

Capítulo 2:

Tecnologías submicrométricas

2.1 Informe sobre tecnologías submicrométricas

2.2.1 Efectos del escalado en la tecnología

2.2.2 Efectos en circuitos Switched-capacitors

2.2 Caracterización del entorno de diseño

Tecnologías submicrométricas

El principal objetivo de esta fase de recopilación de información sobre tecnologías submicrométricas era tanto el conocimiento de los posibles problemas que podían surgir (además de posibles soluciones o alternativas), como la valoración de los riesgos en el diseño a abordar. Es decir, determinar el ámbito de aplicación en el que pudieran afectar al diseño los efectos no deseados ante los que nos íbamos a encontrar.

Esta parte resulta fundamental, puesto que aunque la tecnología lleve asociados algunos errores de difícil solución, no todos ellos tienen por qué afectar al diseño concreto a realizar. Es decir, que no se trataba de una mera búsqueda de información útil acerca de las características de la tecnología, sino más bien de una recopilación, clasificación y evaluación de dicha información.

Por tanto, el resultado de dicha labor constituye la base sobre la que tomar una serie de decisiones:

- Elección de una determinada tecnología submicrométrica (130 o 90 nm) para realizar los circuitos deseados.
- Técnicas de diseño más adecuadas para las especificaciones a cumplir (tiempo continuo, tiempo discreto, etcétera)
- Posible modificación de esquemas para un diseño determinado.

El resultado de dicha búsqueda es un informe que recoge las principales características de la tecnología submicrométrica, centrándose sobre todo en los posibles errores que genera el escalado tecnológico y las consecuencias que de ellos se deriven. También es importante tener en cuenta las peculiaridades de los dispositivos reales disponibles en cada tecnología para nuestro diseño. No obstante, esto es un paso posterior y complementario a dicho informe, puesto que exige una valoración a posteriori, basada fundamentalmente en el comportamiento práctico obtenido mediante simulaciones.

Por tanto, se muestran a continuación los resultados de la recopilación descrita, así como las conclusiones que de dicha información pudieron extraerse.

2.1 Informe sobre tecnologías submicrométricas

Como ya se ha comentado anteriormente, las tecnologías submicrométricas se caracterizan básicamente por su longitud de canal inferior (o cercana) al umbral de los 100 nm. Este hecho podría parecer un simple paso más de un proceso de mejora en los dispositivos, pero existen una serie de efectos que hacen que esto no sea así.

Tradicionalmente, la reducción de longitud de canal de los transistores responde a un proceso gradual de escalado en las dimensiones de los dispositivos. Este proceso es lo que se conoce como el escalado tecnológico y está relacionado con la necesidad de un menor consumo de potencia en los dispositivos. Con ello, se puede conseguir mejorar el rendimiento y alcanzar mayores velocidades de procesado.

Este proceso ha ido evolucionando gracias a las mejoras en las técnicas de fabricación, que hasta hace relativamente poco suponían el principal escollo para su progreso. No obstante, con las tecnologías submicrométricas surgen otros aspectos que obligan a replantear la forma tradicional en que se ha ido desarrollando este proceso.

El escalado tecnológico se basa en una reducción por igual de las dimensiones físicas de los transistores, acompañada a su vez de una disminución proporcional en las magnitudes eléctricas que controlan su comportamiento. El problema surge en las tecnologías submicrométricas cuando nos encontramos con un límite práctico al escalado de las magnitudes que controlan el transistor, con lo que se plantea la necesidad de realizar una disminución no proporcionada en las dimensiones físicas del mismo.

Es decir, que con las tecnologías submicrométricas comienza a romperse el equilibrio que permitía la reducción progresiva de los dispositivos sin grandes cambios en su comportamiento. Como consecuencia de esto, aparecen una serie de efectos no deseados que pueden resultar en un modelado incorrecto de los dispositivos.

Entre las causas fundamentales para estos efectos no deseados pueden destacarse las siguientes:

- Imposibilidad de escalar todos los parámetros por igual
- Disminución de la longitud de canal (efectos de canal corto)
- Disminución del ancho del óxido de puerta

Por otra parte, algunos de los principales efectos no deseados que se producen en estas tecnologías se citan a continuación:

- Reducción del rango dinámico de funcionamiento
- Mayor importancia de las variaciones paramétricas
- Incremento de las corrientes de fuga
- Corriente de puerta no despreciable
- Ganancia en corriente limitada
- “Missmatch” por corriente de fuga dominante frente al tradicional
- Mayor dependencia de efectos parásitos
- Incremento del ruido de sustrato

Además de estos efectos genéricos, también es importante el estudio de su influencia en el tipo de circuitos que se utilicen. En este caso, como los diseños considerados estaban pensados para esquemas en tiempo discreto basados en switches, se recopiló información adicional sobre aspectos concretos del comportamiento y peculiaridades de los mismos en estas tecnologías.

A continuación, se analizan de forma detallada los diversos factores a tener en cuenta para el análisis de estas tecnologías.

2.1.1 Efectos del escalado en la tecnología

En este apartado se van a detallar los principales factores que confluyen en las tecnologías submicrométricas y que determinan un comportamiento diferente al tradicionalmente modelado hasta la fecha en anteriores generaciones comerciales de dispositivos.

Uno de los principales factores a tener en cuenta es el **escalado de la tensión de alimentación** (V_{DD}) de los transistores, ya que la tensión umbral V_{TH} no puede ser escalada en igual medida. Con el paso de una a otra generaciones de dispositivos, se ha ido reduciendo la tensión de alimentación de los mismos ganando en un menor consumo y mejores propiedades de disipación de potencia. Esto ha permitido alcanzar velocidades cada vez mayores sin grandes sacrificios durante varias generaciones de transistores de distintas tecnologías.

No obstante, ello es debido a que la reducción de todas las magnitudes y dimensiones que influyen en el funcionamiento de los transistores se ha podido realizar de un modo más o menos compensado: el problema que se plantea en las tecnologías submicrométricas es que el salto cualitativo al disminuir la alimentación de 2.5 V (para una tecnología 180nm) a 1.2 o 1 V (valores de alimentación para 130 y 90 nm, respectivamente) es mucho mayor en comparación con la disminución llevada a cabo en el valor de la tensión umbral de los transistores, V_{TH} .

El hecho de que no pueda producirse un mayor escalado de la tensión umbral V_{TH} se debe fundamentalmente a la corriente de corte I_{off} de los transistores. En los dispositivos de efecto campo, esta intensidad viene determinada por la ecuación siguiente:

$$I_{off} \approx I_{VT} 10^{-V_T/S} \quad (I)$$

Donde I_{VT} es la intensidad para la que se define V_{TH} , S es la variación de subthreshold (típicamente 90 mV/decada). Como S está fijado por una serie de

parámetros tecnológicos, la única variable que nos permite cambiar su valor es la temperatura, pero esto no es posible en muchas aplicaciones. Por tanto, vemos que el escalado de la tensión umbral está limitado para no superar unos ciertos valores de I_{off} que garanticen un funcionamiento correcto de los dispositivos. Las restricciones que nos imponga esta corriente vendrán dadas por características de consumo o de funcionamiento según las distintas aplicaciones, pero constituye una limitación al escalado que de momento no tiene una solución clara. La alternativa más fiable pasa únicamente por hacer una optimización de los valores de alimentación y umbral en los diseños atendiendo a consideraciones de velocidad, rendimiento y consumo.

Se muestra a continuación una tabla donde se recoge la evolución de los parámetros en el escalado de las sucesivas tecnologías, pudiendo observarse este efecto de imposibilidad del escalado ya descrito.

#	Lmin	Vdd	Tox	Vth	AVth
1	3.0	5.0	700	1.5	35 *
2	2.5	5.0	600	1.2	30
3	2.0	5.0	400	1.1	25
4	1.5	5.0	250	1.0	22 *
5	1.2	5.0	250	1.0	21 *
6	1.0	5.0	250	0.95	20
7	0.8	5.0	200	0.85	13
8	0.5	3.3	135	0.73	11
9	0.35	3.3	100	0.59	9.0
10	0.25	2.5	60	0.52	6.0
11	0.18	1.8	50	0.42	4.2
12	0.12	1.2	42	0.32	3.8
13	0.10	1.2	36	0.31	3.2 *
14	0.07	0.9	30	0.30	2.5 *

Tabla 2.1. - Evolución de los parámetros tecnológicos en tecnologías CMOS

Como puede observarse de manera evidente, se produce un “frenazo” en la reducción de los valores de V_{TH} que limita la próxima evolución de tecnologías submicrométricas y obliga a replantear las técnicas de fabricación. Además, también puede apreciarse que el escalado de las tensiones de

alimentación y umbral resulta algo desproporcionado, lo que genera ciertos efectos no deseados.

No obstante, es complicado separar de un modo claro hasta donde influyen unos y otros parámetros en el funcionamiento de los dispositivos. Así, por ejemplo, que la tensión umbral V_{TH} no pueda reducirse tanto como sería deseable genera que el ancho del óxido de puerta en los transistores tenga que ser distinto de lo que recomendarían las geometrías tradicionales de escalado. Para mantener de algún modo ese equilibrio sería necesario hacerlo más ancho, pero esto a su vez penalizaría en velocidad y también chocaría de algún modo con la disminución de la longitud de canal. Ello repercute en que normalmente los efectos no deseados que aparecen no sean consecuencia de una única causa sino de una serie de ellas íntimamente relacionadas.

En definitiva, puede afirmarse que existe una enorme interrelación entre diversos factores, lo que genera a su vez una serie de compromisos entre distintos parámetros tecnológicos, y que además son enormemente difíciles de analizar de forma aislada.

Esto genera, por un lado, que las técnicas tanto de diseño como de modelado utilizadas hasta la fecha ya no sean aplicables con plenas garantías. Esto se traduce en que ni la aplicación de las mismas reglas de escalado nos llevaría a un funcionamiento correcto del dispositivo, ni los modelos tradicionales proporcionan ya una estimación completamente fiable del comportamiento real de los mismos.

Entre los efectos provocados por la disminución de la tensión de alimentación destaca la disminución del rango dinámico. Como hemos visto, con la evolución tecnológica a menores longitudes de canal, el recorte en la alimentación (V_{DD}) es mucho mayor que el que se consigue para la tensión umbral (V_{TH}). Teniendo en cuenta que la zona de funcionamiento de los transistores va a venir determinada por

$$V_{DD} > V_{TH} \rightarrow V_{DD} - V_{TH} > 0 \quad (II)$$

Donde resulta obvio comprobar que, en las condiciones especificadas anteriormente (V_{DD} se reduce mucho más que V_{TH}), el rango se verá recortado sustancialmente. A esto hay que sumar la influencia de las corrientes de fuga sobre el desapareamiento (mismatch) de los dispositivos, que como veremos más adelante de forma más detenida, en estas tecnologías va a resultar dominante sobre los umbrales de ruido.

Otro problema que se deriva de forma evidente de la reducción de la tensión de alimentación es la disminución de los márgenes de ruido para señales digitales. Es obvio que si disponemos de menor amplitud de señal habrá que ubicar los niveles de decisión digitales algo más cercanos de la zona de funcionamiento intermedia en la que las señales no están definidas a nivel lógico. Por tanto, podemos estar expuestos a un mayor peligro de errores en la decisión ante grandes niveles de ruido.

Otro de los factores que tienen una mayor importancia en las tecnologías submicrométricas es la **reducción de la longitud de canal L_{min}** . Este es el parámetro que se utiliza para la descripción de las diversas tecnologías y uno de los que mayor influencia posee por los efectos que genera su escalado.

Es interesante analizar las consecuencias de su reducción, tanto a nivel de funcionamiento como del propio modelado y aproximación al diseño. Aquí se acentúan una serie de problemas que ya se han venido comentando anteriormente acerca de los modelos de comportamiento del transistor.

Tradicionalmente, las ecuaciones que describen el comportamiento del transistor están basadas en las hipótesis de canal largo, lo que permite una simplificación de los cálculos sin gran pérdida de precisión. En tecnologías submicrométricas el canal se ve reducido a dimensiones en las que esta hipótesis ya no es aplicable, lo que genera imprecisiones no tanto a nivel de simulación (puesto que las herramientas no aplican estas hipótesis) sino más bien en cuanto a que se hace necesario un replanteamiento de la estrategia de diseño para una mejor comprensión del dispositivo.

Por otra parte, el hecho de la propia disminución en la longitud genera una serie de efectos no deseados que sí influyen en el comportamiento de forma más apreciable y han de ser incluidos en simulación convenientemente. Estos son los denominados efectos de canal corto, que se manifiestan principalmente en las siguientes consecuencias:

- Incremento de las corrientes de fuga.
- La intensidad de la puerta deja de ser despreciable.
- Disminución de la resistencia de entrada R_{in} .

Veamos a continuación más detenidamente cada uno de estos efectos y las diversas causas que los generan.

En lo que respecta a las *corrientes de fuga*, se conocen así a las intensidades que circulan por el canal del transistor cuando el dispositivo no está funcionando. Son corrientes residuales que se generan por efectos parásitos no deseados, y cuya principal desventaja es que incrementan el consumo de potencia. En condiciones normales y tecnologías anteriores, el consumo de los dispositivos mientras no están conmutando es despreciable frente a la potencia utilizada en las fases activas. Ahora, se produce un incremento en dichas corrientes que puede resultar en consumos significativos de potencia.

Entre las corrientes de fuga se distinguen básicamente dos tipos:

- Intensidad de subthreshold
- Intensidad de puerta

En el caso de la intensidad de subthreshold, se trata de una corriente que va del drenador a la fuente cuando la tensión a la que se somete al transistor esta por debajo del umbral V_{TH} (en la fase no activa de funcionamiento). Al realizarse el escalado de la longitud de canal en mayor proporción que el de V_{TH} , se produce un incremento en esta intensidad con el consiguiente consumo de potencia adicional.

Hay tres fenómenos que influyen en esta componente de subthreshold: la difusión de electrones en inversión débil, el efecto de “Drain-induced Barrier Lowering”(DIBL) y el Band-to-Band Direct Tunneling (BTBT)

El más dominante es la difusión de electrones en inversión débil, que es tradicionalmente el más significativo y está controlado por la tensión de puerta del transistor de forma exponencial.

El efecto de Drain-induced Barrier Lowering se produce como consecuencia de aplicar una tensión elevada en el drenador que permita que las regiones de depleción de drenador y fuente interactúen entre sí. Este efecto desemboca en una bajada de la barrera de potencial existente entre ambas que permite que pasen más electrones e inyecta corriente al canal.

Por último, el efecto Band-to-Band Direct Tunneling se produce como consecuencia de que el drenador o la fuente estén polarizados a una tensión mucho más elevada que la del sustrato. Si esta tensión es lo suficientemente elevada puede que el salto total a través de la unión supere la barrera de potencial de la banda del semiconductor, produciéndose el paso de una corriente parásita. Esta intensidad de fuga generada puede ser tanto de drenador a sustrato como de drenador a fuente. También es una de las principales contribuciones a la corriente de fuga de la puerta al canal a través del óxido aislante.

Estos efectos de canal corto son en general difícilmente solucionables mediante técnicas de diseño y dependen más bien de la fabricación de los dispositivos (niveles de dopado, estructura de las regiones de difusión, etc). Se estima que están bajo control si la anchura del óxido de puerta no baja de los 30 Å. Actualmente es una de las principales limitaciones a que el escalado de dispositivos siga evolucionando.

Contra el consumo estático debido a corrientes de subthreshold existen una serie de técnicas basadas en distintos esquemas de diseño, tales como usar “*stacks*” de transistores o transistores con varios V_{TH} . Estas técnicas se

utilizan sobre todo para circuitos digitales donde la intensidad de fuga puede resultar muy apreciable en caso de que los transistores no actúen durante periodos de tiempo relativamente grandes.

En el primer caso, se trata de sustituir los transistores aislados por un “*stack*” de dos transistores sin modificar la carga de entrada. Con ello se consigue una compensación del efecto subthreshold y unas corrientes de fuga menores debido al efecto de *stacking*. Este efecto se basa en que al apilar dos transistores en cascada, cuando ambos se cortan la tensión en el punto medio entre ambos sea levemente mayor, disminuyendo así el efecto de bajada de barrera de potencial que provoca un incremento en las corrientes de fuga. No obstante, esta técnica presenta un problema. Aunque se sustituyan los transistores para que no afecte a la carga de entrada y se mantenga el mismo retraso hasta la llegada a ellos, al pasar a través sí sufren una penalización en tiempo. Por tanto, esta técnica sólo puede utilizarse en caminos que no sean críticos para el circuito, en los que un retraso adicional no afecte a su funcionamiento pero ayude a disminuir el consumo estático global del mismo.

La segunda técnica consiste en el uso combinado de transistores con una V_{TH} más elevado que el del resto del circuito. Nuevamente aparece el mismo problema que con la anterior técnica, puesto que en ellos hay un retraso mayor que en los de V_{TH} menor. Por tanto, su empleo es absolutamente análogo, situándose en aquellas zonas del circuito que no formen parte del camino crítico de la señal. No obstante, esto no es siempre posible y se puede hacer necesario redefinir el circuito en base a aquellos transistores que requieran un V_{TH} mayor.

Otro efecto parásito fundamental es la aparición de una corriente de puerta que pasa a no ser despreciable. Esto supone una diferencia clave respecto a otras tecnologías y que obliga a replantear los esquemas de modelado utilizados hasta la fecha en numerosos simuladores de diseño analógico. Su influencia es notable a partir de los 90 nm, donde debe ser estimada porque constituye una parte no despreciable del consumo estático de los circuitos, además de provocar que la resistencia de entrada de los

transistores ya no pueda considerarse infinita como en cálculos a lápiz y papel se hacía tradicionalmente.

La causa principal de la aparición de esta corriente de fuga es la reducción del ancho de óxido de puerta del transistor. Debido a ello, el efecto túnel comentado anteriormente permite el paso de una corriente parásita a través del mismo. Existen tres tipos de mecanismos que controlan el efecto túnel entre la puerta y el sustrato o el canal:

- Efecto túnel de electrones de banda de conducción
- Efecto túnel de electrones de banda de valencia
- Efecto túnel de huecos de banda de valencia

Estos mecanismos se distribuyen según el tipo de transistor utilizado, aunque en todos ellos el funcionamiento es similar y se basa en el paso de electrones a través de la barrera de potencial del canal. Así, en transistores tipo n, se produce un paso de electrones de la banda de conducción del canal a la puerta. En transistores tipo p, la corriente de fuga se genera bien por paso de electrones de la banda de valencia de la puerta a la de conducción del canal, o bien por el paso de huecos del canal a la puerta del transistor si la polarización no es muy elevada.

Como consecuencia de estas corrientes de fuga en la puerta se producen una serie de efectos parásitos no deseados. El principal consiste en la modificación de la capacidad de entrada de los transistores, lo cual puede resultar enormemente perjudicial según la aplicación que se quiera realizar.

En tecnologías submicrométricas, aparece una frecuencia límite, f_{gate} , que determina un umbral que define los cambios en la carga de entrada al transistor. Esta frecuencia viene dada por la expresión siguiente:

$$f_{gate} = \frac{g_{tunnel}}{C_{in} 2\pi} \quad (III)$$

Por debajo de dicha frecuencia, la capacidad de entrada se ve modificada por la adición de un término resistivo y la corriente de fuga en la puerta se hace dominante. Por encima de dicha frecuencia lo que ocurre es que se añade un término capacitivo a la entrada. No obstante, este término es mucho menos influyente debido precisamente a esa naturaleza capacitiva, puesto que cuanto más aumentemos la frecuencia se verá atenuado su efecto. El valor típico de f_{gate} para tecnologías de 90 nm es aproximadamente de 1 MHz, siendo aún menor en tecnologías superiores.

Esto, a efectos prácticos, quiere decir que estas tecnologías van a presentar numerosos problemas a bajas frecuencias, mientras que si trabajamos a frecuencias muy elevadas su comportamiento será más parecido al esperado y darán bastantes menos problemas.

En la actualidad, estas corrientes de fuga constituyen uno de los mayores inconvenientes que presentan estas tecnologías, ya no sólo por el incremento apreciable en el consumo estático sino por la modificación que pueden generar sus efectos no deseados en el correcto funcionamiento de los circuitos a diseñar. Se requiere una correcta estimación de sus valores puesto que en un momento determinado podemos encontrarnos con que la realidad no se corresponda con los comportamientos simulados para su correcto diseño. Especialmente, a partir de una longitud de canal por debajo de los 100nm se hace absolutamente necesario contar con estimaciones precisas puesto que la corriente se hace bastante apreciable.

Como consecuencia de esta corriente de fuga en la puerta se producen una serie de efectos parásitos de cierta importancia:

- Ganancia en corriente limitada
- Efectos de auto-descarga y tasa de caída
- Efectos de desapareamiento (mismatch)
- Ruido

En el diseño tradicional, se tendía a usar transistores de la mayor longitud de canal posible para así conseguir una mayor resistencia de salida o una menor influencia del ruido. No obstante, en tecnologías submicrométricas aparece una ganancia de pequeña señal en corriente limitada, de manera que conforme aumenta la longitud del canal disminuye la ganancia en intensidad haciéndose menor incluso de la unidad. Ello implica que sea recomendable para tecnologías por debajo de 100 nm utilizar transistores de la longitud de canal mínima posible en la tecnología.

Para la corriente de DC, usando estimaciones mediante la frecuencia de puerta f_{gate} , pueden obtenerse resultados para representarla en función de la longitud de los transistores. Vemos por ejemplo una gráfica donde se combinan valores experimentales y estimados para dos tecnologías submicrométricas (en este caso 90 y 65 nm) frente a la longitud de canal de los transistores:

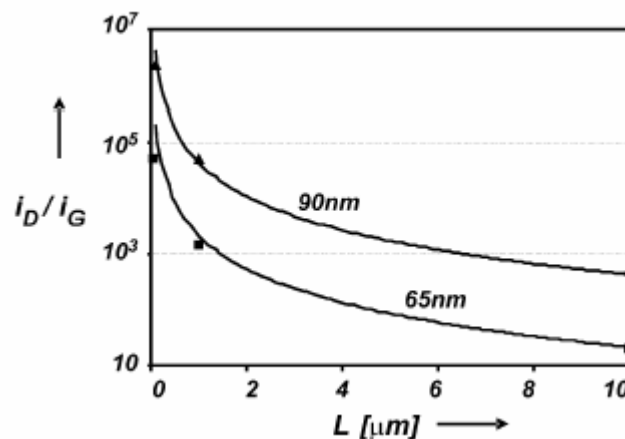


Figura 2.1 – Evolución de la ganancia en corriente en función de L con el escalado

Vemos claramente como se produce una disminución muy importante de la ganancia al incrementar la longitud de los transistores utilizados, y además, que esta particularidad se acentúa cada vez más con el escalado tecnológico a menores tamaños.

Los efectos de auto-descarga se producen especialmente en circuitos donde haya dispositivos CMOS usados para almacenar carga como capacidades, bien sean para mantener una señal o para muestrearla, como en el caso de circuitos que incluyan llaves. El efecto que se produce es que debido

a la corriente de puerta hay unas pérdidas adicionales que hacen que la tensión almacenada tenga una caída en su valor que puede alterar el máximo tiempo de mantenimiento y la frecuencia mínima de funcionamiento.

Aproximadamente, la tasa de caída de carga almacenada en una capacidad CMOS es aproximadamente igual a f_{gate} . En base a esta relación, para un circuito de muestreo y retención, la caída de tensión que se corresponde con un intervalo de tiempo Δt viene dada por la expresión:

$$\Delta t \approx \Delta V / (1V \cdot f_{gate}) \quad (IV)$$

Esto se traduce en que por ejemplo, para una caída de 1mV, el máximo tiempo de mantenimiento para una tecnología de 180 nm estaría en el rango de milisegundos. No obstante, cuando pasamos a tecnologías más recientes como 90 nm e inferiores, para esa misma caída el tiempo de mantenimiento caería al rango del nanosegundo. Esto hace que este tipo de circuitos sean difícilmente realizables con capacidades CMOS para convertidores analógico-digitales de bajo o media tasa de muestreo. En ese caso las capacidades deberían realizarse bien mediante transistores de óxido grueso o con capacidades intermetálicas. Cabe resaltar que en caso de usar transistores de óxido fino, los tipo pMOS son medio orden de magnitud mejores que los nMOS para aplicarlos como capacidades en los circuitos.

Otro problema de vital influencia en estas nuevas tecnologías es el *mismatch* debido a la corriente de fuga de puerta. El *matching* es el nivel de apareamiento que se consigue entre dispositivos idealmente idénticos en el diseño. En teoría, deberían conseguirse dispositivos idénticos utilizando iguales materiales y distribuyendo las distintas capas del layout de un modo perfectamente simétrico. No obstante, en la práctica resulta imposible evitar unas pequeñas desviaciones debidas a procesos de fabricación, tales como desviaciones de temperatura, distintos niveles de dopado, etcétera. Estas variaciones de uno a otro dispositivo es lo que se conoce como *mismatch* o desapareamiento.

El problema de la corriente de fuga en la puerta es que se debe a efecto túnel mecánico-cuántico que depende tanto del ancho de las capas como de la fuerza del campo aplicado. Además, también lleva asociada una dispersión que contribuye a que el nivel de matching puede resultar por debajo de las limitaciones clásicas del mismo, impuestas tradicionalmente por el nivel de ruido. De esta manera, la influencia de las corrientes de fuga de puerta puede llegar a ser más determinante que el ruido en sistemas analógicos, y fijar un límite mínimo para figuras como el offset de los amplificadores o la precisión de los convertidores A/D.

Estos efectos de mismatch producidos por la corriente de fuga se añaden a los del tradicional con una dependencia diferente del área del transistor, de modo que la influencia total viene dada por la expresión siguiente:

$$\frac{\sigma_{ID}^2}{i_D^2} = \left(\frac{\zeta}{\sqrt{WL}} \right)^2 + \left(\frac{\xi \cdot L^2}{\sqrt{WL}} \right)^2 \quad (\text{V})$$

donde el primer término ζ representa la influencia del mismatch clásico y el segundo ξ se corresponde con el mismatch generado a consecuencia de las corrientes de fuga en la puerta. De aquí, vemos que los métodos tradicionales de reducción del mismatch basados en aumentar W y L del transistor (en consecuencia aumentaría el área) ya no serían válidos.

Aunque con ello se seguiría disminuyendo el mismatch tradicional, vemos que existe una dependencia directa con L que aumentaría el mismatch generado por la corriente de fuga en la puerta. Además, como el término en el denominador está dentro de una raíz y el de arriba elevado al cuadrado, el orden del incremento sería casi cuatro veces superior a la disminución. Por tanto, estos métodos de resolución del mismatch no son aplicables, sino que únicamente podemos *aumentar el ancho (W) de los transistores dejando L constante*, lo que aumenta el tanto el área del circuito como su consumo. Además puede venir también fijado por otras especificaciones.

Por tanto, en tecnologías por debajo de los 100 nm se hace necesario imponer una restricción de tamaño sobre los diseños para que esta corriente de fuga en la puerta no nos imponga un límite por debajo del habitualmente fijado por el mismatch clásico. Así, es recomendable que el tamaño total de los diseños no pase de $10^4 \mu\text{m}$ y, en cualquier caso, tratar de usar técnicas de cancelación activa del mismatch o diseños de baja sensibilidad al mismatch.

Otro efecto a tener en consideración es el ruido que puede inducir la corriente de puerta a través del transistor. Por tener una corriente circulando a través de una unión, va a existir una densidad de ruido asociada a ella que viene dada por la expresión:

$$S_q = 2 q I_q \quad (\text{VI})$$

donde q representa las unidades de carga transferidas, que pueden coincidir o no con la carga de un electrón e dependiendo del proceso. Este ruido puede inducirse en los dispositivos y es un factor que debe ser controlado para trabajar en el rango de radiofrecuencia y ser tenido en cuenta como un posible límite para el nivel de ruido en el diseño. La forma de limitarlo pasa únicamente por la disminución en la magnitud de la corriente de puerta.

2.1.2. Efectos en circuitos Switched-Capacitor

Dado que los diseños a realizar para este proyecto se basan en gran parte en sistemas en tiempo discreto, era necesario analizar las peculiaridades concretas que pueden afectar a estos circuitos. Para ello, además de documentación general sobre los efectos del escalado se realizó una búsqueda de información sobre cómo pueden afectar los efectos antes descritos a los componentes básicos de estos sistemas Switched-Capacitors (SC).

En general, el escalado tiene una enorme influencia en el funcionamiento de las llaves de corriente (*switches*) y en consecuencia en los circuitos SC. La reducción del voltaje de alimentación hace que el rango de

señal se ve limitado y que la señal entrada no pueda tener un rango de variación tan elevado como se desearía. No obstante, no es sólo es el problema que se plantea.

Cuando disminuimos el rango de alimentación limitamos la señal de entrada, y eso no sólo provoca una disminución del rango de funcionamiento sino que también se ve afectada la relación señal a ruido (SNR) del dispositivo. En procesados analógicos, para que ésta SNR se maximice debemos conseguir que el rango de señal de entrada en relación a la alimentación sea lo mayor posible. El problema que se plantea con los *switches* es que para tensiones cercanas a los raíles de alimentación su comportamiento es muy bueno, pero cuando las señales de entrada se mueven en los rangos intermedios alejándose de los extremos comienzan a introducirse no linealidades en la resistencia del dispositivo. Estas no linealidades son más acusadas cuanto mayor es el escalado de la alimentación, como puede comprobarse en la gráfica siguiente:

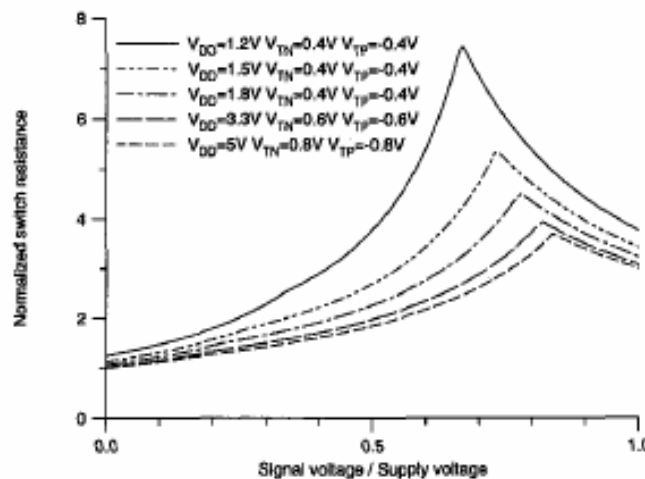


Figura 2.2- No linealidades en switches en función del escalado

Puede observarse como al ir disminuyendo la tensión de alimentación en las llaves, las no linealidades se hacen hasta casi tres veces más acusadas para la zona más desfavorable del rango. Este efecto tiene como consecuencia que cuando los *switches* cargan una capacidad de muestreo a la salida desde

una señal transitoria de entrada, como resultado de estas no linealidades en la resistencia va a introducirse una distorsión es la característica de integración. Y esta distorsión va a provocar un incremento del ruido que puede empeorar la SNR del circuito.

Por tanto, vemos que el uso de capacidades conmutadas en tecnologías submicrométricas va a requerir una serie de cuidados especiales en los diseños. Este efecto por ejemplo se combate principalmente mediante el uso de una técnica denominada Switched OpAmp.

Esta técnica consiste, a grandes rasgos, en la sustitución de capacidades de muestreo que requieran una gran amplitud de señal de entrada por un amplificador al que se le inhabilita la salida durante la fase de reloj correspondiente para así evitar que se introduzca esta distorsión en la fase activa. No obstante, esta técnica presenta el problema de no ser adecuada para altas frecuencias, puesto que el tiempo de recuperación del OpAmp para la fase activa puede acabar retardando el funcionamiento de los esquemas dentro de los que se incluye.

Este problema que se ha descrito se refiere a switches del tipo llaves de transmisión que incluyen transistores tanto tipo n como p. No obstante, dicho efecto no es exclusivo de ellas, sino que también aparece en el caso de interruptores de uno u otro tipo exclusivamente. La resistencia de la fase activa presenta no linealidades fuertemente dependientes de la señal, ya que existe una relación directa entre la transconductancia g_{ds} y las tensiones de entrada y control del interruptor, tanto en el caso de dispositivos n como p. Esta relación viene descrita por la expresión siguiente:

$$g_{ds} = \begin{cases} \mu_n C_{ox} \cdot \left(\frac{W}{L}\right)_n \cdot [V_{GS} - V_{th,n}] & \text{nMOS} \\ \mu_p C_{ox} \cdot \left(\frac{W}{L}\right)_p \cdot [V_{GS} - |V_{th,p}|] & \text{pMOS} \end{cases} \quad \text{(VIII)}$$

Como puede apreciarse, la resistencia en la fase activa va a ir variando en función de la señal de entrada, alterando su comportamiento en distintas

fases del rango y pudiendo generar no linealidades. Además, otro efecto que puede interferir es la dependencia de V_{th} con la entrada a través del efecto cuerpo, aunque éste puede minimizarse siempre que sea posible evitarlo mediante la disposición de los transistores en el esquema. También puede apreciarse que existe una dependencia directa con la tensión de entrada V_{DD} , lo que va a provocar el efecto ya comentado de que las no linealidades sean diferentes para sucesivos escalados tecnológicos de menor longitud de canal en los transistores.

Otro de los problemas que debemos considerar en circuitos SC es el del **ruido**. En general, para sistemas analógicos el ruido con el que vamos a encontrarnos sería tanto el blanco como el ruido flicker ($1/f$). No obstante, a altas frecuencias, que es donde vamos a trabajar, este último es inversamente proporcional a la frecuencia y se hace despreciable. Por tanto, en aquellos sistemas analógicos que usan esquemas SC las principales fuentes de ruido que debemos tener en cuenta son dos:

- ruido blanco (kT/C)
- ruido de entrada al OpAmp

El ruido kT/C es aquel que aparece como consecuencia del plegamiento del ruido blanco a la banda de señal que nos interesa y que es inherente al muestreo. Se trata por tanto de un ruido inevitable que no podemos filtrar sino minimizar en lo posible. LA potencia del ruido asociado al muestreo en una capacidad C responde a la expresión kT/C , donde k es la constante de Boltzmann y T la temperatura absoluta. Por tanto, vemos que el único parámetro libre del que disponemos para preservarnos del ruido es el valor de la capacidad de muestreo C .

De esta expresión (kT/C) podemos deducir que para mantener un valor similar de la relación señal a ruido en tecnologías submicrométricas deberemos de algún modo compensar la variación en la tensión de alimentación debida al escalado tecnológico. Así, si la tensión V_{DD} se ve escalada por un factor α , para

que se mantenga la SNR deberemos escalar la capacidad de forma inversa por un término α^2 .

La otra fuente principal de ruido es el de entrada a los amplificadores. Este efecto aparece como consecuencia de que los OpAmps necesitan consumir potencia estática para alimentar a los dispositivos amplificadores con una transconductancia suficiente para mantener su tensión de entrada por debajo de los niveles de ruido requeridos para la SNR. Para los dispositivos MOSFET, la potencia del ruido a la entrada viene dada (para un ancho de banda determinado B) por la expresión siguiente:

$$(8/3)kTB\gamma/g_m \quad \text{(VII)}$$

donde γ es un factor de degradación para el exceso de ruido atribuido a efectos de calentamiento de electrones. La dependencia de g_m con la intensidad de polarización de los dispositivos es compleja para transistores de canal corto, pero para tensiones de polarización constantes sí puede establecerse una relación más o menos clara. Así, se requiere que la corriente de drenador de los dispositivos se escale por lo menos por el mismo factor por el que se vea escalada la transconductancia.

En términos prácticos, esto quiere decir que cuando la tensión de alimentación V_{DD} se ve escalada por un factor K, la potencia estática del OpAmp deberá ser compensada (en sentido contrario a dicho escalado) por, como mínimo, ese mismo factor K en el valor de la transconductancia g_m . No obstante, en dispositivos cada vez de menor longitud de canal es posible que sea más recomendable hacerlo por un factor de K^2 puesto que son más acusados los efectos de saturación de la movilidad y el factor γ va tendiendo a incrementarse conforme se tiende a longitudes más cortas del canal de los transistores.

Estos valores de ruido pueden constituir el límite para el rango dinámico de los dispositivos, aunque es importante tener en cuenta que no siempre tienen por qué serlo. De hecho, con la disminución de longitud de

canal correspondiente al escalado tecnológico, es muy probable que el futuro límite lo marquen algunos de los efectos descritos anteriormente. En concreto, el mismatch generado por las corrientes de puerta parece en la actualidad uno de los mayores desafíos que afrontan los diseñadores y más claro límite al desarrollo de los nuevos dispositivos.

Otro aspecto fundamental en los circuitos SC analógicos es el error de carga en las capacidades a la salida de los *switches*. Idealmente, siempre vamos a querer cargar un capacidad de muestreo a una tensión lo más exacta posible, pero en la práctica resulta prácticamente inevitable que se produzca un error en la tensión que se pasa a la capacidad de salida. Este error se produce en la conmutación de los *switches* y suele estar provocado principalmente por dos efectos:

- inyección de carga
- *clock feedthrough*

Describamos en primer lugar la *inyección de carga*. Cuando un switch está activo, los transistores están funcionando en la región de triodo y la caída de potencial entre drenador y fuente es aproximadamente cero. No obstante, en realidad existe una carga en el canal por debajo de la puerta que viene dada por la expresión siguiente:

$$Q_{ch} = -WLC_{ox}(V_{GS}-V_T) \quad \text{(IX)}$$

Esta carga en principio no tiene influencia en el funcionamiento del interruptor, pero sí afecta a la tensión que se muestrea en capacidad a la salida. Así, se produce una inyección de carga del canal a dicha capacidad, que se ve afectada por un error que obedece a la expresión siguiente:

$$\Delta V = \frac{kQ_{ck}}{C_k} = -\frac{k(WL)C_{ox}(V_{GS}-V_T)}{C_k} \quad \text{(X)}$$

donde k representa la fracción de carga que se inyecta en la capacidad de salida C_h . El parámetro k es fundamentalmente tecnológico y depende de

numerosos factores, tales como la impedancia vista por cada nodo a tierra o el tiempo de transición del reloj, aunque también de las tensiones de drenador y fuente. La dependencia del error de la tensión de entrada se refleja a través de V_T y V_{GS} , pudiendo ser además causa de distorsión.

También es importante destacar que este error afecta a transistores tipo n y tipo p de forma complementaria, puesto que el comportamiento descrito por la ecuación se corresponde con *switches* tipo n, mientras que en los tipo p la polaridad es la contraria. Es decir, que los transistores tipo n tienden a cargar algo menos de lo ideal, y los p a aportar una tensión levemente superior a la deseada.

El otro mecanismo principal causante de errores en la tensión muestreada en las capacidades de salida de los *switches* es el denominado *clock feedthrough*. Este efecto se produce debido a las capacidades parásitas entre la puerta y el drenador o fuente, que permite que haya una vía de comunicación directa de carga entre la señal de reloj y la tensión que estamos muestreando de la entrada. El error que se introduce debido a este efecto viene expresado por la fórmula:

$$\Delta V_H = -\frac{(V_{DD} - V_{SS})C_{para}}{C_{para} + C_h} \quad (XI)$$

donde C_{para} es la capacidad parásita y V_{DD} , V_{SS} los valores máximo y mínimo de la señal de reloj.

Si sumamos ambos efectos, el error total que se produce en los *switches* analógicos usados en circuitos SC viene dado por la expresión siguiente:

$$\Delta V_{out} = \begin{cases} -k \frac{(WL)_n C_{ox}}{C_H} (V_{DD} - V_{th,n} - V_{in}) - \frac{C_{gd,n}}{C_H + C_{gd,n}} V_{DD} & \text{nMOS} \\ k \frac{(WL)_p C_{ox}}{C_H} (V_{in} - |V_{th,p}|) + \frac{C_{gd,p}}{C_H + C_{gd,p}} V_{DD} & \text{pMOS} \end{cases} \quad (XII)$$

En esta ecuación podemos ver como la señal de entrada tiene diferente influencia en el error que se introduce dependiendo de si estamos usando switches tipo n o tipo p. Así, en los primeros el error disminuye conforme aumenta la señal de entrada, mientras que en los segundos la influencia es la contraria, tendiendo a ser mayor el error que se introduce cuando la entrada disminuye.

Para el caso en que se usen como switches puertas de transmisión con un factor de escalado m para tener igual movilidad en los transistores de tipo n que en los de tipo p ($W_p = m \cdot W_n$), el error total que se produce viene dado por la expresión:

$$\Delta V_{out} = k \frac{(WL)_n C_{in}}{C_H} \left\{ (1+m)V_{in} + V_{in,s} - m|V_{in,p}| - V_{DD} \right\} + C_{gd,s} V_{DD} \left\{ \frac{m}{C_H + mC_{gd,s}} - \frac{1}{C_H + C_{gd,s}} \right\} \quad \text{(XIII)}$$

Vemos que el efecto de inyección de carga varía con la tensión de entrada, mientras que el término de *clock feedthrough* permanece constante. Una de las técnicas más usadas para disminuir la inyección de carga consiste en colocar un transistor dummy suponiendo $k=0.5$, con lo que la regla de diseño a seguir sería hacer dicho dummy la mitad de ancho que el switch principal.

Estos efectos pueden ser minimizados si conseguimos mantener el transistor en inversión fuerte, con lo que se puede utilizar el efecto de pinch-off del canal para aislar el drenador y que sólo se produzca un de inyección de carga en la fuente.

Una mejor solución a estos efectos puede ser conseguida mediante el uso de técnicas de cancelación tales como topologías completamente diferenciales o un esquema de fases adecuado.

Una de las soluciones más extendidas se basa en conseguir que las fases de reloj que controlan el funcionamiento del *switch* no solapen de forma

totalmente simultánea, sino que exista un pequeño desfase entre ellas. Esto permite eliminar los errores de carga en los *switches* de modo que no se almacene en las capacidades en la siguiente fase. De esta manera se realiza una especie de puesta a cero o reseteo capacitivo de los *switches*.

Analizando las ecuaciones que rigen el comportamiento de los efectos mencionados vemos que la dependencia de los mismos con el escalado tecnológico afecta en dos direcciones. Por un lado, el hecho de disminuir la tensión de entrada contribuye a disminuir el error. Pero por otro, vemos que existen términos capacitivos en el denominador que son los que realmente van a determinar el error.

Está claro que conviene hacer que las capacidades de muestreo sean lo mayores posibles dentro del rango más adecuado para un buen funcionamiento a las frecuencias deseadas para los diseños. El problema es que cuánto mayores sean las capacidades, mayores serán los tiempos de carga y deberá llegarse a un compromiso que deberá ser analizado mediante simulaciones y datos prácticos más que de una forma teórica a priori.

Por otra parte, el hecho de que las corrientes de fuga sean cada vez mayores influirá de un modo clave en que a mayor escalado la posibilidad de descargarse los condensadores sea mayor. No obstante, esto dependerá también de las frecuencias a las que trabajemos, y deberá ser analizado a posteriori para ver si realmente afecta o puede ser un efecto despreciable.

2.2 Caracterización del entorno de diseño

Como se ha comentado anteriormente, para la realización de este proyecto se ha contado con un entorno de diseño novedoso diferente del utilizado hasta la fecha con anteriores tecnologías. Se ha dispuesto de un nuevo kit tecnológico con diversas librerías y modelos de simulación adecuados a tecnologías submicrométricas, concretamente para dos de ellas clasificadas según su longitud de canal L_{\min} :

- tecnología de 130 nanómetros [nm]
- tecnología de 90 nm

Una vez realizada la recopilación de información detallada en el anterior apartado, el siguiente paso a seguir consistía en un análisis de los dispositivos disponibles en ambas tecnologías y de sus prestaciones mediante el montaje de una serie de diseños concretos. Así, se toma una decisión sobre en qué tecnología puede resultar más ventajoso trabajar para la realización de los diseños objetivo de este proyecto.

El conjunto de las librerías disponible en ambas tecnologías es bastante similar en cuanto a la variedad de componentes, de manera que inicialmente no se considera como un criterio clave para la decisión a tomar. Tanto en una como en otra tecnología, disponemos de diferentes alternativas de diseño para enfrentarnos a los efectos no deseados ya descritos. Así, existen por ejemplo transistores *Low Leakage* (diseñados para minimizar las pérdidas por corrientes de fuga en la puerta) o *High Speed* (optimizados para funcionar a una mayor velocidad que los anteriores). También podemos disponer de transistores con diferente ancho de óxido para usar una mayor tensión umbral V_{TH} que ayude a minimizar efectos como los provocados por las corrientes de fuga subthreshold que redunden en un mayor consumo estático de los circuitos.

La diferencia más notoria entre ambas tecnologías estriba en el valor de la tensión de alimentación V_{DD} . Para el caso de 130 nm disponemos de

dispositivos alimentados a una tensión nominal de 1.2 V, mientras para el caso de la tecnología de 90 nm este valor se reduce a 1V. Aunque inicialmente no parezca ser un factor determinante, es necesario ver como se traduce esta diferencia en el comportamiento de los circuitos.

Una vez analizado el comportamiento y características de los dispositivos en ambas tecnologías, podemos decir que los criterios seguidos para la elección entre una y otra para el diseño fueron básicamente tres:

- Influencia (a priori) de los efectos del escalado
- Nivel de disponibilidad de herramientas de diseño
- Prestaciones obtenidas en circuitos implementados

En primer lugar, se evalúan las posibilidades que ofrece cada una de las tecnologías y sus características nominales recogidas en la información del fabricante. Inicialmente, sabemos que la reducción de la tensión de alimentación de una a otra va a tener una influencia en los efectos parásitos comentados anteriormente. Sin embargo, no se puede cuantificar de forma exacta si esa variación en V_{DD} resultará determinante o no en el funcionamiento de los diseños a realizar. Sólo podemos establecer un análisis cualitativo acerca del comportamiento que se puede encontrar en ambos casos.

Partiendo del análisis de las características de estas tecnologías descritas en el punto 2.1.1 es fácil de intuir que se va a producir un aumento de la influencia de los efectos no deseados al reducir la tensión de alimentación. Es decir, que tendremos que enfrentarnos a unos efectos parásitos derivados del escalado más acusados en 90 nm que en 130 nm, pero desconocemos si la cuantía de estos efectos será dramática o insignificante. Y al margen de esto, tampoco podemos evaluar de forma rigurosa si las prestaciones en velocidad y/o consumo que obtendremos a cambio del uso de la tecnología de 90 nm pueden llegar a compensar estos efectos no deseados.

Por tanto, de este primer análisis se deriva la necesidad de realizar una serie de pruebas con algunos esquemas a implementar para evaluar sus

prestaciones y tomar la decisión en base a criterios más sólidos. Con esta finalidad, se comienza a realizar simulaciones de algunos esquemas en ambas tecnologías. En concreto, se decide ir realizando diversas pruebas de los dispositivos en una y otra tecnología, y finalmente implementar los esquemas correspondientes a la primera versión del encoder del convertidor pipeline que se describe en el capítulo 3.

No obstante, en el proceso de prueba y simulación de dispositivos apareció un factor que inicialmente no se había tenido en cuenta, o al menos no de una forma decisiva. Como se ha comentado tanto en la introducción como a lo largo de este capítulo, las consecuencias del escalado no sólo afectan al comportamiento de los dispositivos sino que también condicionan el modelado de los mismos mediante las herramientas de diseño. Este punto no había sido considerado como decisivo en principio, puesto que era de esperar que el kit suministrado para estas tecnologías dispusiera de todos los modelos necesarios para un análisis fiable de los dispositivos.

No obstante, en las pruebas realizadas para verificar la correcta simulación de componentes básicos (transistores, puertas lógicas, celdas elementales, etcétera.) se descubrió un detalle de cierta relevancia. Los modelos utilizados para las simulaciones de los transistores no incluyen la intensidad de fuga en la puerta, que se considera a todos los efectos totalmente nula. Esta suposición en otras tecnologías no resulta en absoluto un problema, pero una vez vistos los efectos que se producen en las tecnologías micrométricas (punto 2.1.1) es de suponer que esto puede generar un determinado riesgo no esperado en los diseños. Esto se fundamenta en que la intensidad de puerta tiende cada vez más a ser no despreciable y a provocar efectos no deseados en los dispositivos.

No obstante, el mayor problema que esto genera es que al no estar contemplado este efecto, no ya es que nos enfrentemos a nuevas dificultades que resolver sino que podemos encontrarnos con que existan efectos que modifiquen el funcionamiento pero que están enmascarados por los modelos utilizados. En última instancia y en el peor caso posible, podemos encontrarnos

con un circuito que funcione a la perfección según las herramientas de diseño y simulación, pero que en la práctica, al ser llevado a circuito impreso, su comportamiento no se corresponda con el que se haya diseñado y verificado previamente.

El montaje del encoder se detalla ampliamente en el capítulo 3 y por ello no nos detendremos en explicarlo aquí. La elección de este esquema para estas pruebas responde a que en las simulaciones en tecnologías superiores (250 nm) se llegó a la conclusión de que iba a ser el montaje más crítico de los diseños a realizar, y por tanto aquel en el que mayor aportación podía suponer el uso de una tecnología más avanzada. El diseño del circuito en ambas tecnologías se realizó con los mismos tipos de dispositivos, y tras numerosas simulaciones se determinó el límite máximo de funcionamiento al que podría llegarse con ambas tecnologías para ese montaje.

Sin entrar en muchos detalles, puesto que se especificarán más adelante tanto su funcionamiento como su diseño, simplemente mencionamos el resultado de las simulaciones para ambas tecnologías, que nos permitió hacernos una idea de las prestaciones ofrecidas para los distintos diseños a realizar. Las especificaciones a cumplir por el encoder eran un comportamiento transitorio correcto en el tiempo y en frecuencia unas características de aproximadamente 3 dB de ganancia y 3 GHz de ancho de banda. Esto se consiguió en ambas tecnologías sin mayores dificultades que el dimensionamiento oportuno de los componentes del esquema empleado.

Sin embargo, la verdadera diferencia estriba en la velocidad de funcionamiento que podemos alcanzar. En teoría, el objetivo era conseguir que pudiéramos tener el reloj del circuito funcionando a 1 GHz. En ambos casos se llega a esta frecuencia sin problemas, pero se constata que los efectos no deseados del escalado afectan más de lo previsto en este diseño.

Así, para el montaje del encoder en tecnología de 130 nm, se consigue funcionar a una máxima frecuencia de 1.2 GHz usando dispositivos de bajas pérdidas (*Low Leakage*) para alimentación a 1.2 V, mientras que para 90 nm a

una alimentación de 1 V tan sólo conseguimos llegar a esos mismos 1.2 GHz. Es decir, que el cambio de tecnología no presenta una mejoría sustancial que permita de algún modo compensar los riesgos que se asumen al elegirla.

En consecuencia, y en vista tanto de estos resultados como de las circunstancias antes comentadas sobre efectos de escalado y modelos de simulación, se optó por la decisión de implementar el resto de diseños del proyecto tan sólo en la tecnología de 130 nm. De esta manera, se considera que los efectos nocivos que aparezcan serán menores y que además su influencia estará reflejada más fielmente en el verdadero comportamiento del circuito.