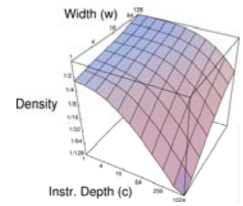
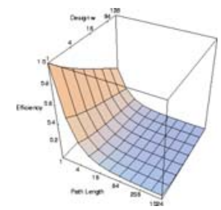
CS184a: Computer Architecture (Structure and Organization)

Day 10: January 28, 2005
Empirical Comparisons

Last Time

- Instruction Space Modeling





Today

- Empirical Data
  - Processors
  - FPGAs
  - Custom
    - Gate Array
    - Std. Cell
    - Full
  - Tasks

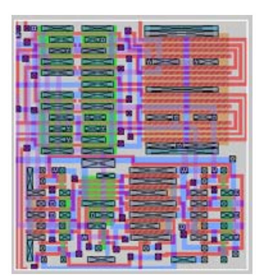
Empirical Comparisons

Empirical

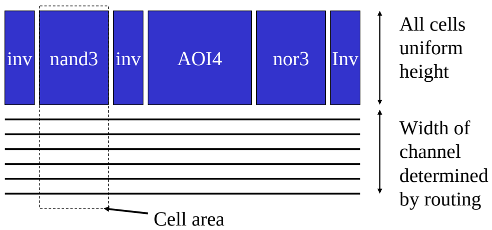
- Ground modeling in some concretes
- Start sorting out
  - custom vs. configurable
  - spatial configurable vs. temporal

Full Custom

- Get to define all layers
- Use any geometry you like
- Only rules are process design rules
- CS181



Standard Cell Area



Identify the full custom and standard cell regions on 386DX die
http://microscope.fsu.edu/chipshots/intel/386dxlarge.html

MPGA

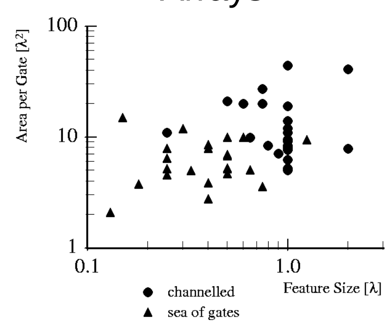
- Metal Programmable Gate Array
- Gates pre-placed (poly, diffusion)
- Only get to define metal connections
  - Cheap – only have to pay for metal mask(s)

MPGA vs. Custom?

- AMI CICC'83
  - MPGA 1.0
  - Std-Cell 0.7
  - Custom 0.5
- Toshiba DSP
  - Custom 0.3
- Mosaid RAM
  - Custom 0.2
- GE CICC'86
  - MPGA 1.0
  - Std-Cell 0.4--0.7
    - FF/counter 0.7
    - FullAdder 0.4
    - RAM 0.2

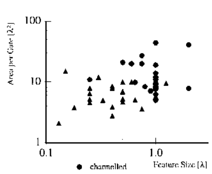
MPGA = Metal Programmable Gate Array
(traditional Gate Array)

Metal Programmable Gate Arrays



MPGAs

- Modern -- "Sea of Gates"
- yield 35--70%
- maybe 5kλ^2/gate ?
  - (quite a bit of variance)



FPGA Table

Year	Design	Organization	Max	λ	λ^2 area	cycle
1986	Xilinx 2K	CLB (4-LUT)	100	1μ	500K	20 ns
1988	Xilinx 3K	CLB (2×4-LUT)	320	0.6μ	1.3M	13 ns
1992	Xilinx 4K	CLB (2×4-LUT +)	1024	0.6μ	1.25M	7 ns
1995	Xilinx 5K	CLB (4×4-LUTS)	484	0.3μ	2.25M	6 ns
1995	Altera 8K	LE (4-LUT)	1296	0.3μ	920K	7.5 ns
1995	ORCA 2C	PLC (4×4-LUT)	900	0.3μ	4.3M	7 ns
1998	HSRA	BLB (5-LUT/2×4-LUT ?)	–	0.2μ	2M	4 ns
	Model	4-LUT	2K	–	800K	–
	Model	4-LUT	16K	–	1M	–

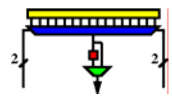
Modern FPGAs

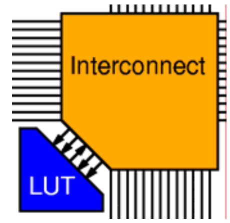
- APEX 20K1500E
  - 52K LEs
  - 0.18μm
  - 24mm × 22mm
  - 1.25Mλ^2/LE
- XC2V1000
  - 10.44mm x 9.90mm
    [source: Chipworks]
  - 0.15μm
  - 11,520 4-LUTs
  - 1. 5Mλ^2/4-LUT

[Both also have RAM in cited area]

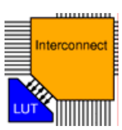
Conventional FPGA Tile

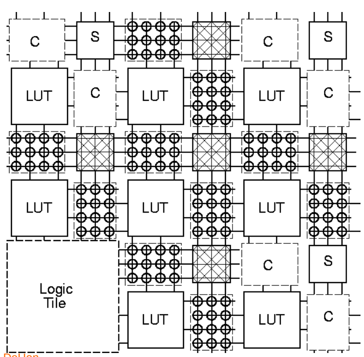
K-LUT (typical k=4)
w/ optional
output Flip-Flop



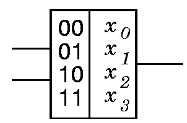


Toronto FPGA Model

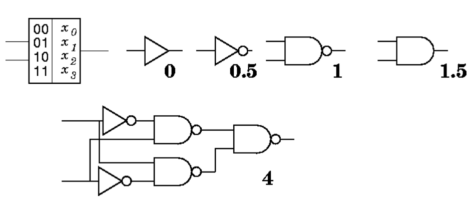




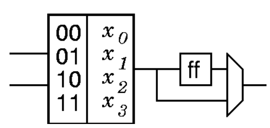
How many gates?

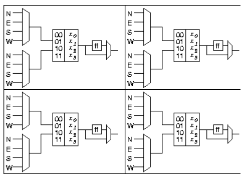


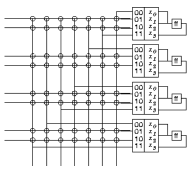
“gates” in 2-LUT



Now how many?





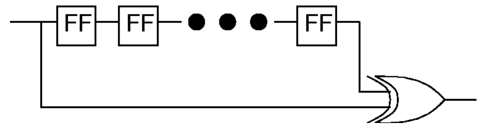


Which gives:

More usable gates?

More gates/unit area?

Gates Required?



Depth=3, Depth=2048?

Gate metric for FPGAs?

- Day8: several components for computations
  - compute element
  - interconnect:
    - space
    - time
  - instructions
- Not all applications need in same **balance**
- Assigning a single "capacity" number to device is an oversimplification

MPGA vs. FPGA

- MPGA (SOG GA)
  - 5Kλ^2/gate
  - 35-70% usable (50%)
  - 7-17Kλ^2/gate net
- Xilinx XC4K
  - 1.25Mλ^2 /CLB
  - 17--48 gates (26?)
  - 26-73Kλ^2/gate net
- Ratio: 2--10 (5)

Adding ~2x Custom/MPGA,
Custom/FPGA ~10x

MPGA vs. FPGA

- MPGA (SOG GA)
  - λ=0.6μ
  - τ_{gd}~1ns
- Xilinx XC4K
  - λ=0.6μ
  - 1-7 gates in 7ns
  - 2-3 gates typical
- Ratio: 1--7 (2.5)

Processors vs. FPGAs

Processors and FPGAs

Metric: $\frac{\text{4 input gate-evaluations}}{\lambda^2 \cdot s}$

Processor: $\frac{2 \times N_{ALU} \times w_{ALU}}{A_{proc} \times t_{cycle}}$ **FPGA:** $\frac{N_{4LUT}}{A_{array} \times t_{cycle}}$

Component Example

- Single die in 0.35µm

XC4085XL-09 3,136 CLBs 4.6ns

682 Bit Ops/ns

Alpha 1996 2×64b ALUs 2.3ns

55.7 Bit Ops/ns

[1 “bit op” = 2 gate evaluations]

Processors and FPGAs

Year	Design	Organization	λ	λ^2 area	cycle	$\frac{ge's}{\lambda^2 \cdot s}$
Microprocessors						
1984	MIPS	1 × 32	1.5µ	15M	250ns	17
1987	MIPS-X	1 × 32	1.0µ	68M	50ns	19
1994	MIPS	1 × 32	0.28µ	1.7G	2ns	19
1992	Alpha	1 × 64	0.38µ	1.7G	5ns	15
1995	Alpha	2 × 64	0.25µ	4.8G	3.3ns	18
1996	Alpha	2 × 64	0.18µ	6.8G	2.3ns	17
Reconfigurable ALUs						
1992	PADDI	8 × 16	0.6µ	126M	40ns	50
1995	PADDI-2	48 × 16	0.5µ	515M	20ns	150
FPGAs						
1986	Xilinx 2K	1 CLB (4 LUT)	1.0µ	500K	20ns	100
1988	Xilinx 3K	64 CLBs (2 4-LUT)	0.6µ	83M	13ns	120
1992	Xilinx 4K	49 CLBs (2 4-LUT)	0.6µ	61M	7ns	230
1995	Xilinx 5K	49 CLBs (4 4-LUT)	0.3µ	110M	6ns	290

Raw Density Summary

- Area
  - MPGA 2-3x Custom
  - FPGA 5x MPGA
- Area-Time
  - Gate Array 6-10x Custom
  - FPGA 15-20x Gate Array
  - Processor 10x FPGA

Raw Density Caveats

- Processor/FPGA may solve more specialized problem
- Problems have different resource balance requirements
  - …can lead to low yield of raw density

Homework

- Day behind
- Current assignment
  - Involves cascades PLAs

Task Comparisons

Broadening Picture

- Compare larger computations
- For comparison
  - throughput density metric: results/area-time
    - normalize out area-time point selection
    - high throughput density
      - → most in fixed area
      - → least area to satisfy fixed throughput target

Multiply

Architecture	Feature Size (λ)	Area and Time	16×16 mpy $\frac{}{\lambda^2 s}$	16×16 scale $\frac{}{\lambda^2 s}$	8×8 mpy $\frac{}{\lambda^2 s}$	8×8 scale $\frac{}{\lambda^2 s}$
Custom 16×16	0.63μm	2.6Mλ^2, 40 ns	9.6	9.6	9.6	9.6
Custom 8×8	0.80μm	3.3Mλ^2, 4.3 ns			70	70
Gate-Array 16×16	0.75μm	26Mλ^2, 30ns	1.3	1.3	1.3	1.3
FPGA	0.60μm	1.25Mλ^2/CLB				
(XC4K)		316 CLBs, 26 ns	0.097			
		84 CLBs, 40 ns		0.24		
		220 CLBs, 12.1 ns			0.30	
		22 CLBs, 25 ns				1.5
16b DSP	0.65μm	350Mλ^2, 50 ns	0.057	0.057	0.057	0.057
RISC	0.75μm	125Mλ^2, 66 ns/cycle				
(no multiplier)		two 16b operands – 44 cycles	0.0028			
		16b constant – 7 cycles		0.017		
		one 8b operand – 24 cycles			0.0051	
		8b constant – 4 cycles				0.030

Example: FIR Filtering

$Y_i = w_1 x_i + w_2 x_{i+1} + \ldots$

Application metric:
TAPs = filter taps
multiply accumulate

Architecture	Feature Size (λ)	$\frac{TAPs}{\lambda^2 s}$
32b RISC	0.75μm	0.020
16b DSP	0.65μm	0.057
32b RISC/DSP	0.25μm	0.021
64b RISC	0.18μm	0.064
FPGA (XC4K)	0.60μm	1.9
(Altera 8K)	0.30μm	3.6
Full Custom	0.75μm	3.6
	0.60μm	3.5
	0.75μm	2.4
(fixed coefficient)	0.60μm	56
(*n.b.* 16b samples)		

IIR/Biquad

Architecture	Feature Size (λ)	Area and Time	16b $\frac{TAPs}{\lambda^2 s}$	10b $\frac{TAPs}{\lambda^2 s}$
16b DSP	0.60μm	200Mλ^2, 500 ns/biquad	0.010	0.010
FPGA	0.60μm	60 CLBs, 320 ns/biquad	0.044	
(XC4K)		43 CLBs, 200 ns/biquad		0.093
Full Custom	0.90μm	68Mλ^2, 11.8 ns/4 biquads		5.0

Simplest IIR: $Y_i = A \times X_i + B \times Y_{i-1}$

DES Keysearch

Architecture	Feature Size (λ)	Area	Keys/Second	$\frac{Keys}{\lambda^2 s}$
DES IC	1.5μm	11.1Mλ^2	310K	0.028
FPGA (Altera 8K)	0.30μm	81188 (930Mλ^2)	800K	0.00086
RISC	0.30μm	1.8Gλ^2	41K	0.000023

<http://www.cs.berkeley.edu/~iang/isaac/hardware/>

DNA Sequence Match

- **Problem:** "cost" of transform $S_1 \rightarrow S_2$
- **Given:** cost of insertion, deletion, substitution
- **Relevance:** similarity of DNA sequences
  - evolutionary similarity
  - structure predict function
- **Typically:** new sequence compared to large databse

DNA Sequence Match

Architecture	Feature Size (λ)	Area	Cell Updates per Second	$\frac{cu}{\lambda^2 s}$
Custom	2.0μm	270Mλ^2	500M	1.9
FPGA				
(SPLASH 2)	0.60μm	43Gλ^2	3,000M	0.070
(SPLASH)	0.60μm	33Gλ^2	370M	0.012
RISC				
(SparcStation I)	0.75μm	273Mλ^2	0.87M	0.0032
(SparcStation 10)	0.40μm	1.6Gλ^2	1.2M	0.00075

N.B. includes memory area for SPLASH

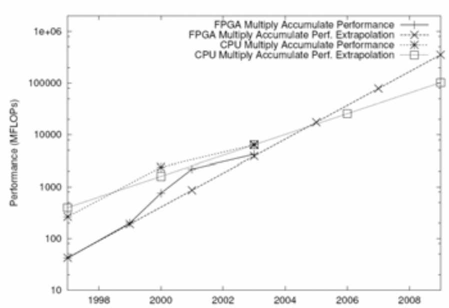
Floating-Point Add (single prec.)

Architecture	Reference	λ	area and time	$\frac{\text{32b FP ADDs}}{\lambda^2 s}$
Custom Macrocells				
32b FP ALU	JSSC92	0.6μ	32Mλ^2, 30 ns	1.0
32b Adder	EUROASIC92	0.4μ	49Mλ^2, 60 ns	0.34
32/64b FPU	CICC92	0.25μ	980Mλ^2, 2 per 12.5 ns	0.16
Coprocessor				
32b FP Processor	ISSCC88	0.75μ	220Mλ^2, 55 ns	0.083
32b FP Processor	SSC89	0.65μ	400Mλ^2, 23 ns	0.11
Processor FP support				
64b RISC w/FPU	ISSCC90	0.4μ	1.4Gλ^2, 2×25 ns cycle	0.014
64b RISC w/FPU	ISSCC92	0.38μ	1.7Gλ^2, 5 ns cycle	0.12
64b RISC w/FPU	ISSCC95	0.25μ	4.8Gλ^2, 3.3ns cycle	0.063
FPGA				
	FCCM96	0.3μ	500 Flex8K LEs (920Kλ^2/LE), 150 ns	0.014
	FCCM98	0.25μ	315 XC4KE CLBs (1.25Mλ^2/CLB), 33 ns	0.077
Processor no FP support				
32b RISC, no FPU	ASPLOS85	1.5μ	15Mλ^2, 16μs	0.0042

Floating-Point Mpy (single prec.)

Architecture	Reference	λ	area and time	$\frac{\text{32b FP MPYs}}{\lambda^2 s}$
Custom Macrocells				
32b mpy	JSSC92	0.6μ	43Mλ^2, 30 ns	0.78
32b mpy	EUROASIC92	0.4μ	81Mλ^2, 42 ns	0.29
32/64b FPU	CICC92	0.25μ	980Mλ^2, 2 per 12.5 ns	0.16
Coprocessor				
dedicated 32b mpy	JSSC84	1μ	33Mλ^2, 79 ns	0.38
dedicated 32b mpy	ISSCC87	0.6μ	160Mλ^2, 67 ns	0.093
32b FP Processor	ISSCC88	0.75μ	200Mλ^2, 55 ns	0.083
32b FP Processor	JSSC89	0.65μ	400Mλ^2, 23 ns	0.11
Processor FP support				
64b RISC w/FPU	ISSCC90	0.4μ	1.4Gλ^2 2×25 ns cycle	0.014
64b RISC w/FPU	ISSCC92	0.38μ	1.7Gλ^2, 5 ns cycle	0.12
64b RISC w/FPU	ISSCC95	0.25μ	4.8Gλ^2, 3.3ns cycle	0.063
FPGA				
	FCCM96	0.3μ	344 Flex8K LEs (920Kλ^2/LE), 755 ns	0.0042
	FCCM98	0.25μ	265 CLBs (1.25Mλ^2/CLB), 200 ns	0.0151
Processor no FP support				
32b RISC, no FPU	ASPLOS85	1.5μ	15Mλ^2, 7μs	0.0096

FPGA vs. Processor FP (Double Precision FP MAC)



[Underwood/FPGA'2004]

Degrade from Peak

Degrade from Peak: FPGAs

- Long path length → not run at cycle
- Limited throughput requirement
  - bottlenecks elsewhere limit throughput req.
- Insufficient interconnect
- Insufficient retiming resources (bandwidth)

Degrade from Peak: Processors

- Ops w/ no gate evaluations (interconnect)
- Ops use limited word width
- Stalls waiting for retimed data

$$E(\text{Functional Density}) = \frac{\text{Gate Evaluations}}{\text{Datapath Bit}} \times \frac{\text{Datapath Bits}}{\text{pinst}} \times \frac{\text{pinsts}}{\text{Issue Slot}} \times \frac{\text{Issue Slots}}{\text{Clock Cycle}} \times \frac{1}{\text{area} \times t_{cycle}}$$

Degrade from Peak: Custom/MPGA

- Solve more general problem than required
  - (more gates than really need)
- Long path length
- Limited throughput requirement
- Not needed or applicable to a problem

Degrade Notes

- We'll cover these issues in more detail as we get into them later in the course

Big Ideas [MSB Ideas]

- Raw densities: custom:ga:fpga:processor
  - 1:5:100:1000
  - close gap with specialization