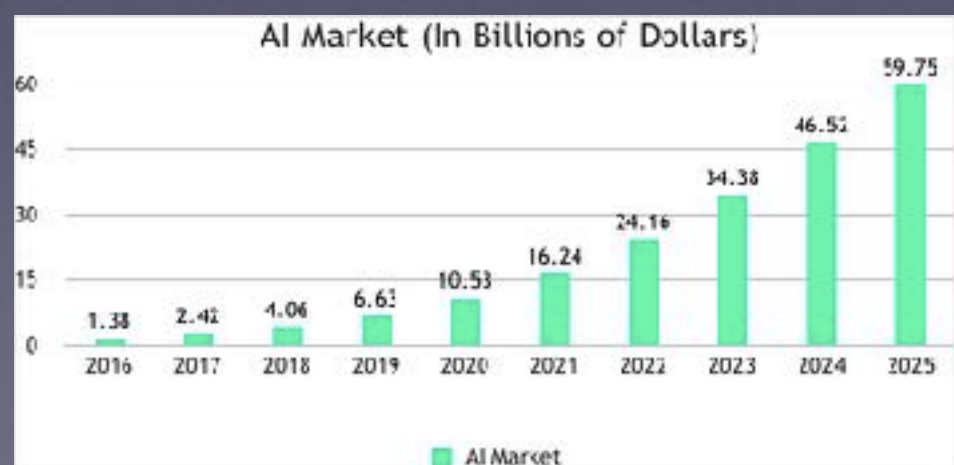
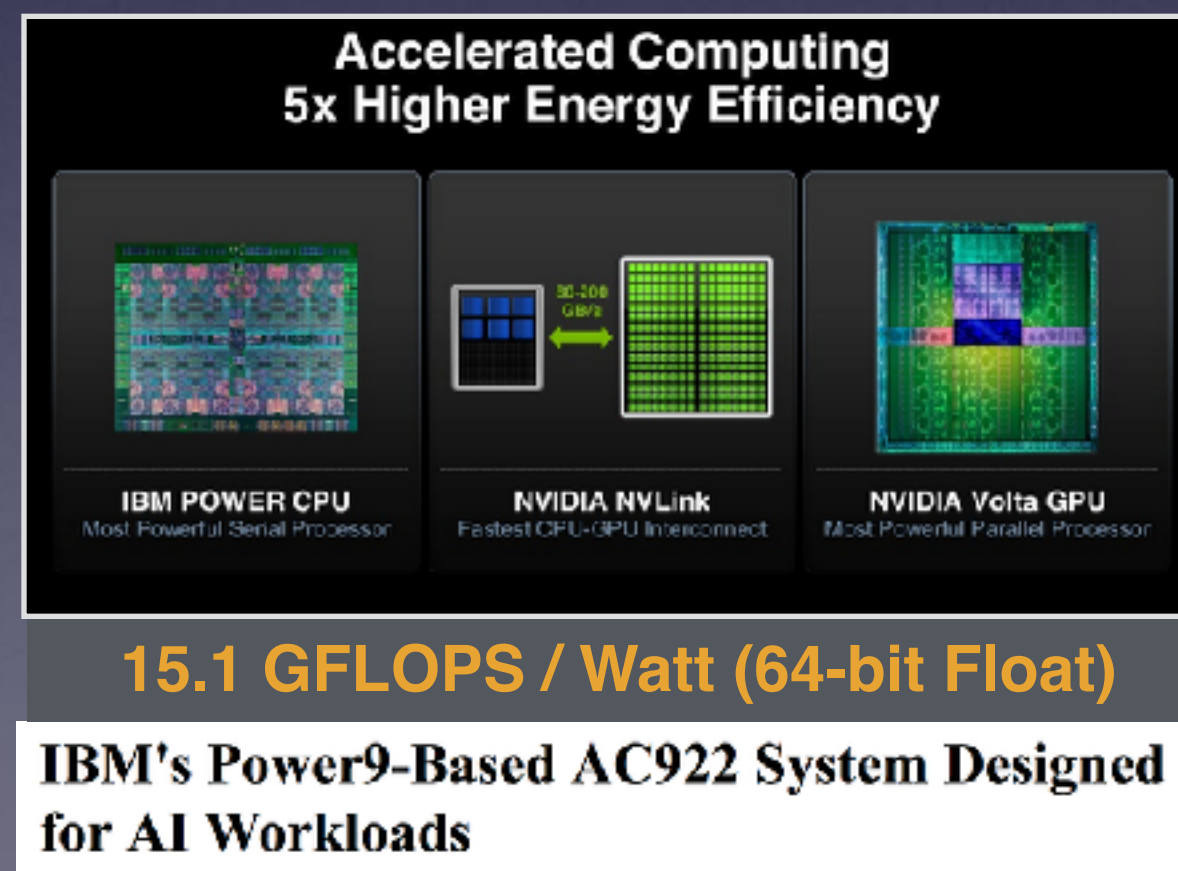
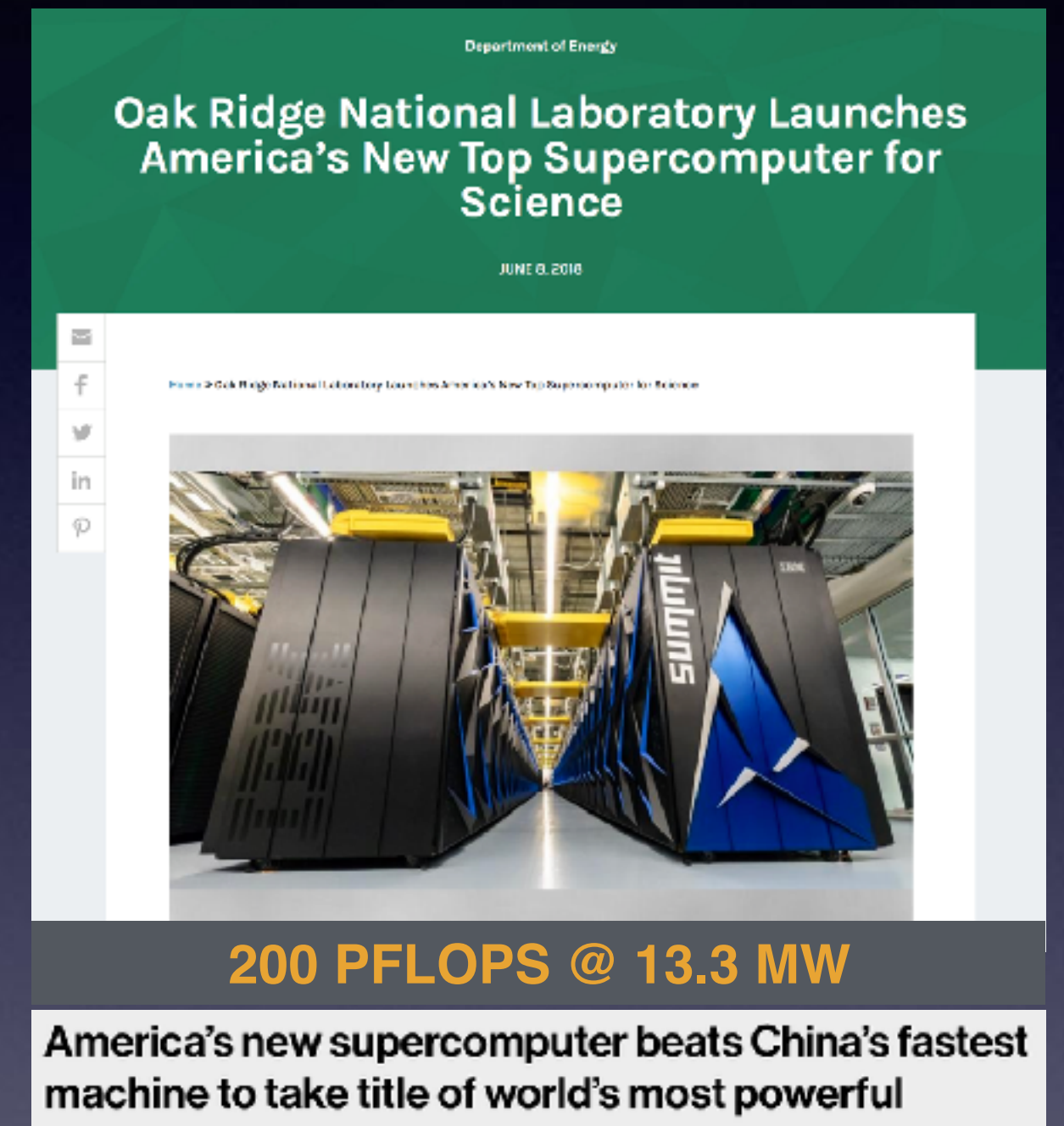
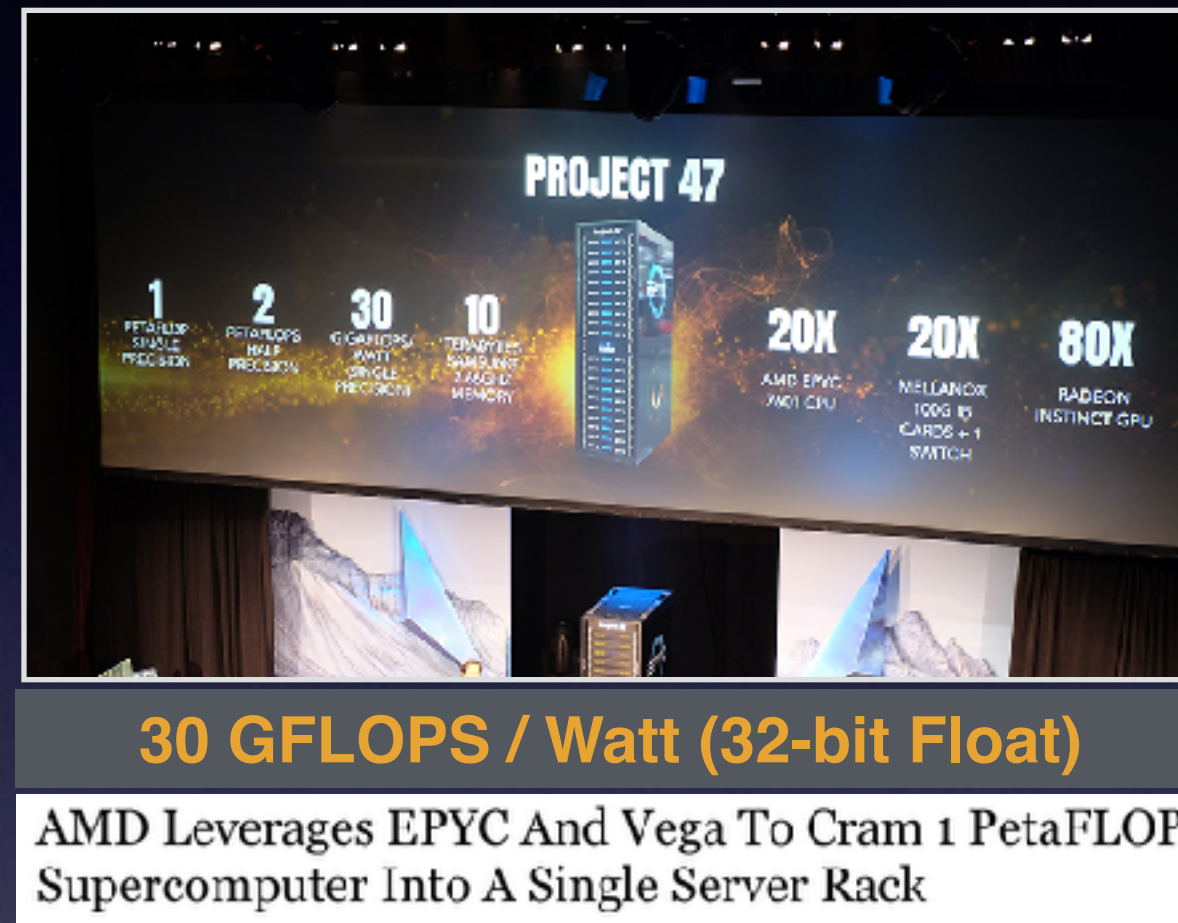
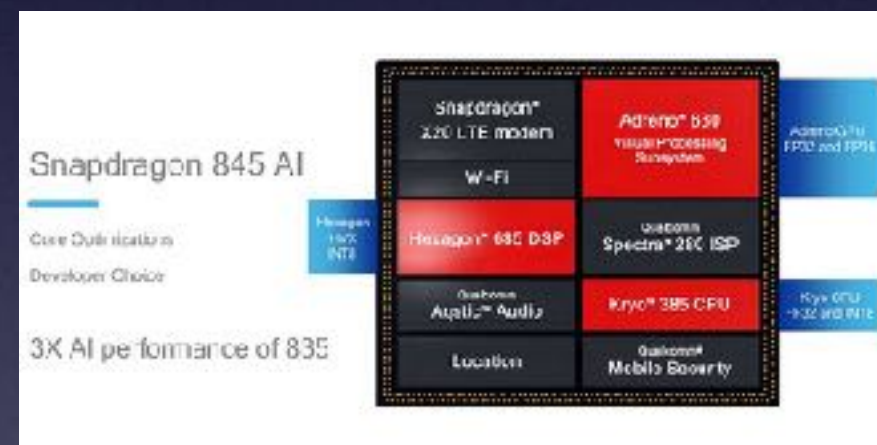
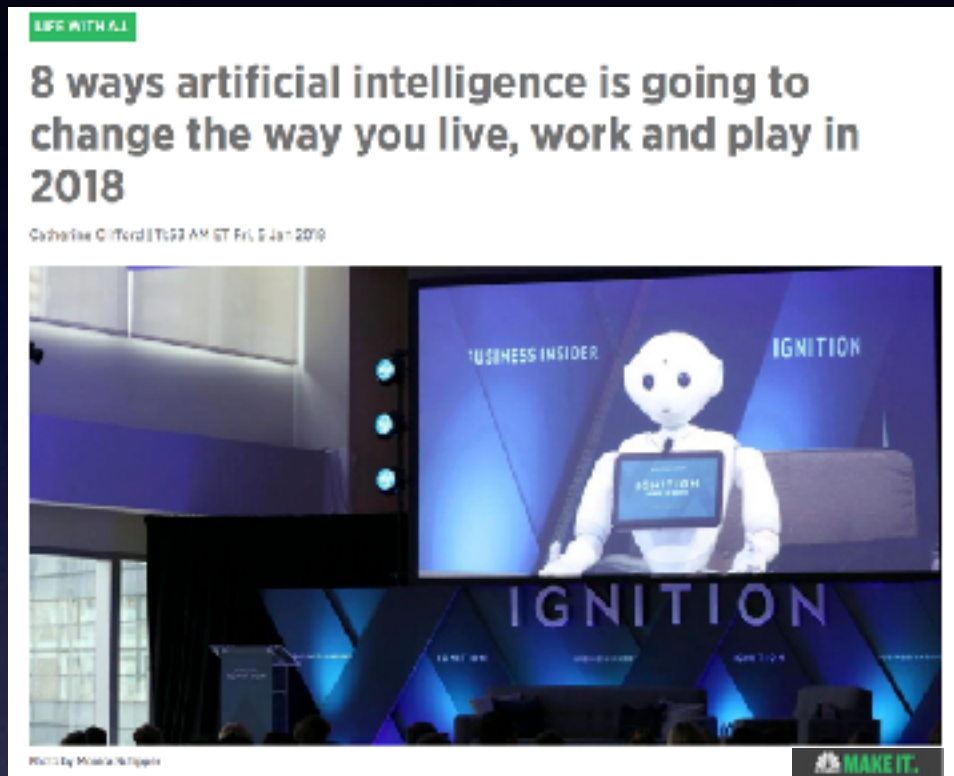


# Using Cellphone Technology to Build **AI** Servers

Peter Hsu, PhD  
pete2222@mac.com

Presented at Barcelona Supercomputer Center, Spain  
25 June 2018

# Energy is Real Limit of AI





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

Your search for the right IT infrastructure for AI is over



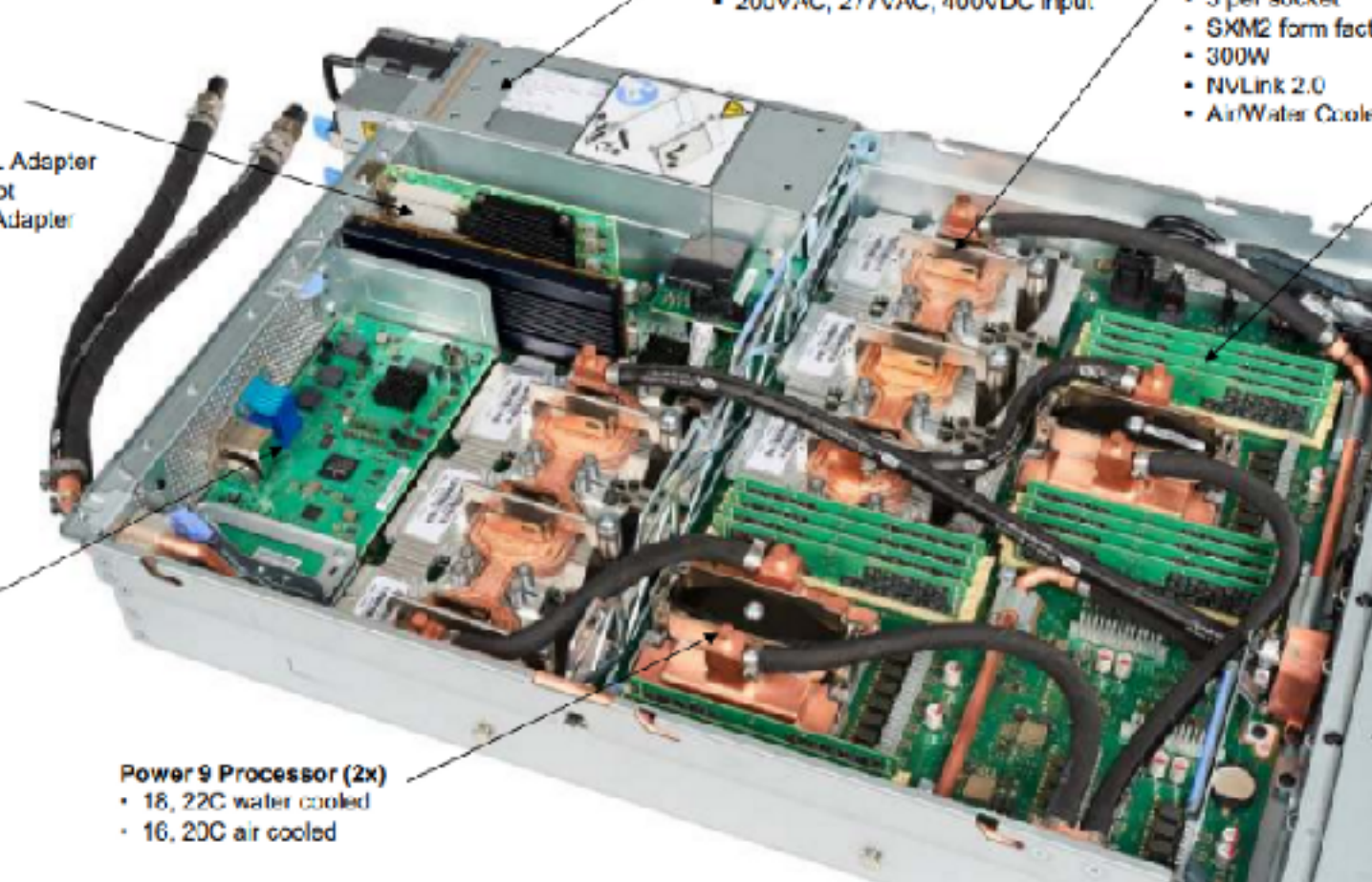
IBM Power Systems

## Deep-Learning for the Enterprise

Sumit Gupta  
VP, HPC, AI, & Machine Learning  
IBM Cognitive Systems

December, 2017

IBM PowerAI



- Power Supplies (2x)**
  - 2200W
  - 200VAC, 277VAC, 400VDC input
- Nvidia Volta GPU**
  - 3 per socket
  - SXM2 form factor
  - 300W
  - NVLink 2.0
  - Air/Water Cooled
- Memory DIMM's (16x)**
  - 8 DDR4 IS DIMMs per socket
  - 8, 16, 32, 64, 128GB DIMMs
- PCIe slot (4x)**
  - Gen4 PCIe
  - 2, x16 HHHL Adapter
  - 1, Shared slot
  - 1 x8 HHHL Adapter
- BMC Card**
  - IPMI
  - 1 GB Fibernet
  - VGA
  - 1 USB 3.0
- Power9 Processor (2x)**
  - 18, 22C water cooled
  - 16, 20C air cooled

"The POWER9 chip is the first chip designed for AI. Of course it goes in a system that is designed to take full advantage of that, and that is the new AC922."

Bob Picciano  
SVP, IBM Cognitive Systems

### VOLTA TO FUEL SUMMIT

New Apex in AI Supercomputing



AI Exascale Today

Performance Leadership

200 PF

20 PF

Accelerated Science

3+EFLOPS  
Tensor Co.

10X  
Per Node, Tile

5-10X  
Application Per Node, Tile

### You Can Also Buy A Smaller Version Of Oak Ridge National Labs Most Powerful AI Supercomputer

Patrick Moorhead, CONTRIBUTOR  
Factor about 10x higher compute, technology and usage models. [FULL BIO](#)

Content expressed by Forbes Contributors is their own




### US to Build Two Flagship Supercomputers



**SUMMIT** **SIERRA**

150-300 PFLOPS Peak Performance

IBM POWER9 CPU + NVIDIA Volta GPU

NVLink High Speed Interconnect

40 TFLOPS per Node, >3,400 Nodes


2017

Major Step Forward on the Path to Exascale

### ANNOUNCING TESLA V100


GIANT LEAP FOR AI & HPC  
VOLTA WITH NEW TENSOR CORE

31B trans | TSMC 12nm FFN | 815mm<sup>2</sup>  
5,120 CUDA cores  
7.5 FP64 TFLOPS | 15 FP32 TFLOPS  
NEW 120 Tensor TFLOPS  
70MB SM RF | 16MB Cache | 16GB HBM2 | 320 GB/s  
300 GB/s NVLink



### POWERING THE AI REVOLUTION

JENSEN HUANG, FOUNDER & CEO | GTC 2017



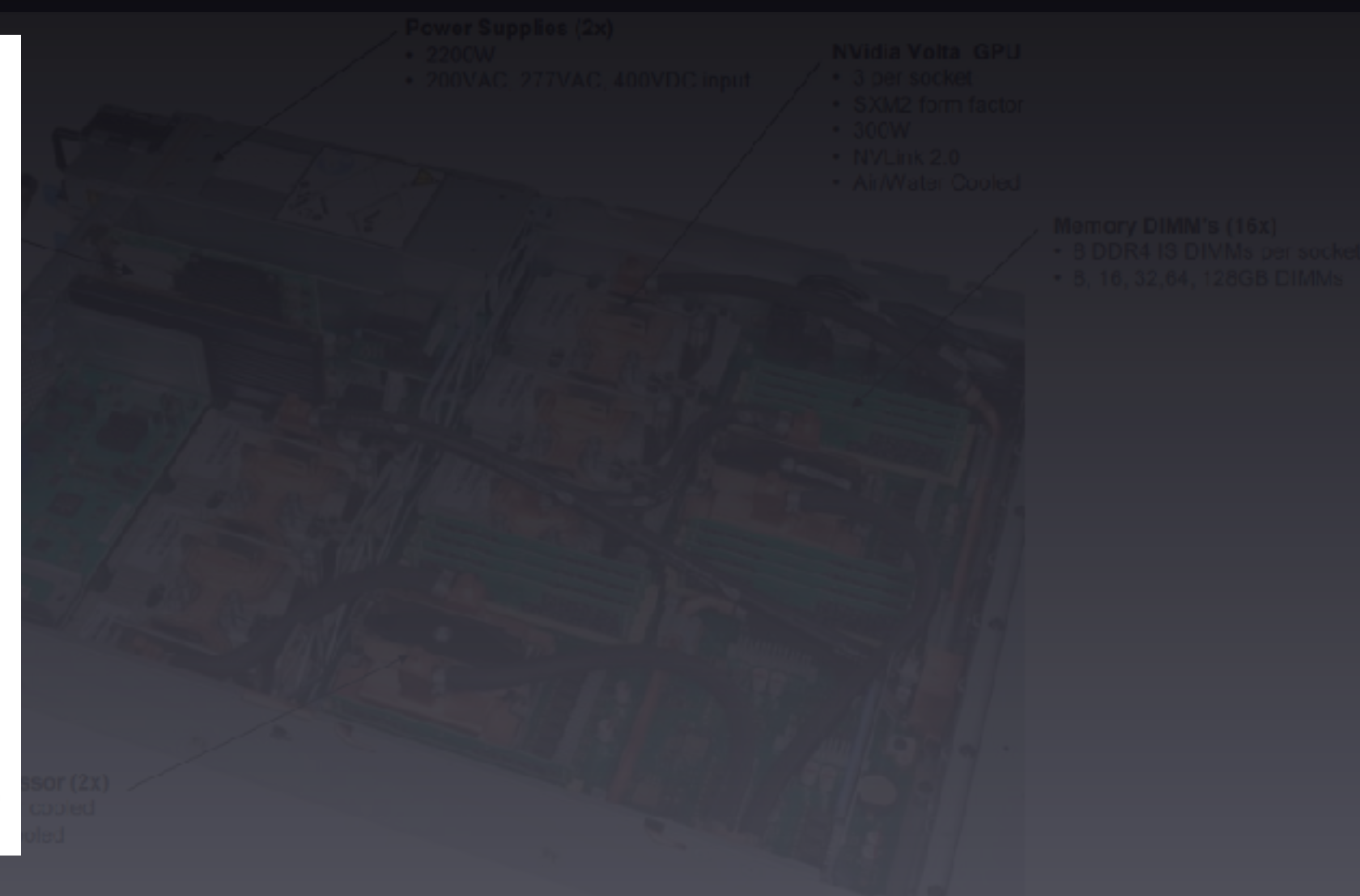
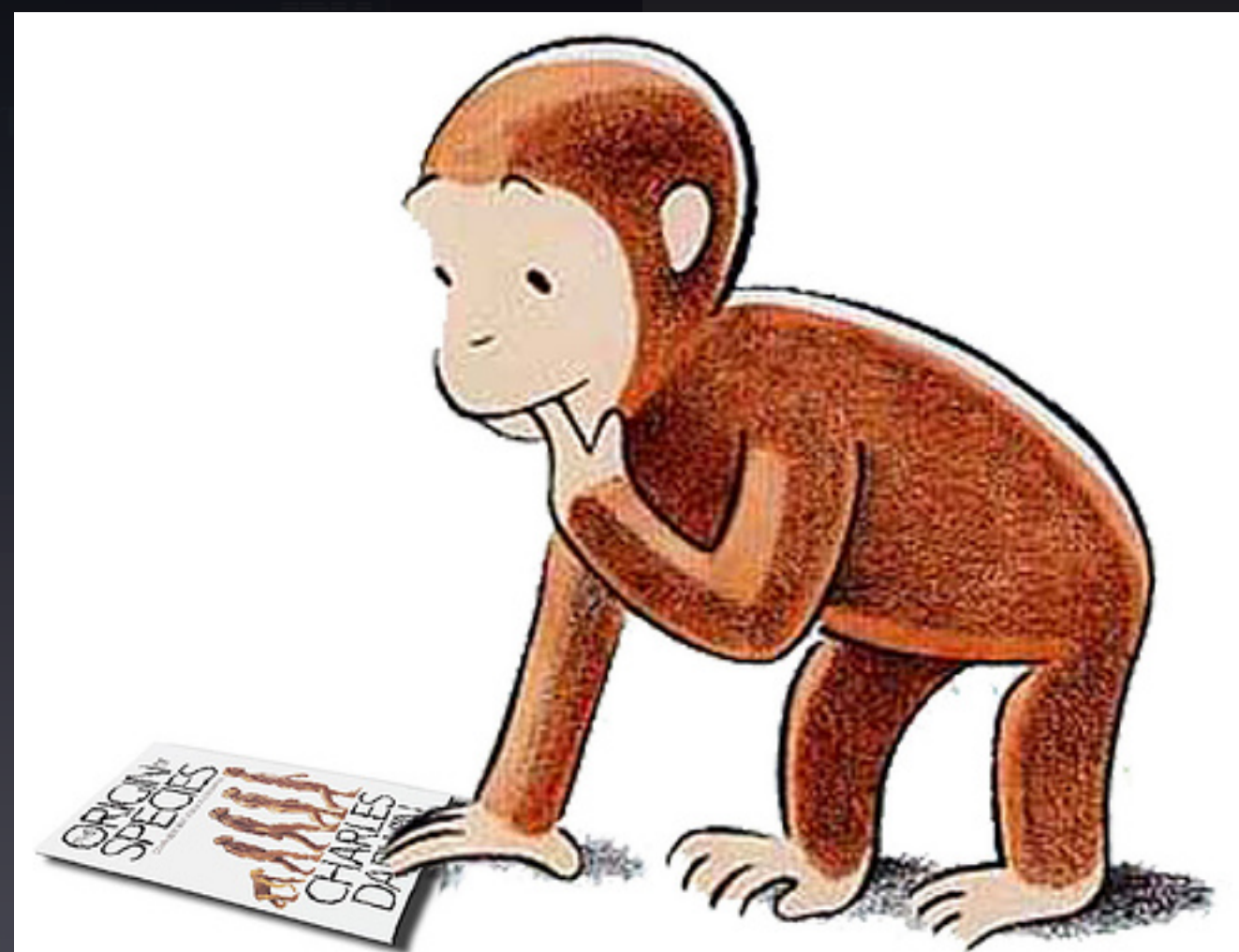


MAY 8, 2018 @ 12:09 PM 3,047 2 Free Issues of Forbes

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

Tirias Research Ad closed by Google

# IBM AC922

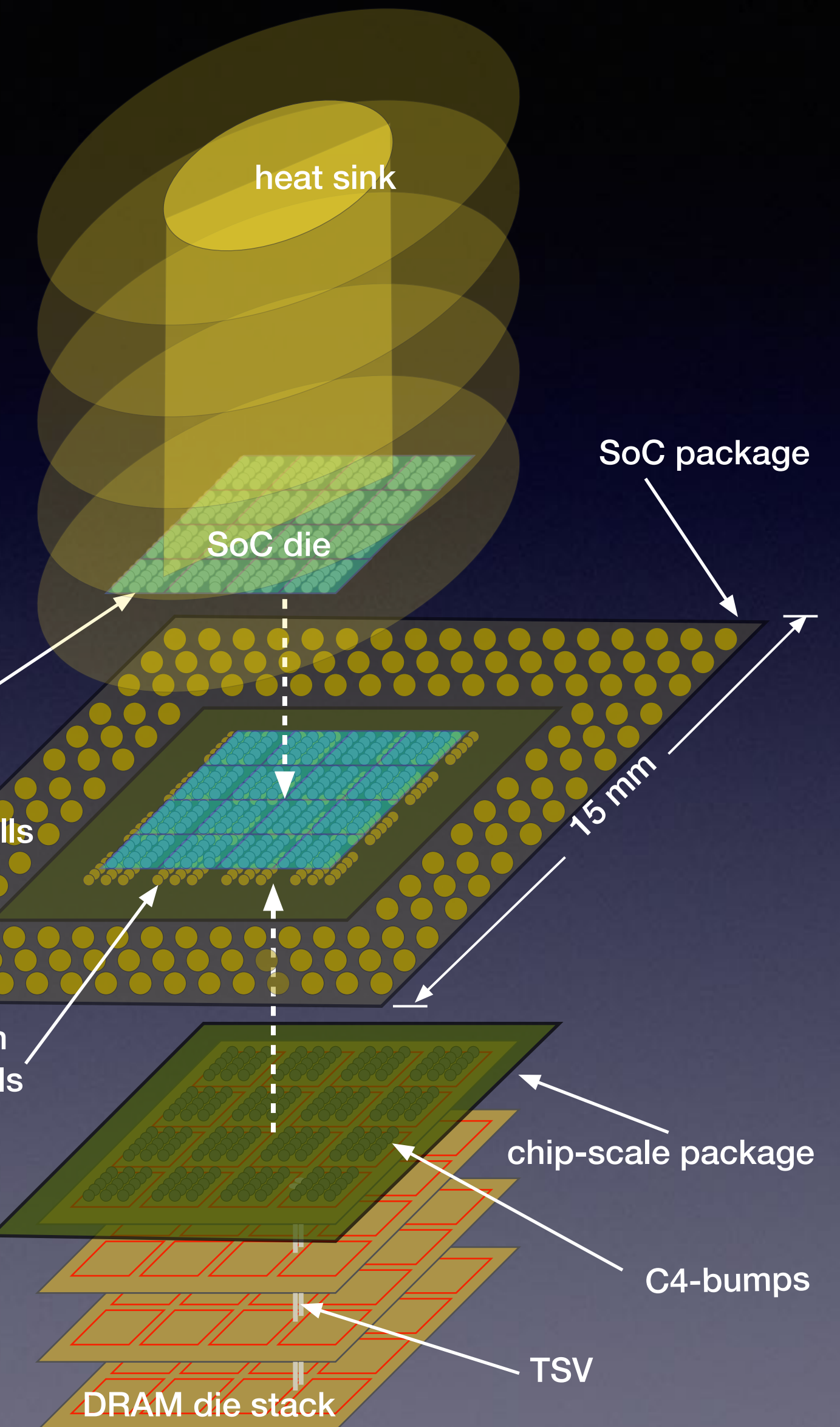
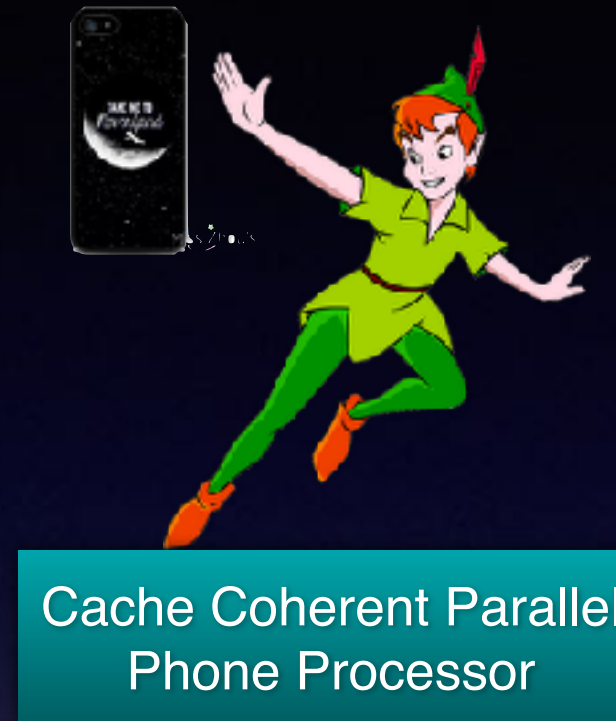
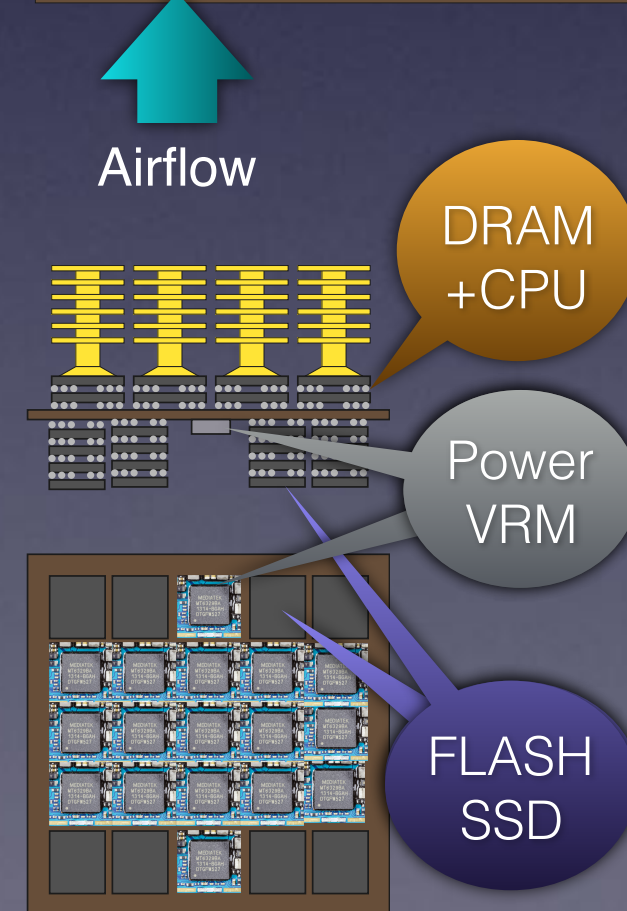
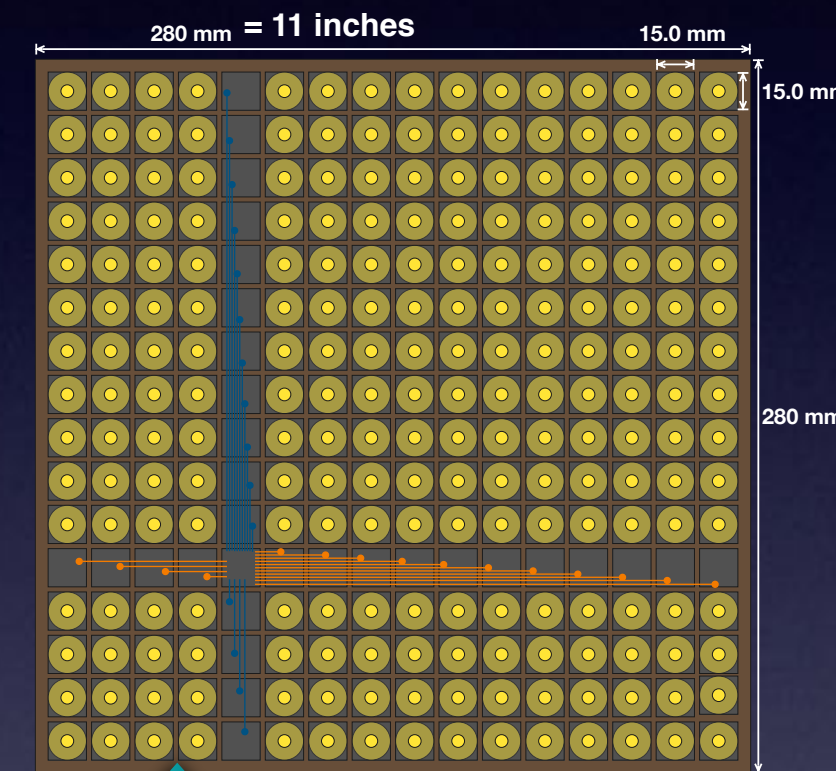


- \* Where does the energy go?
- \* How efficient is it really?
- \* Can we do better?

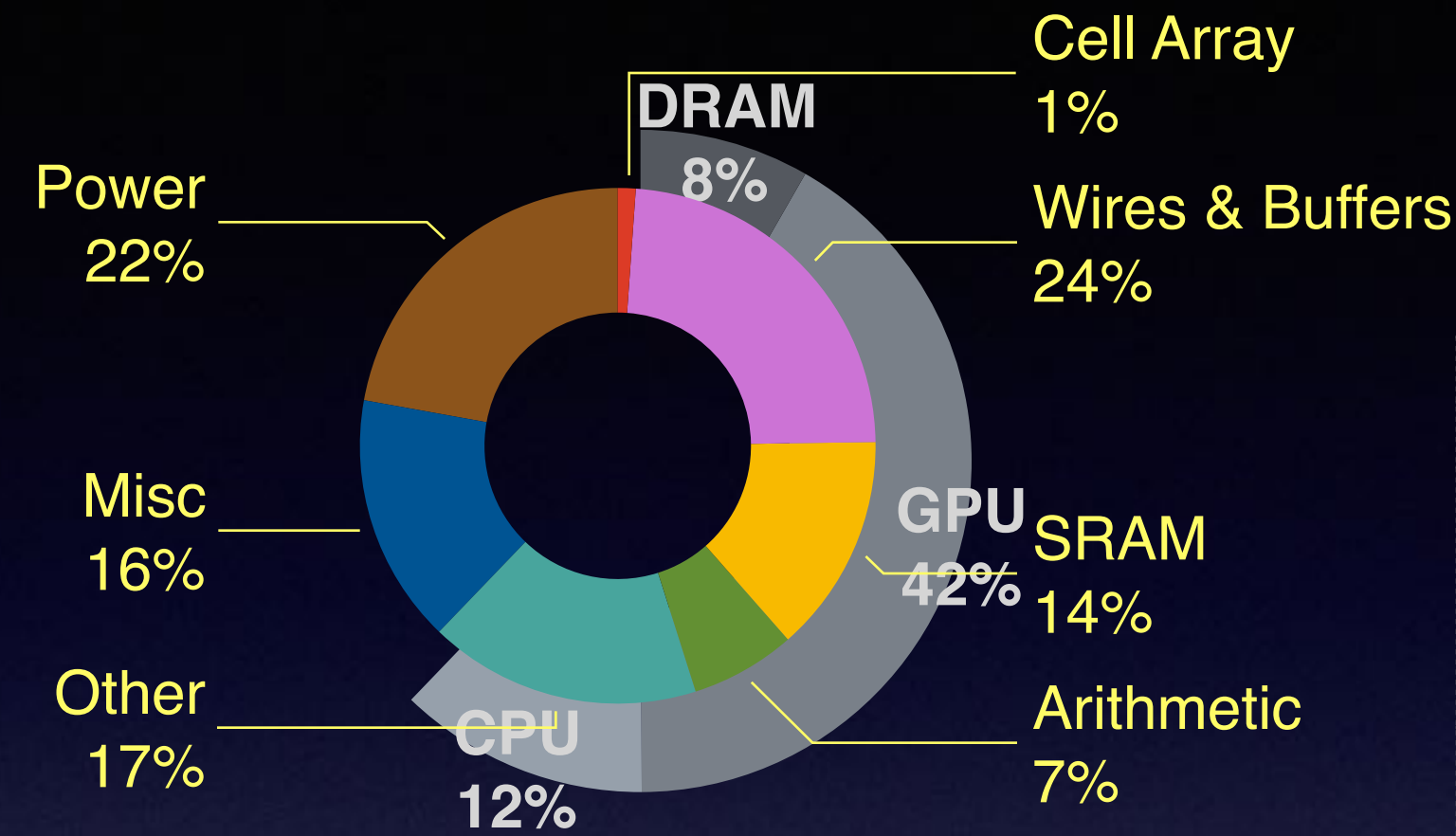
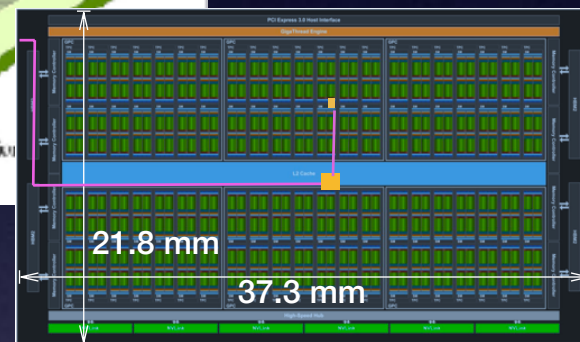
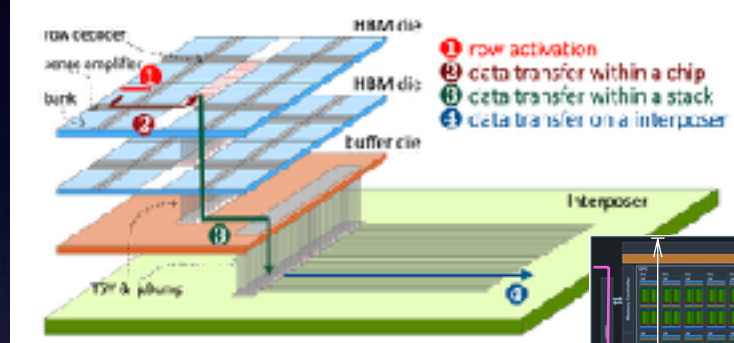
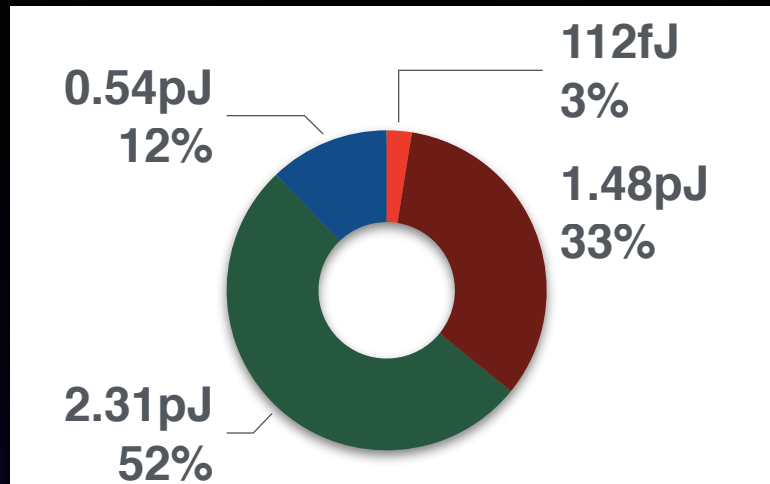
# Summary

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

**10x** more energy efficient  
 i.e. 150 GFLOPS/W 64-bit Float  
 600 GFLOPS/W 16-bit Float



# IBM AC922



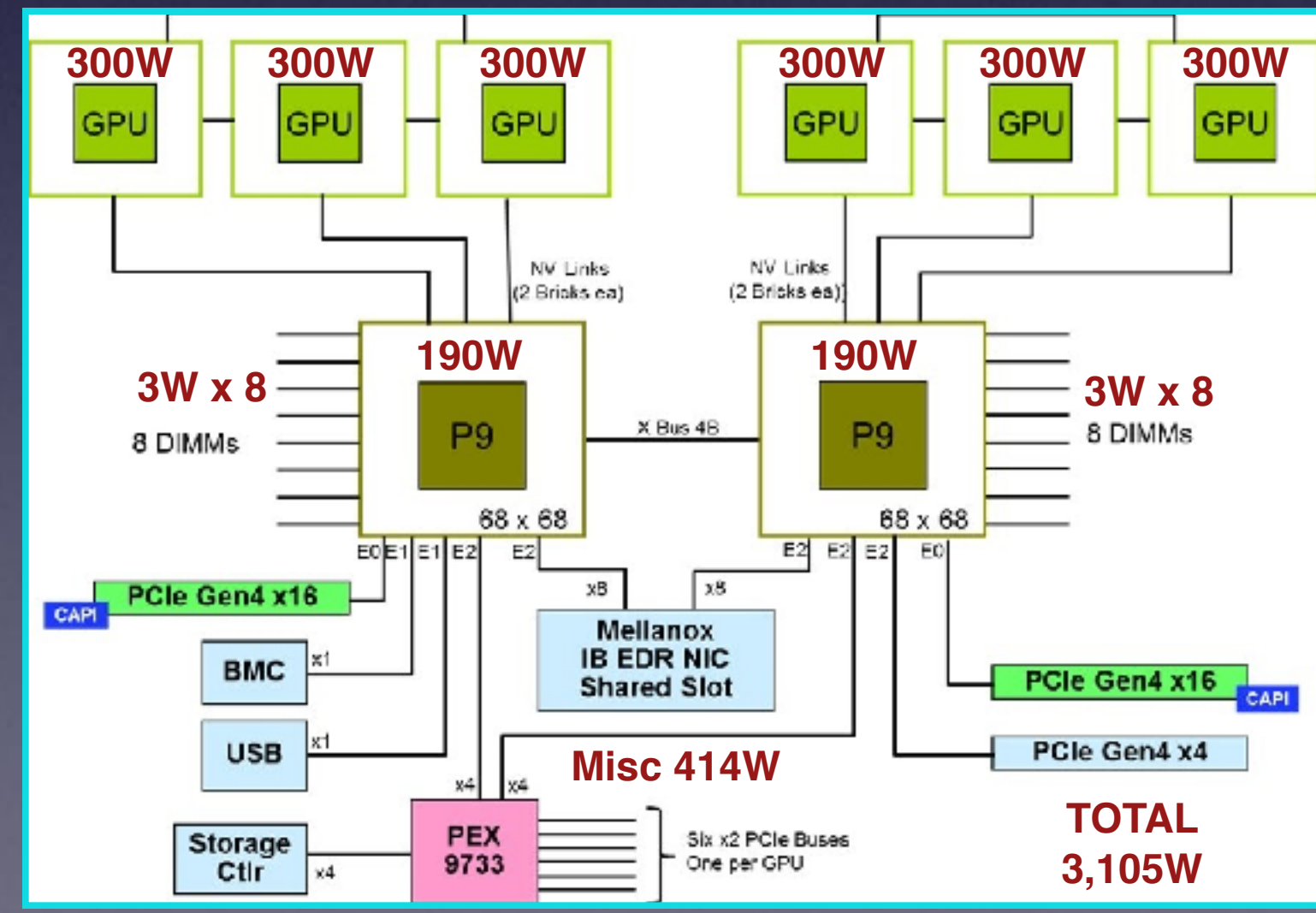
**POWER9 TO THE PEOPLE**  
December 5, 2017 Timothy Prickett Morgan

**Architecting an Energy-Efficient DRAM System For GPUs**  
Niladri 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, mconnor, donghyukl, djohnson, skeckler, nrhu, wdally}@nvidia.com

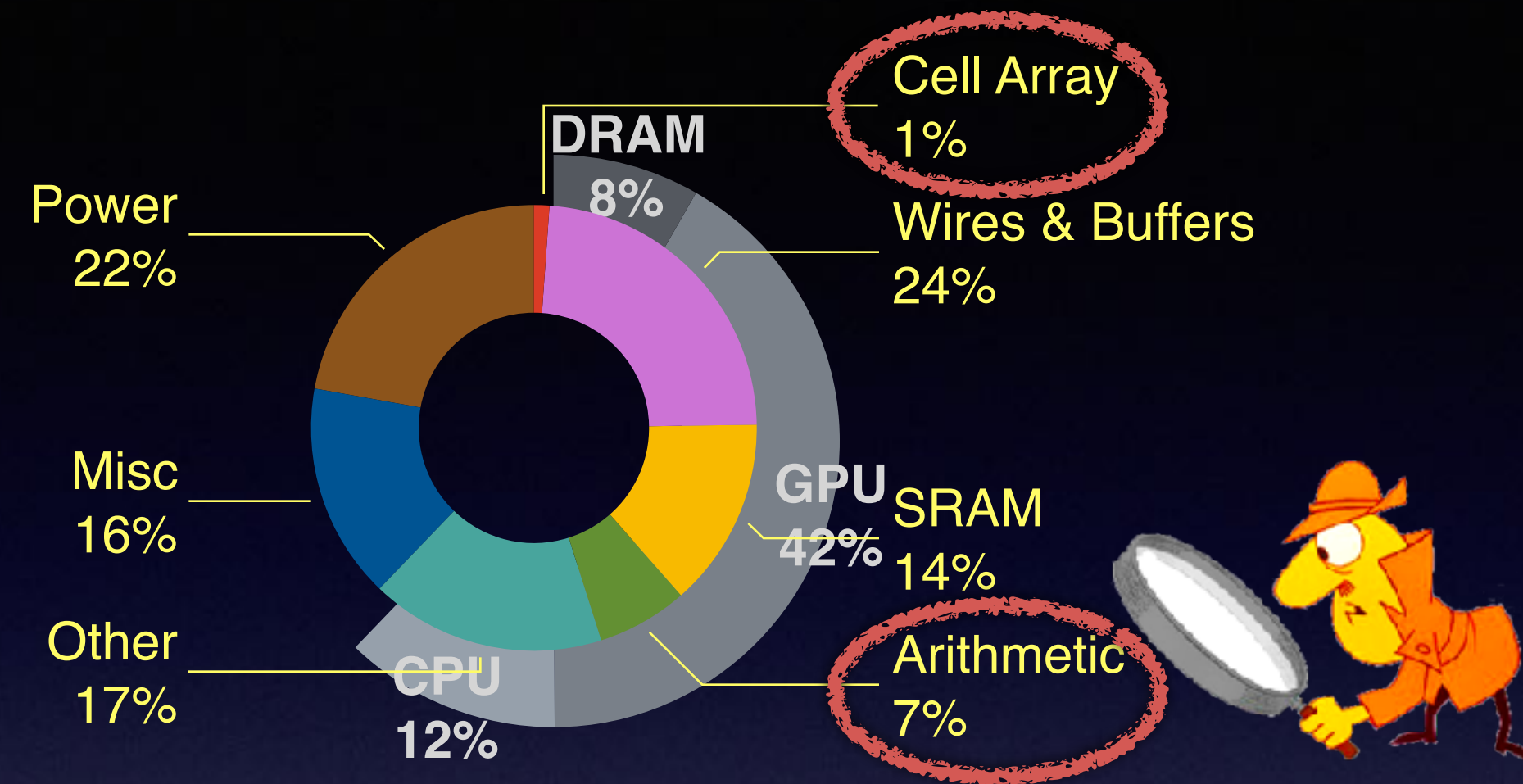
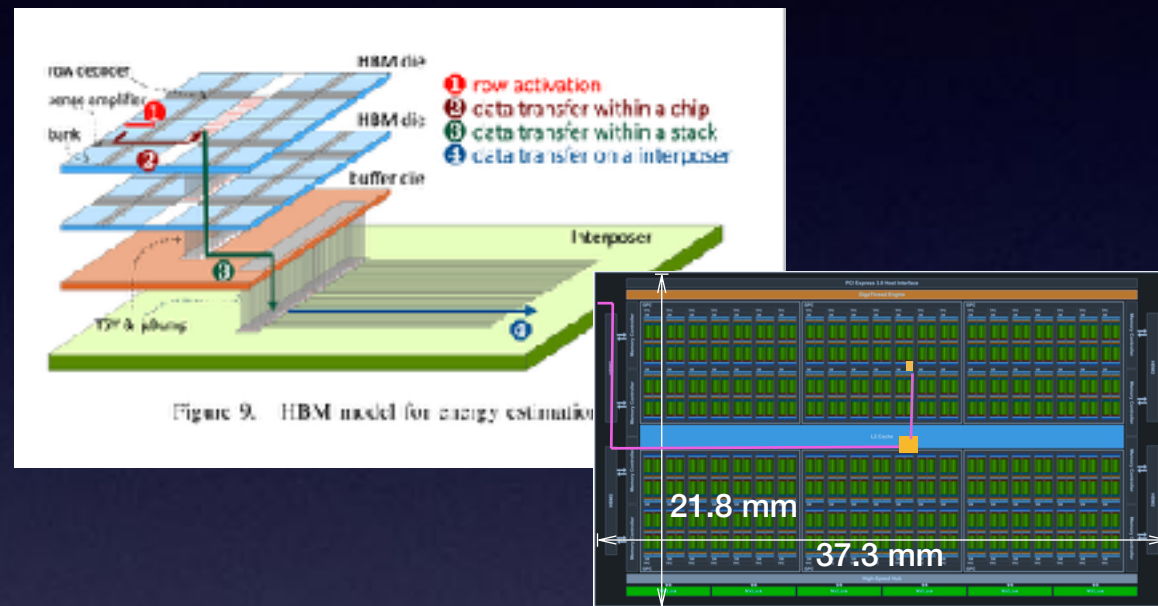
**GPUS AND THE FUTURE OF PARALLEL COMPUTING**  
Stephen W. Keckler, William J. Dally, David Hall, Michael Farber, David Black, Bill Dally

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

Process technology	2010	2017	2017
	40 nm	10 nm, high frequency	10 nm, low voltage
V <sub>DD</sub> (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 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 fJ	150 fJ/bit/mm	115 fJ/bit/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	0.54	1.0	3.9	35	209		
	0.025	row activation (per bit)	1.48	2.7			8.9	10.6
	dword / MADD	data transfer within a chip	2.3	4.2			16.6	
		data transfer within a stack	0.5	1.0			3.9	
NVIDIA V100 GPU	cache miss	2.7	5.0	19.6	215	1,293		
		global wire (35mm x 1 bit)	2.7	0.1			0.3	
		write L2 cache (72-bit SRAM)	2.7	2.7			10.6	
	dload / MADD	read L2 cache (72-bit SRAM)	0.2	12.5			17.6	49.0
		local wire (2.5mm x 1 bits)	2.4	2.4			9.4	9.4
		write vector register file (64-bit)	7.2	7.2			28.2	
	MADD	read 3 vector operands (SRAM)	8.7	8.7			18.3	34.1
		floating-point MADD (64-bit FP)	2.4	2.4			9.4	
		write 1 vector result (SRAM)	14.0	14.0			55	
		other	memory interface, control, etc.	14.0			14.0	14.0
PCIe Card	2	VRM conversion efficiency	85%	9.6	9.6	38	225	
	6	fan (12V, 0.5A) or water pump	6	3.1	3.1	12	72	
	<b>GPU Card Subtotal</b>	<b>300</b>	<b>76.6</b>	<b>76.6</b>	<b>300</b>	<b>1,800</b>		
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		
Chassis	2	other		8.5		100		
	↑	CPU power VRM	85%	2.5	29	58	58	
	#of	misc.—PCIe, fans/pumps, etc.		18		414	414	
	primary power supply	85%	17.2		405	405		
	<b>15.1 GFLOP/W</b>	<b>550</b>	<b>140</b>	<b>Grand Total</b>		<b>3,105</b>		

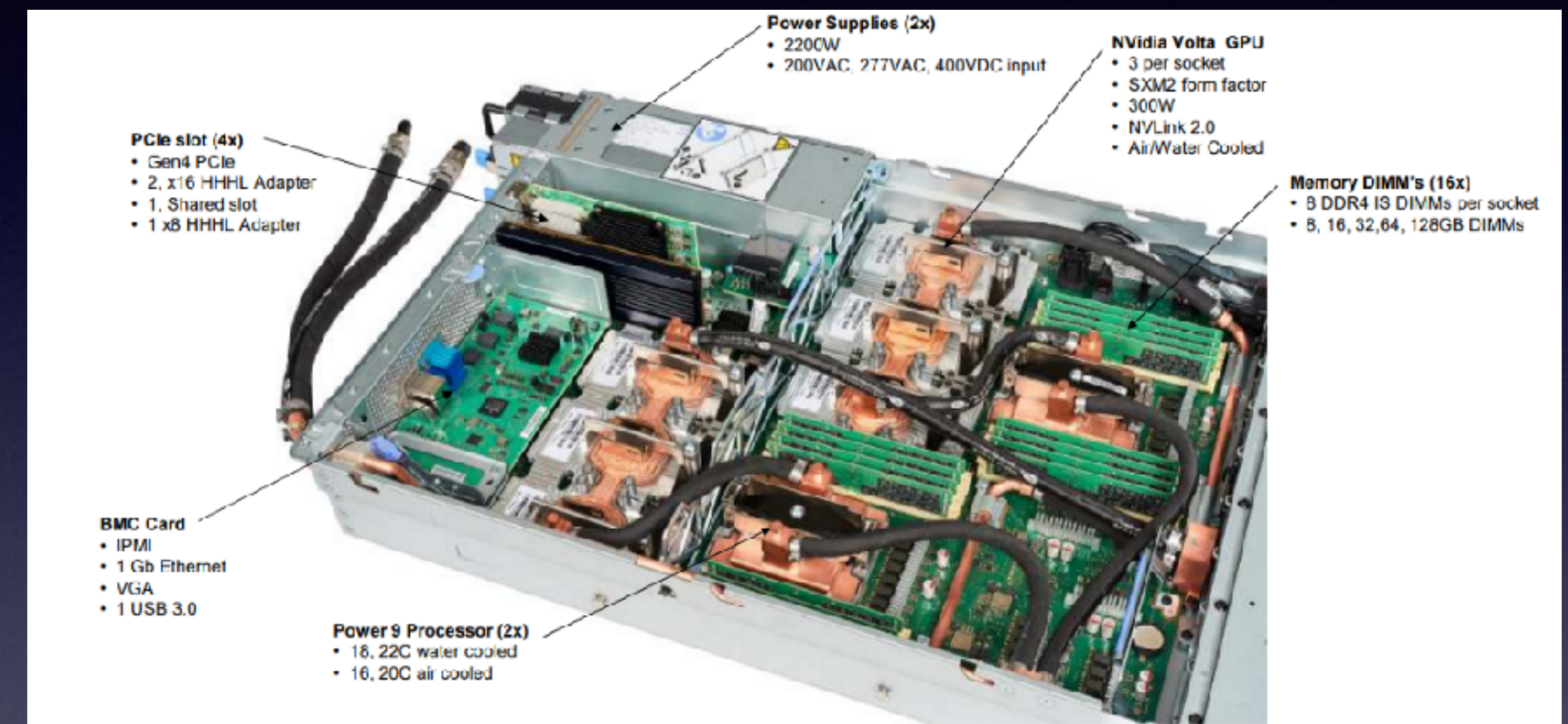
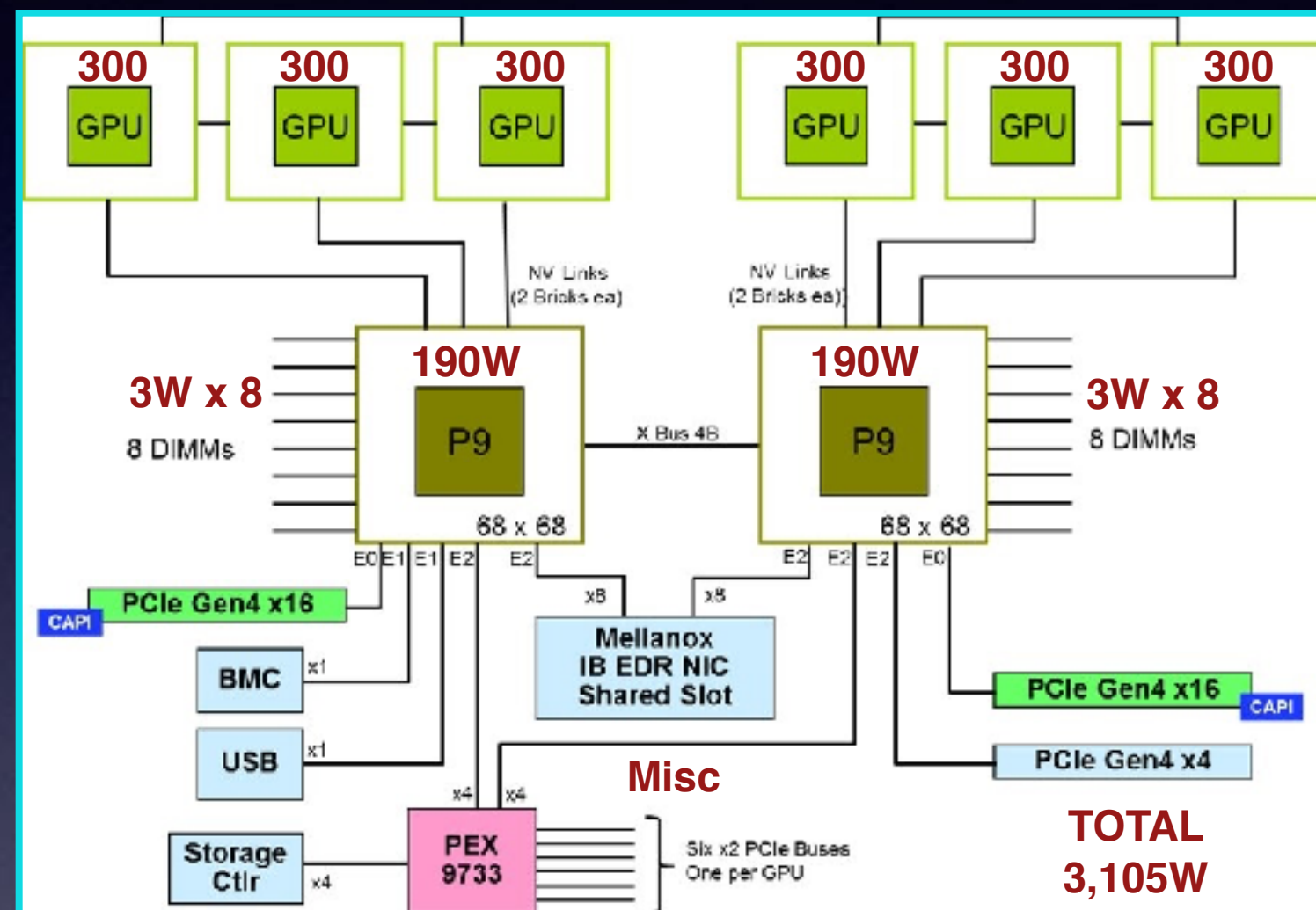


8% DRAM access & arithmetic datapath

38% moving data & staging in SRAM

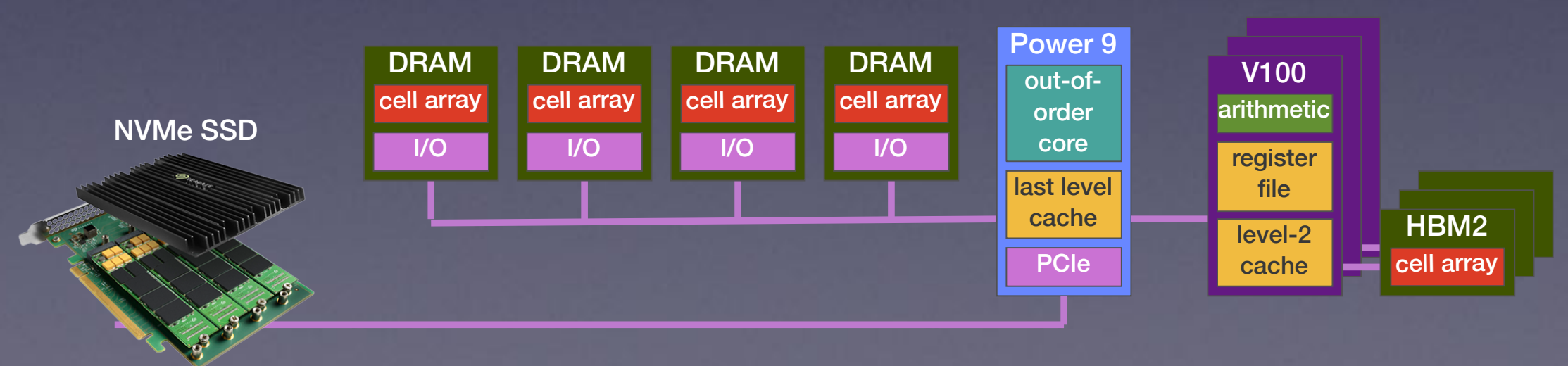
- most of*
- \* Where does the energy go?
  - \* How efficient is it really?
  - \* Can we do better?

# Data-Streaming Architecture



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

**NOTE:** coherence  $\neq$  always communicating through shared memory i.e. asynchronous DMA memcpy





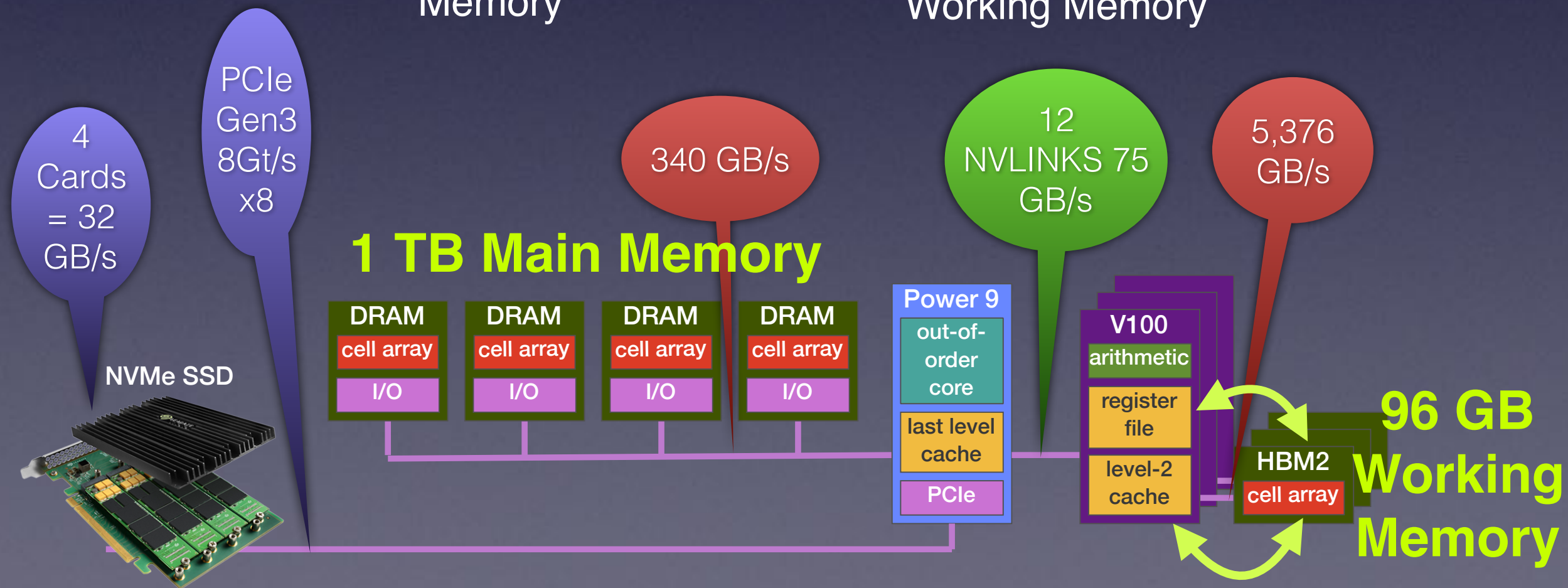
# Data-Streaming Architecture



SSD to Main Memory



Main Memory to Working Memory



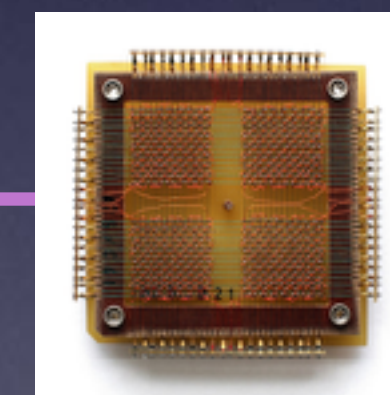
**Example**  
 IBM 7094, \$2M USD  
 0.5 MHz clock cycle  
 0.15 MB main memory



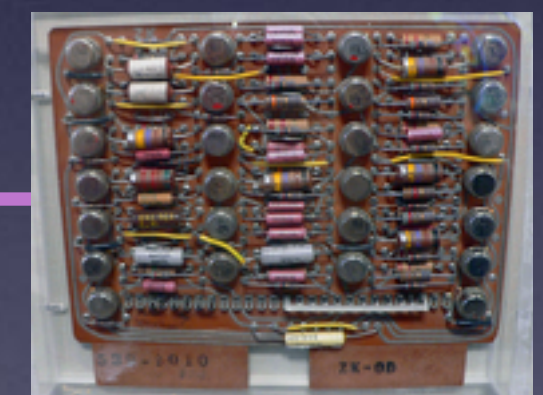
Tape Storage



Disk Drive



Core Memory



Logic Gates

1962

# Architectures

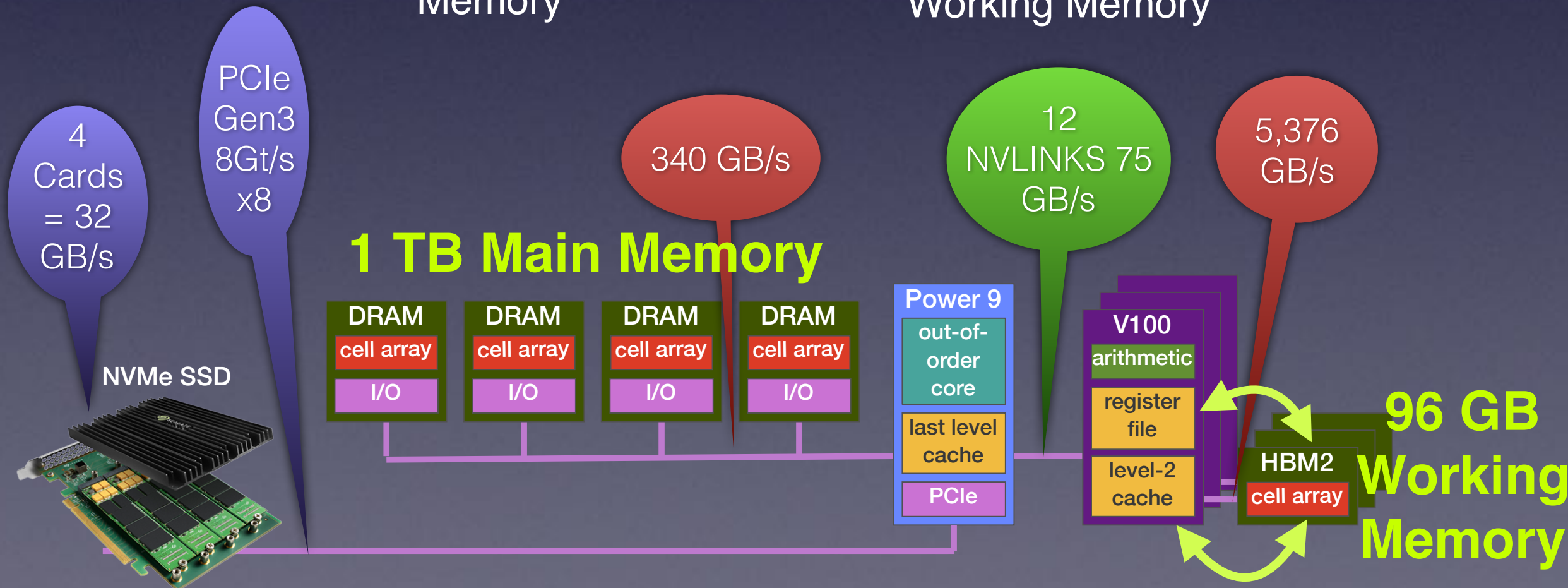
data-streaming



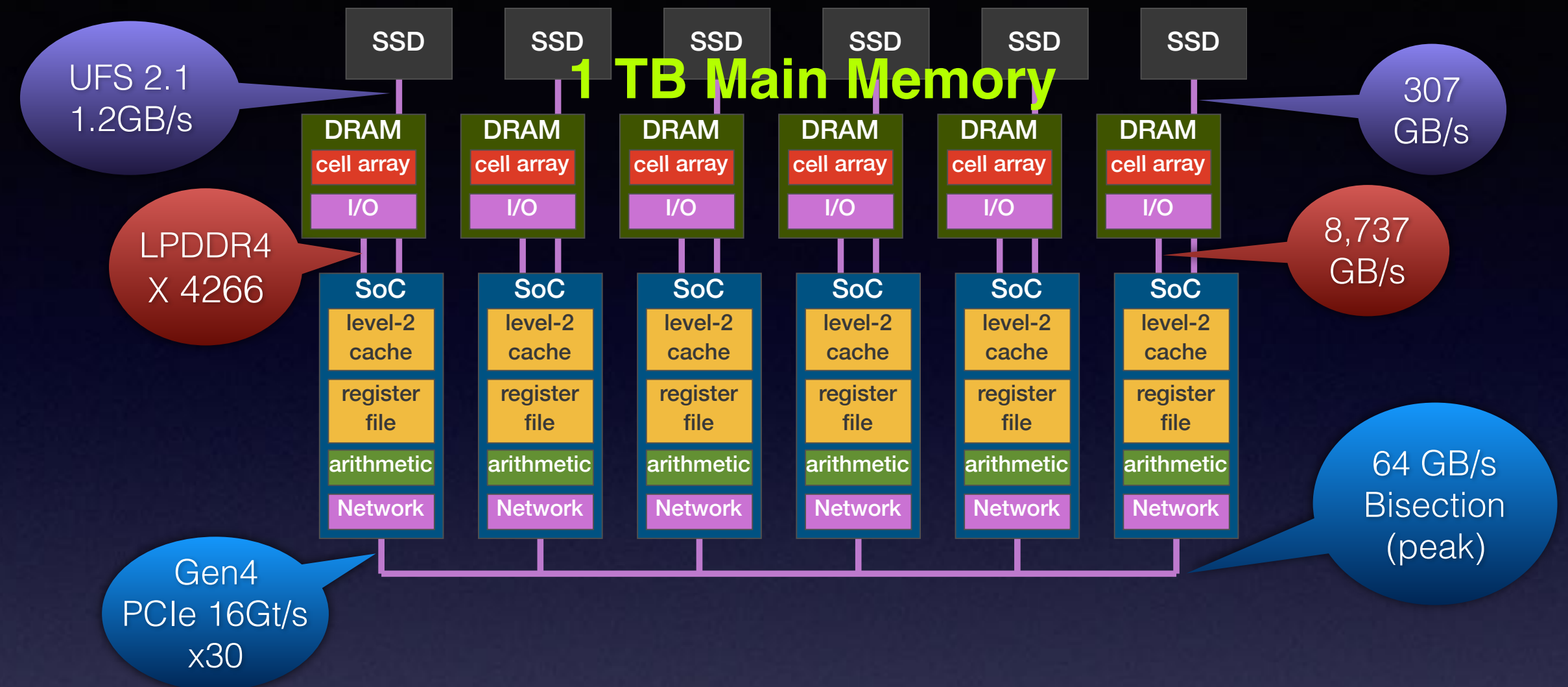
SSD to Main Memory



Main Memory to Working Memory



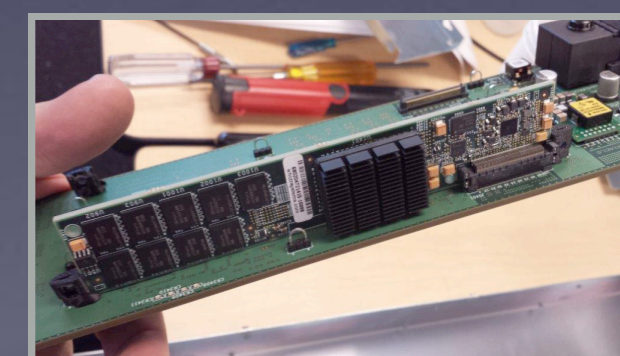
near-data processing



## Near-data processing: What it is and why you need it

Near-data processing (NDP) is a simple concept: Place the processing power near the data, rather than shipping the data to the processor. It's also inevitable. Here's why it's coming to your datacenter.

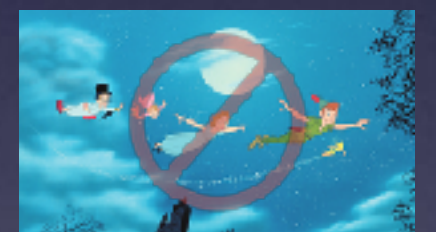
By Robin Harris for Storage Bits | October 19, 2016 -- 12:15 GMT (13:15 BST) | Topic: Storage



Oracle Labs RAPID parallel computer

## A Many-core Architecture for In-Memory Data Processing

- |  |   |  |
|--|---|--|
| Sandeep R Agrawal<br>sandeep.r.agrawal@oracle.com<br>Oracle Labs | Sam Idicula<br>sam.idicula@oracle.com<br>Oracle Labs                    | Arav Bhaswan<br>arav.bhaswan@oracle.com<br>Oracle Labs   |
| Evangelos Vlachos<br>evangelos.vlachos@oracle.com<br>Oracle Labs | Venkatraman Govindaraju<br>venkat.govindaraju@oracle.com<br>Oracle Labs | Venkateshwaran Venkatasubramanian<br>venkateshwaran.venkatasubramanian@oracle.com<br>Oracle Labs |
| Cagri Balcesen<br>cagri.balcesen@oracle.com<br>Oracle Labs       | Georgios Giannakis<br>georgios.giannakis@oracle.com<br>Oracle Labs      | Charlie Roth<br>charlie_roth@atoo.com<br>Oracle Labs   |
| Nipun Agarwal<br>nipun.agarwal@oracle.com<br>Oracle Labs         | Eric Sedlar<br>eric.sedlar@oracle.com<br>Oracle Labs                    |  |



# Cache Coherent Parallel Phone Processor

Qualcomm® Snapdragon® 835  
Mobile PC Platform

First 10nm  
Qualcomm® Processor

Over 3 Billion  
Transistors

Up to 30%  
Performance Improvement  
at Same Power Consumption

Qualcomm makes AI priority second only to 5G

Galaxy S8 | S8+

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

A9

Apple application processor  
fabricated by TSMC using 16nm FinFET process technology

RYZEN EPYC

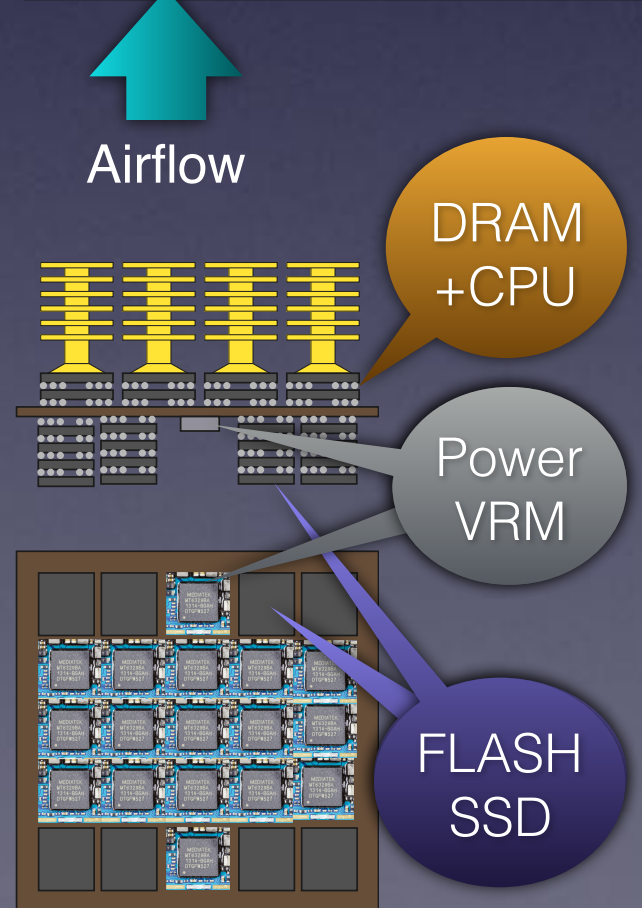
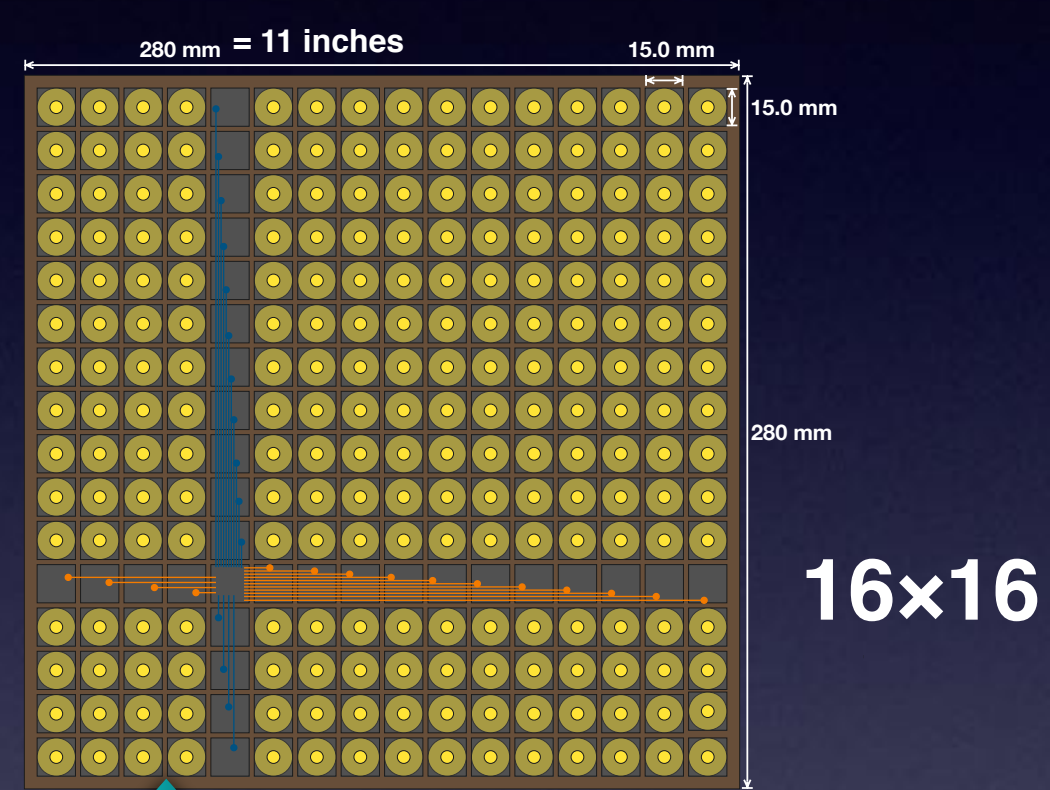
BREAKING CONSTRAINTS OF MOORE'S LAW

- Revolutionary Infinity Fabric
- High-performance, scalable links
- Translates high-level instructions that increase real-world performance
- Maximizes silicon yields

CC-NUMA

- A combination of NUMA and COMA
- Initially static data distribution, then dynamic data migration
- Cache coherency problem is to be solved
- COMA and CC-NUMA are used in newer generation of parallel computers
- Examples: Convex SPP1000, Stanford DASH, MIT Alewife

Intel PC



Gen 4 PCIe

- 16 Gt/s @ 10-12 inches
- Power efficient ( $\approx 5\text{pJ/bit}$ )
- Strong industry support
- Volume manufacturing

Flattened Butterfly: A Cost-Efficient Topology for High-Radix Networks

John Kim, William J. Dally  
Computer Systems Laboratory  
Stanford University, Stanford, CA 94305  
jk12, billd@csa.stanford.edu

Deennis Abis  
Cray Inc.  
Chippewa Falls, WI 54729  
dabts@cray.com

Power & Cooling

- Like 800W processor
- But already spread out

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

Oracle's SPARC T7 and M7 Servers

New Platform for Secure Computing

POWER8 PROCESSOR

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



# Advantages



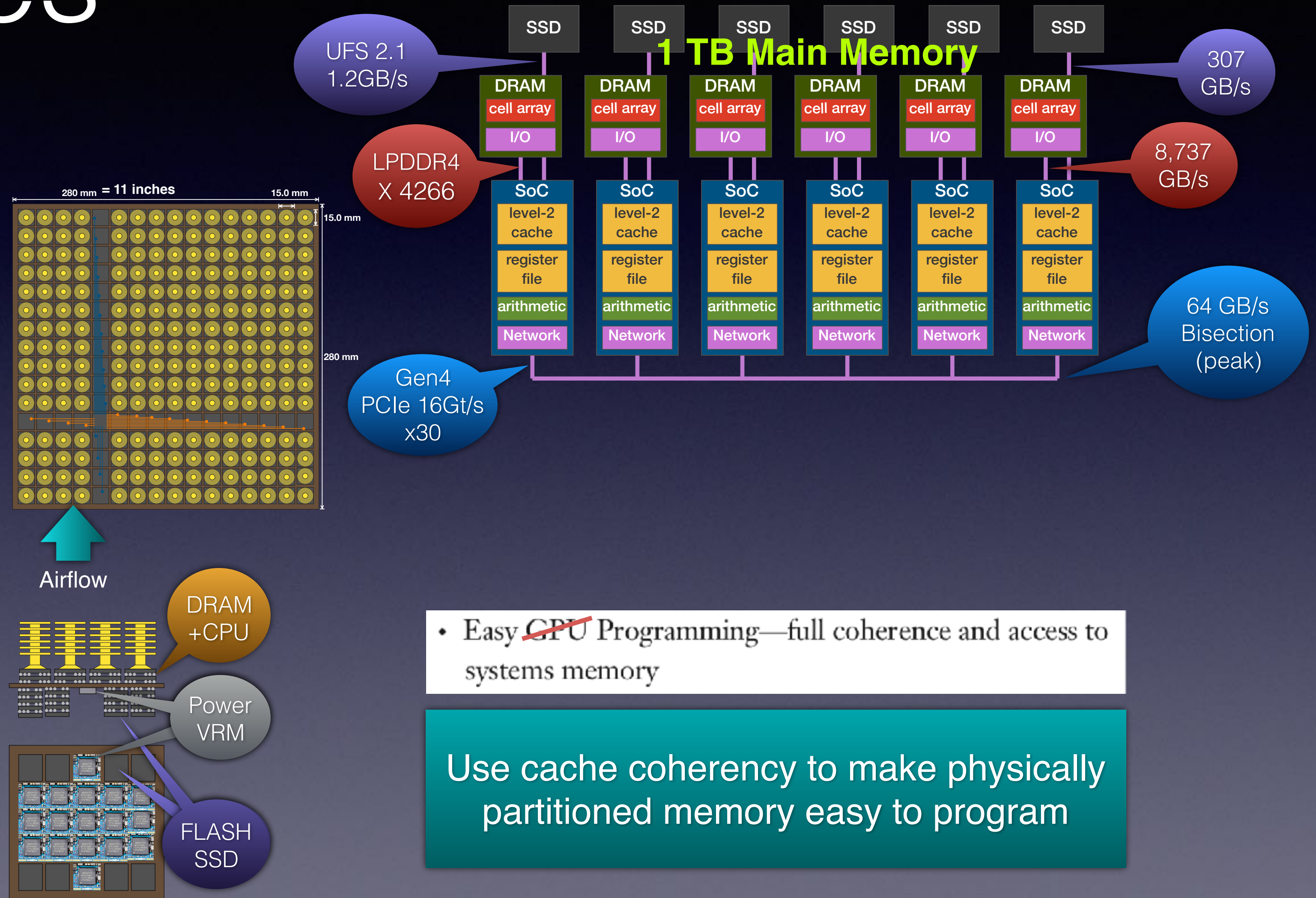
SSD to Main Memory



Main Memory to Working Memory



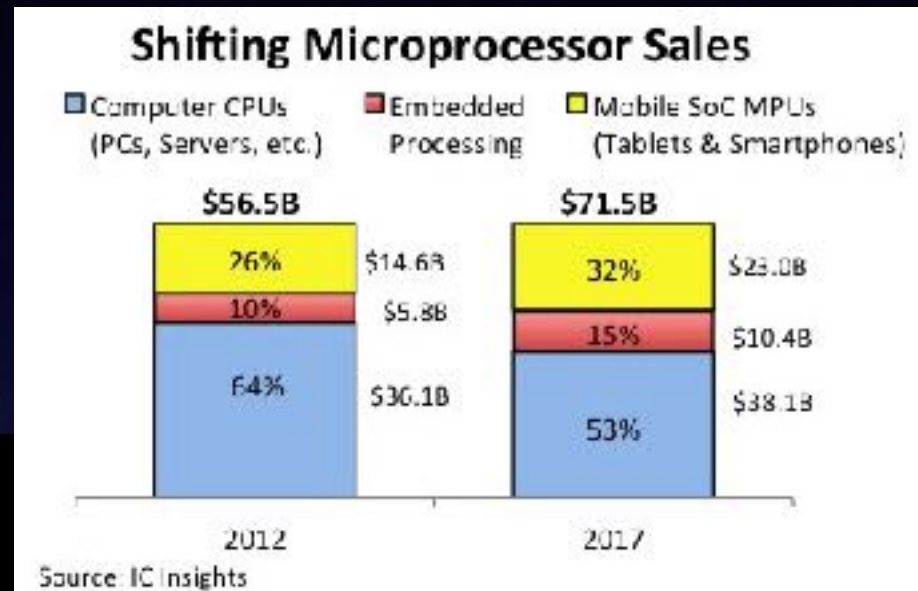
near-data processing



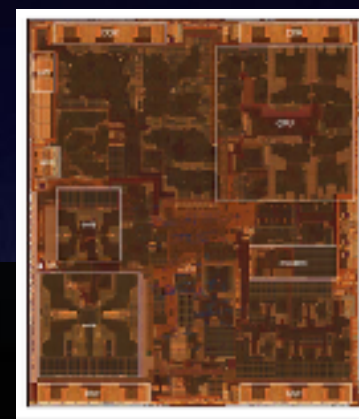
- Easy ~~GPU~~ Programming—full coherence and access to systems memory

Use cache coherency to make physically partitioned memory easy to program

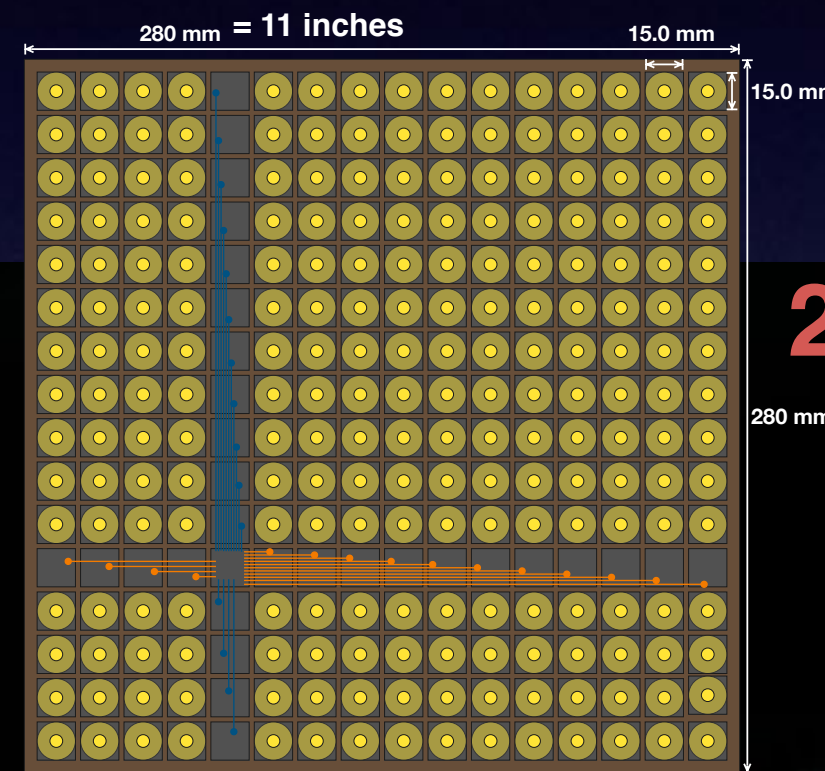
# Best Consumer Technology



Snapdragon 835 on Samsung 10LPE.



72.3mm<sup>2</sup>



27 March 2017

### Qualcomm® Snapdragon™ 835 Mobile PC Platform

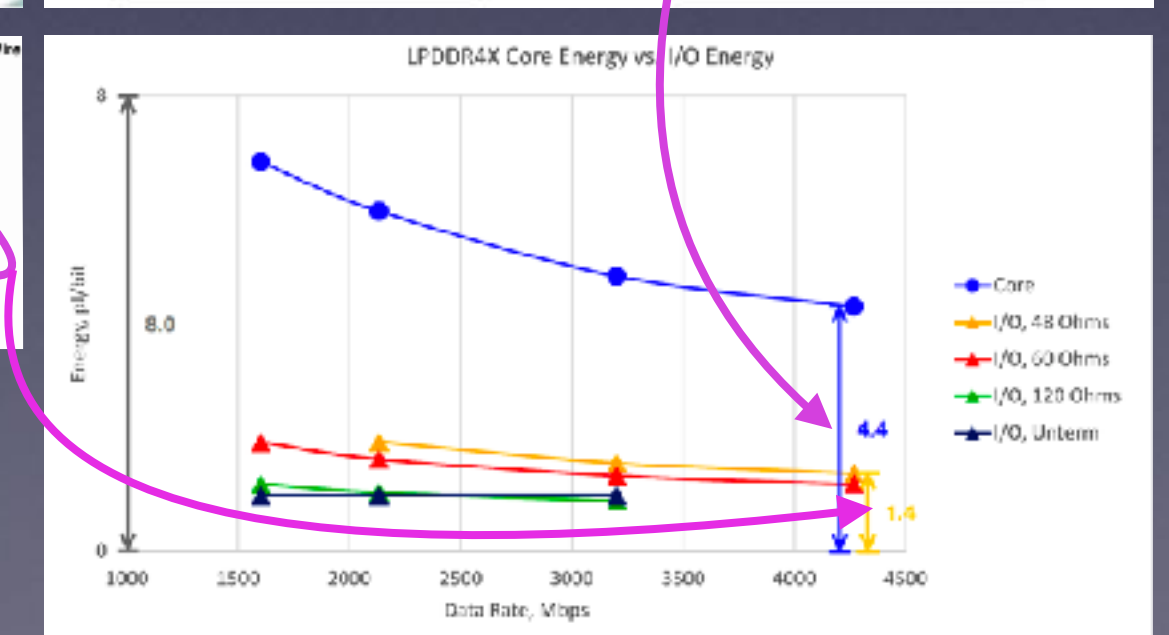
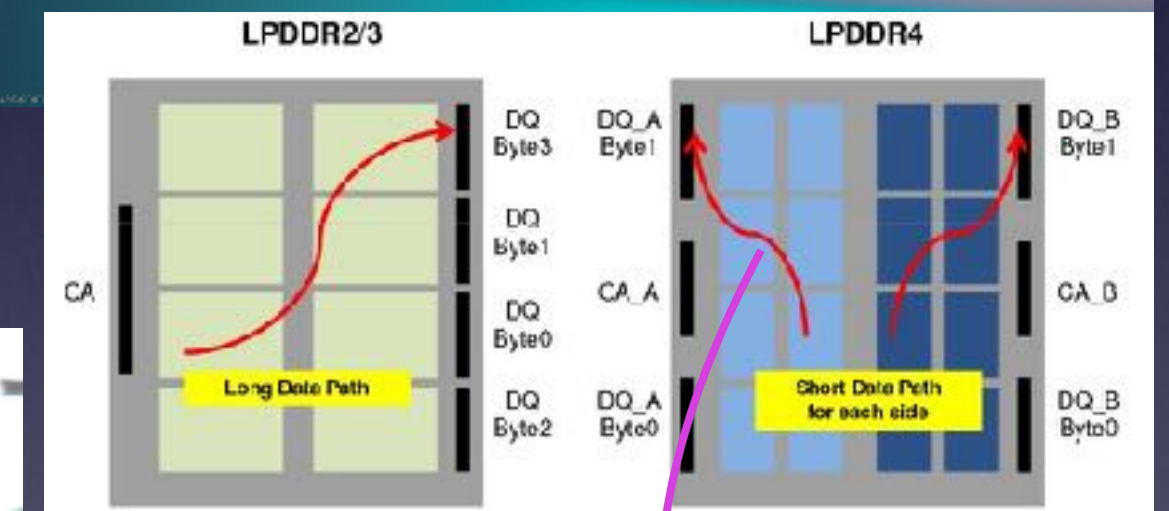
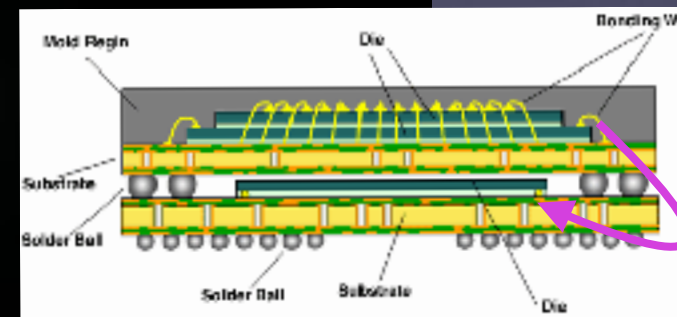
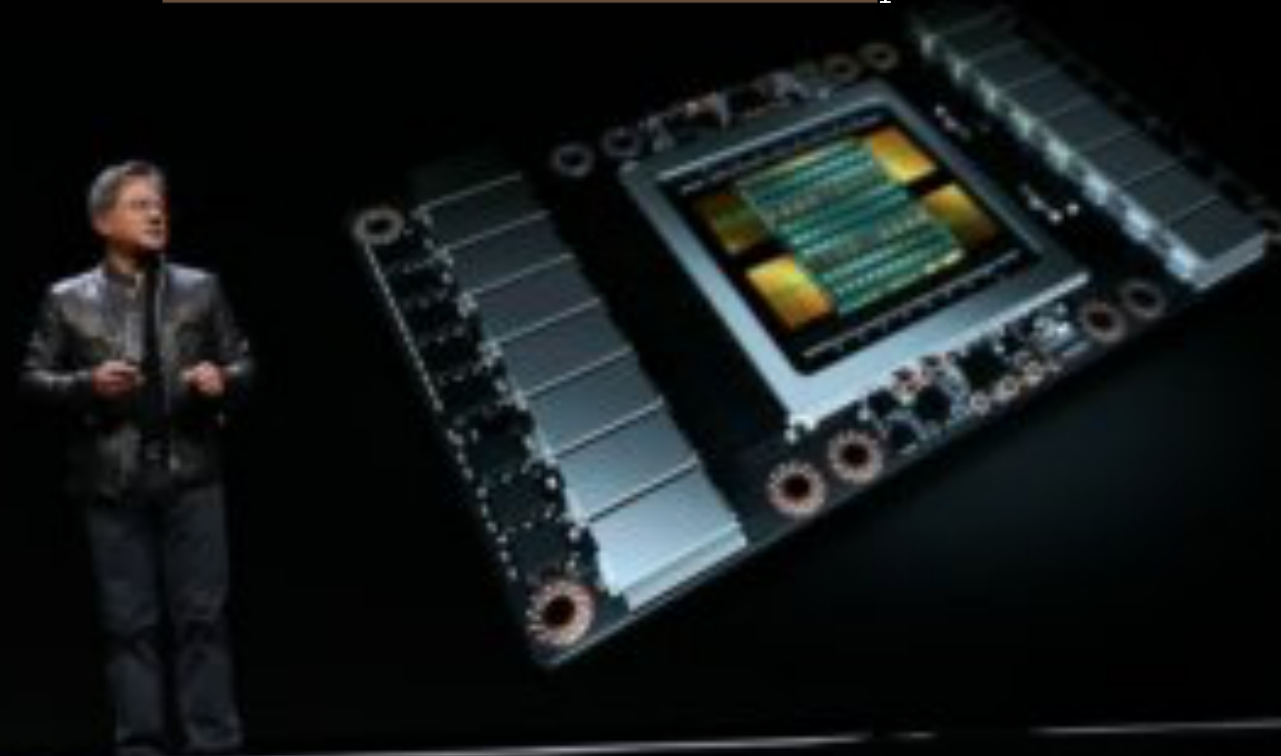
- First **10nm** Commercially Announced SoC
- Over **3 Billion** Transistors
- Up to **30%** more room on the board! Small footprint, thermal efficiency.

## ANNOUNCING TESLA V100

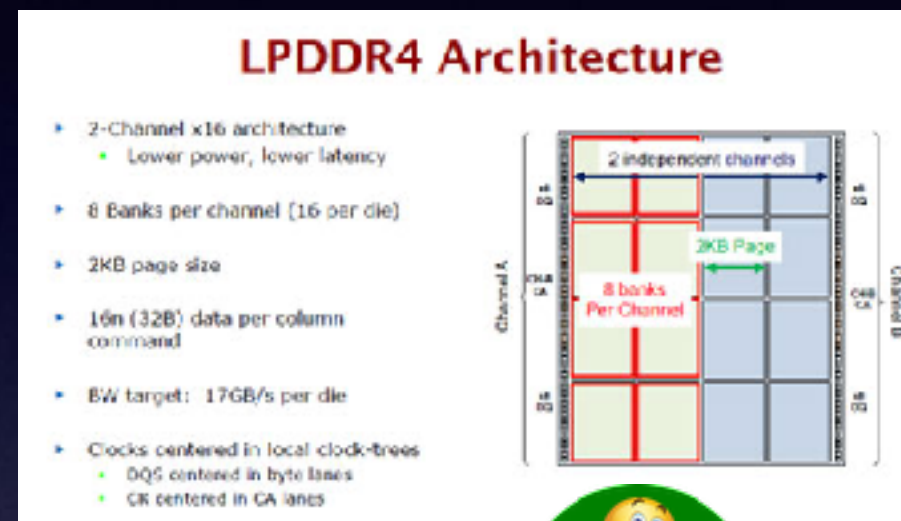
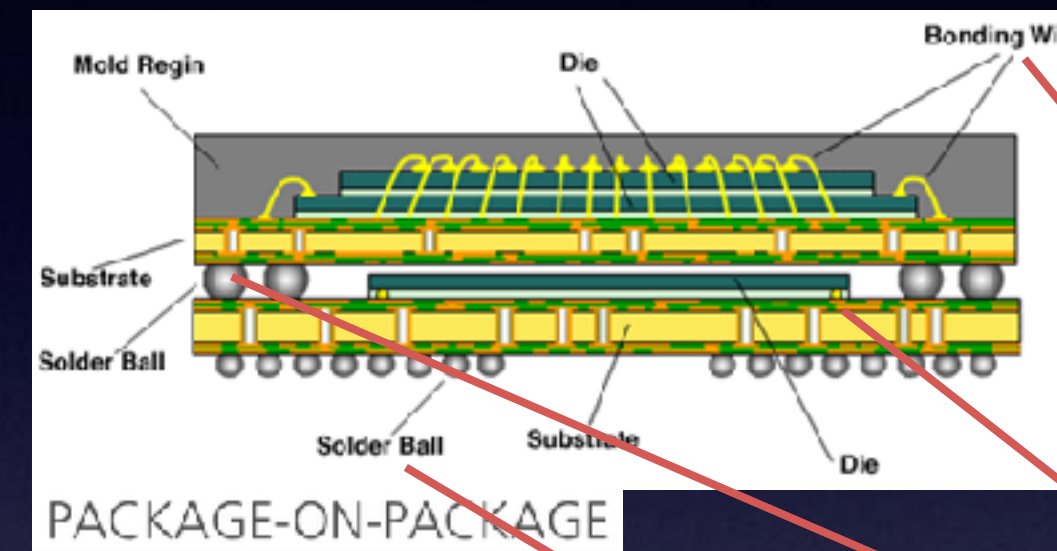
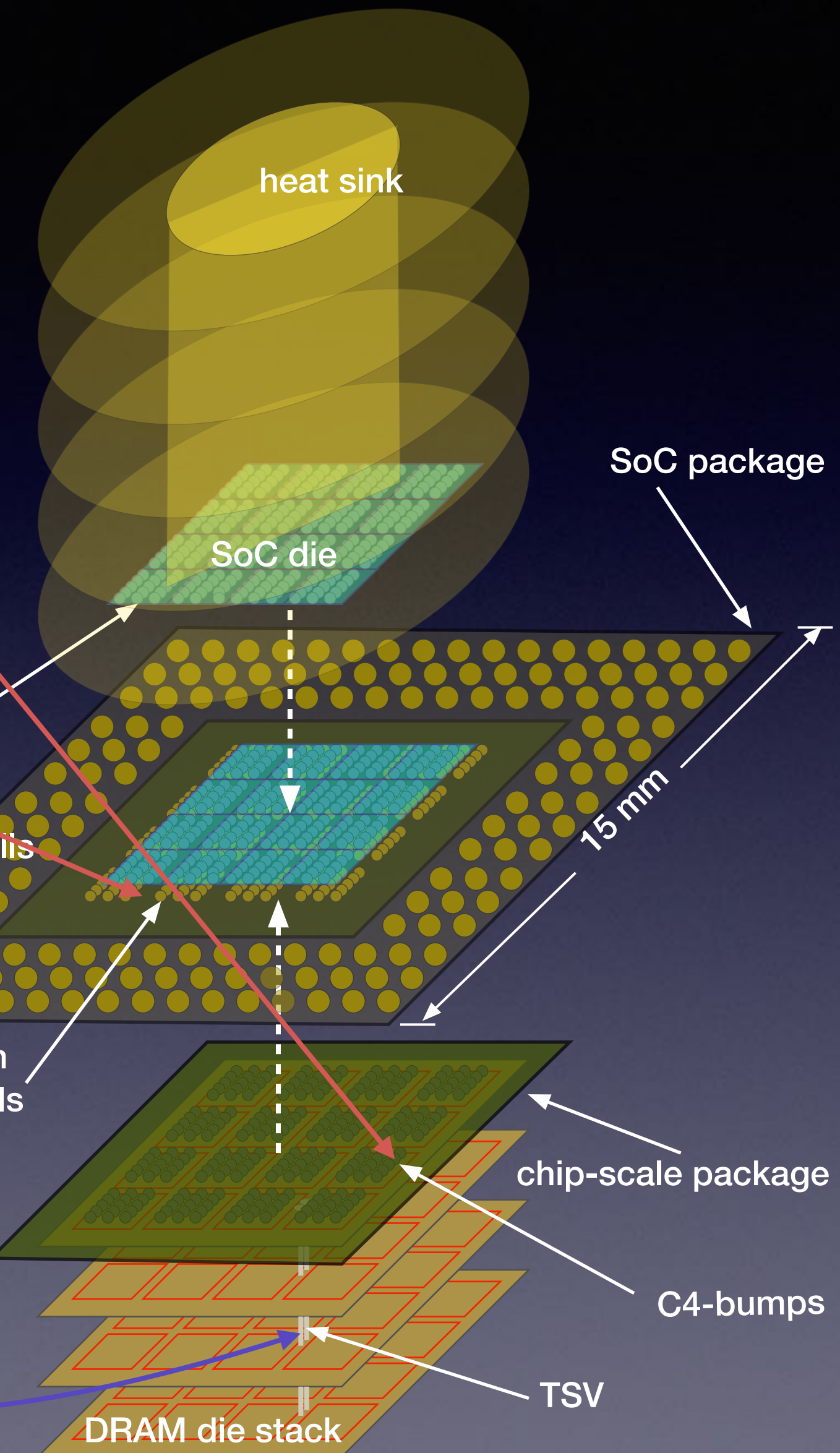
GIANT LEAP FOR AI & HPC  
VOLTA WITH NEW TENSOR CORE

21B xtors - **TSMC 12nm FFN** - 815mm<sup>2</sup>  
5,120 CUDA Cores  
7.5 FP64 TFLOPS | 15 FP32 TFLOPS  
NEW 120 Tensor TFLOPS  
20MB SM RF | 16MB Cache | 16GB HBM2 @ 900 GB/s  
300 GB/s NVLink

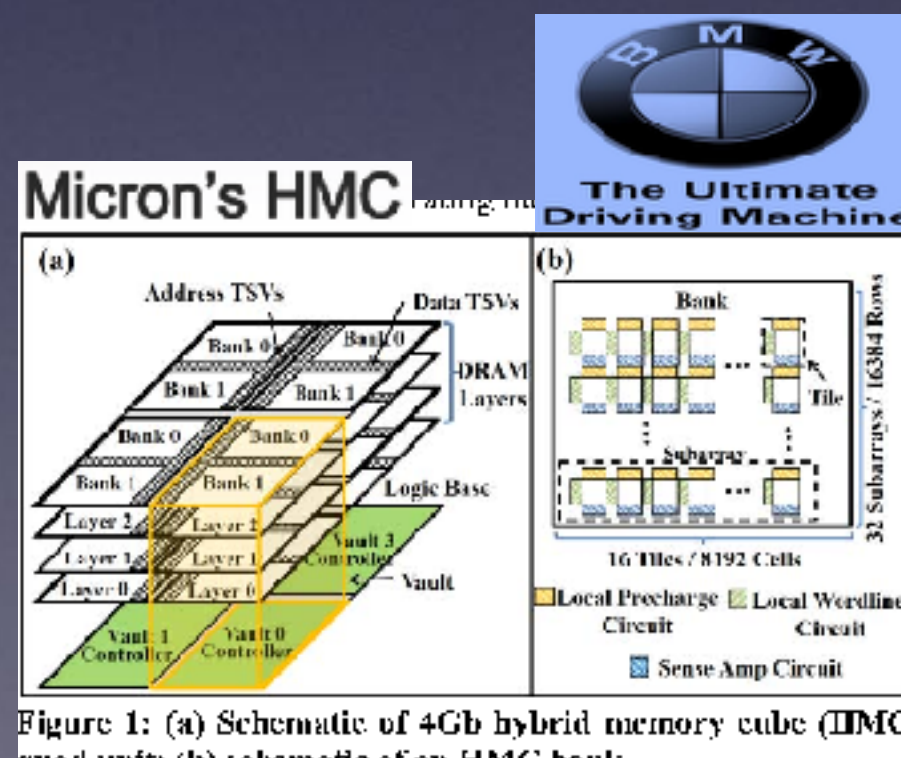
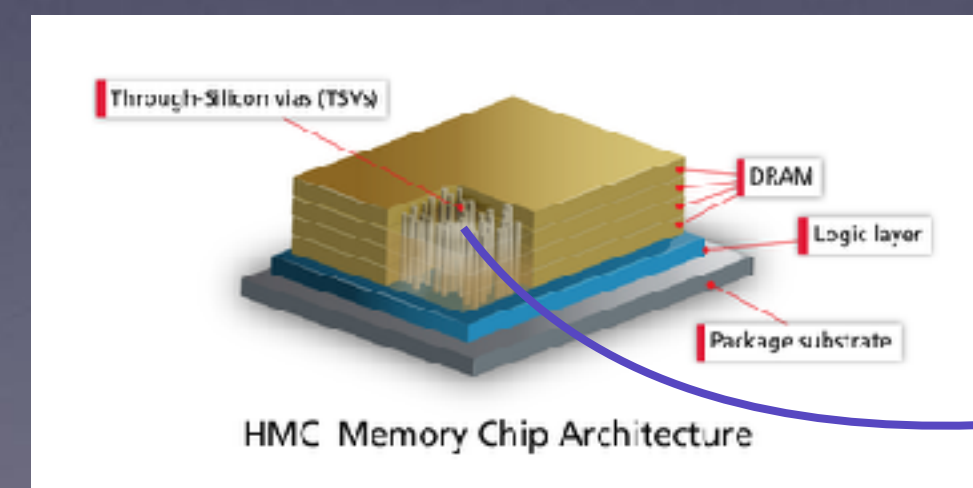
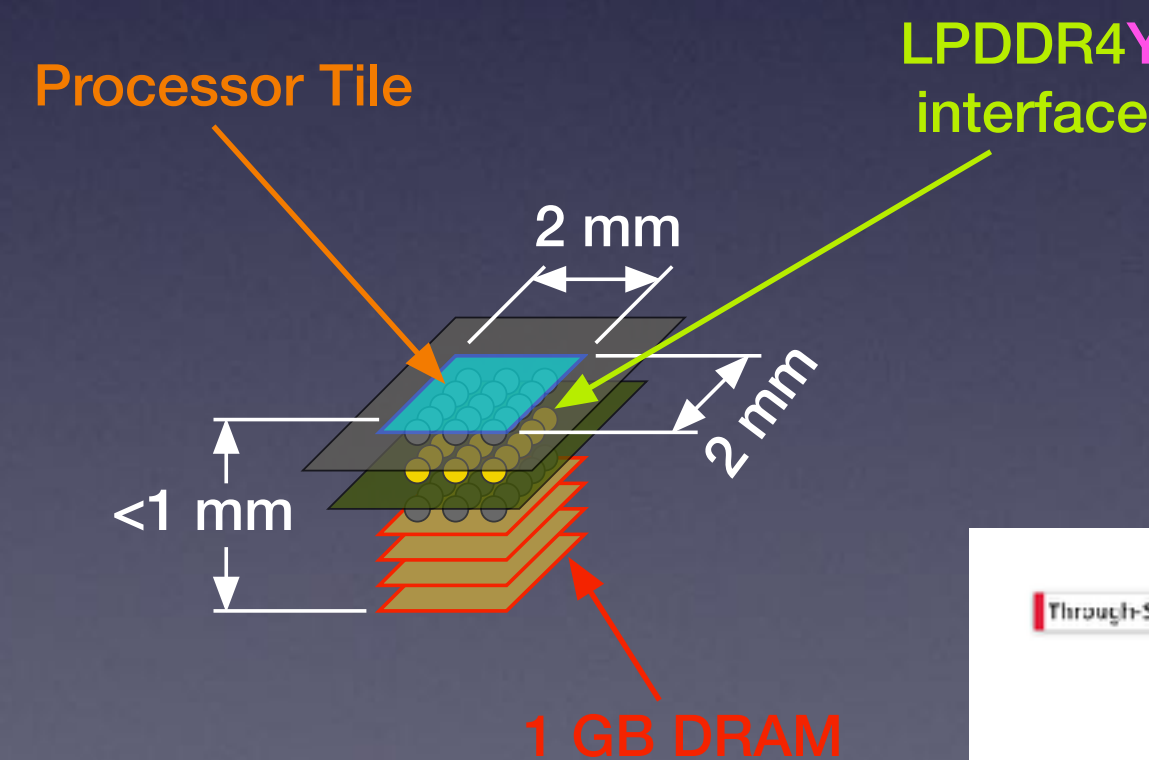
7 December 2017



# Processor over Memory

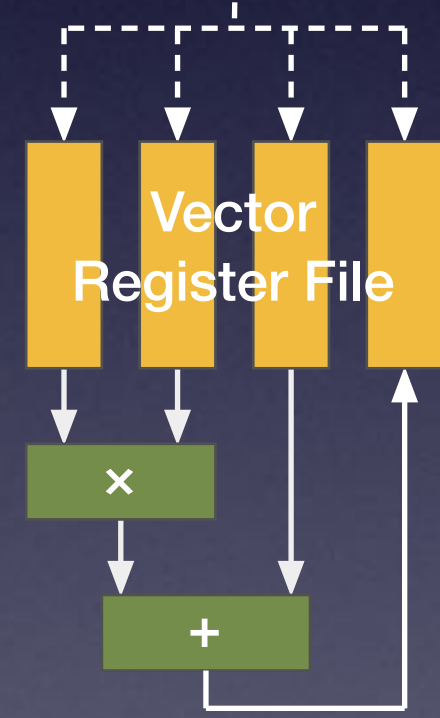


Best of both worlds



# Vector Accumulator Architecture

SRAM main memory

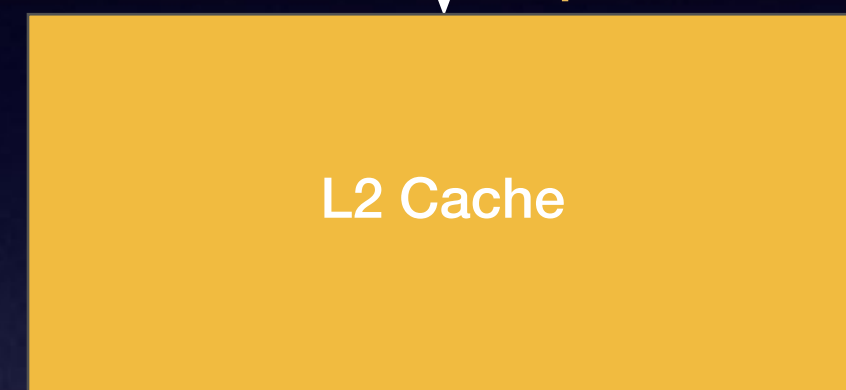


```

r1 = A[1:N]
r2 = B[1:N]
r3 = C[1:N]
r7 = r1 * r2 + r3
r4 = D[1:N]
r5 = E[1:N]
r8 = r4 * r5 + r7
r6 = F[1:N]
r9 = r5 * r6 + r8
G[1:N] = r9
    
```

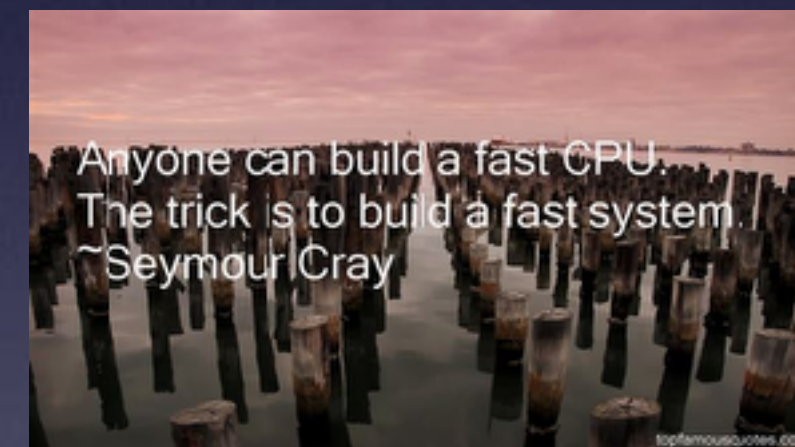
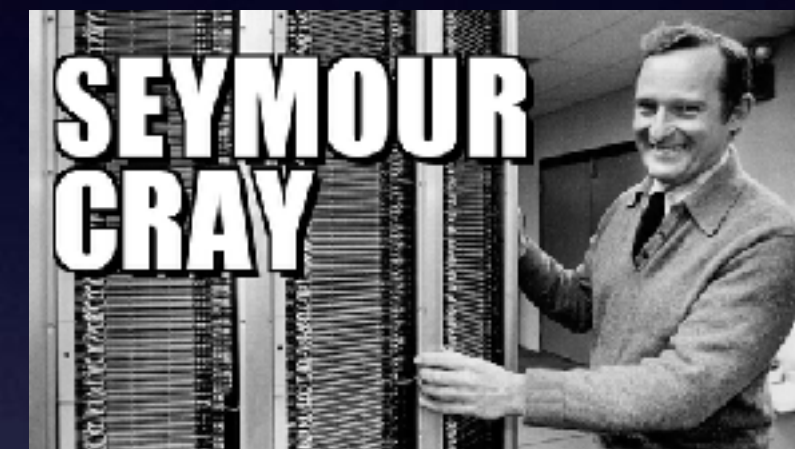
CRAY-1

DRAM main memory



$\Sigma = 4.52\text{pJ}$

GPU



THE FATHER OF SUPERCOMPUTING

"One of my guiding principles" observed Seymour Cray, "is, 'don't do anything that other people are doing.'"

DRAM main memory




```

v1 = A[1:N]
v2 = B[1:N]
vac = v1 * v2
v3 = C[1:N]
vac += v3
v4 = D[1:N]
v5 = E[1:N]
vac += v4 * v5
v6 = F[1:N]
vac += v5 * v6
G[1:N] = vac
    
```

$\Sigma = 1.82\text{pJ}$

C2P3



# Vector Accumulator Architecture

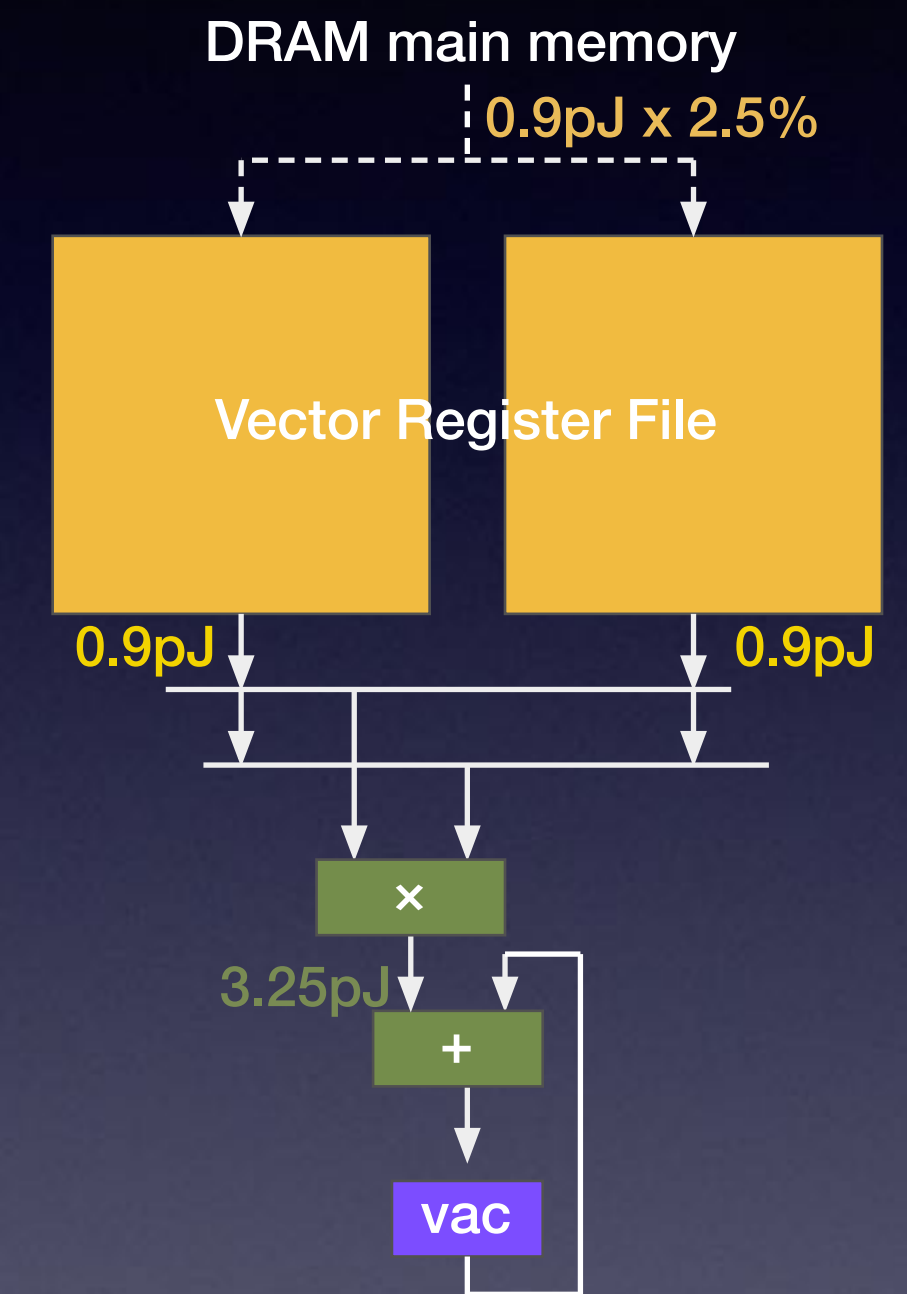

  
 an EPFL research center


Block floating-point DNN training

Mario Drumond  
 Tao Lin  
 Martin Jaggi  
 Babak Falsafi

MSR Contact: Eric Chung


 February 20<sup>th</sup>, 2018
 




  
 an EPFL research center

## Custom arithmetic for DNN

Prior work shows mixed results


- Half-precision floating-point (FP16):
  - 10x worse area/power than fixed-point
- Fixed-point:
  - Limited range
  - Complex techniques to select quantization points
  - Quantization points are static

Key observation:

- Large fraction of DNN computations appear in dot products

Custom arithmetic for dot products is enough

3


  
 an EPFL research center

## Block floating-point (BFP) for DNNs

Compromise between fixed- and floating-point

- Limits range of values within a single tensor
- Wide range of values across tensors
- Dynamically pick quantization points

✓ Dot products in fixed-point  
 ✗ Other operations degenerate to floating-point (FP)

Great candidate for custom DNN representation

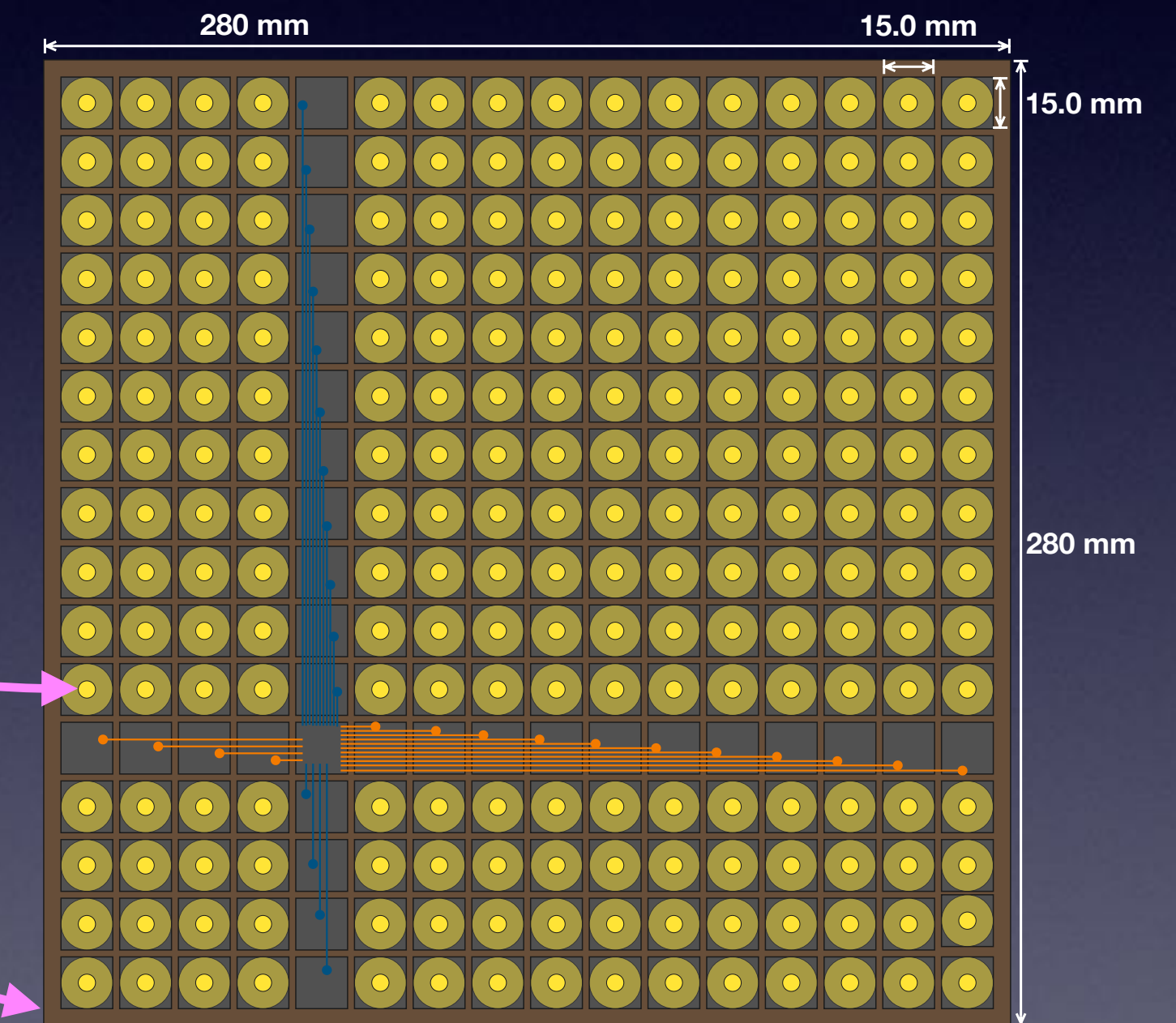
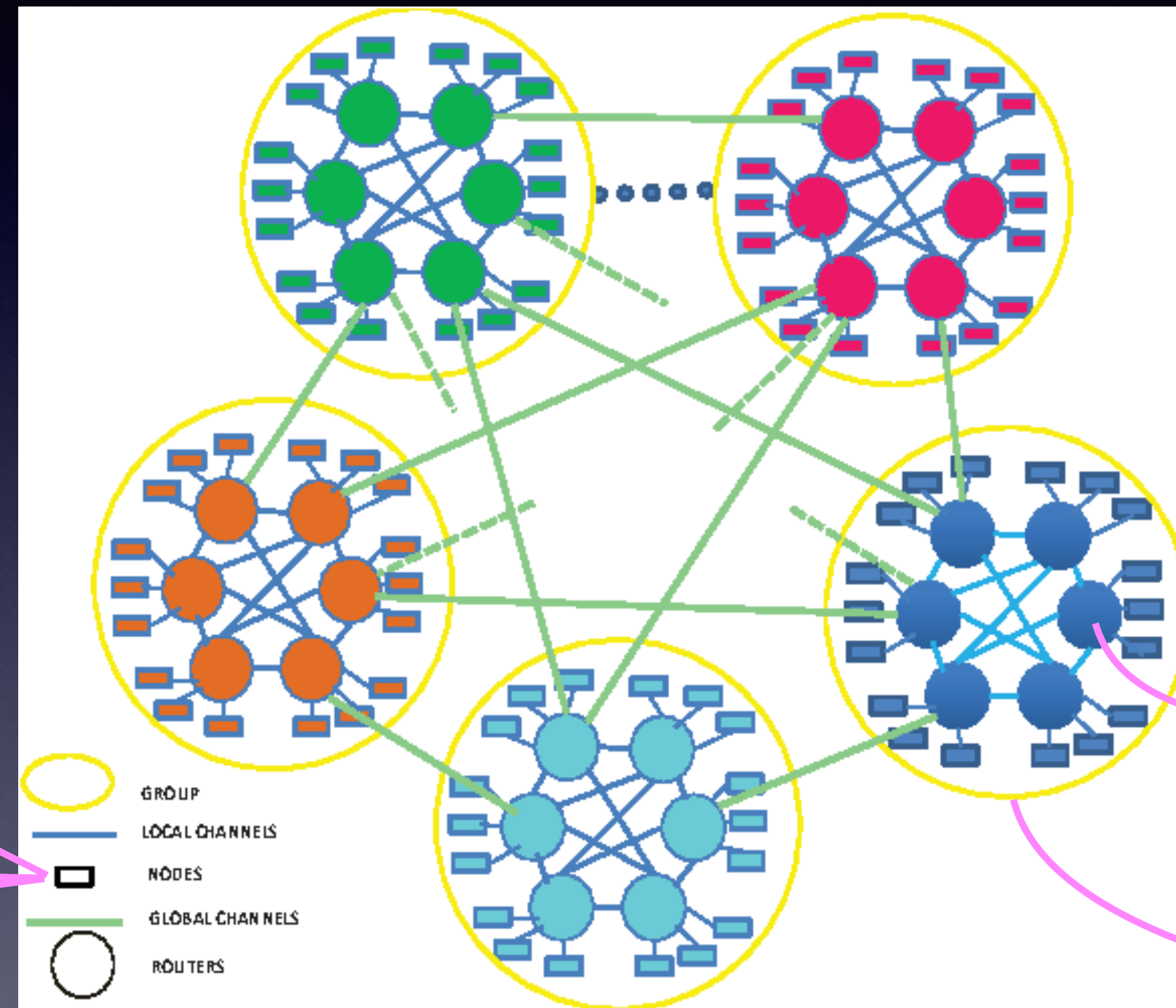
4

C2P3



# Dragonfly Network

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

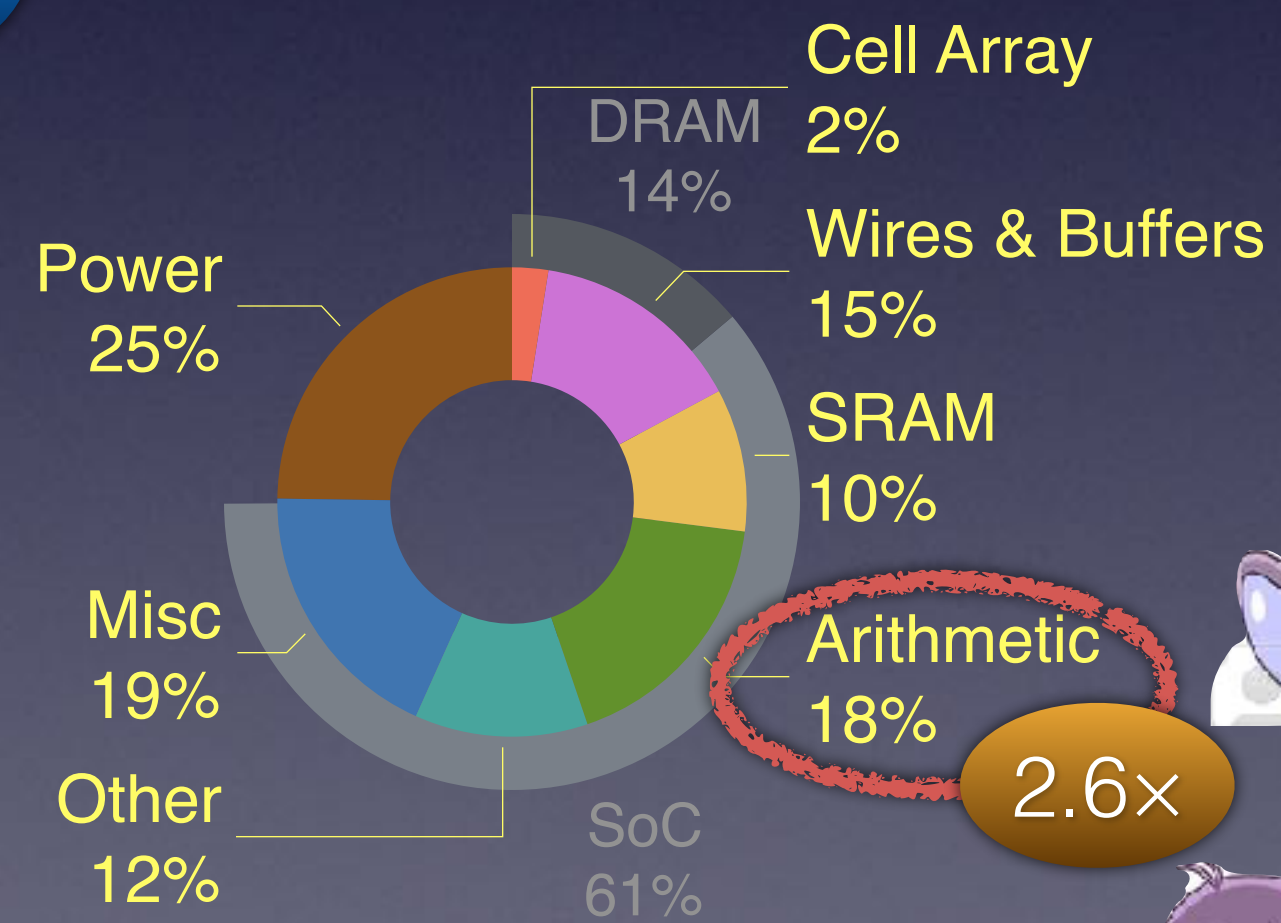
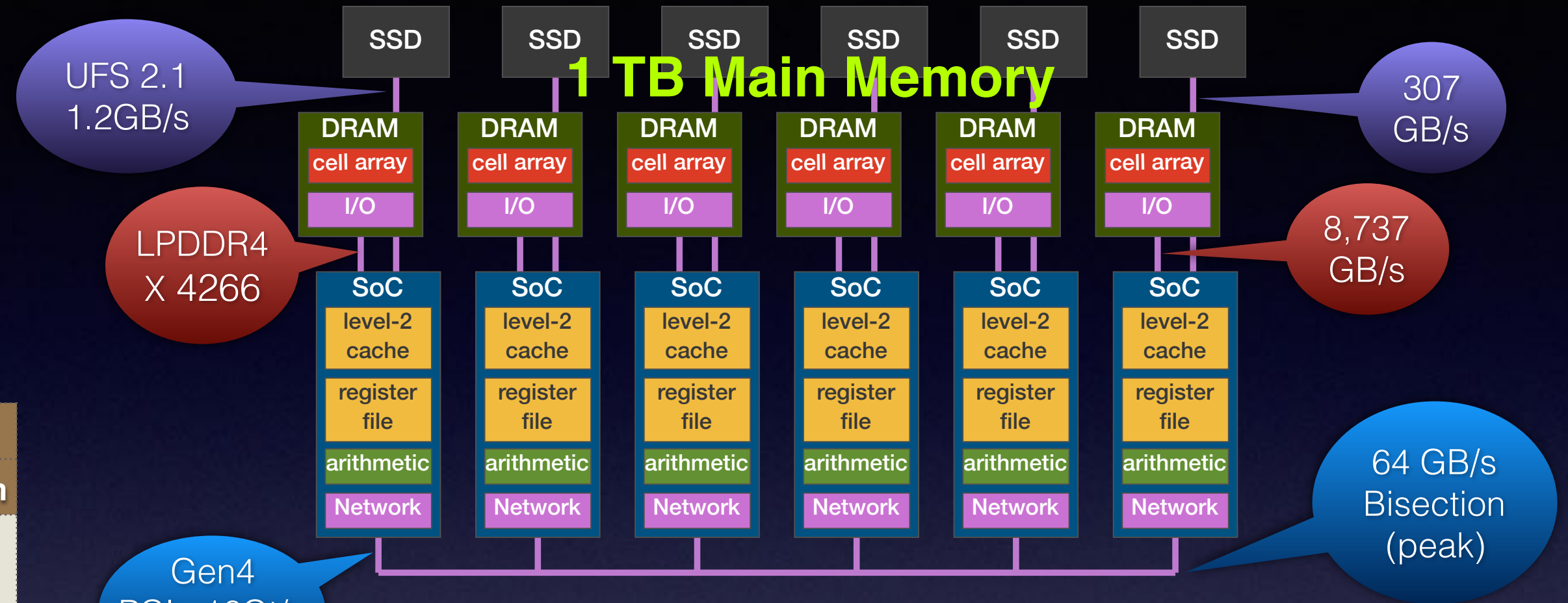


256×256=65,536 processor SoC chips

# 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		0.07		
	0.026	data transfer within a chip	1.93	3.3	5.3	0.25	0.4	106
	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		
	MADD instruction	write register file (64-bit)	1.8	0.05	0.05	0.00		
		read 2 operands (64-bit)	3.6	3.6	10.1	0.28		
	other	floating-point MADD (64-bit)	6.5	6.5		0.51	1.8	468
		memory interface, control, etc.	1.17	1.17		0.09		
		coherency directory, switch	3.19	3.19	4.4	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
<b>TOTAL</b>	<b>52.0</b>	<b>GFLOPS / Watt</b>	<b>3.4</b>	<b>33</b>	<b>32.9</b>	<b>3.00</b>	<b>767</b>	



2.6x

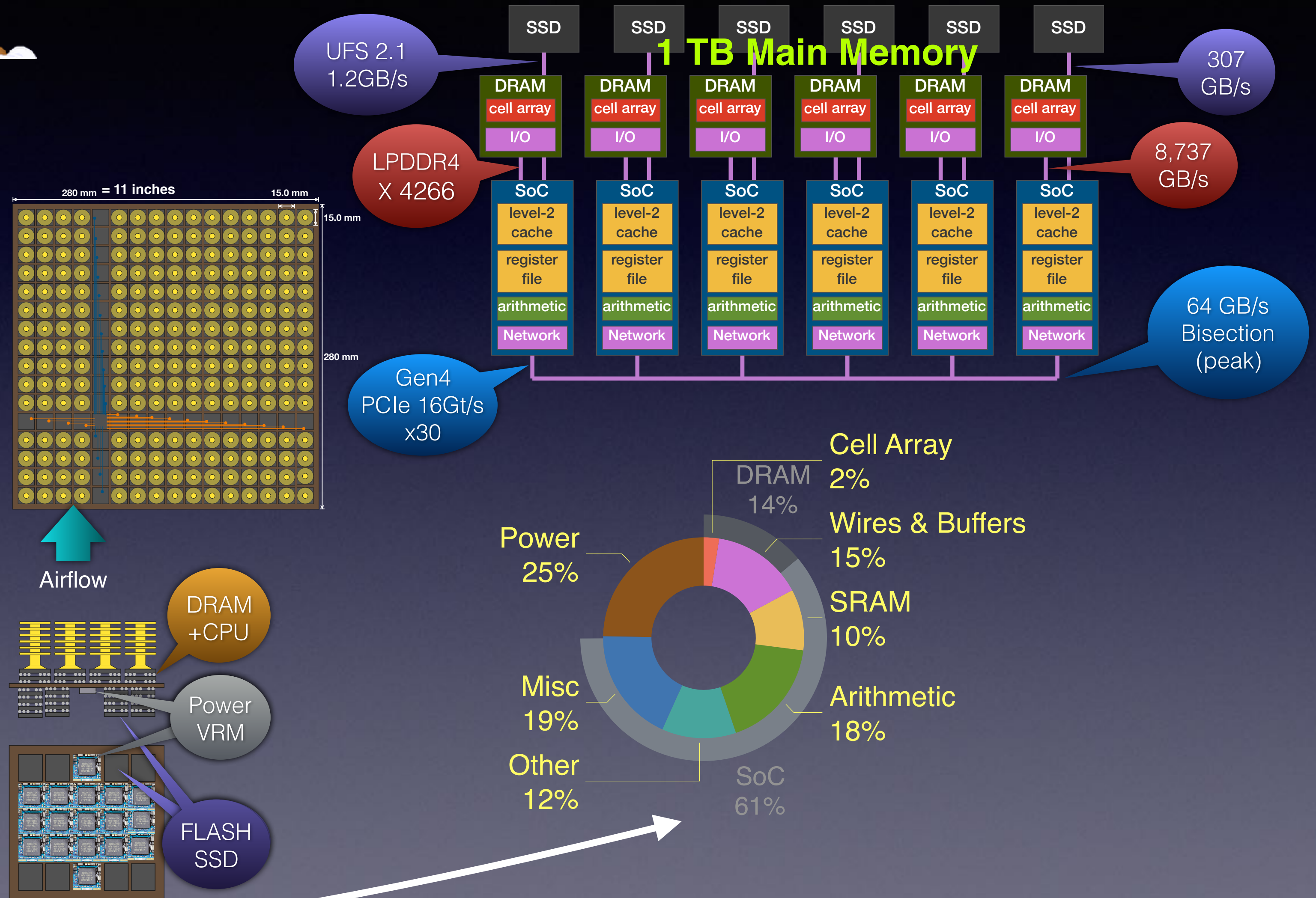


# Together



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

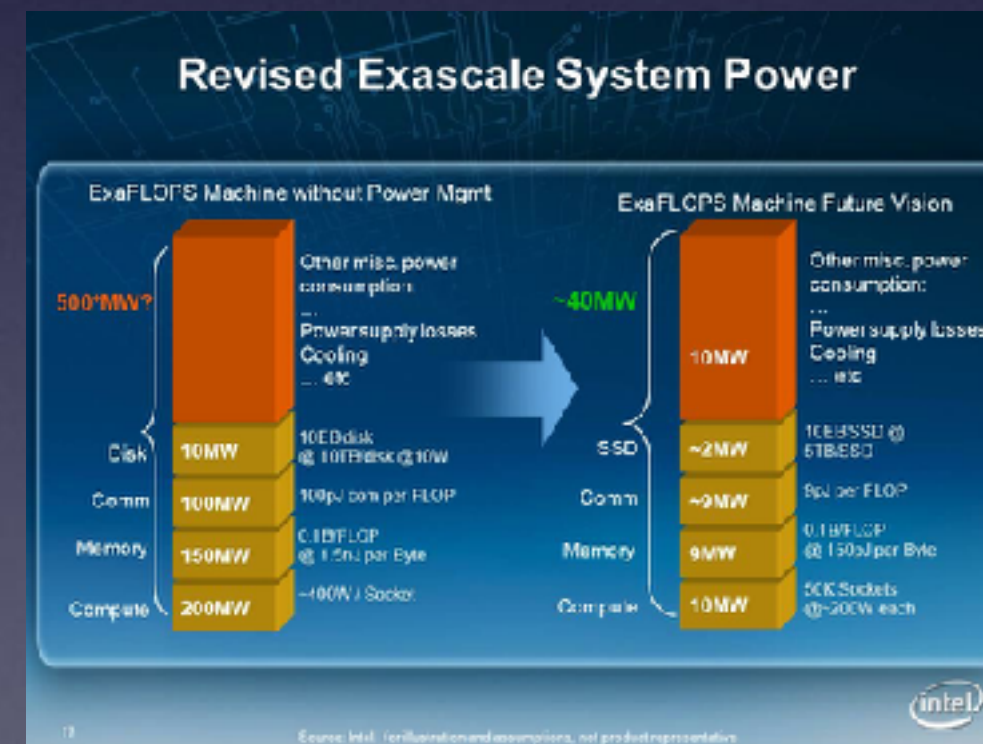
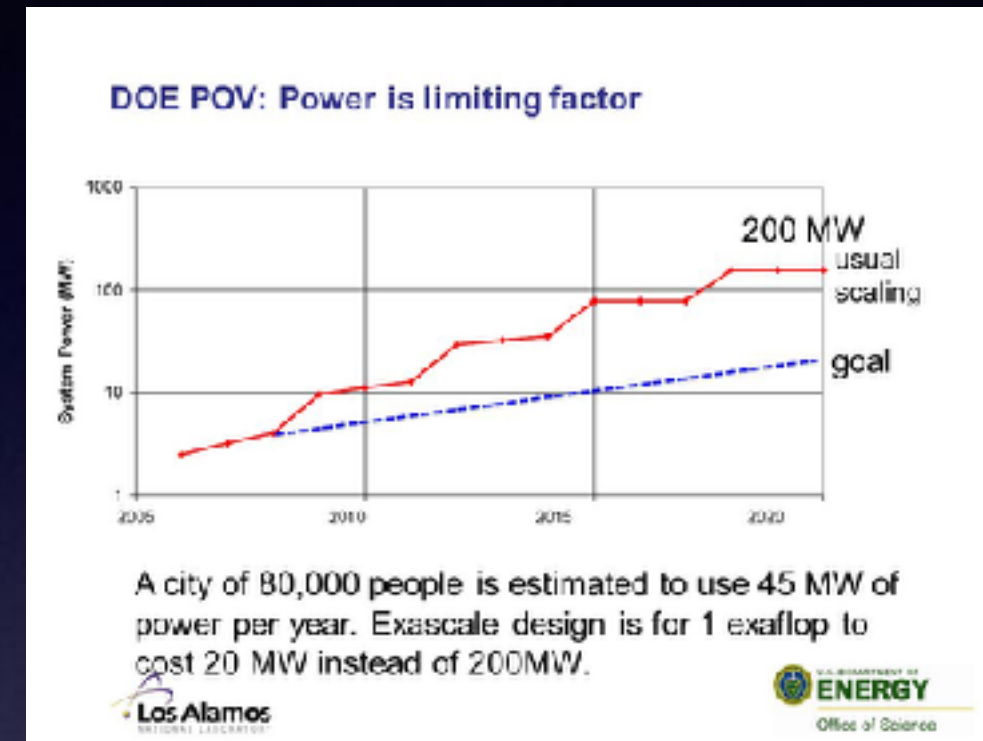
**10x** more energy efficient  
 i.e. 150 GFLOPS/W 64-bit Float  
 600 GFLOPS/W 16-bit Float



# Conclusion

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

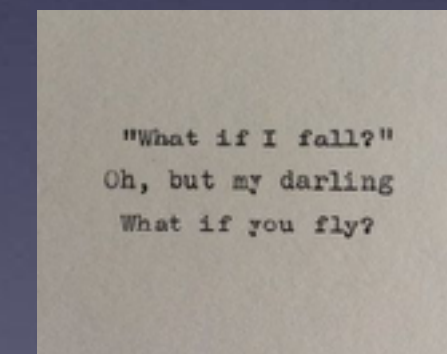
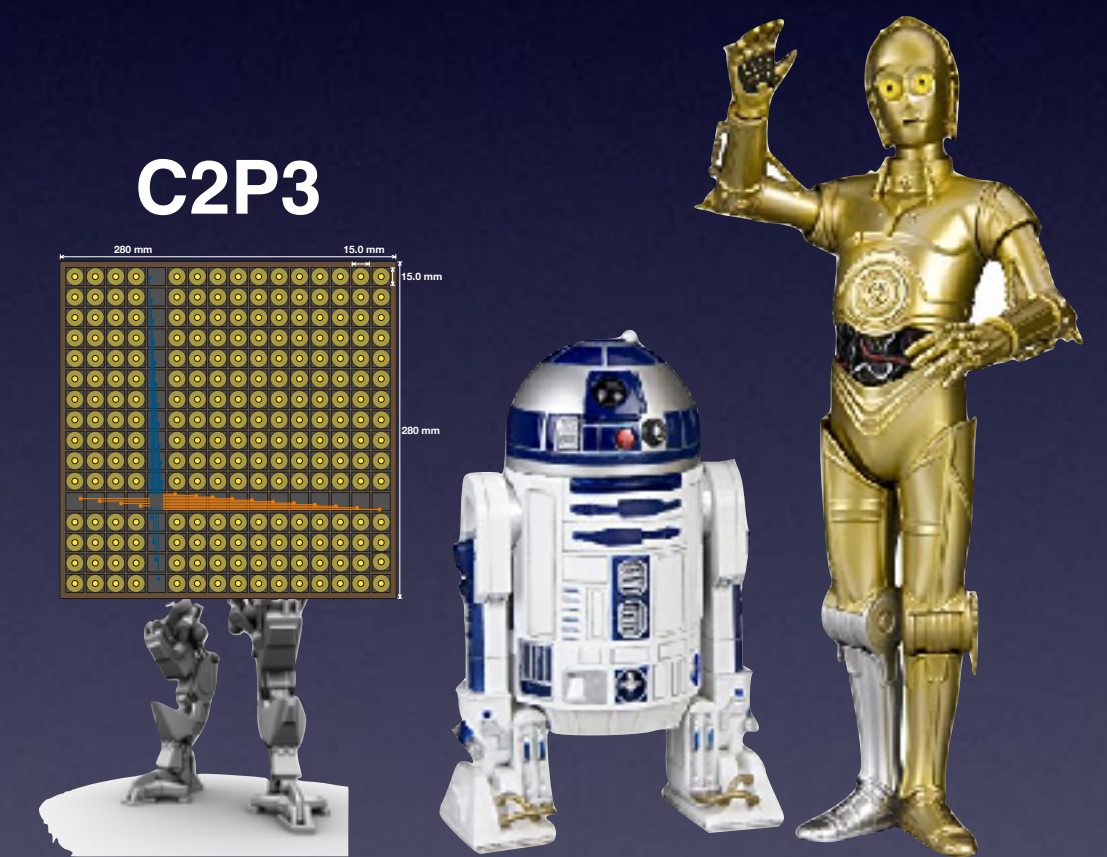
**10x** more energy efficient  
 i.e. 150 GFLOPS/W 64-bit Float  
 600 GFLOPS/W 16-bit Float



(1 ExaFLOPS / 80%)  
 ----- = 8.3 MW  
 150 GFLOPS/W



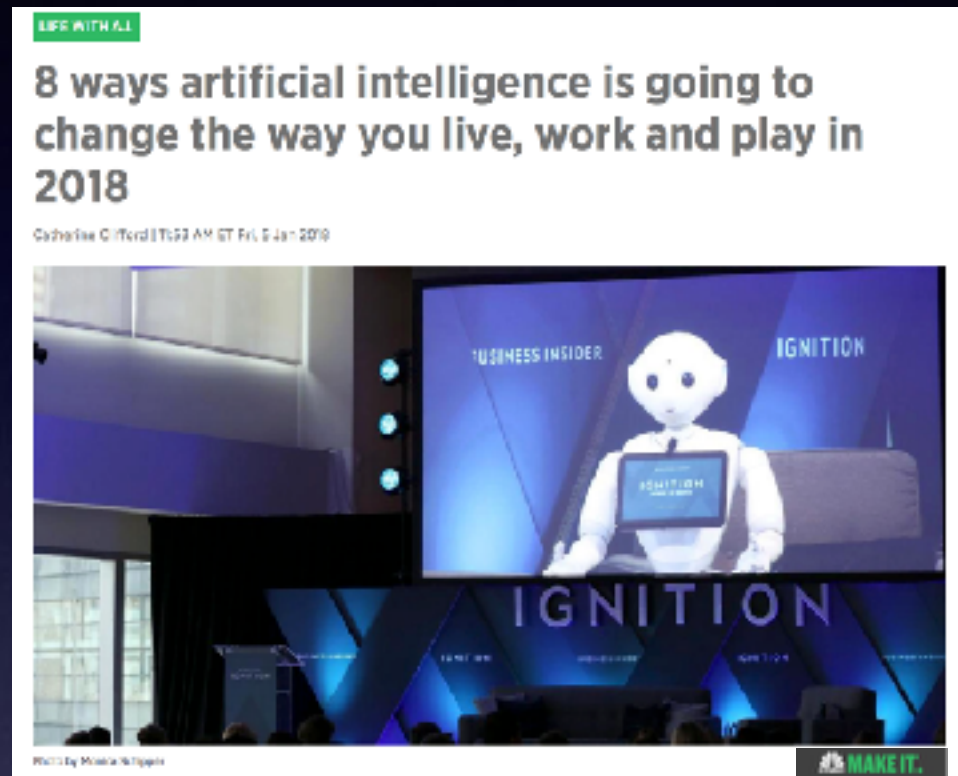
Cache Coherent Parallel Phone Processor



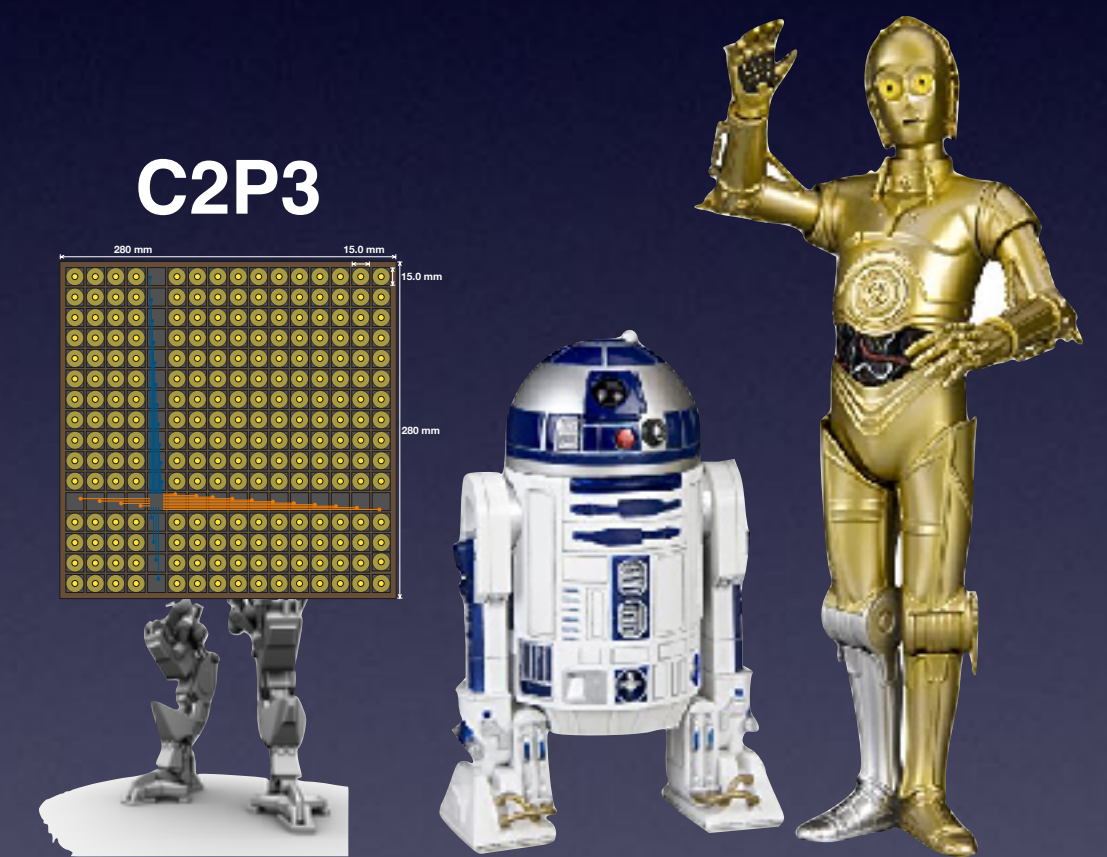
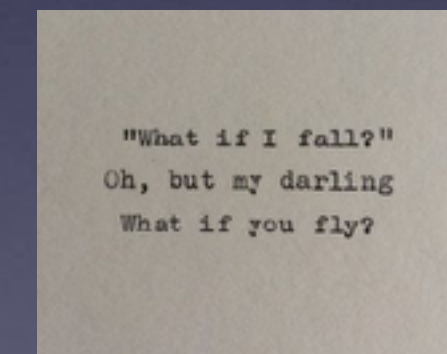
Apple's 'A12' chip reportedly in production using 7nm process from TSMC

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



**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.