Driving Opportunities and Technical Challenges of the Next Wave of Semiconductors Devices

Karim Arabi, Ph.D.

July 4, 2016
1st Federative Event on Design for Robustness (FEDfRo)
Hotel Eden Roc, Sant Feliu de Guixols, Catalunya, Spain
State of the Semiconductor Industry

- Fabless market share is expanding: ~30%
- Semiconductor R&D is growing: $56B in 2015 to $67B in 2018
- R&D to revenue trend is flat at ~14%
- EDA is ~2% of semiconductor revenue

- Wireless is slowing down as primary growth driver of semiconductors
- IoT (Analytic and Security) will be new drivers: ~$19T economy
- Value is shifting away from semiconductor/hardware to software/services
- Vertical integration trend will continue in complex products
State of the Semiconductor Technology

- Moore’s Law slow down creating ripple effect and accelerating trends
- Interconnect delay and process variations overshadows technology scaling value
- DRAM bit cost scaling is slowing down
- Architectural bottlenecks are shifting to memory interfaces driven by massive data movements
- New workloads are emerging including ML and Deep Learning driving computing architectures
- Security requirements are imposing new latency and throughput challenges

- Move towards data flow processors and data stream accelerators
- Near memory and in memory compute architectures are emerging
- Move towards approximate computing, fault tolerance and real time monitoring
- Need for new NV memory technologies and new 3D integration technologies
More Moore (CMOS Scaling)

\[ I_{D,Sat} = \mu W C_{inv} v_s (V_G - V_{th}) \]

\( v_s \) is the saturation velocity for short channel lengths

**Strained Silicon***

*or high mobility channel materials

**Hi-k Metal Gate**

**Appropriately tuned transistors**

Leakage

- LVT
- SVT
- HVT

Performance

**Technology**

ACKNOWLEDGEMENTS: Perlmutter, D. 2012 Sustainability in Silicon and Systems Development. ISSCC Dig. Tech Papers, 30-34.
CMOS Scaling – Toward The End of Roadmap

State of the Technology
Evolution of Connectivity and Computing

Yesterday
<1B Users

Today
~3B Users
~7B Devices

Tomorrow
~50B Devices by 2020
~$19T Revenue by 2025
Exponential Technology Progress

The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies

by Erik Brynjolfsson and Andrew McAfee

- Steam Engine, Moore’s Law, Internet Traffic, Digital Content, …
  - Exponential growth often trigger other exponential growth trends
- Our brains are better at linear extrapolation and often fail to properly size sustained exponential growth
  - 4 billion grains on 32\textsuperscript{nd} square
  - 18 quintillion grains on 64\textsuperscript{th} square (a pile bigger than Mt. Everest)

We are about to move into the second half of the chessboard
Always Aware, Always Connected and Secure

Secure and seamless wireless connectivity amongst things, hub/fog and cloud/clouds

OBSERVES

DISCOVERS

RECOGNIZES

ANTICIPATES

Secure data fusion and data analytic
Sensitive data in transit remains vulnerable

Source: Rambus at GSA Forum, June 2016
IoT Faces New Security and Privacy Challenges

Security is critical for all elements of IoT stack

- Security on device level challenging due to low performance/memory
- IoT partly covering mission critical functionality (e.g., connected car)
- Sensitive private or commercial data (e.g., health data)

Exemplary security risks
- Snooping/blocking of commands/data transmission
- Injection/alteration of data or commands
- Delay of legitimate commands
- Use of hacked devices for other attacks

- End-to-end security across the entire stack
- Compliance to regulatory and privacy requirements
- Safe encryption requiring minimal computing resources
- Agile adaption to new emerging threats without user interaction
- Tampering/hacking detection

SOURCE: Expert interviews, GSA and McKinsey & Company "IoT Collaboration"
Security Requirements for IoT is Market Dependent

- Software-based Security
- Firmware-based Security
- Silicon-based Security

Source: Rambus at GSA Forum, June 2016
Security Requirements for IoT is Market Dependent

<table>
<thead>
<tr>
<th>Level of security</th>
<th>Security concepts/examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light</td>
<td>Public key cryptography, e.g., SSL, VPN, IPsec</td>
</tr>
<tr>
<td>Software</td>
<td>Secure memory, Certificate management</td>
</tr>
<tr>
<td>Hardware</td>
<td>Software obfuscation, Secure bootstrap</td>
</tr>
<tr>
<td></td>
<td>Hardware-implemented de/encoding algorithms, Hardware random number generators</td>
</tr>
<tr>
<td>Physical</td>
<td>Physically unclonable functions, Quantum transmissions</td>
</tr>
<tr>
<td>Strong</td>
<td></td>
</tr>
</tbody>
</table>

Typical requirements by segment:

- **Wearables**
- **Smart home**
- **Industrial**
- **Connected car**

Additional regulatory requirements exist in some regions.

SOURCE: Expert interviews; GSA and McKinsey & Company...
Data Analytic/Machine Learning Expands

Source: CISCO at GSA Forum, June 2016
Computing at the Cloud and Edge

Big Data and abundant computing power are pushing computing to the Cloud

Instant Data generated by users and machines are pushing computing to the Edge
Demand for Processing Efficiency at the Edge is Increasing

Edge Computing
Required to efficiently process instant data from devices and sensors in real-time

Cloud Computing
Required for big data analysis and when instantaneous response at the edge is less critical

Drivers for more processing demand at the edge

More connected things
~8 Billion cumulative smartphone forecast between 2014-2018¹
~50 Billion interconnected device forecast by 2020²

Always-on experiences
Always-on apps running in the background
More sensors in smartphones, wearables and connected objects

Immediate response
More responsive user experience, especially for instant, real-time data

Sources: 1) Gartner, Mar’14, 2) Machina Research/GSMA, Dec. ’12
What do we know about the Brain?

<table>
<thead>
<tr>
<th>Human Brain</th>
<th>IBM Sequoia</th>
</tr>
</thead>
<tbody>
<tr>
<td>~ 3.5 petabytes</td>
<td>1.6 petabytes</td>
</tr>
<tr>
<td>~ 20 petaFLOPS</td>
<td>16.3 petaFLOPS</td>
</tr>
<tr>
<td>~ 20 watts</td>
<td>7.9 mega watts</td>
</tr>
</tbody>
</table>

**The Human Brain**
- is a massively parallel machine with ~86B neurons
- has no system clock, it is event driven
- has no hardware/software distinction
- performs processing and memory by the same components
- is a self-organizing, self healing system
- continuously simplifies itself
Right Task on the Right Processing Engine

- The performance & power benefits of heterogeneity

- 2D -> 3D Video Conversion (View Generation)
  - Prefer GPU

- Character Recognition (MSER)
  - Prefer CPU

- Image Processing (Denoising)
  - Prefer GPU

Source: Qualcomm Technologies Inc.
Mobile Heterogeneous Computing Approach is Needed

Run the appropriate 4K task on the most suitable processing engine

Power reduction by running HEVC decode on specialized video engine versus the CPU

Source: Qualcomm Technologies Inc.
Brain is Heterogeneous Computing Machine

**Left Brain**
- Detail
- Logic
- Math
- Science

**Right Brain**
- Big picture
- Feeling
- Art
- Philosophy

**Frontal lobe**
- Executive functions, thinking, planning, organising and problem solving, emotions and behavioural control, personality

**Motor cortex**
- Movement

**Sensory cortex**
- Sensations

**Parietal lobe**
- Perception, making sense of the world, arithmetic, spelling

**Temporal lobe**
- Memory, understanding, language

**Occipital lobe**
- Vision

**Cerebellum**
- Controls balance and co-ordination

**Brain Stem**
- Breathing, body temperature, heart rate
The Brain is a Massively Parallel Machine

Modern computer
Dense, real-valued data

Human brain
Sparse Events, Approximate

$>10^6$ processing steps
$<10^1$ parallelism

$<10^1$ processing steps
$>10^6$ parallelism
Multi-Core and Heterogeneous Computing: First Step Towards Brain Inspired Computing

- **Power Efficiency**
  - Custom Accelerators
  - GPU & DSP

- **High Flexibility**
  - Neural Processing Unit (NPU)
  - Data Flow CPU

- Custom Accelerators
  - VeNum CPU
  - VeNum GPU & DSP
  - VeNum L2
  - VeNum CPU

Multi-core and heterogeneous computing, through the integration of various processing units and custom accelerators, aim to achieve a first step towards brain-inspired computing, prioritizing high flexibility and power efficiency.
Deep Learning: Second Step Towards Brain Inspired Computing

“Deep Learning” means using a neural network

– With several layers of nodes between input and output
– The series of layers between input & output do feature identification and processing in a series of stages
Deep Learning: Second Step Towards Brain Inspired Computing

LeCun et al., 1998
Deep Learning: Second Step Towards Brain Inspired Computing

Multilayer neural networks have been around for 25 years. What’s actually new?

Abundance of processing power and data in the Cloud
Approximate Computing and In Memory Compute: Third Step Towards Brain Inspired Computing

All pieces of a computation and data are not equivalent (some aspects need to be precise, others can be approximate)

Approximate Computing

- Embracing error
- Relax the abstraction of near-perfect accuracy in general-purpose computing, communication, and storage
- Improve resource utilization efficiency,
  - Energy consumption, area, performance

Fault Tolerance
Focus on reduction of precision for both weights (static value) and activations (dynamic values) versus traditional 32-bit floating approaches

- Physically smaller networks
- 2X improvement in memory access efficiency for network weights

### Neural Network Weight Bit Widths

<table>
<thead>
<tr>
<th>Activation Bit Widths</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>24</th>
<th>32</th>
<th>Float</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>20.0%</td>
<td>1.4%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
</tr>
<tr>
<td>16</td>
<td>20.1%</td>
<td>1.4%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>24</td>
<td>20.1%</td>
<td>1.4%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>32</td>
<td>20.1%</td>
<td>1.4%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Float</td>
<td>20.1%</td>
<td>1.4%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

16-bit values are used with no net increase in top-5 error
Approximate Computing and In Memory Compute: Third Step Towards Brain Inspired Computing

Fully connected layer compression significantly impacts network size

![Graph showing classification performance vs. memory footprint]

- **10X size reduction**
- **1% pt loss in top5 error**

Legend:
- Original
- Network
- FC Layer Compressed
- FC and Conv Layer Compressed
Approximate Computing and In Memory Compute: Third Step Towards Brain Inspired Computing

- Memory bandwidth is becoming the bottleneck
  - Significant amount of computing power is consumed to move data between memory and compute units
- Avoiding memory stalls is a huge architecture burden
- Solution is In Memory Compute (IMC), Computational RAM (CRAM), Processor-in-Memory (PIM)
  - The chief goal of merging the processing and memory components in this way is to reduce memory latency and increase bandwidth
  - Significantly impacts select workloads

Use In Memory Compute for BIST
Neuromorphic Computing: Fourth Step Towards Brain Inspired Computing

• A neuromorphic computer emulates human brain by using granular and abundant neurons and synapses communicating largely using event-driven spikes

• Neuromorphic computing systems excel at computing complex dynamics using a small set of computational primitives (neurons, synapses, spikes)
Neuromorphic Computing: Fourth Step Towards Brain Inspired Computing

- Long term, local Hebbian Learning
  - Neurons that fire together, wire together
  - Neurons that fire out of sync, lost their link
- Spike Timing Dependent Plasticity (STDP) is a modified version of Hebb’s postulate
Neuromorphic Computing: Fourth Step Towards Brain Inspired Computing

• The challenge
  • Learning or training the network to perform various tasks is still an unsolved problem
  • Training a neuromorphic computer requires massive amounts of data, compute and lots and lots of time
    • Human Brain evolved during millions of years and still takes human kids few years before they are independently functional
Technical Bottlenecks Towards Neuromorphic Computing

- Neuron Modeling to enable practical implementations
- Technology (Scaling, Memristors, MRAM, 3D)
**MRAM**

Zero retention/standby power from non-volatility, high density

<table>
<thead>
<tr>
<th>Memory Element</th>
<th>CMOS SRAM</th>
<th>In-plane MRAM</th>
<th>Perpendicular MRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Magnetic Anisotropy</strong></td>
<td>N/A</td>
<td>In-plane</td>
<td>Perpendicular</td>
</tr>
<tr>
<td><strong>Switching Mechanism</strong></td>
<td>Charge Racing</td>
<td>Spin-Transfer Torque (STT)</td>
<td>Spin-Transfer Torque (STT)</td>
</tr>
<tr>
<td><strong>Energy per Cell per Write</strong></td>
<td>0.52 pJ (40nm UMC)</td>
<td>W0: ~2x SRAM W1: ~1x SRAM</td>
<td>W0: ~24% SRAM W1: ~12% SRAM</td>
</tr>
</tbody>
</table>

**MRAM Bitcell and R/W Circuit**
Conclusion

- Semiconductor Industry is going through a tectonic and structural shift
- IoT market is emerging (Home, Auto, Industrial, Health, …)
- New computing technologies/architectures, memory technologies and integration technologies are being adopted
- Security and Data Analytic led by Machine Learning/Deep Learning will create new business opportunities
- Brain Inspired Computing is Emerging
  - First Step: Many Cores and Heterogeneous Computing
  - Second Step: Deep Learning
  - Third Step: Approximate and In-Memory Computing
  - Fourth and Final Step: Neuromorphic Computing