

## # Digital System Synthesis

### # Boolean satisfiability (SAT)

Find an assignment to the variables  $x_1, \dots, x_n$  such that the Boolean function  $f(x_1, \dots, x_n) = 1$ , or prove that no such assignment exists.

### # Ordered Binary Decision Diagrams (OBDDs)

OBDDs provide a representation of a Boolean function in the form of a binary tree. In contrast to truth tables or algebraic representation OBDDs provide an efficient means of recursive analysis that can tackle the solution of complex SAT problems.

In high-level synthesis OBDDs are applied in:

- Digital circuit design and verification
- Sensitivity analysis of digital circuits
- Finite-state system analysis.

### [Uses of OBDDs and SAT]

EDA:	Testing and Verification	AI: Knowledge base deduction
	Logic Synthesis	Automatic theorem problem
	FPGA routing	
	Path delay analysis	

### [Example: SAT $\rightarrow$ solve the logic equivalence problem]

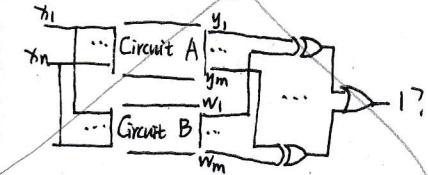
Combinational circuit  $C_A$ , with  $n$  inputs and  $m$  outputs  
Combinational circuit  $C_B$ , with  $n$  inputs and  $m$  outputs

Are the two circuit equivalent?

- A complicated problem formulation:

- Are the outputs equivalent for all input values?
- A simpler alternative:

Are there (at least one) input values that distinguish outputs of two circuits?



### [SAT $\rightarrow$ ATPG Problem in Testing]

ATPG: Automatic Test Pattern Generation

Digital integrated circuits can exhibit defects

Physical defects are modeled as logical faults

Most often used fault mode: single stuck-at fault model

- Circuit lines stuck-at a fixed logic value

\* Sa=0: fixed at 0

\* Sa=1: fixed at 1

Reducing ATPG problem to logic equivalence problem

### Good Circuit

$$\begin{aligned} & \text{Faulty Circuit} \\ & \dots D-1 \\ & (a+l)(\bar{a}+l) \\ & = a\bar{a} + al + \bar{a}l + l \\ & = 0 + al + \bar{a}l + l \end{aligned}$$

### [Conjunctive normal form (CNF)]

A Boolean formula is in CNF if it is a conjunction of clauses  $[k|l:z]$ , where a clause is a disjunction of literals.  $F = ab\bar{c} + ab\bar{c} = (a+b)(\bar{a}+b)(\bar{a}+\bar{c})$

$$\begin{aligned} & = b(\bar{a}+b)(a+b) (\bar{b}+b)(\bar{b}+\bar{c}) (\bar{b}+\bar{c}) \\ & \text{literal clause} (\bar{b}+b) (\bar{b}+\bar{c}) (\bar{b}+\bar{c}) \\ & \text{In other words, CNF is an AND of ORs} \end{aligned}$$

### # Davis-Putnam (DP) SAT algorithm

Recursive Algorithm that creates a binary search tree by making assignments to the remaining variables at each stage of the algorithm.

### [DP Algorithm]

Function DP (clause set S)

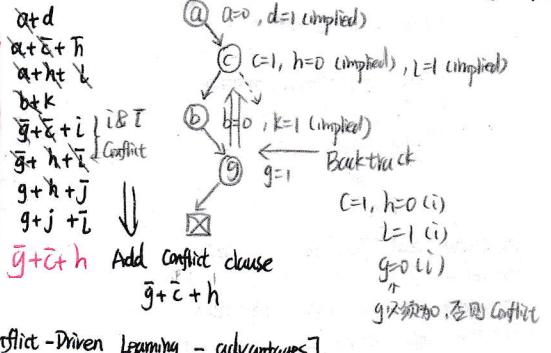
- If S is empty return true
- If S contains a null clause return false
- Choose a variable L in S and split the clause set S recursively for  $L=0$  and  $L=1$ 
  - if  $DP(S(L=0))$  is true (SAT) return true
  - else if  $DP(S(L=1))$  is true (SAT) return true
  - else return false

(Satisfiability Standard)

### # Binary Decision Diagrams

Store the Boolean function in a Directed Acyclic Graph (DAG) representation. Compacted form of the binary decision tree.

### [Conflict Driven Learning and Non Chronological Backtracking Example]



### [Conflict - Driven Learning - advantages]

Learned clause is useful forever

Useful in generating future conflict clauses.

Can restart, i.e. abandon the current search tree and reconstruct a new one

- Add robustness in the solver.

- The clauses learned before the restart are still in the Boolean function after the restart and can help pruning the search space

\* BCP is based on unit clauses (clauses are composed of a Boolean Constraint Propagation (BCP == UP == LDR) style literal)

If a set of clauses contains the unit clause U, the other clauses are simplified by the iterative application:

- Every clause (other than clause U itself) containing U is removed.
- In every clause that contains the negation of  $U = \bar{U}$ , the literal  $\bar{U}$  is removed.

The application of these two rules leads to a new, simpler set of clauses, that is equivalent to the old one.

### Example:

$$F = a(a+b)(\bar{a}+c)(\bar{c}+d)$$

①  $U = a$

$$F = a(a+b)(\bar{a}+c)(\bar{c}+d) = a \cdot c \cdot (\bar{c}+d)$$

$\bar{U}$  itself ↑  
clause that contains U

②  $U = c$

$$F = a \cdot c \cdot (\bar{c}+d) = a \cdot c \cdot d$$

### [Advantages of DLL over PP]

Eliminates the exponential memory requirements

Exponential time is still a problem

Limited practical applicability

Limited size of problems are allowed.

# State how processing time of the DP algorithm is related to the problem size in the worst case.

Briefly outline possible modification of DP that may lead to better computational performance?

$\Rightarrow$  In the worst case DP is NP-complete, i.e. of non-polynomial (exponential) time wrt number of variables.

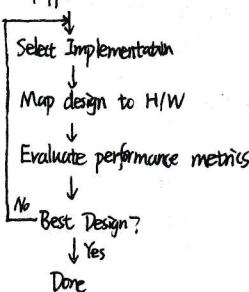
Possible modification include the DLL algorithm with backtracking

Further modification may include conflict driven learning and non-chronological backtracking

后时间先发

## # Design space exploration and synthesis

### Application



Optimisation goal: find the best design in the design space.

Optimisation often has conflict objectives

Typical trade-offs: speed, power, size, cost

{ Evolutionary optimisation (genetic)

Simulated annealing

Integer programming

### # Simulated annealing

- At temperature  $T$ :

- Generate a random perturbation  $x'$  of the current design point  $x$ .

- If the simulated 'energy', i.e. the cost function  $f(x')$  has decreased, move this new design point.

- If the energy  $f(x')$  has increased, the new point is accepted according to the probability of Boltzmann distribution:

$$P = e^{-\frac{f(x') - f(x)}{kT}}$$

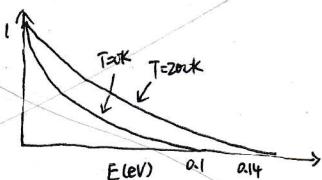
When temperature  $T \rightarrow 0$ ,  $P \rightarrow 0$

$k$  - Boltzmann constant

- When system reaches thermal equilibrium at temperature  $T$ , decrease  $T$  and repeat

- How temperature should change, i.e. the cooling schedule, is important.

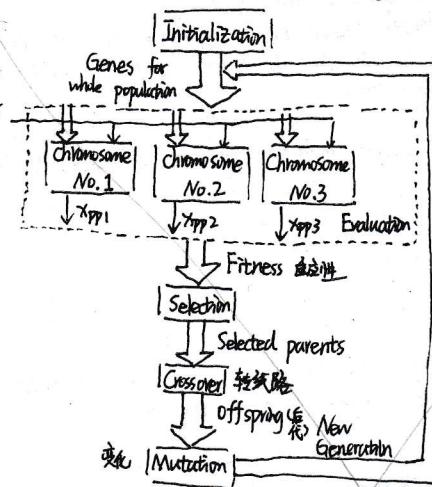
### Boltzmann distribution probabilities



Main adv: guaranteed to find global optimum in the presence of many local optima provided it runs long enough

Main disadv: slow performance, heavily dependent on heuristics (cooling schedule and probability of accepting a worse move).

## # Evolutionary optimisation - basic Genetic Algorithm



### Design $\Rightarrow$ "Chromosome"

- Genes are the design parameters

### Optimization Function $\Rightarrow$ "Fitness"

### Create "Generations" of solutions

- Generation: a population of acceptable (valid) designs

- Select a subset of "fit" designs to breed (繁殖) next generation

- Generate next generation by

- Mutation: modify genes of previous generation

- Crossover: select two (or more) "parents" and create children by combining their genes.

Main advantages of GA:

① Can be used easily in multi-objective optimisation (biggest advantage)

② Is capable of finding the global minimum

③ Easy application to synthesis, both structural and parametric.

④ Lend itself to easy implementations on parallel platforms

Main disadvantages:

① GA can be slow compared with other stochastic methods.

② GA is very sensitive to the fine tuning (微调) of its parameters (e.g. mutation probabilities 变异概率, elitism percentage 精英主义百分比, fitness normalization)

i) Describe how the adaptive body biasing technique is employed in digital designs to provide performance-power trade off.

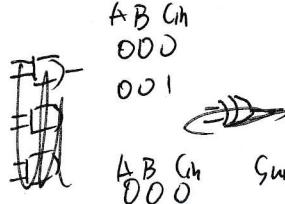
⇒ Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power. Despite its practical complexity, the technique is used in commercial designs such as state-of-the-art - Intel processors.

# include "systemc"

SC-Module (Adder) {

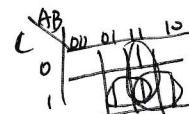
sc\_in <sc\_logic> A, B, Cin;  
sc\_out <sc\_logic> Sum, Cout;

void add();



A	B	Cin	Sum	Cout
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	1	1
1	1	1	0	1

$$\text{Cout} = BC_{in}$$



$$\text{Cout} = BC_{in} + AB + AC_{in}$$

$$= C_{in}(A+B)(A+C)$$

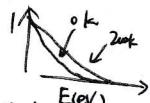
At temperature T,

- generate a random perturbation  $\tau^1$  at the current design point
- If the simulated energy e.g. the cost function  $f(x)$  has decreased, move to this new design point
- If the energy  $f(x)$  has increased, the new design point is accepted according to the probability of boltzmann distribution

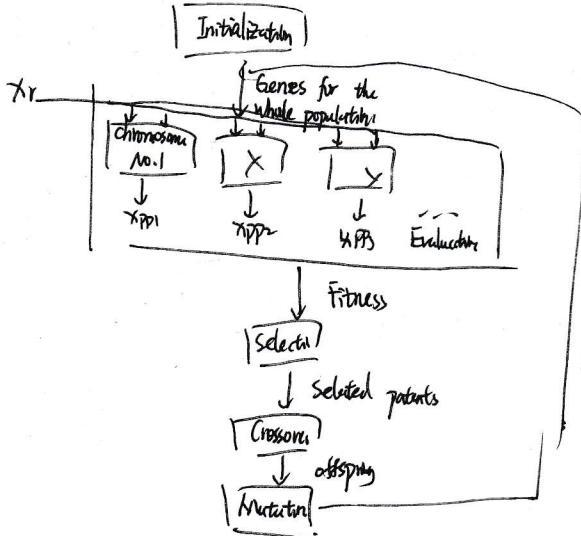
$$P = e^{-\frac{f(x)-f(x')}{kT}}$$

$$P \gg 1, T \gg 1$$

K is the boltzmann constant



When the system reaches thermal equilibrium, reduce temperature and repeat.  
Many local optima proved it runs long enough  
Slow performance, highly dependent on the heuristics controlling the cool scheduling and probability of accepting a worse move.



- { Can be easily used in multi-objective optimization
- Easy to synthesize both structural and parametric
- Is capable of finding a global minimum
- Lend itself easily implementable on parallel platforms
- GA is slow compared to other stochastic methods
- GA is very sensitive to fine tuning of its parameters

At temperature  $T$

- Generate a random perturbation  $x'$  of the current design point  $x$ .

If the simulated 'energy' has decreased, move to this new design point. If the energy first has increased, the new design point is accepted according to the probability of the boltzmann distribution,  $P = e^{-\frac{E_x - E_{x'}}{kT}}$ .

When the system reaches the thermal equilibrium at temperature  $T$ , reduce the temperature and repeat.

How the temperature should change, e.g. the cooling schedule is important to find a global optimum. In the presence of PVT: guaranteed to meet load specification provided it runs long enough. Disadv: slow performance and highly dependent on the heuristic concerning the cooling schedule and the probability of moving to a worse point.

If  $S \leq 0$  accept  
 $\Delta E$  is small change

Choose a width  $L$  in S and split the clause set  $S$  randomly to  $L=0$ .  
 $DPLL(DP)_{L=0}^{unif}$  = tree search for seen and

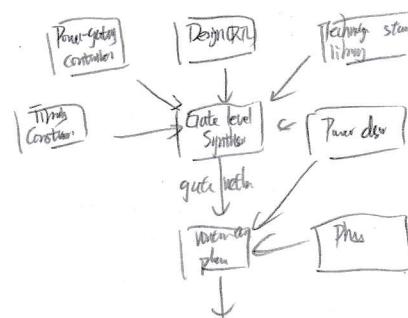
$DPLL(SCL_0)$  tree search through backtrace and implicit graph

Extends this exponential memory requirement  
Limited parallel applicability  
Limited size of problems are solvable

DP edge. At the worst case, DP is an NP-complete problem i.e. of non-polynomial time with regard to size and input

DPLL vs.

Faster resolution includes conflict driven learning and chronological backtracking



ASAP thus we can  
Determine the conditions, criteria is the rule which predicated from basic silicon to  
Schedule cell creation for one cycle like the max cycle of the predication  
Schedule cell creation for one cycle like the max cycle of the predication  
Schedule cell creation for one cycle like the max cycle of the predication

Label each node by the logic net passing through it

Initialize the cycle counter  $L$

Do while  $L < \infty$  for unlinked nodes and until power limit is reached

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

break

Slow S slow start for  $L+1$  as a habitation of new cell

$L = L + 1$  if  $L = N_t$

## # High-level Synthesis steps - Scheduling

### [Automated high-level synthesis]

Determine when to perform each operation

- Scheduling
- Consider alternatives and optimise (e.g. for minimum latency)
- Allocate resources for each operation
- Resource allocation
- Consider alternatives and optimise
- Map operations onto resources
- Binding

### # Scheduling

Scheduling assigns a start time to each operation in DFG

- Start times must not violate dependencies in DFG
- start times must meet performance constraints

#### [ISSUE]

- ① Unconstrained scheduling
- ② Minimum latency scheduling
- ③ Latency constrained scheduling
- ④ Resource constrained scheduling
- ⑤ Minimum-latency, resource constrained
- ⑥ Minimum-resource, latency constrained

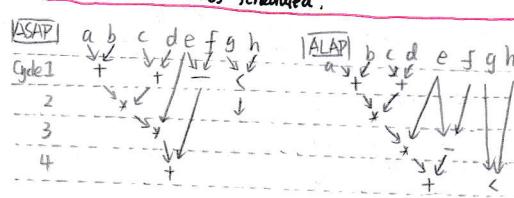
### # Minimum Latency Scheduling

#### ASAP (as soon as possible) algorithm

- Find a candidate node in DFG (candidate node whose predecessors have been scheduled (or there are no predecessors))
- Schedule one node one clock cycle later than max cycle of predecessor.
- Repeat until all nodes scheduled.

#### ALAP (as late as possible) algorithm

- Run ASAP, get minimum latency L
- Find a candidate node in DFG (candidate is node whose successors are scheduled)
- Schedule node one clock cycle before start cycle of successor
  - \* Node with no successors scheduled to last cycle L
- Repeat until all nodes scheduled.



### # Latency constrained scheduling

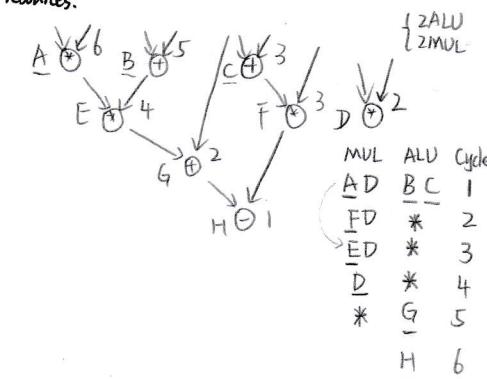
Instead of finding the minimum latency, find latency at most that of a given constraint L.

#### [Solutions]

Use ASAP, verify that minimum latency is at most L  
Use ALAP starting with cycle L instead of minimum latency (no need to run ASAP first, as L is given)

### # Scheduling with Resource Constraints

Schedule must use less than specified number of resources.



### # Minimum-Latency, Resource-Constrained Scheduling

Given resources constraints, find a schedule that has the minimum latency.

[Hu's algorithm] Assume that all resources are of one type (number of resources)

Input: Data flow graph  $G$ , Resource constraints  $R$

① Label each node in  $G$  by the longest path passing through it

② Initialise cycle count:  $L = 1$

③ Repeat: EXAM Standard

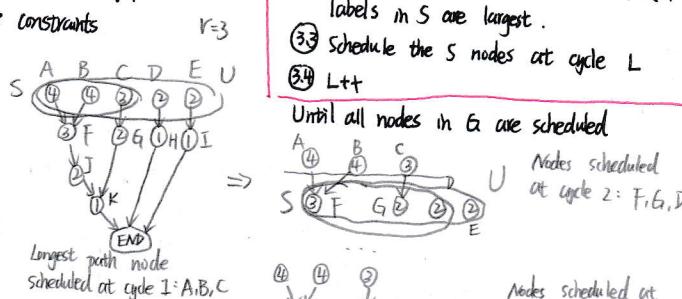
③.1 Determine the subset  $V$  of unscheduled nodes in  $G$  whose predecessors have been scheduled or there are no predecessors

③.2 Select subset  $S$  of  $V$  such that  $|S| \leq R$  and labels in  $S$  are largest.

③.3 Schedule the  $S$  nodes at cycle  $L$

③.4  $L \leftarrow L + 1$

Until all nodes in  $G$  are scheduled



#### [Extensions to Hu's scheduling algorithm]

- { Multiple resource types }
- { Multi-cycle operation }

#### [List Scheduling - minimum latency, constrained]

- Extension for multiple resource types
- Basic idea - run Hu's algorithm for each resource type

1. Input: graph, set of constraints  $R_t$  for each resource type  $t$

2. Label nodes based on max distance to output

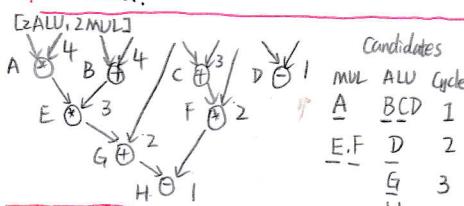
3. For each resource type  $t$

3.1 Determine candidate nodes,  $C$  (no predecessors or scheduled predecessors)

3.2 Schedule up to  $R_t$  operations from  $C$  based on priority, to current cycle

( $R_t$  is the constraint on resource type  $t$ )

4. Increase cycle, repeat from 3 until all nodes scheduled.



#### [Extension for multicycle operations]

- 1. Label the nodes based on max cycle latency to output taking multi-cycle operations into account.

2. For each resource type  $t$

2.1 Determine candidate nodes,  $C$  (those with no predecessors or with scheduled and completed predecessors)

2.2 Schedule up to  $(R_t - N_t)$  operations from  $C$  based on priority, one cycle after the predecessor

( $R_t$  is the constraint on resource type  $t$ )

$N_t$  is the number of resources  $t$  in use from previous cycles.

- Repeat loop 2 until all nodes scheduled.

### # Scheduling for minimum resources with latency constraint.

Optimise: for all schedules that have latency less than the constraint, find the one that uses the fewest resources.

#### [List scheduling for minimum resources]

1. Compute last start times for each node using ALAP with specified latency constraint.

- Latest start times must take into account multicycle operations

2. For each cycle and for each resource type

2.1 Determine candidate nodes

2.2 Compute slack for each candidate

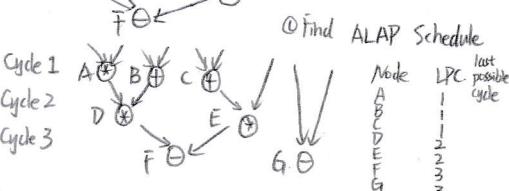
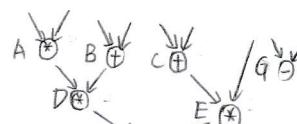
Slack = Current cycle - latest possible value

2.3 Schedule the nodes with 0 slack

Update required number of resources (start with 1 of each at the beginning)

2.4 Schedule the nodes that require no extra resources

3. Repeat loop 2. until all nodes scheduled



① Find ALAP Schedule

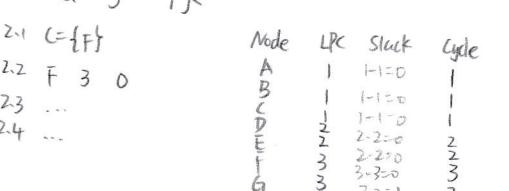
② 2.1  $C = \{A, B, C, G\}$

2.2 Node LPC Slack

Node	LPC	Slack
A	1	1
B	1	1
C	1	1
D	3	3

2.3 Resource = {1 MUL, 2 ALU}

2.4 Schedule the nodes that require no extra resources



#### Minimum-resource, latency-constrained schedule

Step 1: Start with an ALAP schedule and calculate the Latest Possible Start for each node

Step 2: (Extended Hu's algorithm) List scheduling

For each cycle and for each resource type

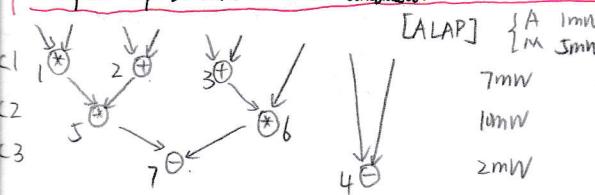
1. Determine candidate nodes
2. Compute slack for each candidate
3. Schedule the nodes with 0 slack
- If necessary, update required number of resources
4. Schedule the nodes that require no extra resources

## # Scheduling for minimum power

[Power constrained scheduling: based on scheduling for minimum resources with latency constraints]

Modification of Hu's algorithm for minimum power consumption:

1. Start with only one type of each resource
2. Complete latest start times for each node using ALAP with specified latency constraint.
3. For each cycle and for each resource type
  - 3.1 Determine candidate
  - 3.2 Compute slack for each candidate
  - 3.3 If there are nodes with 0 slack, schedule all of them - ask for extra resources if necessary
  - 3.4 If there are no nodes with 0 slack, schedule only one node from those with the lowest slack; arbitrary, but first try a node with the highest power consumption.
4. Repeat loop 3. until all nodes scheduled.



Node	LPC	Slack	Cycle
1	1-1	0	1
2	1	0	1
3	1	0	1
4	3/1	2/1/0	3
5	2	2-2	0
6	2	0	2
7	3	3-3	0

$C = \{1, 2, 3, 4\}, R = \{2A + 1M\}$   
 $C = \{4, 5, 6\}, R = \{2A + 2M\}$   
 $C = \{4, 7\}, R = \{2A + 2M\}$

## # Binding 封装

Given a schedule, map operations onto resources such that operations in same cycle do not use same resource

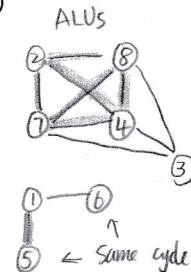
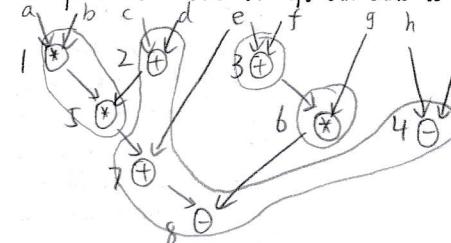
Bad binding may increase resources, require huge steering logic, reduce clock rate.

### [Compatible Graphs]

Each node is an operation

Edges represent compatible operations between a pair of nodes

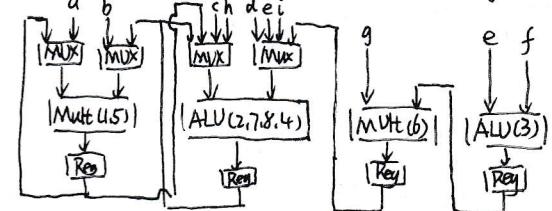
(compatible means that two ops can share a resource)



Note: Fully connected subgraphs can share one resource (all involved nodes are compatible)

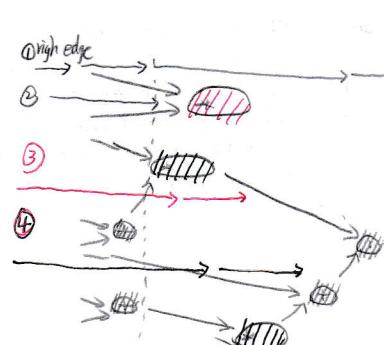
Binding goal: Find minimum number of fully connected subgraphs that cover entire graph

- 1) Add resources and registers
- 2) Add mux for each input
- 3) Add input to left mux for each left input in DFG
- 4) Do same for right mux
- 5) If only one input, remove mux



### # Left Edge Algorithm

- 1) Initialize right\_edge to 0
- 2) Find a node  $N$  whose left edge is  $\geq$  right\_edge
- 3) Bind  $N$  to a particular resource
- 4) Update right\_edge to the right edge of  $N$
- 5) Repeat from 2) for nodes using the same resource type until right\_edge passes all nodes.
- 6) Repeat from 1) until all nodes bound.



Find a valid binding, but they do not consider amount of steering logic required

### [Possible solution]

Use weighted edges/nodes to introduce cost function representing steering logic

Perform clique (clustering) partitioning, finding the set of cliques that minimize weight

0 0 0 0 0 0

## # Low Power

Power of digital circuit:  $P = P_{\text{dynamic}} + P_{\text{leakage}}$   
 $P_{\text{dynamic}} = V_{dd}^2 \cdot F \cdot C_L \cdot E_{SW}$   $\leftarrow$  gate switching activity

Dynamic power reduction happens at various design levels

- Physical layout level  $\Rightarrow$  short interconnect  $\Rightarrow$  smaller  $C_L$
- Transistor level  $\Rightarrow$  transistor with small  $W/L$   $\Rightarrow$  smaller  $C_L$
- Architectural level
  - Switching activity reduction: Clock gating
  - $V_{dd}$  reduction: Multiple  $V_{dd}$

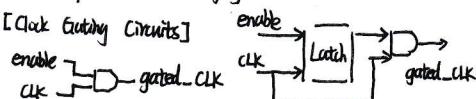
## # Clock Gating (Architectural Level)

When clocked modules (memory, reg files...) are inactive (stored values not changing), use clock gating to disable the clocking feeding module.

[Reason]

FF consumes power due to switching activity even when input is not changing.

### [Clock Gating Circuits]



Latch captures the state of the enable signal and hold it until next clock cycle.

### # Multi $V_{dd}$ (Architectural Level Dynamic Power Reduction)

Modules on critical path use highest  $V_{dd}$  (to meet required timing constraints)

Modules on non-critical paths use lower  $V_{dd}$  (economy of power)

## # Leakage Power Reduction

Leakage power consumption occurs as long as the circuit is powered on.

Sub-threshold current between source and drain in MOS transistor occurs when gate voltage ( $V_{GS}$ ) is below transistor threshold voltage  $V_t$ .

### [Technology Level]

Multiple  $V_t$  technology where fabrication process provides both high and low threshold transistor  $V_t$  (low: fast and leaky and high: slow and less leaky)

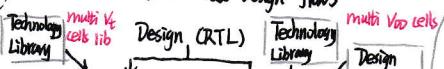
Default design uses high  $V_t$  and careful replacing high  $V_t$  with low  $V_t$ , one can get low  $V_t$  performance and yet significantly lower leakage power.

Example: critical path delay determines performance of the design, all other paths have lower delay. Use low  $V_t$  transistors in modules on the critical path, and high  $V_t$  transistors in modules on non-critical paths

### Many $V_t$ technology

Many libraries are required, hard to use multiple

### [Multi $V_t$ and Multi $V_{dd}$ Design Flow]

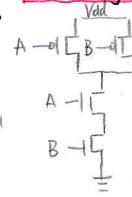


Multi  $V_t$  Design

Multi  $V_{dd}$  Design

## [Stacking Effect]

Sub-threshold voltage current flowing through stacks of series connected transistors reduces when more than one transistor in the stack is turned off. This effect is known as stacking effect.



## [Body Bias] Cost too much

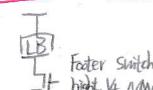
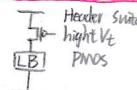
Applying reverse body bias by to transistor reduces leakage and can be used to reduce standby leakage power.

Alternatively apply forward body bias, which reduces  $V_t$  and improves circuit performance but increases leakage

### [Adaptive Body Bias]

Combination of reverse and forward body bias can be applied to reduce leakage power with reduced performance OR improved performance with higher leakage power.

### [Power Gating]



### ① Sleep Transistors

- Insert high  $V_t$  MOSFET between power rail and logic blocks
- Header transistor off lead to logic block floating to near zero
- Footer transistor off lead to logic block floating to near  $V_{dd}$
- Header switches preferred if multiple power rails since gated block logic float to zero irrespective of supply voltages.

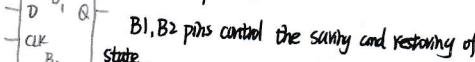
### ② Sleep Transistors to decouple the power-gated logic block from its $V_{dd}$ .

### ③ State Retention Register

State retention registers to restore states after sleep. Typical register will include normal FF with a balloon clutch

Two modes:

1. Sleep mode: state stored in balloon clutch which is powered by  $V_{dd}$
2. Active mode: balloon inactive; do not interfere with normal operation

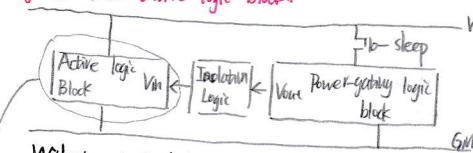


Power gating leads to state logic value being lost due to shut down, some applications require states when active again. Such states can be stored in state retention registers.

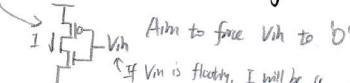
Such registers have two modes of operations.

### ④ Output Isolation

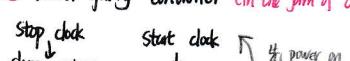
Output isolation to clamp the floating output of the power-gating block and to prevent short circuit current injecting into active logic block.



Without output isolation, floating output causes shortcircuit current in active logics



If  $V_{th}$  is floating, I will be a shortcircuit current



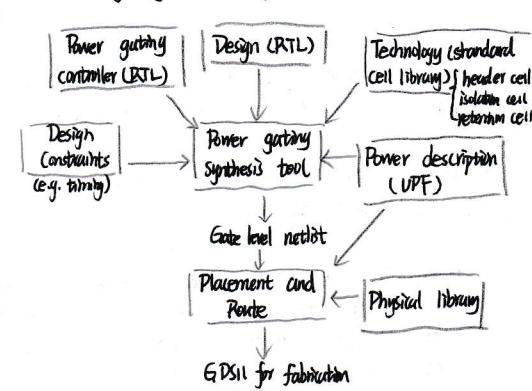
④ Power Gating Controller (in the form of a state machine)

Stop clock  $\downarrow$  clamp output  $\downarrow$  save states  $\downarrow$  power off

Start clock  $\downarrow$  undump output  $\downarrow$  restore states  $\downarrow$  reset

Power on  $\downarrow$  Node-up sequence

## # Power Gating Design Synthesis Flow



Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power.

GDSII for fabrication

Adaptive body biasing applies reverse body bias or forward body bias to transistors to reduce leakage power and reduced performance, or to improve performance with higher leakage power

## # Energy Management

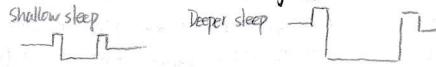
### # Dynamic power management (DPM)

Shut down processing elements (PE) when inactive.

Different algorithms are used, normally engineered as part of the OS:

- Greedy: Go to sleep as soon as processing is finished
- Timedout: Stay on expecting a new request. After time  $t$  of idleness go to sleep.

- Adaptive Timeout: Vary timeout  $t$  depending on activity
- Shallow sleep states are entered and exited quickly (few clock cycles), deep sleep states cost more in time and energy (dozens of clock cycles)



### # Dynamic voltage frequency scaling (DVFS)

Adapt PE performance to requirement by dynamically changing DE clock frequency and  $V_{dd}$

Processing elements have non-uniform workloads which DVFS exploits. The change in frequency or  $V_{dd}$  normally is carried out using energy-efficient operation systems "add-on" such as Linux powersaving governors.

Turning off or reducing  $V_{dd}$  and hence power optimisation is possible due to the presence of idle time and slack time in systems schedule.

In DVFS, power changes cubically with frequency-voltage scaling, and overall performance is roughly linear with clock frequency.

$V_{dd}$	Frequency	Power Savings	Performance Loss
100%	100%	0	0
95%	95%	14%	5%
90%	90%	27%	10%

### # Energy Management : System level

Idle time: periods where DEs do not experience any workload

Slack time: amount of time that a PE is idle during the execution of a function. Results from tasks finishing execution before their deadline.

Why? Slack = deadline - finish time

{ Not all PEs utilised constantly during run-time  
Performance of PEs can not be perfectly during run-time }

### # Interplay of DPM and DVFS

Energy saved with DVFS come at the cost of increased execution time, implying shortened idle period after applying DPM, and greater leakage energy consumption.

Ideal DVFS policy must understand this interplay and hence perform V-f setting selection with the objective of reducing the overall system energy consumption.

This depends on the processor workload's characteristics

- DPM, the characteristic is in terms of the distributions of idle period durations

- DVFS, it is in terms of CPU/memory intensiveness of the executing tasks

DPM and DVFS outperform each other under different workloads, no single PM policy fits perfectly all operating conditions, hence hybrid power management is needed.

// Describe the necessary interaction between the system software and power-gated and dynamic voltage frequency scaling (DVFS) enabled processors in order to achieve power reduction.

⇒ To activate the power gating, sleep signals need to be generated. Such signals are often enabled by the system software (e.g. OS) resulting from monitoring the activities of the processors, inactive processors are candidates for shutting down.

Similarly to activate DVFS, the processor's performance counters such as CPU utilisation give indication how much processing is taking place. The software through special register  $VCO$  will adapt the voltage/frequency of the processors according to tasks

### # SystemC

// 1-Bit Adder , Adder.h

# include "systemc.h"

SC\_MODULE(Adder) {

```
SC_OUT<SC_LOGIC> Sum, Cout;
SC_IN<SC_LOGIC> A, B, Cin;
void add() {
    SC_LOGIC tempC, tempD, tempE;
    tempC = A.read() & B.read();
    tempD = A.read() ^ B.read();
    tempE = Cin.read() & tempD;
    Sum.write(tempD ^ Cin.read());
    Cout.write(tempC | tempE);
}
```

SC\_CTOR(Adder) {

SC\_METHOD(add);

```
sensitive << A << B << Cin;
```

// 8-bit Counter

# include "systemc.h"

SC\_MODULE(Counter8bit) {

```
SC_IN<SC_CLK> cClock;
SC_IN<SC_BOOL> enable;
SC_IN<SC_BOOL> error;
SC_IN<SC_BOOL> reset;
SC_OUT<SC_UINT> counter_out;
```

SC\_UINT<8> count;

void incr\_count() {

while(true)

{

if(enable.read()) {

if(reset.read()) {

count = 0;

counter\_out.write(count);

}

else {

count = count + 1;

counter\_out.write(count);

}

wait();

SC\_CTOR(Counter8bit) {

SC\_THREAD(incr\_count);

sensitive << clock, posc;