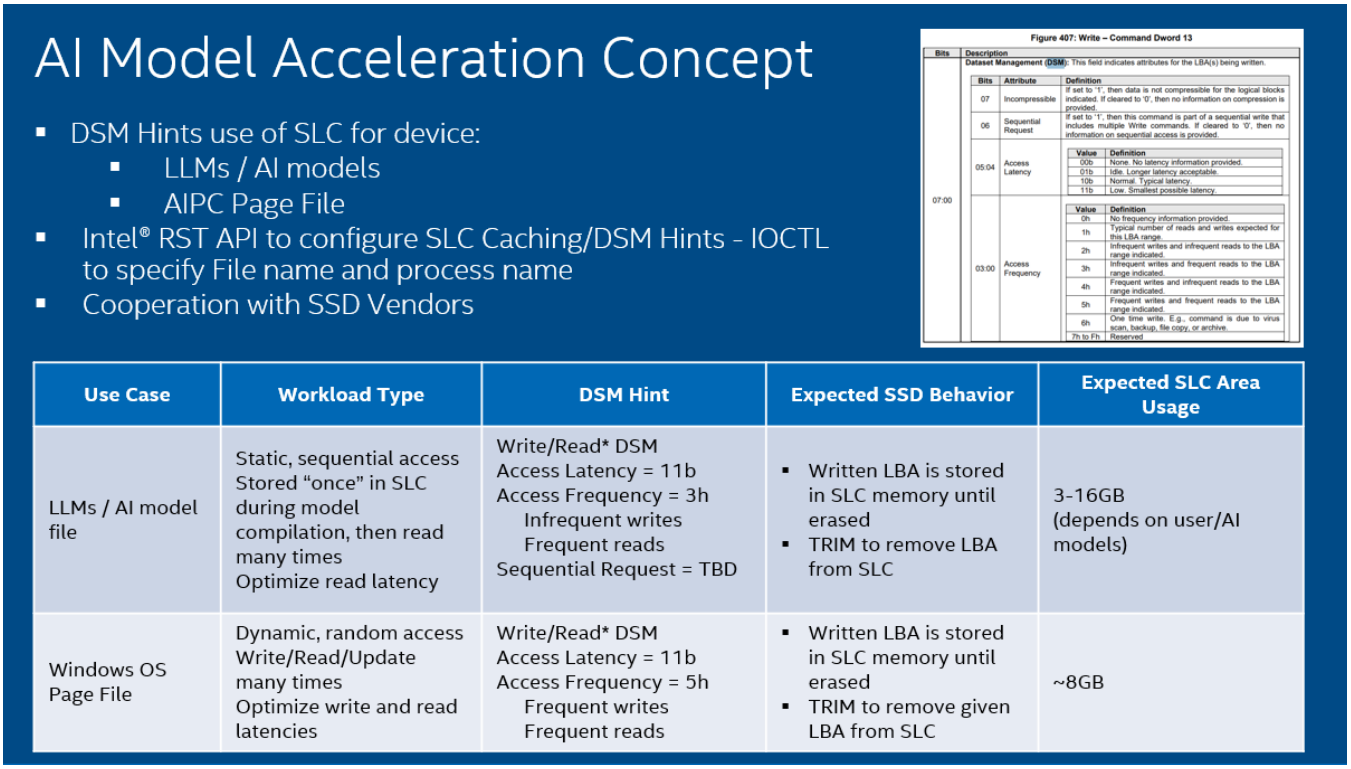
RST POC Driver - DSM Hints

This POC driver is used to evaluate new SSDs which supports DSM Hints feature.

So far, focus on two use cases, refer to below:



AI Model Acceleration Concept

- DSM Hints use of SLC for device:
  - LLMs / AI models
  - AIPC Page File
- Intel® RST API to configure SLC Caching/DSM Hints - IOCTL to specify File name and process name
- Cooperation with SSD Vendors

Figure 407: Write – Command Dword 13

Bits	Description
07:00	**Dataset Management (DSM):** This field indicates attributes for the LBA(s) being written.

Bits	Attribute	Definition
07	Incompressible	If set to '1', then data is not compressible for the logical blocks indicated. If cleared to '0', then no information on compression is provided.
06	Sequential Request	If set to '1', then this command is part of a sequential write that includes multiple Write commands. If cleared to '0', then no information on sequential access is provided.
05:04	Access Latency	Value / Definition: 00b — None. No latency information provided. 01b — Idle. Longer latency acceptable. 10b — Normal. Typical latency. 11b — Low. Smallest possible latency.
03:00	Access Frequency	Value / Definition: 0h — No frequency information provided. 1h — Typical number of reads and writes expected for this LBA range. 2h — Infrequent writes and infrequent reads to the LBA range indicated. 3h — Infrequent writes and frequent reads to the LBA range indicated. 4h — Frequent writes and infrequent reads to the LBA range indicated. 5h — Frequent writes and frequent reads to the LBA range indicated. 6h — One time write. E.g., command is due to virus scan, backup, file copy, or archive. 7h to Fh — Reserved

Use Case	Workload Type	DSM Hint	Expected SSD Behavior	Expected SLC Area Usage
LLMs / AI model file	Static, sequential access Stored "once" in SLC during model compilation, then read many times Optimize read latency	Write/Read* DSM Access Latency = 11b Access Frequency = 3h Infrequent writes Frequent reads Sequential Request = TBD	▪ Written LBA is stored in SLC memory until erased ▪ TRIM to remove LBA from SLC	3-16GB (depends on user/AI models)
Windows OS Page File	Dynamic, random access Write/Read/Update many times Optimize write and read latencies	Write/Read* DSM Access Latency = 11b Access Frequency = 5h Frequent writes Frequent reads	▪ Written LBA is stored in SLC memory until erased ▪ TRIM to remove given LBA from SLC	~8GB

Evaluation Requirements:

1. Intel platform (ex: ARL-H system) - VMD controller enabled and SSDs under VMD
2. RST POC Driver 20.2 and Tools - Link
3. POC SSDs supporting DSM Hints
4. RSTCLI tool in RST 20.2.6.1025 kit

Evaluate Use Case#1 (LLMs files - load/unload)

- Enable NVMe DSM Hints, use below two commands (if NVMe device ID is 2-0-0)

  ```
  NvmePassthroughApp.exe --scsi 0 --path 2 --target 0 --lun 0 configureDsm --enableNvmeHinting 1 --userModeHinting 1 --pageFileHinting 0 --readHinting 1 --writeHinting 0
  NvmePassthroughApp.exe --scsi 0 --path 2 --target 0 --lun 0 addDsmClassification --kind  2 --path \path_to_model_directory\
  ```

- Intel OpenVINO AI app for benchmark:

  openvino.genai/samples/cpp/text_generation/benchmark_genai.cpp at master · openvinotoolkit/openvino.genai · GitHub

  or, use the 2025.4.1.0_x86_64 installer in storage.openvinotoolkit.org
  command options:
  -m <Llama2 7B INT4 OV model> -d GPU -p "The Sky is blue because" --nw 0 -n 1 --mt 20

  ```
  benchmark_genai.exe Help command
  Usage:
    benchmark_vanilla_genai [OPTION...]
  ```

```
  -m, --model arg     Path to model and tokenizers base directory
  -p, --prompt arg    Prompt (default: "")
      --pf arg        Read prompt from file
      --nw arg        Number of warmup iterations (default: 1)
  -n, --num_iter arg  Number of iterations (default: 3)
      --mt arg        Maximal number of new tokens (default: 20)
  -d, --device arg    device (default: CPU)
  -h, --help          Print usage
```

for example:

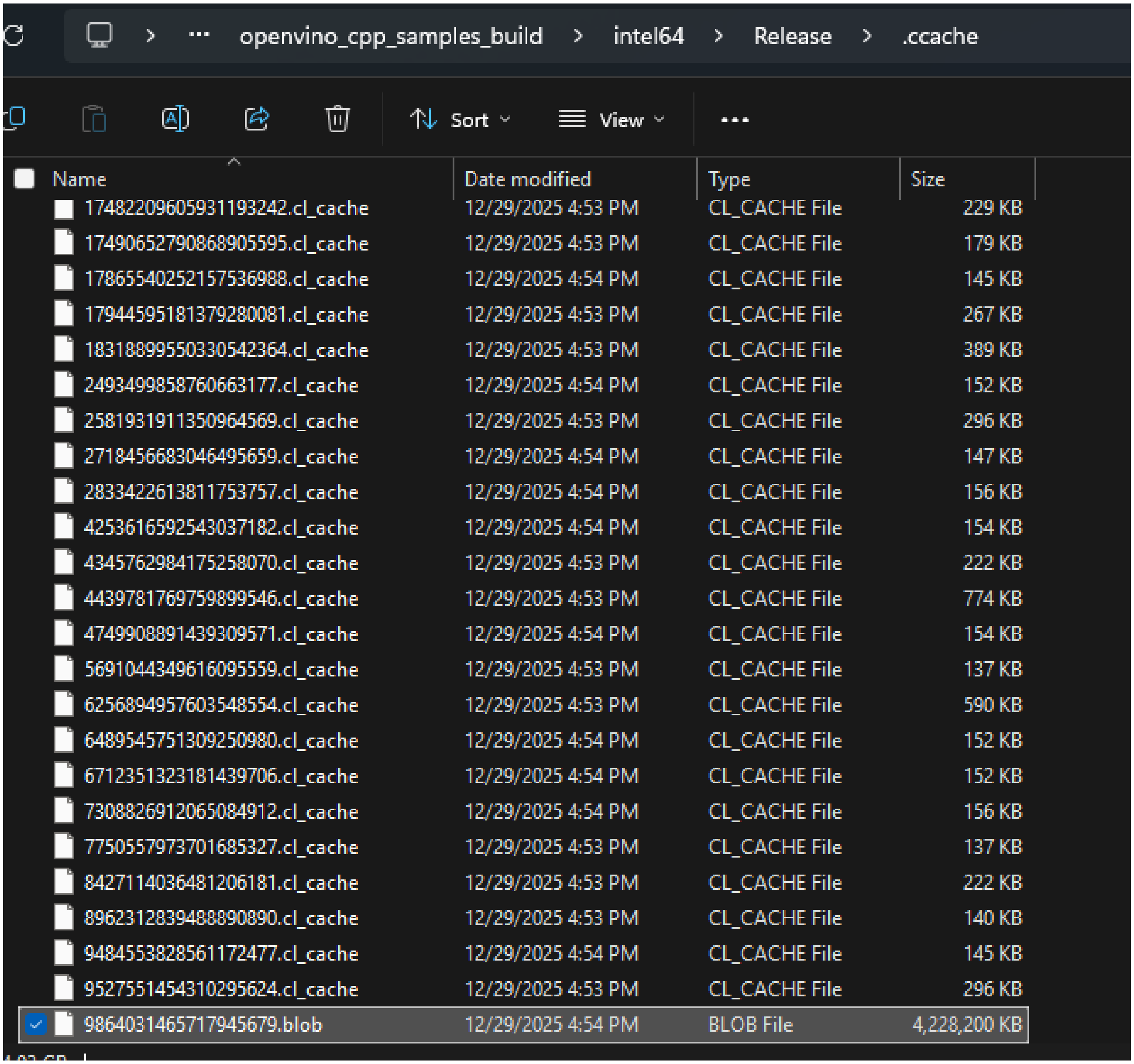
```
benchmark_genai.exe -m C:\dev\pyworkspace\GenAI\models\open_llama_7b_v2-int4-ov -d GPU -p "The Sky is blue because" --nw 0 -n 1 --mt 20
OpenVINO Runtime
    Version : 2025.4.1
    Build   : 2025.4.1-20426-82bbf0292c5-releases/2025/4

Prompt token size:6
Output token size:20
Load time: 11660.00 ms
Generate time: 2532.16  0.00 ms
Tokenization time: 0.31  0.00 ms
Detokenization time: 0.62  0.00 ms
TTFT: 181.35  0.00 ms
TPOT: 123.68  12.68 ms/token
Throughput: 8.09  0.83 tokens/s
```

- Llama2 7B INT4 OV model:

  https://huggingface.co/OpenVINO/open_llama_7b_v2-int4-ov

- Need to modify the code of "benchmark_genai" sample in order to have a compiled model cache in a directory
  benchmark_genai -m .\models\open_llama_7b_v2-int4-ov -d GPU -p "The Sky is blue because" --nw 0 -n 1 --mt 20 --cache_dir ".ccache"



Name	Date modified	Type	Size
17482209605931193242.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	229 KB
17490652790868905595.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	179 KB
17865540252157536988.cl_cache	12/29/2025 4:54 PM	CL_CACHE File	145 KB
17944595181379280081.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	267 KB
18318899550330542364.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	389 KB
2493499858760663177.cl_cache	12/29/2025 4:54 PM	CL_CACHE File	152 KB
2581931911350964569.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	296 KB
2718456683046495659.cl_cache	12/29/2025 4:54 PM	CL_CACHE File	147 KB
2833422613811753757.cl_cache	12/29/2025 4:54 PM	CL_CACHE File	156 KB
4253616592543037182.cl_cache	12/29/2025 4:54 PM	CL_CACHE File	154 KB
4345762984175258070.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	222 KB
4439781769759899546.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	774 KB
4749908891439309571.cl_cache	12/29/2025 4:54 PM	CL_CACHE File	154 KB
5691044349616095559.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	137 KB
6256894957603548554.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	590 KB
6489545751309250980.cl_cache	12/29/2025 4:54 PM	CL_CACHE File	152 KB
6712351323181439706.cl_cache	12/29/2025 4:54 PM	CL_CACHE File	152 KB
7308826912065084912.cl_cache	12/29/2025 4:54 PM	CL_CACHE File	156 KB
7750557973701685327.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	137 KB
8427114036481206181.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	222 KB
8962312839488890890.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	140 KB
9484553828561172477.cl_cache	12/29/2025 4:54 PM	CL_CACHE File	145 KB
9527551454310295624.cl_cache	12/29/2025 4:53 PM	CL_CACHE File	296 KB
☑ 9864031465717945679.blob	12/29/2025 4:54 PM	BLOB File	4,228,200 KB

```
Compiled Cache Dir: compiled_cache
OpenVINO Runtime
    Version : 2025.4.1
    Build   : 2025.4.1-20426-82bbf0292c5-releases/2025/4

Using CACHE_DIR: .ccache
Prompt token size:6
Output token size:20
Load time: 5860.00 ms
Generate time: 1850.92  0.00 ms
Tokenization time: 0.53  0.00 ms
Detokenization time: 0.51  0.00 ms
TTFT: 131.27  0.00 ms
TPOT: 90.47  19.17 ms/token
Throughput: 11.05  2.34 tokens/s
```

Some Discussions/ARs:

1. how to use the driver on ARL-H platform?
2. how to utilize Intel OpenVINO AI app to verify TTFT
3. POC SSDs
4. checkpoint schedule

RST Driver change history:

v3: RST 20.2.0.8335
Fix of ioctl

v2:
change AccessLatency from 2 to 3 for files, processes, directories

v1:
for files, processes, directories: AccessFrequency = 3, AccessLatency = 3

Related articles

- RST POC Driver - DSM Hints