Software-Cooperative Power-Efficient Heterogeneous Multi-core for Media Processing

Hiroaki Shikano*,**, Masaki Ito*, Kunio Uchiyama*,
Toshihiko Odaka*, Akihiro Hayashi**, Takeshi Masuura**,
Masayoshi Mase**, Jun Shirako**, Yasutaka Wada**,
Keiji Kimura**, Hironori Kasahara**

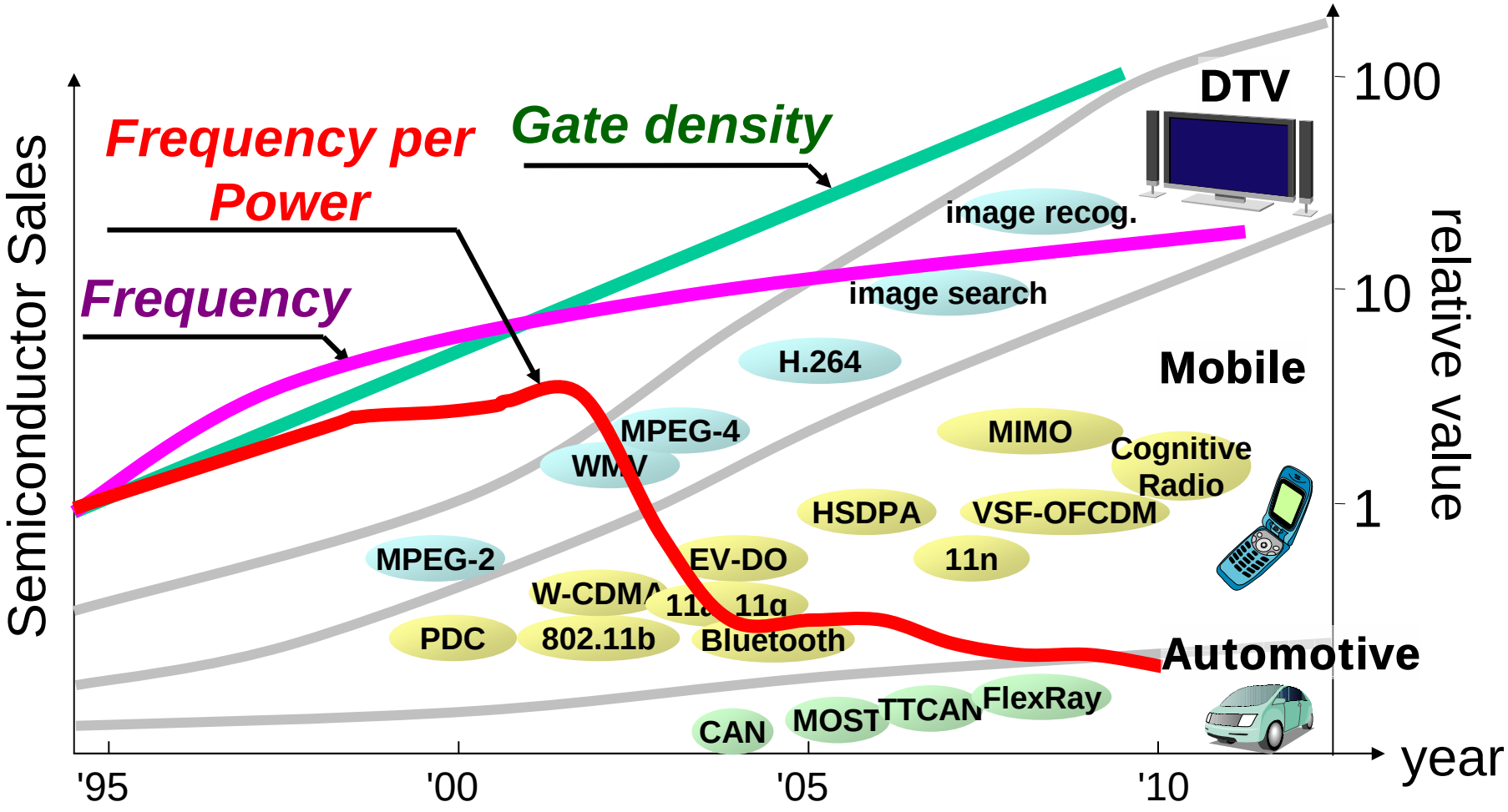
*Hitachi, Ltd. **Waseda University

ASP-DAC 2008

Contents

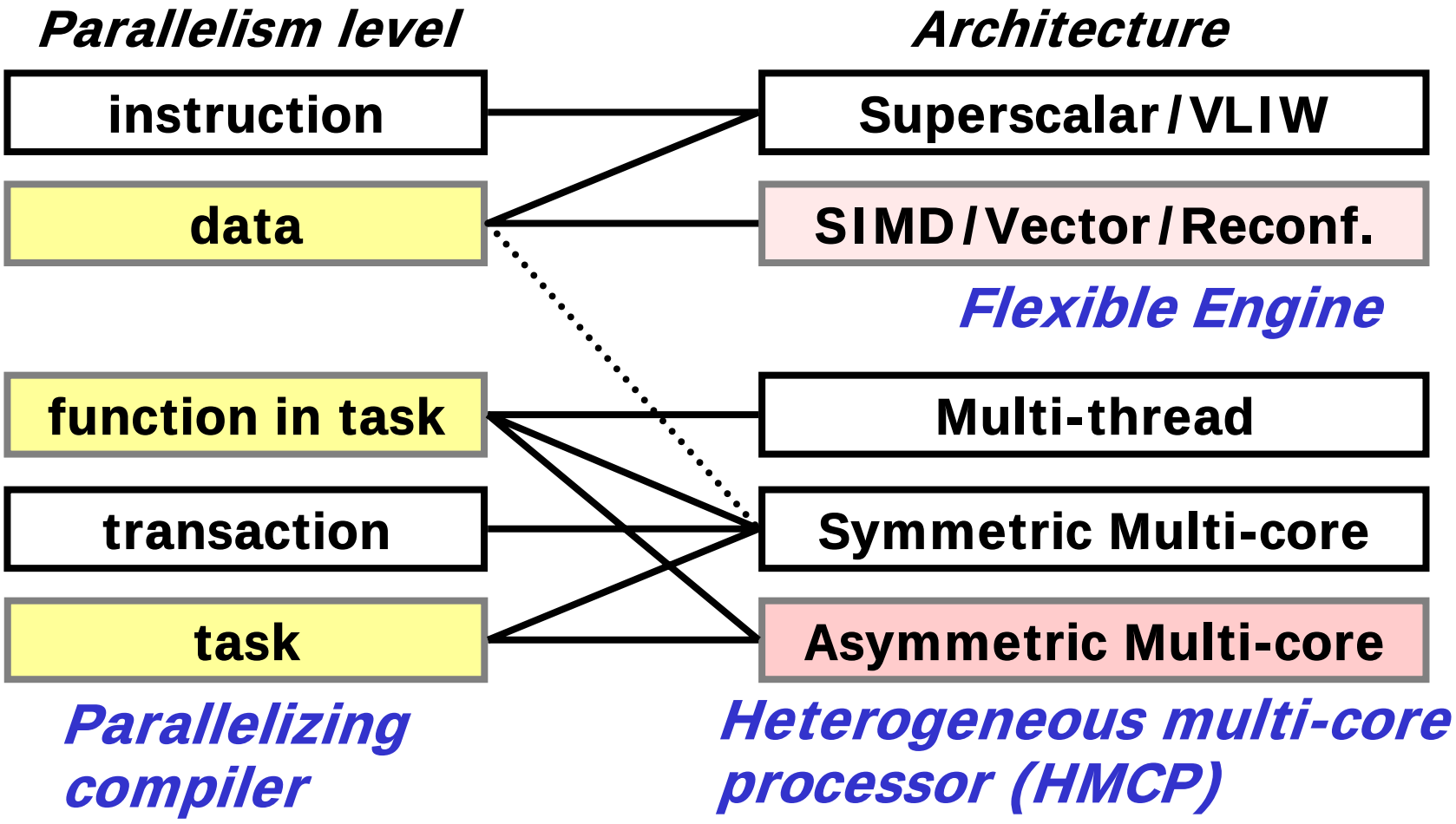
3 Digital semiconductor trends

- Integration of more and more functions into a chip
- Gate density continues to increase (~32nm)
- Frequency saturation and power become issues

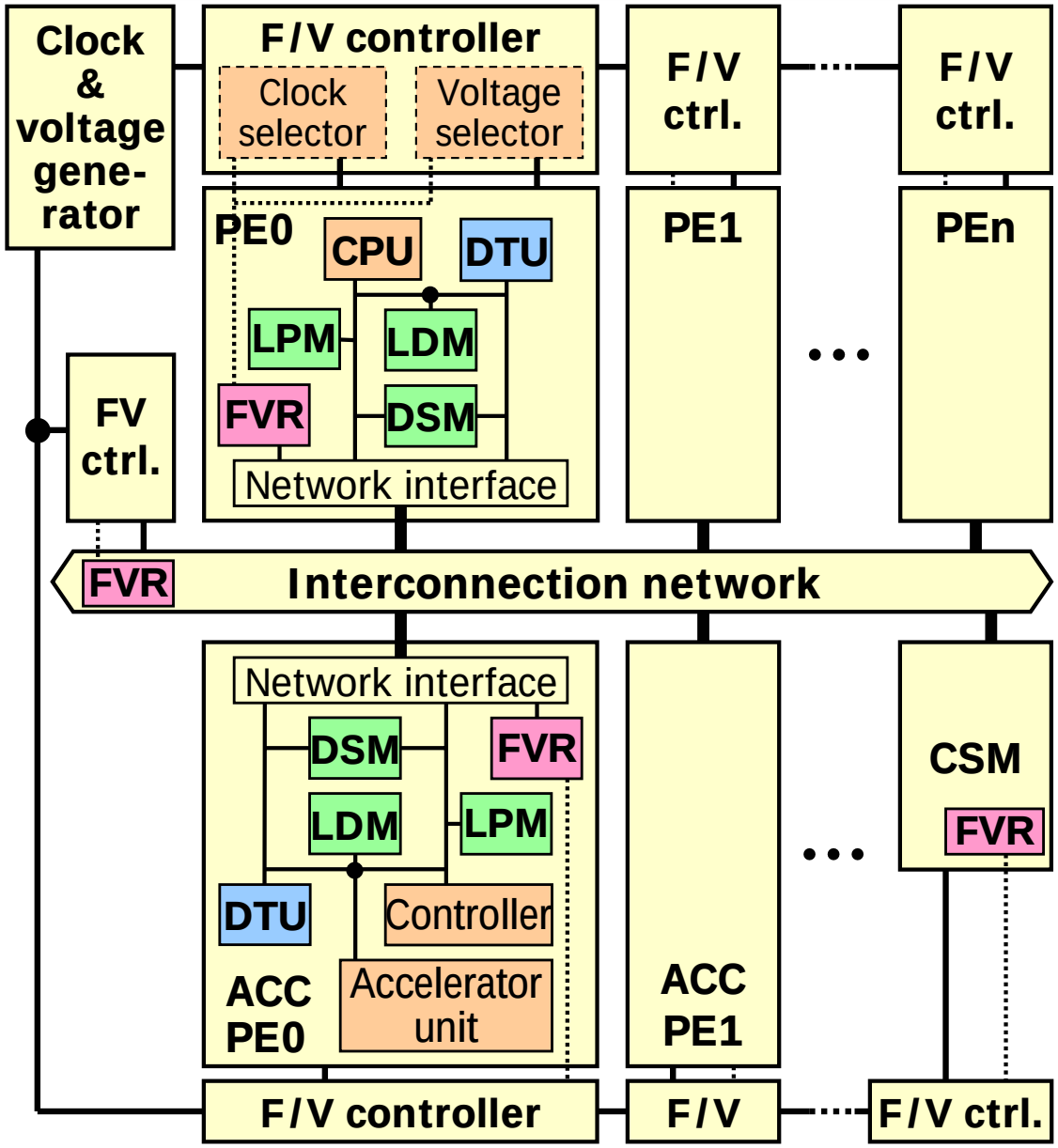


4 Processing architecture direction

- Our approach: to achieve high-performance by parallel architecture utilizing high-gate density with support of parallelizing compiler for high software productivity



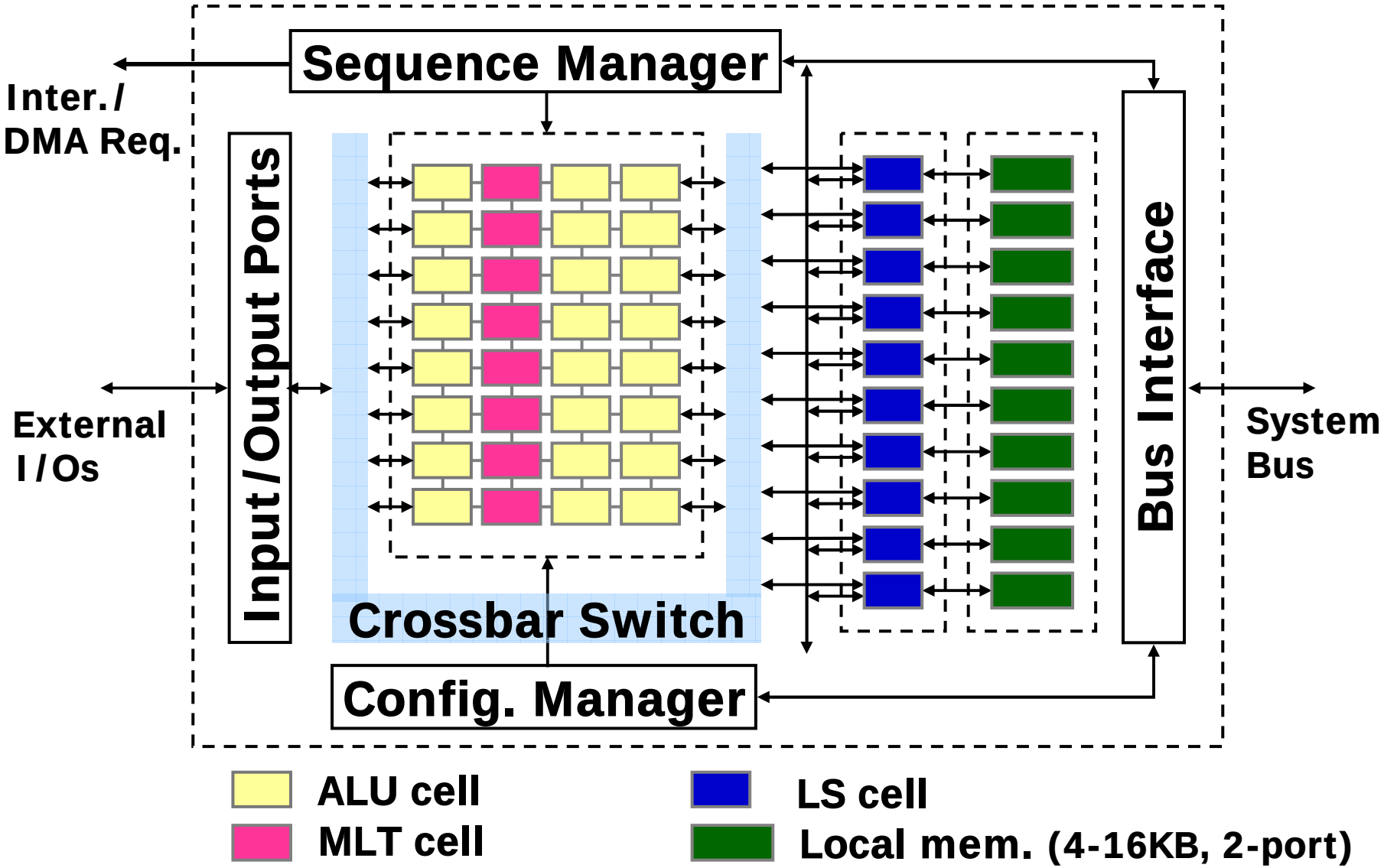
5 Power-aware HMCP architecture



- Multiple types of processors
  - CPUs + accelerators (ACC)
- Unified hierarchical memory architecture
  - LDM: local data memory
  - DSM: distributed shared memory
  - LPM: local program memory
  - CSM: centralized shared memory
- Programmable data transfer unit (DTU)
- Power control register (FVR)

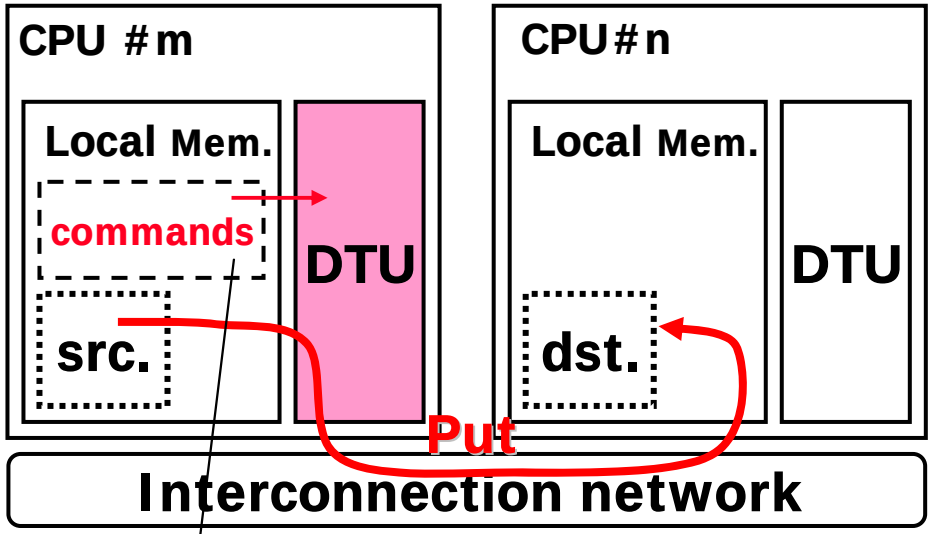
6 Accelerator core; Flexible Engine(FE)

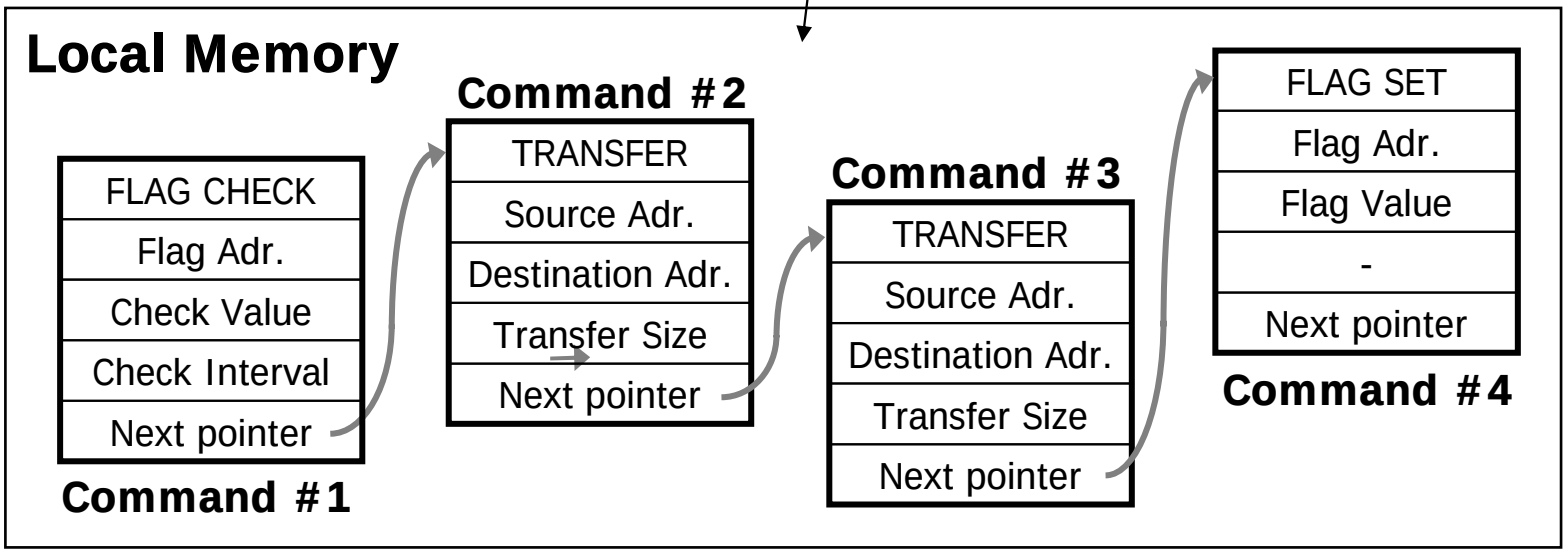
- A dynamically reconfigurable processor as an accelerator



Data transfer unit (DTU)

- Concurrent data transfer with CPU computation
- Programmability by transfer commands on local memories
  - Put/get commands
  - Flag check /set commands





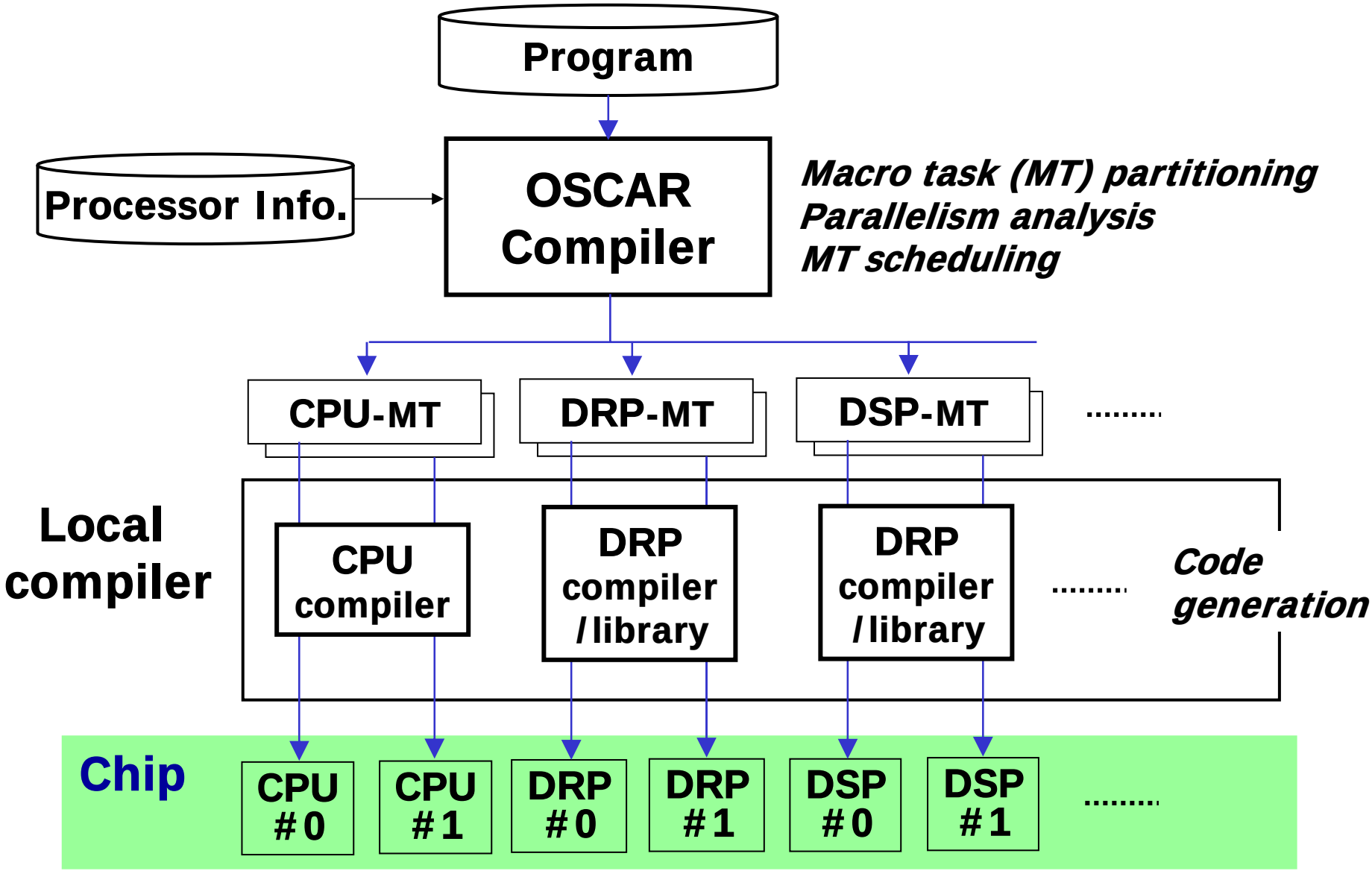
OSCAR parallelizing compiler

Improve effective performance, cost-performance and productivity and reduce consumed power

- **Multigrain parallelization**
  - Exploitation of parallelism from the whole program by use of coarse-grain parallelism among loops and subroutines, near fine grain parallelism among statements in addition to loop parallelism
- **Data localization**
  - Automatic data distribution for distributed shared memory, cache and local memory on multiprocessor systems
- **Data transfer overlapping**
  - Data transfer overhead hiding by overlapping task execution and data transfer using DMA or data pre-fetching
- **Power reduction**
  - Reduction of consumed power by compiler control of frequency, voltage and power shut down with hardware supports

*Optimally SCheduled Advanced multiprocessoR

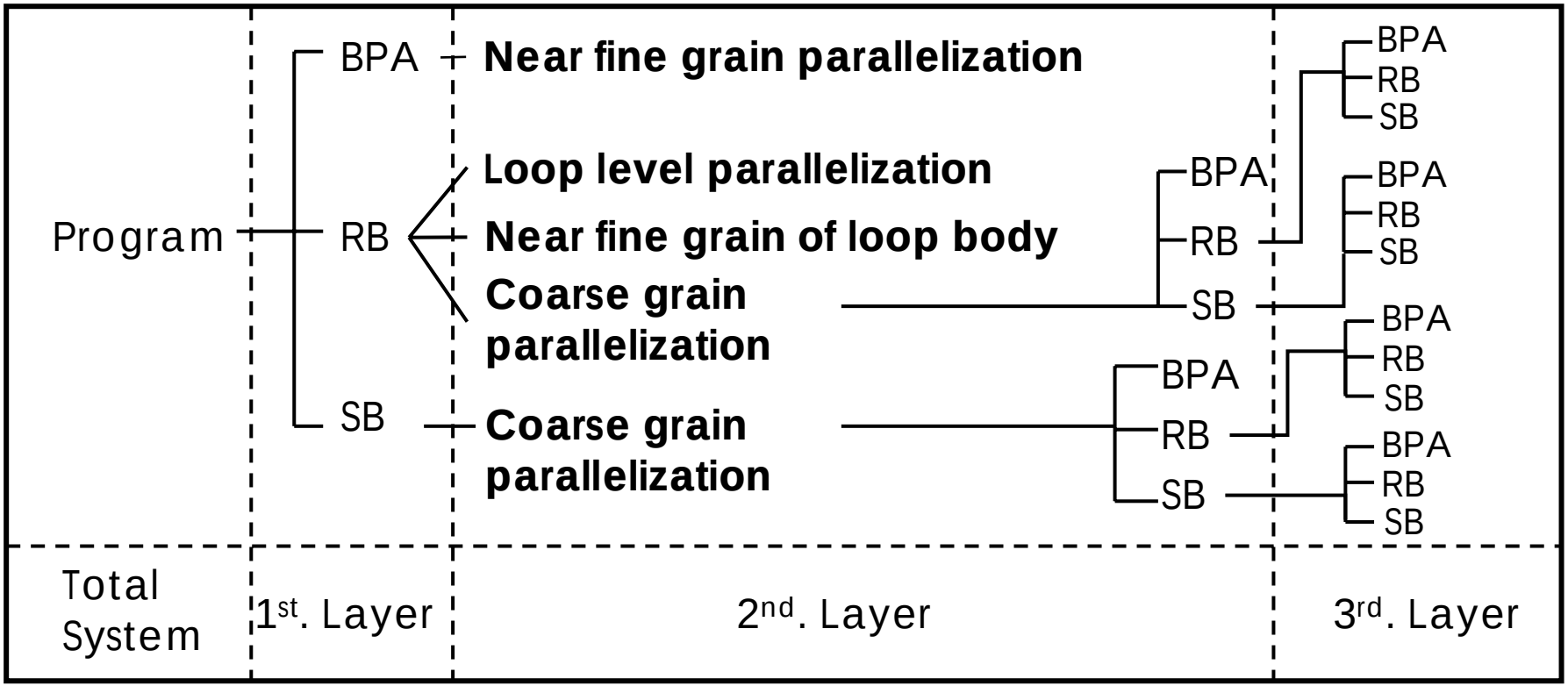
Compiling steps



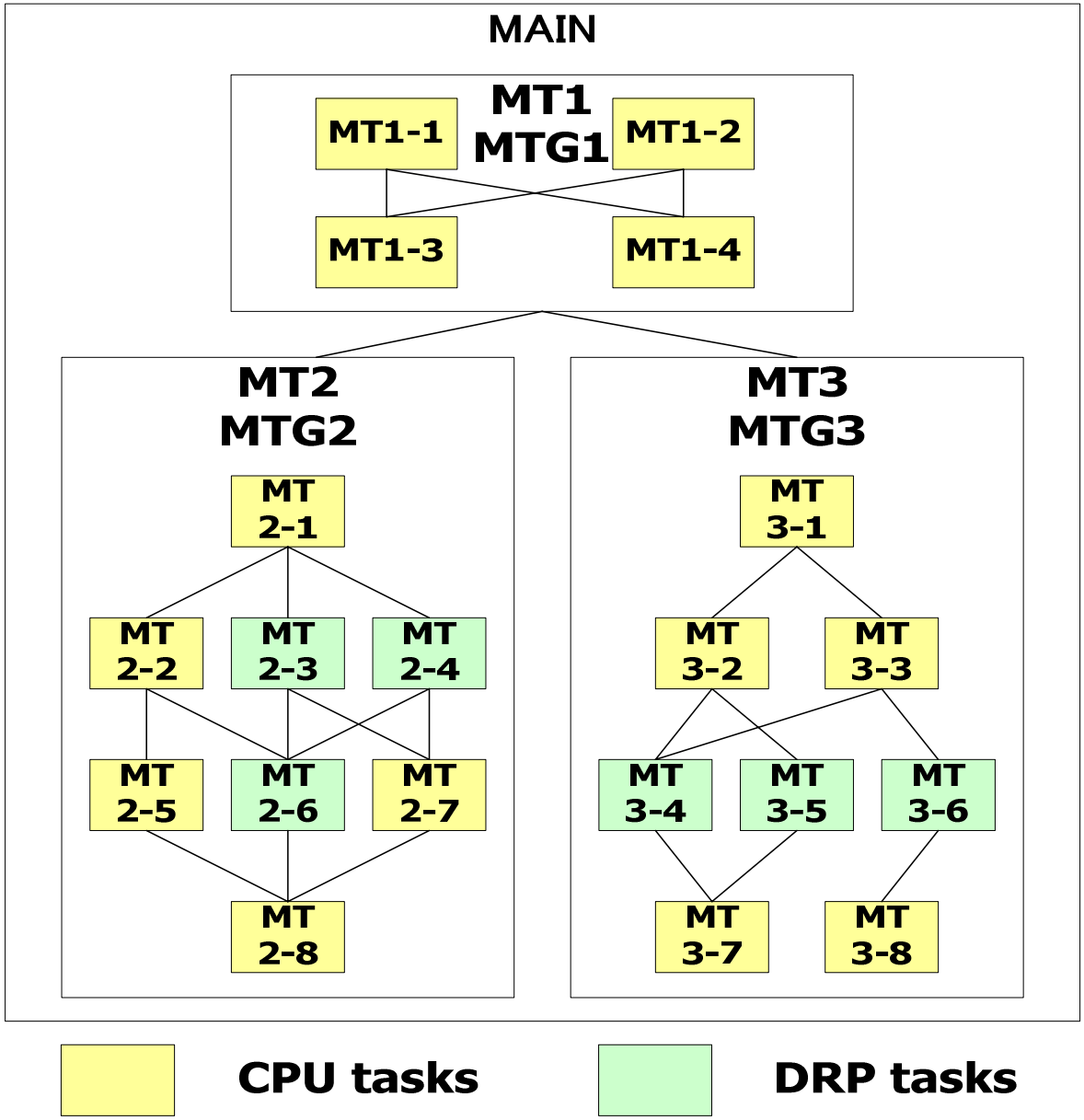
Generation of coarse grain tasks

Macro-tasks (MTs)

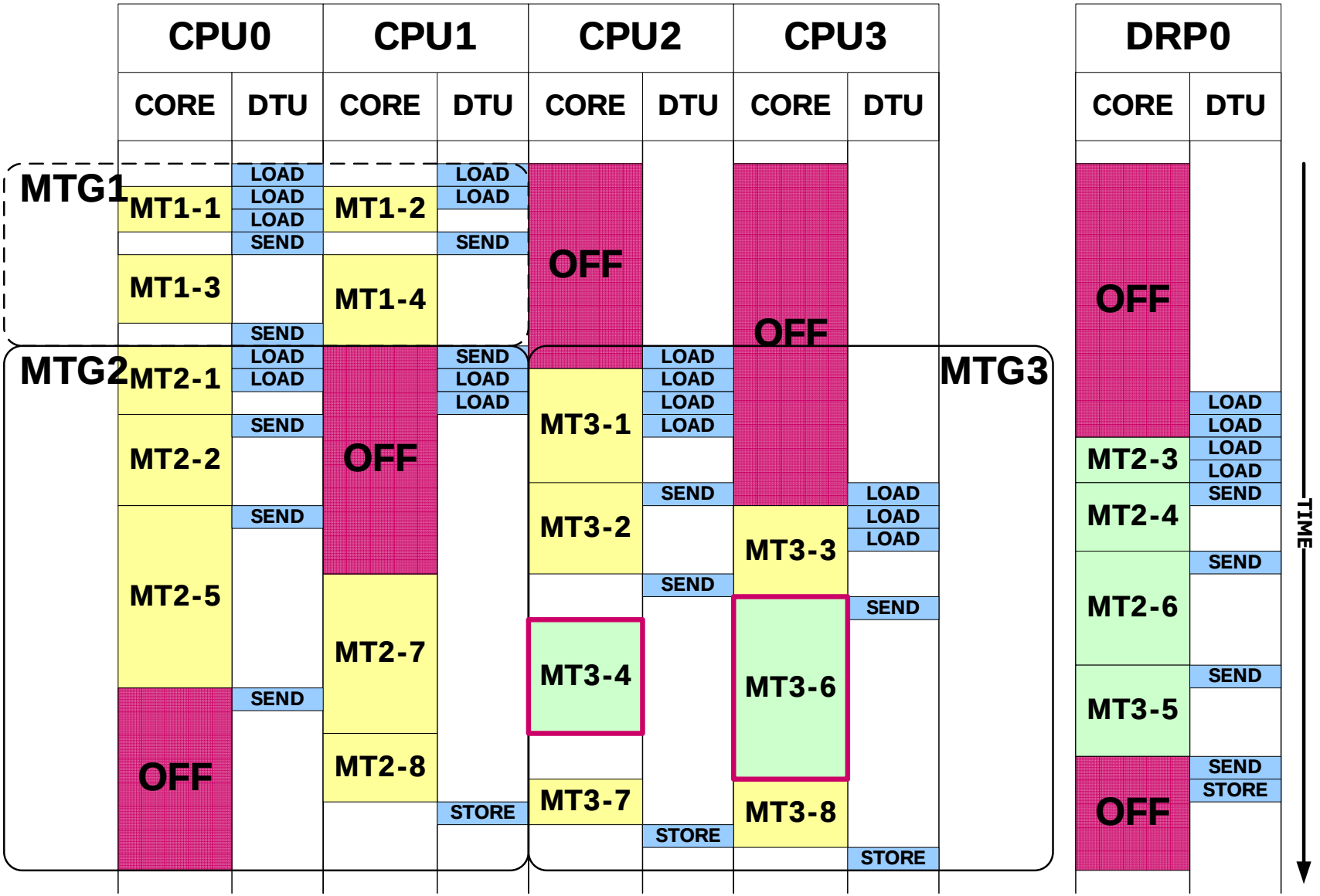
- Block of pseudo assignments (BPA): Basic block (BB)
- Repetition block (RB) : loop
- Subroutine block (SB): subroutine



11 Sample macro-task graph (MTG)



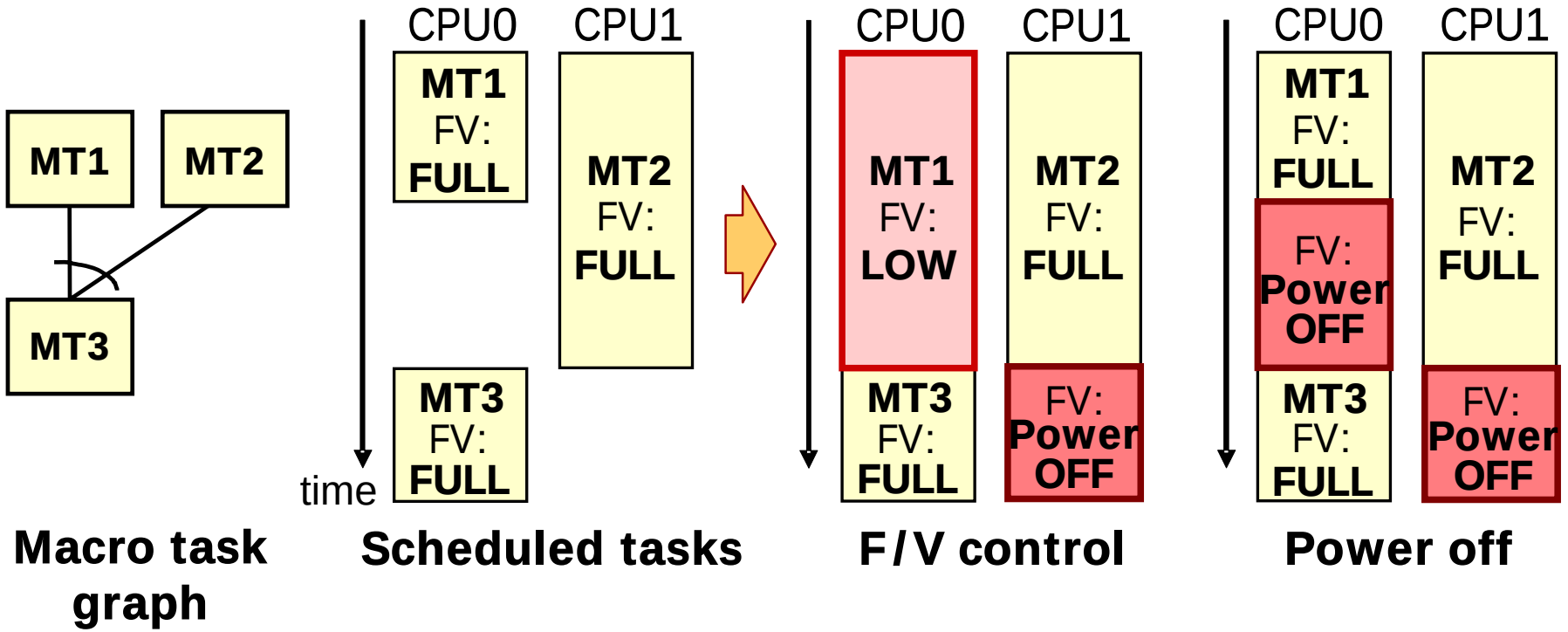
Scheduled-tasks execution



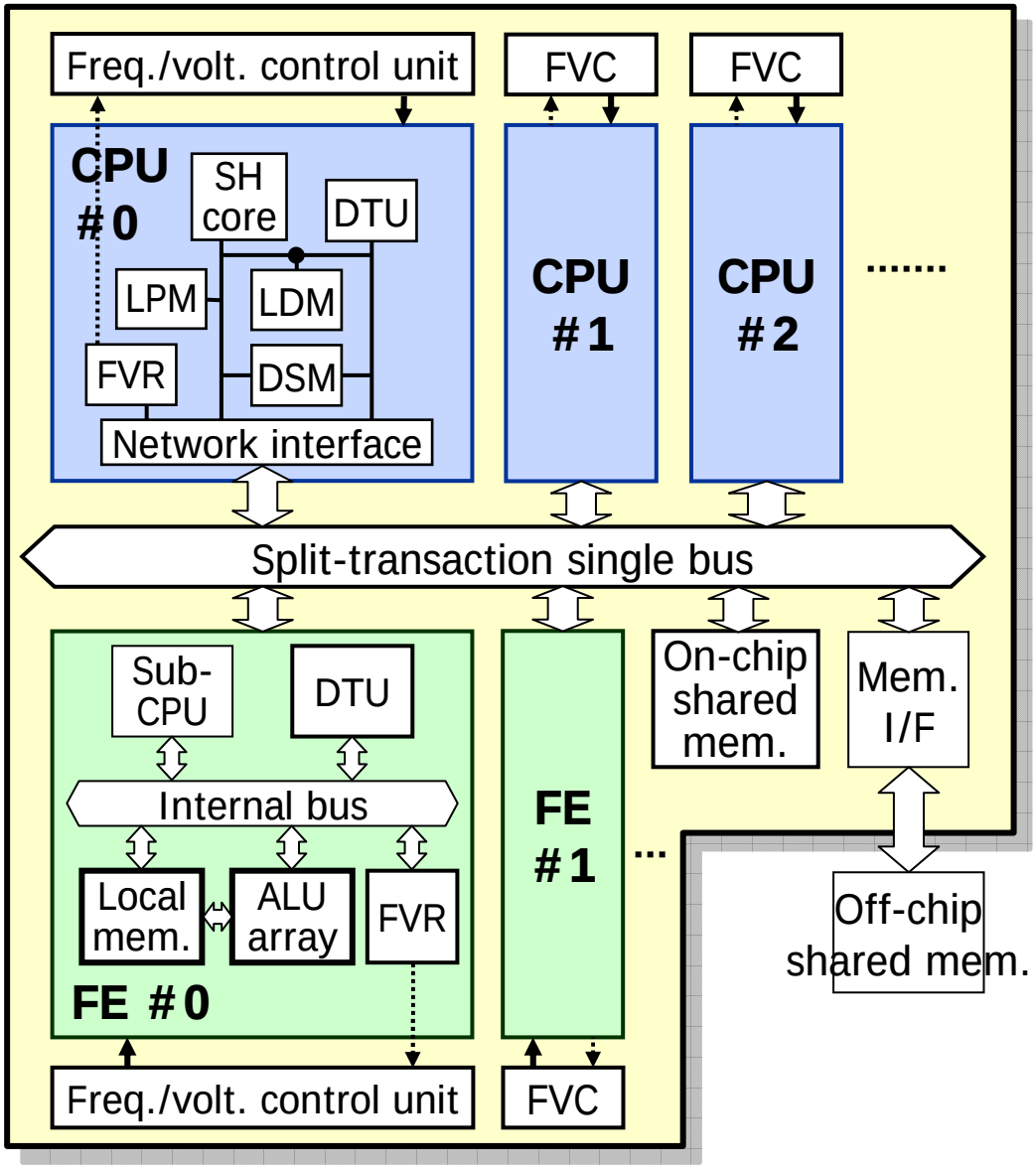
Compiler power saving scheme

Power controlling in parallelized tasks

- Compiler control of frequency / voltage (F/V) and power shut down by utilizing a parallelized task scheduling result
- Reduces power while maintaining parallelized performance



14 Evaluated HMCP architecture

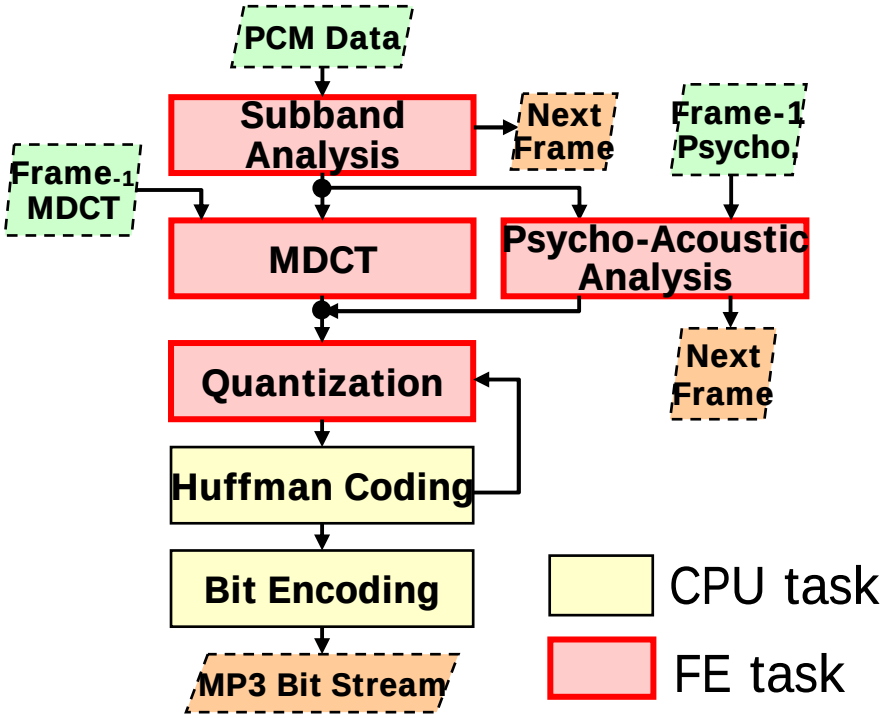


- Cycle-accurate simulator is utilized
- CPU: SuperH (SH)
  DRP: FE w. sub CPU
  - 300 MHz @ 90-nm tech.
- Memory lantecy
  - Local memory: 1-cycle latency (local), 4-cycle latency (remote)
  - On-chip SM: 4-cycle latency
  - Off-chip SM: 16-cycle latency
- Wattch-based power model
  - Parameters were introduced from RTL-level power simulation on SH processors

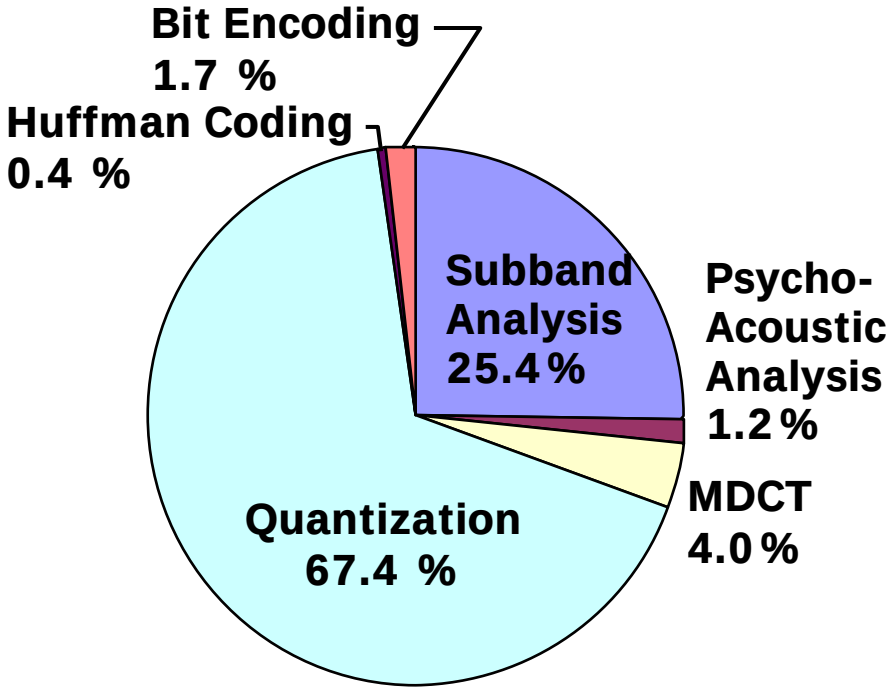
15 Evaluated application; MP3 encoder

- MP3 Audio Encoding as firstly evaluated application targeting car navigation systems as an example
  - Encoding processed in each individual audio frame and inter-frame parallelism exists
  - 16 frames of 16-bit 44.1-KHz input and 128-kbps stream output

Process flow of MP3 encoding



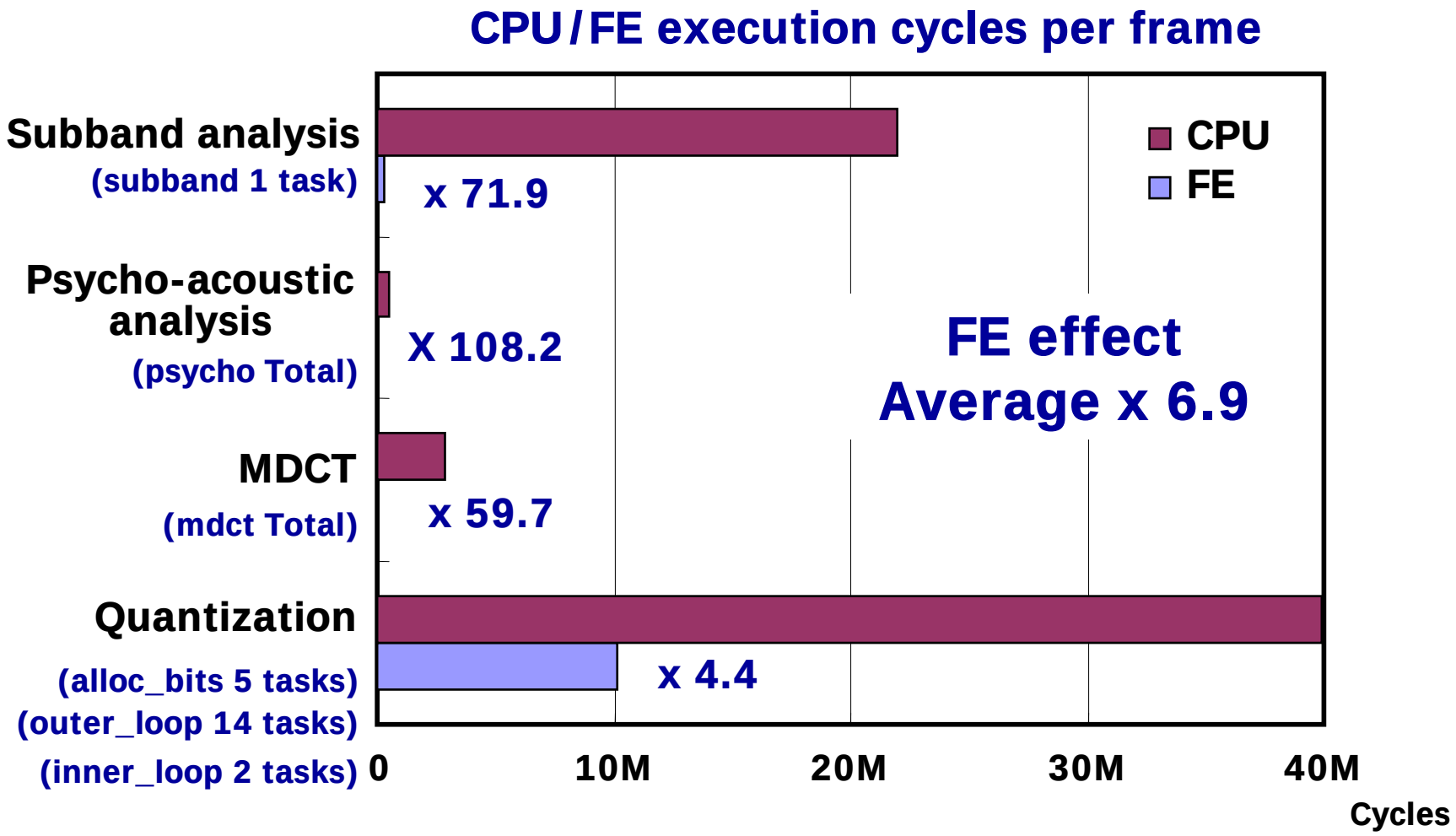
Profiling Result



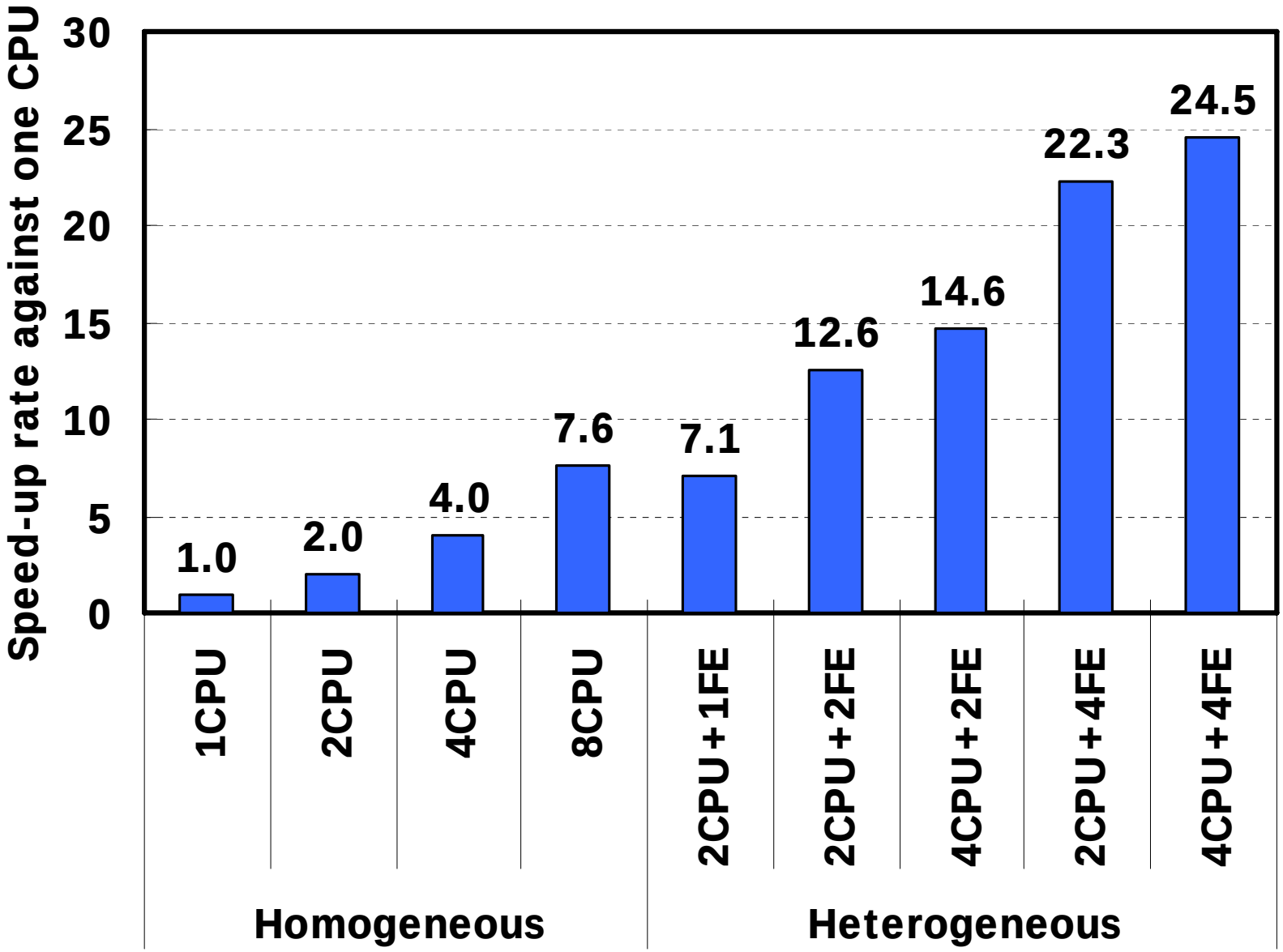
MDCT: Modified Discrete Cosine Transform

Effect of FE

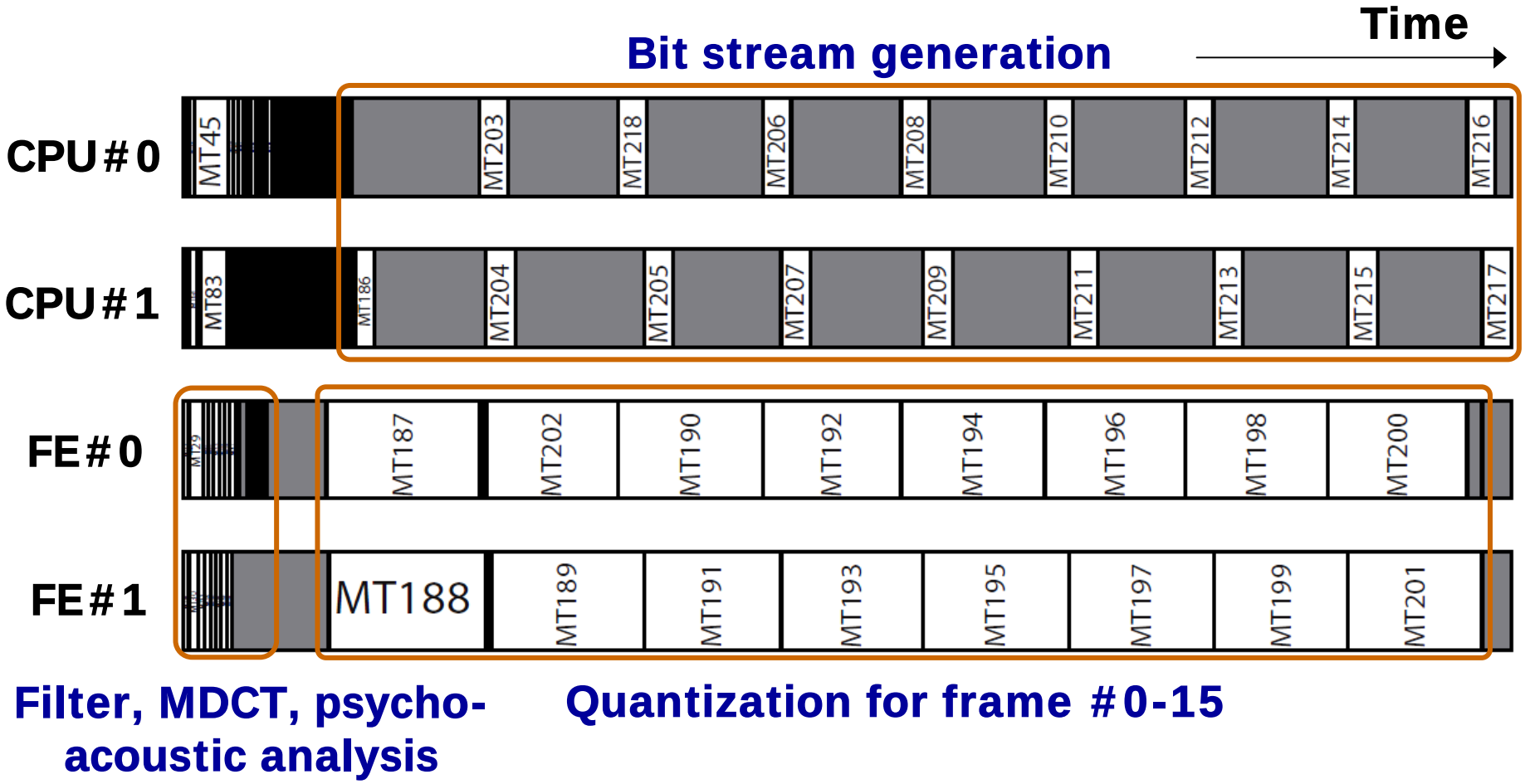
- 24 macro tasks assigned for FEs in MP3 encoder



Performance evaluation



Macro task trace (SHx2＋DRPx2)



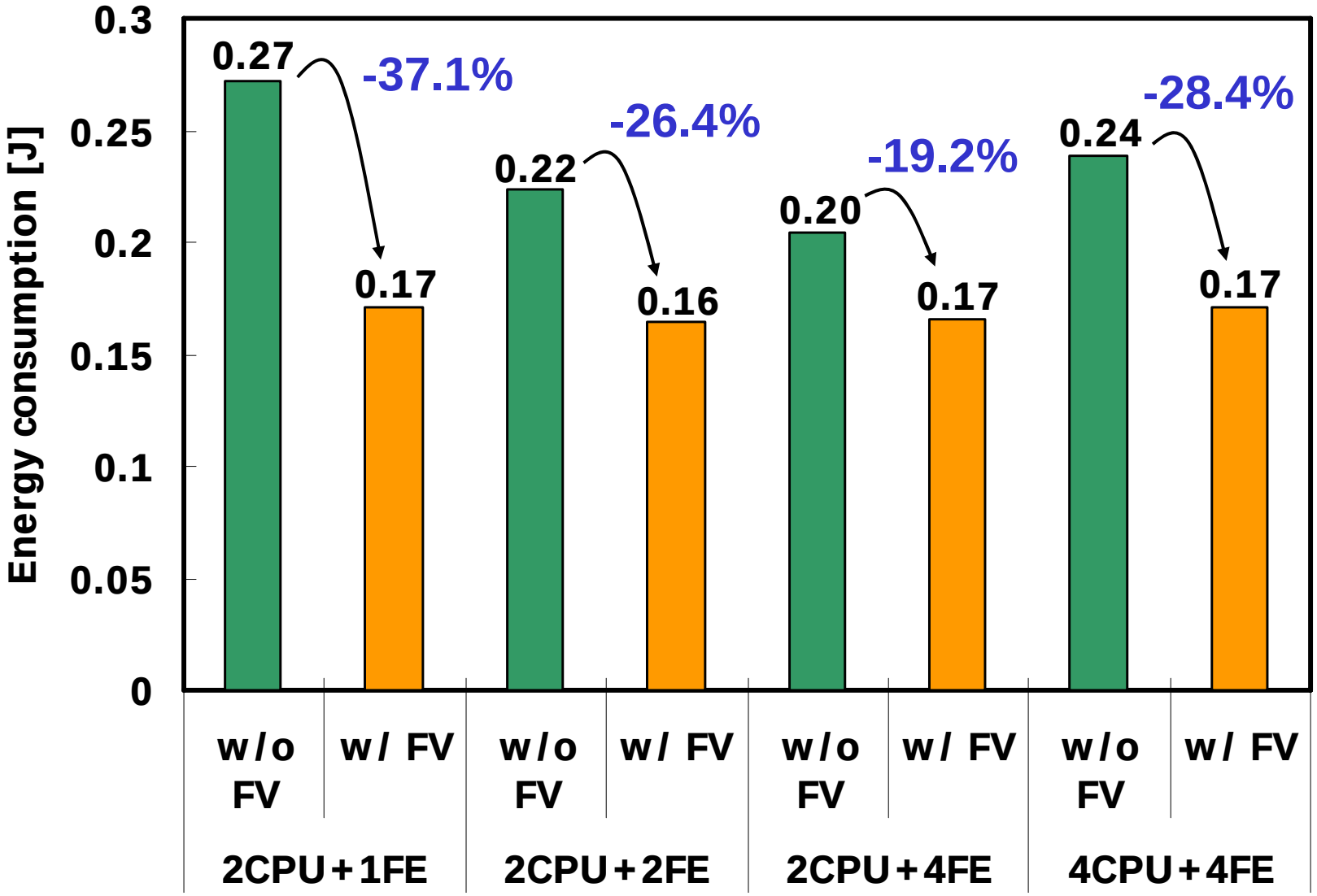
19 Power control mode

Status of the FV power mode

- CPU: Avg. power 150 mW @ 300 MHz, 1.0 V (FULL)
  - FULL, MID, LOW, OFF supported
- FE: Avg. power 210 mW @ 300 MHz, 1.0 V (FULL)
  - FULL, OFF supported

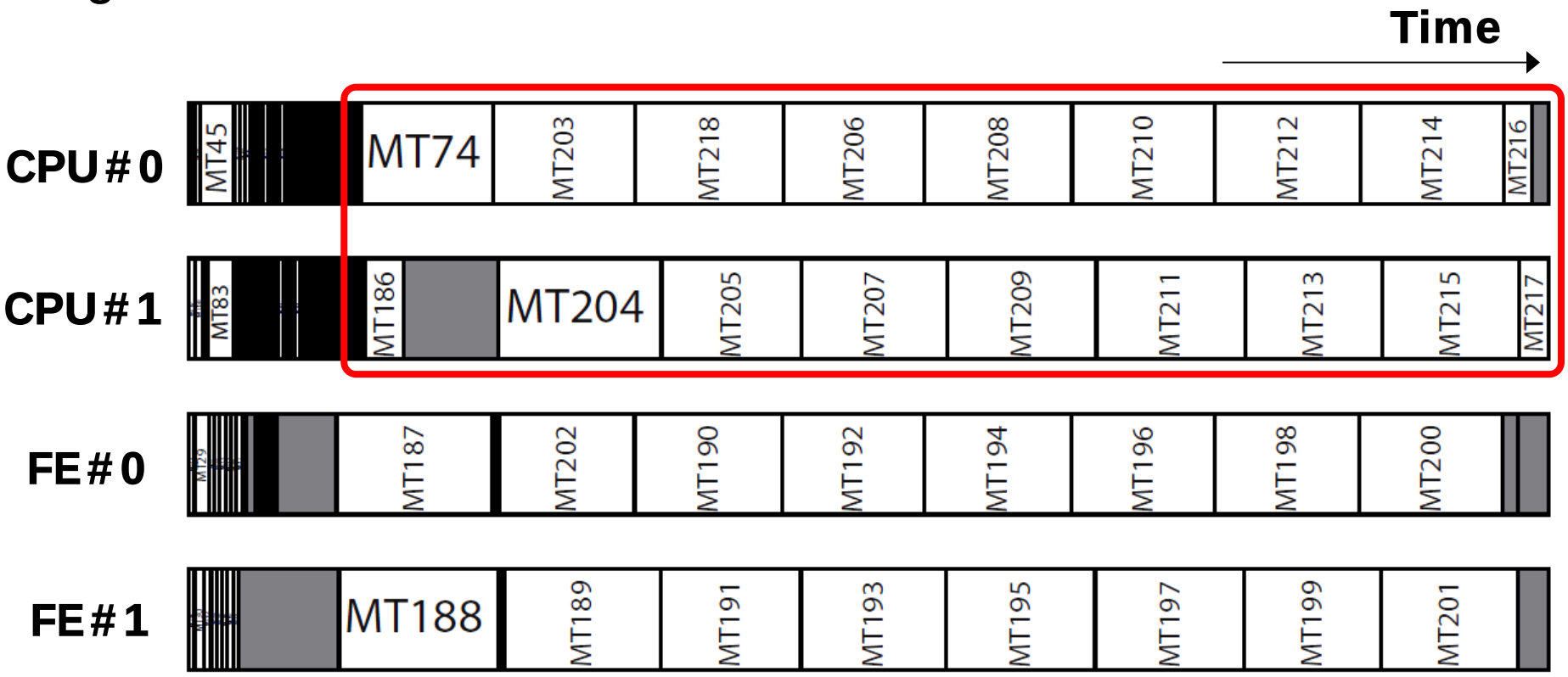
	FULL	MIDDLE	LOW	OFF
Clock frequency	**1** (300 MHz)	**1/2** (150 MHz)	**1/4** (75 MHz)	**0** (Clock off)
Supply voltage	**1** (1.0 V)	**0.87** (0.87 V)	**0.71** (0.71 V)	**0** (Power off)
Leakage power	**1** (1.0 %)	**1** (1.0 %)	**1** (1.0 %)	**0** (No power)

Power control effects by compiler



Macro task trace with power control

- Clock frequency of CPU#0 and #1 is lowered for bit-stream generation tasks



Summary

- **Power-efficient heterogeneous multi-core architecture supported by OSCAR parallelizing compiler was studied**
  - Various types of processor cores on a chip such as CPUs and Flexible Engines as accelerators
  - Unified hierarchical memory architecture throughout the PEs controlled by software to improve performance, power efficiency and programming/compiling efficiency
  - Power control registers reducing power consumption
- **Performance evaluation was performed using MP3 audio encoder**
  - 24.5-folded speed-up in performance on 4 CPUs and 4 FEs against sequential execution on one CPU
  - As much as 37.1% reduction in energy consumption was achieved when the compiler power saving scheme was applied

Thank you for your attention!!