# Workload Characterization of Transformer Text Generation Inference

Puneeth N Naik

Prof. Govindarajan Ramaswamy, Prof. Murali Annavaram

High Performance Computing Lab, Department of Computer Science and Automation

Indian Institute Of Science, Bangalore

puneethnaik@iisc.ac.in

## Motivation

- LLMs use in applications increasing by the day.
- Important to understand the characteristics of LLM inference.
- Jensen Huang(Nvidia CEO): AI language models as-a-service "potentially one of the largest software opportunities ever"
- ChatGPT witnessed 1.8 Billion visits in April 2023.
- Optimizations in LLM inference help reduce the carbon footprint of data centers.
- BLOOM inference consumed 914KWh electricity, of which GPU accounted for 75.3% in just 18 days.

## Related Work

- There has been considerable work to optimize the inference of LLMs in literature.
- Previous work study the effect of only GPU DVFS on CNN inference.

## Setup

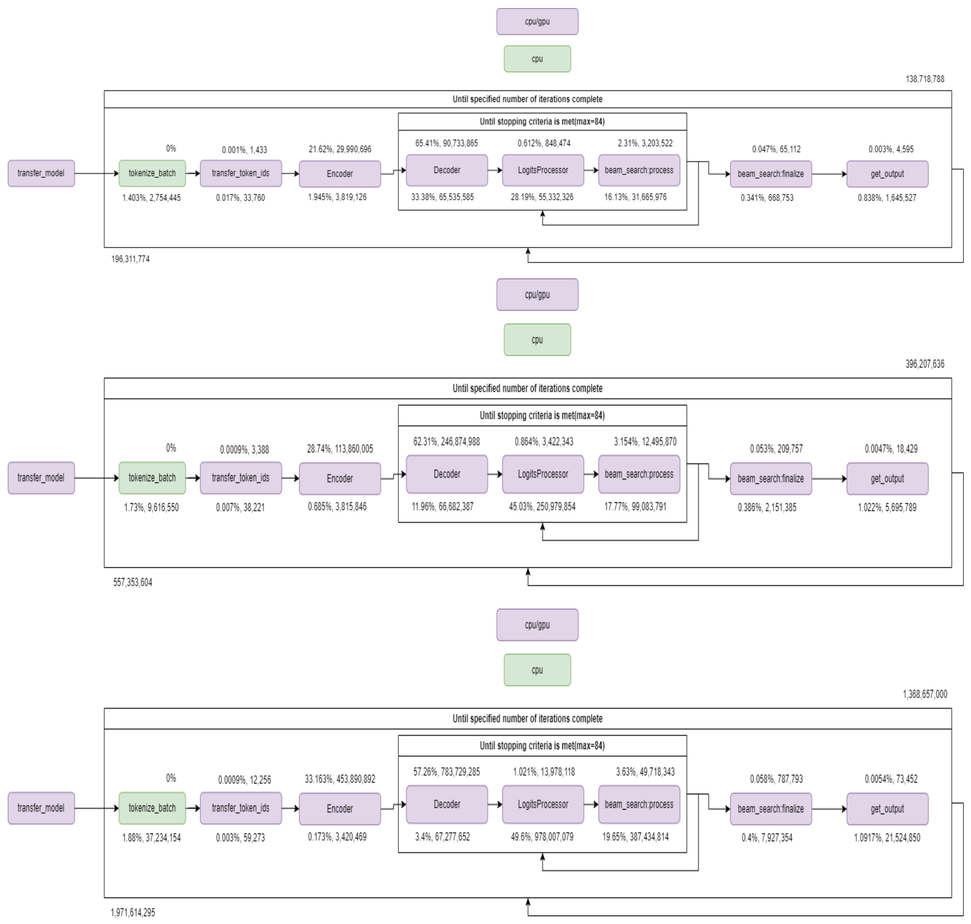
| Feature | indicBART | mBART |
|---|---|---|
| Activation function | Gelu | Relu |
| Vocabulary size | 64K | 250K(3.9x) |
| Number of parameters | 244M | 610M(2.5x) |
| Hidden dimension | 1024 | 1024 |
| Decoder FC dimension | 4096 | 4096 |
| Decoder layers | 6 | 12(2x) |
| Encoder FC dimension | 4096 | 4096 |
| Encoder layers | 6 | 12(2x) |
| Max length | 84 | 300 |

| Name | Component | Spec |
|---|---|---|
| S1 | CPU | AMD Ryzen 5600X |
| S1 | GPU | Nvidia RTX 3060 |

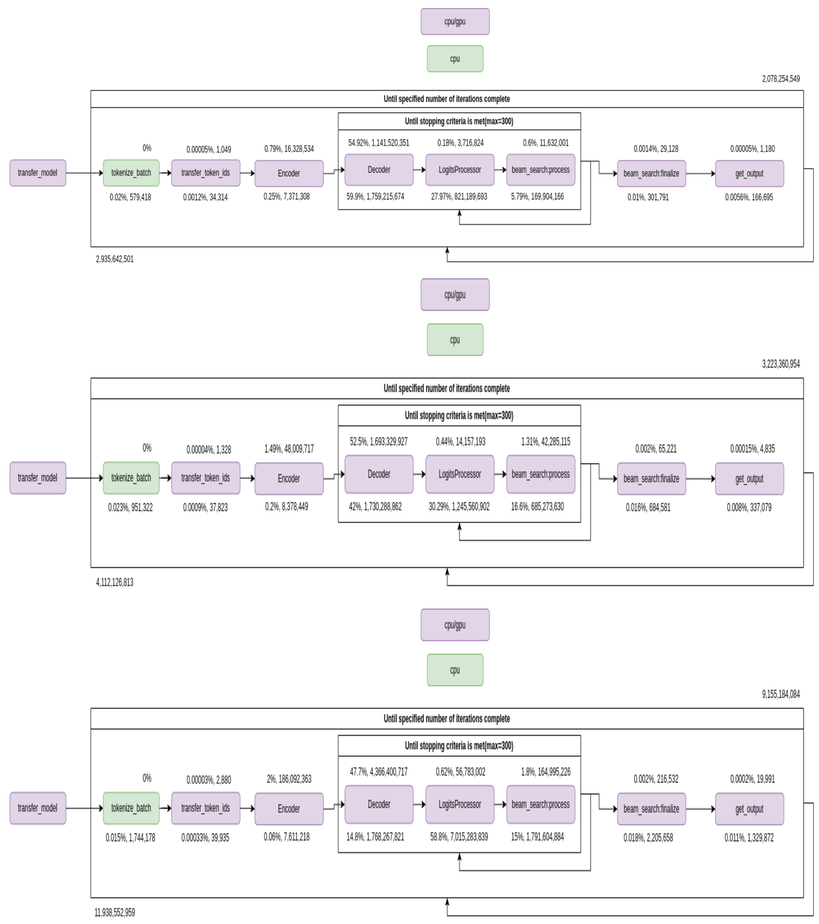
| Software | Version |
|---|---|
| Python | 3.9.15 |
| Transformers | 4.25.1 |
| PyTorch | 1.12.1 |
| CUDA/CUDA Driver | 12.1/570 |

## Results

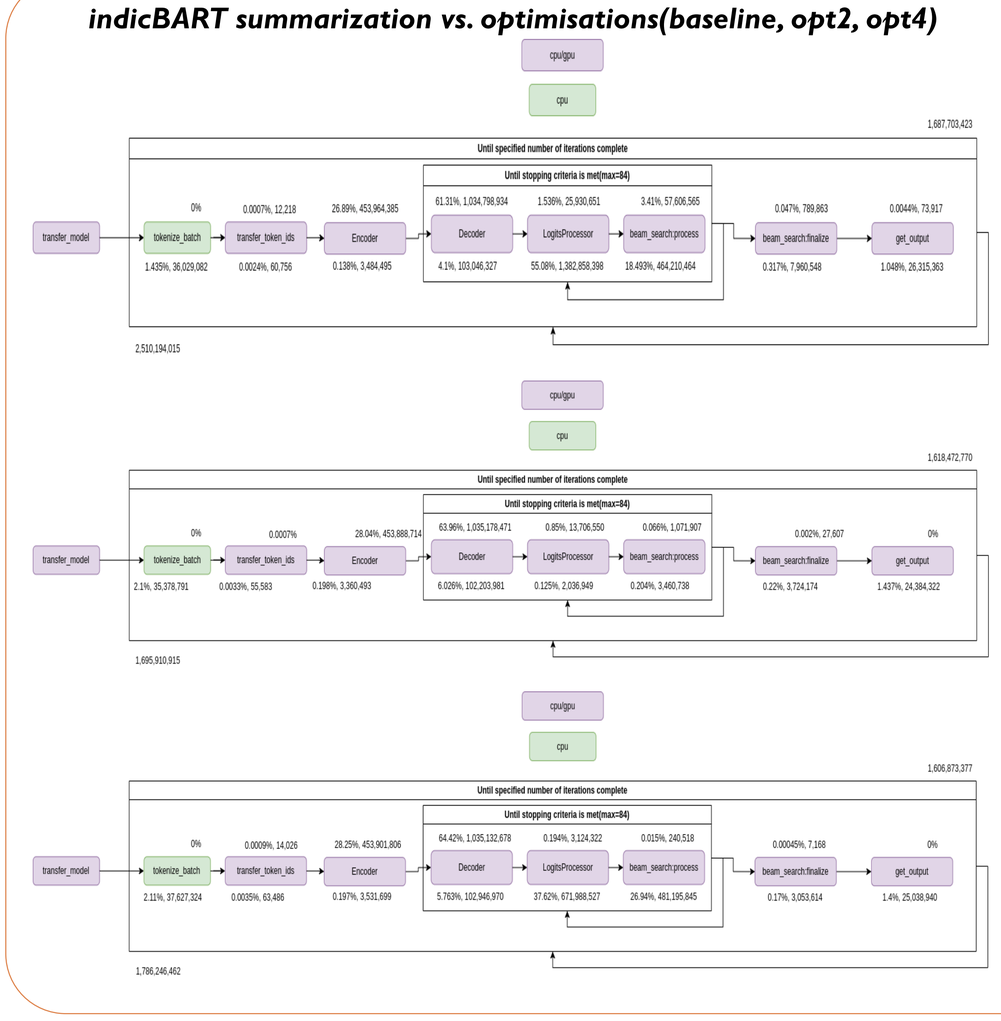
***indicBART summarization vs. batch size(4, 16, 64)***



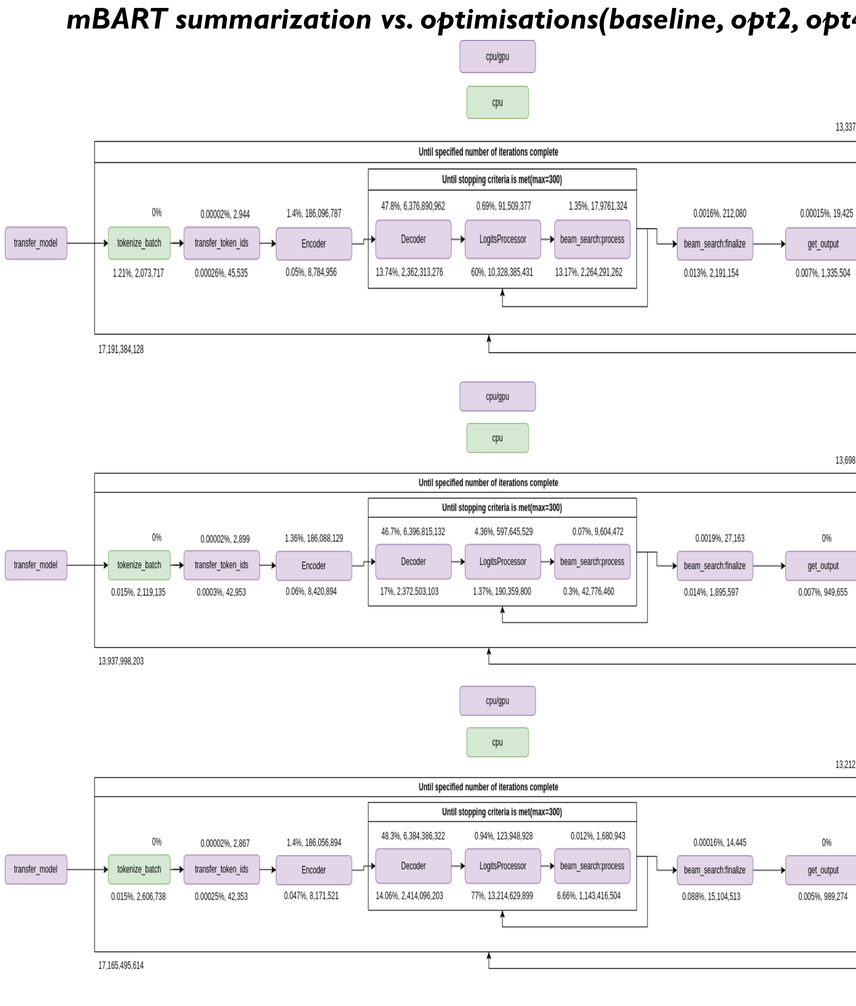

***mBART translation vs. batch size(1, 4, 16)***




***indicBART summarization vs. optimisations(baseline, opt2, opt4)***




***mBART summarization vs. optimisations(baseline, opt2, opt4)***




## Conclusion

- We present a workload characterization of LLM inference for indicBART/summarization and mBART/translation.
- Considerable amount of "unnecessary" device-to-host transfers happen in LogitProcess and BeamSearchProcess.
- Reduce inference latency by upto 32.4% for indicBART/summarization and 19% for mBART/translation by moving logic on the GPU.
- Reduce device-to-host transfer by 66.8% for indicBART/summarization and 99.1% for mBART/translation.
- DVFS plays a role in optimizing the energy efficiency of LLM inference. For some settings, we achieve 15% lower energy consumption at just 5% degradation in performance vs the configuration chosen by the DVFS.

***Baseline***

| Model | Batch Size | Seq Len | D2H Size |
|---|---|---|---|
| indicBART | 64 | 252 | 335KB |
| indicBART | 16 | 252 | 84KB |
| indicBART | 4 | 252 | 21KB |
| mBART | 16 | 203 | 10.5MB |
| mBART | 4 | 203 | 2.63MB |
| mBART | 1 | 203 | 0.66MB |

***Opt2***

| Model | Batch Size | Seq Len | D2H Size |
|---|---|---|---|
| indicBART | 64 | 252 | 111.18KB |
| indicBART | 16 | 252 | 28.12KB |
| indicBART | 4 | 252 | 7.355KB |
| mBART | 16 | 203 | 95.7KB |
| mBART | 4 | 203 | 24.39KB |
| mBART | 1 | 203 | 6.7KB |

***Baseline***

| Model | Batch Size | Seq Len | GPU Util % |
|---|---|---|---|
| indicBART | 64 | 252 | 70.4 |
| indicBART | 16 | 252 | 72.45 |
| indicBART | 4 | 252 | 71.98 |
| mBART | 16 | 203 | 76.9 |
| mBART | 4 | 203 | 78.1 |
| mBART | 1 | 203 | 69.9 |

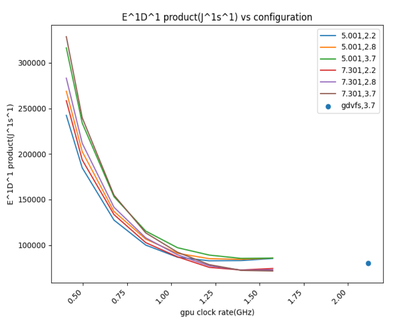
***Opt2***

| Model | Batch Size | Seq Len | GPU Util % |
|---|---|---|---|
| indicBART | 64 | 252 | 99.1 |
| indicBART | 16 | 252 | 97.4 |
| indicBART | 4 | 252 | 89.1 |
| mBART | 16 | 203 | 98.1 |
| mBART | 4 | 203 | 93.7 |
| mBART | 1 | 203 | 76.7 |

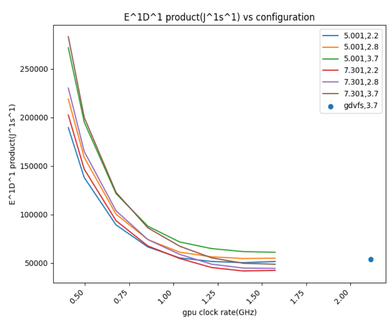
Want to know more?

Please scan me

***Baseline***




***Opt2***




***Baseline***

| Cpu clk (GHz) | GPU clk (GHz) | GPU mem clk(GHz) | Energy (J) | Wall Time(s) | EDP (Js) |
|---|---|---|---|---|---|
| 3.7 | 1.575 | 7.301 | 3,308 | 21.65 | 71,609 |
| 3.7 | 1.395 | 7.301 | 3,154 | 22.92 | 72,299 |
| 3.7 | gdvfs | gdvfs | 3,894 | 20.63 | 80,327 |

***Opt2***

| Cpu clk (GHz) | GPU clk (GHz) | GPU mem clk(GHz) | Energy (J) | Wall Time(s) | EDP (Js) |
|---|---|---|---|---|---|
| 2.2 | 1.395 | 7.301 | 2,223.7 | 18.78 | 41,765 |
| 2.2 | 1.575 | 7.301 | 2,417.7 | 17.54 | 42,401 |
| 3.7 | gdvfs | gdvfs | 3,388.4 | 15.921 | 53,946 |