How We Built Magic Transit

{erich, conjones}@cloudflare.com

Agenda

- Who are we
- Who is Cloudflare
- What is Magic Transit
- Designing The Product
- How Magic Transit Works
- Questions

CLOUDFLARE

{erich, conjones}@cloudflare.com

Who Are We?

{erich, conjones}@cloudflare.com

Erich Heine
Systems Engineer
erich@cloudflare.com

Connor Jones
Systems Engineer
conjones@cloudflare.com

CLOUDFLARE

{erich, conjones}@cloudflare.com

Who is Cloudflare?

CLOUDFLARE

{erich, conjones}@cloudflare.com

“Helping build a better internet”

CLOUDFLARE

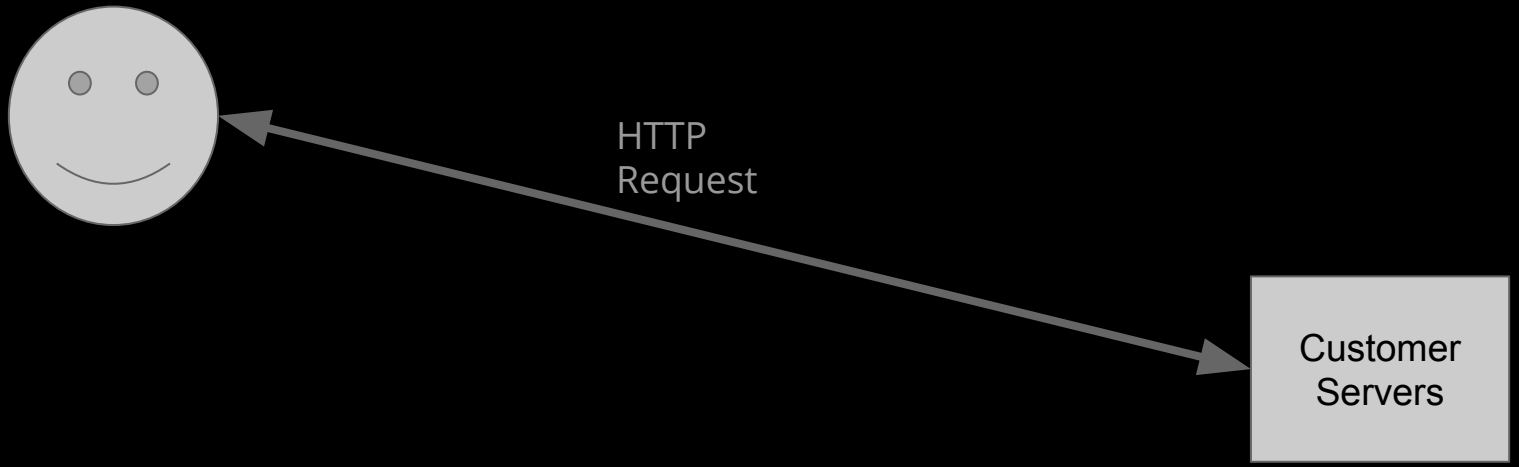
{erich, conjones}@cloudflare.com

But How???

- GLOBAL Anycast network
- Provide security & performance services:
  - DDoS
  - DNS
  - Spectrum
  - CDN, WAF, Workers
  - ...

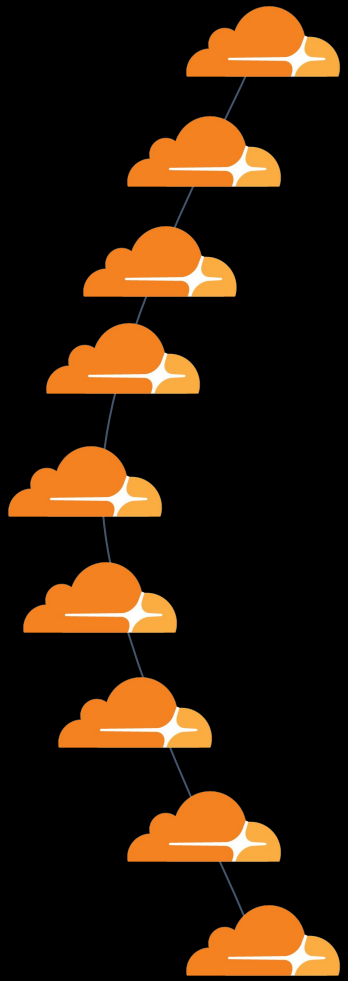
CLOUDFLARE

{erich, conjones}@cloudflare.com



CLOUDFLARE

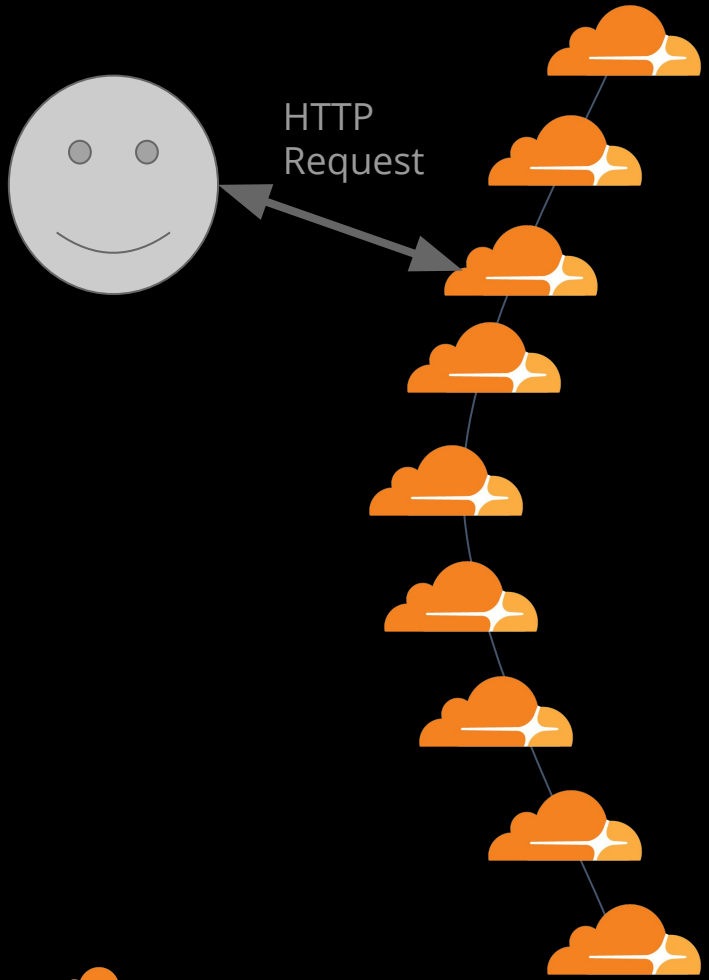
{erich, conjones}@cloudflare.com



Customer Servers

CLOUDFLARE

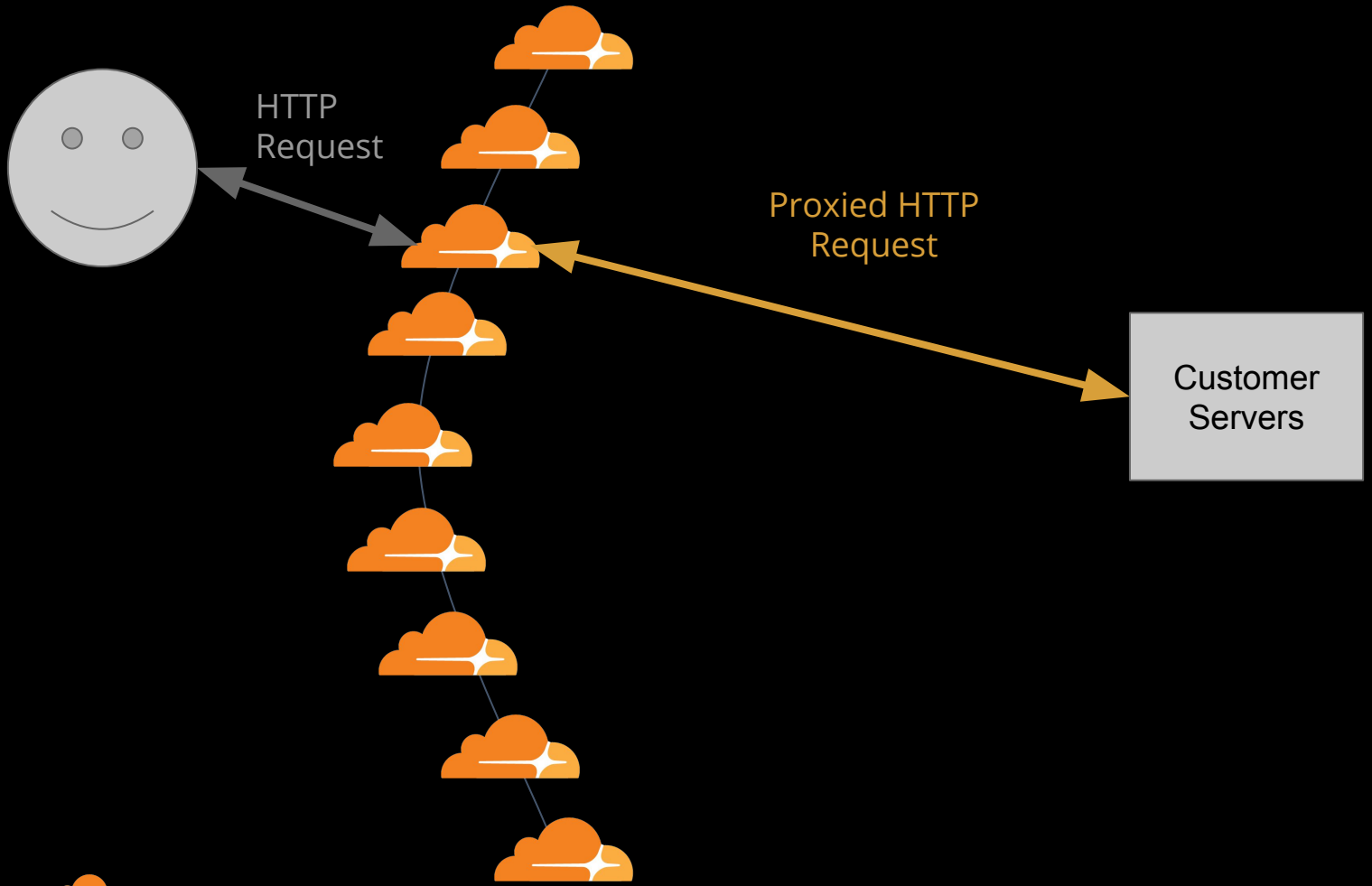
{erich, conjones}@cloudflare.com



Customer
Servers

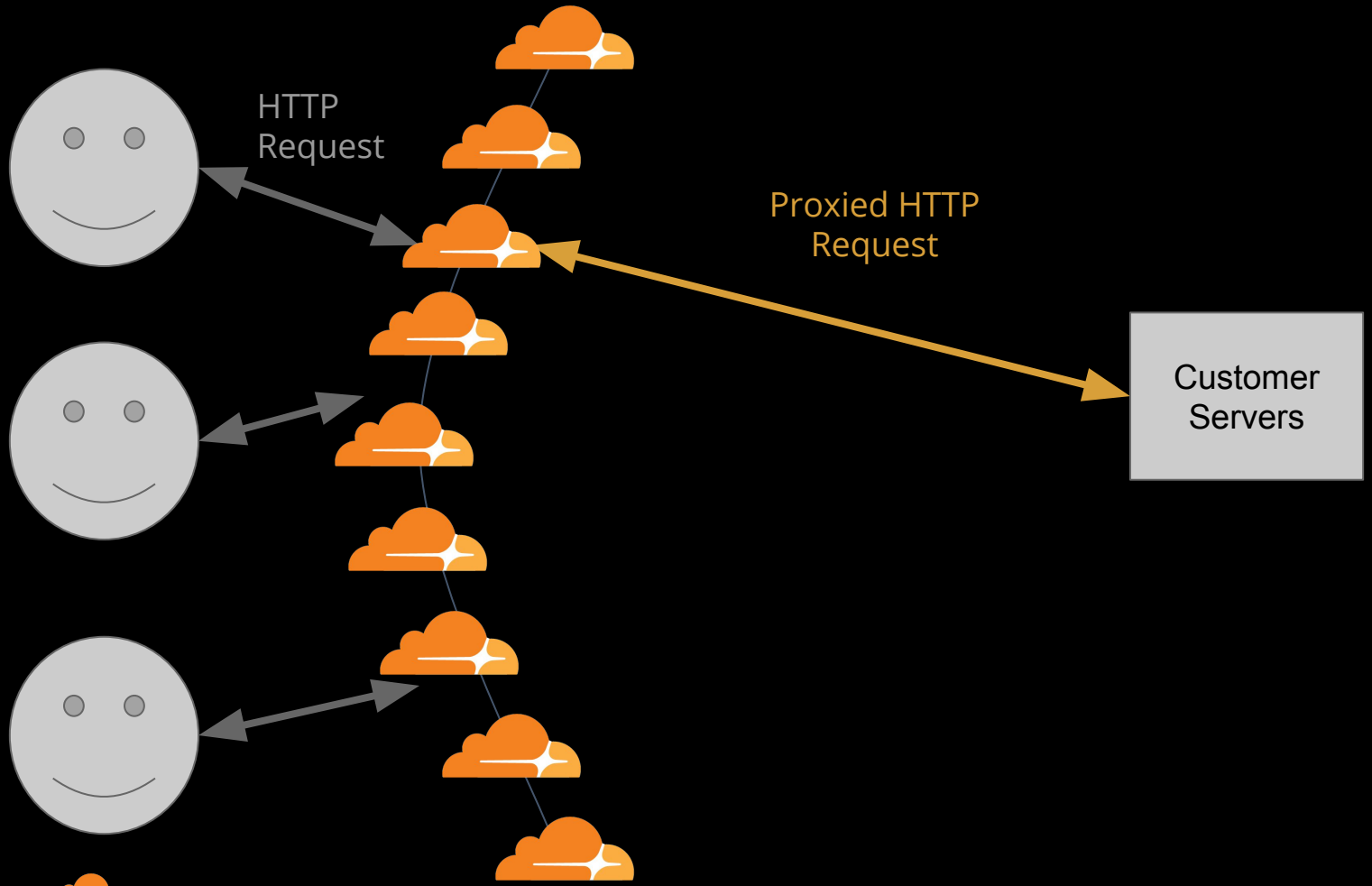
CLOUDFLARE

{erich, conjones}@cloudflare.com



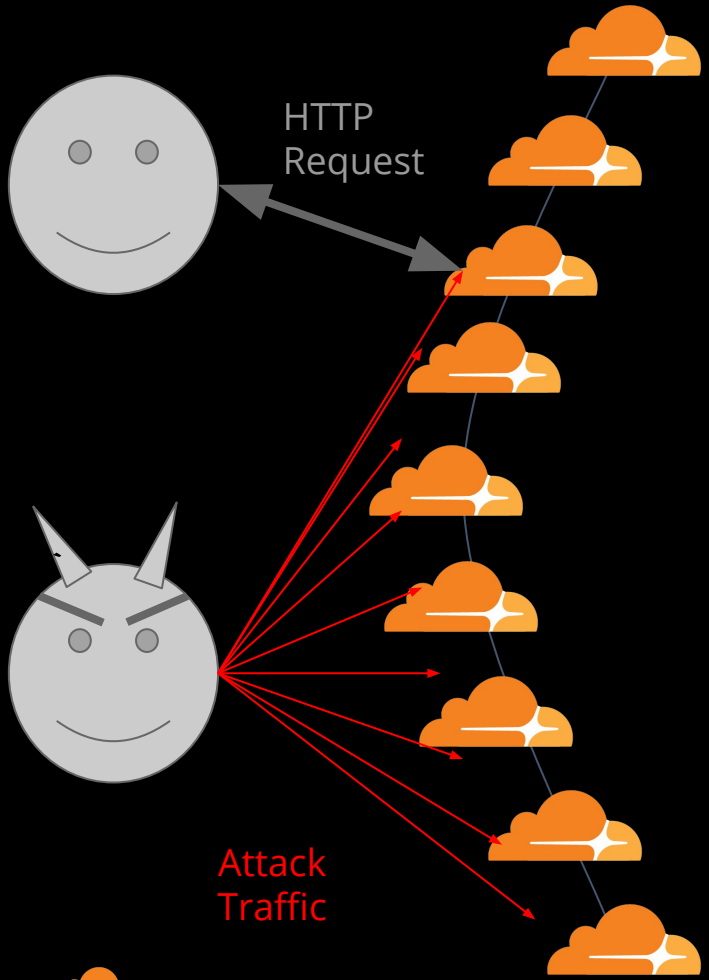
CLOUDFLARE

{erich, conjones}@cloudflare.com



CLOUDFLARE

{erich, conjones}@cloudflare.com

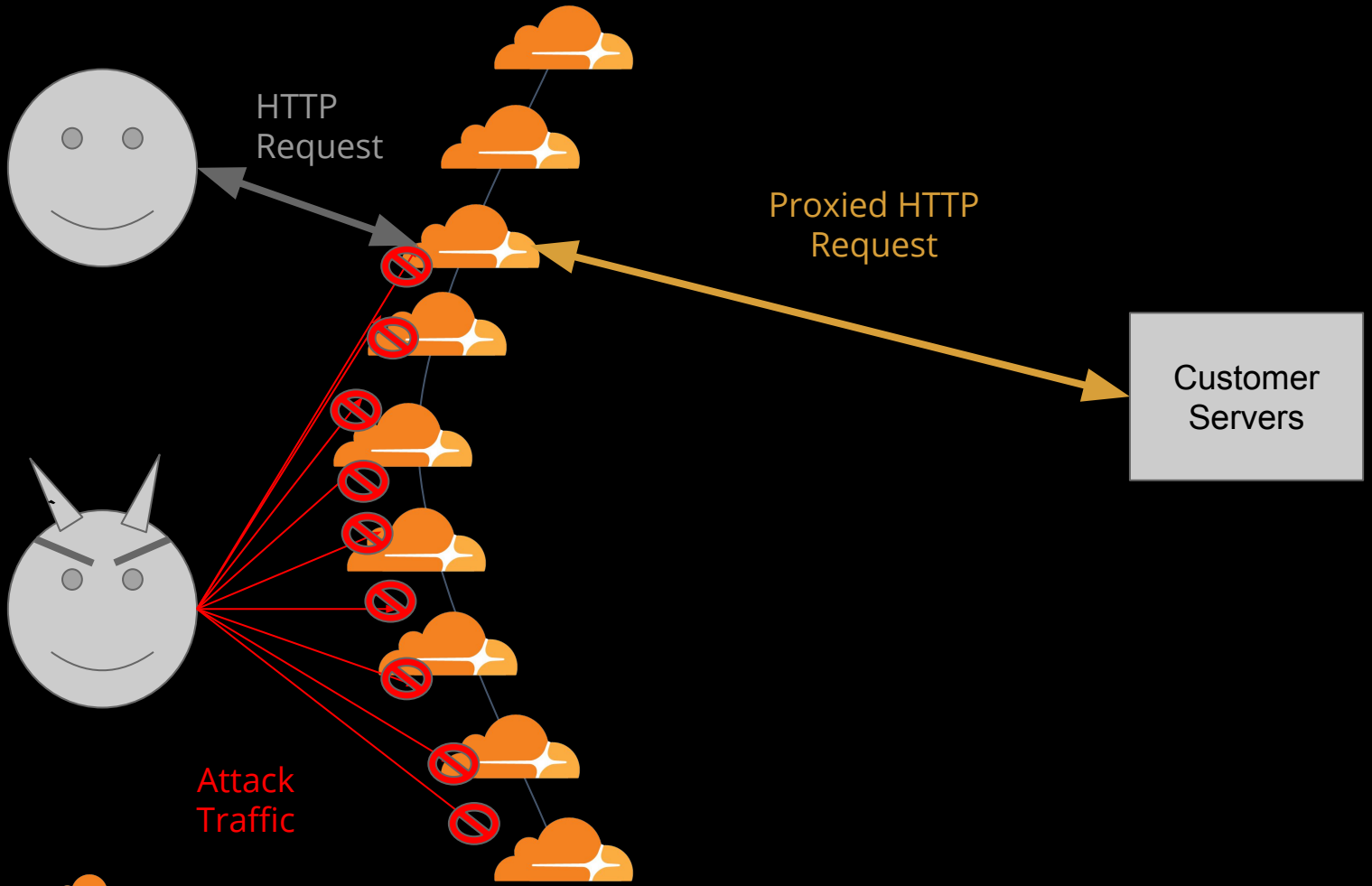


Proxied HTTP
Request

Customer
Servers

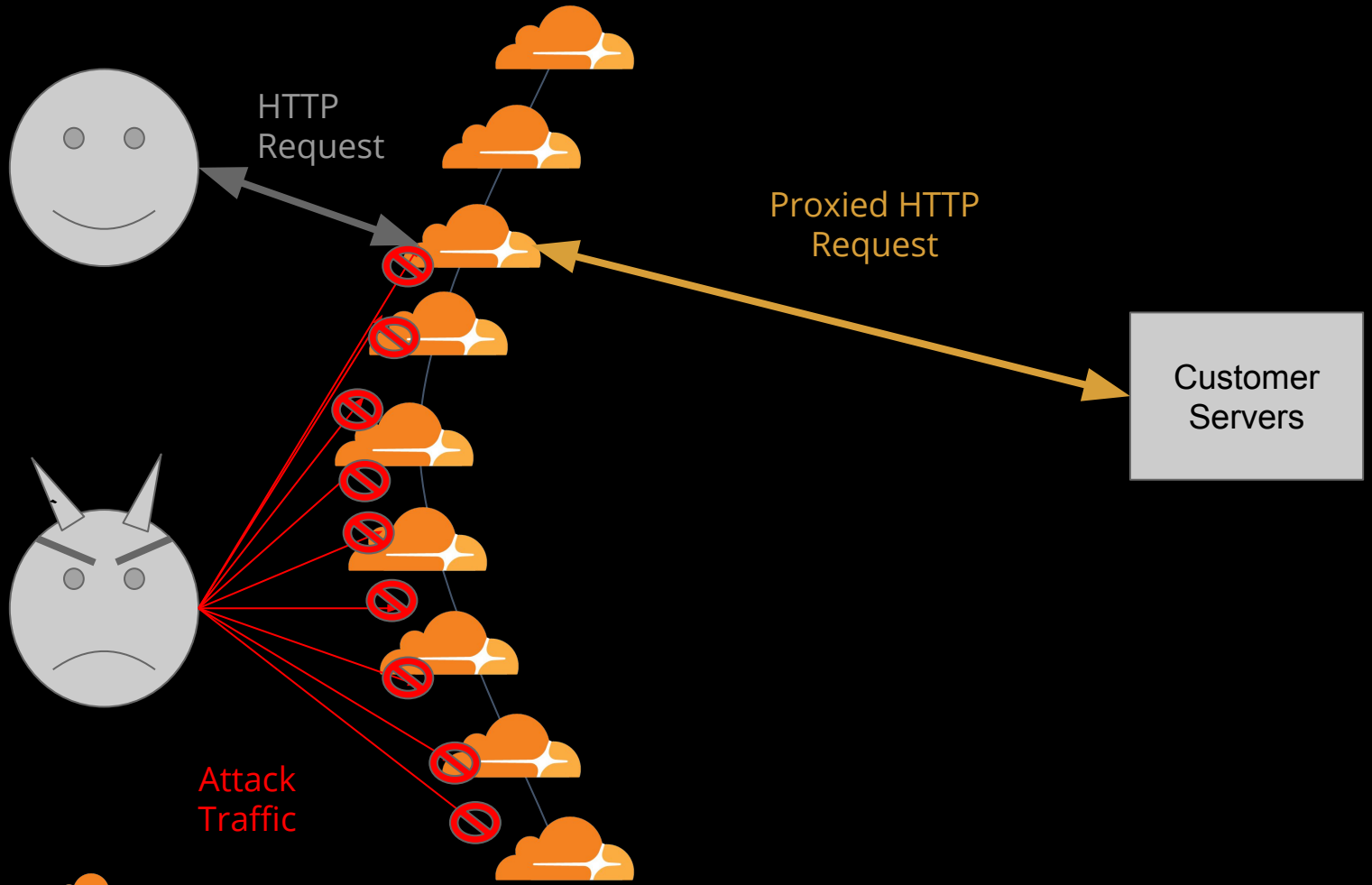
CLOUDFLARE

{erich, conjones}@cloudflare.com



CLOUDFLARE

{erich, conjones}@cloudflare.com



CLOUDFLARE

{erich, conjones}@cloudflare.com

Customers to Cloudflare:

Can you do this with all my traffic?

{erich, conjones}@cloudflare.com

But why?

- Consolidation of vendors
- Network functions in the network (not the boxes)
- Proxying is not always the answer

{erich, conjones}@cloudflare.com

Cloudflare to customers:

- Much of this operates at L3 already
- We have a big network
- We'll build you a Magic Transit

CLOUDFLARE

{erich, conjones}@cloudflare.com

What is Magic Transit?

{erich, conjones}@cloudflare.com

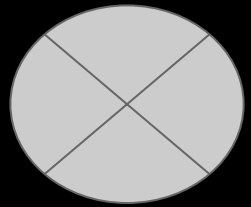
Magic Transit

- Network functions as a service
- L3 DDoS scrubbing
- Firewall

By:

- Existing DDoS solution
- Advertises customer prefixes on BGP

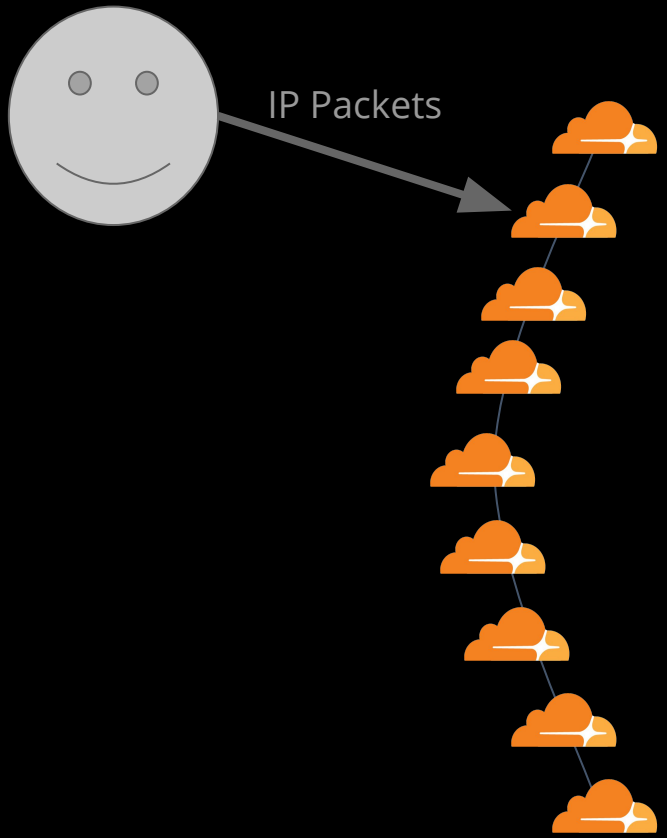
{erich, conjones}@cloudflare.com

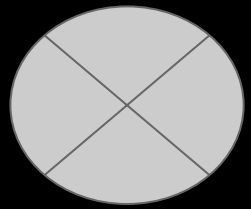


Customer
Router

CLOUDFLARE

{erich, conjones}@cloudflare.com

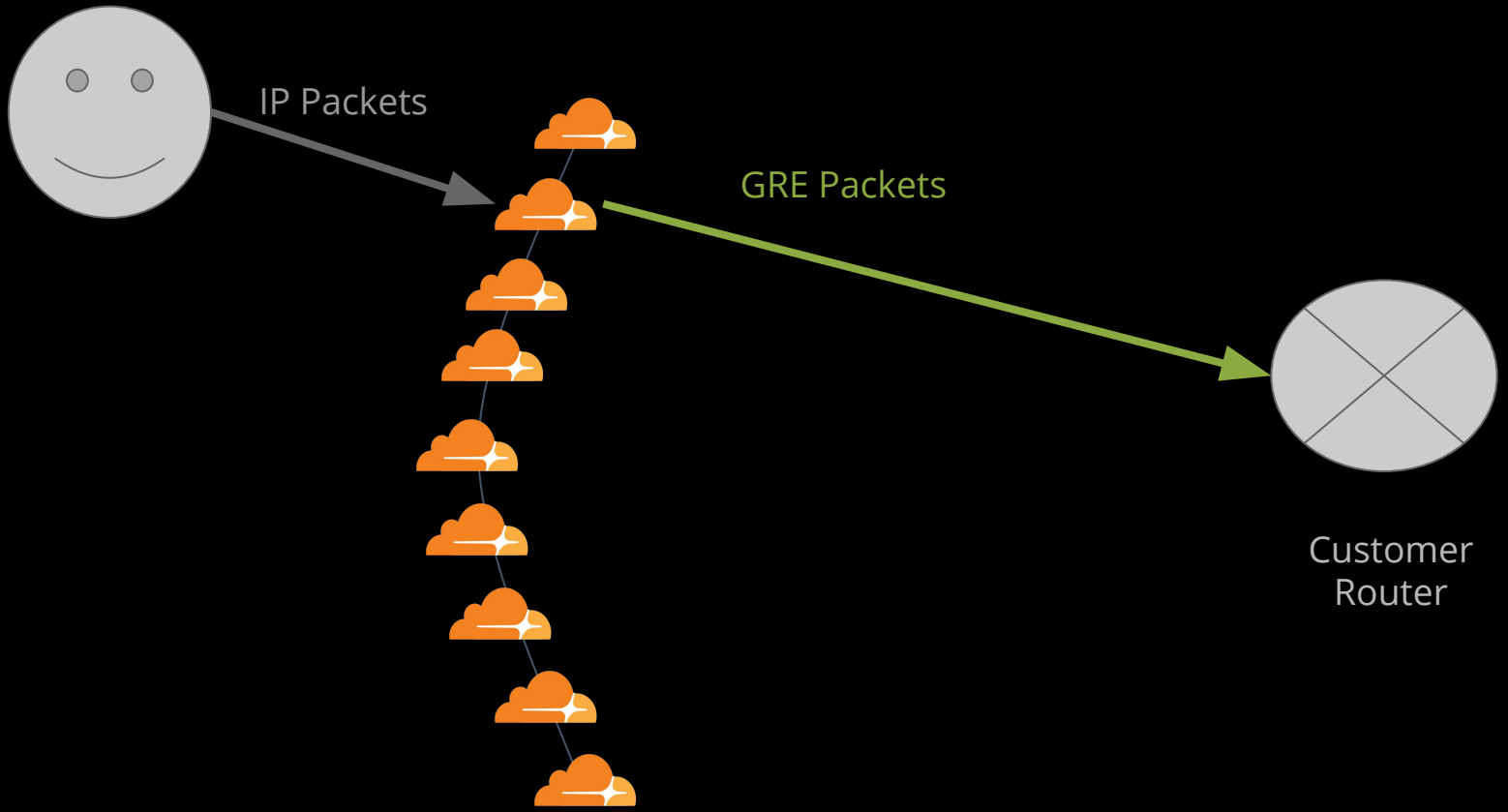




Customer
Router

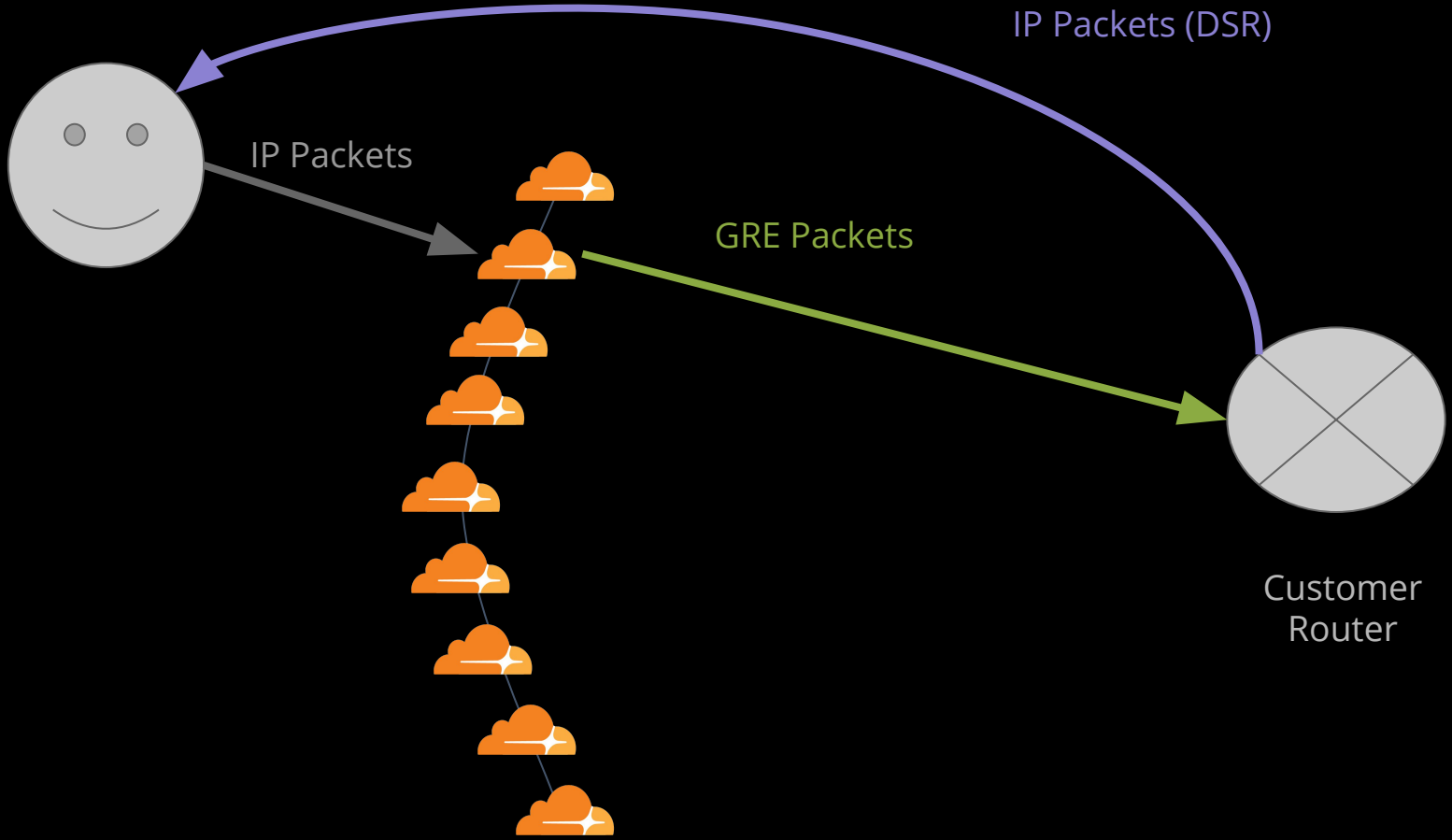
CLOUDFLARE

{erich, conjones}@cloudflare.com



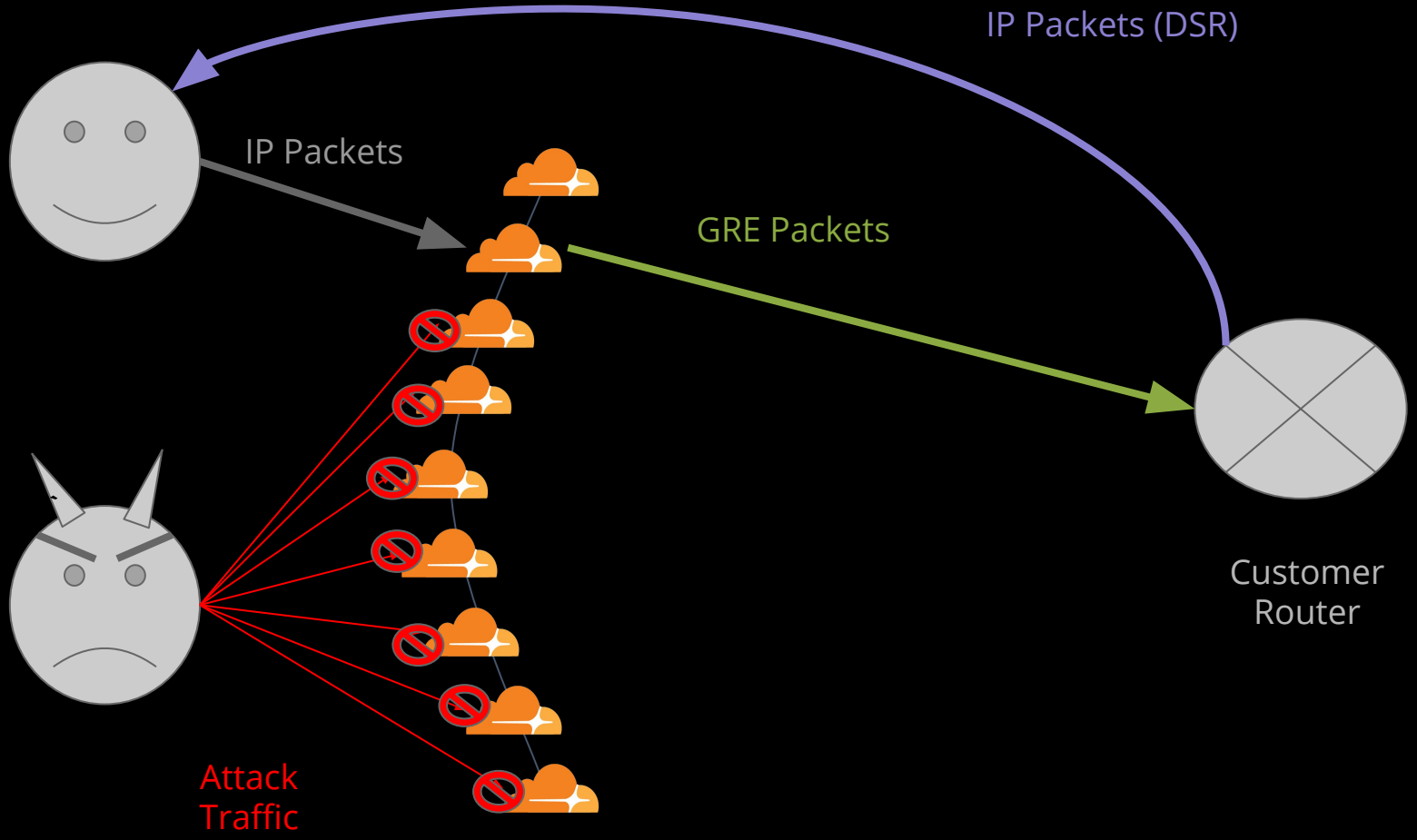
CLOUDFLARE

{erich, conjones}@cloudflare.com



CLOUDFLARE

{erich, conjones}@cloudflare.com



CLOUDFLARE

{erich, conjones}@cloudflare.com

Designing the product

Proof of concept

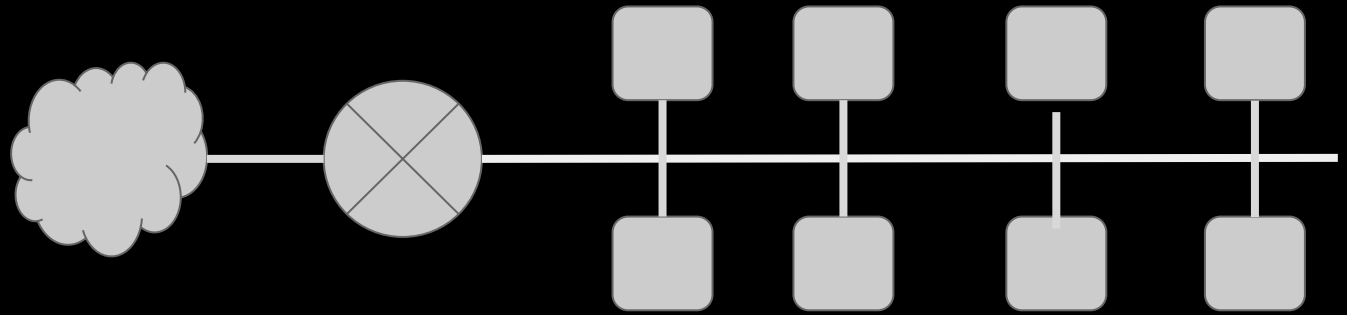
{erich, conjones}@cloudflare.com

Will it even work?

- Does GRE actually work that way?
- Does it scale?
- How do we fit it in our platform?

{erich, conjones}@cloudflare.com

A Cloudflare POP



CLOUDFLARE

{erich, conjones}@cloudflare.com

Challenges

- The metal is a router
- Uniform servers
- Customer isolation
- Security
- Timeline + skillset

{erich, conjones}@cloudflare.com

Technology options

- VRFs
- XDP
- Namespaces

{erich, conjones}@cloudflare.com

VRF

Pros:

- Routing isolation
- Stays at L3

Cons:

- Lacking expertise
- Unclear iptables interaction

{erich, conjones}@cloudflare.com

XDP

Pros:

- Bypass a lot of processing
- Lots of expertise
- “Just Code”

Cons:

- It’s a lot of code
- Time investment

{erich, conjones}@cloudflare.com

Namespaces

Pros:

- Isolation
- Expertise
- Bash!

Cons:

- Lots of veths
- Complex setup

{erich, conjones}@cloudflare.com

And the winner is...

CLOUDFLARE

{erich, conjones}@cloudflare.com

Namespaces

- Proof of concept in a few weeks
- Surprisingly small shell script
- Dataplane performance is pretty good

{erich, conjones}@cloudflare.com

Namespaces

- Proof of concept in a few weeks
- Surprisingly small shell script
- **Dataplane performance is pretty good**

We can use this for the product!

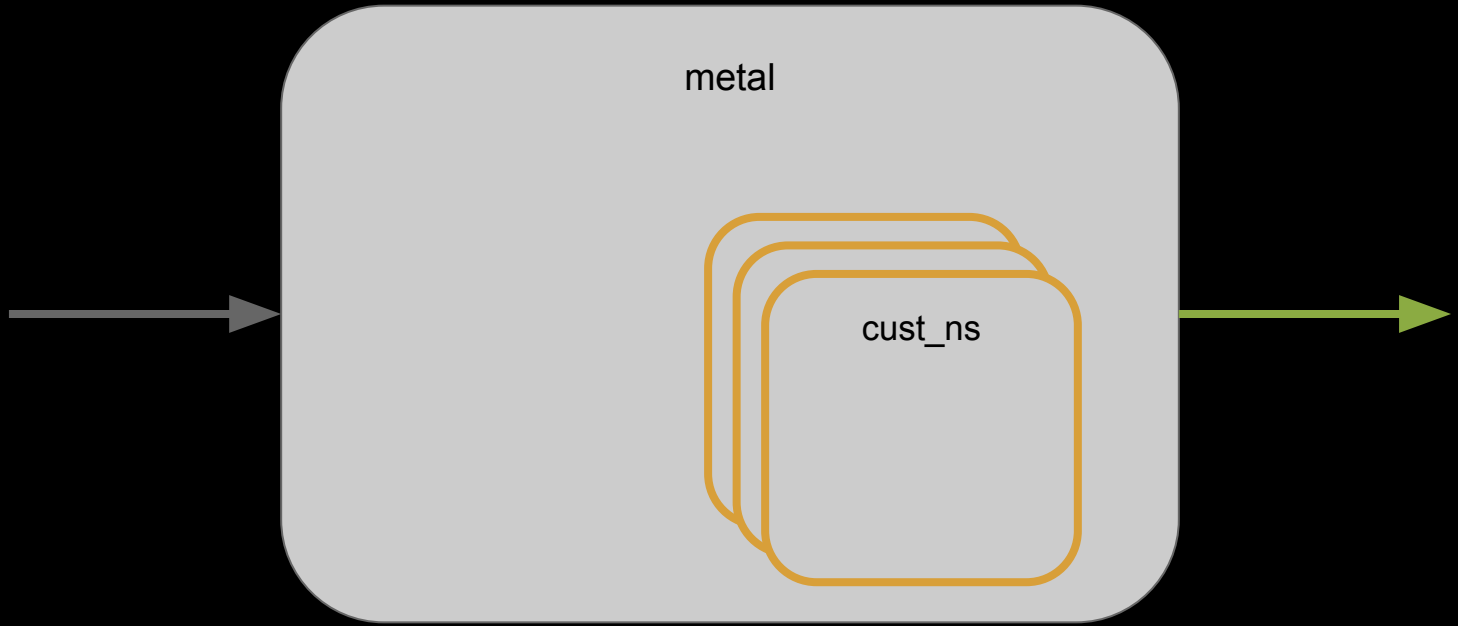
{erich, conjones}@cloudflare.com

Designing the product

The actual product

{erich, conjones}@cloudflare.com

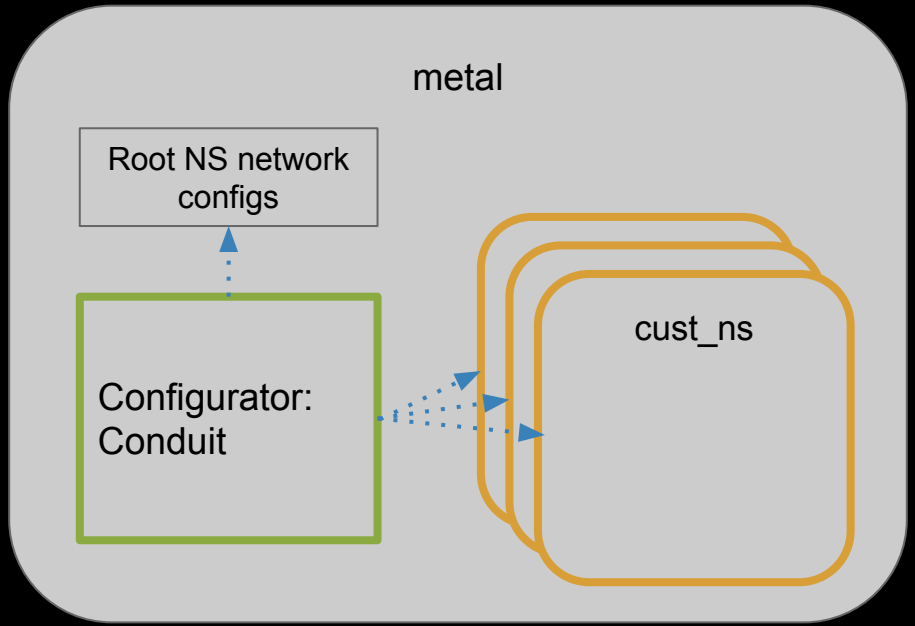
One namespace per customer



CLOUDFLARE

{erich, conjones}@cloudflare.com

Configuration daemon



CLOUDFLARE

{erich, conjones}@cloudflare.com

Configurator: Conduit

- Daemon written in Go
- Uses QS as config distribution
- Can crash without (big) consequence

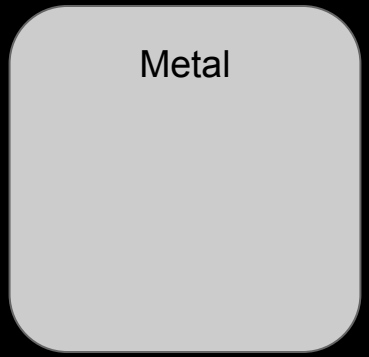
{erich, conjones}@cloudflare.com

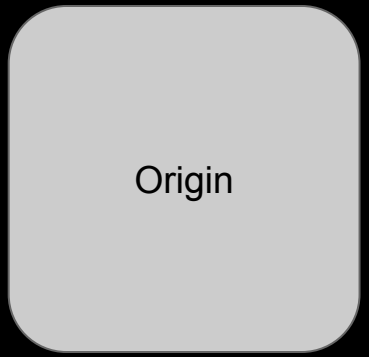
Behind The Curtain Of Magic Transit

Shuffling packets from eyeballs to origins

{erich, conjones}@cloudflare.com

How Packets Get To The Origin





CLOUDFLARE

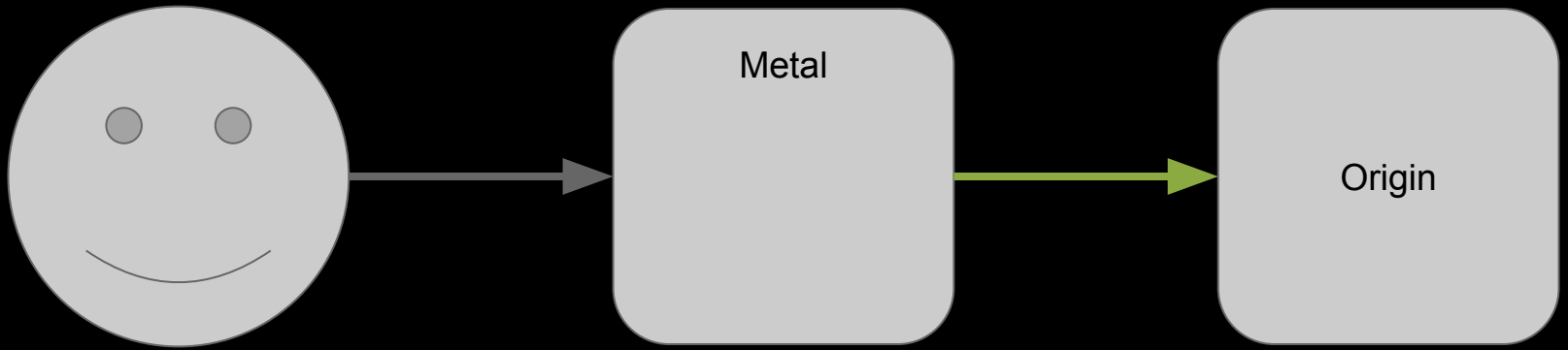
{erich, conjones}@cloudflare.com

Packets Arrive At The Edge

CLOUDFLARE

{erich, conjones}@cloudflare.com

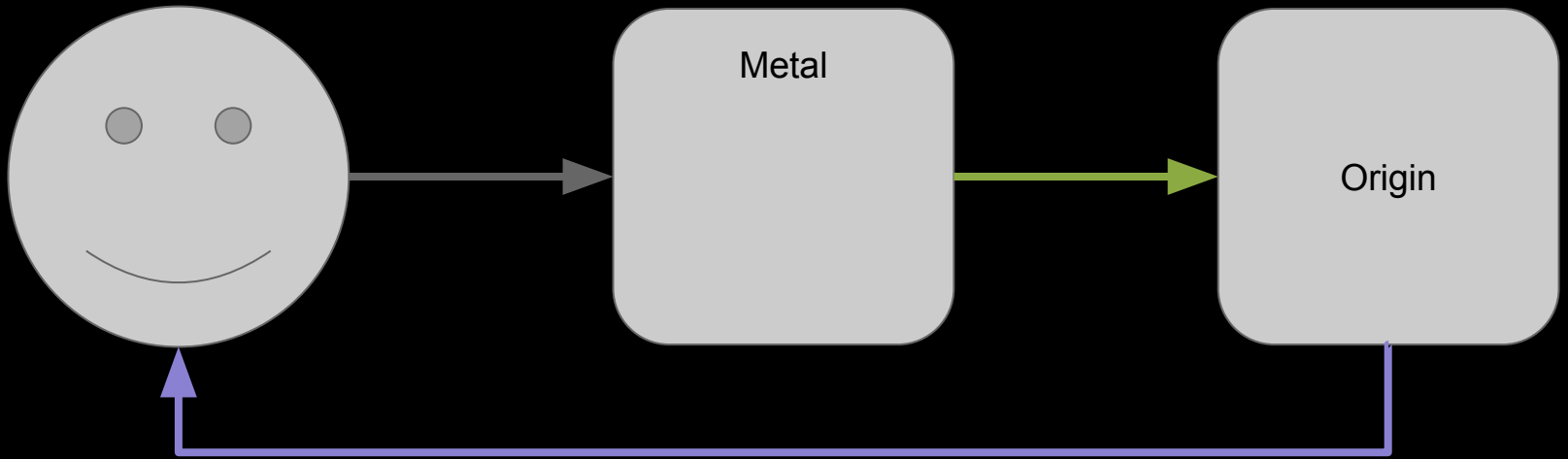
Magic Happens



CLOUDFLARE

{erich, conjones}@cloudflare.com

Origin Responds to Eyeball



CLOUDFLARE

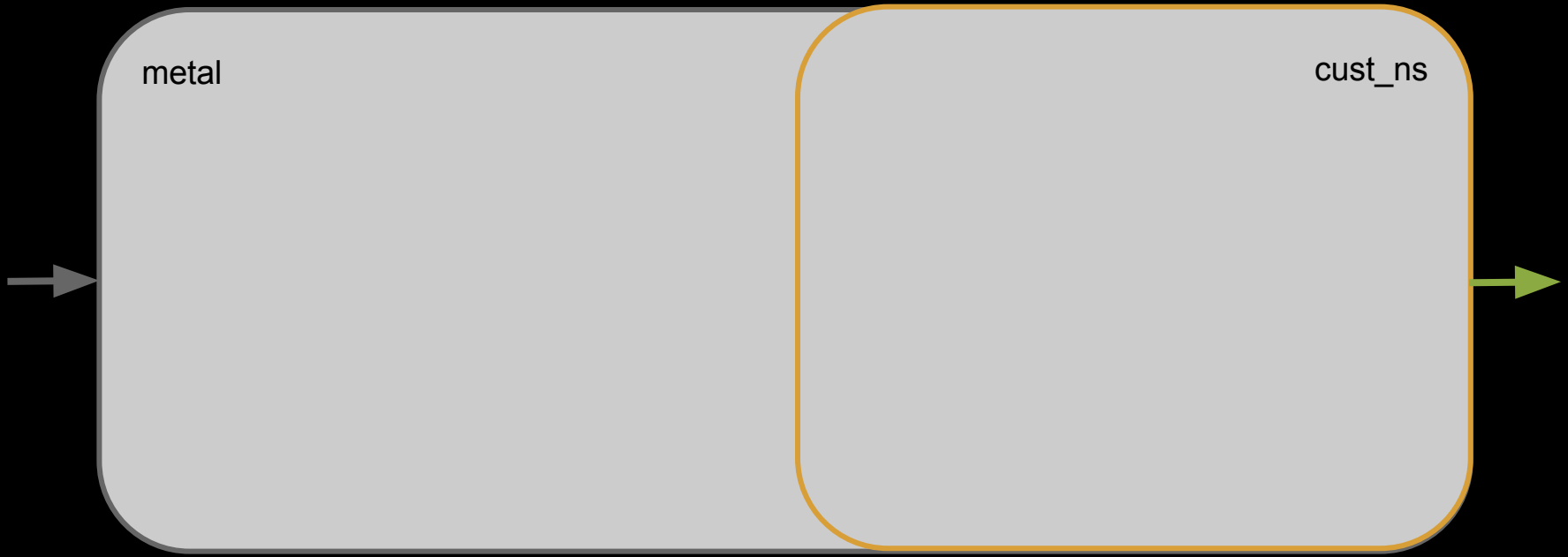
{erich, conjones}@cloudflare.com

An Edge Metal

CLOUDFLARE

{erich, conjones}@cloudflare.com

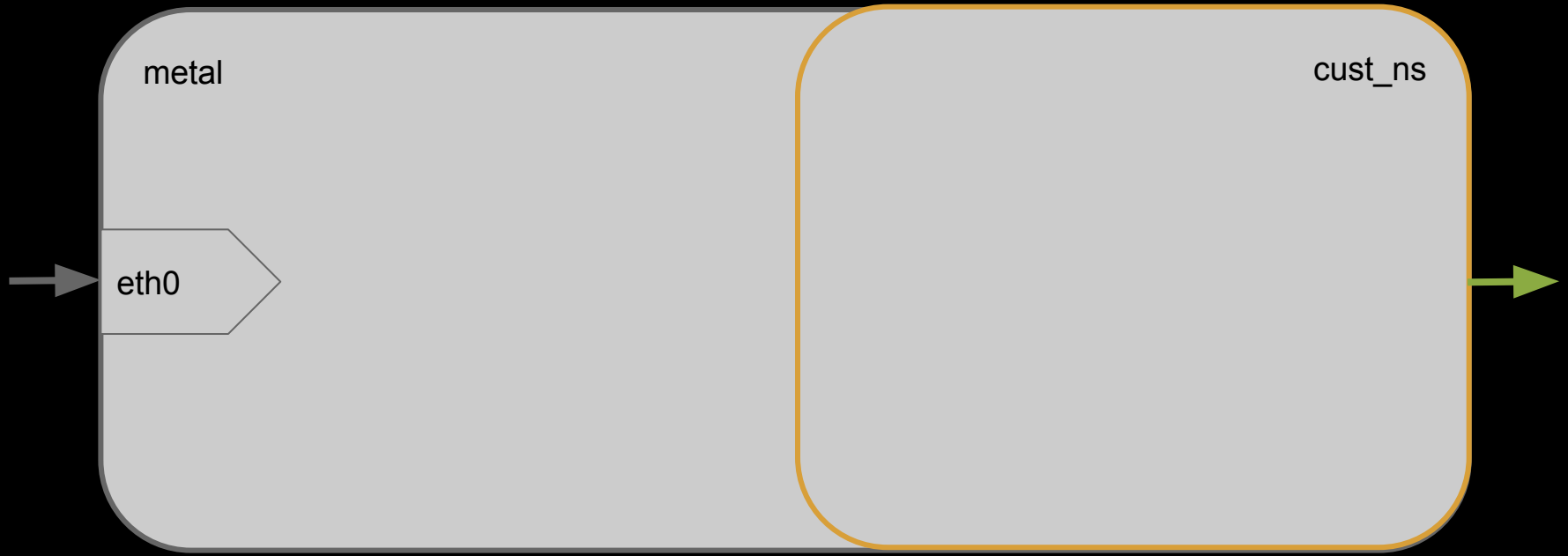
One Namespace Per Customer



CLOUDFLARE

{erich, conjones}@cloudflare.com

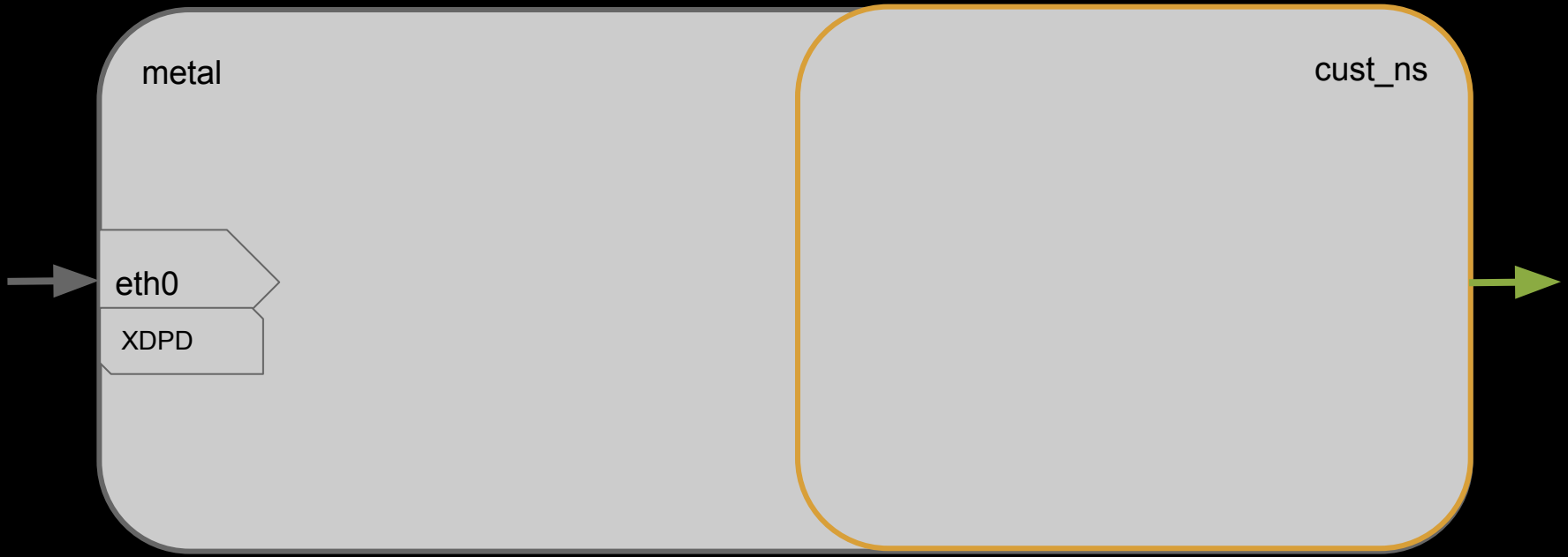
Packets Arrive



CLOUDFLARE

{erich, conjones}@cloudflare.com

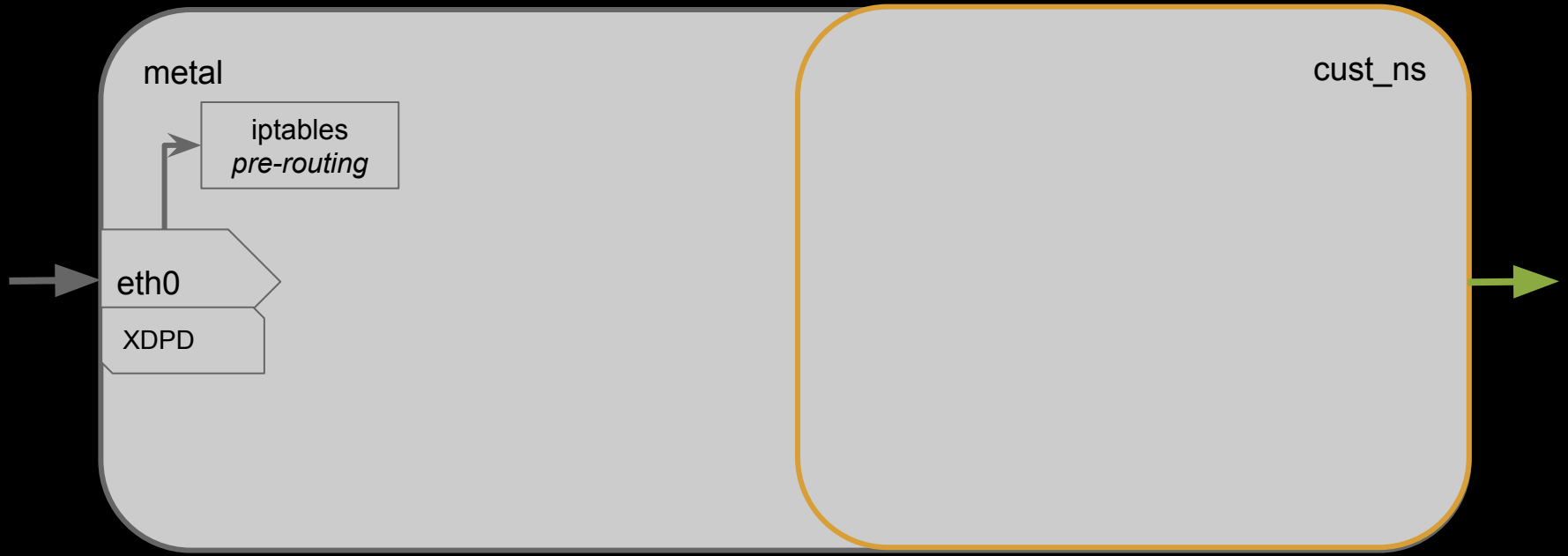
A Lil XDP



CLOUDFLARE

{erich, conjones}@cloudflare.com

Marking Packets



CLOUDFLARE

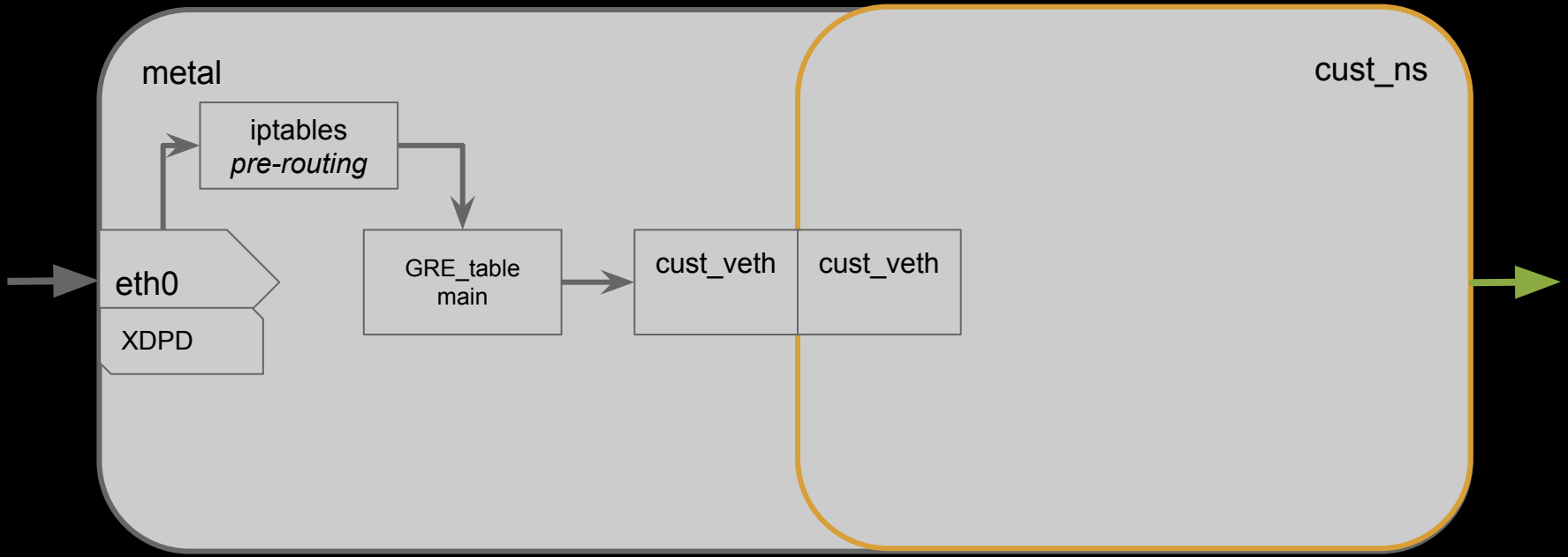
{erich, conjones}@cloudflare.com

Route Tables with an S

CLOUDFLARE

{erich, conjones}@cloudflare.com

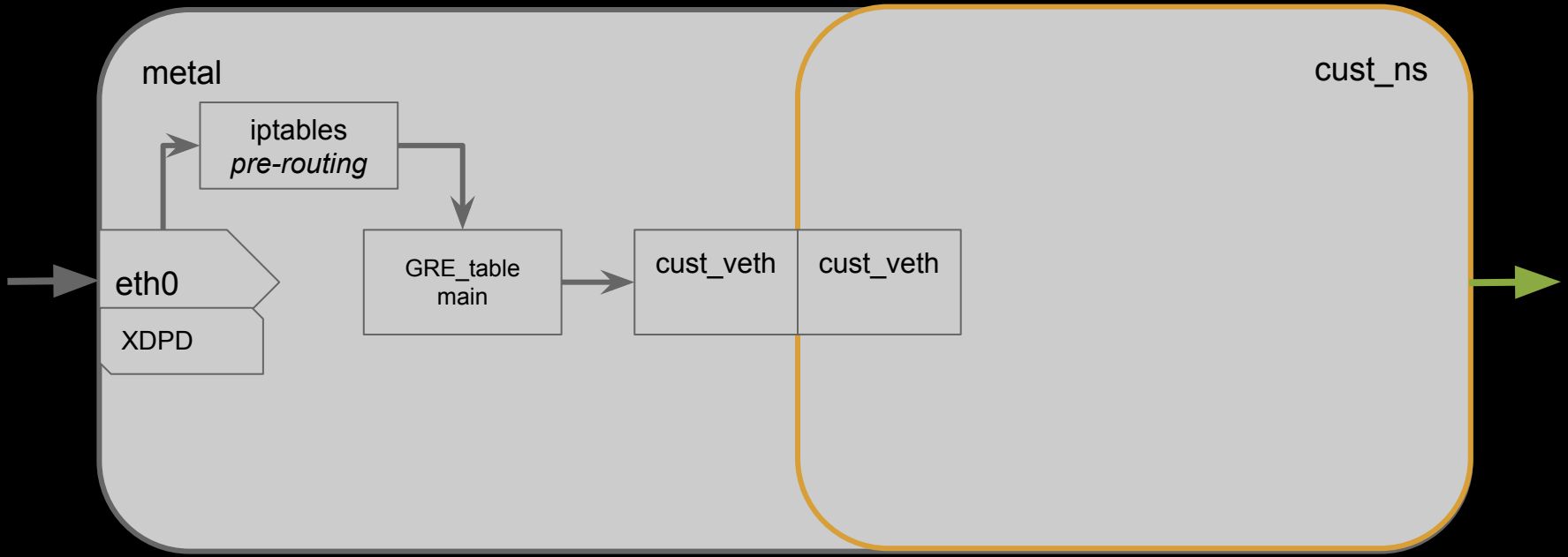
Getting Traffic Into The Namespace



CLOUDFLARE

{erich, conjones}@cloudflare.com

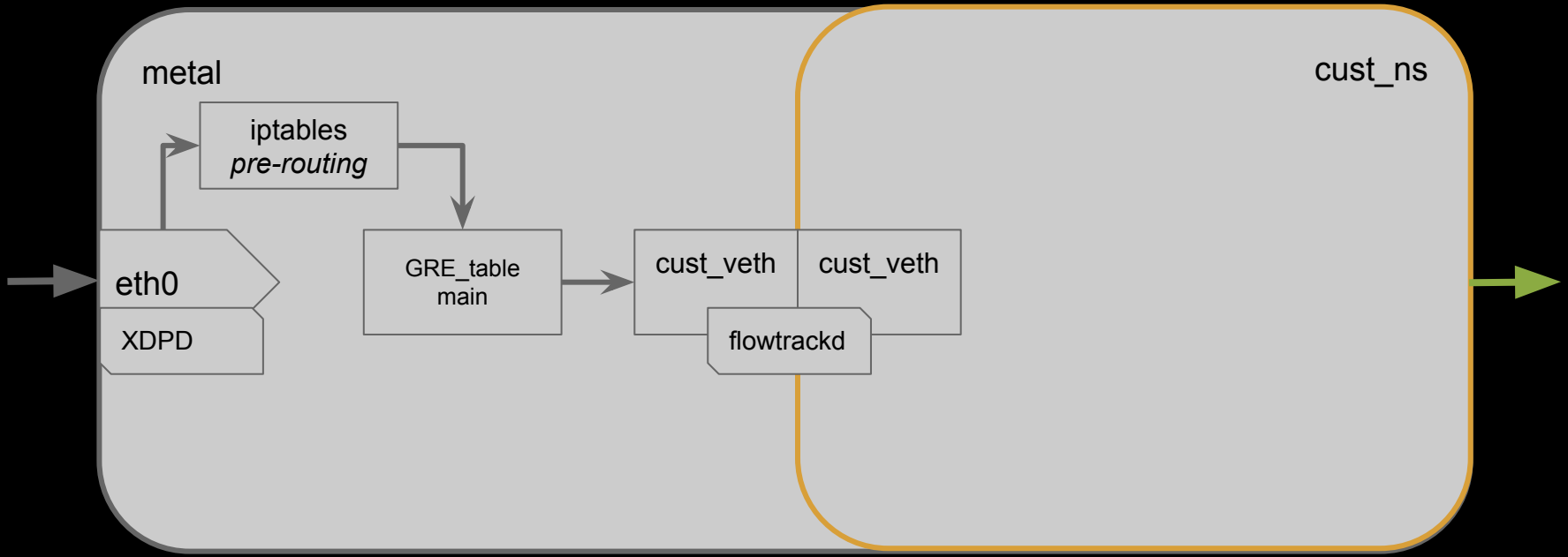
veth IPs...aaand its gone



CLOUDFLARE

{erich, conjones}@cloudflare.com

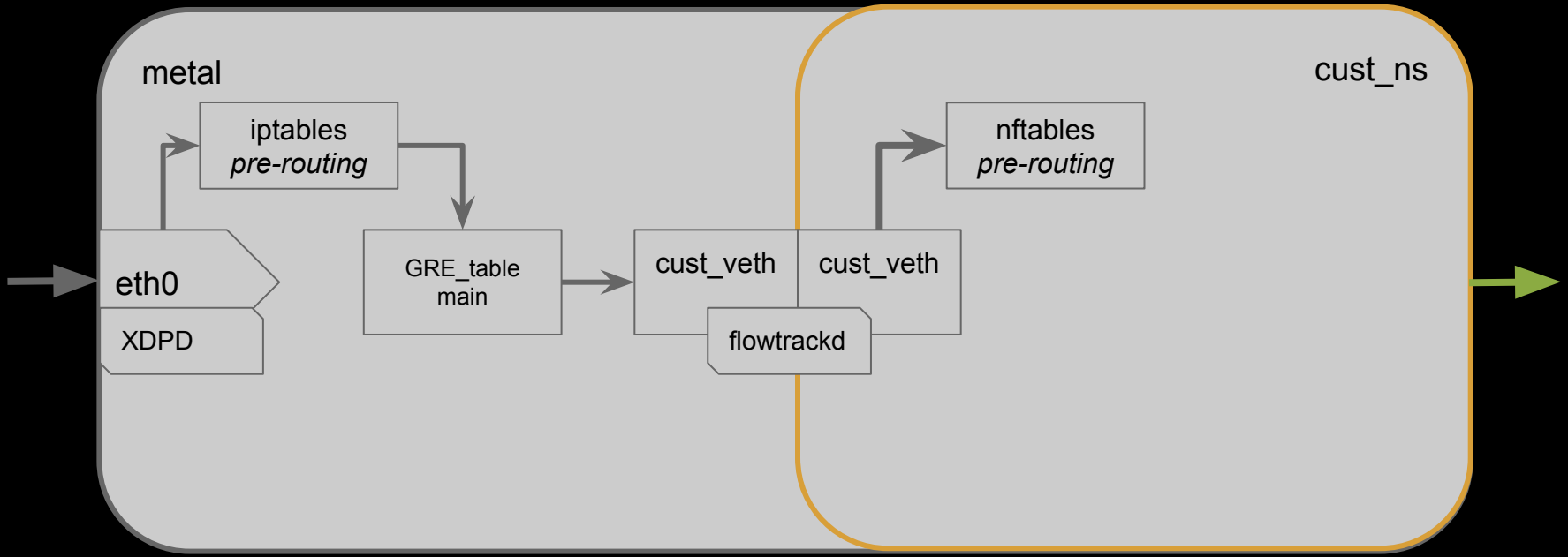
Connection tracking with half the connection, with AF_XDP



CLOUDFLARE

{erich, conjones}@cloudflare.com

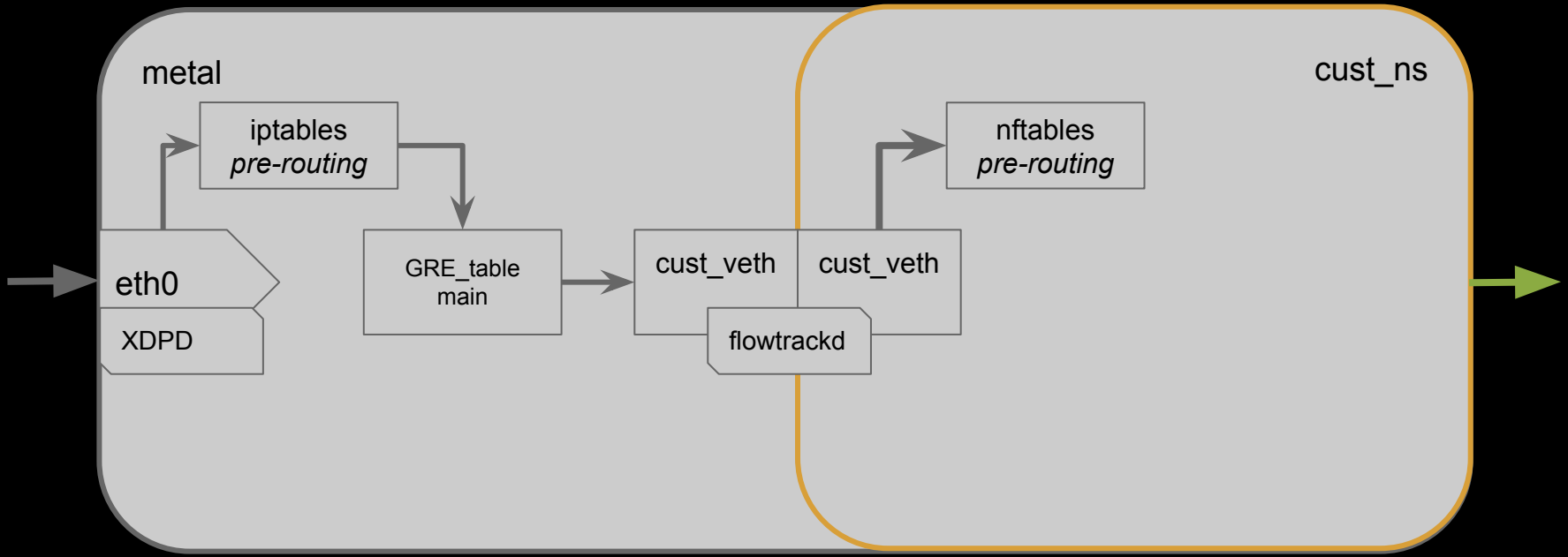
Your Traffic, Your FW Rules



CLOUDFLARE

{erich, conjones}@cloudflare.com

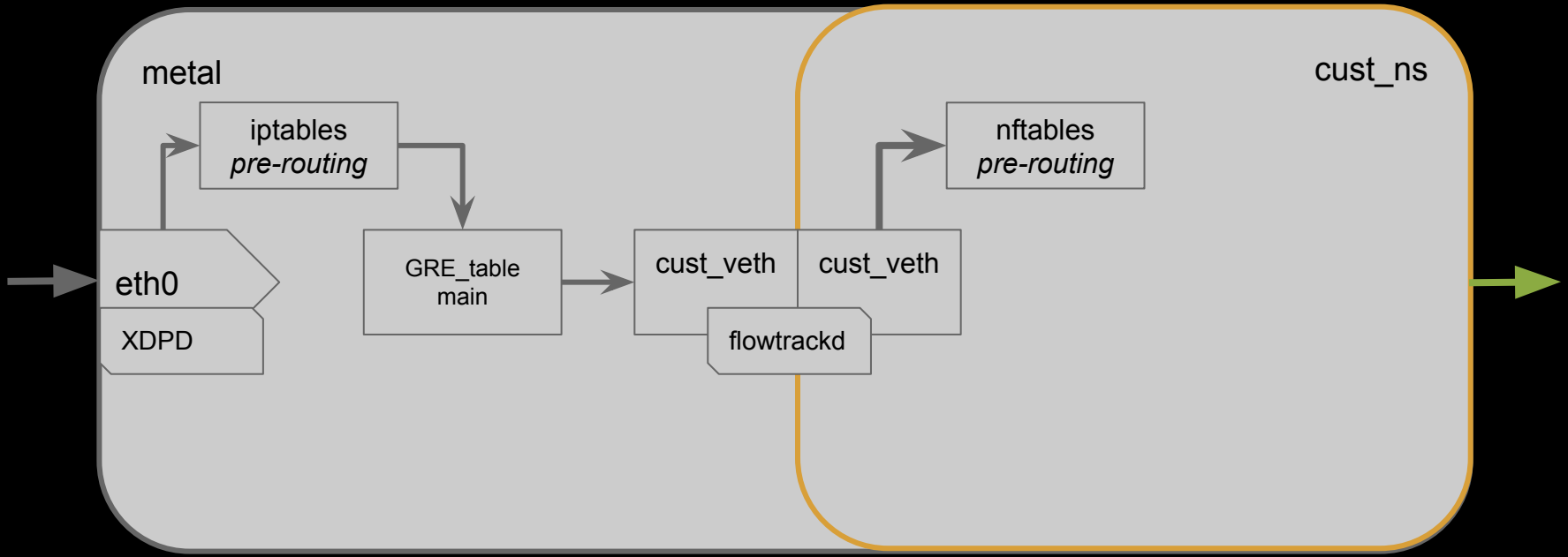
Dummy nft rule + conntrack



CLOUDFLARE

{erich, conjones}@cloudflare.com

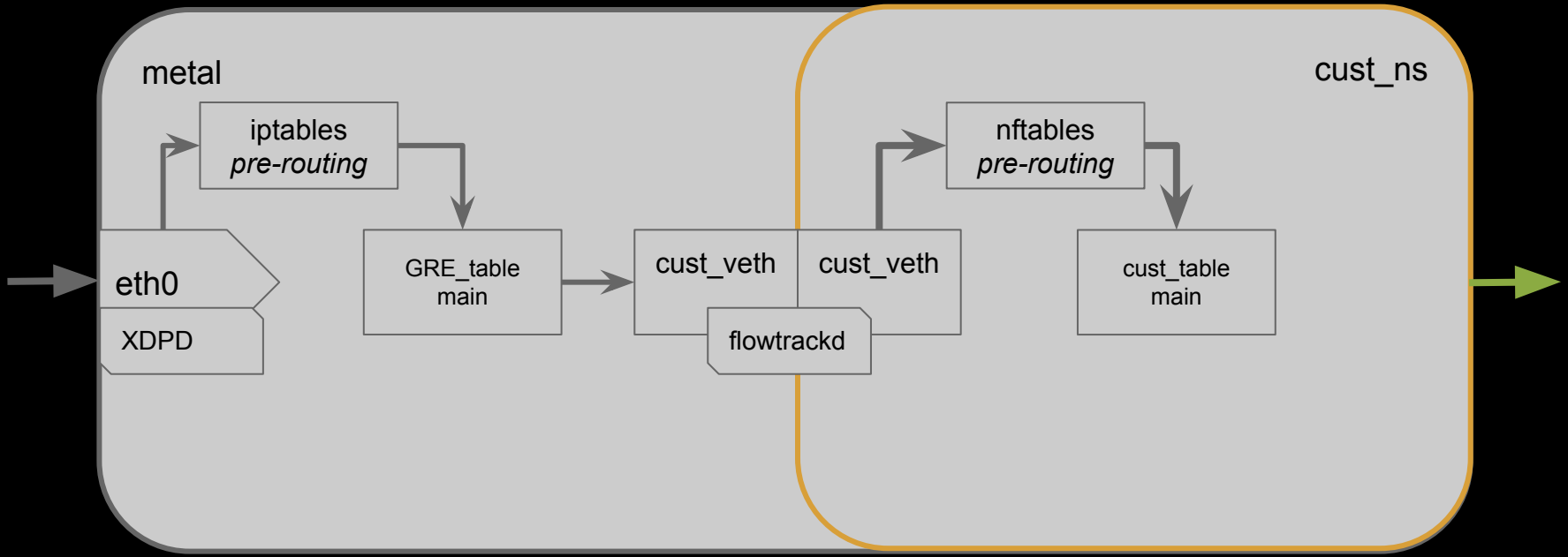
Dummy nft rule + notrack



CLOUDFLARE

{erich, conjones}@cloudflare.com

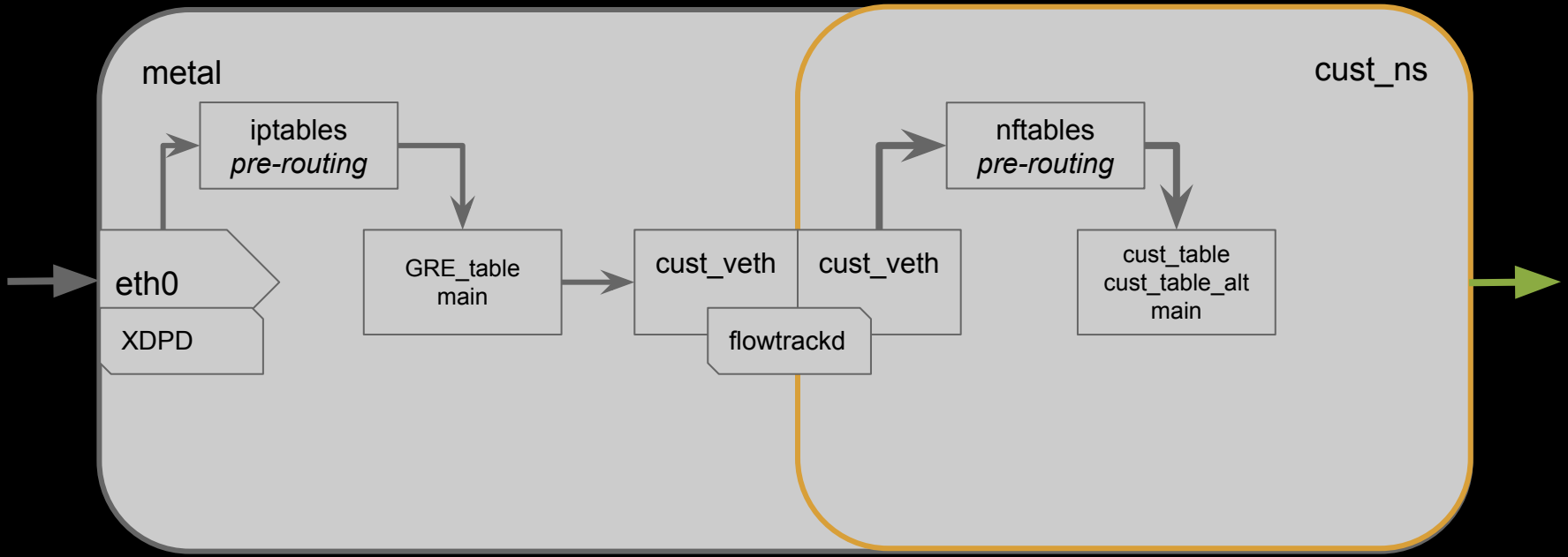
Route Tables With an S, The Second, pt 1



CLOUDFLARE

{erich, conjones}@cloudflare.com

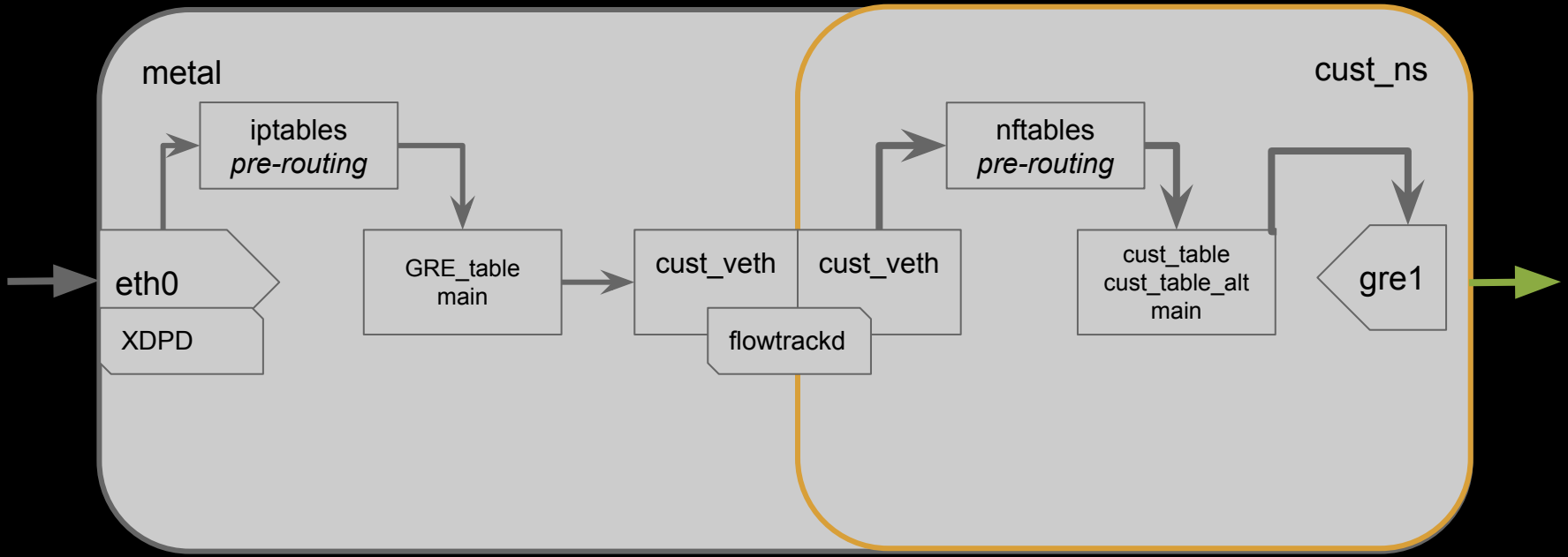
Route Tables With an S, The Second, pt 2



CLOUDFLARE

{erich, conjones}@cloudflare.com

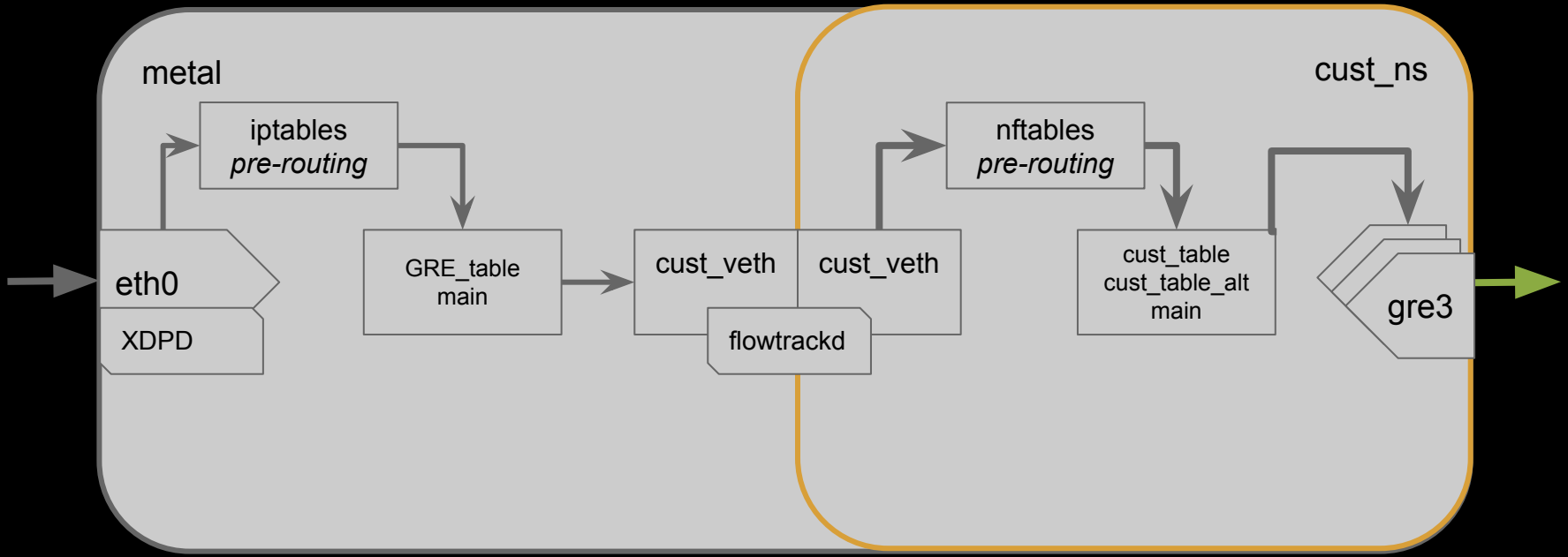
Wrapped With A Bow



CLOUDFLARE

{erich, conjones}@cloudflare.com

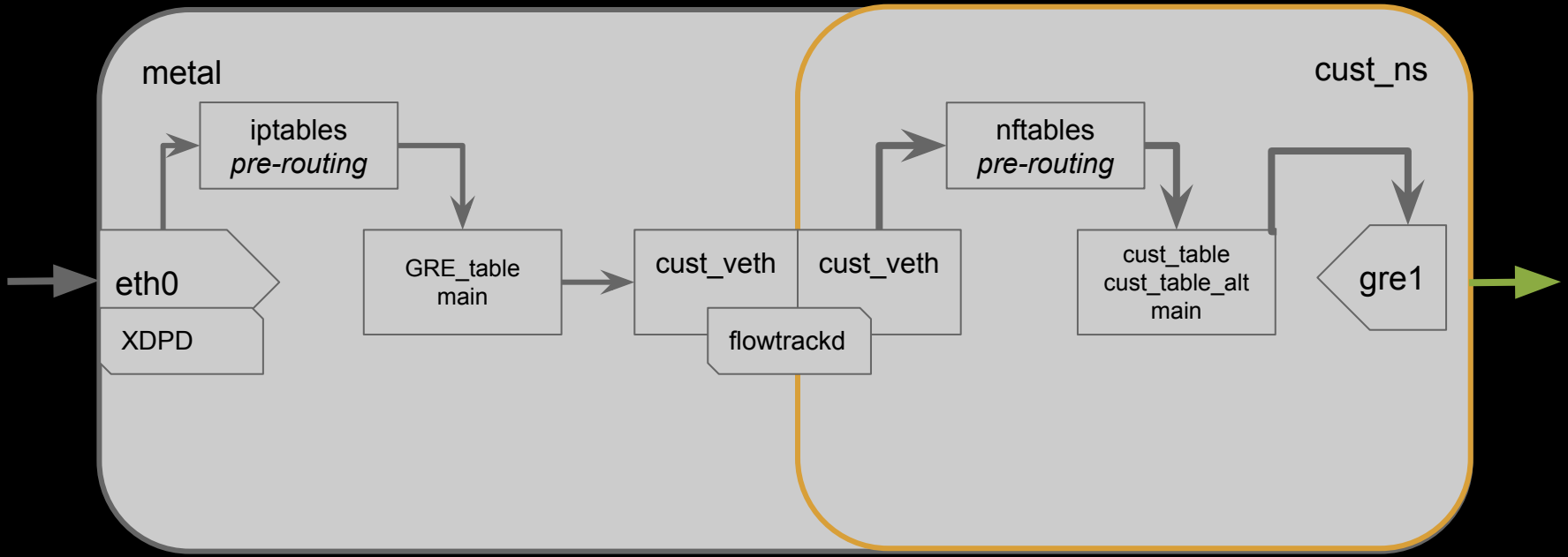
Devices, Devices, Devices
ECMP and Route Priorities



CLOUDFLARE

{erich, conjones}@cloudflare.com

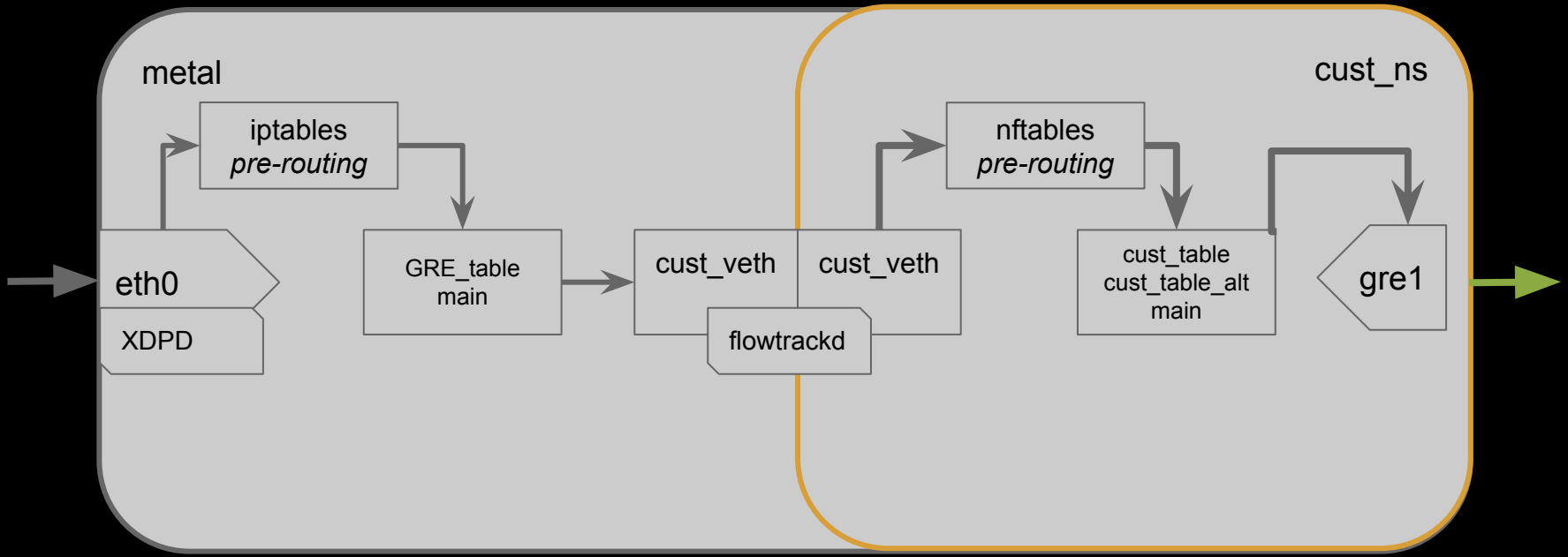
Devices, Devices, Devices
MAGIC-3



CLOUDFLARE

{erich, conjones}@cloudflare.com

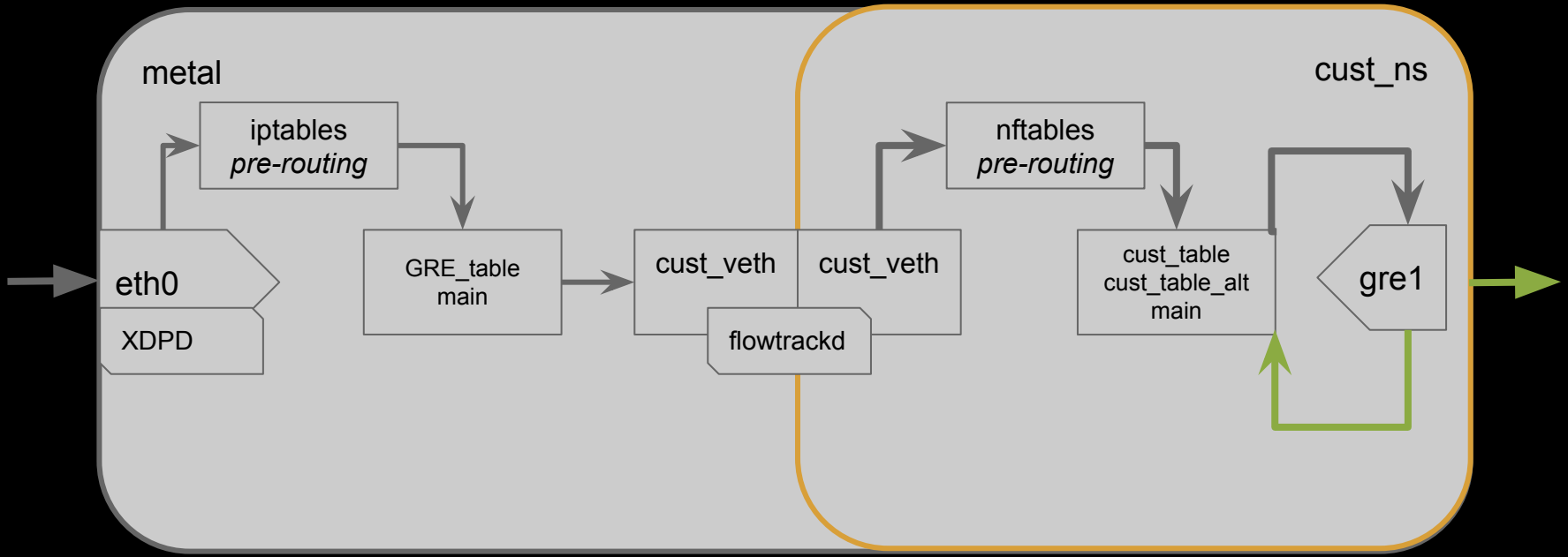
Focusing On One Device



CLOUDFLARE

{erich, conjones}@cloudflare.com

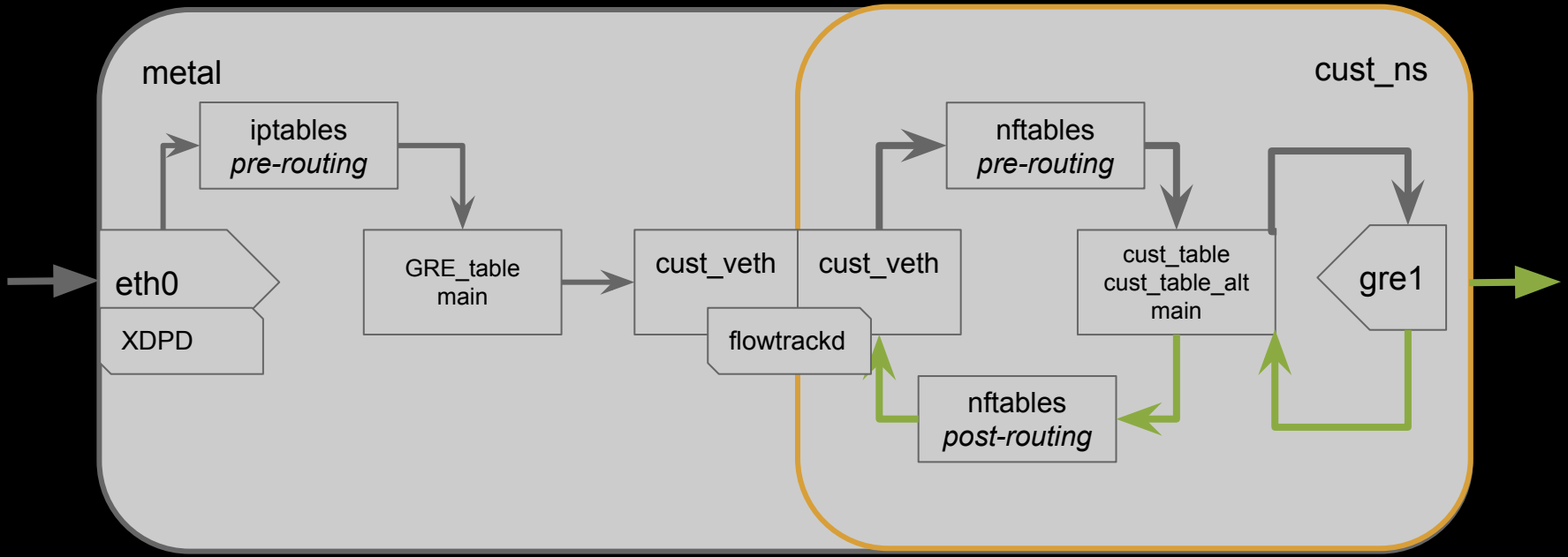
Default Route



CLOUDFLARE

{erich, conjones}@cloudflare.com

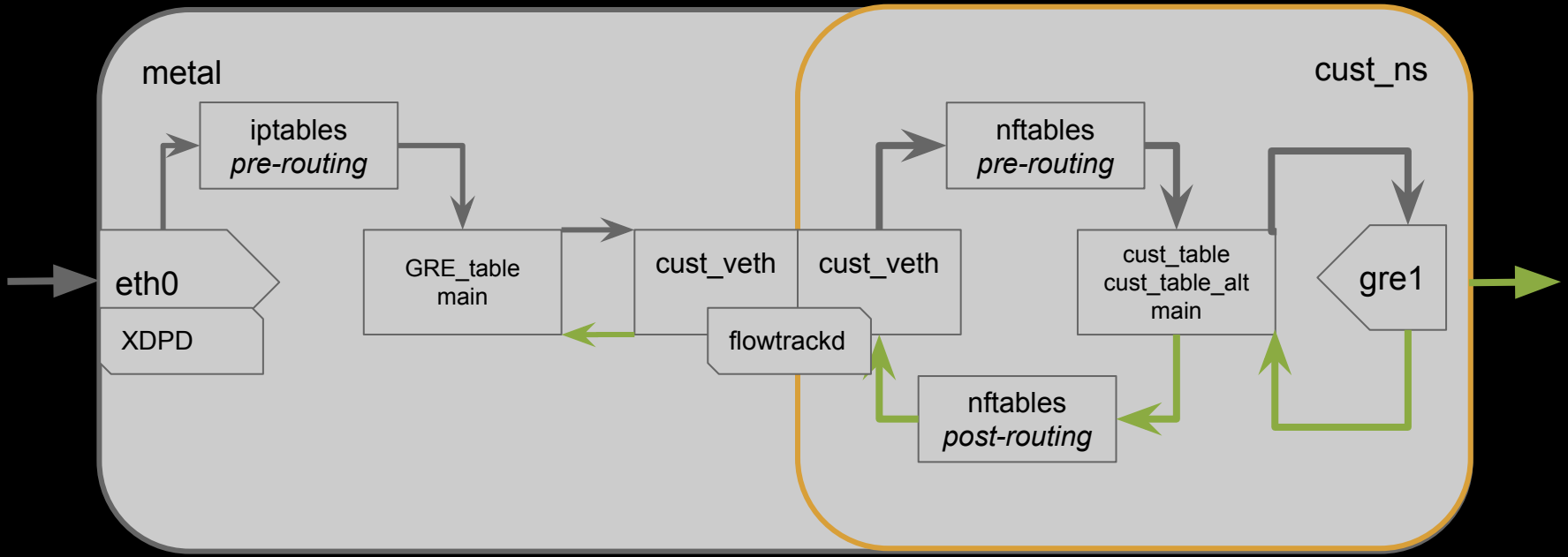
Outgoing Rules



CLOUDFLARE

{erich, conjones}@cloudflare.com

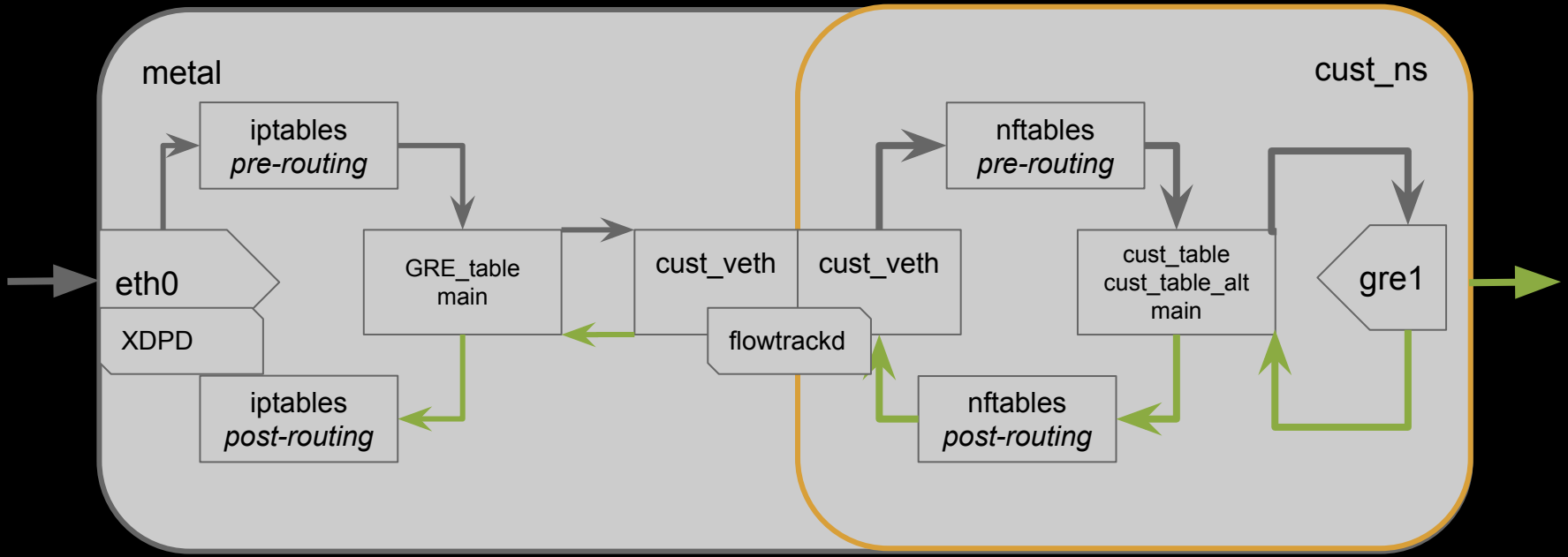
Traverse The veth, again



CLOUDFLARE

{erich, conjones}@cloudflare.com

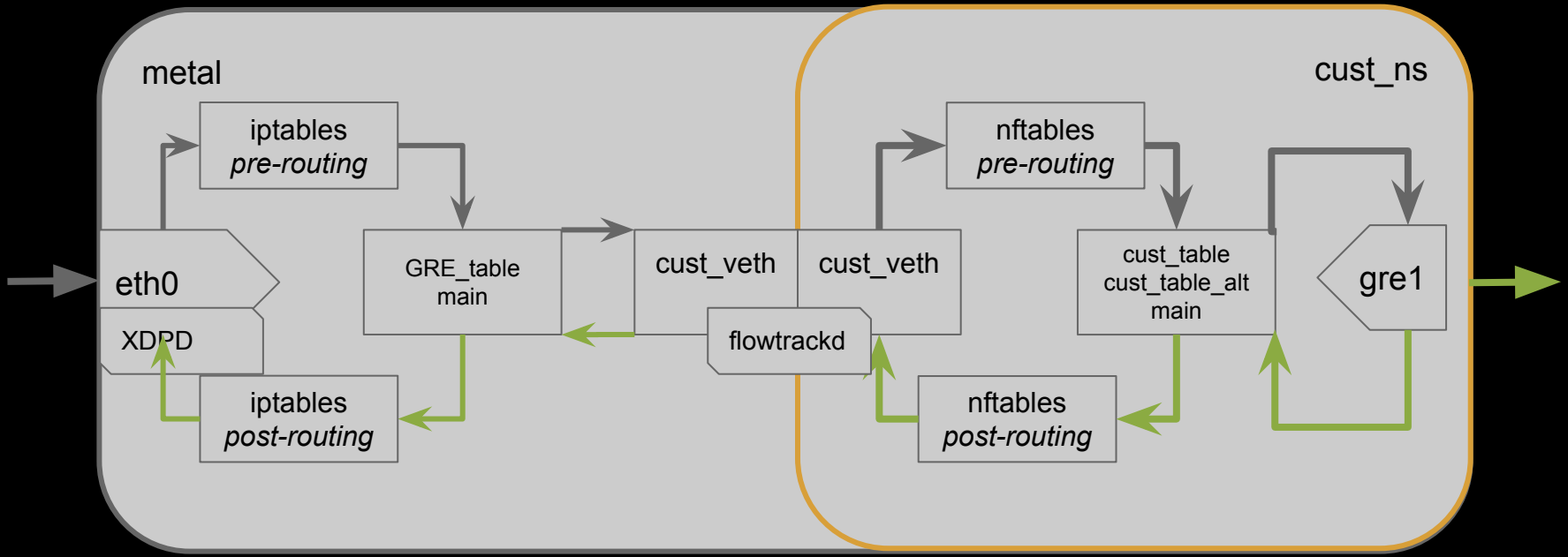
Outgoing rules



CLOUDFLARE

{erich, conjones}@cloudflare.com

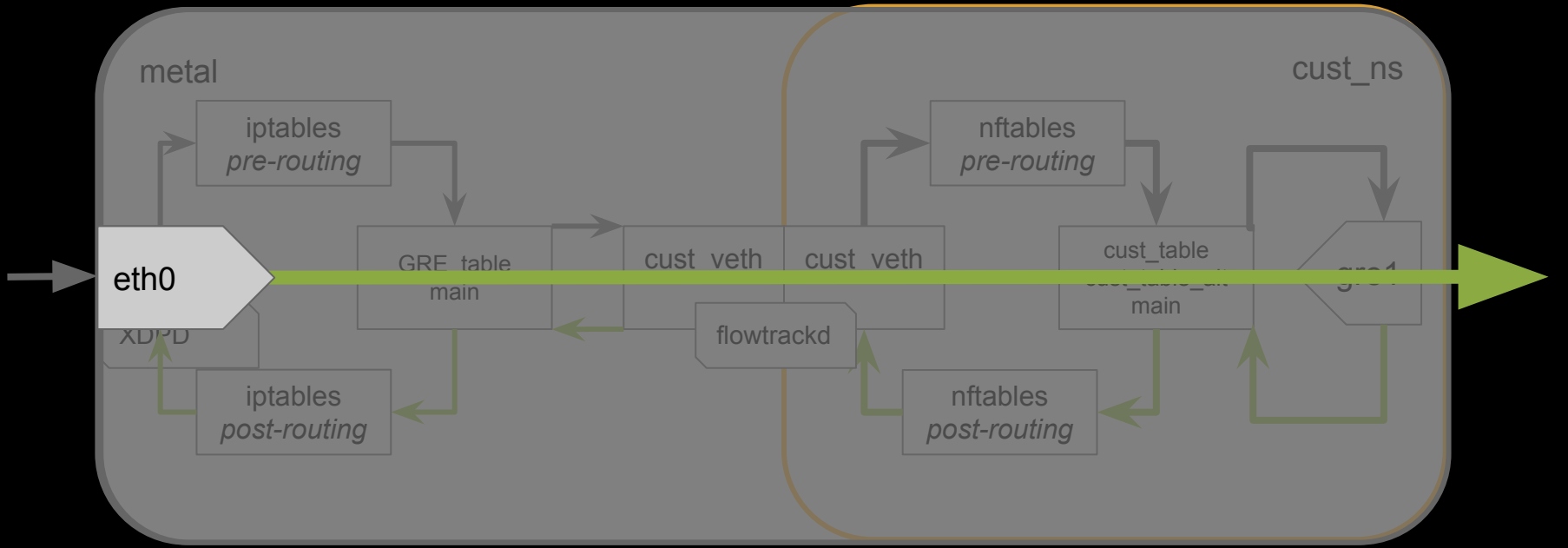
NIC Again



CLOUDFLARE

{erich, conjones}@cloudflare.com

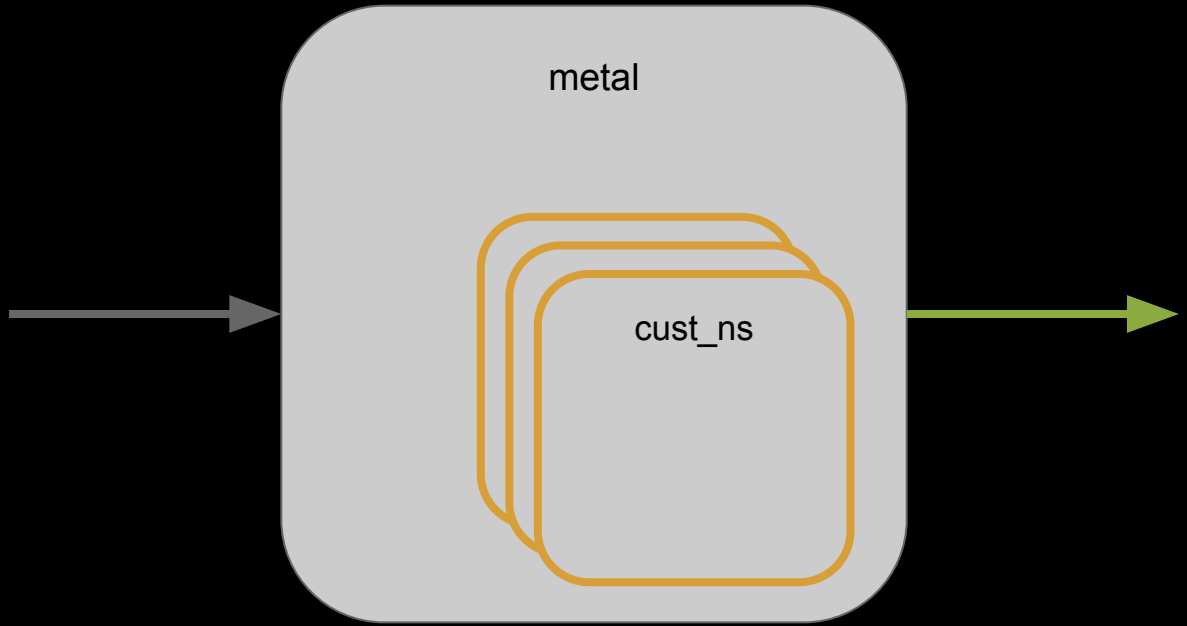
shipit!



CLOUDFLARE

{erich, conjones}@cloudflare.com

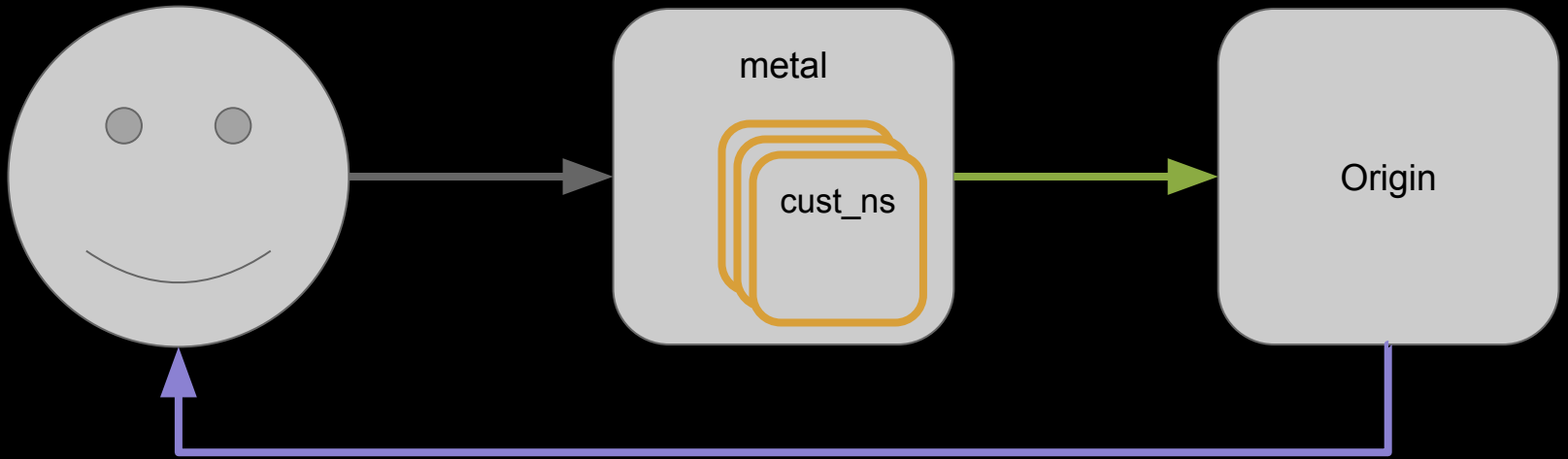
Whoops! We bricked a metal!
Slow-loop/Fast-loop Config Updates



CLOUDFLARE

{erich, conjones}@cloudflare.com

<todo>quippy gre-ping title</todo>



CLOUDFLARE

{erich, conjones}@cloudflare.com

Open Questions

- tx/rx checksum offloading
- Disappearing IPs
- ARP?

{erich, conjones}@cloudflare.com

Questions?

{erich, conjones}@cloudflare.com