**8) Cache**

-Prozessor immer noch schneller als Speicher (ca. 10 mal)
-Maschinenbefehl besteht aus Opcode und ggf Operanden
-Cache gehört zur Mikroarchitektur und nicht zur Instruction Set Architecture (ISA)
-Räumliche Lokalität:Zugriffe häufig auf Adresse in Nähe bereits zuvor benutzer Adressen
-Zeitliche Lokalität:Zugriffe auf dieselbe/benachbarte Adressen zeitlich nahe beieinander
-beide Lokalitätsprinzipien treffen auf Befehlzugriffe (meisten, außer bei jmp), Daten (programmabhängig)
-Working Set: Gesamtheit der Speicherobjekte auf die ein Prozess zugreift. (Stack, evt. Shared Libs, Heap, bss, .data, .text). Working Set besteht aus > 5-6 „Regionen“
- Cache hält Kopien von im Speicher liegenden Objekten (Gefahr: Inkosistenz)
- Konsistent: Alle Cache-Kopien und Originial im Hauptspeicher haben gleichen Wert
- Kohärent: Cache und Hauptspeicher für Objekt liefern gleichen Wert
- Vorübergehende INKOSISTENZ ist tolerabel
- Bei L1: separater Cache für Daten und Befehle (jeweils ca. 64 kB, L2: 4 MB)

-Kohärenzprotokoll
  -Lesezugriff
    - Hit: Daten aus Cache liefern
    - Miss: Daten liefern und in Cache kopieren
  -Schreibzugriff
    -Hit
      -Write Through: Daten in Speicher und Cache schreiben
      -Copy-Back: Daten nur in Cache speicher. Cache Zeile ist „dirty“
    -Miss
      -No Write Allocate: Daten nur in Speicher schreiben
      -Write Allocate:Daten mit umliegender Zeile in Speicher & Cache

-Effektive Wartezeit: $T_{eff} = H * T_{hit} + (1 - H) * T_{miss}$
mehrstufig: $T_{eff} = H * T_{hit1} + (1 - H1) * (H2 * T_{hit2} + (1 - H2) * T_{miss})$

-Assoziativspeicher (auch „inhaltsadressierter Speicher“, Wertepaare: Adresse, Daten)
  - enthält Kopien kleiner (max 100 B) Hauptspeicher-Ausschnitte
  - Cache-Eintrag: Tag (Etikett), Daten (Cache-Zeile), Valid Bit, Dirty Bit
  - Bei Speicherzugriff: gleichzeitiger Vergleich der Adresse mit Cache-Einträgen

-Verdrängungsstrategien (wenn alle Cache Zeilen belegt sind → „Platz schaffen“)
  - Random (einfach, überraschend gut)
  - FIFO (die im längsten im Cache gespeicherte Adresse wird ersetzt → schlecht)
  - LRU (least recently used): die im längsten nicht verwendete Zeile ersetzen
  - LFU (least frequently used): die am wenigsten verwendete Zeile ersetzen
-Organisationformen
  - vollassoziativ: fully associative
  -direkt abbildend: direct-mapped
  -mehrfach assoziativ: N-way set associative