ENVIRONMENTAL LIMITS ON THE PERFORMANCE OF CMOS WAVE-PIPELINED CIRCUITS

Kevin J. Nowka and Michael J. Flynn

Technical Report CSL-TR-94-600

January 1994

Partially supported by an ARPA Fellowship in High Performance Computing administered by the Institute for Advanced Computer Studies, University of Maryland. Additional support for this work from NSF Contract No. MIP88-22961 using facilities provided by NASA under contract NAG2-842.

ENVIRONMENTAL LIMITS ON THE PERFORMANCE OF CMOS WAVE-PIPELINED CIRCUITS

by

Kevin J. Nowka and Michael J. Flynn

Technical Report CSL-TR-94-600

January 1994

Computer Systems Laboratory

Departments of Electrical Engineering and Computer Science

Stanford University

Stanford, California 94305-4055

Abstract

Wave-pipelining is a circuit design technique which allows digital synchronous systems to be clocked at rates higher than can be achieved with conventional pipelining techniques.

Wave-pipelining has been successfully applied to the design of SSI processor functional units[1], a Bipolar Population Counter[9], a CMOS adder[10], CMOS multipliers[3] [8], and several simple CMOS circuits. For controlled operating environments, speed-ups of 2 to 10 have been reported for these designs.

This report details the effects of temperature variation, supply voltage variation, and process variation on wave-pipelined static CMOS designs, derives limits for the performance of wave-pipelined circuits due to these variations, and compares the performance effects with those of traditional pipelined circuits.

This study finds that wave-pipelined circuits designed for commercial operating environments are limited to 2 to 3 waves per pipeline stage when clocked from a fixed frequency source. Variable rate, internal clocking can approach the theoretical limit of waves at a cost of interface complexity.

Key Words and Phrases: Wave-pipelining, pipelining, propagation delay, clocking

Contents

List of Figures

List of Tables

1 Background

Wave-pipelining is a circuit design technique which allows digital synchronous systems to be clocked at rates higher than can be achieved with conventional pipelining techniques. Wave-pipelining relies on the finite propagation delay of a combinational digital circuit to store data. Rather than allowing data to propagate from a synchronizing element, latch or register, through the combinational network to another synchronizing element prior to initiating the subsequent data transfer, wave-pipelined designs apply subsequent data to the network as soon as it can be guaranteed that it will not interfere with the current data wave. In this manner, multiple waves of data are simultaneously propagating through distinct regions of the logic network.

Wave-pipelining has been successfully applied to the design of SSI processor functional units[1], a Bipolar Population Counter[9], a CMOS adder[10], CMOS multipliers[3] [8], and several simple CMOS circuits. These designs have demonstrated speed-ups of 2 to 10 over their non-pipelined counterparts.

Several formalizations of the constraints on clocking of wave-pipelined circuits have been published[9] [7] [20] [12]. This analysis applies these constraints to CMOS systems to arrive at practical limitations on the design and performance of wave-pipelining in CMOS.

This report details the effects of temperature variation, supply voltage variation, and process variation on wave-pipelined static CMOS designs, derives limits for the performance of wave-pipelined circuits due to these variations, and compares the performance effects with those of traditional pipelined circuits. CMOS logic propagation delay models and HSPICE simulations are used to determine the performance impact of temperature and supply variation on wave-pipelined CMOS circuits. Simulations based upon parameters from a variety of fabrication runs and on specified corner parameters are used to assess the effect of process variation on the design of wave-pipelined circuits.

2 Wave-Pipeline Circuit Model

To improve throughput, a logic network can be partitioned into pipeline stages, each of which operates upon data computed in the previous cycle by the previous pipeline stage. When a logic network is pipelined, synchronizing elements, either latches or registers, are inserted to partition the network into stages. These synchronizing elements increase the network area, power, and latency.

Wave-pipelining is a design style which allows overlapped execution of multiple operations without using synchronizing elements. Rather, a knowledge of the signal propagation delay through the network is used to ensure that operations do not interfere with their predecessor nor successor data values.

Figure 1 is a block diagram of an nonpipelined circuit, a pipelined version of the same circuit, and a wave-pipelined equivalent.

For this analysis, the following nomenclature is used:

$T_{max}(V, \tau, P)$	Maximum propagation delay from combinational network inputs to given node in network. Function of supply voltage, temperature, and fabrication process.
$T_{min}(V, \tau, P)$	Minimum propagation delay.
$RF_{min}(V, \tau, P)$	Minimum rise or fall delay of shortest path in network.
$RF_{max}(V, \tau, P)$	Maximum rise or fall delay of longest path in network.
$\Delta C(V, \tau, P)$	Maximum unintentional clock skew.
$\delta C_{io}(V, \tau, P)$	Maximum intentional clock skew of output synchronizer clock with respect to input synchronizer clock.
$T_s(V, \tau, P)$	Minimum setup time.
$T_h(V, \tau, P)$	Minimum hold time.
$T_{clk}, T_{clk}(V, \tau, P)$	Clock Period. May or may not be a function of supply voltage, temperature, and fabrication process.
$T_{trans}, T_{trans}(V, \tau, P)$	Time during which latches are transparent.
$T_{synch}(V, \tau, P)$	Propagation delay through synchronizing elements.

Traditional pipelined synchronous circuits must meet race-through and long-path timing constraints. The race-through constraint requires that in the same clock cycle data cannot propagate out of a synchronizing element, through the combinational network, and into the next synchronizing element. Thus the minimum propagation time though the synchronizing element, through the network, to the next synchronizing element is less that the time from the output initiating edge to the latching edge of the same cycle. Thus the data resulting from the current input data cannot interfere with the previous results in the next synchronizing element.

Long-path constraint requires that the results from the current cycle's inputs are valid at the next synchronizing element prior to the next cycle. Thus the propagation from the synchronizing element, through the data network, to the next synchronizing element is less than the time from the initiating edge of the current clock cycle to the latching edge of the next clock cycle.

In addition to meeting the race-through and long-path constraints, wave-pipelined circuits require that waves of data do not interfere with each other at the output synchronizing element. This constraint result in the following inequality:

$$T_{clk} > T_{max} - T_{min} + 2\Delta C + T_s + T_h + \frac{RF_{min} + RF_{max}}{2} \tag{1}$$

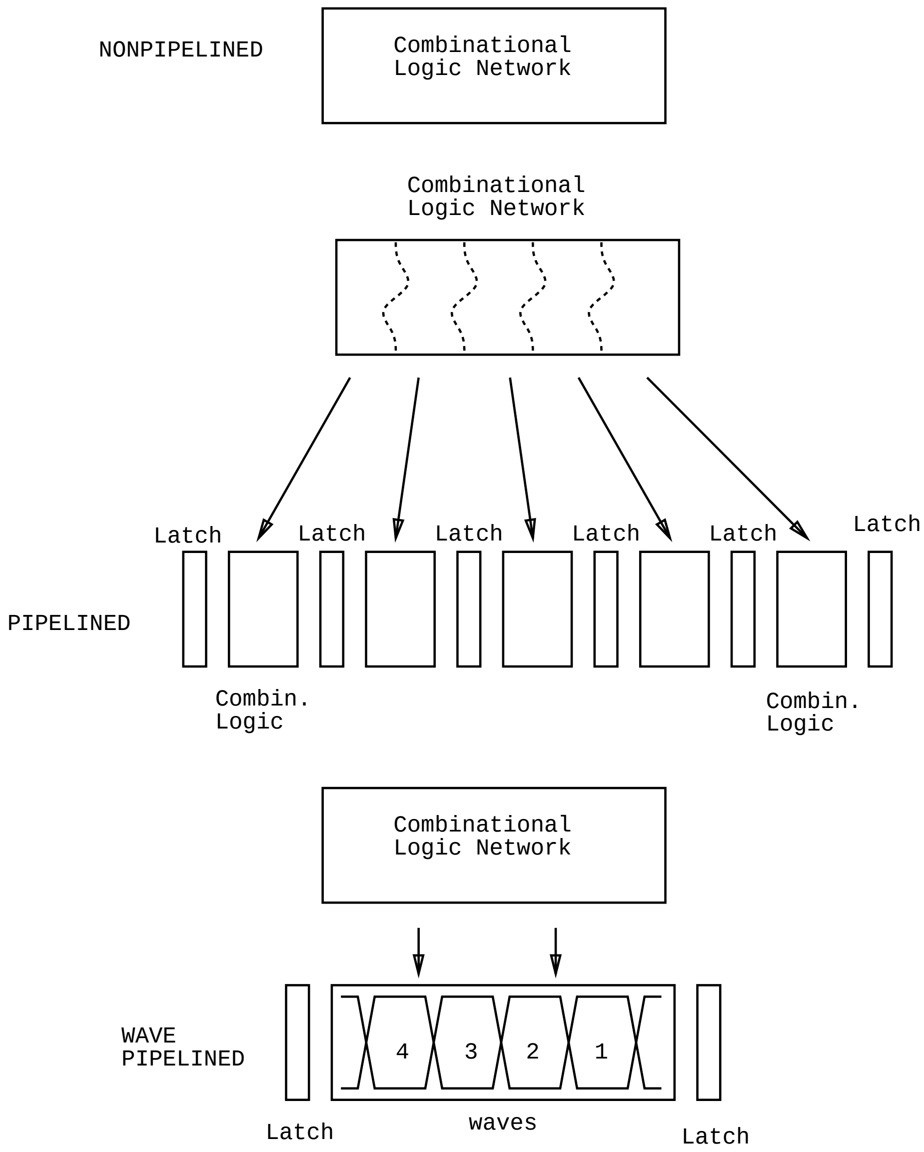


Figure 1: Circuit model

In addition to the output constraint, wave-pipelined circuits can not allow wave interference at any point in the network. This can be represented by the following:

$$T_{clk} > T_{max} - T_{min} + \Delta C + T_{ms} + \frac{RF_{min} + RF_{max}}{2} \qquad (2)$$

where T_{ms} is the minimum amount of time a node voltage must be stable to ensure the subsequent level of logic operates correctly.

Details of the timing constraints for pipelined and wave-pipelined circuits are found in [9].

For wave-pipelined circuits in which the output is latched using the same clock which enables the inputs, the combinational network must hold an integral number of waves. Thus: [20]

$$N * T_{clk} + \delta C_{io} > T_{max} + \Delta C + T_s + \frac{RF_{max}}{2} + T_{synch} \qquad (3)$$

where N is the number of waves of data propagating concurrently through the network.

This inequality ensures that a wave has sufficient time to propagate to the output synchronizing element prior to being latched. In addition, the subsequent wave must not reach the synchronizer prior to the synchronizing clock edge. For edge-triggered registers as synchronizing elements:

$$(N-1) * T_{clk} + \delta C_{io} < T_{min} - \Delta C - T_h - \frac{RF_{min}}{2} + T_{synch} \qquad (4)$$

For transparent latches as synchronizing elements:

$$(N-1) * T_{clk} + T_{trans} + \delta C_{io} < T_{min} - \Delta C - \frac{RF_{min}}{2} + T_{synch} \qquad (5)$$

3 Effects of environmental variation on propagation delay

We now look at how variations in the environment affect the operation of the network and in turn, the wave-pipelining constraints.

The clock rate of wave-pipelined circuits is constrained by the worst-case variation of propagation delay through the network. The sources of variation in the network are:

1. Variations due to differences in propagation of signals along different paths.
2. Variations due to differences in the state of network node voltages (data dependencies.)

3. Variations due to changes in operating temperature.

4. Variations due to supply voltage drift and noise.

5. Variations due to fabrication process variations.

6. Variations due to signal noise.

The variation in propagation delay due to path length differences and data dependencies are discussed in detail in [9] [2]. These studies have shown that in CMOS circuits, by using delay balancing techniques, path length differences variation can be minimized.

By implementing functions in relatively input pattern insensitive logic such as NAND2/INV static CMOS, delay variation can be limited to less than 10% for a 4-bit CLA. [2]

This study focuses upon the limitations of wave-pipelined circuits due to the device operating temperature, supply voltage, and fabrication process.

3.1 Propagation Delay

For simplicity, we define propagation delay as the time from the controlling input reaching 50% of its terminal value to the output reaching 50% of its terminal value.

We use an Elmore Model for the delay of the network [11]. In this model, the propagation delay along a path in a logic network is the sum of the step-input delays of the individual gates along the path.

The step-input propagation delay of a CMOS gate, T_{pd}, consists of the time it takes for a load capacitance to be charged or discharged from its initial voltage to 50

$$T_{pd} = C_l * \int \frac{dV_{ds}}{I_{ds}(V_{ds}, \tau, P)} \tag{6}$$

where C_l is the total load capacitance, τ represents operating temperature, and P distinguishes the fabrication process.

To estimate T_{pd}, gates are represented as a single transistor, sized so as to match the current carrying capacity of the complex gate, charging or discharging a fixed load capacitance.

Using long-channel MOS current equations, we can derive equations for propagation delay for low-going outputs assuming step input:[5]

$$T_{phl} = t_1 + t_2 \tag{7}$$

$$t_1 = \frac{2C_l V_{tn}}{K_n (V_{dd} - V_{tn})^2} \tag{8}$$

$$t_2 = \frac{C_l}{(V_{dd} - V_{tn}) K_n [ln(\frac{3V_{dd} - 4V_{tn}}{V_{dd}})]} \tag{9}$$

For short-channel MOS devices, where velocity saturation limits channel current, the propagation delay for low-going outputs assuming step input is:

if $V_{dmax} > V_{dd}/2$,

$$T_{phl} = t_1 + t_2 \tag{10}$$

$$t_1 = \frac{2C_l (V_{dd} - V_{dmax})}{K_n V_{dmax}^2} \tag{11}$$

$$t_2 = \frac{C_l}{K_n [\frac{1}{V_{dd} - V_{tn}} ln(\frac{V_{dmax}(1.5V_{dd} - 2V_{tn})}{V_{dd}(V_{dd} - V_{tn} - 0.5V_{dmax})}) + (2/V_{sat}) ln(\frac{1.5V_{dd} - 2V_{tn}}{2V_{dd} - 2V_{tn} - V_{dmax}})]} \tag{12}$$

else,

$$T_{phl} = \frac{C_l V_{dd}}{K_n V_{dmax}^2} \tag{13}$$

where,

$$V_{sat} = \frac{L * v_{sat}}{\mu_n} \tag{14}$$

$$V_{dmax} = V_{sat}[(1 + \frac{2(V_{dd} - V_{tn})}{V_{sat}})^{0.5} - 1] \tag{15}$$

$$K_n = \mu_n C_{ox} W/L \tag{16}$$

and W is channel width, L is channel length, μ_n is electron channel mobility, C_{ox} is per area oxide capacitance, v_{sat} is the saturation velocity, and V_{tn} is nmos threshold.

Corresponding equations for the propagation from low to high can be derived by substituting p-channel threshold and gain for the n-channel parameters.

Because wave-pipelining is constrained by relative differences in propagation delay rather than maximum propagation delay, we need only derive ratios of propagation delays to those at nominal operating conditions.

3.2 Temperature Variation

Temperature variation is both spatial and temporal. As a transistor conducts current, heat is conducted through the surrounding die area resulting in changes in local temperature.

The variation in propagation delay due to temperature is primarily the result of the variation of the channel current of the conducting MOS device. The variation of channel current with temperature is strongly related to the change in channel carrier mobility. Therefore, we will model the variations in propagation delay as a function of variations in mobility. Secondary effects such as threshold reduction and junction capacitance variation are ignored for this analysis.

Empirical studies [13] [6] have shown that the temperature dependence of channel carriers can be represented by:

$$\mu(\tau) = \mu_0(\tau) f_v f_h \tag{17}$$

where f_v and f_h represent degradation factors in the vertical and horizontal directions, respectively.

The temperature dependence of the low-field mobility, μ_0, is;

$$\mu_0(\tau_2) = \mu_0(\tau_1) * (\tau_2/\tau_1)^{-M} \tag{18}$$

where M is an empirical constant between 1.5 and 2. HSPICE uses $M = 1.5$ for level 3 IDS MOS device modeling[16]. τ_1 and τ_2 are absolute temperatures.

Figure 2 shows the ratio of channel carrier low-field mobility at 25 C to that for temperatures from 25 C to 125 C as derived from the above mobility formula with M=1.5.

The variation of mobility results in a corresponding variation in channel current, and in turn, propagation delay. Ignoring the secondary temperature effects, and concentrating on the mobility variation, the propagation delay through a network of long-channel devices at a given temperature to that at the nominal temperature should be the inverse of the mobility ratio as suggested by the propagation delay equations in section 3.1.

Figure 2 data suggests that propagation delays of CMOS logic structures can be as much as 50 to 60% slower at 125C than at 25C due to the differences in mobility.

Figure 3 shows HSPICE simulations of propagation delay of two chains of 50 CMOS inverters over a temperature range of 25 C to 125 C. The short-channel chain consists of inverters with $1.5\mu/0.8\mu$ nmos transistors and $3.5\mu/0.8\mu$ pmos transistors. The long-channel chain consists of inverters with $9\mu/3\mu$ nmos transistors and $21\mu/3\mu$ pmos transistors.

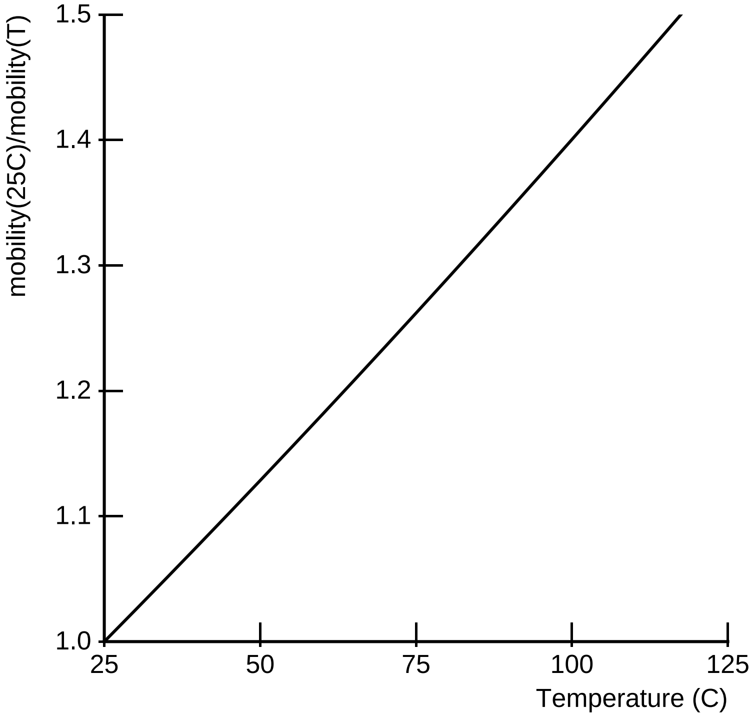


Figure 2: Relative carrier mobilities vs. temperature

Figure 4 shows the ratio of propagation delay of the inverter chains for temperatures from 25 C to 125 C to the propagation delay at 25 C. Superimposed on figure 4 is the ratio of mobilities as given previously. The mobility approximation to relative propagation delay becomes less accurate as temperature is increased due to the assumption of constant thresholds.

Based upon the models of CMOS device behavior and SPICE simulations, the propagation delay of a CMOS network at temperature τ_2 can be approximated by:

$$T_{max}(\tau_2) \approx T_{max}(\tau_1) * (\frac{\tau_2}{\tau_1})^{1.5} \tag{19}$$

$$T_{min}(\tau_2) \approx T_{min}(\tau_1) * (\frac{\tau_2}{\tau_1})^{1.5} \tag{20}$$

For short-channel devices, velocity saturation limits the channel current. Because temperature affects the saturation voltage, the expression for relative propagation delay is more complicated:

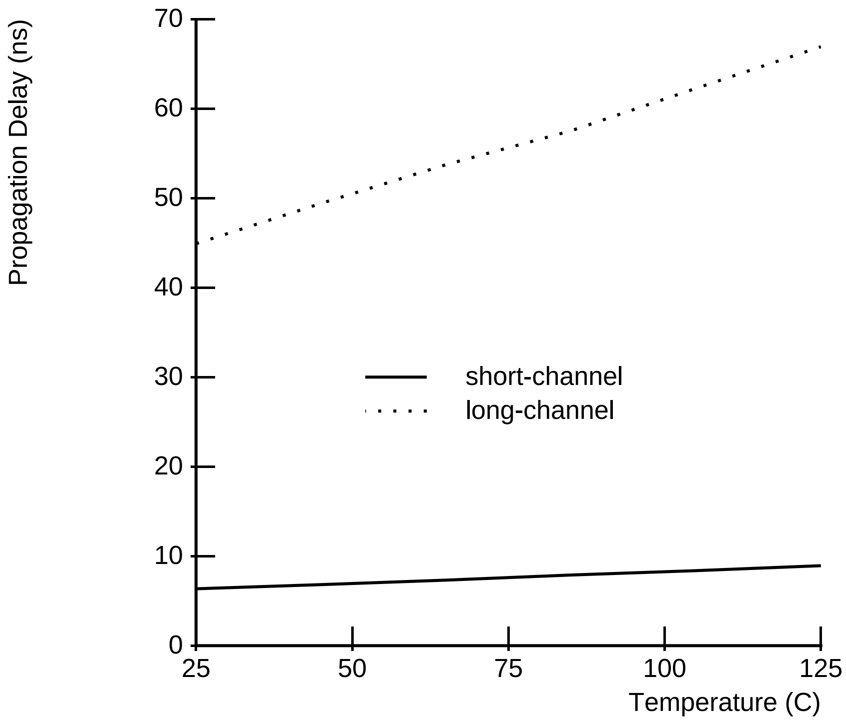


Figure 3: Inverter chain propagation delay vs. temperature

$$T_{max}(\tau_2) \approx T_{max}(\tau_1) * (\frac{\tau_2}{\tau_1})^{1.5} * (\frac{V_{dmax}(\tau_1)}{V_{dmax}(\tau_2)})^2 \tag{21}$$

$$T_{min}(\tau_2) \approx T_{min}(\tau_1) * (\frac{\tau_2}{\tau_1})^{1.5} * (\frac{V_{dmax}(\tau_1)}{V_{dmax}(\tau_2)})^2 \tag{22}$$

Thus, propagation along a given path for a CMOS network will be as much as 50% slower at 125 C than at room temperature.

3.3 Supply Voltage Variation

Supply voltage variation affects propagation delay by altering the channel current and signal voltage swing. Using the delay expressions from section 3.1 for the propagation delay of a capacitor discharging through an n-channel device and charging through a p-channel device, we can derive a first-order expression for the ratio of propagation delay at a given supply voltage to the propagation delay at the nominal supply voltage.

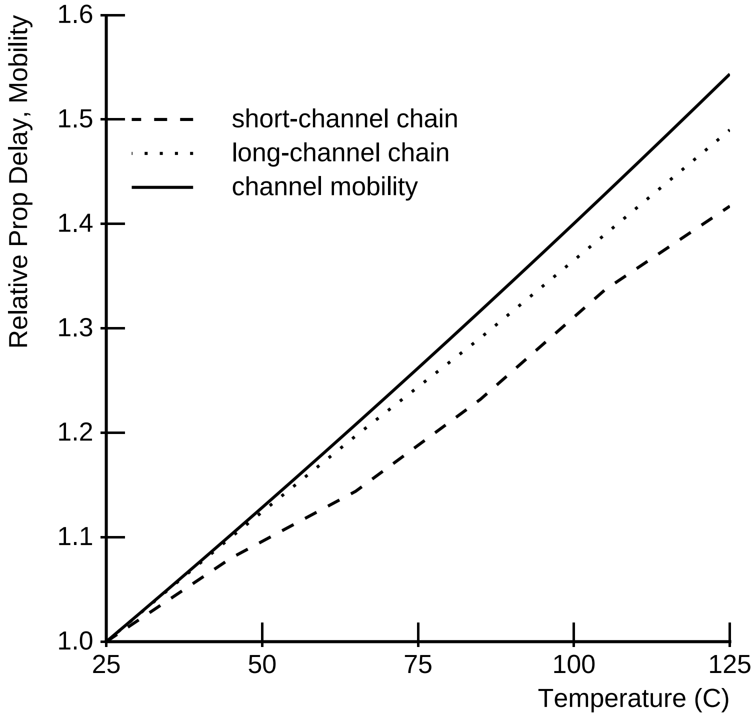


Figure 4: Relative propagation delay vs. temperature

For the simulated process, the model parameters are given in table 1.

Figure 5 shows the propagation delay of a capacitor being driven high and low through a pmos and nmos device, respectively, for a range of supply voltages, relative to the nominal 5V supply.

In the Elmore delay model, the propagation delay through a logic network is the sum of the individual delays. Thus, the ratio of the propagation delays for the network should lie within the charging and discharging ratios.

Figure 6 shows the simulated propagation delay of a minimum-sized balanced inverter driving an identical inverter high and low versus supply voltage.

Figure 7 compares the computed relative propagation delay ratios for rising and falling outputs versus supply voltage. Also included in this figure is the simulated ratios for the short-channel chain of 50 inverters. Figure 8 is a plot of simulated propagation delay verses supply voltage for a nominal supply of 3.3V.

These figures show that propagation delay is less sensitive to fluctuations in supply voltage

Parameter	Value
V_{tn}	0.71V
V_{tp}	-0.90V
V_{dd}	5V
μ_{n0}	572 cm^2/Vs
μ_{p0}	178 cm^2/Vs
C_{ox}	192 nF/cm^2
v_{satn}	1980 cm/s
v_{satp}	3690 cm/s

Table 1: Simulated process parameters

than temperature in the operating ranges. As a first-order approximation the variation in propagation delay due to supply voltage drift is linearly related to the supply voltage. Thus:

$$T_{max}(V_2) \approx T_{max}(V_1) * \frac{V_1}{V_2} \tag{23}$$

$$T_{min}(V_2) \approx T_{min}(V_1) * \frac{V_1}{V_2} \tag{24}$$

The propagation delay of a network shows a variation of 5% to 10% with respect to nominal over an operating supply voltage range of -4.5 to 5.5V.

In addition to dc variations, dynamic power fluctuations have an effect on the propagation delay of CMOS circuits. Supply dI/dt noise due to on-chip circuitry is minimal; however, driver dI/dt noise can have a significant impact on the delay of the driver [19].

With separate power distribution networks, the delay variation is isolated to the driver. Thus, the relative delay variation of a CMOS circuit path due to driver dI/dt noise can be estimated by multiplying the relative delay factor of the driver by the fraction of the nominal delay of the path due to the nominal driver delay.

3.4 Fabrication Process Variation

Fabrication process variation strongly influences the propagation delay of a circuit. For this study, we will characterize process as nominal and corner. Nominal process is the expected process. Corner processes are the limits of acceptable process variation.

Table 2 shows the simulated propagation delay of a chain of 50 inverters for the fabrication corners of a 2 micron MOSIS process[17]. Over these limits, fabrication process variation affects propagation delay by +16% to -19%.

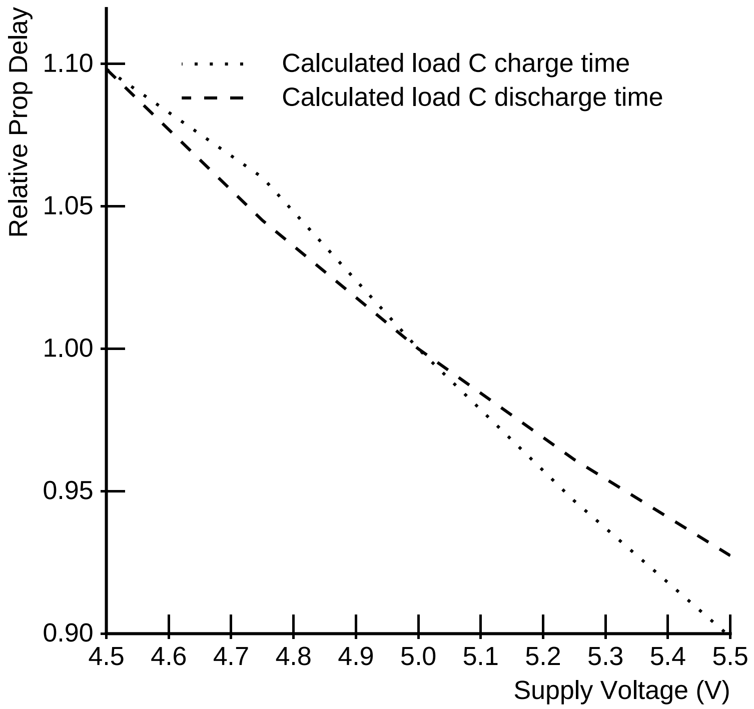


Figure 5: Relative load capacitance charge, discharge delay vs. supply voltage

Figure 9 is a diagram of simulated propagation delay of a chain of 50 inverters for seven MOSIS 0.8 micron fabrication runs. For these runs, the maximum propagation delay is longer than the minimum by a factor of 1.35. When compared to the arithmetic average, the variation is +11% to -18%.

Fan, et. al.[10] performed fabrication process sensitivity analysis on a wave-pipelined adder design. They found delay to be most sensitive to variations in channel oxide thickness, transistor geometry, and device transconductance. Citing Glasser[6], they concluded that the process parameters may be treated as constant within a chip.

We have shown that for static CMOS logic networks temperature variations of from 25 C to 125 C can increase nominal propagation delay by a factor of 1.3 to 1.5, supply voltage variations from 4.5 V to 5.5 V can alter propagation delay by a factor of 0.9 to 1.1, and process can alter nominal delays by a factor of 0.8 to 1.2. We now examine the wave-pipeline performance limits imposed by these variation of propagation delay.

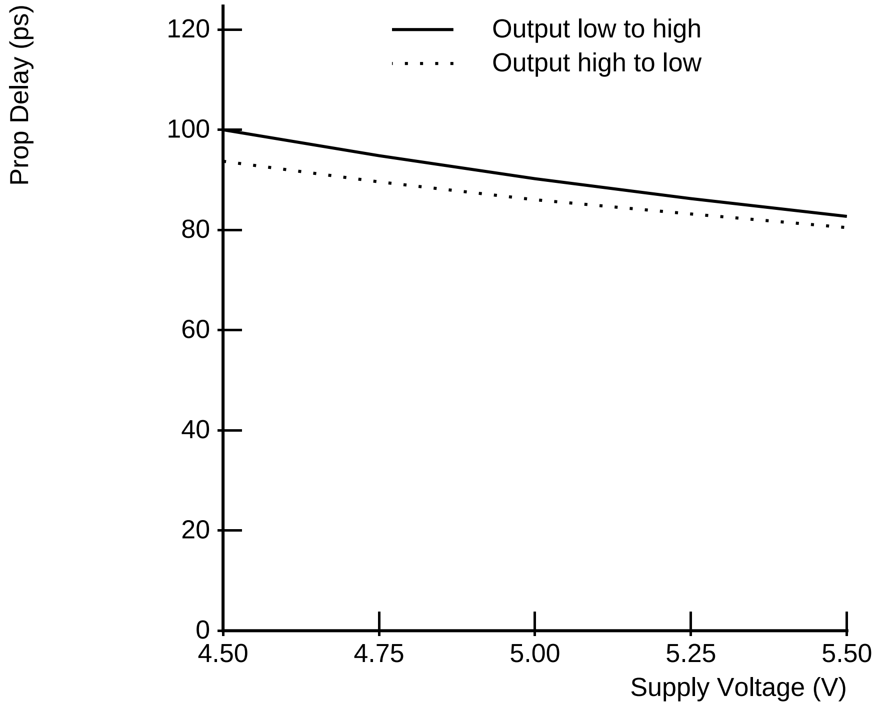


Figure 6: Inverter propagation delay vs. supply voltage

4 Fixed Frequency Clocked Wave-Pipelined Systems

In an fixed frequency clocked synchronous system, a clock with fixed period T_{clk} is supplied to the device. The clock frequency is not a function of chip supply voltage, temperature, or fabrication process. Systems with external clock generation, systems with external clocks which provide timing reference for internally generated clocks, and systems with temperature and supply voltage compensating on-chip VCOs are included in this category. Figure 10 is a block diagram of a synchronous system with an externally supplied clock.

Process	Propagation Delay (ns)
fast	14.6
slow	21.0
typical	18.1

Table 2: Simulated process corner propagation delays

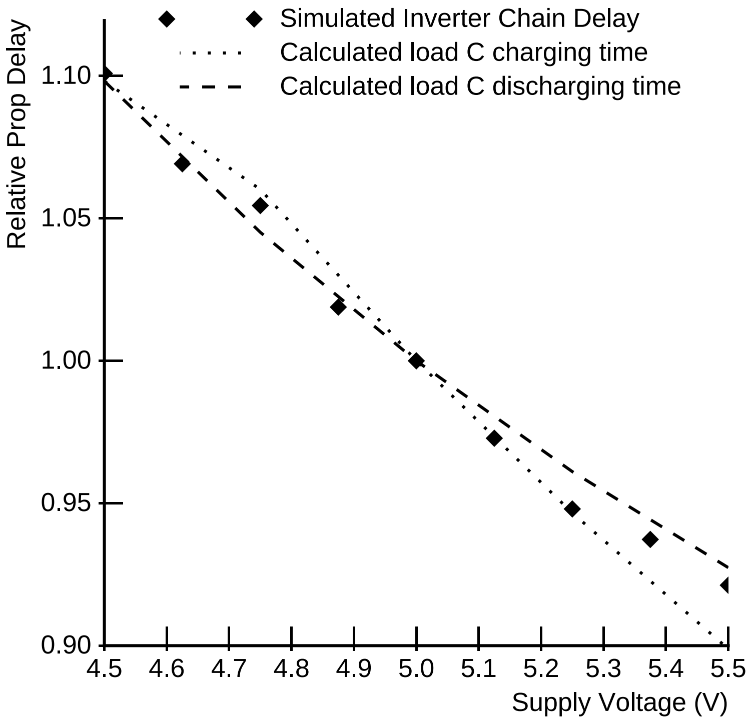


Figure 7: Relative propagation delay vs. supply voltage

For an fixed frequency clocked traditional pipelined system to operate properly, the worst case maximum propagation delay determines the clock rate:

$$T_{max} + RF_{max}/2 + T_s + \Delta C < T_{clk} + \delta C_{io} \tag{25}$$

$T_{max}, RF_{max}, T_s, \Delta C$ and δC_{io} are voltage, temperature, and process dependent. T_{clk} is voltage, temperature, and process independent.

For an fixed frequency clocked wave-pipelined circuit to operate properly, the following two inequalities must hold for edge-triggered registers:

$$\frac{T_{max} + RF_{max}/2 + T_s + \Delta C + T_{synch} - \delta C_{io}}{N} < T_{clk} \tag{26}$$

$$\frac{T_{min} - RF_{min}/2 - T_h - \Delta C + T_{synch} - \delta C_{io}}{N-1} > T_{clk} \tag{27}$$

For flow latches, the following inequalities must hold:

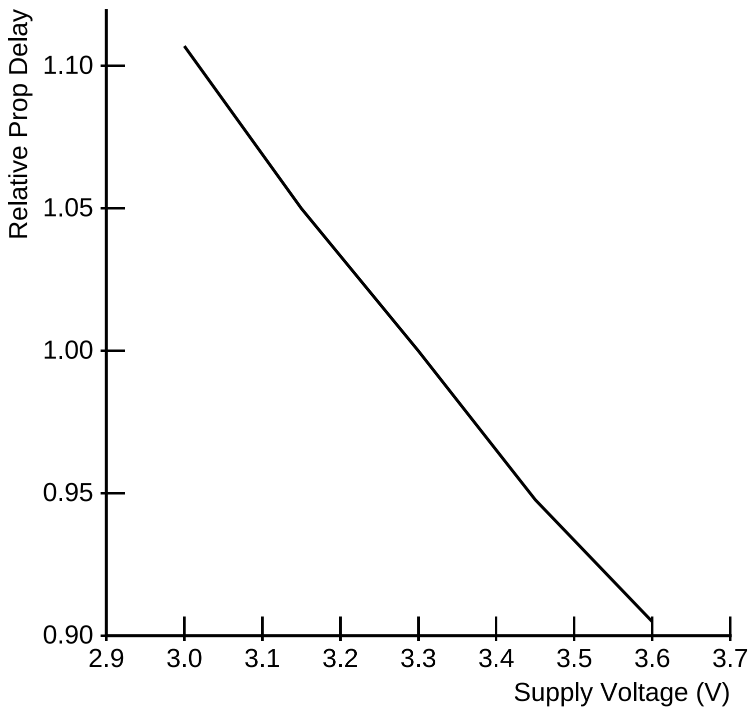


Figure 8: Relative propagation delay vs. supply voltage

$$\frac{T_{max} + RF_{max}/2 + T_s + \Delta C + T_{synch} - \delta C_{io}}{N} < T_{clk} \tag{28}$$

$$\frac{T_{min} - RF_{min}/2 - T_h - \Delta C + T_{synch} - \delta C_{io}}{N-1} > T_{clk} + T_{trans} \tag{29}$$

T_{max}, T_{min}, RF, T_{synch}, T_s and T_h are voltage, temperature, and process dependent. T_{clk} and T_{trans} are voltage, temperature, and process independent.

Deviation of process parameters from their nominal values are relatively time invariant and relatively uniform across an entire die [6]. Thus, once a device is fabricated, its T_{ox}, μ_0, V_{tn}, V_{tp}, etc. can be determined, and an appropriate clock period can be chosen.

Since, however, the particular operating temperature and supply voltage are not known *a priori*, a clock period and an integer number of waves must be specified which satisfy the above inequalities for all valid values of supply voltage and temperature. For the first inequality, the worst condition is minimum supply voltage and maximum temperature. For the second inequality, the worst condition is maximum supply voltage and minimum temperature.

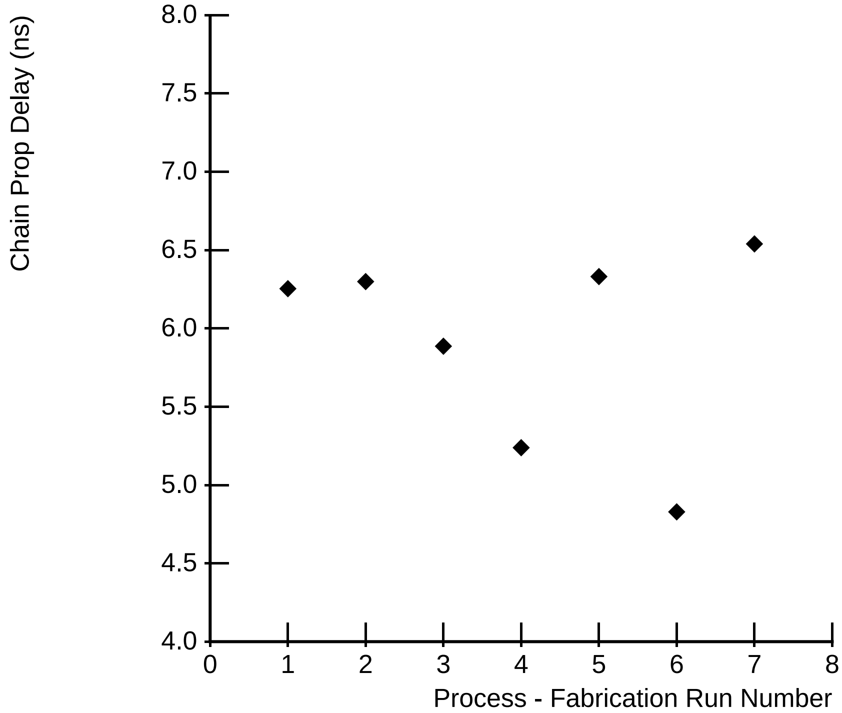


Figure 9: Inverter chain propagation delay vs. fabrication run

The longest path in the network is some factor greater than the shortest path in the network for a given temperature and voltage. This factor, which we define as α, is due to path length differences and data dependent delay variations in the network.

$$T_{max}(V,\tau,P) = \alpha * T_{min}(V,\tau,P), \alpha > 1 \tag{30}$$

Because the relative variation in propagation delay due to temperature and voltage variation is to first order independent of absolute propagation delay, α is a good approximation of the relative path length difference in the network for any temperature and voltage.

The worst-case wave-pipeline timing constraints become:

$$\frac{\alpha T_{min}^{slow} + RF_{max}^{slow}/2 + T_s^{slow} + \Delta C^{slow} + T_{synch}^{slow} - \delta C_{io}^{slow}}{N} < T_{clk} \tag{31}$$

$$\frac{T_{min}^{fast} - RF_{min}^{fast}/2 - T_h^{fast} - \Delta C^{fast} + T_{synch}^{fast} - \delta C_{io}^{fast}}{N-1} > T_{clk} \tag{32}$$

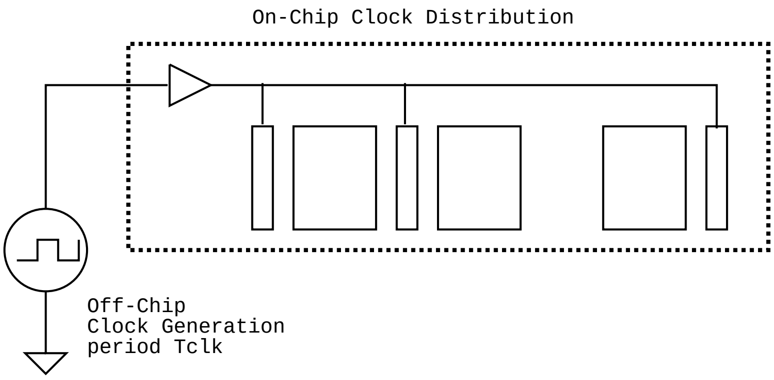


Figure 10: Externally supplied clocked system

where *slow* signifies operating conditions (V_{min}, τ_{max}, P) and *fast* signifies operating conditions (V_{max}, τ_{min}, P).

The propagation delay at worst case operating temperature, supply voltage, and process will be some factor larger than the best case propagation delay. If we define β as:

$$\beta = \frac{T_{min}(V_{min}, \tau_{max}, P)}{T_{min}(V_{max}, \tau_{min}, P)} \tag{33}$$

From section 3 data:

$$\beta \approx (\frac{\tau_2}{\tau_1})^{1.5} * \frac{V_{max}}{V_{min}}) \tag{34}$$

If it is assumed that setup, hold, rise and fall, synchronizer delay, and skew times scale as propagation delay with temperature and voltage, the worst case timing inequalities become:

$$\frac{\alpha\beta T_{min}^{fast} + \beta RF_{max}^{fast}/2 + \beta T_s^{fast} + \beta \Delta C^{fast} - \beta\delta C_{io}^{fast} + \beta T_{synch}^{fast}}{N} < T_{clk} \tag{35}$$

$$\frac{T_{min}^{fast} - RF_{min}^{fast}/2 - T_h^{fast} - \Delta C^{fast} - \delta C_{io}^{fast} + T_{synch}^{fast}}{N-1} > T_{clk} \tag{36}$$

Combining the constraints to solve for N, the number of waves in the wave-pipelined circuit:

$$N < \frac{\alpha\beta T_{min}^{fast} + H_{max} - \beta\delta C_{io}^{fast}}{(\alpha\beta - 1)T_{min}^{fast} + H_{max} + H_{min} - (\beta - 1)\delta C_{io}^{fast}} \tag{37}$$

where,

$$H_{max} = \beta RF_{max}^{fast}/2 + \beta T_s^{fast} + \beta \Delta C^{fast} + \beta T_{synch}^{fast} \tag{38}$$

$$H_{min} = RF_{min}^{fast}/2 + T_h^{fast} + \Delta C^{fast} - T_{synch}^{fast} \tag{39}$$

If $T_{min}^{fast} >> RF, T_s, T_h, \Delta C, T_{synch}$ and the clocks are not intentionally skewed, $\delta C_{io} = 0$, then:

$$N < \frac{\alpha\beta}{\alpha\beta - 1} \tag{40}$$

In a perfectly balanced network $\alpha = 1$, thus:

$$N < \frac{\beta}{\beta - 1} \tag{41}$$

Figure 11 gives the maximum number of waves through a wave-pipelined network versus the environmental delay variation factor, β, for several practical values of the path length variation factor, α.

Table 3 gives the simulated results for the maximum number of waves achievable for the chain of 50 inverters for a range of temperatures and voltages.

Temp Range	**Voltage Range**	α	β	**Max Waves**
25C	5V	1	1	6
25C	4.5-5.5V	1	1.2	4
25-125C	5V	1	1.4	2
25-125C	4.5-5.5V	1	1.7	2

Table 3: Inverter chain simulated maximum number of waves

There are two important implications from equation 40. First, based upon data from section 3.2 and section 3.3 values of β for temperature ranges of 25-125 C and voltage ranges of 4.5-5.5 V for CMOS circuits will be 1.4 to 1.7. Therefore, the number of waves in a static CMOS wave-pipelined logic network, independent of its absolute propagation delay, is three or less. Second, because operating environment changes result in significant changes in propagation delay, extremely accurate path-length balancing may not be necessary to

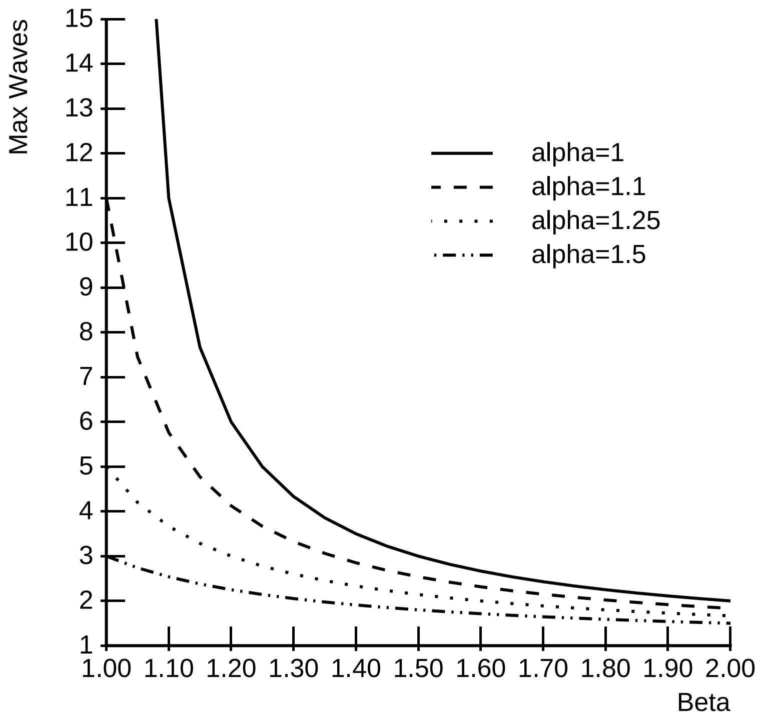


Figure 11: Maximum waves vs. β

achieve the maximum number of waves. For instance, if temperature and supply changes results in a relative propagation delay variation of 60%, i.e. $\beta = 1.6$, the path lengths through the network can differ by as much as 25% for two concurrent waves.

4.1 Environmental Impact Comparison

In this section, we compare the effects of environmental and process variation on traditional pipelines to wave-pipelines with fixed frequency clocking.

For a traditional pipeline, the minimum clock period over all acceptable temperatures and voltages is determined by the maximum propagation delay through the network. Thus:

$$T_{clk}^{min}(\forall V, \forall \tau, P) = \gamma T_{clk}(V_0, \tau_0, P) \tag{42}$$

where,

$$\gamma = \frac{T_{max}(V_{min}, \tau_{max}, P)}{T_{max}(V_0, \tau_0, P)} \tag{43}$$

and,

$$1 \leq \gamma \leq \beta \tag{44}$$

or,

$$\frac{T_{clk}^{min}(\forall V, \forall \tau, P)}{T_{clk}(V_0, \tau_0, P)} = \gamma \tag{45}$$

This factor represents the maximum throughput lost by environmental variation.

For a wave-pipeline,

$$T_{clk}^{min}(V_0, \tau_0, P) = T_{max}(V_0, \tau_0, P) - T_{min}(V_0, \tau_0, P) \tag{46}$$

$$T_{clk}^{min}(\forall V, \forall \tau, P) = \alpha\beta T_{min}(V_{max}, \tau_{min}, P) - T_{min}(V_{max}, \tau_{min}, P) \tag{47}$$

assuming,

$$T_{max}, T_{min} >> \Delta C, T_s, T_h, RF_{min}, RF_{max} \tag{48}$$

Thus,

$$\frac{T_{clk}^{min}(\forall V, \forall \tau, P)}{T_{clk}^{min}(V_0, \tau_0, P)} = \gamma \frac{\alpha\beta - 1}{\alpha\beta - \beta} \tag{49}$$

Figure 12 plots the degradation factors for both traditional and wave pipelines versus γ. It is assumed that for this figure any propagation delay through the network at the nominal environment is approximately equal to the propagation delay at maximum voltage and minimum temperature (i.e. $\gamma \approx \beta$.) Figure 12 is evidence of the need for minimization of environmental fluctuations for wave-pipelined design.

A strategy for maximizing the performance of externally-clocked wave pipelined circuits is tightly controlling the drift of the external power supply and minimizing Vdd and GND noise with numerous supply pins, filter capacitors on the die, and current-limiting I/O drivers. Temperature variation can be minimized by lowering the maximum junction temperature

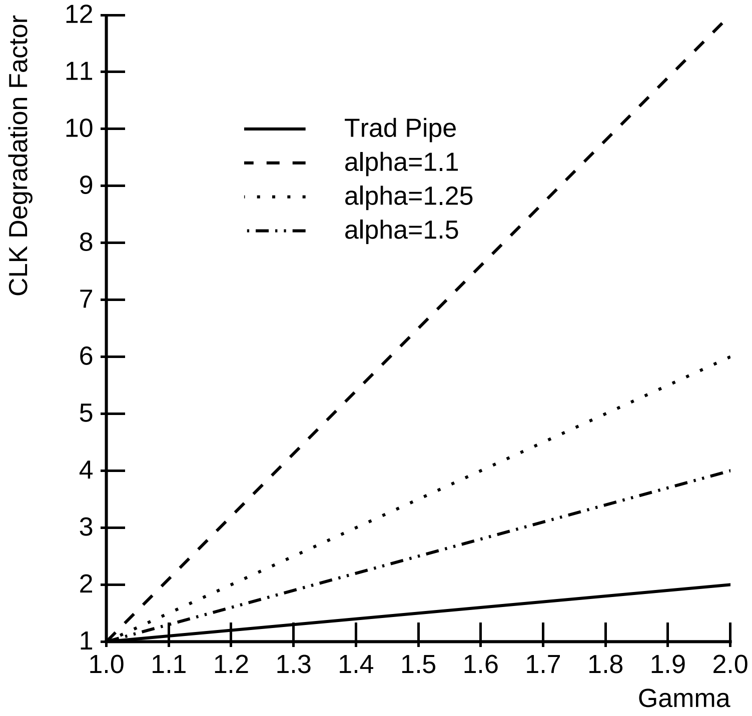


Figure 12: Environmental degradation factor

with low thermal resistivity packaging. Analysis of heat generation and flow could be used in the design process to provide tighter bounds on the expected temporal and spatial propagation delay variation. Lee has suggested integrating thermal analysis in a design environment for improved reliability and performance [14]. Temporal variation can be decreased by raising the minimum operating temperature with "warm-up" cycles.

Without tight controls on temperature and voltage, wave-pipelined fixed-clock circuits are limited to 2-3 waves per stage.

For designs in which full commercial operation is required and tight environmental and process control are not practical, it is unreasonable to expect greater than two waves per wave-pipelined logic block. A useful strategy in this case is to partition the logic into the smallest number of pipeline stages, k, such that inequality 4 with $N = 2$ is satisfied for each section. In this manner, each pipeline stage will be the minimum delay which holds two simultaneous waves. Therefore, the maximum speed-up over a nonpipelined circuit becomes $2 * k$ and the increase in latency will be minimized. Klass[4] analyzes pipelines in which each pipeline stage is in-turn wave-pipelined.

5 Variable Frequency Clocked Systems

In an variable frequency clocked synchronous system, the clock period, $Tclk$, varies so as to match the propagation delay of the logic network.

The clock can be produced by a ring oscillator or voltage controlled ring oscillator. The clock frequency is a function of supply voltage, temperature, or fabrication process. VCOs which compensate for variations in supply voltage and temperature were analyzed with fixed frequency clocked systems. Figure 13 is a block diagram of a synchronous system with an internally generated, variable frequency clock.

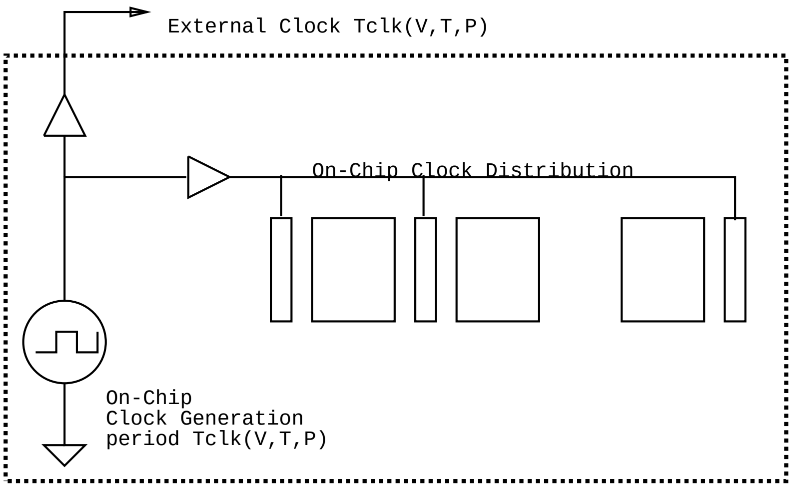


Figure 13: Internally generated variable frequency clocked system

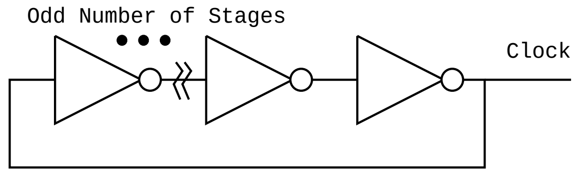
A ring oscillator design and a voltage-controlled ring oscillator design are shown in figure 14.

For a variable frequency clocked traditional pipelined system to operate properly, the worst case maximum propagation delay determines the clock rate:

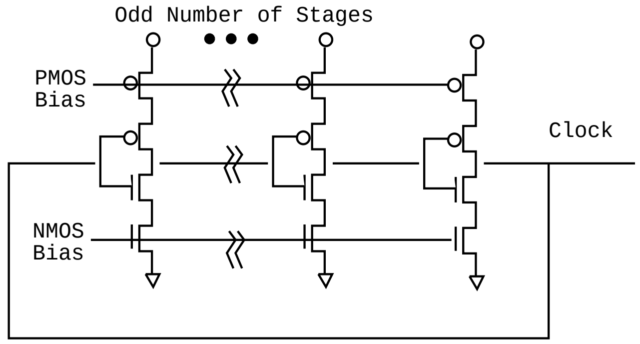
$$T_{max} + RF_{max} + T_s + \Delta C < T_{clk} \tag{50}$$

T_{max}, RF_{max}, T_s, and ΔC are voltage, temperature, and process dependent. T_{clk} is also voltage, temperature, and process dependent.

For a variable frequency clocked wave-pipelined circuit to operate properly, the following two inequalities must hold for edge-triggered registers:



Ring Oscillator



Voltage Controlled Oscillator

Figure 14: Internally generated clocks

$$\frac{T_{max} + RF_{max}/2 + T_s + \Delta C + T_{synch}}{N} < T_{clk} \tag{51}$$

$$\frac{T_{min} - RF_{min}/2 - T_h - \Delta C + T_{synch}}{N-1} > T_{clk} \tag{52}$$

For flow latches, the following inequalities must hold:

$$\frac{T_{max} + RF_{max}/2 + T_s + \Delta C + T_{synch}}{N} < T_{clk} \tag{53}$$

$$\frac{T_{min} - RF_{min}/2 - T_h - \Delta C + T_{synch}}{N-1} > T_{clk} + T_{trans} \tag{54}$$

T_{max}, T_{min}, RF, T_{synch}, T_s and T_h are voltage, temperature, and process dependent. T_{clk} and T_{trans} are also voltage, temperature, and process dependent.

The period of oscillation of a ring oscillator is determined by the propagation delay through the ring. Thus if the temperature, voltage, and process were constant across the device, T_{clk} will vary as the combinational network propagation delay. According to Glasser[6] process parameters can be approximated as constant across a die. Surface temperature profiles of a die tend to be a superposition of a baseline temperature due to average die power dissipation and ambient temperature and hot-spots due to localized device power dissipation[18]. Thus, there is a spatially independent component and a spatially dependent component of temperature variation. For non-uniformly distributed heat sources, the spatially dependent component dominates.

Power supply low frequency voltage variation is also time dependent due to supply drift and spatially dependent due to IR drops across the power distribution network.

Figure 15 compares the variation in propagation delay of a chain of inverters with the variation in clock period for a clock generated by an on-chip ring oscillator. This figure shows that inverter chain propagation delay and the ring oscillator period track if the temperature is spatially uniform.

Figure 16 compares the variation in propagation delay of the inverter chain with variation in period of an on-chip voltage-controlled ring-oscillator for spatially uniform temperature.

Spatial temperature variation depends upon power consumption, device placement, switching behavior, and package design. In the absence of heat flow analysis, worst case spatial temperature variation should be assumed.

With internally generated clocks, the clock frequency is a function of temperature and voltage, and is therefore not time invariant. This may present problems in interfacing a device to other devices in a system.

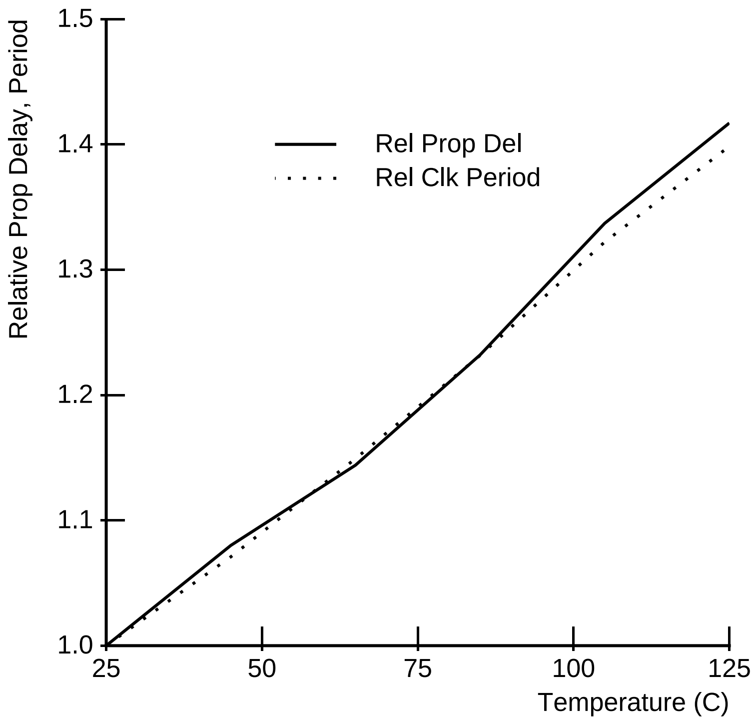


Figure 15: Inverter chain prop. delay and ring-oscillator period vs. temperature

An additional problem for on-chip ring-oscillators is frequency jitter due to noise. Because the clocks used in wave-pipelined circuits are constrained to a range of valid frequencies which becomes increasingly narrow as the number of waves through the logic increases[7], a high degree of clock frequency stability is necessary. This jitter must be included in the ΔC factor in the constraint equations. Low-jitter voltage and current-controlled oscillators minimize jitter through precise capacitance, current, and noise control. Jitter of less than 160 ppm is achievable for on-chip precision CMOS oscillator circuits [15]. They are, however, subject to frequency variation due to supply voltage and temperature changes. Further analysis of the impact of low jitter on-chip oscillators on wave- pipelined designs is warranted.

5.1 Environmental Impact Comparison

In this section, we compare the effects of environmental and process variation on traditional pipelines to wave-pipelines with variable frequency clocking.

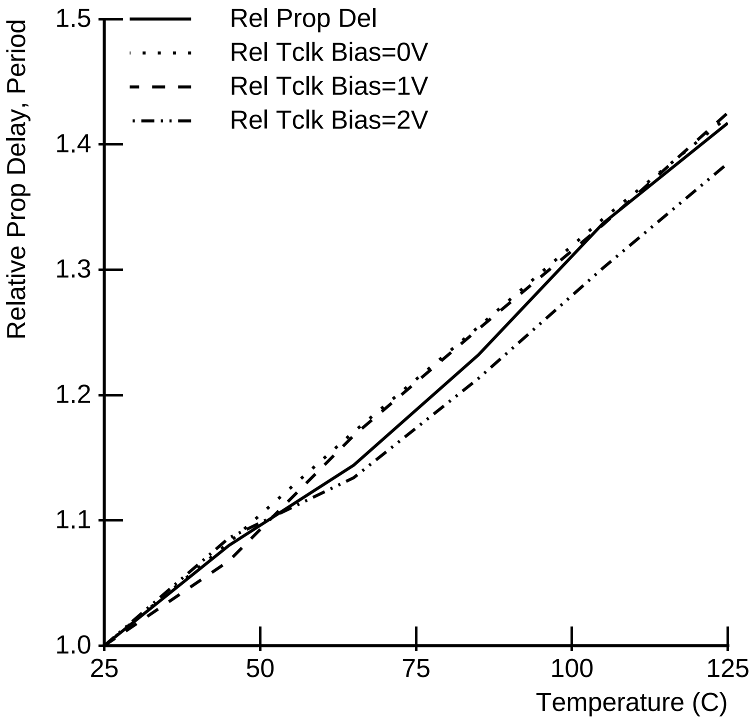


Figure 16: Inverter chain propagation delay and VCO period vs. temperature

For a traditional pipeline, the minimum clock period over all acceptable temperatures and voltages is determined by the worst-cast maximum propagation delay through the network. Thus:

$$T_{clk}^{min}(\forall V, \forall \tau, P) = \gamma T_{clk}(V_0, \tau_0, P) \tag{55}$$

where,

$$\gamma = \frac{T_{max}(V_{min}, \tau_{max}, P)}{T_{max}(V_0, \tau_0, P)} \tag{56}$$

and,

$$1 \leq \gamma \leq \beta \tag{57}$$

or,

$$\frac{T_{clk}^{min}(\forall V, \forall \tau, P)}{T_{clk}^{min}(V_0, \tau_0, P)} = \gamma \tag{58}$$

This factor represents the maximum throughput lost by environmental variation.

For a wave-pipeline,

$$T_{clk}^{min}(V_0, \tau_0, P) = T_{max}(V_0, \tau_0, P) - T_{min}(V_0, \tau_0, P) \tag{59}$$

$$T_{clk}^{min}(\forall V, \forall \tau, P) = \alpha\beta T_{min}(V_{max}, \tau_{min}, P) - \beta T_{min}(V_{max}, \tau_{min}, P) \tag{60}$$

assuming,

$$T_{max}, T_{min} >> \Delta C, T_s, T_h, RF_{min}, RF_{max} \tag{61}$$

Thus,

$$\frac{T_{clk}^{min}(\forall V, \forall \tau, P)}{T_{clk}^{min}(V_0, \tau_0, P)} = \gamma \tag{62}$$

For variable frequency clocking with a uniform surface distribution of supply voltage and temperature, the impact of environmental and process variation affect traditional and wave-pipelines equally. For these circuits, speed-ups like those reported in [3] [10] [8] [12] of 2-10 are achievable.

For non-uniform surface temperature and supply voltage, wave-pipelined circuits with variable-frequency on-chip clocks are subject to the performance constraints of section 4 where β is due to the worst-case spatial variation of environmental conditions.

6 Conclusions

We have shown that temperature, voltage, and process variation can result in much higher performance degradation in wave-pipelined circuits than traditional pipelined equivalents.

We have shown that for fixed-frequency, externally clocked wave-pipelined circuits, the number of waves in the circuit is limited to a maximum of two to three when temperature and supply voltage fluctuations are accounted for.

We suggest several methods of minimizing the performance degradation of wave-pipelined circuits: 1) strict control of operating temperature, voltage, and process; 2) voltage, temperature, and process dependent on-chip clock generation; and 3) partitioning of a wave-pipelined design using traditional pipelining techniques.

The results presented suggest additional research in low-jitter on-chip clock generation, variable clock rate system design, and temperature compensation of logic for CMOS wave-pipelining.

7 Acknowledgements

The authors would like to thank Fabian Klass for review of this work. This work was supported by an ARPA Fellowship in High Performance Computing administered by the Institute for Advanced Computer Studies, University of Maryland and from NSF Contract No. MIP88-22961 using facilities provided by NASA under contract NAG2-842.

References

[1] S. Anderson, J. Earle, R. Goldschmidt, and D. Powers. "The IBM system/360 model 91 floating point execution unit." *IBM Journal of Research and Development,* pp 34-53, January 1967.

[2] F. Klass, and J. Mulder. "CMOS Implementation of Wave Pipelining" Technical Report 1-68340-44(1990)02, Department of Electrical Engineering, Delft University, December 1990.

[3] F. Klass, M. Flynn, and A. J. van de Goor. "Fast Multiplication in VLSI using Wave Pipelining Techniques," To appear in *Journal of VLSI Signal Processing,.*

[4] F. Klass and M. Flynn. "Comparative Studies of Pipelined Circuits," Technical Report CSL-TR-93-579, Stanford University, July 1993.

[5] M. Shoji. *CMOS Digital Circuit Technology* Prentice Hall, 1988. pp. 29-32, 119-125.

[6] L. Glasser and D. Dobberpuhl. *The Design and Analysis of VLSI Circuits.* Addison-Wesley, 1985.

[7] C. T. Gray, W. Liu, and R. K. Cavin III. "Timing Constraints for Wave Pipelined Systems," Technical Report NCSU-VLSI-92-06, North Carolina State University, December 1992.

[8] V. D. Nguyen, W. Liu, C. T. Gray, R. K. Cavin. "A CMOS Signed Multiplier using Wave Pipelining," *IEEE 1993 Custom Integrated Circuits Conference,* 1993.

[9] D. Wong, G. De Micheli, and M. Flynn. "A Bipolar Population Counter Using Wave Pipelining to Achieve 2.5X Normal Clock Frequency" *IEEE International Solid-State Circuits Conference,* San Francisco, CA, February 1992.

[10] D. Fan, C. T. Gray, W. J. Farlow, T. A. Hughes, W. Liu and R. K. Cavin. "A CMOS Parallel Adder Using Wave Pipelining" *MIT Advanced Research in VLSI and Parallel Systems,* March 1992.

[11] W. Elmore. "The Transient Response of Damped Linear Networks with Particular Regard to Wideband Amplifiers" *Journal of Applied Physics,* v. 19, pp55-63, January 1948.

[12] W. Lien and W. Burleson. "Wave-Domino Logic: Timing Analysis and Applications" *MIT Advanced Research in VLSI and Parallel Systems,* March 1992.

[13] A. Sabnis and J. Clemens "Characterization of Electron Mobility in the Inverted <100> Silicon Surface" *1979 IEDM Technical Digest,* 1979.

[14] C. Lee, et al. "Real-time thermal design of integrated circuit devices" *IEEE Transactions on Components, Hybrids and Manufacturing Technology,* v. 11, No. 4, December 1988.

[15] M. P. Flynn and S. Lidholm. "A 1.2-μm CMOS Current-Controlled Oscillator" *IEEE Journal of Solid-State Circuits,* V. 27, No. 7, July 1992.

[16] Meta-Software. *HSPICE User's Manual: Volume 2 Elements and Models,* Meta-Software, Inc. 1992.

[17] The MOSIS Service. MOSIS Parametric Test Results. USC/Information Sciences Institute, University of Southern California, Marina Del Rey, CA. 1993.

[18] K. Hijikata, et. al. "Study on heat transfer from small heating elements in an integrated circuit chip." *Proceedings of the 3rd ASME/JSME Thermal Engineering Joint Conference Part 4,* ASME 1991.

[19] H. B. Bakoglu. *Circuits, Interconnections, and Packaging for VLSI,* Addison-Wesley, 1990.

[20] W. Lam, R. Brayton, and A. Sangiovanni-Vincentelli. "Valid Clocking in Wavepipelined Circuits," *IEEE Conference on Integrated Circuits Computer Aided Design,* 1992.