intel.

3rd Generation Intel® Xeon® Scalable Processor Family
(processors based on Ice Lake microarchitecture)

Instruction Throughput and Latency README

June 2021

Revision 1.0

Revision History		
Document ID	**Description**	**Date**
347348-001US	Initial Release	Jun-21

Legal Notices and Disclaimers

README

Review all information in this README file prior to using the "Throughput and Latency" csv file.

File format:
Iform - "Iform" is the XED term for variants of instructions. Please consider them experimental and subject to change.
Xed url: https://intelxed.github.io/
Example:
VADDPD_XMMdq_XMMdq_MEMdq
Instruction name: VADDPD
Instruction sources: src0 XMM, src1 XMM, src2 MEM
Data element size: dq – 128-bit
Regsize - source register size
Mask - this AVX512 instruction uses a mask (k register)
Throughput definition –
The number of clock cycles required to wait before the issue ports are free to accept the same instruction again.
This number can be lower than 1, e.g., 0.5 or 0.33, when instruction execution time is 1 cycle and the machine can execute multiple instructions within a single cycle. An instruction that executes without occupying a functional unit may have zero throughput value.
Latency definition -
The number of clock cycles that are required for the CPU to complete the execution of all of the μops that form an instruction.

The file contains partial throughput / latency data:
a) Not all 3rd Generation Intel® Xeon® Scalable Processor Family instructions are appearing in the database.
b) Not all instructions that do appear have both their throughput and their latency specified.

If you require data that is missing, we suggest using the values available in the Intel® Xeon® Scalable Processor throughput and latency document (available here: https://software.intel.com/en-us/articles/intel-sdm#optimization), that provides Skylake Server microarchitecture information as a reasonable approximation.

Note that actual throughput measured on Intel processors may vary by up to 0.1 cycles. Also, latency may sometimes vary due to dynamic micro-architectural conditions, and in this case we report the average rounded to the nearest integer value. In all cases we assume data read from DCU or written to store-buffers.