Dokumentation – Manual-Cache ELLPACK Format

Manual-Cache ELLPACK erweitert das ursprüngliche ELLPACK Format. Letzteres liest in jedem Schritt die Werte des Vektors direkt aus dem Global Memory. Diese Vorgehensweise setzt auf einen automatischen Cache, wie er bei Fermi Karten (Compute Capability 2.0 und höher) vorhanden ist, um Kosten durch wiederholte und unstrukturierte DRAM-Zugriffe zu vermeiden.

Für ältere (Prä-Fermi-) Karten ohne diesen automatischen Cache wurde die Manual-Cache ELLPACK Variante entwickelt, die die Werte des Vektors manuell im Shared Memory zwischenspeichert. Dazu werden zunächst im Voraus für jeden Block alle Indizes auf Vektoreinträge, die er benutzt, in einer Subvektor genannten Datenstruktur gesammelt. Die Spaltenindizes verweisen dann nicht mehr global auf den Vektor, sondern geben lokal den Index innerhalb des Subvektors an. Da alle Subvektoren nacheinander in einem Feld stehen und unterschiedlich lang sein können, wird eine vierte Datenstruktur – SubvektorStart – benötigt, die Start und Ende eines Subvektors für jeden Block angibt. Die Größe des Shared Memory wird auf die des größten Subvektors angepasst, sodass immer genug Speicher vorhanden ist.

Im Kernel lädt jeder Block dann zuerst die Werte des Vektors an den Indizes seines Subvektors mit Hilfe von SubvektorStart einmal vom Global in sein Shared Memory, um sie in der darauffolgenden Schleife mit allen Threads gemeinsam von dort zu benutzen. Die geringe Latenz und hohe Geschwindigkeit des Shared Memory sorgen dafür, dass das unstrukturierte und sich überschneidende Herumspringen in den Vektoreinträgen kein Problem mehr darstellt.

Zu beachten ist, dass das einmalige Laden der Werte des Vektors zwar auch nicht unbedingt coalesced und aligned sein muss, die Wahrscheinlichkeit dafür allerdings deutlich höher liegt. Denn die Indizes werden im Subvektor in aufsteigender Reihenfolge abgelegt. Benötigt ein Block zwei aufeinanderfolgende Werte aus dem Vektor, so werden sie von aufeinanderfolgenden Threads geladen, auch wenn sie später an ganz unterschiedlichen Stellen benötigt werden.