[Gentoo mini Howto] Installation d'Ollama avec optimisations matérielles

---

Introduction

Ce guide vous aidera à installer et configurer Ollama sur une machine Gentoo Linux avec des optimisations matérielles spécifiques. Ollama etant une application qui nécessite des ressources matérielles optimales pour fonctionner efficacement.

Prérequis

- Une machine Gentoo Linux installée.
- Accès root ou sudo.
- Connaissances de base de l'administration système Linux.

Étape 1 : Mise à jour du système

Avant de commencer, assurez-vous que votre système est à jour :

```
# emerge -uavDN world
```

Étape 2 : Installation des dépendances

Installez les paquets nécessaires pour Ollama :

```
# emerge dev-util/cmake dev-util/ninja sys-devel/clang sys-devel/llvm
```

Étape 3 : Configuration des optimisations matérielles

Optimisation CPU

```
# emerge -av cpuid2cpuflgas
# echo "sci-ml/ollama $(cpuid2cpuflags)" >> /etc/portage/package.use/ollama
```

Optimisation GPU

- Nvidia Cuda compatibles cards : USE_FLAGS="cuda"
- AMD ROCm compatible cards : USE_FLAGS="rocm"

Selon votre cas : ajoutez les lignes suivantes dans le fichier de configuration **/etc/portage/package.use/ollama** :

```
# echo "sci-ml/ollama cuda" >> /etc/portage/package.use/ollama
```

Ou

```
# echo "sci-ml/ollama rocm" >> /etc/portage/package.use/ollama
```

- Optionnellement vous pouver ajouter les **use flags mkl et blas** pour optimiser les calculs mathématiques. Ces logiciels sont disponibles dans le dépôt gentoo, mais soumis à une **licence propriétaire** :

```
# echo "sci-ml/ollama mkl blas" >> /etc/portage/package.use/ollama
# echo "sci-libs/mkl " >> /etc/portage/package.license/mkl
```

Etape 4 : Installation de la surcouche gentoo/guru

Le dépot gentoo/guru est une surcouche gentoo qui contient des paquets non officiels, mais maintenus par la communauté Gentoo Linux. A ce jour le paquet **ollama** est disponible dans ce dépôt.

```
# emerge -av eselect-repository
# eselect repository enable guru
# emerge --sync
```

Étape 5 : Compilation d'Ollama depuis les sources

Portage se chargera d'installer les dépendances requises, soyez attentif au démasquage des paquets sous license.

Chipset Intel 13th/CUDA 12.4 Nvidia RTX 4060 par l'exemple :

```
strix # emerge -av ollama

Calculating dependencies... done!
Dependency resolution took 1.46 s (backtrack: 0/20).

[ebuild   R   ~] sci-ml/ollama-0.9.6::guru  USE="blas cuda mkl -rocm" AMDGPU_TARGETS="-gfx9(

Total: 1 package (1 reinstall), Size of downloads: 0 KiB

Would you like to merge these packages? [Yes/No]
```

Chipset AMD ROCm :

BETA_TESTER_NEEDED !!! Ne disposant **pas de matériel AMD ROCm compatible**, il serait intéressant d'avoir un **retour d'expérience** sur la compilation d'Ollama avec le **support AMD ROCm**. Si vous êtes intéressé, n'hésitez pas à me contacter à cette adresse : contact@pingwo.org

Chipset Raspberry Pi 4/5 IA embarquée :

Version binaire en cours de développement. Je vous tiendrai informé de l'avancement du projet.