Cluster Management Tool

RESINFO

ANF Outils de déploiement

David DELAVENNAT
CMLS (INSMI / Ecole polytechnique),
GDS Mathrice (INSMI),
Mésocentre PHYMATH (Ecole polytechnique)

Agenda

- Context
- Motivation
- CMT
    - Workflow
    - Server
    - Boot
    - Live
    - Script
    - Systems
- What’s next?
- Questions

Context

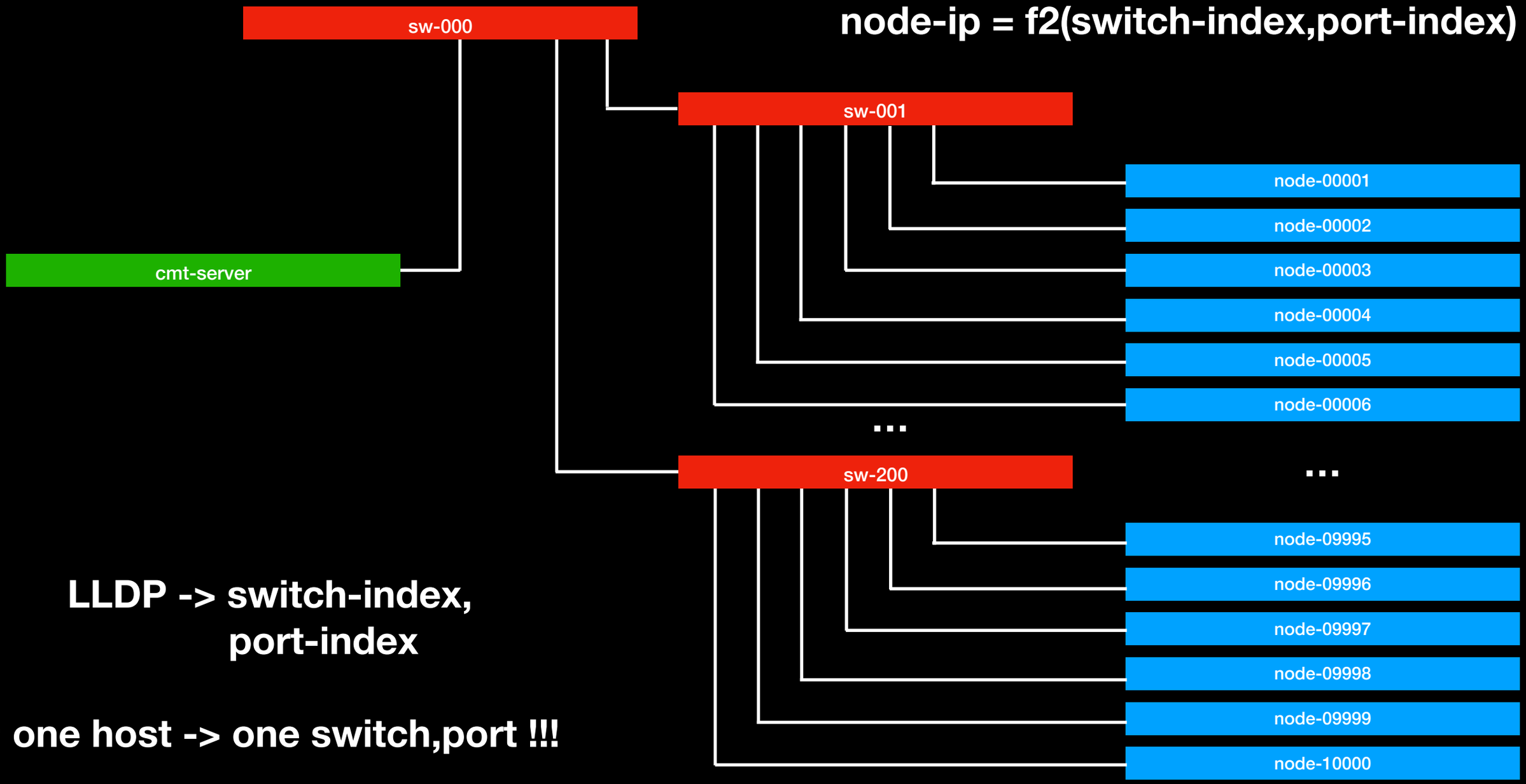
- PHYMATH : CMLS, CPHT, CMAP (Ecole polytechnique/CNRS)
    - RH
        - David DELAVENNAT (CMLS)
        - Jean-Luc BELLON, Danh PHAM-KIM (CPHT)
        - Sylvain FERRAND, Pierre STRAEBLER (CMAP)
- RT
    - 6 clusters, 6k cores, 10 years
        - Serviware Cluster Tool
        - Bright Cluster Manager
        - Trinity

Motivation

- minimize costs
- minimize deployment time
- use open source tools maintainable in time
    - DontRepeatYourself
    - ***no over-engineering***
    - automation, not obfuscation
    - ASR friendly -> team adoption rate
- continuous integration + server unit tests
- stateless computing

CMT-Workflow

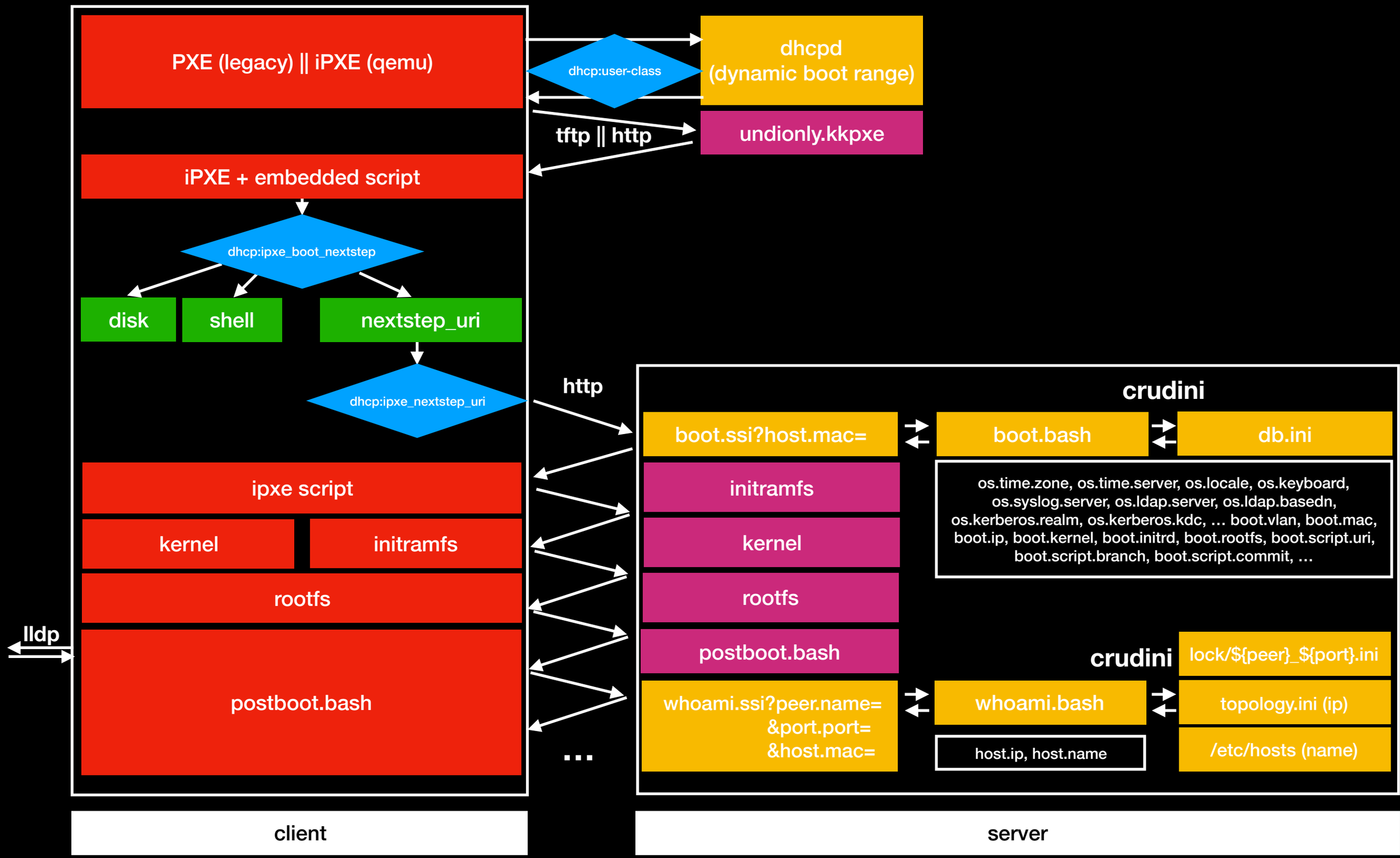
node-name = f1(switch-index,port-index)
node-ip = f2(switch-index,port-index)



LLDP -> switch-index,
port-index

!!! one host -> one switch,port !!!

CMT-Workflow



CMT-Server

- DNS : dnsmasq, bind+LDAP backend, PowerDNS
- TFTP : dnsmasq, tftp-server, tftp-http-proxy
- DHCP : dnsmasq, isc-dhcp + LDAP backend, Kea
- HTTP : Apache HTTP server + ServerSideInclude + Bash 4
- CLI: bash, crudini

CMT-Boot

- iPXE
  - VLAN supported
  - embedded script
    - http://ipxe.org/embed
    - http://ipxe.org/scripting
- http://plmlab.math.cnrs.fr/phymath/undionly.kkpxe
- http://man7.org/linux/man-pages/man7/dracut.cmdline.7.html

CMT-Live

- https://github.com/CentOS/sig-core-livemedia
- Usage
    - image the HPC nodes disks with qcow2 sparse images fetch through HTTP
    - update the DELL firmwares on hardware running unsupported OSes
        - StateLess, InMemory ThinStation
- Tool
    - livemedia-creator
        - use kickstart+ CMT-Modules
        - install a CentOS system into a VM
        - create a PXE-bootable kernel+ramfs+squashfs sparsed root filesystem

CMT-Live : installer.ks

```
keyboard ‘us’
cdrom
network —bootproto=dhcp —activate
repo --name="centos-updates" --baseurl="http://mirrors.ircam.fr/pub/centos/7/updates/x86_64/"
repo --name="epel"           --baseurl="http://dl.fedoraproject.org/pub/epel/7/x86_64/"
timezone Europe/Paris
selinux —disabled
bootloader —location=none
part / —size=8000 —fstype="ext4"
shutdown

%packages
%post
CMT_MODULE_PULL_URL=https://plmlab.math.cnrs.fr/cmt
CMT_MODULE_ARRAY=(
  base
  qemu
)
curl ${CMT_MODULE_PULL_URL}/stdlib/raw/master/bootstrap.sh | bash
cmt.stdlib.module.load_array CMT_MODULE_PULL_URL CMT_MODULE_ARRAY
%end
```

CMT-Live : station.ks

```
keyboard 'us'
cdrom
network —bootproto=dhcp —activate
repo --name="centos-updates" --baseurl="http://mirrors.ircam.fr/pub/centos/7/updates/x86_64/"
repo --name="epel"           --baseurl="http://dl.fedoraproject.org/pub/epel/7/x86_64/"
timezone Europe/Paris
selinux —disabled
bootloader —location=none
part / —size=8000 —fstype="ext4"
shutdown

%packages
%post
CMT_MODULE_PULL_URL=https://plmlab.math.cnrs.fr/cmt
CMT_MODULE_ARRAY=(
  base
  gui
  lmod-profile
)
curl ${CMT_MODULE_PULL_URL}/stdlib/raw/master/bootstrap.sh | bash
cmt.stdlib.module.load_array CMT_MODULE_PULL_URL CMT_MODULE_ARRAY
%end
```

CMT-Script

- no need to pre-register the mac address
    - using LLDP, calculate the node hostname (node-<index>) knowing the port index on the peer switch
    - if needed register the mac address into the CMT-Server
- wipefs
- qemu-img ~1.30min on 1G link
    - fetch sparse qcow2 image using HTTP
        - qemu-img --help | tail -1

          Supported formats: vvfat vpc vmdk vhdx vdi **ssh** sheepdog rbd raw host_cdrom host_floppy host_device file qed qcow2 qcow parallels nbd iscsi **gluster** dmg tftp ftps ftp **https** **http** cloop bochs blkverify blkdebug
    - write to disk
- resize2fs
    - grow the filesystem up to the physical disk constraint
- grubby
- kexec
    - execute the kernel+ramdisk just written on disk without rebooting from the live system

CMT-Systems

- VM creation using Packer from Hashicorp
  - produce sparse file (qcow2)
  - several provider
    - openstack
    - libvirt
    - virtualbox

CMT-Modules

- Lightweight Bash Deployment Tool
  - Continuous Integration friendly
    - works even with AlpineLinux
- software dependencies : bash-4, git
- modules are git repositories
- dependency management : tsort (topological sort)

What next?

- Question : can you upgrade the cluster’s libc?

  - Why not deploy hardware nodes on the fly knowing slurm job low level system needs?

Questions?

TP