Power Efficient Deep Learning Acceleration using Intel Xeon Processors

Xiaofei Jiang
Intel Asia Pacific
Shanghai, China
xiaofei.jiang@intel.com

Mona Minakshi
Intel Corporation
Hillsboro, OR, USA
mona.minakshi@intel.com

Rajesh Poornachandran
Intel Corporation
Hillsboro, OR, USA
rajesh.poornachandran@intel.com

Shamima Najnin
Intel Corporation
Hillsboro, OR, USA
shamima.najnin@intel.com

***Abstract*—With the exponential growth of AI applications in data center, one of the foremost concerns is power consumption. Intel Optimized Power Mode (OPM) aims to lower power and reduce cooling costs when servers are not at full utilization. Most of the data center deployments keep platform workload mix to be around the 30%~40% utilization range for TCO and handling any spikes[13]. In this paper, performance/watt has been measured on Intel 5th Gen Intel Xeon Scalable Processors using both gen-AI and non-Gen-AI workloads to see the impact of OPM on power consumption. It has been seen that OPM mode yields up to 25% improvement on performance/watt at 25% server utilization. Performance or performance/watt improvement varies depending on the use cases running. Meanwhile when hitting 100% server utilization, the performance/watt using the OPM is like out-of-the-box performance.**

Keywords—AI workload, large language models, optimized power mode, performance per watt

I. INTRODUCTION

The success of AI applications followed by high performance conversational agent like Chat-GPT and a number of large language models (LLMs) has revolutionized the data center industry[1]. LLMs are currently deployed into popular services, for example, Slack, Microsoft Office, and intelligent coding assistants such as GitHub Copilot. Training and deployment of both LLMs/Gen-AI models and non-LLMs/non-Gen-AI models is costly[2][3][4]. One of the foremost concerns is the increased power consumption by the surge of popular commercial AI[5]. Goldman Sachs Research estimates that data center power demand will grow 160% by 2030 due to AI revolution[6]. This significantly triggered massive investment in the physical infrastructure requirements of data center facilities[2][6][7]. Three largest hyperscalers - Amazon, Microsoft, and Google reported 54% increases in datacenter capex investment[8][9]. Current planned power grid infrastructure cannot mitigate this energy crisis due to the exponential AI growth in datacenters[2]. Therefore, even a few percentages of power saving can make a great contribution for reduction of operating and maintenance cost of datacenter. Aiming to reduce power consumption, 4th Gen Intel Xeon Scalable processors has introduced "Optimized Power mode (OPM)" (BIOS configurable option) which can save up to 20% system power on select workloads including SpecJBB, SPECINT and NIGNX key handshake[10]. This OPM BIOS option has been enhanced in the 5th Gen Intel Xeon Scalable processors and named as OPM 2.0[11].

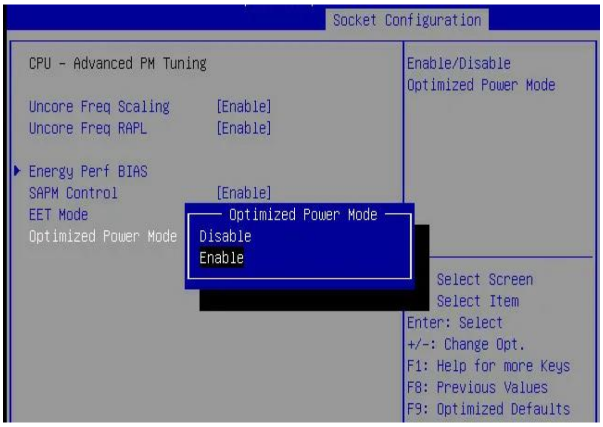
The 5th Gen Intel Xeon Scalable processors[12], formerly codenamed Emerald Rapids (EMR), delivers increased performance for the broad range of data center usages and workloads, including Artificial Intelligence (AI), High Performance Computing (HPC), 5G packet processing, Hyper Converged Infrastructure (HCI), Electronic Design Automation (EDA), and virtualized storage. Predominant contributors of improved performance (perf) and performance per Watt (perf/W) are driven by Core Instruction Per Cycle (IPC) with microarchitecture improvements, larger level 3(L3) cache, memory speed, UPI 2.0 speed, 2 die architecture, and software optimization[11]. Additionally, the 5th Gen Intel Xeon processors are one of most sustainability-enhancing data center processors. Besides of having built-in accelerators for improved perf/W, it can extend power efficiency and savings further with Optimized Power Mode (OPM) enabled in the platform BIOS for workloads.

In this paper, we have considered both Gen-AI (GPT-J-6B and LLaMA-2-7B) and non-Gen-AI (ResNet50-v1.5, Bert-Large, SSD-Resnet34) workloads to analyze the power and performance impact of OPM BIOS knob.

II. DESCRIPTION

With the new Intel 5th Gen Xeon "Emerald Rapids" processors, there is a new feature called the Optimized Power Mode (OPM). This Optimized Power Mode can be enabled via the system BIOS for Emerald Rapids for helping to reduce total cost of ownership (TCO) and improve power efficiency when CPU not running at full utilization. In reality, most of the data center workload mix run at around 30%~40% utilization range for optimal TCO and handling any spikes [13]. Optimized Power Mode 2.0 optimizes for power efficiency in the following ways:

- Enhanced Active Idle Mode (AIM): Improve performance in core-bound applications, by



```
                                        Socket Configuration
CPU - Advanced PM Tuning                            Enable/Disable
                                                    Optimized Power Mode
Uncore Freq Scaling       [Enable]
Uncore Freq RAPL          [Enable]

▶ Energy Perf BIAS
SAPM Control              [Enable]
EET Mode                  ── Optimized Power Mode ──
Optimized Power Mode      Disable
                          Enable
                                                      Select Screen
                                                      Select Item
                                                    Enter: Select
                                                    +/-: Change Opt.
                                                    F1: Help for more Keys
                                                    F8: Previous Values
                                                    F9: Optimized Defaults
```

Fig. 1 BIOS menu to enable/disable OPM mode

Table 1: System & Software Configuration

BIOS Version	EGSDCRB1.SYS.0107.D20.2310211932
OS	CentOS Stream 9
Kernel	6.2.0-emr.bkc.6.2.5.6.33.x86_64
Microcode	0x210001b0
Sockets	2
CPU	INTEL(R) XEON(R) PLATINUM 8592+
Thread(s) per Core	2
NUMA Node(s) per socket	2 (SNC2)
Turbo	Enabled
Power & Perf Policy	Performance
Installed Memory (per socket)	512GB (8x64GB DDR5 5600MT/s rank2)
Huge Pages Size	2048 kB
Transparent Huge Pages	Always
Software Framework	PyTorch +IPEX, https://github.com/intel/models.git

detecting AIM when both core and mesh activities are low. By default, AIM is enabled with Utilization Point (UP)=0; Optimized Power Mode 2.0 enables AIM with UP=3. Here, UP =0 means roughly 0% CPU load while UP=3 means roughly 30% load.

- Cap SOC interconnect frequency at 2.2GHz: limit maximum SOC interconnect frequency because energy efficiency is preferred.
- Disable Perf Power Limit(P-Limit): Decouple SOC interconnect frequency selections on different sockets. This may save power or improve workload performance running on a socket while another socket is idle or runs a different kind of workload.
- Core count aware AIM: improve AIM entry/exit based on cores used and allow dynamic utilization-point threshold setting.

Note that, Optimized Power Mode 1.0 was introduced in the 4th Gen Intel Xeon Processor which included the original AIM with UP=6 and SOC interconnect frequency cap at 2.2GHz.

The BIOS knob, Optimized Power Mode, is found under Socket Configuration > Advanced Power Management Configuration > CPU – Advanced PM (Power Management) Tuning > Optimized Power Mode > Enable/Disable. This feature is disabled by default. The BIOS menu page is shown in Fig. 1.

In this paper, we have considered variety of AI usecases/workloads by enabling and disabling OPM to evaluate the impact on perf/W at different level of CPU utilization on EMR system. Chosen workloads are:

1) *Resnet50_v1.5:* Image Recognition Model
2) *Bert-Large:* Natural Language Process Model
3) *SSD Resnet34:* Object Detection Model
4) *GPT-J-6B, LLaMA-2-7B:* Large Language Models

Our Test System is a two-socket 64 cores EMR system. The detailed system config & software configuration are available in Table 1.

III. RESULTS & ANALYSIS

In this section, we initially present the experiment setup. Then we focus on LLMs analysis, non-LLMs analysis and power consumption at different CPU utilization.

A. *Experiment Setup*

1) *Precision:* The experiment has been carried out using datatypes AVX512 fp32, AMX int8 and AMX bfloat16 for non-LLMs. For LLMs, AMX int8 and AMX bfloat16 are used due to memory size limitation.

2) *Token Size:* Varying input and output token size (1024i128o, 2016i32o) have been considered for LLMs.

3) *Batch Size(BS):* Both BS=1 & BS=x have been considered for LLMs & Non LLMs. For each BS=x case, x is the batch size with the highest performance we found on EMR by batch-size sweeping as shown in Table 2 and Table 3. And BS is kept the same for OPM enabled and disabled for a fair comparison.

4) *CPU Utlization setup:* As OPM mode benefits more while CPU utilization is low, we have run the tests at around 25%, 50%, 75%, and 100% CPU utilization.

Table 2: LLM workload BS information

LLM Workload Name	Token Size	Precision	Batch Size
GPT-J-6B	1024i 128o	amx_bfloat16	10
GPT-J-6B	1024i 128o	amx_int8	18
GPT-J-6B	2016i 32o	amx_bfloat16	10
GPT-J-6B	2016i 32o	amx_int8	12
LLaMA-2-7B	1024i 128o	amx_bfloat16	10
LLaMA-2-7B	1024i 128o	amx_int8	18
LLaMA-2-7B	2016i 32o	amx_bfloat16	10
LLaMA-2-7B	2016i 32o	amx_int8	12

Table 3: non-LLM workload BS information

non-LLM Workload Name	Precision	Batch Size
BERT-LARGE	amx_bfloat16	44
BERT-LARGE	amx_int8	16
BERT-LARGE	avx_fp32	16
ResNet50-v1-5	amx_bfloat16	24
ResNet50-v1-5	amx_int8	64
ResNet50-v1-5	avx_fp32	56
SSD-ResNet34	amx_bfloat16	8
SSD-ResNet34	amx_int8	8
SSD-ResNet34	avx_fp32	56

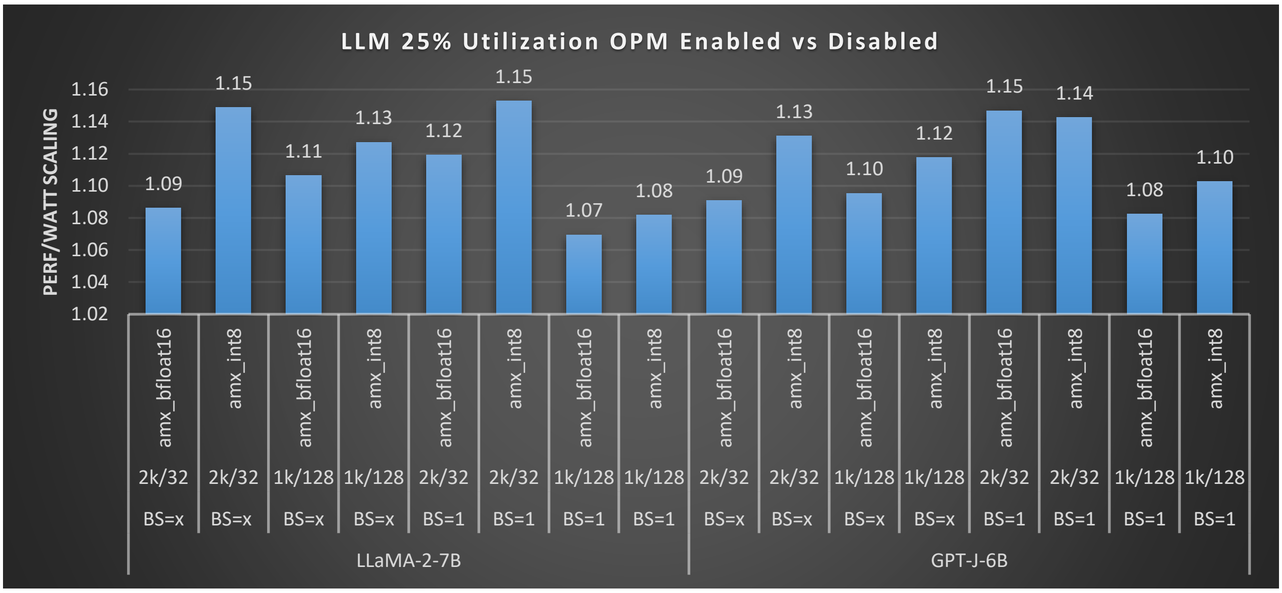


Fig. 2 LLMs performance per watt scaling with OPM enabled vs Disabled at 25% CPU Utilization

Table 4 LLMs uncore frequency and wall power consumption with OPM enabled vs Disabled at 25% CPU Utilization

LLM Workload Name	Token Size	Precision	Batch Size	OPM DIS Uncore (GHz)	OPM EN Uncore (GHz)	OPM Uncore Saving	OPM DIS Wall Power (Watt)	OPM EN Wall Power (Watt)	OPM Wall Power Saving
LLaMA-2-7B	2016i/32o	amx_int8	BS=x	2.4	1.95	18.75%	892.41	759.06	14.94%
LLaMA-2-7B	1024i/128o	amx_bfloat16	BS=x	2.4	2.19	8.75%	946.21	856.00	9.53%
LLaMA-2-7B	1024i/128o	amx_int8	BS=x	2.4	1.92	20.00%	915.09	785.62	14.15%
LLaMA-2-7B	2016i/32o	amx_bfloat16	BS=1	2.4	2.16	10.00%	953.33	836.92	12.21%
LLaMA-2-7B	2016i/32o	amx_int8	BS=1	2.4	2.1	12.50%	938.40	778.75	17.01%
GPT-J-6B	2016i/32o	amx_int8	BS=x	2.4	2.04	15.00%	910.24	773.71	15.00%
GPT-J-6B	1024i/128o	amx_int8	BS=x	2.4	2.14	10.83%	888.99	791.04	11.02%
GPT-J-6B	2016i/32o	amx_bfloat16	BS=1	2.4	2.19	8.75%	947.44	815.53	13.92%
GPT-J-6B	2016i/32o	amx_int8	BS=1	2.4	2.18	9.17%	916.59	770.24	15.97%
GPT-J-6B	1024i/128o	amx_int8	BS=1	2.4	2.19	8.75%	889.90	799.50	10.16%

a) 25% utilization: A single instance of BS=1/BS=x was running on 16 cores out of 64 cores per socket while other cores were idle.

b) 50% utilization: Two instances were running on 32 cores out of 64 cores while leaving 32 cores idle.

c) 75% utilization: Three instances were running on 48 cores out of 64 cores while leaving 16 cores idle.

d) 100% utilization: Four instances were running using all the 64 cores in the socket.

5) Performance and Power collection: For each run, the consumed wall power & socket power have been collected using power meter (for wall power) and EMON (an internal telemetry tool for socket power) along with throughput performance data. Using measured throughput and wall power, perf/W has been computed. To avoid any outlier, all the reported numbers are the average over two runs.

B. LLMs Analysis

1) 25% CPU utlization Analysis: LLMs perf/W scaling of OPM enabled versus disabled at 25% CPU utilization is shown in Fig. 2 where 1k/128 represents1024i/128o token size and 2k/32 represents 2016i32o token size. It can be seen that upto 15% perf/W can be improved at 25% utilization using OPM feature. And all cases had at least 7% perf/W improvement. We show the average uncore frequency and wall power consumption of the LLMs cases with more than 10% perf/W saving in Table 4. At 25% Utilization with OPM feature disabled, uncore frequency was kept 2.4GHz. When enabling OPM feature, uncore frequency was at most 20% reduced and wall power was at most 17.01% reduced. And meanwhile LLMs performance was at most 4.95% reduced. When running LLMs at a CPU utilization level of 25%, OPM feature substantially reduced both the uncore frequency and the wall power consumption. Notebly, this reduction was achieved with minimal impact on performance, thereby, enhancing the efficiency of power usage as measured by perf/W.

2) 50% CPU utlization Analysis: Fig. 3 illustrates LLMs perf/W scaling of OPM feature enabled versus disabled at 50% Utilization. The data indicates that enabling OPM

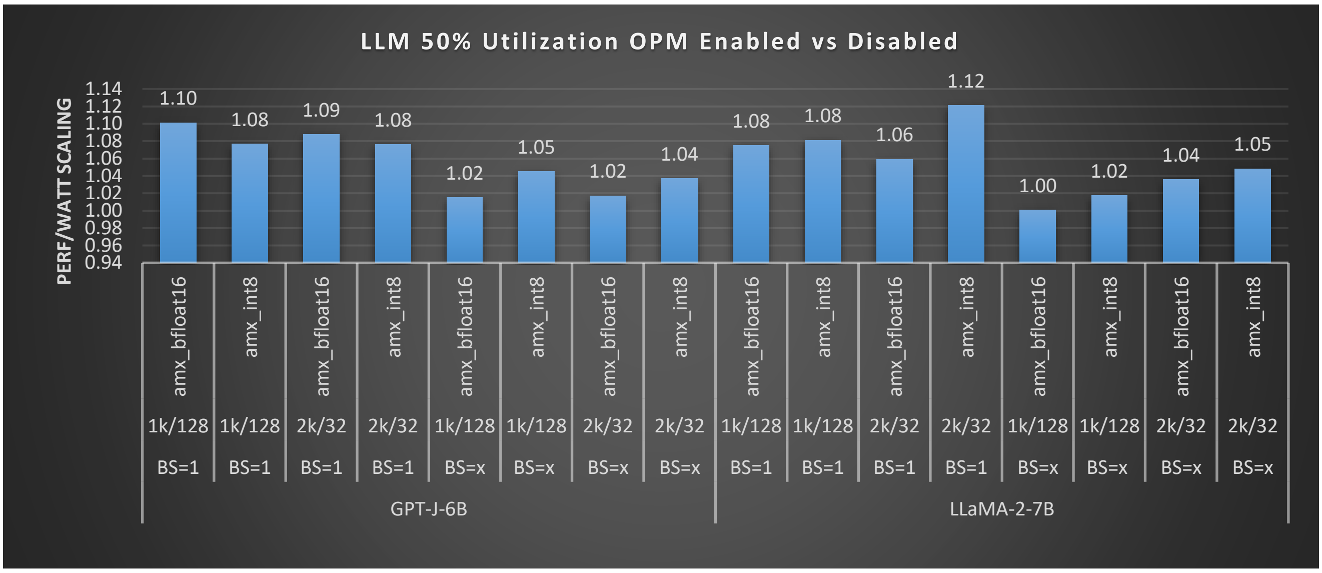


Fig. 3 LLMs performance per watt scaling with OPM enabled vs Disabled at 50% CPU Utilization

3) features yeilded up to 12% perf/W improvement. Furthermore, LLMs perfomrance remained stable and experienced a negligible decline of no more than 1.36%. The activation of OPM feature caused a reduction in wall power consumption by upto 9.8% and a decrease in uncore frequency by as much as 8.33%.

4) 75% and 100% CPU utilization Analysis: There's no obvious performance and power change when the CPU utilization was 75% or 100%.

For LLMs power efficiency with OPM features, it has been noted that perf/W is significantly improved at lower CPU utilization levels, such as 25% and 50%. The power efficiency improvement is primarily coming from reduction of uncore frequency and wall power. When CPU utilization is higher, such as 75% and 100%, OPM feature does not impact power and performance.

C. Non-LLMs Analysis

1) 25% CPU utilization Analysis: Fig. 4 presents perf/W scaling for non-LLMs at 25% CPU utilization, comparing between OPM feature enabled and disabled. There're at least 5% and at most 23% perf/W efficiency improved when OPM feature is enabled. Delving into the specifics of uncore frequency and wall power consumption of the non-LLMs test cases whose perf/W saving were larger than 10% as shown in Table 5, at 25% Utilization, uncore frequency was significantly dropped from 2GHz to around 1.4GHz. There is upto 40.34% uncore frequency reduction. Additionally, power consumption had a maximum reduction of 24.88%. While the impact on performance was relatively minor, with a maximum reduction of 10.39%. This indicates that in the case of running non-LLMs at CPU utilization level of 25%, the decrease in uncore frequency and power consumption is more pronounced than the decrease in performance, thereby resulting in improved perf/W efficiency.

2) 50% CPU utilization Analysis: Non-LLMs perf/W scaling at 50% CPU utilization is shown in Fig. 5. There is up to 27% perf/W saved with OPM features. Non-LLMs had at most 3.33% performance reduction with OPM features. While there are at most 17.83% wall power reduction and 35.56% uncore frequency reduction which contributed to perf/W saving.

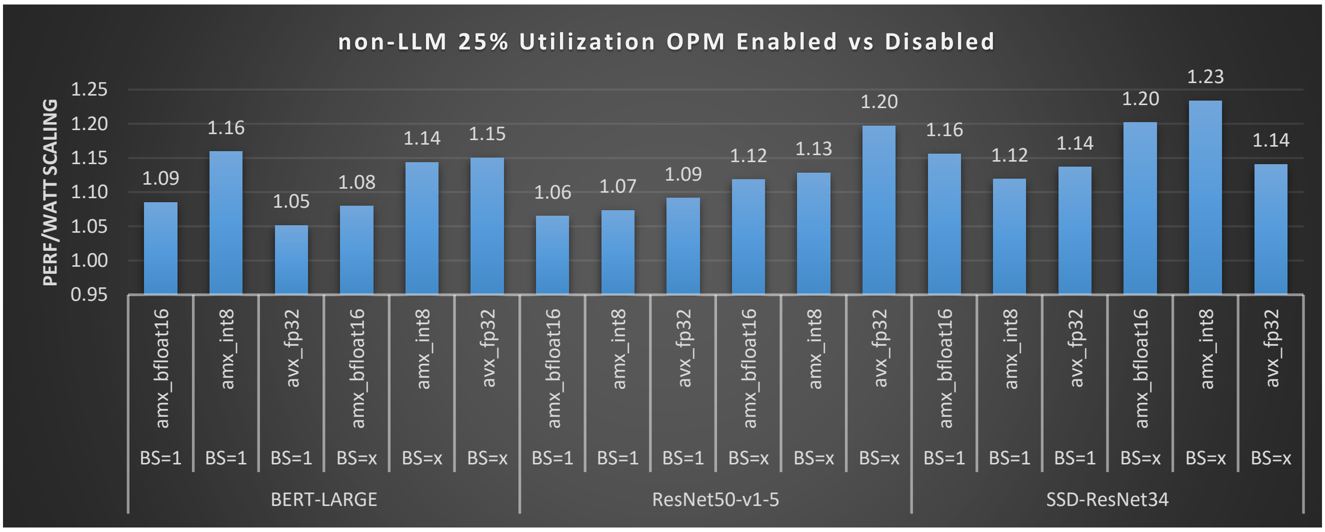


Fig. 4 non-LLMs performance per watt scaling with OPM enabled vs Disabled at 25% CPU Utilization

Table 5 non-LLMs uncore frequency and wall power consumption with OPM enabled vs Disabled at 25% CPU Utilization

non-LLMs Workload Name	Batch Size	Precision	OPM DIS Uncore (GHz)	OPM EN Uncore (GHz)	OPM Uncore Saving	OPM DIS Wall Power (Watt)	OPM EN Wall Power (Watt)	OPM Wall Power Saving
BERT-LARGE	BS=1	amx_int8	2.23	1.48	33.63%	777.29	601.05	22.67%
BERT-LARGE	BS=x	amx_int8	2.35	1.46	37.87%	840.52	658.72	21.63%
BERT-LARGE	BS=x	avx_fp32	2.33	1.46	37.34%	1077.73	890.02	17.42%
ResNet50-v1-5	BS=x	amx_bfloat16	2.28	2.04	10.53%	837.22	752.50	10.12%
ResNet50-v1-5	BS=x	amx_int8	2.4	1.61	32.92%	839.59	680.47	18.95%
ResNet50-v1-5	BS=x	avx_fp32	2.39	1.43	40.17%	1088.56	936.01	14.01%
SSD-ResNet34	BS=1	amx_bfloat16	2.25	1.83	18.67%	792.96	645.03	18.65%
SSD-ResNet34	BS=1	amx_int8	2.27	1.83	19.38%	767.52	667.14	13.08%
SSD-ResNet34	BS=1	avx_fp32	2.38	1.42	40.34%	1070.81	891.85	16.71%
SSD-ResNet34	BS=x	amx_bfloat16	2.33	1.4	39.91%	715.42	543.68	24.01%
SSD-ResNet34	BS=x	amx_int8	2.32	1.41	39.22%	675.66	507.54	24.88%
SSD-ResNet34	BS=x	avx_fp32	2.39	1.43	40.17%	1062.15	881.03	17.05%

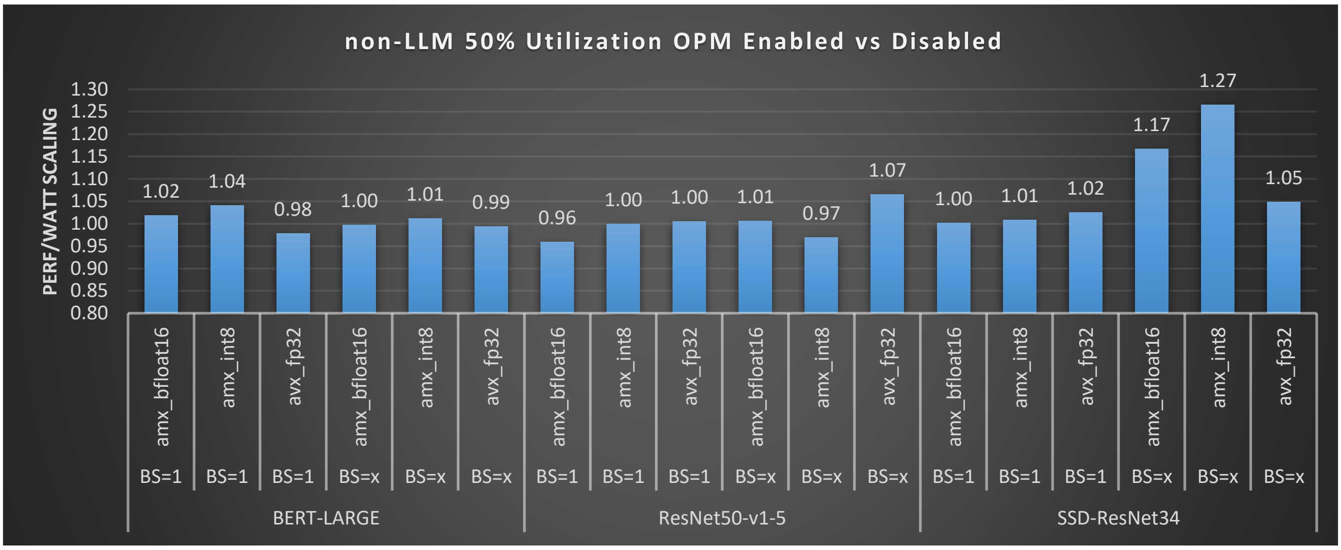


Fig. 5 non-LLMs performance per watt scaling with OPM enabled vs Disabled at 25% CPU Utilization

3) 75% and 100% CPU utilizationo Analysis: we did not observe obvious performance and power change when the CPU utilization was 75% or 100%.

For non-LLMs power efficiency with OPM features, we have observed similar behaviors as LLMs that perf/W was significantly improved at lower CPU utilization levels, such as 25% and 50%. The power efficiency improvement mainly came from reduction of uncore frequency and wall power while performance got slight regression. When CPU utilization was higher, such as 75% and 100%, OPM feature did not impact power and performance.

D. Power Consumption among Different CPU Utilization

Among all the test cases, SSD-Resnet34 BS=8 AMX int8 stood out for achieving the highest perf/W scaling improvement with OPM feature. So, this particular case was selected as a representative example for power consumption analysis among different CPU utilization. SSD-Resnet34 BS=8 AMX int8 wall power and package power consumption with different CPU utilization and different OPM setting are shown in Fig. 6 and Fig. 7. On two sockets, with OPM features both wall power and package power consumption were reduced at 25%, 50%, 75% and 100% CPU utilization. Especially at Utilization 25%, there're up to 162W wall power and 112W package power saved, together with 39.22% uncore frequency reduction. while the performance reduction was within 2%, which contributed to at most 23% perf/W improvement.

We have also observed similar behavior on other test cases as well.

Note that, all the Performance results in this paper are based on testing shown in configurations and may not reflect all publicly available updates. Results may vary depending on hardware and software configuration.

IV. CONCLUSION AND FUTURE PLAN

A. Conclusion

In this paper, we have carried out the performance and power analysis of AI workload to see the impact of OPM feature on EMR systems. Measured perf/W data of non-Gen-AI and Gen-AI workloads shows that OPM feature helps to save a decent amount power while server is running around or

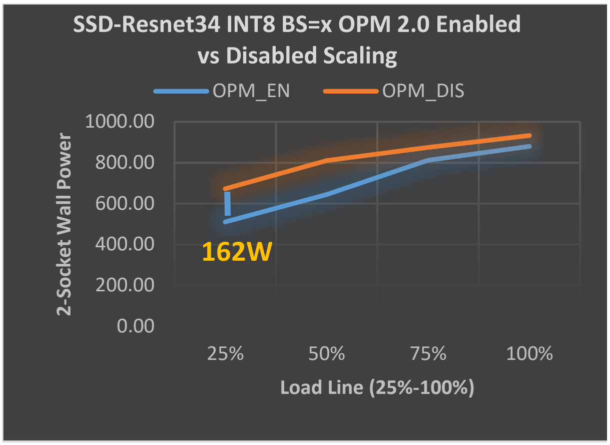


Fig. 6 SSD-Resnet34 2-socket wall power consumption with OPM enabled vs disabled

lower than 50% CPU utilization. With the growing power hungriness due to the increasing demand of AI applications in datacenter, this power saving can play a crucial role to reduce infrastructure cost as well as maintenance cost.

B. Furture Plan

In future work, firstly, we'll target to scale to full end to end AI pipeline and evaluation on lower core count SKUs. Secondly, we'll analyze the percentage of mixed precision and how it varies with the power savings versus accuracy tradeoffs.

REFERENCES

[1] A. Berthelot et al., "Estimating the environmental impact of Generative-AI services using an LCA-based methodology," Procedia CIRP, vol. 122, pp. 707-712, 2024.

[2] S. Luccioni, Y. Jernite, and E. Strubell, "Power hungry processing: Watts driving the cost of AI deployment?" in Proc. 2024 ACM Conf. Fairness, Accountability, and Transparency, 2024.

[3] L. Lin et al., "Exploding AI Power Use: an Opportunity to Rethink Grid Planning and Management," in Proc. 15th ACM Int. Conf. Future Sustainable Energy Syst, 2024.

[4] A. A. Chien et al., "Reducing the Carbon Impact of Generative AI Inference (today and in 2035)," in Proc. 2nd Workshop Sustainable Comput. Syst. (HotCarbon '23), Association for Computing Machinery, New York, NY, USA, Article 11, 7 pages, 2023. [Online]. Available: https://doi.org/10.1145/3604930.3605705

[5] J. Vanian and K. Leswing, "ChatGPT and generative AI are booming, but the costs can be extraordinary," 2023. [Online]. Available: https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html

[6] "AI poised to drive 160% increase in power demand," 2023. [Online]. Available: https://www.goldmansachs.com/intelligence/pages/AI-poised-to-drive-160-increase-in-power-demand.html

[7] A. Radovanović et al., "Carbon-aware computing for datacenters," IEEE Trans. Power Syst., vol. 38, no. 2, pp. 1270–1280, 2022.

[8] "Meta expects first shipments of new Nvidia chips later this year," 2024. [Online]. Available: https://www.reuters.com/technology/meta-does-not-expect-new-nvidia-chips-arrive-until-least-next-year-2024-03-19/

[9] "There's AI in them thar Hills," May 2023. [Online]. https://www.economist.com/business/2023/05/29/nvidia-is-not-the-only-firm-cashing-in-on-the-ai-gold-rush

[10] "Intel Launches 4th Gen Xeon Scalable Processors, Max Series CPUs," 2023. [Online]. Available: https://www.intel.com/content/www/us/en/newsroom/news/4th-gen-xeon-scalable-processors-max-series-cpus-gpus.html

[11] "What Improvements and Features Do the 5th Generation Intel® Xeon® Scalable Processors Deliver?" 2023. [Online]. Available: https://www.intel.com/content/www/us/en/support/articles/000097274/processors.html.

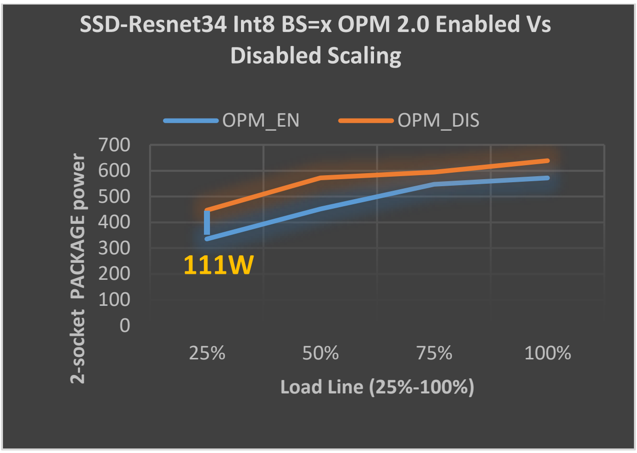


Fig. 7 SSD-Resnet34 2-socket package power consumption with OPM enabled vs disabled.

[12] "5th Gen Intel® Xeon® Processors," 2023. [Online]. Available: https://www.mouser.com/pdfDocs/5th-gen-xeon-processors-product-brief.pdf

[13] Meisner, David, Brian T. Gold, and Thomas F. Wenisch. "Powernap: eliminating server idle power." ACM SIGARCH Computer Architecture News 37.1 (2009): 205-216.