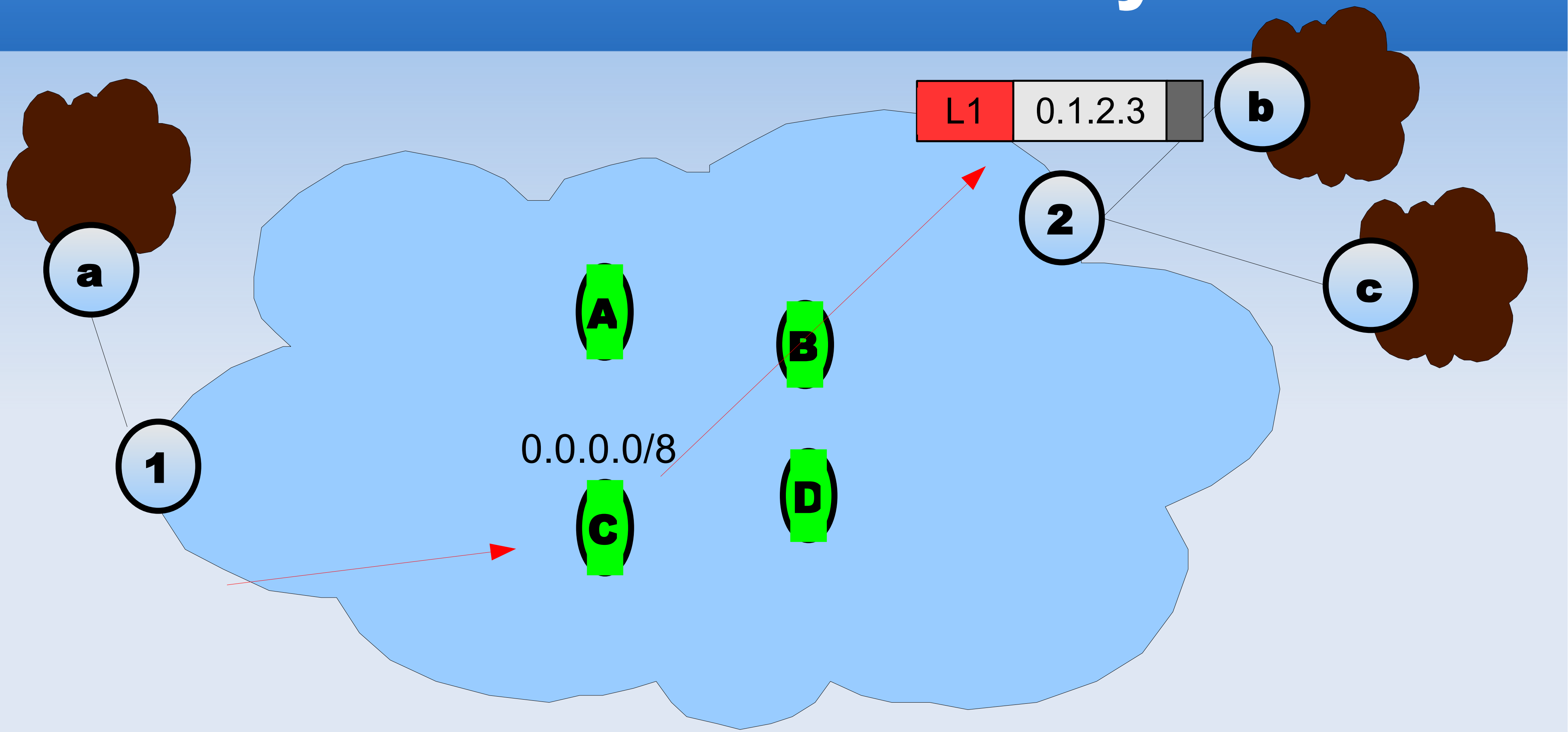
- ND router '1' matches dest address with 0.0.0.0/2, delivers to A.

VA Packet Delivery



- 'A' looks up proper egress point for 0.1.2.3, which is external peer 'b'.
- 'A' encapsulates the packet with the proper label for 'b'.

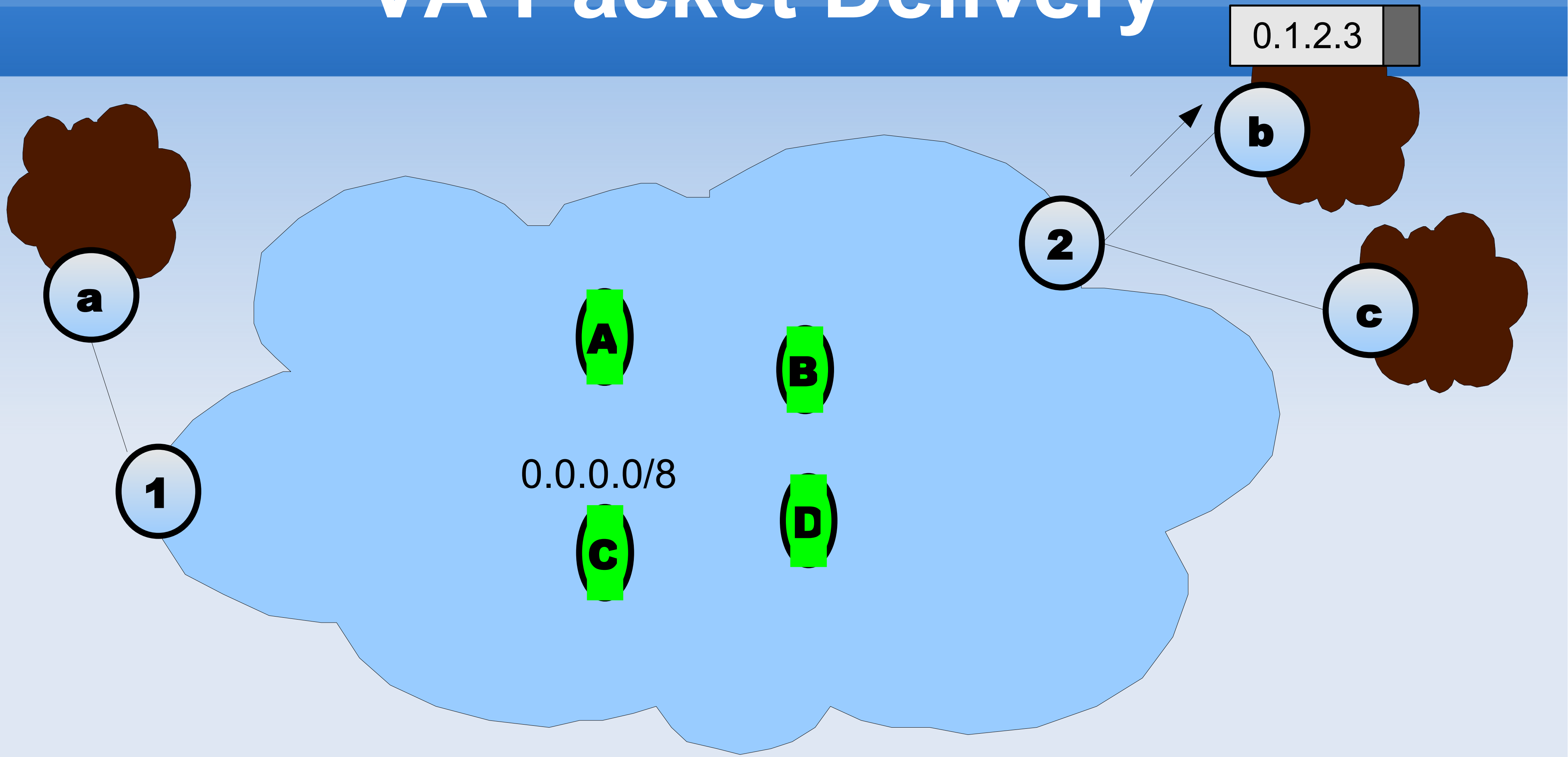
VA Packet Delivery



- Packet is delivered using the label to router 2.

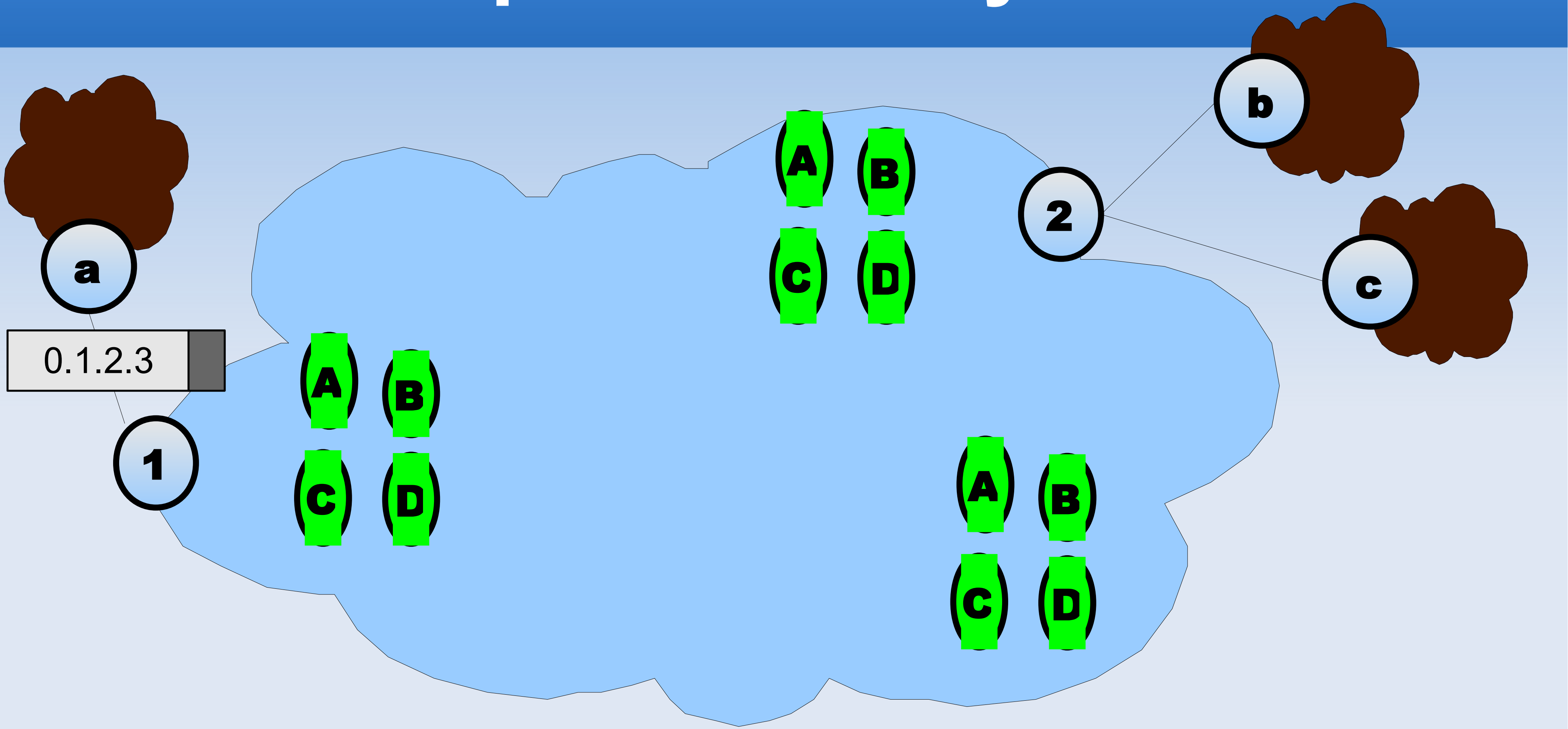
- Note the STRETCH: 1-C-2 instead of 1-2 directly.

VA Packet Delivery



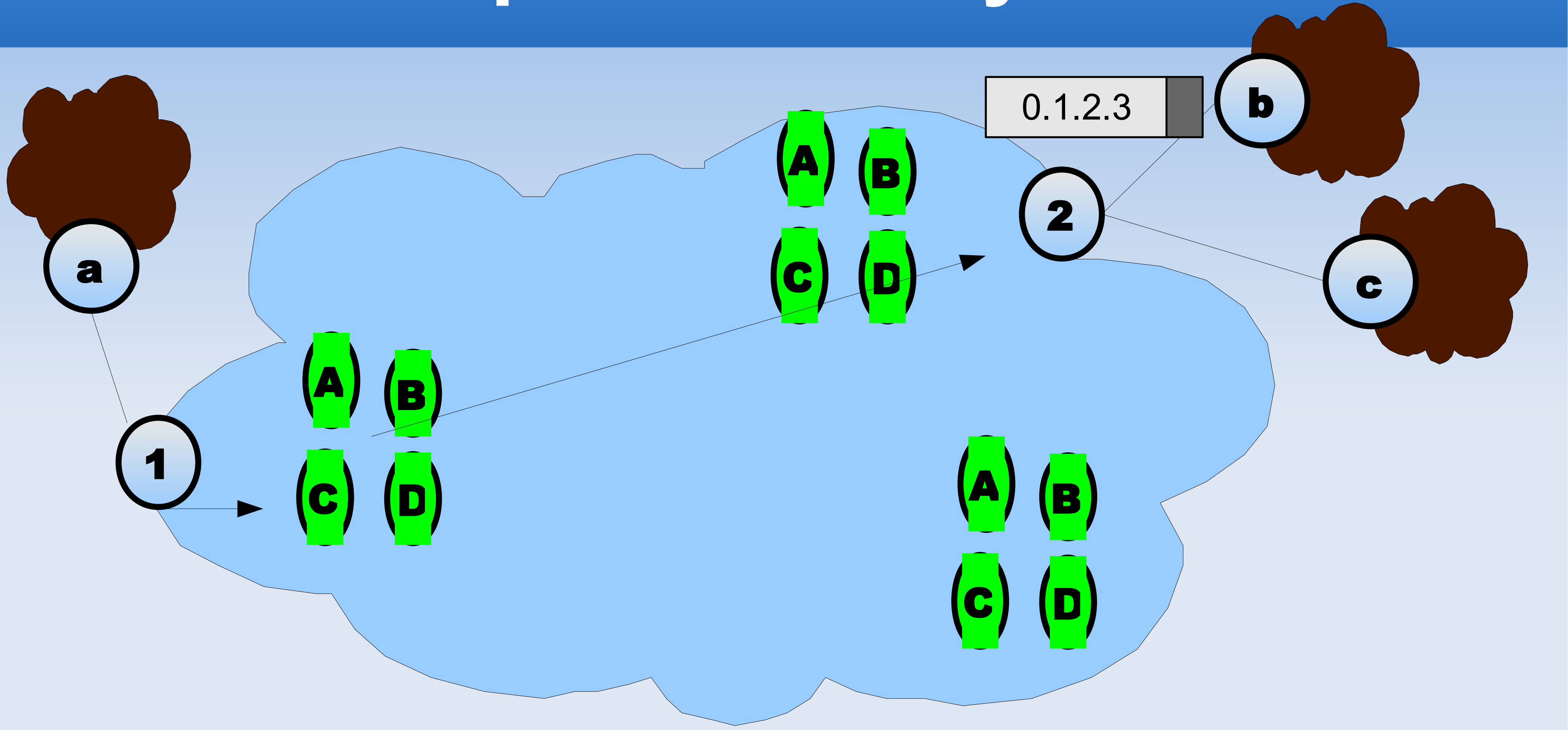
- Router 2 is configured so that any packet encapped with L1 gets decapped and sent to external peer 'b'.

Multiple Directory Sets



- ISPs will likely deploy multiple directory routers for robustness.
- Placement of these directory routers will affect performance!

Multiple Directory Sets



- ND routers send packet to nearest directory router.

- Stretch is reduced.

- But more routers need to be directories (less savings)

VA Tuning Knobs

- # of routers you would like to distribute the FIB over.
    - i.e. # of directory routers in a directory set.
- Number of redundant directory sets to deploy
- Locations of directory sets.

The VA Stretch-Savings Tradeoff: How good is it?

- Do we need optimal values for each knob to realize a good stretch-savings tradeoff?
- Can we realize a good stretch-savings tradeoff without any optimizations?
- Let's find out.

Outline

- Motivation
- VA Primer
- Evaluation Setup
- Evaluation Results
- Concluding Remarks

My Evaluation

- Determine the topology of a real Tier-1 ISP from iBGP feeds and some topology information provided by ISP.
- Choose very basic tuning knob values based on the topology.
  - No optimizations whatsoever
- Analyze the savings and stretch the ISP receives.

Some Topology Characteristics

- For each North American POPs, I counted the number of routers storing the full DFZ.
- Less than 15% of POPs are “major POPs”.
- Other 85% have very few routers with full DFZ
- Exact numbers concealed for confidentiality

Straightfoward Tuning Knob Values

- Let's just put 1 full directory set in each major POP, and see what happens.

- # of routers to distribute the FIB over.
  - 8 (all major POPs have enough routers for this)
- # of redundant directory sets to deploy
  - 1 per major POP (less than 15% of all POPs)
- Locations of directory sets.
  - Same as location of major POPs

Evenly Distributing FIB using /8 VPs

0/8 – 64/8 :	34321 prefixes
65/8 – 74/8 :	35840 prefixes
75/8 – 119/8:	34410 prefixes
120/8 – 189/8:	34836 prefixes
190/8 – 199/8:	36999 prefixes
200/8 – 203/8:	34405 prefixes
204/8 – 210/8:	36069 prefixes
211/8 – 255/8:	29520 prefixes

FIB Savings Calculation

- D router FIB contains:
  - 1/8th of DFZ
  - Virtual Prefixes
  - Egress → Label mappings
- ND router FIB contains:
  - Virtual Prefixes
  - Egress → Label mappings

Stretch Evaluation Methodology

- For each non-major POP
  - Tracerouted to each major POP.
  - Determined the one-way time to nearest major POP
  - Calculated the worst-case stretch the small POP can experience.

Calculating Worst-Case Stretch for POP

- Worst case stretch occurs when directory router is in the opposite direction of destination.
  - Destination ---- Source <-----> Directory.
- Extra stretch is from source to directory and back to source.

Outline

- Motivation
- VA Primer
- Evaluation Setup
- Evaluation Results
- Concluding Remarks

Savings for Directory Routers

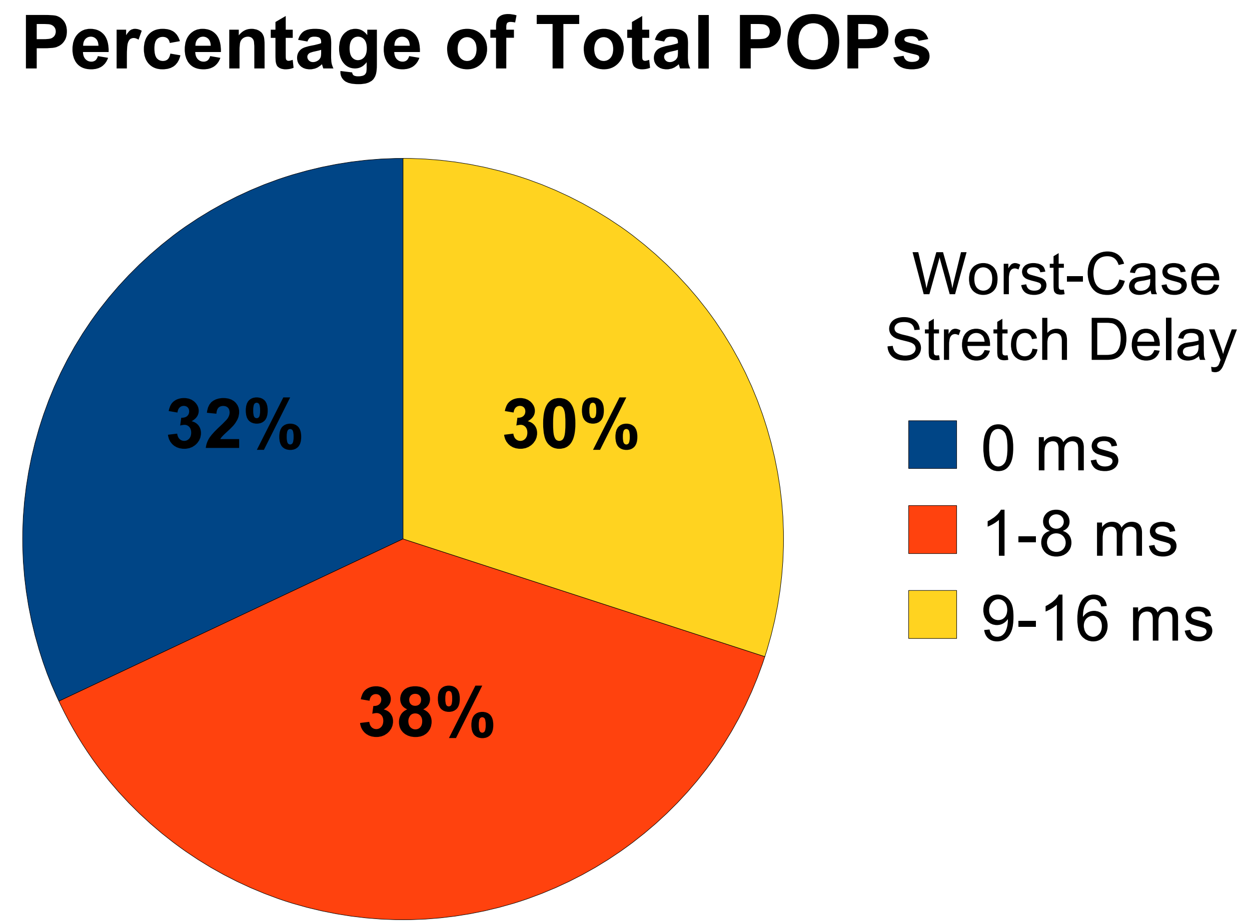
- D router FIB contains:
  - 1/8th of DFZ (~35k, 37k worst case)
  - Virtual Prefixes (256 /8s)
  - Egress → Label mappings (~20k)
- Net Savings: 80% FIB reduction

Savings for Non-Directory Routers

- ND router FIB contains:
  - Virtual Prefixes (256 /8s)
  - Egress → Label mappings (~20k)

- Net Savings: 90% FIB reduction

Stretch Evaluation Results



Conclusions from Stretch Eval Results

- All POPs are within 8ms of major POPs
  - Which is why worst worst-stretch is 16ms
- 32% of POPs experience no additional stretch
  - Some are major POPs
  - Some default to major POPs

Overall Comments on Evaluation Results

- Results apply to a **non-optimized** deployment of VA for the North-American segment of an ISP.
  - Optimizations can change results.
- Results should apply to other ISPs if:
  - ISP has at least a few large POPs containing several backbone routers.
  - Smaller POPs can reach a nearby large POP in short time.

Outline

- Motivation
- VA Primer
- Evaluation Setup
- Evaluation Results
- Concluding Remarks

VA isn't a full RRG solution

- VA just scales FIBs
  - No RIB relief
  - No Churn Insulation
  - No Separation of Locators and Identifiers

But VA has value to RRG

- VA can buy us time to roll out other scalability solutions
  - General consensus that FIB is most immediate concern.
- VA can possibly be one of many small steps which lead us to an overall scalability.
  - http://www.cs.ucla.edu/~lixia/draft-zhang-evolution-01.txt

Acknowledgements

- I'd like to acknowledge all that have assisted me directly and indirectly on this presentation.
- Lixia Zhang, efit Team, Dan Massey, Eric Osterweil, Shane Amante, Chris Morrow, Joel Halpern, Jason Schiller, David Oran, Ricardo Oliviera, Paul Francis

Thanks, RRG

- Q & A starts now.

Backup Slides

- Subsequent Slides not part of presentation.
- Bonus Information

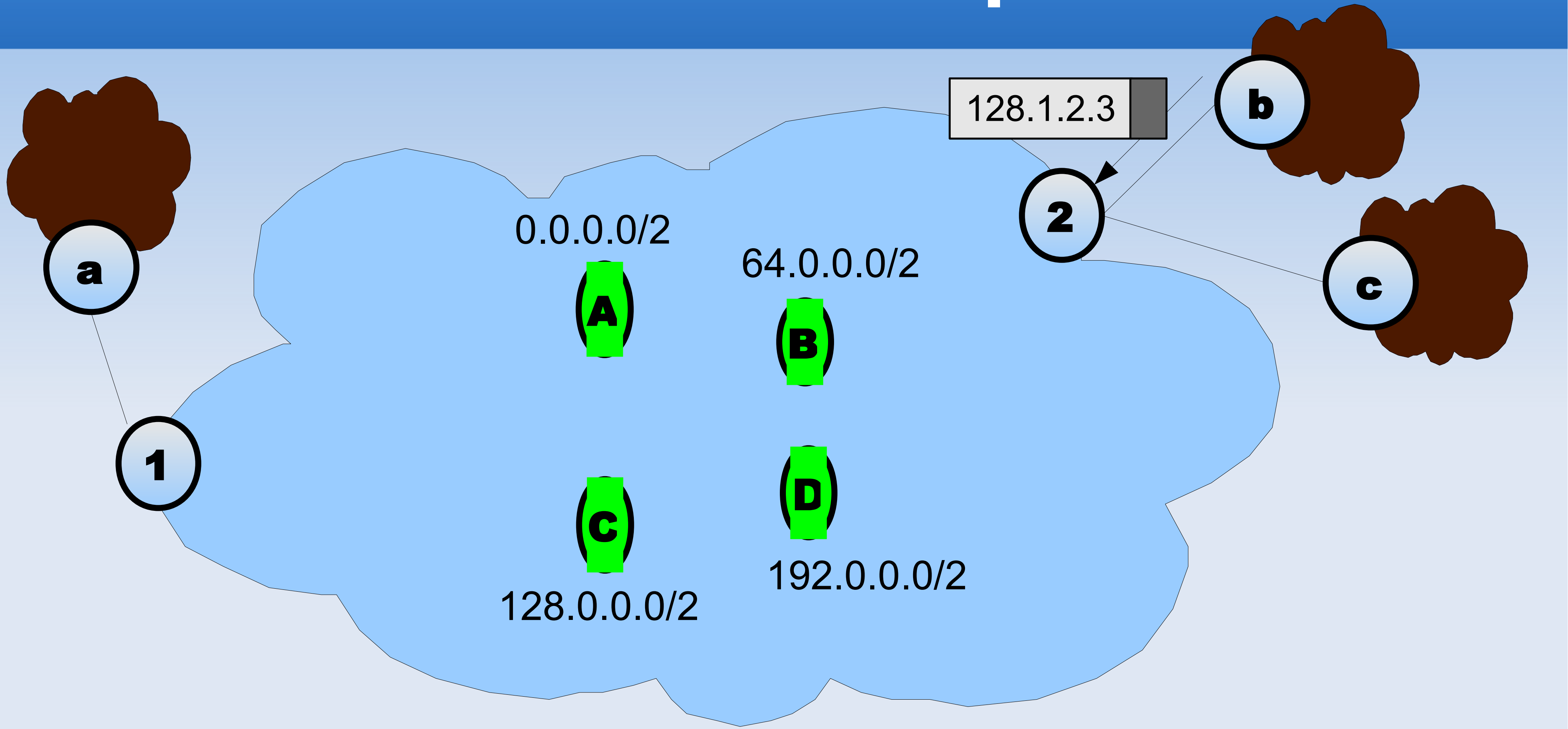
Virtual Aggregation(VA): FIB Resource Pooling

- As Mark Handley has stated in the past, resource pooling is done all the time.
  - Multihoming: pooling reliability.
  - Bittorrent: pooling upstream capacity
- Essentially, VA is resource pooling between the many line cards owned by an ISP.
  - ISPs have many routers, and each store 1 or more copies of the full FIB.
  - VA says: “Why not pool the storage of your routers and store a piece of the FIB on each router?”

This Talk: Evaluation of VA

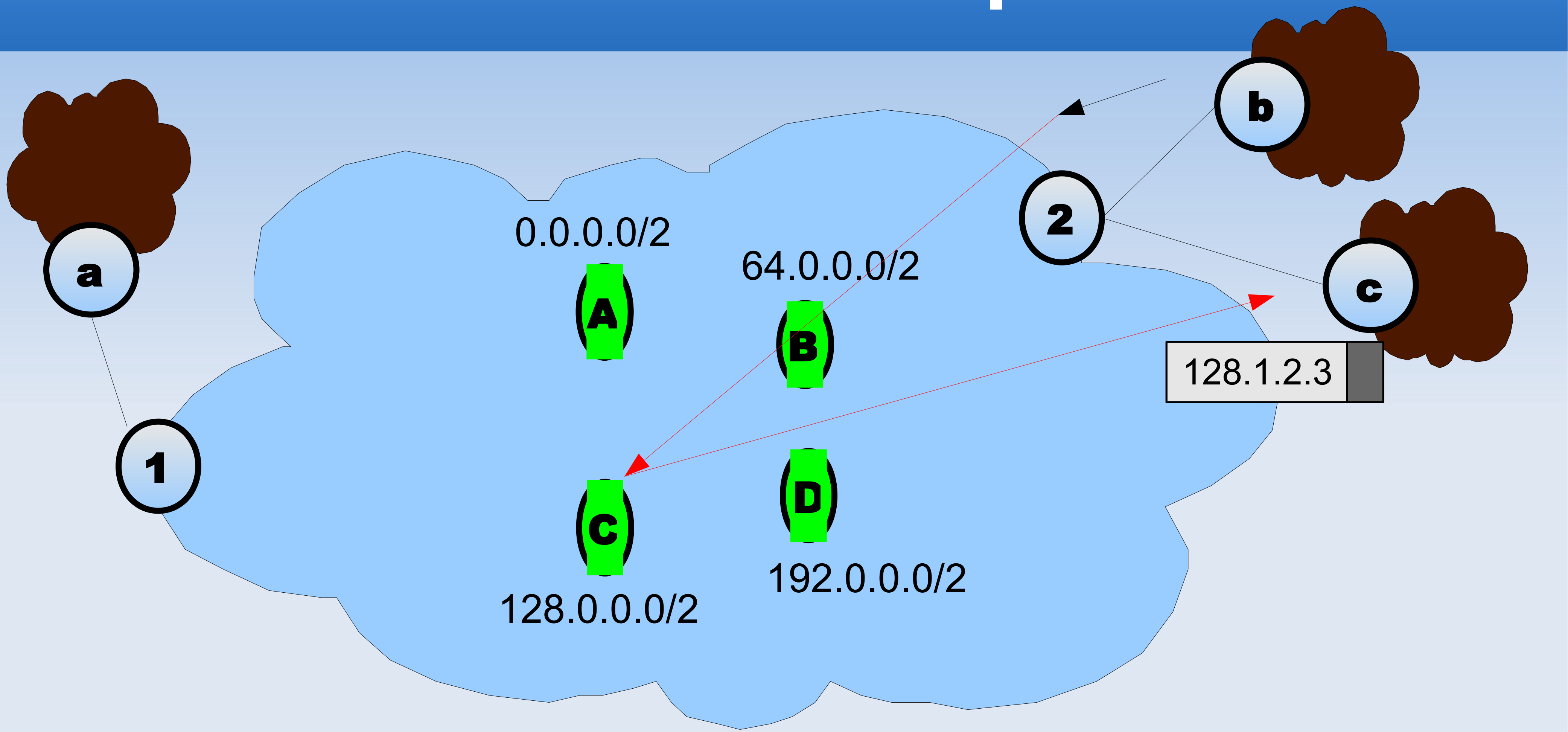
- VA can concentrate a lot of traffic onto a small set of nodes.
  - But how much traffic?
- VA can create suboptimal paths (“stretch”) for packet delivery.
  - But how much stretch?
- This presentation tries to answer these, and now I present results.

Stretch Example



- Assume destination 128.1.2.3 is supposed egress ISP out of external peer 'c'

Stretch Example



- Instead, packet goes from '2' to 'C' to 'c'. Red arrows represent additional 'stretch' due to VA.

Quibbling

- “If you just moved some routers around, you would have THIS topology with no stretch”
  - “That's too much trouble!”
- “You could probably buy a few new directory routers to eliminate stretch”
  - “Could you really? A Directory Set in each POP?”

Some Constraints for Choosing Variable Values

- No unrealistically complex optimizations.
  - Constantly doing an exhaustive search of the best placement of directory routers to minimize stretch at any given time.
  - Constantly monitoring traffic load to directory routers to minimize overloading links.
- Don't move routers around (keep topology).
- Don't purchase new routers.
  - All D and ND routers should be existing routers.

Is Stretch Avoided Entirely?

- Of course not.
- For this to happen, we would need to have a full directory set in every POP.
  - Many POPs have 2 or fewer routers storing the full DFZ in FIB.
  - Putting a full directory set in those POPs would violate our constraint of not purchasing new routers.

How Good are the FIB Savings?

- RAWS report estimate: DFZ increases 30% every 2 years.
  - http://tools.ietf.org/html/draft-iab-raws-report-02#section-4.5

- Assuming 8 VPs, it would take 12 years for directory routers to exceed 200k FIB entries.
- It would take 24 years for directory routers to exceed 1 million FIB entries.

How Good are the FIB Savings? (cont)

- RRG wants solution that scales for the long term. VA does this for FIB size.
- RAWS report: ISPs can increase FIB capacity by 30% each 2 years at constant cost, while DFZ grows 30% each 2 years with occasional bursts.
- With VA and 8 VPs, FIB capacity can be increased 240% each 2 years at constant cost, which exceeds the rate of the DFZ growth.

How Overloaded are Directory Routers?

- Concern that too much traffic will be concentrated to directory routers.
- This could overload the routers as well as their links.

How Overloaded are Directory Routers? (cont)

- Common believe that vast majority of traffic goes to very few destinations.
- VA team observed netflow records from 11/07-1/07 for major tier-1 ISP.
- Results: 90.2% of traffic goes to 5% of destinations.
    - Study to be published at NSDI next month.
- Nearly no change in popular prefixes over this time.

How Overloaded are Directory Routers? (cont)

- Thus load on directory routers can greatly be reduced if popular prefixes are FIB-installed.
- ND routers would still save over 85% of FIB,
- D routers still save over 75%.

How Bad is the Stretch?

- 16ms is the worst case on a very simple, no-cost setup of VA.
  - ISPs could optimize topology to reduce the stretch if it desires.
- Though I assigned VPs to routers, we really just need to assign VPs to line cards.
  - FIB can be divided amongst line cards on the same router, reducing stretch to within a single router!
  - If we do want to go this route, we should look into this option.

Summary of Tradeoff

- Net Savings:
  - D routers: Over 80% FIB reduction
  - ND routers: Over 90% FIB reduction
- Stretch:
  - Worst-Case stretch: 16ms
  - Avg-Case stretch: 8ms

How Applicable Are Eval Results?

- Rocketfuel study showed this to be true of all T1 ISPs studied.
  - http://www.cs.umd.edu/~nspring/talks/sigcomm-rocketfuel.pdf
- While study was from 2002, we believe these properties should still hold for T1 ISPs today.