i-acoma group

VARIUS: A Model of Process Variation and Resulting Timing Errors for Microarchitects

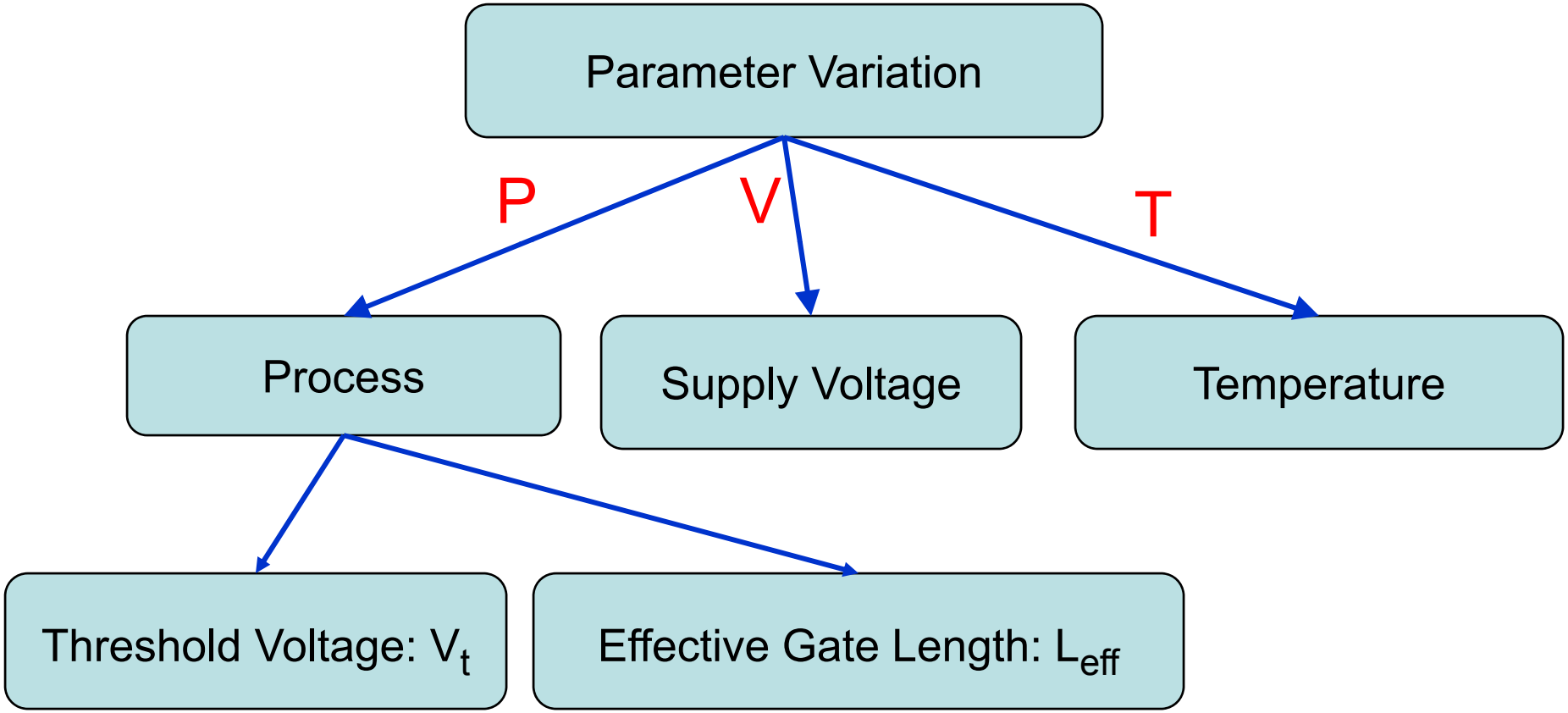
Smruti R. Sarangi, Brian Greskamp, Radu Teodorescu,
Jun Nakano, Abhishek Tiwari and Josep Torrellas

University of Illinois at Urbana-Champaign

http://iacoma.cs.uiuc.edu

ILLINOIS

Parameter Variation



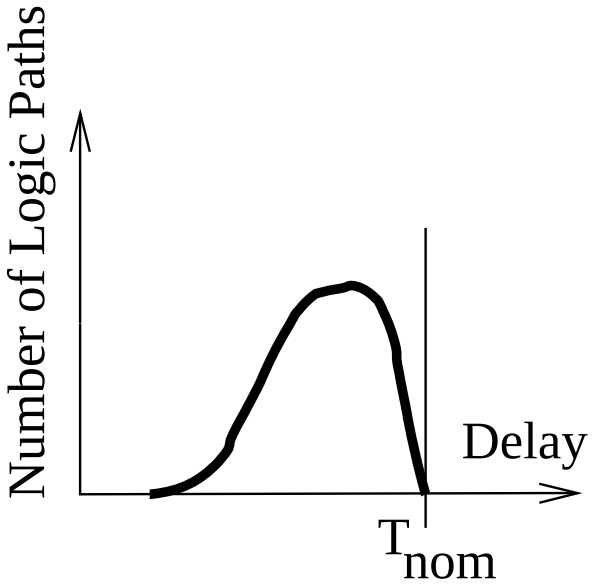
Process Variation is a Problem

Variation of V_t and L_{eff}:

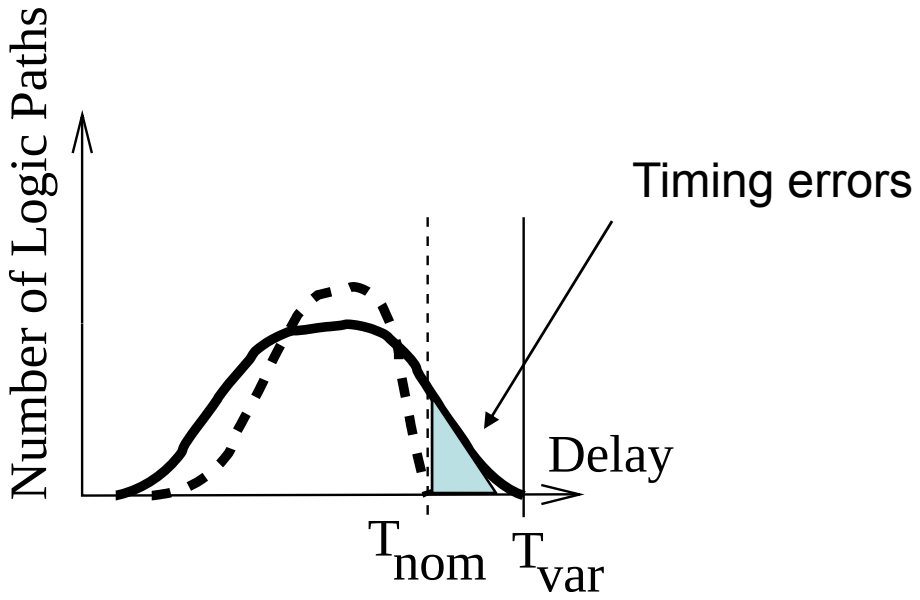
- Chip leakage power ↑
- Chip frequency ↓

Chip Frequency Decreases

$$T_g \propto \frac{V_{dd} L_{eff}}{\mu(T)(V_{dd} - V_t)^{\alpha}}$$



Distribution of path delays in pipe stage: No variation



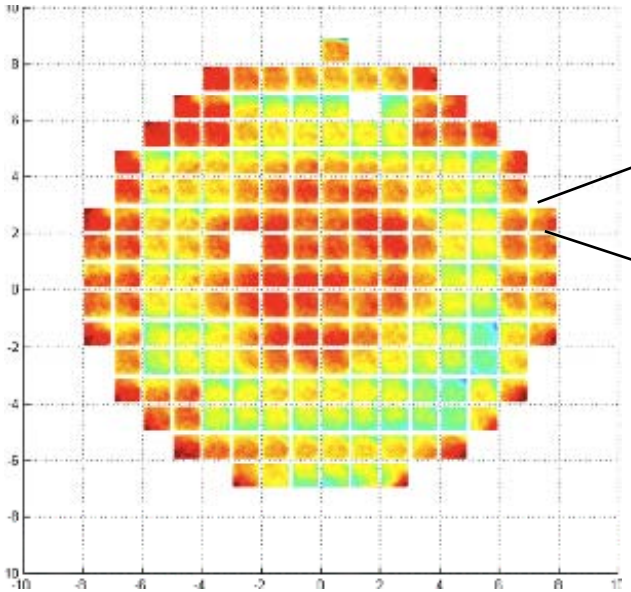
Distribution of path delays in pipe stage: With variation

Implications on Design Decisions

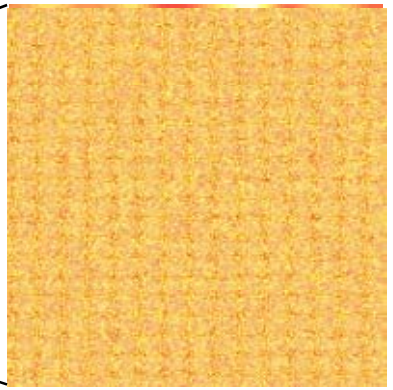
- Unlikely designs will be for worst-case par. values
  - Chips too slow or too costly to design
  - Performance of a generation lost
- Alternative: design closer to avg. par values
  - Some parts of the chip will be too slow: can we live with timing errors?
  - Some parts of the chip will dissipate too much power: can we push it to other parts of the chip?
  - Multi-tiered solution required: circuits, CAD, micro-architecture, software → this talk focuses on μarch.

Variation components

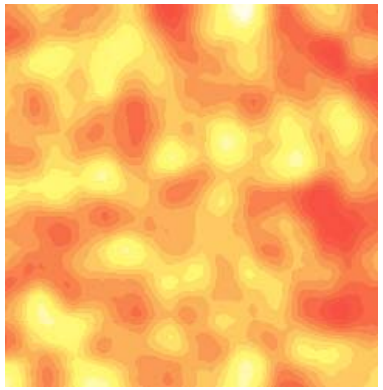
die-to-die



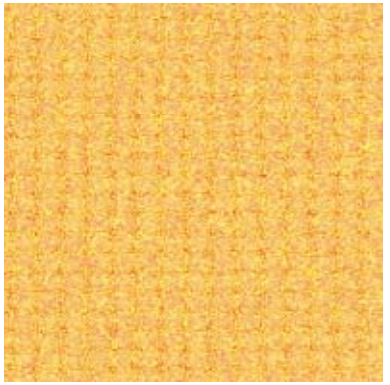
within die



systematic



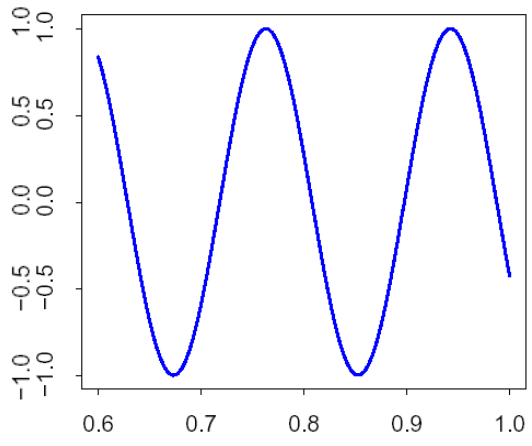
random



spatial correlation

Modeling Process Variation

Process Variation (Not to Scale)



Systematic Variation

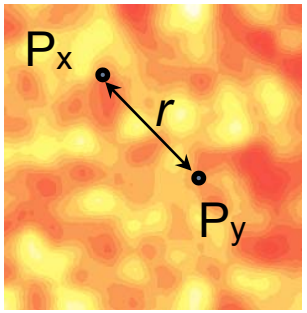
- Lens aberrations
- Mask deformities
- Thickness variation in CMP
- Photo-lithographic effects

Random Variation

- Variable dopant density
- Line edge roughness

Systematic Variation

- We divide the chip into a grid of points
  - Each point has one random value of ΔP_{sys}
- Multivariate normal distribution (μ_{sys}, σ_{sys})
- Characterized by a correlation function:

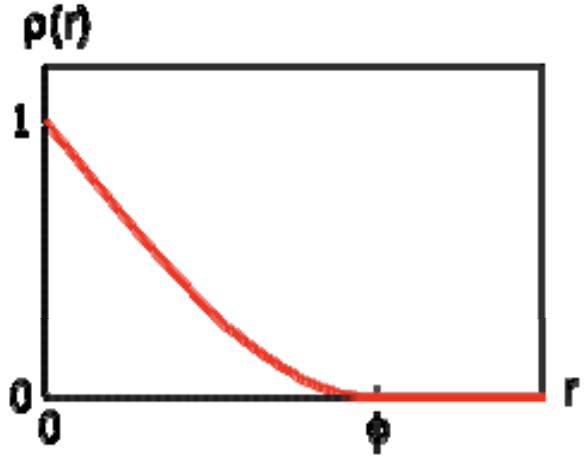


$$corr(P_{\vec{x}}, P_{\vec{y}}) = \rho(r)\ ;\ r = |\vec{x} - \vec{y}|$$

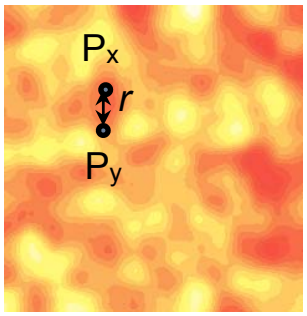
  - Correlation is position independent and isotropic
  - For $\rho(r)$ we choose the spherical model
- Random: modeled analytically at transistor granularity

Spherical Model

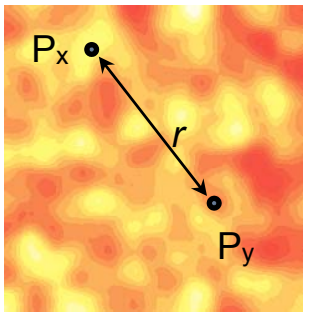
$$\rho(r) = \begin{cases} 1 - \frac{3r}{2\phi} + \frac{r^3}{2\phi^3} & : (r \le \phi) \\ 0 & : \text{otherwise} \end{cases}$$



Stronger correlation



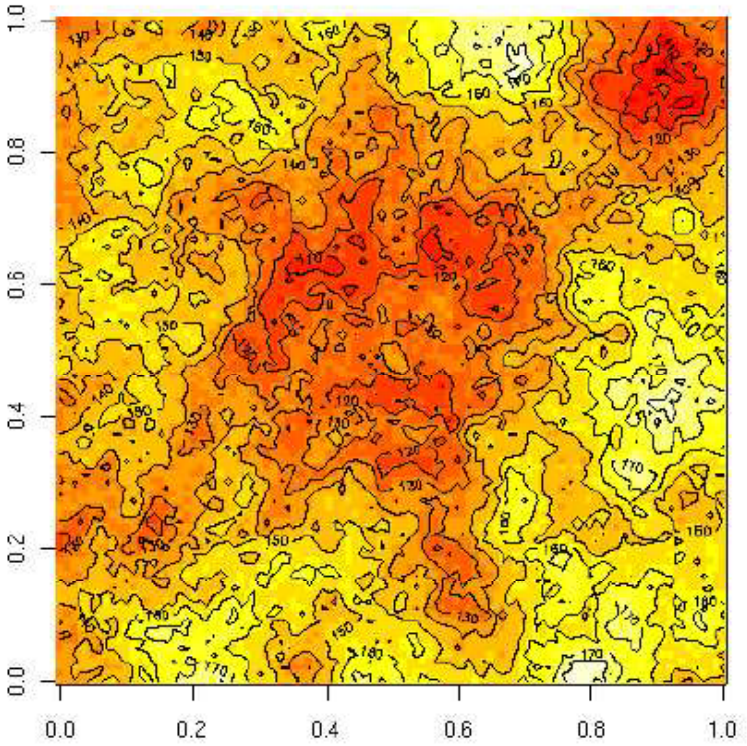
Weaker correlation



- Matches measured data [Friedberg et al. 05]

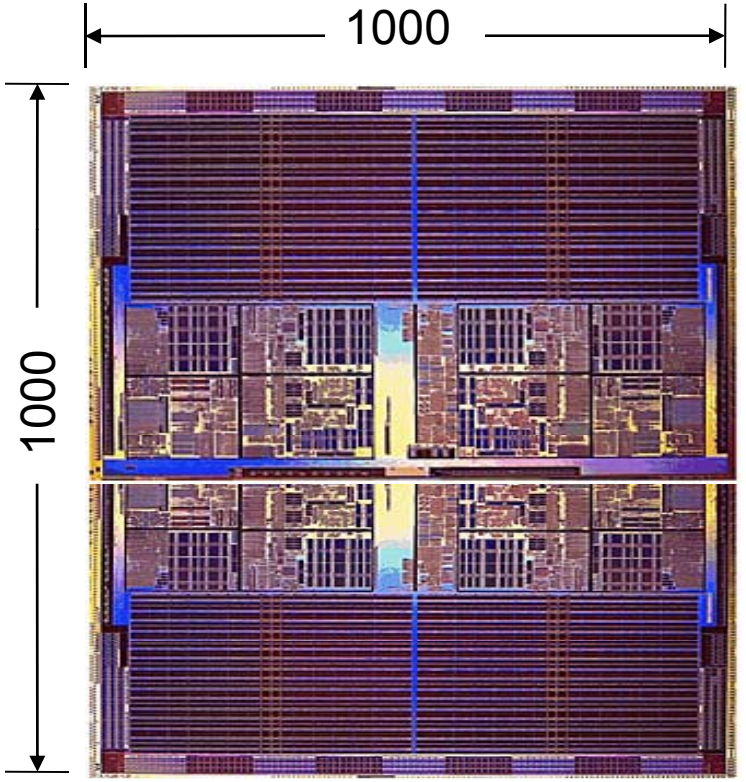
Modeling Systematic Variation

Multivariate normal distribution with
Spherical Spatial Correlation (μ, σ, Φ)

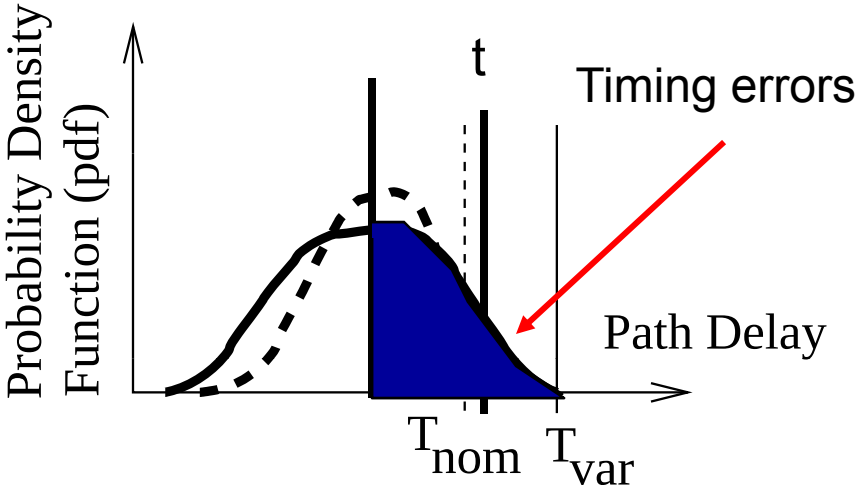


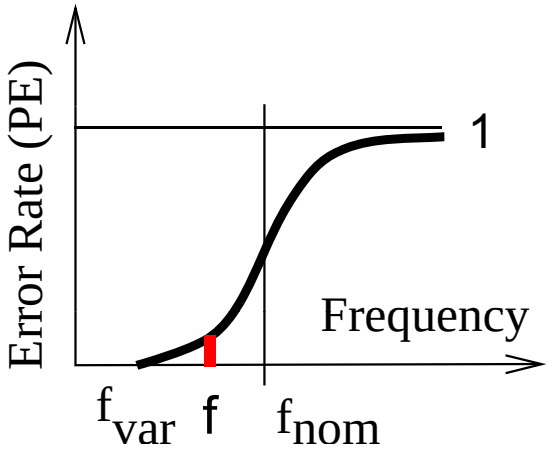
Example variation map

Break into a million cells

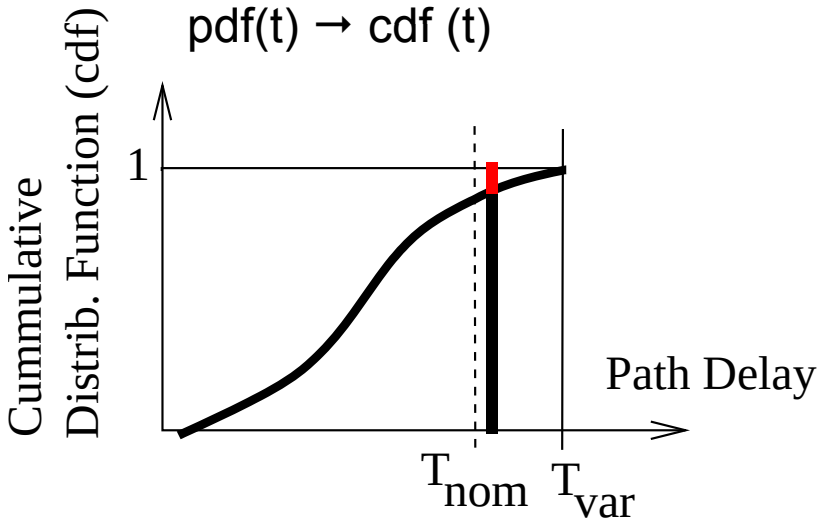


Paths in a Pipeline Stage

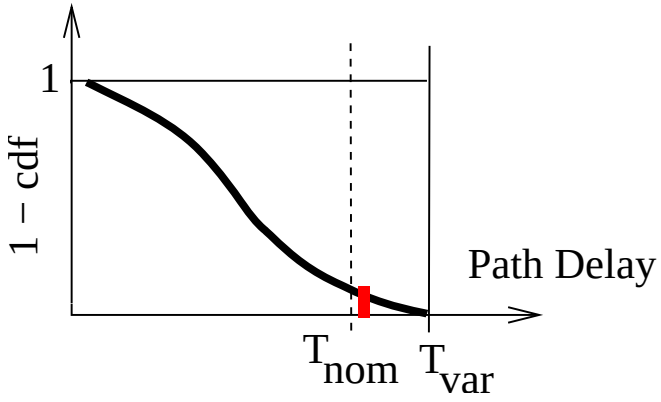




pdf(t) → cdf (t)

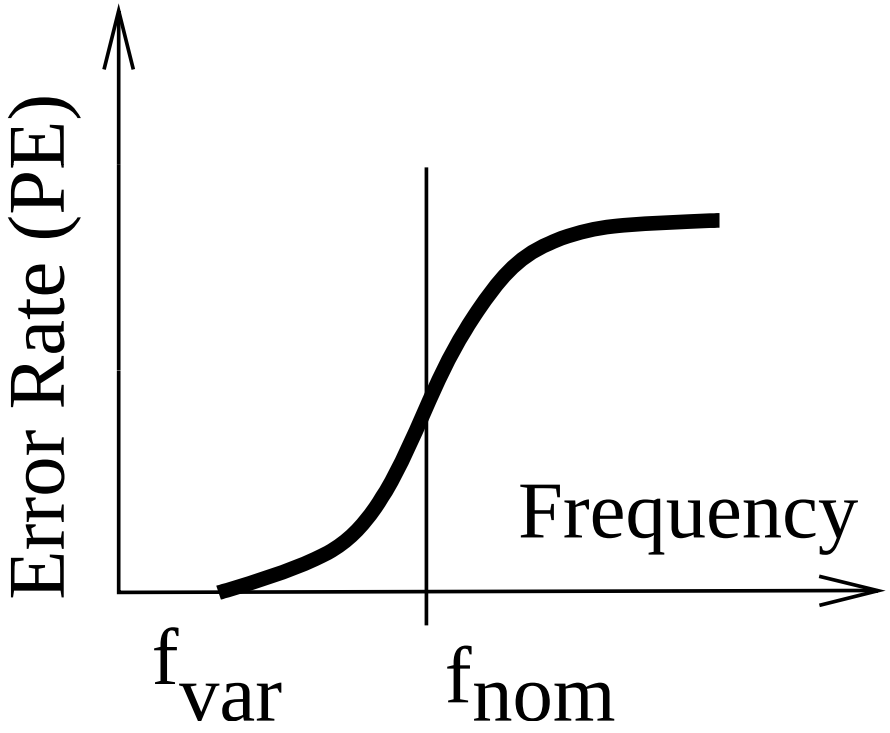


Error rate: P_E (t) = 1 – cdf(t)

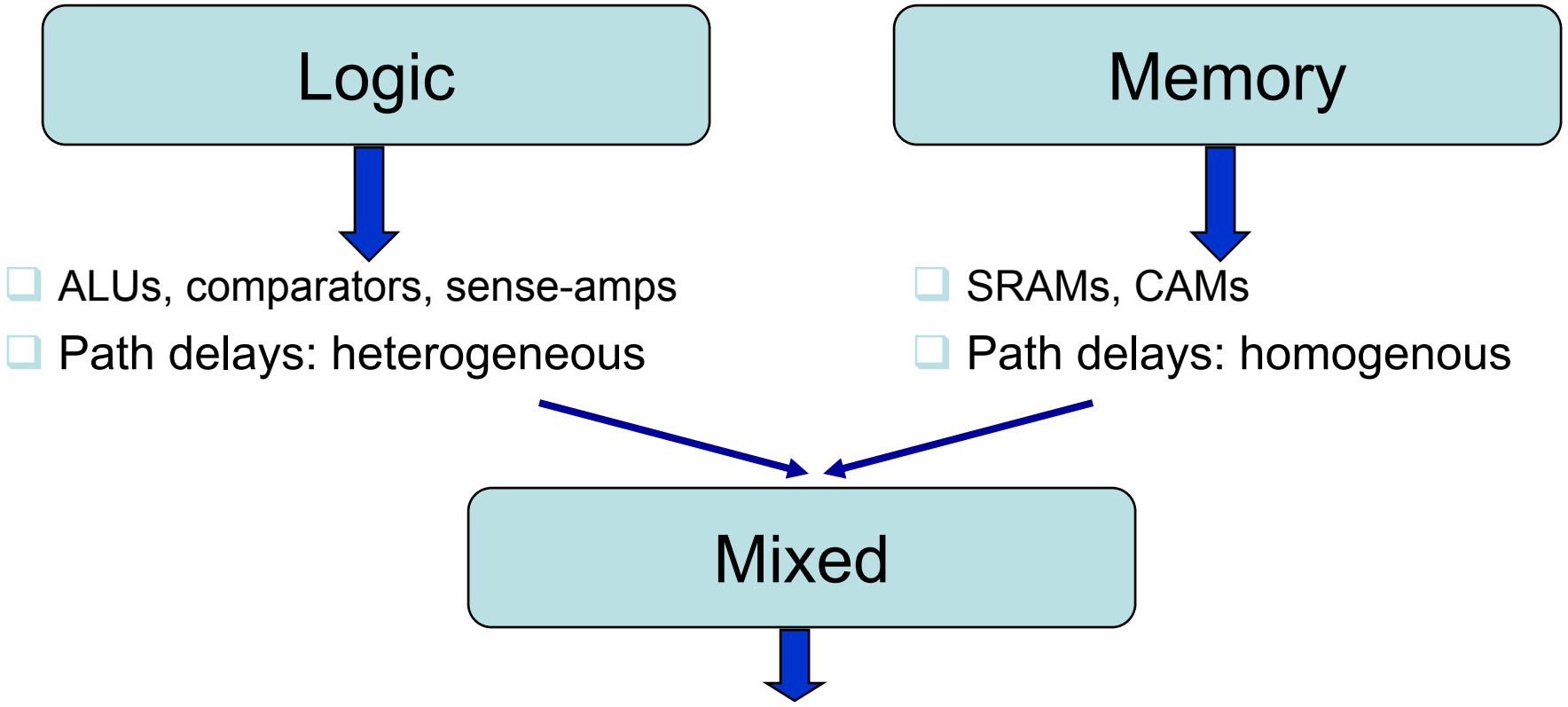


ILLINOIS

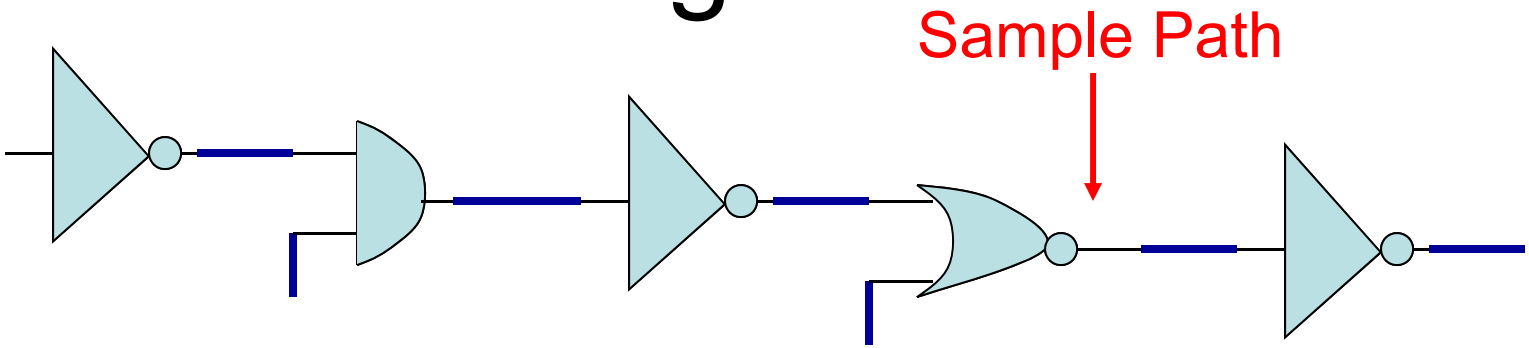
Error-rate vs Frequency



Basic Kinds of Structures



Logic



35% Wiring

65% Logic

Elmore Delay Model

Alpha Power Law

$$T_g \propto \frac{L_{eff} V_{DD}}{\mu(T)(V_{DD} - V_{th})^{\alpha}}$$

Logic Delay

Distribution of path delays – no variation

$d_{wire} + d_{gate} = 1$

$$D_{varlogic} = (d_{wire} + \eta * d_{gate}) * D_{logic} + d_{gate} * D_{extra}$$

Distribution of path delays with variation

Relative gate delay due to systematic variation in P,V, T

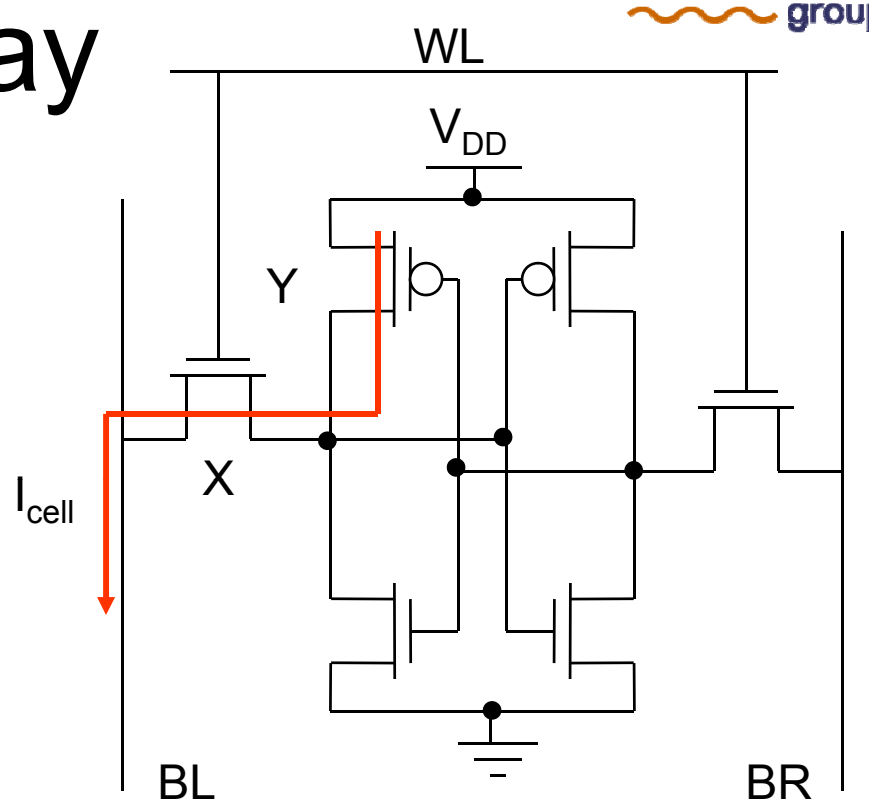
Delay due to random variation

- Obtain D_{logic} using a timing analysis tool

Memory Delay

$$T_{mem} \propto \frac{1}{I_{cell}}$$

- Solve for I_{cell} using long channel eqns.
- $I_{cell} = f(Vt_X, Vt_Y, L_X, L_Y)$
- Vt_X, Vt_Y, L_X and L_Y are gaussian variables

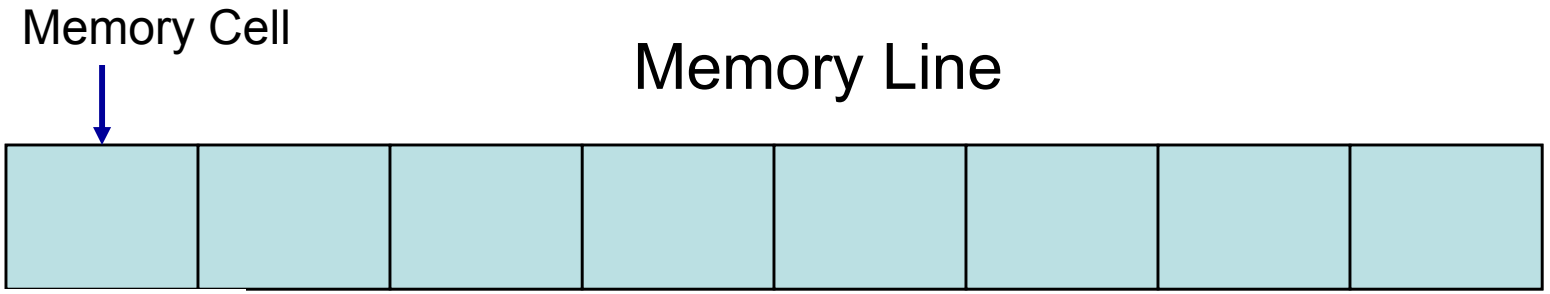


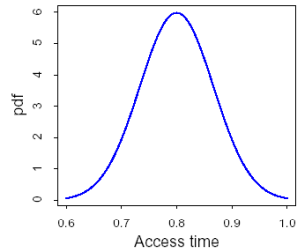
- μ_{vtx}, μ_{vty}, μ_{lx}, μ_{ly} are the systematic components
- σ_{vtx}, σ_{vty}, σ_{lx}, σ_{ly} are the random components

Memory Delay - II

- Find a distribution for T_{mem}
  - T_{mem} is a function of four gaussian variables
  - Model T_{mem} as a normal distribution
  - Find the μ and σ for T_{mem} using multi-variable Taylor expansion
  - This is the access time dist. for 1 bit
- A typical entry has 32-128 bits
  - Find the max distribution of 32-128 normal variables
- Error probability = 1 – cdf(t_{mem})

Memory Delay



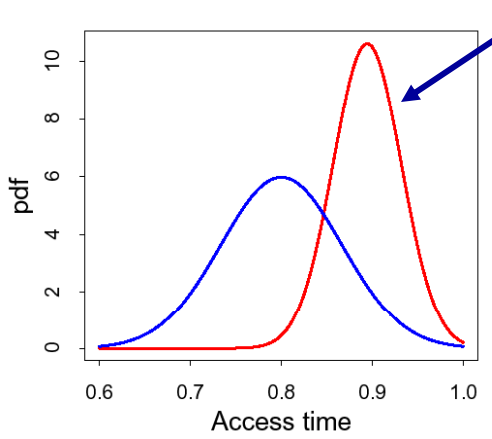


Delay dist.

- Use Kirchoff's equations
- Long channel trans. equations
- Multi-variable Taylor expansion

max. distribution

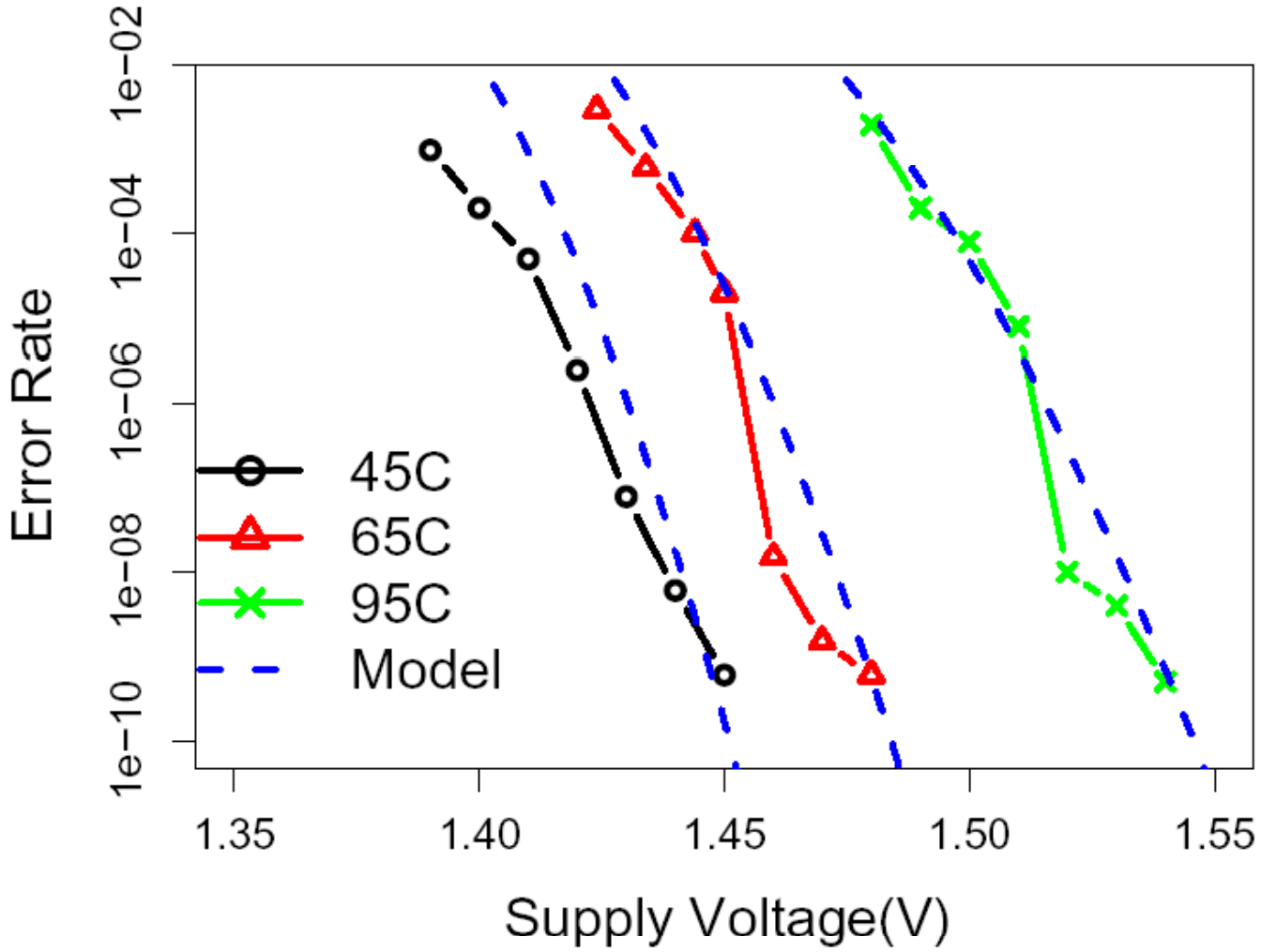
$Delay_{line} = max(Delay_{cell})$



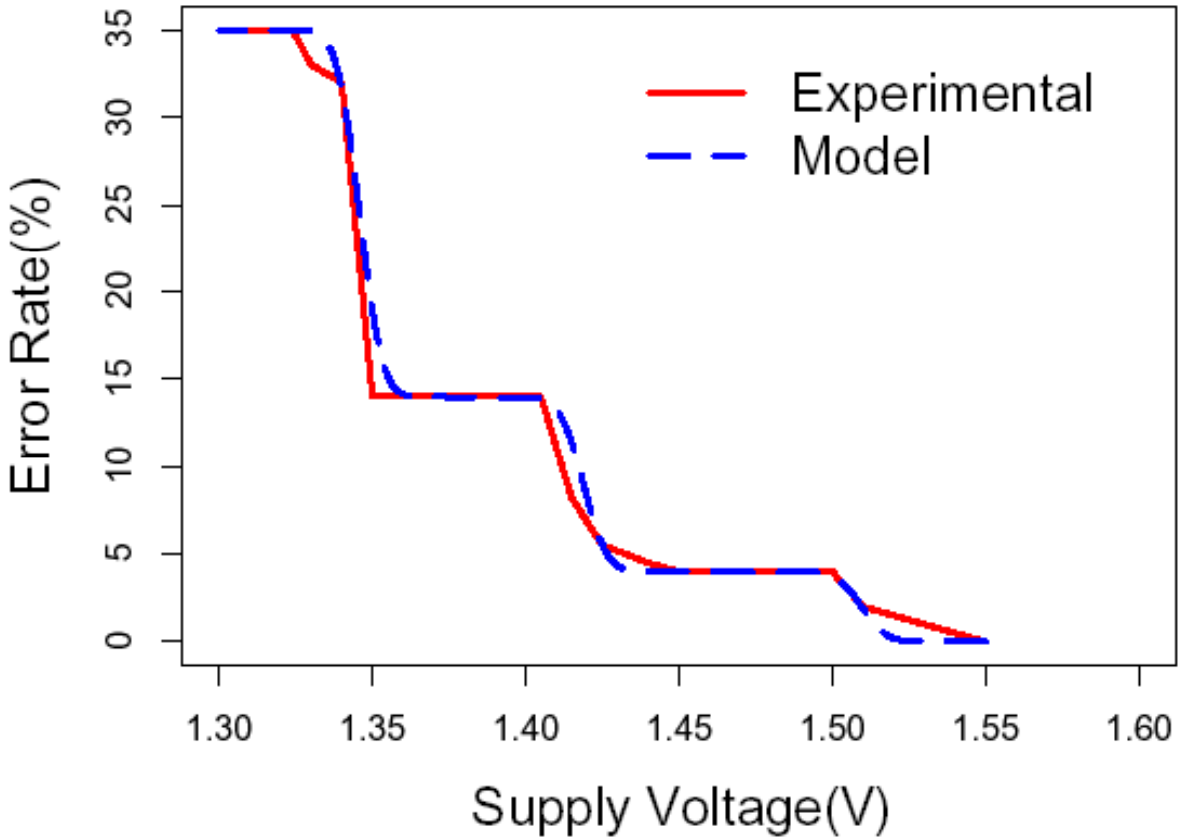
Combined Error Model

- We have the delay distributions
- For each structure
  - per access, P(E) = 1 – cdf(t)
  - P(E) per inst = αP(E) , α=accesses/inst.
- Combined error rate per instruction
  - $P(E)_{total} = \Sigma\ \alpha P(E)$
- CPI penalty per instruction
  - recovery_penalty * $P(E)_{total}$

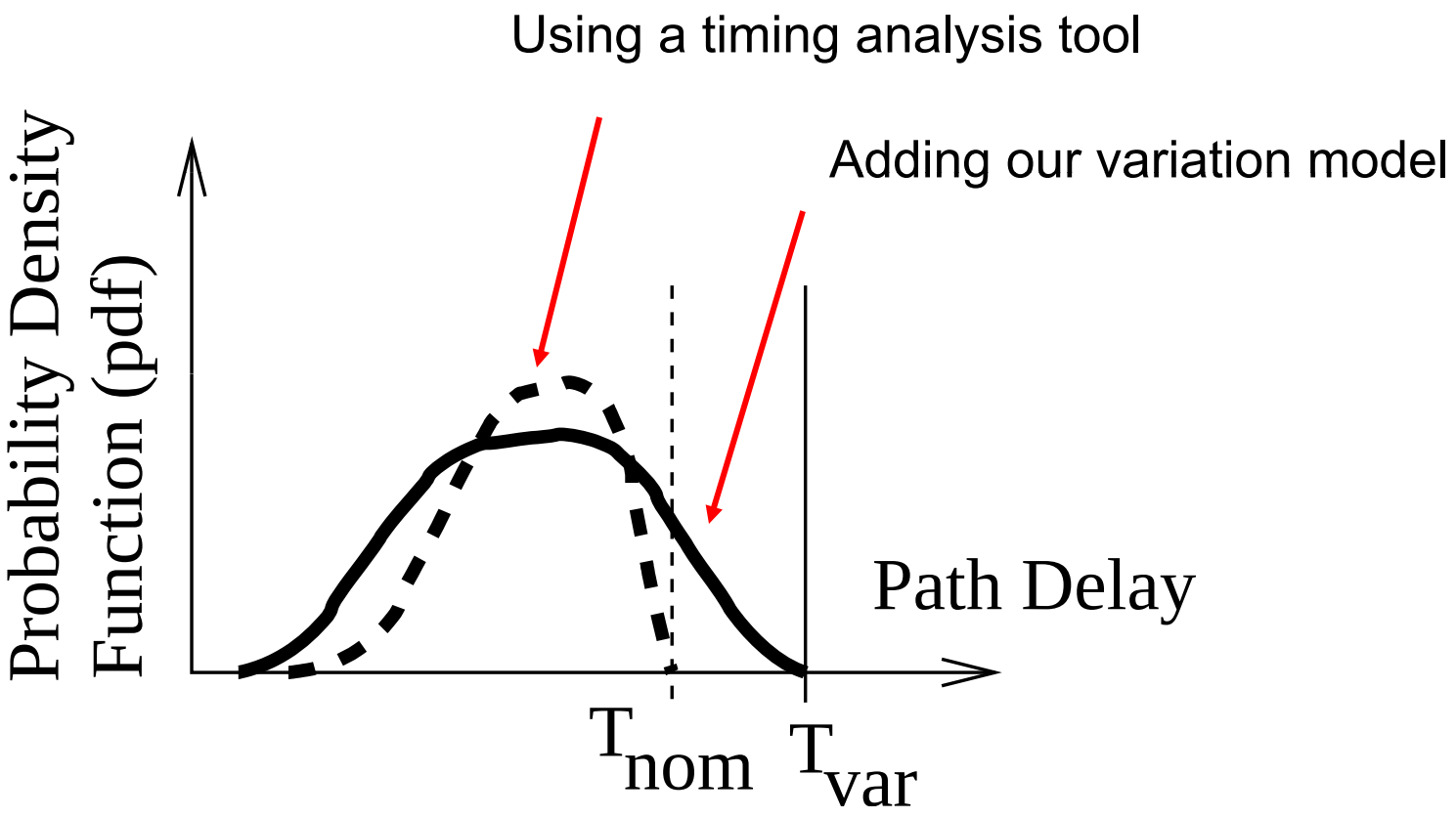
Validation – Logic



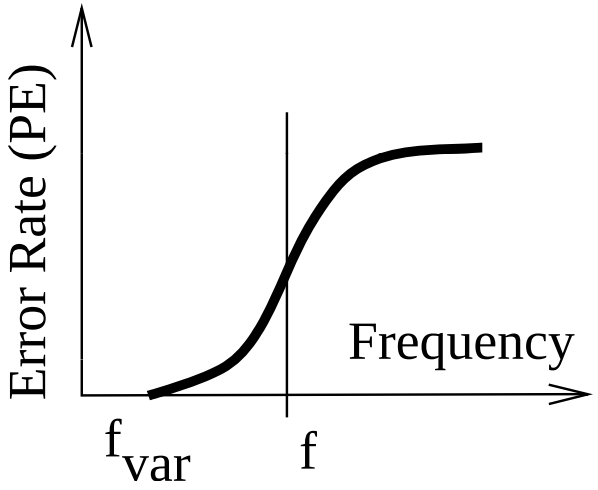
Validation – Memory



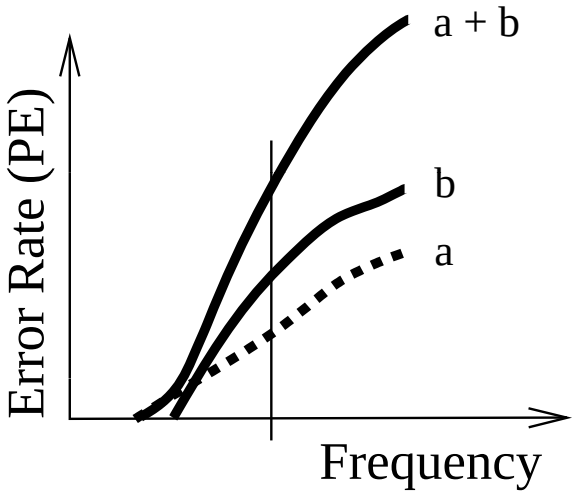
Finding out the Distributions



Adding Up all Pipe Stages

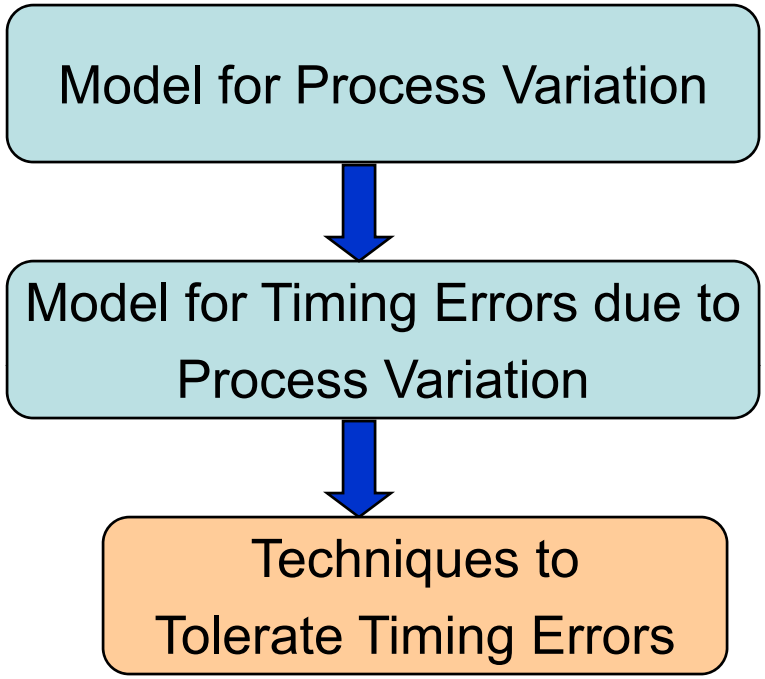


Whole processor

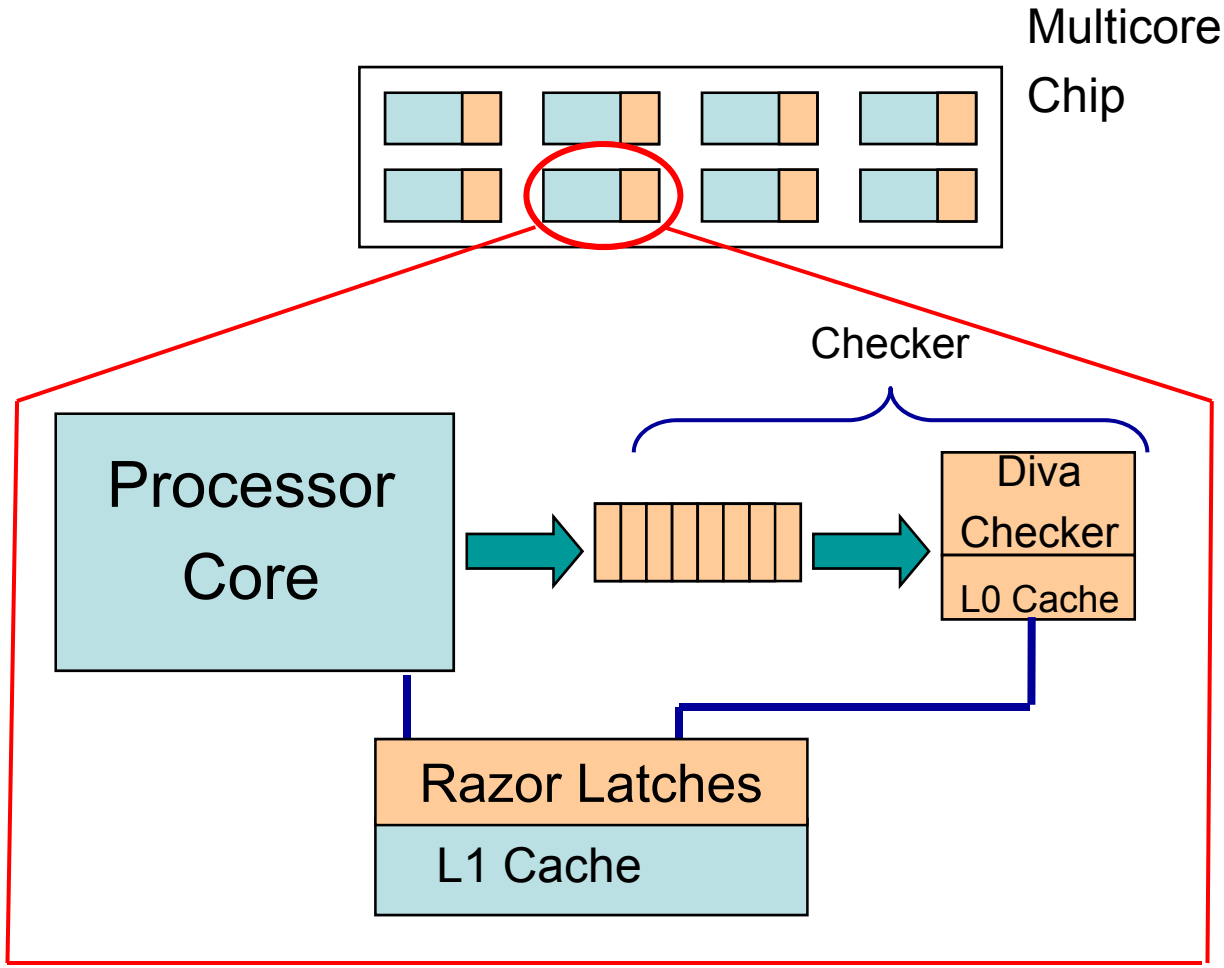


$$P_E(f) = \sum_i (\alpha_i \times P_{Ei}(f))$$

Overview


Model for Process Variation
Model for Timing Errors due to Process Variation
Techniques to Tolerate Timing Errors

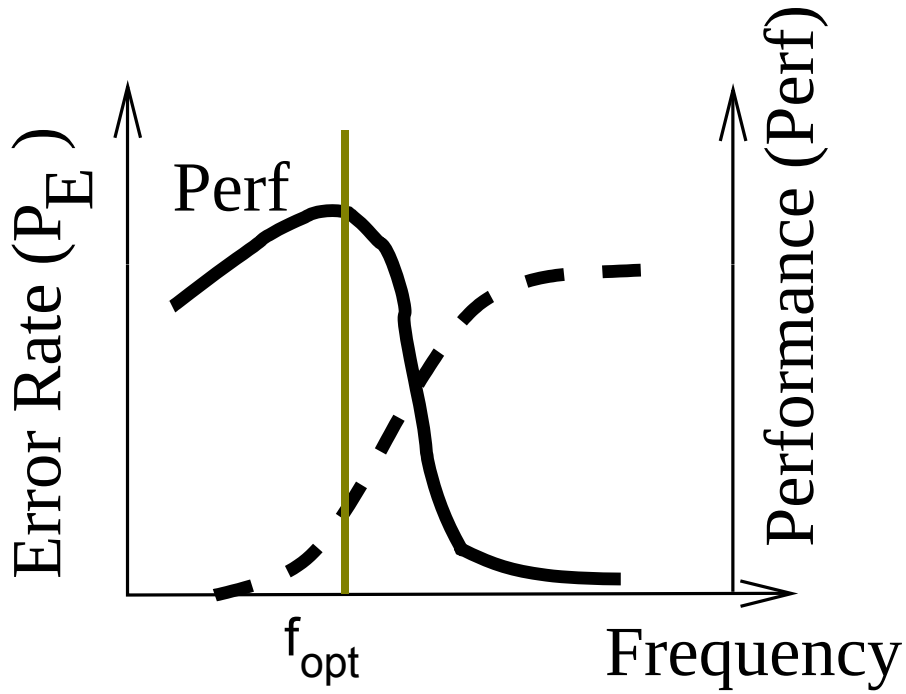
Variation Aware Timing Speculation (VATS)



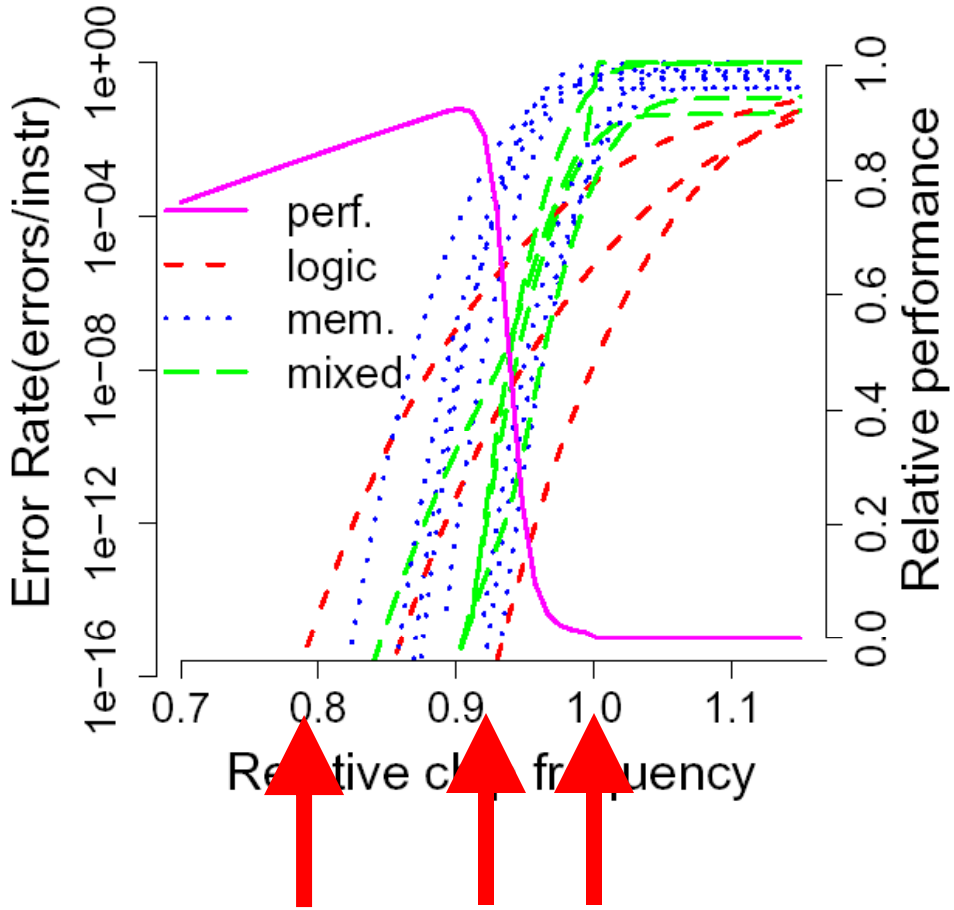
Performance vs Frequency

$$Perf(f) = \frac{f}{CPI_{comp} + CPI_{stall\_mem} + CPI_{rec}}$$

$P_E(f)$ x recovery_penalty



AMD Athlon-like Processor



Conclusion

- Micro-architects can help solve par variation
  - Cores that assume faults occur all the time
  - Frequency / Power / Error rate are tradeable
  - Techniques to mitigate variation-induced errors
  - Develop models that give insights
  - Work with circuits, CAD, and software folks