Optimizing HEVC decoding efficiency on high-end NUMA systems

Tuning MainConcept HEVC Decoder on Intel® Xeon® Scalable Processors

Authors

Björn Taubert
Software Engineering Manager, Intel

Alexander Lazarev
Senior Application Engineer, Intel

Sergey A. Bufalov
Manager, Software Engineering, MainConcept®

Frank Schoenberger
Senior Product Manager, MainConcept®

Intel® Xeon® Scalable Processors have advanced scalability features. To gain workload performance increases developers need to put a strong focus on scalability within their software architecture, revise and adopt multi-processing strategies. MainConcept, together with the Intel Architecture Graphics and Software (IAGS) division at Intel, could achieve up to 1.7x1 performance and efficiency increase for the HEVC decoder (4320p@43Mbit) from the MainConcept® HEVC/H.265 SDK optimizing version 8.2 to 10.0.1 on Intel® Xeon® Platinum 8180 Processor3.

MainConcept Intel Architecture Graphics and Software (IAGS) division

1 Introduction

Latest Intel® Xeon® Scalable Processors have advanced scalability including optimized CPU core design, memory bandwidth, and communication inter-connections. To gain workload performance increases, maxing out the hardware capabilities, software developers need to put a strong focus on scalability within their software architecture, revise and adopt multi-processing strategies. The MainConcept Codec experts have a profound experience in designing high-quality professional products that scale and perform to available computational resources. This paper explains some concepts for effective utilization of Intel products using the MainConcept HEVC/H.265 Decoder.

2 Testing Environment

2.1 Reference System

The reference performance and efficiency are based on an Intel® Core™ i7-8700K processor with 16GB RAM measured for 10-bit 4:2:2 720p to 4320p elementary streams in frames decoded per second (fps).

Table 1 describes the performance of MainConcept's HEVC/H.265 SDK 8.2 decoder on the reference desktop system.

Table of Contents

Inroduction ............................... 1
Testing Environment .................. 1
Problem Statement...................... 3
Optimizations............................ 4
Summary and Discussion .......... 5

Note: The testing environment and configurations used in this document may have been optimized for performance on Intel microprocessors. Performance results are based on testing as of 11/30/2018 and may not reflect all publicly available security updates. See configuration disclosure for details.

1Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

2Testing by MainConcept as of November 30, 2018
Configuration: 4 x Intel® Xeon® Platinum 8180 @ 2.5 GHz, total threads 224, RAM 768 GB, Oracle Linux 7.6, kernel 3.10.0-862.14.4, ucode: 0x2000043


Table 1. MainConcept HEVC SDK 8.2 decoder performance on Intel® Core™ i7-8700K.

Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

For more complete information, visit www.intel.com/benchmarks.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Testing by MainConcept as of November 30, 2018

Configuration: Intel® Core™ i7-8700K @ 3.7 GHz, total threads 12, RAM 32 GB, Centos 7.6, kernel 3.10.0-957.5.1, ucode: 0x96

To achieve comparability for decoder efficiency between Intel® Core™ i7-8700K processor to Intel® Xeon® server processors in a linear manner, the effects of increased number of CPUs and NUMA (Non-Uniform Memory Access) were minimized. To evaluate the decoder scalability on platform \( P \) with total hardware threads \( \text{THREADS}_P \) and base frequency \( \text{FREQUENCY}_P \), its efficiency is compared to the reference value \( \text{EFFICIENCY}_{i7-8700k} \), computed by the following formula:

\[
\text{EFFICIENCY}_P = \frac{\text{EFFICIENCY}_{i7-8700k} \times \text{THREADS}_P}{\text{THREADS}_{i7-8700k}} \times \frac{\text{FREQUENCY}_P}{\text{FREQUENCY}_{i7-8700k}}
\]

Below is the list of Intel® Xeon® processors used in the article and their reference efficiencies, calculated by the formula.

![Figure 1. MainConcept HEVC Decoder reference efficiency on Intel® processors.](image)
The next table describes configurations of test systems for Intel® processors referenced in the article.

<table>
<thead>
<tr>
<th>PROCESSOR</th>
<th>NODES</th>
<th>THREADS</th>
<th>FREQUENCY</th>
<th>RAM</th>
<th>LINUX</th>
<th>KERNEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel® Xeon® Platinum 8180 x 4</td>
<td>4</td>
<td>224</td>
<td>2.5 GHz</td>
<td>768 GB</td>
<td>Oracle 7.6</td>
<td>3.10.0-862.14.4</td>
</tr>
<tr>
<td>Intel® Xeon® Platinum 8168 x 2</td>
<td>2</td>
<td>96</td>
<td>2.7 GHz</td>
<td>192 GB</td>
<td>Oracle 7.6</td>
<td>3.10.0-862.14.4</td>
</tr>
<tr>
<td>Intel® Xeon® E5-2699 v4 x 2</td>
<td>2</td>
<td>88</td>
<td>2.2 GHz</td>
<td>128 GB</td>
<td>Oracle 7.6</td>
<td>3.10.0-862.14.4</td>
</tr>
<tr>
<td>Intel® Xeon® E5-2640 v3 x 2</td>
<td>2</td>
<td>32</td>
<td>2.6 GHz</td>
<td>64 GB</td>
<td>Gentoo</td>
<td>4.10.17</td>
</tr>
<tr>
<td>Intel® Core™ i7-8700K</td>
<td>1</td>
<td>12</td>
<td>3.7 GHz</td>
<td>32 GB</td>
<td>Centos 7.6</td>
<td>3.10.0-957.5.1</td>
</tr>
</tbody>
</table>

Table 2. Test systems configuration.

3 Problem Statement

The MainConcept HEVC Decoder version 8.2 is a well-designed, efficient and performant companion for Intel® Core™ processors. The decoder’s efficiency on UMA (Uniform Memory Access) CPUs is based on SIMD vectorization, memory footprint reduction, cache optimization, lightweight synchronization primitives, and fibers. However, these optimization technics don’t deliver the best performance on Intel® Xeon® server processors.

The gap between real and reference efficiencies in HEVC/H.265 SDK 8.2 is illustrated in the table below.

Figure 2. MainConcept HEVC Decoder reference and SDK 8.2 efficiency on Intel® processors.

The growing number of hardware threads made it possible to decode more pictures in parallel, increasing the memory footprint of the workload. As a result, it caused better synchronization, more scheduling operations, further contention between threads, increased switching between tasks (fibers), improved cache invalidation, and extensive data traffic between NUMA nodes.
On the Intel® Xeon® E5-2640 v3 processor the HEVC Decoder 8.2 achieves the efficiency of 4.3, which is less than the Reference value of 5.1 by 15.6% (refer to the 4th column in the table of Figure 2 above). On the Intel® Xeon® E5-2699 v4 the difference is 23.5% and it gets even worse growing up to 29.4% and 73% on Intel® Xeon® Scalable 8168 and 8180 processors, respectively.

4 Optimizations

According to Intel’s brief introduction of Scalable Platform the expected speedup of technical computations comparing Xeon® E5-2699 v4 versus new Intel® Xeon® Scalable Platinum 8180 is approximate 2.2 times. From figure 2, it is evident that the claimed boost is not achieved in HEVC/H.265 SDK 8.2, and that the Intel® Xeon® Scalable platform is underutilized.

The reason for the observed scalability deficiency was identified inside the MainConcept multi-processing runtime.

Bottleneck identifications and improvement validations were accomplished with Intel® VTune™ Amplifier Task API invoked from inside the MainConcept multi-processing runtime. The growing number of hardware threads made it possible to decode more pictures in parallel increasing the memory footprint of the workload. As a result, it caused better synchronization, more scheduling operations, further contention between threads, increased switching between tasks (fibers), improved cache invalidation, and extensive data traffic between NUMA nodes.

In HEVC/H.265 SDK version 10.0.1, the MainConcept multi-processing runtime was redesigned to eliminate identified deficiencies.

The old features of the multi-processing runtime that were revised and the new features that were added are listed below with the short description of the respective improvement they deliver.

- Less cache invalidation
- Less scheduling contention
- Less scheduling delays
- Less task contention
- Less memory access overhead
- Less thread contention
- Better hardware compatibility
- Better workload compatibility
- Unlimited flexibility in tuning

Table 3. Base features of MainConcept multi-processing runtime.

The results of improved task scheduling in HEVC/H.265 SDK 10.0.1 are summarized below.

---

1 Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Performance results are based on testing as of 11/30/2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Testing by MainConcept as of November 30, 2018

Configurations:
- 4 x Intel® Xeon® Platinum 8180 @ 2.5 GHz, total threads 224, RAM 768 GB, Oracle Linux 7.6, kernel 3.10.0-862.14.4, ucode: 0x2000043
- 2 x Intel® Xeon® Platinum 8168 @ 2.7 GHz, total threads 96, RAM 192 GB, Oracle Linux 7.6, kernel 3.10.0-862.14.4, ucode: 0x200004d
- 2 x Intel® Xeon® E5-2699 v4 @ 2.2 GHz, total threads 88, RAM 128 GB, Oracle Linux 7.6, kernel 3.10.0-862.14.4, ucode: 0xB00002E
- 2 x Intel® Xeon® E5-2640 v3 @ 2.6 GHz, total threads 32, RAM 64 GB, Gentoo, kernel 4.10.17, ucode: 0x2B

Looking at the table in Figure 3, the scalability improved 3 times in the best case (SKL 8180) between HEVC SDK 8.2 (9.3) vs. 10.0.1 (32.0), and the deviation from the Reference values didn’t exceed 12% in the worst case (E5-2699 v4). On SKL 8180 the HEVC Decoder 10.0.1 got more than 3 times faster (9.3 vs. 32.0) compared to E5-2699 v4 (9.1 vs. 10.5).

### 5 Summary and Discussion

MainConcept and Intel Architecture Graphics and Software (IAGS) division could increase performance and efficiency for latest Intel® Xeon® Scalable Processors up to 3 times, getting very close to reference efficiency and reaching Intel’s advertised 2.2 times acceleration in between the architectures.

The multi-processing runtime from HEVC/H.265 SDK 8.2 still performs and scales well on Intel® Core™ processors and older Intel® Xeon® processor generations. Even for older Intel® Xeon® processor generations and Intel® Core™ processors the MainConcept HEVC/H.265 SDK 10.0.1 shows slightly improved efficiency and performance. But the most promising conclusion is that for the current generation of Intel® Xeon® Scalable Processors, the MainConcept HEVC Decoder library version 10 shows the most significant speed-up for 10-bit 4:2:2 decoding.

---

**Figure 3.** MainConcept HEVC Decoder reference, SDK 8.2, and SDK 10.0.1 efficiency on Intel® processors.

---

<table>
<thead>
<tr>
<th>HEVC DECODER EFFICIENCY</th>
<th>INTEL® XEON® PLATINUM 8180 x 4</th>
<th>INTEL® XEON® PLATINUM 8168 x 2</th>
<th>INTEL® XEON® E5-2699 V4 x 2</th>
<th>INTEL® XEON® E5-2640 V3 x 2</th>
<th>INTEL® CORE™ i7-8700K</th>
</tr>
</thead>
<tbody>
<tr>
<td>720p@11Mbit</td>
<td>77.5</td>
<td>35.0</td>
<td>24.8</td>
<td>11.0</td>
<td>6.4</td>
</tr>
<tr>
<td>1080p@21Mbit</td>
<td>33.9</td>
<td>15.8</td>
<td>11.3</td>
<td>5.2</td>
<td>3.0</td>
</tr>
<tr>
<td>2160p@32Mbit</td>
<td>12.7</td>
<td>6.2</td>
<td>4.4</td>
<td>2.0</td>
<td>1.2</td>
</tr>
<tr>
<td>4320p@43Mbit</td>
<td>4.0</td>
<td>2.1</td>
<td>1.4</td>
<td>0.7</td>
<td>0.4</td>
</tr>
<tr>
<td>Reference</td>
<td>34.5</td>
<td>16.0</td>
<td>11.9</td>
<td>5.1</td>
<td>2.7</td>
</tr>
<tr>
<td>HEVC SDK 8.2</td>
<td>9.3</td>
<td>11.3</td>
<td>9.1</td>
<td>4.3</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.