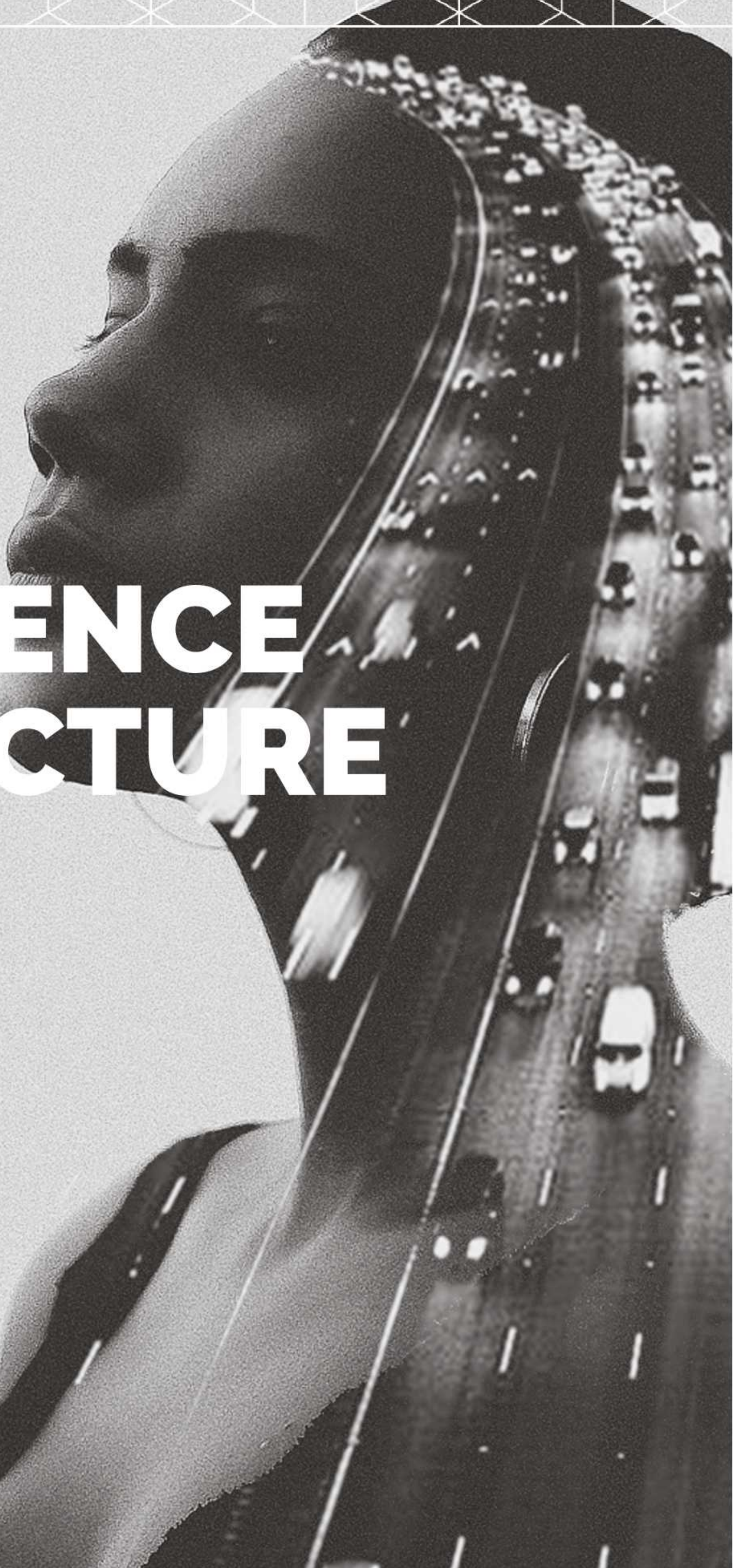
WHITE PAPER

VSCALER AI REFERENCE ARCHITECTURE



VSCALER

BOSTON
Servers | Storage | Solutions

in partnership with

Mellanox
TECHNOLOGIES

TABLE OF CONTENTS

2. Abstract
2. Intro / Executive Summary
3. AI /Deep Learning Workflow
4. Solution Overview
5. vScaler
6. vScaler Storage
6. Ingest/Access
6. Automated Tiering
6. Data Policies
6. Erasure Coding
7. The NVIDIA® DGX-2™ Server
8. Mellanox RDMA Fabric
8. vScaler App. Environment
9. NVIDIA GPU Cloud (NGC)
9. Kubernetes
10. vScaler Multi-Cloud Scaling
11. Reference Architecture
11. Power Considerations
12. Rack Architecture
13. Benchmarking
13. TensorFlow for ResNet
16. Conclusion

"We are at the beginning of a revolution that is fundamentally changing the way we live, work, and relate to one another. In its scale, scope and complexity, what I consider to be the fourth industrial revolution is unlike anything humankind has experienced before..."
Klaus Schwab,
World Economic Forum

ABSTRACT

The convergence of Artificial Intelligence (AI) and High Performance Computing (HPC) has been a driving factor of broader adoption of HPC by a wide variety of industries. This paradigm shift is a result of both computing power advancements, through products like the GPU, and more data being generated and stored than ever before. The time is ripe for AI, and organisations looking to gain an edge in business are turning more and more to AI development to build their next generation of products and services.

INTRO / EXECUTIVE SUMMARY

AI is disrupting the world as we know. Industries across the board are using AI technology to improve quality, performance, functionality and even revenue growth. AI, both as a concept and technology has been around for over half a century, but it is only recently that AI has started to have an impact, revolutionizing industries as diverse as health care, finance, law, automotive, aerospace and manufacturing with the potential to profoundly affect how people live, work, and play.

Over the last two decades we have seen a new industry trend towards simulation and modeling over legacy physical testing, with a focus on leveraging HPC tools and techniques. We only need to look at how clusters have replaced wind tunnels for the simulations of Jet Engine designs and in some cases Formula 1 cars. Healthcare and in particular Genomics, are other example areas that have been embracing HPC technologies to reduce both the time to treatment and result, as well as the cost of sequencing.

When we combine all the above with the fact that data is being generated, captured and analysed by an ever increasing amount of devices we have a situation where we are on the cusp of the 4th industrial revolution with the fusion of technologies between the physical, digital and biological arenas, which will fundamentally alter the way we live, work and relate to one another.

AI (/DEEP LEARNING) WORKFLOW

The development of AI solutions is complex and varies greatly depending on the application, and involves many stages and technologies. In figure 1 we cover a high-level, typical AI workflow.

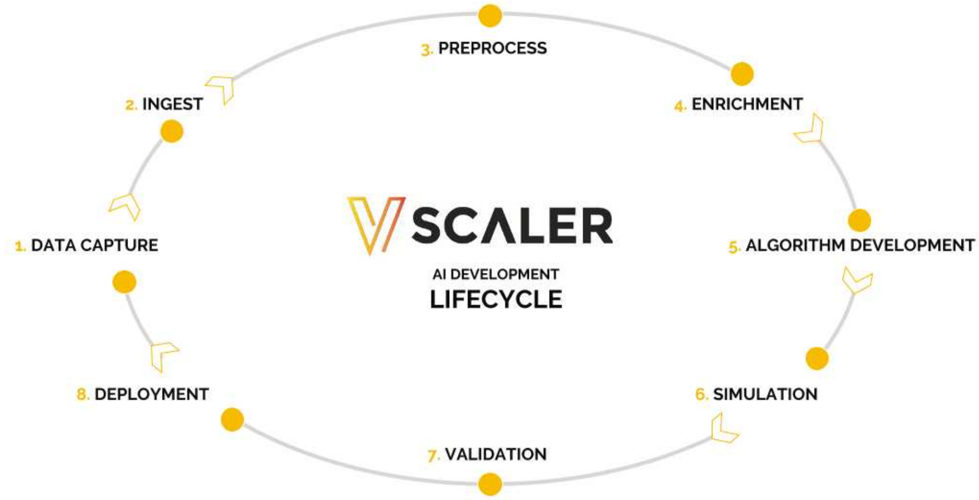


Fig 1. vScaler AI Development Lifecycle

1. Data Capture: Before we can look at training neural networks we need a raw data set. These data sources vary from project to project and can often be publicly available data sets of audio, video, images or sensor data.

2. Ingest: With the data sets selected or data sources known, we then need to ingest this information into our Data Lake for further processing. Considerations should be given to the rate of ingest, frequency (Streaming or periodic blocks) and the protocols used (Ingest via Object S3 interface, or direct to POSIX based file system via FTP/SCP– the Data Lake should accommodate them all).

3. Preprocess: The Data Lake has a number of policy driven automation features which can manipulate or process the data as it is ingested. For example, images can be resized, audio or video can be split up and tagged or enriched with meta-data based on its source.

4. Enrichment: While the preprocess stage will be mainly automated, the enrichment phase can be a more manual process and require human participation. For example creating a data set with known tags and meta-data that can be used to train neural networks.

5. Algorithm Development: During this phase researchers need to test and evaluate different models and frameworks. There are a lot of options and the ability to compare them concurrently and with minimal administration overhead is important. During this stage its also important to be able to run many small simulations (on virtual environments) before scaling up and running larger scale production (bare metal) workloads.

6. Simulation: With the data sets labelled and enriched and the algorithms prepared, its time to train your model. Depending on the size of the data sets, the number of concurrent projects and the depth of your neural networks you will want to be able to scale your environment to reduce the time to results. This scalability is dependant on optimized hardware for high throughput, both in terms of IO and CPU as well as software frameworks that grow as your infrastructure expands.

7. Validation: In order to measure the accuracy of your model you will need to use unknown or unseen data sets and run the models in an inference or classification mode. Consideration for concurrent projects and data sets size is also a critical factor during the validation stage.

8. Deployment: As part of the cyclical nature of software development, data collection and generation, the ability to deploy solutions in a scalable (Inference as a service) and repeatable manner is critical. Rolling updates (and rolling back) become important considerations and managing that process should require minimal overhead.

The dynamic nature of these AI environments typically means that the data sets are constantly growing and algorithms are being optimized for greater accuracy. As soon as your solution is deployed, the team will be busy working on version 2 of the software. An infrastructure that can handle growth seamlessly and grow as your requirements grow is crucial.

.

SOLUTION OVERVIEW

VSCALER™

Today's leading Deep Learning models typically take days or weeks to train, forcing analysts to make compromises between accuracy and time to deployment. In order to speed up training times, GPUs are typically deployed - but can incur complex management challenges as well as increased costs if not utilised efficiently.

vScaler - an optimised cloud platform built with AI and Deep Learning workloads in mind - enables anyone to quickly deploy scalable, production-ready AI environments via an optimised private cloud appliance. The components of the vScaler architecture are presented in figure 2.

VSCALER KEY FEATURES:

- vScaler manages all hardware infrastructure including compute nodes, storage nodes and network switches.
- Built for high performance workloads.
- Application specific software stacks.
- Highly available, with no single point of failure.
- Zero-touch provisioning (of additional hardware – plug and play)/
- Rolling upgrades of core services / Zero downtime.

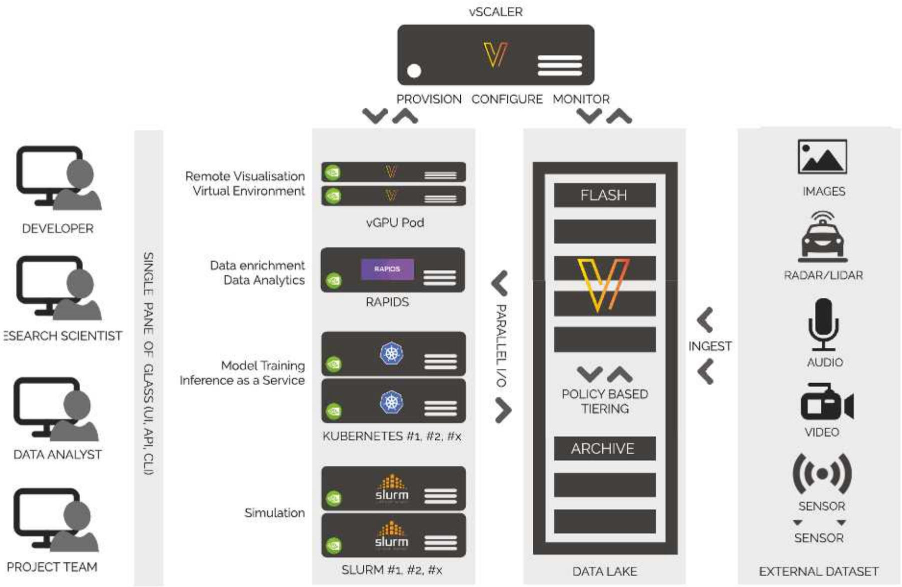


Fig 2. vScaler AI Workflow

VSCALER STORAGE

NVIDIA GPUs are extremely powerful and capable devices, however they need to be configured in a balanced architecture to ensure that they are optimally used. Getting the storage infrastructure correct is a critical component in the overall architecture. Leveraging local NVMe drives within for example, an NVIDIA^{TM} DGX server, is simply not scalable and also creates challenges around data security and management when you try to scale your infrastructure.

With vScaler Storage we centralise all IO and place active files on the fastest tier (Flash/NVMe) but can also leverage the cost benefits of HDD and provide online archival capabilities for longer term storage - all within a single platform.

VSCALER STORAGE KEY FEATURES

- Self-monitoring and self-healing (manage large clusters with minimal staff)
- High IOPs, low-latency client (also provides seamless failover)
- Policy-based data management and placement
- Volume Mirroring (perfect for disaster recovery, or as a backup source)
- Runtime reconfigurable (eliminates downtime, changes on the fly)
- Simple, scripted cluster installation (up and running in less than an hour)
- Compatible with any modern Linux distribution.
- Supports POSIX, NFS, S3, SMB, and Hadoop file access methods (one platform for all workloads)
- Works with HDDs, SSDs, and NVMe devices (no need to buy expensive flash when it's not needed)

Unified Storage Access

VCS implements strong data semantics that can subsume all kinds of unstructured data including block storage and object storage

Single Management Interface

Manage and scale your storage resource with an easy to use web interface

Future Proof

With support for Kubernetes, Mesos, and Rancher you get a solution for rapidly fluctuating workloads

VSCALER STORAGE - INGEST / ACCESS VIA ANY PROTOCOL

vScaler storage features a unified storage layer which allows multiple protocols to access the underlying parallel file system. This enables users to ingest via S3 (Object Interface – e.g. IOT devices uploading file via PUT), perform data processing in Hadoop via the HDFS translator and then run a Kubernetes training workload via the parallel POSIX based filesystem. All on a single volume without having to move any files. See figure 3.

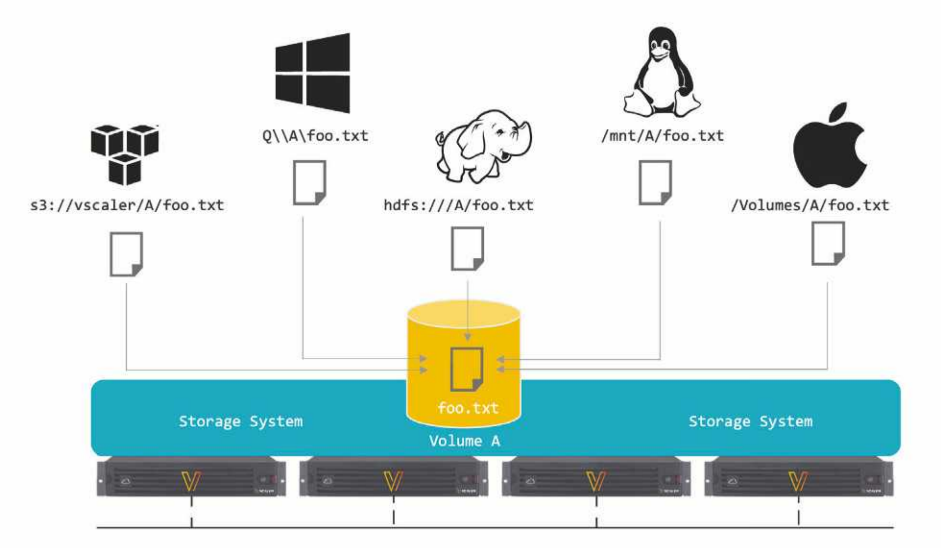


Fig 3. vScaler Storage - Ingest/Access via any protocol

VSCALER STORAGE - AUTOMATED TIERING

As all devices are automatically tagged with the device type, administrators can define a simple age-based placement rule that tiers older data from NVMe to HDD automatically. As a result, archiving old data to a slower tier (Data Lake) becomes an automated process.

VSCALER STORAGE - DATA POLICIES

vScaler's software flexible data options extend beyond simple tiering and policies can be created for specific file types when created. For example, image files that are ingested can automatically be converted to different resolutions for later use during the training or inference stages of development.

VSCALER STORAGE - ERASURE CODING

Erasure coding (EC) is offered as an alternative to replication and is user-configurable on a per-volume or per-file basis. As well as being more space efficient, EC is ideal for large-block sequential workloads, like those found with streaming media, surveillance, analytics, or transcoding. EC is also ideal for long-term archiving, where space efficiency is key.

THE NVIDIA™ DGX-2 SERVER

If you're reading this it's safe to assume you know what a CPU (Central Processing Unit) is and just how powerful these chips are. But if you're an AI developer, CPUs alone are not enough. They can do the processing, but the sheer volume of unstructured data that needs to be analysed to build and train deep learning models can leave them maxed out for weeks on end. Even multi-core CPUs struggle with deep learning, which is where the GPU (Graphics Processing Unit) comes in.

GPUs are particularly adept at processing matrices -- something CPUs have trouble coping with -- and it's this that also suits them to specialised applications like deep learning. NVIDIA® Tesla® GPUs are powered by Tensor Cores, a revolutionary technology that delivers ground-breaking AI performance. Tensor Cores can accelerate large matrix operations, which are at the heart of AI, and perform mixed-precision matrix multiply and accumulate calculations in a single operation. With hundreds of Tensor Cores operating in parallel in one NVIDIA GPU, this enables massive increases in throughput and efficiency.

The DGX-2 server (see figure 4) contains 16 GPUs to create a 2Petaflop (tensor ops) monster of a system. Some of the hardware highlights include:

- 16x V100 32GB GPUs (That's half a TB of GPU HBM2 memory space, when used with the CUDA unified memory and cudaMallocManaged()
- 12x NVSwitch switches providing a non-blocking GPU fabric with 2.4TB/s bisection bandwidth.
- 800GB of a network trunk to get data in and out (seems like overkill for just ssh!)
- 30TB of local NVMe SSD to keep those GPUs busy

The DGX-2 (and also the previous generation DGX-1) is integrated in the vScaler architecture as the building block for computationally intensive workloads.

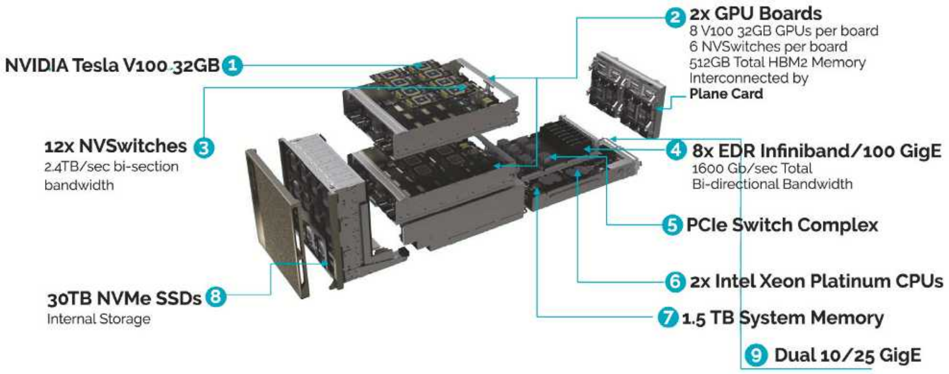


Fig 4. NVIDIA™ DGX-2 Server

MELLANOX RDMA ACCELERATED FABRIC

Mellanox Solutions accelerate many of the world's leading artificial intelligence and machine learning platforms and wide range of applications, ranging from security, finance, and image and voice recognition, to self-driving cars and smart cities. Mellanox solutions enable companies and organizations such as Baidu, NVIDIA, JD.com, Facebook, PayPal and more to leverage machine learning platforms to enhance their competitive advantage.

The latest advancement in GPU-GPU communications is GPUDirect RDMA. This new technology provides a direct P2P (Peer-to-Peer) data path between the GPU Memory directly to/from the Mellanox HCA devices. This provides a significant decrease in GPU-GPU communication latency and completely offloads the CPU, removing it from all GPU-GPU communications across the network. GPU Direct leverages PeerDirect RDMA and PeerDirect ASYNC™ capabilities of the Mellanox network adapters.

Key Benefits:

- Avoid unnecessary system memory copies and CPU overhead by copying data directly to/from pinned GPU memory
- Peer-To-Peer Transfers Between GPU device and Mellanox RDMA devices
- Use high-speed DMA transfers to copy data between P2P devices
- Eliminate CPU bandwidth and latency bottlenecks using direct memory access (DMA)
- With GPUDirect RDMA, GPU memory can be used for Remote Direct Memory Access (RDMA) resulting in more efficient applications
- Boost Message Passing Interface (MPI) Applications with zero-copy support

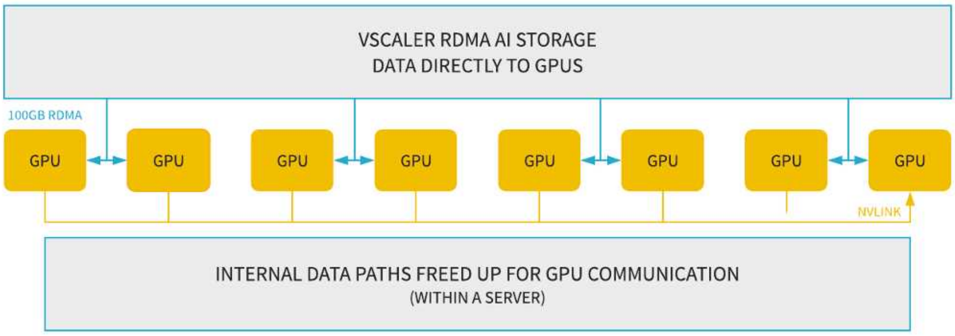


Fig 5. vScaler Mellanox RDMA Fabric for AI

VSCALER APPLICATION ENVIRONMENT

Spin up application specific environments with the appropriate Deep Learning frameworks installed and ready for use, including Tensor- flow, Caffe and Theano*. These frameworks are accelerated using the world's fastest GPUs, purpose-built to dramatically reduce training time for Deep learning and Machine Learning algorithms and AI simulations.

NVIDIA GPU CLOUD (NGC)

User workloads on the vScaler AI Architecture primarily utilize containers from NGC which provides researchers and data scientists with easy access to a comprehensive catalogue of GPU-optimized software for DL, HPC applications, and HPC visualization that take full advantage of the GPUs. The NGC container registry includes NVIDIA tuned, tested, certified, and maintained containers for the top DL frameworks such as TensorFlow, PyTorch, and MXNet. NGC also has third-party managed HPC application containers, and NVIDIA HPC visualization containers.

.

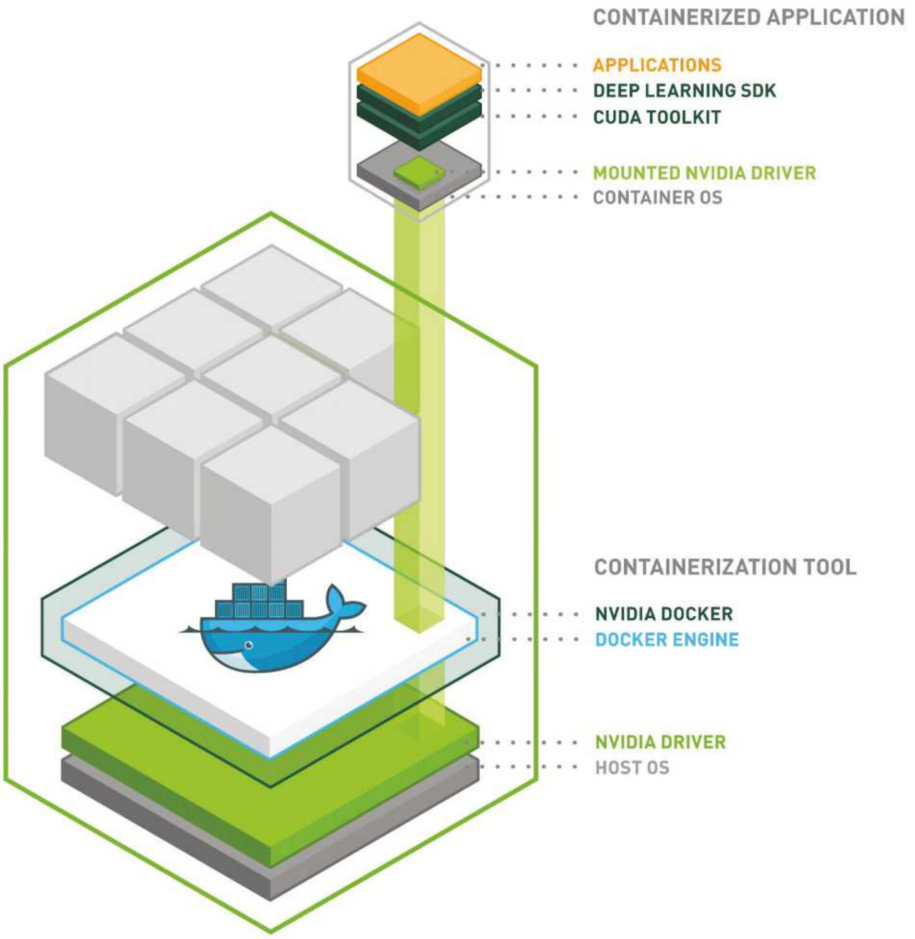


Fig 6. vScaler Software Stack

KUBERNETES CONTAINER ORCHESTRATION ENGINE

Kubernetes is a platform that automates the deployment and management of containerised applications, including complicated workloads like AI and machine learning. Kubernetes and AI represent converging trends. Most companies are running Kubernetes as a platform for their workloads, or plan to soon, and AI is an increasingly important workload. As organizations shift their attention to AI to reduce operating costs, improve decision-making and serve customers in new ways, Kubernetes-based containers are becoming the go-to technology to help enterprises adopt AI and machine learning.

vScaler provides the ability to spin up multiple Kubernetes environments on bare metal (production scale workloads) or in a virtual environment (testing/development) so users can allocate the appropriate resources and the ability to monitor the GPU environment over time with metrics collected to report GPU memory usage, temperature and total utilisaiton over time. This provide a great feedback loop for developers to observe the performance profiles of their applications.

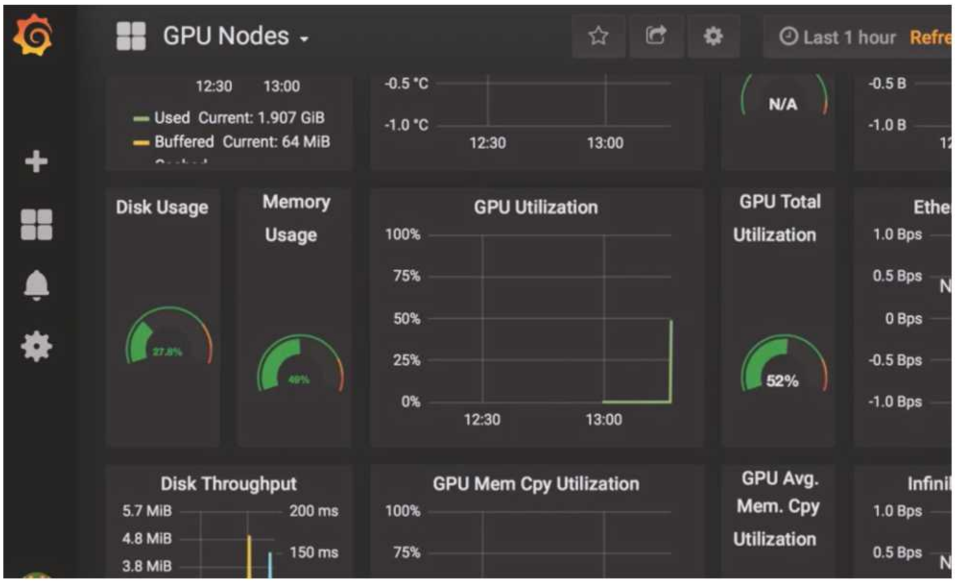


Fig 7. GPU memory usage monitoring

VSCALER MULTI-CLOUD SCALING

If your organization needs to scale out or run multiple environments in separate locations, vScaler can orchestrate and manage remote locations from a single pane of glass. A small core team can roll out consistent features and capabilities to their users globally. The backend storage can also be replicated across zones for resilience and availability.

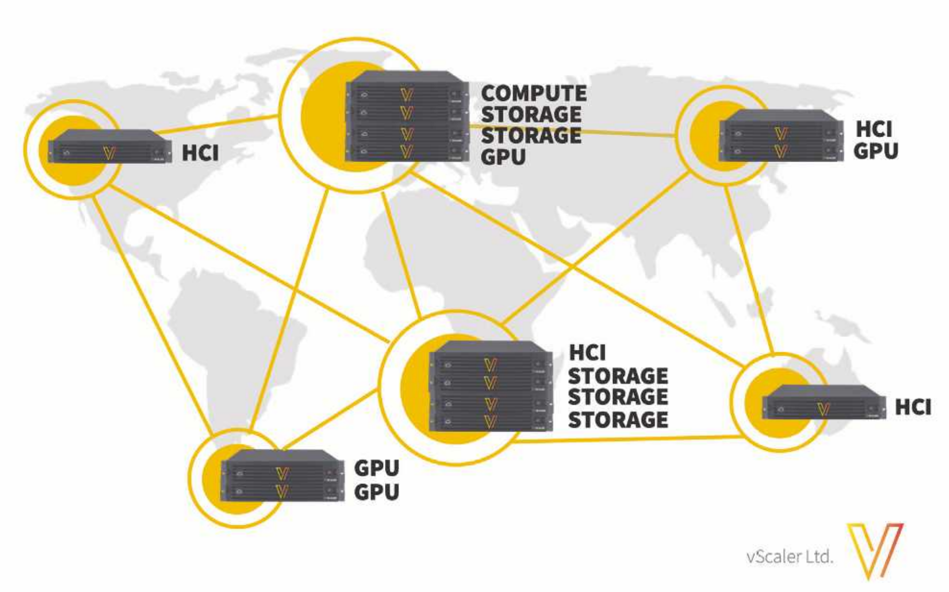


Fig 8. vScaler Multi-Cloud Scaling

REFERENCE ARCHITECTURE

The vScaler AI reference architecture is based on the following building block. There is a 3 node management plane and then after that you add the components required depending on your storage and compute needs. For smaller / edge configurations we can provide hyperconverged platforms which can also be utilized as a starting point for initial setups. As demands grow the solution can then be scaled out with the appropriate building blocks (Either as a hyperconverged solution or as a scale out platform where resources are scaled as required.

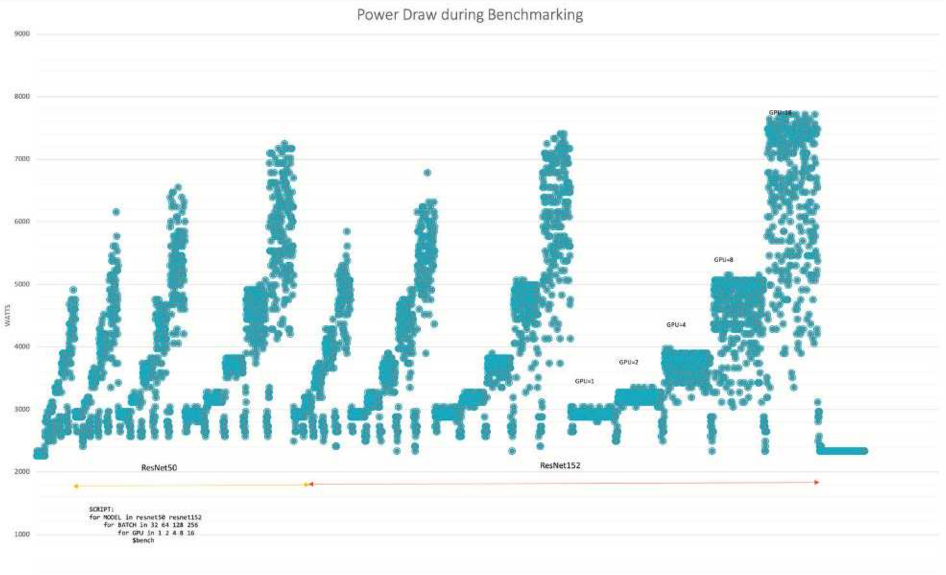
	HCI	Compute	Flash/Hybrid	Stor XL	vGPU	DGX-1	DGX-2
Nodes	4 (2U)	4 (2U)	1 (2U)	1 (4U)	1 (1U)	4U	8U
Cores	288	288	N/A	N/A	36	N/A	N/A
Storage	24 2.5"	N/A	24 2.5" (4 NVMe)	60 3.5" (6 NVMe)	N/A	4x 1.92TB	8x 3.84TB
Memory	1.5TB	1.5TB	192GB	384GB	384GB	512GB	1.5TB
Network	Dual 25G	Dual 25G	Dual 25G	Dual 100G	Dual 25G	4x100G	8x100G
Accelerators	N/A	N/A	N/A	N/A	1-4 V100	8 V100	16V100

Fig 9.vScaler Building Blocks

POWER CONSIDERATIONS

The building blocks used for the computational work are very high performance and high density, which also means they are quite high-power (watts) as well and careful planning when deploying in datacenters is required. In our benchmarking on the DGX-2 server we observed a power draw of 7.8kw (within an 8U) chassis. Please refer to the diagram below for power measured captured during the ResNet benchmarking on the DGX-2 server.

7.8kw max power draw



```
SCRIPT:
for MODEL in resnet50 resnet152
    for BATCH in 32 64 128 256
        for GPU in 1 2 4 8 16
            $bench
```

Fig 10. Power Considerations

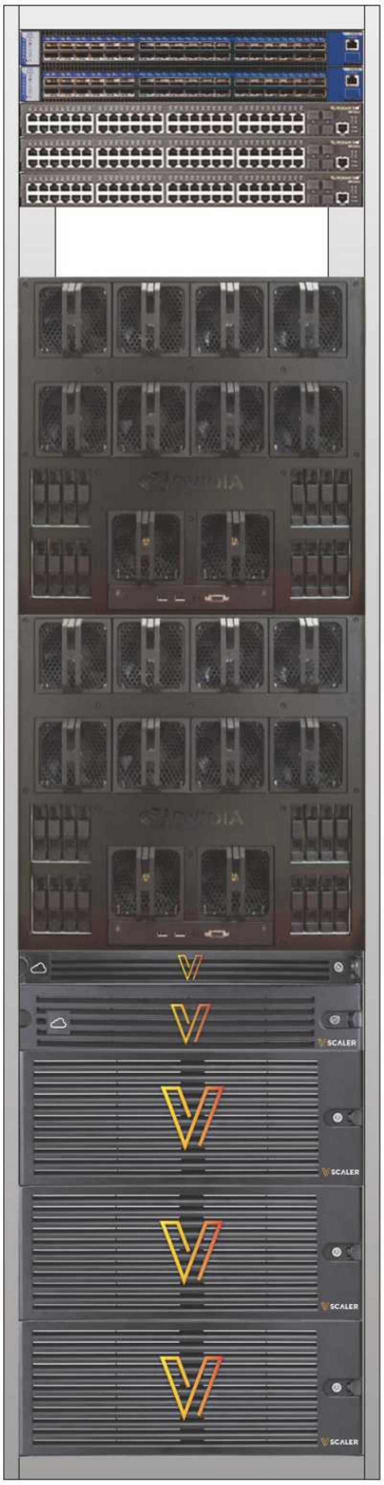


Fig 11. Rack Diagrams

RACK ARCHITECTURE

In the rack architecture diagrams we assume up to 30Kw of power and cooling can be provided. For datacenters not optimized for high density power/cooling we can re-architect this configuration to fit within more modest power budgets per rack.

VSCALER DGX BUILDING BLOCK

As part of our building block, vScaler provides all of the necessary components required to run AI workloads at production scale. This is a suggested configuration only, and can be customised to meet any size and budget to suit customer requirements by simply swapping out the building blocks included.

The suggested configuration in Fig. 11 includes:

- 1x HCI (Control Plane) (2u)
- 1x vGPU Pod (1u)
- 2x DGX-2 (10u each)
- 3x 4U Storage (4U each)
- 2 x Mellanox 100GB Switches
- 2x Supermicro Management
- 1x Supermicro IPMI Management
- Total RackU : 40U (and total power 24kw)

NETWORK ARCHITECTURE

The high-performance fabric is used for NVMe storage delivery from the storage nodes to the GPU server at high bandwidth and low latency. The network design is resilient so there are no single points of failure in the fabric. This fabric can also be used to scale applications across multiple nodes (using MPI or NCCL) which also provides GPU Direct / RDMA between GPUs in separate servers.

MANAGEMENT AND OUT-OF-BAND

The management network and out of band network are used as the control fabric within the architecture. Node provisioning and monitoring is carried out on the management fabric. A physically separate Out-of-Band (IPMI) which provides full system power control, full remote console access and full KVM (Keyboard/Video/Mouse) is also available on this fabric to allow teams remotely manage the entire infrastructure.

BENCHMARKING

There are a wide number of benchmarks that could be run on an architecture like this, but for the purpose of this whitepaper we have focused on AI training and inference performance, using the DGX2 as our building block base with NVMe only vScaler storage. The framework we focused on was ResNet (shorthand for Residual Networks) which many use as a baseline for performance of training/inference on neural nets. We focus on training and inference performance per node.

All benchmarks were run using nvidia-docker, making use of the latest TensorFlow container provided by NVIDIA GPU Cloud (nvidia/tensorflow:18.10-py3), with the imagenet synthetic dataset (provided as part of the tf_cnn_benchmarks).

The benchmark script used was obtained from https://github.com/tensorflow/benchmarks/tree/master/scripts/tf_cnn_benchmarks and we performed a sweep of batch sizes across the tests. All tests were run a number of times and numbers reported are averaged.

TENSORFLOW BENCHMARKING FOR RESNET MODELS

To assess the performance of the system we've employed the commonly used ResNet Model which many use as a baseline for assessing the training/inference performance of systems. ResNet is a shorthand for Residual Network and as the name of the network suggests, the network relies on Residual Learning (which tries to solve the challenges with training Deep Neural Networks. Such challenges include increased difficulty to train as we go deeper, as well as accuracy saturation and degradation. We've selected two common models: ResNet-50 ResNet-152 (Where ResNet50 is a 50 layer Residual Network, and 152 is a 152 layer Residual Network).

Each model was run using various batch sizes to ensure that each GPU was fully utilised, demanding the highest level of performance from the system. Each combination of batch size and GPU count was tested 3 times over 20 epochs and the average result recorded. Results below show the images processed per second during the network training phase.

Training Command:

```
python tf_cnn_benchmarks.py --data_format=NCHW --batch_size=${BATCH_SIZE}
--model=${MODEL} --optimizer=momentum --variable_update=replicated --nodistor-
tions --gradient_repacking=8 --num_gpus=${NUM_GPUS} --num_epochs=10 --weight_de-
cay=1e-4 --data_dir=/workspace/data --use_fp16 \
--train_dir=${CKPT_DIR}
```

Inference Command:

```
python tf_cnn_benchmarks.py --forward_only=True --batch_size=${BATCH_SIZE}
--model=${MODEL} --num_epochs=10 --optimizer=momentum --distortions=True --dis-
play_every 10 --num_gpus=${NUM_GPUS} --data_dir=./test_data/fake_tf_record_data/
--data_name=imagenet
```

THE RESULTS

>12k images per second

ResNet50(Training)

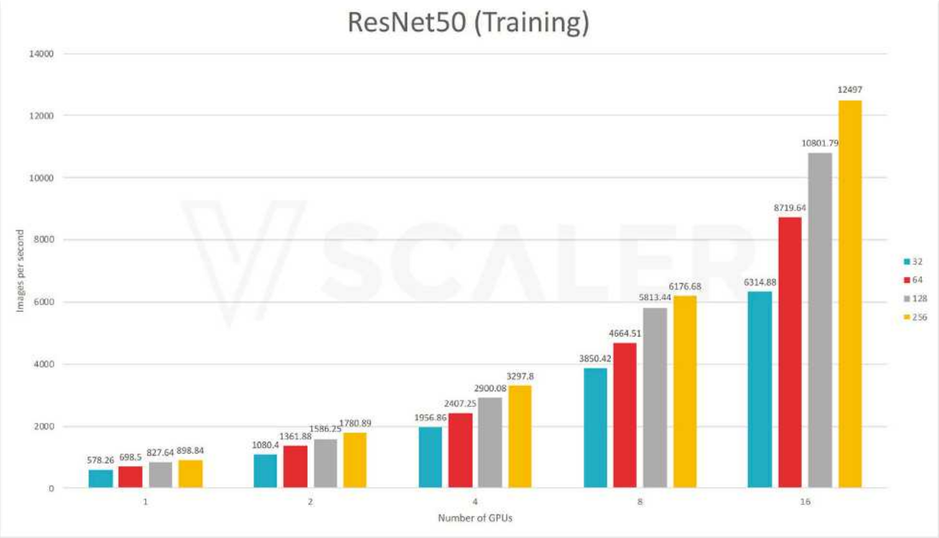


Fig 12. ResNet50 (Training) benchmark results

>5k images per second

ResNet152(Training)

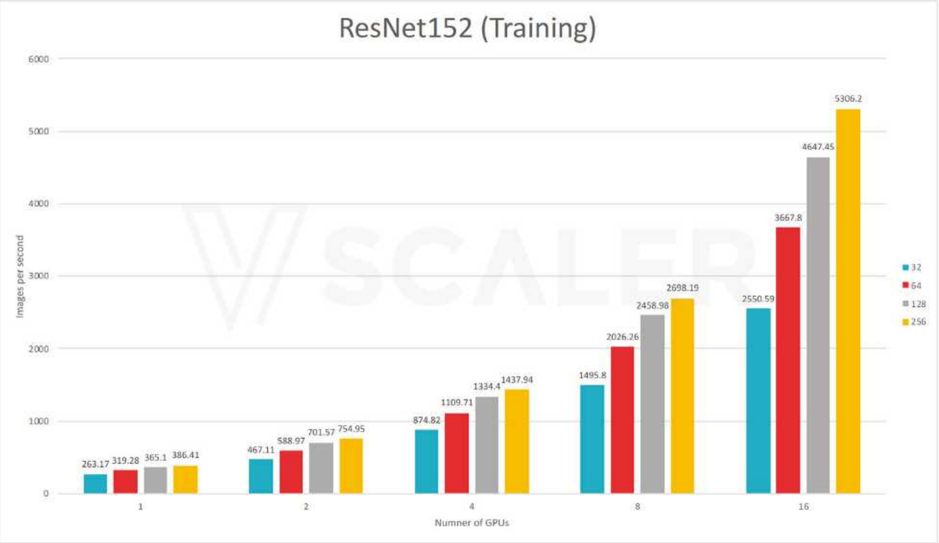


Fig 13. ResNet152 (Training) benchmark results

ResNet Training benchmarking shows that the system scales very well from 1 to 16 GPUs with almost linear performance improvements as we increase the number of GPUs. Its also interesting to note that the batch size also plays an important part in ensuring that the GPUs are kept fully utilised to maximise performance with results favouring larger batch sizes. Training can be the most computationally demanding part of the AI lifecycles so matching that need with an appropriate sized infrastructure is important.

>15k

images

per second

ResNet50(Inference)

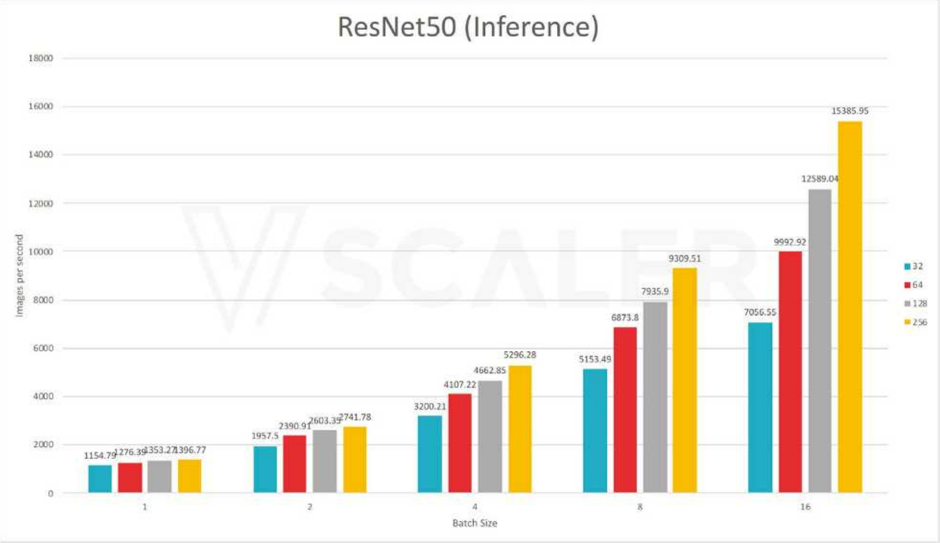


Fig 14. ResNet50 (Inference) benchmark results

>7k

images

per second

ResNet152(Inference)

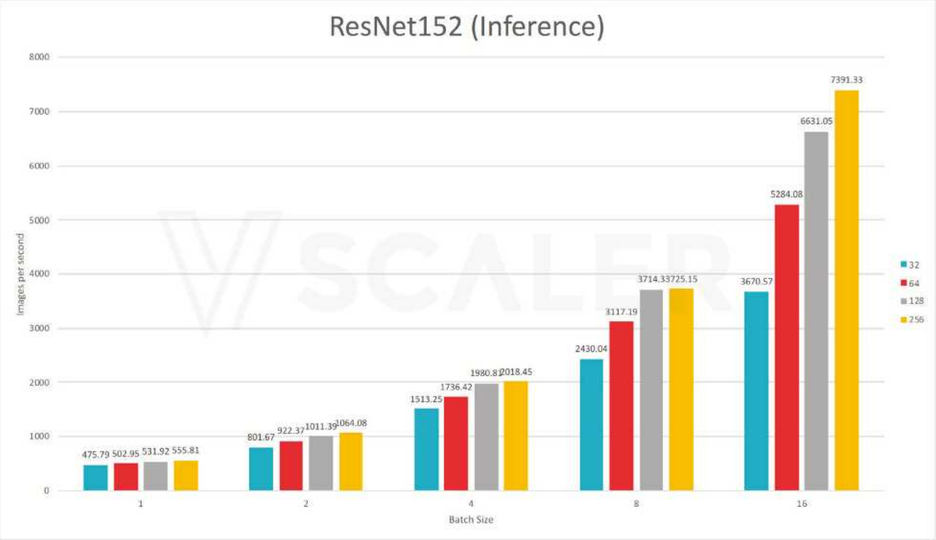


Fig 15. ResNet152 (Inference) benchmark results

The numbers for inference confirm similar results to the training numbers which is important for a number of reasons, firstly this confirms that the solution is flexible and can be used for mixed workloads (training or inference) which is useful at difference stages of the AI lifecycle.

CONCLUSION

With AI becoming more prevalent in almost every industry, it's clear that a platform that can manage the complete AI development life cycle will become more and more critical for organisations. Whether you are starting out on your AI journey or looking to scale an existing infrastructure, vScaler can help with simplifying the configuration and management of all software and storage. vScaler has a number of key design principles:

vScaler enables users to focus on the science, not on the administration.
AI product development and deployment can be a challenging task – not least in the administration and configuration of the software stack and management of the storage backend. The focus of an organization should be on the output of the platform, and not the administration of it. Our platform is designed to minimize the administration burden and allow organisations to focus higher up the stack with an 'as-a-service' model solution where resources are provisioned and configured on demand for end users.

Start small, scale as needed (within a DC, across geographies or to public cloud)
One of our observations from working in the field is that customers usually want to start off with a modest configuration while they explore and understand how AI can benefit the product/organization. vScaler allows you to start with a small HCI (Hyper Converged Infrastructure) and then scale as requirements grow. vScaler can be scaled out within the DC, across geographically distributed organisations and/or out to the public cloud for access to unlimited resources as and when they are required.

Secure Multi-tenant environment and storage.
Certain projects or data sets may need to be securely handled and restricted in terms of access. vScaler's multi-factor and biometric authentication systems can allow strict policies to be placed on the data sets and resources within an organization.

AI will launch the next wave of technology innovations and take us to places we've never dreamed of. The fourth industrial revolution is on our doorstep, and AI will be a driving force behind its arrival. Let vScaler ease your transition to the cloud by allowing us to manage your infrastructure requirements while you focus on delivering world class software and products. Start your AI and Cloud journey today.

Please visit www.vscaler.com or contact your Boston Limited sales representative for more information.

WWW.VSCALER.COM