Reliable Power Delivery for 3D ICs

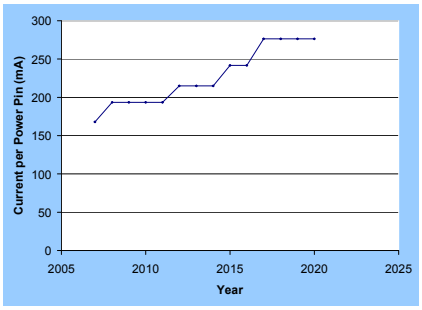
Pingqiang Zhou
Jie Gu
Pulkit Jain
Chris H. Kim
Sachin S. Sapatnekar

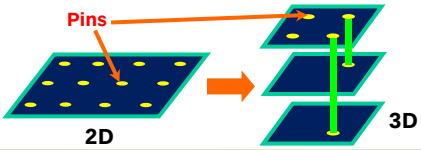
University of Minnesota

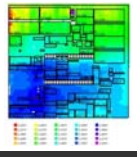
Power Supply Integrity in 3D

- Putting the power in is as important as getting the heat out
- Higher current density, faster current transients worsen supply noise
- Greater challenge in 3D due to via resistance, limited number of supply pins

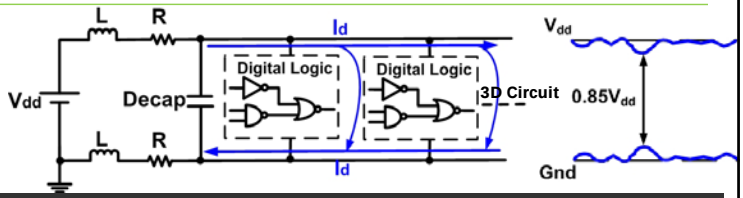
Current per power pin (2D) – ITRS







[IBM]



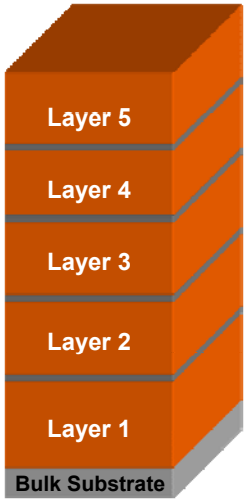
Thermal challenges

- Each layer generates heat
- Heat sink at the end(s)

- Simple analysis
  - Power(3D)/Power(2D) = m
    - m = # layers
  - Let R_{sink} = thermal resistance of heat sink
  - T = Power × R_{sink}
    - m times worse for 3D!

- And this does not account for
  - Increased effective R_{sink}
  - Leakage power effects, T-leakage feedback

- Thermal bottleneck: a major problem for 3D



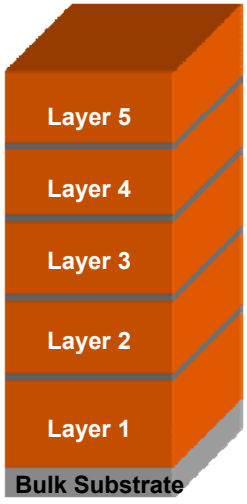
Power delivery challenges

- Each layer draws current from the power grid
- Power pins at the extreme end tier(s)

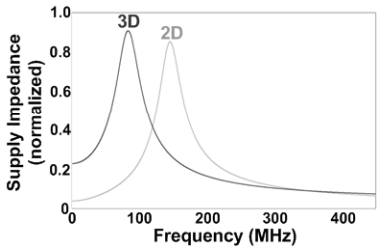
- Simple analysis
  - Current(3D)/Current(2D) = m
    - m = # layers
  - Let R_{grid} = resistance of power grid
  - V_{drop} = Current × R_{grid}
    - m times worse for 3D!

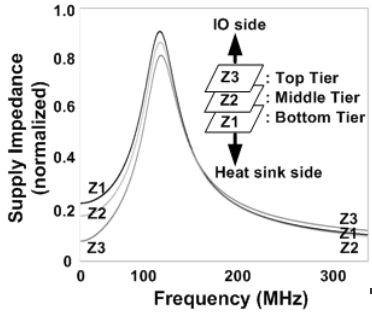
- And this does not account for
  - Increased effective R_{grid}
  - Leakage power effects, increased current due to T-leakage feedback

- Power bottleneck: a major problem for 3D



Power Supply Integrity Characteristics



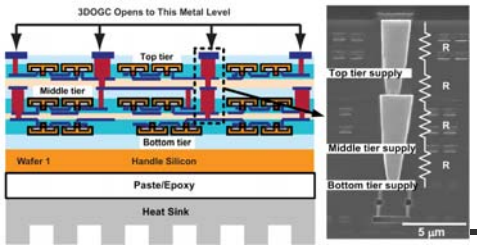


2D versus 3D

- Low frequency noise: 3D > 2D
- Mid frequency noise: 3D ≈ 2D
- High frequency noise: 3D ≈ 2D
- Resonant frequency: 2D > 3D

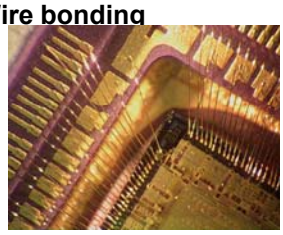
Tier versus tier

- Low/mid frequency noise: Z1 > Z2 > Z3
- High frequency noise: Z3 > Z1 > Z2
- Resonant frequency: Z1 ≈ Z2 ≈ Z3

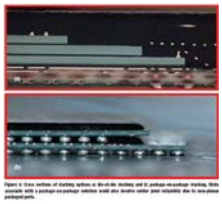


Packaging technologies

- **Wire bonding**

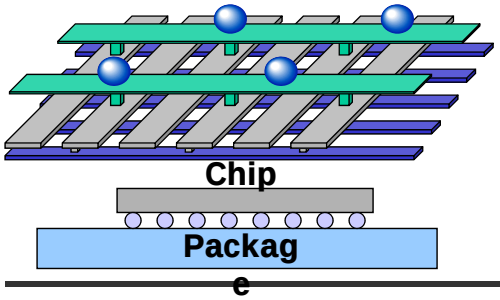


- **Stacked package/stacked die**

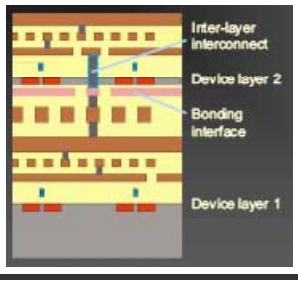


[Steidl, EDN]

- **Flip-chip**



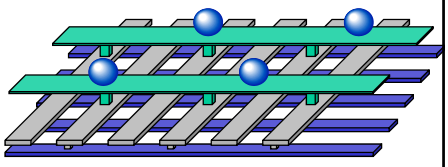
- **3D integration**



[Das *et al.*, ISPD04]

Traditional power delivery

- Requirements
  - V_{dd}, GND signals should be at correct levels (low V drop)
  - Electromigration constraints
    - Current density must never exceed a specification
    - For each wire, $I_i/w_i < J_{spec}$
  - dI/dt constraints
    - Need to manage dI/dt to reduce inductive effects
- Techniques for meeting constraints
  - Widening wires
  - Using appropriate topologies
  - Adding decoupling capacitances

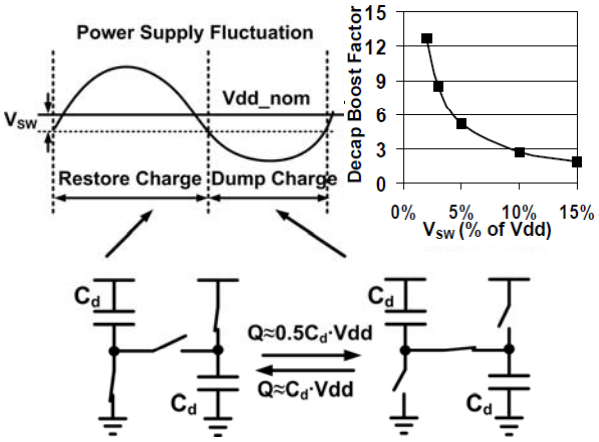


- Already challenged for 2D technologies
  - Reliable power delivery hard
  - Decaps get leaky
- Circuit + CAD approaches necessary

Outline of the talk

- Motivation
- Switched decaps
- Multistory Vdd
- CMOS+MIM decaps

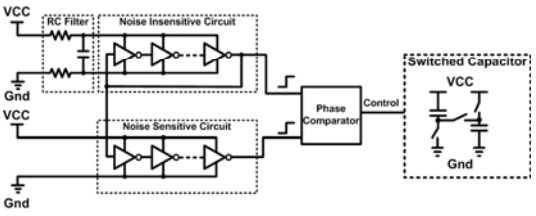
Active supply noise cancellation



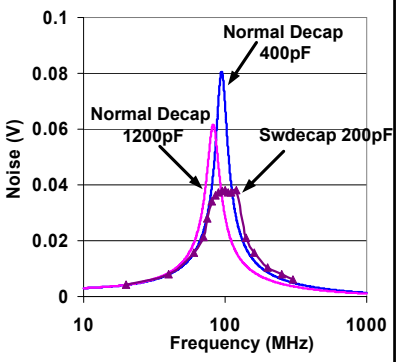
- Charge provided by switched decap (=0.5C·Vdd+CΔVdd/2) much larger than that of a conv. decap (=2C·ΔVdd)
- For a supply noise (ΔVdd) of 5%, effective decap value is boosted by 7.5X

9

Supply noise cancellation: Results



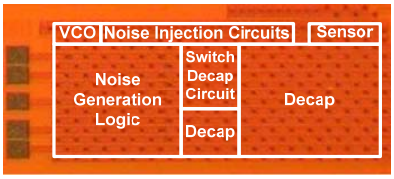
Digital Switched Decap Circuit



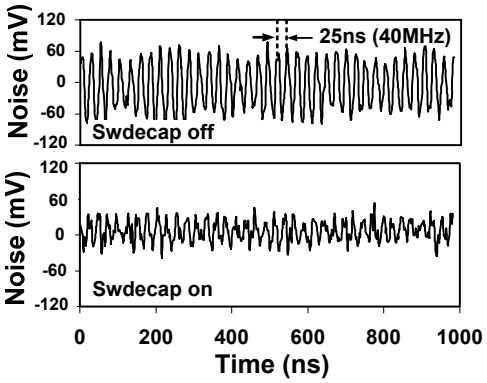
- 200pF switched decap has lower noise than 1200pF conventional decap
- 5−11X boost over passive decaps depending on supply noise magnitude

10

Proof of concept: Switched decap test chip



Technology	0.13µm CMOS
Quiescent Current	0.54mA
Regulation Freq.	10MHz-300MHz
Regulator Area (w/o decap)	100µmx70µm
Regulator Area (w/ 300pF decap)	190µmx220µm
Total Die Area	0.9mmx1.8mm



- **2.2-9.8dB reduction of the 40MHz resonant noise using 100-300pF switched decaps**

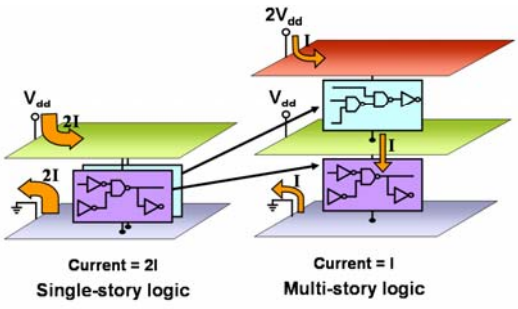
Comparison with passive damping

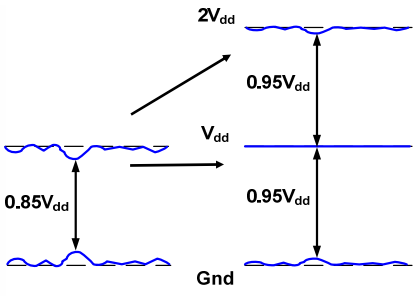
Swdecap Value	Resonant Suppression	Equivalent Passive Decap	Decap Boost
100pF	2.2dB	500pF	5X
200pF	5.5dB	1500pF	7.5X
300pF	9.8dB	3500pF	11X

Outline of the talk

- Motivation
- Switched decaps
- Multistory Vdd
- CMOS+MIM decaps

Multi-story power supply



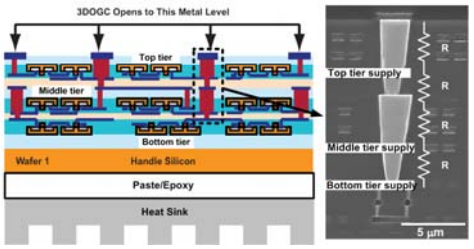


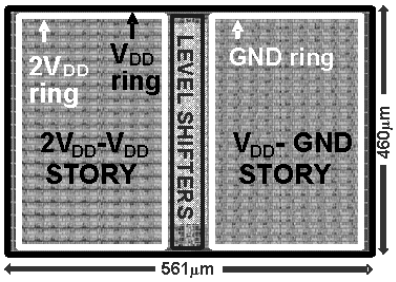
	1-story	2-story
Current	2I	I
Voltage	Vdd	2Vdd
Power	2Vdd·I	2Vdd·I–Δ
Noise	15%Vdd	< 8%Vdd

Improved supply noise due to:
- Reduced current magnitude
- Cleaner middle supply voltage

Attractive for 3D chips:
- Isolated substrate for each tier
- Chip is naturally partitioned

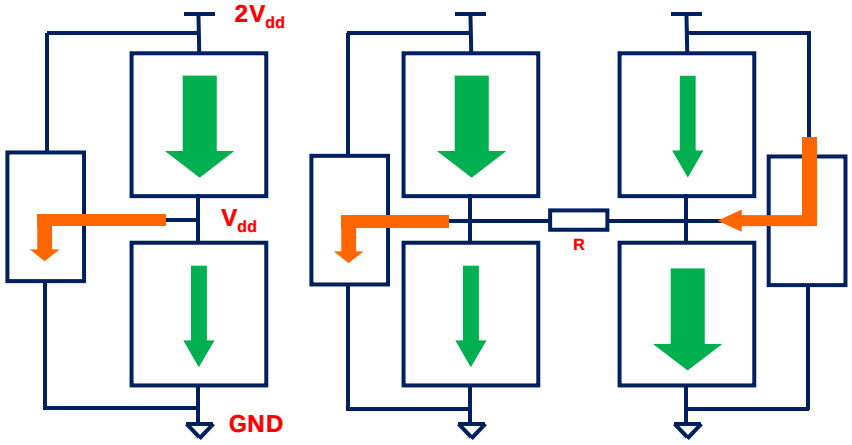
Multi-story power supply: Test layout





- A test layout in MITLL's SOI process shows a 5.3% area overhead

CAD solutions for multi-story circuits



Overall Design Flow

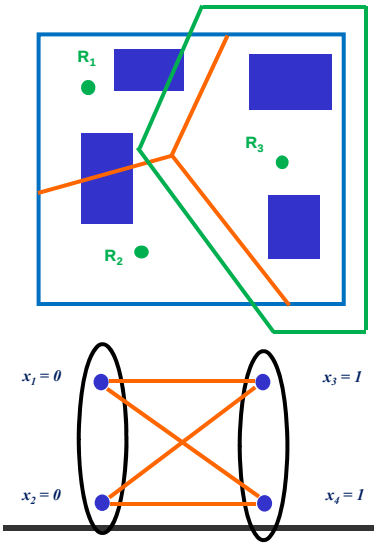
Netlist and block information

Floorplanning involving regular modules and regulators

Assigning modules using a graph partition-based algorithm

Module assignment

Estimating the wasted power

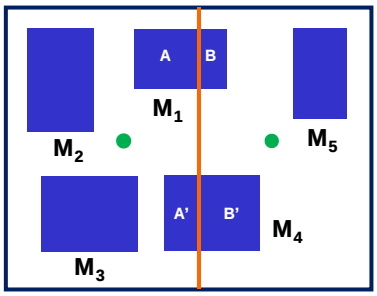


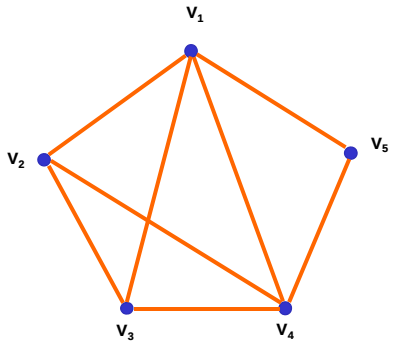
$= 0$ if $x_i = x_j$

$= 1$ if $x_i \neq x_j$

Graph partitioning problem!

Constructing the graph





$$w(V_i, V_j) = \left(\sum_{k=1}^{K} \frac{S_{ik} S_{jk}}{S_i S_j}\right) \overline{I_i(t) I_j(t)}$$

3D benchmarks

- Exercised on GSRC floorplanning benchmarks
- Largest floorplan has 300 modules
- Comparison with (slow)simulated annealing method

Layer	$\frac{\text{WastedPower}}{\text{UsefulPower}}$ (%)		Maximum IR Noise (mV)		Runtime (sec)	
	Partition-Based	Annealing	Partition-Based	Annealing	Partition-Based	Annealing
n100Layer0	3.3	3.1	52.8	62.0	0.03	80
n100Layer1	3.1	3.8	28.9	42.5	0.02	80
n100Layer2	3.7	5.7	45.4	54.6	0.02	80
n200Layer0	8.7	6.4	55.2	88.4	0.31	157
n200Layer1	5.6	6.4	62.1	64.4	0.16	160
n200Layer2	5.6	7.1	77.4	52.7	0.18	165
n300Layer0	4.7	4.5	61.1	56.0	1.83	235
n300Layer1	6.3	6.3	33.4	36.8	0.69	236
n300Layer2	5.4	4.6	46.5	39.5	0.77	236

Runtime Comparison: > 10^3 x speedup over SA

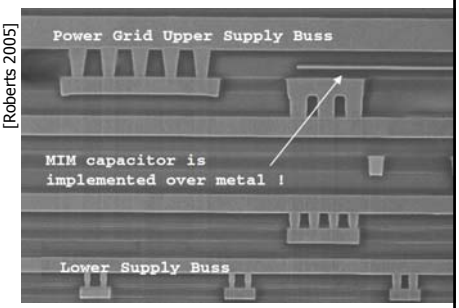
Outline of the talk

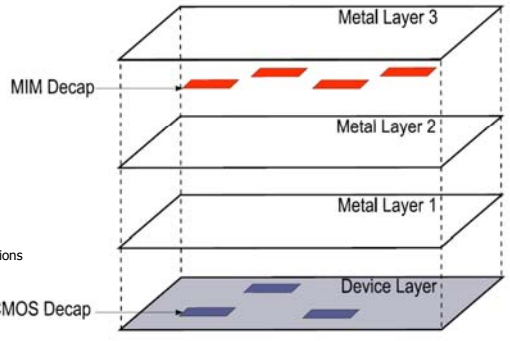
- Motivation
- Switched decaps
- Multistory Vdd
- CMOS+MIM decaps

MIM decaps

- Capacitance density*
  - CMOS – 17.3 fF/μm^2 at 90nm
  - MIM – 8.0 fF/μm^2
- Leakage density*
  - CMOS – 1.45e-4 A/ cm^2
  - MIM – 3.2e-8 A/cm^2
- Congestion
  - MIM – routing blockage

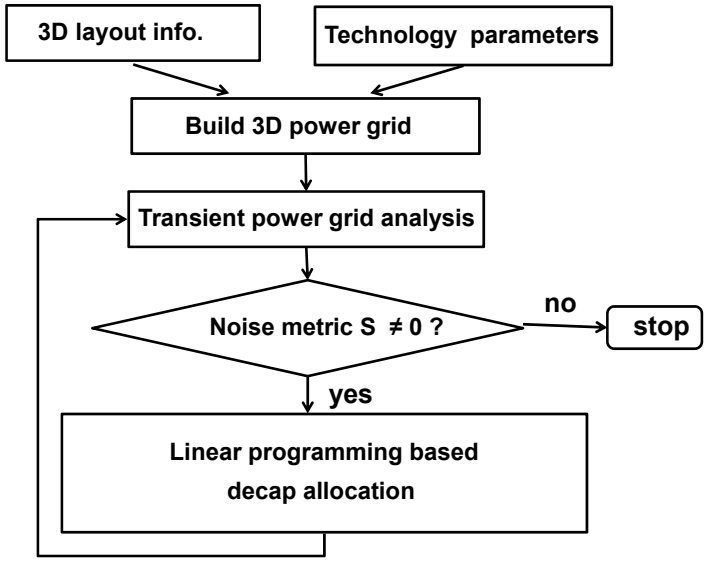
[Roberts 2005]





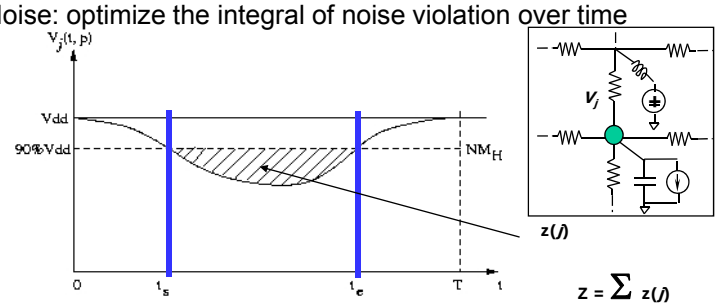
* Numbers deduced from Roberts et al., IEDM05 and PTM simulations

Overall flow of the algorithm



Metrics

- Noise: optimize the integral of noise violation over time



Waveform of node *j* on VDD grid

- Linearized congetion metric

$$\Delta Cong_k = \sum_{i \in R_k} (\lambda_i \cdot \Delta y_i)$$

  - R_k is the set of grid cells adjacent to grid k
  - λ_i reflects the effect on the congestion of grid k after inserting a small MIM decap Δy_i in grid i.

Sequence of linear programs: formulation

- Objective

$$\min \quad \alpha\Delta S + (1-\alpha)\,\Delta P$$

  - $\Delta S = \sum_k (a_k \Delta x_k + b_k \Delta y_k)$ = change of violation area
  - $\Delta P = \sum_k (c_k \Delta x_k + d_k \Delta y_k)$ = change in leakage
  - Δx_k : Newly added CMOS decap to grid k
  - Δy_k : Newly added MIM decap to grid k
- Constraints
  - Congestion constraint

$$\Delta Cong_k \le \gamma \cdot Cong_k$$

  - Decap resource constraint

$$0 \le \Delta x_k \le \min\{\Delta_{CMOS}, C^k_{CMOS}\}$$
$$0 \le \Delta y_k \le \min\{\Delta_{MIM}, C^k_{MIM}\}$$

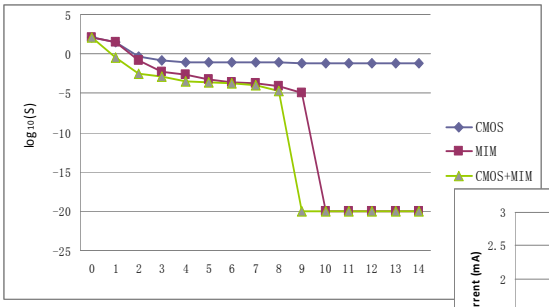
Experimental results

Ckt	# Nodes	Worst V droop (V)	# nodes with noise violations	Violation Area S (V ns)
Ibm123	18,634	0.135	3330	13.739
Ibm05	12,026	0.122	1359	72.260
ibm08	17,030	0.125	3191	41.305
ibm10	29,262	0.159	5935	91.286
ibm18	75,042	0.163	6392	108.649

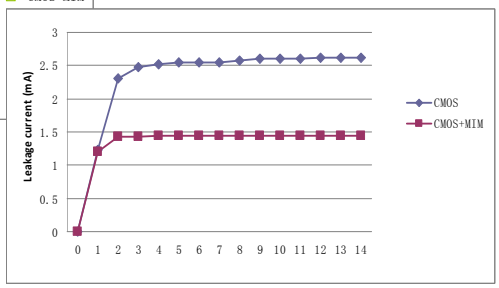
Ckt	CMOS only						MIM only					CMOS + MIM					
	VNs	S (V-ns)	Lkg (mA)	Decap (pF)	#Iter	Time (s)	maxC (%)	avgC (%)	Decap (pF)	#Iter	Time (s)	Lkg (mA)	maxC (%)	avgC (%)	Decap (pF)	#Iter	Time (s)
ibm123	368	0.023	2.1	564	25	130	15.8	3.9	607	7	59	1.1	8.4	1.7	628	4	43
ibm05	24	0.049	2.7	480	5	24	19.7	1.7	550	23	111	2.1	0.0	1.2	546	22	109
ibm08	31	0.010	1.2	313	16	82	30.6	1.5	768	24	134	0.6	0.0	0.9	774	20	116
ibm10	351	0.182	1.6	417	12	108	10.6	5.9	511	11	186	0.9	4.5	2.5	520	4	133
ibm18	130	0.071	2.7	698	14	400	39.5	5.3	812	9	339	1.4	7.0	3.6	826	8	307

Experimental results: ibm18

- **Violation Area**



- **Leakage**



Conclusion

- Power delivery into a 3D chip is a critical problem for next-generation designs
- Incremental solutions will only take us so far
  - Already stretched even for 2D designs
- Need innovative design + CAD solutions