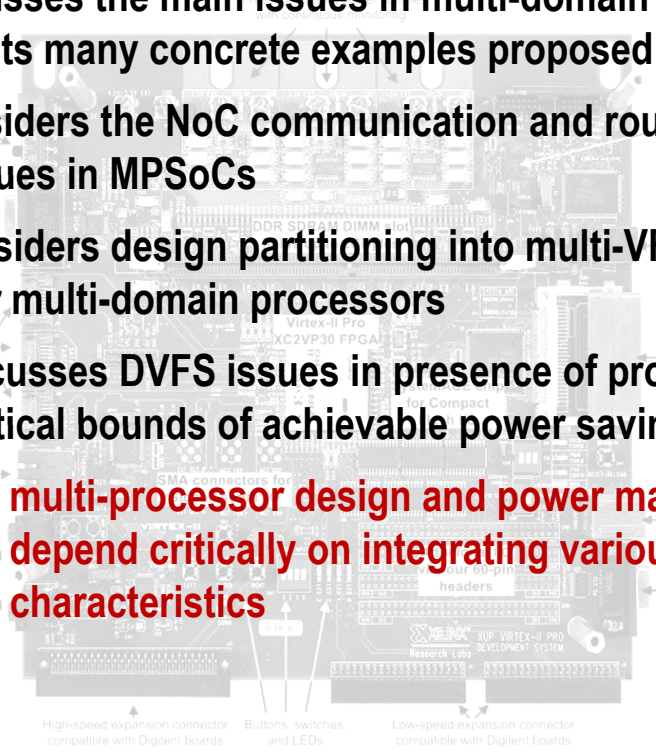
Multi-Domain Processors Design Overview

Stefan Rusu

Big picture

- Part I discusses the main issues in multi-domain processor design and presents many concrete examples proposed to date
- Part II considers the NoC communication and router design, as well as QoS issues in MPSoCs
- Part III considers design partitioning into multi-VFIs and control policies for multi-domain processors
- Part IV discusses DVFS issues in presence of process variations and theoretical bounds of achievable power savings
- Low-power multi-processor design and power management techniques depend critically on integrating various platform and application characteristics



ISCA-2010 Tutorial #2

Multi-Domain Processors Design Overview

Stefan Rusu

Intel Corporation

stefan.rusu@intel.com

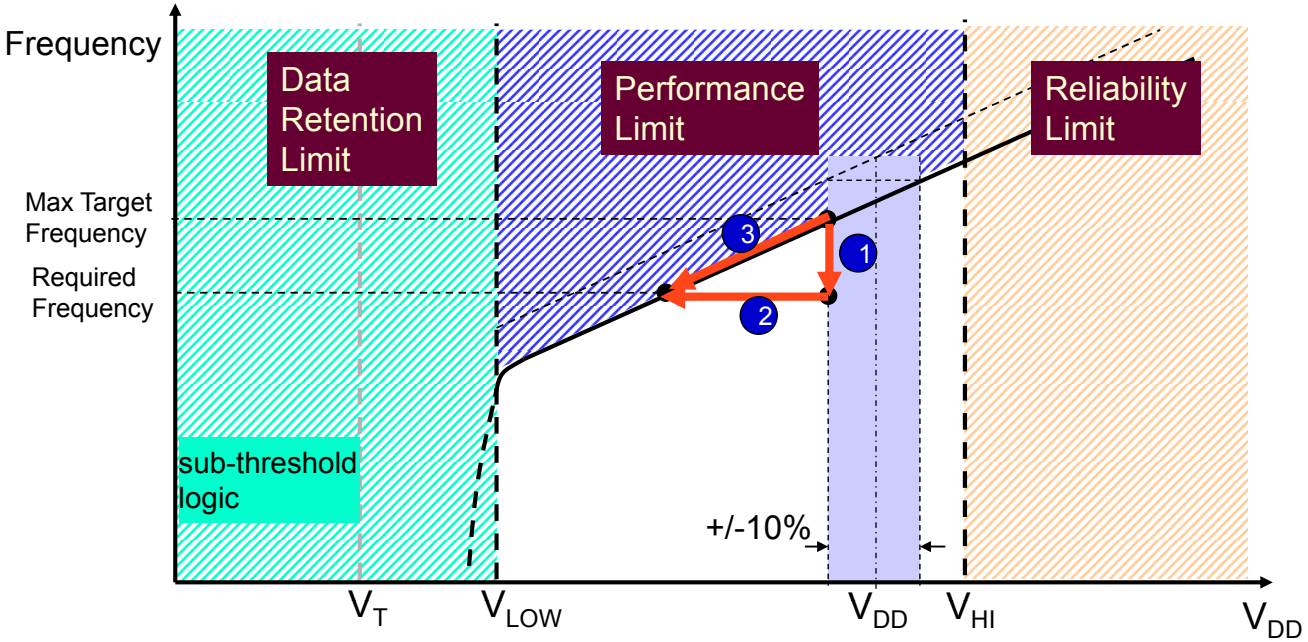
Why Multi-Domain Processors?

- Adapt supply voltage and clock frequency to the underlying circuit needs
  - Special supply voltage for large cache SRAM arrays
  - Adapt clock frequency to standard interface needs (e.g. DDR800)
- Reduce operating voltage and frequency to save power in idle or lower performance blocks
  - Power and clock gating in idle circuit islands
  - Run last level cache at half frequency
- Clock and voltage knobs are usually used together to enable a cubed power reduction factor
  - A 10% reduction in both voltage and frequency provides a 30% reduction in power consumption

Multi-Domain Processors Design Overview

- Voltage / frequency scaling basics
- Multi-domain server processors
  - Power gating
  - Core and cache recovery
  - Split vs. connected supplies
  - Globally Asynchronous, Locally Synchronous (GALS)
- Cell phone processors
- Media processors
- Dual voltage supply at the cell level
- Future directions
- Summary

Voltage and Frequency Scaling



- Case1 - Fixed V_{DD}, Frequency Scaling: Linear Power Reduction
- Case2 - Fixed MHz, V_{DD} Scaling: Square Power Reduction
- Case3 - Voltage and Frequency scaling: Cubic Power Reduction

[J. Rosal, ISSCC 2006]

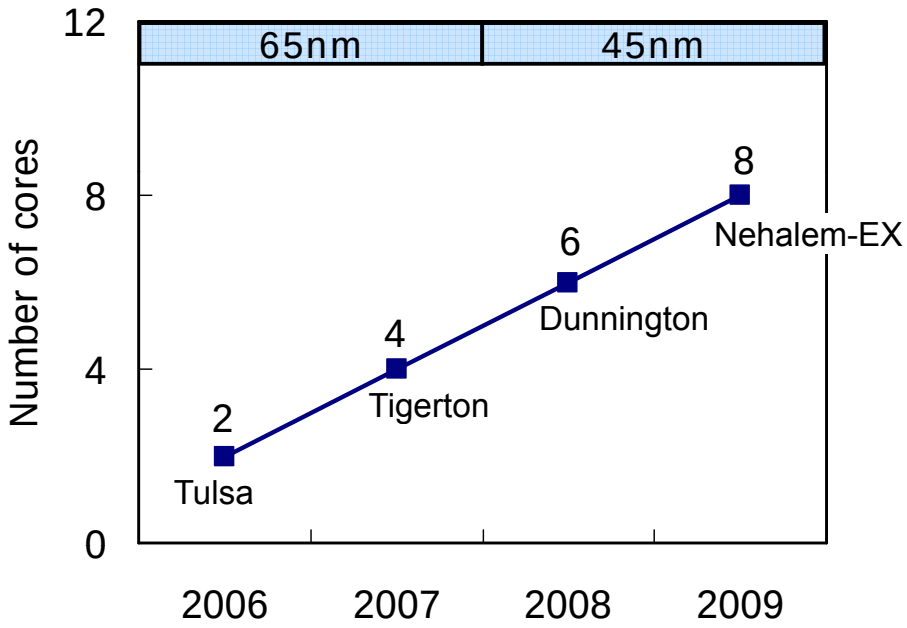
Core Power Management States

	C0 HFM	C0 LFM	C1/C2	C4	C6
Core voltage					
Core clock			OFF	OFF	OFF
PLL				OFF	OFF
L1 caches			flushed	flushed	off
L2 caches				Partial flush	off
Wakeup time	active	active	<1us	<30us	<100us
Power					

Modulating the processor core voltage and frequency enables lower power states

[Gerosa, A-SSCC 2008]

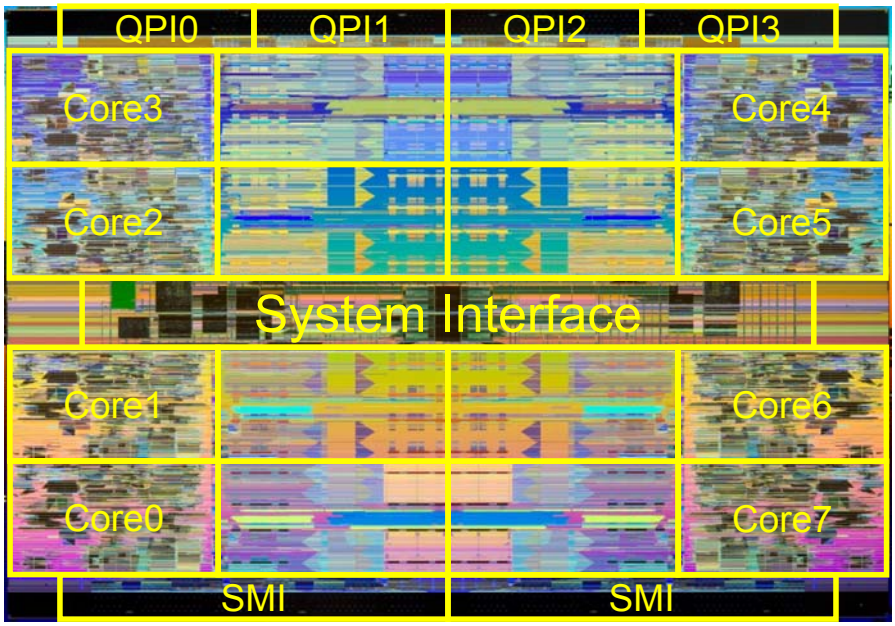
Xeon® EX Multi-Core Scaling Trend



Two additional cores every year

[Rusu, ISSCC 2009]

8-Core Xeon® Processor

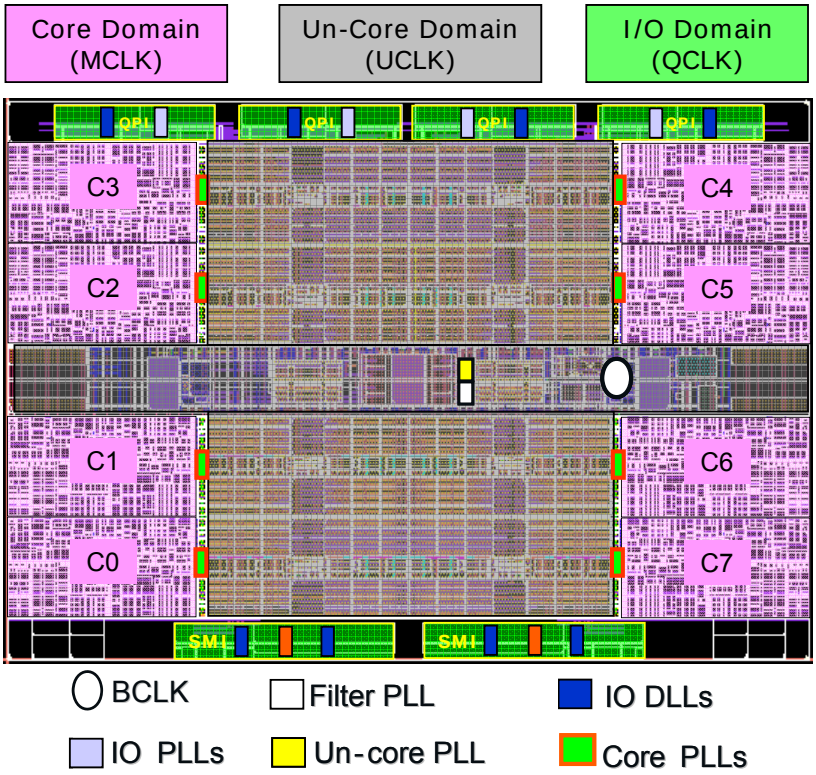


The largest device count reported for a microprocessor

2.3B transistors

[Rusu, ISSCC 2009]

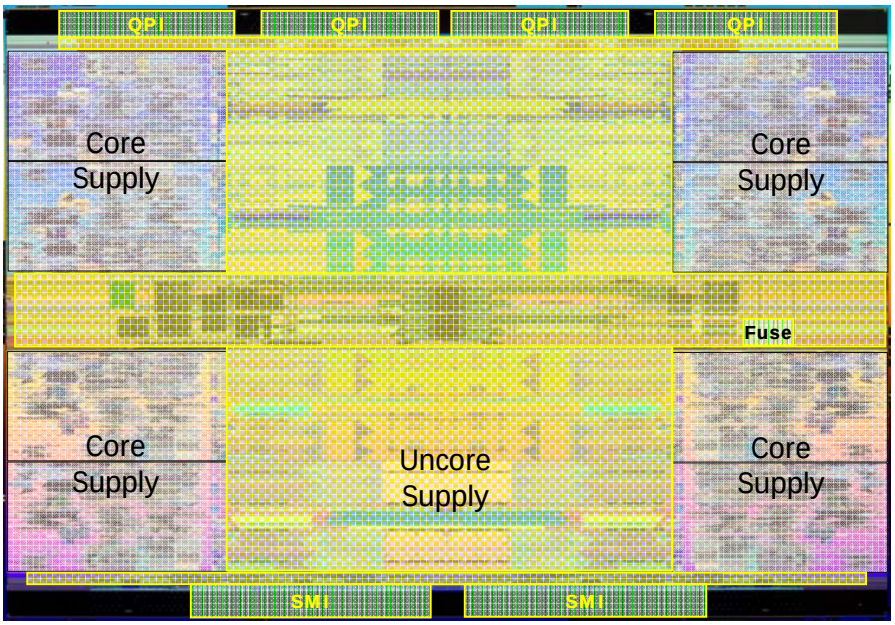
8-Core Xeon® Processor Clock Domains



- 3 primary clock domains:
  - Core
  - Un-core
  - I/O
- 16 PLLs & 8 DLLs
  - Single system clock input (BCLK)

[Rusu, ISSCC 2009]

8-Core Xeon® Processor Voltage Domains



I/O Domain
1.1V
fixed

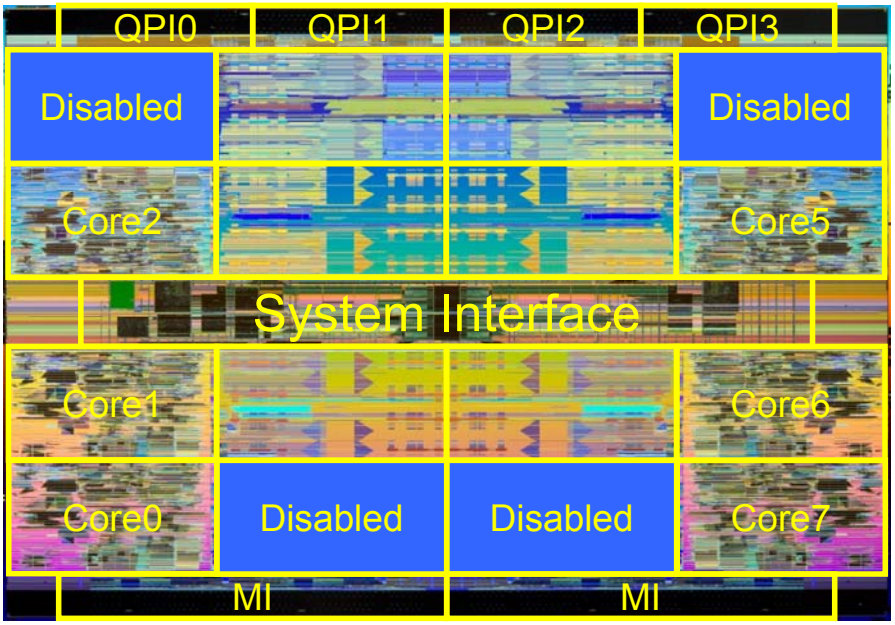
Un-Core Domain
0.9-1.1V
fixed

Core Domain
0.85-1.1V
variable

Multiple voltage domains minimize power consumption across the core and uncore areas

[Rusu, A-SSCC 2009]

Core and Cache Recovery Example

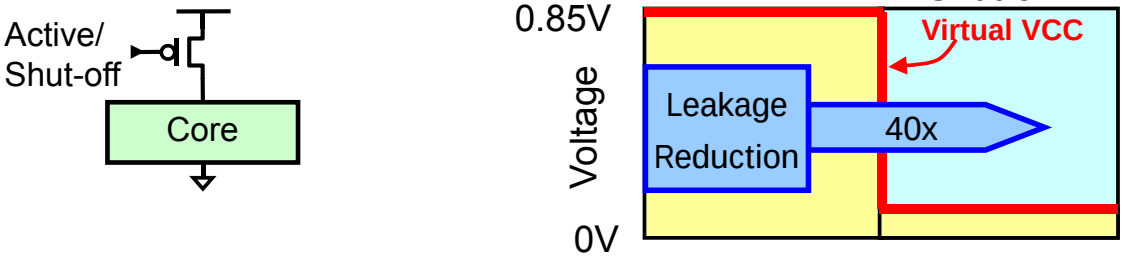


Disabled 2 cores and 2 cache slices

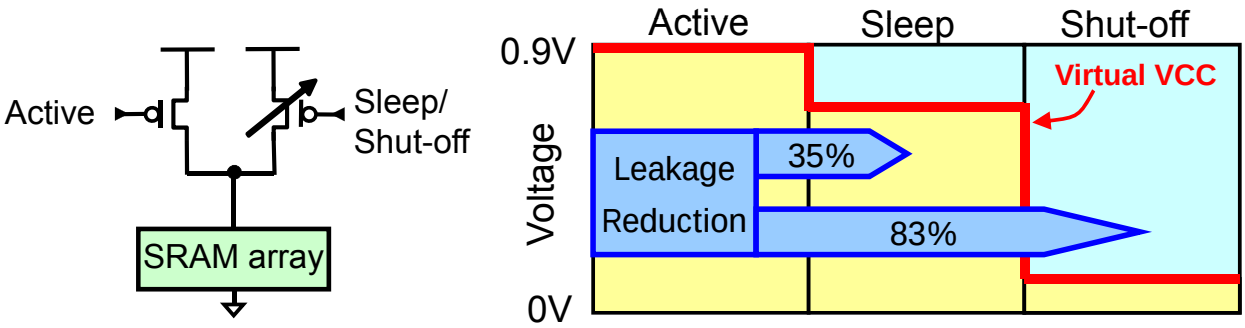
[Rusu, ISSCC 2009]

Minimize Power in Disabled Blocks

- Disabled cores ► Power gated



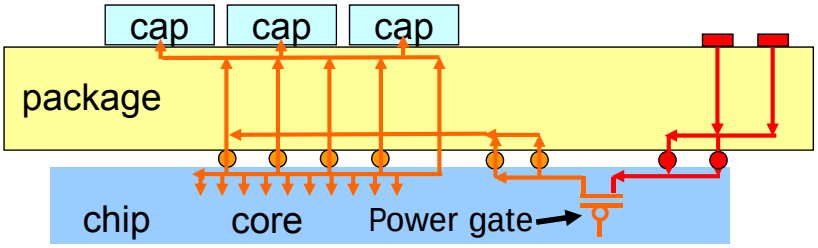
- Disabled cache slices ► All major arrays in shut-off

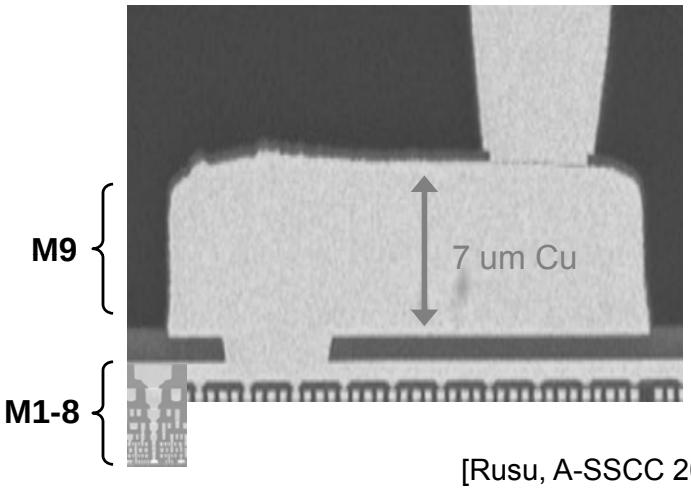


[Rusu, ISSCC 2009]

Core Power Gating

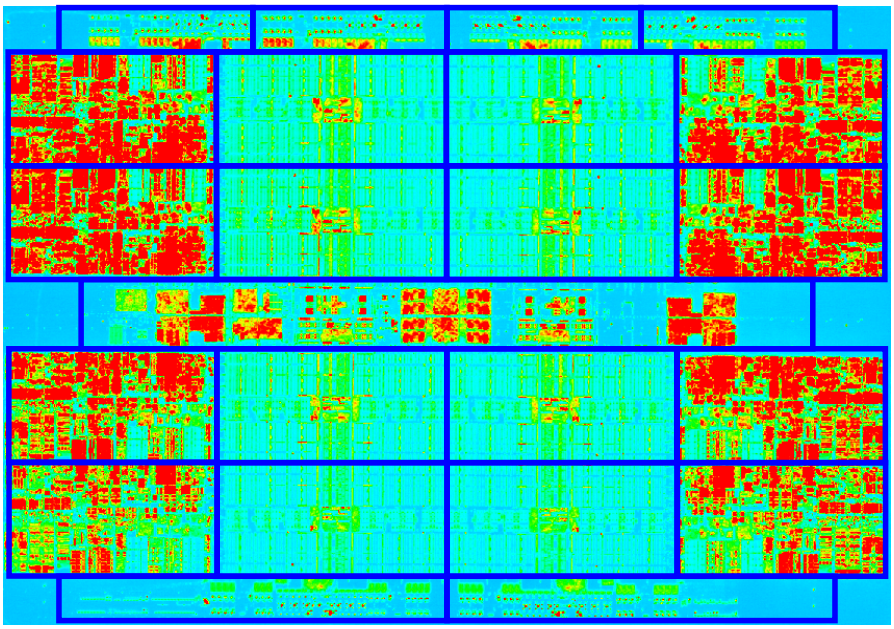
- Resistance target: less than 1% performance loss
- New package-like metal layer on silicon was developed
- M9 has ~10X lower resistance than M8





[Rusu, A-SSCC 2009]

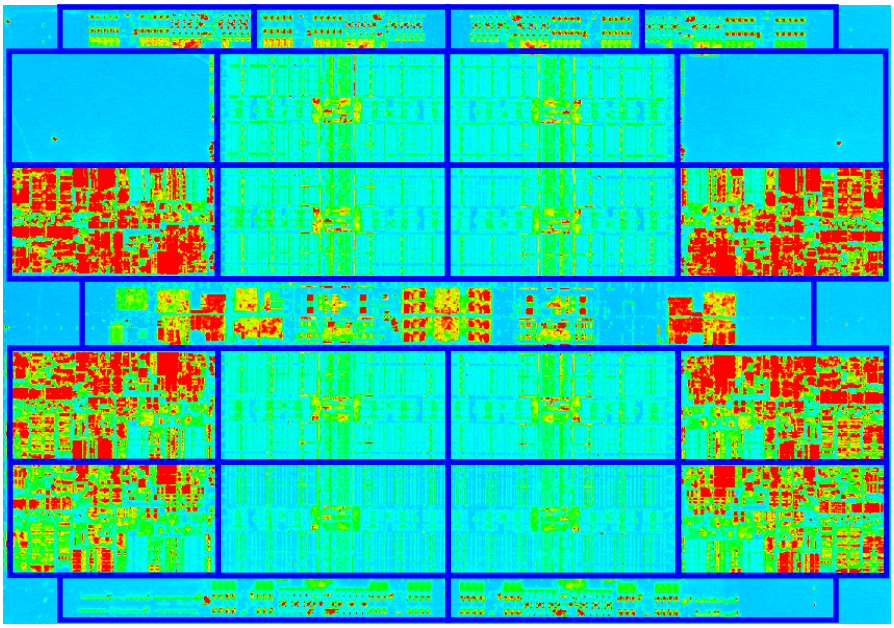
Core and Cache Recovery – Infrared Image



All cores and cache slices are enabled

[Rusu, ISSCC 2009]

Core and Cache Recovery – Infrared Image

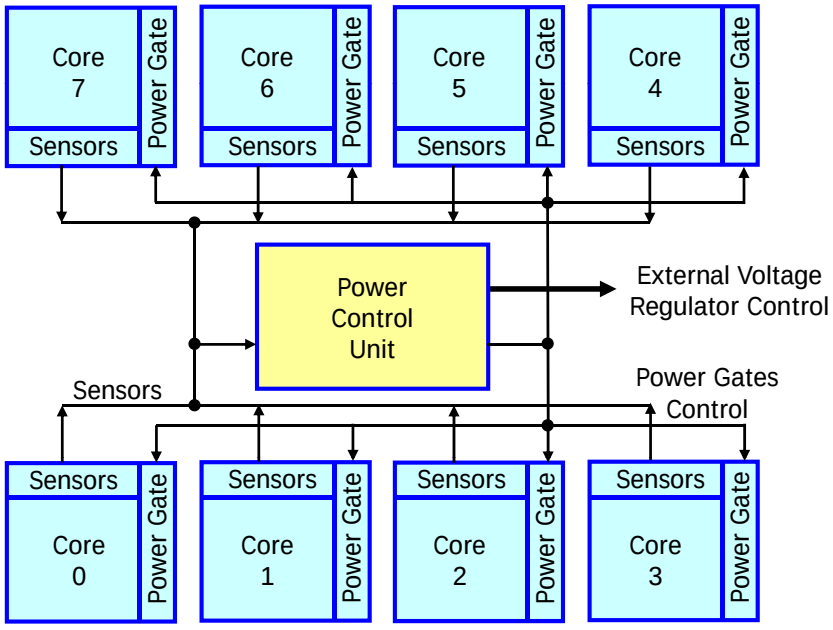


Shut-off 2 cores (top row) and 2 cache slices (bottom row)

Disabled blocks are clock and power gated

[Rusu, ISSCC 2009]

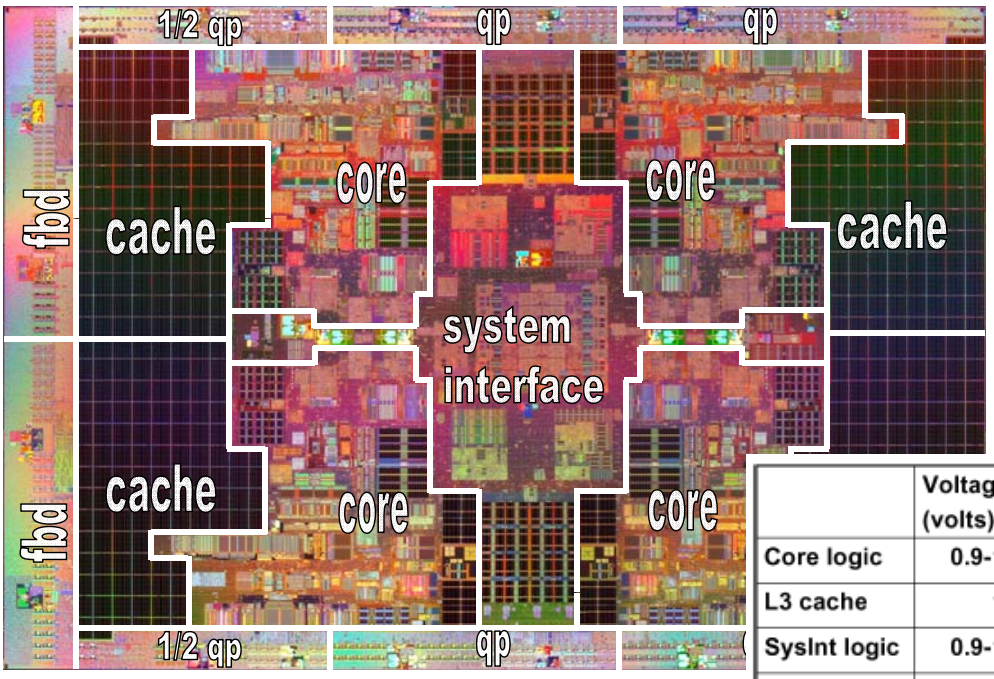
Power Management Unit



PMU controls processor voltage and frequency based on compute loading and thermal data

[Rusu, A-SSCC 2009]

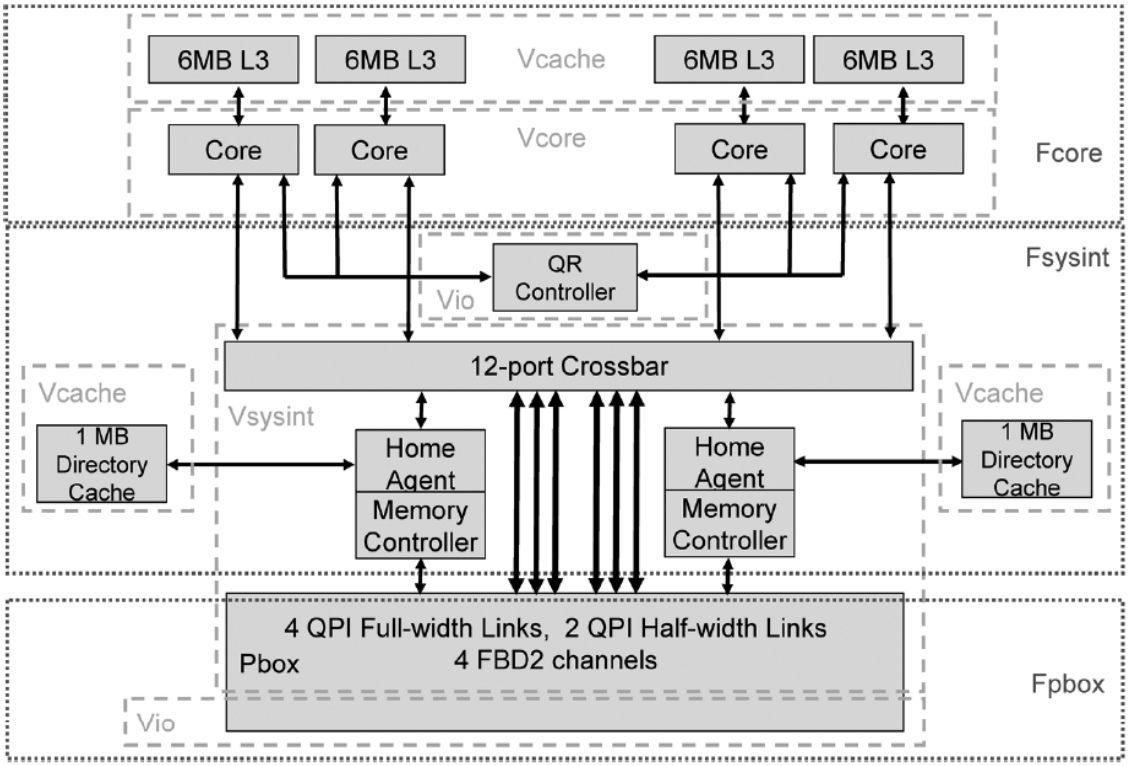
65nm Itanium® Processor Domains



	Voltage (volts)	Frequency (GHz)
Core logic	0.9-1.2	Variable
L3 cache	1.1	Variable
SysInt logic	0.9-1.2	2.4
IO logic	1.1	2.4
QR logic	1.1	0.8

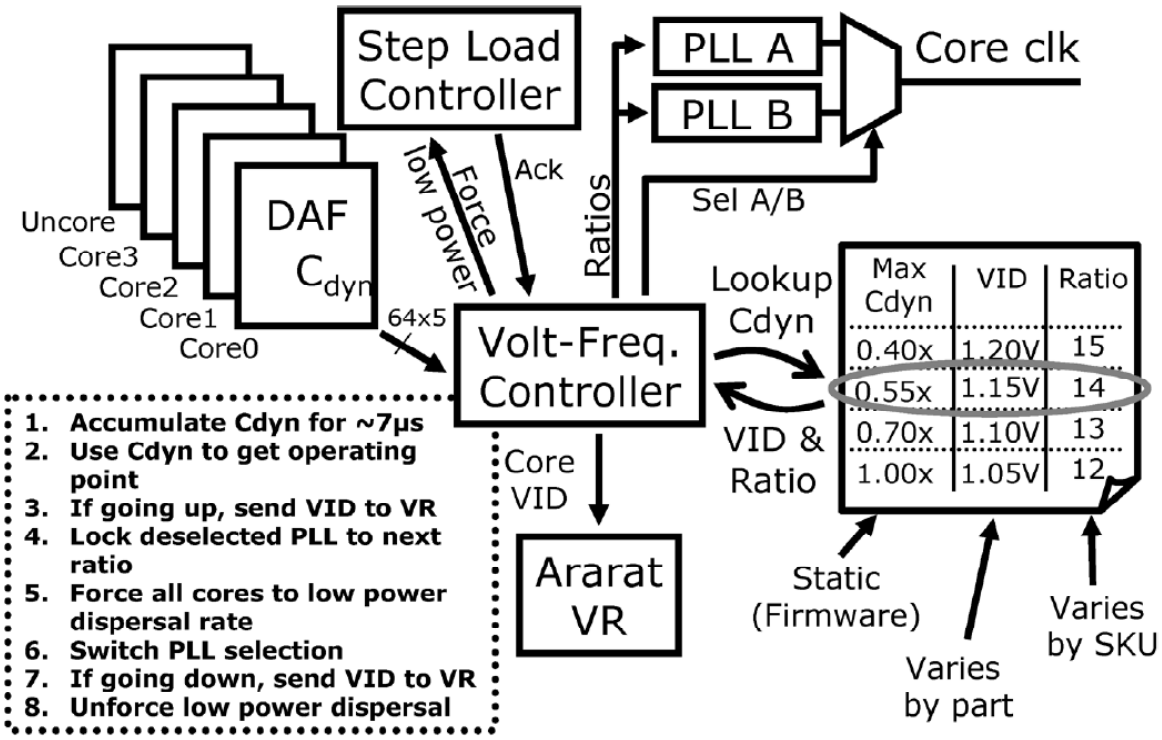
[Stackhouse, JSSC 2009]

65nm Itanium® Processor Domains



[Stackhouse, JSSC 2009]

Adaptive Voltage and Frequency Control



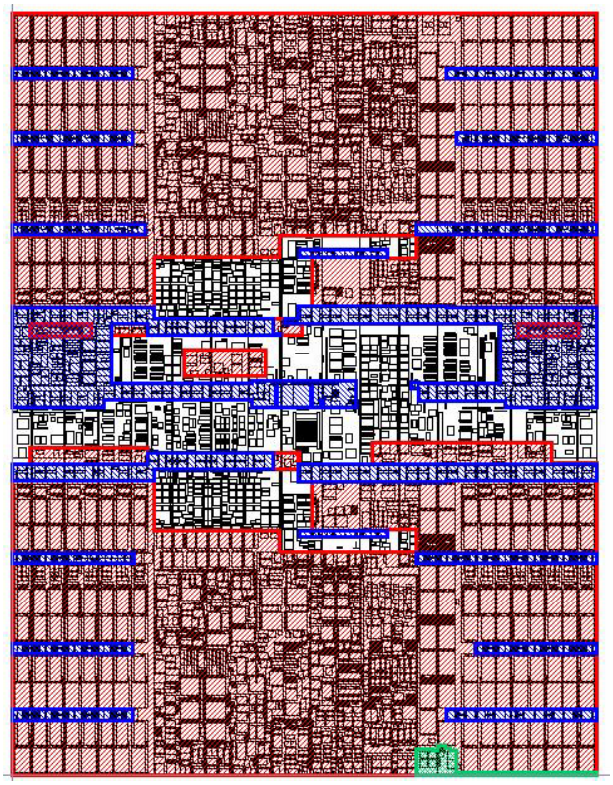
[Stackhouse, JSSC 2009]

IBM POWER6 Multi-Rail Design

- POWER6 infrastructure contains 4 voltage domains

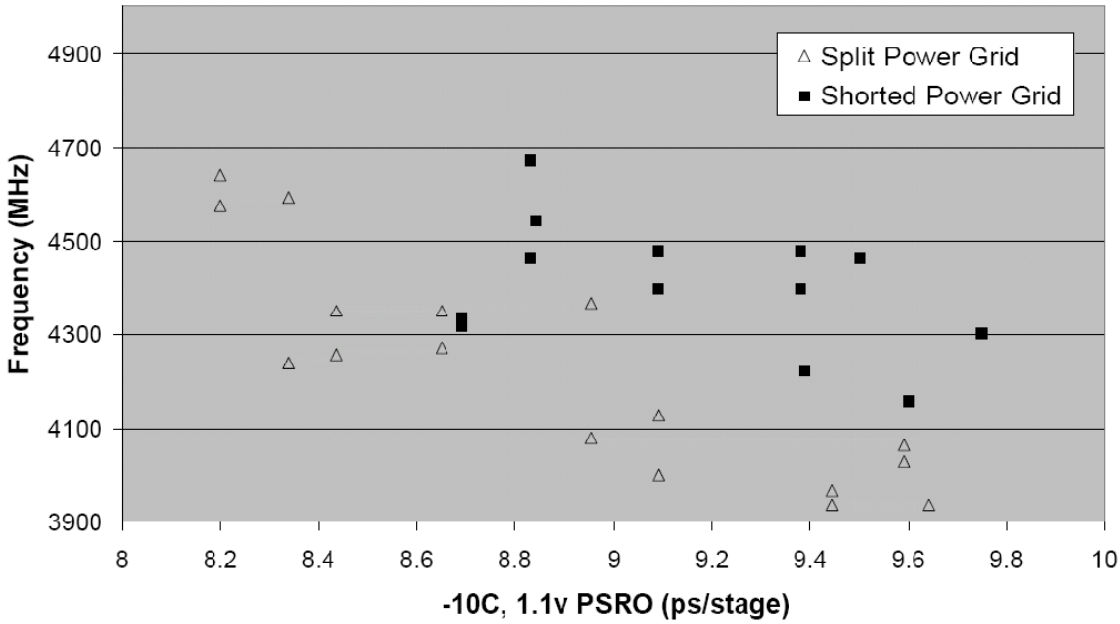
Rail	Purpose	Plot Color
VDD	Logic	All
VCS	Array	Red
VIO	IO, PLL, MC	Blue
VSB	Power-up	Green

- Multi-rail power grid defined based on macro current requirements and iterative IR analysis of each rail.
- Voltage domain of macros and global signals explicitly specified in RTL and validated by checking tools.



[J. Friedrich, ISSCC 2007]

Split vs. Connected Core Supplies



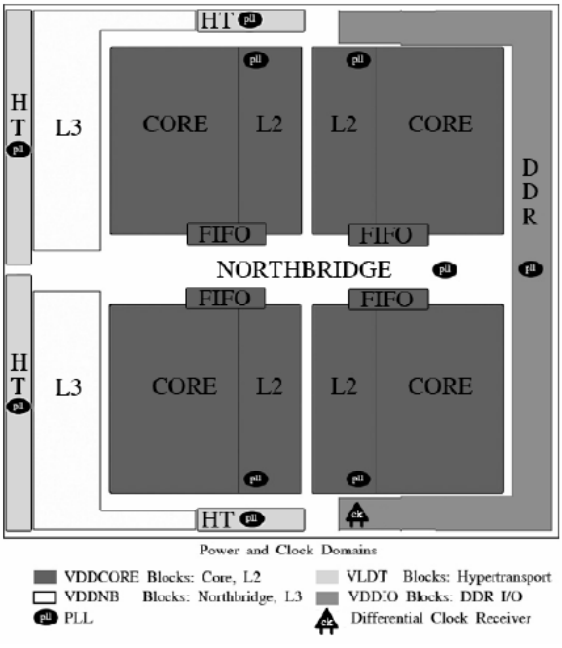
Performance sort ring oscillators (PSRO) show that the connected power chips exhibit a 3-5% fmax improvement

[N. James, ISSCC 2007]

AMD Barcelona Processor Voltage Domains

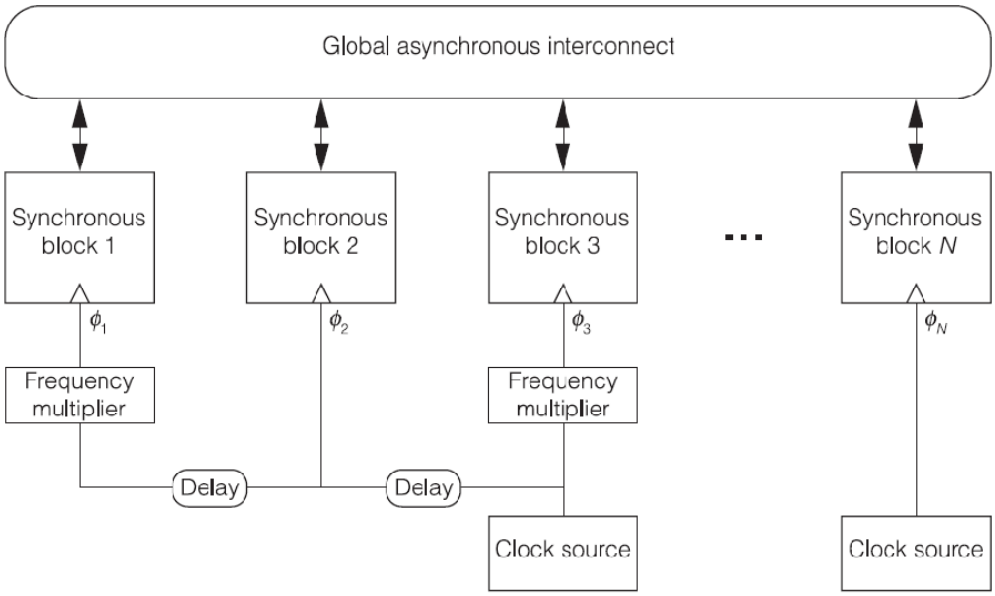
Multiple supplies for power optimization and isolation

- VDDCORE: 0.8V-1.4V
  - Core and L2: 2.0GHz and up
- VDDNB: 0.8V-1.4V
  - Northbridge and L3: 75% of core
- VLDT: 1.2V
  - HyperTransport links
- VDDIO: 1.8V (VTT:0.9V)
  - DDR I/O
- VDDA: 2.5V
  - PLLs (10 across the die) + Thermal



[Dorsey, ISSCC 2007]

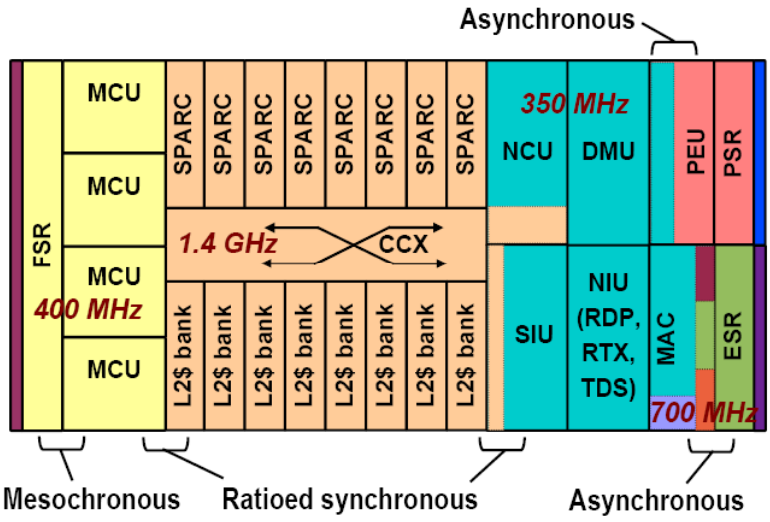
Globally Asynchronous, Locally Synchronous (GALS)



GALS methodology is a natural approach for SoC design, allowing the integration of independently designed blocks operating at different frequencies

[P. Teehan, IEEE D&T of C 2007]

Sun/Oracle Niagara2 Clocking

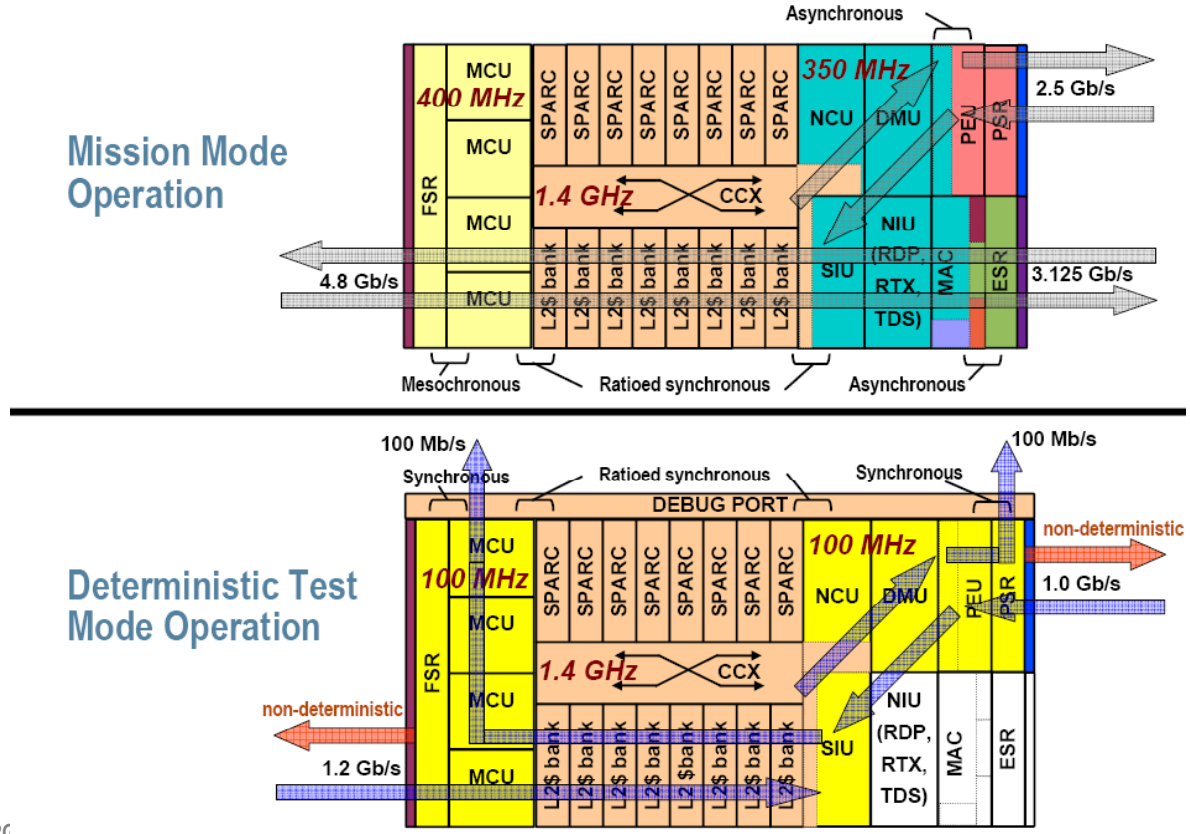


REF	133/167/200 MHz
CMP	1.4 GHz
IO	350 MHz
IO2X	700 MHz
FSR.refclk	133/167/200 MHz
FSR.bitclk	1.6/2.0/2.4 GHz
FSR.byteclk	267/333/400 MHz
DR	267/333/400 MHz
PSR.refclk	100/125/250 MHz
PSR.bitclk	1.25 GHz
PSR.byteclk	250 MHz
PCI-Ex	250 MHz
ESR.refclk	156 MHz
ESR.bitclk	1.56 GHz
ESR.byteclk	312.5 MHz
MAC.1	312.5 MHz
MAC.2	156 MHz

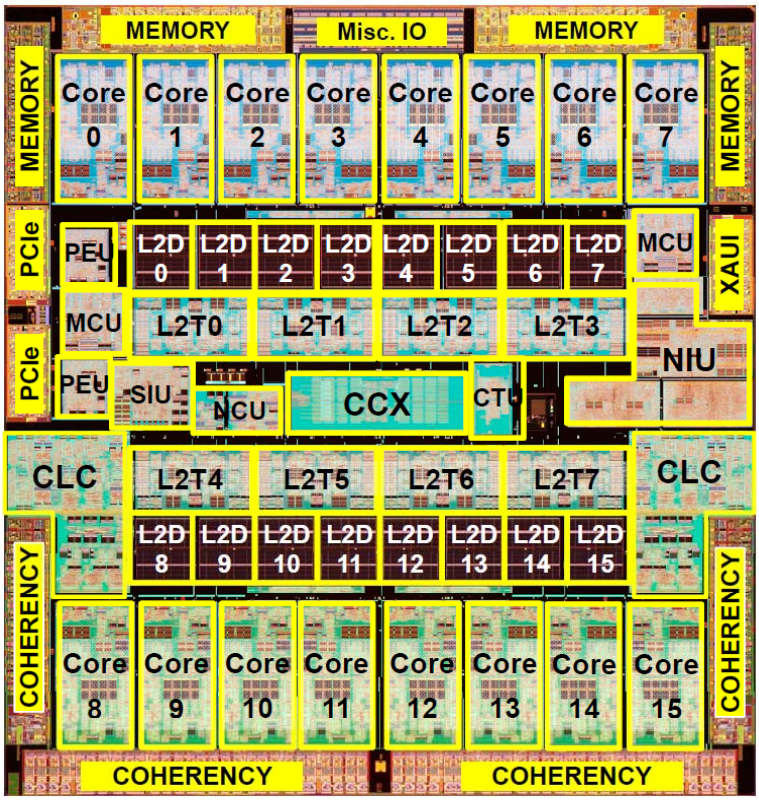
SoC processors must handle multiple clock domains for the various external interfaces

[Nawathe, ISSCC 2007]

Deterministic Test Mode



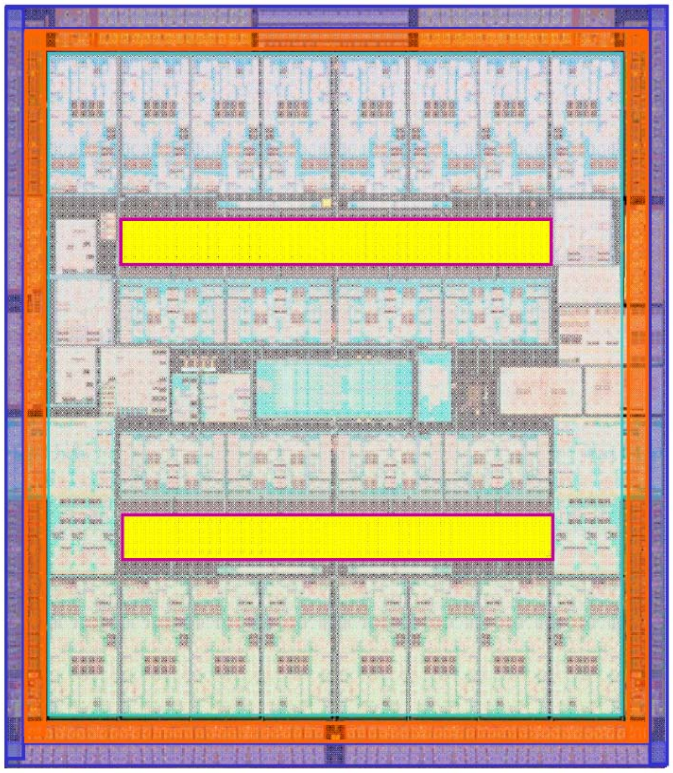
Sun/Oracle Rainbow Falls SoC Processor



- TSMC 40nm process, 11 ML
- 16 8-threaded SPARC® cores
- 16KB 8-way Icache
- 64-entry ITLB
- 8KB 4-way Dcache
- 128-entry DTLB
- Enhanced multiply/add FGU and crypto per core
- Unified 6MB 16-bank 24-way L2 cache
- Hierarchical crossbar
- 4 DDR3 channels at 6.4Gbps
- 6 coherency links at 9.6Gbps
- 2x8 PCIe 2.0 at 5GTS
- 2x10G XAUI Ethernet

[Shin, ISSCC 2010]

Sun/Oracle Rainbow Falls Voltage Domains

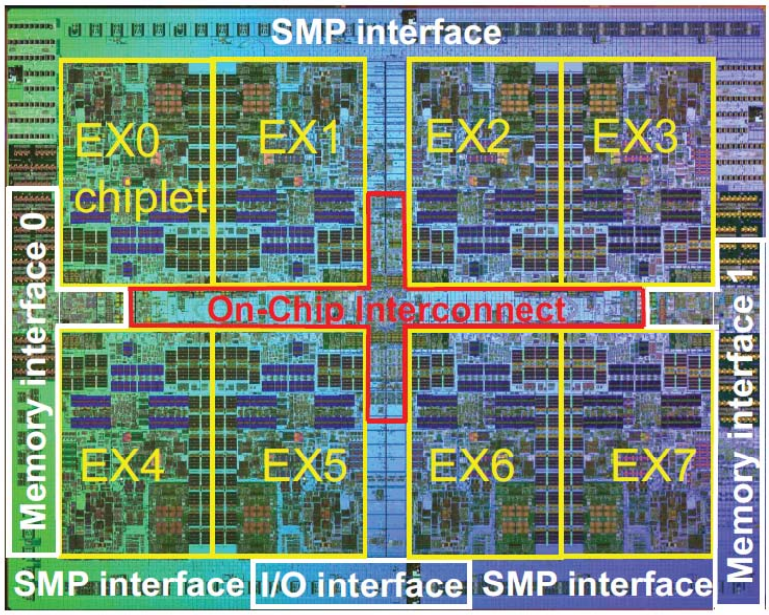


Vdd_core (VDDC) 0.8V ~ 1.1V	All Logic
Vdd_memory 0.9V ~ 1.1V	L2D Memory Cells
Vdd_analog 1.0V	PLL/SerDes
VDDT/VDDR 1.5V	I/O
VNW (VDDC +/- 0.3V)	PMOS Bias
VNWM	L2D Memory Cells PMOS Bias
VSB (VSS +/- 0.3V)	NMOS Bias

[Shin, ISSCC 2010]

IBM POWER7™ Processor

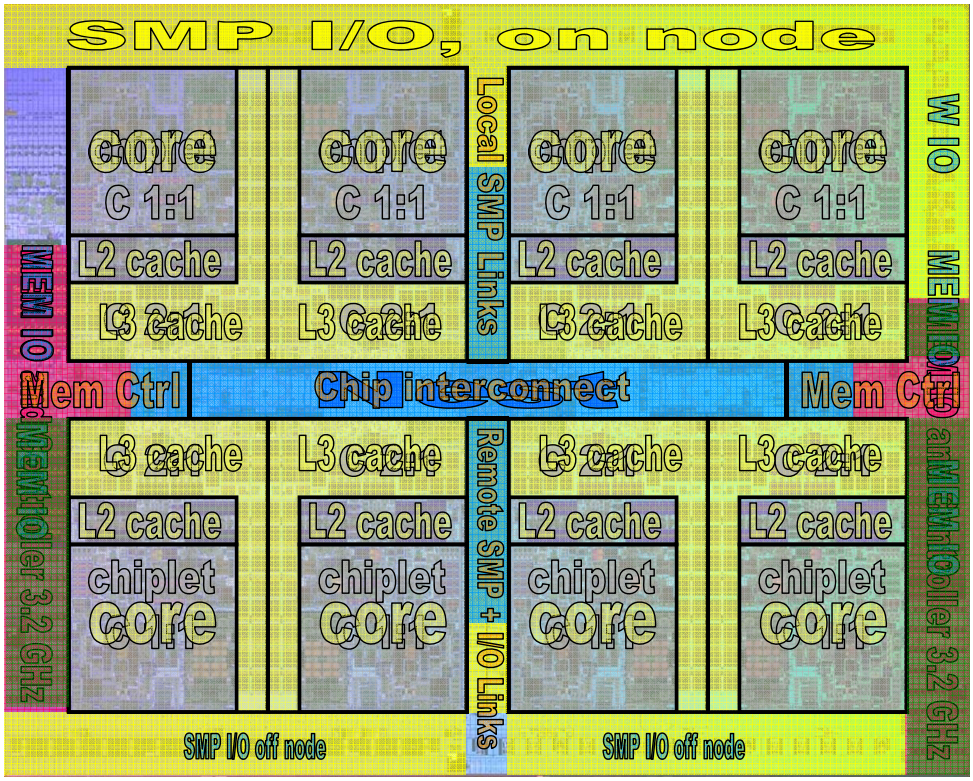
- 3.0 to 4.14 GHz
- 45nm CMOS SOI
- 567mm^2
- 1.2B transistors
- Eight processor cores
  - 12 execution units per core
  - 4 Way SMT per core
  - 32 Threads per chip
  - 256KB L2 per core



- 32MB on chip eDRAM shared L3
- Dual DDR3 Memory Controllers
  - 100GB/s Memory bandwidth per chip sustained

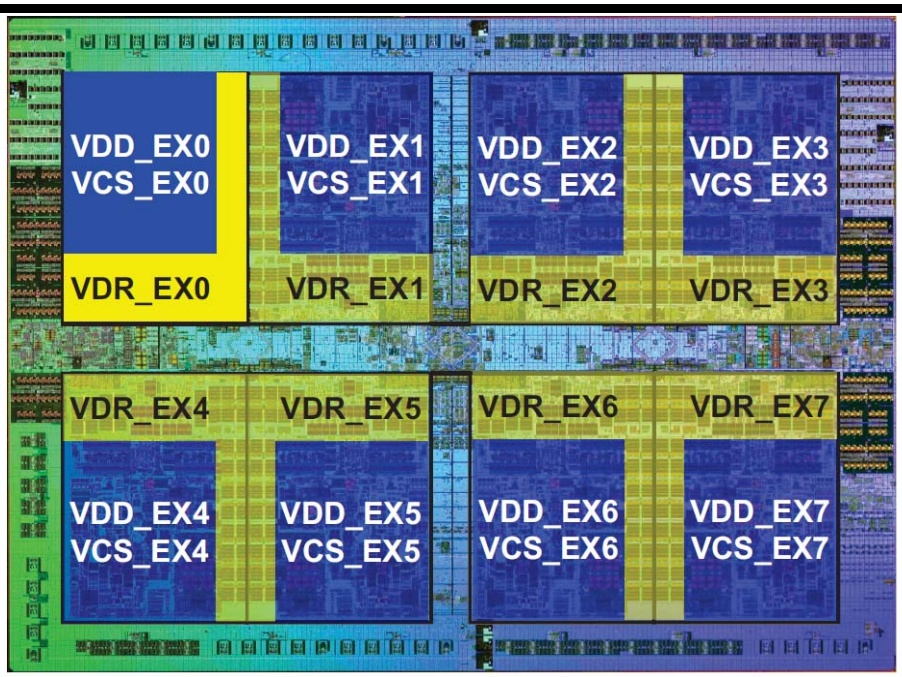
[Wendel, ISSCC 2010]

P7 Clock Domains



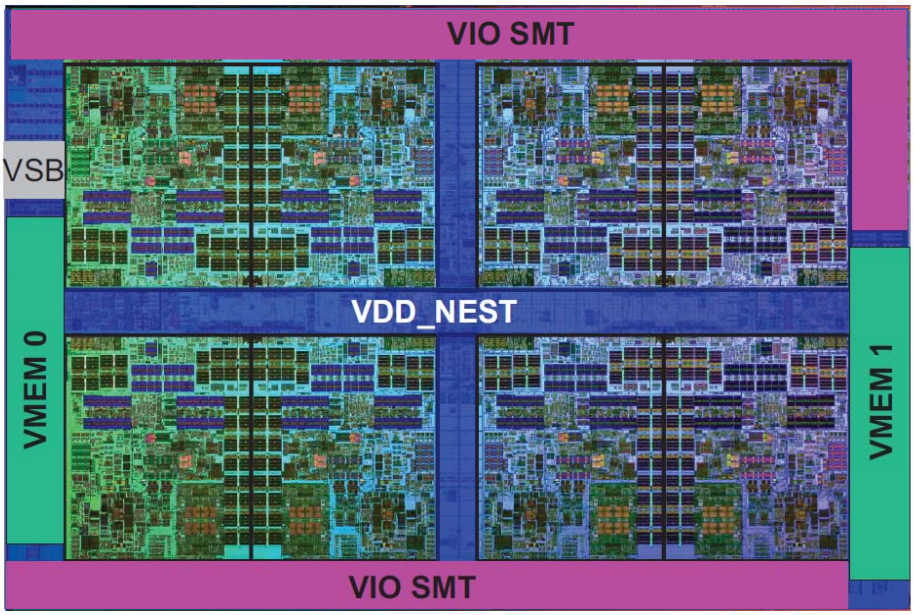
[Wendel, ISSCC 2010]

P7 Voltage Domains (1)



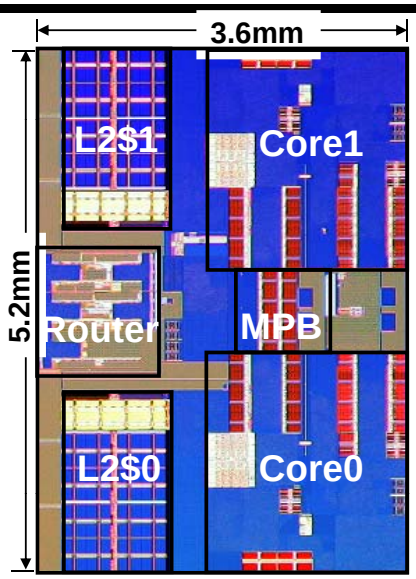
Name: # = 0 to 7	Meaning	Type of Voltage	Min.	Max.
VDD_EX#	Logic supply for Chiplet #	Adaptive, Dynamic	0.7V	1.3V
VCS_EX#, VDR_EX#	Array supply for chiplet#, DRAM supply for chiplet #	Adaptive, Dynamic	0.8V	1.3V

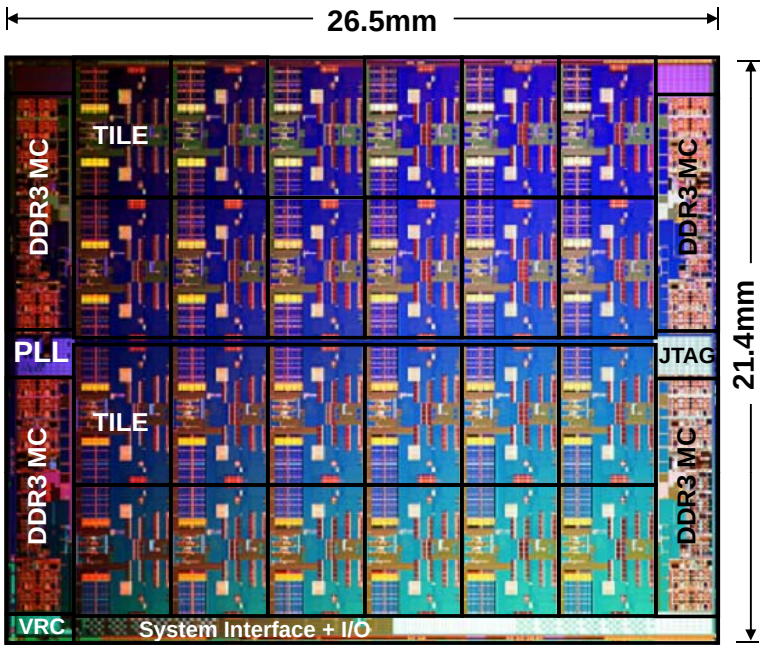
P7 Voltage Domains (2)



Name: # = 0 to 7	Meaning	Type of Voltage	Min.	Max.
VDD_NEST	Logic supply for nest & pervasive logic	Fixed, Static	1.0V	1.2V
VSB	Supply for pervasive	Fixed, Static	1.1V	1.3V
VMEM 0/1	Supply for differential IO	Fixed, Static	0.95V	1.15V
VIO SMT	Supply for IO receiver and transmitter	Fixed, Static	1.0V	1.2V

Intel 48-core SoC Processor

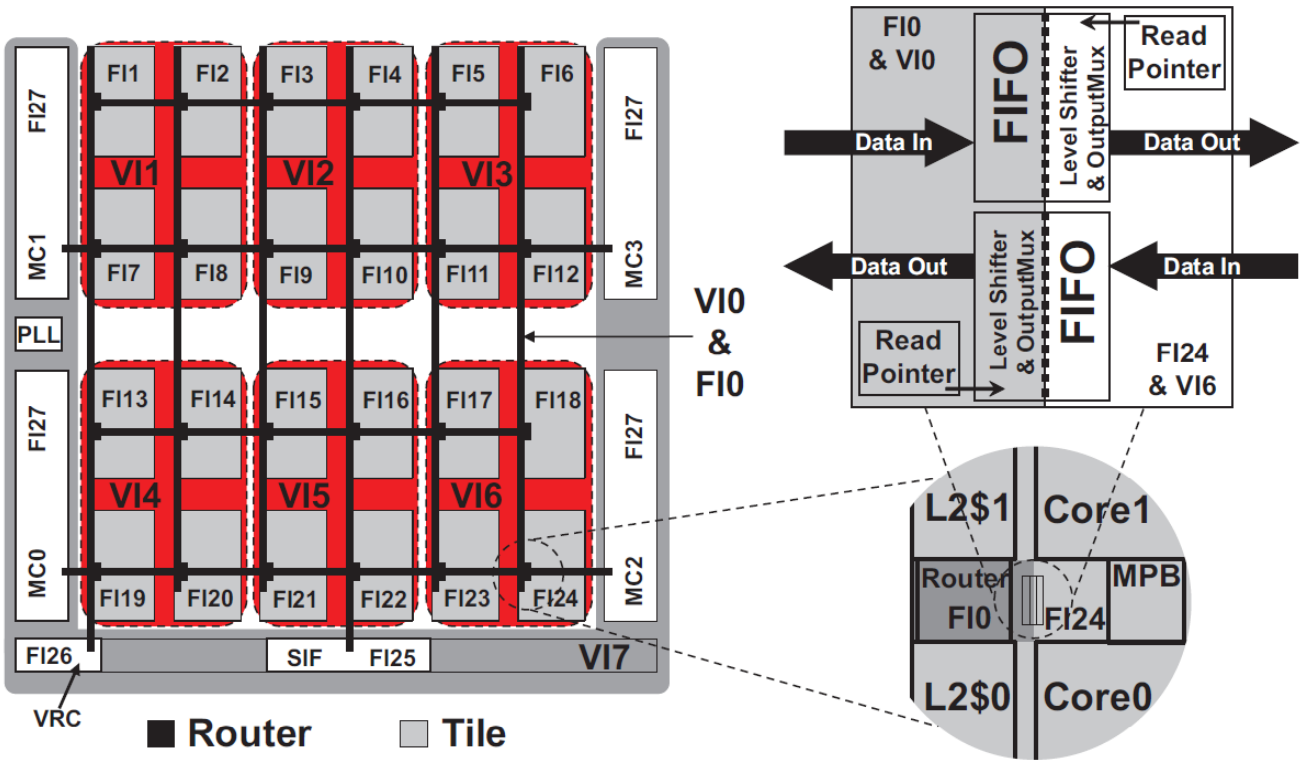




Technology	45nm Hi-K CMOS
Interconnect	9 Metal (Cu)
Transistors	Die: 1.3B, Tile: 48M
Tile Area	18.7mm^2
Die Area	567.1mm^2

[Howard, ISSCC 2010]

Voltage and Frequency Islands



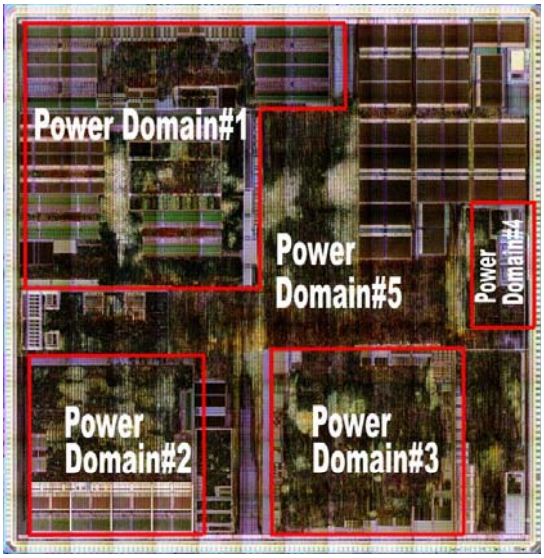
28 frequency and 8 voltage islands

[Howard, ISSCC 2010]

Multi-Domain Processors Design Overview

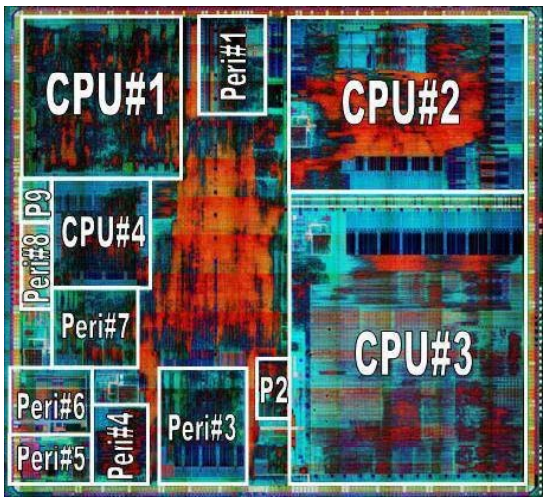
- Voltage / frequency scaling basics
- Multi-domain server processors
- Cell phone processors
- Media processors
- Dual voltage supply at the cell level
- Future directions
- Summary

TI Application Processors



- 90nm OMAP2
- 1 voltage domain
- 5 power domains

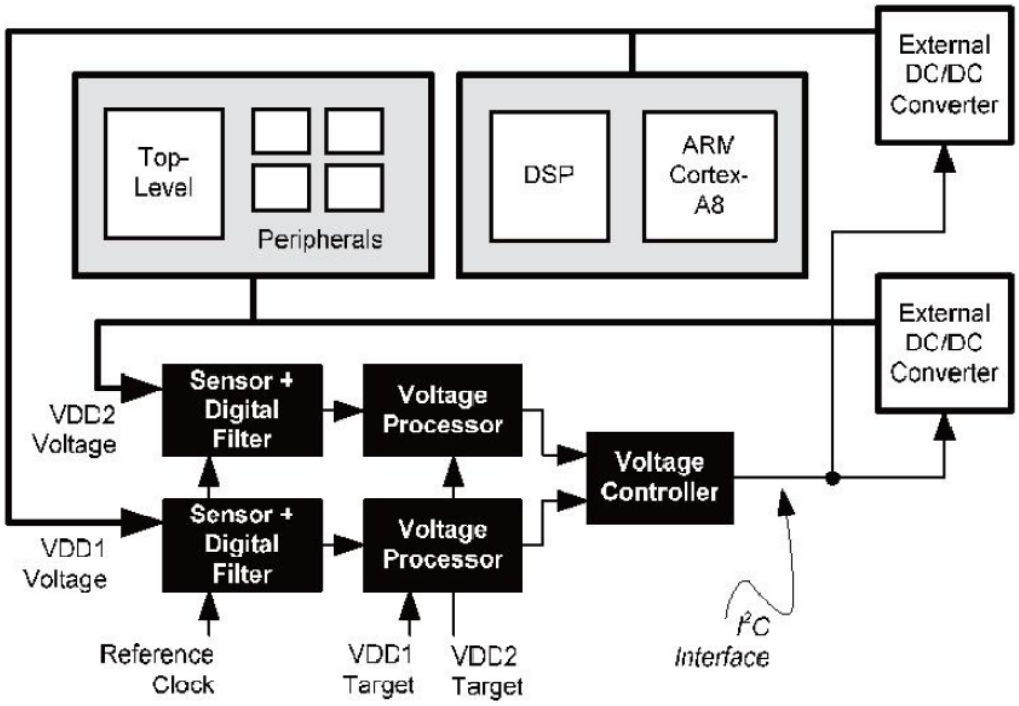
[Royannez, ISSCC 2005]



- 65nm OMAP3
- 2 voltage domains
- 11 major power domains

[Mair, VLSI Symp. 2007]

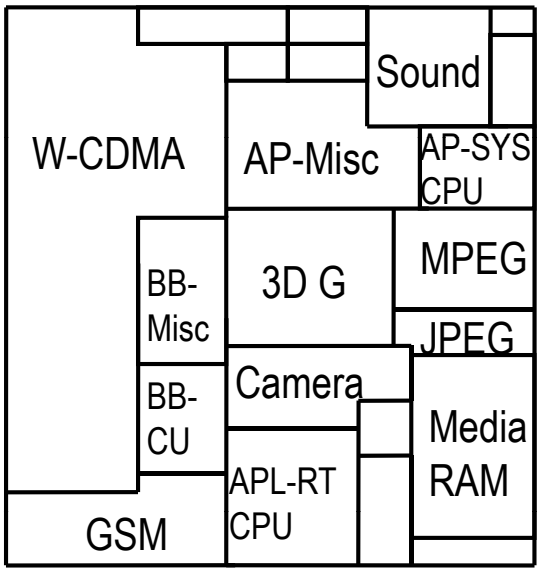
Adaptive Voltage Scaling Control Loop



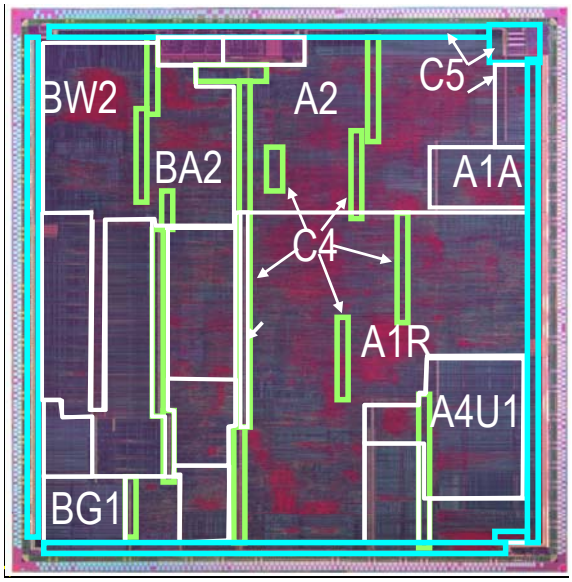
[Mair, VLSI Symp. 2007]

Renesas 90nm Cell Phone Processor

Chip Floorplan



Power Domains



20 power domains for partial power-off,
C4 (common) domain for repeaters

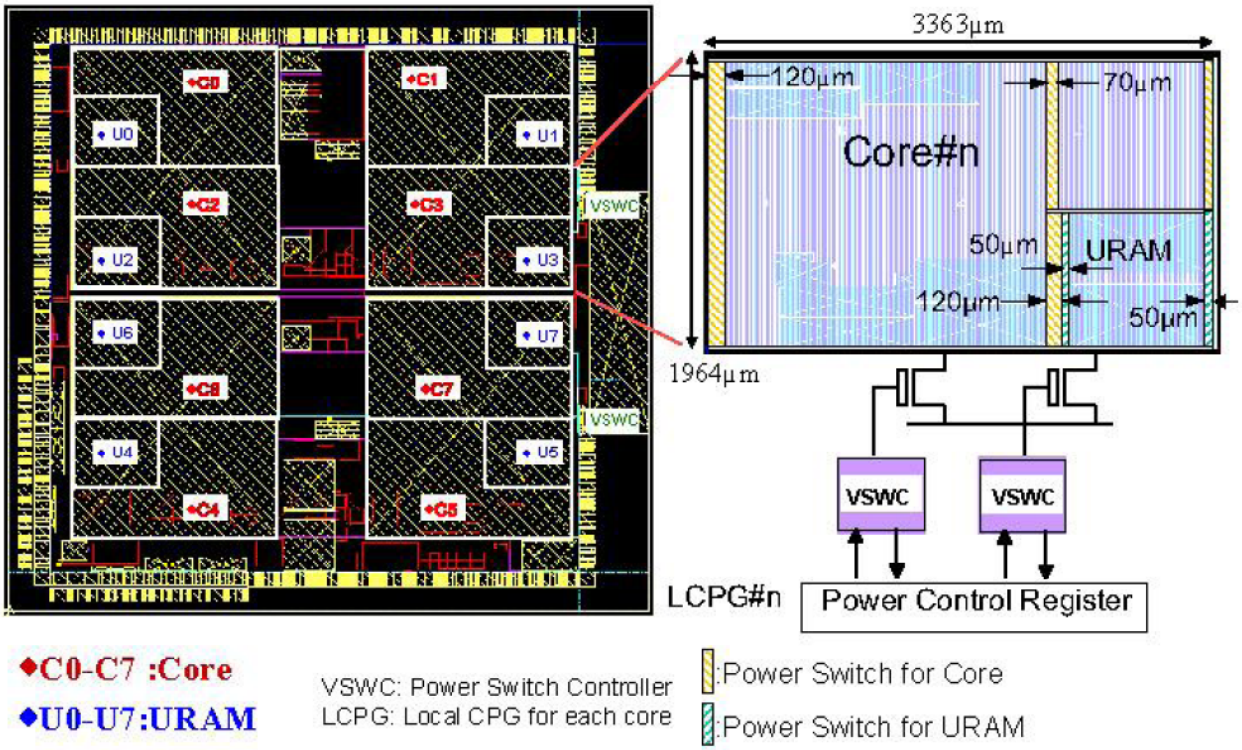
[Hattori, ISSCC 2006]

Implementation Details

# of Power domains	20 domains
# of Islands for C4 (Repeaters,CK buffers, BKUP FFs)	19 islands
# of Repeaters in C4 domain	3100 cells
# of Clock buffers in C4 domain	1600 cells
# of Backup FFs in C4 domain	2300 cells
# of μIOs (isolation cell)	20000 cells
Total area of power switch	4.2 mm^2
Power switch area ratio in the chip	3.4 %
Power-off -> power-on time (one-by-one on)	<100μSec

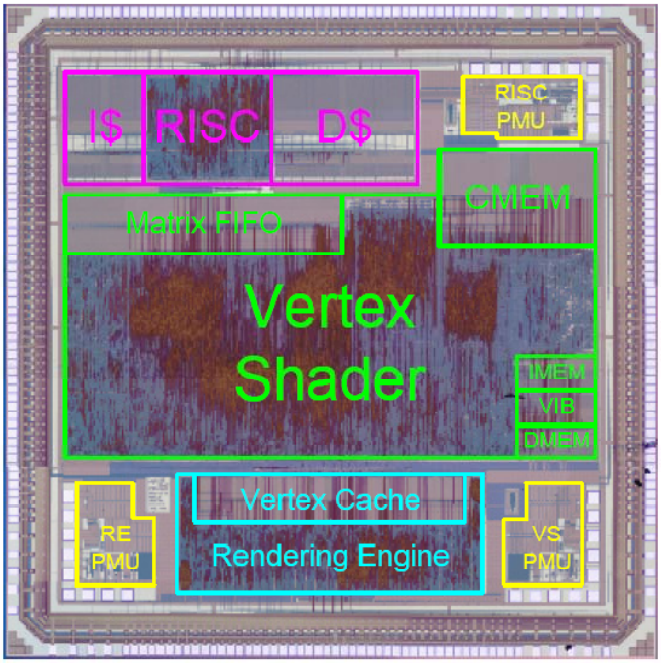
[Hattori, ISSCC 2006]

Renesas 8-Core 90nm Processor



[Ito, ISSCC 2008]

Samsung 3D Graphics Processor



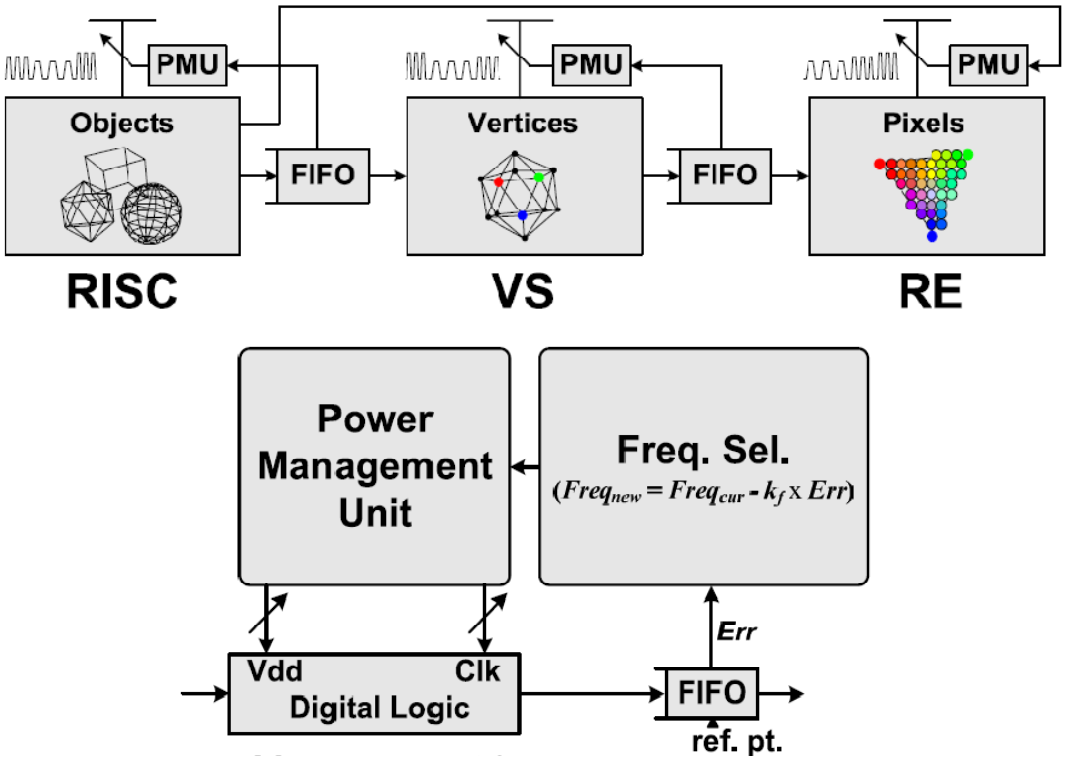
[RAMP-VI]

Chip Characteristics	
Technology	0.18um 1-P 6-M CMOS
Area	Core: 17.2mm² Die: 25mm²
Power supply	Core: 1.0V - 1.8V IO: 3.3V
Frequency	RISC: Max. 200MHz GP: Max. 200MHz RE: Max. 50MHz
Transistors	1.6M logic, 29KB SRAM
Processing speed	RISC: 200MIPS VS: 141Mvertices/s RE: 50Mpixels/s
Power consumption	52.4mW @ 60fps 153mW @ full speed

Three power domains with independent dynamic voltage-frequency scaling

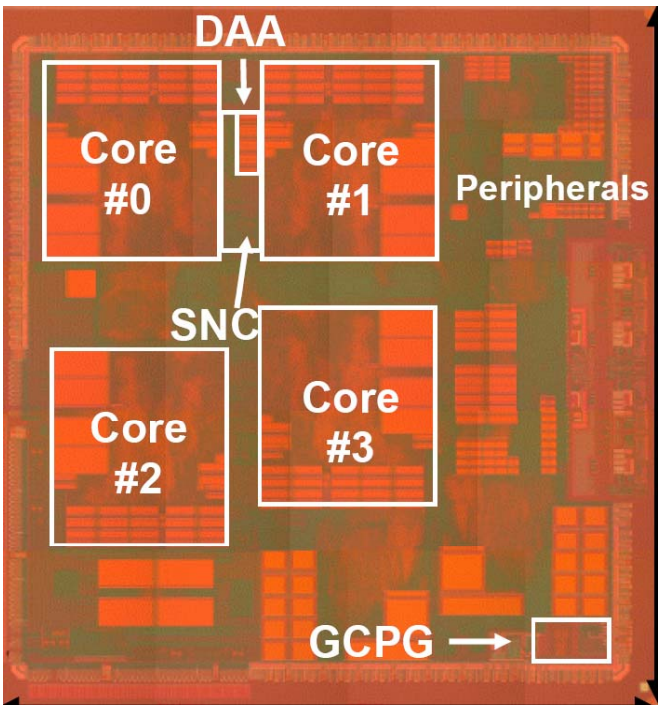
[Nam, ISSCC 2007]

Multiple-Domain Power Management

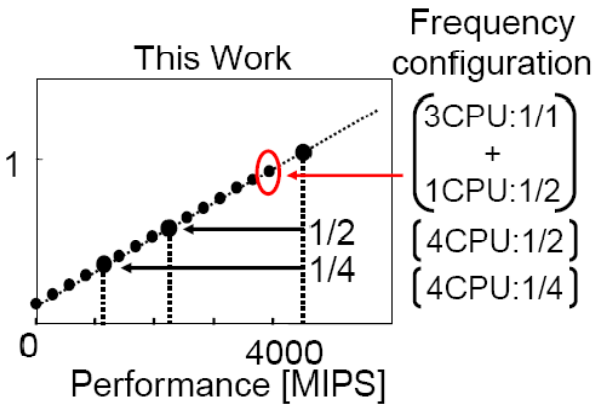


[Nam, ISSCC 2007]

Individually Managed Core Clock Frequency



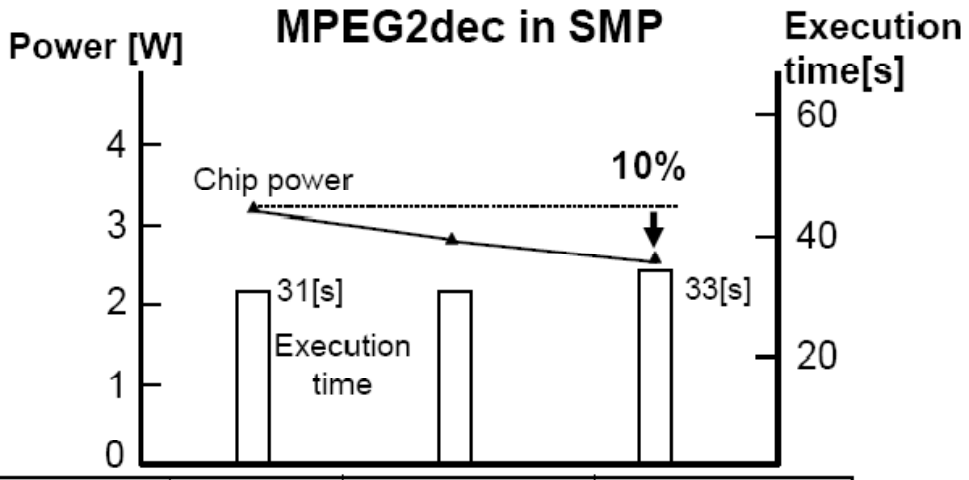
Renesas 90nm



Individual core-level frequency control enables multiple performance points

[Yoshida, ISSCC 2007]

Power Reduction Benefits



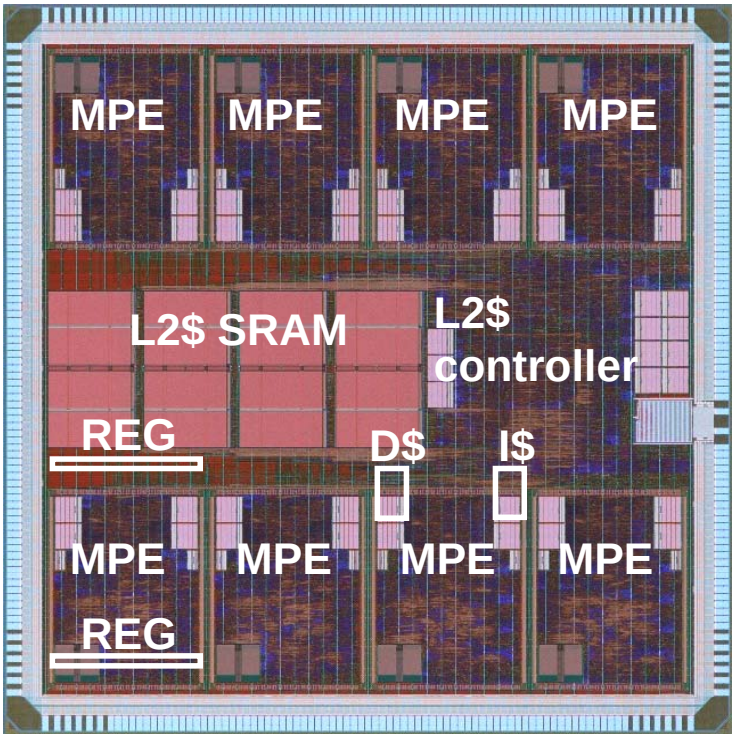
Core #0 - #1	600MHz	600MHz	600MHz
Core #2 - #3	600MHz	600MHz	300MHz
Light sleep	w/o	with light sleep	with light sleep

Individual core clock distribution has modest power reduction benefits

Much higher benefit achievable with separate voltage domains

[Yoshida, ISSCC 2007]

Toshiba's 8-Core Media Processor



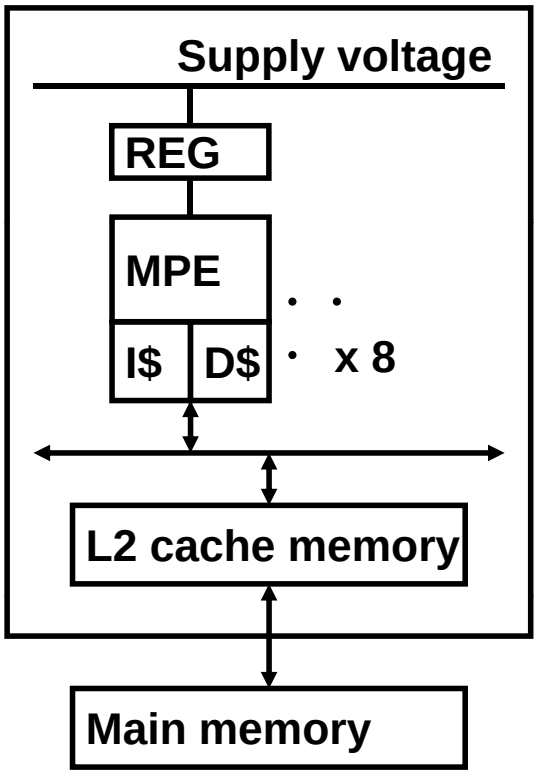
65nm CMOS
8-layer-metal

Supply voltage
2.5V (I/O)
1.2V (core)
1.2V / 0.95V / 0V
(REG output)

Frequency
333MHz
(MPE / L2$ logic)
166MHz
(L2$ SRAM / bus I/F)

[Ueda, A-SSCC 2008]

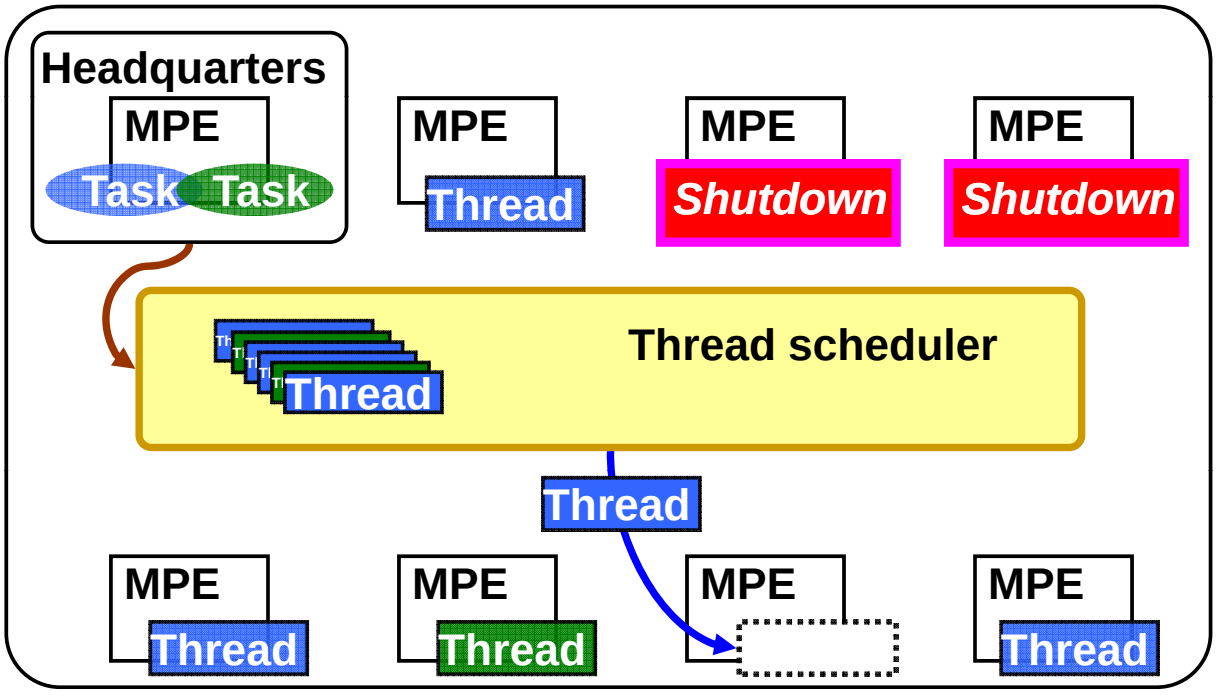
Processor Architecture



- MPE (Media Processing Engine)
  - 5-stage 32b RISC with 64b SIMD 2-way VLIW co-processor
- L1 cache (8KB I$ + 8KB D$)
- L2 cache
- Voltage regulator (REG)
  - Control of supplied voltage
  - 1.2V / 0.95V / 0V

[Ueda, A-SSCC 2008]

Software Power Management



Shutdown MPEs at low work load

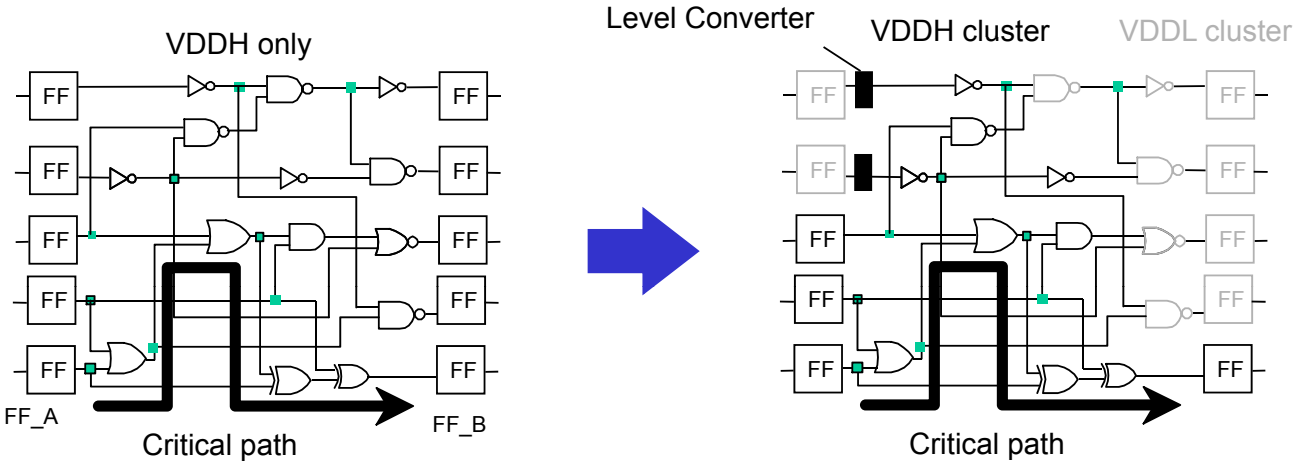
[Ueda, A-SSCC 2008]

Multi-Domain Processors Design Overview

- Voltage / frequency scaling basics
- Multi-domain server processors
- Cell phone processors
- Media processors
- Dual voltage supply at the cell level
- Future directions
- Summary

Cell-Level Dual-VDD Approach

- Use reduced voltage *VDDL* in non-critical paths
- Apply original voltage *VDDH* to timing critical paths

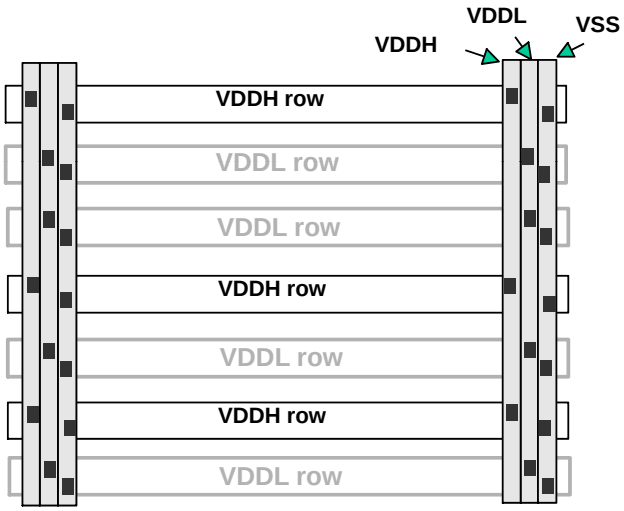


Challenge: minimize number of level converters by clustering

[Usami, DAC 1998]

Cell-Level Dual-VDD (cont)

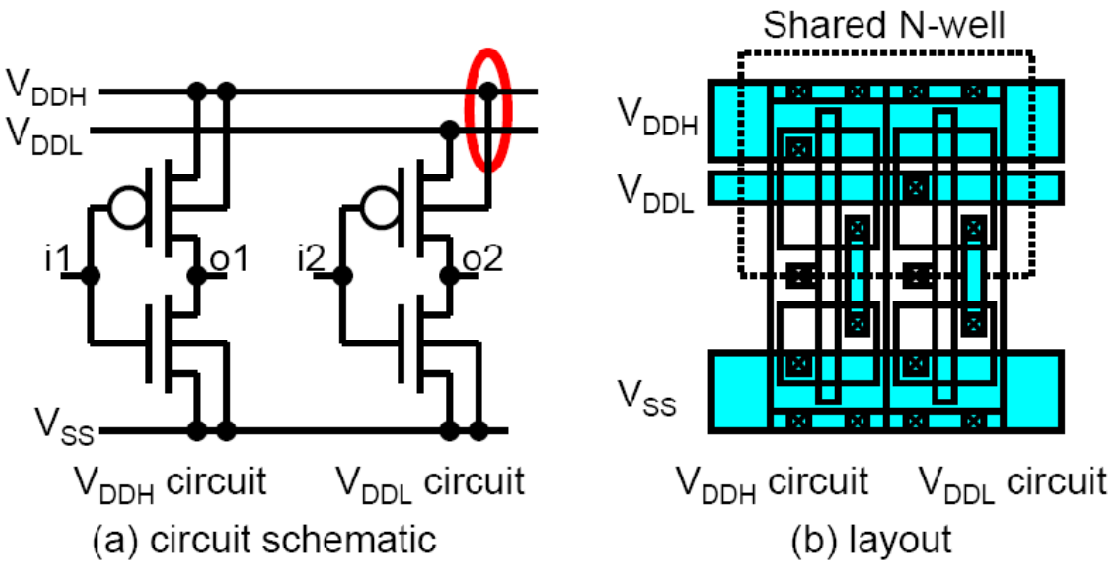
Row-by-Row layout architecture with Dual-V_{DD}



- P&R tool determines which rows should be *VDDL*
- Clock tree synthesis using *VDDL* clock buffers
- 25% power reduction on MPEG4 video codec core

[Usami, DAC 1998]

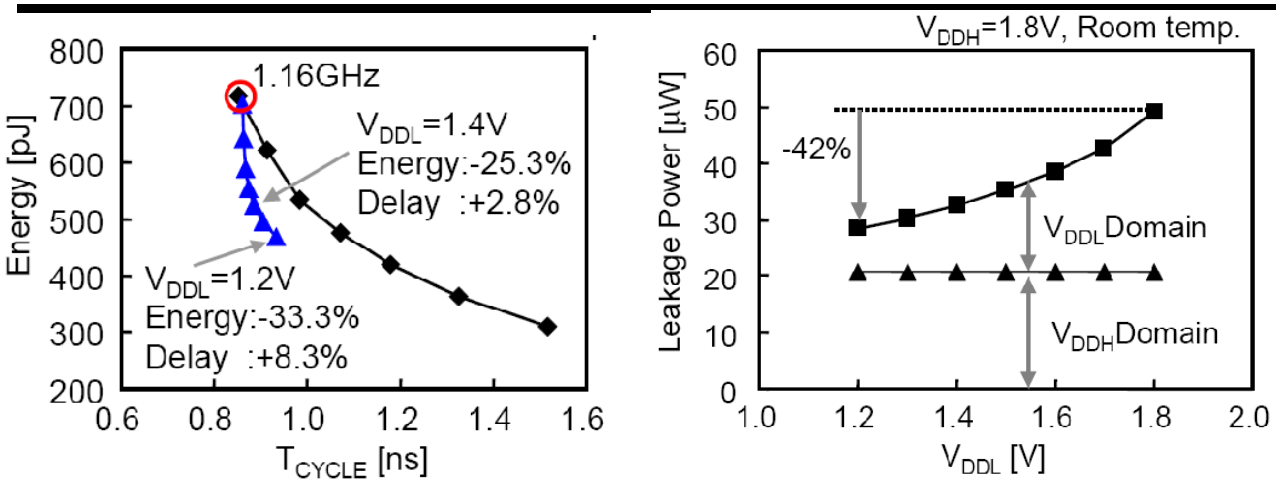
Shared Well Dual Supply



- Both circuits can be placed in the same N-well
- Cell layout becomes complex
- An intrinsic negative back-biasing of PMOS degrades speed

[Shimazaki, ISSCC 2003]

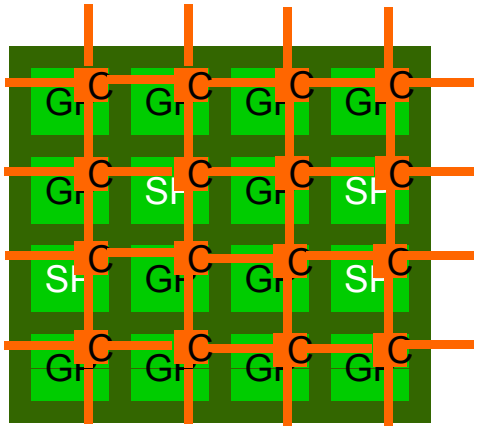
Dual Supply ALU Test Chip Results



- 1.16GHz 64bit ALU in GP 0.18µm bulk CMOS
- 25% energy saving with 2.8% delay increase
- 42% leakage reduction
- Watch for the voltage level converters overhead

[Shimazaki, ISSCC 2003]

Future Directions for Multi-Core Platforms



General Purpose Cores

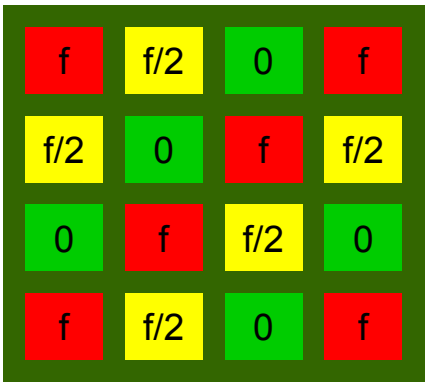
Special Purpose HW

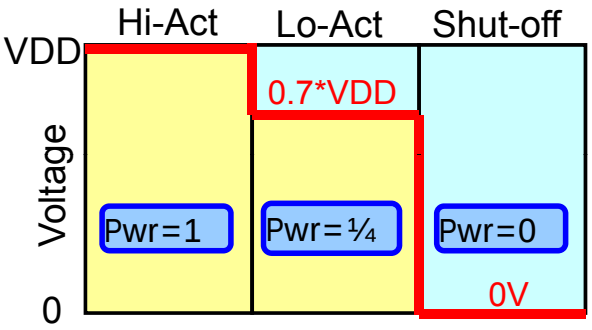
Interconnect fabric

Heterogeneous Multi-Core Platform—SoC

[S. Borkar, DAC 2007]

Fine Grain Power Management





Cores with critical tasks
Freq = f, at Vdd
TPT = 1, Power = 1

Non-critical cores
Freq = f/2, at 0.7xVdd
TPT = 0.5, Power = 0.25

Cores shut down
TPT = 0, Power = 0

[Adapted from S. Borkar, DAC 2007]

Summary

- Multiple voltage and clock domains are widely used in modern processor design to manage power and process scaling issues
- Optimize voltage/frequency operating point for each block to minimize power consumption
- The need to shut-off unused logic is driving a finer granularity clock and power gating
- Core and cache recovery enables multiple product options where disabled cores and cache slices are clock and power gated
- Increased use of Globally Asynchronous Locally Synchronous (GALS) clocking for large SoC designs
- Managing all these voltage and frequency domains requires increased software complexity

Outline

- Part I: Multi-Domain Processors Design Overview (2:00-2:45PM)
  - Multi-domain server, cell phone, and media processors
  - Power management techniques
- Part II: Router Design and Synchronization Issues (2:45-3:30PM)
  - Asynchronous router design
  - Quality of Service and virtual channels in QNoC
- Part III: Control and Power Management in Presence of Workload Variations (4:00-4:45PM)
  - VFI partitioning and voltage assignment
  - Workload modeling and dynamic control of multi-VFI designs
- Part IV: DVFS in Presence of Process Variations (4:45-5:30PM)
  - Impact of process variations on DVFS controller performance
  - Technology-driven limits on DVFS controllability

ISCA-2010 Tutorial #2

The Asynchronous NOC

Ran Ginosar

Technion

ran@ee.technion.ac.il

TECHNION