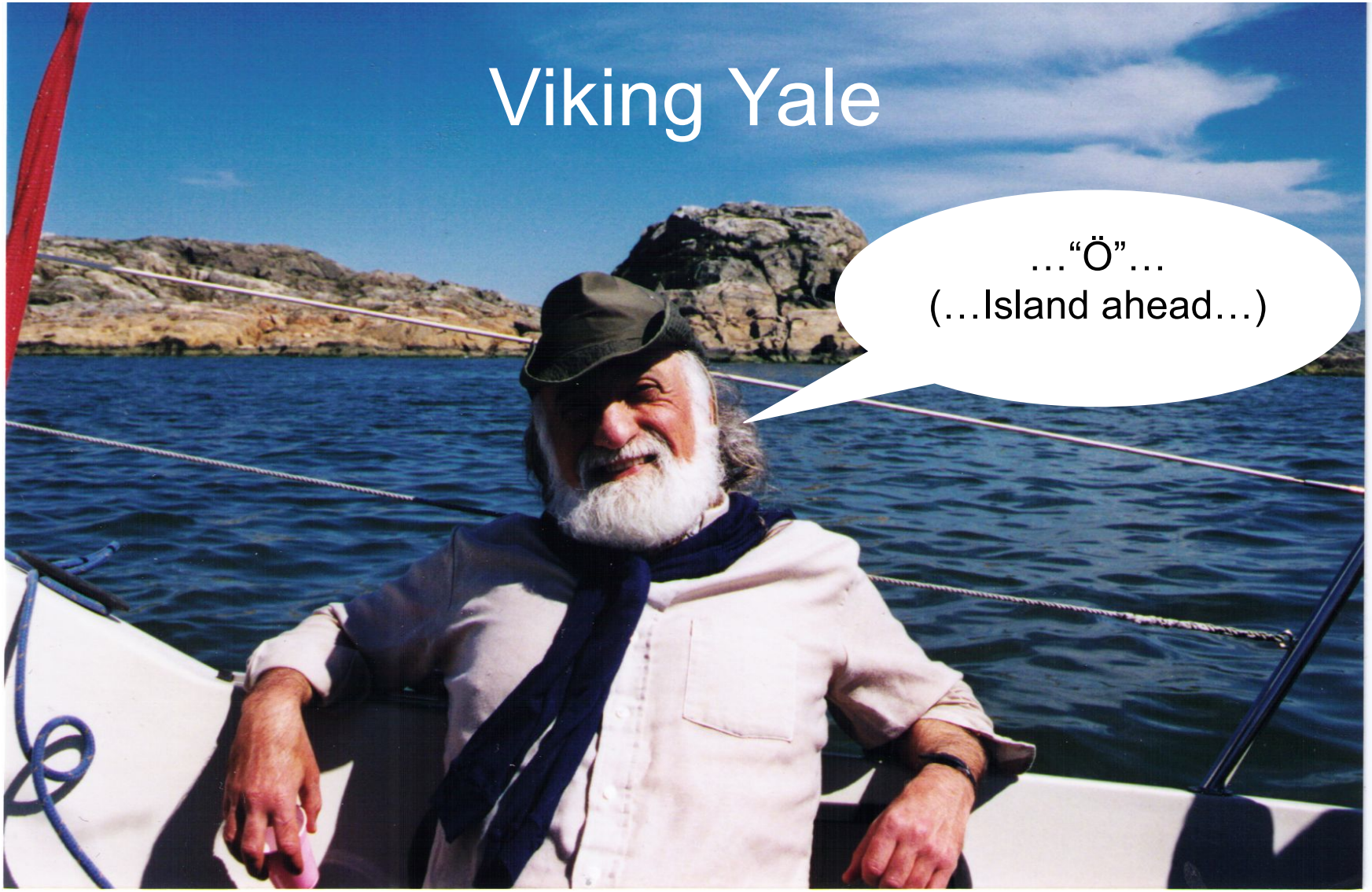
Cross-Layer Driven Computer Architecture Optimizations

Per Stenström

Chalmers University of Technology
Göteborg, Sweden

Viking Yale



MeCCA erc CHALMERS Chalmers University of Technology

Columbus and Huffman (unfairly) got Credit for Viking Discoveries

=

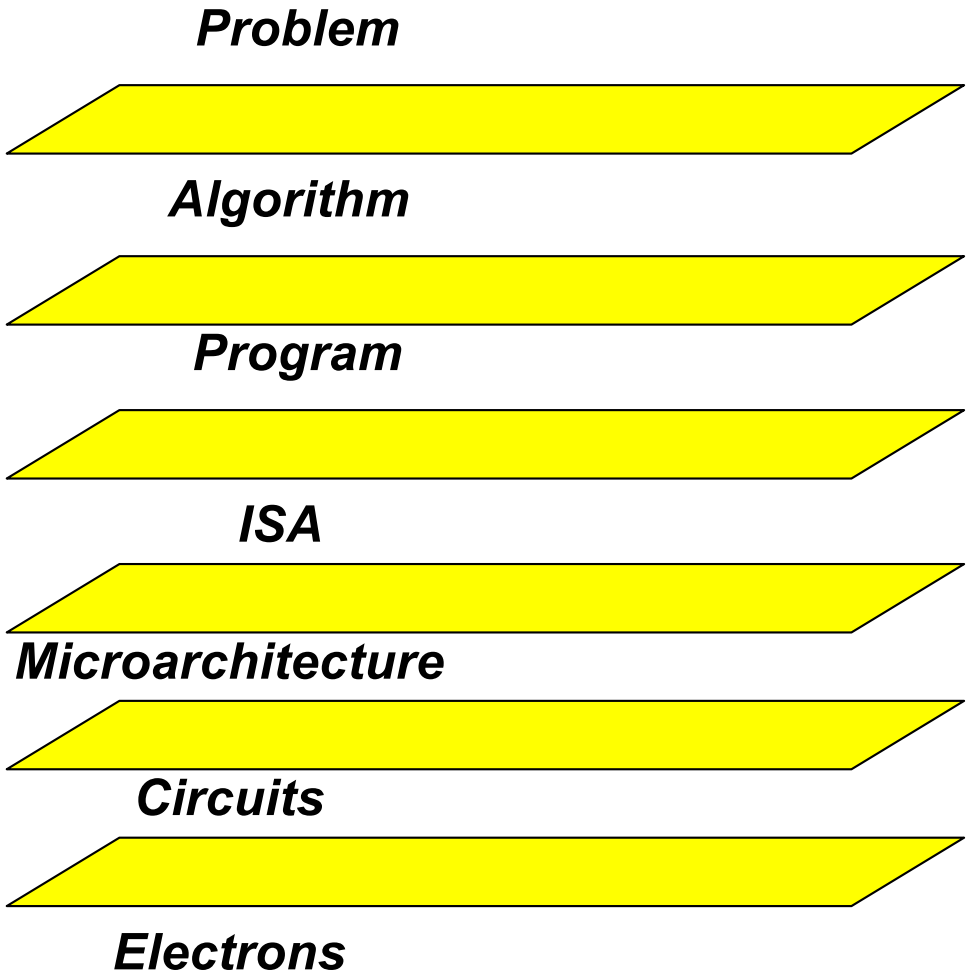
=

Source of inspiration: Yale Patt

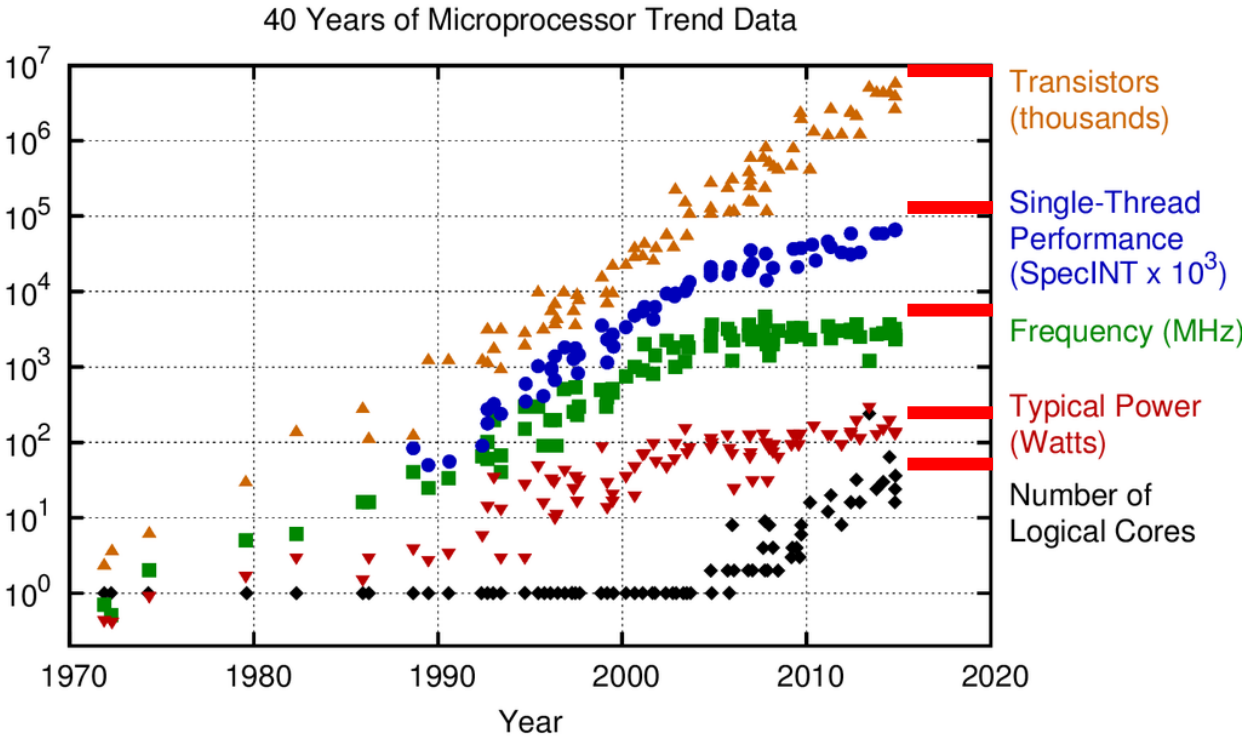
We know that Vikings were on Mars

But Vikings Landed on the Moon before Americans did too

Yale's Transformation Hierarchy



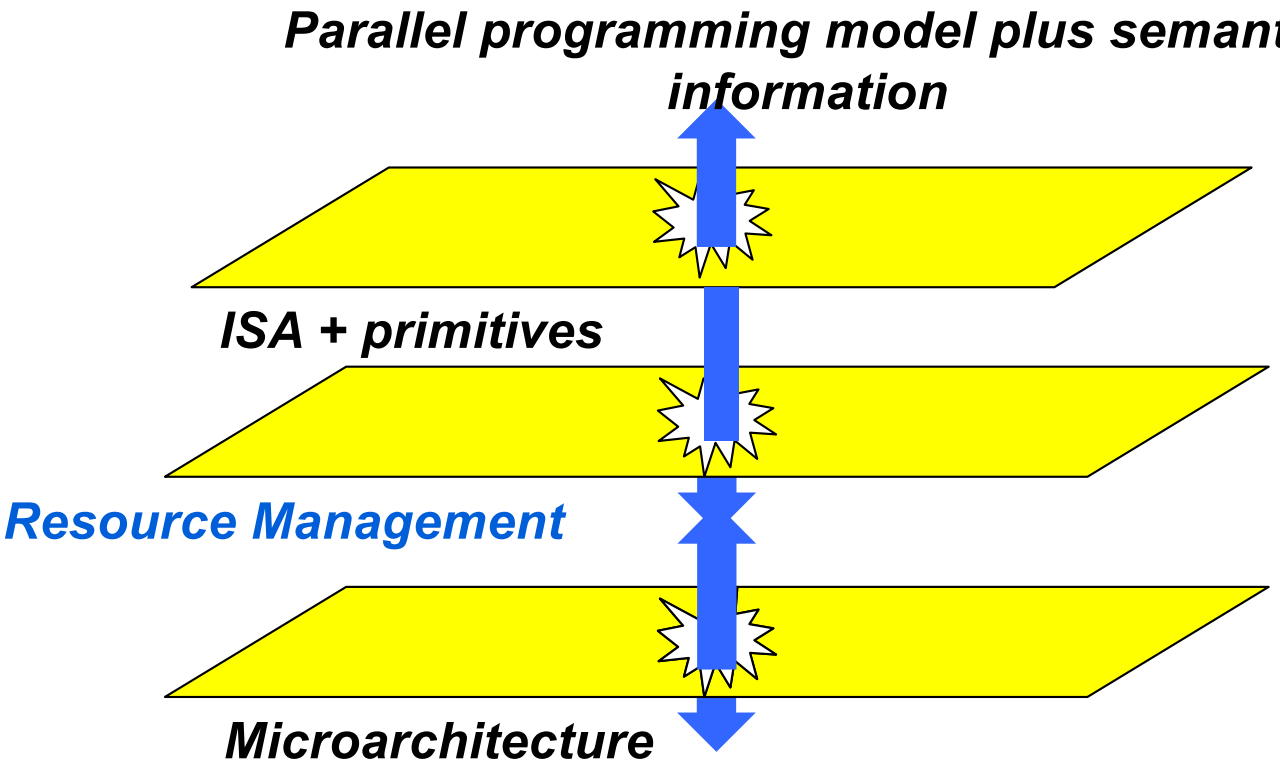
Scaling (as we know it) is ending soon…



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

A radical new way of how we think about compute efficiency is needed

Per's Transformation Hierarchy



Outline

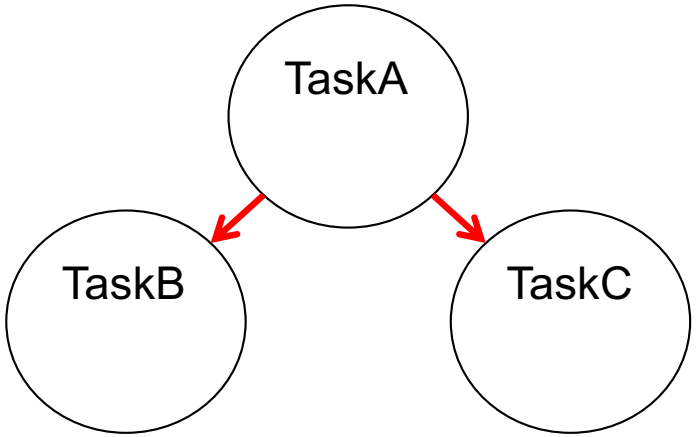
Background

Runtime-Assisted Cache Management

Runtime-Assisted Power Management

Concluding Remarks

Task-based Dataflow Prog. Models




```
#pragma css task output(a)
void TaskA( float a[M][M]);

#pragma css task input(a)
void TaskB( float a[M][M]);

#pragma css task input(a)
void TaskC( float a[M][M]);
```

- Programmer annotations for task dependences
- Annotations used by runtime system for scheduling
- Dataflow task graph constructed dynamically

Convey semantic information to runtime for efficient scheduling

Runtime Management of Cache Hierarchies

View: Runtime is part of the chip and responsible for management of the cache hierarchy

- in analogy with OS managing virtual memory

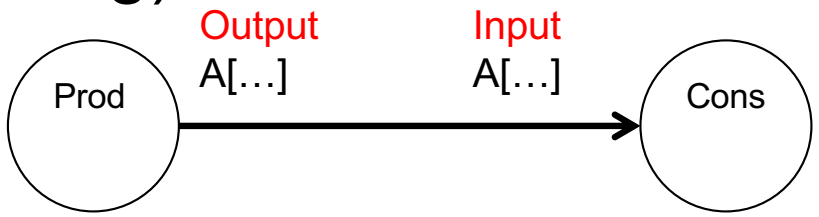
Runtime-assisted cache management:

- Runtime-assisted cache coherence optimizations
- Runtime-assisted dead-block management
- Runtime-assisted global cache management

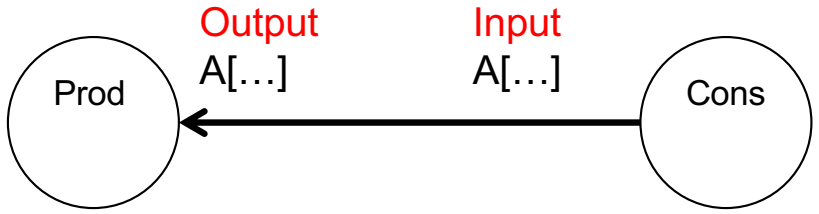
Runtime-Assisted Coherence Management [IPDPS 2014]

Dependency annotations allow for optimizations with high accuracy (like in message passing)

Bulk data transfer

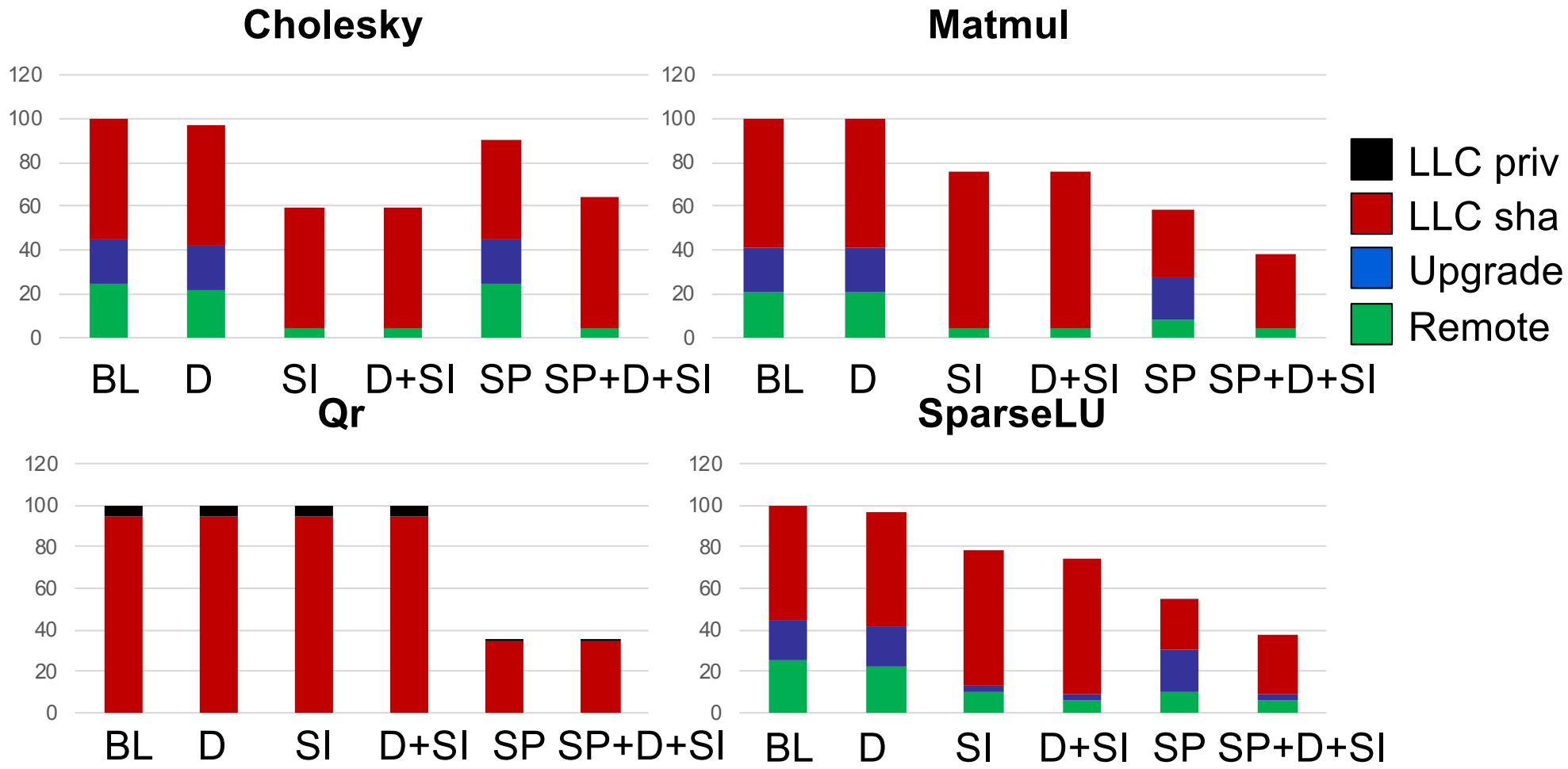


Prefetching



Migratory sharing optimization

Impact on Memory Stall Time

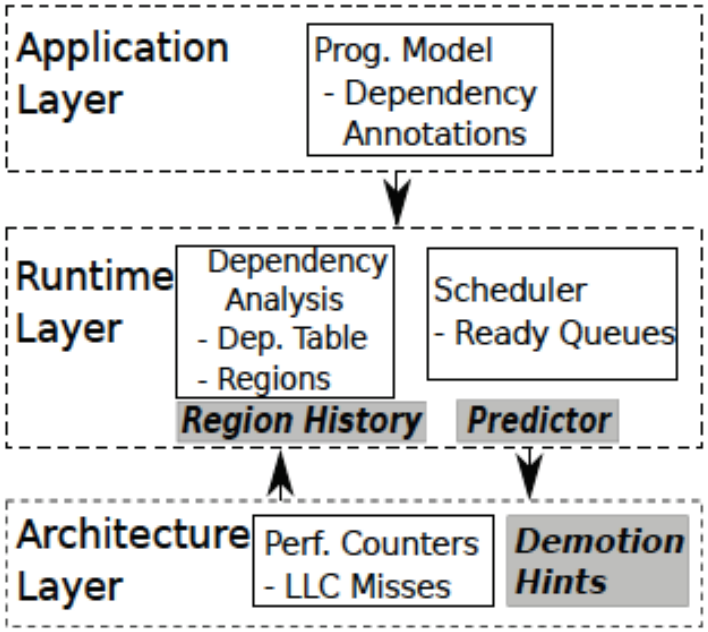


Cross-layer coherence management can yield significant gains

Runtime-Assisted Dead-Block Management – Motivation

- Most Last-Level-Cache blocks are dead consuming precious cache space
- State-of-the-art schemes: prediction based on past behavior
- **RADAR's approach:** Semantic information + prediction

Overview of RADAR [HPCA 2017]



Specify Task and its Region Accesses

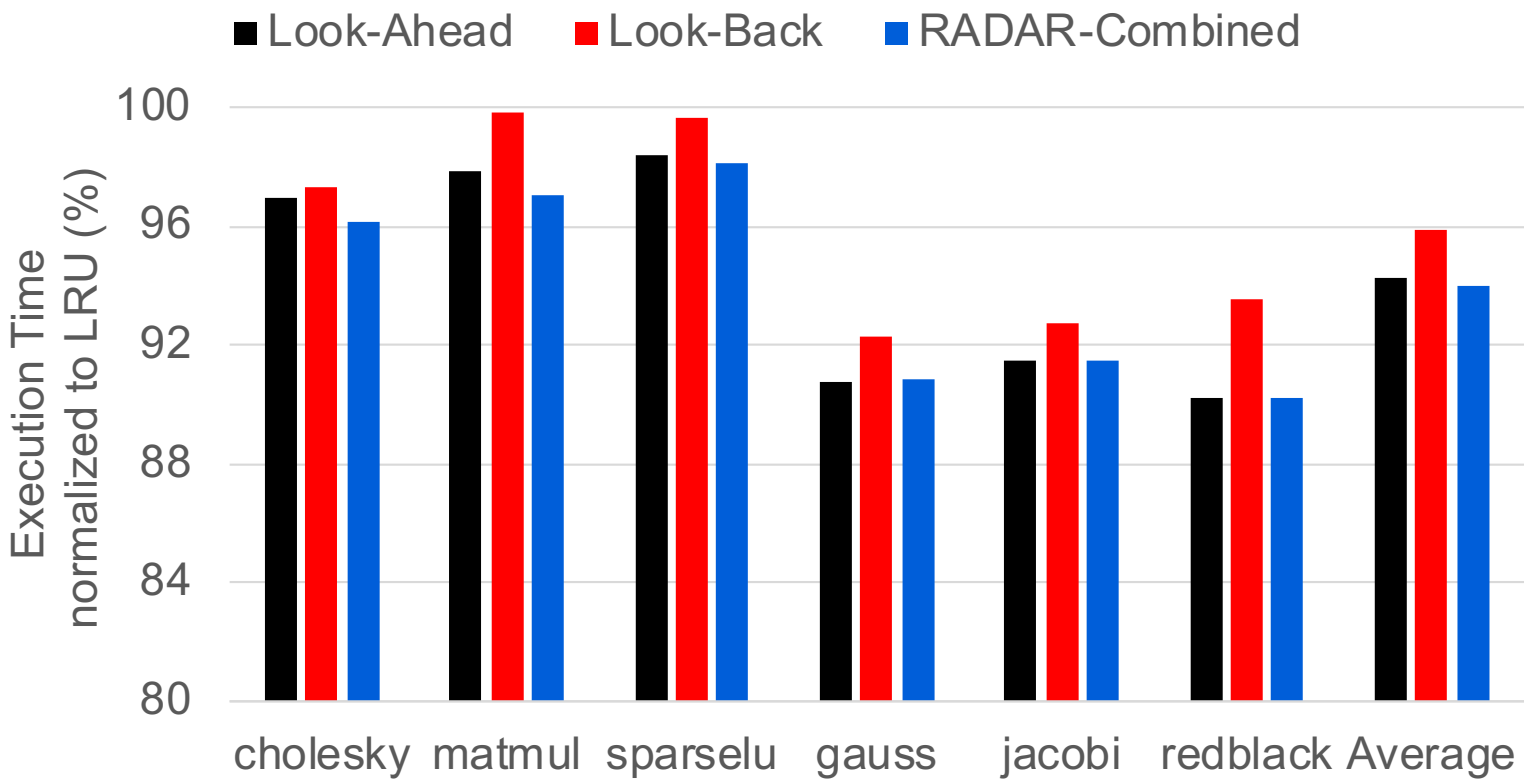
Is Region dead after current access?

Evict (demote to LRU) region from LLC

Three schemes:

- **Look Ahead** (LA): Use data-dependce graph for prediction
- **Look Back** (LB): Use past region access statistics
- **Combined**: LA ∩ LB and LA∪LB

Exec. Time Improvements



Memory bound apps provide significant gains

Outline

Background

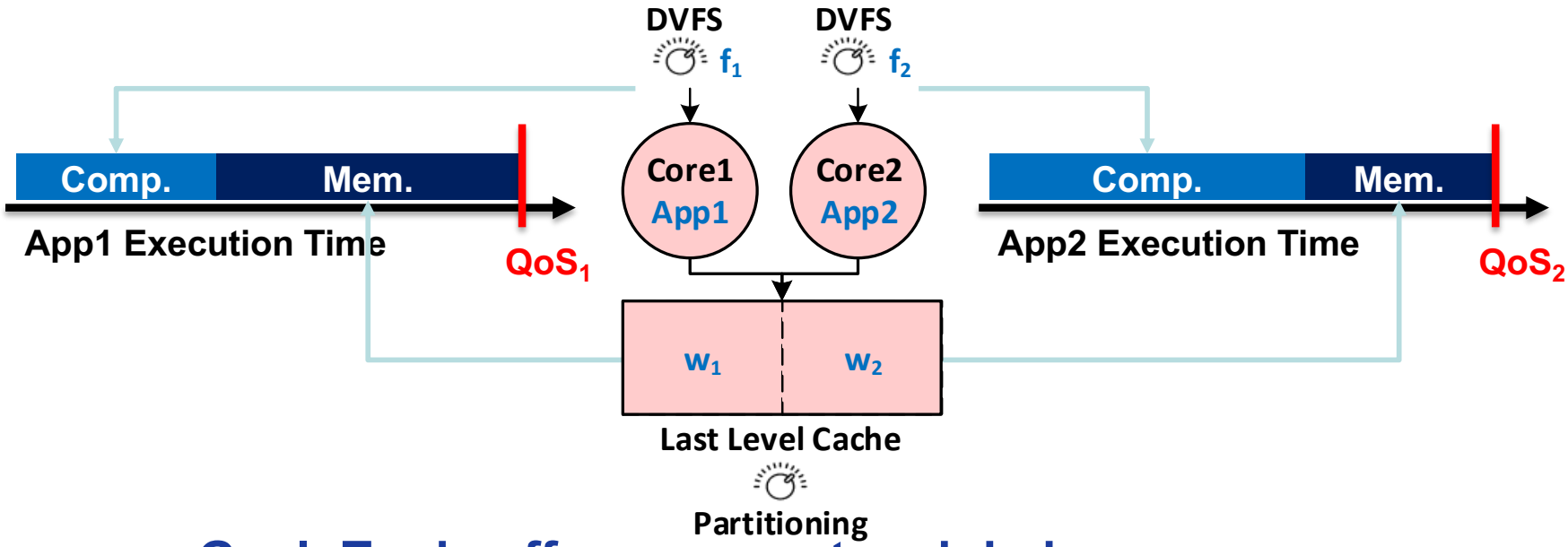
Runtime-Assisted Cache Management

Runtime-Assisted Power Management

Concluding Remarks

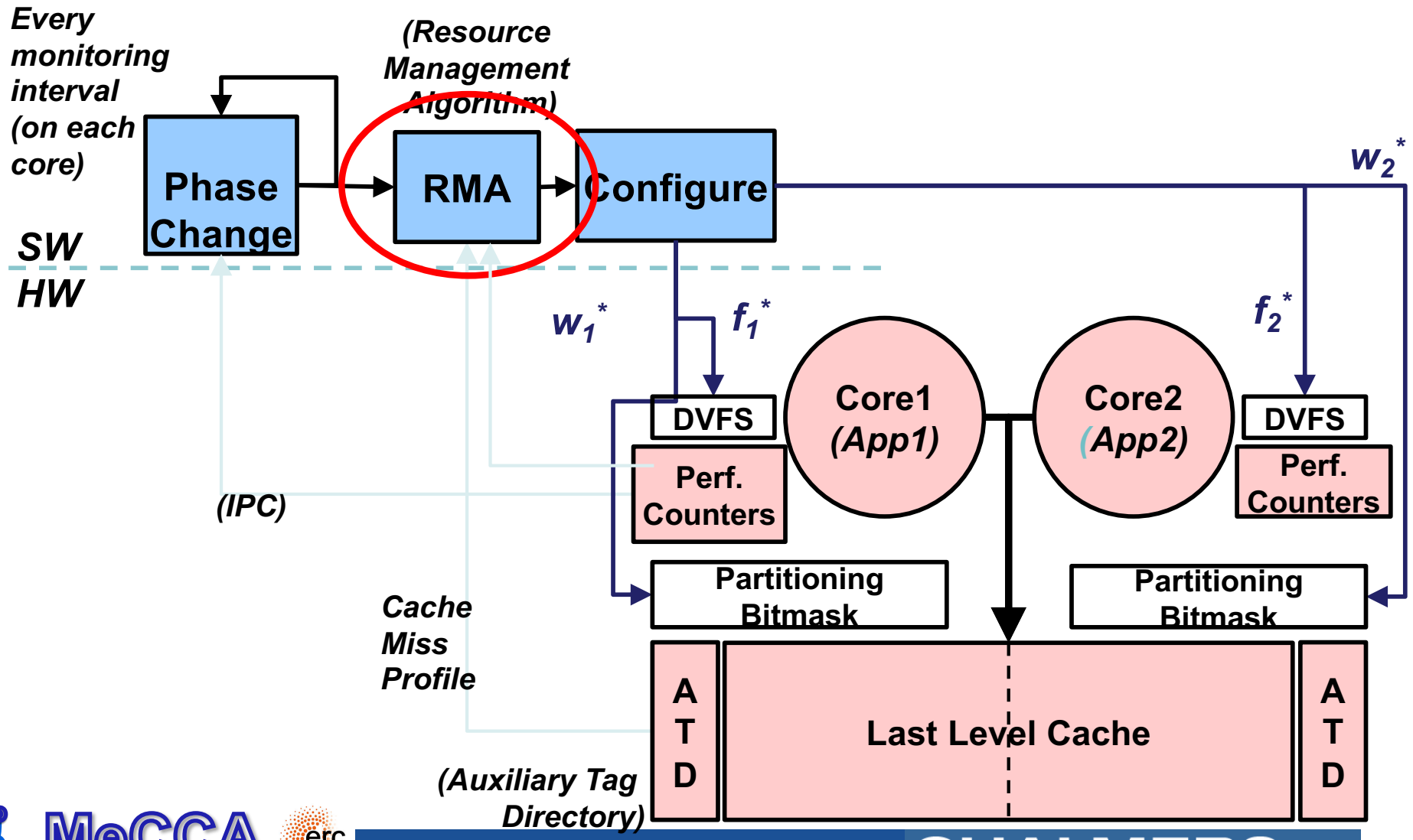
QoS-driven Resource Management – Background [IPDPS 2019]

By using QoS targets, we can throttle processor resources to minimize energy consumption

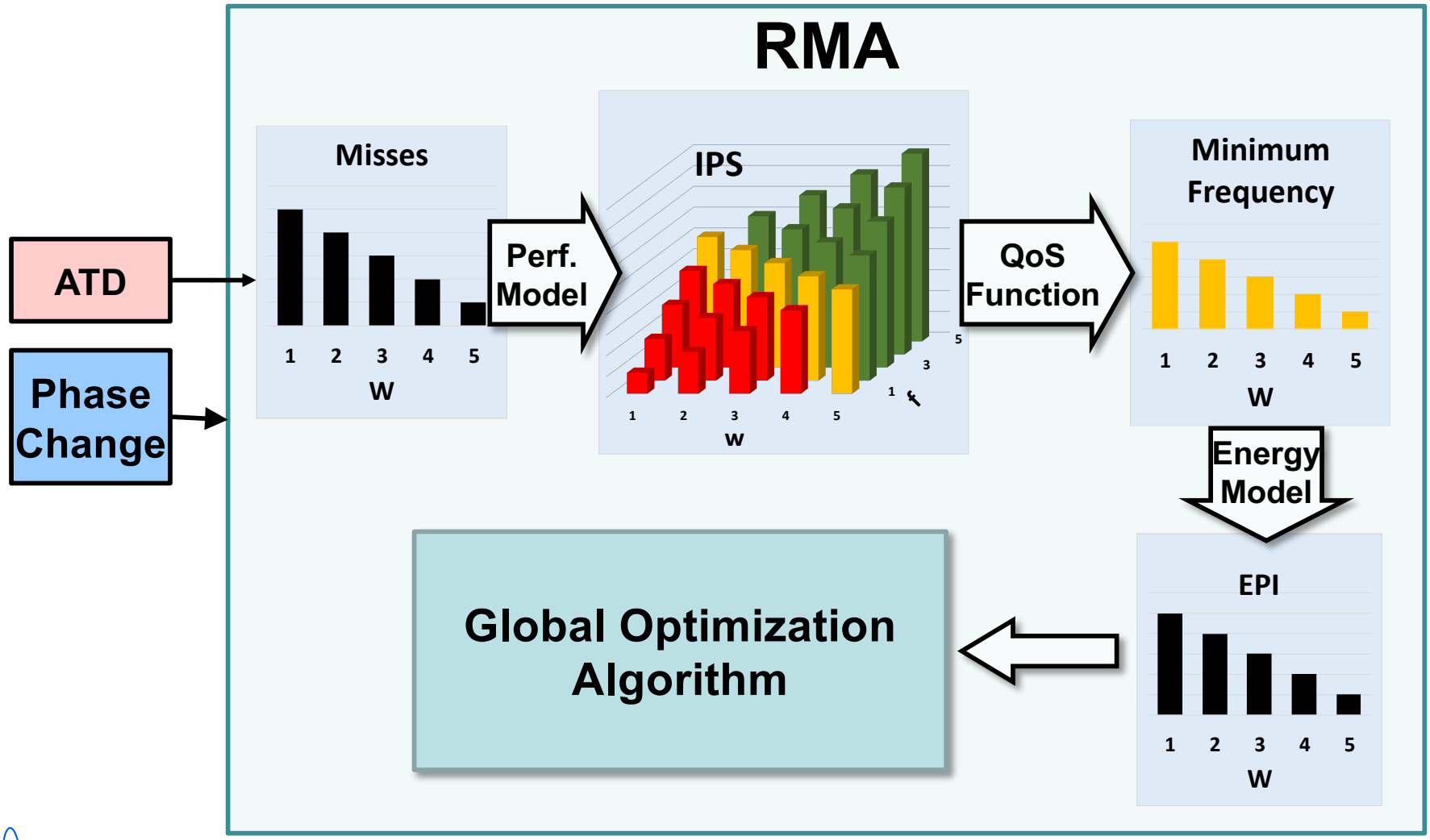


Goal: Trade off resources to minimize energy consumption while meeting QoS targets

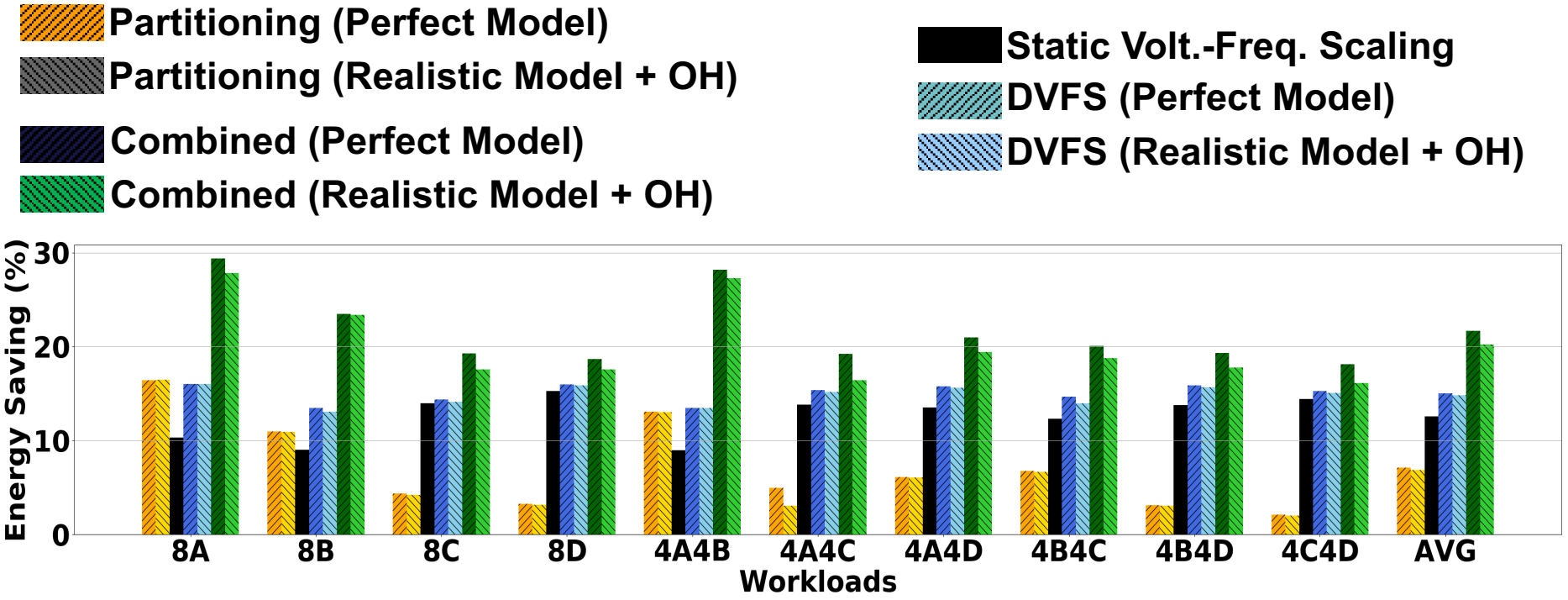
Overview


Every monitoring interval (on each core)
(Resource Management Algorithm)
Phase Change
RMA
Configure
SW
HW
w_2^*
w_1^*
f_1^*
f_2^*
Core1 (App1)
Core2 (App2)
DVFS
Perf. Counters
DVFS
Perf. Counters
(IPC)
Cache Miss Profile
Partitioning Bitmask
Partitioning Bitmask
ATD
Last Level Cache
ATD
(Auxiliary Tag Directory)

Searching the Configuration Space



Energy Saving Results (Relaxed QoS)



Save up to 28% of energy with 30% reduced IPS target (AVG: 20%)

Outline

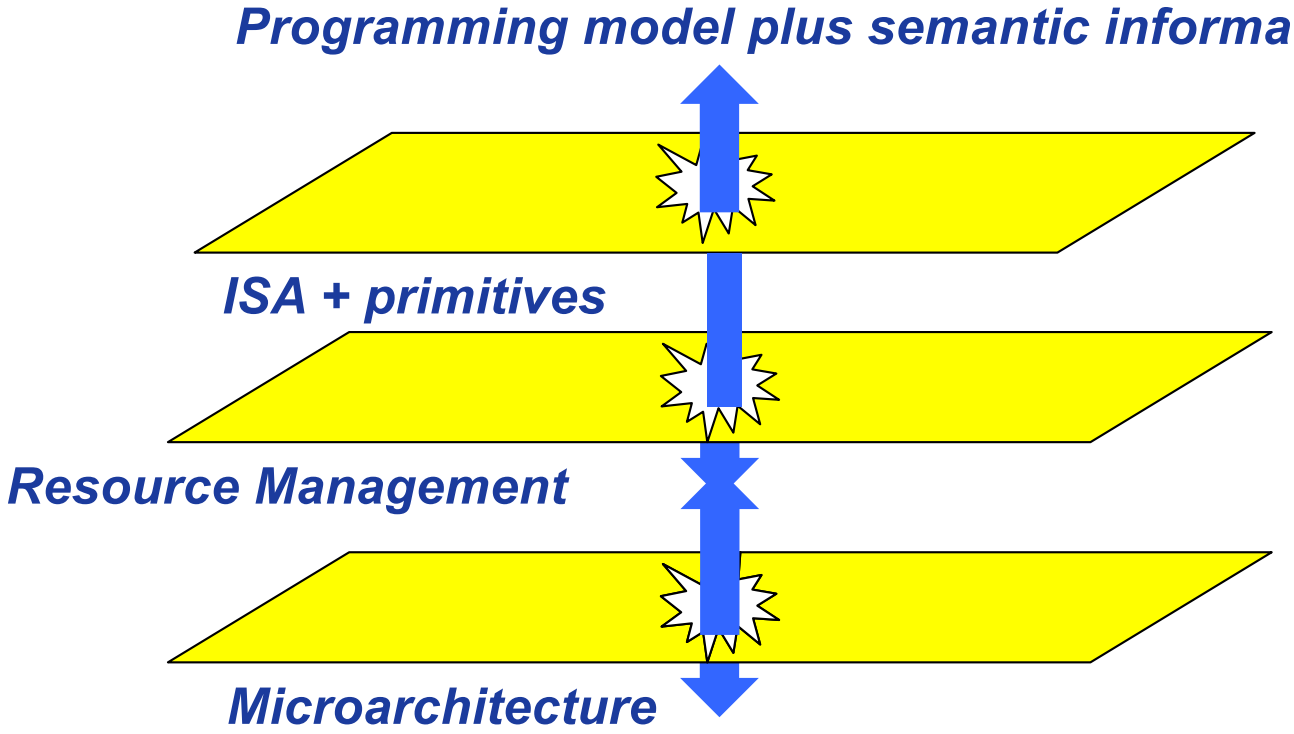
Background

Runtime-Assisted Cache Management

Runtime-Assisted Power Management

Concluding Remarks

Concluding Remarks



MECCA team members: Alexandra Angerd, Mohammad Waqar Azhar, Nadja Holtryd, Madhavan Manivannan, Mehrzad Nejat, Vasileios Papaefstathiou, Risat Pathan, Miquel Pericàs and Petros Voudouris