

Recommendations:**Increase execution time:**

Application execution time is too short. Metrics data may be unreliable. Consider reducing the sampling interval or increasing your application execution time.

Hotspots: Start with Hotspots analysis to understand the efficiency of your algorithm.

Use Hotspots analysis to identify the most time consuming functions. Drill down to see the time spent on every line of code.

Microarchitecture Exploration: There is low microarchitecture usage (8.1%) of available hardware resources. of Pipeline Slots

Run Microarchitecture Exploration analysis to analyze CPU microarchitecture bottlenecks that can affect application performance.

Memory Access: The Memory Bound metric is high (47.2%). A significant fraction of execution pipeline slots could be stalled due to demand memory load and stores. of Pipeline Slots

Use Memory Access analysis to measure metrics that can identify memory access issues.

Elapsed Time: 0.030s

Application execution time is too short. Metrics data may be unreliable. Consider reducing the sampling interval or increasing your application execution time.

CPU:

IPC: 0.015

The IPC may be too low. This could be caused by issues such as memory stalls, instruction starvation, branch misprediction or long latency instructions. Explore the other hardware-related metrics to identify what is causing low IPC.

SP GFLOPS: 0.000

DP GFLOPS: 0.000

x87 GFLOPS: 0.004

Average CPU Frequency: 2.176 GHz

GPU:

Time: 47.4% (0.014s) of Elapsed time

GPU utilization is low. Consider offloading more work to the GPU to increase overall application performance.

IPC Rate: 1.234

Effective Logical Core Utilization: 150.7% (12.058 out of 8)

Effective Physical Core Utilization: 100.0% (4.000 out of 4)

Microarchitecture Usage: 8.1% of Pipeline Slots

You code efficiency on this platform is too low.

Possible cause: memory stalls, instruction starvation, branch misprediction or long latency instructions.

Next steps: Run Microarchitecture Exploration analysis to identify the cause of the low microarchitecture usage efficiency.

Retiring: 8.1% of Pipeline Slots

Front-End Bound: 2.7% of Pipeline Slots

Back-End Bound: 88.0% of Pipeline Slots

A significant portion of pipeline slots are remaining empty. When operations take too long in the back-end, they introduce bubbles in the pipeline that ultimately cause fewer pipeline slots containing useful work to be retired per cycle than the machine is capable to support. This opportunity cost results in slower execution. Long-latency operations like divides and memory operations can cause this, as can too many operations being directed to a single execution port (for example, more multiply operations arriving in the back-end per cycle than the execution unit can support).

Memory Bound: 47.2% of Pipeline Slots

The metric value is high. This can indicate that the significant fraction of execution pipeline slots could

be stalled due to demand memory load and stores. Use Memory Access analysis to have the metric breakdown by memory hierarchy, memory bandwidth information, correlation by memory objects.

Core Bound: 40.8% of Pipeline Slots

This metric represents how much Core non-memory issues were of a bottleneck. Shortage in hardware compute resources, or dependencies software's instructions are both categorized under Core Bound. Hence it may indicate the machine ran out of an 000 resources, certain execution units are overloaded or dependencies in program's data- or instruction- flow are limiting the performance (e.g. FP-chained long-latency arithmetic operations).

Bad Speculation: 1.3% of Pipeline Slots

Memory Bound: 47.2% of Pipeline Slots

The metric value is high. This can indicate that the significant fraction of execution pipeline slots could be stalled due to demand memory load and stores. Use Memory Access analysis to have the metric breakdown by memory hierarchy, memory bandwidth information, correlation by memory objects.

L1 Bound: 26.7% of Clockticks

This metric shows how often machine was stalled without missing the L1 data cache. The L1 cache typically has the shortest latency. However, in certain cases like loads blocked on older stores, a load might suffer a high latency even though it is being satisfied by the L1.

L2 Bound: 0.4% of Clockticks

L3 Bound: 1.6% of Clockticks

DRAM Bound: 4.8% of Clockticks

DRAM Bandwidth Bound: 99.3% of Elapsed Time

The system spent much time heavily utilizing DRAM bandwidth. Improve data accesses to reduce cacheline transfers from/to memory using these possible techniques: 1) consume all bytes of each cacheline before it is evicted (for example, reorder structure

elements and split non-hot ones); 2) merge compute-limited and bandwidth-limited loops; 3) use NUMA optimizations on a multi-socket system. Note: software prefetches do not help a bandwidth-limited application. Run Memory Access analysis to identify data structures to be allocated in High Bandwidth Memory (HBM), if available.

Store Bound: 0.4% of Clockticks

Vectorization: 1.4% of Packed FP Operations

A significant fraction of floating point arithmetic instructions are scalar. Use Intel Advisor to see possible reasons why the code was not vectorized.

Instruction Mix:

SP FLOPs: 0.0% of uOps
Packed: 100.0% from SP FP
128-bit: 100.0% from SP FP

A significant fraction of floating point arithmetic vector instructions is executed with a partial vector load. Make sure you compile the code with the latest instruction set or use Intel Advisor for vectorization help.

256-bit: 0.0% from SP FP
Scalar: 0.0% from SP FP

DP FLOPs: 0.0% of uOps
Packed: 69.2% from DP FP
128-bit: 69.2% from DP FP

A significant fraction of floating point arithmetic vector instructions is executed with a partial vector load. Make sure you compile the code with the latest instruction set or use Intel Advisor for vectorization help.

256-bit: 0.0% from DP FP
Scalar: 30.8% from DP FP

A significant fraction of floating point arithmetic instructions are scalar. Use Intel Advisor to see possible reasons why the code was not vectorized.

x87 FLOPs: 0.1% of uOps

Non-FP: 99.9% of uOps

FP Arith/Mem Rd Instr. Ratio: 0.038

The metric value is low. This can be a result of unaligned access to data for vector operations. Use Intel Advisor to find possible data access inefficiencies for vector operations.

FP Arith/Mem Wr Instr. Ratio: 0.060

The metric value is low. This can be a result of unaligned access to data for vector operations. Use Intel Advisor to find possible data access inefficiencies for vector operations.

GPU Active Time: 47.4%

GPU utilization is low. Consider offloading more work to the GPU to increase overall application performance.

GPU Utilization when Busy: 15.2%

The percentage of time when the EUs were stalled or idle is high, which has a negative impact on compute-bound applications.

IPC Rate: 1.234
EU State: 15.2%
Active: 15.2%
Stalled: 38.7%

A significant portion of GPU time is lost due to stalls. For compute-bound code, this could indicate that performance is limited by memory or sampler accesses.

Idle: 46.1%

A significant portion of GPU time is spent idle. This is usually caused by imbalance or thread scheduling problems.

Occupancy: 31.1% of peak value

Low value of the occupancy metric may be caused by inefficient work scheduling. Make sure work items are neither too small nor too large.

Collection and Platform Info:

Application Command Line: ./codecs/HHI-VVC/decoder/vvdecapp "-b" ".bin/HHI-VVC/randomaccess_fast.cfg/CLASS_C/RaceHorses_416x240_30_QP_22_HHI-VVC.bin"

Operating System: 5.4.0-72-generic DISTRIB_ID=Ubuntu
DISTRIB_RELEASE=18.04 DISTRIB_CODENAME=bionic
DISTRIB_DESCRIPTION="Ubuntu 18.04.5 LTS"

Computer Name: eimon

Result Size: 3.7 MB

Collection start time: 22:28:22 18/04/2021 UTC

Collection stop time: 22:28:22 18/04/2021 UTC

Collector Type: Event-based sampling driver,Event-based counting driver

CPU:

Name: Intel(R) Processor code named Kabylake ULX

Frequency: 1.992 GHz

Logical CPU Count: 8

Max DRAM Single-Package Bandwidth: 10.000 GB/s

Cache Allocation Technology:

Level 2 capability: not detected

Level 3 capability: not detected

GPU:

Name:	Display controller: Intel Corporation Device 22807
Vendor:	Intel Corporation
EU Count:	24
Max EU Thread Count:	7
Max Core Frequency:	1.150 GHz

Intel® oneAPI VTune™ Profiler 2021.1.1 Gold**Recommendations:****Increase execution time:**

Application execution time is too short. Metrics data may be unreliable. Consider reducing the sampling interval or increasing your application execution time.

Hotspots: Start with Hotspots analysis to understand the efficiency of your algorithm.

Use Hotspots analysis to identify the most time consuming functions. Drill down to see the time spent on every line of code.

Microarchitecture Exploration: There is low microarchitecture usage (3.5%) of available hardware resources. of Pipeline Slots

Run Microarchitecture Exploration analysis to analyze CPU microarchitecture bottlenecks that can affect application performance.

Memory Access: The Memory Bound metric is high (42.3%). A significant fraction of execution pipeline slots could be stalled due to demand memory load and stores. of Pipeline Slots

Use Memory Access analysis to measure metrics that can identify memory access issues.

Elapsed Time: 0.031s

Application execution time is too short. Metrics data may be unreliable. Consider reducing the sampling interval or increasing your application execution time.

CPU:**IPC:** 0.060

The IPC may be too low. This could be caused by issues such as memory stalls, instruction starvation, branch misprediction or long latency instructions. Explore the other hardware-related metrics to identify what is causing low IPC.

DP GFLOPS: 0.000

x87 GFLOPS: 0.002

Average CPU Frequency: 1.849 GHz

GPU:

Time: 32.5% (0.010s) of Elapsed time

GPU utilization is low. Consider offloading more work to the GPU to increase overall application performance.

IPC Rate: 1.325

Effective Logical Core Utilization: 124.6% (9.968 out of 8)

Effective Physical Core Utilization: 100.0% (4.000 out of 4)

Microarchitecture Usage: 3.5% of Pipeline Slots

You code efficiency on this platform is too low.

Possible cause: memory stalls, instruction starvation, branch misprediction or long latency instructions.

Next steps: Run Microarchitecture Exploration analysis to identify the cause of the low microarchitecture usage efficiency.

Retiring: 3.5% of Pipeline Slots

Front-End Bound: 9.4% of Pipeline Slots

Back-End Bound: 85.7% of Pipeline Slots

A significant portion of pipeline slots are remaining empty. When operations take too long in the back-end, they introduce bubbles in the pipeline that ultimately cause fewer pipeline slots containing useful work to be retired per cycle than the machine is capable to support. This opportunity cost results in slower execution. Long-latency

operations like divides and memory operations can cause this, as can too many operations being directed to a single execution port (for example, more multiply operations arriving in the back-end per cycle than the execution unit can support).

Memory Bound:

42.3% of Pipeline Slots

The metric value is high. This can indicate that the significant fraction of execution pipeline slots could be stalled due to demand memory load and stores. Use Memory Access analysis to have the metric breakdown by memory hierarchy, memory bandwidth information, correlation by memory objects.

Core Bound:

43.4% of Pipeline Slots

This metric represents how much Core non-memory issues were of a bottleneck. Shortage in hardware compute resources, or dependencies software's instructions are both categorized under Core Bound. Hence it may indicate the machine ran out of an OOO resources, certain execution units are overloaded or dependencies in program's data- or instruction- flow are limiting the performance (e.g. FP-chained long-latency arithmetic operations).

Bad Speculation:

1.4% of Pipeline Slots

Memory Bound:

42.3% of Pipeline Slots

The metric value is high. This can indicate that the significant fraction of execution pipeline slots could be stalled due to demand memory load and stores. Use Memory Access analysis to have the metric breakdown by memory hierarchy, memory bandwidth information, correlation by memory objects.

L1 Bound:

21.9% of Clockticks

This metric shows how often machine was stalled without missing the L1 data cache. The L1 cache typically has the shortest latency. However, in certain cases like loads blocked on older stores, a load might suffer a high latency even though it is being satisfied by the L1.

L2 Bound: 0.5% of Clockticks
L3 Bound: 6.5% of Clockticks

This metric shows how often CPU was stalled on L3 cache, or contended with a sibling Core. Avoiding cache misses (L2 misses/L3 hits) improves the latency and increases performance.

DRAM Bound: 2.5% of Clockticks
DRAM Bandwidth Bound: 98.0% of Elapsed Time

The system spent much time heavily utilizing DRAM bandwidth. Improve data accesses to reduce cacheline transfers from/to memory using these possible techniques: 1) consume all bytes of each cacheline before it is evicted (for example, reorder structure elements and split non-hot ones); 2) merge compute-limited and bandwidth-limited loops; 3) use NUMA optimizations on a multi-socket system. Note: software prefetches do not help a bandwidth-limited application. Run Memory Access analysis to identify data structures to be allocated in High Bandwidth Memory (HBM), if available.

Store Bound: 0.9% of Clockticks

Vectorization: 0.0% of Packed FP Operations

A significant fraction of floating point arithmetic instructions are scalar. Use Intel Advisor to see possible reasons why the code was not vectorized.

Instruction Mix:

SP FLOPs:	0.0% of uOps
Packed:	0.0% from SP FP
128-bit:	0.0% from SP FP
256-bit:	0.0% from SP FP
Scalar:	0.0% from SP FP
DP FLOPs:	0.0% of uOps
Packed:	0.0% from DP FP
128-bit:	0.0% from DP FP
256-bit:	0.0% from DP FP
Scalar:	100.0% from DP FP

This code has floating point operations and is not vectorized. Consider using Intel Advisor to vectorize the loops.

x87 FLOPs: 0.1% of uOps

Non-FP: 99.9% of uOps

FP Arith/Mem Rd Instr. Ratio: 0.004

The metric value is low. This can be a result of unaligned access to data for vector operations. Use Intel Advisor to find possible data access inefficiencies for vector operations.

FP Arith/Mem Wr Instr. Ratio: 0.007

The metric value is low. This can be a result of unaligned access to data for vector operations. Use Intel Advisor to find possible data access inefficiencies for vector operations.

GPU Active Time: 32.5%

GPU utilization is low. Consider offloading more work to the GPU to increase overall application performance.

GPU Utilization when Busy: 19.6%

The percentage of time when the EUs were stalled or idle is high, which has a negative impact on compute-bound applications.

IPC Rate: 1.325
EU State: 19.6%
Active: 19.6%
Stalled: 31.9%

A significant portion of GPU time is lost due to stalls. For compute-bound code, this could indicate that performance is limited by memory or sampler accesses.

Idle: 48.6%

A significant portion of GPU time is spent idle. This is usually caused by imbalance or thread scheduling problems.

Occupancy: 32.9% of peak value

Low value of the occupancy metric may be caused by inefficient work scheduling. Make sure work items are neither too small nor too large.

Collection and Platform Info:

Application Command Line: ./codecs/HHI-VVC/decoder/vvdecapp "-b" ".bin/HHI-VVC/randomaccess_fast.cfg/CLASS_C/RaceHorses_416x240_30_QP_22_HHI-VVC.bin"

Operating System: 5.4.0-72-generic DISTRIB_ID=Ubuntu
DISTRIB_RELEASE=18.04 DISTRIB_CODENAME=bionic
DISTRIB_DESCRIPTION="Ubuntu 18.04.5 LTS"

Computer Name: eimon

Result Size: 3.7 MB

Collection start time: 07:48:14 19/04/2021 UTC

Collection stop time: 07:48:14 19/04/2021 UTC

Collector Type: Event-based sampling driver,Event-based counting driver

CPU:

Name: Intel(R) Processor code named Kabylake
ULX

Frequency: 1.992 GHz

Logical CPU Count: 8

Max DRAM Single-Package Bandwidth: 10.000 GB/s

Cache Allocation Technology:

Level 2 capability: not detected

Level 3 capability: not detected

GPU:	
Name:	Display controller: Intel Corporation
Device 22807	
Vendor:	Intel Corporation
EU Count:	24
Max EU Thread Count:	7
Max Core Frequency:	1.150 GHz