INTERNET-DRAFT
Expires: July 24, 2026

Alexander Montrose

January 25, 2026

BGP Optional Transitive Attribute for Advertising GPU and AI Accelerator Capabilities

draft-montrose-idr-gpu-capability-00

Abstract

This document defines a new BGP path attribute, GPU_CAPABILITY, to allow network devices to advertise the availability, capacity, and characteristics of GPU and AI accelerators within a data center or AI fabric. This optional, transitive attribute enables schedulers, orchestration systems, and control-plane applications to discover GPU resources directly through BGP, integrating resource-awareness into routing and placement decisions. The attribute is TLV-based, extensible, and vendor-neutral.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

Copyright Notice

Copyright (c) 2026 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (https://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Table of Contents

1. Introduction

Modern data centers and AI fabrics deploy large numbers of GPU and AI accelerators (e.g., NVIDIA, AMD, Qualcomm) to support high-performance

workloads. Existing BGP mechanisms allow for network reachability and overlay distribution but do not provide a standardized mechanism to advertise resource characteristics such as GPU type, available memory, number of free accelerators, NVLink locality, or congestion awareness.

The GPU_CAPABILITY attribute provides an optional, transitive mechanism to expose GPU and AI accelerator metadata via BGP. It is TLV-based, allowing extensibility to future accelerator types or metrics.

2. Terminology

GPU: Graphics Processing Unit or AI accelerator
TLV: Type-Length-Value
DCQCN: Data Center Quantized Congestion Notification
EVPN: Ethernet VPN
VXLAN: Virtual Extensible LAN
SLURM: Simple Linux Utility for Resource Management (widely used resource manager for HPC clusters)
Slinky-based frameworks: Orchestration frameworks designed for AI/GPU fabrics
Scheduler: Orchestration system that places jobs on GPU resources

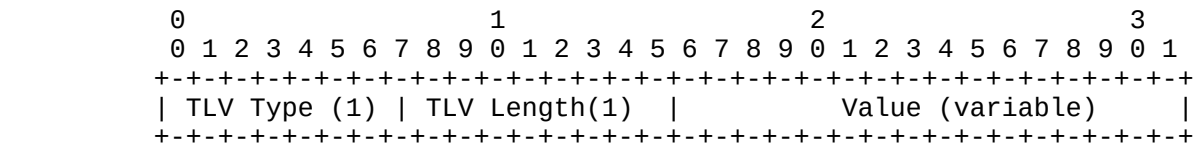
3. GPU_CAPABILITY Attribute

3.1. Attribute Flags and Type

Attribute Type Code: TBD (to be assigned by IANA)
Flags: Optional (O), Transitive (T)
Format: Variable length, TLV-based

3.2. TLV Encoding

Each TLV in GPU_CAPABILITY consists of:



TLV Type: Identifies the metric or property
TLV Length: Length in bytes of the Value field
Value: Metric value or vendor-specific encoding

4. TLV Types

Type	Name	Description
0x01	Vendor ID	1=NVIDIA,2=AMD,3=Qualcomm,4=Intel
0x02	Model ID	Accelerator / NIC model identifier
0x03	GPUs Free	Free GPUs / accelerators
0x04	GPUs Total	Total GPUs / accelerators
0x05	Memory Free	Free memory (GB)
0x06	Memory Total	Total memory (GB)
0x07	NVLink Domain	NVLink / NVSwitch domain ID
0x08	Island ID	GPU island / pod / cell ID
0x09	Rack ID	Optional rack identifier
0x0A	Node ID	Node identifier
0x0B	GPU Health	0=bad,1=ok,2=excellent
0x0C	Power Headroom	Percent (0-100)
0x0D	Thermal Headroom	Percent (0-100)
0x0E	Job Affinity Group	Scheduler / placement grouping
0x0F	Vendor-Specific TLV	Vendor-defined extensions
0x10	RoCEv2 Enabled	1=enabled,0=disabled
0x11	RoCEv2 Link Speed	NIC port speed (Gbps)

0x12	RoCEv2 Port State	0=down,1=up
0x13	RoCEv2 MTU	Ethernet MTU (bytes)
0x14	RoCEv2 VLAN ID	VLAN carrying RoCEv2 traffic
0x15	RoCEv2 DSCP	DSCP value for RoCEv2 packets
0x16	RoCEv2 IP Version	4=IPv4,6=IPv6,46=dual-stack
0x17	RoCEv2 IPv4 Address	Source IPv4 address
0x18	RoCEv2 IPv6 Address	Source IPv6 address
0x19	RoCEv2 GID	IPv6-based Global Identifier
0x1A	RoCEv2 UDP Destination Port	UDP port (default 4791)
0x1B	DCQCN Enabled	1=enabled,0=disabled
0x1C	ECN Profile	0=none,1=aggressive,2=moderate
0x1D	PFC Enabled	1=enabled,0=disabled
0x1E	PFC Priority Mask	Bitmap of lossless priorities
0x1F	ETS Profile	Traffic class -> bandwidth mapping
0x20	RoCEv2 Congestion Control	DCQCN, HPCC, TIMELY
0x21	Ultra Ethernet Link Speed	UE port speed (Gbps)
0x22	Ultra Ethernet Port State	0=down,1=up
0x23	Ultra Ethernet Congestion	ECN/CC mode
0x24	Ultra Ethernet Fabric ID	UE fabric / domain identifier
0x25	InfiniBand Port LID	Local Identifier
0x26	InfiniBand Link Speed	NDR/XDR speed (Gbps)
0x27	InfiniBand MTU	MTU size (bytes)
0x28	InfiniBand Port State	1=Active,0=Inactive
0x29	InfiniBand SL / VL Profile	Service / Virtual lane mapping
0x2A	InfiniBand Subnet Prefix	IB subnet identifier
0x2B	Optical Interface Type	DR4, FR4, ZR, ZR+
0x2C	Optical Wavelength	Lambda / DWDM channel
0x2D	Optical Modulation	NRZ, PAM4, QPSK, 16QAM
0x2E	Optical Reach	Max reach (km)
0x2F	Optical FEC Mode	KP4, RS(544,514), SD-FEC
0x30	Optical Line Rate	100G/200G/400G/800G
0x31	Optical Power Budget	Tx/Rx budget (dB)
0x32	MPLS Enabled	1=enabled,0=disabled
0x33	MPLS Vendor ID (PEN)	IANA Private Enterprise Number
0x34	MPLS Vendor Capability TLV	Vendor-defined MPLS extensions
0x35	MPLS Label Range	Allocated / private label space
0x36	MPLS Traffic Engineering	RSVP-TE / SR-TE
0x37	MPLS Protection Mode	FRR, TI-LFA, None
0x38	MPLS OAM Capability	BFD, LSP Ping
0x39	MPLS QoS Mapping	EXP -> DSCP/TC mapping
0x40	NCCL Path Type	NVLink / NVLS / IB / RoCE / UE
0x41	NCCL Path Bandwidth	Effective bandwidth for graph edge scoring
0x42	NCCL Path Latency	One-way latency estimate
0x43	NCCL Rail ID	Multi-rail / dual-rail identifier
0x44	GPUDirect RDMA Capable	1=enabled,0=disabled
0x45	SHARP Capable	In-network reduction support
0x46	CollNet Capable	Hierarchical collective support
0x47	NVLS Capable	NVLink Switch collectives
0x48	NCCL Preferred Transport	Hard / soft transport bias

```
| 0x49 | NCCL Failure Domain          | GPU / node / rack / pod
 |
| 0x4A | NCCL Max Channels            | Parallel channel hint
  |
| 0x4B | NCCL Topology Symmetry Group | Identical nodes for ring cloning
 |
| 0x4C | NCCL Cross-Island Penalty    | Cost factor for island crossing
 |
| 0x4D | NCCL Inter-Rack Penalty      | Cost factor for rack crossing
 |
| 0x4E | NCCL Inter-DC Penalty        | Strongly discouraged paths
 |
| 0x4F | NCCL Algorithm Mask          | Ring / Tree / CollNet enable mask
 |
+------+------------------------------+-------------------------------------------+
```

4.1. MPLS Vendor Capability TLV

The MPLS Vendor Capability TLV (Type 0x34) is used to advertise vendor-specific MPLS features. It contains:

```
   0                   1                   2                   3
   0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  | TLV Type      | TLV Length                                    |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                   Vendor ID (IANA PEN, 32 bits)               |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                         Capability Data (variable)            |
  ~                                                                 ~
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Devices that do not recognize the PEN MUST ignore the TLV. Vendors are encouraged to structure the Capability Data as sub-TLVs for interoperability and future extensibility. This follows the mechanism described in RFC 8029 and RFC 4360.

4.2. Capability Sub-TLV Structure

Vendors SHOULD structure the payload as sub-TLVs:

```
+--------+--------+----------------------+
| Sub-T  | Sub-L  | Sub-Value            |
+--------+--------+----------------------+
```

Where:

```
Field   Size    Description
Sub-T   1 byte  Sub-Type: Identifies the capability being advertised.
                     Each vendor can define its own sub-types.
Sub-L   1 byte  Sub-Length: Length of the Sub-Value field in bytes.
Sub-Value variable The actual value/data for the capability.
```

Sub-TLV Examples

```
Sub-Type  Name               Description
0x01      AI Traffic Class   AI / RDMA traffic marking
0x02      Optical Awareness  Cross-layer optical hints
0x03      Fast Reroute Variant Proprietary FRR behavior
0x04      Telemetry Export   Streaming / in-band telemetry
0x05      ECMP Hash Seed     Vendor hash control
```

5. Examples

Example 1: AMD MI300X Node

```
   TLV 0x01: 2        # AMD
   TLV 0x02: 3001     # MI300X
   TLV 0x03: 6        # GPUs free
   TLV 0x04: 8        # Total GPUs
   TLV 0x05: 160      # HBM free (GB)
   TLV 0x11: 1001     # Job Affinity Group
```

Example 2: NVIDIA H100 Node

```
   TLV 0x01: 1        # NVIDIA
   TLV 0x02: 1001     # H100
   TLV 0x03: 4        # GPUs free
   TLV 0x04: 8        # Total
   TLV 0x05: 320      # Memory free (GB)
   TLV 0x34: <Vendor PEN + Capability Data>  # MPLS vendor extension
```

6. Deployment Considerations

The GPU_CAPABILITY attribute is optional. Devices that do not recognize this attribute MUST ignore it and continue normal BGP processing.

Leaf switches originate GPU_CAPABILITY attributes on behalf of attached GPU servers. Spine switches propagate the attribute transparently.

Schedulers and orchestration platforms MAY ingest the BGP RIB to make placement decisions based on proximity, interconnect domain, congestion state, and GPU availability.

The attribute integrates with EVPN/VXLAN overlays and supports multi-vendor fabrics including NVIDIA, AMD, Qualcomm, Intel, and future accelerators.

7. Security Considerations

The GPU_CAPABILITY attribute introduces operational resource data into the BGP control plane. Incorrect or malicious advertisements could mislead schedulers and orchestration systems.

Implementations SHOULD restrict origination of this attribute to trusted devices. Standard BGP security mechanisms such as TCP-AO, GTSM, and RPKI SHOULD be used where applicable.

8. IANA Considerations

- Allocate BGP Path Attribute Type Code for GPU_CAPABILITY
  - Flags: Optional, Transitive
  - Type Code: TBD

- Create and maintain a registry of GPU_CAPABILITY TLV Types

- Reserve TLV Type 0x34 for MPLS Vendor Capability
  - Use IANA Private Enterprise Number (PEN) for Vendor ID
  - Follow RFC 8029 ignore-if-unknown processing

9. References

9.1 Normative

[RFC4271] Y. Rekhter, T. Li, S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC7432] A. Sajassi, et al., "BGP MPLS-Based Ethernet VPN", RFC 7432, February 2015.

[RFC8092] K. Patel, et al., "BGP Large Communities", RFC 8092, February 2017.

[RFC8029] A. Farrel, et al., "MPLS LSP Ping Vendor-Specific TLVs", RFC 8029, February 2017.

9.2 Informative

ROCm SMI Documentation
NVIDIA DCGM Documentation
Qualcomm AI Accelerator Architecture

Author’s Address

Alexander Montrose
Email: alexandermontrose.ietf@gmail.com