The resurgence of Shared Memory Systems

Steve Pawlowski

Corporate Vice President, Advanced Computing Solutions

March 2023

Some things to consider …

Solving the energy efficiency problem means that one must address data movement

"… (system) profiling revealed that 25-35% of all CPU time was spent just moving bytes around… If data movement were faster, more work could be done on the same processors.

- Richard L. Sites; Computer Architecture Today Blog, ACM SIGARCH, December 19, 2022

The industry will continue to innovate on the current computing paradigm. This should be a key focus while looking for the next 'BIG' thing.

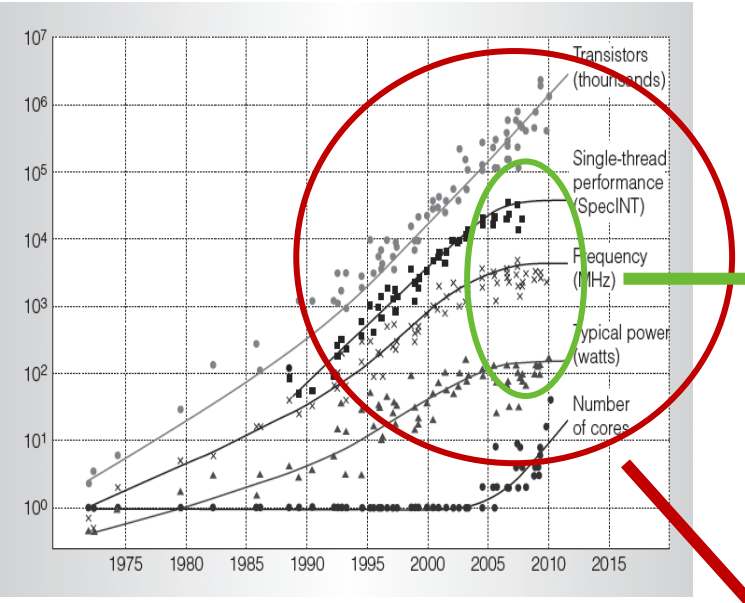
- System improvements can yield nearly two orders of magnitude efficiency improvement
- 40 years of SW will not be changed overnight. Need to execute existing code
  - Amdahl's Law reigns – sequential performance is STILL important
- It takes "Two Olympic Cycles" for SW to 'catch-up' with HW.
- Any changes to the computing model need investments in Workforce Development.
- General Purpose Computing as we know it today will still be the dominate architecture 20 years from now.

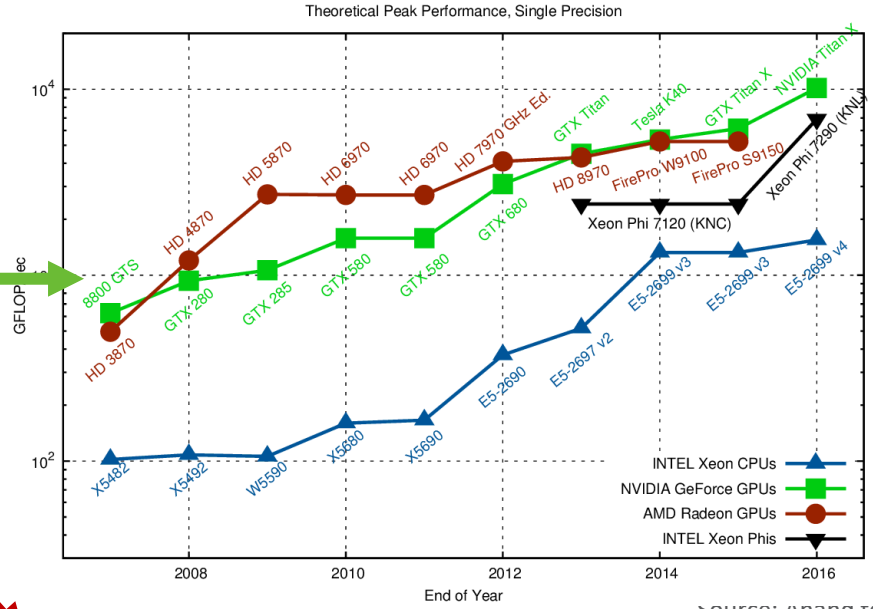
In general, DRAM is a hard technology to beat in terms of performance and activation energy.

Comparison of various emerging memory technologies

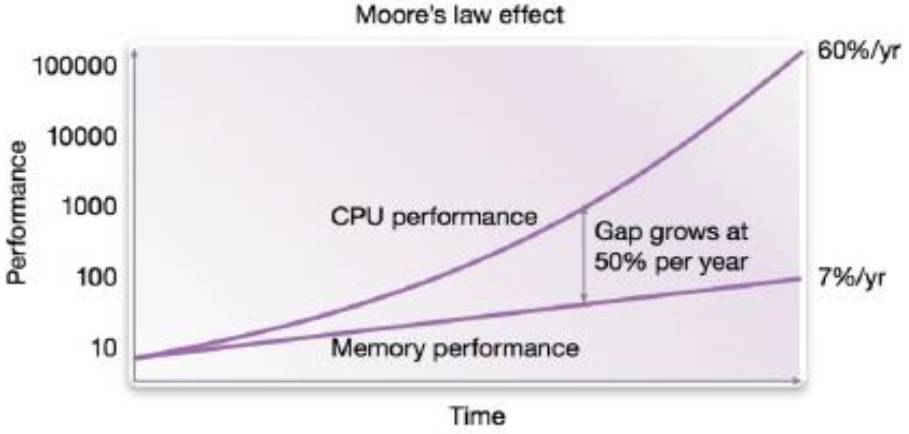
	DRAM	STTRAM	PCM/ 1T1R	Cross Point RRAM	NAND
Read Latency	20ns	~50ns	~100ns-200ns	~100ns-200ns	~10us
Write Latency	20ns	~50ns	~1us	~1us	~10us
Read Endurance	>1e15	$>10^{11}$	$>10^{7}$	$>10^{7}$	$>10^{7}$
Write Endurance	>1e15	$>10^{11}$	$>10^{6}$	$>10^{6}$	2K-100K
Write/Read Energy/Bit	<10pJ/bit	~25pJ/bit	~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

Moore's Law and Dennard's Scaling Law reductions are the reason we're here today





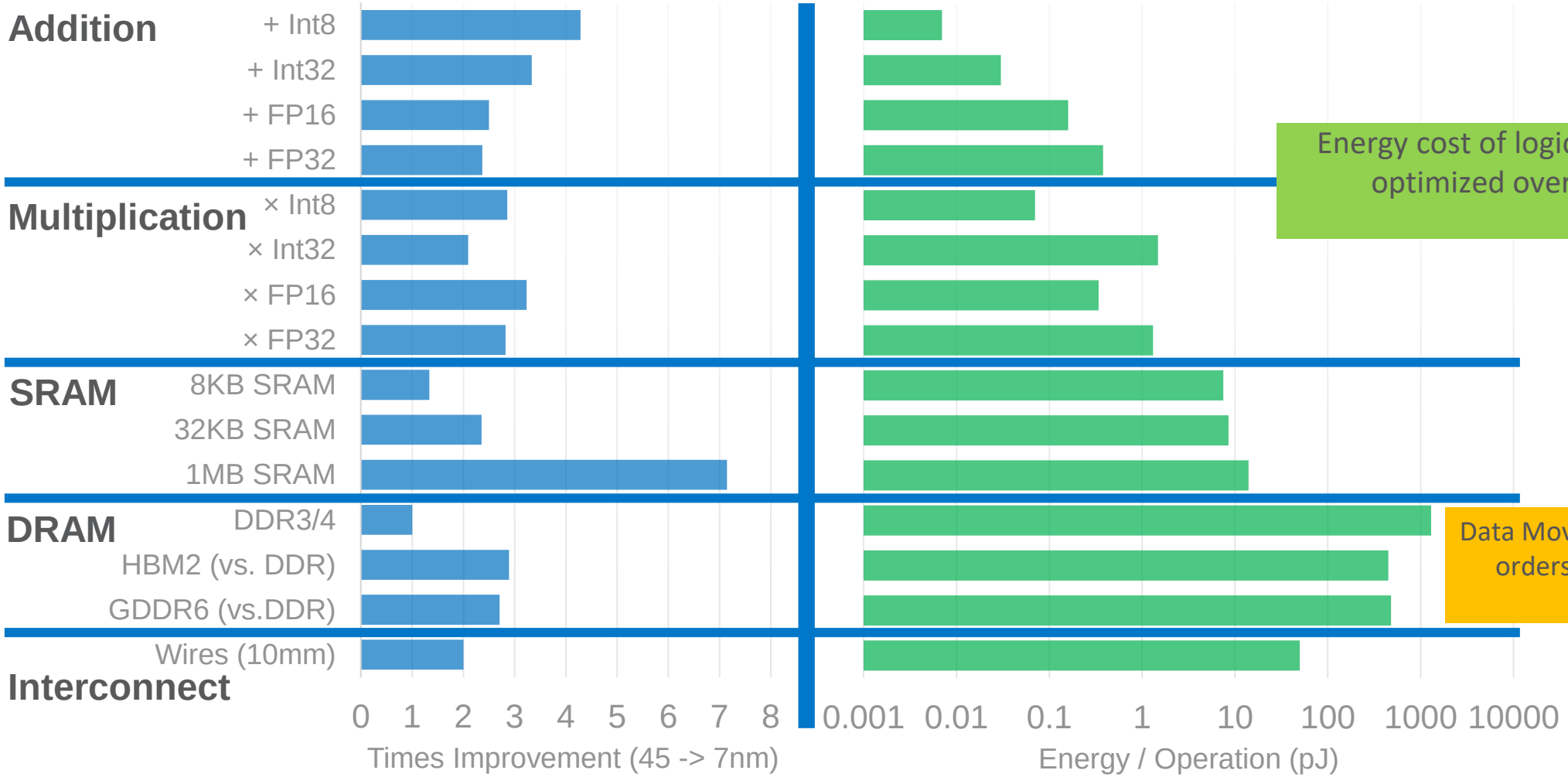
Source: Anand tech.com



http://www.extremetech.com/wp-content/uploads/2014/07/140364245678419.jpg

Moving Data Dominates Energy costs

Energy numbers from Jouppi, et. al. “Ten Lessons From Three Generations Shaped Google’s TPUv4i” and Keckler et al. “GPUs and the Future of Parallel Computing”



Energy cost of logic has been well optimized over the years.

Data Movement energy costs are 2 – 3 orders of magnitude higher than compute.

November 2022: The TOP 10 Systems

Source: Jack Dongara, “A Not So Simple Matter of Software”, SC’22 Keynote, 2021 ACM A.M Turing Lecture

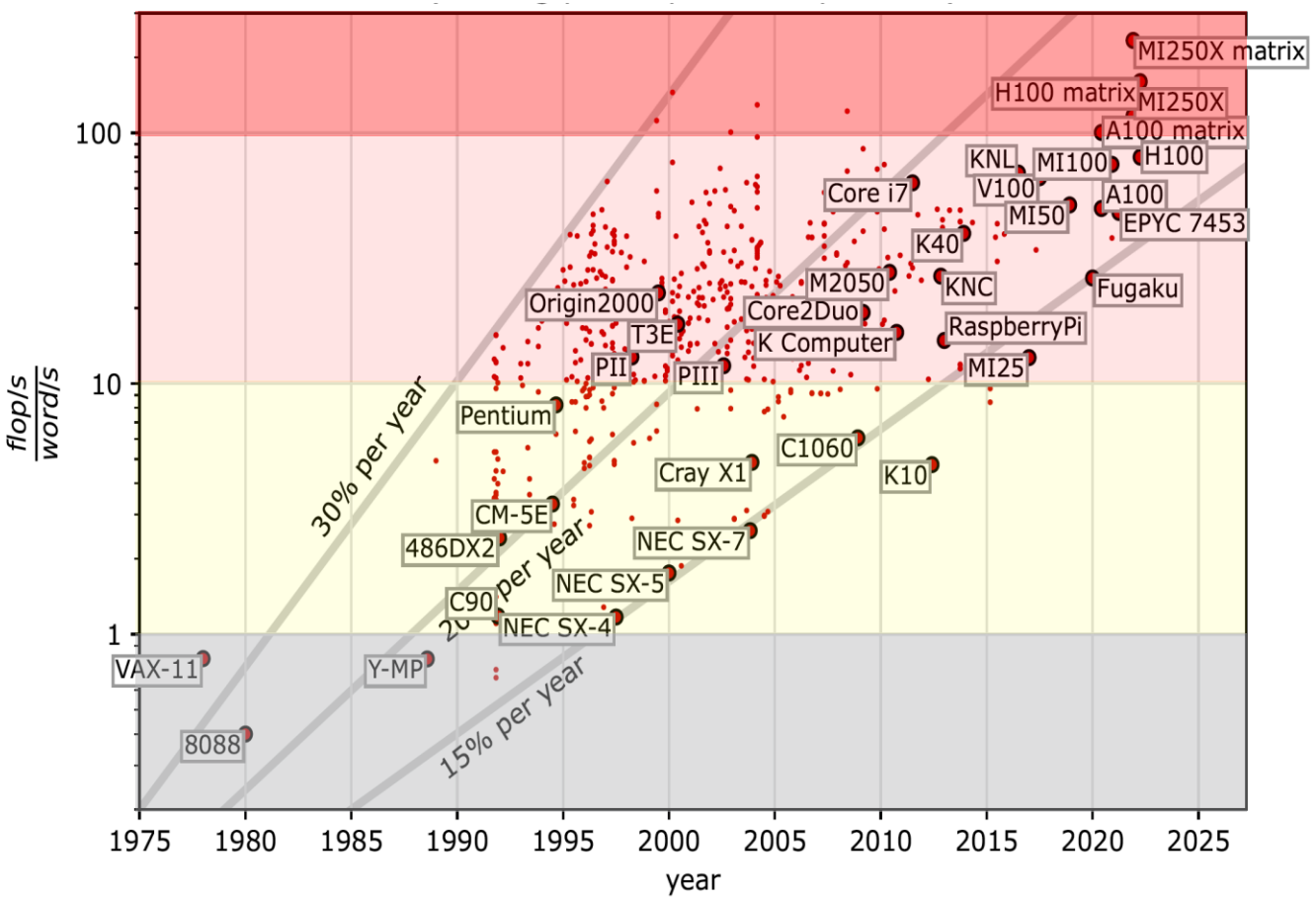
Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	GFlops/ Watt
1	DOE / OS Oak Ridge Nat Lab	Frontier, HPE Cray Ex235a, AMD 3rd EPYC 64C, 2 GHz, AMD Instinct MI250X, Slingshot 10	USA	7,733,248	1,102	65	21.1	52.2
2	RIKEN Center for Computational Science	Fugaku, ARM A64FX (48C, 2.2 GHz), Tofu D Interconnect	Japan	7,299,072	442.	82	29.9	14.8
3	EuroHPC /CSC	LUMI, HPE Cray EX235a, AMD 3rd EPYC 64C, 2 GHz, AMD Instinct MI250X, Slingshot 10	Finland	1,268,736	304.	72	2.94	52.3
4	EuroHPC/CINECA	BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 (108C), Quad-rail NVIDIA HDR100	Italy	1,463,616	175.	68	5.6	31.1
5	DOE / OS Oak Ridge Nat Lab	Summit, IBM Power 9 (22C, 3.0 GHz), NVIDIA GV100 (80C), Mellonox EDR	USA	2,397,824	149.	74	10.1	14.7
6	DOE / NNSA L Livermore Nat Lab	Sierra, IBM Power 9 (22C, 3.1 GHz), NVIDIA GV100 (80C), Mellonox EDR	USA	1,572,480	94.6	75	7.44	12.7
7	National Super Computer Center in Wuxi	Sunway TaihuLight, SW26010 (260C), Custom Interconnect	China	10,649,000	93.0	74	15.4	6.05
8	DOE / OS NERSC - LBNL	Perlmutter HPE Cray EX235n, AMD EPYC 64C 2.45GHz, NVIDIA A100, Slingshot 10	USA	706,304	64.6	71	2.59	27.4
9	NVIDIA Corporation	Selene NVIDIA DGX A100, AMD EPYC 7742 (64C, 2.25GHz), NVIDIA A100 (108C), Mellanox HDR	USA	555,520	63.4	80	2.64	23.9
10	National Super Computer Center in Guangzhou	Tianhe-2A NUDT, Xeon (12C) , MATRIX-2000 (128C) + Custom Interconnect	China	4,981,760	61.4	61	18.5	3.32

Performance/BW mismatch in Numerical Computations.

- Data movement has a big impact
- Performance comes from balancing floating point execution (Flops/sec) with memory->CPU transfer rate (Words/sec)
  - "Best" balance would be 1 flop per word-transfered
- Today's systems are close to 100 flops/sec per word-transferred
  - Imbalanced: Over provisioned for Flops

Machine Balance

Ratio of Fl Pt Ops per Data Movement over Time



Graph from Mark Gates

Plot for 64-bit floating point data movement & operations

(Bandwidth from CPU or GPU memory to registers)

Source: Jack Dongara, "A Not So Simple Matter of Software", SC'22 Keynote, 2021 ACM A.M Turing Lecture

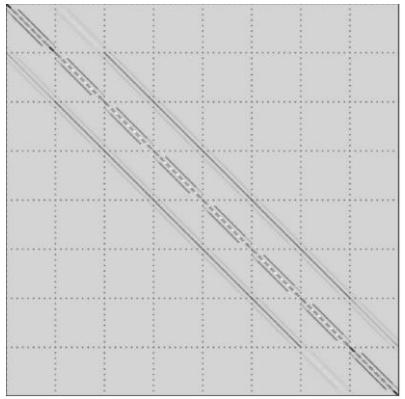
Performance and Benchmarking Evaluation Tools

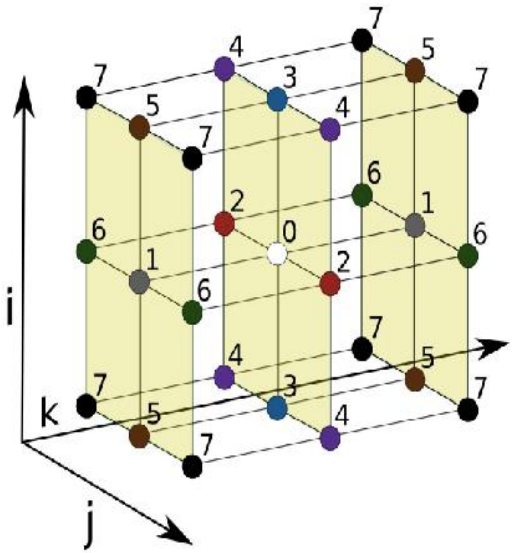
- Linpack Benchmark - Longstanding benchmark started in 1979
  - Lots of positive features; easy to understand and run; shows trends
- However, much has changed since 1979
  - Arithmetic was expensive then and today it is over-provisioned and inexpensive
- Linpack performance of computer systems is no longer strongly correlated to real application performance
  - Linpack benchmark based on dense matrix multiplication
- Designing a system for good Linpack performance can lead to design choices that are wrong for today's applications

Source: Jack Dongara, "A Not So Simple Matter of Software", SC'22 Keynote, 2021 ACM A.M Turing Lecture

HPCG Results; The Other Benchmark

- High Performance Conjugate Gradients (HPCG)
- Solves *Ax=b*, *A* large, sparse, *b* known, *x* computed
- An optimized implementation of PCG contains essential computational and communication patterns that are prevalent in a variety of methods for discretization and numerical solution of PDEs





27-point stencil operator

- Patterns:
  - Dense and sparse computations
  - Dense and sparse collectives
  - Multi-scale execution of kernels via MG (truncated) V cycle.
  - Data-driven parallelism (unstructured sparse triangular solves)
- Strong verification (via spectral properties of PCG)

hpcg-benchmark.org With Piotr Luszczek and Mike Heroux

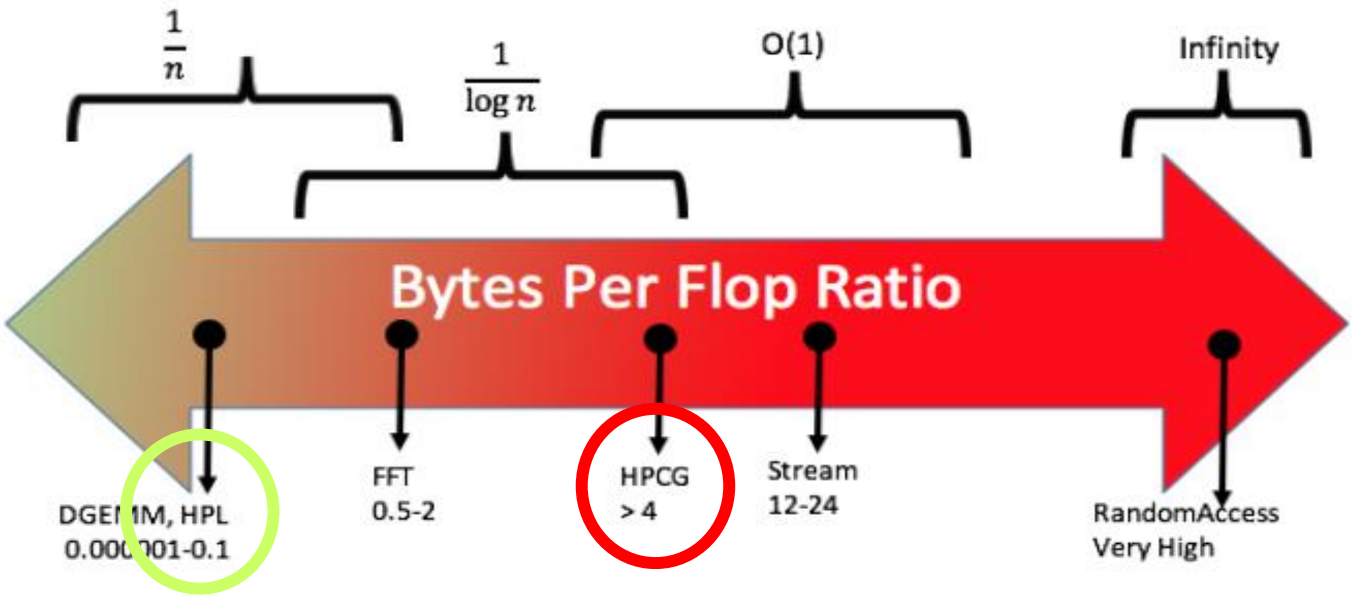
Slide Source: Jack Dongara, “A Not So Simple Matter of Software”, SC’22 Keynote, 2021 ACM A.M Turing Lecture

HPCG Top 10, November 2022

Slide Source: Jack Dongara, “A Not So Simple Matter of Software”, SC’22 Keynote, 2021 ACM A.M Turing Lecture

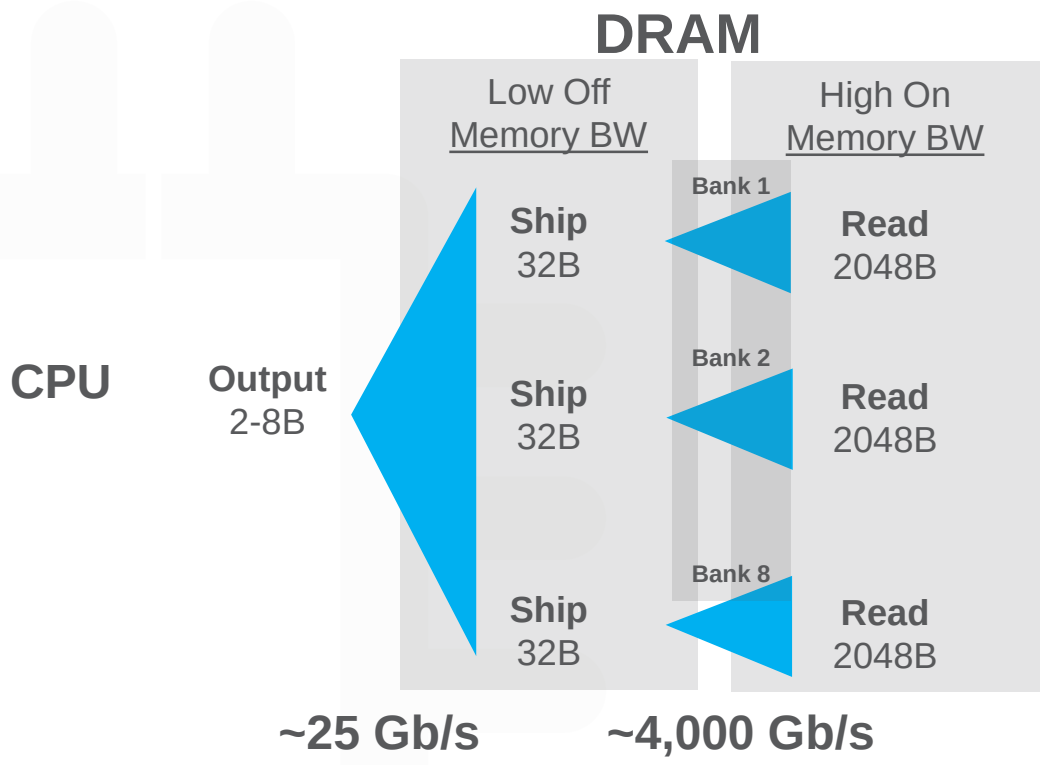
Rank	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	Fraction of Peak
1	RIKEN Center for Computational Science **Japan**	**Fugaku**, Fujitsu A64FX 48C 2.2GHz, Tofu D, Fujitsu	7,630,848	442	2	16.0	**3.0%**
2	DOE/SC/ORNL **USA**	**Frontier,** HPE Cray Ex235a, AMD 3rd EPYC 64C, 2 GHz, AMD Instinct MI250X, Slingshot 10	8,730,112	1,102	1	14.1	**0.8%**
3	EuroHPC/CSC **Finland**	**LUMI**, HPE Cray EX235a, AMD Zen-3 (Milan) 64C 2GHz, AMD MI250X, Slingshot-11	2,174,976	304	3	3.41	**[illegible]%**
4	DOE/SC/ORNL **USA**	**Summit**, AC922, IBM POWER9 22C 3.7GHz, Dual-rail Mellanox FDR, NVIDIA Volta V100, IBM	2,414,592	149	5	2.93	**[illegible]%**
5	EuroHPC/CINECA **Italy**	Leonardo, BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 40 GB, Quad-rail NVIDIA HDR100 Infiniband	1,463,616	175	4	2.57	**1.0%**
6	DOE/SC/LBNL **USA**	**Perlmutter**, HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10	761,856	70.9	8	1.91	**2.0%**
7	DOE/NNSA/LLNL **USA**	**Sierra**, S922LC, IBM POWER9 20C 3.1 GHz, Mellanox EDR, NVIDIA Volta V100, IBM	1,572,480	94.6	6	1.80	**1.4%**
8	NVIDIA **USA**	**Selene**, DGX SuperPOD, AMD EPYC 7742 64C 2.25 GHz, Mellanox HDR, NVIDIA Ampere A100	555,520	63.5	9	1.62	**2.0%**
9	Forschungszentrum Juelich (FZJ) **Germany**	**JUWELS Booster Module**, Bull Sequana XH2000 , AMD EPYC 7402 24C 2.8GHz, Mellanox HDR InfiniBand, NVIDIA Ampere A100, Atos	449,280	44.1	12	1.28	**1.8%**
10	Saudi Aramco **Saudi Arabia**	**Dammam-7**, Cray CS-Storm, Xeon Gold 6248 20C 2.5GHz, InfiniBand HDR 100, NVIDIA Volta V100, HPE	672,520	22.4	20	0.88	**1.6%**

For the more real world numerical applications, need from 100x-300x Reduction in FLOP/BW over current solutions



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

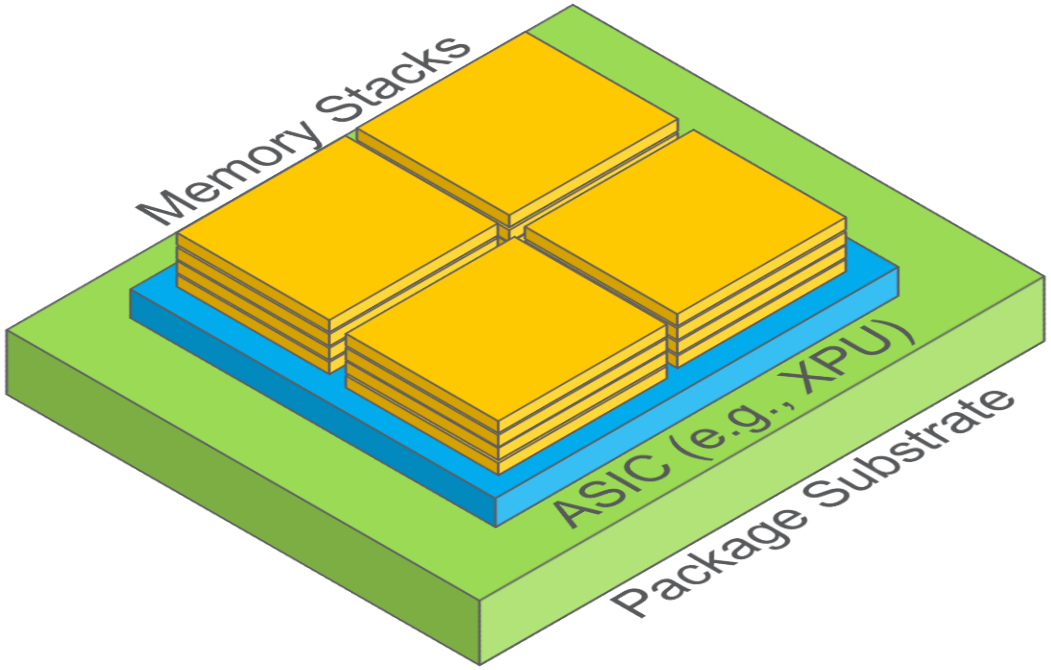
The path memory data takes to its destination…



Narrow busses are driven by system/package cost, power and standardization.

What if…. We revisit the Hybrid Memory Cube (HMC) concept with advanced packaging innovations

- 10s of TB/s at significantly reduced energy/bit over state of the art
- 3D-stacked memory and logic for optimized bandwidth and energy efficiency
- Increased bandwidth at lower power enabled by hybrid bonding
- Significantly greater number of connections between logic and memory
- Co-optimized logic and memory architectures and designs

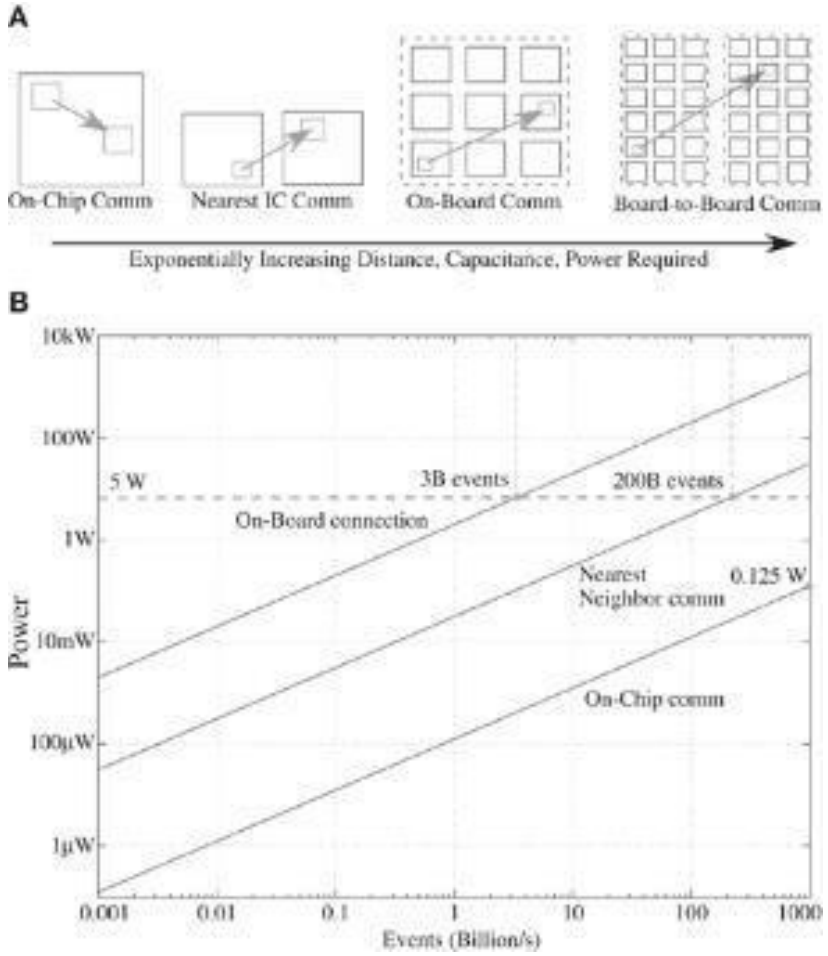


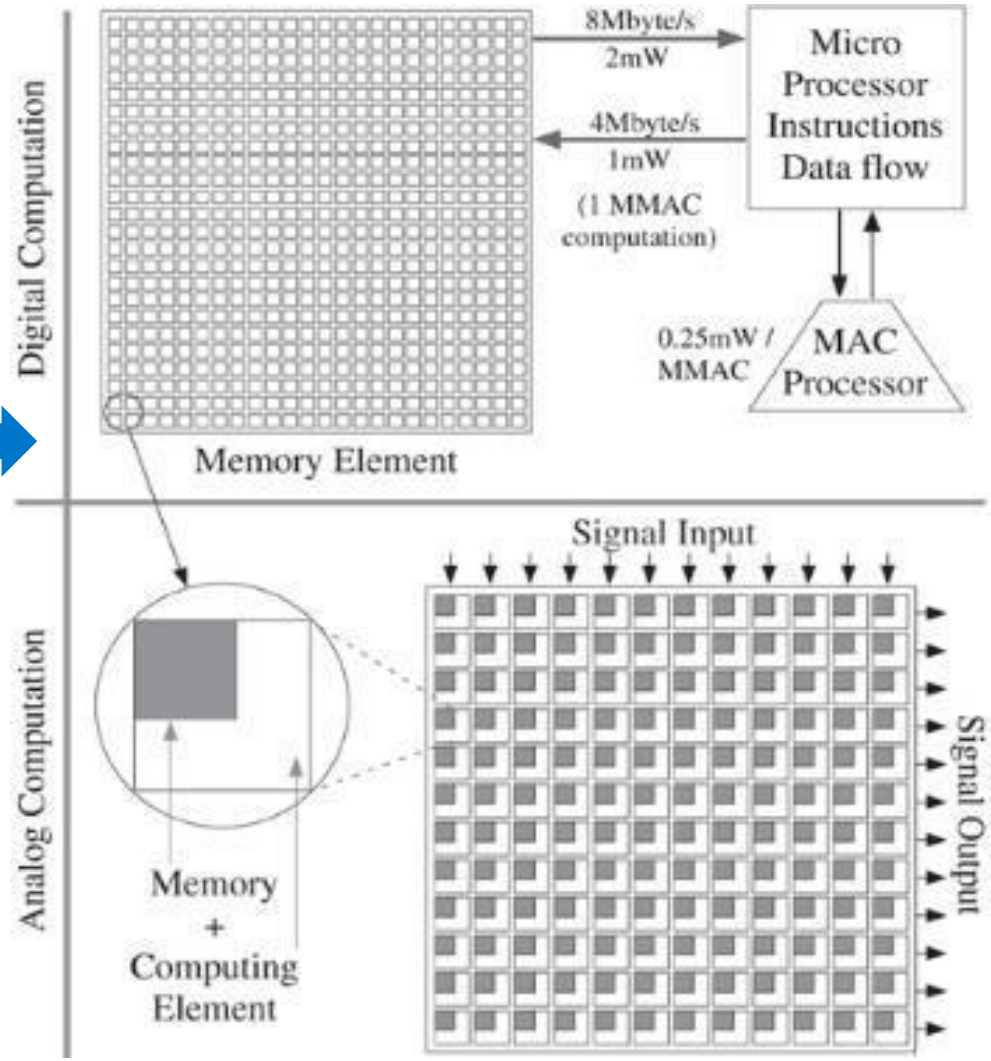
Stacking RAM w/logic reverses the FLOP/BW mismatch (The example assumes GPT-3, batch size of 1, 3.5ms latency)

	Design Target for GPT-3 (example)	HBM	Memory-on-Logic Optimized Solution
Memory Bandwidth	100TB/s	0.82TB/s	>10x HBM
Est. Energy/bit	1.5pJ/b	2.75pJ/b	0.75 - 1.00pJ/b
User Capacity Range @ 100 TB/s	~350GB	3900GB (32GB/stk) **~11x Extra capacity**	352GB (32GB/stk) **1X capacity**
Memory stacks for 350GB @ 100TB/s (min)	11 @ 32GB	121	11
Memory System Power at >350GB / >50TB/s	Target: <= 800W	~2200W	660W - 880W

With a change in the memory/logic relationship, an improvement in energy efficiency can be achieved.

Co-Locating Memory and computing for highest efficiency.

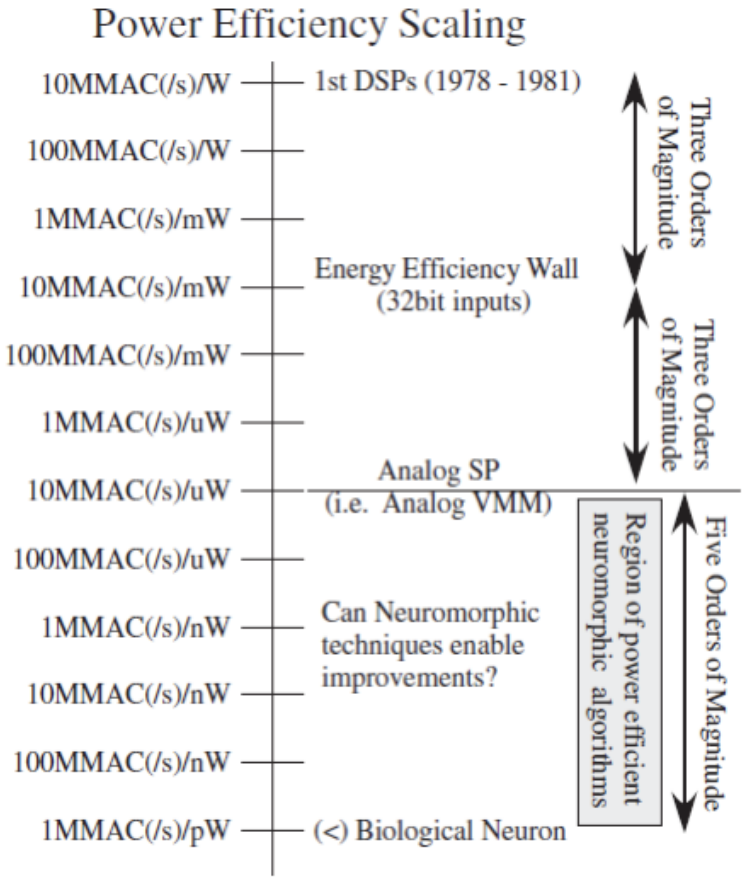




J. Hasler, B. Marr; " Finding a Roadmap to achieve large neuromorphic hardware systems"; Frontiers in Neuroscience, Sept 10, 2013
http://journal.frontiersin.org/article/10.3389/fnins.2013.00118/full

Approaching the efficiency of biological systems…

There is roughly a five order of magnitude in Energy efficiency gap that needs to be closed



Massively parallel (slow) compute engines where computation's occurring at the data in the Analog domain.

We are 'around' here.

This is our challenge

Source: Hasler,and Marr, "Finding a roadmap to achieve large neuromorphic hardware systems", Frontiers in Neuroscience, Sept. 2013.

Micron®