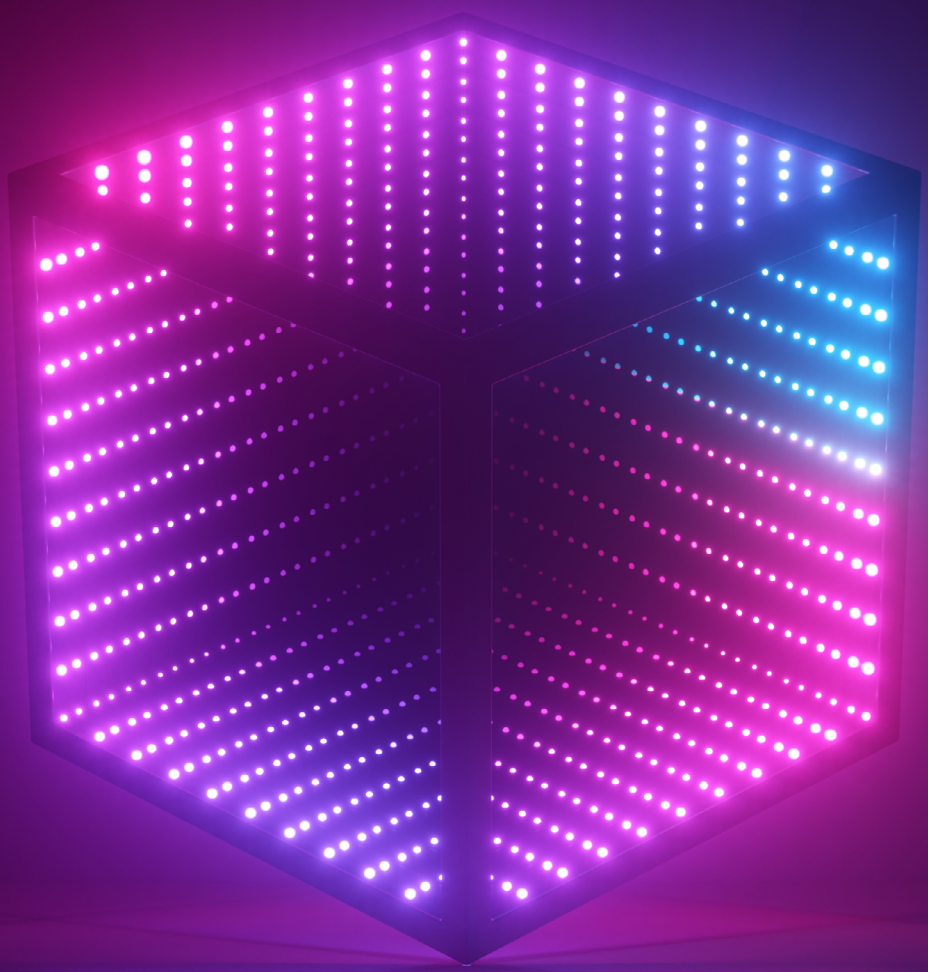
WEKA | GRAPHCORE

Graphcore IPU Based Systems With Weka Data Platform

April 2022

Reference Architecture



Summary
and Revisions

This white paper provides a technical overview of the reference architecture with Graphcore next generation IPU (Intelligence Processing Units) based compute systems and Weka Data Platform for AI, based on WekaFS™ filesystem. This document highlights the features and management, and independent performance validation of the joint solution. The combined solution lets innovators create the next breakthroughs in machine intelligence to enhance human potential in Finance, BioTech, Scientific Research and Consumer Internet markets.

Revisions

Date	Description
Mar. 2022	1.0 Initial release

Contents

1 Introduction

Data and AI teams need simple yet powerful infrastructure in place to take ideas from experimentation to production rapidly. They need end-to-end infrastructure that provides a performant platform that is both easy to set up and simple to use. Graphcore and WEKA have brought intelligent compute and storage together to create a converged infrastructure solution to serve machine learning workloads of all sizes while maintaining simplicity and performance at scale.

Graphcore provide machine intelligence compute systems. These are built around the Intelligence Processing Unit (IPU), a new processor specifically designed for machine learning compute. The IPU's unique architecture enables the exploration and deployment of entirely new types of workloads, to drive advances in machine intelligence.

Storage is a crucial component of a deployable datacentre solution. Choosing and optimally configuring the right storage components is critical for users that want efficient and reliable infrastructure solutions for machine learning.

This document describes an example reference architecture which has been developed by WEKA and Graphcore in partnership. This architecture enables system deployers to configure optimal solutions so they can extract maximum performance and value from their IPU Pod configurations as they build and scale their AI compute capability.

The information in this document applies to Graphcore IPU Pod systems, which covers both classic IPU-POD™ systems (such as the IPU-POD$_{64}$) and Bow™ Pod systems (such as Bow Pod$_{64}$). The term IPU-Machine refers to the blades installed in your system, so IPU-M2000™ in IPU-POD™ systems and Bow-2000™ in Bow Pod systems.

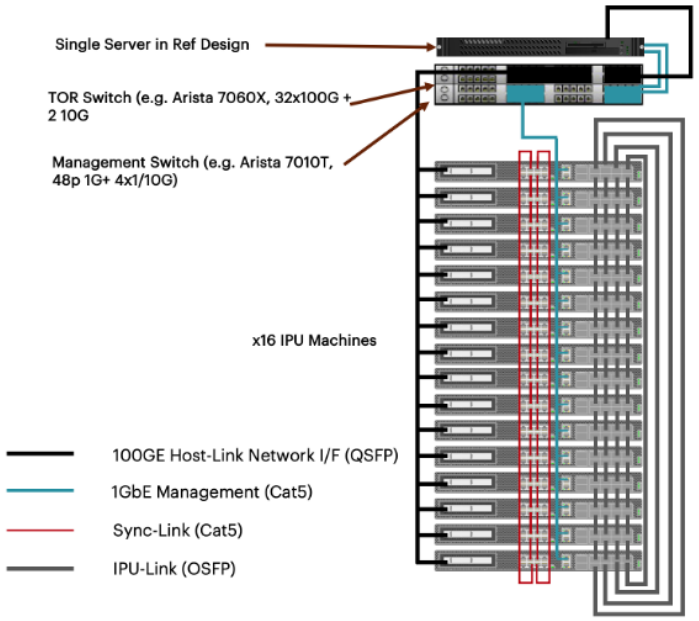
All integration and set-up information in this document is also valid for new systems built using Bow-2000 IPU-Machines, however performance with Bow-2000s will be up to 40% higher.

2 Graphcore IPU Pod Overview

IPU Pod systems are created by connecting multiple IPU-Machines, third party CPU server units, and storage appliances, allowing powerful and flexible AI infrastructure designs for machine intelligence training and inference workloads.

2.1 Innovate at Massive Scale

The core building block of an IPU Pod is the IPU-Machine, contained in a slim 1U blade. This is the fundamental compute engine for machine intelligence from Graphcore, containing four IPU processors, accelerators designed from the ground up for AI. An individual IPU-Machine can deliver up to 1 petaFLOPS (IPU-M2000) or 1.4 petaFLOPS (Bow-2000) of AI compute and has up to 260GB Memory (3.6GB In-Processor-Memory™ and up to 256 GB Streaming Memory™), enabling it to handle the most demanding of machine intelligence workloads.

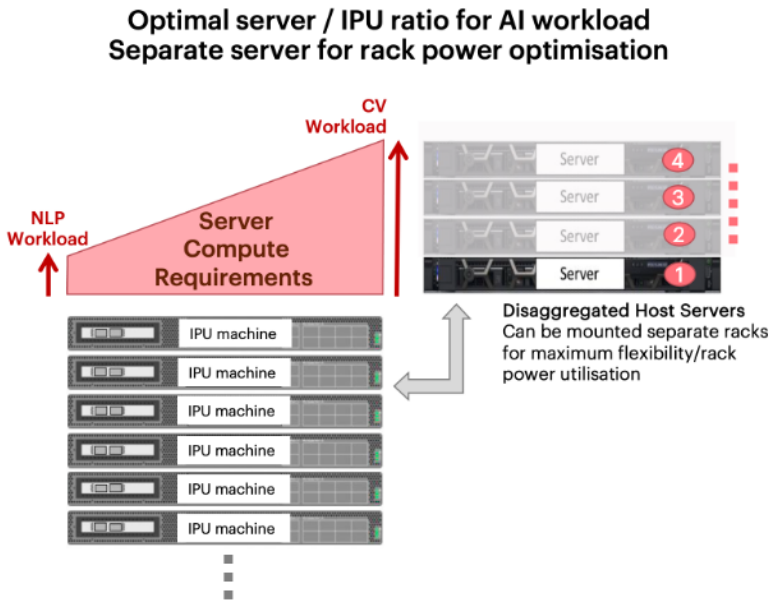


The IPU-Machine has a flexible, modular design, so you can start with one and scale to thousands. IPU-Machines can work as standalone systems or can be interconnected in racks. Individual IPU Pod racks can then be interconnected with other Pods using the 2.8Tbps high-bandwidth, near-zero latency IPU-Fabric™ interconnect architecture, to grow to supercomputing scale.

The classic IPU-POD$_{64}$ reference design is a rack solution containing 16 IPU-M2000 IPU-Machines, one to four host servers (the default is one host server in the reference configuration), network switches and platform software. It is designed to deliver 16 petaFLOPS of AI compute in an efficient, flexible and pre-qualified configuration. The equivalent Bow based system (Bow Pod$_{64}$) adopts the same form factor and architecture, with 16 Bow-2000 IPU-Machines, delivering an increase in performance of up-to 40% whilst reducing energy consumption by up-to 16%.

2.2 Disaggregated to Scale With Your Needs

AI workloads have very different compute demands. For production deployment, optimizing the ratio of AI to host compute can maximize performance and efficiency, and improve the total cost of ownership. IPU Pods are disaggregated systems, this means that the ratio between the number of host servers and switches and the number of IPU-Machine units is not fixed. The system can be built so that it is ideally matched to the production workload.



For example, NLP models require very little server CPU interaction and utilization, whereas CNN-based workloads such as CV require a larger proportion of scalar computing and would benefit from more server CPU being involved. The system can be tailored for the workload.

2.3 Unmatched Scale-Out With IPU-Fabric

IPU-Fabric is Graphcore's innovative low-latency, all-to-all IPU interconnect. Eliminating communication bottlenecks with reliable deterministic performance it is highly efficient whatever your scale.

2.4 Datacentre Compatibility

IPU Pod systems bring together powerful IPU compute with a choice of best-in-class datacentre technologies and systems from leading technology providers in flexible, pre-qualified configurations, to ensure your datacentre is operating with maximum efficiency and performance, while making your datacentre AI deployments simpler and faster.

In this document, the WEKA limitless data platform is evaluated as a storage solution to support AI workloads running on an IPU-POD$_{64}$.

3 WEKA overview

WEKA offers a limitless data platform that was built to address the storage challenges posed by modern machine learning applications. The software-defined architecture is secure, robust and operating system-agnostic with flexible deployment options. The limitless data platform is built on the WEKA file system (WEKAFS) which can be run on any commodity Intel x86-based hardware platform. WEKAFS is available in the AWS cloud, as well as many storage server platforms from HPE, Hitachi Vantara, Lenovo, Dell, and Cisco.

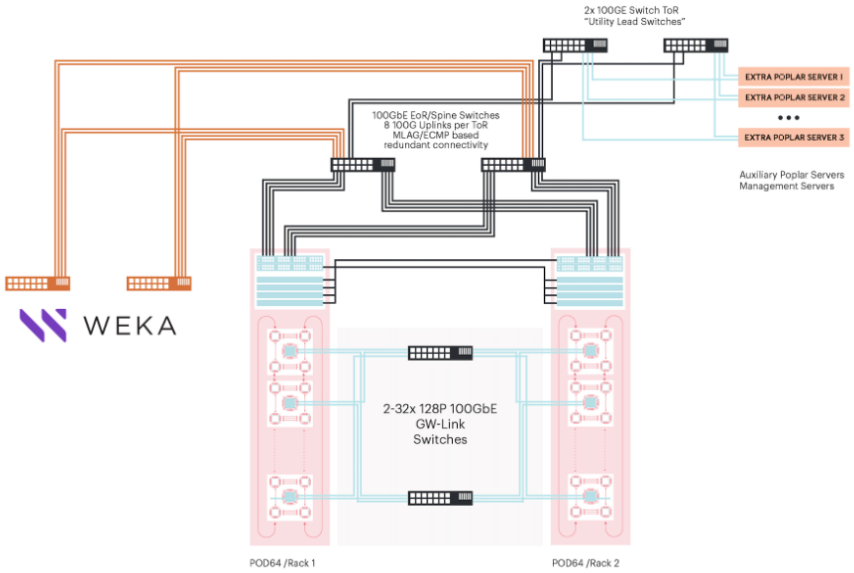
WEKAFS delivers on multi-tenancy, multi-protocol support, end-to-end encryption, authenticated mounts, hybrid cloud management, snapshots, cloud bursting, archive, backup and disaster recovery. There is support for fast object use case in addition to high performance filesystem implementation, extending the benefits of parallel, distributed filesystems to cloud native applications which need strong consistency, multi-protocol access as well as high performance and low latency, particularly for small objects.

The WEKA data platform can provide:

- Reduced epoch times while delivering low inference times in a containerized deployment
- Strong consistency and ability to handle small files which is important for fast object use cases such as databases and cloud native applications
- Agility for data management across edge, core, and cloud
- Ability to ingest data for applications needing multi-protocol access
- Scale and ease of deployment – performance scales linearly and WEKA can be installed on any x86 and AMD based storage servers as well as in AWS EC2 instances
- Leverage of hybrid media, providing fast object over NVMe flash and capacity object with HDD; burst workloads in cloud as needed
- A single global namespace consolidating storage and data silos; unifies your data with a single storage platform for entire data pipeline, allowing easy access and management of billions of files from a single directory
- Hybrid cloud functionality – WEKA extends native S3 access both on-premises as well as for AWS deployments, uniquely extending AWS S3 for fast object use cases and frameworks such as AWS Sagemaker

4 Reference Architecture

This section describes the hosts, storage and networking configuration used in the IPU-POD$_{64}$ reference architecture featuring a WEKA storage solution. All information in this section is also applicable to a Bow Pod$_{64}$ system.



4.1 Poplar Hosts

The Poplar host is the head node for the IPU Pod. For the purpose of this evaluation, the Poplar host consists of the following:

- Hardware: Dell R640 Intel Skylake 768GB RAM
- Mellanox connectX-5 100Gb Ethernet network cards
- Software: Ubuntu 18.04 LTS 5.4.0-66-generic kernel
- Poplar SDK 2.4.0 application environment

4.2 Network Connections

The network configuration recommended for IPU-POD$_{64}$ and Bow Pod$_{64}$ systems is as follows:

- 1G management network: Arista DCS-7010T-48-F
- Top-of-Rack for 100G connectivity: Arista DCX-7060CX-32S-ES-F
- Connected via 8way 100G LAG to: Arista 7060PX4-32-F2 as SPINE 400G

4.3 Storage

To support the IPU-POD$_{64}$, a WEKA storage solution was architected as follows:

- 8x SuperMicro BigTwin with 2x Intel Xeon Gold 6126 CPUs
- 8x Mellanox Connect-X 6 100Gb network interfaces
- 48x 3.84TB Micron 9300 NVMe

4.4 System Configuration

This section describes the system configuration tested by Graphcore and WEKA for the IPU-POD$_{64}$ reference platform. The information below is also applicable to Bow Pod$_{64}$ systems.

This section includes configuration settings for:

- Poplar host
- Storage
- Networking

4.4.1 Poplar Host Configuration

The Poplar framework can be used to define graph operations and control the execution and profiling of code on the IPU, as well as to configure the host.

A single 100G ethernet interface for the storage connection is configured with a Mellanox OFED 5.1 device driver.

Create a DPDK mount with 4 dedicated cores:

```
mount -t wekafs -o num_cores=4 -o net=ethX 1.1.1.1/default /mnt/weka
```

4.4.2 Storage Configuration

The following is the script required in order to create and configure the cluster for the WEKA storage system. The system was configured in a 6+2 protection scheme, reserving 19 CPU cores for WEKA, utilizing 6 NVMe drives per server and using a single network interface for data access.

```
weka cluster create weka1 weka2 weka3 weka4 weka5 weka6 weka7 weka8 --host-ips 10.103.0.241,10.103.0.24
2,10.103.0.243,10.103.0.244,10.103.0.245,
10.103.0.246,10.103.0.247,10.103.0.248
sleep 10
weka cluster update --data-drives=6 --parity-drives=2 --cluster-name=Weka-Graphcore
for i in {0..7}; do weka cluster host net add $i enp24s0 --netmask=16; done
for i in {0..7}; do weka cluster host dedicate ${i} on; done
for i in {0..7}; do weka cluster drive add $i /dev/nvme{0..5}n1; done
for i in {0..7}; do weka cluster host cores $i 19 --frontend-dedicated-cores=2 --drives-dedicated-
cores=6; done
weka cloud enable
for i in {0..7}; do weka cluster host failure-domain $i --auto; done
weka cluster hot-spare 1
weka cluster host apply --all --force
sleep 90
weka cluster start-io
```

4.4.3 Networking Configuration

The following information refers to the system networking configuration.

Clients connected to the WEKA appliance via VLAN network:

- MTU 9000 on 100G VLAN
- 100G VLAN configured in ToR switch to provide untagged connection to client
- 2x4-way MLAG using Arista MLAG active/active technology from ToR to pair of SPINE switches
- 2x4-way MLAG using Arista MLAG active/active technology from storage LEAF to pair of SPINE switches

5 Performance Results

All benchmarking results in this section are based on the IPU-POD$_{64}$ classic system with IPU-M2000 IPU-Machines.

Performance tests were undertaken using:

- ResNet 50
- BERT Large
- Standardized storage benchmark for AI workloads
- FIO
- S3 storage

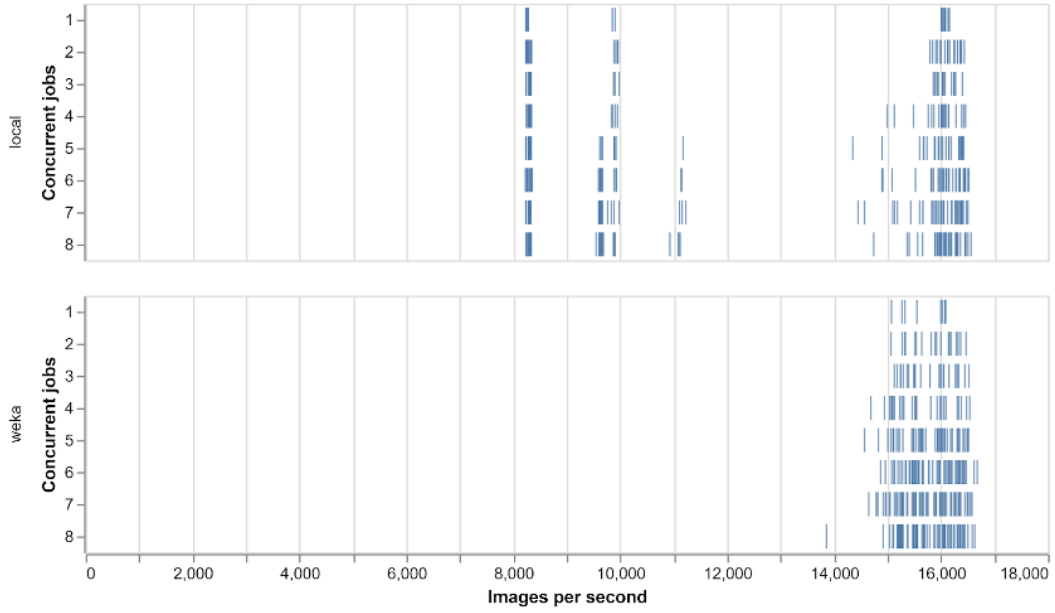
For the ResNet and BERT tests, it is important to keep in mind that the focus is on the IPU performance. The storage solution needs to be fast enough to keep the processors fed with all the data they can consume.

5.1 ResNet 50

ResNet 50 tests were performed using 8 hosts concurrently and produced the following results when compared to local NVMe drives in a RAID6 configuration:

5.1.1 ResNet 50 Performance Tests

The following graphs summarise the benchmark results in images per second.

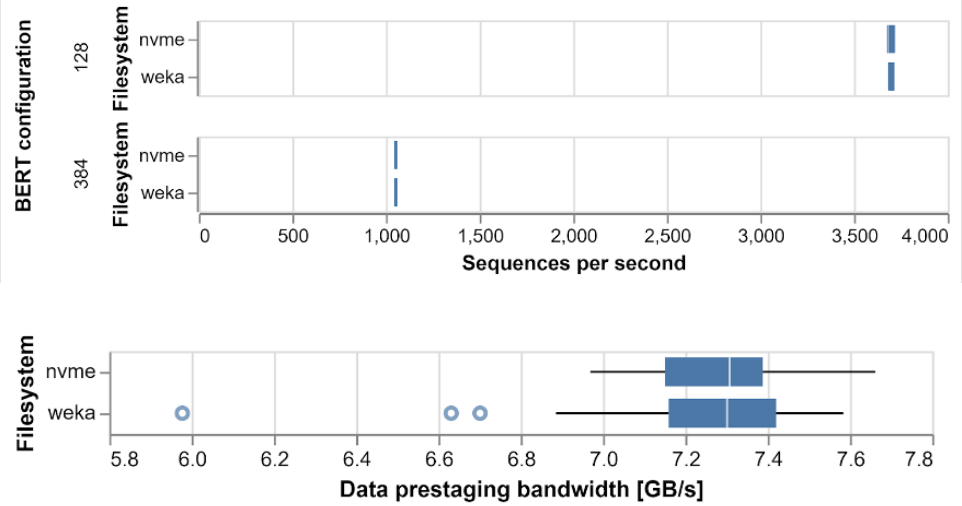


The results illustrate that the performance of the WEKA appliance is reliably around 16k images/sec, and in line with the best performance we have observed using local NVMe.

5.2 BERT Large

BERT Large pretraining using 8 hosts concurrently. The graphs shown are per-host.

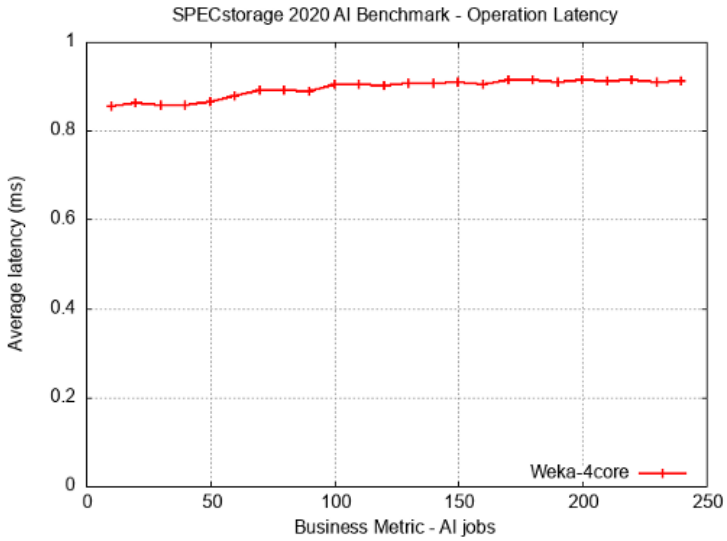
5.2.1 BERT Large Performance Tests



These results show that the performances of the local and WEKA.IO filesystems are comparable. The average bandwidth for each of the two filesystems is just above 7.3GB/s, with a similar distribution around this average. The WEKA.IO results show some outliers (shown as circles in the plot) where the performance dips to around 20% below the average.

5.3 Standardized Storage Benchmark (AI Workload)

The following results refer to the SPECstorage solution 2020 benchmark. The benchmark simulates an AI application and runs an increasing number of concurrent workloads against the storage appliance.



AI Jobs	Storage Ops/sec	Latency (msec)	Storage MB/sec
10	4350	0.856	921
20	8700	0.862	1841
30	13050	0.858	2762
40	17400	0.857	3681
50	21750	0.866	4600
60	26100	0.878	5520
70	30450	0.892	6440
80	34800	0.892	7360
90	39150	0.888	8280
100	43500	0.904	9198
110	47850	0.903	10119
120	52200	0.902	11034
130	56550	0.908	11957
140	60900	0.907	12876
150	65250	0.909	13796
160	69600	0.905	14712
170	73950	0.914	16238
180	78300	0.915	17191
190	82650	0.91	18146

5.4 FIO and Application-Level Sanity-Checks

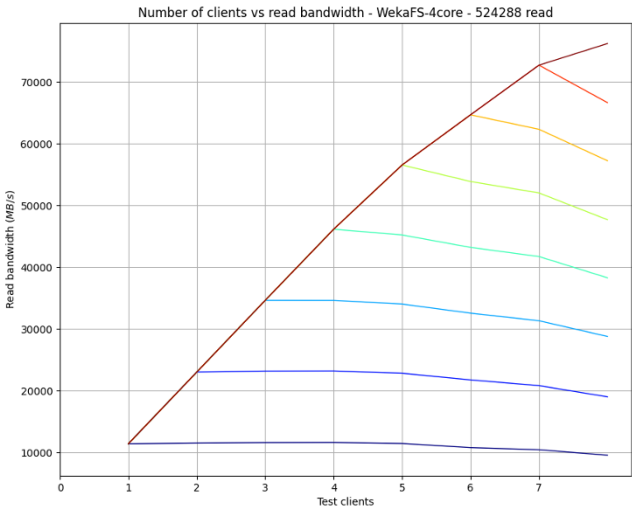
5.4.1 File-Based Tests

Flexible IO Tester (FIO) was used for file-based testing. FIO is a storage benchmarking and workload simulation tool.

For file-based tests, the results were:

- Aggregate peak throughput of 76GB/s using 512KB block random reads, 8 clients and 4 WEKA cores per client; no RDMA in use
- Peak single client performance of single test (as above) is 10GB/s

The following diagram shows the read throughput (MB/s) recorded by increasing the number of clients for the WEKA storage solution connected to the IPU Pod.

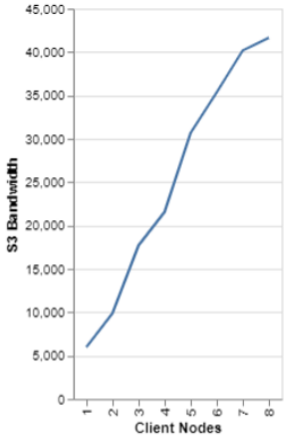


5.5 S3 Object Storage Tests

Amazon Simple Storage Service (Amazon S3) is an object storage service offering industry-leading scalability, data availability, security, and performance. The following section refers to S3 object storage benchmarks (not the storage, just the protocol).

The S3 object storage tests used a block size of 120MB and 48 client threads per server. The test results below show that with 8 storage clients the system is able to read at an aggregate bandwidth of approximately 42 GB/s.

These results show that object storage scales linearly, with a slight drop when using the largest number of client nodes.



6 Conclusion

Artificial Intelligence is a data-centric field where storage requirements are both highly demanding and somewhat idiosyncratic compared to compute as we have known it previously.

Delivering best-in-class performance for machine learning requires highly optimised solutions, co-designed by compute system and storage providers.

Graphcore and WEKA have worked closely together to develop this reference architecture, drawing on deep expertise in their respective fields.

The coming together of the Intelligence Processing Unit (IPU) with WEKA's storage solution, with its wide range of file system protocols, gives users flexibility and choice, with the reassurance that however they configure their setup, it will deliver superlative machine learning performance.

Technology from both these companies - and artificial intelligence in general - are advancing quickly, and the ongoing partnership between Graphcore and WEKA will ensure that every advance and technological refinement we make in future will be in close collaboration, for the benefit of our mutual customers and their evolving machine learning needs.

Contact Graphcore and WEKA sales for more information.

WKA315-01 03/2022