



Implementing In-Kernel toFQDN Policies with Cilium & eBPF

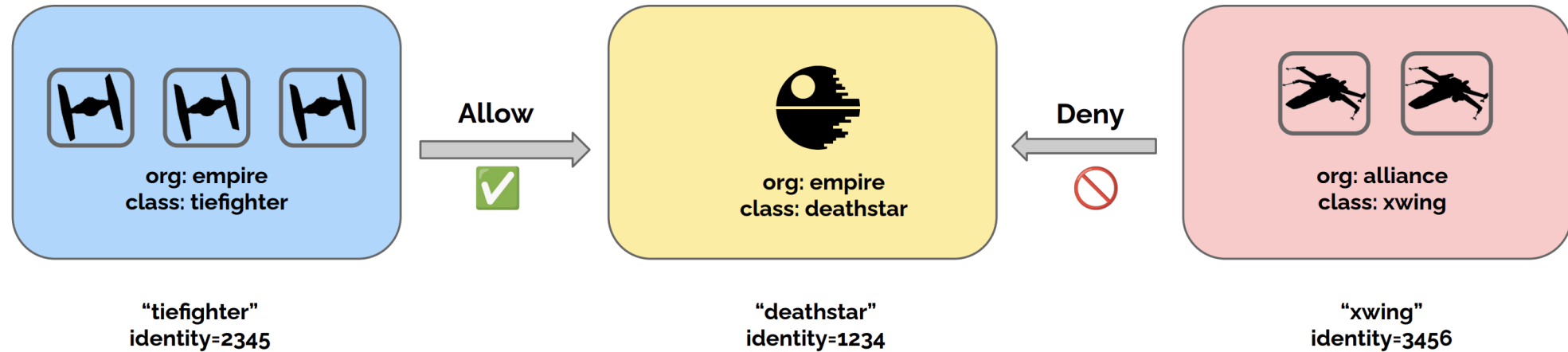
Contents

- 01 Why ?
- 02 Intro to Cilium's toFQDN Policy
- 03 Limitations and HA Mode
- 04 Challenges with eBPF Parser
- 05 Closer look at wire protocol
- 06 Looping in eBPF
- 07 Stream Parsing
- 08 Next Steps

DNS Egress Policy | toFQDN

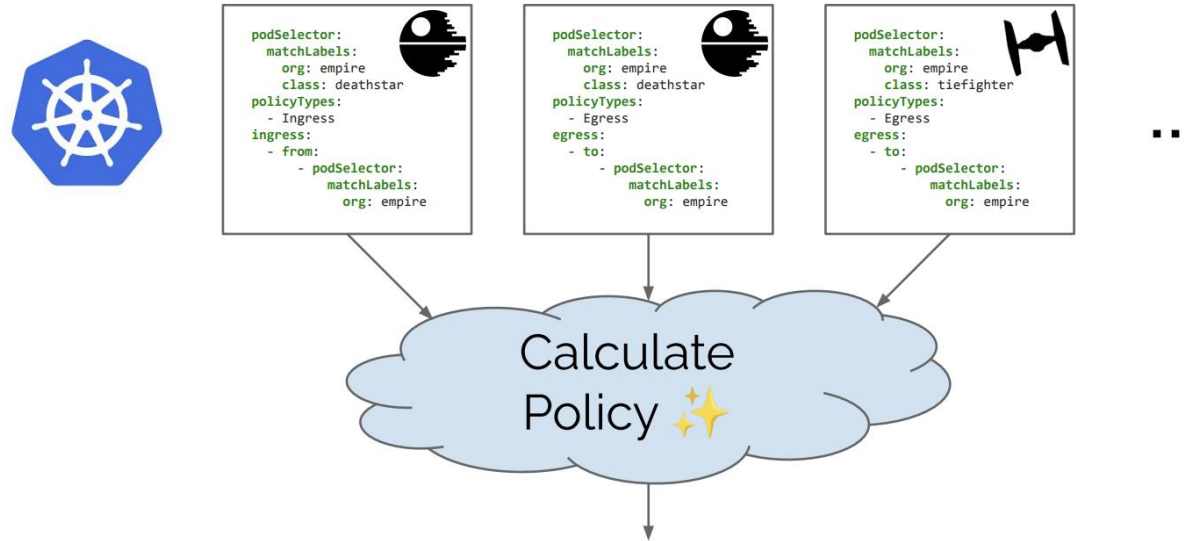
```
kind: CiliumNetworkPolicy
metadata:
  name: "fqdn"
spec:
  endpointSelector:
    matchLabels:
      org: empire
      class: mediabot
  egress:
    - toFQDNs:
      - matchName: "api.github.com"
    - toEndpoints:
      - matchLabels:
          "k8s:io.kubernetes.pod.namespace": kube-system
          "k8s:k8s-app": kube-dns
  toPorts:
    - ports:
      - port: "53"
        protocol: ANY
  rules:
    dns:
      - matchPattern: "*"
```

Identity based Policy



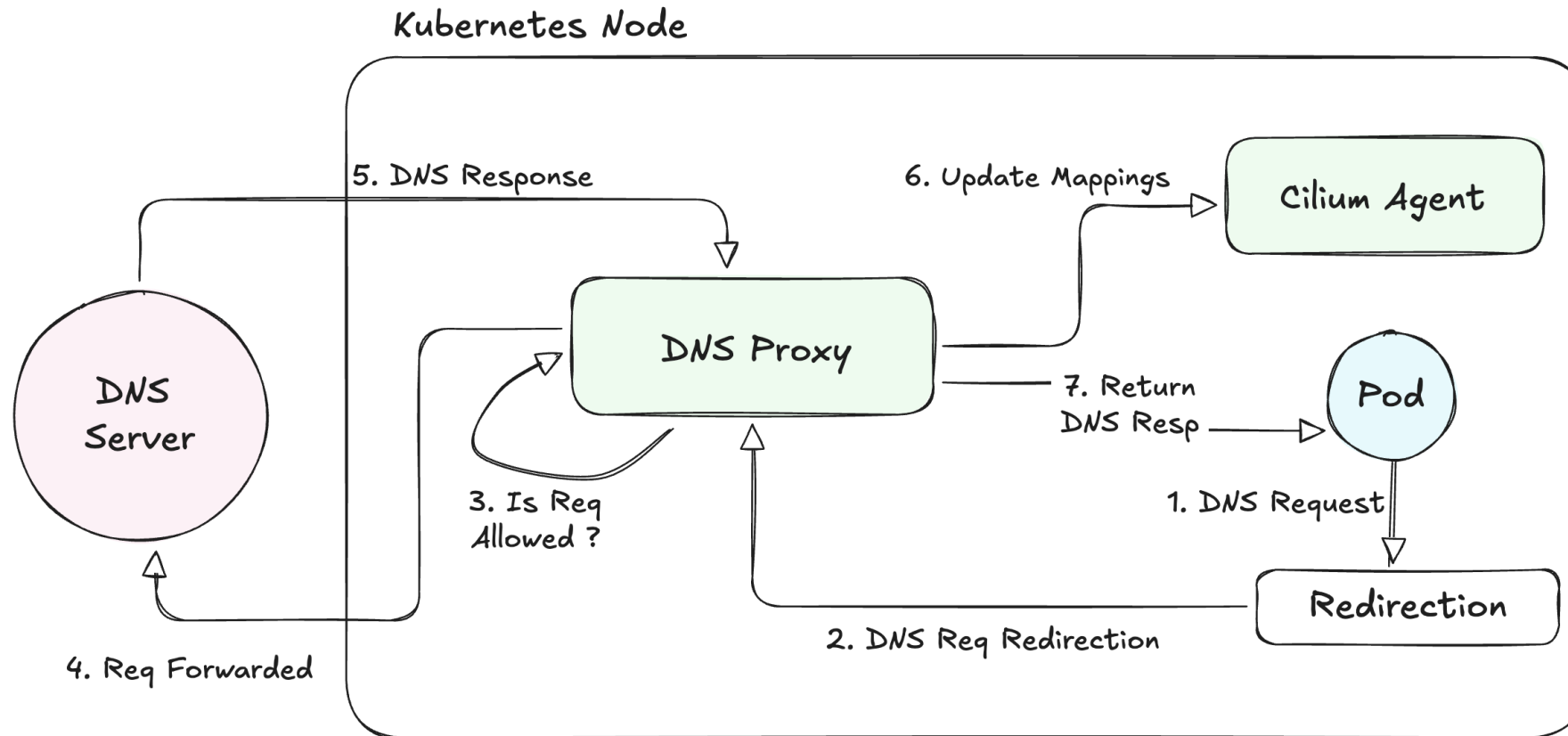
Protection against pod churn from existing / identical workloads

Identity based Policy



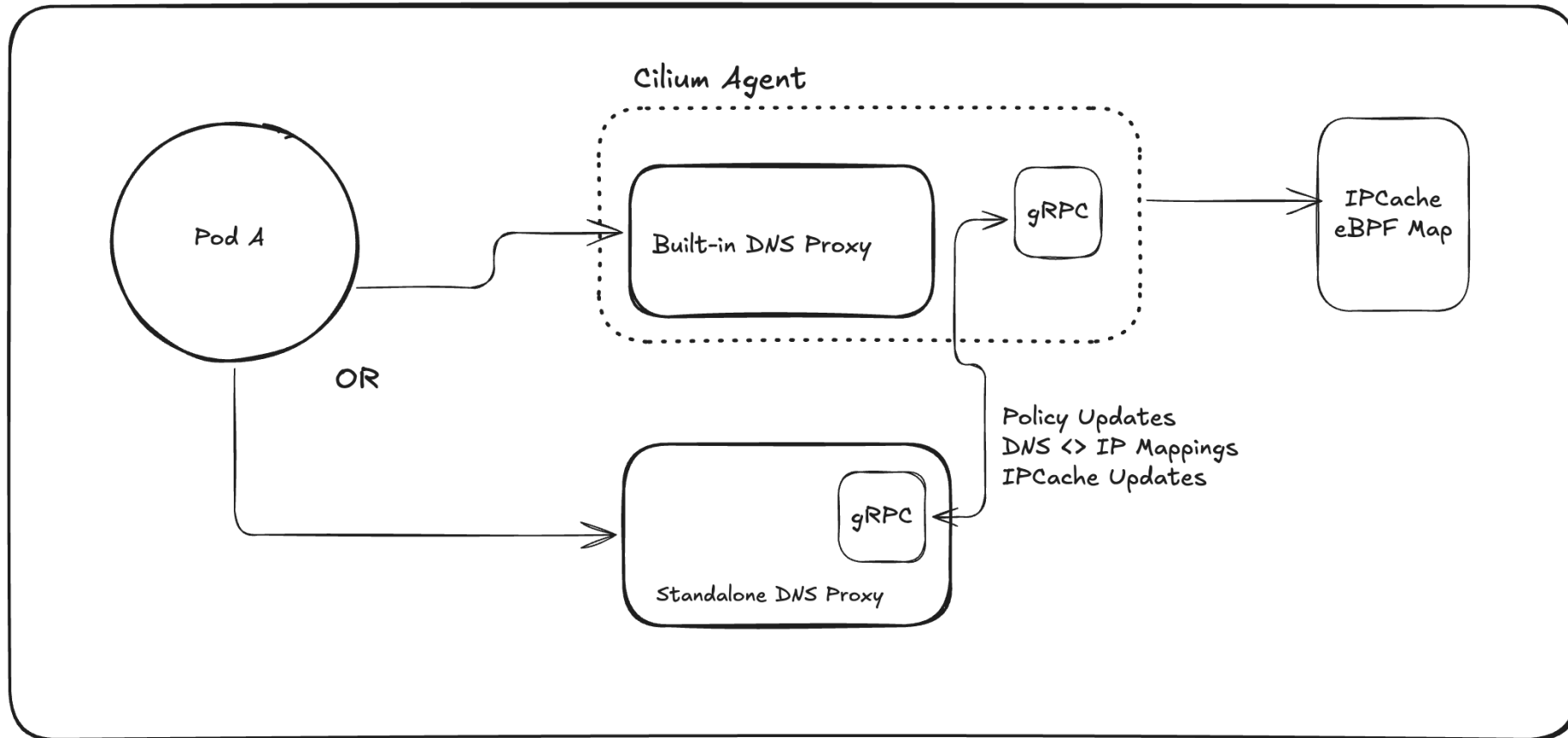
Local Identity	Peer Identity	Destination Port	Proto	Direction	Action*
1234	2345	*	*	Ingress	ALLOW
1234	3456	*	*	Ingress	DROP
*	*	*	*	Egress	ALLOW

DNS Egress Policy | toFQDN



toFQDN HA-ish Mode

Node



<https://github.com/cilium/design-cfps/pull/54>

How about a native eBPF implementation ?

Thanks to Mahé for planting the seed

Challenges with parsing DNS in eBPF

- Unknown number of DNS answers
- Label compression
- UDP Truncation
- DNS over TCP – Stream Handling
- Complexity issues
- Allocating Identity from bpf ?

FQDN Identities

FQDN: preallocate identities for fqdn selectors #39868

Merged **squeed** merged 4 commits into `cilium:main` from `squeed:namemanager-preallocate-identities` on Jun 17

Conversation 36

Commits 4

Checks 53

Files changed 20



squeed commented on Jun 3

Member ...

Background: the ToFQDN policy feature works by assigning labels with source `fqdn:` to IP addresses as names are learned. For example, the IP address 1.1.1.1 may have the labels `(reserved:world, fqdn:one.one.one.one)`. (Note that IPs no longer have per-IP labels unless also selected by toCIDR policies.)

Before this change, identities were allocated when the first IP was discovered that matched a selector. Now, identities are allocated when ToFQDN policies are first created.

The goal is to reduce tail latency. Because allocating an identity requires a policy update, all endpoints must lock, apply incremental changes, and update envoy *before* the DNS packet can be returned to the requesting pod.

By pre-allocating identities, newly learned IPs require only an ipcache write, not a policymap and envoy update. This reduces DNS response latency.

Reduces ToFQDN selector tail latency by pre-allocating selected identities. This slightly increases bpf p

RFC 1035

RFC 1035

Network Working Group
Request for Comments: 1035

P. Mockapetris
ISI
November 1987

Obsoletes: RFCs 882, 883, 973

DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION

1. STATUS OF THIS MEMO

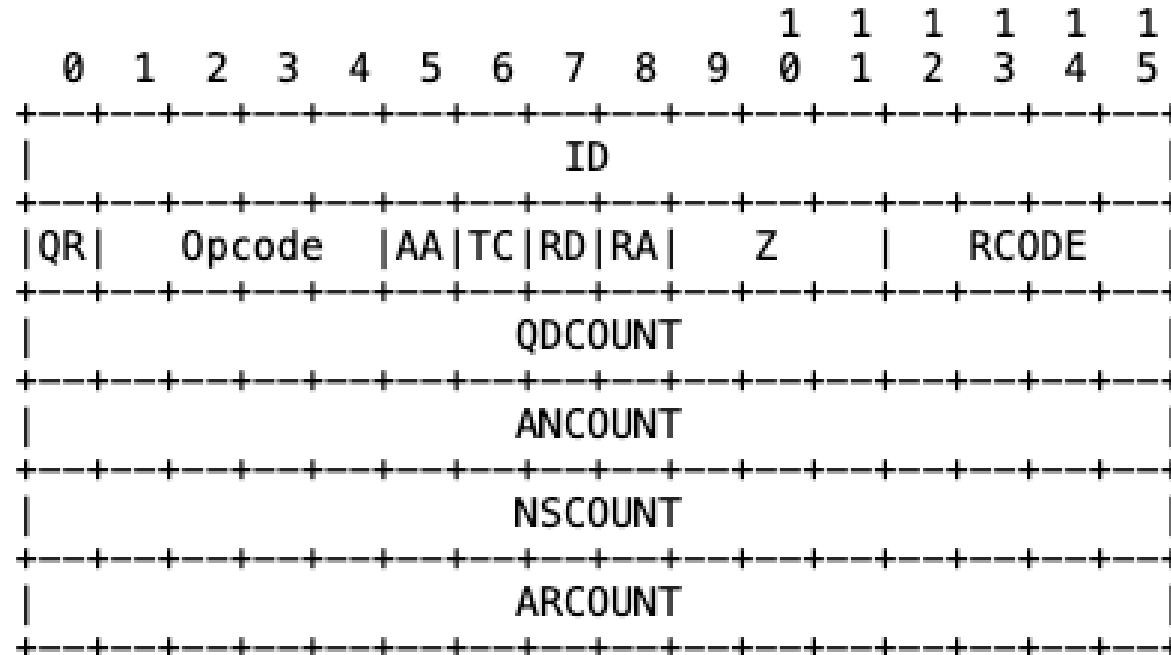
This RFC describes the details of the domain system and protocol, and assumes that the reader is familiar with the concepts discussed in a companion RFC, "Domain Names - Concepts and Facilities" [RFC-1034].

The domain system is a mixture of functions and data types which are an official protocol and functions and data types which are still experimental. Since the domain system is intentionally extensible, new data types and experimental behavior should always be expected in parts of the system beyond the official protocol. The official protocol parts

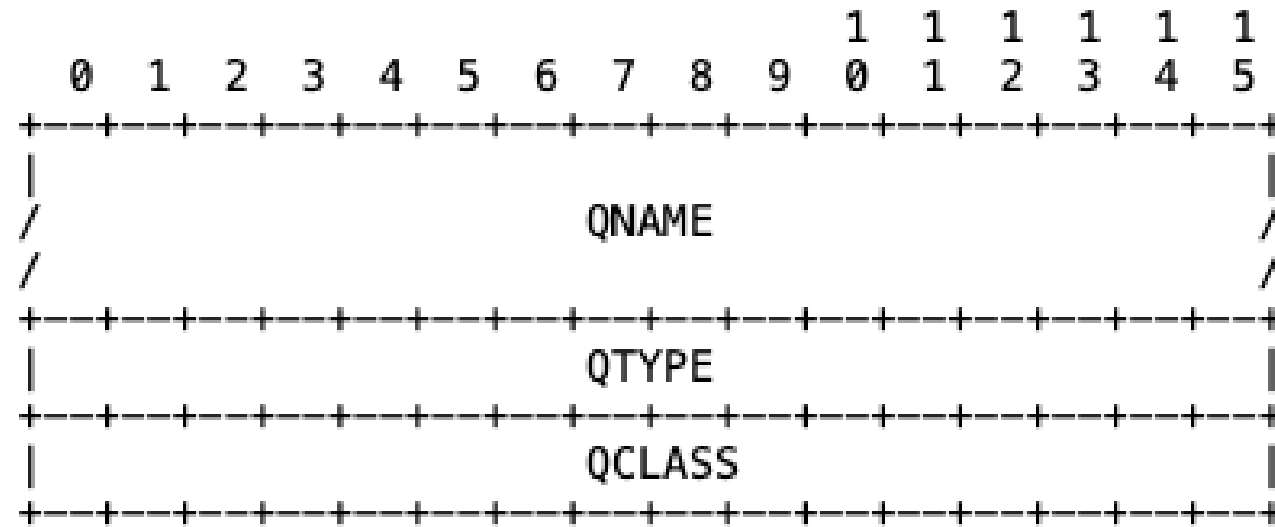
Wire Format – DNS Message

Header	
Question	the question for the name server
Answer	RRs answering the question
Authority	RRs pointing toward an authority
Additional	RRs holding additional information

Wire Format – DNS Header

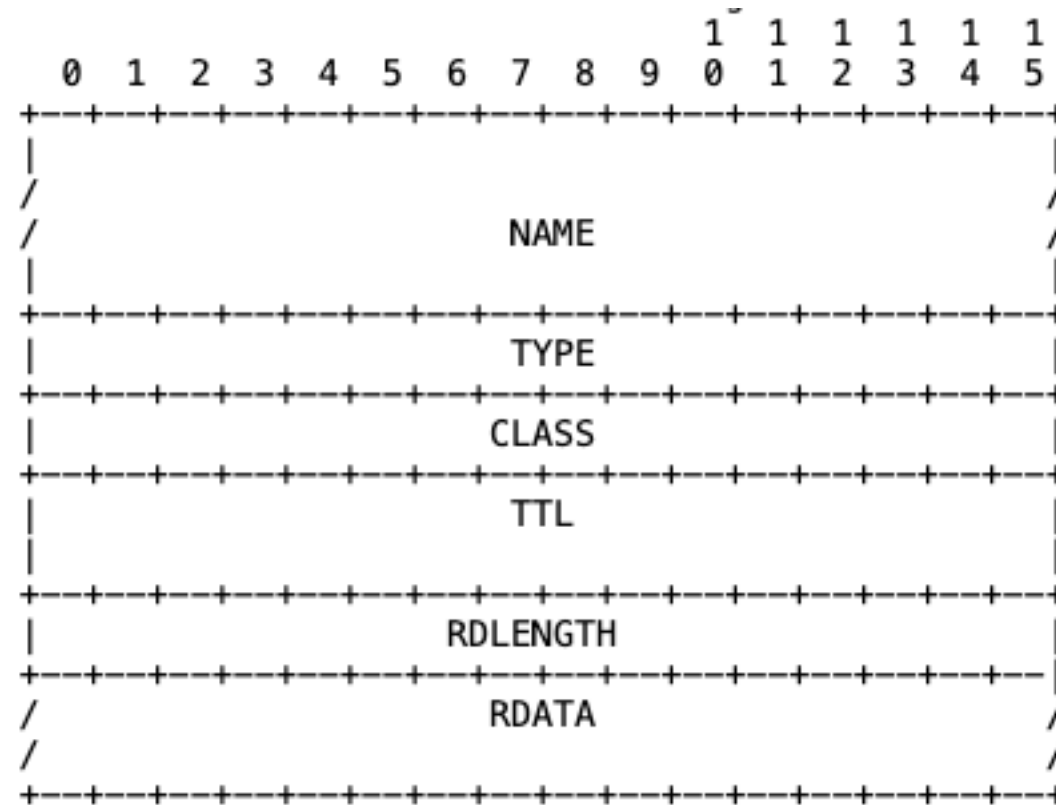


Wire Format – DNS Question Section



Label Length – Label – Repeat until zero length

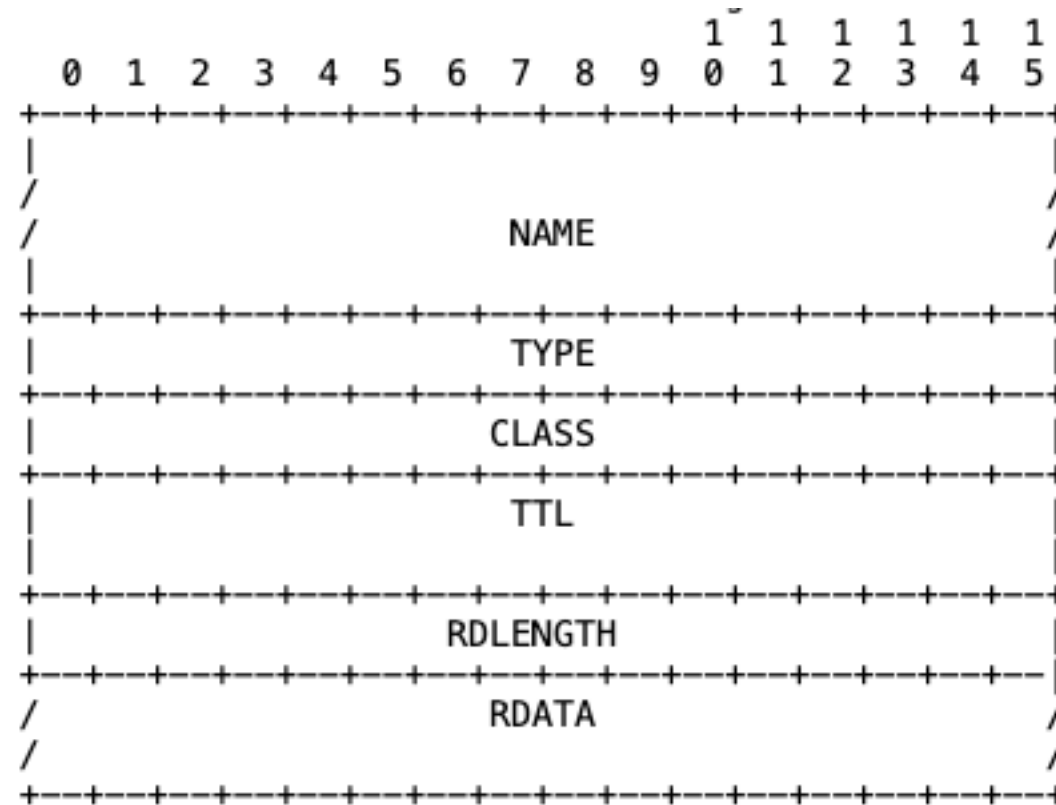
Wire Format – Resource Record



Wire Format – Label Compression

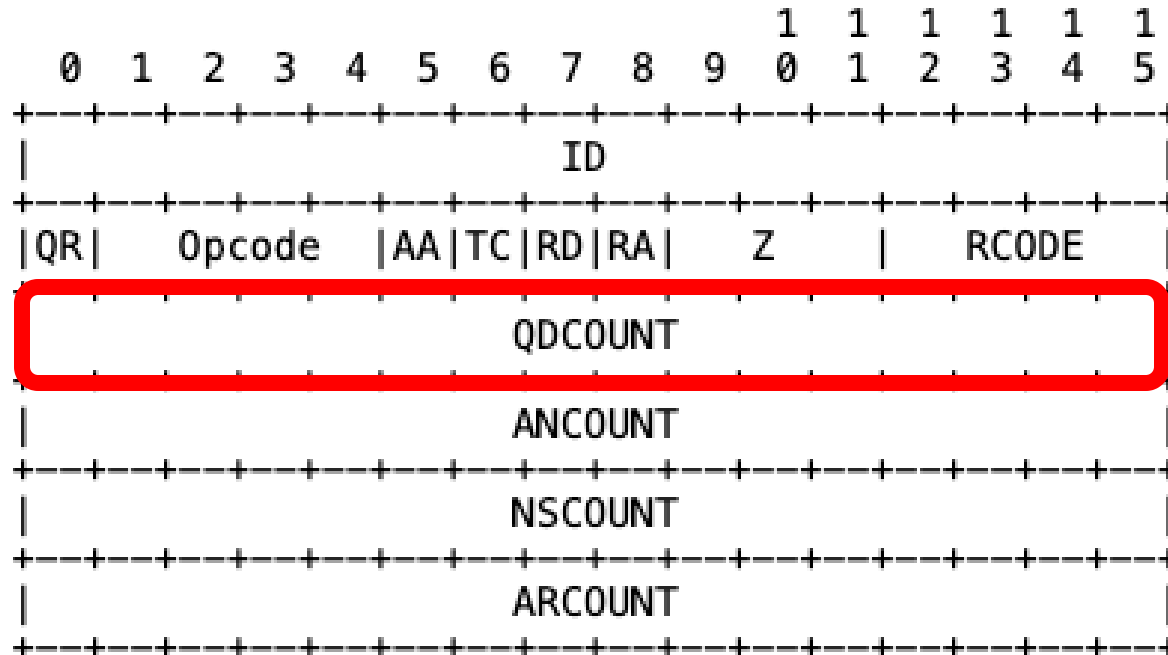


Wire Format – RDLENGTH and RDATA



32 Bits – IPv4 || 128 Bits IPv6

Wire Format – Multiple Questions ?



Wire Format – Multiple Questions ?

RFC 9619


Internet Engineering Task Force (IETF)
Request for Comments: 9619
Updates: 1035
Category: Standards Track
ISSN: 2070-1721

R. Bellis
ISC
J. Abley
Cloudflare
July 2024

In the DNS, QDCOUNT Is (Usually) One

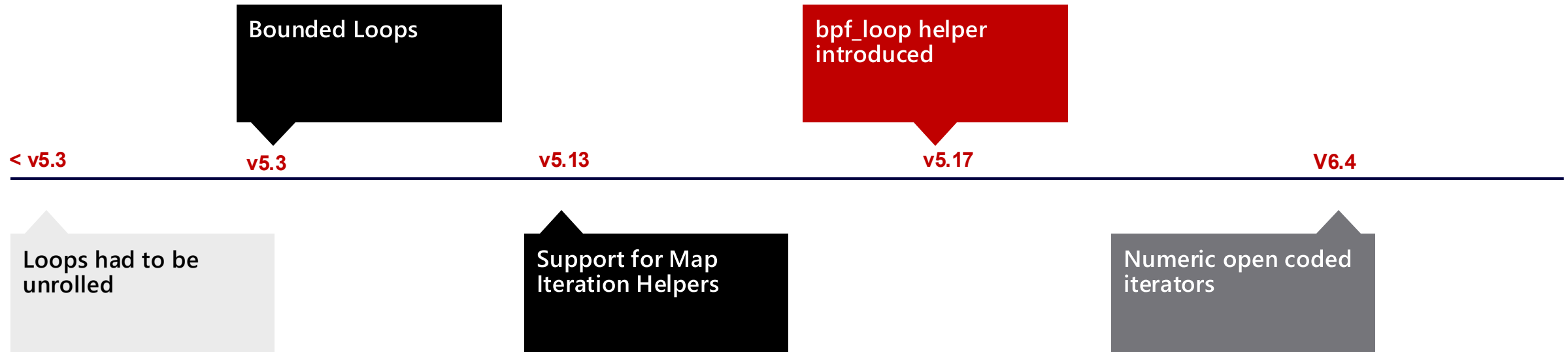
Abstract

This document updates RFC 1035 by constraining the allowed value of the QDCOUNT parameter in DNS messages with OPCODE = 0 (QUERY) to a maximum of one, and it specifies the required behavior when values that are not allowed are encountered.



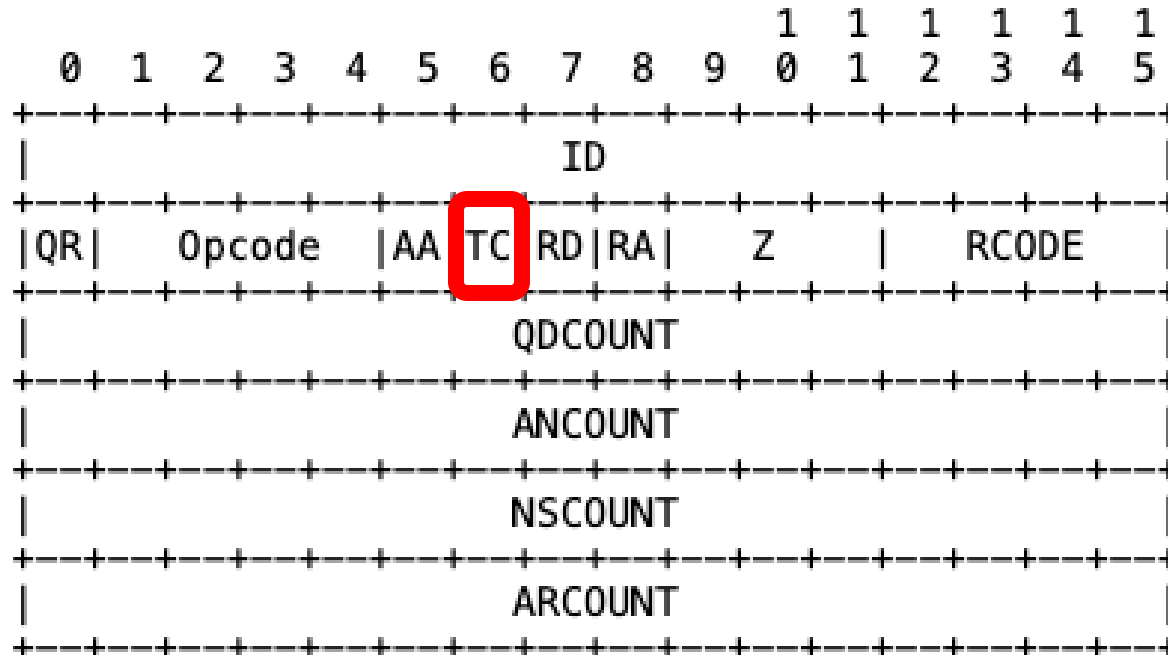
But we'll need to solve for multiple
answers

Loops in eBPF



Bpf_loop from v5.17+ Multiple answers

Wire Format – UDP Truncation

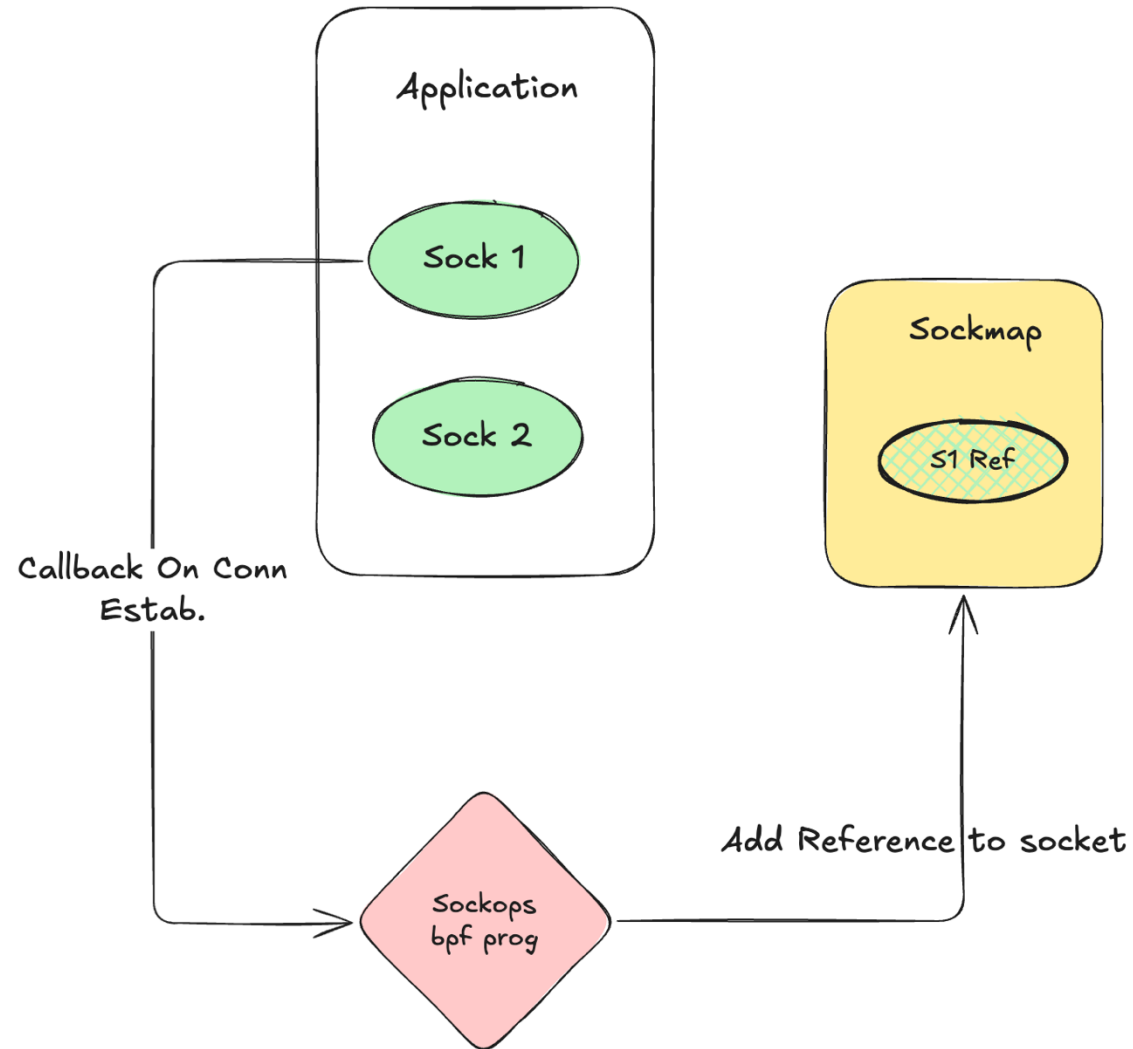


SK_SKB + SOCK_OPS eBPF Prog Types

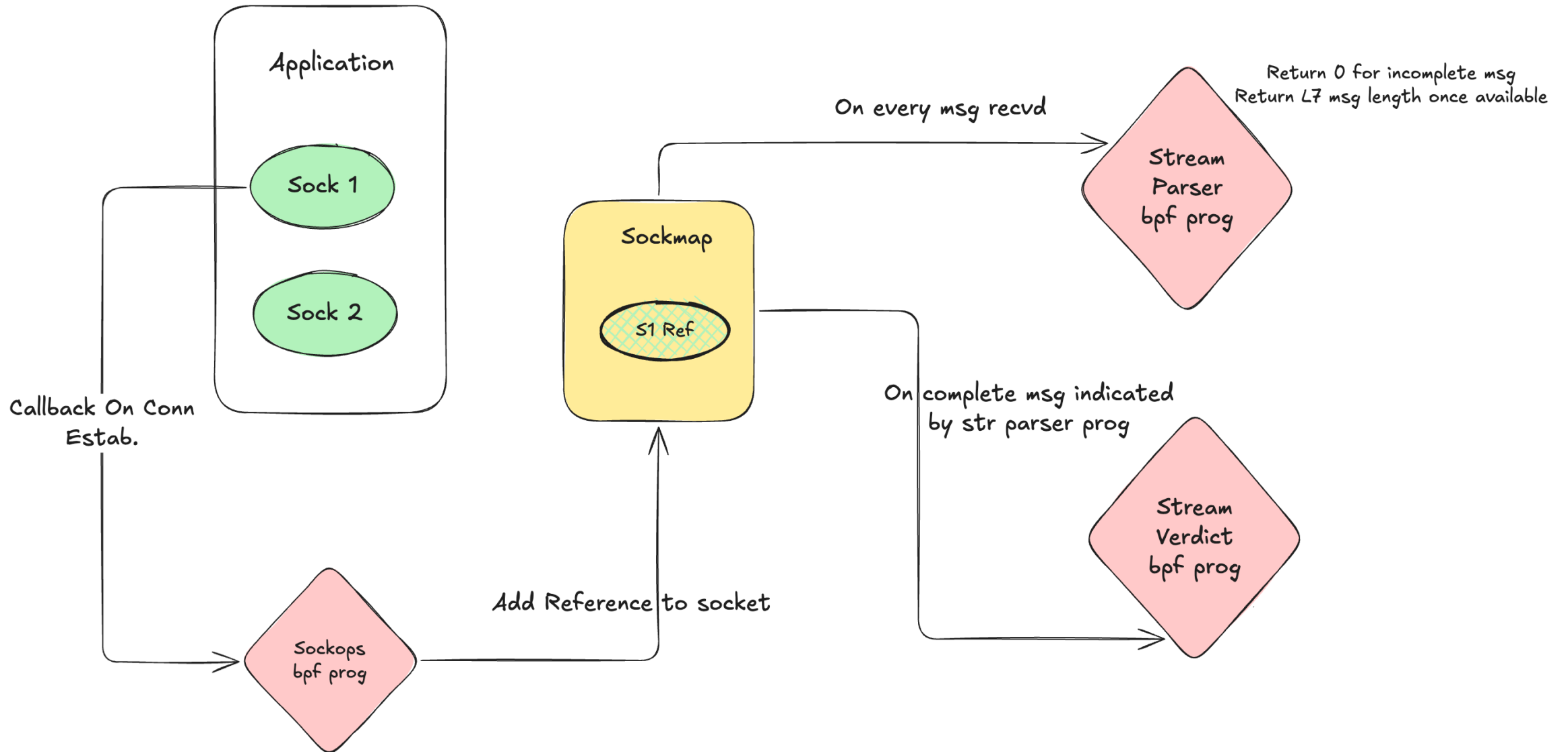
DNS over TCP Parsing with eBPF

- Sockmap + Sockops to register sockets
- Stream Parser attach type to check if full msg is received
- Stream verdict attach type for parsing and verdict

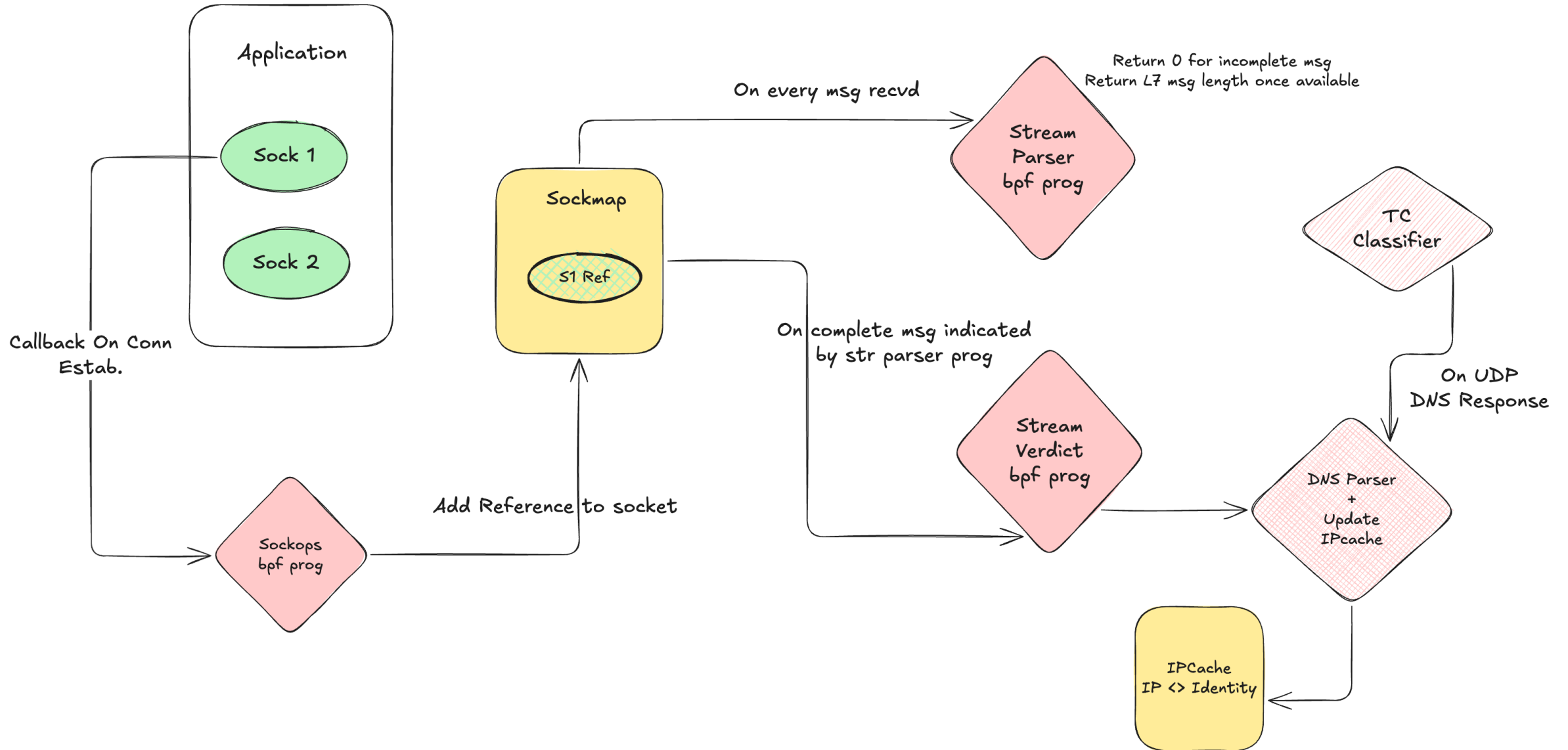
DNS over TCP Parsing with eBPF



DNS over TCP Parsing with eBPF



DNS over TCP Parsing with eBPF





Challenge #1

Non-Linear SKB Data

Non Linear Data

- Comparing DNS msg length with SKB len in stream parser is not enough
- bpf_skb_pull_data to the rescue
- Should this happen automatically ?

```
__section("sk_skb/stream_parser")
int dns_stream_parser(struct __sk_buff *skb) {
    __u16 msg_len;

    // Check if we can read the first 2 bytes
    if (skb->len < 2)
        return 0; // Not enough data, wait

    // Read first 2 bytes as DNS length prefix
    if (skb_load_bytes(skb, 0, &msg_len, 2) < 0)
        return SK_DROP;

    msg_len = builtin_bswap16(msg_len);
    if (skb_pull_data(skb, msg_len + 2) < 0)
        return SK_DROP;
    return msg_len + 2; // Total length including prefix
}
```



Challenge #2

Match Patterns with *

Current Approach

- Only care about patterns like *.<> or *.*.foo
- LPM_TRIE with reversed strings
- On length mismatch check number of dots
- Can we do better ?

Challenge #3

CNAME Chasing



Misc. Challenges

Loops + Data End Ptr

Verifier Improvements



Demo

Next Steps

- Garbage collection
- Proxy less DNS visibility
- Performance Tests
- DoH / DoT ?
- Try regex support from Justin's work ?



Questions ?



東京 **2025**

**LINUX
PLUMBERS
CONFERENCE**

TOKYO, JAPAN / DECEMBER 11-13, 2025

