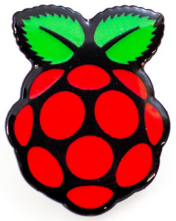
Cloud Native and Machine Learning at the Edge:
ARM and NVIDIA

Overview

- Cloud Native
- Containers
- Kubernetes
- Raspberry Pi



- Machine Learning
- CPU: ARM64
- GPU: NVIDIA
- Jetson Family

Cloud Native

“Cloud native technologies empower organizations to build and run scalable applications in modern, dynamic environments such as public, private, and hybrid clouds. Containers, service meshes, microservices, immutable infrastructure, and declarative APIs exemplify this approach.”

https://www.cncf.io/about/faq/#what-is-cloud-native

CLOUD NATIVE COMPUTING FOUNDATION

Here we will just talk about the technology: Containers and Kubernetes

“Barry Dingle” https://dilbert.com/strip/2017-11-08

https://fsfe.org/nocloud

https://kubernetes.io

Containers?

- 1982: Chroot (Unix)
- 2000: Jails (FreeBSD)
- 2002: Zones (Solaris)
- 2005: OpenVZ (Linux)
- 2008: LXC (also Linux)
- 2013: Docker (Linux)
- 2016: Docker (Windows)

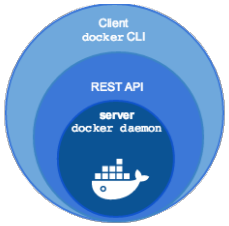
image: Freepik.com

Docker / Moby

2013: Docker announced

https://www.docker.com/

- Docker Engine
  - Client (docker)
  - Server (daemon)



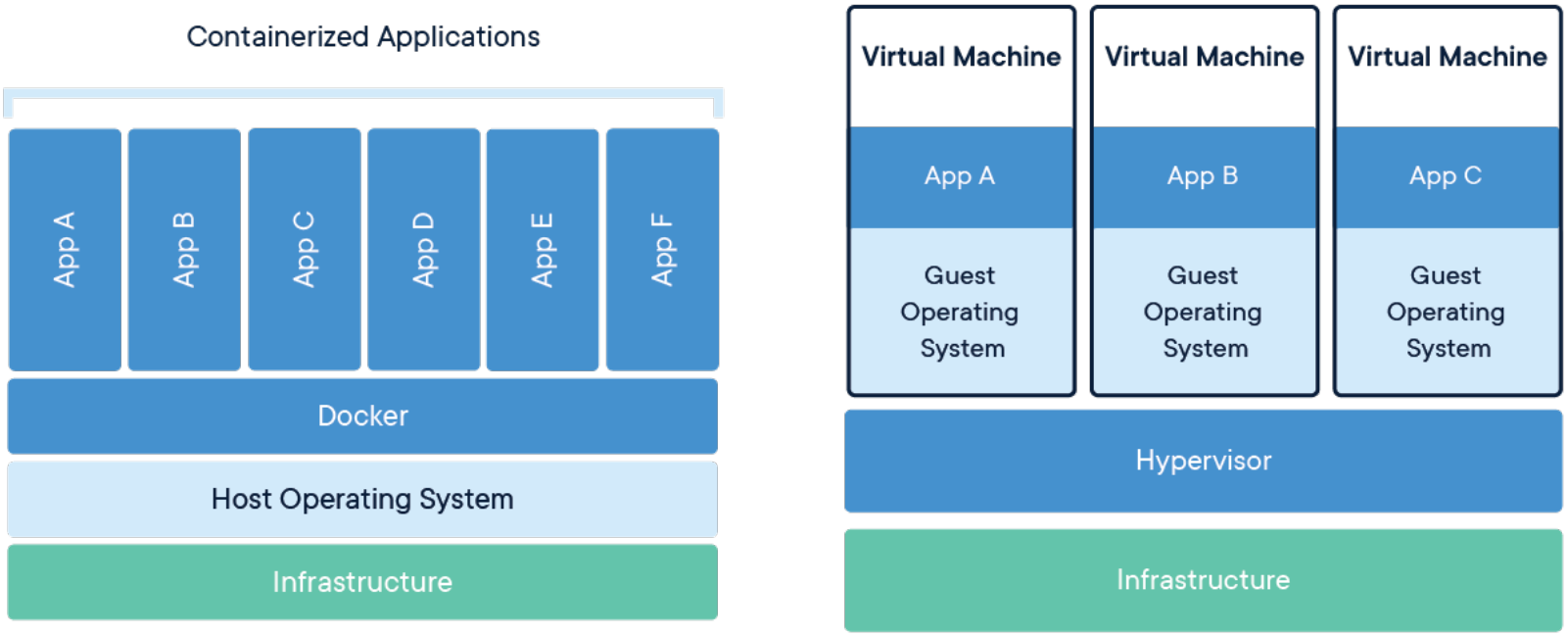
- Docker Desktop
  - macOS (Linux VM)
  - Windows (Linux VM)

2017: Moby introduced

https://mobyproject.org/

Containers vs. Virtual Machines

”A container is a standard unit of software that packages up code and all its dependencies ...”

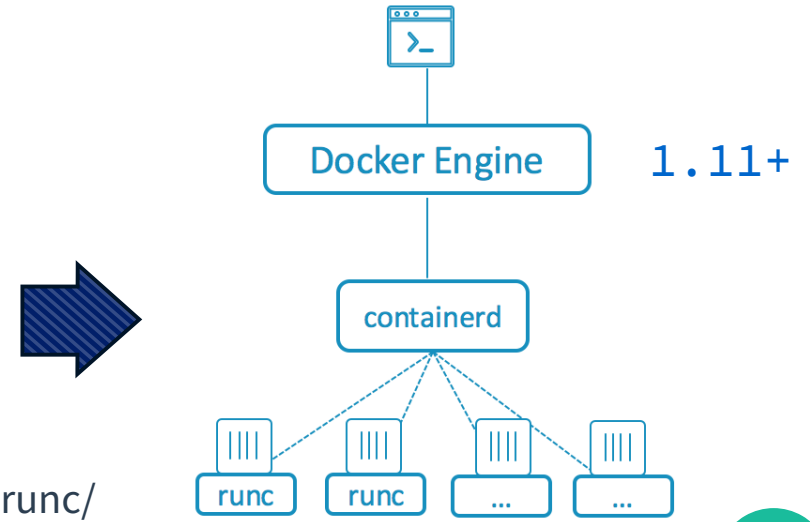


https://www.docker.com/resources/what-container

containerd

”Containerd is a daemon providing a GRPC API to manage containers on the local system. Containerd leverages runC to [...] provide advanced functionality.”

https://github.com/containerd/containerd

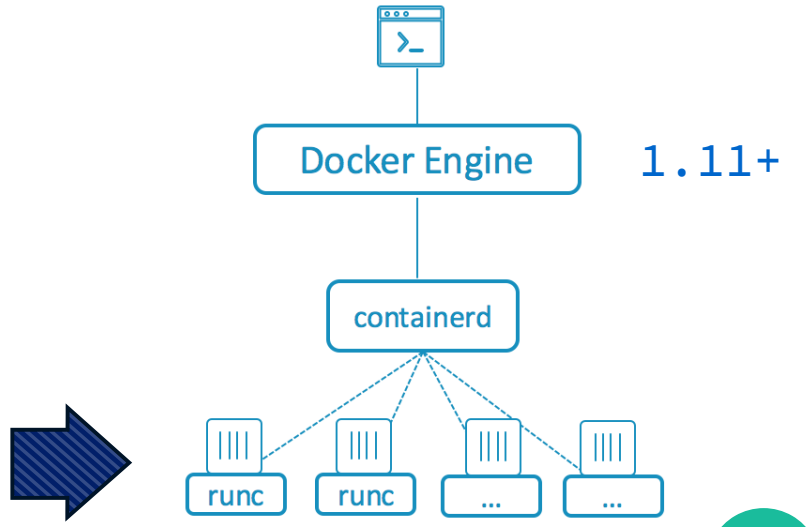


https://blog.docker.com/2015/12/containerd-daemon-to-control-runc/

runC

”runC is a lightweight, portable container runtime. It includes all of the plumbing code used by Docker to interact with system features related to containers.”

https://github.com/opencontainers/runc



https://blog.docker.com/2015/06/runc/

Kubernetes

Kubernetes is an open-source container orchestration system for automating deployment, scaling, and management of containerized applications.

https://kubernetes.io

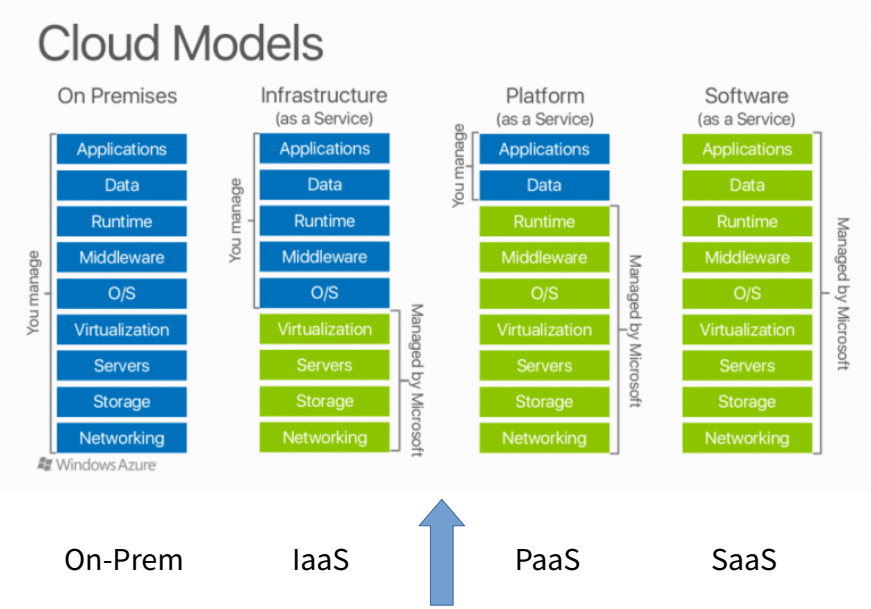
Google open-sourced the Kubernetes project in 2014.

“ Kubernetes builds upon a decade and a half of experience that Google has with running production workloads at scale, combined with best-of-breed ideas and practices from the community. “

Kubernetes

- ## Why containers ? (Containers-as-a-Service, CaaS)

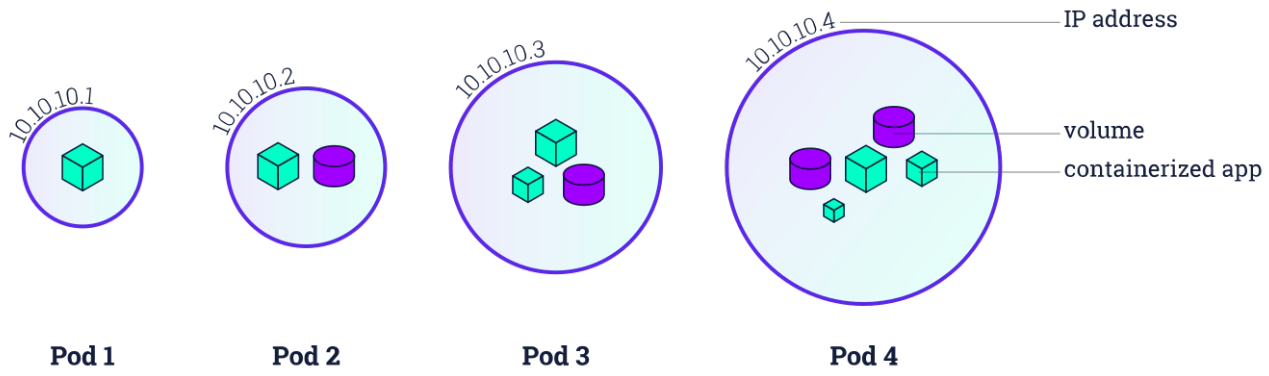
" Kubernetes provides a container-centric management environment. It orchestrates computing, networking, and storage infrastructure on behalf of user workloads. This provides much of the simplicity of Platform as a Service (PaaS) with the flexibility of Infrastructure as a Service (IaaS), and enables portability across infrastructure providers. "



https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/#why-containers

Pod?

A pod (as in a pod of seals or peas) is a group of one or more containers, with shared storage/network, and a specification for how to run the containers.



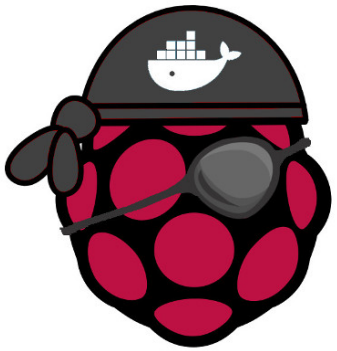
https://kubernetes.io/docs/tutorials/kubernetes-basics/

Hypriot and Kubecloud

2015

https://hypriot.com/

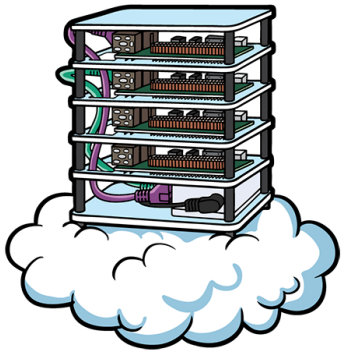
“HypriotOS - the fastest way to get Docker up and running on any Raspberry Pi.”



2016

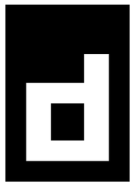
https://kubecloud.io/

“KubeCloud is a tangible cloud computing environment … a small-scale datacenter.”



Installation

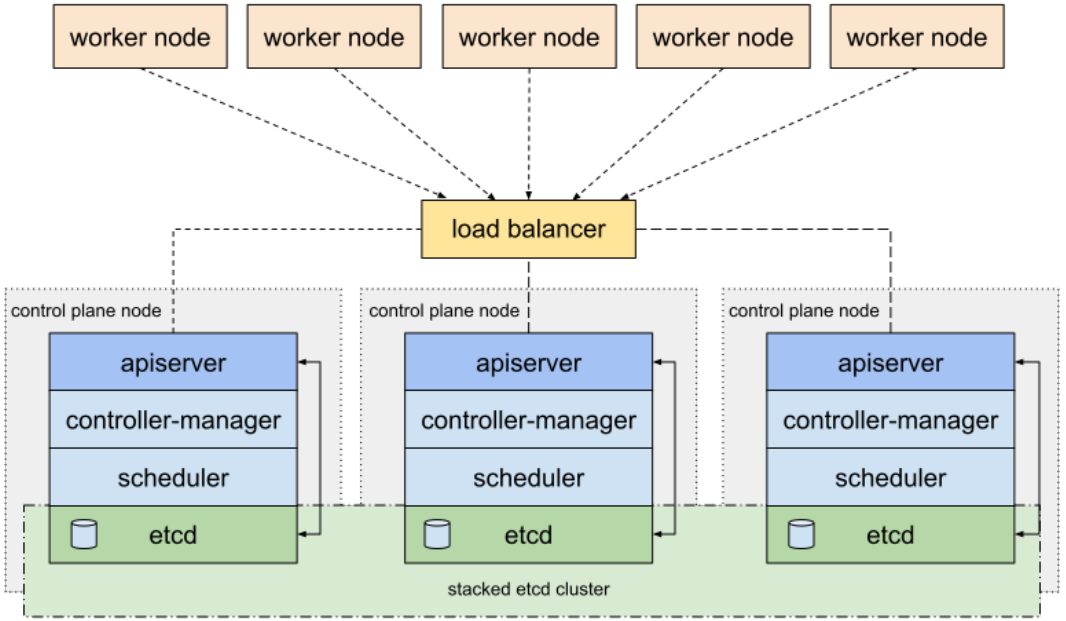
- **cloud-init**
  - add ssh/sudo
- **ansible**
  - needs python
- **K8s (kubeadm)**
  - docker/moby
- **K3s (rancher)**
  - containerd



High Availability

kubeadm

- **Stacked CP**
  - control plane
  - etcd (database)
- **Virtual IP**
  - keepalived
  - haproxy



https://kubernetes.io/docs/setup/production-environment/tools/kubeadm/high-availability/

Raspberry Pi

1 B	2 B	3 B	4 B
armv6 700 MHz	armv7 900 MHz	armv8 1.2 GHz	armv8 1.5 Ghz
512 MB RAM	1 GB RAM	1 GB RAM	2-8 GB RAM
2012	2015	2016	2019
$35	$35	$35	$35*

* 2 GB

Demo – Raspbernetes

Machine Learning

- Artifical Intelligence (AI)
- **Machine Learning**
- Deep Learning

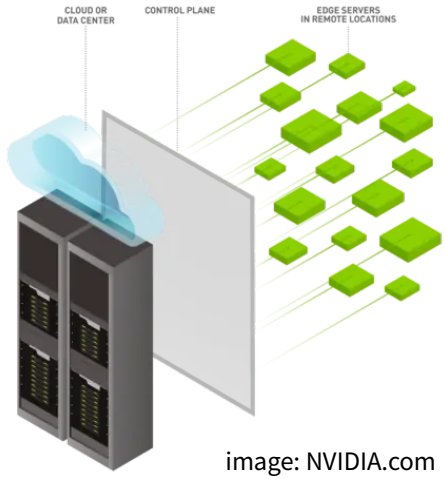
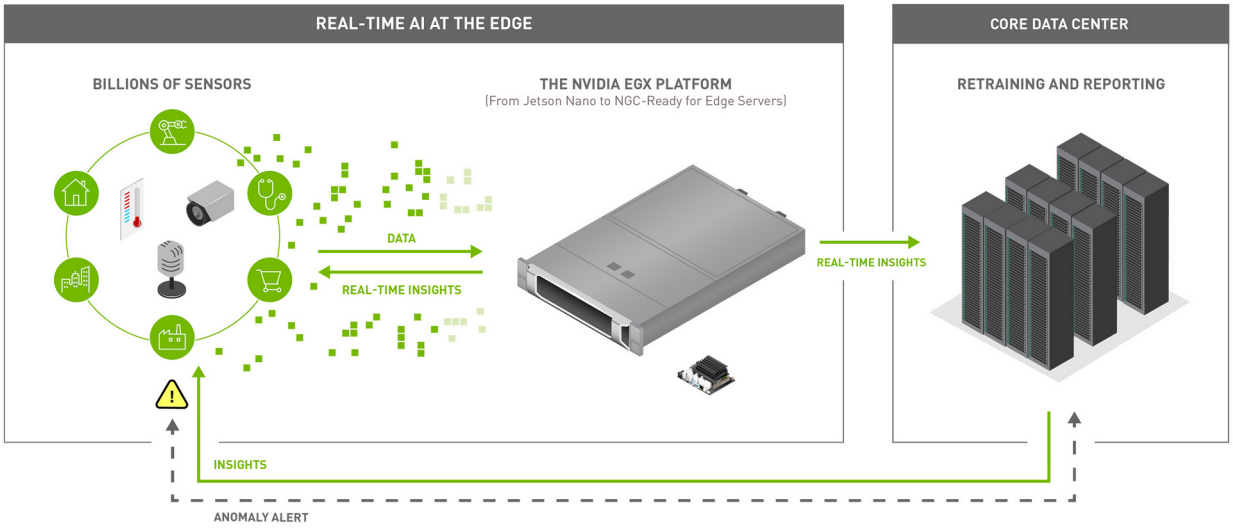


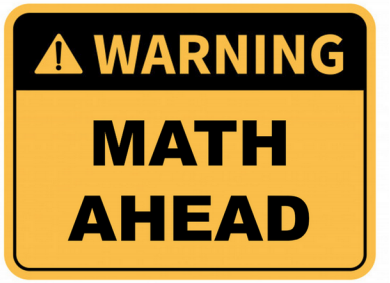
image: NVIDIA.com

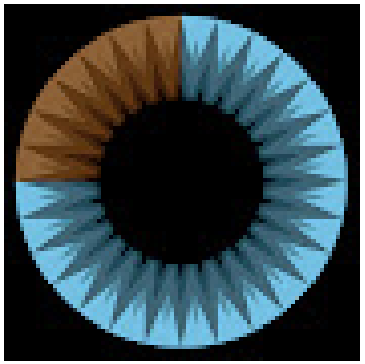


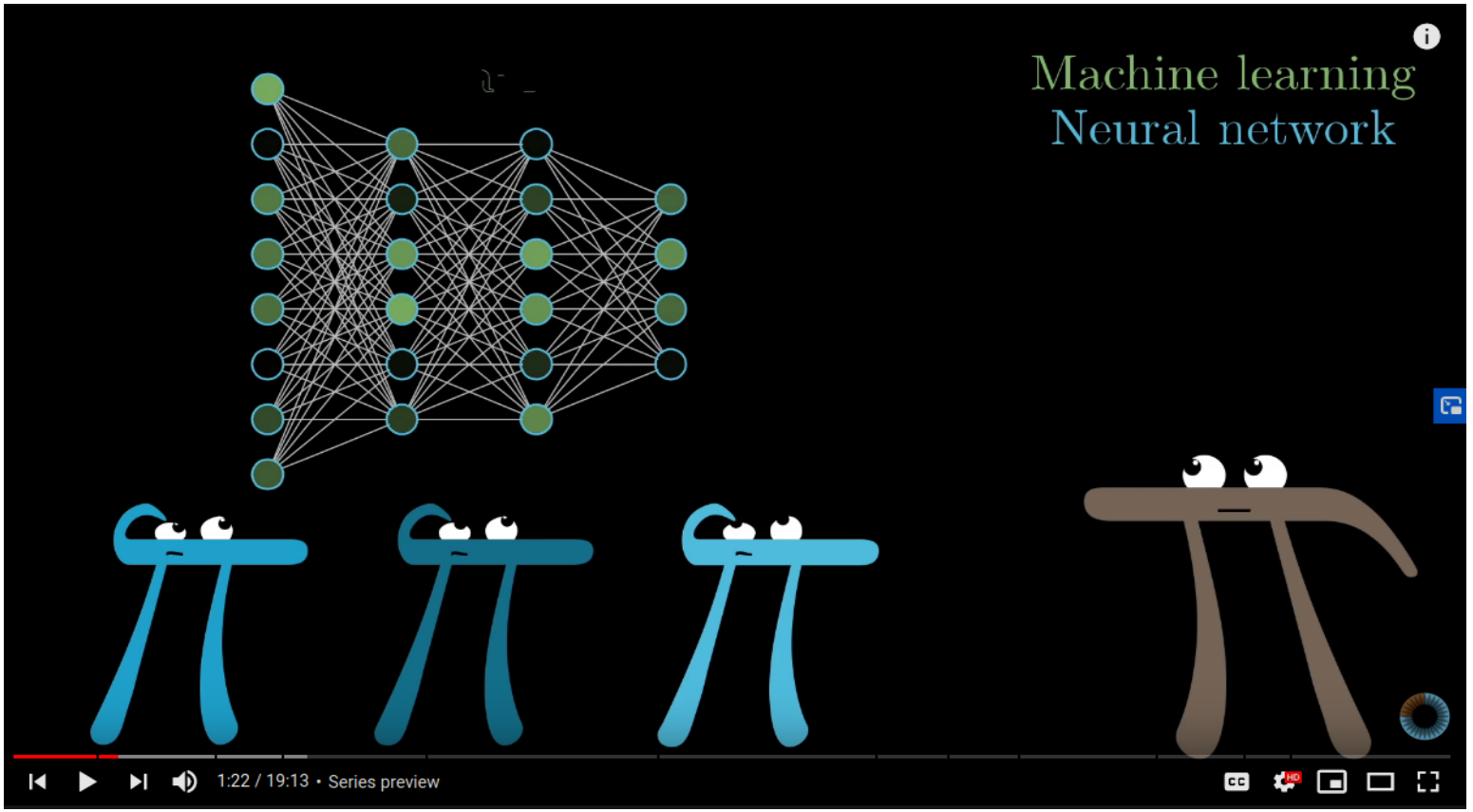
Cloud
(Data Center)

Edge
(Remote Location)

Neural Network







“But what is a Neural Network? | Deep learning, chapter 1”

https://www.youtube.com/c/3blue1brown/playlists

Free Courses

- Elements of AI — https://www.elementsofai.se/
- AI of Sweden — https://ai.se/
- NVIDIA Deep Learning Institute — https://developer.nvidia.com/
- Practical Deep Learning for Coders — https://fast.ai/

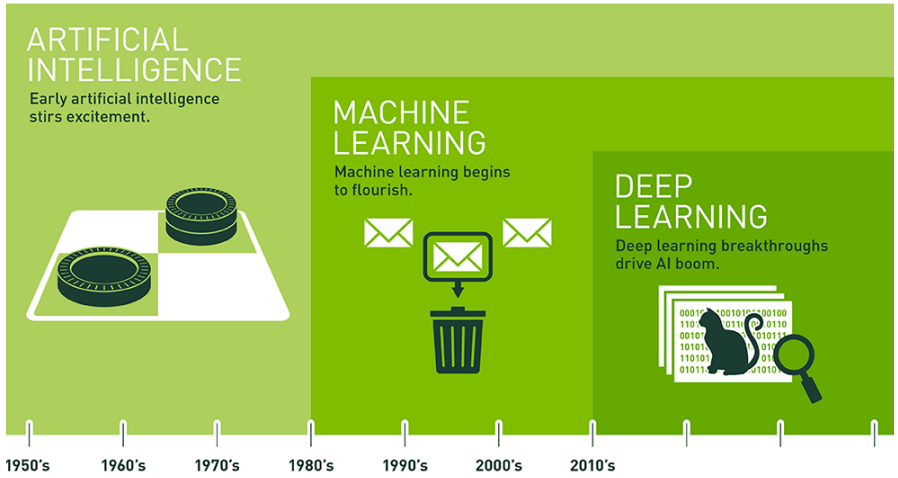
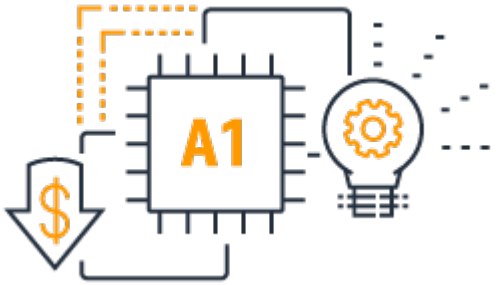


image: NVIDIA.com

CPU

- **amd64** (`x86_64`)
- **arm64** (`aarch64`)
- **Apple**
- **Amazon**
- **NVIDIA!**

2018



Amazon EC2 A1 Instances: AWS Graviton Processors

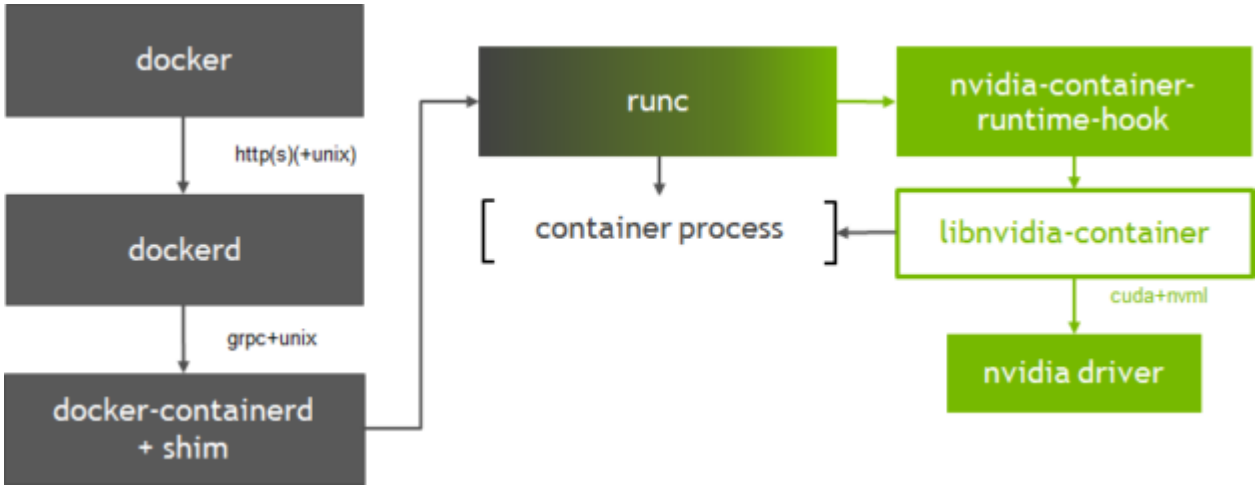
2020

Fugaku supercomputer: 8 million ARM cores

https://nvidianews.nvidia.com/news/nvidia-to-acquire-arm-for-40-billion-creating-worlds-premier-computing-company-for-the-age-of-ai

GPU

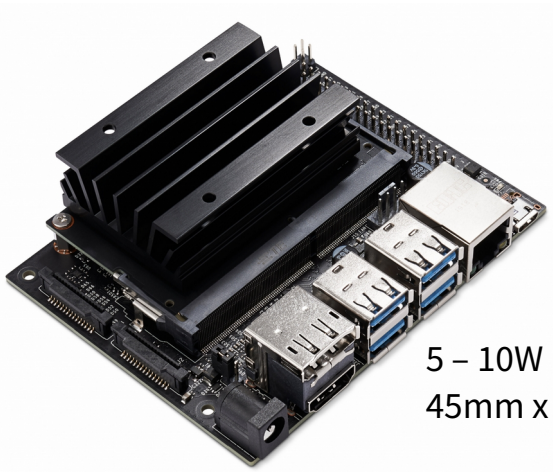
- CUDA (nvidia)
- nvidia-docker2



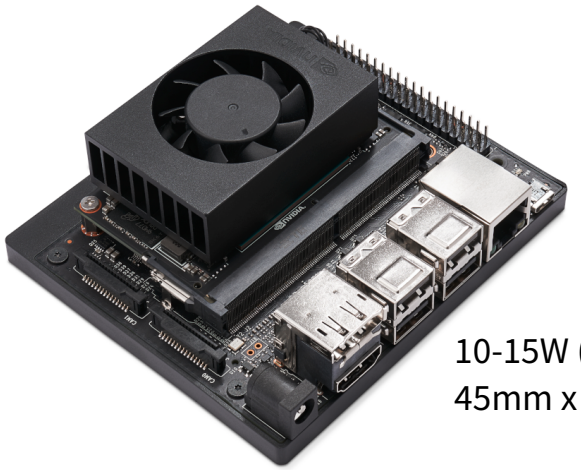
https://developer.nvidia.com/blog/gpu-containers-runtime/

Jetson Family

	Jetson Nano	Jetson TX2	**Jetson Xavier NX**	Jetson AGX Xavier
CPU	4 core ARM A57	6 core Denver and A57	6 Core Carmel ARMv8.2	8 core Carmel ARM v8.2
RAM	4 GB	8 GB	8 GB	16 GB
GPU	128 core Maxwell	256 core Pascal	384 core Volta	512 core Volta
	0.5 TFLOPS	1.3 TFLOPS	6 TFLOPS	11 TFLOPS
Dev Kit	$99		$399	



5 – 10W (5V)
45mm x 70mm



10-15W (19V)
45mm x 70mm

Demo – Jetson Nano

Questions?

Anders F Björklund
github.com/afbjorklund

For more info on boot2podman:
https://boot2podman.github.io/