Understanding CoreDNS in Kubernetes

Public link to this document

Speakers

- Cricket Liu
  - EVP Engineering and Chief DNS Architect, Infoblox
  - Co-author of *DNS and BIND, 5th Edition*
  - Co-author of all of O'Reilly Media's books on DNS
- François Tur
  - Senior Software Manager at Infoblox
  - CoreDNS maintainer, @fturib
- John Belamaric
  - Senior Staff Software Engineer, Google
  - CoreDNS maintainer, @johnbelamaric

Agenda

- Introduction
- Cluster DNS Default Configuration
- Outside the Defaults
- More Resources

Introduction

- Flexible DNS server written in Go
- Plugin-based architecture, easily extended
  - To support different cloud-native stacks, for example
- Supports DNS, DNS over TLS (DoT), DNS over gRPC
- Started and led by Miek Gieben, author of SkyDNS and SkyDNS2
- Originally a fork of the Caddy HTTP server (“Caddy DNS”)

CoreDNS

- Native support of service discovery for Kubernetes
  - Generally available with Kubernetes 1.11
  - Now the default in 1.13
- Integration with *etcd* and cloud vendors (e.g., AWS's Route 53)
- Support for Prometheus metrics
- Forwarding to recursive DNS server

Why CoreDNS (vs. kube-dns)?

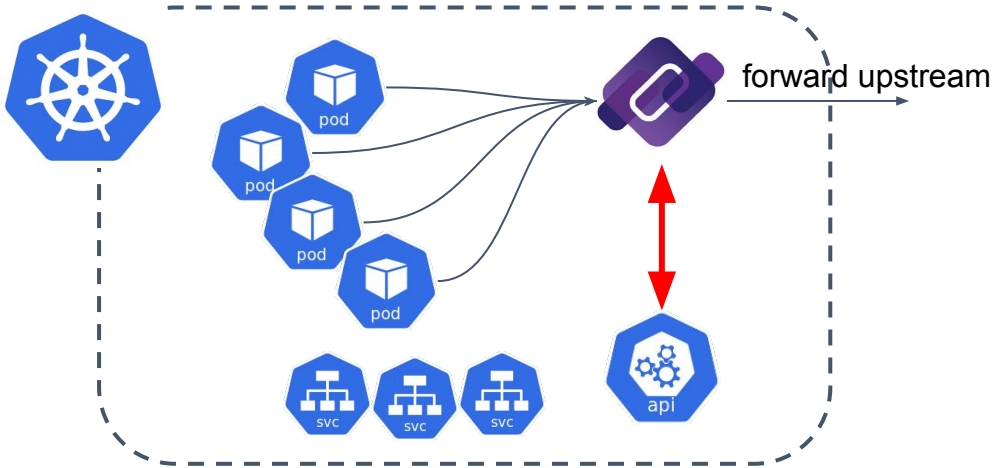
- Easily extensible plugin architecture
- Rich set of (~34) plugins, with new ones being developed all the time
- Simpler, with fewer moving parts (single executable and process)
  - And all written in Go
- Customizable DNS entries in and out of the cluster domain
- Experimental server-side search path to reduce query volume

Project Status

- Version 1.2.6 (released 11/5/2018)
- Incubating project in CNCF
  - Graduation vote underway
- Growing community
  - 112 contributors (big thanks!)
  - 16 maintainers
  - 29+ public adopters
  - 3000+ stars

CoreDNS as Cluster DNS

- CoreDNS Kubernetes Resources
- Default Corefile
- Resolving a Query
- Stub Domains
- Cache Tuning



Kubernetes Resources

coredns -n=kube-system

kube-dns -n=kube-system


```
data:
 Corefile: |
  .:53 {
  …
 }
```


```
spec:
 replicas: 2
 Template:
  metadata labels:
    k8s-app: kube-dns
  spec:
    containers:
     image: k8s.gcr.io/coredns:1.2.6
     resources:
      limits:
       memory: 170Mi
      requests:
       cpu: 100m / memory: 70Mi
     …
     livenessProbe: http://8080
    dnsPolicy: Default
```


```
spec:
 selector:
  k8s-app: kube-dns
  clusterIP: x.x.x.10
 Ports:  …
  53 - UDP/TCP
  9153 - TCP
```

Default Corefile

Enable error logging

Serve liveness status on http 8080

Backend to k8s for cluster.local and reverse domains
- Mimic kube-dns pod records behavior
- Resolve CNAME targets upstream
- Continue searching reverse zones

Serve prometheus metrics

Forward other domains to /etc/resolv.conf ns

Cache for up to 30 seconds

DNS protocol loop check

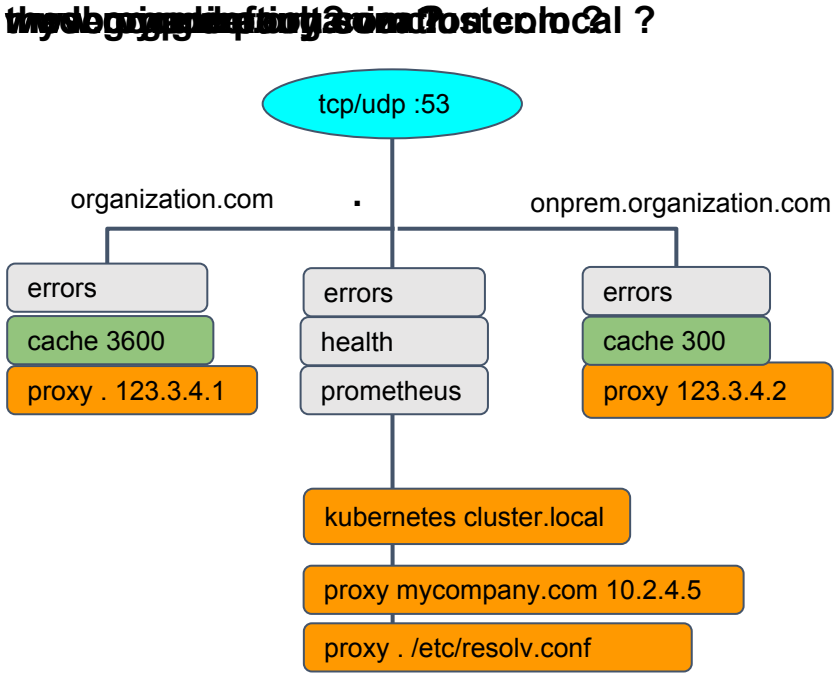
Reload server if the Corefile change

Shuffle order of returned records

```
.:53 {
    errors
    health
    kubernetes cluster.local in-addr.arpa ip6.arpa {
       pods insecure
       upstream
       fallthrough in-addr.arpa ip6.arpa
    }

    prometheus :9153
    proxy . /etc/resolv.conf
    cache 30

    loop
    reload

    loadbalance
}
```

Resolving a Query


```
organization.com:53 {
    errors
    cache 3600
    proxy . 123.3.4.1
}
onprem.organization.com:53 {
    errors
    cache 300
    proxy . 123.3.4.2
}
.:53 {
    errors
    health
    kubernetes cluster.local ... {
        ...
    }
    proxy mycompany.com 10.2.4.5
    proxy . /etc/resolv.com
}
```



Stub Domains

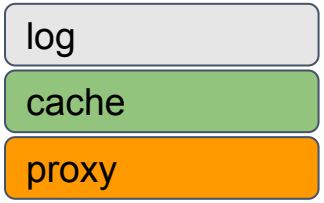

```
corporate.com :53 {
    proxy . <DNS-IP-corporate>
}
local.corporate.com :53 {
    proxy . <DNS-IP-local-corporate>
}

.:53 {
    ...
    kubernetes cluster.local ... {
       ...
    }
    ...
    proxy . /etc/resolv.conf
    ...
    cache 30
}
```

proxy

- Forward some out-of-cluster domains directly to the right authoritative DNS server
- Handle internal corporate domains

Cache Tuning


```
corporate.com {
    errors
    log innerservices.corporate.com
    proxy local.corporate.com <DNS-IP-local>
    proxy . <DNS-IP-corporate>
    cache 3600
}
.:53 {
    ...
    kubernetes cluster.local ... {
        ...
    }
    proxy . /etc/resolv.conf
    ...
    cache 30
}
```



- Allow specific configuration for known zone. cache, logging ...

When using plugin, check description and options at https://coredns.io/plugins/

Outside the Defaults

- Kubernetes Plugin Options
- Kubernetes-related Plugins
  - Autopath, external, kubernetai, and federation
- Adding Static Records
- Query Rewrites

Common Options for Cluster DNS

pods [**disabled** | insecure | verified]	When receiving query for pod : <IP-like>.<ns>.pod.<cluster.local> - `disabled` - unknown domain - `insecure` - reply with the IP provided - `verified` - reply only if IP is a pod IP - Watches all pods. Costly in memory & api server load
endpoint_pod_names	Use the pod name for endpoint address name, if the hostname is empty
ttl TTL	Change the TTL of kubernetes records
fallthrough [ZONES]	Second chance option : try with the next plugin in chain for resolving ‘unknown’ domains.

All Kubernetes Plugin Options

kubernetes [ZONES...] {

Connection to k8s API

- **resyncperiod** DURATION # API Server resync DURATION period
- **endpoint** URL [URL...] # API Server URL
- **tls** CERT KEY CACERT # API Server connection TLS config.
- **kubeconfig** KUBECONFIG CONTEXT # use kubeconfig context

Filtering/Exposing information

- **namespaces** NAMESPACE... # exposed ns
- **labels** EXPRESSION # selector for exposed services
- **ignore_empty_service** # do not expose service which no healthy endpoints
- **noendpoints** # don't watch endpoints resource - therefore ignore Headless services
- **upstream** [ADDRESS...] # resolve External services and define specific nameserver

Tune domain naming

- **pods** POD-MODE # pods mode: disabled,insecure,verified
- **endpoint_pod_names** # allow using pod names to identify Headless service domain

DNS workflow

- **ttl** TTL # TTL default: 5s
- **fallthrough** [ZONES...] # no record found in the domain, second chance with next plugin
- **transfer** to ADDRESS… # allow transfer to secondary DNS server

}

Plugins

Complete list of plugins available at coredns.io

Complete list of external plugins

Configuration	Middleware	Backends
metadata	loadbalance	template
tls	cache	host
reload	rewrite	route53
bind	dnssec	kubernetes
health	autopath	file
prometheus		auto
errors		etcd
log	loop	forward
		proxy

External
kubernetai
redisc

- Only most commonly used are mentioned
- They are shown in order of chaining in the default image (top to bottom, left to right)

Other Kubernetes-related Plugins

- Autopath
  - Server-side search path resolution
- K8s External
- Kubernetai
- Federation

Autopath - the problem

- Kubernetes has a long DNS search path and ndots value
  - `<namespace>.svc.cluster.local`
  - `svc.cluster.local`
  - `cluster.local`
  - plus the nodes search path
- Enables flexible use of names, but leads to extra queries

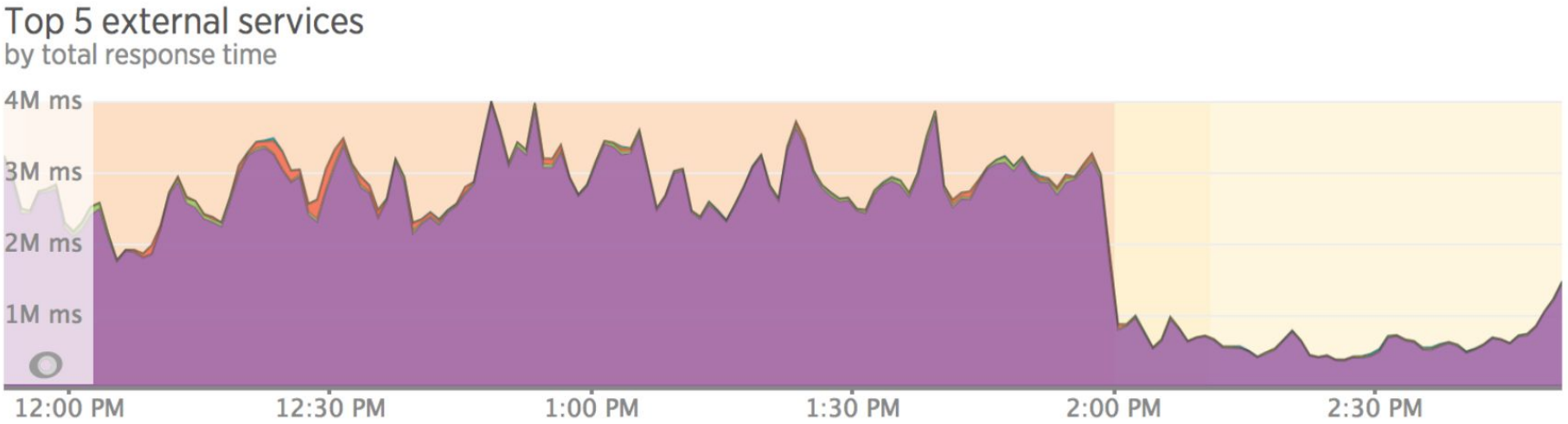
```
dnstools# host -v google.com
Trying "google.com.default.svc.cluster.local"
Trying "google.com.svc.cluster.local"
Trying "google.com.cluster.local"
Trying "google.com"
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 62752
;; flags: qr rd ra; QUERY: 1, ANSWER: 1, AUTHORITY: 4, ADDITIONAL: 4
...
```

Autopath - the solution

- kubernetes pods verified + autopath
- Since CoreDNS knows the namespace of the source pod IP, it knows the search path
- Execute the search path server-side

```
dnstools# host -v google.com
Trying "google.com.default.svc.cluster.local"
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 38177
;; flags: qr rd ra; QUERY: 1, ANSWER: 2, AUTHORITY: 0, ADDITIONAL: 0

;; QUESTION SECTION:
;google.com.default.svc.cluster.local. IN A

;; ANSWER SECTION:
google.com.default.svc.cluster.local. 13 IN CNAME google.com.
google.com.        13   IN   A    172.217.9.142
...
```

Autopath - results



So, why not default??

- Requires **`pods verified`** - too much memory, too much API server load
- Edge case can result in a pod getting the wrong response from the cache
  - Mitigated by enabling only negative cache…?

Sneak Peek: Kubernetes External

- Coming in version 1.3.0
- Service ExternalIPs and LoadBalancer IPs published in another zone
- Add to cluster DNS or (preferred) run a separate CoreDNS
- For example, external zone configured as “apps.example.com”:

```
apiVersion: v1
kind: Service
metadata:
  name: foo
  namespace: bar
spec:
  type: LoadBalancer
...
status:
  loadBalancer:
    ingress:
    - ip: 203.0.113.10
```

```
dnstools# host foo.bar.apps.example.com
foo.bar.apps.example.com has address 203.0.113.10
```

Kubernetai


```
. {
    errors
    log
    kubernetai cluster.local {
      endpoint http://192.168.99.100
    }
    kubernetai assemblage.local {
      endpoint http://192.168.99.101
    }
    kubernetai conglomeration.local {
      endpoint http://192.168.99.102
    }
}
```

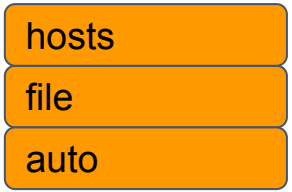
- Single CoreDNS serves Kubernetes service names for multiple clusters
- Remember cluster IPs are not routable
- Only really useful for headless services with routable pod IPs
- ...or for some Istio magic

A Couple Interesting Use Cases

- Add Static Data
- Query Rewriting

Add Static Records


```
myservices.com {
    file /etc/coredns/myservice.db myservices.com
    cache 3600
}
.:53 {
    ...
    hosts svc.cluster.local {
       167.8.9.2 undeployed.default.svc.cluster.local
       fallthrough
    }
    kubernetes cluster.local ... {
       ...
       fallthrough svc.cluster.local
    }
    auto cluster.local {
       directory /etc/coredns/cluster.local
    }
    proxy . /etc/resolv.conf
    ...
}
```



- Host non-cluster zones
- Override IP for specific name
- Fallback for missing services not yet migrated
- Populate subdomains other than **svc** and **pod** (a little dangerous)

Query Rewriting


```
.:53 {
    errors
    health
    rewrite {
        name regex (.*)\.demo\.com\.$ {1}.default.svc.cluster.local
        answer name (.*)\.default\.svc\.cluster\.local\.$ {1}.demo.com
      }

    kubernetes cluster.local ... {
        ...
    }
    ...
    proxy . /etc/resolv.conf
    ...
}
```

rewrite

- During migration of services, translate old naming into the Kubernetes DNS scheme.
- Use the same server cert without adding cluster name

Resources

DNS in Kubernetes Documentation

- Customizing DNS Service
- Using CoreDNS for Discovery Service
- Debugging DNS resolution

Kubernetes Blog

- CoreDNS GA for Kubernetes Cluster DNS

CoreDNS Blogs on the Kubernetes Plugin

- Cluster DNS : CoreDNS versus kube-dns
- Scaling CoreDNS in a Kubernetes cluster
- Migration from kube-dns to CoreDNS
- Deploying CoreDNS with kubeadm
- How queries are process in CoreDNS
- CoreDNS for Kubernetes Service discovery

Github Resources

- CoreDNS github
- Manual deployment of CoreDNS in kubernetes
- Kubernetes / DNS github

Community and Support

Thank you!

Issues/Questions/Support

github: http://github.com/coredns/coredns (also kubernetes/dns)

slack: https://slack.cncf.io #coredns

security related: security@coredns.io

Documentation/Resources

http://coredns.io - plugin docs. blogs.

Backup Slides

Kubernetes DNS Schema

ClusterIP Service

```
<service>.<ns>.svc.<zone>. <ttl> IN A <cluster-ip>
_<port>._<proto>.<service>.<ns>.svc.<zone>. <ttl> IN SRV
                    <weight> <priority> <port-number> <service>.<ns>.svc.<zone>.
<d>.<c>.<b>.<a>.in-addr.arpa. <ttl> IN PTR <service>.<ns>.svc.<zone>.
```

Headless Service

```
<service>.<ns>.svc.<zone>. <ttl> IN A <endpoint-ip>
<hostname>.<service>.<ns>.svc.<zone>. <ttl> IN A <endpoint-ip>
_<port>._<proto>.<service>.<ns>.svc.<zone>. <ttl> IN SRV
                    <weight> <priority> <port-number> <hostname>.<service>.<ns>.svc.<zone>.
<d>.<c>.<b>.<a>.in-addr.arpa. <ttl> IN PTR <hostname>.<service>.<ns>.svc.<zone>.
```

External Service

```
<service>.<ns>.svc.<zone>. <ttl> IN CNAME <extname>.
```

Pod (deprecated)

```
<a>-<b>-<c>-<d>.<ns>.pod.<zone>. <ttl> IN A <a>.<b>.<c>.<d>
```

DNS Version

```
dns-version.<zone>. <ttl> IN TXT <schema-version>
```

CoreDNS - kubernetes plugin
API connection options

resyncperiod DURATION	API Server resync duration (default is 5mn)
endpoint URL [URL...]	Define the url to connect to the API. If several defined, plugin will use one that is healthy if not define, use the cluster service account available on the pod
tls CERT KEY CACERT	TLS certificate for the connection to API (requires endpoint)
kubeconfig KUBECONFIG CONTEXT	authenticates the connection to a remote k8s cluster using a kubeconfig file. (requires endpoint and ignore tls)

CoreDNS - kubernetes plugin
Filtering k8s domains

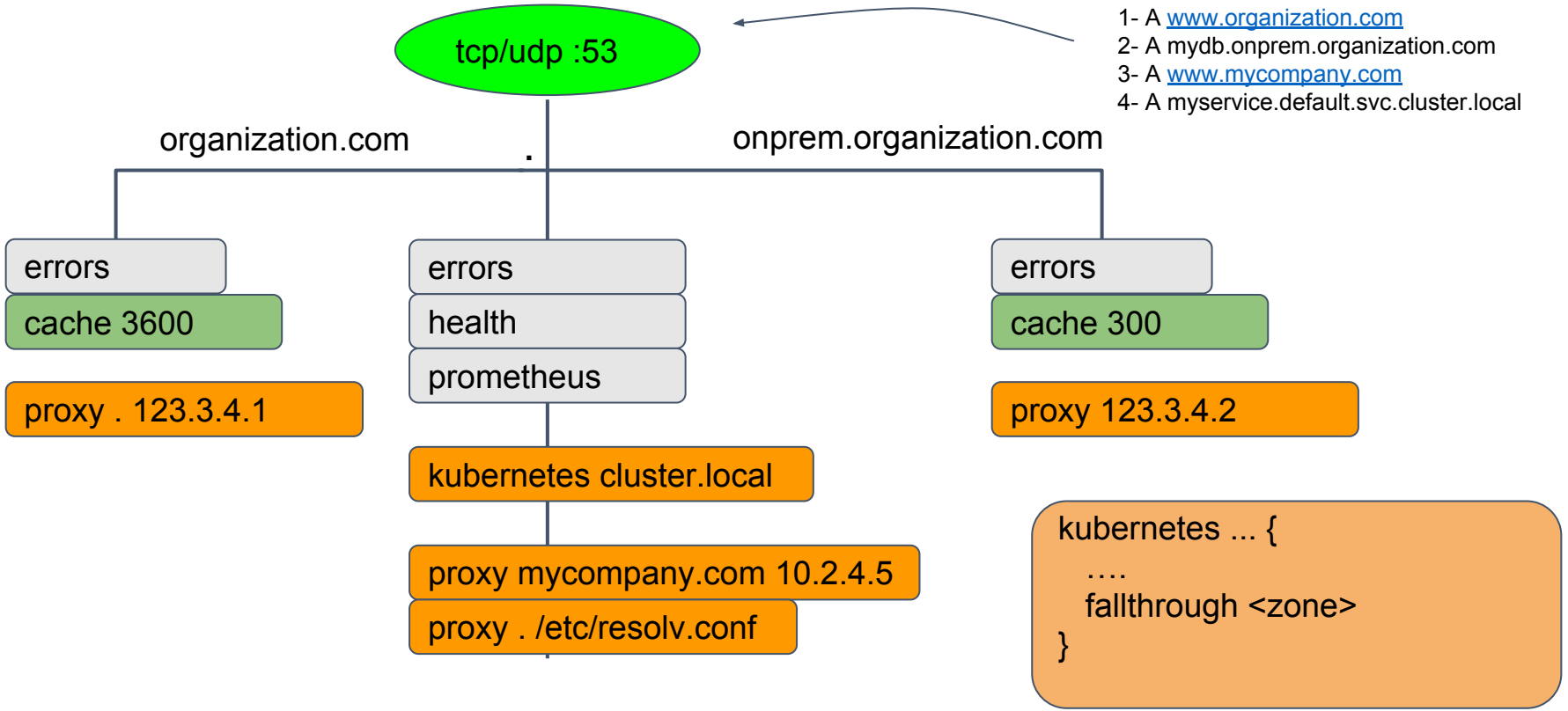
namespaces NAMESPACE...	Only the namespaces indicated will be exposed. All domains in othe domains will be replied “unknown”
labels SELECTOR-EXPRESSION	Only the objects (pod, service, endpoint) matching the selector will be exposed
ignore_empty_service	consider “unknwon” any service that have no ready pod instead of returning a valid record with no data
noendpoints	Do not watch endpoints. Any domains that includes a endpoints will be considered “unknown”. All headless service will be “unknown”
upstream [ADDRESS...]	If not present, will not resolve the external services.

These filters can be combined.
To be used in conjunction with “fallthrough” and with another way to resolve the unexposed elements

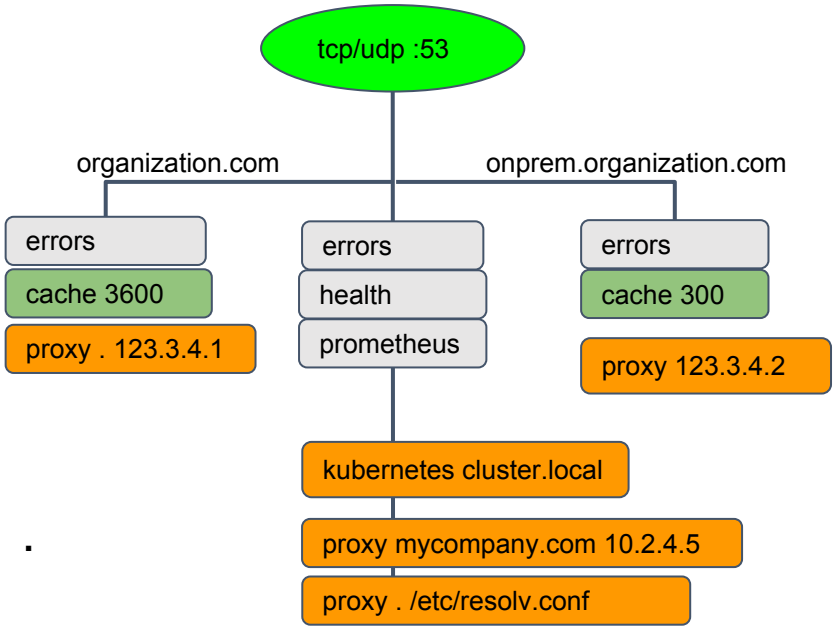
Kubernetes DNS Schema

<service>.<ns>.svc.<czone> kubernetes.default.svc.cluster.local kube-dns.kube-system.svc.cluster.local *.default.svc.cluster.local	A AAAA PTR	If ClusterIP service => the Cluster IP If Headless service => all the sub-domains as <ep>.<service>.... resolved.
<service>.<ns>.svc.<czone> **_<port>._<proto>.<service>.<ns>.svc.<czone>** _https._tcp.kubernetes.default.svc.cluster.local	SRV	=> all the port / proto supported by the service.
<ep>.<service>.<ns>.svc.<czone> <ep> is the hostname, if not and if endpoint_pod_names the name of the pod, if not the <ip-like> of the endpoint	A AAAA PTR	Only for Headless Services => the IP of the corresponding endpoint
<ip-like>.<ns>.pod.<czone> 10-96-0-65.kube-system.pod.cluster.local	A AAAA	Change the TTL of kubernetes records

Resolving a Query



Resolving a Query



1- A www.organization.com
2- A mydb.onprem.organization.com
3- A www.mycompany.com
4- A myservice.default.svc.cluster.local

```
kubernetes ... {
  ….
  fallthrough <zone>
}
```


KubeCon
CloudNativeCon
North America 2018