Timing-Aware Decoupling Capacitance Allocation in Power Distribution Networks

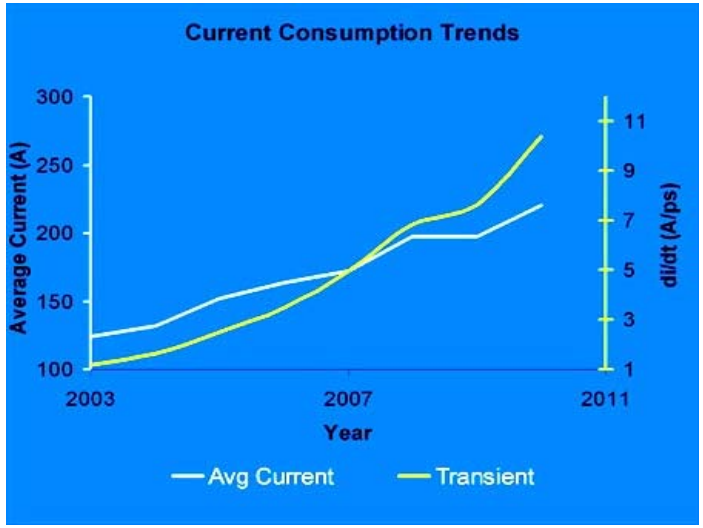
Sanjay Pant, David Blaauw

Electrical Engineering and Computer Science

University of Michigan

Motivation

- Power supply integrity issues
  - Functional failure
  - Performance failure
- Ldi/dt drop becoming significant
  - Large amounts of extrinsic decap added to suppress L*di*/d*t*
- Explicitly added decap is not free
  - Decap oxide leakage increasing with each technology generation
  - Decap leakage may limit the amount of extrinsic decap
- Proposed work
  - Decap added optimally to improve circuit performance
  - Utilizes the timing slack available in the circuit
    - Non-critical gates can tolerate relatively larger supply drop



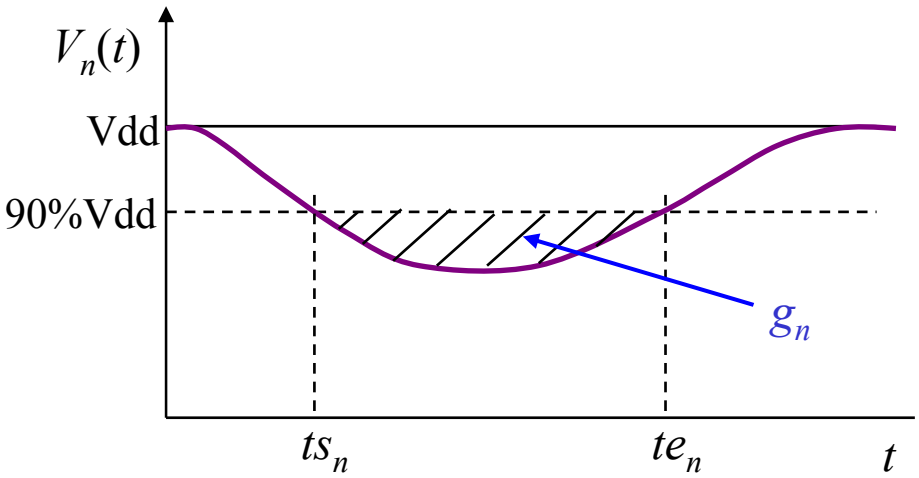
[*Apache*]-*ITRS 2004*

Outline

- Traditional Methods and Prior Work
- Proposed Approach
- Experimental Results
- Conclusion

Prior Works On Decap Allocation

- Allocate decap with objective of minimizing drop at all nodes
- Decap sizes w_j are the opt. variables



g_n = violation at node n

$$Noise\ Metric = \sum_{nodes} g_n$$

Minimize *Noise Metric*

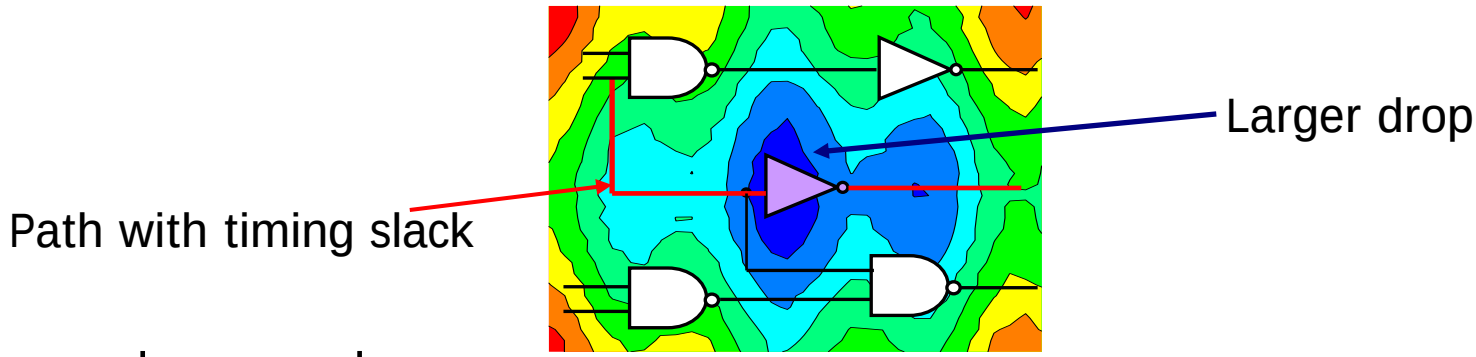
Subject to *constraints on decap sizes*

- Adjoint sensitivity method for sensitivity of noise metric to decap sizes

Sapatnekar [ISPD-02], Roy [DAC-00], Li [DAC-05]

Proposed Approach

- Prior approaches
  - Constrain voltage drop at all nodes regardless of connected gates being critical
  - May not be optimal for maximum performance
- Observation
  - Gates which are not timing critical can afford relatively larger voltage drop
  - Lesser decap area and leakage for same performance if circuit has timing slack

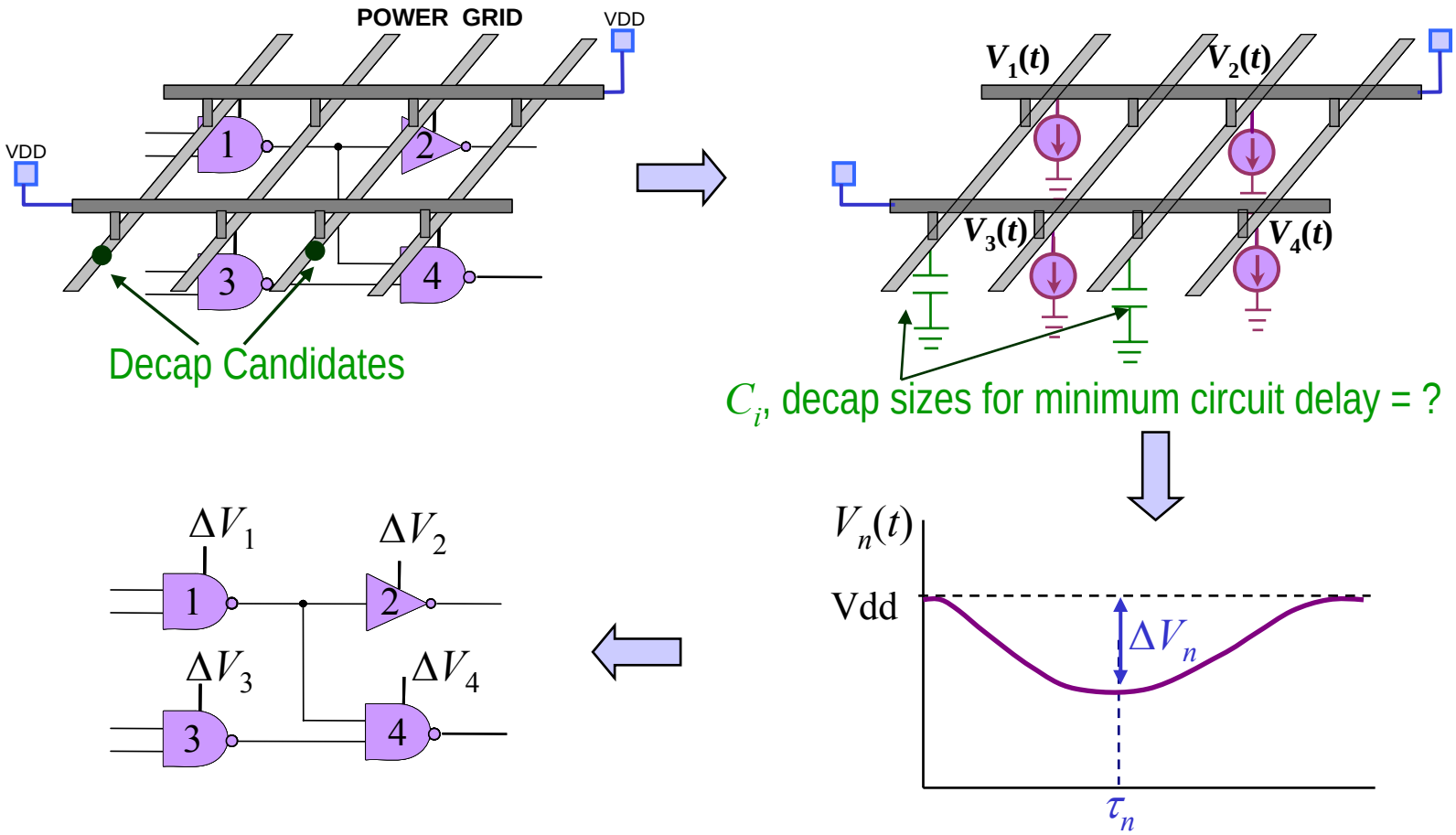


- Proposed approach
  - Allocate decaps in order to minimize circuit delay
  - Not focused on minimizing the drop at all the power grid nodes
  - Utilizes timing slacks for driving the decap allocation optimization problem

Outline

- Traditional Methods and Prior Work
- Proposed Approach
  - Primal Problem
  - Lagrangian Relaxation and Gradient Computation
  - Path-based greedy algorithm
- Experimental Results
- Conclusion

Problem Definition



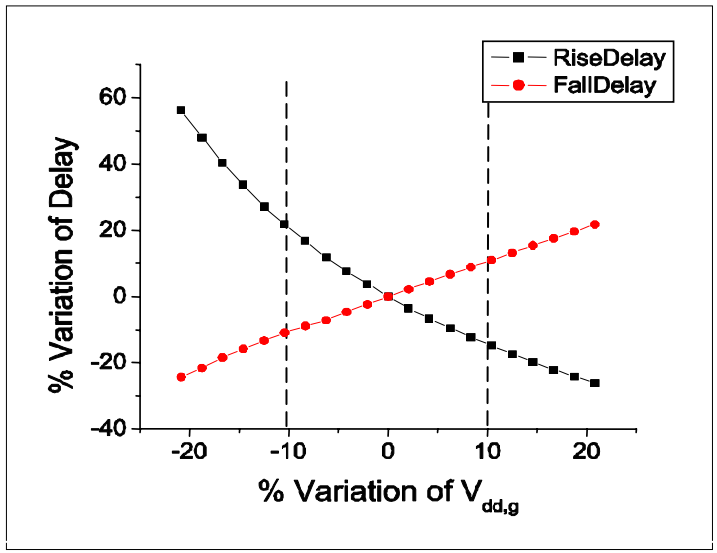
- Gate n assumed to be operating at supply voltage $(\mathrm{Vdd} - \Delta V_n)$
  - Conservative analysis which assumes all gates switching with local worst drops

Gate Delay Model

- Characterize delay of gate, i from its input j as a linear function of
  - Local supplies, Vdd_i and Vss_i (reduction in drive strength)
  - Input driver's supplies, Vdd_j and Vss_j (input signal swing)

$$D_{ji} = D_{ji}^{0} + k_{ji}\Delta Vdd_i + l_{ji}\Delta Vss_i + m_{ji}\Delta Vdd_j + n_{ji}\Delta Vss_j$$

$$tr_{ji} = tr_{ji}^{0} + p_{ji}\Delta Vdd_i + q_{ji}\Delta Vss_i + r_{ji}\Delta Vdd_j + s_{ji}\Delta Vss_j$$

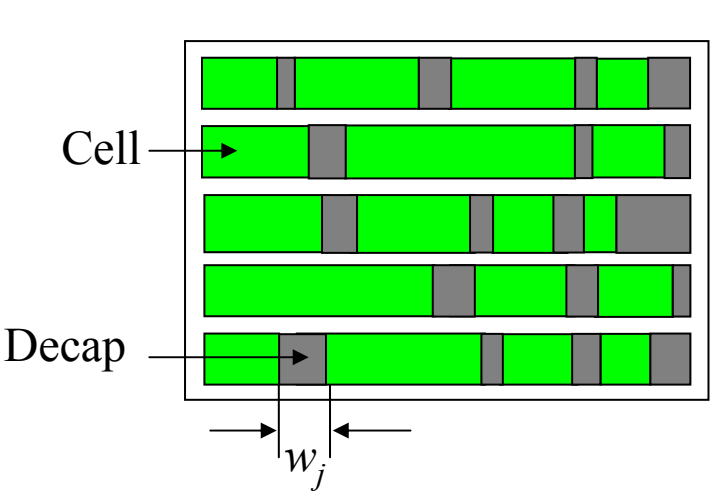


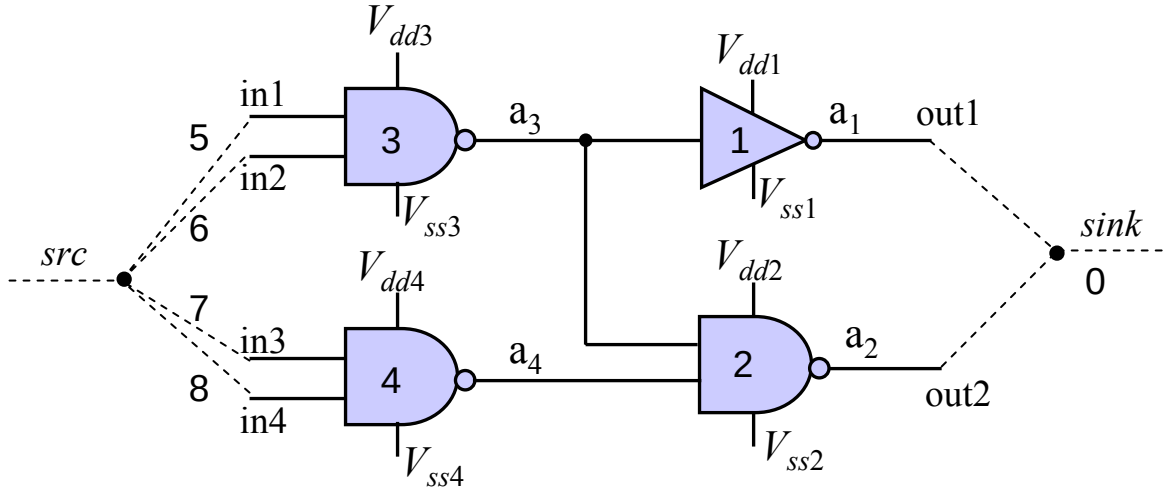
Library re-characterization

A load-slope based 7x7 table containing D_{ji}^{0}, k, l, m, n

Delay measured w.r.t. $50\%Vdd_{nominal}$ *point*

Primal Problem (PP)





Minimize ΣC_i

Subject to

(1) $a_j \leq T_0 \qquad \forall\, j = input(0)$

(2) $a_j + D_{ji} \leq a_i \qquad \forall\, j = input(i), \quad \forall\, gates\ i$

(3) $D_{ji} = D_{ji}^{\,0} + k_{ji}\Delta Vdd_i + l_{ji}\Delta Vss_i + m_{ji}\Delta Vdd_j + n_{ji}\Delta Vss_j$
$\forall\, i = input(j), \quad \forall\, gates\ j$

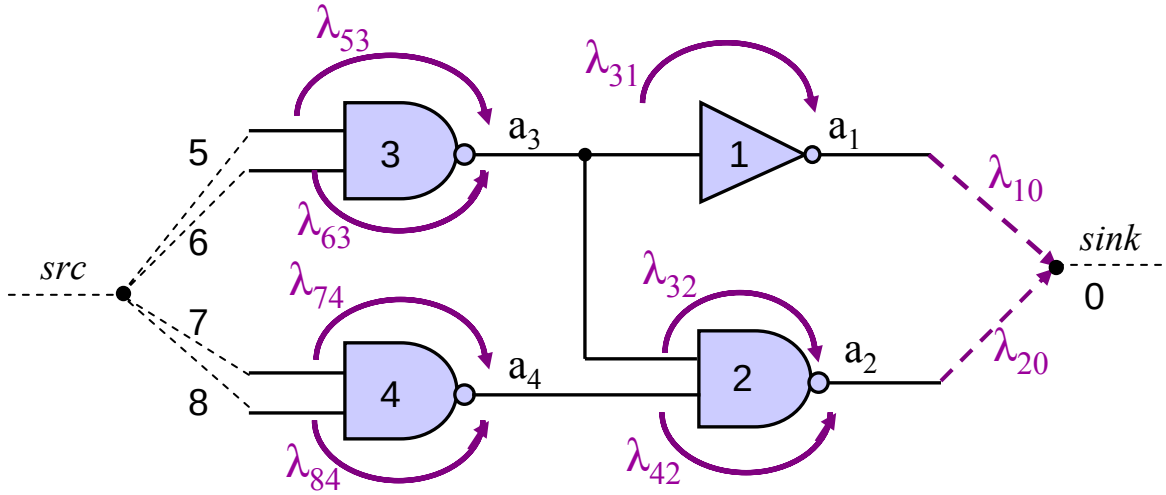
(4) $0 \leq C_i \leq w_{max}\,,\ i = 1..N_{decap}$

(5) *Voltage Supplies a fn of decap sizes* $\qquad Gx(t) + C\dot{x}(t) = u(t)$

Outline

- Traditional Methods and Prior Work
- Proposed Approach
  - Primal Problem
  - Lagrangian Relaxation and Gradient Computation
  - Path-based greedy algorithm
- Experimental Results
- Conclusion

Lagrangian Relaxation Problem (*LRP*)



Minimize $\Sigma\, C_i + \Sigma\, \lambda_{j0}(a_j - T_0) + \Sigma\,\Sigma\, \lambda_{ji}(a_j + D_{ji} - a_i)$

Subject to

(1) $D_{ji} = D_{ji}^{\,0} + k_{ji}\Delta Vdd_i + l_{ji}\Delta Vss_i + m_{ji}\Delta Vdd_j + n_{ji}\Delta Vss_j$
$\forall\ i = input(j), \quad \forall\ gates\ j$

(2) $0 \le C_i \le w_{max}\ ,\ i = 1..N_{decap}$

(3) *Voltage Supplies a fn of decap sizes* $\quad Gx(t) + C\dot{x}(t) = u(t)$

(4) $\lambda \ge 0$

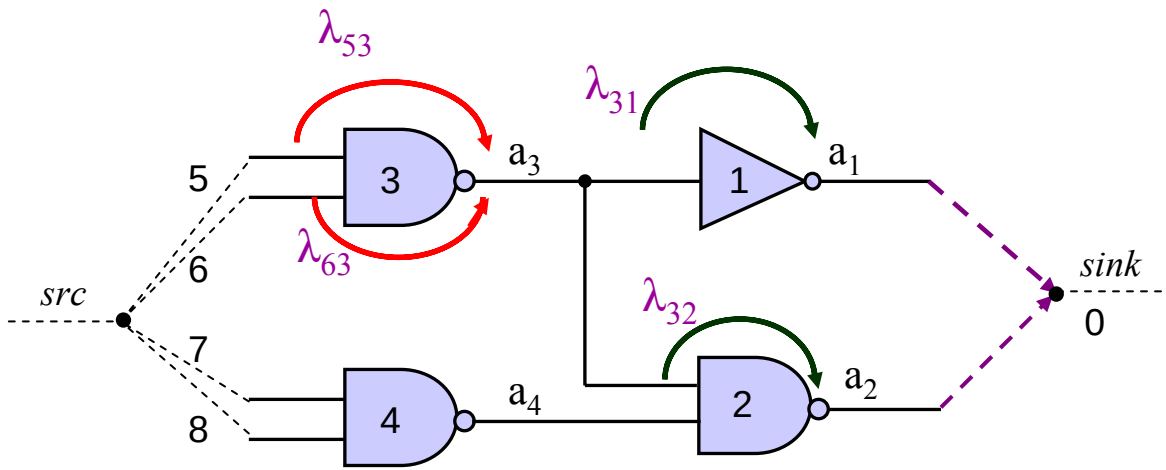
- λ_{ji} denotes the criticality of gate i from its input j

Kuhn Tucker Conditions

- If λ is optimal, sensitivity of objective fn wrt. arrival times = 0

$$\frac{\partial obj}{\partial a_i} = 0 \quad \forall \text{ gates } i$$

$$\Rightarrow \quad \sum_{k=output(i)} \lambda_{ik} = \sum_{j=input(i)} \lambda_{ji} \quad \forall \text{ gates } i$$



- Using KT conditions, obj. becomes independent of a_i for given set of λ

Minimize $\quad \Sigma C_i + \Sigma \Sigma \lambda_{ji} D_{ji} - \Sigma \lambda_{j0} T_0$

Solving the Lagrangian Relaxation Problem

- Using the delay model expression, objective function becomes a linear function of supply voltages

Minimize $\Sigma C_i + \Sigma \Sigma \lambda_{ji} D_{ji} - \Sigma \lambda_{j0} T_0$

$D_{ji} = D_{ji}^0 + k_{ji}\Delta Vdd_i + l_{ji}\Delta Vss_i + m_{ji}\Delta Vdd_j + n_{ji}\Delta Vss_j$

Minimize $\Sigma C_i + \Sigma \alpha_i \Delta Vdd_i + \Sigma \beta_i \Delta Vss_i - \Sigma \lambda_{j0} T_0$

Subject to (1) $0 \le C_i \le C_{max}$, $i = 1..N_{decap}$

(2) *Voltage Supplies a fn of decap sizes*

Where, $\alpha_i = \sum_{j=input(i)} \lambda_{ji} . k_{ji} + \sum_{k=output(i)} \lambda_{ik} . m_{ik}$ and $\beta_i = \sum_{j=input(i)} \lambda_{ji} . l_{ji} + \sum_{k=output(i)} \lambda_{ik} . n_{ik}$

- Needed: Gradients of the objective function wrt decap sizes

$$\frac{\partial}{\partial C}(C_i + \Sigma \alpha_i \Delta Vdd_i + \Sigma \beta_i \Delta Vss_i)$$

Gradient Computation

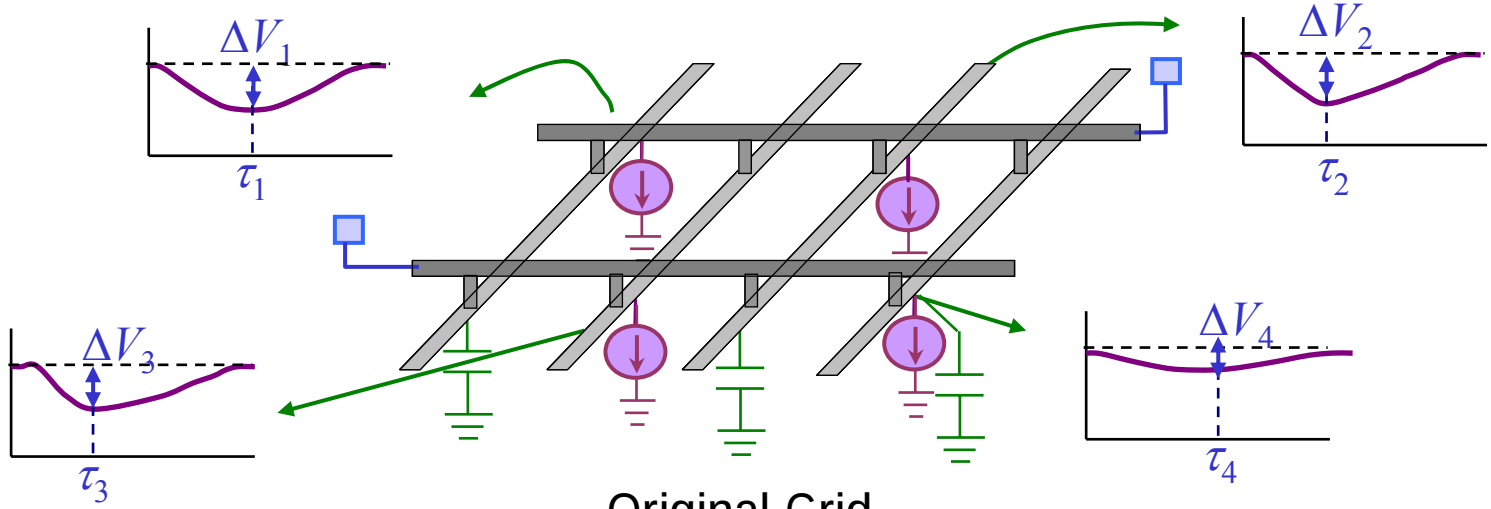
- Direct method: Single variable parameter, many measurement variables
  - Total # of simulations = # of nodes in the grid
- Adjoint method: Many variable parameters, single measurement objective
  - Total # of simulations = # of gates in the circuit
- Proposed Approach: Many variable parameters, many measurement variables
  - Modified adjoint sensitivity method
  - Measure derivative of voltage in the original circuit at all decap locations
  - Simulate adjoint circuit with multiple current excitations simultaneously applied

$$\frac{\partial Z}{\partial C} = \int_0^T \psi_C(T-t)\dot{v}_C(t)dt$$

- Sensitivity of objective function obtained in only one simulation of adjoint grid

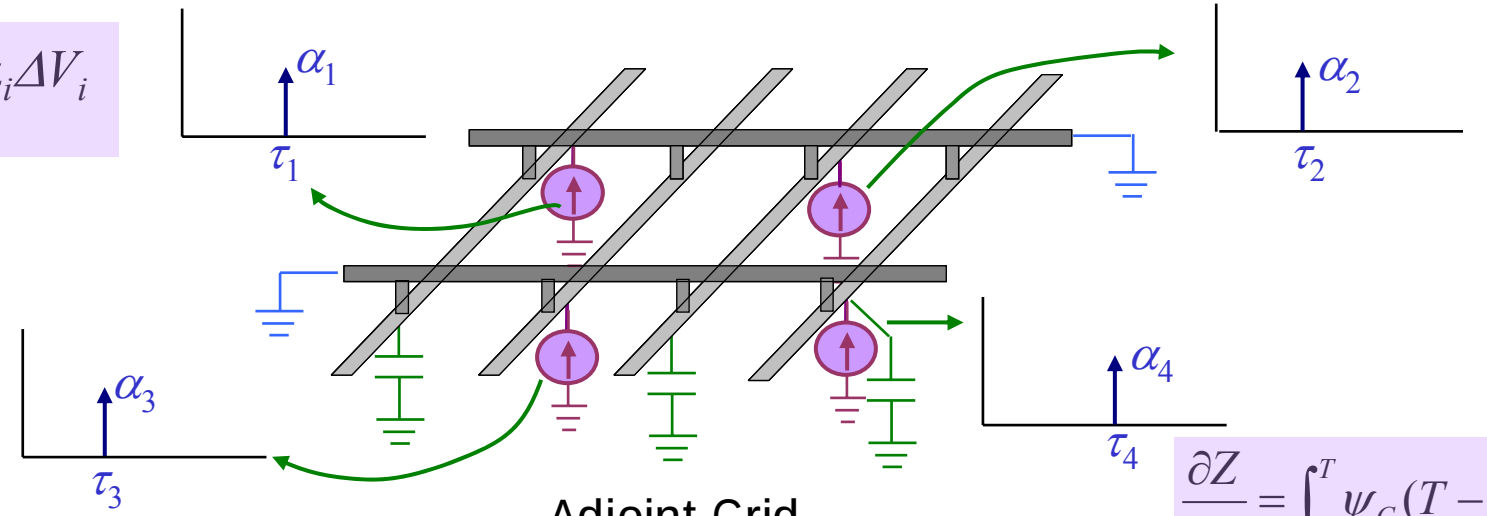
Conn, Visweswariah [SPECS, Jiffytune]

Gradient Computation Illustration



Original Grid

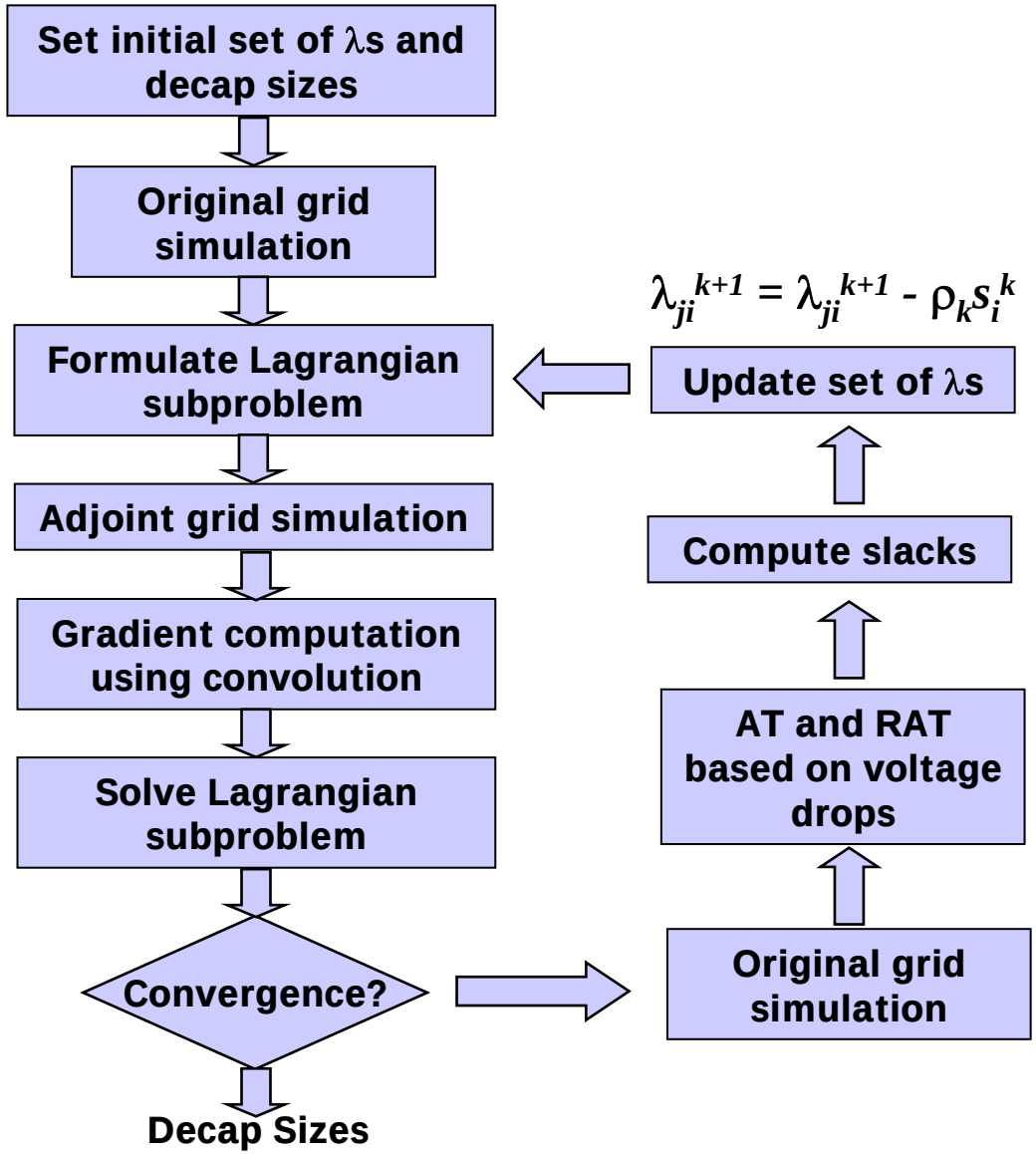
$$\frac{\partial \sum \alpha_i \Delta V_i}{\partial C}$$



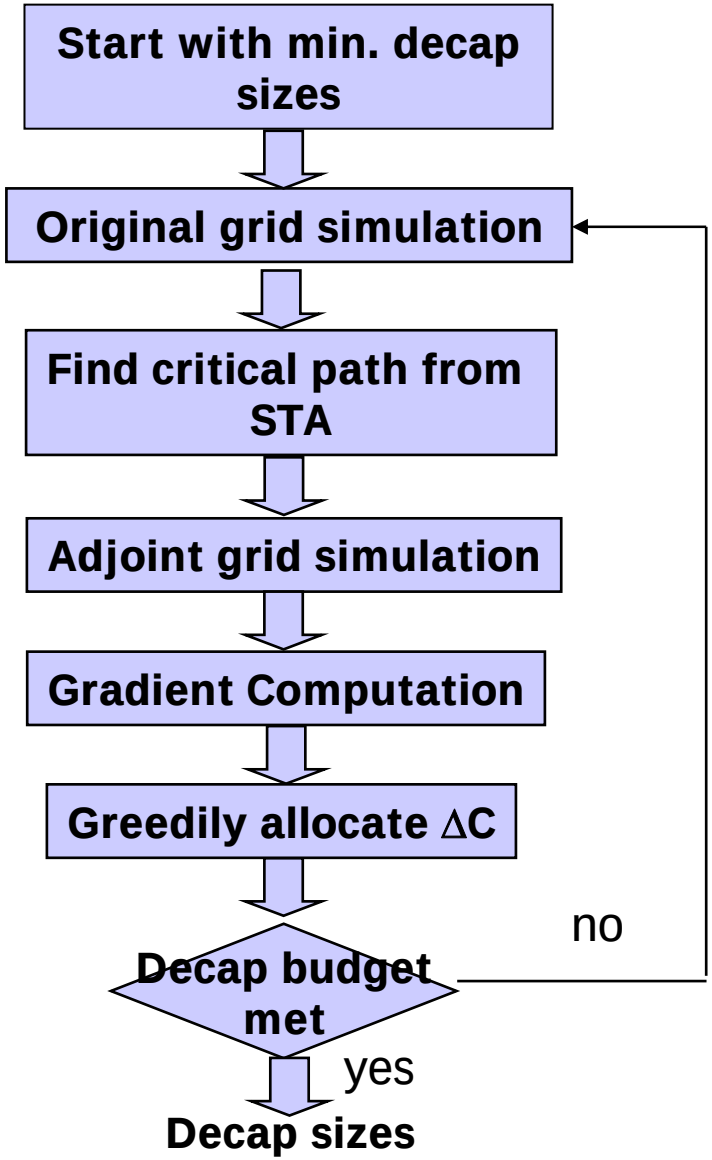
Adjoint Grid

$$\frac{\partial Z}{\partial C} = \int_0^T \psi_C(T-t)\dot{v}_C(t)dt$$

Overall Global Optimization Flow



Path Based Heuristic



Outline

- Traditional Methods and Prior Work
- Proposed Approach
  - Primal Problem
  - Lagrangian Relaxation and Gradient Computation
  - Path-based greedy algorithm
- Experimental Results
- Conclusion

Experimental Setup

- Current profiles - A triangular waveform applied at all the gates
- Gates placed through APR
- LANCELOT used for non-linear optimization
- C++ MNA solver used for original and adjoint grid simulation
- ISCAS85 benchmarks synthesized in 0.13μ library
- Power Grid description

Name	# Layers	Die area	# nodes	# elements	#C4s
Grid1	4	700μx700μ	10,804	17,468	12Vdd, 12Vss
Grid2	4	1.2mmx1.2mm	17,530	29,746	28Vdd, 28Vss

Experimental Results

- Iso-decap comparison with uniform decap distribution

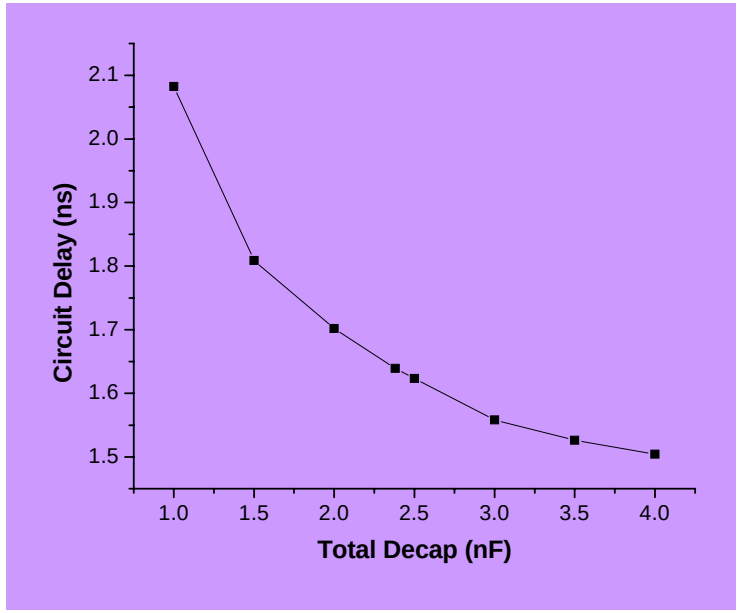
Grid	ckt	# gates	# decaps	decap budget	Circuit delay				% delay redn.		runtimes	
					Nom.	uniform	Global opt.	Greedy opt.	Global opt.	Greedy opt.	Global opt.	Greedy opt.
Grid1	c432	212	476	2.38nF	1.498ns	1.798ns	1.621ns	1.640ns	9.84%	8.79%	11m15s	1m15s
Grid1	c499	553	595	2.98nF	1.233ns	1.480ns	1.308ns	1.394ns	11.62%	5.81%	9m41s	1m57s
Grid1	c1355	654	793	3.97nF	1.839ns	2.207ns	1.878ns	1.913ns	14.90%	13.32%	11m43s	2m58s
Grid1	c1908	543	579	2.89nF	2.088ns	2.506ns	2.251ns	2.256ns	10.17%	9.98%	20m24s	25.83s
Grid1	c2670	1043	1190	3.57nF	1.622ns	1.946ns	1.754ns	1.764ns	9.86%	9.35%	52m33s	8m41s
Grid2	c3540	1492	1559	7.79nF	2.301ns	2.761ns	2.498ns	2.564ns	9.52%	7.31%	109m59s	23m49s
Grid2	c5315	2002	2217	6.65nF	2.080ns	2.769ns	2.409ns	2.416ns	13.00%	12.75%	221m24s	61m18s
Grid2	c6288	3595	3712	8.15nF	5.186ns	6.223ns	-	5.820ns	-	6.48%	>4hrs	188m36s
Grid2	c7552	2360	2571	7.18nF	2.975ns	3.571ns	-	3.262ns	-	8.65%	>4hrs	63m03s

Avg Delay Reduction: 10.11%

Experimental Results

- Iso-delay comparison with uniform decap distribution

Grid	Ckt	Nom. Delay	Delay constraint	Decap Allocated		% decap redn.
				Uniform	Optim.	
Grid1	c432	1.498ns	1.640ns	3.55nF	2.38nF	32.98%
Grid1	c499	1.233ns	1.394ns	3.49nF	2.98nF	17.60%
Grid1	c1355	1.839ns	1.913ns	6.65nF	3.97nF	40.33%
Grid1	c1908	2.088ns	2.256ns	6.15nF	2.89nF	52.92%
Grid2	c2670	1.622ns	1.764ns	6.96nF	3.57nF	95.80%
Grid2	c3540	2.301ns	2.564ns	10.04nF	7.80nF	22.37%
Grid2	c5315	2.080ns	2.416ns	12.20nF	6.65nF	45.56%
Grid2	c6288	5.186ns	5.820ns	9.74nF	8.15nF	16.31%
Grid2	c7552	2.978ns	3.262ns	13.26nF	7.18nF	45.85%



Avg Decap Reduction: 35.51%

Conclusions and Future Work

- Proposed a method for timing aware decap allocation
  - Efficient sensitivity computation of circuit delay to decap sizes
  - Utilizes timing slacks available in the circuit
  - Iso-decap comparison with uniform decap distribution demonstrates 10% improvement in circuit timing
  - Iso-delay comparison with uniform decap distribution demonstrates 35% reduction in total decap

- Future Work
  - Validation on larger circuits
  - Explore convergence and better optimization algorithms such as IPOPT
  - Exploit grid-locality for reducing run-time of grid simulations