Dell PowerScale: Key Performance Prediction using Artificial Intelligence for IT Operations (AIOps)

Prediction of PowerScale NFS Latency Spikes using AI Dynamics NeoPulse

September 2022

H19311

White Paper

Abstract

This white paper reveals the results of a benchmarking test with software from ISV partner AI Dynamics that can predict NFS latency. A traditional software system can use these predictions to alert users of potential performance bottlenecks.

Dell Technologies

D&LL Technologies

Contents

Executive summary

Overview

AIOps introduces the intelligence needed to manage the complexity of modern IT environments. The NeoPulse platform from AI Dynamics makes AIOps easy to implement. In an all-flash configuration of Dell PowerScale clusters, performance is one of the key considerations. As hundreds and thousands of performance logs are generated every day, it is very easy for AIOps to consume the existing logs and provide insight into potential performance bottlenecks. Dell servers with GPUs are great platforms to perform training and inference, not just for this scenario but any other new AI challenge the company wishes to tackle, across dozens of problem types.

Revisions

Date	Description
July 2022	Initial release
September 2022	Updated template

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email.

Author: Vincent Shen, Ryan Monsurate, CTO, AI Dynamics

Note: For links to other documentation for this topic, see the PowerScale Info Hub.

Eliminating IT operations challenges with AIOps

Architecting IT infrastructure has never been more challenging than it is today. Many end-user workloads, such as semiconductor design, autonomous driving development, and life sciences and healthcare are growing more complex every day. With that complexity comes growing demand on IT. A further complication is the fact that data footprints are also growing – often exponentially. It is no wonder that IT Administrators struggle to maintain consistent infrastructure performance, with rogue applications (and users) occasionally overloading infrastructure, creating performance issues across the organization. While there are tools available to monitor performance, these tools lack domain knowledge – they are one-size-fits-all – and thus are mostly used to alert administrators after a problem has occurred. Ideally, IT admins would like a means to predict, based on their specific mix of workloads, when performance challenges are going to arise – with enough time to avoid them altogether.

AIOps is the methodology to address these challenges. AIOps refers to the use of artificial intelligence (AI), specifically machine learning (ML) techniques to ingest, analyze, and learn from the large volumes of data (performance logs, usage reports, audits, and so on) that is now available from modern infrastructure, spanning every corner of the IT environment. AIOps can reduce the complexity of managing IT infrastructure by bridging data silos autonomously with the means to filter them, detect patterns, and to cluster meaningful information for more efficient and effective actionable insights. With AIOps, it is now possible to create, with ease, an IT alert solution that adapts automatically to each data center's unique workloads, to provide accurate predictions of performance issues – to provide administrators with a way to prevent them proactively.

The goal of AIOps is to allow IT departments to manage their assets and tackle performance challenges proactively, in real-time (or better), before they become system-wide issues. Along with the flexibility needed to find and fix issues faster, AIOps provides IT professionals with predictive insights to prevent issues from happening in the first place.

A recent study[1] revealed that the global AIOps market will surpass $3 billion by 2025 at a CAGR of 43.7% from 2020 to 2025. This observation leads many to anticipate that most AIOps tools will turn into more innovative and developed platforms from 2022 onwards.

Solution Benefits   The solution benefits are:

- Predict NFS latency spikes up to four hours in advance.

- Generate predictions with only four hours of historical data.

- This AI model can easily be adapted to predict other performance metrics, such as NFS throughput spike, SMB latency spike, and so on.

- The AI learns to improve its predictions week by week as it adapts to each customer's nuanced usage patterns.

---

[1] https://www.analyticsinsight.net/analytics-insight-predicts-global-aiops-market-to-value-at-us-3127-44-million-in-2025/

Dell PowerScale Benefits

The PowerScale benefits are:

- PowerScale enables innovation at scale with unstructured data

- Architecture designed for massive capacity, performance, and scalability

- Multi-protocol access to the data, supporting SMB, NFS, S3, HDFS, HTTP, and FTP

- Seamlessly tier data across all-flash, hybrid, and archive nodes

- With new NVMe drives, PowerScale delivers incredible throughput and performance

AI Dynamics Benefits

The benefits of AI dynamics are:

- Automate AI workflows: data cleaning, transformation, feature engineering, model training, fine tuning, a/b testing, and deployment in a point and click interface

- No-code/low code machine learning architecture design for many types of ML problems

- Scale-out to hundreds of containerized deployments

- Perform many tasks on a single platform – data analytics, image segmentation, image classification, video classification, text analysis, audio analysis, time series forecasts, and anomaly detection

Benchmark results: PowerScale key performance prediction using AIOps

Overview

We conducted our testing using [Dell PowerScale storage](#) with the NeoPulse platform from AI Dynamics to study how to predict performance using AIOps. We identified NFS latency as the performance indicator for which customers would like to see predictive reporting. We studied historical system activity to predict the NFS latency at five minute intervals for four hours in the future. A traditional software system can use these predictions to alert users of the potential performance bottleneck based on the user's specified latency threshold level and spike duration. In the future, AI models can be built that help diagnose the source of these issues so that both an alert and the best recommended solution can be reported to the user.

The entire training process consists of these steps:

- **Data preparation** – to get the raw data and extract the useful features as the input for training and validation.

- **Training the model** – to pick up a proper AI architecture and tune the parameters for accuracy.

- **Model validation** – to validate the AI model based on the data set withheld from the training.

The following sections provide details about each of these steps.

Data preparation Dell Secure Remote Services (SRS) collects the raw performance data from 12 different all-flash PowerScale clusters on-site at a semiconductor design customer. Data from these clusters, which was used to support electronic design automation (EDA) workloads, was collected every week. We identify and extract 26 performance key metrics from the raw data, most of which are logged and updated every five minutes. NeoPulse is used to extract some additional fields, such as day of week and time of day from the UNIX timestamp fields, to allow the model to make better predictions. Every week new data was collected from the PowerScale cluster to increase the size of the training dataset and to improve the AI model. During each training run we also withheld 10% of the data, which was used to test the AI model in the testing phase. This was separate from the 10% of training data withheld for validation.
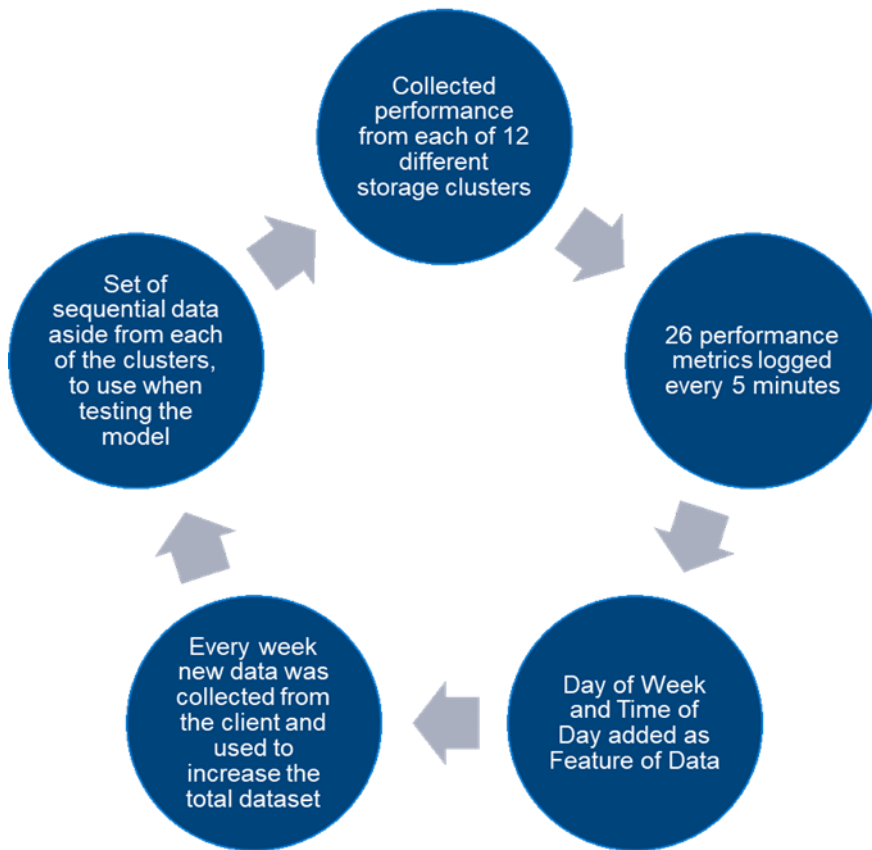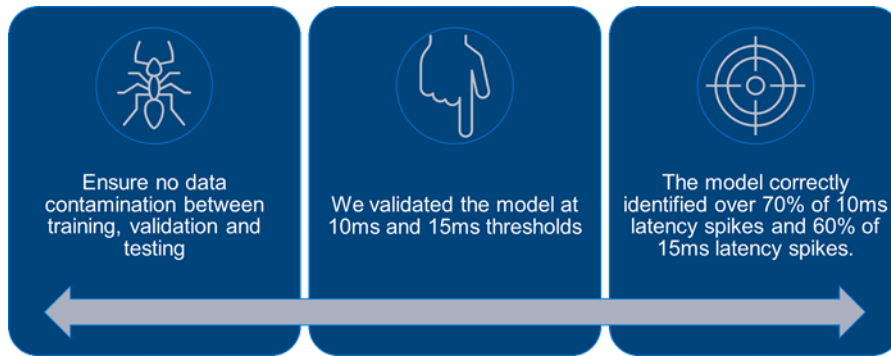

Figure 1. Data preparation process

Training the model Over a period of two months, more than 50 different AI models were trained using a variety of different time series architectures, varying model architecture parameters, hyperparameters, and data engineering techniques to maximize performance, without overfitting to existing data. When these training pipelines were created in NeoPulse, they could be reused easily as new data arrived from the client every week, to rerun training and testing to quantify the performance of the model. At the end of the two-month period we had built a model that could predict whether this one performance metric (NFS3 latency) would be above a threshold of 10ms, correctly for 70% of each one of the next 48 five-minute intervals (4 hours total).

Model validation

As noted earlier, in the data preparation phase we withheld 10% of the total data set to be used for AI model validation and testing. With the current AI model, end-users can easily configure the threshold of the latency as they want. In this case, we validated the model at 10ms and 15ms thresholds latency. The model can correctly identify over 70% of 10ms latency spikes and 60% of 15ms latency spikes over the entire ensuing four-hour period.

Results

In this solution, we use NFS latency from PowerScale as the indicator to be predicted. The AI model uses the performance data from the previous four hours to predict the trend and spikes of NFS latency in the next four hours. If the software identifies a five-minute period when a >10ms latency spike would occur more than 70% of the time, it will trigger a configurable alert to the user.

The following diagram shows an example. At 8:55 a.m., the AI model predicts the NFS latency from 8:55 a.m. to 12:55 p.m., based on the input of performance data from 4:55 a.m. to 8:55 a.m. The AI model makes predictions for each five-minute period over the prediction duration. The model predicts a few isolated spikes in latency, with a large consecutive cluster of high latency occurrences between around 12 p.m. and 12:55 p.m. A software system can use this prediction to alert the user about the expected increase in latency, giving them over three hours to get ahead of the problem and reduce the server load. In the graph, the dotted line shows the AI model's prediction, whereas the solid line shows actual performance.
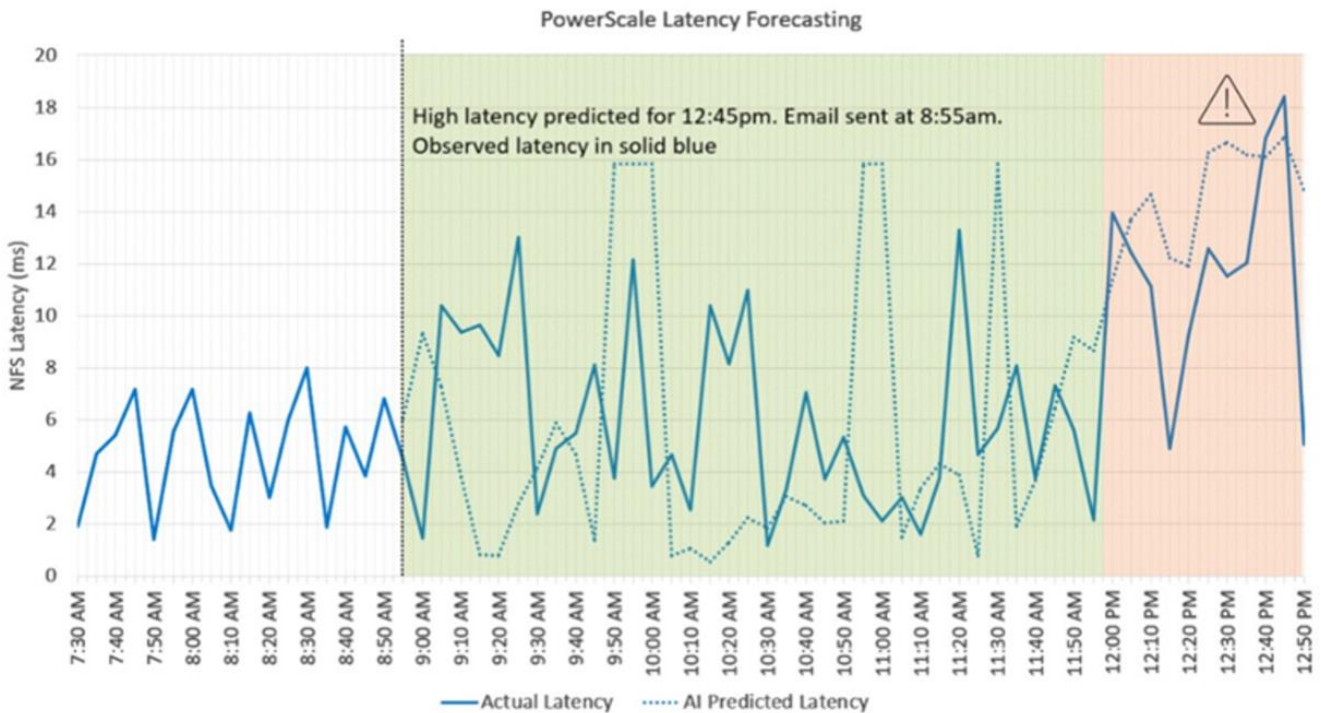
Figure 2.   Dell PowerScale NFS Latency Forecasting

In summary, the solution has achieved the following:

- By using the previous four hours of PowerScale performance data, the solution can forecast the next four hours of any specified metric.

- For NFS3 latency, the solution was benchmarked as correctly identifying periods when latency would be above 10ms 70% of the time.

- The data and model training pipelines created for this task can easily be adapted to predict other performance metrics, such as NFS throughput spike, SMB latency spike, and so on.

- The AI learns to improve its predictions week by week as it adapts to each customer's nuanced usage patterns, creating customized models for each customer's idiosyncratic workload profiles.

Dell PowerScale all-flash storage platform

Dell PowerScale all-flash storage platforms, powered by the PowerScale OneFS operating system, provide a powerful yet simple scale-out storage architecture to speed access to massive amounts of unstructured data, while dramatically reducing cost and complexity. With a highly dense design that contains four nodes within a single 4U chassis, PowerScale all-flash delivers extreme performance and efficiency for the most demanding unstructured data applications and workloads.

Dell Data Science Workstation with GPUs

Data scientists need powerful computing resources to extract valuable insights from vast amounts of data. Dell's Precision line, the world's number one workstation[2], has partnered with NVIDIA to deliver fully integrated AI hardware and software solutions. The Dell Precision Data Science Workstation (DSW) data science platform provides outstanding performance and reliability.

Figure 3. **Dell Precision Data Science Workstation (DSW)**

A member of the DSW family, the Dell Precision 7920 Tower handles learning model training and larger solution frameworks with ease.

- Up to 10x Hard Disk Drive (HDD)/Solid State Drive (SSD) storage

- Ubuntu or Windows OS factory installed

- NVIDIA RTX A6000 certified for NVIDIA Data Science Software and Intel® one API Analytics Tool Kit

For more information, see the article Artificial Intelligence (AI) technologies powered by Dell Precision workstations.

Dell PowerScale all-flash nodes

The Dell PowerScale family includes the following all-flash nodes:

PowerScale F900:

- Maximum performance of all-NVMe drives in a cost-effective configuration

- Each node is 2U in height and hosts 24 NVMe SSDs. Minimum 3 nodes – maximum 252 nodes

- Scale raw storage capacity from 46 TB to 368 TB per node, and up to 93 PB of raw capacity per cluster

- Includes in-line compression and deduplication

Dell PowerScale F200:

- Provides the performance of flash storage in a cost-effective form factor

- Scale raw storage capacity from 3.84 TB to 15.36 TB per node, and up to 3.8 PB of raw capacity per cluster. Minimum 3 nodes – maximum 252 nodes

- Includes in-line compression and deduplication

---

[2] https://www.dell.com/en-us/shop/workstations-isv-certified/sc/workstations

Dell PowerScale F600:

- With new NVMe drives, the F600 provides larger capacity with massive performance

- Scale raw storage capacity from 15.36 TB to 61.4 TB per node, and up to 15.48 PB of raw storage per cluster. Minimum 3 nodes – maximum 252 nodes

- Includes in-line compression and deduplication

Dell PowerScale F800:

- Delivers up to 250,000 IOPS and 15 GB/s aggregate throughput in a single chassis configuration, and up to 15.75M IOPS and 945 GB/s of aggregate throughput in configurations of up to a 252-nodes cluster

- Each chassis houses 60 SSDs with a capacity choice of 1.6 TB, 3.2 TB, 3.84 TB, 7.68 TB, or 15.36 TB per drive

- Scale raw storage capacity from 96 TB to 924 TB in a single 4U chassis, and up to 58 PB in a single cluster

Dell PowerScale F810:

- Delivers up to 250,000 IOPS and 15 GB/s aggregate throughput in a single chassis configuration, and up to 15.75M IOPS and 945 GB/s of aggregate throughput in a 252-node cluster

- Each F810 chassis houses 60 SSDs with a capacity choice of 3.84 TB, 7.68 TB or 15.36 TB per drive

- Scale raw storage capacity from 230 TB to 924 TB in a 4U chassis, and up to 58 PB of raw storage in a single cluster

- Inline data compression and deduplication can deliver up to a 3:1 reduction in storage requirements

Figure 4.   Dell PowerScale F900

For more information, refer to the article: PowerScale All-Flash Scale-Out NAS Specification Sheet.

Dynamics NeoPulse Platform

The NeoPulse Platform enables organizations to manage their entire AI workflow and infrastructure from one place.

This means that DevOps, data engineers, and ML engineers work from one interface instead of using separate applications. Using NeoPulse, a data engineer can assemble training data sets, the machine learning engineer can create and train hundreds of AI models with ease, and the DevOps engineer can deploy and manage the models without ever leaving the NeoPulse environment.
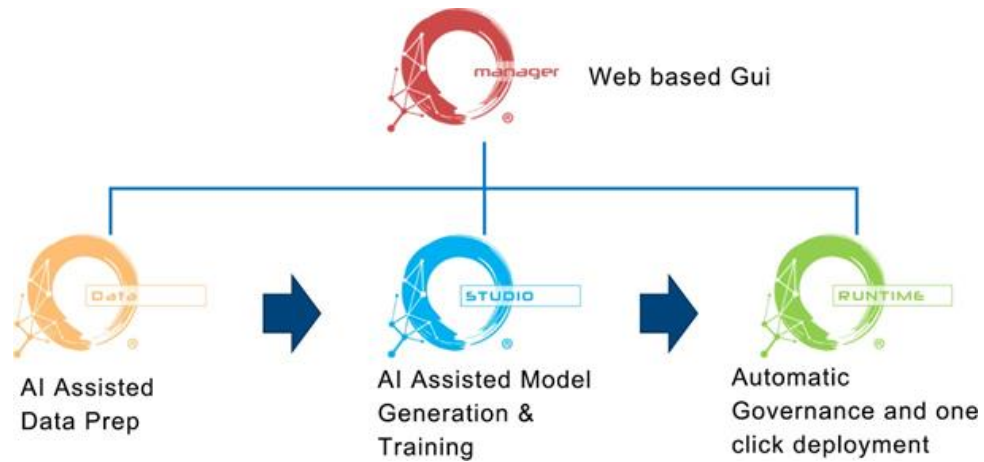
Figure 5.   AI Dynamics NeoPulse Platform

NeoPulse on Dell DSW or PowerEdge servers allows enterprises to automate the process by drawing data from the edge, selecting data most likely to improve AI models, automatically triggering model retraining when new data is available, and then validating and deploying the models. When using NeoPulse application deployment capabilities, all solution deployment and management can be completely automated when certain trigger conditions are met (such as the availability of new data, AI model performance thresholds being breached, and so on) using Dell Technologies PowerEdge. AI Models and applications can also be hot swapped with zero downtime.
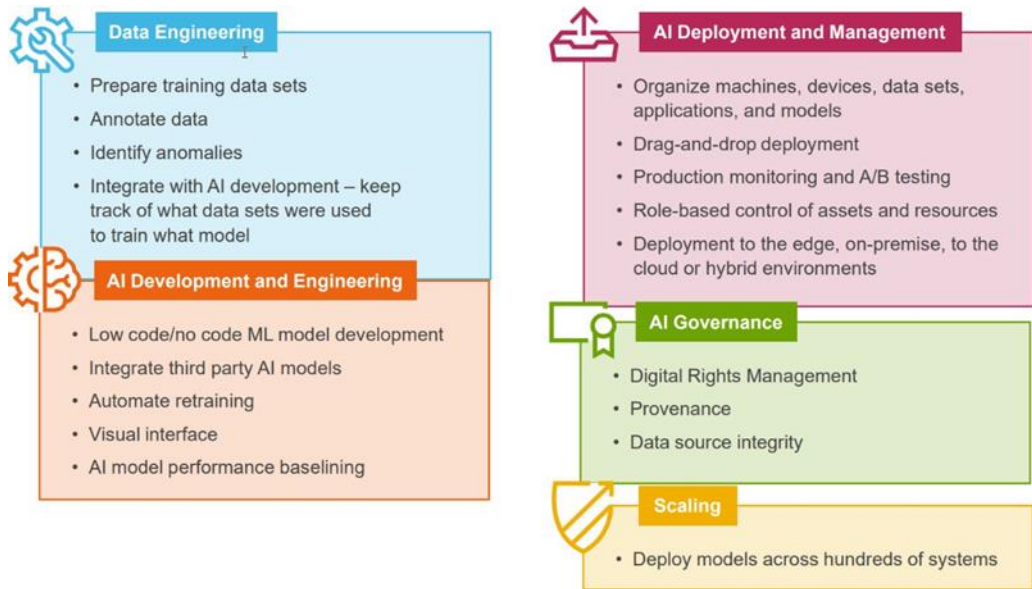

Figure 6.   NeoPulse Features

About Dell PowerScale

Dell PowerScale provides the leading enterprise-grade, scale-out NAS platform that scales from terabytes to 10s of PBs of capacity in a single file system. The PowerScale OneFS file system has unmatched storage efficiency[3] with high utilization capacity and enterprise features such as data deduplication to save even more space to further lower overall TCO. PowerScale has industry-leading data protection[4] that can lose up to four nodes and still operate with no loss of data. PowerScale stays simple to manage, regardless of how your environment grows — allowing you to manage your business and not your storage.

To learn more about Dell PowerScale for Semiconductor Design, visit https://delltechnologies.com/Semi. To learn more about Dell Technologies solutions for Artificial Intelligence/Machine Learning, visit https://www.dell.com/en-us/dt/solutions/artificial-intelligence.

About AI Dynamics

AI Dynamics is a global organization founded on the belief that everyone should have access to the power of artificial intelligence (AI) to change the world. The company's NeoPulse® Framework is a scalable end-to-end AI platform transforming how organizations innovate and grow. Considered to be a true 'operating system for AI,' NeoPulse® is a new AI development standard that empowers anyone at any skill level to create AI solutions with minimum coding and at a fraction of the cost traditionally associated with artificial intelligence. NeoPulse® is the only AI platform that enables a standardized, reproducible process to create, distribute, and manage AI in the enterprise. A trusted solution, NeoPulse® also provides data governance and can be deployed on any on-premises device, cloud, or IoT environment. For more information, visit https://AIDynamics.com.

---

[3] https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/h10541-ds-isilon-platform.pdf

[4] See the section "OneFS high availability and data protection suite", in https://infohub.delltechnologies.com/t/high-availability-and-data-protection-with-dell-powerscale-scale-out-nas/.