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
Power delivery architecture can be quite complex in multi/many core systems

- Many rails
- High current / wide voltage requirements

![Diagram showing power delivery components and supplies](image)
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)

- 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)

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

![Diagram of power saving techniques with equations and graphs.]
High Performance Computing is carried out using massive number of processors running in parallel
- Cray XT5 in ORNL: 224,256 AMD Opteron processors (18,688 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
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

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

- Mitigation of die to die and core to core variations

  Die to die variations: causes deviations in product performance
  Core to core variations: voltage is set by the slowest core to hit performance target → the other cores run at higher voltage than necessary*

*package unbalances add a systematic error to the random variations

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

Sample wafer scale $V_{th}$ variation
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
More insight: performance benefit from per-core voltage regulation
– what if each core could operate at its optimum (f,V)

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

\[ P_{\text{ref}} = C(V) \cdot V^2 \sum_{i=0}^{N-1} f_i + N \cdot V \cdot I_{\text{leak}}(V, T) \]

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

Note that IVR efficiency is not accounted for here
- More insight: performance benefit from per-core voltage regulation
  - what if each core could operate at its optimum (f,V)

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

Statistical analysis of power reduction*

*high leakage technology, random P-state with uniform distribution, 8 possible P-states 0.75-1.2V equally spaced, power is delivered with 100% efficiency, 100C
More insight: performance benefit from per-core voltage regulation
– now consider VRM and IVR efficiencies

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

Statistical analysis of power reduction*

For 16 cores, average gain can be \(~25\%\)

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

- 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

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

Fan at inlet

44 mm

High-fin density heatsink

37 mm

G34 processor package

200 mm

VRM region

Exhaust

41.6855
36.4748
31.2641
26.0534
20.8427
15.6321
10.4214
5.21069
0.00000
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
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

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

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.

\[ \eta_{LDO} = \frac{V_{\text{out}}}{V} \]

1.1V to 1V → \( \eta_{LDO} = 91\% \)

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

- Linear gain

\[
P_{\text{ref}} = C(V) \cdot V^2 \sum_{i=0}^{N-1} f_i + N \cdot V \cdot I_{\text{leak}}(V, T)\]

- Quadratic gain

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

- Exponential * linear gain
### LDO vs switching IVR

<table>
<thead>
<tr>
<th></th>
<th>Switching IVR</th>
<th>LDO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>High</td>
<td>Low-medium</td>
</tr>
<tr>
<td>Chip area</td>
<td>Increase</td>
<td>No impact</td>
</tr>
<tr>
<td>Efficiency</td>
<td>High</td>
<td>Medium-high ($V_{in}/V &gt; 0.9$)</td>
</tr>
<tr>
<td>Thermal impact</td>
<td>Medium-high</td>
<td>Small or no impact</td>
</tr>
<tr>
<td>Custom design required</td>
<td>High</td>
<td>Low-medium</td>
</tr>
</tbody>
</table>
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 CONCEPT
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]

Used to supply a low power microcontroller 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

- 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
- 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.

©2014 Advanced Micro Devices, Inc. All rights reserved.