MP-BGP Extension and the Procedures for IPv4/IPv6 Mapping Advertisement

draft-xie-idr-mpbgp-extension-4map6

Chongfeng Xie	China Telecom
Guozhen Dong	
Xing Li	CERNET Center/Tsinghua University
Congxiao Bao	
Guoliang Han	Indirection Network Inc.
Davey Song	Alibaba Cloud
Zhongfeng Guo	

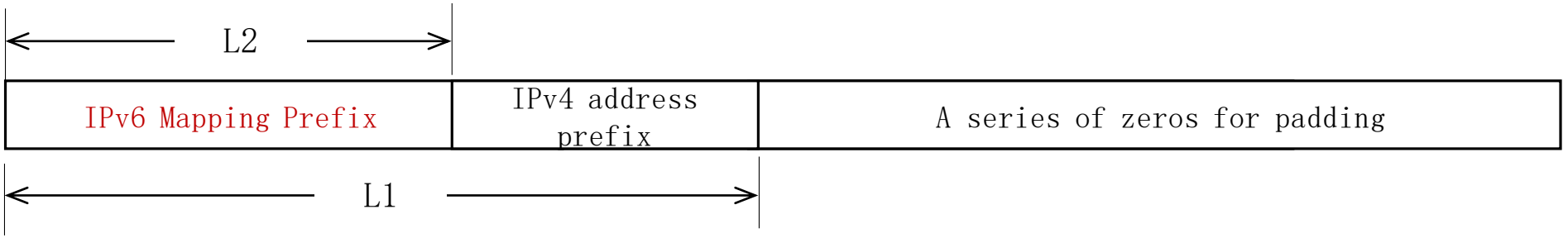
idr@IETF 117, July 2023

Overview

- [draft-ietf-v6ops-framework-md-ipv6only-underlay] proposes a framework, in which IPv4 packets will be stateless translated or encapsulated into IPv6 ones for transmission across multi-domain IPv6-only underlay.
- This document defines MP-BGP extension and the procedures for IPv4 service delivery in multi-domain IPv6-only underlay networks.
- It was proposed in Jan. 2023, firstly presented in IETF 116, current version is -03.
- Existing approaches, e.g. RFC5565(Softwire Mesh Framework) and RFC6992 (Routing for IPv4-Embedded IPv6 Packets), are mainly aimed at a single domain.

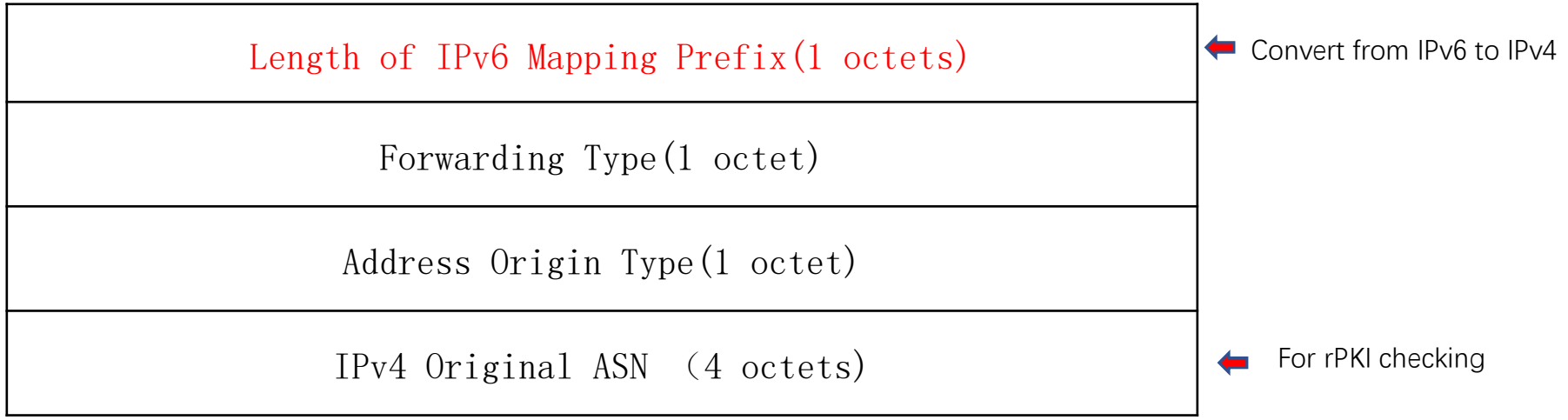
MP_REACH_NLRI

- The existing AFI/SAFI combination is used to identify the reachability of IPv4 address block in IPv6-only network.
  - AFI = 2 (IPv6)
  - SAFI = 1 (Unicast)
  - Length of Next Hop
  - Network Address of Next Hop
  - NLRI : Composite IPv6 address prefix, which is composed of a **IPv6 mapping prefix, the original IPv4 address prefix**, and a series of padding bits



4map6 BGP path attribute

- It specifies the IPv6 mapping prefix and other additional information needed to properly transform the IPv4 packets.
- It is optional and transitive.

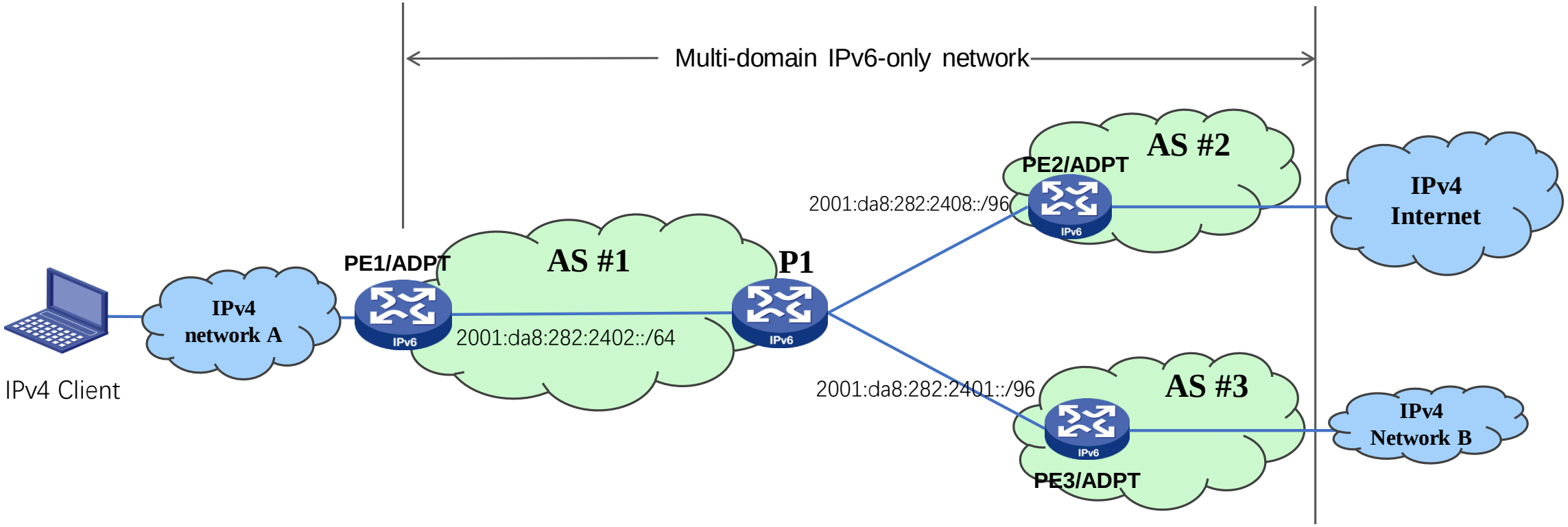


Note: ATTR_ SET attribute of RFC6368 can also be used to transfer the BGP routing information of the IPv4 in IPv6-only networks, including IPv4 ASN,

Revisions made since IETF 116

- Based on the comments of Jeff Hass, ATTR_ SET attribute of RFC6368 can be used to transfer the BGP routing information of the IPv4 in IPv6-only networks, this has been added in section 3.2.
- The use case of IPv6-only DC for AI-fabric is added in the appendix.
- Several editorial changes have been made.
- Davey Song and Zhongfeng Guo of Alibaba Cloud have joined as co-authors.

System Implementation and Test



Original IPv4 Route vs New IPv6 Route

IPv4 route in PE2

```
BGP table version is 11, local router ID is 192.168.66.189, vrf id 0
Default local pref 100, local AS 64589
Status codes:  s suppressed, d damped, h history, * valid, > best, = multipath,
               i internal, r RIB-failure, S Stale, R Removed
Nexthop codes: @NNN nexthop's vrf id, < announce-nh-self
Origin codes:  i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*> 110.242.68.0/24  0.0.0.0                  0         32768 i
*> 192.168.8.0/24   0.0.0.0                  0         32768 i
*> 192.168.66.173/32
                    192.168.66.173           0             0 64520 i
   192.168.66.175/32
                    192.168.66.175           0             0 64550 i

Displayed  4 routes and 4 total paths
xlat#
```

IPv6 route received by P router

```
xlat# show bgp ipv6 2001:da8:282:2408::6ef2:4400/120
BGP routing table entry for 2001:da8:282:2408::6ef2:4400/120
Paths: (1 available, best #1, table default)
  Advertised to non peer-group peers:
  2001:da8:282:ff0c:62fd:f174:f6c3:189 2001:da8:282:ff0c:9ec5:d104:c664:171
  64589
    2001:da8:282:ff0c:62fd:f174:f6c3:189 from 2001:da8:282:ff0c:62fd:f174:f6c3:189 (192.168.66.189)
      Origin IGP, metric 0, valid, external, best (First path received)
      Last update: Thu Jul 13 02:30:19 2023
xlat#
```

110.242.68.0/24 is in IPv4 Internet
(110.242.68.3 is www.baidu.com)

2001:da8:282:2408:: 6e f2 : 44 00/120

IPv6 Mapping Prefix (/96) 110. 242. 68. 0/24

4map6 BGP Path Attribute

```
Path Attribute - Unknown (64)
  > Flags: 0xc0, Optional, Transitive, Complete
    Type Code: Unknown (64)
    Length: 3
    Unknown Path attributes
```

```
aa 4f 44 c9 84 a0 62 fd  f1 74 f6 c3 86 dd 6c 01   ·OD···b· ·t····l·
b6 0d 00 a1 06 01 20 01  0d a8 02 82 ff 0c 62 fd   ····· · ·····b·
f1 74 f6 c3 01 89 20 01  0d a8 02 82 ff 0c aa 4f   ·t···· · ·····O
44 c9 84 a0 01 70 c2 22  00 b3 af f1 fc ea 42 a7   D····p·" ·····B·
57 db 80 18 01 f5 63 be  00 00 01 01 08 0a 15 8c   W·····c· ·····
37 ac 96 55 5b 63 ff ff  ff ff ff ff ff ff ff ff   7··U[c·· ·····
ff ff ff ff ff ff 00 57  02 00 00 00 4d 90 0e 00   ·····W ····M···
26 00 02 01 10 20 01 0d  a8 02 82 ff 0c 62 fd f1   &···· ·· ·····b··
74 f6 c3 01 89 00 80 20  01 0d a8 02 82 24 08 00   t·····  ·····$··
00 00 00 c0 a8 42 ad 40  01 01 00 50 02 00 0a 02   ·····B·@ ···P····
02 00 00 fc 4d 00 00 fc  08 c0 40 03 60 02 00      ····M··· ··@·`··
```

Value	Description
0xc0	Path Attribute Flags (optional transitive)
0x40	Path Attribute Type Code (only for trial)
0x03	DATA length (3 bytes)
0x60	IPv6 Mapping Prefix Length (96)
0x02	Forwarding Type (translation)
0x00	Address Origin Type (local)

IPv6-IPv4 auto-config in PE1

PE1 received IPv6 route 2001:da8:282:2408::6ef2:4400/120 (attribute plen=96)

```
xlat# show bgp ipv6 2001:da8:282:2408::6ef2:4400/120
BGP routing table entry for 2001:da8:282:2408::6ef2:4400/120
Paths: (1 available, best #1, table default)
  Advertised to non peer-group peers:
  2001:da8:282:ff0c:62fd:f174:f6c3:189 2001:da8:282:ff0c:9ec5:d104:c664:171
  64589
    2001:da8:282:ff0c:62fd:f174:f6c3:189 from 2001:da8:282:ff0c:62fd:f174:f6c3:189 (192.168.66.189)
      Origin IGP, metric 0, valid, external, best (First path received)
      Last update: Thu Jul 13 02:30:19 2023
xlat#
```

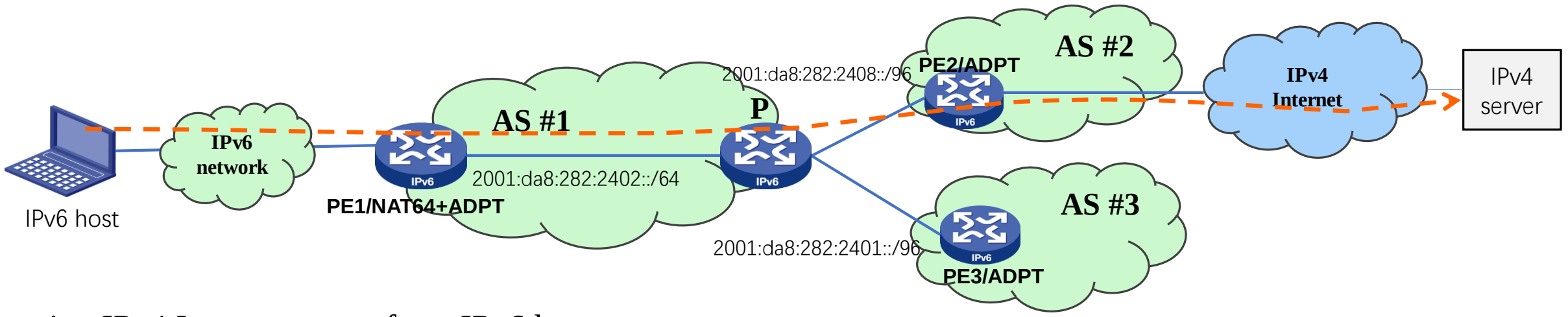
PE1 map this route to an IPv4-IPv6 translation rule:

111.242.68.0/24 ---> rule3 (source-IPv6-prefix: 2001:da8:282:2402::/64; dest-IPv6-prefix: 2001:da8:282:2408::/96)

```
local IPv6 prefix:2001:da8:282:2402::/64
local map6id:fec0::1/64
local mapid:172.2.0.1/24
local nexthop:2001:da8:282:ff0c:aa4f:44c9:84a0:170
local prefix6 address:64:ff9b::/96
local dnat ns name:dnat66
local dnat ipset name:iviDns64Ipset
index:1        spl:2001:da8:282:2402::/64      dp:2001:da8:282:2401::/96       map6id:fec0::2  mapid:172.2.0.2 peerid:192.168.66.171 pref:73
index:2        spl:2001:da8:282:2402::/64      dp:2001:da8:282:2408::/96       map6id:fec0::3  mapid:172.2.0.3 peerid:192.168.66.189 pref:3
```

IPv6 host→IPv4 server visit

Support IPv6→IPv6, IPv4→IPv4, IPv6→IPv4, etc.



ping IPv4 Internet server from IPv6 hosts

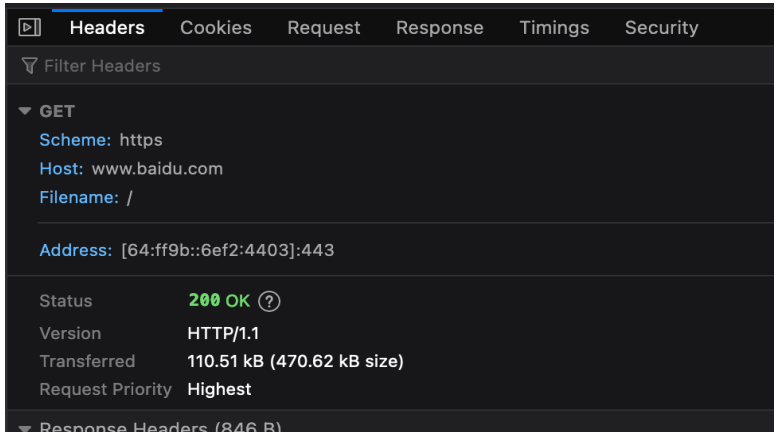
windows IPv6 host

```
C:\Users\ivi>ping www.baidu.com

正在 Ping www.a.shifen.com [64:ff9b::6ef2:4403] 具有 32 字节的数据:
来自 64:ff9b::6ef2:4403 的回复: 时间=26ms
来自 64:ff9b::6ef2:4403 的回复: 时间=24ms

64:ff9b::6ef2:4403 的 Ping 统计信息:
    数据包: 已发送 = 2, 已接收 = 2, 丢失 = 0 (0% 丢失),
往返行程的估计时间(以毫秒为单位):
    最短 = 24ms, 最长 = 26ms, 平均 = 25ms
Control-C
^C
```

MacOS IPv6 host



Android IPv6 host

```
$ ping6 www.baidu.com
PING www.baidu.com(64:ff9b::6ef2:4404) 56 data bytes
64 bytes from 64:ff9b::6ef2:4404: icmp_seq=1 ttl=42 time
=18.7 ms
64 bytes from 64:ff9b::6ef2:4404: icmp_seq=2 ttl=42 time
=43.7 ms
64 bytes from 64:ff9b::6ef2:4404: icmp_seq=3 ttl=42 time
=19.2 ms
64 bytes from 64:ff9b::6ef2:4404: icmp_seq=4 ttl=42 time
=30.1 ms
64 bytes from 64:ff9b::6ef2:4404: icmp_seq=5 ttl=42 time
=17.2 ms
64 bytes from 64:ff9b::6ef2:4404: icmp_seq=6 ttl=42 time
=20.5 ms
64 bytes from 64:ff9b::6ef2:4404: icmp_seq=7 ttl=42 time
=38.7 ms
64 bytes from 64:ff9b::6ef2:4404: icmp_seq=8 ttl=42 time
=29.5 ms
64 bytes from 64:ff9b::6ef2:4404: icmp_seq=9 ttl=42 time
=36.1 ms
^C
--- www.baidu.com ping statistics ---
9 packets transmitted, 9 received, 0% packet loss, time
8021ms
rtt min/avg/max/mdev = 17.247/28.249/43.798/9.263 ms
$
```

(termux)

Next Steps

- Comments and suggestions are welcome, and make further refinement to improve the document
- Authors would like to ask for WG adoption of this document

Thank you !
Q&A