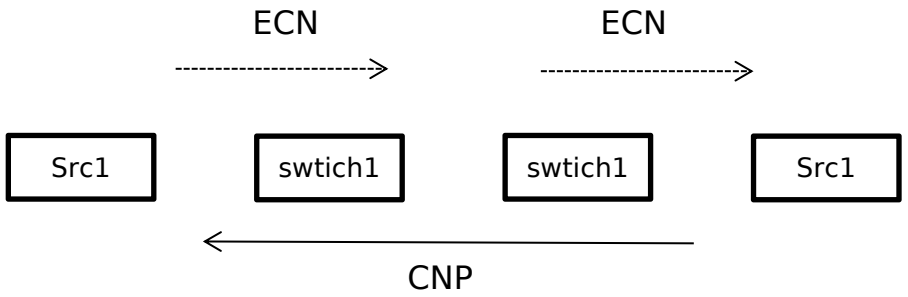
Multi-level Congestion Response Framework with Long-haul Congestion Notification for DCI Networks

draft-tian-ccwg-long-haul-cnp-00

Background

In long-haul transmission scenarios, network congestion causes severe throughput degradation and link underutilization. While standard CNP mechanisms perform well in low-latency DCNs, in high-latency DCI networks, the slow feedback loop leads to congestion propagationand performance collapse.

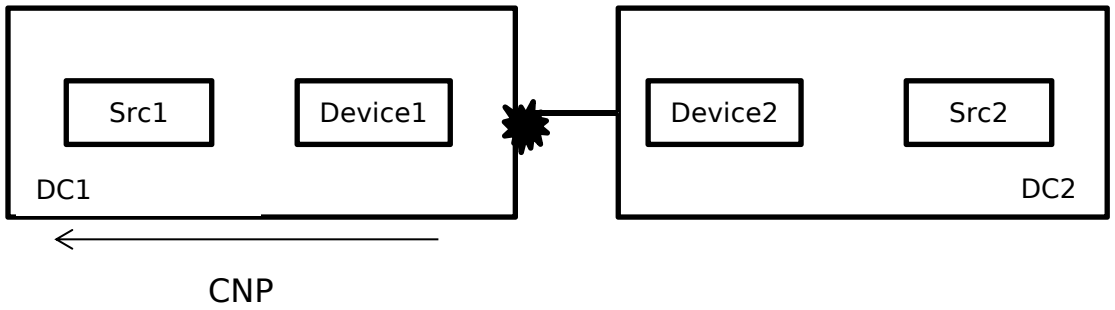
Standard CNP path: Switch marks ECN → Receiver generates CNP → CNP returns to sender.



Problem: In DCI, WAN RTT can be 10ms+, causing >20ms feedback delay → buffer overflow, throughput collapse

Related Work

Fast CNP （draft-xiao-rtgwg-rocev2-fast-cnp-04）：Enables switch-originated CNP, cuts latency in half



Our Extensions:

- Multi-level response (over-reacts to micro-bursts)
- Explicit control instructions (only "reduce")
- Multi-node coordination (duplicate CNPs)

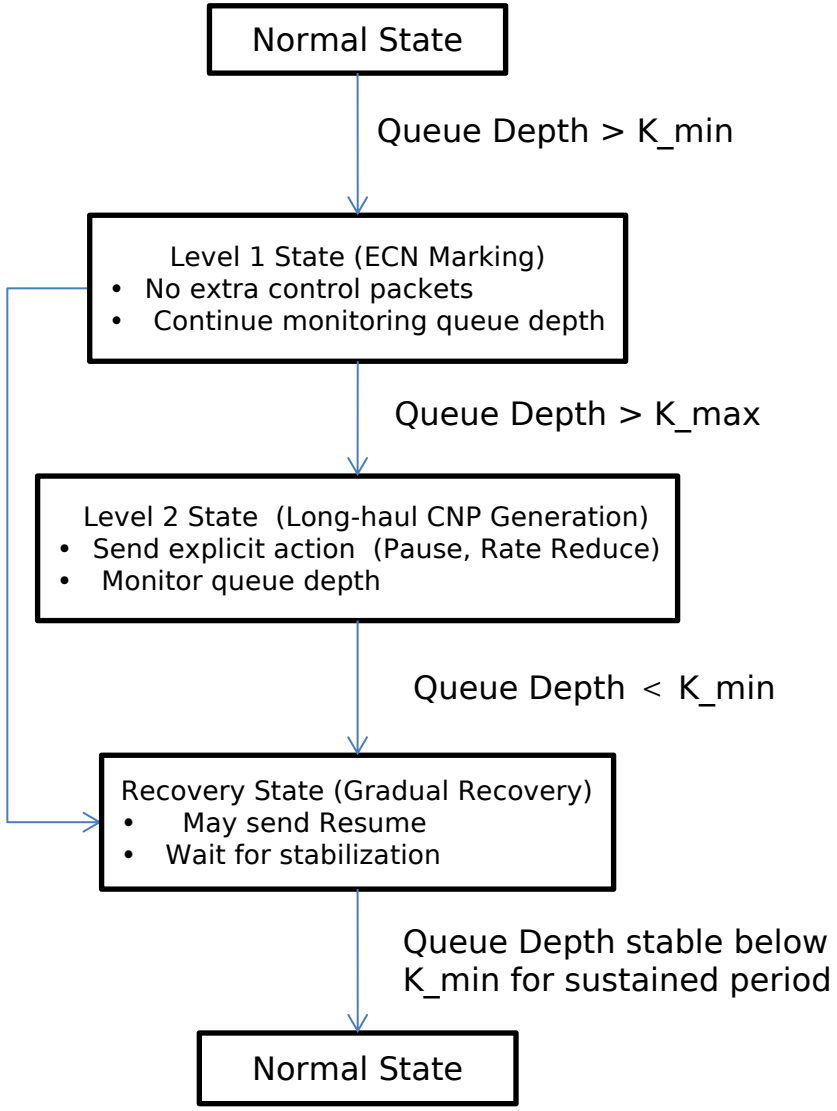
Solution Overview

- Two-level trigger:

  Level 1 (Incipient): ECN marking

  Level 2 (Severe): Long-haul CNP generation

- BDP-adaptive thresholds: K_max dynamically calculated based on link RTT & bandwidth
- Explicit control instructions: Notify, Pause, Rate Reduce, Resume (with parameters)
- Multi-device coordination: Upstream responds first, downstream defers



BDP-Adaptive Dynamic Threshold Calculation

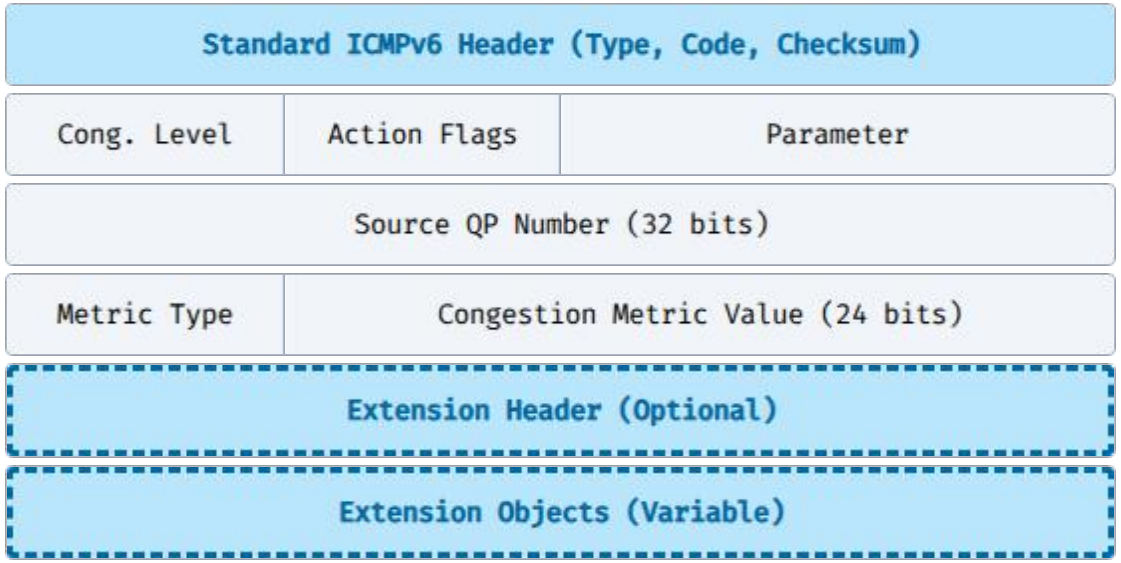
Because DCI link characteristics (bandwidth, distance) vary widely, static queue thresholds are inadequate. The upper queue depth threshold (**K_max**) adapt based on the Bandwidth-Delay Product (BDP).

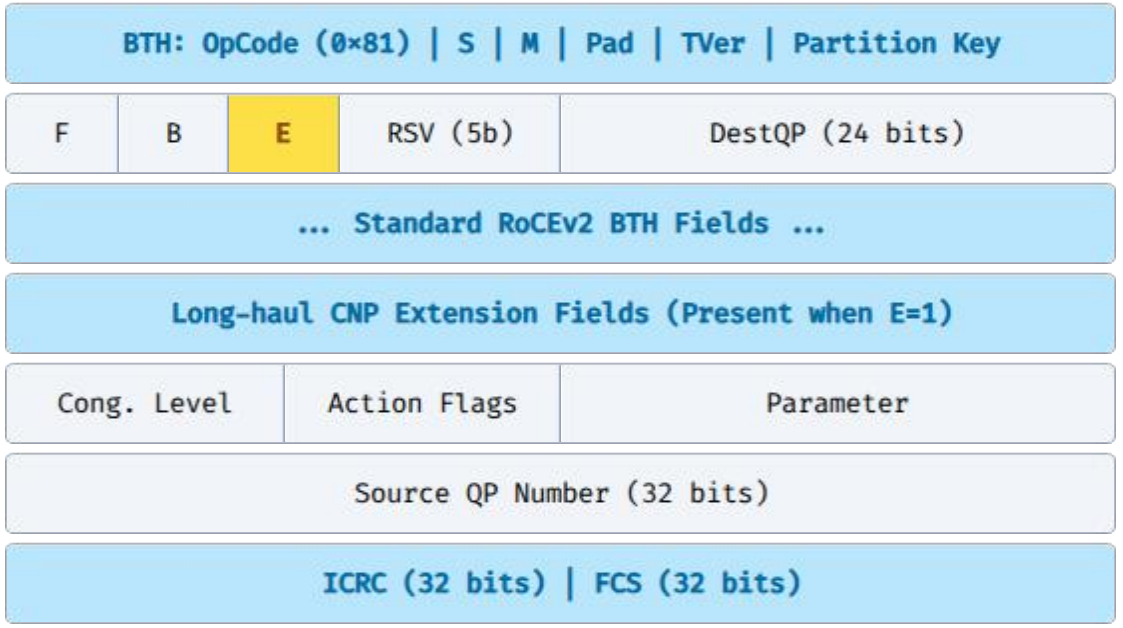
$$K_{max} = max(K_{base},\ \alpha \times \frac{R_{port} \times RTT_{est}}{8})$$

- **R_port:** Egress port rate (bps)
- **RTT_est:** Estimated round-trip time
- **K_base:** Baseline threshold floor to protect short-distance/low-speed links from starvation
- **α :** Adjustment coefficient (default 1.0)

The minimum threshold (**K_min**) for Level 1 ECN marking is typically recommended to be K_max / 2.

Packet Format Details





ICMPv6 Extension Format

- Fixed 12-byte body – avoids RFC 4884 length ambiguity
- Optional Extension Header – follows RFC 4884 object structure (Class-Num, C-Type, Length)
- Metric Value masking – Type=0, Value=0 for security/privacy-sensitive deployments

RoCEv2 Backward-Compatible Format

- Reuses standard CNP Opcode (0x81) – no new IBTA Opcode required
- Defines E bit – most significant bit of 6-bit reserved field after BECN
  E=0: Standard CNP
  E=1: Long-haul CNP extension follows BTH
- ICRC covers extensions natively – legacy RNICs pass CRC check, fallback gracefully

Questions?

Thank you for your time and attention.