Delay Metrics for the Next 50 Years

Larry Pileggi
Carnegie Mellon University
Department of Electrical and Computer Engineering
Outline

◆ Introduction
  – Interconnect delay dominance
  – Back-end models and analyses
  – Front-end metrics
◆ Elmore delay
  – Introduced in 1948
  – Applied to digital IC problems in early 1980’s
  – Somewhat ineffective for deep submicron (DSM)
◆ Probability Interpretation of Moments (PRIMO)
◆ Stable n-Pole Models (SnP)
◆ Conclusions
Interconnect Dominance

- Metal resistance per unit length is increasing, while gate output resistance is decreasing, with scaling
- Average wire lengths are not scaling, so portion of delay associated with the interconnect is increasing

- Gate delay is further decreasing with increasing metal resistance due to shielding effects
- DSM interconnect dominance impacts all aspects of the top-down design flow
Model order reduction via moment matching can be used effectively for interconnect verification.

Orthonormalized moments, or Krylov subspace methods were recently proposed for increased numerical accuracy.

\[ Y(s) = m_0 + m_1 s + m_2 s^2 + \ldots \]
Krylov Reduction Methods

- Same as moment matching if we have infinite precision
- Can capture dozens of dominant poles
- Approximations to the 10’s of gigahertz is straightforward
- Some issues remain to be solved with regard to passivity

Coupled RLC Lines

Frequency (GHz)

Y11(s) 60 poles

exact
MPVL
PRIMA

Frequency (GHz)
But there are very few applications which require this level of detail.

There is a greater need for improved interconnect modeling at the front-end and physical design levels.
Catching all of the interconnect problems at back-end is too late!
Front-end Metrics

- Even with an approximate interconnect topology and values, moment matching and Krylov subspace methods are inappropriate for the front-end of design.
- Higher order moments can be calculated at a fraction of the cost [RICE] required to calculate the first one.

\[
V(s) = m_0 + m_1s + m_2s^2 + \ldots
\]

- But calculating the delays requires nonlinear iterations.

\[
v(t) = 0.5V_{DD} = \sum_{i=1}^{n} k_i e^{p_i t_d}
\]
The Elmore Delay

- Metric of choice for front-end applications and performance-driven physical design
- Explicit delay metric, yet can still capture interconnect resistance effects
- Primarily applied to RC tree circuits [Penfield & Rubenstein]

\[ T_D = R_1(C_1 + C_2 + C_3 + C_4) + R_2(C_2 + C_3 + C_4) + R_4C_4 \]

- The first moment of the impulse response

\[ H(s) = m_0 + m_1s + m_2s^2 + ... \]
Elmore (1948) proposed to treat the derivative of a monotonic step response as a PDF, and estimate the median (50% delay point) by the mean.
The Elmore Delay

- Exact only if $h(t)$ is symmetrical
- We’ve proven that RC tree impulse responses have positive skew

mean = 134 ps
median = 100 ps
Central Moments

- The circuit response moments

\[ H(s) = \frac{1+a_1 s + \ldots + a_n s^n}{1+b_1 s + \ldots + b_m s^m} = m_0 + m_1 s + m_2 s^2 + \ldots \quad \rightarrow \quad m_q = \frac{(-1)^q}{q!} \int_0^{\infty} t^q h(t) dt \]

are related to the Central Moments of the \( h(t) \) Distribution by:

\[ \mu_1 = m_1 \equiv \text{mean} \quad \quad \mu_n = \sum_{k=0}^{n} \binom{n}{k} m_k (-m_1)^{n-k} \]

\[ \mu_2 = 2m_2 - m_1^2 \equiv \text{variance} \quad \mu_3 = -6m_3 + 6m_1 m_2 - 2m_1^3 \]

- Roughly speaking:

\[ \text{Skew} = \frac{\mu_3}{\mu_2^{1.5}} = \frac{\text{Mean} - \text{Median}}{\sqrt{\mu_2}} \]
Elmore Delay Bound

- Skew is a measure of the asymmetry

- We proved that all RC interconnect trees:
  - have unimodal impulse responses, \( h(t) \)
  - and that the \( h(t) \) distributions have positive skew

- It is then easily shown for such a distribution that
  \[
  \text{Mode} \leq \text{Median} \leq \text{Mean}
  \]

- The Elmore delay is an upper bound on the 50% step response delay
Elmore Bound

- Bounds get tighter toward the interconnect loads
- Repeated convolutions make the distributions more “normal” --- positive skew decreases toward a constant value
Finite Rise Times

- Any input voltage with a unimodal derivative will also make the response more normal (finite $t_{in}$) --- and the first moment bound still holds.

- For finite rise times, the pulse response distribution becomes more symmetrical as the rise time increases.

\[ V_{in}(t) \xrightarrow{\text{derivative}} V'_{in}(t) \xrightarrow{\text{response}} v'_{out}(t) \]

- In the limit, the mean of the pulse response equals the median and the Elmore delay becomes exact.

- A large percentage of responses will fall into this category.
Elmore Errors

- 1200 response nodes for 700 nets from a 0.35 micron CMOS µP

100ps rise times

max
Dominant Time Constant

- The Elmore delay as a dominant time constant

\[ H(s) = \frac{(s-z_1)(s-z_2)\ldots(s-z_n)}{(s-p_1)(s-p_2)\ldots(s-p_m)} \quad \rightarrow \quad m_1 = \sum_{i=1}^{m} \frac{1}{p_i} - \sum_{i=1}^{n} \frac{1}{z_i} \]

- If one time constant dominates all others, and there are no low frequency zeros, we can approximate the dominant pole by \( m_1 \)

\[ m_1 \approx \tau_1 \]

- This approximation only scales the step response delay by a constant factor

\[ t_{delay} = \ln(0.5)m_1 \approx 0.7 \cdot m_1 \]
Max error is reduced, but ramp follower responses are optimistic.
Using The Elmore Delay

- Not a good approximation for general DSM trees
- Worst case error for busses with near- and far-end loads
- Works well when the rise time is slow
- Or for balanced interconnects such as clock trees
Given the following floorplan for a μP clock tree, optimize the metal widths in terms of the Elmore delays to balance the skew.

R and C per unit length values are pre-layout estimates.
µP Clock Tree

- Widths for zero Elmore skew produced 8 ps of skew with moment-matching models

- But correlations for optimization of signal paths are not as good

- Signal paths require small absolute errors, whereas clock trees require only small relative errors
Higher Order Metrics

- For signal nets it would appear that we should match 3 moments minimally
- Capture shapes for good relative errors
- But we can’t afford nonlinear iterations for most delay metric applications

- Two potential approaches:
  - PRIMO
  - SnP
PRIMO - Gamma Functions

- Extend Elmore’s idea to matching other distribution properties
- Requires selection of some representative distribution
- Incomplete gamma is similar to RC impulse responses

\[ g_{\lambda,n}(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{\Gamma(n)} \]

\[ \Gamma(x) = \int_0^\infty y^{x-1} e^{-y} \, dy \]

- Moment matching \( m_1, \mu_2, \) and \( \mu_3 \) for time-shifted incomplete gamma is provably stable

\[ n = \frac{4(\mu_3)^3}{(\mu_3)^2} \quad \lambda = \frac{2\mu_2}{\mu_3} \]
Since provably stable, a gamma integral table can be used for delays

With rise time a 2D table is required

For this example the step-delay error is < 1%
Gamma Fitting

- Gamma approximation **struggles** for some cases
- DSM interconnects can have complex low frequency zero effects
- Step delay error is underestimated by 8% for this example
We can build provably stable n-pole approximations.

Driving point pole approximations are provably stable.

k’s are fitted by matching moments at the response nodes of interest.

Generates stable n-exponential distribution model which permits table lookup evaluation.

\[ I(s) = m_0 + m_1 s + m_2 s^2 + \ldots \]

\[ h(t) = \sum_{i=1}^{n} k_i e^{-t/\tau_i} \]
A two-pole model, or double exponential distribution function, can be used with a 3D table to evaluate finite rise time response delays.

Step delay error is less than 1.5% in this example.
S2P: Double Exponential

- CMOS µP example

![Graph showing percentage error](image-url)
It can be shown that three exponentials are minimally required to fit some unimodal impulse responses.

S2P step delay error is 14% in this example.

But is a 4D table practical?
**Inductance**

- On-chip inductance is becoming a reality for long lines
- Impulse responses are no longer unimodal
- Skew measure ($\mu_3$) can be used to control damping

**Packaging Example**

- $R = 1\ \text{ohm/cm}$
- $L = 0.335\ \text{nH/cm}$
- $C = 0.134\ \text{pF/cm}$
- $Z_0 = 50\ \text{ohms}$

**Graph**

- $\mu_3$ (scaled)
- $R$

3/6/98
The delays are accurately predicted by the moment metrics once the damping is controlled.
Conclusions

- Some progress has been made on more accurate delay metrics
- But more work remains to be done for the most difficult DSM problems
- Similar metrics for coupling are necessary
- But coupled line responses are provably not unimodal for the general case