Network Isolation for multi-IP exposure in XRootD

Frank Würthwein, Jonathan Guiang, Aashay Arora, **Diego Davila**, John Graham, Dima Mishin, Thomas Hutton, Igor Sfiligoi, Harvey Newman, Justas Balcas, Preeti Bhat, Tom Lehman, Xi Yang, Chin Guok, Oliver Gutsche, Phil Demar, Marcos Schwarz

XRootD and FTS Workshop Sep, 2024

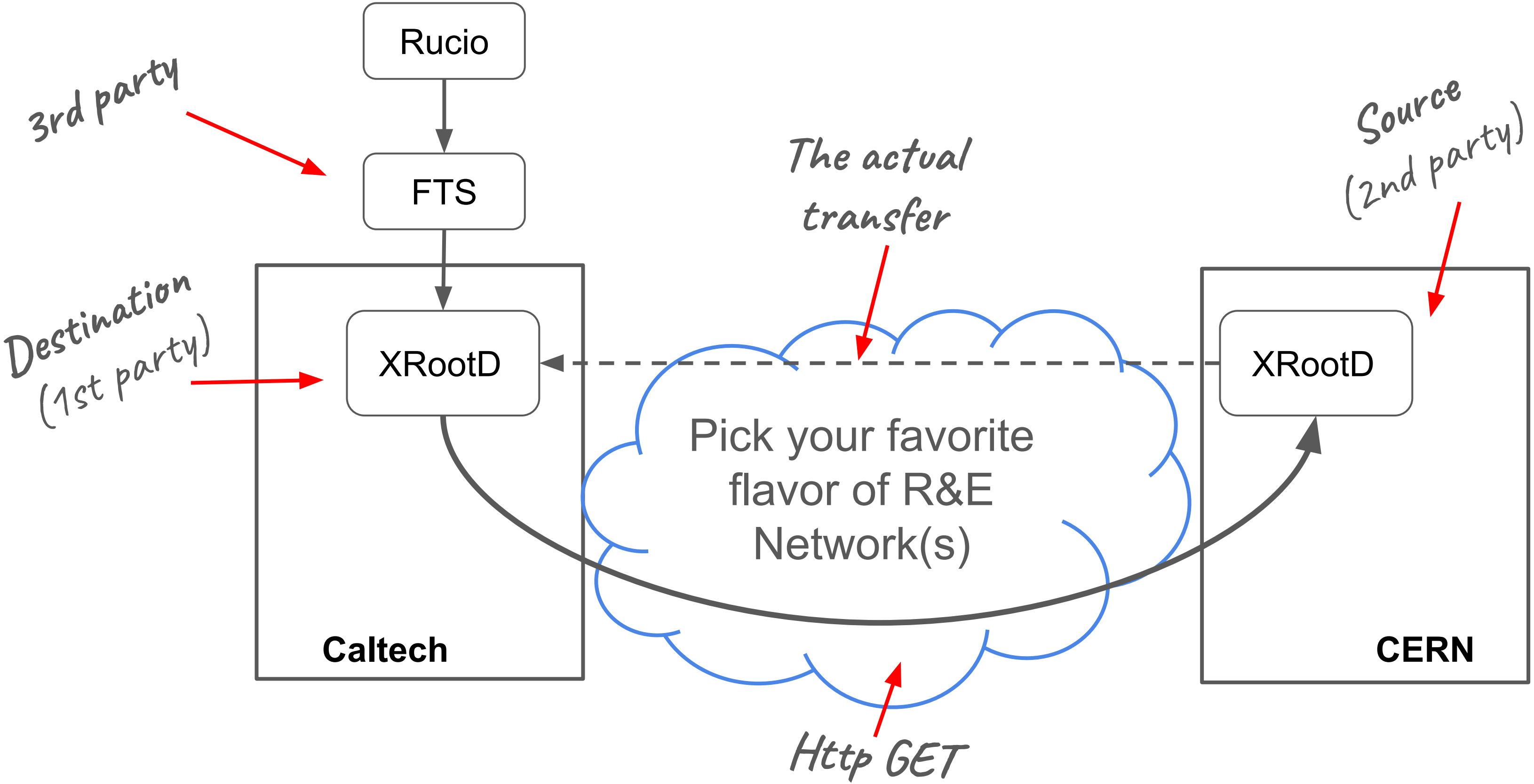
Why in hell are we doing this?

Short answer: *“To have better control over our transfers”*

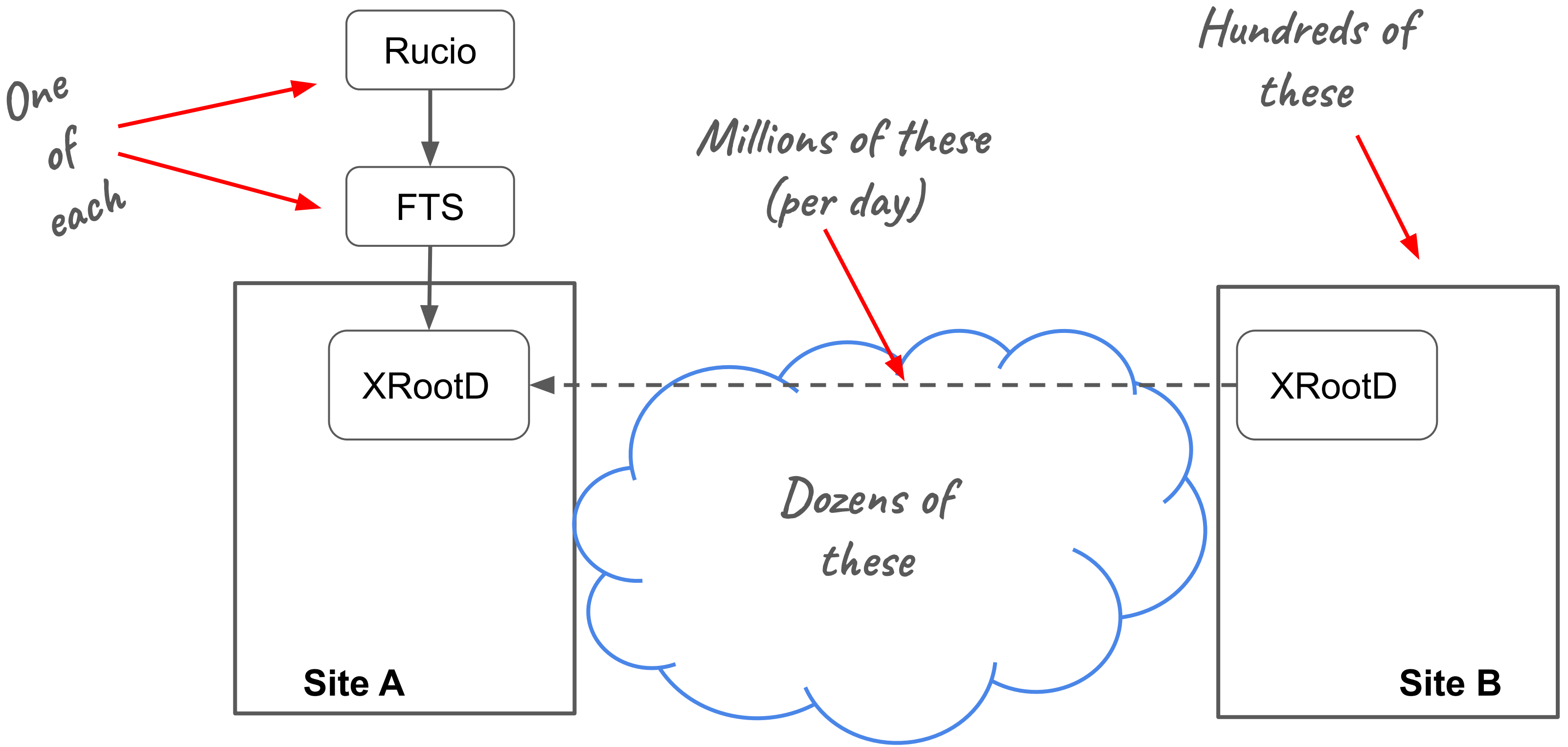
Context:

- This is R&D work we do for the LHC experiments (CMS and ATLAS)
  - Motivated by High Luminosity LHC (more data = more transfers)
- Focus on HTTP-TPC transfers, i.e. server to server full-file transfers

A Third-Party-Copy (TPC) transfer



Ballpark Numbers (CMS)



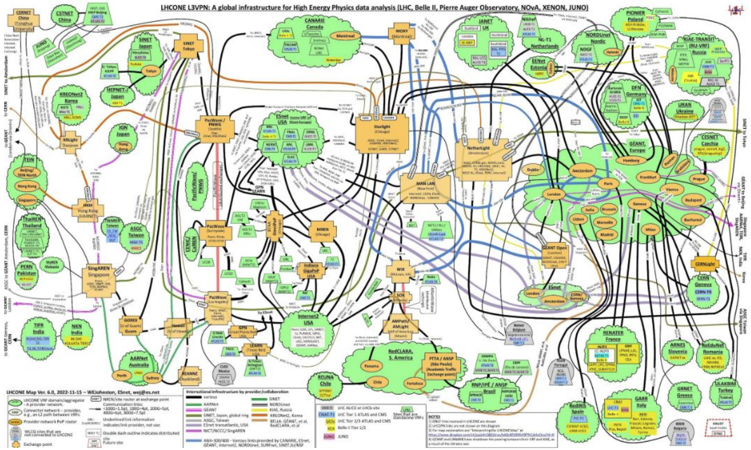
The problem

Network-wise **we treat all these Millions of transfers equally** i.e. they all get the same share of the network

..but we know **they are not all equally important**

Not a black box

Regardless of what like to think of; the Network it is NOT a black box and it’s neither an **infinite resource**


LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO)

≠

LHCONE network
http://nejohnston.org/LHCONE/Interpreting%20the%20LHCONE%20Map,%20LHCONE,%202020-09-14.pdf

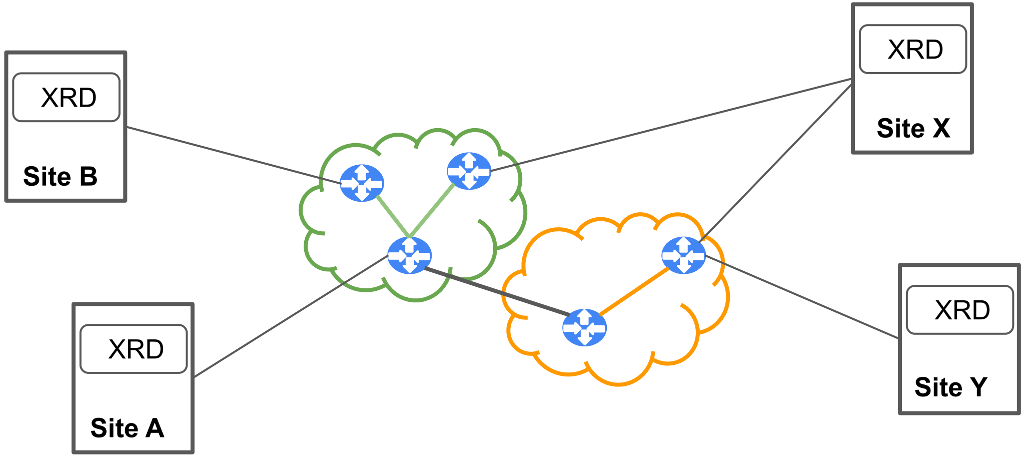
Just a Black Box

What if we could negotiate with the Network?

… and get special network services for special transfers

SENSE: Software Defined Networking (SDN) for End-to-End Networked Science at the Exascale

ESnet

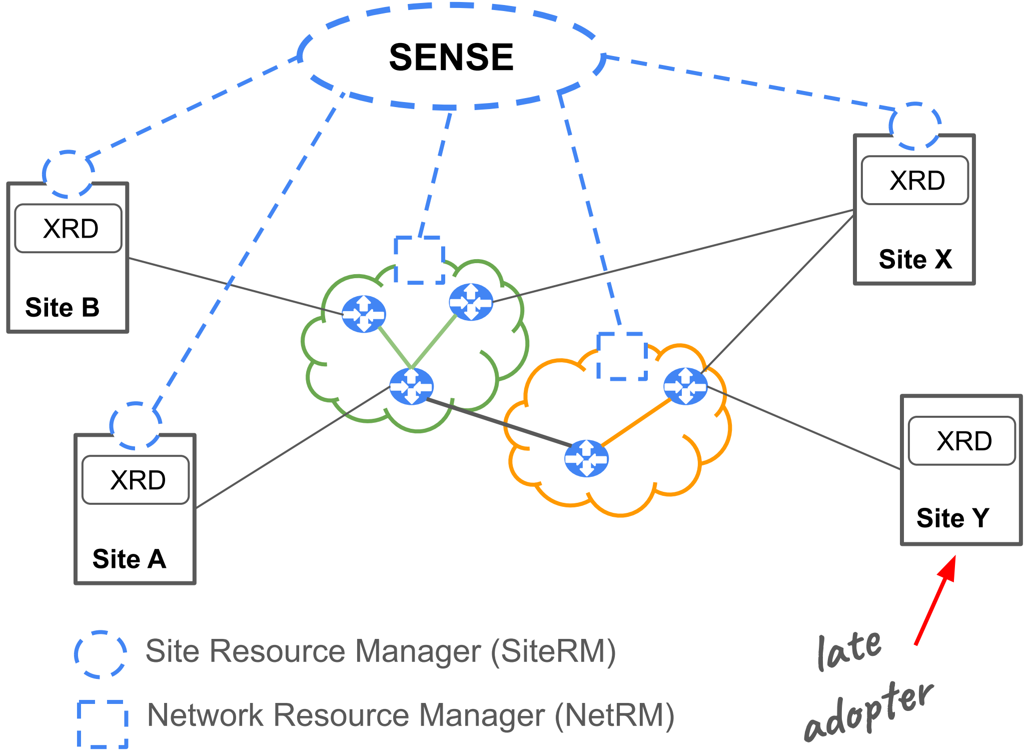


SENSE

Like a puppet master.

*It has agents both at the **Site** and the **Network** level*

*Uses these agents to configure **routing** and **bandwidth allocation** <u>on-demand</u>*



○ Site Resource Manager (SiteRM)

□ Network Resource Manager (NetRM)

late adopter

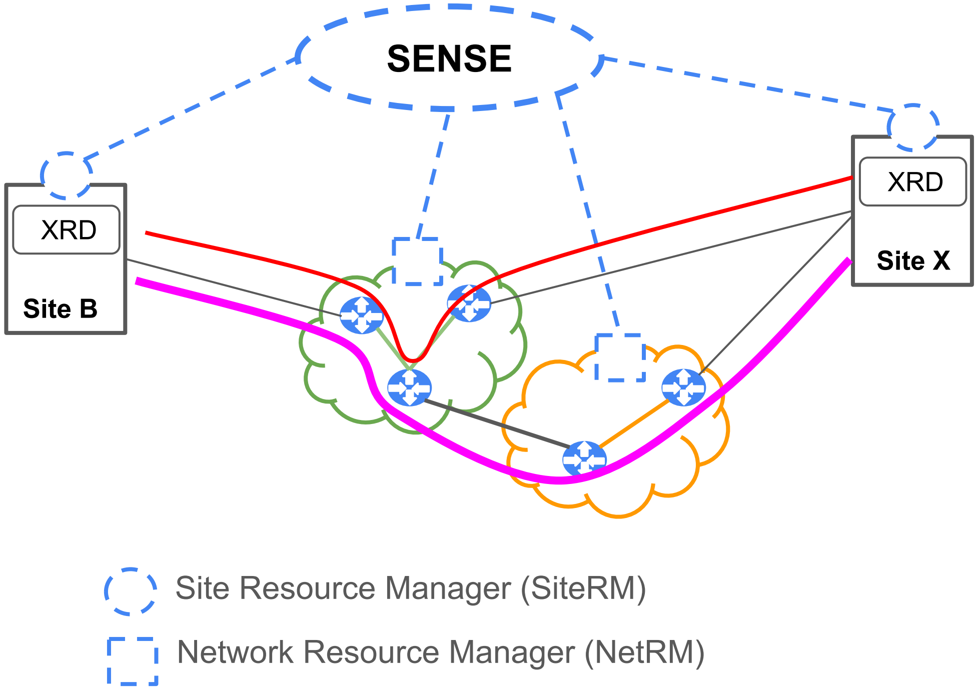
SENSE

For example:

we can ask SENSE to build a “special” path for a given LARGE transfer from B => X

And keep the rest of the transfers over the “best effort” path.

Once the LARGE flow is done the path is destroyed

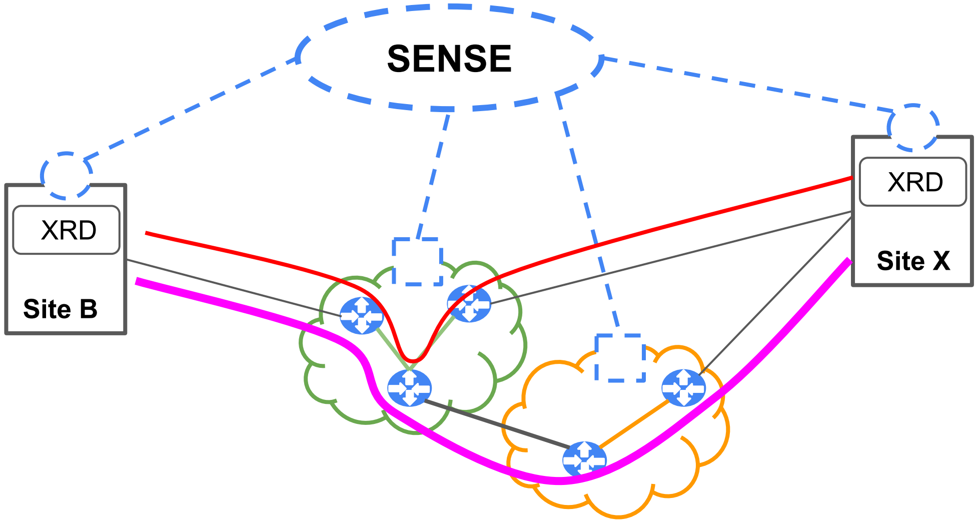

SENSE
XRD
Site B
XRD
Site X
Site Resource Manager (SiteRM)
Network Resource Manager (NetRM)

Why multiple-IPs?

SENSE builds these special routes based on subnets

Having **multiple "special paths"** on a given site, requires **multiple subnets**

In the following we show examples using 2 subnets: red and pink but in Prod we foresee to have 16 different subnets per site

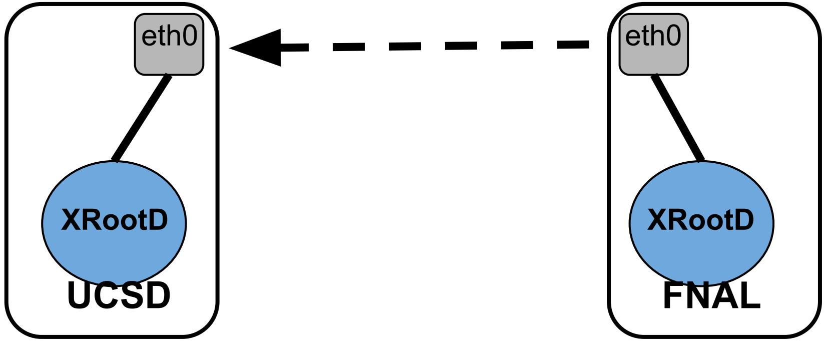


The Red and Pink paths connect to 2 different IPs on each Site

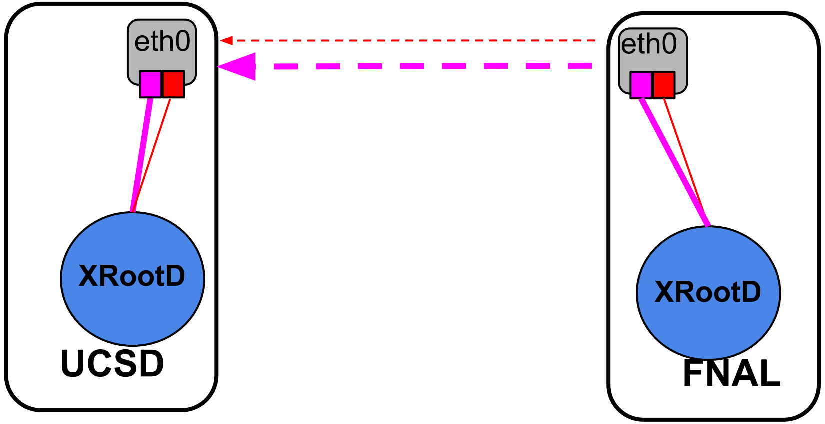
Single vs Multiple

NOTE: for sake of simplicity let's assume a Site is composed of a single server

In order to leverage from SENSE "magic" we need to go from **a)** to **b)**



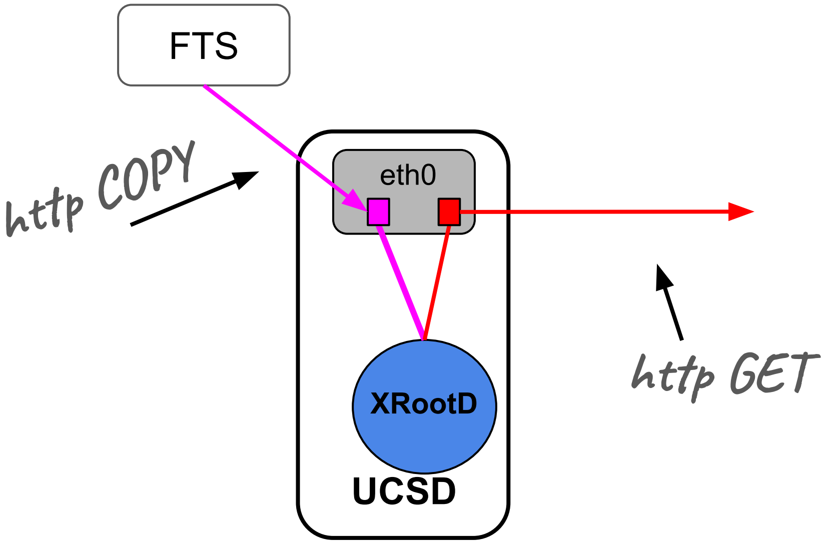
a) Single IP per server, all transfers travels between the **same pair** of IPs



b) Multiple IPs per server, transfers can travel between **different pairs** of IPs

It’s not that simple :(

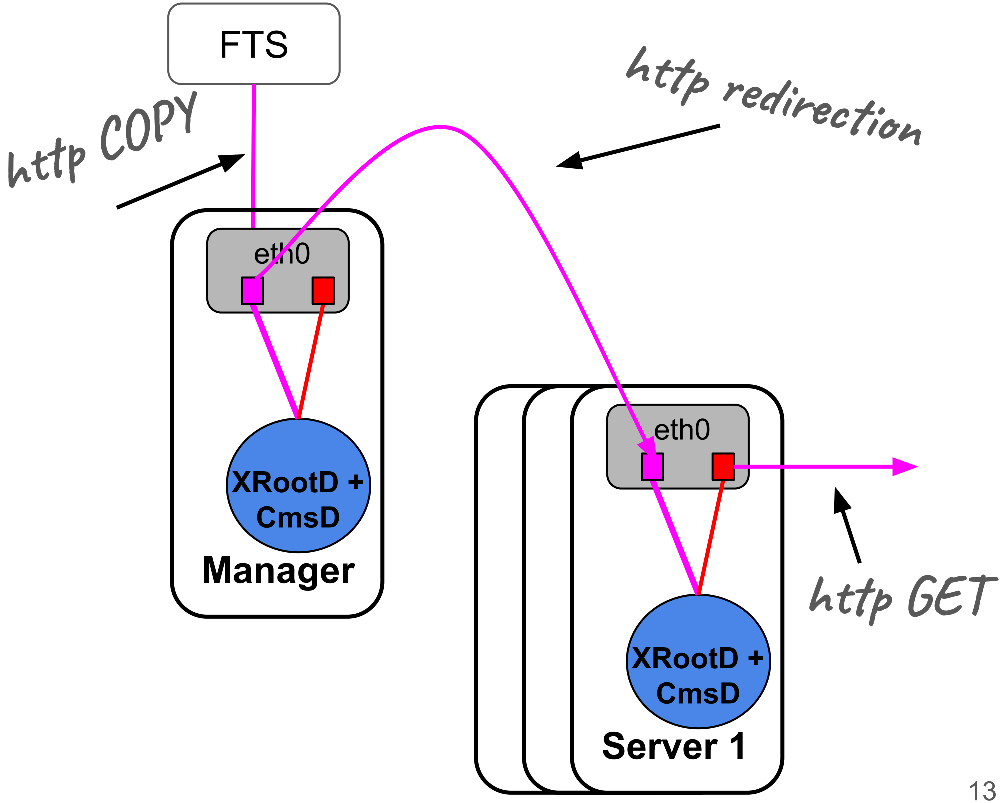
Naively, we thought that doing a TPC request to the pink IP will produce a GET from the pink IP… well it didn’t



Note: this has been fixed (kind of)

Gets more complicated in a cluster

Here we need the transfer request (COPY), the redirection and the GET to stay in the same subnet

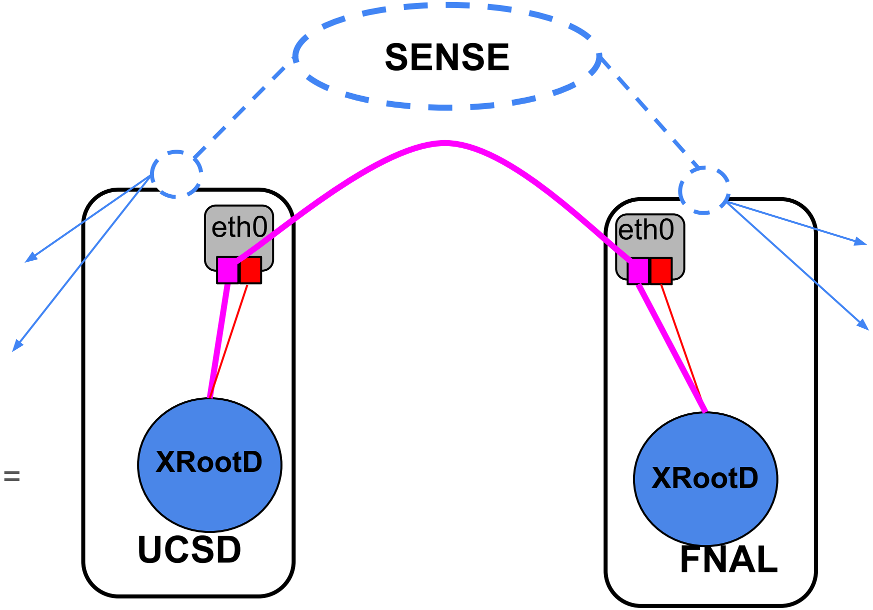


Note: this is still missing :(

Solution #1

Use SiteRM to Insert routing rules on both sides of the “special path”

```
==== Routing Table ====
TO   FNAL-pink USE pink

FROM FNAL-pink USE pink

DEFAULT use red
========================
```

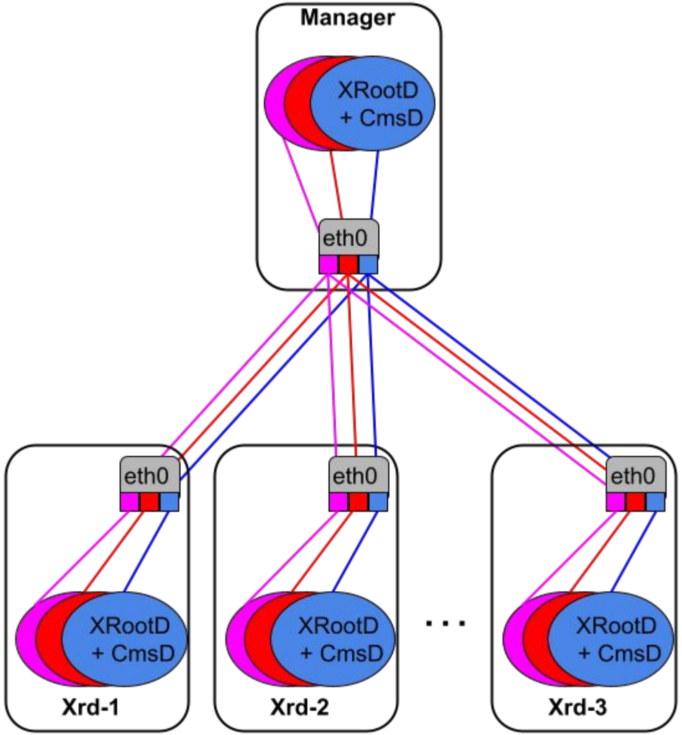


```
==== Routing Table ====
TO   UCSD-pink USE pink

FROM UCSD-pink USE pink

DEFAULT use red
========================
```

Solution #2

Use Network Namespaces to isolate multiple XRootD/CmsD instances, each of them attached to a different subnet

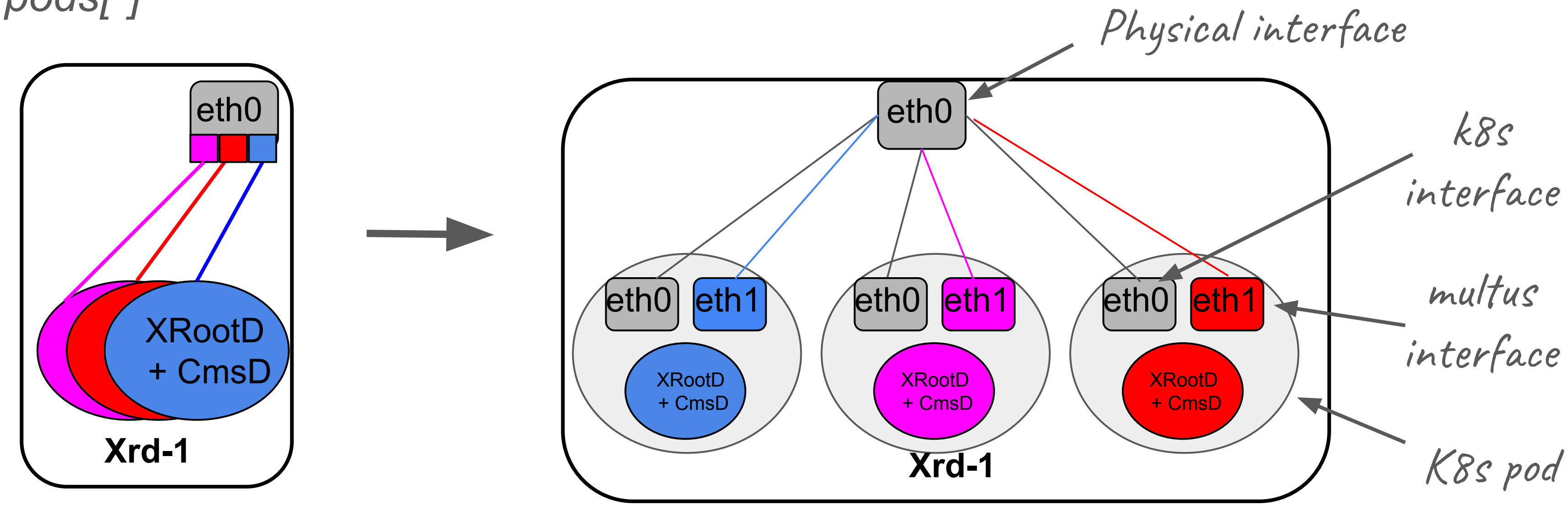
Each instance only sees 1 IP and its own (very simple) Routing Table



Each color globe represents an XRootD/CmsD instance in a separated network namespace

Solution #3

Similar to #2 but using Kubernetes and **Multus**: *a container network interface (CNI) plugin for Kubernetes that enables attaching multiple network interfaces to pods[*]*



[*] Multus: https://github.com/k8snetworkplumbingwg/multus-cni

Pros and Cons

Solution #	Pros	Cons
1	Least overhead for admins	Not a good idea to mess that much with the Routing Table
2	Significant overhead for initial set up	No changes required after initial setup
3	Easy if you are used to k8s	Hard if you are not used to k8s

Thanks!

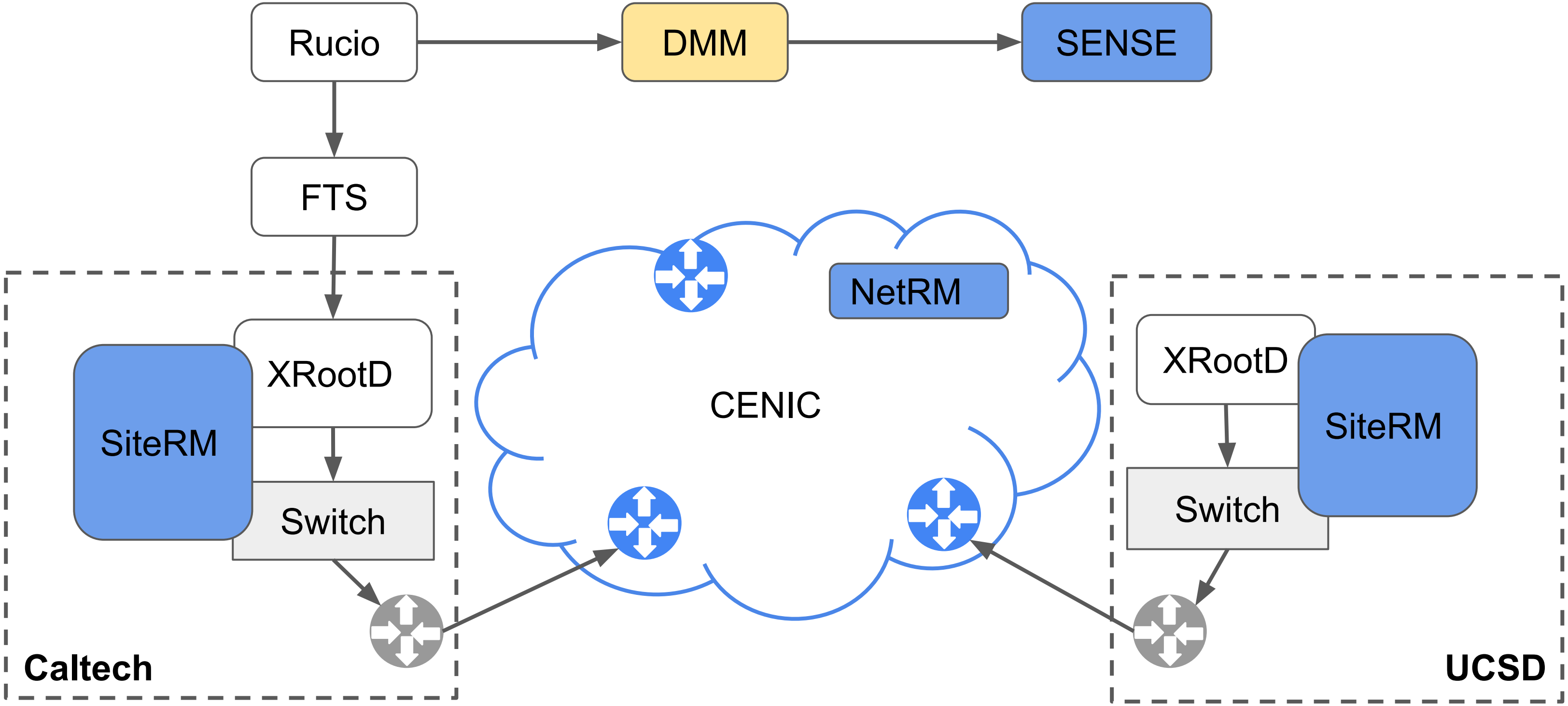
Questions?

ACKNOWLEDGMENTS

This ongoing work is partially supported by the US National Science Foundation (NSF) Grants OAC-1836650, MPS-1148698, and PHY-1624356. In addition, the development of SENSE is supported by the US Department of Energy (DOE) Grants DE-SC0015527, DE-SC0015528, DE-SC0016585, and FP-00002494. Finally, this work would not be possible without the significant contributions of collaborators at ESNet, Caltech, FNAL and SDSC.

Background slides

This is how Rucio + DMM + SENSE looks like


Rucio
DMM
SENSE
FTS
NetRM
CENIC
XRootD
SiteRM
Switch
Caltech
XRootD
SiteRM
Switch
UCSD