

1-bit AI Infra: Part 1.1, Fast and Lossless BitNet b1.58 Inference on CPUs

Jinheng Wang*, Hansong Zhou*, Ting Song*, Shaoguang Mao,
Shuming Ma, Hongyu Wang, Yan Xia, Furu Wei[◇]
Microsoft Research
<https://aka.ms/GeneralAI>

Abstract

Recent advances in 1-bit Large Language Models (LLMs), such as BitNet [WMD⁺23] and BitNet b1.58 [MWM⁺24], present a promising approach to enhancing the efficiency of LLMs in terms of speed and energy consumption. These developments also enable **local LLM** deployment across a broad range of devices. In this work, we introduce **bitnet.cpp**, a tailored software stack designed to unlock the full potential of 1-bit LLMs. Specifically, we develop a set of kernels to support **fast** and **lossless** inference of ternary BitNet b1.58 LLMs on CPUs. Extensive experiments demonstrate that bitnet.cpp achieves significant speedups, ranging from 2.37x to 6.17x on x86 CPUs and from 1.37x to 5.07x on ARM CPUs, across various model sizes. The code is available at aka.ms/bitnet.

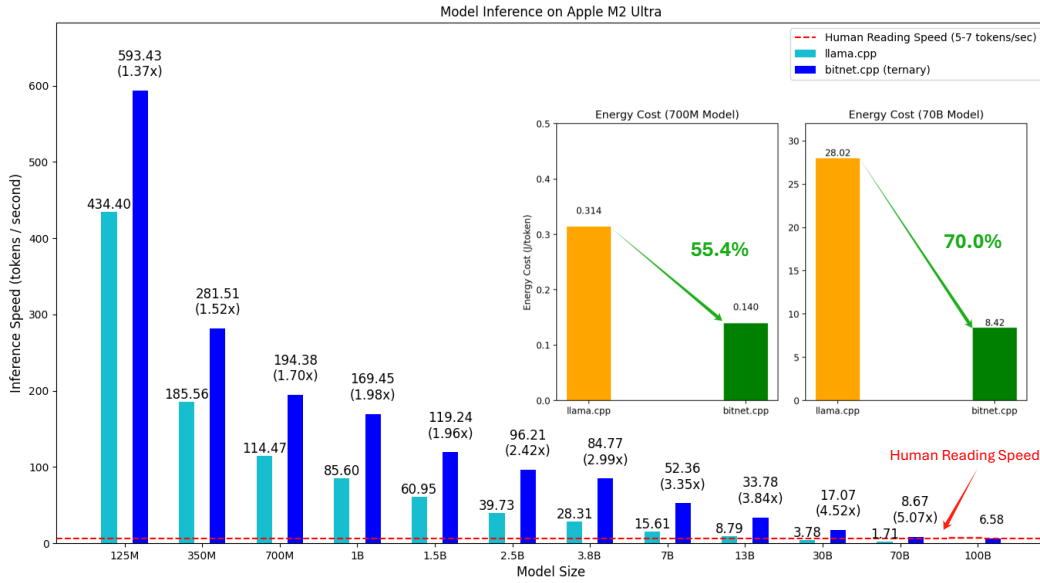


Figure 1: Comparison of **inference speed** and **energy consumption** for various BitNet b1.58 model sizes on an Apple M2 Ultra (ARM CPU) using llama.cpp(fp16) [lla] versus bitnet.cpp (ternary kernels). The results demonstrate that bitnet.cpp can achieve human reading speed, even for a 100B model on a single CPU. Notably, bitnet.cpp significantly reduces energy consumption across different model sizes.

* Equal contribution. [◇] Corresponding author. T. Song, S. Mao, S. Ma, Y. Xia, F. Wei are with Microsoft Research. J. Wang and H. Zhou are with Peking University. H. Wang is with University of Chinese Academy of Sciences.

1 bitnet.cpp

bitnet.cpp is an inference framework for 1-bit LLMs (e.g., BitNet b1.58 models). It provides **lossless inference** while optimizing both **speed and energy consumption**. The initial release of bitnet.cpp supports inference on CPUs.

As illustrated in Figure 1, bitnet.cpp achieves speedups ranging from 1.37x to 5.07x on ARM CPUs, with larger models experiencing greater performance gains. Additionally, it reduces energy consumption by 55.4% to 70.0%, further boosting overall efficiency. On x86 CPUs, speedups range from 2.37x to 6.17x with energy reductions between 71.9% and 82.2%. Furthermore, bitnet.cpp can run a 100B BitNet b1.58 model on a single CPU, achieving speeds comparable to human reading (5-7 tokens per second)[Bry19], thus significantly enhancing the potential for running LLMs on local devices.

To start using bitnet.cpp for inference, follow these steps:

```
1 # Clone the repo
2 git clone git clone --recursive https://github.com/microsoft/BitNet.git
3 cd BitNet
4
5 # Create a new conda environment (Recommended)
6 conda create -n bitnet-cpp python=3.9
7 conda activate bitnet-cpp
8 pip install -r requirements.txt
9
10 # Download the model from Hugging Face, convert it to quantized gguf
    format, and build the project
11 # These models were neither trained nor released by Microsoft. We used
    them to demonstrate the inference capabilities of bitnet.cpp
12 python setup_env.py --hf-repo HF1BitLLM/Llama3-8B-1.58-100B-tokens -q
    i2_s
13
14 # Or you can manually download the model and run it using a local path
15 huggingface-cli download HF1BitLLM/Llama3-8B-1.58-100B-tokens --local-dir
    models/Llama3-8B-1.58-100B-tokens
16 python setup_env.py -md models/Llama3-8B-1.58-100B-tokens -q i2_s
17
18 # Run inference with the quantized model, use -m to specify the model
    path, -p to specify the prompt
19 python run_inference.py -m models/Llama3-8B-1.58-100B-tokens/ggml-model-
    i2_s.gguf -p "Daniel went back to the the the garden. Mary travelled
    to the kitchen. Sandra journeyed to the kitchen. Sandra went to the
    hallway. John went to the bedroom. Mary went back to the garden.
    Where is Mary?\nAnswer:"
```

2 Optimized Kernels for 1.58-bit Models

bitnet.cpp offers a suite of optimized kernels, including I2_S, TL1 and TL2. The kernels are designed for fast and lossless inference of 1.58-bit models on both x86 and ARM architectures.

Unpack	Pack
-1	00
0	01
1	10

Table 1: I2_S Kernel transforms each full-precision weight into a 2-bit value to save memory and bandwidth. When performing computation, the 2-bit weights are unpacked to the original values.

I2_S Kernel adopts the vanilla multiply-then-addition manner to perform the matrix multiplication. As shown in Table 1, it transforms each full-precision weight into a 2-bit representation offline.

During computation, it transforms the weights back to their original values and performs the vanilla GEMV operations. We recommend using it with sufficient threads, since it allows the compiler to generate efficient pipelined instruction sequences.

Unpack		Pack
-1	-1	0000
-1	0	0001
-1	1	0010
0	-1	0011
0	0	0100
0	1	0101
1	-1	0110
1	0	0111
1	1	1000

Table 2: TL1 Kernel transforms every two full-precision weights into 4-bit index and performs LUT computation.

TL1 Kernel preprocesses every two full-precision weights by packing them into 4-bit index (see Table 2), and pre-computes their corresponding activations into $3^2 = 9$ values. The index-value pairs are stored in a lookup table to perform LUT computation [PPK⁺22, WCC⁺24]. GEMV processing is performed using an int16 LUT and accumulation through addition. We recommend using it with a limited number of threads when serving large models.

Unpack			Pack
-1	-1	-1	1 1101
-1	-1	0	1 1100
-1	-1	1	1 1011
-1	0	-1	1 1010
...			
0	0	0	0 0000
...			
1	0	1	0 1010
1	1	-1	0 1011
1	1	0	0 1100
1	1	1	0 1101

Table 3: TL2 Kernel compresses every three full-precision weights into a 1-bit sign (0 or 1) and a 4-bit index.

TL2 Kernel is similar to TL1. The major difference is that it compresses every three weights into a 5-bit index, while TL1 compresses every two weights into a 4-bit index. Therefore, TL2 achieves a higher compression ratio than TL1. We recommend using it in environments with limited memory or bandwidth, since it employs LUT and reduces model size by 1/6 compared to TL1 Kernel, thereby lowering bandwidth requirements.

3 Evaluation

3.1 Inference Performance

We evaluated bitnet.cpp in terms of both **inference speed** and **energy cost**. Comprehensive tests were conducted on models with various parameter sizes, ranging from 125M to 100B. specific configurations for each model are detailed in the Appendix A. These sizes represent popular LLM configurations. Additionally, systematic tests were performed on both ARM and x86 architectures. For ARM, we used a Mac Studio with an Apple M2 Ultra processor and 64GB of memory for

end-to-end tests. For x86, a Surface Laptop Studio 2 with an Intel Core i7-13700H processor (14 cores, 20 threads) and 64GB of memory was used.

We tested two scenarios for each device: one with inference limited to two threads, and the other without thread restrictions, reporting the optimal inference speed. It was to consider the limited thread availability on local devices, providing a more accurate performance assessment of BitNet b1.58 in local environments.

Inference Speed: Table 4 and Table 5 demonstrate significant performance advantages of bitnet.cpp over llama.cpp on both ARM (Apple M2) and x86 (Intel i7-13700H) architectures, especially as model sizes increase. bitnet.cpp consistently outpaces llama.cpp, with speedups ranging from 1.37x to 6.46x, depending on the model and architecture. On the Apple M2, speedups peak at 5.07x in the unlimited thread scenario, while on the Intel i7-13700H, bitnet.cpp achieves up to 6.46x in thread-limited scenarios, making it particularly effective for local inference on resource-constrained systems.

The performance gap widens as models scale up, with larger models (13B and above) benefiting the most from bitnet.cpp’s optimizations. On the Intel i7-13700H, bitnet.cpp provides substantial speed improvements, making it well-suited for x86 architecture, even with limited threads. While smaller models (125M to 1B) also see meaningful gains, the advantages of bitnet.cpp become especially critical for larger, more complex models, underscoring its efficiency in handling demanding workloads. Adding to the observed performance differences, bandwidth limitations play a significant role in the varying efficacy of bitnet.cpp across different architectures, particularly when comparing the Apple M2 and Intel i7-13700H. Due to the larger bandwidth of the Apple M2, it achieves significantly faster speed improvements with bitnet.cpp compared to the Intel i7-13700H, especially when running larger models.

CPU	Kernel	125M	350M	700M	1B	1.5B	2.5B	3.8B	7B	13B	30B	70B	100B
APPLE M2	llama.cpp	434.40	186.56	114.47	85.60	60.95	39.73	28.31	15.61	8.79	3.78	1.71	N/A
	bitnet.cpp	593.43 (1.37x)	281.51 (1.51x)	194.38 (1.70x)	169.45 (1.98x)	119.24 (1.96x)	96.21 (2.42x)	84.77 (2.99x)	52.36 (3.35x)	33.78 (3.84x)	17.07 (4.51x)	8.67 (5.07x)	6.58 (N/A)
Intel i7-13700H 20C 64G	llama.cpp	164.04	56.67	30.73	22.31	15.02	11.07	5.85	3.30	1.78	N/A	N/A	N/A
	bitnet.cpp	389.08 (2.37x)	172.95 (3.05x)	119.08 (3.88x)	86.50 (3.88x)	67.12 (4.47x)	46.33 (4.19x)	30.51 (5.22x)	18.75 (5.68x)	10.99 (6.17x)	5.10 (N/A)	2.44 (N/A)	1.70 (N/A)

Table 4: Comparison of inference speed across different CPUs (Unit: Tokens/Second) in an unlimited thread setting. "N/A" indicates that the tested CPU cannot host the specified model size with the given kernel.

CPU	Kernel	125M	350M	700M	1B	1.5B	2.5B	3.8B	7B	13B	30B	70B	100B
APPLE M2	llama.cpp	251.95	95.18	53.93	41.36	26.67	17.61	11.88	6.71	3.73	1.60	0.71	N/A
	bitnet.cpp	401.76 (1.59x)	168.88 (1.77x)	96.34 (1.79x)	79.23 (1.92x)	56.31 (2.11x)	37.29 (2.12x)	26.75 (2.25x)	15.26 (2.27x)	8.75 (2.35x)	3.89 (2.43x)	1.76 (2.48x)	1.27 (N/A)
Intel i7-13700H 20C 64G	llama.cpp	119.84	41.57	18.56	13.92	8.99	6.95	3.49	1.92	1.30	N/A	N/A	N/A
	bitnet.cpp	316.35 (2.64x)	137.68 (3.31x)	80.13 (4.32x)	57.76 (4.15x)	44.69 (4.97x)	29.41 (4.23x)	20.51 (5.88x)	12.41 (6.46x)	7.09 (5.45x)	3.23 (N/A)	1.51 (N/A)	0.97 (N/A)

Table 5: Comparison of inference speed across different CPUs (Unit: Tokens/Second) in a thread-limited setting, where the number of available inference threads is set to 2. "N/A" indicates that the tested CPU cannot host the specified model size with the given kernel.

Energy Cost: We ran 700M, 7B and 70B models and reported the energy cost (J/token) with the best inference speed in the unlimited thread setting. Table 6 demonstrates a clear advantage of bitnet.cpp in reducing energy consumption. For the Apple M2, bitnet.cpp reduces energy usage by 55.4% to 70.0% depending on the model size. As model size increases, bitnet.cpp’s energy efficiency becomes more pronounced, with the largest model (70B) showing a 70.0% reduction in energy consumption compared to llama.cpp. This highlights bitnet.cpp’s ability to deploy large-scale inference more efficiently, both in terms of speed and energy usage, which is crucial for energy-constrained environments such as mobile devices or edge computing.

On the Intel i7-13700H, energy savings with bitnet.cpp are even more dramatic, ranging from 71.9% to 82.2% for models up to 7B. Although energy consumption data for the 70B model on the Intel

CPU is unavailable, the results for smaller models clearly show that bitnet.cpp can significantly lower the energy demands of large language model inference on high-performance, multi-core processors.

CPU	Kernel	700M	7B	70B
APPLE M2	llama.cpp	0.314	3.013	28.02
	bitnet.cpp	0.140	1.068	8.42
	saving	55.4%	64.6%	70.0%
Intel i7-13700H 20C 64G	llama.cpp	1.367	11.305	N/A
	bitnet.cpp	0.384	2.017	17.33
	saving	71.9%	82.2%	N/A

Table 6: Comparison of Energy Costs Across CPUs (Unit: J/Token). "N/A" indicates that the specific model size cannot be hosted on the tested CPU with the given kernel.

3.2 Inference Accuracy

The bitnet.cpp framework enables lossless inference for ternary BitNet b1.58 LLMs. To evaluate **inference accuracy**, we randomly selected 1,000 prompts from WildChat [ZRH⁺24] and compared the outputs generated by bitnet.cpp and llama.cpp to those produced by an FP32 kernel. The evaluation was conducted on a token-by-token basis, with a maximum of 100 tokens per model output, considering an inference sample lossless only if it exactly matched the full-precision output.

This evaluation used a 700M BitNet b1.58 model². The results confirm that bitnet.cpp achieves accurate, lossless inference for 1-bit LLMs.

Kernel	llama.cpp		bitnet.cpp		
	TQ1_0	TQ2_0	I2_S	TL1	TL2
Accuracy	1.4%	1.4%	100%	100%	100%

Table 7: Comparison of inference accuracy between llama.cpp and bitnet.cpp. TQ1_0 and TQ2_0 are kernels in llama.cpp, while I2_S, TL1, and TL2 are kernels in bitnet.cpp. Accuracy indicates the proportion of lossless inference samples, where outputs matched exactly with the full-precision baseline.

4 Future Work

We are expanding bitnet.cpp to support a broader range of platforms and devices, including mobile devices (e.g., iPhone and Android), NPUs, and GPUs. We will also work on 1-bit LLM training optimization in the future. Furthermore, we are interested in the co-design of customized hardware and software stacks for 1-bit LLMs.

Acknowledgement

We express our sincere gratitude to Georgi Gerganov and the entire llama.cpp community, whose work served as the foundation for our implementation of BitNet b1.58 kernels. Additionally, we extend our thanks to our colleagues and fellow interns at Microsoft Research Asia for their invaluable discussions and feedback. In particular, we acknowledge Jiangyu Wei, Shijie Cao, and Ting Cao for introducing the LUT method for low-bit LLM inference on CPUs in their T-MAC work. Xiaoyan Hu provided extensive system and device support, as well as insightful discussions on operating system integration.

²We used [bitnet_b1_58-large](#) available on HuggingFace to demonstrate the inference capabilities of bitnet.cpp. This model and the 8B BitNet b1.58[MSvWW24] in the quick start were **neither trained nor released by Microsoft**.

References

- [Bry19] Marc Brysbaert. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of memory and language*, 109:104047, 2019.
- [lla] llama.cpp. <https://github.com/ggerganov/llama.cpp>.
- [MSvWW24] Mohamed Mekkouri, Marc Sun, Leandro von Werra, and Thomas Wolf. 1.58-bit llm: A new era of extreme quantization, 2024.
- [MWM⁺24] Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-bit llms: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 2024.
- [PPK⁺22] Gunho Park, Baeseong Park, Minsub Kim, Sungjae Lee, Jeonghoon Kim, Beomseok Kwon, Se Jung Kwon, Byeongwook Kim, Youngjoo Lee, and Dongsoo Lee. Lutgemm: Quantized matrix multiplication based on luts for efficient inference in large-scale generative language models. *arXiv preprint arXiv:2206.09557*, 2022.
- [WCC⁺24] Jianyu Wei, Shijie Cao, Ting Cao, Lingxiao Ma, Lei Wang, Yanyong Zhang, and Mao Yang. T-mac: Cpu renaissance via table lookup for low-bit llm deployment on edge. *arXiv preprint arXiv:2407.00088*, 2024.
- [WMD⁺23] Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. Bitnet: Scaling 1-bit transformers for large language models. *arXiv preprint arXiv:2310.11453*, 2023.
- [ZRH⁺24] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024.

A Model Config

The tested models are dummy setups used in a research context to demonstrate the inference performance of bitnet.cpp. The specific configuration is as follows:

```
1 {
2   "125M": {
3     "hidden_size": 768,
4     "intermediate_size": 3072,
5     "num_hidden_layers": 11,
6     "num_attention_heads": 12
7   },
8   "350M": {
9     "hidden_size": 1024,
10    "intermediate_size": 3072,
11    "num_hidden_layers": 24,
12    "num_attention_heads": 16
13  },
14  "700M": {
15    "hidden_size": 1536,
16    "intermediate_size": 4096,
17    "num_hidden_layers": 24,
18    "num_attention_heads": 16
19  },
20  "1B": {
21    "hidden_size": 2048,
22    "intermediate_size": 3584,
23    "num_hidden_layers": 24,
24    "num_attention_heads": 32
25  },
26  "1.5B": {
27    "hidden_size": 1536,
28    "intermediate_size": 9216,
29    "num_hidden_layers": 28,
```

```

30     "num_attention_heads": 32
31 },
32 "2.5B": {
33     "hidden_size": 2560,
34     "intermediate_size": 6912,
35     "num_hidden_layers": 30,
36     "num_attention_heads": 20
37 },
38 "3.8B": {
39     "hidden_size": 3840,
40     "intermediate_size": 8192,
41     "num_hidden_layers": 24,
42     "num_attention_heads": 32
43 },
44 "7B": {
45     "hidden_size": 4096,
46     "intermediate_size": 12032,
47     "num_hidden_layers": 32,
48     "num_attention_heads": 32
49 },
50 "13B": {
51     "hidden_size": 5120,
52     "intermediate_size": 13824,
53     "num_hidden_layers": 40,
54     "num_attention_heads": 40
55 },
56 "30B": {
57     "hidden_size": 6656,
58     "intermediate_size": 16384,
59     "num_hidden_layers": 60,
60     "num_attention_heads": 52
61 },
62 "70B": {
63     "hidden_size": 8192,
64     "intermediate_size": 24576,
65     "num_hidden_layers": 80,
66     "num_attention_heads": 64
67 },
68 "100B": {
69     "hidden_size": 8192,
70     "intermediate_size": 45568,
71     "num_hidden_layers": 72,
72     "num_attention_heads": 64
73 }
74 }

```

We hope the release of bitnet.cpp will inspire the development of 1-bit LLMs in large-scale settings in terms of model size and training tokens.

B Statement of Contribution

All co-authors contributed to discussions, provided input on various aspects of the project, and assisted with experimental design, paper writing, and resource coordination. In addition to these contributions, J. Wang and T. Song crafted multiple kernels and laid out the overarching architecture for the BitNet inference framework. J. Wang took on the full task of kernel implementation. H. Zhou invested considerable time running experiments focused on inference speed, accuracy, and energy consumption.