VLSI, very large scale integration
IC, integrated circuit 集成电路
HPC，high performance computer
IDC, Internet data center
UPS, uninterruptable power supply
PDU, power distribution units
TCO, total cost of ownership
CapEx，capital expenditure 资本支出
OpEx, operational expenditure 运营开销

ISA， instruction set architecture
CISC, complex instruction set computer
RISC, reduced instruction set computer
PC, program count
GPR, general-purpose register 通用寄存器
SPR, special-purpose register
ILP, instruction-level parallelism
uOps, micro-ops

RAW, read after write
WAR, write after read
WAW, write after write
FU, function unit
IF, instruction fetch
ID, instruction decode
RO, read operands
EX, execution
WR, write result
RS, reservation station
CDB, common data bus

IPC, instruction per cycle
VLIW, very long instruction word
DB, dispatch buffer
BHT, branch history tables
BTB, branch target buffer
BIA， branch instruction address
BTA, branch target address
OoO, out-of-order execution
ROB, reorder buffer
EPIC, explicitly parallel instruction computing

PTE， page table entry
MMU, memory management unit
VPN, virtual page number
PPN, physical page number
3C model, compulsory/capacity/conflict miss
DIMM, dual in-line memory module
MC, memory controller
RAS, row access table
CAS, column access table
SDRAM, synchronous DRAMS
DDR, double data rate SDRAM
IPM, instruction per miss
MLP, memory-level parallelism

ATA，Advanced Technology Attachment
高级技术附件
SATA，Serial ATA，串行高级技术附件
SCSI，Small Computer System Interface
SAS，Serial Attached SCSI
FC，Fiber Channel，光纤信道协议
JBOD，Just a bunch of disks
RAID，redundant array of inexpensive drives
ECC，Error correcting coding，错误校验码
DAS，Direct Access Storage，直接存取存储器
NAS，Network Attached Storage 网络附属存储
SAN，Storage Area Network，存储区域网络
NAND Flash cell
SLC, single-level cell
MLC, multi-level cell
SSD，Solid-State Drive
WA，Write amplification

CPI，Clock cycle Per Instruction
PCA，Principal Component Analysis 主成分分析
NTT， normalized turnaround time
ANTT, average NTT

SISD，Single Instruction Single Data
SIMD，Single Instruction Multiple Data
MISD，Multiple Instruction Single Data
MIMD，Multiple Instruction Multiple Data
UMA架构，Centralized Shared-Memory
SMP，symmetric multiprocessors，对称多处理器
NUMA架构，DSM，Distributed Shared-Memory
MSI, MESI

TLP, thread-level parallelism
CMP，Chip Multi-Processor，片上多处理器
SMT，simultaneous multithreading，超线程
LLC，Last Level Cache
MIC，Intel Many Integrated Core 架构
SCC，Intel Single-Chip Cloud Computer

DLP，Data-level parallelism，数据级并行
MVL，maximum vector length
VLR，Vector Length Register
GPU，Graphics processing unit
GPGPU，general-purpose computation on GPU
GPU上的通用计算
SPMD，Single Program Multiple Data
SIMT，Single Instruction Multiple Thread
CUDA, Stands for Compute Unified Device
Architecture
TPC，texture/processor clusters
SM，streaming multiprocessors，流多处理器
SP，streaming processor，流处理器
GPC，Graphics Processor Clusters，
图形处理器集群
RF，Register File

Flit, flow control unit
MIN, multistage interconnection network

ICT, information and communication
technology
WSC, warehouse-scale machine
PUE, power usage effectiveness
WUE, water usage effectiveness
CUE, carbon usage effectiveness
ATS, automatic transfer switch
STS, static transfer switch equipment
UPS, uninterruptable power supply
PDU, power distribution unit
PSU, power supply unit
CRAC, computer room air conditioning
COP系数, coefficient of performance
TOR, top of rack
EOR, end of rack
MDA, main distribution area
HAD, horizontal distribution area
EDA, equipment distribution area
SLA, service level agreement
DFS,
DVFS, dynamic voltage and frequency scale
MDC, modular data center

ACPI, advanced configuration and power
interface
TDP, thermal design power
D2D, die to die
C2C, core to core
MPP, maximal power point
PTP, performance time product
TPR, throughput power ratio
ROI, return on investment

MTTF, mean time to failure
MTBF, mean time between failure
MTTR, mean time to repair
ACE. Architecturally correct execution
AVF, architectural vulnerability factor
PSU,

Summary 1
- What is Computer Architecture
- History of IC(integrated circuit)
- Transistor basics
- Feature length
- HPC vs IDC
- Scale up/out
- Energy/power issues
- The trend of Computer Architecture research

Summary 2
- Architecture vs microarchitecture
- Evolution of instruction set
- CISC(IA32) vs RISC(MIPS)
- Machine interfaces：ISA
- User/System ISA
- MIPS instruction field
- Single-cycle MIPS
- Ideal pipeline
- Stage quantization
- Pipeline slot

Summary 3
- Pipeline stall and bubble
- Dependency and hazards
- RAW, WAR, WAW
- Forwarding and pipeline interlock
- Functional units
- Dynamic scheduling and OoO
- Scoreboarding
- Tomasulo's Algorithm

Summary 4
- Superscalar Pipeline
  - Limitations of scalar processor
  - Basic feature of superscalar pipeline
  - Multi-Issue Processor
  - Dispatch Buffer and Completion Buffer
  - Classification of ILP Machines
  - Rationale of branch prediction
  - 2-bit prediction
- Speculation
  - Precise exception: in-order completion
  - Reorder buffer (ROB)
  - Tomasulo's Algorithm with ROB
- VLIW and EPIC
  - CISC vs RISC vs VLIM
  - Loop unrolling
  - The concept of EPIC

Summary 5
- Memory hierarchy, uncore and off-chip
- Cache line, block, address
- Virtual memory, page table, PTE, TLB
- Locality principle, Inclusive and exclusive relationship
- Miss caching, victim caching, prefetching
- Cache write policies, write buffer/cache
- rank, bank, array, channel, MC, parallelism in DRAM
- 1T1C DRAM cell, data access, DRAM refresh
- DRAM access cost
- synchronous/asynchronous design
- Memory design challenges, memory wall
- MLP

Summary 6
- Disk concept; platter/track/sector
- Design good drive Interfaces
- Parallel/Serial ATA; Parallel/Serial SCSI
- RAID Organization
- DAS, NAS, SAN
- Flash memory cell, SLC/MLC
- SSD advantages, hybrid storage

Summary 7
- Amdahl's Law
- Calculating CPI （公式）
- Analyzing memory access time （公式）
- Little's Law
- Estimating server power
- Trace-/Execution- driven simulation
- Simulation acceleration
- Concepts of workload characterization
- Multi-programmed workload

Summary 8
- SIMD, MIMD, TLP
- Multiprocessors, UMA and NUMA
- Definition of cache coherency
- Cache coherency and memory consistency
- Basic facts of the snooping protocol
- A simple write-through invalidation protocol
- 3-state MSI protocol

Summary 9
- Thread, Multithreading, SMT
- CMP and multicore
- Benefits of multicore
- Multicore system architecture
- Heterogeneous multicore system
- Heterogeneous-ISA CMP
- Multicore and manycore
- Design challenge

Summary 10
- Throughput computing and data-level parallelism
- Vector processor and vector instruction
- VMIPS, vector registers, DAXPY, execution latency
- SIMD lanes, chaining, vector length register
- GPGPU, SIMT, CUDA programming model
- TPC, SM, SP, warp and warp scheduling
- Branch divergence
- GPU register file

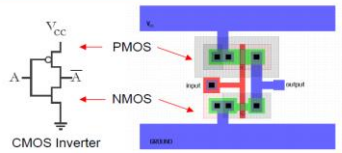
Summary 11
- basic concepts, link/channel/buffer
- switch degree, average distance
- non-blocking network, direct/indirect network
- network performance, latency estimation
- network switch and switch strategy
- bus and crossbar
- array ring mesh torus tree butterfly hypercube

Summary 12
- what is a data center
- major metrics of data center design
- data center infrastructure
- the long tail concept
- data center capacity utilization
- types of power provisioning
- modular data center and cooling

Summary 13
- G-States, S-States, C-States, P-States
- TDP, Turbo Boost
- Power management can be challenging

Summary 14
- Faults, error, and failure
- MTTF, MTBF, MTTR
- Availability, reliability
- ACE, AVF
- Redundancy

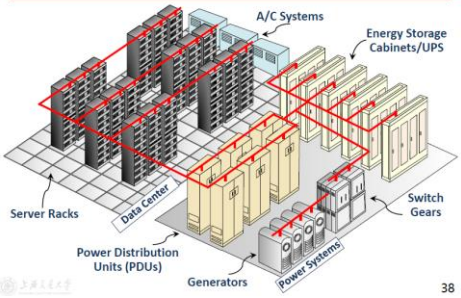
From Circuit to Layout



Active areas for thin oxide region
Polysilicon for the gate
Metal for interconnection
Contact (Via) for inter-layer connection

25

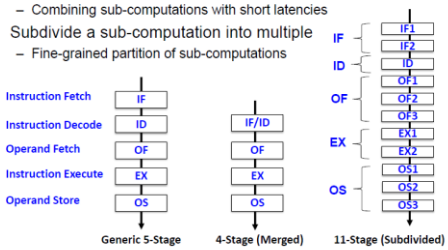
A Typical Data Center



38

Stage Quantization

- Merge multiple sub-computations into one
  - Combining sub-computations with short latencies
- Subdivide a sub-computation into multiple
  - Fine-grained partition of sub-computations



Q: the more (stages), the better ?

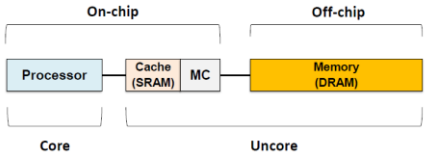
44

Architecture Comparison

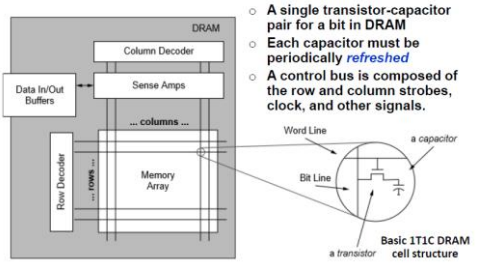
Category	CISC	RISC	VLIW
Inst Size	Varies	Fixed (typically 32bits	Fixed
Inst Format	Field placement varies	Regular, consistent placement of fields	
Inst Semantics	Complex; possibly many dependent ops per instr	Almost always one simple operation	Multiple, independent simple operations
Registers	Few, sometimes special	Many, general purpose	
Memory	Reg-Mem architecture	Load/Store architecture	
Hardware	Exploit microcode	Not microcoded implementation	
Example:			

Basic Concepts

4



1T1C DRAM Cell



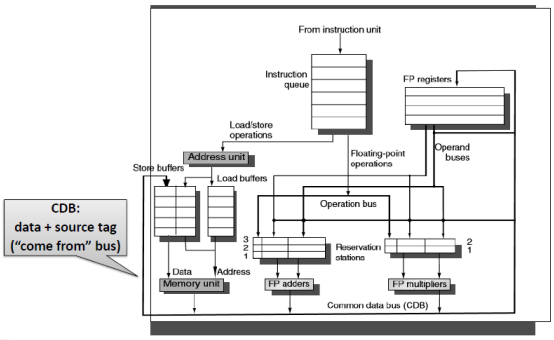
- A single transistor-capacitor pair for a bit in DRAM
- Each capacitor must be periodically *refreshed*
- A control bus is composed of the row and column strobes, clock, and other signals.

Simulation Approach Comparison

	Function Simulation	Trace-Driven Simulation	Execution-Driven Simulation
Development Time	Excellent	Poor	Very Poor
Evaluation Time	Good	Poor	Very Poor
Accuracy	Excellent	Very good	Excellent
Coverage	Poor	Excellent	Excellent

- Full system simulation
  - Trace-driven simulation and execution-driven simulation

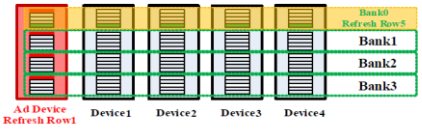
Basic Structure Implementing Tomasulo's Alg.



0

Refresh Mechanism

- A Row is the smallest refresh unit in bank.
- Typically, the retention time of data in DRAM cell is 64ms if the ambient temperature is less than 85 degree Celsius.
- Refresh operation can implement at rank level (per-rank refresh) or at bank level (per-bank refresh)

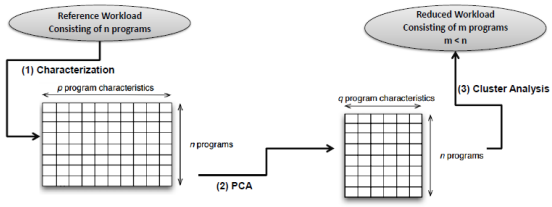


In per-rank refresh, all banks are locked and inaccessible during refresh cycle.

42

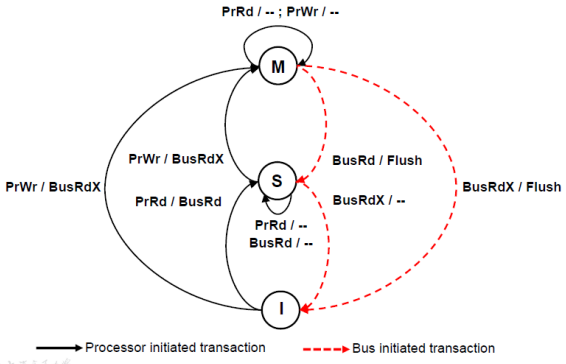
Workload Design (Cont'd)

- Determine a reduced but representative workload

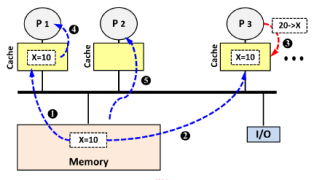


Schematic overview of the PCA-based workload reduction method

A 3-state (MSI) Write-Back Invalidation Protocol



A 3-state (MSI) Write-Back Invalidation Protocol



Processor Action	State in P1	State in P2	State in P3	Bus Action	Data Supplied By
1. P1 reads x	S	-	-	BusRd	Memory
2. P3 reads x	S	-	S	BusRd	Memory
3. P3 writes x	I	-	M	BusRdx	Memory
4. P1 reads x	S	-	S	BusRd	P3 Cache
5. P2 reads x	S	S	S	BusRd	Memory

40

Calculating CPI

$$CPI = \sum \frac{IC_i \times CPI_i}{Instruction\ count}$$

$$CPU\ time = (\sum IC_i \times CPI_i) \times Clock\ cycle\ time$$

Example:

Suppose we have made the following measures:

Frequency of FP operations = 25%
Average CPI of FP operations = 4.0
Average CPI of other instructions = 1.33

Then:

CPI original = (4×25%)+(1.33×75%) = 2.0

6

Memory Performance Analysis

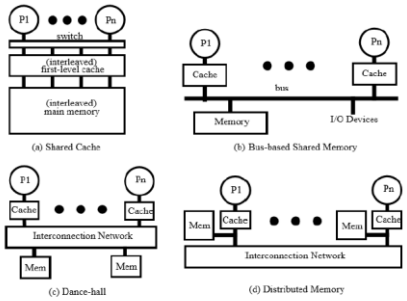
$$\frac{Misses}{Instruction} = Miss\ rate \times \frac{Memory\ accesses}{Instruction}$$

$$Average\ memory\ access\ time = Hit\ time + Miss\ rate \times Miss\ Penalty$$

$$Average\ memory\ access\ time = Hit\ time_{L1} + Miss\ rate_{L1} \times Miss\ Penalty_{L1}$$

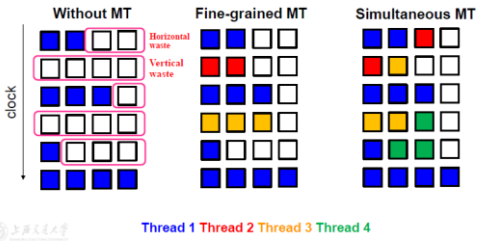
$$Miss\ Penalty_{L1} = Hit\ time_{L2} + Miss\ rate_{L2} \times Miss\ Penalty_{L2}$$

Common Memory Hierarchies in Multiprocessors



Impacts of SMT on Utilization

- Multithreaded processor improves hardware utilization in different dimensions



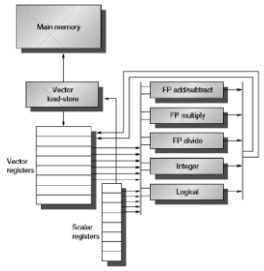
9

Classification of Heterogeneous Multicore

	Same Cores	Different Cores
Different ISA	??	Heterogeneous-ISA Chip Multiprocessors
Same ISA	Homogeneous Multi-/Many- Core	Cores of Different Capabilities

Case Study: VMIPS

- Vector registers
  - 64-element register
  - Register file has 16 read ports and 8 write ports
- Vector functional units
  - 5 fully pipelined FUs
  - Hazards detection
- Vector load-store unit
  - Loads/stores a vector
  - 1 word per clock cycle



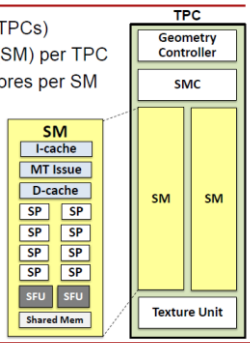
12

Case Study: Tesla GPU Architecture

- 8 texture/processor clusters (TPCs)
- 2 streaming multiprocessors (SM) per TPC
- 8 streaming processor (SP) cores per SM

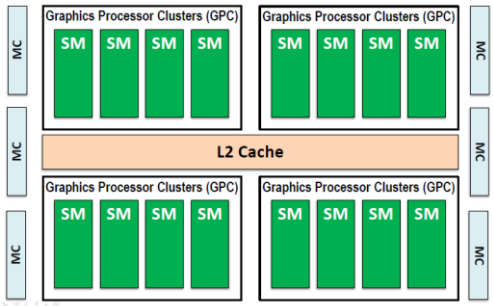
- SM is a unified graphics and computing engine
  - SM executes thread blocks

- Streaming processor (SP)
  - Scalar ALU for a single CUDA thread
  - SP core is a SIMD lane



Case Study: Fermi GF100 Architectural Overview

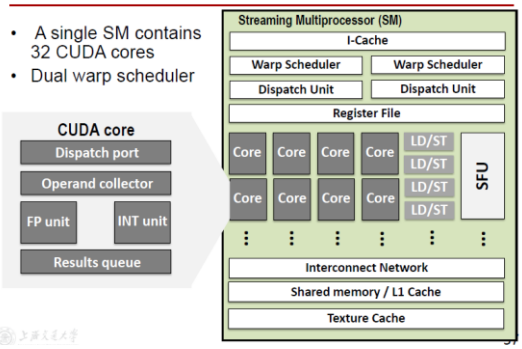
- Each GPC contains four streaming multiprocessors (SMs)



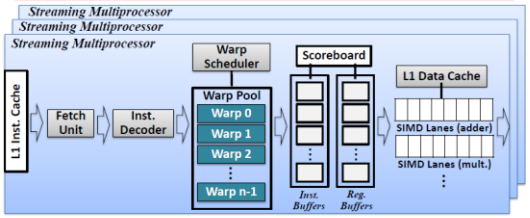
36

Case Study: Fermi GF100 Architectural Overview

- A single SM contains 32 CUDA cores
- Dual warp scheduler



Warp Scheduling (Cont'd)



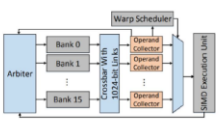
- Potential factors that can delay a warp's execution
  - Scheduling policies
  - Instruction/Data cache miss
  - Structural/Control/Data hazard
  - Synchronization primitives

41

Resource Limits

- The maximum parallelism in GPUs is often limited by the register file capacity
  - Applications with high TLP triggers more active warps

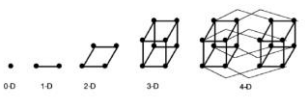
- Register file is a large SRAM structure
  - Fastest memory block available to the processor
  - Store intermediate results from units such as ALU
  - Power hungry structure



45

Hypercube (Optional)

- Also known as binary d-cubes $N = 2^d$
- Switch degree equals dimension: $d = \log_2 N$
- Good bisection bandwidth
- **k-ary d-cube** is a d dimensional torus with k elements along each dimension: $N = k^d$
  - Each node is addressed by a d-vector



24

Performance Evaluation

- Link width: w
- Unit interval: τ
- Signaling rate: $f = 1/\tau$
- Channel bandwidth: $b = w \cdot f$
- Total bandwidth of all the channels (or links)
  - the number of channels times the bandwidth per channel

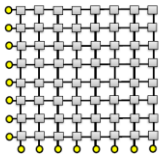
Latency (Lower Bound)

- Sending overhead: $Overhead_s$
- Receiving overhead: $Overhead_r$
- Total routing time: T_R
- Arbitration time: T_A
- Switching time: T_S
- Total time of flight of the packet $T_{TotalProp}$

$$Latency = Overhead_s + (T_{TotalProp} + T_R + T_A + T_S) + \frac{Packet\ size}{Bandwidth} + Overhead_r$$

Crossbar

- Crossbar switch
  - A type of fully-connected network
  - Every node connected to all others
  - O(N) bandwidth
  - Cost of interconnect: O(N^2)
  - Good for small number of nodes



- Crosspoint switch complexity increases quadratically with the number of ports

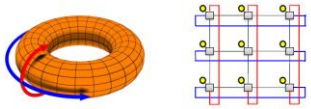
- Multistage interconnection networks reduces complexity

16

Multidimensional Topology: 2D Torus

- Reduces the diameter of a mesh network

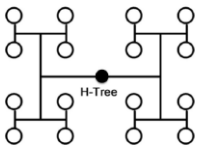
- Torus: adding end-round direct connections
  - An extension of the 2D mesh
  - A regular torus has long warp-around links
  - Slightly lower latency while higher cost



19

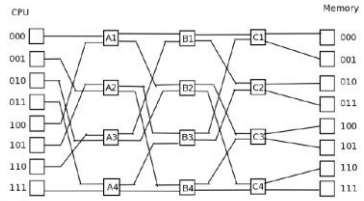
Trees (Optional)

- Trees features planar, hierarchical topology
- Employed as indirect networks with hosts as leaves
- Routing distance grows only logarithmically



Multistage Interconnection Network (Optional)

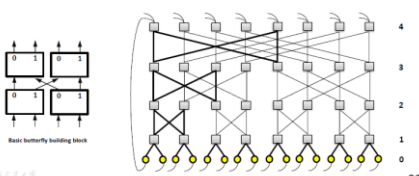
- Indirect networks with multiple layers of switches
- Omega network:
  - Log(N) number of stages and N/2 switching units



22

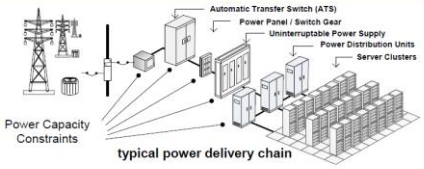
Butterfly Topology (Optional)

- Butterfly is an important logarithmic network
  - Can be viewed as a tree with multiple roots
- A d-dimensional indirect butterfly:
  - Connects $N = 2^d$ nodes (d≥2)
  - $d = \log_2 N$ levels of switches



23

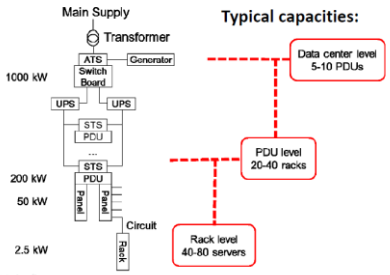
Datacenter Infrastructure – Power System



- ATS: basically a fast, mechanical switch
- STS: basically a superfast, electronic switch
- UPS: basically a battery with control interfaces
- PDU: a power converter and distribution unit

8

Datacenter Infrastructure – Power System



9

Global System States (G-states)

G-states are high-level description of the platform states

- **G0** (working)
  - The working system state
- **G1** (sleeping)
  - No computational task is performed
- **G2** (soft off)
  - Powered down but can be restarted by interrupts
- **G3** (hard off)
  - Mechanical off

Sleep States (S-states)

6

S-states are set in the BIOS and configured by the system

- **G0-S0**: normal operation
- **G1**
  - S1: processor clock is off
  - S2: processor is off
  - S3: suspend to RAM
  - S4: suspend to disk
- **G2-S5**: halt state

Processor Power States (C-states)

G0 and S0 together define a working platform state, at which a range of C-states are defined to save power

- **C0** State (normal operating state)
  - code is being executed
- **C1** State (auto halt):
  - The clock is gated, i.e., prevented from reaching the core
- **C3** State (sleep):
  - Maintains architectural state but flushes data to shared LLC
  - Shut down the clock generators
- C6 State:
  - Architectural states are saved to a dedicated SRAM
  - Core voltage reduced to zero volts

Processor Performance State (P-states)

9

P-States talk about different operational modes (freq.)

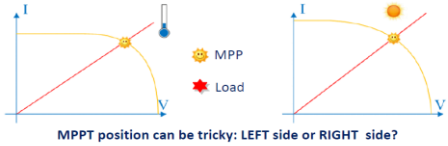
- Multiple levels of clock frequency
  - From *P0* (the highest performance) to *Pn* (the lowest performance)
- Sub states of **C0**
  - Defines dynamic voltage and frequency scaling (DVFS) steps
- Switching latencies are negligible for most purposes

Frequency	Voltage	P-State
1.6 GHz	1.484 V	P0
1.4 GHz	1.420 V	P1
1.2 GHz	1.276 V	P2
1.0 GHz	1.164 V	P3
800 MHz	1.036 V	P4
600 MHz	0.956 V	P5

Intel Pentium M at 1.6GHz

Tracking Coordination

- **Move load I-V curve to MPP**
  - Tune the power converter
  - Tune the multi-core processor



MPPT position can be tricky: LEFT side or RIGHT side?

19