Virtual Memory System (Cone Family)

- The processor works with virtual addresses.
- All caches work with physical addresses.
- Both address spaces are organized in pages.

Page size: 4 KB (can be changed to 2 KB and even 1 KB on the latest processor; change in OS settings)

Address translation: virtual page number (VPN) → physical page number (PPN)

L1 cache (32 KB)

64 sets

64 bits

4 = associativity

Page size

L1 cache lookup can start concurrently with address translation.
Address translation

- uses a cache called translation lookaside buffer (TLB)
- In Sandybridge and later:

ITLB (instructions): 128 entries
DTLB (data): 64 entries

STLB (shared level 2): 512 - 1536 entries
(Sandy Bridge - Skylake)

STLB hit: no penalty
DTLB miss, STLB hit: few cycles penalty
STLB miss: can be very expensive

Impact on performance

Repeatedly accessing a working set spread out over too many pages (uncached) yields TLB misses and can yield significant slow down.

Example: STLB = 1024 (Haswell)

accessing repeatedly 2048 doubles spread out in
memory = working set of 2048 x 64 = 128 KB

↑
cache block

⇒ fits into L2 cache

set TLB misses
We look for parts in working set spread in memory
- block row of a: contiguous
- all of s: contiguous
- block of c: if M > 512 (512 doubles = 4096)
  then spread over > Nb pages

Typically Nb is in the 10s, so no big problem

But: the BLAS function dgemm has this interface:

\[
dgemm(a, s, c, N, K, M, lda, ldc, ide)
\]

Leading dimensions: enable use on matrices inside matrices.
\[
\text{assume } l_2a, l_2b, l_2e > 512 \\
- \text{ block row of } a: \text{ spread our } \geq N_3 \text{ pages} \\
- \text{ all of } b: \text{ spread our } \geq N \text{ pages} \\
- \text{ block of } c: \text{ spread over } \geq N_3 \text{ pages} \\
\]

So copying to contiguous memory may pay off!

\[
\begin{align*}
&// \text{ all of } b \text{ reused: possibly } \text{ copy} \\
&\text{ for } i = 0: N_3: N-1 \\
&// \text{ block row of } a \text{ reused: possibly } \text{ copy} \\
&\text{ for } j = 0: N_3: N-1 \\
&// \text{ block of } c \text{ reused: possibly } \text{ copy} \\
&\text{ for } k = 0: N_3: K-1 \\
&\ldots 
\end{align*}
\]