Services for Machine Learning (Part 1)

CERN School of Computing on IT Services 2024
https://indico.cern.ch/event/1441237/timetable/

Ricardo Rocha, IT-CD

Cluster

```
ssh lxplus.cern.ch

openstack coe cluster create --cluster-template kubernetes-1.30.5-1 --node-count 1 csc-demo

openstack coe cluster list

openstack coe cluster config csc-demo

export KUBECONFIG=`pwd`/config
```

Sessions

Services for Machine Learning (Part 1), Ricardo Rocha

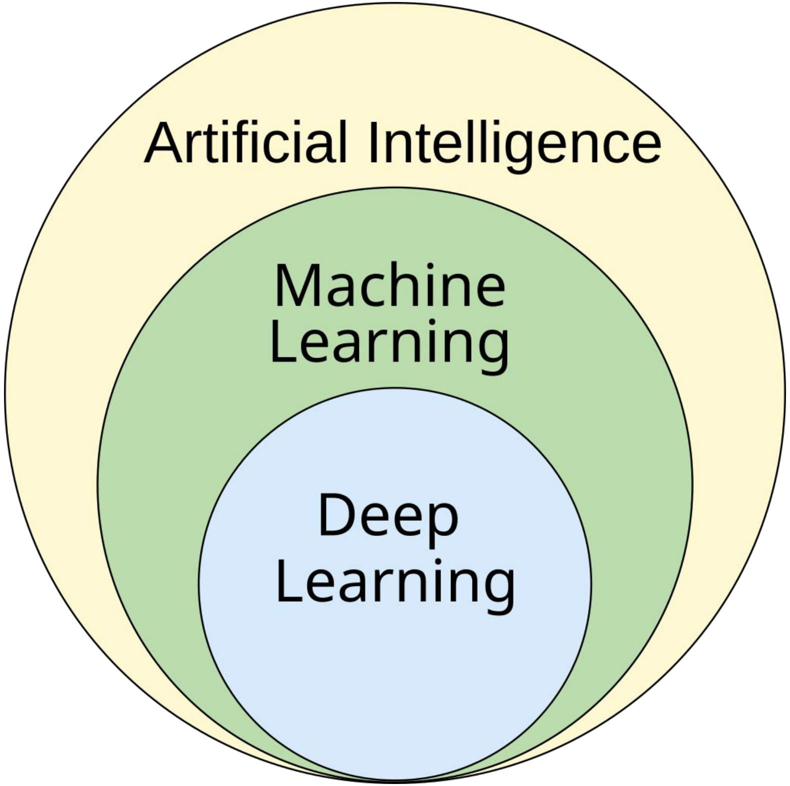
Available services and hardware, use cases, containerization, scaling

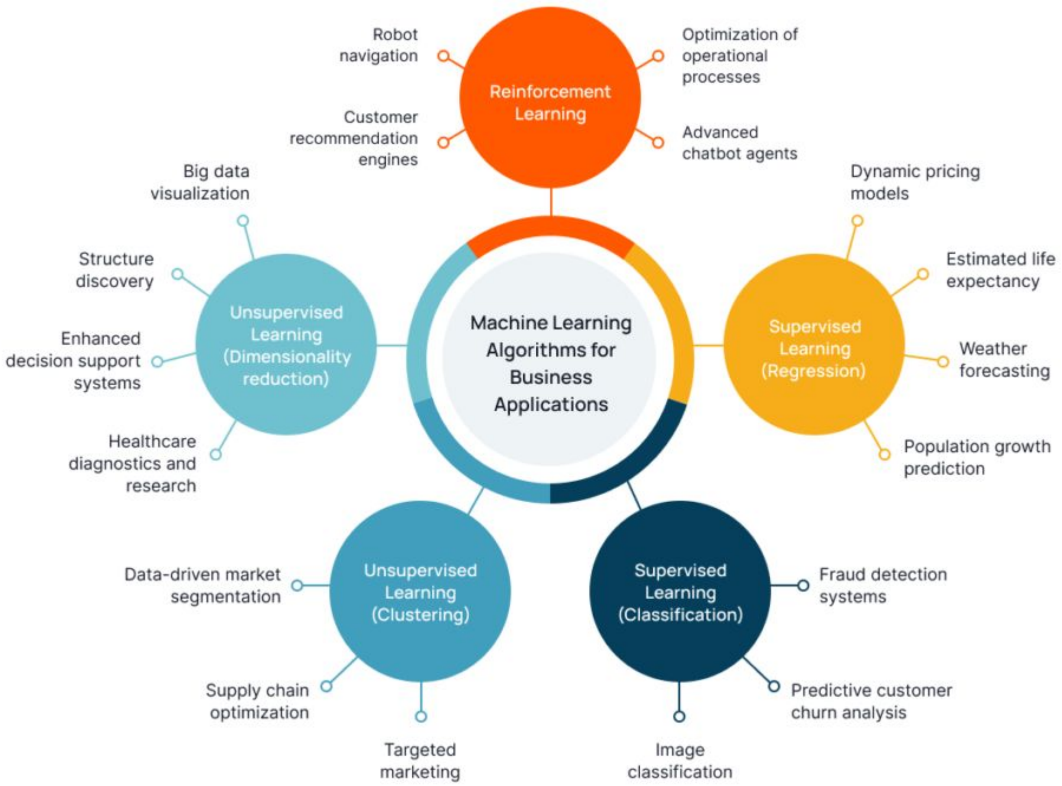
Services for Machine Learning (Part 2), Diana Gaponcic

Efficient usage of shared GPU resources, partitioning, slicing, …

Services for Machine Learning (Part 3), Raul Chiorescu

Managing your ML lifecycle with ml.cern.ch


Artificial Intelligence
Machine Learning
Deep Learning


Machine Learning Algorithms for Business Applications
Reinforcement Learning
Robot navigation
Customer recommendation engines
Optimization of operational processes
Advanced chatbot agents
Supervised Learning (Regression)
Dynamic pricing models
Estimated life expectancy
Weather forecasting
Population growth prediction
Supervised Learning (Classification)
Fraud detection systems
Predictive customer churn analysis
Image classification
Unsupervised Learning (Clustering)
Data-driven market segmentation
Supply chain optimization
Targeted marketing
Unsupervised Learning (Dimensionality reduction)
Big data visualization
Structure discovery
Enhanced decision support systems
Healthcare diagnostics and research

MISTRAL
AI_

OpenAI

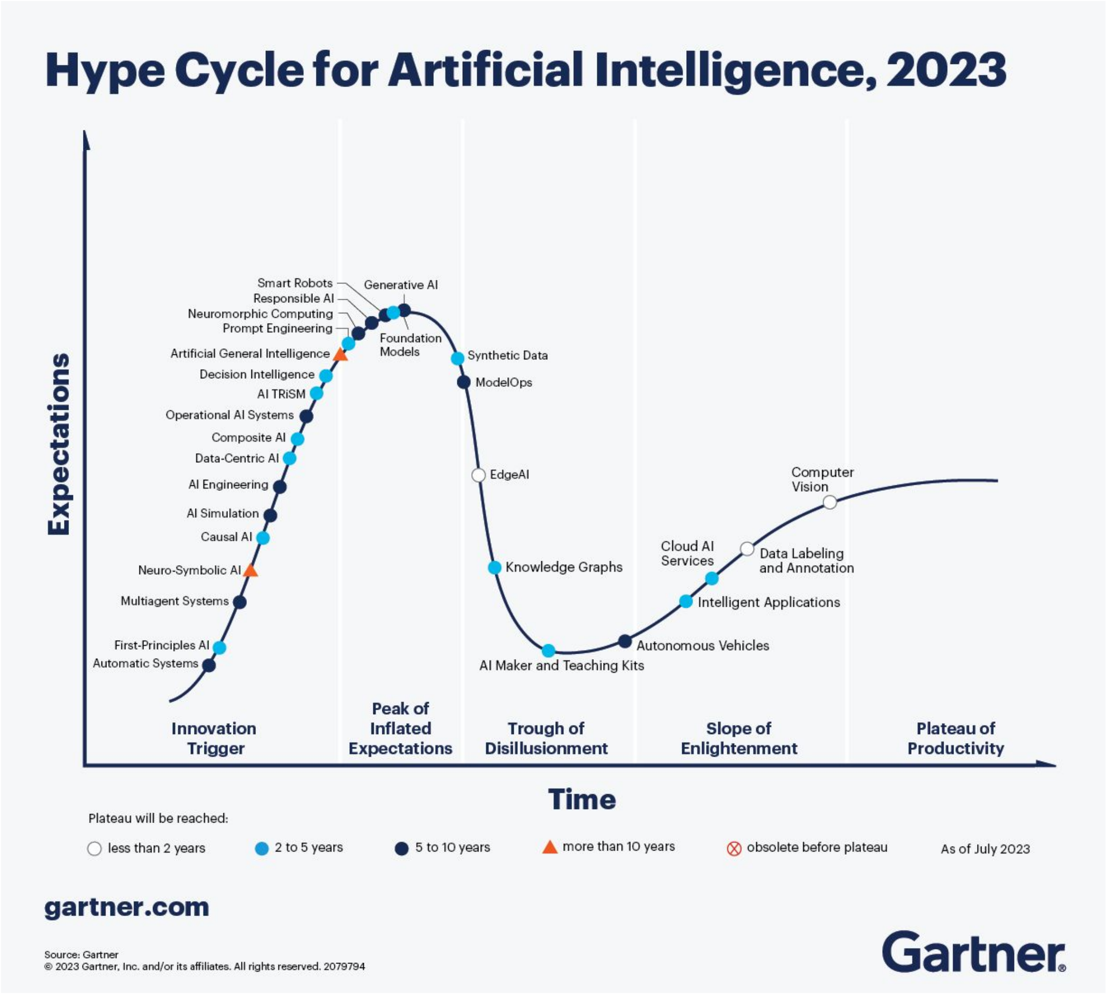
November 30, 2022

Introducing ChatGPT

Try ChatGPT ↗ Download ChatGPT desktop › Learn about ChatGPT ›

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.


Hype Cycle for Artificial Intelligence, 2023
Expectations
Smart Robots
Responsible AI
Neuromorphic Computing
Prompt Engineering
Generative AI
Foundation Models
Artificial General Intelligence
Decision Intelligence
AI TRiSM
Operational AI Systems
Composite AI
Data-Centric AI
AI Engineering
AI Simulation
Causal AI
Neuro-Symbolic AI
Multiagent Systems
First-Principles AI
Automatic Systems
Synthetic Data
ModelOps
EdgeAI
Knowledge Graphs
AI Maker and Teaching Kits
Autonomous Vehicles
Intelligent Applications
Cloud AI Services
Data Labeling and Annotation
Computer Vision
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
Plateau will be reached:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
As of July 2023
gartner.com
Source: Gartner
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. 2079794
Gartner

Use Cases

09:00 → 09:10 **Introduction** 10m

Speakers: Ricardo Rocha (CERN), Dr Sofia Vallecorsa (CERN)

2nd IT ML infrastruc...

09:10 → 09:30 **ATLAS** 20m

Speaker: Daniel Thomas Murnane (Lawrence Berkeley National Lab. (US))

ATLAS ML Resourc...

09:30 → 09:50 **CMS** 20m

Speaker: Davide Valsecchi (ETH Zurich (CH))

23_10_10 - CERN IT ...

09:50 → 10:10 **ALICE** 20m

Speaker: Fabio Catalano (CERN)

CERN_IT_workshop...

10:10 → 10:30 **LHCb** 20m

Speaker: Simon Akar (University of Cincinnati (US))

IT_Workshop_LHCb....

10:30 → 10:50 **ATS** 20m

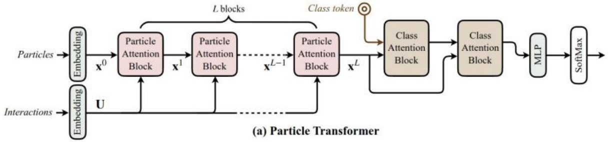
Speaker: Verena Kain (CERN)

IT_ML_ATS_11Oct2...

Jet taggers

Jet tagging = playground for ML architectures

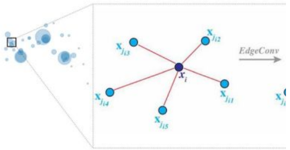
- 2 leading models emerged in CMS: *ParticleNet* (graph conv.), *ParticleTransformer* (ParT)



(a) Particle Transformer

- Probably we still haven't reached the *ultimate* performance, but CMS focus is also shifting more on consolidation:
    - energy / pT calibration for jet and taus
    - working on jet flavor, tau decay mode and lepton tagging
    - Exploring adv attack and data adaptation to improve stability and minimize efficiency corrections

ParticleNet

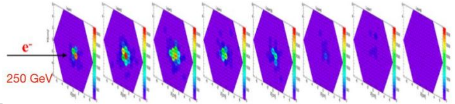


End-to-end reconstruction

Heavy R&D on end-to-end high granularity calorimeter (HGCAL) reconstruction arxiv2204.01681

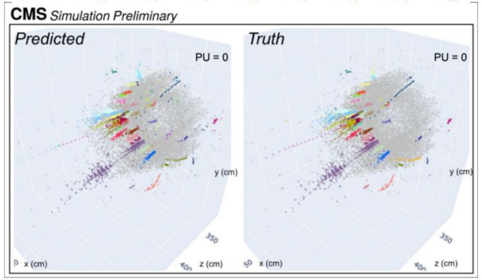
From **hits → clusters position and properties:** perfect application of GNNs and *object condensation.*

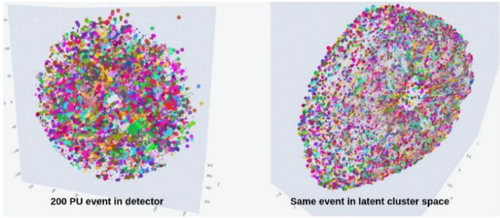
- GravNet architecture with dynamical graph building applied successfully arxiv 2106.01832



Challenges:

- large input dimensionality and non-regular data structure
- Implemented custom kNN kernel for in-memory operations (200x faster than pytorch geometric
- Multi-GPU training neede to speed up prototyping

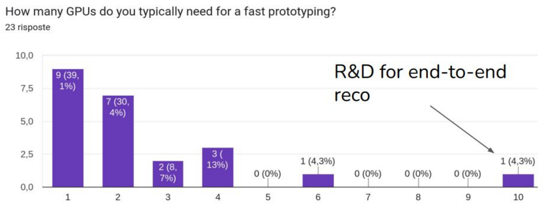




Training resources - survey

ML development in CMS is carried out independently by many groups: no central training infrastructure in place

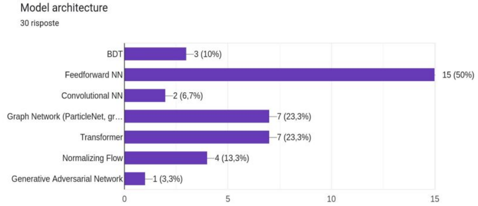
- Analysis, reconstruction, trigger, DQM, anomaly detection, simulation → **many different requirements**
- Organized a **survey** to collect feedback about **training resources:**
  - ~small amount of answers (30 for now) but **main R&D efforts included**
  - the bulk of the distribution (model training for analysis) is still not covered
- Investigated GPU resources needed, resources provider, frequency of retraining, etc

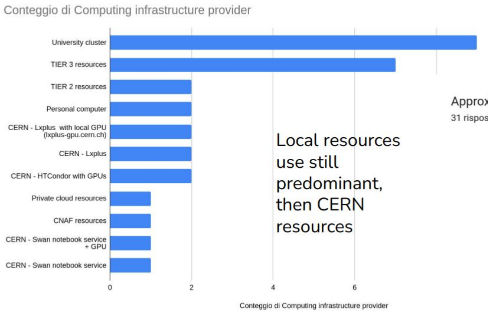


Training resources - survey

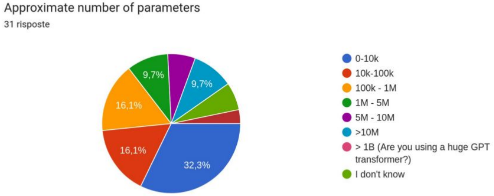
Most of the ML efforts haven't still performed an extensive hyper-parameters optimization due to **resources limitation.**

Modern architectures (graph, transformer, NF), which are heavier to train, are now common





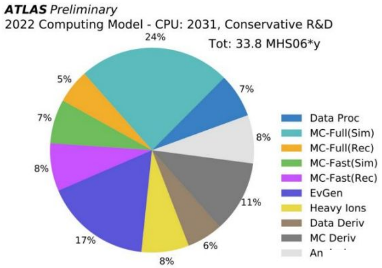
Local resources use still predominant, then CERN resources



Hardware: GPUs

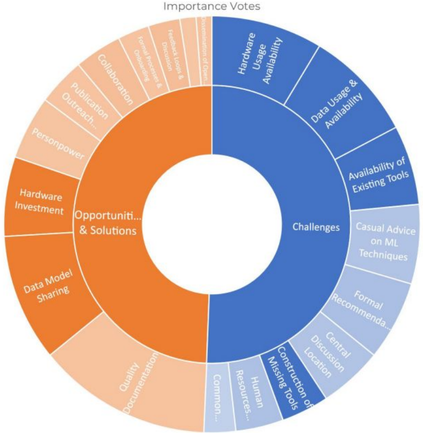
Production Inference

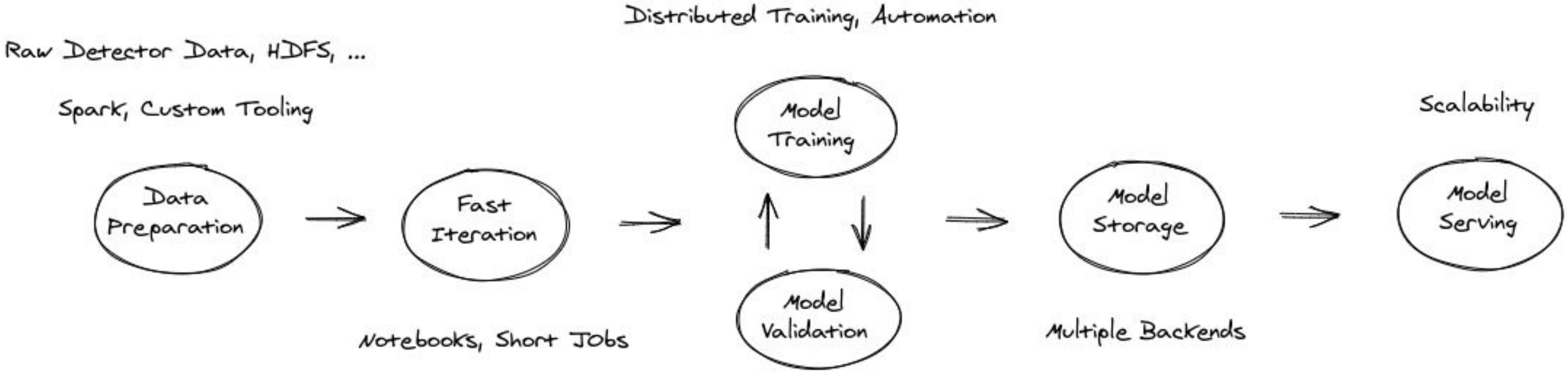
- Consider that the bulk of MC simulation and reconstruction could be targeted by GPU-based ML solutions, and some part of analysis and derivations could be accelerated with GPU ML
- We could estimate that O(50%) of the ATLAS computing model could be accelerated by GPU-based ML by 2031



ML Infrastructure C&Os

- Approximately 30% of interest is relevant to ML hardware, software & tooling
- Can be mostly summarized as **GPU access** & **data access**

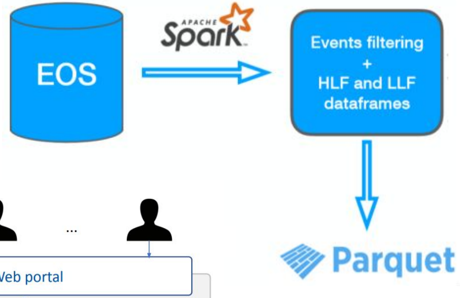


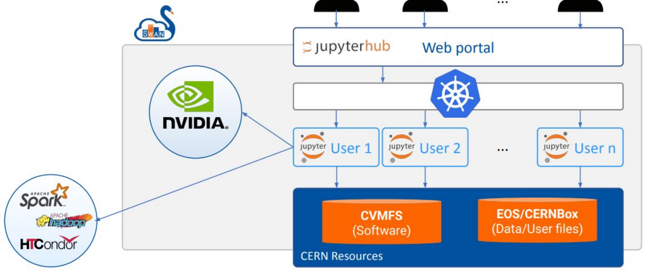

Distributed Training, Automation
Raw Detector Data, HDFS, ...
Spark, Custom Tooling
Scalability
Data Preparation
Fast Iteration
Model Training
Model Validation
Model Storage
Model Serving
Notebooks, Short Jobs
Multiple Backends

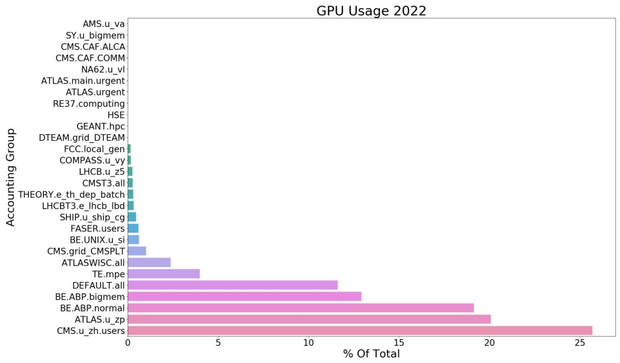
Services

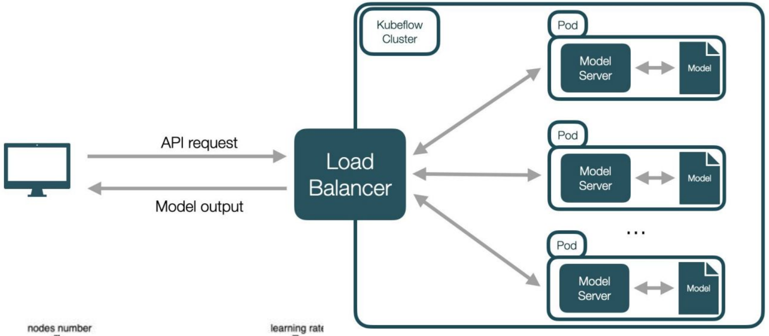
CERN IT ML Infrastructure Workshop

09:00	**Welcome and Introduction** *31/3-004 - IT Amphitheatre, CERN*	*Ricardo Rocha et al.*	09:00 - 09:10
	SWAN *31/3-004 - IT Amphitheatre, CERN*	*Diogo Castro et al.*	09:10 - 09:35
	Batch and Lxplus *31/3-004 - IT Amphitheatre, CERN*	*Laurence Field*	09:35 - 10:00
10:00	**Spark Ecosystem for Machine Learning** *31/3-004 - IT Amphitheatre, CERN*	*Luca Canali*	10:00 - 10:25
	Coffee Break *31/3-004 - IT Amphitheatre, CERN*		10:25 - 10:40
	Kubeflow - ml.cern.ch *31/3-004 - IT Amphitheatre, CERN*	*Dejan Golubovic et al.*	10:40 - 11:05
11:00	**Public Cloud** *31/3-004 - IT Amphitheatre, CERN*	*Dr Sofia Vallecorsa*	11:05 - 11:30

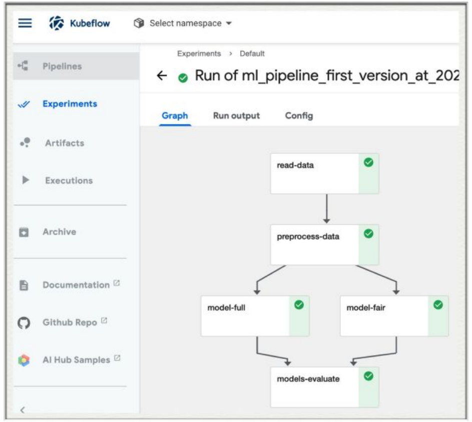


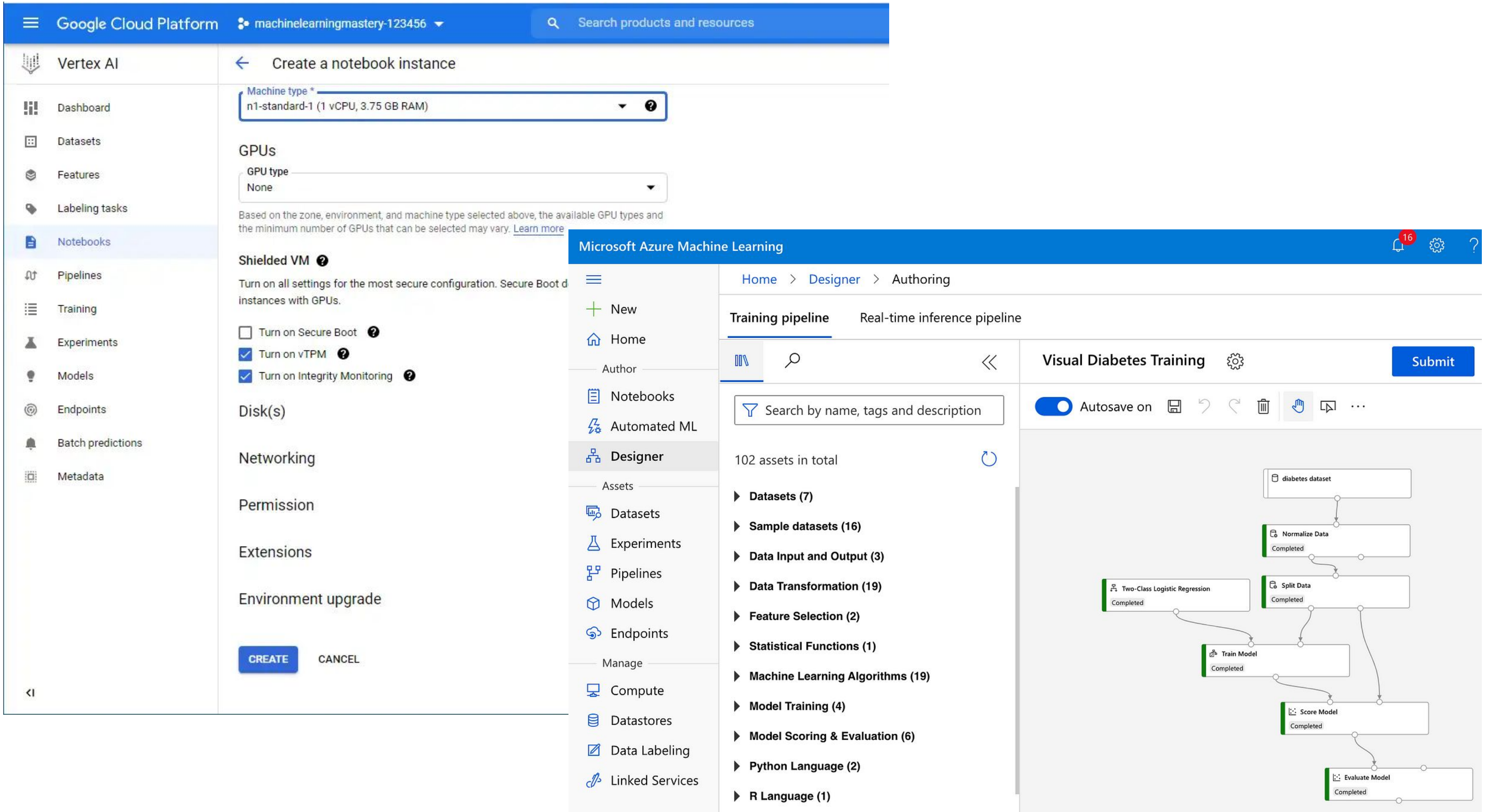






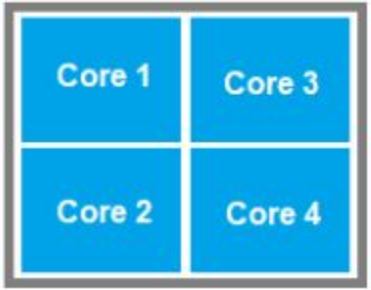
Trial name	Status	Test acc	Nodes number	Learning rate
test-wze6q-bmmwpbc	Succeeded	0.99593	32	0.001
test-wze6q-kqvz9zcp	Succeeded	0.98831	32	0.01
test-wze6q-lnbhttzs	Succeeded	0.98932	32	0.0001
test-wze6q-qvhz6pm8	Succeeded	0.99796	64	0.001

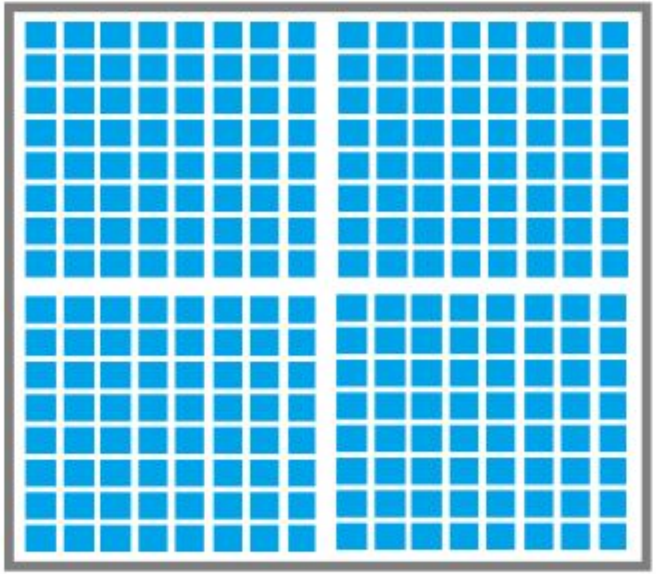




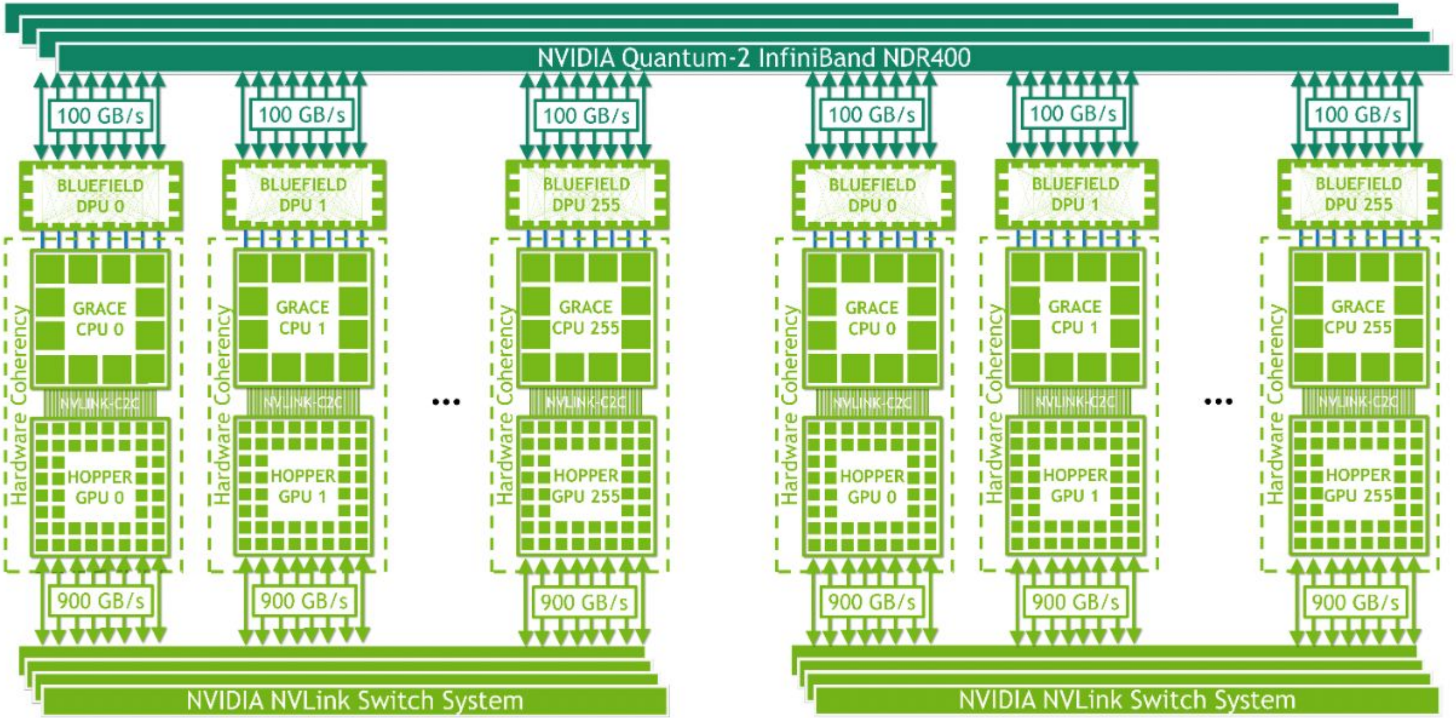
Google Cloud Platform
machinelearningmastery-123456
Search products and resources
Vertex AI
Create a notebook instance
Dashboard
Datasets
Features
Labeling tasks
Notebooks
Pipelines
Training
Experiments
Models
Endpoints
Batch predictions
Metadata
Machine type *
n1-standard-1 (1 vCPU, 3.75 GB RAM)
GPUs
GPU type
None
Based on the zone, environment, and machine type selected above, the available GPU types and the minimum number of GPUs that can be selected may vary. Learn more
Shielded VM
Turn on all settings for the most secure configuration. Secure Boot d
instances with GPUs.
Turn on Secure Boot
Turn on vTPM
Turn on Integrity Monitoring
Disk(s)
Networking
Permission
Extensions
Environment upgrade
CREATE
CANCEL
Microsoft Azure Machine Learning
16
New
Home
Author
Notebooks
Automated ML
Designer
Assets
Datasets
Experiments
Pipelines
Models
Endpoints
Manage
Compute
Datastores
Data Labeling
Linked Services
Home > Designer > Authoring
Training pipeline
Real-time inference pipeline
Search by name, tags and description
102 assets in total
Datasets (7)
Sample datasets (16)
Data Input and Output (3)
Data Transformation (19)
Feature Selection (2)
Statistical Functions (1)
Machine Learning Algorithms (19)
Model Training (4)
Model Scoring & Evaluation (6)
Python Language (2)
R Language (1)
Visual Diabetes Training
Submit
Autosave on
diabetes dataset
Normalize Data
Completed
Two-Class Logistic Regression
Completed
Split Data
Completed
Train Model
Completed
Score Model
Completed
Evaluate Model
Completed

Hardware


Core 1
Core 3
Core 2
Core 4

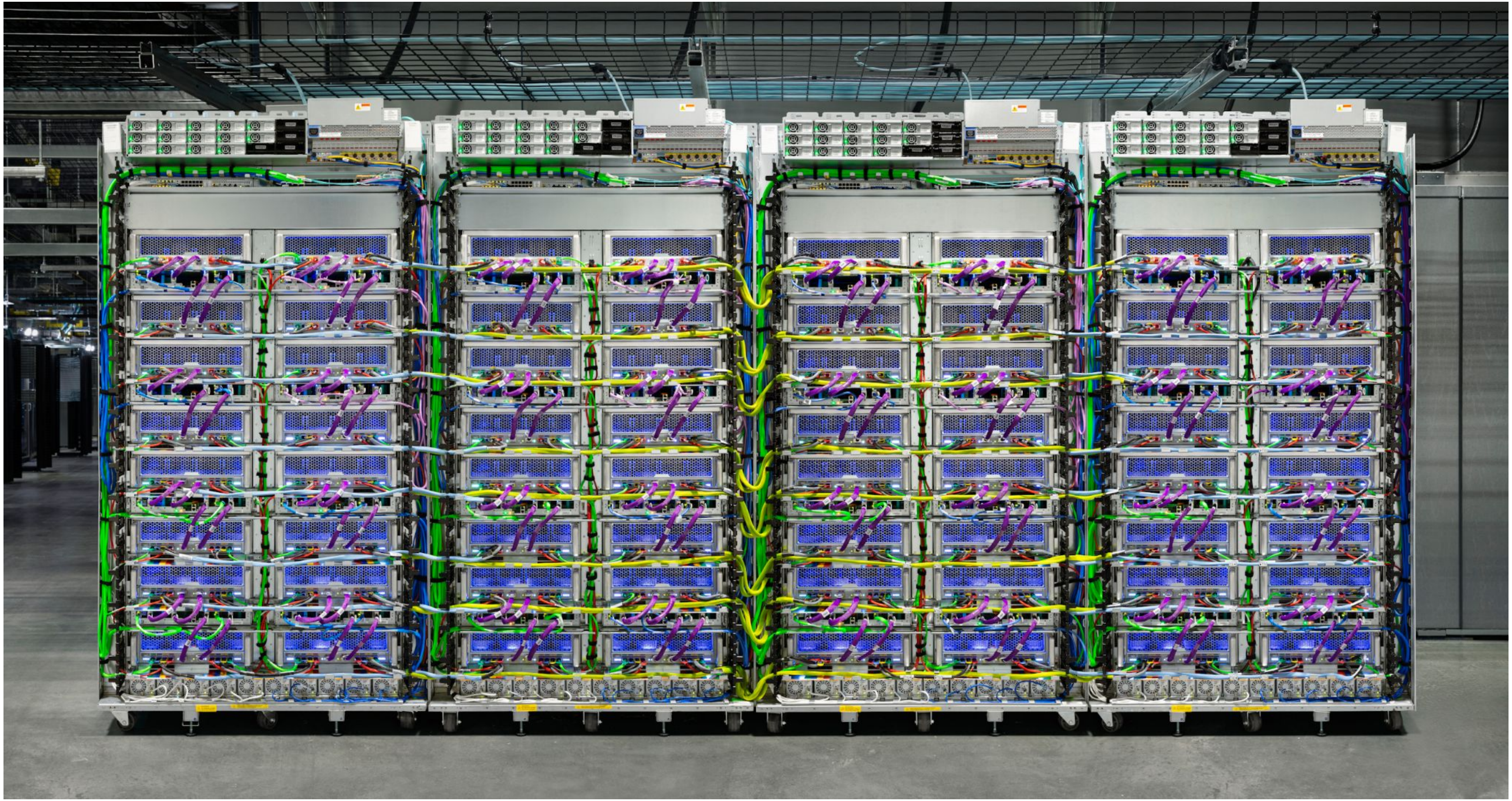


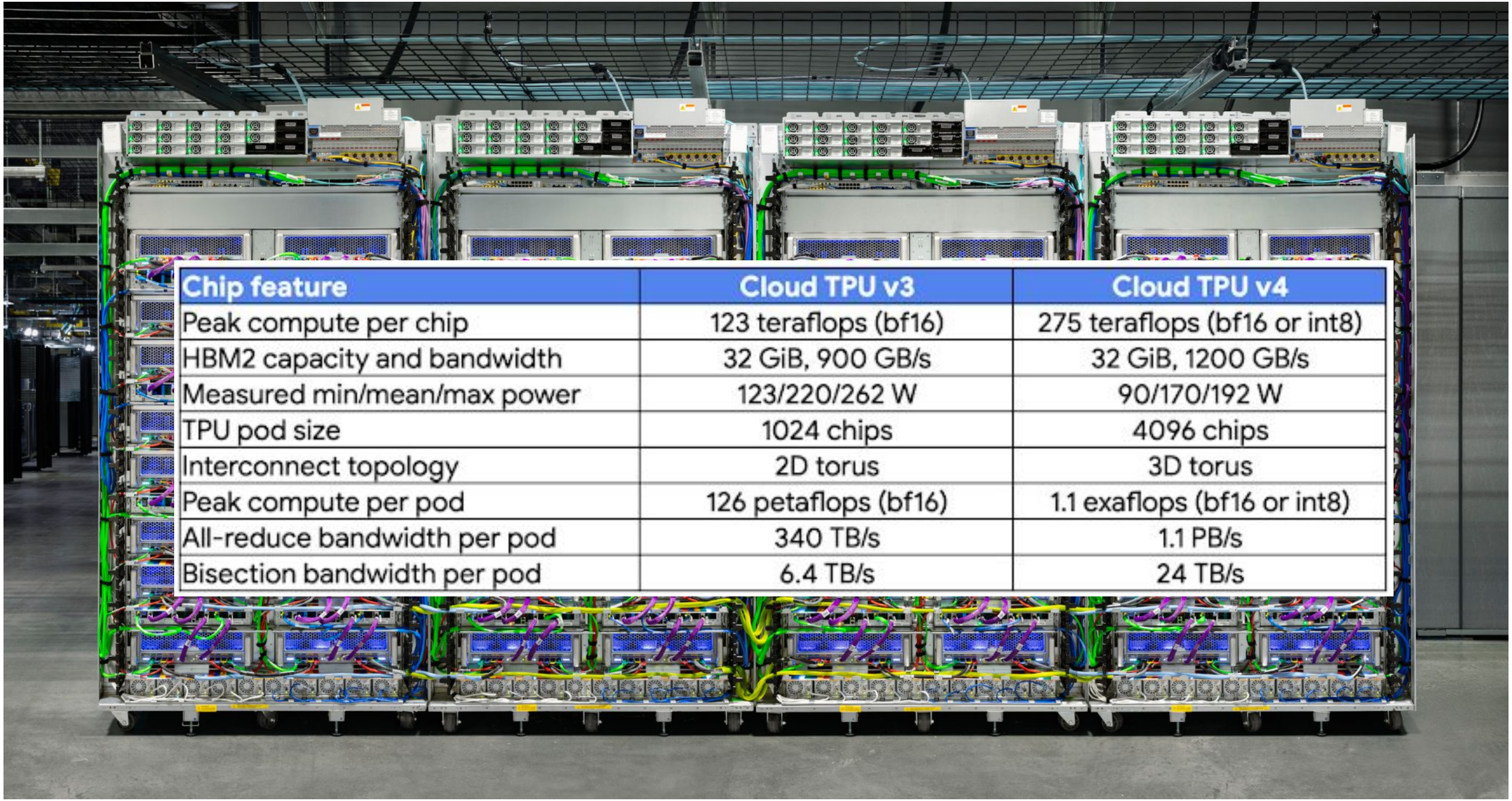
	Supported CUDA Core Precisions									Supported Tensor Core Precisions								
	FP8	**FP16**	**FP32**	**FP64**	**INT1**	**INT4**	**INT8**	**TF32**	**BF16**	**FP8**	**FP16**	**FP32**	**FP64**	**INT1**	**INT4**	**INT8**	**TF32**	**BF16**
NVIDIA Tesla P4	No	No	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No
NVIDIA P100	No	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No
NVIDIA Volta	No	Yes	Yes	Yes	No	No	Yes	No	No	No	Yes	No	No	No	No	No	No	No
NVIDIA Turing	No	Yes	Yes	Yes	No	No	Yes	No	No	No	Yes	No	No	Yes	Yes	Yes	No	No
NVIDIA A100	No	Yes	Yes	Yes	No	No	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
NVIDIA H100	No	Yes	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes


NVIDIA Quantum-2 InfiniBand NDR400
100 GB/s
BLUEFIELD DPU 0
GRACE CPU 0
NVLINK-C2C
HOPPER GPU 0
Hardware Coherency
900 GB/s
BLUEFIELD DPU 1
GRACE CPU 1
HOPPER GPU 1
BLUEFIELD DPU 255
GRACE CPU 255
HOPPER GPU 255
...
NVIDIA NVLink Switch System
NVIDIA NVLink Switch System

Nvidia Datacenter GPU	Nvidia A100 SXM	Nvidia H100 SXM	Nvidia H100 PCIe
GPU codename	GA100	GH100	GH100
GPU architecture	Ampere	Hopper	Hopper
GPU board form factor	SXM4	SXM5	PCIe Gen5
Launch date	May 2020	March 2022	March 2022
GPU process	TSMC 7nm N7	custom TSMC 4N	custom TSMC 4N
Die size	826mm2	814 mm2	814 mm2
Transistor Count	54 billion	80 billion	80 billion
FP64 CUDA cores	3,456	8,448	7,296
FP32 CUDA cores	6,912	16,896	14,592
Tensor Cores	432	528	456
Streaming Multiprocessors	108	132	114
Peak FP64	9.7 teraflops	30 teraflops	24 teraflops
Peak FP64 Tensor Core	19.5 teraflops	60 teraflops	48 teraflops
Peak FP32	19.5 teraflops	60 teraflops	48 teraflops
Peak FP32 Tensor Core	156 teraflops | 312 teraflops*	500 teraflops | 1,000 teraflops*	400 teraflops | 800 teraflops*
Peak BFLOAT16 Tensor Core	312 teraflops | 624 teraflops*	1,000 teraflops | 2,000 teraflops*	800 teraflops | 1,600 teraflops*
Peak FP16 Tensor Core	312 teraflops | 624 teraflops*	1,000 teraflops | 2,000 teraflops*	800 teraflops | 1,600 teraflops*
Peak FP8 Tensor Core	-	2,000 teraflops | 4,000 teraflops*	1,600 teraflops | 3,200 teraflops*
Peak INT8 Tensor Core	624 TOPS | 1,248 TOPS*	2,000 TOPS | 4,000 TOPS*	1,600 TOPS | 3,200 TOPS*
Peak INT4 Tensor Core	1,248 TOPS | 2,496 TOPS*	-	-
Interconnect	NVLink: 600GB/s PCI Gen4: 64GB/s	NVLink: 900GB/s PCI Gen5: 128GB/s	NVLink: 600GB/s PCI Gen5: 128GB/s
Max TDP	400 watts	700 watts	350 watts

*Effective TFLOPS or FLOPS using the Sparsity feature





Chip feature	Cloud TPU v3	Cloud TPU v4
Peak compute per chip	123 teraflops (bf16)	275 teraflops (bf16 or int8)
HBM2 capacity and bandwidth	32 GiB, 900 GB/s	32 GiB, 1200 GB/s
Measured min/mean/max power	123/220/262 W	90/170/192 W
TPU pod size	1024 chips	4096 chips
Interconnect topology	2D torus	3D torus
Peak compute per pod	126 petaflops (bf16)	1.1 exaflops (bf16 or int8)
All-reduce bandwidth per pod	340 TB/s	1.1 PB/s
Bisection bandwidth per pod	6.4 TB/s	24 TB/s

Scaling

January 18, 2018

Scaling Kubernetes to 2,500 nodes

We’ve been running Kubernetes for deep learning research for over two years. While our largest-scale workloads manage bare cloud VMs directly, Kubernetes provides a fast iteration cycle, reasonable scalability, and a lack of boilerplate which makes it ideal for most of our experiments. We now operate several Kubernetes clusters (some in the cloud and some on physical hardware), the largest of which we’ve pushed to over 2,500 nodes. This cluster runs in Azure on a combination of D15v2 and NC24 VMs.

https://openai.com/index/scaling-kubernetes-to-2500-nodes/

OpenAI Research Products Safety Company

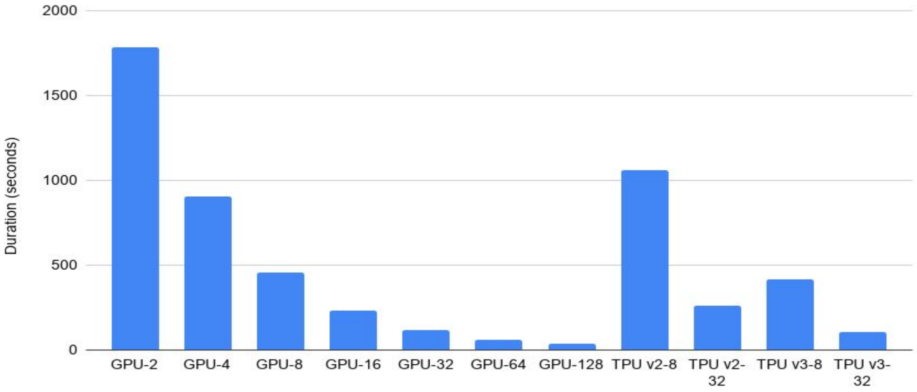
January 25, 2021

Scaling Kubernetes to 7,500 nodes

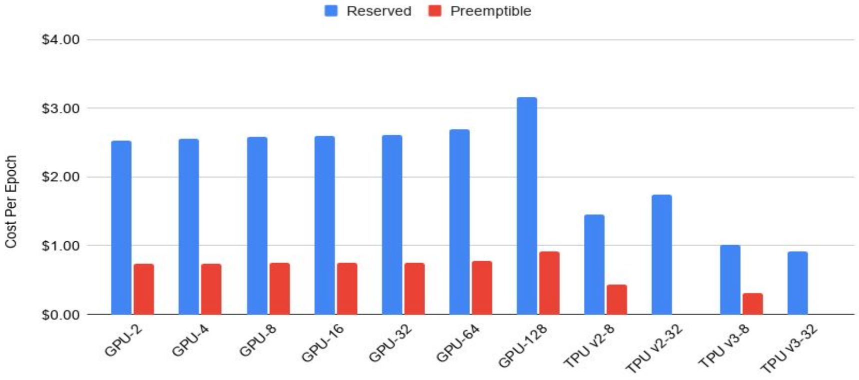
We've scaled Kubernetes clusters to 7,500 nodes, producing a scalable infrastructure for large models like GPT-3, CLIP, and DALL·E, but also for rapid small-scale iterative research such as Scaling Laws for Neural Language Models.

Scaling a single Kubernetes cluster to this size is rarely done and requires some special care, but the upside is a simple infrastructure that allows our machine learning research teams to move faster and scale up without changing their code.

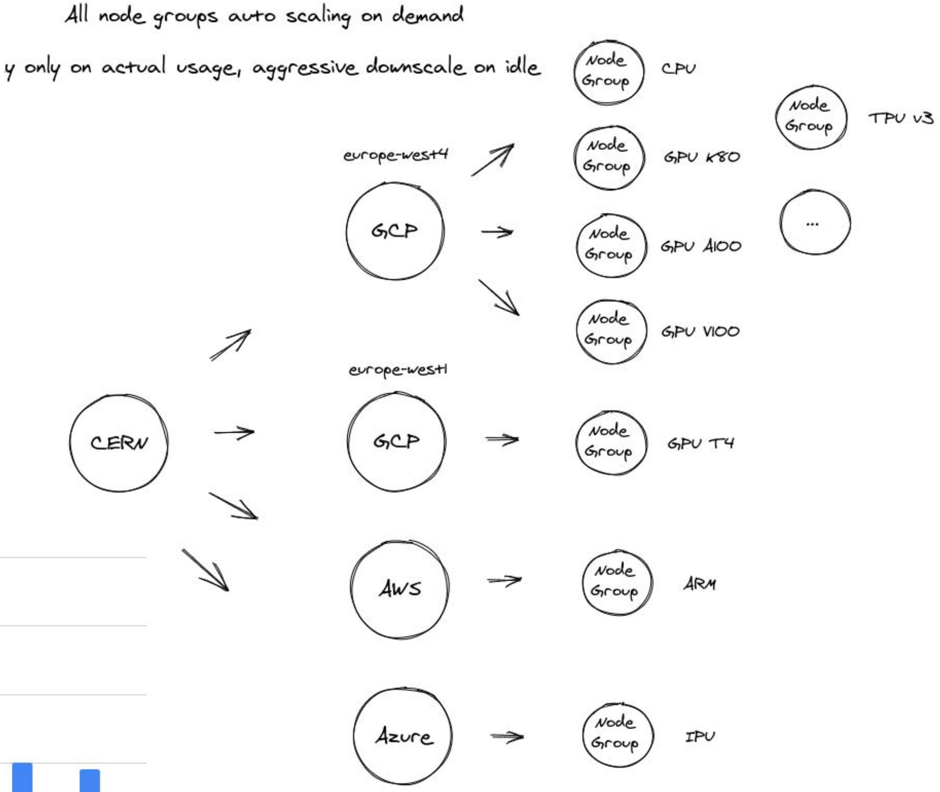
https://openai.com/index/scaling-kubernetes-to-7500-nodes/



Duration (seconds)
2000
1500
1000
500
0
GPU-2
GPU-4
GPU-8
GPU-16
GPU-32
GPU-64
GPU-128
TPU v2-8
TPU v2-32
TPU v3-8
TPU v3-32



Reserved
Preemptible
Cost Per Epoch
$4.00
$3.00
$2.00
$1.00
$0.00
GPU-2
GPU-4
GPU-8
GPU-16
GPU-32
GPU-64
GPU-128
TPU v2-8
TPU v2-32
TPU v3-8
TPU v3-32



All node groups auto scaling on demand
y only on actual usage, aggressive downscale on idle
CERN
europe-west4
GCP
europe-west1
GCP
AWS
Azure
Node Group
CPU
Node Group
GPU K80
Node Group
GPU A100
Node Group
GPU V100
Node Group
GPU T4
Node Group
ARM
Node Group
IPU
Node Group
TPU v3
...

Best Practices

Code and Datasets Storage

Containers are **ephemeral**, notebook servers can crash

Keep code and datasets on **persistent storage** that can be easily accessed

- Code - **Github** or **GitLab**, commit regularly, generate container images
- Datasets - **EOS** or **S3** object storage

Model Training

Develop models with scalability in mind

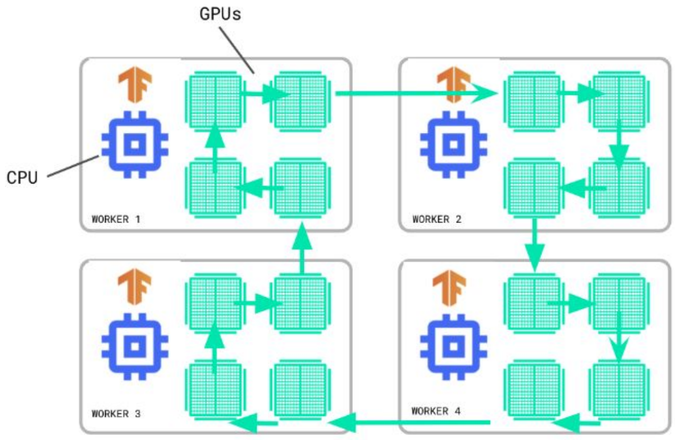
Develop model training to be **distributed** across multiple GPUs

Only **prototype** using a single GPU

TF Distributed - TFJob

Pytorch Distributed - PyTorchJob

Ray, MPI, PaddlePaddle, …



Containerised Workloads

Build Docker images for your ML workloads

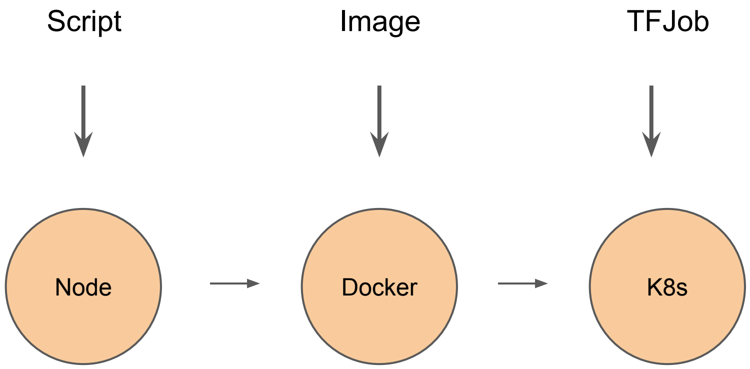
Allows for **reproducibility**

Mobility - build once, run anywhere

Fast deployment

Integration with other ML components

- Pipelines
- Distributed training
- Hyperparameter optimization
- Serving



Demo & Exercises

Cluster

```
ssh lxplus.cern.ch

openstack coe cluster create --cluster-template kubernetes-1.30.5-1 --node-count 1 csc-demo

openstack coe cluster list

openstack coe cluster config csc-demo

export KUBECONFIG=`pwd`/config
```

Basic Job

wget https://kubernetes.io/examples/controllers/job.yaml

kubectl apply -f job.yaml

```yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: pi
spec:
  template:
    spec:
      containers:
      - name: pi
        image: perl:5.34.0
        command: ["perl",  "-Mbignum=bpi", "-wle", "print bpi(2000)"]
      restartPolicy: Never
  backoffLimit: 4
```

Job Parallelism

https://kubernetes.io/docs/concepts/workloads/controllers/job/

```
spec:
  completions: 10
  parallelism: 3
```

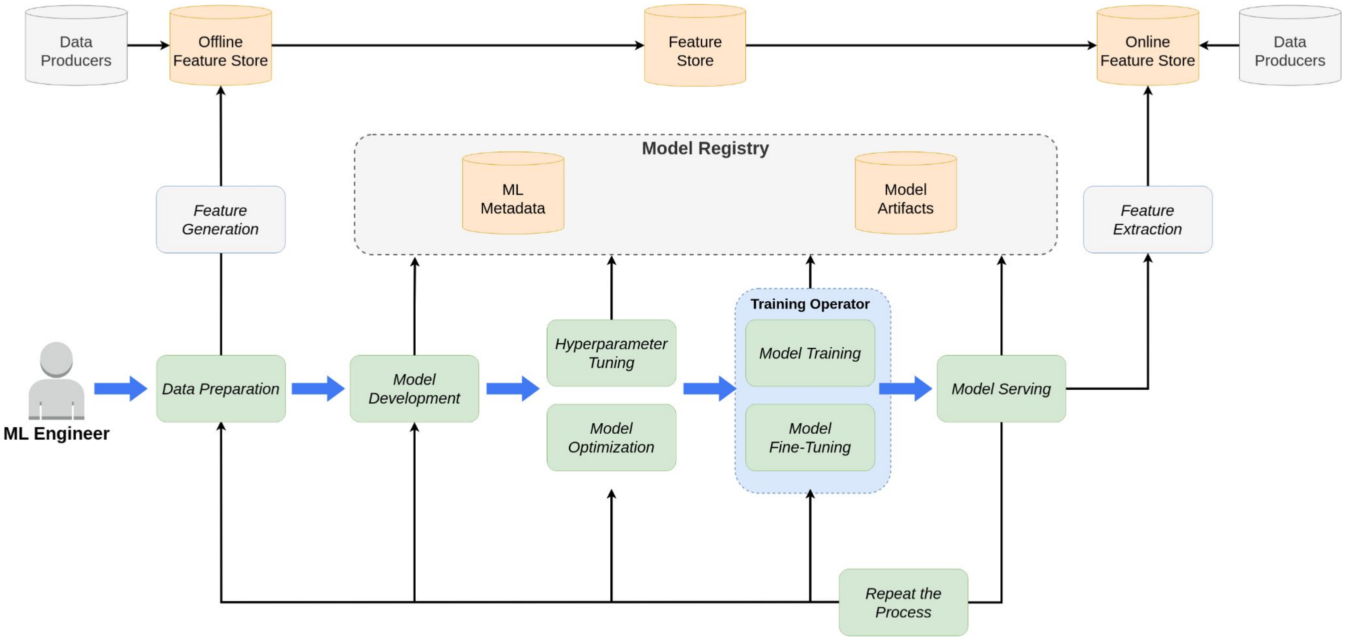
Indexing and Failure Policy

https://kubernetes.io/docs/concepts/workloads/controllers/job/#pod-backoff-failure-policy

```yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: job-backoff-limit-per-index-example
spec:
  completions: 10
  parallelism: 3
  completionMode: Indexed  # required for the feature
  backoffLimitPerIndex: 1  # maximal number of failures per index
  maxFailedIndexes: 5      # maximal number of failed indexes before terminating the .
  template:
    spec:
      restartPolicy: Never # required for the feature
      containers:
      - name: example
        image: python
        command:           # The jobs fails as there is at least one failed index
                           # (all even indexes fail in here), yet all indexes
                           # are executed as maxFailedIndexes is not exceeded.
        - python3
        - -c
        - |
          import os, sys
          print("Hello world")
          if int(os.environ.get("JOB_COMPLETION_INDEX")) % 2 == 0:
            sys.exit(1)
```

Training Operator

https://www.kubeflow.org/docs/components/training/overview/



Training Operator

https://www.kubeflow.org/docs/components/training/user-guides/

TFJob: TensorFlow Training

PyTorchJob: PyTorch Training

PaddleJob: PaddlePaddle

XGBoostJob: XGBoost

JAXJob: JAX Training

MPIJob: MPI Training


```yaml
apiVersion: kubeflow.org/v1
kind: PyTorchJob
metadata:
  clusterName: ""
  creationTimestamp: 2018-12-16T21:39:09Z
  generation: 1
  name: pytorch-tcp-dist-mnist
  namespace: default
  resourceVersion: "15532"
  selfLink: /apis/kubeflow.org/v1/namespaces/default/pytorchjobs/pytorch-tcp-dist-mnist
  uid: 059391e8-017b-11e9-bf13-06afd8f55a5c
spec:
  cleanPodPolicy: None
  pytorchReplicaSpecs:
    Master:
      replicas: 1
      restartPolicy: OnFailure
      template:
        metadata:
          creationTimestamp: null
        spec:
          containers:
          - image: gcr.io/kubeflow-ci/pytorch-dist-mnist_test:1.0
            name: pytorch
            ports:
            - containerPort: 23456
              name: pytorchjob-port
            resources: {}
    Worker:
      replicas: 3
      restartPolicy: OnFailure
      template:
        metadata:
          creationTimestamp: null
        spec:
          containers:
          - image: gcr.io/kubeflow-ci/pytorch-dist-mnist_test:1.0
            name: pytorch
            ports:
            - containerPort: 23456
              name: pytorchjob-port
            resources: {}
```

PyTorch Job

```
$ kubectl apply -k "github.com/kubeflow/training-operator.git/manifests/overlays/standalone?ref=v1.8.0"

$ wget https://raw.githubusercontent.com/kubeflow/training-operator/master/examples/pytorch/simple.yaml

$ kubectl apply -f simple.yaml

$ kubectl -n kubeflow get pytorchjobs

$ kubectl -n kubeflow logs -f pytorch-simple-worker-0

2024-11-07T08:50:33Z INFO     Train Epoch: 1 [0/60000 (0%)] loss=2.2975

2024-11-07T08:50:36Z INFO     Train Epoch: 1 [640/60000 (1%)]loss=2.2965

2024-11-07T08:50:40Z INFO     Train Epoch: 1 [1280/60000 (2%)]    loss=2.2948

2024-11-07T08:50:43Z INFO     Train Epoch: 1 [1920/60000 (3%)]    loss=2.2833
```

Queues and Multi-Cluster

https://kueue.sigs.k8s.io/docs/concepts/

Local Queue

Cluster Queue

Resource Flavor

Workload Priority

MultiKueue and MultiKueueCluster

Ongoing Work

Ongoing Work

Improved documentation on availability and access to GPUs

Consolidation of GPU pools among services

Integration with public cloud providers

Q & A