Baidu ABC Platform

Vice President , Baidu
Watson Yin

Baidu Introduction

Baidu Cloud

| Most Popular Chinese Search Engine | Most influential Internet Company |

79.84%
Baidu Mobile+PC
Search Market Share
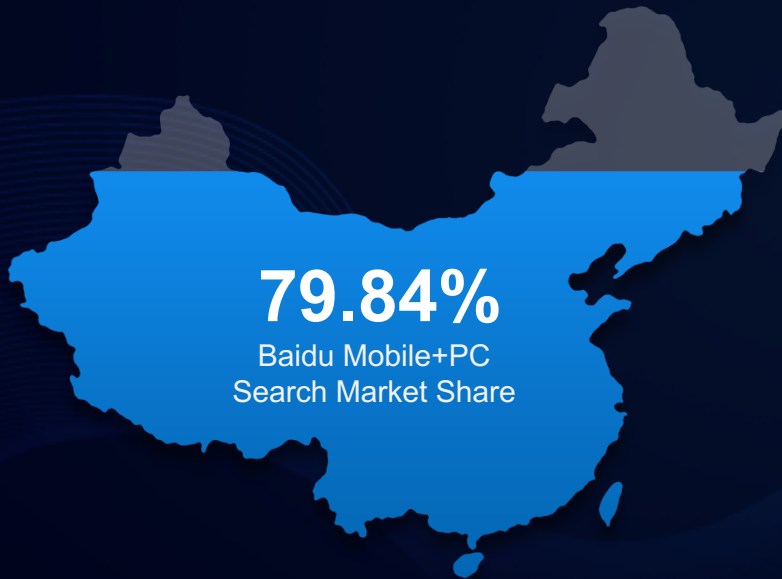
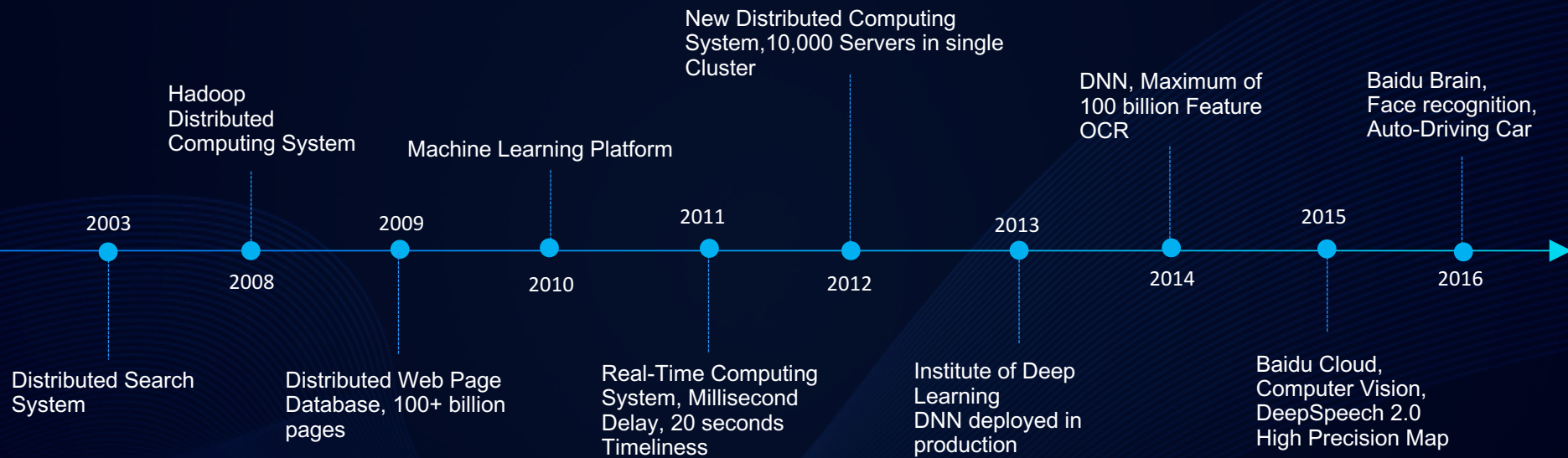**92.1%** Baidu penetration rate of PC Internet users

667 Millions Baidu Mobile search MAU

90.3% Baidu penetration rate of wireless Internet users

30 Billions Daily responded location requests on Baidu Map LBS open platform

6 Billions Daily responded search requests

Journey of Baidu Technology

Baidu Cloud

New Distributed Computing System,10,000 Servers in single Cluster

Hadoop Distributed Computing System

DNN, Maximum of 100 billion Feature OCR

Baidu Brain, Face recognition, Auto-Driving Car

Machine Learning Platform

2003

2009

2011

2013

2015

2008

2010

2012

2014

2016

Distributed Search System

Distributed Web Page Database, 100+ billion pages

Real-Time Computing System, Millisecond Delay, 20 seconds Timeliness

Institute of Deep Learning DNN deployed in production

Baidu Cloud, Computer Vision, DeepSpeech 2.0 High Precision Map

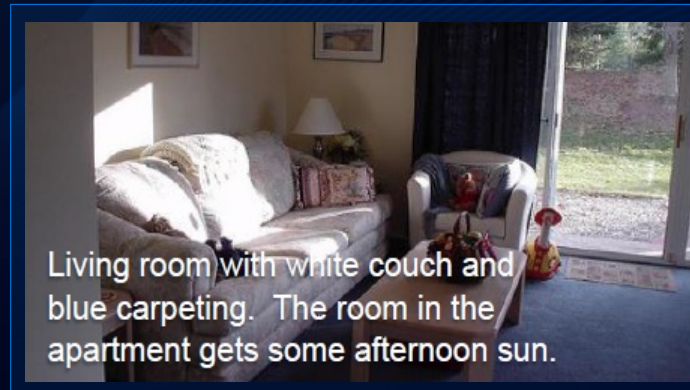**Baidu Brain: 'Crown Jewel' of Baidu Artificial Intelligence**

Baidu Cloud

One of Largest Deep Neural Network in the world

Trillion level parameter

Support hundreds of billions of samples and feature training

Baidu ABC Cloud Platform
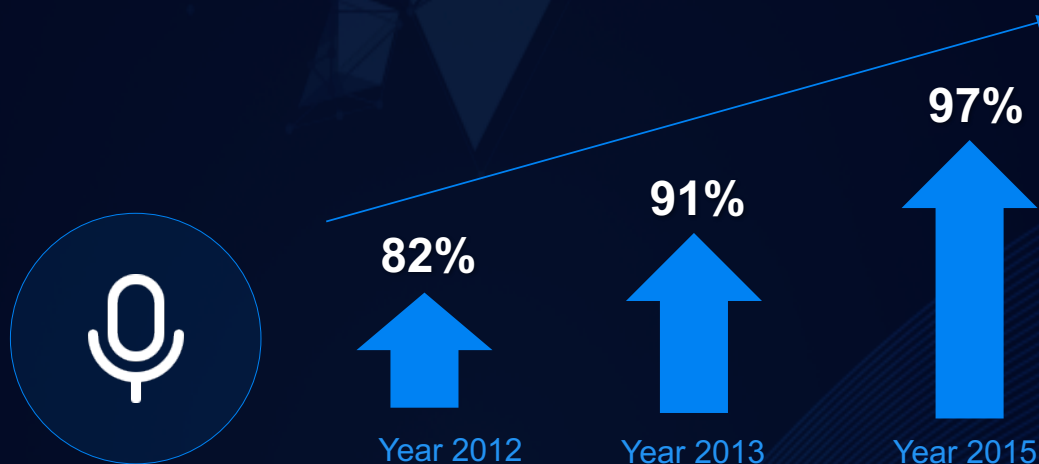
Baidu Cloud

Cloud Computing + Big Data + AI = Baidu Cloud

Baidu Products

External Customers

AI
One of World Largest Deep Neural Network (Trillion Parameter)
One of World Biggest Deep Learning Open Source Platform

Big Data
One of World best General Recommendation Engine
ZB Data Storage, 100+ PB Data Processing per Day

Cloud Computing
IDC Resource Management Platform, 100,000+ Servers Management
PUE < = 1.11 , 10,000+ Servers Installation in One Day

Asia Largest IDC: Yangquan

Baidu Cloud

120,000 m^2 , 75 MW, 200+ Patents

MIIT&TGG
First IDC with 5A Certificate of Operation and Design in China

WWF&IDC
Best IT Innovation in Energy Efficiency

CDCC
Energy Efficiency Award 2015

HVDC Offline

Water Cooling

OCU

Photovo -ltaics System

Power Supply Efficiency > 99.5%

65% Free Cooling Time

AC Efficiency ↑ 30%

CO_2 Emission ↓ 110T

Modular Data Center
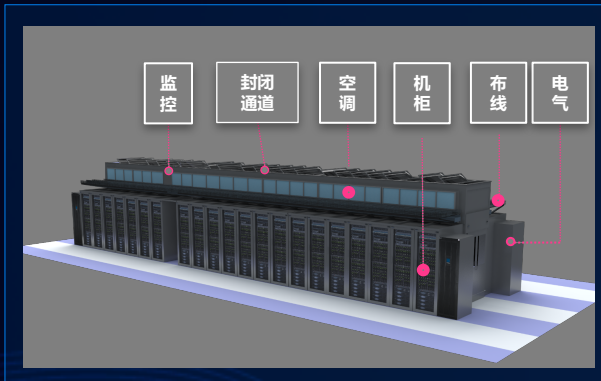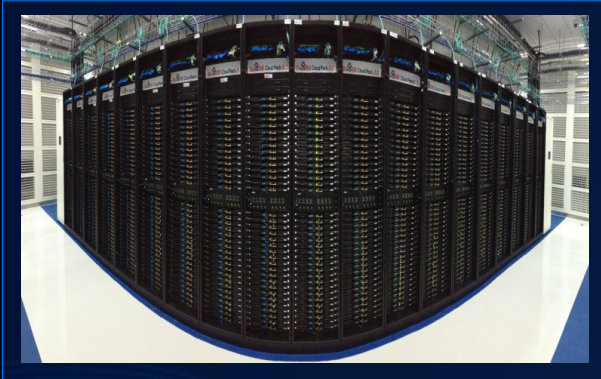

Modular: OCU cooling unit

15 components, deployment time ↓50%+

Modular：Prefabricated Data Center

First deployment in Internet IDCs in China

Productized, Fast delivery with high reliability, On-demand deployment, Eco-friendly

Project Scorpio

First open source hardware solution in China

Centralized power and thermal, AI solution for management

Power efficiency ↑15％, Thermal efficiency ↑70％

Delivery efficiency ↑20 times

Iceberg Cold Storage Server

HDD deep hibernate strategy

18 HDD/U （Biggest storage density）

Power consumption ↓50%

Failure rate ↓60%

Baidu Cloud

South-North、South-East network across China

Features :

- End to end in house design

- Industry largest deployment volume

- Hardware open source

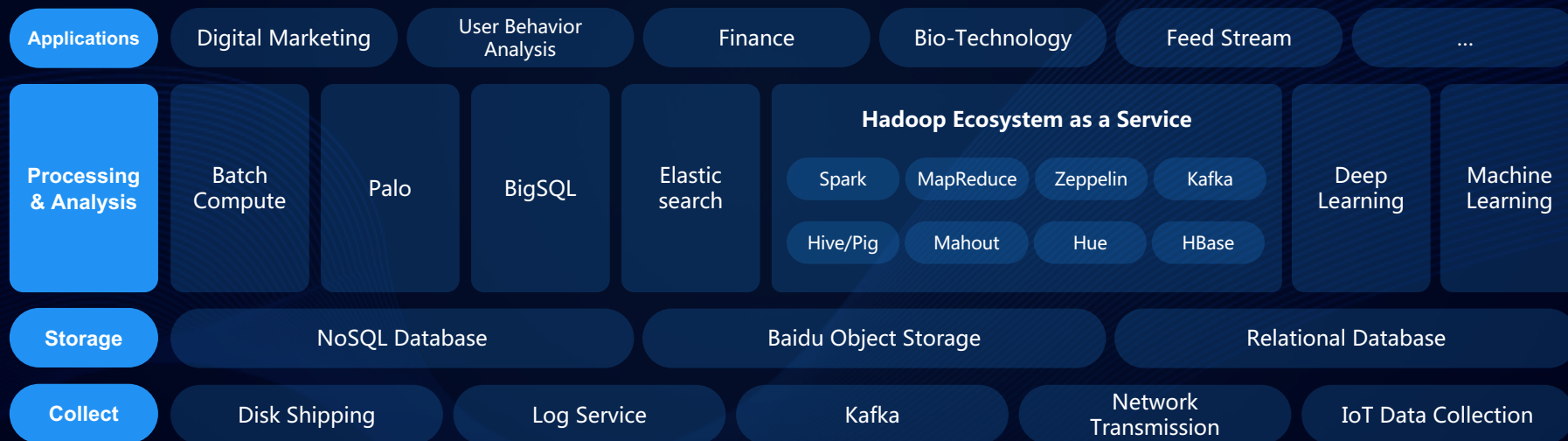25G TOR（Kevin）

100G Core switch（Jerry）

Big Data at Baidu

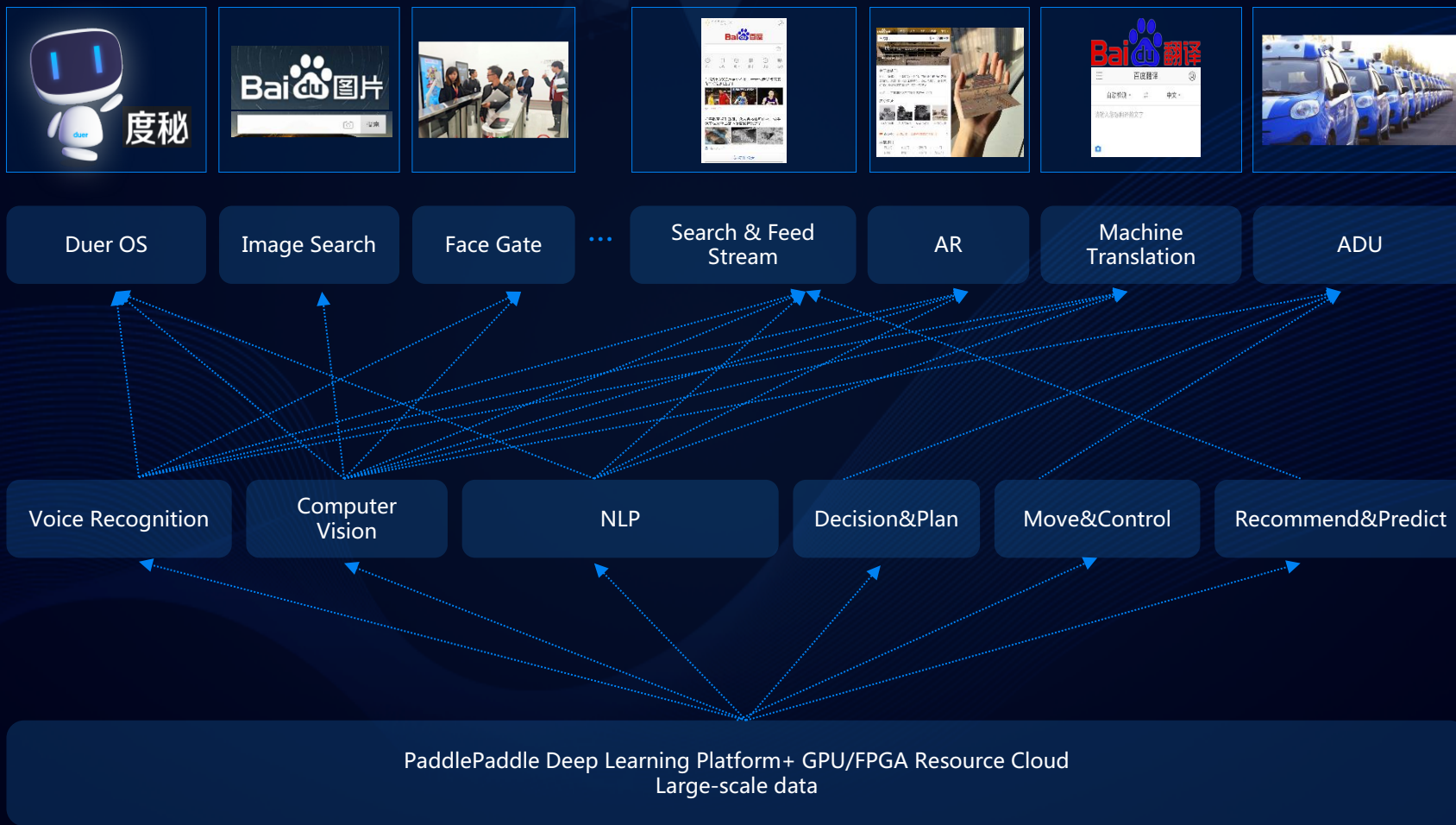**Baidu Cloud**

1,000,000,000,000+
web pages

1,000,000,000+
searches/day

10,000,000,000+
images/videos

10,000,000,000+
POI data

Applications	Digital Marketing	User Behavior Analysis	Finance	Bio-Technology	Feed Stream	...

Processing & Analysis	Batch Compute	Palo	BigSQL	Elastic search	Hadoop Ecosystem as a Service	Deep Learning	Machine Learning

Hadoop Ecosystem as a Service

Spark	MapReduce	Zeppelin	Kafka
Hive/Pig	Mahout	Hue	HBase

Storage	NoSQL Database	Baidu Object Storage	Relational Database

Collect	Disk Shipping	Log Service	Kafka	Network Transmission	IoT Data Collection

Baidu AI Architecture

Baidu Cloud

| Duer OS | Image Search | Face Gate | ... | Search & Feed Stream | AR | Machine Translation | ADU |

| Voice Recognition | Computer Vision | NLP | Decision&Plan | Move&Control | Recommend&Predict |

PaddlePaddle Deep Learning Platform+ GPU/FPGA Resource Cloud
Large-scale data

AI Computing Engine

Baidu Cloud

Baidu Brain Computing Engine

	GPU		FPGA
Distributed Communication	MPI, GPU Peer to Peer HGCP model training cluster		Self research network
Development Environment	Cuda	OpenCL	FPGA LIB/API
Computing Library	GEMM, Conv		FPGA Logic
Network	40/100G RDMA		FPGA high-speed interconnection
Server	1 ~ 64 cardboards per server		
GPU/FPGA	Training : 10T Inference : 5.2T		24/16bit: 4T 8bit: 9.6T

GPU   **FPGA**

GPU

RDMA Network

100Gb RDMA

100G NIC — Compute Node-1

Host CPU

CPU-A1 — QPI — CPU-A2

Compute Node-n

0.5/1/2x128Gb

PCIe Network

PCIe core switch — Cluster Manager

4x128Gb

Devices

X-Man-1 — 2x128Gb

GPU /FPGAx4 — GPU /FPGAx4

GPU/FPGA x4 — GPU /FPGAx4

X-Man - 4

Highlights

- In house design GPU Solution

- PCI-E Fabric, micro-seconds Latency

- 100G RDMA Ready

- 640TFlops/Suite

THANK YOU

cloud.baidu.com