CS184a:
Computer Architecture
(Structures and Organization)

Day20: November 29, 2000

Review

Today

- Review content and themes
- *N.B.* EOT Feedback Questionnaire
  - return end of class in basket
  - or later to Cynthia (256 JRG)

Physical Implementation of Computation: Engineering Problem

- Implement a computation:
  - with least resources (in fixed resources)
    - with least cost
  - in least time (in fixed time)
  - with least energy
- Optimization problem
  - how do we do it best

Architecture Not Done

- Not here to just teach you the forms which are already understood
  - (though, will do that and give you a strong understanding of their strengths and weaknesses)
- **Goal**: enable you to design and synthesize new and better architectures
- Engineering not Biology

Authority/History

- ``Science is the belief in the ignorance of experts." -- Richard Feynman
- **Goal**: Teach you to think critically and independently about computer design.

Content Overview

- This quarter:
  - building blocks and organization
  - raw components and their consequences
- Next two quarters:
  - abstractions, models, techniques, systems
  - *e.g.* ISA, Control and Data Flow, caching, VM, processor pipeline, branching, renaming….RISC, VLIW, SuperScalar, Vector, SIMD, ….

Content (this quarter)

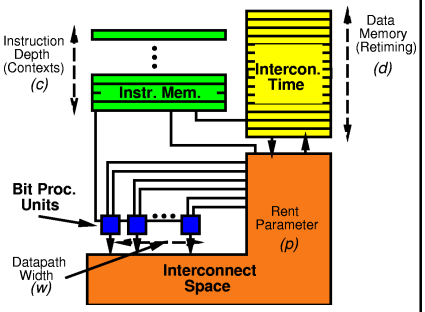
- Requirements of Computation
- Key components:
  - Instructions
  - Interconnect
  - Compute
  - Retiming
  - Control

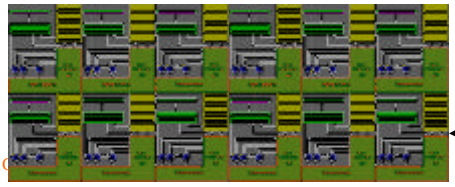
Themes (this quarter)

- Implementation techniques
- Costs
- Structure in Computations
- Design Space
  - identify and model
- Parameterization
- Metrics and Figures of Merit
- Tradeoffs, analysis
- Change

Computing Device

- Composition
  - Bit Processing elements
  - Interconnect: space
  - Interconnect: time
  - Instruction Memory

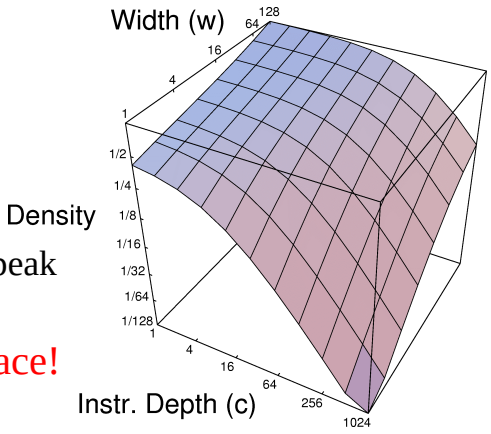




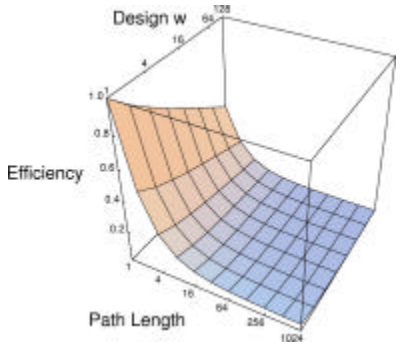
Tile together to build device

Peak Computational Densities from Model

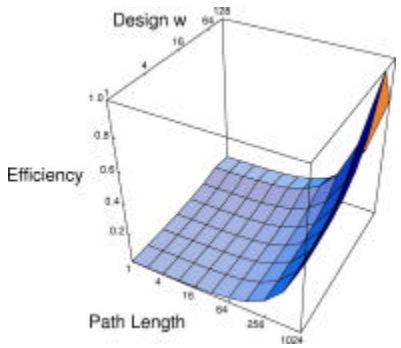
- Small slice of space
  - only 2 parameters
- 100× density across
- Large difference in peak densities
  - large design space!



Yielded Efficiency



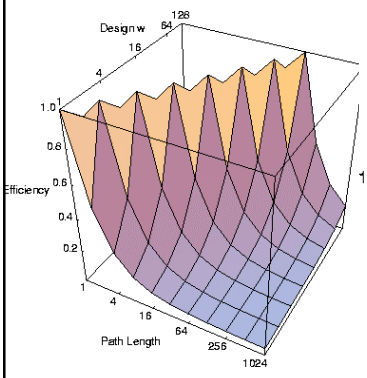
FPGA (c=w=1)

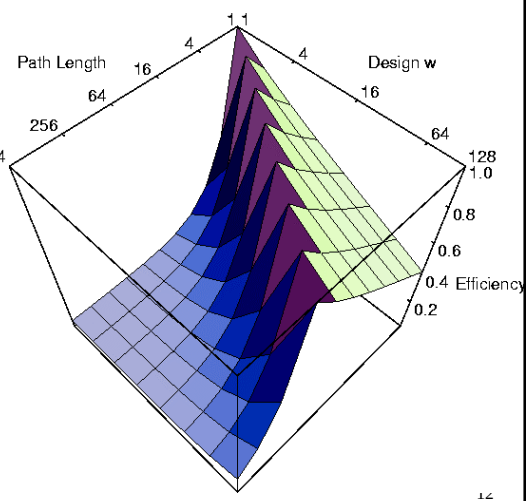


"Processor" (c=1024, w=64)

- Large variation in **yielded** density
  - large design space!

Throughput Yield





Same graph, rotated to show backside.

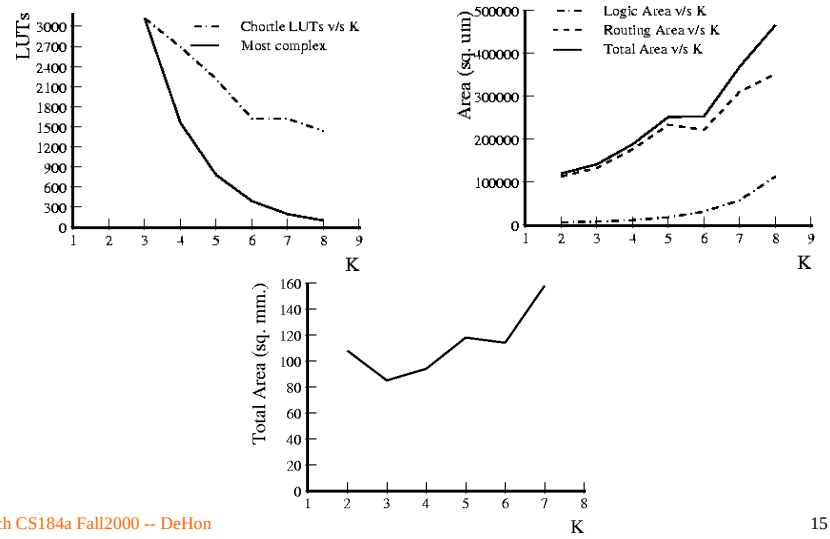
Architecture Instr. Taxonomy

Control Threads (PCs)	*pinsts* per Control Thread	Instruction Depth	Granularity	Architecture/Examples
0	0	0	n/a	Hardwired Functional Unit (*e.g.* ECC/EDC Unit, FP MPY)
			1	FPGA
	n	1	w	Reconfigurable ALUs
1	1	c	$n_v \cdot 1$	Bitwise SIMD
			w	Traditional Processors
			$n_v \cdot w$	Vector Processors
	n	c	1	DPGA
		8	16	PADDI
		c	w	VLIW
m	n	1	1	HSRA/SCORE
	1	c	$n_v \cdot w$	MSIMD
m	1	c	1	VEGA
		8	16	PADDI-2
		c	w	MIMD (traditional)

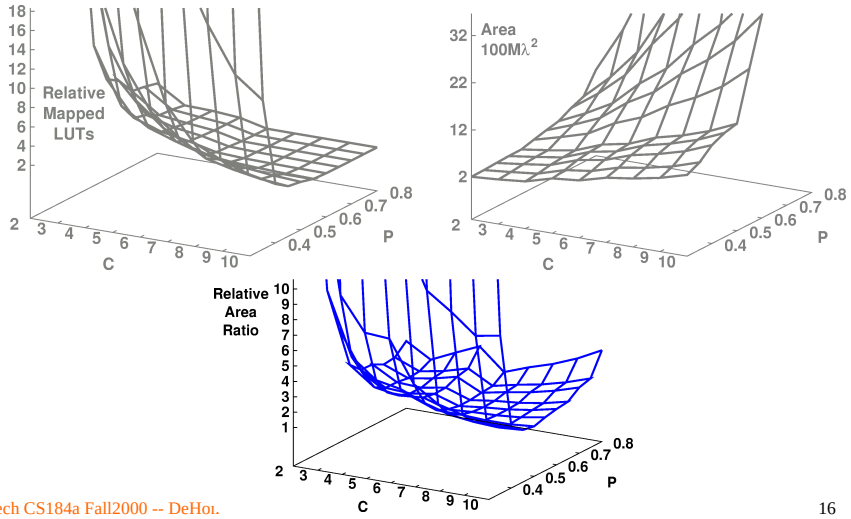
Methodology

- Architecture model (parameterized)
- Cost model
- Important task characteristics
- Mapping Algorithm
  - Map to determine resources
- Apply cost model
- Digest results
  - find optimum (multiple?)
  - understand conflicts (avoidable?)

Mapped LUT Area

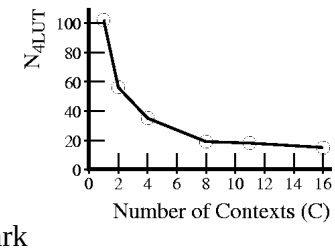


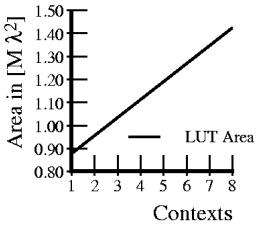
Resources × Area Model ⇒ Area

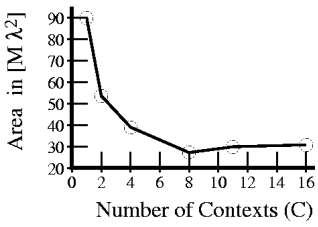


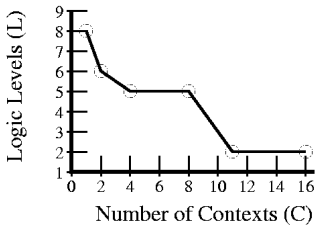
Control: Partitioning versus Contexts (Area)

CSE benchmark









Design Space

- Mindset
- Methodology
- Decomposition
  - fundamental building blocks
  - basis set
- Build Intuition on Space
  - grounded in quantifiable instances

Change

- A key feature of the computer industry has been rapid and continual change.
- We must be prepared to adapt.
- For our substrate:
  - capacity (orders of magnitude more)
    - what can put on die, parallelism, need for interconnect and virtualization, homogeneity
  - speed
  - relative delay of interconnect and gates

Fountainhead Parthenon Quote

"Look," said Roark. "The famous flutings on the famous columns---what are they there for? To hide the joints in wood---when columns were made of wood, only these aren't, they're marble. The triglyphs, what are they? Wood. Wooden beams, the way they had to be laid when people began to build wooden shacks. Your Greeks took marble and they made copies of their wooden structures out of it, because others had done it that way. Then your masters of the Renaissance came along and made copies in plaster of copies in marble of copies in wood. Now here we are making copies in steel and concrete of copies in plaster of copies in marble of copies in wood. Why?"

What About Computer Architecture?

Are we making copies in submicron CMOS VLSI of copies in NMOS of copies in TTL of early vacuum tube computer designs?

Mainframe->Mini->super microprocessors ?

CDC->Cray1->i860->Vector microprocessors?

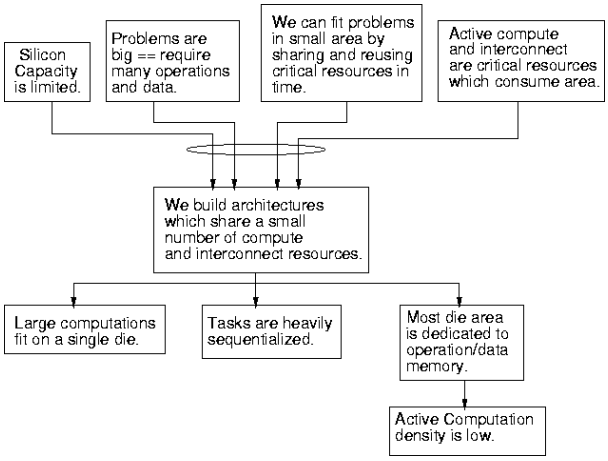
1983 Computer Architecture

- VLSI is “new” to the computer architect
- you have 15Mλ^2 in 4μm NMOS
- want to run “all” programs
- What do you build?

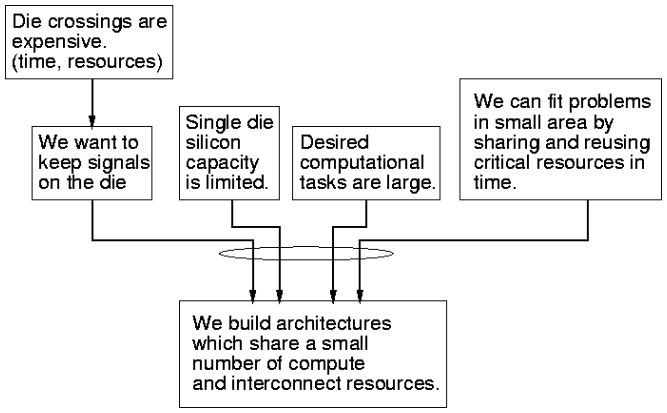
What can we build in 15Mλ^2?

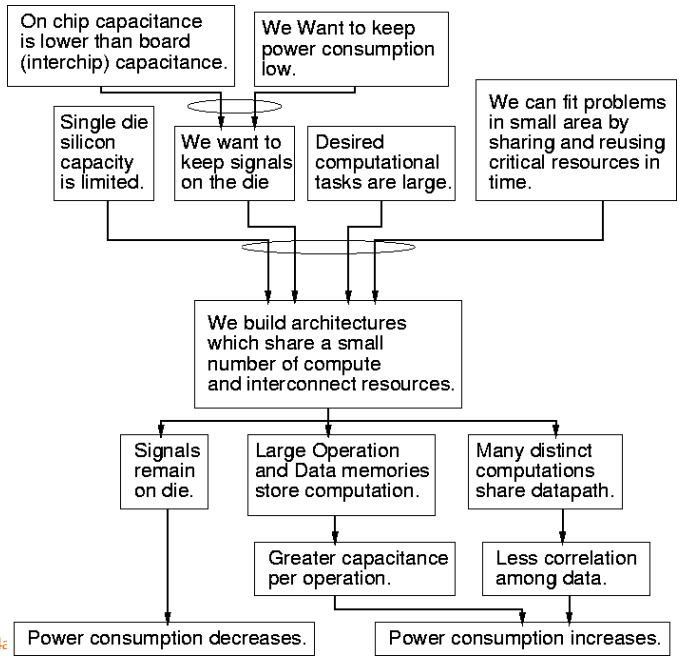
- 12Kb SRAM (1.2Kλ^2/bit)
- 1500 Gate-Array Gates (10Kλ^2/gate)
- 30 4-LUTs (500Kλ^2/4LUT)
- 32b ALU+RF+control

What…1983?



More Why?


Die crossings are expensive. (time, resources)
We want to keep signals on the die
Single die silicon capacity is limited.
Desired computational tasks are large.
We can fit problems in small area by sharing and reusing critical resources in time.
We build architectures which share a small number of compute and interconnect resources.

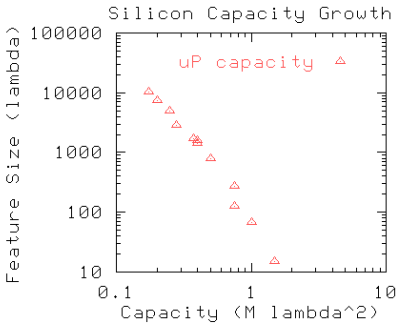

On chip capacitance is lower than board (interchip) capacitance.
We Want to keep power consumption low.
Single die silicon capacity is limited.
We want to keep signals on the die
Desired computational tasks are large.
We can fit problems in small area by sharing and reusing critical resources in time.
We build architectures which share a small number of compute and interconnect resources.
Signals remain on die.
Large Operation and Data memories store computation.
Many distinct computations share datapath.
Greater capacitance per operation.
Less correlation among data.
Power consumption decreases.
Power consumption increases.

1983

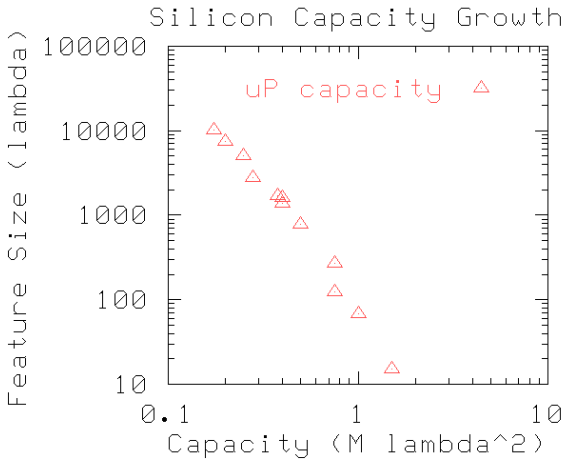
- RISC II
- MIPs

What has changed in 17 years?

- Technology (0.18µm CMOS)
- Capacity (50Gλ^2)
- Architecture?



Capacity



Architecture (last 17 years)

- Moved memory system on chip
- 32->64b datapath
- +FPU, moved on chip
- 1->4 or 8 compute units
- …lots of “hacks” to preserve sequential model of original uP

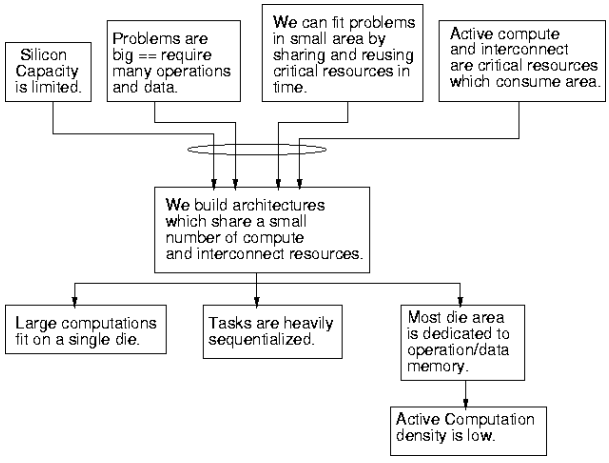
Have our assumptions changed?

- Beware of cached answers.
- Always check your assumptions.

To stay young requires unceasing cultivation of the ability to unlearn old falsehoods.

-- Lazarus Long

1983 Design Landscape



Should we still build computers the way we did in 1983?

Yesterday's solution becomes today's historical curiosity.

-- Goldratt

Example

- HP PA-RISC8500 (Hot Chips X)
- SPEC fits in on-chip cache
- What next?
- Does it make sense to keep this architecture and balance as capacity continues to grow?
- Hopefully, this class has given you some ideas of what else you could do with $100+G\lambda^2$
- …continue with next quarter...

Also Ask...

- What happened in early 1980's to make RISC possible / the right answer?
  - Compared to 70's ?

What do I want?

- Develop systematic design
- Parameterize design space
  - adapt to costs
- Understand/capture req. of computing
- Efficiency metrics
  - (similar to information theory?)
  - [related to last time: how much really need to compute]

Big Ideas

- Matter Computes
- Efficiency of architectures varies widely
- Computation design is an engineering discipline
- Costs change ⇒ Best solutions (architectures) change
- Learn to cut through hype
  - analyze, think, critique, synthesize

Big Ideas

- Design Space
- Effects of organization:
  - Instructions
  - Interconnect
  - Compute Block
  - Retiming
  - Control
- Key components of computing device