BGP MultiNexthop Attribute - Status

https://datatracker.ietf.org/doc/draft-kaliraj-idr-multinexthop-attribute/11/

IETF IDR Interim Meeting (**interim-2024-idr-02**)

Jan 29 2024

Kaliraj Vairavakkalai
(behalf of co-authors)

Juniper Networks

Agenda

- MNH - Recap
- Changes to the draft – since IDR 118.
- Summary - WG Adoption responses
- Summary - comments and discussions points
- Next steps.

Background: Expressing nexthops in BGP (Recap 1/3)

- What is a nexthop?
  - Instructions on how to forward a payload specified in BGP NLRI.

Nexthop information is extracted from BGP PDU/Route from various portions:

- Endpoint Identifier (Where to forward?)
  - Nexthop attribute (code 3)
  - MP_REACH_NLRI attribute (code 14) : “Network Address of Next Hop”
  - Redirect to IP extended community attribute.
  - Tunnel Encap Attribute.
  - Color-only community attribute.
  - Redirect to VRF extended community attribute.
- Encap to use:
  - MP_REACH_NLRI attribute (code 14) : “Label in NLRI portion”
  - Prefix-SID attribute.
  - Tunnel Encap Attribute.
  - Repair-Label attribute.
  - **Secondary-Label attribute.** (new since idr interim, Oct-2022)
  - **FSv2 Redirect to * actions.**
- Constraints:
  - Color community or Mapping community attribute.
  - Link bandwidth community attribute.

Problems (Recap 2/3)

- ❑ Inability to advertise more than one nexthop in a route.
- ❑ Not easily extensible to newer endpoint types, encapsulation types.
- ❑ Addpath unable to express relationship between different nexthops (active/backup, UCMP etc), Scaling heavy.
- ❑ Inability to signal encap-information uniformly across families (e.g. cannot signal Labels for SAFI 1 routes).
- ❑ Inability to signal multiple labels in a route.

  Helpful in some multihomed cases to avoid label oscillation.

- ❑ Semantics of a downstream allocated label is not known to receiver.

  This info may be useful for some scenarios, e.g. network visualization, EPE decisions.

These problems are solved by MultiNexthop Attribute.

MNH – bird's eye view (Recap 3/3)

```
MNH Attribute: {
    PrimaryPath {
        [Forwarding Instruction 1],
         ..
        [Forwarding Instruction n]
    }
    RepairPath {
        [Forwarding Instruction 1],
         ..
        [Forwarding Instruction n]
    }
}
```

```
Forwarding Instruction : {
    FwdAction, FwdArguments
}
```

Changes to the draft – since IDR 118

- Removed Capability Negotiation
  - Being a negotiated Open Capability causes BGP session flap whenever config changes.
  - It also significantly increases rib-out Implementation complexity
  - Being an Optional Non-Transitive attribute stops propagation as unrecognized attribute.
  - Adding Receive side rule may be enough to stop unintended propagation across supported node also.

BGP MNH – WG Adopt Status (IDR Interim 1/29/2024)

- Call for WG Adoption
  - https://datatracker.ietf.org/doc/draft-kaliraj-idr-multinexthop-attribute/10 , version 11
- Support: 10
  - Natrajan Venkataraman
  - Aravind Prabhakar
  - Minto Jeyananth
  - LINGALA, AVINASH
  - Reshma Das
  - Robert Raszuk
  - Tony Przygienda
  - Nat Kao
  - Mohan Nanduri
  - Gyan Mishra
- See value in work, but also having questions/discussions trying to shape the draft: 4
  - Adrian Farrel
  - Igor Malyushkin
  - Satya Mohanty
  - Ketan Talaulikar
- Have queries/concerns about the proposal: 2
  - Donatas Abraitis
  - Swadesh Agrawal

Comments and discussion points – 1/2

- This draft provides an efficient and concise framework to signal a list of ordered next-hops with corresponding instructions
- operationally useful approach provides new functionalities currently unavailable with ADD-PATH
- It's great to see some OpenConfig-AFT-like approaches progressing in IETF.
- An alternative to help with RIB scaling issues in Addpath deployments
- Label oscillation prevention in Multihomed Labeled-PE/ASBRs protecting each other.
- provides an unique way to carry the label, so that it can be decoupled from the RFC-8277 NRLI in the future
- the ability to specify ordered Next hops helps in use cases where symmetric load balancing is needed
- helps to carry ordered set of one or more NHs with forwarding information scoped on a per next hop basis
- what is in the document I would consider just a framework for what it could contain down the road. I also admit we were talking about this in the past a number of times, but never came with a killer use case
- The MNH solution could provide a nice alternative to reflect all the paths without having to rely on add paths RFC 7912 and the propagation of paths maybe be in a more controlled fashion

Comments and discussion points – 2/2

- Remove Capability Negotiation?
- Why Non-Transitive attribute? Why not Transitive?
- But can't we already achieve this using Tunnel encap attributes?
- Clarify text to be more Normative?
- Split use cases into separate documents?
- Remove Label descriptor role out of base doc, to confine scope?
- Is allowing carrying Labels for today unlabeled families safe?
- From my understanding, MNH looks like a use-case of downloading forwarding information to FIB from controller using BGP TCP channel to specific node. This does not looks to be appliable for generic BGP routing
- Precedence and interaction between Link-BW per MNH-leg and per route Link-BWC
- Make the TLV/SubTLV relationship more evident
- this use-case is not listed in the document (especially coming from a controller) and I think it would help
- didn't understand why this could not be a dynamic capability?

Next Steps

- ❑ Consider comments received, and rework draft text.
- ❑ Work on Implementation.
- ❑ Try out with customers who are eager to experiment, test and iterate with any available code.
- ❑ Iterate.

Thank you.

Backup Slides

Summary of Email Threads

Email threads numbered: T1 to T16
Queries numbered: T<n>.Q<n>
Comments numbered: T<n>.C<n>

Threads:

T1: https://mailarchive.ietf.org/arch/msg/idr/srOj-rxRKGqUCfBlM-tLqtHybZk/

Donatas Abraitis

T1.Q1:

Any changes to how we determine NH-family?

Ans: MNH explicitly carries the nexthop-encoding-type using EP-Type.

So AFI/SAFI based derivation or rfc5549/rfc8950 is not required.

T1.Q2:

Can MNH carry LLA addresses?

Ans: EP-type 2 (IPv6 addresss) an carry either a global address or LLA.

And the Proximity check can be used to speficy that the EP be confined to be directly connected, or whether multihop is OK.

T1.Q3:

Mention OAD?

Ans: mentioning it as per neighbor import/export configuration control as a generic way covers it.

OAD can be considered as one of those config controls.

T1.Q3:

While this is something like "overthinking", it looks like a complicated chain of TLVs

Ans: explained the building block and general format

T1.Q4:

"One or Two bytes field stating length of attribute value in bytes" - how to decide how many bytes to use?

Ans: his is based on the usual rfc4271 ‘Extended Length bit’

T2: https://mailarchive.ietf.org/arch/msg/idr/HppmMhZlCHI-i2M0MoEolvV7eeU/

Natrajan Venkataraman

T2.C1:

- I support the adoption of this draft.
- comes with numerous operational improvements
- solving the label oscillation problem in Multihomed CEs
- achieves the capability of sending load balanced nexthop sets
- also provides an unique way to carry the label, so that it can be decoupled from the RFC-8277 NRLI in the future
- also helps several Intent Driven Service Mapping use-cases as it provides a way to carry the “Transport Class ID”

T3: https://mailarchive.ietf.org/arch/msg/idr/cttkDIlM2z1kI7iTA3pJSH6lSGk/

Aravind Prabhakar

T3.C1:

- I support the adoption of the draft
- the ability to specify ordered Next hops helps in use cases where symmetric load balancing is needed.

T4: https://mailarchive.ietf.org/arch/msg/idr/tql6SQN_S98ILC7CJHLjueGZfRc/

Minto Jeyananth

T4.C1:

- support adoption of this draft

T5: https://mailarchive.ietf.org/arch/msg/idr/bL9dQJb11oDhEz2hzBRA6G-AXJw/

LINGALA, AVINASH

T5.C1:

- I support this draft.

T6: https://mailarchive.ietf.org/arch/msg/idr/k8qkS-YaIxG1PGpnIHq_K5OjCXY/

Reshma Das

T6.C1:

- I support the adoption of this draft.
- helps to carry ordered set of one or more NHs with forwarding information scoped on a per next hop basis

T7: https://mailarchive.ietf.org/arch/msg/idr/3l6hiFr84UFLpRLHNhe1IUMIFBI/

Robert Raszuk

T7.C1:

- I fully support adoption of this document.
- what is in the document I would consider just a framework for what it could contain down the road
- I also admit we were talking about this in the past a number of times, but never came with a killer use case. Likely there can be IPRs on this already
- I recall we also discussed using such an approach instead of rolling out Add-Paths at some point, but I can't recall if we dropped it due to political or technical reasons.

T8: https://mailarchive.ietf.org/arch/msg/idr/EmdUjA1IU7S3yNy8AXoEErZMPaM/

Tony Przygienda

T8.C1:

- support as well.
- Robert puts it well, it's an abstract way to group PEs/nexthops basically and that helps with lots things.
- There were proposals like doing recursive resolution over hierarchies of nexthops that are on their own RR hierarchy and so on but this is far more practical and seems to be finding new relevant usecases along the way

T9: https://mailarchive.ietf.org/arch/msg/idr/AgZTd0zMdH8w-FNK8tHVtF8iAH8/

Nat Kao

T9.C1:

- I support the adoption of this draft.
- Also thanks for this useful draft.
- This draft provides an efficient and concise framework to signal a list of ordered next-hops with corresponding instructions.
- operationally useful approach provides new functionalities currently unavailable with ADD-PATH.
- It's great to see some OpenConfig-AFT-like approaches progressing in IETF.
- Suggestions and clarifying questions:

T9.Q1:

+ clarifications and discussions on Label Descriptor
  - indicated Label-descriptor is being removed from this document, will be added in a separate document if interest persists.

T9.Q2:

+ clarifying interaction of labels/label indices carried in a Type-3 FA-TLV interact with the label(s) in a Labeled-NLRI?
  - labels in NLRI get pushed first followed by the ones in FI-TLVs.

T9.Q3:

+ clarify interaction of MNH with per-destination steering procedures defined in RFC9256?

KV> the more specific/granular FI-TLV scoped TC/Color will take precedence over the per-route scoped Color-EC.

- this text will update the precedence order given in bgp-ct document

T9.Q4:

+ Does each FI-TLV resolve to the corresponding SR-Policy identified by <Endpoint, TC/Color>?

KV> Yes, each FI-TLV resolves over a <EP, TC/Color> path provided by a transport protocol like SRTE/RSVP-TE.

T9.Q5:

+ Since we can also push labels/SRv6 SIDs in MNH via FI-TLVs, how do these TLVs interact with the matching SR Policy?

KV> Labels resolved by FA Type 1,2 TLVs (EP, TC-ID) will be outer labels, and FA Type 1(Encap) will be inner labels.

T9.Q6:

+ Is it better to use "Weight" instead of "Balance Percentage"?

- OK.

T9.Q7:

+ Would it be better to use the term "Relative Metric" or "Relative Discriminator" instead of "Relative Preference"?
It aligns better with terms like LOCAL_PREF, IGP Metric, or MED.

KV> - may be settle on "Relative ECMP Ordering"?
- have two fields? "ECMP level", "Order within the ECMP level"
- One thing to clarify: we don’t intend to carry things like Local-Pref or MED inside the MNH.
Only things that qualify/relate to a NH (like AIGP, Bandwidth, etc) are intended to be carried in the MNH.

T9.Q8:

+ I would suggest that we explicitly declare the limitations on the combinations of TLVs/Sub-TLVs.
(ex: some Sub-TLVs cannot be attached to some types of its parent TLVs; some Sub-TLVs might interfere/conflict with each other)

KV> Sure. We need to work on that.

T10: https://mailarchive.ietf.org/arch/msg/idr/w2LMd-HmDRmx-4zQm4QPlDRpCxU/

Adrian Farrel

T10.C1

- Useful function that it is worth investigating how to provide.

T10.Q1

- But can't we already achieve this using Tunnel encap attributes?
  Do we need to introduce another mechanism into our armoury?

KV> https://mailarchive.ietf.org/arch/msg/idr/EEPFjxGAFBxuQml6xt51dJH9V-k/

T11: https://mailarchive.ietf.org/arch/msg/idr/pj1N2q71xaU0T65OY5YnwWs8twk/

Igor Malyushkin

T11.Q1

- Concern on carrying label in Internet service families, which are unlabeled today.

KV> This allows not redistributing between AFs. for 6PE like usecases.

T11.C1

- Thanks for your response! In general, I agree with the list above and personally find your work promising.
  I just want to clarify several things.

/* Please check email thread above for complete discussed, only brief summary is provided here */

T12: https://mailarchive.ietf.org/arch/msg/idr/NzPwO1MYAmUgDMDY7uS55AArstg/

Gyan Mishra

T12.C1:

- I support WG adoption of MNH draft.
- A possible use case for MNH could be: .. ORR like usecase
- The MNH solution could provide a nice alternative to reflect all the paths without having to rely on add paths RFC 7912 and the propagation of paths maybe be in a more controlled fashion.
- There maybe as well load balancing use cases ECMP or UCMP that could take advantage of MNH.

T13: https://mailarchive.ietf.org/arch/msg/idr/nDtoqtaJQX_tkoJJR7hxd2S6hJM/

Swadesh Agrawal

T13.Q1

- Now when Path is advertised with multiple nexthops then for each next hop these attributes need to be signaled.

KV> No. MNH only intends to carry properties that relate to the NH. So AIGP, Link-Bandwidth yes.
But not Local-Pref or MED, they will remain a per-path attribute. IOW, there is no path-selection between different nexthop-legs in a MNH.

T13.Q2

- This is drastic change from current BGP routing model where individual paths eligible for ADDPATH are advertised separately. It has implications on years of BGP implementation specially update generation, attribute formatting etc

KV> yes, there will be changes required to implement the new functionality, like for any new attribute.

T13.Q3

- This also means any changes to attribute of one next hop, all next hops along with its associated attributes need to be readvertised instead of just affected next hop.

KV> Conversely if NH properties on multiple routes are changing, one route-update with a MNH is needed to convey the effective change.

with huge scale, that kind of property can be benefitial. Instead of sending route-updates for all the routes that changed.

T13.Q4:

- How will path resolvability work when it carries multiple next hops. What if one next hop goes down?

KV> The nexthop-legs that are unreachable are not programmed to FIB. They are not part of the ECMP nexthop.

T13.Q5:

- Do we run best path among the multiple nexthops contained within MNH of a path as well across paths from multiple source?

KV> No. we don't run any pathselection between NH-legs in a MNH. Just that the MNH can contribute to some path-selection steps

T13.Q6:

- How does receiver chooses ECMP/backup. All these work on BGP path currently. Now with single path having multiple NHs, is receiver not expected to perform best path/ECMP?

KV> A single route expresses ECMP Active/Backup using the Relative Pref field on Forwarding-Instructions.

No best-path computation happens between different legs of a MNH.

Explained with Illustration, how A/A and A/B work

T13.Q7:

- From my understanding, MNH looks like a use-case of downloading forwarding information to FIB from controller using BGP TCP channel to specific node. This does not looks to be appliable for generic BGP routing

KV> Yes MNH is a way to specify forwarding information in a more expressive manner to an ingress BGP device.

The originator of the MNH can be a controller or a BGP speaker who has central view of the forwarding-information being conveyed.

This draft just specifies the encoding on how the forwarding-information can be expressed on a single BGP route.

There can be more than one ways of using it, including those involving a controller.

T13.Q8:

- If requirement is enhanced forwarding information then Tunnel Encaps should be considered for a given nexthop.

KV> As discussed in the following thread https://mailarchive.ietf.org/arch/msg/idr/EEPFjxGAFBxuQml6xt51dJH9V-k/

I think TEA does not fit this purpose.

T14: https://mailarchive.ietf.org/arch/msg/idr/AUUqnvBRRwgufdvjIX11rHgDhgs/

Satya Mohanty

T14.Q1:

- can't the relationships between add-paths be indicated by something like LP rather than embedding the information in the MNH (Relative Pref in the MNH Forward Information TLV).

KV> The difference is in the RIB scaling, sending 1 path with N nexthops,

vs sending N paths and running path-selection on all receiving nodes.

if the decision is already made on sender, communicating the result using the MNH on one route seems enough. It alleviates pressure on the RIB scale.

T14.Q2:

- Add-path which is already deployed solution for about 12-15 years now. My understanding of the MNH is that it “should NOT” be a rip & replace strategy for Add-Path, but rather use it where it makes sense.

KV> That’s right. It is not a rip and replace. Actually MNH and Addpath may also work together as indicated in sec 4.3.

T14.Q3:

- Why use MNH in addition to add-paths between the same BGP peering? I think it may be better to make them mutually exclusive;

KV> I don’t think we should disallow it. if the individual routes have MNH, they may be advertised in addpath.

T14.Q4:

- Section 4.5.1, I see a fundamental issue. We are considering highest cost among the next-hop-legs. So, let’s say if one next-hop (not the primary one matching with the PNH) is inaccessible at the receiver, can we assume that this (means the best-path) is unresolvable for RIB resolution purposes then and all the information needs to be discarded by the receiver? I didn’t see a mention of whether we can discard this inaccessible non-best forwarding next-hop.

Receiver: when a nexthop in MNH is unresolvable, it is ignored when installing to FIB. It does not contribute to the IGP-cost calculation. I will clarify this in the draft.

T14.Q5:

- What is the rationale of the highest cost amongst the next-hop-legs representing the MNH ?

KV> Basically the (A)IGP-cost of the route is the cost of the weakest NH-leg.
For different properties, this calcalation will vary. E.g. for Bandwidth, the effective Bandwidth
of the route will be aggregation of bandwidths specified on the active (equal lowest Relative-Pref) NH-legs.

T14.Q6:

- You also mention about AIGP. As I understand AIGP will need to be put inside this MNH TLV.
There is now no placeholder for that in the draft and so this part is incomplete for this adoption
call purposes. So, assuming this is addressed, what happens in the situation I pointed out above?
Again, If the next-hop in a Forward Information TLV is not resolvable, why not just drop it?

KV> Yes, it needs to be added. This adoption call is for the WG to start working together and shape it.
AIGP will be similar to IGP-cost wrt resolvability. If the NH-leg is not reachable, it will be ignored.
And effective AIGP-cost will be worst of the lot.

T14.Q7:

- Today, without add-path, any non-best path change will not result in an update generation to the peer.
But now, with this MNH, and AIGP, a change in the metric of a MNH Forwarding Info which is not best,
will result in an update?

KV> with AIGP, a change in igp-metric will change the effective-AIGP and that will cause in an update.
And yes that includes inactive-paths also. when MNH carries those nexthop. Basically, if a NH property that
is carried on MNH changes on inactive-path, that may result in a MNH-update.

T14.Q8:

- As regards 5.1.2, we already are proposing the secondary label and there are cases in EVPN for carrying the redirect label. For the scenario you described, I feel MNH is a “heavy machinery” for solving this.

KV> I’d say it is a generic machinery, that’s useful in multiple usecases, including the secondary-label one. There are some extra effort that come with any generic machinery. And, the extra things encoded (e.g., the EP or Color) in Backup/RepairPath may be useful when you want to signal a backup-path with distinct EP also. which can provide further path-diversity and separation for repair traffic. gave e.g.

T14.Q9:

- As regards 5.4.2.3, today, the concept of load balancing is at the per-path level. Admittedly, for many cases, this will indirectly translate to the Nexthop, but I can think of situations where two routes with the same MNH can be given different link bandwidths. This causes a conflict with the link bandwidths extended community use-case. I feel link bandwidth, despite its name, should be treated as a path property because operators are using it in deployed networks for load balancing purposes and the link bandwidth today can be configured per-path via policy in many existing vendor implementations.

KV> The bandwidth per NH-leg in MNH takes precedence over the per route link-bandwidth community. As it is more specific scope (Each leg in MNH vs the PNH). I see the per-route link-bw to apply to the per-route PNH (attcode 3, 14).

And the bandwidth inside MNH legs to refer to the respective legs. To keep it simple, I am thinking not to have the per-route link-bw to apply to MNH legs, when the per-NH bw don’t exist.

T14.Q10:

- I don’t see a need for a capability Negotiation (sec 4.1). As per the draft, this needs to be done for every BGP hop by hop even those that do not reset the next-hop. I think this will be a burden if operators would want to deploy this feature at a future time and so would upgrades. At this time, my own thinking is pass/drop policy configuration should be sufficient.

KV> Ok, I agree.

T14.Q11:

- I think we need to define it as an optional transitive, otherwise this goes against the lines of next-hop-dependent-capability aka entropy label draft (Sec 4.2). Both accrue to the next-hop and scope-wise should be treated similarly. If we want to restrict to the same AS, attaching a no-export community should suffice.

KV> Non-transitive looks safer. I agree needing the RR upgrades is a little bit of hassle.
But not a showstopper IMO since some of the usecases we're talking about are at the RR (using in lieu of addpath).

T14.Q12:

- Also, there must have been a good reason not to include MVH within Tunnel Encapsulation Attribute and go for its own independent structure. What would be the reasons? Others have also mentioned this, but this thought logically comes when reading this draft.

KV> Yes, there are good reasons. I'll refer to the other thread where this is discussed:
https://mailarchive.ietf.org/arch/msg/idr/EEPFjxGAFBxuQml6xt51dJH9V-k/

T15: https://mailarchive.ietf.org/arch/msg/idr/z0AYDWv8oeBLD3BCbPEHQS3v884/

Mohan Nanduri

T15.C1:

- I support adoption of this draft and I am not aware of any undisclosed IPR.

T16: https://mailarchive.ietf.org/arch/msg/idr/96RAxTWBjHpg2kCGsSAowLIw2fE/

Ketan Talaulikar

T16.Q1:

- In my opinion the draft is NOT READY for WG adoption in its current state (v11) and needs significant updates before being considered as explained below.

KV> I think at this point of WG-adoption, we are just considering if it is a problem worth working together on. 😊

This is adoption of the doc in initial form, not final form waiting for approval for publishing. Ref: RFC-7221 Section 2.2

And, as mentioned in my previous responses, I expect the draft to evolve based on these WG discussions.

KT2> following from sec 2.2 of RFC7721:

* Is the purpose of the draft sufficiently clear?
* Does the document provide an acceptable platform for continued effort by the working group?

KV2> (Reply pending) From the discussions on IDR, answers to above two are 'Yes'.

T16.Q2:

- see MNH as two parts:
  a) Consolidation of ‘forwarding-property carrying path-attributes’ into the new attribute
  b) ability to encode multiple of those forwarding instructions in an attribute.
- At this point, to me, there seems to be some value of (a) by itself. The recent work on NHC Attribute and other work/discussions is the reason for my belief - however, we need to be careful with this (e.g. use only for newer attributes).
- Coming to (b), it has a significant impact on the base and core BGP implementation (across all AFI/SAFI). There needs to be a much higher bar for clarity and scoping for this work.

KV>: Agree it has significant impact, and also significant benefit avoiding redundant work also, provided we can get it right. This is an attempt to do just that. I am OK with setting a higher bar for clarity and collaborating towards that.

KT2> Sounds good. Would be great if you could bring out the two high-level problem statements to solve. I will leave the naming to you (and if you want to use NH-set or NHv2 or something else). I think it would benefit the discussions.

T16.Q3:

- The document refers to almost everything as TLVs (very few sub-TLVs), yet the pictures indicate there are multiple levels of TLVs/sub-TLVs. Even the sections do not reflect this hierarchy. Please consider updating the text in this section for more clarity - as also the IANA section.

KV> OK, I’ve tried to name the items more descriptive, so that it is easy to read and comprehend their purpose, than conveying its nesting level in the name.
I will see how to make the TLV/SubTLV relationship more evident, while keeping the simple descriptive names of the items.

T16.Q4:

- Clarify text to be more in form of Normative procedures.

KT>This is not clearly reflected in the draft text. What I would suggest is that the document have a normative section that describes the procedures and implications/changes to base BGP update handling (RFC4271 Sec 9). This is important since it will bring out the true nature and impact of this proposal.

KV> I will clarify/add text explaining more on the operations and procedures.

KV> I will clarify more on the normative text of procedures for sender and receiver of MNH.

T16.Q5:

- whether to separate usecases into a separate document?

KT> I would suggest retaining the use cases in this document (even if in the appendix)
since it will provide helpful context and in the main text we can refer to those use-cases as an example.

KT2> Once there is a normative text and procedure, then the use-cases serve just as an example. That should be a good balance IMO.

T16.Q5:

- use of MNH when doing NH-self:

KT> 3) The document does not explain clearly how a router can decide what NHs to "group" together as a NH-Set and importantly when they cannot be grouped together. I assume that putting together a NH-Set makes sense when the router is not setting itself as the NH (e.g., an RR) or are there cases where the router can originate a NH-set even when doing NH-self?

KV> MNH can be sent with NH-self also. This is described in section 4.1 :

KT> Are you referring to NHv2 or NH-Set here?

KV> both variants. NH-set usecase e.g. A.2, A.3, A.6, A.9

T16.Q6:

- Some normative text on the origination of this attribute and forming of a NH-set is required.
  Note that there are some use cases listed where there is actually a single NH but carrying more
  than one label (i.e. forwarding info) for it - this is not really a NH-set.

KV> In such cases, a separate Forwarding Instruction TLV is used, which has both Endpoint-address
and encap Label. Yea, the EP-address in two F.I.TLVs can be the same, with distinct labels.

KT> I only meant to say that this was NHv2 and not NH-Set.

KV> In NHv2 context, we can carry only one label/stack in the Encap portion only (referring tree diagram I have above).

T16.Q7:

- 4) Similarly, what happens when a router receives two paths each with NH-set and
  perhaps additional paths with a single NH? Can that router perform any aggregation
  into a single NH-Set? What is the implication of other attributes...

KV> The router can perform ECMP computation across the nh-legs in MNH and NH of
the contributing routes, based on path-selection. And that computed ECMP nexthop
may be conveyed in MNH to the receiving speakers. And yea, the attributes on the
active route will be conveyed. I think it will not cause any loops, BGP free core,
receiver will use tunnels to reach the reachable nh-legs in MNH. Pls let me know
if you can think of cases where it can cause loops.

T16.Q8:

- Have you considered BGP sessions over interfaces (EBGP or BGP) and the
  possibility that there are some parts of the network where tunneling is not used?

KV> yes, A.6 is BGP sessions over interfaces. And no, the document considers no-tunneling (not bgp-free core) networks out of scope.

T16.Q9:

- 5) The correlation between add-paths and NH-set needs to be specified. One aspect is (4) above.
  Another important aspect is clear guidelines on when to use add-paths and when to use NH-set.
  Yet another aspect is that the router needs to know if the neighbor supports/understands MNH so it can send either a single path with MNH or add-paths to it.

KV> sec 4.3 specifies one interaction. Sure, will add more text clarifying it.
KV> as specified in sec 4.3, addpath can be used with MNH as-well. MNH is expected to be used in lieu of Addpath, where RIB-scaling is a problem. This is described in A.1

T16.Q10:

- So, what are the other problems that NH-set is going to solve other than RIB-scaling? Note: I am not talking about the NHv2 use-cases here.

KV> some usecases as replied above. In general, if there is an off-box decision that has already computed
the NH-set, it just needs to be conveyed to the receivers rather than conveying all paths (consuming bandwidth)
and computing ecmp/ucmp again at all receivers (consuming networkwide CPU).
also, in some cases a consolidated view from multiple egress nodes can be consumed at a ‘single place’,
and the final result conveyed to multiple ingress nodes. this ‘single place’ can be a controller, RR,
or the egress-PE itself (where it is consolidating/exposing its view towards CEs/InternetPeers).

KT2> Yes, this use-case is not listed in the document (especially coming from a controller) and I think it would help.
The challenges in working with NH-set are in handling transient conditions when performing aggregation. Then also,
if the NH-set is unpacked and then repacked during the propagation. This needs further clarification and discussion.

T16.Q11:

- KT> Ah but it is not that simple! As a sender, I need to know whether to do add-paths or NH-Set. Right?

KV> addpath uses capability as today, no change. Sending MNH is config controlled. And it is not a strict either/or, as explained above.

KT2> The draft has use-cases where one can do MNH instead of add-paths. For those use-case, I think having a capability will help make things easier.

T16.Q12:

KT> So, only in BGP free core within a single provider/administrative domain but not over the Internet. Right?

KV> It is not prohibited per-se. But it is config controlled, so not sent by default. The non-transitive nature,
and the following text in sec 4.1:

KT2> This is all good, but is orthogonal to the point that I was trying to make. A capability would help decide
(in an automated manner) when to use MNH or add-paths in specific use-cases - this will ease transition/introduction
rather than the operator having to configure/manage this via something like a per-peer config. Again, just a suggestion ...
also, didn't understand why this could not be a dynamic capability?