

# ESE532: System-on-a-Chip Architecture

Day 22: April 10, 2017  
Energy



## Today

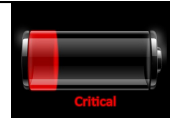
### Energy

- Today's bottleneck
- What drives
- Efficiency of
  - Processors, FPGAs, accelerators

## Message

- Energy dominates
  - Including limiting performance
- Make memories small and wires short
  - Small memories cost less energy per read
- Accelerators reduce energy
  - Compared to processors

## Energy



- Growing domain of portables
  - Less energy/op → longer battery life
- Global Energy Crisis
- Power-envelope at key limit
  - E reduce → increase compute in P-envelope
  - Scaling
    - Power density **not** transistors limit sustained ops/s
  - Server rooms
    - Cost-of-ownership **not** dominated by Silicon
    - **Cooling**, **Power** bill

## Preclass 1--4

- 20,000 gates/mm<sup>2</sup>
- 2.5\*10<sup>-15</sup> J/gate switch
- Gates on 1cm<sup>2</sup>
- Energy to switch all?
- Power at 1GHz?
- Fraction can switch with 1W/cm<sup>2</sup> power budget?

## Challenge: Power

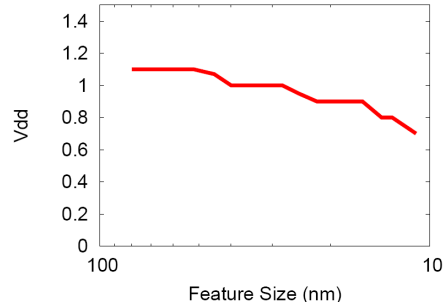
## Origin of Power Challenge

- Limited capacity to remove heat
  - ~100W/cm<sup>2</sup> force air
  - 1-10W/cm<sup>2</sup> ambient
- Transistors per chip grow at Moore's Law rate =  $(1/F)^2$
- Energy/transistor must decrease at this rate to keep constant power density
- $P/tr \propto CV^2f$
- $E/tr \propto CV^2$ 
  - ...but V scaling more slowly than F

Penn ESE532 Spring 2017 -- DeHon

7

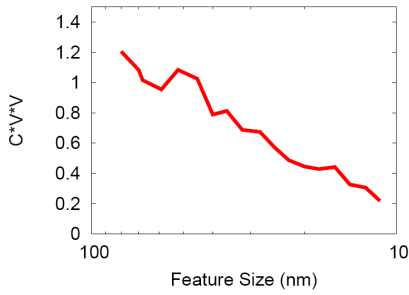
## ITRS V<sub>dd</sub> Scaling: More slowly than F



Penn ESE532 Spring 2017 -- DeHon

8

## ITRS CV<sup>2</sup> Scaling: More slowly than $(1/F)^2$

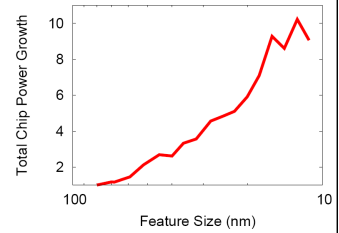


Penn ESE532 Spring 2017 -- DeHon

9

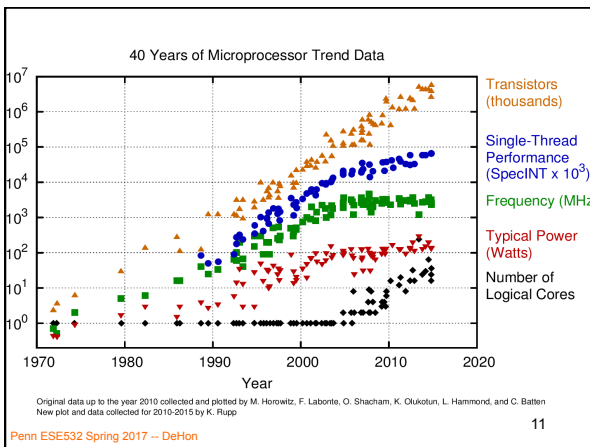
## Origin of Power Challenge

- Transistors per chip grow at Moore's Law rate =  $(1/F)^2$
- Energy/transistor must decrease at this rate to keep constant
- $E/tr \propto CV^2$



Penn ESE532 Spring 2017 -- DeHon

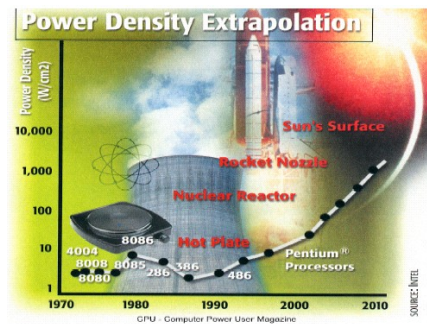
10



Penn ESE532 Spring 2017 -- DeHon

11

## Intel Power Density

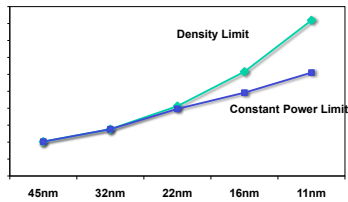


Penn ESE532 Spring 2017 -- DeHon

12

## Impact

### Power Limits Integration



Source: Carter/Intel

Penn ESE532 Spring 2017 -- DeHon

13

13

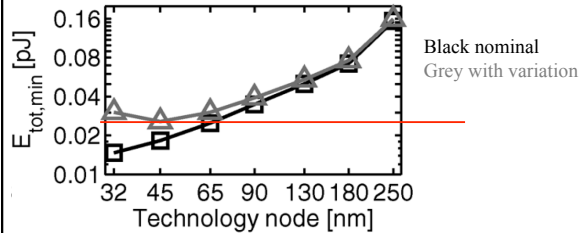
## Impact

- Power density is limiting scaling
  - Can already place more transistors on a chip than we can afford to turn on!
- Power is potential challenge/limiter for all future chips.
  - Only turn on small percentage of transistors?
  - Operate those transistors as much slower frequency?
  - Find a way to drop  $V_{dd}$ ?

Penn ESE532 Spring 2017 -- DeHon

14

## Variation threatens E/Op reduction



Min-Energy for multiplication (typically subthreshold)

[Bol et al., IEEE TR VLSI Sys 17(10):1508—1519]

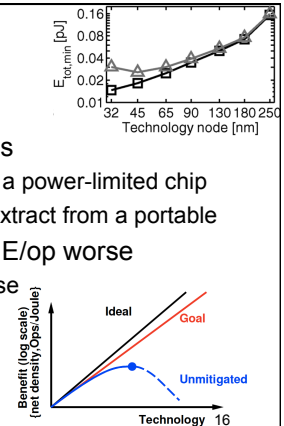
Penn ESE532 Spring 2017 -- DeHon

15

## Energy Limited

- It is Energy that defines
  - Ops/s can extract from a power-limited chip
  - Ops/battery-hour can extract from a portable
- If a technology makes E/op worse
  - That technology is worse

**–End-of-scaling**



Penn ESE532 Spring 2017 -- DeHon

## Energy

$$E_{total} = E_{switch} + E_{leak}$$

Penn ESE532 Spring 2017 -- DeHon

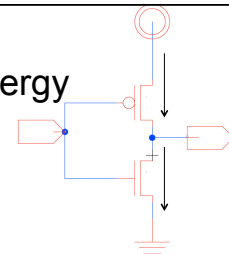
17

## Leakage Energy

- $I_{leak}$ 
  - Subthreshold leakage
  - (possibly) Gate-Drain leakage

$$P_{leak} = I_{leak} \times V$$

$$E_{leak} = P_{leak} \times T$$



Penn ESE532 Spring 2017 -- DeHon

18

## Switching Energy

$$E_{switch} \propto \alpha CV^2$$

- C – driven by architecture
  - Also impacted by variation, aging
- V – today, driven by variation, aging
- $\alpha$  – driven by architecture, coding/information

## Energy

$$E_{total} = E_{switch} + E_{leak}$$

$$E_{switch} \propto \alpha CV^2$$

$$E_{leak} = I_{leak} \times V \times T$$

## Preclass 6

Memory bank

- Leaks at  $8\mu W$
- Switches 24pJ/read

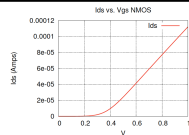
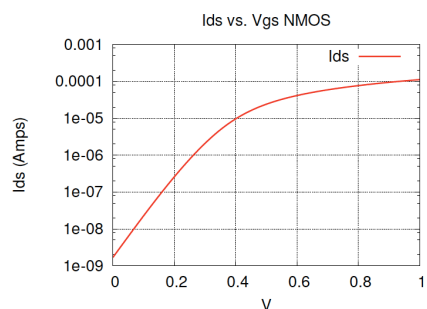
- At what rate of reads does  $E_{switch} > E_{leak}$ ?

## Voltage

$$E_{switch} \propto \alpha CV^2 \quad E_{leak} = I_{leak} \times V \times T$$

- We can set voltages
  - $V_{TH}$  – threshold voltage
  - $V_{dd}$  – switching/operating voltage
- Typically need
  - $V_{dd} > V_{TH} > 0$
  - (can have  $V_{dd} \sim V_{TH}$ , maybe even below)

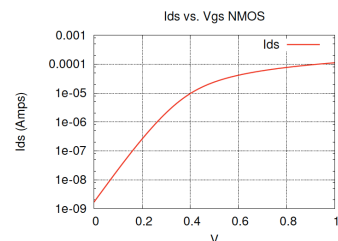
## $I_{DS}$ vs. $V_{GS}$



## Voltage

$$E_{switch} \propto \alpha CV^2 \quad E_{leak} = I_{leak} \times V \times T$$

- What happens to T if V set low? [ $T=CV/I$ ]
- E.g.  $V \rightarrow 400mV, 300mV, 200mV$ ?



## Operating a Transistor

- Concerned about  $I_{on}$  and  $I_{off}$
- $I_{on}$  drive (saturation) current for charging
  - Determines speed (latency):  $T_{gd} = CV/I$
- $I_{off}$  leakage current
  - Determines leakage power/energy:
    - $P_{leak} = V \times I_{leak}$
    - $E_{leak} = V \times I_{leak} \times T_{cycle}$

Penn ESE532 Spring 2017 -- DeHon

25

## Leakage

- To avoid leakage want  $I_{off}$  very small
- Switch  $V$  from  $V_{dd}$  to 0
- $V_{gs}$  in off state is 0 ( $V_{gs} < V_{TH}$ )

$$I_{sub} = I_{VT} \times 10^{((V_{gs} - V_{TH})/S)}$$

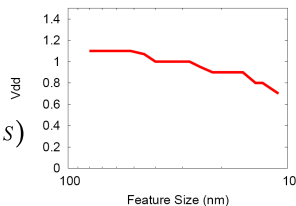
$$I_{off} = I_{VT} \times 10^{-((V_{TH})/S)}$$

Penn ESE532 Spring 2017 -- DeHon

26

## Leakage

$$I_{off} = I_{VT} \times 10^{-((V_{TH})/S)}$$

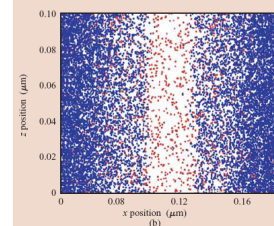
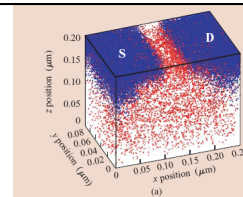


- $S \approx 90\text{mV}$  for single gate
- $S \approx 70\text{mV}$  for double gate (FinFET)
- For lowest leakage, want  $S$  small,  $V_{TH}$  large
- 4 orders of magnitude  $I_{VT}/I_{off} \rightarrow V_{TH} > 280\text{mV}$

Leakage limits  $V_{TH}$  in turn limits  $V_{dd}$  27

Penn ESE532 Spring 2017 -- DeHon

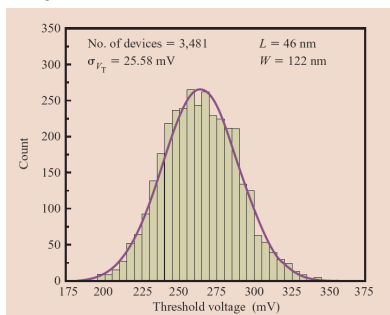
## Statistical Dopant Count and Placement



[Bernstein et al, IBM JRD 2006]

Penn ESE532 Spring 2017 -- DeHon

## $V_{th}$ Variability @ 65nm



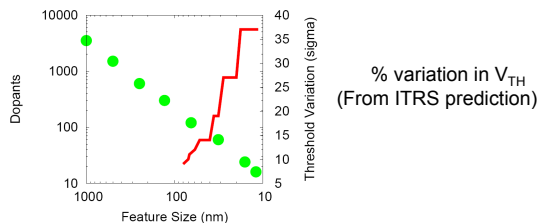
Penn ESE532 Spring 2017 -- DeHon

[Bernstein et al, IBM JRD 2006]

29

## Variation

- Fewer dopants, atoms  $\rightarrow$  increasing Variation
- How do we deal with variation?



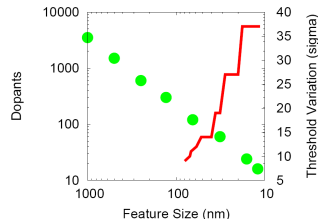
Penn ESE532 Spring 2017 -- DeHon

30

30

## Variation

- Increasing variation forces higher voltages
  - On top of our leakage limits



Penn ESE532 Spring 2017 -- DeHon

31

## Variations

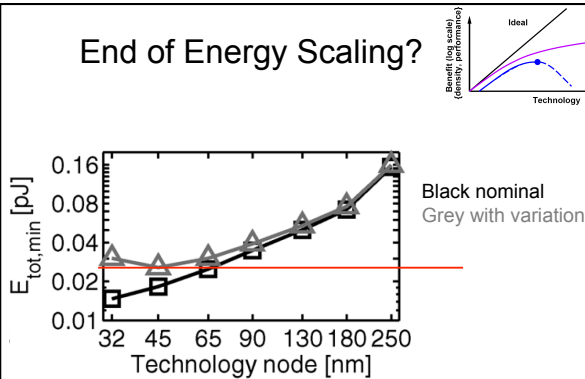
- Margins growing due to increasing variation
- Margined value may be worse than older technology?



Penn ESE532 Spring 2017 -- DeHon

32

## End of Energy Scaling?



Penn ESE532 Spring 2017 -- DeHon

[Bol et al., IEEE TR VLSI Sys 17(10):1508—1519]<sup>33</sup>

33

## Scaling

- Voltage scaling mostly over
  - Need ~300mV for Ion/Ioff
  - Plus variation and noise margin

$$E_{switch} \propto \alpha CV^2 \quad E_{leak} = I_{leak} \times V \times T$$

Penn ESE532 Spring 2017 -- DeHon

34

## Extending Preclass 6 @ 22nm

- |   |  |
|---|--|
| <p>Low Power process</p> <ul style="list-style-type: none"> <li>Higher <math>V_{TH}</math> to reduce leakage</li> <li>(Higher <math>V_{dd}</math>)</li> </ul> <p>Memory Bank</p> <ul style="list-style-type: none"> <li>Leaks at 21μW</li> <li>Switches 7.4pJ/read</li> </ul> | <p>High Performance process</p> <ul style="list-style-type: none"> <li>Lower <math>V_{TH}</math> <ul style="list-style-type: none"> <li>So <math>V_{dd}-V_{TH}</math> larger</li> <li>Runs faster</li> </ul> </li> <li>(Lower <math>V_{dd}</math>)</li> </ul> <p>Memory Bank</p> <ul style="list-style-type: none"> <li>Leaks at 5mW</li> <li>Switches 7.1pJ/read</li> </ul> |
|---|--|

Crossover for each? When HP preferred?

Penn ESE532 Spring 2017 -- DeHon

35

## Switching Energy

$$E_{switch} \propto \alpha CV^2$$

- C – driven by architecture
  - Also impacted by variation, aging
- V – today, driven by variation, aging
- $\alpha$  – driven by architecture, information

Penn ESE532 Spring 2017 -- DeHon

36

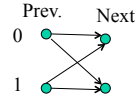
## Data Dependent Activity

- Consider an 8b counter
  - How often do each of the following switch?
    - Low bit?
    - High bit?
  - Average switching across all 8 output bits?
- Assuming random inputs
  - Activity at output of nand4?
  - Activity at output of xor4?

Penn ESE532 Spring 2017 -- DeHon

37

## Gate Output Switching (random inputs)



$$P_{switch} = P(0@i) * P(1@i+1) + P(1@i) * P(0@i+1)$$

Penn ESE532 Spring 2017 -- DeHon

38

## Switching Energy

$$E_{switch} = \left( \sum_i \alpha_i C_i \right) V^2$$

$C_i$  == capacitance driven by each gate (including wire)

Penn ESE532 Spring 2017 -- DeHon

39

## Switching Rate ( $\alpha_i$ ) Varies

- Different logic (low/high bits, gate type)
- Different usage
  - Gate off unused functional units
- Data coded
- Entropy in data
- Average  $\alpha$  5--15% plausible

$$E_{switch} = \left( \sum_i \alpha_i C_i \right) V^2$$

Penn ESE532 Spring 2017 -- DeHon

40

## Switching Energy

$$E_{switch} \propto \alpha C V^2$$

- $C$  – driven by architecture
  - Also impacted by variation, aging
- $V$  – today, driven by variation, aging
- $\alpha$  – driven by architecture, information

Penn ESE532 Spring 2017 -- DeHon

41

## Wire Driven

$$E_{switch} = \left( \sum_i \alpha_i C_i \right) V^2$$

- Gates drive
  - Self
  - Inputs to other gates
  - Wire routing between self and other gates
- Typically:
  - $C_{wire} > C_{self} + C_{load}$

Penn ESE532 Spring 2017 -- DeHon

42

## Wire Capacitance

- How does wire capacitance relate to wire length?

## Wire Capacitance

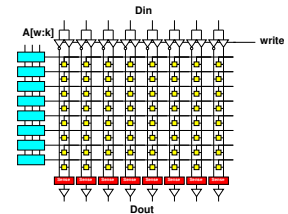
- $C = \epsilon A/d = \epsilon W * L_{\text{wire}}/d = C_{\text{unit}} * L_{\text{wire}}$
- Wire capacitance is linear in wire length
- E.g. 1.7pF/cm (preclass)
- Remains true if buffer wire
  - Add buffered segment at fixed lengths

## Wire Driven Implications

- Care about locality
  - Long wires are higher energy
  - Producers near consumers
  - Memories near compute
  - Esp. for large  $\alpha_i$ 's
- Care about size/area
  - Reduce (worst-case) distance must cross
- Care about minimizing data movement
  - Less data, less often, smaller distances
- Care about size of memories

## Preclass 5

- Primary switching capacitance in wires
- How does energy of a read grow with capacity (N) of a memory bank?
- Energy per bit?



## Memory Implications

- Memory energy can be expensive
- Small memories cost less energy than large memories
  - Use data from small memories as much as possible
- Cheaper to re-use data item from register than re-reading from memory

## Architectural Implications



## Component Numbers

# TABLE 1

Operation	Energy
32-bit arithmetic operation	5 pJ
32-bit register read	10 pJ
32-bit 8KB RAM read	50 pJ
32-bit traverse 10mm wire	100 pJ
Execute instruction	500 pJ

Energy Per Operation (0.13µm, 1.2V)

Penn ESE532 Spring 2017 -- DeHon [Dally, March 2004 ACM Queue] 49

## Component Numbers

- Processor instruction 100x more than arithmetic
- Register read 2x
- RAM read 10x
- Why processor instruction > arith operation?

# TABLE 1

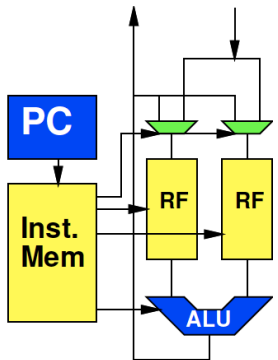
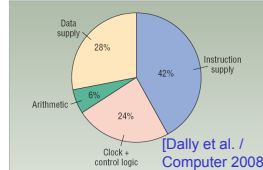
Operation	Energy
32-bit arithmetic operation	5 pJ
32-bit register read	10 pJ
32-bit 8KB RAM read	50 pJ
32-bit traverse 10mm wire	100 pJ
Execute instruction	500 pJ

Energy Per Operation (0.13µm, 1.2V)

Penn ESE532 Spring 2017 -- DeHon [Dally, March 2004 ACM Queue] 50

## Processors and Energy

- Very little into actual computation
- Determine and Fetch Instruction
- Read and Write data from memories

[Dally et al. / Computer 2008]

Penn ESE532 Spring 2017 -- DeHon

## ARM Cortex A9

Estimate find: 0.5W at 800MHz in 40nm

- $0.5/0.8 \times 10^{-9}$  J/instr
- ~600pJ/instr
- Scale to 28nm
  - maybe  $0.7 \times 600$ — $0.5 \times 600$
  - 300—400pJ/instr ?
- Is superscalar w/ neon, so not as simple a processor as previous example

Penn ESE532 Spring 2017 -- DeHon 52

## ARM Cortex A7, A15 (Samsung 28nm)

Instruction	Cortex-A7		Cortex-A15	
	min EPI	max EPI	min EPI	max EPI
Simple Integer	50	80	200	450
Simple Float/Double	90	200	250	1500
Multiplication	80	340	360	1730
Division	150	1200	1270	1960
Load (L1 hit)	150	195	450	450
Store (L1 hit)	185	195	680	750
Store (L1 miss)		200		700
Load (L1 miss)		270		1000

[Evangelos Vasilakis, Technical Report FORTH-ICS/TR-450, March 2015]  
<http://www.ics.forth.gr/carv/greenvm/files/tr450.pdf> 53

Penn ESE532 Spring 2017 -- DeHon

## Processor Differences

- What different among A7, A9, A15?

Penn ESE532 Spring 2017 -- DeHon 54

## ARM Cortex A7, A15 (Samsung 28nm)

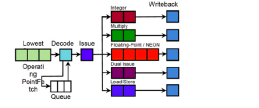
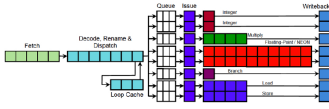


Figure 5.2: ARM Cortex-A7 Pipeline

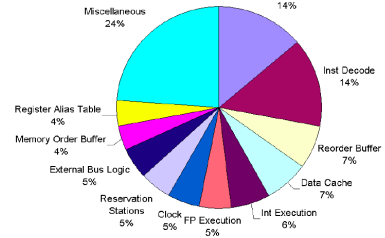


[Evangelos Vasilakis, Technical Report FORTH-ICS/TR-450, March 2015]

<http://www.ics.forth.gr/carv/greenvm/files/tr450.pdf> 55

Penn ESE532 Spring 2017 -- DeHon

## Pentium Pro Energy Breakdown



[Bose, Martonosi, Brooks / Sigmetrics 2001]

Penn ESE532 Spring 2017 -- DeHon

## Implications

- Complex, multi-issue superscalars
  - Cost more energy per operation
  - Spend energy on issue logic, etc. that does not go into computation for the task
- Even if can get performance from superscalar processors
  - For energy reasons, benefit getting it elsewhere

Penn ESE532 Spring 2017 -- DeHon

57

## Zynq

Operation	PL Resource	ARM A9 Resource	ARM A9 energy/OP (pico Joules or mW/GOP/sec)	PL energy/OP (pico Joules or mW/GOP/sec)
Logical Op of 2 var	LUT/FF	ALU		1.3
32-bit ADD	LUT/FF	ALU		1.3
16x16 Mult	DSP	ALU		8.0
32-bit Read/Write register	LUTRAM	L1		1.4
32-bit Read/Write AXI register	LUT/FF	AXI		30
32-bit Read/Write local RAM	BRAM	L2		23.7/17.2
32-bit Read/Write OCM	AXI/OCM	CPU/OCM		44
32-bit Read/Write DDR3	AXI/DDR	CPU/DDR		541/211

- ARM A9 instruction 300—400pJ
- ARM A9 L1 cache read 23pJ

Penn ESE532 Spring 2017 -- DeHon

Xilinx UG585 – Zynq TRM 58

## Compare

- Assume ARM Cortex A9 executes 4x32b Neon vector add instruction for 300pJ
- Compare to 32b adds on FPGA?

Operation	PL Resource	ARM A9 Resource	ARM A9 energy/OP (pico Joules or mW/GOP/sec)	PL energy/OP (pico Joules or mW/GOP/sec)
Logical Op of 2 var	LUT/FF	ALU		1.3
32-bit ADD	LUT/FF	ALU		1.3
16x16 Mult	DSP	ALU		8.0
32-bit Read/Write register	LUTRAM	L1		1.4
32-bit Read/Write AXI register	LUT/FF	AXI		30
32-bit Read/Write local RAM	BRAM	L2		23.7/17.2
32-bit Read/Write OCM	AXI/OCM	CPU/OCM		44
32-bit Read/Write DDR3	AXI/DDR	CPU/DDR		541/211

Penn ESE532 Spring 2017 -- DeHon

## Compare

- Assume ARM Cortex A9 executes 8x16b Neon vector multiply instruction for 300pJ
- Compare to 16x16 multiplies on FPGA?

Operation	PL Resource	ARM A9 Resource	ARM A9 energy/OP (pico Joules or mW/GOP/sec)	PL energy/OP (pico Joules or mW/GOP/sec)
Logical Op of 2 var	LUT/FF	ALU		1.3
32-bit ADD	LUT/FF	ALU		1.3
16x16 Mult	DSP	ALU		8.0
32-bit Read/Write register	LUTRAM	L1		1.4
32-bit Read/Write AXI register	LUT/FF	AXI		30
32-bit Read/Write local RAM	BRAM	L2		23.7/17.2
32-bit Read/Write OCM	AXI/OCM	CPU/OCM		44
32-bit Read/Write DDR3	AXI/DDR	CPU/DDR		541/211

Penn ESE532 Spring 2017 -- DeHon

## Programmable Datapath

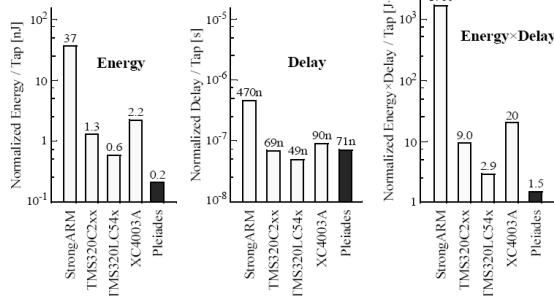
- Performing an operation in a pipelined datapath can be orders of magnitude less energy than on a processor
  - ARM 300pJ vs. 1.3pJ 32b add
  - Even neon 300pJ vs. 4x1.3pJ for 4x32b add
  - 300pJ vs. 8x8pJ for 8 16x16b multiplies

## Zynq

Operation	PL Resource	ARM A9 Resource	ARM A9 energy/OP (pico Joules or mW/GOP/sec)	PL energy/OP (pico Joules or mW/GOP/sec)
Logical Op of 2 var	LUT/FF	ALU		1.3
32-bit ADD	LUT/FF	ALU		1.3
16x16 Mult	DSP	ALU		8.0
32-bit Read/Write register	LUTRAM	L1		1.4
32-bit Read/Write AXI register	LUT/FF	AXI		30
32-bit Read/Write local RAM	BRAM	L2		23.7/17.2
32-bit Read/Write OCM	AXI/OCM	CPU/OCM		44
32-bit Read/Write DDR3	AXI/DDR	CPU/DDR		541/211

- Reading from OCM order of magnitude less than from DRAM
- ...and BRAM half that

## Energy



[Abnous et al, *The Application of Programmable DSPs in Mobile Communications*, Wiley, 2002, pp. 327-360] 63

## FPGA vs. Std Cell Energy

- 90nm
- FPGA: Stratix II
- STMicro CMOS090
- eASIC (MPGA) claim
  - 20% of FPGA power
  - (best case)

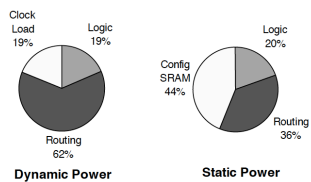
TABLE VI  
DYNAMIC POWER CONSUMPTION RATIO (FPGA/ASIC)

Name	Method	Logic Only	Logic & DSP	Logic Memory	Logic, Memory & DSP
booth	Sim	26			
rs.encoder	Sim	52			
cordic18	Const	6.3			
cordic8	Const	5.7			
des.area	Const	27			
des.perf	Const	9.3			
fir_restruct	Const	9.6			
mael1	Const	19			
aes192	Sim	12			
fir3	Const	12	7.5		
diffeq	Const	15	12		
diffeq2	Const	16	12		
molecular	Const	15	16		
rs.decoder1	Const	13	16		
rs.decoder2	Const	11	11		
atm	Const			15	
aes	Sim			13	
aes.inv	Sim			12	
ethernet	Const			16	
serialproc	Const			16	
fib24	Const				5.3
pipeproc	Const				8.2
raytracer	Const				8.3
Geomean		14	12	14	7.1

[Kuon/Rose TRCADv26n2p203--215 2007]

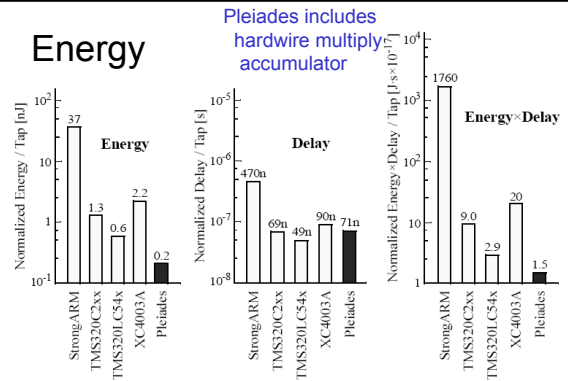
## FPGA Disadvantage to Custom

- Interconnect Energy
  - Long wires → more capacitance → more E
  - Switch Energy is an overhead



[Tuan et al./FPGA 2006]

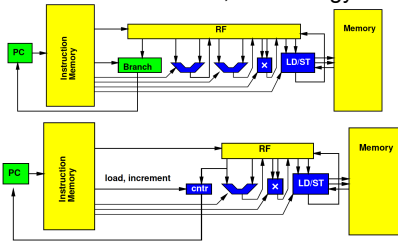
## Energy



[Abnous et al, *The Application of Programmable DSPs in Mobile Communications*, Wiley, 2002, pp. 327-360] 66

## Zero-Overhead Loop Simplify

- TI DSPs specialized w/ tricks like ZOL...
  - Fewer instructions, less energy/instruction

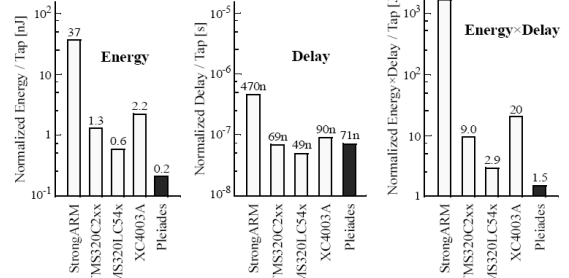


Penn ESE532 Spring 2017 -- DeHon

67

## Energy

Pleiades includes  
hardwire multiply  
accumulator



[Abnous et al, *The Application of Programmable DSPs in Mobile Communications*, Wiley, 2002, pp. 327-360]

Penn ESE532 Spring 2017 -- DeHon

68

## Simplified Comparison

- Processor two orders of magnitude higher energy than custom accelerator
- FPGA accelerator in between
  - Order of magnitude lower than processor
  - Order of magnitude higher than custom

Penn ESE532 Spring 2017 -- DeHon

69

## Big Ideas

- Energy dominance
- With power-density budget
  - The most energy efficient architecture delivers the most performance
- Make memories small and wires short
- SoC, accelerators reduce energy by reducing processor instruction execution overhead

Penn ESE532 Spring 2017 -- DeHon

70

## Admin

- Project energy Milestone
  - Due Friday

Penn ESE532 Spring 2017 -- DeHon

71