The Future of Computing from a Memory/Storage Centric Point-of-view

Steve Pawlowski, Vice President, Advanced Computing Solutions

November 4th, 2019
Emergence of the Data Economy

Virtuous Cycle Driven by Increased Data Value
- Creates continuous need to capture, process, move & store data
- Generates ever-increasing demand for memory & fast storage

Demand for Memory Density Growth Insatiable

<table>
<thead>
<tr>
<th>Year</th>
<th>Internet Era – 250B GB/Year</th>
<th>Mobile Era – 7,000B GB/Year</th>
<th>Mobile Era – 22,000B GB/Year</th>
<th>Mobile Era – 62,000B GB/Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000-2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008-2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Can classical computing provide 2x performance gain every two years?

Legacy Memory Model support impacts the architecture efficiency...

Instead of more transistors for less gain in instruction level performance
Add more cores for greater parallel performance – Good for AI

More Cores = More Memory IO
Efficiencies fall off for BW intensive workloads.

Restoring ‘system’ balance is critical.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer</th>
<th>Cores</th>
<th>HPL Rmax (Pflop/s)</th>
<th>TOP500 Rank</th>
<th>HPCG (Pflop/s)</th>
<th>Fraction of Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/SC/ORNL (USA)</td>
<td><strong>Summit</strong> – AC922, IBM POWER9 22C 3.07GHz, dual-rail Mellanox EDR Infiniband, NVIDIA Volta V100 (IBM)</td>
<td>2,397,824</td>
<td>143.500</td>
<td>1</td>
<td>2.926</td>
<td>1.5%</td>
</tr>
<tr>
<td>2</td>
<td>DOE/NNSA/LLNL (USA)</td>
<td><strong>Sierra</strong> – S922LC, Power9 22C 3.1GHz, Mellanox EDR, NVIDIA Tesla V100 (IBM / NVIDIA / Mellanox)</td>
<td>1,572,480</td>
<td>94.640</td>
<td>2</td>
<td>1.796</td>
<td>1.4%</td>
</tr>
<tr>
<td>3</td>
<td>RIKEN Advanced Institute for Computational Science (Japan)</td>
<td><strong>K computer</strong> – , SPARC64 VIIIfx 2.0GHz, Tofu interconnect (Fujitsu)</td>
<td>705,024</td>
<td>10.510</td>
<td>18</td>
<td>0.603</td>
<td>5.3%</td>
</tr>
<tr>
<td>4</td>
<td>DOE/NNSA/LANL/SNL (USA)</td>
<td><strong>Trinity</strong> – Cray XC40, Intel Xeon E5-2698 v3 16C 2.3GHz, Aries, Intel Xeon Phi 7250 68C 1.4GHz (Cray)</td>
<td>979,072</td>
<td>20.159</td>
<td>6</td>
<td>0.546</td>
<td>1.3%</td>
</tr>
<tr>
<td>5</td>
<td>National Institute of Advanced Industrial Science and Technology (AIST) (Japan)</td>
<td><strong>Al Bridging Cloud Infrastructure (ABCI)</strong> – PRIMERGY CX2570M4, Intel Xeon Gold 6148 20C 2.4GHz, Infiniband EDR, NVIDIA Tesla V100 (Fujitsu)</td>
<td>368,640</td>
<td>16.859</td>
<td>10</td>
<td>0.509</td>
<td>1.7%</td>
</tr>
<tr>
<td>6</td>
<td>Swiss National Supercomputing Centre (CSCS) (Switzerland)</td>
<td><strong>Piz Daint</strong> – Cray XC50, Intel Xeon E5-2690v3 12C 2.6GHz, Cray Aries, NVIDIA Tesla P100 16GB (Cray)</td>
<td>387,872</td>
<td>21.230</td>
<td>5</td>
<td>0.497</td>
<td>1.8%</td>
</tr>
<tr>
<td>7</td>
<td>National Supercomputing Center in Wuxi (China)</td>
<td><strong>Sunway TaihuLight</strong> – Sunway MPP, SW26010 260C 1.45GHz, Sunway (NRCPC)</td>
<td>10,649,600</td>
<td>93.015</td>
<td>3</td>
<td>0.481</td>
<td>0.4%</td>
</tr>
</tbody>
</table>
Many Workloads Require higher BW/FLOP, Not lower

Assume a 24-core chip, 512-bit-wide vector unit, @ 3GHz.

1.15 Peak TFLOPs
Peak Memory BW needed - ~9TB/s to ~14TB/s
Peak memory power (@ 6 pJ/b) – ~432W to ~650W

To improve system efficiency, we need to improve the BW to Flops ratio of memory/compute systems

AND...

*Reduce Data Movement power*
High device defect rate (>15%) may become a fact of life

- Functional Redundancy…memory has been doing this for a LONG time!

Dynamic Reconfigurability

With 100’s of replicated cores on die, performance and functionality can be maintained.

Memory technologies we have today will still be around for some time.

<table>
<thead>
<tr>
<th></th>
<th>DRAM</th>
<th>STTRAM</th>
<th>PCM/ 1T1R</th>
<th>Cross Point RRAM</th>
<th>NAND</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read Latency</strong></td>
<td>20ns</td>
<td>~50ns</td>
<td>~100ns-200ns</td>
<td>~100ns-200ns</td>
<td>~10us</td>
</tr>
<tr>
<td><strong>Write Latency</strong></td>
<td>20ns</td>
<td>~50ns</td>
<td>~1us</td>
<td>~1us</td>
<td>~10us</td>
</tr>
<tr>
<td><strong>Read Endurance</strong></td>
<td>&gt;1e15</td>
<td>&gt;10^11</td>
<td>&gt;10^7</td>
<td>&gt;10^7</td>
<td>&gt;10^7</td>
</tr>
<tr>
<td><strong>Write Endurance</strong></td>
<td>&gt;1e15</td>
<td>&gt;10^11</td>
<td>&gt;10^6</td>
<td>&gt;10^6</td>
<td>2K-100K</td>
</tr>
<tr>
<td><strong>Alterability</strong></td>
<td>~2KB</td>
<td>&lt;2KB</td>
<td>~10’s B</td>
<td>~10’s B</td>
<td>Large Blocks</td>
</tr>
<tr>
<td><strong>Retention@RT</strong></td>
<td>~milli seconds</td>
<td>Months</td>
<td>~Years</td>
<td>~Years</td>
<td>Years</td>
</tr>
<tr>
<td><strong>Areal Density</strong></td>
<td>1X</td>
<td></td>
<td></td>
<td></td>
<td>~30x</td>
</tr>
</tbody>
</table>
A challenge is not the memory device, but the way it’s used.

Intrinsic, on die, Memory BW is high, but is constrained by the off die system bus.

If we stay with today’s paradigm, the memory bottleneck continues.
- Memory energy is interconnect dominated.
Higher memory BW = higher power.
Reduce the interconnect distance.

J. Hasler, B. Marr; "Finding a Roadmap to achieve large neuromorphic hardware systems”; Frontiers in Neuroscience, Sept 10, 2013
To improve system performance and power efficiency – MOVE compute to where the data is stored.

Bytes/FLOP could improve by over 10x

The opportunity is deciding the type of computation to put near/in memory
When considering an ‘architectural’ change...

Likely the best ‘product’ advice I’ve ever received...

“The architecture that wins is the one that’s EASIEST to program”

So the architecture should have:
- High Performance efficiency for memory intensive workloads.
- Bring the ‘Compute to the Memory’.
- Scalable to handle today’s and future algorithms.
- Robust operation even with high device failure rates
- Forward compatibility… ‘preserve’ 40+ years of SW investment.

… Scrutinize measures of goodness carefully.
Artificial Neural Networks

Supporting the Basic Functionality is One Key to HW Scalability

only matrix multiplication, no feedback loop, low-latency, scalable, easily programmable, low-power consumption

\[
\begin{pmatrix}
  a & b & c \\
  d & e & f \\
  g & h & i \\
\end{pmatrix}
\begin{pmatrix}
  x \\
  y \\
  z \\
\end{pmatrix}
=
\begin{pmatrix}
  ax + by + cz \\
  dx + ey + fz \\
  gx + hy + iz \\
\end{pmatrix}
\]
AI/Machine Learning provides the capability to get more insight
With the larger volumes of data.
The ratio of compute to memory BW is different for different networks.
Example: BW demands for a ResNet-50 Network vary significantly depending on image resolution.

Flexibility of the architecture to ‘tune’ a network is a must for an optimal solution.

<table>
<thead>
<tr>
<th>resnet50 input image sizes w/Optimization</th>
<th>Gops</th>
<th>BW/Image (GB/s)</th>
<th>BW (30 Images/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>224x224x3</td>
<td>7.4</td>
<td>0.17</td>
<td>5.1</td>
</tr>
<tr>
<td>640x480x3</td>
<td>45.9</td>
<td>1.03</td>
<td>30.9</td>
</tr>
<tr>
<td>1920x1080x3</td>
<td>314.0</td>
<td>7.07</td>
<td>212.1</td>
</tr>
<tr>
<td>3840x2160x3</td>
<td>1256.3</td>
<td>28.3</td>
<td>849</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>resnet50 input image sizes w/o optimization</th>
<th>Gops</th>
<th>BW/Image (GB/s)</th>
<th>BW (30 Images/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>224x224x3</td>
<td>7.4</td>
<td>0.37</td>
<td>11.1</td>
</tr>
<tr>
<td>640x480x3</td>
<td>45.9</td>
<td>2.3</td>
<td>69</td>
</tr>
<tr>
<td>1920x1080x3</td>
<td>314.0</td>
<td>15.7</td>
<td>471</td>
</tr>
<tr>
<td>3840x2160x3</td>
<td>1256.3</td>
<td>62.82</td>
<td>1884.6</td>
</tr>
</tbody>
</table>
Combining memory and processing resources in a single device has huge potential to increase the performance and efficiency of DNNs... (to) achieve... performance in a system that can be generally useful across all problem sets.

Looking Forward – stacking memory on top of the Compute fabric, we can get high bandwidth, low energy and...yes...modest capacity.

“Combining memory and processing resources in a single device has huge potential to increase the performance and efficiency of DNNs... (to) achieve... performance in a system that can be generally useful across all problem sets.”

Q1. Neural computing layer should meet thermal and area constraint in 3D stacked DRAM
Q2. NeuroCube should be programmable to cover different types of neural network

Memory architecture provides insight into the next generation of AI Accelerators

- Exploit the unique physics of “emerging memory” technologies for in memory neural fabrics.
  - Summing (threshold) and sigmoid (triggering) behavior
  - Analog “weight” storage
  - Many recent papers based on resistive, magnetic, and floating gate technologies

$$y_i = \sum_{i=0}^{N} x_i w_{ij}$$