Cognitive Computing:
The Next Wave of Computing Innovation

Antonio González

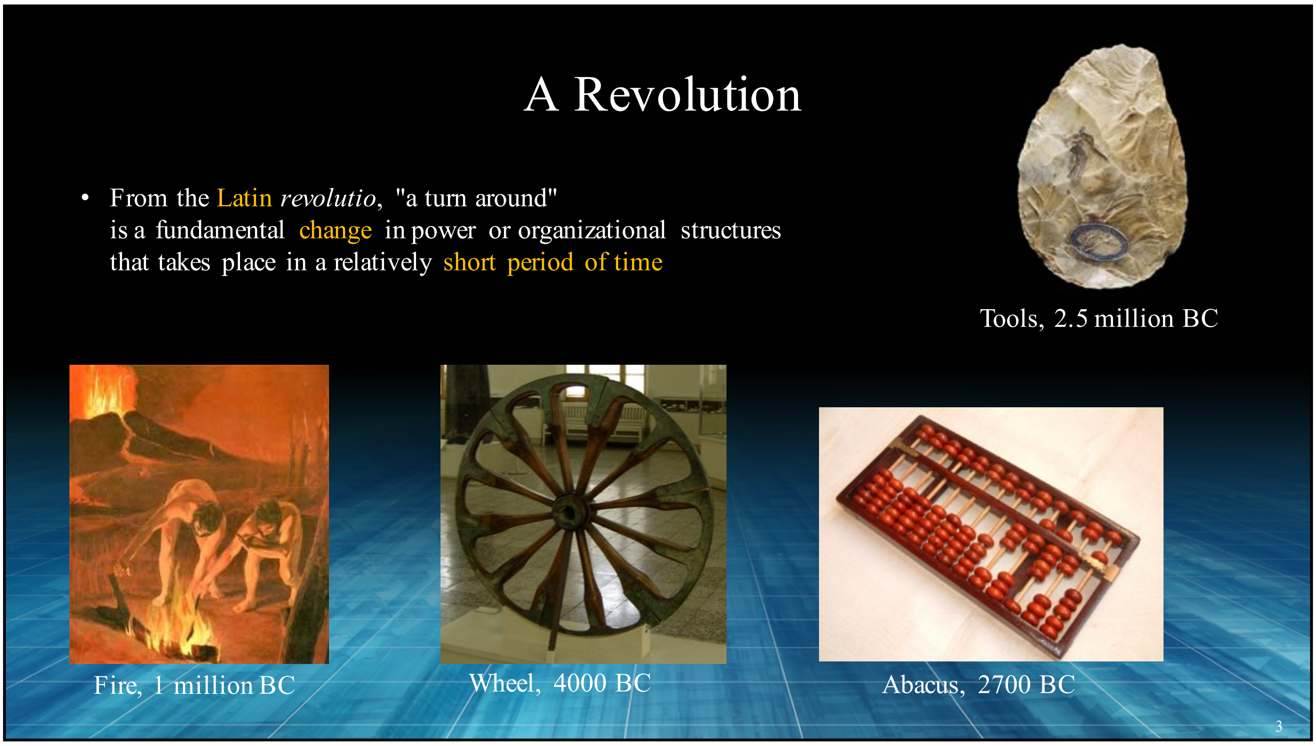
Director, ARCO Research Group
Professor, Computer Architecture Department, UPC

Facultad de Informática - Universidad Complutense de Madrid, Madrid (Spain), May 9, 2016

Agenda

- The next revolution in computing
- Key innovations to make it happen
- Concluding remarks

2



A Revolution

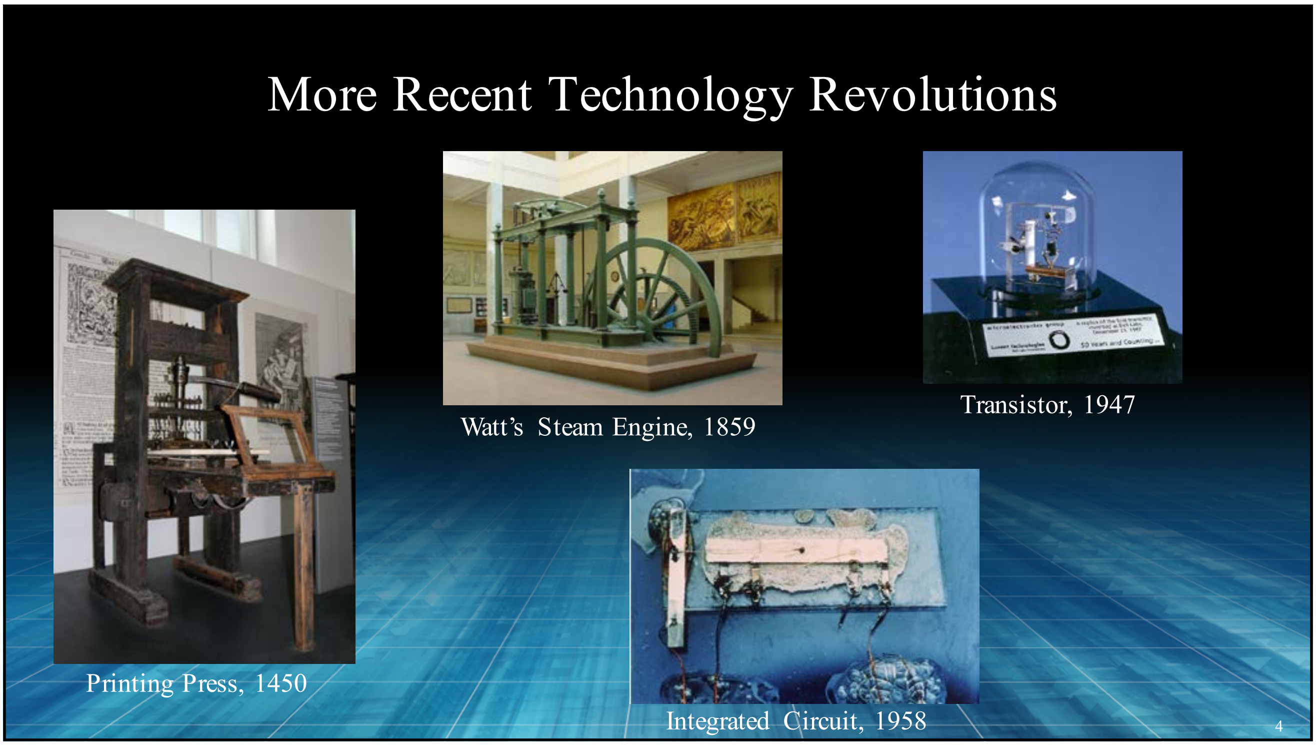
- From the Latin *revolutio*, "a turn around" is a fundamental change in power or organizational structures that takes place in a relatively short period of time

Tools, 2.5 million BC

Fire, 1 million BC

Wheel, 4000 BC

Abacus, 2700 BC



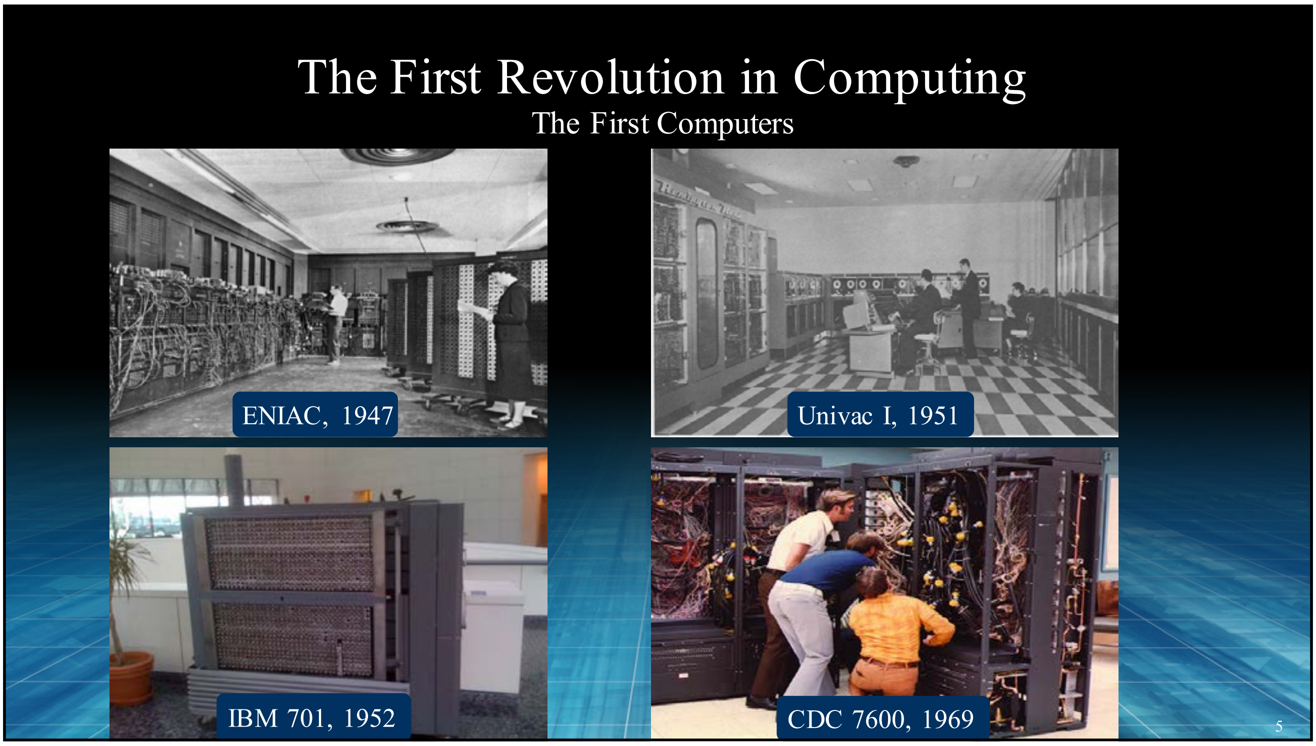
More Recent Technology Revolutions

Printing Press, 1450

Watt's Steam Engine, 1859

Transistor, 1947

Integrated Circuit, 1958



The First Revolution in Computing

The First Computers

ENIAC, 1947

Univac I, 1951

IBM 701, 1952

CDC 7600, 1969

The Second Revolution

The Personal Computers

PC

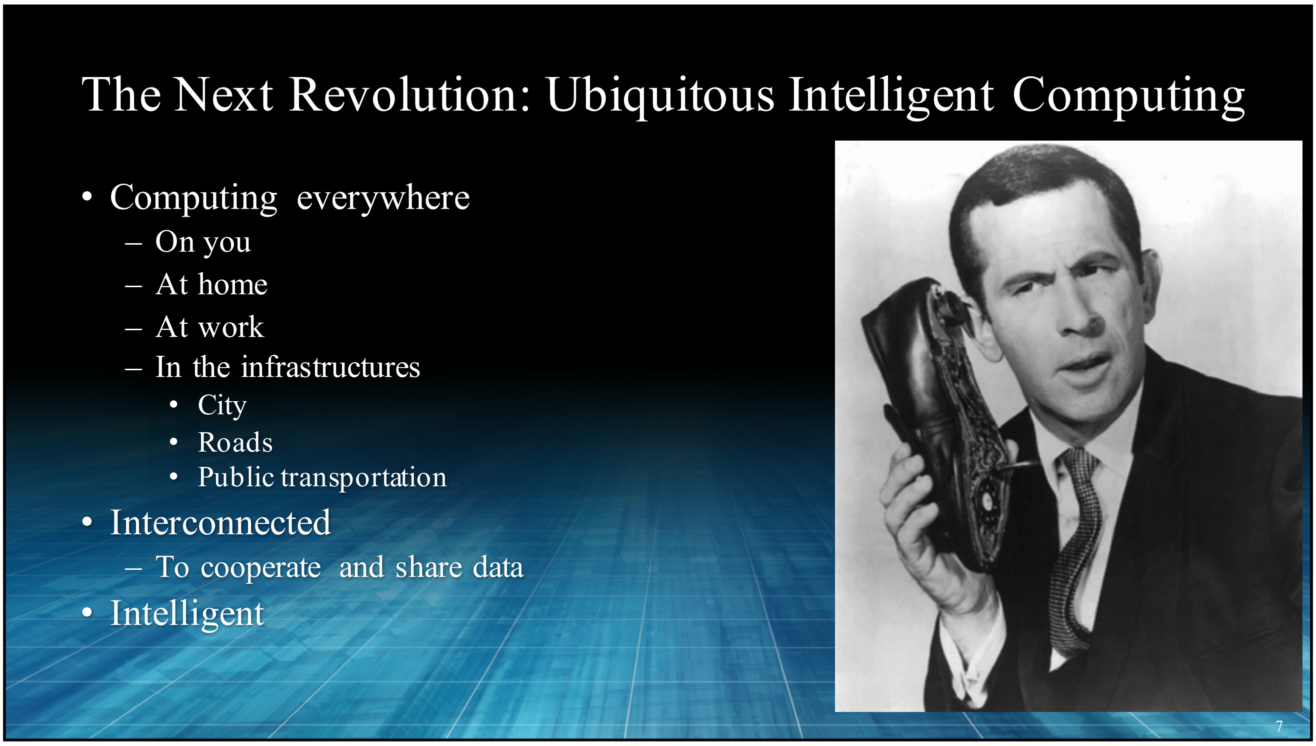
Laptop

Ultrabook

Tablet

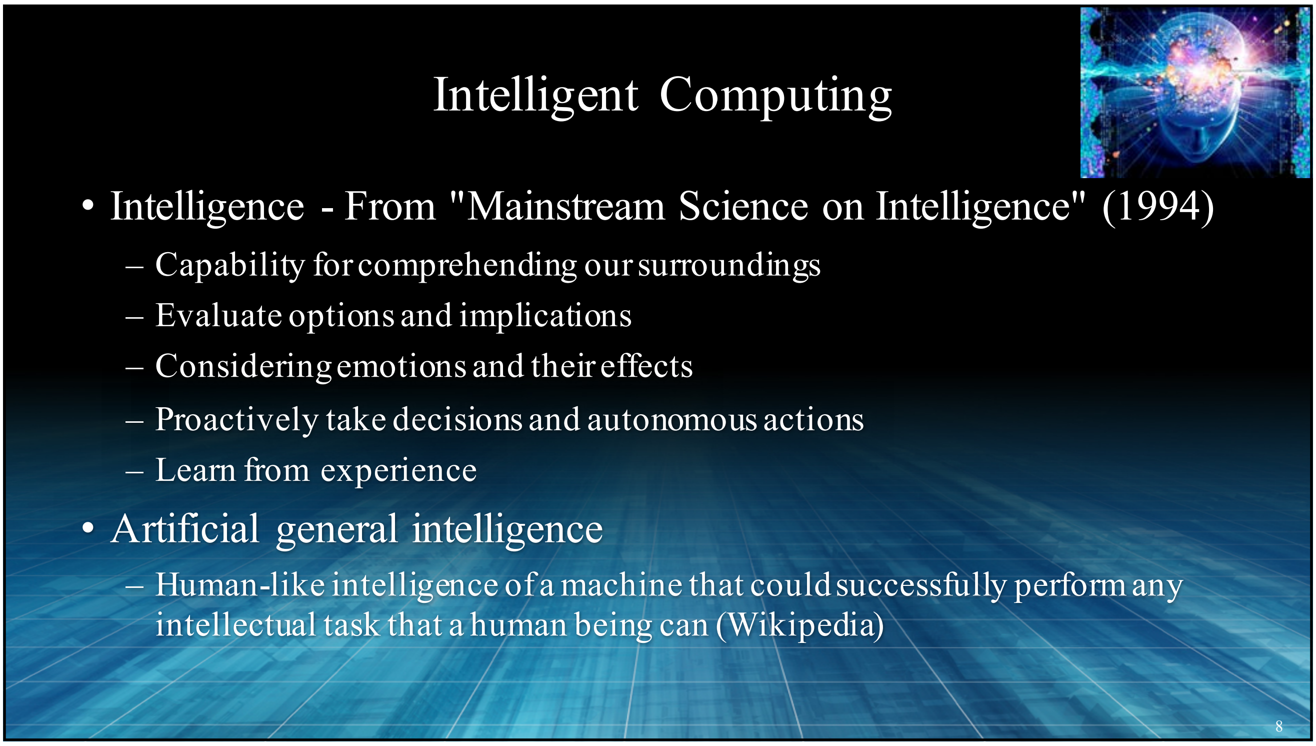
Convertible

Smartphone



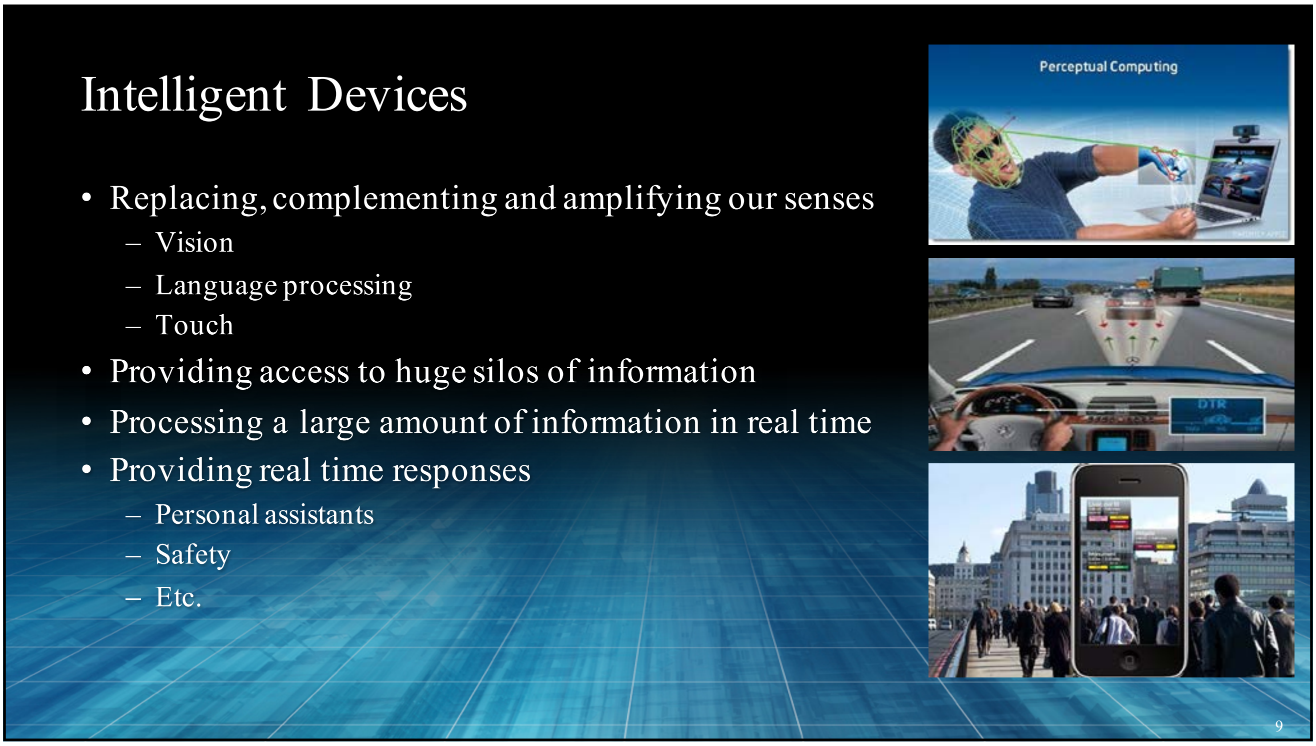
The Next Revolution: Ubiquitous Intelligent Computing

- Computing everywhere
  - On you
  - At home
  - At work
  - In the infrastructures
    - City
    - Roads
    - Public transportation
- Interconnected
  - To cooperate and share data
- Intelligent



Intelligent Computing

- Intelligence - From "Mainstream Science on Intelligence" (1994)
  - Capability for comprehending our surroundings
  - Evaluate options and implications
  - Considering emotions and their effects
  - Proactively take decisions and autonomous actions
  - Learn from experience
- Artificial general intelligence
  - Human-like intelligence of a machine that could successfully perform any intellectual task that a human being can (Wikipedia)



Intelligent Devices

- Replacing, complementing and amplifying our senses
  - Vision
  - Language processing
  - Touch
- Providing access to huge silos of information
- Processing a large amount of information in real time
- Providing real time responses
  - Personal assistants
  - Safety
  - Etc.

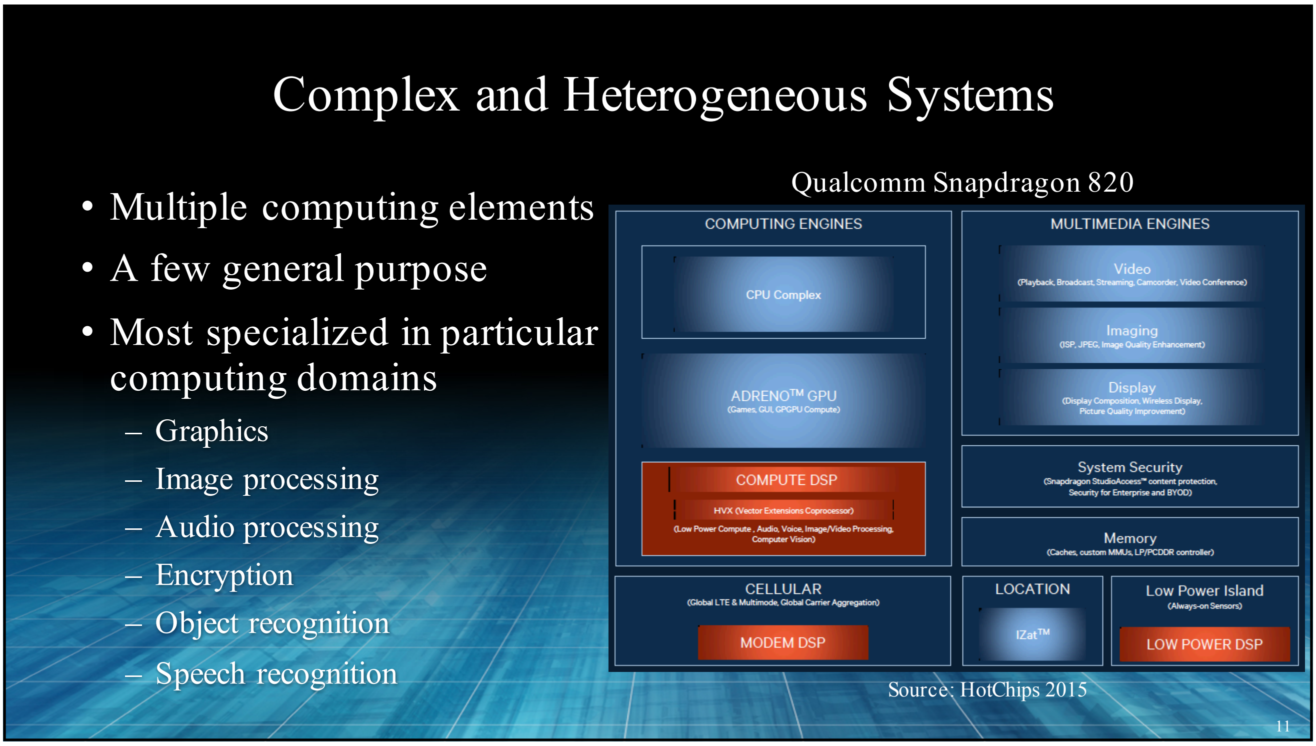
9

Very Diverse

- Worn devices
- Body sensors
- Driving devices
- Home robots
- Healthcare devices
- Energy management
- Smart consumer electronics

hello
9
47
nest
74
intel
NXP
MDV

10



Complex and Heterogeneous Systems

- Multiple computing elements
- A few general purpose
- Most specialized in particular computing domains
  - Graphics
  - Image processing
  - Audio processing
  - Encryption
  - Object recognition
  - Speech recognition

Qualcomm Snapdragon 820

COMPUTING ENGINES

CPU Complex

ADRENO™ GPU (Games, GUI, GPGPU Compute)

COMPUTE DSP

HVX (Vector Extensions Coprocessor)

(Low Power Compute , Audio, Voice, Image/Video Processing, Computer Vision)

MULTIMEDIA ENGINES

Video (Playback, Broadcast, Streaming, Camcorder, Video Conference)

Imaging (ISP, JPEG, Image Quality Enhancement)

Display (Display Composition, Wireless Display, Picture Quality Improvement)

System Security (Snapdragon StudioAccess™ content protection, Security for Enterprise and BYOD)

Memory (Caches, custom MMUs, LP/PCDDR controller)

CELLULAR (Global LTE & Multimode, Global Carrier Aggregation)

MODEM DSP

LOCATION

IZat™

Low Power Island (Always-on Sensors)

LOW POWER DSP

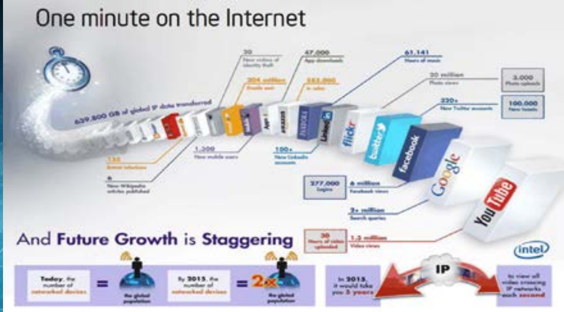
Source: HotChips 2015

Key Enabling Technologies

- Data analytics
- Device and data security
- Energy-efficient high performance

Data Analytics

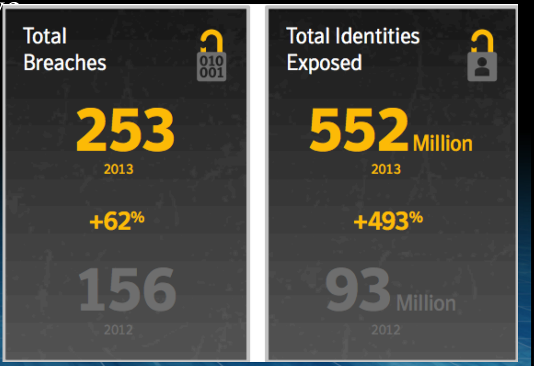
- Huge amounts of unstructured data (“big data”)
- The challenge
  - Find the useful data (a tiny percentage of this huge volume)
  - Derive useful information from data

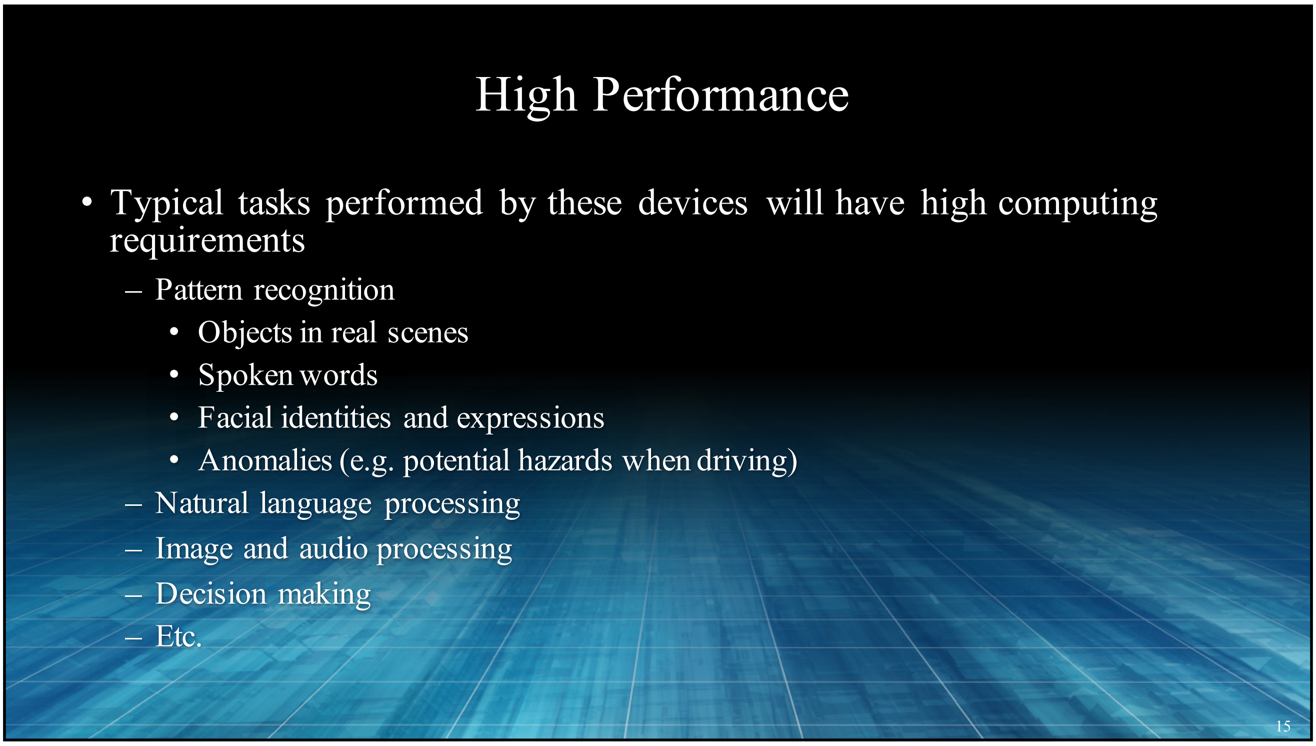


Security

- Interoperability implies accessibility
- These devices will be used for very sensitive activities
  - Private data
    - Digital wallet
    - House key
    - Personal data
  - Control systems
    - Health care
    - Car driving
    - Access control (e.g. home)
- Threats are increasing

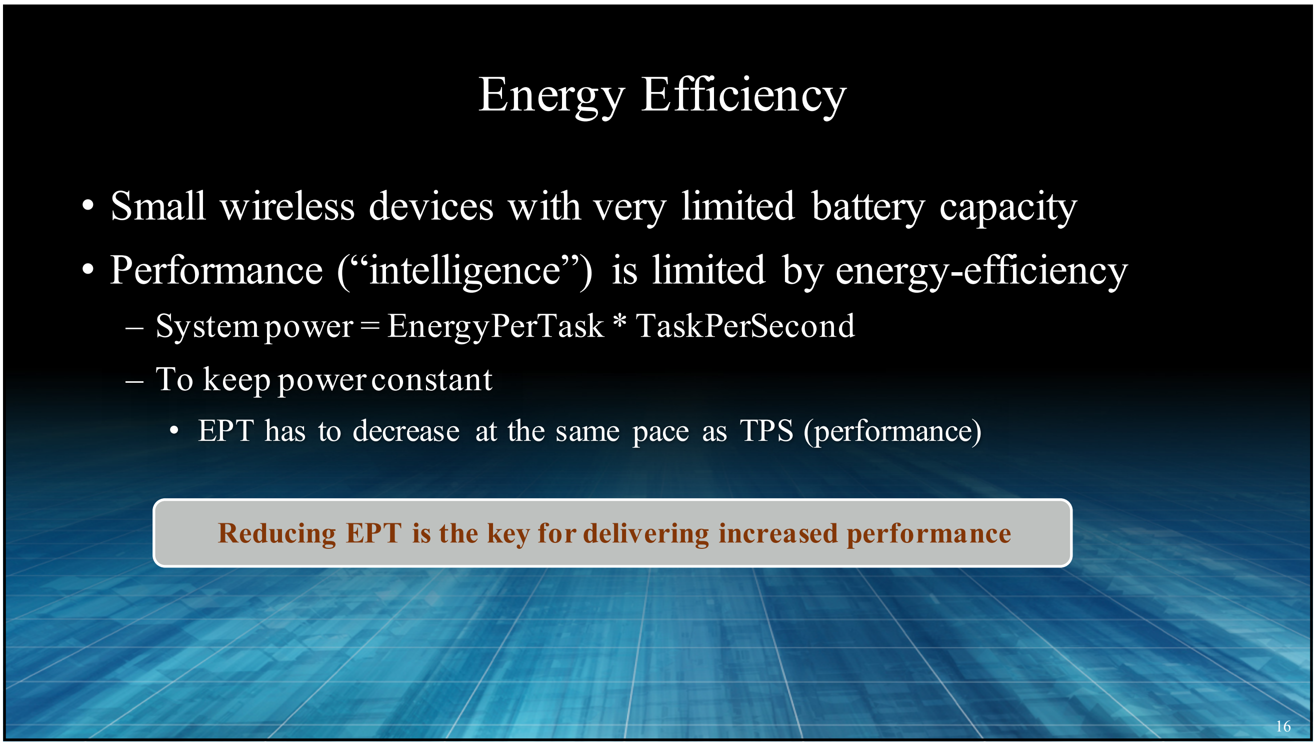
Source: Symantec





High Performance

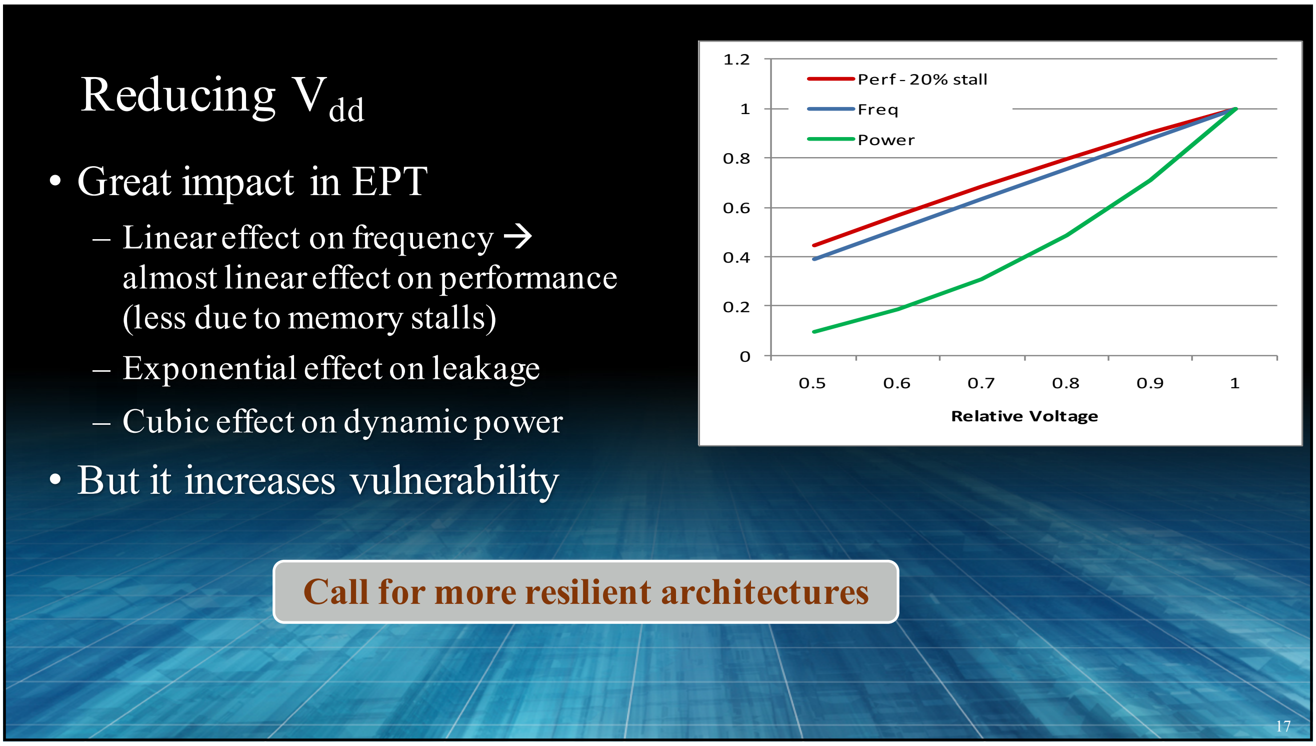
- Typical tasks performed by these devices will have high computing requirements
  - Pattern recognition
    - Objects in real scenes
    - Spoken words
    - Facial identities and expressions
    - Anomalies (e.g. potential hazards when driving)
  - Natural language processing
  - Image and audio processing
  - Decision making
  - Etc.



Energy Efficiency

- Small wireless devices with very limited battery capacity
- Performance ("intelligence") is limited by energy-efficiency
  - System power = EnergyPerTask * TaskPerSecond
  - To keep power constant
    - EPT has to decrease at the same pace as TPS (performance)

Reducing EPT is the key for delivering increased performance

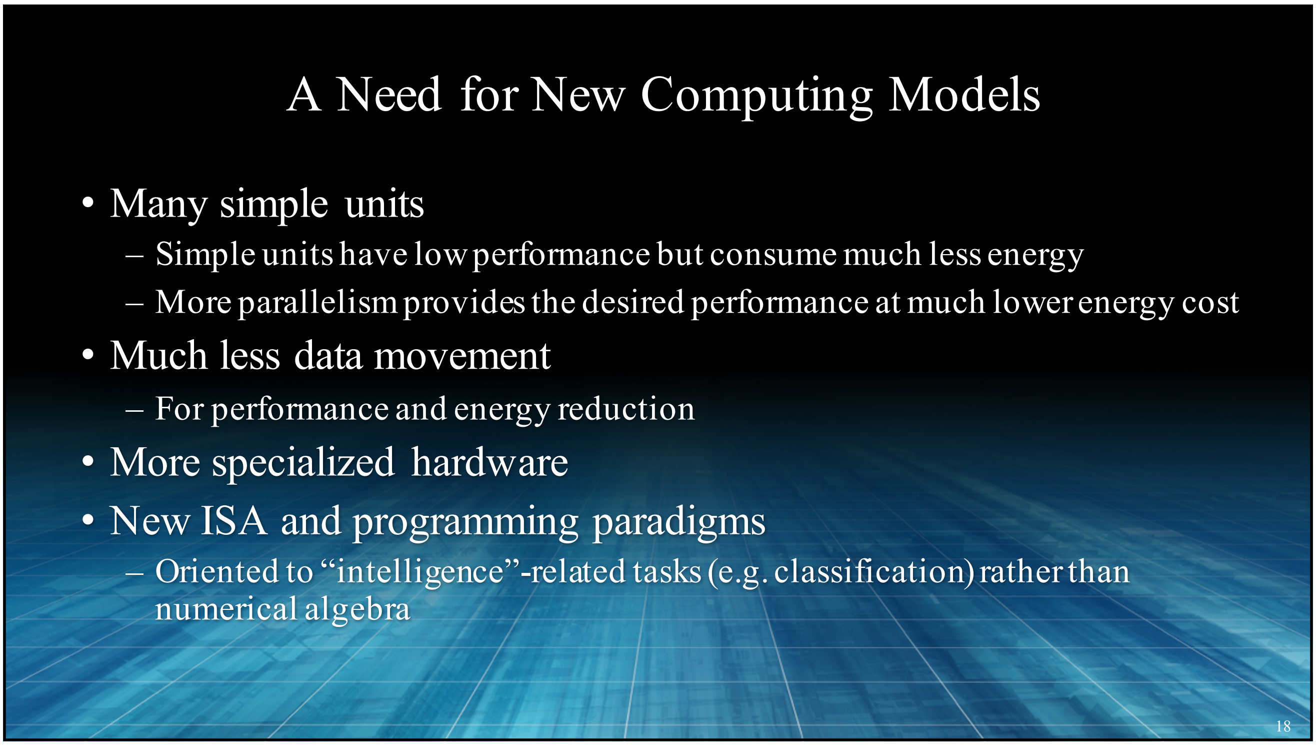


Reducing V_{dd}

- Great impact in EPT
  - Linear effect on frequency → almost linear effect on performance (less due to memory stalls)
  - Exponential effect on leakage
  - Cubic effect on dynamic power
- But it increases vulnerability

Call for more resilient architectures

17



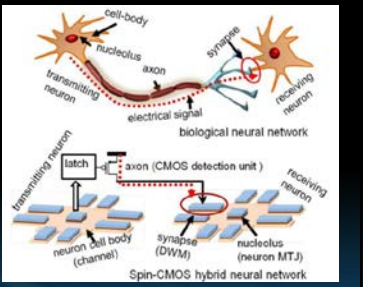
A Need for New Computing Models

- Many simple units
  - Simple units have low performance but consume much less energy
  - More parallelism provides the desired performance at much lower energy cost
- Much less data movement
  - For performance and energy reduction
- More specialized hardware
- New ISA and programming paradigms
  - Oriented to “intelligence”-related tasks (e.g. classification) rather than numerical algebra

18

Example: Brain-Inspired Computing

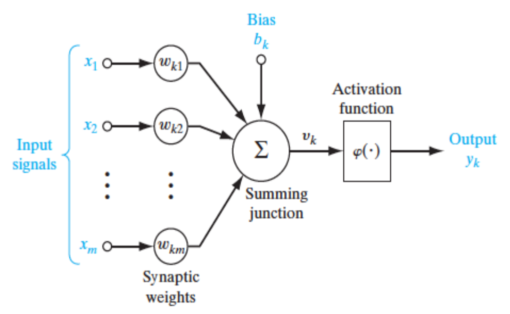
- Human brain is very good at some of these intelligence-related tasks
  - E.g. object recognition
- Human brain uses a very different computing model with many good properties
  - Composed of many simple units
  - Highly parallel
  - Fault tolerant
  - With a very different programming paradigm: learning



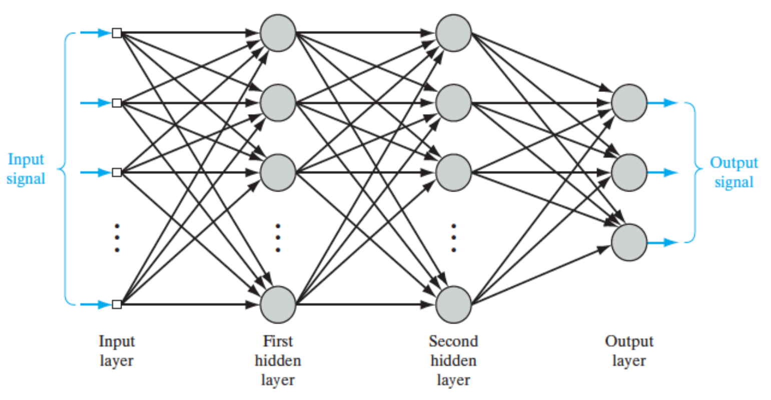
M. Sharad, C. Augustine, G. Panagopoulos, K. Roy, "Spin-Based Neuron Model with Domain Wall Magnets as Synapse," IEEE Transactions on Nanotechnology, 20

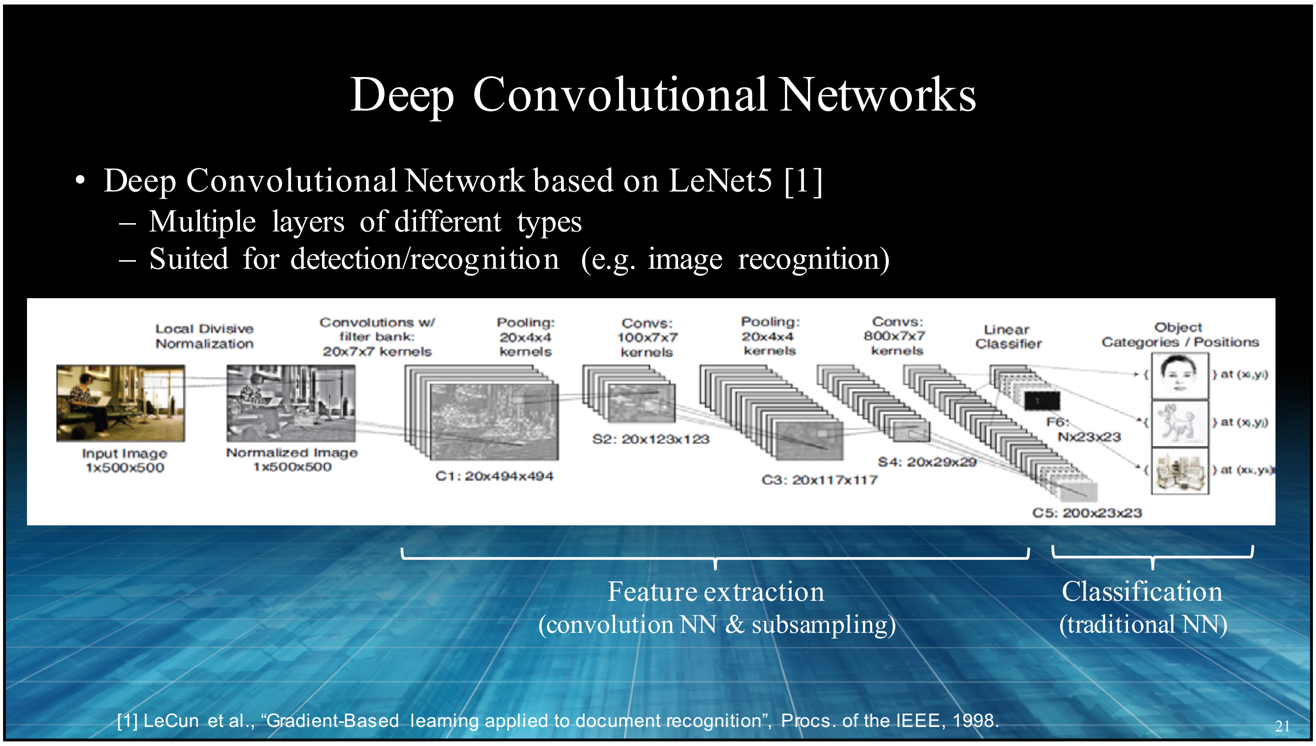
Example of an Architecture

A neuron



A feed-forward neural network





Deep Convolutional Networks

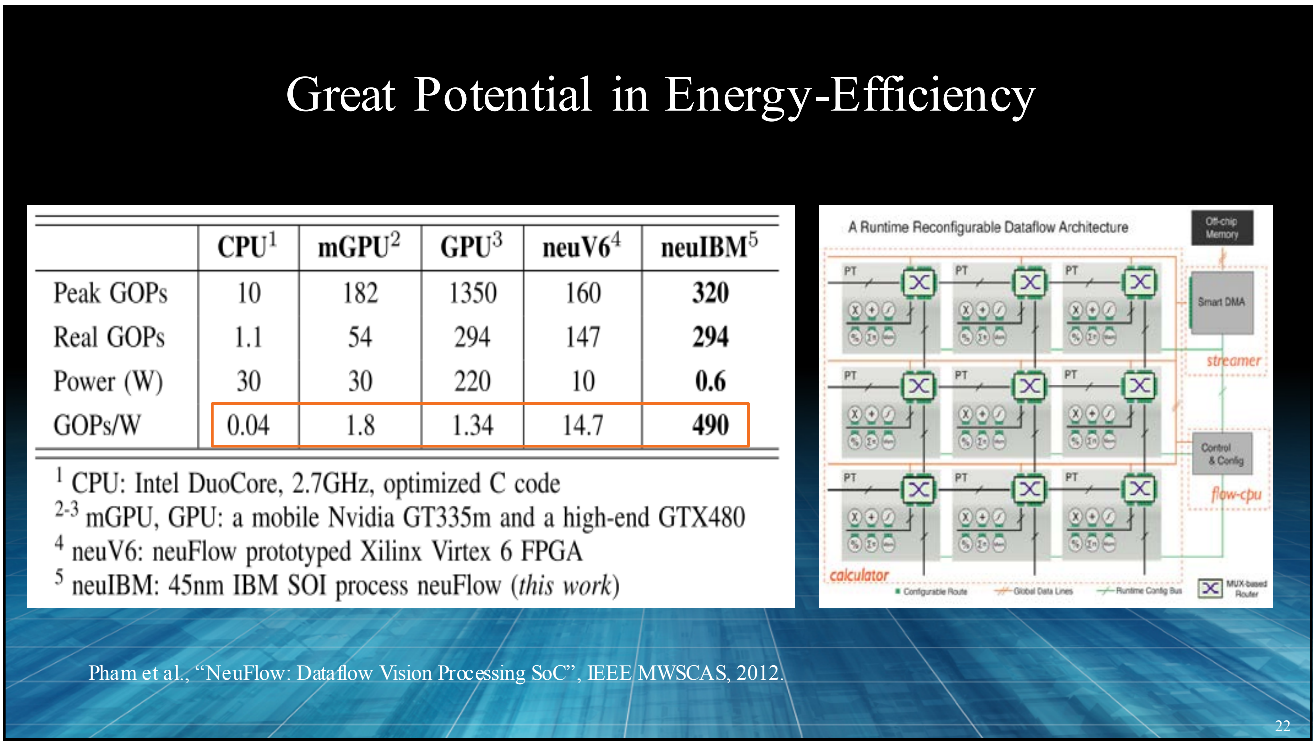
- Deep Convolutional Network based on LeNet5 [1]
  - Multiple layers of different types
  - Suited for detection/recognition (e.g. image recognition)

Feature extraction
(convolution NN & subsampling)

Classification
(traditional NN)

[1] LeCun et al., “Gradient-Based learning applied to document recognition”, Procs. of the IEEE, 1998.

21



Great Potential in Energy-Efficiency

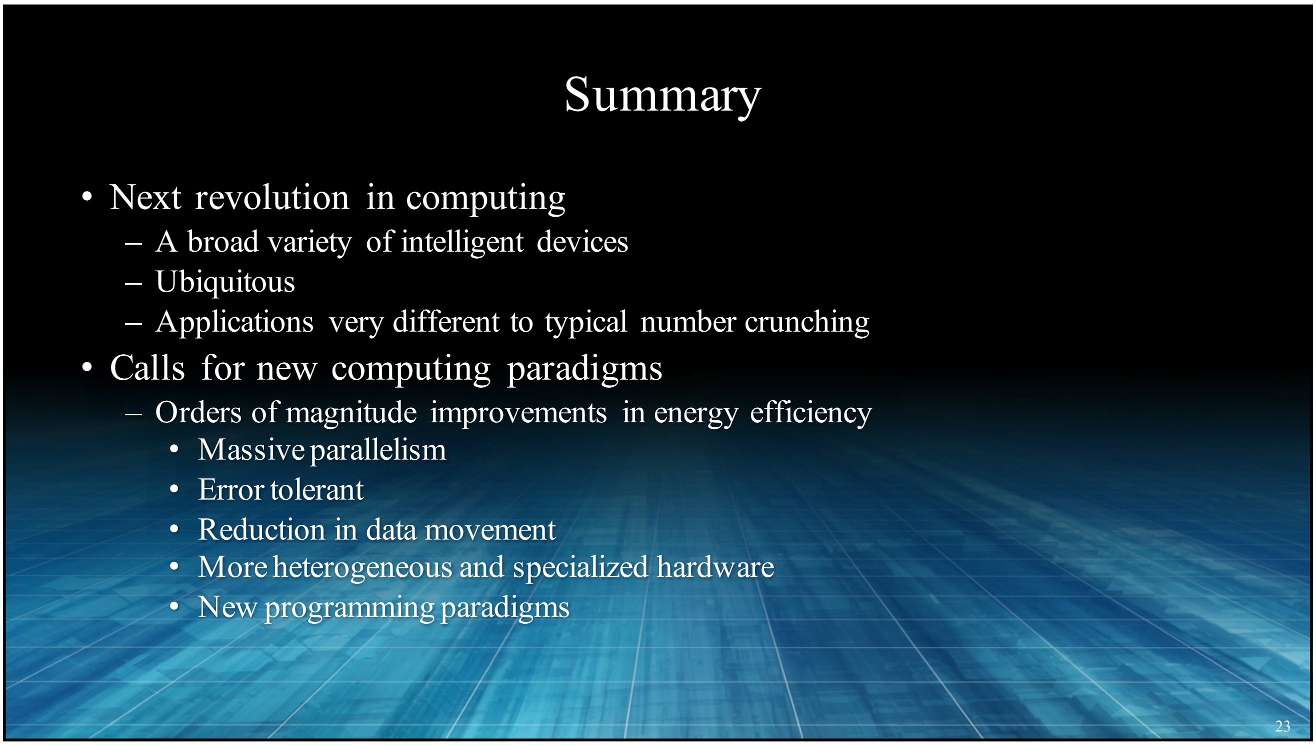
	CPU[1]	mGPU[2]	GPU[3]	neuV6[4]	neuIBM[5]
Peak GOPs	10	182	1350	160	**320**
Real GOPs	1.1	54	294	147	**294**
Power (W)	30	30	220	10	**0.6**
GOPs/W	0.04	1.8	1.34	14.7	**490**

[1] CPU: Intel DuoCore, 2.7GHz, optimized C code
[2-3] mGPU, GPU: a mobile Nvidia GT335m and a high-end GTX480
[4] neuV6: neuFlow prototyped Xilinx Virtex 6 FPGA
[5] neuIBM: 45nm IBM SOI process neuFlow (*this work*)

A Runtime Reconfigurable Dataflow Architecture

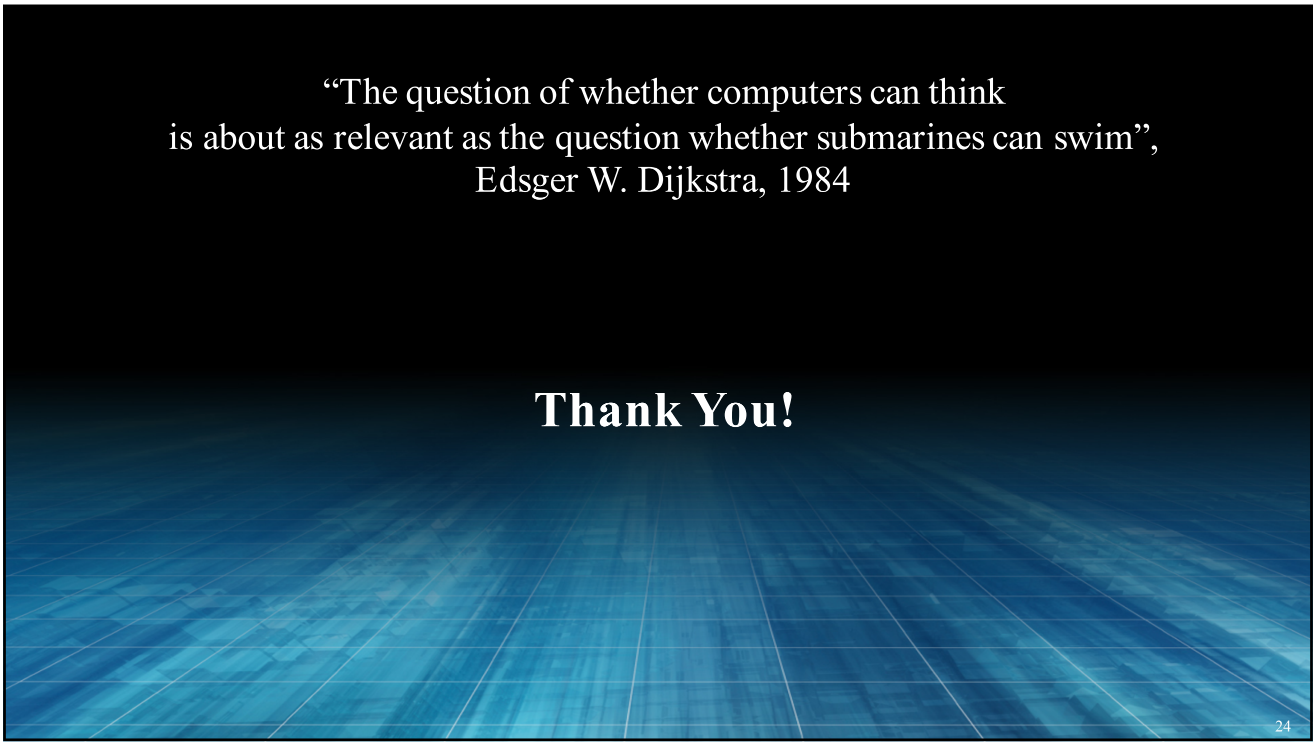
Pham et al., “NeuFlow: Dataflow Vision Processing SoC”, IEEE MWSCAS, 2012.

22



Summary

- Next revolution in computing
  - A broad variety of intelligent devices
  - Ubiquitous
  - Applications very different to typical number crunching
- Calls for new computing paradigms
  - Orders of magnitude improvements in energy efficiency
    - Massive parallelism
    - Error tolerant
    - Reduction in data movement
    - More heterogeneous and specialized hardware
    - New programming paradigms



"The question of whether computers can think
is about as relevant as the question whether submarines can swim",
Edsger W. Dijkstra, 1984

Thank You!