Domain Specific Processors Drive Changing Outlook for Semiconductor Memory

Walden C. Rhines
CEO Emeritus
Mentor, a Siemens Business
SEMICON Korea 2019
Acceleration of Semiconductor Revenue Growth

IC Annual Growth

Source: VLSI Research, December 2018

2.8 % 5yr CAGR


Source: VLSI Research, December 2018
MEMORY WAS THE PRIMARY SEMICONDUCTOR GROWTH DRIVER IN 2017 & 2018
Memory Now Almost 40% of IC Revenue
Versus 27% in 2016

Source: IC Insights 2018 Market Drivers
Memory Average Selling Prices Increased Dramatically in 2017 and 2018

Source: VLSI Research
Memory IC Unit Shipments Historically Follow a Trend-line With Few Variations

Source: VLSI Research

WCR, SEMICON Korea 2019
Memory IC Unit Shipments Historically Follow a Trend-line With Few Variations

Source: VLSI Research

WCR, SEMICON Korea 2019
Memory IC Unit Shipments Historically Follow a Trend-line With Few Variations

Total Memory IC Units
Millions

\[ y = 45.088x^2 - 205.8x + 1186.7 \]
\[ R^2 = 0.9776 \]

Source: VLSI Research
Despite Memory Revenue Growth in 2017 & 2018, Memory Transistor Volume Growth Is Below the Long Term Trend Line

\[ y = 1E+16e^{0.4868x} \]
\[ R^2 = 0.9861 \]

\[ y = 1994e^{0.0005x} \]
\[ R^2 = 1 \]

Source: VLSI Research
Transistor Unit Volume Should Begin Accelerating in 2019 & 2020 Growing Faster than 2017 and 2018
Learning Curve

Cost measured in constant currency assuming free markets

Doubling Cumulative Unit Volume Decreases Cost by a Fixed %

Rate of improvement tied to rate of unit volume growth

As cumulative unit production increases over time...

Cumulative Unit Production
Transistor cost reduction continues at ~32% per Year

Source: VLSI Research, Federal Reserve
Note: Revenue adjusted for Inflation... 1957-2018
Memory Transistor Unit Volume Dominates the Long Term Semiconductor Learning Curve

Semiconductor (excluding DRAM & FLASH)

DRAM & FLASH

Learning Curve CAGR 1995-2018
- DRAM & FLASH: -35%
- Remaining SEMI: -23%

Cumulative Transistors Shipped

Source: VLSI Research, Federal Reserve
Note: Revenue adjusted for Inflation... 1957-2018
Market Responds Quickly to Learning Curve Variation With Equal and Opposite Corrections

Revenue/Transistor vs. Cumulative Transistors Shipped

- DRAM & FLASH
- Remaining SEMI

Learning Curve CAGR 1995-2018
- DRAM & FLASH: -35%
- Remaining SEMI: -23%

Source: VLSI Research, Federal Reserve
Note: Revenue adjusted for Inflation... 1957-2018
Market Responds Quickly to Learning Curve Variation with Equal and Opposite Corrections

- Increased Transistor Shipments
- Increased Units
- Oversupply Correction

Source: VLSI Research, Federal Reserve
Note: Revenue adjusted for Inflation... 1957-2018
Memory Prices Peaked in Q1 2018, Expected to Drop in 2019
Negative Memory ASP Year-to-Year Growth in Q4 2018 and Forecast for Next 4 Quarters

- Units from trend-line assumptions
- Transistors from trend-line assumptions
- Revenue/transistor from learning curve
- Transistors/unit calculation
- Revenue/unit calculation = ASP

Source: VLSI Research, TCI Graphics December 2018 & Mentor
2018 DRAM/SRAM CapEx Growth Est 56%

Capital Spending by Product Type
$ Millions

Source: McClean Report, IC Insights, November 2018 Update
WILL THE MARKET NEED THE INCREASED MEMORY CAPACITY?
Memory Now Dominates Transistors Manufactured

Memory as a Percentage of All Transistors

Source: VLSI Research, 2016
3D NAND Allows Memory to Scale Faster than Logic or SoC’s

<table>
<thead>
<tr>
<th>Supplier</th>
<th>2012</th>
<th>2013</th>
<th>1H14</th>
<th>2H14</th>
<th>1H15</th>
<th>2H15</th>
<th>2016</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samsung</td>
<td>21nm</td>
<td>19nm MLC</td>
<td>TLC</td>
<td>16nm MLC</td>
<td>TLC</td>
<td>14nm</td>
<td>12nm 10nm</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>24L MLC</td>
<td>SSD</td>
<td></td>
<td>32L MLC</td>
<td>TLC</td>
<td>48L MLC</td>
<td>TLC 64L</td>
</tr>
<tr>
<td>Toshiba/SanDisk</td>
<td>24nm</td>
<td>19nm</td>
<td>TLC</td>
<td>SSD</td>
<td>15nm</td>
<td>12nm</td>
<td>10nm</td>
<td></td>
</tr>
<tr>
<td>Micron/IM Flash</td>
<td>20nm</td>
<td>16nm</td>
<td>SSD</td>
<td>TLC</td>
<td>12nm</td>
<td>32L MLC</td>
<td>48L</td>
<td></td>
</tr>
<tr>
<td>SK Hynix</td>
<td>25nm</td>
<td>20nm</td>
<td>16nm</td>
<td>TLC</td>
<td>12nm</td>
<td>32L MLC</td>
<td>48L</td>
<td></td>
</tr>
</tbody>
</table>

Source: IC Insights, McClean Report, 2016
Traditional Von Neumann Computer Architectures Are Not Efficient for Pattern Recognition
Evolution of Non-von Neumann Computer Architectures Will Improve Processing Speed, Reduce Power and Integrate More Memory
MAJOR WAVE OF NEW “DOMAIN SPECIFIC” ARCHITECTURES
The End of Moore’s Law & Faster General Purpose Computing, and a New Golden Age

- **SW-centric**
  - Modern scripting languages are interpreted, dynamically-typed and encourage reuse
  - Efficient for programmers but not for execution

- **HW-centric**
  - Only path left is *Domain Specific Architectures*
  - Just do a few tasks, but extremely well

- **Combination**
  - Domain Specific Languages & Architectures
Venture Capital Investment in Fabless Semiconductor Startups

Fabless Semiconductor Venture Capital Investment by Year, All Rounds ($Millions)

Source: Global Semiconductor Alliance (GSA), IMF, VentureSource, Pitchbook, Crunchbase, & Mentor Graphics Analysis Rev 01/09/19
Startups Dominated by Domain Specific Architectures
Worldwide Fabless Company Venture Capital Funding (1-3)

Source: Global Semiconductor Alliance (GSA), VentureSource, PitchBook & Mentor Graphics Analysis
Revised 01/09/19
## Market Segments Funded 2012 – 2018 By Funding Dollars ($M)

<table>
<thead>
<tr>
<th>Market Segment</th>
<th>Funding Dollars ($M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI &amp; Machine Learning</td>
<td>$1,834</td>
</tr>
<tr>
<td>Crypto Currency</td>
<td>$465</td>
</tr>
<tr>
<td>Comm. High Speed / 5G</td>
<td>$348</td>
</tr>
<tr>
<td>Memory</td>
<td>$312</td>
</tr>
<tr>
<td>Analog &amp; Power</td>
<td>$264</td>
</tr>
<tr>
<td>Sensors &amp; MEMS</td>
<td>$232</td>
</tr>
<tr>
<td>MPU &amp; MCU</td>
<td>$134</td>
</tr>
<tr>
<td>Video Compression</td>
<td>$73</td>
</tr>
<tr>
<td>Silicon Photonics</td>
<td>$123</td>
</tr>
<tr>
<td>LED</td>
<td>$55</td>
</tr>
<tr>
<td>Waveguide Optics</td>
<td>$25</td>
</tr>
<tr>
<td>Flexible Transistors-TFT</td>
<td>$25</td>
</tr>
<tr>
<td>VCSCEL Image Laser Chip</td>
<td>$15</td>
</tr>
<tr>
<td>Image Processor</td>
<td>$4</td>
</tr>
<tr>
<td>Hardware Security SoC</td>
<td>$2</td>
</tr>
</tbody>
</table>

Source: Global Semiconductor Alliance (GSA), VentureSource, PitchBook & Mentor Graphics Analysis
Revised 1/9/19
ARTIFICIAL INTELLIGENCE IS NOT A NEW TREND
“Artificial Intelligence Heavyweights”, July 1986
A Technology that Arrived Before Its Time
Reasons for AI Adoption Delay in the 1980s

- Lack of big data to analyze
  - No Internet or IoT to collect sizable data sets

- Limited computing power
  - Limitation of traditional computer chip architectures

- Need for more advanced algorithms

- Lack of ‘killer’ applications to make money
What’s Different Today?
All these limitations are going away

- Lack of big data to analyze
  - No Internet or IoT to collect sizable data sets

- Limited computing power
  - Limitation of traditional computer chip architectures

- Need for more advanced algorithms

- Lack of ‘killer’ applications to make money
24 Fabless AI Companies Received Venture Capital Funding 2018
Worldwide Fabless Company Venture Capital Funding (Total Rounds)

AI, Deep Learning Portion of Funding $1,844M - 2018

Selected companies

AIMotive Gmbh, $38M – ADAS Acceleration Chip
Beijing Intengine Tech – $30M – AI chips for end-point devices in multiple industries
Hailo Technologies – $12.5M – AI chips data center edge devices
Kneron, $18M – AI Network Processor (NPU)
Metawave, $10M – AI chips driving radar chip
Mythic Inc, $40M – IPU for adding intelligence to any device
NextVPU, $2M – AI engine for autonomous driving radar chip
Sensetime, $1,220M – AI architecture with facial recognition technology
Syntiant, $30.1M – Neural Decision Processors (NDP) - reducing data movement
Xanadu Quantum Tech, $9M – Quantum processor chip using Silicon Photonics
Hangzhou Fabu Technologies, $Undisclosed – AI adaptive chip architecture for ADAS
Cambricon, $Undisclosed – Processor for AI, datacenter
Tachyum Inc, $Undisclosed – Processor for AI, data center and HPC platforms

Source: Global Semiconductor Alliance (GSA), VentureSource, PitchBook & Mentor Graphics Analysis
Revised 1/9/19
Domain-Specific AI/Deep Learning Controllers

- Vision/Facial Recognition: 39
- Voice/Speech/Sound Pattern Recognition: 9
- Autonomous Driving/ADAS: 17
- Disease Diagnosis AI: 3
- Optical Computing AI: 6
- Smell/Odor Recognition: 2
- Data Center/Cloud AI/HPC: 23
- Unknown/Stealth Mode: 6
- Space/Military Applications: 3
- Cryptocurrency: 2
- Edge Computing: 21
- Robotics/Motion Control/Collision Avoidance: 4
- Deep Learning - Training: 8
- Intelligent Wireless Control: 1
Pattern Recognition Dominates New AI Designs

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision/Facial Recognition</td>
<td>39</td>
</tr>
<tr>
<td>Voice/Speech/Sound Pattern Recognition</td>
<td>9</td>
</tr>
<tr>
<td>Autonomous Driving/ADAS</td>
<td>17</td>
</tr>
<tr>
<td>Disease Diagnosis AI</td>
<td>3</td>
</tr>
<tr>
<td>Optical Computing AI</td>
<td>6</td>
</tr>
<tr>
<td>Smell/Odor Recognition</td>
<td>2</td>
</tr>
<tr>
<td>Data Center/Cloud AI/HPC</td>
<td>23</td>
</tr>
<tr>
<td>Unknown/Stealth Mode</td>
<td>6</td>
</tr>
<tr>
<td>Space/Military Applications</td>
<td>3</td>
</tr>
<tr>
<td>Cryptocurrency</td>
<td>2</td>
</tr>
<tr>
<td>Edge Computing</td>
<td>21</td>
</tr>
<tr>
<td>Robotics/Motion Control/Collision Avoidance</td>
<td>4</td>
</tr>
<tr>
<td>Deep Learning - Training</td>
<td>8</td>
</tr>
<tr>
<td>Intelligent Wireless Control</td>
<td>1</td>
</tr>
</tbody>
</table>
Facial Recognition

- KAIST Convolutional Neural Network Processor
  - Custom image sensor
  - Facial recognition software
  - 97% accuracy, 0.62mW

- K-Eye device “always on” facial detection
# Data Center/Cloud AI Chip Architectures for Performance

<table>
<thead>
<tr>
<th>Application</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision/Facial Recognition</td>
<td>39</td>
</tr>
<tr>
<td>Voice/Speech/Sound Pattern Recognition</td>
<td>9</td>
</tr>
<tr>
<td>Autonomous Driving/ADAS</td>
<td>17</td>
</tr>
<tr>
<td>Disease Diagnosis AI</td>
<td>3</td>
</tr>
<tr>
<td>Optical Computing AI</td>
<td>6</td>
</tr>
<tr>
<td>Smell/Odor Recognition</td>
<td>2</td>
</tr>
<tr>
<td>Data Center/Cloud AI/HPC</td>
<td>23</td>
</tr>
<tr>
<td>Unknown/Stealth Mode</td>
<td>6</td>
</tr>
<tr>
<td>Space/Military Applications</td>
<td>3</td>
</tr>
<tr>
<td>Cryptocurrency</td>
<td>2</td>
</tr>
<tr>
<td>Edge Computing</td>
<td>21</td>
</tr>
<tr>
<td>Robotics/Motion Control/Collision Avoidance</td>
<td>4</td>
</tr>
<tr>
<td>Deep Learning - Training</td>
<td>8</td>
</tr>
<tr>
<td>Intelligent Wireless Control</td>
<td>1</td>
</tr>
</tbody>
</table>
Datacenter AI IC Examples

NVIDIA Bets On Turing For Datacenter AI

POST WRITTEN BY
Karl Freund

Karl Freund is a Moor Insights & Strategy Senior Analyst for deep learning & HPC

NVIDIA’s datacenter business has been on a tear, doubling every year for the past several years. It hit $1.93 billion in the fiscal year, an increase of nearly 130% over the previous year. The increase has been largely driven by the pervasive use of NVIDIA GPUs in HPC neural network training for Artificial Intelligence research and development.

Xilinx Moves FPGAs Deeper into Datacenters

October 17, 2018

Three Chinese infrastructure vendors are embracing FPGA technology as a way of accelerating datacenter workloads.

AI Chip Startup Graphcore Rings Up $200M in Funding

Graphcore, the U.K. AI chip developer, has attracted more investors as it rolls out its intelligent processing unit to enterprise datacenters and cloud services providers. The GPU rival reported this week it has closed a $200 million funding round led by the London-based venture firm Atamis and other existing investors. New investors included automaker BMW and Microsoft, who join Robert Bosch Venture Capital, Dell Technologies Capital and Samsung Electronics as auto and cloud investors.
# Domain-Specific AI/Deep Learning Controllers

<table>
<thead>
<tr>
<th>Domain</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision/Facial Recognition</td>
<td>39</td>
</tr>
<tr>
<td>Voice/Speech/Sound Pattern Recognition</td>
<td>9</td>
</tr>
<tr>
<td>Autonomous Driving/ADAS</td>
<td>17</td>
</tr>
<tr>
<td>Disease Diagnosis AI</td>
<td>3</td>
</tr>
<tr>
<td>Optical Computing AI</td>
<td>6</td>
</tr>
<tr>
<td>Smell/Odor Recognition</td>
<td>2</td>
</tr>
<tr>
<td>Data Center/Cloud AI/HPC</td>
<td>23</td>
</tr>
<tr>
<td>Unknown/Stealth Mode</td>
<td>6</td>
</tr>
<tr>
<td>Space/Military Applications</td>
<td>3</td>
</tr>
<tr>
<td>Cryptocurrency</td>
<td>2</td>
</tr>
<tr>
<td>Edge Computing</td>
<td>21</td>
</tr>
<tr>
<td>Robotics/Motion Control/Collision Avoidance</td>
<td>4</td>
</tr>
<tr>
<td>Deep Learning - Training</td>
<td>8</td>
</tr>
<tr>
<td>Intelligent Wireless Control</td>
<td>1</td>
</tr>
</tbody>
</table>
Automotive AI Chip Examples

- Ambarella/Smart Eye Monitoring System - track driver actions and intentions
- Eyeris AI chip for the auto industry’s AEC-Q100 standard
- Israeli startup Hailo expects to release automotive deep learning processor 1H 2019
HOW LONG CAN SILICON TECHNOLOGY SUPPORT THESE CHIP ARCHITECTURE INNOVATIONS?
The Gompertz Curve

- Benjamin Gompertz, 1825
- Mathematical model for time series
  - Maximum Growth at 36.8% of Asymptote
- Example
  - Growth of tumors
  - Population growth
  - Mobile phone uptake
  - Market impact in finance

\[ y = ae^{-be^{-ct}} \]
Total Cell Phone Subscribers
Gompertz Forecast 2017

Source: International Telecommunications Union

Source: The Guardian
Total Cell Phone Subscribers
Gompertz Forecast 1997

Source: International Telecommunications Union

Source: The Guardian
Total Cell Phone Subscribers
Gompertz Forecast 1997 vs. 2017

Source: The Guardian

Source: International Telecommunications Union
PC Notebooks
Gompertz Forecast 2016

Sources: Semico Research, IC Insights & Mentor Graphics
PC Notebooks
Gompertz Forecast 2001

Cumulative Unit Volume Billions

Sources: Semico Research, IC Insights & Mentor Graphics
PC Notebooks
Gompertz Forecast 2001 vs. 2016

Cumulative Unit Volume Billions

Sources: Semico Research, IC Insights & Mentor Graphics
Data Center Processing Will Grow for Decades

Server Semiconductor Consumption
4.2% of Total Semiconductor Revenue 2015

Sources: Gartner Group, IC Insights & Mentor Graphics analysis
Total Transistors Gompertz Curve

Source: VLSI Research, 2016
Need an Alternative to the Silicon Transistor ‘Switch’ Before 2040

Source: VLSI Research, 2016

Silicon Transistor Peak Growth (36.8% of Asymptote)

Slowing Growth

Asymptote

Total Transistors

Forecast

Cumulative Transistor Unit Volume

0.0E+00 5.0E+23 1.0E+24 1.5E+24 2.0E+24 2.5E+24 3.0E+24 3.5E+24 4.0E+24 4.5E+24


Today

2038

WCR, SEMICON Korea 2019
Summary: IC Design Activity Is Increasing Rapidly

- Sudden **acceleration** of growth of semiconductor revenue in 2017
- Memory **unit volume** is **below** the long term **trend** but unit **prices** are **above** the long term learning curve
  - Decreasing memory selling **prices** in 2019 and 2020 will be offset somewhat by increased unit volume
- Continued **non-memory** market strength due to:
  - Introduction of “domain specific processors”
  - Acceleration in start-up activity
  - Entry of new participants in IC design
  - Chinese investment in semiconductors
- **Domain specific architectures** and **memory unit growth** will enable next wave of semiconductor industry growth
  - Lower power, higher performance, lower cost
  - Reduction in design time and cost of design