

Diseño Digital Avanzado

Unidad 3 - Mapeo de Arq. De cadas

Dr. Ariel L. Pola apola@fundacionfulgor.org.ar September 25, 2021

Tabla de Contenidos

- 1. Contenidos Temáticos
- 2. Sistemas Discretos de Tiempo Real
- 3. Sistema de Hardware Digital Síncrono
- 4. Redes de Procesamiento tipo Kahn
- 5. Métodos para Representar Sistemas DSP
- 6. Medidas de Rendimiento
- 7. Arquitecturas Dedicadas





Presentación del Curso

Contenidos Temáticos

Unidad 3 Mapeo de Arquitecturas Dedicadas

- Sistemas discretos de tiempo real.
- Sistemas síncronos.
- Redes de procesamiento tipo Kahn para modelar aplicaciones de streaming.
- Métodos para representar Sistemas DSP.
- Diagramas de bloque.
- Gráficos de flujo de señal.
- Diagramas de flujo de datos.

- Single, multi-rate y homogeneous SD-FGs.
- Gráficos de control de flujo.
- Maquinas de estado finitas.
- Medidas de rendimiento: Periodo de iteración, periodo de muestro y velocidad de transmisión, latencia, disipación de potencia.
- Arquitecturas dedicadas.





Sistemas Discretos de Tiempo Real

Concepto

- Un sistema de tiempo real discreto está limitado por la velocidad de muestreo de la señal de entrada y la cantidad de muestras necesarias para producir muestras de salida a una velocidad especifica.
- En un receptor de comunicación digital, la señal de entrada en tiempo real puede ser voz modulada, datos o vídeo y la salida es la señal demodulada respectiva.
- La señal analógica se convierte en una señal de tiempo discreto utilizando un conversor analógico a digital (Analog-to-Digital Converter - ADC).
- En muchos diseños esta señal discreta en tiempo real se procesa en bloques de tamaño fijo.
- El tiempo que se tarda en adquirir un bloque de datos y el tiempo necesario para procesar este representan una dura restricción en el diseño.
- El diseño debe ser lo suficientemente rápido para completar su procesamiento antes de que el próximo bloque de datos esté listo para su turno de procesamiento.



Sistemas Discretos de Tiempo Real

Concepto

- El tamaño del bloque también es importante en muchas aplicaciones, ya que provoca un retraso inherente.
 - Un bloque grande aumenta los requerimientos de retardo y memoria.
 - Un bloque más pequeño aumenta la sobrecarga de procesamiento.
- En muchas aplicaciones el tamaño mínimo del bloque está restringido por el algoritmo seleccionado.
- En un transmisor de comunicación, una señal en tiempo real es digitalizada y luego procesada por GPPs, ASICs o FPGAs, o cualquier combinación de estos.
- Sistemas de procesamiento
 - Tasa Unica (Single-Rate): El número de muestras por segundo a la entrada y salida del sistema es el mismo, y el número de muestras por segundo no cambia cuando las muestras se mueven de un bloque a otro para su procesamiento.
 - Multi-Tasa (Multi-Rate): Los datos se procesan a diferentes velocidades en bloques diferentes, en donde para cada bloque se especifica el número de muestras por segundo y dependiendo de si el sistema es un transmisor o un receptor, el número de muestras por segundo puede aumentar o disminuir respectivamente para el procesamiento posterior.





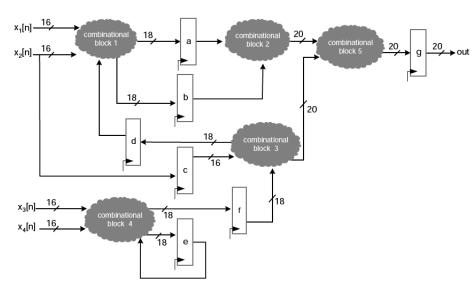
Sistema de Hardware Digital Síncrono

Camino Crítico

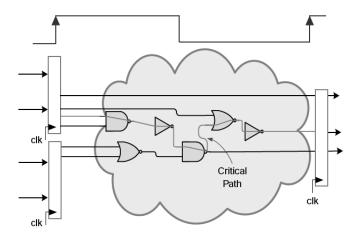
- En muchos casos, las aplicaciones de procesamiento de señales se mapean en lógica digital síncrona.
- Síncrono significa que todos los cambios en la lógica son controlados por un reloj de circuito.
- La lógica digital síncrona suele estar diseñada en el nivel de transferencia de registros (Register Transfer Level - RTL).
- En este nivel el diseño consiste en nubes combinacionales que ejecutan cómputos en el algoritmo y registros síncronos que almacenan valores intermedios y realizan retrasos algorítmicos de la aplicación.
- La nube combinatoria consiste en compuertas lógicas, donde la entrada a la nube son señales discretas almacenadas en registros.
- Al pasar por la lógica combinatoria estas señales experimentan diferentes retrasos en sus respectivos caminos, en donde el camino más largo o lento define el camino crítico (critical path).
- El camino crítico limita la frecuencia de reloj más rápida posible de un diseño.



Sistema de Hardware Digital Síncrono Ejemplo



Sistema de Hardware Digital Síncrono Camino Crítico







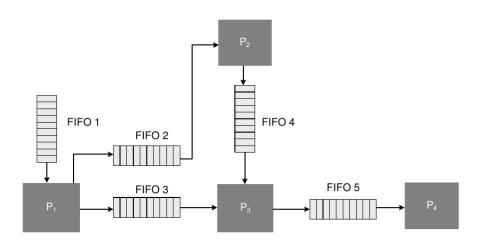
Redes de Procesamiento tipo Kahn

Qué es una red KPN?

- Un sistema que implementa una aplicación de flujo continuo es mejor representarla como componentes que funcionan de forma autónoma tomando entradas de FIFOs y generando salida en FIFOs.
- Las redes de procesamiento tipo Khan (Khan Process Network KPN) proporcionan un método formal para estudiar como se comporta un sistema que considera sus entradas y salidas interconectadas a FIFOs y su posterior mapeo en el diseño digital.
- El KPN es un conjunto de procesos autónomos que se ejecutan simultáneamente y que se comunican entre sí de una manera punto a punto sobre FIFOs, donde la sincronización en la red se logra mediante una operación de lectura de bloqueo y todas las escrituras a las FIFO son no bloqueantes.
- Un proceso espera en un modo de lectura de bloqueo para las FIFOs en cada uno de sus enlaces entrantes para obtener un número predefinido de muestras.
- Todos los nodos de la red se ejecutan después de que sus FIFOs de entrada asociadas hayan adquirido suficientes datos, donde dicha ejecución de un nodo se llama disparo (firing), y las muestras se llaman tokens.



Redes de Procesamiento tipo Kahn Ejemplo





Redes de Procesamiento tipo Kahn

Limitaciones

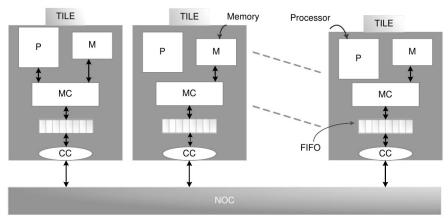
- La lectura de datos requiere estricta adhesión a FIFO, que obliga a las lecturas a seguir un orden secuencial desde el primer valor escrito en el buffer hasta el último. Varios algoritmos de procesamiento de señal no siguen esta secuenciación estricta, un ejemplo es un algoritmo FFT de decimación en el tiempo que lee datos en orden de direccionamiento bit-reverse.
- Asume que una vez que se lee un valor de la FIFO, se elimina. En muchos algoritmos de procesamiento de señales, los datos se utilizan varias veces. Por ejemplo, un algoritmo de convolución simple requiere múltiples iteraciones del algoritmo para leer los mismos datos una y otra vez.
- Supone que todos los valores serán leídos, mientras que en muchos algoritmos puede haber algunos valores que no requieren ninguna lectura y los datos se leen escasamente.



Redes de Procesamiento tipo Kahn

Modificaciones

Una forma de solucionar las limitaciones anteriores es usar la memoria local M en el nodo del procesador D para mantener una copia de sus datos FIFO de entrada.







Descripciones ejecutables basadas en lenguaje

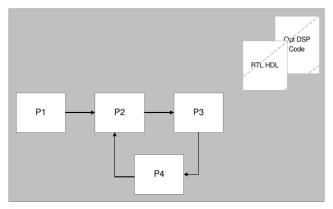
- Los métodos basados en el lenguaje se utilizan para el desarrollo de software, y se utilizan lenguajes de alto nivel para codificar algoritmos.
- Los idiomas son interpretativos o ejecutables (Ejemplo Matlab).
- Para algoritmos computacionalmente intensivos, el diseñador prefiere escribir el código en C/C++.
- Como en estos idiomas el código se compila para la ejecución, el ejecutable funciona mucho más rápidamente que su simulación equivalente de MATLAB.



Especificación dirigida por gráficos o diagrama de flujo

- Los métodos gráficos son especialmente convenientes para el mapeo de HW y la comprensión del funcionamiento y el flujo del algoritmo.
- Una representación gráfica es el método de elección para desarrollar hardware optimizado, generación de código y síntesis.
- Algunas herramientas que usan esta metodología son Simulink de Mathworks, Advanced Design Systems (ADS) de Agilent, Signal Processing Worksystem (SPW) de Cadence, Cocentric System Studio de Synopsys, LabVIEW de National Instruments y DSP Station de Mentor Graphics.
- También soportan el flujo de diseño jerárquico.
- Estos métodos también hacen hincapié en el diseño de arquitectura basada en componentes.
- Los componentes pueden estar parametrizados para ser reutilizados en una serie de instancias de diseño.
- Cada componente puede describirse adicionalmente en diferentes niveles de abstracción.



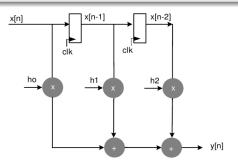


Representación gráfica de un algoritmo de procesamiento de señal donde el nodo P3 se describe en RTL y es optimizado para FPGA o un DSP particular.



Diagrama en Bloque

- Un diagrama de bloques es un método gráfico muy simple que consiste en bloques funcionales conectados con bordes dirigidos.
- Un borde conectado representa el flujo de datos del bloque de origen al bloque de destino.

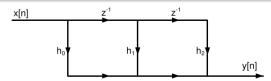


Filtro FIR de tres coeficientes ($y[n] = h_0x[n] + h_1x[n-1] + h_2x[n-2]$).



Gráfico de Flujo de Señal (SFG)

- Es una versión simplificada de un diagrama de bloques.
- El funcionamiento de la multiplicación con una constante y los retrasos están representados por aristas, mientras que los nodos representan operaciones de suma, resta y de entrada y salida (I/O).
- SFGs se utilizan principalmente en la descripción de los algoritmos DSP.



Filtro FIR de tres coeficientes ($y[n] = h_0x[n] + h_1x[n-1] + h_2x[n-2]$).



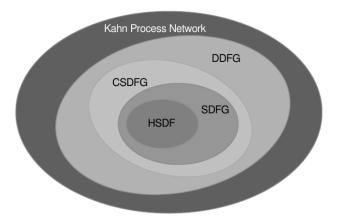
Gráfico de flujo de datos (DFG)

- Se describe un algoritmo de procesamiento de señales mediante un grafo dirigido $G = \langle V, E \rangle$, donde un nodo $v \in V$ representa una unidad computacional o, en un diseño jerárquico. Un borde dirigido $e \in E$ desde un nodo fuente a un nodo destino representa un buffer FIFO o simplemente precedencia de ejecución. También se utiliza para representar los retrasos algorítmicos introducidos en los datos mientras se mueve desde el nodo de origen al nodo de destino.
- Es de especial interés para los diseñadores de hardware, ya que captura la propiedad impulsada por datos de un algoritmo DSP.
- Además, expone la concurrencia oculta entre diferentes partes del algoritmo. Un DFG puede utilizarse para representar algoritmos DSP síncronos, asíncronos y de tasa múltiple.
- Motiva al diseñador a pensar en términos de componentes, mejorando la reutilización de hardware.
- También ayuda en la optimización, testing y verificación a nivel de módulos.





Representaciones Gráficas



Kahn Process Network - Dynamic DFG (DDFG) - Cycle Static DFG (CSDFG) - Synchronous DFG (SDFG) - Homogeneous SDFG (HSDFG).



Gráfico de flujo de datos (DFG)

Síncrono DFG

- Un ejemplo de una aplicación de transmisión es un sistema de procesamiento multimedia que consiste en procesos o tareas en las que cada tarea opera sobre un número predefinido de muestras y luego produce un número fijo de valores de salida.
- Estas tareas se ejecutan periódicamente en un orden definido.
- Aquí se conocen a priori el número de tokens consumidos por un nodo en cada uno de sus bordes y como resultado de su disparo el número de tokens que produce en sus bordes de salida.

Detalle



- S: Fuente
- D: Destino
- T_S/T_D : Tiempo de ejecución

- P_S/C_D: Tasa de producción y consumo de los datos
- •: Retardo



Gráfico de flujo de datos (DFG)

Filtro IIR como un Síncrono DFG

- Ecuación en diferencia $y[n] = x[n] + a_1y[n-1] + a_2y[n-2]$
- El gráfico consiste en dos nodos para multiplicaciones y un nodo para la adición, tomando cada uno un token como entrada y produciendo un token en la salida.
- Los caminos de realimentación desde la salida a los dos multiplicadores cada uno requieren un retardo.
- Estos retrasos se muestran con puntos negros en los respectivos bordes.

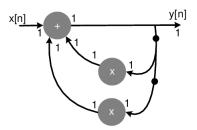




Gráfico de flujo de datos (DFG)

Disparo Auto-temporizado (Self-timed Firing)

- En un disparo auto-temporizado, un nodo se activa tan pronto como se obtiene el número necesario de tokens en sus bordes entrantes.
- Para implementar SDFG en hardware dedicado, se puede usar la ejecución autotemporizada de nodos o un vector de repetición se puede calcular primero usando un conjunto equilibrado de ecuaciones.
- Este vector calcula disparos múltiples de cada nodo para hacer el SDFG consistente.
- Una implementación SDFG auto-temporizada generalmente da como resultado una fase transitoria, y después de eso la secuencia de disparo se repite en una fase periódica.



Gráfico de flujo de datos (DFG)

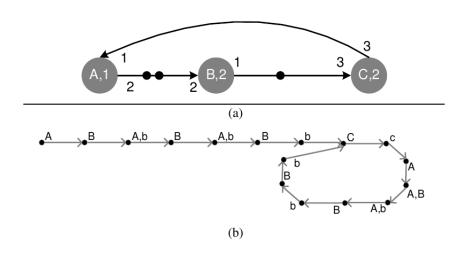




Gráfico de flujo de datos (DFG)

Simple y multi Tasa (Single-rate and Multy-rate) SDFGs

- En un SDFG de tasa única, la tasa de consumo r_c es la misma que la tasa de producción r_p .
- En un SDFG multi tasa, estas tasas no son iguales y uno es un múltiplo racional del otro.
- Un sistema es de tasa múltiple de decimación cuando $r_c > r_p$.
- Un sistema es de interpolación cuando $r_p < r_c$.

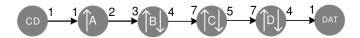




Gráfico de flujo de datos (DFG)

Homogeneous SDFGs

- Es un caso especial de un gráfico de tasa única en el que cada nodo produce un token o valor de datos en todos sus bordes salientes.
- Extendiendo esta cuenta a bordes entrantes, cada nodo también consume un valor de datos de todos sus bordes entrantes.
- Cualquier SDFG consistente se puede convertir en HSDFG.
- La conversión da una medida exacta del rendimiento, aunque la conversión puede dar lugar a un aumento exponencial en el número de nodos y por lo tanto puede ser muy complejo para la interpretación y la implementación.
- La forma más sencilla de convertir un SDFG coherente a HSDFG es encontrar primero el vector de repetición y luego hacer copias de cada nodo tal como se da en el vector de repetición y dibujar apropiadamente los bordes de los nodos de origen a los nodos de destino de acuerdo con el SDFG original.



Homogeneous SDFG - Ejemplo

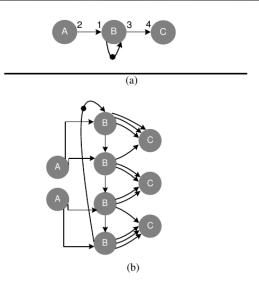




Gráfico de flujo de datos (DFG)

Cyclo-static DFG

- En un DFG ciclo-estático, el número de tokens consumidos por cada nodo, aunque varía de iteración a iteración, exhibe periodicidad y repite el patrón después de un número fijo de iteraciones.
- Adecuado para modelar varias aplicaciones de procesamiento de señal, ya que proporciona flexibilidad de diferentes tasas de producción y consumo de cada nodo siempre que el patrón se repita después de un número finito de iteraciones.
- Esta representación también funciona bien para diseños en los que una secuencia periódica de funciones está asignada en el mismo bloque HW.





Gráfico de flujo de datos (DFG)

Gráficos de control de flujo

- Es adecuado para procesar un algoritmo de control.
- Estos algoritmos se encuentran generalmente en la implementación de protocolos de comunicación o diseños de controladores para rutas de datos.
- Combina funcionalidad controlada por datos y control específica de un algoritmo.
- Cada nodo del DFG representa una operación matemática, y cada borde representa la precedencia o una dependencia de datos entre las operaciones.
- Un CDFG puede cambiar el número de tokens producidos y consumidos por los nodos en diferentes ajustes de entrada.
- Un DFG con tasas variables de producción y consumo se denomina gráfico de flujo de datos dinámico (DDFG).



Gráfico de flujo de datos (DFG)

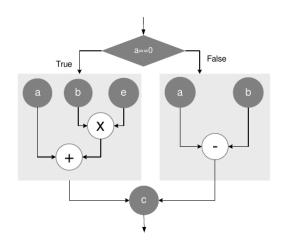




Gráfico de flujo de datos (DFG)

Máquina de estado finita (FSM)

- Se utiliza para proporcionar señales de control a un camino de datos de procesamiento de señales para ejecutar un cálculo con una selección de operandos de un conjunto de operandos.
- Un FSM genérico asume que un sistema está en uno de un número finito de estados.
- Un FSM tiene un estado actual y, basado en un evento interno o externo, calcula el siguiente estado.
- De esta manera, una vez que el FSM transita de un estado a otro, el datapath mantiene la implementación de diferentes porciones del algoritmo.
- Además de implementar un planificador, el FSM también funciona bien para implementar protocolos donde se sigue un conjunto de procedimientos y sincronización entre varios componentes del sistema, como con el arbitraje de buses compartidos.



Gráfico de flujo de datos (DFG)

Transformaciones en un gráfico de flujo de datos

- Las transformaciones matemáticas convierten un DFG en un DFG más apropiado para la implementación de hardware.
- Estas transformaciones cambian la implementación del algoritmo de tal manera que el algoritmo transformado satisface mejor las metas específicas del diseño.
- De un conjunto de objetivos de diseño, el diseñador puede querer minimizar el retardo de la ruta crítica o el número de registros.
- Retiming, plegado, despliegue y look-ahead son algunas de las transformaciones utilizadas comúnmente.







Clasificación

Período de iteración

- Para un sistema de procesamiento de señal de tasa única, una iteración del algoritmo adquiere una muestra de un convertidor A/D y realiza un conjunto de operaciones para producir una muestra de salida correspondiente.
- El tiempo que tarda el sistema en calcular todas las operaciones en una iteración de un algoritmo se denomina período de iteración y se mide en unidades de tiempo o en número de ciclos.
- Para un sistema digital genérico, la relación entre la frecuencia de muestreo f_s y la frecuencia de reloj de circuito f_c es importante.
- Cuando éstos son iguales, el período de iteración es determinado por el camino crítico.
- En los diseños donde f_c > f_s, el periodo de iteración se mide en términos del número de ciclos de reloj requeridos para calcular una muestra de salida.



Clasificación

■ Período de muestreo y Throughput

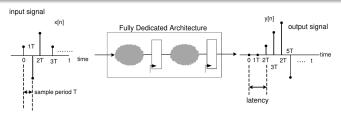
- El periodo de muestreo T_S se define como el tiempo promedio entre dos muestras de datos sucesivas.
- El periodo especifica el número de muestras por segundo de cualquier señal. El requisito de frecuencia o de frecuencia de muestreo ($f_s = 1/T_s$) es específico de una aplicación y, posteriormente, obliga al diseñador a producir hardware que puede procesar los datos que se introducen en el sistema a esta tasa.
- A menudo, esta restricción requiere que el diseñador minimice los retrasos en la ruta crítica.
- Pueden reducirse utilizando unidades de cálculo más optimizadas o añadiendo retrasos en la lógica
- Los retrasos de pipelining añaden latencia en el diseño.
- En los diseños donde f_s < f_c, el diseñador digital explora vías de intercambio de recursos para la reutilización óptima de bloques computacionales.



Clasificación

Latencia

- La latencia se define como el retardo de tiempo para que el algoritmo produzca una salida y[n] en respuesta a una entrada x[n].
- En muchas aplicaciones los datos se adquieren en un buffer y después se introducen para su procesamiento, introduciendo latencia.
- Además de los retrasos algorítmicos, los registros de pipelining son la fuente principal de latencia en una sistema.
- Generalmente hay una relación inversa entre el camino crítico y la latencia.
- Con el fin de reducir la ruta crítica, se añaden registros de pipeline que dan lugar a un aumento de la latencia del diseño.





Clasificación

Potencia

- La disipación de potencia estática se debe a la corriente de fuga en la lógica digital.
- La disipación de potencia dinámica se debe a toda la actividad de conmutación, la cual constituye la mayor parte de la disipación de potencia en un diseño.
- La disipación de potencia dinámica es específica del diseño, mientras que la disipación de potencia estática depende de la tecnología.
- En un FPGA, la disipación de potencia estática se debe a la corriente de fuga a través de diodos de polarización inversa.
- En el mismo FPGA, el uso de la energía dinámica depende de la frecuencia de reloj, la tensión de alimentación, la actividad de conmutación y la utilización de recursos.
- En el nivel de transferencia de registro (RTL), el diseñador puede determinar las partes del diseño que no están realizando cálculos útiles y se pueden apagar para ahorrar energía.
- Se utiliza la técnica gated clock para detener selectivamente el reloj en áreas que no están realizando cálculos en el ciclo actual.





El espacio de diseño

- Para diseñar una lógica óptima para un problema dado, el diseñador explora el espacio de diseño donde hay varias opciones disponibles para mapear algoritmos en tiempo real en hardware.
- La frecuencia de reloj máxima alcanzable del circuito f_c y la tasa de muestreo requerida de la entrada de datos al sistema f_s juegan un papel crucial en la determinación y selección de la mejor opción.
- Los dispositivos digitales como FPGAs y ASICs pueden ejecutar la lógica a velocidades de reloj en el rango de 30 MHz a 800 MHz aproximadamente.
- En mucho diseño la arquitectura esta definida pero es posible mejorar el área aplicando otras técnicas en las operaciones de adición.
- Lo mismo se aplica a los multiplicadores y shifters. Después de un mapeo uno a uno del DFG a la arquitectura, el diseño se evalúa para satisfacer la tasa de datos de entrada del sistema.
- En ocasiones, en las arquitecturas óptimas para las operaciones básicas, el diseño sintetizado no cumple los requisitos de tiempo.
- El diseñador necesita emplear transformaciones matemáticas apropiadas o puede agregar registros de pipeline para obtener una mejor sincronización.

Pipelining

- La asignación de un gráfico de flujo de datos (DFG) transformado a una arquitectura totalmente dedicada es trivial, ya que cada operación se asigna a su operador de hardware coincidente.
- En muchos casos, esta asignación puede consistir en rutas con lógica combinatoria que viola las restricciones de tiempo.
- Por lo tanto, es imperativo romper estas nubes combinacionales con registros.
- Para las rutas de feedforward, el diseñador también tiene la opción de agregar registros de pipeline en el camino principal.
- Mantener la coherencia de los datos en el gráfico es un problema crítico en el pipelining.
- El diseñador debe cerciorarse de que el datapath trazado de cualquier entrada primaria a cualquier salida primaria pasa con el mismo número de registros de la pipeline.
- En contraste, no existe una forma sencilla de añadir registros de pipeling en los caminos realimentados (feedback).



Arquitecturas Dedicadas Ejemplo

Ecuación en diferencia:

$$d[n] = a[n] + b[n]$$
 (1)

$$out[n] = (d[n-1]-c[n])*e[n]$$
 (2)

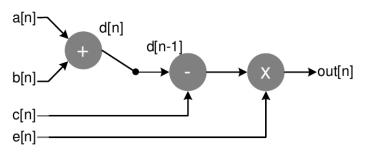


Diagrama de flujo de datos.



Ejemplo - Pipelinig

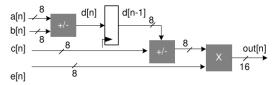
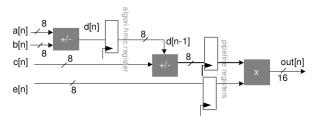


Diagrama en Bloque.



Pipeling



Selección de Bloques Diseño Básicos

- Partiendo de un gráfico de flujo de datos que representa un algoritmo de procesamiento de señales se debe elegir los bloque básicos que menos recursos utilicen en el diseño.
- Se busca diseñar una arquitectura óptima con un área mínima y la mejor sincronización mientras mapea el DFG a una arquitectura totalmente dedicada.

Basic building blocks	Relative timing	Relative area
Adder 1 (A1)	T	2.0A
Adder 2 (A2)	1.3T	1.7A
Adder 3 (A3)	1.8T	1.3A
Multiplier 1 (M1)	1.5T	2.5A
Multiplier 2 (M2)	2.0T	2.0A
Multiplier 3 (M3)	2.4T	1.7A



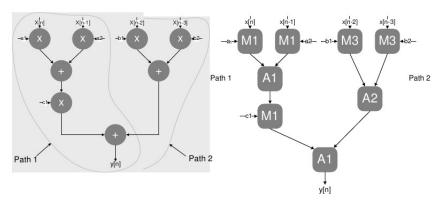


Diagrama de Flujo de Datos y Arquitectura Totalmente Dedicada (FDA).



DFG a Hardware

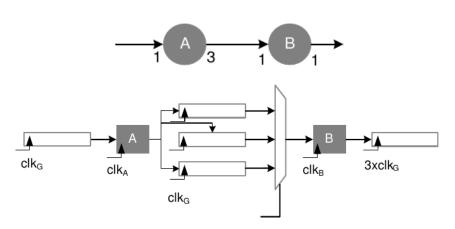
- Un gráfico de flujo de datos proporciona una representación visual completa a los algoritmos de procesamiento de señales.
- Cada nodo se caracteriza por las tasas de producción y consumo en cada puerto.
- En muchos casos, existen varias opciones para un componente representado como un nodo en el DFG en la biblioteca de diseño.
- Estos componentes intercambian el área con el tiempo de ejecución o el número de ciclos necesarios para procesar los datos de entrada.



Mapeo de hardware de un sistema Multi-Rate

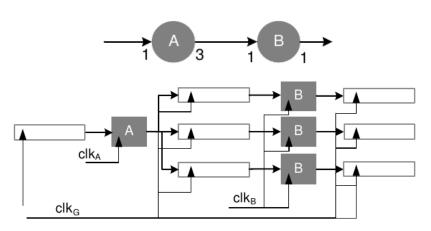
- Cada etapa se considera con sus índices de producción y consumo de muestras.
- Para una ventaja que requiere una producción de tasa múltiple para un consumo de tasa única, es posible una implementación paralela o secuencial.
- Una implementación paralela invoca cada nodo de destino varias veces, mientras que un ajuste secuencial almacena los datos en registros y lo introduce secuencialmente en un único nodo de destino.





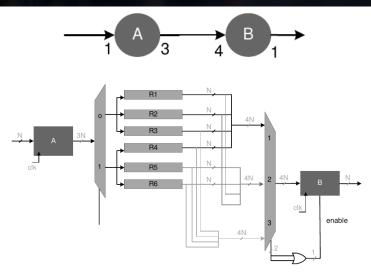
Multi-Rate en formato Secuencial.





Multi-Rate en formato Paralelo.



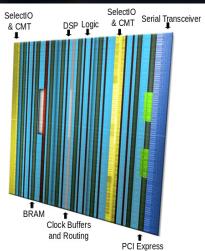


Multi-Rate con tasa de producción menor a la tasa de consumo.



Field Programmable Gate Array

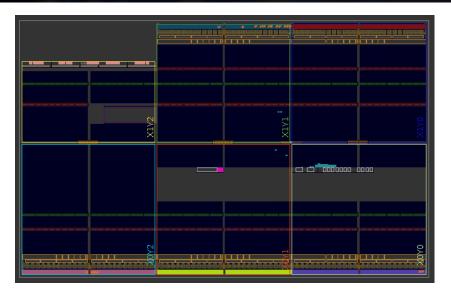




Esquema simplificado de una FPGA (fuente obtenida de "11_7-Series Architecture Overview").

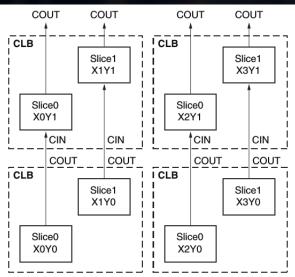


Arquitecturas Dedicadas FPGA - Regiones de Clock



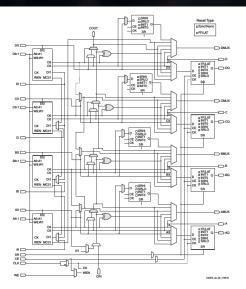


FPGA - Configurable Logic Block



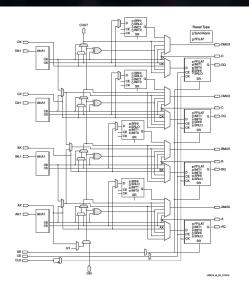


FPGA - SliceM



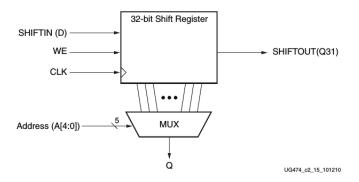


FPGA - SliceL



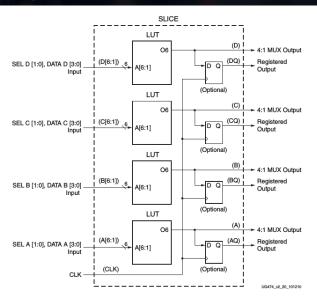


FPGA - Shift Register



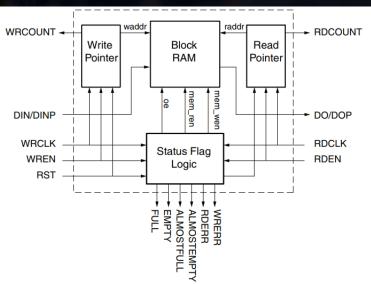


FPGA - Multiplexor



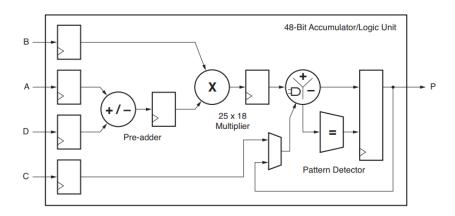


FPGA - Block RAM



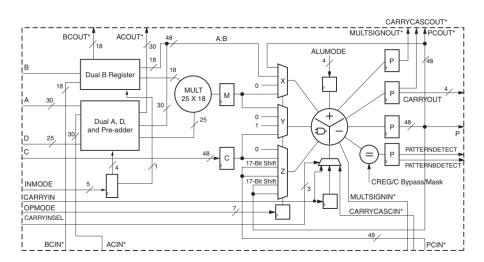


FPGA - DSP



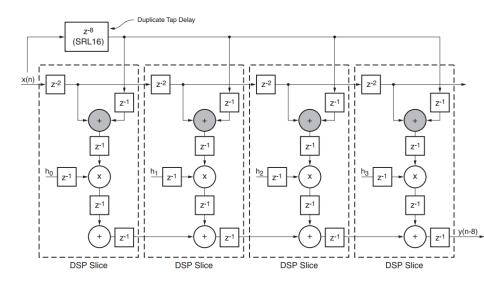


FPGA - DSP



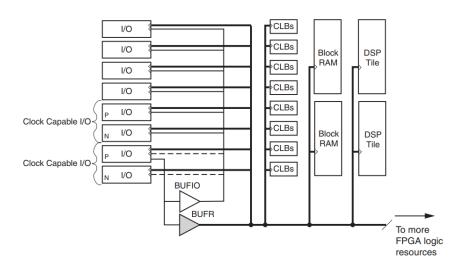


FPGA - DSP FIR



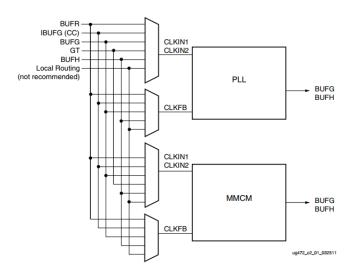


FPGA - Distribución del clock



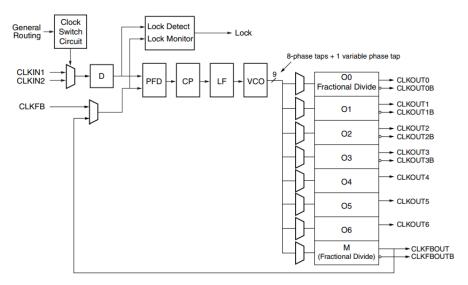


FPGA - Divisores de reloj





FPGA - Mixed-Mode Clock Manager





FPGA - PLL

