SERVER POWER DELIVERY CHALLENGES AND OPPORTUNITIES

MIGUEL RODRIGUEZ, STEPHEN KOSONOCKY,
ALI A. MERRIKH

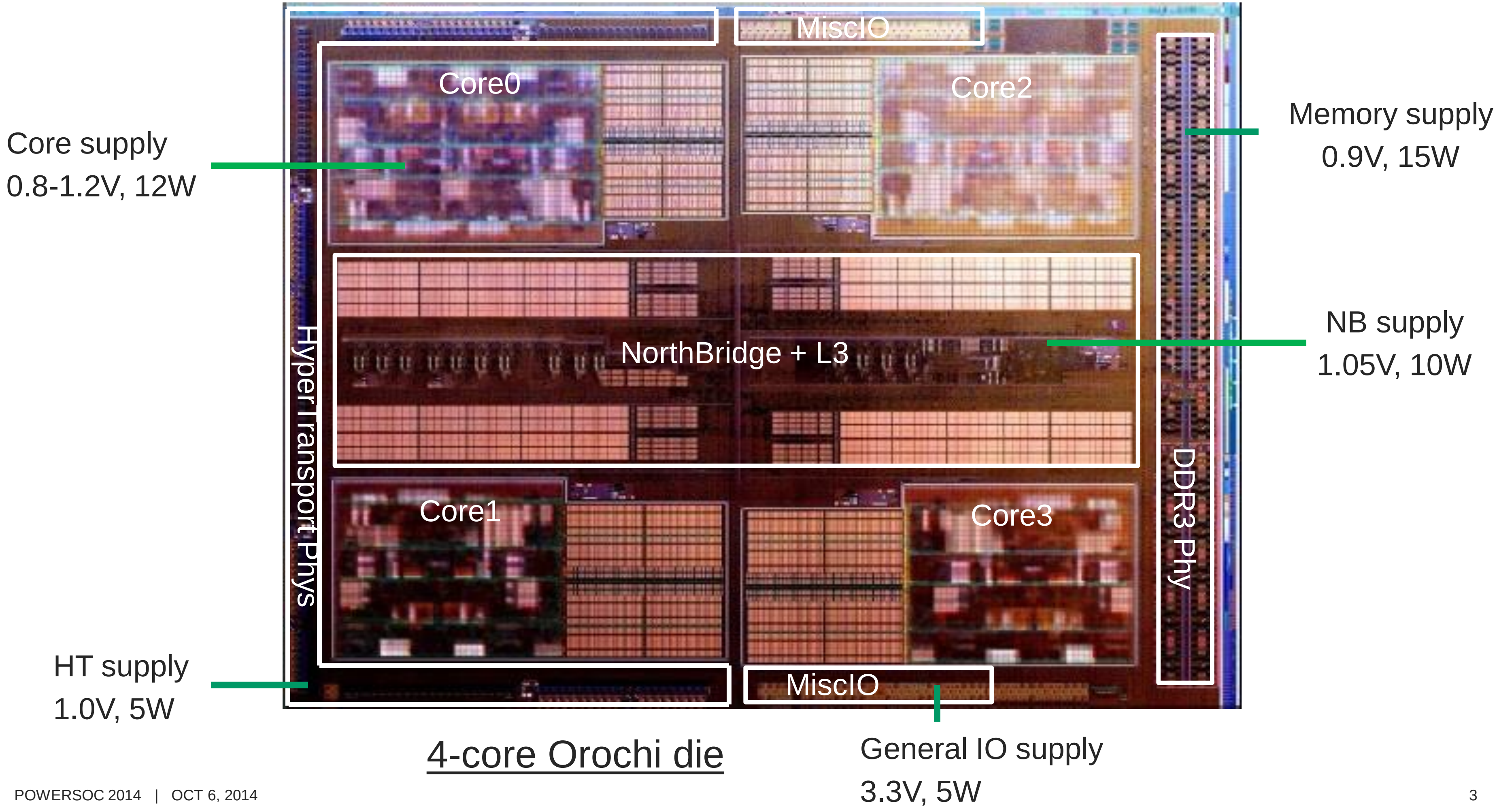
OCT. 6, 2014

OUTLINE

- Introduction
  - Server power delivery
  - Traditional power saving techniques
  - Power saving limitations in server and HPC
- Performance improvement in multi/many core systems: Integrated Voltage Regulation
  - The IVR concept: benefits and concerns
  - P-state optimization
- Limitations of switching IVR solutions in HPC and server systems
  - Performance limitations
  - Thermal limitations
- Using low-dropout linear regulators as IVRs in HPC and servers systems
  - Performance benefit of linear IVRs
  - LDO IVR architectures
- Conclusions

SERVER DIE POWER DELIVERY

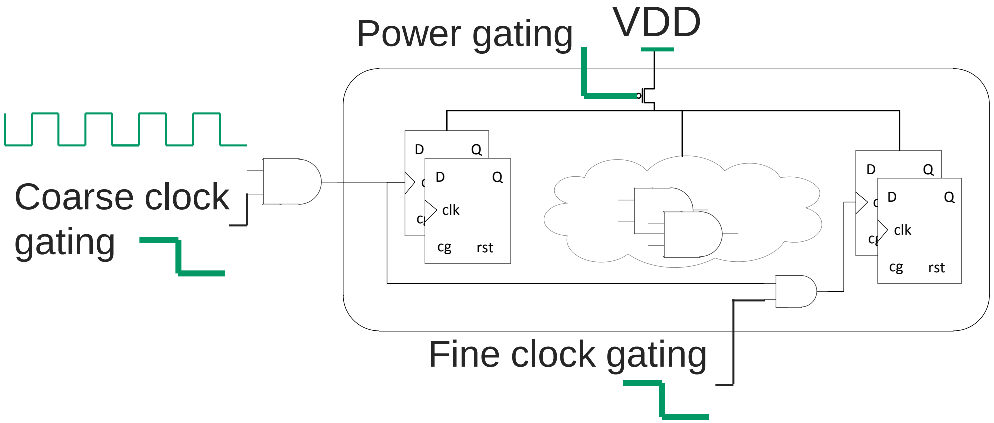
- Power delivery architecture can be quite complex in multi/many core systems
  - Many rails
  - High current / wide voltage requirements



4-core Orochi die

TRADITIONAL POWER SAVING TECHNIQUES

- Basic rule: optimize each section
  - ‘turn off’ unused features
  - Optimize voltage supply
- Clock gating and power gating
  - Use flops only when data is changing (spatially and temporal fine-grained)
  - Turn off the complete clock tree inside an IP (spatially and temporally mid-grained)
  - Idle IP: gate the supply (spatially and temporally coarse-grained)



Power gating

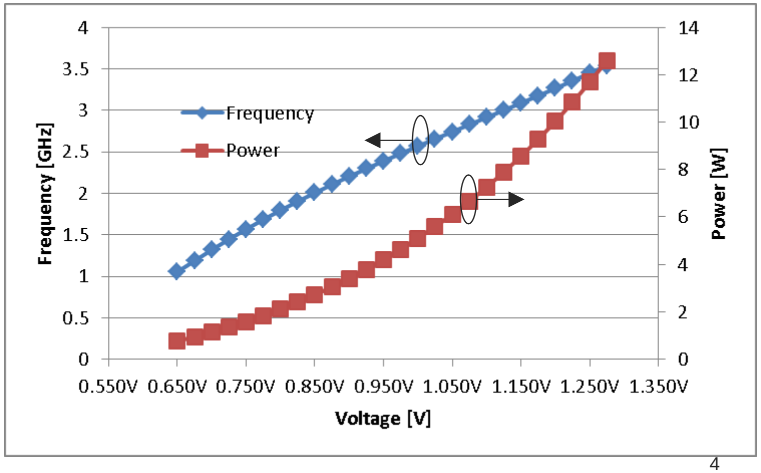
VDD

Coarse clock gating

Fine clock gating

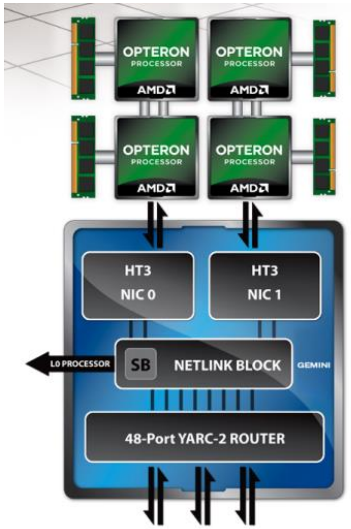
$$P = \underbrace{C \cdot V^2 f}_{\text{Activity, supply}} + \underbrace{V \cdot I_{leak}(V,T)}_{\text{Supply, temperature}}$$

- Core power savings through P-state adjustments
  - A core operates at an optimal v-f pair
  - Frequency is defined by required performance (voltage is adjusted later)



POWER SAVING LIMITATIONS IN SERVER AND HPC SYSTEMS

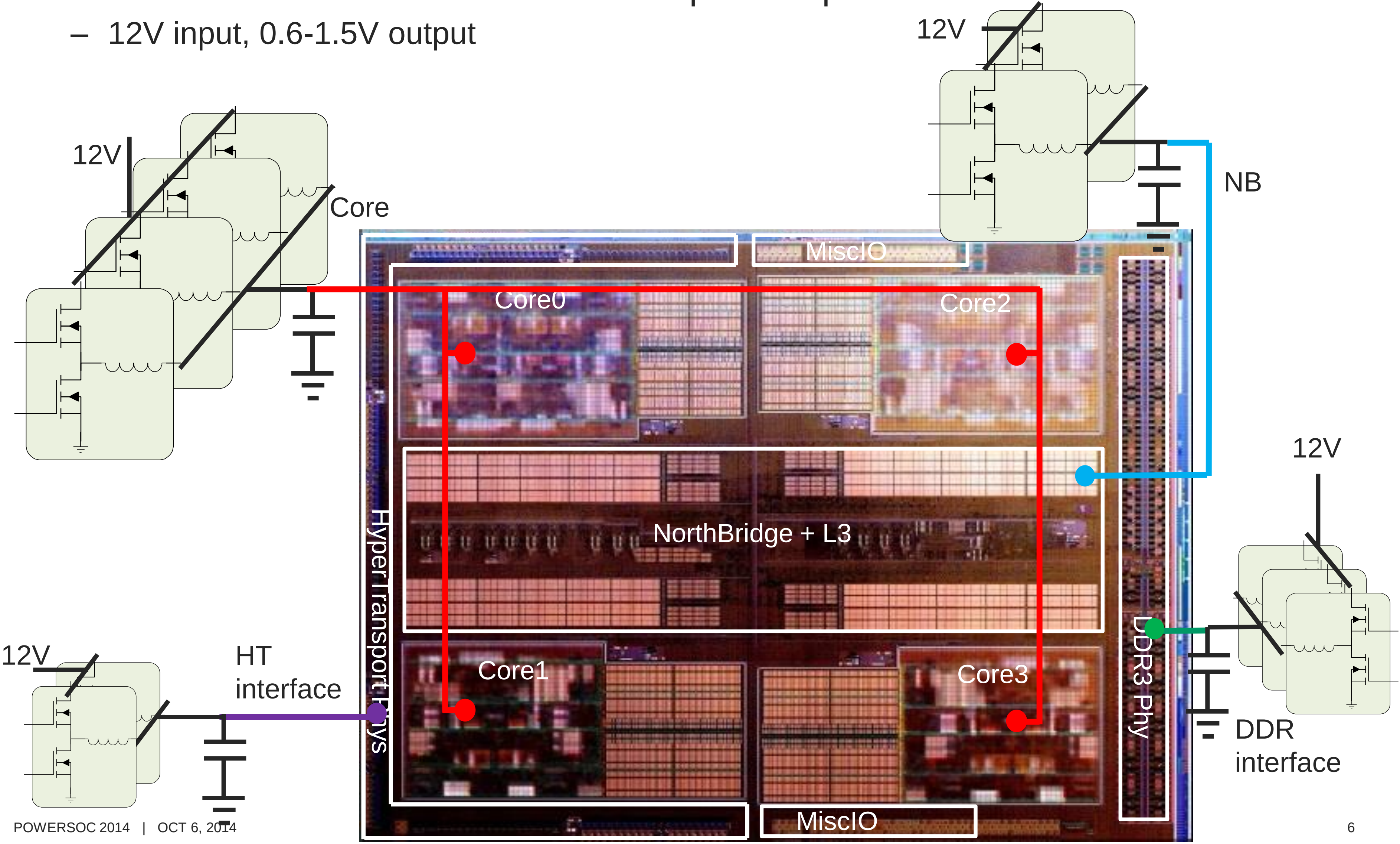
- High Performance Computing is carried out using massive number of processors running in parallel
  – Cray XT5 in ORNL: 224.256 AMD Opteron processors (18688 CU, each is a dual hex-core)



- Very intensive resource utilization: always doing something!
  – Multi-threading is extensively used to maximize throughput
- Coarse techniques do not work well
  – Coarse clock gating or power gating are not effective, as most of the time everything is working at near-full capacity
  – Power gating can even be disabled
- Fine clock-gating is still useful

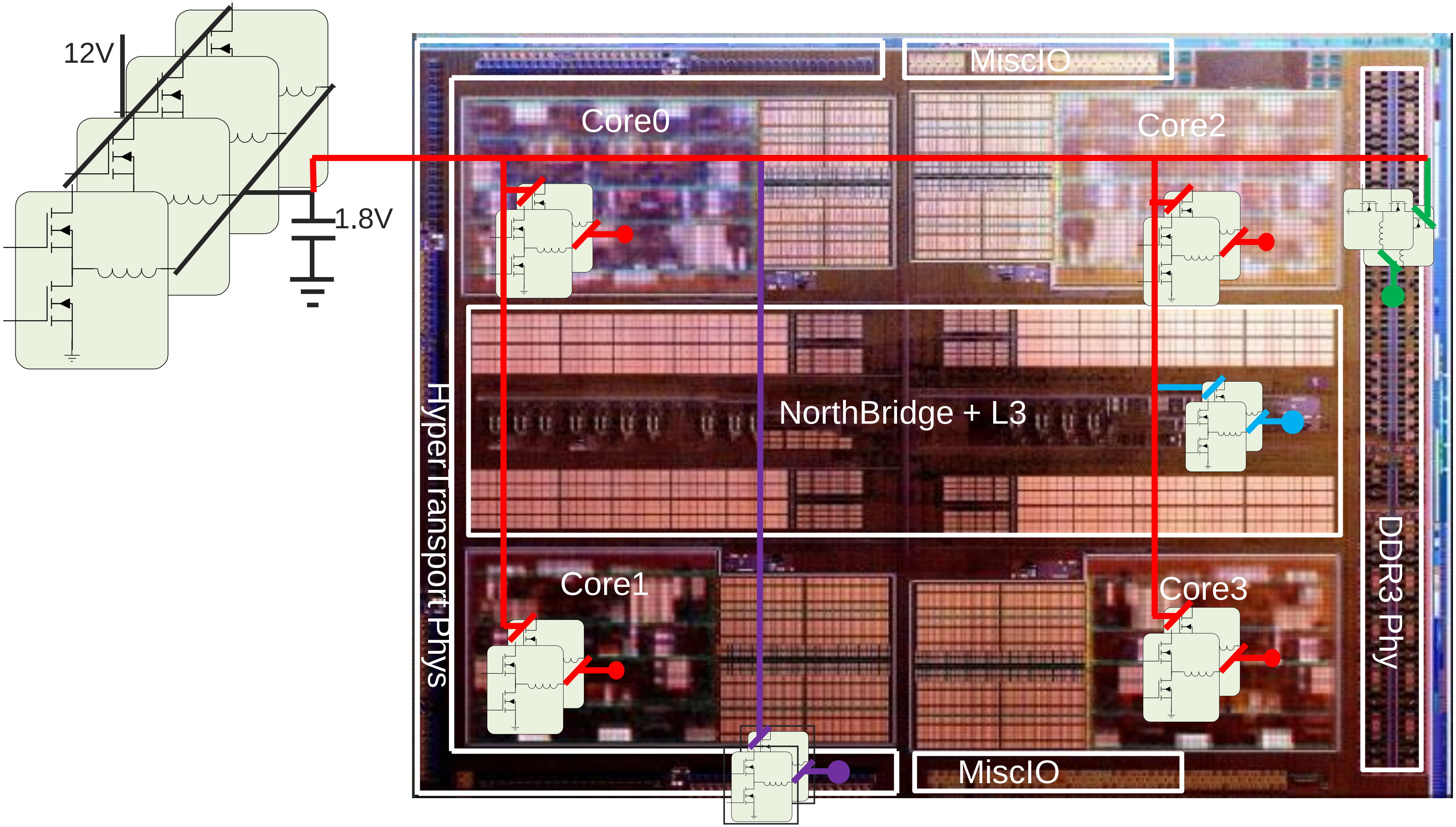
THE IVR CONCEPT

- Traditional VRM structure in multicore chips: multiphase buck converter
  - 12V input, 0.6-1.5V output



THE IVR CONCEPT

- IVR in multicore chips (assuming just a single input rail)

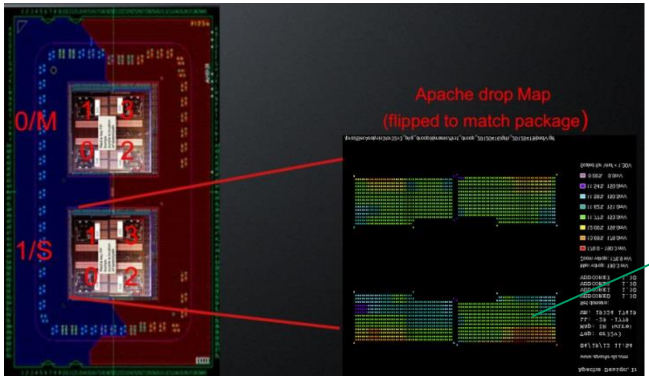


THE IVR CONCEPT

- Move VRM from the board to the chip (IVR)
- General benefits enabled by IVR
  - Improved transient response (lower voltage droops), eliminate interconnection parasitics
  - P-state optimization: critical for multi/many core systems
  - Cost benefit: eliminate significant PCB real state and BOM
  - Reduction of package power distribution unbalances and hot-spots

<u>More subtle problem</u>

complex package power distribution in multicore dies can cause die supply unbalances

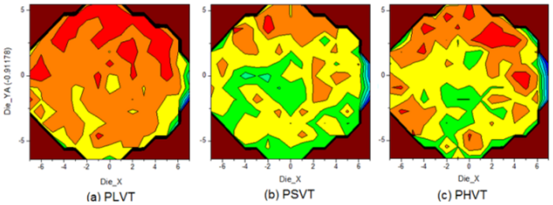


These 2 cores can have worse droops than the other 2 cores

- Mitigation of die to die and core to core variations

<u>Die to die variations</u>: causes deviations in product performance

<u>Core to core variations</u>: voltage is set by the slowest core to hit performance target → the other cores run at higher voltage than necessary*



Sample wafer scale V_{th} variation

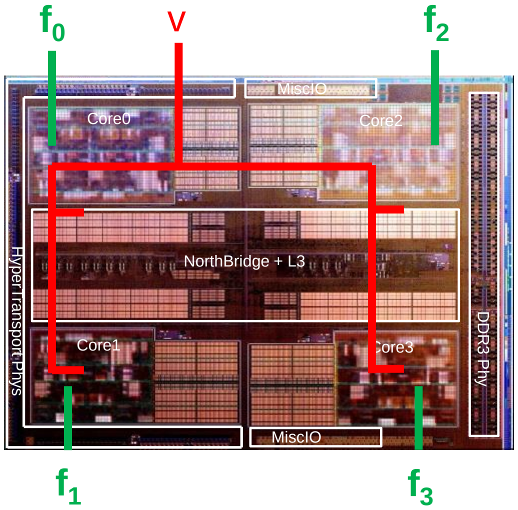
*package unbalances add a systematic error to the random variations

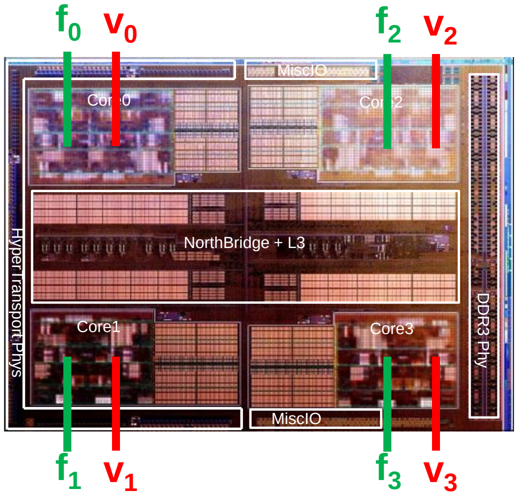
THE IVR CONCEPT

- But IVR does not come for free: there are trade-offs that have to be carefully considered
  - Increased silicon area: higher cost (especially in deep submicron technologies)
  - Increased complexity: on-die inductors? package inductors? control loop? efficiency optimization?
  - Increased package complexity
  - Switching noise/EMI impact
  - Thermal impact
- Furthermore, performance benefits heavily depend on use cases
  - Typical P-state usage
  - Thermally-limited scenarios

P-STATE OPTIMIZATION

- More insight: performance benefit from per-core voltage regulation
  - what if each core could operate at its optimum (f,V)





$$P = C \cdot V^2 f + V \cdot I_{leak}(V,T)$$

$$\Big\downarrow$$

$$P_{ref} = \underbrace{C(V) \cdot V^2 \sum_{i=0}^{N-1} f_i}_{\text{quadratic gain}} + \underbrace{N \cdot V \cdot I_{leak}(V,T)}_{\text{linear*exp gain}}$$

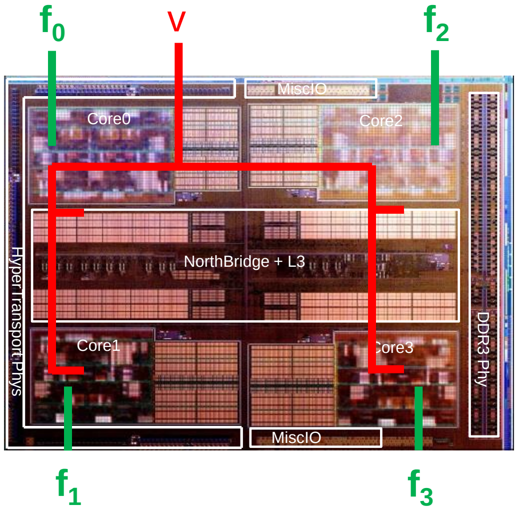
$$P_{ivr} = \sum_{i=0}^{N-1} C(V_i) \cdot V_i^2 \cdot f_i + \sum_{i=0}^{N-1} V_i \cdot I_{leak}(V_i, T)$$

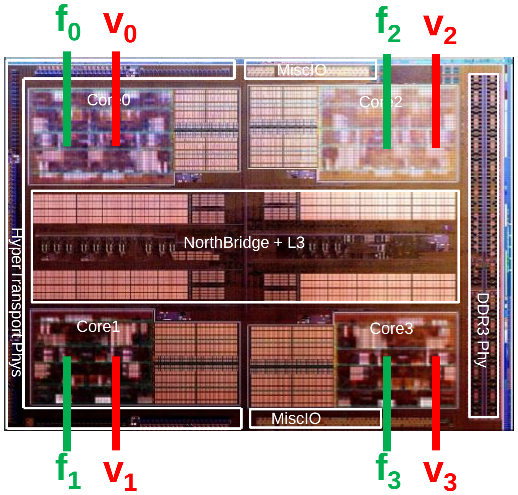
$$I_{leak}(V) \propto a e^{b \cdot V}$$

Note that IVR efficiency is not accounted for here

P-STATE OPTIMIZATION

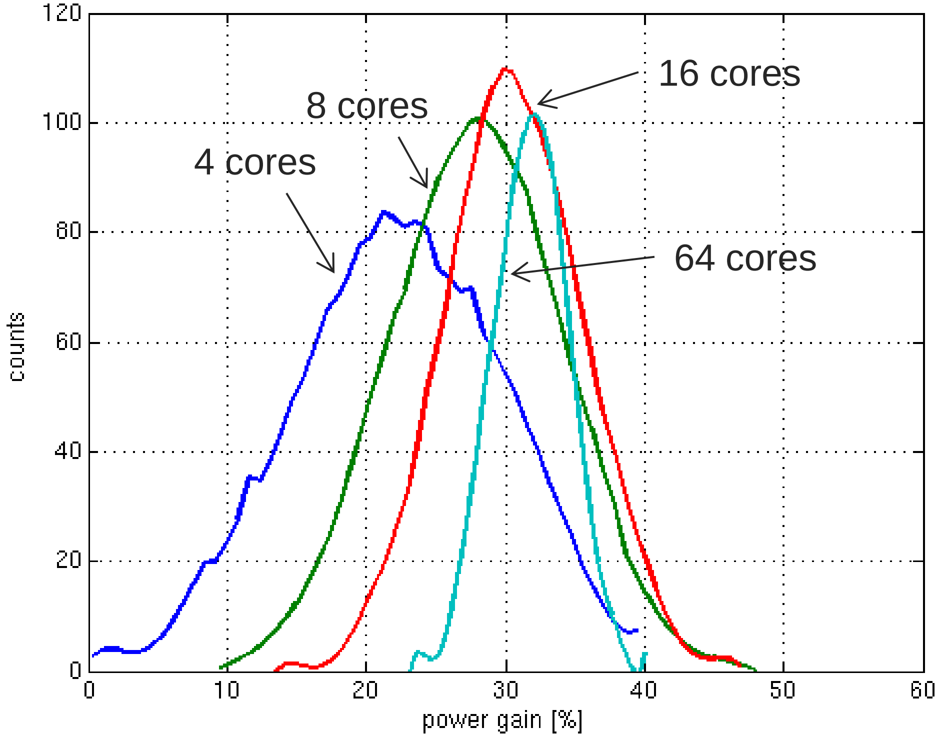
- More insight: performance benefit from per-core voltage regulation
  - what if each core could operate at its optimum (f,V)





$$P = C \cdot V^2 f + V \cdot I_{leak}(V,T)$$

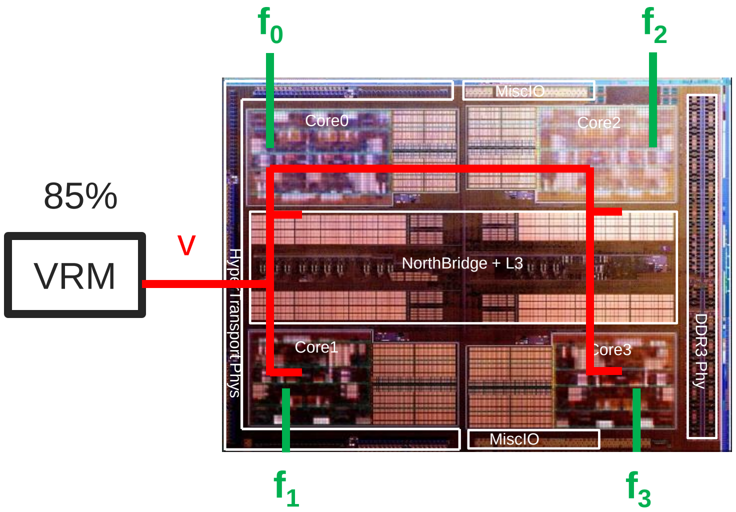
Statistical analysis of power reduction*



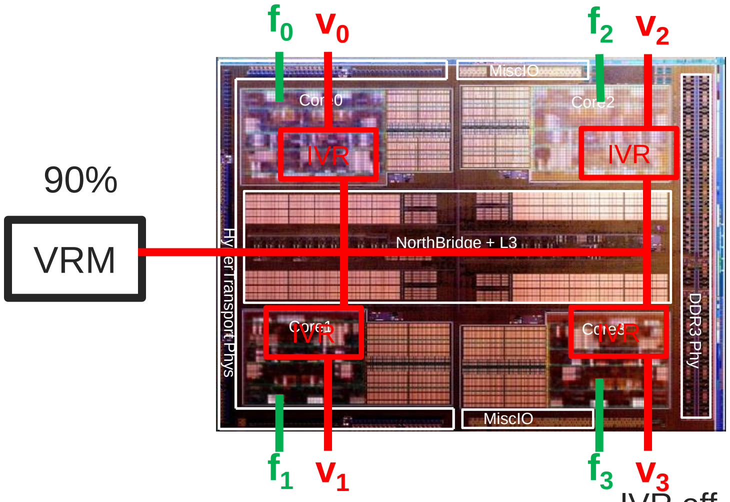
*high leakage technology, random P-state with uniform distribution, 8 possible Pstates 0.75-1.2V equally spaced, power is delivered with 100% efficiency, 100C

P-STATE OPTIMIZATION

- More insight: performance benefit from per-core voltage regulation
  - now consider VRM and IVR efficiencies

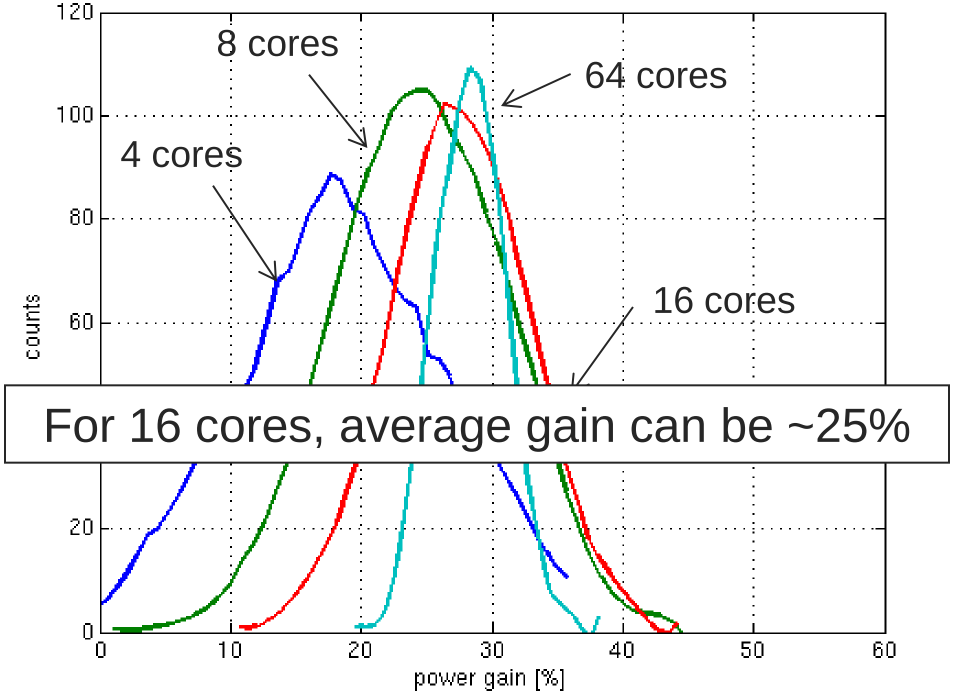


$$P = C \cdot V^2 f + V \cdot I_{leak}(V,T)$$



IVR eff. is 90%

Statistical analysis of power reduction*



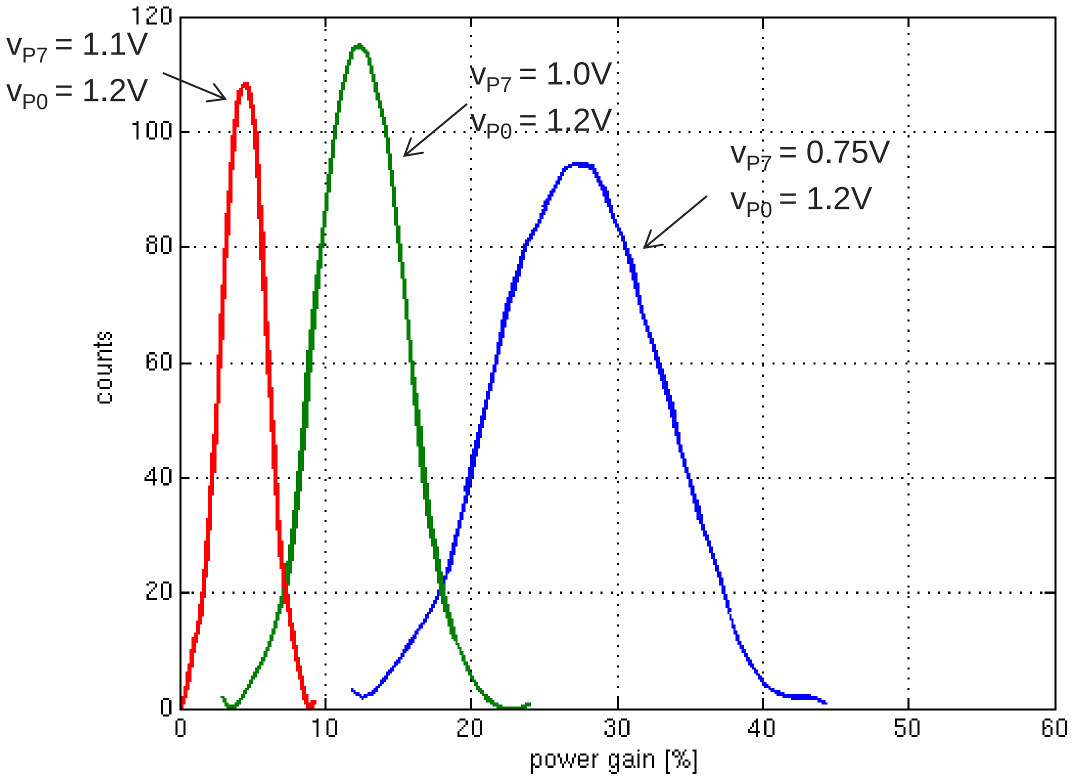
For 16 cores, average gain can be ~25%

*high leakage technology, random P-state with uniform distribution, 8 possible Pstates 0.75-1.2V equally spaced, 100C

PERFORMANCE LIMITATIONS

- Performance gains offered by IVR depend on workload
  - In server and HPC systems, high-performance P-states are used the vast majority of the time
  - This leads to a significant reduction of achievable power gains

Statistical analysis of power reduction*



*high leakage technology, 16 cores, 8 Pstates, uniform distribution over indicated voltage range, 100C

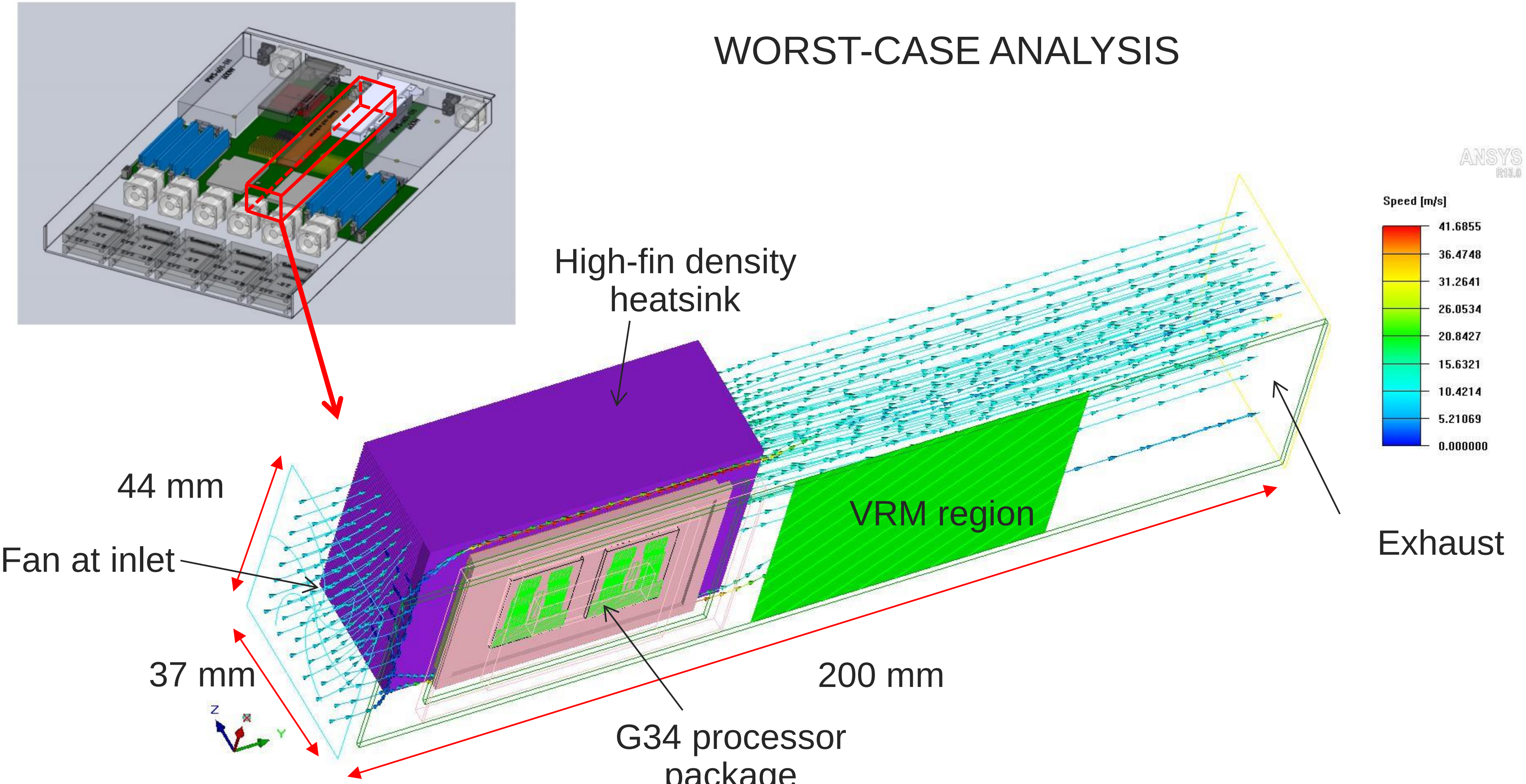
- Less than <15% power reduction
- Somewhat optimistic conditions (IVR efficiency 90%)
- Increase in area and complexity (inductors, control) might not be worth anymore

MIGHT DISCOURAGE IVR SOLUTION IN THESE SYSTEMS

THERMAL LIMITATIONS

- Server and HPC system are typically thermally limited
  - This further impacts performance gains: when all cores are running at the same P-state, losses have shifted from VRM to the die

WORST-CASE ANALYSIS



High-fin density heatsink

44 mm

Fan at inlet

37 mm

VRM region

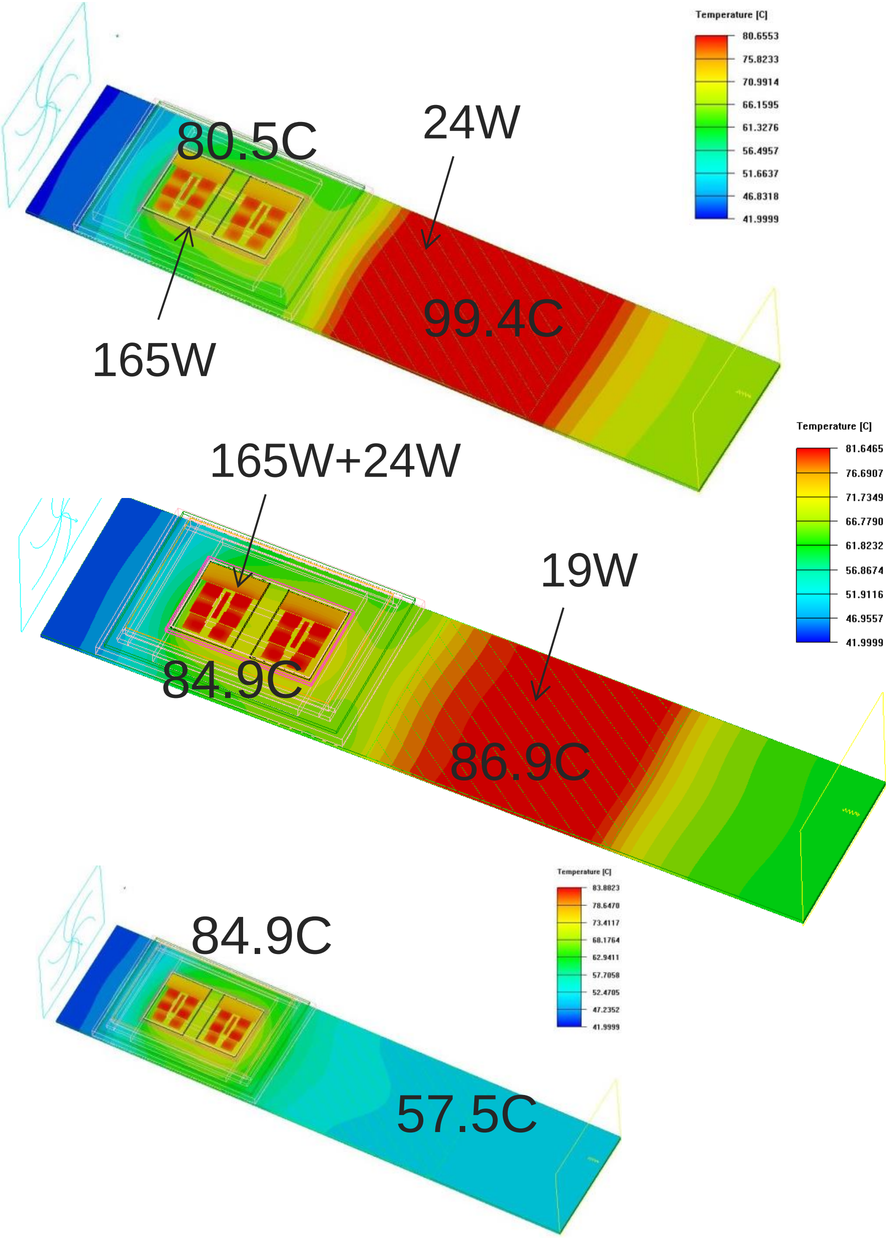
200 mm

Exhaust

G34 processor package

THERMAL LIMITATIONS

- Scenario 1: traditional VRM design
  - Per package TDP: 165W
  - VR Power loss: 24W (~87% efficiency)
  - Fan Speed: 30 CFM
- Scenario 2: VRM+IVR
  - Per package TDP: 165W
  - IVR Power loss: 24W (~87% efficiency)
  - Fan Speed: 30 CFM
  - All cores running at full speed (max P-state)
  - Extra heat uniformly distributed
- Scenario 3: IVR only (as a guideline)
  - Per package TDP: 165W
  - VR Power loss: 24W (~87% efficiency)
  - Fan Speed: 30 CFM
  - All cores running at full speed (max P-state)
  - Extra heat uniformly distributed
  - No VRM required



THERMAL LIMITATIONS

- Assuming equivalent junction and package temperatures
  - Adding IVR results in ~24W core power (non-IVR) deficit (at worst-case operating point)
  - Impact of 24W power deficit on performance is -10.9% assuming leakage constitutes 25% of the total core power

Max. core power needs to be brought down 24W to reach same T_{jmax} → ~11% performance hit

Scenario	Fan flow rate (CFM)	T_{amb}	Heatsink R_{ca} (C/W)	Heatsink R_{ja} (C/W)	T_c	T_j	T_{pcb}	$T_{j\ delta}$	Power compensation	Performance deficit
1- No IVR (165W)	30	42	0.172	0.23	70.4	80.5	99.4			
2- With IVR (190W)	30	42	0.166	0.22	73.4	84.9	86.9	4.3	-24W	-10.9%
3- With IVR (190W)	30	42	0.166	0.22	73.4	84.9	57.5	4.3	-24W	-10.9%

- This could also be addressed with a different thermal solution
  - Heat sink design, package heat transfer, increase fan speed
  - Modify die floorplan

All these have a significant system-level/cost impact

PERFORMANCE BENEFIT OF LINEAR IVR

- We have seen that switching IVRs can add substantial power dissipation to the die, as well as significant complexity
- If the cores are going to operate most of the time in a narrower voltage range, why not use low dropout regulators (LDOs)?
  - In power electronics, this is counterintuitive due to low linear efficiency
  - However, power gain can still be achieved

$$\eta_{LDO} = \frac{V_{out}}{V}$$

1.1V to 1V → η_{LDO}=91%

$$P_{ldo} = \sum_{i=0}^{N-1} C(V_i)\cdot V_i \cdot V \cdot f_i + \sum_{i=0}^{N-1} V \cdot I_{leak}(V_i, T)$$

$$P_{ref} = \underbrace{\overbrace{C(V)\cdot V^2 \sum_{i=0}^{N-1} f_i}^{\text{> linear gain}}}_{\text{> quadratic gain}} + \underbrace{\overbrace{N \cdot V \cdot I_{leak}(V,T)}^{\text{exponential gain}}}_{\text{exponential * linear gain}}$$

$$P_{ivr} = \sum_{i=0}^{N-1} C(V_i)\cdot V_i^2 \cdot f_i + \sum_{i=0}^{N-1} V_i \cdot I_{leak}(V_i, T)$$

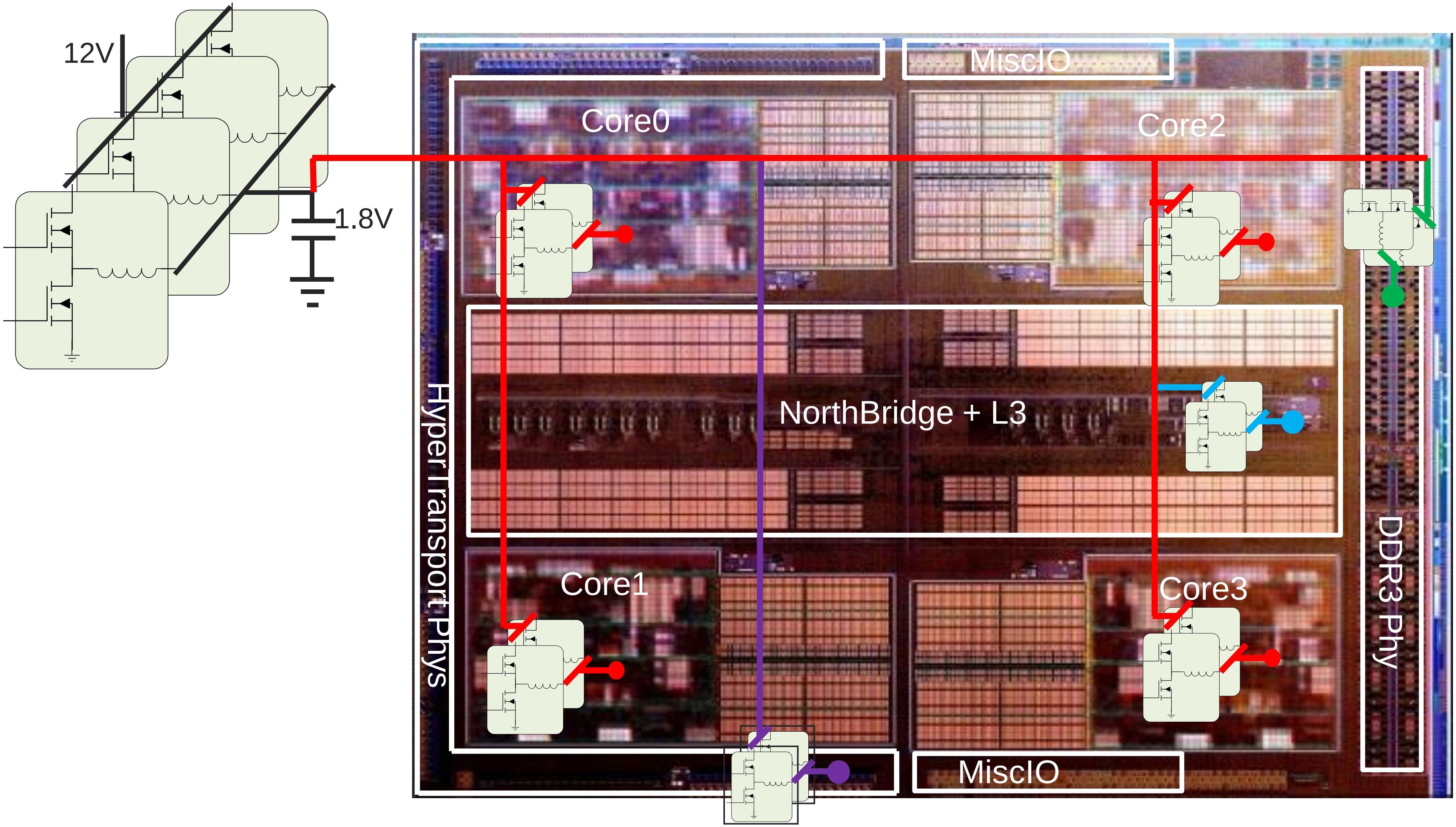
USING LDO AS IVR IN SERVER AND HPC

- LDO vs switching IVR

	Switching IVR	**LDO**
Complexity	High	Low-medium
Chip area	Increase	No impact
Efficiency	High	Medium-high (V_{in}/V > 0.9)
Thermal impact	Medium-high	Small or no impact
Custom design required	High	Low-medium

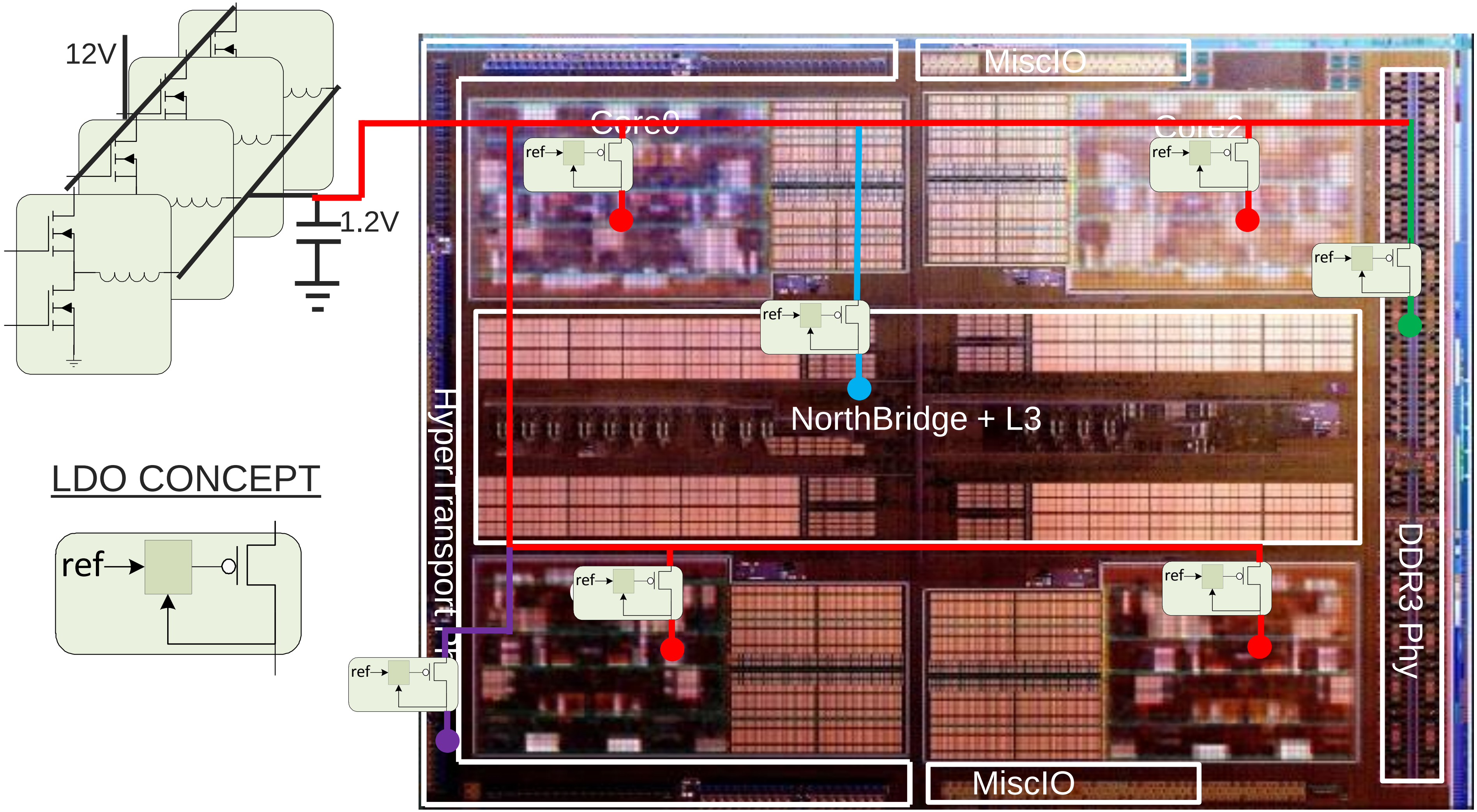
THE IVR CONCEPT

- IVR in multicore chips (assuming just a single input rail)



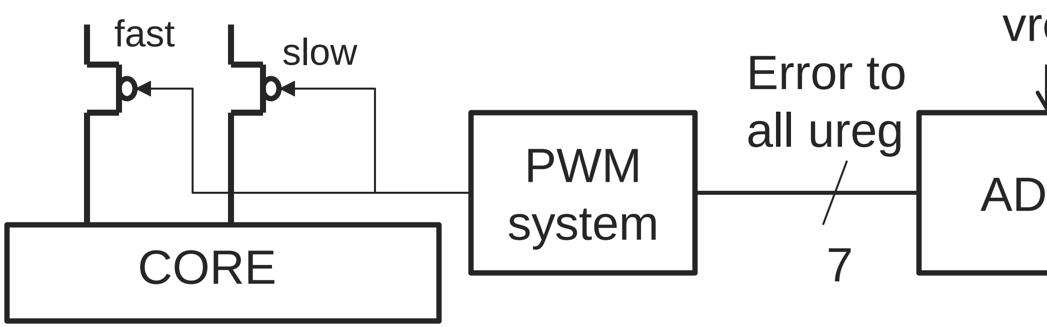
THE IVR CONCEPT

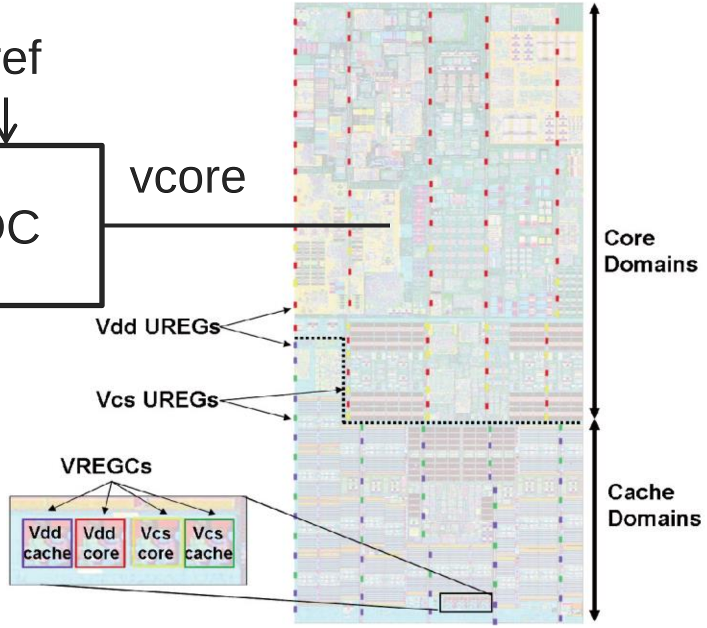
- IVR in multicore chips (assuming just a single input rail)



LDO IVR ARCHITECTURES

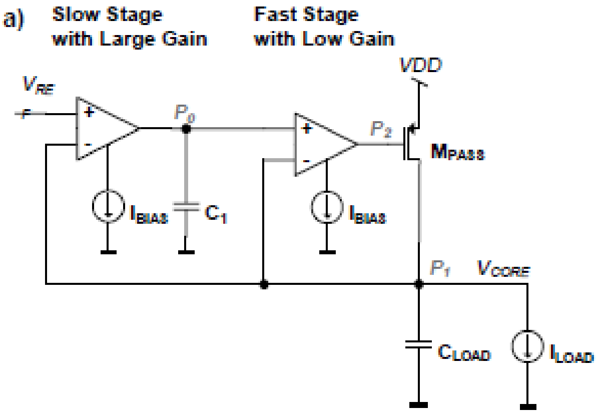
- Distributed:
  - [ISSCC14, “Distributed System of Digitally Controlled Microregulators Enabling Per-Core DVFS for the Power8 Microprocessor]





- Traditional
  - [Fully-integrated LDO voltage regulator for digital circuits]

  M. Luders et al, Adv. Radio Sci., 9, 263–267, 2011



Used to supply a low power micro-controller core

- Traditional analog approach
- Any-load stable

CONCLUSIONS

- Power delivery in multicore systems is challenging: many rails with different requirements
- Per-core voltage regulation can be advantageous in these systems, but certain trade-offs have to be considered
  - P-state performance gains
  - Thermal limitations
- Server and HPC systems have very specific constraints that can discourage switching IVR implementations
  - Typical workloads yield low benefit from per-core P-state optimization
  - Thermal impact in thermally-limited systems can be intolerable
- LDOs can be a good alternative solution to switching IVRs
  - High efficiency when dropout is low
  - Relatively simple, low design / chip area impact, almost no overhead
  - Several approaches already demonstrated in literature and commercially

REFERENCES

- Distributed System of Digitally Controlled Microrregulators Enabling Per-Core DVFS for the Power8 Microprocessor, Toprak, Deniz et. al., Solid-State Circuits Conference Digest of Technical Papers (ISSCC), pp. 98-99, Feb. 2014
- Fully-integrated LDO voltage regulator for digital circuits, M. Luders et al, Adv. Radio Sci., 9, 263–267, 2011
- Thermal Management of Fujitsu's High-performance servers, Jie Wei, Fujitsu Sci. Tech J., 43, 1, p. 122-129, 2007
- System Level Analysis of Fast, Per-Core DVFS using On-Chip Switching Regulators, Wonyoung Kim et al., High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symposium on , pp. 123-134, Feb. 2008
- FIVR — Fully integrated voltage regulators on 4th generation Intel® Core™ SoCs, Burton, E.A et. al., Applied Power Electronics Conference and Exposition (APEC), 2014 Twenty-Ninth Annual IEEE , pp. 432-439, March 2014
- A Switched-Inductor Integrated Voltage Regulator With Nonlinear Feedback and Network-on-Chip Load in 45 nm SOI, N. Sturcken et. al., IEEE Journal of Solid State Circuits, vol. 47, no. 8, pp. 1935-1945, August 2012
- Evaluation of Fully-Integrated Switching Regulators for CMOS Process Technologies, Jaeseo Lee et. al., IEEE Transactions on VLSI Systems, vol. 15, no. 9, pp. 1017-1027, September 2007
- Digitally Controlled Low-Dropout Regulator with Fast-Transient and Autotuning Algorithms, Yen-Chia Chu et. al., IEEE Transactions on Power Electronics, vol. 28, no. 9, pp 4308-4317, September 2013
- Full On-Chip CMOS Low-Dropout Voltage Regulator, R. J. Milliken et. al., IEEE Transactions on Circuits and Systems I, vol. 54, no. 9, pp. 1879-1890, September 2007
- 0.5-V input digital LDO with 98.7% current efficiency and 2.7-µA quiescent current in 65nm CMOS, Y. Okuma, Custom Integrated Circuits Conference (CICC), 2010 IEEE , pp. 1-4, Sept. 2010

Disclaimer and Attribution

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Trademark Attribution

AMD, the AMD Arrow logo and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. Other names used in this presentation are for identification purposes only and may be trademarks of their respective owners.