IEEE Electronics Packaging Society Seminar

Challenges & Innovations in Microprocessor Power Delivery

Kaladhar Radhakrishnan
Intel Fellow

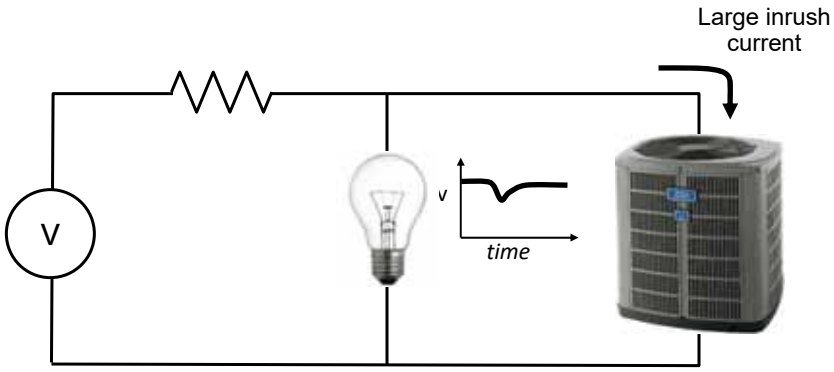
intel

Outline

- Power Delivery 101
  - Microprocessor Power Delivery Overview
  - Power Delivery Network & Decoupling Capacitors
- Microprocessor Scaling Trends
  - Break down of Dennard Scaling
  - Introduction of Multi-core Architecture
- Integrated Voltage Regulators
  - Types of IVR
  - Fully Integrated Voltage Regulator (FIVR)
  - Technology Ingredients for High Voltage IVR
- Summary

The Power Delivery Problem

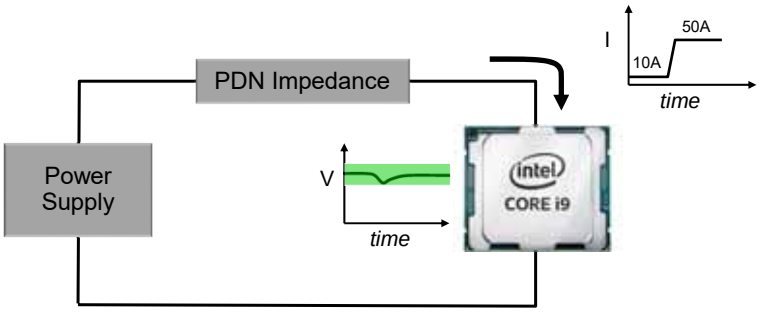
- What happens when you turn on the air conditioner?



IEEE Electronics Packaging Society Seminar intel 3

Microprocessor Power Delivery

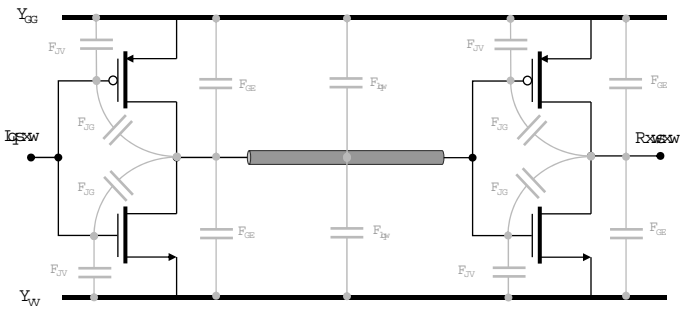
- Typical Microprocessor power delivery problem
  - Load can go from 10A to 50A in ~2ns (di/dt = 20 A/ns)
  - Keep die voltage within ~10% of nominal voltage at all times



IEEE Electronics Packaging Society Seminar intel 4

Microprocessor Power

- Power is consumed by a micro-processor due to the charging and discharging of various parasitic capacitors in a CMOS circuit
- The different types of parasitic capacitance in a circuit are
  - Gate to source capacitance
  - Gate to drain capacitance
  - Drain to body (source) capacitance
  - Interconnect (wire) capacitance
- Each capacitor is charged and discharged during a clock cycle



Dynamic Power = AF x $C_{dyn} V_{DD}^2$ f

Where is the Power Supply?

Desktop

Power Supply Unit
(120V AC → 12V DC)

Desktop VR
(12V DC → 1V DC)

Laptop

AC Adapter
(120V AC →10V DC)

Mobile VR
(10V DC → 1V DC)

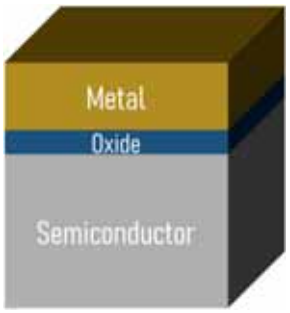
Battery
2S or 3S

Power Distribution Network

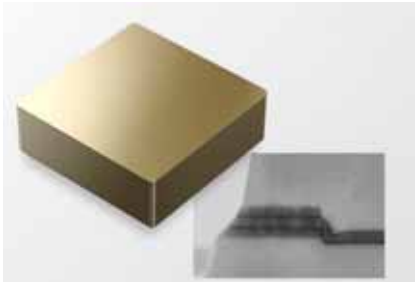
- The buck regulator on the MB has a finite response time which can be modeled as an effective inductance
  - The inductance of the path from the platform VR to the die is of the order of ~1nH
- However, the typical dI/dt for a microprocessor is around ~20A/ns
- Capacitors are used to manage the short-term current draw before the power supply on the MB can respond
  - Multiple stages of decoupling capacitors are used to manage the transient current

On-die Capacitors

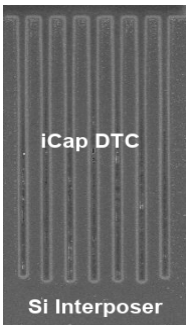
MOS Capacitor



MIM Capacitor



Trench Capacitor

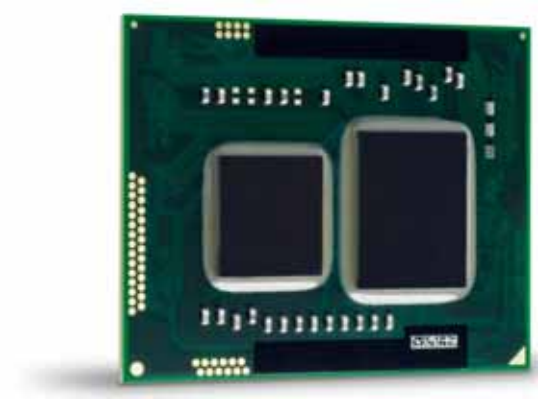


Increasing Capacitance Density

Faster Response Time

Package Capacitors

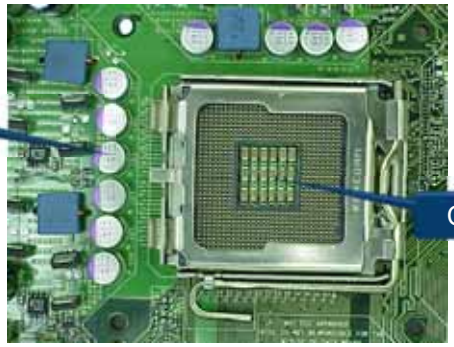
Package with Dieside Caps



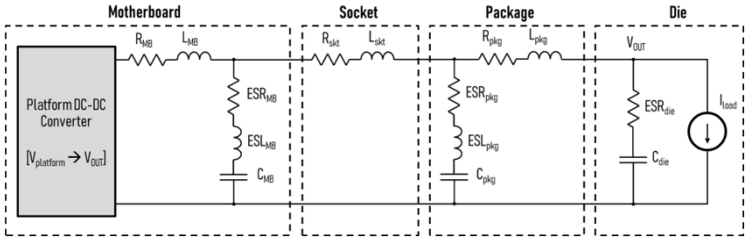
Package with Landside Caps

MB Ceramic & Electrolytic Caps

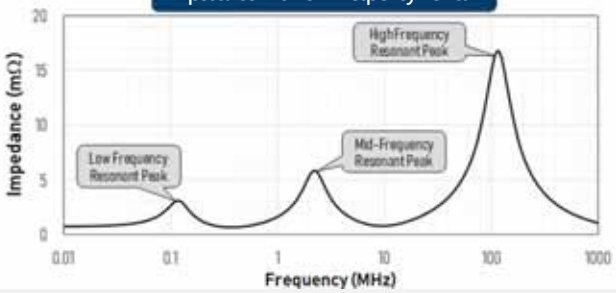
- The MB capacitors are usually ceramic or electrolytic
  - Ceramic caps are used to provide mid frequency decoupling
  - Bulk caps are used to provide low frequency decoupling
- Capacitance and ESL are higher than package caps



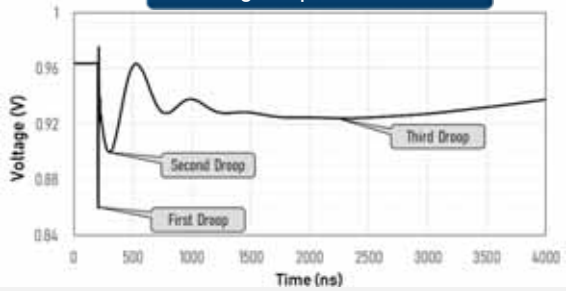
Power Delivery Network



Impedance Profile – Frequency Domain



Voltage Droops – Time Domain



Impact of Voltage Droop on F_{max}

- F_{max} is a function of gate delay and RC delay

$$F_{max} \; \alpha \; \frac{1}{\tau_{gate} + \tau_{RC}}$$

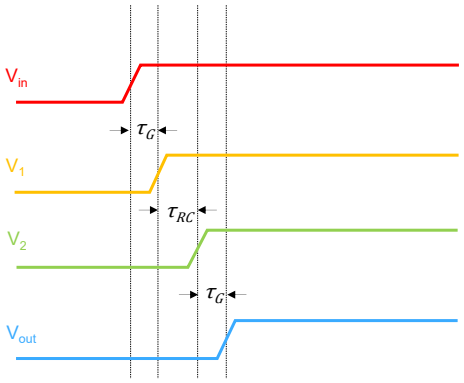
- Gate delay has a strong dependence on gate length and voltage

$$\tau_{gate} \; \alpha \; \frac{L}{V_{cc}}$$

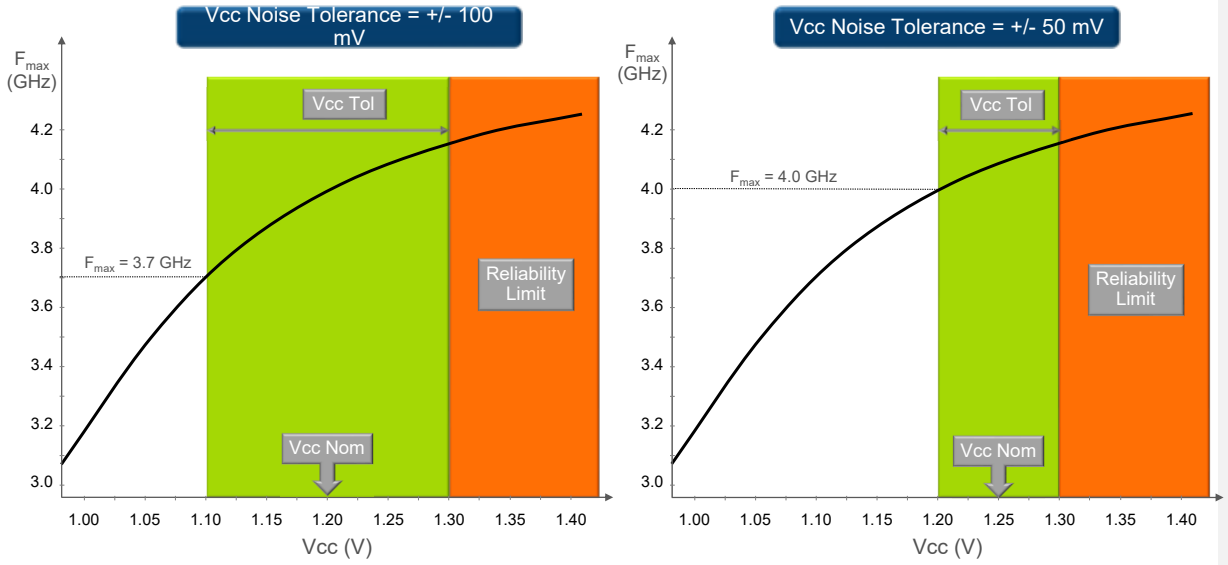
- RC delay is a function of the resistance and capacitance of the on-die interconnects

$$\tau_{RC} \; \alpha \; RC$$

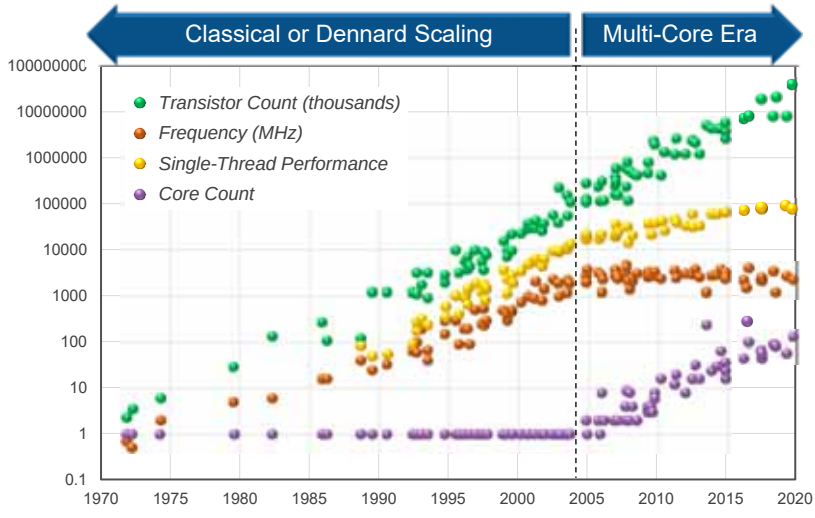
- Fmax can be increased by
  - Reducing gate length (L)
  - Minimizing voltage droops (maximize Vcc)
  - Reducing temperature (reduces R)



Impact of Voltage Droop on F_{max}



Microprocessor Scaling Trends



https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/

Dennard Scaling

- Dennard Scaling to enable Moore's Law
  - $k = \sqrt{2}$; $1/k \sim 0.7$
- Device Linear Dimension scaling
  - L, W, t_{ox} scale by $1/k \sim 0.7\times$
- Transistor Area Scaling — Moore's Law
  - $L \times W = 1/k^2 \sim 0.5\times$
- Voltage Scaling
  - V_G, V_T, V_{DD} scale by $1/k \sim 0.7\times$
- Gate Capacitance Scaling
  - $C_G = \varepsilon_{ox} WL/t_{ox} = 1/k \sim 0.7\times$
- Transistor Drive Current (ON current) Scaling
  - $I_D = \frac{W}{2L} \cdot \mu_n \cdot \frac{\varepsilon_{ox}}{t_{ox}} \cdot (V_G - V_T)^2 = 1/k \sim 0.7\times$
- Gate Delay Scaling — More Fmax
  - $t_d = C_G \cdot V_G / I_D = 1/k \sim 0.7\times$
- Power Density Scaling — Constant Power Density
  - $P_D = CV^2 f/A =$ Constant

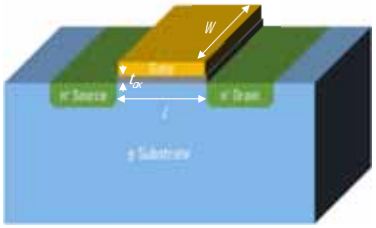
Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions

ROBERT H. DENNARD, MEMBER, IEEE, FRITZ H. GAENSSLEN, HWA-NIEN YU, MEMBER, IEEE, V. LEO RIDEOUT, MEMBER, IEEE, ERNEST BASSOUS, AND ANDRE R. LEBLANC, MEMBER, IEEE

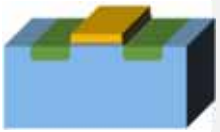
TABLE I
SCALING RESULTS FOR CIRCUIT PERFORMANCE

Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_a	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\epsilon A/t$	$1/\kappa$
Delay time/circuit VC/I	$1/\kappa$
Power dissipation/circuit VI	$1/\kappa^2$
Power density VI/A	1

NMOS – Gen N



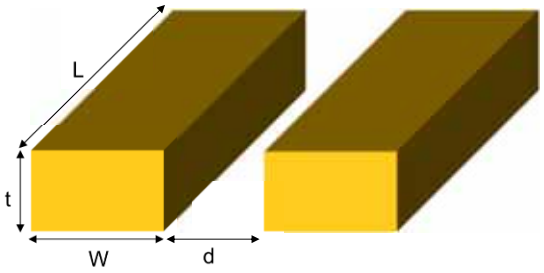
NMOS – Gen N+1



Interconnect Scaling

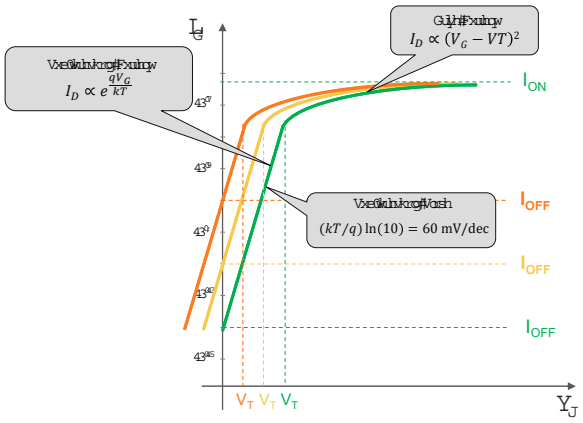
- Interconnect linear dimensions
  - L, W, t, d scale by $1/k \sim 0.7\times$
- Line Resistance
  - $R = \frac{\rho L}{tW}$ scales by $k \sim 1.4\times$
- Normalized voltage drop — IR drop gets worse
  - $\frac{IR}{V}$ scales by $k \sim 1.4\times$
- Capacitance
  - $C = \frac{\epsilon L t}{d}$ scales by $1/k \sim 0.7$
- Interconnect delay — RC delay does not scale
  - RC is constant
- Current density — Current density gets worse
  - I/A scales by $k \sim 1.4\times$

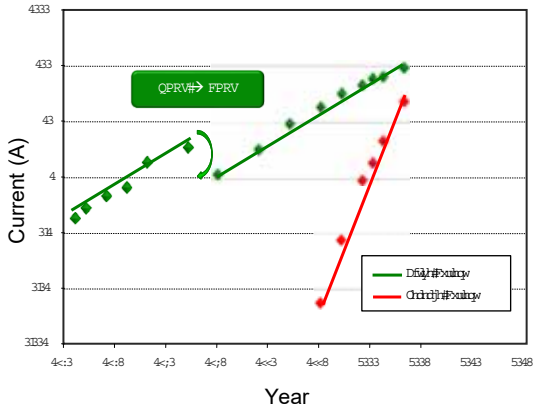
Interconnect – Gen N



Interconnect – Gen N+1

Threshold Voltage Scaling



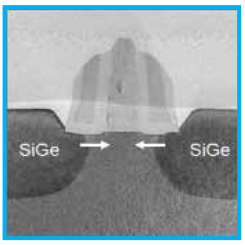


- Threshold voltage scaling causes the leakage current to go up significantly
- As the channel gets shorter, the sub-threshold slope goes up and leakage increases

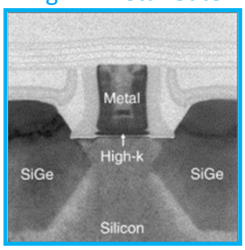
Post-Dennard Scaling

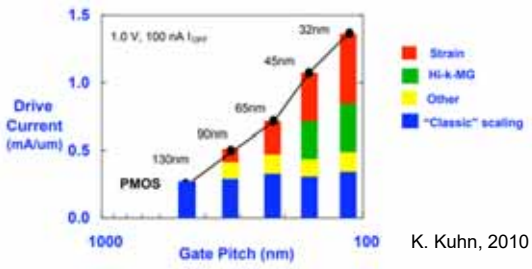
- Strained Silicon
  - Introduces compressive on tensile strain on the silicon atoms in the channel
  - Increases I_D by improving electron/hole mobility
  - $I_D = \frac{W}{2L} \cdot \mu_n \cdot \frac{\varepsilon_{ox}}{t_{ox}} \cdot (V_G - V_T)^2$
- Hi-K Metal Gate
  - Scaling gate oxide thickness results in significant leakage due to tunneling
  - HKMG increases the gate oxide permittivity to offset the lack of gate oxide thickness scaling
  - $I_D = \frac{W}{2L} \cdot \mu_n \cdot \frac{\varepsilon_{ox}}{t_{ox}} \cdot (V_G - V_T)^2$

Strained Silicon



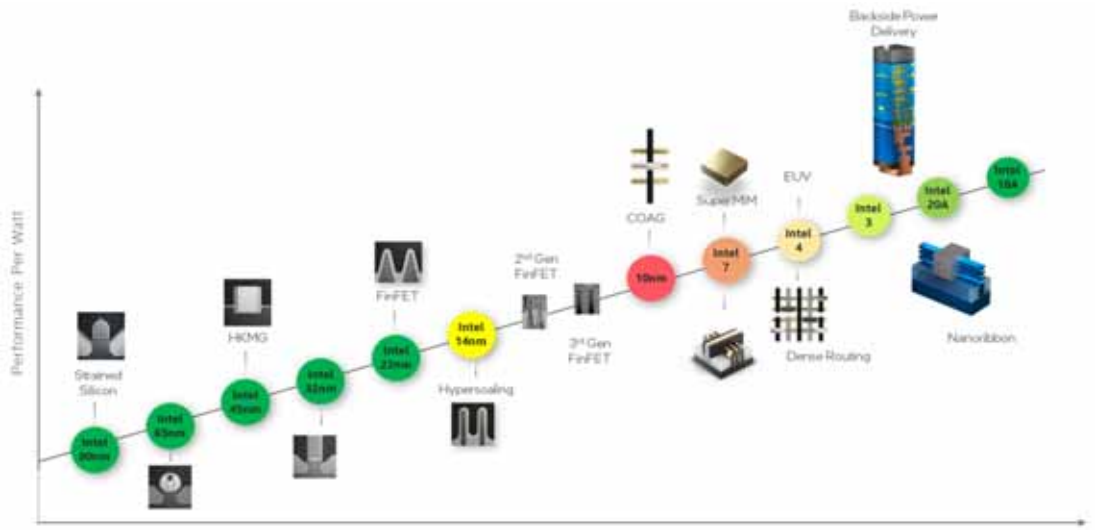
High-K Metal Gate





K. Kuhn, 2010

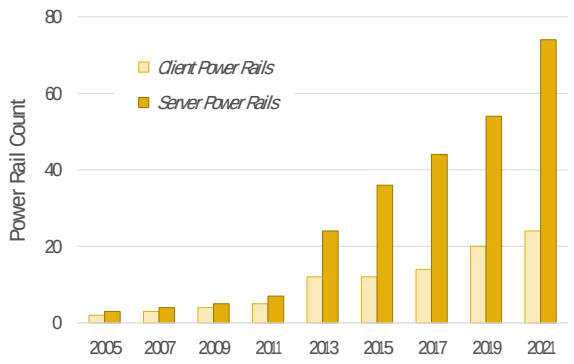
Post-Dennard Scaling

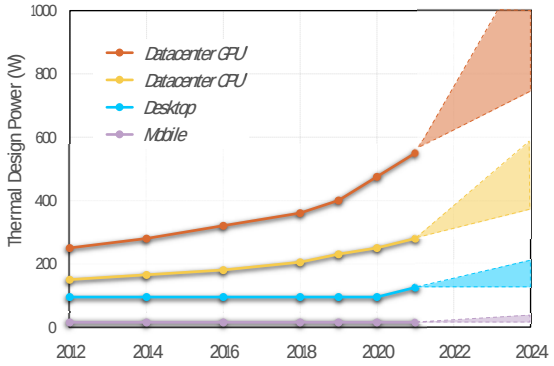


Multi-Core Architecture

- Classical or Dennard Scaling (1970s to early 2000s)
  - Frequency scaling to maximize performance
  - Scale V_{th} & device dimensions
  - Faster (& leaky) transistors enable higher frequency
  - Increased leakage and active power
- Shift to Multi-Core (mid 2000s onwards)
  - Slow down V_{th} scaling to control leakage
  - Use process improvements to control leakage as well
  - Slow down frequency scaling
  - Add more cores for performance

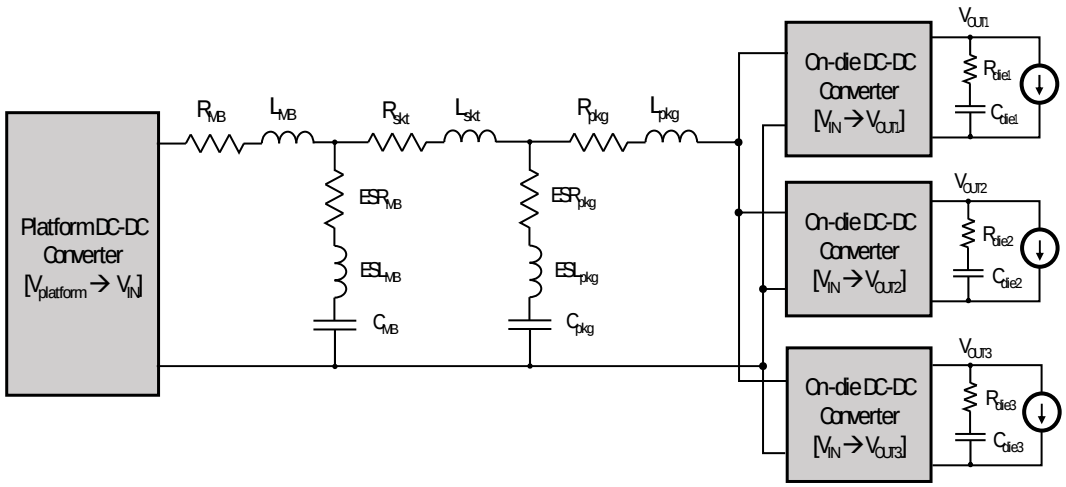
PD Challenges in Multi-Core Era





- The number of power rails has been scaling up with the core count
- Overall power levels have been going up in the high-power segments

Integrated Voltage Regulators

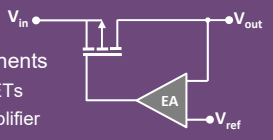


Integrated Voltage Regulators

Power Gate

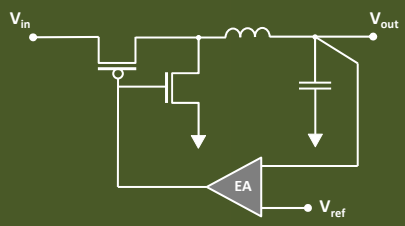
- Key Components
  - Power FETs
- Easy to integrate on Silicon
- No voltage regulation

LDO Regulator



- Key Components
  - Power FETs
  - Error Amplifier
- Easy to integrate on Silicon
- Restricted to low power rails

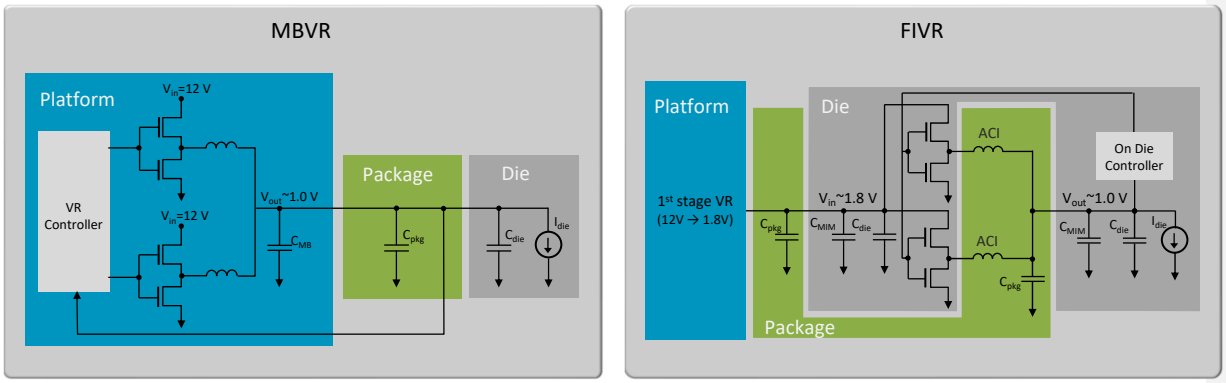
Buck Regulator



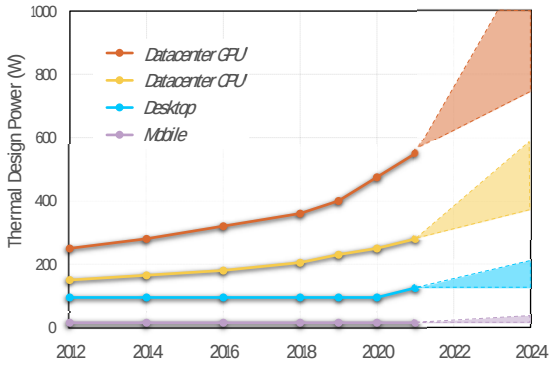
- Key Components
  - Power FETs
  - Inductor
  - Output Filter Capacitor
- More expensive & difficult to integrate
- Typically used for high power rails to take advantage of their high efficiency

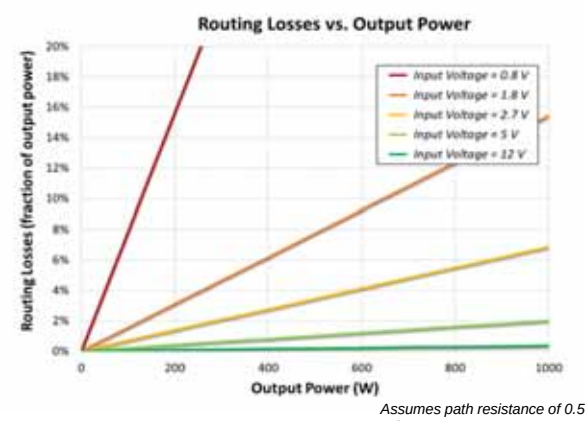
Fully Integrated VR

- Fully Integrated Voltage Regulator (FIVR) was first introduced on the 4th Generation Intel® Core™ Processor
  - The high frequency of operation enabled the use of Air Core Inductors (ACI) as the energy storage element



Impact of Routing Losses

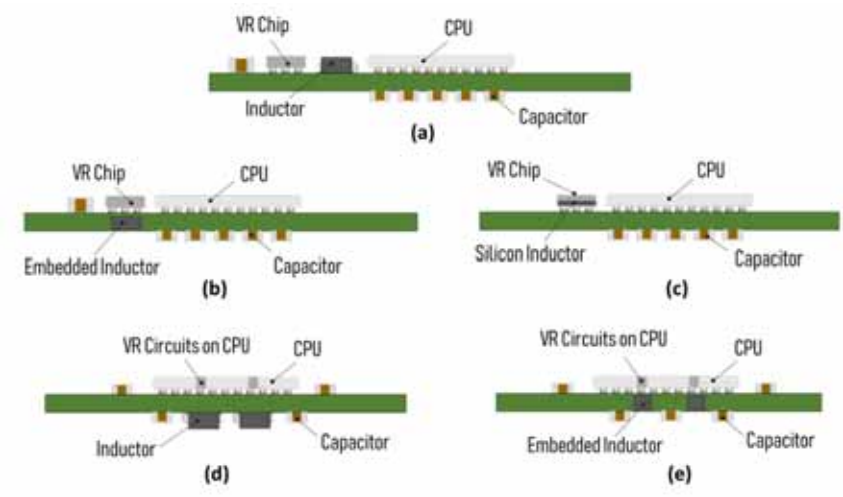




Assumes path resistance of 0.5 mΩ

- As the output power goes up, IVRs are required to bring in power at a higher voltage and keep routing losses manageable

HV-IVR Configurations



IVR Design Targets

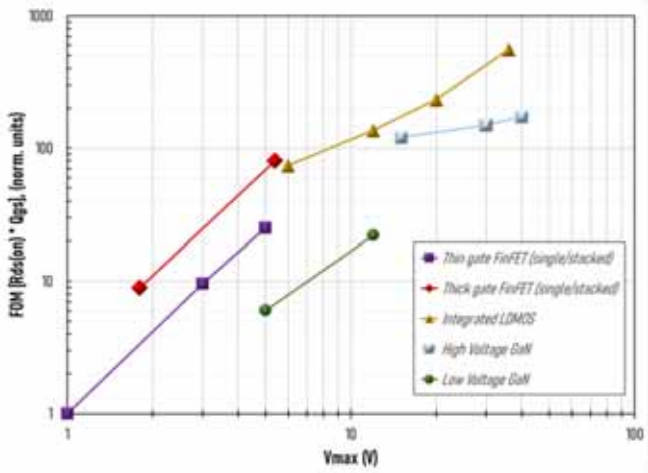
Metric	Target
Input Voltage	> 5 V
Output Voltage	0.6 – 1.2 V
Current Density	> 10 A/sqmm
Target Efficiency	> 90 %
Switching Freq	10-25 MHz

Technology Ingredients

- Device Technology for HV-IVR
- Magnetic Inductors
- Advanced Decoupling Solutions
- Advanced Packaging Technology

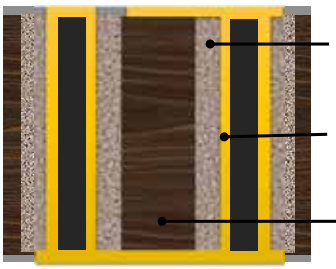
Device Technology for HV-IVRs

- Device performance for power FETs is quantified using Figure of Merit
  - FoM = $R_{ds(ON)}$ (Conduction Loss) x Q_{gs} (Switching Loss)
- Thin gate FinFET and low voltage GaN are promising candidates for switching FETs in an IVR

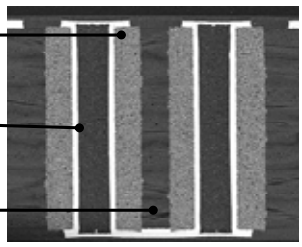


HV-IVR Technology – Coax MIL

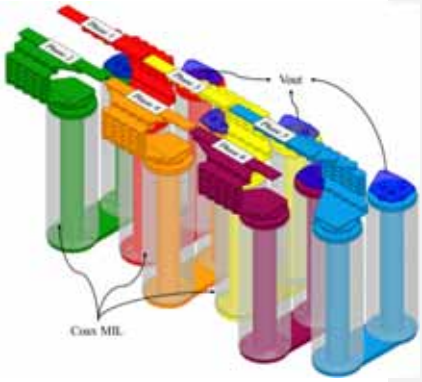
Coax MIL Concept



Coax MIL Prototype



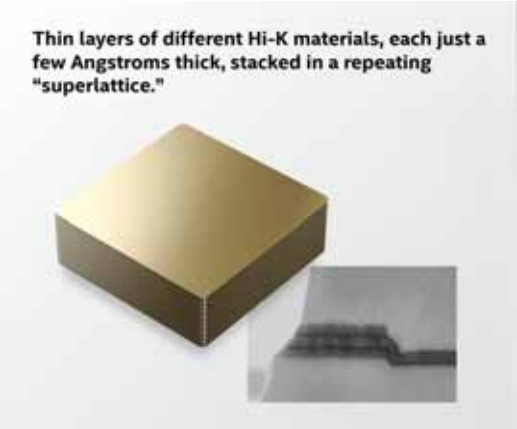
Product Implementation



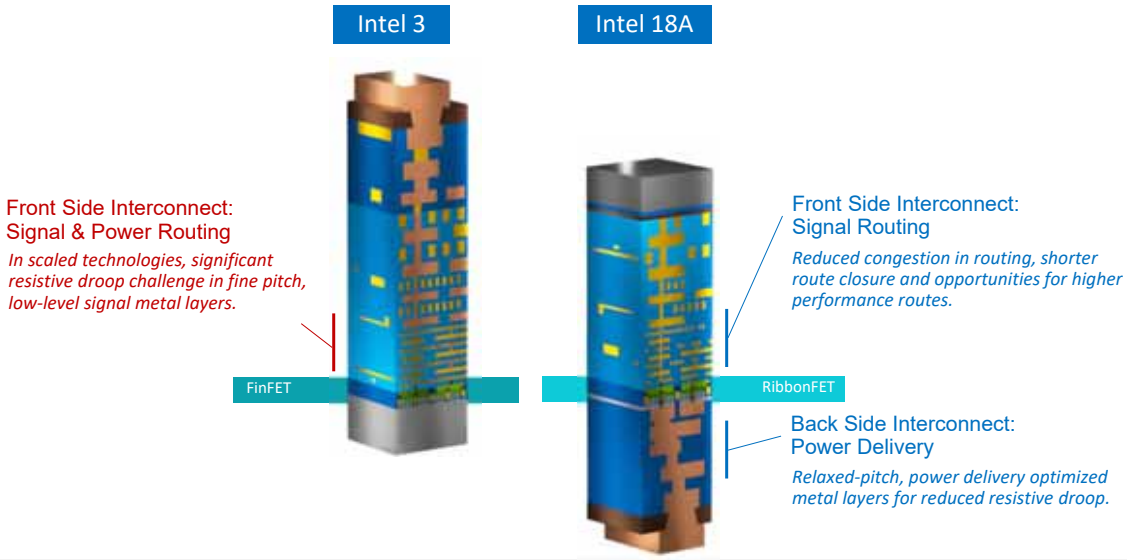
HV-IVR Technology – SuperMIM

SuperMIM Capacitor

5 x increase in MIM capacitance

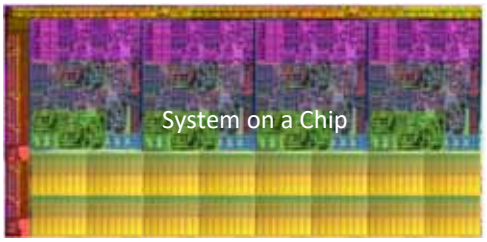


PowerVIA



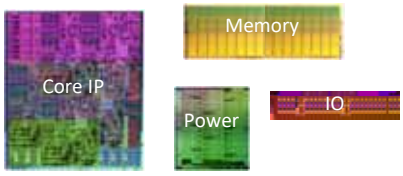
Monolithic Scaling Challenges

Monolithic SoC



Metric	Monolithic SoC
Yield	Inverse relationship to die size
IP Customization	Not possible
Performance	Good
Post-Silicon	Significant debug work
Cost	High on latest process nodes

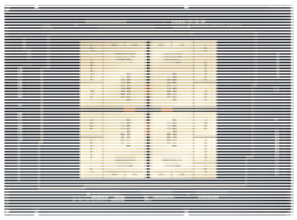
Heterogeneous System



Metric	Heterogeneous System
Yield	Good due to small die size
IP Customization	Each IP is on an optimized node
Performance	Need to manage D2D IO
Post-Silicon	Fewer bugs, improved TTM
Cost	Overall cost is lower

Advanced Packaging

Heterogenous Systems



Intel Sapphire Rapids Next Gen Xeon

Built for the Datacenter

~1600 mm^2 Silicon Footprint
4+ Individual Chiplet Tiles

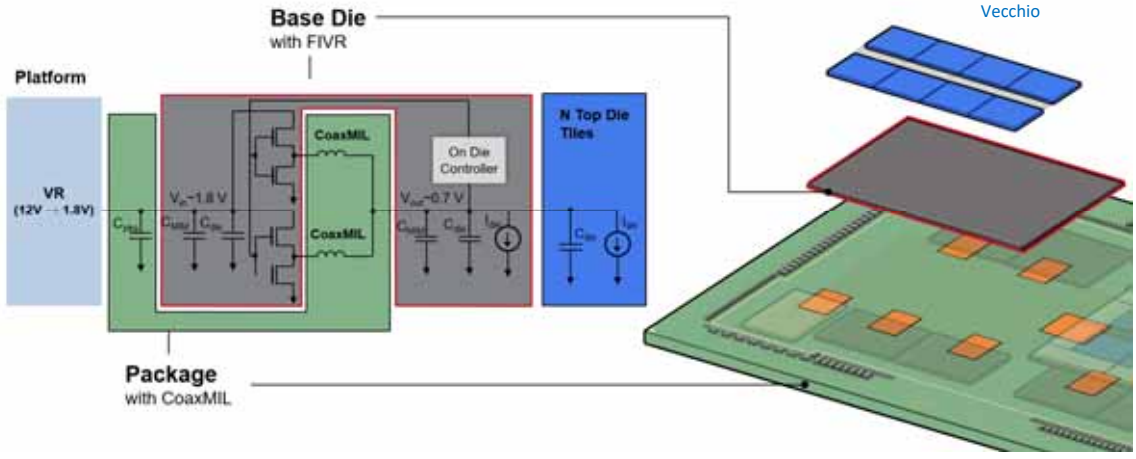
Intel Ponte Vecchio XPU

Built for HPC Clusters
3100 mm^2 Total Silicon Footprint
47 Individual Chiplet Tiles
5 Process Technologies

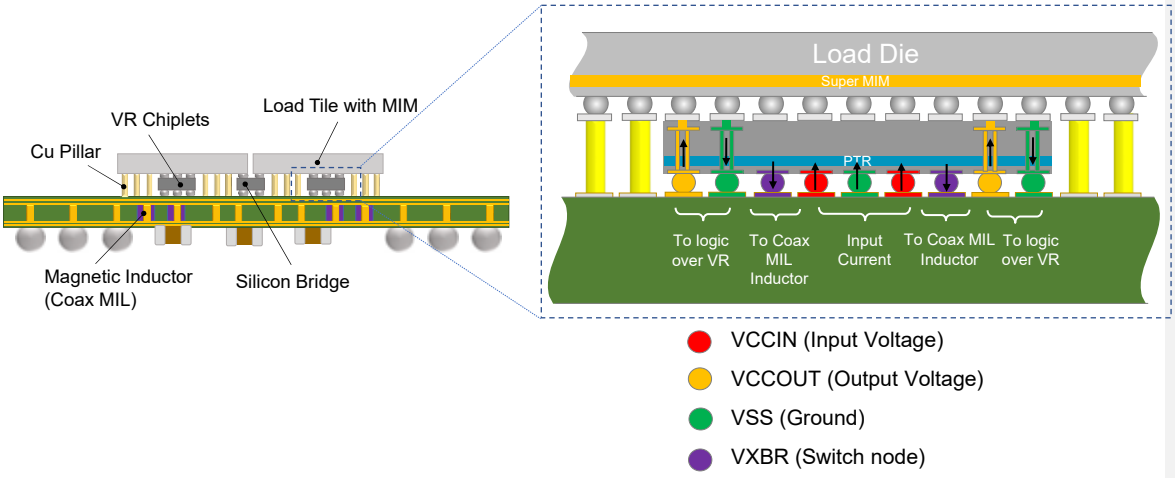
Large systems are moving beyond reticle limits to reduce data movement
Expect to continue to increase in-package silicon footprint to drive energy efficient systems

3D FIVR with Foveros

Ponte Vecchio



HV-IVR with Foveros Omni



Summary

- Delivering a clean power supply to the microprocessor is important to optimize microprocessor performance
- Classical or Dennard scaling ensured exponential performance growth without increasing power density until the mid-2000s
  - Multi-core architecture introduced to scale performance in the post-Dennard scaling era
  - Heterogeneous Integration has emerged as a key vector in extending Moore's law
- The increase in number of power rails & overall microprocessor power levels introduces new power delivery challenges
- Development of high efficiency, high density IVRs will be critical to meet the requirements of future high-performance microprocessors
  - Intel is developing a number of building block power delivery technologies such as PowerVIA, SuperMIM, Coax MIL Inductors, and GaN VR to enable this