Utilizing Intel® Turbo Boost Technology for Managing Dynamic Change in Transcode Requirements on Intel® Quick Sync Video-Enabled Transcoding Servers

Abstract

When serving multiple clients, video transcoding servers must be able to dynamically adjust to changing load conditions on the network. For instance, servers will need to provide a variety of bit streams according to requirements that could vary over time. Requirements may include sizing content according to network capacity, thus dictating the transcode capability and maximum/peak throughput of the servers.

Helping equipment manufacturers gain more performance, video transcoding servers based on Intel® Core™ i7 processors can potentially unlock 30 percent or more additional capacity. This is achievable with Intel® Turbo Boost Technology 2.0, which can increase the frequency of the processor's on-chip GPU when the processor is running below its power limits. This capability, combined with Intel® Quick Sync Video, allows servers to deliver maximum performance to satisfy peak demand, or pull back during non-peak periods to minimize power consumption. This paper details how to manage Intel Turbo Boost Technology for graphics with the Intel® Media SDK for Linux® servers in order to deliver required average and peak capacities on Intel® Quick Sync Video-enabled transcoding servers.

Introduction

As the delivery of live, linear video over IP networks continues to increase, it is important for video transcoding servers to adapt to changing network conditions. As the delivery of live, linear video over IP networks continues to increase, it is important for video transcoding servers to adapt to changing network conditions. Service providers could simply size network capacity to handle the maximum number of transcoded streams requested by clients and transcode every stream into all the various bitrates, codecs, resolutions, etc. However, this isn’t an efficient use of available resources and obviously, is quite expensive.

An alternative to transcoding every stream into a superset of formats is to tailor the output of the transcoders to match the real-time requirements in the network. Consider the example in Figure 1, where a service provider delivers streams to N types of devices; there could be Y unique outputs (light blue) and X common outputs (dark blue), corresponding to a selection of outputs used by all device types. Depending on the time of the day, the output of the transcoder should change to reflect the different ways clients consume content. For instance, people wake up and turn on their TVs and/or tablets to check the news, weather, etc. At work, they use a more diverse set of mobile devices to view different types of content. In the evening, they may stream movies to any of a number of devices.
When sizing a network to support the first model mentioned previously, transcoders must output the maximum common outputs and the maximum unique transcodes, although at some point in the day, some of the content won’t be consumed.

A better use of resources is to dynamically adjust the transcoders in the network, both in terms of the outputs being served and the number of transcoders that are active at any time. Bringing transcoders in and out of service takes time. Using Intel Turbo Boost Technology for graphics, a service provider can temporarily add more load to an existing transcoder until other transcoders can be brought back into service. Afterwards, the transcode load can be moved to the newly running transcoder node in the network, ensuring high quality of service (QoS) for the end consumer.

Similarly, if a transcoder node is transcoding a particular channel and a new version is needed while it is 100 percent loaded, the system can temporarily activate Intel Turbo Boost Technology, while powering up a new transcoder in the background.

There are other ways Intel Turbo Boost Technology can be used to take workload from nodes that are being brought out of service; however, this will be discussed in a later paper.

---

What is Intel® Turbo Boost Technology 2.0 and how does it work to increase graphics performance?

Intel Turbo Boost Technology 2.0 increases graphics performance in much the same way it increases general-purpose computing performance. When the processor has sufficient thermal and power headroom, the frequency can be increased to allow the GPU to run much faster and achieve higher performance. The downside is that as the frequency changes, the supply voltage is increased, and thus power consumption increases and the thermal headroom is used up. More information can be found at the Intel Turbo Boost Technology 2.0 web portal [1].

It is important to note that Intel Turbo Boost Technology provides a potential speed up that is dependent upon the operating conditions. The speed up is determined, in part, by the available thermal headroom, which is impacted by the server cooling solution and the amount of headroom being consumed by the processor and other parts of the device that are in operation. With good thermal conditions and when the processor is not fully utilized, the GPU should be able to run at full speed.

Intel Quick Sync Video, built into the 4th generation Intel® Core™ processors, uses dedicated media processing to increase the performance of video...
processing. When the server runs Intel Quick Sync Video, the processor offloads most of the encoding and decoding to the GPU, so the workload on processor cores decreases. This can lower the heat generated by the processor, allowing the GPU to run at full speed, cooling solution permitting. Given all the variables impacting thermal conditions, the performance gain from Intel Turbo Boost Technology is non-deterministic.

Running in Intel Turbo Boost Technology raises the supply voltage periodically, and thus over time the Intel® Quick Sync Video performance increases, which can lower the heat generated by the processor. However, this can also increase the wear on the silicon. This is reflected in the operating frequencies supported by consumer and embedded SKUs, illustrated in the example shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>INTEL® CORE™ i7-4850HQ PROCESSOR</th>
<th>INTEL® CORE™ i7-4850EQ PROCESSOR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SKU</strong></td>
<td>Consumer</td>
<td>Embedded</td>
</tr>
<tr>
<td>Maximum GPU Frequency without Intel® Turbo Boost Technology</td>
<td>200 MHz</td>
<td>650 MHz</td>
</tr>
<tr>
<td>Maximum GPU Frequency with Intel® Turbo Boost Technology</td>
<td>1.2 GHz</td>
<td>1 GHz</td>
</tr>
</tbody>
</table>

### Table 1: Operating Frequency Comparison of Two Intel® Processors

Both processor SKUs are physically the same device, but the lifetime and usage profiles are very different. The consumer Intel® Core™ i7-4850HQ processor has a nominal mean time before failure (MTBF) (see data sheets for the 4th Generation Intel Core processor of interest) that is suitable for a consumer-grade device. The usage profile for such a device typically keeps the processor cores and GPU idle for significant amounts of time, allowing the processor to default to higher (lower power) C states. For instance, when the system is playing movies and not doing any heavy processing with Intel® HD Graphics or Intel® Iris™ graphics, the processor will fall back to C6 low power state, reduce the supply voltage, and keep the display warmed up and the decode pipeline active.

On the other hand, the Intel® Core™ i7-4850EQ processor was developed for embedded applications that could be active all or most of the time, which may result in a higher MTBF. Consequently, the allowed maximum frequencies for the processor are lower in order to provide a greater guard band on expected lifetime. Hence, the embedded SKUs (Intel® Core™ i7-4860EQ and Intel Core i7-4850EQ processors) are limited to 1 GHz with Intel Turbo Boost Technology enabled compared to the consumer SKU (Intel Core i7-4850HQ processor) at 1.2 GHz.

The performance graph will vary somewhat when comparing higher quality modes, like TU4 and TU1, as well as other resolutions. In general, the performance is similar when it is determined by the operating frequency of the GPU’s Motion Estimation Engine.

As stated earlier, Intel Turbo Boost Technology does not guarantee the GPU will operate above its base frequency; however, with thermal

---

**Maximizing Intel® Quick Sync Video Performance with Intel® Turbo Boost Technology 2.0**

Intel Quick Sync Video is a combination of fixed-function acceleration and some programmable units driven by embedded firmware on the integrated GPU’s execution units. This effectively means the performance of transcoding scales with the frequency of GPU itself. Consider the transcoder operating at Target Usage 7 (TU7), which is the quickest mode of operation. Mapping the base frequency and the turbo frequency against performance for the Intel Core i7-4860EQ processor, Figure 2 shows that with Intel Turbo Boost Technology active, the graphics performance scales from a base of seven simultaneous 1080p30 transcodes to 10 – a potential (depending on the thermal headroom) increase of approximately 40 percent capacity per device.
headroom, the Intel Core i7-4860EQ processor frequency could increase by 250 MHz (above its base of 750 MHz). If deterministic performance is required, it is recommended to size the system using the base frequency with Intel Turbo Boost Technology disabled, and only employ Intel Turbo Boost Technology for non-essential spare capacity.

Controlling the frequency on the GPU

To allow the processor to operate above its base frequency and deliver potentially higher performance, Intel Turbo Boost Technology must be enabled in the platform BIOS.

The system's operating system (OS) has very little control over Intel Turbo Boost Technology. OS tools can be used to determine the current frequency of the GPU, and whether Intel Turbo Boost Technology is permitted, but there will be no information about whether Intel Turbo Boost Technology is active. Also the Intel® 915 graphics driver by default will place the GPU in a low frequency power saving mode, so observing the GPU frequency just after booting will typically show a GPU not in turbo or normal modes of operation; and thus, the frequency may be much lower than the base frequency.

The frequencies are controlled by the RP_STATE_CAP register in the system as described in the Mobile 4th Gen Intel® Core™ Processor Family: Datasheet, Vol. 2 [2].

CAGF (Current Actual Graphics Frequency) will display the current frequency, while the Lowest, Nominal, and Max parameter will provide the low power, base, and turbo frequencies, respectively.

As mentioned previously, the Intel 915 graphics driver will by default set the GPU min frequency to RPN. Then, once the GPU is activated and Intel Turbo Boost Technology is enabled, it will continue to the turbo frequency. When performance isn’t a concern, the system can set the minimum frequency to RP1 to ensure the GPU always runs at the base frequency. To disable Intel Turbo Boost Technology, the system simply sets the max frequency to the same value as the RP1 frequency. To re-enable Intel Turbo Boost Technology,

RP-RP_STATE_CAP

This register contains the maximum base frequency capability for the Integrated GFX Engine (GT).

<table>
<thead>
<tr>
<th>B/D/F/Type: 0/0/0/ME</th>
<th>Access: RO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size: 32</td>
<td></td>
</tr>
<tr>
<td>Default Value: 00000000h</td>
<td>Address Offset: 5998h</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bit Range</th>
<th>Acronym</th>
<th>Description</th>
<th>Default</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>31:24</td>
<td>RSVD</td>
<td>Reserved.</td>
<td></td>
<td>RO</td>
</tr>
<tr>
<td>23:16</td>
<td>RPN_CAP</td>
<td>This field indicates the maximum RPN base frequency capability for the Integrated GFX Engine (GT). Values are in units of 100 MHz.</td>
<td>00h</td>
<td>RO</td>
</tr>
<tr>
<td>15:8</td>
<td>RP1_CAP</td>
<td>This field indicates the maximum RP1 base frequency capability for the Integrated GFX Engine (GT). Values are in units of 100 MHz.</td>
<td>00h</td>
<td>RO</td>
</tr>
<tr>
<td>7:0</td>
<td>RP0_CAP</td>
<td>This field indicates the maximum RP0 base frequency capability for the Integrated GFX Engine (GT). Values are in units of 100 MHz.</td>
<td>00h</td>
<td>RO</td>
</tr>
</tbody>
</table>

RP0 is the graphics frequency with Intel Turbo Boost Technology enabled; RP1 is the base frequency; and RPN is the low power frequency. The values of these registers can be determined by reading a debugfs parameter for i915_cur_delayInfo.

The system can then use the debugfs parameters of the i915 driver to control the min and max frequencies to limit Intel Turbo Boost Technology and also set up the base frequency. Assuming the Intel 915 graphics is the only graphics in the system, the driver parameters can be found under dri/0, as shown in the following:

```
/sys/kernel/debug/dri/0# ls -l…
-r--r--r-- 1 root root 0 Mar 10 10:49 i915_cur_delayInfo
-rw-r--r-- 1 root root 0 Mar 10 10:49 i915_max_freq
-rw-r--r-- 1 root root 0 Mar 10 10:49 i915_min_freq
```

The capability and current GPU frequency can be determined from 1915_cur_delayInfo, as shown in the following:

```
# cat 1915_cur_delayInfo
GT_PERF_STATUS: 0x00000000
RPSTAT1: 0x00048400
Render p-state ratio: 8
Render p-state VID: 0
Render p-state limit: 255
CAGF: 400MHz
RP CUR UP EI: 13228us
RP CUR UP: 5us
RP PREV UP: 0us
RP CUR DOWN EI: 327221us
RP CUR DOWN: 7us
RP PREV DOWN: 0us
Lowest (RPN) frequency: 200MHz
Nominal (RP1) frequency: 750MHz
Max non-overclocked (RP0) frequency: 1000MHz
```
the system resets the max frequency to RP0, and the system will return to turbo frequency when an application makes a new request to the GPU (current requests already made will remain in non-turbo mode). To verify that Intel Turbo Boost Technology has been activated, the system can monitor the current GPU frequency. When the frequency exceeds RP1, the GPU is considered to be operating with Intel Turbo Boost Technology active.

**Putting it all together**

This paper discussed why Intel Turbo Boost Technology is useful, how the GPU performance scales with frequency, and how an application can control the frequency of an Intel Quick Sync Video-enabled device. Using the controls mentioned, a system can be designed to always have the minimum number of live and active transcoders to deliver the required number of transcodes to consumers. As consumers become more active, or change their transcode profiles, the system can use Intel Turbo Boost Technology on Intel Quick Sync Video-enabled nodes to temporarily handle increases in load while other transcode nodes are brought online to handle the additional requests.

To clarify, the figure from the beginning of the paper shows a nominal set of minimum transcodes required. The system can simply have enough transcoders powered up and active for all the transcodes with each transcoder running at base frequency (e.g., if the system must provide 700 transcodes, it just needs just 100 active nodes). At a later point in time, the system needs to allow for two scenarios

a) A consumer migrates from one transcode type to another. Here, the system can enable Intel Turbo Boost Technology on a node serving that consumer and add an additional transcode output while the consumer migrates. Once the consumer has migrated, it can terminate the no longer required transcode and then disable Intel Turbo Boost Technology to return back to a steady state.

b) Additional consumers come online and the system needs additional transcode outputs to be created. The system can enable Intel Turbo Boost Technology on a node serving that content to temporarily provide the output while the network brings on new transcoding nodes. Once the new transcoding node is ready, the transcoding from the Intel Turbo Boost Technology-enabled system is migrated, Intel Turbo Boost Technology can once again be disabled.

Once nodes are no longer required, they can simply be powered down, increasing the lifetime of the processor and reducing operational costs.

**Conclusion**

Utilizing Intel Turbo Boost Technology to manage temporary fluctuations in demand for transcode outputs in a dynamic network can provide much needed assistance in maintaining reliability and increasing cost efficiency while at the same time ensuring a high quality of service to the end consumer.

**References**

[1] Intel® Turbo Boost Technology 2.0