Many voltage domains enabling an energy efficient graphics processor in 14nm

Rinkle Jain, Pascal Meinerzhagen, Vivek De

October 19 2018

Radio Circuits Integration Lab, Intel Corporation
Hillsboro, OR

Outline

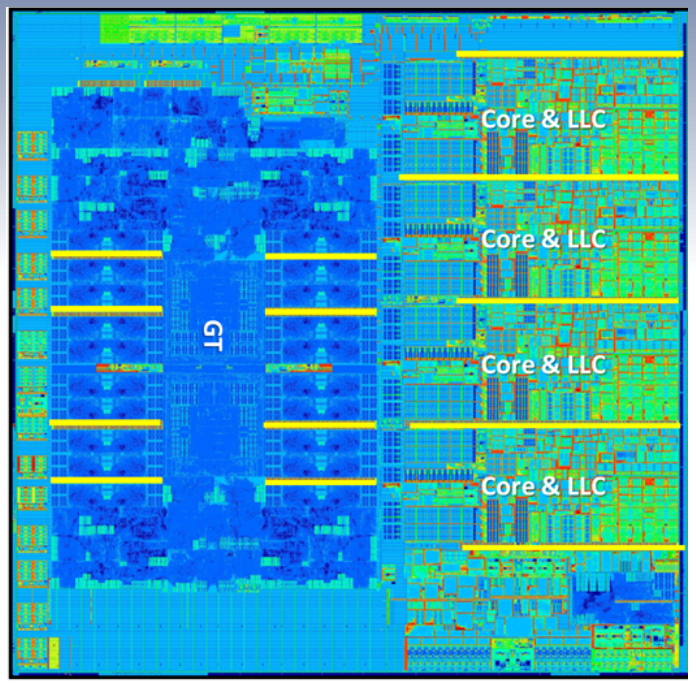
State of the Art

Fine grained DVFS: Costs and benefits

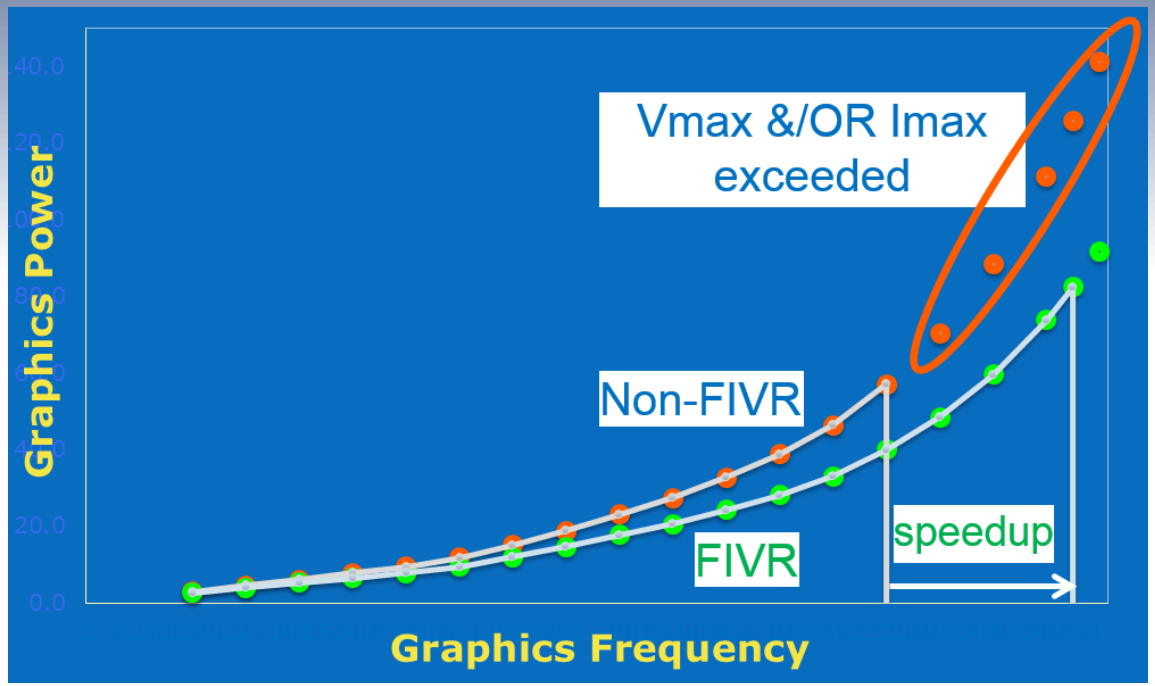
Co-design with load

Summary and Conclusion

The SoC Side of the Story - the FIVR



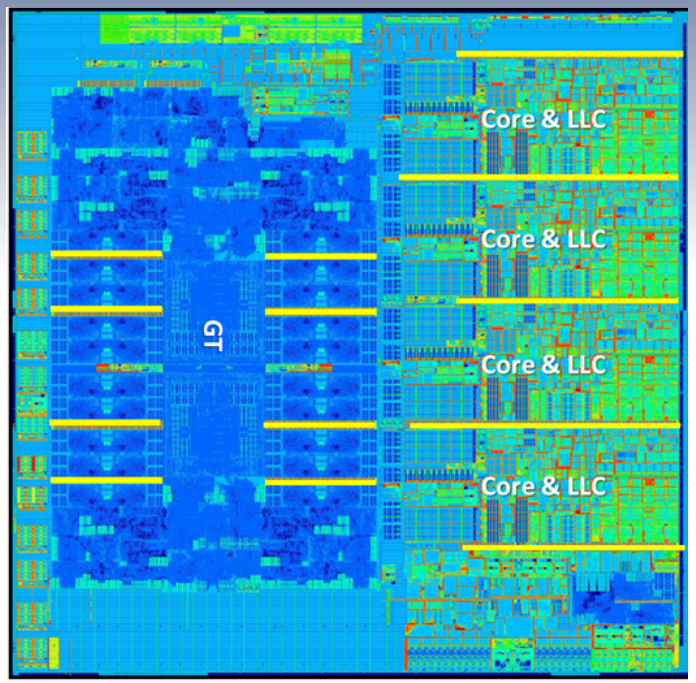
[Burton et.al APEC '14]

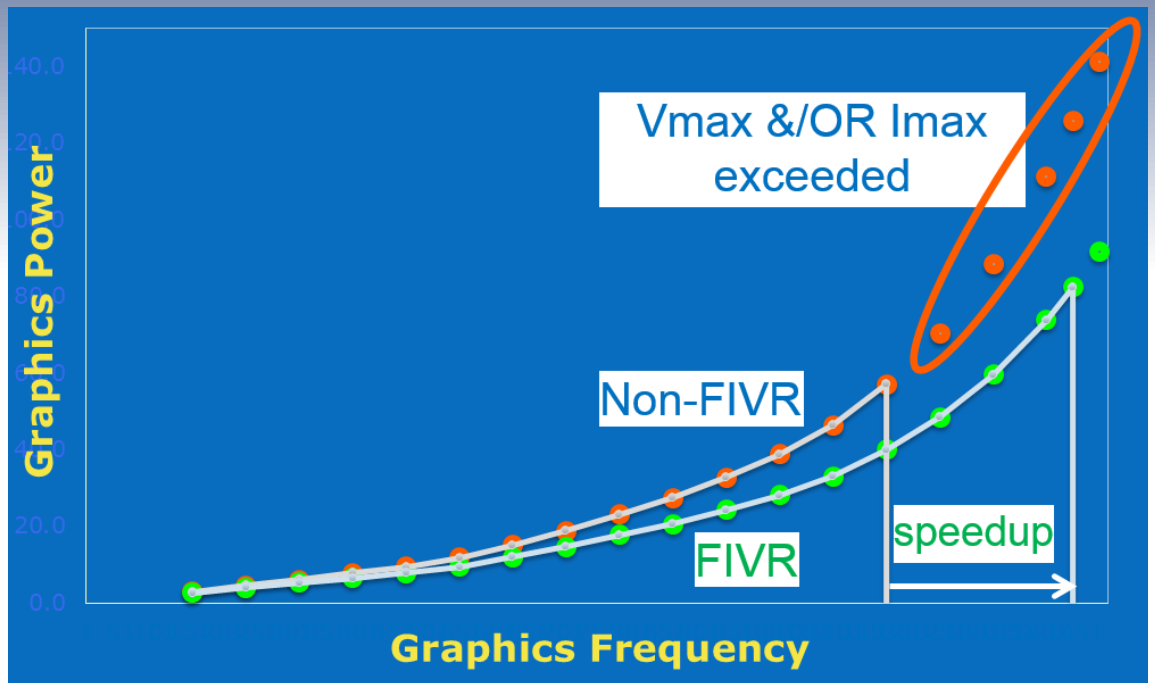


[Kurd et.al ISSCC '14]

- Faster state transitions by 25%, higher performance per watt
- Overall idle power slashed by 20x, battery life improvement by $>$ 50%
- BOM savings helped all segments

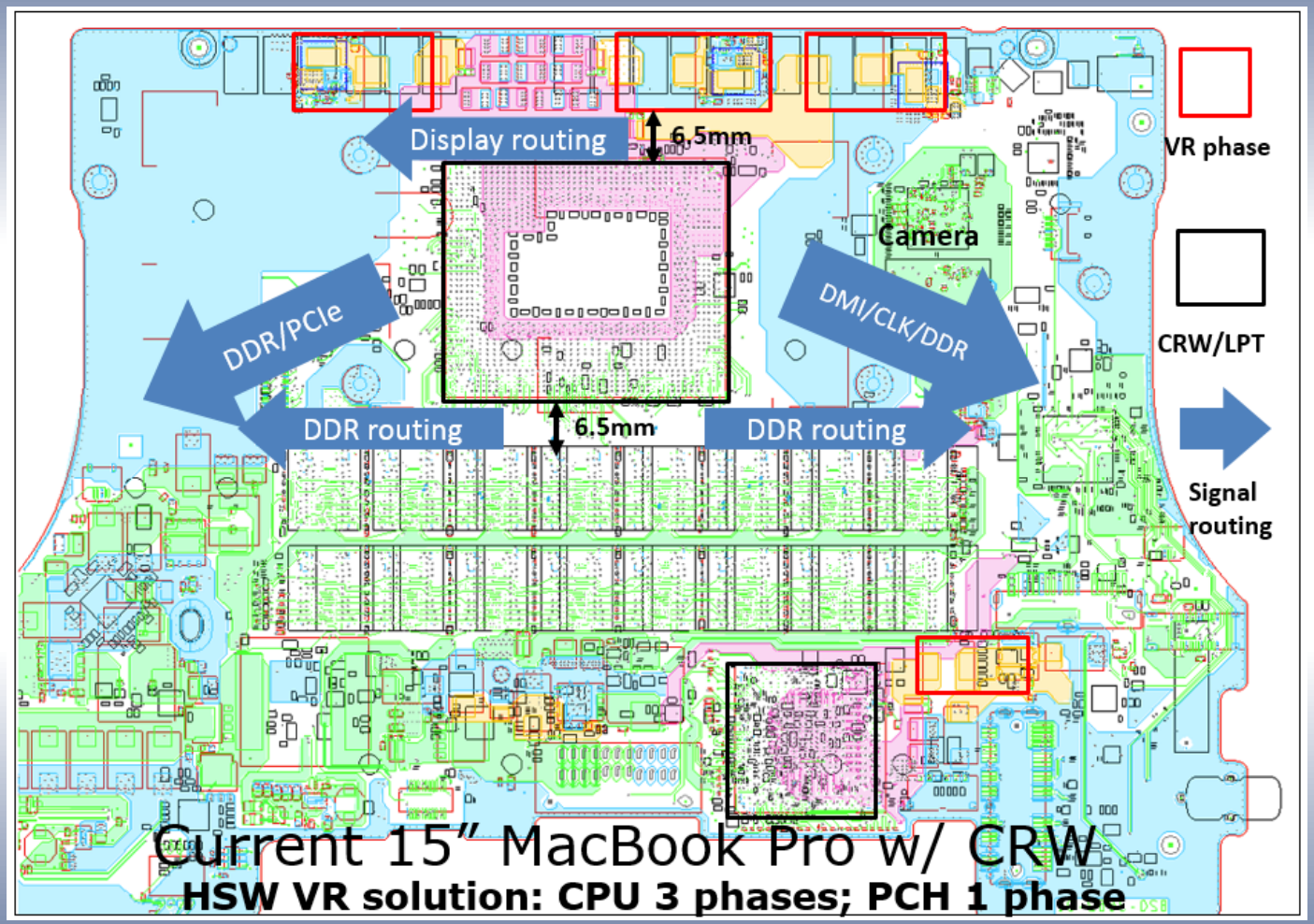
The SoC Side of the Story - the FIVR



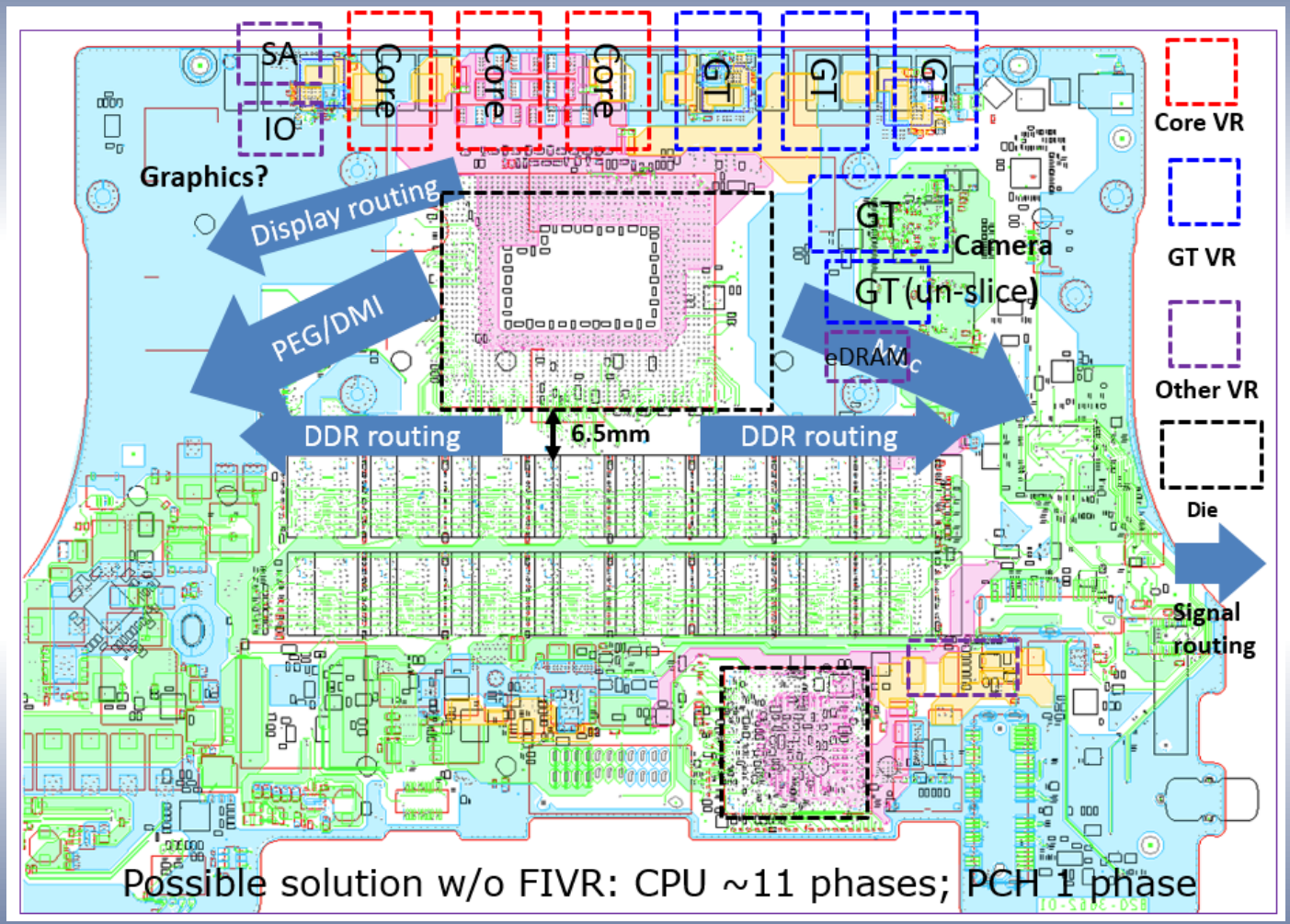


Segment	Key Challenge	What helped the most
Servers	Bump Imax	I_{in} reduction
Laptops	Board Area, Iccmax	High bandwidth/frequency
Desktops	Performance	High bandwidth

The Platform Perspective



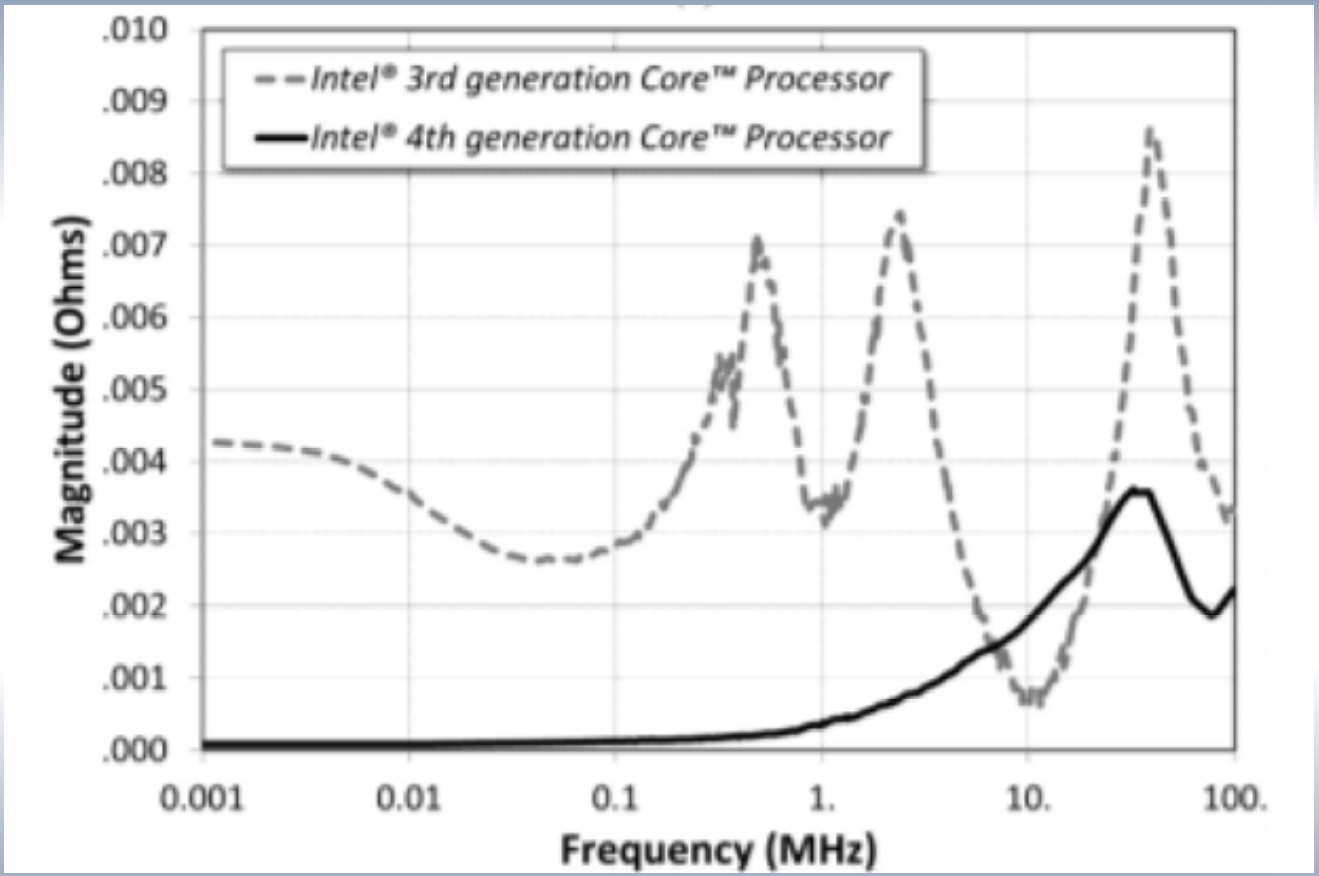
The Platform Perspective



SoCs inherently require several voltage rails

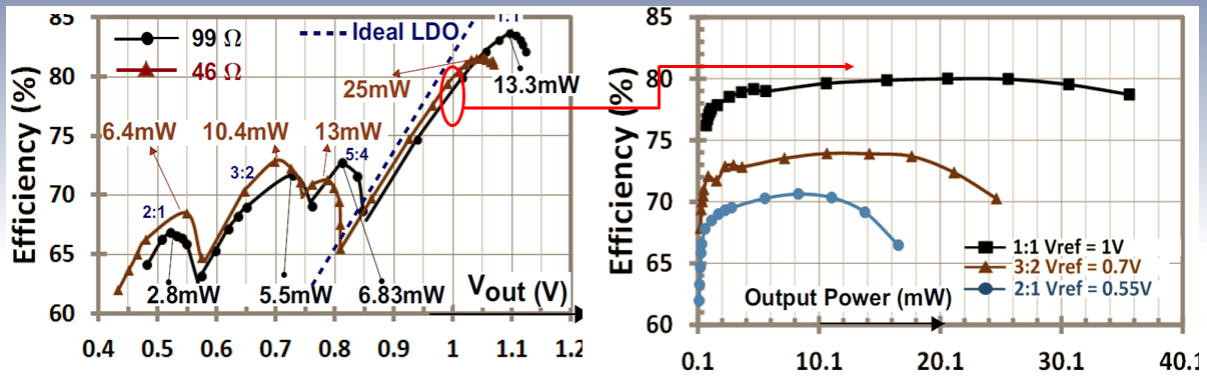
Input rail consolidation simplifies power delivery significantly

Loadline improvements

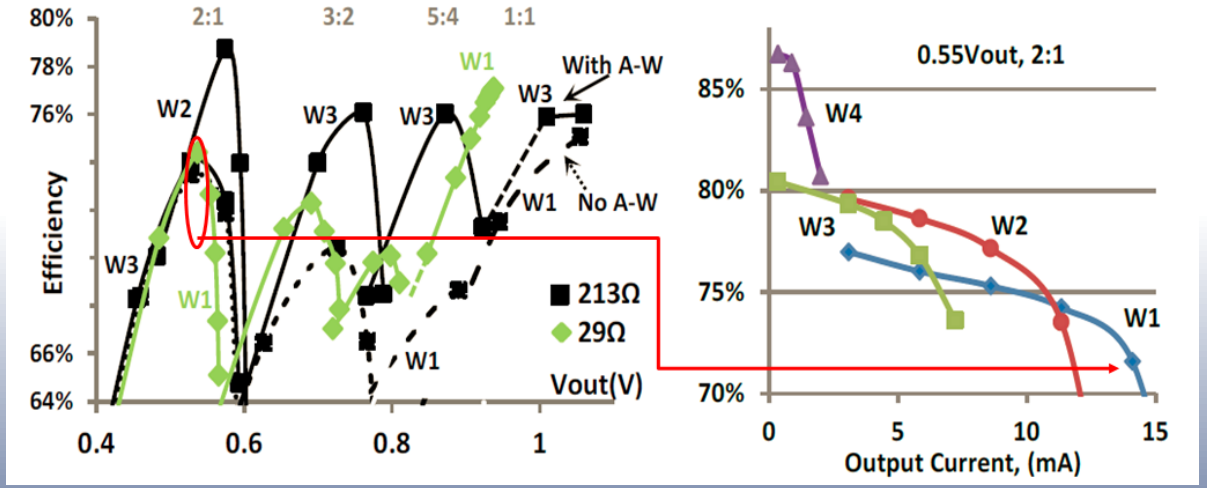


Translates to battery life, area and/or performance benefits

MIM-based Switched Capacitor VR



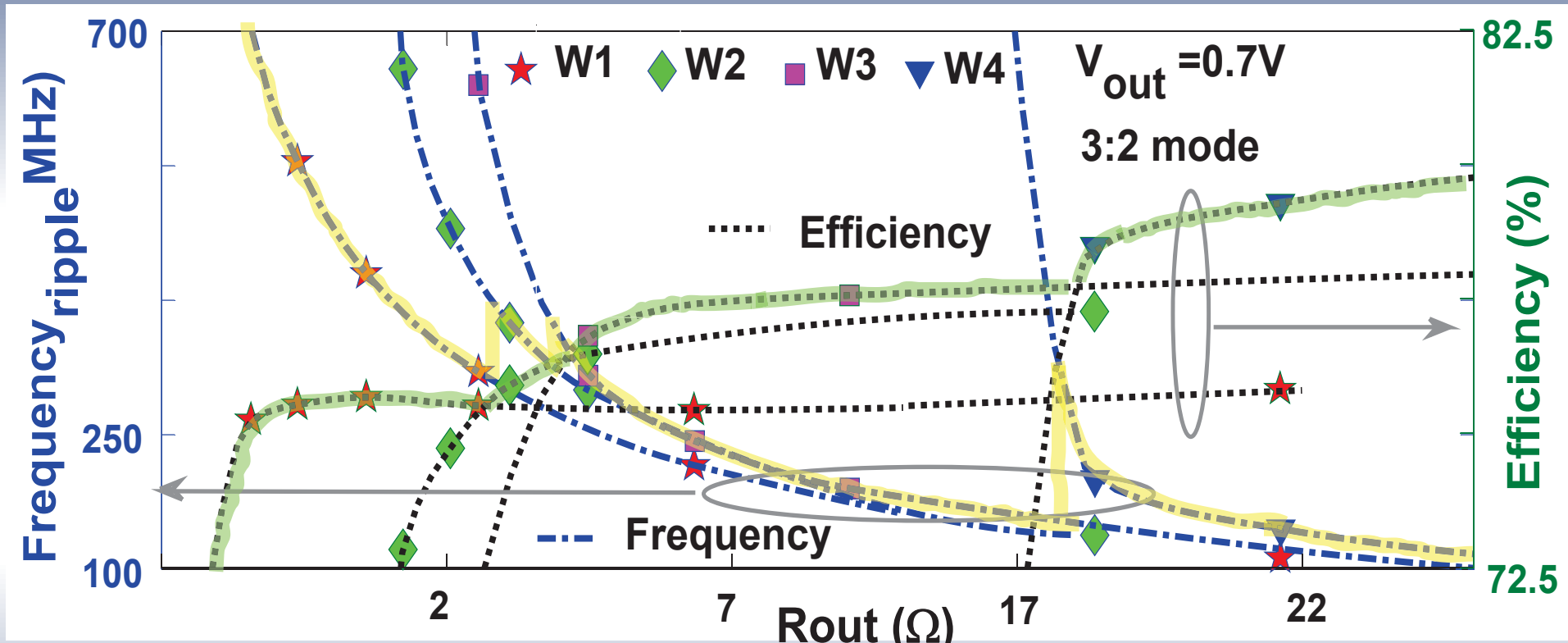
[R.Jain et al. JSSC '14]



[R.Jain et al. JSSC '15]

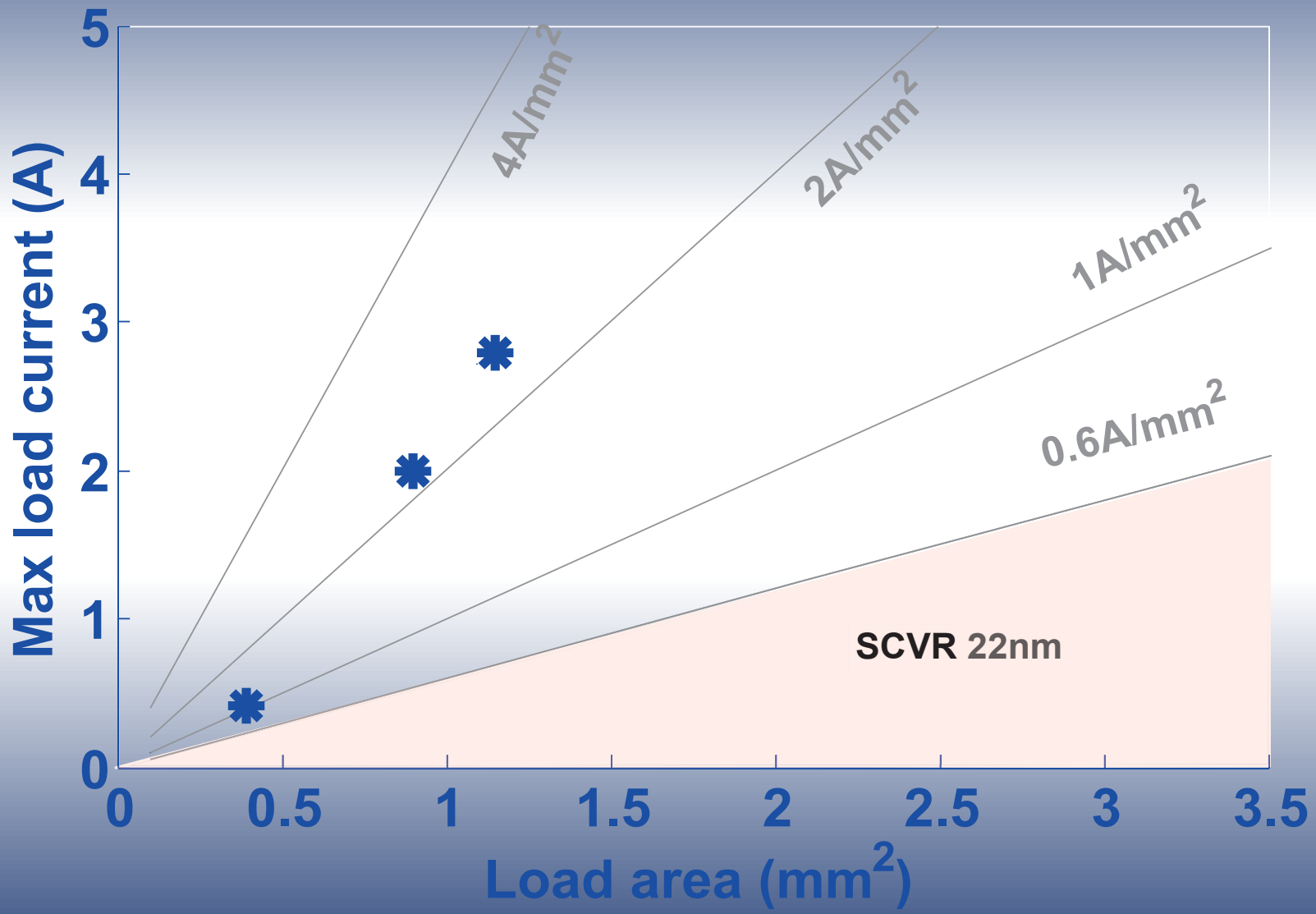
- Regulation, few ns response, low area overhead, 880mA/mm^2

Proposed Control Law for Adaptive Widths

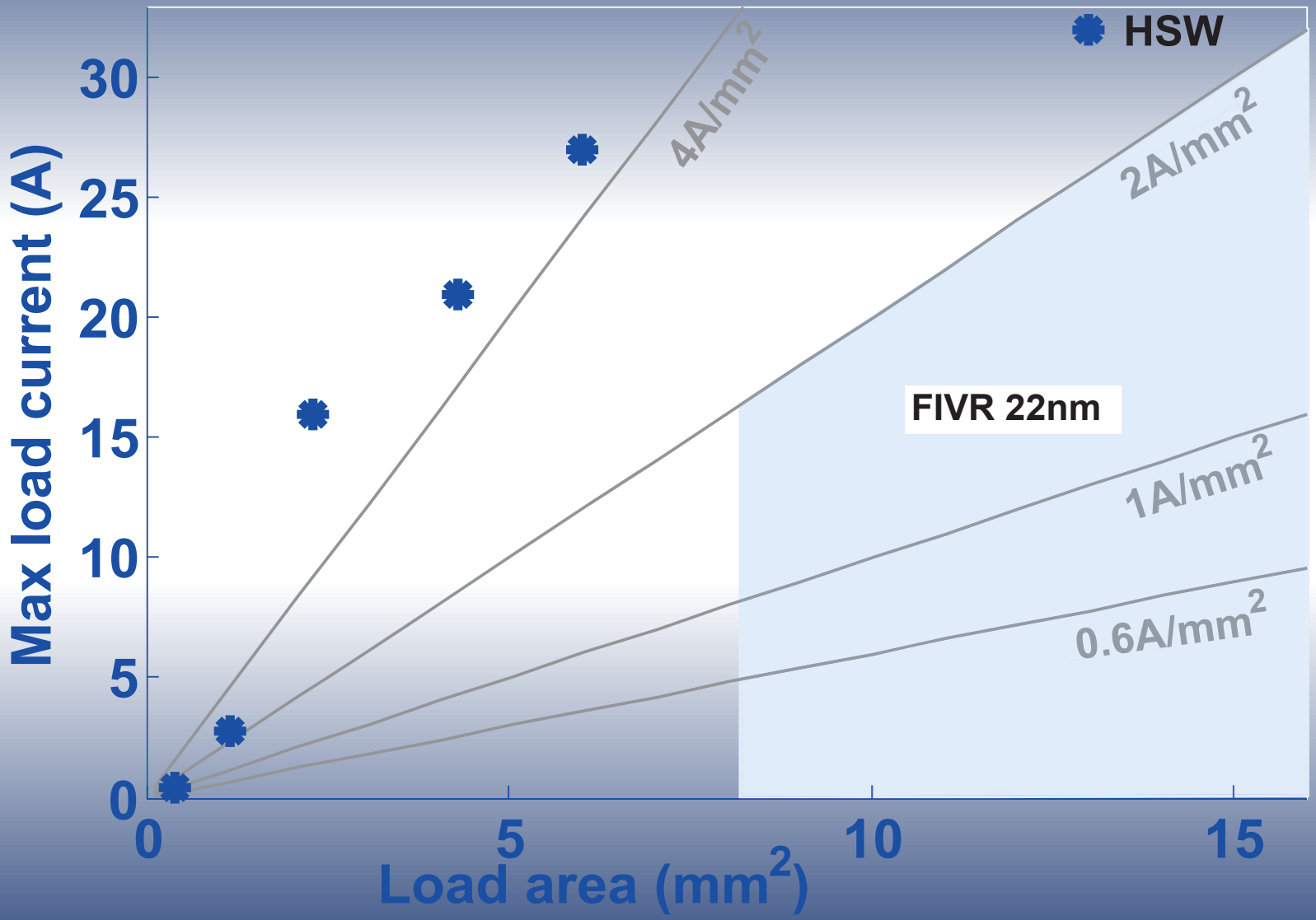


- fsw $<$ Fth$_W$ $= \frac{bW'}{a}$ implies higher efficiency at W' (W'=W/n)
- fsw is a good indicator of low voltage and light load conditions

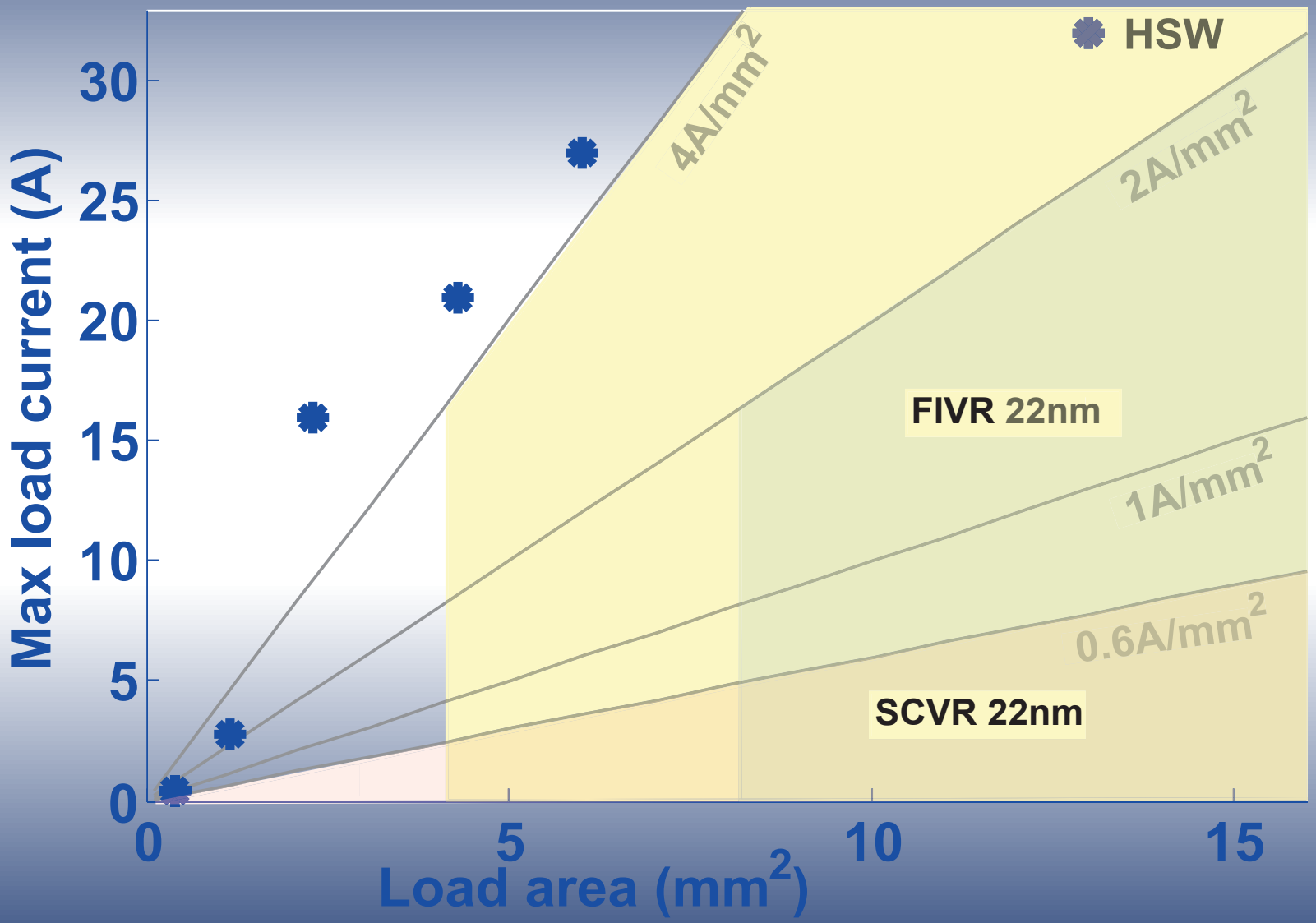
Current density & domain area-power trends



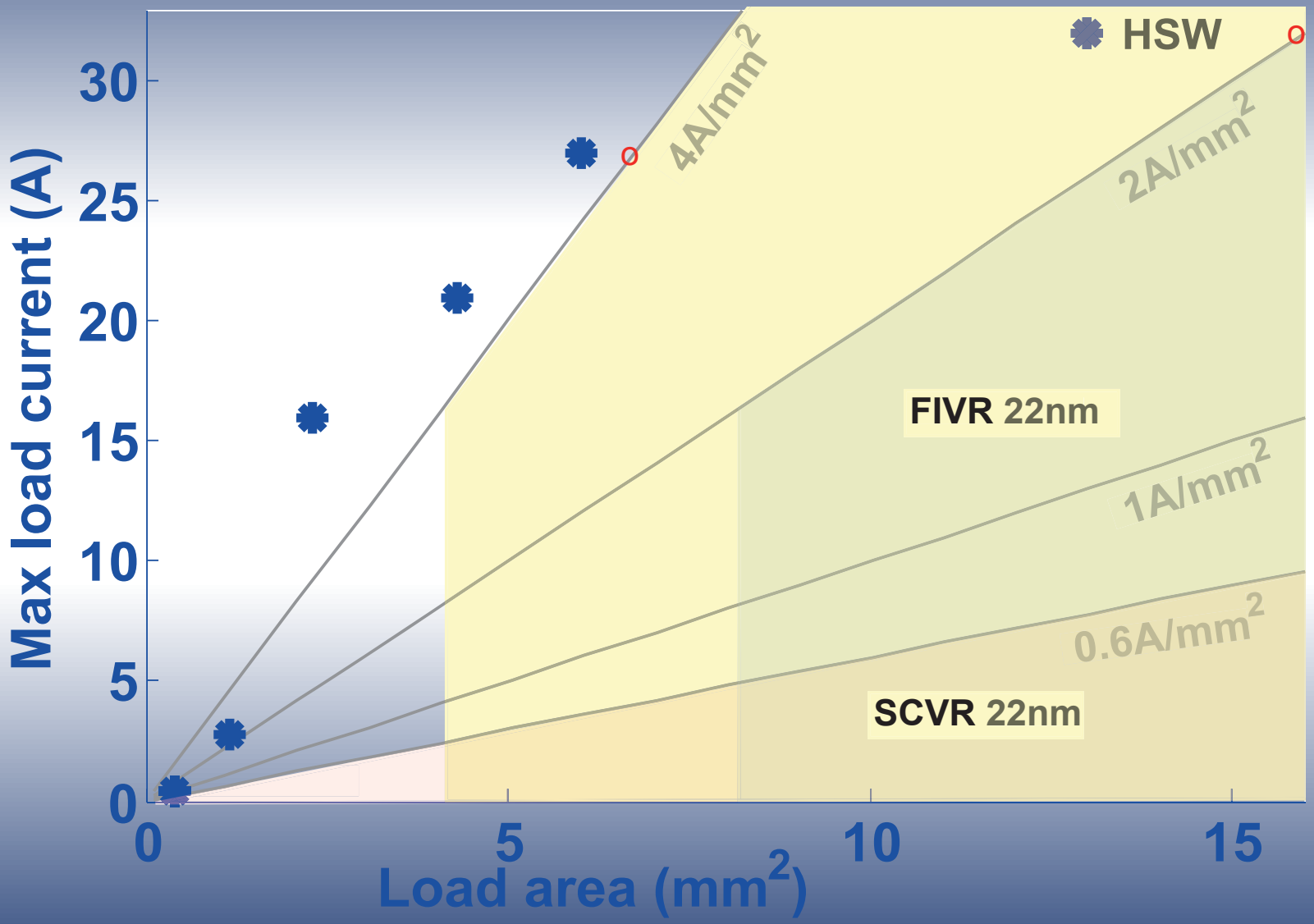
Current density & domain area-power trends



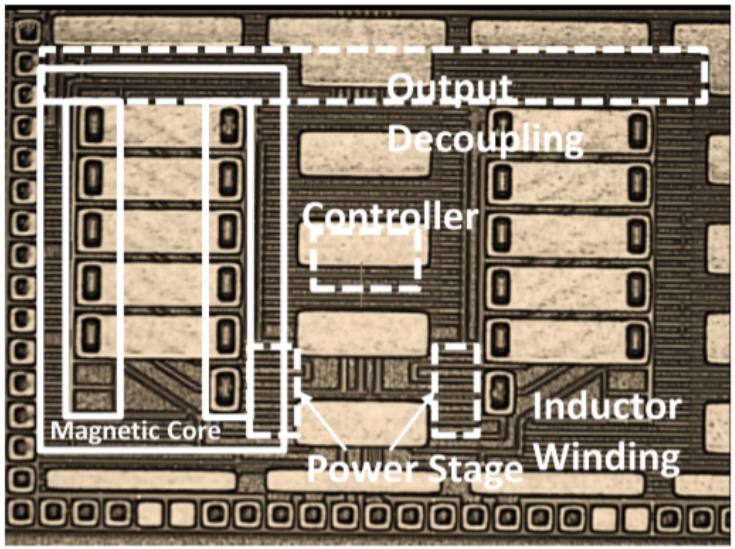
Current density & domain area-power trends



Current density & domain area-power trends

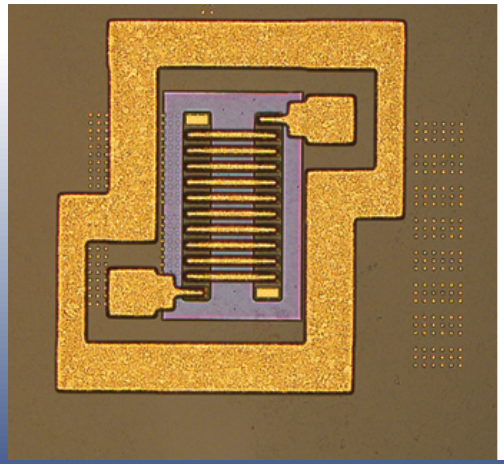


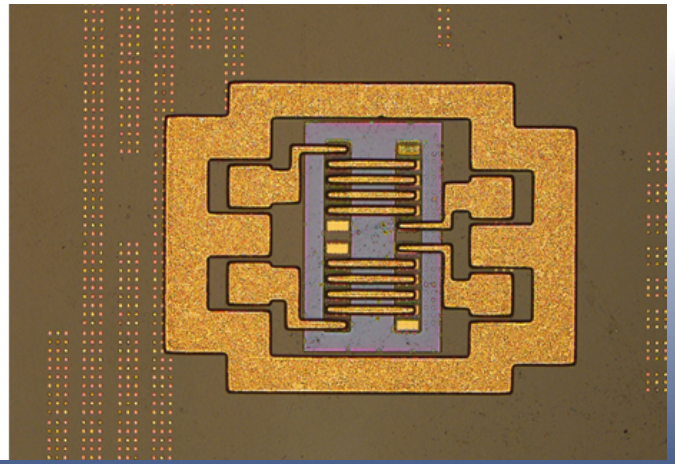
Buck converter with thin-film magnetics



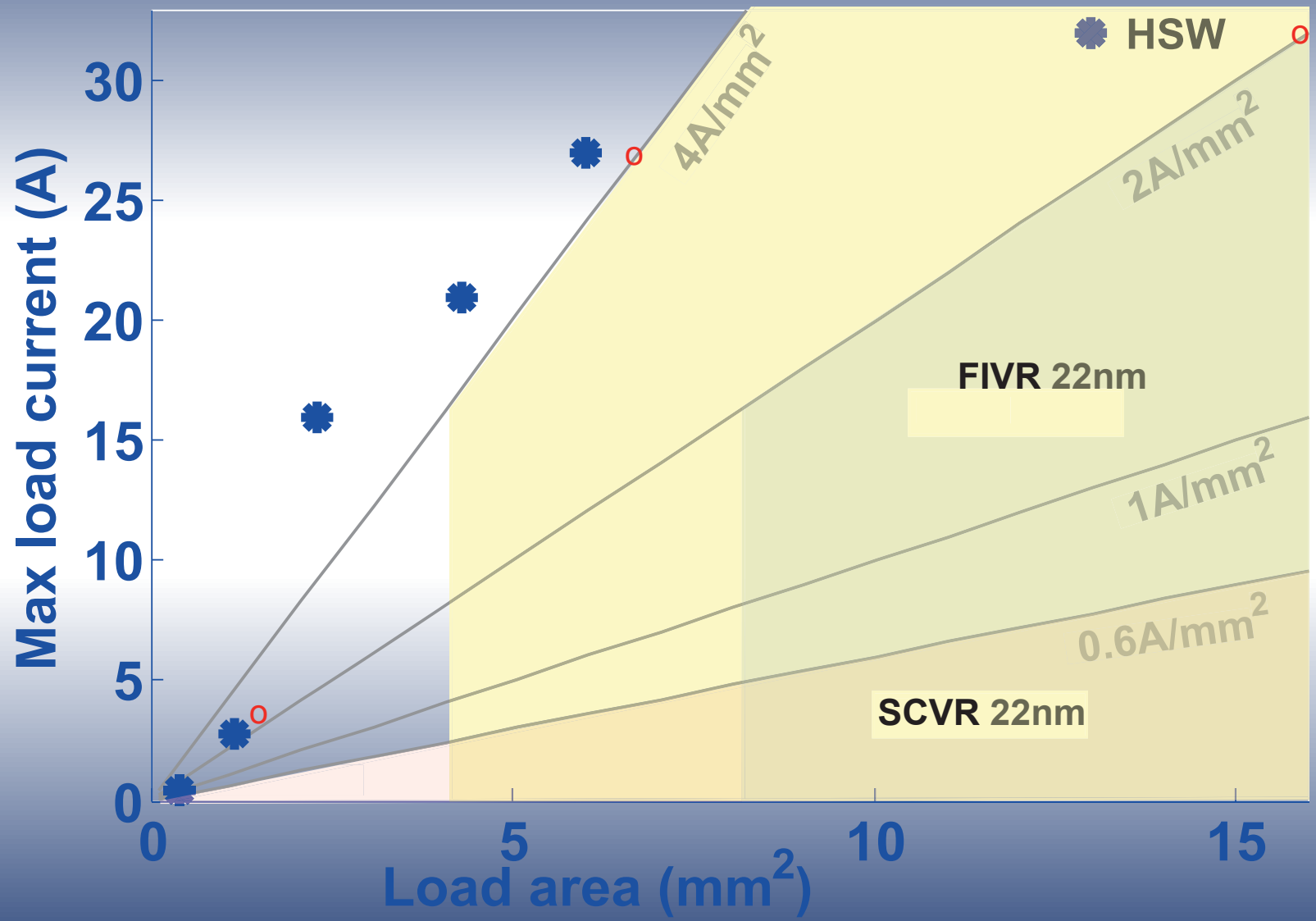
Characteristics Table	
Input Voltage	1.5V
Output Voltage	0.4V-1.15V
Load Current	0-400mA
Inductors	On-Die Non-Planar Solenoid with Magnetic Core & External Discrete Inductor in Membrane Probe Card
Input Decoupling	~ 1nF
Output Decoupling	~ 5nF
No of Phases	2
Switching Frequency	100MHz
Controller	Digital

H. Krishnamurthy et al. JSSC 2017

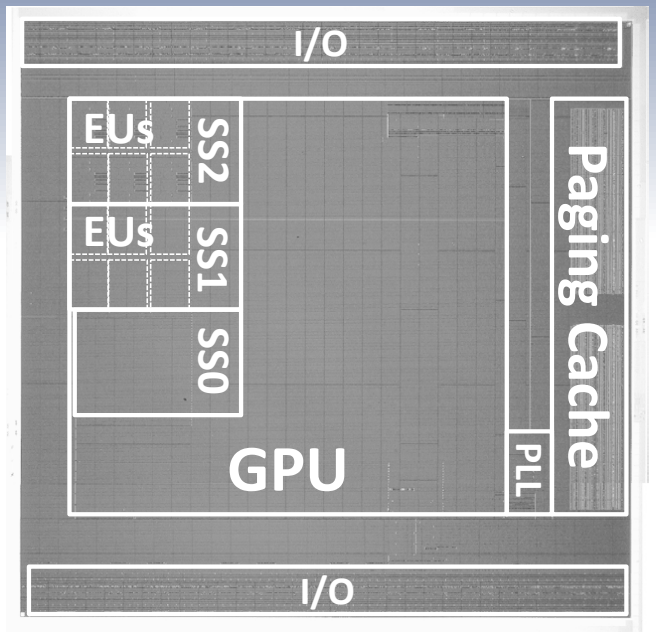




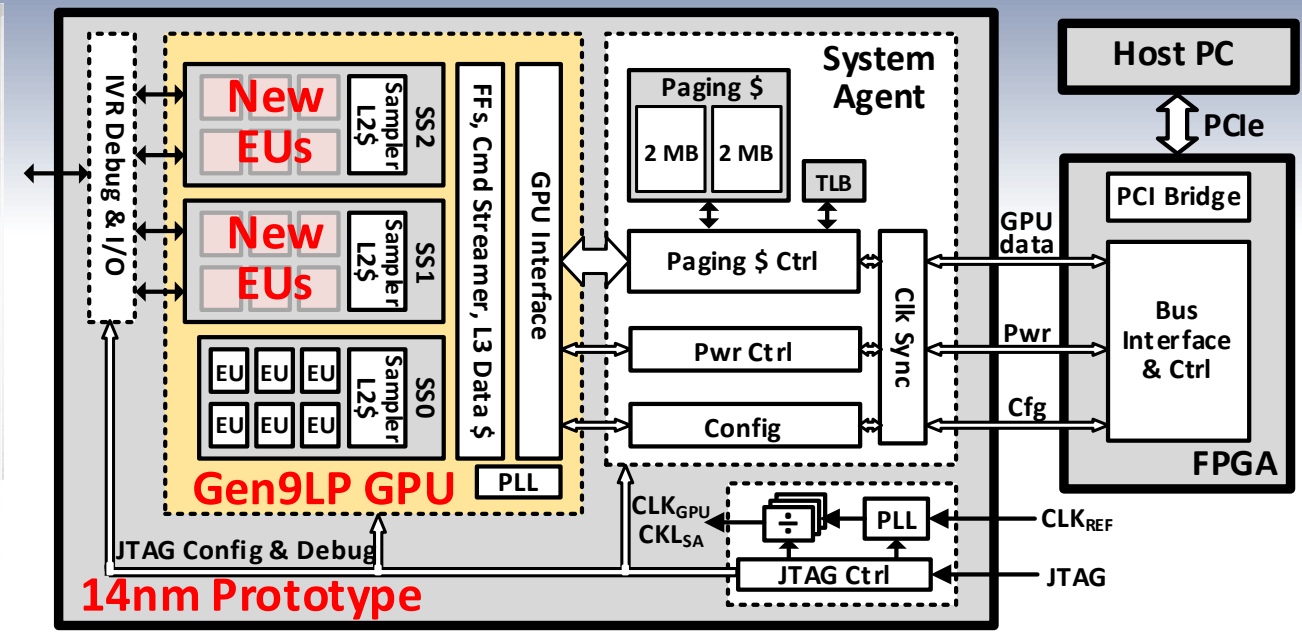
Current density & domain area-power trends



14nm Graphics Processor

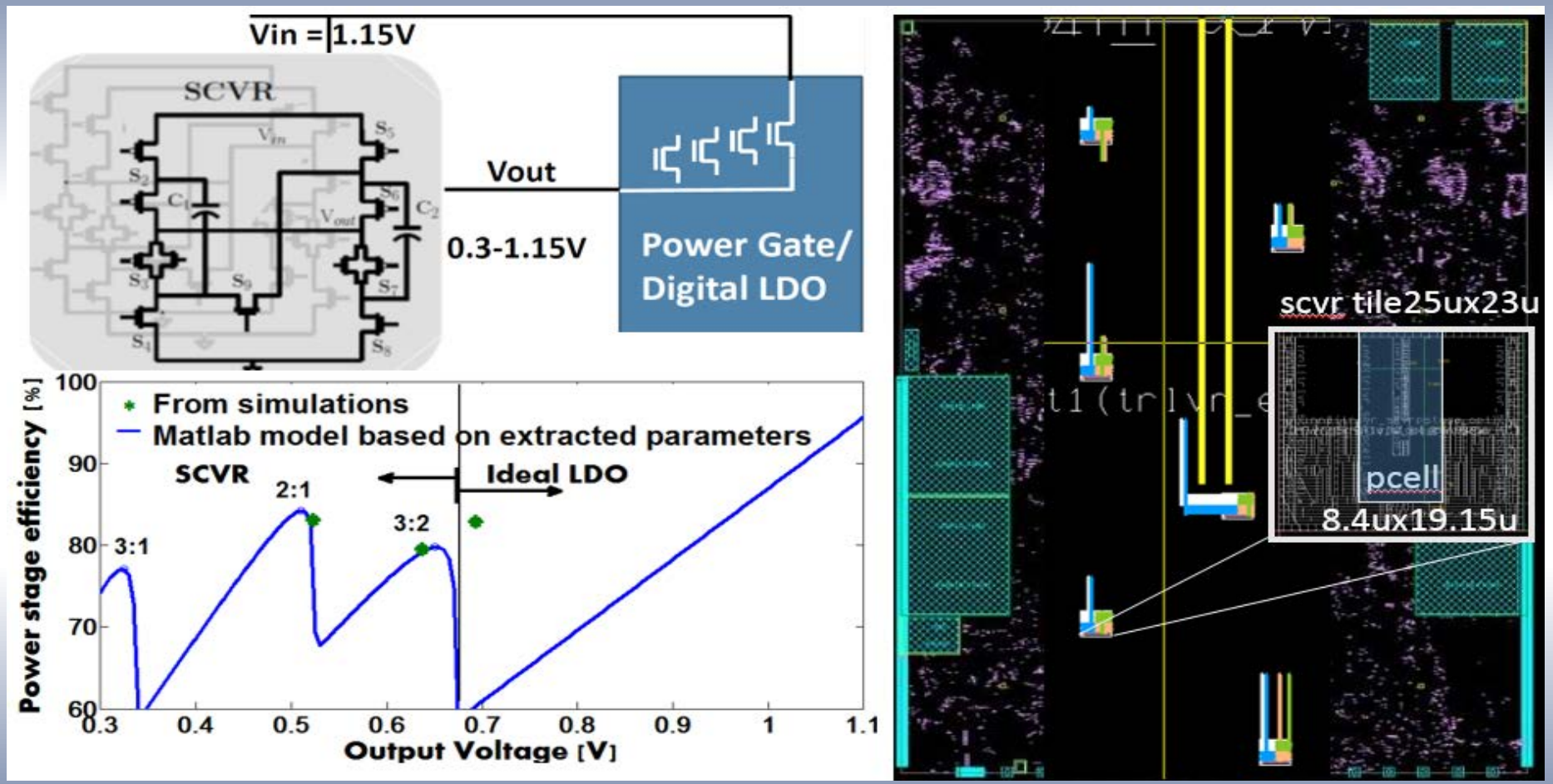


8.0 x 8.0 mm²



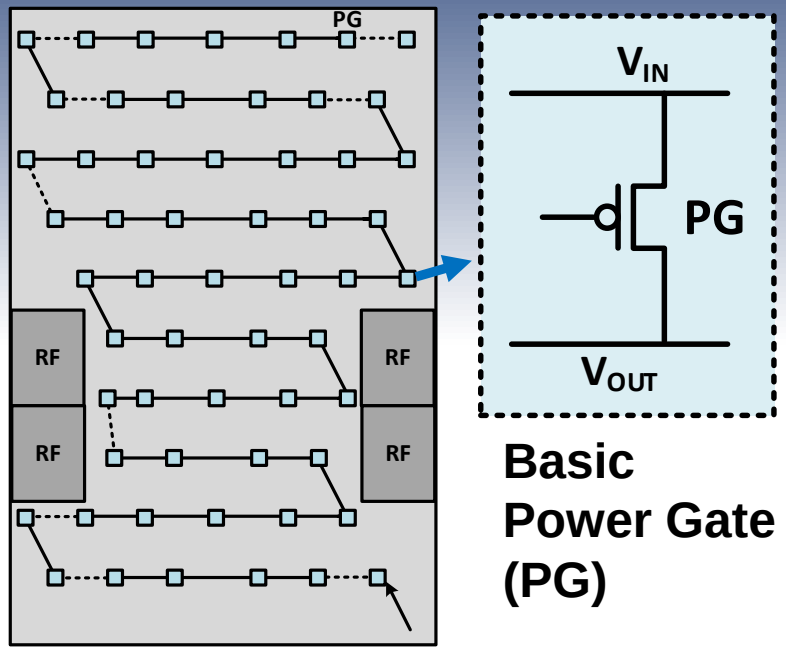
New execution unit (EU) has a dedicated integrated voltage regulator

Switched Capacitor Voltage Regulator

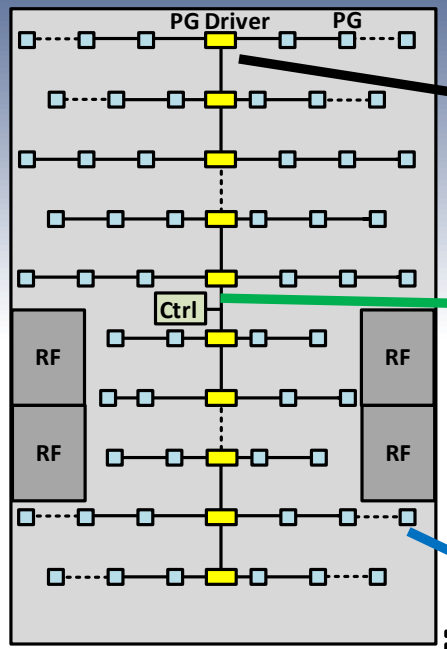


APR friendly design, 6 distributed tiles with area overhead of 1.1% (680umx520um)

Modified EU: Digital LDO (DLDO)



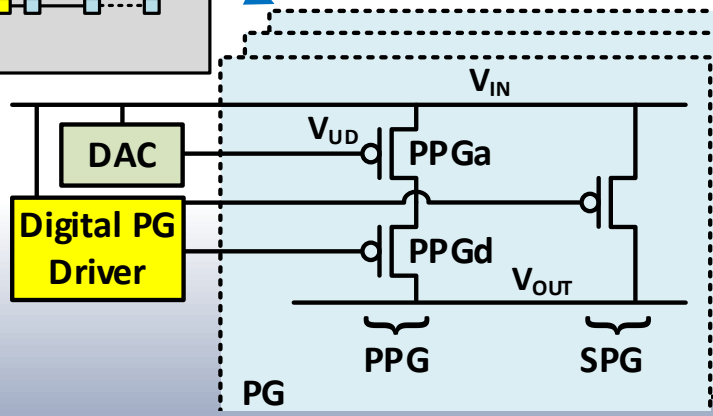
Basic Power Gate (PG)



Driver

Controller

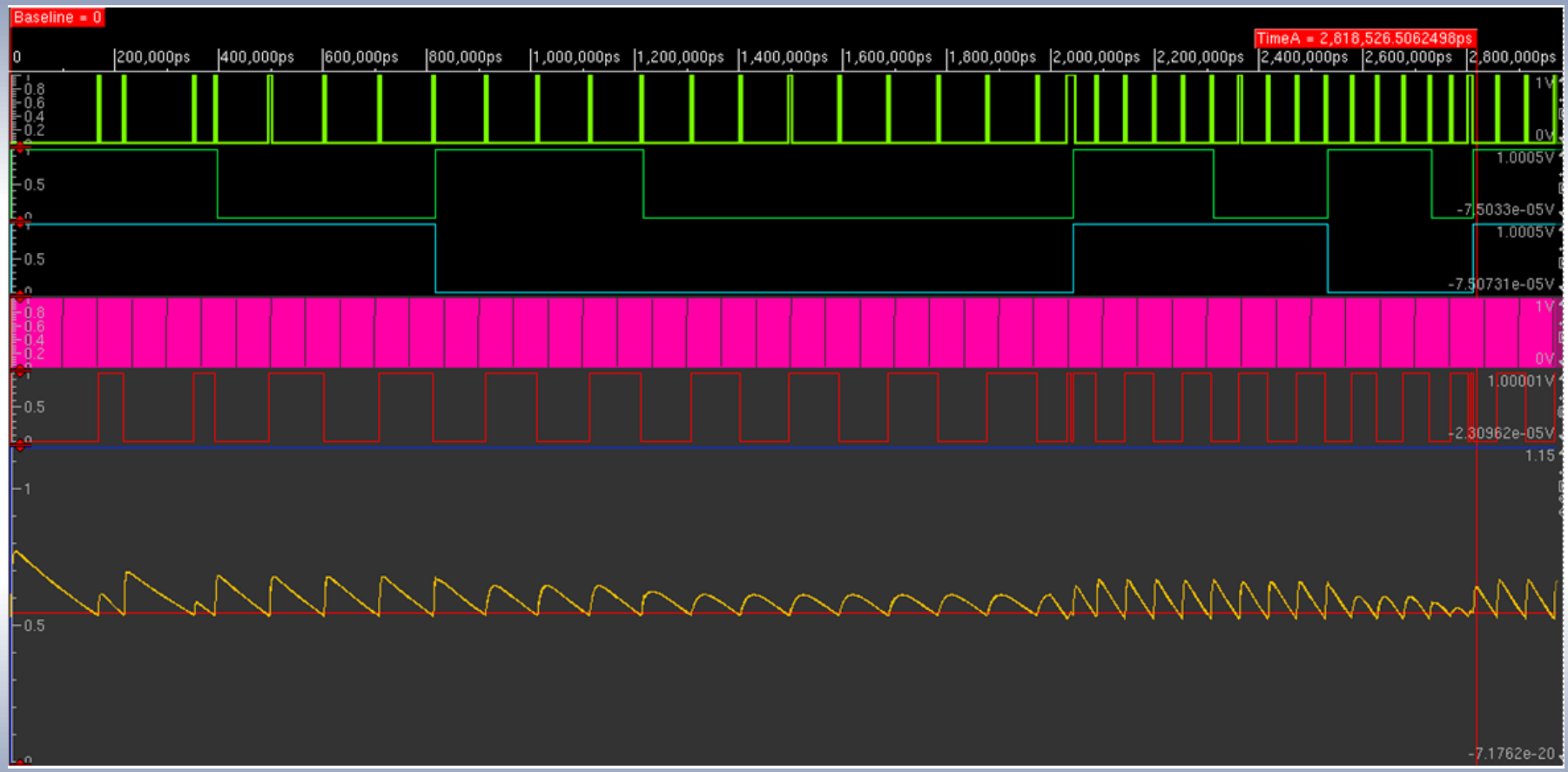
DLDO PG



Area overhead: 0.5%

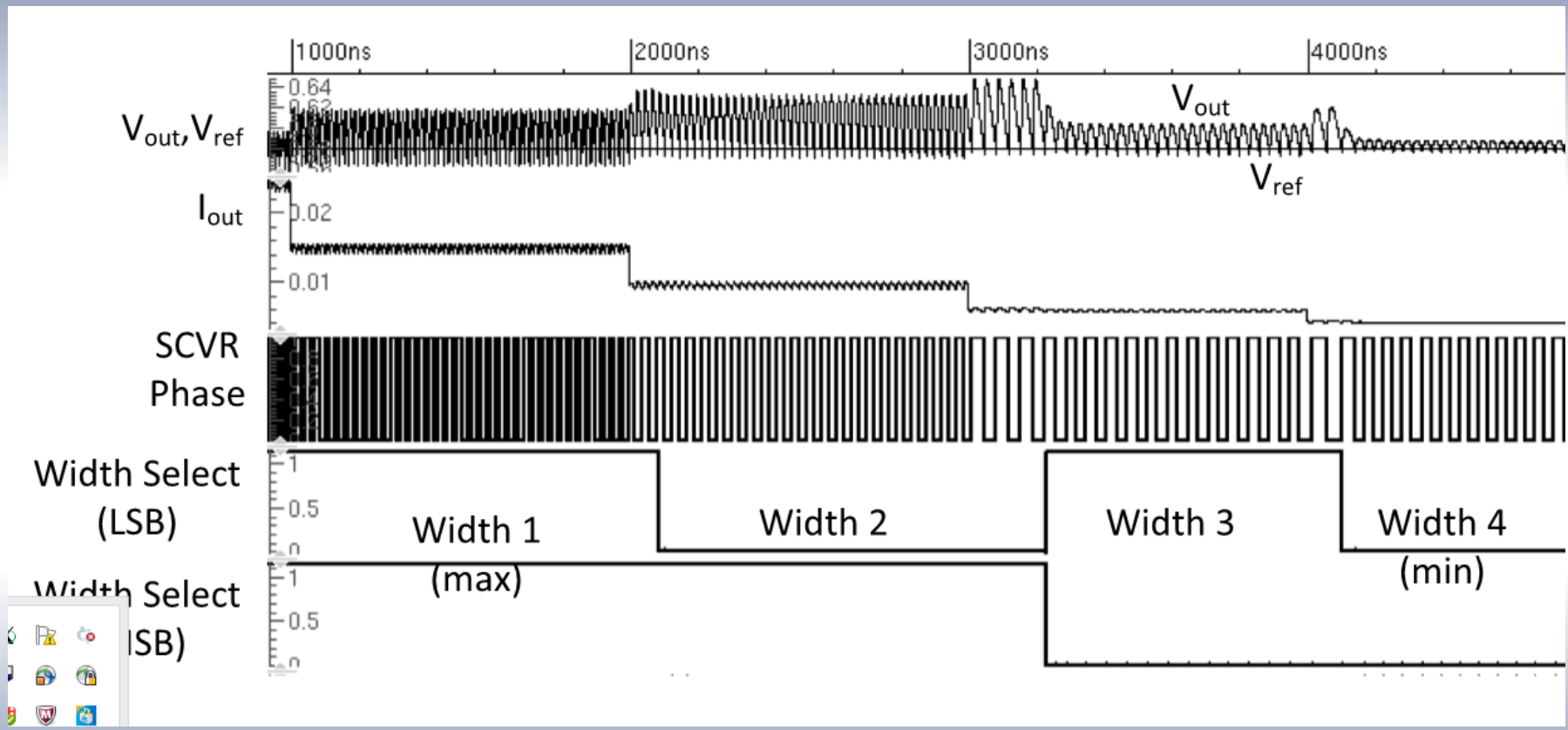
Digital LDO demonstrated in *S.Kim et al. JSSC '15*

Adaptive Width Simulations



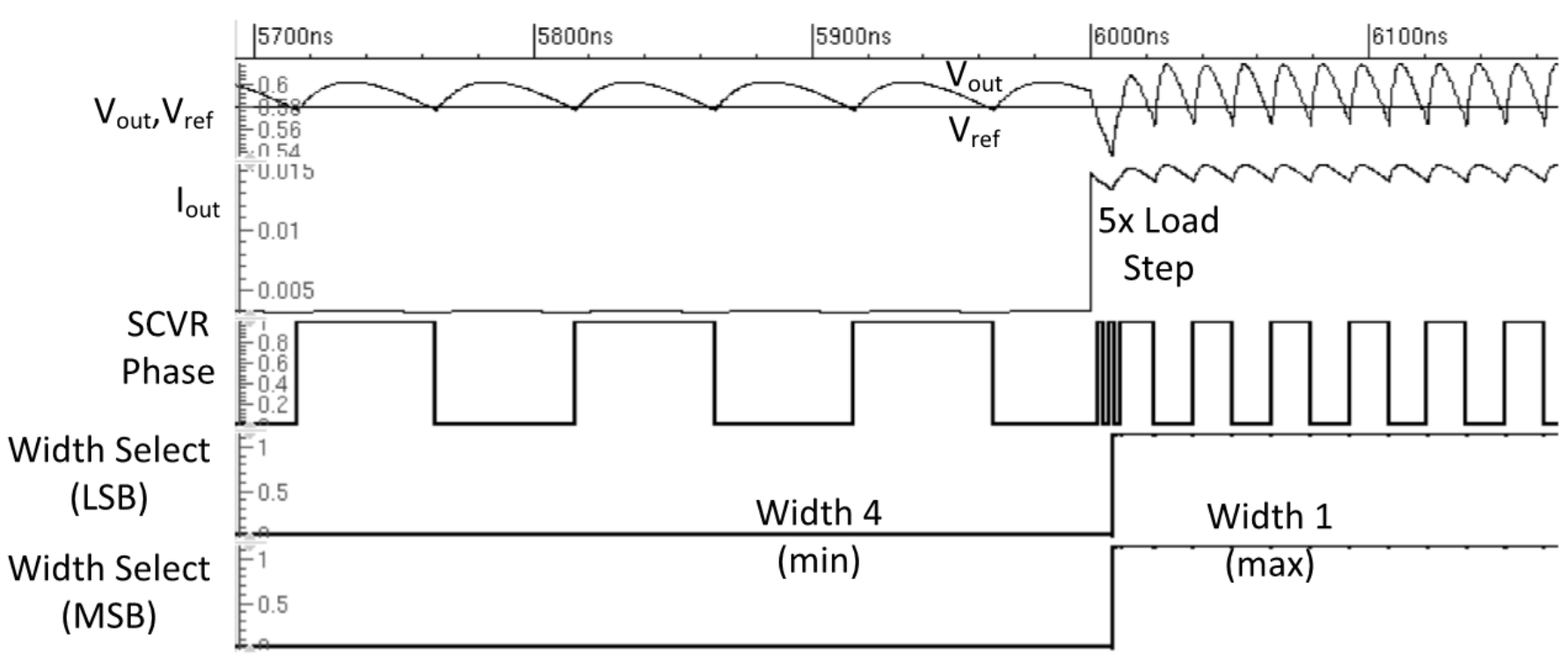
- Fast and stable adaptation

Adaptive Width Simulations



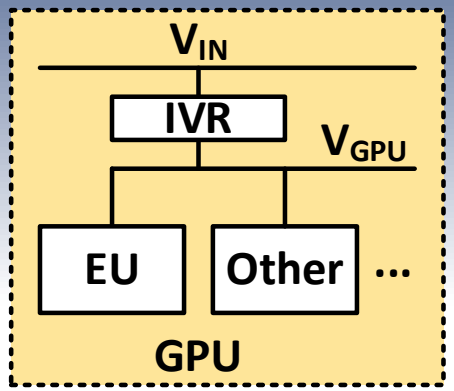
- Fast and stable adaptation

Adaptive Width Simulations



- Fast and stable adaptation

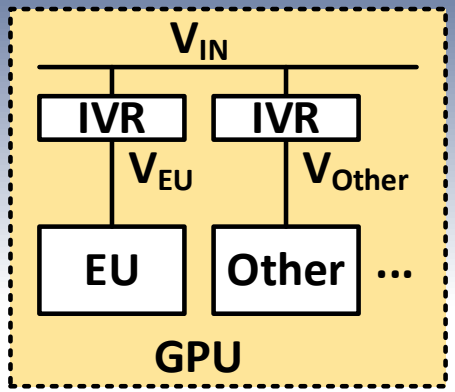
Fine grained DVFS



Shared voltage rail

All blocks at high V/F

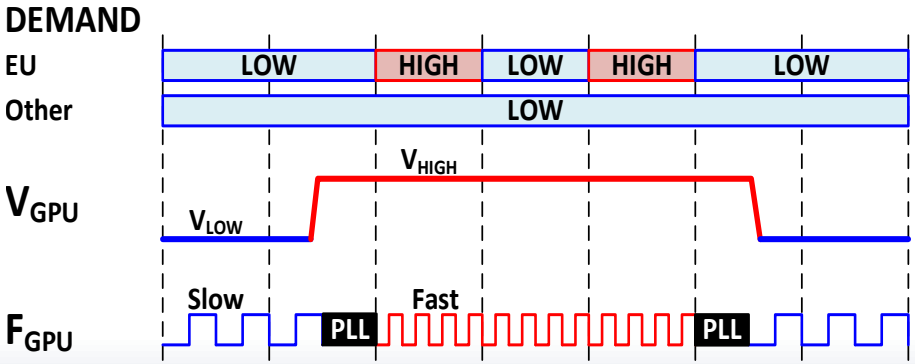
PLL re-lock time

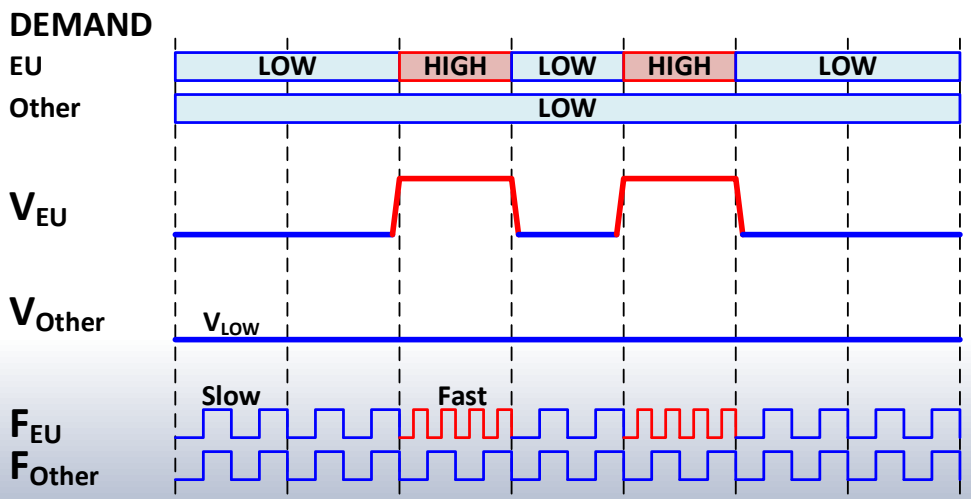


Dedicated rails

Fast EU clock

No PLL re-lock



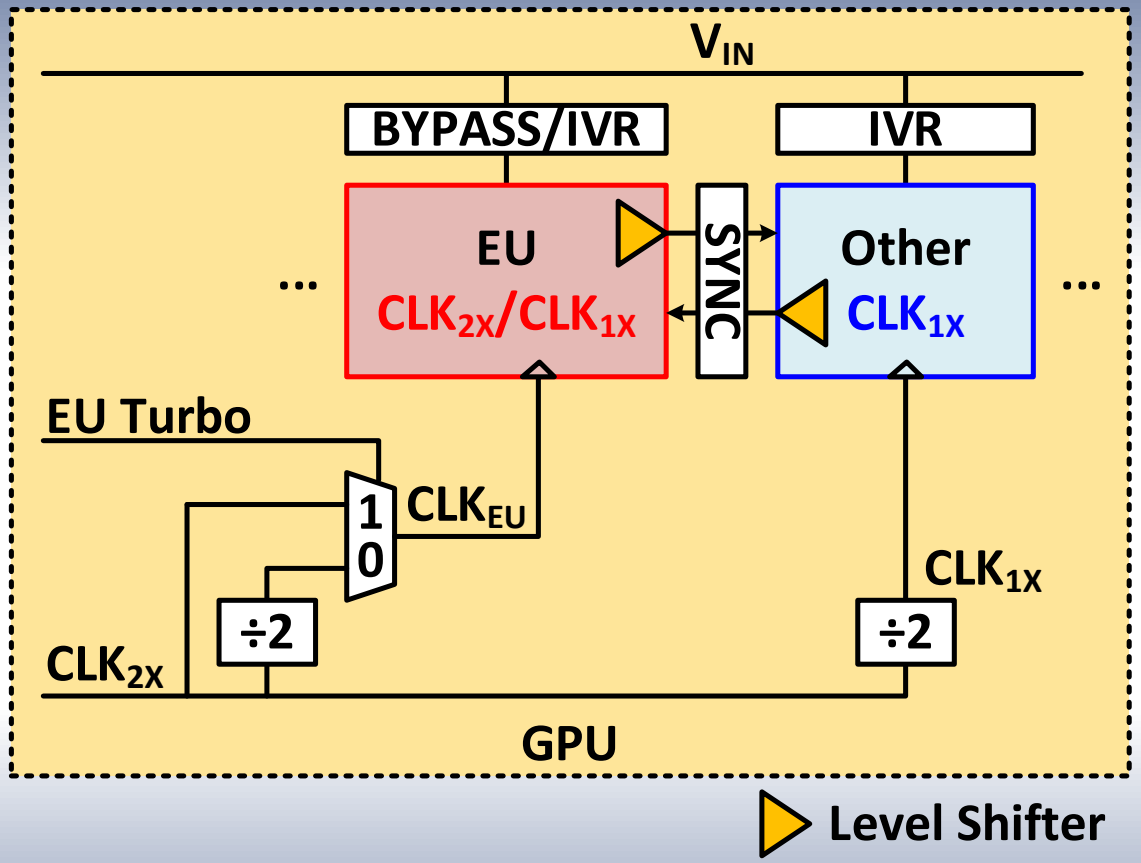


EU Turbo

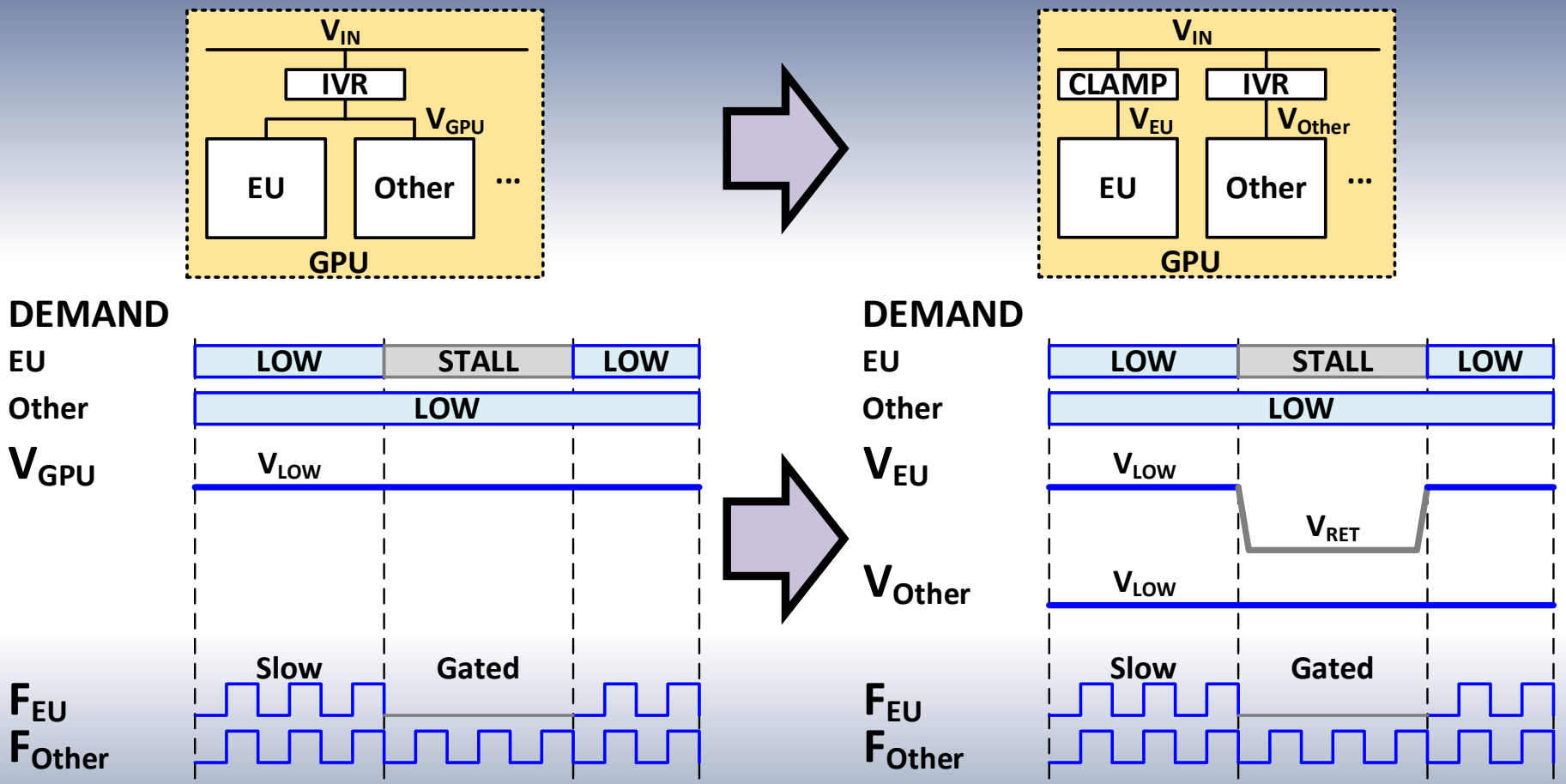
EUs at 2X clock (CLK_{2X}) as needed

Sync logic at 1X/2X clock boundaries

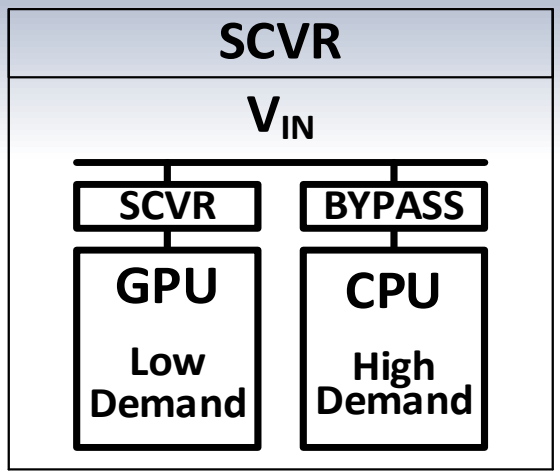
IVR for fast turbo/ un-turbo transitions

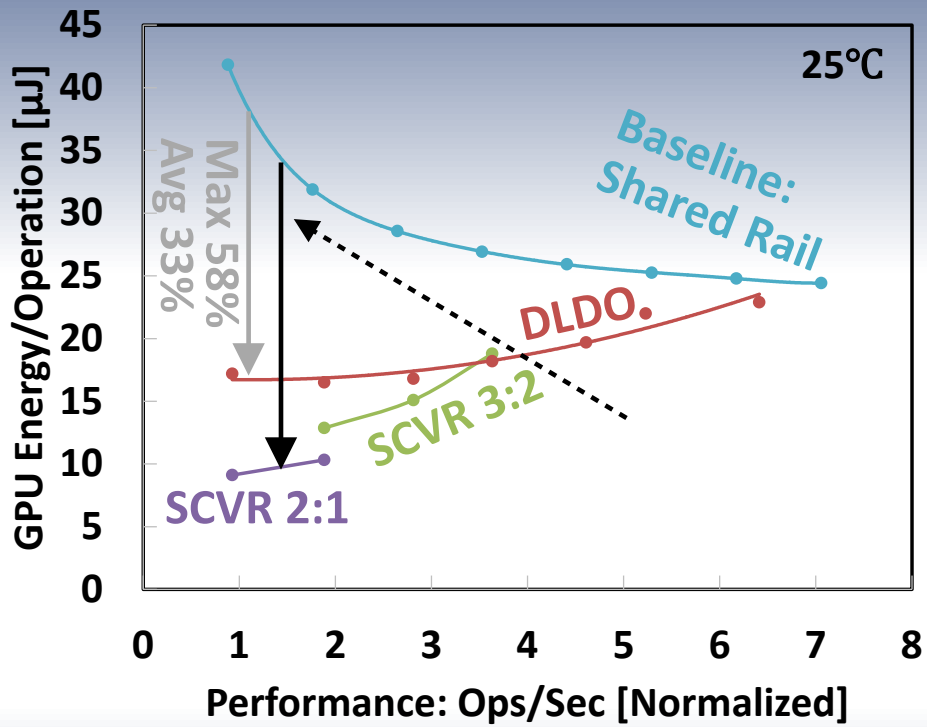


Retentive Sleep



Independent V/F Domains at SoC Level

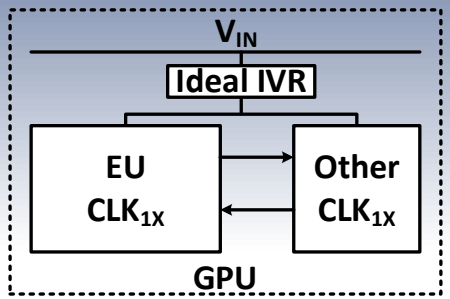




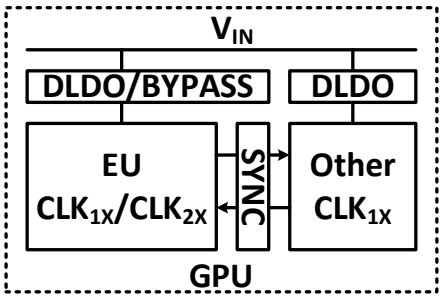
Energy Savings at Iso-Performance: Max 77%, avg 62%

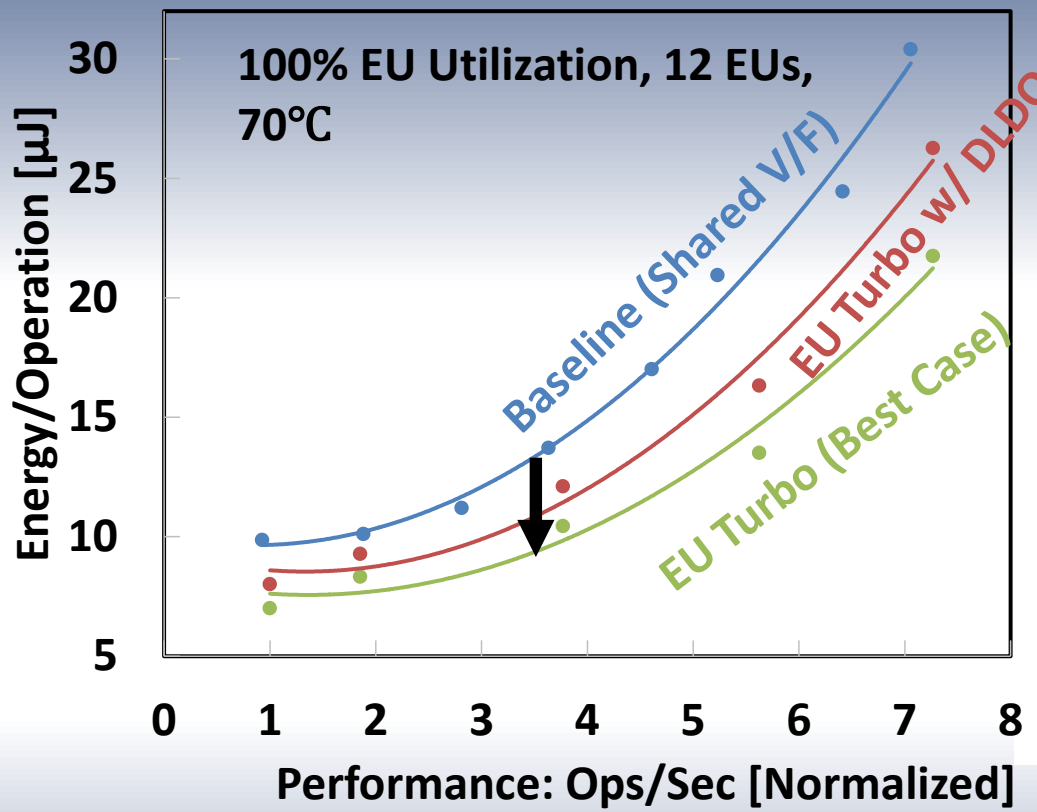
EU Turbo Performance Gain

Baseline: Shared V/F



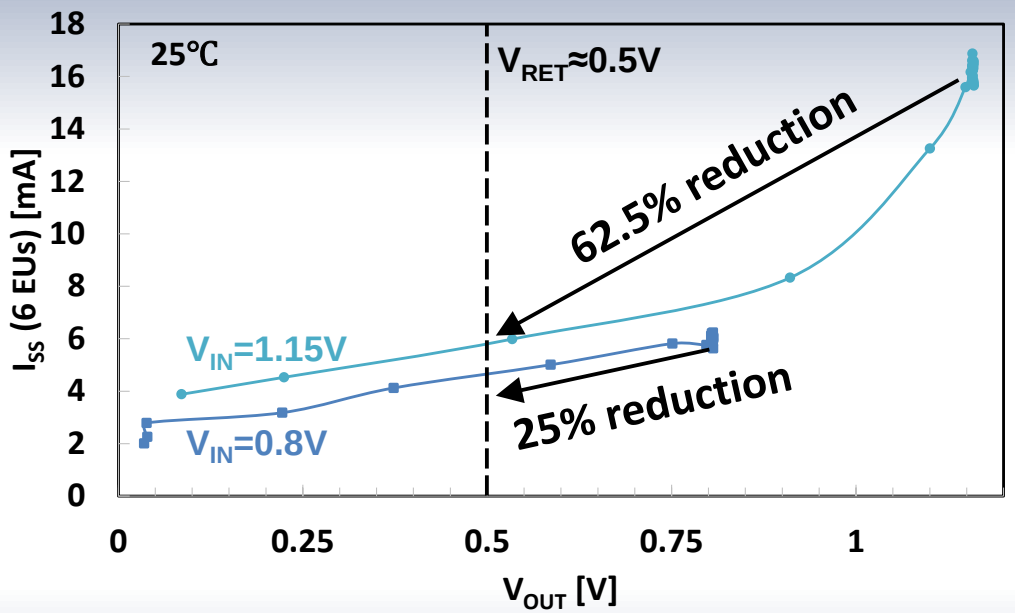
EU Turbo

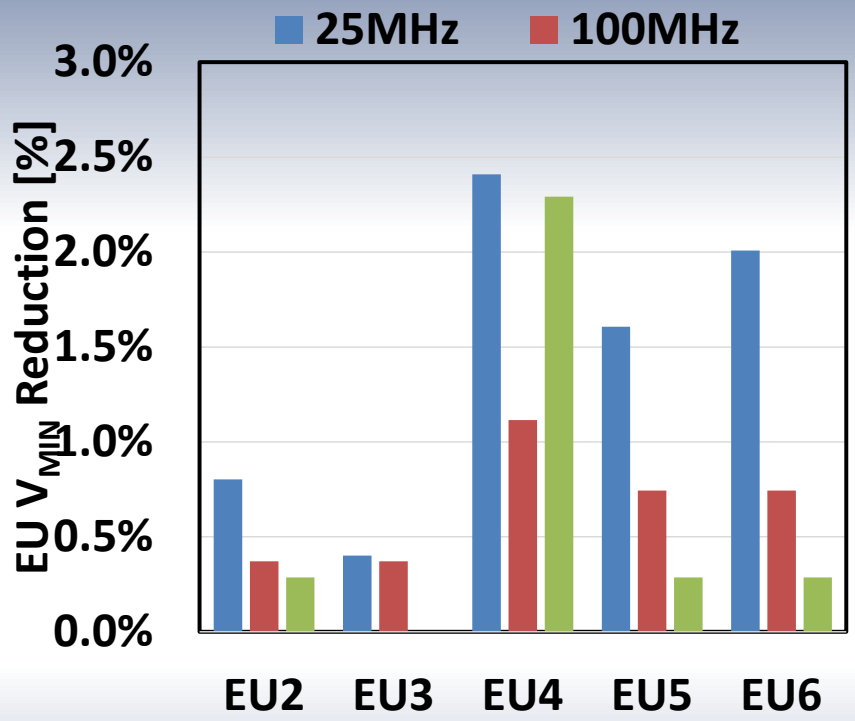




Energy Savings at Iso-Performance: Max 32%, avg 29%

Retentive Sleep: Leakage Savings





Conclusions

Complete 14nm GPU features energy efficiency techniques

Hybrid IVR (DLDO + SCVR) for independent V/F domains

- **Within GPU: EU turbo provides up to 68% performance gain, 50% EU area reduction, or 32% energy savings**
- **SoC Level: SCVR enables up to 77% GPU energy savings**

Additional DLDO usages

- **Retentive sleep: Reduces EU leakage by 63%**
- **V_{MIN} per block: Up to 2.4% EU V_{MIN} reduction**

Summary

Power delivery: VRs inside SoC is fully integrated(>100MHz, >2:1)

- Buck for mainstream SoC, SC for relatively small rails
- Reduced area => less L, C; worse interconnects with every node; efficiency reduces
- Need: Better passives utilization, current density, die stacking

Power management: Inductor less approach for down-scalability

- Distributed, APR-friendly designs desirable for SoC integration and wide adoption

Need down-scalable high current density/die stacking for point-of-load VRs!

Need battery/source-to-SoC VR integration(3-5 MHz bandwidth, discrete passives)

Summary

Power delivery: VRs inside SoC is fully integrated(>100MHz, >2:1)

- Buck for mainstream SoC, SC for relatively small rails
- Reduced area => less L, C; worse interconnects with every node; efficiency reduces
- Need: Better passives utilization, current density, die stacking

Power management: Inductor less approach for down-scalability

- Distributed, APR-friendly designs desirable for SoC integration and wide adoption

Need down-scalable high current density/die stacking for point-of-load VRs!

Need battery/source-to-SoC VR integration(3-5 MHz bandwidth, discrete passives)

Summary

Power delivery: VRs inside SoC is fully integrated(>100MHz, >2:1)

- Buck for mainstream SoC, SC for relatively small rails
- Reduced area => less L, C; worse interconnects with every node; efficiency reduces
- Need: Better passives utilization, current density, die stacking

Power management: Inductor less approach for down-scalability

- Distributed, APR-friendly designs desirable for SoC integration and wide adoption

Need down-scalable high current density/die stacking for point-of-load VRs!

Need battery/source-to-SoC VR integration(3-5 MHz bandwidth, discrete passives)

Summary

Power delivery: VRs inside SoC is fully integrated(>100MHz, >2:1)

- Buck for mainstream SoC, SC for relatively small rails
- Reduced area => less L, C; worse interconnects with every node; efficiency reduces
- Need: Better passives utilization, current density, die stacking

Power management: Inductor less approach for down-scalability

- Distributed, APR-friendly designs desirable for SoC integration and wide adoption

Need down-scalable high current density/die stacking for point-of-load VRs!

Need battery/source-to-SoC VR integration(3-5 MHz bandwidth, discrete passives)

Summary

Power delivery: VRs inside SoC is fully integrated(>100MHz, >2:1)

- Buck for mainstream SoC, SC for relatively small rails
- Reduced area => less L, C; worse interconnects with every node; efficiency reduces
- Need: Better passives utilization, current density, die stacking

Power management: Inductor less approach for down-scalability

- Distributed, APR-friendly designs desirable for SoC integration and wide adoption

Need down-scalable high current density/die stacking for point-of-load VRs!

Need battery/source-to-SoC VR integration(3-5 MHz bandwidth, discrete passives)

Summary

Power delivery: VRs inside SoC is fully integrated(>100MHz, >2:1)

- Buck for mainstream SoC, SC for relatively small rails
- Reduced area => less L, C; worse interconnects with every node; efficiency reduces
- Need: Better passives utilization, current density, die stacking

Power management: Inductor less approach for down-scalability

- Distributed, APR-friendly designs desirable for SoC integration and wide adoption

Need down-scalable high current density/die stacking for point-of-load VRs!

Need battery/source-to-SoC VR integration(3-5 MHz bandwidth, discrete passives)

Summary

Power delivery: VRs inside SoC is fully integrated(>100MHz, >2:1)

- Buck for mainstream SoC, SC for relatively small rails
- Reduced area => less L, C; worse interconnects with every node; efficiency reduces
- Need: Better passives utilization, current density, die stacking

Power management: Inductor less approach for down-scalability

- Distributed, APR-friendly designs desirable for SoC integration and wide adoption

Need down-scalable high current density/die stacking for point-of-load VRs!

Need battery/source-to-SoC VR integration(3-5 MHz bandwidth, discrete passives)

Thank you for your attention!

Acknowledgement

- Vaibhav Vaidya
- Tri Hyunh
- Chung-Ching Peng
- George Matthew
- Carlos tokunaga
- Don Gardener
- Kaladhar Radhakrishnan
- Paul Fischer
- Kevin O'brien
- Muhammad Khellah
- Jim Tschanz
- Ravi Mahajan
- Jonathan Douglas
- Takao Oshita

This research was, in part, funded by the U.S. Government (DARPA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.