CS184a: Computer Architecture (Structures and Organization)

Day6: October 11, 2000

Instruction Taxonomy

VLSI Scaling

Last Time

- Computing requirements
- Instruction requirements
- Structure

Today

- Instruction Taxonomy
- VLSI Scaling

Instruction Distribution

- Beyond 64 PE, instruction bandwidth dictates PE size

$$\frac{\sqrt{PE_{area}} \times 4 \times \sqrt{N}}{(64 \times 8\lambda)} = N$$

$$PE_{area} = 16K\lambda^2 \times N$$

- Build larger arrays

  $\Rightarrow$processing elements become less dense

Instruction Memory Requirements

- **Idea:** put instruction memory in array
- **Problem:** Instruction memory can quickly dominate area, too
  - Memory Area = $64 \times 1.2K\lambda^2$/instruction
  - $PE_{area} = 1M\lambda^2 + (\text{Instructions}) \times 80K\lambda^2$

Instruction Pragmatics

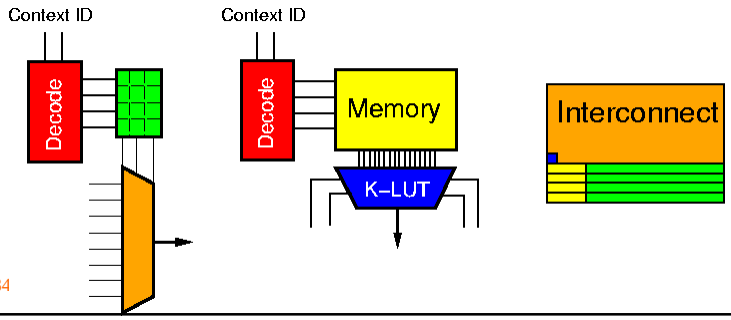
- Instruction requirements *could* dominate array size.
- Standard architecture trick:
  - Look for structure to exploit in "typical computations"

Two Extremes

- SIMD Array (microprocessors)
  - Instruction/cycle
  - share instruction across array of PE s
  - uniform operation in space
  - operation variance in time
- FPGA
  - Instruction/PE
  - assume temporal locality of instructions (same)
  - operation variance in space
  - uniform operations in time

Hybrids

- VLIW (SuperScalar)
  - Few *pinsts*/cycle
  - Share instruction across *w* bits
- DPGA
  - Small instruction store / PE



Architecture Instruction 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)
	n	1	1	FPGA
			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)

Instruction Message

- Architectures fall out of:
  - general model too expensive
  - look for structure in common problems
  - exploit structure to reduce resource requirements
- Architectures can be viewed in a unified design space

VLSI Scaling

Why Care?

- In this game, we must be able to predict the future
- Rapid technology advance
- Reason about changes and trends
- re-evaluate prior solutions given technology at time X.

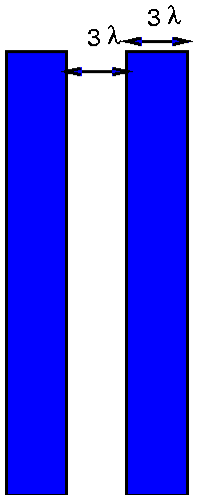
Why Care

- Cannot compare against what competitor does today
  - but what they can do at time you can ship
- Careful not to fall off curve
  - lose out to someone who can stay on curve

Scaling

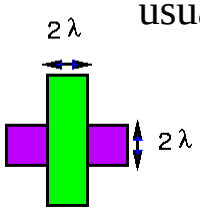
- **Premise:** features scale "uniformly"
  - everything gets better in a predictable manner
- **Parameters:**
  - λ (lambda) -- Mead and Conway (class)
  - S -- Bohr
  - 1/κ -- Dennard

Feature Size



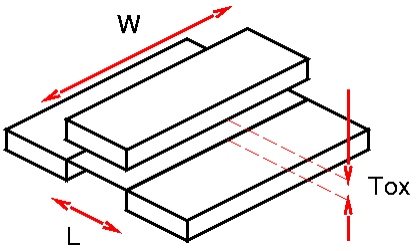
λ is half the minimum feature size in a VLSI process

[minimum feature usually channel width]



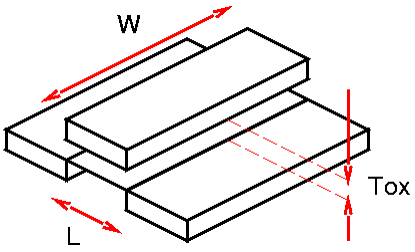
Scaling

- Channel Length (L)
- Channel Width (W)
- Oxide Thickness (T_{ox})
- Doping (N_a)
- Voltage (V)



Scaling

- Channel Length (L) λ
- Channel Width (W) λ
- Oxide Thickness (T_{ox}) λ
- Doping (N_a) $1/\lambda$
- Voltage (V) λ



Effects?

- Area
- Capacitance
- Resistance
- Threshold (V_{th})
- Current (I_d)
- Gate Delay (τ_{gd})
- Wire Delay (τ_{wire})
- Power

Area

$\lambda \rightarrow \lambda/\kappa$

$A = L * W$

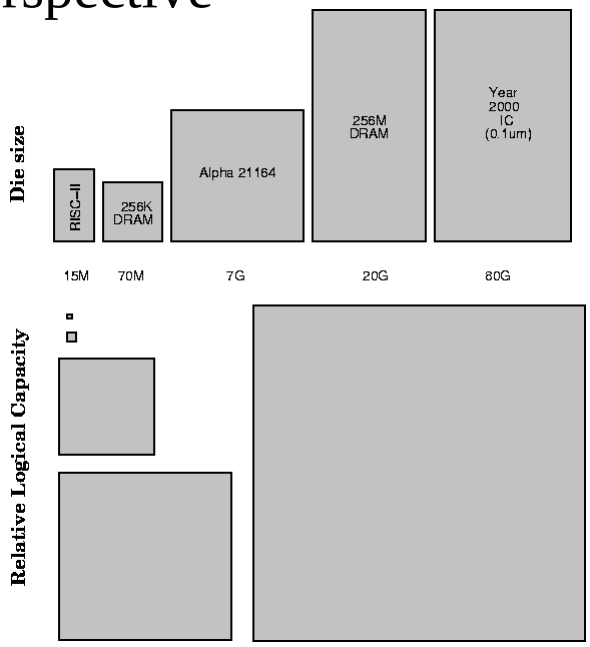
$A \rightarrow A/\kappa^2$

$0.35\mu m \rightarrow 0.25\mu m$

- 50% area
- 2x capacity same area

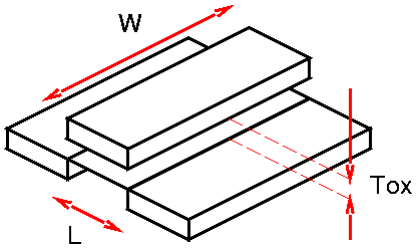
Area Perspective

[2000 tech.]
18mm×18mm
0.18µm
60G λ^2



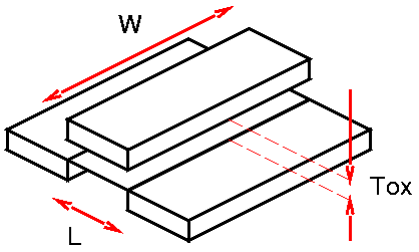
Capacitance

- Capacitance per unit area
  - $C_{ox} = \varepsilon_{SiO_2}/T_{ox}$
  - $T_{ox} \rightarrow T_{ox}/\kappa$
  - $C_{ox} \rightarrow \kappa C_{ox}$



Capacitance

- Gate Capacitance
  - $C_{gate} = A * C_{ox}$
    $A \rightarrow A/\kappa^2$
  - $C_{ox} \rightarrow \kappa C_{ox}$
  - $C_{gate} \rightarrow C_{gate}/\kappa$



Threshold Voltage

Before:

$$V_{th} = \frac{1}{C_{OX}}\left(-Q_{eff} + \left(2\epsilon_{Si} q N_a \left(\phi_s + V_{s\text{-}sub}\right)\right)^{1/2}\right) + \left(W_f + \phi_s\right)$$

$$\left(W_f + \phi_s\right) \approx 0$$

adjust $V_{s\text{-}sub}$ **so** $\left(\phi_s + V_{s\text{-}sub}\right) \rightarrow \frac{\left(\phi_s + V_{s\text{-}sub}\right)}{\kappa}$

After:

$$V'_{th} = \frac{1}{\kappa C_{OX}}\left(-Q_{eff} + \left(2\epsilon_{Si} q \kappa N_a \frac{\left(\phi_s + V_{s\text{-}sub}\right)}{\kappa}\right)^{1/2}\right)$$

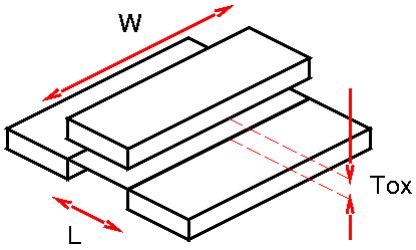
$$V'_{th} \approx \frac{V_{th}}{\kappa}$$

Threshold Voltage

- $V_{TH} \rightarrow V_{TH}/\kappa$

Current

- Saturation Current
  - $I_d=(\mu C_{OX}/2)(W/L)(V_{gs}-V_{TH})^2$
  - $V_{gs=}V \rightarrow V/\kappa$
  - $V_{TH} \rightarrow V_{TH}/\kappa$
  - $W \rightarrow W/\kappa$
  - $C_{ox} \rightarrow \kappa C_{ox}$
  - $I_d \rightarrow I_d/\kappa$

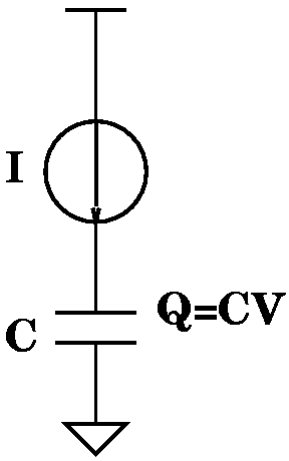


Gate Delay

$\tau_{gd}=Q/I=(CV)/I$

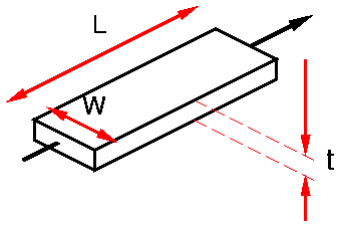
- $V \rightarrow V/\kappa$
- $I_d \rightarrow I_d/\kappa$
- $C \rightarrow C/\kappa$

$\tau_{gd} \rightarrow \tau_{gd}/\kappa$



Resistance

- $R=\rho L/(W*t)$
- $W \rightarrow W/\kappa$
- L, t similar
- $R \rightarrow \kappa R$



Wire Delay

$\tau_{wire}=R_L C$

- $R \rightarrow \kappa R$
- $C \rightarrow C/\kappa$

$\tau_{wire} \rightarrow \tau_{wire}$

- …assuming (logical) wire lengths remain constant...

Power Dissipation (Static)

- Resistive Power
  - $P=V*I$
  - $V \rightarrow V/\kappa$
  - $I_d \rightarrow I_d/\kappa$
  - $P \rightarrow P/\kappa^2$

Power Dissipation (Dynamic)

- Capacitive (Dis)charging
  - $P=(1/2)CV^2f$
  - $V \rightarrow V/\kappa$
  - $C \rightarrow C/\kappa$
  - $P \rightarrow P/\kappa^3$
- Increase Frequency?
  - $f \rightarrow \kappa f$?
  - $P \rightarrow P/\kappa^2$

Effects?

- Area $1/\kappa^2$
- Capacitance $1/\kappa$
- Resistance κ
- Threshold (V_{th}) $1/\kappa$
- Current (I_d) $1/\kappa$
- Gate Delay (τ_{gd}) $1/\kappa$
- Wire Delay (τ_{wire}) 1
- Power $1/\kappa^2$–> $1/\kappa^3$

Delays?

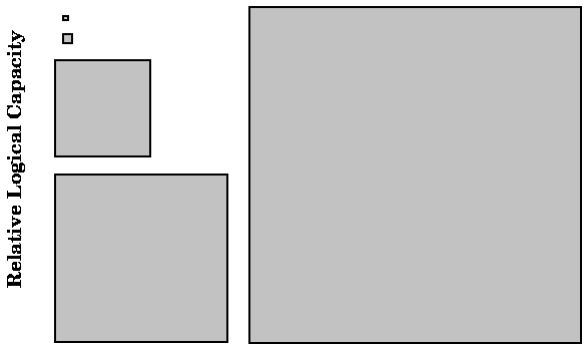
- If delays in gates/switching?
- If delays in interconnect?
- Logical interconnect lengths?

Delays?

- If delays in gates/switching?
  - Delay reduce with $1/\kappa$ [λ]

Delays

- Logical capacities growing
- Wirelengths?
  - No locallity–>κ
  - Rent's Rule
    - L –>$n^{(p-0.5)}$
    - [p>0.5]



Capacity

- Rent: $IO=C*N^p$
- $p>0.5$
- $A= C*N^{2p}$
- Logical Area $\rightarrow\kappa^2$

  $\kappa^2 A= C*N_2^{2p}$

  $\kappa^2 N^{2p} = N_2^{2p}$

  $N_2 = \kappa^{(1/p)} N$

- Sanity Check
  - $p=1$
  - $N_2 = \kappa N$
  - $p\sim0.5$
  - $N_2 \sim \kappa^2 N$

Compute Density

- Density = compute / (Area * Time)

  κ^3>compute density scaling>κ

  κ^3: gates dominate, $p<0.5$

  κ^2: moderate p, good fraction of gate delay

  κ: large p (wires dominate area and delay)

Power Density

- P–> P /κ^2 (static, or increase frequency)
- P–> P/κ^3 (dynamic, same freq.)
  A –> A/κ^2
- P/A –> P/A … or … P/κA

Physical Limits

- Doping?
- Features?

Physical Limits

- Depended on
  - bulk effects
    - doping
    - current (many electrons)
    - mean free path in conductor
  - localized to conductors
- Eventually
  - single electrons, atoms
  - distances close enough to allow tunneling

Finishing Up...

Big Ideas [MSB Ideas]

- Instruction organization induces a design space (taxonomy) for programmable architectures
- Moderately predictable VLSI Scaling
  - unprecedented capacities/capability growth for engineered systems
  - change
  - be prepared to exploit
  - account for in comparing across time

Big Ideas [MSB-1 Ideas]

- Uniform scaling reasonably accurate for past couple of decades
- Area increase κ^2
  - Real capacity maybe a little less?
- Gate delay decreases ($1/\kappa$)
- Wire delay not decrease, maybe increase
- Overall delay decrease less than ($1/\kappa$)