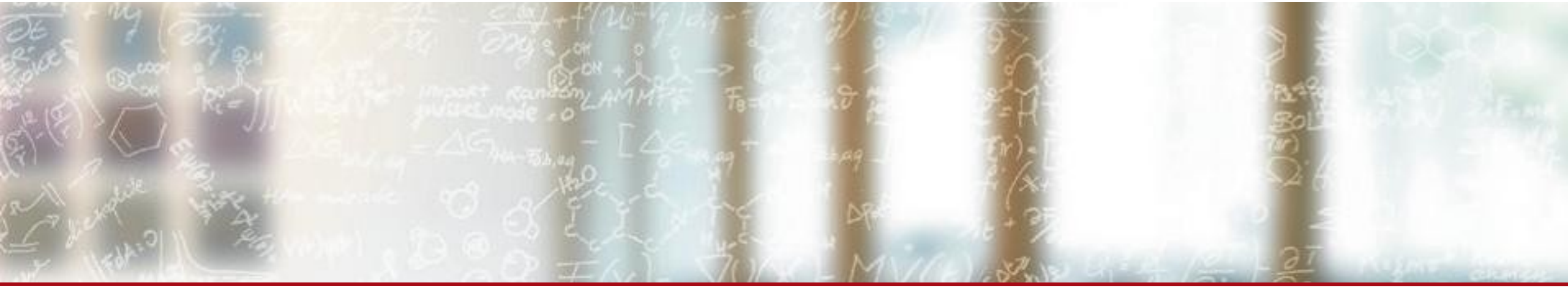
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Journey in slingshot HSN segmentation using VLANs

CUG2023 - Helsinki

Chris Gamboni, CSCS

May 11th, 2023

Table of Contents

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

CSCS Introduction

CSCS INTRO

- ALPS is an HPE Cray EX with CSM and Slingshot 11

- Currently running CSM 1.2 and Slingshot 2.0.1
- Alps will be the new CSCS flagship HPC system (2023-2024)

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

CSCS Multitenancy and VLANs

CSCS Multitenancy and VLAN

- The goal of CSCS is to have a flexible way to create different platforms
  - with different IdP
  - on different networks
  - on different security zones

⇨ **We need to segment slingshot network using VLANs**

- VLAN tests started on Q1/2023, after the upgrade of slingshot to 2.0.1
  - First on PreAlps (Alps TDS), then on Alps
- Work in progress

CSCS Use cases

CSCS Use cases for multitenancy: use case 1

- CSCS platform
  - Most nodes will stay on the default network segment (VLAN1)
  - Many vClusters, but on the same VLAN and same security zone
  - Traditional HPC vClusters
  - Experimental vClusters
  - Storage , shared storage

CSCS Use cases for multitenancy: use case 2

- Dedicated platform
  - Dedicated VLAN
  - CSCS network / address space
  - Some customers may need public IP address space
    - Note: All CSCS machines currently use public IP addresses on HSN
  - Different security zones
  - Dedicated storage
  - Separated from other vClusters or platform
  - IdP may be different from CSCS

CSCS Use cases for multitenancy: use case 3

- Dedicated plaform for tenants on customer network
  - Dedicated VLAN on customer network
  - On customer IP Address space
  - Customer IdP
  - Dedicated storage
  - Separated from other vClusters or platform

CSCS Technical setup

CSCS Data Center network (1)

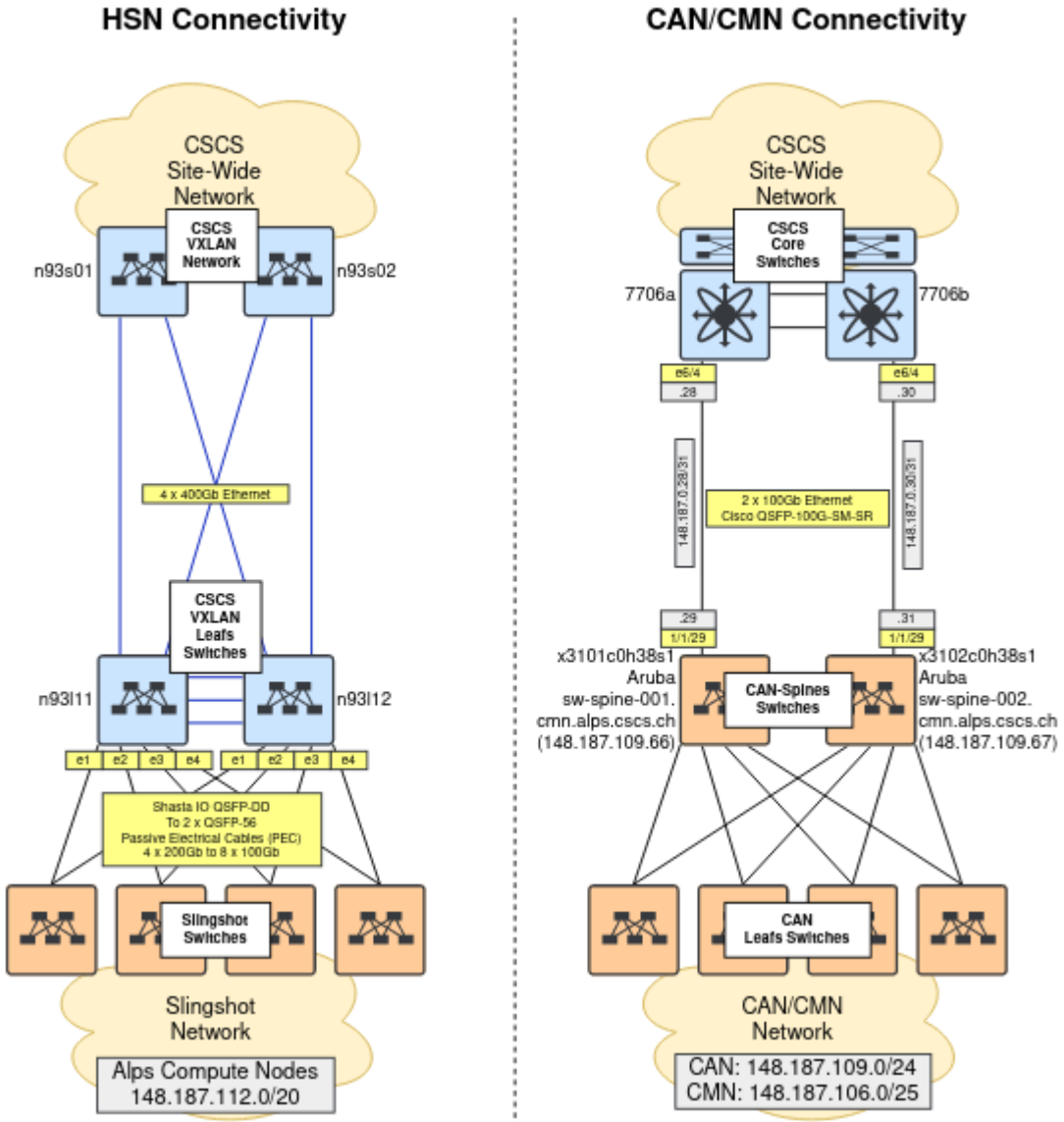
- 2 x 100Gbit/s to INTERNET
- CSCS Data Center network with 400GbE VXLAN / EVPN
  - 2 spines with 64 ports 400GbE switches
  - Many leaf switches with 400/100 Gbit/s connections to spine switches
- ALPS slingshot network is currently connected with an MLAG of 8 x 100Gbit/s to CSCS Data Center network
- L3 on CSCS Data Center network

- Next steps:
  - INTERNET connection will be upgraded to 400Gbit/s
  - Double the spine switches: from 2 to 4 spine 64 port 400GbE switches
  - Increase the capacity between ALPS and CSCS Data Center network
    - From 800 Gbit/s to 1600 Gbit/s or more

CSCS Data Center network – HSN and CAN/CMN connections

4 x 400G to CSCS spine switches

8 x 100G using the provided copper cables (previously attached to the edge gateways)



2 x 100G-CWDM (routed) between management network and CSCS Data Center network

CSCS Data Center network – HSN VLAN support

- Slinghshot network is directly connected to CSCS Data Center network
  - no dedicated “edge gateways”
- New VLANs have been created on CSCS Data Center network
- New SVIs (routed interfaces) on CSCS Data Center network
- Dedicated ACL (Access lists) for the new VLANs
- MLAG on CSCS DC network has been configured as 802.1q VLAN trunks

Slingshot fabric configuration

- Add the new VLANs to the slingshot fabric manager

```
fmctl create vlans name=vlan337 status=ONLINE id=337 --raw | jq .
fmctl create vlans name=vlan396 status=ONLINE id=396 --raw | jq .
```

- Enable VLAN support in the slingshot fabric manager

```
fmctl update topology-policies/template-policy fabricPropertyMap.vlanEnabled=true --raw | jq .
```

Slingshot Switch configuration (1)

- Create a new JSON file for the new VLAN

```
{
  "nativeVlanId": "/fabric/vlans/337",
  "documentKind": "com:services:fabric:models:PortPolicyState",
  "documentSelfLink": "/fabric/port-policies/vlan-337",
  "allowedVlans": [
    "/fabric/vlans/337"
  ],
  "isUntaggedAllowed": true
}
```

- Create a new port policy for the new VLAN

```
fmctl create port-policies --file 337.json --raw | jq .
```

Slingshot Switch configuration (2)

- Create a new JSON file with 802.1q support for the MLAG

```
{
  "state": "ONLINE",
  "autoneg": true,
  "speed": "BJ_100G",
  "precode": "AUTO",
  "mac": "02:00:00:00:00:00",
  "loopback": "NONE",
  "nativeVlanId": "/fabric/vlans/1",
  "documentKind": "com:services:fabric:models:PortPolicyState",
  "documentSelfLink": "/fabric/port-policies/cisco-vlan",
  "allowedVlans": [
    "/fabric/vlans/1",
    "/fabric/vlans/337",
    "/fabric/vlans/396"
  ],
  "isUntaggedAllowed": true
}
```

- Create a new port policy for the new VLAN

```
fmctl create port-policies --file cisco-vlan.json --raw | jq .
```

Slingshot Switch configuration (3)

- Apply the port policy to enable 802.1q support on MLAG interfaces

```
fmn-update-port-policy -n cisco-vlan,edge-policy-disable-autoneg,qos-ll_be_bd_et-ethernet-policy $PORTS
```

- Apply the new VLAN port policy to the desired switch port of the node

```
fmn-update-port-policy -n vlan-337,cassini-policy,qos-ll_be_bd_et-cassini-policy x1500c4r3j107p1
```

Nodes configuration (1)

- All nodes start with a default IP configuration as described in the fabric template, even if the nodes are configured on a dedicated VLAN

- Important: DVS is on nmn

Nodes configuration (2)

- With an ansible script we do the following steps, for each node:
  - Delete the current HSN IP Address(es)
  - Delete the current IP rules
  - Configure the new IP Address(es) on HSN interface(s)
  - Configure the new IP rules
  - Configure the correct IP routes, including the default gateway
  - If needed: change the DNS server address

Challenges

- Nmn network is still fully open – all nodes may talk to each other
  - Workaround/Solution will be based on a mix of:
    - ACLs on the nmn network switches to avoid traffic between nodes on the nmn network
    - Firewall on compute nodes (nmn interface)
- Shared storage access, routed via L3 CSCS Data Center Network
- DNS configuration for nodes on a different VLAN

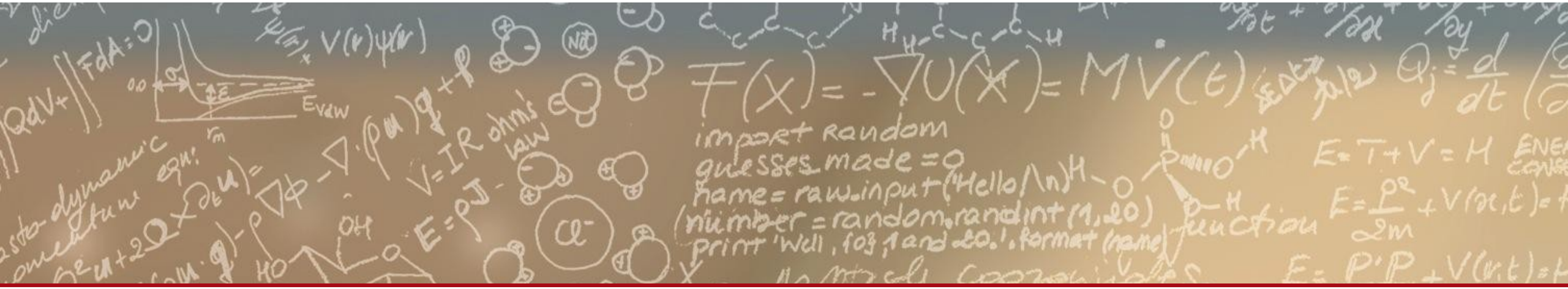
Next steps

- ACL / FW for the management network
- Automate the VLAN configuration of the slingshot switches
  - Some manual configuration steps still needed, room for improvement
- Use a central configuration for nodes, VLANs, new IP addresses and other info
- VNI and QoS configuration
- Add some flexibility, for example to dinamically move nodes between VLANs

Many thanks to:

- Miguel Gila (CSCS)
- Jérôme Tissières (CSCS)

- Davide Tacchella (HPE)



Thank you for your attention.