ESE534:
Computer Organization
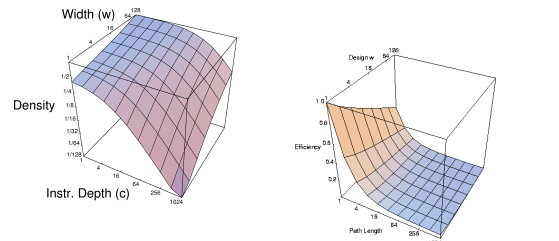
Day 14: March 12, 2012
Empirical Comparisons

Penn

Previously

• Instruction Space Modeling

Width (w)

Density

Instr. Depth (c)

Before Break

• Programmable compute blocks
  – LUTs, ALUs, PLAs
  – …still working on PLA assignment

Today

• What if we just built a completely custom circuit?
• What cost are we paying for programmability?
• Can we afford to build completely custom circuits?
  – Can we afford not to?

Today

• Empirical Data
  – Custom
    • Gate Array
    • Std. Cell (ASIC)
    • Full
  – FPGAs
  – Processors
  – Tasks

Preclass 1

• How big?
  – 2-LUT?
  – 2-LUT w/ Flip-flop?
  – 2-LUT w/ 4 input sources?
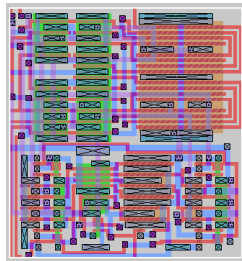 – 2-LUT w/ 1024 input sources?

1

Empirical Comparisons

Empirical

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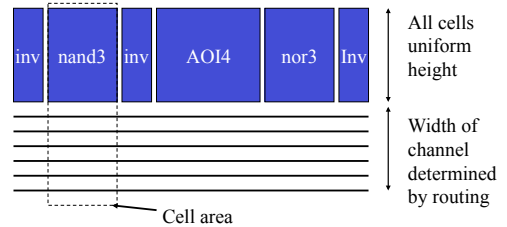- Start sorting out
  - custom vs. configurable
  - spatial configurable vs. temporal

Full Custom

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

- ESE570

Standard Cell Area

All cells uniform height

Width of channel determined by routing

Cell area

Standard Cell Area

All cells uniform height

Width of channel determined by routing
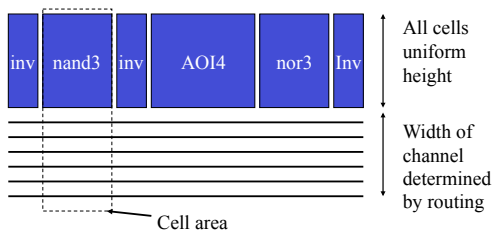
Cell area

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

Standard Cell Area

All cells uniform height

Width of channel determined by routing

Cell area
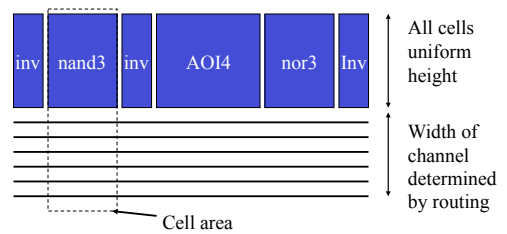
What freedom have we removed? Impact?

MPGA
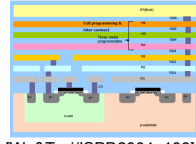
[Wu&Tsai/ISPD2004p103]

- Metal Programmable Gate Array
  – Resurrected as "Structured ASICs"
- Gates pre-placed (poly, diffusion)
- Only get to define metal connections
  – Cheap (low NRE)
    - only have to pay for metal mask(s)
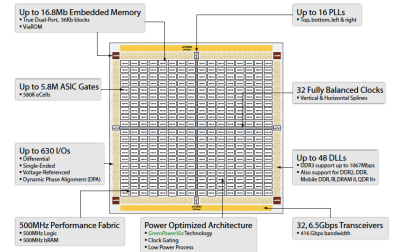
Structured ASIC: eASIC

- 2011
- 45nm
- 1.4M LUTs
- 500MHz?

http://www.easic.com/wp-content/uploads/2011/02/eASIC-Nextreme-2T-Product-Brief.pdf

Structured ASIC

- Maybe think about it as an FPGA with vias instead of configurable switches?
- Ratio of SRAM to via design?

200 of these
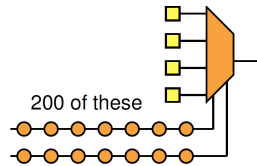
What do we expect?

- Comparing density/delay/energy
  – Full custom
  – Standard Cell (ASIC)
  – MPGA / Structured ASIC
  – FPGA
  – Processor

Why it isn't trivial?

- Different logic forms
- Interconnect
- Logic Balance
- Mix of Requirements

MPGA/SA vs. Custom?

- AMI CICC'83
  – MPGA 1.0
  – Std-Cell 0.7
  – Custom 0.5
- AMI CICC'04
  – Custom 0.6 (DSP)
  – Custom 0.8 (DPath)
- 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

Metal Programmable Gate Arrays

MPGAs

- Modern -- "Sea of Gates"
- yield 35--70%
- maybe $5k\lambda^2$/gate = $1.25F^2$/gate ?
  - (quite a bit of variance)
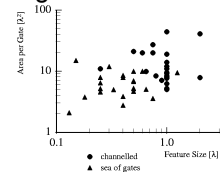
Conventional FPGA Tile

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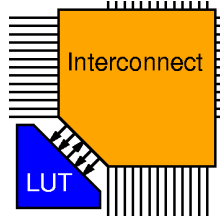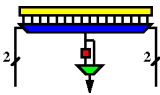
Interconnect

LUT

Toronto FPGA Model
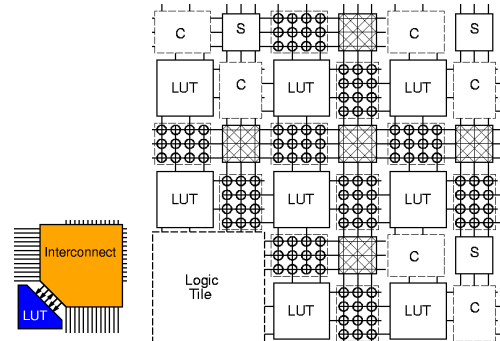
FPGA Table

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	--

(semi) 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
  - (~375KF^2/4-LUT)

[Both also have RAM in cited area]

4

How many gates?
(Prelcass 2)
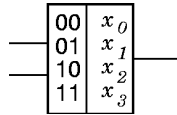
"gates" in 2-LUT
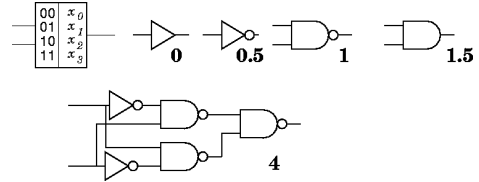
Now how many?
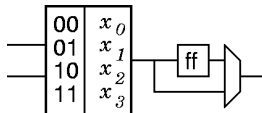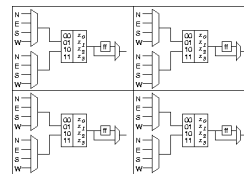
Which gives:
  More usable gates?

  More gates/unit area?
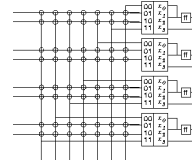
Gates Required?
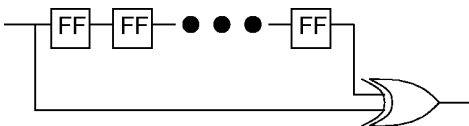
Depth=3, Depth=2048?

Gate metric for FPGAs?

- Day11: 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)
  - $5K\lambda^2$/gate
  - 35-70% usable (50%)
  - $7\text{-}17K\lambda^2$/gate net
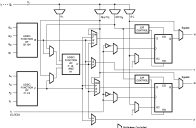
- Ratio: 2--10 (5)

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

- Xilinx XC4K
  - $1.25M\lambda^2$ /CLB
  - 17--48 gates (26?)
  - $26\text{-}73K\lambda^2$/gate net
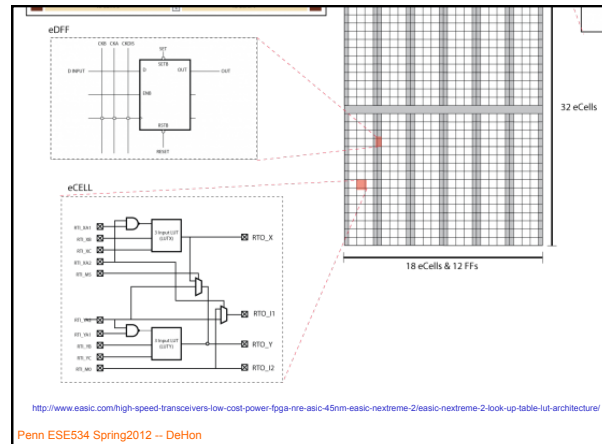
FPGA vs. Structure ASIC

- Virtex 6
- 40nm
- 470K 6-LUTs
- Largest device

- eASIC
- 45nm
- 580K eCells
- Probably smaller die

FPGA vs. Std Cell

- 90nm
- FPGA: Stratix II
- STMicro CMOS090
  - Standard Cell
    - Full custom layout
    - …but by tool

TABLE II
AREA RATIO (FPGA/ASIC)

Name	Logic Only	Logic & DSP	Logic & Memory	Logic, Memory & DSP
booth	33			
rs_encoder	32			
cordic18	19			
cordic8	25			
des_area	42			
des_perf	17			
fir_restruct	28			
mac1	43			
aes192	47			
fir3	45	17		
diffeq	41	12		
diffeq2	39	14		
molecular	47	36		
rs_decoder1	54	58		
rs_decoder2	41	37		
atm			70	
aes			24	
aes_inv			19	
ethernet			34	
serialproc			36	
fir24				9.5
pipe5proc				23
raytracer				26
Geomean	35	25	33	18

[Kuon/Rose TRCADv26n2p203--215 2007]

MPGA vs. FPGA (Delay)

- MPGA (SOG GA)

  $\lambda = 0.6\mu$

  $\tau_{gd} \sim 1ns$

- Ratio: 1--7 (2.5)
  - Altera claiming 2×
    - For their Structured ASIC [2007]
  - LSI claiming 3×
    - 2005

- Xilinx XC4K

  $\lambda = 0.6\mu$

  1-7 gates in 7ns

  2-3 gates typical

FPGA vs. Std. Cell Delay

- 90nm
- FPGA: Stratix II
- STMicro CMOS090

TABLE IV
CRITICAL-PATH-DELAY RATIO (FPGA/ASIC)—FASTEST SPEED G

Name	Logic Only	Logic & DSP	Logic & Memory	Logic, Memory & DSP
booth	5.0			
rs_encoder	3.8			
cordic18	3.7			
cordic8	1.9			
des_area	2.0			
des_perf	3.1			
fir_restruct	4.0			
mac1	3.8			
aes192	4.4			
fir3	3.9	3.5		
diffeq	4.0	4.1		
diffeq2	3.9	4.0		
molecular	4.6	4.7		
rs_decoder1	2.5	2.9		
rs_decoder2	2.2	2.4		
atm			2.9	
aes			3.8	
aes_inv			4.3	
ethernet			4.3	
serialproc			2.8	
fir24				2.6
pipe5proc				2.9
raytracer				3.5
Geomean	3.4	3.5	3.5	3.0

[Kuon/Rose TRCADv26n2p203--215 2007]

FPGA vs. Std Cell Energy

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			
mac1	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	
fir24	Const				5.3
pipe5proc	Const				8.2
raytracer	Const				8.3
Geomean		14	12	14	7.1

- 90nm
- FPGA: Stratix II
- STMicro CMOS090

- eASIC (MPGA) claim
  – 20% of FPGA power
  – (best case)

[Kuon/Rose TRCADv26n2p203--215 2007]

Processors vs. FPGAs

Processors and FPGAs

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

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

Component Example

- Single die in $0.35\mu m$

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	$\frac{ge's}{\lambda^2 s}$
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
    - FPGA:std-cell custom ~ 15-30x
- Area-Time
  – Gate Array 6-10x Custom
  – FPGA 15-20x Gate Array
    - FPGA:std-cell custom ~ 100x
  – 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

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 $\frac{mpy}{\lambda^2 s}$
Custom 16 $\times$16	0.63μm	2.6Mλ^2, 40 ns	9.6
Custom 8 $\times$8	0.80μm	3.3Mλ^2, 4.3 ns	
Gate-Array 16 $\times$16	0.75μm	26Mλ^2, 30ns	1.3
FPGA (XC4K)	0.60μm	1.25Mλ^2/CLB 316 CLBs, 26 ns 84 CLBs, 40 ns 220 CLBs, 12.1 ns 22 CLBs, 25 ns	0.097
16b DSP	0.65μm	350Mλ^2, 50 ns	0.057
RISC (no multiplier)	0.75μm	125Mλ^2, 66 ns/cycle two 16b operands - 44 cycles 16b constant - 7 cycles one 8b operand - 24 cycles 8b constant - 4 cycles	0.0028

Preclass 4

- Efficiency of 8×8 multiply on 16×16 multiplier?

Multiply

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

8

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)	$\frac{TAP_s}{\lambda^2 S}$
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)		

Mixed Designs

- Modern FPGAs include hardwired multipliers (Virtex 25x18)

Table 1: Virtex-6 FPGA Feature Summary by Device

Device	Logic Cells	Configurable Logic Blocks (CLBs)		DSP48E1 Slices[2]	Block RAM Blocks		
		Slices[1]	Max Distributed RAM (Kb)		18 Kb[3]	36 Kb	Max (Kb)
XC6VLX75T	74,496	11,640	1,045	288	312	156	5,616
XC6VLX130T	128,000	20,000	1,740	480	528	264	9,504
XC6VLX195T	199,680	31,200	3,040	640	688	344	12,384
XC6VLX240T	241,152	37,680	3,650	768	832	416	14,976
XC6VLX365T	364,032	56,880	4,130	576	832	416	14,976
XC6VLX550T	549,888	85,920	6,200	864	1,264	632	22,752
XC6VLX760	758,784	118,560	8,280	864	1,440	720	25,920
XC6VSX315T	314,880	49,200	5,090	1,344	1,408	704	25,344
XC6VSX475T	476,160	74,400	7,640	2,016	2,128	1,064	38,304

http://www.xilinx.com/products/silicon-devices/fpga/virtex-6/index.htm

FPGA vs. Std Cell (revisit)

- 90nm
- FPGA: Stratix II
- STMicro CMOS090

TABLE II
AREA RATIO (FPGA/ASIC)

| Name	Logic Only	Logic & DSP	Logic & Memory	Logic, Memory & DSP
booth	33			
rs_encoder	32			
cordic18	19			
cordic8	25			
des_area	42			
des_perf	17			
fir_restruct	28			
mac1	43			
aes192	47			
fir3	45	17		
diffeq	41	12		
diffeq2	39	14		
molecular	47	36		
rs_decoder1	54	58		
rs_decoder2	41	37		
atm			70	
aes			24	
aes_inv			19	
ethernet			34	
serialproc			36	
fir24				9.5
pipe5proc				23
raytracer				26
Geomean	35	25	33	18

[Kuon/Rose TRCADv26n2p203--215 2007]

Energy

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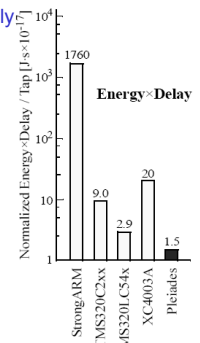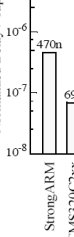
[Abnous et al, *The Application of Programmable DSPs in Mobile Communications*, Wiley, 2002, pp. 327-360]

FPGA vs. Std Cell Energy (revisit)

- 90nm
- FPGA: Stratix II
- STMicro CMOS090

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			
mac1	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	
fir24	Const				5.3
pipe5proc	Const				8.2
raytracer	Const				8.3
Geomean		14	12	14	7.1

[Kuon/Rose TRCADv26n2p203--215 2007]

IIR/Biquad (Infinite Impulse Response)

Architecture	Feature Size (λ)	Area and Time	16b TAPs $\lambda^2 S$	10b 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: $Y_i = A \times X_i + B \times Y_{i-1}$

9

DES Keysearch

Architecture	Feature Size (λ)	Area	Keys/Second	$\frac{\text{Keys}}{\lambda^2 \text{s}}$
DES IC	1.5μm	$11.1M\lambda^2$	310K	0.028
FPGA (Altera 8K)	0.30μm	81188 ($930M\lambda^2$)	800K	0.00086
RISC	0.30μm	$1.8G\lambda^2$	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	$\frac{\text{cu}}{\lambda^2 \text{s}}$
Custom FPGA	2.0μm	$270M\lambda^2$	500M	1.9
(SPLASH 2)	0.60μm	$43G\lambda^2$	3,000M	0.070
(SPLASH)	0.60μm	$33G\lambda^2$	370M	0.012
RISC (SparcStation I)	0.75μm	$273M\lambda^2$	0.87M	0.0032
(SparcStation 10)	0.40μm	$1.6G\lambda^2$	1.2M	0.00075

N.B. includes memory area for SPLASH

Degrade from Peak

How do various architecture degrade from peak?

- FPGA?
- Processor?
- Custom?

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

Admin

- Grading through 6.1-2 done during break
  - Check feedback
- HW6.3-4 due Wednesday
- HW7 out
- Wednesday Reading on web
  - Classic Paper

Big Ideas [MSB Ideas]

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