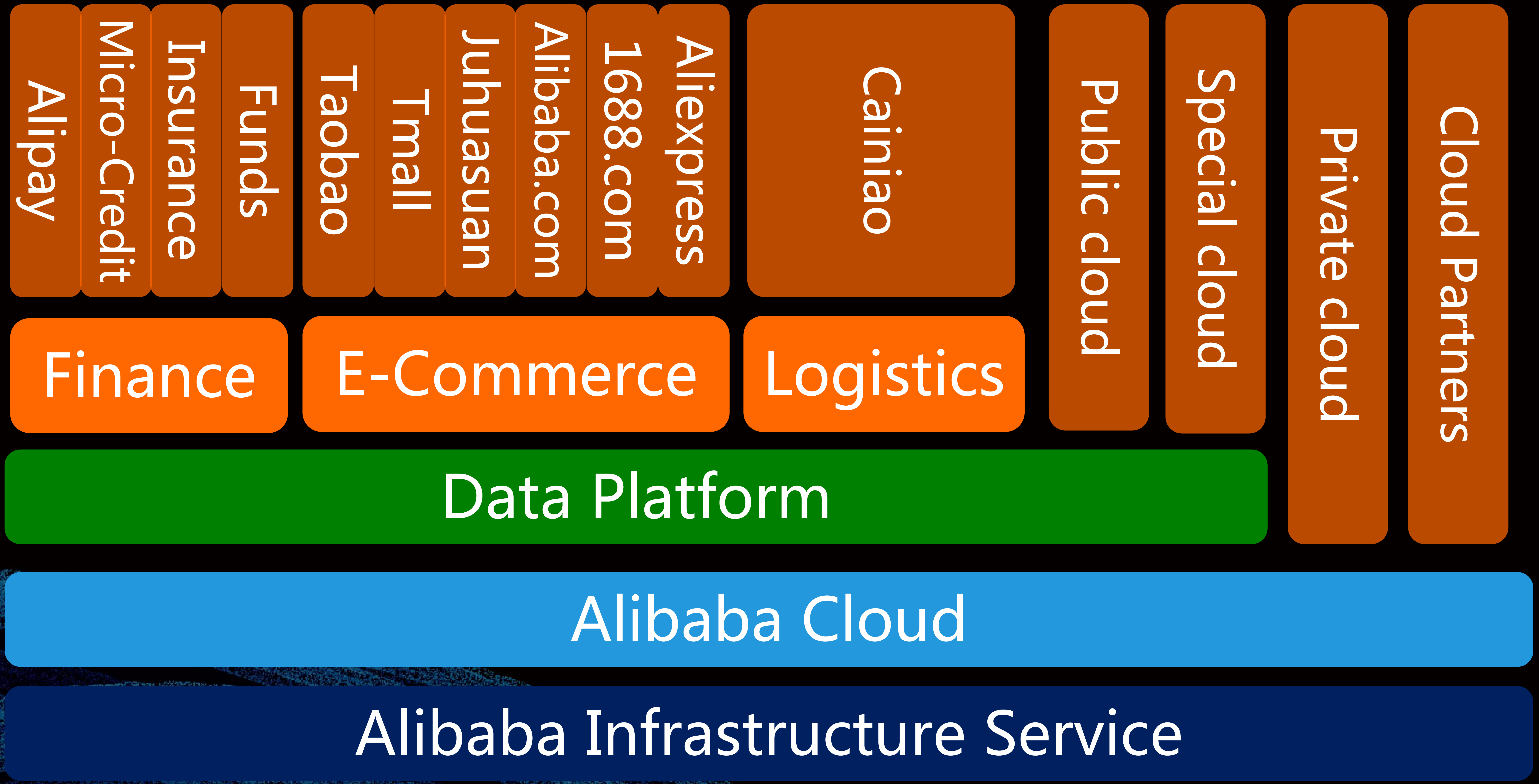
Applied AI Architecture @ Alibaba Infrastructure

Lingjie XU, Director
Heterogeneous Computing
AIS

Who I am

- Joined Alibaba Spring 2017
- Leading Applied AI Architecture team, focusing on AI HW acceleration and HW/SW co-play
- Held multiple senior architect and management roles in GPU domain

Who we are



Alipay

Micro-Credit

Insurance

Funds

Taobao

Tmall

Juhuasuan

Alibaba.com

1688.com

Aliexpress

Cainiao

Public cloud

Special cloud

Private cloud

Cloud Partners

Finance

E-Commerce

Logistics

Data Platform

Alibaba Cloud

Alibaba Infrastructure Service

Alibaba Infrastructure

IDC

Modular
Eco-Friendly
Automation

Server

High Perf
Low Power
Scalability

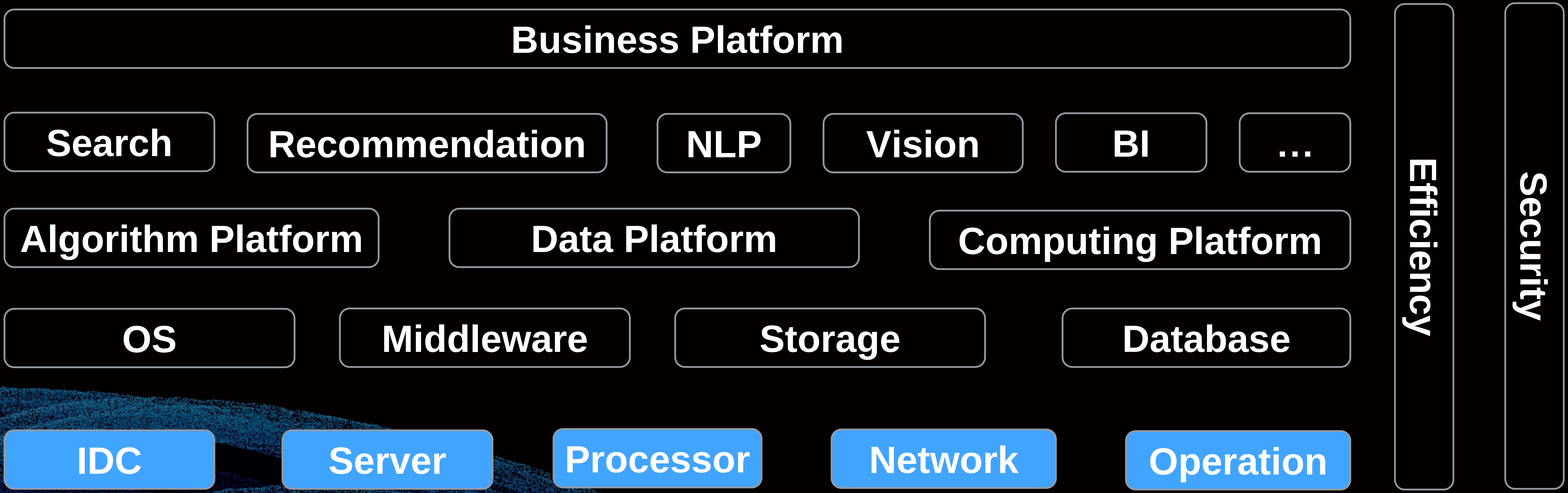
Network

100G
SDN
Security

GOC

Monitor
Analyze
Act

Technology Overview



Business Platform

Search | Recommendation | NLP | Vision | BI | …

Algorithm Platform | Data Platform | Computing Platform

OS | Middleware | Storage | Database

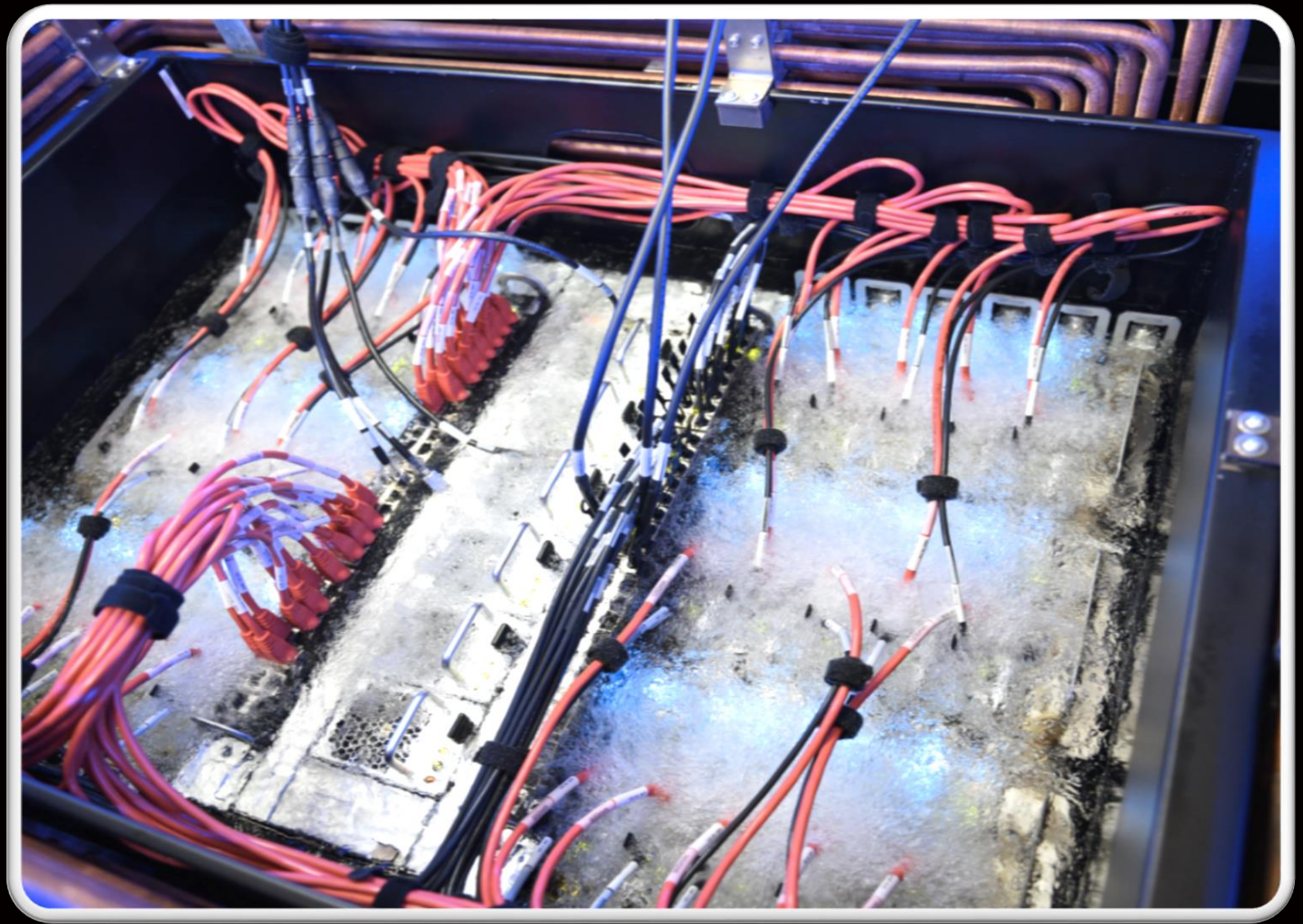
IDC | Server | Processor | Network | Operation

Efficiency

Security

Datacenters

Zhangbei Datacenter
(Fresh air cooling system)
Best PUE <1.2

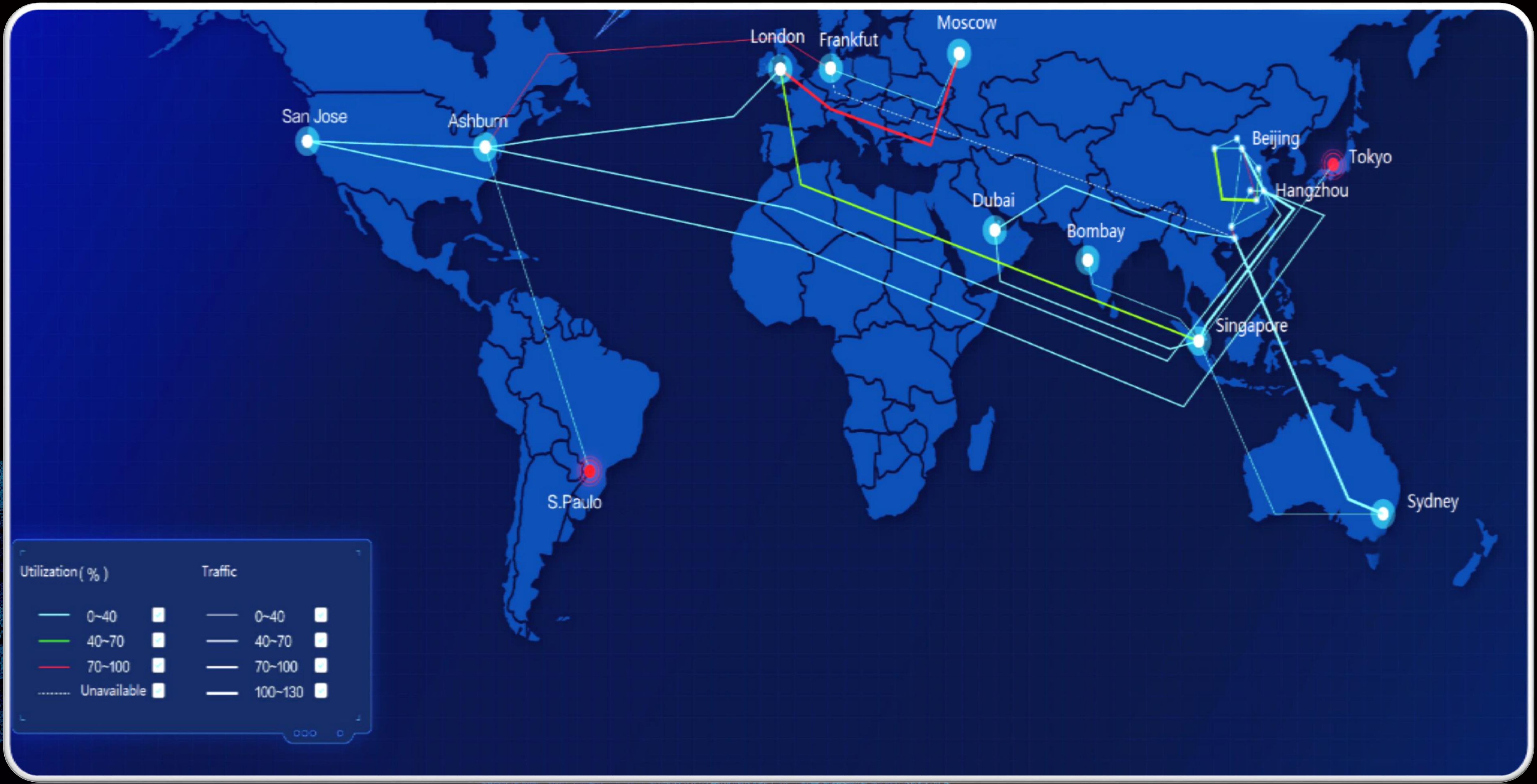


New Frontier
Server immersion cooling
PUE ~1.0

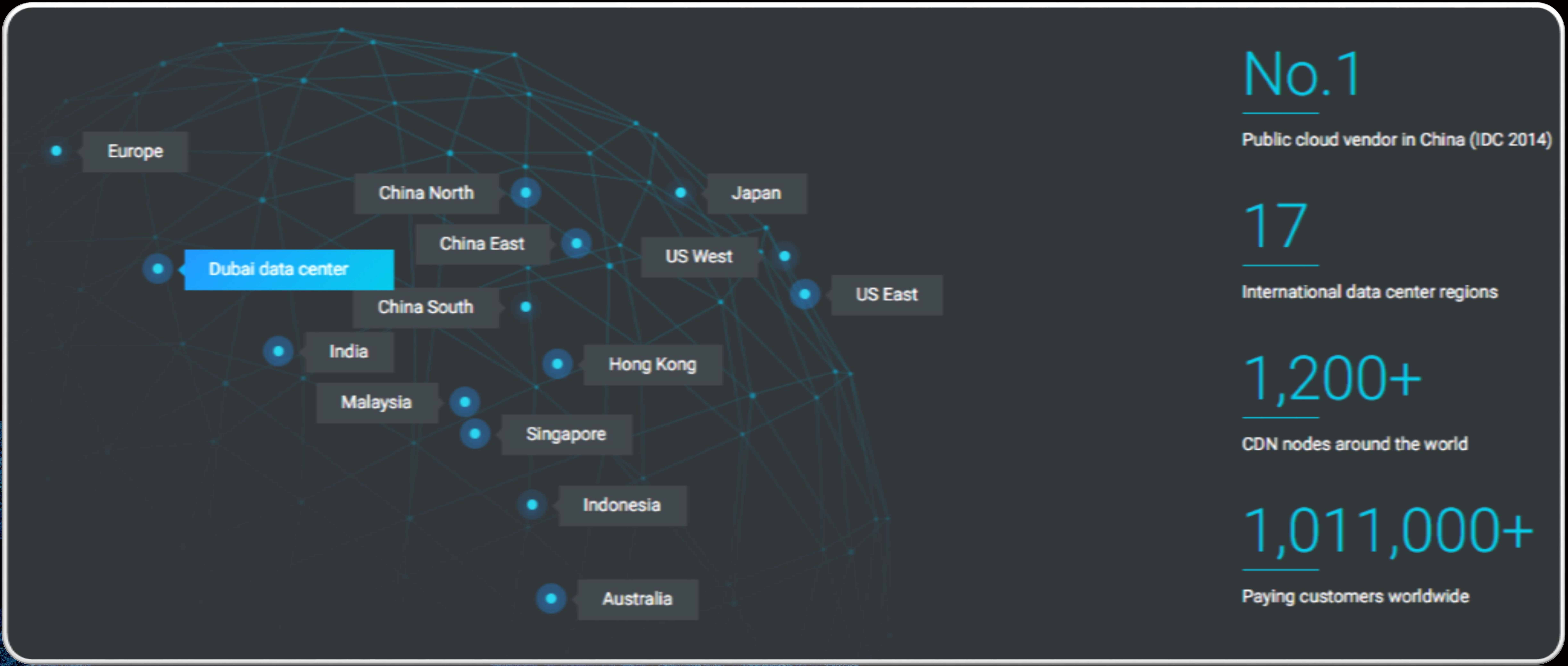
Qiandaohu Datacenter
(Lake water cooling system)
PUE < 1.3

Network

Massive Scale + Diverse Applications + Bursty Traffic + Fast Growth



Global Infrastructure



No.1

Public cloud vendor in China (IDC 2014)

17

International data center regions

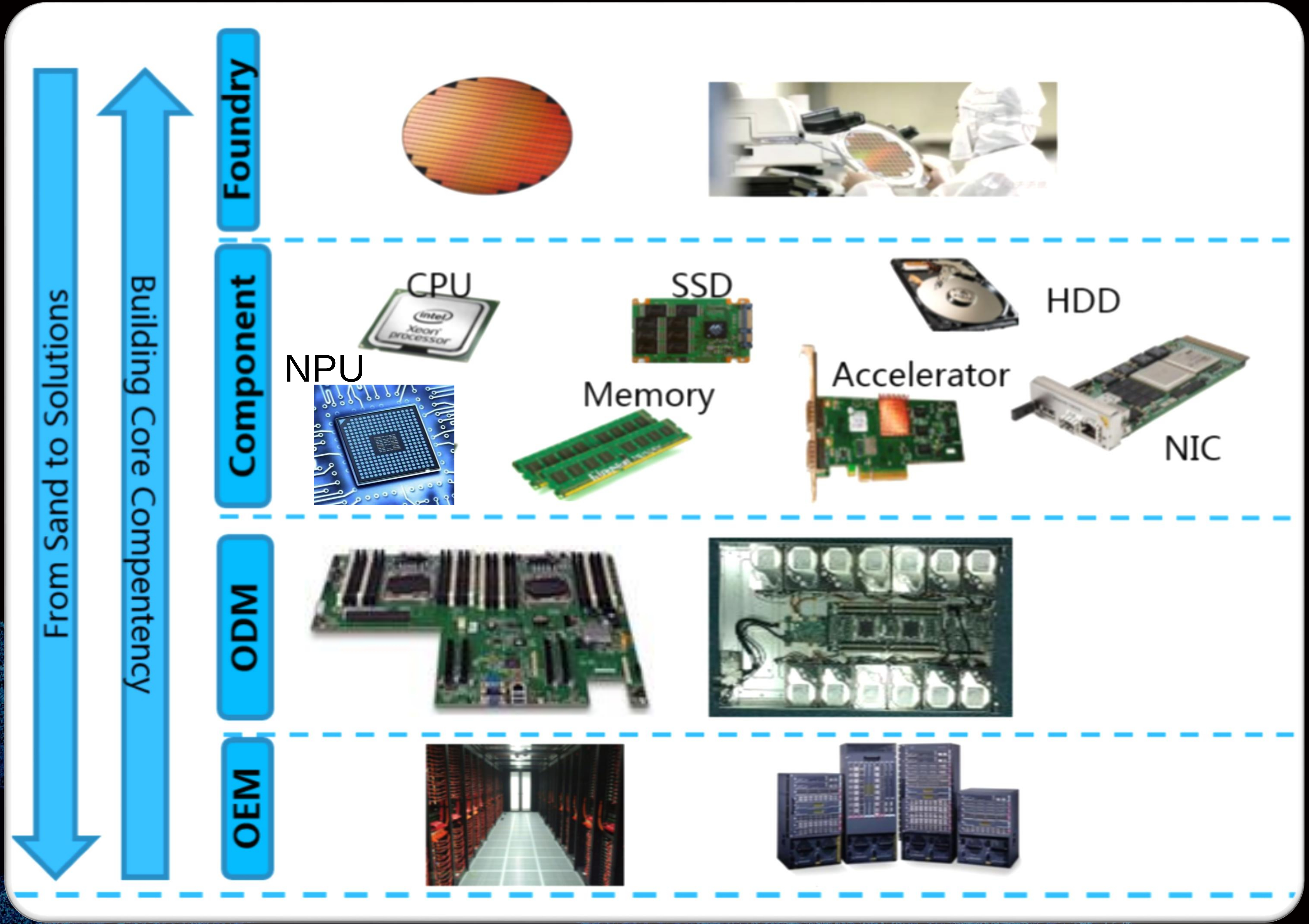
1,200+

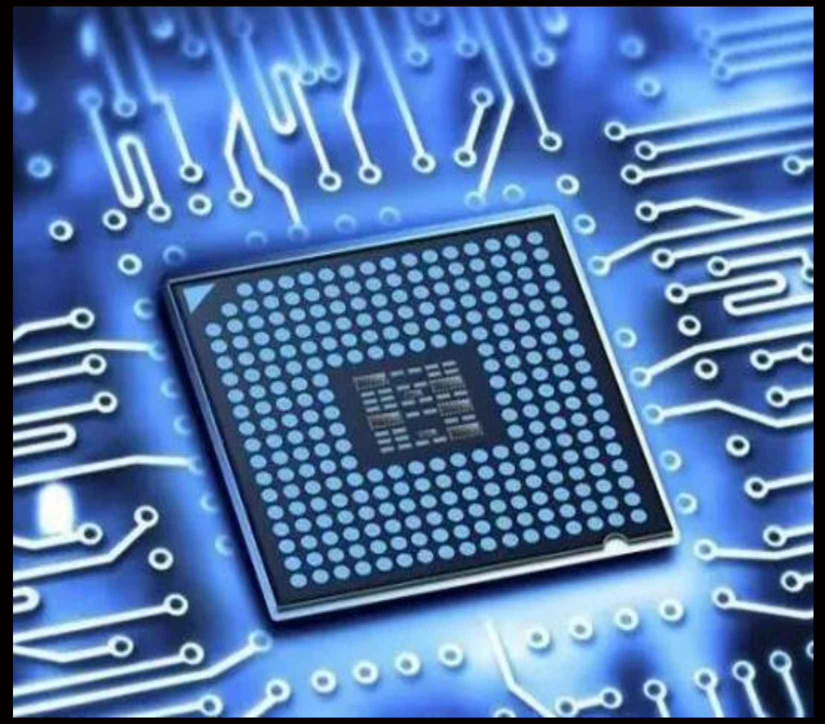
CDN nodes around the world

1,011,000+

Paying customers worldwide

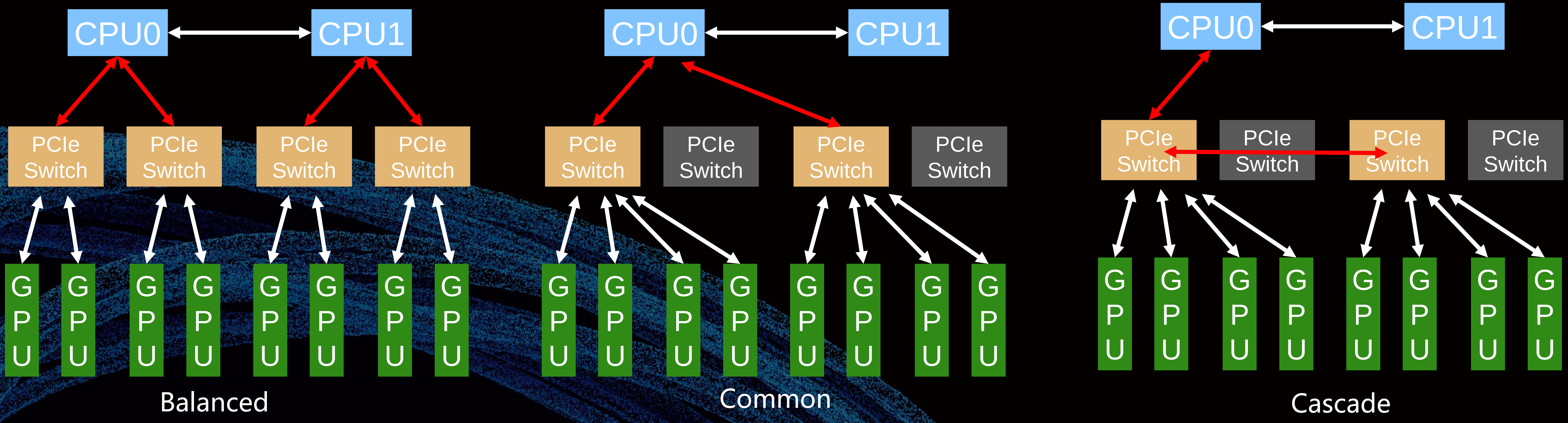
Compute & Storage





GN6: 8-way GPU Server

- SXM2 or PCIe
- Decoupled modular design
- Configurable topology



The Wave of AI Revolution

The Economist

Crunch time in France

Ten years on: banking after the crisis

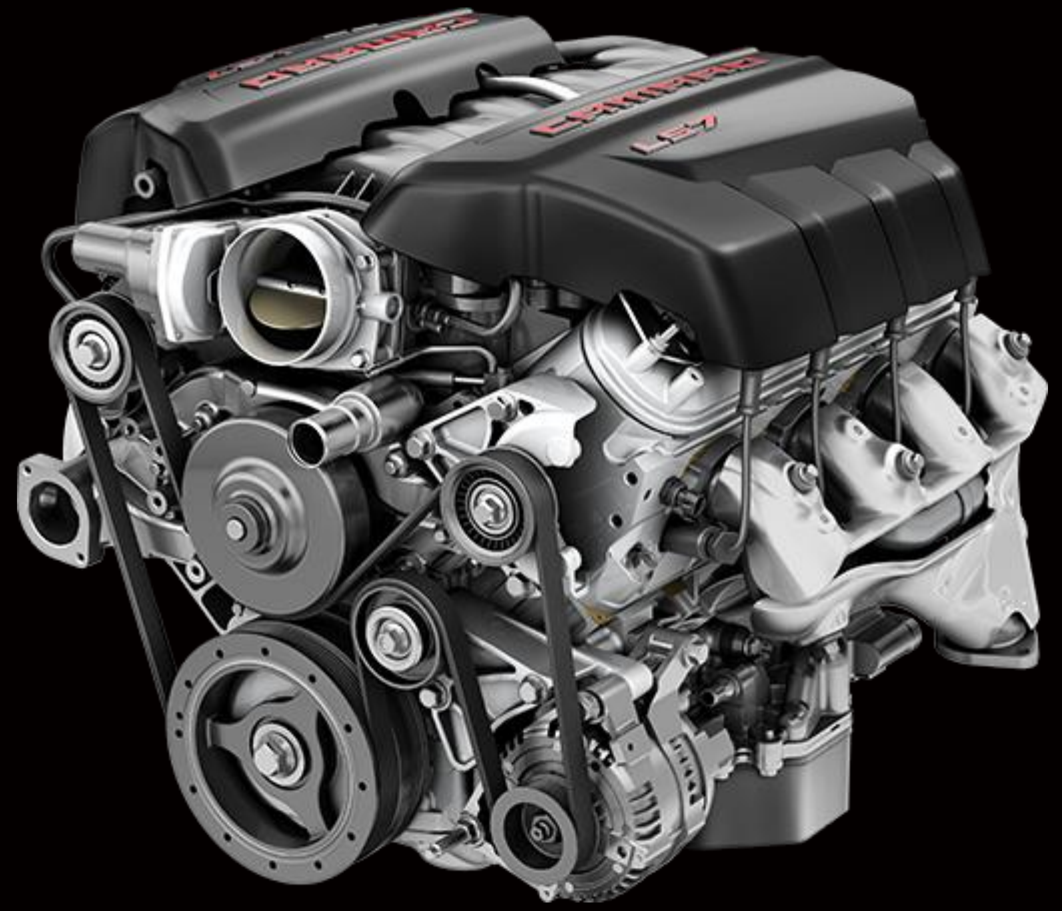
South Korea's unfinished revolution

Biology, but without the cells

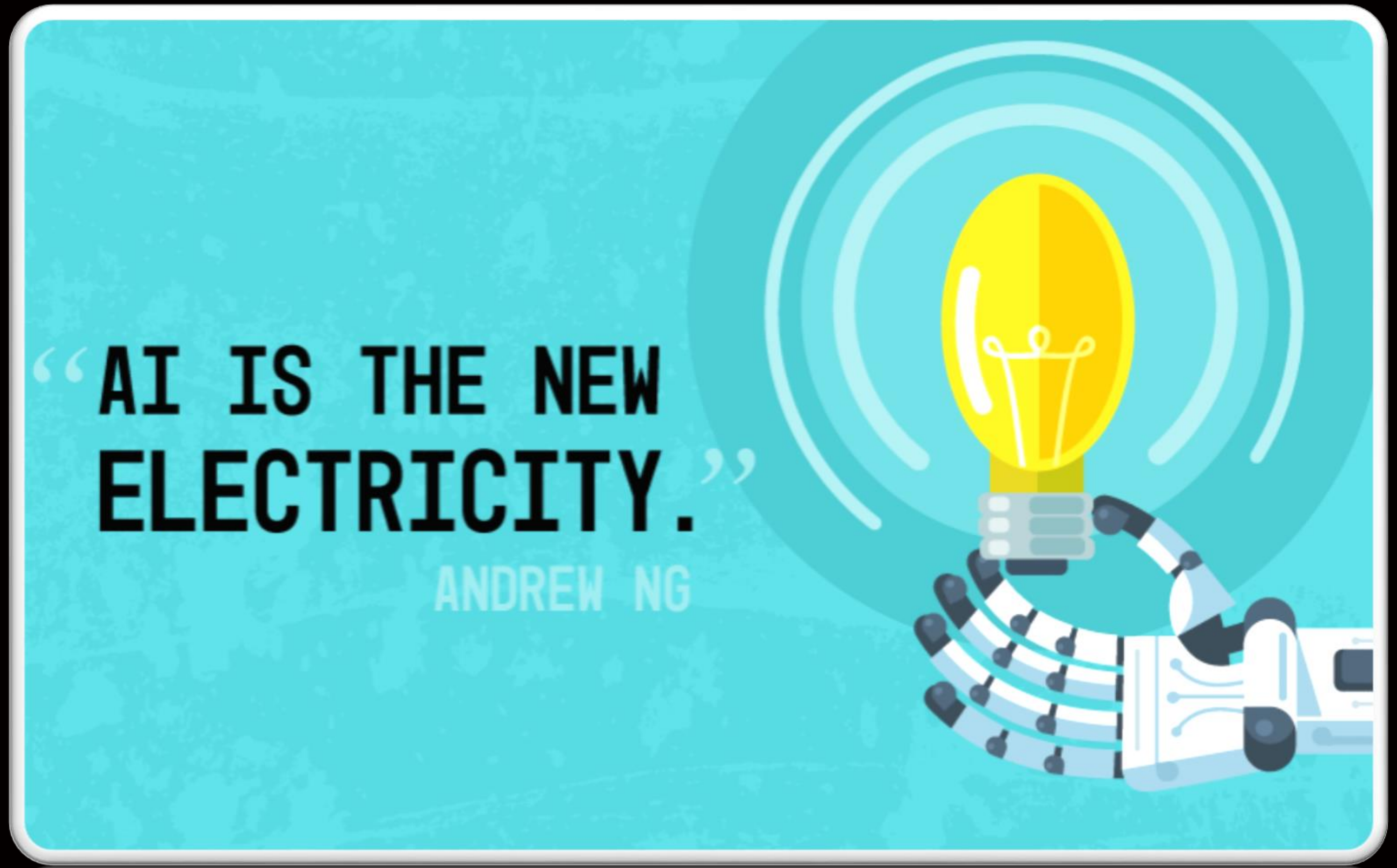
The world's most valuable resource

Data and the new rules of competition

Data

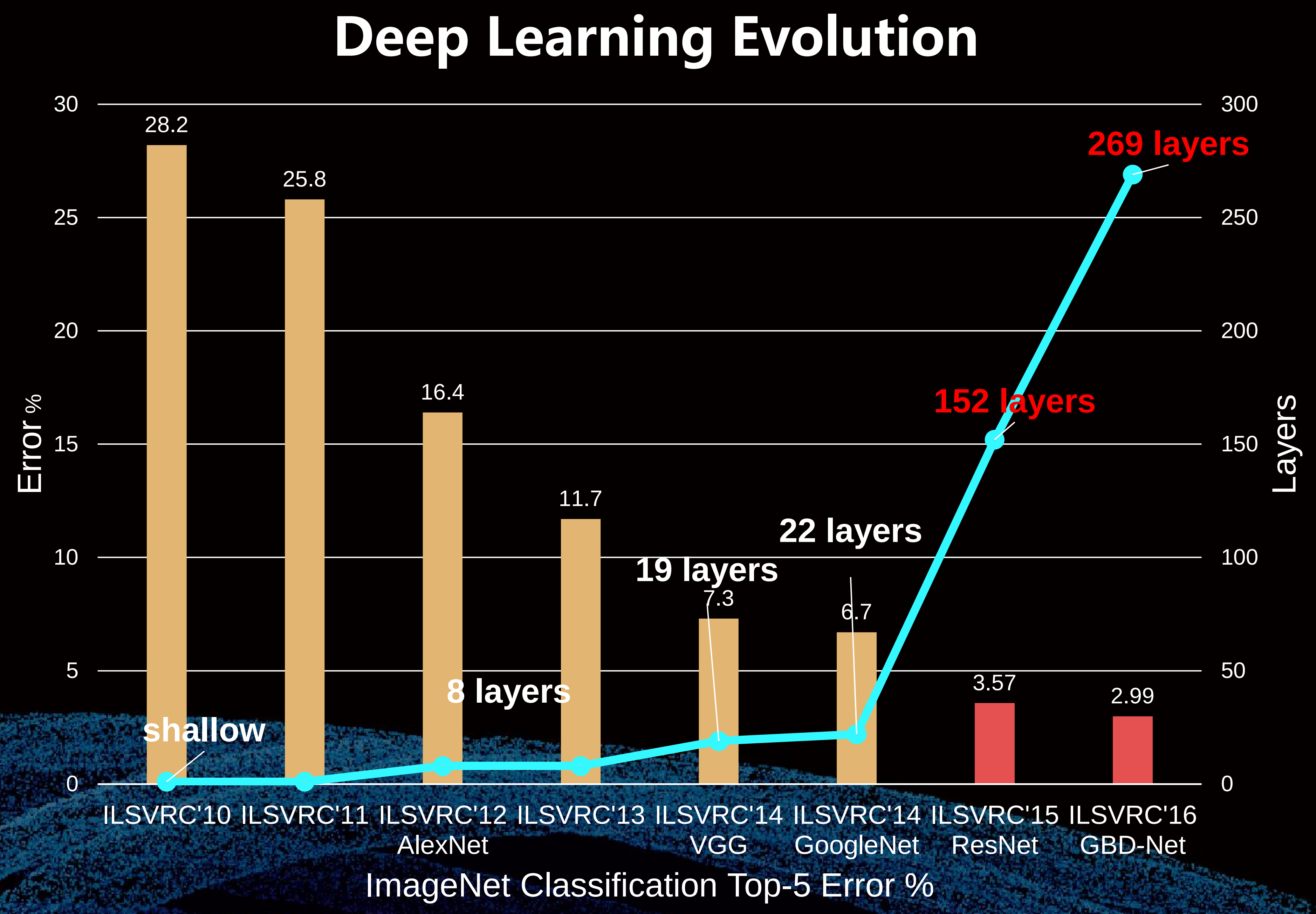


Computing Power

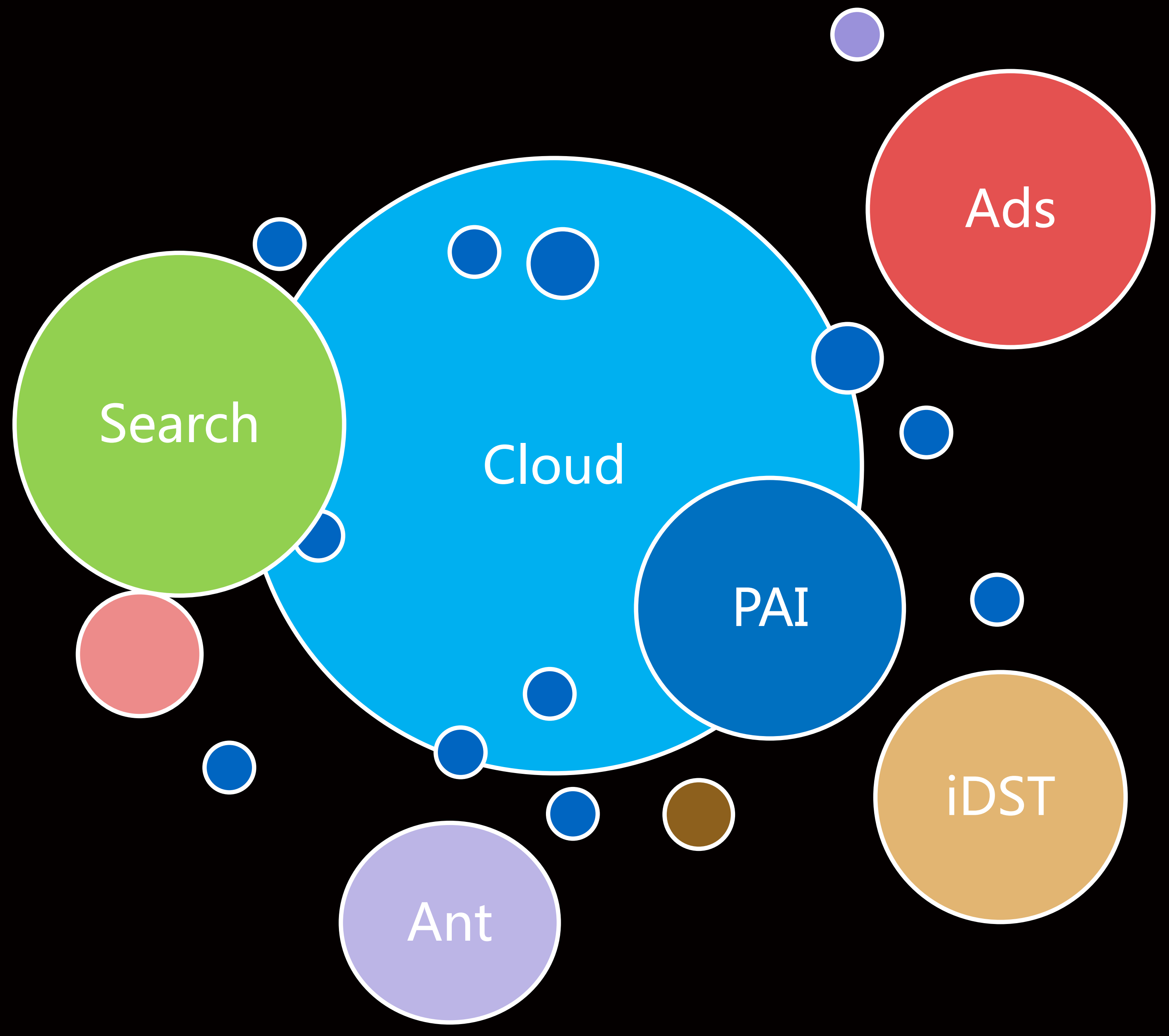


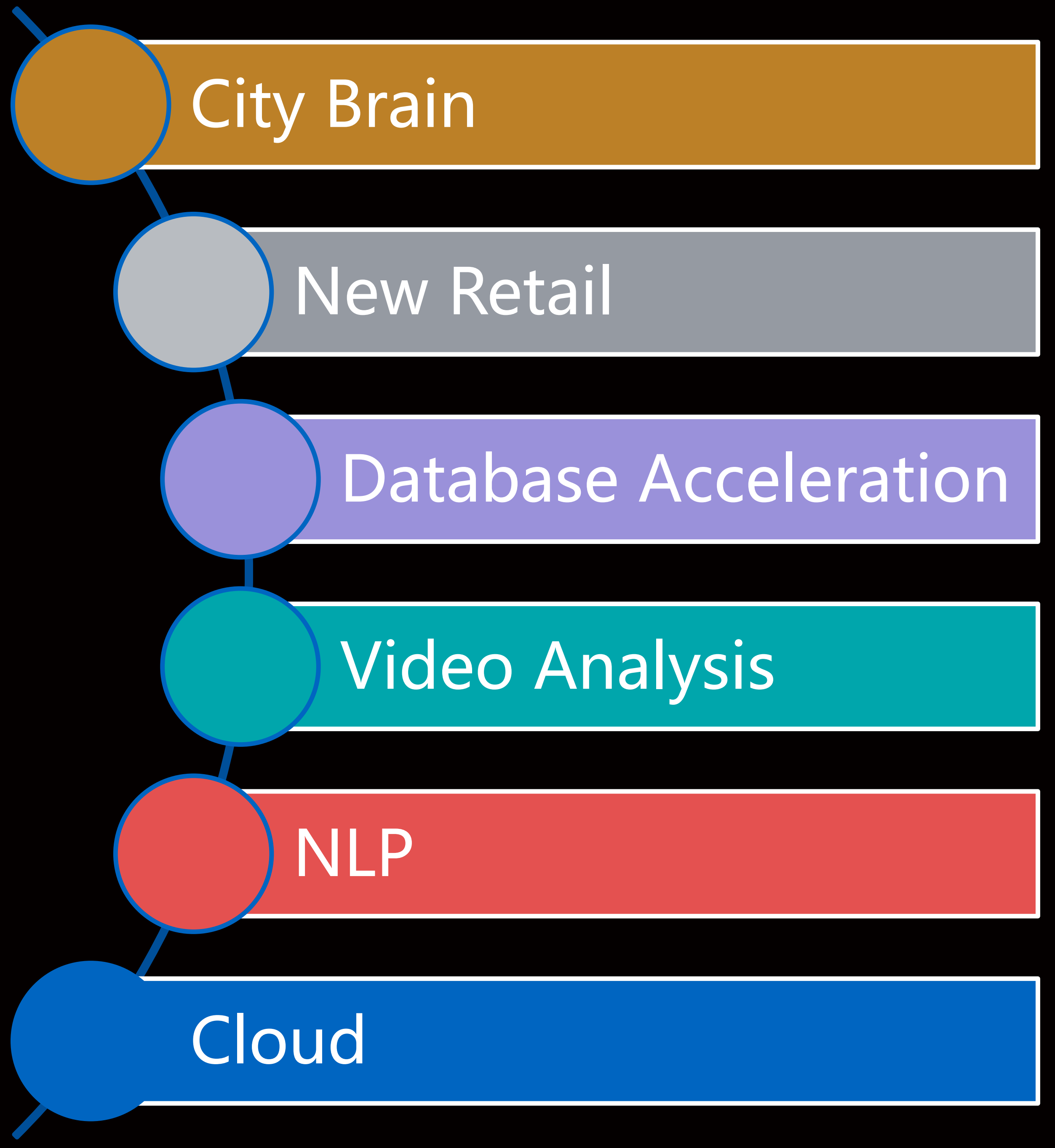
Algorithm

Deep Learning @ Alibaba



Deep Learning Evolution





City Brain

New Retail

Database Acceleration

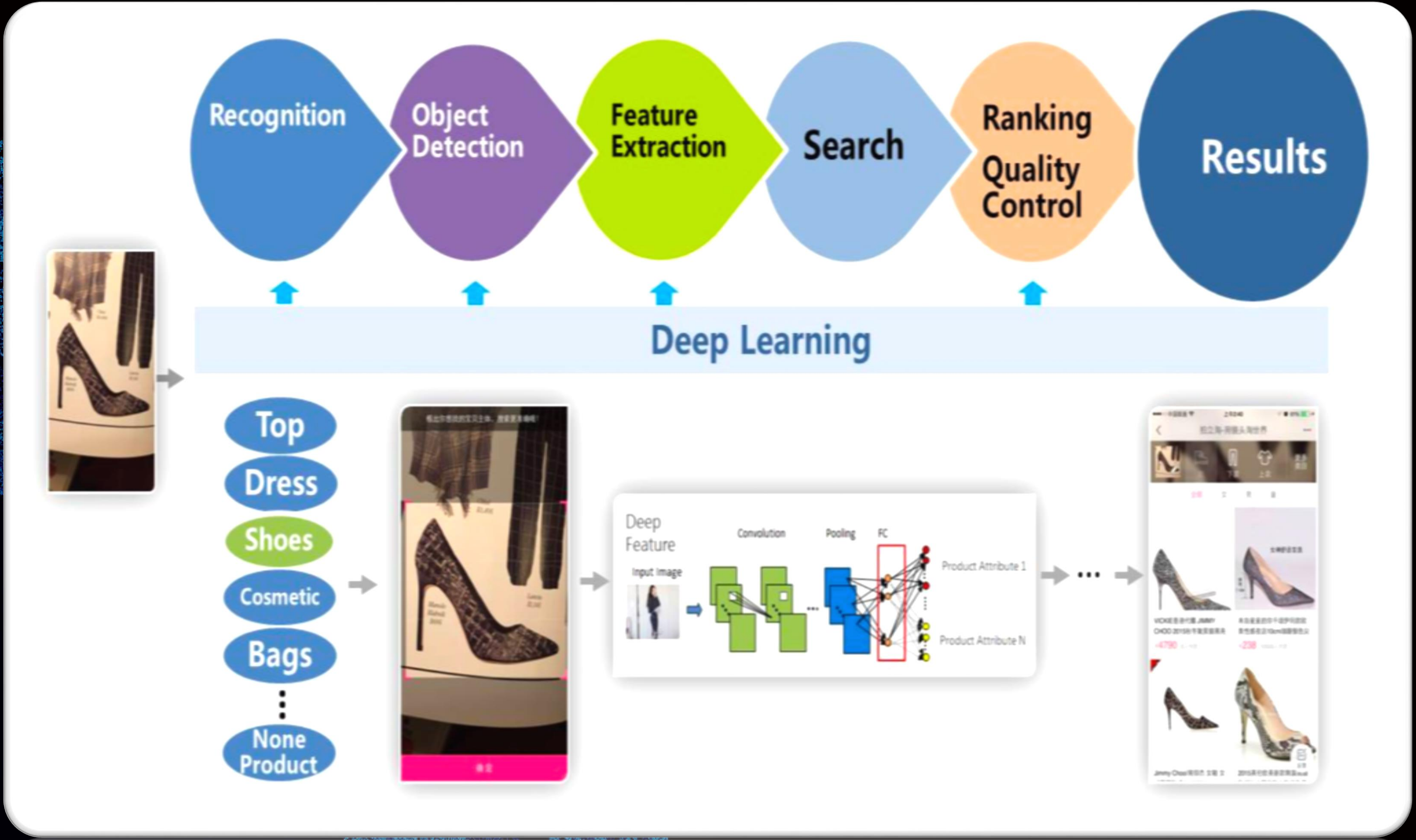
Video Analysis

NLP

Cloud

PaiLiTao

- Category Prediction
- Object Detection
- Feature Extraction
- Index Searching
- Soring & Output



OCR

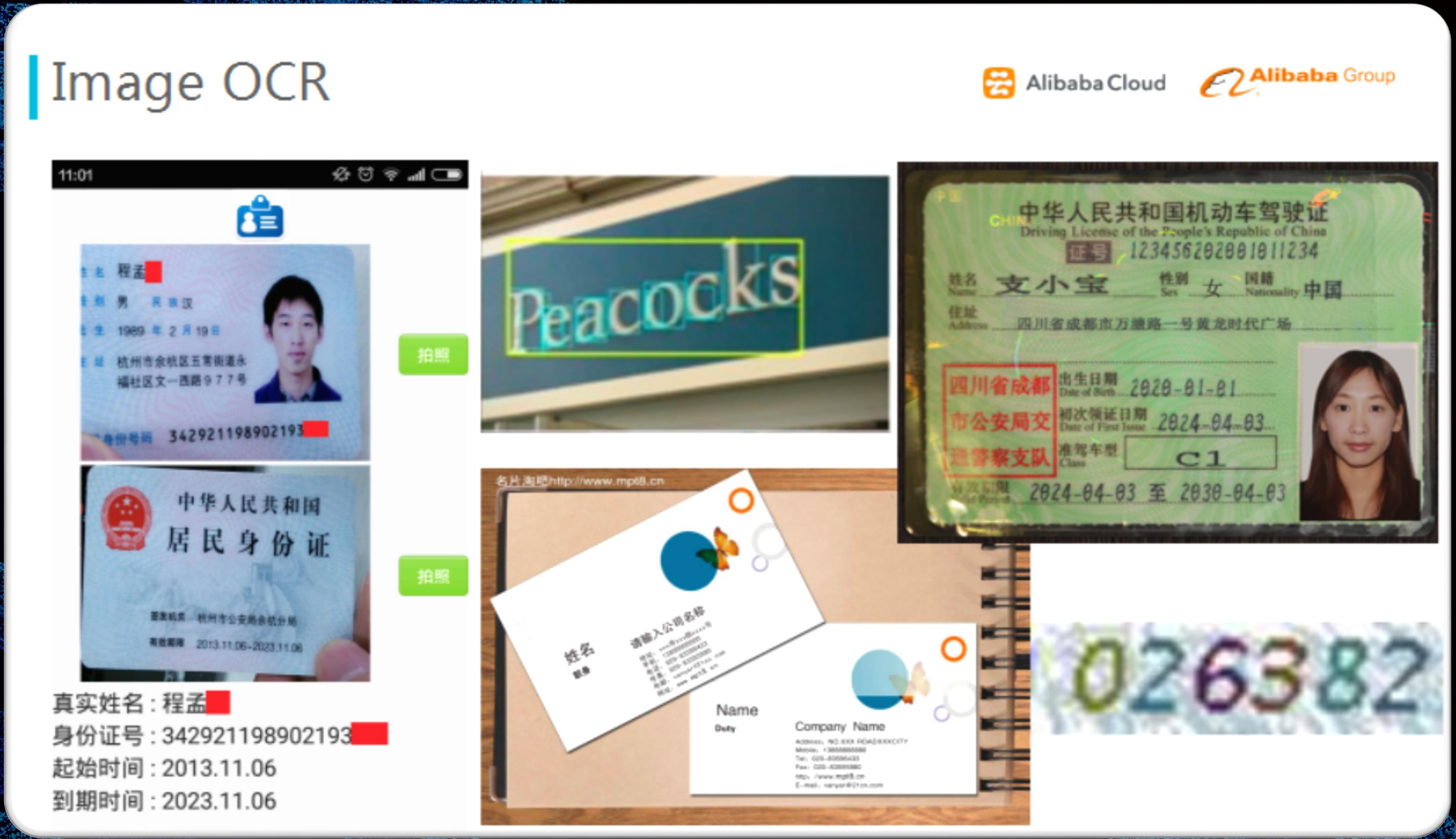


Image OCR
Alibaba Cloud
Alibaba Group
11:01
拍照
拍照
中华人民共和国
居民身份证
Peacocks
名片海吧http://www.mp8.cn
姓名
Name
Company Name
中华人民共和国机动车驾驶证
Driving License of the People's Republic of China
证号 123456202001011234
姓名 支小宝 性别 女 国籍 中国
住址 四川省成都市万腾路一号黄龙时代广场
出生日期 2020-01-01
初次领证日期 2024-04-03
准驾车型 C1
有效期限 2024-04-03 至 2030-04-03
四川省成都市公安局交通警察支队
026382

真实姓名：程孟[redacted]

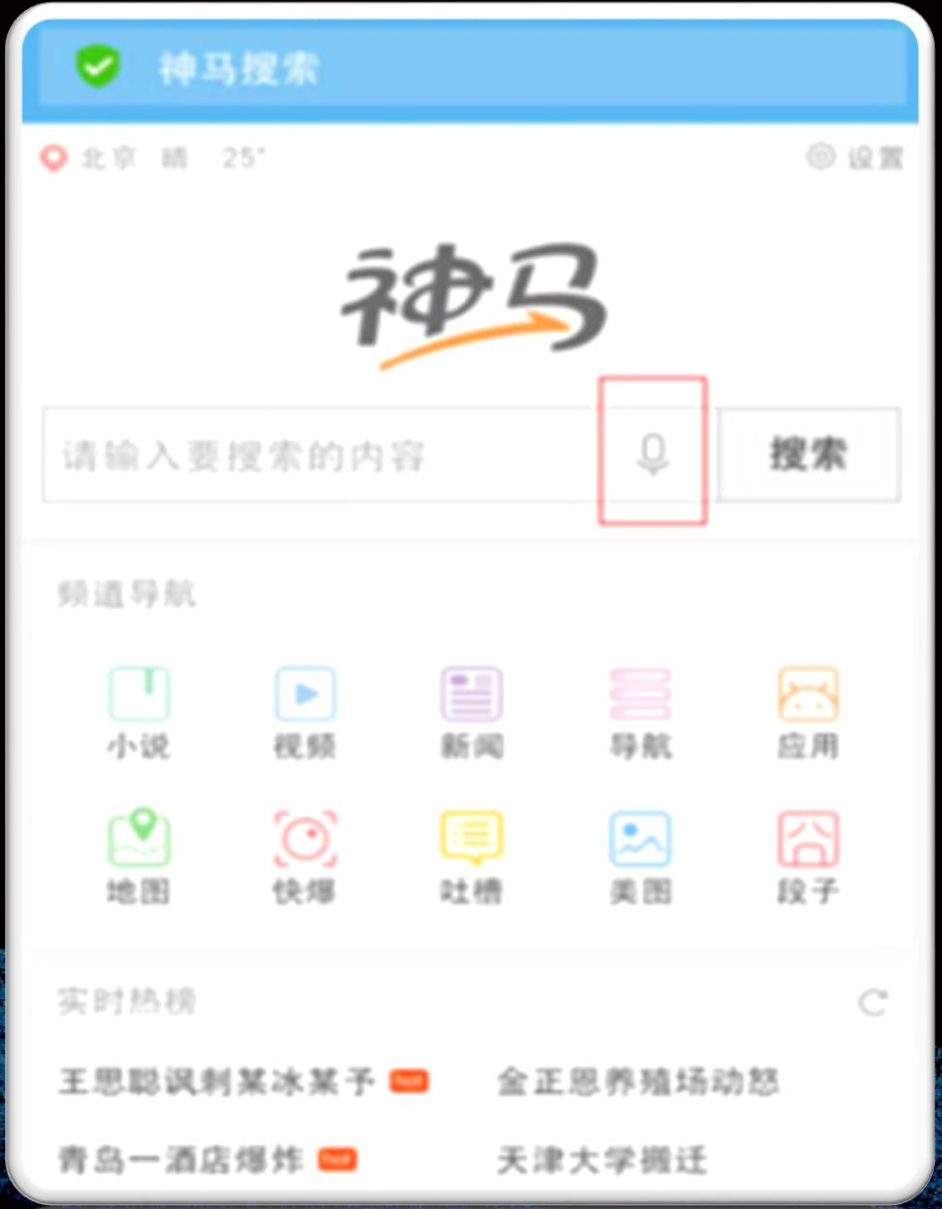
身份证号：342921198902193[redacted]

起始时间：2013.11.06

到期时间：2023.11.06

- 10s Millions of Image
- CNN Model
- Single character accuracy 99.6%
- Overall accuracy 93%
- 8 way distributed GPU solution
- 7x training speed

Deep Learning Everywhere



神马搜索

北京 晴 25°　设置

神马

请输入要搜索的内容　搜索

频道导航

实时热榜

王思聪讽刺某冰某子 hot　全正恩养殖场动怒

青岛一酒店爆炸 hot　天津大学搬迁

Translation

Voice

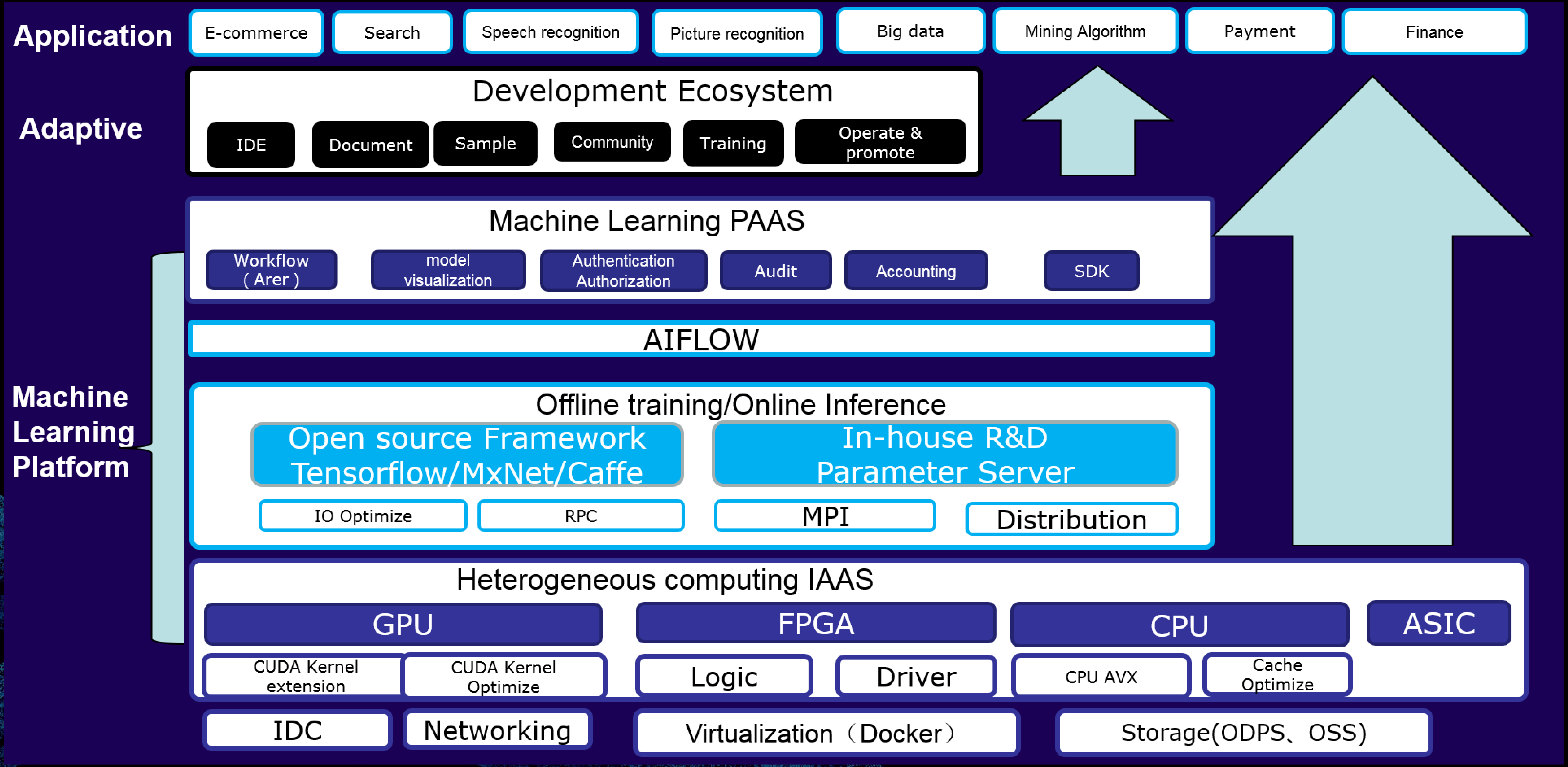
定损原理

行业专家与算法工程师通力合作，精细化打磨产品，实现车辆定损

全链路深度学

Insurance

Heterogeneous Machine Learning Platform



Application

E-commerce | Search | Speech recognition | Picture recognition | Big data | Mining Algorithm | Payment | Finance

Adaptive

Development Ecosystem

IDE | Document | Sample | Community | Training | Operate & promote

Machine Learning Platform

Machine Learning PAAS

Workflow (Arer) | model visualization | Authentication Authorization | Audit | Accounting | SDK

AIFLOW

Offline training/Online Inference

Open source Framework Tensorflow/MxNet/Caffe

In-house R&D Parameter Server

IO Optimize | RPC | MPI | Distribution

Heterogeneous computing IAAS

GPU: CUDA Kernel extension | CUDA Kernel Optimize

FPGA: Logic | Driver

CPU: CPU AVX | Cache Optimize

ASIC

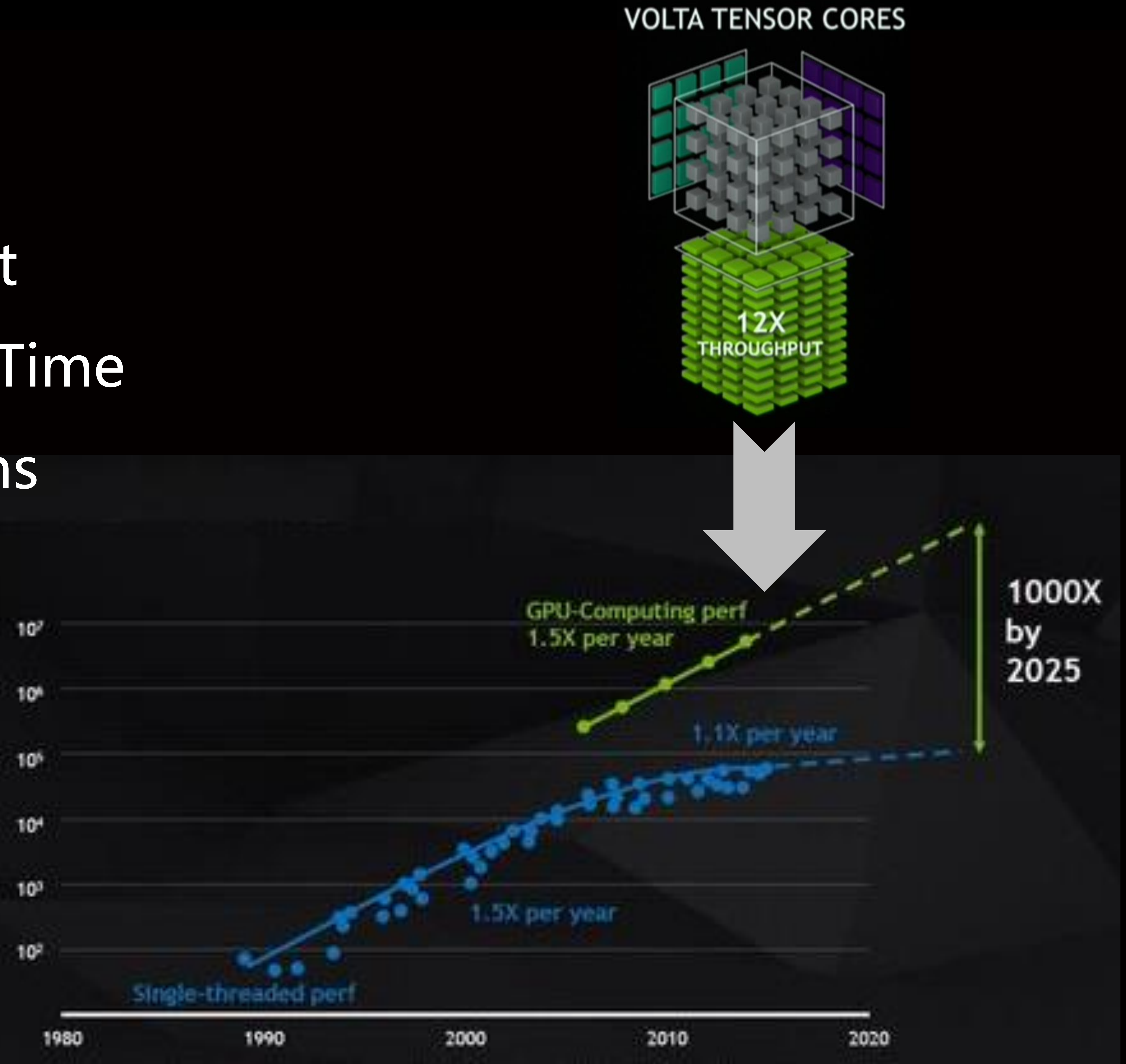
IDC | Networking | Virtualization（Docker） | Storage(ODPS、OSS)

Hardware Accelerated AI

- Training: Compute Intensive, Time Cost
- Inference: Service Oriented, Response Time
- Eco-System: Framework, Libs, Precisions
- Hardware Dividends for everyone

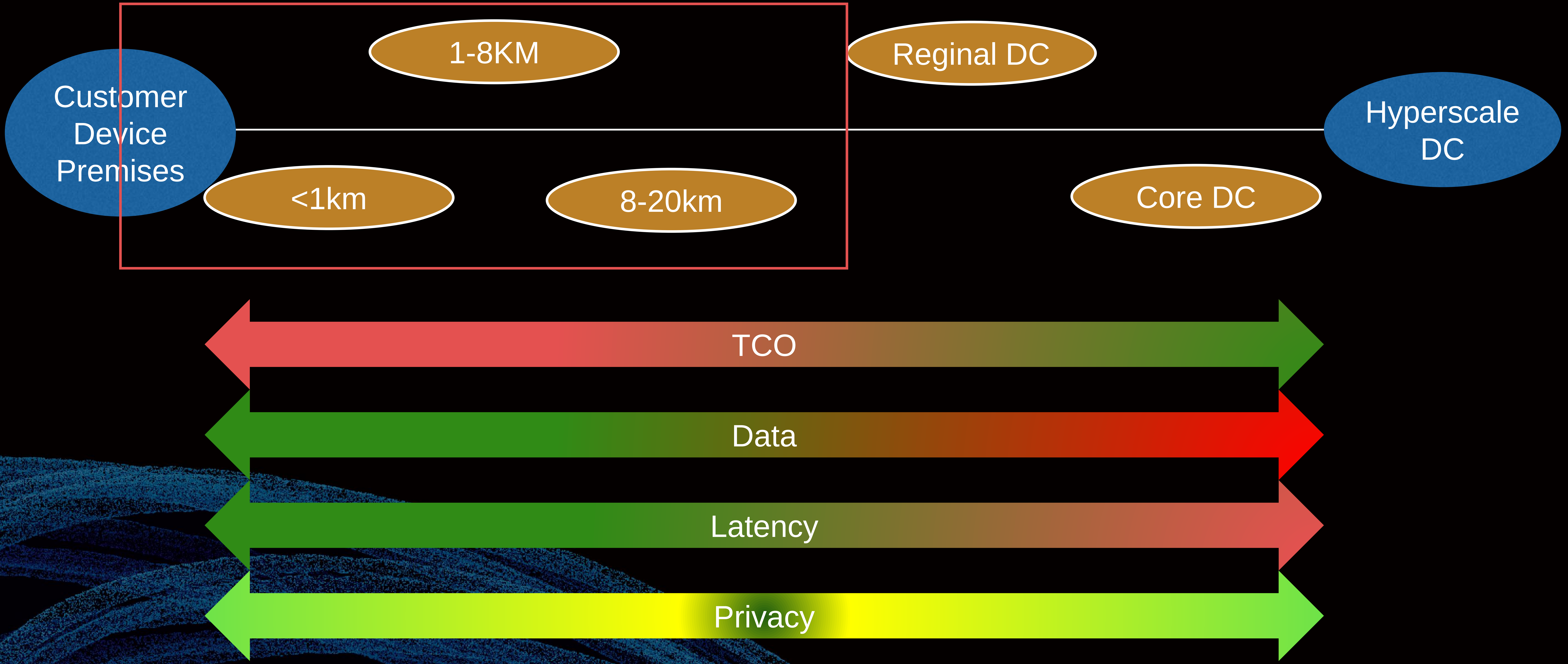
Tipping point:

- Google TPUs
- Volta TensorCore
- New hardware accelerators for AI

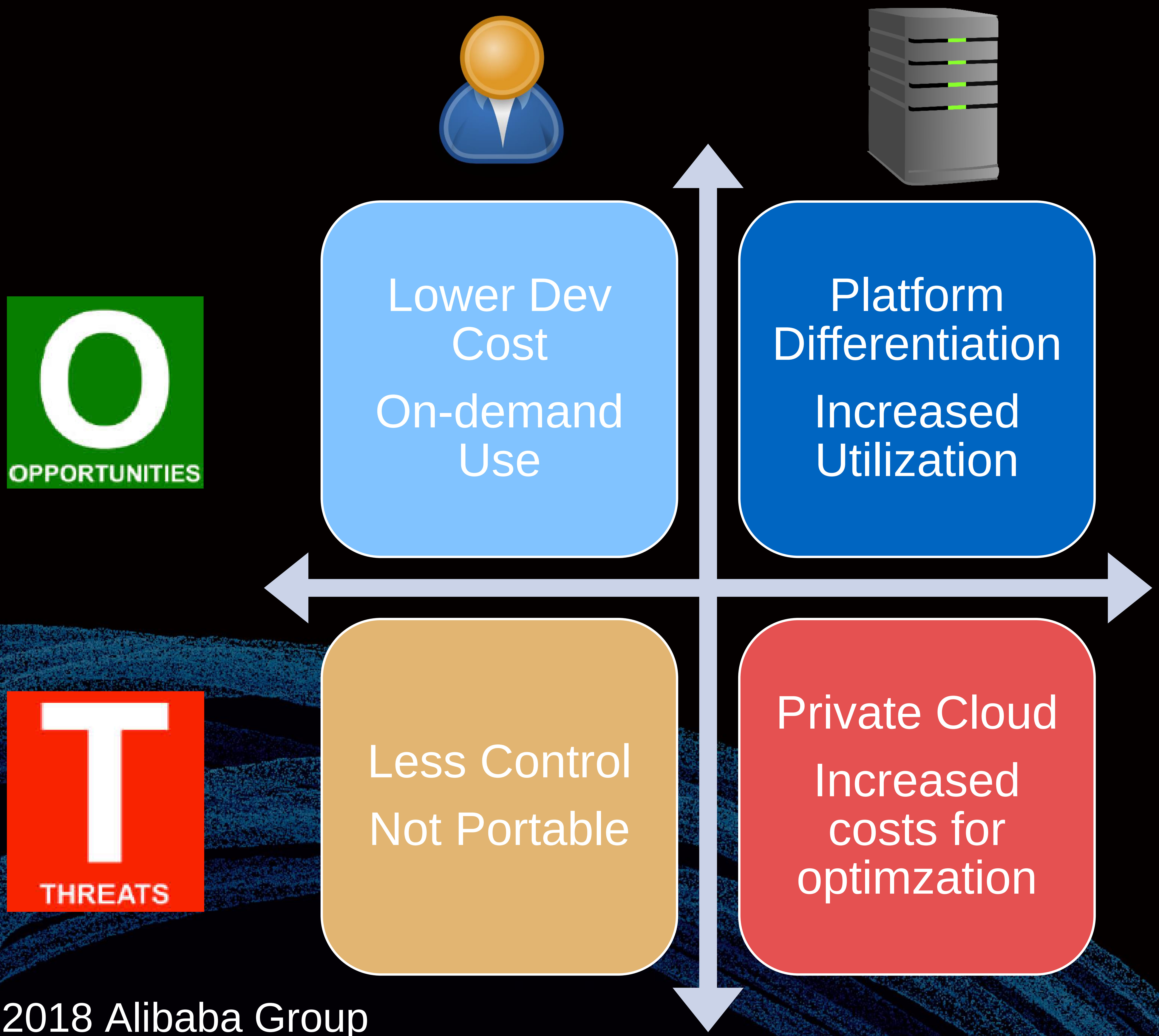


*data from NVIDIA GTC 2017

Edge – Forces of Gravity



Function Computing



O OPPORTUNITIES

Lower Dev Cost

On-demand Use

Platform Differentiation

Increased Utilization

T THREATS

Less Control

Not Portable

Private Cloud

Increased costs for optimzation

Technology Challenges

- Quality of Service
- Infrastructure Utilization
- Accelerator Efficiency
- Capacity Granularity
- Multi-Tenancy Management
- Demand Projection
- Scheduling
- Compatibility

Opportunities & Challenges

Perf
Power
Stability

Fine Grained
Monitoring
For Efficiency

Deep Customization
Best TCO
Competitive

Empower
“Traditional”
Algorithms

Thank You!

E-mail: lingjie.xu@alibaba-inc.com