Monitoring du noyau Linux

sur une architecture NUMA

Kevin Gallardo
Eric Lombardet
Pierre-Yves Péneau

Université Pierre et Marie Curie

12 Mai 2014

Introduction

Problématique

- architecture NUMA
- systèmes non efficaces
- gain de performances

Introduction

Problématique

- architecture NUMA
- systèmes non efficaces
- gain de performances

Objectifs

- évaluation d'activité
- mesures d'évènements

Architecture NUMA

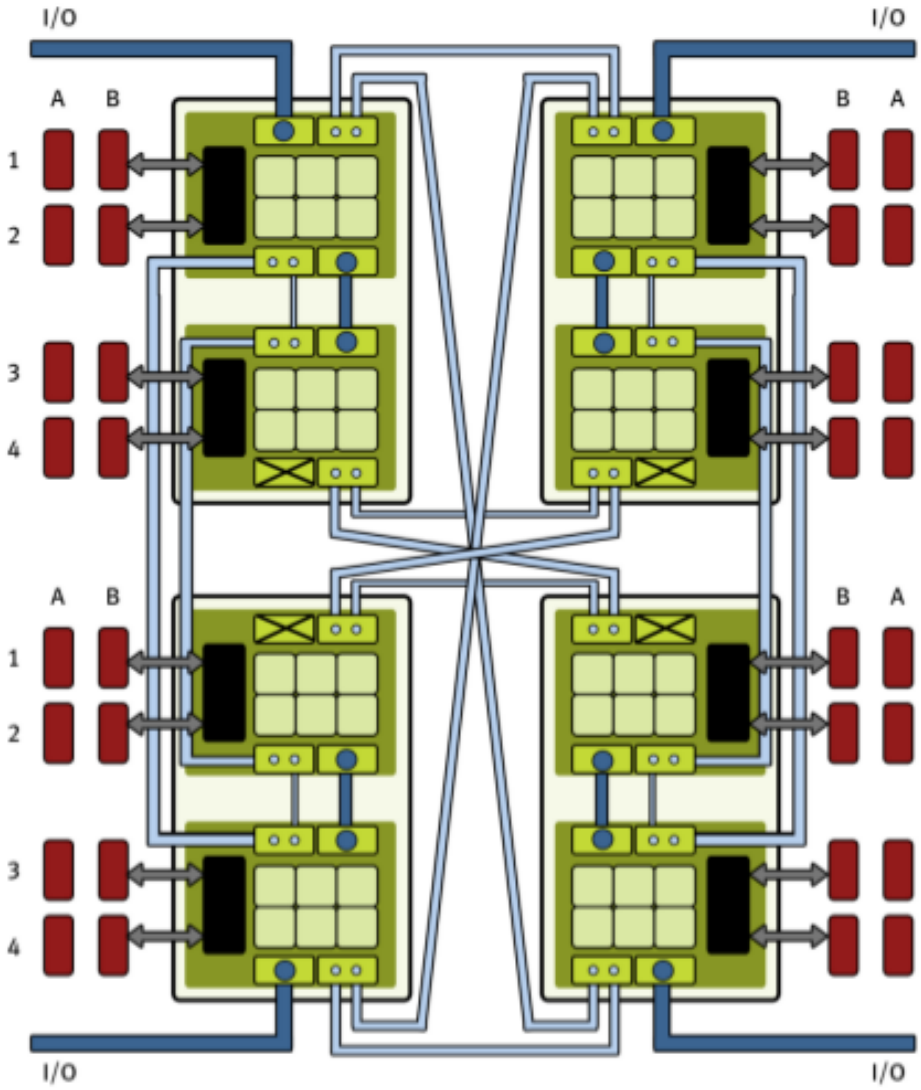
Présentation

Objectifs

- accélerer les temps de traitement
- répondre aux besoins d'applications spécifiques

Moyens mis en œuvre

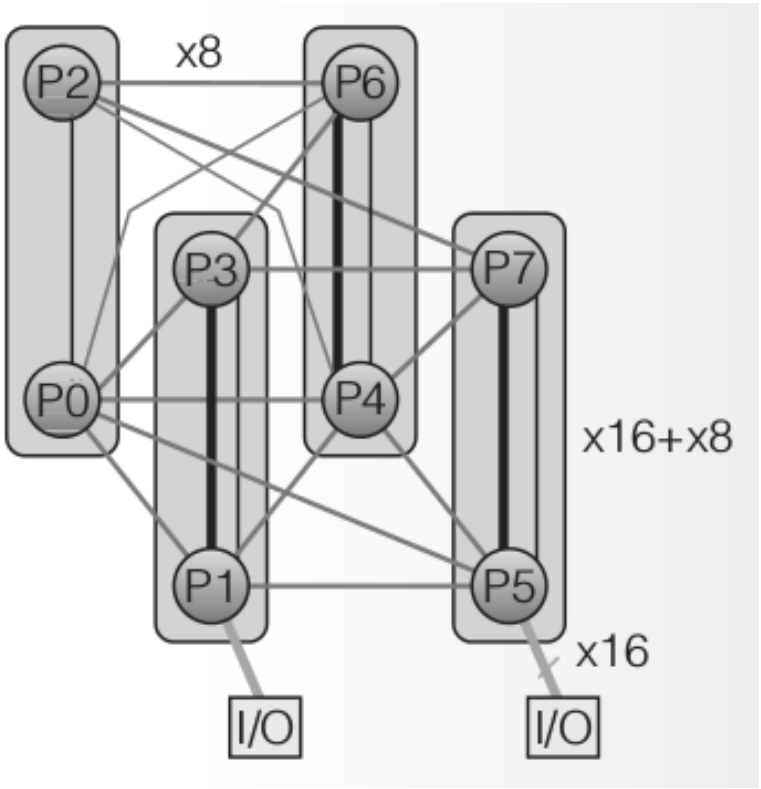
- découpe en noeuds
- placement des contrôleurs d'E/S
- liens d'interconnexions
- mise en place d'une topologie



Architecture NUMA

Enjeux

- placement mémoire
- placement des threads
- activité d'entrées/sorties



Infrastructure de tests

- utilisation mutualisée du Magny Cour $\rightarrow$ machines virtuelles
- compilation du noyau avec KGDB
- problème: aucune émulation du monitoring par Qemu

Infrastructure de tests

- utilisation mutualisée du Magny Cour $\rightarrow$ machines virtuelles
- compilation du noyau avec KGDB
- problème: aucune émulation du monitoring par Qemu

Conséquence

- Travail en réel sur le noyau pour 50% du projet

Monitoring

Qu'est-ce que c'est ?

- étude bas niveau du comportement matériel et système
- très utile pour le débugage ou l'optimisation poussée
- différentes solutions de monitoring existent
  - outils en mode utilisateur

Monitoring

Qu'est-ce que c'est ?

- étude bas niveau du comportement matériel et système
- très utile pour le débugage ou l'optimisation poussée
- différentes solutions de monitoring existent
  - outils en mode utilisateur
  - Performance Monitoring Counters

Monitoring

Qu'est-ce que c'est ?

- étude bas niveau du comportement matériel et système
- très utile pour le débugage ou l'optimisation poussée
- différentes solutions de monitoring existent
  - outils en mode utilisateur
  - Performance Monitoring Counters
  - Instruction Based Sampling (IBS)

Monitoring

Instruction Based Sampling - Présentation

- technologie AMD
- informations plus précises car IBS spécifique à une famille de processeur
- problème: plus difficile à mettre en place

Monitoring

Instruction Based Sampling - Fonctionnement

Principe :

- taguer aléatoirement une instruction (découpage des instructions)
- suivi de l'exécution
- deux types de mesures: fetch/execution sampling

Monitoring

Instruction Based Sampling - Utilisation

- beaucoup d'informations remontées par IBS
- sélection des plus utiles: cache hit/miss
- informations stockées dans des registres MSR

Monitoring

Instruction Based Sampling - Utilisation

- beaucoup d'informations remontées par IBS
- sélection des plus utiles: cache hit/miss
- informations stockées dans des registres MSR

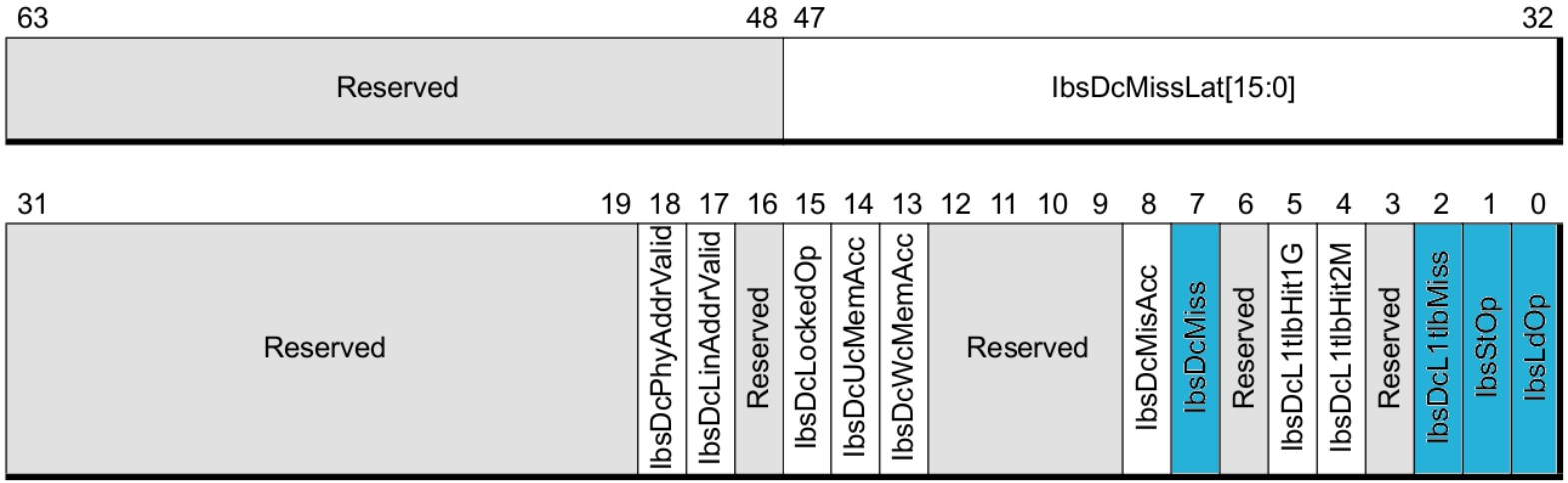


Figure: Schéma du registre MSR IbsOpData3

Monitoring

Mise en place(1)

- les interruptions matérielles - APIC

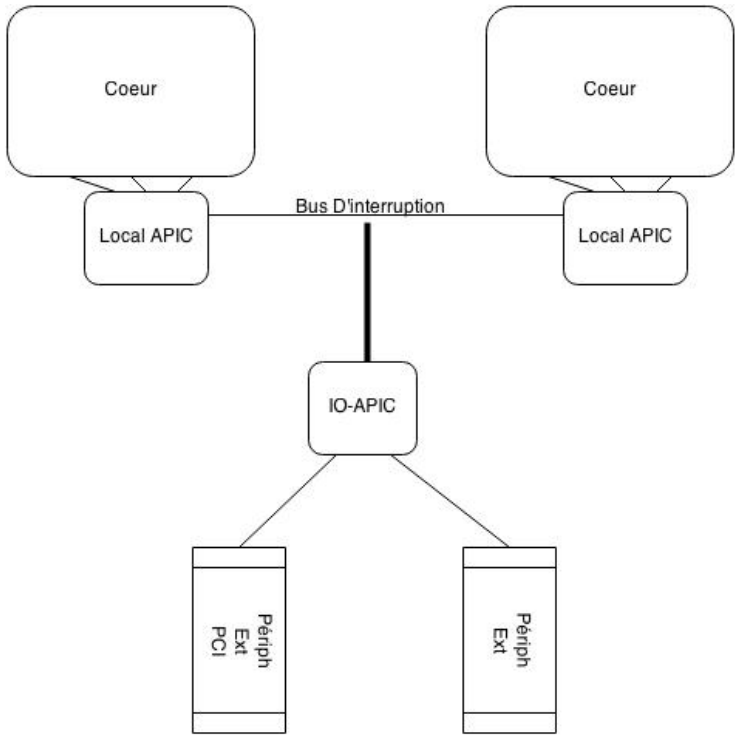


Figure: Schéma des composants matériels qui gèrent les interruptions

Monitoring

Mise en place(1)

- les interruptions matérielles - APIC

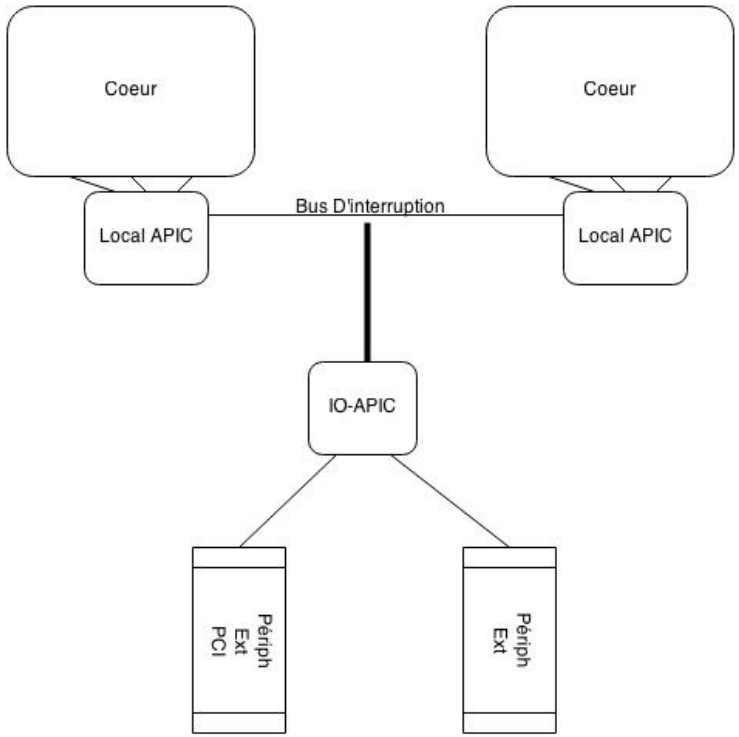


Figure: Schéma des composants matériels qui gèrent les interruptions

- interruptions IBS sont de type NMI (Non Maskable Interrupt)

Monitoring

Mise en place(2)

Pour configurer les mesures :

- configuration de l'APIC
  - informer l'APIC de la présence d'interruptions IBS
  - à faire pour chaque coeur
- enregistrement d'un handler NMI
  - appelé à chaque interruption IBS
  - récolte les informations dans les registres MSR

Monitoring

Mise en place(3)

Lancer les mesures (IbsOpCtl) :

- configurer le taux d'échantillonnage
- bit IbsOpEn $= 1$
- à faire sur le coeur du thread concerné

Monitoring

Instruction Based Sampling - Défauts

- overhead: traitement coûteux des mesures
- pas de vision d'ensemble

Mesures sur le noyau Linux

Chaleur d'un thread

- un compteur représente l'activité d'un thread
- différents critères d'activité:
  - état: (in)actif
  - taux d'utilisation mémoire
  - nombre d'entrées/sorties
  - commnunications entre threads
  - ...

Mesures sur le noyau Linux

Méthodes de tri envisagées

- nécessité d'une structure dédiée
- utilisation d'un tableau ou d'une liste chainée
  - insertion de nouveaux threads
  - difficulté à trouver les threads morts
  - tri peu performant

Mesures sur le noyau Linux

Méthodes de tri envisagées

- nécessité d'une structure dédiée
- utilisation d'un tableau ou d'une liste chainée
  - insertion de nouveaux threads
  - difficulté à trouver les threads morts
  - tri peu performant

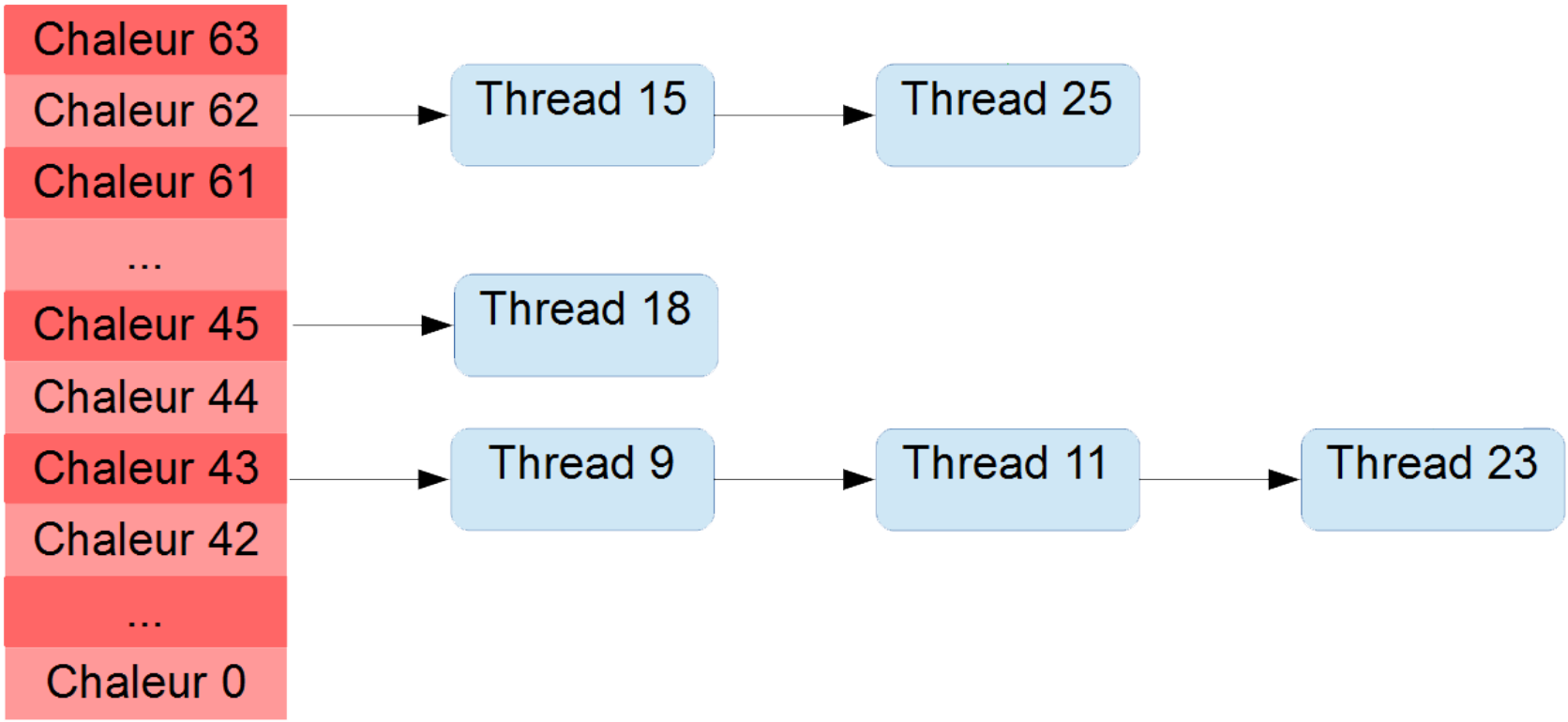
Conclusion

Solution abandonnée

Mesures sur le noyau Linux

Méthodes de tri envisagées

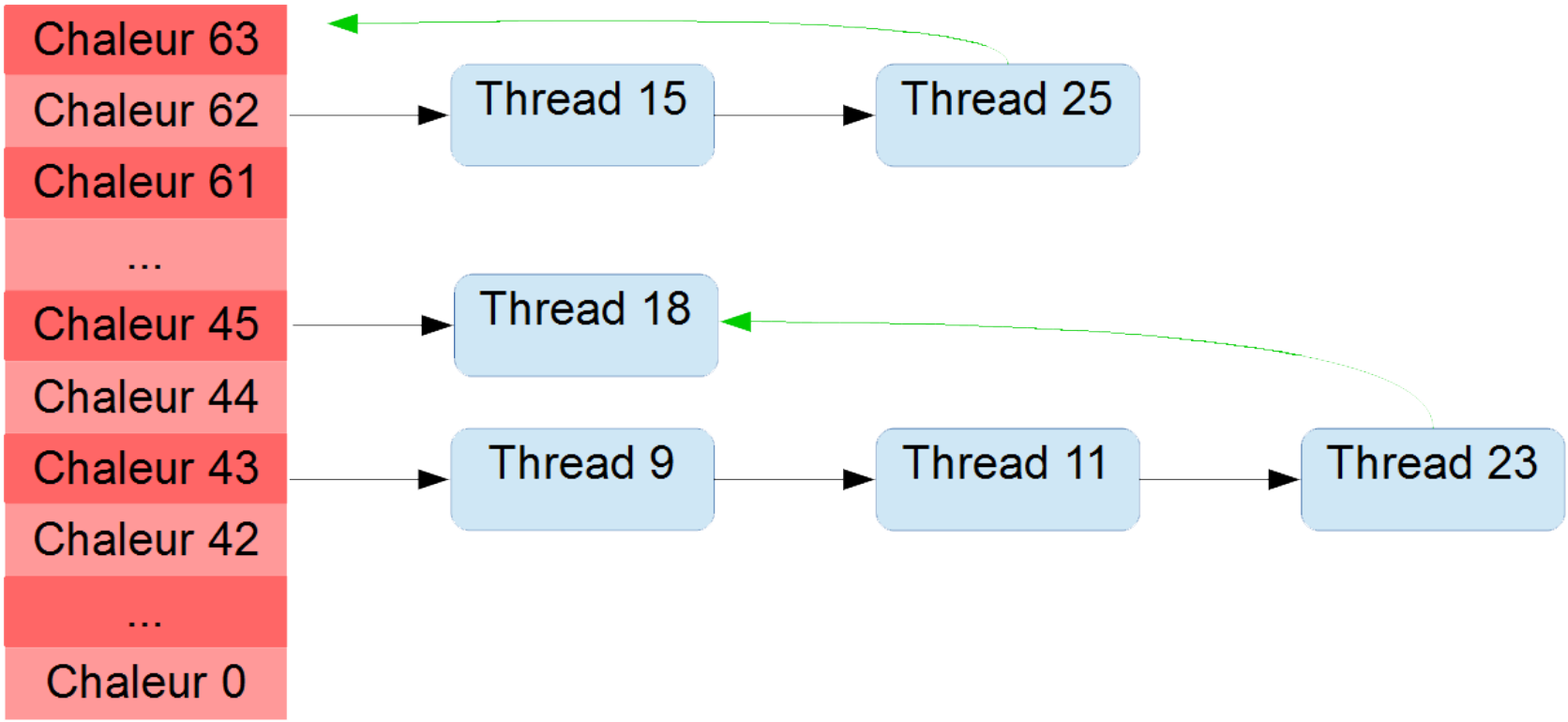
Utilisation d'un tableau de chaleur



Mesures sur le noyau Linux

Méthodes de tri envisagées

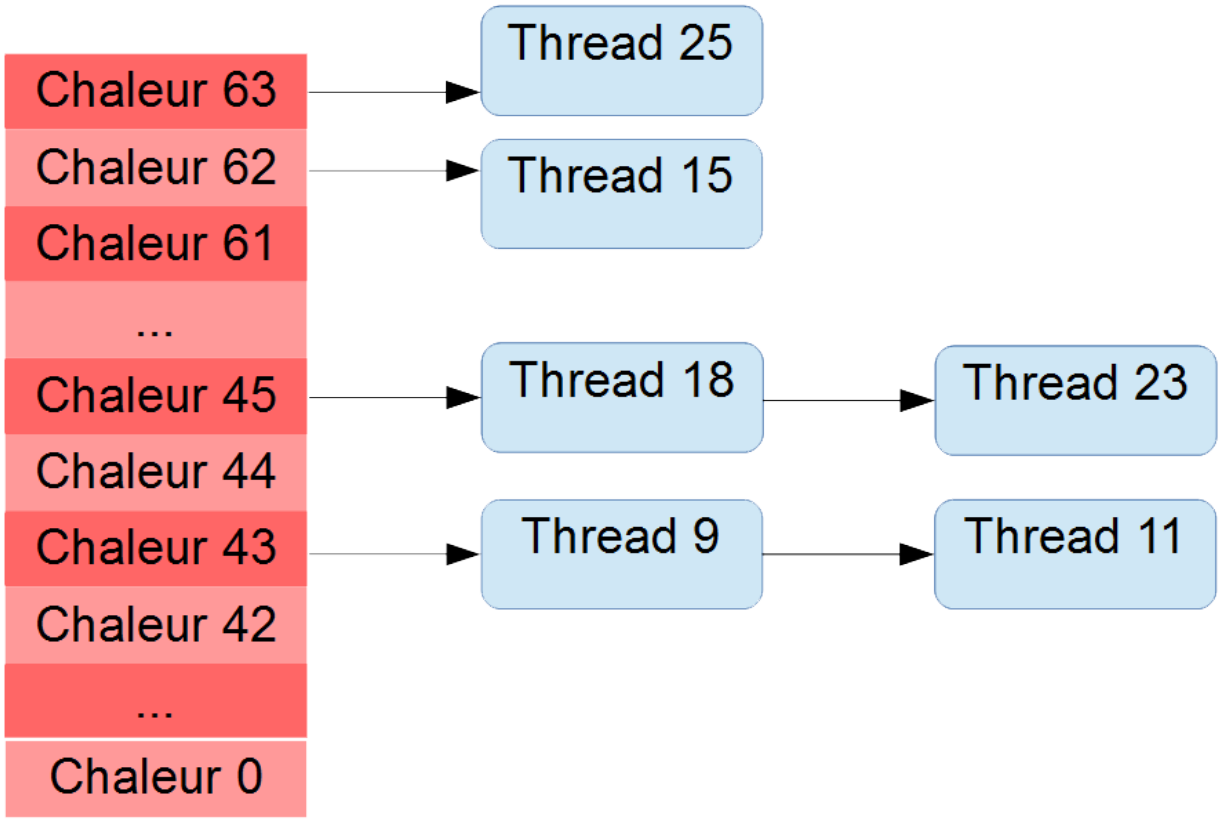
Utilisation d'un tableau de chaleur



Mesures sur le noyau Linux

Méthodes de tri envisagées

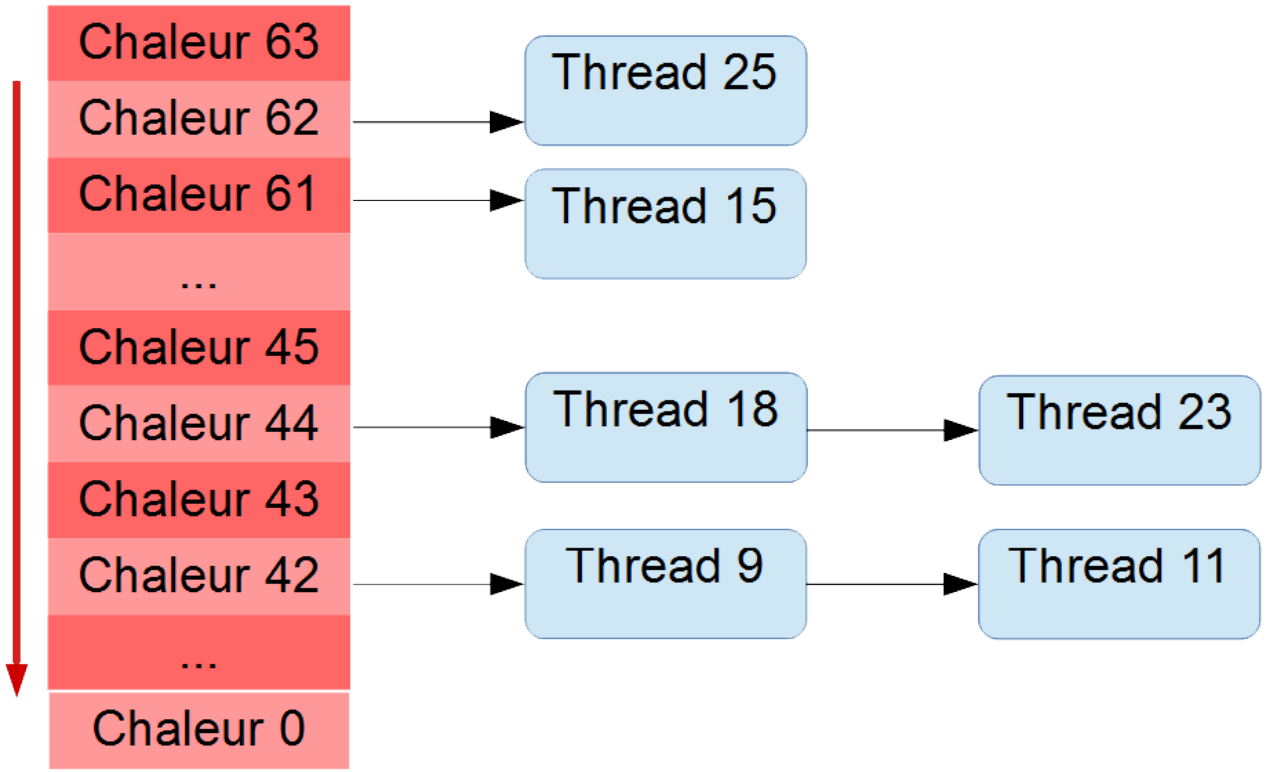
Utilisation d'un tableau de chaleur



Mesures sur le noyau Linux

Méthodes de tri envisagées

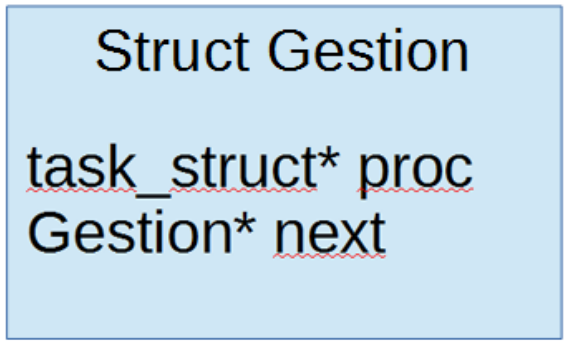
Utilisation d'un tableau de chaleur



Mesures sur le noyau Linux

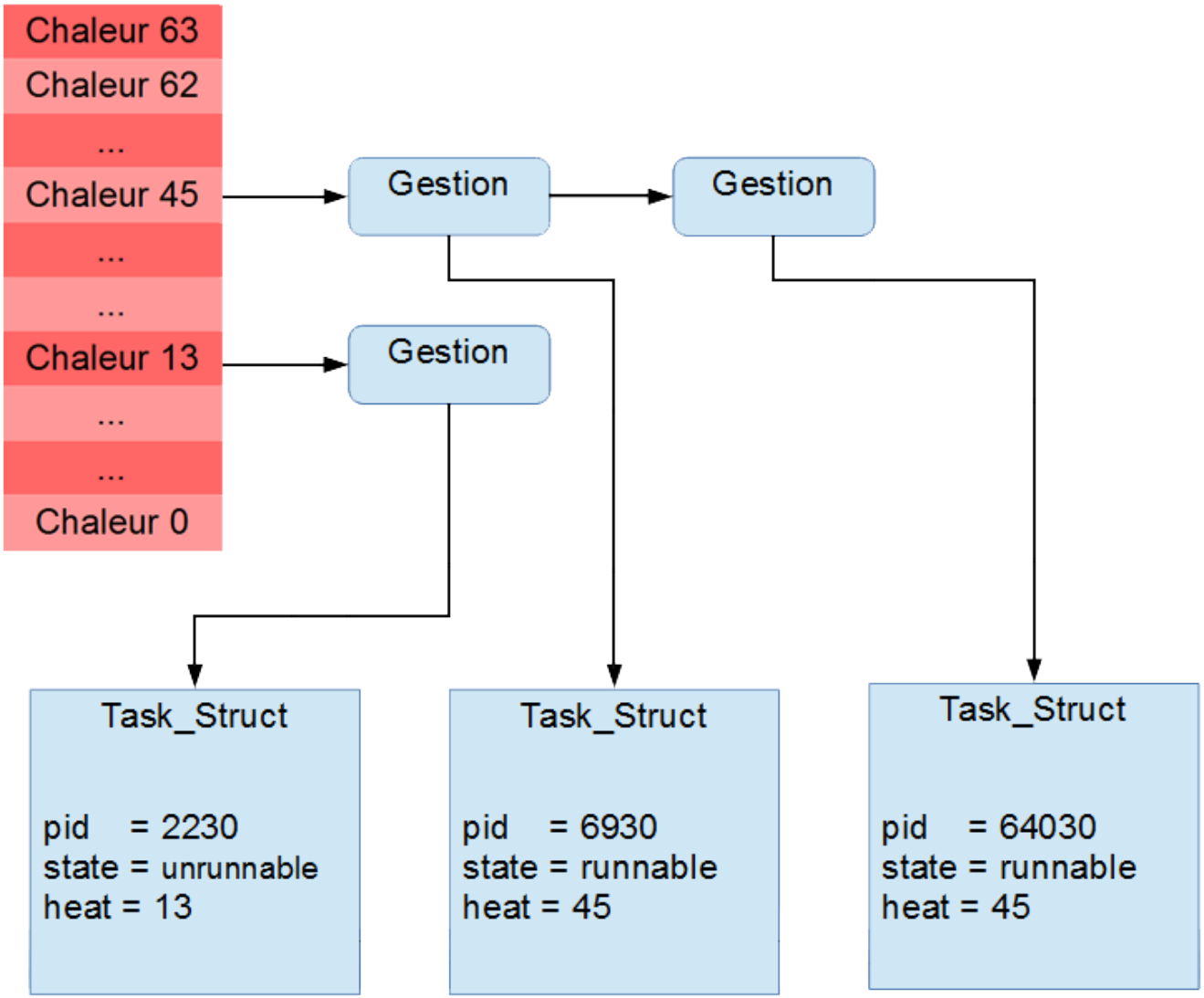
Solution retenue

- ajout du compteur dans la task_struct
- on conserve le tableau de chaleur précédent
- structure Gestion pour les listes



Mesures sur le noyau Linux

Solution retenue



Mesures sur le noyau Linux

Réalisation

Algorithme:

1. parcourir la task_struct
   a. si RUNNING $\rightarrow$ incrémentation du compteur de chaleur
   b. sinon décrémentation
2. stopper IBS
3. vider le tableau de chaleurs
4. générer le tableau de chaleurs
5. lancer les mesures sur les threads chauds

Mesures sur le noyau Linux

Réalisation

Optimisation

- Utilisation d'un facteur d'incrémentation et de décrémentation dynamique

Mesures sur le noyau Linux

Réalisation

Optimisation

- Utilisation d'un facteur d'incrémentation et de décrémentation dynamique

Problèmes

- pas d'IBS avec qemu $\rightarrow$ merge impossible sur Magny Cour

Conclusion

Apports personnels

- beaucoup de connaissances acquises
- utile pour l'année prochaine
- découverte d'une nouvelle architecture prometteuse

Conclusion

Apports personnels

- beaucoup de connaissances acquises
- utile pour l'année prochaine
- découverte d'une nouvelle architecture prometteuse

Ce qu'il reste à faire

- merger les deux parties du projet sur Magny Cour
- mettre en place un traitement des données
- améliorer l'algorithme de tri d'activités

THE END

To be continued...
No questions please.