CS294-7: Reconfigurable Computing

Day 7: Tuesday, February 11

- Admin
- Comparing Computing Devices

Admin

- CAD Tools for Project 1:
-
- Project 1 Proposals due this Thursday (next class)

Quotes

An Engineer is someone who can do for a dime what everyone else can do for a dollar.

– Petroski (paraphrased)

If it can't be expressed in figures, it is not science; it is opinion.

– Lazarus Long (Heinlein)

Motivation

We want to know:

- How big (costly) a particular solution is
- How it compares to alternatives
- What do I pay/gain for flexibility

Since we generally have a time-space tradeoff, size alone is seldom sufficient:

- What performance achievable at size X?
- What is the performance density?
- What does the area-time curve look like for this device/application?

Difficulty

In lectures 2 and 3 we saw that the balance among

1. compute elements
2. spatial interconnect
3. temporal interconnect
4. instructions

...can vary among architectures and computational tasks.

In lecture 2, we saw that this made it hard to assign a single metric (*i.e.* gate count) to tasks or devices.

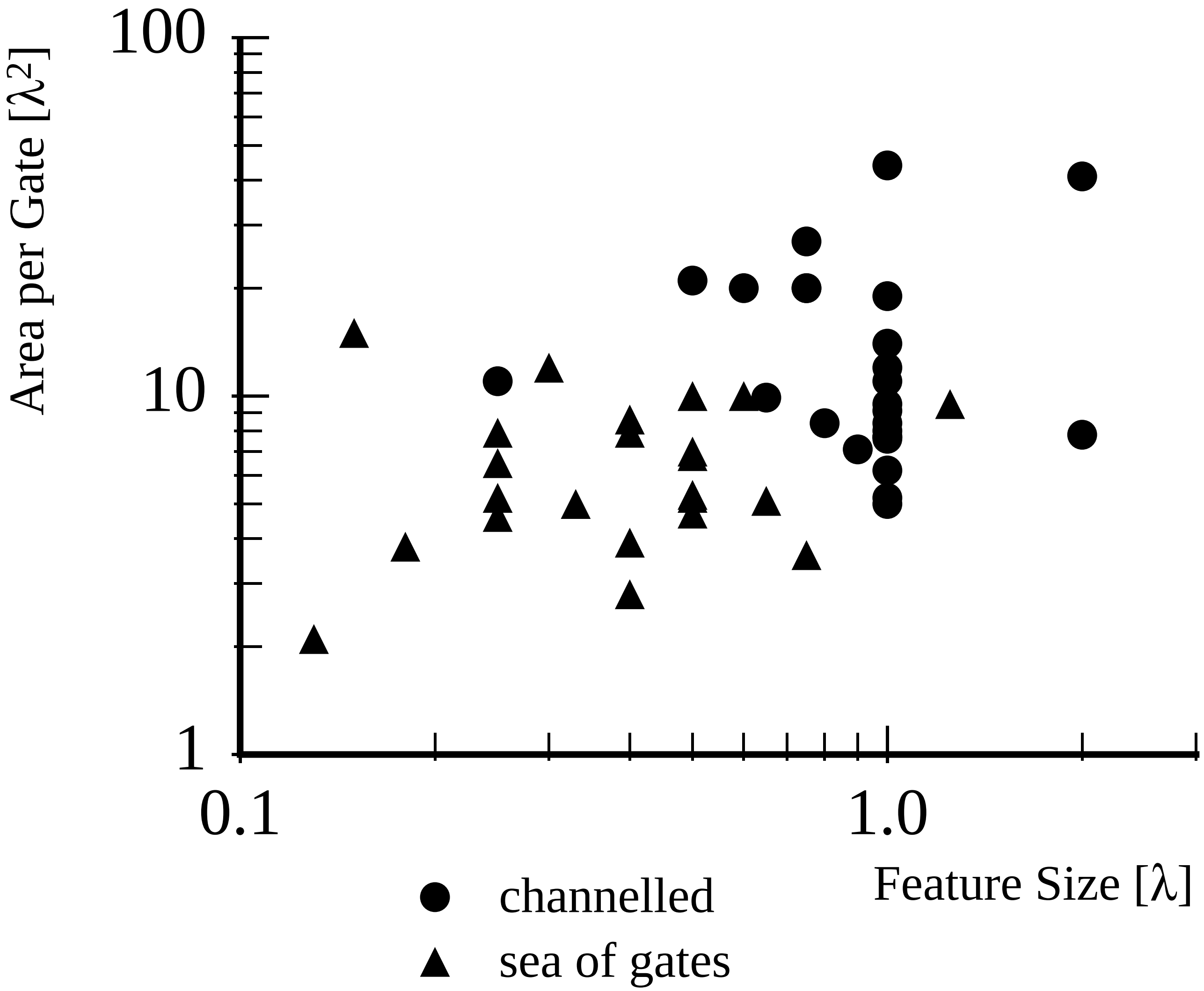
...nonetheless, worthwhile to try to understand when a device style might be smaller/denser/faster.

Computation Limiting – Gate Evaluations and Bit-ops

1. Metal-masked Gate array density
   (a) relate to custom
   (b) relate to FPGAs

2. FPGA compute density
   (a) relate to general-purpose portion of processors
   (b) how both degrade from peak

Metal-Programmable Gate Array Density



Sea-of-gates: $\approx$5Kλ^2, 35-70% usable?

MPGA versus Custom (anecdotal)

Area:

	AMI CICC'83	GE CICC'86
Full	0.5	
std-cell	0.7	0.4-0.7
MPGA	1.0	1.0

	Function	std-cell	Full
GE	FF/counter	0.7	
	FA	0.4	
	RAM	0.2	
Toshiba	DSP		0.3
EPFL	counter/mpy/adder		0.5
MOSAID	memory		0.2

MPGA versus FPGAs

Area:

SOG GA	5Kλ^2/gate	35-70% usable (50%)	7-17K	10K
XC4K	1.25Mλ^2/CLB	17-48 gates (26 gates)	26-73K	48K
ratio			2-10	5

Speed: (@ λ=0.6μ)

GA	$\tau_{gd} \approx 1$ns
XC4K	1-7 gates in 7ns
ratio	1-7 (2.5)

Processors and FPGAs: Computational Density

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

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

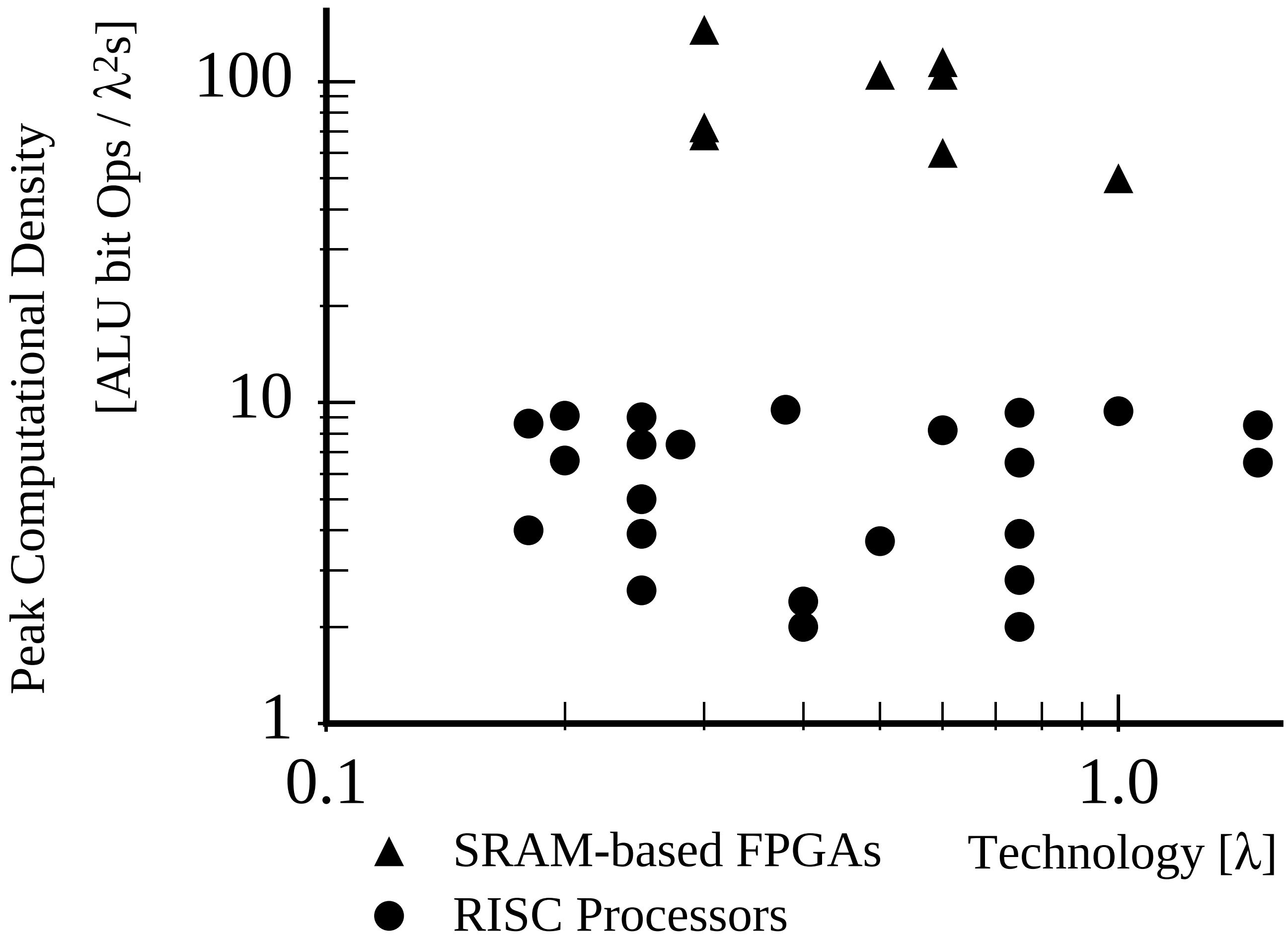
Caveats:

- Ignores special-purpose coprocessors (*e.g.* multipliers, FPUs)
- Assumes adequate interconnect
- Does not metric diversity

Raw Computational Densities

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

Proc/FPGA GP Density



Processor Degradation from Peak

- Operations do only data movement
- Operations not use full word width
- Stalls waiting for retimed data

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

FPGA Degradation from Peak

- Long path length
- Limited throughput requirement
  (bottlenecks elsewhere limit throughput,
  or need for highly diverse computational functions)
- Insufficient interconnect
- Insufficient retiming resources

Custom/MPGA Degradation from Peak

- Solve more general problem than required at a particular time
- Long path length
- Limited throughput requirement (bottlenecks elsewhere limit throughput need for highly diverse computational functions)
- (not needed / applicable to problem)

Yielded Throughput Density

As a complement to looking at raw density,
we can look at relative densities on specific applications.

Applications represent a particular resource mix
— PDF of characteristics:

- path length
- data width
- interconnect richness
- diversity

Throughput Density Metric

Figure of Merit when:

- throughput requirement is high
- available concurrency is high
- easy scaling through replication

E.g. Solution with highest value:

- will satisfy a throughput target with least area
- will provide most throughput out of a fixed area

$$\text{Throughput Density} = \frac{\text{Results}}{\text{Area} \times \text{Time}}$$

Multiply

Architecture	Feature Size (λ)	Area and Time	16×16 mpy/λ^2s	16×16 scale/λ^2s	8×8 mpy/λ^2s	8×8 scale/λ^2s
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

FIR – 8b sample, 8b coefficient

Architecture	Feature Size (λ)	Area and Time	$\frac{TAPs}{\lambda^2\mathbf{s}}$
32b RISC	0.75μm	125Mλ^2, 66 ns/cycle$\times$6+cycles/TAP	0.020
16b DSP	0.65μm	350Mλ^2, 50 ns/TAP	0.057
32b RISC/DSP	0.25μm	1.2Gλ^2, 40 ns/TAP	0.021
64b RISC	0.18μm	6.8Gλ^2, 2.3 ns/TAP	0.064
FPGA (XC4K)	0.60μm	240 CLBs, 14.3 ns/8-TAPs	1.9
(Altera 8K)	0.30μm	30 LEs$\times$0.92Mλ^2/LE, 10 ns/TAP	3.6
Full Custom	0.75μm	400Mλ^2, 45 ns/64 TAPs	3.6
	0.60μm	140Mλ^2, 33 ns/16 TAPs	3.5
	0.75μm	82Mλ^2, 50 ns/10 TAPs	2.4
(fixed coefficient)	0.60μm	114Mλ^2, 6.7 ns/43 TAPs	56
		(*n.b.* 16b samples)	

IIR

Architecture	Feature Size (λ)	Area and Time	16b $\frac{\text{TAPs}}{\lambda^2 \text{s}}$	10b $\frac{\text{TAPs}}{\lambda^2 \text{s}}$
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

DES Key Search

Architecture	Feature Size (λ)	Area	Keys/Second	$\frac{\text{Keys}}{\lambda^2 \text{s}}$
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/>

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 precision)

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

Floating Point Multiply (single precision)

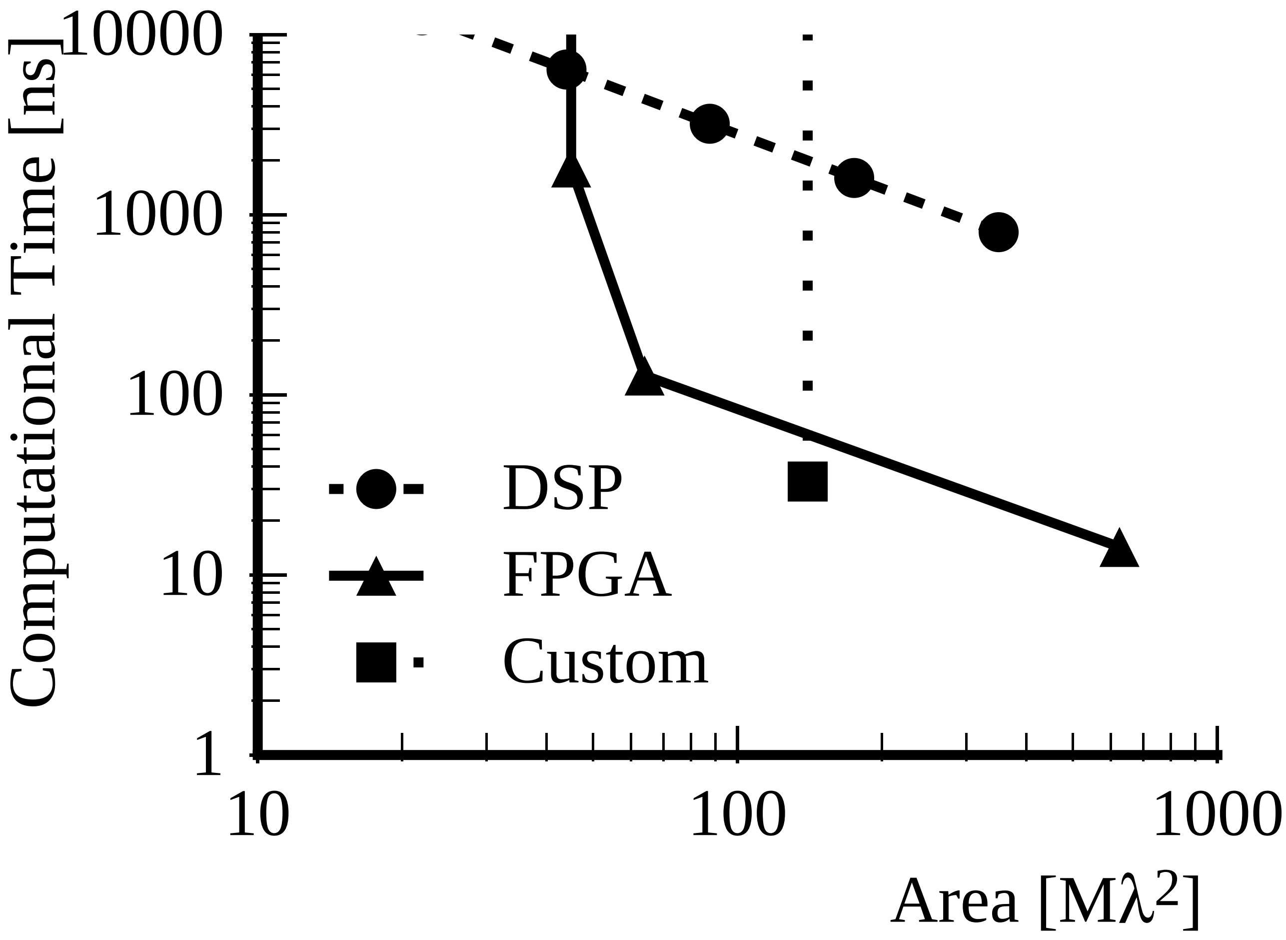
Architecture	Reference	λ	area and time	$\frac{\text{32b FP MPYs}}{\lambda^2 \cdot s}$
Custom Macrocells				
32b mpy	JSSC92	0.6μ	$43M\lambda^2$, 30 ns	0.78
32b mpy	EUROASIC92	0.4μ	$81M\lambda^2$, 42 ns	0.29
32/64b FPU	CICC92	0.25μ	$980M\lambda^2$, 2 per 12.5 ns	0.16
Coprocessor				
dedicated 32b mpy	JSSC84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

Other Figures of Merit

- area required to achieve low, fixed throughput rate
- performance achievable with fixed area budget
- performance achievable with fixed energy budget

...complicated by non-idealness of area-time curves...

Area-Time Curve for 16-TAP, 8-bit FIR



Area-Time Curve and DES Encryption

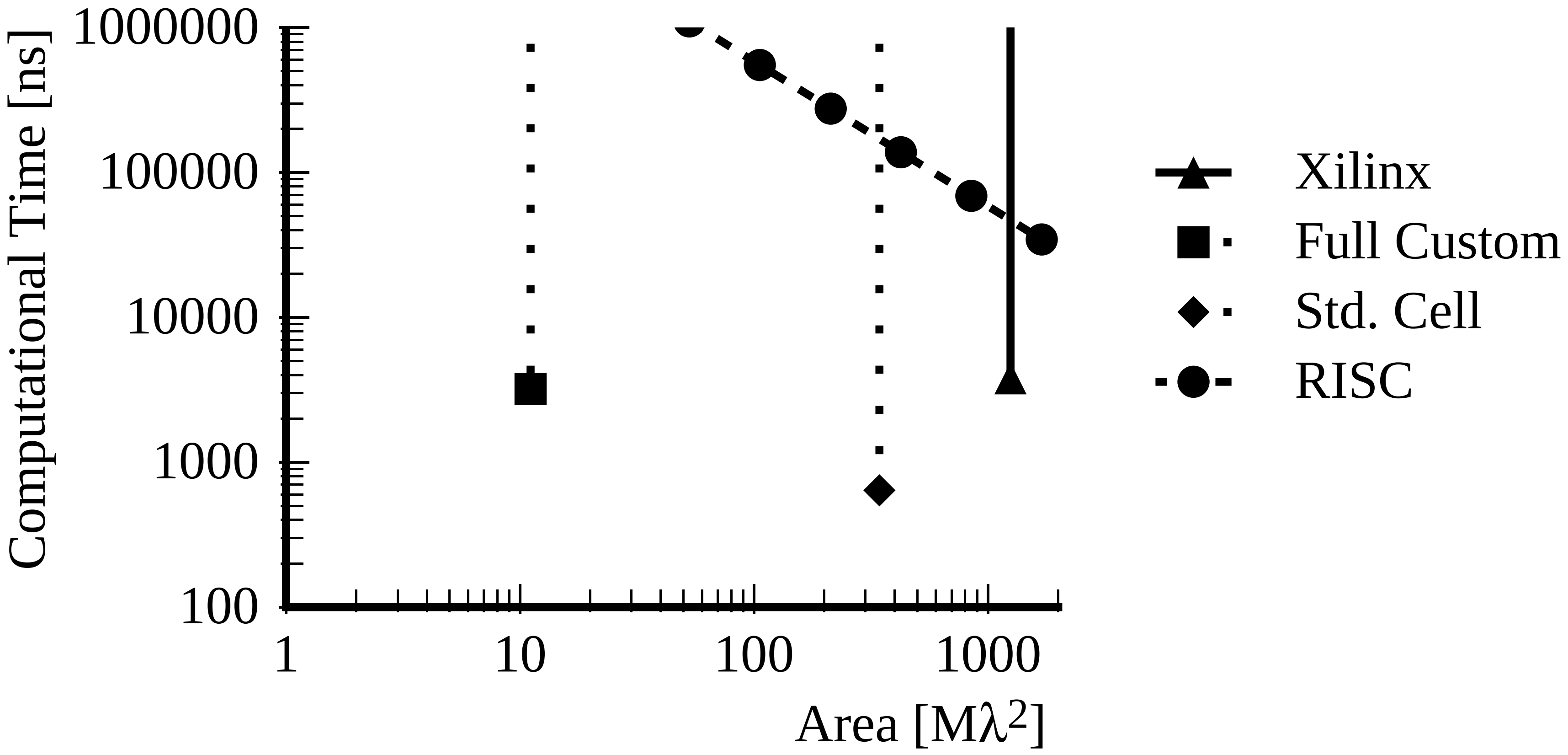


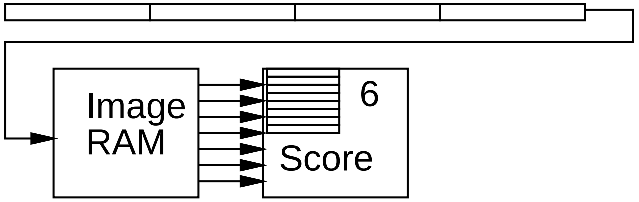
Image Scoring

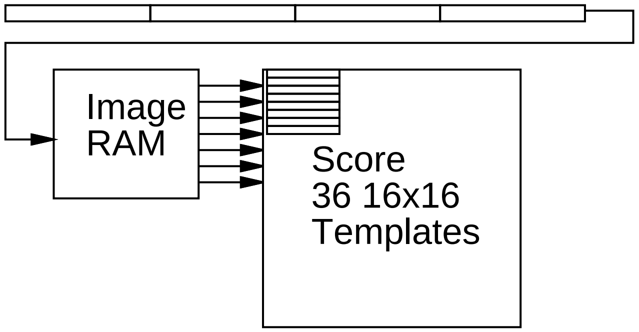
Task:

- Inside a 64×64 image window
- Score n 16×16 image templates

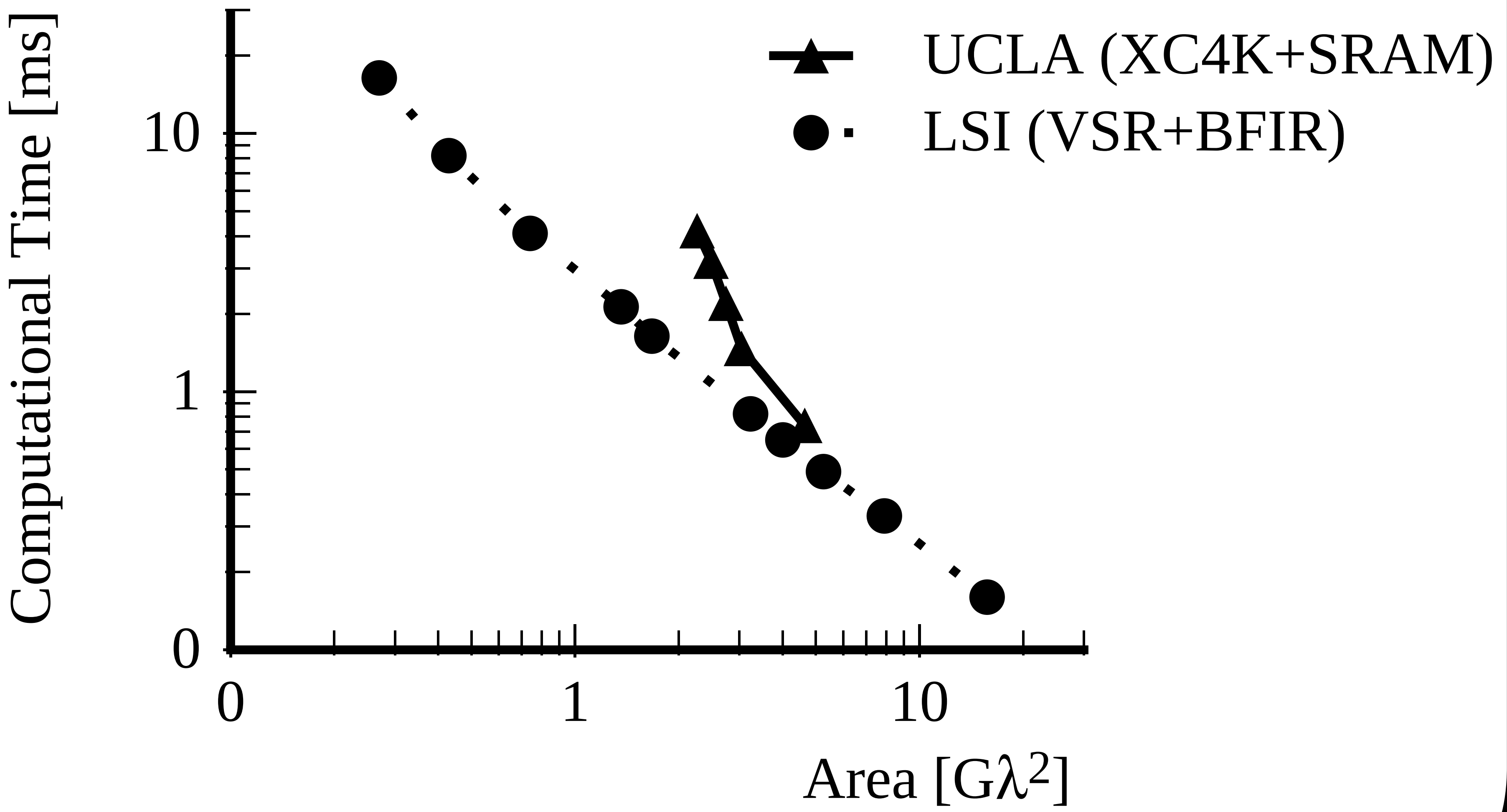
UCLA

- 16 $32K \times 8$ SRAMs ($1.76G\lambda^2$) (coarse retime)
- 150 CLBs shift/retime active pixels
- $\approx$ 50CLBs/template (average)





Area-Time Curve for Image Scoring



Common Non-idealities in Area-Time

1. **Fixed component:**

$$A(T) = A_f + \left(\frac{1}{T}\right) A_v$$

2. **Critical Path**

$$T = \mathrm{max}(\frac{\mathrm{Ops}}{A}, \mathrm{Path\ Length})$$

3. **Interconnect Requirements Grow w/ Parallelism**

$$\begin{aligned} A &= C \cdot N^{2p} \\ N &= N_f + \left(\frac{1}{T}\right) N_v \end{aligned}$$

$$N_f = 0 \rightarrow AT^{2p} = C$$

Highlights

- Area
  - Gate Array 2-3$\times$ Custom
  - FPGA 5$\times$ Gate Array
- Area-Time
  - Gate Array 6-10$\times$ Custom
  - FPGA 15-20$\times$ Gate Array
  - Processors 10+$\times$ FPGA
- Processor FPGA may solve more specialized problem than GA/custom
- Non-idealness in Area-Time curves

Much left to map out

...there is quite a bit still to be understood and mapped out here.

While the primary goal of project 1 is education, may contribute more datapoints.

There are a number of good projects (project 2, theses...) advancing our understanding of these issues.