CAICT 中国信通院

The evaluation of Artificial Intelligent chips

China Academy of Information and Communication Technology (CAICT)
Artificial Intelligence Department

Zhang Weimin

2022.1

中国信息通信研究院
http://www.caict.ac.cn

CONTENT

1 AI CHIP Industry

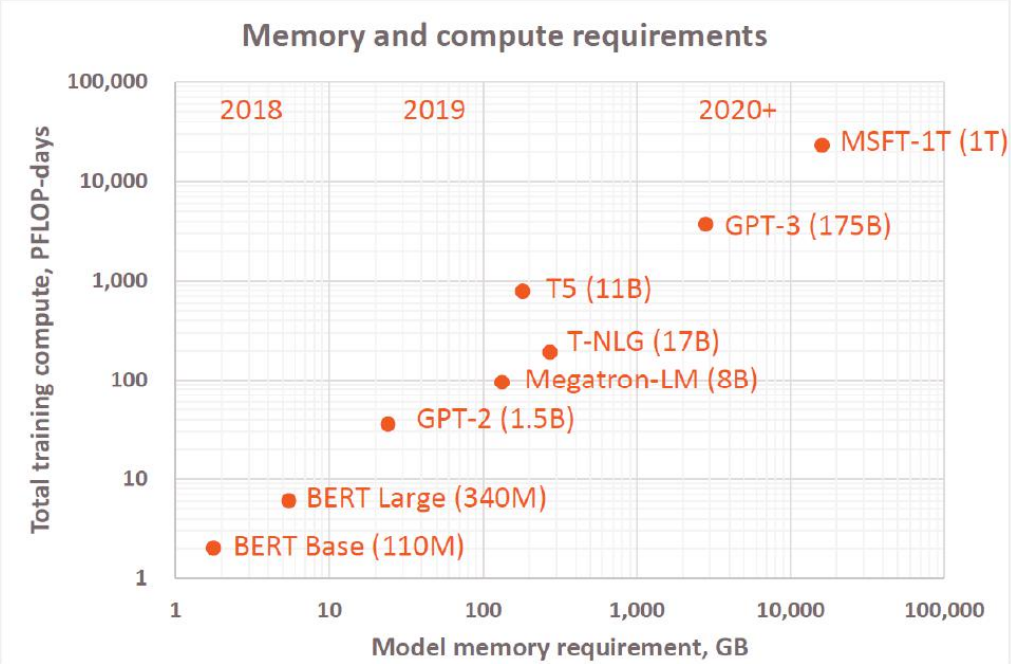
- **Background**
- **Industry analysis**

Background： huge & diversity of computing demand

Internet of everything and Intelligence are the trend. The demand for computing is increasing hundreds of times. In many typical application scenarios, such as autonomous driving, machine learning and augmented reality, computing has become the bottleneck restricting innovation.

High performance ： x-10x TFLOPS
Greater specificity ： INT+FLOAT+parallel +Serial
More Scenarios： Low power consumption, high reliability and low delay

Huge Model & Dataset & Compute



1000x **larger models**

1000x **more compute**

In just **2 years**

Today, GPT-3 with 175 billion params trained on 1024 GPUs for 4 months.

Tomorrow, **multi-trillion** parameter models and beyond.

Source: sambanova@HC33

The demand of AI compute

DSAs for DNNs need to perform at least these matrix-oriented operations well: vector-matrix multiply, matrix-matrix multiply, and stencil computations. They will also need support for the nonlinear functions, which include at a minimum ReLU, Sigmoid, and tanh.

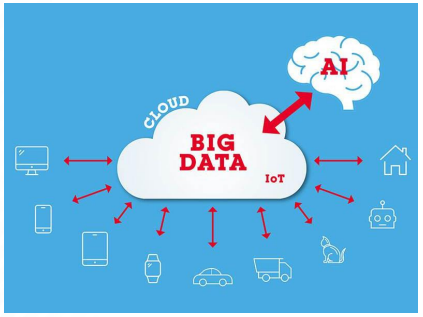
demand

• High-dimensional, large vector data type • Domain-specific computing • Graph topology	High memory
• Massive parallel computing • High throughput • Huge data relocation	High bandwidth
• Low latency • Real-time demands	Strong computing power
• Algorithm fast iteration • Model diversification and differentiation • The difference between cloud and end computing requirements	Reconfigurable ability

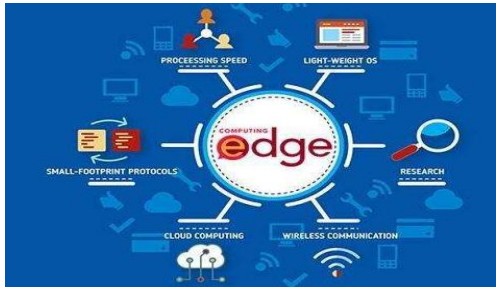
AI Chips Have Enabled Many Fields

At present, AI chips are mainly used in cloud training, cloud inference, terminal inference and other fields. At the same time, AI chips have been widely used in cloud computing, automatic driving, intelligent security, smart phones and other fields.

Cloud Computing



Edge Computing



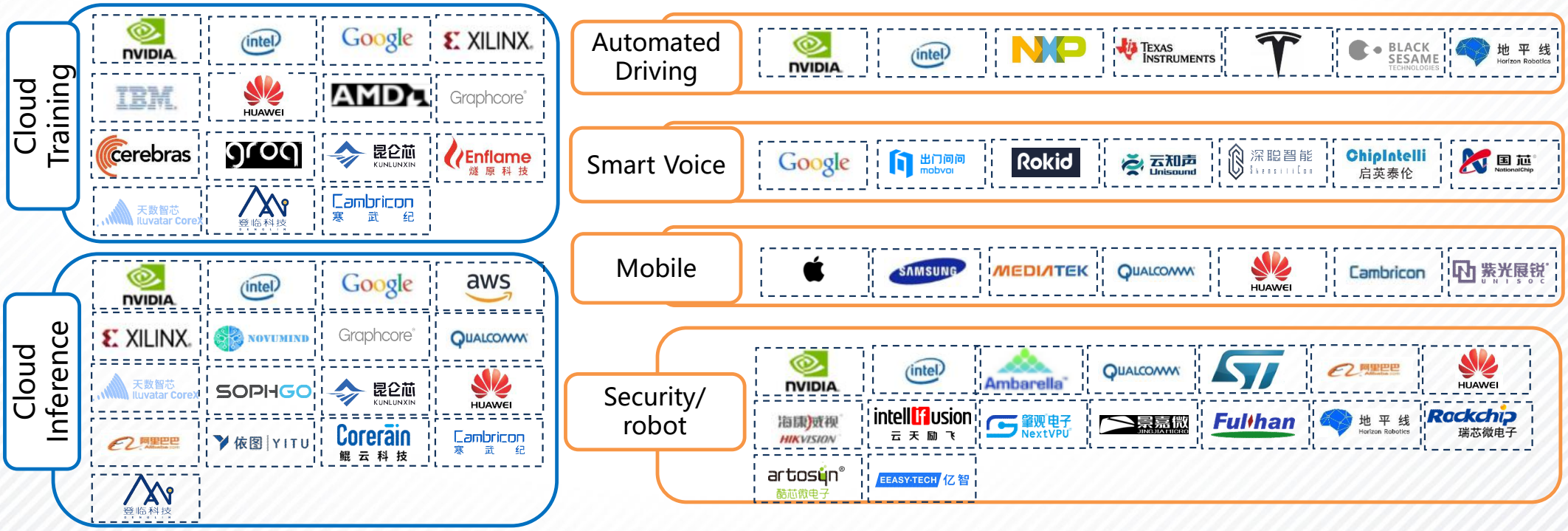
Automated Driving

Intelligent security

Smartphone

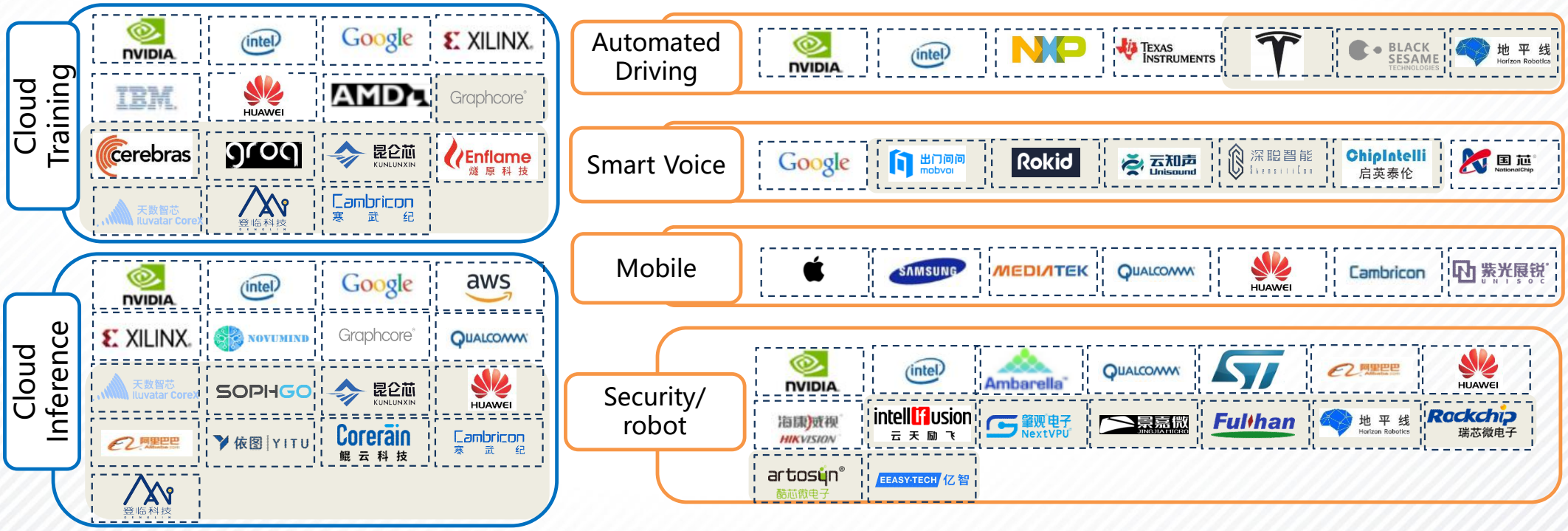
AI chip landscape

Observations:
Training company 20+, inference company 100+



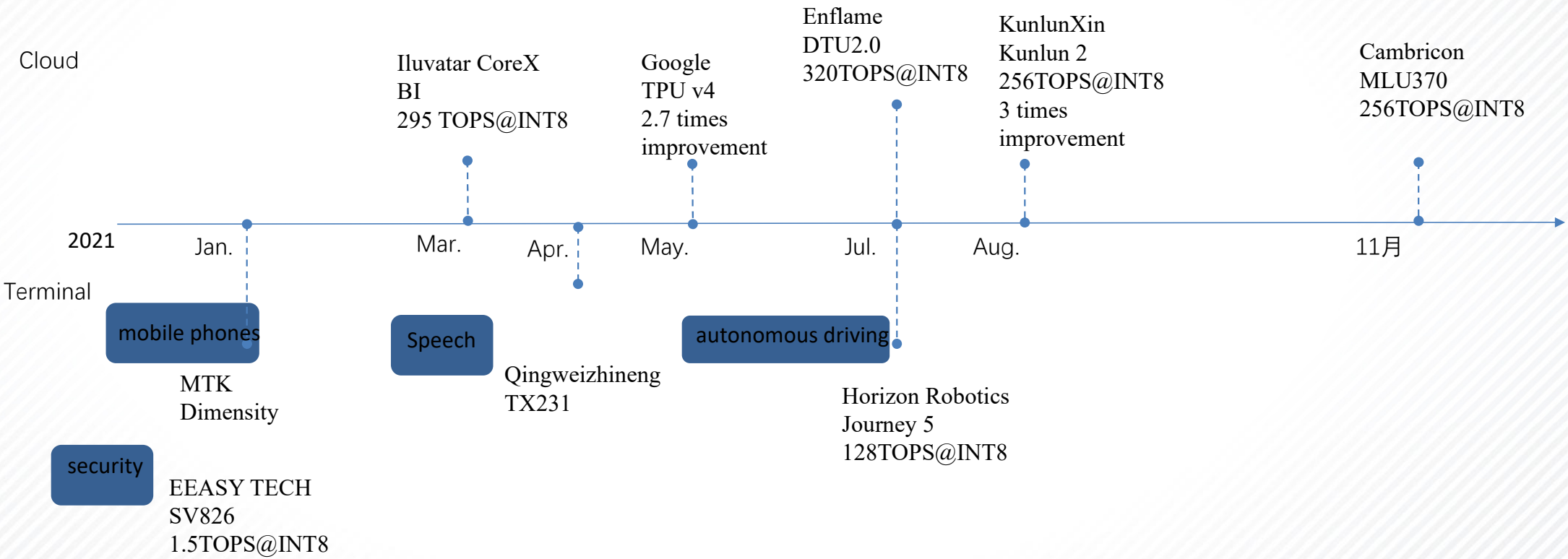
AI chip landscape

Observations:
Training company 20+, inference company 100+
Non-traditional players play important roles
Huge number of startups

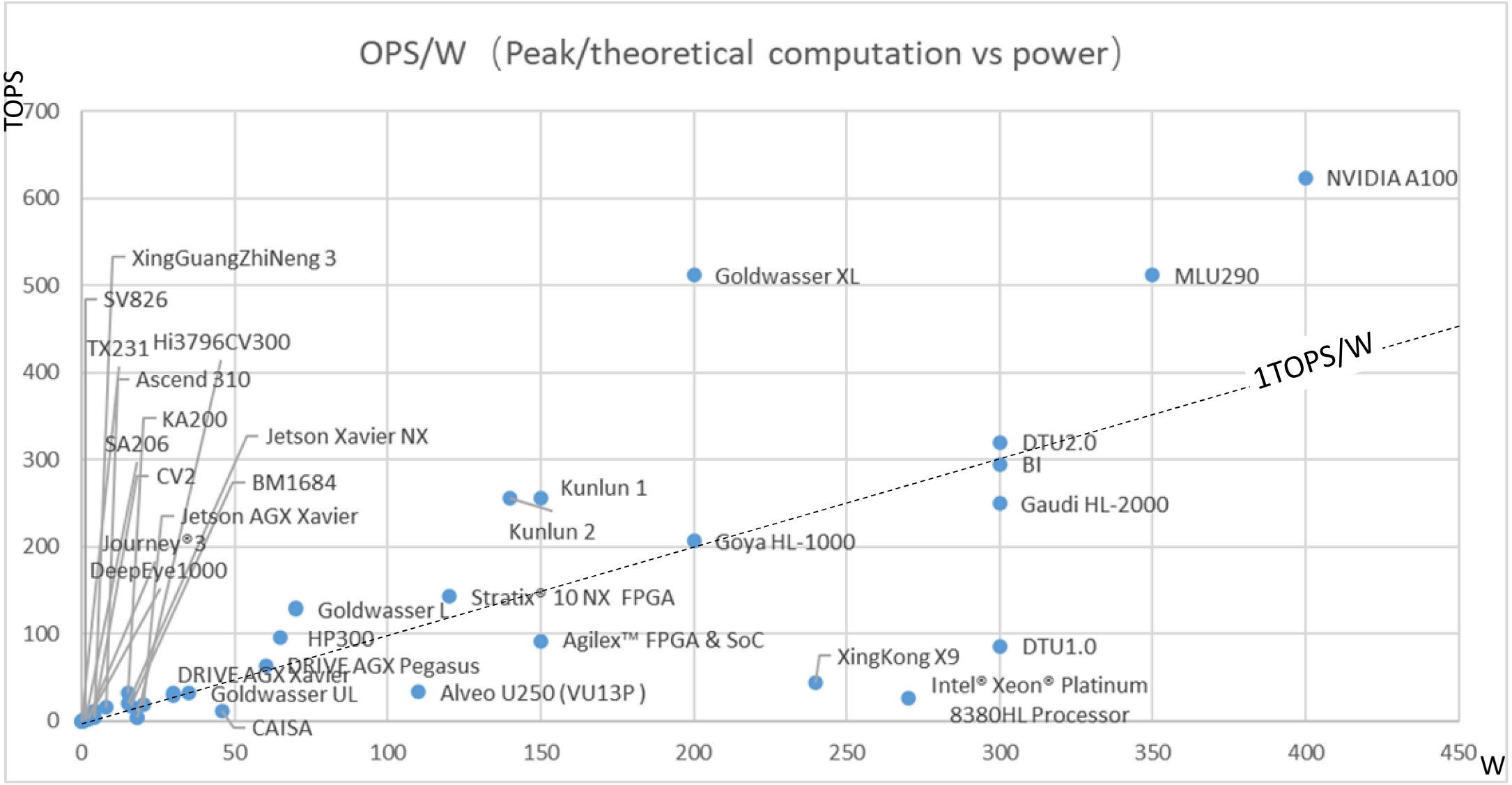


Deeply Empowering Different Scenarios, AI Enters the Age of Computing ability Customization

Cloud training chip has become the focus of domestic layout, with 2-4 times higher performance than the previous generation



Inference chips focuses on vertical application scenario customization, enabling mobile phones, security, smart homes, autonomo driving , etc.



Note: The data source is the schematic diagram of the relationship between theoretical peak computing power (INT8) and maximum power consumption/peak power consumption in the Catalogue of AI Chip Technology Selection (July 2020) and (2021), which does not show that the product is an undisclosed product.

AI chip industry challenge

First…Up to x times the current level of improvement.

manufacturer

High performance，xxxTPOS ！

DSA, Optimal hardware architecture design

Third party evaluation
To build a bridge between users and AI chip companies

Deployment, adaptation, and scalability

USER

Satisfy the needs of real computing.

Price–performance ratio

2 AI CHIP evaluation

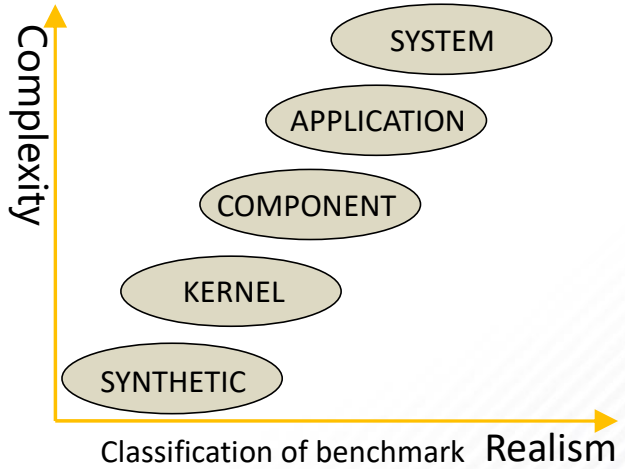
- **Significance**
- **Status**

Defination

- ✓ to assess the relative performance……
- ✓ normally by running a number of standard tests and trials against it
- ✓ Evaluation method+ metrics
- ✓ provide a method of comparing the performance of various subsystems across different chip/system architectures.

Types of benchmark

- ✓ Real program
- ✓ Component Benchmark / Microbenchmark
- ✓ ……



Source: SPEC 2016 summit

Significance

- ✓ Important reference metrics for user selection
- ✓ Guide the direction of innovation for manufacturer

AI chip benchmark challenge

Fragmentation of application scenarios 04

- ✓ Speech semantics, image and video, etc.
- ✓ Mobile terminal, video surveillance, etc.

Algorithm iterative fast 02

- ✓ The algorithm is constantly updated.
- ✓ Algorithm security is not considered.

Benchmarking test data 05

- ✓ Public + private data set
- ✓ Standard test data set

Multiple frameworks 03

- ✓ Kinds of training framework
- ✓ Fragmentation of inference framework

Diverse product 01

- ✓ Chip types and functions
- ✓ Device types and functions

AI chip benchmark challenge

At present, the application scenarios of deep learning, such as vision, language, sound, recommendation, are wide, different and fragmented.

Area	Vision	Language	Audio	Commerce	Action / RL	Other
Problem	Image Classification Object Detection / Segmentation Face ID HealthCare (Radiology) Video Detection Self-Driving	Translation Language Model Word Embedding	Speech Recognition Text-to-Speech Question Answering Keyword Spotting Language Modeling Chatbots Speaker ID Graph embeddings Content ID	Rating Recommendations Sentiment Analysis Next-action Healthcare (EHR) Fraud detection Anomaly detection Time series prediction Large scale regression	Games Go Robotics Health Care Bioinformatics	GANs 3D point clouds Word embeddings
Datasets	ImageNet COCO	WMT English-German	LibriSpeech SQuAD LM-Benchmark	MovieLens-20M Amazon IMDB	Atari Go Chess Grasping	
Models	ResNet-50 TF Object Detection Detectron	Transformer OpenNMT	Deep Speech 2 SQuAD Explorer	Neural Collaborative Filtering CNNs	DQN PPO	
Metrics	COCO mAP Prediction accuracy	BLEU	WER Perplexity	Prediction accuracy	Prediction accuracy Win/Loss	

AI benchmark landscape

Before 2017

Deepbench

Cambricon 寒武纪

DAWNBench

An End-to-End Deep Learning Benchmark and Competition

BenchIP

Baidu 百度

2017~

AI - Benchmark

AI Matrix.

Mobile AI Bench

DNN benchmark

ML C

AIBench

AI-perf

...

The attention of benchmark is constantly increasing.

More and more research institutions and companies are aware that the establishment of benchmark is of great significance to promote the healthy development of the industry.

no generally acknowledge benchmark methods and metrics

At present, only reaching a local agreement. A fair and scientific benchmark still needs efforts from all sides.

3 AIIA DNN benchmark

- **Brief introduction**
- **Standard-F.748.11**

Formation of AIIA

China Artificial Intelligence Industry Alliance was founded on Oct.13th, 2017. Until now, There are 728 memberships, including 31 vice-chairman entities, 191 directors entities, 506 ordinary member entities after the foundation of the alliance.

Main Task: To gather momentum from industry ecology, to establish AI technology, standards and industry synergy, to explore new patents and new mechanisms to promote technology, industry and applications development, make pilot demonstrations, promote international cooperation and establish worldwide cooperative platforms.

To conduct AI related research on policy, regulations, technology, industry, applications and security.

To carry out AI related standards pre-research and testing, promote high standard development and innovations.

International and domestic cooperation and communication

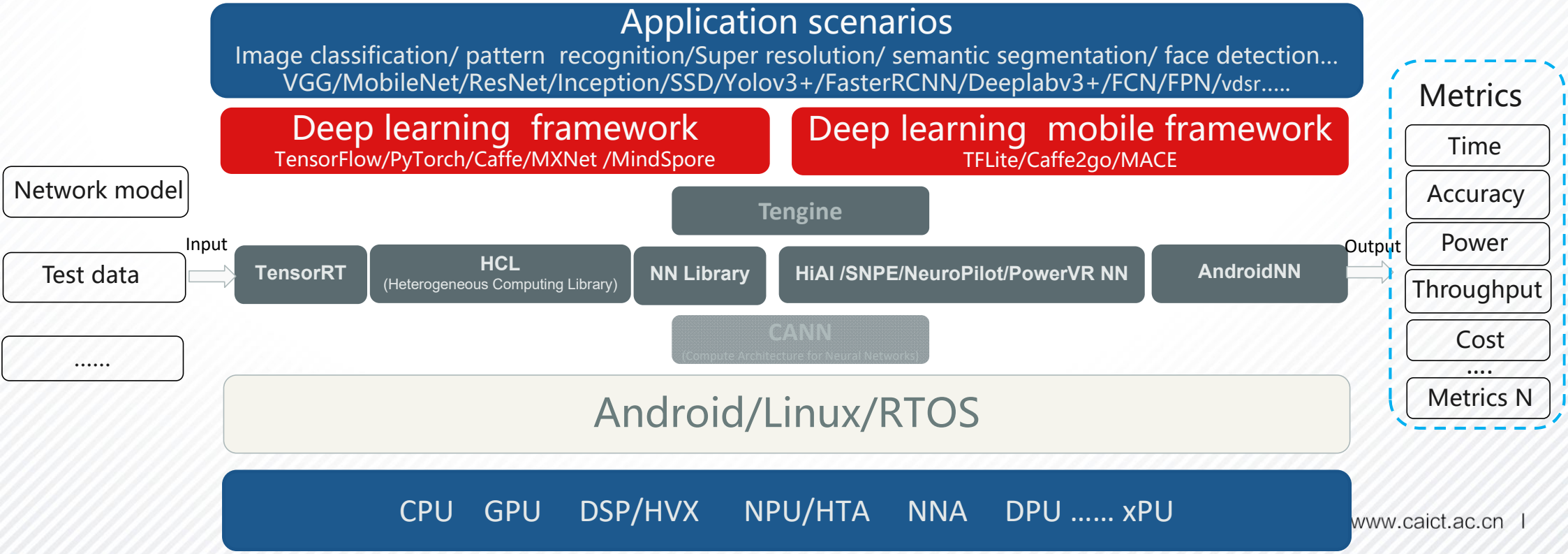
AI related IP research and resource sharing

AIIA DNN benchmark Brief introduction

About AIIA DNN benchmark : Initiated in 2017**,** Provide selection reference for application companies, and provide third-party evaluation results for chip companies.

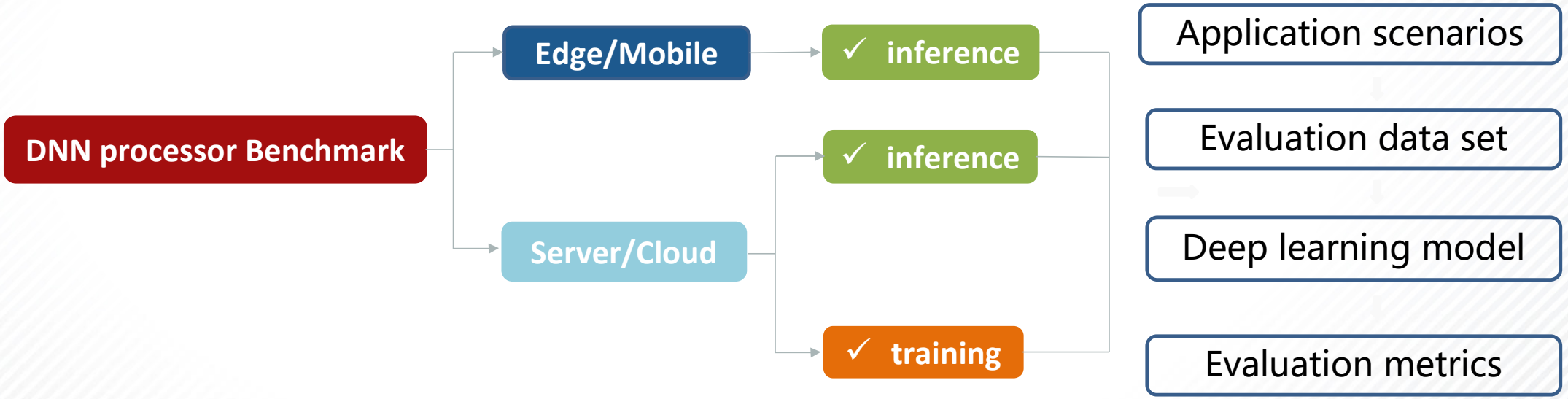
Aims: To objectively reflect the current state of AI accelerator capabilities, and all metrics are designed to provide an objective comparison dimension.

Evaluation Framework:



AIIA DNN benchmark evaluation method

Evaluation method: step-by-step, version iterations, training and inference, mobile and cloud.



AIIA DNN benchmark standard

AIIA evaluation standard

Guidelines of artificial intelligence chip benchmark:

Part 1: Metrics and evaluation methods for edge-based deep neural network processor benchmark (AIIA/P 0056-2020, **completed**)

Part 2: Metrics and evaluation methods for cloud-based inference deep neural network processor benchmark (AIIA/P 0057-2020, **completed**)

Part 3: Metrics and evaluation methods for cloud-based training deep neural network processor benchmark (AIIA/PG 0061-2021, **completed**)

Application-oriented Guidelines of artificial intelligence chip benchmark:

Part 1:Metrics and evaluation methods of neural network processor benchmark in Media Intelligence Center(AIIA/PG 0062-2021, **completed**)

Part2: Metrics and evaluation methods of artificial intelligence chip benchmark in TinyML(ongoing)

Part3: Metrics and evaluation methods of artificial intelligence edge computing box(ongoing)

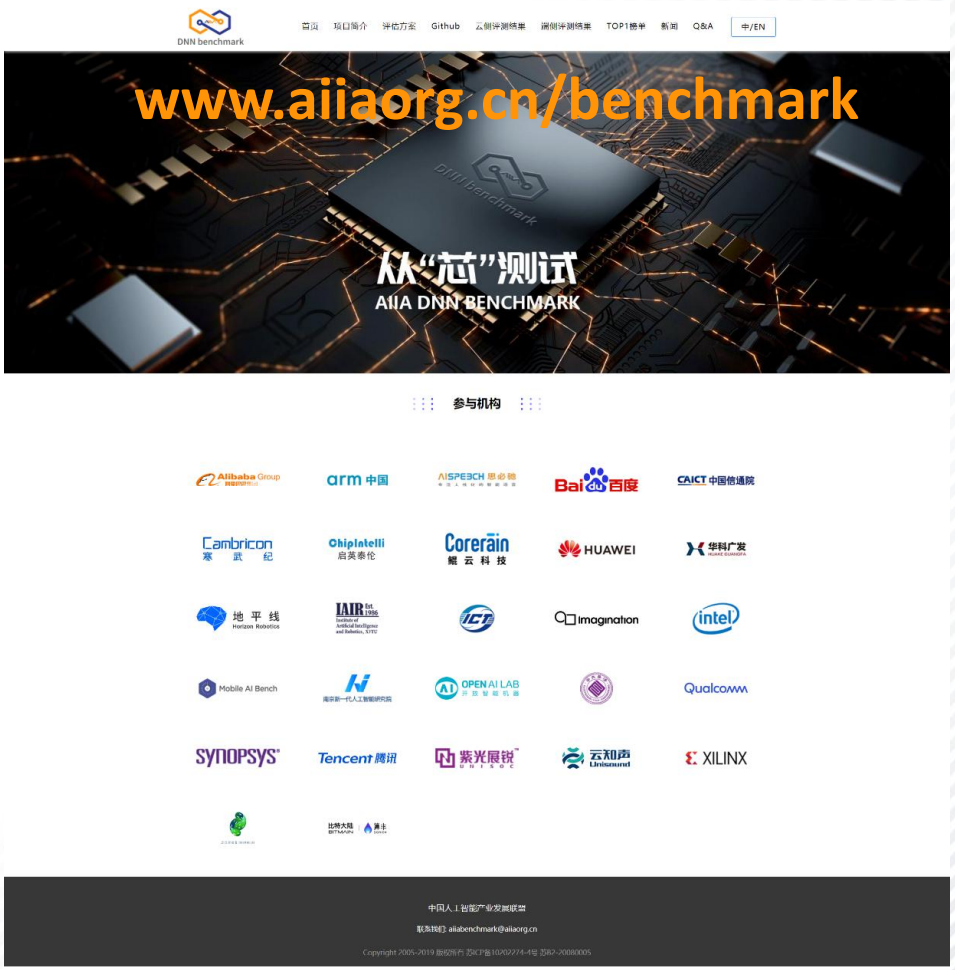
Industry standard：CCSA

WG:TC1 WG1

Project No.:2019-1009T-YD

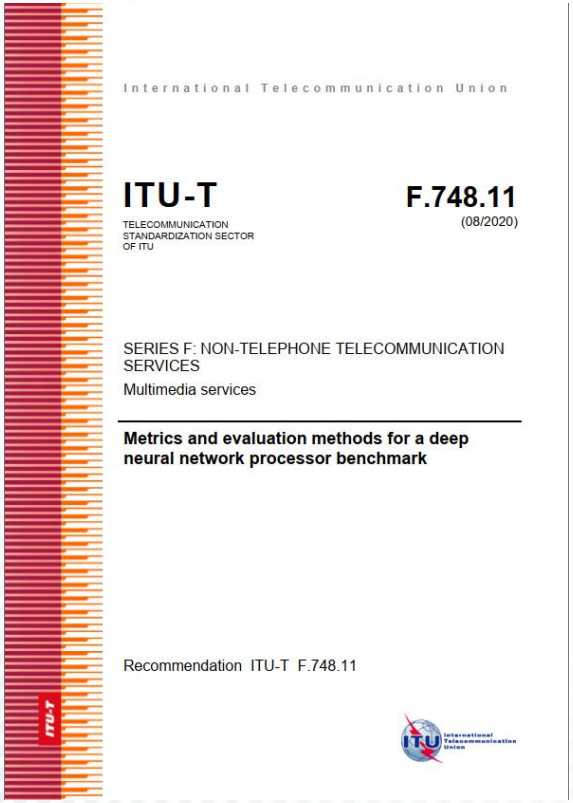
Proposal: Evaluation method for artificial intelligence chip benchmark (Released)

■ Official website



F.748.11

WG: SG16/WP3/Q5
Project No.: F.748.11
Approval : 2020-08-13



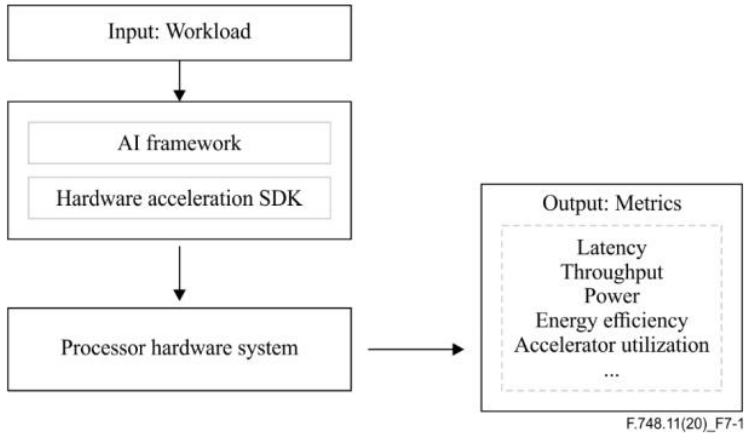


Figure 7-1 – Architecture framework of the deep neural network processor benchmark

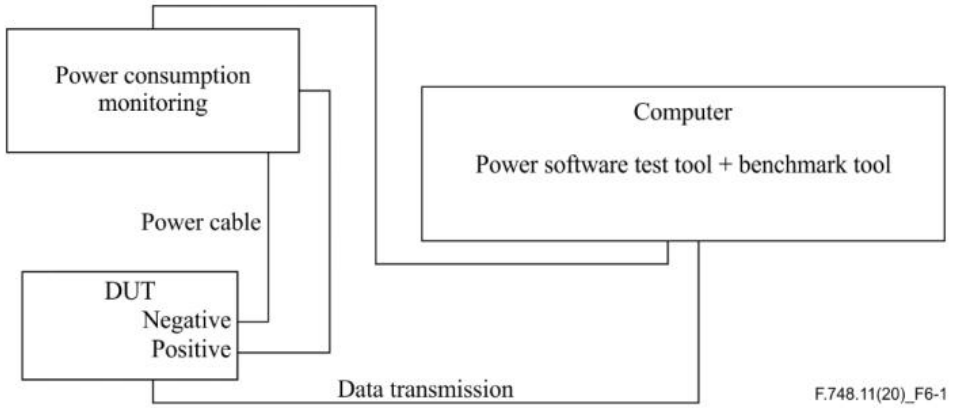


Figure 6-1 – Benchmark test environment

Recommendation ITU-T F.748.11 provides the benchmarking framework, evaluation metrics and methods, and application scenarios for deep neural network processors operating training and inference tasks.

Application scenarios	Test Model	Test data set	Metrics							
			Training (under specific Batch Size)			Inference (under specific Batch Size)				
			Accuracy+Training time	Throughput	Power	Accuracy	Inference time	Throughput	Power	Throughput /power
classification	ResNet50	ImageNet	TOP1/TOP5			TOP1/TOP5				
object detection	SSD	VOC2012	mAP/mIoU			mAP/mIoU				
	Yolo V3	COCO	mAP/mIoU			mAP/mIoU				
semantic segmentation	Deeplabv3+	VOC2012				mIoU				
NLP	BERT	Wikipedia	0.712 Mask-LM accuracy			Sequence length：fixed 128				
machine translation (recursion)	GNMT	WMT16 E-German	24.0 SacreBLEU							
machine translation (non-recursion)	Transformer	WMT16 E-German	25.0 BLEU			mAP/mIoU				
Recommendation system	DLRM	1TB Click Logs	0.8025 AUC							

CAICT 中国信通院 AIIA DNN benchmark project

2018 （v0.5）

Oct.
Edge/Mobile –based inference standard released

2019 (v0.5)

Mar.
Edge/Mobile –inference （v0.5a） evaluation result released

May.
Cloud –Inference standard released

Jun.
Edge/Mobile –inference (v0.5b) evaluation result released

Nov.
Cloud –Inference evaluation result released

2020 （v0.6）

Service 7 companies, including Horizon, Huawei HiSilicon , Corerain, Cambrian, Baidu, Bitmain, ChipIntelli, etc. a total of 10 chip tests.

2021 （v0.7）

Service 4 companies include 6 chip tests.

Technical selection Catalog of AI Chip

- Include **46** AI chips from **19** enterprises.
- 20 AI chips were submitted with verification test results, 6 AI chip were evaluated by AIIA DNN Benchmark

- Include **44** AI chips from **27** enterprises.
- 2 AI chip were evaluated by AIIA DNN Benchmark

CAICT 中国信通院

A specialized think-tank for the government and an innovation and development platform for the industry

Thank you!

中国信息通信研究院
http://www.caict.ac.cn