Breaking the 3D IC power delivery walls using voltage stacking

IEEE CAS DL program
Mircea Stan mircea@virginia.edu
HPLP lab. http://hplp.ece.virginia.edu
ECE Dept., University of Virginia

May 2013

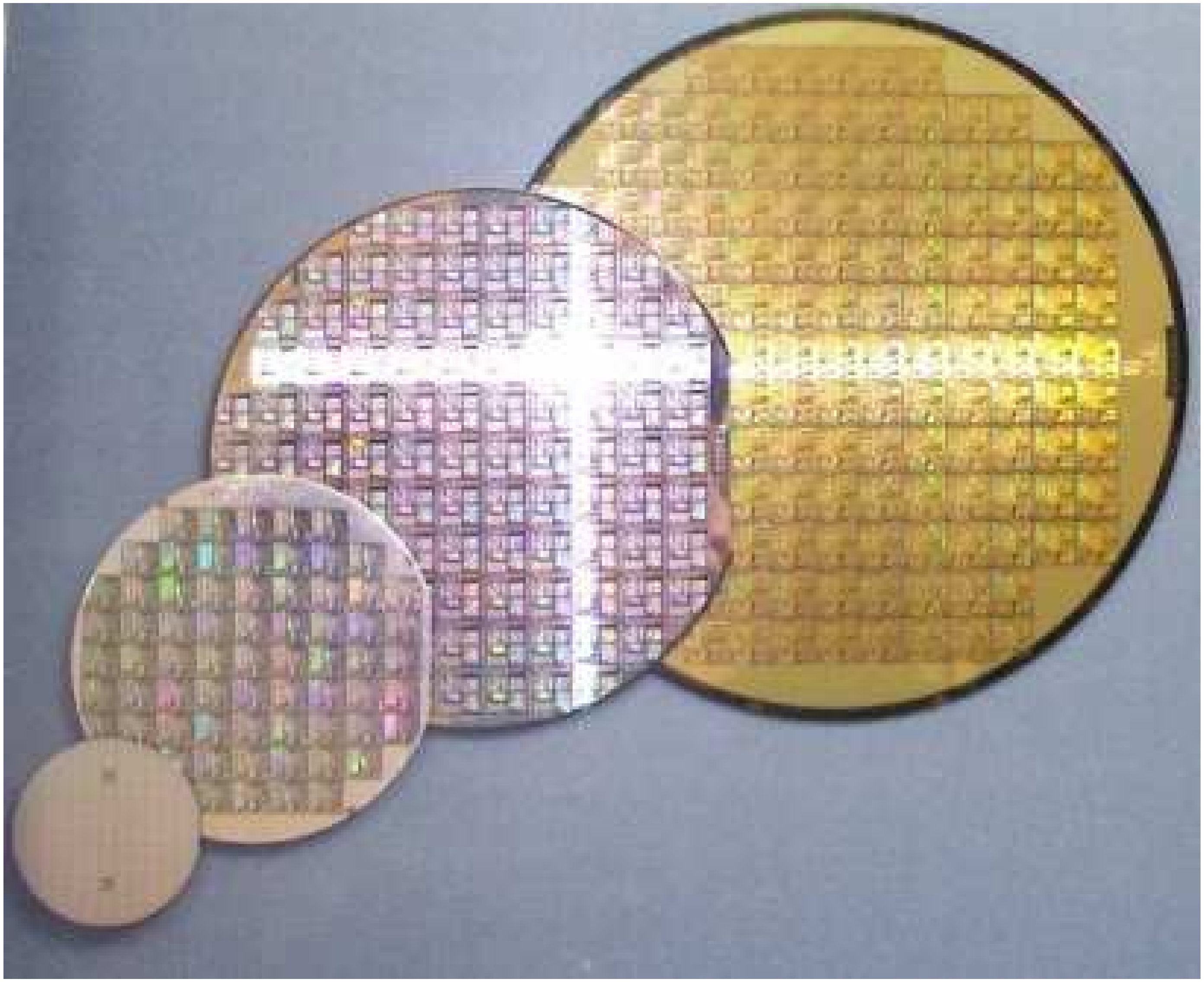
University of Virginia
School of Engineering and Applied Science

hplp
High-Performance Low-Power

Outline

- Why, What, How?
- 3D IC
- Power Walls
- Voltage Stacking
- Voltage Regulation for Stacking
- Voltage stacking in 3D IC

Integrated circuits – 2D

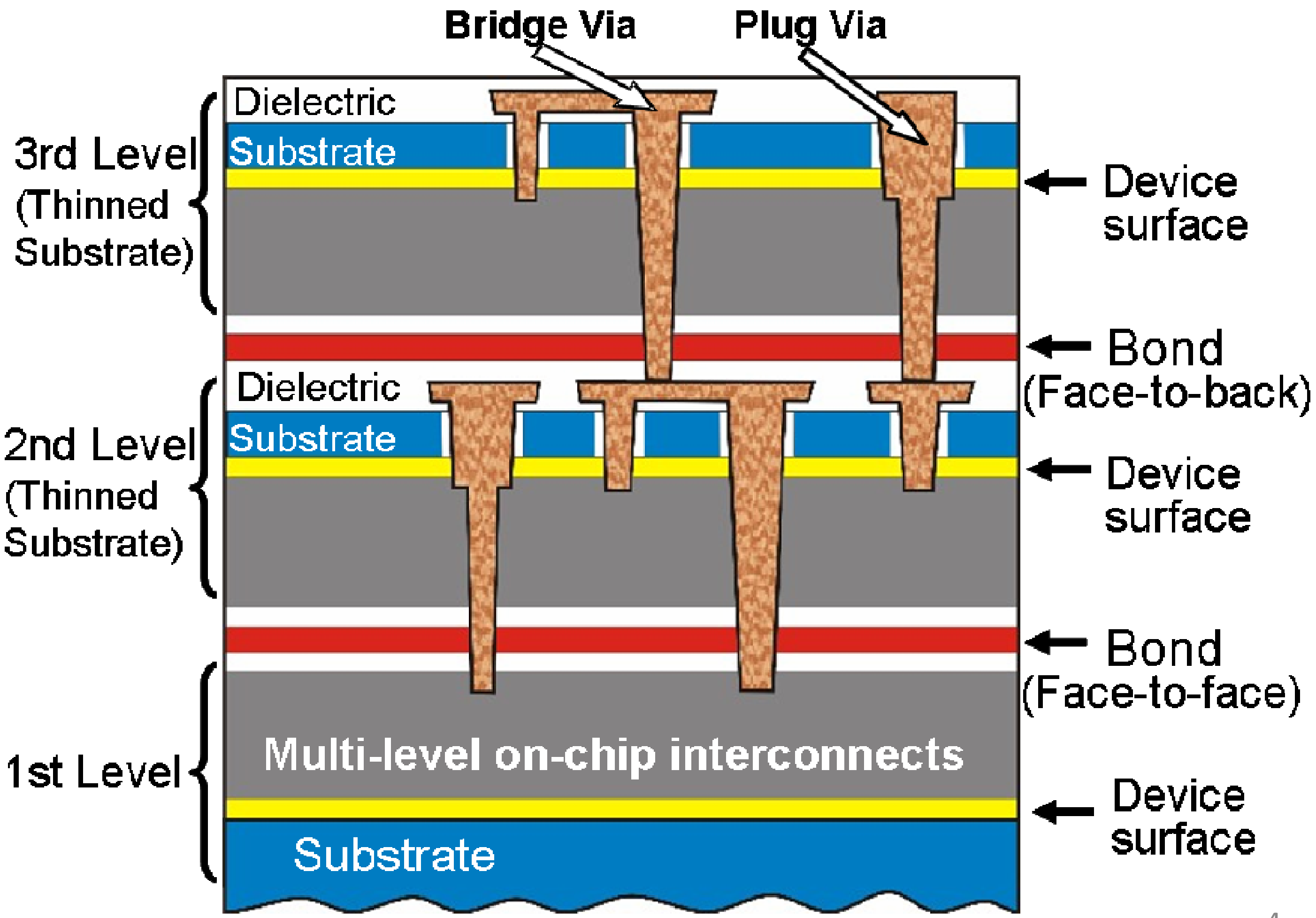


Si wafers

Source: Wikipedia

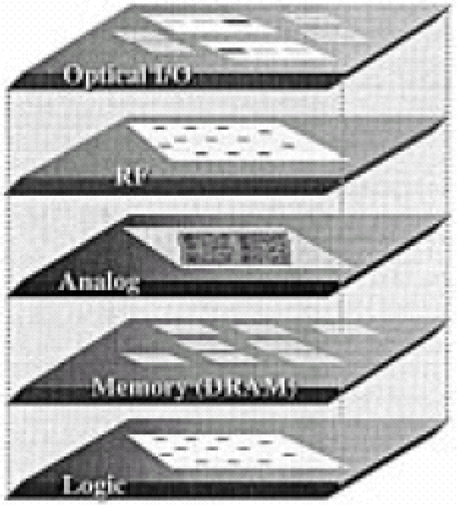
What is 3D IC?

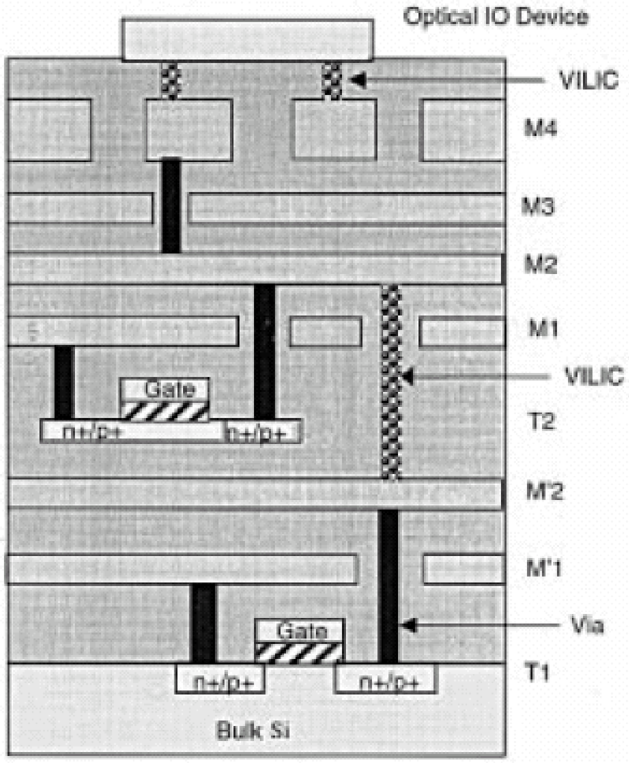
Multiple *active* layers - continue Moore's law



Source: J.-Q. Lu et al., 2002 IITC, IEEE, 2002, pp. 78-80

Why 3D IC? – Moore's law!



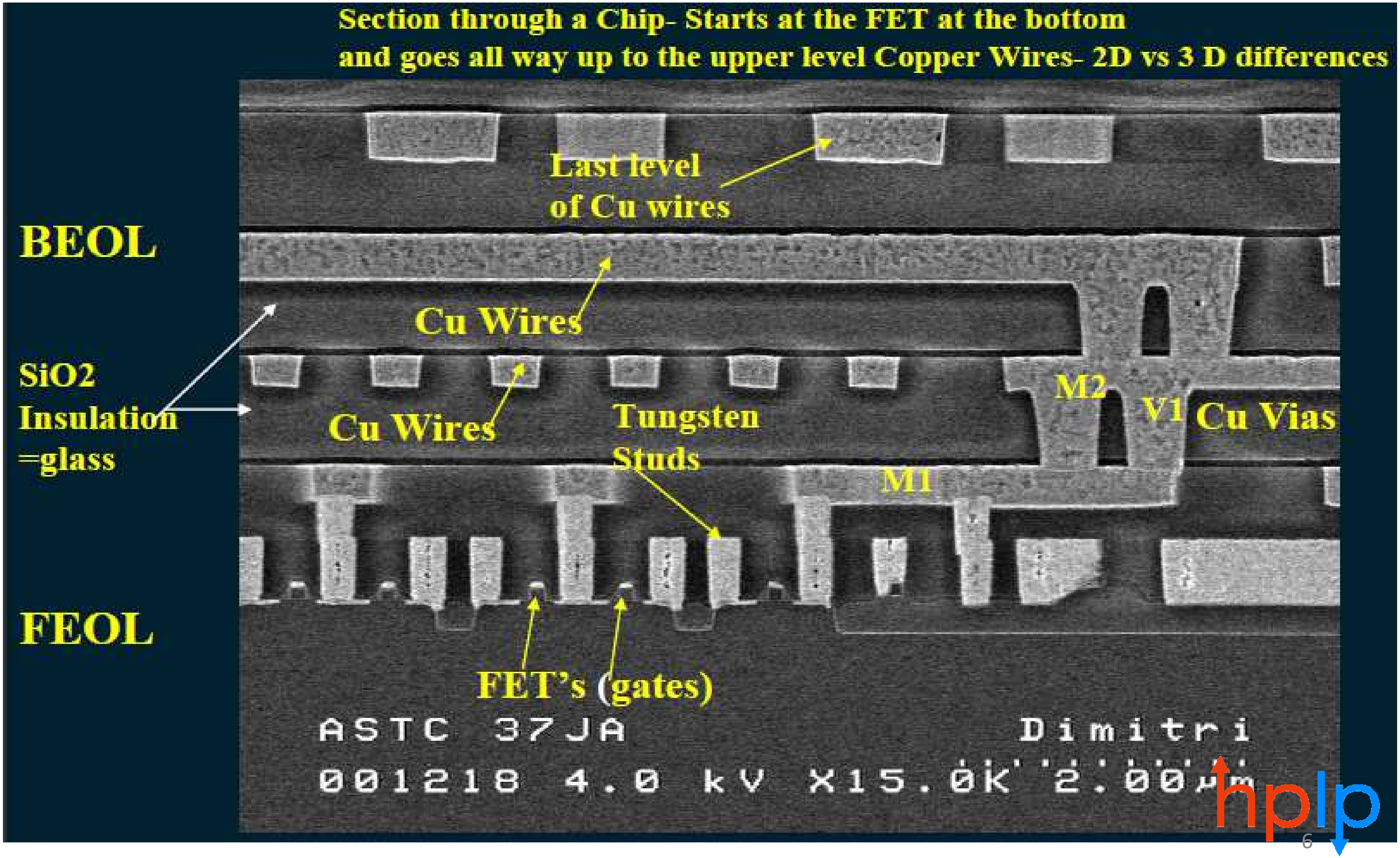


Could be Heterogeneous…

"Stacked" 2D (Conventional) ICs

Pouya Dormiani, Christopher Lucas ,"3D IC Technology"

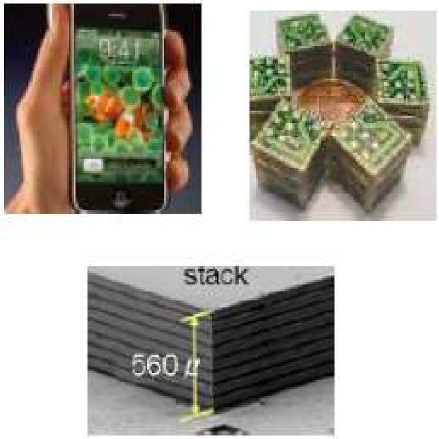
CMOS process cross section



Section through a Chip- Starts at the FET at the bottom and goes all way up to the upper level Copper Wires- 2D vs 3 D differences

Source: E. Levine – IC Fabrication and Yield Control

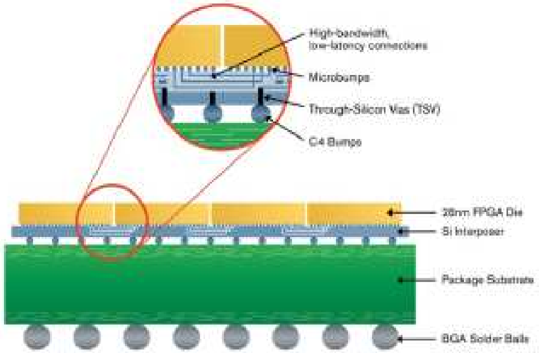
Benefits of 3D IC



Density

high capacity

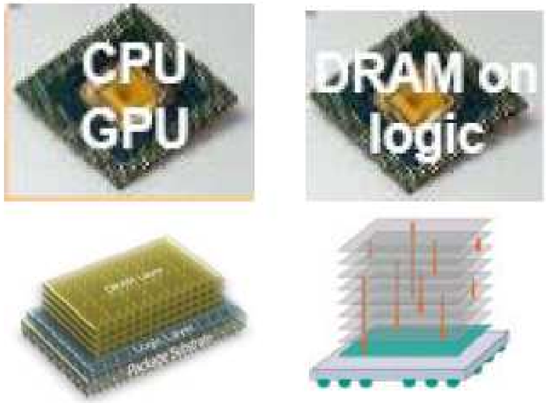
Small footprint



Cost

yield/cost improvement

For large die and new technology



Performance/Energy

fast interconnect (latency)

High bandwidth

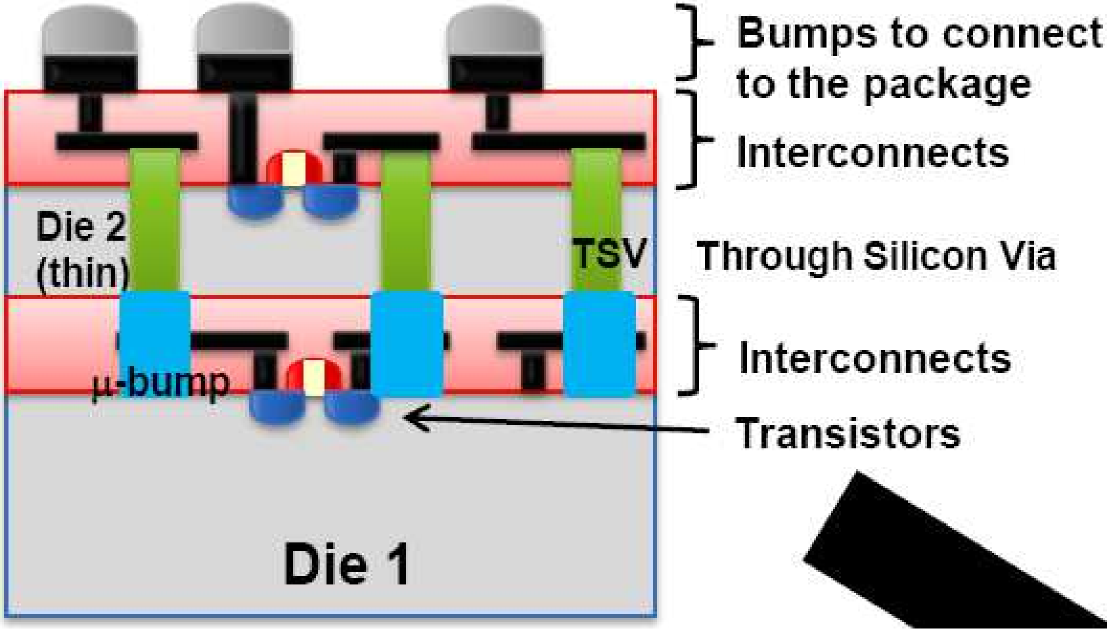
"More than Moore"

Heterogeneous integration with Logic+memory+RF+optical etc.

Yuan Xie, " Cost/architecture/application Implication for 3D Stacking Technology

3D IC “Power Walls”

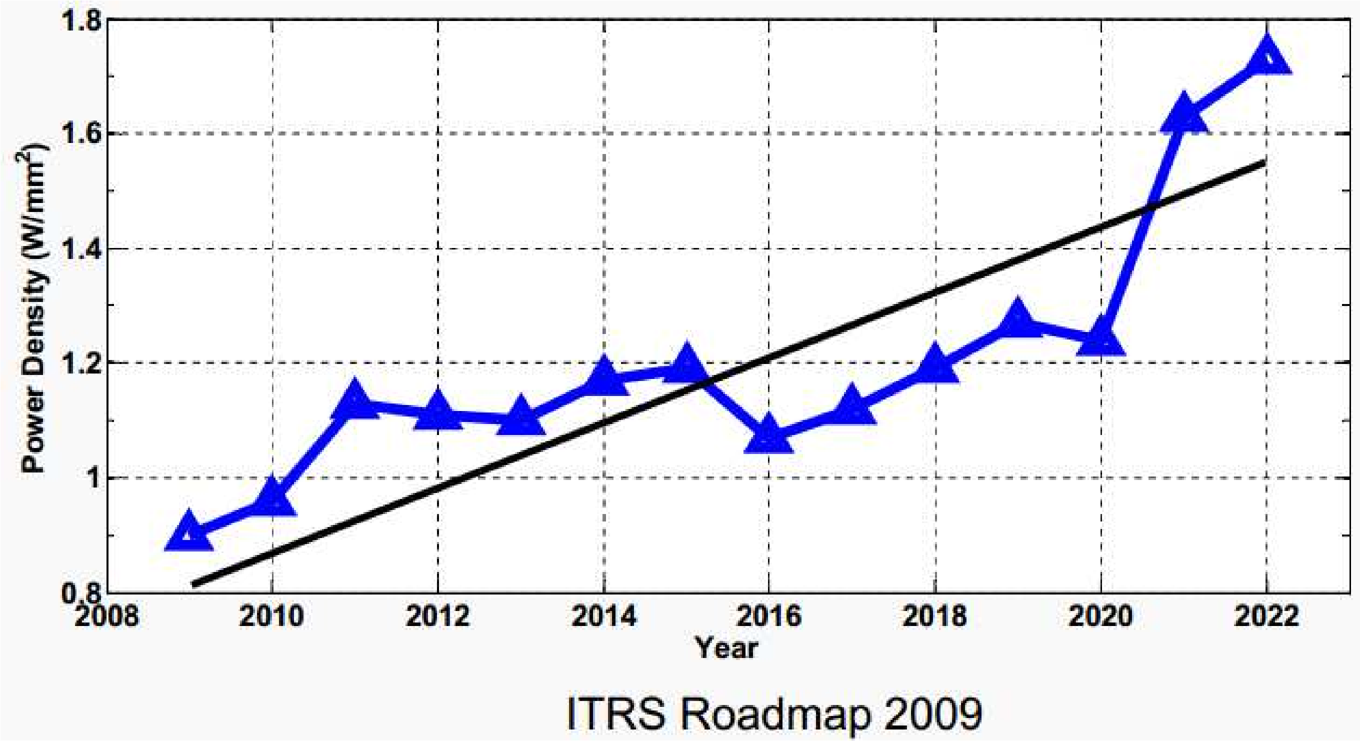
- Physical stacking in 3rd dimension exacerbates the two-dimensional power density explosion
- k-layered 3D IC : k-times supply current, Lower power pads
- TSV (Through silicon-vias) : Adds resistance to the PDN impedance
- TSV Area Over-head



Source :Synopsis ,Sematech Symposium 2009

Power Density "Wall"

- Technology scaling => **Increased** Power Density
- Physical stacking in 3rd dimension exacerbates the two-dimensional power density explosion
- Overhauling **Heat Dissipation** Capacity

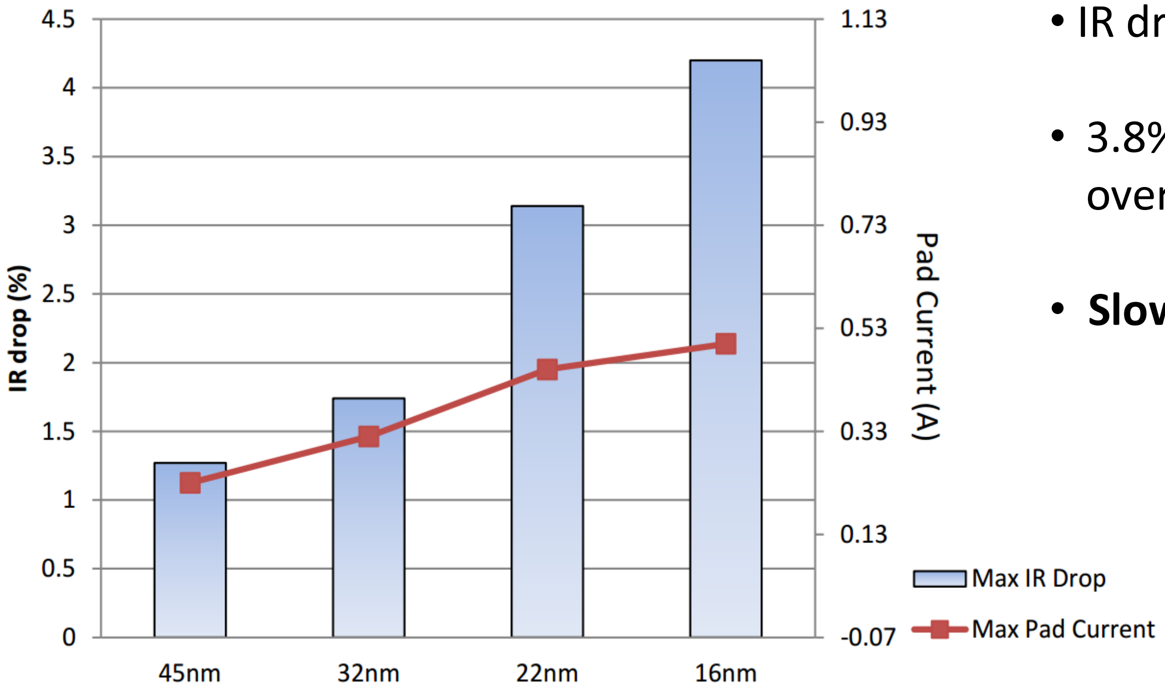


ITRS Roadmap 2009

Zhiyu Zeng et al, " Tradeoff analysis and optimization of power delivery networks with on chip voltage regulation", DAC, 2010.

Power Noise “Wall”

- EM effect, IR drop, Ldi/dt : ↑ with increasing current density
- Voltage Scaling : Noise margin ↓
- Increased current demand => **Lower PDN impedance** needed



- IR drop triples from 45nm to 16nm
- 3.8% increase in IR drop -> 51% delay overhead
- **Slow scaling** of PDN Impedance

Source: Runjhie Zhang et al, “Some Limits of Power Delivery in the Multicore Era”

hplp
High-Performance Low-Power

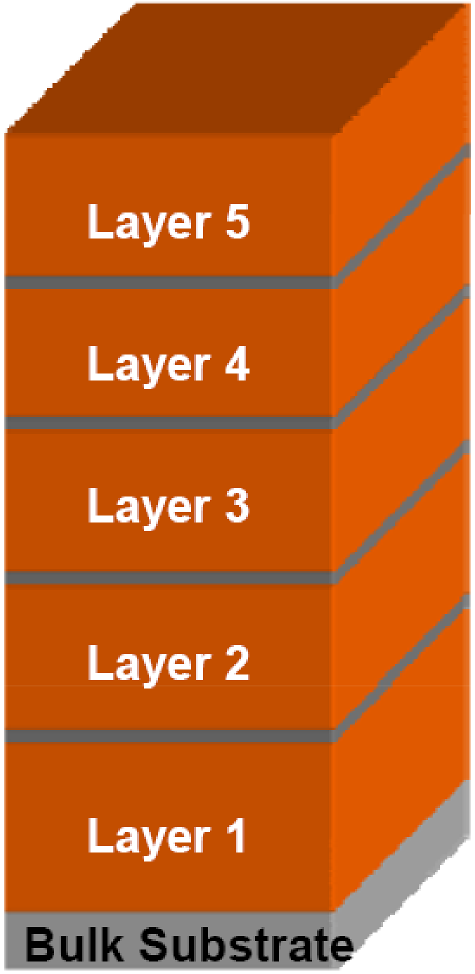
Thermal 3D IC bottleneck

- Power Pins at one end of the tiers, heat sink the other end

- Current starved components placed near heat sink, farthest from the power pins

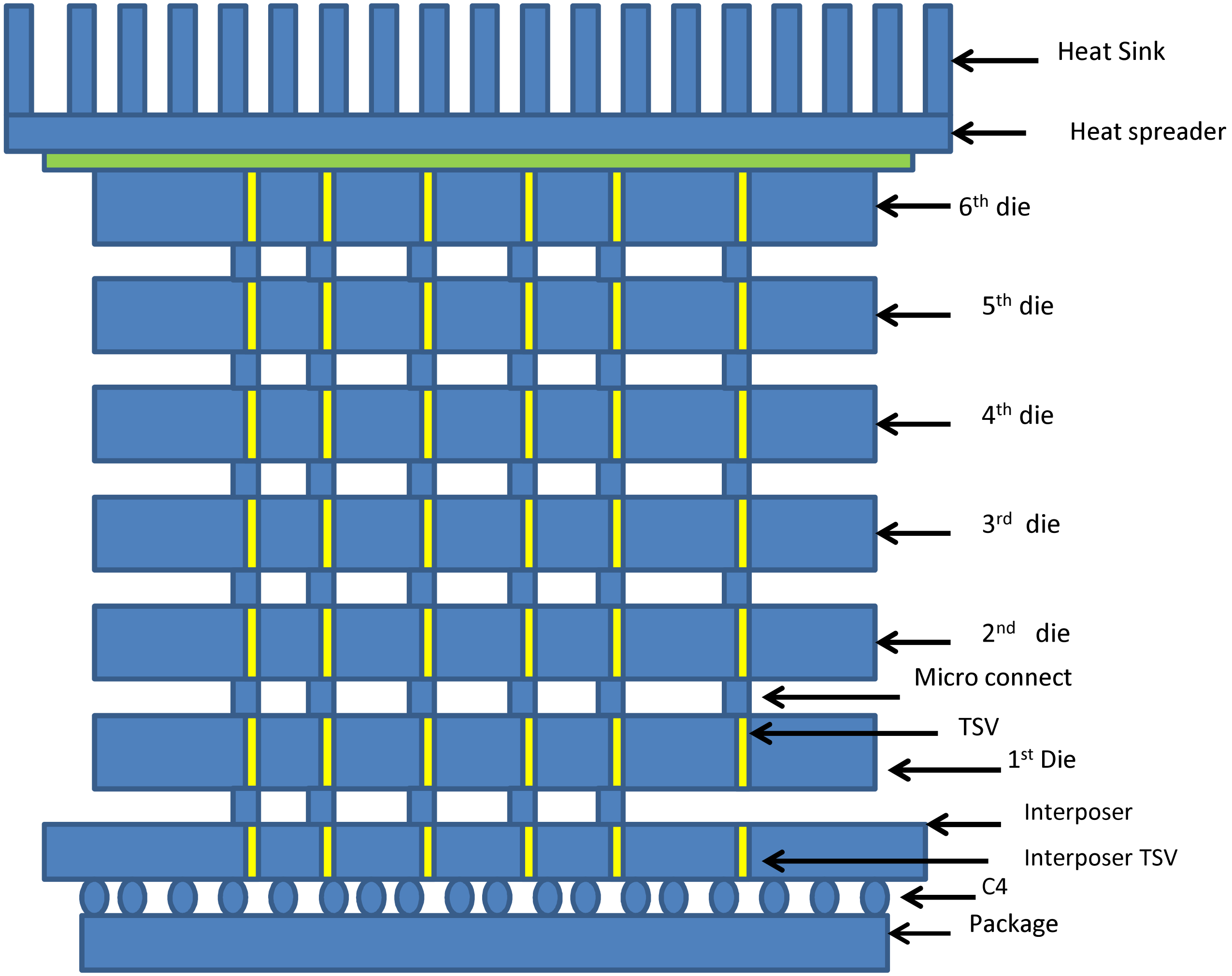
- Current (3D)/Current(2D) = n, n number of layers
  Let R_{grid} = resistance of power grid
  $V_{drop\text{-}3D}$ = n*Rgrid*Current(2D)

- **Current starved layers getting lower voltage headroom**



High-Performance Low-Power

3D IC Power Delivery “Wall”



Power ~ O(Vol)

Vdd = ct.

I ~ O(Vol)

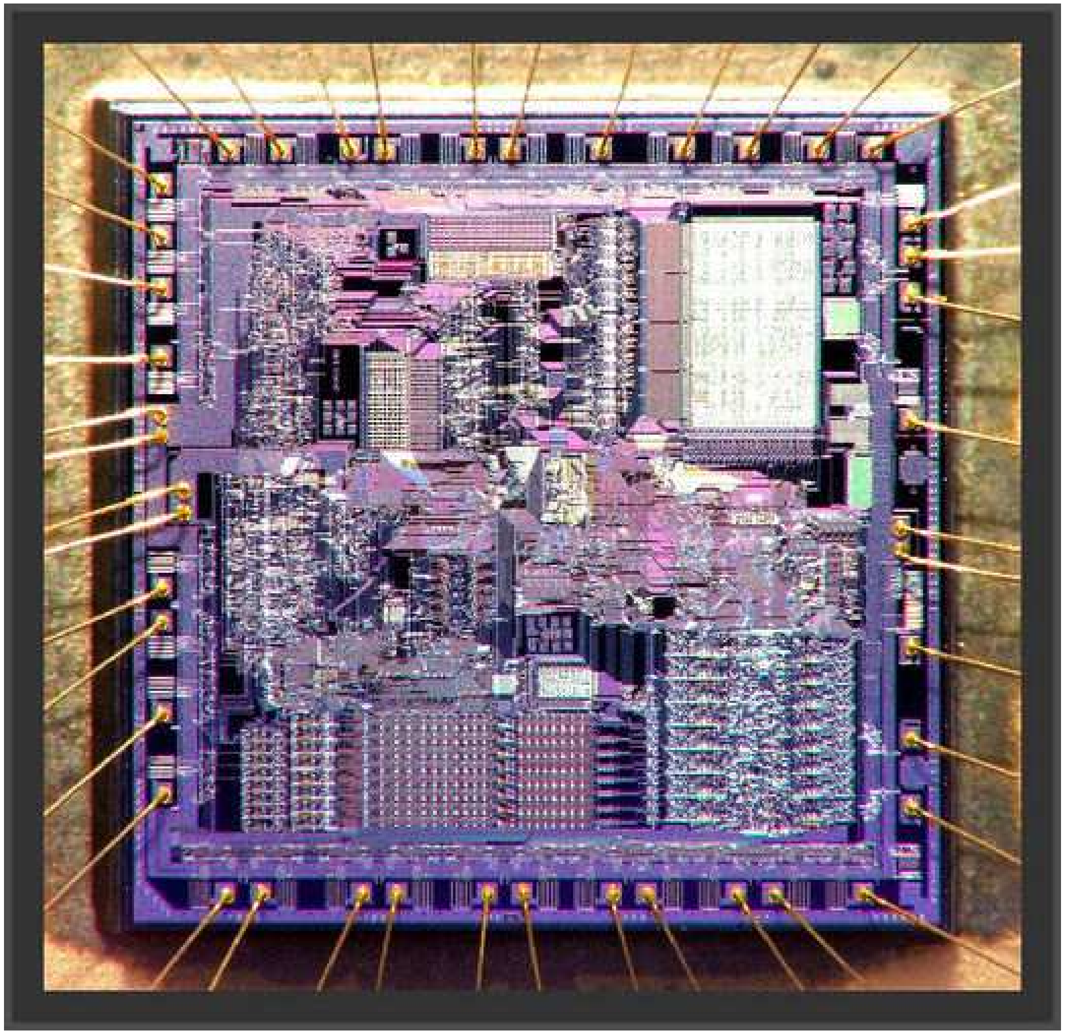
but,

C4s ~ O(Area)

TSVs ~ O(Area)

Unsustainable!

In the past: 2D power delivery wall



Power ~ O(Area)

Vdd = ct.

I ~ O(Area)

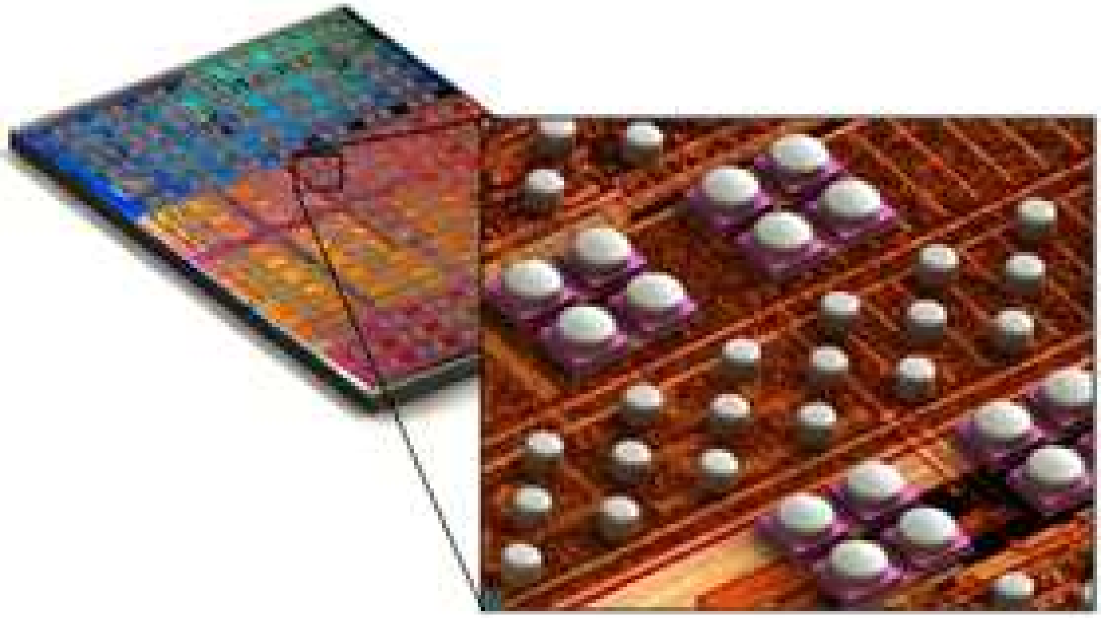
but,

pads ~ O(Perim)

Unsustainable!

Source: http://gadgets.boingboing.net/2008/04/28/power-on-self-test-l.html

2D: Flip-chip to the rescue



Physical solution

Power ~ O(Area)

Vdd = ct.

I ~ O(Area)

C4s ~ O(Area)

2D Solved!

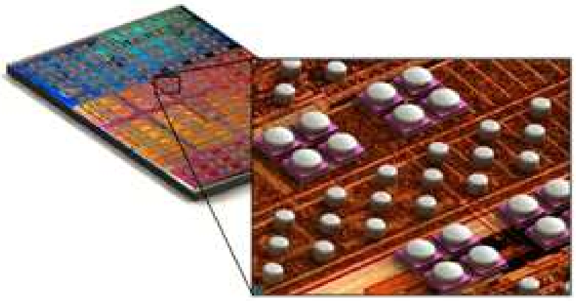
Not 3D though…

C4 - Controlled Collapse Chip Connection

Source: Nextreme, Inc.

Power Pin "Wall"

- C4 Count ~ **constant** : ↑current/pad
- For n layer 3DIC , Power pin count ~ 1/nth of 2D IC
- **Electro migration** can cause open/short circuit
  => Chip failure



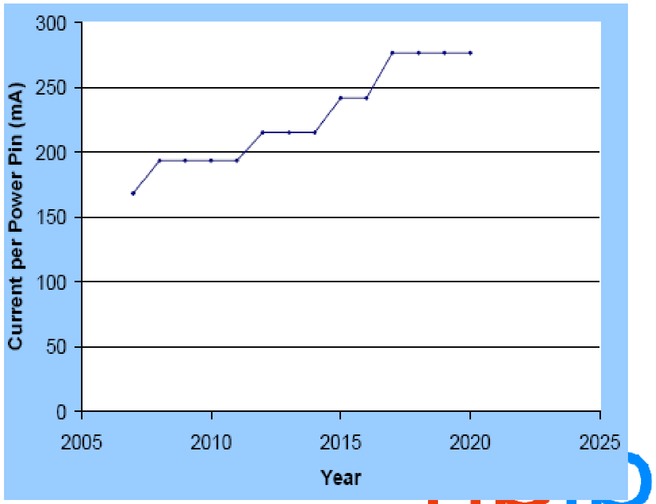
C4: controlled collapse chip connection

ITRS Roadmap 2009

Parameter	2009	2012	2015
Average Current Density (A/cm^2)	64	108	150
Power Pins (% of total pins)	66	66	66
Off chip Data rate (Gb/s)	8	14	30

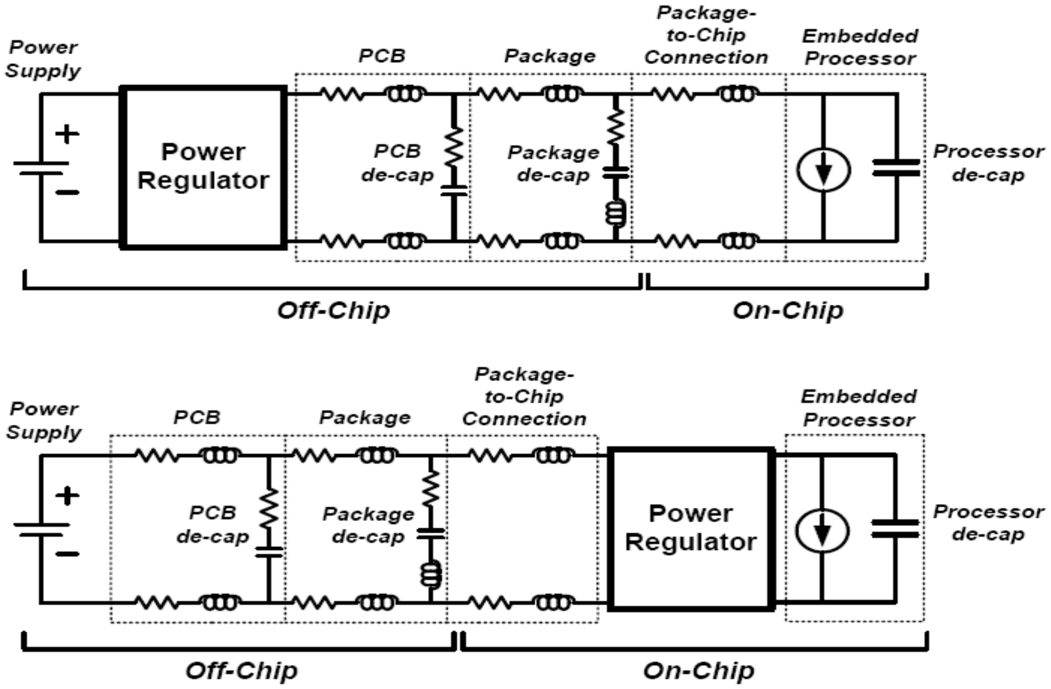
Source: Nextreme, Inc

Current Per Power Pin (2D), ITRS



Source : Pingqiang Zhou et al, " Reliable Power Delivery for 3D ICs"

Off-Chip Versus On-Chip Regulation



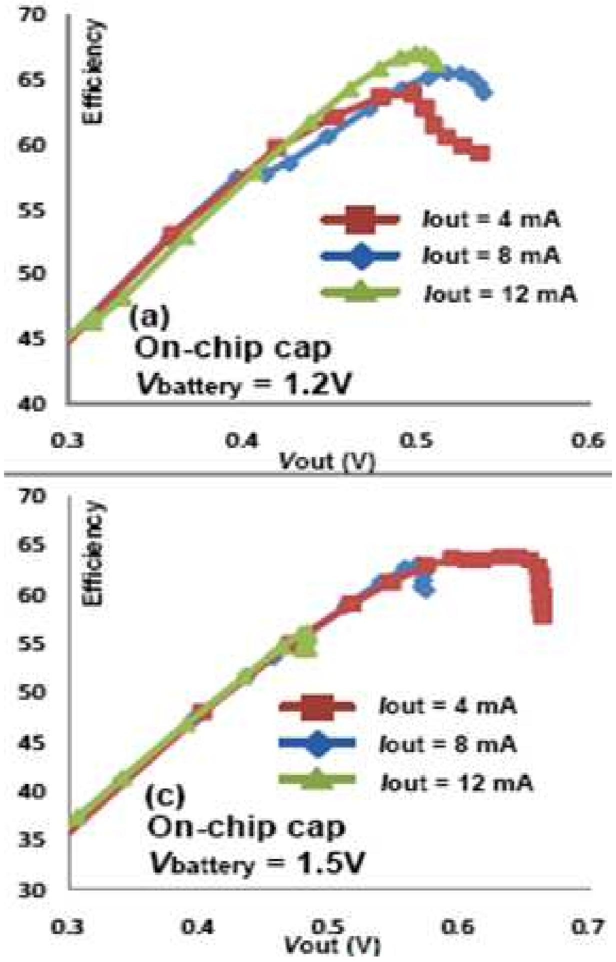
- Reduces Off-Chip I^{2}R **losses** in PDN parasitic

- **Fast** On-Chip voltage scaling

Zhiyu Zeng et al, “ Tradeoff analysis and optimization of power delivery networks with on chip voltage regulation”, DAC, 201

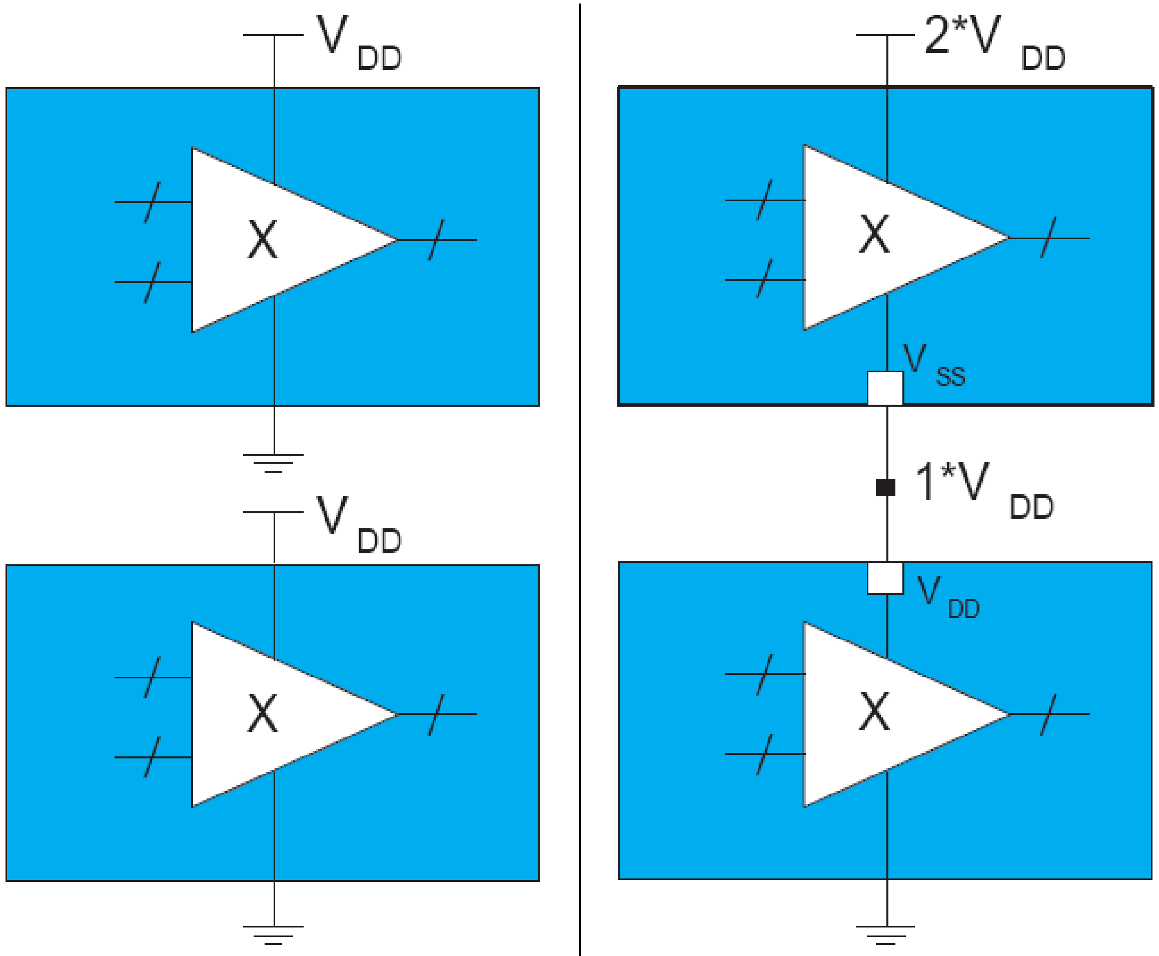
On-Chip-Regulation Efficiency “Wall”

- Off-Chip High Voltage => On-Chip Low Voltage : **Power loss**
- Switching Regulator : High efficiency, Difficult to integrate On-Chip
- LDO efficiency : constrained by V_{out}/V_{in}
- Switched Capacitor : Switch Conductance/ Switching Loss
- Existing regulation techniques **not energy efficient** to generate low voltage and exploit DVFS



Source : Yu Pu, “Misleading energy and performance claims in Sub-Near Threshold Digital System”, ICCAD 2010

3D IC: Voltage Stacking to the rescue



$$I_{tot} = \frac{V}{R} + \frac{V}{R} = \frac{2V}{R}$$

$$I_{tot} = \frac{2V}{2R} = \frac{V}{R}$$

Electrical solution

Kirchoff’s current and voltage laws!

Power ~ O(Vol)

I = O(Area)

Also for N > 2

Vdd = O(N) = O(Vol/Area)

Voltage Stacking

- 3D IC power delivery walls arise due to unsustainable increase in *current*
- Solution for delivering increased power without increase in current is to increase *voltage*
- Essentially the same idea used in macroscopic power distribution grids
- Simply increasing voltage for high voltage on-chip require explicit on-chip DC/DC regulators
- Voltage stacking uses *implicit* power regulation based on Kirchoff’s voltage law (Ohm’s law)

Pros and Cons

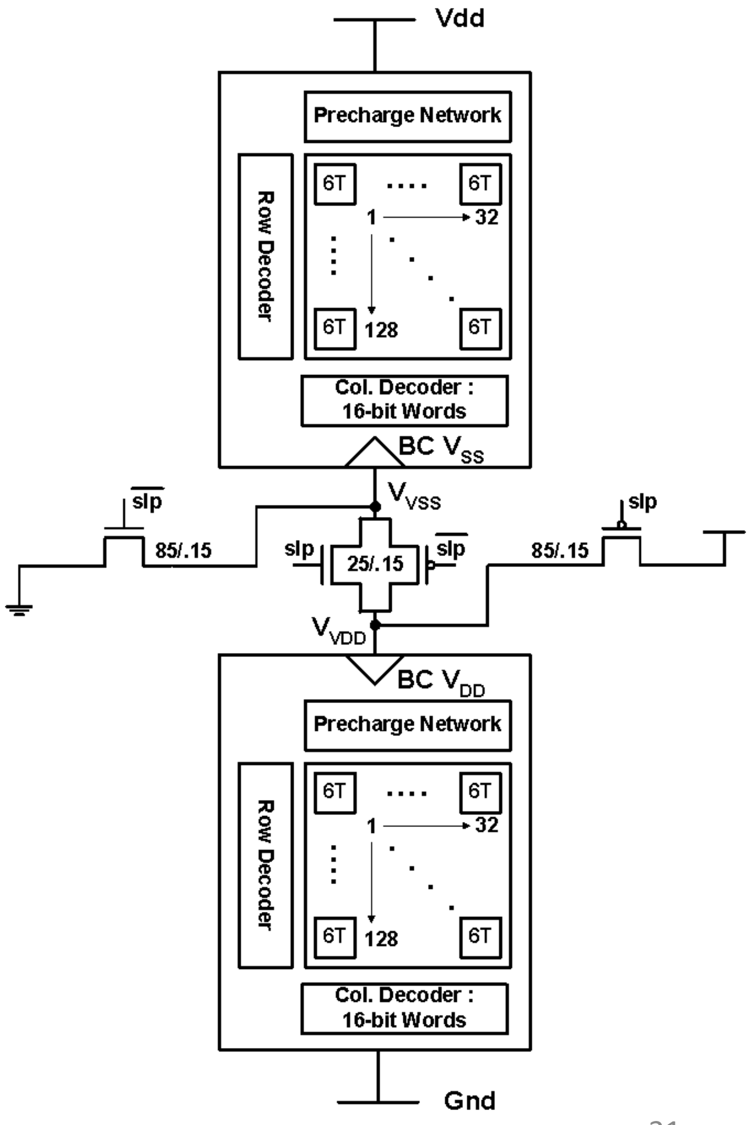
- Power Pin : **Implicit Regulation**; k cores stacked need same/less number of power pins as single core
- Off-Chip I^2R Power Loss : k^2 times ↓
- IR Drop : k times ↓
- Efficiency : Depending on "Imbalance" ↑
- 3D IC : Physical layering of 3D IC naturally maps to voltage stacked solution reducing TSV count
- Inter-layer core activity mismatch : **Internal voltage noise**

Stacking for other power walls

Power efficiency wall

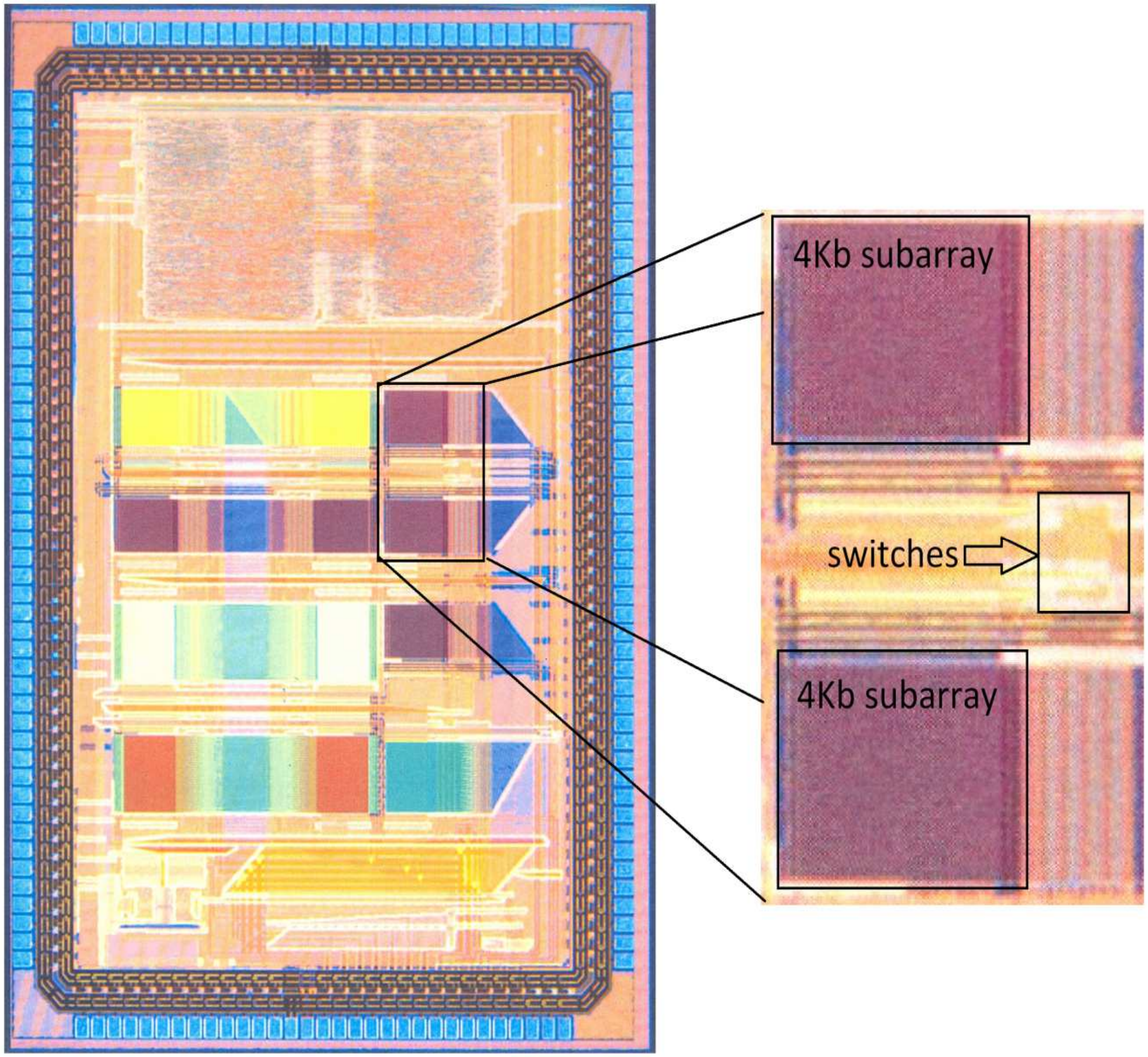
Active mode: parallel

During sleep: stacked



hplp
High-Performance Low-Power

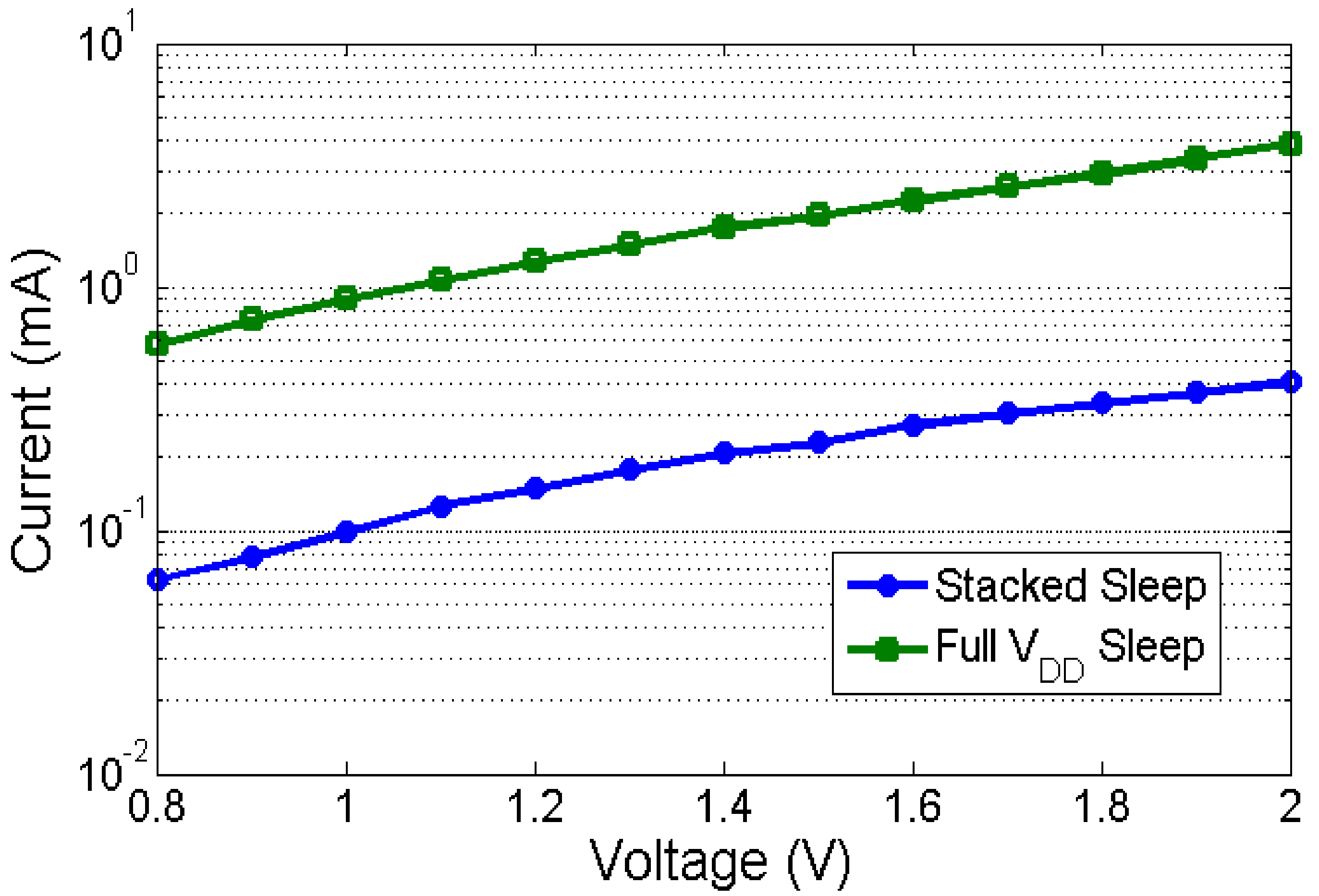
Fabricated chip die photo



A. Cabe, M. Stan, “Standby Power Reduction using Voltage Stacking” GLSVLSI 2011

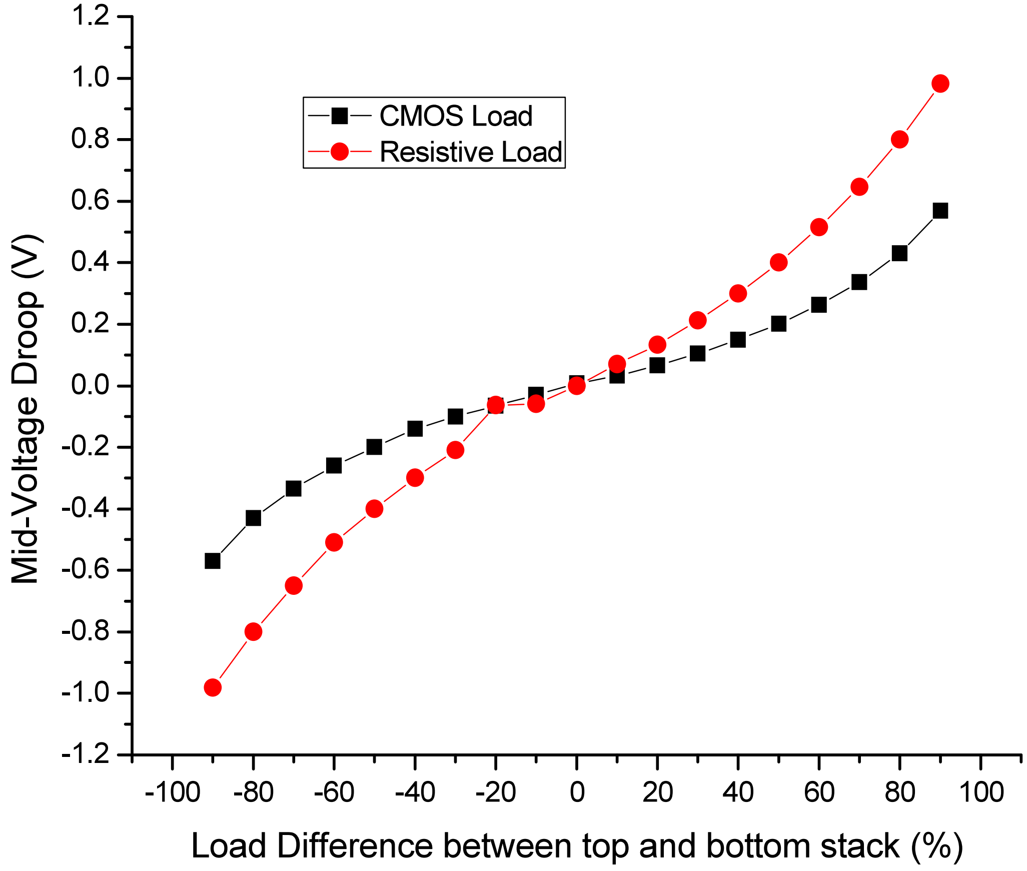
Power Savings during Sleep

- 1 order of magnitude savings!



hplp
High-Performance Low-Power

Implicit Regulation : Resistive Versus CMOS Stacked Load



Resistive load: V α I
CMOS load : V α √I

For CMOS load, less dependency of Voltage droop on load current

Voltage droop α I_{Load} difference between the stacked layers

Stacked CMOS Load

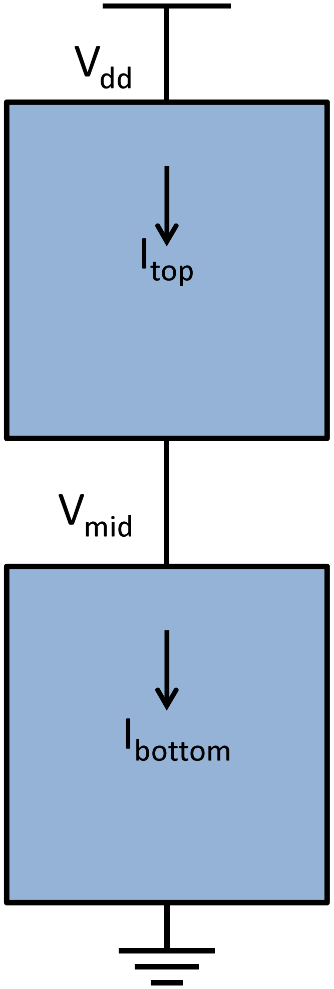
Charge conservation $\mathbf{I_{top} = I_{bottom}}$

$\mathbf{I_{top} = \alpha_{top}\ C_L(V_{dd} – V_{mid})Fc}$ $\mathbf{I_{bottom} = \alpha_{Bottom}\ C_L V_{mid}\ Fc}$

$\mathbf{V_1 = V_{dd} – V_{mid} = \alpha_{bottom}/\ (\alpha_{top} + \alpha_{bottom})}$
$\mathbf{V_2 = V_{mid} = \alpha_{top}/\ (\alpha_{top} + \alpha_{bottom})}$

α_{top} ,α_{bottom} : Top and bottom core activity factors,
F_c: Core frequency C_L: Capacitive load
V_{mid} the output voltage delivered.

$\mathbf{\alpha_{top} = \alpha_{Bottom}}$ $\mathbf{V_{mid} = 0.5\ V_{dd}}$



High-Performance Low-Power

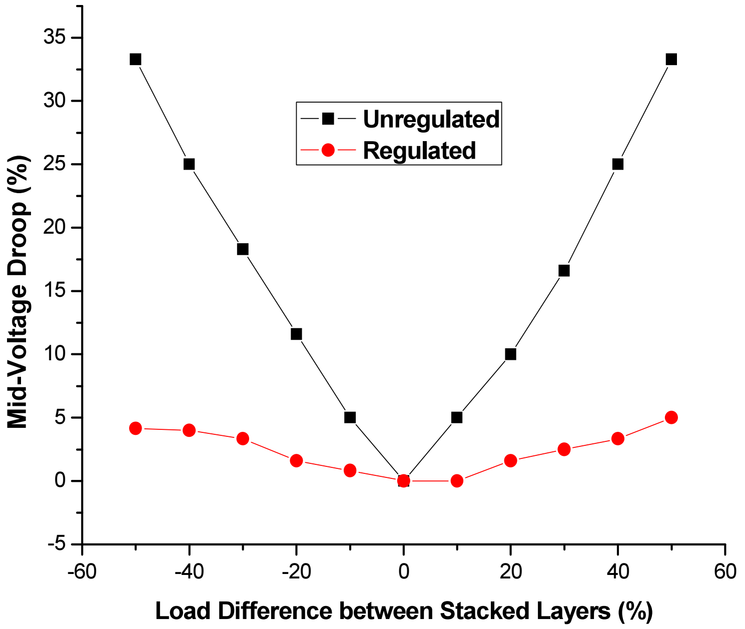
Explicit Regulation needed ?

$\alpha_{top} > \alpha_{Bottom}$ $V_{mid} > 0.5\ V_{dd}$:

Self-Regulation forces lower voltage headroom for high activity cores

Unregulated Voltage Stacking **oppose DVFS**

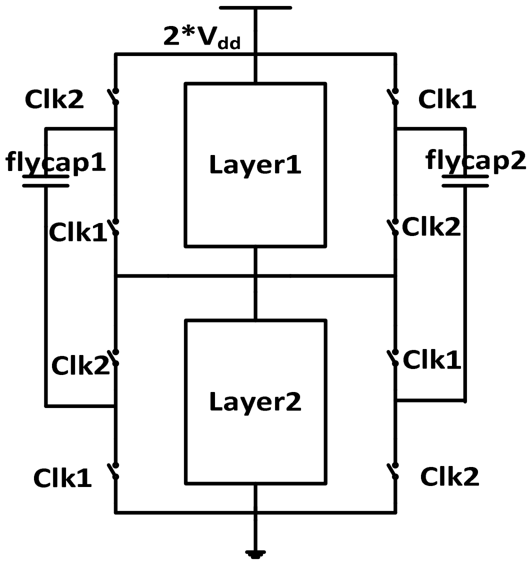
Explicit Regulator: **Sink/Source "imbalance"** and compensate for the natural "feedback"

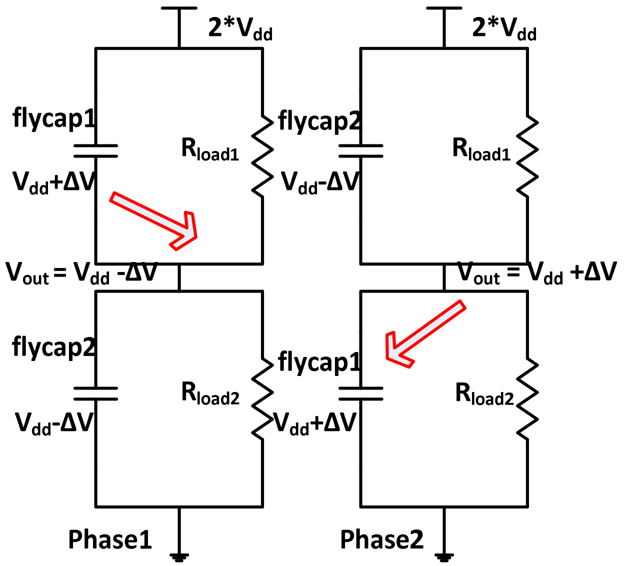


Kaushik Mazumdar et al, " Charge Recycling On Chip DC-DC Conversion for Near-Threshold Operation", IEEE SubVt, Boston, 2012

On-Chip Regulator

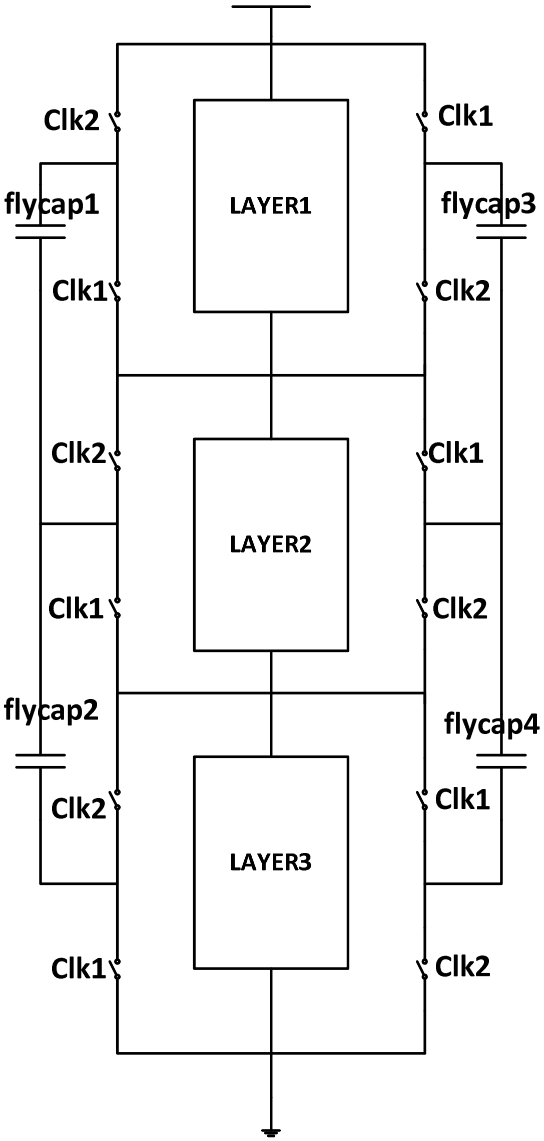
- Switched Capacitor (SC) :





Assuming Current offset: V_{out} droops below V_{dd}
Phase 1 : flycap1 charges to $V_{dd} + \Delta V$ while flycap2 to $V_{dd} - \Delta V$
Phase 2 : flycap1 and flycap2 swap, redirecting charge to V_{out}

Voltage Stacking for more than 2 Layers



hplp
High-Performance Low-Power

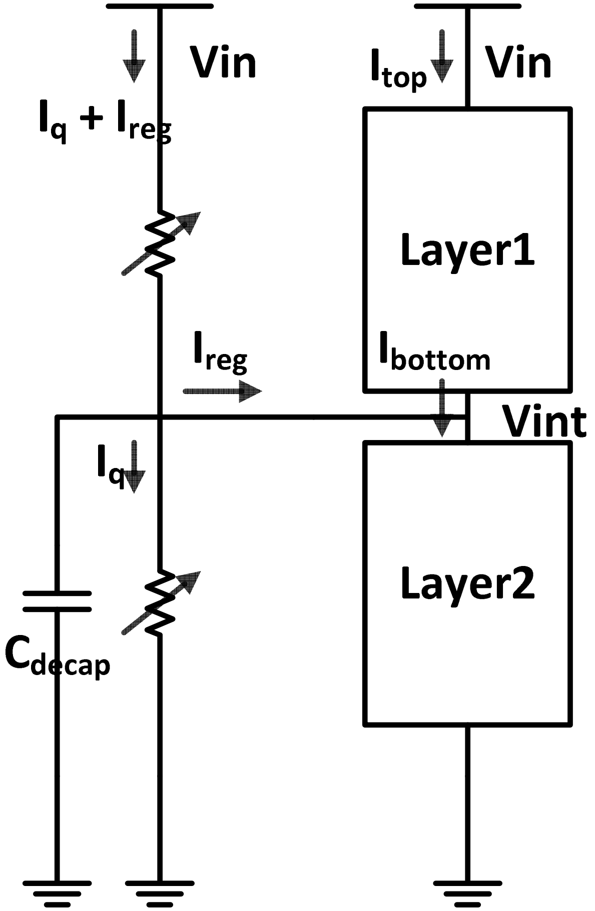
Efficiency of V-S Regulated Technique

Efficiency: depends on **mismatch** between the stacked domains

Mismatch :

- Activity of the circuits,
- Evaluation node capacitance
- Voltage swing in the domains.

$$\text{Efficiency} = \frac{\text{Power\_logic}}{\text{Power\_system}} = \frac{V_{IN}\, I_{top} + V_{INT}\, |I_{reg}|}{V_{IN}\, (\, I_{top} + I_{reg}| + I_q\,)}$$

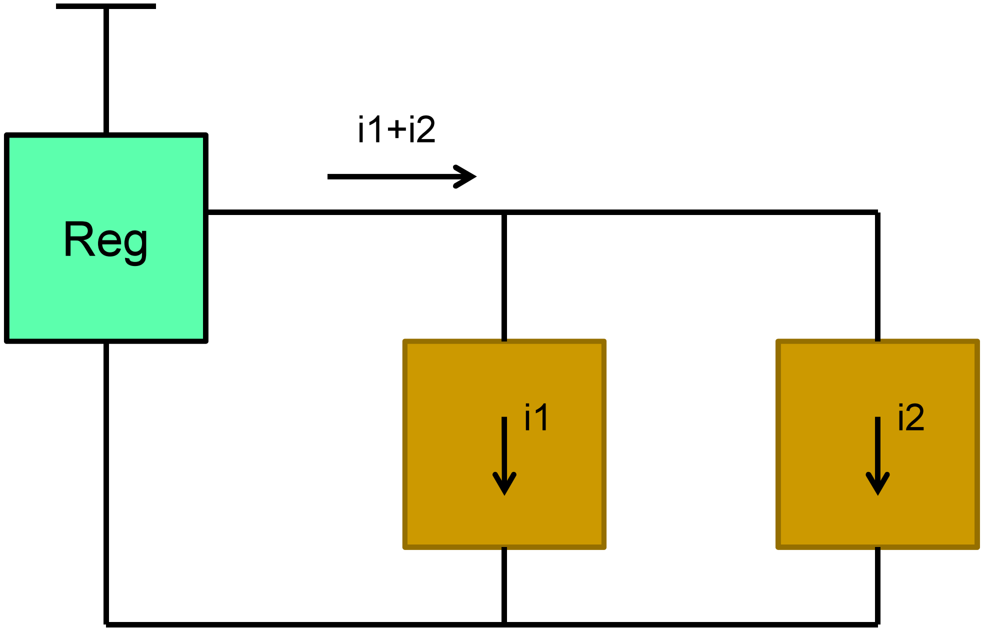


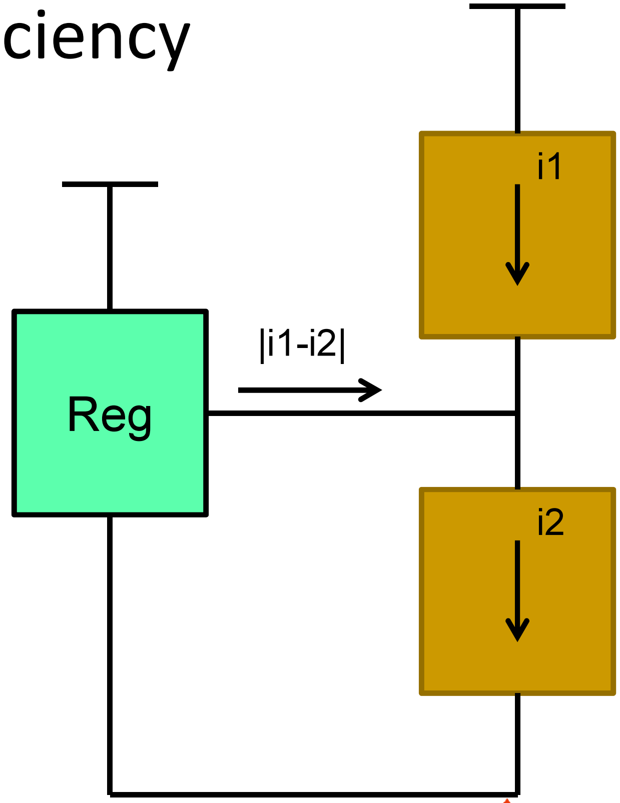
Higher efficiency in Voltage Stacking

Implicit vs. Explicit

Regulate the current difference, not the sum!

Lower imbalance leads to higher efficiency

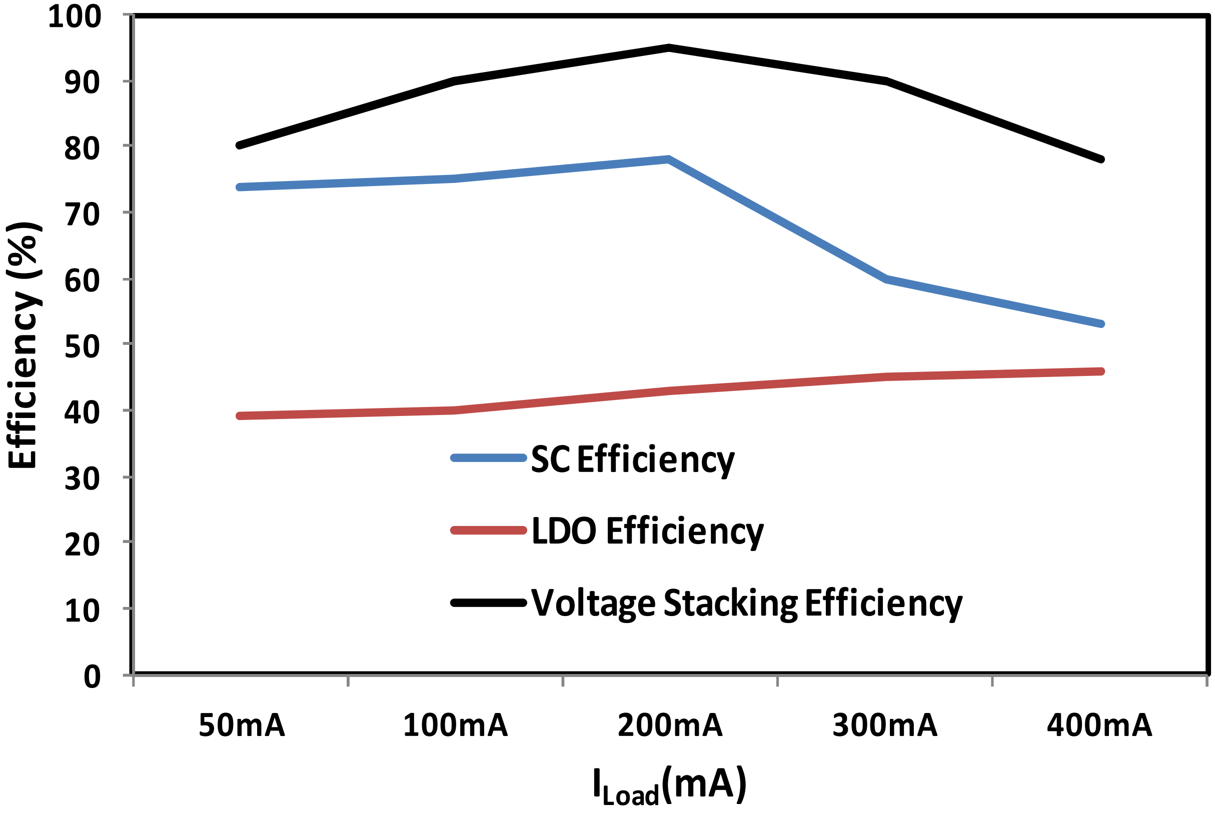




hplp

High-Performance Low-Power

Efficiency Comparison



Voltage Stacking Efficiency dependent on **mismatch** : More than 90% Efficiency for closely matched stacked load

Worst case V-S Efficiency ~ SC Efficiency

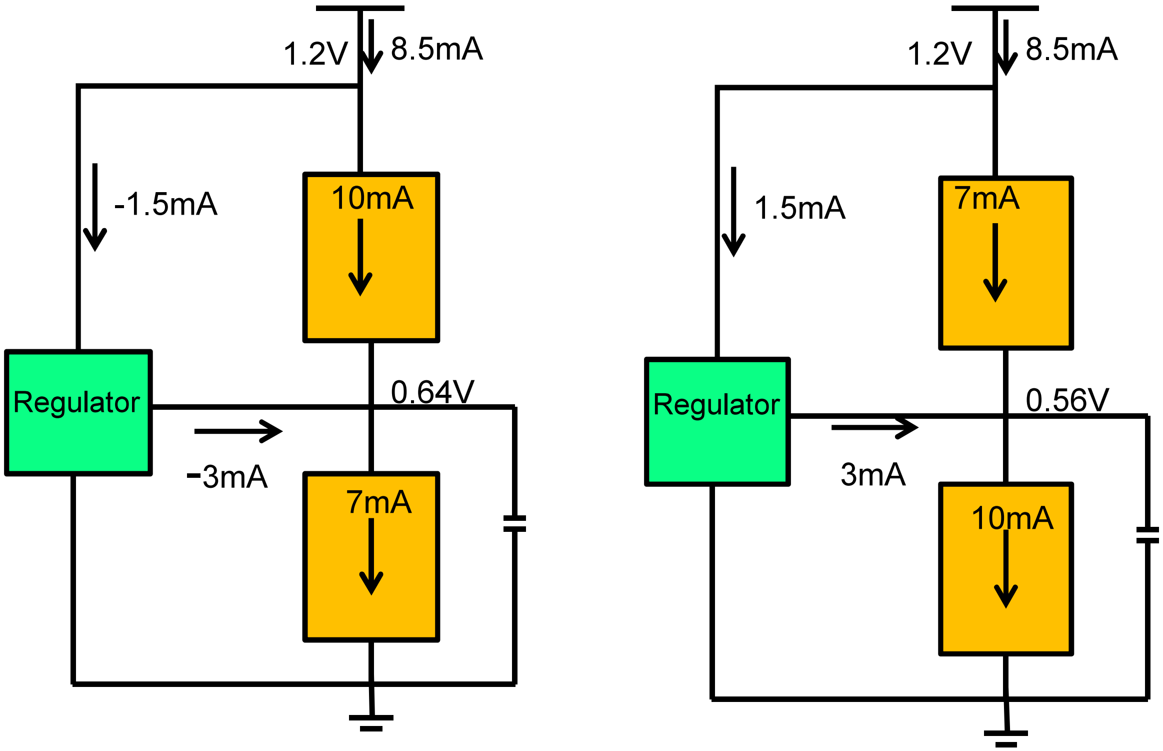
hplp

High-Performance Low-Power

Positive Vs. Negative Imbalance

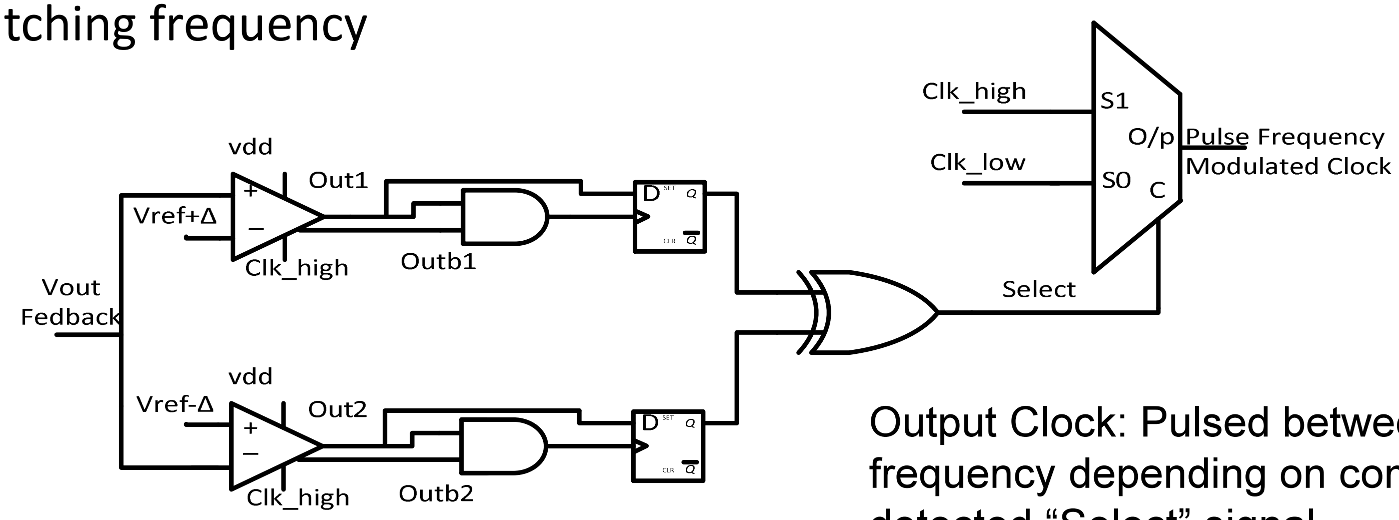
Positive imbalance: similar to conventional regulator (Sourcing I_{Load})

Negative imbalance: regulator absorbs current (Sinking I_{Load})



Feedback Control Circuitry

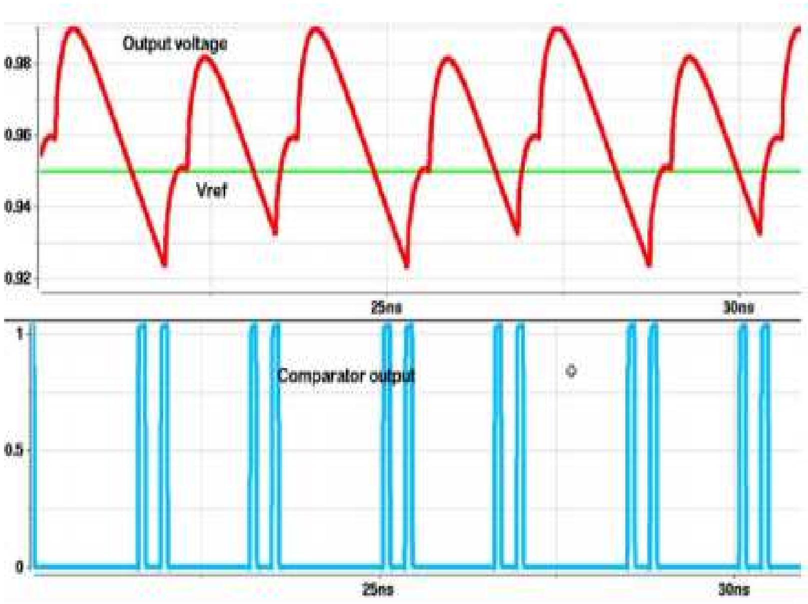
- $V_{out} = nV_{in} – i_{out}\ Rout\ (f_{sw}, D_i, G_i)$
- Hysteretic feedback scheme with lower and upper bounds to modulate the switching frequency



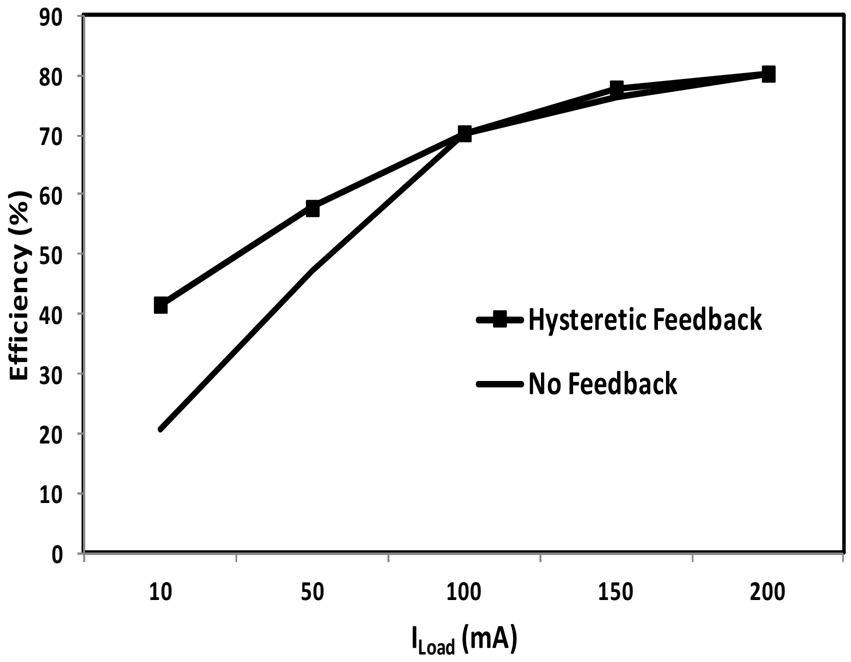
Output Clock: Pulsed between high/low frequency depending on comparator detected "Select" signal

State of O/P	Out1	Out2	Select	O/P Clock
$V_{out} > V_{ref}+\Delta$	Toggle	Low	1	Clk_high
$V_{ref-\Delta}<V_{out}< V_{ref+\Delta}$	Low	Low	0	Clk_low
$V_{out} <V_{ref-\Delta}$	Low	Toggle	1	Clk_high

Feedback with Conventional Load



Comparator o/p acting as frequency modulated clock

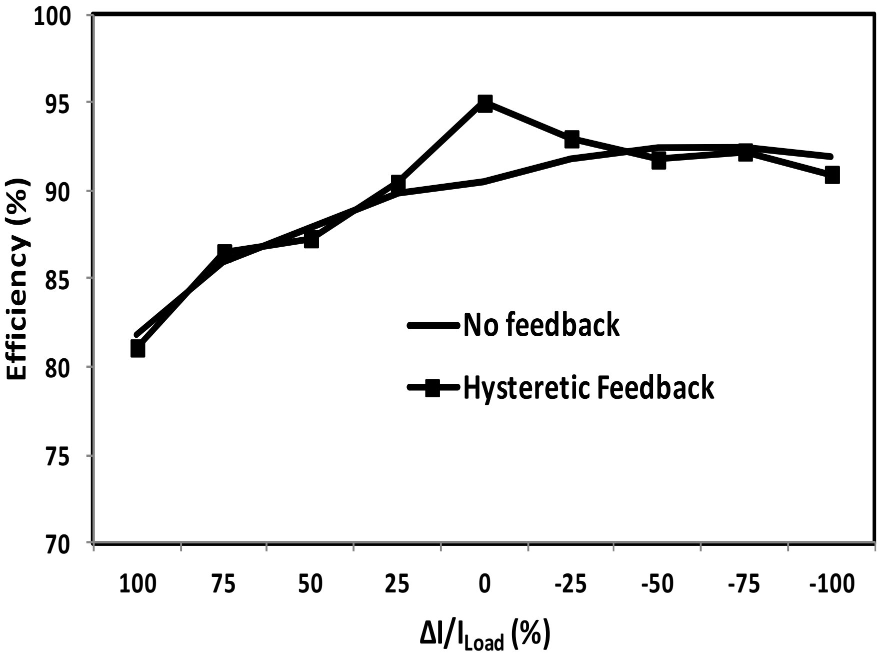


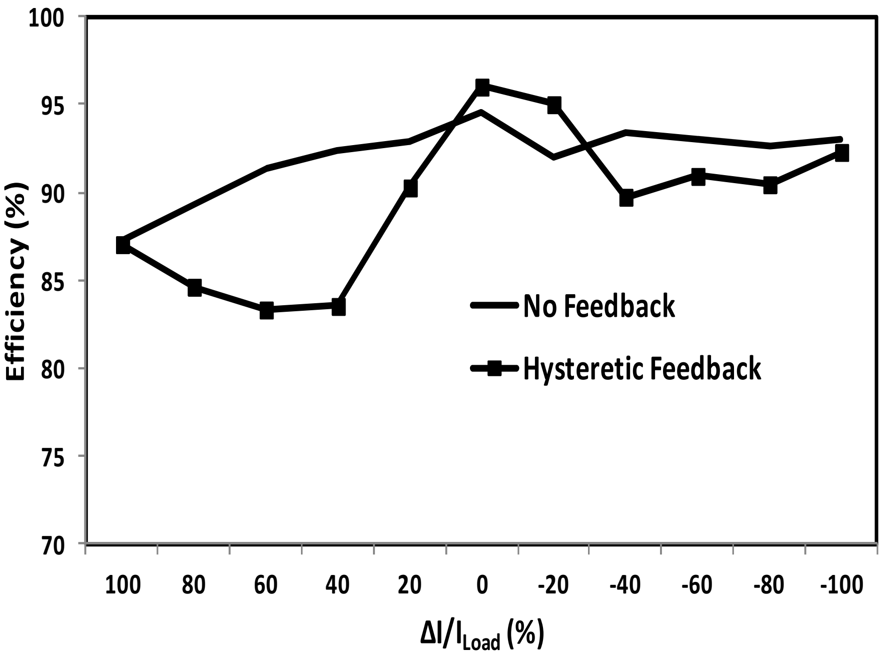
Efficiency : Improves at low power with feedback

Jain, R : **200mA switched capacitor voltage regulator on 32nm CMOS and regulation schemes to enable DVFS,*(EPE 2011)***

hplp
High-Performance Low-Power

Open/Closed Loop for Stacked Load ?

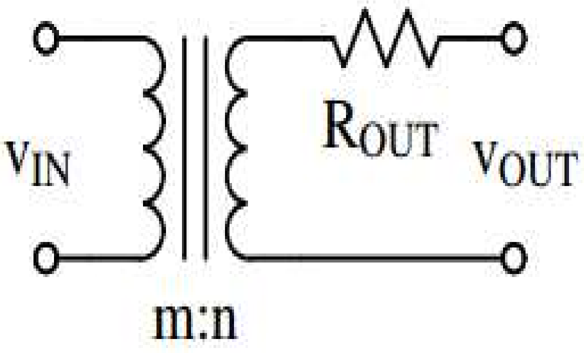




- Comparison of open-loop/close-loop SC circuit for high power (left: 10mW-400mW, 2V→1V) and low power (right: 0.5mW-10mW, 1.2V→0.6V) loads

- **Higher efficiency for Open loop regulation** (low power loads)

Switch Capacitor Model



m:n No-Load Conversion ratio

R_{OUT} has 2 asymptotic limits : Slow Switching Limit (R_{SSL}) and Fast Switching Limit (R_{FSL})

R_{SSL} => **Ideal Switches**, Current **Impulsive** in nature, Impedance **inversely proportional** to Switching Frequency

R_{FSL} => Switches and capacitance resistance dominate, capacitance act as fixed voltage source, Impedance **independent** of Switching frequency

$$\mathbf{R_{OUT} \sim \sqrt{(R_{FSL}{}^2 + R_{SSL}{}^2)}}$$

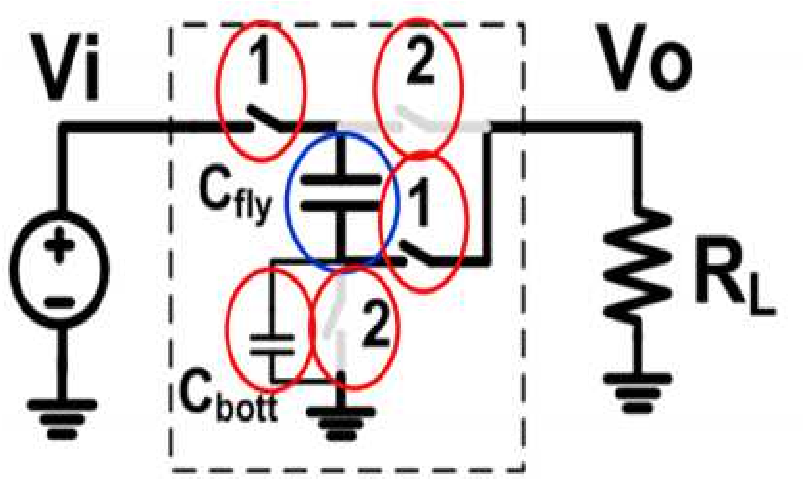
Source : Seeman, " A Design Methodology for Switched-Capacitor DC-DC Converter", PhD Dissertation , UC Berkeley , May 2009

Switch Capacitor Power Loss

- **SSL** impedance Loss : Charge transfer related loss => $\mathbf{I_{Load}^2.R_{SSL}}$
- FSL impedance Loss : Switch conductance loss => $\mathbf{I_{Load}^2.R_{FSL}}$
- Switch Drive Loss : Parasitic loss in the switches => $\mathbf{V_{swing}^2.N.W_{switch}.C_{gate}.F_{sw}}$
- Bottom Plate Loss
- ESR Loss in Capacitor

High-Performance Low-Power

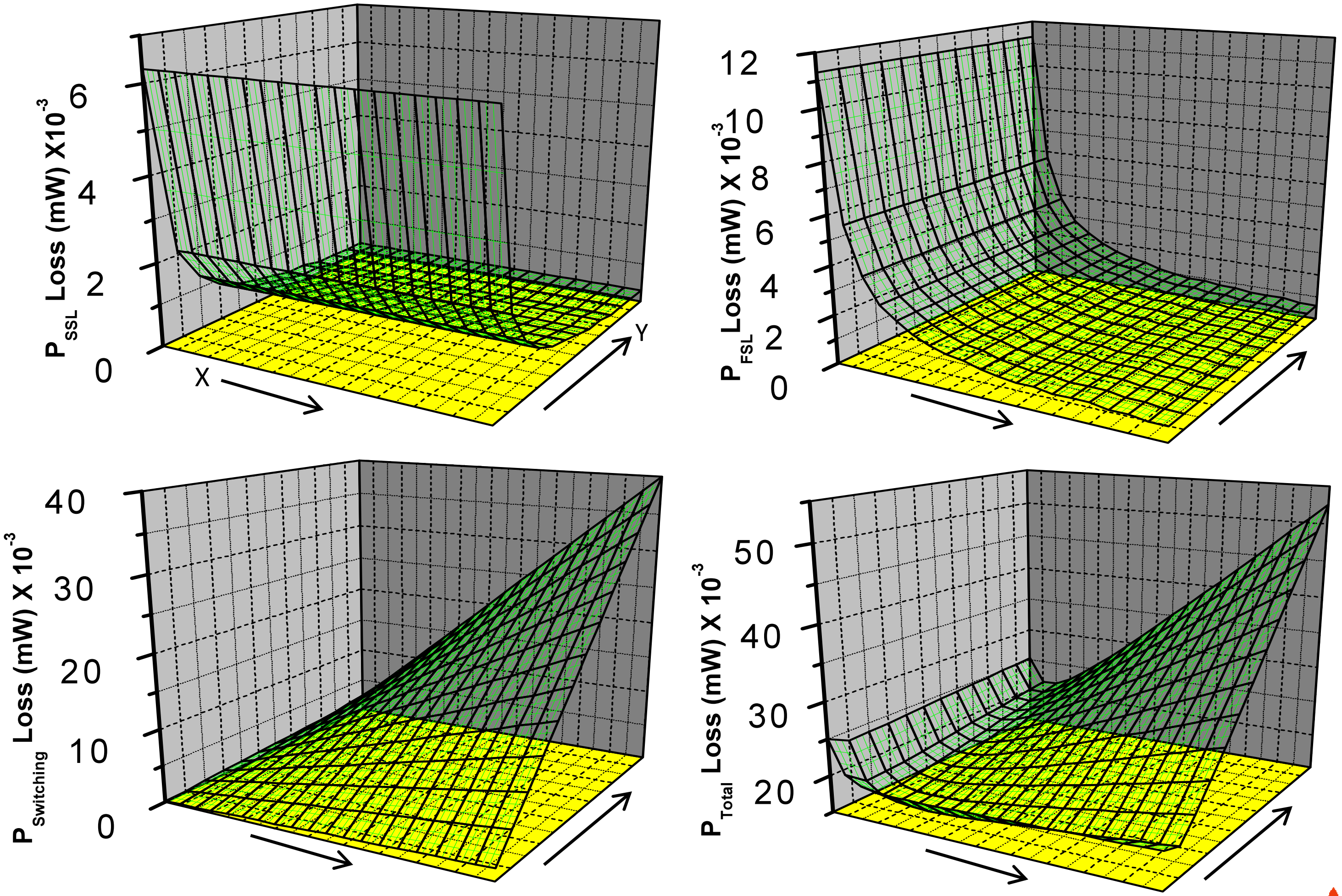
Power Loss Optimization



- **Intrinsic loss**
  - Reduced by ↑ C density
  - Reduced by ↑ f_{sw}
- **Switch/parasitic loss**
  - Reduced by ↑ switch f_T
  - Increased by ↑ f_{sw}

Source : Hanh-Phuc et al, "Design Techniques for Fully Integrated Switched-Capacitor DC-DC Converters"

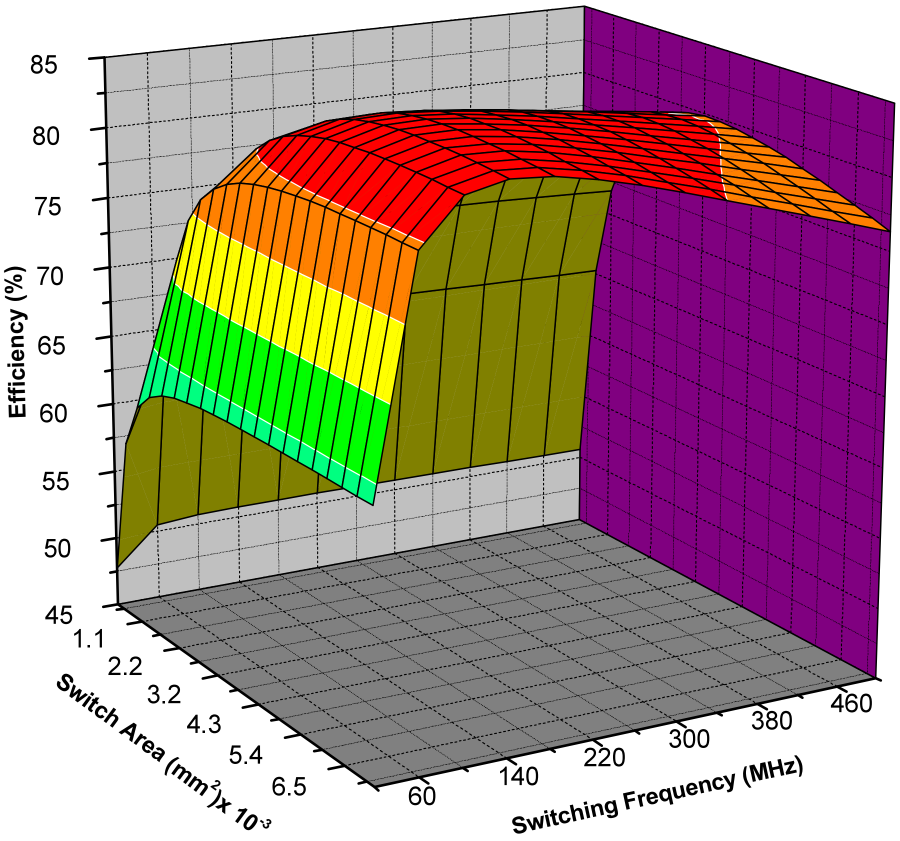
Power Loss Breakdown



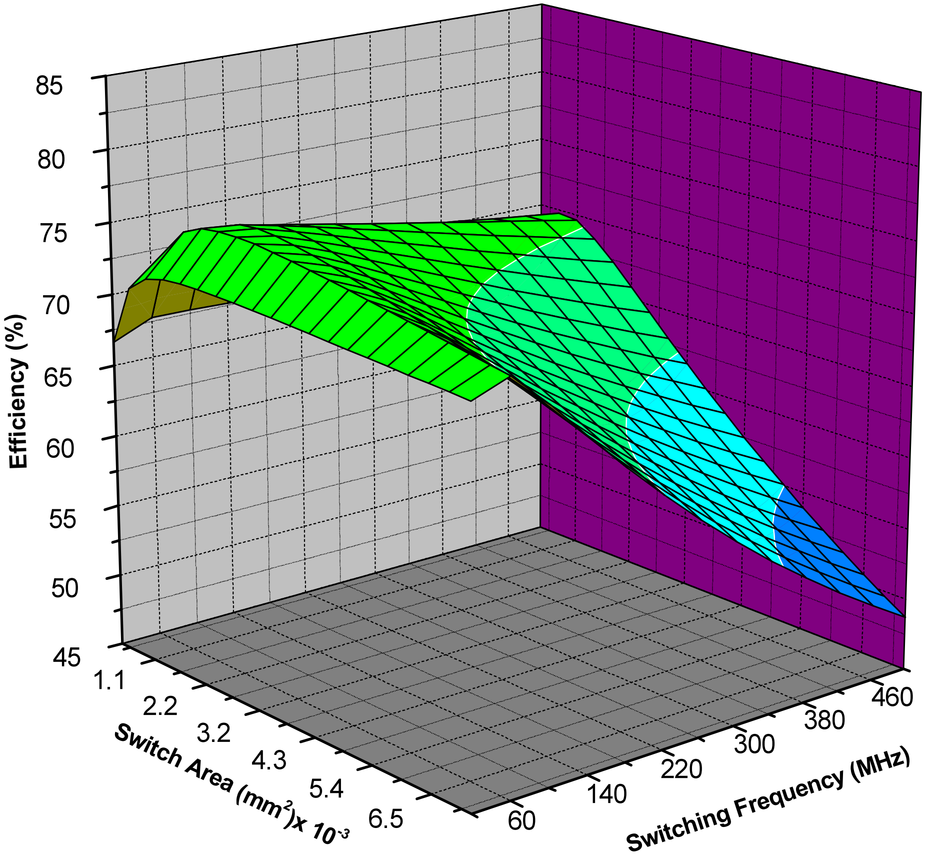
X-Axis => Switch Area

Y-Axis => Switching Frequency

Efficiency Versus Design Knobs



I_{Load} = 200mA

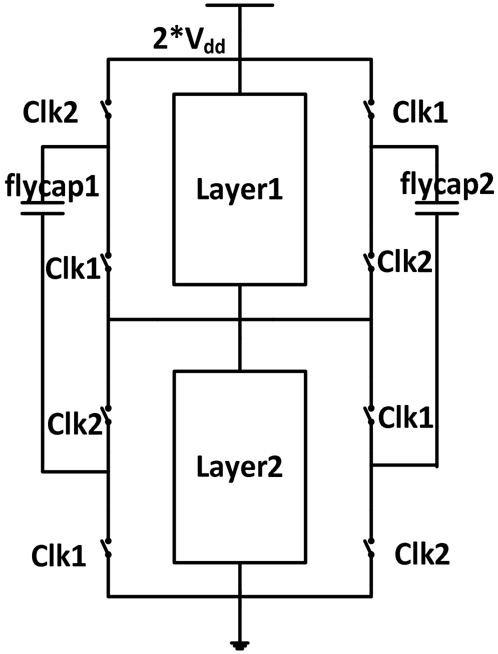


I_{Load} = 50mA

hplp

High-Performance Low-Power

Output Impedance with imbalance



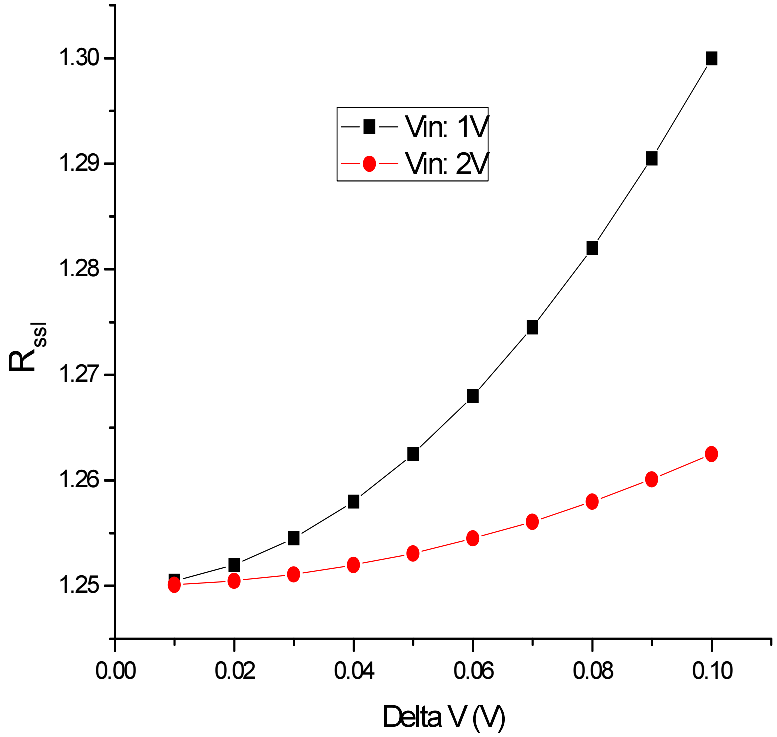
i_1 (Layer1 current) = i_2 (Layer 2 current)
$R_{ssl} = 1/(2*f_{sw}*C_{fly})$
$R_{fsl} = 4R$

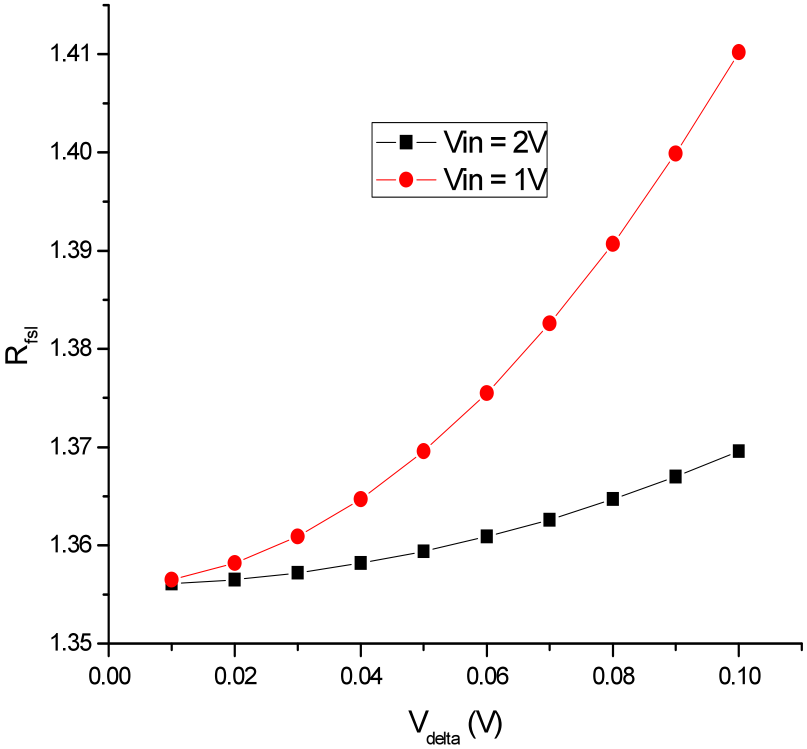
i_1 > i2 or i_1<i2
$R_{ssl} = 1/(2*f_{sw}*C_{fly}) + [2(\Delta V/V_{in})]^2$
$R_{fsl} = 4R[1+4\ (\Delta V/V_{in})^2]$

Increase in R_{out} with increase in imbalance and lowering of Vin

High-Performance Low-Power

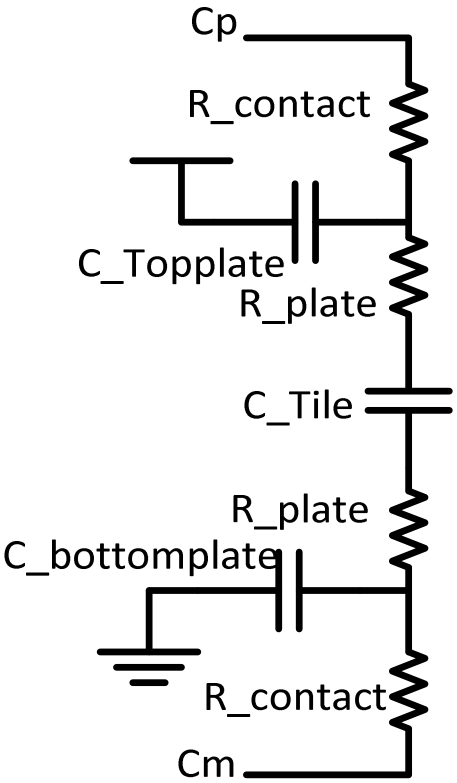
Output Impedance

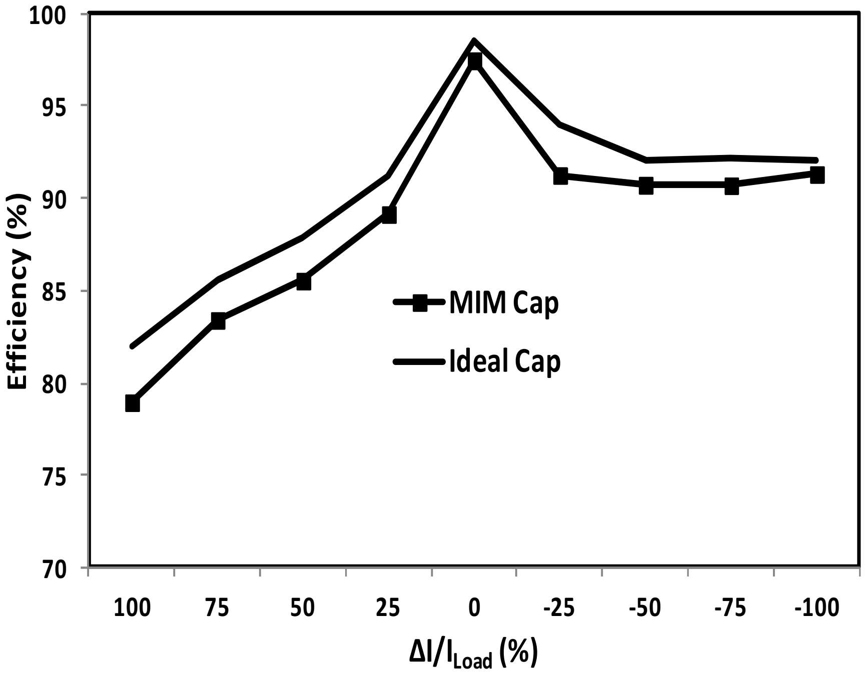




More the imbalance, more V_{delta} and more the loss

Impact of Capacitor Parasitic



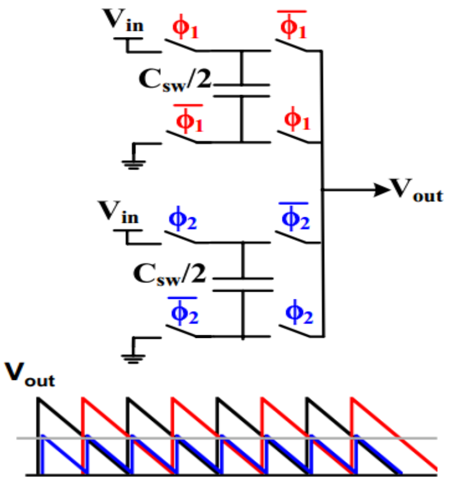


MIM cap Model (Including top/bottom plate capacitance, plate and contact resistance

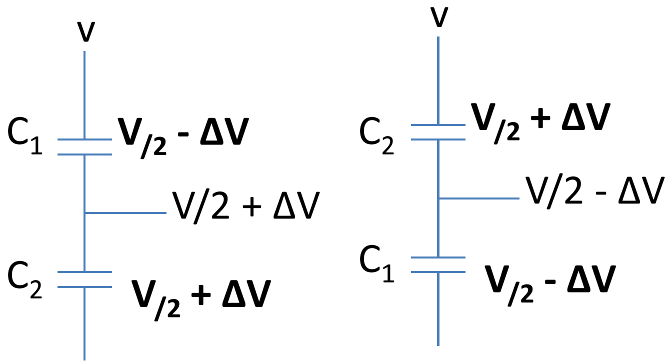
MOS Cap -> Highest Density(12nF/mm^2), Max Bottom Plate Parasitic (7-10%)
MIM Cap -> Lower Density (2nF/mm^2), Less parasitic (2-3%)

hplp
High-Performance Low-Power

Interleaving – Ripple Mitigation



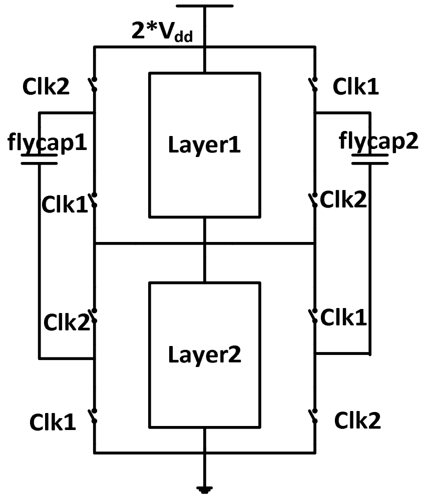
No interleaving



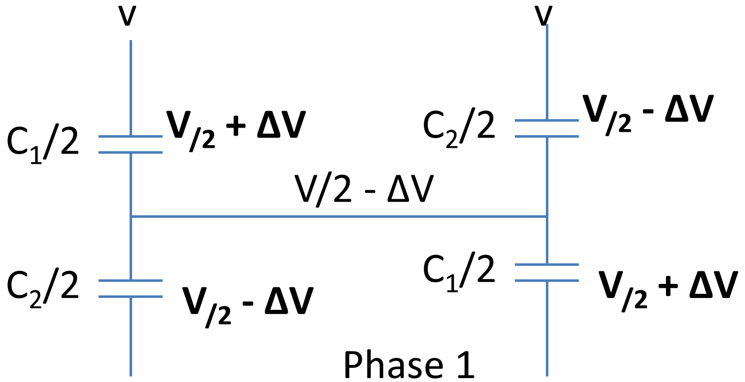
Phase 1

Phase 2

Fly caps never come parallel,
No energy loss through charge sharing



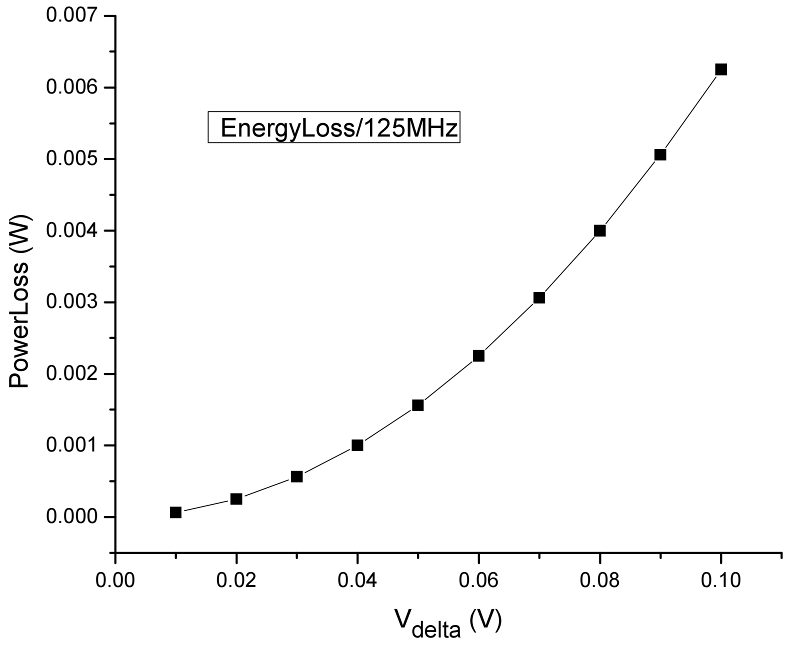
2-way interleaving



Phase 1

Fly caps come parallel to each other sharing 2ΔV of charge between them, leading to energy Loss

Power Loss with Interleaving



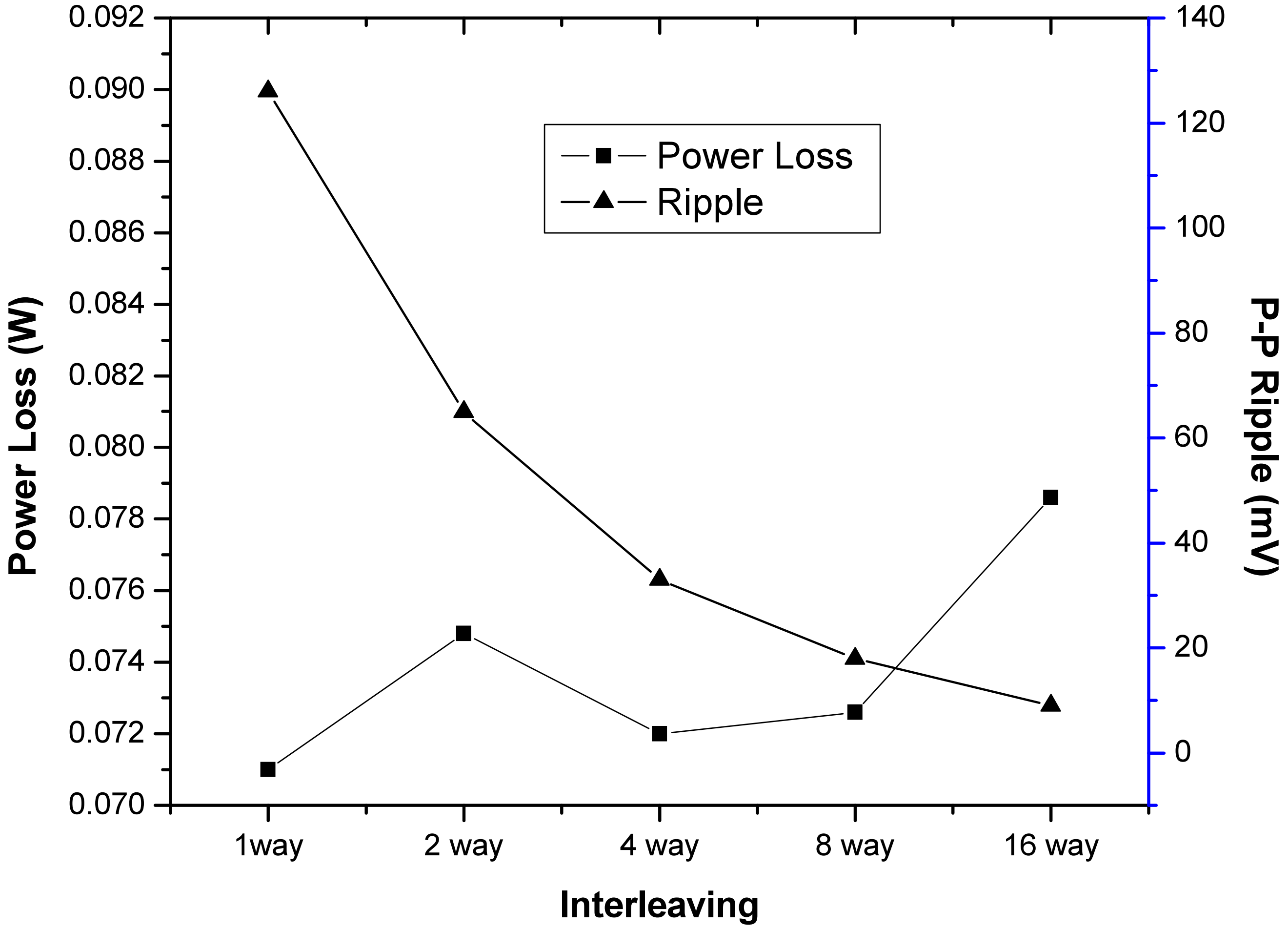
Energy Loss (interleaving 2 way) =
1/2*c*(1/2*vin+v_{del})^2+1/2*c*(1/2*vin-
v_{del})^21/2*c*(vin1/2*(c*(1/2*vin+v_{del})+
c*(1/2*vin-v_{del}))/c)^2-
1/8*(c*(1/2*vin+v_{del})+c*(1/2*vin-v_{del}))^2/c

As ΔV increases, Powerloss due to charge sharing increases

More interleaving, less ΔV and less the intrinsic loss, but more the extrinsic loss (from additional buffers and control circuitry needed for interleaving)

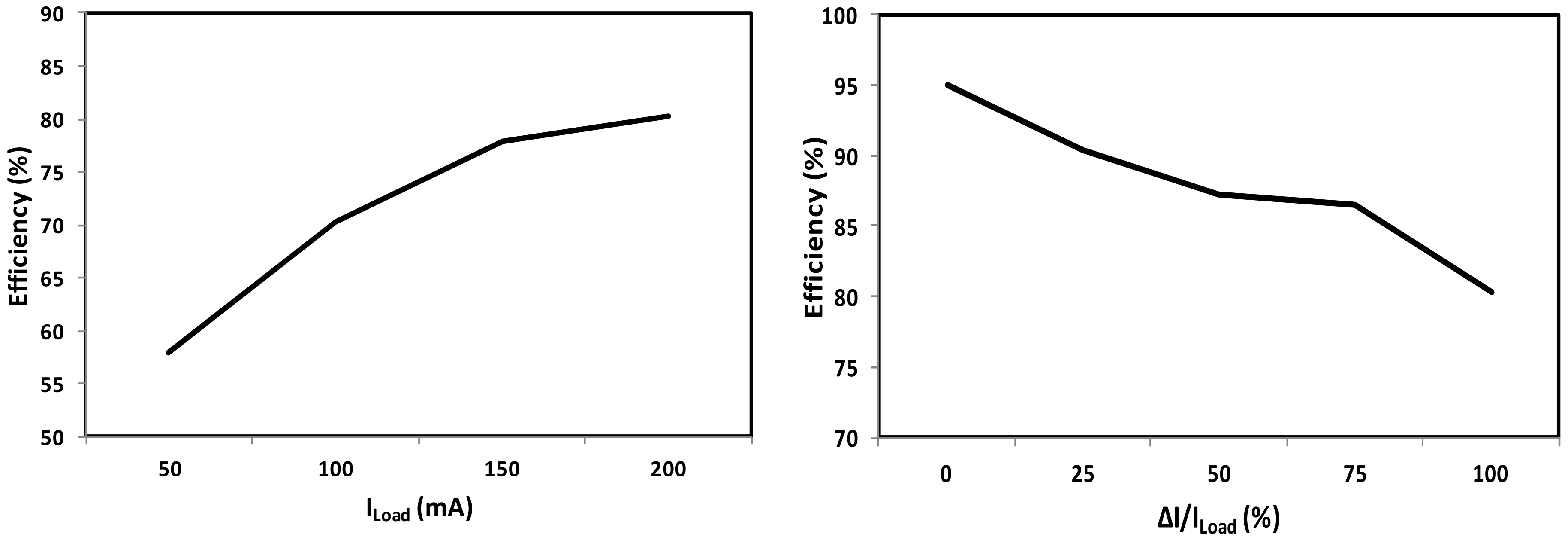
hplp
High-Performance Low-Power

Finding Optimum Interleaved Stages



Tradeoff between P-P Ripple (performance) and Power Loss (Efficiency

Efficiency : Conventional Versus Stacked Load



(a)

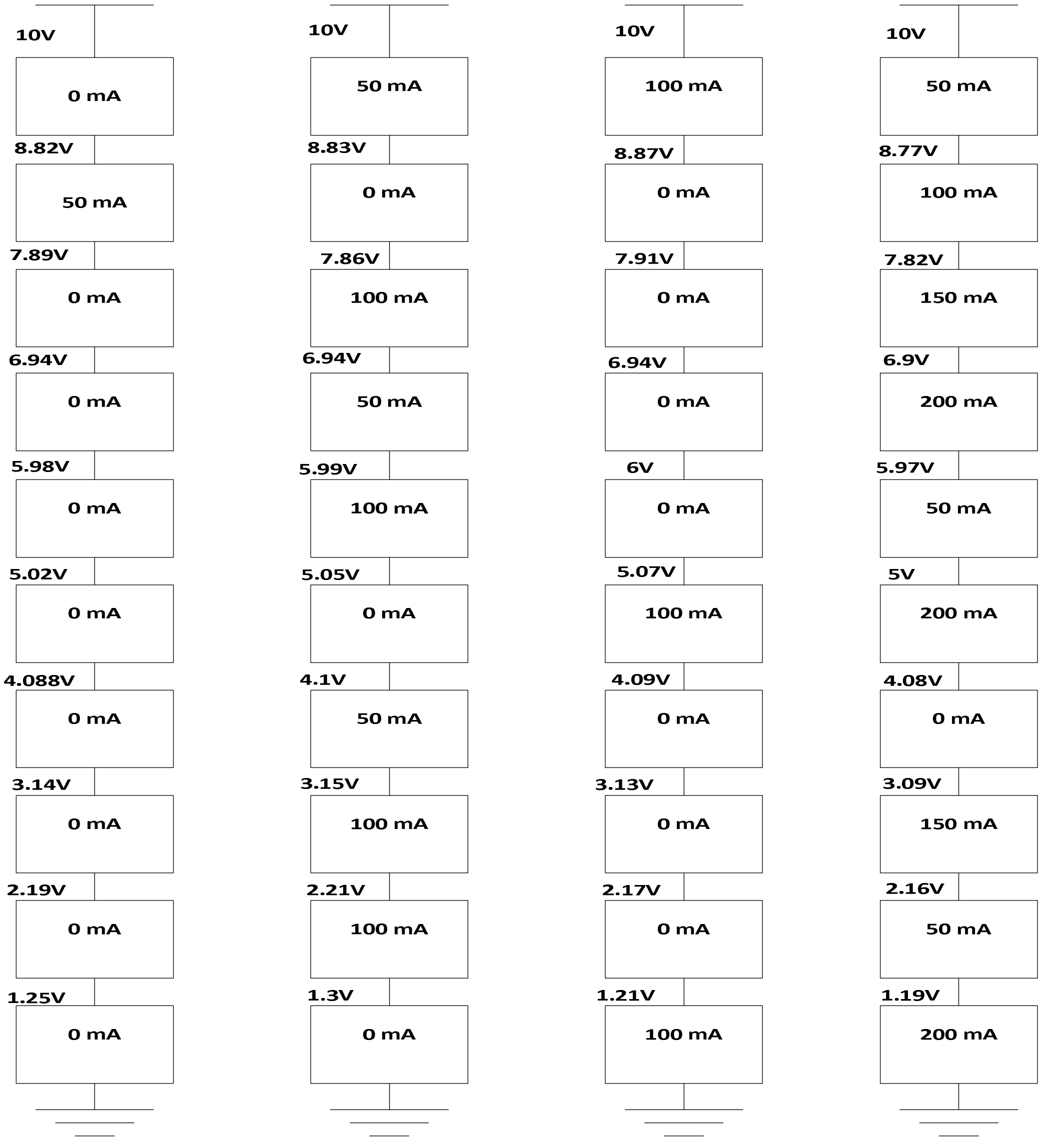
(b)

Efficiency with varying conventional load (left) and stacked load (right). In Figure (b), X-axis indicates relative imbalance (%) between the domains.

hplp

High-Performance Low-Power

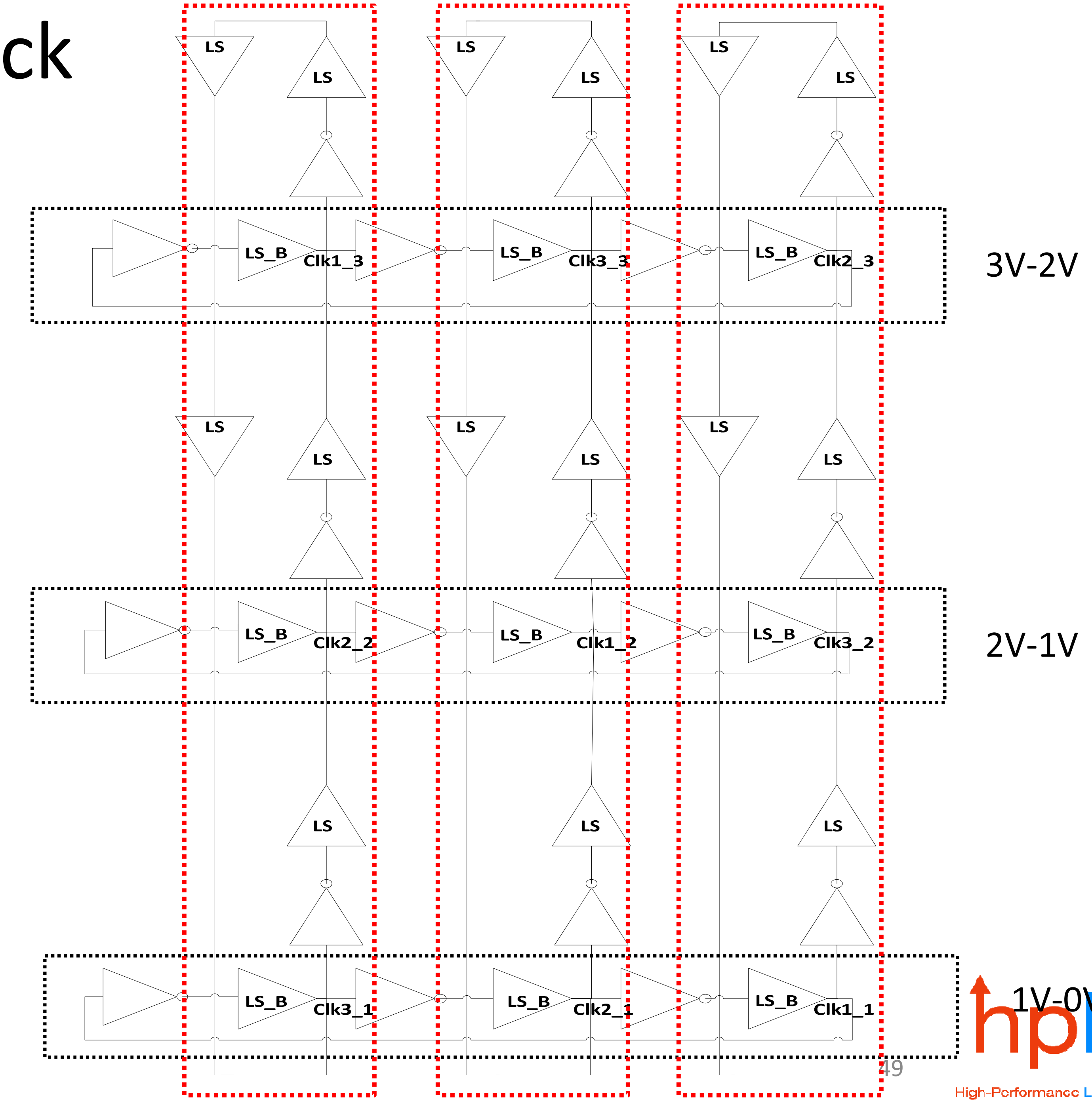
3D IC scaling: more stacked layers



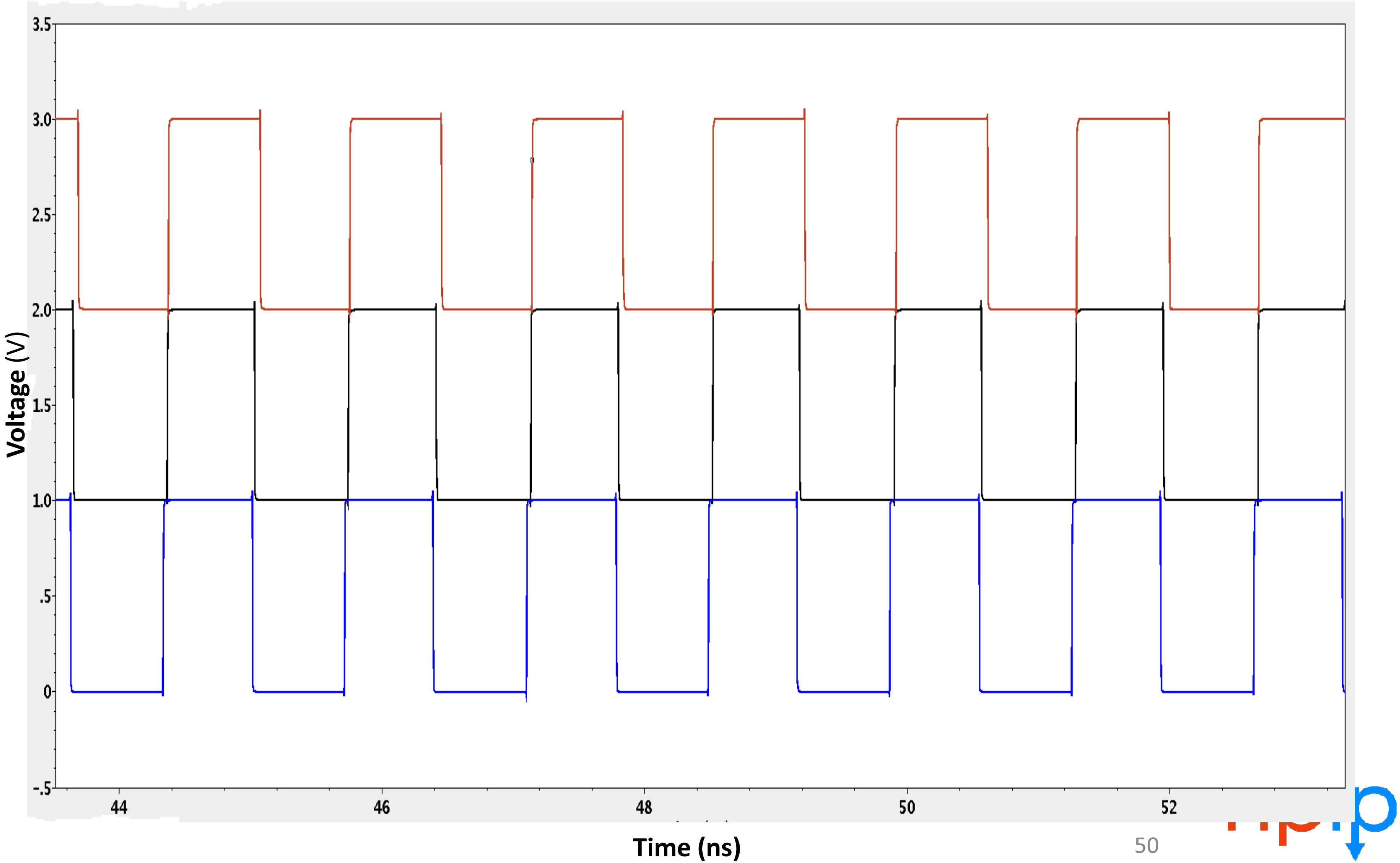
SC Clock

Cross coupled rows of oscillators lock in Phase

Horizontal and vertical row buffer-inverter delay equivalent



Phase-Frequency Locked Clock



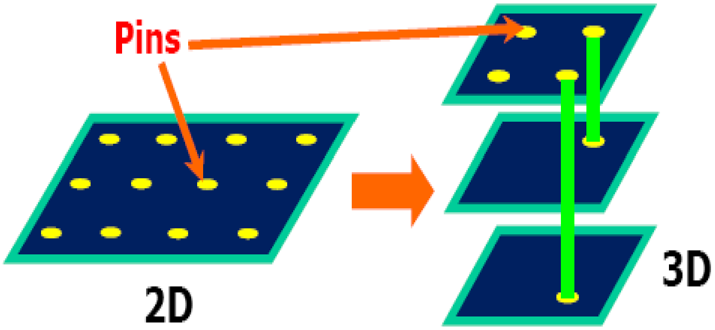
High-Performance Low-Power

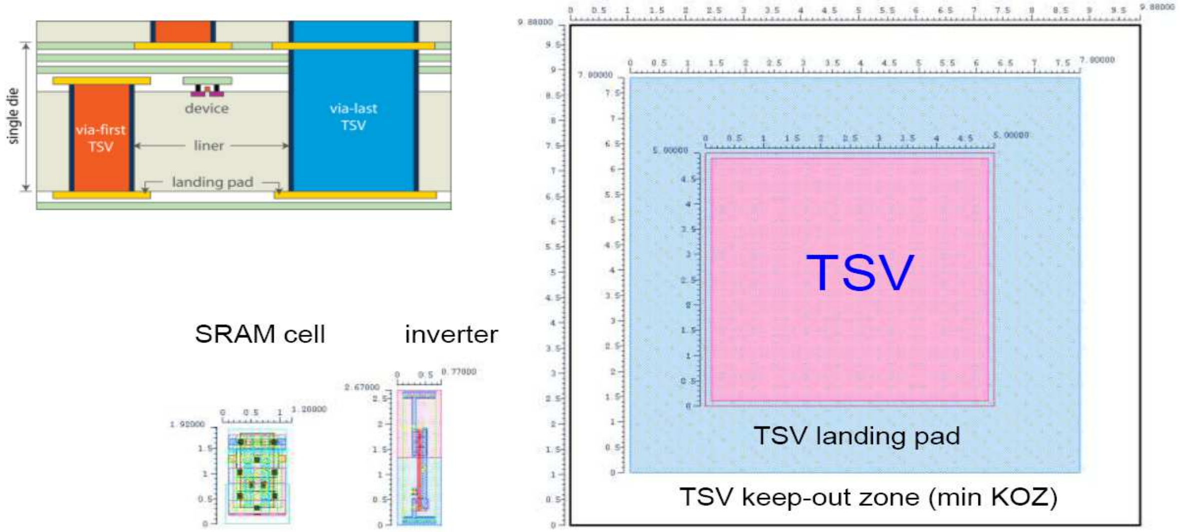
3D IC Power Delivery- TSV Bottleneck

- Smaller footprint : Fewer Power bumps

- Big P/G TSVs to deliver power to all the stacked layers, causing congestion

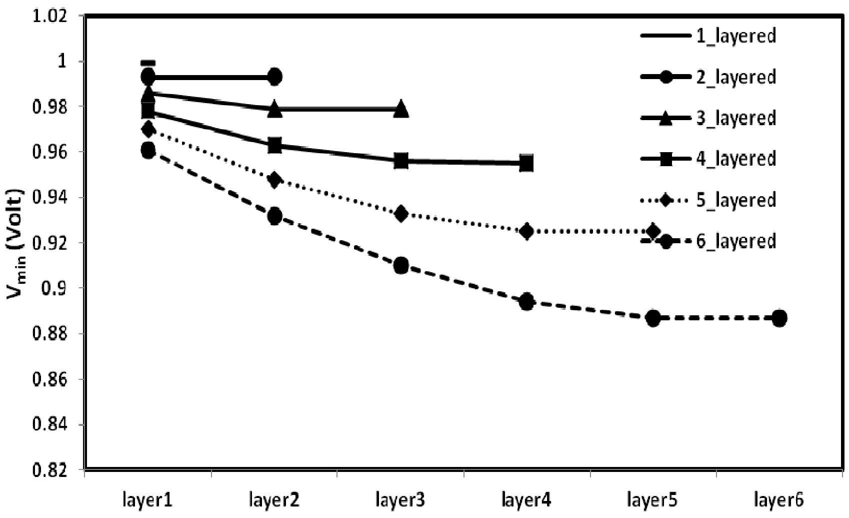
- TSVs contribute to IR drop, reducing supply rail integrity

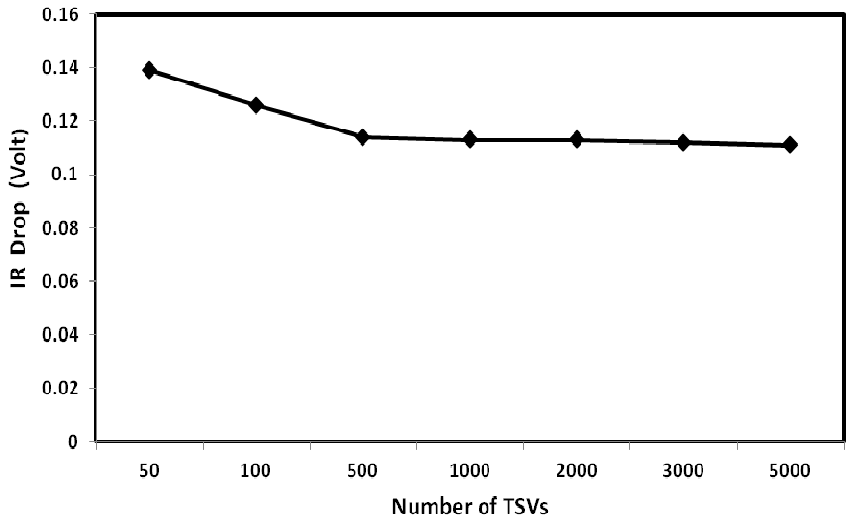


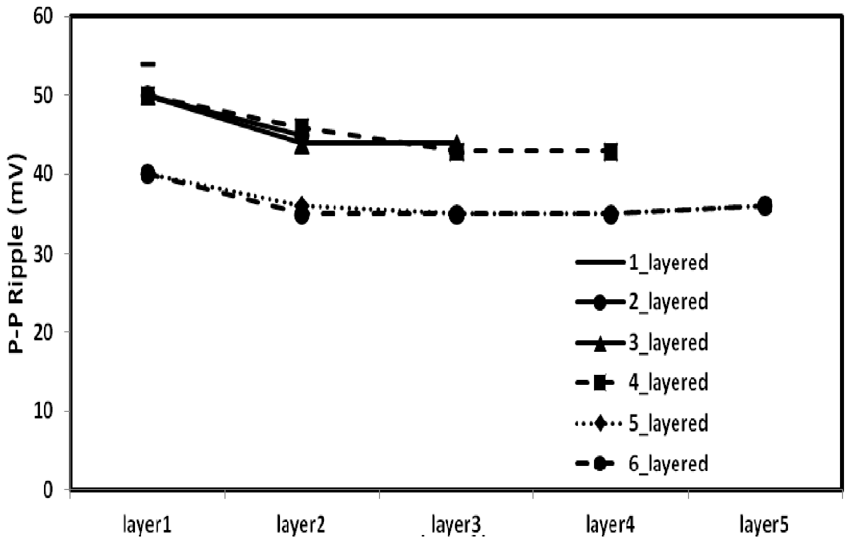


Source: Sung Kyu Lim, “ 3D IC Circuit Design with Through-Silicon-Via : Challenges and Opportunities “, GTCAD Laboratory

TSV Trade-Offs





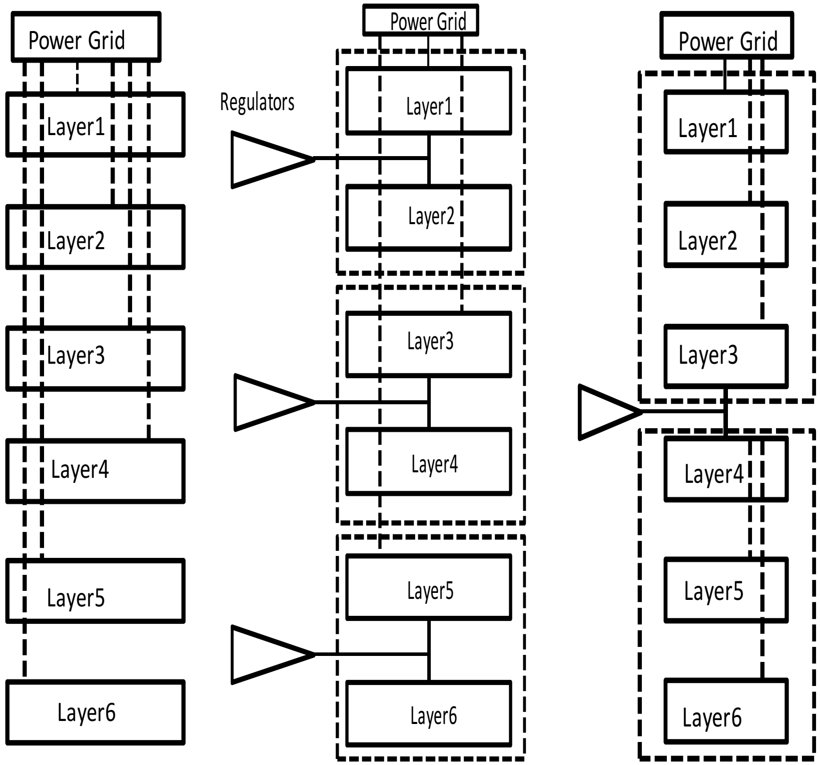


Change in IR drop with increasing number of 3D IC layers and TSV density

P-P Ripple improves due to TSV/3D layers capacitive effect

hplp
High-Performance Low-Power

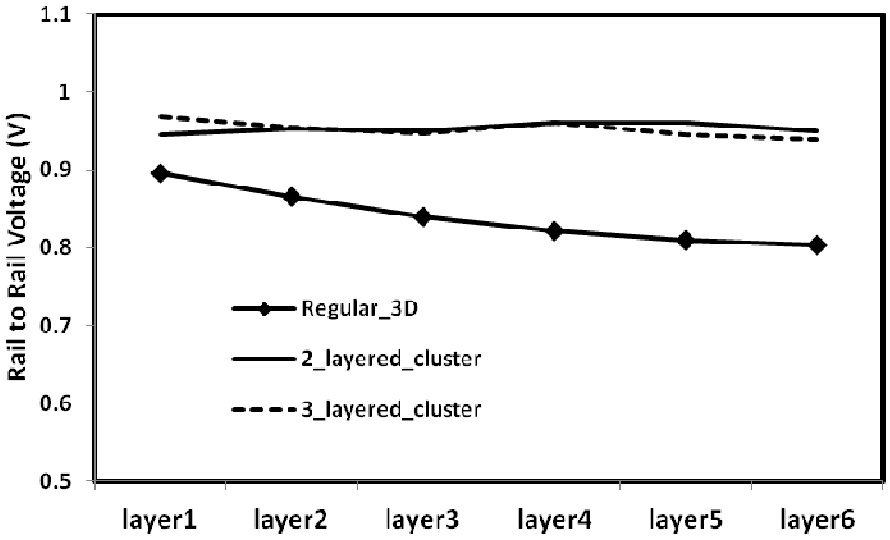
Clustered Voltage Stacking



Conventional 3D:
Max No of TSVs

2_Layered 3D
Least No of TSVs

3_Layered 3D



Tradeoffs between TSV count and regulators

Summary : Voltage Stacking in 3D IC

- 3D IC power delivery wall: at constant voltage cubic increase in power/current but only quadratic area/pins

- Voltage stacking can help break wall: quadratic current and linear voltage

- Implicit regulation + explicit for imbalance

- Clustered Voltage Stacking

Acknowledgments

- Funding from SRC, NSF, AMD, Intel
- Collaborators: Kaushik Mazumdar, Runjie Zhang, Kevin Skadron
- IEEE CAS
- Universities of Victoria, British Columbia, Washington and Portland State