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

**Steve Pawlowski, Vice President, Advanced Computing Solutions**

**November 4th , 2019**

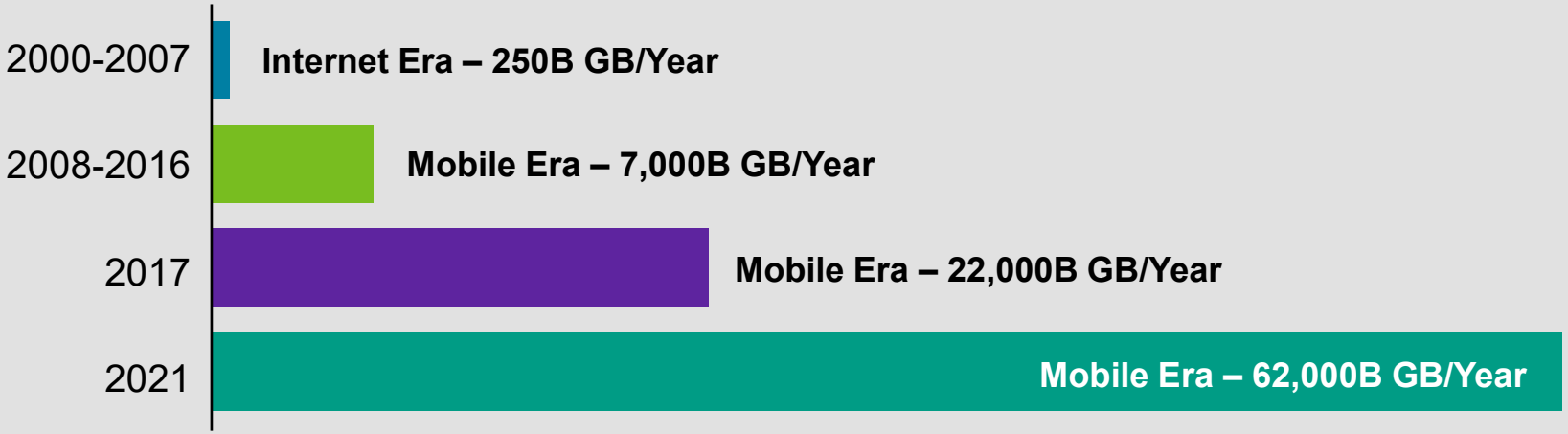
©2019 Micron Technology, Inc. All rights reserved. Information, products, and/or specifications are subject to change without notice. All information is provided on an "AS IS" basis without warranties of any kind. Statements regarding products, including regarding their features, availability, functionality, or compatibility, are provided for informational purposes only and do not modify the warranty, if any, applicable to any product. Drawings may not be to scale. Micron, the Micron logo, and all other Micron trademarks are the property of Micron Technology, Inc. All other trademarks are the property of their respective owners.

# Emergence of the Data Economy

**Virtuous Cycle Driven by Increased Data Value**

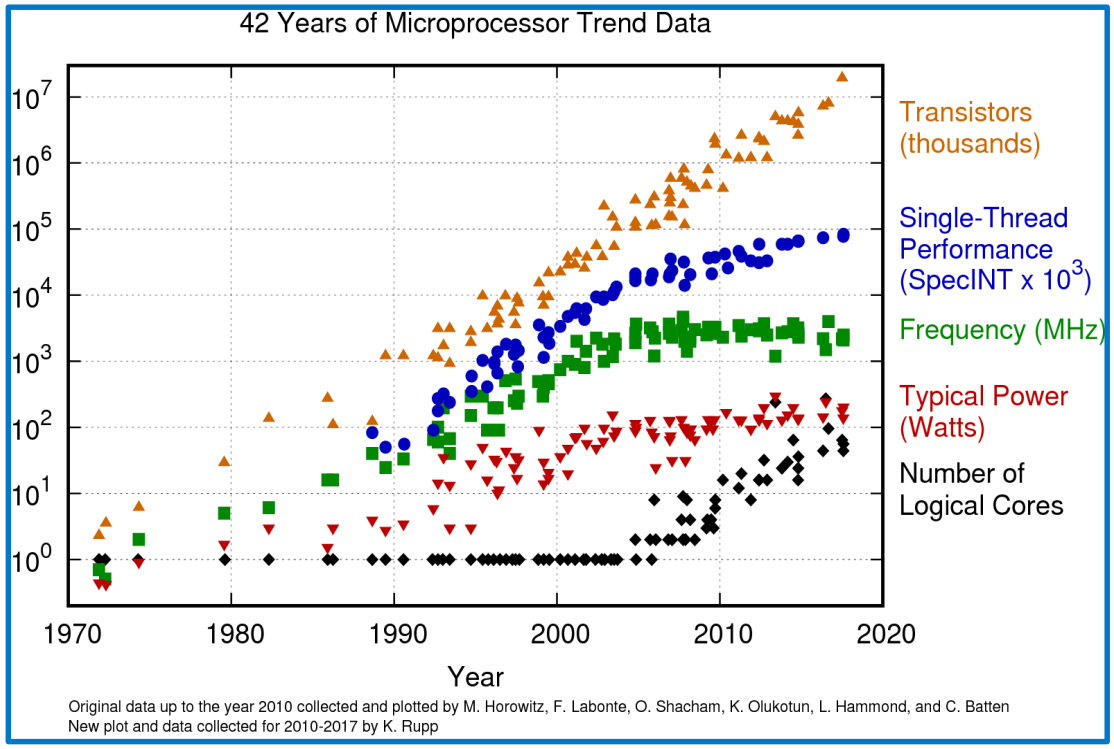
- Creates continuous need to capture, process, move & store data
- Generates ever-increasing demand for memory & fast storage

**Demand for Memory Density Growth Insatiable**



# 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**



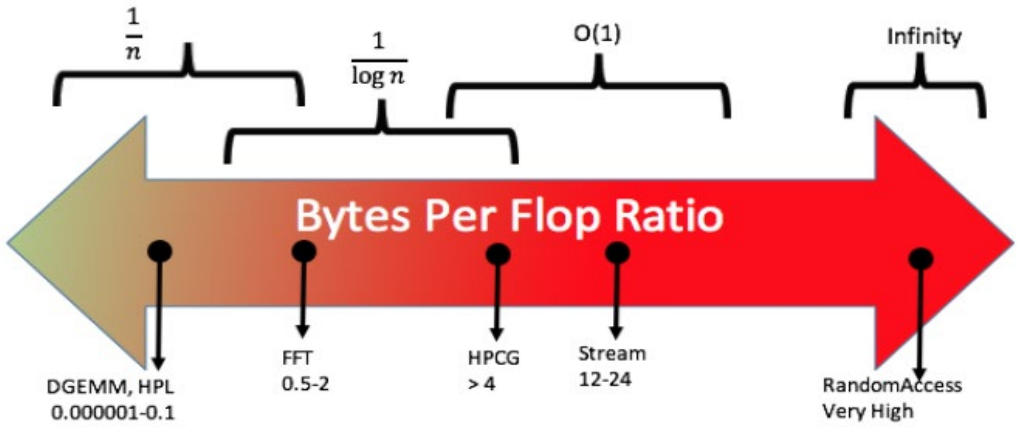
Micron

# Efficiencies fall off for BW intensive workloads.

## Restoring ‘system’ balance is critical.

| Rank | Site | Computer | Cores | HPL Rmax (Pflop/s) | TOP500 Rank | HPCG (Pflop/s) | Fraction of Peak |
|---|---|---|---|---|---|---|---|
| 1 | DOE/SC/ORNL (USA) | **Summit** – AC922, IBM POWER9 22C 3.07GHz, dual-rail Mellanox EDR Infiniband, NVIDIA Volta V100 (IBM) | 2,397,824 | 143.500 | 1 | 2.926 | 1.5% |
| 2 | DOE/NNSA/LLNL (USA) | **Sierra** – S922LC, Power9 22C 3.1GHz, Mellanox EDR, NVIDIA Tesla V100 (IBM / NVIDIA / Mellanox) | 1,572,480 | 94.640 | 2 | 1.796 | 1.4% |
| 3 | RIKEN Advanced Institute for Computational Science (Japan) | **K computer** – , SPARC64 VIIIfx 2.0GHz, Tofu interconnect (Fujitsu) | 705,024 | 10.510 | 18 | 0.603 | 5.3% |
| 4 | DOE/NNSA/LANL/SNL (USA) | **Trinity** – Cray XC40, Intel Xeon E5-2698 v3 16C 2.3GHz, Aries, Intel Xeon Phi 7250 68C 1.4GHz (Cray) | 979,072 | 20.159 | 6 | 0.546 | 1.3% |
| 5 | National Institute of Advanced Industrial Science and Technology (AIST) (Japan) | **AI Bridging Cloud Infrastructure (ABCI)** – PRIMERGY CX2570M4, Intel Xeon Gold 6148 20C 2.4GHz, Infiniband EDR, NVIDIA Tesla V100 (Fujitsu) | 368,640 | 16.859 | 10 | 0.509 | 1.7% |
| 6 | Swiss National Supercomputing Centre (CSCS) (Switzerland) | **Piz Daint** – Cray XC50, Intel Xeon E5-2690v3 12C 2.6GHz, Cray Aries, NVIDIA Tesla P100 16GB (Cray) | 387,872 | 21.230 | 5 | 0.497 | 1.8% |
| 7 | National Supercomputing Center in Wuxi (China) | **Sunway TaihuLight** – Sunway MPP, SW26010 260C 1.45GHz, Sunway (NRCPC) | 10,649,600 | 93.015 | 3 | 0.481 | 0.4% |

# Many Workloads Require higher BW/FLOP, Not lower




Assume a 24-core chip, 512bit-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

| Kernel Name | Computation Complexity | Number of computation | Number of Bytes | Bytes / Flop Ratio |
|---|---|---|---|---|
| SYMGS | O(nrows * nnz/row) | 2 *(2*nnz/row +3)* nrows | 2 * ( nnz/row * (2*8+4) + 5*8+2*4 ) *nrows | 10.32 |
| SPMV | O(nrows * nnz/row) | 2 * nnz/row * nrows | (nnz/row * (2*8+4)+2*8+2*4) * nrows | 10.44 |
| WAXPBY | O(nrows) | 2 * nrows | nrows * 3 * 8 | 12 |
| DDOT | O(nrows) | 2 * nrows | nrows * 2 * 8 | 8 |

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

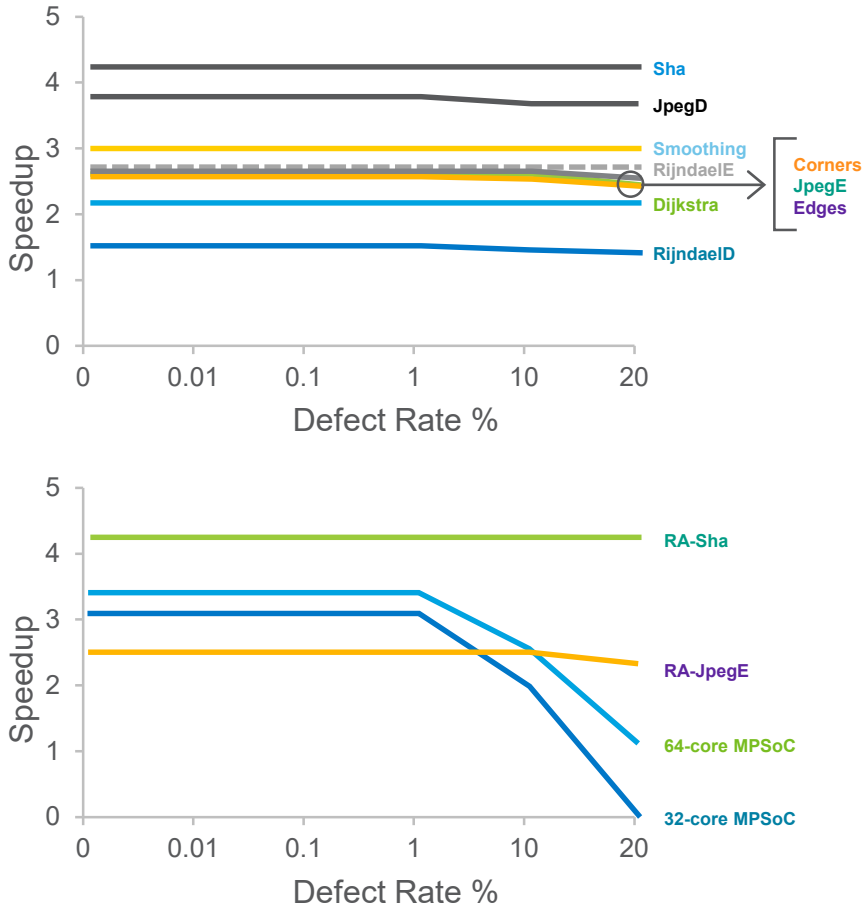
***Reduce Data Movement power***

# 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.




Source: Monica Magalhaes Periera and Luigi Carro, "Dynamic Reconfigurable Computing: The Alternative to Homogeneous Multicores under Massive Defect Rates", *International Journal of Reconfigurable Computing*, Vol. 2011.

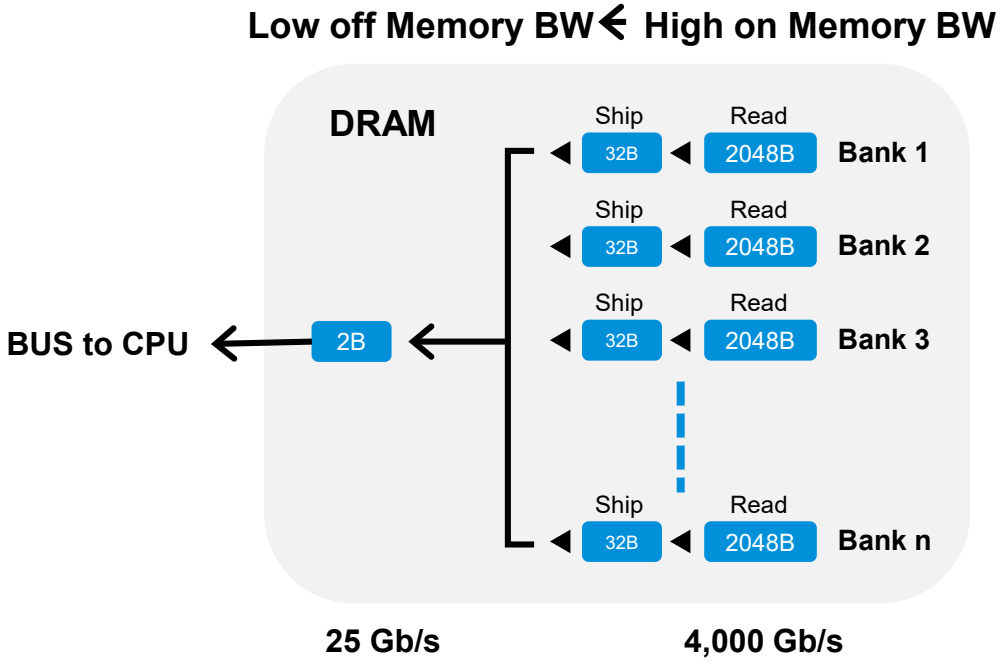
# Memory technologies we have today will still be around for some time.

| | DRAM | STTRAM | PCM/ 1T1R | Cross Point RRAM | NAND |
|---|---|---|---|---|---|
| Read Latency | 20ns | ~50ns | ~100ns-200ns | ~100ns-200ns | ~10us |
| Write Latency | 20ns | ~50ns | ~1us | ~1us | ~10us |
| Read Endurance | >1e15 | $>10^{11}$ | $>10^{7}$ | $>10^{7}$ | $>10^{7}$ |
| Write Endurance | >1e15 | $>10^{11}$ | $>10^{6}$ | $>10^{6}$ | 2K-100K |
| Write/Read Energy/Bit | <10pJ/bit | ~25pJ/bit | ~100-200 pJ/bit | ~100-200 pJ/bit | >100pJ/bit |
| Alterability | ~2KB | <2KB | ~10's B | ~10's B | Large Blocks |
| Retention@RT | ~milli seconds | Months | ~Years | ~Years | Years |
| Areal Density | 1X | | | | ~30x |



Micron

# A challenge is not the memory device, but the way it's used.



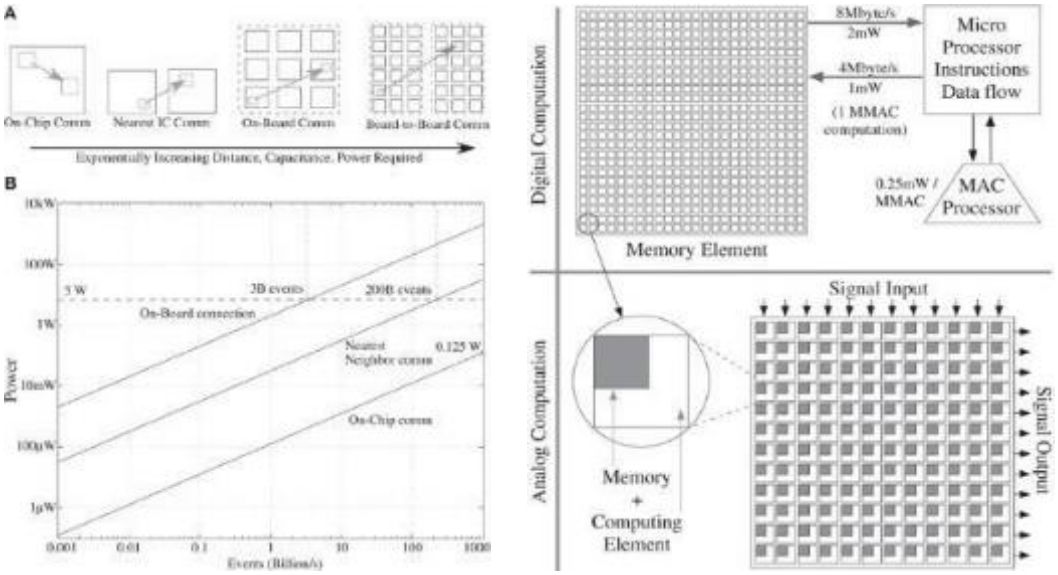

Intrinsic, on die, Memory BW is high, but is constrained by the off die system bus.

**If we stay with today's paradigm, the memory bottleneck continues.**

- Memory energy is interconnect dominated

Micron

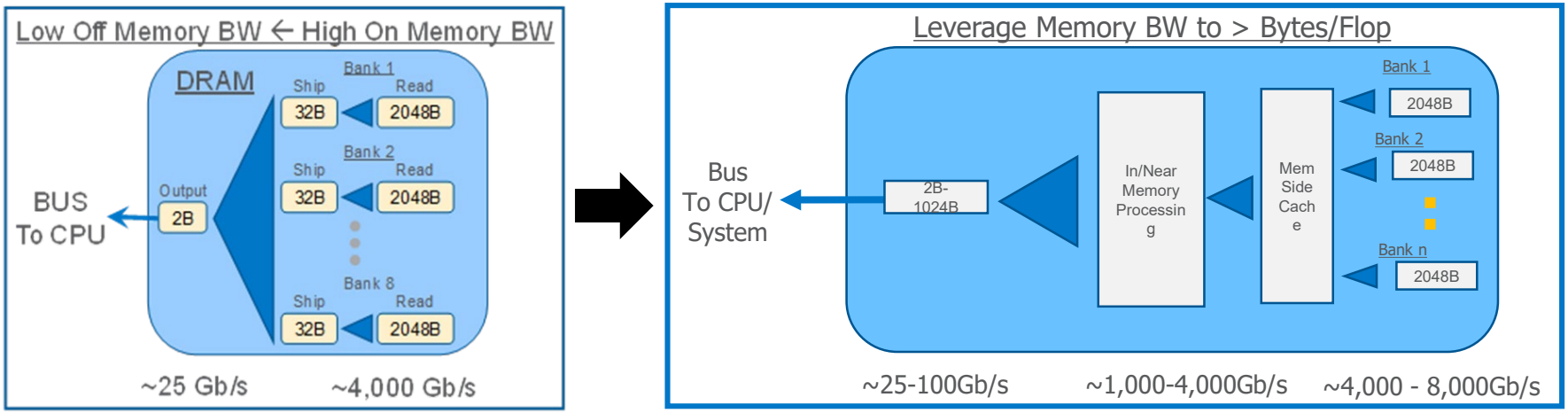
# Higher memory BW = higher power. Reduce the interconnect distance.



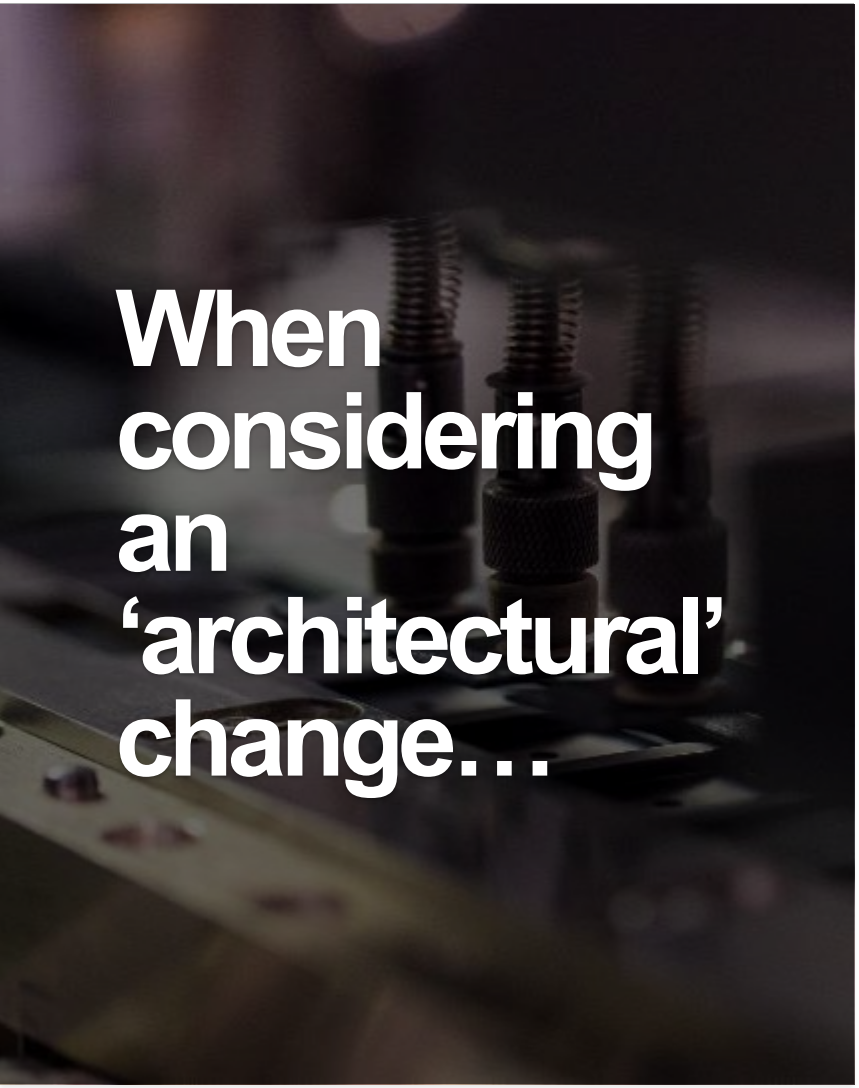

J. Hasler, B. Marr; " Finding a Roadmap to achieve large neuromorphic hardware systems"; Frontiers in Neuroscience, Sept 10, 2013
http://journal.frontiersin.org/article/10.3389/fnins.2013.00118/full

# Improved System Performance and Power Efficiency

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




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

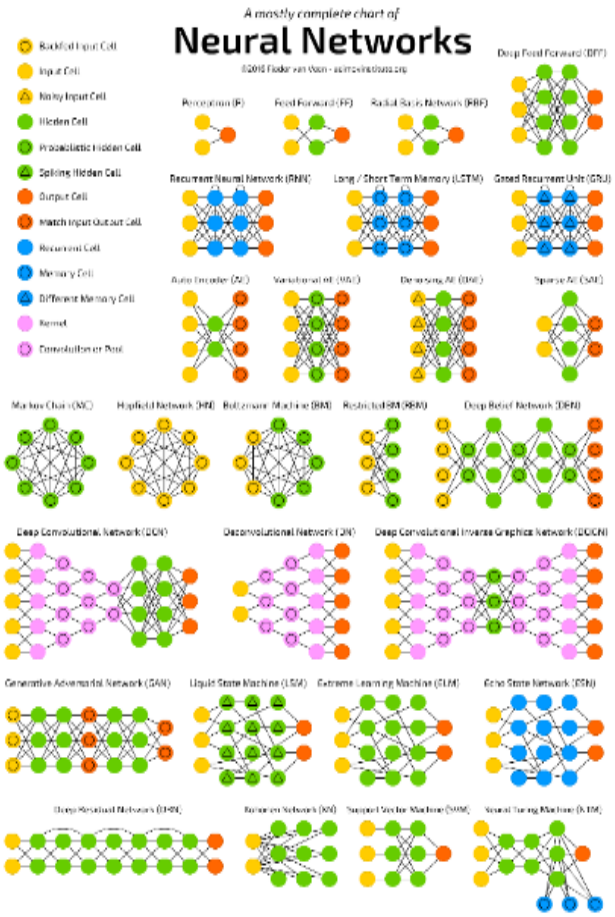
## “The architecture that wins is the one that’s EASIEST to program”

**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.

Micron



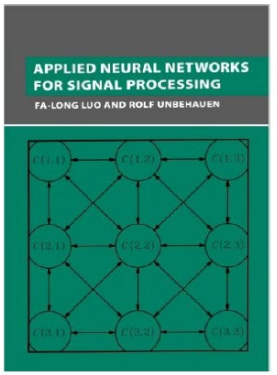

http:www.asimovinstitute.org/neural-network-zoo

# 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

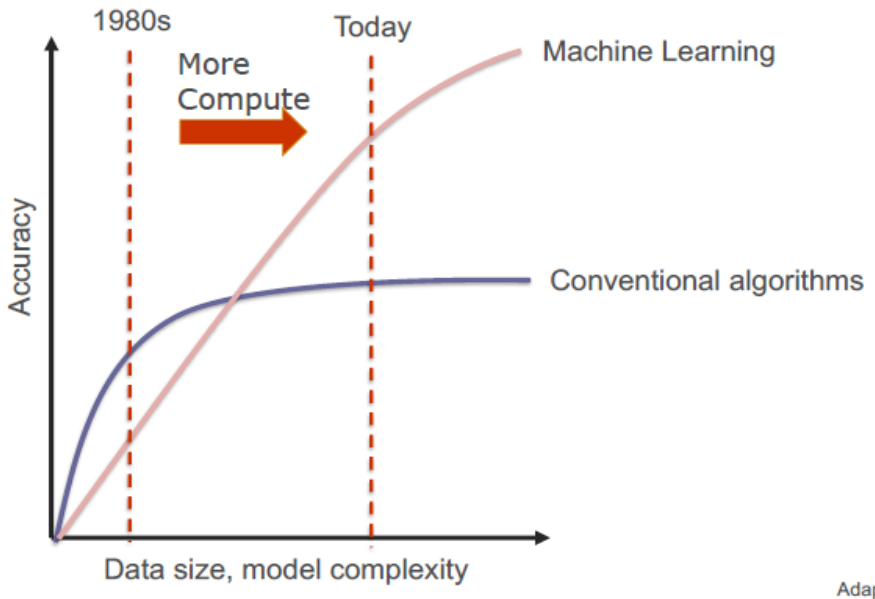
$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ax + by + cz \\ dx + ey + fz \\ gx + hy + iz \end{pmatrix}$$




CAMBRIDGE
UNIVERSITY PRESS

Micron

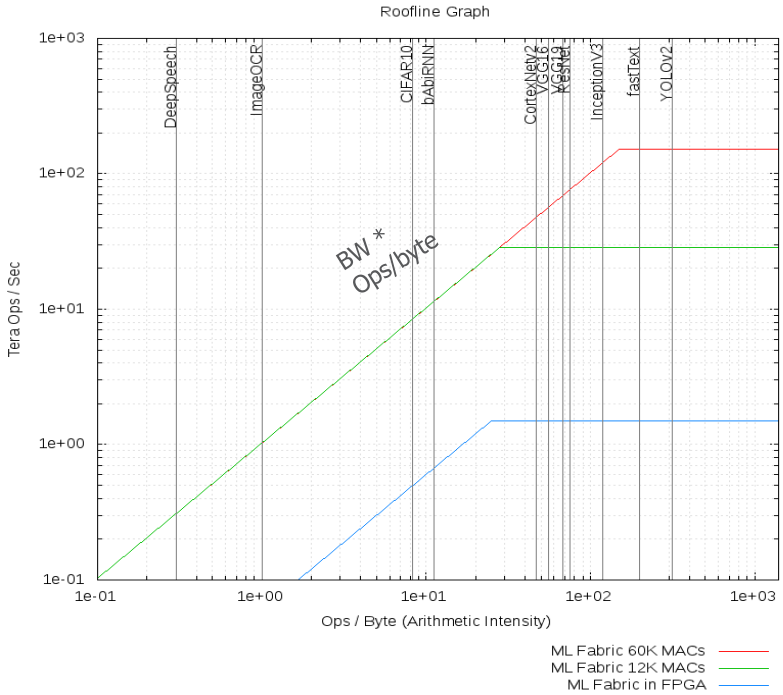
# AI/Machine Learning provides the capability to get more insight With the larger volumes of data.

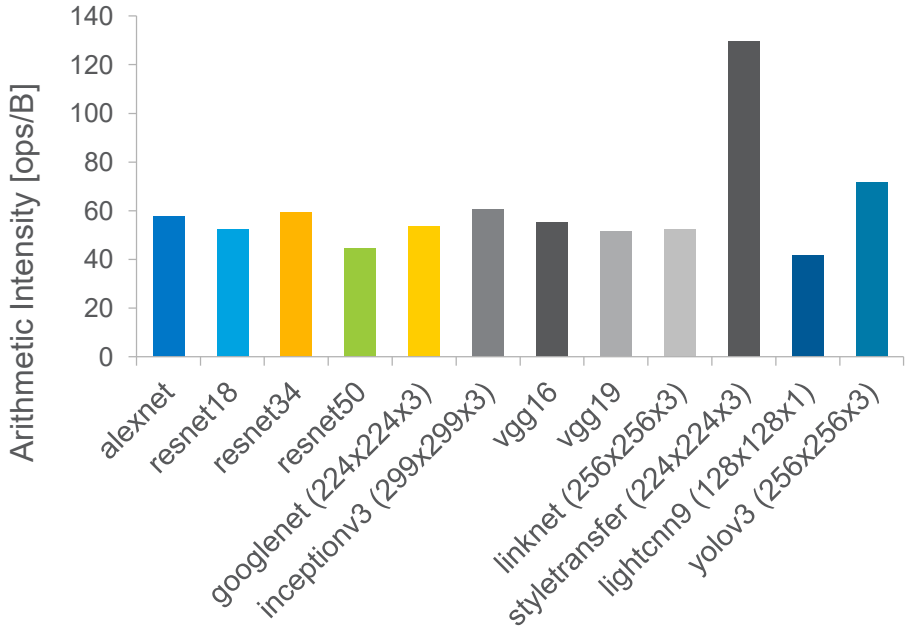



Kunle Olukotun ISCA'18 Keynote, June 15, 2018

Adapted from Jeff Dean
HotChips 2017

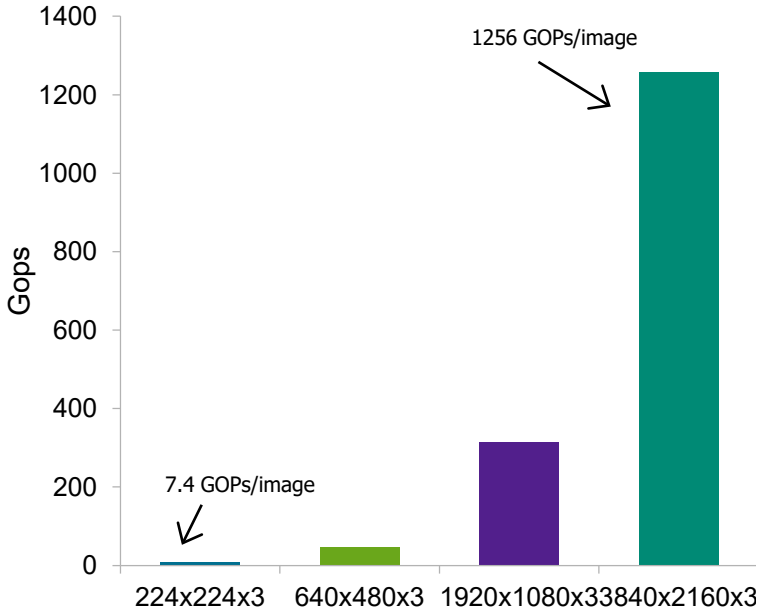
# The ratio of compute to memory BW is different for different networks.



# 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.

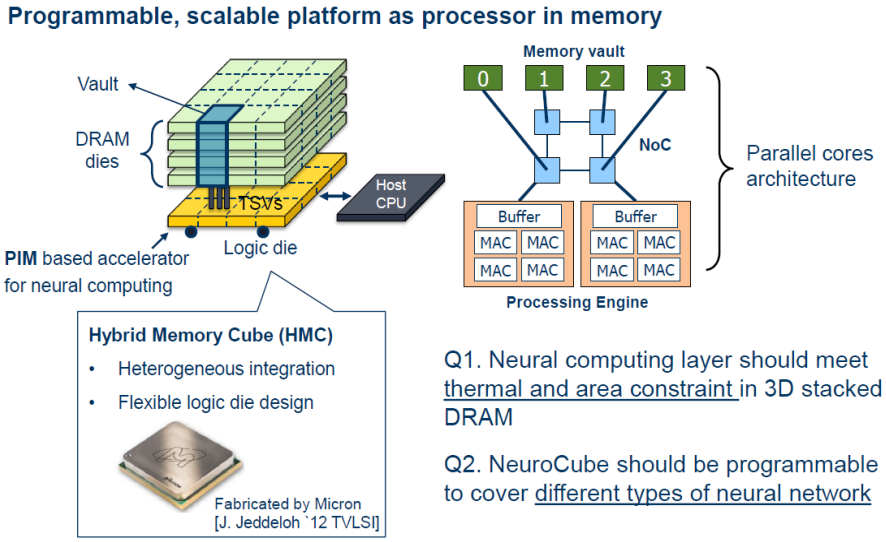



| resnet50 input image sizes w/Optimization | Gops | BW/Image (GB/s) | BW (30 Images/s) |
|---|---|---|---|
| 224x224x3 | 7.4 | 0.17 | 5.1 |
| 640x480x3 | 45.9 | 1.03 | 30.9 |
| 1920x1080x3 | 314.0 | 7.07 | 212.1 |
| 3840x2160x3 | 1256.3 | 28.3 | 849 |

| resnet50 input image sizes w/o optimization | Gops | BW/Image (GB/s) | BW (30 Images/s) |
|---|---|---|---|
| 224x224x3 | 7.4 | 0.37 | 11.1 |
| 640x480x3 | 45.9 | 2.3 | 69 |
| 1920x1080x3 | 314.0 | 15.7 | 471 |
| 3840x2160x3 | 1256.3 | 62.82 | 1884.6 |

# 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.

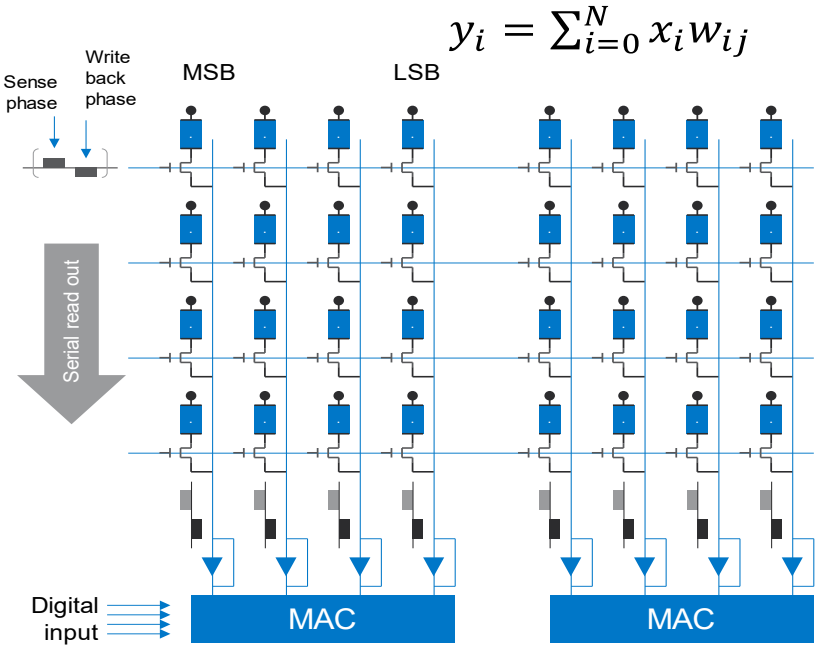



https://www.graphcore.ai/blog/why-is-so-much-memory-needed-for-deep-neural-networks

# 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




Micron®