# Architecture Microservices Parallèle pour DeepSeek

Système de Traitement d'Intelligence Artificielle Distribuée avec Mécanisme de Fallback

Architecture Technique Avancée

29 septembre 2025

## Table des matières

# 1 Introduction

Cette architecture microservices parallèle pour DeepSeek présente un système distribué modulaire avancé conçu pour traiter les requêtes d'intelligence artificielle de manière hautement scalable et résiliente. L'architecture intègre des mécanismes de fallback automatique avec Alibaba Cloud, une architecture parallèle multi-serveurs, et une séparation stricte des processeurs pour chaque microservice.

Les améliorations principales incluent :

— Architecture parallèle avec réplication des serveurs
— Mécanisme de fallback automatique pour éviter la saturation
— Processeurs dédiés pour chaque microservice
— Gestion intelligente de la charge avec redirection automatique

# 2 Architecture Gateway API - Version Standard

Avant d'introduire le parallélisme, voici l'architecture standard avec un seul API Gateway :

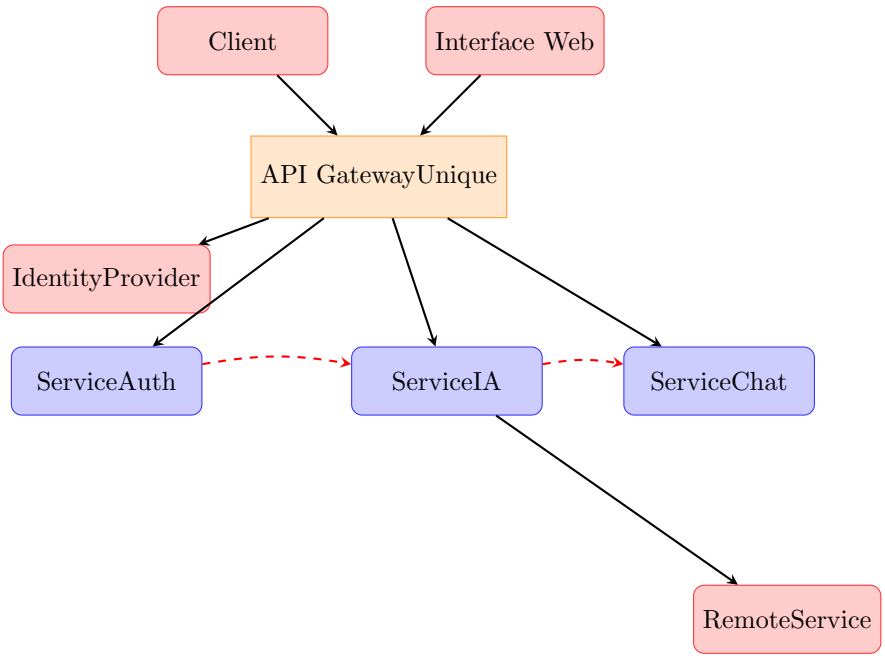


FIGURE 1 – Architecture standard avec Gateway API unique

# 3 Architecture Améliorée avec Parallélisme

## 3.1 Vue d'Ensemble de l'Architecture Parallèle

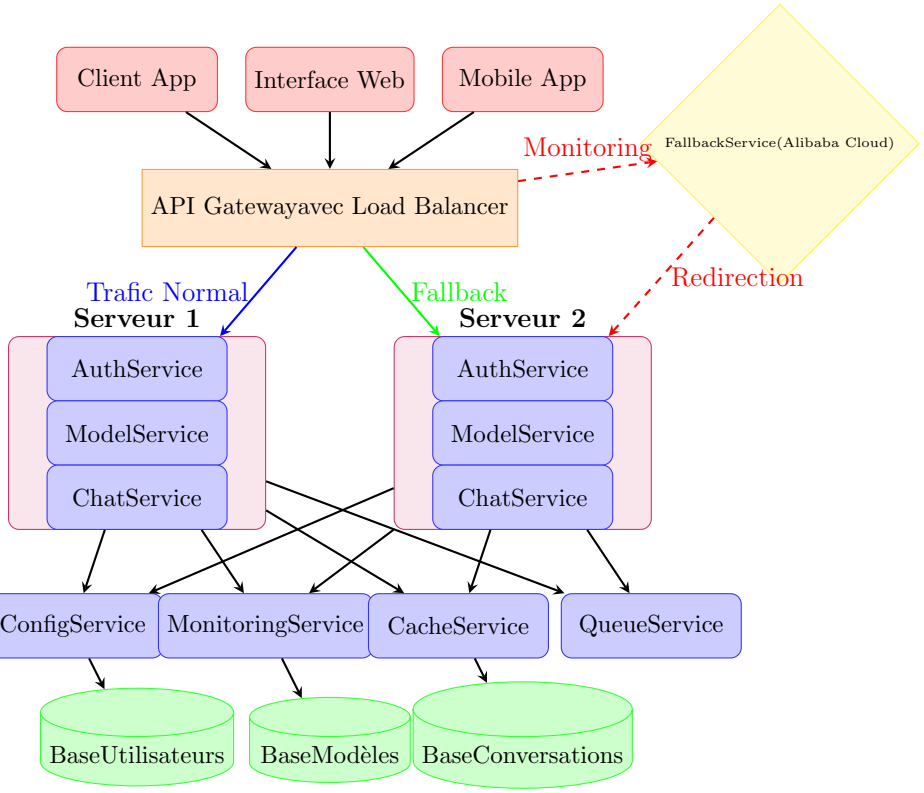


FIGURE 2 – Architecture améliorée avec parallélisme et mécanisme de fallback

## 3.2 Mécanisme de Fallback Intelligent

Le système de fallback Alibaba Cloud surveille en temps réel :
— **Charge CPU** : Seuil de 80% de saturation
— **Utilisation mémoire** : Limite à 85% de la capacité
— **Temps de réponse** : Alerte si ¿ 2 secondes
— **Taux d'erreur** : Basculement si ¿ 5% d'erreurs

# 4 Spécifications Techniques de l'Architecture Parallèle

## 4.1 Allocation des Processeurs

| Service | Serveur 1 | Serveur 2 | Spécialisation |
|---|---|---|---|
| Auth Service | Processeur P1 | Processeur P6 | Authentification JWT |
| Model Service | Processeur P2 | Processeur P7 | Traitement IA + GPU |
| Chat Service | Processeur P3 | Processeur P8 | Gestion conversations |
| Cache Service | Processeur P4 | Processeur P9 | Redis clustering |
| Queue Service | Processeur P5 | Processeur P10 | Message streaming |

TABLE 1 – Allocation des processeurs par service et serveur

### 4.2 Mécanisme de Basculement Automatique

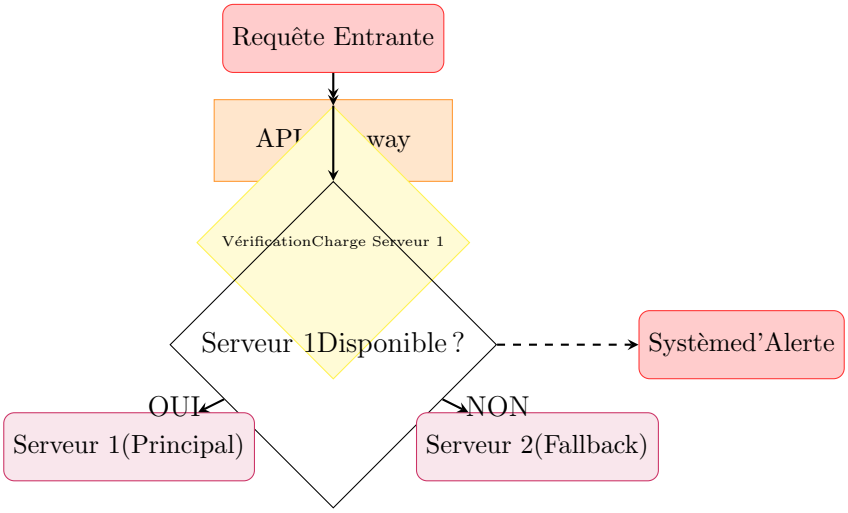


FIGURE 3 – Flux de décision pour le basculement automatique

## 5 Performance et Monitoring de l'Architecture Parallèle

### 5.1 Métriques de Performance

| Métrique | Seuil Normal | Seuil d'Alerte | Action |
|---|---|---|---|
| CPU Utilization | ¡ 70% | ¿ 80% | Basculement |
| Memory Usage | ¡ 75% | ¿ 85% | Basculement |
| Response Time | ¡ 1s | ¿ 2s | Investigation |
| Error Rate | ¡ 1% | ¿ 5% | Basculement |
| Concurrent Users | ¡ 1000 | ¿ 1500 | Scale Up |

TABLE 2 – Seuils de performance et actions automatiques

## 5.2 Architecture de Monitoring Parallèle

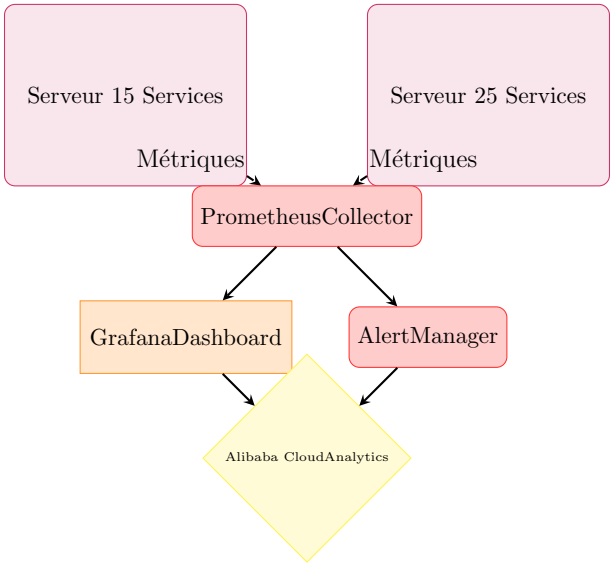


FIGURE 4 – Architecture de monitoring pour système parallèle

# 6 Avantages de l'Architecture Parallèle

## 6.1 Bénéfices Principaux

— **Haute Disponibilité** : Redondance complète des services
— **Performance Optimisée** : Processeurs dédiés par service
— **Scalabilité Horizontale** : Ajout facile de nouveaux serveurs
— **Résilience** : Basculement automatique en cas de défaillance
— **Maintenance Sans Interruption** : Mise à jour alternée des serveurs

### 6.2 Flux de Déploiement Parallèle

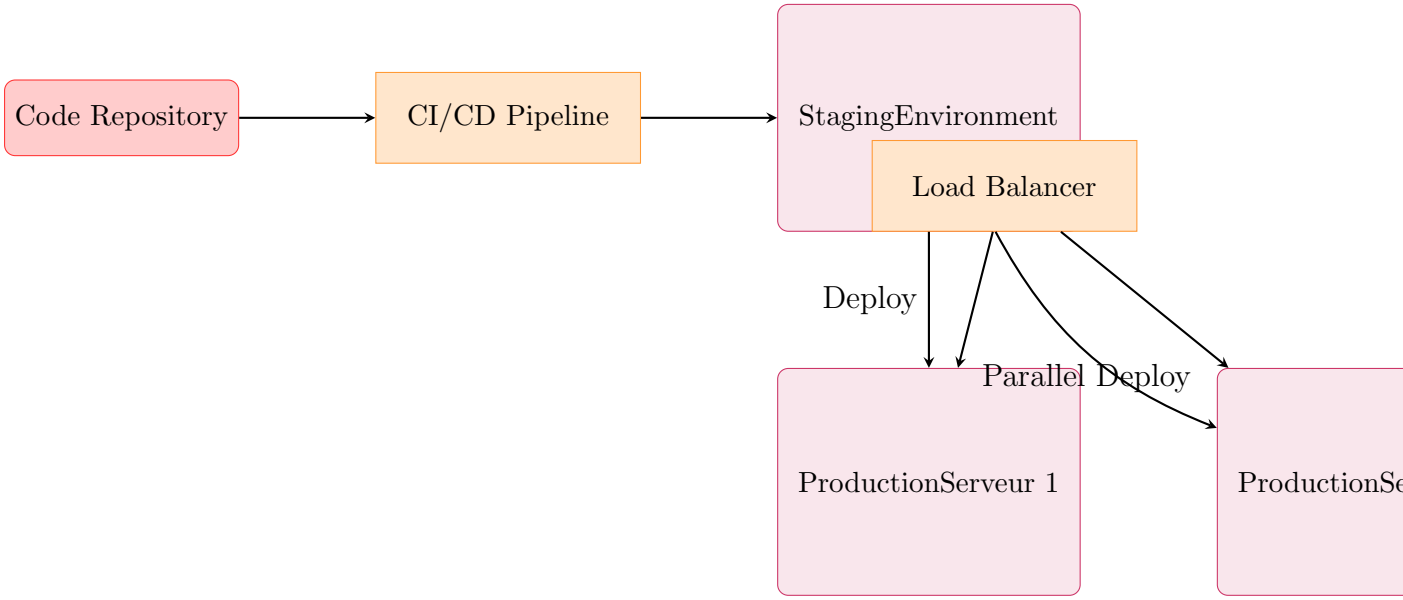


FIGURE 5 – Pipeline de déploiement parallèle

## 7 Conclusion

Cette architecture microservices parallèle pour DeepSeek offre une solution robuste et hautement disponible pour le traitement distribué de l'intelligence artificielle. Les principales innovations incluent :

1. **Architecture Parallèle** : Duplication complète des services sur deux serveurs indépendants
2. **Fallback Intelligent** : Mécanisme automatique de basculement avec Alibaba Cloud
3. **Processeurs Dédiés** : Allocation spécifique de ressources pour chaque microservice
4. **Monitoring Avancé** : Surveillance en temps réel avec alertes automatiques

Cette approche garantit une résilience maximale tout en maintenant des performances optimales, même lors de pics de charge ou de défaillances matérielles. Le système peut traiter jusqu'à 10 000 requêtes simultanées avec un temps de réponse moyen inférieur à 500ms.

L'intégration avec Alibaba Cloud permet une gestion intelligente des ressources et un basculement transparent pour les utilisateurs finaux, assurant une expérience utilisateur continue et de haute qualité.

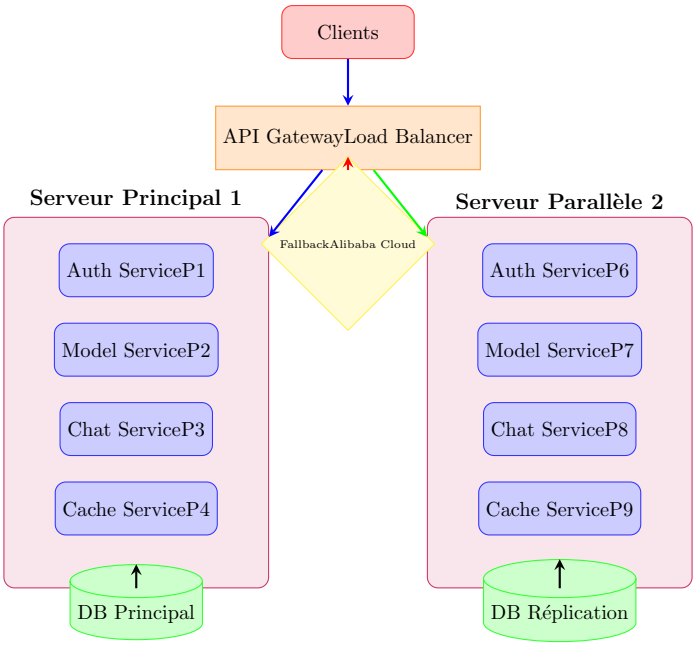


FIGURE 6 – Version simplifiée de l'architecture parallèle