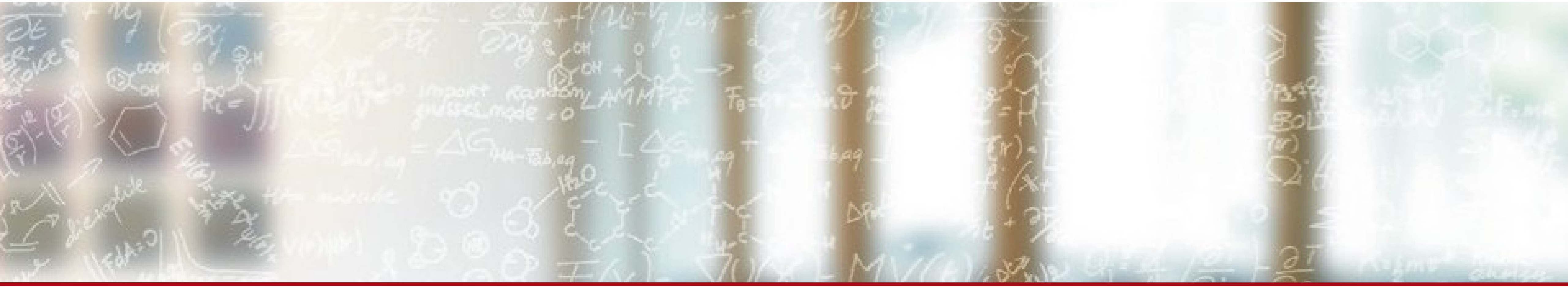
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Separating concerns: Decoupling the Slingshot Fabric Manager from Cray System Management

CUG25

CSCS: Chris Gamboni, Riccardo De Maria, Mark Klein

HPE: Davide Tacchella, Isa Wazirzada

May 2025

Synopsis

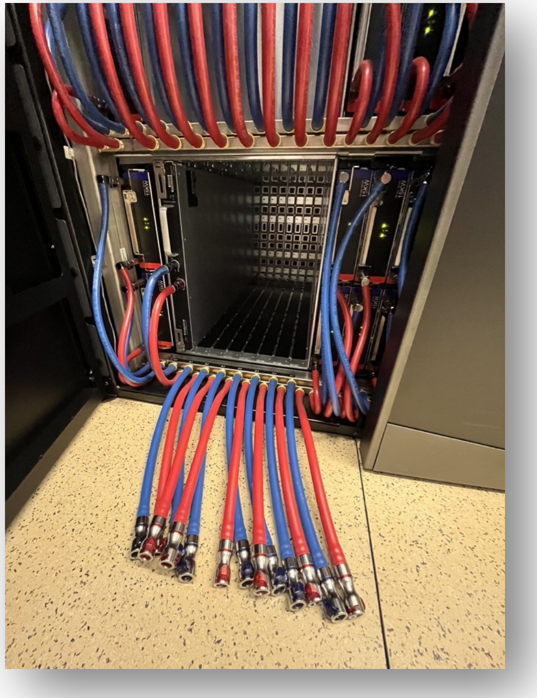
- Motivation
- Objective and requirements
- Technical developments
- Future plans
- Lessons learned and conclusion

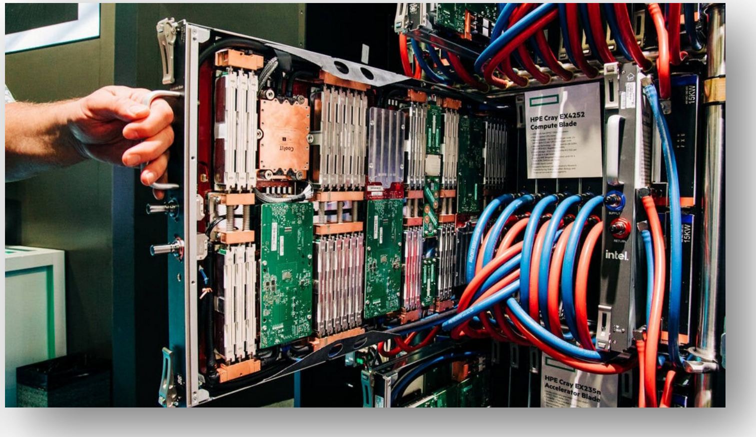
Motivation

- Decouple Slingshot Fabric Manager from CSM
- Improve fabric manager resiliency and fabric stability
- Desire to switch to a High Availability (HA) model that supports failover
- Part of an overall strategy to improve resiliency

Alps Research Infrastructure

- Alps is an HPE Cray EX supercomputer being our new flagship infrastructure
- Some specs
  - ***24 Cray EX 4000 Cabinets***
  - ***8 Cray EX 3000 Cabinets***
  - ***984 Slingshot Switches***
  - ***2688 Grace-Hopper nodes***
  - 1024 AMD Rome-7742 nodes 256/512GB
  - 144 Nvidia A100 GPU nodes
  - 128 AMD MI300A GPU nodes
  - 24 AMD MI250x GPU nodes
  - Two availability zones (HA, non-HA)
  - 100+10 PiB HDD
  - 5+1 PiB SSD (RAID10)
  - 100s of PiB tape library
  - ~10 MW (envelope for power and cooling)
  - 8x 100 Gb/s connection to CSCS network

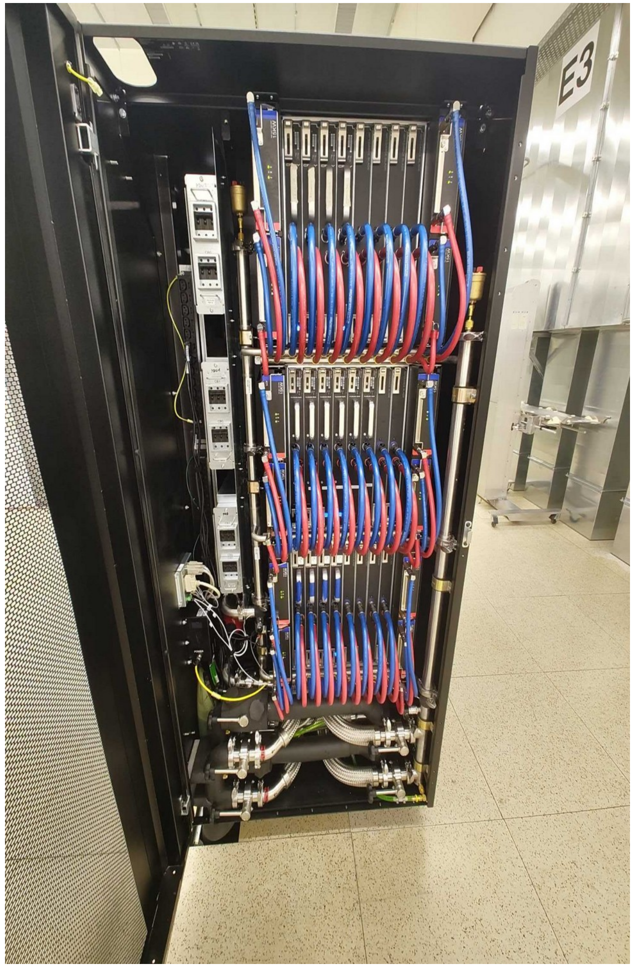




Water cooled blades

PreALPS

- PreAlps is our Staging system
- Some specs
  - 1 cabinet EX2500
  - CSM 1.6.1
  - Slingshot FM 2.3.0
  - GH200, MI250 and AMD CPU nodes
  - 16 Slingshot switches
  - 8x 100 Gb/s connection to CSCS network



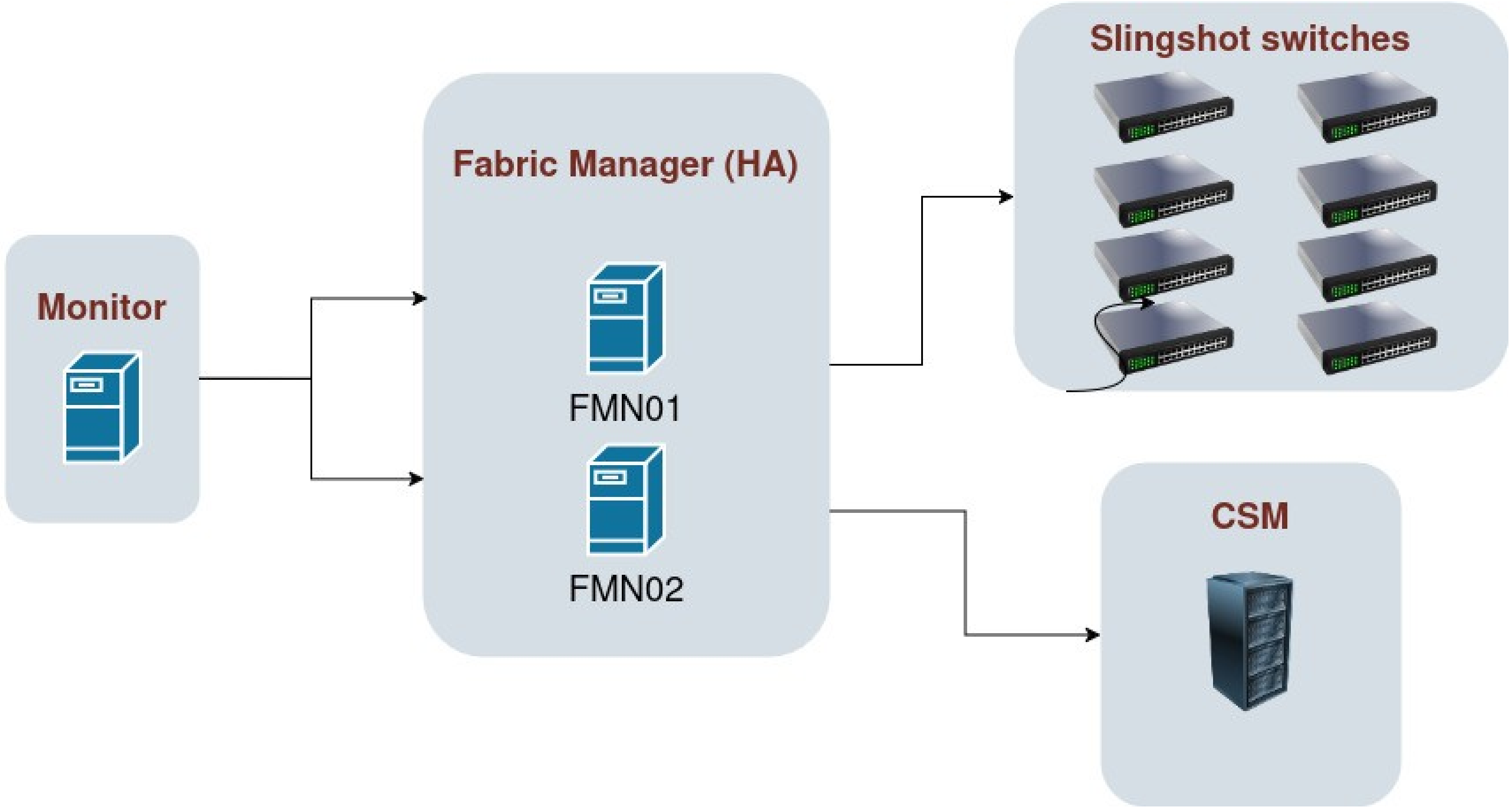
Objectives

- Investigate operational fabric dependencies on CSM services
- Install Slingshot fabric manager**s** on bare metal, decoupling from CSM
- Highly available fabric management, with failover
- Migrate fabric management to the new fabric manager HA cluster
- Bring up fabric to healthy operational state in new environment

Hardware

- Two identical hosts
  - HPE ProLiant DL325 Gen10 Plus
  - 64GB RAM, AMD EPYC 7302P 16-Core Processor
- OS
  - Rocky Linux release 9.5 (Blue Onyx)
- Management Network Connections:
  - Node Management Network (NMN) - To interact with K8s services
  - Hardware Management Network (HMN) - connectivity to switches
  - Customer Management Network (CMN) - SSH access to the fabric managers
- VM for monitoring HA cluster
  - Resides on the CSCS datacenter network

Solution Design



Technology: Installation

- Stop fabric management on CSM
- Backup the CSM fabric manager
- Install fabric manager software on the two bare metal nodes
- Restore fabric manager from backup on bare metal
- Set up certificates/SSH
- Create entries for each fabric manager in `/etc/hosts`
- Update the `active-standby-domain.json` file
- Enable fabric management on the first node
- Synchronize the fabric managers
- Verify `fmn-show-clusters` output

Wait, what about the fabric manager pod?

Technology: What about DHCP and DNS?

- Slingshot switch DHCP IP assignment is managed by CSM
- DNS entries are also managed by CSM – still need to lookup switch hostnames
- To DHCP or not to DHCP?
  - Current lease Time To Live (TTL) is 60 minutes
  - Does it make sense to move to static switch IPs/hosts file
  - **Answer** [redacted]

- `fmn-update-dns`
  - Can create, delete, and update edge port DNS entries (ex. nid001000-hsn0 maps to 10.253.0.24)
  - Usually leveraged during the fabric bring up process
  - Interacts with the System Layout Service (SLS) in CSM

Source: HPE documentation

Monitoring host

- The scope of the monitoring host is to check the status of the fabric
  - Restart fabric manager if it is down
  - Move the active fabric manager to the healthy host in case of failure
  - Caveat: If fabric management has been disabled on the standby HA node, failover won’t work
- Monitoring host has been installed on a VM in a separated network
- Monitoring host needs to access both FMN nodes via ssh and https (tcp 8443)

Source: HPE documentation

Operational view: fabric operations

- All fabric commands should be entered on the node with role ACTIVE

```
Prealps-fmn01:~ #  fmn-show-cluster

address          : state           : switchStateUpdates : clusterRole  : local

10.254.1.201     : FM_UP           : ENABLED            : ACTIVE       : TRUE

10.254.1.202     : FM_UP           : ENABLED            : STANDBY      : FALSE
```

- In that case, we need to issue all commands to fmn01
- The command fmn-show-cluster works on both nodes

```
Prealps-fmn02:~ # fmn-show-cluster

address          : state           : switchStateUpdates : clusterRole  : local

10.254.1.201     : FM_UP           : ENABLED            : ACTIVE       : FALSE

10.254.1.202     : FM_UP           : ENABLED            : STANDBY      : TRUE
```

Operational view: fabric operations

```
Prealps-fmn02:~ # fmn-show-status
------------------------------------------
Topology Status
Policy: template-policy
Health
------
Runtime:HEALTHY
Configuration:HEALTHY
Traffic:HEALTHY
Security:HEALTHY
For more detailed Health - run 'fmctl get health-engines/template-policy'

ERROR: Fabric management is not enabled. (clusterRole: STANDBY switchStateUpdates: True)
Exiting...
```

Operational view: fabric configuration

- In case we need to apply changes to the fabric:
  - Disable the monitor on the monitoring host
  - Apply the changes to the fabric
  - Verify the state of the fabric
  - Synchronize the cluster
  - Enable the monitor

Operational view: fabric configuration synchronization

- Fabric configuration changes need to be synchronized between FMNs
- This is accomplished via the `fmn-synchronize-active-standby` command
  - `fmn-synchronize-active-standby --node-source 10.254.1.1 --node-target 10.254.1.2`
- Feature request
  - Have an easy way to verify the configuration consistency between fabric managers
  - Have configuration changes synchronize automatically - within reason ☺

Operational view: manual failover

- Forcing a failover
- In case of maintenance on the active node, we can force a failover to the other node

```
fmn01:~ #  fmn-show-cluster
address         : state           : switchStateUpdates : clusterRole  : local
10.254.1.201    : FM_UP           : ENABLED            : ACTIVE       : TRUE
10.254.1.202    : FM_UP           : ENABLED            : STANDBY      : FALSE

fmn01:~ # fmn-failover-active-standby --node-active 10.254.1.202 --node-standby 10.254.1.201
address         : state           : switchStateUpdates : clusterRole  : local
10.254.1.201    : FM_UP           : ENABLED            : STANDBY      : TRUE
10.254.1.202    : FM_UP           : ENABLED            : ACTIVE       : FALSE
```

Operational view: automatic failover

- Monitoring host checks the fabric status every 10 seconds
- In case of problems with the active fabric manager node
  - The monitoring host automatically assigns the active fabric manager to the healthy fabric managers
  - This will happen only if the standby node has fabric management enabled
- Important – One needs to remember which FMN is “in control”
  - If one has failed over to another FMN one has to remember to SSH to it – **habits are hard to break!**
  - Feature request: Make it easier for admins to only interact with the active FMN

Lessons learned

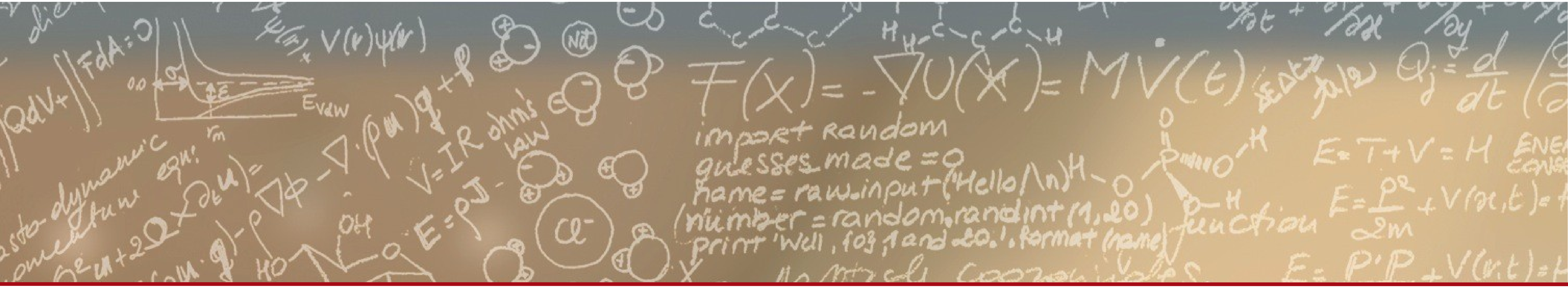
- Setting up new servers is always “fun” (cabling, power, booting)
- Getting the monitoring VM configuration required careful planning
- Admin experience - single pod to HA FMN pair is a mindset shift
  - Need to make sure you’re on the correct FMN
  - Need to remember to sync changes every time
- Improve configuration backup process

Future work

- Improve automatic failover
- Investigate interaction with the K8s based Slingshot network operator
- Explore – how to remove dependencies on management services
- End goal: Mirror PreALPS, in production on ALPS

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich


F(x) = -∇U(x) = MV̇(t)
import random
guesses_made = 0
name = raw_input('Hello\n')
number = random.randint(1, 20)
print 'Well, {0}, 1 and 20.'.format(name)
V=IR ohm's law
E=T+V=H
function
E_vdW

Thank you for your attention.