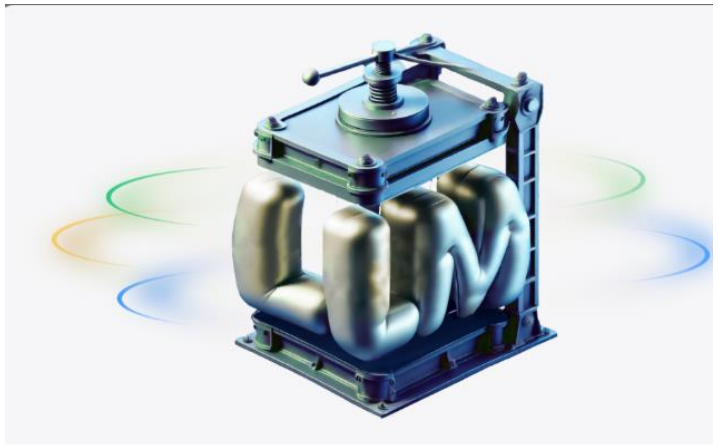


Btech Project

Performance of LLMS locally on edge

Dr. Nitin Auluck
Indian Institute of Technology Ropar

Group No.36

Team Members:
Jai Anurag Y
Nishant Yadav

Overview

- **Project Goal**: Benchmarking compressed LLMs on edge devices
- **Key Technologies**:
  - NVIDIA Jetson platform
  - CUDA
  - Quantization techniques
  - GGUF model format
- **Tools & Frameworks**:
  - PyTorch on Jetson
  - Hugging Face Transformers
  - Energy monitoring

Challenges and Reasons

Challenges :

- Limited computational resources
- Power constraints
- Memory limitations
- Real-time processing requirements
- Thermal considerations

Why Edge Computing for LLMs Matters

•Privacy Preservation : Local data processing eliminates network transmission vulnerabilities

•Latency Reduction : Critical for real-time applications

•Reliability : Independence from network connectivity

•Cost Efficiency : Long-term cost amortization vs. API call expenses

Performance Metrics

1. Prompt Processing Speed
2. Generation Speed
3. Time to first token

Prompt Processing Speed

Prompt processing speed refers to how quickly our system process the input . Its measured in tokens per second.

Several factors influence this speed, including model size, context length, batch size, tokenizer performance, and the underlying hardware (e.g., GPU vs CPU).

Generation Speed

Generation Speed refers to how fast the system generates new text . It is measured in tokens per second.

We need generation speed to ensure responsive, real-time interactions and efficient scalability in LLM-powered applications. Generation speed is essential for optimizing user experience and reducing computational costs during inference.

Time to first token

Time to first token refers to the latency and time before the first response appears . It is measured in milli seconds.

It reflects the model’s initialization, context processing, and early decoding efficiency, making it a key metric for real-time responsiveness. A lower ttft indicated faster model responsiveness which is crucial for interactive applications and chatbots .It is influenced by model architecture , input length and hardware performance .

Custom decided LocalScore

Taking into consideration all of the metrics , we define

Local Score = 10 x (avg_prompt_tps . avg_gen_tps . 1000/avg_ttft_ms)^1/3

As a general guideline :

- A score of 1000+ is excellent
- A score of 250 is acceptable to good for most people
- A score of 100 is relatively poor

Command to run localscore

```
usage: localscore [options]

options:
  -h, --help                                  Show this help message
  -m, --model <filename>                      Model to benchmark (default: path/to/default)
  -c, --cpu                                   Disable GPU acceleration (alias for --gpu=disabled)
  -g, --gpu <auto|amd|apple|nvidia|disabled>  GPU backend to use (default: "auto")
  -i, --gpu-index <i>                         Select GPU by index (default: 0)
  --list-gpus                                 List available GPUs and exit
  -o, --output <csv|json|md>                  Output format (default: md)
  -v, --verbose                               Enable verbose output
  -y, --send-results                          Send results without confirmation
  -n, --no-send-results                       Disable sending results
  -e, --extended                              Run 4 repetitions (shortcut for --reps=4)
  --long                                      Run 16 repetitions (shortcut for --reps=16)
  --reps <N>                                  Set custom number of repetitions
```

Understanding the output

```
================== Active GPU (GPU 0) Information ==================

GPU Name:                  NVIDIA GeForce RTX 3060 Laptop GPU
VRAM:                      6.0 GiB
Streaming Multiprocessors: 30
CUDA Capability:           8.6

====================================================================

Loading model... Model loaded.
Warming up..... Warmup complete.

+-----------------------------------------------------------------------------------------+
|                          NVIDIA GeForce RTX 3060 Laptop GPU - 6.0 GiB                   |
|                              Llama 3.2 1B Instruct - Q4_K - Medium                      |
+-----------------------------------------------------------------------------------------+
|          test | run number | avg time   | tokens processed | pp t/s     | tg t/s     | ttft       |
| ------------- | ---------- | ---------- | ---------------- | ---------- | ---------- | ---------- |
|    pp1024+tg16 | 4/4        | 279.11 ms  | 4160 / 4160      | 6206.32    | 140.36     | 171.86 ms  |
|   pp4096+tg256 | 4/4        | 3.06 s     | 17408 / 17408    | 4365.22    | 120.88     | 948.63 ms  |
|   pp2048+tg256 | 4/4        | 2.13 s     | 9216 / 9216      | 5748.78    | 144.27     | 363.38 ms  |
|   pp2048+tg768 | 4/4        | 5.73 s     | 11264 / 11264    | 5630.06    | 143.14     | 371.09 ms  |
|  pp1024+tg1024 | 4/4        | 6.89 s     | 8192 / 8192      | 6671.50    | 152.11     | 160.09 ms  |
|  pp1280+tg3072 | 4/4        | 23.07 s    | 17408 / 17408    | 6243.96    | 134.40     | 212.98 ms  |
|   pp384+tg1152 | 4/4        | 8.41 s     | 6144 / 6144      | 7557.95    | 137.77     | 58.14 ms   |
|    pp64+tg1024 | 4/4        | 7.55 s     | 4352 / 4352      | 4428.96    | 135.87     | 21.37 ms   |
|    pp16+tg1536 | 4/4        | 11.48 s    | 6208 / 6208      | 1294.04    | 133.96     | 22.94 ms   |
+-----------------------------------------------------------------------------------------+

Token Generation:          138.09 tok/s
Prompt Processing:         5349.64 tok/s
Time to First Token:       258.94 ms
```

What the columns signify

Run Number : Shows number of test runs executed vs planned (e.g., 4/4 = 4 out of 4 completed).
Confirms consistency and reliability of the benchmarking data.

Avg Time : Average total time per run including prompt processing and generation.
Gives an overall sense of response time for the given token configuration.

Tokens processed : Number of prompt and generation tokens processed (Prompt / Total).
Validates data volume handled in each benchmark case.

pp t/s (Prompt Processing Tokens/sec) :
Speed of encoding and embedding the input prompt.
Higher is better; influenced by model efficiency and input length.

tg t/s (Token Generation Tokens/sec) :
Measures how fast the model generates output tokens.
Reflects decoding speed and output fluency performance.

Ttft (Time to first token) : Latency between sending the prompt and receiving the first token.

Prefill Phase

- This phase occurs before any output has been generates
- All prompt tokens (e.g. pp1024, pp4096) are processed at once through the full model.
- The time taken here affects the time to first token directly since generation doesn’t begin until the prompt is fully processed

- you will notice longer prompts have higher time to first token.
- This is due to the quadratic attention cost with prompt length

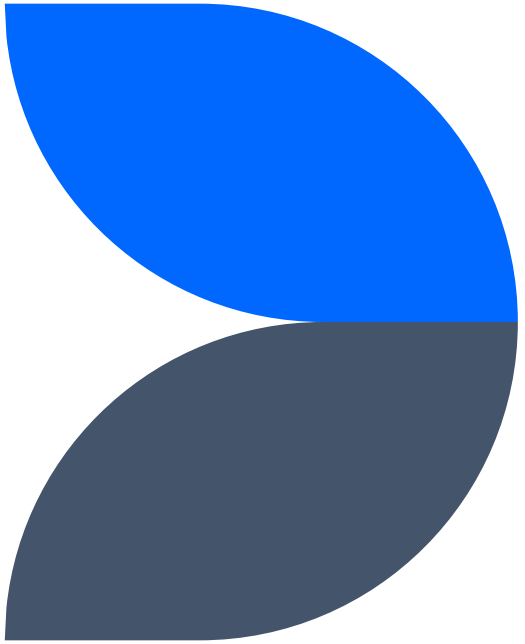
Token Generation Phase

The token generation phase begins after the prompt has been processed and is represented by the tg t/s.

After the prefill, tokens are generated sequentially, using the KV cache populated during the prompt phase.

Here, the model generates one token at a time using cached key-value pairs from the prefill step. This phase is relatively consistent in performance across different test cases, with values ranging around 130–150 tokens per second

TESTING and INFERENCES



google_gemma-3-1b-it-bf16

```
================== Active GPU (GPU 0) Information ==================

GPU Name:                  NVIDIA GeForce RTX 3060 Laptop GPU
VRAM:                      6.0 GiB
Streaming Multiprocessors: 30
CUDA Capability:           8.6

====================================================================

Loading model... Model loaded.
Warming up...... Warmup complete.
```

NVIDIA GeForce RTX 3060 Laptop GPU - 6.0 GiB Gemma 3 1b It - BF16						
test	run number	avg time	tokens processed	pp t/s	tg t/s	ttft
pp1024+tg16	1/1	979.61 ms	1040 / 1040	3911.12	22.29	305.23 ms
pp4096+tg256	1/1	12.51 s	4352 / 4352	3754.20	22.42	1.14 s
pp2048+tg256	1/1	11.73 s	2304 / 2304	3975.31	22.82	563.14 ms
pp2048+tg768	1/1	33.85 s	2816 / 2816	4197.85	23.02	529.80 ms
pp1024+tg1024	1/1	45.07 s	2048 / 2048	4139.28	22.85	292.81 ms
pp1280+tg3072	1/1	142.13 s	4352 / 4352	3663.40	21.67	390.95 ms
pp384+tg1152	1/1	53.99 s	1536 / 1536	3424.61	21.38	157.40 ms
pp64+tg1024	1/1	48.47 s	1088 / 1088	838.69	21.16	122.27 ms
pp16+tg1536	1/1	72.84 s	1552 / 1552	187.87	21.11	136.39 ms

```
Token Generation:        22.08 tok/s
Prompt Processing:       3121.37 tok/s
Time to First Token:     403.99 ms
```

Local Score
555

google_gemma-3-1b-it-Q8_0

```
================== Active GPU (GPU 0) Information ==================

GPU Name:                  NVIDIA GeForce RTX 3060 Laptop GPU
VRAM:                      6.0 GiB
Streaming Multiprocessors: 30
CUDA Capability:           8.6

====================================================================

Loading model... Model loaded.
Warming up..... Warmup complete.

+-----------------------------------------------------------------------------------------------+
|                          NVIDIA GeForce RTX 3060 Laptop GPU - 6.0 GiB                         |
|                                    Gemma 3 1b It - Q8_0                                       |
+-----------------------------------------------------------------------------------------------+
|          test | run number | avg time   | tokens processed | pp t/s     | tg t/s     | ttft       |
| ------------- | ---------- | ---------- | ---------------- | ---------- | ---------- | ---------- |
|   pp1024+tg16 | 1/1        | 351.18 ms  | 1040 / 1040      | 8531.73    | 69.22      | 132.66 ms  |
|  pp4096+tg256 | 1/1        | 4.40 s     | 4352 / 4352      | 7349.01    | 66.56      | 575.04 ms  |
|  pp2048+tg256 | 1/1        | 4.04 s     | 2304 / 2304      | 7745.13    | 67.85      | 276.59 ms  |
|  pp2048+tg768 | 1/1        | 11.85 s    | 2816 / 2816      | 7720.07    | 66.30      | 284.53 ms  |
| pp1024+tg1024 | 1/1        | 15.37 s    | 2048 / 2048      | 7801.35    | 67.19      | 148.57 ms  |
| pp1280+tg3072 | 1/1        | 45.26 s    | 4352 / 4352      | 7185.23    | 68.14      | 196.12 ms  |
|  pp384+tg1152 | 1/1        | 16.85 s    | 1536 / 1536      | 8559.81    | 68.56      | 61.94 ms   |
|   pp64+tg1024 | 1/1        | 15.06 s    | 1088 / 1088      | 2301.82    | 68.14      | 41.67 ms   |
|   pp16+tg1536 | 1/1        | 22.57 s    | 1552 / 1552      | 567.65     | 68.14      | 43.11 ms   |
+-----------------------------------------------------------------------------------------------+

Token Generation:        67.79 tok/s
Prompt Processing:       6417.98 tok/s
Time to First Token:     195.58 ms
```

google_gemma-3-1b-it-Q4_K_M

```
================== Active GPU (GPU 0) Information ==================

GPU Name:                  NVIDIA GeForce RTX 3060 Laptop GPU
VRAM:                      6.0 GiB
Streaming Multiprocessors: 30
CUDA Capability:           8.6

====================================================================

Loading model... Model loaded.
Warming up...... Warmup complete.
```

NVIDIA GeForce RTX 3060 Laptop GPU - 6.0 GiB Gemma 3 1b It - Q4_K - Medium						
test	run number	avg time	tokens processed	pp t/s	tg t/s	ttft
pp1024+tg16	1/1	326.98 ms	1040 / 1040	8299.06	78.59	134.15 ms
pp4096+tg256	1/1	4.03 s	4352 / 4352	7217.31	73.93	579.51 ms
pp2048+tg256	1/1	3.66 s	2304 / 2304	7992.86	75.27	267.73 ms
pp2048+tg768	1/1	10.38 s	2816 / 2816	8002.79	75.88	270.01 ms
pp1024+tg1024	1/1	13.77 s	2048 / 2048	7988.66	75.08	142.07 ms
pp1280+tg3072	1/1	43.17 s	4352 / 4352	7714.57	71.44	182.14 ms
pp384+tg1152	1/1	16.12 s	1536 / 1536	8182.64	71.68	61.73 ms
pp64+tg1024	1/1	14.43 s	1088 / 1088	2223.93	71.12	39.90 ms
pp16+tg1536	1/1	21.57 s	1552 / 1552	669.50	71.28	35.84 ms

```
Token Generation:        73.81 tok/s
Prompt Processing:       6476.81 tok/s
Time to First Token:     190.34 ms
```

Local Score
1359

Llama-3.2-1B-Instruct-Q4_K_M

```
=================== Active GPU (GPU 0) Information ===================

GPU Name:                  NVIDIA GeForce RTX 3060 Laptop GPU
VRAM:                      6.0 GiB
Streaming Multiprocessors: 30
CUDA Capability:           8.6

======================================================================

Loading model... Model loaded.
Warming up..... Warmup complete.
```

NVIDIA GeForce RTX 3060 Laptop GPU - 6.0 GiB Llama 3.2 1B Instruct - Q4_K - Medium						
test	run number	avg time	tokens processed	pp t/s	tg t/s	ttft
pp1024+tg16	1/1	277.74 ms	1040 / 1040	6568.59	131.32	165.96 ms
pp4096+tg256	1/1	3.15 s	4352 / 4352	4220.22	117.36	979.26 ms
pp2048+tg256	1/1	2.30 s	2304 / 2304	5505.63	132.61	378.84 ms
pp2048+tg768	1/1	6.26 s	2816 / 2816	5511.64	130.41	380.40 ms
pp1024+tg1024	1/1	8.06 s	2048 / 2048	6297.76	129.63	169.39 ms
pp1280+tg3072	1/1	25.95 s	4352 / 4352	5912.30	119.39	225.02 ms
pp384+tg1152	1/1	9.18 s	1536 / 1536	6967.29	126.27	62.13 ms
pp64+tg1024	1/1	8.23 s	1088 / 1088	3767.65	124.72	27.16 ms
pp16+tg1536	1/1	12.54 s	1552 / 1552	1109.64	122.62	20.63 ms

```
Token Generation:        126.04 tok/s
Prompt Processing:       5095.64 tok/s
Time to First Token:     267.64 ms
```

Local Score
1339

tiny-vicuna-1b.q2_k

```
=================== Active GPU (GPU 0) Information ==================

GPU Name:                  NVIDIA GeForce RTX 3060 Laptop GPU
VRAM:                      6.0 GiB
Streaming Multiprocessors: 30
CUDA Capability:           8.6

======================================================================

Loading model... Model loaded.
Warming up...... Warmup complete.

+-----------------------------------------------------------------------------------------+
|                          NVIDIA GeForce RTX 3060 Laptop GPU - 6.0 GiB                   |
|                                    active - Q2_K - Medium                               |
+-----------------------------------------------------------------------------------------+
|          test | run number | avg time   | tokens processed | pp t/s     | tg t/s     | ttft       |
| ------------- | ---------- | ---------- | ---------------- | ---------- | ---------- | ---------- |
|   pp1024+tg16 | 1/1        | 349.73 ms  | 1040 / 1040      | 5118.26    | 106.92     | 207.69 ms  |
|  pp4096+tg256 | 1/1        | 3.96 s     | 4352 / 4352      | 3252.84    | 94.74      | 1.27 s     |
|  pp2048+tg256 | 1/1        | 2.91 s     | 2304 / 2304      | 4267.17    | 105.19     | 488.76 ms  |
|  pp2048+tg768 | 1/1        | 8.00 s     | 2816 / 2816      | 4181.61    | 102.25     | 499.35 ms  |
| pp1024+tg1024 | 1/1        | 10.10 s    | 2048 / 2048      | 4759.48    | 103.56     | 223.47 ms  |
| pp1280+tg3072 | 1/1        | 32.65 s    | 4352 / 4352      | 4413.00    | 94.93      | 299.34 ms  |
|  pp384+tg1152 | 1/1        | 11.55 s    | 1536 / 1536      | 5467.18    | 100.31     | 82.66 ms   |
|   pp64+tg1024 | 1/1        | 10.15 s    | 1088 / 1088      | 3425.68    | 101.03     | 26.10 ms   |
|   pp16+tg1536 | 1/1        | 15.19 s    | 1552 / 1552      | 844.65     | 101.22     | 26.65 ms   |
+-----------------------------------------------------------------------------------------+

Token Generation:        101.13 tok/s
Prompt Processing:       3969.99 tok/s
Time to First Token:     347.34 ms
```

Local Score
1049

Quantization Inferences

Here are the key inferences from the benchmarks of the three differently quantized **Gemma 3 1B It** models:

1.Quantization Improves Speed: Both Q8_0 and Q4_K quantized models show significantly faster token generation and prompt processing speeds compared to the BF16 (full-precision) model. For example, token generation improves from 22.08 tok/s (BF16) to 73.81 tok/s (Q4_K), over a 3× speedup.

2.Lower TTFT with Quantization: Time to First Token drops sharply from 403.99 ms (BF16) to ~190 ms for both Q8_0 and Q4_K. This indicates faster prompt processing and reduced model initialization time due to smaller model size and reduced compute.

3.Q4_K Is the Fastest: Among the quantized models, Q4_K achieves the highest token generation rate (73.81 tok/s), slightly outperforming Q8_0 (67.79 tok/s), despite having lower precision. This suggests that Q4_K strikes a better balance between speed and acceptable accuracy.

Quantization Inferences

4.Prompt Processing Speeds Are Similar for Q8_0 and Q4_K: Both models achieve prompt processing rates above 6400 tok/s, suggesting that prompt throughput saturates around that point regardless of quantization level once below BF16.

5.BF16 Is Unsuitable for Latency-Sensitive Tasks: With over 400 ms TTFT and much slower generation speed, BF16 is impractical for real-time applications on this hardware setup.
Overall, quantization—especially to Q4_K—offers significant performance gains with minimal trade-offs in prompt and generation latency.

Overall Inferences

1. Quantization Greatly Improves Performance
All quantized models (Q2_K, Q4_K, Q8_0) significantly outperform the BF16 version of Gemma in both token generation and prompt processing speed. For instance, token generation jumps from **22.08 tok/s (BF16)** to over **70 tok/s** in quantized versions, and prompt throughput doubles or more.

2. Q4_K Offers Best Speed-to-Quality Tradeoff
Q4_K quantization consistently delivers the **highest performance** among all configurations tested:
- **Gemma Q4_K:** 73.81 tok/s
- **Llama Q4_K:** 126.04 tok/s

It balances aggressive compression with good runtime speed and relatively low degradation in model quality.

Overall Inferences

3. Llama Q4_K Outperforms All Others
Among all models, **Llama 3.2 1B Instruct - Q4_K** achieves the **highest token generation speed** (126.04 tok/s) and strong prompt processing (5095.64 tok/s), making it ideal for high-throughput, low-latency applications. Its TTFT of 267.64 ms is also moderate.

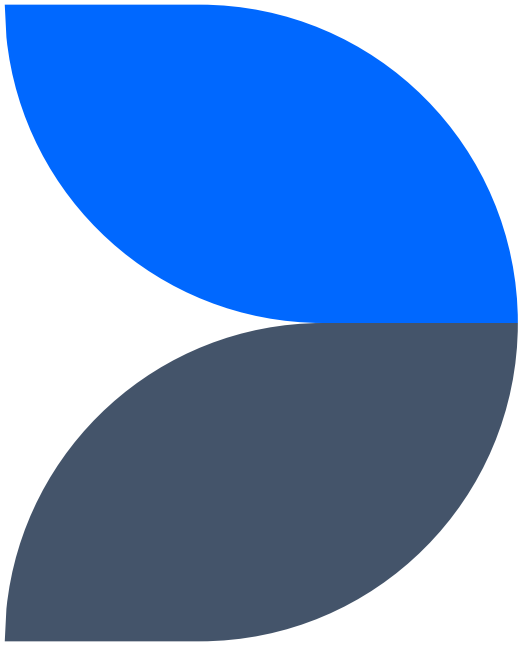
4. Lower Precision Reduces TTFT and Improves Latency
Quantized models consistently show lower **Time to First Token** compared to BF16. TTFT drops from **403.99 ms (Gemma BF16)** to as low as **190.34 ms (Gemma Q4_K)**, showing quantization reduces early latency.

5. Model and Format Matter
The Llama Q4_K model outperforms Gemma Q4_K by a wide margin in token generation speed (126 vs 73 tok/s). This highlights that **architecture and tokenizer design** also play a key role in inference performance, not just quantization level.

Other Findings

We experimented with a wide range of models and hardware, including both integrated and discrete GPUs such as RTX 3060, RTX 3050, and Intel Iris Xe across multiple devices. Due to time constraints, we are only presenting a subset of the results here.

Why NO JetsonNano



JetPack Archive

This page includes access to previously released versions of **JetPack**. The latest version of JetPack is always available under the main NVIDIA JetPack product page.

> JetPack 6.2
  > Jetson AGX Orin Series, Jetson Orin NX Series, Jetson Orin Nano Series [L4T 36.4.3]
> JetPack 6.1
  > Jetson AGX Orin Series, Jetson Orin NX Series, Jetson Orin Nano Series [L4T 36.4]
> JetPack 6.0
  > Jetson AGX Orin Series, Jetson Orin NX Series, Jetson Orin Nano Series [L4T 36.3]
> JetPack 6.0 DP
  > Jetson AGX Orin Series, Jetson Orin NX Series, Jetson Orin Nano Series [L4T 36.2]
> JetPack 5.1.5
  > Jetson AGX Orin Series, Jetson Orin NX Series, Jetson Orin Nano Series, Jetson Xavier NX series, Jetson AGX Xavier Series, [L4T 35.6.1]
> JetPack 5.1.4
  > Jetson AGX Orin Series, Jetson Orin NX Series, Jetson Orin Nano Series, Jetson Xavier NX series, Jetson AGX Xavier Series, [L4T 35.6.0]
> JetPack 5.1.3
  > Jetson AGX Orin Series, Jetson Orin NX Series, Jetson Orin Nano Series, Jetson Xavier NX series, Jetson AGX Xavier Series, [L4T 35.5.0]
> JetPack 5.1.2
  > Jetson AGX Orin Series, Jetson Orin NX Series, Jetson Orin Nano Series, Jetson Xavier NX series, Jetson AGX Xavier Series, [L4T 35.4.1]
> JetPack 5.1.1
  > Jetson AGX Orin Series, Jetson Orin NX Series, Jetson Orin Nano Series, Jetson Xavier NX series, Jetson AGX Xavier Series, [L4T 35.3.1]
> JetPack 5.1
  > Jetson AGX Orin Developer Kit, Jetson AGX Orin 32GB, Jetson Orin NX 16GB, Jetson Xavier NX series, Jetson AGX Xavier Series, [L4T 35.2.1]
> JetPack 5.0.2
  > Jetson AGX Orin Developer Kit, Jetson Xavier NX series, Jetson AGX Xavier Series, [L4T 35.1]
> JetPack 5.0.1 Developer Preview
  > Jetson AGX Orin Developer Kit, Jetson Xavier NX series, Jetson AGX Xavier Series, [L4T 34.1.1]
> JetPack 5.0 Developer Preview
  > Jetson AGX Orin Developer Kit, Jetson Xavier NX series, Jetson AGX Xavier Series, [L4T 34.1]
> JetPack 4.6.6
  > Jetson Xavier NX Series, Jetson TX2 Series, Jetson AGX Xavier Series, Jetson Nano, Jetson TX1, [L4T 32.7.6]

PyTorch for Jetson

Categories > Robotics & Edge Computing > Jetson & Embedded Systems > Announcements | python | pytorch | kb

Log In

JetPack 5

- ▶ PyTorch v2.1.0
- ▶ PyTorch v2.0.0
- ▶ PyTorch v1.14.0
- ▶ PyTorch v1.13.0
- ▶ PyTorch v1.12.0
- ▶ PyTorch v1.11.0

JetPack 4

- ▼ PyTorch v1.10.0
  - JetPack 4.4 (L4T R32.4.3) / JetPack 4.4.1 (L4T R32.4.4) / JetPack 4.5 (L4T R32.5.0) / JetPack 4.5.1 (L4T R32.5.1) / JetPack 4.6 (L4T R32.6.1)
    - Python 3.6 - `torch-1.10.0-cp36-cp36m-linux_aarch64.whl` 25.7k
    - This is the final PyTorch release supporting Python 3.6.
- ▶ PyTorch v1.9.0
- ▶ PyTorch v1.8.0
- ▶ PyTorch v1.7.0
- ▶ PyTorch v1.6.0
- ▶ PyTorch v1.5.0
- ▶ PyTorch v1.4.0
- ▶ PyTorch v1.3.0
- ▶ PyTorch v1.2.0
- ▶ PyTorch v1.1.0
- ▶ PyTorch v1.0.0

Instructions

Mar 2019

1 / 1367

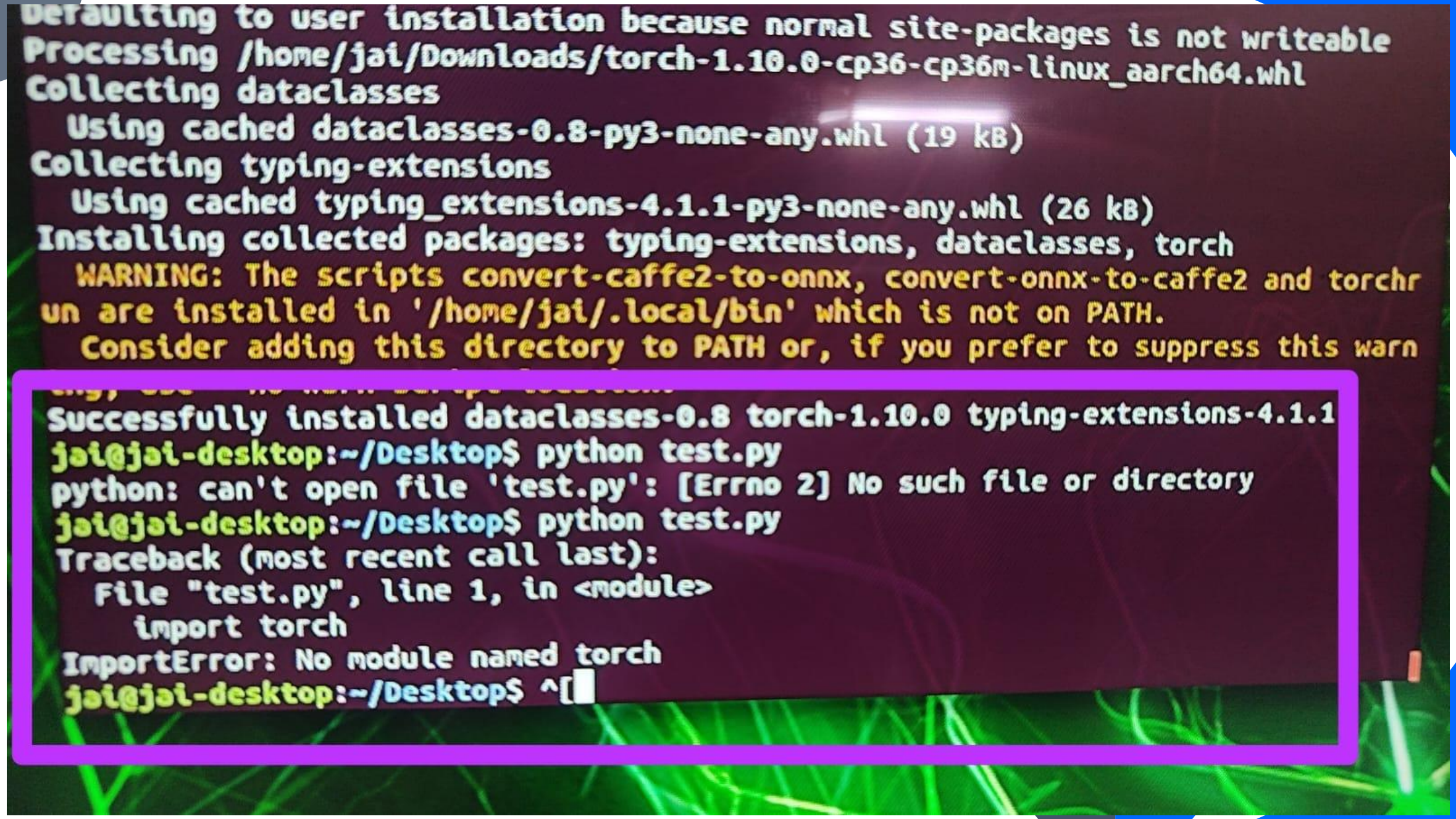
Mar 2019

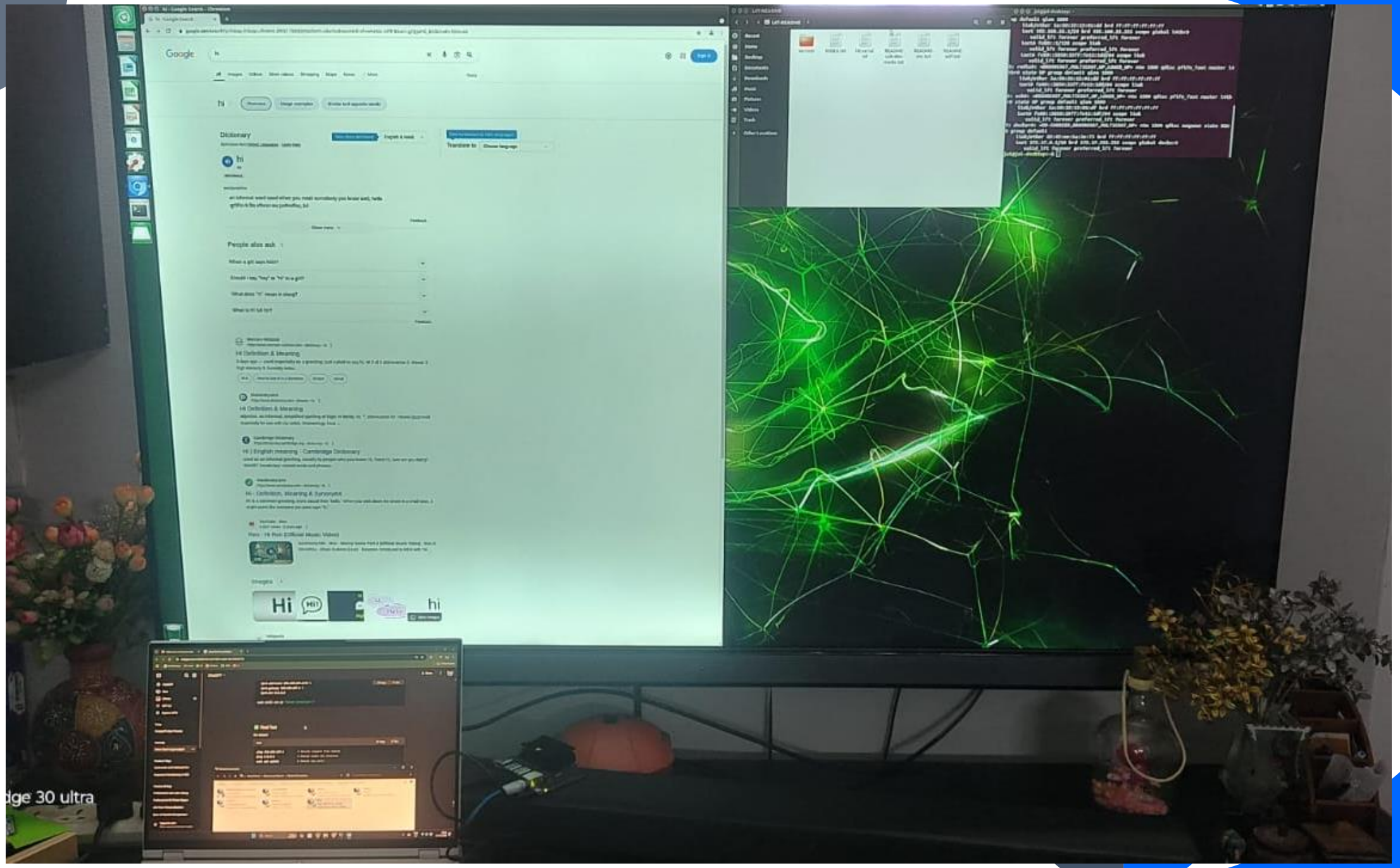
May 15

Search

ENG IN

22:15 14-05-2025


Defaulting to user installation because normal site-packages is not writeable
Processing /home/jai/Downloads/torch-1.10.0-cp36-cp36m-linux_aarch64.whl
Collecting dataclasses
Using cached dataclasses-0.8-py3-none-any.whl (19 kB)
Collecting typing-extensions
Using cached typing_extensions-4.1.1-py3-none-any.whl (26 kB)
Installing collected packages: typing-extensions, dataclasses, torch
WARNING: The scripts convert-caffe2-to-onnx, convert-onnx-to-caffe2 and torchrun are installed in '/home/jai/.local/bin' which is not on PATH.
Consider adding this directory to PATH or, if you prefer to suppress this warn
Successfully installed dataclasses-0.8 torch-1.10.0 typing-extensions-4.1.1
jai@jai-desktop:~/Desktop$ python test.py
python: can't open file 'test.py': [Errno 2] No such file or directory
jai@jai-desktop:~/Desktop$ python test.py
Traceback (most recent call last):
File "test.py", line 1, in <module>
import torch
ImportError: No module named torch
jai@jai-desktop:~/Desktop$ ^[


Google
hi
Dictionary
People also ask
Hi
hi
dge 30 ultra

Thank you