AI驱动的又一星球级计算

Planet Scale computing driven by AI

比特大陆

汤炜伟

2018.4

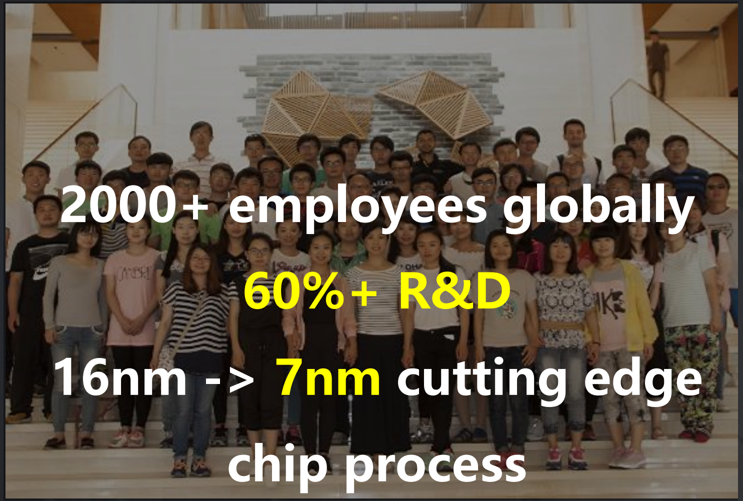
BITMAIN – AI and Blockchain chip company

2013

Founded in Beijing

HPC technology

Silicon chip design technology



2000+ employees globally

60%+ R&D

16nm -> 7nm cutting edge chip process

2015 year end

Start AI

Billions chips shipped

Top 1 Digital currency chip provider

85%+Market share globally

Exascale computing

Billion Watt datacenter

2017

Shipped SOPHON AI Chip, card and server

Industries + AI, significantly changing everything

	语音识别/合成	图像识别/机器视觉	自然语言处理	机器人
安防		人脸识别，视频结构化		
互联网	客服	精准营销	精准投放	仓储机器人
消费电子	智能音箱	手机		
汽车		ADAS		
医疗	问诊	影像诊断	电子病历	手术机器人
金融	客服	人脸、身份验证	智能投顾推荐	服务机器人

资料来源：中金公司，BITMAIN内部分析

Computing driven by AI – Internet Photos and Videos

Bitcoin Network Hashrate

~10 Exa DHash

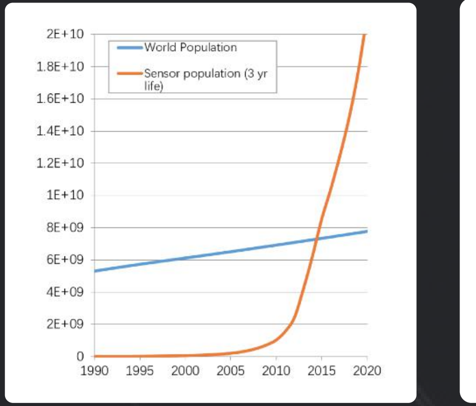
- 3.4 Billion global Internet users
- Assume 20 minutes video per user per day, from mobile phone, video call, robot …

Assume drawing 2 images per second, 0.1T FLOP/image Total operations: 50 Exa Flops

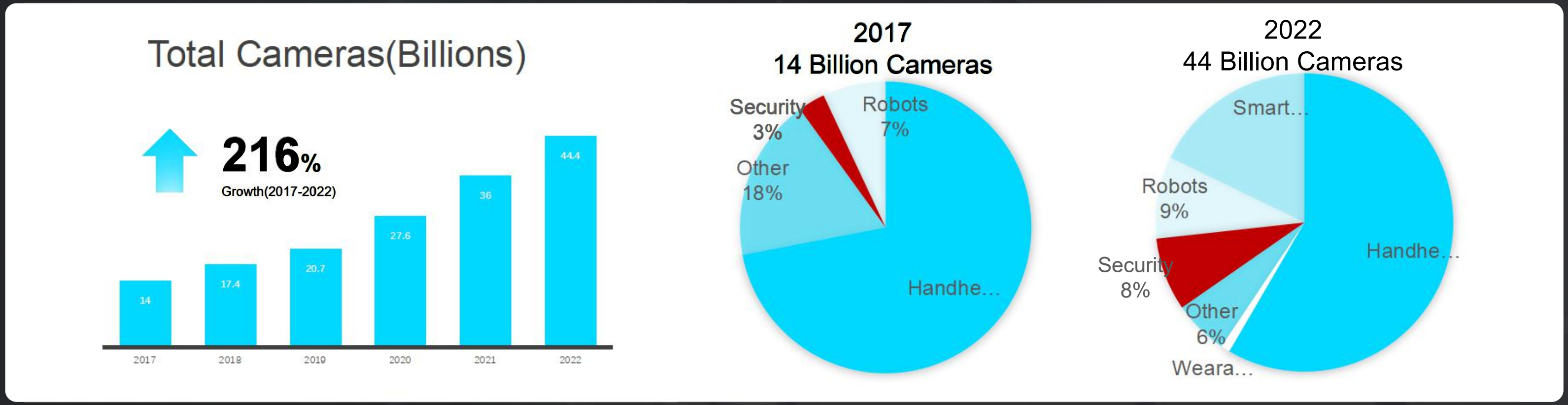
Computing driven by AI –Voices and NLP

- 3.4 Billion global Internet users
- Assume 30 minutes voice data per user per day
- Assume 1 second voice data requires1T FLOP
- Total operations: 71 Exa Flops

Computing driven by AI – Surveillance Cameras



Source: Shared by Chris Rowen, founder of Cognite Ventures LLC

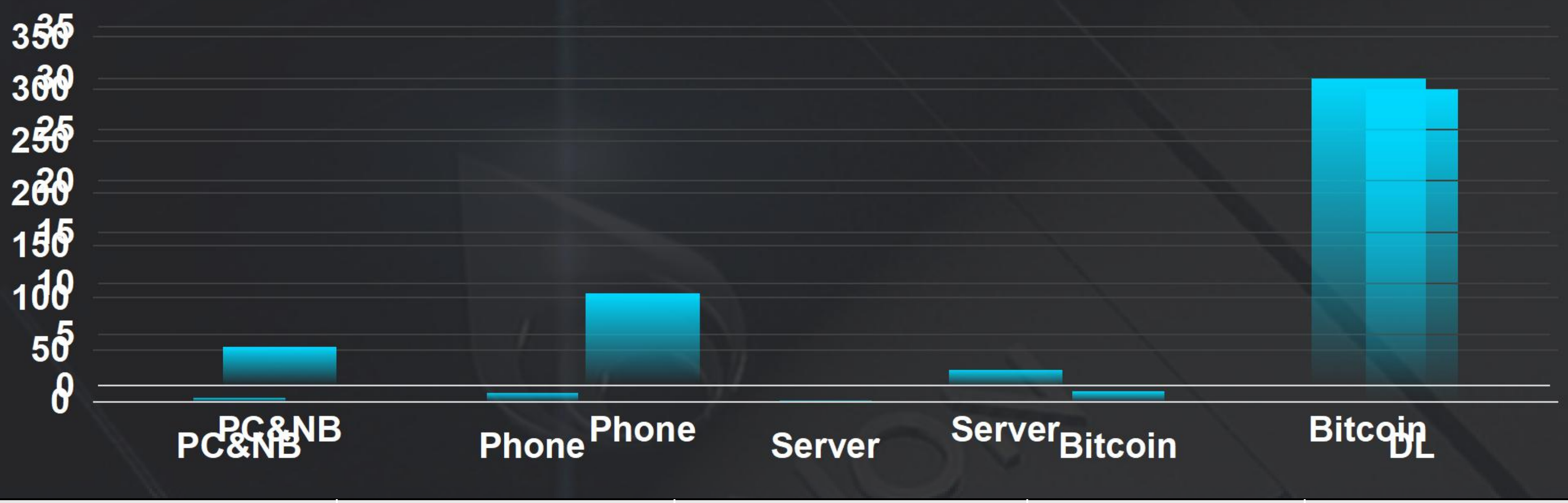


Source: Growth in cameras in the next five years will be primarily driven by the move toward depth/3D capture and the integration of cameras into a wide range of existing products ©LDV Capital

Assume:

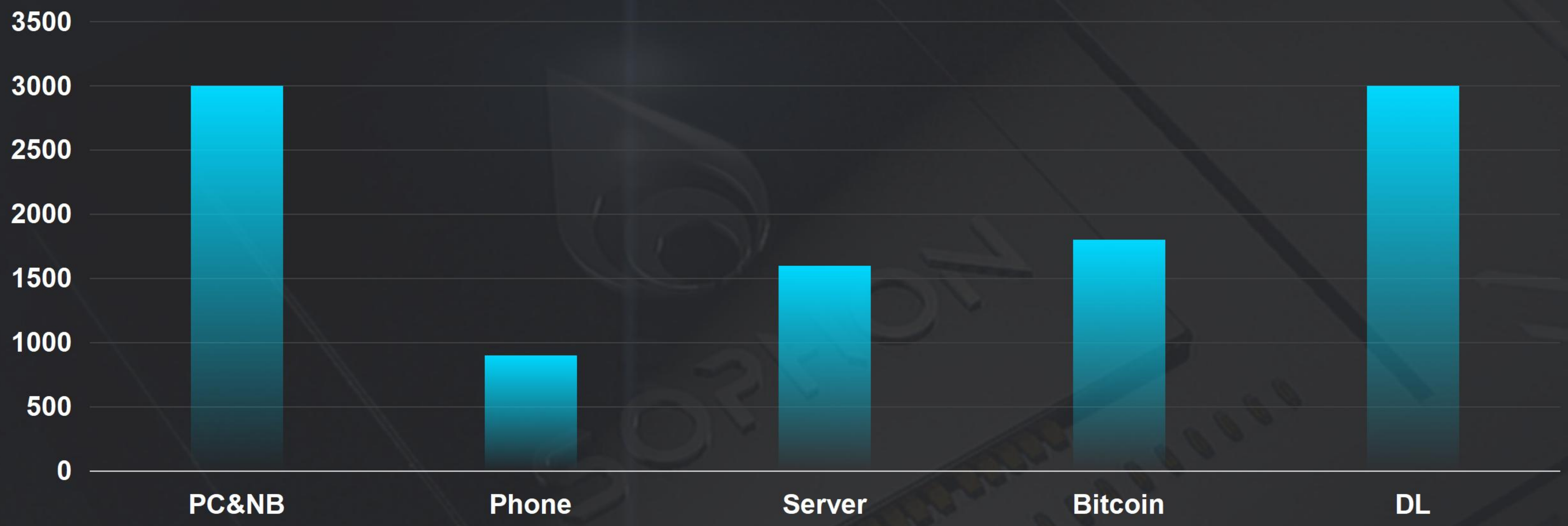
0.5T Flops/surveillance camera 2017: 210 Exa Flops 2022: 1,760 Exa Flops

Planet Scale Computing : Exa operations



-	PC&Notebook	Mobile Phone	Server	Bitcoin
CPU#/year	250,000,000	1,500,000,000	20,000,000	
core#	2	2	8	
frequency	2.5	1.5	2	
retention years	3	2	5	
total operation (Exa)	3.8	9.0	1.6	30
power/CPU	40	2	100	
active time %	30%	30%	80%	100%
total power(MW)	3,000	900	1,600	1,800

Planet Scale Computing : mega watt operations



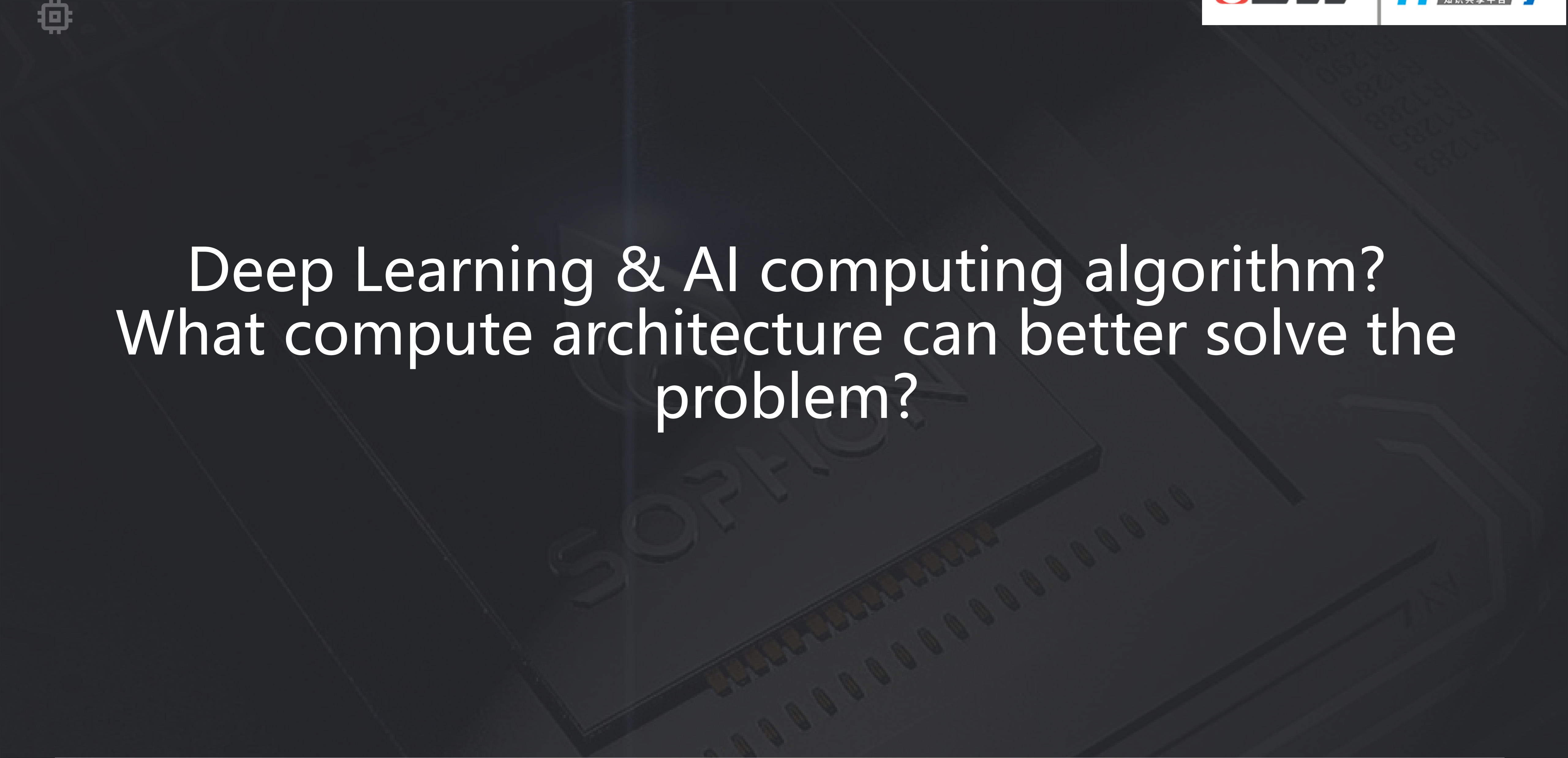
Assume DL power efficiency: 10W per T Flops

Industry computing driven by AI

	50 Exaflops	71 Exaflops	210 Exaflops
CPU计算	50 Million CPU	70 million CPU	200Million CPU
	50 Billion USD	70 Billion USD	200 Billion USD
	5 Billion Watt**	7 Billion Watt**	20 Billion Watt**
GPU计算	10 Million GPU	14 Million GPU	40 Million GPU
	20 Billion USD	20 Billion USD	80 Billion USD
	1 Billion Watt**	1.4 Billion Watt*	4 Billion Watt**

基于现有的CPU、GPU的AI计算带来高成本、高能耗

**以常见产品为例1CPU 100W 1Tflops， 1GPU 100W 5Tflops



Deep Learning & AI computing algorithm? What compute architecture can better solve the problem?

Deep Learning requires 4-dimensional tensor computing

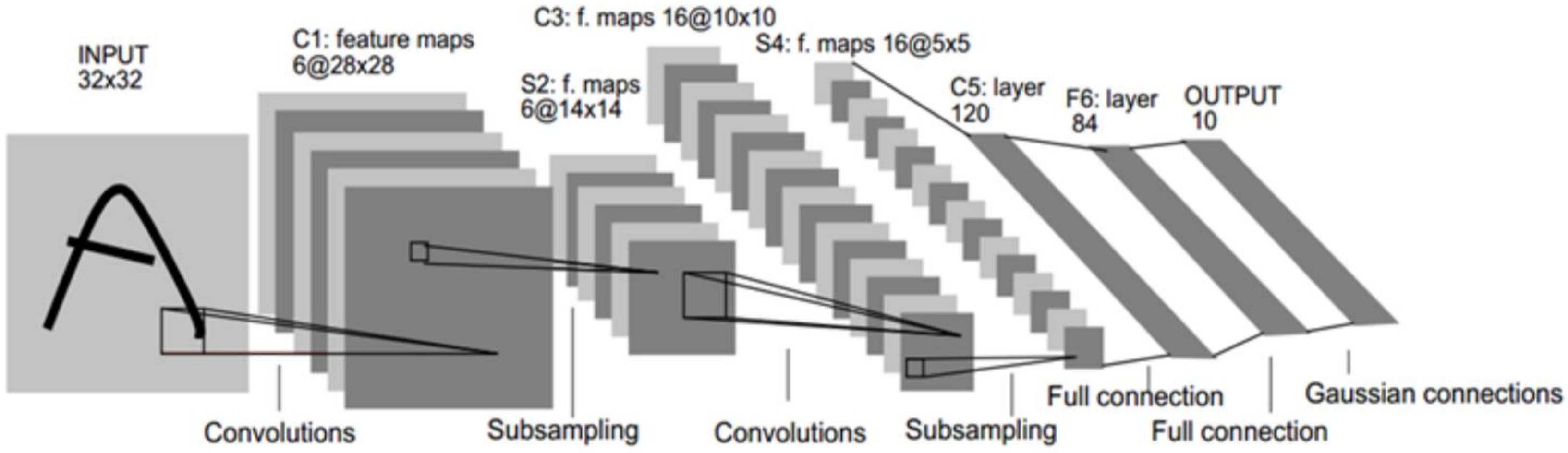
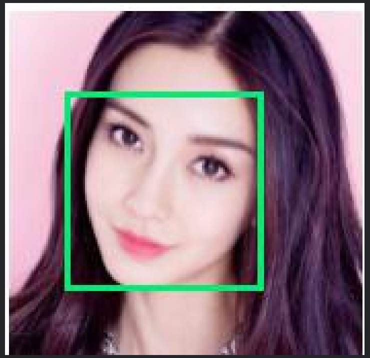


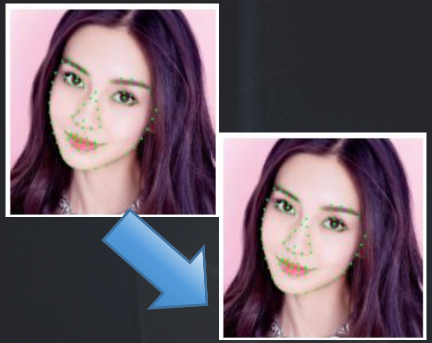
Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Source: Architecture of LeNET-5, a Convolution Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

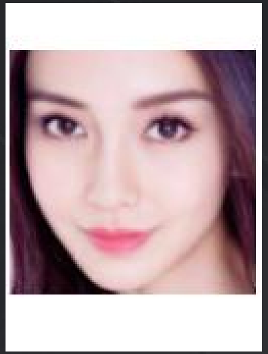
Example：Face recognition need several neural ne
以脸识别为例：多种神经网络模型的密集张量计算



检测
人脸框



跟踪



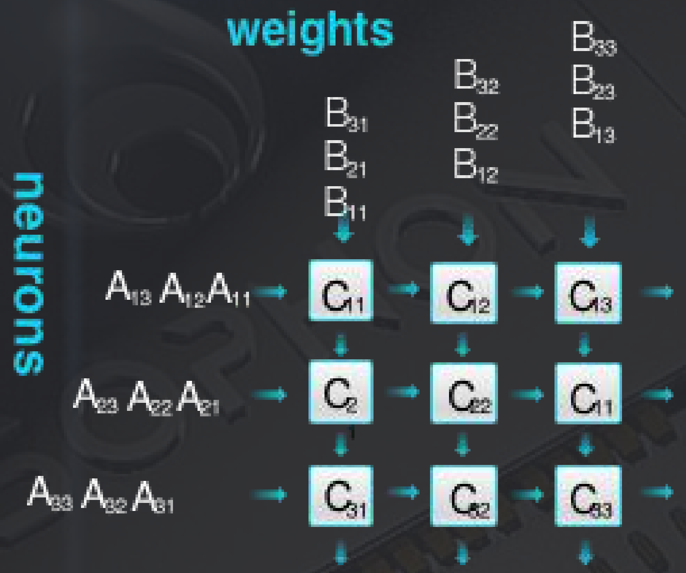
识别
特征提取和属性识别

静态比对1:N
动态比对m:N

	检测	跟踪	识别	比对
应用	人脸框	轨迹和图像选取	特征和属性提取	
典型网络/技术	MTCNN SSH	轻量级NN/ 卡尔曼滤波	RESNET GOOGLENET	多维空间搜索
每张图计算量	~10Gflops	~Gflops	1~0Gflops	
每秒计算力	约为 500Gflops (0.5Tflops)			取决于库规模和速度要求

High Performance Systolic Tensor Processor

- Huge systolic MAC array
- 4-dimensional data moving automatically
- Several dozen Watts TDP


weights
neurons
B_{31} B_{21} B_{11}
B_{32} B_{22} B_{12}
B_{33} B_{23} B_{13}
A_{13} A_{12} A_{11}
A_{23} A_{22} A_{21}
A_{33} A_{32} A_{31}
C_{11}
C_{12}
C_{13}
C_2
C_{22}
C_{11}
C_{31}
C_{32}
C_{33}

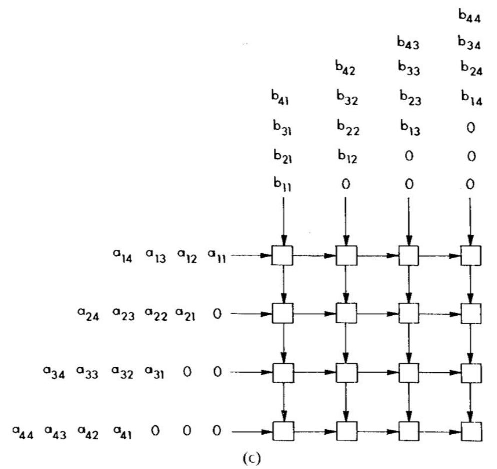

b_{44}
b_{43} b_{34}
b_{42} b_{33} b_{24}
b_{41} b_{32} b_{23} b_{14}
b_{31} b_{22} b_{13} 0
b_{21} b_{12} 0 0
b_{11} 0 0 0
a_{14} a_{13} a_{12} a_{11}
a_{24} a_{23} a_{22} a_{21} 0
a_{34} a_{33} a_{32} a_{31} 0 0
a_{44} a_{43} a_{42} a_{41} 0 0 0
(c)

Fig. 8. (a) An SFG for matrix multiplication. (b) The detailed diagram of the processing nodes. (c) A systolic array for matrix multiplication.

Source: Paper ‘On Supercomputing with Systolic/Warefront Array Processor’ published in 1984, by SUN-YUAN KUNG, SENIOR MEMBER, IEEE

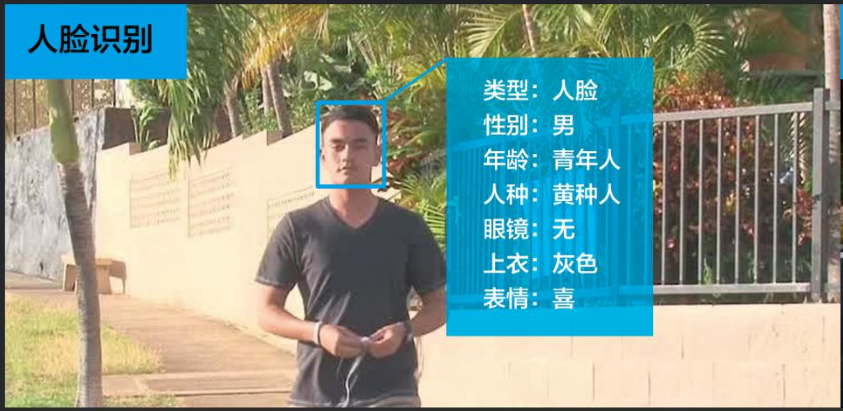
SOPHON-BM1680 : Deep Learning ASIC

- Enhanced Systolic
- Developed since late 2015, taped out in April 2017
- Samples in June 2017
- 1st DL ASIC succeeding to the Google TPU. Available production now.

SOPHON-SS1 : Intelligent Video Analysis Server

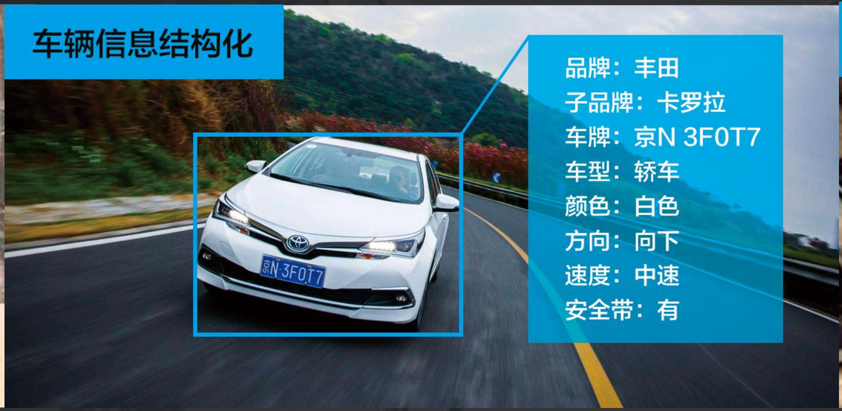
Available applications:

- Face Recognition
- Face/Pedestrian detection and attributes analysis
- Vehicle/Pedestrian detection and classification



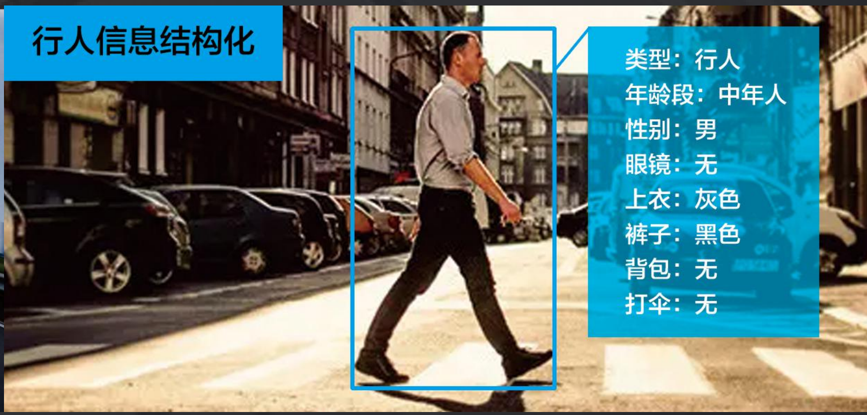
人脸识别

类型：人脸
性别：男
年龄：青年人
人种：黄种人
眼镜：无
上衣：灰色
表情：喜



车辆信息结构化

品牌：丰田
子品牌：卡罗拉
车牌：京N 3F0T7
车型：轿车
颜色：白色
方向：向下
速度：中速
安全带：有

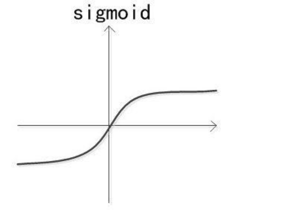


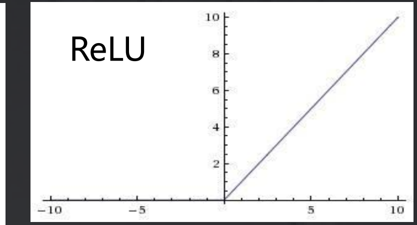
行人信息结构化

类型：行人
年龄段：中年人
性别：男
眼镜：无
上衣：灰色
裤子：黑色
背包：无
打伞：无

张量计算/矩阵计算算法能够设计优秀的POW算

矩阵的非线性运算 - HASH性质的不可逆加密（比如BTM）





- DNN的线性运算：矩阵的加法和数乘运算（Conv, FC)
- DNN的非线性运算：非线性映射函数 SIGMOD、TanH、ReLU
- Tensority的非线性举例: Scrypt, compress32to8, FNV, Random Proof

Cache Calculation:
Seed extent (SHA256)
Cache Extent (Scrypt)
32x1024x128

Matrix operation:
compress32to8
Convert int32 of data b =
(B0B1B2B3) ((big endian)) into int8

$$D = (B_2 + B_3) mod 2^8$$

Work Generation:
Hash Matrix
Binary Forwarded FNV

The entropy H(x) of
discrete random
variable X with
probability

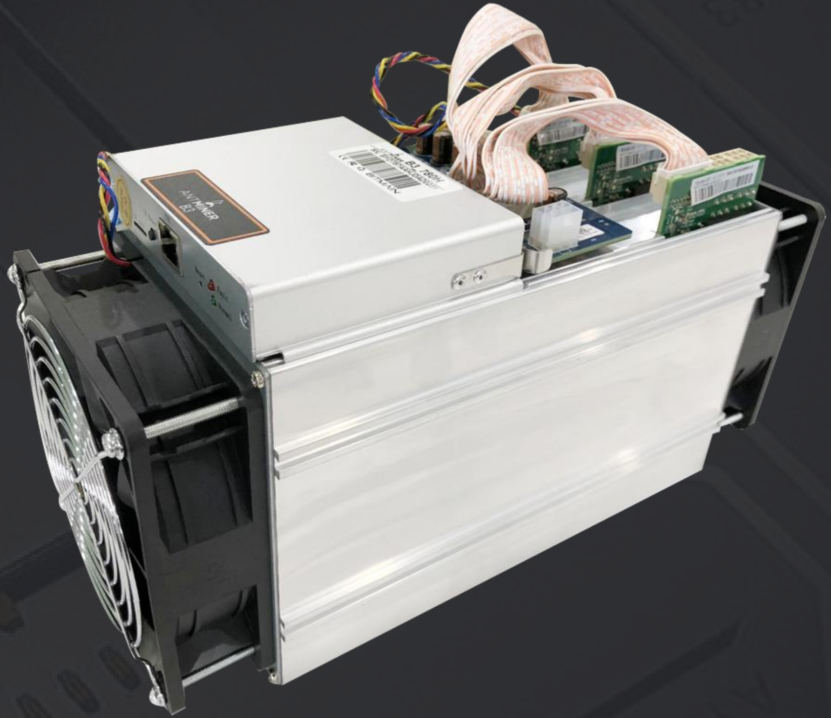
$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

人工智能芯片 + BYTOM POW公链

- 用人工智能芯片的强大算力，为POW公链保驾护航
  - BYTOM的POW算法充分利用了张量计算/矩阵计算的非线性特性
  - POW公链算法具有优秀的不可逆加密特性
  - 基于张量计算的POW公链算法具有很好的前景和生命力
- 比特大陆未来的每一代AI产品都会支持Bytom挖矿计算
  - AI芯片加速，更有助于BYTOM和AI的结合，生命力更强
  - 用AI智能机器赋能更强生产力，用Blockchain智能机器更好维护信任的公链

BTM的基础设施：B3矿机介绍

- **蚂蚁矿机B3 780H**
- 1.额定算力：780H / s ±5%
- 2.墙上功耗：360W+7%（不标配电源，可选配比特大陆APW3-1600瓦电源，AC / DC 93 %的效率，25℃环境温度测试）。
- 3.电源效率：0.46 J/H ±7%（墙上，AC / DC 93%的效率，25°C的环境温度测试）。



THANKS