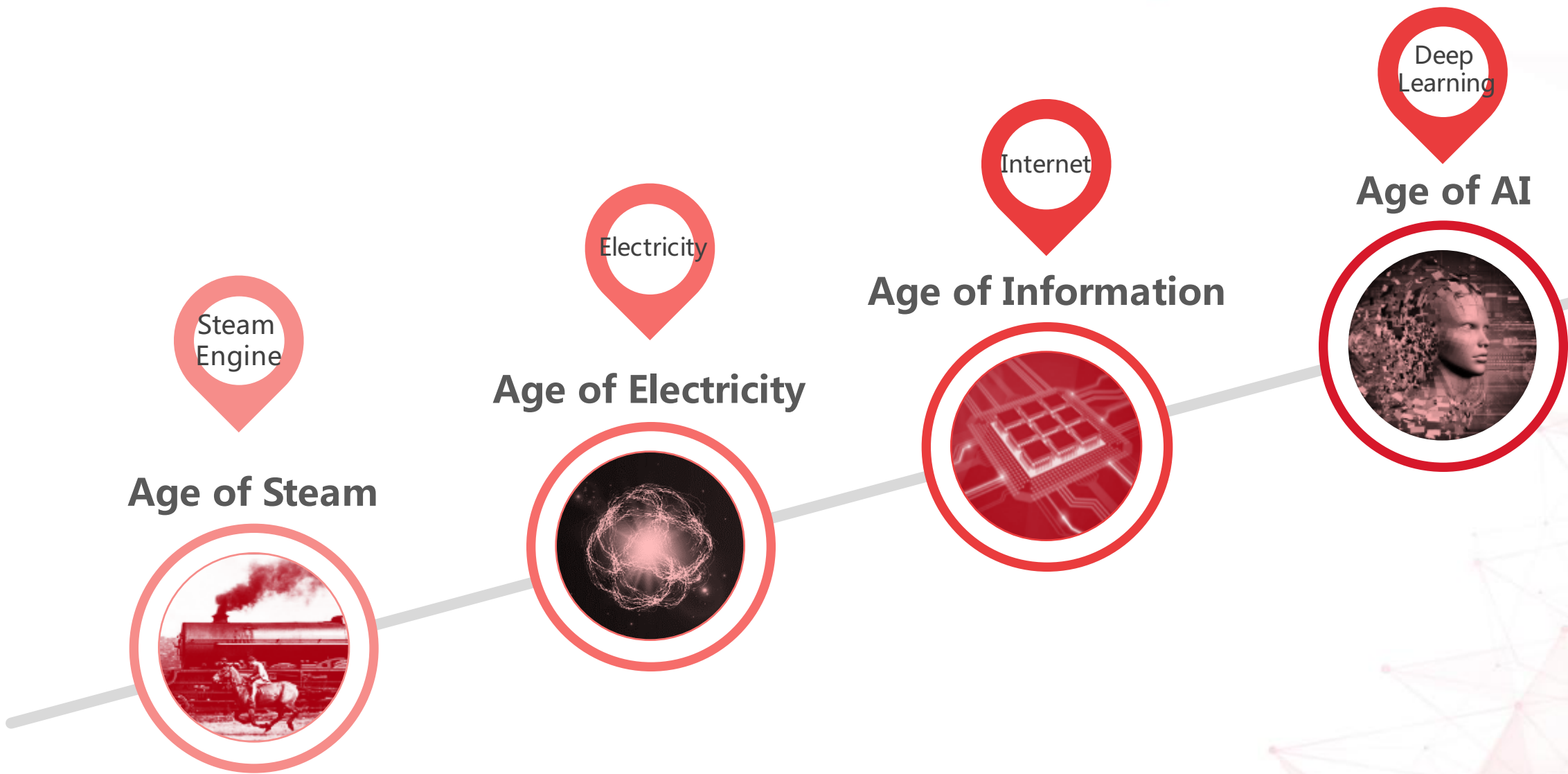


Supercomputing Power

Booster for Artificial Intelligence

Dahua Lin

Artificial Intelligence: A Technical Revolution in Sight



Deep Learning Enables AI Breakthroughs

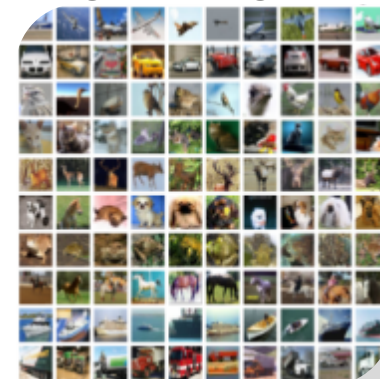
Voice Recognition



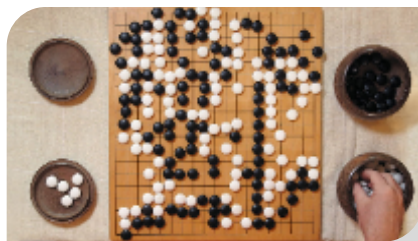
Face Recognition



Image Recognition



Game Playing



Deep Learning



Autonomous Driving



Intelligent Financial Services



Natural Language Processing

World Economy: Artificial Intelligence Is The New Lever

Till **2030**

AI will

Increase global GDP by

14%

Contribute to the
world economy by

\$15.7 trillion

Data source : PwC

Company Profile

20 years
Research
Experience

2300+
Staff

150+
AI PhDs

Proprietary
Deep Learning
Platform

Performance

Valuation
Highest
in the World

Core Tech
World Class

Revenue
Industry No.1

700+
Major Clients
and Partners

Leading Technologies



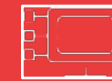
AI+Fintech



AI+Smart City



AI+Mobile



AI+Chips



AI+
Autonomous Car



AI+Health

Partnership with 700+ Major Corporations

Revenue & Market share #1

AI + Fintech



Revenue & Market share #1

AI + Mobile Internet



Strategic Partners



AI + Smart City



AI + IOT



AI+ Remote Sensing



AI + Smart Phones



AI + Car



AI + Retail



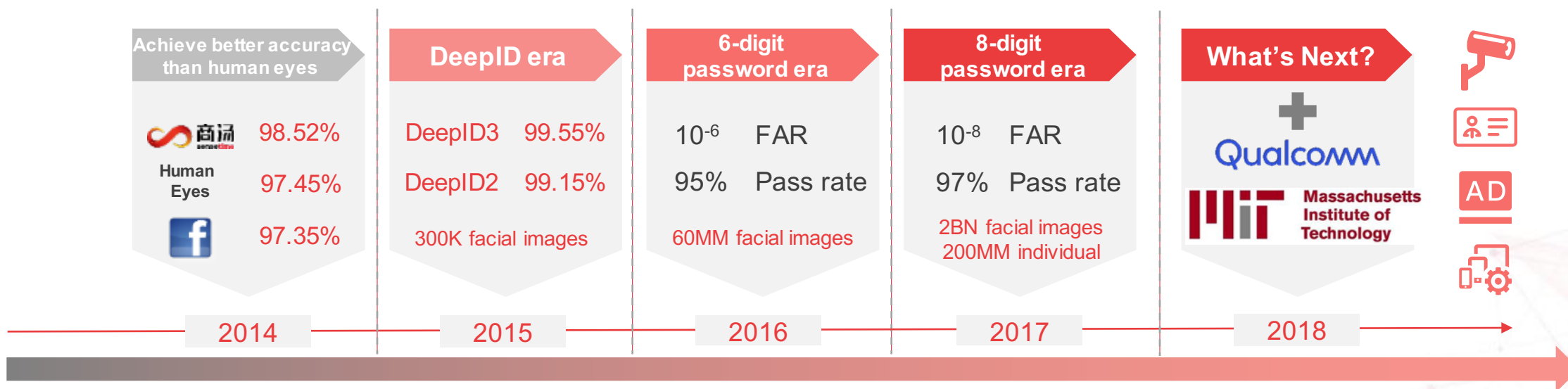
Revenue & Market share #1



Revenue & Market share #1

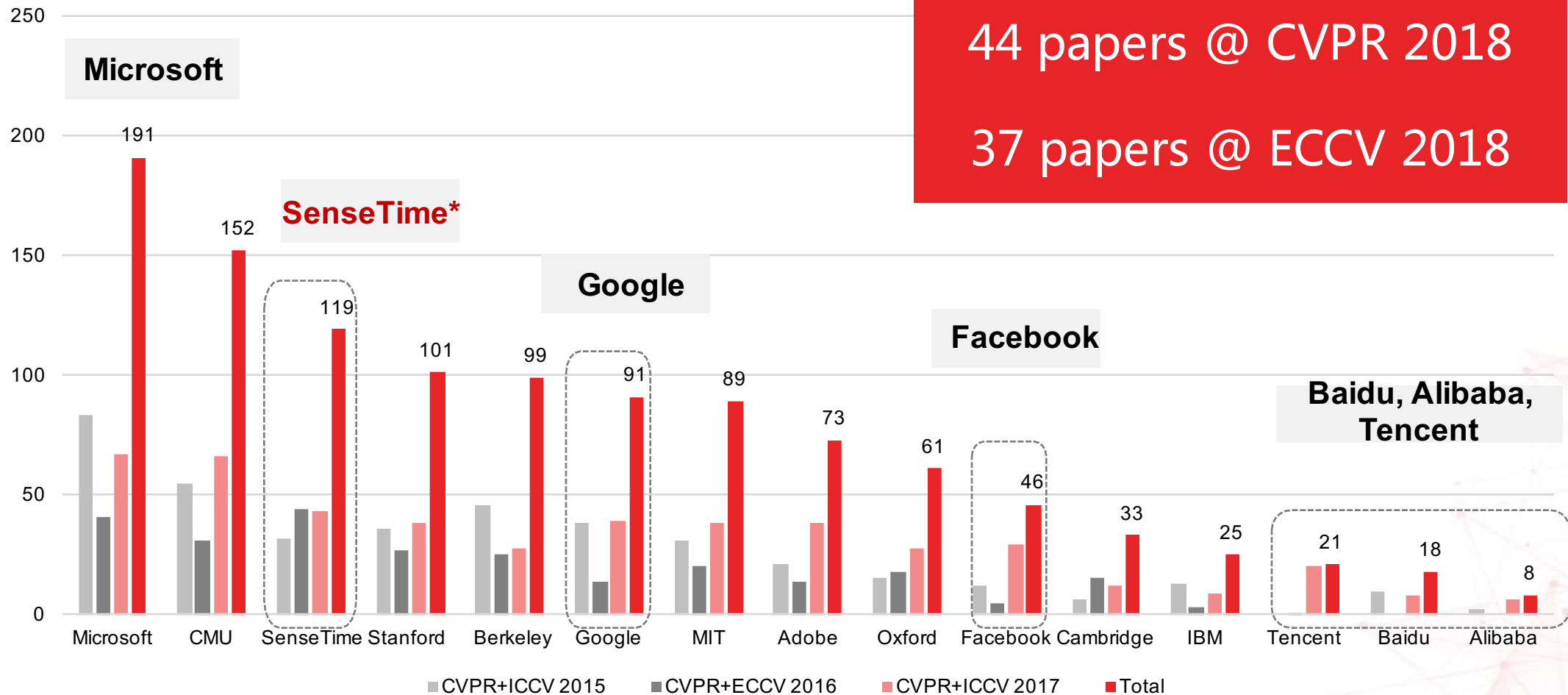


The Breakthroughs in Face Recognition



Academic Publication in Comparison with Tech Giants

Dominance in No. of top papers published leads to rich transform into intellectual properties



44 papers @ CVPR 2018
37 papers @ ECCV 2018

*SenseTime co-R&D with CUHK

CVPR, ICCV, ECCV are the top 3 computer vision conferences worldwide with highest impact factor. They accept the best work on AI and deep learning.

Top Performances on International Competitions

2015-2017

- Detection and Tracking : KITTI 2015 Winner
- Depth Estimation : KITTI Stereo 2015 Benchmark No.1
- Action Recognition : ActivityNet Challenge 2016 No.1
- Scene segmentation : Cityscapes Challenge 2016 Winner
- Multiple Object Tracking : MOT Challenge 2016 Winner
- Lane Detection : TuSimple in CVPR 2017 workshop Winner
- Segmentation : Pascal VOC Dataset No.1
- Video Object Segmentation : DAVIS Challenge 2017 No.1

2018

- COCO2018 : Object Detection Champion
- VOT Challenge : Champion
- Face Identification & Verification : MegaFace2018 Two Winners

SenseParrots



High performance

High computational speed ,
low consumption of memory



High Scalability

Linear speed up for training on hundreds of
GPU cards



High flexibility

Highly modular design, can effectively adopt novel
training tasks



High productivity

Rapid prototyping, easy deployment

SensePetrel



Multiple Storage Types

Object Storage, File Storage, Middle-ware



High System Stability

High Availability Architecture + QoS, High Stability



Linear Expansion

Expand capacity and performance as needed



Optimization for AI

A large number of optimization such as RDMA and SSD

World's Top Ten AI Lab | CUHK – SenseTime Joint Lab





Trevor Darrel




Tomaso Poggio










Russ Salakhutdinov




Yoshua Bengio








Xiao'ou Sean Tang
Founder of SenseTime






Nandode Freitas












Yann LeCun
Leader of Facebook AI Lab




Geoff Hinton
Leader of Google deep learning research







Fei-Fei Li
Director of Stanford AI Lab, founder of ImageNet Competition





Jurgen Schmidhuber









SenseTime Becomes the 5th National Platform for Next-gen AI

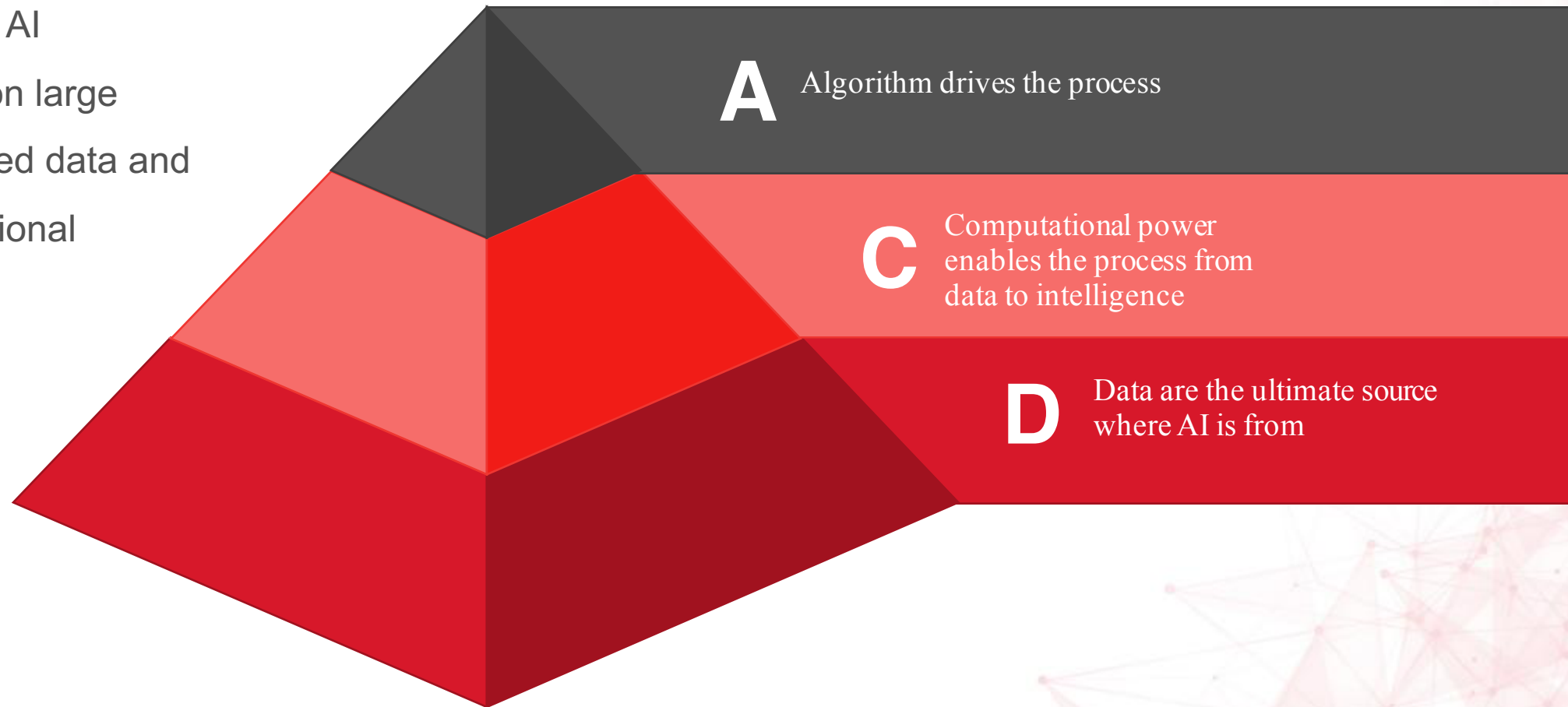


National Open Innovation Platform for Next Generation Artificial Intelligence on Intelligent Vision

On Sep. 20, 2018, the Ministry of Science and Technology of China declared SenseTime as the fifth National AI Platform.

AI is not a magic

The success of an AI technology relies on large amount of annotated data and plenty of computational resources.



Tremendous computing power will be needed in the future

10x/yr.

Since 2012

The computation resources required by a single model training increases by over 10.85 per year.

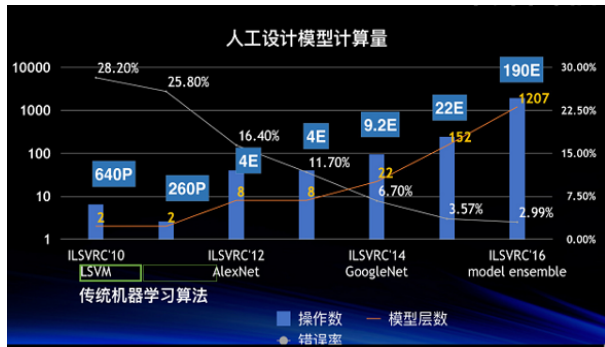
300,000

Since 2012

This metric has grown by more than 300,000x

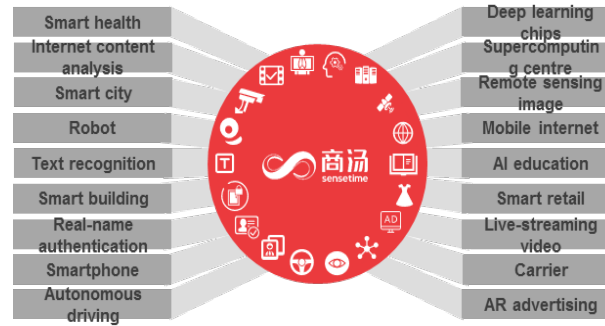
➤ Computing Power

Industry-grade applications require reliable infrastructures that can deliver the needed computing power.



➤ Diverse Demand

Well-designed components and middle-wares are needed, in a timely and scalable way.



➤ Reduced Cost

Computing infrastructures requires highly professional skills and is costly.



➤ Top Performances on International Competitions





 Union Pay  Banks & Financial Institutions  OPPO, Xiaomi  Public Security & Surveillance companies  China Mobile  Qualcomm, NVIDIA ...

Application






 Face Recognition  Image Recognition  Autonomous Driving  Human-machine interaction  Medical Image  AI Chip ...

A variety of core AI technologies



 SensePetrel  PPL  SenseParrots

Platform Layer

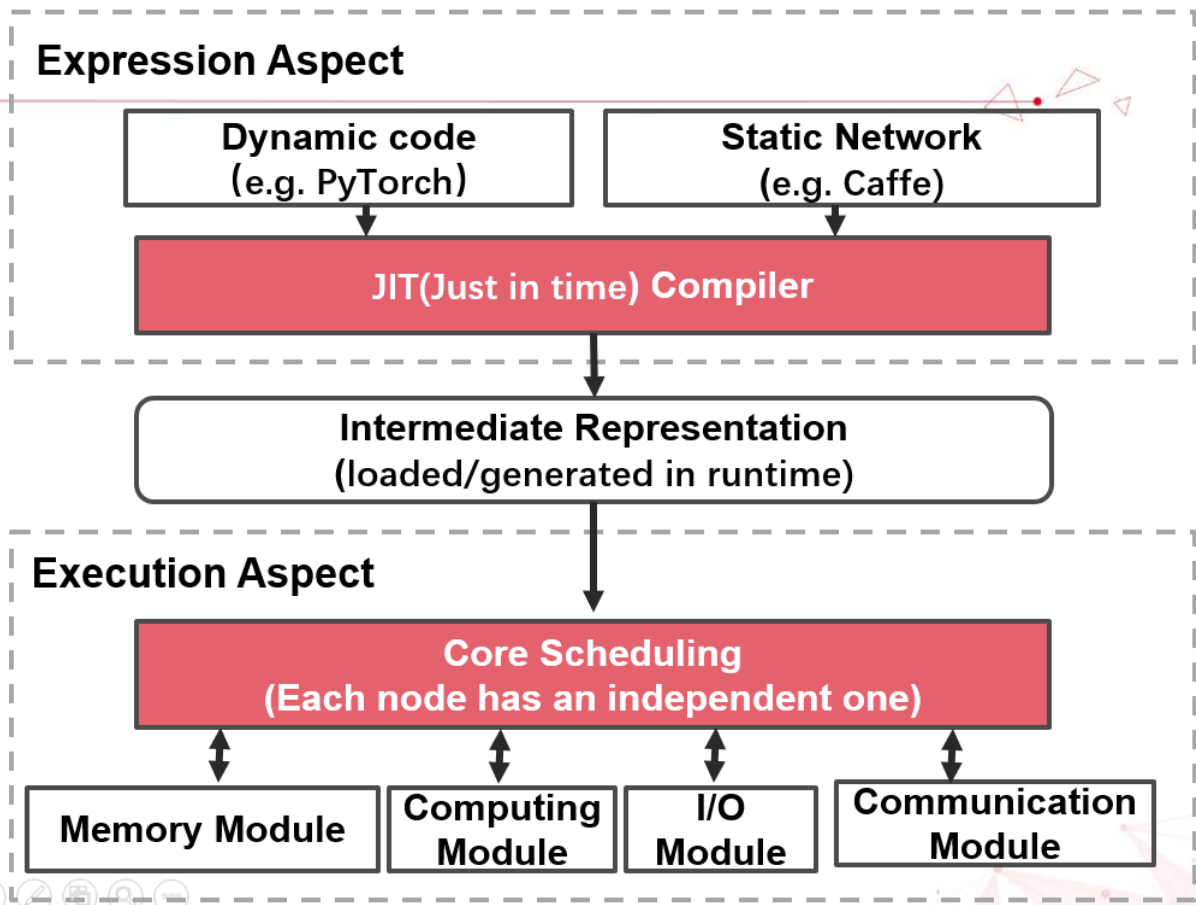
 High Speed Comm System  Virtualization  High performance Computing System

Infrastructure Layer

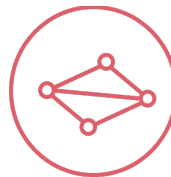


Supercomputing Platform in SenseTime

SenseParrots: A Training System Designed for the Future



System Architecture



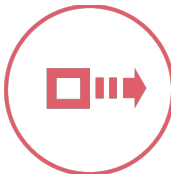
Dynamic Graph Parallelism

Embrace the future with high flexibility, supporting arbitrary models, defined in advance or created on the fly



Scalability over Thousand GPUs

Scale to 1000+ GPUs, with nearly linear speed-up and just in one click



Non-blocking Execution Engine

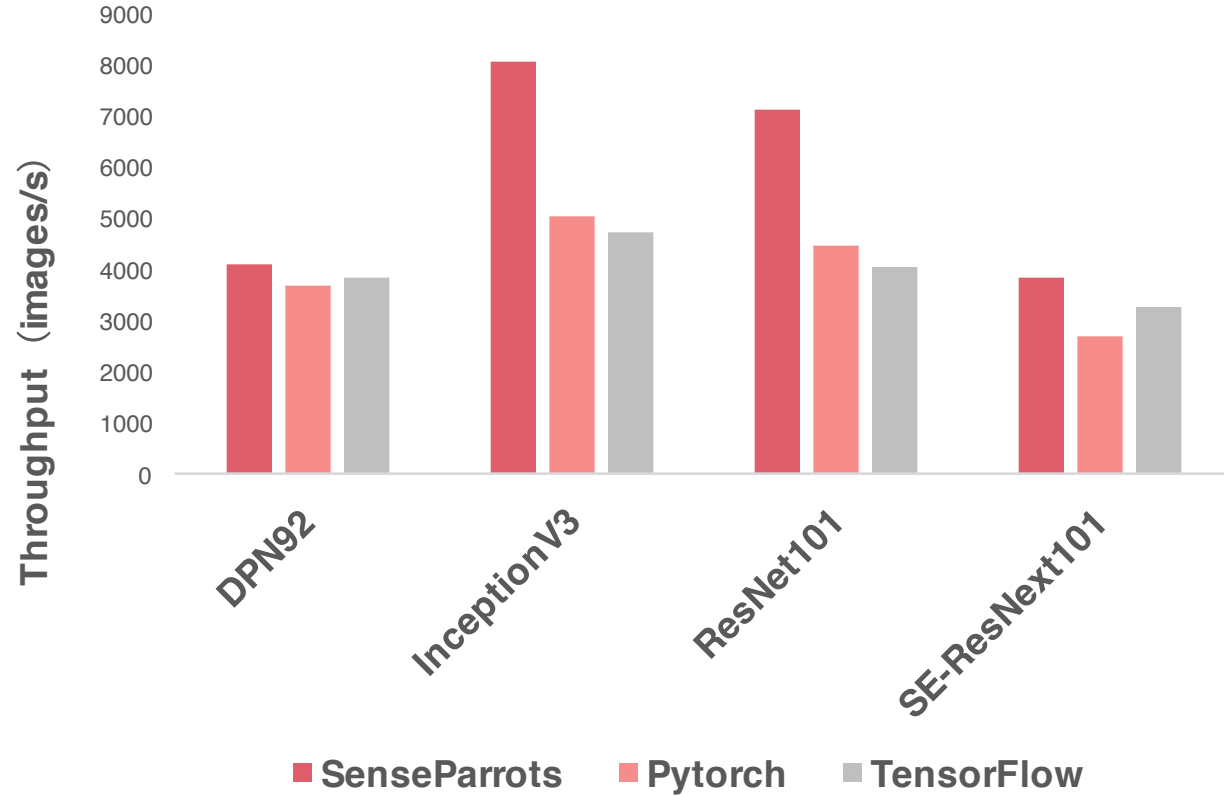
No longer limited by the interpreter — computation, communication, and IO, all running in parallel yet reliably



JIT Compilation

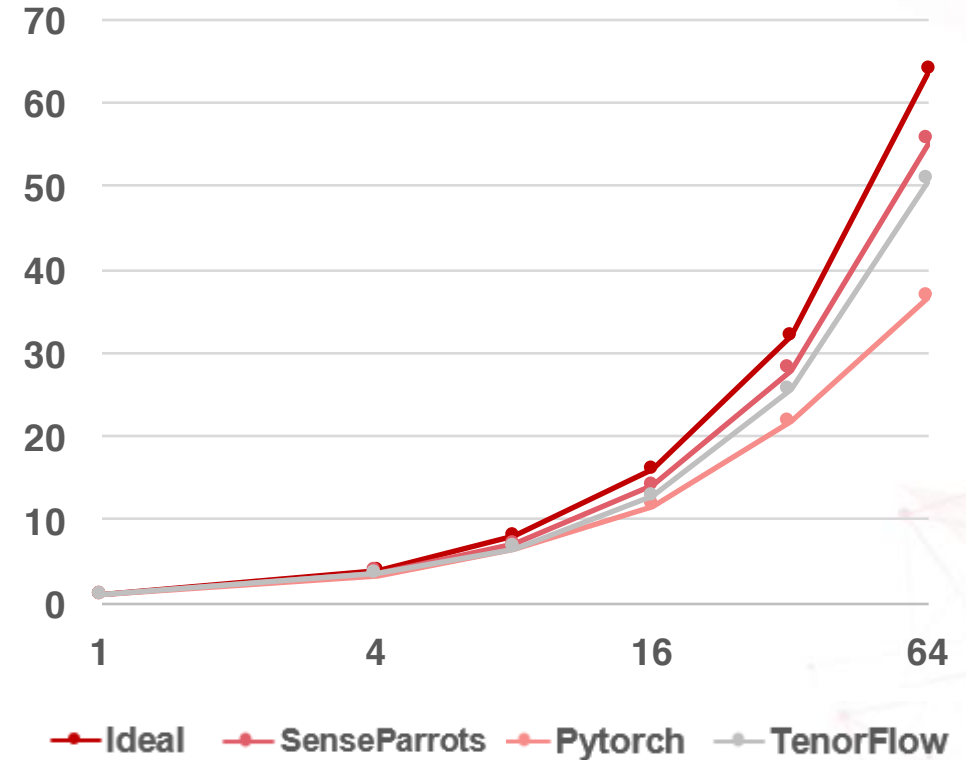
Compile, schedule, and execute codes just in time, exploiting the computing resources to the maximum

SenseParrots: Performance in Comparison



Training throughput on 64GPUs (images/s)

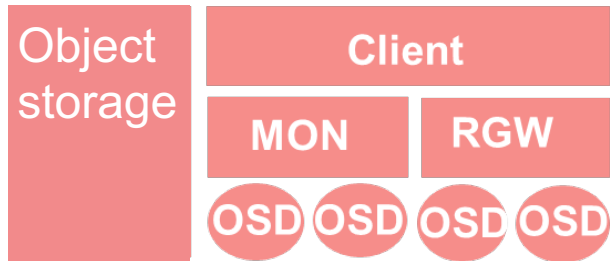
CPU: Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.40GHz
GPU: NVIDIA 1080ti



Speedup on InceptionResNetV2 (Batchsize = 32)

CPU: Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz
GPU: NVIDIA 1080ti

Technical proposal



Unified Interface

Unifies Storage interfaces with different types of storage system



Expansion on Demand

Expands capacity and performance whenever needed



Enhanced Availability and Stability

Uses High Availability Architecture and QoS



Optimized IO Performance

Aggregates small files, increases batch read-write interface, optimizes client pre-read cache,

1.3 million QPS

Usually small file system are 100,000 QPS

100 billion

Horizontal expansion of architecture to store small files

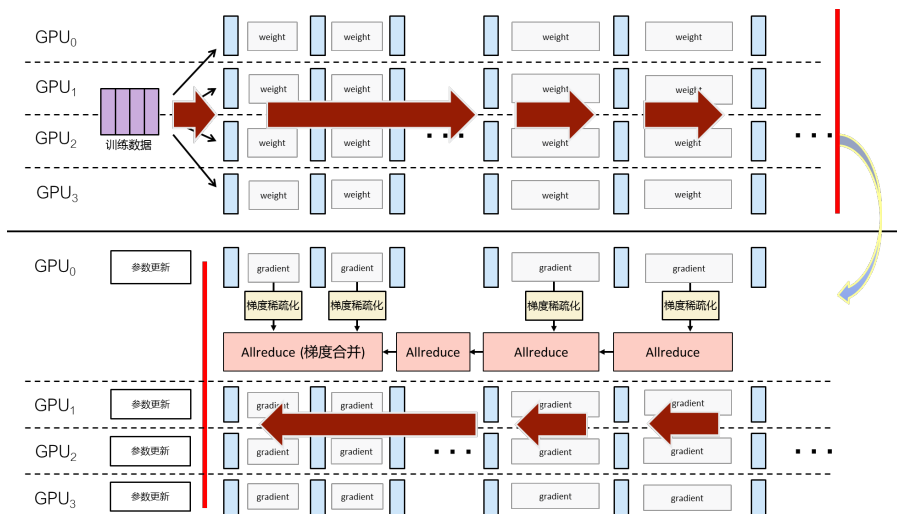
6 times ↑

Writing Speed of S3 interface of 10 billion 10K files

36% ↑

Throughput compare with memory caching systems

Training Deep Networks in Minutes



1.5 min Training **AlexNet**
7.5 min Training **ResNet50**

ImageNet Dataset, 512 V100 GPUs, 56 Gbps Network

128 | **90%+** | **512** | **86%**
GPUs Efficiency GPUs Efficiency

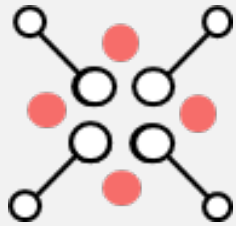
Lazy Allreduce

Coarse-Grained sparsity

LARS & Warm up

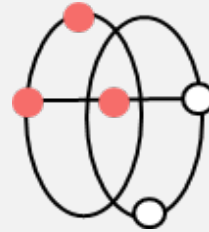
State-of-the-Art Approach:

the training speed of ImageNet reaches **1 epoch/s** on GPU cluster.



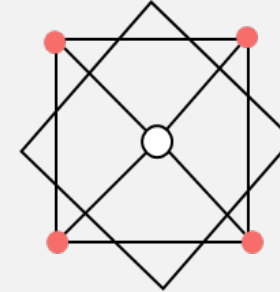
Architecture Search

- Finding a good new architecture takes thousands of trials. Minute-level training makes automatic search a reasonable option



Rapid Iteration

- Revolutionize the workflow of model design, allowing a new model to be tested in minutes and reducing the design period from months to hours



Automated Adaptation

- You no longer need a team of experts in response to a new customer. Everything can be done in one click, with the support of the supercomputing

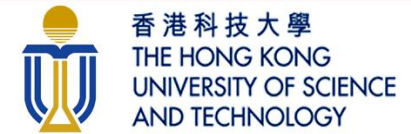
GPU-based Supercomputing Platform



- ✓ 17 clusters
- ✓ 14,000+ GPUs
- ✓ Computing power :160PFlops+
- ✓ Beijing, Shanghai, Shenzhen, Hong Kong, Tokyo, Singapore



Universities



Enterprises



Lead AI Innovation

Power the Future