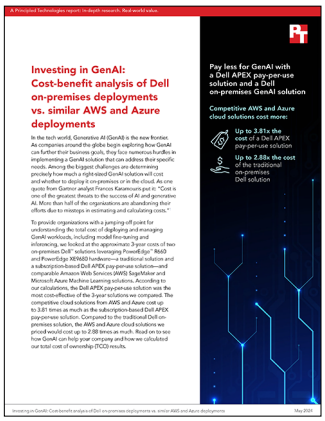

The science behind the report:

Investing in GenAI: Cost-benefit analysis of Dell on-premises deployments vs. similar AWS and Azure deployments

This document describes the details of our research. To learn how these facts translate into real-world benefits, read the report Investing in GenAI: Cost-benefit analysis of Dell on-premises deployments vs. similar AWS and Azure deployments.

We concluded our research on March 29, 2024. The results in this report reflect configurations that we finalized and acquired pricing data for on March 29, 2024 or earlier. Unavoidably, these configurations and their costs may not be up to date when this report appears.

System information

Dell on-premises solutions

The traditional and managed on-premises Dell solutions both include the following hardware:

- 3 x PowerEdge R660 head node
- 2 x PowerEdge XE9680 GPU worker node
- 2 x PowerSwitch S5232-ON network infrastructure
- 1 x PowerSwitch N3200-ON OOB management

Table 1: : Detailed configuration information for each PowerEdge XE9680 GPU worker node.

Configuration information	Dell PowerEdge XE9680 GPU worker node
Number of nodes in solution	2
Chassis	
Chassis	XE9680 6U Chassis with 8 GPU 8x 2.5 NVMe Only
Processor	
Number of processors	2
Vendor and model	Intel® Xeon® Platinum 8468
Core count (per processor)	48 cores and 96 threads
GPU(s)	
Number of GPUs	1 8-GPU Assembly
Vendor and model	NVIDIA® HGX H100 8-GPU SXM 80GB 700W GPUs Assembly

Configuration information	Dell PowerEdge XE9680 GPU worker node
Memory module(s)	
Total memory in system (GB)	1,024
Number of memory modules	16
Type	RDIMM, 4800MT/s Dual Rank
Size (GB)	64
Storage controller	
Vendor and model	BOSS-N1 controller card + with 2 M.2 480GB (RAID 1)
Local storage (type A)	
Total size of drives in system (TB)	44.8
Number of drives	7
Drive size (TB)	6.4
Drive information (speed, interface, type)	Enterprise NVMe™ Mixed Use AG Drive U.2 Gen4 with Carrier
NIC	
Number and type of ports	2 x 10/25GbE
Vendor and model	Intel E810-XXV Dual Port 10/25GbE SFP28, OCP NIC 3.0
Network adapter	
Number and type of ports	2 x 10/100GbE
Vendor and model	Mellanox ConnectX-6 DX Dual Port 100GbE QSFP56 Network Adapter, Full Height
Cooling fans	
Number of cooling fans	6
Vendor and model	Very High Performance Fan
Power supplies	
Number of power supplies	1
Vendor and model	Fully Redundant 5 + 1 (or 3+3 FTR), Hot-Plug PSU, 2800W MM HLAC (200-240Vac) Titanium, C22 Connector
Wattage of each (W)	2800
ProSupport and ProDeploy	
ProSupport (5 years)	ProSupport and Next Business Day Onsite Service
ProDeploy Plus	ProDeploy Plus Dell Server XE Series 5U/6U
Embedded system management	
iDRAC9	iDRAC9, Datacenter 16G
OpenManage	OpenManage™ Enterprise Advanced Plus

Table 2: Detailed configuration information for each PowerEdge R660 head node.

Configuration information	Dell PowerEdge R660 head node
Number of nodes in solution	3
Chassis	
Chassis	2.5" chassis with up to 10 hard drives (SAS/SATA), PERC11, 1CPU
Processor	
Number of processors	1
Vendor and model	Intel Xeon Gold 6426Y
Core count (per processor)	16 cores and 32 threads
Memory module(s)	
Total memory in system (GB)	192
Number of memory modules	12
Type	16
Size (GB)	RDIMM, 4800MT/s Single Rank
Storage controller	
Vendor and model	BOSS-N1 controller card + with 2SED M.2 960GB (RAID 1)
Local storage (type A)	
Total size of drives in system (TB)	5.76
Number of drives	6
Drive size (TB)	960
Drive information (speed, interface, type)	vSAS Read Intensive SSD 12Gbps 512e 2.5in Hot-Plug ,AG Drive SED, 1DWPD
NIC	
Number and type of ports	2 x 10/25GbE
Vendor and model	NVIDIA ConnectX-6 Lx Dual Port 10/25GbE SFP28, No Crypto,OCP NIC 3.0
Cooling fans	
Number of cooling fans	4
Vendor and model	Very High Performance Fan
Power supplies	
Number of power supplies	1
Vendor and model	Dual, Redundant(1+1), Hot-Plug Power Supply,1100W MM(100-240Vac) Titanium
Wattage of each (W)	1,100
ProSupport and ProDeploy	
ProSupport (5 years)	ProSupport and Next Business Day Onsite Service
ProDeploy Plus	ProDeploy Plus PowerEdge R Series 1u2u
Embedded system management	
iDRAC9	iDRAC9, Datacenter 16G
OpenManage	OpenManage Enterprise Advanced Plus

AWS SageMaker solution instances

Table 3: Detailed configuration information for the AWS instances.

Configuration information	ml.t3.medium (notebooks)	ml.r5.16xlarge (processing)	ml.p5.48xlarge (inference and fine-tuning)
Number instances	20	2	2
Cloud service provider (CSP)	AWS	AWS	AWS
Region	US East (Ohio)	US East (Ohio)	US East (Ohio)
Processor			
Number of vCPU	2	64	192
Memory module(s)			
Total memory in system (GiB)	4	512	2,048
Storage controller			
Vendor and model			
Local storage (type A)			
Number of drives	1	1	1
Drive size (GB)	5GB	Default	Default
Drive information (speed, interface, type)	EBS	EBS	EBS
GPU			
Number of GPUs	N/A	N/A	8
Vendor and Model	N/A	N/A	NVIDIA H100
Additional features	N/A	N/A	3,200 Gbps of networking bandwidth[1]

Azure Machine Learning solution instances

Table 4: Detailed configuration information for the Azure instances.

Configuration information	D2 v2 (notebooks)	M64 (processing)	ND96amsr A100 v4 (inference and fine-tuning)
Number instances	20	1	8
Cloud service provider (CSP)	Azure	Azure	Azure
Region	East US 2	East US 2	East US 2
Processor			
Number of vCPU	2	64	96
Memory module(s)			
Total memory in system (GiB)	7	1,000	1,900
Local storage (type A)			
Number of drives	1	1	1
Drive size (GB)	100	7,168	6,400
Drive information (speed, interface, type)	Temporary	Temporary	Temporary

Configuration information	D2 v2 (notebooks)	M64 (processing)	ND96amsr A100 v4 (inference and fine-tuning)
Number of GPUs	N/A	N/A	8
Vendor and model	N/A	N/A	NVIDIA A100
Additional information	N/A	N/A	Each GPU has a dedicated 200 GB/s NVIDIA Mellanox HDR InfiniBand connection[2]

Introduction

To provide an example for AI solution costs, we created an AI scenario using the open-source Llama 2 13B model and compared the cost to run the workload in four different environments. We sized and estimated the costs for four solutions:

- Traditional on-premises Dell solution
- Managed on-premises solution with a Dell APEX pay-per-use solution
- AWS SageMaker solution
- Microsoft Azure Machine Learning solution

Both the traditional and managed on-premises Dell solutions use the same hardware. With the traditional plan, the enterprise purchases the hardware upfront; with the Dell APEX pay-per-use solution, Dell installs hardware in the customer's data center and bills the enterprise monthly based on "Committed" and "Buffer Capacity."

For this analysis, we tried to create a broadly applicable example scenario to estimate cost differences across environments. We chose the Llama 2 13B GenAI model because it is a widely available, open-source model, and we built our scenario around a single, relatively small AI workload. We included costs for data scientists' machine learning development notebooks, data processing tasks, continuous model fine-tuning, and real-time inference.

We sized the hardware for each solution based on assumptions about hours of work and hardware capabilities needed. We used public sources for that research. We used online pricing calculators for AWS SageMaker and Azure Machine learning and requested and received quotes from Dell Technologies for the costs using Dell Recommended Pricing for the two Dell solutions. We did not do any hands-on testing of any of the solutions for this paper.

Our findings

In the main report, we show the normalized comparisons each of the two on-premises Dell solutions compared to the AWS SageMaker and Azure Machine Learning cloud solutions. Tables 5 and 6 show the cost basis for those normalizations. The normalized value is the result of dividing each value by the cost for the Dell on-premises solution shown in the table. Table 5 shows that cloud solutions cost up to 2.88x the Dell traditional on-premises solution over three years. The breakeven row shows that 3 years of the traditional on-premises Dell solution costs roughly similar to the cost of just one year of the two cloud solutions.

Table 5: Normalized three-year TCO for Dell traditional on-premises solution compared to SageMaker and Azure Machine Learning solutions.

	Dell traditional on-premises 3-year costs	AWS SageMaker 3- year commitment	Azure Machine Learning 3-year commitment
Total	$817,880.00	$2,357,549.00	$2,231,805.00
Normalized	1	2.88	2.72
Breakeven in months for Dell solution compared to cloud solutions	N/A	12.5	13.3

Table 6 shows that the cloud solutions cost up to 3.81x the cost of a Dell APEX pay-per-use solution over 3 years and that 3 years of the Dell traditional on-premises solution costs less than the cost of just 1 year of the two cloud solutions.

Table 6: Normalized 3-year TCO for a Dell APEX pay-per-use solution compared to SageMaker and Azure Machine Learning solutions.

	Dell APEX pay-per-use solution 3-year costs	AWS SageMaker 3-year commitment	Azure Machine Learning 3-year commitment
Total	$618,648.00	$2,357,549.00	$2,231,805.00
Normalized	1	3.81	3.60
Breakeven in months for Dell solution compared to cloud solutions	N/A	9.5	10

Storage considerations

We did not include costs for storage beyond that which is needed for the servers or instances to do their tasks.

The Dell on-premises solutions include 106.88TB storage:

- 5.76TB of SSD capacity on each of the three PowerEdge R660 head nodes
- 44.8TB of SSD capacity on each of the two PowerEdge XE9680 GPU worker nodes

Cluster management and notebook tasks share the storage on the head nodes. Processing, model fine-tuning, and inferencing tasks share the storage on the GPU worker nodes. We provisioned the Dell PowerEdge clusters with some additional storage relative to the cloud solutions to ensure room for management tasks on the head nodes and some room for growth if needed.

The AWS SageMaker solution includes a total of 63.444TB storage:

- 7,000 GB EBS gp2 storage purchased for each of the two ml.r5.16xlarge processing instances
- 8 x 3084 GB NVMe SSDs for each of the ml.p5.48xlarge instances
- 1 x 5GB EBS temporary storage for each notebook instance included with the instance

The Azure Machine Learning solution included a total of 64.82 TB temporary storage:

- 7168GiB temporary storage for the M64 processing instance
- 6400GiB temporary storage for each of the eight ND96amsr A100 v4 instances
- 1 x 100GiB temporary storage for each notebook instance included with the instance

Storage needs vary for notebook instances, but typically do not require much for this type of workload, so we opted to leave the cloud instances with the storage that came by default. We included data transfer costs for the EBS data transfer in the AWS SageMaker and for Blob storage transfer in the Azure Machine learning solutions. We did not include data transfer costs for the Dell on-premises solutions, which would be using on-board SSDs.

Usage hours

We sized the solutions based on the following estimates of hours per month of notebook, data processing, model fine-tuning, and inference usage. We used these hours to calculate hours of instance usage for the cloud solutions and to size those instances and the servers for the on-premises solutions.

We sized the solutions with the assumption that there are 22 workdays in each month, with workloads set to run overnight to maximize usage. Thus, each server and cloud instance would have 528 hours of runtime available each month. (See Table 7.)

Table 7: Usage hours for the four tasks.

Task	Total hours per month	Usage calculations
Notebook	3,520	We sized each solution to support 20 data professionals, with one notebook instance each, 8 hours a day, 22 days a month, a total of 176 hours each a month, a total of 3,520 for all 20 data professionals. Minimum requirements are small cloud notebook instances with 2vCPU and at least 4GiB memory.
Processing	1,056	Data processing tasks would run during the 528 uptime hours on two Dell PowerEdge XE9680 servers for a total of 1,056 hours runtime and would require 1,056 hours runtime on cloud instances. Minimum requirements for the cloud instances were 64vCPU, 1000GiB memory, 7000 GB storage. We sized the Dell PowerEdge servers to support these requirements plus those of the fine-tuning and inferencing tasks.
Model fine-tuning	792	The combined runtime of 1,056 hours for fine-tuning and inferencing tasks requires two Dell PowerEdge XE9680 servers. All jobs use eight H100 GPUs and up to a half TB of memory. The cloud solutions require the same number of hours on instances with eight H100 GPUs or equivalent. For AWS we used two H100 instances; for Azure Machine learning we substituted eight A100 instances.
Inferencing	264	

Notebooks details

Data scientists would need small cloud notebook instances with 2vCPU and at least 4GiB memory. While some notebook tasks might perform better with more memory, we opted for the AWS ml.t3.medium instance based on the AWS SageMaker TCO guide[3] suggestions, and then chose a similar sized instance for Azure. For Dell on-premises solutions, we assumed the one core of an equivalent processor and 4.3GB memory per notebook with the notebooks running on the 3x PowerEdge R660 management servers along with management tasks.

AWS SageMaker and Azure Machine Learning notebook instances

For the AWS SageMaker and Azure Machine Learning solutions, we selected notebook instances that had 2vCPU and at least 4GiB memory.

Table 8: Key configuration information for the AWS SageMaker and Azure Machine Learning notebook instances.

Instance type	Instance	Number of vCPU per instance	Memory (GiB) per instance
SageMaker notebooks	ml.t3.medium	2	4
Azure ML notebooks	D2 v2	2	7

Dell on-premises solutions notebooks

The Dell solutions support these notebook workloads on the three PowerEdge R660 head nodes, which combined have 576GB memory and 48 processor cores, enough to support both cluster management tasks and the 20 notebook workloads. Our assumptions for making that sizing decision are as follows:

- Management tasks take up less than half of the processor capacity of these head nodes and less than 500GB memory, with remaining capacity available to run these tasks.
- During the 176 hours each notebook workload runs each month, it would use a single processor core, the equivalent of the 2 vCPU for the SageMaker and Azure ML notebooks, assuming a 1 thread: 1vCPU ratio, and 4.3 GB memory to match the 4GiB we defined in sizing the cloud notebooks. All 20 notebooks running at the same time would use less than half of the 48 cores and 15% of the memory on these systems.
- All work on all systems occurs during less than 72.3% of the total hours in each month, based 22 workdays a week with 24 hours available each day.

Processing details

AWS SageMaker and Azure Machine Learning processing instances

Processing tasks run best on CPU rather than GPU and thrive on a high memory to core ratio,[4] so we focused on memory-optimized instances for the AWS SageMaker and Azure Machine Learning processing instances. Our target was at least 64vCPU, 1,000GiB memory, and 7,000 GB storage, which the Azure Machine Learning M64 processing instance closely matched with 64vCPU, 1,000GiB memory and 7,168GiB of temporary drive space. AWS SageMaker memory-optimized processing instances didn't offer an instance matching our specification, so instead we selected a pair of ml.r5.16xlarge SageMaker memory-optimized processing instances with a combined 1,024GiB of memory. Together these SageMaker instances exceed our needs with double the vCPU capacity and almost twice the storage capacity (with EBS storage) of our targets and of the Azure Machine Learning instance.

Table 9: Key configuration information for the AWS SageMaker and Azure Machine Learning processing instances.

Instance configuration information	SageMaker ml.r5.16xlarge (processing)	Azure Machine Learning M64 (processing)
Number instances	2	1
Cloud service provider (CSP)	AWS	Azure
Number of vCPU	64 (128 for 2 instances)	64
Total memory in system (GiB)	512 (1,024 for two instances)	1,000
Number of drives	1 (2 for two instances)	1
Drive size (GiB)	7,000 (14,000 for two instances)	7,168
Drive information (speed, interface, type)	EBS	Temporary

Dell on-premises solutions

The two Dell PowerEdge XE9680 worker nodes have more capacity than the fine-tuning and inference workloads require, enough to support the processing workloads running in parallel. The processing workloads would rely on CPUs, and the model fine-tuning and inference on GPUs. To match our target specs, the processing workloads would use half of the combined 2TB of memory of the two servers, 32CPU cores and about 7TB of the 44.8TB of storage of the two servers.

Model fine-tuning and inference details

The solutions require GPU instances or servers for fine-tuning and inference workloads with eight NVIDIA HGX H100 GPUs or equivalent GPU capacity and 512GB of memory.

AWS SageMaker and Azure Machine Learning instances

The ml.p5.48xlarge inferencing instance is the only SageMaker instance that has the NVIDIA HGX H100 GPUs.[5] At the time of our study, the best Azure VM had eight NVIDIA A100 GPU VMs, so we set the Azure environment to run four times the number of instance hours to meet roughly the same performance as the H100s in the AWS SageMaker environment.[6] Both instance types have more than our 512GB total memory minimum requirement.

Table 10: : Key configuration information for the AWS SageMaker and Azure Machine Learning fine-tuning and inference instances.

Inference and training instances	SageMaker ml.p5.48xlarge	Azure Machine Learning ND96amsr A100 v4
Number of instances	2 (1 for fine-tuning and 1 for inference)	8 (4 for fine-tuning and 4 for inference)
Cloud service provider (CSP)	AWS	Azure
Number of vCPU	192	96
Total memory in system (GiB)	2,048	1,900
Number of drives	8	1
Drive size (GiB)	3,084	6,400
Drive information (speed, interface, type)	NVMe SSD	Temporary
Number of GPUs	8	8
Vendor and Model	NVIDIA H100	NVIDIA A100

Dell on-premises solutions

For the Dell on-premises solutions, we sized the two PowerEdge XE9680 GPU worker nodes to handle the fine-tuning and inference workloads using GPU resources and to support the processing workloads using spare CPU and memory resources. The two PowerEdge XE9680 GPU worker nodes each have two 48-core Intel Xeon Platinum 8468 processors, 1,024GB memory, 44.8 TB of storage, and an NVIDIA HGX H100 8-GPU assembly.

Cost analysis

For the cloud solutions, we included the licensing cost for the instances we needed for notebooks, processing, fine-tuning, and inference. For the Dell on-premises solutions we include two payment options for the hardware: an upfront-purchase model and a Dell APEX pay-per-use solution monthly payment plan. For both on-premises solutions, we added server administration costs for the hardware and OS, and data center costs for rack space and energy costs for power and cooling, costs that aren't relevant to the two cloud solutions.

We omitted some costs, for example:

- We omitted costs of work that could be similar on all four solutions such as installing and maintaining open-source software, data transferring and backup, and the salaries of the data scientists.
- We did not include software costs for any of the solutions. SageMaker and Azure Machine Learning instances include some software and services such as Jupyter Notebooks on the notebook instances and processing APIs with the processing instances. We assumed any additional software and tools the data scientists install there would be open source. With the Dell on-premises solutions, data scientists would exclusively use open-source software and tools such as Jupyter Notebook, Python, and PyTorch, and the servers would run Ubuntu or another open-source OS.

- We did not include sales taxes because those vary state to state and business to business. We do not include migration costs or end-of-life costs.

We focused on a 3-year lifecycle for each of these generative AI solutions. We chose 3 years because AWS and Azure pricing calculators capped commitments at three years. Three years is also a reasonable lifecycle for an on-premises generative AI solution that requires state-of-the-art hardware.[7] Organizations could re-purpose the Dell hardware they purchased after that, and would have 2 remaining years of the included 5-years of Dell ProSupport services to help keep it productive.

3-year costs for Dell on-premises solutions

For the Dell on-premises solutions, we included the following costs over a 3-year period:

- The Dell Recommended Price for Dell servers and switches, including ProSupport and Next Day Onsite Service and ProDeploy Plus for the servers
- System administrator to maintain and secure the hardware and OS
- Energy costs for power and cooling
- Data center costs for rack space

The two solutions included the same hardware and would incur the same costs for system administration, energy costs for power and cooling, and data center rack space costs.

Table 11: 3-year costs for traditional on-premises solution.

Dell traditional on-premises	3-year costs (rounded up to dollar)
Dell hardware with 5-year ProSupport and ProDeploy Plus (for servers)	$680,251
System administration	$6,531
Energy costs for power and cooling	$88,978
Data center costs for rack space	$42,120
Total	$817,880

For the traditional on-premises solution, we requested a quote from Dell Technologies Sales for the recommended price of the on-premises Dell solution. On March 20, 2024, Dell quoted $680,251 as the purchase price for the hardware.

The quote included 5-year ProSupport and Next Business Day Onsite Service for the servers and switches and ProDeploy Plus for the servers. Dell defines recommended price as a price that serves as a starting point for potential buyers. It represents the cost immediately accessible to companies, even if they are not existing customers, and essentially functions as a suggested retail price for their products.

Table 12: 3-year costs for a Dell APEX pay-per-use solution.

Dell APEX pay-per-use solution	3-year costs (rounded up to dollar)
Dell hardware with 5-year ProSupport and ProDeploy Plus (for servers)	$481,019
System administration	$6,531
Energy costs for power and cooling	$88,978
Data center rack space costs	$42,120
Total	$618,648

On-premises Dell APEX pay-per-use solution

Based on that recommended price, Dell Technologies provided a 3-year cost estimate for a Dell APEX pay-per-use solution of $481,019. Dell sent PT a quote for the Dell APEX pay-per-use solution for the same hardware as above for a 36-month term commitment, and a 75% capacity commitment. We received that quote on April 2, 2024. The 75% capacity estimate was the closest capacity option that would cover the 528 uptime hours that we size the solutions to deliver.

System administration

Server administrators monitor and ensure performance, availability, functionality, and security of the hardware and OS, and in this case install the OS. These are services that the on-premises solution requires but the cloud solutions do not because they are included in their service agreement. We kept those time and cost estimates low for the on-premises solution because we assume they are mostly automated and because ProSupport for Infrastructure offloads some of the monitoring tasks and the physical repair tasks from the on-premises system administrators.

We estimated a three-year cost of $6,531.00 for this server administration based on the total compensation for a mid-level system administrator[8] who is able to maintain 300 servers and associated switches and OSs using automated tools and processes and who is aided by ProSupport and ProDeploy Plus services.

Dell ProDeploy Plus for Infrastructure

We did not include a separate estimate for deployment, instead relying on Dell ProDeploy Plus for Infrastructure, a service we included in the hardware quote, to provide onsite hardware and software deployment.[9] A Principled Technologies report shows that ProDeploy Plus for Infrastructure can "Save valuable in-house admin time by using a Dell Technologies-certified engineer for installation and configuration of a Dell."[10]

That service might not cover some planning tasks, unboxing and racking the switches, or installing the OS; those tasks would take little time, and are included in the estimate of system administration time to maintain and secure the solution.

Dell ProSupport for Infrastructure

We included 5 years of Dell ProSupport and Next Day Onsite Service. The 5-year support is longer than the 3-year time-period of our analysis but gives the purchasing enterprise the added value of a longer lifecycle for the Dell hardware.

Energy costs for power and cooling

The cloud solutions included energy costs for power and cooling in their prices. For the on-premises solutions, we estimated 3-year power and cooling costs using the Dell Enterprise Infrastructure Planning tool.[11] To get an estimate, we entered in the specifications for the servers and switches included in the Dell Technologies Sales price quote. We provided two other inputs that affected calculations:

- 1.58 power usage effectiveness (PUE) multiplier of power costs to get combined power and cooling costs. That PUE was the industry average in 2023, according to the Uptime Institute, an organization that surveys and tracks data center costs.[12]
- 12.74 cents per kilowatt-hour energy cost based on US Energy Administration (EIA) reported average retail price of electricity for the commercial sector in 2023.[13]

We calculated costs for power and cooling separately for the devices running idle and computational workloads. We weighed the results based on the 528 runtime hours (about 72.3 percent of an average month) that we sized the solutions to deliver.

Table 13: 3-year energy cost for power and cooling.[14]

Workloads	Energy cost for 3 years for on-premises solution	Weighting multiplier	Weighted energy cost for power and cooling
Computational	$114,439.69	72.3%	$82,739.90
Idle	$22,516.69	27.7%	$6,237.12
3-year weighted energy cost for power and cooling			$88,977.02

Data center rack costs

The enterprise would incur additional cost for housing the servers and racks in the data center. These costs include the costs of the rack, networking, and other data center facility costs. We estimated those costs at $1,800 per rack per month for racks with a usable capacity of 28u.[15] The servers and switches in this solution fill 18u, 65 percent of one of those racks. The three-year cost for that usage is $42,120.

While we included data transfer for the two cloud solutions to access the AWS S3 storage and the Azure Block Storage, we did not add those costs for the on-premises solutions because they would be using their onboard disks or local storage arrays for storage.

AWS SageMaker solution

We configured the AWS SageMaker solution to match the quoted on-premises Dell solution as closely as possible for the four tasks we outlined previously: notebooks, processing, model fine-tuning, and inference. The AWS Pricing Calculator for SageMaker lists each task as a separate pricing module you can toggle to add to the estimate. We added SageMaker Studio Notebooks, SageMaker Processing, SageMaker Training, and SageMaker Real-Time Inference. For each module, we filled in the necessary fields to determine the hourly cost of each instance we chose for each task. We then used that hourly cost to determine how much a user would spend to run each instance for the pre-calculated number of hours we determined based on the Dell systems. We calculated twice as many processing instances and therefore processing hours to ensure to match the processing capacity of the other solutions. We also added EBS storage to the processing instances, as they do not spin up with storage outside the OS volume. After applying the all upfront 3-year commitment savings plan, our total costs came to $2,357,549 for 3 years. See Table 14 for the full instance details.

Table 14: 3-year SageMaker solution instance costs.

Service	Instance type	Instance $/hr	Run time (hours/mo)	Cost for 3 years
SageMaker Studio Notebooks	ml.t3.medium	$0.02244	3520	$2,843.60
SageMaker Processing	ml.r5.16xlarge*	$2.2908	2112	$174,174.11
SageMaker Processing EBS storage (7TB a month)				$25,804.80
SageMaker Training	ml.p5.48xlarge**	$56.5340	792	$1,611,897.41
SageMaker Real-Time Inference	ml.p5.48xlarge**	$56.5340	264	$537,299.14
S3 data transfer (1 in and 15 out)				$5,529.60
Total (rounded up)				$2,357,549

*We include to processor instances to get 1TB memory for the processing tasks.
**Prices estimated based on public P5 (non ml.p5) commitment percentage because calculator did not include this instance.

Azure Machine Learning solution

We used the Azure Pricing Calculator to plug in each instance type to determine the hourly cost of each in the Machine Learning service. Since the best GPU instance Azure offered at the time of this study included eight A100 GPUs, we calculated the costs of running four times the number of those GPU instances per task to provide a closer approximation of the performance the eight H100 GPUs that the AWS instances and Dell PowerEdge XE9 servers provide. We then used that hourly cost to determine how much a user would spend to run each instance for the pre-calculated number of hours we determined based on the Dell systems. After applying the Azure 3-year reserved commitment pricing, our total costs for 3years came to $2,231,805. See Table 15 for full instance details.

Table 15: 3-year Azure Machine Learning solution instance costs.

Service	Instance type	Instance $/hr	Run time (hours/mo)	Cost for 3 years
Azure ML Notebooks	D2 v2	$0.0476	3,520	$6,027.72
Azure ML Processing	M64	$1.8258	1,056	$69,407.81
Azure ML Training	ND96amsr A100 v4	$14.1475	3,168	$1,613,491.01
Azure ML Real-Time Inference	ND96amsr A100 v4	$14.1475	1,056	$537,830.34
Azure Block Blob Storage data transfer operations (10,000,000)				$5,047.20
Total (rounded up)				$2,231,805

1. AWS, "Get started with P5 instances," accessed April 29, 2024, https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/p5-instances-started.html.
2. Microsoft Build," NDm A100 v4-series," accessed April 29, 2024, https://learn.microsoft.com/en-us/azure/virtual-machines/ndm-a100-v4-series
3. Amazon, "Total Cost of Ownership of Amazon SageMaker," accessed April 29, 2024, https://pages.awscloud.com/rs/112-TZM-766/images/Amazon_SageMaker_TCO_uf.pdf.
4. StackOverflow, "Why should preprocessing be done on CPU rather than GPU?" accessed April 20, 2024, https://stackoverflow.com/questions/44377554/why-should-preprocessing-be-done-on-cpu-rather-than-gpu and Hugging Face, "Model Memory Requirements," accessed April 29, 2024, https://huggingface.co/NousResearch/Llama-2-70b-hf/discussions/2.
5. AWS, "New – Amazon EC2 P5 Instances Powered by NVIDIA H100 Tensor Core GPUs for Accelerating Generative AI and HPC Applications," accessed April 29, 2024, https://aws.amazon.com/blogs/aws/new-amazon-ec2-p5-instances-powered-by-nvidia-h100-tensor-core-gpus-for-accelerating-generative-ai-and-hpc-applications/.
6. Comet, "Comparison of NVIDIA A100, H100 + H200 GPUs," accessed April 29, 2024, https://www.comet.com/site/blog/comparison-of-nvidia-a100-h100-and-h200-gpus/.
7. The Jerusalem Post, "Maximizing Efficiency: Your 2023 Guide to GPU Servers," accessed April 8, 2024, https://www.jpost.com/insights/article-770858.
8. Systems Administrator II total compensation (salary and benefits) of $130,616 per year. Source: Salary.com, "Systems Administrator II," accessed March 25, 2024, https://www.salary.com/tools/salary-calculator/systems-administrator-ii-benefits.
9. Dell, "The market's most complete deployment offer," accessed March 28, 2024, https://www.delltechnologies.com/asset/en-us/services/deployment/briefs-summaries/prodeploy_plus_deployment_unification_ds.pdf.
10. Principled Technologies, "Using Dell ProDeploy Plus for Infrastructure can improve deployment times for Dell technology," accessed March 28, 2024, https://www.delltechnologies.com/asset/en-us/products/cross-company/industry-market/principled-technologies-prodeploy-plus-for-infrastructure-services-whitepaper.pdf.
11. Dell, "Dell Enterprise Infrastructure Planning Tool," accessed March 24, 2024, https://dell-ui-eipt.azurewebsites.net/#/
12. Uptime Institute, "Large data centers are mostly more efficient, analysis confirms," accessed March 29, 2024, https://journal.uptimeinstitute.com/large-data-centers-are-mostly-more-efficient-analysis-confirms/.
13. EIA, "Electricity Data Browser," accessed March 29, 2024, https://www.eia.gov/electricity/data/browser/#/topic/7?agg=0,1&geo=g&endsec=vg&linechart=ELEC.PRICE.US-ALL.A~ELEC.PRICE.US-RES.A~ELEC.PRICE.US-COM.A~ELEC.PRICE.US-IND.A&columnchart=ELEC.PRICE.US-ALL.A~ELEC.PRICE.US-RES.A~ELEC.PRICE.US-COM.A~ELEC.PRICE.US-IND.A&map=ELEC.PRICE.US-ALL.A&freq=A&ctype=linechart&ltype=pin&rtype=s&maptype=0&rse=0&pin=.
14. Dell, "Dell Enterprise Infrastructure Planning Tool," accessed March 25, 2024, https://dell-ui-eipt.azurewebsites.net/#/.
15. VMware by Broadcom partner, Softchoice, uses this rack cost in a TCO analysis comparing costs of running VMware on-premises or in the cloud. Source: Softchoice, "VMware Cloud on AWS," accessed March 24, 2024, https://www.softchoice.com/technology-partners/vmware/cloud-on-aws-tco-calculator.

Read the report ▶

This project was commissioned by Dell Technologies.

Principled Technologies®

Facts matter.®