

AI on Nvidia Jetson Edge System

Jeong-Gun Lee

AI Accelerator Computing (AIAC) Lab

Division of Software, College of Info. Science, Hallym University

Jeonggun.lee@gmail.com



차례



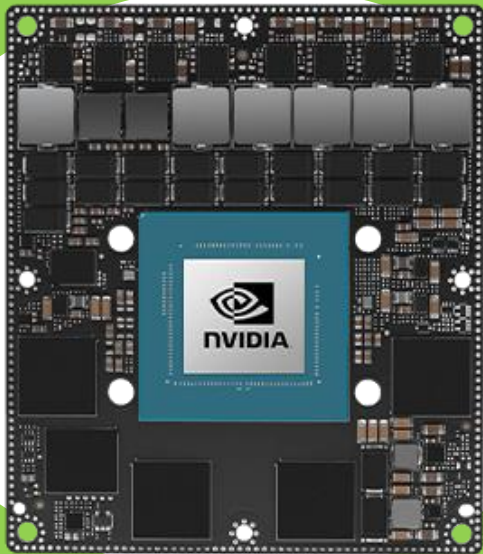
- Nvidia Jetson System 소개
- Jetson Nano System 설정
- Jetson Nano에서 즐기는 Parallel Computing (CUDA)
- Jetson Nano에서 즐기는 Deep Learning



AIAC lab



Nvidia Jetson System 소개



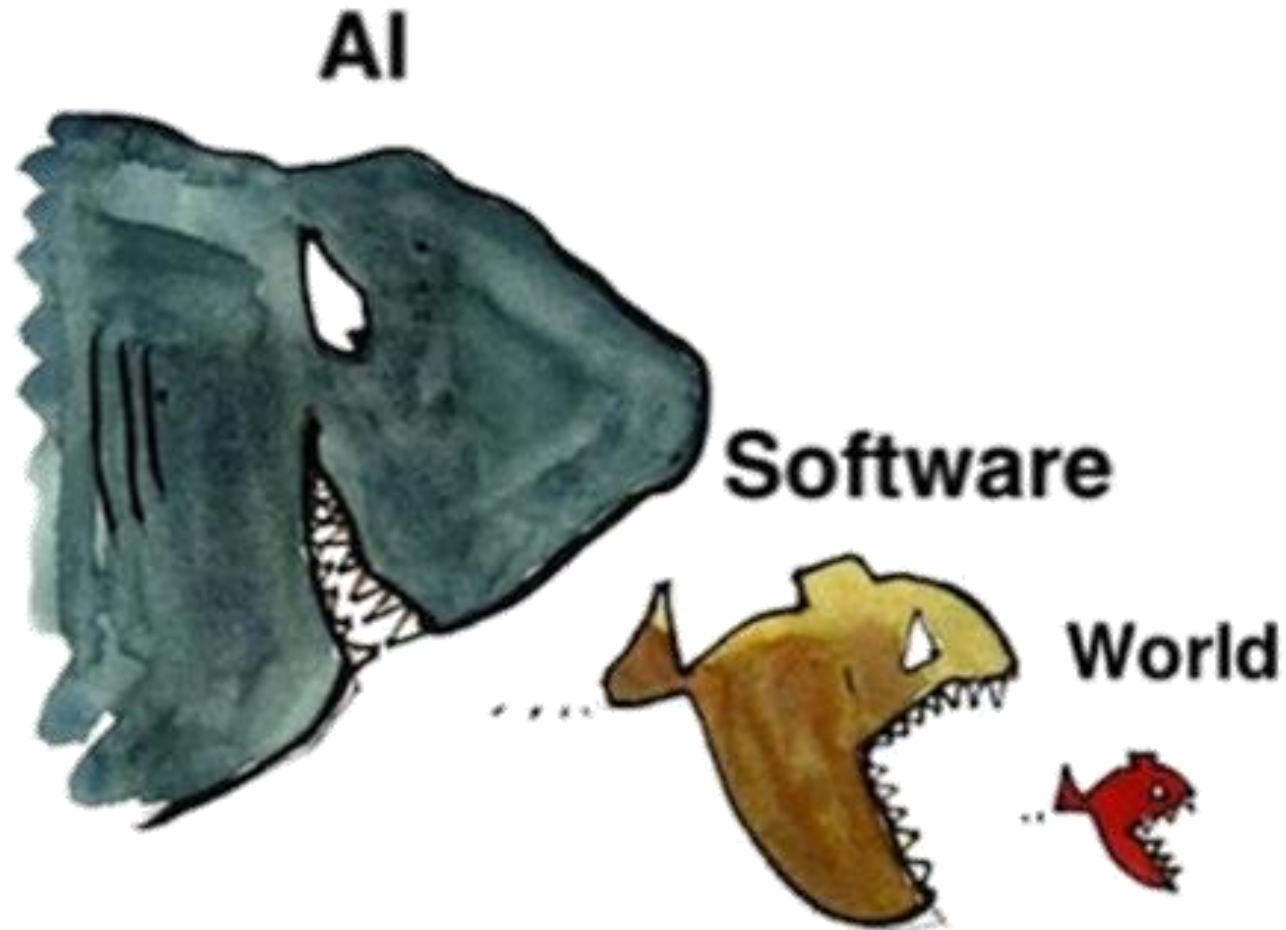
지금 세상을 생각하면...

Software



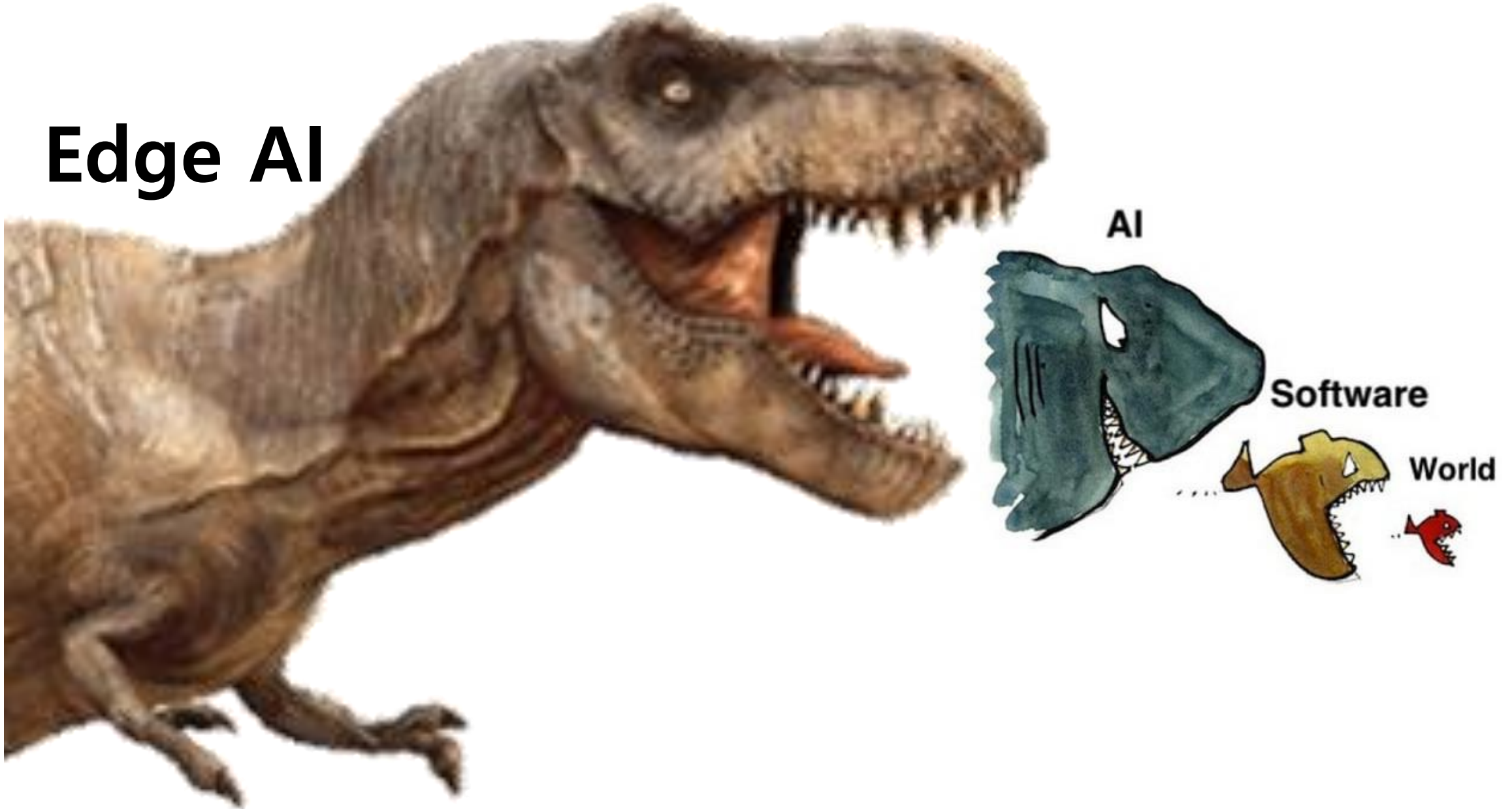
지금 세상을 생각하면...

AI Revolution & What is the NEXT?

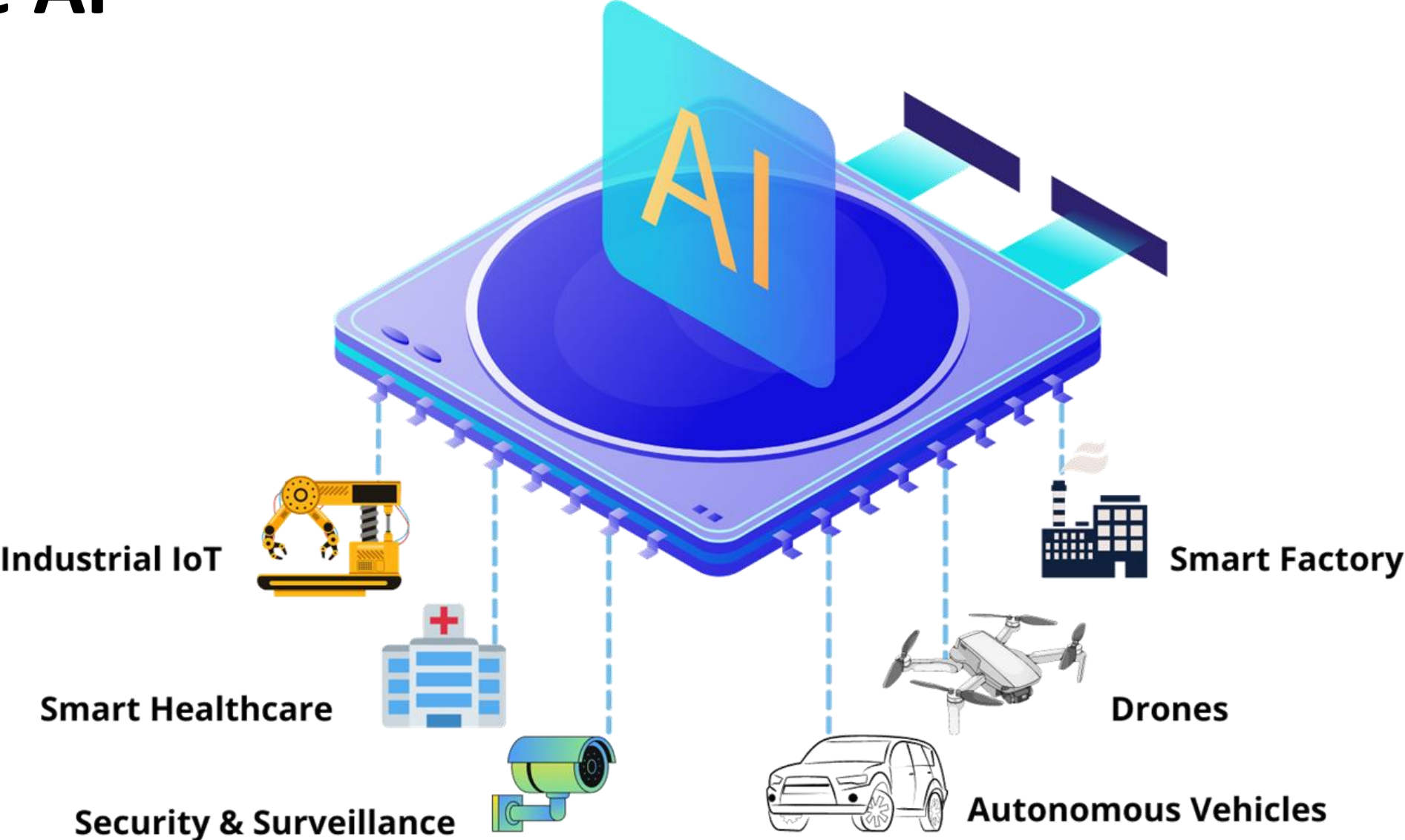


지금 세상을 생각하면...

Edge AI



Edge AI



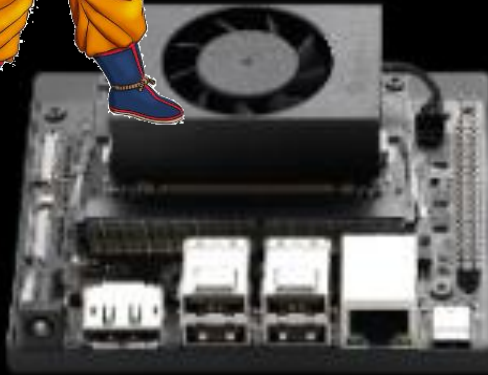
AI용 Edge 장치: 내가 하나 소개해줄께!



Jetson Nano Developer Kit

0.472 TFLOPS
\$149

LEARNER



Jetson Orin Nano Developer Kit

40 TOPS (INT8)
\$499

INTERMEDIATE



Jetson AGX Orin Developer Kit

275 TOPS (INT8)
\$1999

ADVANCED

AI용 Edge 장치: 기본 용어 정리~

TOPS: Tera Operations Per Second

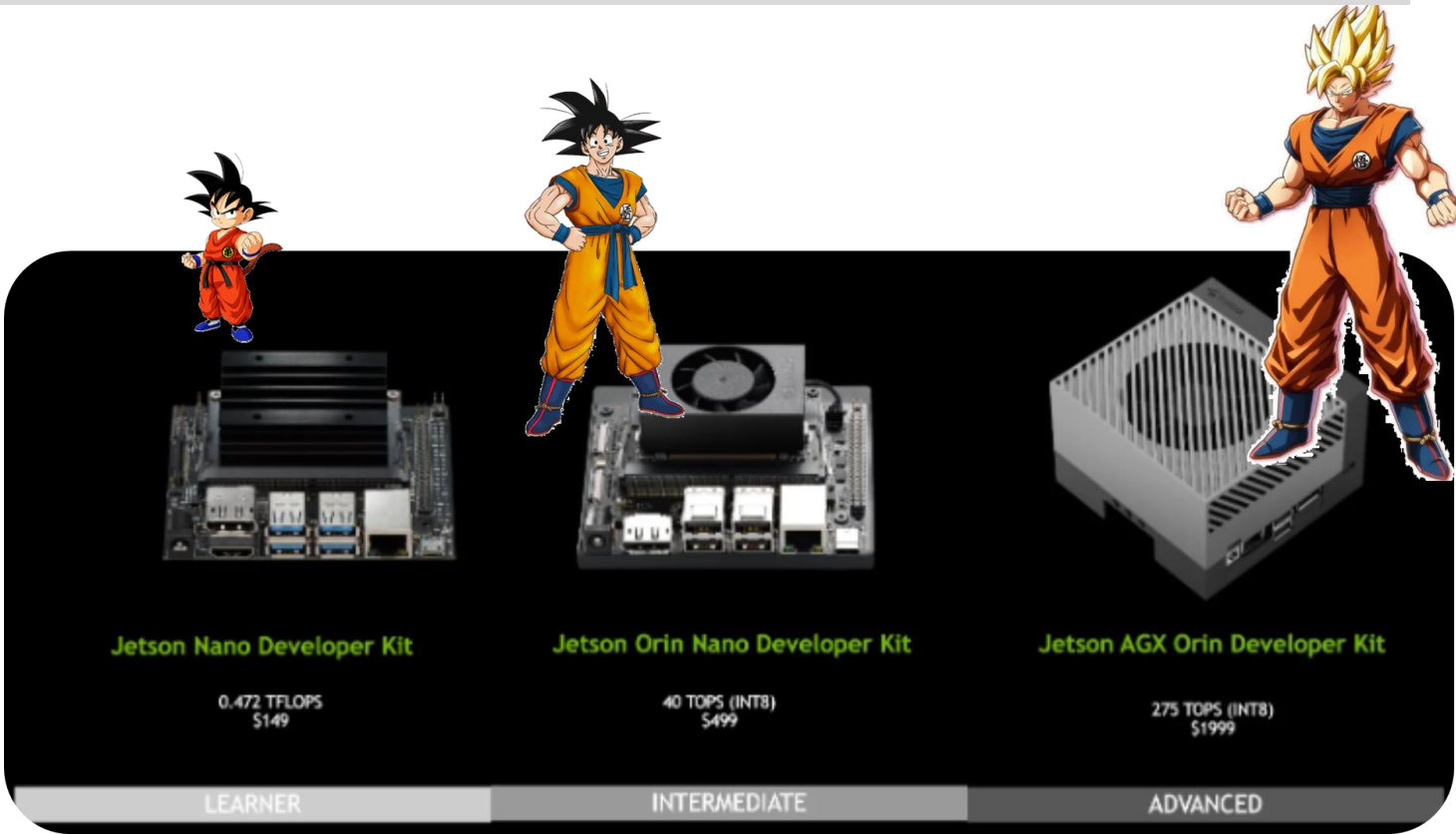
GFLOPS: Giga Floating-Point Operations Per Second

TFLOPS: Tera Floating-Point Operations Per Second

국제단위계(SI)에서 지정한 표준 접두어

접두어	기호	한자표기	10진법표기
데카(deca)	da	십(十)	10
헥토(hecto)	h	백(百)	100
킬로(kilo)	k	천(千)	1,000
메가(mega)	M	백만(百萬)	1,000,000
기가(giga)	G	십억(十億)	1,000,000,000
테라(tera)	T	조(兆)	1,000,000,000,000
페타(peta)	P	천조(千兆)	1,000,000,000,000,000
엑사(exa)	E	백경(百京)	1,000,000,000,000,000,000
제타(zetta)	Z	십해(垓)	1,000,000,000,000,000,000,000
요타(yuotta)	Y	자(秭)	1,000,000,000,000,000,000,000,000

<https://it.donga.com/7949/>



Nvidia Jetsons

REF: <https://www.techpowerup.com/gpu-specs/?mfgr=NVIDIA&generation=Tegra&sort=name>

• Nvidia Jetson Family

Year	Version	Performance	GPU	CPU	Memory	Power
2019	Jetson Nano	235.8 GFLOPS	<div>128-core Nvidia Maxwell architecture GPU</div> <div><div>GM20B</div><div>128</div><div>16</div><div>16</div><div>4 GB</div><div>LPDDR4</div><div>64 bit</div><div>GRAPHICS PROCESSOR</div><div>CORES</div><div>TMUS</div><div>ROPS</div><div>MEMORY SIZE</div><div>MEMORY TYPE</div><div>BUS WIDTH</div></div>	Quad-core ARM Cortex-A57 MPCore processor	4 GiB	5–10 W
2023	Jetson Orin Nano	20–40 TOPS - 1,280 GFLOPS	<div>from 512-core Nvidia Ampere architecture GPU with 16 Tensor cores</div> <div><div>GA10B</div><div>1024</div><div>32</div><div>16</div><div>8 GB</div><div>LPDDR5</div><div>128 bit</div><div>GRAPHICS PROCESSOR</div><div>CORES</div><div>TMUS</div><div>ROPS</div><div>MEMORY SIZE</div><div>MEMORY TYPE</div><div>BUS WIDTH</div></div>	6-core ARM Cortex-A78AE v8.2 64-bit CPU 1.5MB L2 + 4MB L3	4-8 GiB	7–10 W
2023	Jetson Orin NX	70–100 TOPS - 1.880 TFLOPS	<div>1024-core Nvidia Ampere architecture GPU with 32 Tensor cores</div> <div><div>GA10B</div><div>1024</div><div>32</div><div>16</div><div>8 GB</div><div>LPDDR5</div><div>128 bit</div><div>GRAPHICS PROCESSOR</div><div>CORES</div><div>TMUS</div><div>ROPS</div><div>MEMORY SIZE</div><div>MEMORY TYPE</div><div>BUS WIDTH</div></div> <div><div>GA10B</div><div>1024</div><div>32</div><div>16</div><div>16 GB</div><div>LPDDR5</div><div>128 bit</div><div>GRAPHICS PROCESSOR</div><div>CORES</div><div>TMUS</div><div>ROPS</div><div>MEMORY SIZE</div><div>MEMORY TYPE</div><div>BUS WIDTH</div></div>	up to 8-core ARM Cortex-A78AE v8.2 64-bit CPU 2MB L2 + 4MB L3	8–16 GiB	10–25 W
2023	Jetson Orin AGX	200-275 TOPS - 5.325 TFLOPS	<div>up to 2048(1792)-core Nvidia Ampere architecture GPU with 64 Tensor cores</div> <div><div>GA10B</div><div>1792</div><div>56</div><div>24</div><div>32 GB</div><div>LPDDR5</div><div>256 bit</div><div>GRAPHICS PROCESSOR</div><div>CORES</div><div>TMUS</div><div>ROPS</div><div>MEMORY SIZE</div><div>MEMORY TYPE</div><div>BUS WIDTH</div></div> <div><div>GA10B</div><div>2048</div><div>64</div><div>32</div><div>64 GB</div><div>LPDDR5</div><div>256 bit</div><div>GRAPHICS PROCESSOR</div><div>CORES</div><div>TMUS</div><div>ROPS</div><div>MEMORY SIZE</div><div>MEMORY TYPE</div><div>BUS WIDTH</div></div>	up to 12-core ARM Cortex-A78AE v8.2 64-bit CPU 3MB L2 + 6MB L3	32–64 GiB	15–60 W

FP16 (half):	471.6 GFLOPS (2:1)
FP32 (float):	235.8 GFLOPS
FP64 (double):	7.368 GFLOPS (1:32)

FP16 (half):	2.560 TFLOPS (2:1)
FP32 (float):	1,280 GFLOPS
FP64 (double):	640.0 GFLOPS (1:2)

FP16 (half):	3.760 TFLOPS (2:1)
FP32 (float):	1.880 TFLOPS
FP64 (double):	940.0 GFLOPS (1:2)







FP16 (half):	6.666 TFLOPS (2:1)
FP32 (float):	3.333 TFLOPS
FP64 (double):	1.667 TFLOPS (1:2)

FP16 (half):	10.65 TFLOPS (2:1)
FP32 (float):	5.325 TFLOPS
FP64 (double):	2.662 TFLOPS (1:2)

Nvidia Jetsons

REF: <https://www.techpowerup.com/gpu-specs/?mfgr=NVIDIA&generation=Tegra&sort=name>

- Nvidia Jetson Family

Graphics cards using the NVIDIA GA10B GPU								
Name		Chip	Memory	Shaders	TMUs	ROPs	GPU Clock	Memory Clock
NVIDIA Jetson AGX Orin 32 GB			32 GB	1792	56	24	930 MHz	1600 MHz
NVIDIA Jetson AGX Orin 64 GB			64 GB	2048	64	32	1300 MHz	1600 MHz
NVIDIA Jetson Orin Nano 4 GB			4 GB	512	16	8	625 MHz	1067 MHz
NVIDIA Jetson Orin Nano 8 GB			8 GB	1024	32	16	625 MHz	1067 MHz
NVIDIA Jetson Orin NX 16 GB		TE980M-A1	16 GB	1024	32	16	918 MHz	1600 MHz
NVIDIA Jetson Orin NX 8 GB		TE980M-A1	8 GB	1024	32	16	765 MHz	1600 MHz

- Shaders ~ CUDA cores
- TMUs ~ Texture mapping units
- ROPs ~ Render output units

Nvidia GPUs

- GPU ???
- CPU ???
- Cores ???



CPU	GPU
Central Processing Unit	Graphics Processing Unit
4-8 Cores	100s or 1000s of Cores
Low Latency	High Throughput
Good for Serial Processing	Good for Parallel Processing
Quickly Process Tasks That Require Interactivity	Breaks Jobs Into Separate Tasks To Process Simultaneously
Traditional Programming Are Written For CPU Sequential Execution	Requires Additional Software To Convert CPU Functions to GPU Functions for Parallel Execution

Nvidia Jetsons



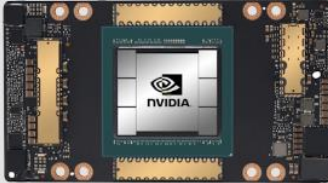

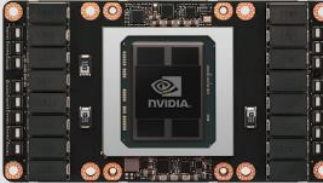
REF: <https://www.techpowerup.com/gpu-specs/?mfgr=NVIDIA&generation=Tegra&sort=name>

- Nvidia **RTX 3090, 4090**, A100, H100

	RTX 4090	RTX 3090 Ti	RTX 3090
CUDA Cores	16384	10,752	10,496
Boost Clock	2.52 GHz	1.86 GHz	1.7 GHz
Base Clock	2.23 GHz	1.67 GHz	1.4 GHz
Memory Size	24 GB	24 GB	24 GB
Memory Type	GDDR6X	GDDR6X	GDDR6X
Memory Interface	384-bit	384-bit	384-bit

Nvidia Jetsons

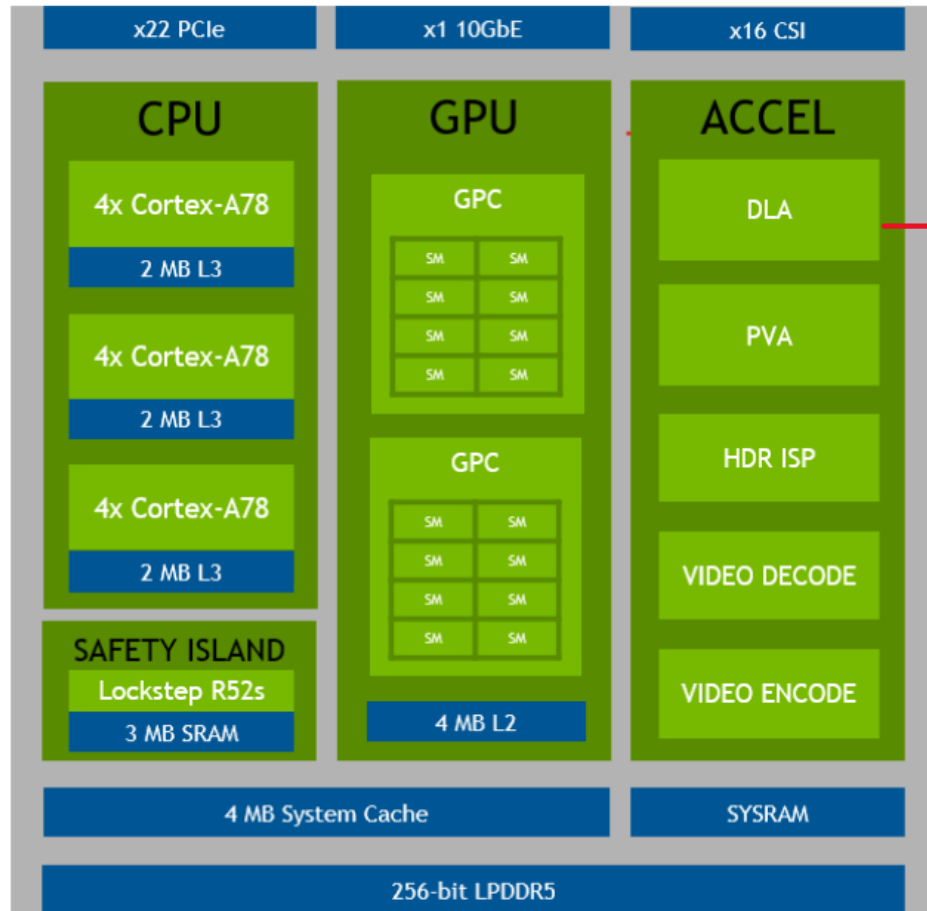
- Nvidia RTX 3090, 4090, **A100**, **H100**

 NVIDIA Data-Center GPUs Specifications				
VideoCardz.com	NVIDIA H100	NVIDIA A100	NVIDIA Tesla V100	NVIDIA Tesla P100
Picture				
GPU	GH100	GA100	GV100	GP100
Transistors	80B	54.2B	21.1B	15.3B
Die Size	814 mm ²	828 mm ²	815 mm ²	610 mm ²
Architecture	Hopper	Ampere	Volta	Pascal
Fabrication Node	TSMC N4	TSMC N7	12nm FFN	16nm FinFET+
GPU Clusters	132/114*	108	80	56
CUDA Cores	16896/14592*	6912	5120	3584
L2 Cache	50MB	40MB	6MB	4MB
Tensor Cores	528/456*	432	320	–
Memory Bus	5120-bit	5120-bit	4096-bit	4096-bit
Memory Size	80 GB HBM3/HBM2e*	40/80GB HBM2e	16/32 HBM2	16GB HBM2
TDP	700W/350W*	250W/300W/400W	250W/300W/450W	250W/300W
Interface	SXM5/*PCIe Gen5	SXM4/PCIe Gen4	SXM2/PCIe Gen3	SXM/PCIe Gen3
Launch Year	2022	2020	2017	2016

Nvidia Jetsons

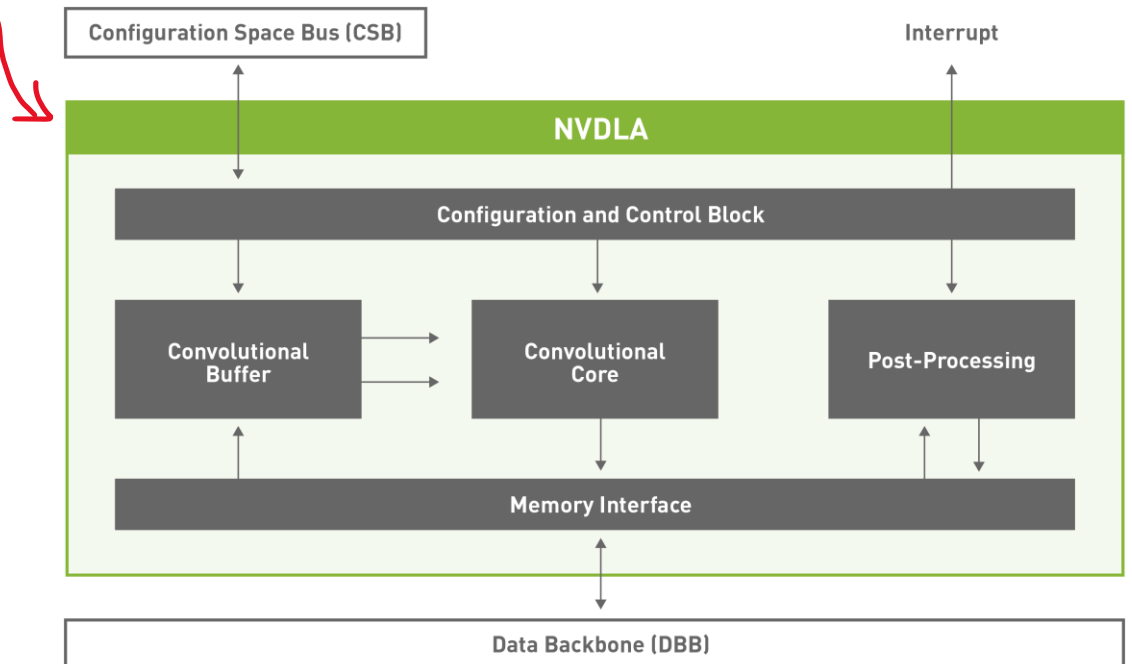
- Nvidia Jetson Family

Figure 2: Orin System-on-Chip (SoC) Block Diagram



NOTE: Jetson AGX Orin 32GB will have 2x 4 Core Clusters, and 7 TPCs with 14 SMs

- GPC:** Graphics Processing Cluster
- DLA:** Deep Learning Accelerator
- PVA:** Programmable Vision Accelerator is a processor in NVIDIA® Jetson AGX Xavier™ and NVIDIA® Jetson Xavier™ NX devices that is specialized for image processing and computer vision algorithms



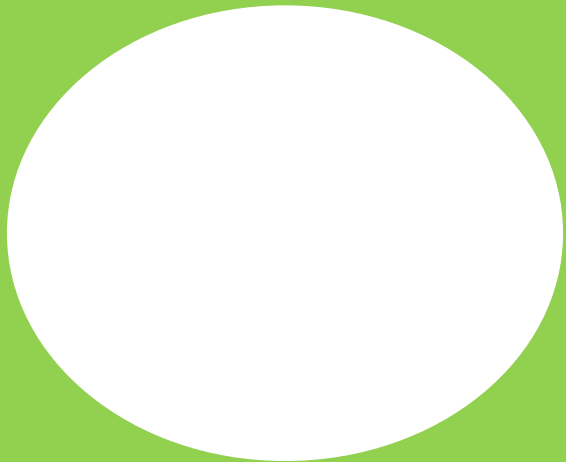
AI Model Performance on a Jetson (FPS)

	사용 모델	고성능		중간 성능				저성능			
Model	Jetson AGX Orin 32GB	Jetson AGX Orin 64GB	성능 비율	Jetson Orin NX 8GB	성능 비율	Jetson Orin NX 16GB	성능 비율	Jetson Orin Nano 4GB	성능 비율	Jetson Orin Nano 8GB	성능 비율
Inveption_V4	1337.8	1702.6	1.27	593	0.44	769	0.57	182	0.14	361	0.27
VGG19	937	1471	1.57	442	0.47	532	0.57	174	0.19	361	0.39
Super_resolution	610	882	1.45	280	0.46	386	0.63	102	0.17	203	0.33
UNET-sgmentation	387	584	1.51	183	0.47	217	0.56	76	0.20	148	0.38
Pose Estimation	1424	2048	1.44	665	0.47	800	0.56	280	0.20	546	0.38
Yolov3-tiny	2611	3179	1.22	1156	0.44	1440	0.55	371	0.14	731	0.28
Resnet50	3717	4834	1.30	1725	0.46	2183	0.59	621	0.17	1158	0.31
SSD-Mobilnet	6415	7671	1.20	2893	0.45	3457	0.54	1094	0.17	2156	0.34
SSD_Resnet34_1200 x1200	120	163	1.36	52	0.43	72	0.60	18	0.15	34	0.28
Yolov5m	342	519	1.52	162	0.47	193	0.56	69	0.20	131	0.38
Yolov5s	785	1135	1.45	379	0.48	449	0.57	158	0.20	301	0.38
평균 성능 비율	1		1.39		0.46		0.57		0.17		0.34

- These Benchmarks were run using [Jetpack 5.1.1](#)
- Each Jetson module was run with maximum performance (Max Frequencies in MAXN for JAO64, JAO32, ONX16, ONX8; and 15W mode for JON8, and 10W mode for JON4)
- Steps to reproduce these results can be found [here](https://github.com/NVIDIA-AI-IOT/jetson_benchmarks/) (https://github.com/NVIDIA-AI-IOT/jetson_benchmarks/)

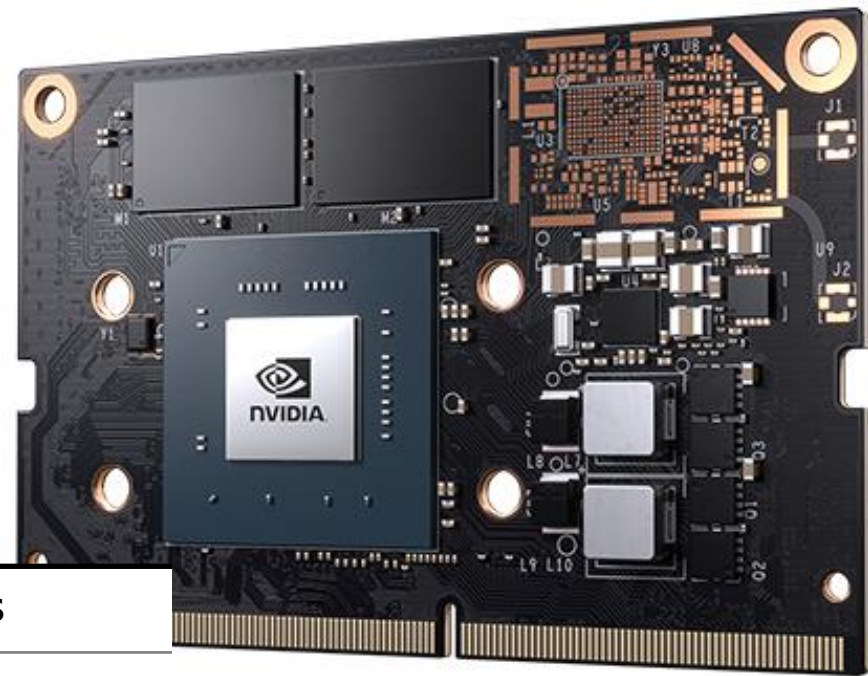


Jetson Nano System 설정



Jetson Nano에 대해...

- Nvidia Jetson Nano

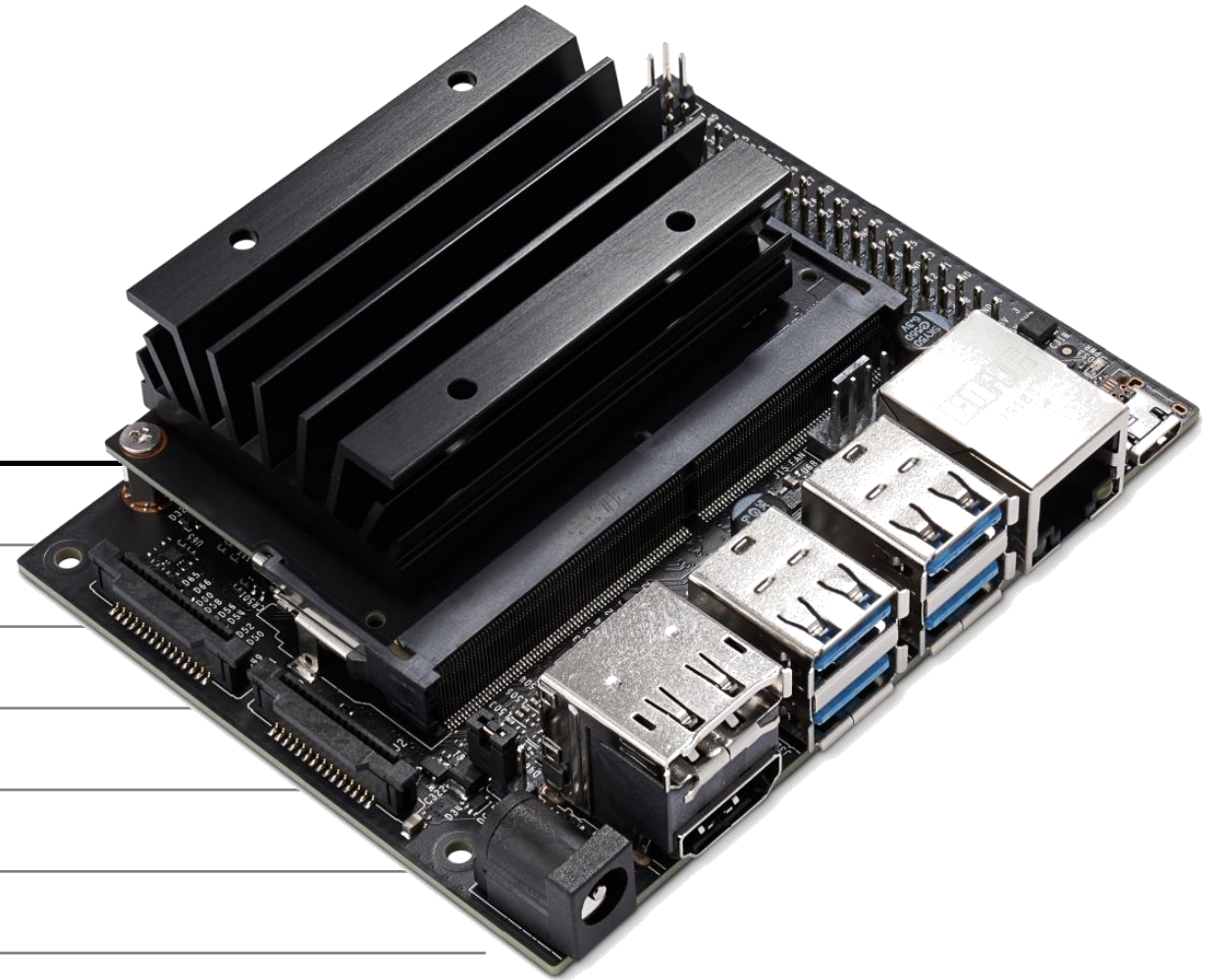


GPU	NVIDIA Maxwell architecture with 128 NVIDIA CUDA® cores
CPU	Quad-core ARM Cortex-A57 MPCore processor
Memory	4 GB 64-bit LPDDR4, 1600MHz 25.6 GB/s
Storage	16 GB eMMC 5.1
Camera	12 lanes (3x4 or 4x2) MIPI CSI-2 D-PHY 1.1 (1.5 Gb/s per pair)
Connectivity	Gigabit Ethernet, M.2 Key E
Display	HDMI 2.0 and eDP 1.4
USB	4x USB 3.0, USB 2.0 Micro-B
Others	GPIO, I ² C, I ² S, SPI, UART
Mechanical	69.6 mm x 45 mm 260-pin edge connector

Jetson Nano에 대해...

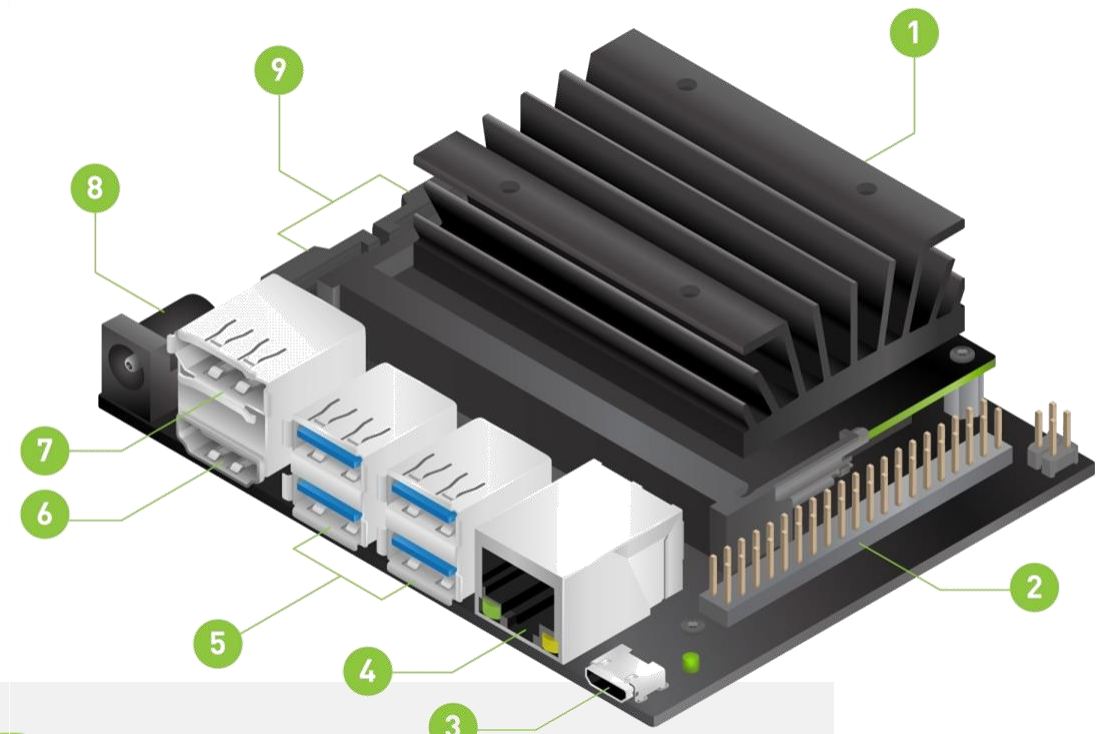
- Nvidia Jetson Nano

GPU	128-core Maxwell
CPU	Quad-core ARM A57 @ 1.43 GHz
Memory	4 GB 64-bit LPDDR4 25.6 GB/s
Storage	microSD (not included)
Camera	2x MIPI CSI-2 DPHY lanes
Connectivity	Gigabit Ethernet, M.2 Key E
Display	HDMI and display port
USB	4x USB 3.0, USB 2.0 Micro-B
Others	GPIO, I ² C, I ² S, SPI, UART
Mechanical	69 mm x 45 mm, 260-pin edge connector



Jetson Nano에 대해...

- Nvidia Jetson Nano



1 microSD card slot for main storage

2 40-pin expansion header

3 Micro-USB port for 5V power input, or for Device Mode

4 Gigabit Ethernet port

5 USB 3.0 ports (x4)

6 HDMI output port

7 DisplayPort connector

8 DC Barrel jack for 5V power input

9 MIPI CSI-2 camera connectors

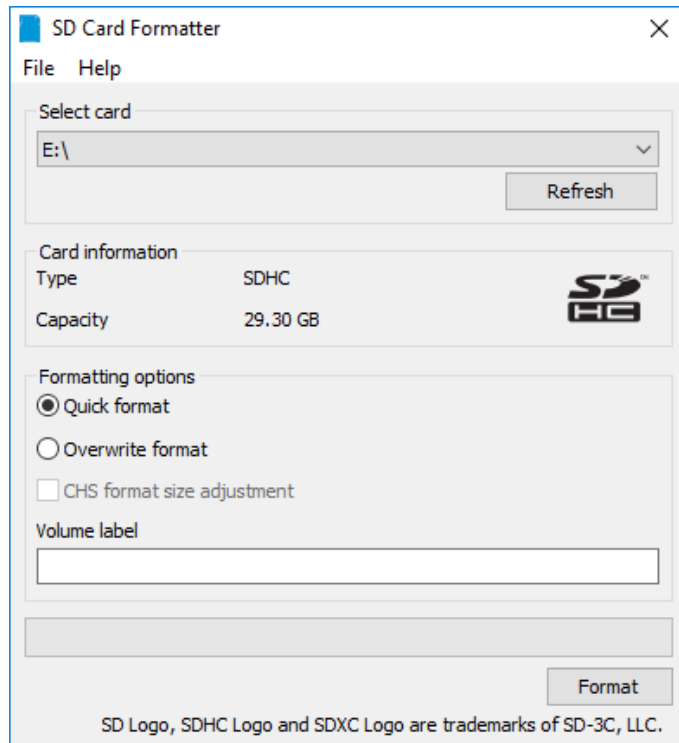
<https://developer.nvidia.com/embedded/learn/get-started-jetson-nano-devkit>

Jetson Nano에 대해...

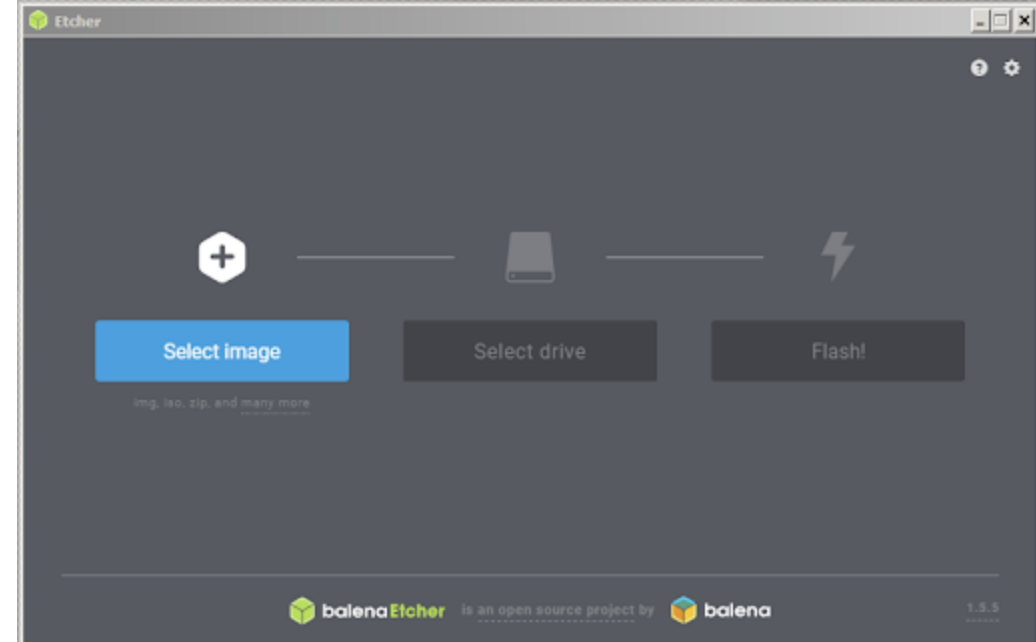
- Nvidia Jetson Nano에 OS 설치하자!

<https://developer.nvidia.com/embedded/learn/get-started-jetson-nano-devkit>

- Download the [Jetson Nano Developer Kit SD Card Image](#)
- Write the image to your microSD card



<https://etcher.balena.io/>



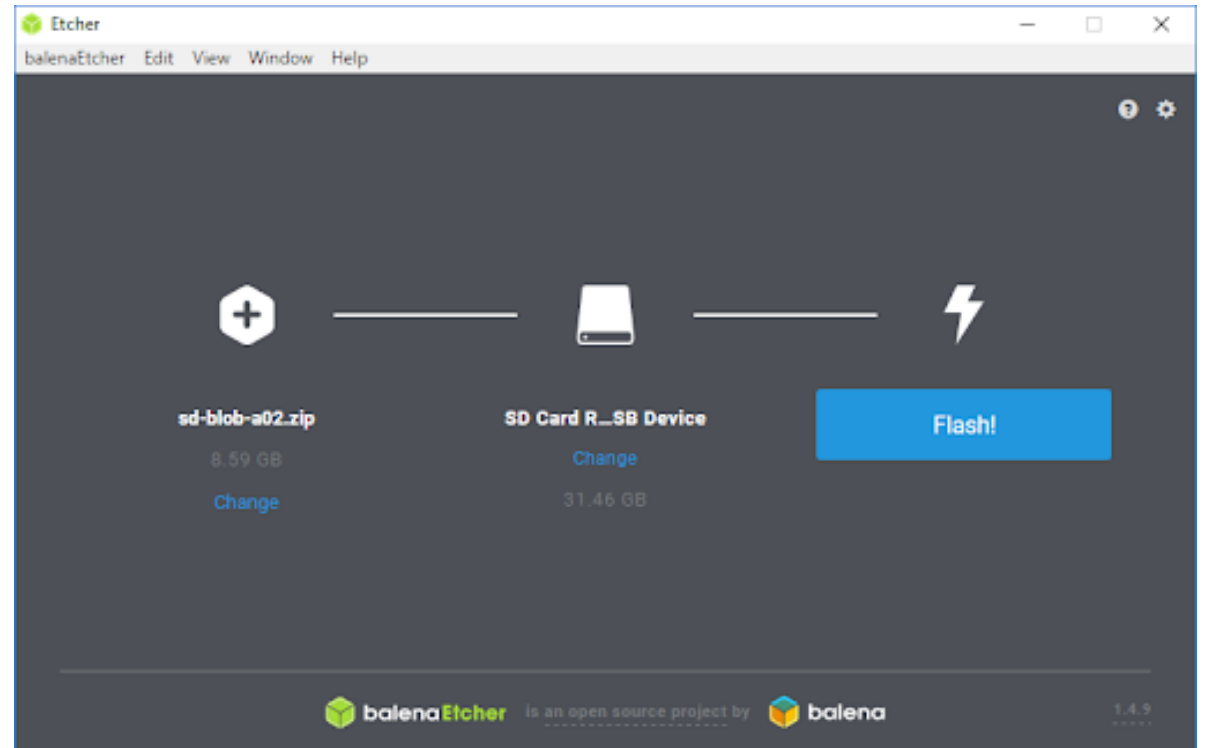
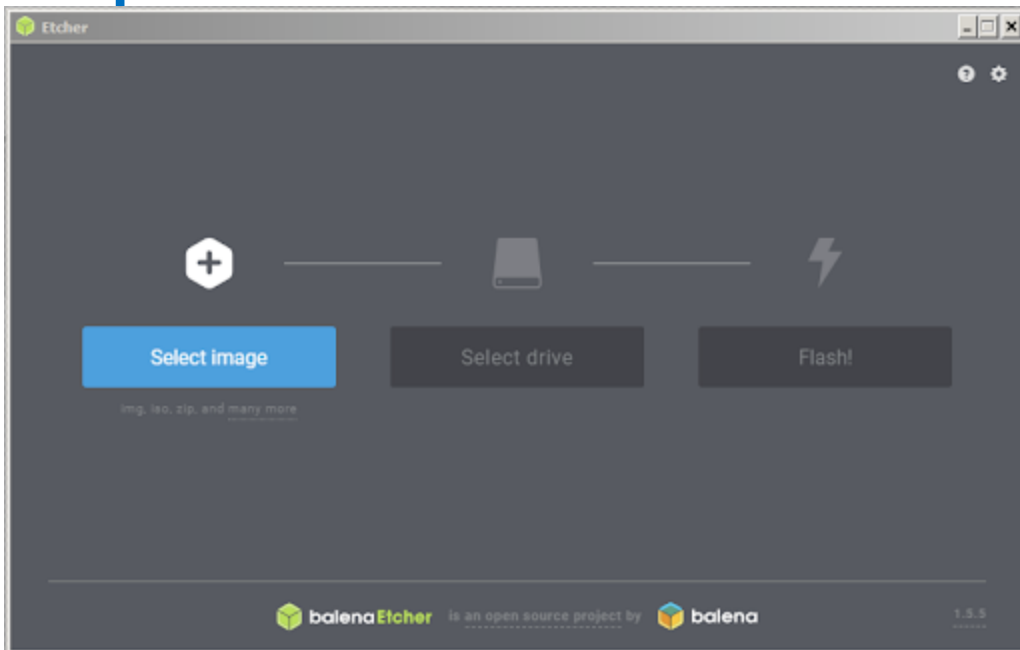
Jetson Nano에 대해...

- Nvidia Jetson Nano에 OS 설치하자!

<https://developer.nvidia.com/embedded/learn/get-started-jetson-nano-devkit>

- Download the [Jetson Nano Developer Kit SD Card Image](#) (**Jetpack 4.6.1**)
- Write the image to your microSD card

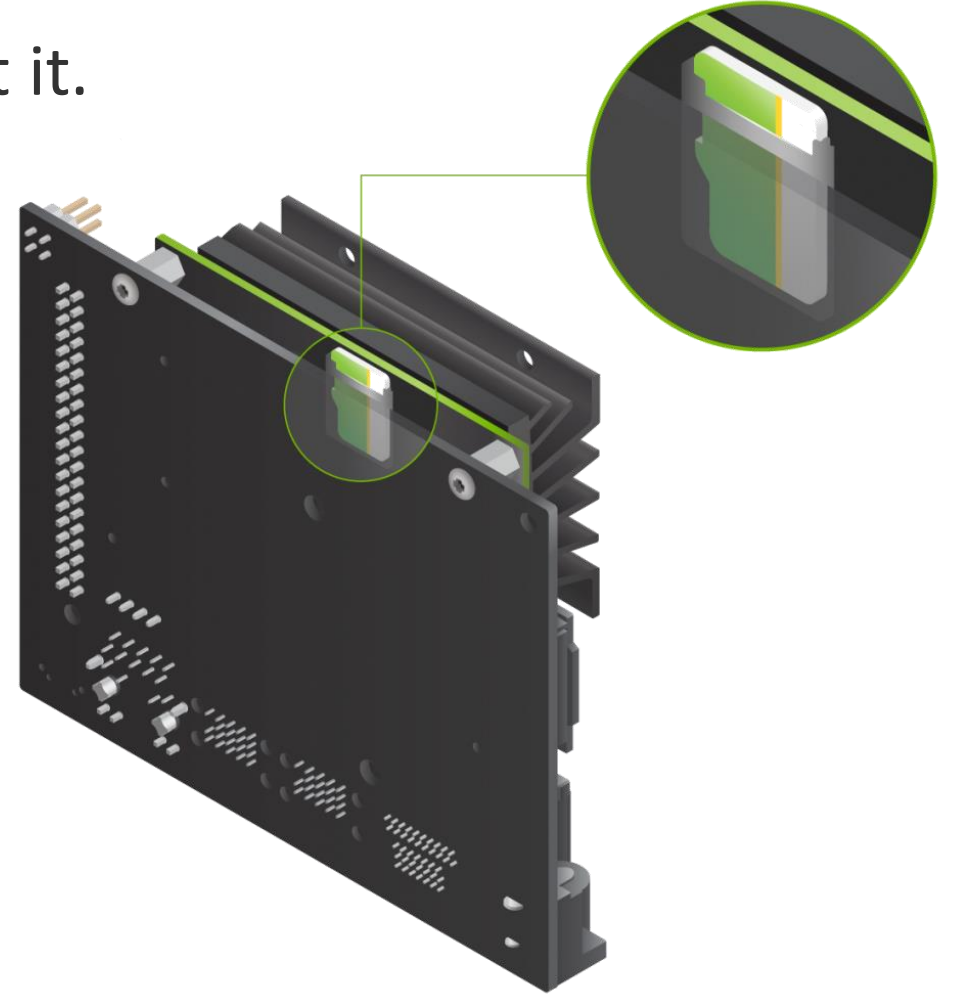
<https://etcher.balena.io/>



Jetson Nano의 전원을 올려볼까요 ?

Jetson Nano의 전원을 올려볼까요 ?

- Insert the microSD card.
- Power on your computer display and connect it.
- Connect the USB keyboard and mouse.
- Connect your Micro-USB power supply



Jetson Nano 부팅부팅~

- A green LED next to the Micro-USB connector will light as soon as the developer kit powers on.
- When you boot the first time, the developer kit will take you through some initial setup, including:
 - Review and accept NVIDIA Jetson software EULA (사용자 라이선스 동의)
 - Select system language, keyboard layout, and time zone
 - Create username, password, and computer name
 - Select APP partition size—it is recommended to use the max size suggested



<https://developer.nvidia.com/embedded/learn/get-started-jetson-nano-devkit#setup>

Jetson Nano 부팅부팅~

• jtop 설치

sudo apt-get update

sudo apt-get upgrade

sudo apt-get install python-pip

jetson-stats 설치

sudo -H pip install -U jetson-stats

재부팅

sudo reboot now

jetpack 버전 확인 및 cpu, 메모리 cuda,

opencv 정보까지 확인 가능

```
Model: NVIDIA Jetson Nano Developer Kit - Jetpack 4.6.4 [L4T 32.7.4]
1 [|||||] 11.0%] 204MHz 3 [|||] 3.0%] 307MHz
2 [|||||] 15.6%] 307MHz 4 [|||] 1.0%] 614MHz
Mem [|||||] 631M/3.9G] FAN [|||||] 0.0%] 0RPM
Swp [|||||] 0.0k/1.9G] Jetson Clocks: inactive
Emc [|||||] 1.6GHz 0% NV Power[0]: MAXN
Iram [|||||] 0.0k/252k](lfb 252kB) Uptime: 8 days 7:4:47
GPU [|||||] 0.0%] 76.8MHz
Dsk [|||||] 15.9G/58.4G]

#####
PID USER GPU TYPE PRI S CPU% MEM [GPU MEM] Command
4444 root I G 20 S 0.0 7.8M 17.8M Xorg
5086 root I G 20 S 0.1 28.1M 6.4M gnome-shell
3997 root I G 20 S 0.0 4.7M 76.0k nvargus-daemon

[HW engines] [Sensor] [Temp] [Power] [Inst] [Avg]
APE: 25.5MHz GPU 17.00C POM 5V CPU 416mW 403mW
NVENC: [OFF] NVDEC: [OFF] thermal 14.75C POM 5V GPU 0.0mW 0.0mW
NVJPG: [OFF] SE: [OFF] PLL 13.00C POM_5V_IN 1.5W 1.5W
CPU 13.50C
A0 15.50C

1ALL 2GPU 3CPU 4MEM 5ENG 6CTRL 7INFO Quit (c) 2023, RB
```

port MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

Jetson Nano 부팅부팅~

- Jtop을 통해 시스템의 정보를 살펴보기

Model: NVIDIA Jetson Nano Developer Kit - Jetpack 4.6.4 [L4T 32.7.4]
Temperatures: GPU 17.00C
iGPU gpu 0% - gov: nvhost_podgov

GPU Shared RAM 24.2M/3.9GB

100%
91%
83%
75%
66%
58%
50%
41%
33%
25%
16%
8%

3.9G
3.6G
3.2G
2.9G
2.6G
2.3G
2.0G
1.6G

L -8s L -6s L -4s L -2s

3D scaling:[Active] Railgate: Active

Frq [76.8MHz]

PID	USER	GPU	TYPE	PRI	S	CPU%	MEM
4444	root	I	G	20	S	0.0	7.8M
5086	root	I	G	20	S	0.1	28.1M
3997	root	I	G	20	S	0.0	4.7M

1ALL 2GPU 3CPU 4MEM 5ENG 6CTRL 7INFO Quit

jtop 4.2.3 - (c) 2023, Raffaello Bonghi [raffaello@rnext.it]
Website: https://rnext.it/jetson_stats

Platform

Machine: aarch64
Release: 4.9.337-tegra
Distribution: Ubuntu 18.04 bionic
Python: 2.7.17
System: Linux

Serial Number: [s|XX CLICK TO READ XXX]
Hardware

SoC: tegra210
L4T: 32.7.4
Jetpack: 4.6.4
P-Number: p3448-0000
699-level Part Number: 699-13448-0000-402 K.0
Module: NVIDIA Jetson Nano (4 GB ram)
CUDA Arch BIN: 5.3
BoardIDs: p3448
Model: NVIDIA Jetson Nano Developer Kit
Codename: Porg

Libraries

cuDNN: 8.2.1.32
TensorRT: 8.2.1.8
CUDA: 10.2.300
VPI: 1.2.3
OpenCV: 4.1.1 with CUDA: NO

Hostname: aiacnano-desktop
Interfaces
docker0: 172.17.0.1
eth0: 192.168.0.10

Model: NVIDIA Jetson Nano Developer Kit - Jetpack 4.6.4 [L4T 32.7.4]
ALL [|||||]
1 7% Schedutil

3 3% Schedutil

9.1%

100%
87%
75%
62%
50%
37%
25%
12%

L -8s L -6s L -4s L -2s

RMv8 Processor rev 1 (v8l)

Frq [102MHz]

1.5GHz] 825MHz

24% Schedutil

100%
87%
75%
62%
50%
37%
25%
12%

L -8s L -6s L -4s L -2s

RMv8 Processor rev 1 (v8l)

Frq [102MHz]

1.5GHz] 921MHz

it (c) 2023, RB

Libraries

cuDNN: 8.2.1.32
TensorRT: 8.2.1.8
CUDA: 10.2.300
VPI: 1.2.3
OpenCV: 4.1.1 with CUDA: NO

SoC: tegra210
L4T: 32.7.4
Jetpack: 4.6.4

ort MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>