System Architecture

The Future of Computing

David Kanter
dkanter@gmail.com
www.realworldtech.com
@TheKanter

IVR 2015

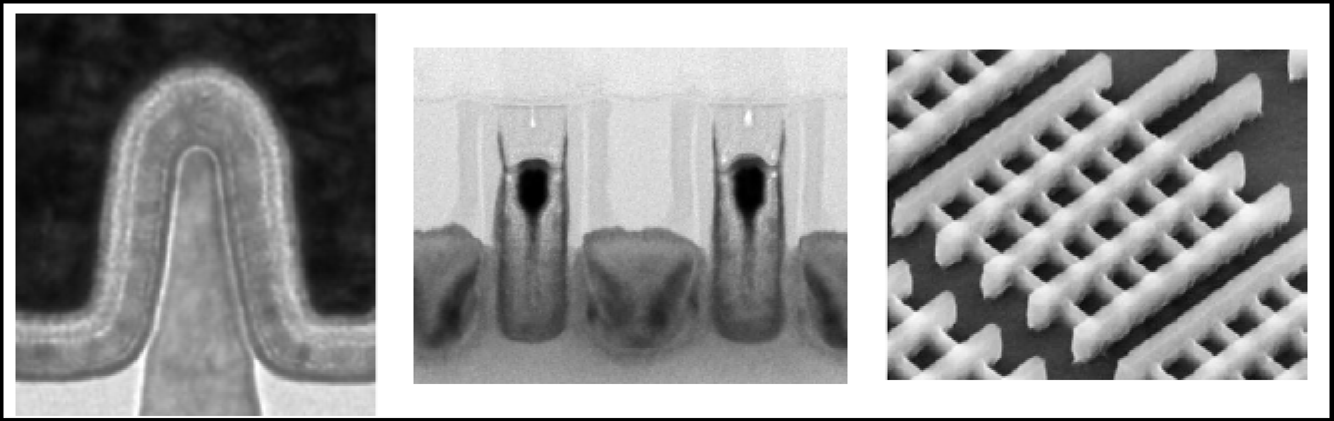
My Background

- Real World Tech
  - CSI/QPI; CPU, process, GPUs, etc.
  - Technology and IP consulting
  - Microprocessor Report
- Strandera
  - Speculative multi-threading for x86
- Several aerial imaging/camera patents

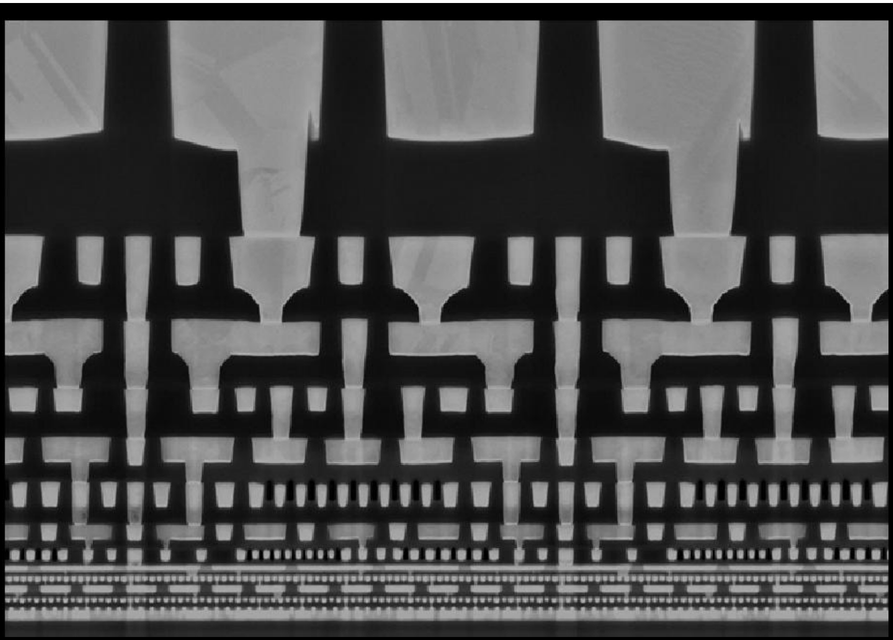
- **Garden of Eden**
- (Silicon) Paradise Lost
- Case Studies
- What Can You Do?

CMOS is Made of These

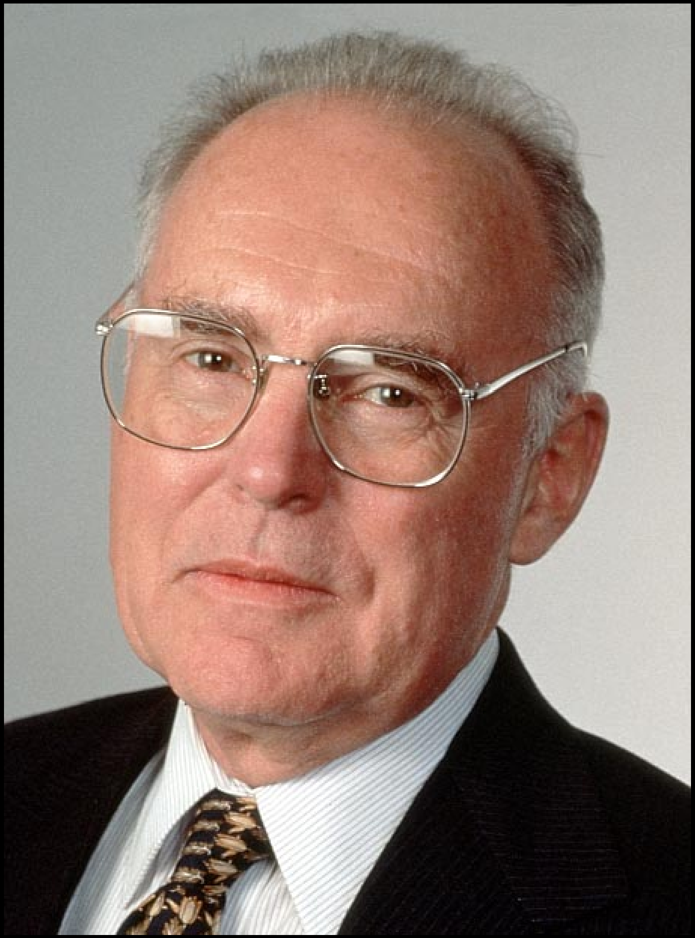
Transistors
(1 layer)

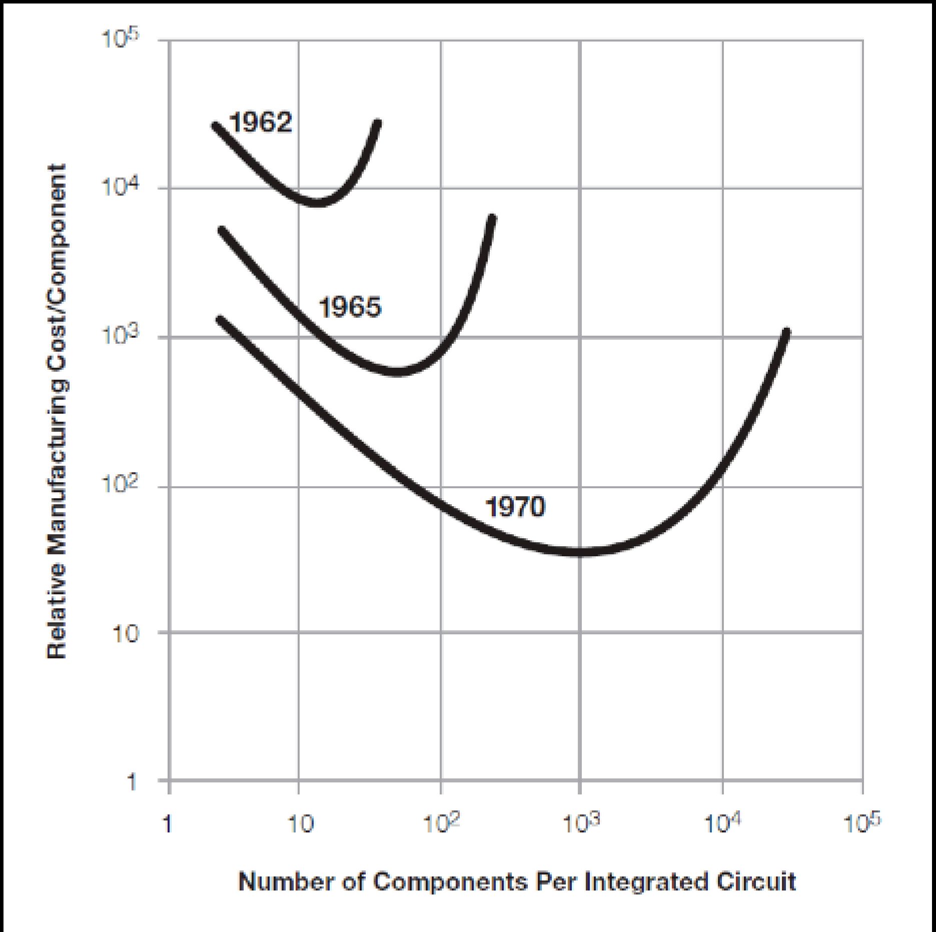


Metal interconnects
(e.g., 7-13 layers)



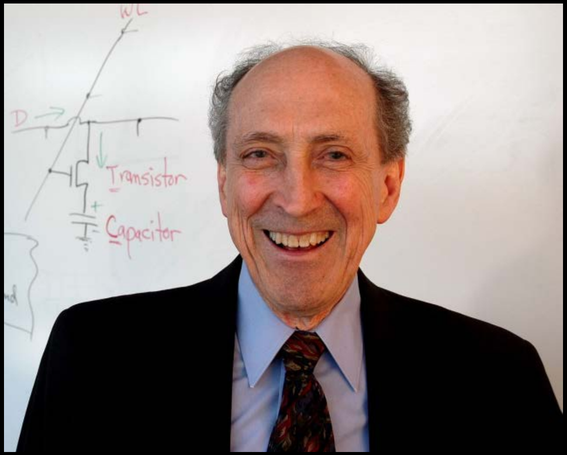
Moore's Law Scaling: Cost

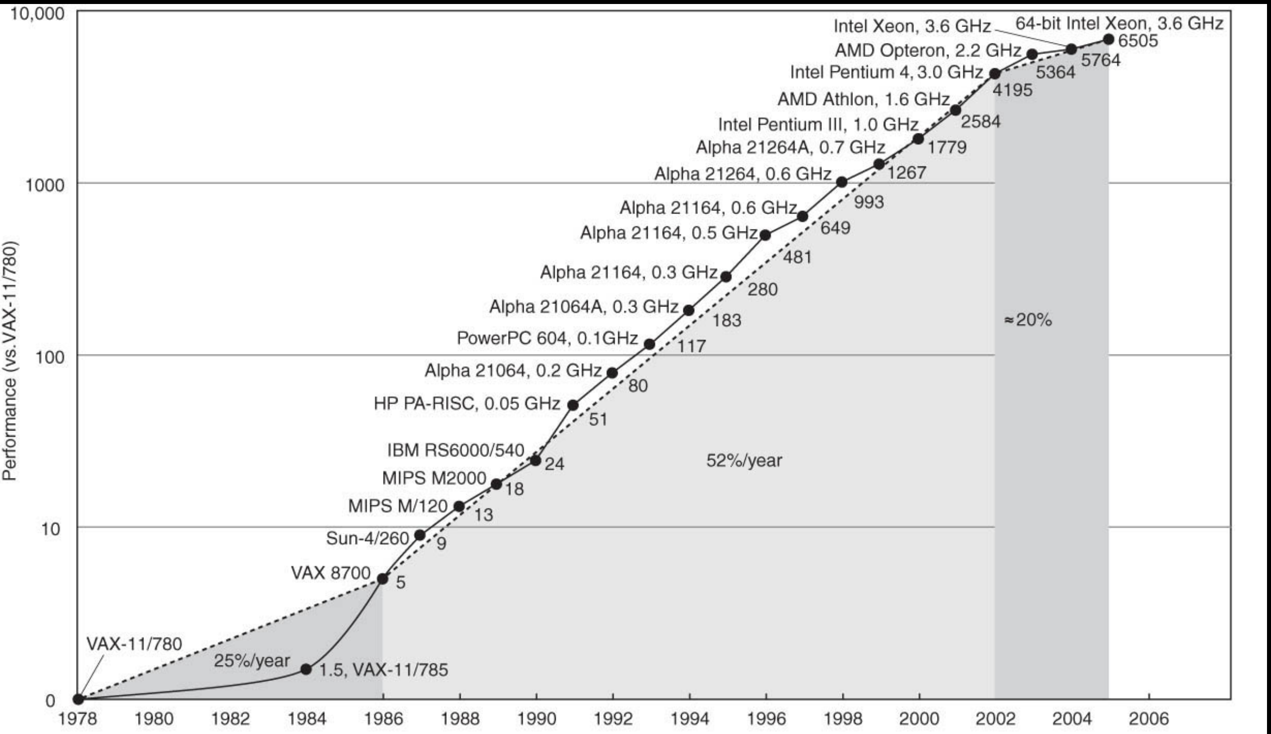


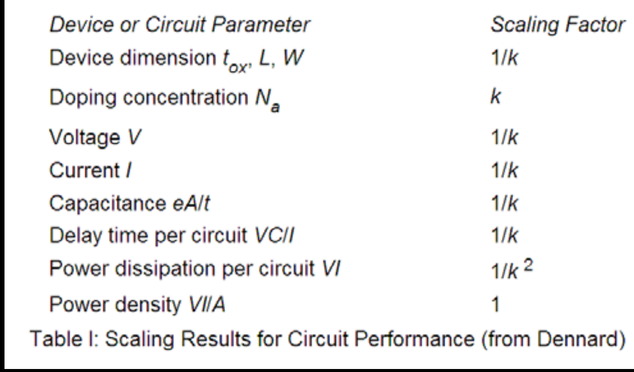


Relative Manufacturing Cost/Component
10^5
10^4
10^3
10^2
10
1
1962
1965
1970
1
10
10^2
10^3
10^4
10^5
Number of Components Per Integrated Circuit

Dennard Scaling: Performance







Device or Circuit Parameter	*Scaling Factor*
Device dimension t_{ox}, L, W	$1/k$
Doping concentration N_a	k
Voltage V	$1/k$
Current I	$1/k$
Capacitance eA/t	$1/k$
Delay time per circuit VC/I	$1/k$
Power dissipation per circuit VI	$1/k^2$
Power density VI/A	1

Table I: Scaling Results for Circuit Performance (from Dennard)

CPU Performance over time

Frequency increased by 3,500X

© Hennessey & Patterson, 2009

- Garden of Eden
- **(Silicon) Paradise Lost**
- Future of Computing
- Case Studies
- What Can You Do?

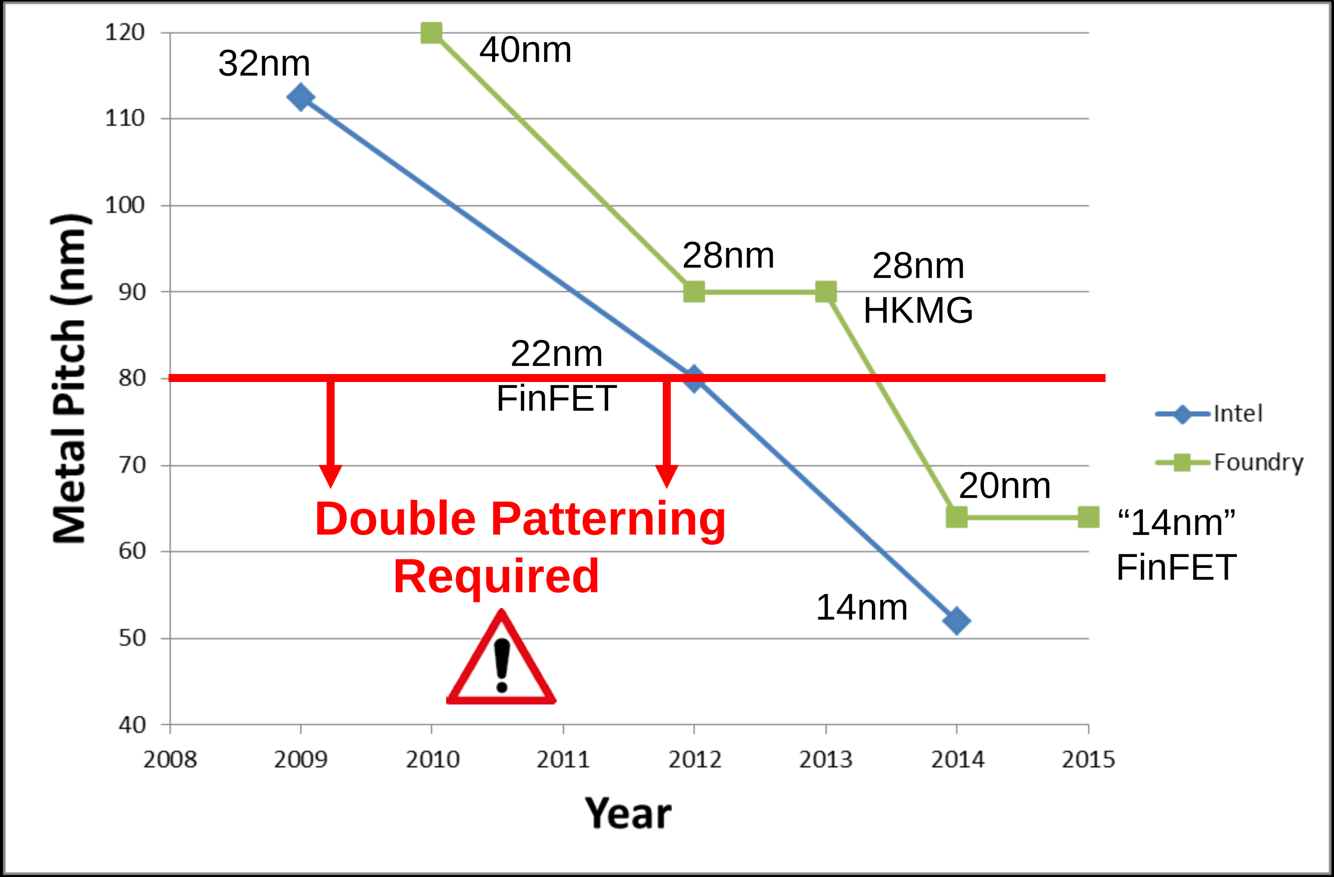
Physics is a Harsh Mistress

Geometric scaling is increasingly difficult

Dennard scaling stopped around 2000-2005

Silicon performance and density are decoupled

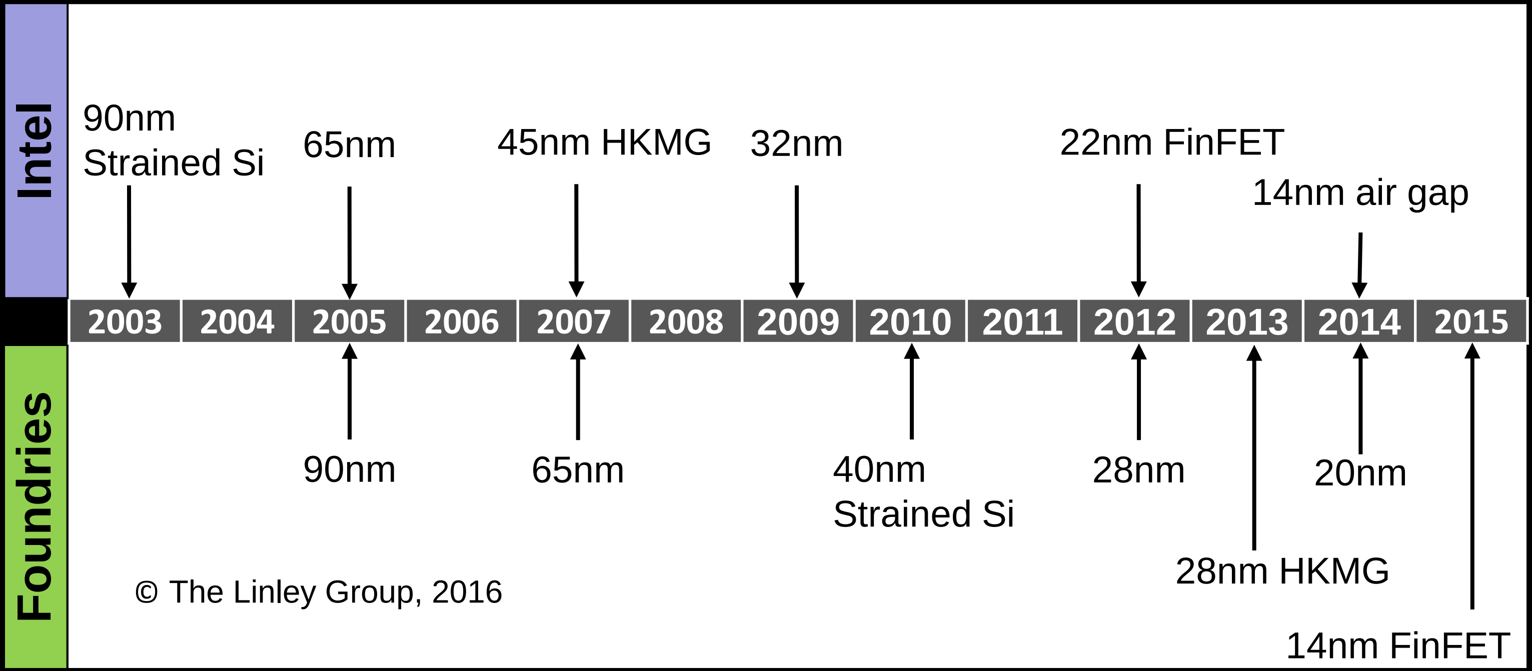
193nm Lithography Woes



Increased wafer cost erodes Moore’s Law

Prepared for EEC 277

A Decade of Ingenuity



© The Linley Group, 2016

- Material science improves performance
  - Strain, High-k/metal gate, FinFET, air gaps
  - Scaling continues, cost/complexity increases

No More Free Lunch

Reality: Cannot improve **everything**

- Multi-dimensional trade-offs
  - Performance, active power, idle power, area, time-to-market, NRE, yield
  - Analog & RF very different from digital

Must focus on value for **each product**

- Servers, ASICs, phones different

- Garden of Eden
- (Silicon) Paradise Lost
- **Future of Computing**
- Case Studies
- Where Next?

Systems: Better Together

Software — Hypervisor, OS, driver, libs, app, languages

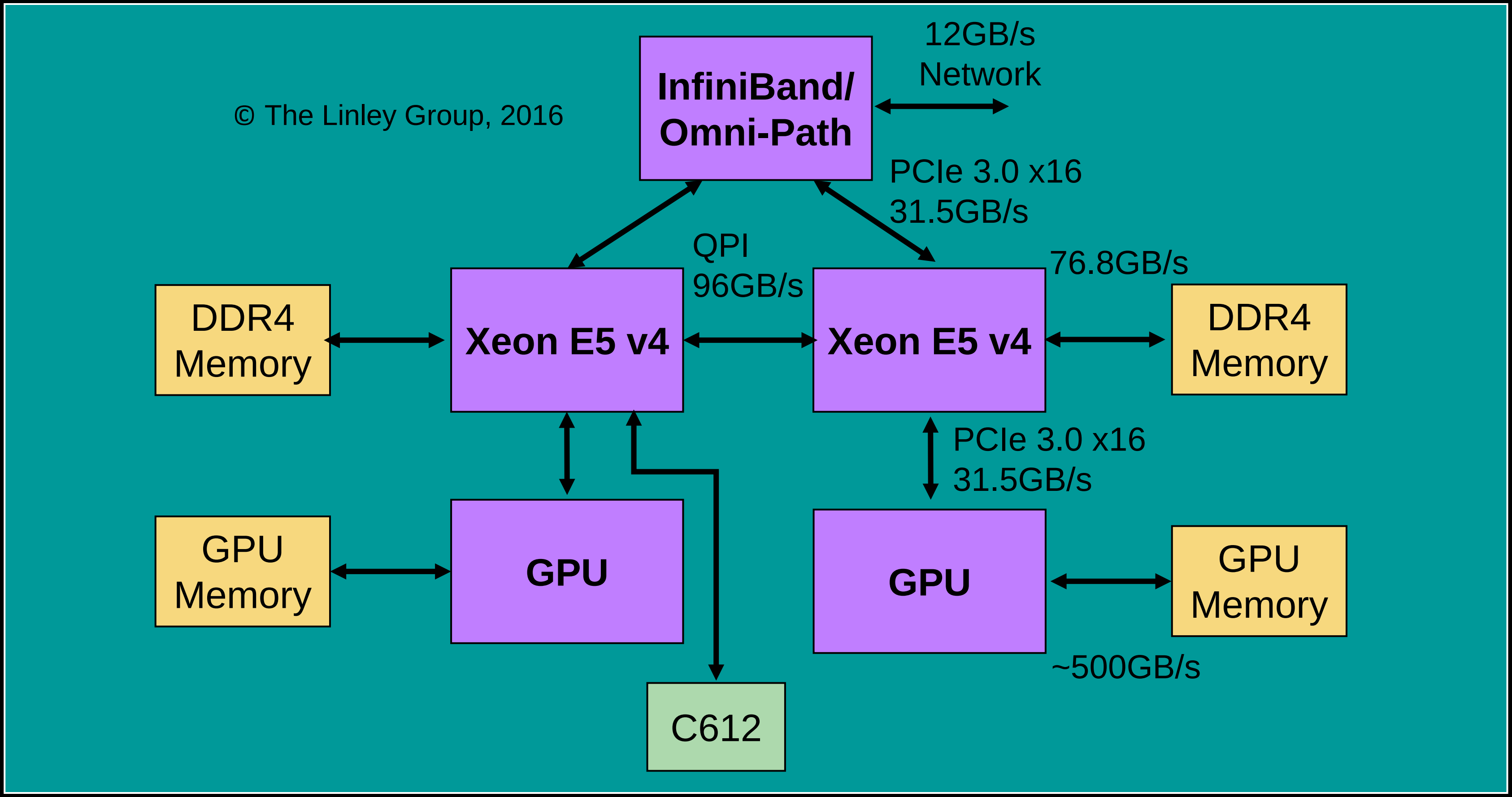
IP — CPU, cache, GPU, fabric, PCIe, memory, storage

Silicon — Logic, DRAM, NAND, analog, RF, interposers

Abstraction and isolation are costly

A unified view enables macro-optimization

Systems Drive Performance



Power Limits Performance

What Drives Power?

System Power = Chip + Cooling + Delivery

$$\overbrace{\text{Power}_{\text{DYN}} + \text{Power}_{\text{STATIC}}}$$

$$\overbrace{\text{Activity Factor} * \text{Cap.} * V_{dd}^{2} * \text{Frequency}}$$

$\text{Power}_{\text{STATIC}}$ = Complicated

0.15pJ/bit per mm in copper with 1V signal

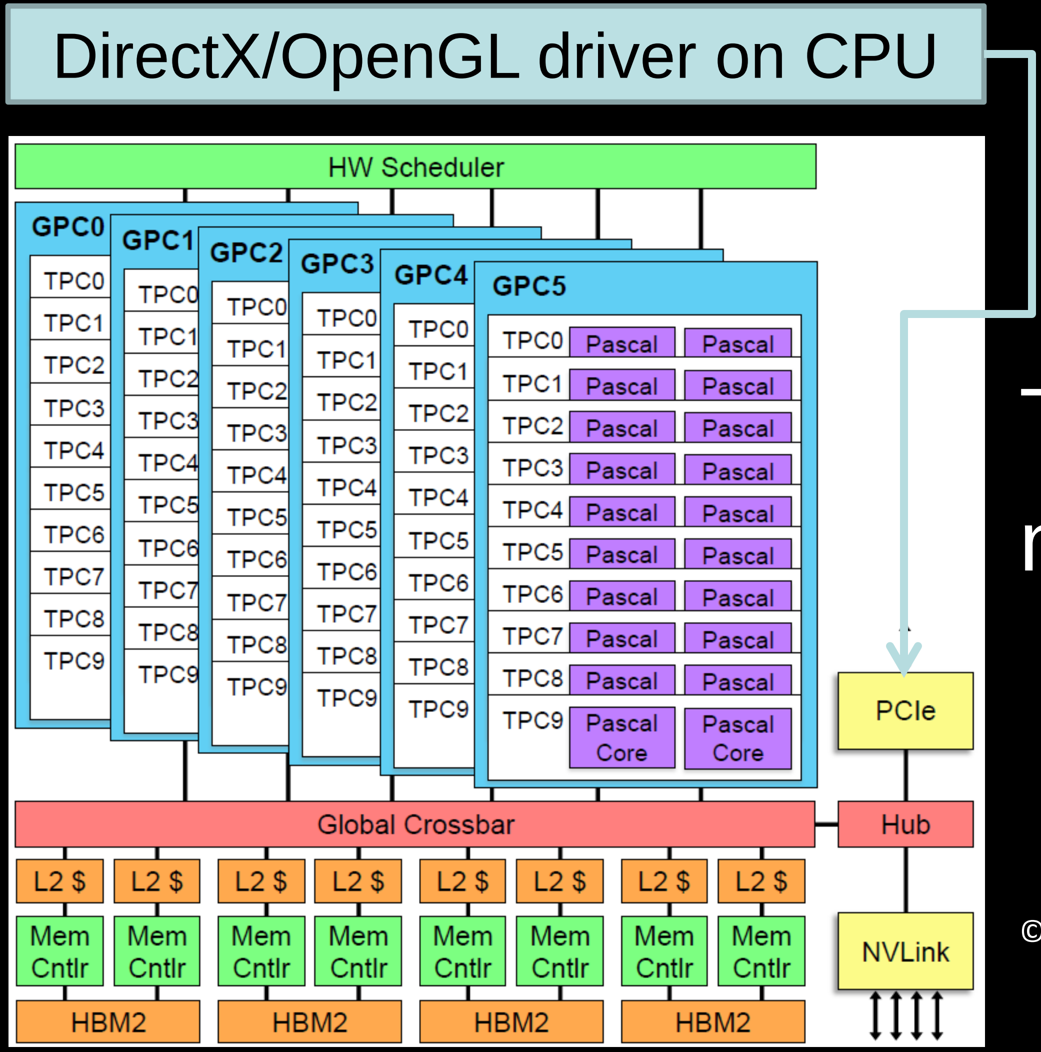
Creating Value

Performance is obvious, but also need:

- Security
- Reliability/robustness/resilience
- Programmability
- Form factor (temperature, volume, etc.)

- Garden of Eden
- (Silicon) Paradise Lost
- Future of Computing
- **Case Studies**
- What Can You Do?

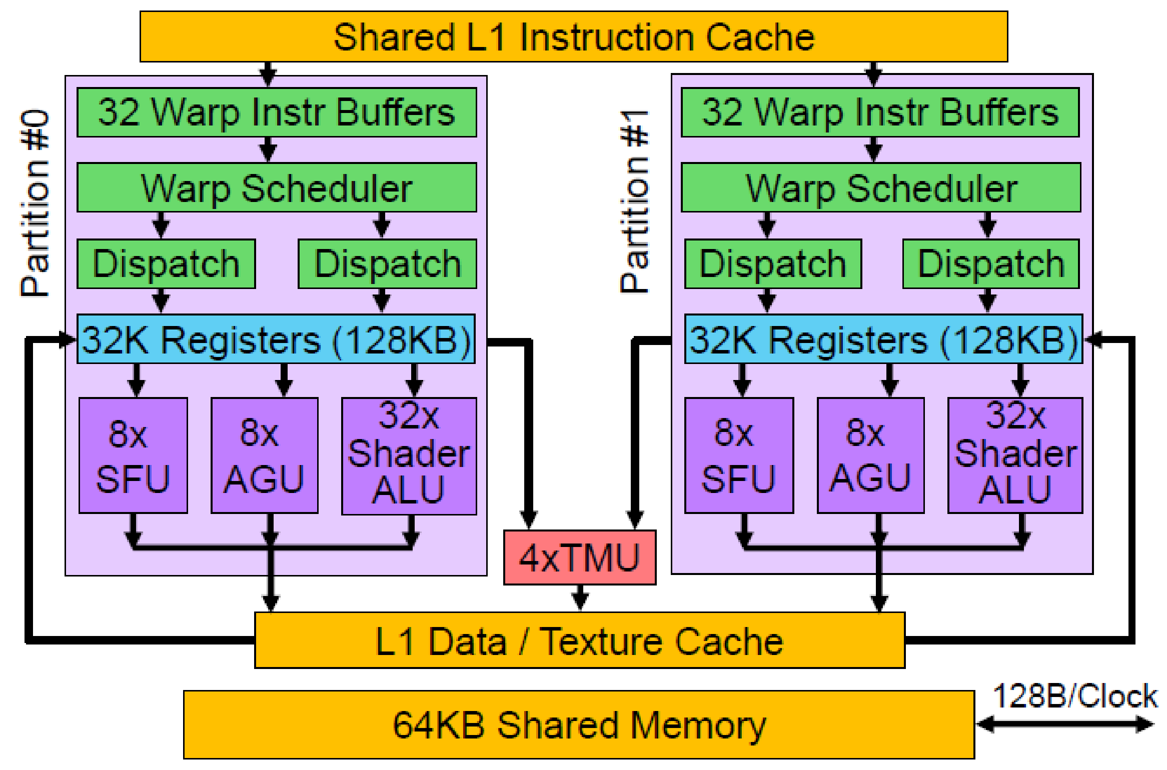
Modern Graphics Systems



Tiled rasterizer not shown

© The Linley Group, 2016

Modern GPU core



© The Linley Group, 2016

Customized for different products

Nvidia Pascal Cores

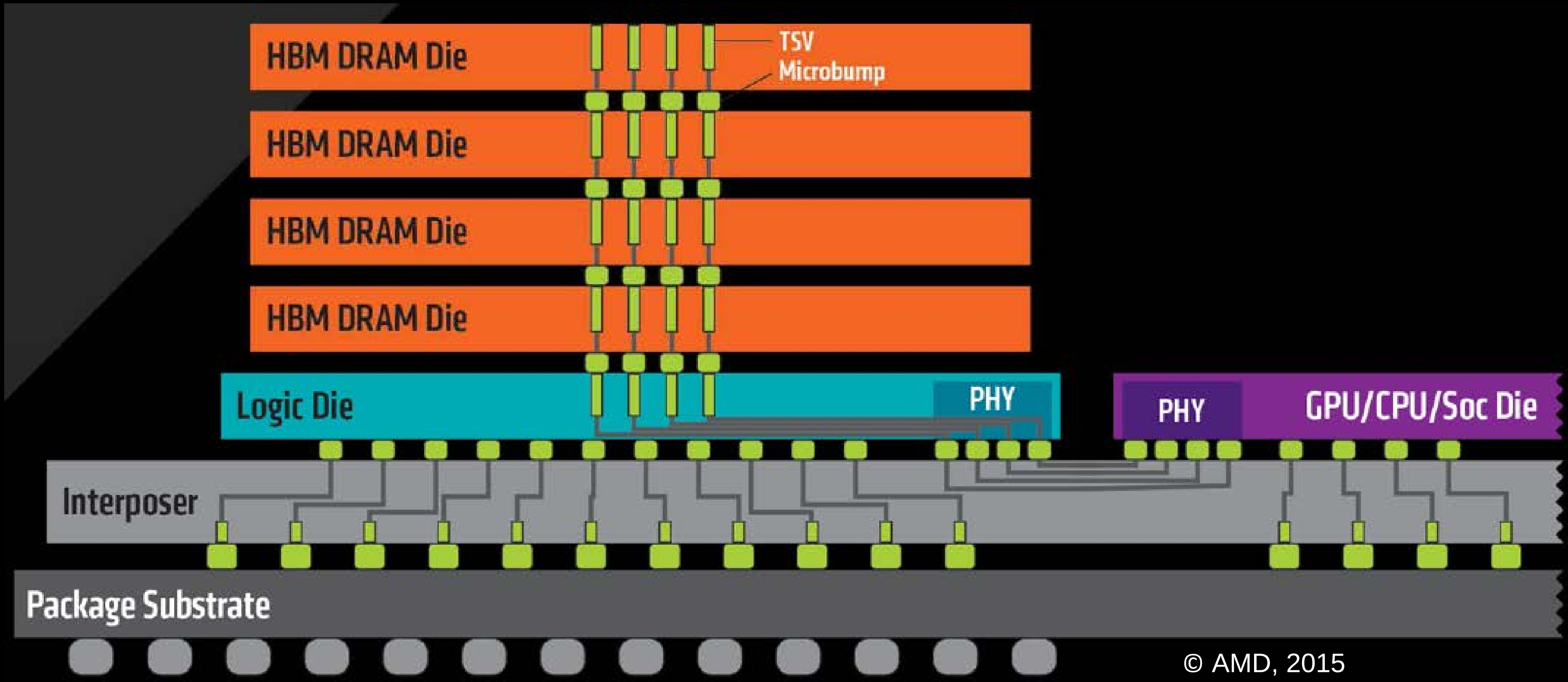
Instruction	**FMA, ADD, MUL**			**special functions**	**add, sub, boolean**	**compare, min/max, SAD, int32 shift, int32 bitfield**	**warp shuffle**	**64b convert**	**other convert, popcount**	**int32 mul/muladd, int24 mul, LZCOUNT, SIMD video**
Datatype	**HP**	**SP**	**DP**	**SP**	**int32**					
P100	128	64	32	16	64	32	32	16	16	Multiple instructions
GP10x	2	128	4	32	128	64	32	4	32	Multiple instructions

Throughput changes by product

L1 cache, shared memory also change

Some instructions are very slow

HBM instead of GDDR



© AMD, 2015

Transmit data over ~mm of silicon...

instead of ~cm of package/circuit board

HBM Advantages

- Wide and slow interface
  - Very simple PHY, no training, easy to idle
  - Lower voltage
- No ESD for on-package interconnect

	R 290X GDDR5	R Fury HBM
Bus width	16x 32-bit	4x 1,024-bit
Clock/data rates	1.25GHz/5Gbps	0.5GHz/1Gbps
Bandwidth	320GB/s	512GB/s
Voltage	1.5V	1.3V
Power	30W	12W

GPUs as Accelerators

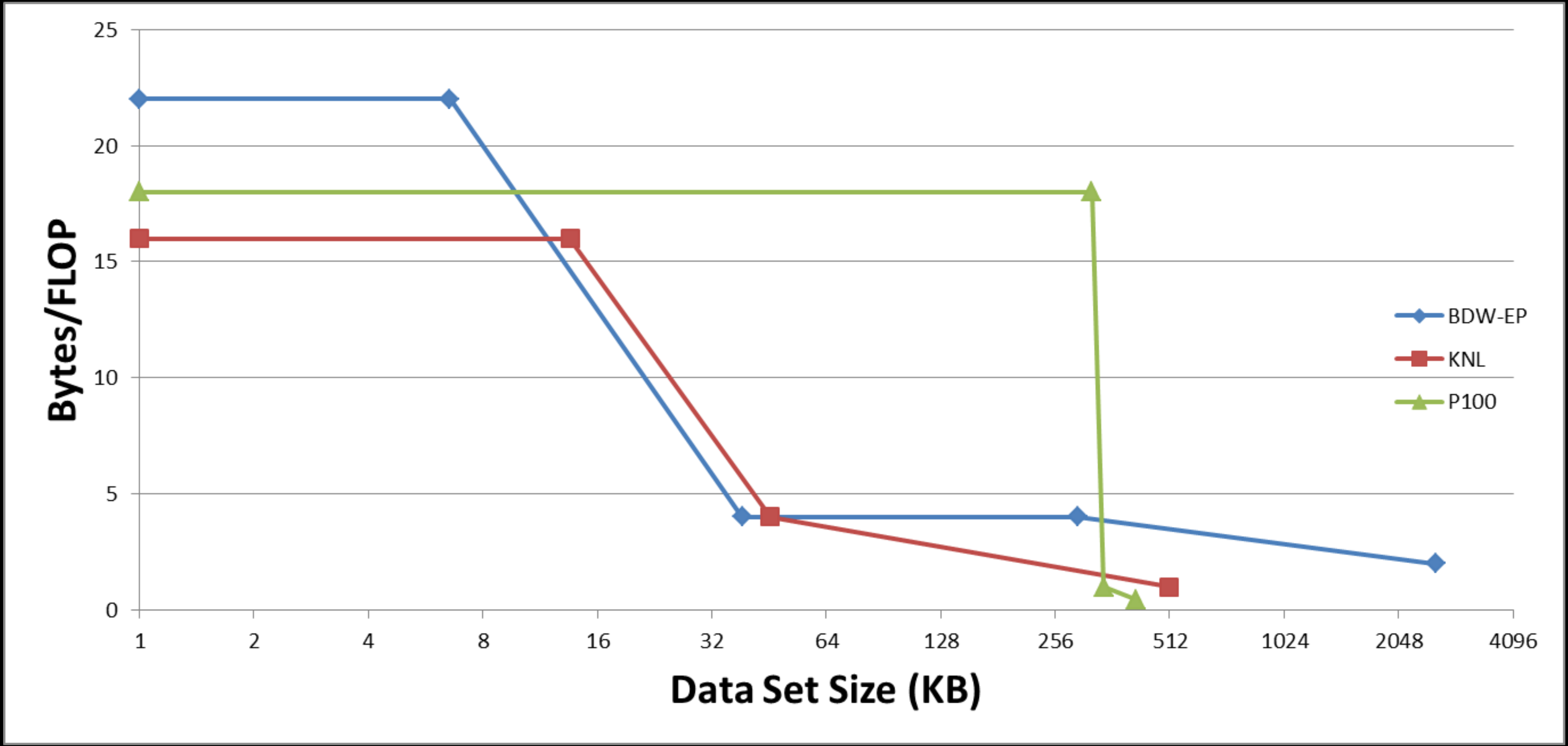
Leverage volume/economics of graphics

- Great for FP matrix multiplication
  - E.g., some HPC, neural network training, etc.

- Tolerable for image and video (mostly ints)
  - But eclipsed by dedicated hardware

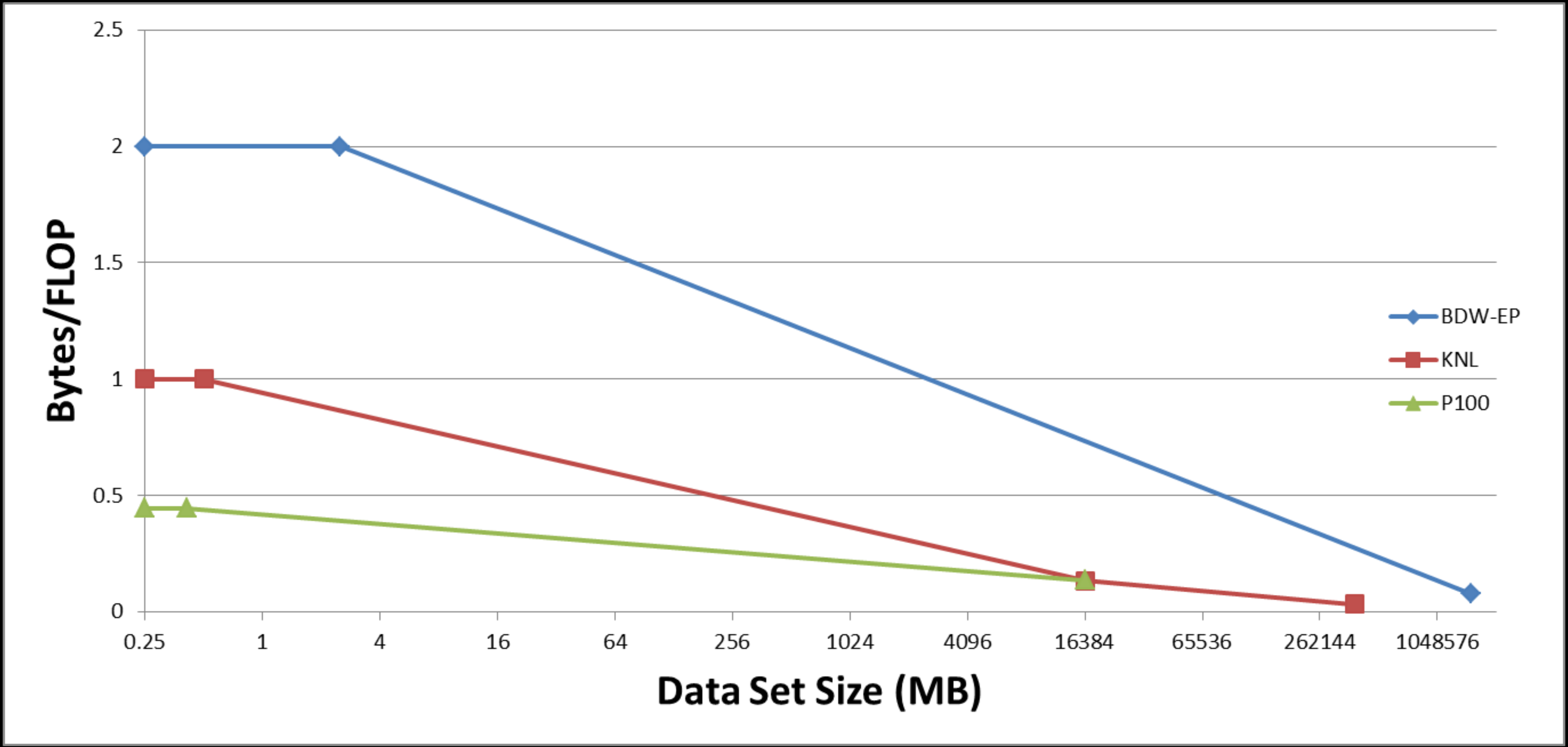
Avoid control flow, pick right problem size

On-Chip Data: CPUs and GPU



GPU registers are impressive, mind the cliff!

Off-Chip Data: CPUs and GPU



GPUs have small but fast memory

Prepared for EEC 277

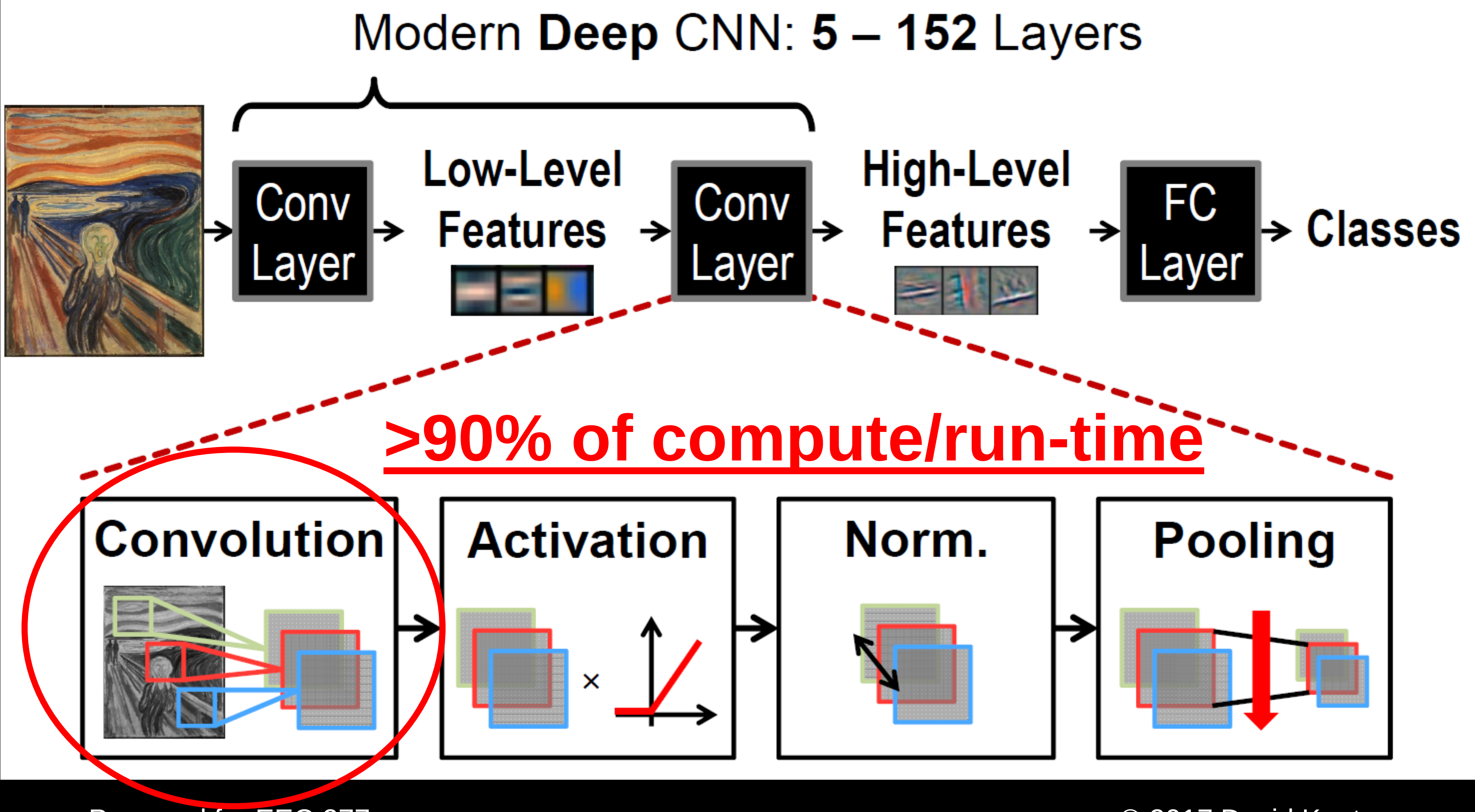
Workload Specific Accelerators

Bespoke hardware for the largest gains

Example: Eyeriss project at MIT

- Reconfigurable convolutional neural network accelerator
- Chen, et. al. from ISSCC 2016

Convolutional Neural Networks



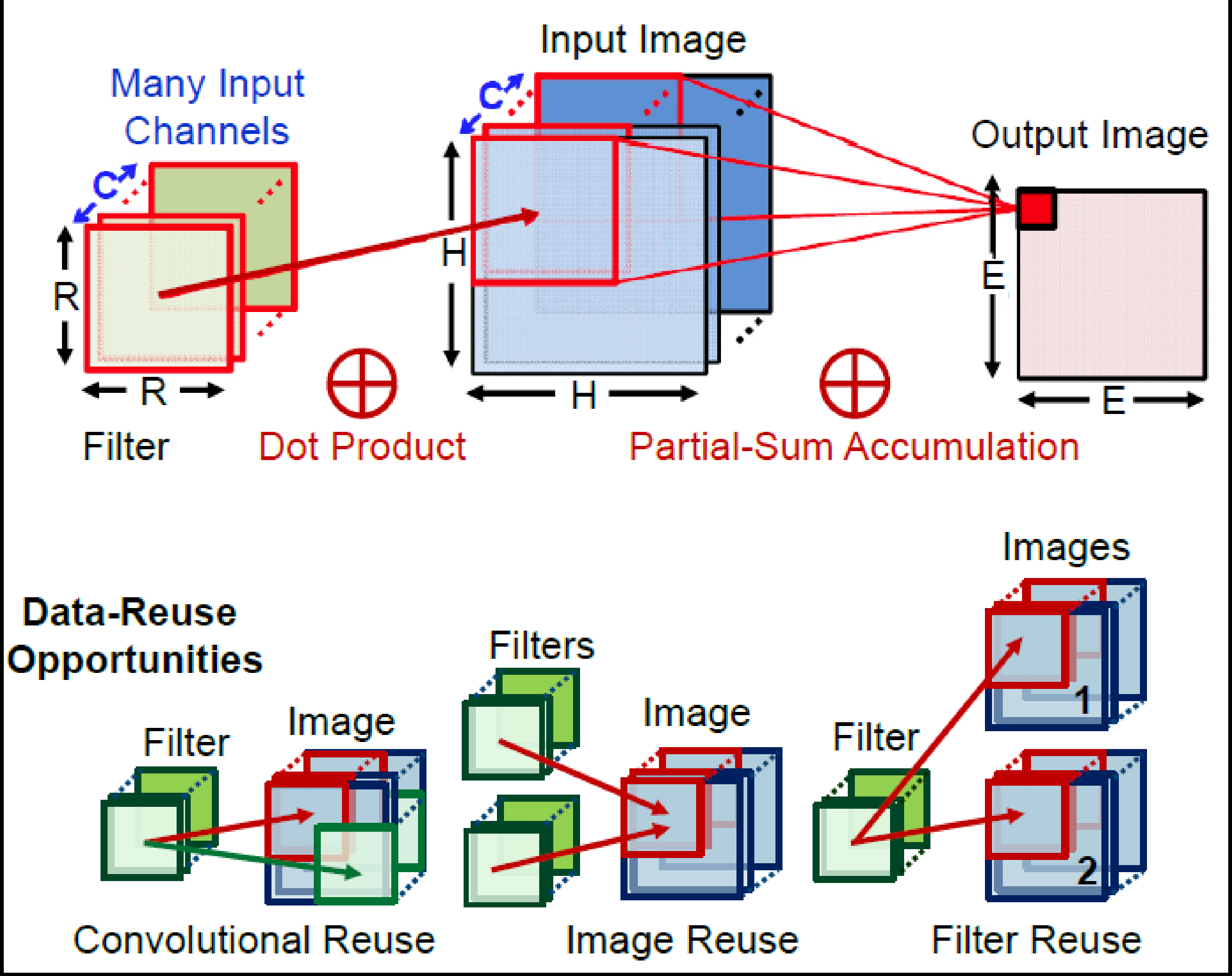
Modern **Deep** CNN: **5 – 152** Layers

Conv Layer → Low-Level Features → Conv Layer → High-Level Features → FC Layer → Classes

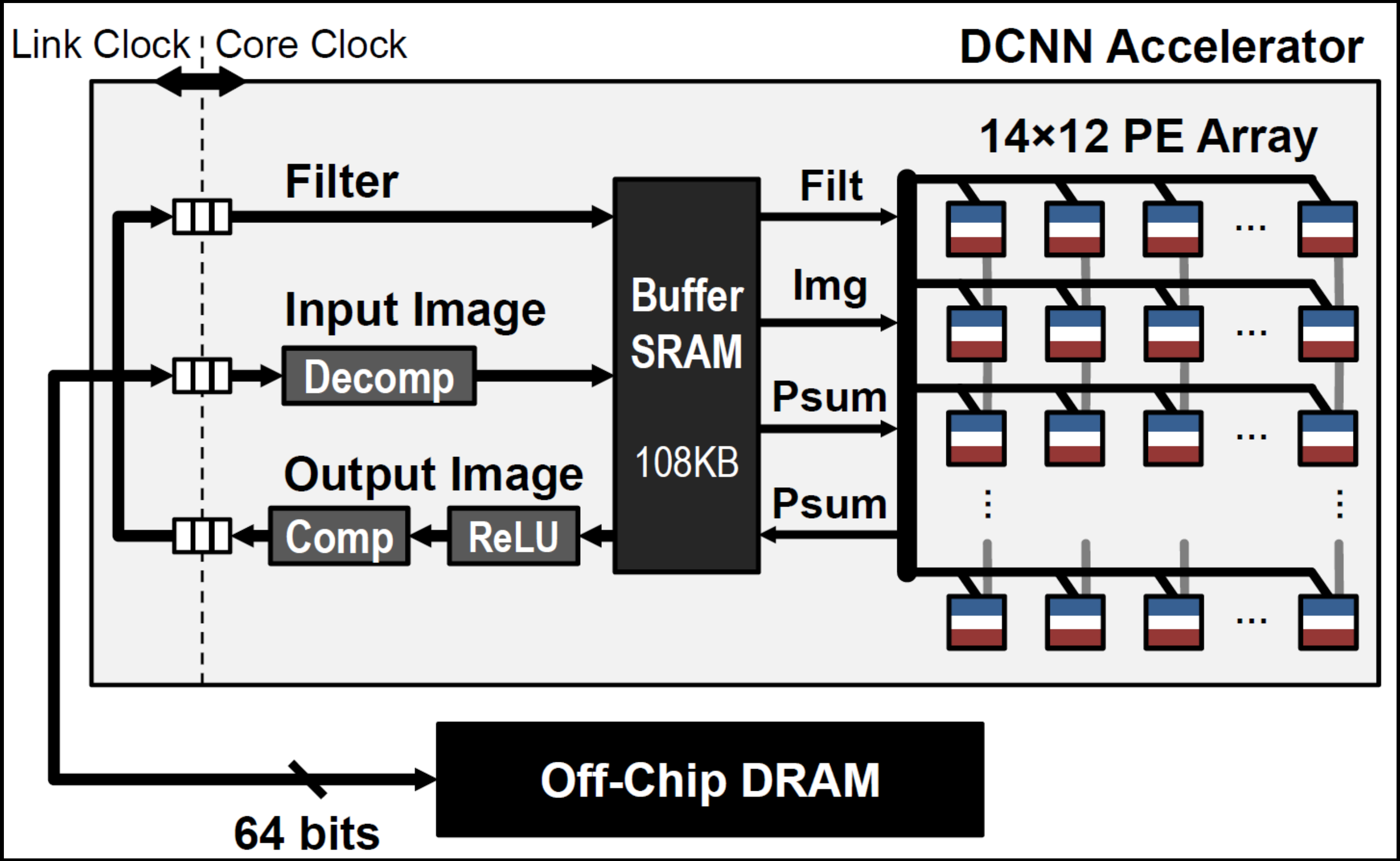
>90% of compute/run-time

Convolution → **Activation** → **Norm.** → **Pooling**

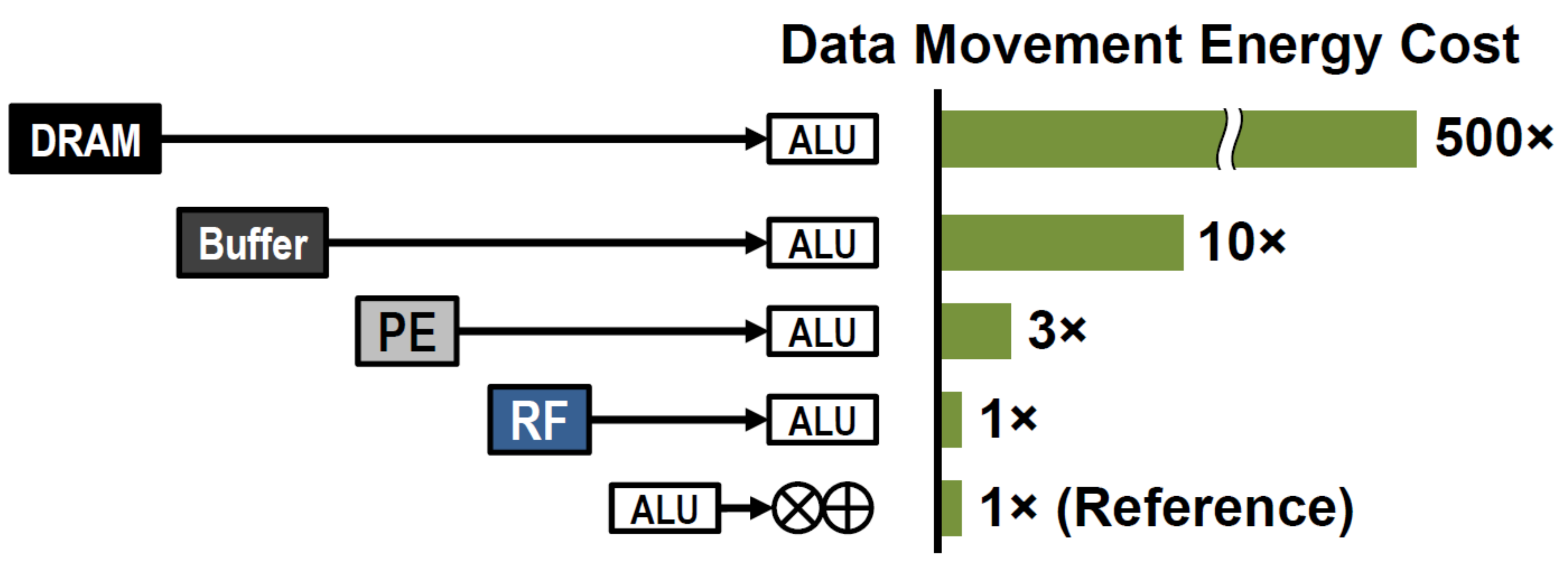
Convolution Dataflow



CNN Inference Accelerator



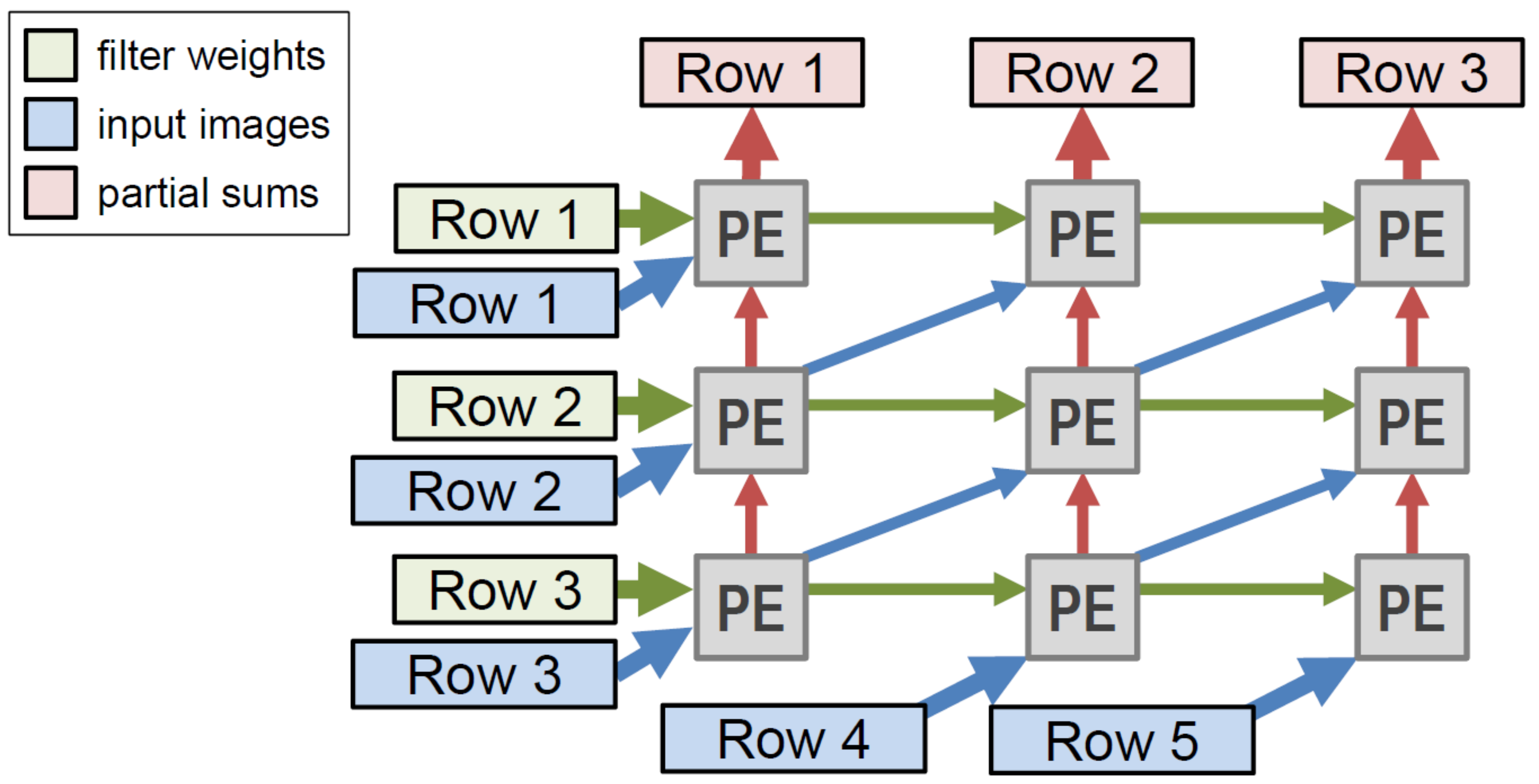
Minimize Data Movement



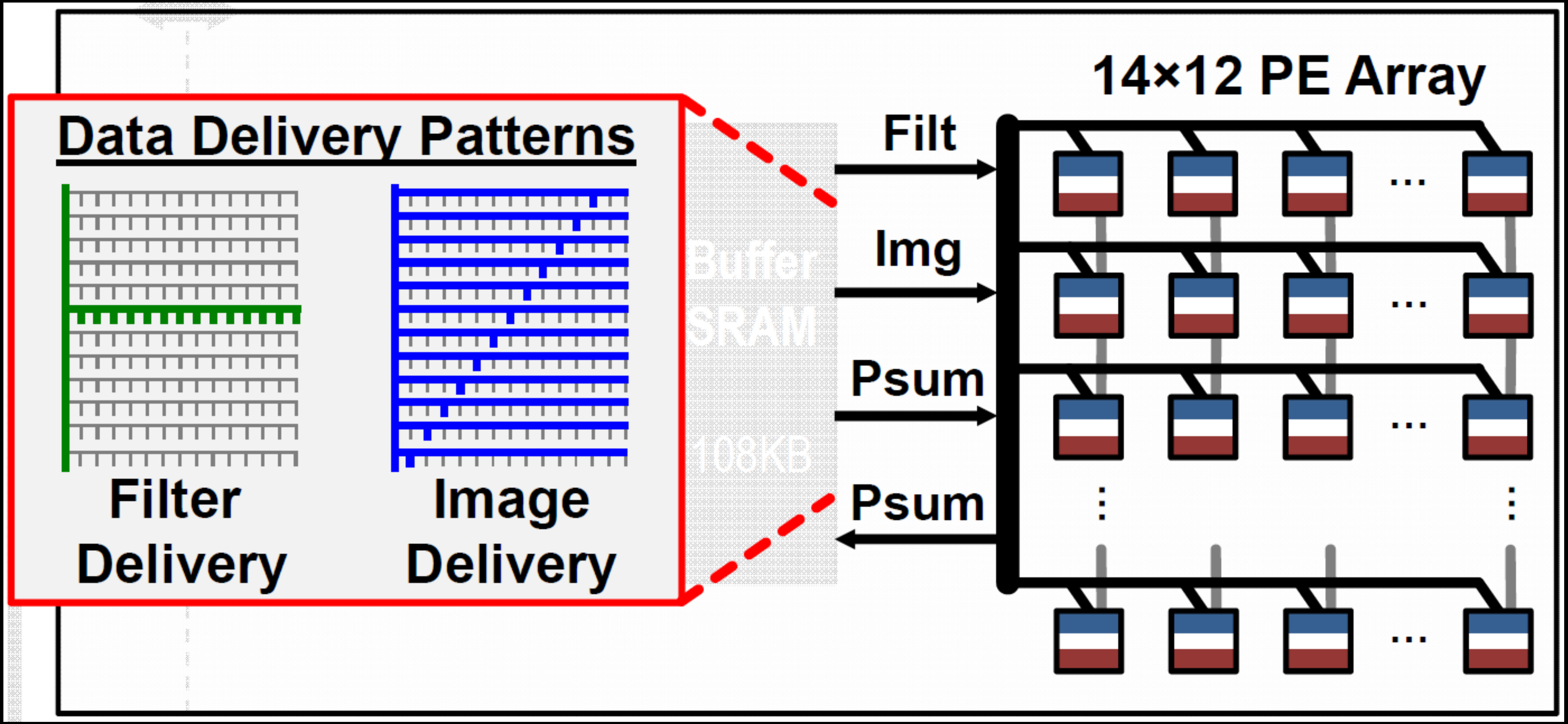
Don’t move data!

Keep input row * filter row within a PE

Dataflow on Accelerator



Dataflow in Practice



Multicast, P2P network saves power

Compress out zeroes (sparse data)

CNN Accelerator Recap

Beats general-purpose GPUs easily!

- 10X less DRAM bandwidth
- Much better perf/watt

Why? Specialized architecture!

- Custom network-on-chip
- Minimal data movement/compression
- Simple cores

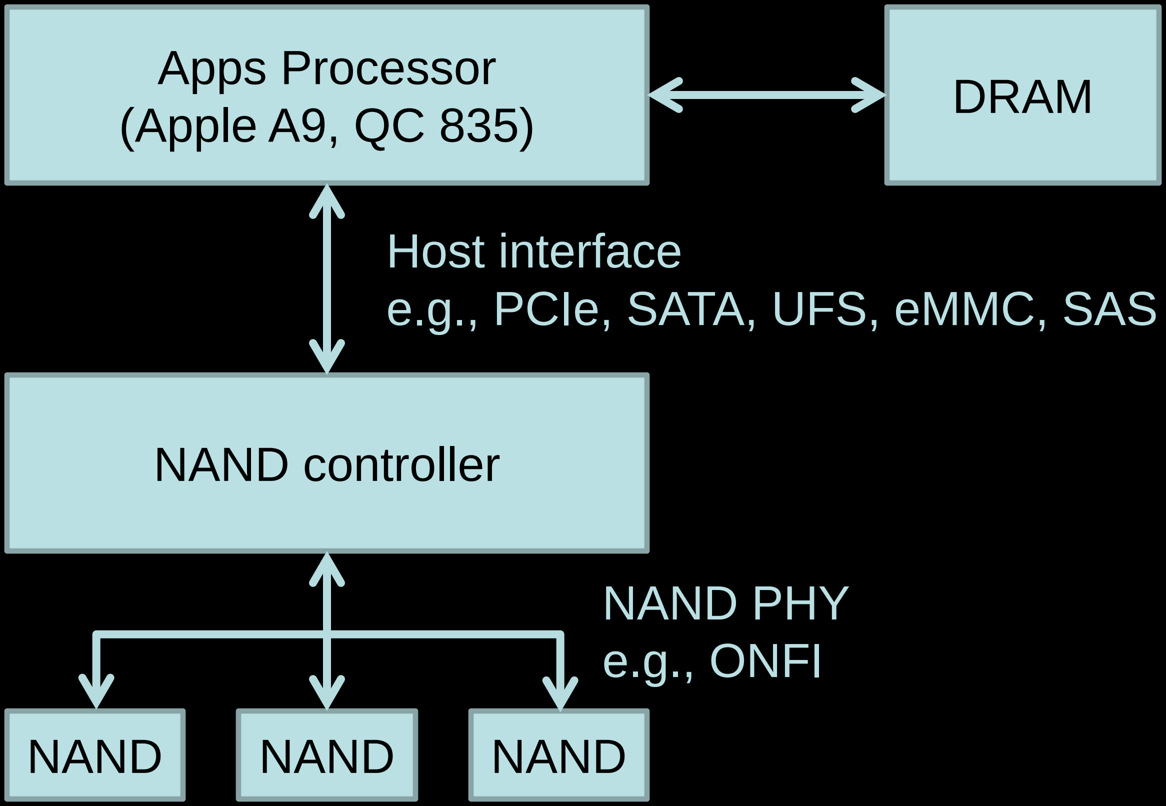
Mobile Full Disk Encryption

Security, privacy important for customers

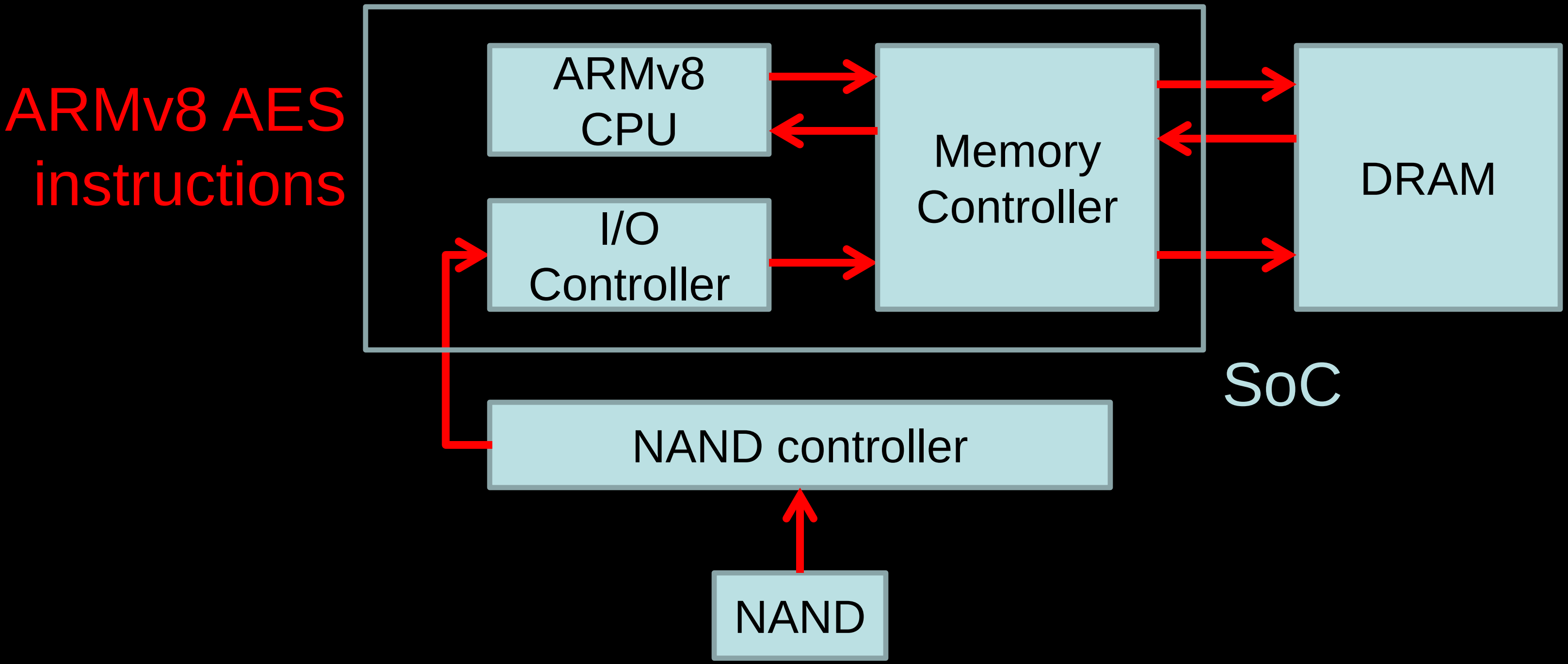
Challenges

- Performance overhead
- Power/energy cost
- Application adoption

Mobile System Architecture



Software Crypto

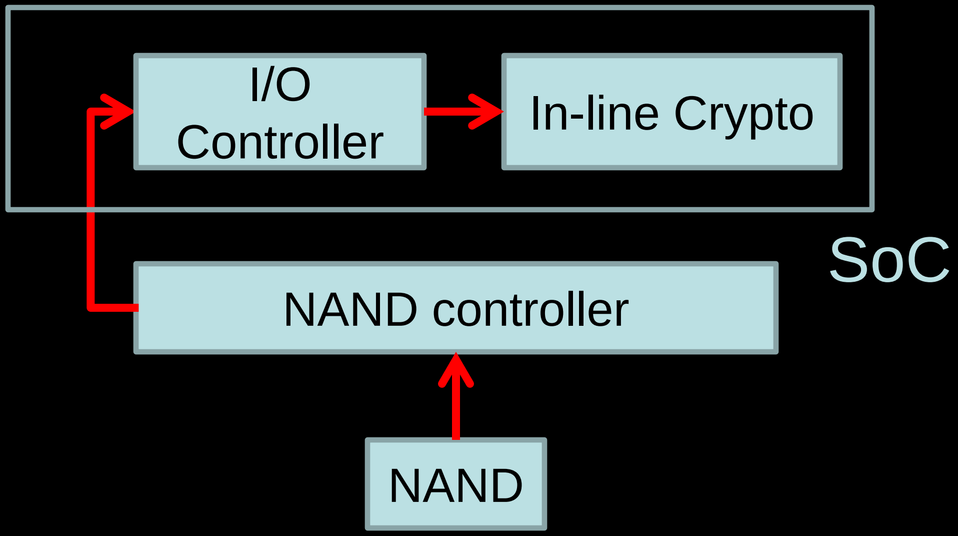


Works on all **ARMv8** systems... ☺

...slowly and ineffeciently ☹

Prepared for EEC 277

Hardware Crypto

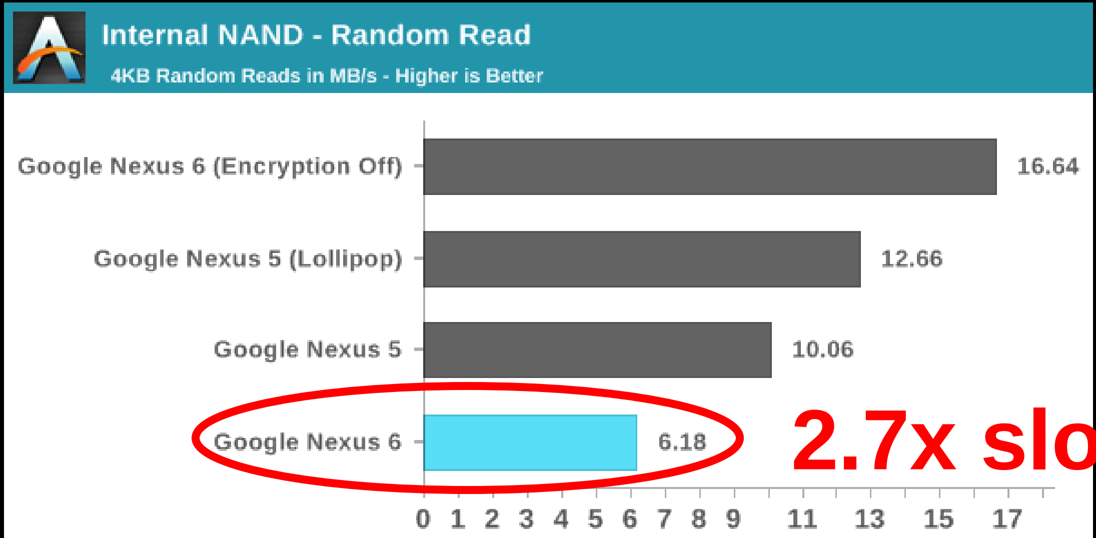


Must spend area on crypto module ☹

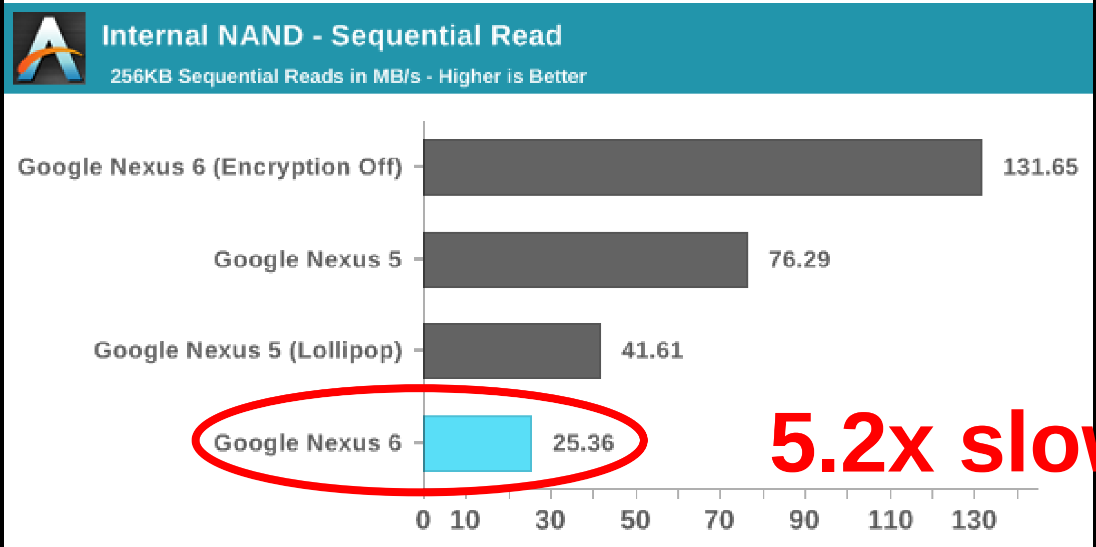
No change to NAND or controller ☺

...actually works well! ☺

SW Crypto Performance

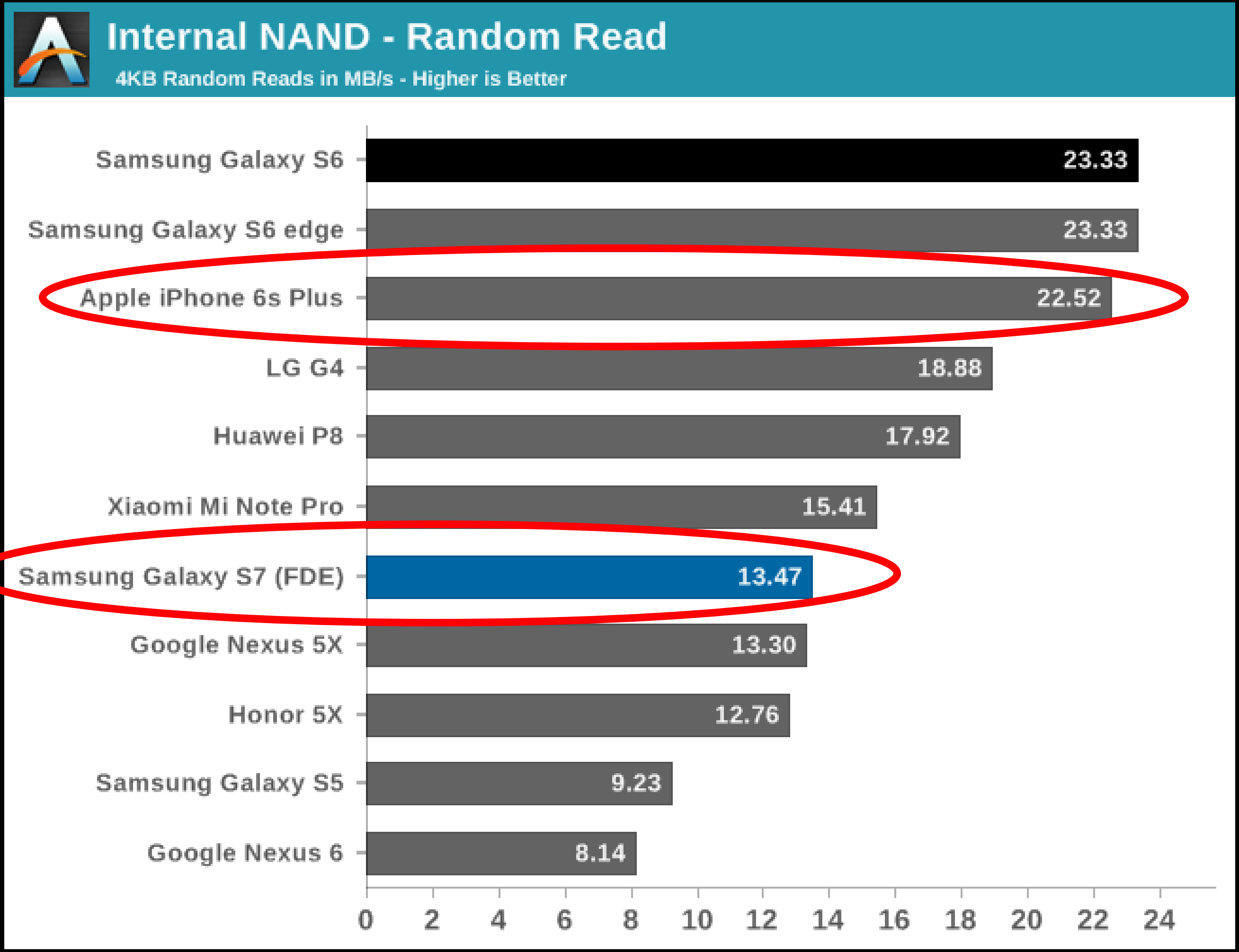


2.7x slower



5.2x slower

HW Crypto Performance



Accelerator Challenges I

- Is the workload stable?
  - May need programmable solution
  - E.g., Intel Gen GPU for media pipeline
  - Can harden over time

- How do we expose accelerators?
  - Instructions, e.g., AESNI, TSX
  - DSLs, e.g., Halide; APIs, e.g., Tensorflow
  - Don’t enable 3rd parties (e.g., Apple DSPs)

Accelerator Challenges II

- How often can we use an accelerator?
  - Determined by market/customer/use case
  - Will this run on GPU, DSP (existing HW)?
  - Developer adoption usually necessary

- How do I get paid?
  - Pay attention to volume and economics
  - Enabling new capabilities is best
  - Selling better perf/power is OK

- Garden of Eden
- (Silicon) Paradise Lost
- Future of Computing
- Case Studies
- **What Can You Do?**

Focus on the System

- Intersection of SW and HW
  - – Understand hardware and underlying tech
    - Don’t be that guy who defrags an SSD!
  - – Cross layers of abstraction
    - Technical, business, company, etc.

- Learn new technologies
  - – Computers were vacuum tubes in the 60’s
  - – Have a critical eye and remember economics

Future Options to Ponder

- Computing
  - Quantum, near-threshold, approximate, FPGA
- Memory
  - ReRAM, MRAM, 3D Xpoint, CBRAM, HMC
- Exotic semiconductors (e.g., GaN)
- MEMs
- Low power communications
- Displays, sensors

Q&A

Suggested Reading

RWT: www.realworldtech.com

Accelerators, NTV: http://bit.ly/2lhOZtv

CBRAM: http://bit.ly/2mt7E5u

Hot Chips: http://www.hotchips.org/archives/