Thermal Analysis of FinFETs and its Application to Gate Sizing

Brian Swahn

Soha Hassoun

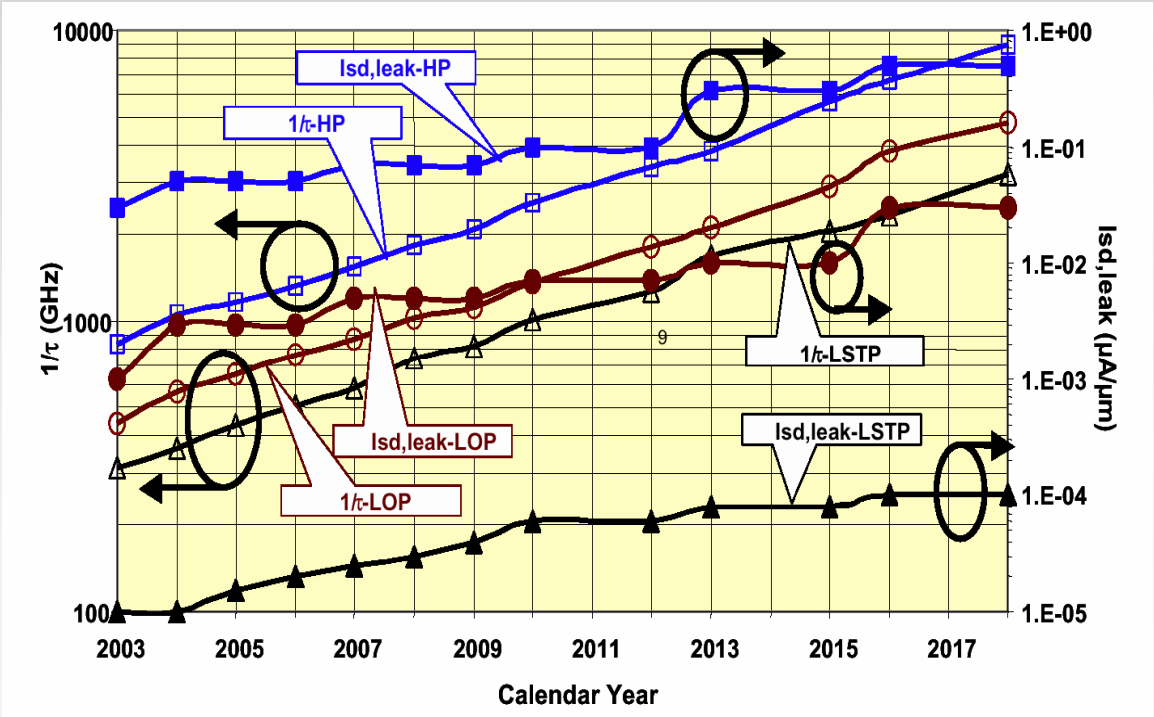
Syed Alam*, David Botha and Arvind Vidyarthi

Tufts University

**Freescale Semiconductor*

Current Technology Trends

- Continuous scaling of traditional CMOS devices leads to problems
  - Increased leakage currents

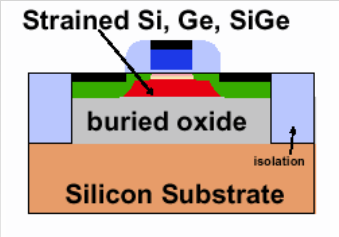


International Technology Roadmap for Semiconductors, 2003

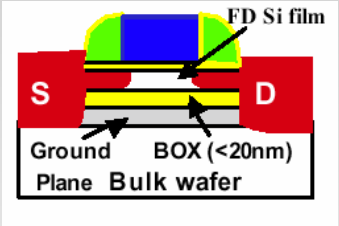
HP – High performance
LOP – Low operating power
LSTP – Low standby power
$\tau = CV/I$ – MOSFET delay
ISD,Leak – Subthreshold source/drain leakage current

ITRS Recommends New Devices

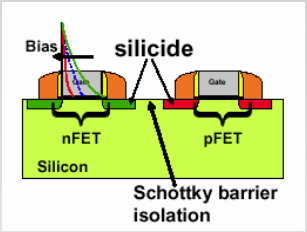
"Non-classical" CMOS device predictions for beyond the 45nm node (ITRS 2003)



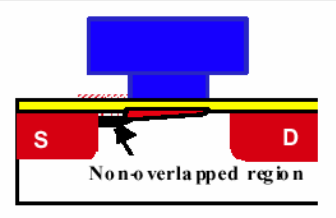
Strained Si



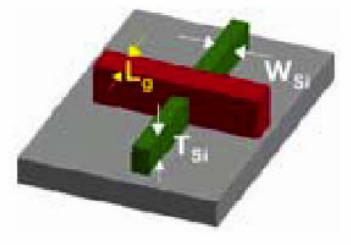
Ultra-thin bodies



Schottky junctions

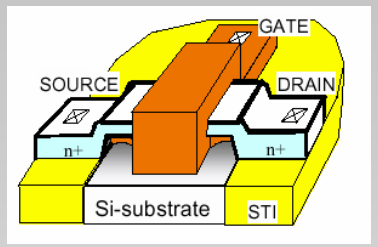


Non-overlapping S/D extensions

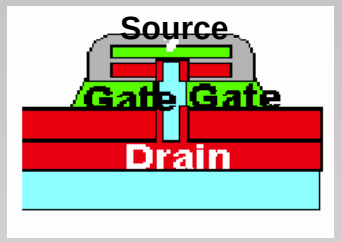


Multiple gates

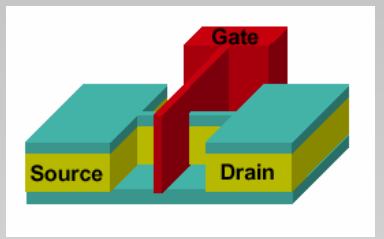
Multiple Gate Devices



Planar double gate



Vertical double gate

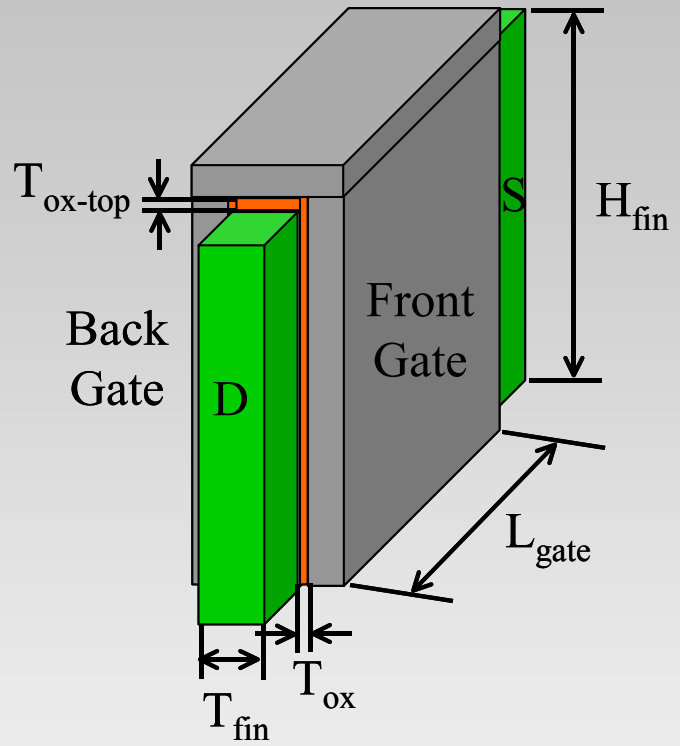


FinFET

- High drive currents
- Improved subthreshold slope
  - Two gates control the channel
- Self-aligned gates
- Fabricated by introducing a few new mask steps in a traditional MOSFET process flow*

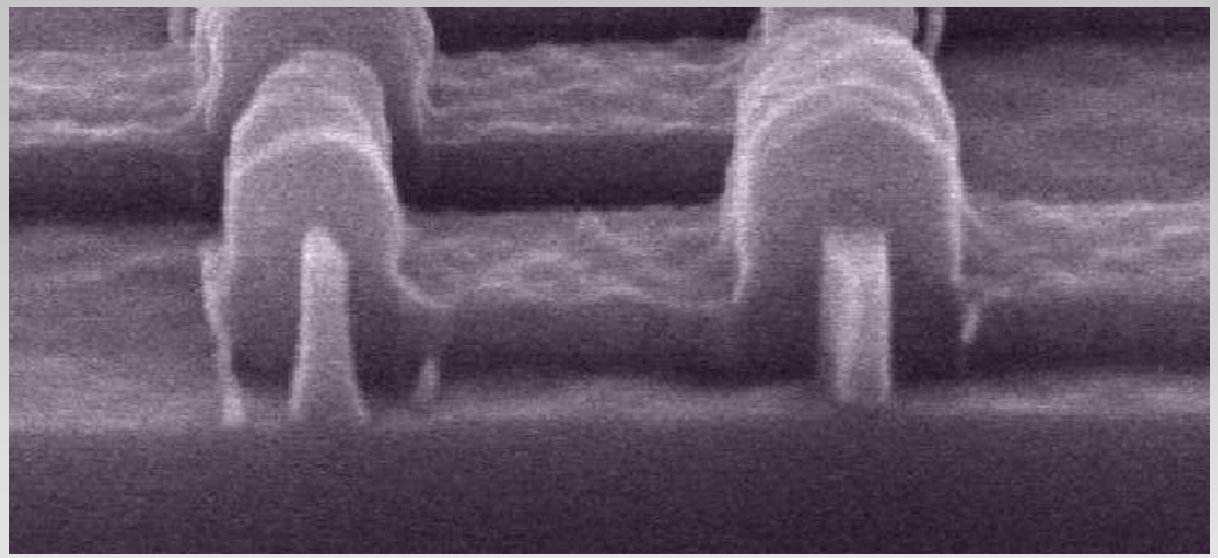
* T. Ludwing, I. Aller, V. Gernhoefer, J. Keinert, E. Nowak, R. Joshi, A. Mueller, and S. Tomaschko. “FinFET Technology for Future Microprocessors”. *IEEE International SOI Conference*, pages 33—4, 2003

FinFET Device Structure



- The FinFET consists of a channel, source, drain, and gate
- Quasi-planar device – current flows parallel to the plane of wafer
- Fin height is a process-fixed parameter
- Device width: $W = 2 \times H_{fin}$

Building Wider FinFETs



Multiple fin device*

- Wider FinFETs are achieved by placing multiple fins in parallel
- Total device width: $W_T = 2 \times H_{fin} \times n$

* K. Bernstein and C. Chaung and R. Joshi and R. Puri, “Design and CAD Challenges is Sub-90nm CMOS Technologies”, *International Conference on Computer-Aided Design*, 129—36, 2003.

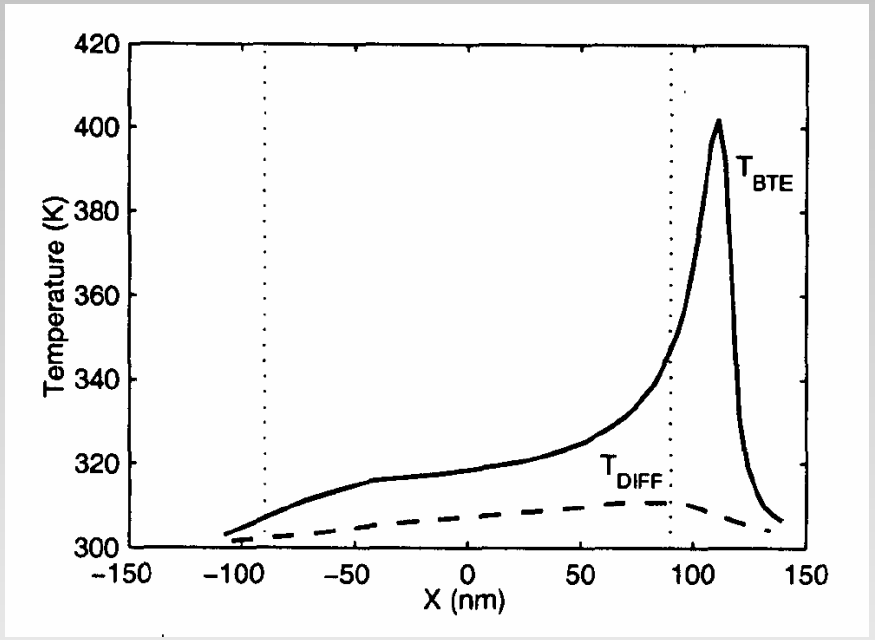
Outline

- Device thermal analysis
  - Single fin (Pop model)
  - Multiple fin
- FinFET gate sizing
- Preliminary experimental results
- Future work

Device Thermal Analysis

- Confined geometries require device thermal analysis
- Wider devices (multiple tightly packed parallel fins) hinder efficient heat removal
- Classical effects
  - Heat generated in devices is due to recombination of electron-hole pairs in the drain region
  - Generated heat causes temperature gradients within device
  - Heat generated can be approximated as: $Q=I_{on}V_{gs}$
- Sub-continuum effects
  - Small device dimensions and material types reduce thermal conductivity
  - Heat diffusion equations fail to capture dominate heat transport due to phonons – Boltzmann Transport Equation (BTE) is needed

BTE Experimental Results*

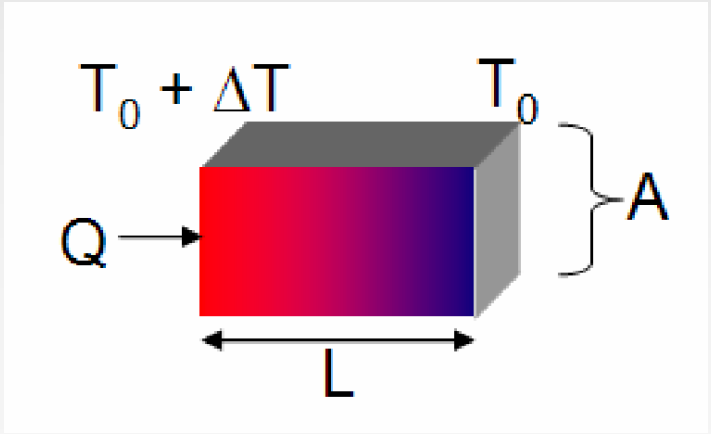


- BTE can be used to estimate phonon distributions within a device
- Captures phonon-phonon interactions
- Estimates localized hot spots

* E. Pop, K. Banerjee, P. Servdrup, and K. Goodson. “Localized Heating Effects and Scaling of Sub-0.18 Micron CMOS Devices”. *IEEE International Electron Devices Meeting 2001*, pages 677—80, 2001.

Thermal Analysis

- Complicated geometries can be modeled by a thermal resistance circuit
- Electrical analysis tools (SPICE) can be used to solve the thermal network
- Ohm's law $\Leftrightarrow$ Fourier's law
  - $\Delta V = I \cdot R \Leftrightarrow \Delta T = Q \cdot R$

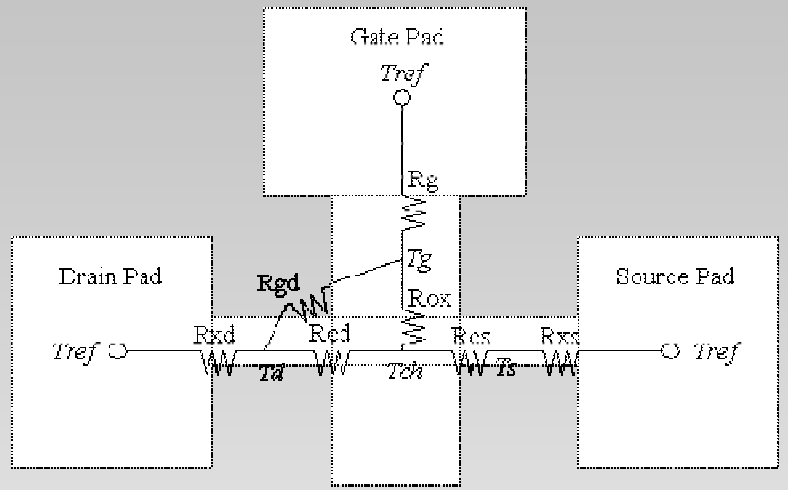


Fourier's law of
1-D heat conduction:

$$\Delta T = \frac{L}{k \cdot A} \cdot Q$$

R_{th}

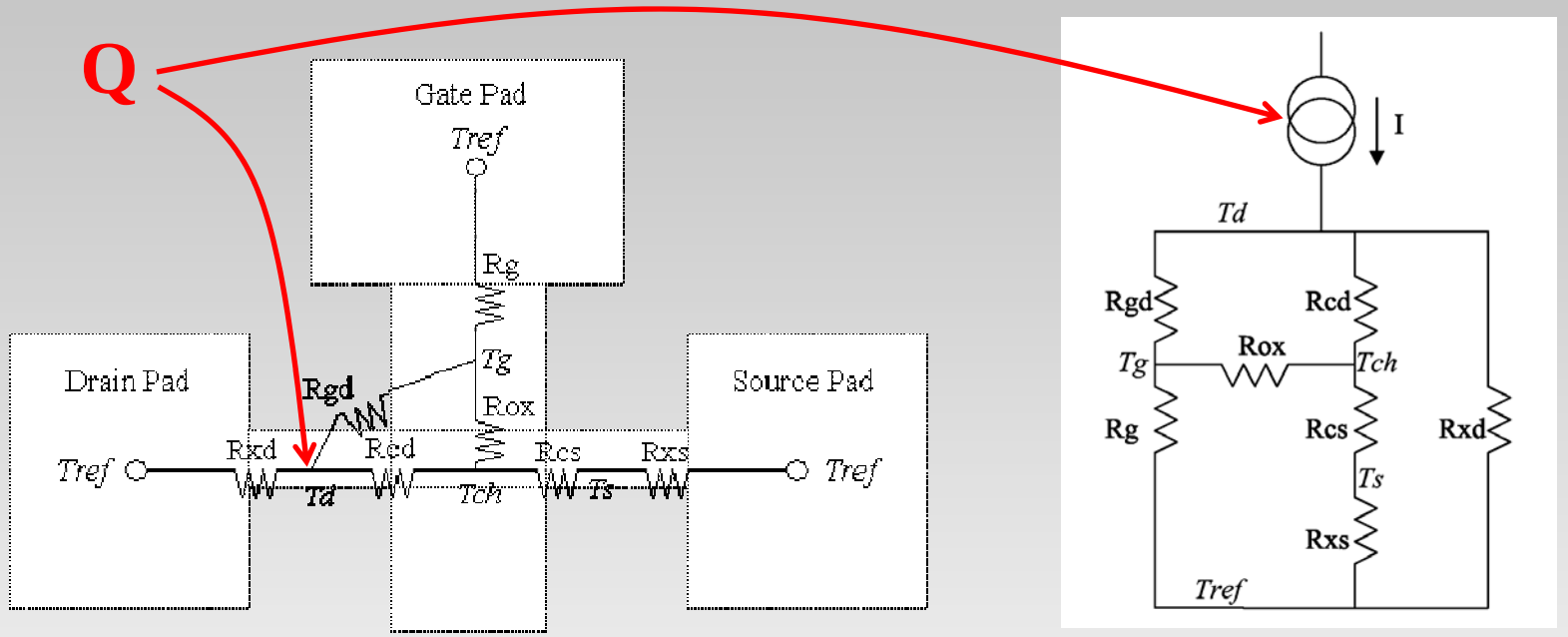
Pop's Single-Fin Thermal Model*



- Thermal analysis of single fin device
  - Calculate thermal resistances based on device geometries and reduced thermal conductivities due to sub-continuum effects**
  - Assume drain, source and gate pads are at a known reference temperature
  - Solve the thermal network for temperatures at four points: drain, source, channel, and gate

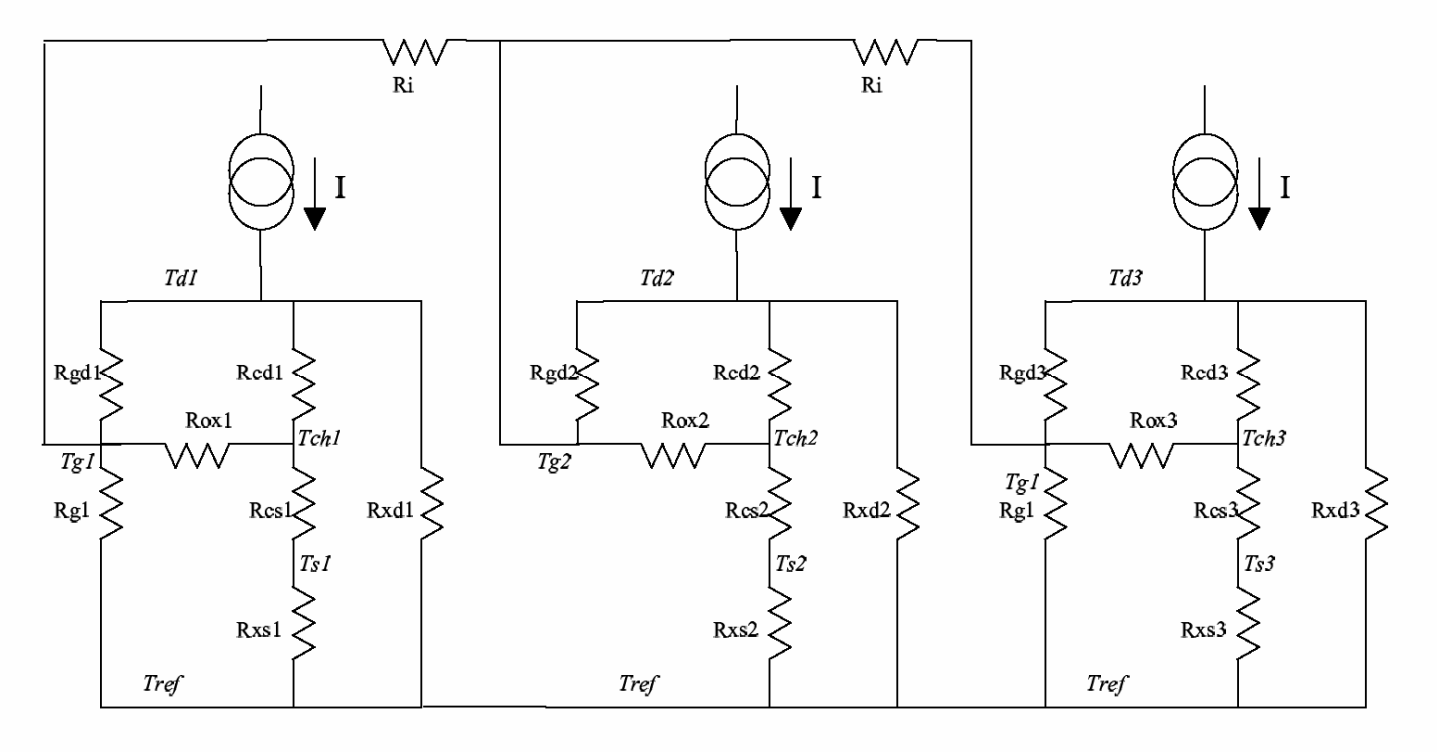
* E. Pop, R. Dutton, and K. Goodson. "Thermal Analysis of Ultra-Thin Body Scaling". *IEEE International Electron Devices Meeting 2003*, pages 36.6.1—4, 2003.
** P. Sverdrup, Y.S. Ju, and K. Goodson. "Sub-continuum Simulations of Heat Conduction in Silicon-On-Insulator Transistors". *Journal of Heat Transfer*, pages 130—7, February 2001.

Pop's Single-Fin Thermal Model



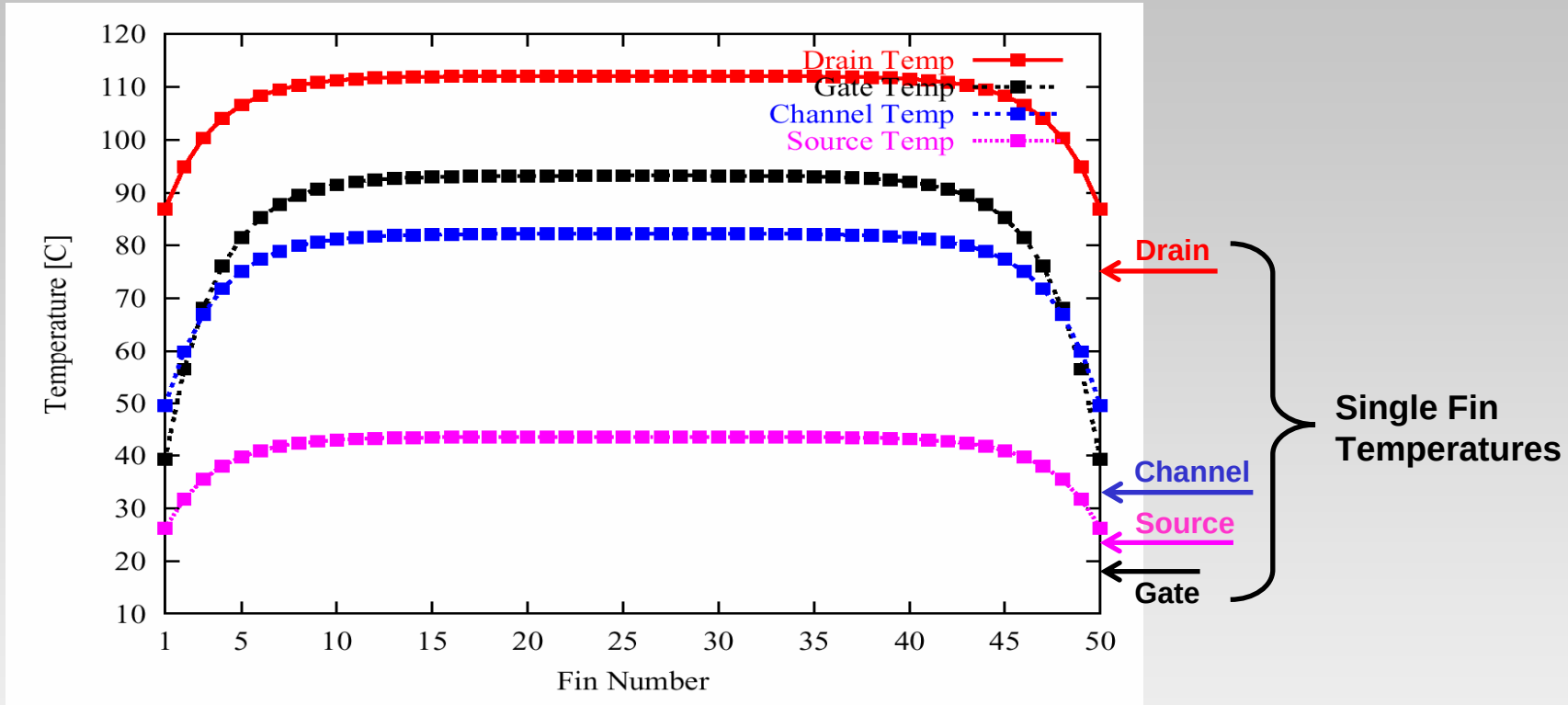
- Heat injected: $Q=I_{on}V_{gs}$
- Use SPICE to solve for nodal voltages (temperatures)

Multi-Fin Thermal Model



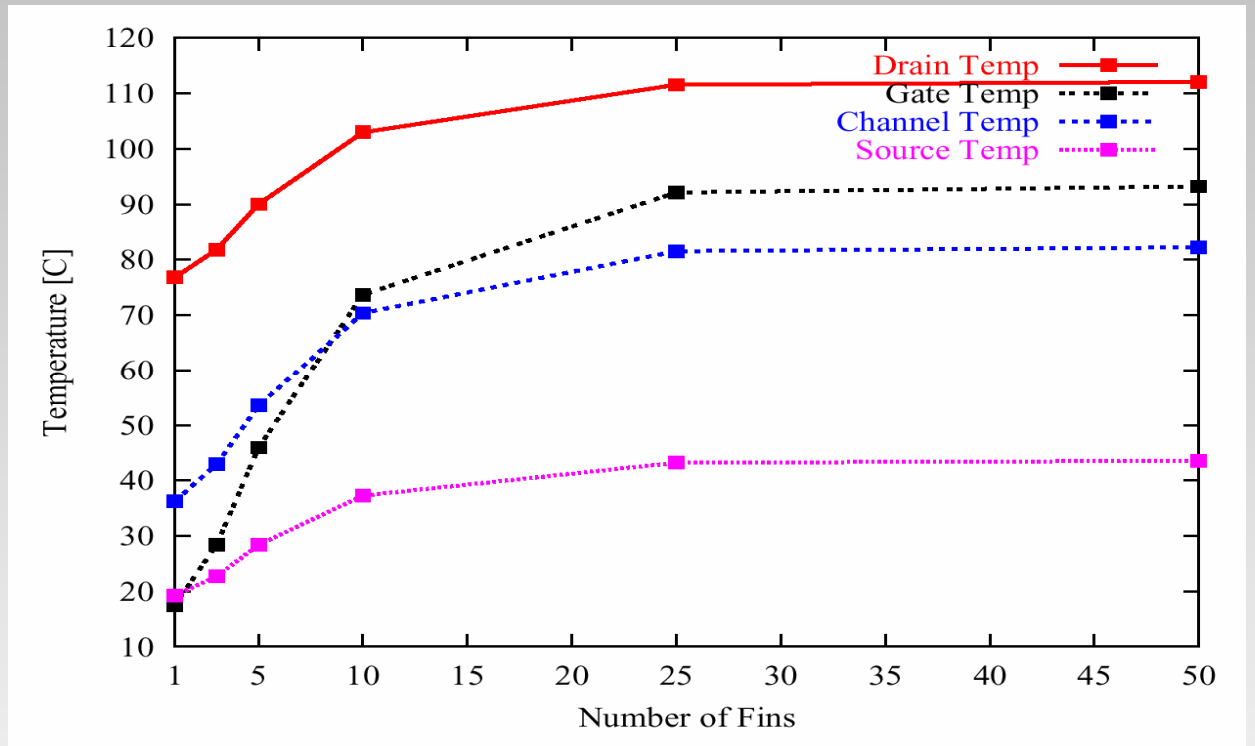
- Extended single-fin model to multiple fins
  - Assume fins are spaced a fixed distance apart (W_{space})
  - Assume two gate pads will be present

Thermal Analysis For a 50-Fin Device



- Inner fins are hotter than outer fins due indirect access to gate pads
- End fins are hotter than single fin device temperatures due to heat spreading
- Results correlate well with Pop's experiments

Multi-Fin Thermal Analysis Results



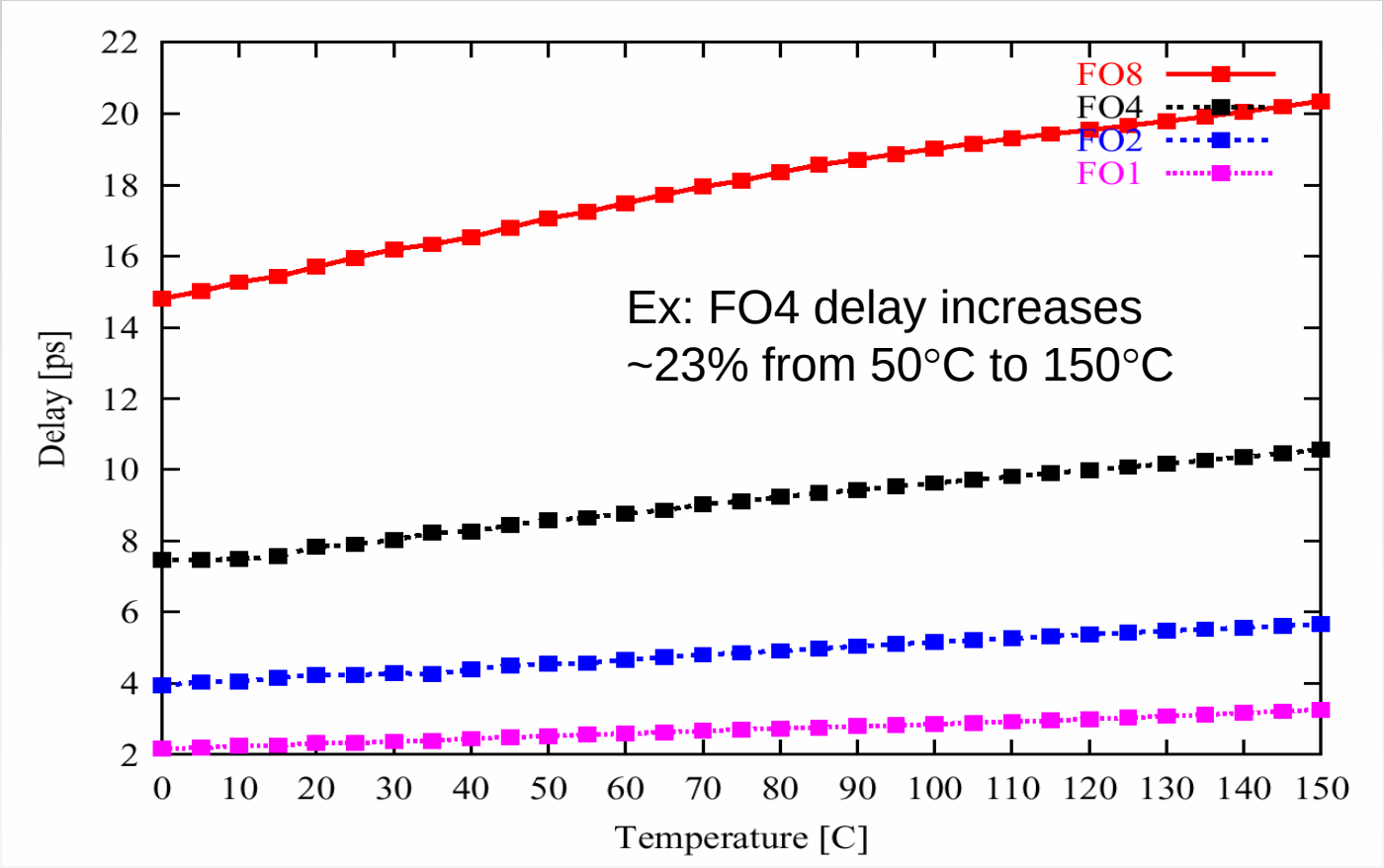
- Temperature has a dependence on the number of fins, for devices with less than 20 fins
- Gate temperature exceeds channel temperature for devices with greater than 10 fins

FinFET Gate Sizing

- Must meet given constraints
- Traditional gate sizing has been studied extensively over the past several decades, resulting in numerous approaches
  - Exact vs. heuristic
  - Convex vs. linear
  - Simplified delay models vs. lookup tables
- Unlike traditional gate sizing, FinFET gate sizing is a discrete optimization problem
  - Width of device based on number of fins

Delay is a Function of Temperature

- Mobility and threshold voltage are temperature dependent
  - As T↑, μ↓ and V_T↓



Modeling Using Logical Effort*

$$d_{gate} = \tau\left(g \cdot h + p\right)$$

- Model describes gate delays from capacitive loading
- Delay of a gate consists of four quantities
  - Logical effort (g)
  - Electrical effort (h)
  - Parasitic delay ($p = \alpha \cdot p_{inv}$)
  - Process delay unit (τ)
- Process (τ) and parasitic inverter (p_{inv}) delays are applied to all gates
  - Obtained by curve fitting process simulation data

* I. Sutherland, B. Sproull, and D. Harris. “Logical Effort: Designing Fast CMOS Circuits”. Morgan Kaufmann Publishers Inc., 1999.

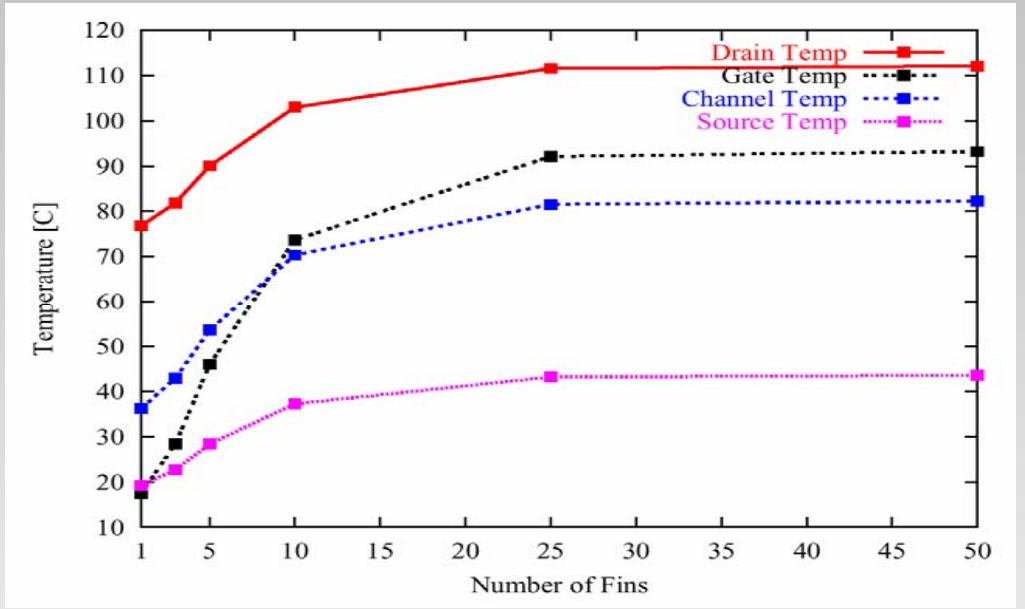
Delay Modeling Based on Logical Effort

$$d_i = \tau(g \cdot h + p), \text{where}$$

$$h = \frac{C_{out}}{C_{in}} = \frac{\sum_{j \in FO(i)} n_j}{n_i}$$

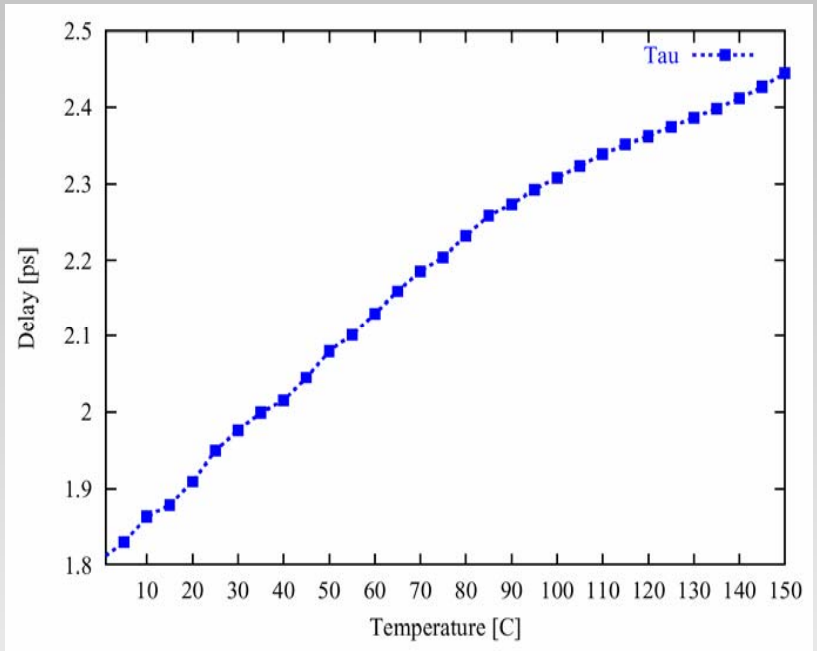
$$n_i \in \{1,2,3,\ldots,n_{max}\}$$

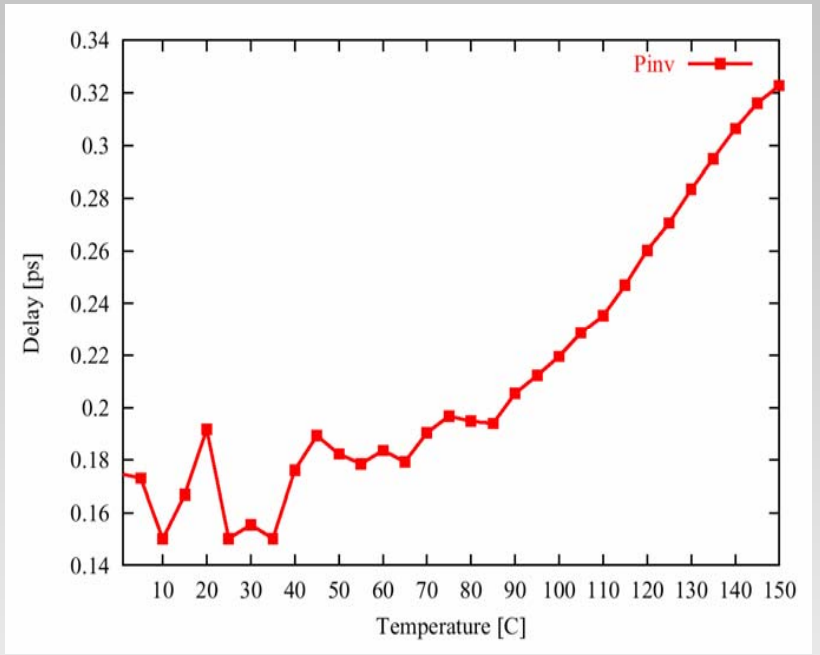
$$d_{ko} = d_k + \max_{j \in FI(k)}(d_{jo})$$



- Integer restrictions are placed on gate scale factors (n_i)
- From thermal analysis, maximum device temperature rise can be controlled (setting n_{max})

Model Calibration





- We calibrate the temperature dependent parameters, τ and P_{inv}, at different temperatures*
- The parameters are then used during gate sizing, based on a maximum expected source temperature

* I. Sutherland, B. Sproull, and D. Harris. “Logical Effort: Designing Fast CMOS Circuits”. Morgan Kaufmann Publishers Inc., 1999.

Optimization

$$\text{Minimize} \quad \sum n_i$$

$$\text{Subject to} \quad d_{PO} \le T_{clk}, \forall PO$$

$$d_{ko} = d_k + \max_{j \in FI(k)}(d_{jo})$$

$$1 \le n_i \le n_{max}$$

$$n_i \in Z^*$$

- Problem formulated as a Mixed Integer Non-Linear Program (MINLP)
  - We solve via the SBB package in the General Algebraic Modeling System (GAMS)
- Optimization minimizes area based on delay and temperature constraints
  - Leakage also minimized

Experimental Setup

- Utilized earlier thermal analysis to derive maximum number of allowable fins
  - Established thermal constraints, based on temperature rise above reference temperature and number of fins
- Used combinational circuits from the LGSynth93 benchmarks
- Mapped each circuit in SIS (minimizing area), to produce a gate-level netlist

Preliminary Experimental Results

Name	# Gates	$T_{source} = 60°C$ Area $T_{Drain} = 125°C$ $n_{max} = 8$	$T_{source} = 60°C$ Run Time(s)	$T_{source} = 25°C$ Area $n_{max} = \infty$	$T_{source} = 25°C$ Run Time (s)
C17	8	109	0.3	99	1
C432	198	1560	6	1541	4
apex7	287	2081	31.5	2080	35
alu2	376	3278	9	3212	11
t481	717	6167	13	6075	5
alu4	753	5898	45	5815	27

- Further experimentation is required
  - Modeling with interconnect
  - Sizing with source temperatures greater than 60°C
  - Heuristic optimization methods

Improvements

- FinFET thermal model
  - Consider heat loss to surrounding dielectric, oxide, and interconnect
  - Modify assumption that drain, source, and gate pads are at reference temperature
- Thermal-electrical co-simulations
  - Currently, heat generation is assumed uniform for each fin
  - Coupling the thermal and electrical networks will improve the temperature estimations in the thermal network

Summary

- Investigated using thermal device modeling to drive circuit optimization
- Extended the existing single fin FinFET thermal model
  - Performed detailed analysis of self-heating in multiple fin devices
- More work on FinFET gate sizing