VSCALER AI REFERENCE ARCHITECTURE

A scalable cloud infrastructure

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 dynamic nature of these AI environments typically means that the datasets are constantly growing and algorithms are being optimised for greater accuracy. An infrastructure that can handle growth seamlessly and grow as your requirements grow is crucial.

VSCALER™ AI ARCHITECTURE

vScaler - an optimised cloud platform built with AI and Deep Learning workloads in mind - provides you with a production ready enivronment with integrated Deep Learning application stacks and optimised NVMe storage, eliminating the administrative burden of setting up these complex AI environments manually.

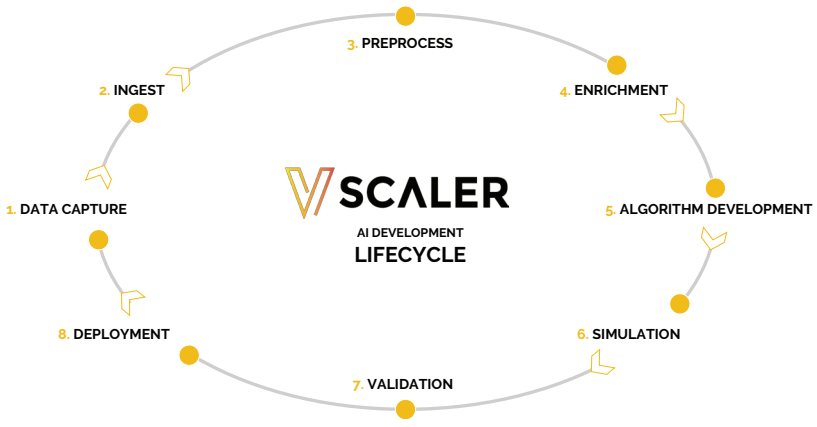

Figure 1. vScaler can manage all steps in a typical AI Development Lifecycle

"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

NVIDIA VSCALER

VSCALER STORAGE

With vScaler Storage we centralise all I/O 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.

THE NVIDIA® DGX-2™ SERVER

Increasingly complex AI demands unprecedented levels of compute. NVIDIA DGX-2 - the world's first 2 peta-FLOPS system, packing the power of 16 of the world's most advanced GPUs, has been integrated in the vScaler architecture as the building block for computationally intensive workloads. In comparison with legacy x86 architectures, DGX-2's ability to train ResNet-50 would require the equivalent of 300 servers with dual Intel® Xeon® Gold CPUs costing over $2.7 million dollars.

MELLANOX® RDMA ACCELERATED FABRIC

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.

VSCALER APPLICATION ENVIRONMENT

Spin up application specific environments with the appropriate Deep Learning frameworks installed and ready for use, including TensorFlow, Caffe and Theano* orchestrated on individual VMs or in a scale-out manner using Kubernetes. 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.

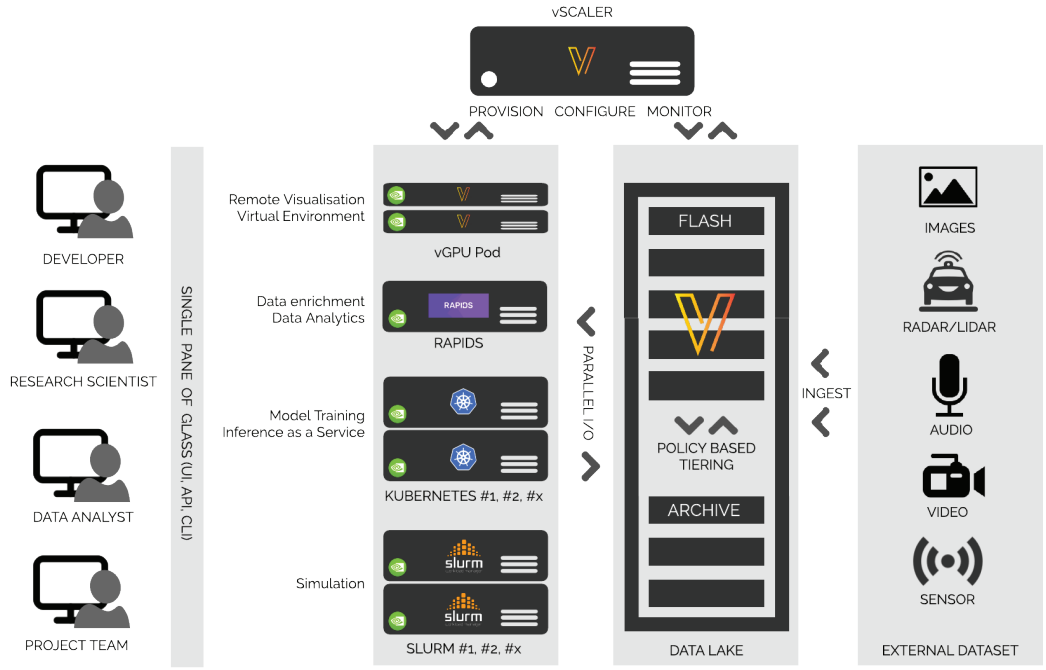

Figure 2. vScaler AI Workflow

For More information,

Contact Boston Ltd.

BOSTON
Servers | Storage | Solutions

sales@boston.co.uk | www.boston.co.uk | +44 (0) 1727 876 100