SOLUTION BRIEF

Diamanti Machine Learning Platform

KEY BENEFITS

TURNKEY AND SCALABLE PLATFORM TO ACCELERATE AI/ML WORKLOADS

Experiment, train and run production ML models on a single Diamanti ML platform.

EFFICIENCY

Deploy GPU and CPU targeted workloads in the same cluster by leveraging Kubernetes capabilities to ensure effective use of resources.

MAXIMIZE RESOURCE UTILIZATION AND ROI

Achieve more with less. Reduce infrastructure footprint by more than 50%.

Artificial intelligence (AI) and machine learning (ML) have become an integral part of every digital strategy and are driving value across the business, especially in process automation, improving customer experience, cost reduction, and revenue growth. However, the journey towards building a pain-free pipeline for deploying models from experimentation into production is not an easy one. Contrary to popular belief, the ML ecosystem isn't just about the models. It comes with a myriad of components such as data collection, data verification, resource management, analysis tools, server infrastructure, and monitoring. Figure 1 shows that ML code is a much smaller part of the ML ecosystem than the various infrastructure components required for its support.

As ML projects move from experimentation to large-scale production environments, data scientists and data engineers spend a significant amount of time overcoming many of the DevOps challenges to support provisioning, configuring, scaling and managing these ML ecosystem components. Additionally, making the choice between building the ML infrastructure on-premises, in the cloud, or a hybrid environment isn't a one-time decision and can change during the project lifecycle. Data scientists need the flexibility to easily operationalize and migrate ML models across hybrid cloud environments.

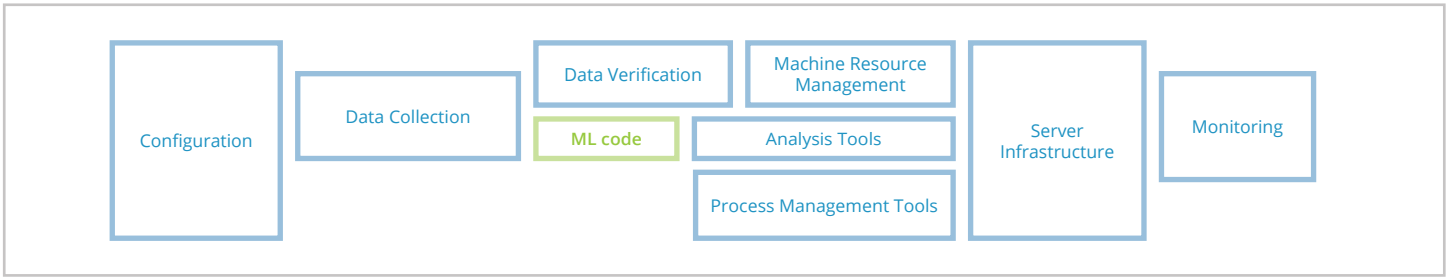

Figure 1: The ML code is just a small part of the ML ecosystem[1]

1 Sculley, Holt et al: Hidden Technical Debt in Machine Learning Systems

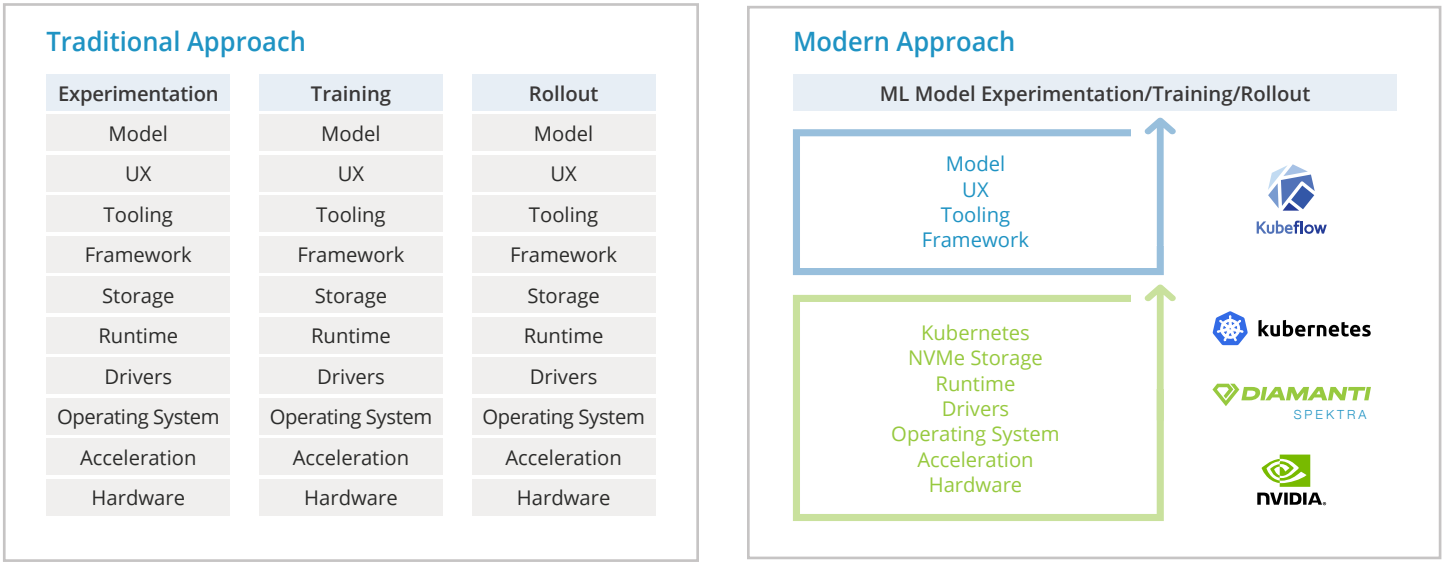

Figure 2: Leap forward to the modern approach of running ML workloads with Kubeflow and Diamanti ML Platform

To address these DevOps challenges data scientists and data engineers need a scalable, composable, and portable solution. Containers and Kubernetes help exactly with this, but there is a lot to learn about containers, Kubernetes, networking, persistent storage, services, deployments, GPUs, drivers and more. This gap has been partially addressed by open source projects such as Kubeflow[2] which helps in managing the ML lifecycle. Kubeflow is an open-source Kubernetes-native platform that provides support for spawning and managing components such as Jupyter notebooks, scalable training services (for TensorFlow, PyTorch) and ML pipelines. Kubeflow is gaining a lot of traction both in terms of commits (1850+) and contributors. Organizations such as Google, Intel, Microsoft, LinkedIn, IBM, and more are actively contributing to the Kubeflow project.

While Kubeflow makes it easy for data scientists and data engineers to learn, deploy, and manage portable, distributed ML workloads on Kubernetes, the Diamanti ML Platform delivers a turnkey Kubernetes solution and the underlying infrastructure. Together, Diamanti and Kubeflow help data science teams to leap forward from the traditional to modern approach of running ML workloads. With the legacy/traditional approach, the journey from experimentation to building production ready ML systems requires data scientists to spend most of their valuable time rearchitecting the infrastructure and manually tweaking deployments. This results in complex infrastructure setup and increases the cost of building and maintaining a deep learning infrastructure. With the modern approach, data scientists can experiment, train and run production ML models on a single scalable platform with minimal infrastructure work. This enables data science teams to dedicate more time building models and delivering value to their organizations.

Diamanti ML Platform

The Diamanti ML platform is a powerful combination of:

- Diamanti Spektra, a pre-validated, pre-packaged and fully-featured software stack including Kubernetes, container runtime, operating system, enterprise-class DP/DR features, access controls and Management UI
- Diamanti Ultima I/O acceleration cards that offload networking and storage traffic to deliver dramatically improved performance
- Diamanti D20 series of modern hyperconverged platforms offering multiple configurations consisting of Intel CPUs, NVIDIA GPUs, memory, and NVMe storage.

2 https://www.kubeflow.org/

The Diamanti ML Platform provides the flexibility for data scientists to accelerate model training with GPUs and deploy CPU-optimized trained data models and any other workloads to the same cluster. With Diamanti's highly available storage system, data is readily accessible for model training as well as to run analytics to draw insights. This avoids back and forth movement of massive datasets between the online processing system and off-the-cluster storage hence saving significant time for data scientists.

The Diamanti ML Platform combines the power of Diamanti Ultima I/O acceleration cards, NVIDIA GPUs and lightning-fast NVMe shared storage to provide a bare metal Kubernetes platform for data scientists to run ML workloads by maximizing resource utilization and return on investment. In addition to data science tools, Diamanti Spektra can host various data analytics tools, databases and monitoring/logging solutions such as Kafka, Crunchy, MongoDB, PostgreSQL, Microsoft SQL, Splunk, Spark, Elasticsearch and more.

Scalable AI architecture

The foundation of AI/ ML workloads is data. Recent research[3] suggests that model accuracy increases logarithmically based on the volume of training data. Data science teams require multiple terabytes of data to get small improvements in model accuracy. Legacy scale-up storage arrays don't fit into modern scale-out container based deployments. Diamanti meets the storage needs of AI/ML workloads with its scale-out architecture. Each Diamanti cluster dynamically pools NVIDIA GPUs and low-latency, high-performance NVMe flash storage, extending NVMe across the cluster, offering data mobility without compromise.

Easy to manage and monitor GPU infrastructure

Diamanti Spektra management dashboard provides an at-a-glance view of the health and status of both G20 (GPU enabled) and D20X nodes. The dashboard displays important metrics such as the GPU utilization, energy, power consumption, GPU memory usage and temperature.

Open-Source Flexibility

Diamanti provides open-source, vendor-agnostic interfaces for networking and storage, and is committed to enabling as many choices as possible in how cloud-native applications are deployed. With an open-source operating system, container runtime and Kubernetes included, there's no vendor lock-in.

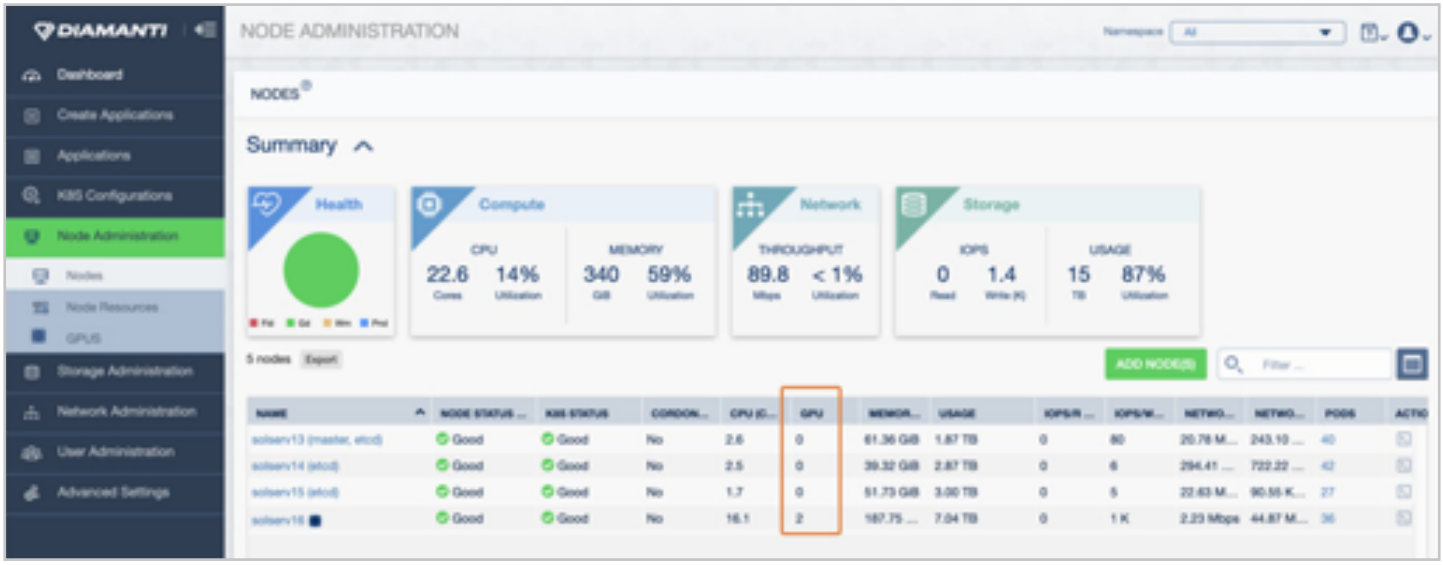

Figure 3: Diamanti Spektra Management Dashboard displaying GPU enabled nodes

3 Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, C. Sun, A. Shrivastava, S. Singh, and A. Gupta.

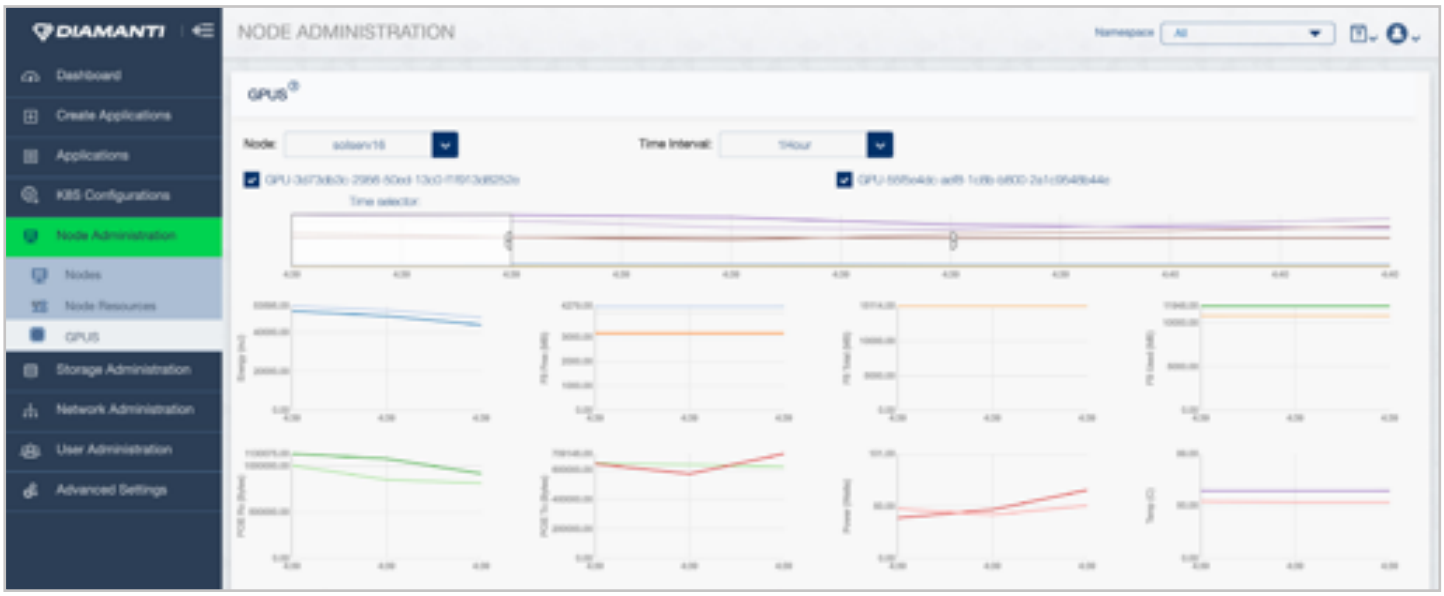

Figure 4: Insights into GPU metrics

Customer Case Study: Optical Character Recognition (OCR) for Invoice Management

A major energy company turned to Diamanti for a new workload leveraging AI/ML for OCR to scan invoices. The customer needed to scan more than 15,000 invoices a day. The legacy infrastructure could not keep up with the demand and eventually accrued a backlog of more than 200,000 invoices. Deploying the Diamanti solution with GPU support eliminated the backlog within hours.

Customer Case Study: Improved Web Conversions

A major online travel company wanted to double its conversion rate on website visitors looking at offers and booking travel. They are using AI modeling to determine the best approach, but unsuccessfully ran proof of concept trials with several major vendors before turning to Diamanti. With Diamanti, they are meeting their new performance requirements without any dependencies on proprietary software. They can also run both their GPU-dependent applications and standard containerized applications in the same environment, minimizing the complexity of managing multiple stacks.

Summary

Using the power of Kubernetes on bare-metal, data scientists can tap into performance, ease of use, and flexibility of deploying AI/ML workloads in containers. Diamanti ML Platform with NVIDIA GPU support provides a turnkey solution for deploying containerized workloads on Kubernetes, ideal for the demanding requirements of emerging AI/ML applications. It supports easy scaling of AI/ML workloads, allows for seamless application portability and reuse, and provides an intuitive portal for deploying and managing Kubernetes infrastructure.