BGP path 'hinting': A New Way to Influence *Return* Routing

Brent Sweeny
Global Research NOC at Indiana University (USA)
sweeny@iu.edu

Terena Network Conference 19 April 2014 (Dublin)

Topics

Purpose (what's this all about?)

Why? (why might this be a *good* idea?)

Why not? (why might this be a *bad* idea?)

How? (how might it be done?)

How? (possible variations)

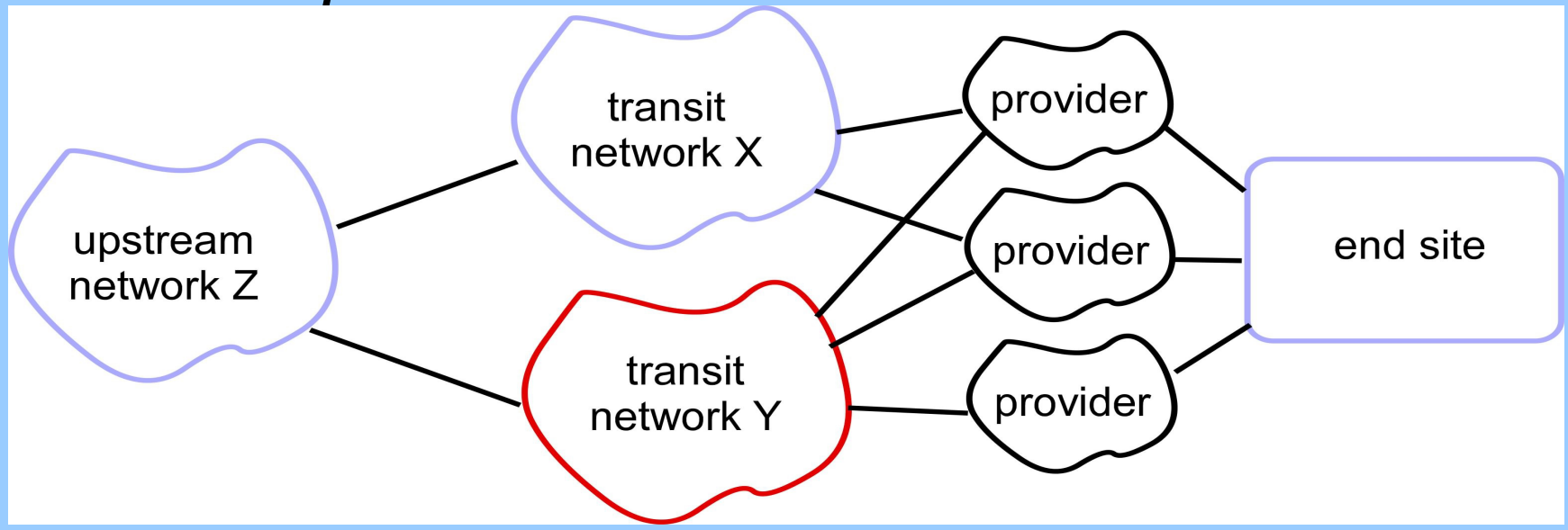
What's necessary to make it work?

Demonstration

What next?

Purpose

Allow *end sites* to 'hint' or suggest to intermediate networks a path the end site would *prefer* traffic take back toward them



Use well-known BGP community values

Its use is optional to intermediate networks

Why? (p.1)

Some user-oriented reasons:

Unequal paths toward user, end-user wants to direct

Varying criteria for preference of a path—even by the same site at different times—for example:

- Lowest latency, irrespective of other measures
- Highest bandwidth, irrespective of other measures
- Symmetry
- Avoid lossy path
- Move traffic from a path, e.g. for a large demo

Existing alternative methods aren't sufficient: there is no *good* way for end-users in multi-tiered, multihomed networks to indicate to a network more than a layer away how best to return traffic. (see “what's wrong”)

Why? (p.2)

Some 'community' and implementation considerations:

'R&E network community' approach may help establish a new convention (as with jumbo-MTU BCP)

Common, documented approach easier to debug through net than current potpourri

'Normalized' approach could allow for selection to be programmed for general cases, not one-off exceptions to your normal routing policies

Intermediate networks may use (default) criteria for path selection different than end-site's but may take requests into consideration

Why #2: What's wrong with current alternatives?

Existing methods which work in some simple cases are insufficient, or Bad Ideas... for example:

- MEDs—only communicates to next AS
- AS-prepending—blunt tool: affects all who hear it
- Sending more-specifics—blunt tool: affects all who hear it, some nets refuse them (and no scheme will work with very small netblocks for that reason)
- Withdrawing announcement in some directions—*very* blunt tool
- Local-pref 'nudging' using per-network hint schemes (e.g. Internet2, NLR, carriers each have different schemes, others have none). Per-vendor communities are scoped only per-vendor: a more generalized solution would be useful.

There's no other good way to do this!

Why not?

What problems *could* it cause?

- Does anyone remember “ip source-route”?
- Do end-sites know something about topology and policy (esp for intermediate networks) that the transit networks don't? (answer: sometimes, yes, but enough to override your policy?)
- Wrong people making the decision about traffic engr?
- How do you know the actual end-site really requested this? (the “rogue transit network”: could it be a DoS? Do you trust your peers?)
- Added complexity to routing, troubleshooting
- Will it scale to lots of networks? Is it maintainable?
- Some networks filter “local” BGP communities

How? The BGP Community

Defined in RFC1997, 'extended' in RFC4360.

A 32-bit (or 64-bit 'extended') value commonly encoded 16bits:16bits with an *Autonomous System number* in the 1st field and a *value* with meaning *to that AS* in the 2nd: e.g. 11537:950.

Used to 'mark' prefixes with values that can be used in policies for arbitrary special treatment (higher, lower, refused, classification, etc.)

Generally (but not always) transitive.

There are some pre-defined “well-known community” values.

How?

Very simple first step: Propose a 'well-known' consensus-agreed-upon set of unique BGP communities that work in the same way among any transits who choose to participate

Attached to hinted prefixes by originator of the BGP announcement, who is the owner of the 'hinted' destination

Format: (*"well-known" hinting value)* : *(AS-path-selection)*

e.g. 60000:11537

Where the *hinting value* is an well-known number, and *AS-path-selection* is the 'hint' to what network to prefer for return traffic toward the originator of the prefix.

Read as "please, whoever sees this, when you're deciding which available path to send traffic to me, I'd prefer you use AS11537".

Some networks in the path must pass the signal upstream

Participating transit-nets modify their policy to implement hinting *any way they choose*

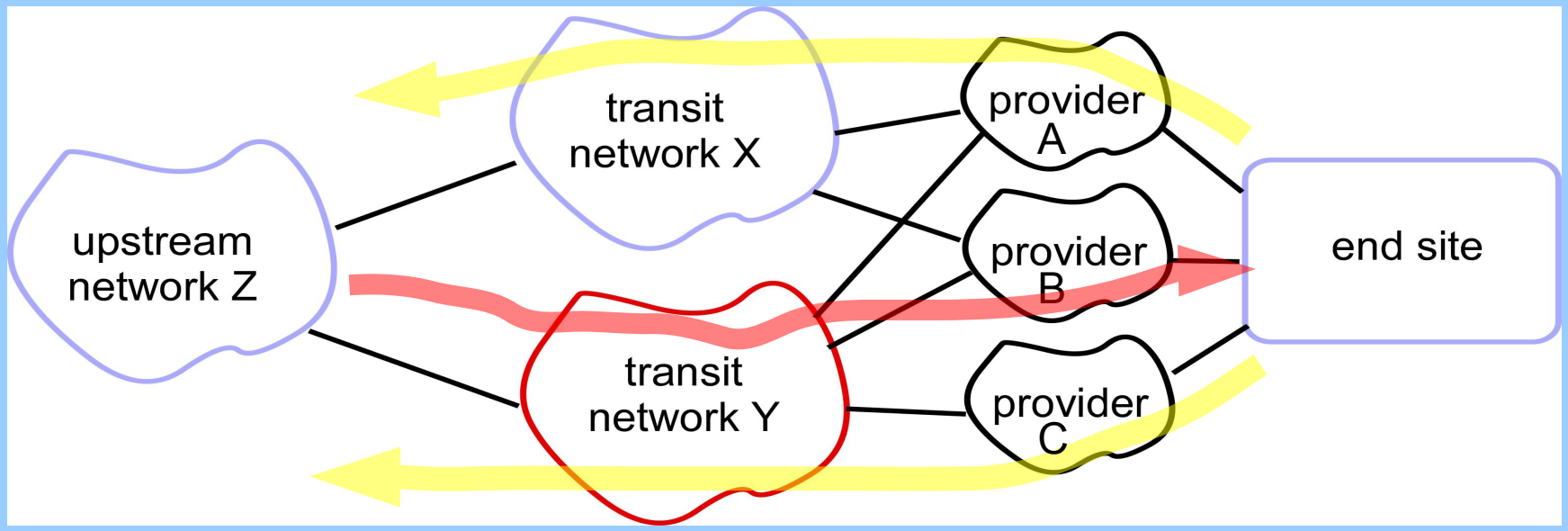
Possible Enhancements

Degrees of preference (prefer path through network A less, B more)

Hierarchy of preference (prefer path through network A first, network E second, network K third)

Mark or signal continents or countries to avoid suboptimal paths

How? (example)



End site sends 'hint' upstream: prefer path via transit network 'Y'

Upstream network 'Z' honors the request and prefers 'Y' path *for end site's requested prefixes*

Endsite could change preference at any time (and so could 'Z'!)

What's necessary to make it work?

1)General agreement on a method (“critical mass”)

2)Willingness at some user sites to use this method for hinting

3)Agreement by some R&E transit networks—ideally those who carry traffic for users in #2 above—to implement that method (#1) and to honor at least *some* requests

It's (slightly) past the talking phase...

It's been done, as a proof-of-concept, in a multipath, multi-AS network

Demonstration

Logic:

If I want to to honor hinting requests from peers Q & M

On inbound policy from each of Q & M:

If a prefix has the *hinting* community 'marking'

Then give it a high local-preference, e.g. (JunOS):

```
policy-statement HINTS
 term HINT-I2
    from community HINT-I2  <which is 60000:11537 today>
    then local-preference 5000
 term HINT-NLR
    from community HINT-NLR  <60000:19401>
    then local-preference 5000
```

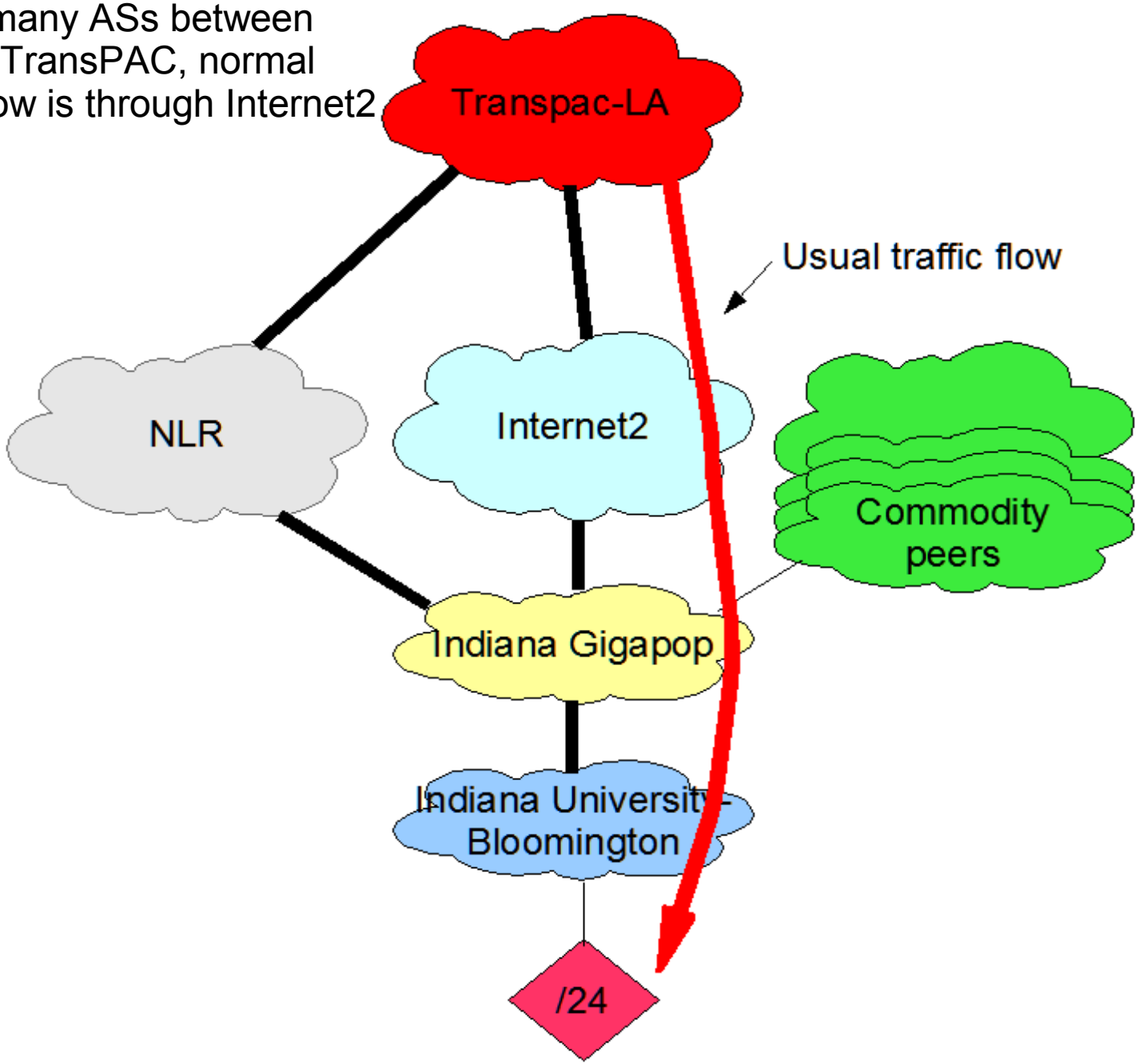
Result: prefer path through I2 or NLR toward that prefix

Else (other peers or no hint) leave unchanged, apply normal policies and BGP path-selection

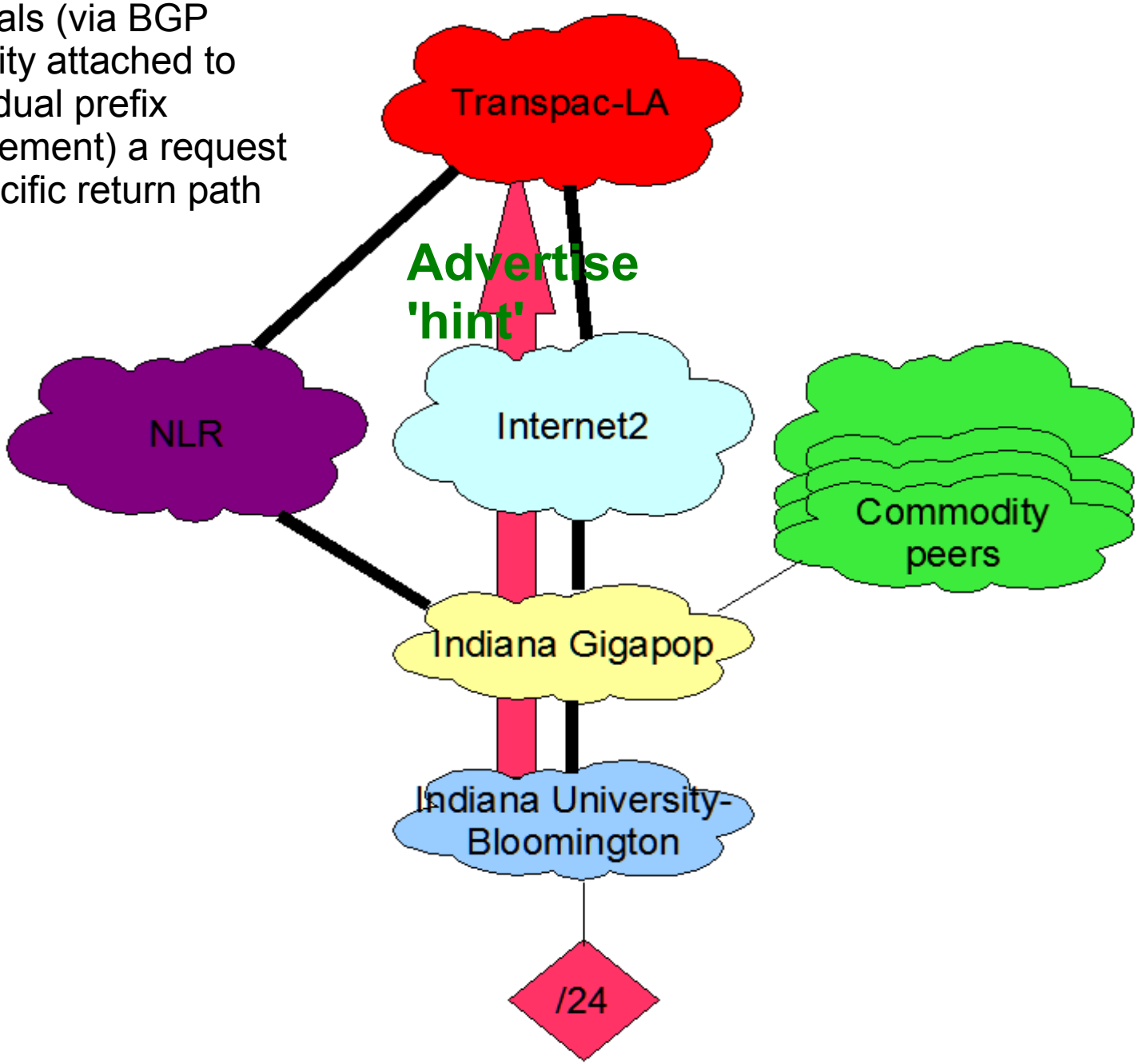
How does it work?

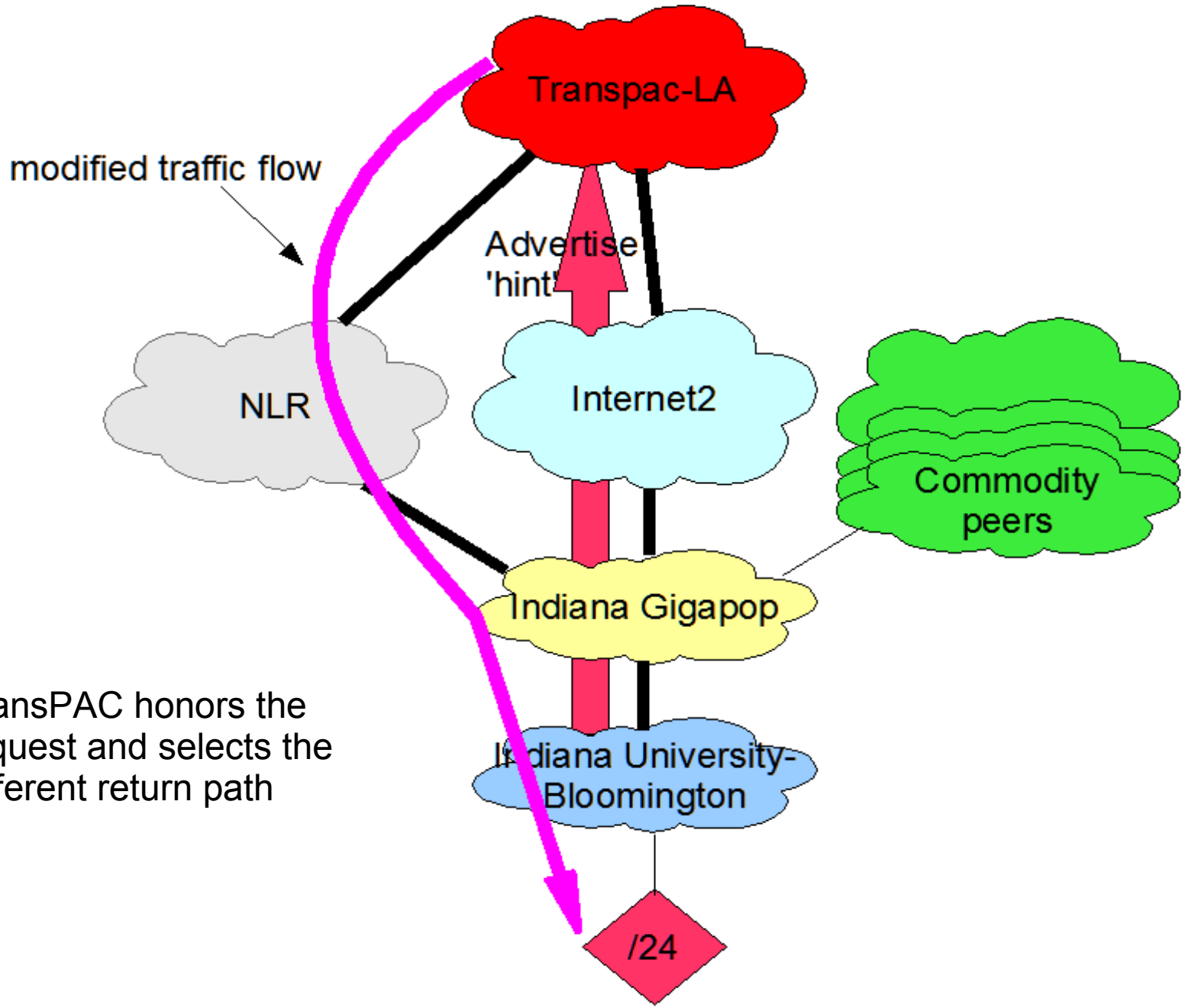
(bird's-eye view)

Before: many ASs between
IUB and TransPAC, normal
Traffic flow is through Internet2



IUB signals (via BGP community attached to an individual prefix announcement) a request for a specific return path





TransPAC honors the request and selects the different return path

How does it work?

(the details)

Construction of the 'hint' signal

- Originator of the prefix attaches a BGP community to the prefix containing the 'hint'
  - BGP communities conventionally have two parts:
    - 16 bits of autonomous-system number, usually a marker for who should listen
    - 16 bits arbitrary to the listener
  - The hinting community uses a unique 'well-known' ASN specific to this function: 27198; the 2nd field is the ASN of the network through which you request upstream network operators to send the traffic back towards you.
  - e.g. Please send back through APAN: 27198:7760.

Requesting site marks traffic with 'hint'

```
(JunOS:)
from {
    route-filter 129.79.9.1/32 exact;
    # could also use prefix-list
}
then {
    community add 27198:19401;
}
```

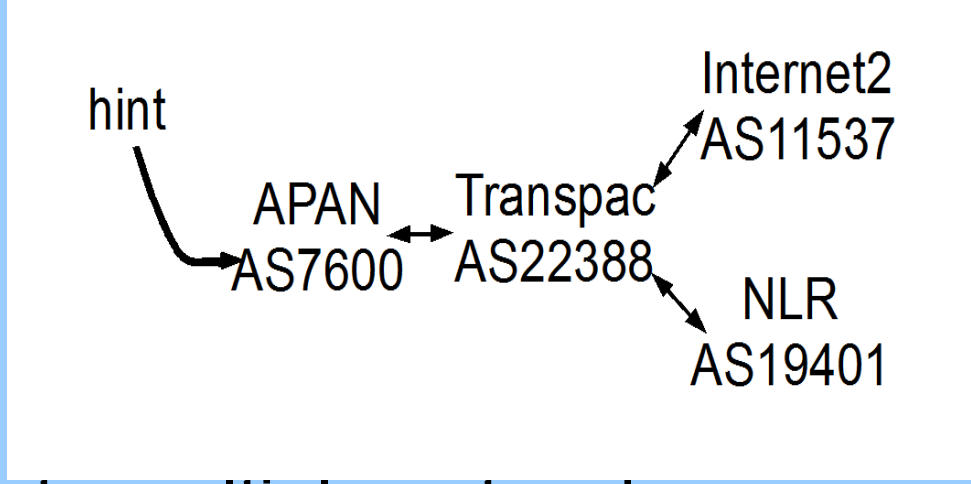
Upstream site honors 'hint'

(JunOS configuration in the TransPAC router):

```
    community HINT-I2 members 27198:11537;     (11537 is Internet2's AS)
    community HINT-NLR members 27198:19401;  (19401 is NLR's AS)

policy-statement HINT-IN {                    (add to BGP import policies)
    term hint-I2 {
        from community HINT-I2;
        then {
            local-preference 5000;
            next-hop 207.231.240.131;
            next term;
        } ..
    term hint-NLR {
        from community HINT-NLR;
        then {
            local-preference 5000;
            next-hop 207.231.241.14;
            next term;
        } ..
```

Watch out for loops!



- A loop is possible IFF:
  - 'signaling' site requests hints to multiple networks
  - Transit site is connected to >1 of them
- To avoid this:
  - Be careful initiating multiple hints
  - Transit site may be able to avoid with more complicated policy, by examining AS-path or by examining hint-requests for directional validity—you want to send traffic only toward requester

Current status

- Did proof-of-concept from Indiana University through Indiana Gigapop, NLR & Internet2, to TransPAC in Los Angeles to influence altered return path
- Worked with ARIN to procure unique AS27198 used for 'signaling'; received Feb'09. *(This is a 'paid' ASN; it'd be better to be able to get a long-term experimental one.)*
- Several expressions of interest from R&E nets; need to convert those to action; working to publicize more
- Plan to document as draft RFC
- Goal to have path-hinting in place in some R&E networks for Supercomputing conference

Mailing list

- A Sympa mailing list for those interested in further discussion of this path-hinting idea has been created.
- Its name is bgp-hinting-l@indiana.edu
- To subscribe, send a note to list@list.indiana.edu with this *subject line*:

  `subscribe bgp-hinting-l <firstname> <lastname>`

  and nothing in the body.

Selected References 1

Chandra, Traina, Li, RFC1997, “BGP Communities Attribute” (1996).

Chen & Bates, RFC1998 “An Application of the BGP Community Attribute in Multi-home Routing” (1996).

Sangli, Tappan, Rekhter, RFC4360, “BGP Extended Communities Attribute” (2006).

Meyer, RFC4384/BCP114, “BGP Communities for Data Collection” (2006).

IANA, “Data Collection Standard Communities per RFC4360” (2007).

Olivier Bonaventure et al., Internet Draft (Draft-bonaventure-bgp-redistribution), “Controlling the redistribution of BGP Routes” (2002).

Selected References 2

Jin Tanaka (JP-NOC/KDDI), “BGP Routing with Communities”, presented at the 24 APAN Meeting Network Engineering Workshop, August 2007 and the October 2007 Internet2 Member Meeting RENOG session with additional suggestion from Akira Kato.

Robert Raszuk (Cisco) “BGP Wide Communities”, presented at NANOG49 (2010) see http://www.ietf.org/proceedings/78/slides/idr-2.pdf

Brent Sweeny, “BGP path 'hinting' proposal”, presented as it evolved to JET, Internet2/ESnet Joint Techs, and Internet2 Member meetings 2006-2012.

Questions? Comments? Discussion?

Thank you for your attention!

My email: sweeny@iu.edu