1. Cache

- *Hvad er et memory hieraki?*
- *Hvad er cache, og hvad bruger vi den til?*
- *Hvordan er en cache opbygget?*
- *Hvilke begrænsninger har en cache og hvordan minimerer vi disse?*
- *Hvordan kan man som programmør optimere brugen af cache?*

Hvad er et memory hieraki?

- Hvilken memory er tjekket foerst?
  - L1, L2, L3, RAM, Devices(Lagringsmedium)
- L1 er taettest paa CPU'en, men mindst
- L2 er langsommere, men stoerre
- L3(hvis tilstede) er endnu langsommere, men ogsaa tilsvarende stoerre
- RAM er stoerst, men langsomt
- RAM er ogsaa kaldet **Main Memory**
- Eksterne devices(som HDD) er **Secondary Memory**

Hvad er cache, og hvad bruger vi den til?

- Lille, hurtig hukommelse tæt på CPU'en
- Bliver brugt sammen med en FIFO write buffer
  - Frigør processor resource
- Er usynlig for software
- Indeholder de mest brugte data
- Reducerer hastigheds problemer med langsom main memory

Hvordan er en cache opbygget?

- To forskellige arkitekturer: **Von Neumann og Harward**
- Von Neumann:
  - En enkelt cache til både instruktioner og data
  - Unified cache
- Harward:
  - Seperat instruktion og data bus
  - Bedre overordnet performance
  - Split cache
- Cache er delt i to hoveddele:

- **Cache controller**
- **Cache hukommelse**

Hvilke begrænsninger har en cache og hvordan kan vi minimere disse?

- Hit / miss rate
- Linjer
- Størelse

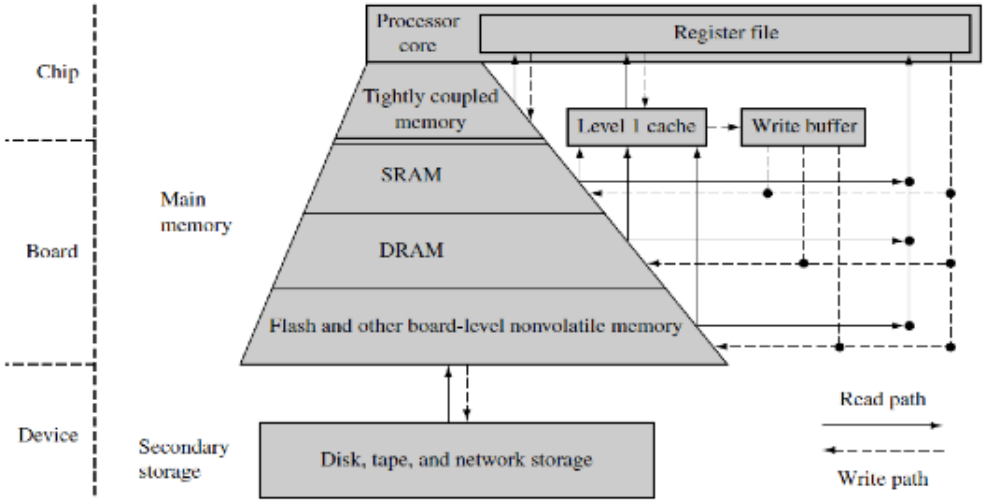
Hvordan kan man som programmør optimere brugen af cache?

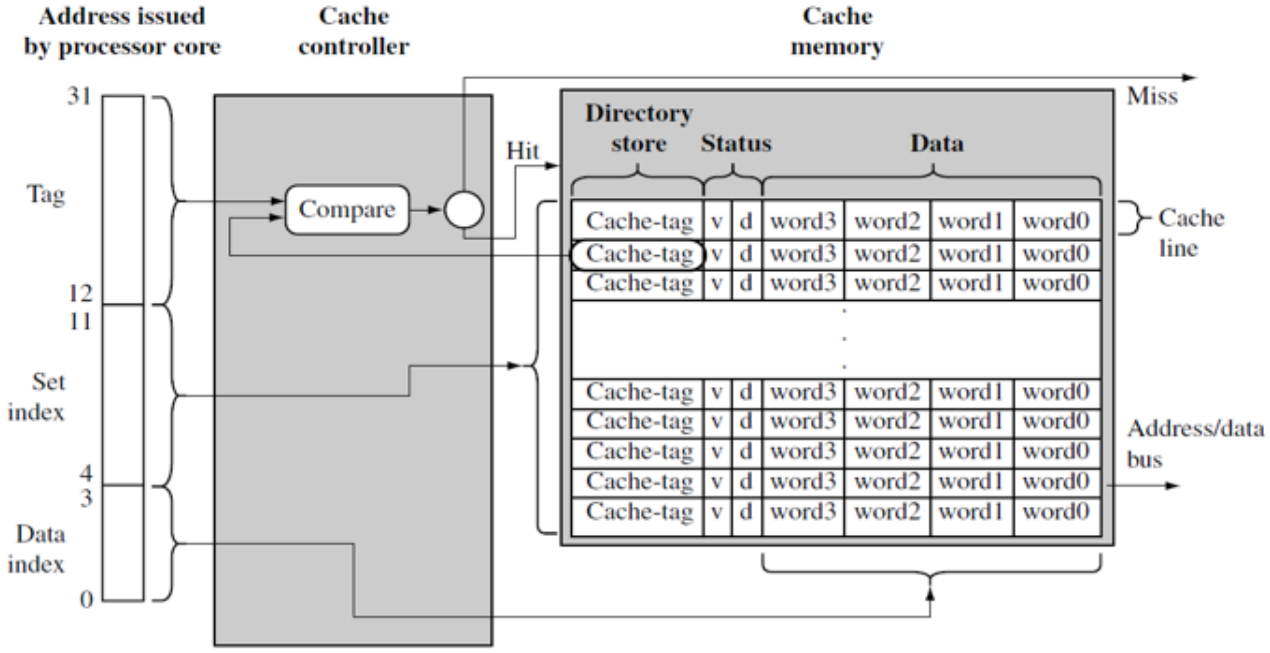
- Compiler optimisering
- Omarangere matricer

Langsom, hurtig

```
for (i = 0; i < N; ++i)
  for (j = 0; j < N; ++j)
    for (k = 0; k < N; ++k)
      res[i][j] += mul1[i][k] * mul2[k][j];
```

```
double tmp[N][N];
for (i = 0; i < N; ++i)
  for (j = 0; j < N; ++j)
    tmp[i][j] = mul2[j][i];
for (i = 0; i < N; ++i)
  for (j = 0; j < N; ++j)
    for (k = 0; k < N; ++k)
      res[i][j] += mul1[i][k] * tmp[j][k];
```




Address issued by processor core
Cache controller
Cache memory
31
Tag
12
11
Set index
4
3
Data index
0
Compare
Hit
Miss
Directory store
Status
Data
Cache-tag
v
d
word3
word2
word1
word0
Cache line
Address/data bus