A Static Power Model for Architects

J. Adam Butts and Guri Sohi

University of Wisconsin-Madison

{butts,sohi}@cs.wisc.edu

33rd International Symposium on Microarchitecture

Monterey, California

December, 2000

Overview

The static power problem

- Leakage current
- Scaling trends

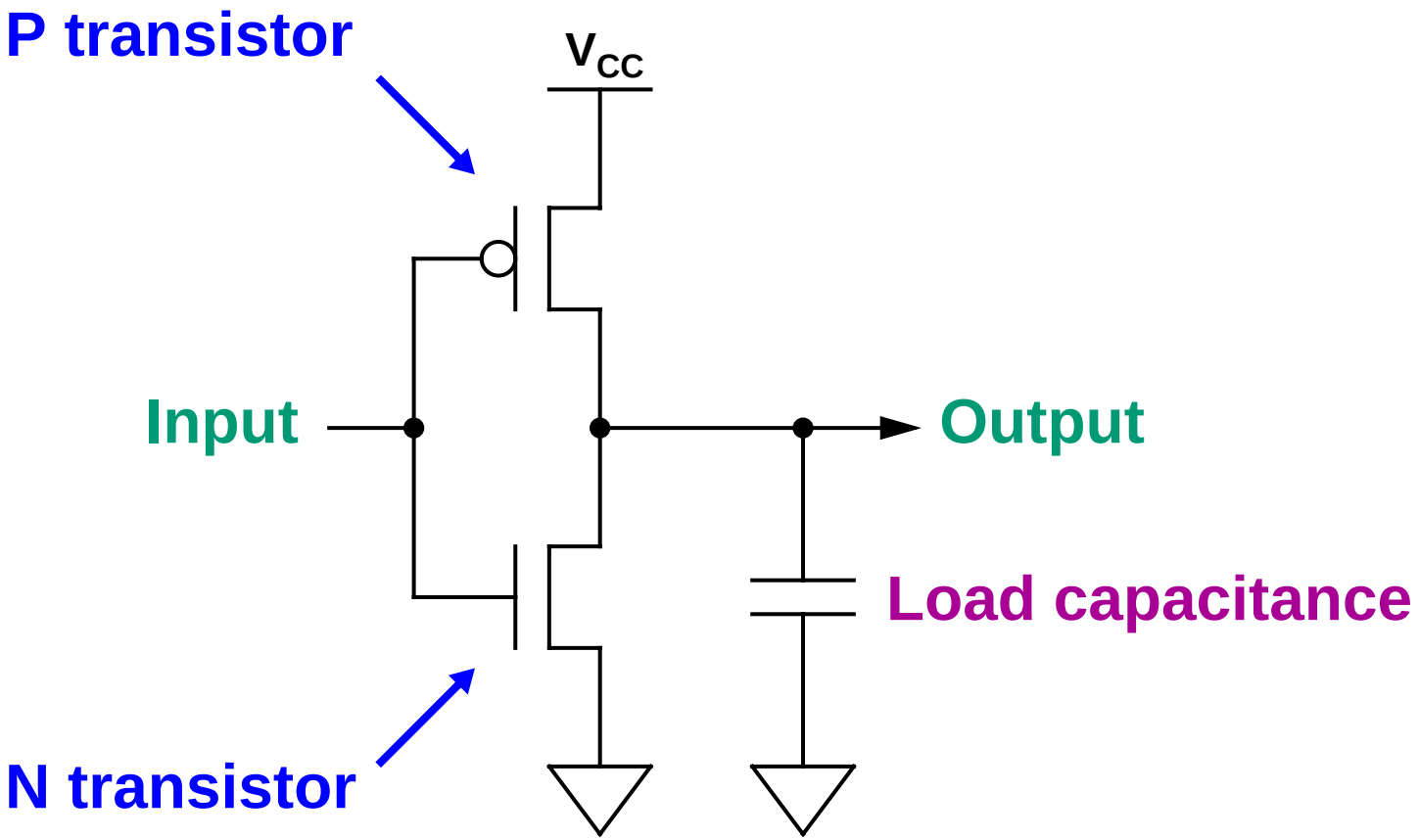
A static power model: $P_{static} = V_{CC} \cdot I_{leak} \cdot N \cdot k_{design}$

Attacking static power

- Power gating
- Using slower devices
- Applying speculation

Conclusion

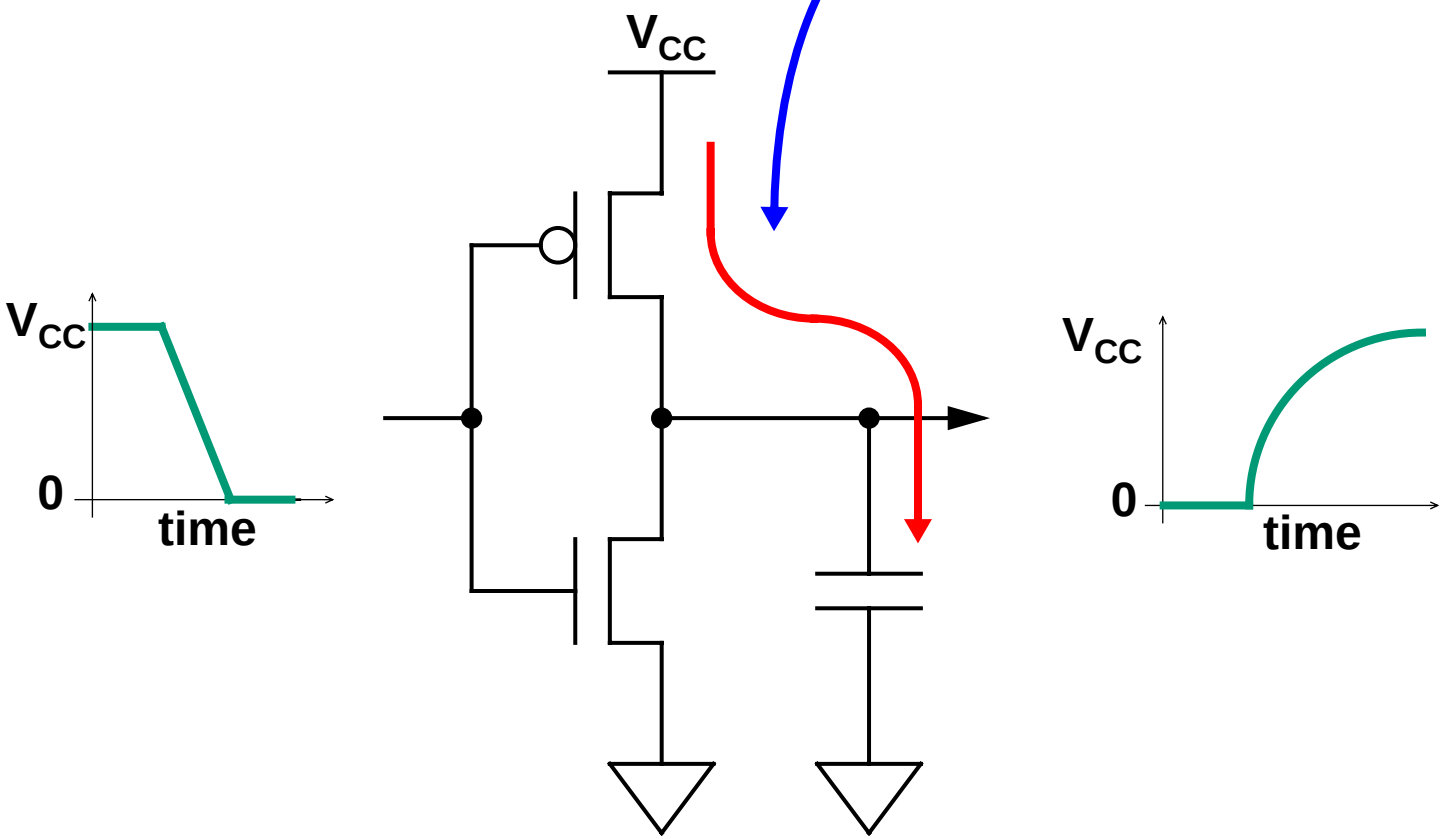
A CMOS Gate


P transistor
V_{CC}
Input
Output
Load capacitance
N transistor

Sources of Power Consumption

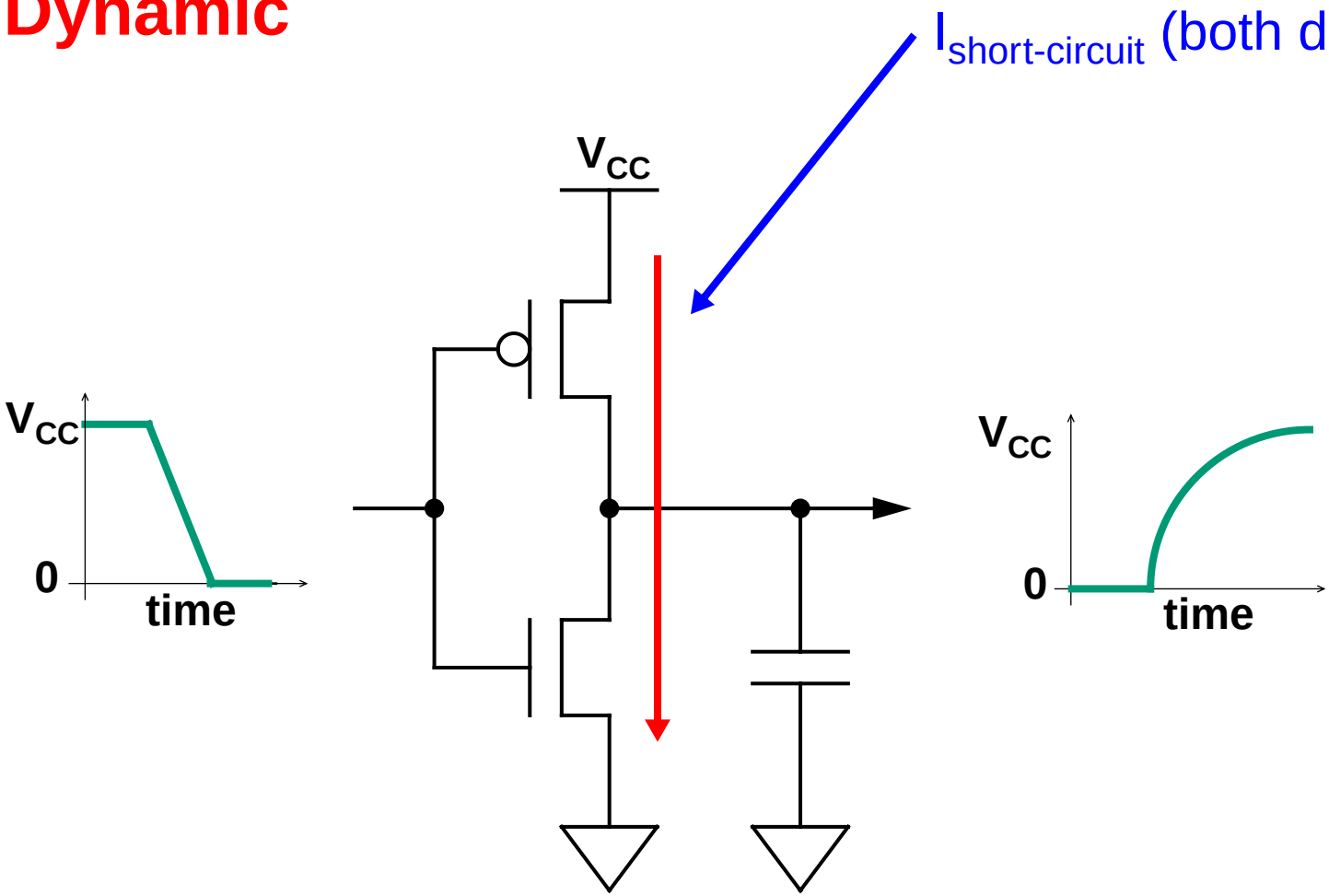
Dynamic

C dV/dt (charging of capacitative load)



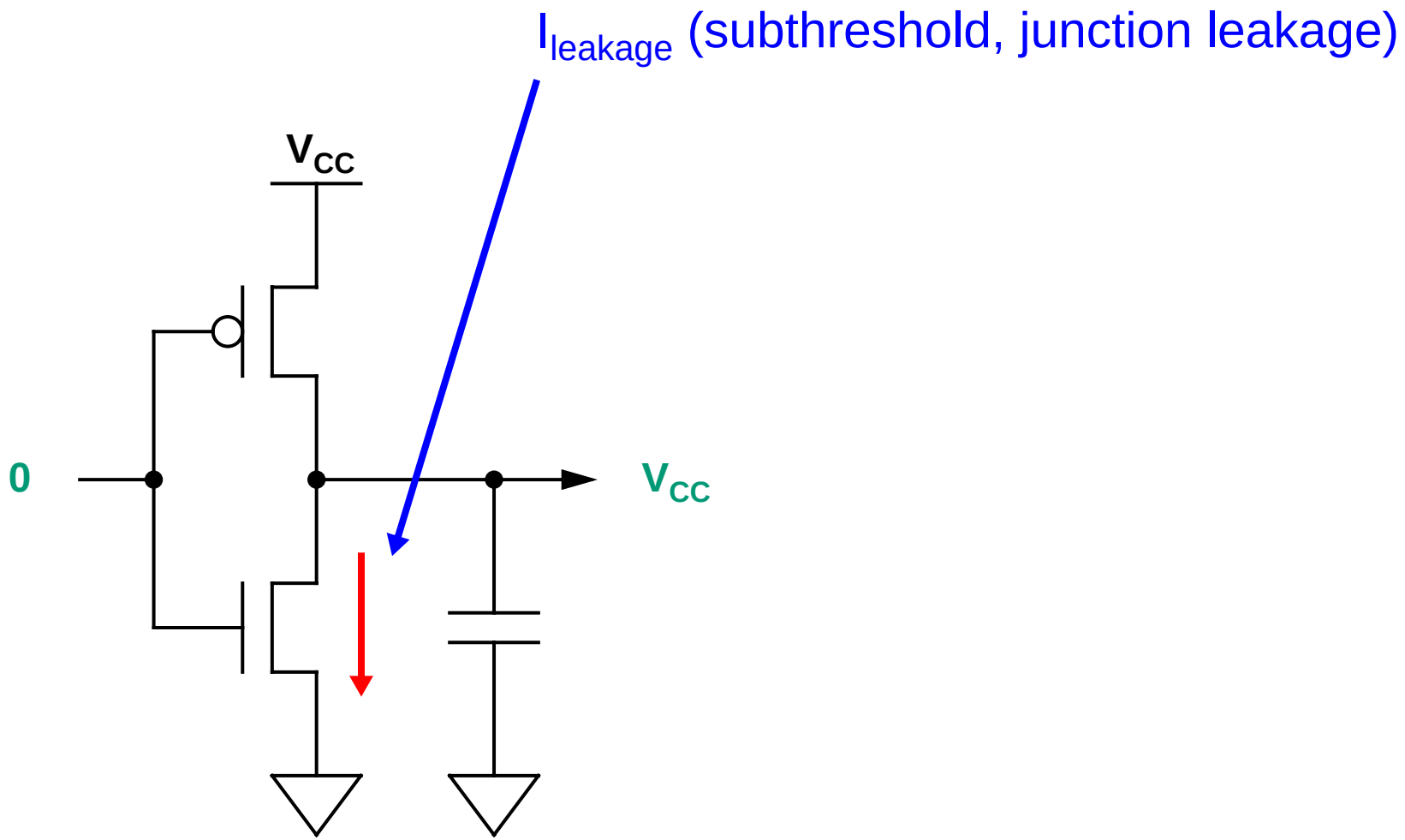
Sources of Power Consumption

Dynamic



Sources of Power Consumption

Static



Technology Scaling

Dimensions reduced to increase performance and density

V_{CC} decreases each generation...

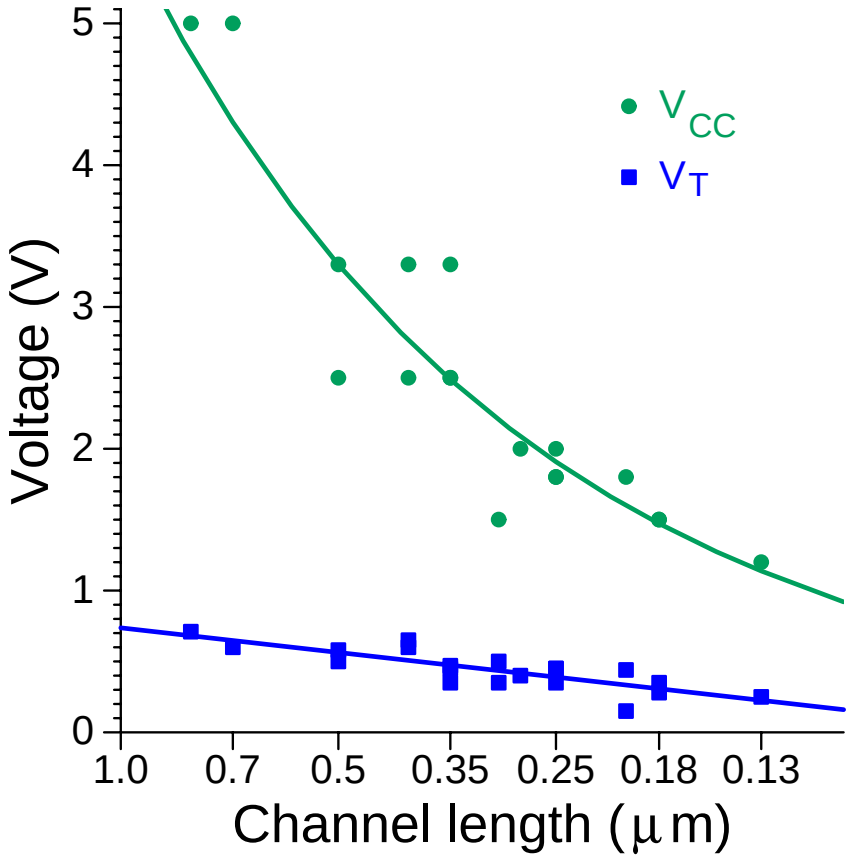
- Limit dynamic power
- Limit electric fields

...requiring lower V_T

- Gate overdrive = $V_{CC} - V_T$

Leakage increases exponentially

- $P_{static} = V_{CC} I_{leak} \sim exp(-V_T)$



Static Power Projections

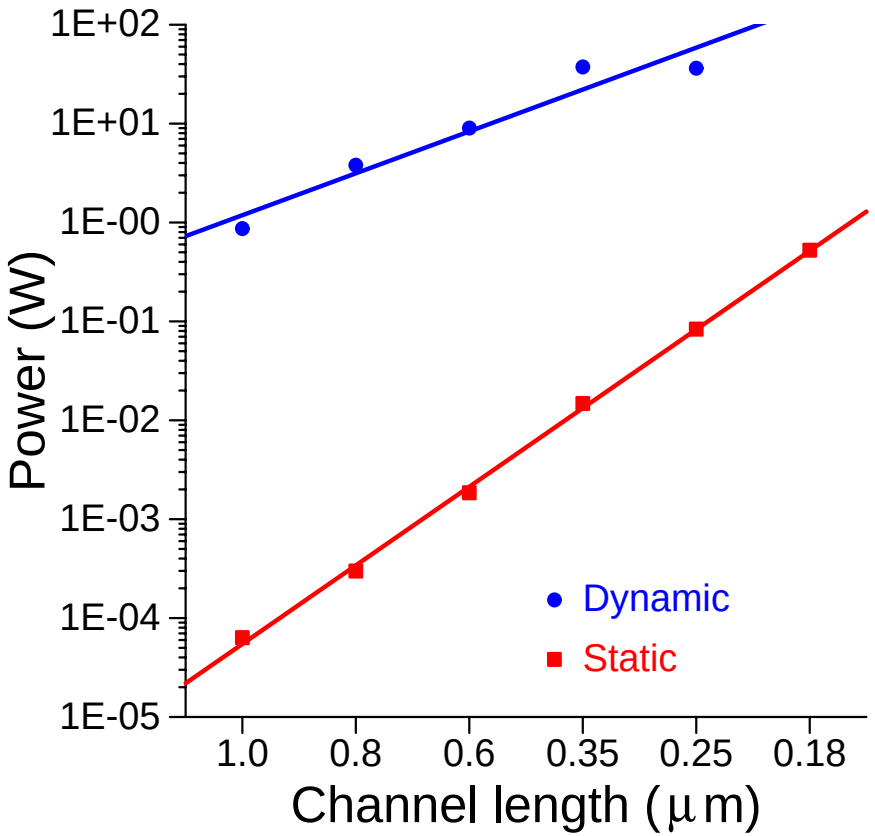
Static power is an increasing fraction of total power

Today: Pentium III 1.13 GHz

- P_{total} (peak) = 41.4 watts
- P_{static} = 5.4 watts
- Static power is **13 %** of total
- Higher contribution on average

This is only getting worse

- $P_{static} = P_{dynamic}$ in 3 generations



Important Characteristics of Static Power

❶ Exponentially increasing due to V_T scaling

- Increasing faster than dynamic power

❷ Adds to average power, not peak power

- More expensive than dynamic power

❸ Independent of transistor utilization

- Transistors are not free

Model Derivation

Want an equivalent of $C \cdot V_{CC}^2 \cdot f$ for static power

Develop model from the bottom-up

- Lack of data precludes a top-down “data-driven” approach
- Start from BSIM3v3.2 transistor model

$$I_{Dsub} = I_{s0}' \cdot \frac{W}{L} \cdot \left(1 - e^{\frac{-V_{ds}}{v_t}}\right) \cdot e^{\frac{V_{gs} - V_T - V_{off}}{n \cdot v_t}}$$

BSIM3 model eq.

Aspect ratio

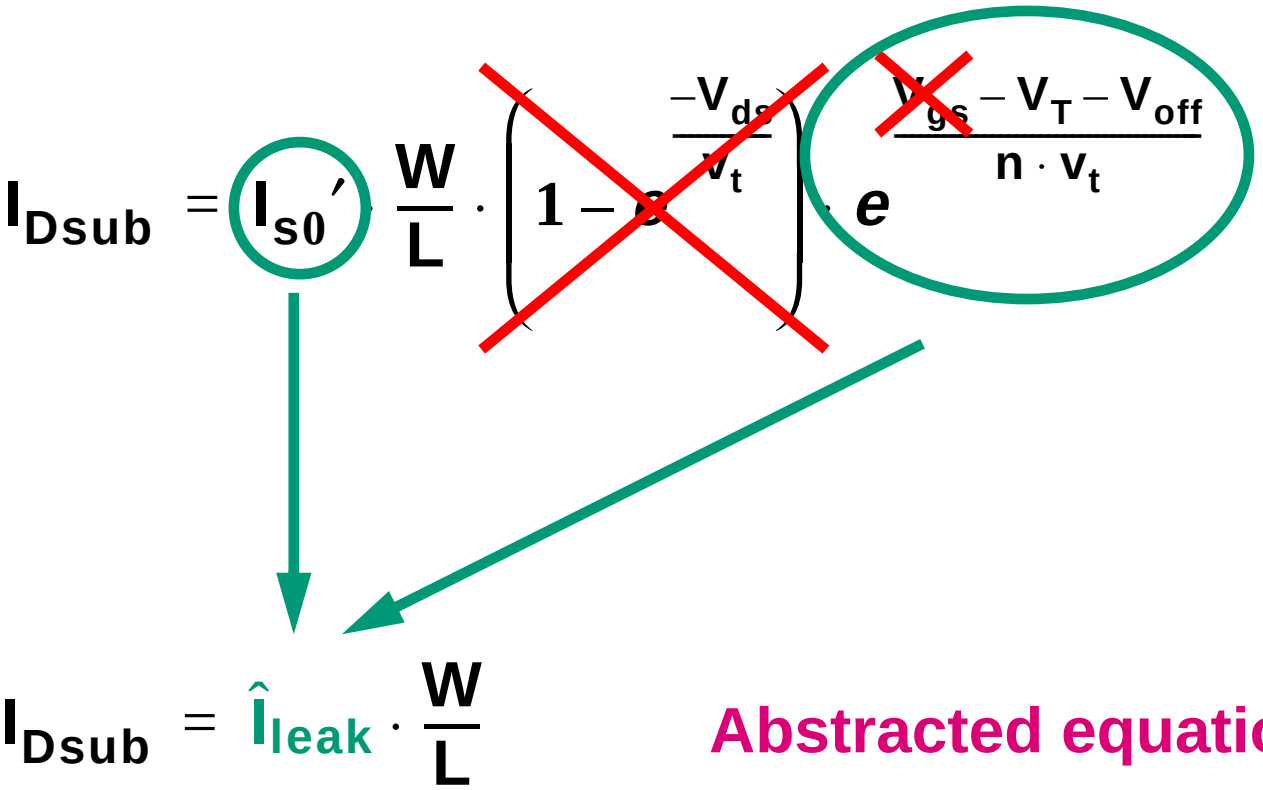
V_T dependence

Model Derivation

$$I_{Dsub} = I_{s0}' \cdot \frac{W}{L} \cdot \left(1 - e^{\frac{-V_{ds}}{v_t}}\right) \cdot e^{\frac{V_{gs} - V_T - V_{off}}{n \cdot v_t}}$$

❶ Apply BSIM to a single “off” (leaking) device

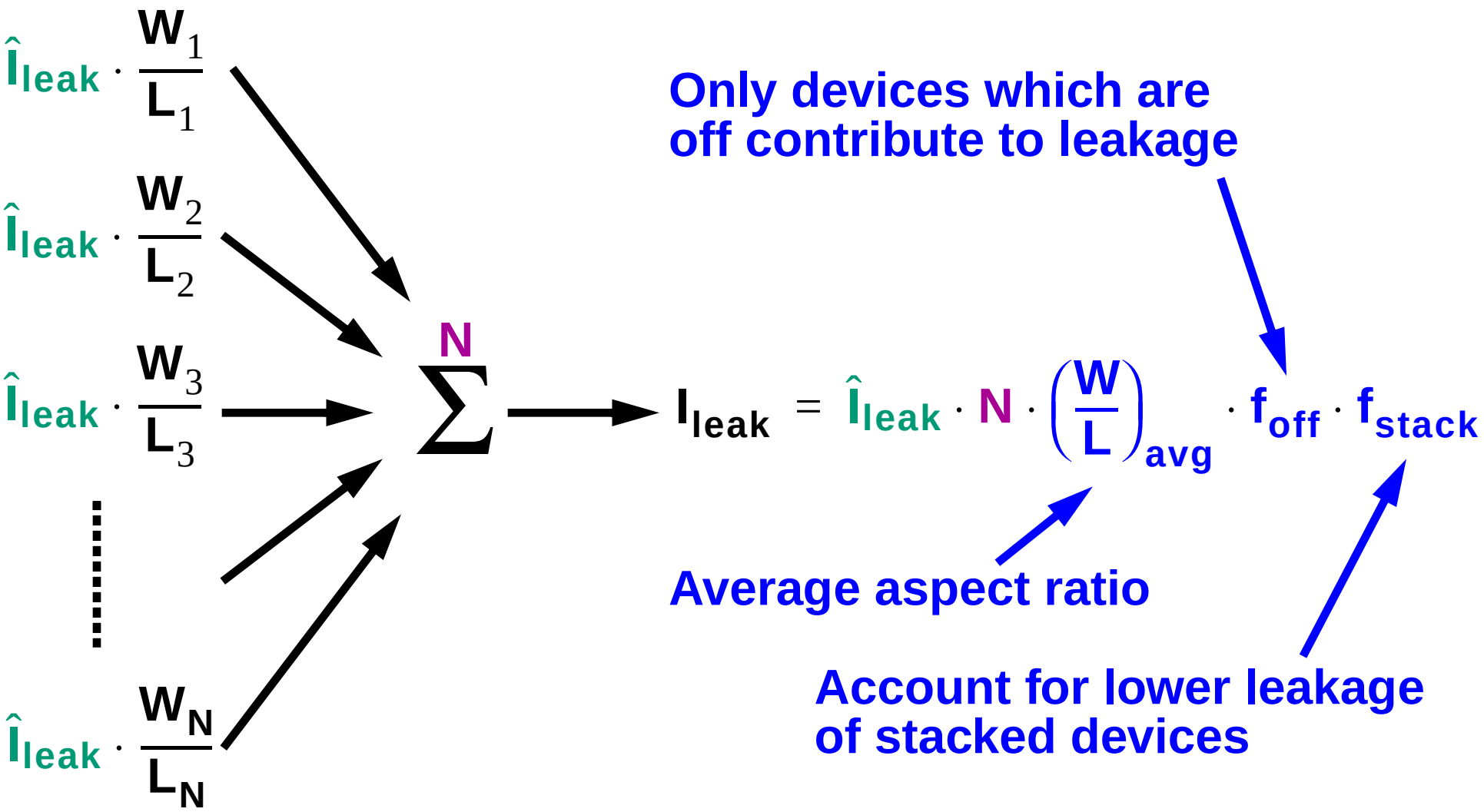
Model Derivation



Abstracted equation for a single device

❷ Group technology-dependent parameters together

Model Derivation



❸ Apply to large numbers of devices

Model Derivation

$$I_{leak} = \hat{I}_{leak} \cdot N \cdot \left(\frac{W}{L}\right)_{avg} \cdot f_{off} \cdot f_{stack}$$

$$I_{leak} = \hat{I}_{leak} \cdot N \cdot k_{design}$$

❹ Group design-dependent parameters together

Static Power Model

Resulting power model has four parameters

- Technology-dependent (from scaling, process data)
- Design-dependent (from estimates, past designs)

$$P_{static} = V_{CC} \cdot \hat{I}_{leak} \cdot N \cdot k_{design}$$

Supply voltage

Speed of transistor

Circuit style

Number of transistors

The Design Constant

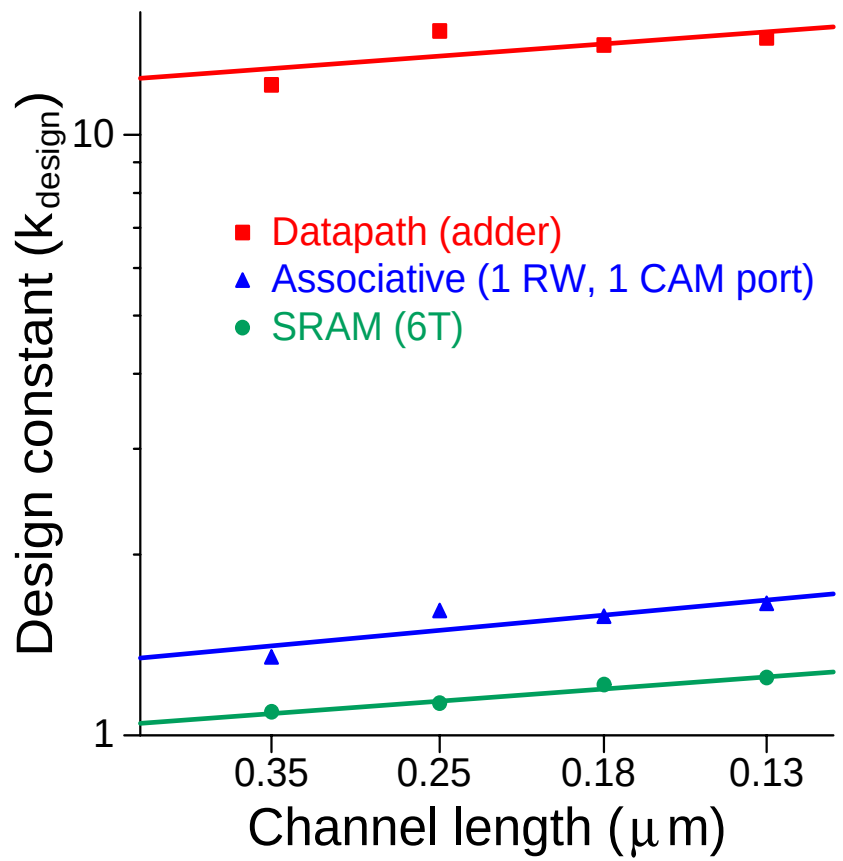
Represents an “average” device

- Aspect ratio (device size)
- Fraction of leaking devices
- Stacking factor

Depends on design style

Independent of technology

- → Allows for forward projection



Attacking Static Power

Power reduction techniques address factors in the model equation:

$$P_{static} = V_{CC} \cdot \hat{I}_{leak} \cdot N \cdot k_{design}$$

Use power aware microarchitecture

- Use fewer devices
- Power gating

Employ slow devices

- Enables supply voltage reduction (voltage partitioning)
- Enables use of higher threshold voltage devices

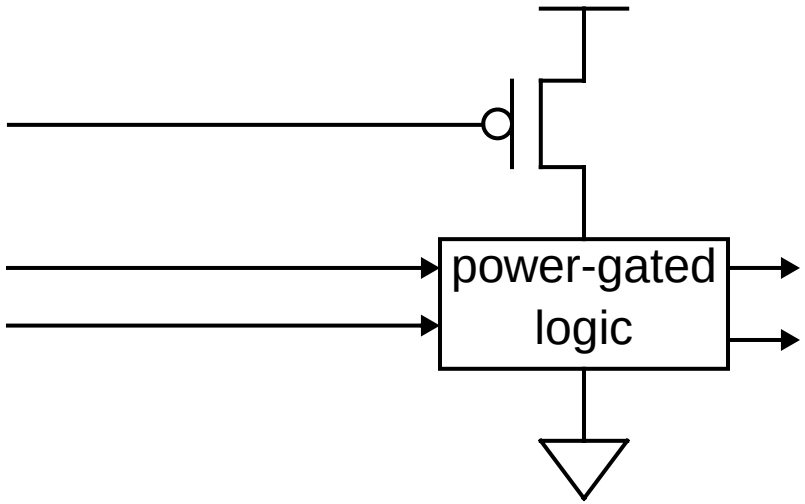
Power Gating

Eliminate leakage by removing power to unused devices

- Analogous to clock gating
- Requires logic to determine power down/up conditions

Many power gating possibilities

- Floating point hardware
- Rare instruction decode logic
- Interrupt handling hardware



Power-up prediction problem

- Large decoupling capacitance
- Limited charging current & dI/dt
- → **Several cycles of power-up latency**

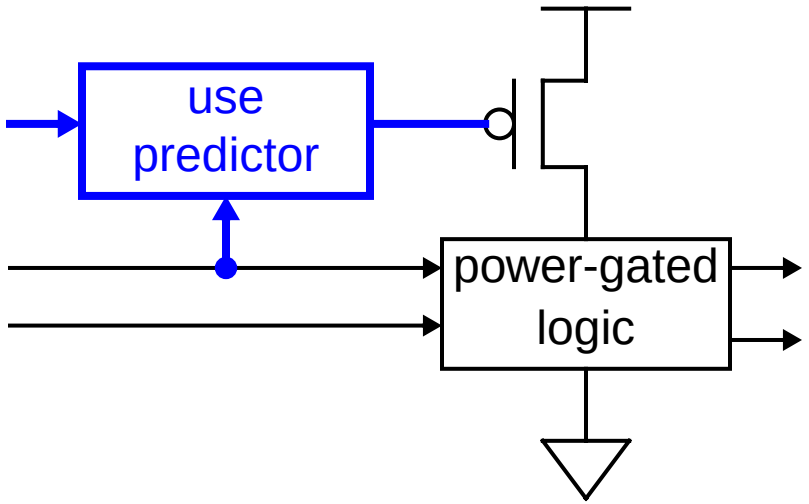
Speculative Power Gating

Power-up latency limits power gating potential

- ❶ Do not gate power (**no power savings**)
- ❷ Accept power-up latency (**lower performance**)
- ❸ **Build predictor for power-up condition**

Adjustable misprediction penalties

- Power/performance bias



Sample Applications

- PC based prediction for special instruction needs
- PC based prediction for L1 miss handler (L1-L2 interface)

Using Slower Devices

Trade latency and area for power

- 2× devices at 0.5× frequency
  → Equivalent throughput with higher latency and lower total power

Reducing clock frequency helps only dynamic power

- Multiple threshold voltage technology (multiple frequency domains)
- Variable supply voltage (multiple supply voltage domains)

Architectural Issues

- Interdomain communication
- Latency tolerance

Using Slower Devices with Speculation

Speculation is a latency tolerance technique

- Generate speculative result more quickly than it can be determined
- Check accuracy off critical path, recover when wrong
- → Average latency is **decreased**

Using Slower Devices with Speculation

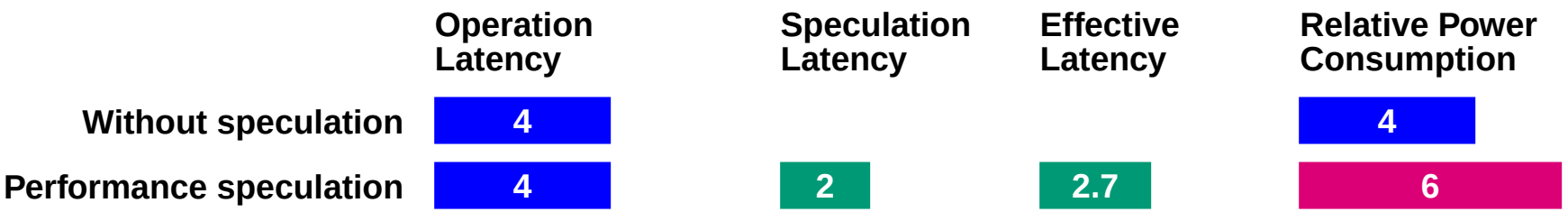
Speculation is a latency tolerance technique

- Generate speculative result more quickly than it can be determined
- Check accuracy off critical path, recover when wrong
- → Average latency is **decreased**

Using Slower Devices with Speculation

Speculation is a latency tolerance technique

- Generate speculative result more quickly than it can be determined
- Check accuracy off critical path, recover when wrong
- → Average latency is **decreased**

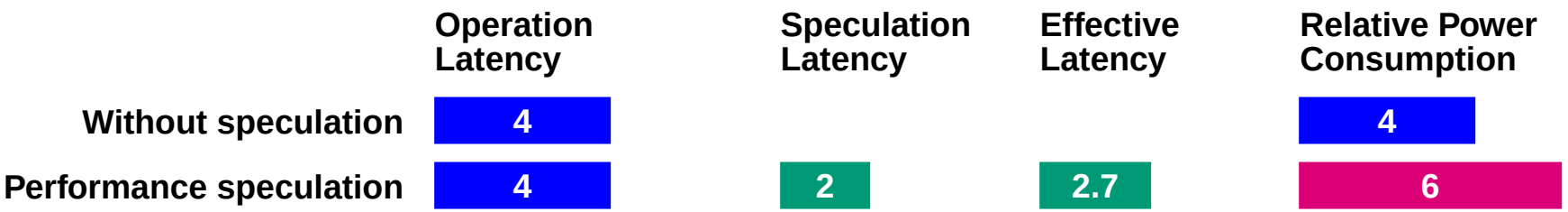


Using Slower Devices with Speculation

Speculation is a latency tolerance technique

- Generate speculative result more quickly than it can be determined
- Check accuracy off critical path, recover when wrong
- → Average latency is **decreased**

Use slower devices to save power and speculation to tolerate increased latency



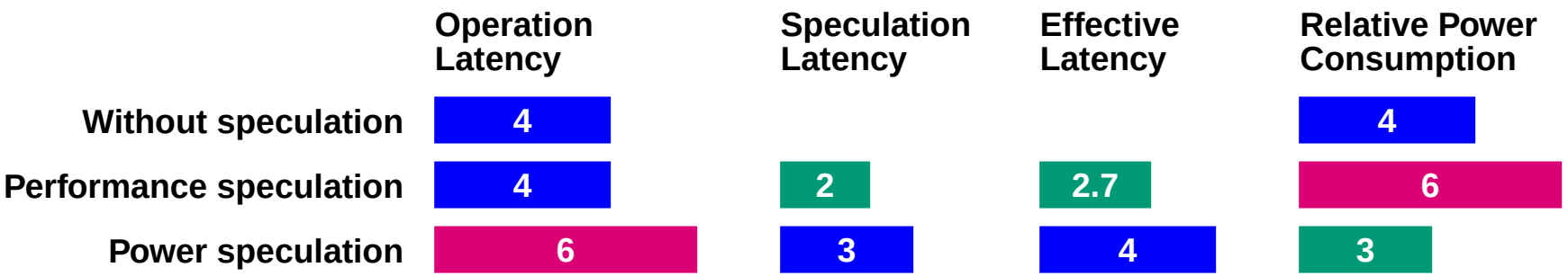
A Static Power Model for Architects - J. Adam Butts and Guri Sohi
33rd International Symposium on Microarchitecture, December 2000

Using Slower Devices with Speculation

Speculation is a latency tolerance technique

- Generate speculative result more quickly than it can be determined
- Check accuracy off critical path, recover when wrong
- → Average latency is **decreased**

Use slower devices to save power and speculation to tolerate increased latency



A Static Power Model for Architects - J. Adam Butts and Guri Sohi
33rd International Symposium on Microarchitecture, December 2000

Conclusions

Static power will become important (V_T scaling)

A high-level model is available: $P_{static} = V_{CC} \cdot I_{leak} \cdot N \cdot k_{design}$

Reducing static power also reduces dynamic power

Speculation as a power savings technique

- Speculative power gating
- Allows use of slower devices with controlled performance penalty

What can architects do to impact static power dissipation?

- Latency/throughput tradeoffs
- Design partitioning (voltage/frequency domains)
- Identify idle resources, predict the need for them
- Identify opportunities for power speculation