Future of Electronics in the AI Era

Dr. Alessandro Curioni
IBM Fellow, Vice President, Europe and Director, IBM Research – Zurich
@Ale_Curioni
25 gigabytes of data per hour is generated by a connected car.

90% of cars will be connected by 2020.

2.5 quintillion bytes of data generated daily by connected machines.

80 million wearable health devices will be available by 2017.

153 exabytes of healthcare data generated by devices in 2013.

Increasing to 2,314 exabytes in 2020.

25 gigabytes of data per hour is generated by a connected car.

90% of cars will be connected by 2020.

2.5 quintillion bytes of data generated daily by connected machines.

80 million wearable health devices will be available by 2017.

153 exabytes of healthcare data generated by devices in 2013.

Increasing to 2,314 exabytes in 2020.

There will be 28 times more sensor-enabled devices than people by the year 2020.

1.7 megabytes of data per second generated by every human being on the planet by 2020.
1971
Moore’s Law

1995
Metcalfe’s Law

Today
Watson’s Law
AI impacts all industries
Helping China reduce carbon intensity with AI + IoT

1 billion people touched by benefits of this project

367 cities in China have deployed IoT sensors linked with AI

34.3 exabytes of data ingested per day
Tracking the Progression of Chronic Obstructive Pulmonary Disease (COPD)

1 in 3 leading cause of death by 2030

90% of deaths by 2030 in low and middle income countries

$90 billion expected cost of medical care for adults with COPD
AI foundations

Data

Algorithms

Compute
AI foundations

AI Workloads

Networks and Algorithms

Data

Algorithms

Compute

End to End Software & Hardware Optimization for AI

Heterogeneous Integration

New Architecture & Materials for AI
Narrow AI: Initial Value
Creation

2010 2015

AI learns to solve specific tasks, or focuses on **individual domains** or modalities primarily using **human-curated**, **training data sets** & **manually-crafted architectures**
AI acquires knowledge by reading, discussion, observation, experiments. Broad transfer of knowledge across tasks. Cross-domain reasoning is common. Broad autonomy within human-managed teams.
AI selectively acquires and integrates knowledge from **multiple modalities and sources**, including interaction; develops and **retains skills that it adapts and combines** to complete new tasks; learning is adaptive, using automatically-constructed architectures.
Advancing Broad AI

Signal Comprehension: From video and text to rich human perception

“A green bird sitting on top of a bowl”

Learning and Reasoning: From scalable machine learning to making a case

Interaction: Understanding language, tone, emotion and context
AI today is where semiconductors were in the 1960s

Pioneering phase (1960s)
- New material: silicon
- New component: transistor
- New process: manual circuit design
- New infrastructure: custom made computing systems

Scaling phase (ensuing 50 years)
- Incredible range of chemistry extensions around silicon
- Constant innovations in device structure while undergoing many orders of magnitude of miniaturization
- Absolutely essential advances in logic synthesis, design automation and process integration
- Massive scale out occurring over many waves, from early glasshouse systems to planetary-scale cloud
### CMOS Scaling Supports AI…

<table>
<thead>
<tr>
<th>POWER8 Family</th>
<th>POWER9 Family</th>
<th>7nm Bulk FinFET</th>
<th>5nm: New Device Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z13 22nm</td>
<td>Z14 14nm</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Si:C**
- **embedded Stressors w/ eDRAM**
- **SOI FinFET with RMG and eDRAM**
- **Bulk FinFET and EUV**
- **NanoSheet and EUV**

...But Requires Innovations Across the Stack

Heterogeneous integration (3D)

Systems and Architecture
Processors specifically designed for the AI era

**POWER9**

- **4x**
  Threads per core vs x86

- **9.5x**
  Up to 9.5x more I/O bandwidth than x86

- **2.6x**
  More RAM possible vs x86

- **1st**
  CPU to deliver PCIe gen4

- **Enhanced on-chip acceleration**

- **Nvidia NVLink 2.0**: High bandwidth

- **CAPI 2.0**: Coherent accelerator and storage attach (PCIe G4)

- **New CAPI**: Improved latency and bandwidth, open interface
Summit: World’s Most Powerful Supercomputer

- 200,000 trillion calculations per second
- 9216 Power 9 processors
- 27,648 GPUs

Move Over, China: U.S. Is Again Home to World’s Speediest Supercomputer

IBM and the DoE launch the world’s fastest supercomputer
World’s smallest computer

The world’s smallest computer is an edge device that can monitor, analyze, communicate, and even act on data.

1mm X 1mm computer size

10 cents expected manufacturing cost

Solar cell allows for standalone power
The world’s smallest computer is an edge device that can monitor, analyze, communicate, and even act on data. It measures 1mm X 1mm and is expected to have a manufacturing cost of 10 cents. A solar cell allows for standalone power. IBM Research © 2018 IBM Corporation.

**IoT Crypto Anchors**
Distributed Deep Learning

100s of servers with GPUs
scale of the computational infrastructure enabled by IBM’s communication library for Distributed Deep Learning training

95%
scaling efficiency achieved by IBM @ 256 P100 GPUs

+4%
increase in image recognition accuracy over previous best result
Deep Learning Performance/Watt

Current Industry Trend

Trend: x 2.5 / year

Near-term accelerators in the industry are based on existing CMOS-based technologies
- GPU Roadmap
- Startups (ASICs, FPGAs)

Another roadmap running out of steam?
AI Accelerators
Extending the 2.5x / year through 2025

Beyond Exact Computing
Beyond Digital
Beyond Classical

Deep Learning Performance/Watt
IBM Research Projection

Trend Extend through 2025:  x 2.5 / year
Approximate Computing

- Data Sampling
- Reduce Communication Load
- Reduced Precision Computation
Approximate Computing

- 64 and 32 bit floating point arithmetic is overkill for DNN training and inference
- 16 bit formats shown to be sufficient for wide array of Deep Learning tasks
- Cores with 16 bit precision 4X smaller than cores with 32 bit precision

Reduced Precision Computation

- Trade numerical precision for computational efficiency
- Algorithmic improvements to retain model accuracy

S. Gupta et al, Deep Learning with Limited Numerical Precision, ICML, '15
Approximate Computing

Coming soon..
“A Scalable Multi-TeraOPS Deep Learning Processor Core for AI Training and Inference”
Fleischer et al
VLSI 2018 (#C4.2)

- Programmable architecture and custom ISA
- Deep learning accelerator for training and inferencing

Scratchpad Example

- High-performance, multi-TOPS
- >90% utilization across the range of neural network topologies
Beyond Digital Computing

Analog Computing

Neuromorphic computing using non-volatile memory

Digital
- Precision: Fixed
- Accuracy: Exact

Analog
- Precision: Infinite
- Accuracy: Approximate
Removing the “von Neumann Bottleneck”
Go beyond von Neumann Computing

Minimize the time and distance to memory access

Neuromorphic computing

Merolla *et al.*, *Science*, 2014

Computational memory

Borghetti *et al.*, *Nature*, 2010
New devices for machine learning

Translating neural network structure and weights

Source: arxiv.org/abs/1603.07341

Resistive Processing Unit (RPU)

Arrays, using analog memory elements with variable resistance to store weights
Computational Memory for Deep Learning

Exploit the crystallization dynamics of Phase Change Materials

- Synaptic weights stored in computational memory
- The matrix-vector multiplications performed in place with low precision
- The desired weight updates are accumulated in high precision

Nandakumar et al., arXiv:1712.01192, 2017
"Mixing" digital computing with in-memory computing

Applications: Optimization, machine learning, deep learning

“The key breakthrough of Le Gallo and colleagues is to circumvent the problems caused by device variability by combining, in a seamless fashion, in-memory and conventional processing.”

Prof. C. David Wright, University of Exeter
17 April 2018

Hybrid design: the computational memory unit performs the bulk of the computational tasks, whereas the von Neumann machine implements a method to iteratively improve or refine the accuracy of the solution.

Optimize RPU for symmetric switching

Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations

Tayfun Gokmen * and Yuri Vlasov †
IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

TABLE 2 | Summary of RPU device specifications.

<table>
<thead>
<tr>
<th>Specs</th>
<th>Parameter</th>
<th>Value</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulse duration</td>
<td></td>
<td>1 ns</td>
<td></td>
</tr>
<tr>
<td>Operating voltage</td>
<td>±V_S</td>
<td>1 V</td>
<td></td>
</tr>
<tr>
<td>Maximum device area</td>
<td></td>
<td>0.04 μm²</td>
<td></td>
</tr>
<tr>
<td>Average device resistance</td>
<td>R_device</td>
<td>24 MΩ</td>
<td>7 MΩ</td>
</tr>
<tr>
<td>Maximum device resistance</td>
<td>max (g_f)</td>
<td>112 MΩ</td>
<td>7 MΩ</td>
</tr>
<tr>
<td>Minimum device resistance</td>
<td>min (g_f)</td>
<td>14 MΩ</td>
<td>7 MΩ</td>
</tr>
<tr>
<td>Resistance on/off ratio</td>
<td>max (g_f)/min (g_f)</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Resistance change at ±V_S</td>
<td>Δg^+_min</td>
<td>100 KΩ</td>
<td>30 KΩ</td>
</tr>
<tr>
<td>Resistance change at ±V_S/2</td>
<td>Δg^-_min</td>
<td>10 KΩ</td>
<td></td>
</tr>
<tr>
<td>Storage capacity</td>
<td>(max (g_f) - min (g_f))/Δg^-_min</td>
<td>1000 levels</td>
<td></td>
</tr>
<tr>
<td>Device up/down asymmetry*</td>
<td>Δg^+_min/Δg^-_min</td>
<td>1.05</td>
<td>2%</td>
</tr>
</tbody>
</table>

Note that these numbers are derived from the radar diagram in Figure 4A and correspond to the shaded area. *Global asymmetry in up/down responses can be to a large extend compensated by proper adjustment of pulse widths and/or pulse amplitude.
Optimize RPU for symmetric switching

Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations

Tayfun Gokmen* and Yuri Vlassov
IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

doi: 10.3389/fnins.2016.00333

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Summary of RPU device specifications.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specs</td>
<td>Parameter</td>
</tr>
<tr>
<td>Pulse duration</td>
<td>$V_s$</td>
</tr>
<tr>
<td>Operating voltage</td>
<td>$V$</td>
</tr>
<tr>
<td>Maximum device area</td>
<td>$A_{device}$</td>
</tr>
<tr>
<td>Average device resistance</td>
<td>$R_{device}$</td>
</tr>
<tr>
<td>Maximum device resistance</td>
<td>$\max(g_l)$</td>
</tr>
<tr>
<td>Minimum device resistance</td>
<td>$\min(g_l)$</td>
</tr>
<tr>
<td>Resistance on/off ratio</td>
<td>$\frac{\max(g_l)}{\min(g_l)}$</td>
</tr>
<tr>
<td>Resistance change at $V_s/2$</td>
<td>$\Delta g_{min}^+$</td>
</tr>
<tr>
<td>Resistance change at $V_s$</td>
<td>$\Delta g_{min}^-$</td>
</tr>
<tr>
<td>Storage capacity</td>
<td>$\left(\frac{\max(g_l) - \min(g_l)}{\Delta g_{min}^+}\right) / \Delta g_{min}^-$</td>
</tr>
<tr>
<td>Device up/down asymmetry*</td>
<td>$\frac{\Delta g_{min}^+}{\Delta g_{min}^-}$</td>
</tr>
</tbody>
</table>

Note that these numbers are derived from the radar diagram in Figure 4A and correspond to the shaded area. *Global asymmetry in up/down responses can be to a large extend compensated by proper adjustment of pulse widths and/or pulse amplitude.
Optimize RPU for symmetric switching

By exploring new materials and devices

- Invert polarity of pulse
- Mature materials and devices today
- Ideal RPU

**Graphs:**
- Conductivity (inverse of resistance) vs. # of voltage pulses
- Conductance (S) vs. Pulse #
  - IBM Research prototype:
    - Pulse Up
    - Pulse Down
  - Data points showing conductivity changes with pulse number and polarity.

**Legend:**
- Blue dots: Mature materials and devices today
- Red dots: Ideal RPU
Beyond Classical Computing

Quantum Computing

- Optimization
- Chemistry
- Machine Learning

Hard Problems for Classical Computers

Quantum Possible

Easy Problems
The power of quantum computing

**Classical Computers**

One-to-one relationship between number of transistors and processing power

Potential power doubles when you double the number of transistors

**Quantum Computers**

Rough equivalent of a transistor is a quantum bit, or a qubit

Potential power doubles when you add one additional qubit

~ $N$

~ $2^N$
Forget the qubit race, it's all about quantum volume!

1
Qubits added: 0
Error rate decrease: 10x
Quantum volume increase: 500x

2
Qubits added: 50
Error rate decrease: 0x
Quantum volume increase: 0x
Future Quantum Applications

- Cryptography & Security
- Designing Drugs & New Materials
- Machine Learning
- Searching Big Data
IBM Q Experience: Access a Quantum Computer in the Cloud

- 85,000 users
- 143 countries
- 4.4 million executions
- +60 scientific papers
- 1,500 universities
- 300 high schools
Summary
Artificial intelligence (AI) will rapidly evolve from narrow capabilities to broader intelligence with integrated knowledge and adaptive learning.

AI will permeate every discipline and all industries from finance to manufacturing to healthcare boosting productivity and enabling brand new opportunities.
Transform Industry Though AI
AI deployed at scale in enterprise: Interpretability/ explainability of AI, Learning from small datasets, federated learning for privacy, bias-free

Exponential Improvements in AI Computation
Execute roadmap of exponential improvement in power-performance for AI training & inferencing.

Apply Quantum to solve intractable AI problems

Inclusion
"If we truly want unbiased technology, we need to make sure everyone has an equal opportunity to participate in its creation" - Ginni Rometty
“Does AI Have a Hardware Problem?”
Editorial | 17 April 2018

AI has a hardware solution
Grand Challenges

Transform Industry Though AI
AI deployed at scale in enterprise: Interpretability/ explainability of AI, Learning from small datasets, federated learning for privacy, bias-free

Exponential Improvements in AI Computation
Execute roadmap of exponential improvement in power-performance for AI training & inferencing.

Apply Quantum to solve intractable AI problems

Inclusion
"If we truly want unbiased technology, we need to make sure everyone has an equal opportunity to participate in its creation" - Ginni Rometty
AI for Semiconductors

Industry 4.0
Design Automation
Industry 4.0 for Semiconductors

Semiconductor Fab (moderate size) Data

- Many Terabytes per hour
- Over 5 Million transactions per day from more than 500 tools
- More than 500,000 sensors
- More than 400 Terabytes of storage supporting real-time applications
AI for Industry 4.0
Some use cases for the Semiconductor Industry

Value Chain

- predict quality in the field
- trace defective parts in the field to production signature & supply chain
- limit warrantees/recall issues

Equipment

Anomaly detection example

Early Anomaly detection

+ Sparse Learning for Diagnosis

good
pre-failure
AI for design automation

Machine Learning is exploited across the Design Cycle

- Concept
- High-Level Design
- Implementation
- Test
- System Build
- In-Field Use

SynTunSys

- Use ML to automatically set design parameters
  - 60% reduction in latch-to-latch delay
  - 7% power reduction

M.M. Ziegler et al DATE 2016

- Machine Learning is exploited across the Design Cycle
- AI for design automation

Verification Cockpit

- Platform for planning, tracking, analysis, and optimization of a large scale verification project
  - 30% reduction in Hard-to-Hit (H2H) Events
  - Steer the verification to less verified areas

R. Gal et al ASP DAC tutorial 2017

- Examples from z13 processor development

Data Warehouse

- Failure Tracking
- Test Submission
- Bug Tracking
- Verification Plan
- Work Plan
- Coverage Tracking
- Test Bench
- Reporting and Analytics Engines
- Optimization Directives

Data Warehouse

- Design
- Functional Verification
- Reporting and Analytics Engines
- Optimization Directives

- 30%
What is AI?

- Artificial Intelligence
  - Machine Learning
    - Neural Networks
      - Deep Learning
Your Typical Machine Learning Workflow

Data

Data Pre-Processing

Build, Train, Optimize

Deploy & Inference

Maintain Model Accuracy

Structured & Unstructured Data

- Sensors / Devices
- Operations & Process Data
- Yield, Performance
- Quality
- Resources, Costs / Budget
- Maintenance Logs, Images, Reports

Data Pre-Processing

Train Model

Deep Learning:
- Deep neural networks
  - CNN, RNN, LSTM, ...
- Deep stacking networks

Traditional Machine Learning:
- Linear regression
- Logistic regression
- Support vector machine
- Naïve Bayes classifier
- K-means clustering
- Random forests
- Decision trees

Knowledge Graph & Inference Scoring

Category: Actors Who Direct

Clue: "A Bronx Tale"

Assess and Address Data Quality

Align DataSets

Merged Time Series

Sync Build Time-Run Time Updates

Assess and Address Data Quality

Align DataSets

Merged Time Series

Sync Build Time-Run Time Updates

Assess and Address Data Quality

Align DataSets

Merged Time Series

Sync Build Time-Run Time Updates
New Architecture & Approximate Computing

- Trade numerical precision for computational efficiency
- Algorithmic improvements to retain model accuracy

Materials Innovation: Analog Cross-point Devices

- Circumvents the von-Neumann bottleneck
- Path for 10,000x improvement in deep learning

S. Gupta et al, Deep Learning with Limited Numerical Precision, ICML, '15

The next decade of system leadership

Heterogeneous

Multi-chip Module

Computing

Data

Secure

x

(3b₁ + 2b₂)

b₁

b₂

Novel devices

Phase-change Memory

Neuromorphic Array

Carbon Nanotubes

Resistive Processing Unit

FHE MODEL

STATUS QUO MODEL

Neuromorphic Accelerators

RRAM Materials
AI foundations

Data

Algorithms

Compute
Al foundations

1900’s

Santiago Ramón y Cajal
AI foundations

1940’s

Artificial Neural Networks
Deep learning 1.0

Training
Inferencing

Data input (MNIST database)

Forward propagation
Fully Trained Network

Update weights

Backward propagation: this is a four!
Deep learning 1.0

explosion of networks

Convolutional Neural Networks (CNNs)

Alexnet

Resnet

LeNet

Source: Goodfellow 2016
Deep Learning 2.0

Learns more from less data
Adaptation to new domains
Interpretability and explainability
Deep Learning 2.0: Adaptation to New Domains

Training Across Modalities

Example: Automatic Curation of Golf Highlights

How to Measure Excitement?

CVPR 2017
"Found in Translation"
Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models

Take the Tour!