HP Z8 G4 with NVIDIA GPU's or Public Cloud for heavy data science tasks

Which solution is a better investment?

Powered by

Intellerts

NVIDIA

HP Z

Data science is a computationally heavy task. As methods improve and breakthroughs are made, new state-of-the-art technologies push the boundaries of modern hardware. Keeping up with the latest advancements, implementing the newest models and testing their capabilities has become increasingly difficult.

Data science is a computationally heavy task. As methods improve and breakthroughs are made, new state-of-the-art technologies push the boundaries of modern hardware. Keeping up with the latest advancements, implementing the newest models and testing their capabilities has become increasingly difficult.
From an IT perspective, there are two main ways you can support your data scientists: invest in local hardware or use a public cloud's computation power. In this business case, we will compare using an HP Z8 G4 machine with two NVIDIA-GPUs with similar services from a major cloud provider. Our goal is to evaluate which approach is a better long-term investment: purchasing an HP Z8 G4 machine, which requires a sizable upfront investment, or using the cloud with a monthly service fee.
As a starting point for our comparison, we will run some data science experiments with a defined, realistic workload. We will calculate the costs of running the same experiment on the HP Z8 G4 machine and a comparable cloud service.

In this white paper, we will describe the data science workload used as a basis for our calculations. Next, we will outline the hardware required. First, we will describe how the workload runs on a desktop machine and then describe how the same workload runs on a comparable cloud service. Finally, we will calculate the total cost of both options to determine which approach is a better long-term investment: purchasing an HP Z8 G4 machine or paying a monthly cloud service subscription fee.

Starting from a realistic workload

For our comparison, we needed a realistic data science workload. We decided to experiment with Named Entity Recognition (NER) tasks. NER is an advanced subsection of the natural language processing (NLP) field with a myriad of fascinating applications. While the underlying models for named entity recognition are complex, the principle is not.

With NER, an algorithm extracts information from unstructured text by locating and classifying so called named entities. Here's a simple example:

""ig Business Corp. has paid 10,000 dollars in regard to an invoice sent by the Spanish division of Desktop Inc."

Using NER, an algorithm identifies "Big Business Corp." and "Desktop Inc." as company names, "10,000 dollars" as a sum of money, and "Spanish" as a location identifier. Figure 1 illustrates this principle further by applying NER to text from Wikipedia.

Using NER on a large amount of unstructured data could generate useful insights for analysts involved in auditing, investing, and identifying suspicious events in financial reports. Further analysis of the structured data generated by NER could lead to unknown insights and connections.

In a real-life scenario, data scientists would train, run, and fine-tune models to achieve actionable results for the business. This task is continuous and iterative, which is addressed in more detail further down in this white paper.

The Netherlands (GPE) (Dutch (NORP) : Nederland (GPE) ['ne:dərlant] (About this soundlisten)); informally Holland (GPE) , is a country primarily located in Western Europe (LOC) and partly in the Caribbean (LOC) . It is the largest of four (CARDINAL) constituent countries of the Kingdom of the Netherlands (GPE) . In Europe (LOC) , the Netherlands (GPE) consists of twelve (CARDINAL) provinces, bordering Germany (GPE) to the east, Belgium (GPE) to the south, and the North Sea (LOC) to the northwest, with maritime borders in the North Sea (LOC) with those countries and the United Kindom (GPE) . In the Caribbean (LOC) , it consists of three (CARDINAL) special municipalities: the islands of Bonaire (GPE) , Sint Eustatius (PERSON) and Saba (GPE) . The country's official language is Dutch (NORP) , with West Frisian (NORP) as a secondary official language in the province of Friesland (GPE) , and Englisch (LANGUAGE) and Papiamento (GPE) as secondary official languages in the Caribbean Netherlands (LOC) . Dutch (NORP) Low Saxon (PERSON) and Limburgish (NORP) are recognised regional languages (spoken in the east and southeast respectively), while Sinte Romani (PERSON) and Yiddish (GPE) are recognised non-territorial languages.

- Person
- Norp
- Org
- Gpe
- Loc
- Product
- Event
- Work of Art
- Language
- Date
- Time
- Percent
- Money
- Quantity
- Ordinal
- Cardinal

Fig.1: NER extracts information from unstructured text by classifying entities.

Setting up the test

For our test, we analysed the cost of running some popular NER models. Recent advances in natural language processing rely heavily on large pretrained models using state-of-the-art deep learning techniques. In addition, they are very resource-demanding and require multiple iterations to achieve prominent results. There are numerous advanced models available. For our experiment, we selected the following models:

- Google's BERT (Bidirectional Encoder Representations from Transformers), the original transformer model which sparked the whole transformer revolution;
- Facebook's RoBERTa, a more robust and optimised version of the original BERT model;
- ELMo model, with deep contextualised word representations, created by Allen Institute;
- Flair contextual string embeddings created by Zalando. For better performance, we used both forward and backward embeddings;
- XLNet model, which applies autoregressive pretraining and overcomes BERT in a multitude of tasks;
- XLM-Roberta, a large multilingual model based on Facebook's RoBERTa.

For training the models, we use data from Ontonotes. Ontonotes contains a massive amount of textual data from telephone conversations, newswire, newsgroups, broadcast news, broadcast conversations, and blogs. We excluded the available religious texts, as they were irrelevant for training our benchmark models.

In addition, we used data from the US Security and Exchange Commission (SEC). The documents in this dataset were highly relevant for training our financial NER model. Unfortunately, the SEC data contained only basic types of entities such as Person, Location, Organisation, or Miscellaneous. The Ontonotes data contained 20 types of entities. We also used FinBERT, a transformer model that was pretrained by Hong Kong University of Science and Technology using a large set of financial documents.

The specifics of each model are relevant to data scientists. From a business point of view however, they are all similar. Each model needs to be trained to perform NER on a set of unstructured data. Training, running and fine-tuning the models consumes a lot of system resources. We performed the experiment with different models to ensure there was no bias towards any particular hardware.

The hardware

First, we ran our experiments on local hardware. NER is a resource-heavy task, requiring a capable machine. For this experiment, we selected the HP Z8 G4 Data Science Workstation with two state-of-the-art NVIDIA RTX8000 GPUs, equipped with 48GB of VRAM each. The system has 376GB of RAM, which enabled us to use the largest versions of all the language models mentioned above. It is also relevant to note that the system is equipped with dual Intel(R) Xeon(R) Gold 6242R CPUs (20 cores each), a 4 TB HP Z Turbo data drive, and a 1,450 watt power supply, and it runs on Ubuntu 20.04.

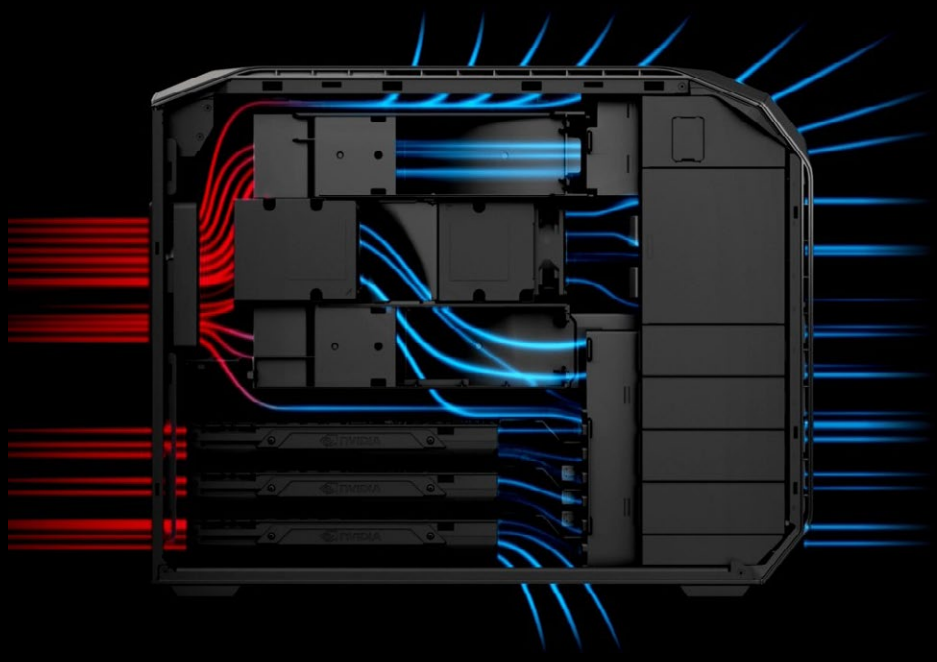


Fig.2: The HP Z8 G4 configuration used in our experiment.

NER, like other natural language applications and deep learning in general, performs most efficiently when the software models have access to hardware accelerators like the NVIDIA GPUs. Training and inference are not necessarily complex tasks in and of themselves. More specifically, a calculation for one step in the process is not very difficult. However, to run a complex model within an acceptable time frame, the hardware must complete a large number of small calculations simultaneously.

GPUs excel in the parallel execution of smaller instructions, which is what we needed in our experiment. The amount of memory available determines the size of the model and the amount of data a system can handle. The HP Z8 G4 is well equipped for complex tasks, having a sizeable amount of memory and two powerful GPUs.

Testing and tuning

During our experiment, we optimised the models as much as possible, by gradually increasing the size of the training models' data batches until some models ran out of available resources.

Generally in data science, and specifically in our NER experiment, a finite state or perfect answer is not attainable. Data scientists run an experiment, evaluate the results, and then adjust the model according to their findings. Creating an accurate and efficient model that delivers relevant results may take hundreds of attempts.

On average, running one experiment with a single model took three hours and 46 minutes. Running all models during the first iteration took 23 hours. From there, we started fine-tuning the models. After each experiment, we evaluated the model's performance and running time and iterated further. The experiment entailed a total of 20 iterations. In the end, running all the models required eight and a half hours using the HP Z8 G4's full GPU capacity.

This process of running and iterating the full set of models is our benchmark for calculating the cost of processing data on both the HP Z8 G4 and in the cloud.

To the cloud

The public cloud offers an interesting alternative to high-end data science machines. The initial costs are significantly lower. However, as previously mentioned, data science requires continuous iteration and improvement. How does the cloud compare to the desktop when evaluating the entire lifecycle of an NER project?

First, we needed a cloud service comparable to the HP Z8 G4 workstation. Amazon Web Services (AWS) offers the g4db.metal instance. A bare metal instance enabled us to run the same software and did not limit us to a specific data science service from one cloud provider.

The bare metal instance is powered by up to eight NVIDIA T4 Tensor core GPUs with 320 Turing Tensor cores and 2,560 CUDA cores. The GPUs each have 16GB of memory. We selected an instance with 384GiB of RAM. The CPU is a custom Xeon Scalable processor with up to 64 vCPUs, which is less than the HP workstation, but the difference was not significant enough to impact the experimental workload.

Generally, the HP Z8 G4 and the AWS g4dn.metal will perform on equal levels. However, there are some key differences to consider. AWS g4dn.metal has approximately double the amount of GPU cores but the GPU cores in the HP Z8 G4 operate nearly twice as fast. This means that the AWS instance can perform twice as many parallel computations, but each calculation takes twice as long. In theory, these differences would more or less cancel each other out but we did not verify the exact performance differences in a practical setting. The HP Z8 G4 is available with flexible configurations, while the AWS instance is not: users must adapt to whatever resource closest matches their needs.

Technical Capability	Type	HP Z8 G4 configuration using 2 X Quadro RTX 8000	AWS g4dn.metal configuration using 8 x Tesla T4
CUDA Parallel-Processing Cores	Total	4608 x 2 = 9216	2560 x 8 = 20480
NVIDIA Tensor Cores	Total	576 x 2 = 1152	320 x 8 = 2560
Core clock speed		1395 MHz	585 MHz
Boost clock speed		1770 MHz	1590 MHz
Number of transistors		18,600 million	13,600 million
Thermal design power (TDP)		260 Watt	70 Watt
Interface		PCIe 3.0 x16	PCIe 3.0 x16
Memory type		GDDR6	GDDR6
Maximum RAM amount	Total	48GB x 2 = 96GB	16GB x 8 = 128GB
Memory bus width		384 Bit	256 Bit
Memory clock speed		14000 MHz	10000 MHz
Memory bandwidth		672.0 GB/s	320.0 GB/s

Fig.3: Although the AWS bare metal instance and the HP Z8 G4 workstation have different configurations, their performance is similar.

It took the same amount of time to run the experiment's final iteration on the AWS instance as it did on the HP G8 Z4: eight and a half hours.

To compare costs, we referenced the cloud's on-demand pricing, because this flexible structure is one of the cloud's most appealing features.

Using the AWS g4dn.metal instance for one hour costs 9.78 dollars. Running one iteration of the experiment took more than eight hours. The total cost for this iteration in the cloud would have been 106.50 dollars (9 hours multiplied by the hourly rate, plus standard 21% VAT).

8 hours and 30 minutes

The full experiment entailed 320 hours ((23h +9h)/2 * 20 iterations) of calculations performed with the HP Z8 G4 machine. Using the AWS g4db.metal instance to do the same calculations would have cost 3,129.60 dollars excluding VAT, and 3,789.82 dollars including VAT. This amount does not include additional costs of networking, storage, or extra layers of security.

Comparison

The data science workstation used costs approximately 30,000 dollars. Operating costs also need to be considered. The HP Z8 G4 utilises 1,450 Watt of electricity per hour. During the experiment, the entire 320-hour runtime resulted in an electricity bill of 53.36 dollars in the Netherlands, which is negligible in this context. Having a desktop requires some maintenance, estimated at 300 dollars per month. With these expenses, the initial investment for running this experiment locally is 30,353 dollars.

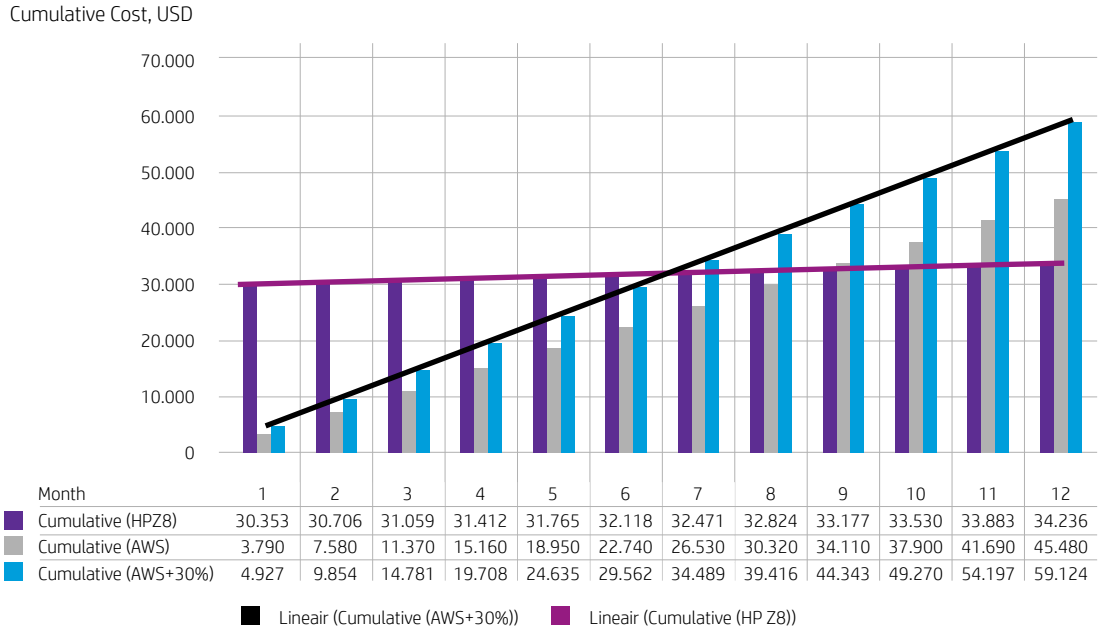
Using AWS for 320 hours costs 3,790 dollars. Including costs for other necessary services, such as storage and networking, the monthly price comes to 4,927 dollars. In practice, data scientists continually fine-tune models and iterate, meaning that our experiment might need to be run month after month, each time requiring 320 hours of calculations.

Using the HP Z8 G4 workstation, the total cost of running the experiment for seven months including the initial hardware investment, maintenance, and electricity is 32,471 dollars. Relying on the AWS g4db.metal instance for seven months would cost 34,489 dollars. These numbers demonstrate that for a dedicated data science project it would only take seven months to start seeing a return on investment from purchasing an HP Z8 G4 desktop. Beyond then, the savings only increase. Running experiments for an entire year would have cost 34,236 dollars with the HP Z8 G4. Using the cloud, the cost could easily rise to nearly 60,000 dollars.

If the workstation is used to its maximum load level for two years, the total cost of ownership would not exceed 40,000 dollars (38,472 dollars to be precise). Using the cloud for a similar workload and amount of time would cost 118,248 dollars.

	Using HP Z8 G4 machine				Using AWS g4dn.metal On-Demand Instance		Using AWS g4dn.metal On-Demand Instance (+ other services)	
Month	Cost of Machine (HPZ8)	Maintenance (HPZ8)	Electricity Costs (HPZ8)	Cumulative (HPZ8)	Monthly Price (AWS)	Cumulative (AWS)	Monthly Price (AWS+30%)	Cumulative (AWS+30%)
1	30,000	300	53	30,353	3,790	3,790	4,927	4,927
2		300	53	30,706	3,790	7,580	4,927	9,854
3		300	53	31,059	3,790	11,370	4,927	14,781
4		300	53	31,412	3,790	15,160	4,927	19,708
5		300	53	31,765	3,790	18,950	4,927	24,635
6		300	53	32,118	3,790	22,740	4,927	29,562
7		300	53	32,471	3,790	26,530	4,927	34,489
8		300	53	32,824	3,790	30,320	4,927	39,416
9		300	53	33,177	3,790	34,110	4,927	44,343
10		300	53	33,530	3,790	37,900	4,927	49,270
11		300	53	33,883	3,790	41,690	4,927	54,197
12		300	53	34,236	3,790	45,480	4,927	59,124
13		300	53	34,589	3,790	49,270	4,927	64,051
14		300	53	34,942	3,790	53,060	4,927	68,978
15		300	53	35,295	3,790	56,850	4,927	73,905
16		300	53	35,648	3,790	60,640	4,927	78,832
17		300	53	36,001	3,790	64,430	4,927	83,759
18		300	53	36,354	3,790	68,220	4,927	88,686
19		300	53	36,707	3,790	72,010	4,927	93,613
20		300	53	37,060	3,790	75,800	4,927	98,540
21		300	53	37,413	3,790	79,590	4,927	103,467
22		300	53	37,766	3,790	83,380	4,927	108,394
23		300	53	38,119	3,790	87,170	4,927	113,321
24		300	53	38,472	3,790	90,960	4,927	118,248

Fig.4: With the HP Z8 G4 workstation, businesses start saving money after seven months compared to using similar cloud services. Savings increase from then on.



Month	1	2	3	4	5	6	7	8	9	10	11	12
Cumulative (HPZ8)	30.353	30.706	31.059	31.412	31.765	32.118	32.471	32.824	33.177	33.530	33.883	34.236
Cumulative (AWS)	3.790	7.580	11.370	15.160	18.950	22.740	26.530	30.320	34.110	37.900	41.690	45.480
Cumulative (AWS+30%)	4.927	9.854	14.781	19.708	24.635	29.562	34.489	39.416	44.343	49.270	54.197	59.124

Lineair (Cumulative (AWS+30%)) Lineair (Cumulative (HP Z8))

Fig.5: The initial investment in a data science machine pays off after approximately seven months, depending on usage.

Conclusion

As our experiment shows, businesses planning to run data science tasks on a regular basis will save money over time by purchasing the HP Z8 G4. The more the machine is used, the faster savings are realised compared to using similar cloud services. Our experiment used a 320-hour runtime each month as a baseline. There are certainly possibilities for further optimising the usage of the machine in daily operations.

In less than a year, companies will save money by purchasing the HP Z8 G4 compared to using cloud services. In two years, the total cost of comparable cloud services will be nearly 120,000 dollars. This amount will buy you four HP Z8 G4 machines delivering four times the performance.

Contact

For more information about HP, HP products and services, and support, please visit:

Workstationspecialist.nl