Interférences temporelles au sein d'un GPU

T. Carle

Université de Toulouse, IRIT, CNRS

ETR, 10 Septembre 2026

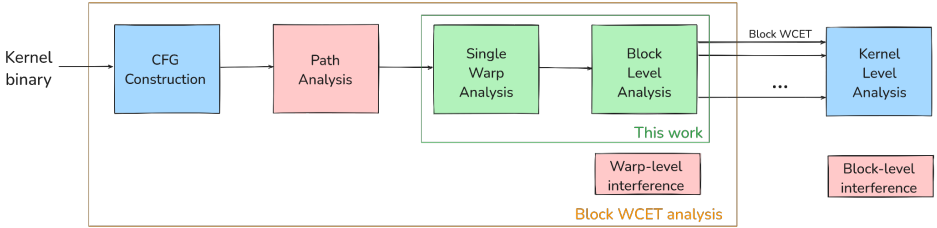






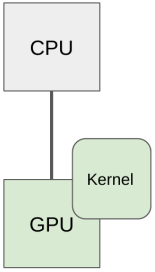
Objectifs

- Calculer le WCET de programmes accélérés par GPU
  - Analyse statique → WCET en isolation + interférences (préemptions, accès mémoire partagée)
  - Noyau → Bloc → Warp → Thread : quel est le WCET de base ? Qu'est-ce qu'une interférence ?
  - WCET du warp + interférences au niveau warp + interférences au niveau bloc

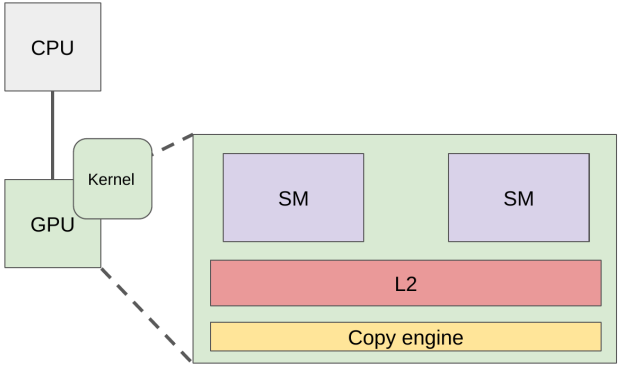


Reproduit depuis Jeanmougin et al. *Bounding the WCET of a GPU Thread Block with a Multi-Phase Representation of Warps Execution*. ECRTS 25

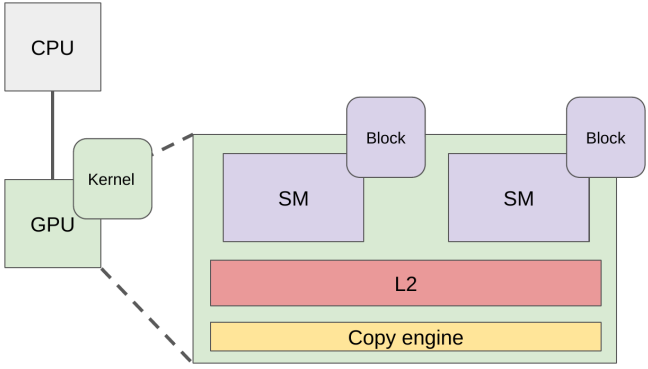
Introduction aux GPU : vue du programmeur



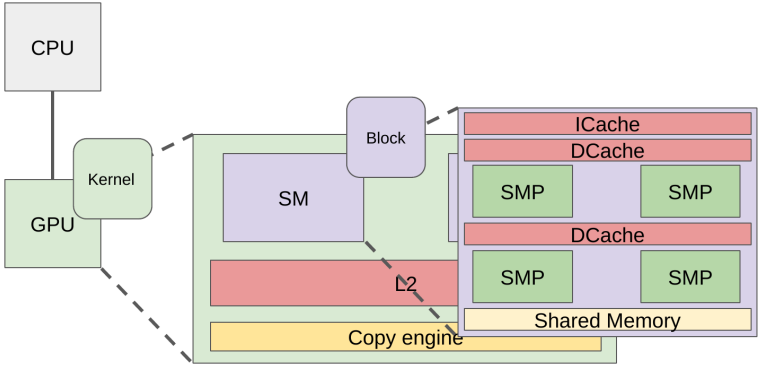
Introduction aux GPU : vue du programmeur



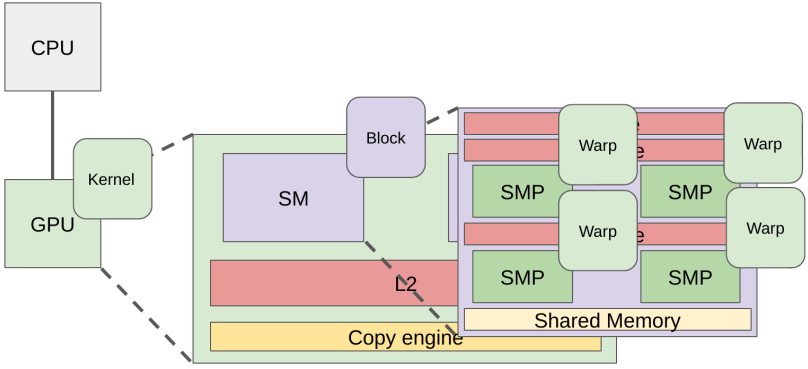
Introduction aux GPU : vue du programmeur



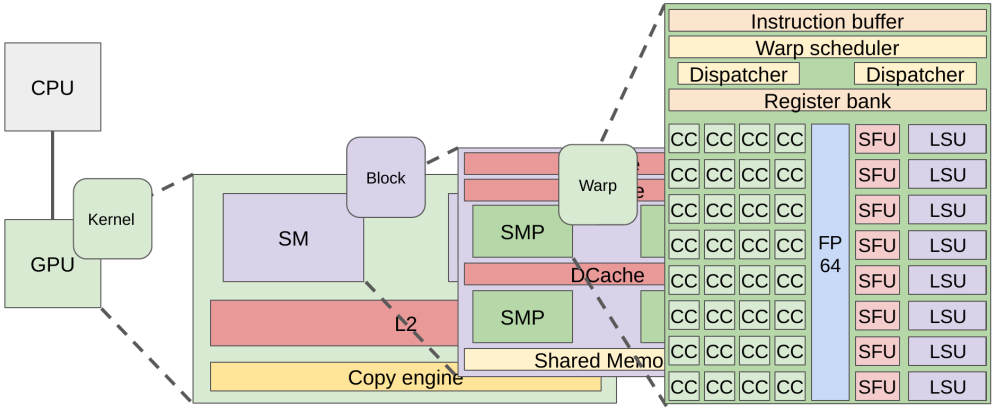
Introduction aux GPU : vue du programmeur



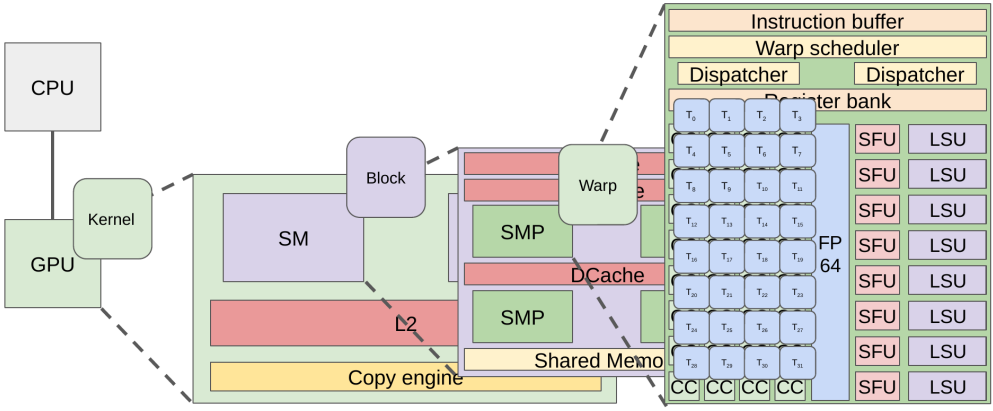
Introduction aux GPU : vue du programmeur



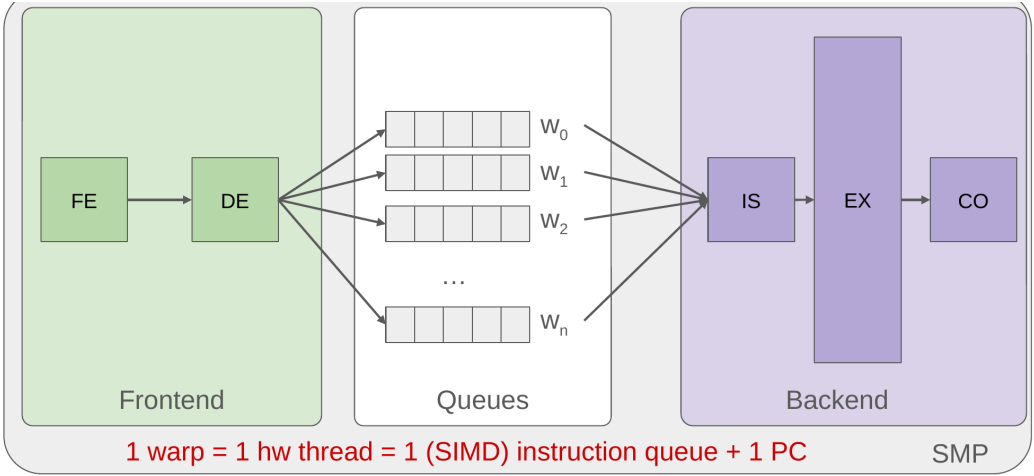
Introduction aux GPU : vue du programmeur



Introduction aux GPU : vue du programmeur



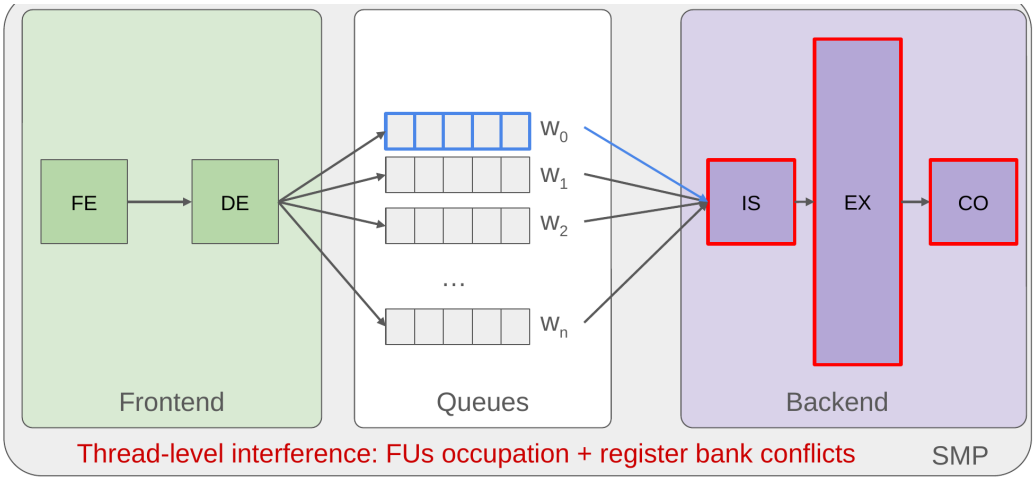
Introduction aux GPU : vue architecturale



Analyse du WCET d'un warp en isolation

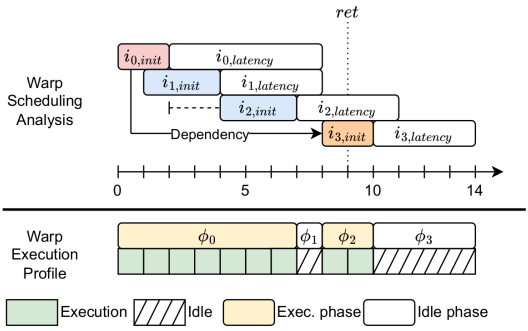
- Calculer le WCET du warp en isolation, considérer le reste comme interférence

Analyse du WCET d'un warp en isolation

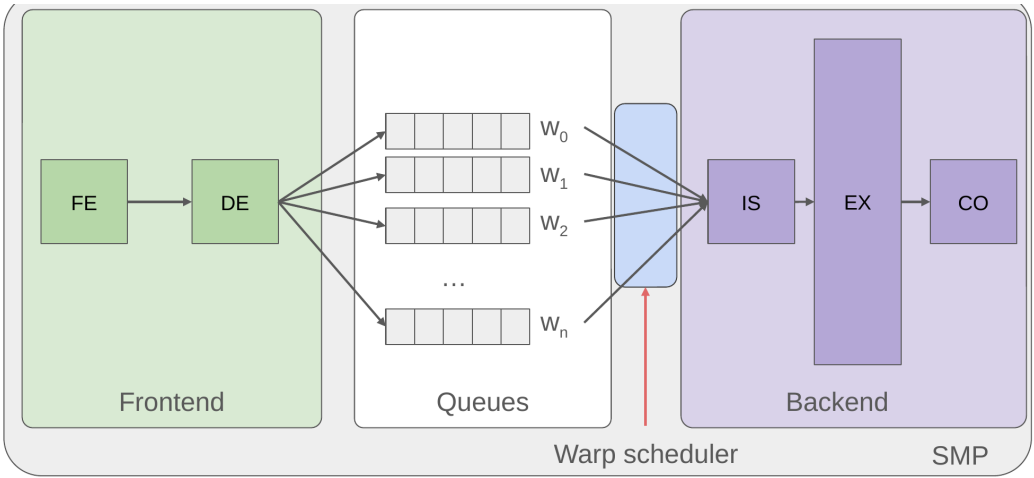


Analyse du WCET d'un warp en isolation

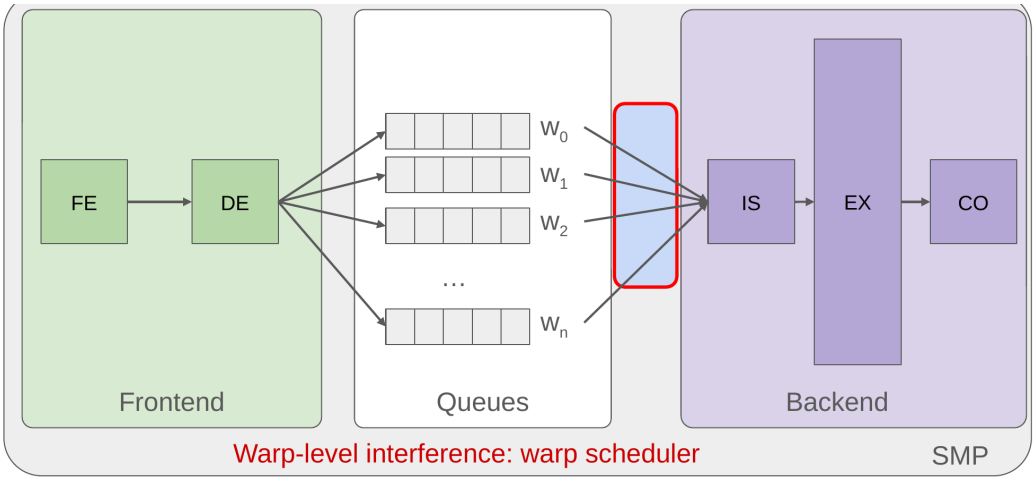
- Calculer le WCET du warp en isolation, considérer le reste comme interférence
- Focus sur le back-end
  - Hypothèse : les files ne sont jamais vides sauf après des événements statiquement prévisibles (ex. défaut de cache d'instructions)



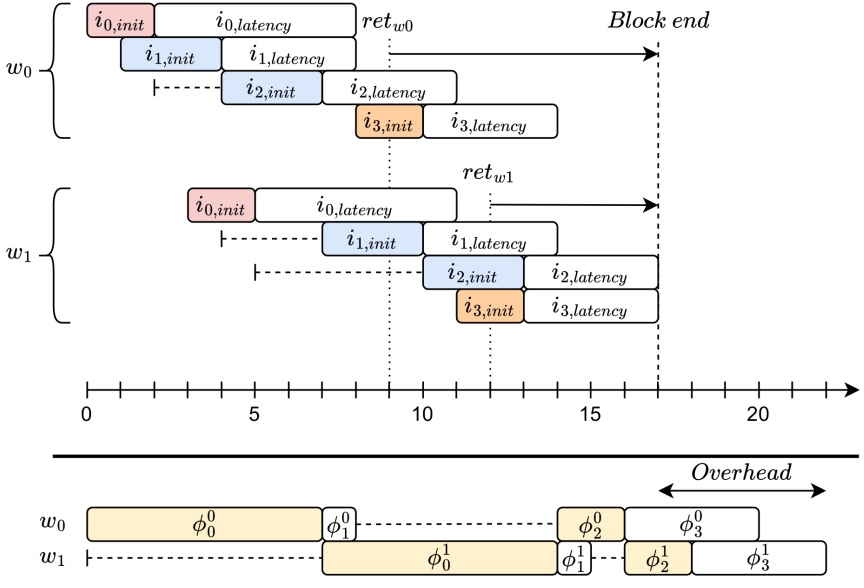
Analyse du WCET : combinaison des warps



Analyse du WCET : combinaison des warps



Analyse du WCET : combinaison des warps



Analyse du WCET : combinaison des warps

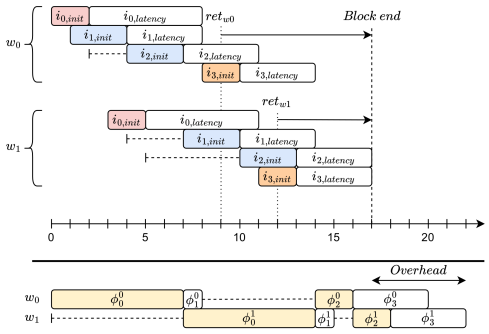
Définition 9. *Borne supérieure du warp (WUB) :*
Pour tout warp w dans un bloc $\mathcal{B}$ donné :

$$WUB(w) = \sum_{i=0}^{\Phi_w - 1} \phi_i^w.dur + \sum_{W \in \mathcal{B} \setminus \{w\}} \Big(\sum_{\substack{j \\ \phi_j^W.\textbf{kind}=\textbf{exec}}} \phi_j^W.dur \Big)$$

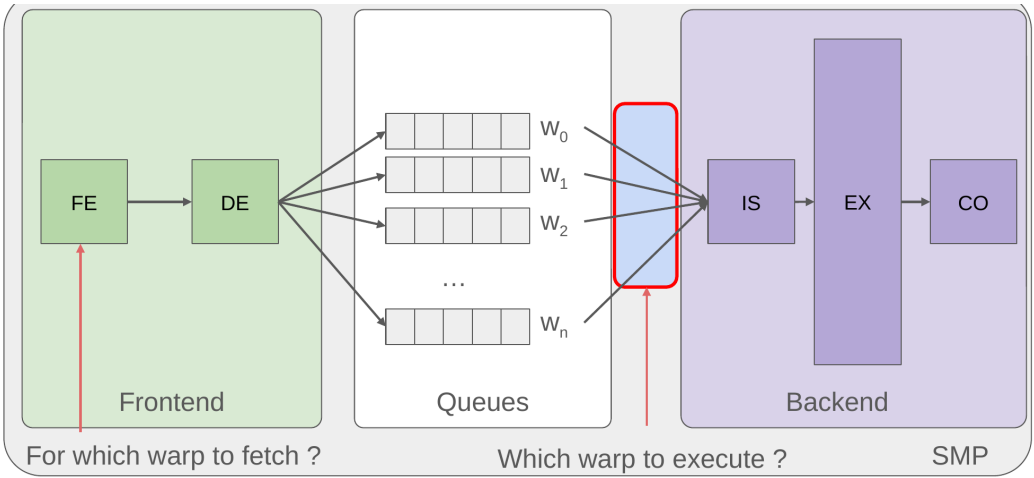
Définition 10. *Borne supérieure générale (GUB) :*
Pour tout bloc de warps $\mathcal{B}$:

$$GUB(\mathcal{B}) = \max_{w \in \mathcal{B}} WUB(w)$$

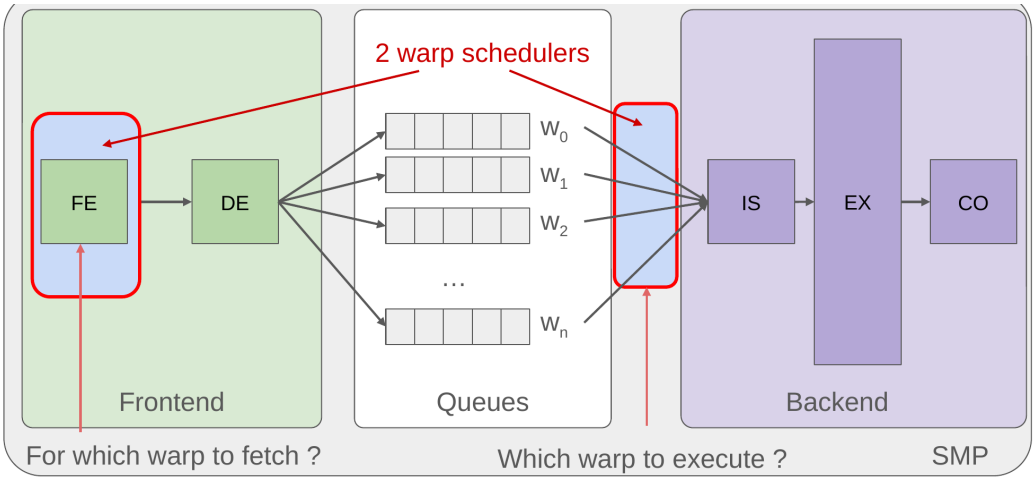
Théorème 11. *Pour tout bloc de warps $\mathcal{B}$ exécutant un noyau $\mathcal{K}$, $WCET(\mathcal{B}) \leq GUB(\mathcal{B})$*



Analyse du WCET : combinaison des warps



Analyse du WCET : combinaison des warps



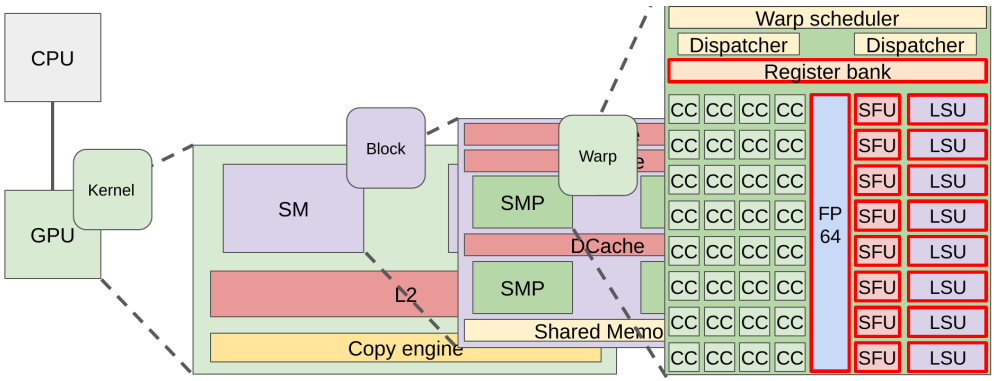
Garantir la validité des hypothèses

- L'analyse du warp en isolation suppose que la file n'est jamais vide sauf dans des cas statiquement prévisibles (ex. défaut de cache d'instructions)
- **Problèmes :**
  - Les deux ordonnanceurs ne sont pas coordonnés
  - Au stade Issue, l'ordonnanceur ne sélectionne que les files ayant des instructions prêtes
  - Au stade Fetch, des interférences entre requêtes cache sont possibles (1 MSHR + 1 succès cache provenant de warps différents)

  $\Rightarrow$ Difficile à prédire statiquement
  - Une politique au Fetch, une autre à l'Issue
  - Les files peuvent être temporairement vides
  - Des cycles peuvent être perdus à cause de conflits au Fetch
- **Solution architecturale :**
  - Garantir qu'une instruction prête est toujours disponible dans chaque file $\Rightarrow$ coordonner les ordonnanceurs et traiter les branchements + insérer un nop à chaque défaut de cache d'instructions
  - Garantir que le Fetch arbitre les conflits entre warps de façon prévisible $\Rightarrow$ un warp ne peut émettre tant que le défaut n'est pas résolu

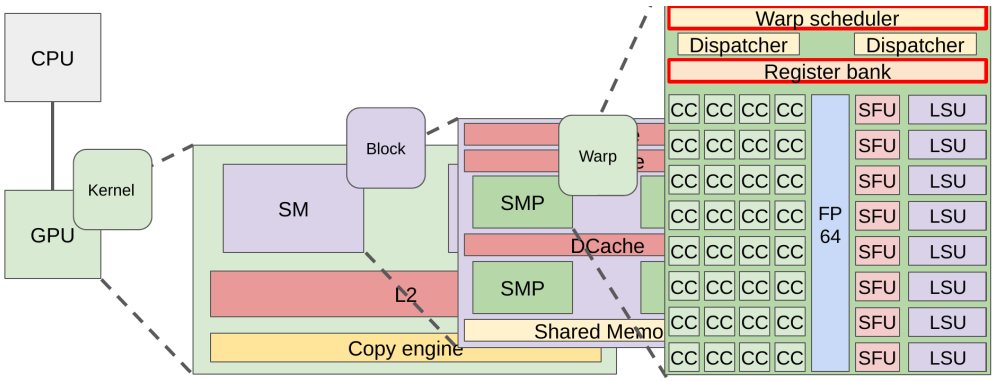
Niveaux de parallélisme dans le calcul du WCET

- Interférence au niveau thread
  - Constitue le WCET du warp en isolation



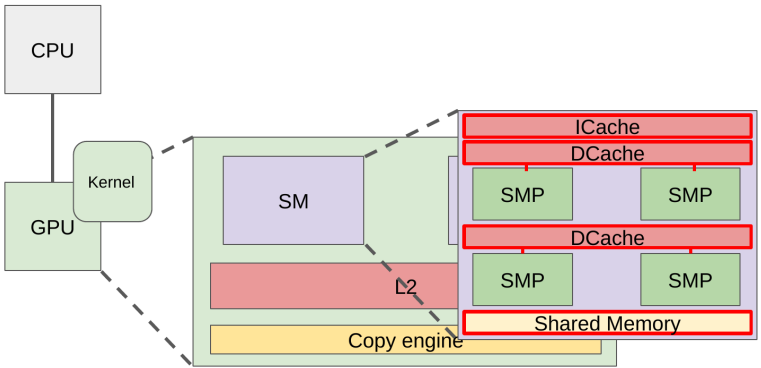
Niveaux de parallélisme dans le calcul du WCET

- Interférence au niveau warp
  - Combiner les warps pour obtenir le WCET du bloc



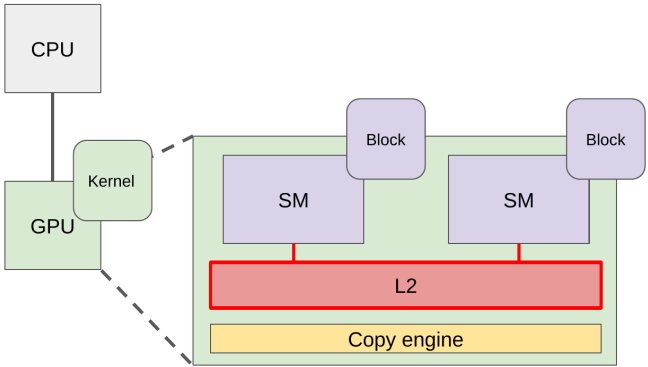
Niveaux de parallélisme dans le calcul du WCET

- Interférence au niveau warp
  - Borner et ajouter des pénalités au WCET du bloc (comme dans les multi-cœurs)



Niveaux de parallélisme dans le calcul du WCET

- Interférence au niveau bloc
  - Borner et ajouter des pénalités au WCET du noyau (accès mémoire + ordonnancement des blocs)



Conclusion

- Analyse statique réalisée comme suit :
  - Analyse temporelle au niveau pipeline pour 1 warp en isolation
  - Analyse de type WCRT pour combiner les warps d'un bloc
  - Les interférences mémoire supplémentaires (niveau warp et niveau bloc) peuvent être modélisées
  - Support matériel nécessaire pour garantir certaines hypothèses
- Réduire la sur-estimation due à l'abstraction
  - Construire un matériel plus prévisible (ex. unités fonctionnelles pipelinées, bancs de registres multi-ports + forwarding d'opérandes)
  - S'appuyer sur la « mémoire partagée » (SPM) avec un style de programmation à double tampon