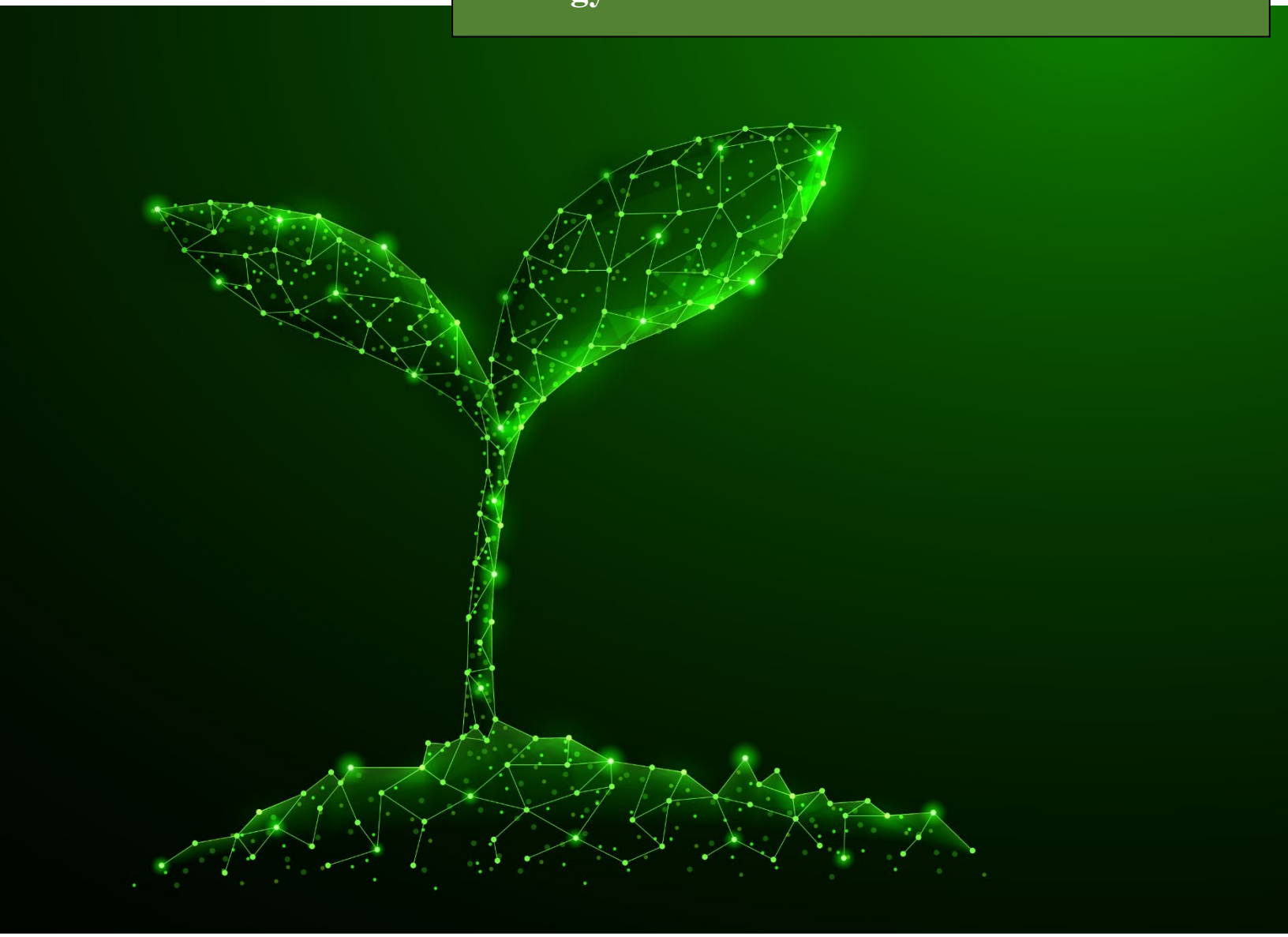
GSI TECHNOLOGY

Flexible Computation: Paving the Way for Energy-Efficient LLMs and a Greener Future



www.gsitechnology.com

Table of Contents

Introduction

The popularity of Generative AI (GenAI) applications, such as ChatGPT, is growing rapidly. A study by UBS found that two months after its launch, ChatGPT had 100 million monthly active users.[1] That makes it the fastest-growing consumer internet application to date.

A study by Accenture[2] stated that large language models (LLMs), like ChatGPT, could impact 40% of employee working hours. This is because the capability of LLMs is growing rapidly—due largely to their size.

LLMs have been growing by orders of magnitude with each successive generation. For example, GPT-2 has 1.5B parameters and GPT-3 has 175B parameters. It's believed that GPT-4 has more than 1T parameters.[3]

While this growth in model size has enabled more capable models, it also requires more computation, which means higher energy usage—negatively impacting climate change.

The number of parameters in LLMs is growing by orders of magnitude—leading to higher energy consumption, which impacts climate change.

The Move Towards Lower Carbon Footprints

To address the issue of climate change, many companies are looking to reduce their carbon footprint. For example, Microsoft stated that they're committed to being carbon negative by 2030.[4]

In a paper that was published by Hugging Face and others, it was stated that since LLMs are among the largest models, their impact on the carbon footprint should be understood.[5]

A lot of focus regarding carbon emissions has been on the training of LLMs since training them requires a lot of GPUs. According to an article by Bloomberg News, GPUs "are among the most power hungry the chip industry makes."[6] In that same article, they state that Sasha Luccioni, a researcher at Hugging Face, thinks that the total emissions from GPUs will be equal to that of a small country.

Inference—the Key to Reducing Emissions

While the impact of model training on carbon emissions is clearly a concern, it turns out that focusing on inference might have a greater impact on reducing emissions.

Researchers from Google and UC Berkeley state that "Most companies spend more energy on serving a DNN model (performing inference) than on training it."[7] They cite sources stating that inference accounts for 80-90% of machine learning workloads and demand.

In another paper published by researchers from Google and UC Berkeley, they state that inference accounts for more machine learning energy usage at Google than training, with inference representing about 3/5 and training about 2/5. "Inference represents about ⅗ of total ML energy usage at Google, owing to the many billion-user services that use ML." [8]

Quantization

Quantization is one of the main ways to reduce model inference energy. With quantization, you use fewer bits to represent data. For example, FP32 (floating point 32 bits) can be compressed to FP8 (floating point 8 bits).

When data is compressed, fewer data bits need to be transmitted to and from external memory. Access to external memory is the most power intensive access in the memory hierarchy, so reducing the amount of data accessed from it is critical to managing energy consumption.

Quantization is a key technique for reducing LLM energy use. However, traditional quantization formats are becoming less effective. Researchers require flexible solutions to explore different ways of quantizing data.

This has led to focused research on different quantization methods. An example of this is research by Microsoft and Meta on a Block Data Representations (BDR) framework that explores quantized formats based on shared microexponents (MX).[9]

In a paper detailing that research, Microsoft and Meta stated that scaling down traditional scalar floating point formats is providing diminishing returns. They say that alternate approaches to quantization are needed. [9]

In that paper, they mention that a diversity of data formats allows them to navigate the subtle trade-offs of efficiency (compute efficiency for energy management and memory efficiency to minimize the number of bits) and accuracy (model quality).

They explore formats where the model weights and activations have different quantization formats from one another. Thus, a highly flexible solution is needed to explore this diversity of formats.

Lack of Flexibility

Unfortunately, traditional solutions, like CPU and GPU, lack this flexibility. For example, as seen in **Figure 1**, GPUs, which are widely used for LLM inference and training, are limited to a small, fixed set of data formats (FP32, etc.). This means that they provide limited ability for researchers to deploy new methods such as shared microexponents, which restricts researchers' ability to innovate.

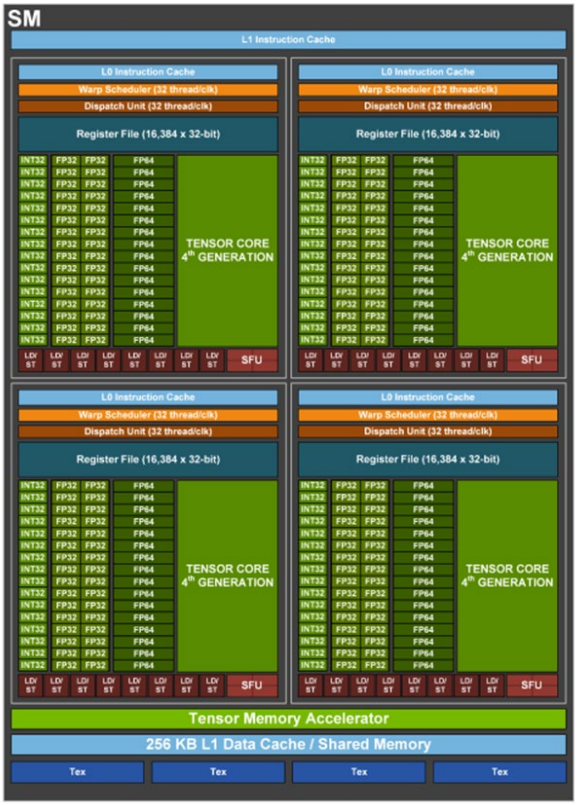


Figure 1 *GPUs have blocks of fixed data formats—limiting researchers' ability to explore new quantization methods. [Source: Nvidia]*

Additionally, some of those blocks could go unused if they do not match a user's needs. Not only does this impact performance, but it also wastes area that could have been used for valuable memory resources. This exacerbates the fact that GPUs already have a small cache memory and thus need to perform many energy-expensive data transfers to external main memory.

This is because LLMs have large weight matrices that they reuse frequently. GPUs have limited cache memory, so the large weight matrices need to be broken down into smaller "tiles" to fit in the cache. Since the weight matrices are large, and GPU caches are small, the tiles need to be re-fetched many times from main memory. Not only does this impact performance, but it also impacts power because main memory accesses require a lot of energy.

Another issue with architectures such as GPU stems from the fact that their caches are far from their processing elements. The farther that data must travel, the greater the energy needed to transfer that data. This is another hit to energy consumption.

What LLMs Need to Reduce Energy Use

What LLMs need to reduce energy use is a solution with a flexible data format that allows researchers to explore innovative quantization methods. The solution should also minimize data movement and make that movement more efficient.

Data movement can be minimized and made more efficient by:

- **Enabling compute on nonstandard data formats—** Traditional solutions, such as GPU, are limited to standard binary data formats (8-bit, 16-bit, 32-bit, etc.). If the data format being explored deviates from these standard binary formats, not all the bits transferred will be usable.

  For example, if a researcher were exploring 5-bit quantized formats, if they are forced to pack those 5 bits into a standard 8-bit format, they would be wasting 3/8ths of the bits. More data transfers to external memory would be needed to compensate for those wasted bits, resulting in increased energy use.

  A flexible data format addresses this waste by allowing for data to be packed more efficiently on the external memory interface. This minimizes data transfers, which reduces energy use.

- **Tightly coupling local memory to processing—**This minimizes the distance between the processing elements and their local memory, which reduces energy consumption.

  In addition to improving energy efficiency, tightly coupling local memory to processing also better utilizes silicon space. This allows for more memory to be placed on the die, which means fewer accesses to external memory, and thus lower energy use.

In-Memory Associative Computing—The Flexible Solution

GSI Technology's in-memory associative computing APU technology is based on flexible bit processing. The computation is performed at the bit level, so the architecture allows for computation on any sized data element—with a resolution as fine as single bit.

The flexibility of the bit processor allows:

- **Innovation—**The APU can process mixed data types on a cycle-by-cycle basis, and process different data types in different sections of the chip. This processing of mixed data types allows for exploration of innovative formats, such as the shared microexponent research from Microsoft and Meta, where the weights and activations have different quantization formats from one another.

- **Efficient Data Transfer—**Single-bit resolution allows data to be packed onto the external memory interface more efficiently. It eliminates unused bits, which reduces energy consumption by minimizing data transfer.

- **Full Utilization of all Processing Elements—**As **Figure 2** shows, the APU does not have fixed data formats. Its processors can handle any data format, unlike GPUs which may waste fixed-data format blocks if they don't match the user's needs. All processing elements in an APU can be fully utilized.

Also, as seen in **Figure 2**, the APU architecture has a tight coupling of local memory to the millions of bit processors on an APU chip. The local memory is aligned on the same column pitch as the bit processors to ensure tight coupling.

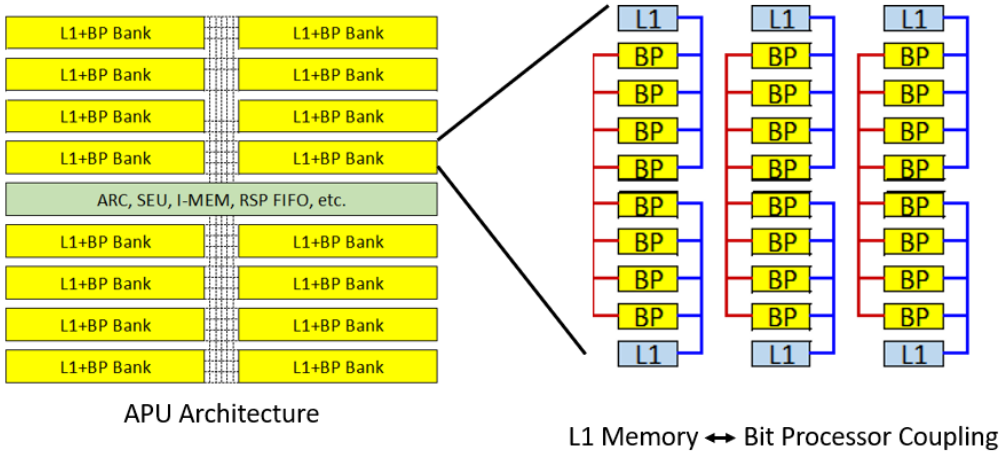


Figure 2 *APU architecture: Tight coupling of local memory (L1) to bit processing increases memory density and reduces the distance between memory and processing—resulting in lower energy use.*

This tight coupling and alignment between memory and processing provides multiple benefits:

- **Reduced Energy Consumption—**It results in a much shorter distance between the memory and processing. The distance between local memory and the bit processors in the APU is 1–2 orders of magnitude less than that between L1 cache and the register files in GPU—this significantly reduces energy consumption.

- **Higher Memory Density—**It allows for a more compact layout, which means more memory can be placed on a die. This higher local memory density means fewer accesses to external memory—lowering energy use.

Conclusion

GenAI applications, like ChatGPT, are becoming more capable with each generation. This improved capability is driven largely by the orders of magnitude growth in the number parameters used by the LLMs that power them.

Unfortunately, more parameters means more compute, which leads to higher energy usage and ultimately a larger carbon footprint.

To help combat carbon footprint growth, researchers are exploring new methods of quantization to compress data. Data compression helps reduce data transmission from external memory. This

significantly reduces energy use because accessing data from external memory is the most power intensive access in the memory hierarchy.

There are subtle system-level tradeoffs between the diverse set of quantized data formats that researchers are investigating. To allow researchers the freedom to navigate these subtle tradeoffs and explore innovative quantization methods, they need a flexible solution.

Unfortunately, traditional solutions, like CPU and GPU, lack this flexibility and are limited to a small, fixed set of data formats.

GSI Technology's APU technology provides researchers with the flexibility they need to explore new quantization methods. It allows for computation to be performed at the bit level, so computation can be performed on any sized data element—with a resolution as fine as single bit.

To learn more about how the APU's bit-level processing can lower the carbon footprint of your LLM applications, contact us at associativecomputing@gsitechnology.com.

---

[1] Paris, Martine. "ChatGPT Hits 100 Million Users, Google Invests In AI Bot And CatGPT Goes Viral." *Forbes*, Forbes, 3 Feb. 2023, https://www.forbes.com/sites/martineparis/2023/02/03/chatgpt-hits-100-million-microsoft-unleashes-ai-bots-and-catgpt-goes-viral/?s.

[2] Accenture. "A New Era of Generative AI for Everyone." *Accenture*, https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf. Accessed 18 July 2023.

[3] Lutkevich, Ben. "What Is GPT-4? Everything You Need to Know | TechTarget." *WhatIs.Com*, TechTarget, 20 Mar. 2023, https://www.techtarget.com/whatis/definition/GPT-4.

[4] "Our Sustainability Journey | Microsoft CSR." *Microsoft*, https://www.microsoft.com/en-us/corporate-responsibility/sustainability-journey. Accessed 18 July 2023.

[5] Luccioni, Alexandra Sasha, et al. "ESTIMATING THE CARBON FOOTPRINT OF BLOOM, A 176B PARAMETER LANGUAGE MODEL." *Arxiv*, Cornell University, 3 Nov. 2022, https://arxiv.org/pdf/2211.02001.pdf.

[6] "Bloomberg - Are You a Robot?" *Bloomberg - Are You a Robot?*, https://www.bloomberg.com/news/articles/2023-03-09/how-much-energy-do-ai-and-chatgpt-use-no-one-knows-for-sure?leadSource=uverify%20wall. Accessed 18 July 2023.

[7] Patterson, David, et al. "Carbon Emissions and Large Neural Network Training." *Arxiv*, Cornell University, https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf. Accessed 18 July 2023.

[8] Patterson, David, Joseph Gonzalez, Urs Holzle, et al. "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink." *Arxiv*, Cornell University, https://arxiv.org/ftp/arxiv/papers/2204/2204.05149.pdf. Accessed 18 July 2023.

[9] Rouhani, Bita, et al. "With Shared Microexponents, A Little Shifting Goes a Long Way." *Arxiv*, Cornell University, 13 Apr. 2023, https://arxiv.org/pdf/2302.08007.pdf.

