Power and Leakage Reduction in the Nanoscale Era

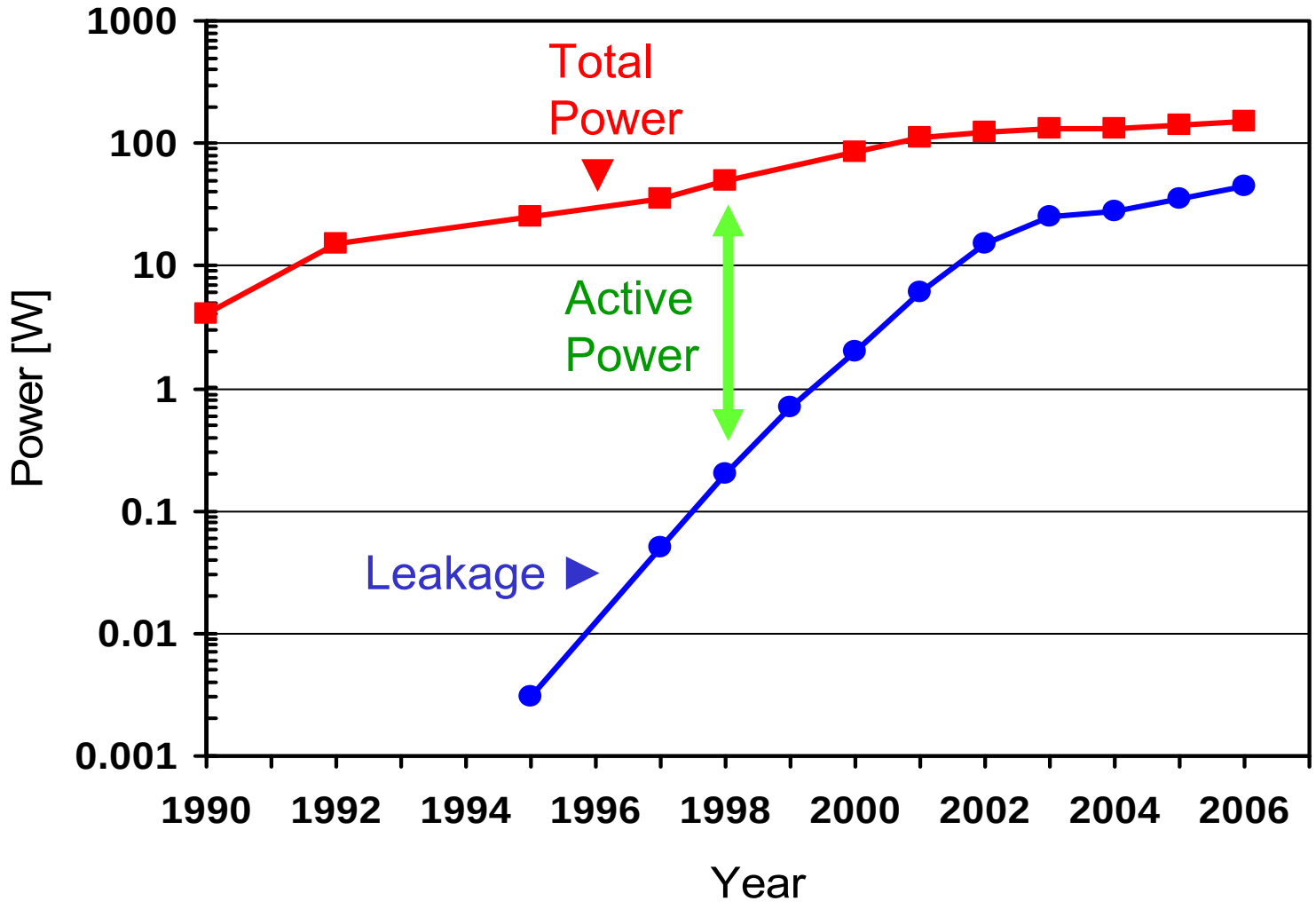
Stefan Rusu
Senior Principal Engineer
Intel Corporation

August 21st, 2008

Outline

- Power components and trends
- Active power reduction techniques
- Leakage reduction techniques
- Power management methods
- Summary

Server Processor Power Trends



Power Components

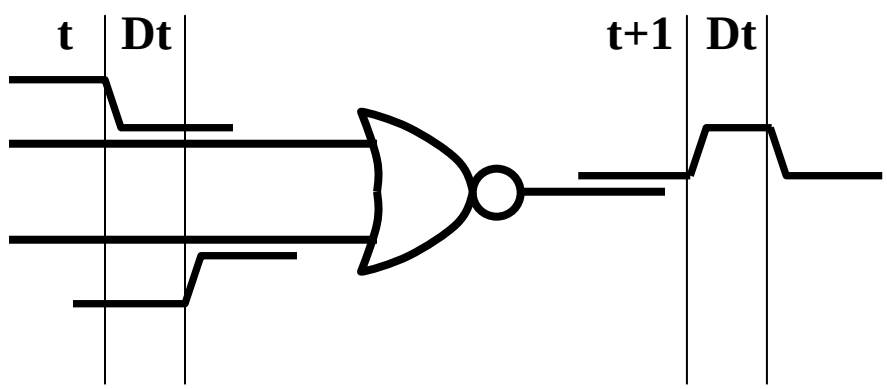
- Total power includes switching, short-circuit and leakage:

$$P = P_{sw} + P_{short} + P_{leakage}$$

$$P_{sw} = f \cdot V_{cc}^2 \cdot \sum_{i=1}^{n} AF_i \cdot C_i$$

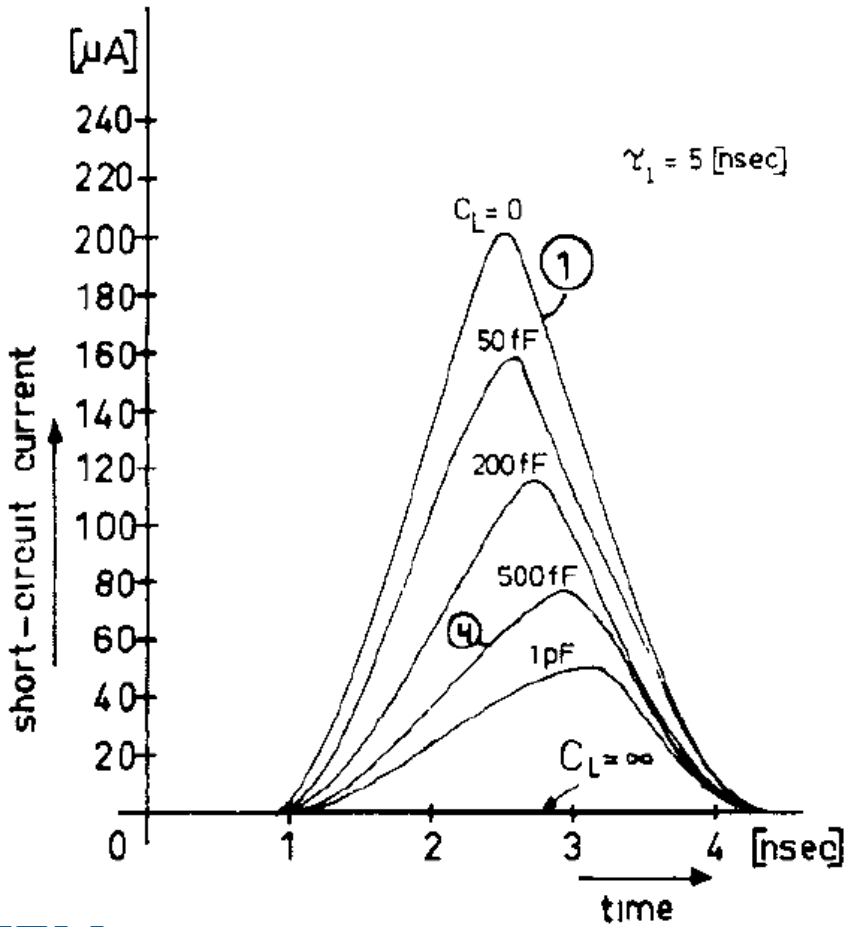
$$AF_i = AF_i^{0-delay} + AF_i^{glitch}$$

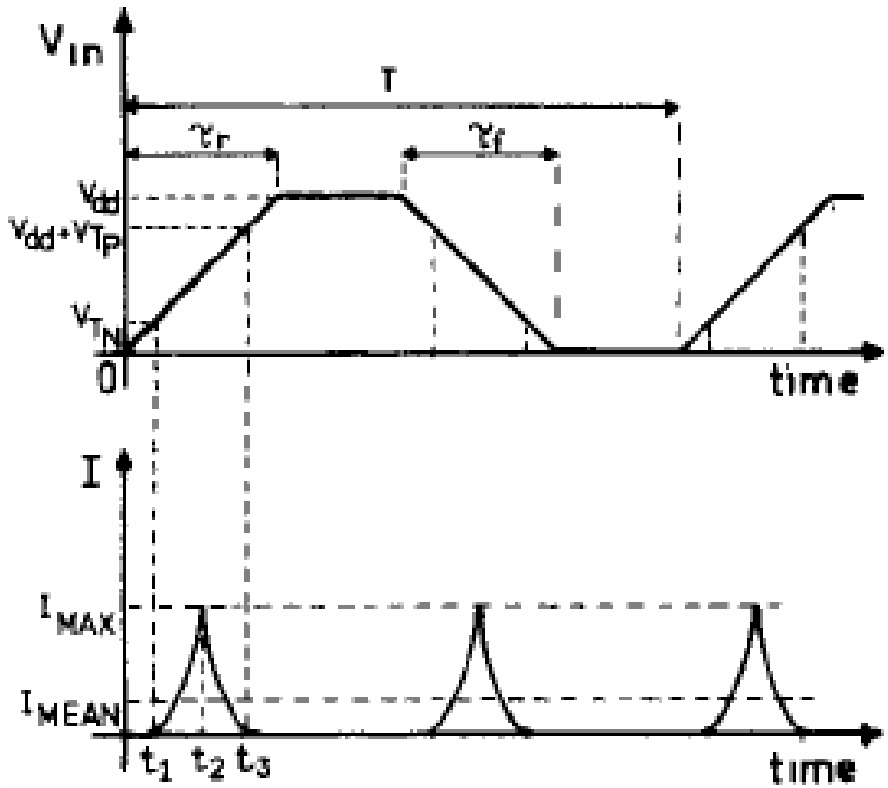
- Glitches are a significant contributor to power as illustrated in the NOR gate example below



Short Circuit Power

- Short circuit power is a function of $(Vcc - 2Vt)^3$
- Linearly increases with input slope ► Avoid large slopes

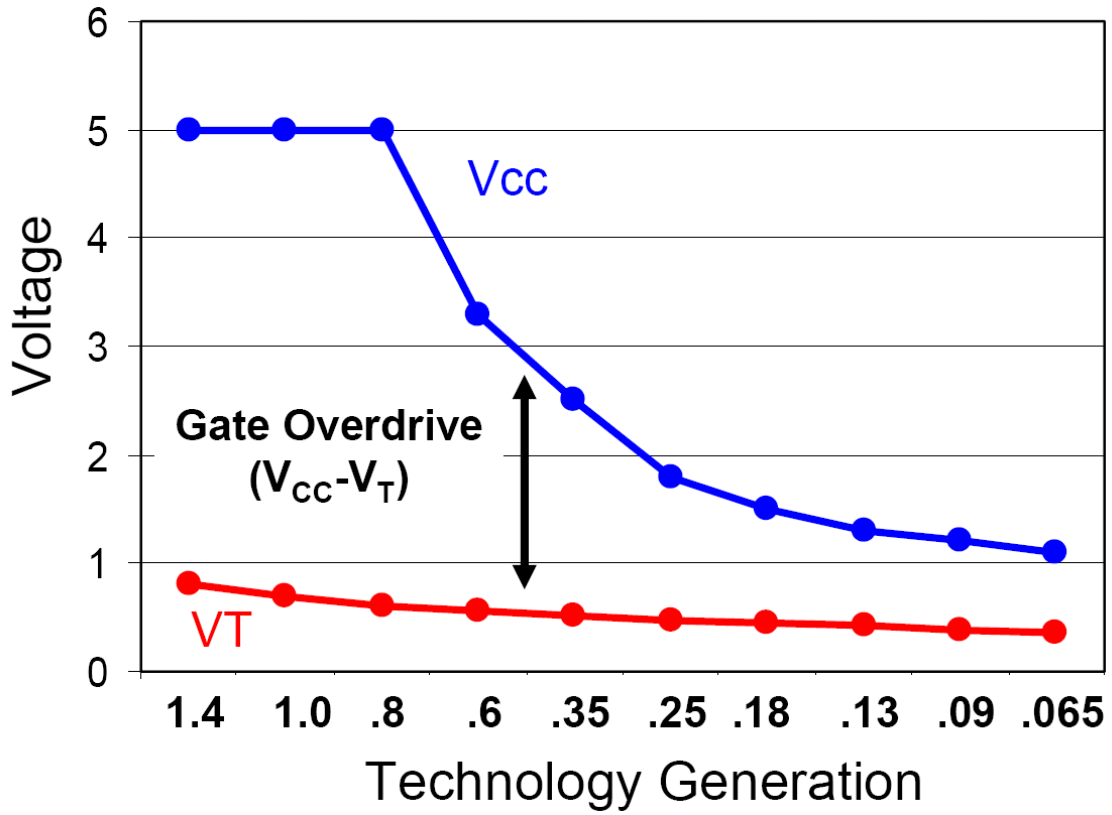




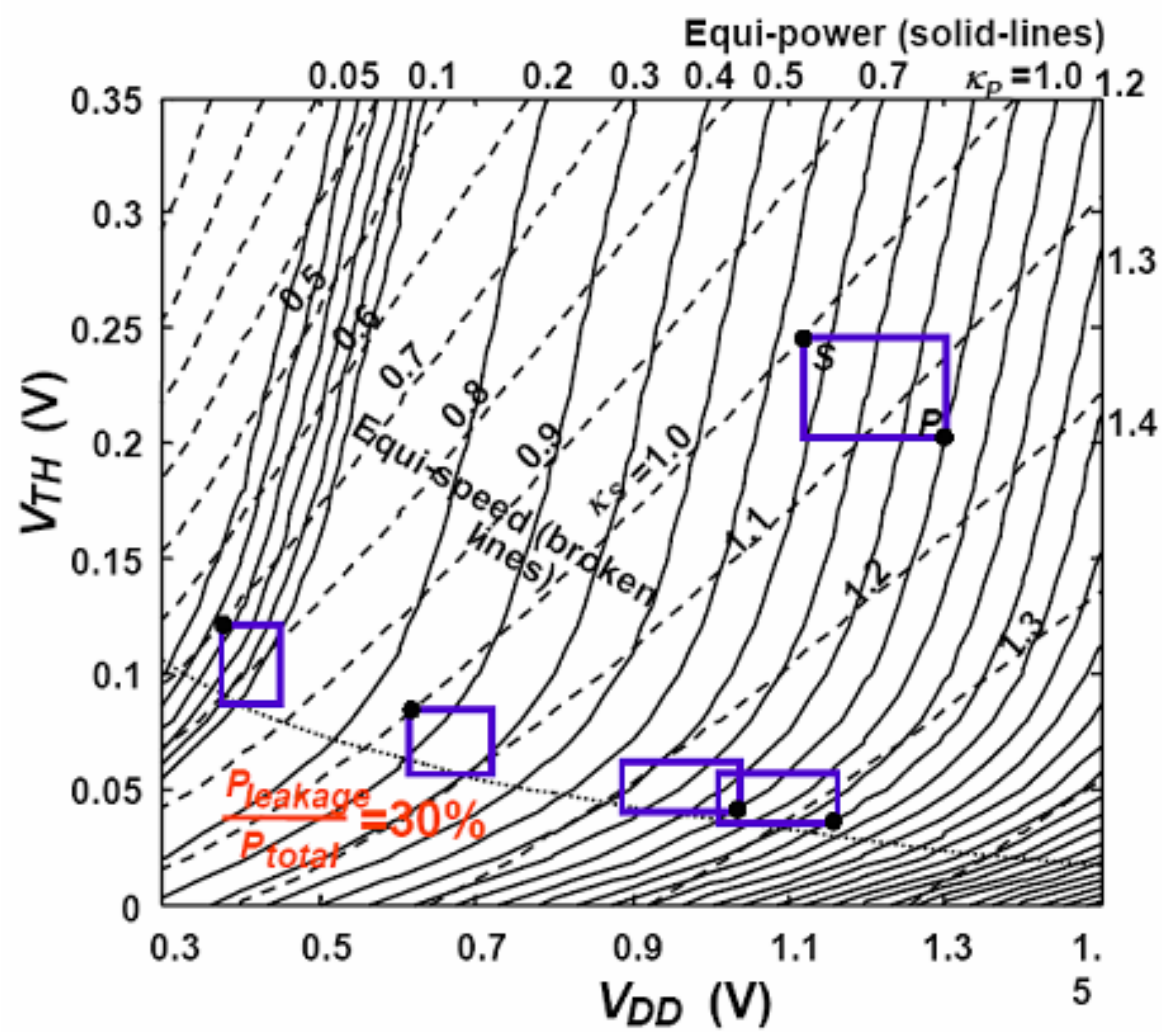
H. Veendrick (NXP), JSSC, 1984

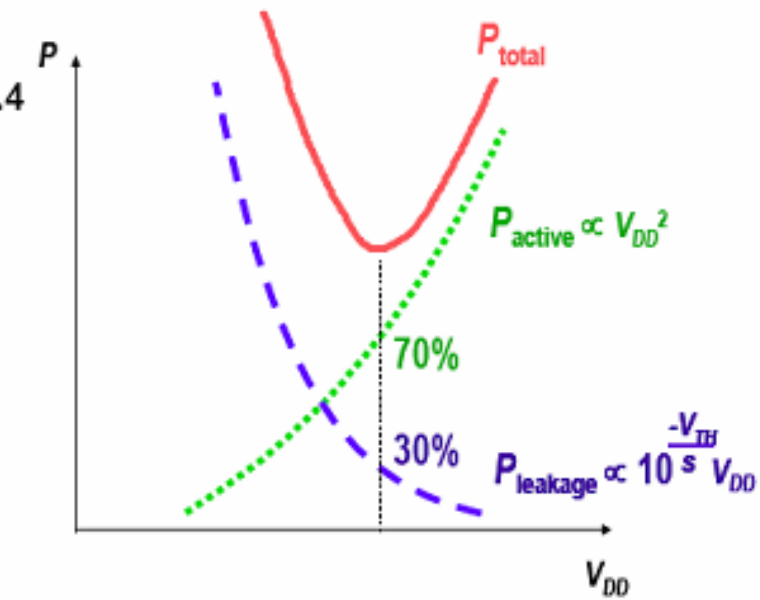
Voltage Scaling Trends

- Vcc scaling has been driven by power and oxide reliability
- Gate overdrive is decreasing with each technology generation
- VT is scaling very slowly
- Vcc scaling trend is decreasing due to performance concerns



Optimal Active / Leakage Power Ratio

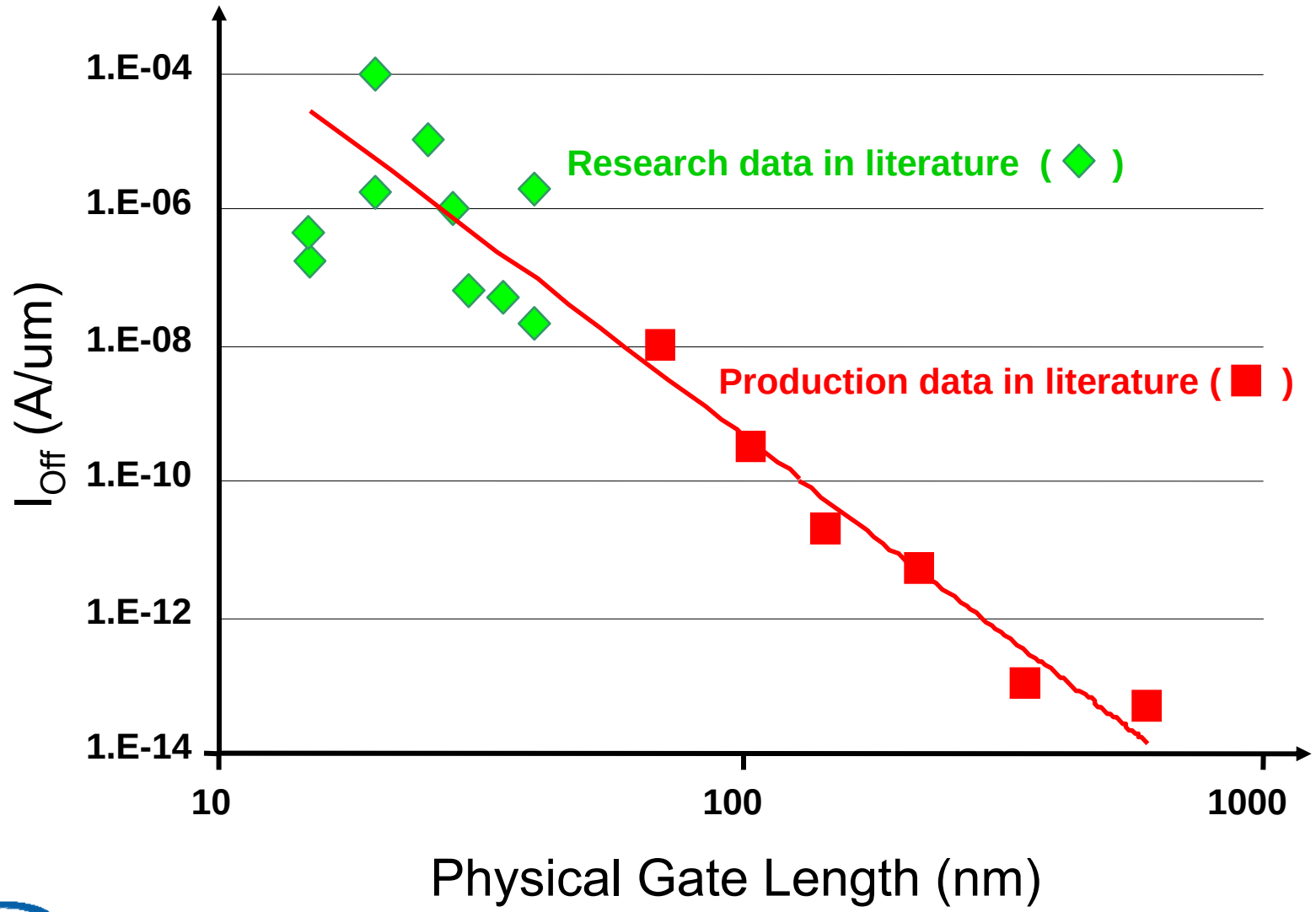




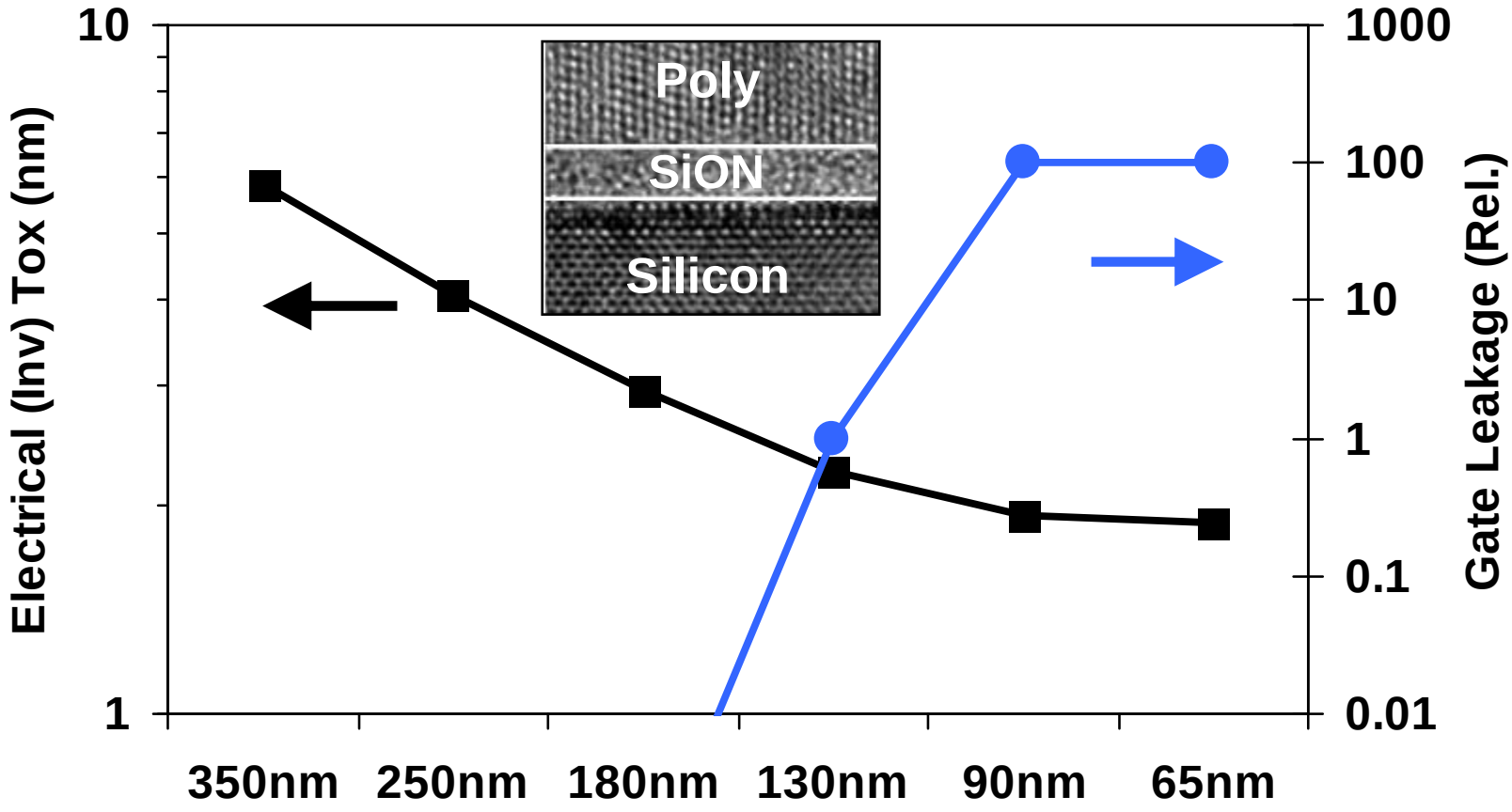
Optimal active/leakage power ratio is 70/30

Kuroda (Keio Univ.), ICCAD 2002

Source/Drain Leakage (I_{off})



Gate Leakage Trends



- SiON scaling running out of atoms
- Poly depletion limits inversion T_{OX} scaling

K. Mistry, et. al, IEDM 2007

45nm High-K + Metal Gate Transistors

Metal Gate

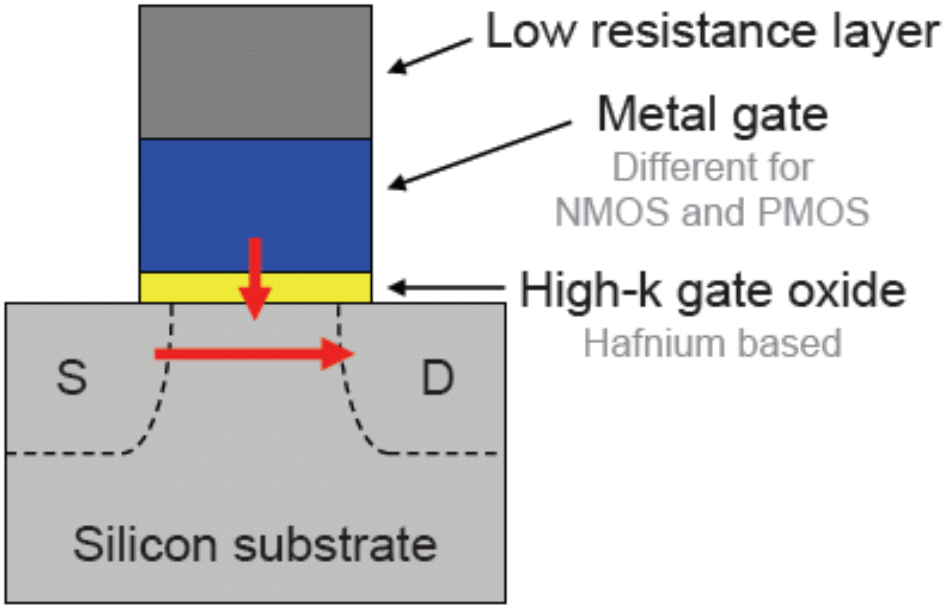
- Increases the gate field effect

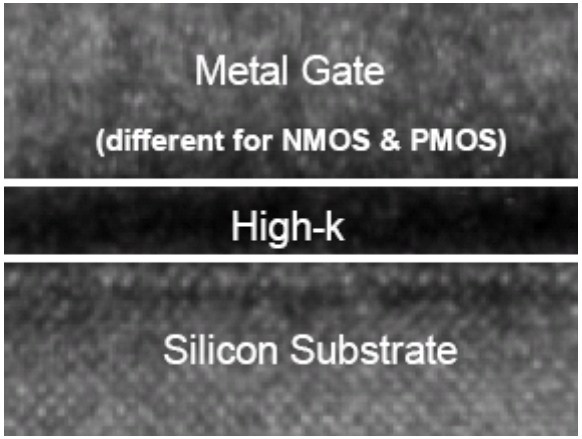
High-K Dielectric

- Increases the gate field effect
- Allows use of thicker dielectric layer to reduce gate leakage

HK + MG Combined

- Drive current increased >20%
- Or source-drain leakage reduced >5x
- Gate oxide leakage reduced

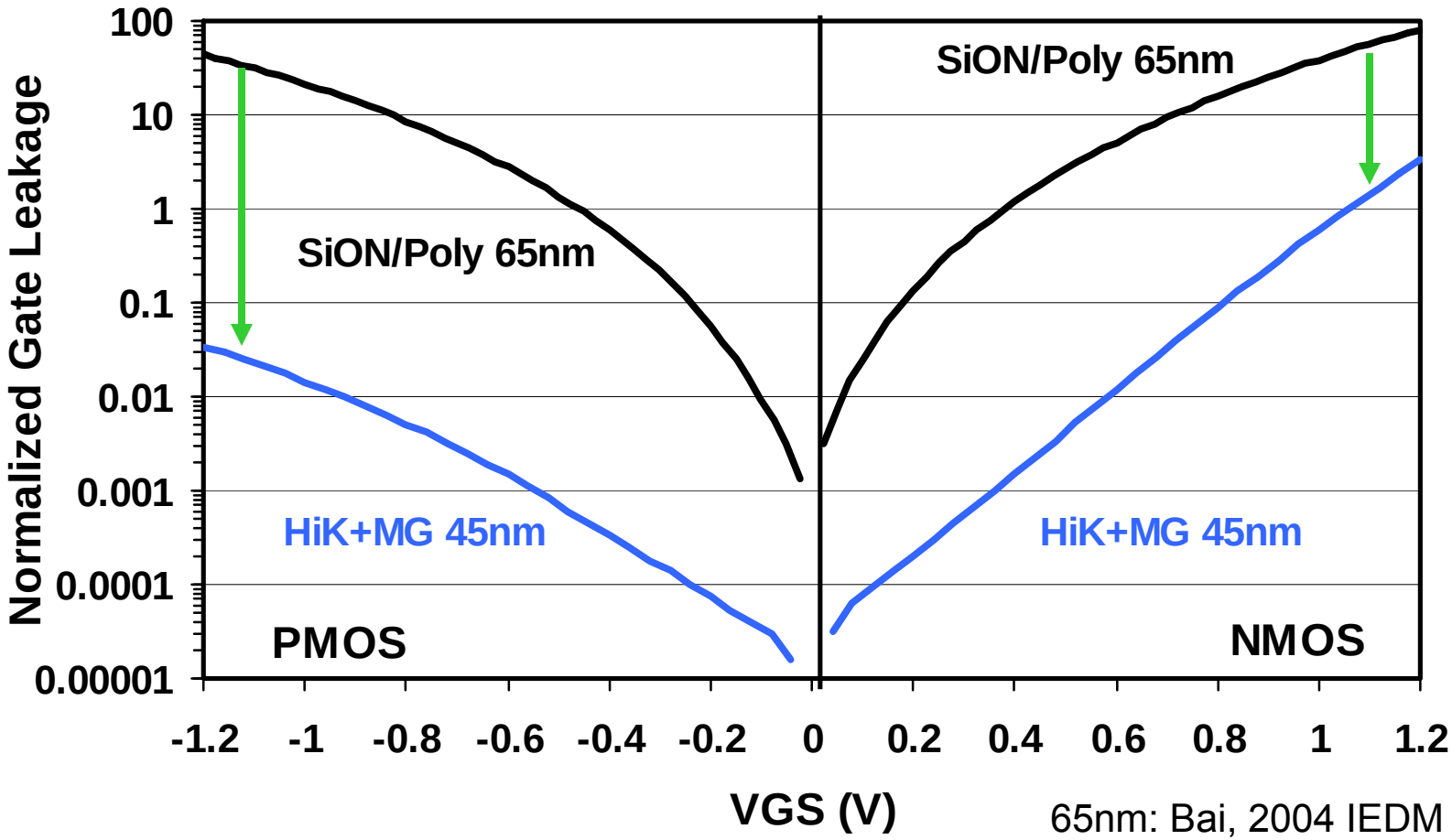




http://download.intel.com/pressroom/kits/45nm/Press%2045nm%20107_FINAL.pdf

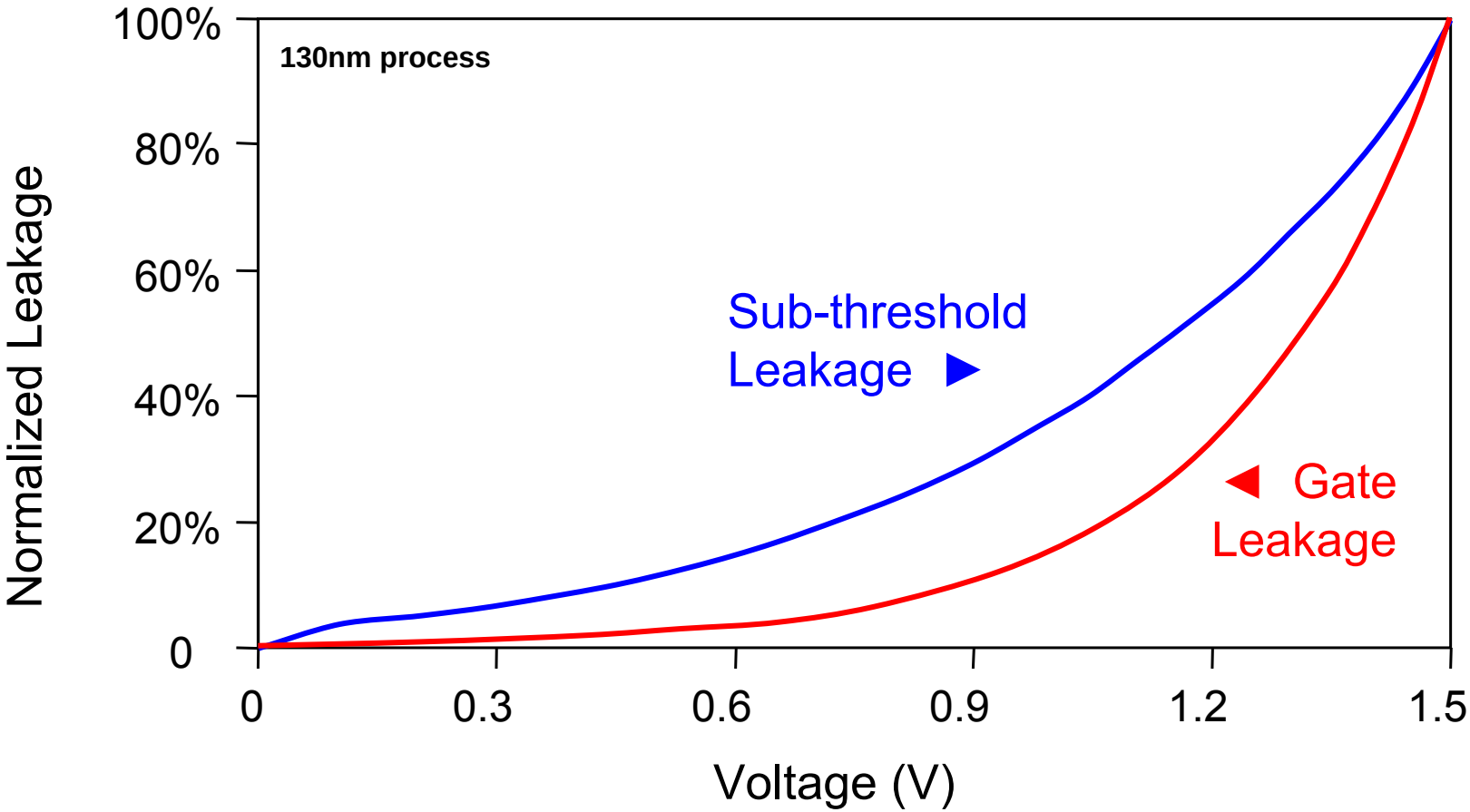
HK+MG Gate Leakage Reduction

- Gate leakage is reduced >25X for NMOS and 1000X for PMOS



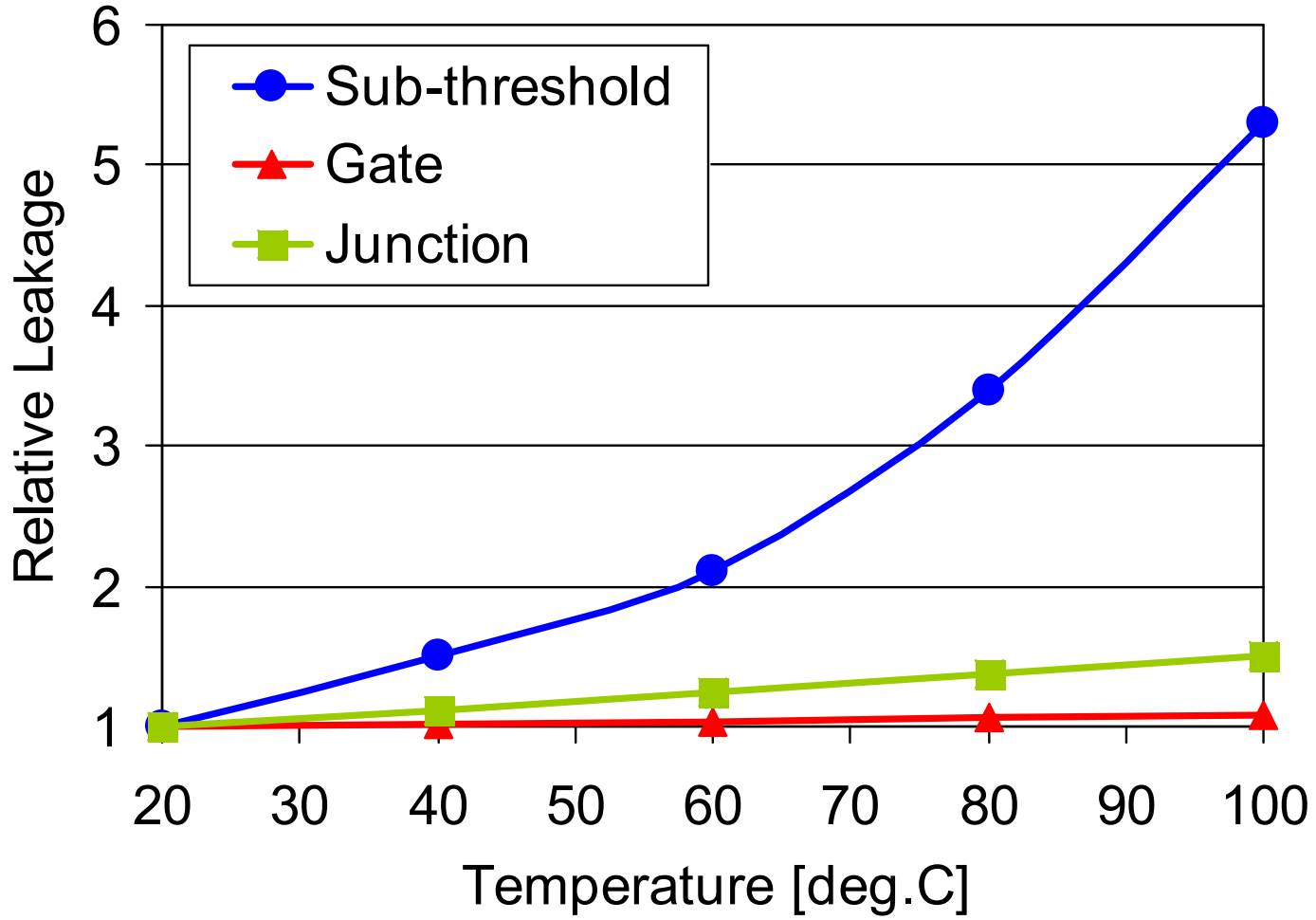
65nm: Bai, 2004 IEDM
45nm: Mistry, 2007 IEDM

Leakage Dependency on Voltage



[Krishnamurthy, et. al, ASICON 2005]

... And Temperature



[Mukhopadhyay, et al., VLSI Symposium 2003]

Outline

- Power components and trends
- Active power reduction techniques
  - Clock gating
  - Reduce clock loading
  - Multiple cores
  - Multiple voltage domains
- Leakage reduction techniques
- Power management methods
- Summary

Active Power Reduction

Reduce switched capacitance:
- Minimize diffusion, wire and gate loading, particularly in high activity factor nodes (clocks, domino)
- Use more efficient layout techniques

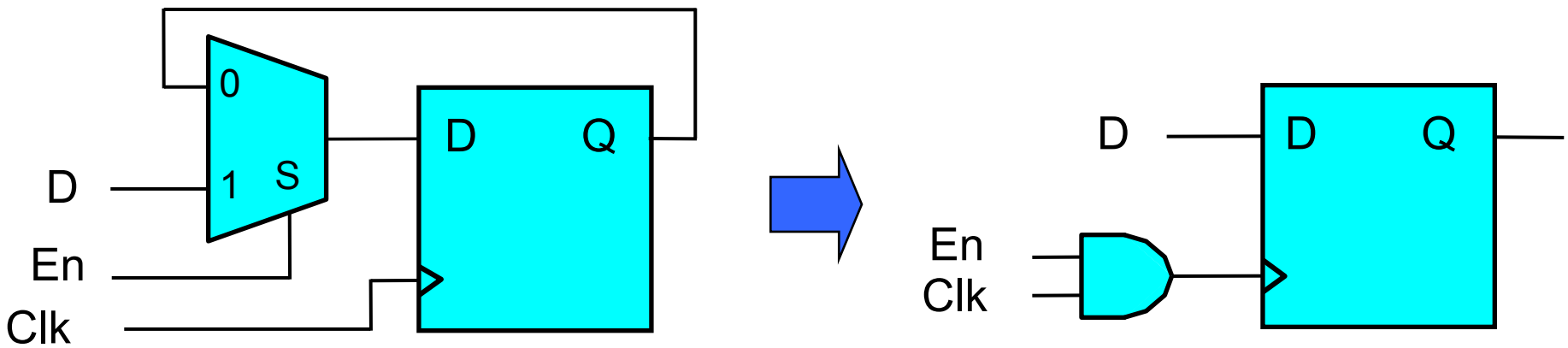
Technology scaling:
- Dynamic voltage scaling
- Supply voltage scaling is slowing down
- Thresholds don't scale

$$P = \alpha C_L V^2 f_{CLK}$$

Reduce switching activity:
- Conditional execution
- Conditional clocking
- Conditional precharge
- Turn off inactive blocks
- Reduce toggling of high capacitance nodes/busses

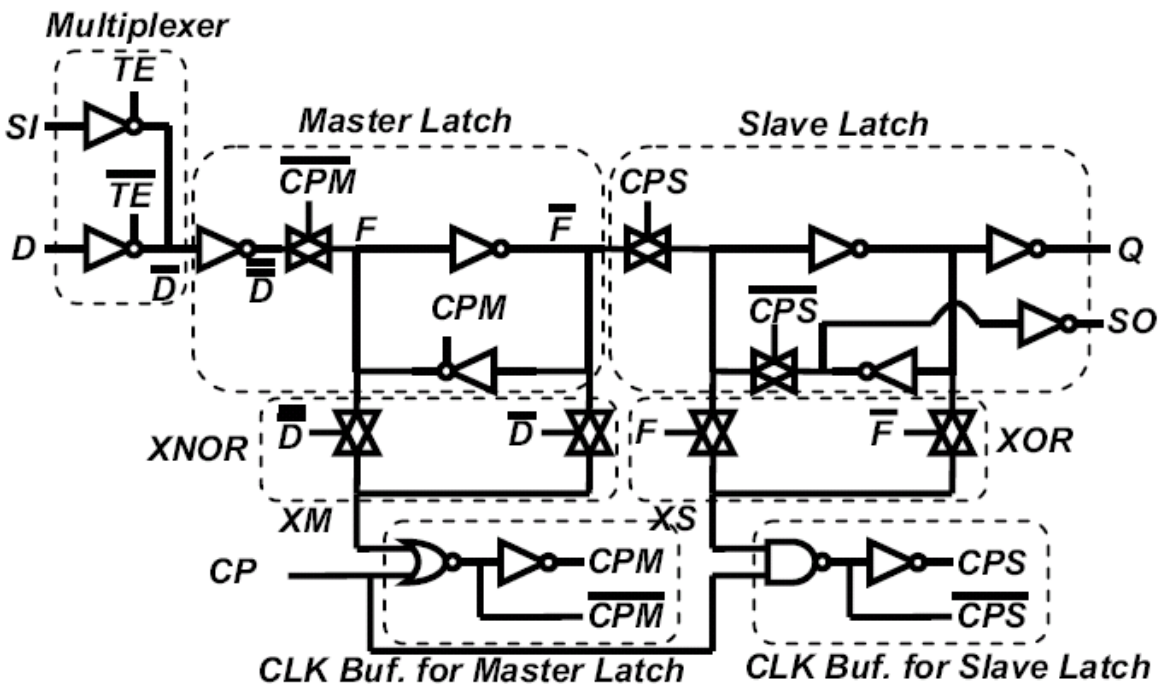
Reduce clock frequency:
- Use parallelism
- Less pipeline stages
- Use double-edge flip-flops

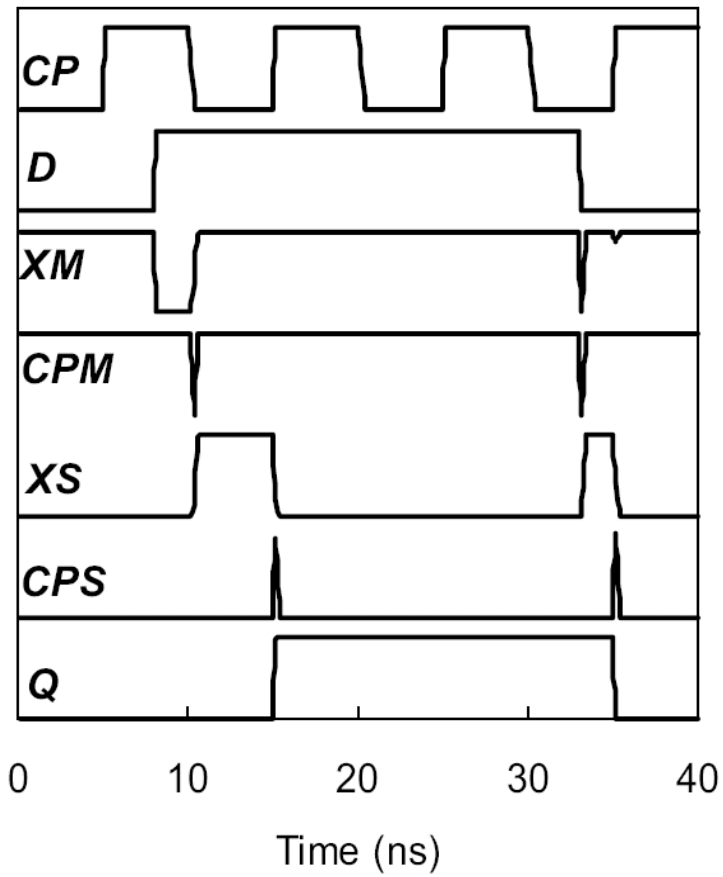
Clock Gating



- Save power by gating the clock when data activity is low
- Widest used switching power reduction technique
- Requires early En signal arrival, as well as detailed timing and logic validation

Conditional Clocking Flip-Flop



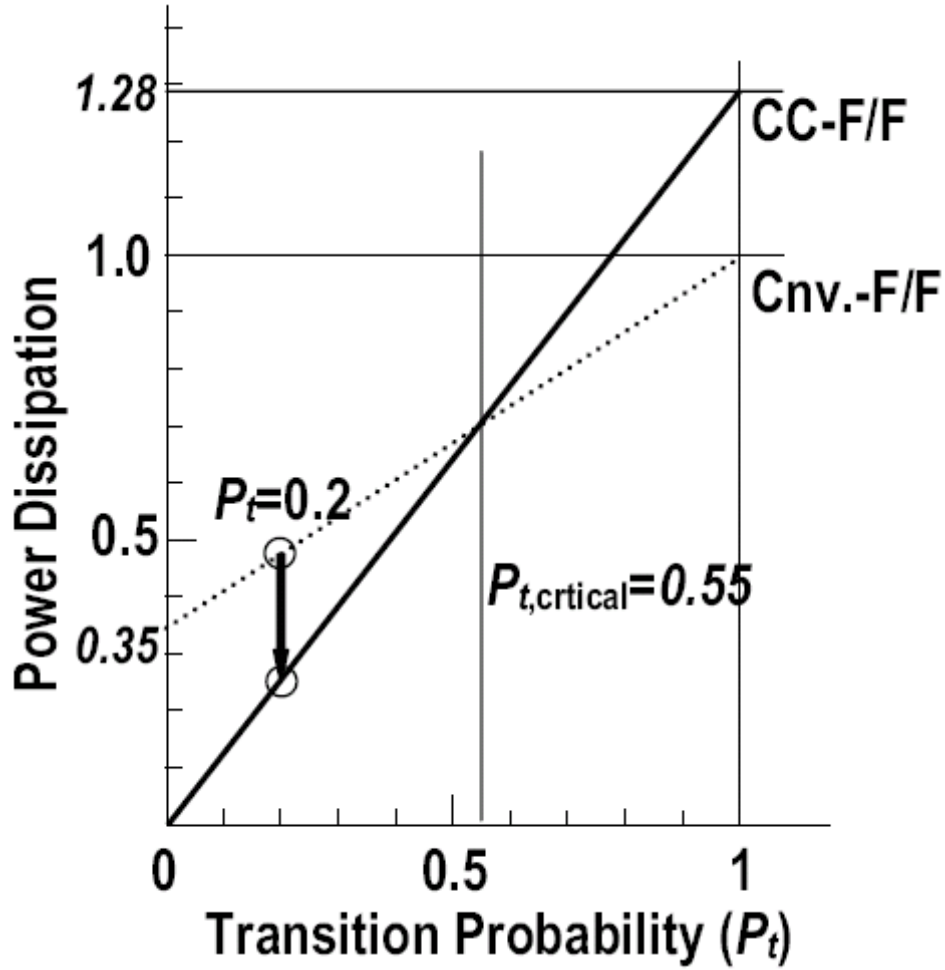


- FF does not consume active power when the data input does not change its state

M. Hamada (Toshiba), CICC, 2005

Conditional Clocking Flip-Flop (2)

		conventional	conditional clk
Power	$P_{LH/HL}$	1.00	0.35
	$P_{LL/HH}$	1.28	0.00
Delay (ps)	*CP*-to-*Q*	82	86
	Setup	84	199
	Hold	-72	-195
Area		1.00	1.33

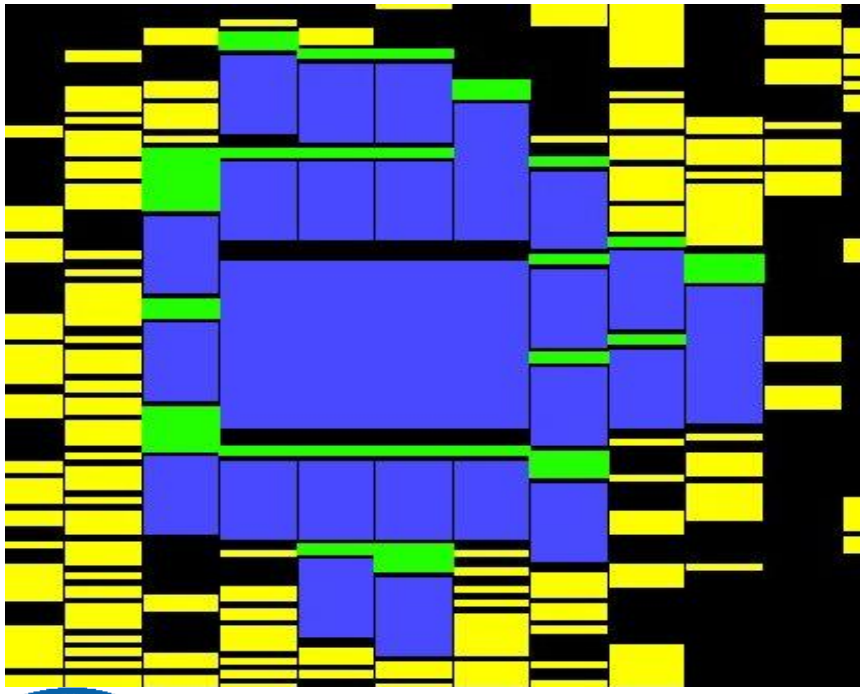


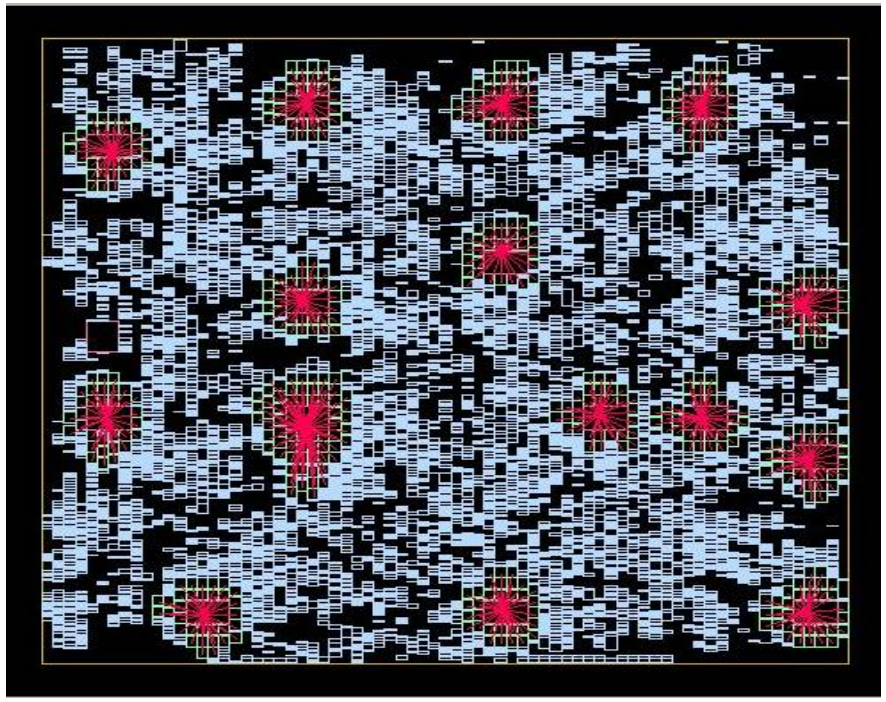
- Taking into account the overhead of the auxiliary circuits, the flip-flop consumes less power than the conventional flip-flop when the data transition probability is less than 55%
- Issues: leakage, setup time

M. Hamada (Toshiba), CICC, 2005

Latch Clustering

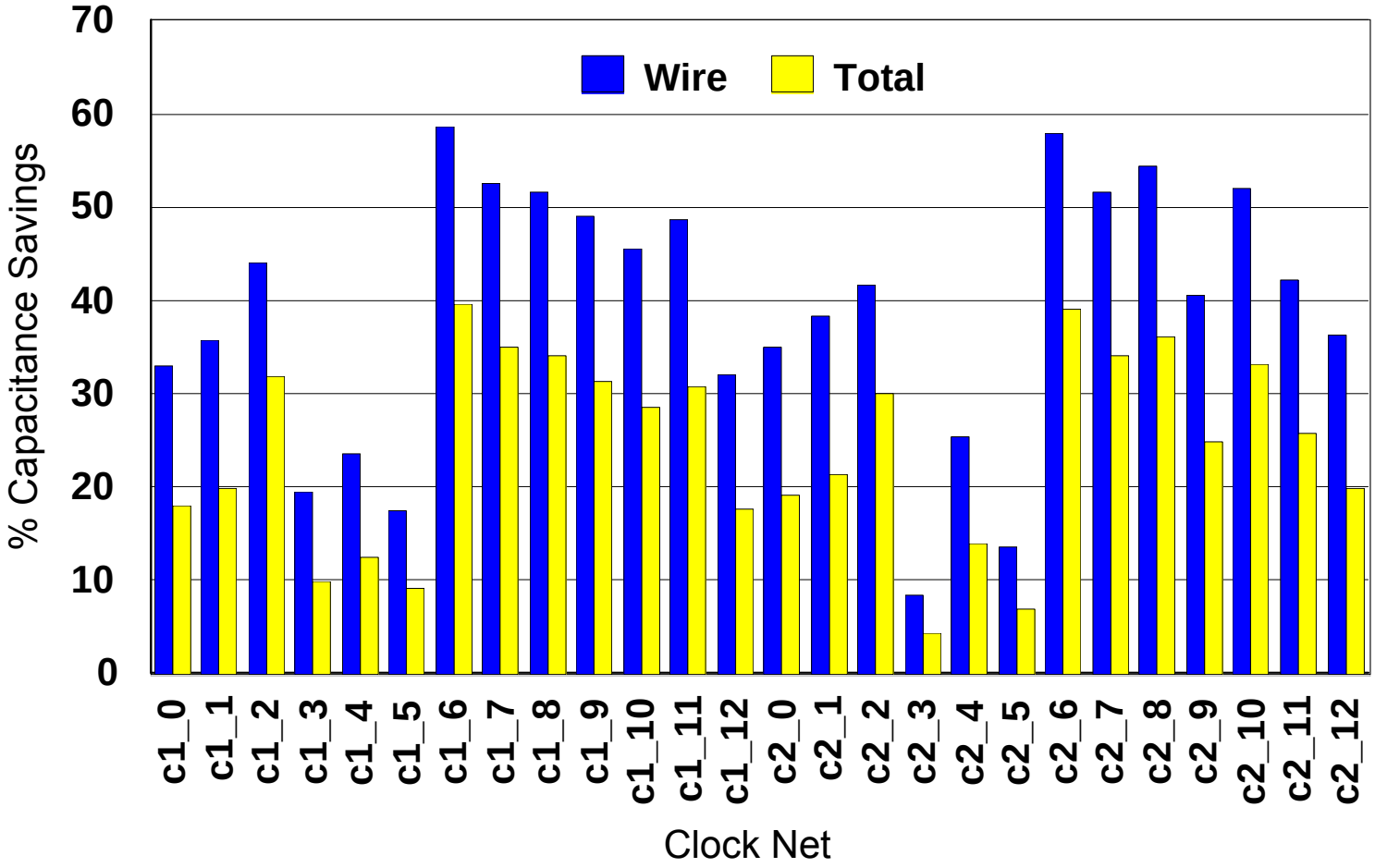
- Minimize the capacitive loading on local clock buffers by clustering latches around them
  - Tradeoff between latch placement flexibility and clock power savings
  - Reduction in clock skew between capturing and launching latch compensates for loss in latch placement flexibility





R. Puri (IBM), DAC, 2005

Clock Power Savings

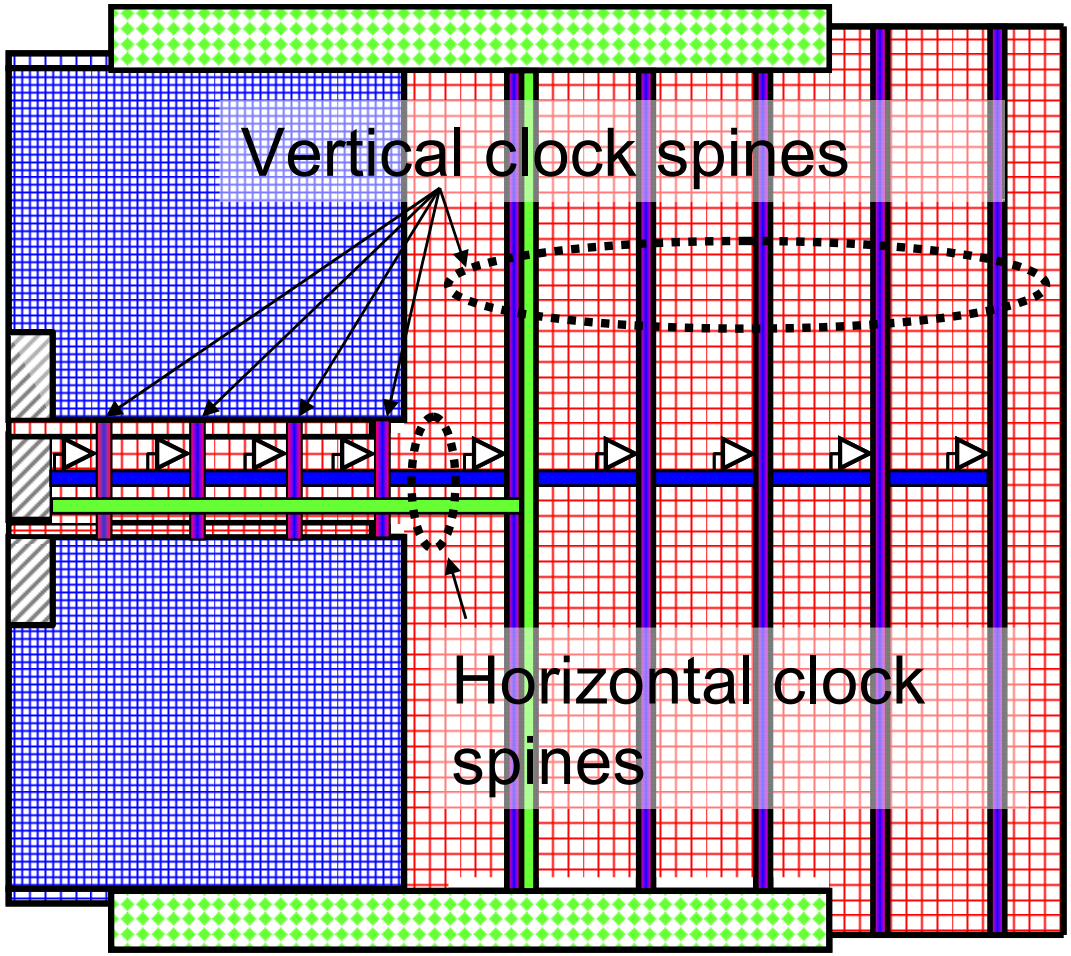


Latch clustering reduces local clock net capacitance by 25%

R. Puri (IBM), DAC, 2005

Multiple Clock Grid Types

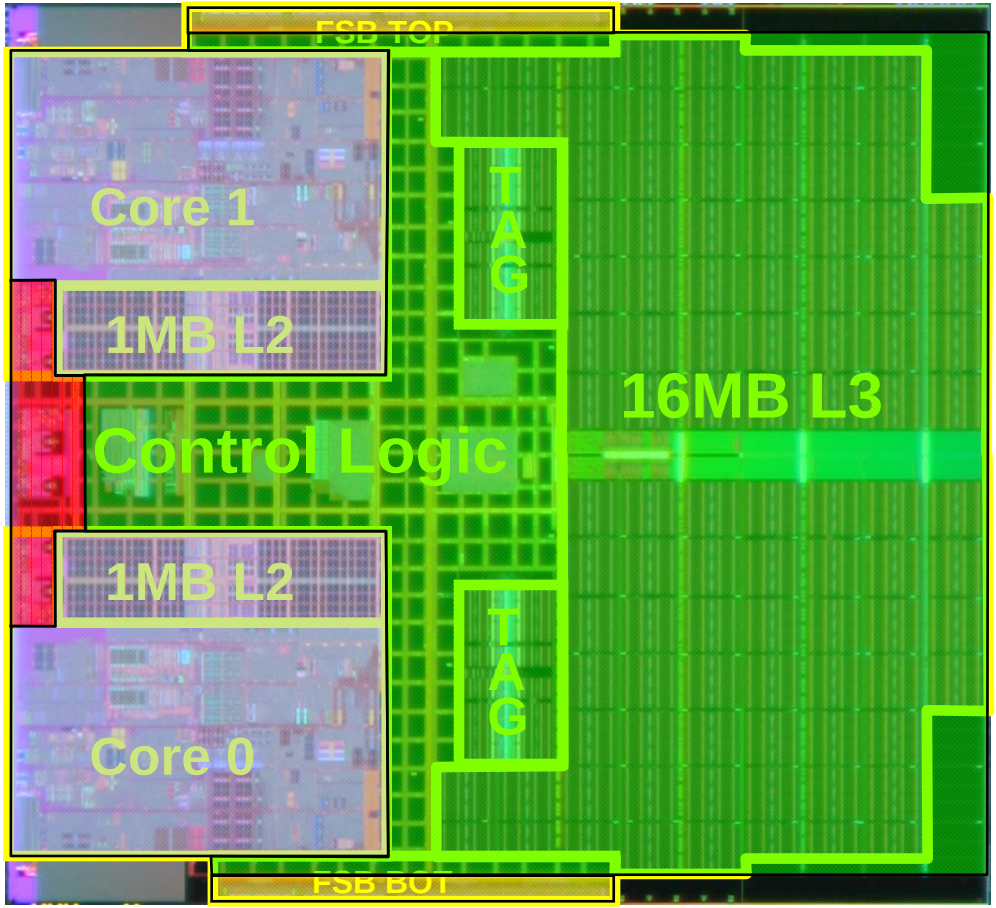
- PLL (Clock Generator)
- Core dense MCLK grid
- Un-Core ZCLK grid
- Un-Core pre-global ZCLK spine
- Un-Core sparse SCLK grid
- Un-Core pre-global MCLK spine
- De-skew buffer



Match the clock grid to the underlying circuits to reduce clock loading

S. Tam, ISSCC 2006

Multiple Voltage Domains

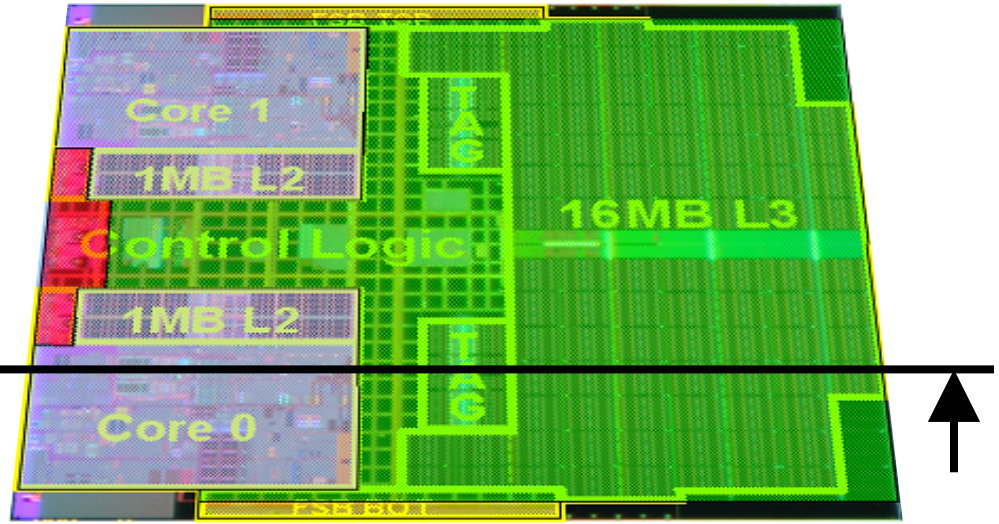


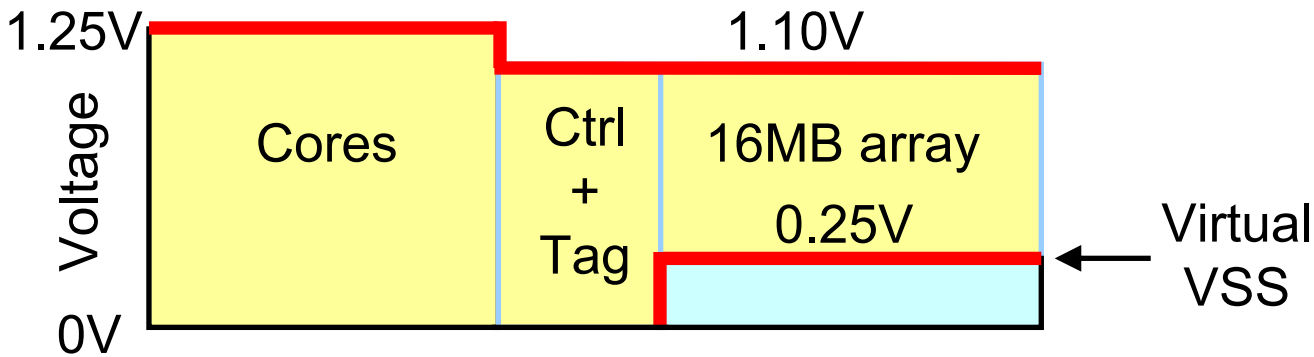
Legend: Core PLL Uncore I/O

S. Rusu, ISSCC 2006

Voltage Profile

Voltage Profile
Cut Line

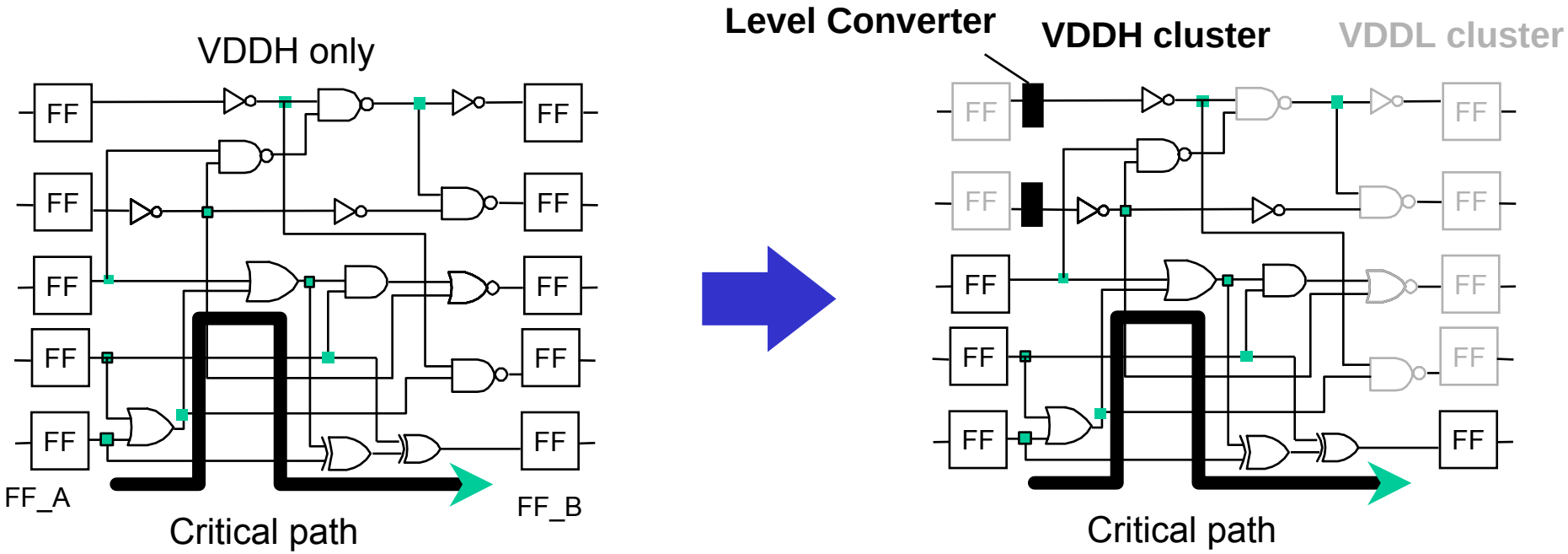




Operate each block at the lowest possible voltage

Cell-Level Dual-VDD Approach

- Use reduced voltage *VDDL* in non-critical paths
- Apply original voltage *VDDH* to timing critical paths

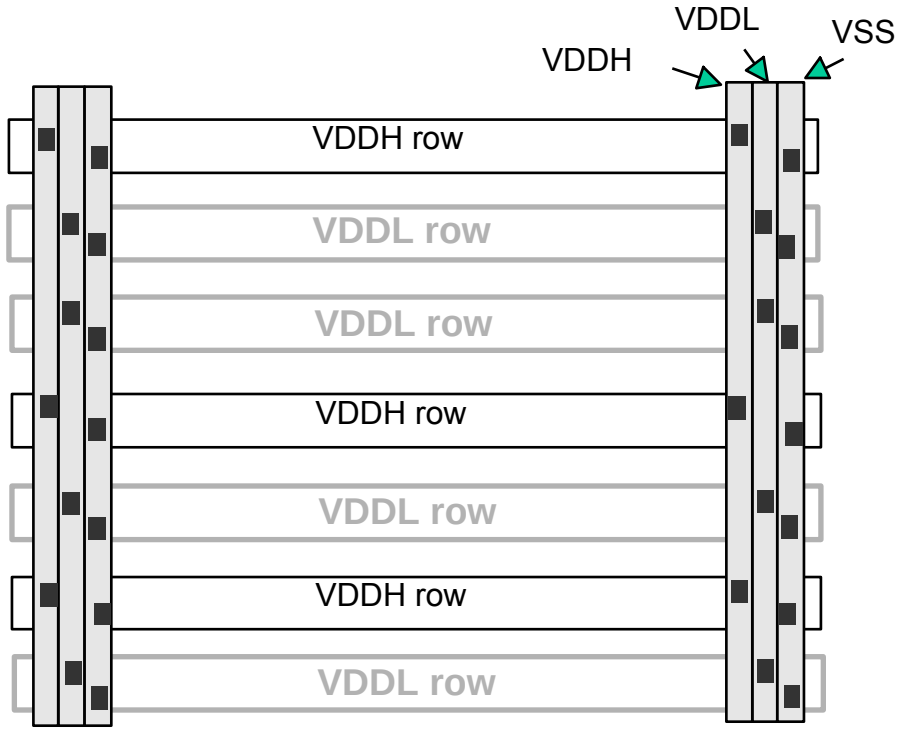


- Challenges: minimize # of level converters by clustering

K. Usami (Toshiba), DAC 1998

Cell-Level Dual-VDD (cont)

Row-by-Row layout architecture with Dual-V_{DD}



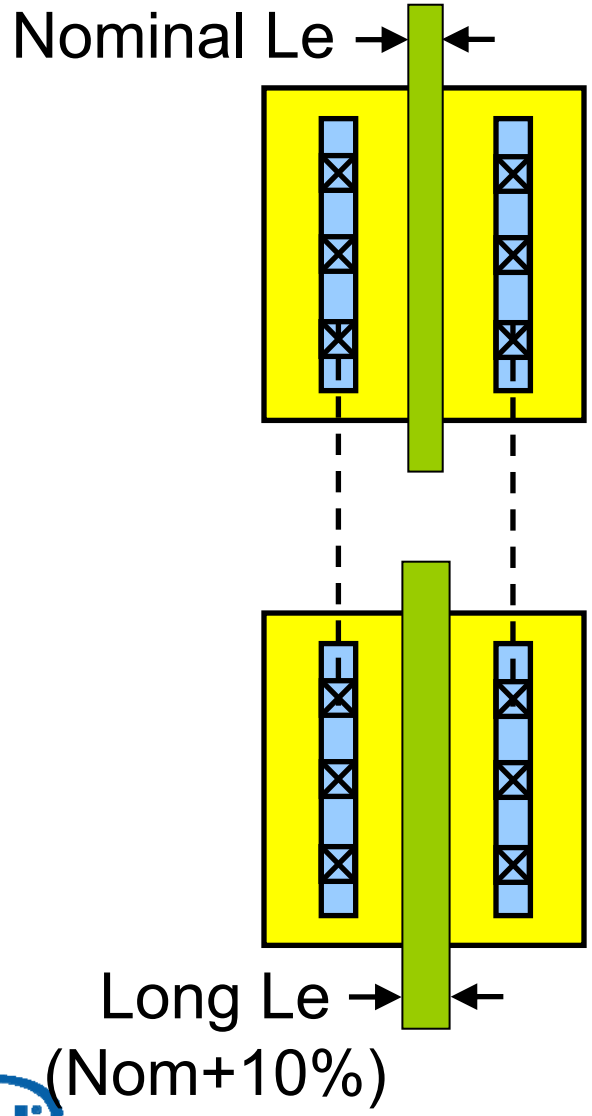
- P&R tool determines which rows should be *VDDL*
- Clock tree synthesis using *VDDL* clock buffers
- 25% power reduction demonstrated on MPEG4 video codec core

K. Usami (Toshiba), DAC 1998

Outline

- Power components and trends
- Active power reduction techniques
- Leakage reduction techniques
  - Long channel devices
  - High-Vt transistors
  - Body bias
  - Transistor stacking
  - Cache leakage reduction
  - Power gating and multiple supplies
- Power management methods
- Summary

Long-Le Transistors

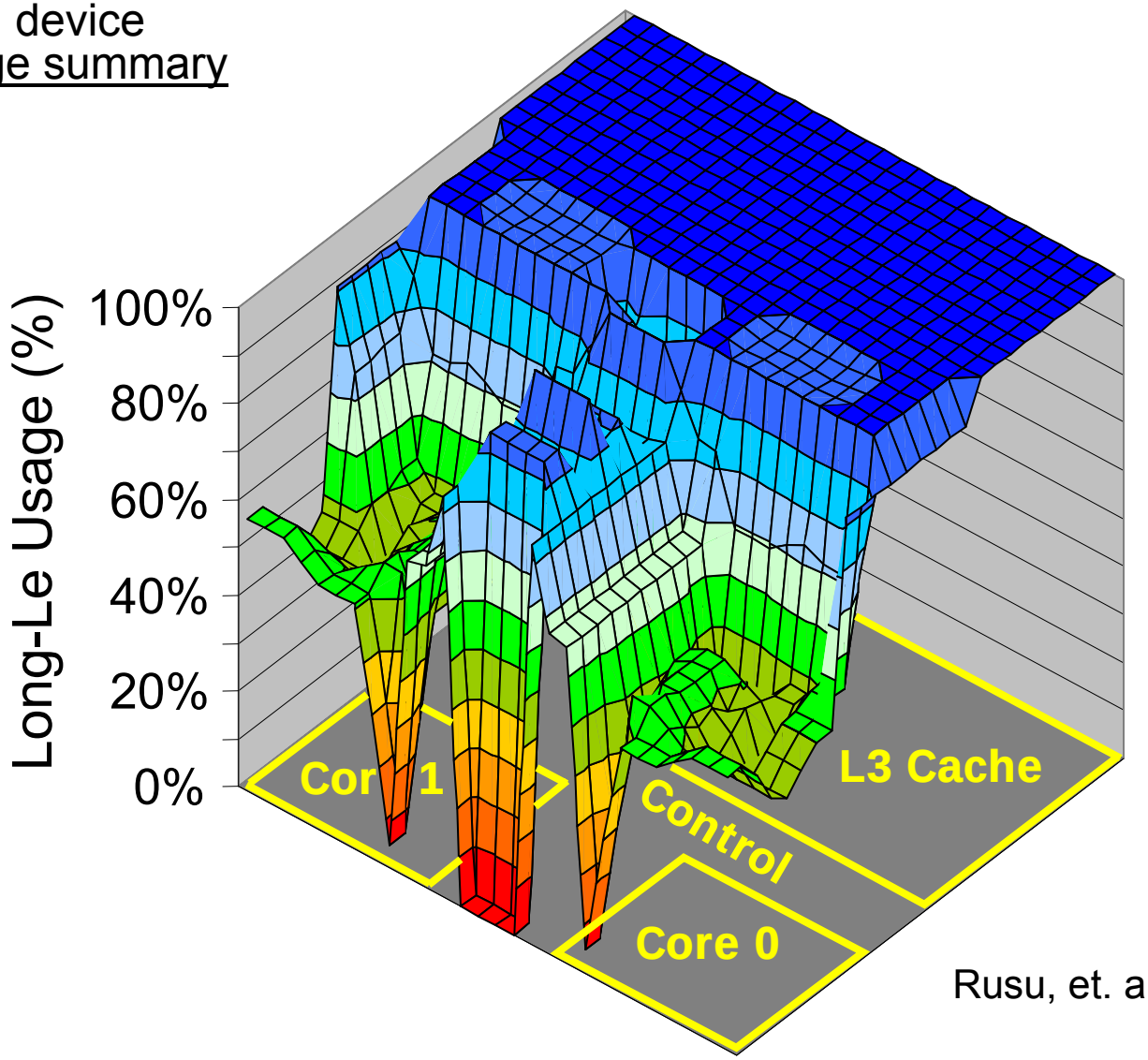


- All transistors can be either nominal or long-Le
- Most library cells are available in both flavors
- Long-Le transistors are ~10% slower, but have 3x lower leakage
- All paths with timing slack use long-Le transistors
- Initial design uses only long channel devices

Rusu, et. al, ISSCC 2006

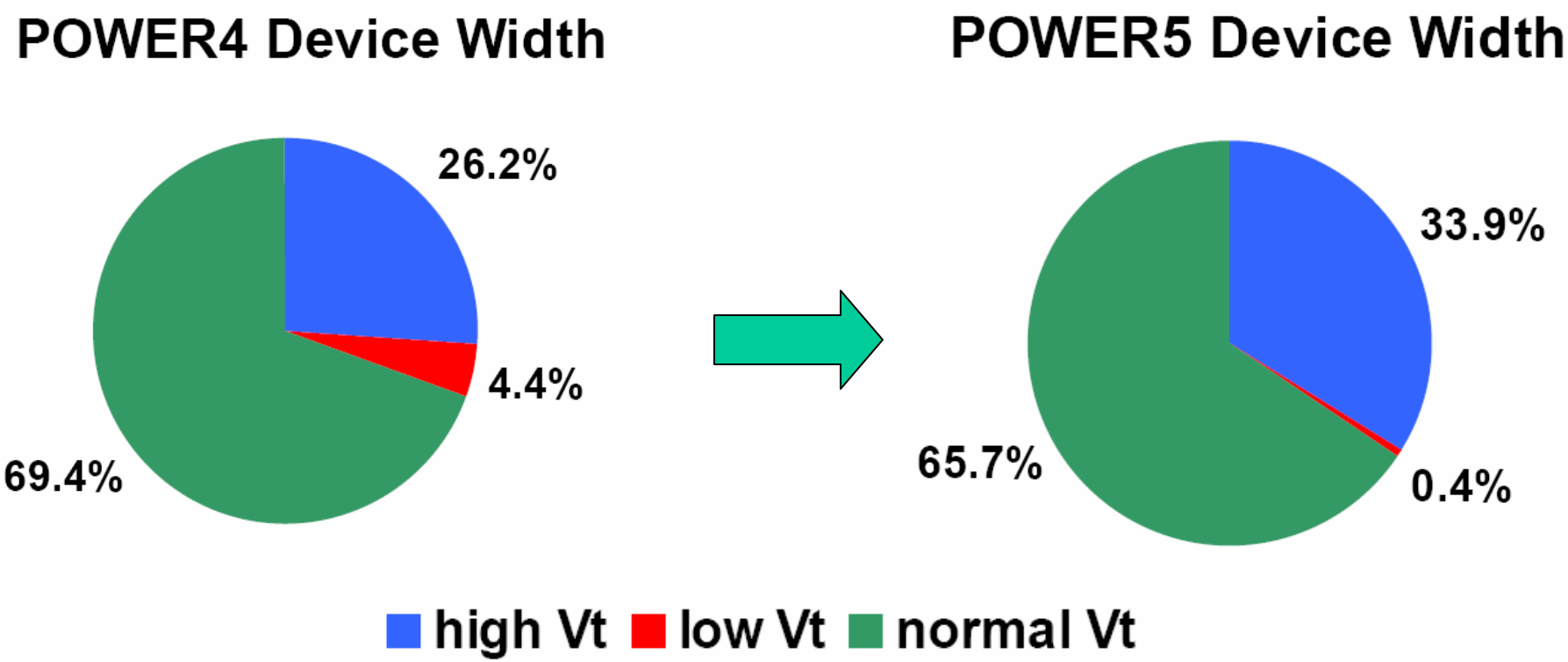
Long-Le Transistors Usage

Long channel device
<u>average usage summary</u>
Cores 54%
Uncore 76%
Cache 100%



Rusu, et. al, ISSCC 2006

High-Vt Transistors

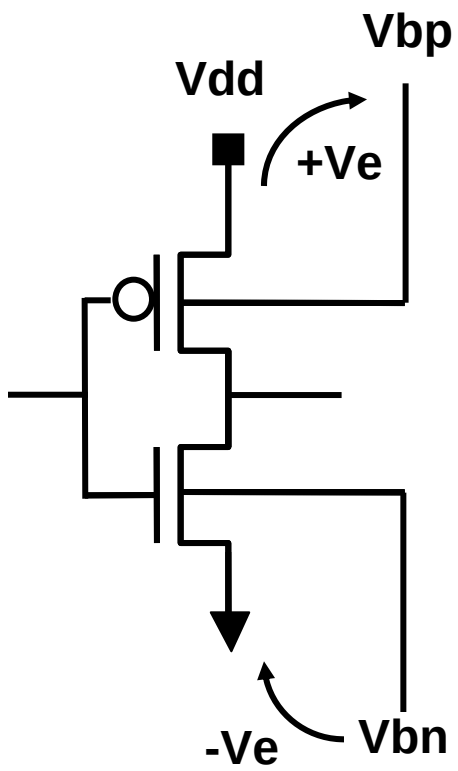


IBM's Power Processors are leveraging triple Vt process option

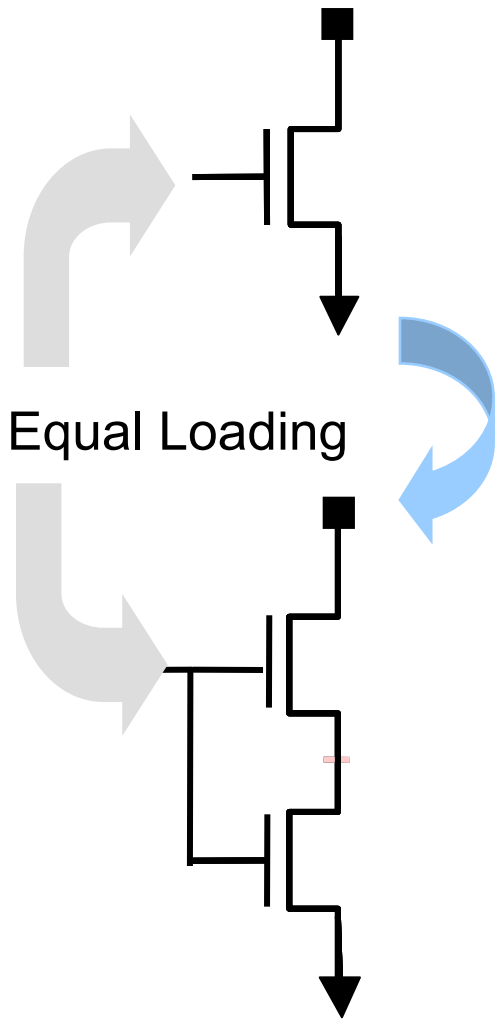
Clabes, et al. (IBM), ISSCC 2004

Leakage Reduction Circuit Techniques

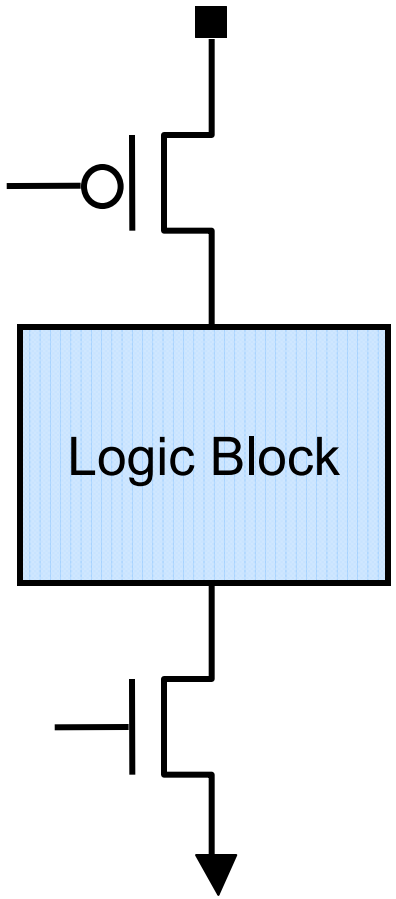
Body Bias



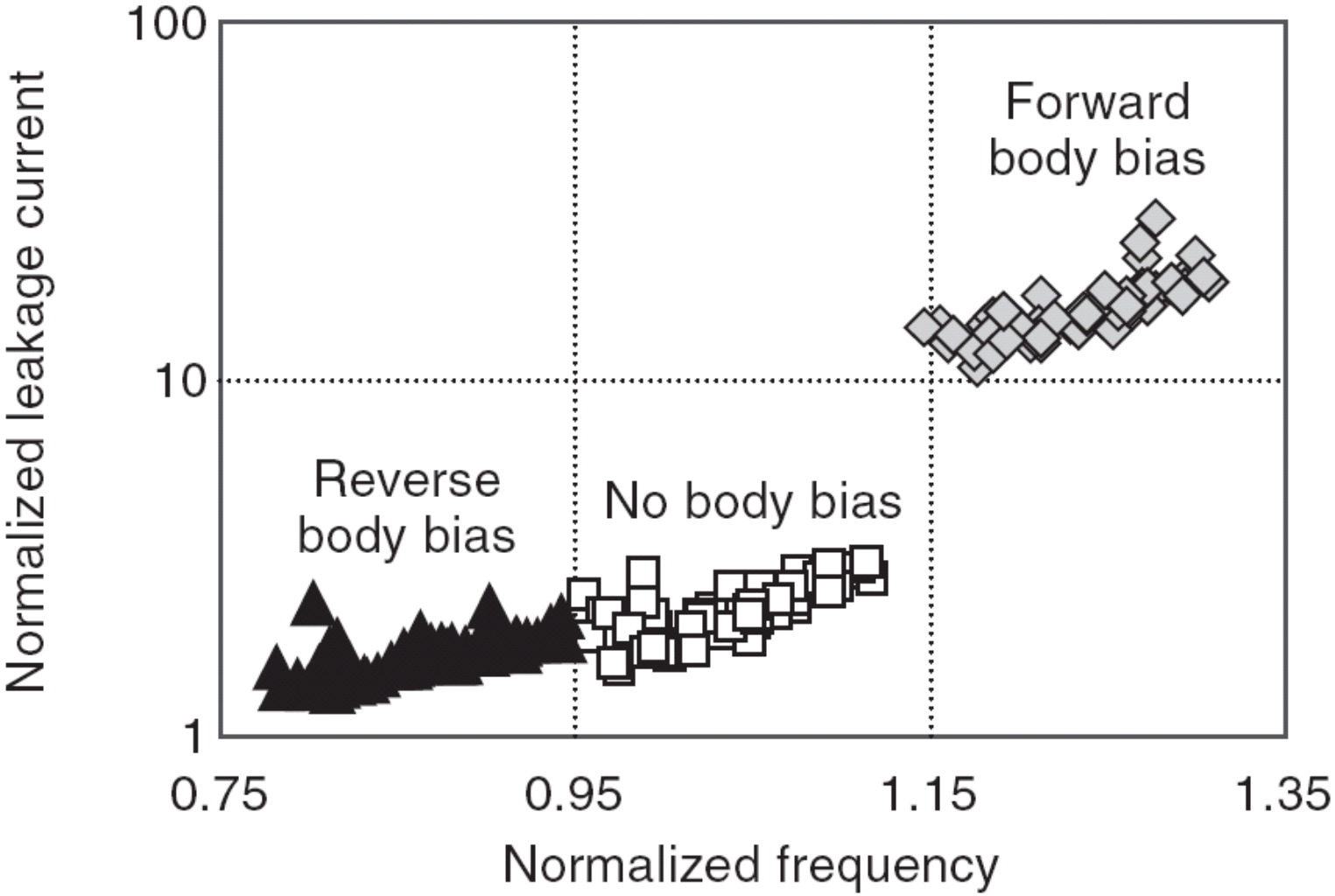
Stack Effect



Sleep Transistor

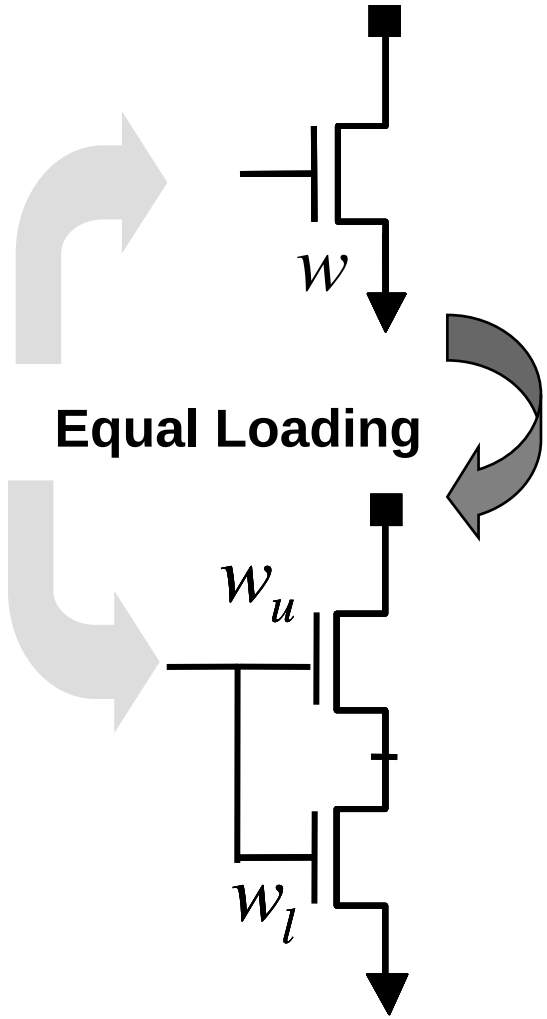


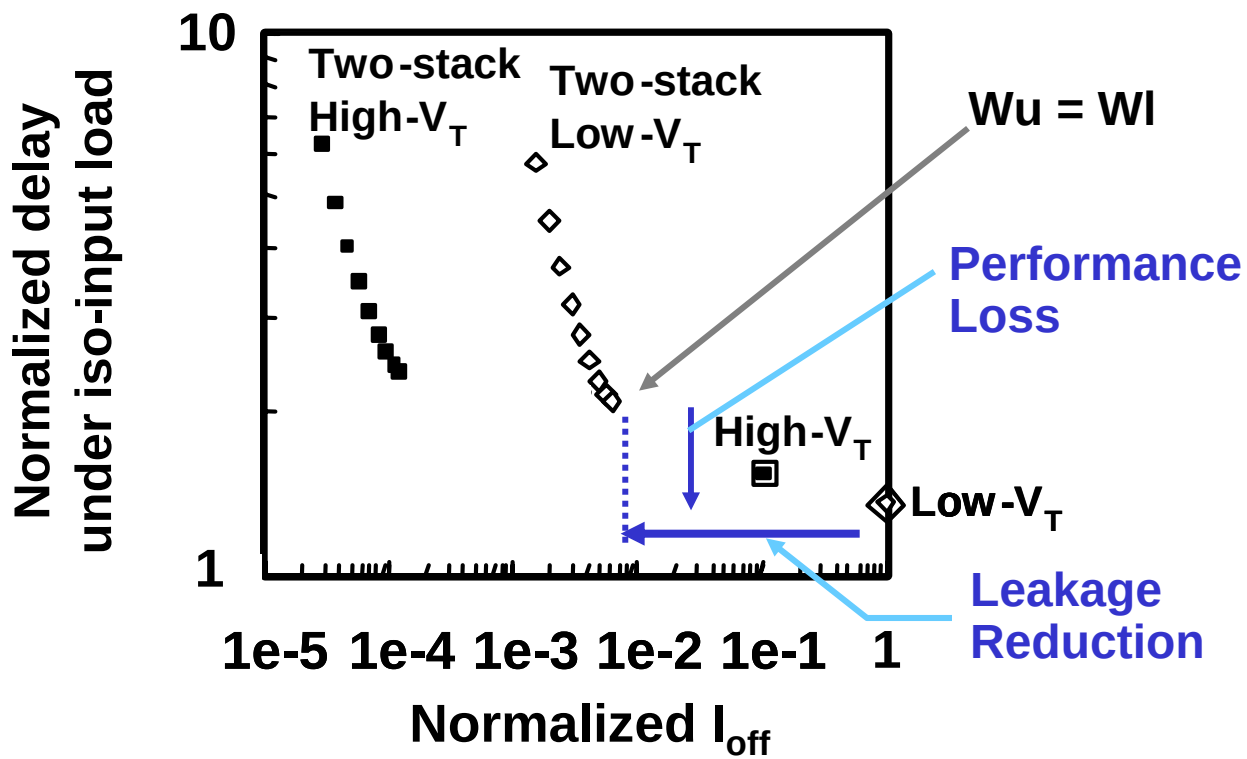
Body Bias Leakage Reduction



Keshavarzi, et al., D&TC 2002

Stack Forcing

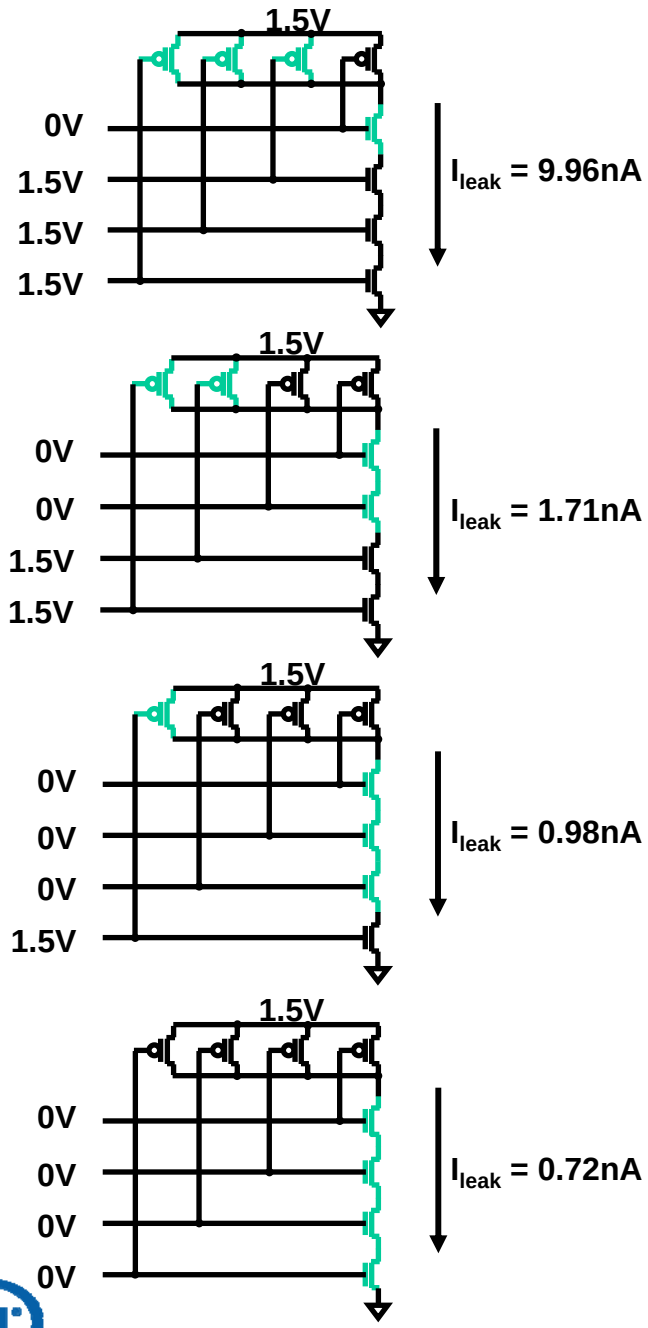




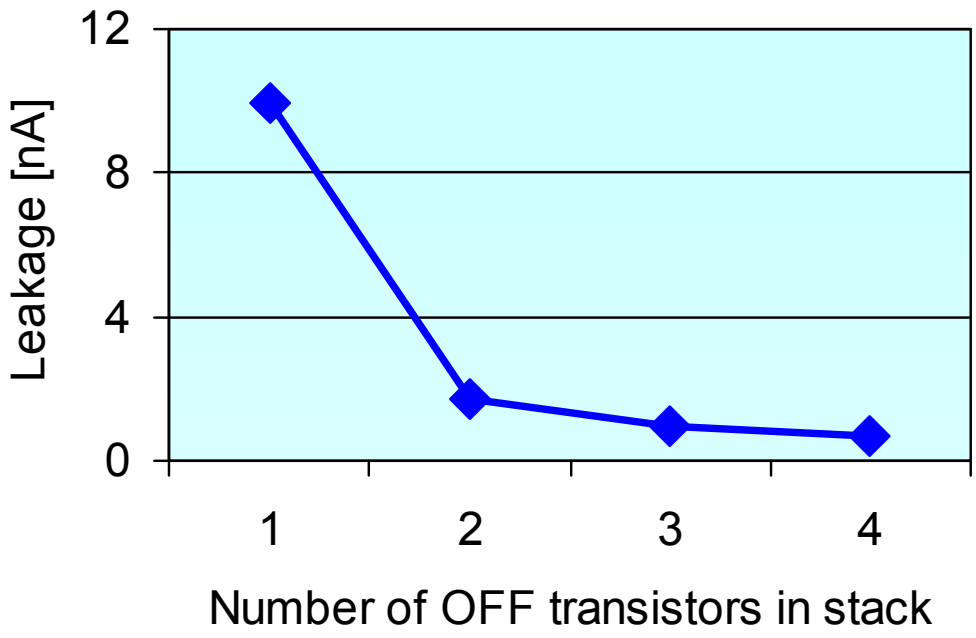
- Force one transistor into a two transistor stack with the same input load
- Can be applied to gates with timing slack
- Trade-off between transistor leakage and speed

Narendra, et al, ISLPED 2001

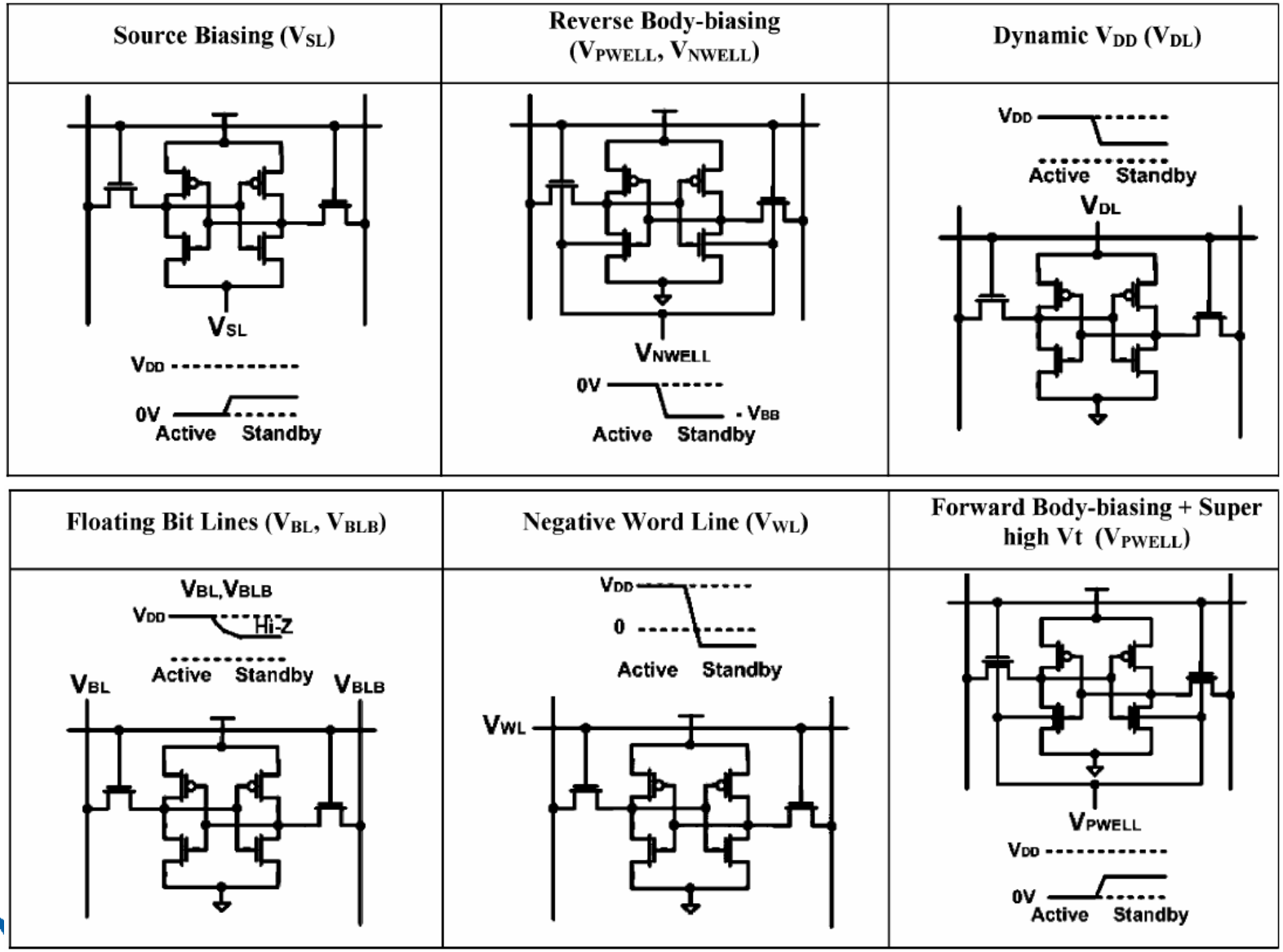
Natural Stacks



- Leakage reduced significantly when two transistors are off in a stack
- Educate circuit designers, monitor average stacking factor

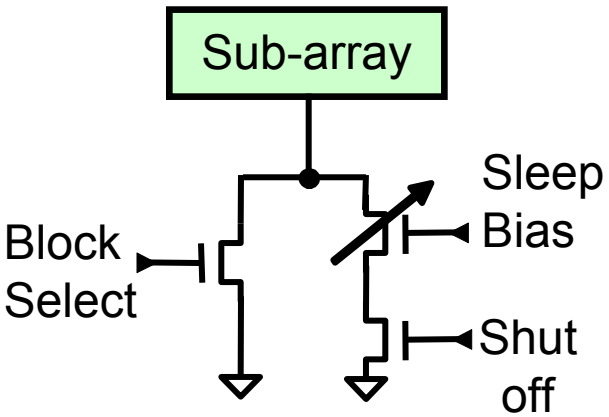


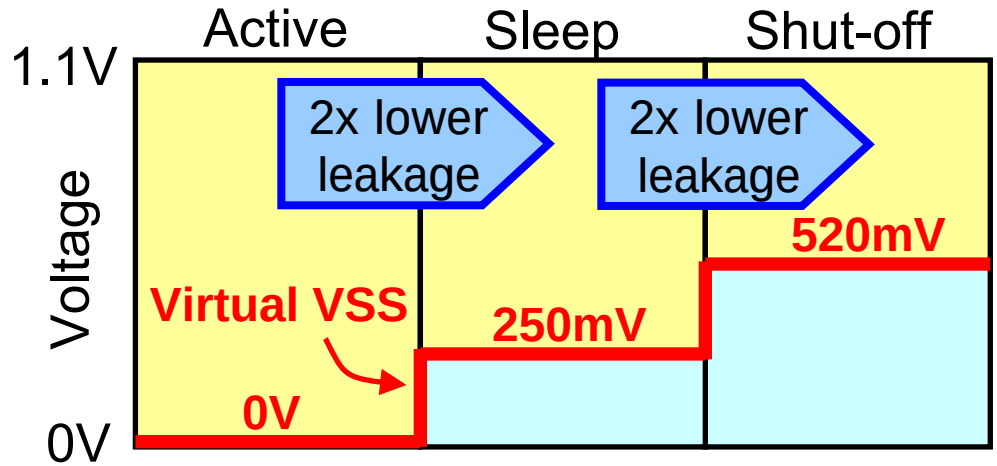
Cache Leakage Reduction Techniques

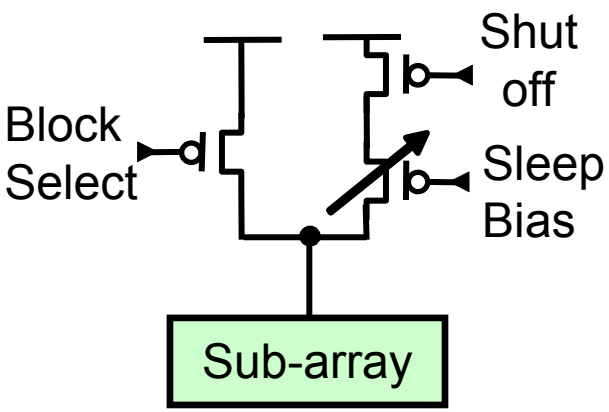


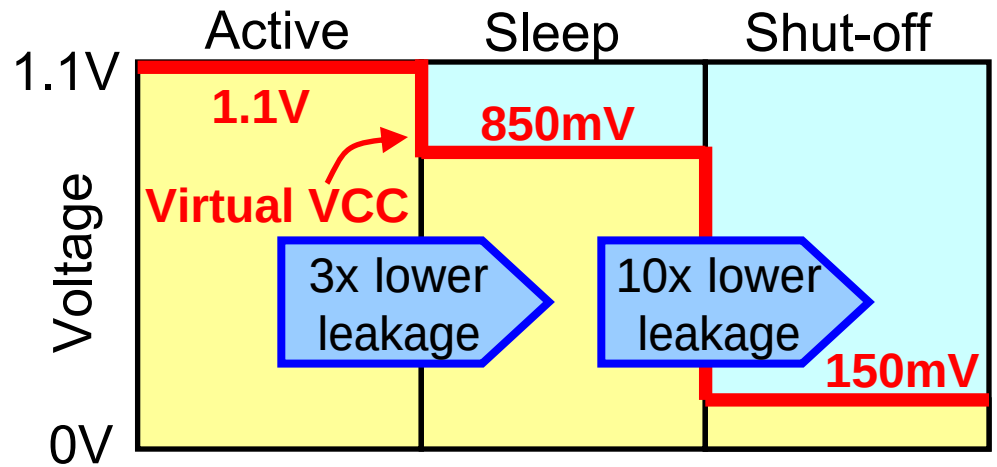
[Kim, et al., IEEE Trans. VLSI Sys.,2005]

Cache Sleep and Shut-off Modes







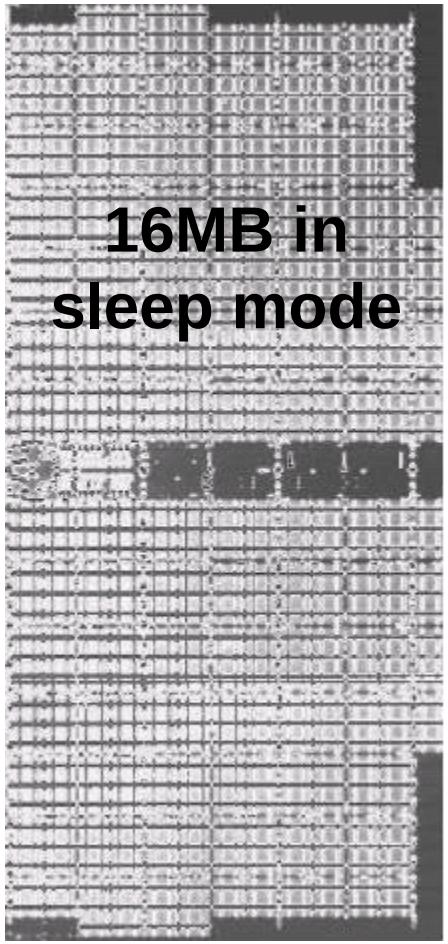


PMOS reduces junction leakage and has better shut-off

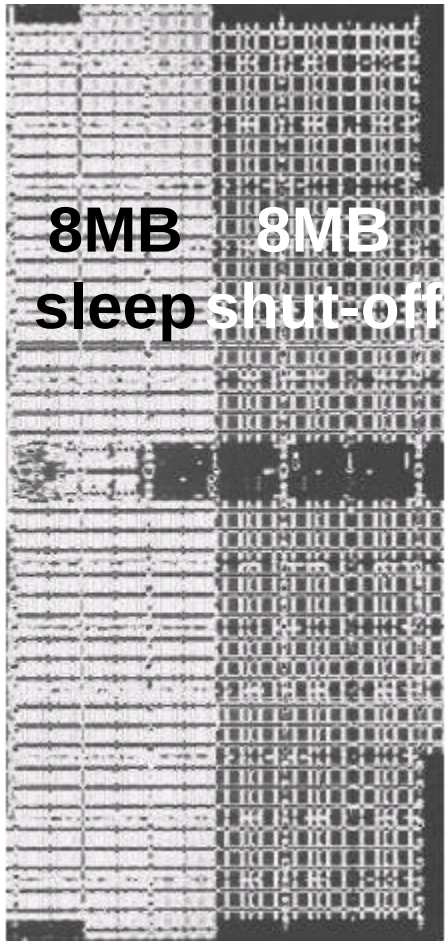
S. Rusu, et. al, US Pat App 20070005999, 6/2005

Leakage Shut-off Infrared Images

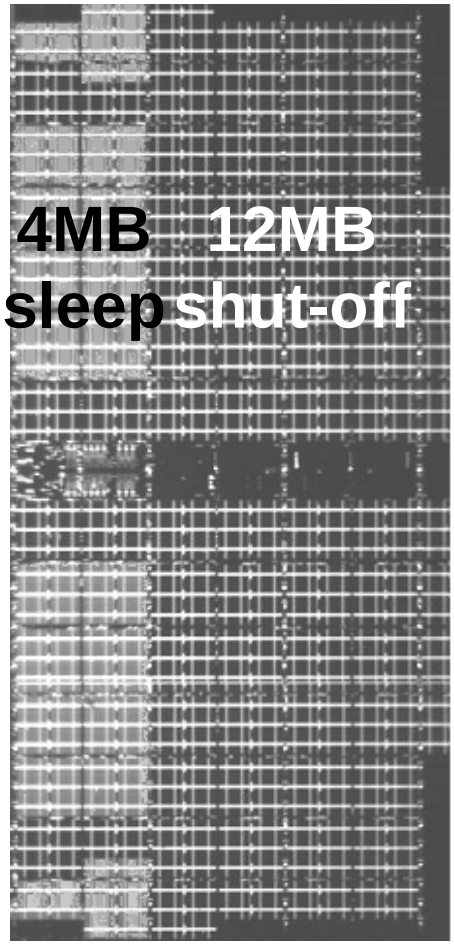
16MB part



8MB part



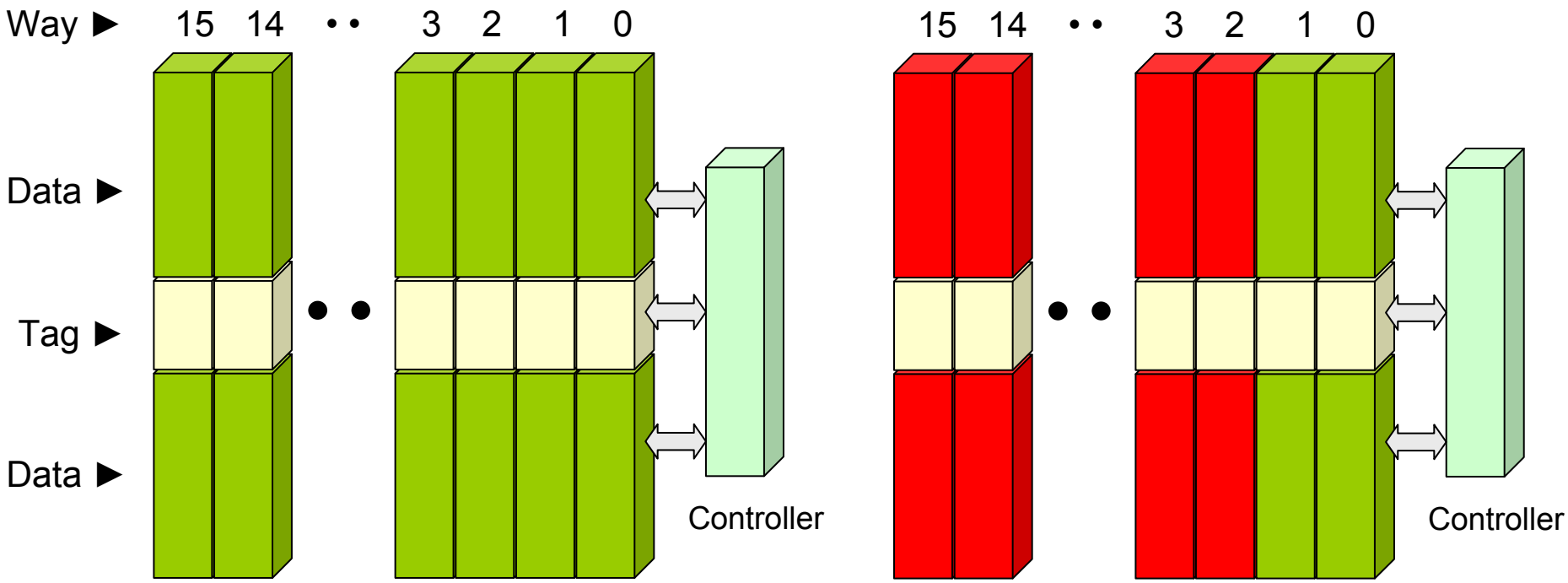
4MB part



Leakage reduction ► 3W (8MB) 5W (4MB)

Rusu, et al., ISSCC 2006

Cache Dynamic Shut-off



Normal Operation

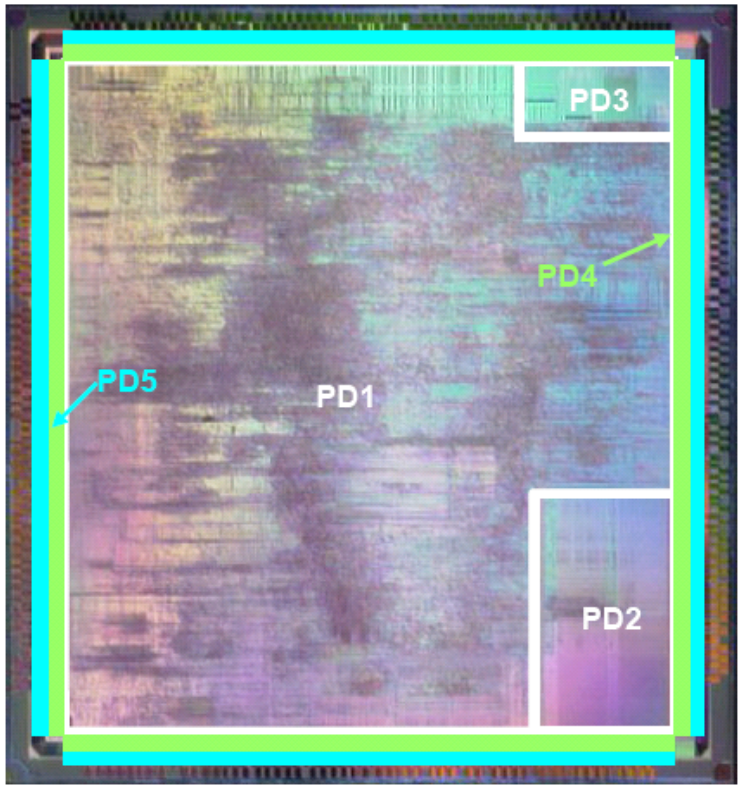
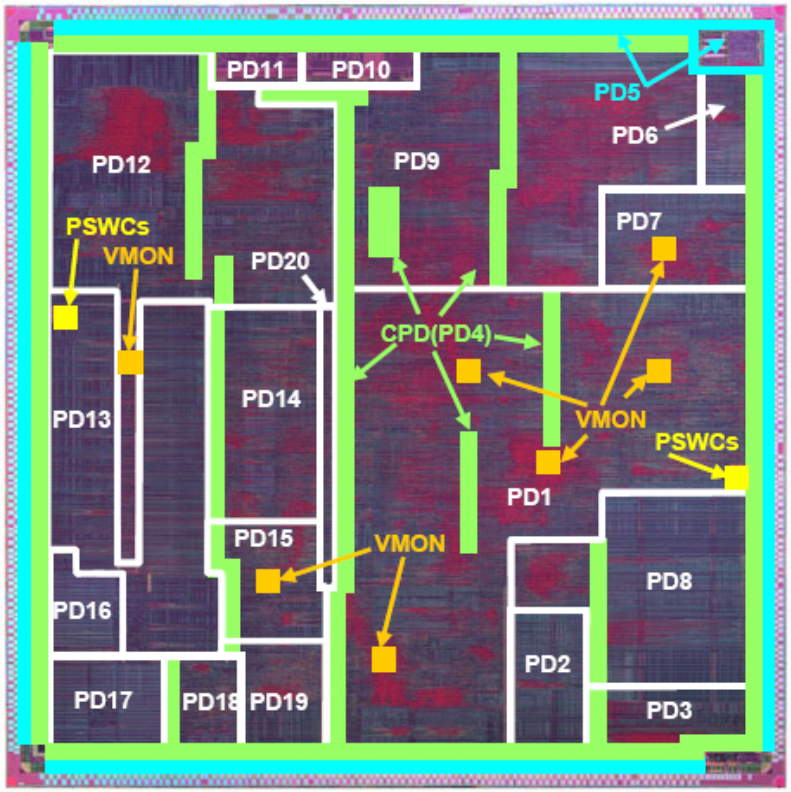
- In the full-load state, all 16 ways are enabled (green)

Cache-by-Demand Operation

- Under idle or low-load states, cache ways are dynamically flushed out and put in shut-off mode (red)

Sakran, et al., ISSCC 2007

Multiple Power Domains

Implementation example of conventional power domain.	Implementation example of dozens of power domains.
	

Kanno, et. al, ISSCC-2006
Hitachi + Renesas

Power Domains Activation Examples

Power off in white area

Kanno, et. al, ISSCC-2006
Hitachi + Renesas

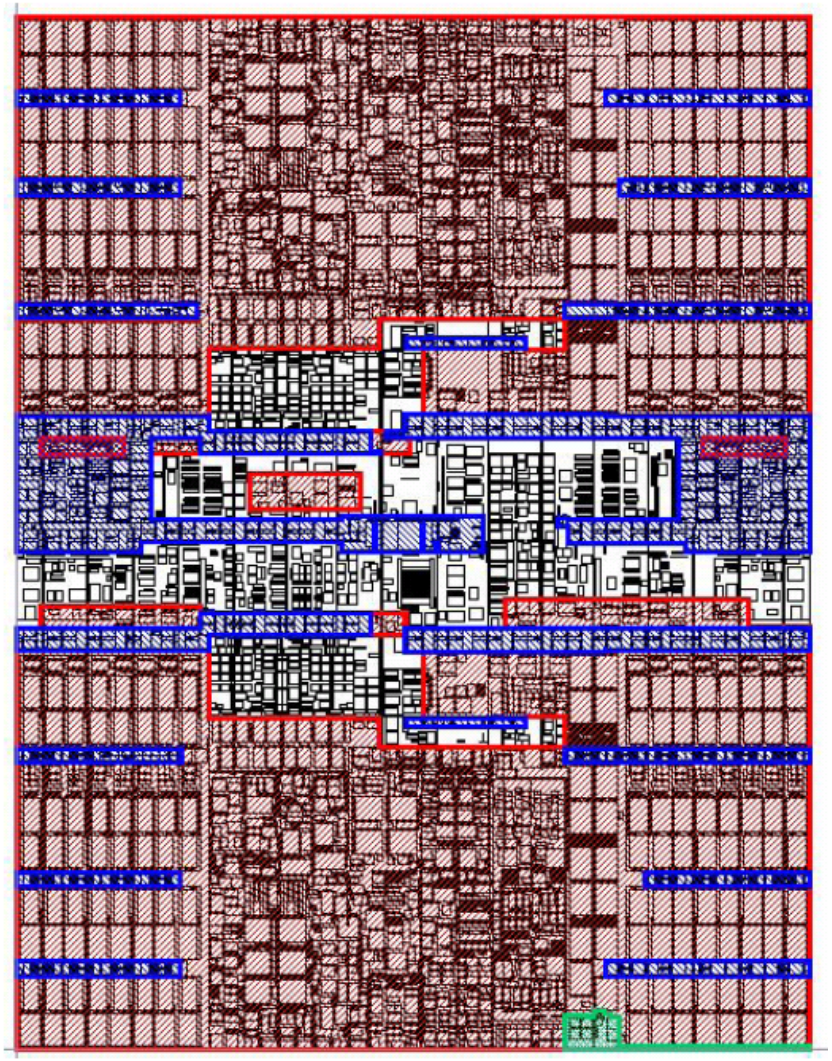
IBM POWER6 Voltage Domains

- POWER6 infrastructure contains 4 voltage domains

Rail	Purpose	Plot Color
VDD	logic	all
VCS	Array	Red
VIO	IO, PLL, MC	Blue
VSB	Powerup	Green

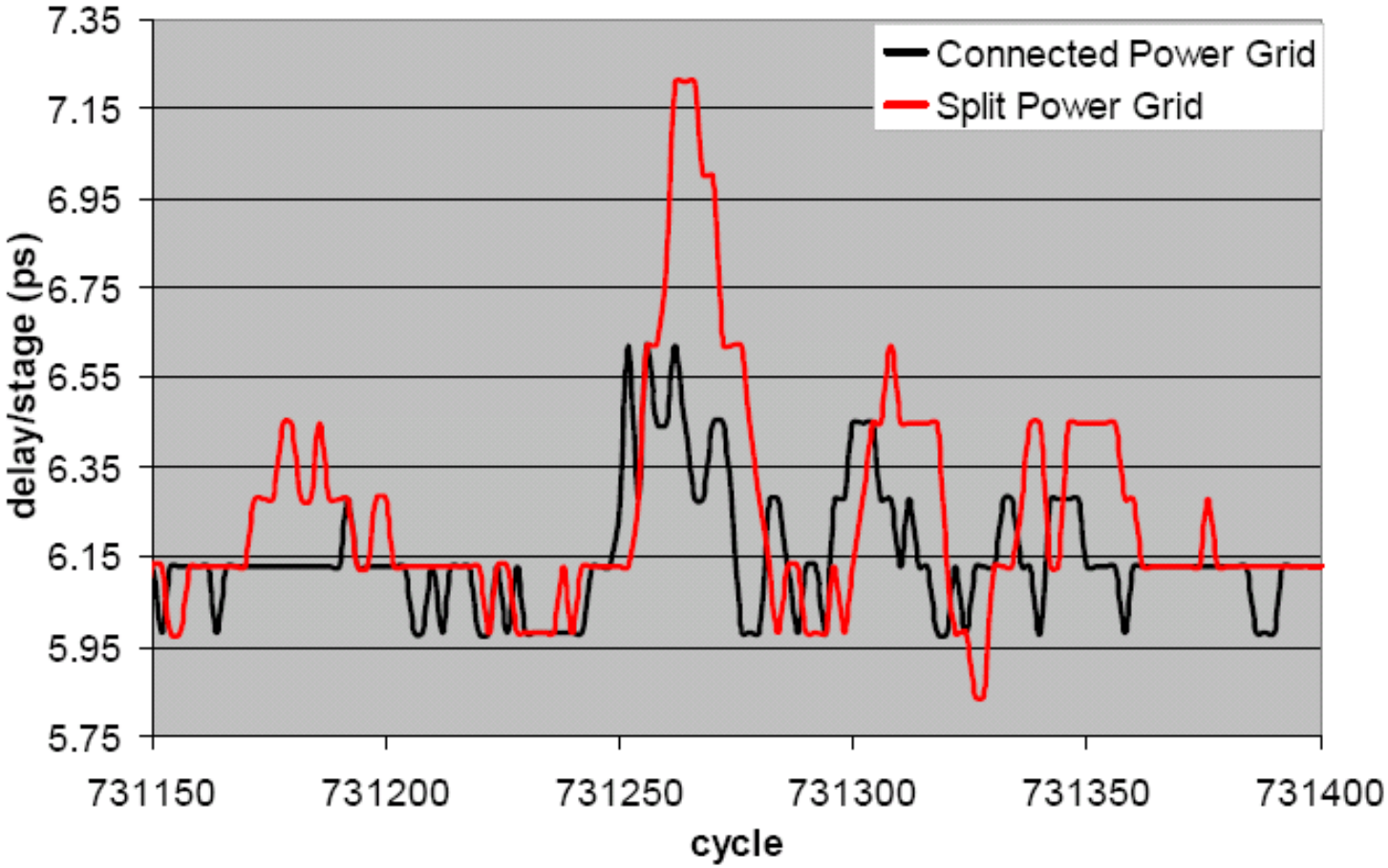
- Multi-rail power grid defined based on macro current requirements & iterative IR analysis of each rail.

- Voltage domain of macros and global signals explictly specified in RTL and validated by checking tools.



J. Friedrich (IBM), ISSCC 2007

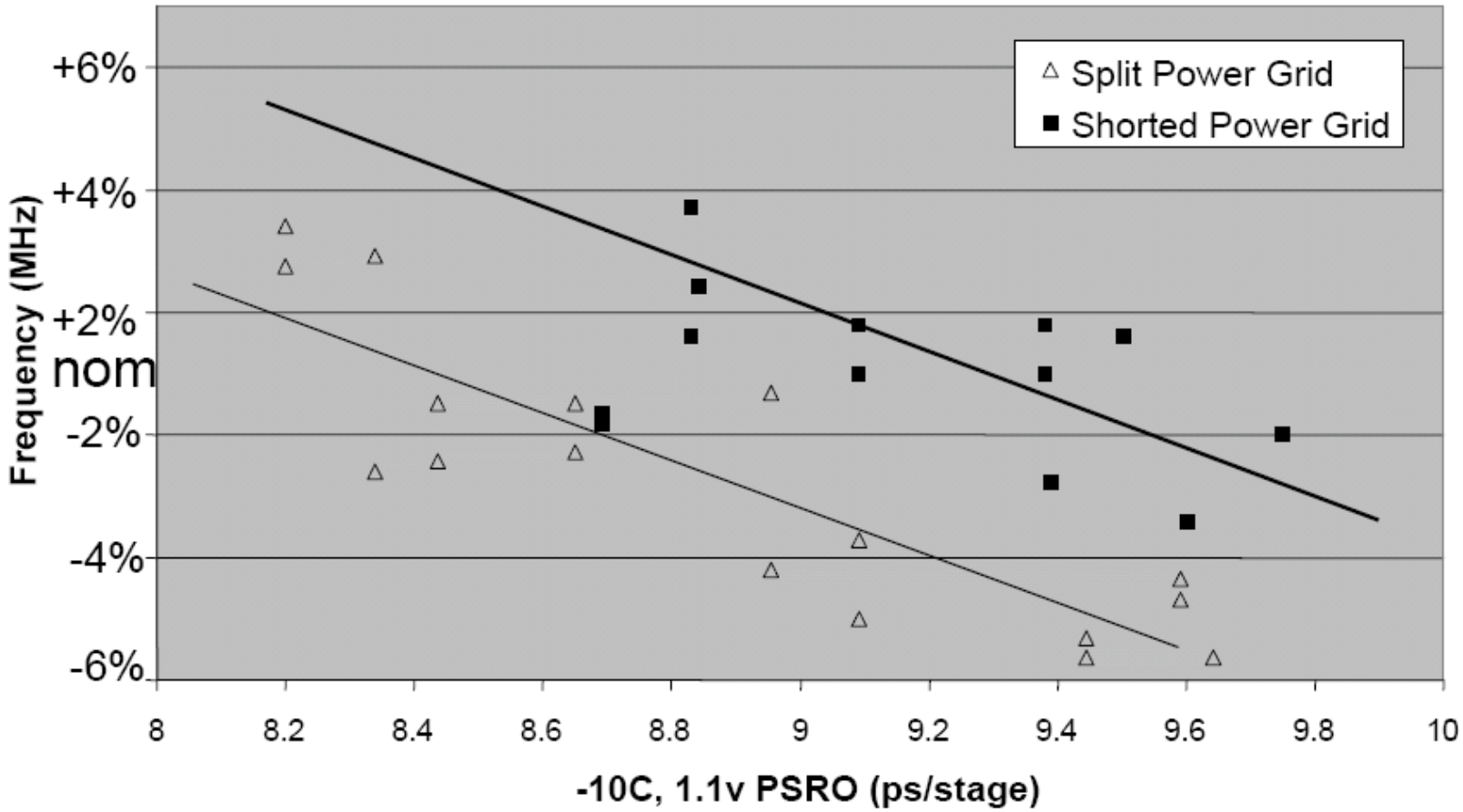
Split vs. Connected Power Grid



- Chips are roughly same process speed
- 17% to 7% droop by connecting power grids

N. James (IBM), ISSCC 2007

Split vs. Connected Core Supplies



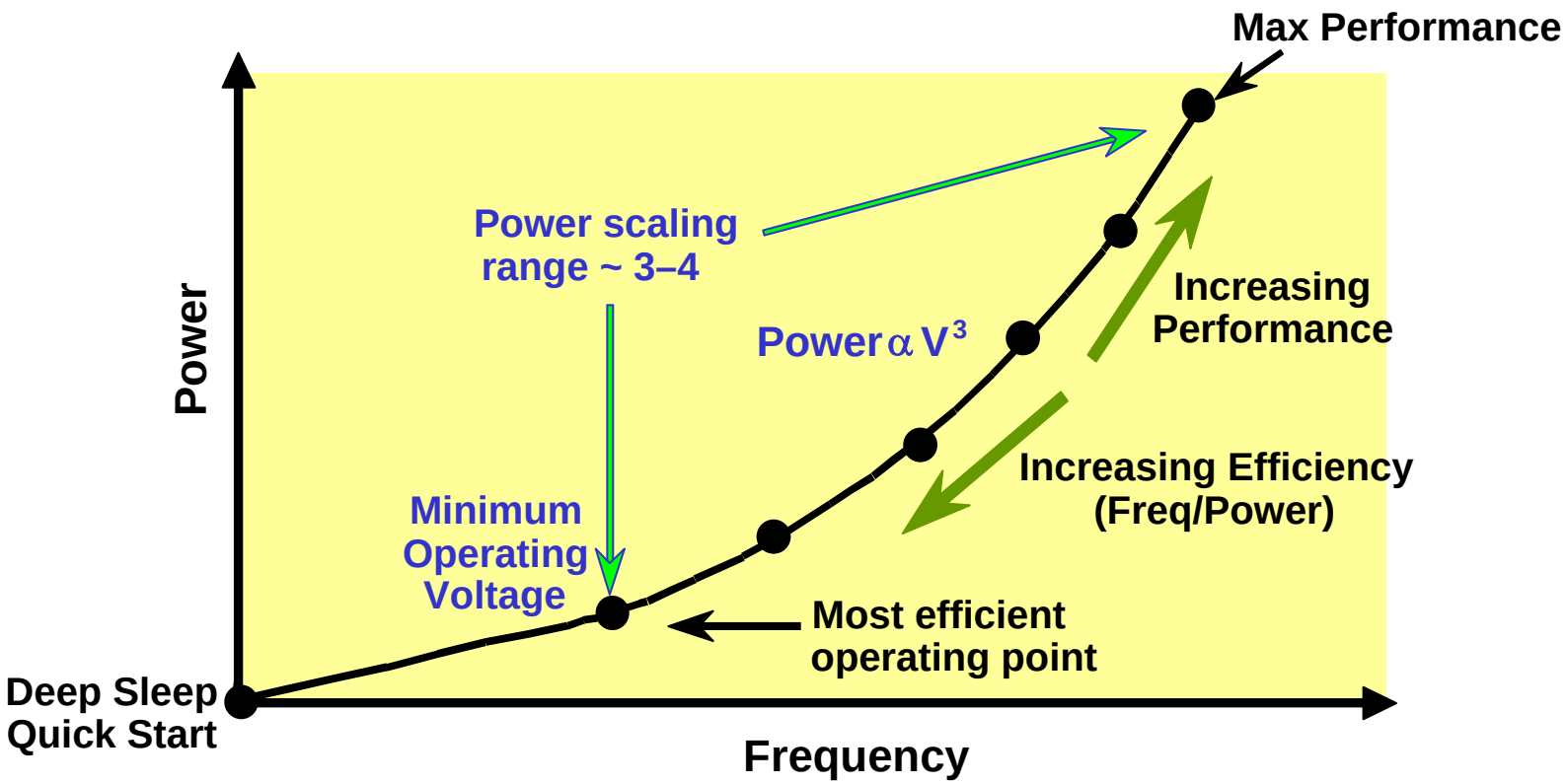
- Normalized to Process Sensitive Ring Oscillator (PSRO), the Fmax is ~5-10% higher on chips with connected core power grids

N. James (IBM), ISSCC 2007

Outline

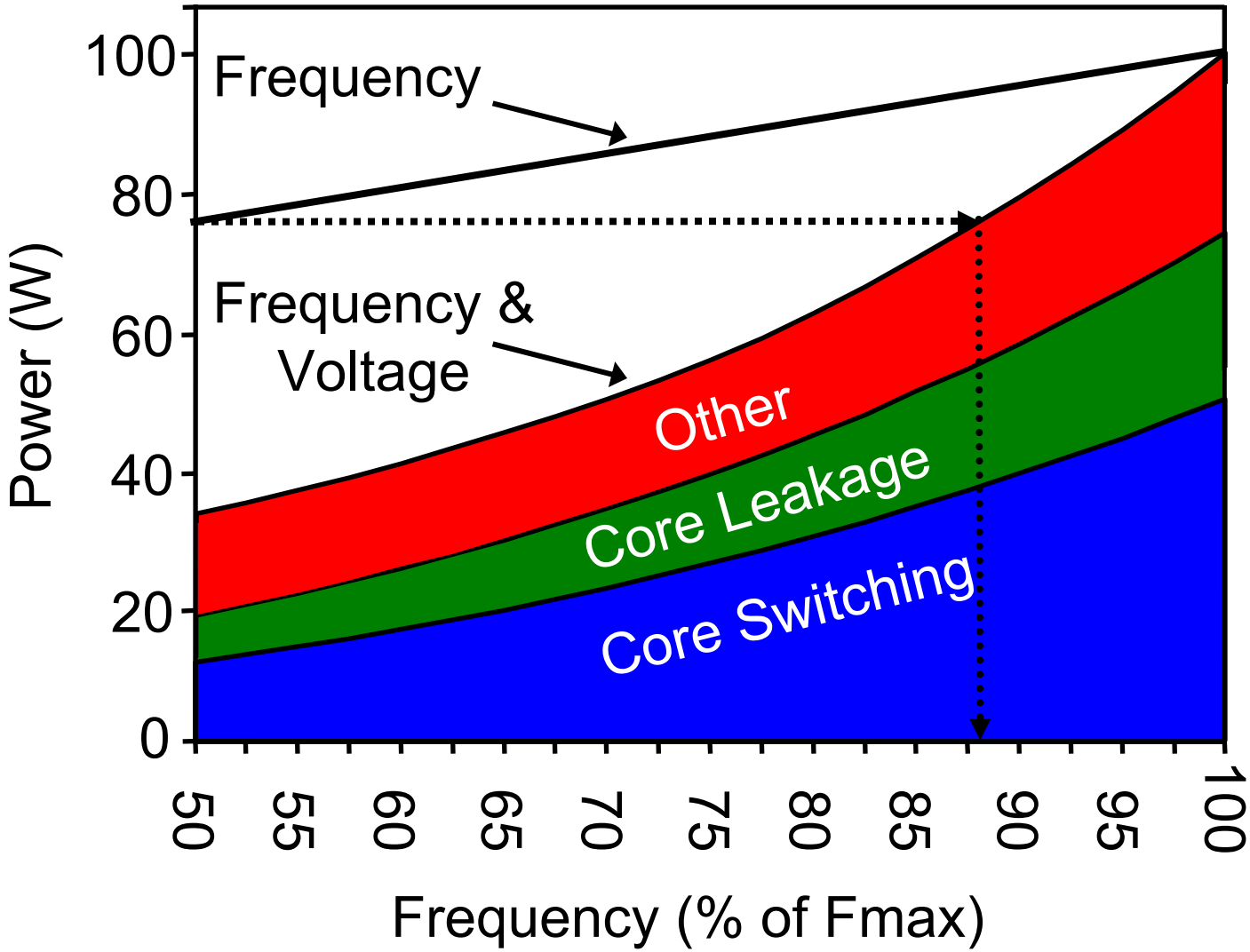
- Power components and trends
- Active power reduction techniques
- Leakage reduction techniques
- Power management methods
  - Voltage / Frequency Scaling
  - Deep Power Down Technology
  - Enhanced Dynamic Acceleration Technology
  - Power Throttling
  - Future Directions
- Summary

Voltage / Frequency Scaling



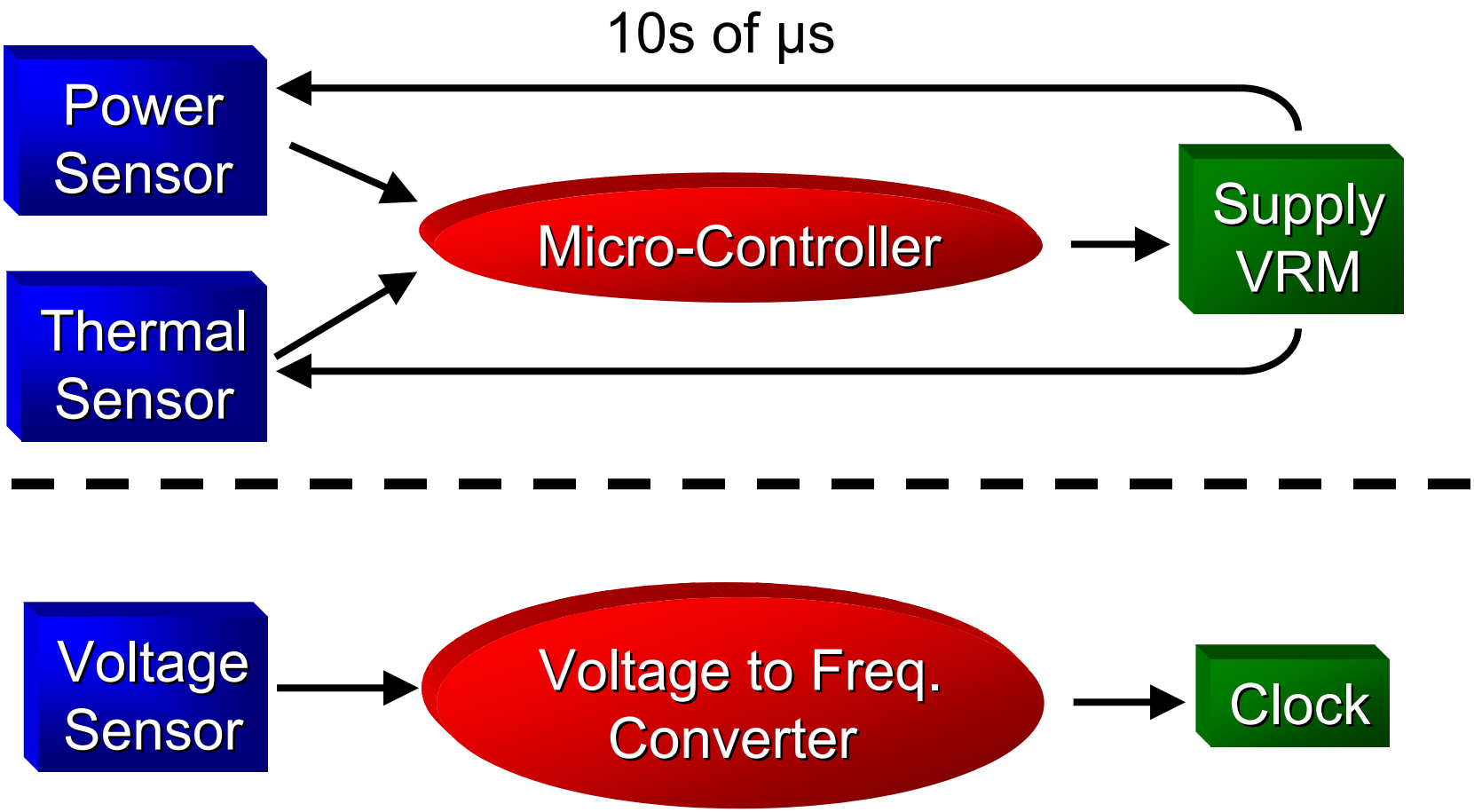
- Voltage-frequency scaling with active thermal feedback
- Multi-operating states from high performance to deep sleep
- Power management reduces average and peak power

Itanium® Processor V/F Scaling



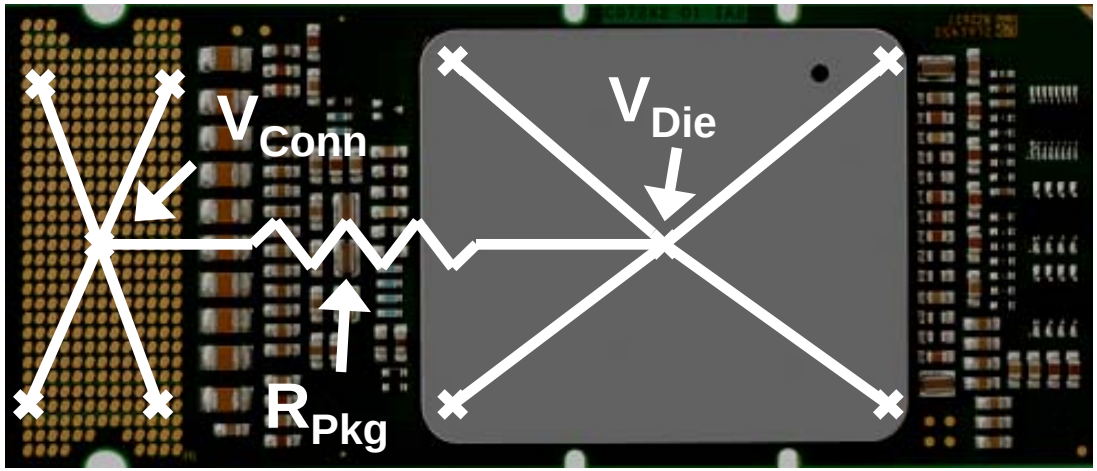
R. McGowan, ICCAD 2005

V / F Control System

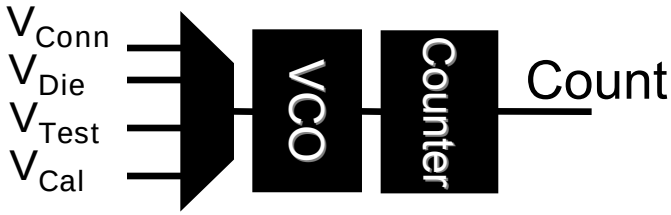


R. McGowan, ICCAD 2005

Power Measurement



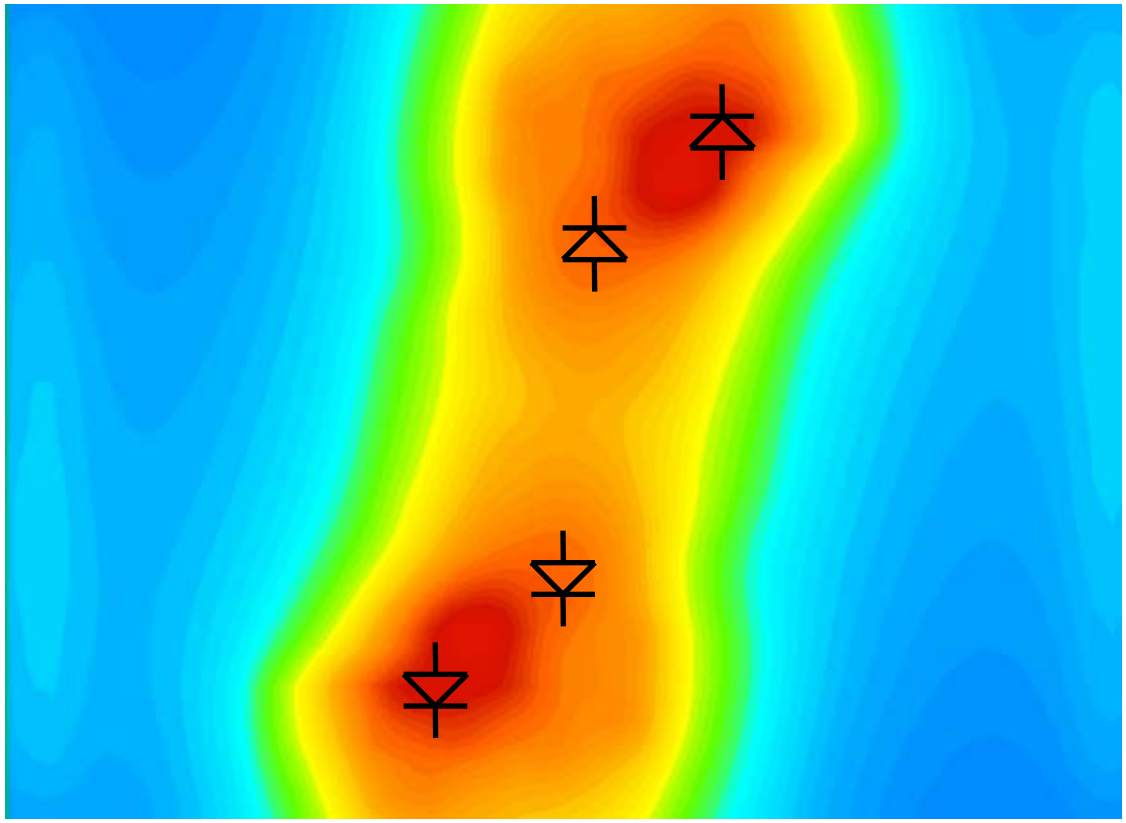
On Die Measurement



$$Power = V_{Die}\frac{(V_{Conn} - V_{Die})}{R_{Pkg}}$$

- Uses package resistance to measure power
  - Widely variable, changes with temperature
- VCO speed changes with process, temperature
- Uses a lookup table created with reference V
  - Unique to each part / operating condition
  - Linear interpolation for entries not in the table
- On die microcontroller software generates table, calibration and computes final power measurement

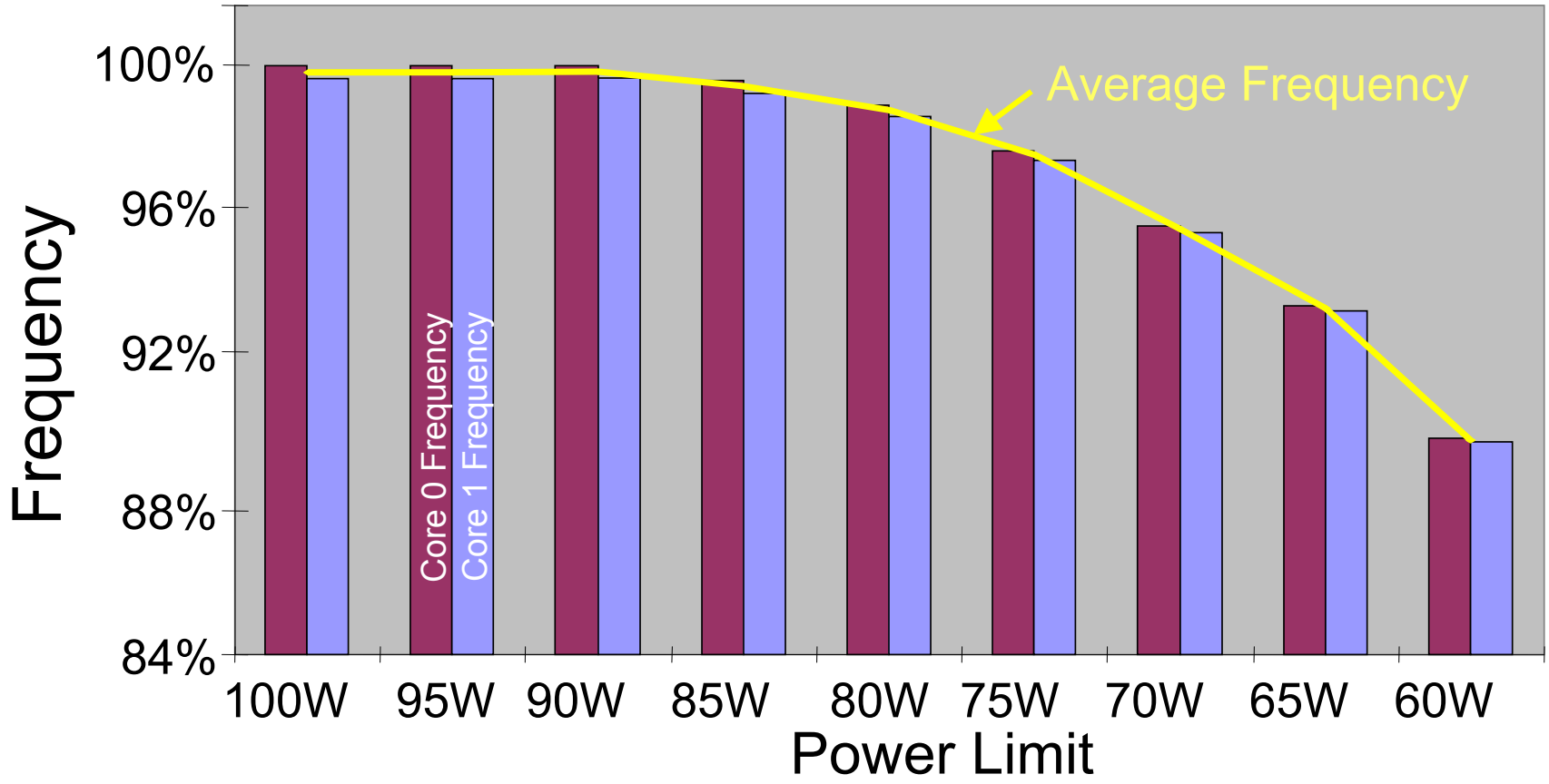
Temperature Measurement



- Two thermal sensors per core
- Mux thermal diodes into VCOs to measure temperature

R. McGowan, ICCAD 2005

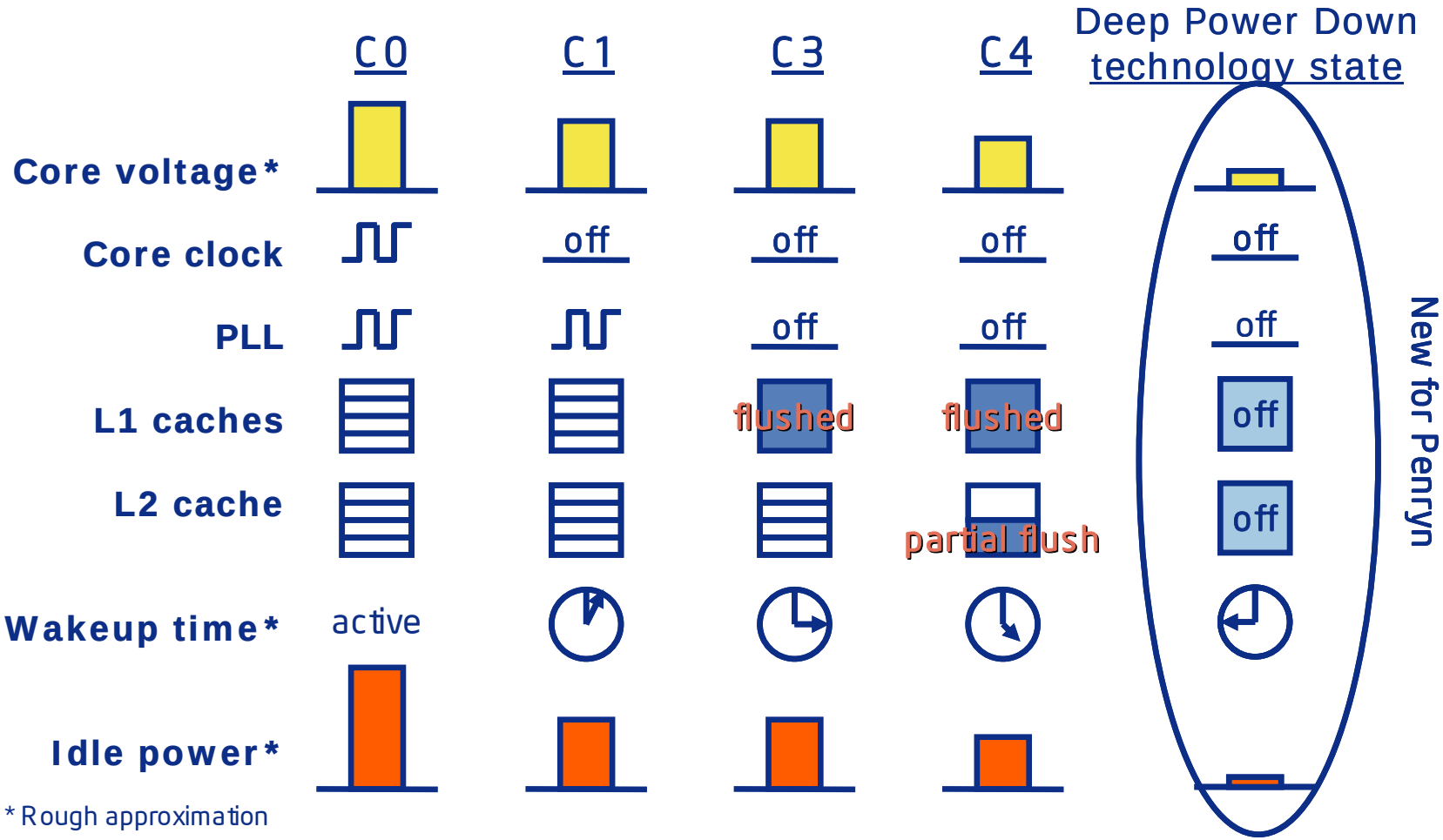
Frequency vs. Power Limit



31% power reduction for only 10% frequency drop

R. McGowan, ICCAD 2005

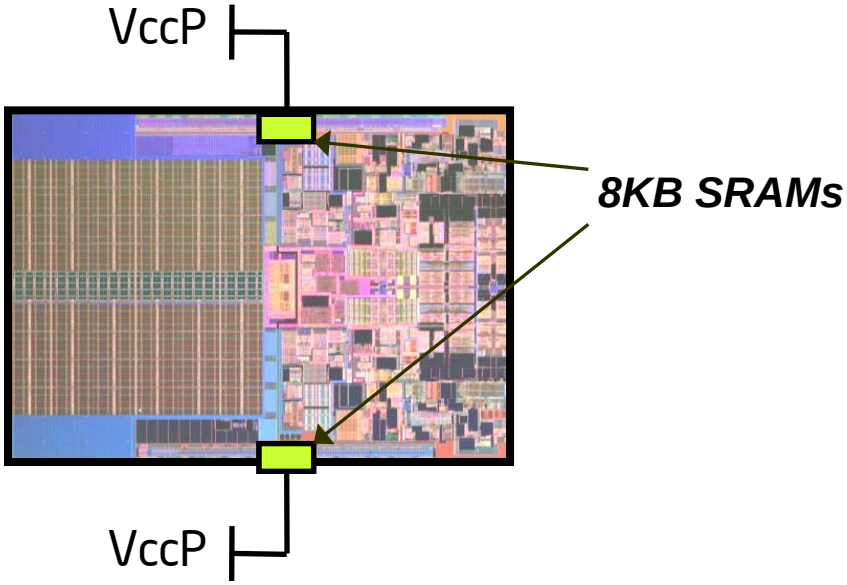
Deep Power Down Technology



DPD enables reaching lower limit of CPU idle power of 0 W

V. George, et al., Hot Chips 2007

Penryn DPD Implementation



STATE STORAGE:

– 8KB per core, ECC protected

– Powered from I/O Vcc (VccP)

STATE DEFINITION:

– What to include?

– Criteria: *"Software seamless"*

– Inclusions:

- All Architectural state
- Most micro-architectural state

– Exclusions:

- Temp registers used by ucode
- Some others on a case by case basis

MICROCODE:

– State save and restore

– Core synchronization

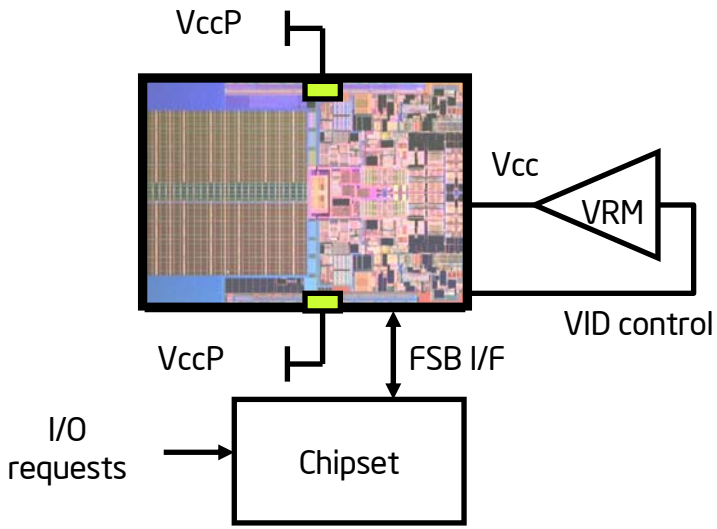
Power Management Unit:

– Manages the DPD power-up sequence

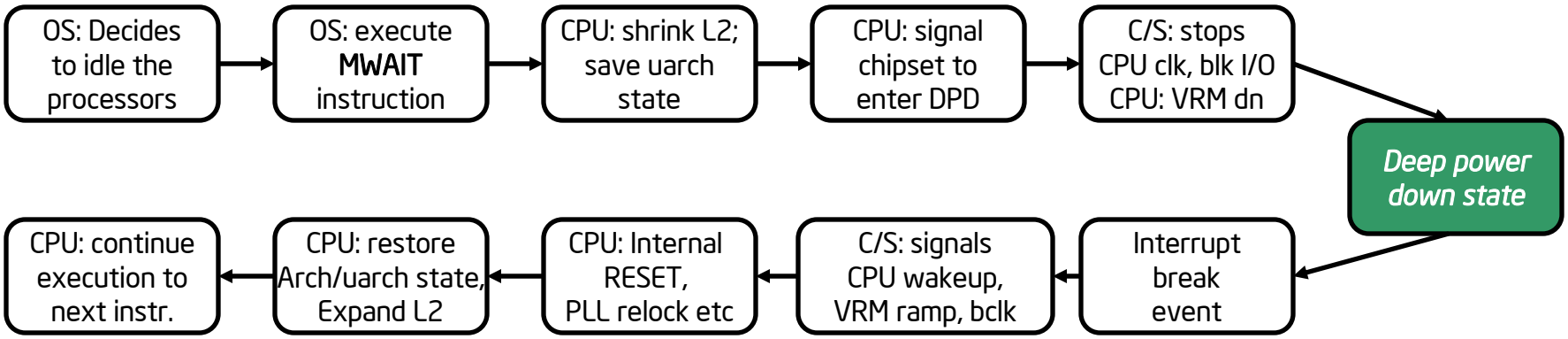
– Manages entry/exit protocol with platform

V. George, et al., Hot Chips 2007

DPD Technology Entry/Exit

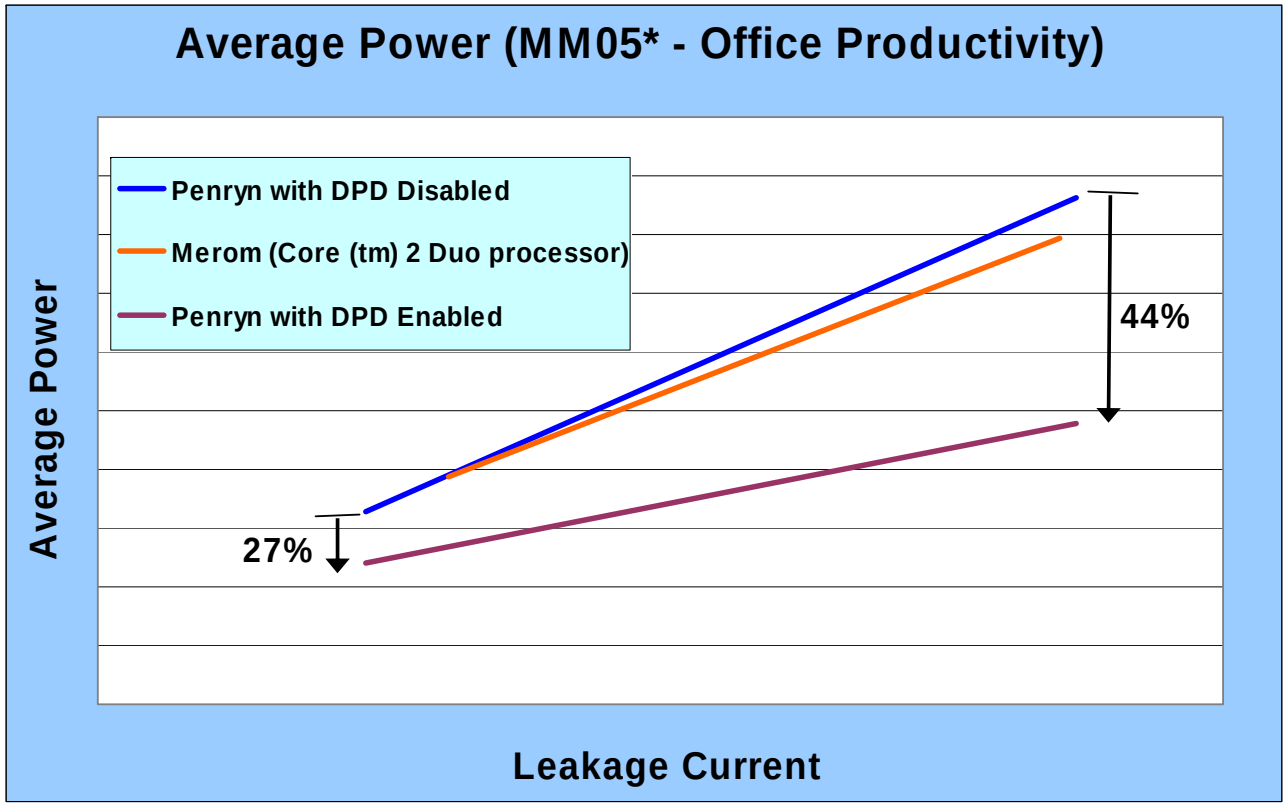


- S/W instruction initiates processor DPD entry
- CPU does rest of sequencing with platform
- Protocol with chipset to block snoops (no CPU wakeup required) while in DPD state
- Exit initiated by a break event (int) in platform
- CPU drives VID to VRM, internal hardware reset, state restore and execution resumption



V. George, et al., Hot Chips 2007

DPD Results (Average Power)



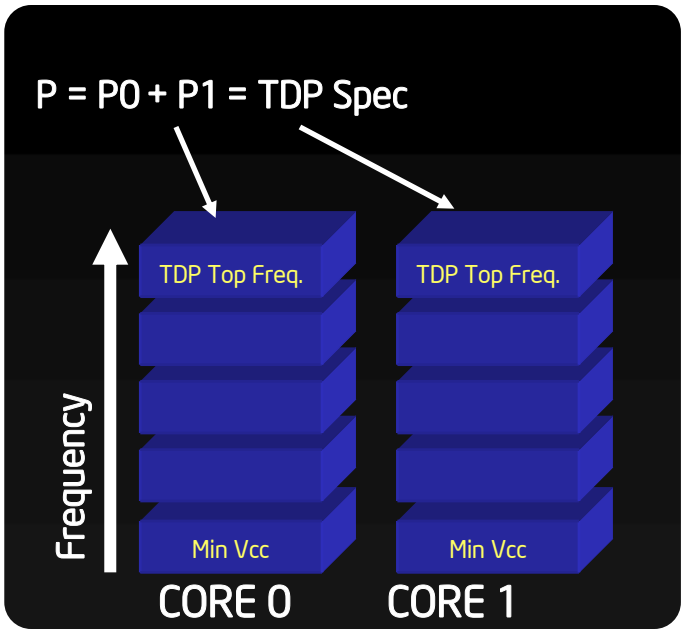
- 27% to 44% (based on the leakage of the part) Average Power reduction as measured by Mobile Mark – Office Productivity benchmark due to DPD feature
- Significant improvement compared to previous generation (Merom)
- Measured Exit latency for DPD state: ~ 150 - 200 us => In expected range

V. George, et al., Hot Chips 2007

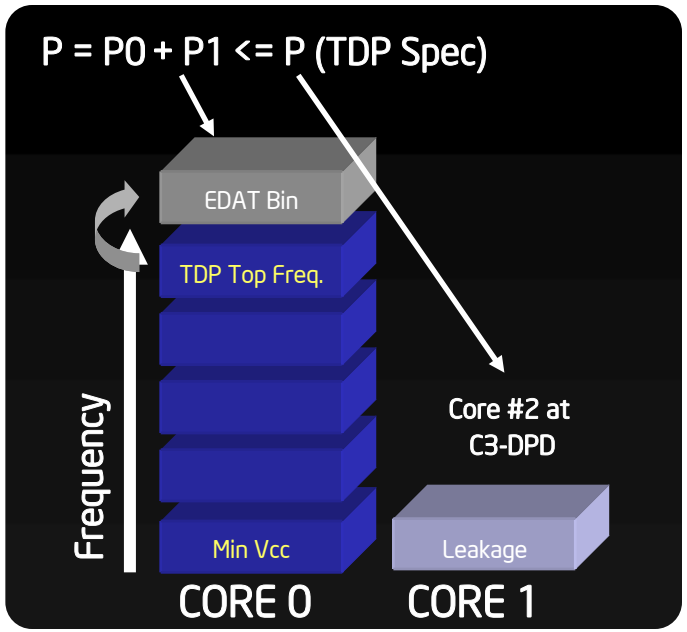
Enhanced Dynamic Acceleration Technology (EDAT)

Concept: In multi-core CPUs, use the power headroom of idle core to boost performance of the active core

Two cores active:
Marked frequency



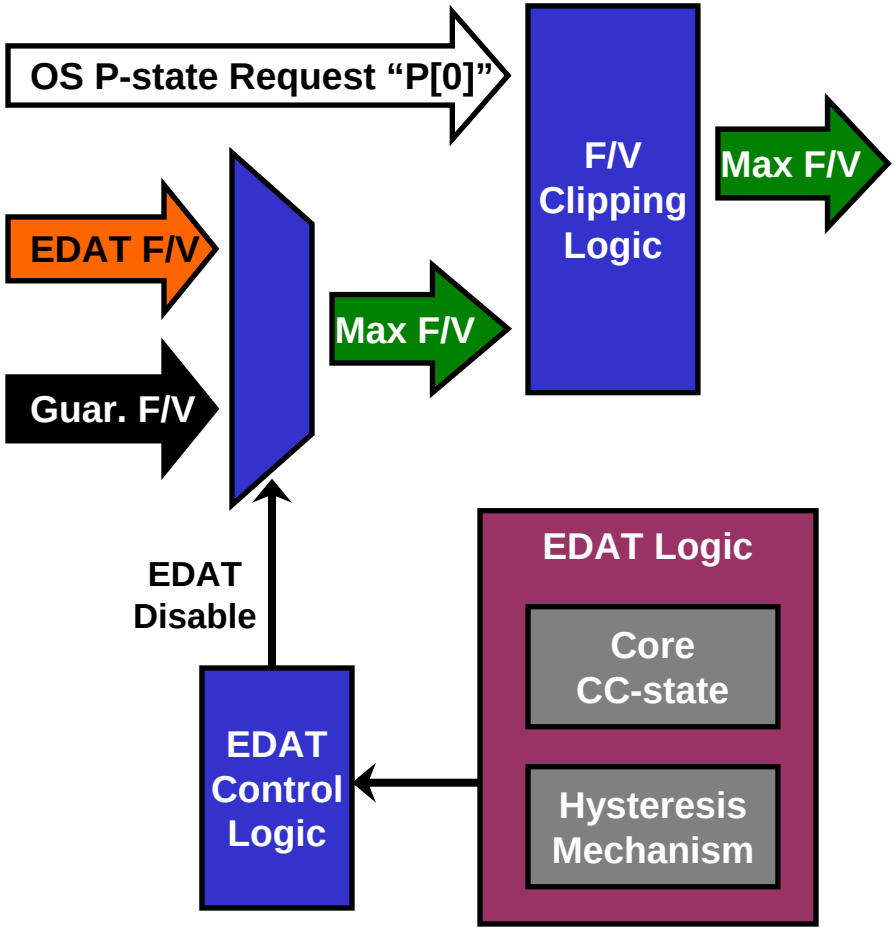
Single core active:
EDAT freq



EDAT provides single-threaded performance boost

V. George, et al., Hot Chips 2007

EDAT Implementation Overview



Microarchitecture

- Entry on OS request AND other core idle
- Idle core defined as "CC3" or deeper C-state
- EDAT Freq pre-programmed in chip based on power, reliability and other constraints
- Exit EDAT mode when Idle core wakes up

Hysteresis mechanism

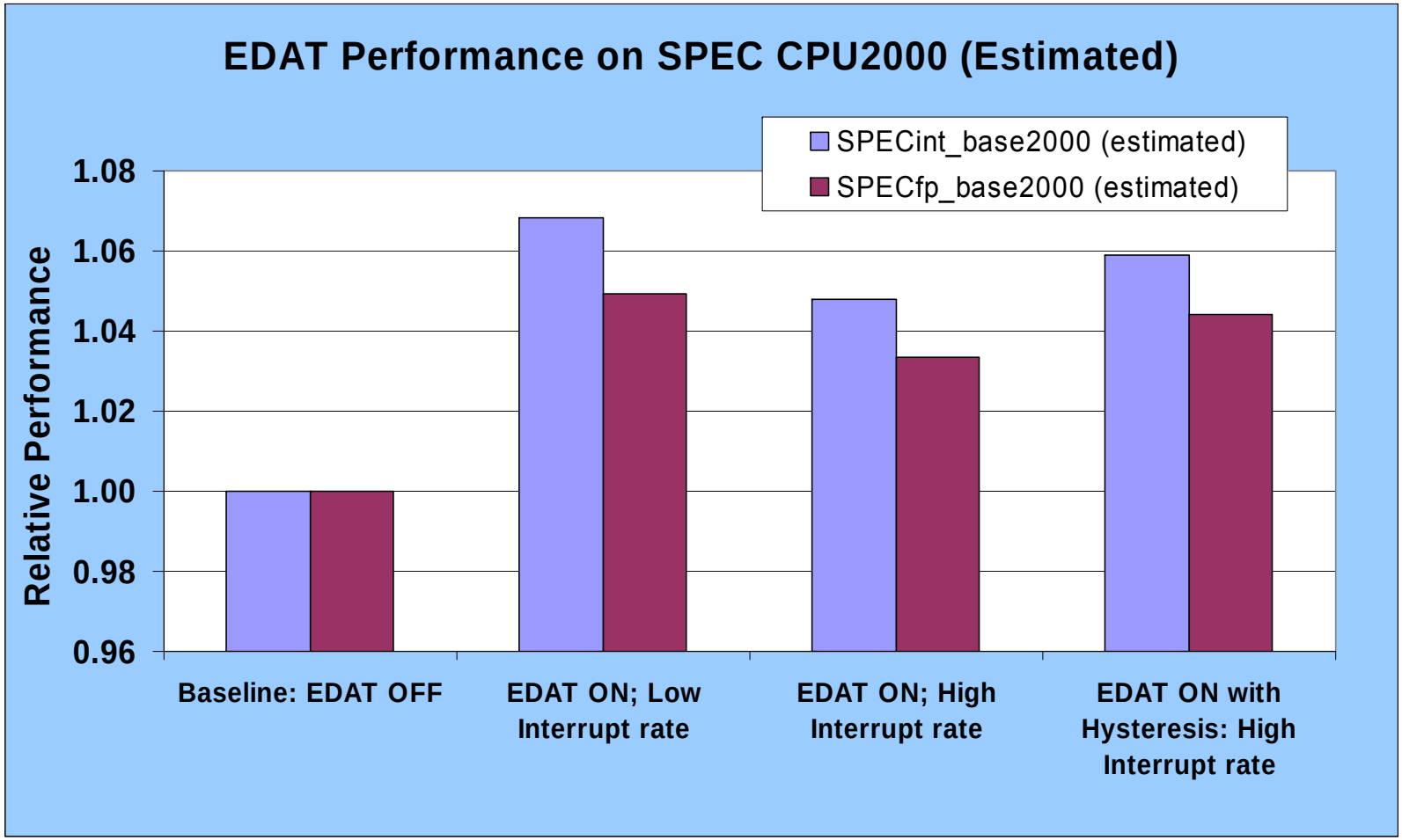
- Allows short durations where 2 cores active
- Reduces perf loss for low activity wake-ups
- Implemented using a few counters
- Voltage Regulator needs to provide for this
- Benefits most at high timer tick rates

OS interface

- OS requests P[0] state if perf demand exists
- EDAT logic grants it if power headroom exists

V. George, et al., Hot Chips 2007

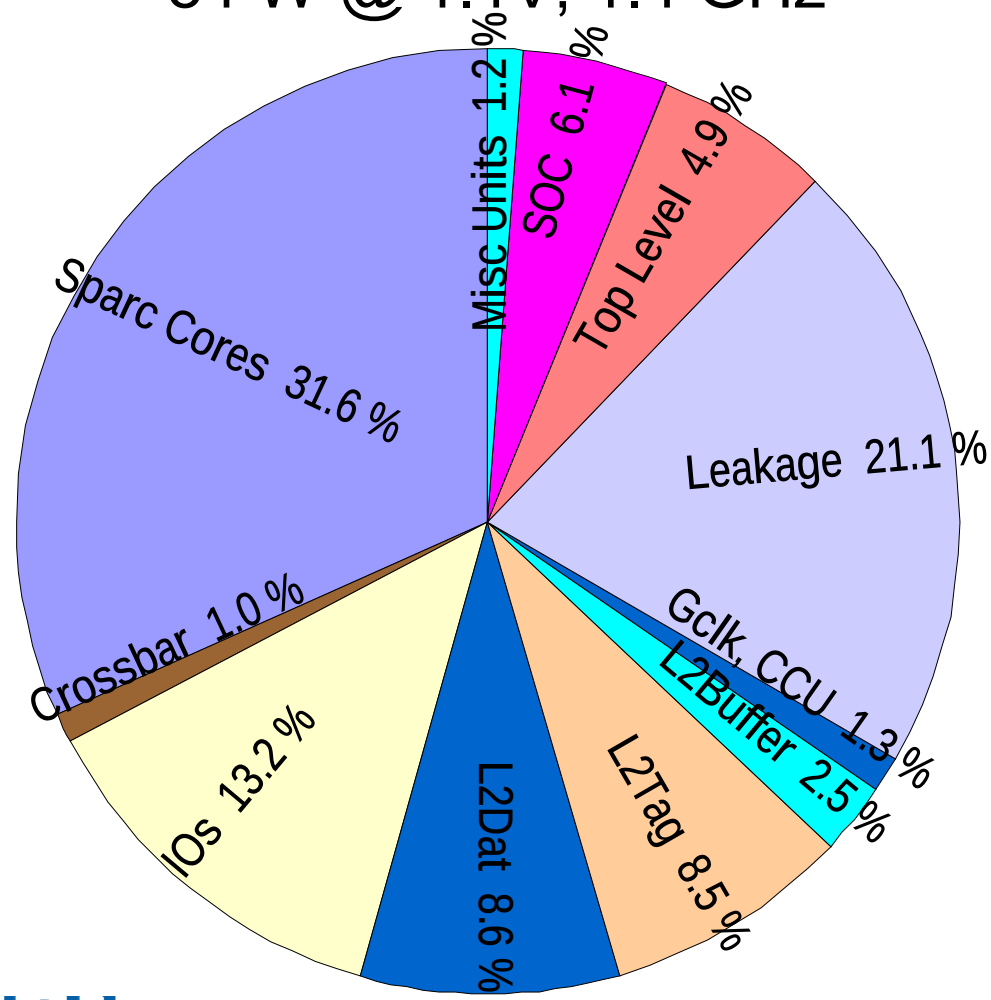
EDAT Performance Results



Performance gains of about 5% on SPECfp_base2000 and 7% on SPECint_base2000 due to EDAT within the same TDP power envelope

Sun's Niagara 2 Power

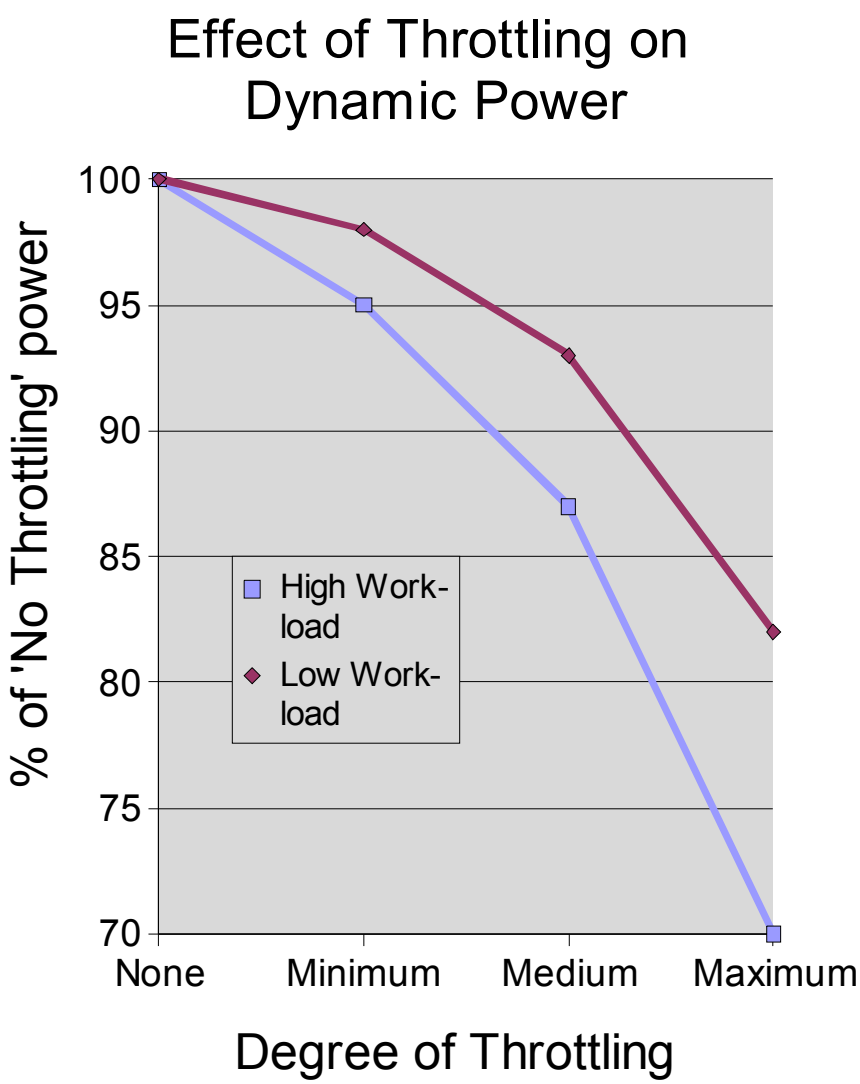
Niagara2 Worst Case Power = 84 W @ 1.1V, 1.4 GHz



- CMT approach used to optimize the design for performance/watt.
- Clock gating used at cluster and local clock-header level.
- 'GATE-BIAS' cells used to reduce leakage.
  - ~10 % increase in channel length gives ~40 % leakage reduction.
- Interconnect W/S combinations optimized for power-delay product to reduce interconnect power.

U. Nawathe (Sun Micro), ISSCC 2007

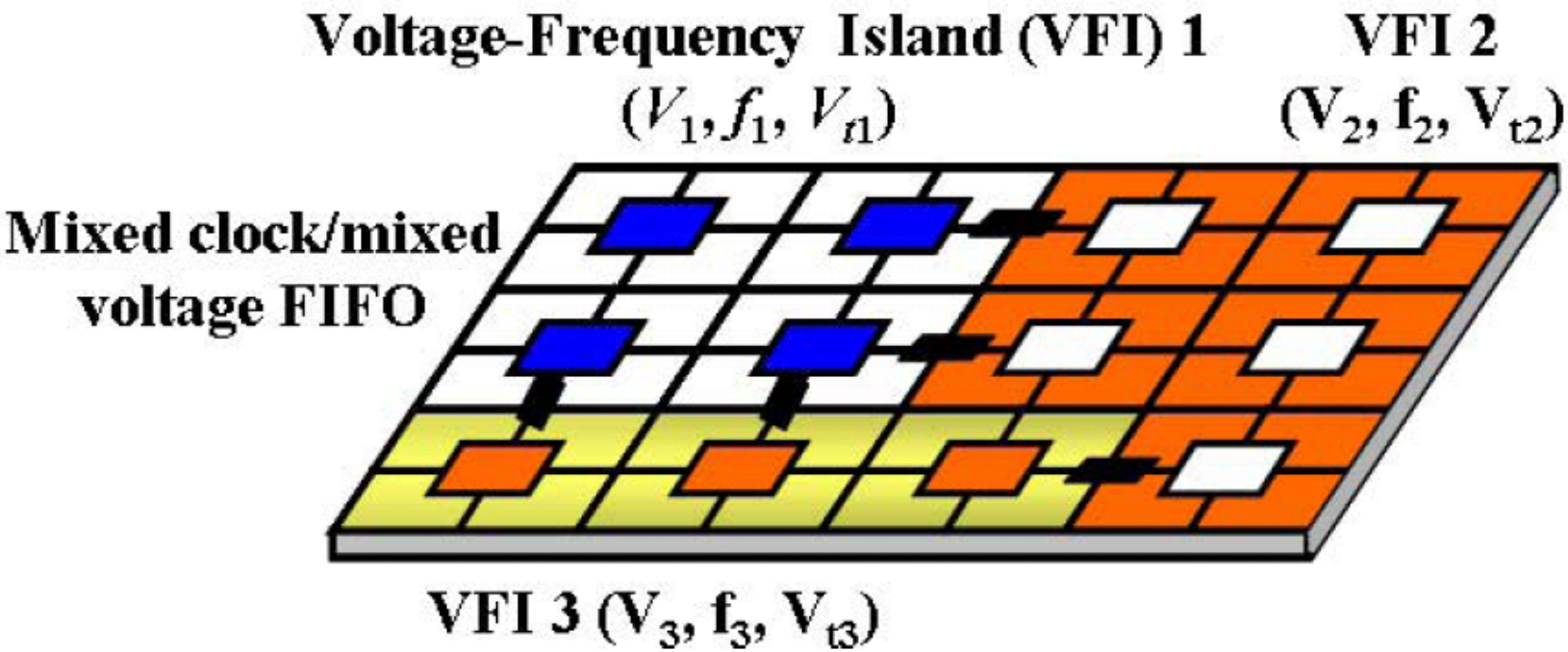
Niagara2 Power Management



- Software can turn threads on/off.
- 'Power Throttling' mode controls instruction issue rates to manage power consumption.
- On-chip thermal diodes monitor die temperature.
  - Helps ensure reliable operation in case of cooling system failure.
- Memory Controllers enable DRAM power-down modes and/or control DRAM access rates to control memory power.

U. Nawathe (Sun Micro), ISSCC 2007

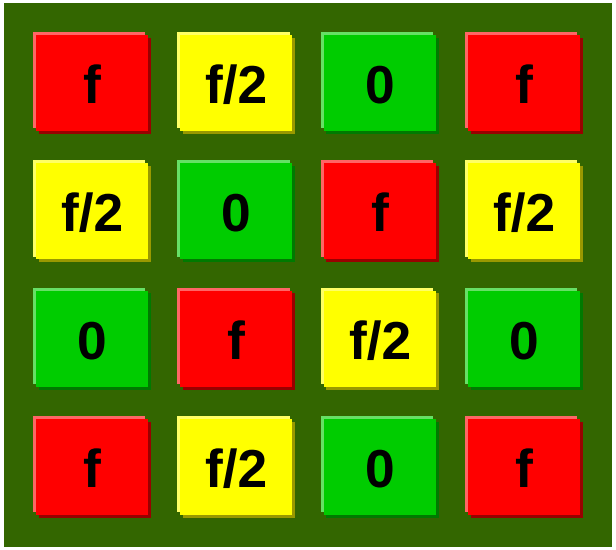
Future Directions

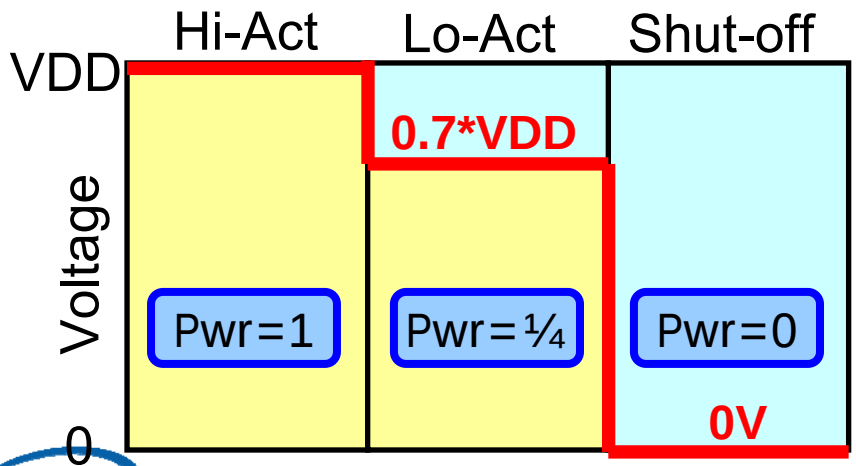


- A sample 2D mesh network with three Voltage / Frequency Islands
- Communication across different islands is achieved through mixed clock / mixed voltage FIFOs

U. Ogras (CMU), DAC 2007

Fine Grain Power Management





Cores with critical tasks
Freq = f, at Vdd
TPT = 1, Power = 1

Non-critical cores
Freq = f/2, at 0.7xVdd
TPT = 0.5, Power = 0.25

Cores shut down
TPT = 0, Power = 0

Adapted from S. Borkar, DAC 2007

Summary

- Low power design is essential for modern computing from hand-held all the way to servers
- Major low-power technology directions:
  - Advanced process technology features: High-K + metal gate, strained silicon
  - Multiple clock and voltage domains
  - Advanced voltage / frequency scaling
  - Operate at the lowest possible voltage
  - Turn off blocks that are not in use (clock and power gating)
- Low-power design techniques are becoming a way of life at all levels of chip and platform design!