

Memory Technologies and the Impact on System Architecture

Steve Pawlowski

Vice President, Advanced Computing Solutions

June 18th, 2019



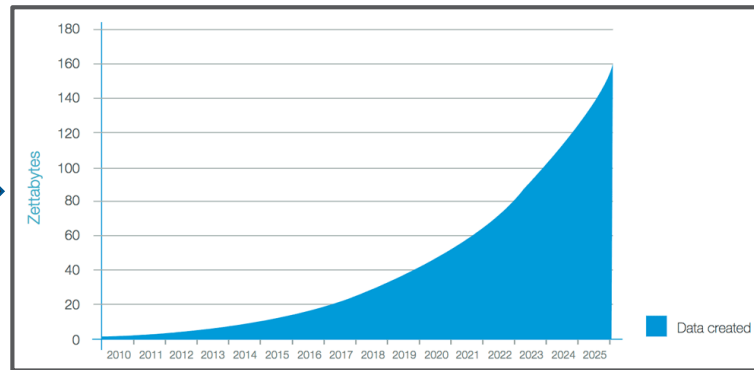
©2019 Micron Technology, Inc. All rights reserved. Information, products, and/or specifications are subject to change without notice. All information is provided on an "AS IS" basis without warranties of any kind. Statements regarding products, including regarding their features, availability, functionality, 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.

Estimated that roughly 5-8% of the data generated by 2025 will be analyzed.

Business demand for
Real Time decision
making

Rapid Data
Growth

More data to store and to process



<https://www.futuretimeline.net/blog/images/1140-optical-disk-10tb.gif>

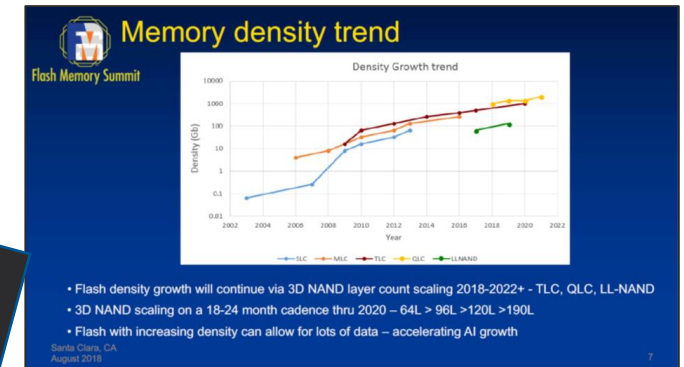
Today, images and image sequences (videos) make up about 80 percent of all corporate and public unstructured big data. As growth of unstructured data increases, analytical systems must assimilate and interpret images and videos as well as they interpret structured data such as text and numbers.

Images & Videos, really big data, Analytics Magazine, 11/2012

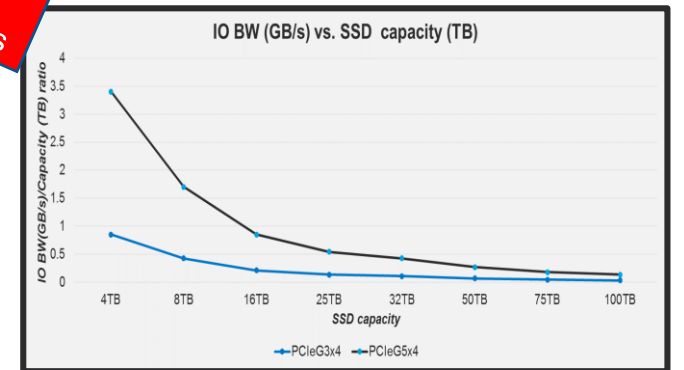
Market
demand

Arch
Bottlenecks

Technology innovation



Larger
SSD



Lagging data access rates

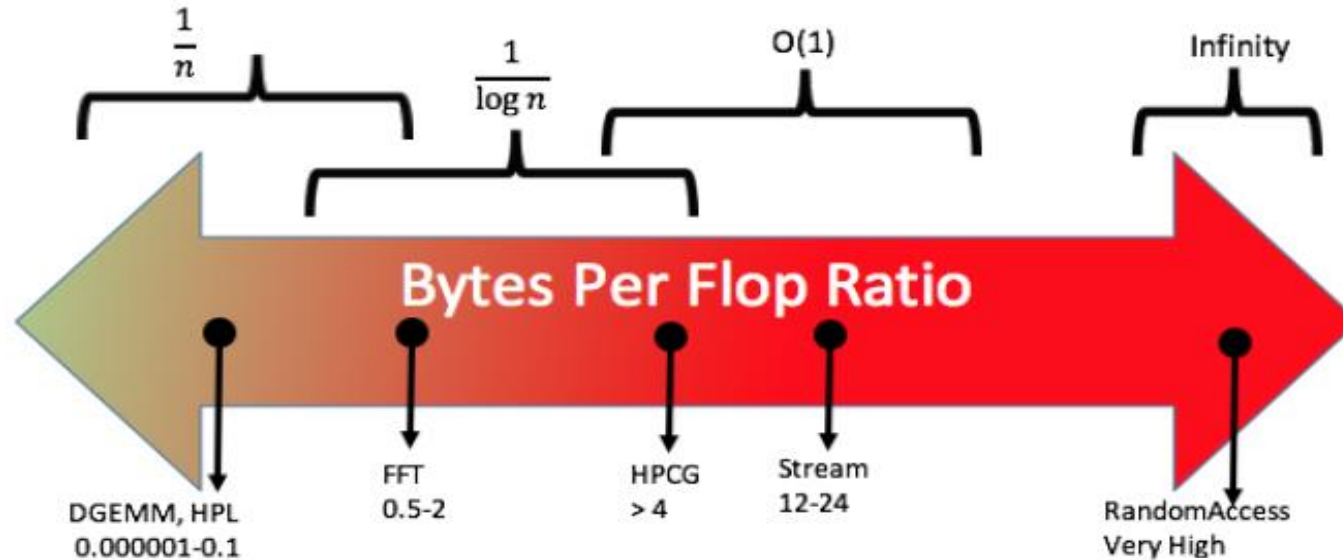
Challenge – Moving the data takes time and energy!



Source – Horst Simon 2013

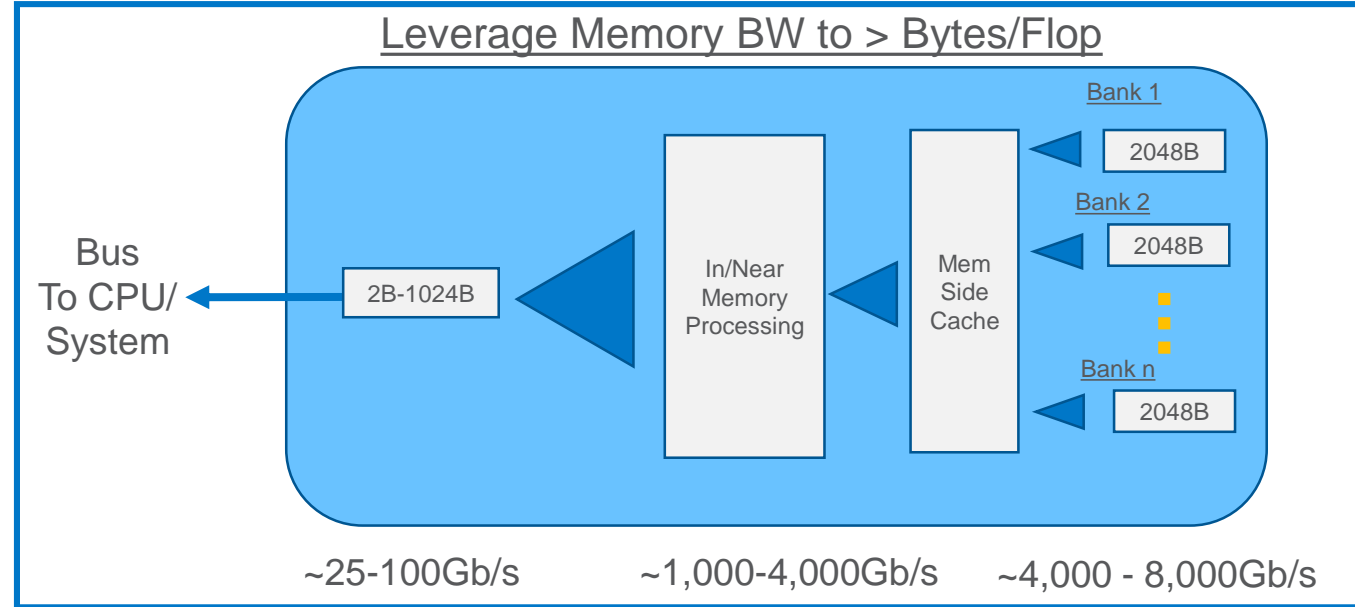
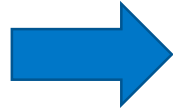
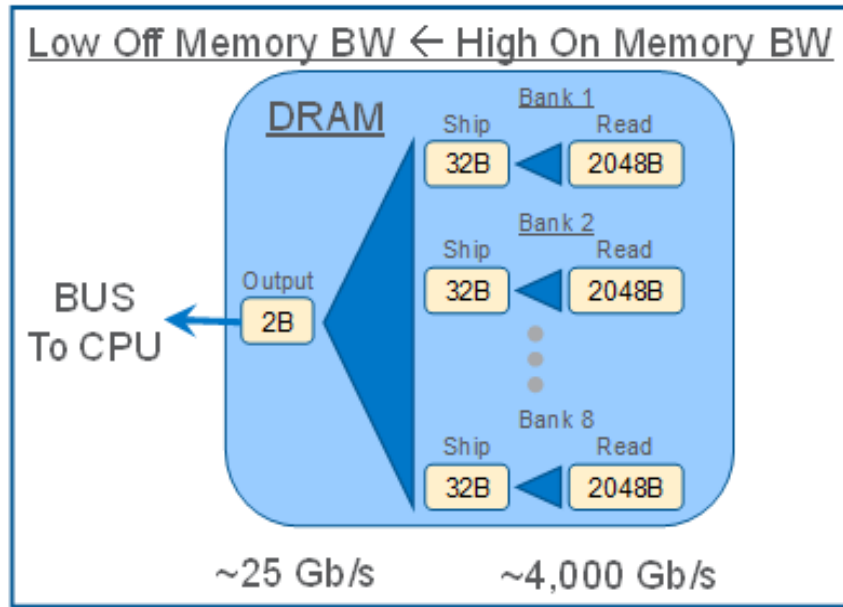
Distributed systems and processing closer to where the data is generated/captured is the best way to gain a orders of magnitude improvement in performance and efficiency

Many real world applications (not all mind you) are memory-bound on standard HW



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

The bandwidth and parallelism is available in memory.

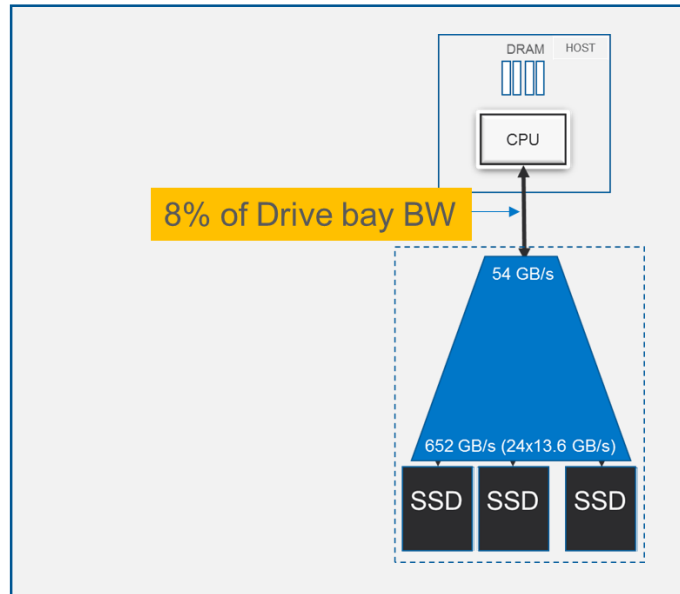


Bytes/FLOP could improve by over 10x

The opportunity is deciding the type of computation to put near/in memory

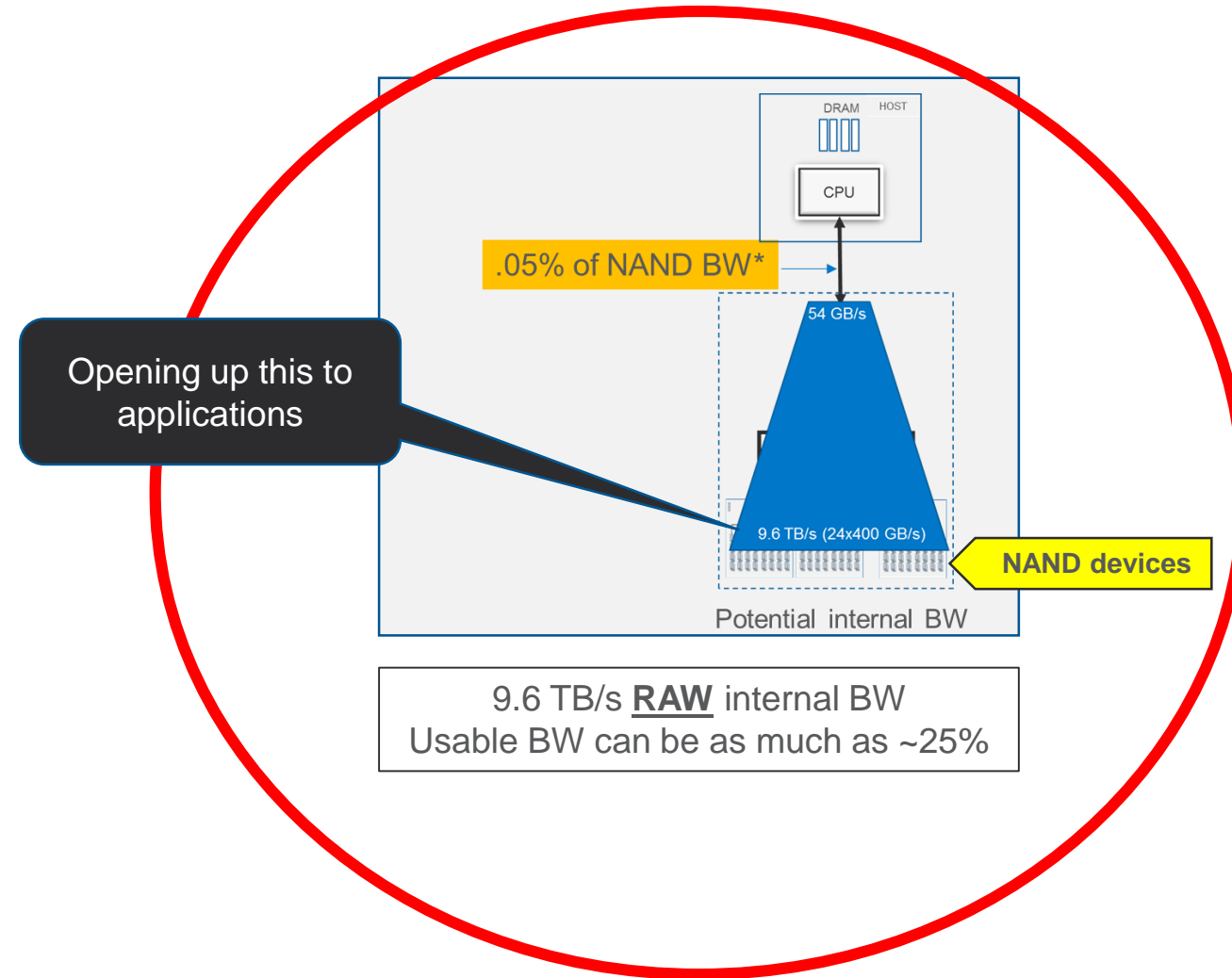
In Storage...Utilize the Media device IO BW

There is an even larger BW funnel seldom discussed



652 GB/s
IO BW at SSD connector

Current SSD architectures are focusing on capacity scaling within IO constraints



9.6 TB/s RAW internal BW
Usable BW can be as much as ~25%

*Note: in both cases there are 24 drives per JBOF, identical PCIe G5 NVMe interface

