Green Computing with Remote GPU Service for Generative AI / LLM Use Case

- Light Speed Data Transfer for AI Training -

Classification: APPROVED REFERENCE DOCUMENT

Confidentiality: [PUBLIC]

Version 1.0

August 21, 2024

[GC with Remote GPU Use Case]

Legal

THIS DOCUMENT HAS BEEN DESIGNATED BY THE INNOVATIVE OPTICAL AND WIRELESS NETWORK GLOBAL FORUM, INC. (“IOWN GLOBAL FORUM”) AS AN APPROVED REFERENCE DOCUMENT AS SUCH TERM IS USED IN THE IOWN GLOBAL FORUM INTELLECTUAL PROPERTY RIGHTS POLICY (THIS “REFERENCE DOCUMENT”).

THIS REFERENCE DOCUMENT IS PROVIDED ”AS IS” WITH NO WARRANTIES WHATSOEVER, WHETHER EXPRESS, IMPLIED, STATUTORY, OR OTHERWISE, INCLUDING WITHOUT LIMITATION ANY WARRANTY OF MERCHANTABILITY, NONINFRINGEMENT OF THIRD PARTY RIGHTS, TITLE, VALIDITY OF RIGHTS IN, FITNESS FOR ANY PARTICULAR PURPOSE, OR ANY WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, REFERENCE DOCUMENT, SAMPLE, OR LAW. WITHOUT LIMITATION, IOWN GLOBAL FORUM DISCLAIMS ALL LIABILITY, INCLUDING WITHOUT LIMITATION LIABILITY FOR INFRINGEMENT OF ANY PROPRIETARY RIGHTS AND PRODUCTS LIABILITY, RELATING TO USE OF THE INFORMATION IN THIS REFERENCE DOCUMENT AND TO ANY USE OF THIS REFERENCE DOCUMENT IN CONNECTION WITH THE DEVELOPMENT OF ANY PRODUCT OR SERVICE, AND IOWN GLOBAL FORUM DISCLAIMS ALL LIABILITY FOR COST OF PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES, LOST PROFITS, LOSS OF USE, LOSS OF DATA OR ANY INCIDENTAL, CONSEQUENTIAL, DIRECT, INDIRECT, PUNITIVE, EXEMPLARY, OR SPECIAL DAMAGES, WHETHER UNDER CONTRACT, TORT, WARRANTY OR OTHERWISE, ARISING IN ANY WAY OUT OF USE OR RELIANCE UPON THIS REFERENCE DOCUMENT OR ANY INFORMATION HEREIN.

EXCEPT AS EXPRESSLY SET FORTH IN THE PARAGRAPH DIRECTLY BELOW, NO LICENSE IS GRANTED HEREIN, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS OF THE IOWN GLOBAL FORUM, ANY IOWN GLOBAL FORUM MEMBER OR ANY AFFILIATE OF ANY IOWN GLOBAL FORUM MEMBER. EXCEPT AS EXPRESSLY SET FORTH IN THE PARAGRAPH DIRECTLY BELOW, ALL RIGHTS IN THIS REFERENCE DOCUMENT ARE RESERVED.

THIS DOCUMENT IS AN APPROVED REFERENCE DOCUMENT AND IS SUBJECT TO THE REFERENCE DOCUMENT LICENSING COMMITMENTS OF THE MEMBERS OF THE IOWN GLOBAL FORUM PURSUANT TO THE IOWN GLOBAL FORUM INTELLECTUAL PROPERTY RIGHTS POLICY. A COPY OF THE IOWN GLOBAL FORUM INTELLECTUAL PROPERTY RIGHTS POLICY CAN BE OBTAINED BY COMPLETING THE FORM AT: www.iowngf.org/join-forum. USE OF THIS REFERENCE DOCUMENT IS SUBJECT TO THE LIMITED INTERNAL-USE ONLY LICENSE GRANTED ABOVE. IF YOU WOULD LIKE TO REQUEST A COPYRIGHT LICENSE THAT IS DIFFERENT FROM THE ONE GRANTED ABOVE (SUCH AS, BUT NOT LIMITED TO, A LICENSE TO TRANSLATE THIS REFERENCE DOCUMENT INTO ANOTHER LANGUAGE), PLEASE CONTACT US BY COMPLETING THE FORM AT: https://iowngf.org/contact-us/

The IOWN GLOBAL FORUM mark and IOWN GLOBAL FORUM & Design logo are trademarks of Innovative Optical and Wireless Network Global Forum, Inc. in the United States and other countries. Unauthorized use is strictly prohibited. IOWN is a registered and unregistered trademark of Nippon Telegraph and Telephone Corporation in the United States, Japan, and other countries. Other names and brands appearing in this document may be claimed as the property of others.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Opportunities for Generative AI / LLM

Since ChatGPT was launched in late 2022, the potential for AI-driven business transformation has expanded globally. Rather than simply analyzing and classifying existing data, Generative AI can create entirely new things, including text, images, audio, and synthetic data. It brings breakthroughs in human creativity and productivity to business, science, and society.

Generative AI (GenAI) applications are built by foundation models trained with complex deep learning systems at scale on vast amounts of data. Most of today's foundation models are large-scale language models (LLMs) trained on natural language processing.

The following opportunities are examples of use cases where LLMs can be applied.

1. Business meetings

LLM applications can summarize the content of business meetings, especially extracting the points of discussion, what should be done, action items, and next steps, as well as clarifying the meeting’s results.

2. Call centers

After converting the conversation information between customers and call center staff into text, the LLM application can evaluate the content of customer’s inquiries and search for similar responses to suggest potential next steps and responses.

3. Medical record information

An LLM application can review and summarize patient examination results and other vital information such as, medications and interview responses. It can also support doctors’ diagnoses based on similar symptoms and follow-up observations.

1.2. Purpose of Generative AI / LLM with IOWN project

These large-scale calculations for Generative AI / LLM consume considerable electric power, so we need to effectively use locally installed computing resources in the green datacenter as much as possible. Previous GPU computing services assumed that computing and data storage resources were located in the same data center. To effectively utilize the large amounts of data located beyond the data center, end-users will require a high-performance network, such as an IOWN-based All-Photonic Network (APN). The purpose of this project is to assume a use case in which large amounts of training data related to Generative AI / LLM are referenced between the green datacenter and user location using an IOWN APN, which will contribute to a reduction in training time, power consumption, and cost for generating an LLM.

1.3. Scope of Generative AI / LLM with IOWN project

We are attempting to show the practicality of an Open APN (All Photonic Network) when applied to this use case by comparing the training time and power consumption using a typical WAN line versus an Open APN, while changing the training data size, model size, and the number of calculations. In addition, the security of the entire system leveraging IOWN infrastructure (including the Open APN) must be considered for the data owner to maintain data sovereignty in this use case.

2. Use Case

The envisioned use case is for users to generate an LLM by performing training on GPU computing located on Open APN-connected green data centers using the training data that exists at their own location. It is assumed that many packets used for data reference will be transferred on the Open APN and that the network latency will affect model training time and power consumption regarding GPU computing (Figure 2-1).

For this example, we will focus on the training phase of building an LLM and will separately consider the federated learning with multi-site and the inference processing using an LLM later.

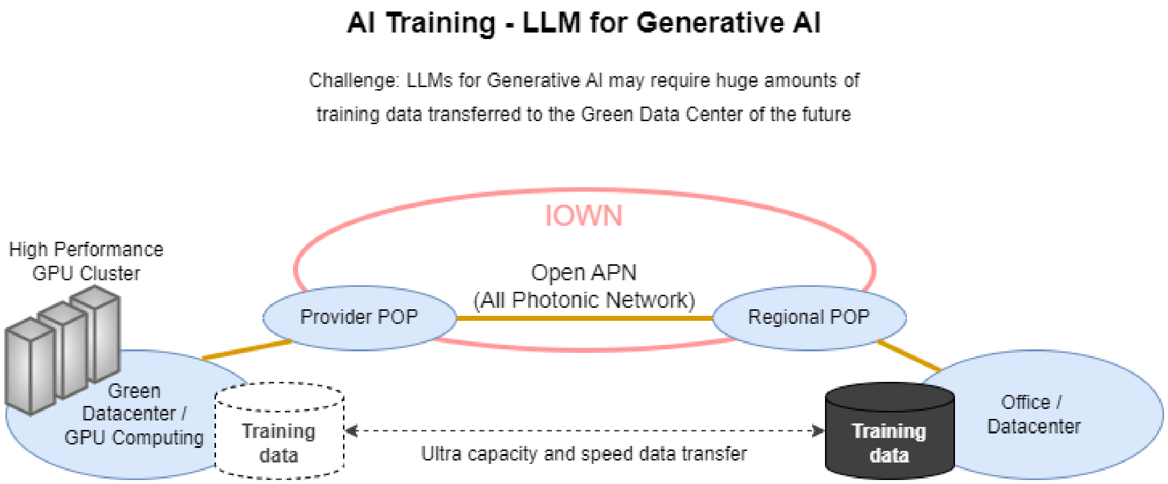
AI Training - LLM for Generative AI
Challenge: LLMs for Generative AI may require huge amounts of
training data transferred to the Green Data Center of the future
IOWN
High Performance
GPU Cluster
Open APN
(All Photonic Network)
Provider POP
Regional POP
Green
Datacenter /
GPU Computing
Training
data
Ultra capacity and speed data transfer
Training
data
Office /
Datacenter

Figure 2-1. Training Data Transfer on the Open APN to GPU Clusters for LLMs

GPU-based computing is becoming exponentially faster, and training of the LLM model that used to take months will soon be completed in only a few days. When this happens, it is expected that it will become more common for LLMs to be trained and customized using a company's or individual's internal data.

On the other hand, high-performance GPU computing consumes more power and has higher heat density in proportion to its performance, so a green power supply and water-cooling system will be essential. The above Green Datacenter, which has both of these, is the core of efficient operation of high-performance GPU computing.

For use cases where multiple users train their respective AI models, the Green Data Center becomes an AI factory, where the customer's data and models are received, trained, and the parameters are transferred. From a data-centric perspective, the image is "the required GPU computing resources are delivered from the Green Data Center to the customer's data storage, and the models are trained in the customer's environment.” This use case requires that the GPU computing in the green data center and the Open APNs connected to them should be available on-demand for as long as needed.

In addition, protection of training data and models must be considered in this use case. Confidentiality and auditability of the entire data lifecycle including the computing space are the key to achieving this use case.

3. Key Requirements

In the training phase of the LLM, the LLM tends to increase the parameter size considered, and accordingly, the data size tends to increase as well. As a result, total traffic size of the data reference will also increase. Some generic LLMs have more than 100B parameters. However, LLMs larger than 100B require a large number of GPUs and are too large to customize with a variety of user data.

Performance requirements for storage greatly depend on the types of AI models and data formats used. The guidelines in Table 1 and Table 2 are provided to help determine the I/O levels required for different types of AI models.

Table 1. Characterizing different I/O workloads

Storage Performance Level Required	Example Workloads
Good	NLP
Better	Image processing with compressed images, ImageNet/ResNet-50
Best	Training with 1080p, 4K, or uncompressed images

Table 2. Storage Performance Requirements for General Purpose AI Platform

Performance Characteristic	Good (Gbps)	Better (Gbps)	Best (Gbps)
Single-node read	32	64	320
Single-node write	16	32	160

Considering the LLM model training workload, this project assumes the “Better” level performance in the table above.

Since remote training via Open APN is slower than local training, the rate of increase in the overall training time is the most important metric. For example, an **Open APN should NOT increase the overall training time more than 10%**, because the acceptable increase time for a typical month of training time for an LLM is about 3 days.

In an environment without an Open APN, the green data center and data storage are connected via the Internet. Storage caching and replication technologies are used to eliminate Internet latency.

Caching is a method of anticipating data to be read and copied it to local storage in advance. Replication is a method of pre-copying the required data sets to the local storage.

One strong incentive to avoid copying the training data used for AI/LLM customization out of your site is that it many contain sensitive data such as company-specific technical Intellectual Property (IP) or personal employee information. In this case, when training an AI/LLM using GPUs in a remote green computing center as a third-party service, the confidentiality of the training data and model must be guaranteed within the communication path and the actual computing space, including within the green data center. Confidentiality here means that training data and models cannot be accessed by anyone other than the data owner, including the operators of the green data center. Ideally, the confidentiality of the entire system, including user sites, should be maintained.

To achieve the above, we need to seamlessly connect mechanisms to protect data in motion and data in use from data originally stored, such that there are no unprotected moments in the sequence of transferring training data from storage to the GPU servers, training the model, and returning the result parameters to storage. As an additional level of confidentiality, it must be secure against quantum computers, a new threat coming in near future.

In order for the data owner to obtain such assurance of confidentiality, authentication of the data transfer destination, provision of authenticity such as remote attestation, and auditability must also be considered.

4. Technology Evaluation Criteria

4.1. Reference Case

As shown in the Use Case chapter, file access throughput increases in proportion to the LLM parameter size. To simplify the problem, consider technology evaluation criteria for a small LLM of size 10B, which is easy to train.

- **Our targeting LLM: 10B parameters**.

The simplest architecture for the training LLM with 10B parameters has a 100 Gbps Ethernet connection between GPU computing and remote data storage, as shown in Figure 4-1.

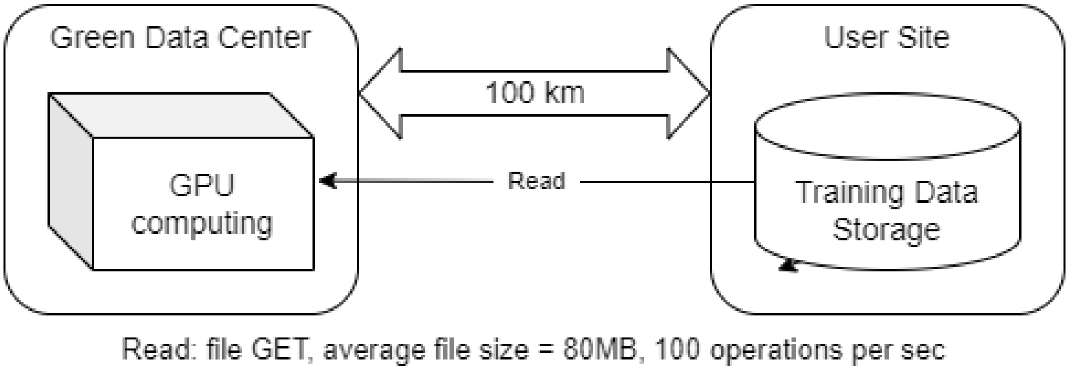

Figure 4-1. Reference Architecture for Remote GPU over Open APN

4.2. Key Benchmarks

Assuming a multimodal LLM with a large training data size, the following conditions for the workload are considered:

Table 3 shows file access conditions for general purpose training, that is 64Gbps read and 32Gbps write. We will assume 100Gbps is good enough for single node GPU and Storage to do training.

Classification of LLM Training Environments in Green Datacenter with Remote Data Access

The following is a summary of the methods used to connect a remote Green Datacenter to the company's owned data for training.

Table 3. Guideline for using the recommended training scale by configuration

Training Data Access Methods	Inter-site network	Data copy	Comparison with local training	Recommended training scale (training time, dataset)
Replication	Internet	Required	All data must be copied remotely to Replication storage in advance	- Long hours (>weeks) - Small (>10TB)
Cache	Internet	Required	It takes some time for data to be copied to cache, but after that it is fast	- Long hours (>weeks) - Medium (>100TB)
Cache	Open APN	Required	If RTT is less than a few ms, there is almost no difference	- Short time (<1 week) - Medium (>100TB)
None (Direct access)	Open APN	Not Required	If RTT is less than a few ms, there is almost no difference	- Short time (<1 week) - Large (<100TB)

4.2.1. Benchmark

Each access method and description is shown below.

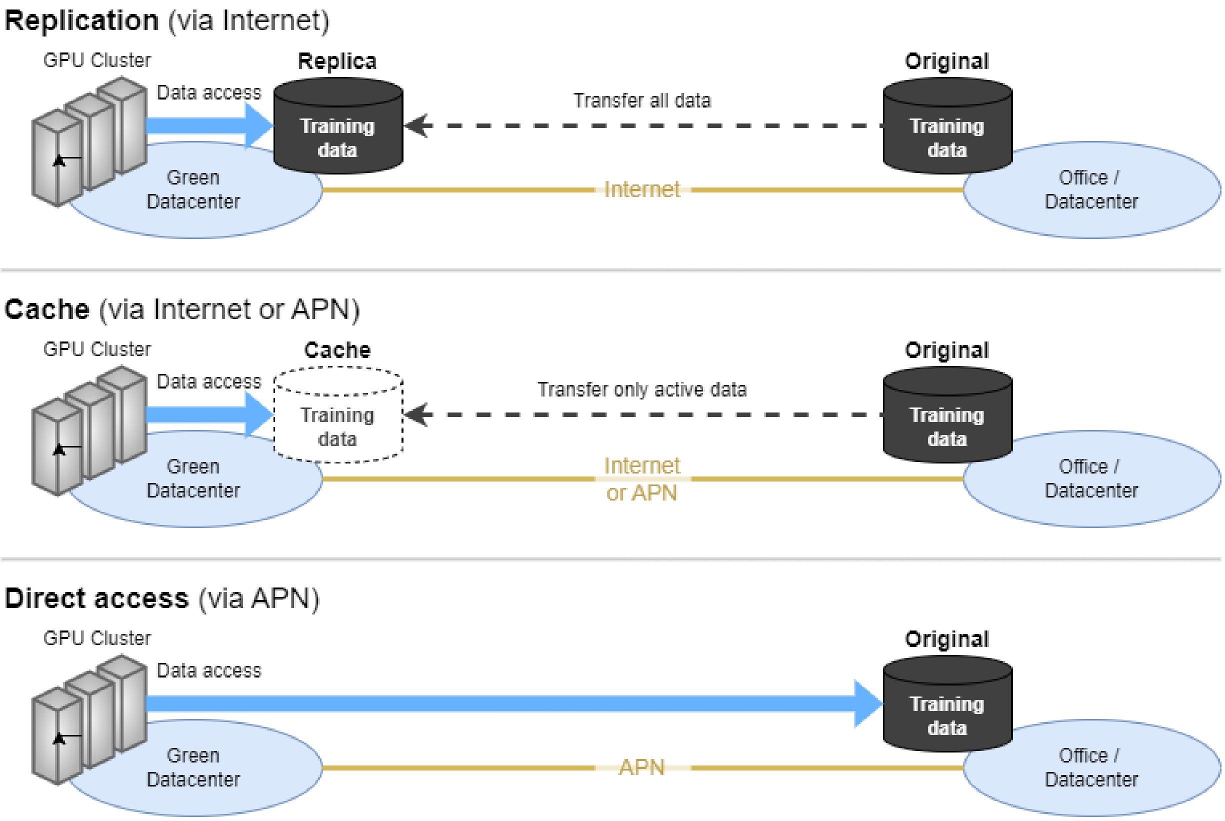

Figure 4-2 GPU/Storage - three storage access methods

- Replication
  - This method copies all training data to Green DC and GPU accesses Replica in Green DC. This eliminates network latency between sites in data access and reduces training time.
  - This assumes a connection via the Internet, since network latency between sites does not affect training time in this method.
- Cache
  - This method caches only the data actually needed by GPU for training in Green DC, same as “Replication”, this method eliminates network latency between sites and reduces training time.
  - From GPU, Cache appears to have all of Origin's data, but in reality, there is no all data on the cache, and only the data read by GPU is cached (if GPU accesses non-cached data, the target data block is transferred from Origin to Cache in the background).
  - Data written to Cache is transferred to Origin at any time in the background.
  - This assumes a connection via the Internet or Open APN, since network latency between sites affects the training time in this method.

- Direct access
  - In this method, GPU accesses training data on Office side via Open APN. Since Open APN latency slows down data read speed, training time is expected to be slower than other methods.
  - This assumes a connection via Open APN, since network latency between sites affects the training time in this method.
  - This method also needs to confirm that even if Open APN is encrypted with post-quantum security, it does not have a significant impact on network latency.

4.2.2. Metrics

Since remote training via Open APN is slower than local, the rate of increase in training time is the most important metric. For example, Open APN should increase training time by less than 10%, because the acceptable increase time for a typical month of training time in LLM is about 3 days, which are required to copy tens of terabytes of data offline or online.

1. Training Time
2. Storage access speed (Throughput, Latency)
3. Power consumption
4. Equipment cost

4.2.3. Other conditions

In addition, since green data centers will be prepared in urban areas, the conditions for the network are as follows.

1. Datacenter distance between GPU computing and Training data: 100 km or longer
2. Network bandwidth: 100 Gbps or higher

5. Conclusion

IOWN enables the ability to generate your own customized LLMs without storing your most valuable data outside of your site. In other words, IOWN accelerates your digital transformation and keeps your digital sovereignty.

A) Appendix

Figure A-1 is a reference architecture for the storage configuration of a large GPU cluster. In the Leaf-Spine topology, the architecture recommends 400Gpbs of bandwidth in the Spine portion and 200Gbps in the Leaf portion where traffic is concentrated.

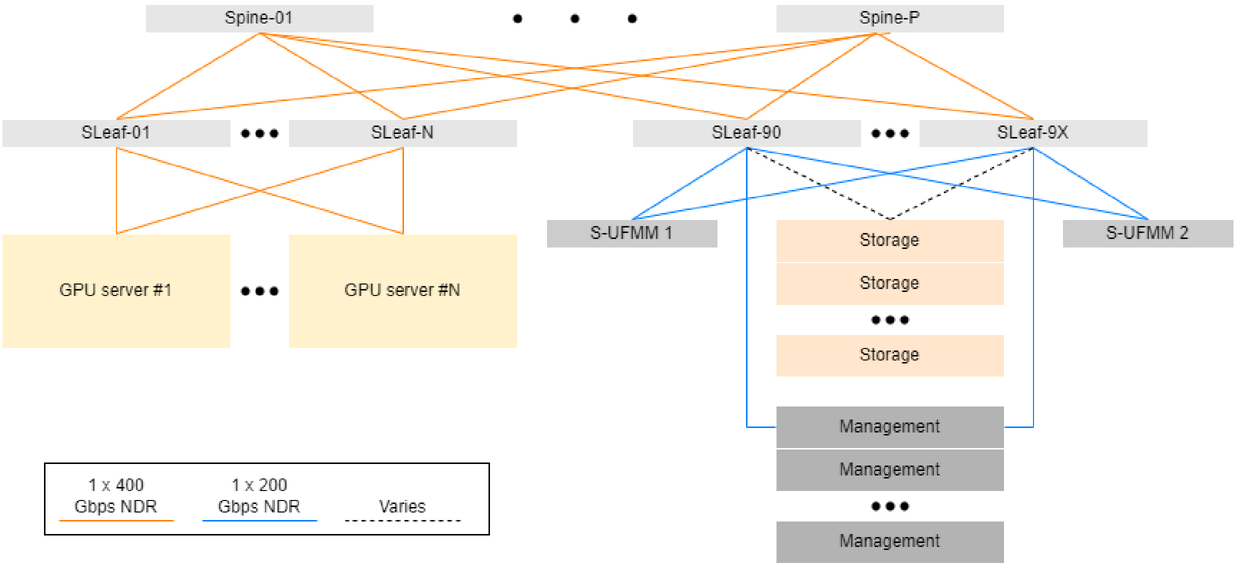

Figure A-1. Sample Architecture for Storage Network in a large GPU cluster

Acknowledgments

The management team is listed here, followed by contributors in alphabetical order.

Fumiaki Kudoh
Hideaki Tagami
Joao Kluck Gomes
Kei Karasawa
Nobu Sasaki
Yoshinobu Nakayama

History

Revision	Release Date	Summary of Changes
1.0	August 21, 2024	Initial Version