

BGP path 'hinting' update

Brent Sweeny
Global Research NOC at Indiana University
sweeny@grnoc.iu.edu

to:

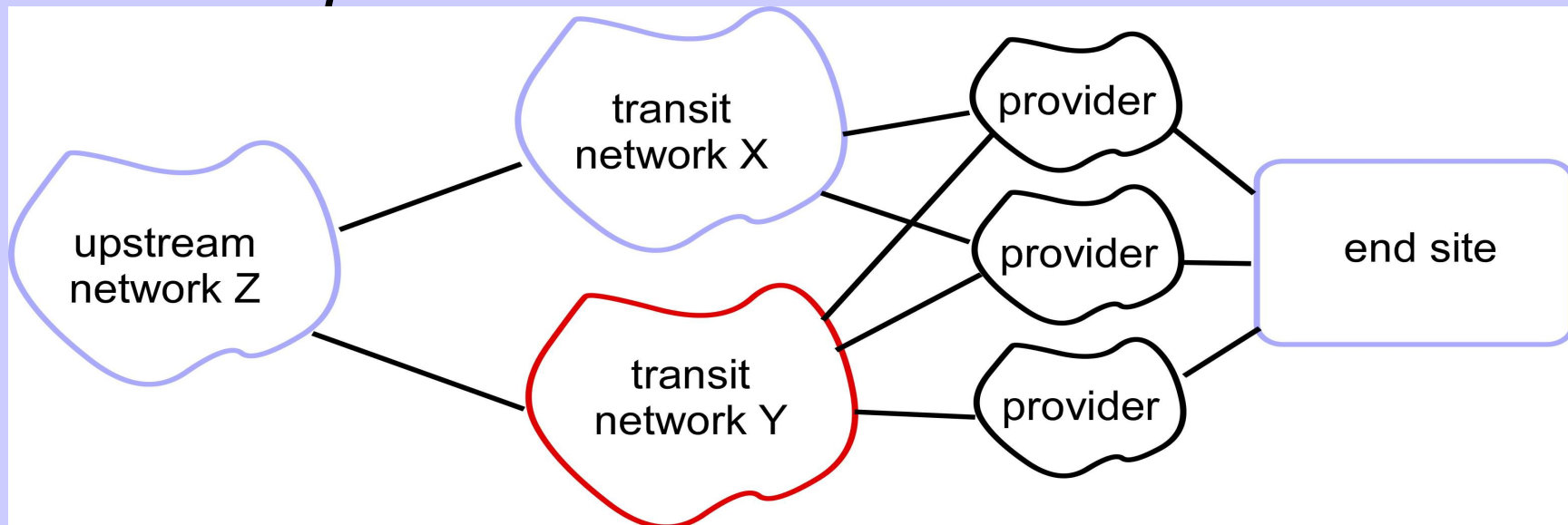
Joint Techs (University of Hawaii), 21 January 2008

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 later slide)

Why? (p.2)

Some 'community' and implementation considerations:

- 'R&E network community' approach may work (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-prepend—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)

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.

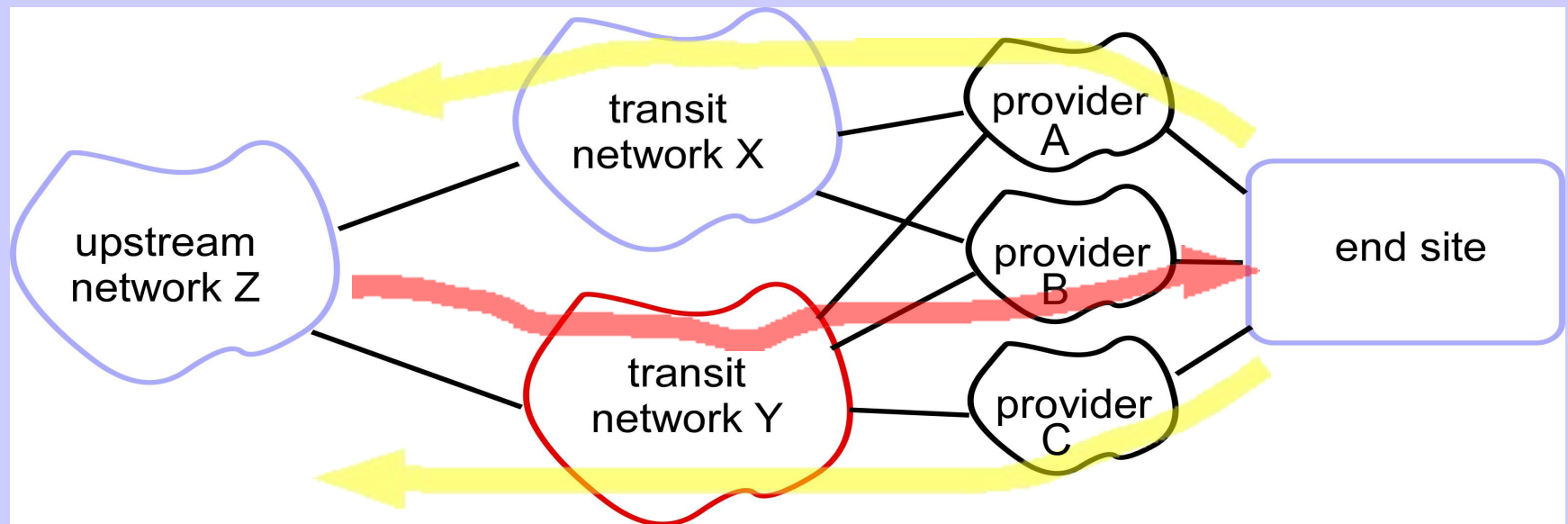
How?

- 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 arbitrary but agreed-upon number, and *AS-path-selection* is the hint to what network to prefer.
 - 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)



1. End site sends 'hint' upstream: prefer path via transit network 'Y'
2. Upstream network 'Z' honors the request and prefers 'Y' path *for end site's requested prefixes*
3. *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 (does this condition need to exist before #2 will happen?)
- (3) Agreement by some R&E transit networks—ideally those who carry traffic for users in #2 above—to implement that method and to honor at least *some* requests

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

- It's now been done, as a proof-of-concept

Demonstration

Logic:

If I want 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>
```

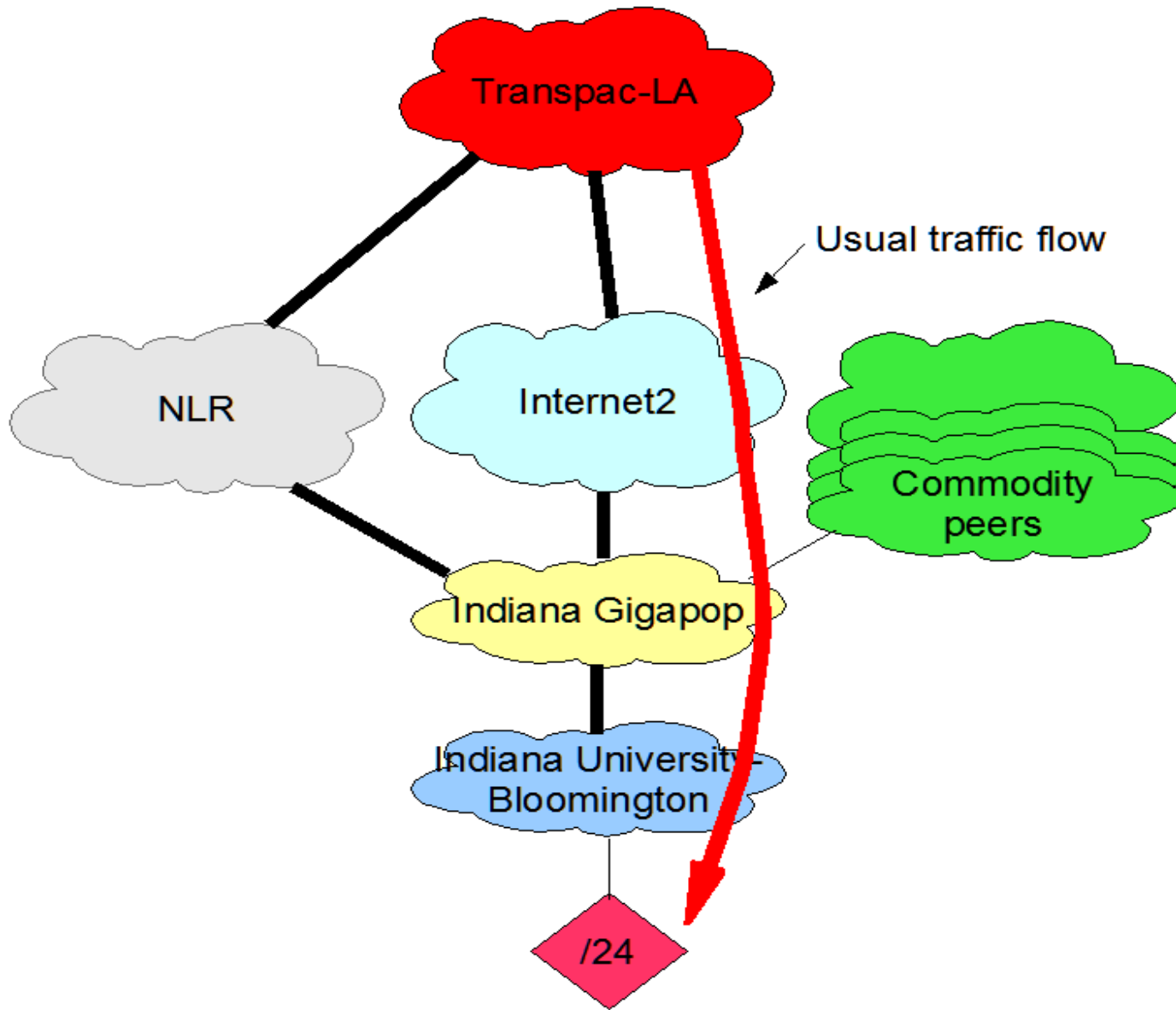
```
then local-preference 5000
```

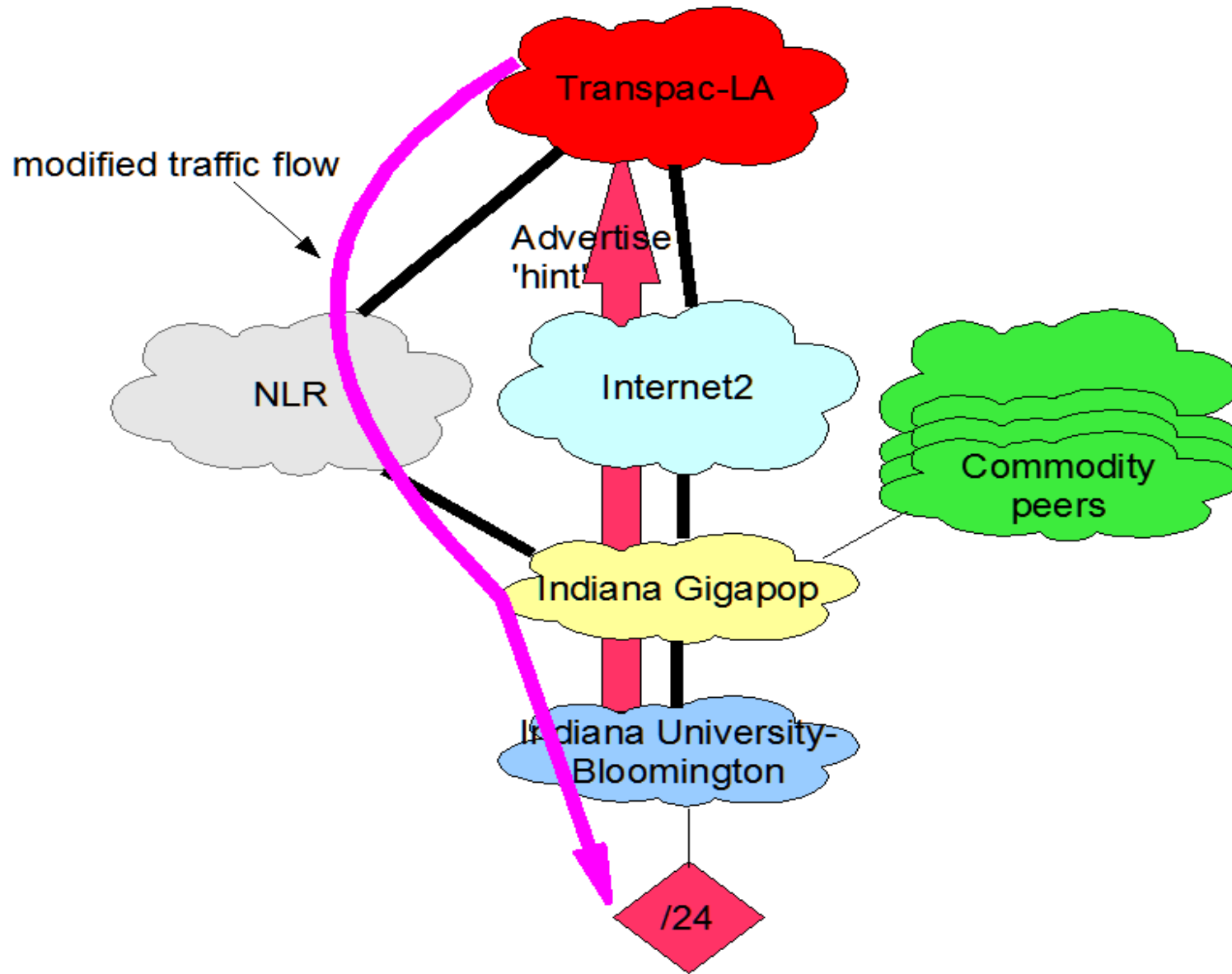
```
term HINT-NLR
```

```
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





What next?

- Further discussion by community:
 - Assume hinting is worth further investigation...
 - Where is it needed most? SC08 likely for some upstreams
 - Currently two major directions, → simpler || →more features
 - Try one? Try both and compare? (going w/*simpler* for now)
 - How to keep it manageable? (scale)
 - Discuss where? RENOG.org list?
 - How do we know when we agree? Consensus? Lack of argument?
Lack of interest? *Someone willing to try something?*
- Continue discussion: member mtg, JET, Joint Techs (NANOG?)
- Agreement on 'well-known' approach
- Public documentation of the consensus scheme (local or RFC?)
- Provide example policies for netOSs (e.g. IOS, JunOS)
- Some trial adopters
 - Need both requesting-sites and transit-nets
- Re-evaluation: Is it working as desired?
- Revision path: publicizing (RENOG?) & implementing tweaks
- [Could need/implementation extend into the 'general NetOps' space?¹⁷]

Selected References

- 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).
- 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.
- Brent Sweeny, “BGP path 'hinting' proposal”, presented as it evolved to JET, Internet2/ESnet Joint Techs, and Internet2 Member meetings 2006-2008.

Yet to be resolved

- What 'hinting' marker to use?
- Normal 32bit BGP community is 16bits:16bits, where the first segment is normally the AS of the network being signaled. In hint scheme, it's a signal to any listener that a hint is tagged.
- We don't need a "real" ASN for that signal field, but need something that won't conflict, so must avoid the 'legal' ASN range (1-64511) and some nets use private ASNs(64512-) internally.
- The demo used 60000, in IANA-unallocated.
- Possibly get extASN in 'data-collection' range?