Resum

En els últims anys, la creixent necessitat de capacitat de còmput ha suposat un repte que ha portat a la indústria a buscar arquitectures alternatives als processadors superescalars amb execució fora d'ordre convencionals, amb l'objectiu d'incrementar la potència de còmput alhora que s'aconsegueix una major eficiència energètica.

Les arquitectures GPU, les quals fins fa només una dècada es dedicaven exclusivament a l'acceleració dels gràfics en els computadors, han sigut una de les alternatives més utilitzades durant alguns anys per a aconseguir l'esmentat objectiu. Una de les característiques particulars de les GPU és el seu elevat ample de banda per a accedir a memòria principal, la qual cosa permet executar un gran nombre de fils de forma molt eficient. Aquesta característica, així com la seua elevada potència computacional executant operacions de coma flotant, ha originat l'aparició del paradigma de computació anomenat *GPGPU computing*, paradigma on les GPU realitzen còmput de propòsit general. Les citades característiques converteixen a les GPU en dispositius especialment apropiats per a l'execució d'aplicacions massivament paral·leles que tradicionalment s'havien executat en processadors convencionals d'altes prestacions.

El treball desenvolupat en aquesta tesi persegueix ajudar a millorar les prestacions de les GPU en l'execució de les aplicacions GPGPU. A aquest efecte, com a primer pas, es realitza un estudi de caracterització on s'identifiquen les característiques més importants d'aquestes aplicacions des del punt de vista de la jerarquia de memòria i el seu impacte en les prestacions. Per a això s'utilitza un simulador detallat cicle a cicle on es modela l'arquitectura d'una GPU recent. L'estudi revela que és necessari modelar de forma més detallada alguns components crítics de la jerarquia de memòria de les GPU per a obtindre resultats precisos. Els resultats obtinguts

mostren que les prestacions aconseguides poden variar fins i tot en un factor de 3× depenent de com es modelen aquests components crítics.

Per aquest motiu, com a segon pas abans d'elaborar la proposta de millora, el treball se centra en determinar quins components de la jerarquia de memòria de la GPU necessiten modelar-se amb major detall per a millorar la precisió dels resultats del simulador i en millorar els models existents d'aquests components. A més, es realitza un estudi de validació que compara els resultats obtinguts amb els models millorats contra els d'una GPU comercial real. Les millores implementades redueixen la desviació dels resultats del simulador sobre els resultats reals al voltant d'un 96%.

Finalment, una vegada millorada la precisió del simulador, en aquesta tesi es presenta una proposta innovadora, denominada FRC (sigles en anglés de *Fetch and Replacement Cache*), que millora en gran manera la potència computacional de la GPU, gràcies a que augmenta el paral·lelisme en l'accés a memòria principal. La proposta incrementa el nombre d'accessos en paral·lel a memòria principal mitjançant l'acceleració de la gestió de les accions de recerca i reemplaçament relacionades amb els accessos que fallen en la cache. La proposta FRC es basa en una xicoteta estructura cache auxiliar que descongestiona el subsistema de memòria eficientment, augmentant les prestacions de la GPU fins a un 118% de mitjana respecte al sistema base. A més, també redueix, al voltant d'un 57%, el consum energètic de la jerarquia de memòria.