LEGAL DISCLAIMER:

THIS DOCUMENT AND THE INFORMATION CONTAINED HEREIN IS PROVIDED ON AN “**AS IS**” BASIS. TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, NVM EXPRESS, INC. (ALONG WITH THE CONTRIBUTORS TO THIS DOCUMENT) HEREBY DISCLAIM ALL REPRESENTATIONS, WARRANTIES AND/OR COVENANTS, EITHER EXPRESS OR IMPLIED, STATUTORY OR AT COMMON LAW, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE, VALIDITY, AND/OR NONINFRINGEMENT.

NVM Express Workgroup
c/o VTM, Inc.
3855 SW 153rd Drive
Beaverton, OR 97003
USA
info@nvmexpress.org

NVM Express Technical Proposal for New Feature

Technical Proposal ID	**8009 Automated Discovery of NVMe-oF Discovery Controllers for IP Networks**
Change Date	**2022-01-11**
Builds on Specification	**NVM Express™ Base Specification Revision 2.0a**
References Specification	**TP-8010 - NVMe-oF Centralized Discovery Controller** **TP-8013 - Unique Discovery Controller ID** **IETF RFC 1034** **IETF RFC 1035** **IETF RFC 2136** **IETF RFC 8766** **IETF RFC 6762** **IETF RFC 6763**

Technical Proposal Author(s)

Name	**Company**
Erik Smith	Dell EMC

This proposal intends to do the following:

- Describe how to interpret Domain Name System Service Definition (DNS-SD) records to determine the IP Address of one or more NVMe-oF Discovery Service instances.
- Describe when and how to use the multicast Domain Name System (mDNS) protocol to automatically retrieve the DNS-SD records that are available on the local Broadcast Domain.
- Describe when and how to use unicast DNS requests to retrieve all DNS-SD records known to the DNS.

Revision History

Revision Date	**Change Description**
2020-05-21	Initial version
2020-06-12	Incorporated HPE’s comments
2020-06-23	Remove TP 8010 references.
2020-10-06	Updated to leverage Subtypes are defined in RFC6763
2020-10-7	Updated numbering of section 5.6 and added references.
2020-11-03	Updated references and incorporated feedback from FMDS meeting on 10/27
2020-11-10	Added DNS information and incorporated feedback from FMDS meeting on 11/3
2020-11-11	Incorporated feedback from FMDS meeting on 11/10
2021-06-02	Updated format of TP for incorporation into NVMe 2.0 spec.

2021-07-01	Updated “References Specifications”
2021-08-31	Incorporated Phase 3 comments from Mike A
2021-09-15	Incorporated Phase 3 comments from review during FMDS meeting on 2021-09-14
2021-09-21	Incorporated Phase 3 comments from review during FMDS meeting on 2021-09-21
2021-10-20	Incorporated 30 day member review comments from Mike Allison (Samsung)
2021-11-02	Incorporated Phase 3 comments from review during FMDS meeting on 2021-11-02
2021-11-03	Incorporated Phase 3 comment from Cisco related to a typo.
2022-01-10	Integration
2022-01-11	Added quotes to strings that were not quoted.

Description for NVM Express Base Specification, revision 2.0a Changes Document

This technical proposal defines the changes being requested to the NVM Express Base Specification, revision 2.0a. The changes impact one section:

1. Add a new subsection 8.NEW.A that describes “Automated Discovery of NVMe-oF Discovery controllers for IP Networks”.

Markup Conventions:

Black:	Unchanged (however, hot links are removed)
~~Red Strikethrough~~:	Deleted
Blue:	New
Blue Highlighted:	TBD values, anchors, and links to be inserted in new text.
<Green Bracketed>:	Notes to editor

Description of NVMe 2.0a Base Specification Changes

Add a new section 8.NEW.A as shown below:

<Editor: The new section in this document (i.e., 8.NEW.A) and the section 8.NEWA described in TP-8010 are the same.>

8.NEW.A Automated Discovery of NVMe-oF Discovery Controllers for IP Based Fabrics

When operating in an IP based fabric, before transmitting a Fabrics Connect command, the IP address of the fabric interface of the Discovery controller is determined by one of the following methods:

a. administrative configuration;
b. discovered using DNS-SD (refer to RFC 6763); or
c. obtained by some means not defined in this specification.

DNS-SD information may be retrieved using mDNS (refer to RFC 6762) or from a DNS server (refer to RFC 1034 and RFC 1035).

When DNS-SD is used as described in this section, hosts, CDCs and DDCs may use DNS-SD to perform automated discovery of Discovery controllers in IP fabrics consisting of:

a. Two IP interfaces (i.e., a host and a NVM subsystem) physically connected to one another (refer to Figure NEW.ES1);
b. Many IP interfaces participating in one or more Broadcast Domains (refer to Figure NEW.ES2); or
c. A Centralized Discovery controller and many IP interfaces that may reside in multiple Broadcast Domains (refer to Figure NEW.ES3).

mDNS is a multicast protocol that allows discovery of IP interfaces within the same Broadcast Domain. Discovering IP interfaces outside of the Broadcast Domain using mDNS requires either the use of RFC 8766 or an mDNS NVMe-oF proxy. An mDNS NVMe-oF proxy is an mDNS responder that is responsive to queries for either the "_nvme-disc" service or the "_cdc._sub._nvme-disc" service and responds with the information defined in section 8.NEW.A.1.2.

Figure NEW.ES1: Configuration A - Two IP Interfaces

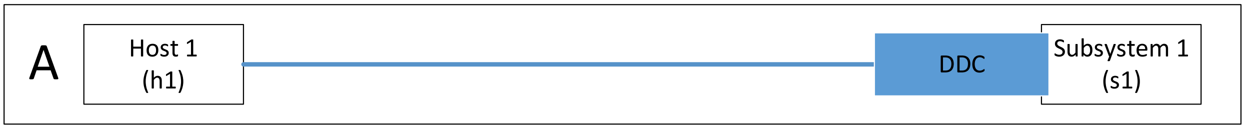


Figure NEW.ES2: Configuration B – Multiple IP Interfaces without a CDC

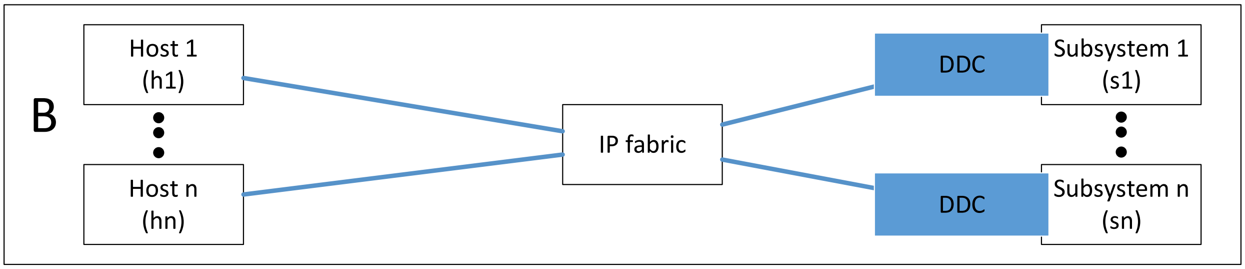
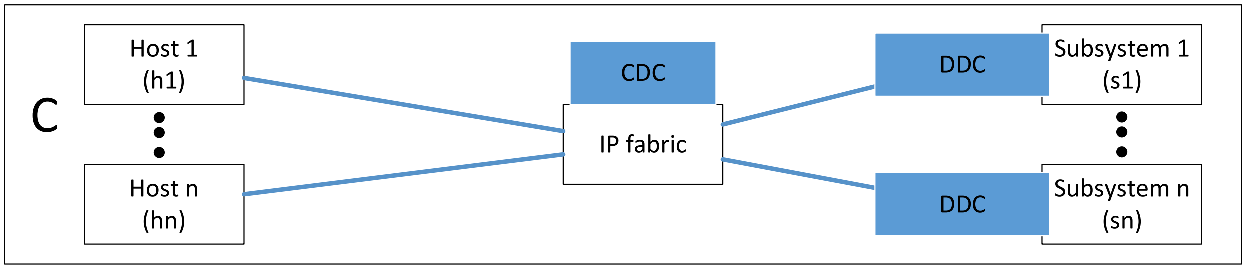


Figure NEW.ES3: Configuration C – Multiple IP interfaces with a CDC



8.NEW.A.1 Discovery of NVMe-oF Discovery Controllers

8.NEW.A.1.1 Query

To facilitate the discovery of Discovery controller IP fabric interface addresses, hosts, CDCs and DDCs may transmit an mDNS query (refer to RFC 6762) or a DNS query (refer to RFC 1034 and RFC 1035) that includes a DNS PTR record (refer to RFC 6763) with the name in the form of:

“<Service>.<Domain>”.

The <Service> portion of the name can be further broken down into:

“<service name>.<protocol>”.

For NVMe over Fabrics, the DNS PTR record included in the mDNS or DNS query shall be in the form of:

“_nvme-disc.<protocol>.<domain>”; or

“_<subtype>._sub._nvme-disc.<protocol>.<domain>”

The protocol field shall be set as shown in Figure NEW.ES4.

The subtype field shall be set as shown in Figure NEW.ES5.

The domain field shall be set as shown in Figure NEW.ES6.

Figure NEW.ES4: mDNS Protocol Field

NVMe-oF Transport	Fabric Protocol	mDNS <protocol> Field
TCP	TCP	“_tcp”
RDMA	RoCE	“_udp”
RDMA	iWARP	“_tcp”

Figure NEW.ES5: mDNS Subtype

Subtype	Usage
“_cdc”	Used by CDC and DDC instances to detect the presence of a CDC service.
“_ddcpull”	May be used by CDC instances to detect the presence of DDC instances that are requesting a pull registration.

Figure NEW.ES6: mDNS Domain

Domain	Usage
“local”	Used for mDNS.
<FQDN>	A fully qualified domain name may be used when DNS-SD information is retrieved from a DNS server.

8.NEW.A.1.2 Response

The responses that may be received for an mDNS or DNS query include the following records as described in DNS-Based Service Discovery (refer to RFC 6763):

- A DNS PTR record (refer to section 3.3.12 in RFC 1035);
- a DNS SRV record (refer to RFC 2782);
- a DNS TXT record (refer to section 3.3.14 in RFC 1035); and
- an A record and/or AAAA record providing the IPv4 and IPv6 IP addresses respectively.

8.NEW.A.1.2.1 DNS PTR record

The DNS PTR record included in the mDNS or DNS response shall be in the form of:

“<Service>.<Domain>”.

The <Service> portion of the name can be further broken down into:

“<service name>.<protocol>”.

For NVMe over Fabrics, the DNS PTR record included in the mDNS or DNS response shall be in the form of:

“_nvme-disc.<protocol>.<domain>”; or

“_<subtype>._sub._nvme-disc.<protocol>.<domain>”.

The protocol field shall be set as shown in Figure NEW.ES4.

The subtype field shall be set as shown in Figure NEW.ES5.

The domain field shall be set as shown in Figure NEW.ES6.

8.NEW.A.1.2.2 DNS SRV record

The DNS SRV record provides the TCP port where the service instance can be reached and shall have a name in the form of:

“<Instance>.<Service>.<Domain>”.

Instance (Instance name) is a vendor defined human readable string. Although use of a serial number is discouraged in DNS-Based Service Discovery (refer to RFC 6763), an instance name that may be meaningful to an IT administrator is a combination of the Model Number (MN), Serial Number (SN) and Physical Fabric Interface (Physical Ports). For example:

“SubsystemIF-SerialNumber-SubsystemModel”.

The maximum length of the Instance Name field is 63 bytes (refer to RFC 6763).

The <Service> portion of the name (see section 4.1 of RFC 6763) can be further broken down into:

“<service name>.<protocol>”.

8.NEW.A.1.2.3 DNS TXT record

The DNS TXT record provides additional information about the instance using key/value pairs in the form of “key=value” separated by commas (refer to section 6.4 in RFC 6763).

For NVMe over Fabrics, the DNS TXT record shall include a key/value pair for protocol (p) and may include a key/value pair describing an NQN.

DNS TXT record key/value pairs:

Protocol (p): the protocol field shall indicate the IP transport protocols that are supported by the Discovery controller being advertised. For example:

“p=tcp”, “p=roce”, or “p=iwarp”.

NQN: the DNS TXT record may contain an nqn key/value pair. When it is included the NQN provided shall be set to the unique NQN of the Discovery subsystem if one is available for use. Otherwise, the well-known Discovery Service NQN (nqn.2014-08.org.nvmexpress.discovery) may be used.

As described in RFC 6763, the format of the data within a DNS TXT record is one or more strings, packed together without any intervening gaps or padding bytes for word alignment.

The format of a TXT record that may be provided in a response is:

“<length byte>p=tcp<length byte>nqn=NQN.of.Discovery.subsystem”.

Using this format, an example of a TXT record that may be provided in a response is:

“05p=tcp1Enqn=NQN.of.Discovery.subsystem”.

8.NEW.A.2 Host Operation

8.NEW.A.2.1 Host Query

As described in section 8.NEW.A.1.1, a host may transmit an mDNS or DNS query to discover CDC and DDC instances that are present on the transport network. When used, the mDNS or DNS query shall include a DNS PTR record (refer to RFC 6763) with the name in the form of:

“_nvme-disc.<protocol>.<domain>”.

The protocol field shall be set as shown in FigureNEW.ES4: mDNS Protocol Field

The domain field shall be set as shown in FigureNEW.ES6: mDNS Domain

8.NEW.A.2.2 Host Processing of DNS-SD Records

Upon reception of an mDNS or DNS response that contains a DNS SRV record with the service name set to “_nvme-disc.<protocol>.local”. The host interface may use the IP address in the A or AAAA record as the destination IP address for a subsequent Fabrics Connect command and attempt to perform Discovery Information Registration (refer to section 8.NEW.2). <Note to editor: 8.NEW.2 is located in TP-8010>

8.NEW.A.3 DDC Operation

8.NEW.A.3.1 DDC mDNS Initialization

During initialization (e.g., following a link transition or power cycle), before the DDC’s mDNS responder function is enabled, the DDC shall probe to ensure the unique resource records the DDC are responsible for are unique on the local link (refer to section 8.1 in RFC 6762).

Upon successful completion of the probe, the DDC shall Announce (refer to section 8.2 in RFC 6762) its newly registered resource records. If a DDC is configured for pull registration, the service name of “_ddcpull._sub._nvme-disc._<protocol>.local” is one of these resource records.

Upon announcing its resource records, if a DDC:

a. has not been configured to perform push registration (refer to section 8.NEW.2.1) <Note to editor: 8.NEW.2.1 is located in TP-8010>, or has not been configured to request a pull registration (refer to section 8.NEW.2.2) <Note to editor: 8.NEW.2.2 is located in TP-8010> from a CDC, it may respond to mDNS queries for the service name of “_nvme-disc.<protocol>.local”;
b. has not been configured to request a pull registration (refer to section 8.NEW.2.2) <Note to editor: 8.NEW.2.2 is located in TP-8010>, it may respond to queries for the service name of “_ddcpull._sub._nvme-disc._<protocol>.local” as described in section 8.NEW.A.3.5; or
c. has been configured to perform push registration (refer to section 8.NEW.2.1) <Note to editor: 8.NEW.2.1 is located in TP-8010> with a CDC, it should not respond to mDNS queries for the service name of “_nvme-disc.<protocol>.local”, unless it has been administratively configured to do so or until it has performed a query and determined a CDC is not present as defined in section 8.NEW.A.3.3.

8.NEW.A.3.2 DDC DNS Initialization

During initialization (e.g., following a link transition or power cycle), a DDC may dynamically update DNS records (refer to RFC 2136) by providing an update that includes the Resource Records defined in section 8.NEW.A.1.2.

8.NEW.A.3.3 DDC Query

A DDC may determine if a CDC is present by transmitting a query that includes a DNS PTR record (refer to RFC 6763) with the name in the form of:

> “_cdc._sub._nvme-disc.<protocol>.<domain>”.

The protocol field shall be set as shown in Figure NEW.ES4.

The domain field shall be set as shown in Figure NEW.ES6.

8.NEW.A.3.4 DDC Processing of DNS-SD records

Upon reception of an mDNS or DNS response that contains a DNS SRV record with the service name set to “_cdc._sub._nvme-disc”, the DDC may use the IP address in the A or AAAA record as the destination IP address to either:

a. Perform push registration (refer to section 8.NEW.2.1) <Note to editor: 8.NEW.2.1 is located in TP-8010> with the CDC; or
b. Request a pull registration (refer to section 8.NEW.2.2) <Note to editor: 8.NEW.2.2 is located in TP-8010> from the CDC (e.g., using Kickstart Discovery Request PDU (KDReq) section in the NVMe Transport Specification).

If a DDC supports mDNS and has been configured to perform push registration (refer to section 8.NEW.2.1) <Note to editor: 8.NEW.2.1 is located in TP-8010>, or has been configured to request a pull registration (refer to section 8.NEW.2.2) <Note to editor: 8.NEW.2.2 is located in TP-8010> from a CDC, the DDC should cease responding to mDNS requests for the service name of “_nvme-disc.<protocol>.local” if a CDC is detected as defined in section 8.NEW.A.3.3.

8.NEW.A.3.5 DDC response to mDNS queries

A DDC may respond to mDNS queries for the service names of either:

> “_nvme-disc.<protocol>.local”; or
>
> “_ddcpull._sub._nvme-disc.<protocol>.local”.

mDNS responses to queries for either of these service names shall contain the information described in section 8.NEW.A.1.2.

DDCs should only respond to mDNS queries for the service name of “_ddcpull._sub._nvme-disc.<protocol>.local” if the DDC is requesting a pull registration (refer to section 8.NEW.2.2) <Note to editor: 8.NEW.2.2 is located in TP-8010> be performed.

8.NEW.A.4 CDC Operation

8.NEW.A.4.1 CDC mDNS Initialization

During initialization (e.g., following a link transition or power cycle), before the CDC’s mDNS responder function is enabled, the CDC shall probe to ensure the unique resource records the CDC are responsible for are unique on the local link (refer to section 8.1 in RFC 6762).

Upon successful completion of the probe, the CDC shall announce (refer to section 8.2 in RFC 6762) its newly registered resource records.

Upon announcing its resource records, the CDC's mDNS responder function may be enabled and respond to queries for the service name of "_cdc._sub._nvme-disc._<protocol>.local" as described in section 8.NEW.A.4.5.

A CDC should query for the presence of another CDC as defined in section 8.NEW.A.4.3 and process responses as defined in section 8.NEW.A.4.4.

8.NEW.A.4.2 CDC DNS Initialization

During initialization (e.g., following a link transition or power cycle), a CDC may dynamically update DNS records (refer to RFC 2136) by providing an update that includes the Resource Records defined in section 8.NEW.A.1.2.

8.NEW.A.4.3 CDC Query

A CDC may query for both CDC and DDC instances. When performed the mDNS or DNS query shall include a DNS PTR record (refer to RFC 6763) with the name in the form of:

"_cdc._sub._nvme-disc.<protocol>.<domain>"; or

"_ddcpull._sub._nvme-disc.<protocol>.<domain>".

The protocol field shall be set as shown in Figure NEW.ES4.

The domain field shall be set as shown in Figure NEW.ES6.

8.NEW.A.4.4 CDC Processing of DNS-SD records

Upon reception of an mDNS or DNS response that contains a DNS SRV record with the service name set to "_cdc._sub._nvme-disc.<protocol>.local" the CDC may provide an alert to the administrator to indicate the presence of more than one CDC in a broadcast domain.

Upon reception of an mDNS or DNS response that contains a DNS SRV record with the service name set to "_ddcpull._sub._nvme-disc.<protocol>.local" the CDC may choose to perform a pull registration (refer to section 8.NEW.2.2) <Note to editor: 8.NEW.2.2 is located in TP-8010> for the responding DDC. If the CDC performs a pull registration, it shall use the IP address in the A or AAAA record as the destination IP address for a subsequent connect command.

8.NEW.A.4.5 CDC response to mDNS queries

A CDC may respond to mDNS queries for the service names of either:

"_nvme-disc.<protocol>.local"; or

"_cdc_sub._nvme-disc.<protocol>.local".

mDNS responses to this query shall contain the information described in section 8.NEW.A.1.2.