Panthalia: Verifiable Compute for Superintelligence

Goliath
Ritser Labs
goliath@ritser.com

Godking
Ritser Labs
godking@ritser.com

Abstract

As machine learning models grow in scale, so too does the demand for computational resources, driving the emergence of peer-to-peer compute marketplaces. However, these platforms face key challenges in ensuring the reliability and integrity of computational services, potentially leading to inefficiencies and dishonest practices. In response, we introduce *Panthalia*, a modular, verifiable compute marketplace that employs optimistic staking and probabilistic verification to ensure computational integrity while minimizing costs and latency. Panthalia's architecture incentivizes honest behavior through economic mechanisms, reducing the risk of dishonest actors. We demonstrate the viability of Panthalia by training a 124-million-parameter nanoGPT language model on the FineWeb dataset using four distributed nodes, showcasing its potential to decentralize and democratize large-scale model training.

1 Introduction

Scaling laws have shown that as the computational resources devoted to training increase, language models exhibit superior generalization [Rich Sutton, 2019]. Consequently, there has been a sharp rise in the demand for compute, as organizations compete to train larger and more sophisticated models.

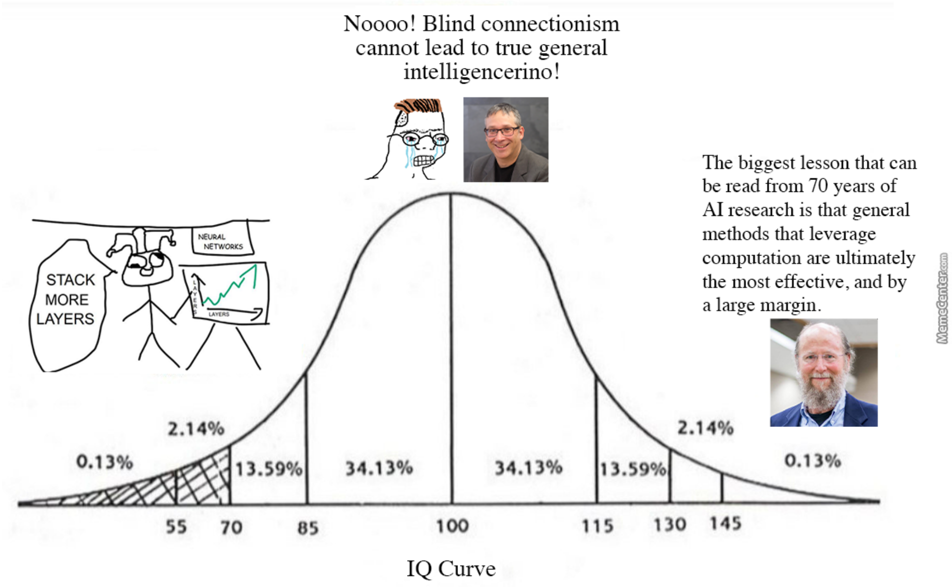

Noooo! Blind connectionism cannot lead to true general intelligencerino!
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.
STACK MORE LAYERS
NEURAL NETWORKS
LAYERS
LAYERS
MemeCenter.com
0.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
0.13%
55
70
85
100
115
130
145
IQ Curve

Figure 1: On the scaling of models [roon [@tszzl], 2020]

While traditional cloud providers dominate the market, peer-to-peer compute marketplaces have emerged as an alternative, allowing individuals and small businesses to rent out their computational resources. However, these marketplaces face a critical issue: ensuring that compute providers remain honest. Providers may allocate resources across multiple instances, significantly reducing actual performance compared to what is advertised.

We propose *Panthalia*, a marketplace that addresses this challenge by requiring compute providers to stake tokens to ensure honesty. Compute results are sampled probabilistically, and if a result is proven incorrect via a voting mechanism, the staked tokens are slashed. The stake is set high enough to ensure that providers are incentivized to act honestly, even when the sampling probability is low [Zhang and Wang, 2024].

We benchmark existing compute providers to illustrate the need for Panthalia. We then simulate Panthalia's functionality and apply it to train a 124-million-parameter nanoGPT model on the FineWeb dataset using distributed nodes [Andrej Karpathy, 2023, Penedo et al., 2024].

2 Benchmarking Existing Compute Providers

We assessed the reliability of current compute providers by renting 19 instances:

- five from A, a cloud provider selling compute from "T3/T4 data centers";
- five from B, invite-only, peer-to-peer compute offered by the same company as A;
- four from C, another provider reselling third-party cloud compute; and
- five from D, a peer-to-peer compute marketplace.

Using `speedtest-cli`, we measured the download and upload bandwidth of each instance and compared it to the advertised bandwidth. The average advertised and measured bandwidth for each provider is shown in Figure 2. Detailed results are available in the appendix. Instance 7, an instance from provider B, is excluded from Figure 2 as it was unable to execute a speed test.

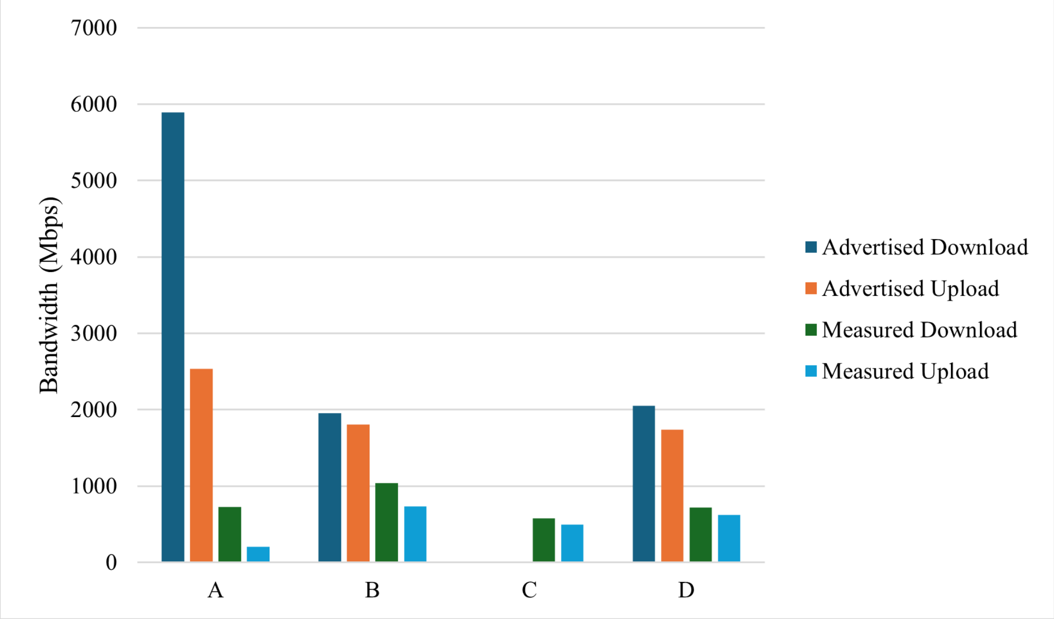


Figure 2: Average Advertised vs. Measured Bandwidth for Existing Compute Providers

The results reveal substantial discrepancies between advertised and measured bandwidth, with some instances showing extreme bandwidth reductions. For example, one instance advertised a download bandwidth of 7,334 Mbps but measured only 459.1 Mbps, likely due to providers sharing bandwidth among multiple instances. The absence of penalties for inaccurate advertising may contribute to this behavior.

We also benchmarked GPU performance by timing tensor operations across 100 runs, with the results summarized in Figure 3.

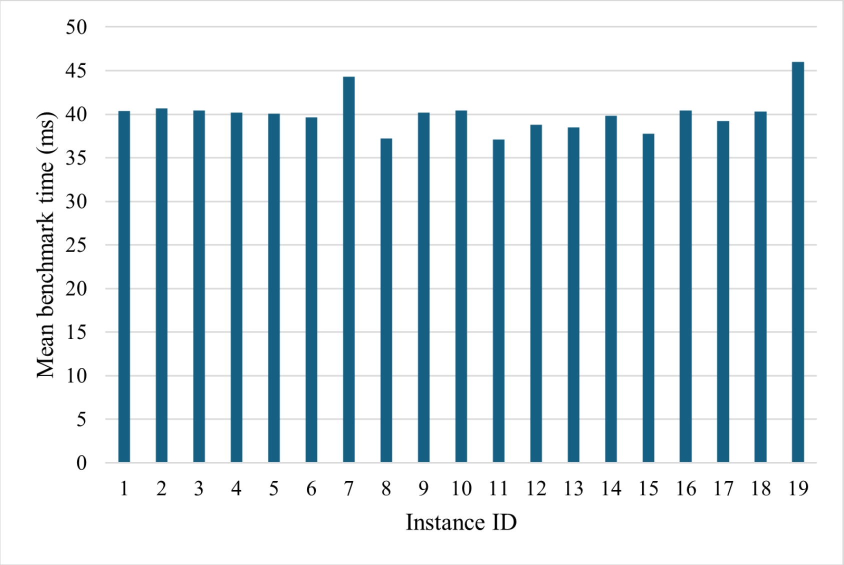


Figure 3: Mean Benchmark Time for 100 Runs Per Instance of Tensor Operations

GPU performance results showed smaller discrepancies between advertised and measured values, indicating that bandwidth sharing is more problematic than GPU performance degradation.

3 Panthalia

Panthalia is a decentralized compute marketplace that ensures honesty by requiring providers to stake tokens. Compute results are probabilistically sampled and verified through a voting mechanism, where dishonest providers have their stakes slashed.

The operation of Panthalia follows the steps illustrated in Figure 4:

1. **Task Submission:** Buyers submit tasks to Panthalia, specifying the computational requirements.
2. **Solver Processing:** A solver is randomly selected from the pool of staked providers, computes the task, and submits the result for a reward.
3. **Dispute:** If any participant suspects an incorrect solution, they can initiate a dispute by paying a fee. The dispute is resolved via a voting mechanism, where verifiers assess the correctness of the result. If the dispute is successful, the stakes of the solver and losing verifiers are slashed.
4. **Additional Rounds:** Disputes can be escalated through further rounds as needed.

Panthalia enables the creation of subnet-specific instances, each optimized for distinct computational tasks (e.g., AI inference, rendering, reinforcement learning). This architecture is modular and highly customizable, allowing users to define plugins tailored for specific PyTorch models, enabling composable workflows across various machine learning tasks. Task-solving and verification are performed client-side, while payments and dispute resolution are securely managed on-chain via the blockchain. To ensure deterministic outcomes, manual seeding is required for any random number generation within the system.

Although Panthalia's probabilistic verification shares some similarities with Proof-of-Sampling-Protocol (PoSP) [Zhang and Wang, 2024], our approach to probabilistic verification, which was developed in March 2024, distinguishes itself through its strong emphasis on reputation-based heuristics, modular subnets, and participant-driven disputes, offering task-specific optimizations and flexibility across a wide range of computational needs.

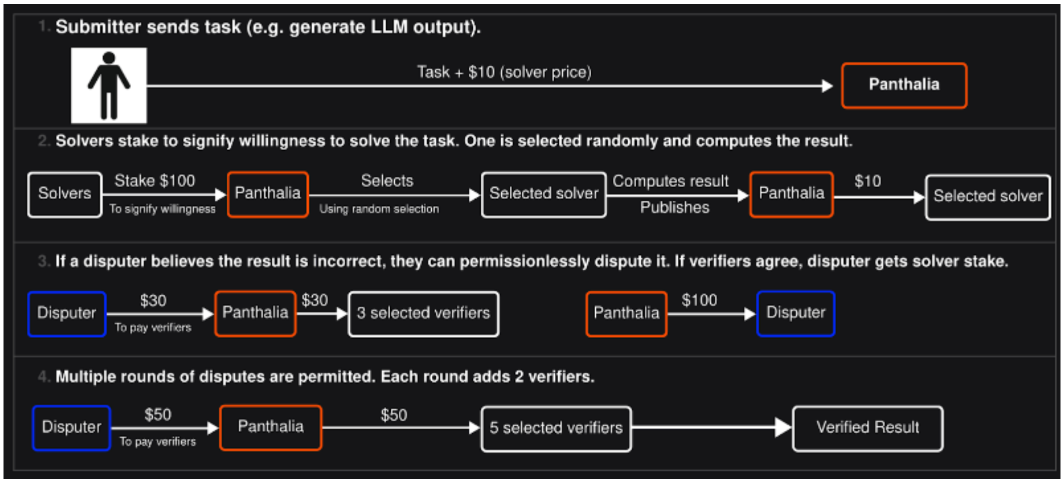
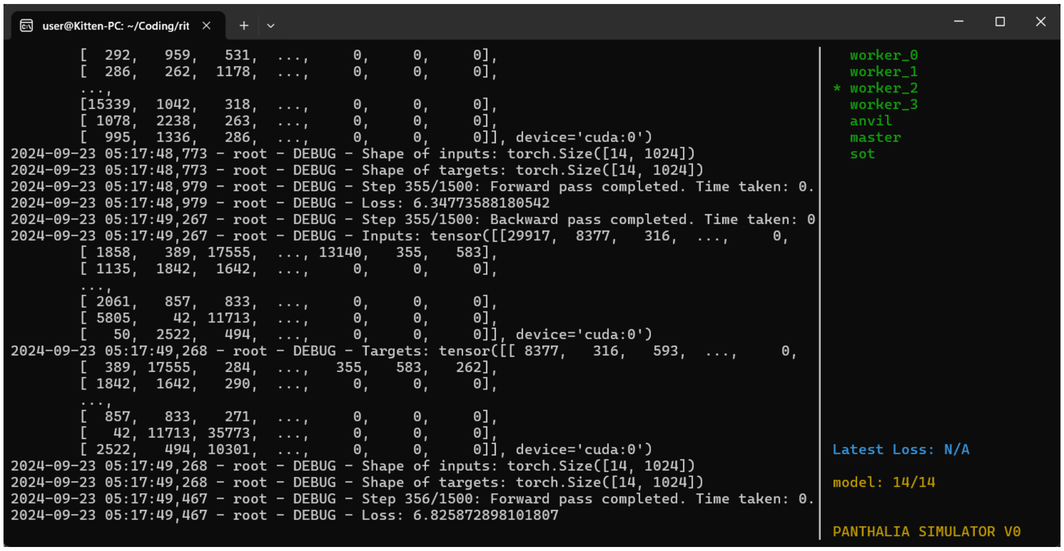

1. Submitter sends task (e.g. generate LLM output).
Task + $10 (solver price)
Panthalia
2. Solvers stake to signify willingness to solve the task. One is selected randomly and computes the result.
Solvers
Stake $100
To signify willingness
Panthalia
Selects
Using random selection
Selected solver
Computes result
Publishes
Panthalia
$10
Selected solver
3. If a disputer believes the result is incorrect, they can permissionlessly dispute it. If verifiers agree, disputer gets solver stake.
Disputer
$30
To pay verifiers
Panthalia
$30
3 selected verifiers
Panthalia
$100
Disputer
4. Multiple rounds of disputes are permitted. Each round adds 2 verifiers.
Disputer
$50
To pay verifiers
Panthalia
$50
5 selected verifiers
Verified Result

Figure 4: The Stages of Panthalia

3.1 Implementation

We implemented and simulated Panthalia to train arbitrary PyTorch models over the internet. The simulation includes:

- **Anvil:** A blockchain testnet client from Foundry.
- **Source of Truth (SOT):** A server that manages model weights and implements the DiLoCo outer optimizer [Douillard et al., 2024].
- **Master:** The entity submitting task requests to Panthalia.
- **Workers:** Compute providers who stake tokens and execute tasks as solvers. They implement the DiLoCo inner optimizer.



```
        [  292,   959,   531,  ...,     0,     0,     0],
        [  286,   262,  1178,  ...,     0,     0,     0],
        ...,
        [15339,  1042,   318,  ...,     0,     0,     0],
        [ 1078,  2238,   263,  ...,     0,     0,     0],
        [  995,  1336,   286,  ...,     0,     0,     0]], device='cuda:0')
2024-09-23 05:17:48,773 - root - DEBUG - Shape of inputs: torch.Size([14, 1024])
2024-09-23 05:17:48,773 - root - DEBUG - Shape of targets: torch.Size([14, 1024])
2024-09-23 05:17:48,979 - root - DEBUG - Step 355/1500: Forward pass completed. Time taken: 0.
2024-09-23 05:17:48,979 - root - DEBUG - Loss: 6.34773588180542
2024-09-23 05:17:49,267 - root - DEBUG - Step 355/1500: Backward pass completed. Time taken: 0
2024-09-23 05:17:49,267 - root - DEBUG - Inputs: tensor([[29917,  8377,   316,  ...,     0,
        [ 1858,   389, 17555,  ..., 13140,   355,   583],
        [ 1135,  1842,  1642,  ...,     0,     0,     0],
        ...,
        [ 2061,   857,   833,  ...,     0,     0,     0],
        [ 5805,    42, 11713,  ...,     0,     0,     0],
        [   50,  2522,   494,  ...,     0,     0,     0]], device='cuda:0')
2024-09-23 05:17:49,268 - root - DEBUG - Targets: tensor([[ 8377,   316,   593,  ...,     0,
        [  389, 17555,   284,  ...,   355,   583,   262],
        [ 1842,  1642,   290,  ...,     0,     0,     0],
        ...,
        [  857,   833,   271,  ...,     0,     0,     0],
        [   42, 11713, 35773,  ...,     0,     0,     0],
        [ 2522,   494, 10301,  ...,     0,     0,     0]], device='cuda:0')
2024-09-23 05:17:49,268 - root - DEBUG - Shape of inputs: torch.Size([14, 1024])
2024-09-23 05:17:49,268 - root - DEBUG - Shape of targets: torch.Size([14, 1024])
2024-09-23 05:17:49,467 - root - DEBUG - Step 356/1500: Forward pass completed. Time taken: 0.
2024-09-23 05:17:49,467 - root - DEBUG - Loss: 6.825872898101807
```

```
  worker_0
  worker_1
* worker_2
  worker_3
  anvil
  master
  sot

Latest Loss: N/A

model: 14/14

PANTHALIA SIMULATOR V0
```

Figure 5: The Panthalia Simulator Interface

The simulation demonstrated the feasibility of distributing computational tasks across decentralized nodes. For this simulation, we used four cloud nodes, each equipped with a NVIDIA RTX 4090 GPU, as the worker nodes. However, the Source of Truth (SOT) was also run on an instance equipped with an RTX 4090, as the compute provider's RTX 4090 nodes are less bandwidth-constrained, even though the SOT itself did not require the compute power of an RTX 4090.

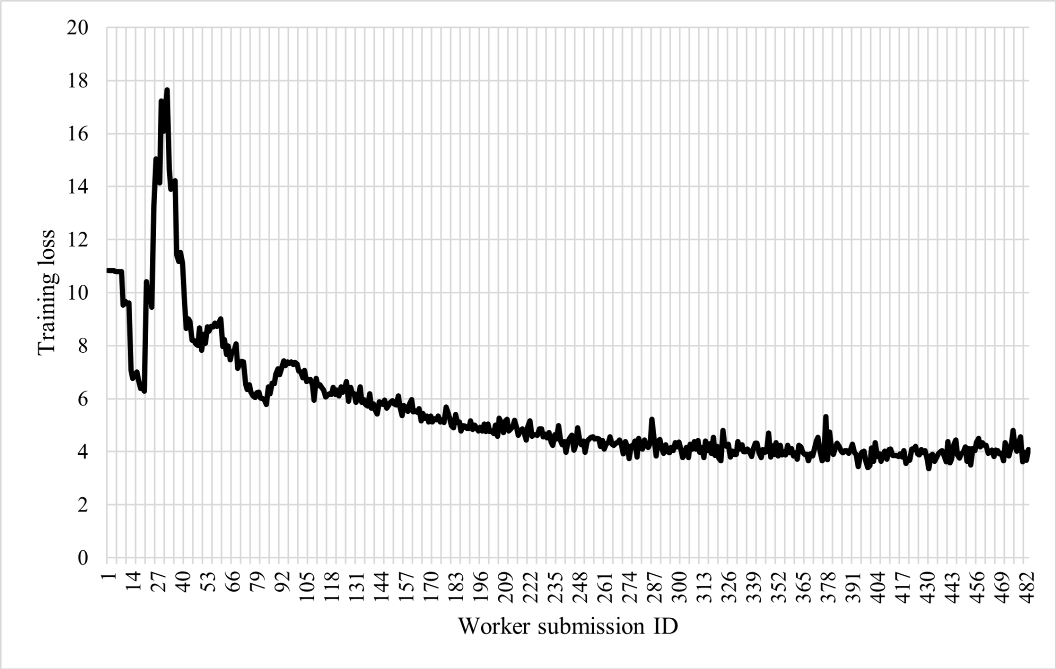


Figure 6: nanoGPT Training Loss Over Time

Although the training was interrupted after 36 hours, with the loss stagnating around 4.0, the experiment confirmed Panthalia's capability to decentralize large-scale model training. We note that we did not invoke the dispute functionality of the smart contracts in this simulation.

The input is coloured dark blue in the samples.

Sample 1:

```
\nThis is for a second chance to discover their relationship.
The mutual relationship is here and here are the essential bodies of love.
What is one way when you love someone. Let them be a friend of you, and share them.
In this blog, we will talk about some of the best relationships of romantic types.
We are on the other side of the boat, the ones that have been known since the love
    affair that resulted from each other.
For any relationship from my partner
```

Sample 2:

```
HAMLET:  ((CLAUNCH CORK)-
We have an opportunity to enhance the success of the technology industry through
    innovative innovative solutions. In today's digital age with the growing demand
     for technology, we've witnessed a range of technologies, from tech to mobile,
    to smartphones, and many others. For this, we've been harnessing the power of
    tech.
A great solution for existing and future needs is with the ability to innovate,
    innovate, and innovate in new ways.
```

Sample 3:

```
This product is made with 100% 100% free of all natural ingredients.
- Purety Coconut Oil
Each day, Honeyy Buttery Herb butter (not gluten free) contains 100% pure coconut
    and 50% fresh ginger peas.
Ingredients: Coconut, Ginger
- Coconut Oil: Stir
Ingredients: Coconut, Coconut and Lime
This product is made from 100% gluten free.
We are not responsible for any loss of flavor.
-Ingredients: Coconut Oil
- Vegan friendly
-Non
```

4 Conclusion

In this report, we evaluated the shortcomings of current compute marketplaces and introduced *Panthalia*, a verifiable compute marketplace that leverages probabilistic verification and staking mechanisms to ensure integrity. Our simulation, which successfully trained a 124-million-parameter nanoGPT model across distributed nodes, demonstrates Panthalia's potential to support large-scale machine learning tasks over decentralized networks.

Panthalia represents a significant step forward in creating reliable, decentralized compute platforms, with future work focused on advanced verification techniques, dispute mechanisms, and a more decentralized Source of Truth. As demand for compute continues to grow, Panthalia offers a promising path forward for accessing, verifying, and utilizing distributed compute resources globally.

References

Rich Sutton. The Bitter Lesson, March 2019. URL http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

roon [@tszzl]. https://t.co/1RGoLxT3Gl, July 2020. URL https://x.com/tszzl/status/1288383752067067904.

Yue Zhang and Shouqiao Wang. Proof of Sampling: A Nash Equilibrium-Secured Verification Protocol for Decentralized Systems, May 2024. URL http://arxiv.org/abs/2405.00295. arXiv:2405.00295 [cs].

Andrej Karpathy. karpathy/nanoGPT: The simplest, fastest repository for training/finetuning medium-sized GPTs., 2023. URL https://github.com/karpathy/nanoGPT.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, June 2024. URL http://arxiv.org/abs/2406.17557. arXiv:2406.17557 [cs].

Arthur Douillard, Qixuan Feng, Andrei A. Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, Marc'Aurelio Ranzato, Arthur Szlam, and Jiajun Shen. DiLoCo: Distributed Low-Communication Training of Language Models, September 2024. URL http://arxiv.org/abs/2311.08105. arXiv:2311.08105 [cs] version: 2.

A Appendix

Screenshots and Jupyter notebooks containing the benchmarking code and output can be found at `https://panthalia.com/appendix.zip`.

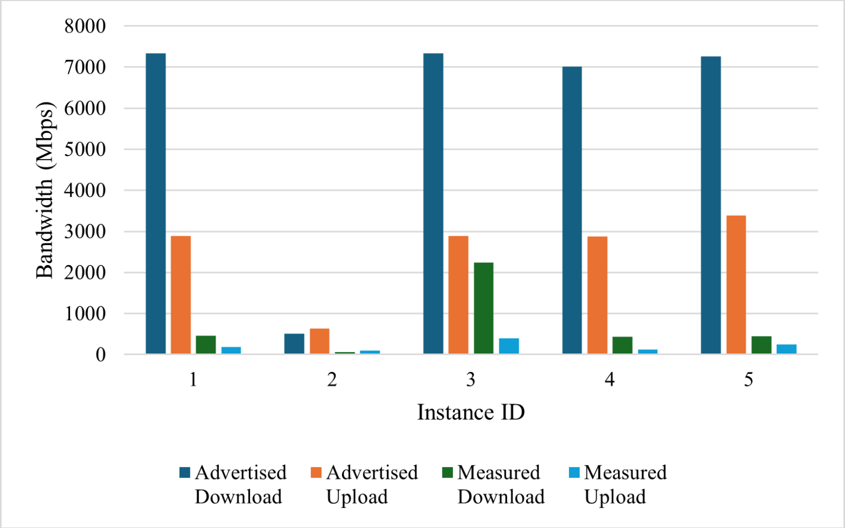


Figure 7: Advertised vs. Measured Bandwidth for Compute Provider A's Instances

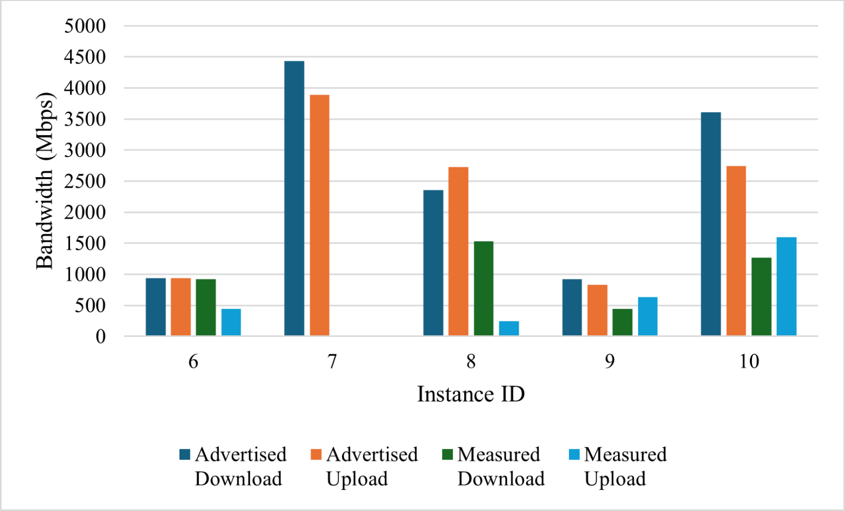


Figure 8: Advertised vs. Measured Bandwidth for Compute Provider B's Instances

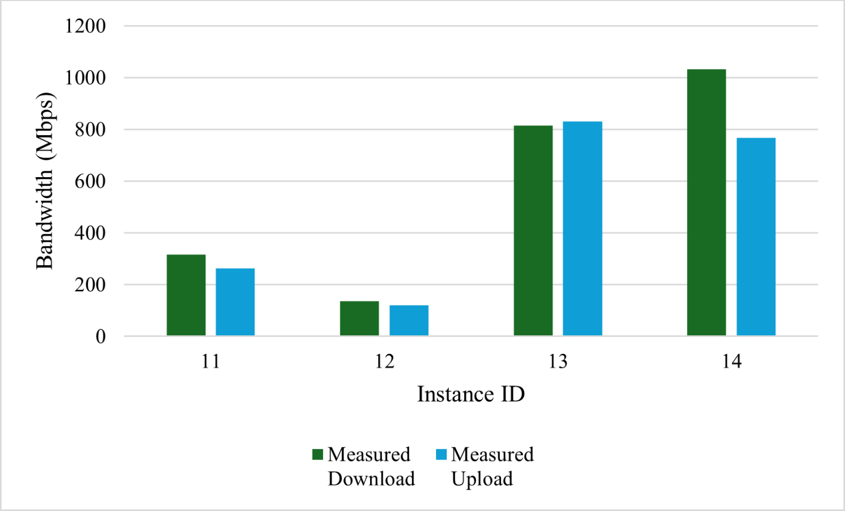


Figure 9: Measured Bandwidth for Compute Provider C's Instances

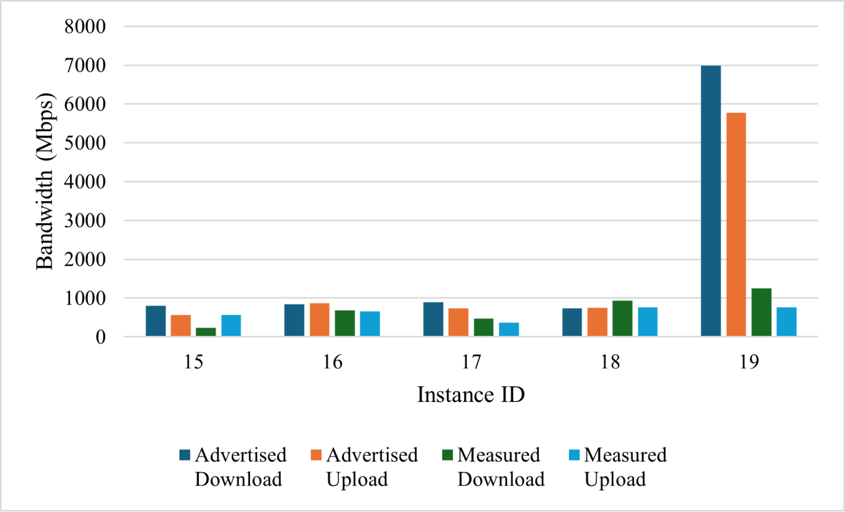


Figure 10: Advertised vs. Measured Bandwidth for Compute Provider D's Instances