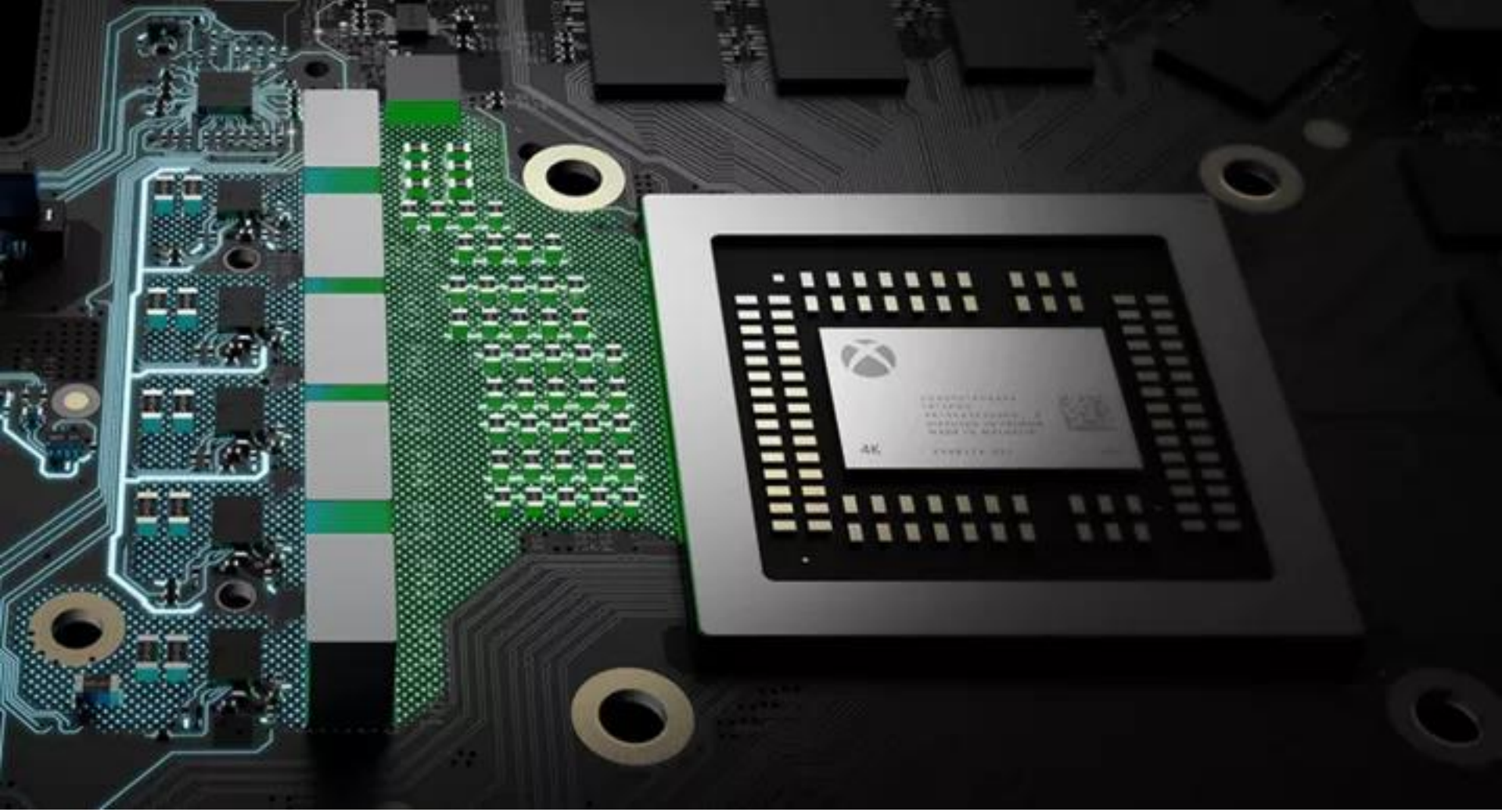


PROGRAMACIÓN en PARALELO

FUNDAMENTOS

COMPUTO en PARALELO

tiene

DISEÑO de PROGRAMAS: Ideas clave

paso

1 Particionamiento

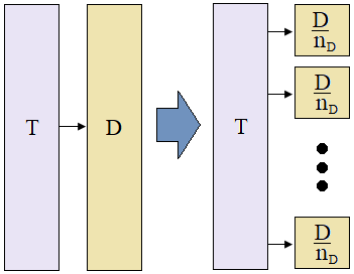
si se necesita o no

2 Coordinación

tipo

Paralelismo de **DATOS**

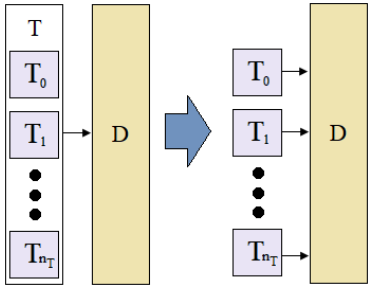
Descomposición del DOMINIO



tipo

Paralelismo de **TAREAS**

Descomposición del FUNCIONAL



en

Comunicación

es

Transferencia de Datos entre Tareas

en

Sincronización

causa

Necesidad de **Cooperación**

hay

Dependencia entre tareas

en

Balance de Carga

es

Cada Procesador debe tener la misma cantidad de trabajo

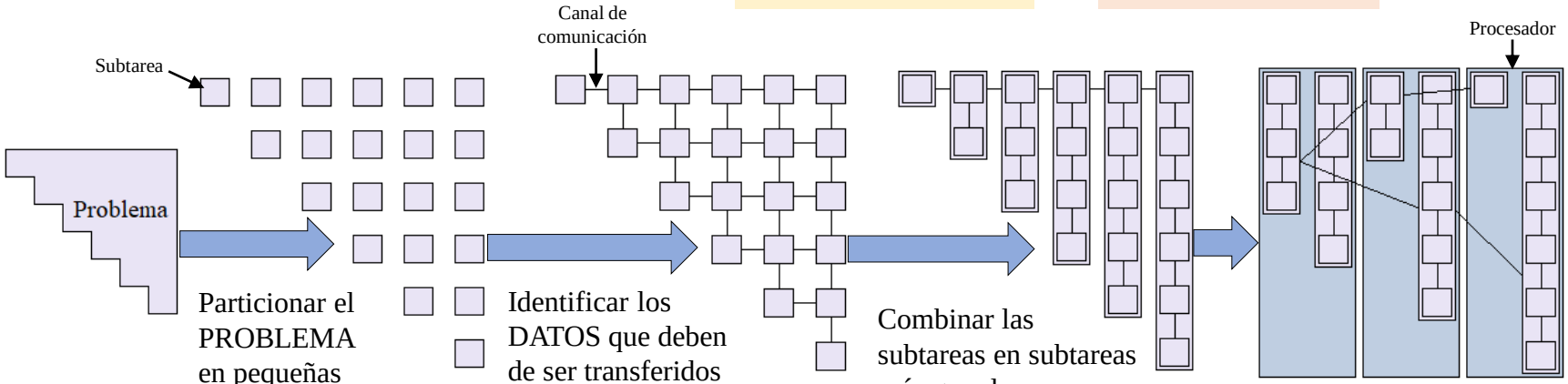
COMPUTO en PARALELO

tiene

DISEÑO de PROGRAMAS: Método de Foster

paso

1 Particionamiento → 2 Comunicación → 3 Unión [Sincronización] → 4 Mapeo [Balance de Carga]



Particionar el PROBLEMA en pequeñas subtareas

Identificar los DATOS que deben de ser transferidos entre cada par de subtareas A y B en términos de un canal de comunicación que conecta A y B

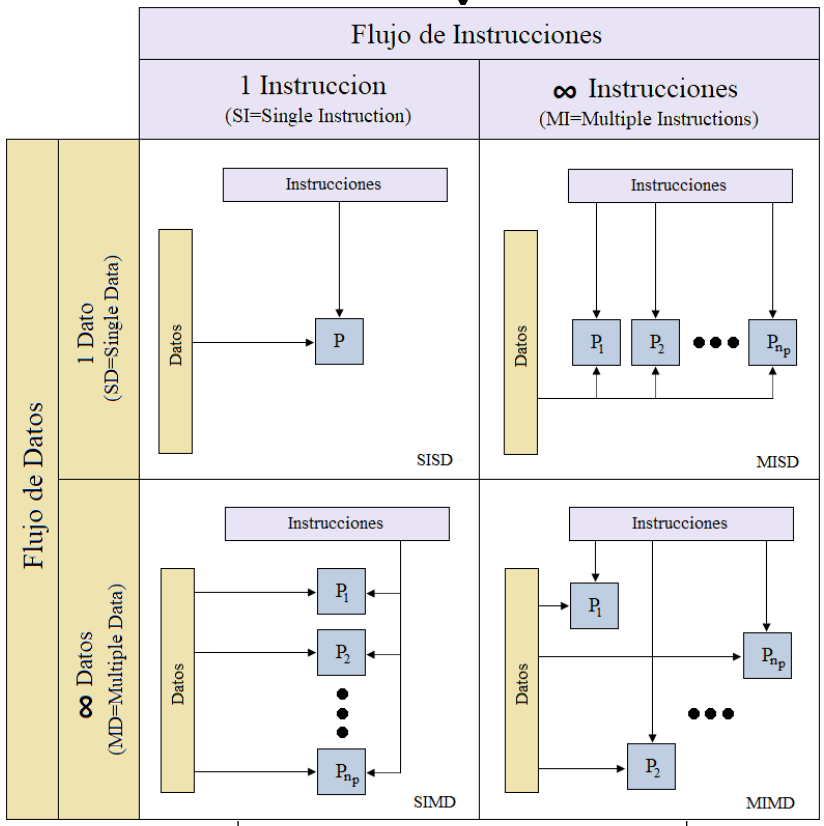
Combinar las subtareas en subtareas más grandes:

- regla: Si la subtarea B depende de la subtarea A entonces unirlas: A→B
- razón: ∞ subtareas → sobrecarga de comunicación
- reduce: Cantidad de DATOS a ser transmitidos entre subtareas
- disminuye

Asignar las subtareas a cada uno de los procesadores de tal forma que se distribuya la misma cantidad de trabajo por procesador y se minimice la comunicación entre procesadores.

CLASIFICACIÓN: Procesamiento (Taxonomía de Flynn)

es



característica

uso en

Paralelismo de Datos

uso en

GPU=Graphics PU
TPU=Tensor PU
VPU=Vector PU

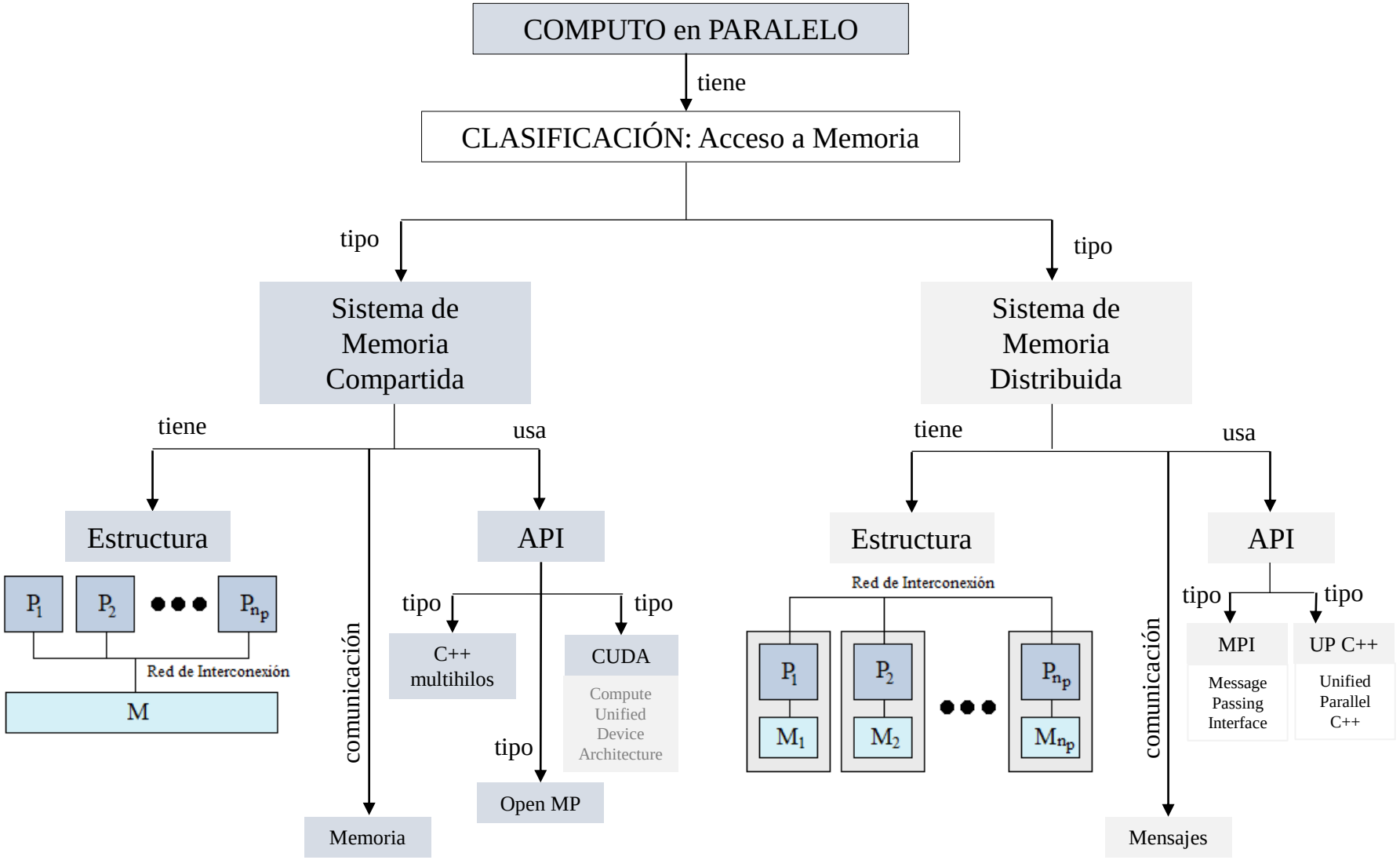
$P_1, P_2, \ldots P_{np}$ son síncronos

característica

uso en

Procesador Multinúcleo

$P_1, P_2, \ldots P_{np}$ son síncronos



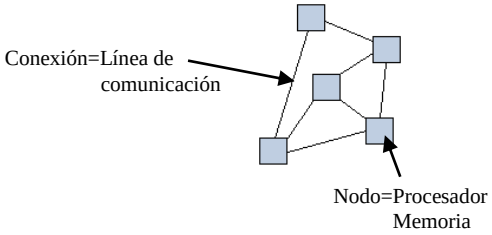
NOTA: P_i=*i*-esima Unidad de Procesamiento
M_i=*i*-esima Memoria
n_p=Numero de Procesadores

RED de INTERCONEXIÓN

es

GRAFO

Conjunto de conexiones entre 2 o más elementos



Determina el nivel de

Paralelización de la comunicación

es

Transferencia simultanea de varios mensajes entre diferentes pares de nodos

definición

Ancho de Banda

Razón a la cual el destino recibe los Datos después de haber recibido el primer byte

Ancho de Banda

Número de bytes en Δt

$$B = \left\lceil \frac{n_b}{\Delta t} \right\rceil$$

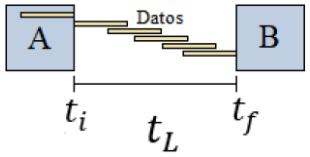
Intervalo de tiempo (s)

definición

t_L=Tiempo de Latencia

es

Tiempo que transcurre entre la transmisión del primer byte de los Datos en A al tiempo t_i hasta la recepción del primer byte de los Datos en B al tiempo t_f

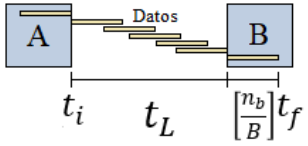


definición

t_T=Tiempo de Transmisión de n_b bytes

es

Tiempo que transcurre entre la transmisión del primer byte de los Datos en A al tiempo t_i hasta la recepción del ultimo byte de los Datos en B al tiempo t_f



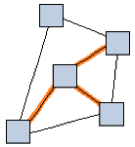
$$t_T = t_L + \left\lceil \frac{n_b}{B} \right\rceil$$

definición

N_G=Grado

es

Número máximo de nodos vecinos



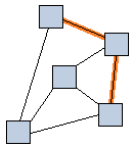
$N_G = 3$

definición

$\downarrow N_D$=Diámetro

es

La máxima distancia de entre todas las distancias mínimas entre pares de nodos



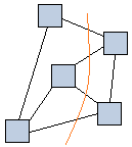
$N_D = 2$

definición

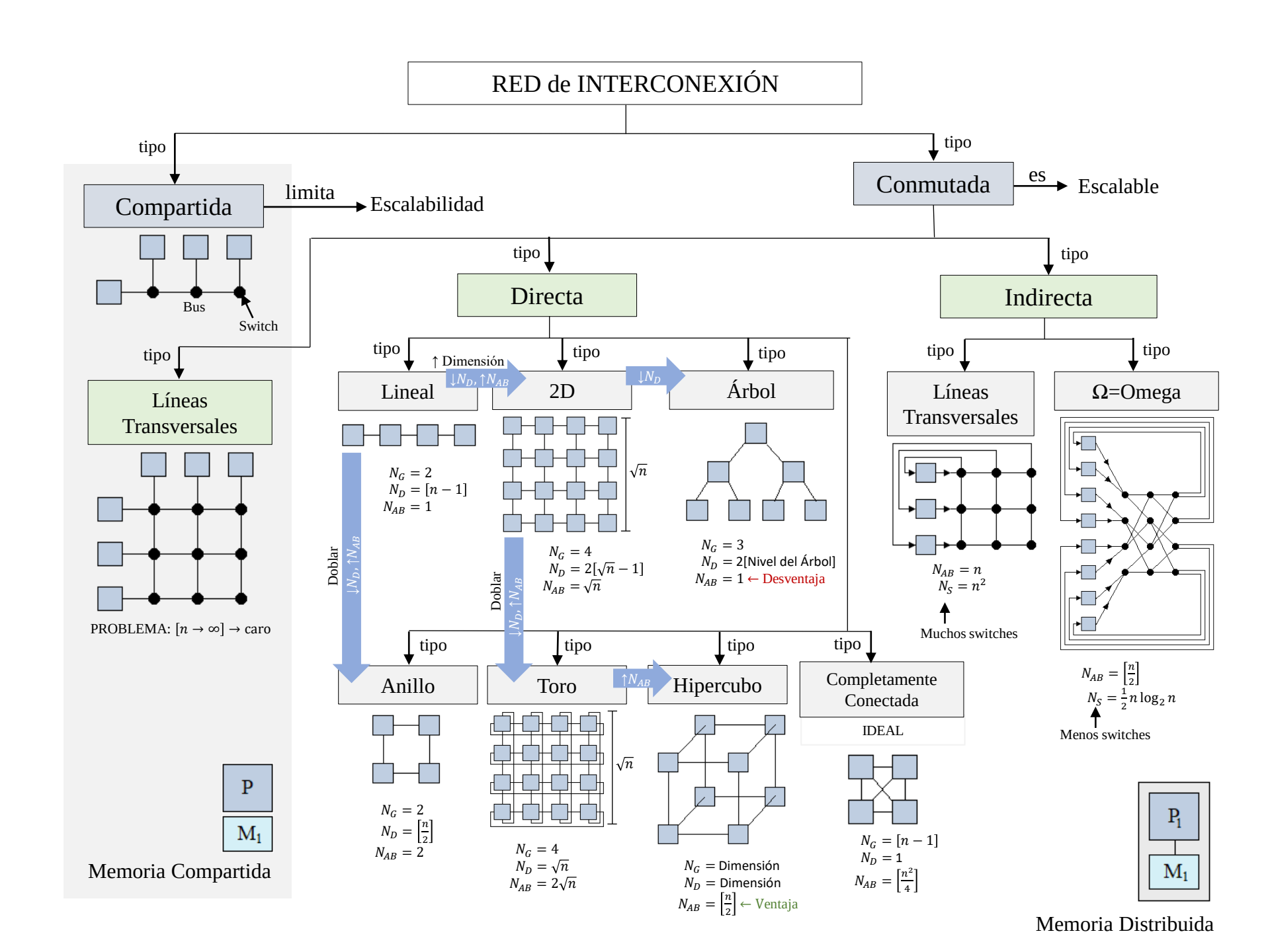
$\uparrow N_{AB}$=Ancho de Bisección

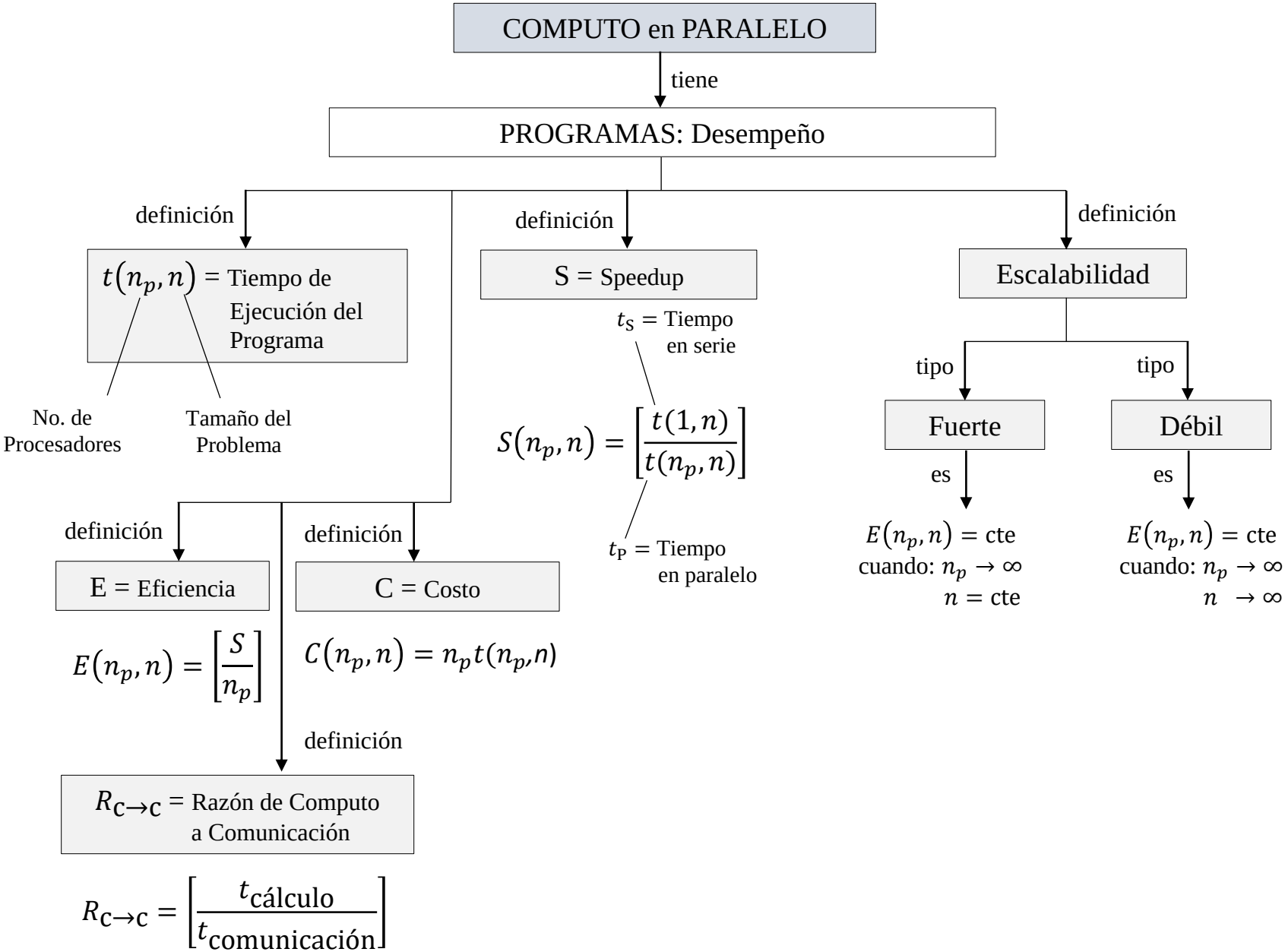
es

Número mínimo de aristas a remover para dividir la red en 2 partes iguales



$N_{AB} = 4$


RED de INTERCONEXIÓN
tipo
tipo
Compartida
limita
Escalabilidad
Conmutada
es
Escalable
Bus
Switch
tipo
tipo
Directa
Indirecta
tipo
tipo
tipo
tipo
tipo
Líneas Transversales
↑ Dimensión
↓N_D, ↑N_{AB}
Lineal
2D
↓N_D
Árbol
Líneas Transversales
Ω=Omega
PROBLEMA: $[n \to \infty] \to$ caro
$N_G = 2$
$N_D = [n-1]$
$N_{AB} = 1$
$\sqrt{n}$
$N_G = 4$
$N_D = 2[\sqrt{n}-1]$
$N_{AB} = \sqrt{n}$
$N_G = 3$
$N_D = 2$[Nivel del Árbol]
$N_{AB} = 1$ ← Desventaja
$N_{AB} = n$
$N_S = n^2$
Muchos switches
$N_{AB} = \lceil \frac{n}{2} \rceil$
$N_S = \frac{1}{2} n \log_2 n$
Menos switches
Doblar
↓N_D, ↑N_{AB}
Doblar
↓N_D, ↑N_{AB}
tipo
tipo
tipo
tipo
Anillo
Toro
↑N_{AB}
Hipercubo
Completamente Conectada
IDEAL
$N_G = 2$
$N_D = \lceil \frac{n}{2} \rceil$
$N_{AB} = 2$
$N_G = 4$
$N_D = \sqrt{n}$
$N_{AB} = 2\sqrt{n}$
$N_G =$ Dimensión
$N_D =$ Dimensión
$N_{AB} = \lceil \frac{n}{2} \rceil$ ← Ventaja
$N_G = [n-1]$
$N_D = 1$
$N_{AB} = \lceil \frac{n^2}{4} \rceil$
P
M_1
Memoria Compartida
P_1
M_1
Memoria Distribuida


COMPUTO en PARALELO
tiene
PROGRAMAS: Desempeño
definición
definición
definición
$t(n_p, n)$ = Tiempo de Ejecución del Programa
No. de Procesadores
Tamaño del Problema
S = Speedup
t_S = Tiempo en serie
$S(n_p, n) = \left[\frac{t(1,n)}{t(n_p,n)}\right]$
t_P = Tiempo en paralelo
Escalabilidad
tipo
tipo
Fuerte
Débil
es
es
$E(n_p, n)$ = cte cuando: $n_p \to \infty$ n = cte
$E(n_p, n)$ = cte cuando: $n_p \to \infty$ $n \to \infty$
definición
definición
E = Eficiencia
C = Costo
$E(n_p, n) = \left[\frac{S}{n_p}\right]$
$C(n_p, n) = n_p t(n_p, n)$
definición
$R_{C \to C}$ = Razón de Computo a Comunicación
$R_{C \to C} = \left[\frac{t_{cálculo}}{t_{comunicación}}\right]$

COMPUTO en PARALELO

tiene

PROGRAMAS: Desempeño

definición

Tiempo de Ejecución Ideal en Paralelo

$$t_{\mathrm{p}}^{\mathrm{Ideal}} = \left\lceil \frac{t_{\mathrm{S}}}{n_{\mathrm{p}}} \right\rceil$$

definición

Tiempo de Sobrecarga

$$t_{\mathrm{sc}} = \left[t_{\mathrm{p}} - t_{\mathrm{p}}^{\mathrm{Ideal}} \right]$$

en

en

Memoria Compartida

es

Comunicación

$\uparrow [\mathrm{Datos}] \rightarrow \uparrow [t_{\mathrm{sc}}]$

Memoria Distribuida

es

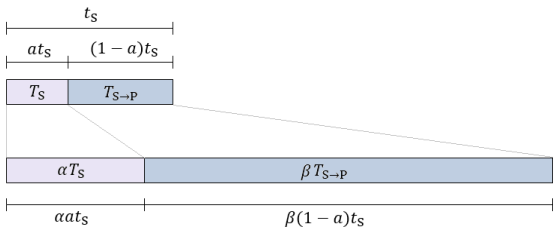
Secciones Críticas

$\uparrow \left[\begin{matrix} \text{secciones} \\ \text{críticas} \end{matrix} \right] \rightarrow \uparrow [t_{\mathrm{sc}}]$

definición

Ley General

$$S = \left\lceil \frac{a + \frac{\beta}{\alpha}[1-a]}{a + \frac{\beta}{\alpha}\left[\frac{1-a}{n_p}\right]} \right\rceil$$

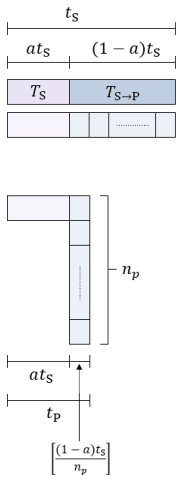


$\left[\frac{\beta}{\alpha}\right]=1$

$\left[\frac{\beta}{\alpha}\right]=n_p$

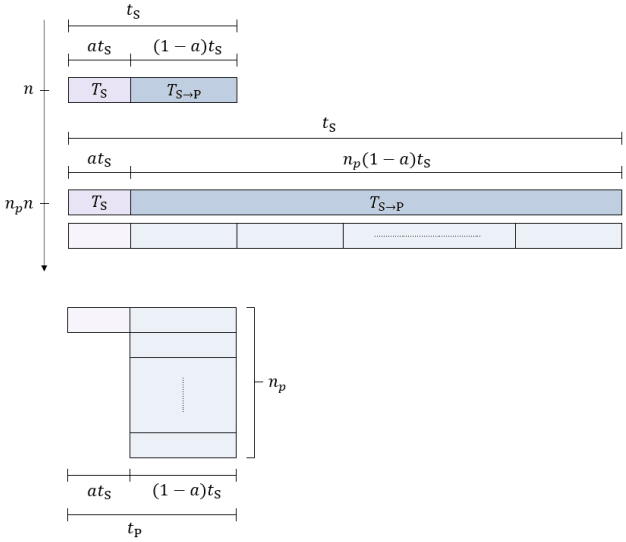
Ley de Amdahl

$$S = \left\lceil \frac{1}{a + \left[\frac{1-a}{n_p}\right]} \right\rceil$$

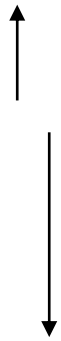


Ley de Gustafson

$$S = [1-a]n_p + a$$



$\left[\frac{(1-a)t_S}{n_p}\right]$ $\left[\frac{1}{n_p}\right](1-a)t_S$



T_S $T_{S\to P}$ $\alpha a t_S$ $\beta(1-a)t_S$ n $n_p(1-a)t_S =$ Tiempo en paralelo

T_S $T_{S\to P}$ $\alpha a t_S$ $\beta(1-a)t_S$ n $n_p n$ $n_p(1-a)t_S =$ Tiempo en paralelo

$\left[\frac{(1-a)t_S}{n_p}\right]$ $\left[\frac{1}{n_p}\right](1-a)t_S$