Caches in Intel Broadwell (2014-...)

- 14nm shrink of 22nm Haswell, i.e., basically same architecture as Haswell
- L1 cache: 64 KB (=32KB data+32KB instruction) per core; data: 8-way associative, write-back
- L2 cache: 256 KB per core, 8-way associative, writeback
- L3 cache: may vary (2 MB to 20 MB), shared among cores

Details can be found in “Intel 64 and IA-32 Architectures Optimization Reference Manual”, Sept. 2014

```
http://www.intel.com/content/dam/www/public/us/en
      /documents/manuals/64-ia-32-architectures-optimization-manual.pdf
```

On Linux systems

```
grep . /sys/devices/system/cpu/cpu0/cache/index*/*
```

will give you information on the caches of your machine.

Cache Parameters of the Haswell Microarchitecture (2013))

Cache Level	1st (data)	1st (instr)	2nd	3rd
Capacity	32KB	32KB	256KB	Can vary
Associativity	8-way	8-way	8-way	Can cary
Line Size	64 bytes	64 bytes	64 bytes	64 bytes
Fastest Latency	4 cycles		11 cycles	
Throughput	0.5 bytes/cycle			
Peak Bandwidth	64 (load) + 32 (store)		64 bytes/cycle	
Update Policy	Writeback		Writeback	Writeback

Associative Memory (German: Assoziativspeicher): Content Addressable Memory (CAM)

In standard random access memory (RAM) the memory is presented an address, and it returns the data stored in this address.

- CAM does not use addresses but uses part of the content (denoted “tag”) to find stored data.
- The tag is searched in a table of stored data, and then the full matching data is returned.
- It is possible that no entry matches.
- If several entries match they are all returned.

Cache Line

The portion of data that is exchanged between RAM and cache (or between levels of Caches). A cache line is valid or invalid (see MESI), there are not partially invalid cache lines.

If a CPU wants to read a single byte the complete cache line of 64 bytes is loaded to the cache. RAM is usually faster in providing cache lines that follow each other in memory (burst).

Ways (German: Weg)

The number of ways is the number of places where a cache line can be stored. In a 8-way set associative cache, a given cache line can be stored in 8 different places in the cache. A 1-way set associative cache is also called directly mapped cache. Here, a given cache line can be stored in only a single place of the cache. Many other cache lines will be stored in the same place.

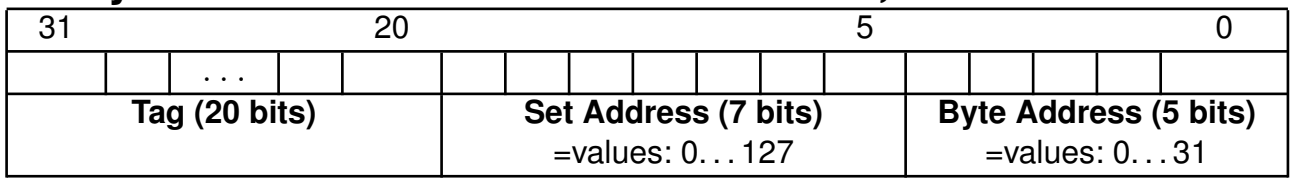
Set (German: Satz)

A set has a number and consists of a tag and the cache line.

Cache Organization (here: 32-bit Pentium from 1993)

The 66 MHz Intel Pentium of 1993 was a 32 bit processor, produced in an 800nm process and famous for the fdiv bug. [Remark: Reincarnated, extended to 64bit, as Intel Xeon Phi with 60 cores. Rank 1 of the current TOP500 uses 3M Xeon Phis.]

2-way set associative L1 cache with 128; $8\,\mathrm{KB} = 2 \cdot 32 \cdot 128$ bytes (ways*line*sets):



Cache-Directory		
Set	Way 0	Way 1
0	**Tag**	**Tag**
1	**Tag**	**Tag**
2	**Tag**	**Tag**
⋮	⋮	⋮
126	**Tag**	**Tag**
127	**Tag**	**Tag**

Cache-Directory-Entry	Cache Line
Tag + Valid bit + . . .	32 bytes of data: byte 0. . . byte 31

The set address is the number of the set in the cache directory.
The byte adress is the number of the entry in the cache line.

Haswell 2014: 8-way associative with 32Kb size; cache line size of 64 bytes.