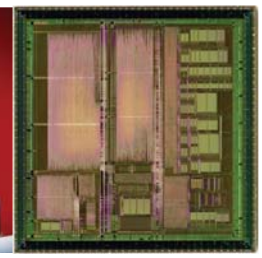

## Dr. Ren Wu

- **Independent researcher**
- Distinguished Scientist, Baidu
- HSA Chief Software Architect, AMD
- PI, HP Labs CUDA Research Center
- World Computer Xiangqi Champion

- **AI expert**
- **Heterogeneous Computing expert**
- **Computational scientist**

## Eighteen Years Ago - 05/11/1997

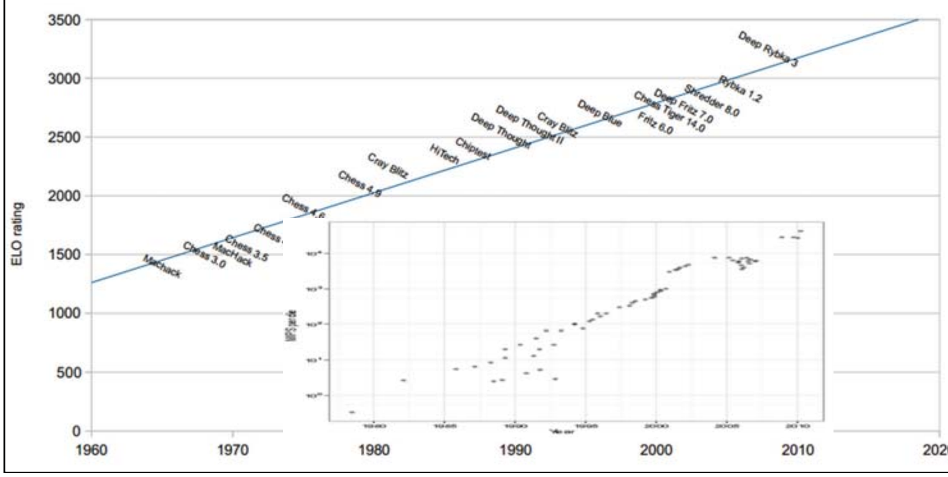




## Deep Blue


**A classic example of application-specific system design** comprised of an IBM supercomputer with 480 custom-made VLSI chess chips, running massively parallel search algorithm with highly optimized implementation.

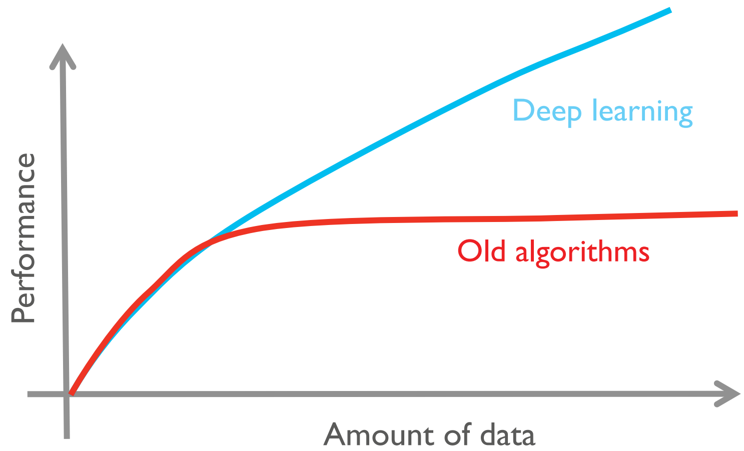
## Computer Chess and Moore's Law




## Deep Learning and Artificial Intelligence

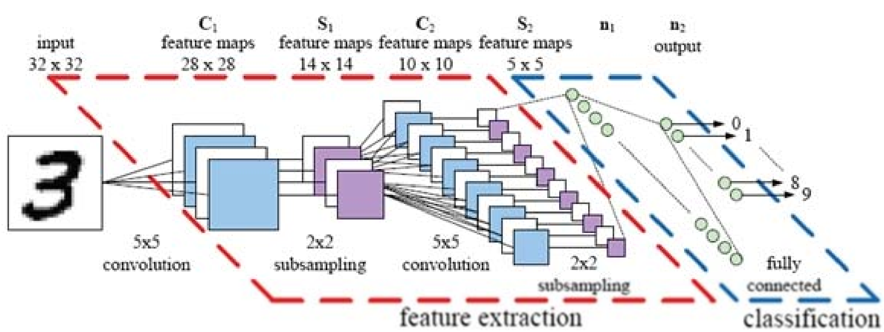
The Economist **Rise of the machines**

# Deep Learning






# Deep Convolutional Neural Networks




* Efficient mapping of the training of Convolutional Neural Networks to a CUDA-based cluster
courtesy of Jonatan Ward, Sergey Andreev, Francisco Heredia, Bogdan Lazar, Zlatka Manevska

## Big Data

| | |
|---|---|
| Storage | • >2000PB |
| Processing | • 10-100PB/day |
| Webpages | • 100b-1000b |
| Index | • 100b-1000b |
| Update | • 1b-10b/day |
| Log | • 100TB~1PB/day |



## Heterogeneous Computing

2000 world #1
ASCI White (IBM RS/6000SP)
6MW power, 106 tons
12.3 TFlops

1993 world #1
Think Machine CM5/1024
131 GFlops

2013
Samsung Note 3 smartphone
(Qualcomm SnapDragon 800)
129 Gflops

2013
Two MacPro workstation
(dual AMD GPUs each)
14 TFlops

# History is repeating itself!



# Deep Learning: Two Step Process

**Supercomputers used for training**

**And then deploy the trained models everywhere!**

Datacenters

Tablets, smartphones

Wearable devices

IoTs

# Deep Learning: Training

**Big data + Deep learning + High performance computing =**
**Intelligence**

**Big data + Deep learning + Heterogeneous computing =**
**Success**

Dr. Ren Wu @ MPSoC

# Insights and Inspirations

| | | |
|---|---|---|
| 多算胜少算不胜 | 元元本本殚见洽闻 | 明足以察秋毫之末 |
| 孙子 计篇 (Sun Tzu, 544-496 BC) | 班固 西都赋(Gu Ban, 32-92 AD) | 孟子梁惠王上 (Mencius, 372-289 BC) |
| More calculations win, few calculation lose | Meaning the more you see the more you know | ability to see very fine details |

Dr. Ren Wu @ MPSoC

## Project Minwa (敏娲)

- Minerva + Athena + 女娲
- Athena: Goddess of Wisdom, Warfare, Divine Intelligence, Architecture, and Crafts
- Minerva: Goddess of wisdom, magic, medicine, arts, commerce and defense
- 女娲: 抟土造人, 炼石补天, 婚姻, 乐器

World's Largest Artificial Neural Networks

❖ Pushing the State-of-the-Art
❖ ~ 100x bigger than previous ones
❖ **New kind of Intelligence?**

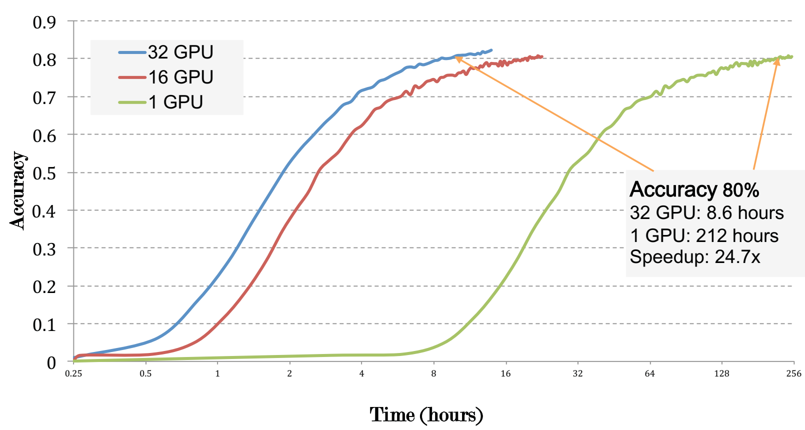
Dr. Ren Wu @ MPSoC

## Hardware/Software Co-design

- Stochastic gradient descent (SGD)
  - High compute density ⟹ **GPUs**
- Scale up, up to 100 nodes
  - High bandwidth low latency ⟹ **Infiniband**
- 36 nodes, 144 GPUs, 6.9TB Host, 1.7TB Device
  - **0.6 PFLOPS**
- **Highly Optimized software stack**
  - RDMA/GPU Direct
  - New data partition and communication strategies



## **Speedup** (wall time for convergence)




**Validation set accuracy for different numbers of GPUs**

Dr. Ren Wu @ MPSoC

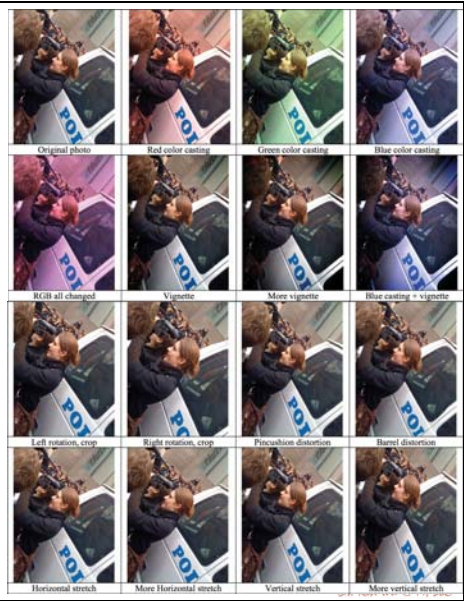
# Data Augmentation

Never have enough training examples!

Key observations
- Invariant to illuminant of the scene
- Invariant to observers

Augmentation approaches
- Color casting
- Optical distortion
- Rotation and cropping etc

**“见多识广”**



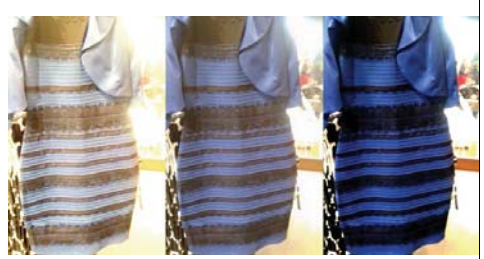

# The Color of the Dress

## And the **Color Constancy**

Key observations
- **Invariant to illuminant of the scene**
- Invariant to observers

Augmentation approaches
- **Color casting**
- Optical distortion
- Rotation and cropping etc

*“Inspired by the color constancy principal. Essentially, this ‘forces’ our neural network to develop its* ***own color constancy ability****.”*



# Data Augmentation

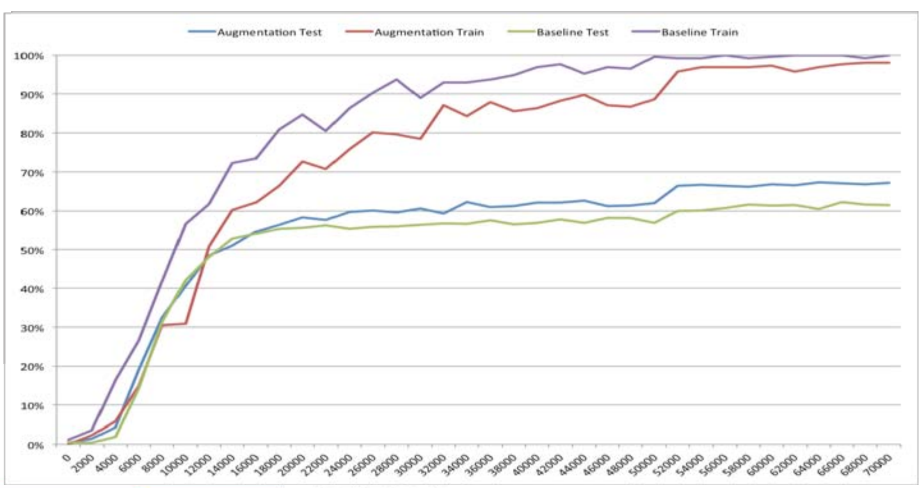
**Possible variations**

| Augmentation | The number of possible changes |
|---|---|
| Color casting | 68920 |
| Vignetting | 1960 |
| Lens distortion | 260 |
| Rotation | 20 |
| Flipping | 2 |
| Cropping | 82944(crop size is 224x224, input image size is 512x512) |

The Deep Image system learned from **~2 billion** examples, out of **90 billion** possible candidates.

Dr. Ren Wu @ MPSoC

# Data Augmentation vs. Overfitting




Dr. Ren Wu @ MPSoC

## Examples


Bathtub


Isopod


Indian elephant


Ice bear

Some hard cases addressed by adding our data augmentation.

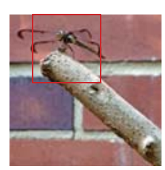


## Multi-scale Training

- Same crop size, different resolution
  - Fixed-size 224*224
- Downsized training images
  - Reduces computational costs
  - But not for state-of-the-art
- Different models trained by different image sizes

- High-resolution model works
  - 256x256: top-5 7.96%
  - 512x512: top-5 7.42%
- Multi-scale models are complementary
  - Fused model: 6.97%


256*256


512*512

“明查秋毫”

# Multi-scale Training

| Rank | Score | Class |
|---|---|---|
| 01 | 0.2287 | ant |
| 02 | 0.0997 | damselfly |
| 03 | 0.0570 | nematode |
| 04 | 0.0546 | chainlink fence |
| 05 | 0.0522 | long-horned beetle |
| 06 | 0.0307 | walking stick |
| 07 | 0.0287 | dragonfly |
| 08 | 0.0267 | tiger beetle |
| 09 | 0.0225 | doormat |
| 10 | 0.0198 | flute |
| 11 | 0.0198 | grey whale |
| 12 | 0.0178 | mantis |
| 13 | 0.0171 | lacewing |
| 14 | 0.0161 | radiator |
| 15 | 0.0161 | scabbard |
| 16 | 0.0157 | slide rule |
| 17 | 0.0148 | fly |
| 18 | 0.0129 | leafhopper |
| 19 | 0.0101 | cucumber |
| 20 | 0.0094 | velvet |

| Rank | Score | Class |
|---|---|---|
| 01 | 0.1026 | lacewing |
| 02 | 0.0742 | dragonfly |
| 03 | 0.0742 | damselfly |
| 04 | 0.0632 | walking stick |
| 05 | 0.0390 | long-horned beetle |
| 06 | 0.0272 | leafhopper |
| 07 | 0.0248 | nail |
| 08 | 0.0228 | grasshopper |
| 09 | 0.0191 | ant |
| 10 | 0.0151 | mantis |
| 11 | 0.0146 | fly |
| 12 | 0.0127 | hammer |
| 13 | 0.0120 | American chameleon |
| 14 | 0.0119 | gar |
| 15 | 0.0110 | chainlink fence |
| 16 | 0.0108 | padlock |
| 17 | 0.0108 | tree frog |
| 18 | 0.0105 | cicada |
| 19 | 0.0098 | screwdriver |
| 20 | 0.0096 | harvestman |

Tricycle

Washer


Backpack

Little blue heron

Dr. Ren Wu @ MPSoC

Tricycle

# Single Model Performance

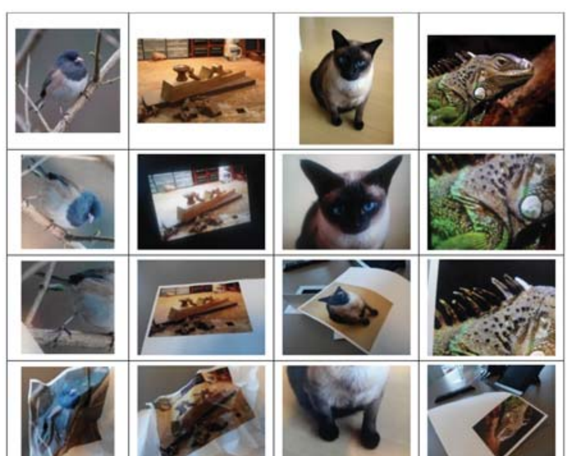
- One basic configuration has 16 layers


- The number of weights in our configuration is 212.7M
  - About 40% bigger than VGG's

| Team | Top-5 val. error |
|---|---|
| VGG | 8.0% |
| GoogLeNet | 7.89% |
| BN-Inception | 5.82% |
| MSRA, PReLU-net | 5.71% |
| **Deep Image** | **5.40%** |



# Robustness



Junco | Plane | Siamese cat | Common iguana

# Robustness

Summer tanager | Bengal | Brown creeper | English setter



| Rank | Score | Class |
|---|---|---|
| 01 | 0.6621 | jay |
| 02 | 0.0946 | European gallinule |
| 03 | 0.0628 | indigo bunting |
| 04 | 0.0456 | junco |
| 05 | 0.0101 | guenon |
| 06 | 0.0071 | Italian greyhound |
| 07 | 0.0070 | little blue heron |
| 08 | 0.0058 | bee eater |
| 09 | 0.0048 | patas |
| 10 | 0.0036 | chickadee |
| 11 | 0.0032 | macaw |
| 12 | 0.0030 | handkerchief |
| 13 | 0.0025 | coucal |
| 14 | 0.0025 | mosquito net |
| 15 | 0.0025 | water ouzel |
| 16 | 0.0024 | crayfish |
| 17 | 0.0020 | whippet |
| 18 | 0.0019 | Siberian husky |
| 19 | 0.0017 | koala |
| 20 | 0.0016 | wallaby |

| Rank | Score | Class |
|---|---|---|
| 01 | 0.3687 | king crab |
| 02 | 0.2159 | hotdog |
| 03 | 0.1031 | pizza |
| 04 | 0.0575 | burrito |
| 05 | 0.0406 | bagel |
| 06 | 0.0307 | Dungeness crab |
| 07 | 0.0234 | crayfish |
| 08 | 0.0133 | goldfish |
| 09 | 0.0114 | American lobster |
| 10 | 0.0114 | potpie |
| 11 | 0.0094 | strawberry |
| 12 | 0.0089 | carbonara |
| 13 | 0.0085 | plate |
| 14 | 0.0079 | ice cream |
| 15 | 0.0065 | orange |
| 16 | 0.0064 | butcher shop |
| 17 | 0.0063 | corn |
| 18 | 0.0062 | butternut squash |
| 19 | 0.0046 | sea cucumber |
| 20 | 0.0045 | mashed potato |

Dr. Ren Wu @ MPSoC

# Benchmark Results

| Benchmark | Measurement | Previous Best | Deep Image |
|---|---|---|---|
| Caltech CUB200-2011 | Top-1 accuracy | 85.4% | **85.6%** |
| Oxford Flowers | Top-1 accuracy | 95.3% | **98.7%** |
| Oxford-IIIT Pets | Top-1 accuracy | 91.6% | **93.1%** |
| FGVC-aircraft | Top-1 accuracy | 81.5% | **85.2%** |
| MIT Indoor Scene | Top-1 accuracy | 81.1% | **82.4%** |
| ImageNet ILSVRC | Top-5 error | 4.82% | **4.54%** |

Dr. Ren Wu @ MPSoC

## ImageNet ILSVRC Results

| Team | Date | Top-5 test error |
|---|---|---|
| GoogLeNet | 2014 | 6.66% |
| Deep Image | 01/12/2015 | 5.98% |
| Deep Image | 02/05/2015 | 5.33% |
| Microsoft | 02/05/2015 | 4.94% |
| Google | 03/02/2015 | 4.82% |
| Deep Image | 05/10/2015 | **4.58%** |



## Major Differentiators

- Customized built **supercomputer** dedicated for DL
- **Simple**, **scalable** algorithm + **Fully optimized** software stack
- **Larger** models
- **More Aggressive** data augmentation
- Multi-scale, include **high-resolution** images

**Scalability + Insights and push for extreme**

## Deep Learning: Deployment

**Big data + Deep learning + High performance computing =**
**Intelligence**

**Big data + Deep learning + Heterogeneous computing =**
**Success**

Dr. Ren Wu @ MPSoC

## Owl of Minwa (敏鹗)

Models trained by supercomputers
Trained models will be deployed in many ways
data centers (cloud), smartphones, and even wearables and IoTs

OpenCL based, light weight and high performance

**DNNs everywhere !**

Supercomputers

Datacenters

Tablets, smartphones


**knowledge, wisdom, perspicacity and erudition**

Dr. Ren Wu @ MPSoC

## DNNs Everywhere

| Supercomputers | Datacenters | Tablets, smartphones | Wearable devices IoTs |
|---|---|---|---|
|  |  |  |  |
| 1000s GPUs | 100k-1m servers | 2b (in China) | 50b in 2020? |

Supercomputer used for training
Trained DNNs then deployed to data centers (cloud), smartphones, and even wearables and IoTs

Dr. Ren Wu @ MPSoC

## Offline Mobile DNN App


- Image recognition on mobile device
- Real time and no connectivity needed
- directly from video stream, what you point is what you get
- Everything is done within the device
- OpenCL based, highly optimized
- Large deep neural network models
- Thousands of objects, flowers, dogs, and bags etc
- Unleashed the full potential of the device hardware
- Smart phones now, Wearables and IoTs tomorrow

Dr. Ren Wu @ MPSoC


Dr. Ren Wu @ EVS15

## Cloud Computing: What's Missing?

| Operation | Energy, pJ | Relative cost |
|---|---|---|
| 16b Int ADD | 0.06 | 1 |
| 16b Int MULT | 0.8 | 13 |
| 16b FP ADD | 0.45 | 8 |
| 16b FP MULT | 1.1 | 18 |
| 32b FP ADD | 1.0 | 17 |
| 32b FP MULT | 4.5 | 80 |
| Register File, 1kB | 0.6 | 10 |
| L1 Cache, 32kB | 3.5 | 58 |
| L2 Cache, 256kB | 30.2 | 500 |
| on-chip DRAM | 160 | 2667 |
| DRAM | 640 | 10667 |
| Wireless transfer | 60000 | 1000000 |

Bandwidth?
Latency?
and
Power consumption?

*Artem Vasilyev: CNN optimizations for embedded systems and FFT

Moving data around is expensive, **very expensive!**

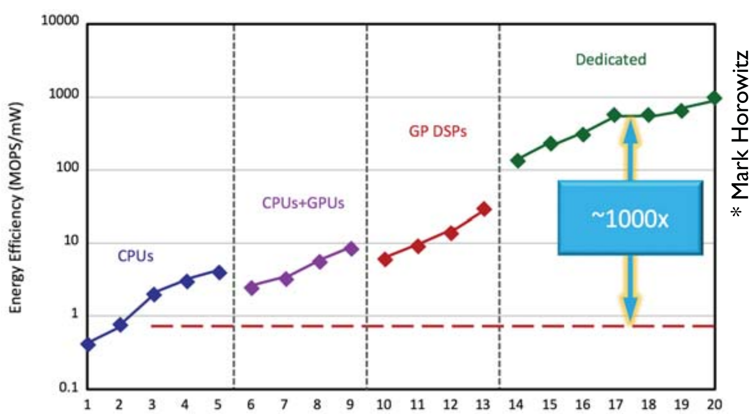
Dr. Ren Wu @ MPSoC

## Cloud Computing: What's Missing?


**How about**
**privacy?**

Dr. Ren Wu @ MPSoC

## What's Next?




Dedicated Hardware + Heterogeneous Computing

Dr. Ren Wu @ MPSoC

# Heterogeneous Computing

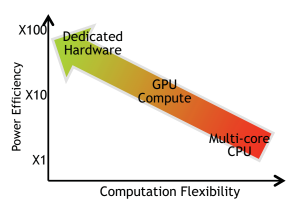
"Human mind and brain is not a single general-purpose processor but a collection of highly specialized components, each solving a different, specific problem and yet collectively making up who we are as human beings and thinkers." – Prof. Nancy Kanwisher

Dr. Ren Wu @ MPSoC

# Vision Processing Power Efficiency

- **Wearables will need 'always-on' vision**
  - With smaller thermal limit / battery than phones!
- **GPUs have x10 imaging power efficiency over CPU**
  - GPUs architected for efficient pixel handling
- **Dedicated Hardware/DSPs can be even more efficient**
  - With some loss of generality
- **Mobile SOCs have space for more transistors**
  - But can't turn on at same time = Dark Silicon
  - Can integrate more gates 'for free' if careful how and when they are used

**Potential for dedicated sensor/vision silicon to be integrated into Mobile Processors**
**But how will they be programmed for PORTABILITY and POWER EFFICIENCY?**



## OpenCL Ecosystem

**Implementers**
Desktop/Mobile/FPGA

Apple, IBM, AMD, Intel, NVIDIA, ARM, Imagination, Qualcomm, Vivante, MediaTek, ST, Samsung, Altera, Xilinx

**SYCL**
Single Source C++ Programming

**OpenCL**
Core API and Language Specs

**SPIR**
Portable Kernel Intermediate Language

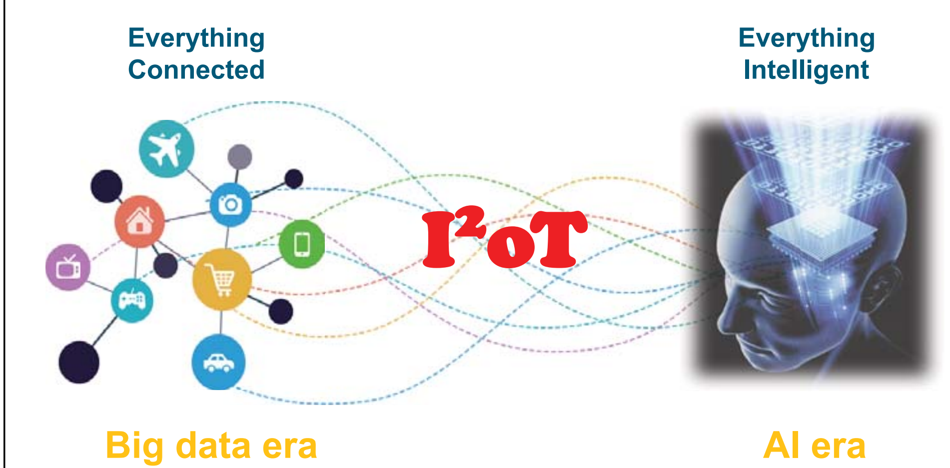
**Working Group Members**
Apps/Tools/Tests/Courseware

Adobe, codeplay, Huawei, mobica, Broadcom, cognivue, Sony, University of Bristol, Texas Instruments, vmware, Multicore Ware

KHRONOS GROUP



## Intelligent Internet of Things




Dr. Ren Wu @ MPSoC


Thank you!