# Notas Ocotillo y Ollama

April 16, 2024

## Contents



## 1 Ejecutar el servidor de Ollama en un nodo de Ocotillo

Consideramos que existe el usuario `eacuna` en Ocotillo.

```
ssh eacuna@148.225.111.150
```

Puedes establecer la siguiente relación en `/etc/hosts` de tu local:

```
148.225.111.150 ocotillo.acarus
```

El comando queda entonces como:

```
ssh eacuna@ocotillo.acarus
```

Para ejecutar Ollama en un nodo del cluster, debemos utilizar un script de Slurm, el manejador de colas que utiliza ACARUS.

## 1.1 Script de SLURM para servidor Ollama

Consideramos el siguiente script de Bash llamado `ollama-serve.slrm`.

```
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --job-name=ollama-serve
#SBATCH --ntasks=40
#SBATCH --time=24:00:00
#SBATCH --partition=general
#SBATCH --constraint=broadwell

cluster=$(hostname -f)
echo -e "
Directorio compartido de modelos = /LUSTRE/home/mlg/ollama/models

Para conectarse a esta instancia del servidor Ollama se debe invocar
con la variable de entorno

OLLAMA_HOST=${cluster}:11434

Por ejemplo, para correr el chat con el modelo Llama 2 con 7B
parámetros debemos ejecutar:

$ OLLAMA_HOST=${cluster}:11434 ollama run llama2:7b
"

OLLAMA_DEBUG=1 \
    OLLAMA_MODELS=/LUSTRE/home/mlg/ollama/models \
    OLLAMA_HOST=0.0.0.0 \
    ollama serve
```

## 1.2 Ejecutar servidor

Desde tu cuenta de Ocotillo, ejecutar el script anterior:

```
sbatch ollama-serve.slrm
```

Se mostrará un mensaje diciendo algo como:

```
Submitted batch job 58630
```

El número al final del mensaje es el identificador del proceso de Ollama.

Un archivo llamado `slurm-<id>.out` será creado, en este ejemplo es `slurm-58630.out`.

Al inicio del contenido de este archivo se encuentra la variable de entorno que debemos usar para conectarnos al servidor de OLLAMA, por ejemplo, usando el comando:

```
head slurm-58630.out
```

obtenemos:

```
Directorio compartido de modelos = /LUSTRE/home/mlg/ollama/models

Para conectarse a esta instancia del servidor Ollama se debe invocar
con la variable de entorno

OLLAMA_HOST=nodo18.ocotillo.unison.mx:11434
```

### 1.3 Probando conexión con el servidor de Ollama

Para probar que estos pasos han funcionado, ejecutamos una inferencia simple sobre el modelo `gemma:2b`, especificando a `ollama` el valor apropiado de `OLLAMA_HOST`.

```
OLLAMA_HOST=nodo18.ocotillo.unison.mx:11434 \
    ollama run \
    gemma:2b \
    "Escribe los primeros 5 números primos" \
    --verbose
```

### 1.4 Gestión de la cola con Slurm

Puedes determinar qué tareas hay en la cola con el comando `squeue`. Si quieres filtrar únicamente las que ha enviado un usuario puedes usar `squeue -u <usuario>`.

Una vez que terminemos de utilizar el servidor de Ollama, debemos cancelar la tarea de la cola utilizando el identificador de tarea:

```
scancel 58630
```

## 2 Establecer un tunel entre Ocotillo y local

Al ejecutar el servidor de Ollama en Ocotillo e identificar la dirección del nodo, podemos salir de la conexión SSH y abrir una nueva con un tunel que redireccione las peticiones locales al puerto 11434 para la dirección del nodo en el mismo puerto:

```
ssh -L11434:nodo18.ocotillo.unison.mx:11434 \
    eacuna@ocotillo.acarus \
    -o ServerAliveInterval=60
```

Debemos evitar cerrar esta terminal mientras utilizamos Ollama localmente. Al acceder a `localhost:11434` utilizaremos el servidor de Ollama corriendo en Ocotillo.

## 3 Medición de tiempos

Comando:

```
ollama run mixtral "Escribe los primeros 100 números primos" --nowordwrap --verbose
```

| máquina | modelo | total duration | load duration | prompt count | prompt duration | promp |
|---|---|---|---|---|---|---|
| furiosa | mixtral | 2m25.584702349s | 16.105427486s | 24 token(s) | 912.969ms | 26.29 |
| ocotillo | mixtral | 3m26.917136239s | 54.706411014s | 24 token(s) | 3.328061s | 7.21 t |
| furiosa | mistral:7b | 45.295963367s | 874.828221ms | 23 token(s) | 345.833ms | 66.51 |
| ocotillo | mistral:7b | 32.959572945s | 2.089187529s | 23 token(s) | 866.565ms | 26.54 |
| furiosa | gemma:2b | 4.704500452s | 860.642973ms | 20 token(s) | 210.688ms | 94.93 |
| ocotillo | gemma:2b | 8.961845446s | 4.355539941s | 20 token(s) | 289.785ms | 69.02 |
| furiosa | gemma:7b | 1m20.937150876s | 1.154339854s | 20 token(s) | 378.748ms | 52.81 |
| ocotillo | gemma:7b | 1m45.699217015s | 8.135698374s | 20 token(s) | 928.112ms | 21.55 |