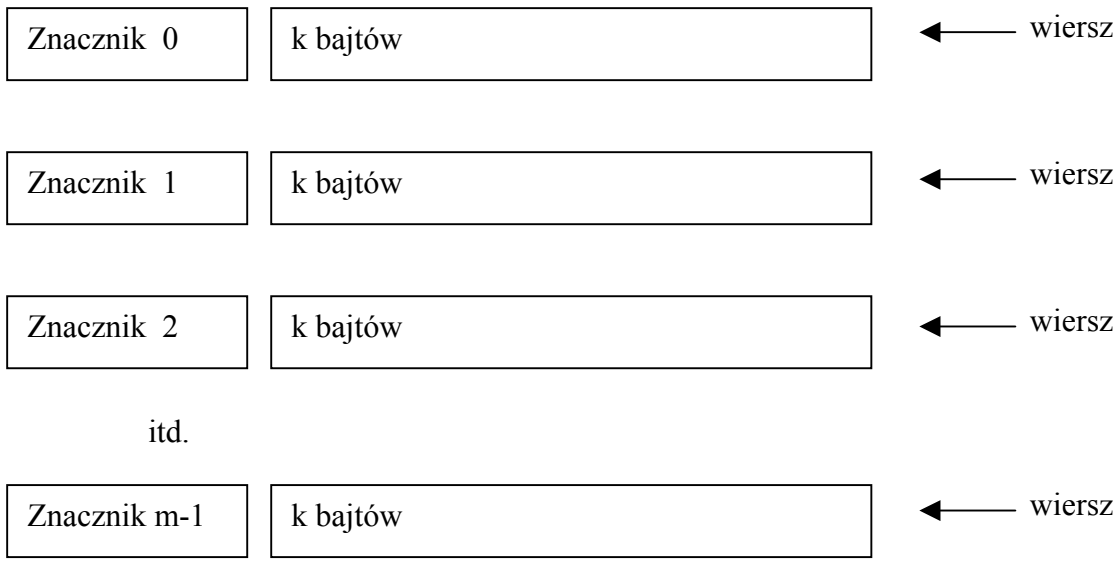
9.3 Współpraca pamięci cache z pamięcią operacyjną

1. Pamięć cache

Pamięć cache to inaczej *pamięć podręczna*. Na ogół pamięć cache wykonywana jest jako pamięć statyczna SRAM i jest znacznie szybsza od pamięci głównej systemu komputerowego. W szerszym sensie pamięć cache to każda szybka pamięć RAM o stosunkowo niewielkiej pojemności (pamięć szybka jest z reguły droższa przy tej samej pojemności).

Będzie nas teraz interesować współpraca pamięci cache z pamięcią główną. Pamięć cache zorganizowana jest na ogół w m wierszy k-bajtowych nazywanych czasem *linijkami* (por. rys. 1). Znacznik jest w pamięci cache adresem wiersza. Pamięć cache ma więc organizację bajtową. Przyjmiemy, że pamięć główna ma również organizację bajtową i podzielona jest na bloki o długości równej długości wiersza. Pomiędzy pamięcią cache a pamięcią główną dane wymieniamy blokami.



Rys. 1. Organizacja pamięci cache; wygodnie jest przyjąć, że *m* i *k* są potęgami 2

Omówimy współpracę pamięci cache z pamięcią główną na przykładzie pamięci cache z tzw. *odwzorowaniem bezpośrednim*. Oznacza to, że jeśli i jest numerem (adresem) wiersza pamięci cache, j jest numerem bloku pamięci głównej, a m ilością wierszy pamięci cache, to

$$i = j \pmod{m}$$

Powyższa funkcja odwzorowania może być łatwo zrealizowana za pomocą podziału adresu pamięci głównej na 3 pola (znacznik, znacznik wiersza, pole pozycji bajtu). Najmniej znaczące bity adresu (czyli bity pola pozycji bajtu) określają jednoznacznie pozycję bajtu w bloku. Pozostałych s bitów określa jeden z 2^s bloków pamięci głównej. Układy logiczne pamięci cache interpretują te s bitów jako znacznik złożony z s-r bitów (bardziej znaczące

bity) i r bitów (mniej znaczących) określających jeden z 2^r wierszy pamięci cache, czyli znacznik wiersza.

Kiedy procesorowi potrzeby jest bajt spod określonego adresu zwraca się on najpierw do pamięci cache. Jeśli znajduje tam poszukiwany bajt, to mówimy, że nastąpiło trafienie. Jeśli poszukiwanego bajtu nie ma w pamięci cache, to procesor przenosi do pamięci cache odpowiedni blok z pamięci głównej, gdzie znajduje się poszukiwany bajt.

Możemy mieć kilka poziomów pamięci cache. Najczęściej stosowane są jeden lub dwa poziomy: L1 i L2 (odpowiednio cache 1-go i 2-go poziomu) .

Często jako pamięć cache stosowana jest pamięć skojarzeniowa. Umożliwia ona operowanie wierszami w wygodniejszy sposób niż w przypadku odwzorowania bezpośredniego.

Stosowanie pamięci cache współpracującej z pamięcią główną to przykład hierarchicznej organizacji (i współpracy) pamięci o różnych szybkościach i pojemnościach. Podstawowy pomysł polega na tym by z pary:

- wolna (ale tania) pamięć o dużej pojemności
- szybka (i droga) pamięć o małej pojemności

zrobić pamięć względnie tanią, szybką i o dużej pojemności.

Realizację tej koncepcji stanowi również tzw. *pamięć wirtualna*. Istota rzeczy polega tu na tym, że programista widzi (tzn. może wykorzystać w pisanym prze siebie programie) znacznie większą pamięć operacyjną niż ta, jaką dysponuje w rzeczywistości. Tak się dzieje dzięki hierarchicznej współpracy zewnętrznej pamięci dyskowej (o bardzo dużej pojemności) z pamięcią operacyjną. Mechanizmem który umożliwia stosunkowo łatwą realizację pamięci wirtualnej jest tzw. *stronicowanie* pamięci czyli podział pamięci operacyjnej na strony.

Inną alternatywną organizacja pamięci jest tzw. *segmentacja* pamięci. czyli podział pamięci operacyjnej na segmenty.

W typowym mikroprocesorze zarządzaniem pamięcią zajmuje się specjalizowany układ, tzw. układ zarządzania pamięcią (MMU od ang. Memory Management Unit).