Instrukcja obsługi klastra Slurm

Instrukcja dla administratora

Spis treści

Instrukcja instalacji systemu operacyjnego

Tworzenie nowej partycji na dysku

System operacyjny służący do obliczeń należy zainstalować na jednej z partycji głównego dysku komputera. Aby stworzyć nową partycję można posłużyć się programem *parted*. Poniższy przykład ilustruje jak dokonać zmiany rozmiaru istniejącej partycji o identyfikatorze idx, istniejącej na dysku sdX. Zmianę rozmiaru partycji należy rozpocząć od wykonania polecenia

```
parted /dev/sdX
```

Które otworzy konsolę narzędzia parted, umożliwiającego edycję partycji na dysku sdX. Następnie, należy dokonać zmiany rozmiaru partycji (ilość i rozmiar poszczególnych partycji można wylistować za pomocą polecenia *print*) oraz utworzenia nowej partycji, na której umieścimy nasz system. Aby tego dokonać należy wykonać poniższe polecenia, zastępując XXX wartością, w której chcemy aby partycja idx się kończyła. Naszą nowo utworzoną partycję będziemy oznaczać jako sdXY, oraz nadamy jej etykietę Slurm.

```
(parted) > resizepart idx XXXGB
(parted) > mkpart ext4 XXXGB XXX+70GB
(parted) > quit
e2fsck -f /dev/sdX
resize2fs /dev/sdX
mkfs.ext4 /dev/sdXY -L SLURM
```

Instalacja systemu na nowej partycji

Kiedy mamy już gotową pustą partycję to możemy rozpocząć proces instalacji systemu. Najpierw należy zamontować nową partycję wraz z wszystkimi potrzebnymi do sprawnego działania katalogami. Jako partycję sdYA oznacza się partycję, na której znajduje się EFI (można to sprawdzić za pomocą polecenia *lsblk*)

```
mount /dev/disk/by-label/SLURM /mnt
mkdir /mnt/dev /mnt/proc /mnt/sys /mnt/run /mnt/boot
/mnt/boot/efi
mount --bind /dev /mnt/dev
```

```
mount --bind /proc /mnt/proc
mount --bind /sys /mnt/sys
mount --bind /run /mnt/run
mount /dev/disk/by-label/EFI /mnt/boot/efi
```

Po zamontowaniu potrzebnych katalogów można przystąpić do instalacji systemu i najpotrzebniejszych bibliotek:

```
zypper --root /mnt ar http://download.opensuse.org/distribution/leap/15.6/repo/oss/ main
zypper --root /mnt refresh
zypper --root /mnt install --no-recommends bash coreutils glibc zypper rpm filesystem vim ca-certificates coreutils glibc-locale openssh wicked dhcp_client sssd shim kernel-default ca-certificates-mozilla timezone iproute2 less ethtool nettools net-snmp iputils kernel-firmware hostname
```

Należy obserwować wynik przetwarzania żądania. Może się pojawić komunikat *Please run "COMMAND" as soon as your system is complete.* W takim wypadku należy wykonać zalecaną komendę.

Konfiguracja nowego systemu

Po stworzeniu nowej partycji, zamontowaniu katalogów i zainstalowaniu systemu operacyjnego można przystąpić do jego konfiguracji. Aby ją wykonać można posłużyć się narzędziem chroot. Przed użyciem chroot należy jednak dokonać konfiguracji działania interfejsów sieciowych:

```
cp /etc/sysconfig/network/if* /mnt/etc/sysconfig/network/
cp /etc/udev/rules.d/70-persistent-net.rules
/mnt/etc/udev/rules.d/70-persistent-net.rules
cp /etc/resolv.conf /mnt/etc/resolv.conf
```

Następnie można wykonać polecenie chroot:

```
chroot /mnt
```

Następnie należy dokonać konfiguracji gruba

```
      echo -e 'GRUB_DISABLE_OS_PROBER="false"
      GRUB_TERMINAL="console"
      GRUB_TIMEOUT="6"
      GRUB_ENABLE_CRYPTODISK="n"
      GRUB_GFXMODE="auto"
      GRUB_DISABLE_RECOVERY="true"
      GRUB_DISTRIBUTOR=
      GRUB_DEFAULT="saved"
      SUSE_BTRFS_SNAPSHOT_BOOTING="false"
      GRUB_HIDDEN_TIMEOUT="0"
      GRUB_CMDLINE_LINUX_DEFAULT="console=ttyS4,115200n8  preempt=full
mitigations=auto quiet"
      GRUB_CMDLINE_XEN_DEFAULT="vga=gfx-1024x768x16"
      GRUB_BACKGROUND=
      GRUB_THEME=/boot/grub2/themes/openSUSE/theme.txt'              >
/etc/default/grub

      echo -e 'LOADER_TYPE="grub2"
      SECURE_BOOT="no"
      TRUSTED_BOOT="no"
      UPDATE_NVRAM="yes"' > /etc/sysconfig/bootloader
```

Oraz dodać wpis o nowo stworzonej partycji do pliku, z którego korzysta bootloader:

```
      echo -e "LABEL=SLURM   /   ext4    defaults  0 0" > /etc/fstab
```

Następnie należy dokonać konfiguracji ustawień serwisów

```
      systemctl enable wickedd
      systemctl enable wicked
      systemctl enable sshd
      timedatectl set-ntp true
      timedatectl set-timezone Europe/Warsaw
      timedatectl --adjust-system-clock
```

Następnie należy zaktualizować certyfikaty repozytoriów, co pozwoli na bezproblemowe korzystanie z narzędzia *zypper* w nowym systemie

```
update-ca-certificates
```

W ostatnim kroku należy ustawić nowe hasło dla użytkownika root, poleceniem

```
passwd
```

Ten krok kończy etap konfiguracji.

Czyszczenie po konfiguracji

Po zakończeniu konfiguracji należy wyjść z chroot'a poleceniem

```
exit
```

Następnie należy odmontować zamontowane wcześniej katalogi. UWAGA - nie wykonanie tego kroku grozi uszkodzeniem zawartości dysku!

```
umount -R /mnt
```

Ostatnim krokiem jest dodanie systemu do konfiguracji gruba:

```
echo -e 'menuentry "SLURM compute node" {
  load_video
  set gfxpayload=keep
  insmod gzio
  insmod part_gpt
  insmod ext2
  search --label --set=root SLURM
    linux    /boot/vmlinuz  root=/dev/disk/by-label/SLURM  quiet
splash=silent reboot=pci fastrestore quiet
  initrd /boot/initrd
}' >> /boot/grub2/grub.cfg
```

Uruchomienie nowego systemu

Następnie należy zrestartować komputer, wybierając jako system, który ma się uruchomić po restarcie nasz świeżo stworzony OS. Aby tego dokonać można posłużyć się poleceniem grub2-once. Jako X należy podstawić numer porządkowy, pod którym nowo zainstalowany system jest widoczny w grubie (można to sprawdzić poleceniem *grub2-once --list)*

```
grub2-once X
reboot
```

Po tym kroku komputer powinien uruchomić się w nowym systemie.

Instrukcja konfiguracji nowych węzłów

Konfiguracja uwierzytelniania za pomocą munge

Każdy komputer należący do klastra (w tym zarządca) muszą posiadać ten sam klucz munge - służy on do uwierzytelniania i jest niezbędny do prawidłowego działania klastra. Taki klucz można wygenerować za pomocą następującej komendy:

```
dd if=/dev/random bs=1 count=1024 > /etc/munge/munge.key
```

Po wygenerowaniu należy go rozpowszechnić na wszystkich komputerach klastra, na przykład za pomocą komendy:

```
scp /etc/munge/munge.key other@host:/folder/with/config/files/
```

zastępując nazwę hosta oraz ścieżkę do folderu poprawnymi wartościami.

Należy również rozpowszechnić klucz pozwalający na korzystanie z LDAP:

```
scp  /etc/pki/trust/anchors/cs.local.pem other@host:/folder/with/config/files/
```

Oraz ustawić prawidłową nazwę sieciową komputera poleceniem

```
hostnamectl set-hostname lab-xxx-ab
```

Przygotowanie komputera do bycia węzłem obliczeniowym

Potrzebne paczki

Należy zainstalować potrzebne paczki następującą komendą:

```
zypper install --no-recommends autofs nss_ldap openldap2-client
pam_ldap slurm slurm-munge slurm-pam_slurm sssd nfs-utils
```

W przypadku pojawienia się błędu

```
/usr/sbin/groupadd -r -g 65533 nogroup
groupadd: GID '65533' already exists
```

należy wybrać opcję ignore (i).

Konfiguracja

Aby dokonać konfiguracji komputera można wykorzystać skrypt konfiguracyjny setup.sh, który automatycznie wykona praktycznie całość potrzebnej konfiguracji dla node'a. W tym celu należy:

- Zalogować się na komputer jako konto z uprawnieniami administratora
- Pobrać z repozytorium pliki konfiguracyjne, znajdujące się w folderze system-config
- Przejść do miejsca na dysku, do którego pobraliśmy folder system-config
- Umieścić w nim klucz munge (np. za pomocą polecenia scp, lub generując nowy klucz - patrz wyżej, rozdział 'Klucz munge')
- Umieścić w nim certyfikat dla LDAPa w pliku cs.local.pem
- Uruchomić skrypt konfiguracyjny komendą sh setup.sh
- Wybrać tryb konfiguracji node
- UWAŻNIE PRZECZYTAĆ WYGENEROWANY PRZEZ SKRYPT OUTPUT - szczególnie zwrócić uwagę na linie zaczynające się od WARNING lub ERROR
- Spróbować z innej konsoli połączyć się do konfigurowanego komputera, żeby sprawdzić czy modyfikacja plików PAM nie wprowadziła niechcianych zmian
- W razie potrzeby wycofać zmiany skryptem rollback_setup.sh

Logowanie

Należy zadbać aby użytkownik slurm na komputerze zarządcy miał możliwość zdalnego logowania się na węzły. Można to zapewnić generując klucze ssh, na przykład w podany poniżej sposób.
Najpierw należy zalogować się jako użytkownik slurm, oraz przejść do odpowiedniego folderu:

```
su slurm
cd ~/.ssh
```

Następnie wykonać polecenia generujące klucze:

```
ssh-keygen -t ed25519 -f nazwa_klucza -N haslo_do_klucza
ssh-add ~/.ssh/nazwa_klucza
ssh-copy-id -i ~/.ssh/nazwa_klucza user@hostname

echo -e 'Host alias_nazwy_wezla
  HostName           adres_ip_wezla
  User               username
  IdentityFile        ~/.ssh/nazwa_klucza
  UserKnownHostsFile     ~/.ssh/nazwa_pliku
  Port               22
  PasswordAuthentication  no
  PreferredAuthentications publickey
 ' >>  ~/.ssh/config
```

Przygotowanie komputera do bycia zarządcą

Potrzebne paczki

Na komputerze zarządcy należy zainstalować wszystkie paczki potrzebne do działania node'a oraz dodatkowo:

```
zypper install --no-recommends mariadb grafana-server nginx prometheus-slurm-exporter
```

InfluxDb OSS w wersji 2 zgodnie z instrukcją oraz Influx CLI zgodnie z instrukcją

Konfiguracja

Rezerwacje

Podstawowe kroki konfiguracyjne są takie same jak dla node'a, tylko odpalając skrypt setup.sh należy po odpaleniu wybrać tryb master. Oprócz tego należy:

- Utworzyć cron job, który będzie odpalał skrypt czyszczący stan węzłów po rezerwacji. W tym celu należy wykonać komendę

```
crontab -e
```

a następnie dopisać do pliku crona linię:

```
0 22 * * * /etc/slurm/post_reservation_cleanup.sh
```

- Należy również ustawić rezerwację zasobów przypadającą na godziny, w których komputery będą wykorzystywane do zajęć dydaktycznych. Można to zrobić poleceniem:

```
scontrol create Reservation="zajecia_dydaktyczne"
StartTime=07:00:00 Duration=14:59:00 user=root
flags=ignore_jobs,daily Nodes=ALL
```

Sacctmgr

Aby dodać do bazy kont Slurma użytkowników LDAP, których nazwa zaczyna się od przedrostka inf, a numer indeksu ma 6 cyfr i rozpoczyna się od liczb z zakresu 140-170 można użyć poniższej pętli bashowej:

```
for idx in $(seq 140 170); do
   for user in $(ldapsearch -LLL -ZZ -x "(uid=inf$idx*)" dn |
awk -F'uid=|,' '{print $2}'); do
    if echo $user | grep -q -E 'inf[0-9]{6}'; then
      sacctmgr add user name=$user cluster=dcc account=ldap -i;
    fi
  done
done
```

Wykonanie powyższej pętli chwilę zajmie, więc należy uzbroić się w cierpliwość... Aby dodać pojedyńczych użytkowników można po prostu wykorzystać polecenie

```
sacctmgr add user name=MY_USER cluser=dcc account=MY_ACCOUNT
```

MySQL

Należy odpalić skrypt konfiguracyjny mysql, w którym należy ustawić hasło dla roota oraz usunąć tymczasowych, testowych użytkowników i struktury. Skrypt można odpalić poleceniem: $ mysql_secure_installation Następnie, należy zalogować się do

bazy danych i stworzyć w niej użytkownika i struktury, które będą przechowywały dane dotyczące wykorzystania zasobów:

```
$ sudo mysql -u root -p
> CREATE DATABASE slurm_acct_db;
> CREATE USER 'slurm'@'localhost' IDENTIFIED BY 'phahbaShei6f';
>    GRANT    ALL    PRIVILEGES    ON    slurm_acct_db.*    TO
'slurm'@'localhost';
> FLUSH PRIVILEGES;
```

InfluxDB

Aby móc monitorować wykorzystanie zasobów na przestrzeni czasu należy skonfigurować bazę danych Influx, która będzie przechowywać informacje o zużyciu zasobów w sposób, który umożliwia ich łatwy eksport do narzędzi do wizualizacji np. Grafany. Aby to zrobić należy postępować zgodnie z instrukcją z https://docs.influxdata.com/influxdb/v2/get-started/setup/

Grafana

Aby móc wyświetlać dane dotyczące zużycia zasobów w Grafanie należy dodać influx jako źródło informacji. W tym celu należy:

- Wejść na stronę grafany pod adresem http://<adres-komputera-zarzadcy>:3000
- Zalogować się jako admin (domyślne hasło admin)
- Wejść w ustawienia, a następnie w Data Sources
- Dodać data source - wybrać InfluxDB
- Jako url do Influx podać http://<adres-hosta-influx>:8086
- Jako bazę danych podać slurm
- Jako użytkownika podać slurm
- Podać hasło użytkownika slurm (ustawione w ramach wykonywania instrukcji paragraf wyżej)
- Zapisać źródło danych
- Stworzyć, korzystając z plików admin_dashboard.json oraz public_dashboard.json dashboardy, które będą wyświetlały interesujące nas informacje