Computer Architecture
计 算 机 体 系 结 构

Lecture 13. Design for Power/Energy Efficiency
第十一讲、面向功效和节能的设计

Chao Li, PhD.
李超 博士

SJTU-SE346, Spring 2019

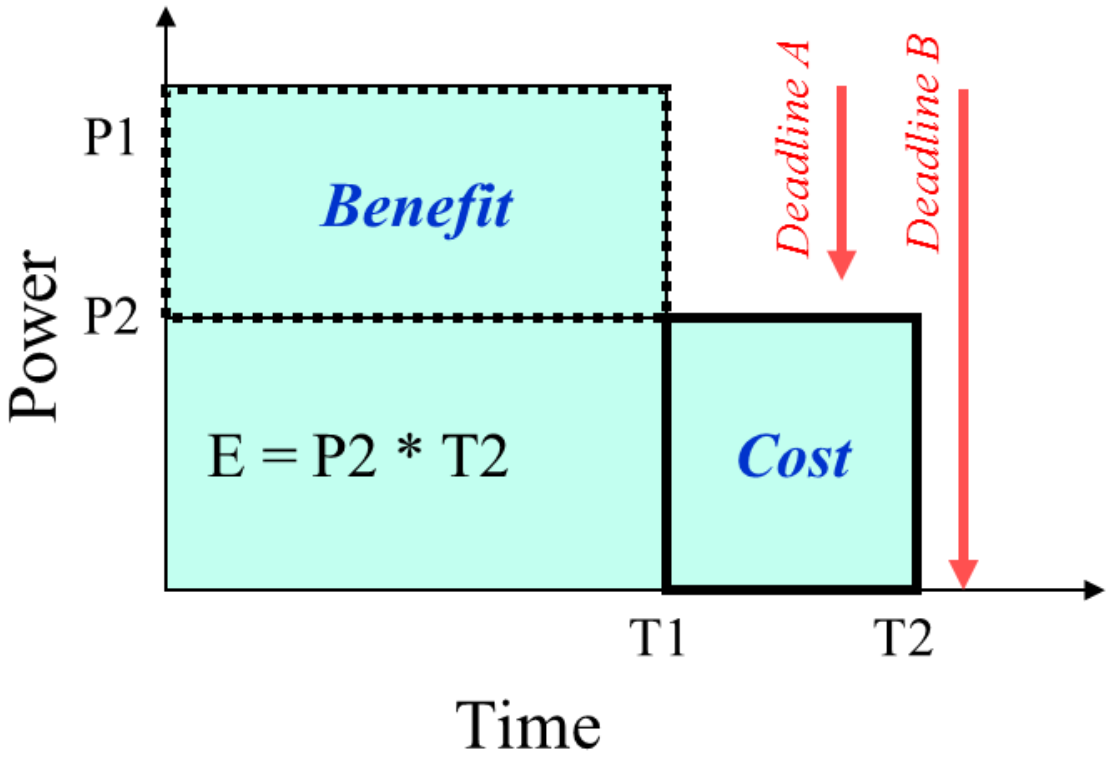
Review

- Sever-level, rack-level, cluster-level, facility-level
- Major metrics of data center design
- Data center infrastructure: Power/Cooling/ICT
- The long tail concept
- Data center capacity utilization
- Types of power provisioning
- Modular data center and cooling

Outlines

- Computer Power Management Basics
- Discussion and Case Studies

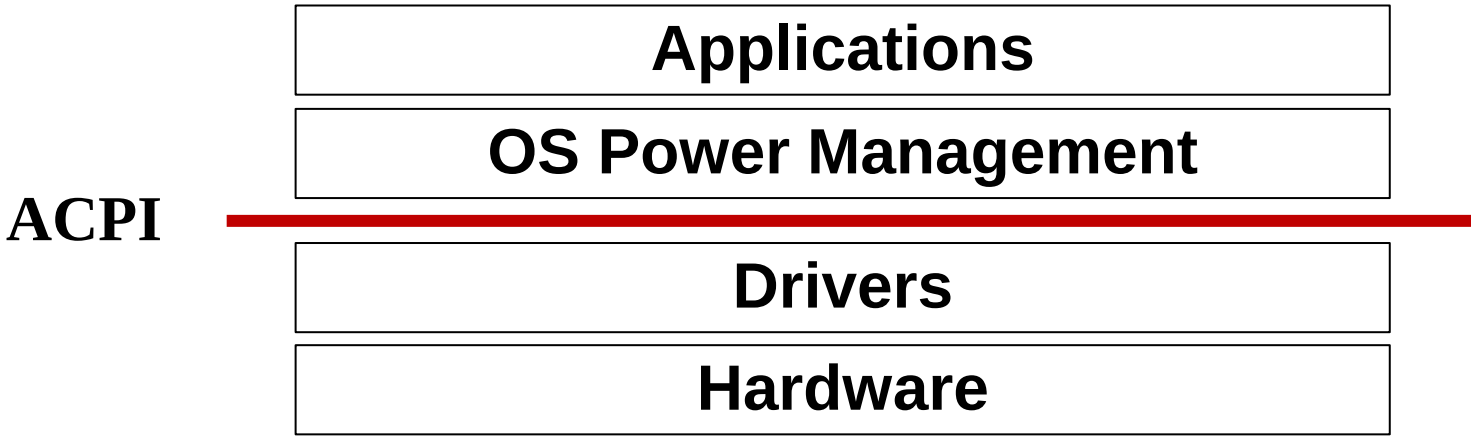
Frequency Scaling



- Benefit vs. Cost
  - power demand $\propto$ overall performance

Intel ACPI Specification

- ACPI: Advanced Configuration and Power Interface
  - An open standard
  - OS can perform power management though it



Global System States (G-states)

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

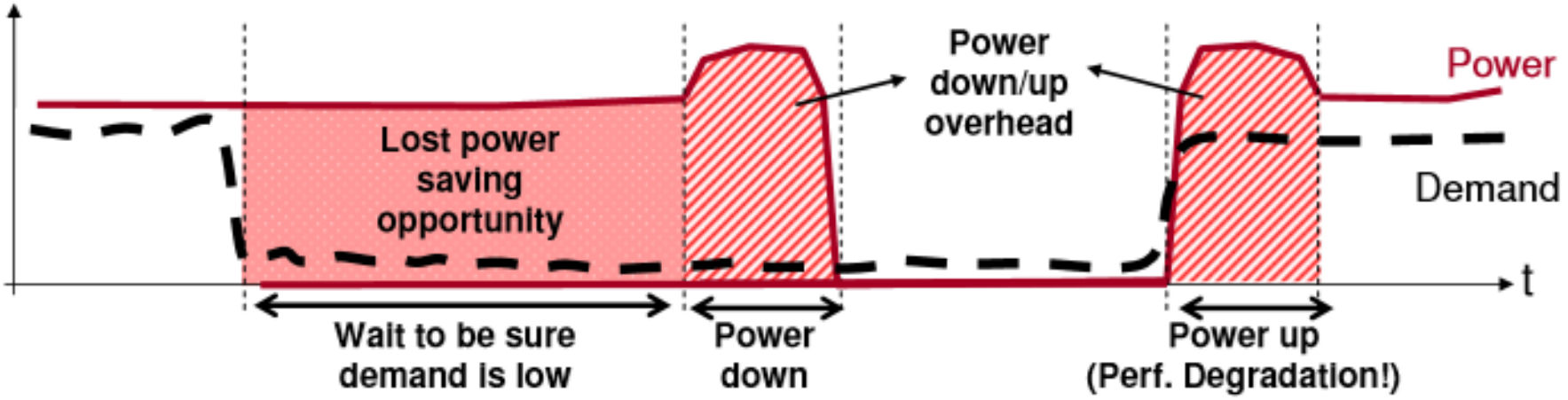
- **G0** (working)
  - The working system state
- **G1** (sleeping)
  - No computational task is performed
- **G2** (soft off)
  - Powered down but can be restarted by interrupts
- **G3** (hard off)
  - Mechanical off

Sleep States (S-states)

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

S-State Transition Latency



- Resume times from S3 can be an order of magnitude better than those with S4 or S5
- The power-off times for S3 are significantly better than for S4 and S5

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)

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

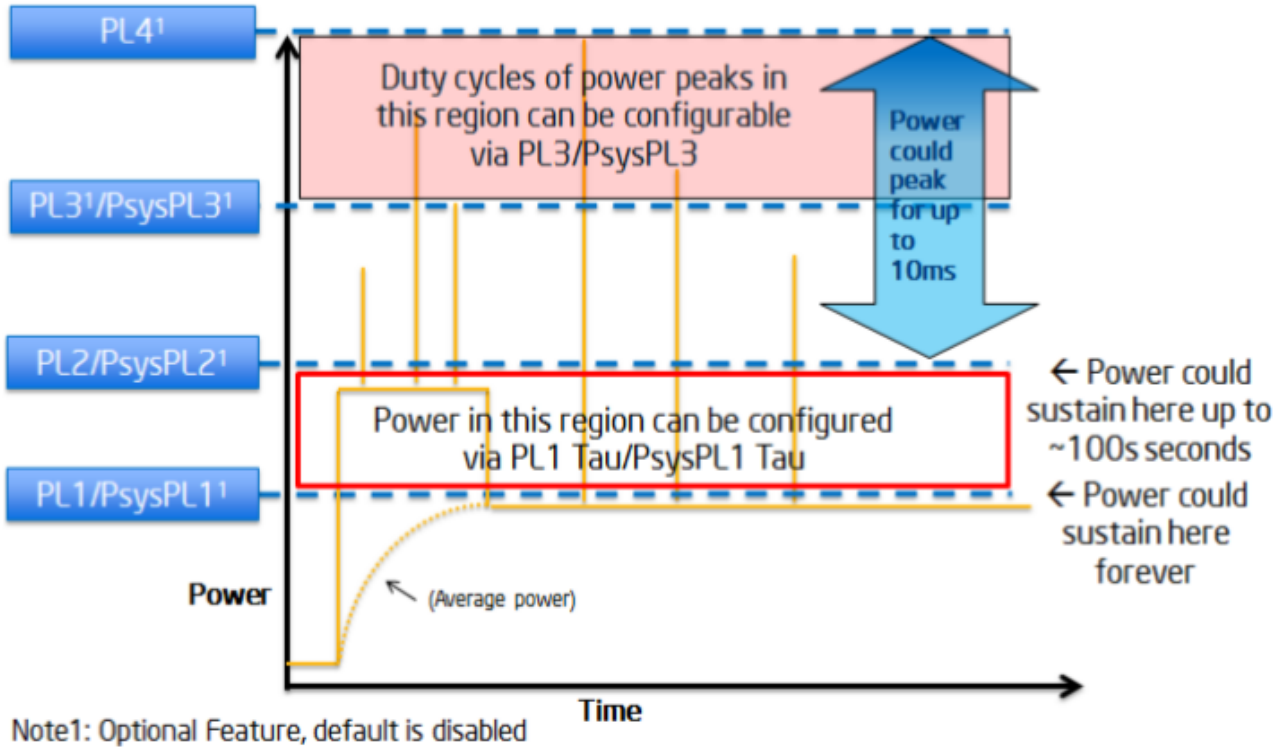
- Multiple levels of clock frequency
  - From ***P0*** (the highest performance) to ***Pn*** (the lowest performance)
- Sub states of **C0**
  - Defines dynamic voltage and frequency scaling (DVFS) steps
- Switching latencies are negligible for most purposes

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

Thermal Limitations

- Thermal design power (TDP)
  - The maximum sustained power that should be used for design of the processor thermal solution

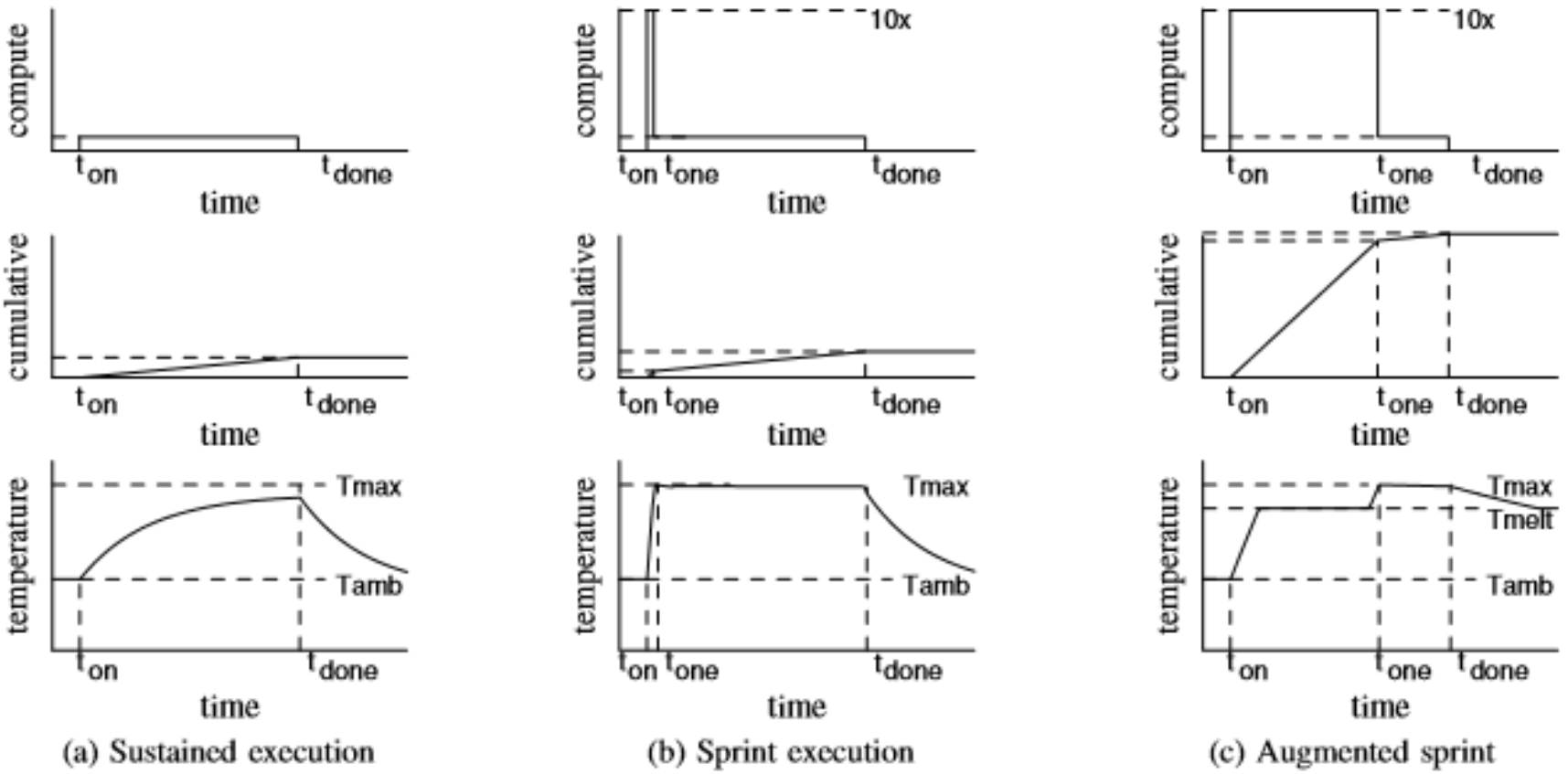


Package Power Control

Outlines

- Computer Power Management Basics
- Discussion and Case Studies

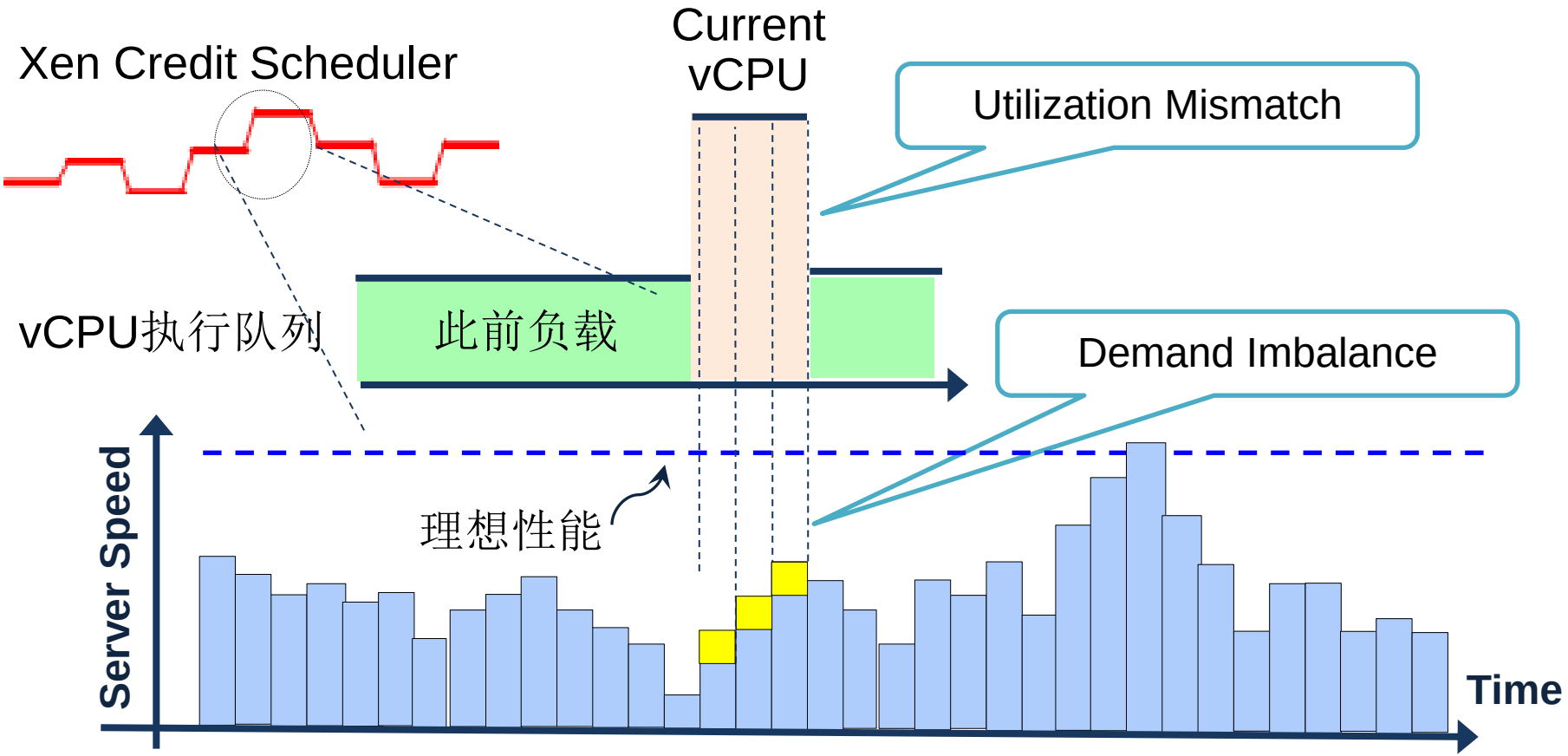
Discussion: Emerging Apps and Computational Sprinting



Cores active (top row), cumulative computation (middle row) and temperature (bottom row) over time for three execution modes: (a) sustained, (b) sprint, and (c) sprint augmented with phase change material.

Discussion: DVFS in Virtualized Cloud Environment

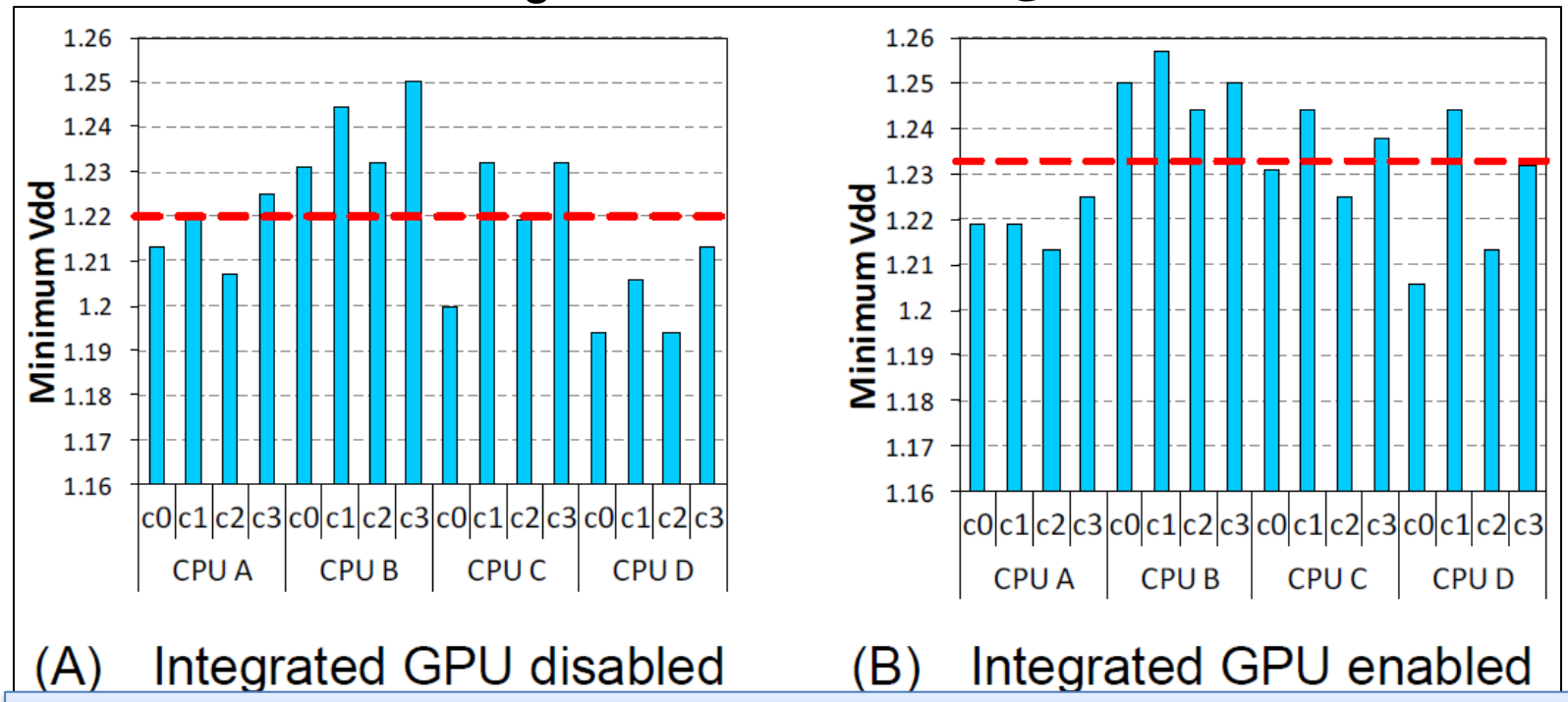
Assume the default time slice (credit) for vCPU is 30 ms and the minimum sampling interval of frequency adjustment is 10 ms



Discussion: Non-uniform Hardware Power Characteristics

Core to core (C2C) variation has been identified and the maximum difference in core frequencies is estimated to be 20%

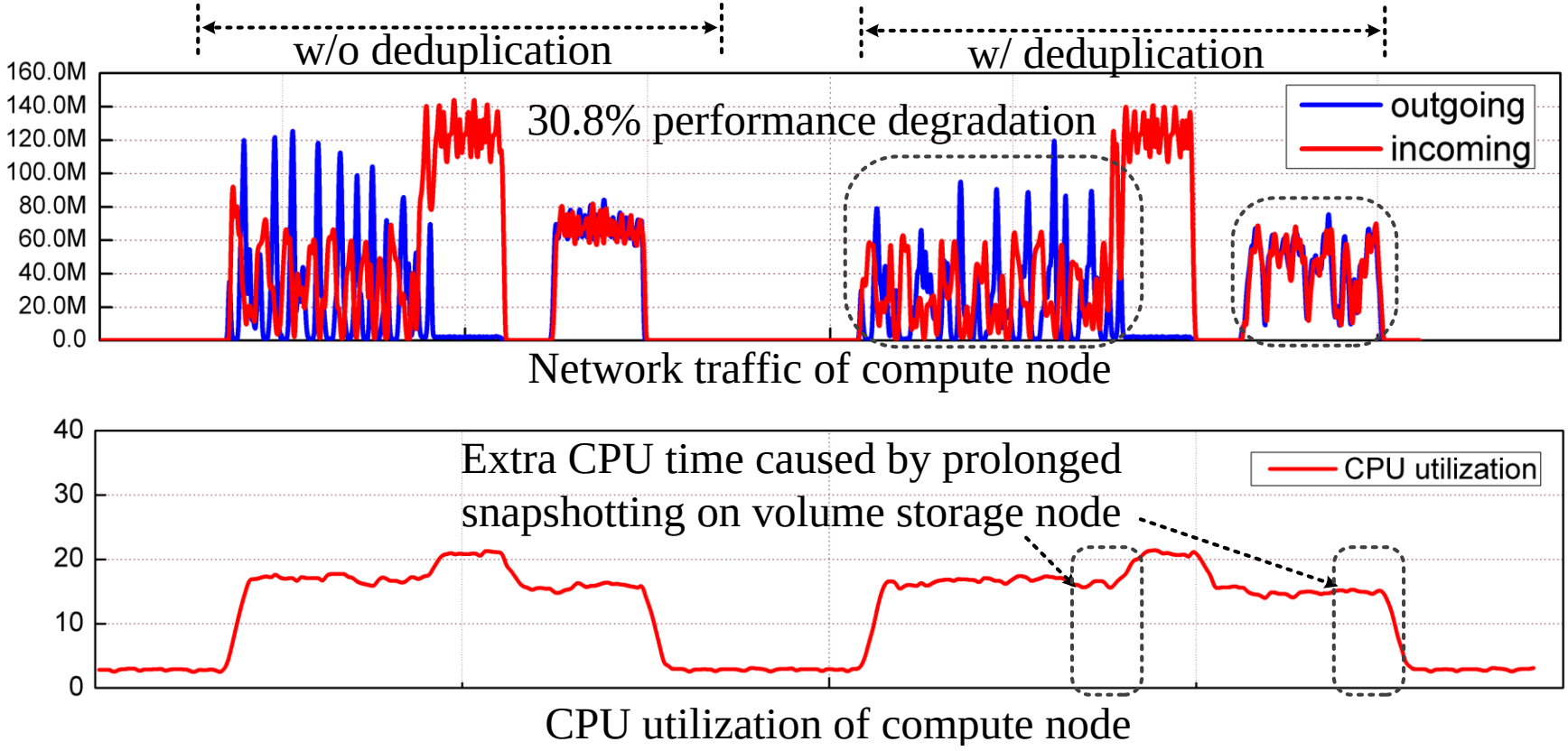
Min Voltage of AMD A10-5800K@3.8GHz



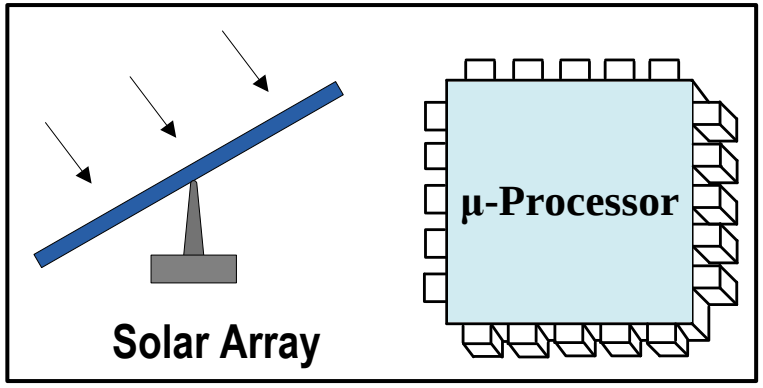
同样的处理器节点，其最低可承受供电电压不同，在**1.19V**至**1.25V**之间。

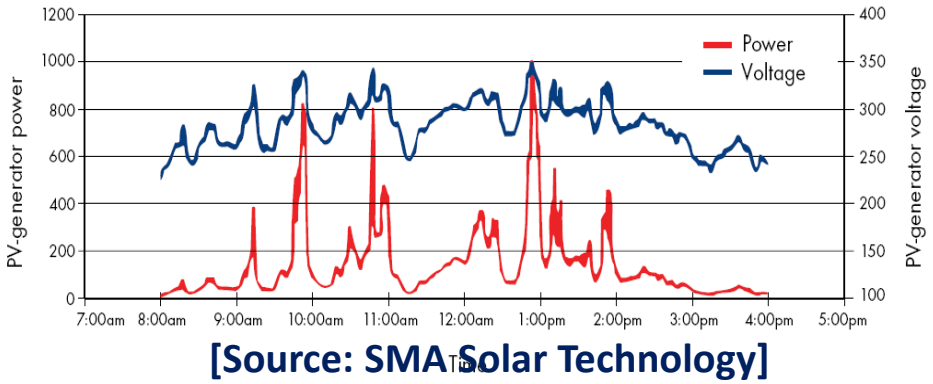
Discussion: Identifying the Power Management Bottleneck

In highly complex computing environment, a background task can become the efficiency bottleneck if other jobs depend on it



Case Study： SolarCore





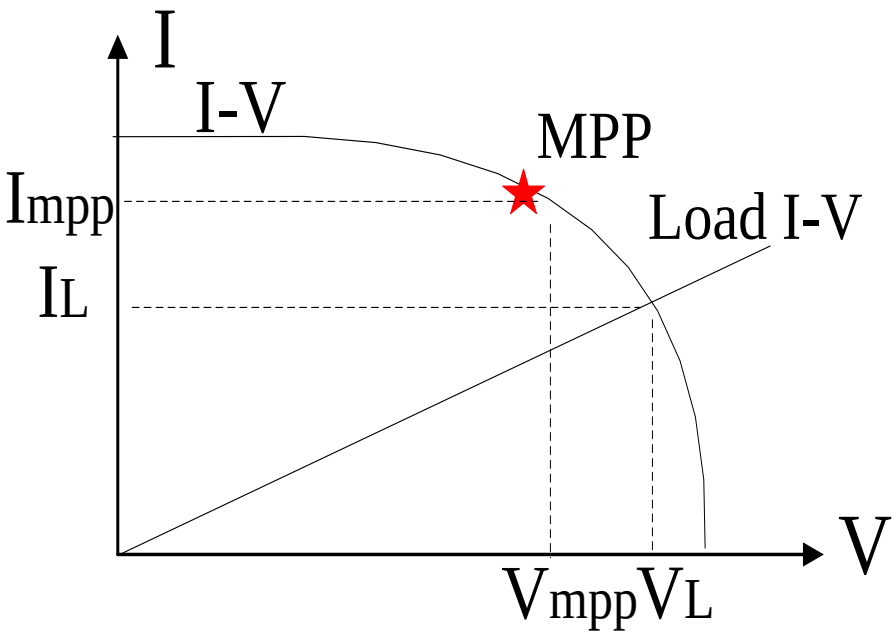
[Source: SMA Solar Technology]

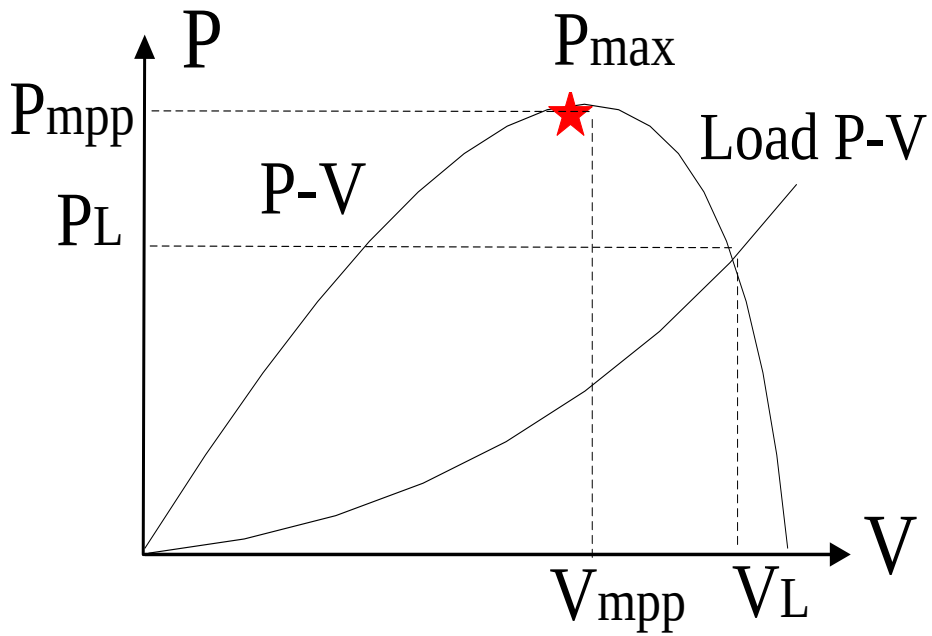
Chao Li, Wangyuan Zhang, Chang-Burm Cho, and Tao Li. “SolarCore: Solar Energy Driven Multi-core Architecture Power Management”. Proc. the 17th IEEE Int. Symp. on High-Performance Computer Architecture (HPCA), Feb. 2011. (Best Paper Award)

- Rethinking SolarCore’s Power Management Strategy
  - What we can learn?
  - What is the key limitations?

Unique Solar Power Behavior

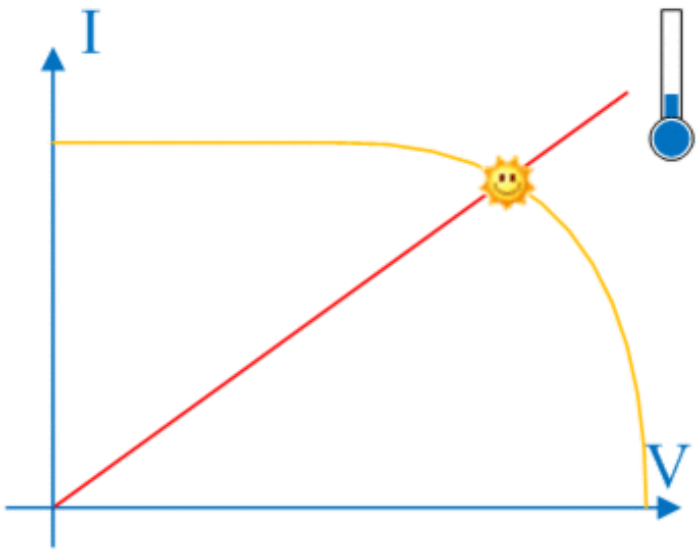
- **Variable, non-linear power output**
- **Maximal power point (MPP)**
  - A special operation point that delivers maximum electrical power

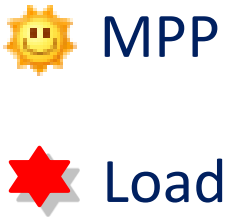


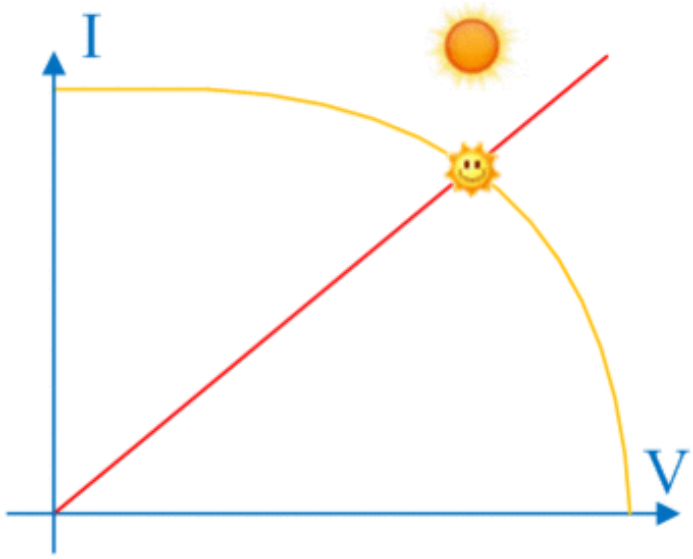


Tracking Coordination

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







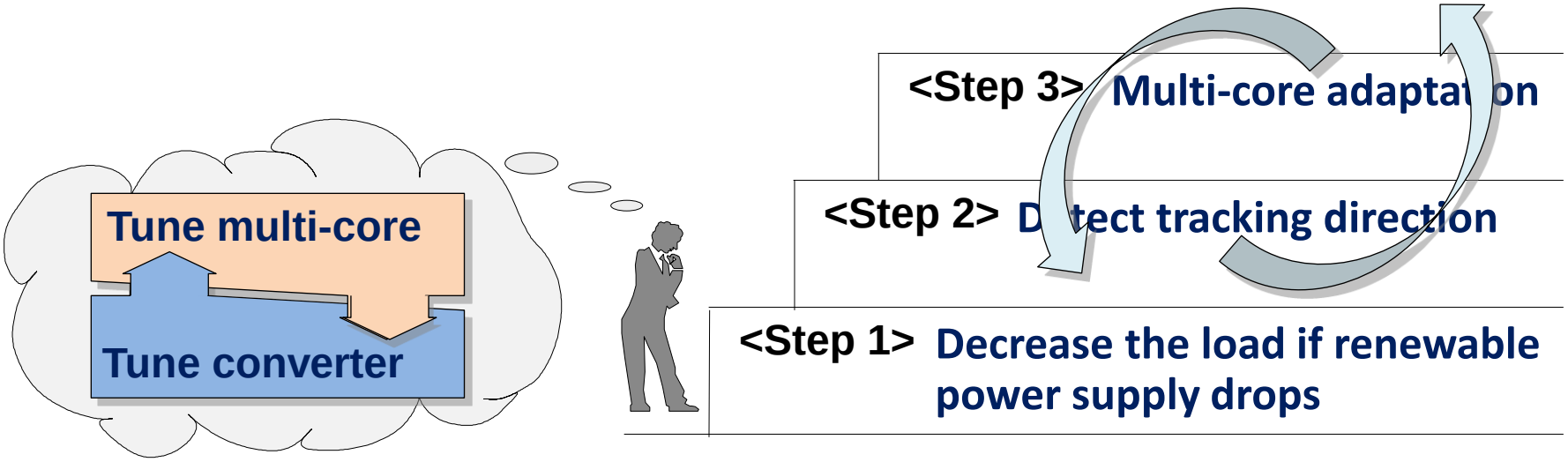
MPPT position can be tricky: LEFT side or RIGHT side?

Multi-core Aware MPP Tracking

- **Architecture support**
  - Per-core DVFS

6 V/F × 8 core=48 load levels

- **Stepwise, successive tracking**
  - Progressively move the multi-core load to MPP

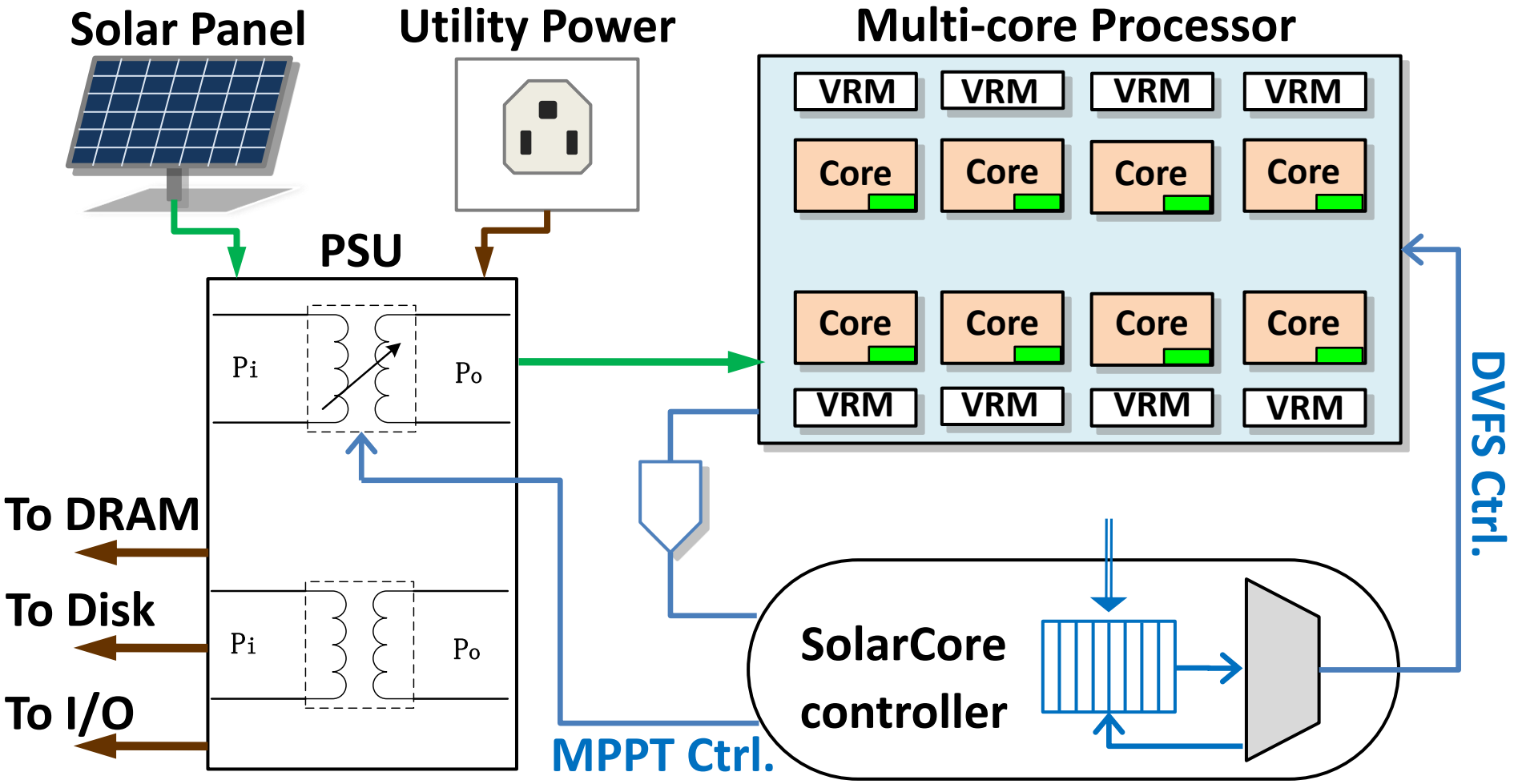


Tracking ≠ Performance

- **Metrics: performance-time-product (PTP)**
  - Throughput × Runtime/Day = Total instructions committed
  - Needs effective optimization along with the tracking

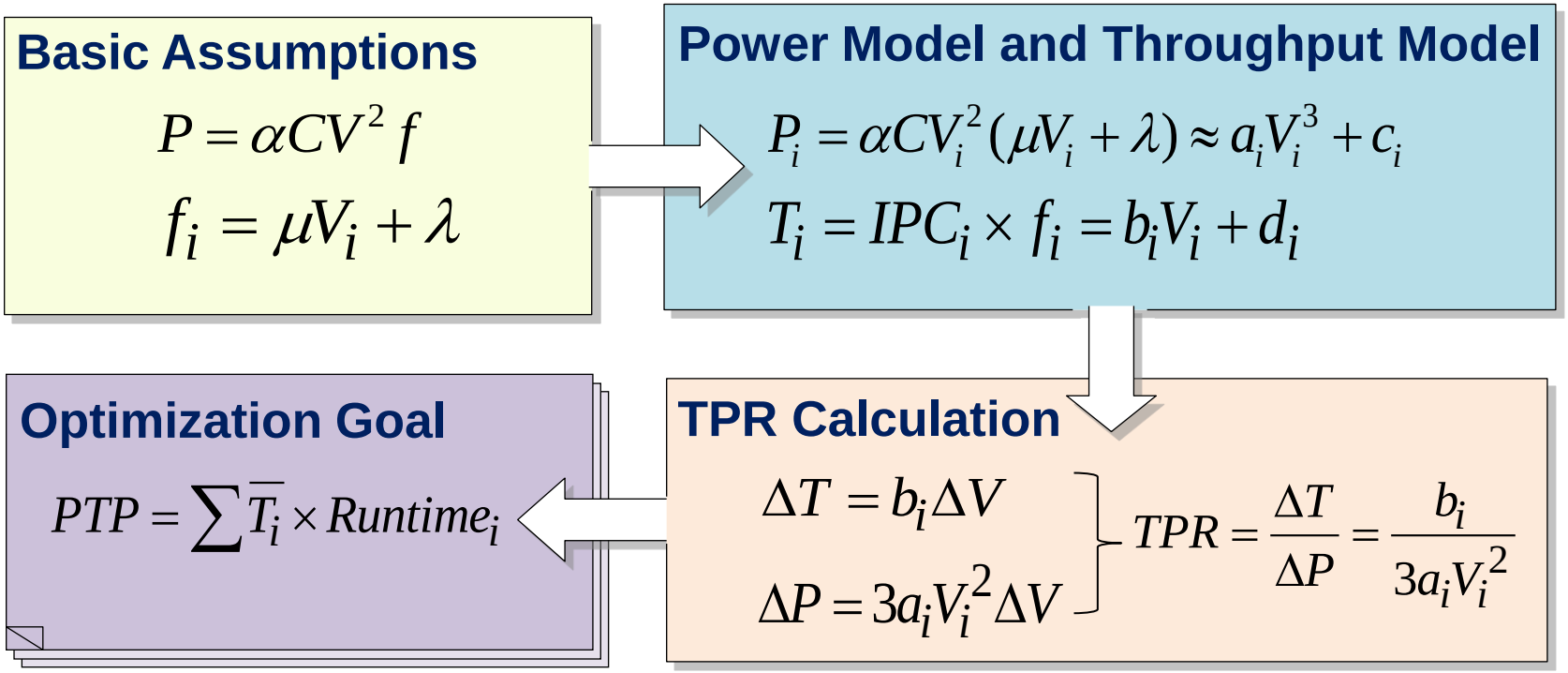
- **Improve PTP using throughput-power ratio (TPR)**
  - Performance-per-watt evaluates computation efficiency

Per-core Load Adaptation



Calculate Throughput-Power Ratio

- **Allocate precious power to high productive cores**
  - Predict the return on investment (ROI)



TPR Calculation

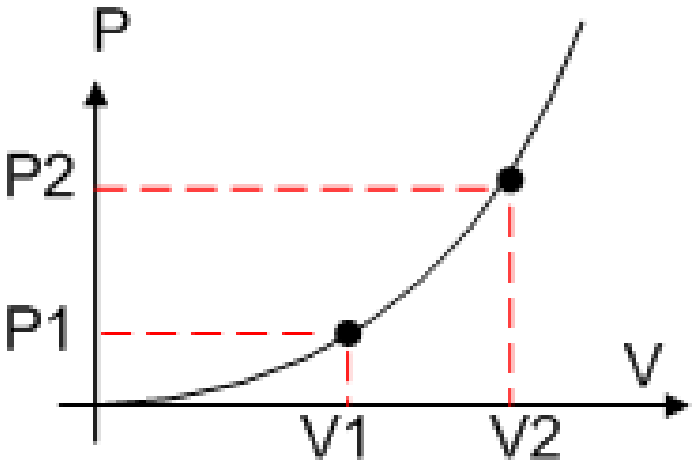
- **Profiling**
  - Two operation points: (V1, P1), (V2, P2)
  - IPC

$$P = \alpha C V^2 f \qquad f_i = \mu V_i + \lambda$$

$$P_i = \alpha C V_i^2 (\mu V_i + \lambda) \approx a_i V_i^3 + c_i$$

$$T_i = IPC_i \times f_i = b_i V_i + d_i$$

$$\left.\begin{aligned} \Delta T &= b_i \Delta V \\ \Delta P &= 3 a_i V_i^2 \Delta V \end{aligned}\right\} \quad TPR = \frac{\Delta T}{\Delta P} = \frac{b_i}{3 a_i V_i^2}$$



Compute a and c using performance counter statistics

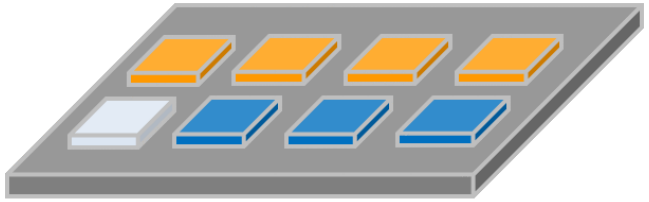
Per-core Load Adaptation Policy

- **Tune one core at each timestamp**
  - Increase/decrease on V/F level

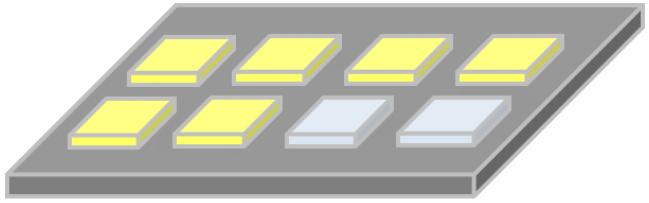
Keeps tuning individual core until reaching its highest or lowest V/F level

Distribute the additional renewable power evenly across all the cores in a round-robin fashion

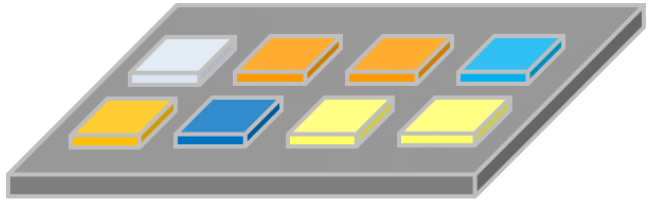
Selects cores based on the throughput-power ratio (TPR) metrics



Tuning individual core (IC)



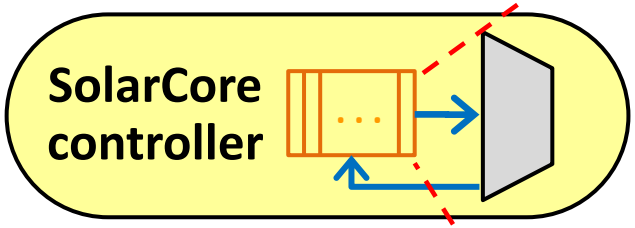
Round-robin allocation (RR)



Performance oriented (Opt)

Performance Oriented Core Selection

- **Core tuning table**
  - Tracks core status
  - Periodically updated



- **Operation rule**
  - Front pointer
    - Voltage level < 6
    - Chooses higher TPR
  - Rear pointer
    - Voltage level > 0
    - Chooses lower TPR

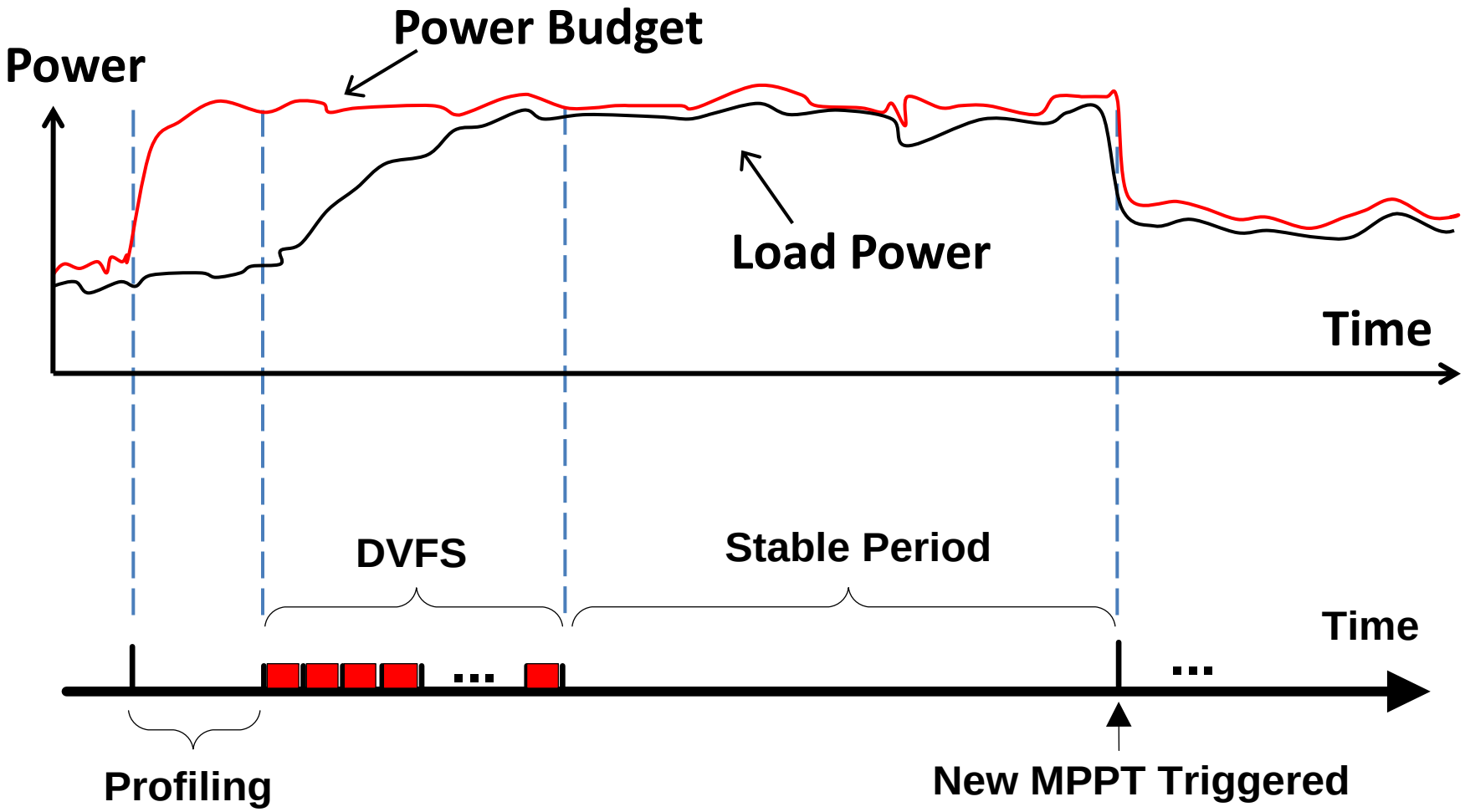
Table format

Core ID	Voltage level
0	6
5	6
7	6
6	6
4	3
3	6
1	5
2	0

F
F
R
R
high
TPR
low

Core tuning table

SolarCore Management Timeline



Summary, Reference, and Exercises

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

Reference: 6th Generation Intel® Processor Datasheet for S-Platforms

Question: What would happen if a computer in state S1/S2/S3 loses all its AC connection or battery power?