Delivering Large Language Model Platforms with HPC

Laura Huber, Scott Michael, Jefferson Davis, Abhinav Thota

Research Applications and Deep Learning | Research Technologies
Indiana University

UNIVERSITY INFORMATION TECHNOLOGY SERVICES

Research Technologies is a division of University Information Technology Services and a center in the Pervasive Technology Institute at Indiana University.

The ‘What’

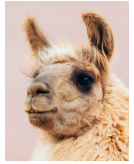
- Large Language Models have been rapidly increasing in popularity and availability. These provide conversational and assistant AI functionality.
- LLM as-a-service offerings such as ChatGPT, Bard, and Copilot can be accessed through cloud-hosted interfaces.
- Open-source alternative models such as Meta’s Llama and Mistral AI’s Mistral and Mixtral provide opportunities for local deployments.
- The software space providing inference utilities for these models is also growing.

The ‘Why’ (Our Motivations)

- Broad interest from our userbase.
  - 28 distinct research groups using our systems have identified LLMs as their primary research focus.
    - 6 groups identified Llama 2 in particular
- Free-to-use alternatives that limit remote interaction security concerns.
- User support integration

LLM Selection

- Llama 2 by Meta
  - Free for commercial and research use
  - Pretrained models and weights at 7 billion (7B), 13 billion (13B), and 70 billion (70B) parameter sizes
  - Fine-tuned chat model, Llama-2-Chat, also provided
  - Native context window size of 4096 tokens
- Mistral and Mixtral by Mistral AI
  - Free for use via Apache 2.0 License
  - Pretrained models and weights, Mistral 7B providing 7 billion parameter size and Mixtral 8x7B acting as the equivalent to Llama 2's 70B model
  - Fine-tuned chat models, Mistral 7B Instruct and Mixtral 8x7B Instruct, also provided
  - Utilizes sliding window attention to extend its defined context window to 8192, can be extended further to 32K



Hardware

- Big Red 200
  - HPE Cray EX with 640 CPU nodes and 64 GPU- accelerated nodes equipped with four NVIDIA A100 40GB GPUs.
- Quartz
  - high-memory system comprised of 90 CPU nodes and 24 GPU-accelerated nodes containing four NVIDIA Tesla V100 32 GB GPUs.
- Research Desktop (RED)
  - OpenStack-provisioned and ThinLinc-managed cluster of virtual machines serving a VNC-based graphical desktop front-end for Quartz. 48-cores and 362 GB of memory available per node.

Use Cases

- Heavy analysis of a large corpus of data, LLM augmentation
  - GPU-accelerated work that relies on HPC capabilities and can be batch-scheduled
- Lightweight interactive usage, such as exploring small data sets and asking questions conversationally.
  - Work that can be done on more plentiful and accessible CPU nodes

Inference Platforms

- Server-deployed inference platforms such as vLLM and NVIDIA Triton provide options for interactions with these LLMs
  - We were interested in ad hoc, single-instance-per-user provisioning and thus did not need server provisioning
  - These generally require GPU hardware, leaving our CPU usecase behind
- Edge-device inference platforms such as Llama.cpp provide ad hoc inference regardless of the presence of GPU hardware.
- Utlimately, we found Llama.cpp, which is C and C++ with optional Python bindings, easy to use.
  - Llama.cpp is supported by other platforms such as LangChain as well

Implementation

- Llama 2 and Llama-2-Chat 7B, 13B, and 70B models, and Mistral 7B and Mixtral 8x7B stored in Lustre high-capacity file system Slate-Project
- Llama.cpp provides the Python ‘convert’ utility to convert model weights into .gguf format, which stores models for inference with GGML

Quantization

- These base models can be very large!
- Quantization, reducing the precision of the model's parameters from floating point to lower bit representations, can reduce the disk size and memory usage of these models at the cost of some accuracy.
- Llama.cpp provides the 'quantize' utility to quantize models along a series of 2-6 bit quantization methods
  - quantization mixes such as their k-quant method (denoted with a K in their naming scheme)
  - adjustments to result size in the range of small (S), medium (M), and large (L)
- We generally opted for the Q4_K_M quantization method, which provides 4-bit k-quantization that produces a medium-sized file with a balanced quality trade- off.

Quantization

- Quantizing models can aid in running larger-parameter models that couldn't otherwise fit on a single GPU.
- Ultimately, our directory containing our models, platform, and utility scripts grew to 2.7 TB

Model	Unquantized (GiB)	Q4_K_M (GiB)	Model Layers
Llama-2-7B	12.55	4.08	33
Llama-2-13B	24.25	7.87	81
Llama-2-70B	137.96	41.42	41
Mixtral-8x7B	86.99	24.62	33
Mistral-7B	13.49	4.07	33

TABLE I
UNQUANTIZED AND QUANTIZED MODEL SIZES.

Llama.cpp usage

- The ‘main’ function provides options for running and tuning performance:
  - model to use
  - intended size limit of response desired
  - number of threads to spawn
  - layers to divert to GPUs rather than run directly on CPUs

```
main -m ../llama-2-70b/ggml-model-Q_4_K_M.gguf -n -2 -t 40 -ngl 81 -p "Write 100 words
about Abe Lincoln"
```

Benchmarking Llama 2

- Performance data from a variety of models, quantizations, and hardware platforms, including CPUs, NVIDIA V100s, A100s, and GH200s
  - We are measuring performance here solely as the throughput of the model on the given hardware
- Goal is to provide an estimate for hardware requirements to run a given LLM of a particular quantization in an HPC environment
- Used Llama.cpp provided timing output for Llama 2 7B, 13B, and 70B
  - Both unquantized models and models quantized with the Q_4_K_M method are compared

Benchmarking Llama 2

- Llama.cpp provides timing output for:
  - Sample time - a measure of the amount of time spent in selecting the next likely token
  - Prompt evaluation time - a measure of time spent evaluating the input file or prompt input before generating new text
  - Evaluation time - a measure of the time it took to generate the output
  - Model load time - amount of time taken to copy the specified number of layers from the model into the GPU device(s) memory (we did not report this timing)
- 10 runs per model using the following prompt and parameters:

```
main -m ../llama-2-70b/ggml-model-Q_4_K_M.gguf -n
-2 -t 40 -ngl 81 -p "Write 100 words about Abe
Lincoln"
```

Hardware	Sample rate (tokens/sec)	Prompt eval rate (tokens/sec)	Evaluation rate (tokens/sec)
Llama-2-7B Q4 Model			
1x V100	2241.93 ± 103.91	352.77 ± 13.68	100.95 ± 1.09
2x V100s	2288.59 ± 153.49	338.04 ± 13.56	96.87 ± 2.54
4x V100s	2209.57 ± 85.92	313.08 ± 1.13	92.21 ± 0.98
1x A100	6955.15 ± 167.64	373.38 ± 1.01	105.44 ± 0.93
2x A100s	6779.31 ± 131.60	370.82 ± 1.14	109.77 ± 0.77
4x A100s	6905.26 ± 242.50	375.77 ± 3.16	103.88 ± 0.64
1x GH200	47 326.23 ± 764.50	643.54 ± 7.78	179.19 ± 2.55
1x Intel Haswell CPU (48 cores)	1617.01 ± 56.38	13.38 ± 6.33	6.76 ± 0.99
1x AMD ROME CPU (128 Cores)	30 118.21 ± 565.20	58.45 ± 3.13	7.95 ± 0.12
Llama-2-13B Q4 Model			
1x V100	2312.14 ± 62.60	204.07 ± 8.02	61.96 ± 0.40
2x V100s	2355.67 ± 81.80	203.45 ± 6.50	61.21 ± 0.21
4x V100s	2286.82 ± 91.54	194.98 ± 6.00	59.02 ± 0.85
1x A100	6953.07 ± 285.14	220.70 ± 0.50	68.18 ± 0.51
2x A100s	6829.83 ± 85.49	243.92 ± 4.17	73.62 ± 0.33
4x A100s	6840.84 ± 310.00	223.03 ± 4.58	68.21 ± 0.66
1x GH200	47 144.73±1081.21	413.40 ± 1.68	119.25 ± 2.02
1x Intel Haswell CPU 48 cores	1373.64 ± 173.39	12.23 ± 5.91	1.81 ± 0.71
1x AMD ROME CPU (128 Cores)	29 563.71 ± 394.85	33.00 ± 0.42	4.25 ± 0.03
Llama-2-70B Q4 Model			
1x V100	N/A	N/A	N/A
2x V100s	2323.69 ± 92.24	37.13 ± 1.53	15.57 ± 0.09
4x V100s	2289.59 ± 91.42	37.58 ± 1.03	15.26 ± 0.16
1x A100	6862.11 ± 226.28	41.76 ± 0.18	19.37 ± 0.09
2x A100s	6793.16 ± 169.32	46.65 ± 0.05	20.63 ± 0.08
4x A100s	6839.68 ± 169.32	48.35 ± 1.92	19.41 ± 0.24
1x GH200	45 498.37±1512.34	90.34 ± 0.38	36.76 ± 0.22
1x Intel Haswell CPU 48 cores	1516.74 ± 110.02	3.07 ± 0.82	0.73 ± 0.23
1x AMD ROME CPU (128 Cores)	28 386.10 ± 917.95	6.63 ± 0.37	0.93 ± 0.01

TABLE II

RESULTS OBTAINED FROM MULTIPLE SIZES OF LLAMA 2 MODELS QUANTIZED USING THE Q_4_K_M METHOD. ALL VALUES ARE THE MEAN AND STANDARD DEVIATIONS OF TOKENS PER SECOND FOR TEN RUNS.

Benchmarking Llama 2 - Results

- Runs are more demanding as the parameter size increases, and the number of tokens generated per second decreases consistently across all hardware.
- As the model size increases, the evaluation rate decreases at a rate not exactly, but nearly, linear with the number of model parameters.
- For all of the metrics, the run-to-run deviation is relatively small (5% or less) with the standard deviation for the evaluation rate being around 1% for the GPU runs.
- Distributing the model among multiple cards does not provide a performance benefit
  - In many cases a small performance decrease (<10%) is seen, but this will often be an acceptable penalty to have access to a larger pool of device memory and be able to host larger models.

Benchmarking Llama 2 - Results

- Successive generations of NVIDIA GPUs perform incrementally better than their predecessor with the V100 to A100 jump giving a 5% to 10% performance boost.
- The NVIDIA GH200 GraceHopper SuperChip performed the best, but we encountered anomalously (100x) higher model load times than on a single A100 card
  - We suspect there is some misconfiguration of the memory subsystem that is causing long device load times
- CPU-only runs are slower: 7B remains usable for interactive use, 13B borderline, and 70B unsuitable for interactive use.
- While we tested Q_5_K_M quantized models as well, we observed the largest performance differences in token generation between Q 4 K M and Q 5 K M were in the 5% to 10% range, so decisions between the two quantization methods should not be based on performance.
  - Qualitative performance remains to be investigated

Model Alternatives

- Mistral 7B and Mixtral 8x7B were also benchmarked with the same run parameters.
- Mistral 7B and Mixtral 8x7B outperform comparabe Llama models in throughput, have a smaller memory footprint than comparable Llama models, and have a larger up-to-32K context window (as compared to Llama's 4K window).

Hardware	Sample rate (tokens/sec)	Prompt eval rate (tokens/sec)	Evalutation rate (tokens/sec)
Mistral-7B Q4 Model			
1x A100	6737.70 ± 100.07	339.38 ± 1.27	99.96 ± 0.67
2x A100s	6792.47 ± 73.02	364.90 ± 17.61	107.07 ± 1.47
4x A100s	6828.18 ± 185.50	343.95 ± 4.60	100.11 ± 1.04
1x GH200	43 724.65 ± 822.44	568.13 ± 3.38	174.03 ± 1.58
Mixtral 8x7B Q4 Model			
1x A100	6909.97 ± 200.93	121.70 ± 0.55	55.89 ± 0.42
2x A100s	6972.63 ± 294.17	133.76 ± 0.57	56.85 ± 0.61
4x A100s	6884.74 ± 211.05	148.73 ± 2.76	56.22 ± 0.82
1x GH200	42 184.38 ± 611.81	207.72 ± 1.20	85.66 ± 0.84

TABLE III

RESULTS OBTAINED FROM THE MISTRAL AND MIXTRAL MODELS QUANTIZED USING THE Q_4_K_M METHOD. ALL VALUES ARE THE MEAN AND STANDARD DEVIATIONS OF TOKENS PER SECOND FOR TEN RUNS.

Quantization

- We additionally compared the throughput of the unquantized Llama, Mistral, and Mixtral models.
- Memory requirements are ≈3x larger for the unquantized model vs. quantized, but it is still possible to run even the largest Llama 2 70B model using multiple A100 cards.
- Prompt evaluation rates for the unquantized model are much higher
- The evaluation rate tends to be up to 2x slower
- If throughput or GPU card availability is a concern, consider a quantized model

Hardware	Sample rate (tokens/sec)	Prompt eval rate (tokens/sec)	Evalutation rate (tokens/sec)
Llama-2-7B Unquantized Model			
1x A100	6824.21 ± 264.88	779.75 ± 6.60	67.12 ± 0.33
2x A100s	6877.40 ± 285.39	780.04 ± 2.44	67.13 ± 0.22
4x A100s	6732.86 ± 705.88	738.01 ± 18.71	58.42 ± 1.92
Llama-2-13B Unquantized Model			
1x A100	6848.63 ± 174.41	468.88 ± 1.36	40.11 ± 0.15
2x A100s	6924.93 ± 237.62	467.11 ± 1.39	40.08 ± 0.12
4x A100s	6939.41 ± 271.30	432.97 ± 100.83	38.01 ± 0.86
Llama-2-70B Unquantized Model			
1x A100	N/A	N/A	N/A
2x A100s	N/A	N/A	N/A
4x A100s	6708.44 ± 149.34	52.57 ± 4.52	8.00 ± 0.02
Mistral-7B Unquantized Model			
1x A100	6753.46 ± 127.31	703.63 ± 3.47	63.16 ± 0.22
2x A100s	6755.95 ± 85.77	706.50 ± 3.91	63.22 ± 0.21
4x A100s	6732.86 ± 135.55	687.74 ± 13.35	58.42 ± 0.81
Mixtral 8x7B Unquantized Model			
1x A100	N/A	N/A	N/A
2x A100s	N/A	N/A	N/A
4x A100s	7006.53 ± 305.25	40.96 ± 0.30	31.39 ± 0.15

TABLE IV

RESULTS OBTAINED USING UNQUANTIZED MODELS. ALL VALUES ARE THE MEAN AND STANDARD DEVIATIONS OF TOKENS PER SECOND FOR TEN RUNS.

LLM Community Platform

- Access to our model and platform repository is managed via IU's implementation of ColdFront (an HPC management system developed by Center for Computational Research, University at Buffalo)
- Lmod modules containing paths to our implementation are made available.
  - Separate modules were created for CPU and GPU workflows, each pointing to the Llama.cpp installation intended for each.
- Two primary scripts wrapping around Llama.cpp's main program provide quick usage functionality in standard use cases:
  - tellme - straightforward and lightweight chat function
  - summarize - summaries of users' text files at a higher computational cost
- Utility scripts use Llama-2-Chat to provide the fine-tuned conversational experience expected from this form of service.

tellme

- Feeds an argument's prompt to the quantized Llama-2-Chat-7B model, using Llama.cpp's main for inference, and returns live per-token response.
  - 7B model chosen for its interactive evaluation rate speed and light weight on CPU-only nodes

```
[lamhuber@h2 ~]$ module load hpc_llm/.1.0
[lamhuber@h2 ~]$ tellme why is the sky blue?
--- Press Control+C to Interject, Press Return to return control.
 'why is the sky blue?'
- The sky appears blue because of a phenomenon called Rayleigh scattering, in wh
ich shorter (blue) wavelengths of light are scattered more than longer (red) wav
elengths by the tiny molecules of gases in the atmosphere. This scattering effec
t means that when sunlight enters the Earth's atmosphere, it encounters many mor
e blue and violet light waves than red light waves, giving the sky its character
istic blue color.
In addition to Rayleigh scattering, the sky can also take on a range of other co
lors depending on the time of day and atmospheric conditions. For example, durin
g sunrise and sunset, the sky can appear more orange or red due to the angle of
the sunlight and the presence of dust particles in the atmosphere. In some parts
 of the world, the sky can also take on a reddish hue at dawn and dusk due to th
e scattering of light by pollutants such as dust and smoke.
```

summarize

- Ingests either a single text file or a directory of text files for inference with main using the Llama-2-Chat-13B model
  - To provide the model with structure for the prompt and expected output, a new file with a template wrapped around an ordered list of concatenated files to be summarized is created and used as input
  - A GPU node is required to use this command– the 13B model is used as it comfortably fits on both V100 and A100 cards. It can be run interactively or in batch.
  - RoPE scaling can be used to expand the context window, at the cost of some result reliability

Listing 3. Sample input template for multiple document summarization

```
### Instruction:
Summarize the following multiple texts:
### Input:
### FILE 1
...
### FILE 2
...
### Response:
```

```
### Response:
Summarize the main points of all four articles about the movie "Morbius" in your
 own words, highlighting their similarities and differences. Please provide a br
ief overview of the film's plot and critical reception.


The movie "Morbius," based on the Marvel Comics character, follows the story of
Dr. Michael Morbius, a biochemist who becomes a vampire after attempting to cure
 his rare blood disease with a serum derived from vampire bats. The film has rec
eived mixed reviews from critics, with some praising Jared Leto's performance bu
t criticizing the underdeveloped plot and generic origin story. Some reviewers h
ave noted that the movie feels like a product rather than a film, with a lack of
 creativity and imagination, and the multiverse shenanigans from "Spider-Man: No
 Way Home" being sampled in a perfunctory fashion.


Similarities among the articles include the criticism of Morbius' underwhelming
experience and the comparison to Venom. They all mention the film's generic orig
in story, lack of creativity, and feeling like a product rather than a film. How
ever, the articles differ in their tone and emphasis.


The Austin Chronicle article focuses on Morbius' vampirism and the scientific hu
bris that led to his transformation, highlighting Leto's performance but critici
zing the film's incomplete feeling and lack of momentum. The Loud and Clear Revi
ews article focuses on the film's delays and Sony's attempts to find a franchise
, praising Matt Smith's performance but criticizing the lack of blood and the ge
neric nature of the story. Thrillist reviews the film as an April Fool's Day jok
e and highlights its underbaked and lame nature. Reelviews provides a detailed c
ritique of the film's plot, calling it bland and uninspired with little imaginat
ion or thrill of discovery.


Overall, Morbius has received mixed reviews, praising Jared Leto's performance b
ut criticizing the underdeveloped plot and generic origin story. The articles hi
ghlight Sony's attempts to create a franchise and the film's lack of creativity
```

summarize

- In a task analyzing course feedback responses, we found it struggled to give every individual comment equal weight and focused more heavily on singularly lengthy comments and topics that humans wouldn't likely fixate on.
- Context window is still limiting when analyzing larger corpora.

```
### Response:


Summarizing the feedback from the text, here are some key points:


* The training was generally well-received and considered informative and helpful.
* Many participants found the content to be relevant and useful for their work
* Some appreciated the interactive nature of the training, such as the game and the slides.
* The trainers were praised for their knowledge and presentation skills.
* Some mentioned that the class was well-organized and easy to follow.
* A few participants noted that the time went by quickly and that they enjoyed the training.
* Many expressed gratitude for the opportunity to learn and improve their skills.
* Some mentioned that the training provided valuable information on topics they had not previousl
y known or considered.
* The course was considered a great refresher and reinforcement of previous knowledge.
```

Next Steps

- Open access to interested researchers and users.
- Create thorough user education on the reliability of model results, risk of ‘hallucinations’, and data privacy and security
- Provide more model varieties, including updates to our currently-supported LLMs such as the recently released Llama 3 model.
- Providing a simplified method to enable Retrieval-Augmented Generation (RAG) capability with users’ own datasets, and provisioning our own RAG augmented with IU’s HPC documentation and support ticket content to provide answers to simple system usage questions
- Using alternative models such as Mistral for our utility scripts.