Resumen

Las memorias cache han sido implementadas normalmente con tecnología *Static Random-Access Memory* (SRAM) ya que es la tecnología de memoria electrónica más rápida. Sin embargo, esta tecnología consume una gran cantidad de corrientes de fuga, lo cual es un problema de diseño importante porque el consumo de corrientes de fuga incrementa a medida que el tamaño del transistor encoge. Se están considerando tecnologías alternativas para reducir este consumo. Entre ellas, la tecnología *embedded Dynamic* RAM (eDRAM) ofrece por diseño un área y corrientes de fuga mínimas pero las lecturas son destructivas y no es tan rápida como SRAM.

En esta tesis, ambas tecnologías SRAM y eDRAM se mezclan para conseguir las ventajas que cada una de ellas ofrece. En primer lugar, se combinan a nivel de celda para implementar una *macrocelda* de *n*-bits consistente en una celda SRAM y *n-1* celdas eDRAM. La macrocelda se utiliza para construir caches híbridas de datos de primer nivel (L1) asociativas por conjuntos de *n*-vías que tienen una vía SRAM y *n-1* vías eDRAM. Una sola vía SRAM es suficiente para conseguir buenas prestaciones dado que la localidad de los datos en caches L1 es elevada. Mecanismos arquitectónicos como predicción de vía, intercambio de datos (*swaps*) y operaciones de *scrub* se consideran para evitar lecturas eDRAM innecesarias, mantener los datos más recientemente utilizados (MRU) en la vía SRAM *rápida* y eliminar completamente la lógica de refresco. Los resultados experimentales muestran que, comparado con una cache convencional SRAM, las corrientes de fuga y área se reducen considerablemente con un impacto escaso en las prestaciones.

El estudio de los beneficios de las caches híbridas también se ha llevado a cabo en caches de segundo nivel (L2) actuando como caches de último nivel (LLCs). En este caso, las tecnologías se combinan a nivel de banco y se identifica el ratio óptimo de bancos SRAM y eDRAM que consigue el mejor compromiso entre prestaciones, energía y área. Como en las caches L1, los bloques MRU se mantienen en los bancos SRAM y se acceden primero para evitar lecturas destructivas innecesarias. Sin embargo, la lógica de refresco no se elimina ya que la localidad de los datos difiere ampliamente en este nivel de cache. Los

resultados experimentales muestran que una LLC híbrida con un octavo de sus bancos implementados con tecnología SRAM es suficiente para conseguir el mejor compromiso.

Esta disertación también se ocupa de las prestaciones de las políticas de reemplazo en LLCs heterogéneas centrándose principalmente en la sobrecarga de energía incurrida por las operaciones de refresco. En esta tesis se define un concepto nuevo, llamado MRU-Tour (MRUT), que ayuda a la estimación de información de reuso de los bloques de cache. Basándose en este concepto, se propone una familia de algoritmos de reemplazo basados en MRUT que seleccionan aleatoriamente los bloques víctima entre aquellos que tienen un solo MRUT. Estas políticas se mejoran para hacer uso de la recencia de información de unos pocos bloques y adaptarse a los cambios en el comportamiento de las aplicaciones. Los resultados muestran que las políticas MRUT propuestas, con menor complejidad *hardware*, mejoran las prestaciones de *Least Recently Used* (LRU) y de un conjunto representativo del estado del arte de algoritmos de reemplazo para las LLC.

Las operaciones de refresco representan una fracción importante del consumo total de energía dinámica de las LLC eDRAM. Esta fracción incrementa con la capacidad de cache, ya que una cantidad mayor de bloques tienen que ser refrescados en un periodo de tiempo dado. Algunos trabajos anteriores han atacado la energía de refresco teniendo en cuenta las variaciones de los componentes entre celdas. A diferencia de estos trabajos, esta tesis propone una política de refresco selectiva basada en el concepto de MRUT. La política ideada tiene en cuenta el número de MRUTs de un bloque para seleccionar si el bloque se refresca. De esta manera, muchos refrescos realizados en una política de refresco típica y distribuida se omiten, es decir, en aquellos bloques que tienen un solo MRUT. Este mecanismo de refresco se aplica en la memoria LLC híbrida. Los resultados muestran que el consumo de energía de refresco se reduce ampliamente respecto a una cache convencional eDRAM, mientras que la degradación de prestaciones es mínima respecto a una cache convencional SRAM.