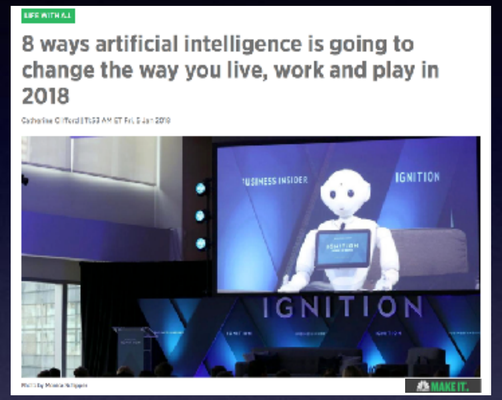
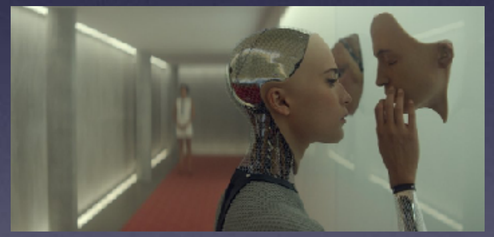
Using Cellphone Technology to Build **AI** Servers

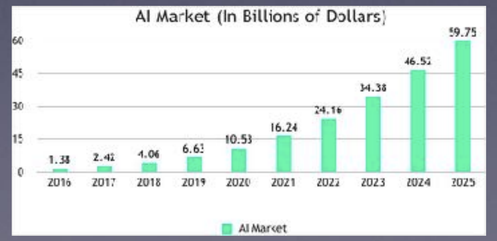
Peter Hsu, PhD
pete2222@mac.com

Presented at Barcelona Supercomputer Center, Spain
25 June 2018

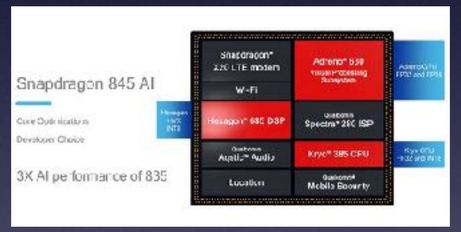
Energy is Real Limit of **AI**





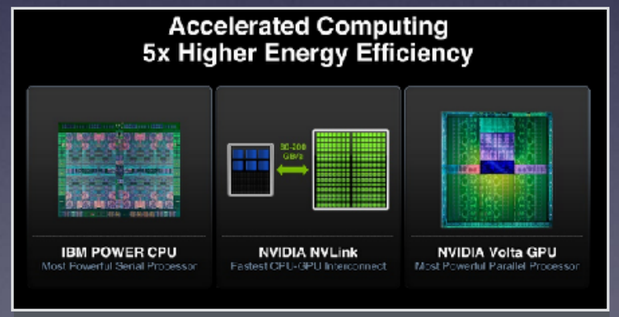


Qualcomm makes AI priority second only to 5G



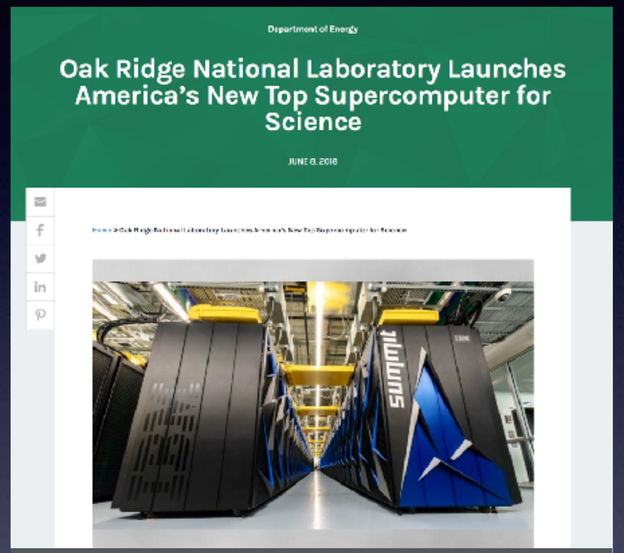
30 GFLOPS / Watt (32-bit Float)

AMD Leverages EPYC And Vega To Cram 1 PetaFLOP Supercomputer Into A Single Server Rack



15.1 GFLOPS / Watt (64-bit Float)

IBM's Power9-Based AC922 System Designed for AI Workloads



200 PFLOPS @ 13.3 MW

America's new supercomputer beats China's fastest machine to take title of world's most powerful

China's Next Supercomputer May Spoil America's Plans to Retake TOP500 Crown

Michael Feldman | January 29, 2018 16:00 CET

IBM AC922

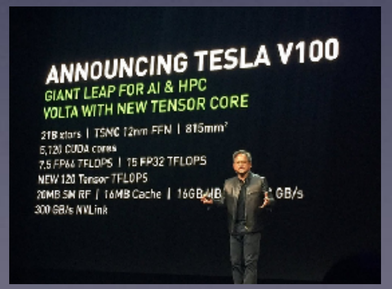
MAY 8, 2018 @ 12:09 PM 3,047

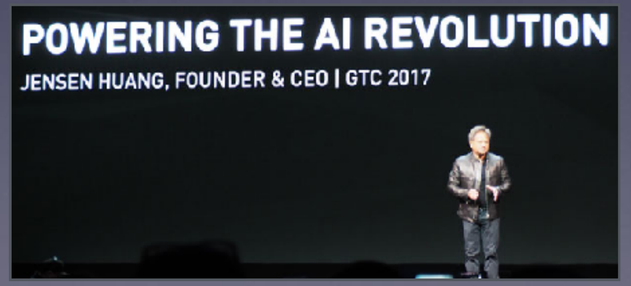
2 Free Issues of Forbes

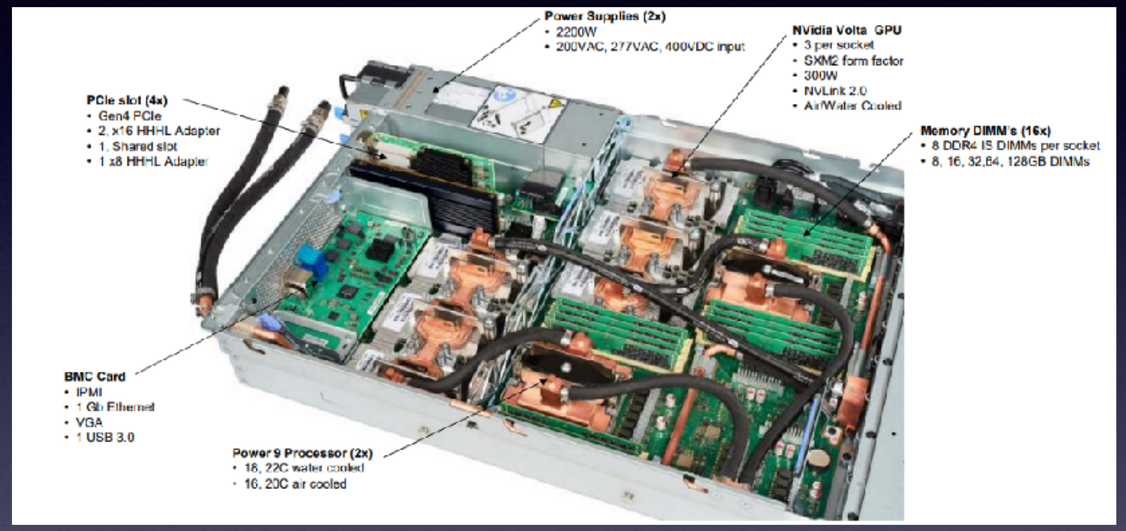
IBM Power Systems For AI and Big Data: Aimed at the Enterprise

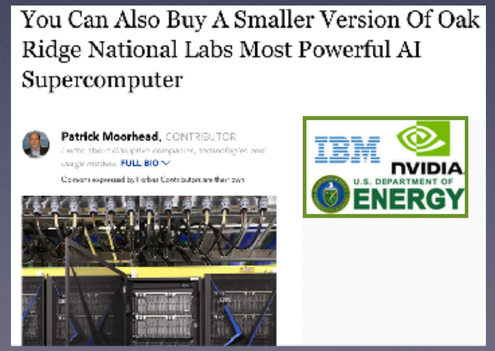
Tirias Research

Ad closed by Google









MAY 8, 2018 @ 12:09 PM 3,047

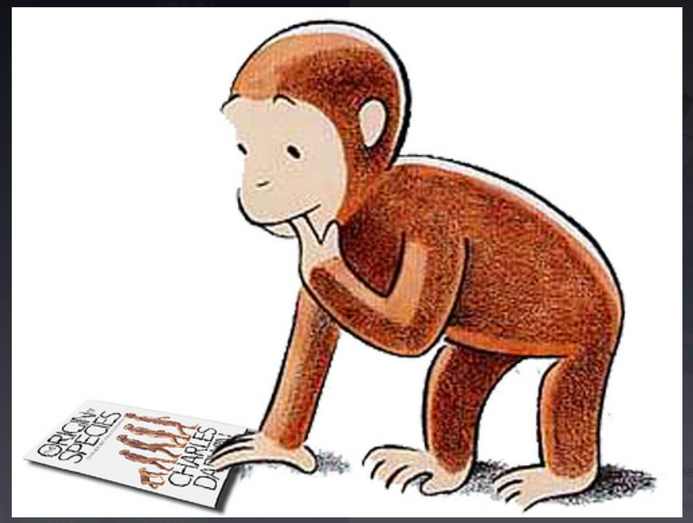
2 Free Issues of Forbes

IBM Power Systems For AI and Big Data: Aimed at the Enterprise

Tirias Research

Ad closed by Google

IBM AC922



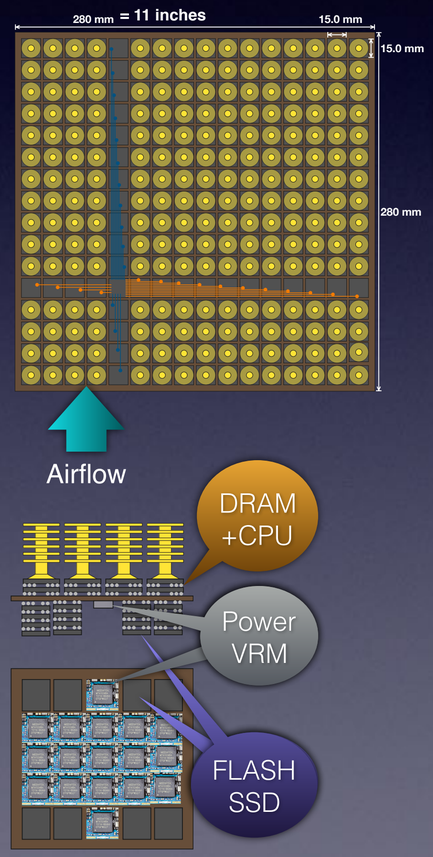
- Where does the energy go?
- How efficient is it really?
- Can we do better?

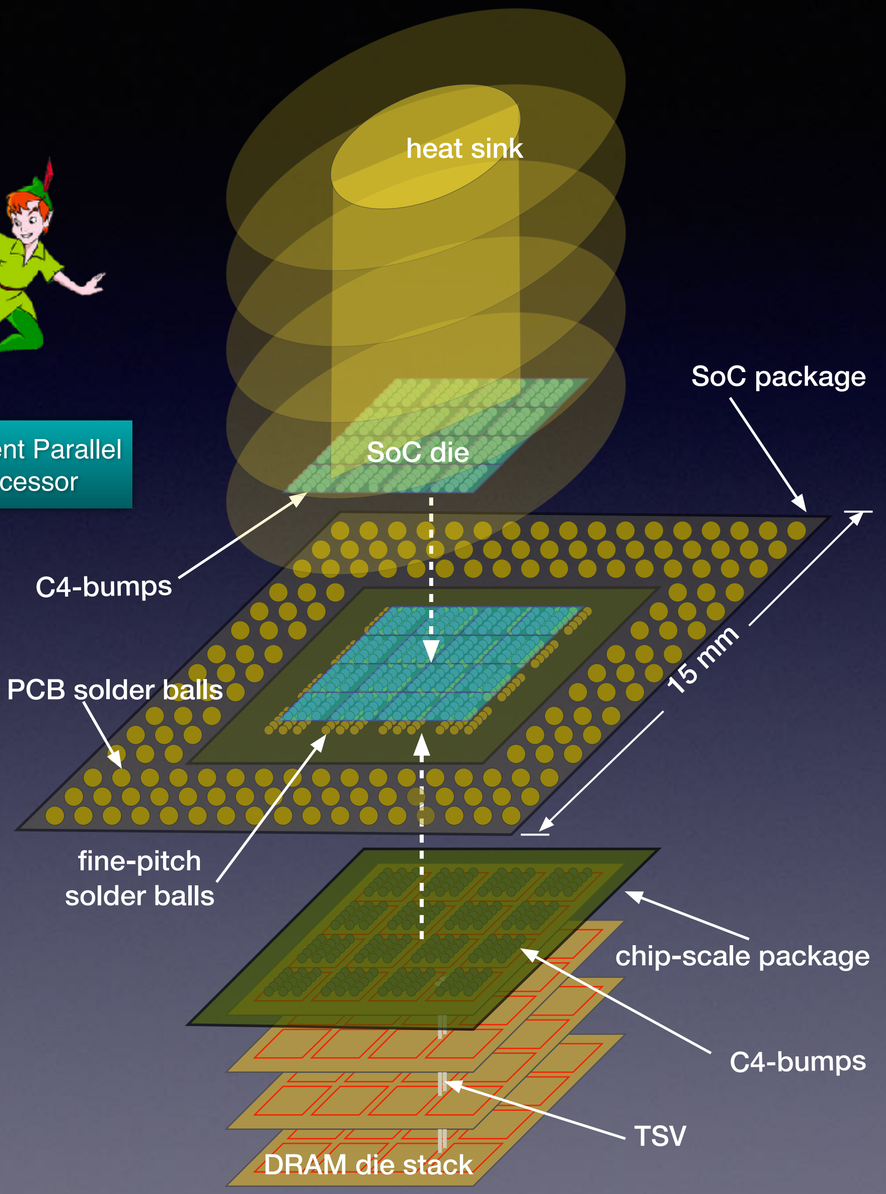
Summary

- 2× — near-data processing
- 1.6× — SoC/DRAM 3D layout/package co-design
- 1.6× — vector accumulator
- 1.4× — best consumer electronics process
- 1.4× — 10nm → 7nm

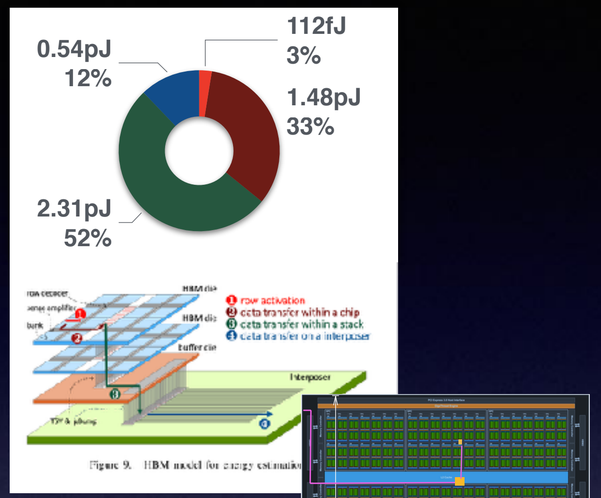
10x more energy efficient

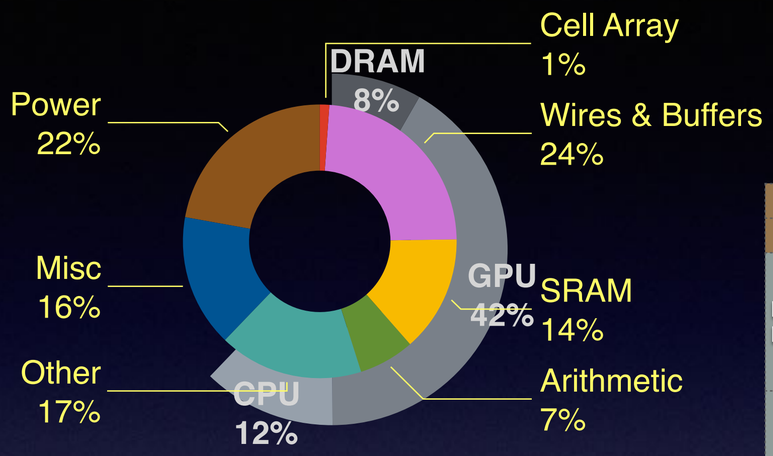
i.e. 150 GFLOPS/W 64-bit Float
600 GFLOPS/W 16-bit Float





IBM AC922





POWER9 TO THE PEOPLE

December 5, 2017 Timothy Prickett Morgan

Architecting an Energy-Efficient DRAM System For GPUs

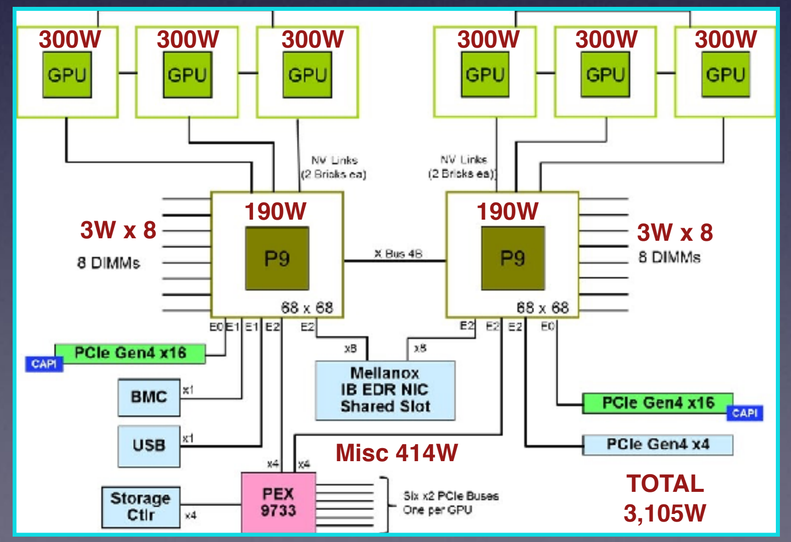
Niladrish Chatterjee*, Mike O'Connor*†, Donghyuk Lee*, Daniel R. Johnson*, Stephen W. Keckler*†, Minsoo Rhu*, William J. Dally*
*NVIDIA †The University of Texas at Austin
{nchatterjee, moconnor, donghyuki, djohnson, skeckler, mrhu, bdally}@nvidia.com

GPUs AND THE FUTURE OF PARALLEL COMPUTING

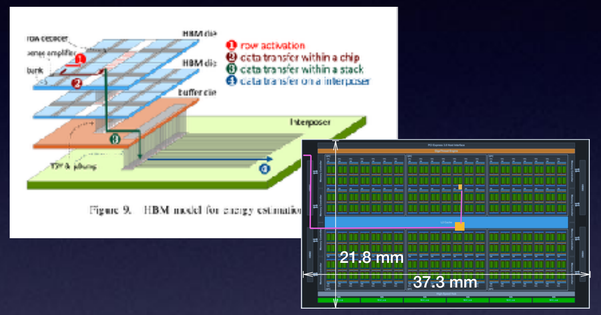
Stephen W. Keckler
William J. Dally
Brucek Khailany
Michael Garland
David Glasco
NVIDIA

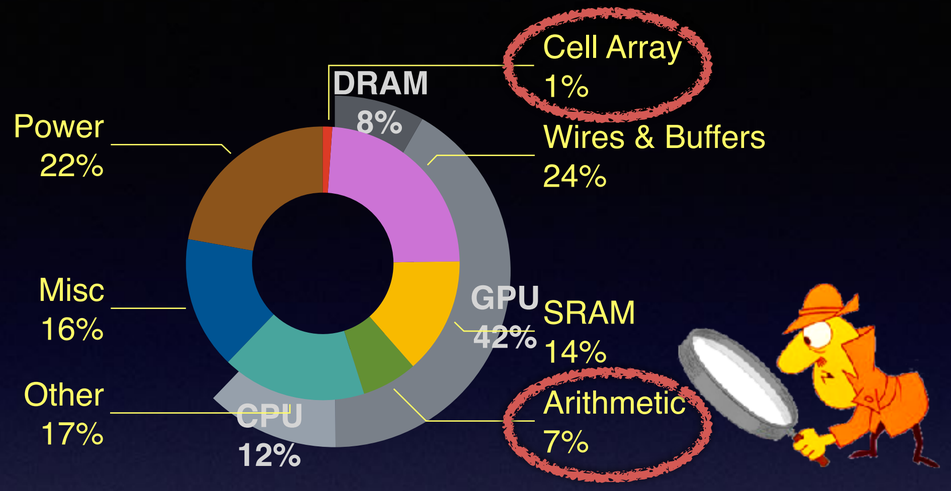
Table 1. Technology and circuit projections for processor chip components.

	2010	2017	
Process technology	40 nm	10 nm, high frequency	10 nm, low voltage
V_{DD} (nominal)	0.9 V	0.75 V	0.65 V
Frequency target	1.6 GHz	2.5 GHz	2 GHz
Double-precision fused-multiply add (DFMA) energy	50 picojoules (pJ)	8.7 pJ	6.5 pJ
64-bit read from an 8 Kbyte static RAM (SRAM)	14 pJ	2.4 pJ	1.8 pJ
Wire energy (per transition)	240 femtojoules (fJ) per bit per mm	150 fJ/b/mm	115 fJ/b/mm
Wire energy (256 bits, 10 mm)	310 pJ	200 pJ	150 pJ



GPU Subsystem and Full Server			Energy (pJ)			Power (W)		
Component		Operation	per Op	per MADD		per Op	Chip	System
HBM2 DRAM	read	row activation (per bit)	0.54	1.0	8.9	3.9	35	209
	0.025	data transfer within a chip	1.48	2.7		10.6		
	dword / MADD	data transfer within a stack	2.3	4.2		16.6		
		data transfer across interposer	0.5	1.0		3.9		
NVIDIA V100 GPU	cache miss	global wire (35mm x 1 bit)	2.7	5.0	5.1	19.6	215	1,293
		write L2 cache (72-bit SRAM)	2.7	0.1		0.3		
	dload / MADD	read L2 cache (72-bit SRAM)	2.7	2.7	17.6	10.6		
		local wire (2.5mm x 1 bits)	0.2	12.5		49.0		
		write vector register file (64-bit)	2.4	2.4		9.4		
	MADD	read 3 vector operands (SRAM)	7.2	7.2	18.3	28.2		
		floating-point MADD (64-bit FP)	8.7	8.7		34.1		
		write 1 vector result (SRAM)	2.4	2.4		9.4		
	other	memory interface, control, etc.	14.0	14.0	14.0	55		
PCIe Card		VRM conversion efficiency	85%	9.6	9.6	38		225
	2	fan (12V, 0.5A) or water pump	6	3.1	3.1	12		72
	6	**GPU Card Subtotal**	300	**76.6**	76.6	**300**		**1,800**
DDR4 DRAM	16	row activation (per bit)	0.54	2.5	10	0.7	3.0	48
		data transfer within chip, I/O	1.71	7.9		2.3		
IBM Power 9 CPU	2	SRAM	just guesses	3.4	16	40	190	380
		wires & buffers		4.3		50		
		other		8.5		100		
Chasis	2	CPU power VRM	85%	2.5	29	58		58
	⬆	misc.—PCIe, fans/pumps, etc.		18		414		414
	#of	primary power supply	85%	17.2		405		405
	15.1	**GFLOP/W**	550	**140**	**Grand Total**			**3,105**



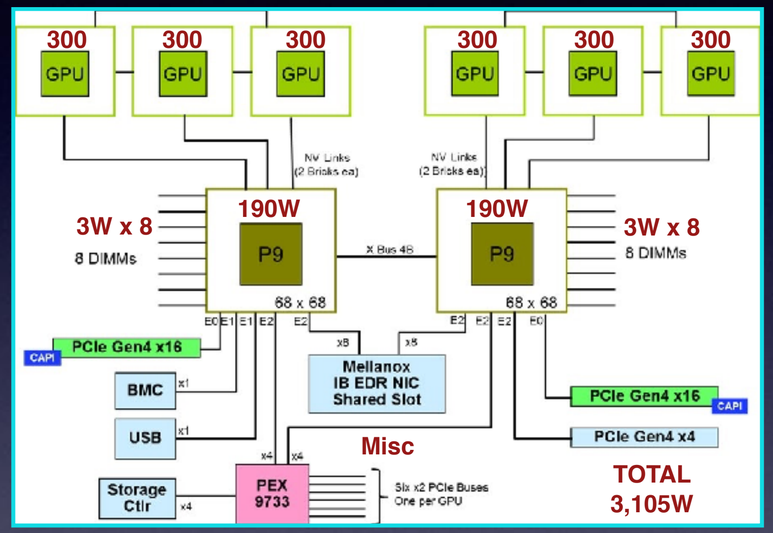


8% DRAM access & arithmetic datapath

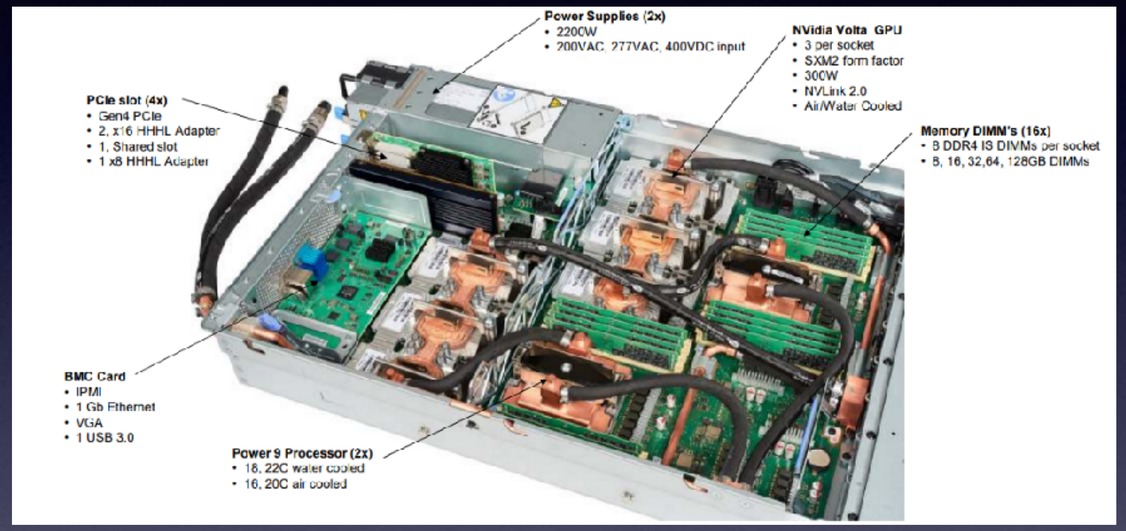
38% moving data & staging in SRAM

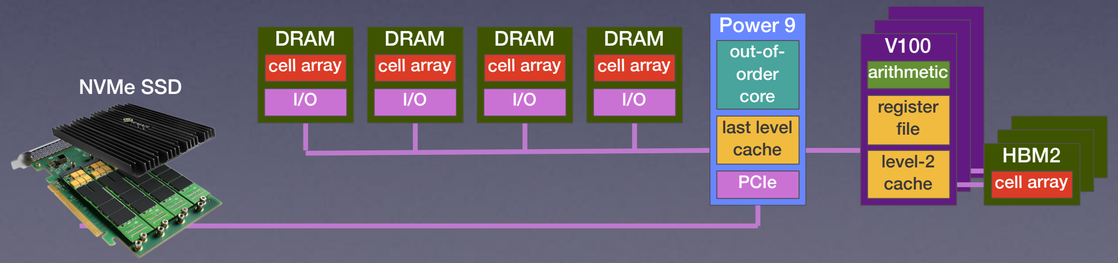
- **Where does most of the energy go?**
- **How efficient is it really?**
- **Can we do better?**

Data-Streaming Architecture



CPU+GPU Architecture

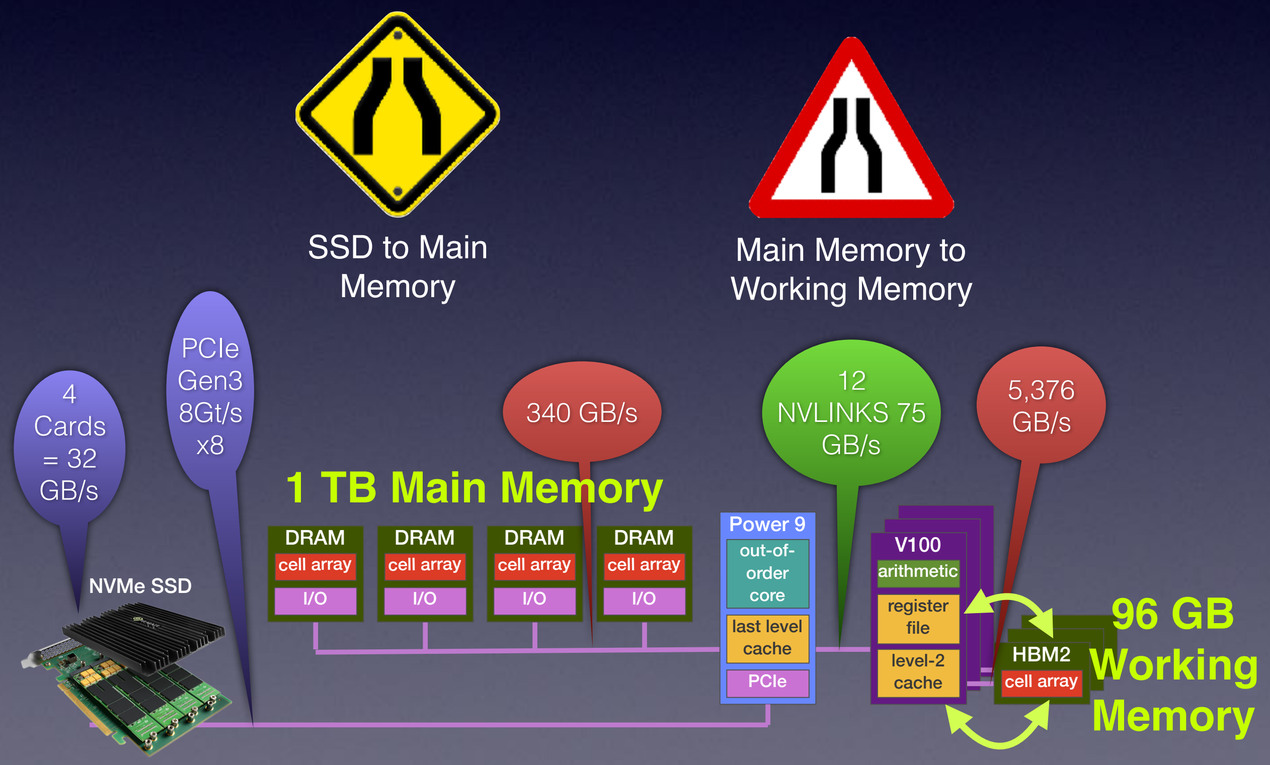


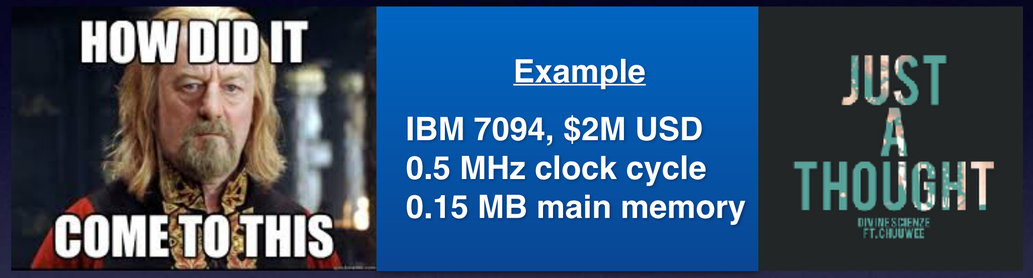


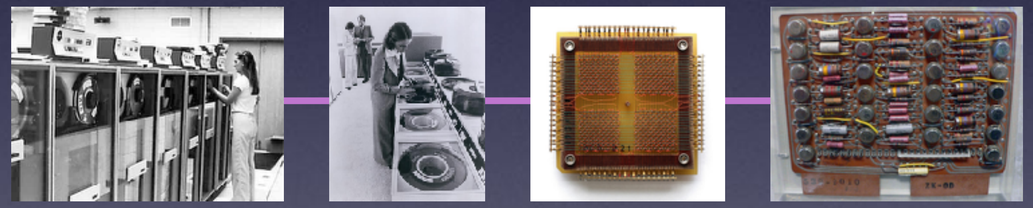
- Easy GPU Programming—full coherence and access to systems memory

NOTE: coherence ≠ always communicating through shared memory
i.e. asynchronous DMA memcpy

Data-Streaming Architecture





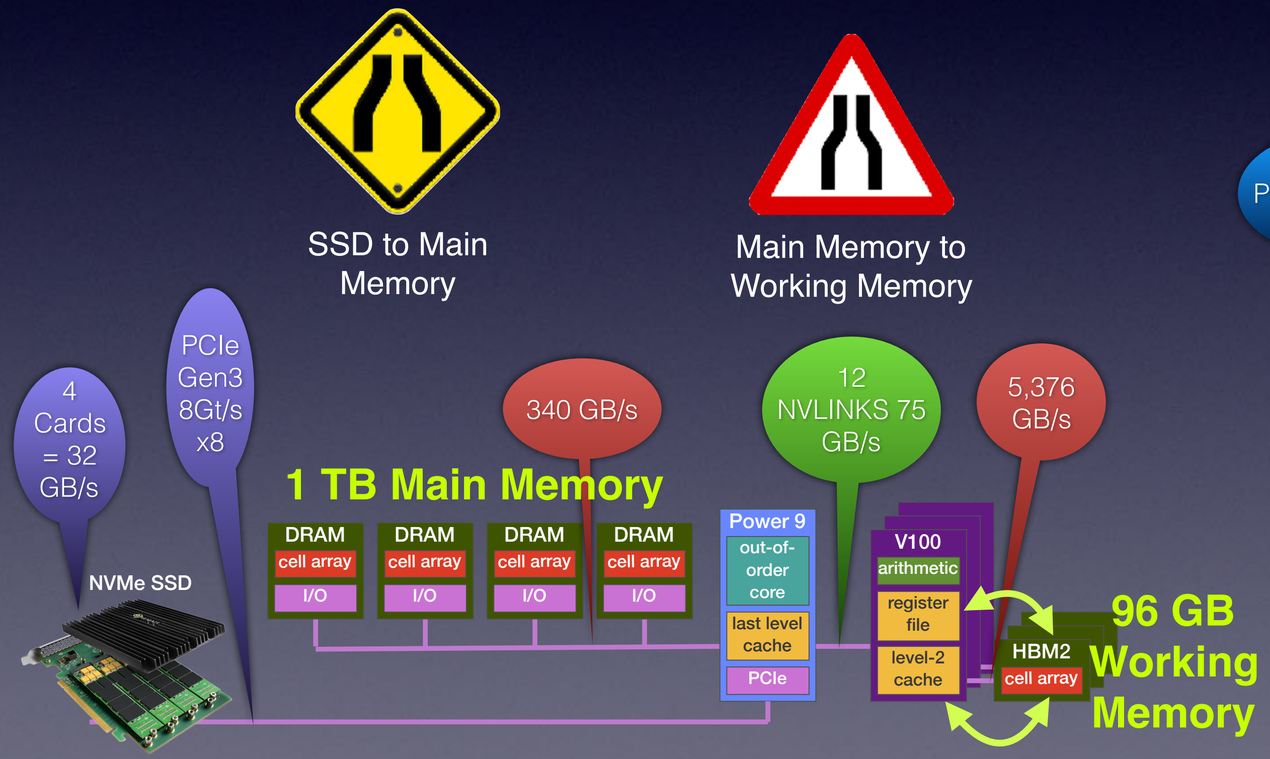


Tape Storage | Disk Drive | Core Memory | Logic Gates

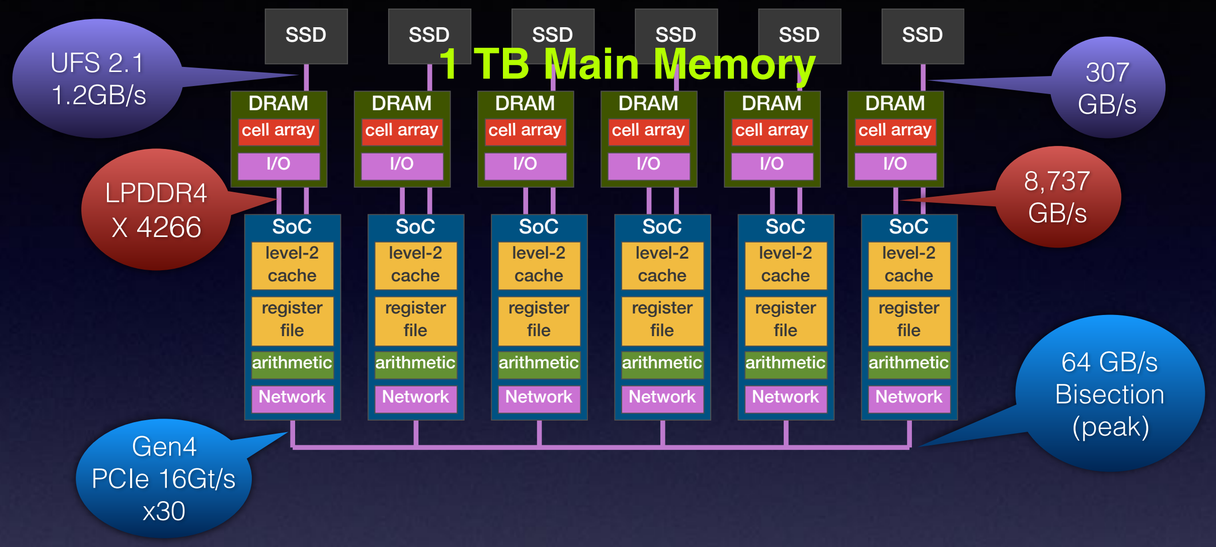
1962

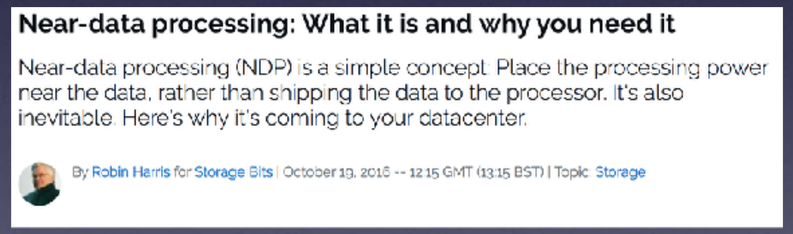
Architectures

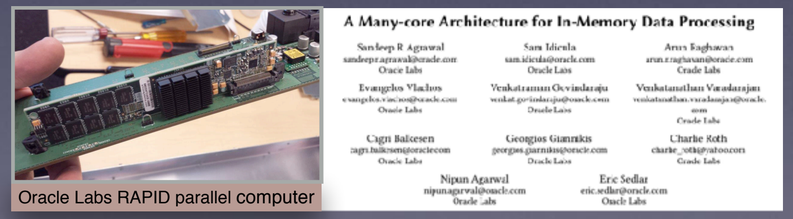
data-streaming



near-data processing



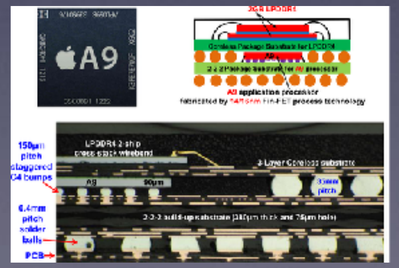


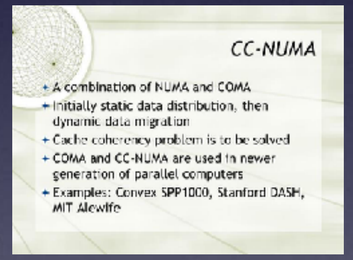


Oracle Labs RAPID parallel computer

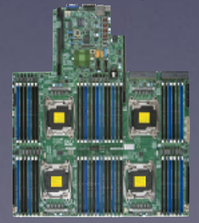
Cache Coherent Parallel Phone Processor

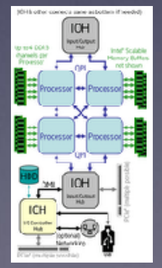
4 GB DRAM, 30GB/s BW
0.567 TFLOPS 16-bit Float

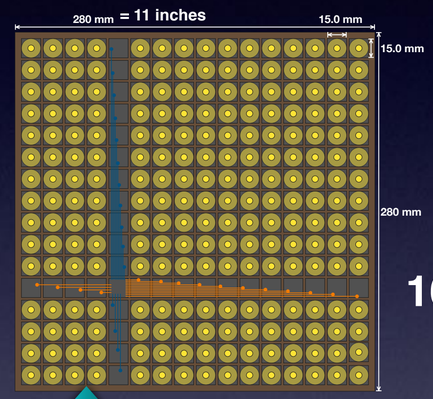




Intel PC

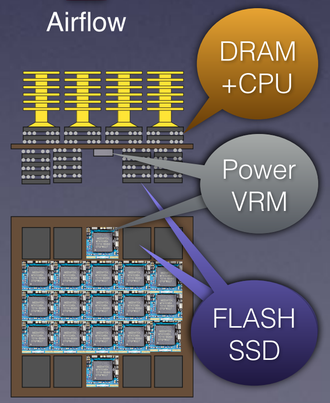






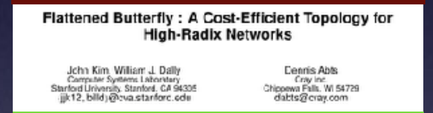
16×16

Airflow



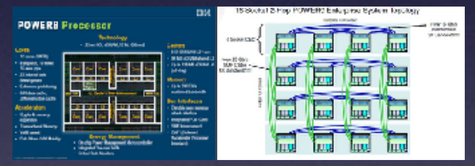
Gen 4 PCIe

- 16 Gt/s @ 10-12 inches
- Power efficient (≈5pJ/bit)
- Strong industry support
- Volume manufacturing



Power & Cooling

- Like 800W processor
- But already spread out



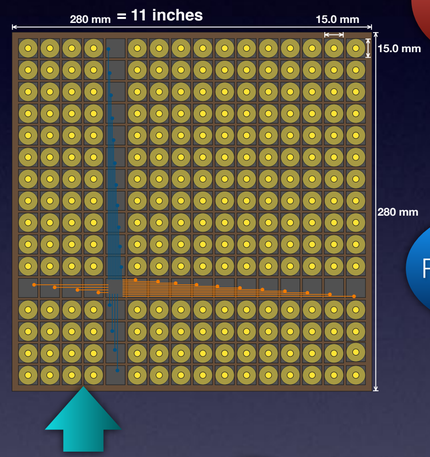
Industry has much experience with CC-NUMA, 256 nodes just bigger number than usual

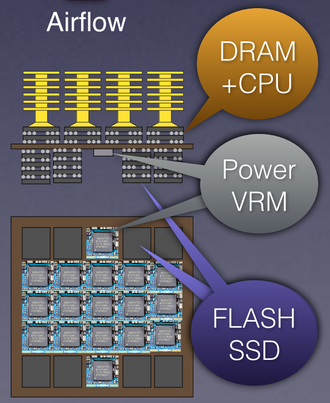
Metric	AC922	This	Unit
DRAM Capacity	1.1	1.0	TB
DRAM Bandwidth	5.7	7.6	TB/s
16-Bit Compute		145	TFLOPS
64-Bit Compute	47.0	36.3	TFLOPS

Advantages

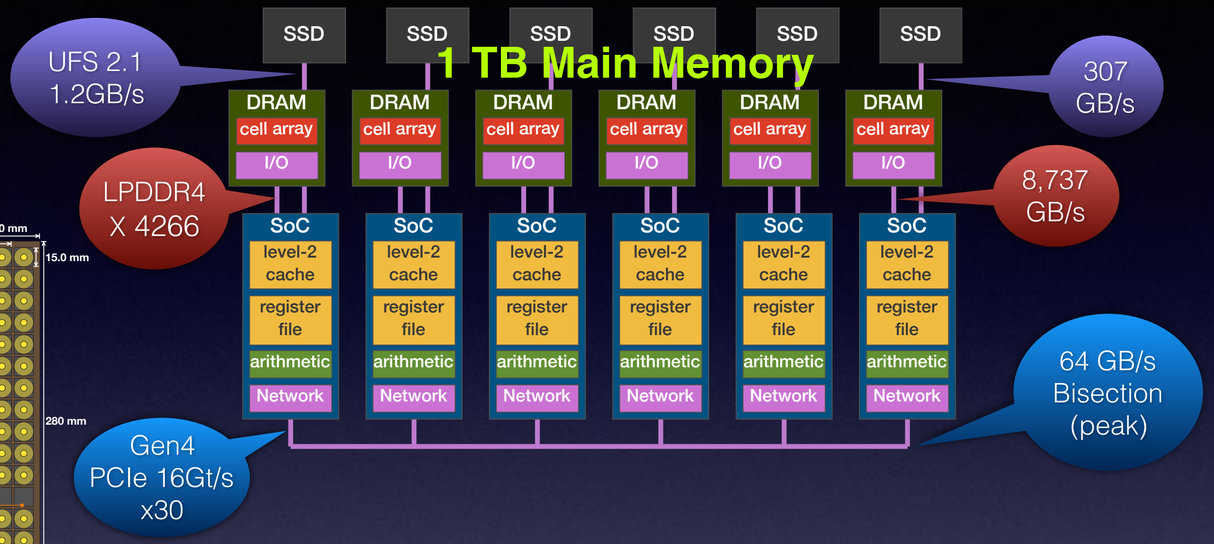
SSD to Main Memory

Main Memory to Working Memory





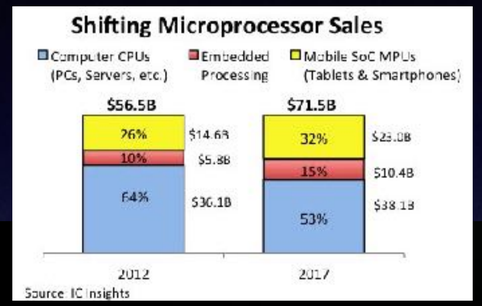
near-data processing



- Easy ~~GPU~~ Programming—full coherence and access to systems memory

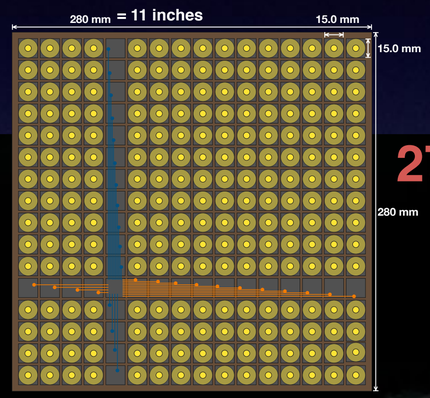
Use cache coherency to make physically partitioned memory easy to program

Best Consumer Technology

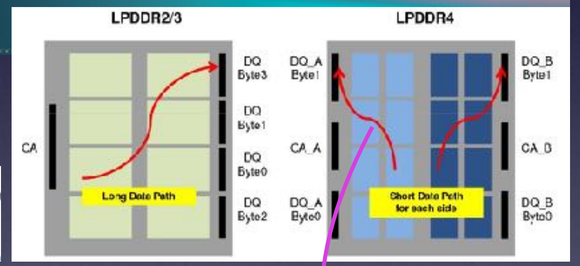


Snapdragon 835 on Samsung 10LPE.

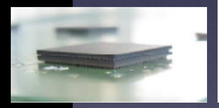
72.3mm²

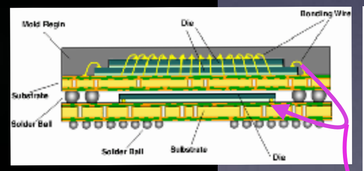


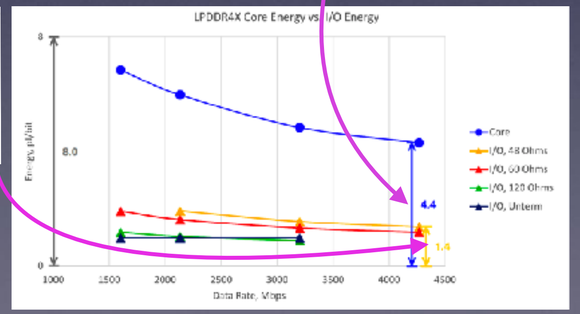
27 March 2017



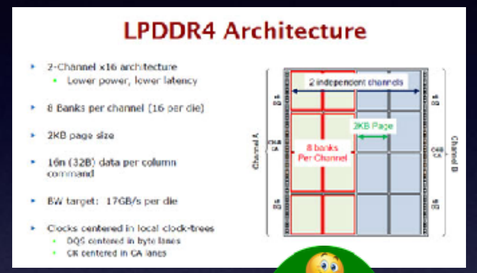
7 December 2017



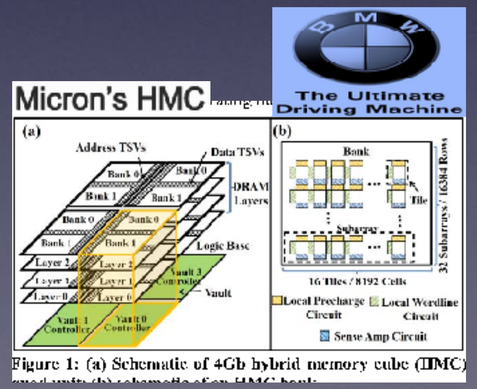


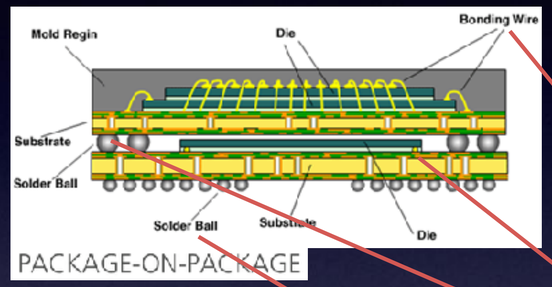


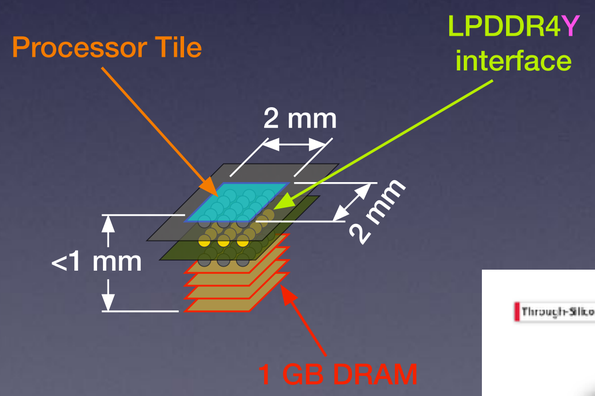
Processor over Memory

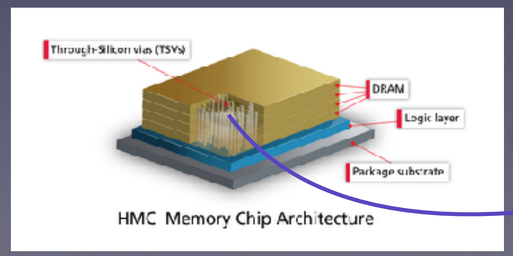


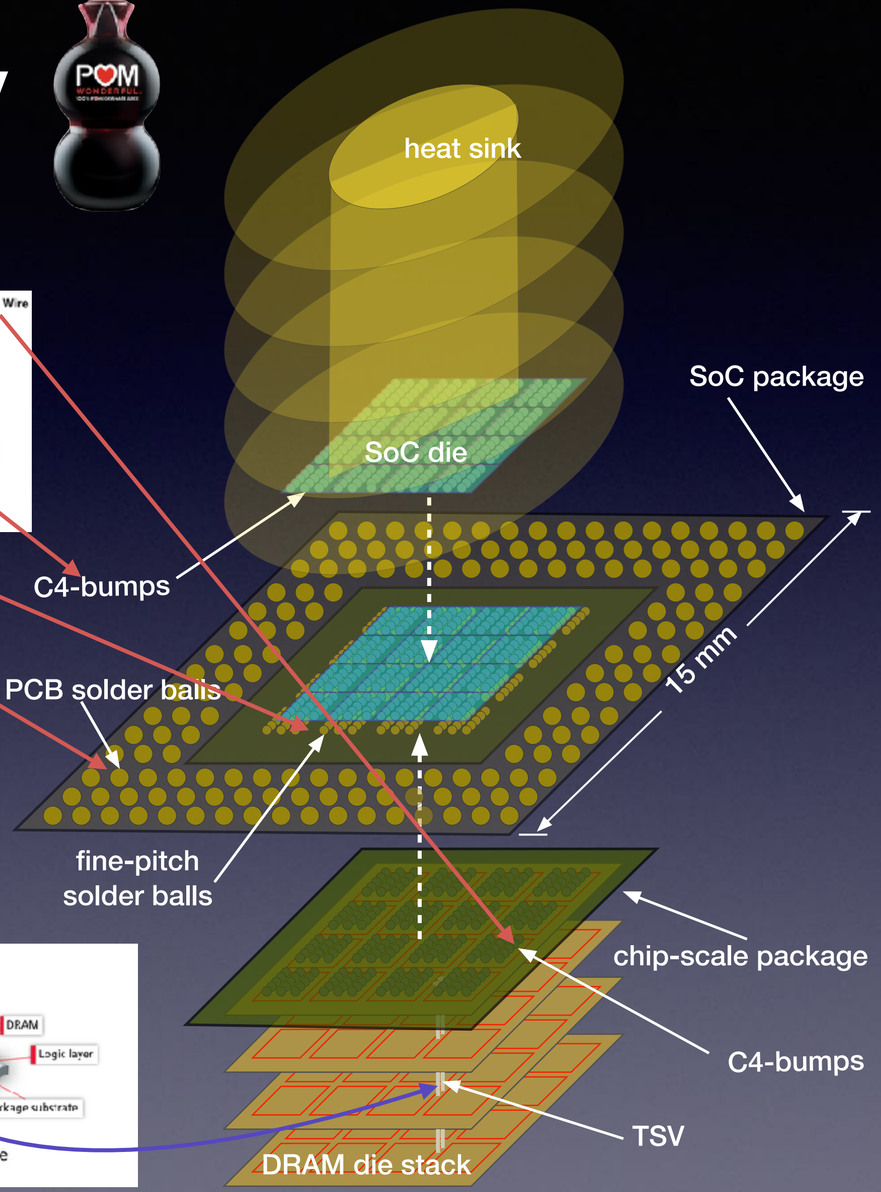
Best of both worlds



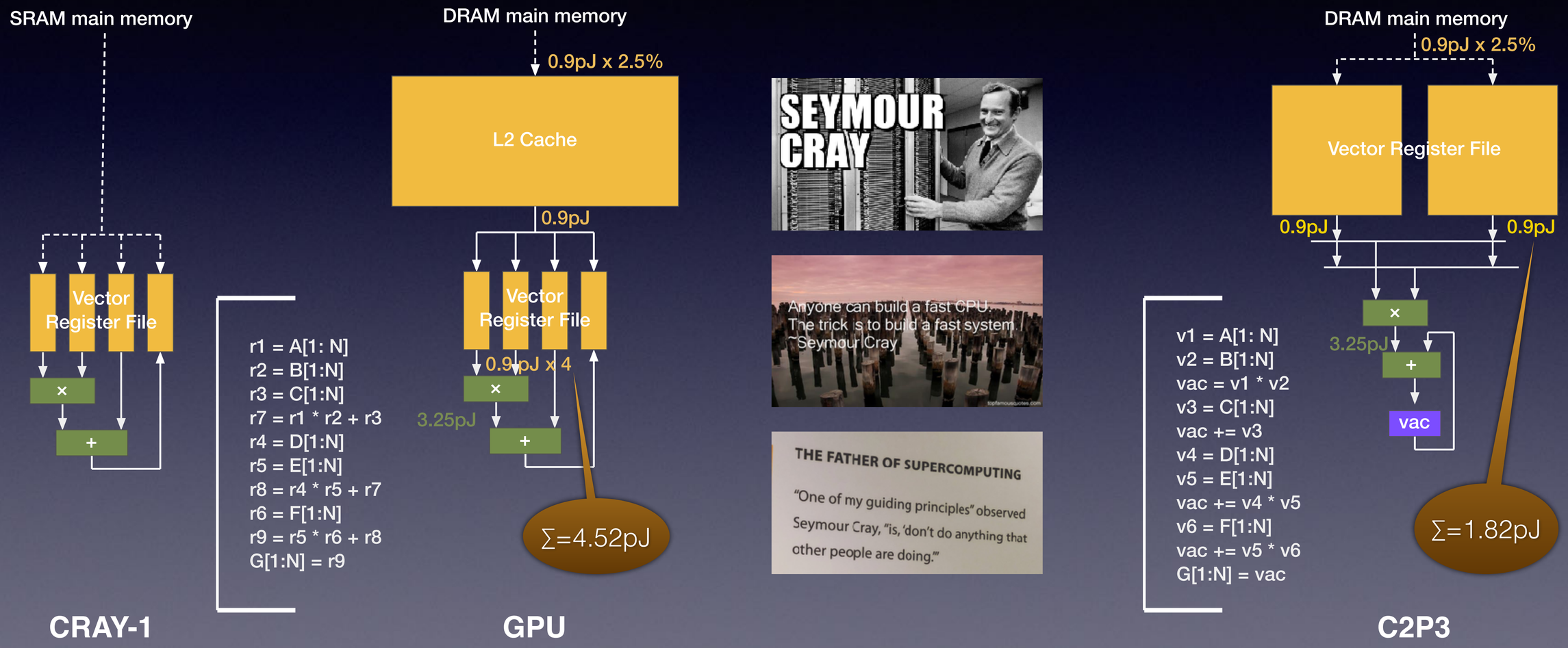




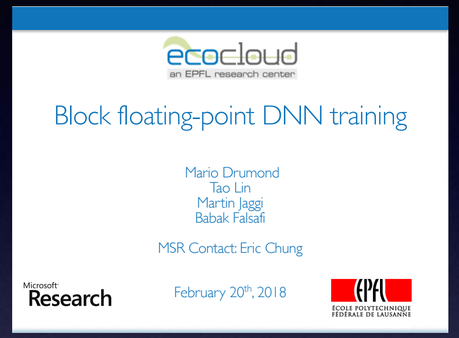


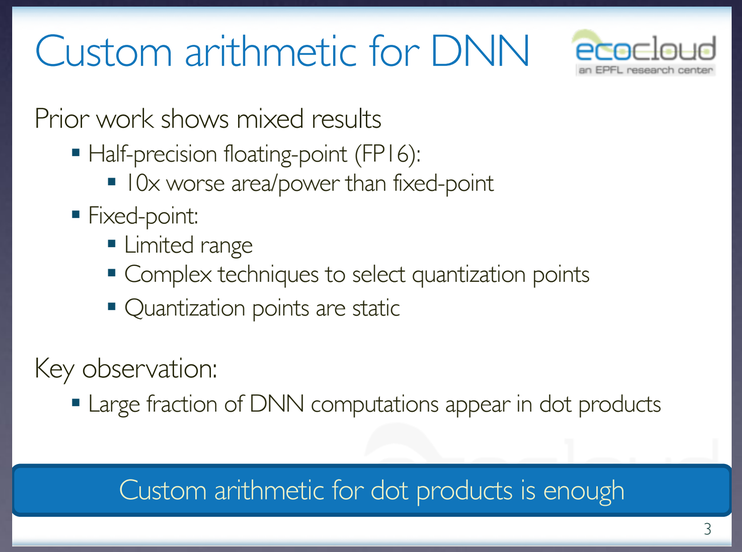


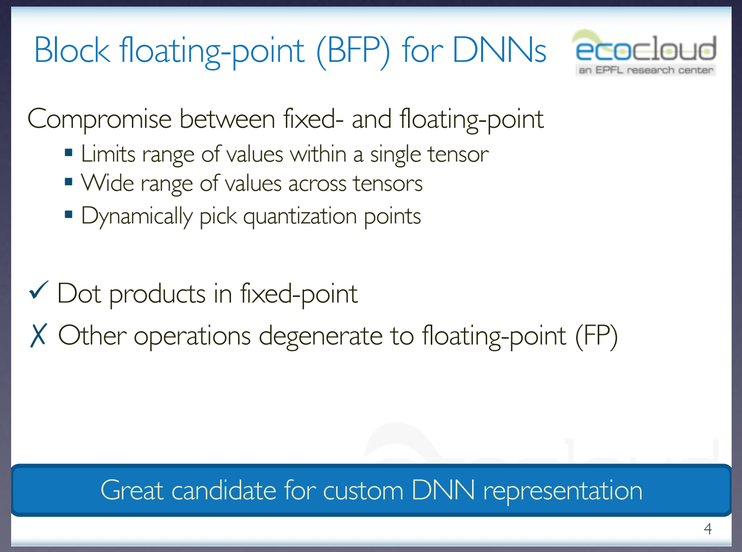
Vector Accumulator Architecture

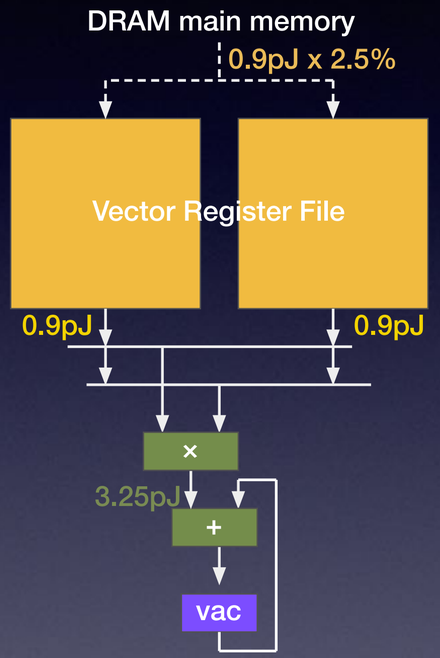


Vector Accumulator Architecture





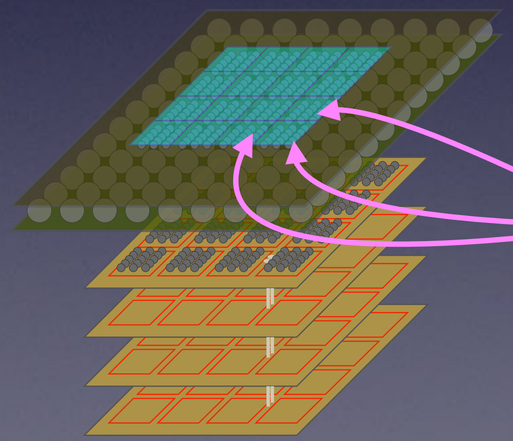


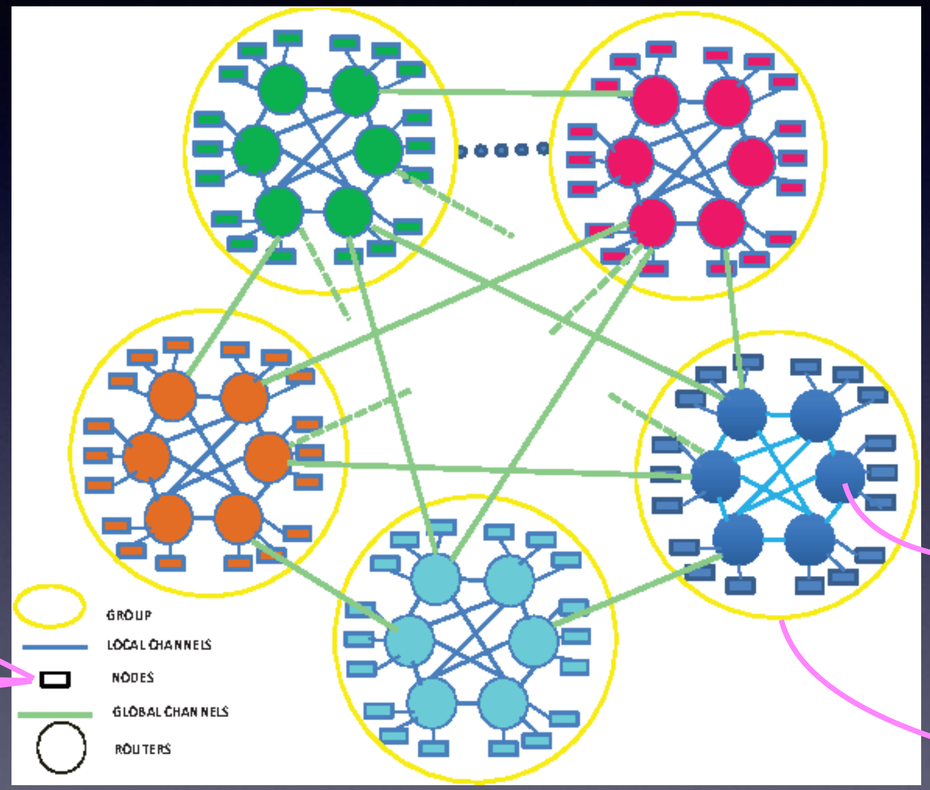


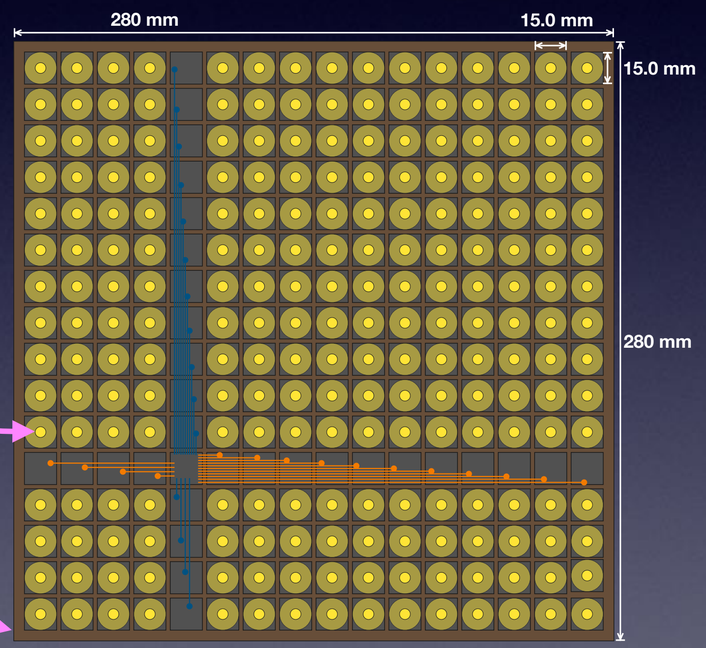
C2P3

Dragonfly Network

Key idea: leverage on-chip and off-chip coherence networks



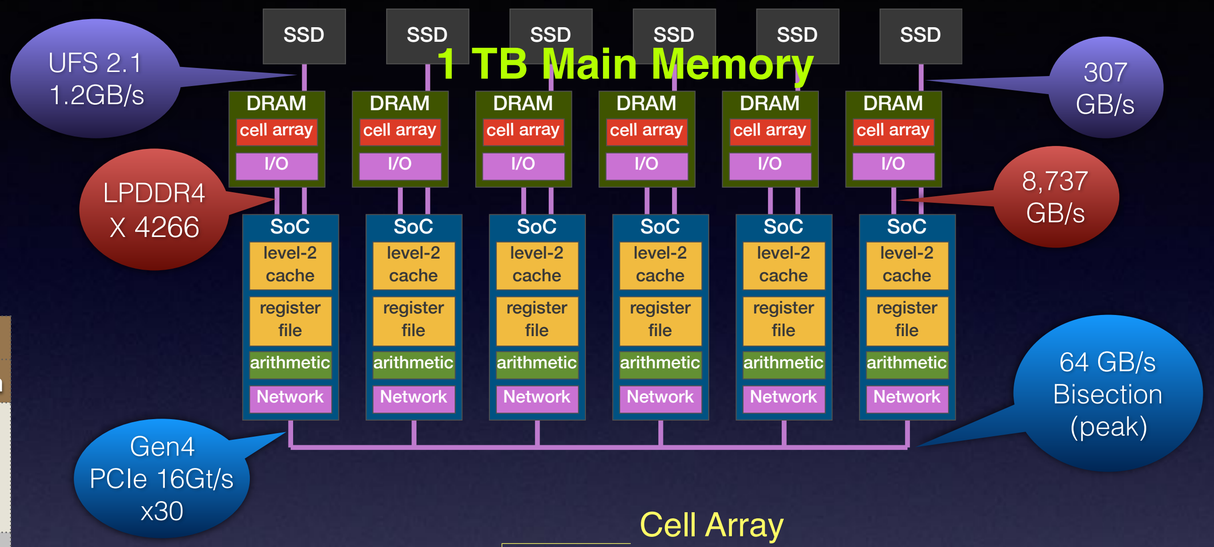


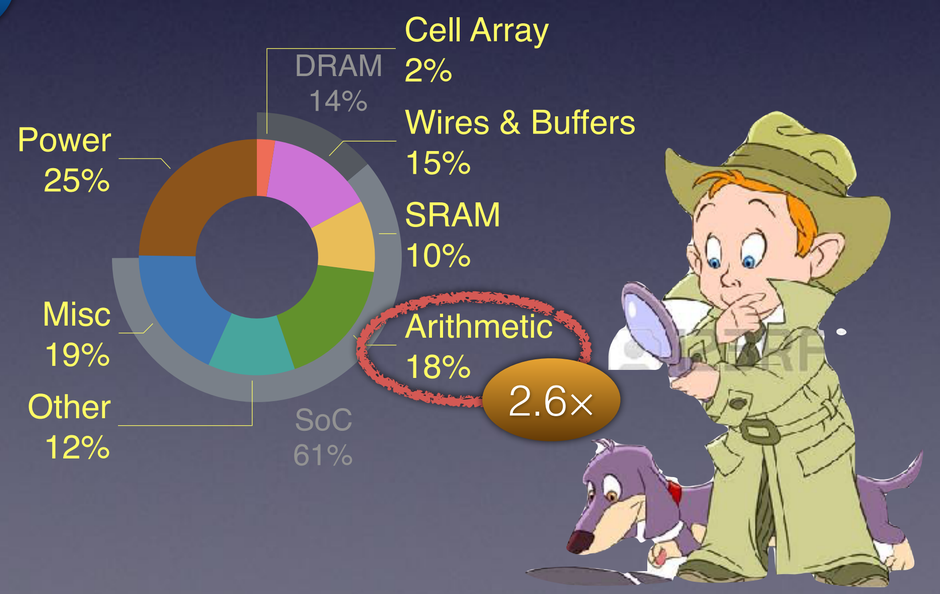


256×256=65,536 processor SoC chips

Together

C2P3 SoC and System			Energy			Power (W)		
Component		Operation	pj	pJ/MADD		per Chip		System
LPDDR4X DRAM	read	row activation (per bit)	0.54	0.91	5.3	0.07	0.4	106
	0.026	data transfer within a chip	1.93	3.3		0.25		
	DW/MADD	off chip I/O SERDES	0.7	1.2		0.09		
C2P3 Processor	load DW	local wire (0.5 mm x 1 bits)	0.03	0.9	0.9	0.07	1.8	468
		write register file (64-bit	1.8	0.05	0.05	0.00		
	MADD instruction	read 2 operands (64-bit	3.6	3.6	10.1	0.28		
		floating-point MADD (64-bit)	6.5	6.5		0.51		
	other	memory interface, control, etc.	1.17	1.17	4.4	0.09		
		coherency directory, switch	3.19	3.19		0.25		
	PCIe	30 links @ 10% (12GB/s)	5	6.2	6.2	0.48		
	Ethernet	1 link (10 Gb/s)	15	1.9	1.9	0.15		
Power		VRM conversion efficiency	85%	4.0	4.0	0.31	0.3	80
Chasis	256	misc—fans, etc.				0.05		12
	SoC	primary power supply	0.85			0.39		100
TOTAL	**52.0**	**GFLOPS / Watt**	3.4	**33**	32.9	**3.00**		**767**



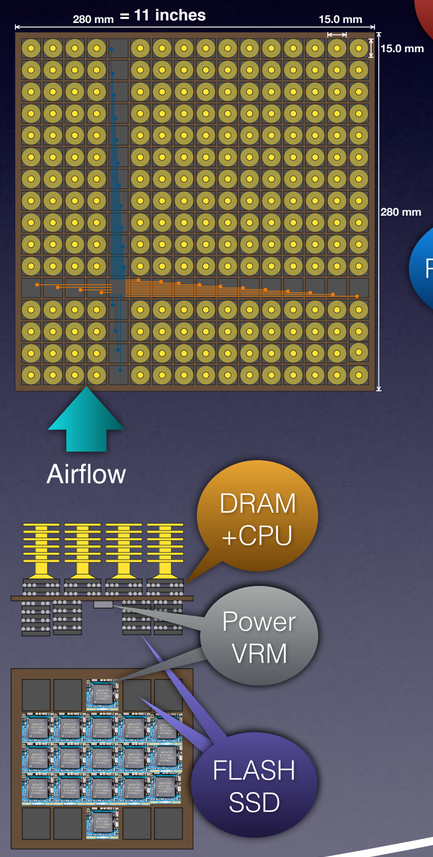


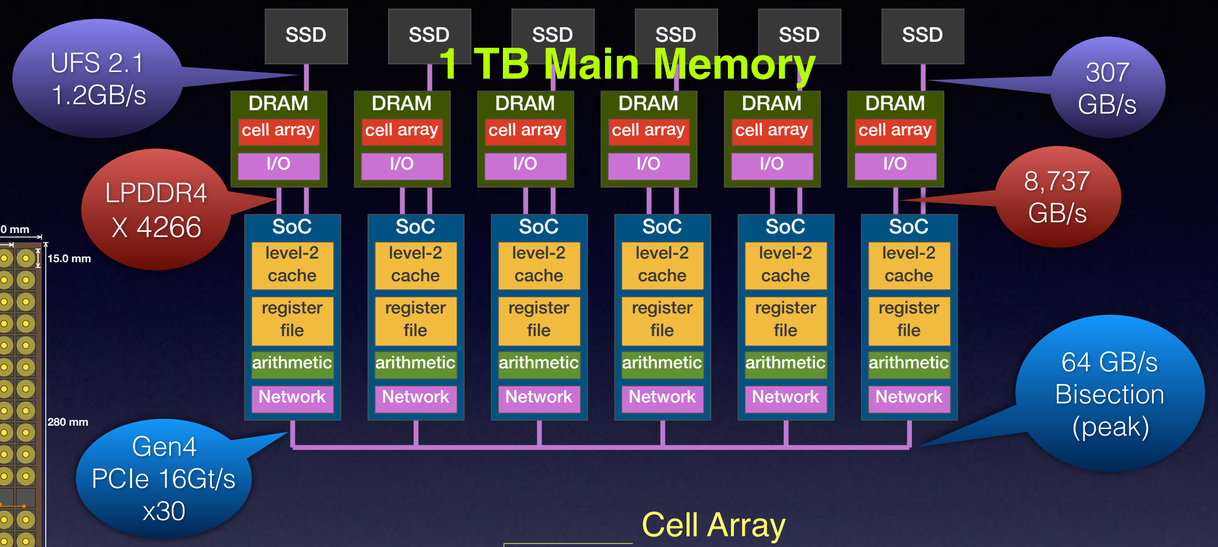
Together

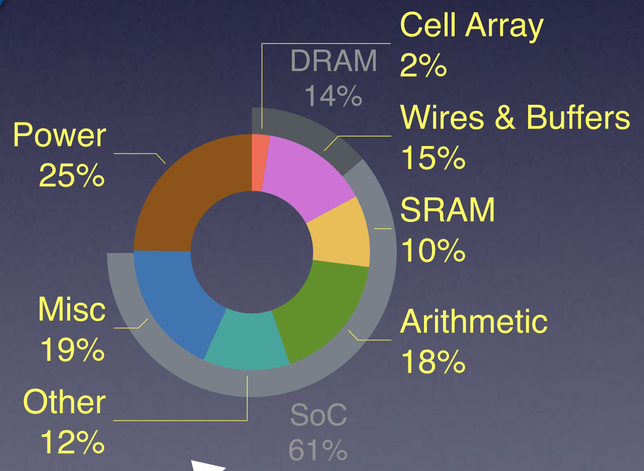
- 2× — near-data processing
- 1.6× — SoC/DRAM 3D layout/package co-design
- 1.6× — vector accumulator
- 1.4× — best consumer electronics process
- 1.4× — 10nm → 7nm

10x more energy efficient

i.e. 150 GFLOPS/W 64-bit Float
600 GFLOPS/W 16-bit Float





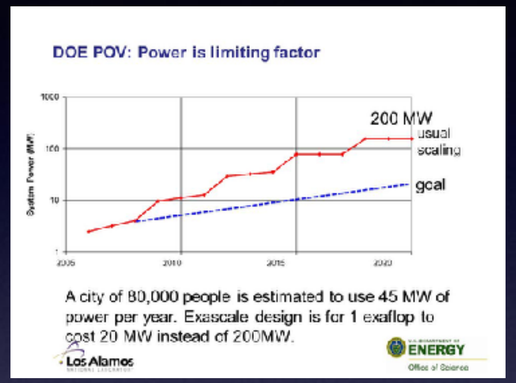


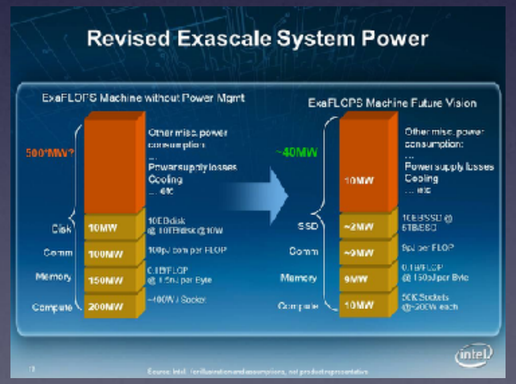
Conclusion

- 2× — near-data processing
- 1.6× — SoC/DRAM 3D layout/package co-design
- 1.6× — vector accumulator
- 1.4× — best consumer electronics process
- 1.4× — 10nm → 7nm

10x more energy efficient

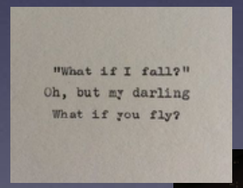
i.e. 150 GFLOPS/W 64-bit Float
600 GFLOPS/W 16-bit Float





$$\frac{(1\ \text{ExaFLOPS} / 80\%)}{150\ \text{GFLOPS/W}} = 8.3\ \text{MW}$$

Cache **C**oherent **P**arallel **P**hone **P**rocessor



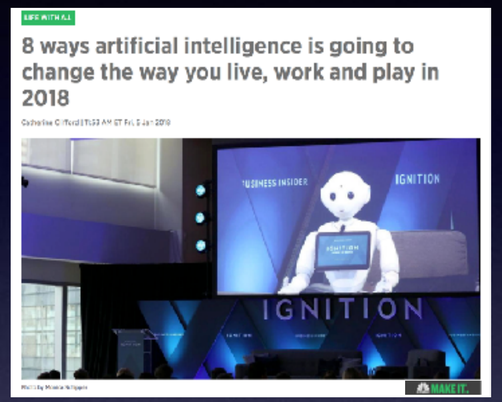
C2P3

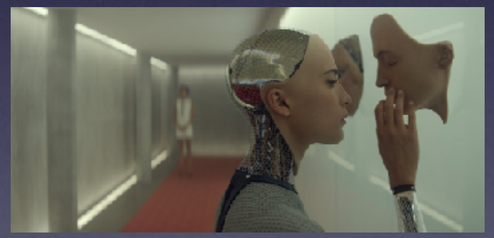
Apple's 'A12' chip reportedly in production using 7nm process from TSMC

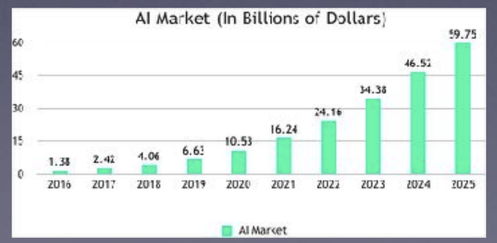
by Paul Lilly — Sunday, June 10, 2018

Samsung's 7nm Exynos 9820 Mongoose M4 Could Crush Mighty ARM Cortex-A76

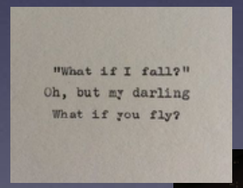
Thank You

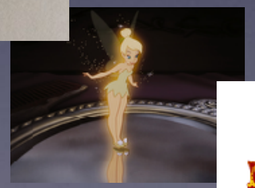






To Build The Fastest Computer In The World





C2P3

Abstract: It is said artificial intelligence is going to change the way we live, work and play in 2018. Certainly the market for AI technology is growing rapidly. Some of us believe excessive energy consumption is holding back even more revolutionary advances in AI software. This talk begins by looking at how energy is consumed in the IBM AC922 server, marketed for enterprise AI computing and used in the world's fastest supercomputer, US DOE Summit. The AC922 is a CPU+GPU data-streaming architecture. This talk proposes a near-data processing architecture using low-power consumer cellphone technology. By combining architecture and 3D SoC/DRAM chip layout/packaging co-design ideas, I suggest it may be possible to improve energy efficiency by an order of magnitude within one process generation. This talk presents on-going work I hope to continue during my visit to EPFL University.

Bio: Peter Hsu was born in Hong Kong and moved to the United States as a teenager. He received a B.S. degree from the University of Minnesota at Minneapolis in 1979, and the M.S. and Ph.D. degrees from the University of Illinois at Urbana-Champaign in 1983 and 1985, respectively, all in Computer Science. His first job was at IBM T. J. Watson Research Center from 1985-1987, working on code generation techniques for superscalar and out-of-order processors with the 801 compiler team. He then joined one of his former professor at Cydrome, which developed an innovative VLIW computer. In 1988 he moved to Sun Microsystems and tried to build a water-cooled gallium arsenide SPARC processor, but the technology was not sufficiently mature and the effort failed. He joined Silicon Graphics in 1990 and designed the MIPS R8000 TFP microprocessor. The R8000 was released in 1994 and shipped in the SGI Power Challenge servers and Power Indigo workstations. Fifty of the TOP500.org list of supercomputer systems used R8000 chips in 1994. Peter became a Director of Engineering at SGI, then left in 1997 to co-found his own startup, ArtX, best known for designing the Nintendo GameCube. ArtX was acquired by ATI Technologies in 2000. He left ArtX in 1999 and worked briefly at Toshiba America, where he developed advanced place-and-route methodologies for high frequency microprocessor cores in SoC designs, then became a visiting Industrial Researcher at the University of Wisconsin at Madison in 2001. Throughout the 2000's he consulted for various startups, attended the Art Academy University and the California College of the Arts in San Francisco where he learned to paint oil portraits, attended a Paul Mitchell school where he learned to cut and color hair. In the late 2000's he consulted for Sun Labs, which lead to discussions about the RAPID research project, a power-efficient massively parallel computer for accelerating big data analytics in the Oracle database. He was with Oracle Labs as an Architect from 2011 to 2016. In 2017 Dr. Hsu founded CAVA Computers, Inc. in an attempt to bring to market high-performance hyper-converged storage with computing. Peter will be a visiting researcher at EPFL University in Lausanne, Switzerland fall of 2018.