Energy and thermal management in MPSOCs

Prof. Tajana Šimunić Rosing
Dept. of Computer Science

System Energy Efficiency Lab
see
seelab.ucsd.edu

QUALCOMM
CISCO
YAHOO!
National Institutes of Health
The Nation's Medical Research Agency
CNS
it²
hp
invent
Texas Instruments Inc.
Sun
microsystems
DARPA
U.S. DEPARTMENT OF ENERGY
Los Alamos
NATIONAL LABORATORY
EST. 1943
NSF
SRC
Semiconductor Research Corporation

Future of IT

- Energy consumption is a critical issue:
  - Wireless systems: maximize battery life, optimize energy harvesting
  - High performance systems: minimize operational costs

Reference: J. Rabaey, “A Brand New Wireless Day,” Keynote Presentation, ASPDAC Jan. 08

What are we doing about it?

NSF Projects GreenLight & FlashGordon

- Green cyber-infrastructure in energy-efficient mobile facilities
- Closed-loop power and thermal management

Dynamic power management (DPM)

- Optimal DPM for a given class of workloads
- Machine learning to adapt
  - Select among specialized policies
  - Use sensors and performance counters to monitor
  - Multitasking/within task adaptation of voltage and frequency

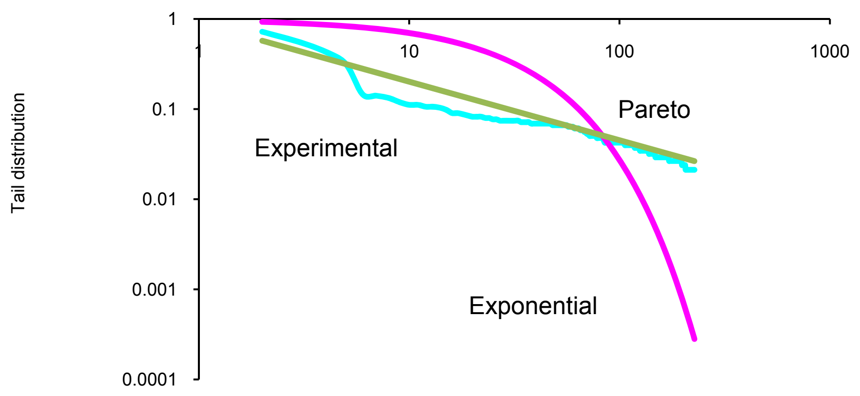
Dynamic thermal management (DTM)

- Workload scheduling:
  - Power vs. thermal management
  - Runtime adaptation to get best temporal and spatial profiles using closed-loop sensing
  - Negligible performance overhead
- Machine learning for dynamic adaptation
- Proactive thermal management

it²　SDSC SAN DIEGO SUPERCOMPUTER CENTER　Sun microsystems　Texas Instruments Inc.　YAHOO!　hp invent　CISCO　DARPA　NSF　SRC Semiconductor Research Corporation

DPM: Workloads - Idle State

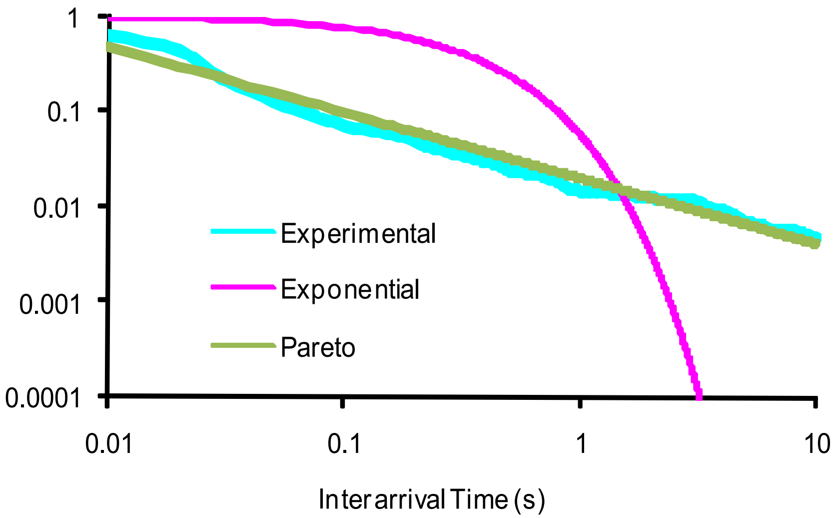
Hard Disk Trace



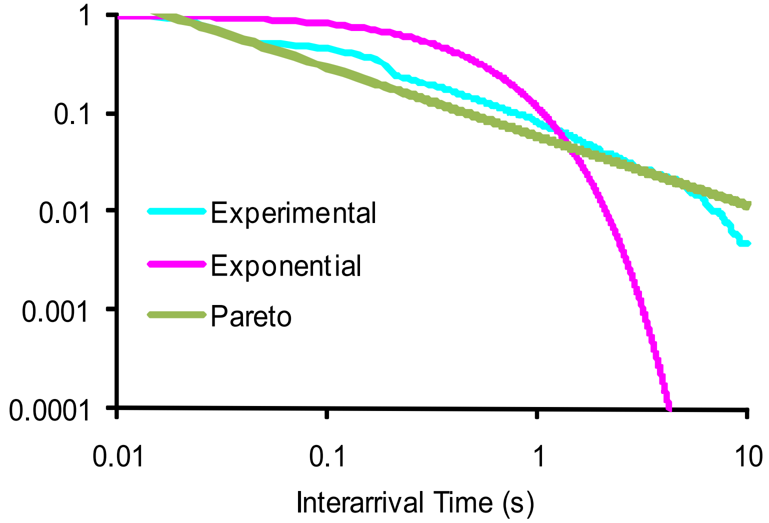
Pareto Distribution:

$$E_{user} = 1 - a \cdot t^{-b}$$

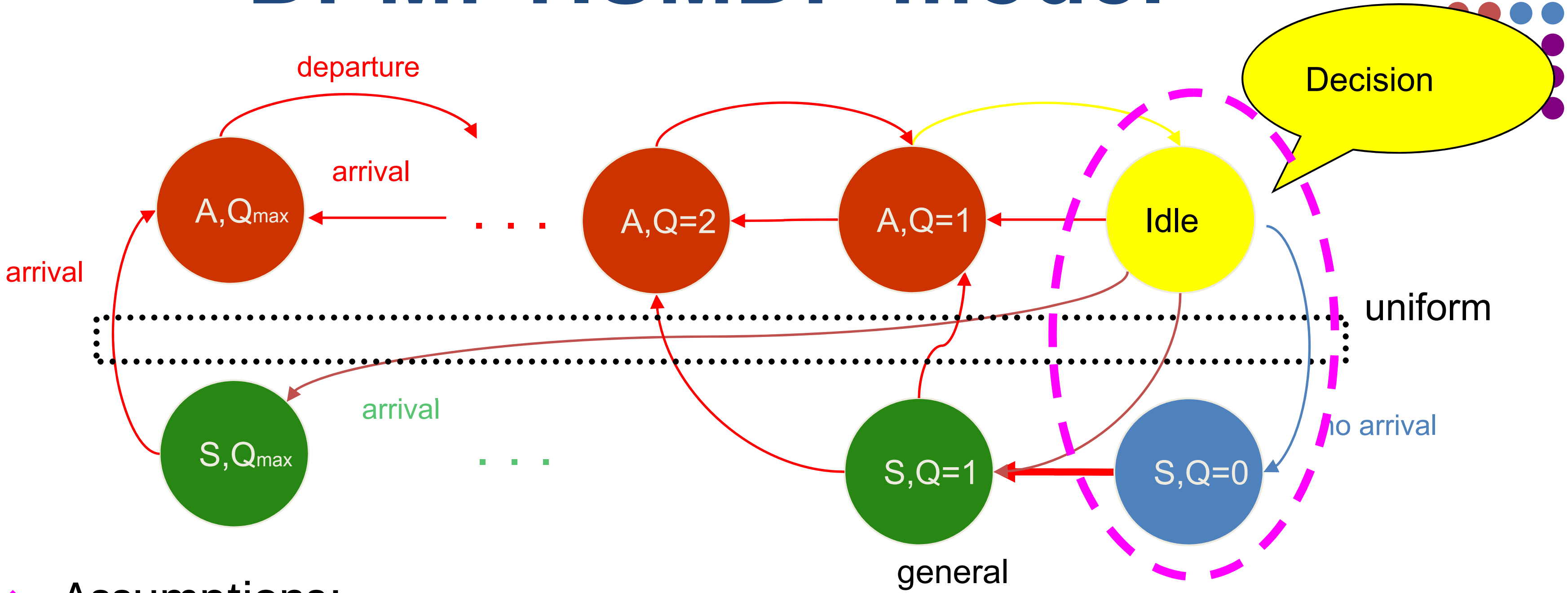
WWW Trace



Telnet Trace



DPM: TISMDP model



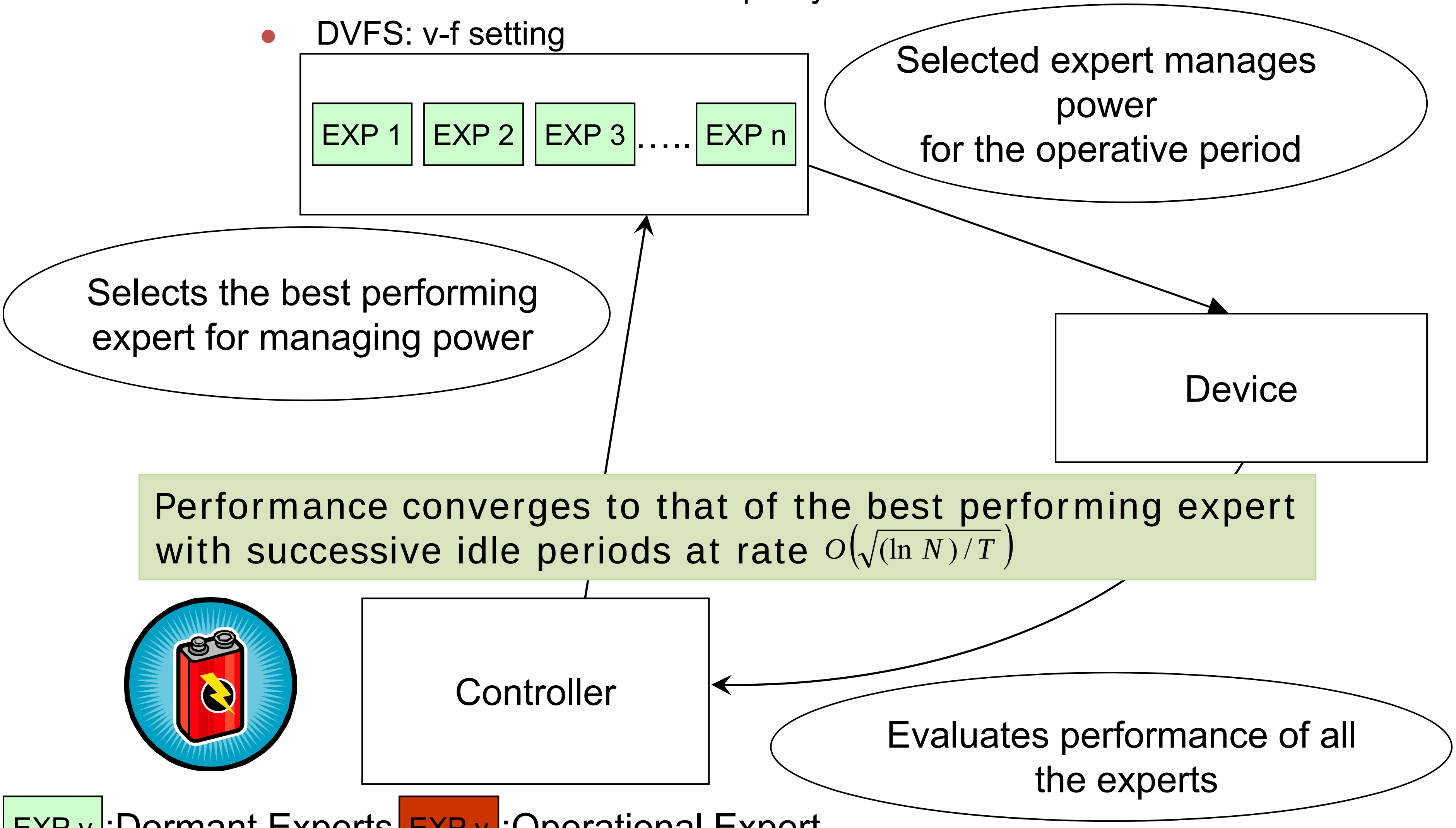
- Assumptions:
  - general distribution governs the first request arrival
  - exponential distribution represents arrivals after the first arrival
  - user, device and queue are stationary

Obtain globally optimal policy using linear programming

Measurements on hard disk within 11% of ideal oracle policy
factor of 2.4 lower than always-on
factor of 1.7 lower than default time-out

Online Learning for Power Management

- Experts:
  - DPM: A state of the art DPM policy
  - DVFS: v-f setting



EXP 1 EXP 2 EXP 3 ….. EXP n

Selected expert manages power for the operative period

Selects the best performing expert for managing power

Device

Performance converges to that of the best performing expert with successive idle periods at rate $O\left(\sqrt{(\ln N)/T}\right)$

Controller

Evaluates performance of all the experts

EXP y :Dormant Experts EXP y :Operational Expert

Policies used in experiments

- Hard disk drive

Expert	Characteristics
Fixed Timeout	Timeout = $7*T_{be}$
Adaptive Timeout	Initial timeout = $7*T_{be}$; Adjustment = $+0.1T_{be}/-0.1T_{be}$
Exponential Predictive	$I_{n+1} = a\, i_n + (1 - a).I_n$, with a = 0.5
TISMDP	Optimized for delay constraint of 3.5% on HP-1 trace

Trace Name	Duration (in sec)	$\overline{t}_{RI}$	$\sigma_{t_{RI}}$
HP-1Trace	32311	20.5	29
HP-2 Trace	35375	5.9	8.4
HP-3 Trace	29994	17.2	2

$\overline{t}_{RI}$: Average Request Inter-arrival Time (in sec)

- CPU: Xscale
- Workloads:
  - qsort, djpeg, blowfish, dgzip

Freq (MHz)	Voltage (V)
208	1.2
312	1.3
416	1.4
520	1.5

Measurements on HDD

With Individual Experts

Policy	HP1 Trace		HP2 Trace		HP3 Trace	
	%delay	%energy	%delay	%energy	%delay	%energy
Oracle	0	68.17	0	65.9	0	71.2
Timeout	4.2	49.9	4.4	46.9	3.3	55
Ad Timeout	7.7	66.3	8.7	64.7	6	67.7
TISMDP	3.4	44.8	2.26	36.7	1.8	42.3
Predictive	8	66.6	9.2	65.2	6.5	68

Converges to Predictive

With Controller

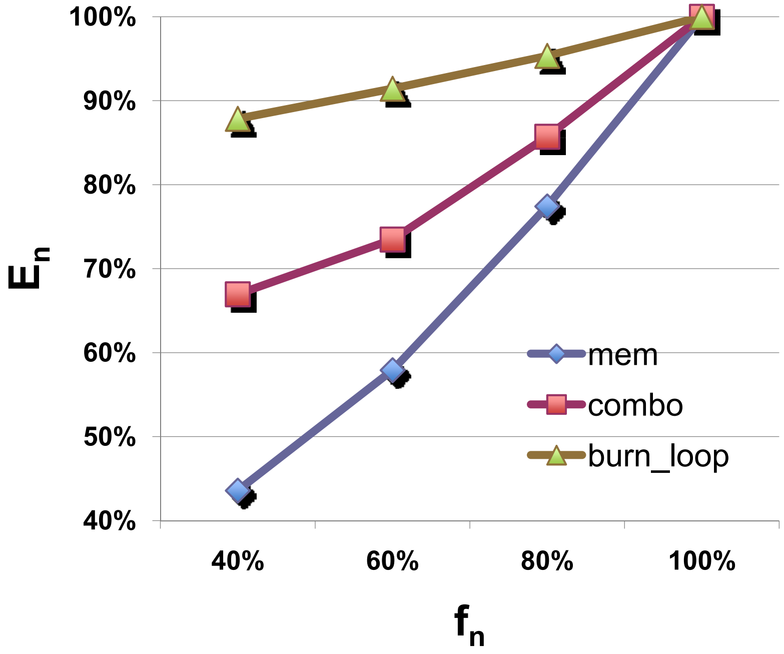
Maximum Energy Savings

Least Delay

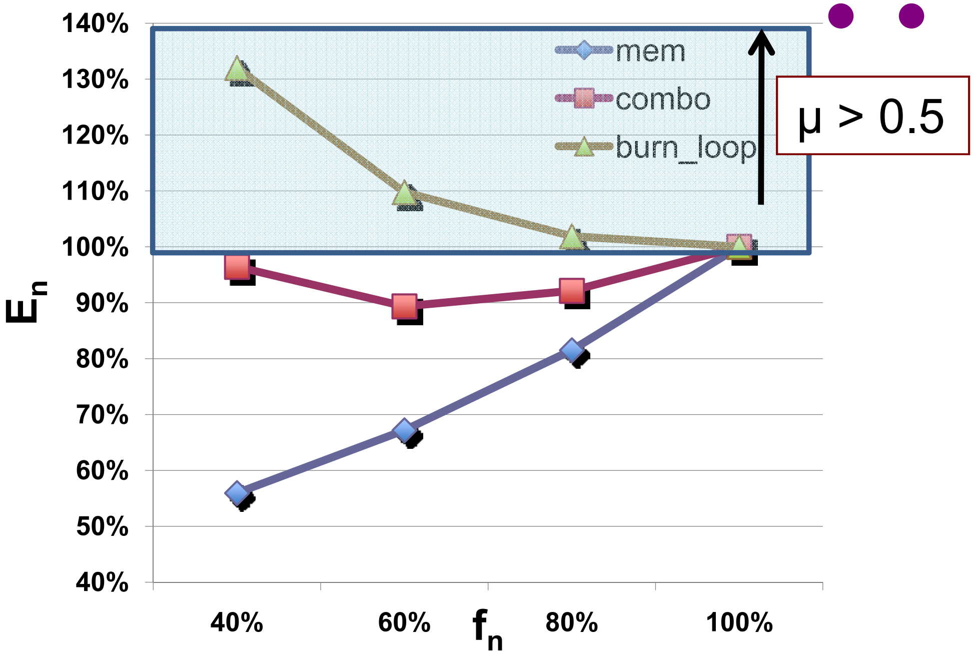
Preference	HP-1 Trace		HP-2 Trace		HP-3 Trace	
	%delay	%energy	%delay	%energy	%delay	%energy
Low delay	3.5	45	2.61	37.41	2.55	49.5
↓	6.13	60.64	5.86	54.2	4.36	61.02
High energy savings	7.68	65.5	8.59	64.1	5.69	66.28

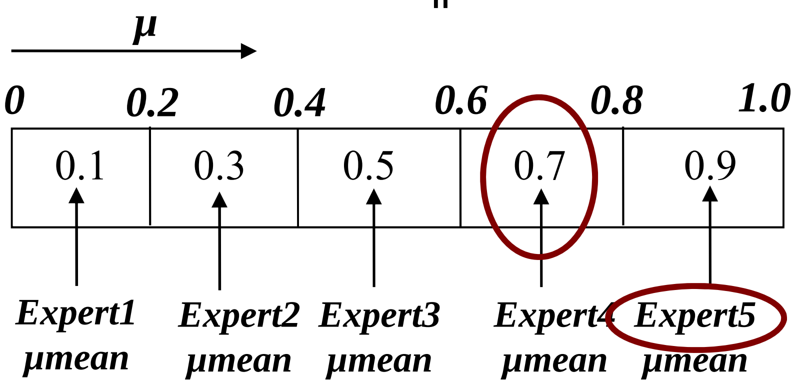
DVFS: Mem vs. CPU

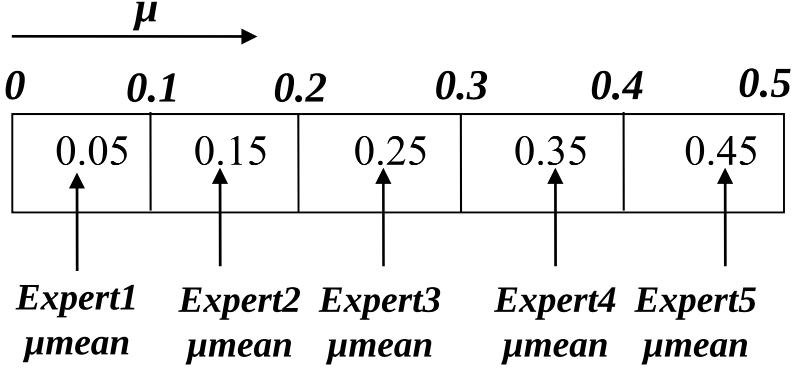
Static power ratio = 30%



Static power ratio = 50%







Energy Loss = (0.9 – 0.7) = 0.2
Performance Loss = 0

CPU: Higher utilization tasks

- Single task: within 7% of the max possible energy savings

Bench.	Low perf delay -------> Higher energy savings					
	%delay	%energy	%delay	%energy	%delay	%energy
qsort	6	17	16	32	25	41
djpeg	7	21	15	37	26	45
dgzip	15	30	21	42	27	49
bf	6	11	16	27	25	40

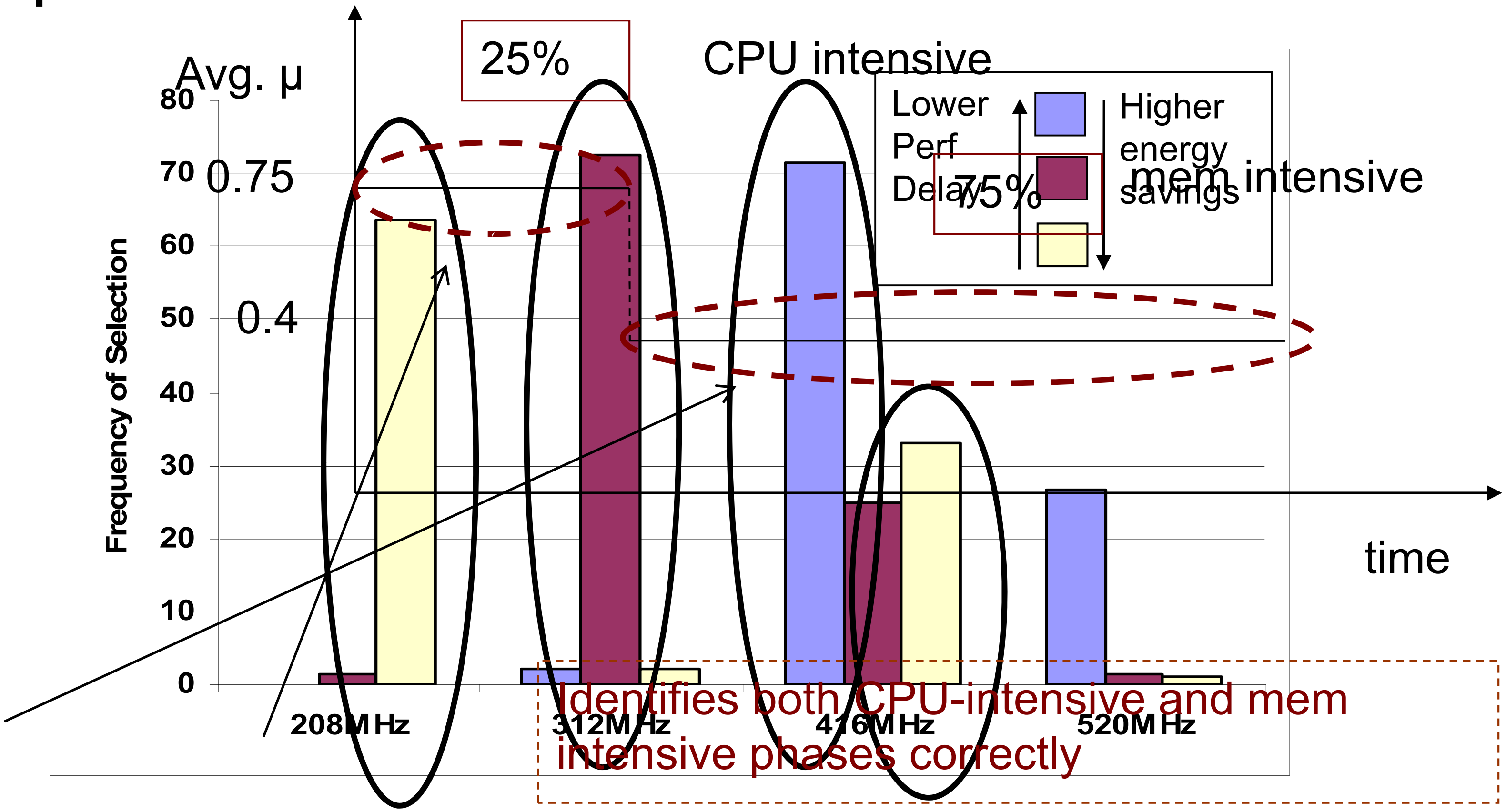
Bench.	208MHz/1.2V	
	%delay	%energy
qsort	56	48
djpeg	34	54
dgzip	33	54
bf	40	51

- Multitasking environment: energy savings 20-50% maximum
  - energy savings are average of per thread savings (e.g. djpeg & dgzip)

Bench.	Low perf delay -------> Higher energy savings					
	%delay	%energy	%delay	%energy	%delay	%energy
qsort+djpeg	6	17	15	33	25	41
djpeg+dgzip	13	24	19	39	27	48
qsort+djpeg	7	20	18	35	26	42
dgzip+bf	13	18	22	32	27	44

CPU: Frequency of Selection

For qsort



Identifies both CPU-intensive and mem intensive phases correctly

Performance vs. Energy

- Assume a simple static-DVFS policy
  - AMD Opteron (four v-f settings):
    - 1.25V/2.6GHz, 1.15V/1.9GHz, 1.05V/1.4GHz, 0.9V/0.8GHz
- Compare against a base system with no DVFS and three simple idle PM policies:

Policy	Description
PM-1	switch CPU to ACPI state C1 (remove clock supply) and move to lowest voltage setting
PM-2	switch CPU to ACPI state C6 (remove power)
PM-3	switch CPU to ACPI state C6 and switch the memory to self- refresh mode

$$\%E_{savings_{PM-i}} = \frac{E_{DVFS_f}}{E_{PM-i}}$$

Results

Benchmark	**Freq**	**% delay**	**% Energy$_{savingsPM-i}$**		
			PM-1	**PM-2**	**PM-3**
mcf	1.9	29	5.2	0.7	-0.5
	1.4	63	8.1	0.1	-2.1
	0.8	163	8.1	-6.3	-10.7
bzip2	1.9	37	4.7	-0.6	-2.1
	1.4	86	7.4	-2.4	-5
	0.8	223	7.8	-9.0	-14
art	1.9	32	6	1	-0.1
	1.4	76	7.3	-1.7	-4
	0.8	202	8	-8	-13
sixtrack	1.9	37	5	-0.5	-2
	1.4	86	6	-4.3	-7.2
	0.8	227	7	-11	-16.1

Key points

- Simple power management policies provide better energy performance tradeoffs
- Lower v-f setting offer worse e/p tradeoffs due to high performance delay
- DVFS still useful for:
  - Peak power reduction
  - Thermal management
  - Systems with simpler memory controllers and low power system components

Evaluating Thermal Management Policies

- Combination of temperature characteristics and performance:
  - Hot Spots: % time spent above threshold
  - Thermal Cycles: % time cycles above ΔT_{cyc} are observed
  - Spatial Gradients: % time gradients above ΔT_{spat} are observed across the die
  - Performance: Load average

    (sum of run queue length and number of jobs currently running)

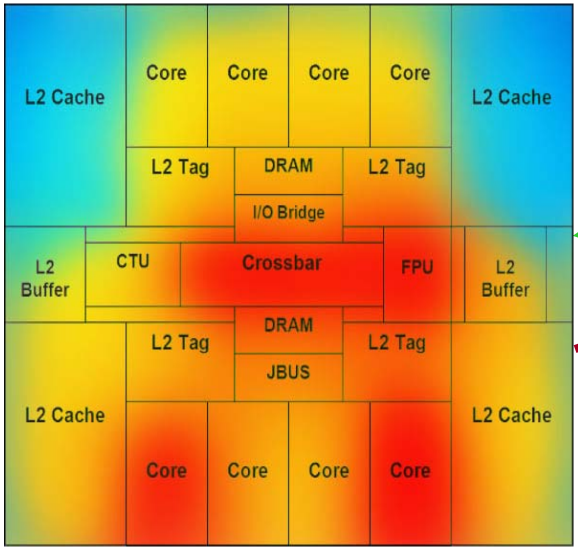
DTM: Evaluation Framework

Inputs:
- Workload – collected at a data center
- Floorplan, temperature (for dynamic policies)

Resource manager

Static: Fixed allocation (ILP)

Dynamic: Dependent on the policy



Power Manager

DPM, DVS

Inputs:
- Power trace for each unit
- Floorplan, package and die properties (Niagara-1)

Thermal Simulator

HotSpot [Skadron, ISCA'03]

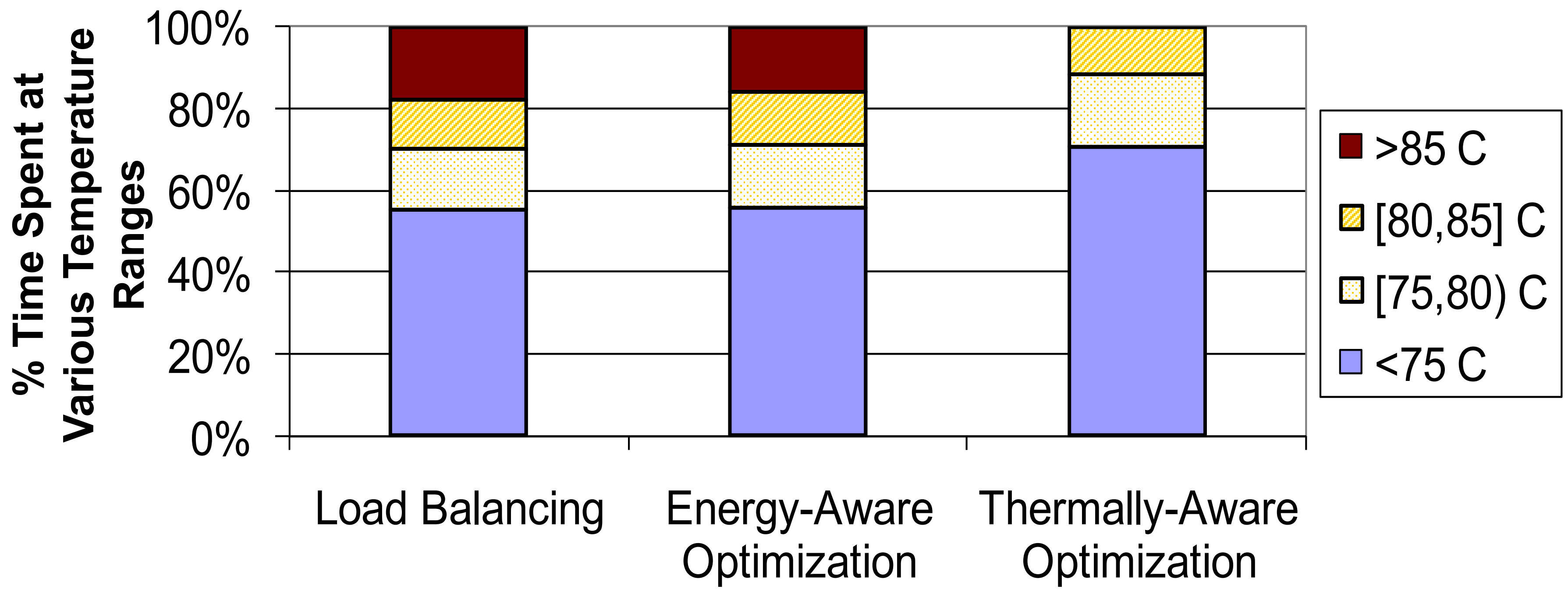
Transient Temp. Response for Each Unit

DTM: Policies compared

- Optimal and static:
  - ILP-energy
    - minimizes the overall energy consumption
  - ILP-comb
    - minimizes the thermal hot spots and the temperature gradients
- Dynamic:
  - Load balancing
    - Balances threads for performance only
  - Adaptive-Random Policy
    - Minimizes & balance temperature with low scheduling complexity
    - *Probability* of sending a workload to a core based on temperature history
    - Adapts to changes in temperature dynamics
  - DVFS, DPM, Thread migration
- Online learning (OL)
  - Various specialist/expert combinations

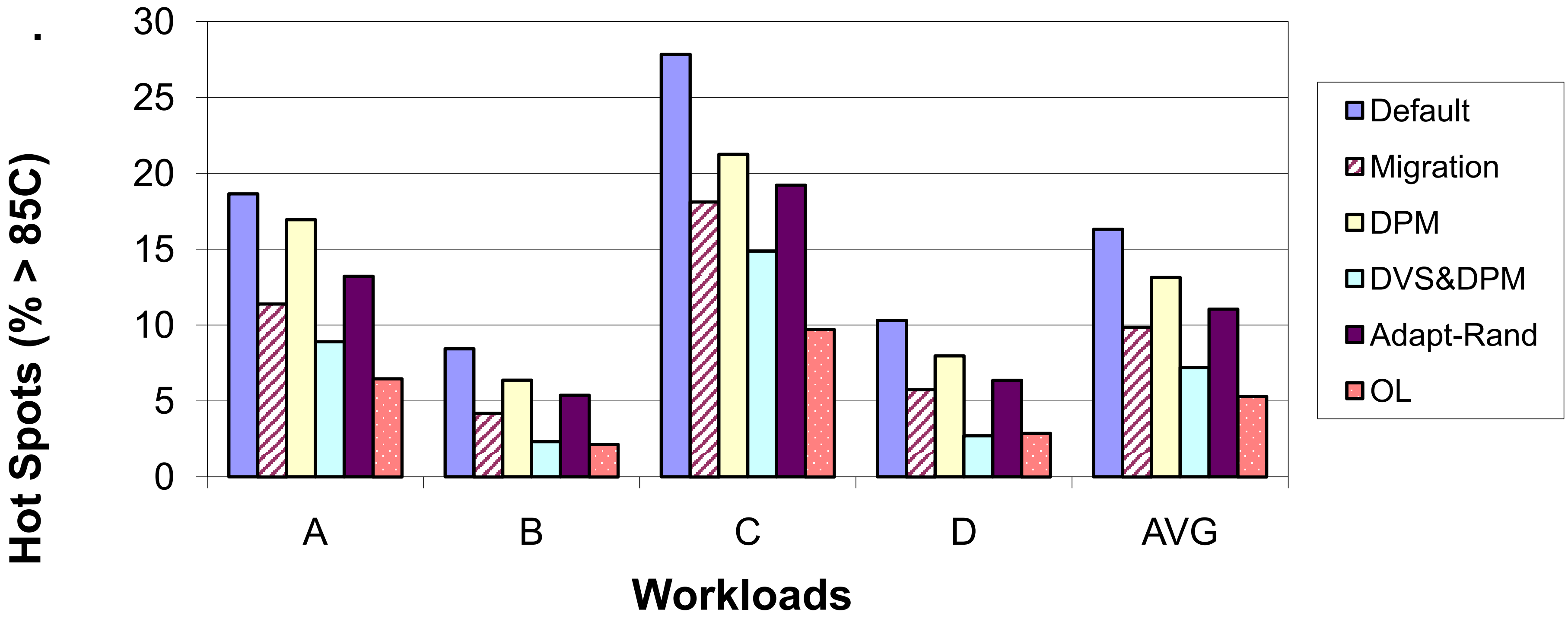
Load balancing vs. optimal policies

- Energy or performance-aware methods are not always sufficient to manage temperature.



Dynamic Policies: Thermal Hot Spots

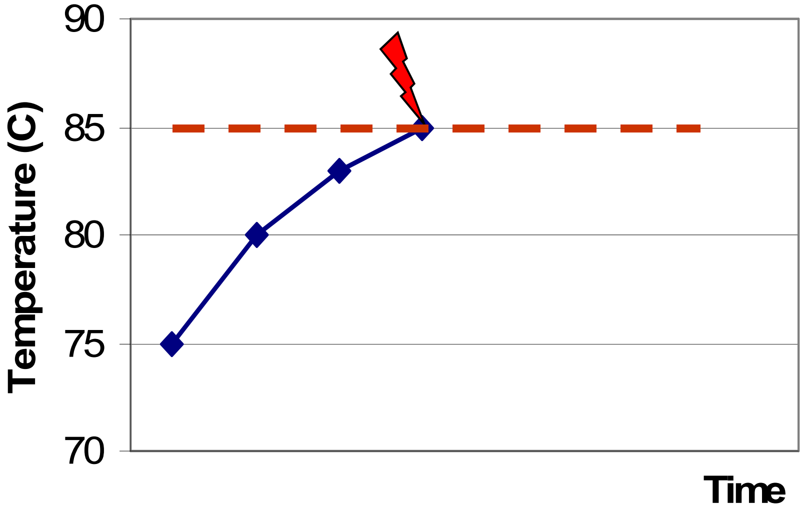
- Workloads collected at an operational datacenter over a period of a week; concatenated 1hr of each day to show adaptation



Online learning gives 20% hot spot reduction in average in comparison to the best policy

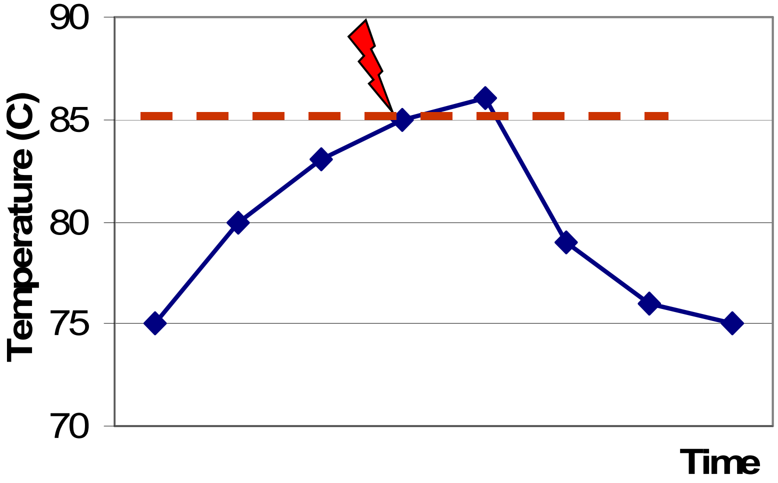
Reactive vs. Proactive Management

- **Reactive**

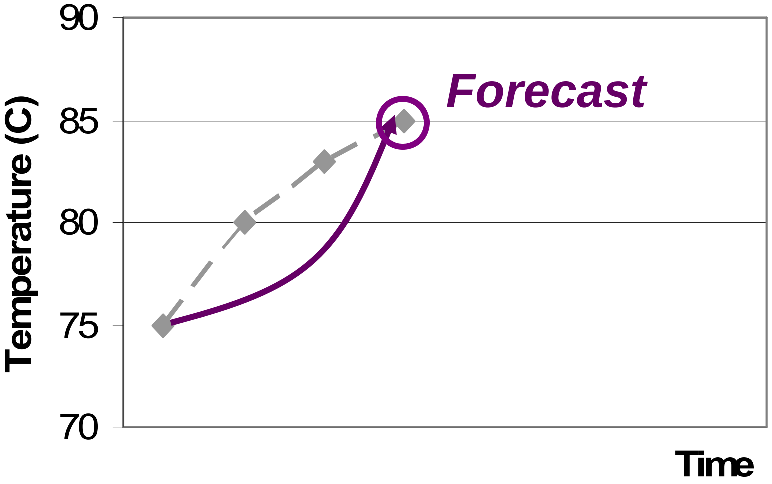


Reactive vs. Proactive Management

- **Reactive**



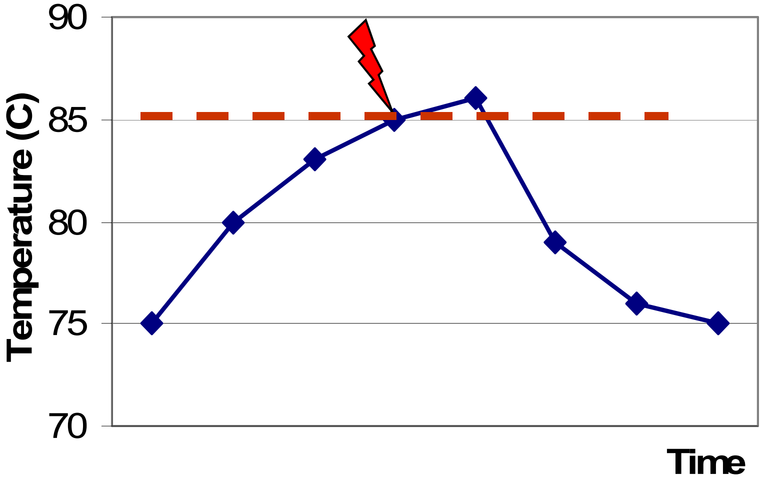
- **Proactive**



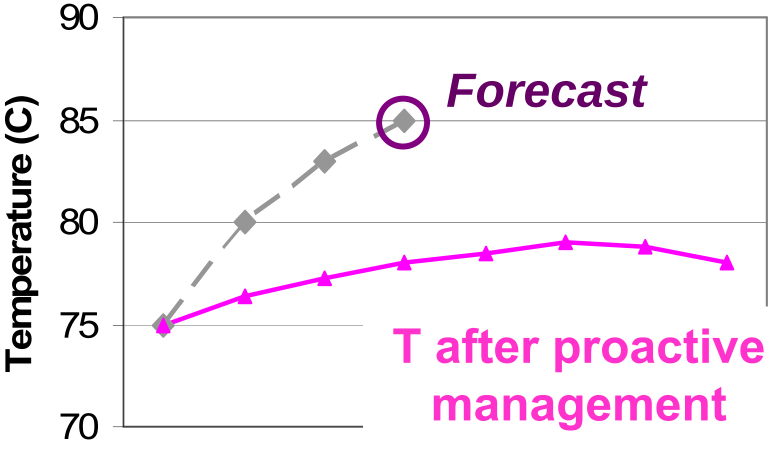
- e.g., DVFS,
  fetch-gating,
  workload migration,
  …

Reactive vs. Proactive Management

- **Reactive**



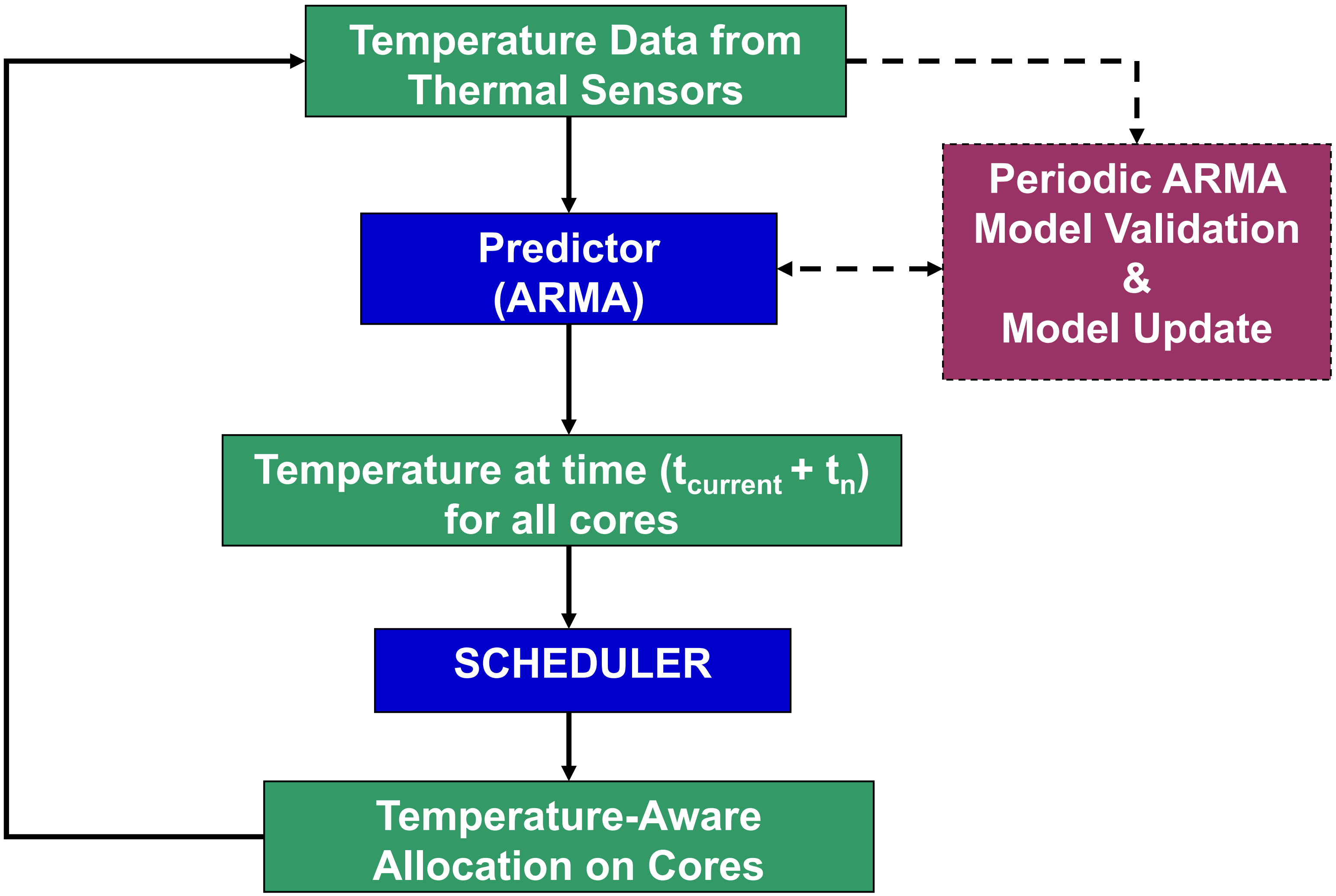
- **Proactive**



- e.g., DVFS,
  fetch-gating,
  workload migration,
  …

- Reduce and balance temperature
  - Adjust workload, V/f setting, etc.

Proactive management flow



Temperature Data from Thermal Sensors

Predictor (ARMA)

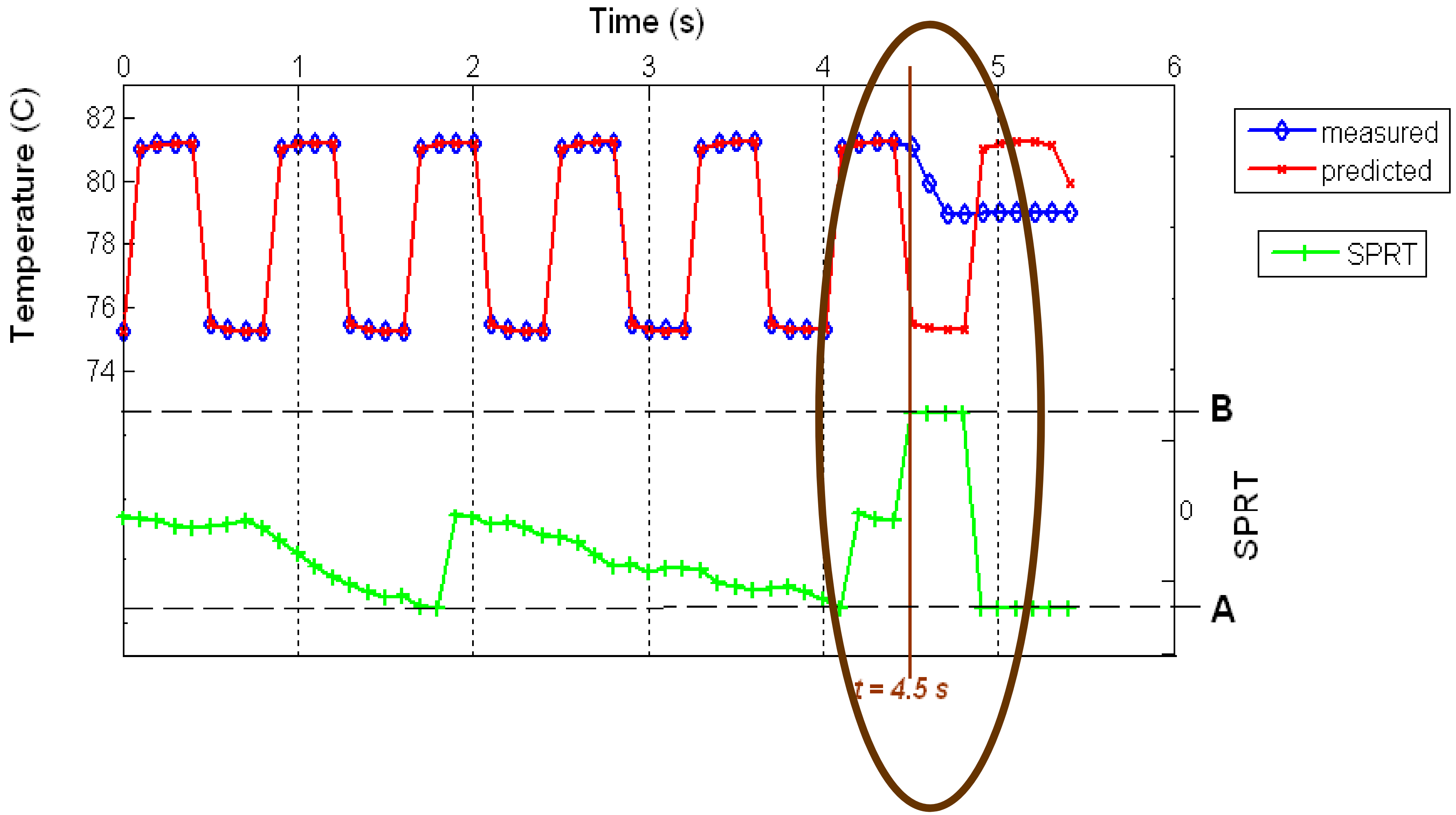
Periodic ARMA Model Validation & Model Update

Temperature at time ($t_{current} + t_n$) for all cores

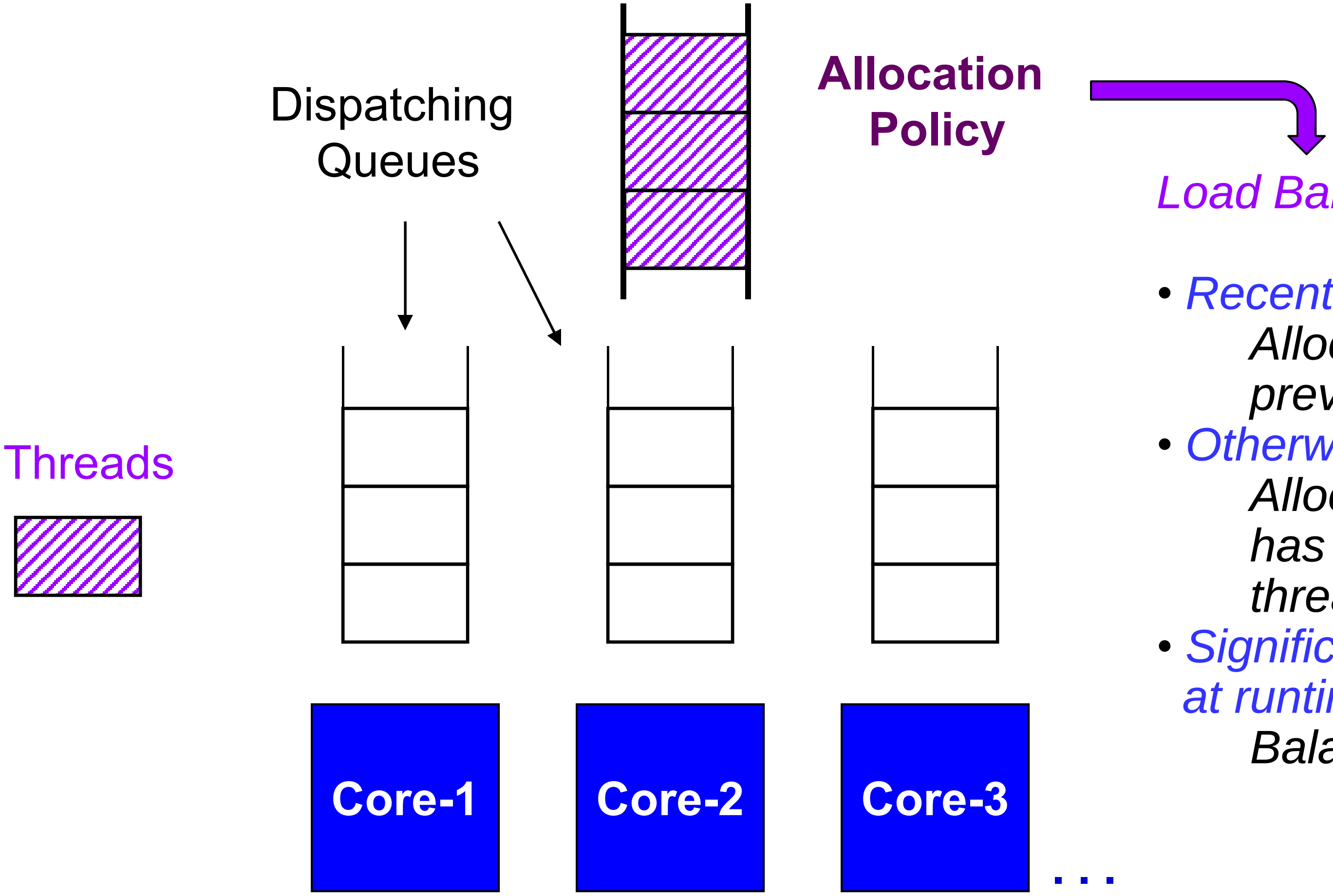
SCHEDULER

Temperature-Aware Allocation on Cores

Detection with SPRT



System Model



Dispatching Queues

Allocation Policy

Threads

Core-1

Core-2

Core-3

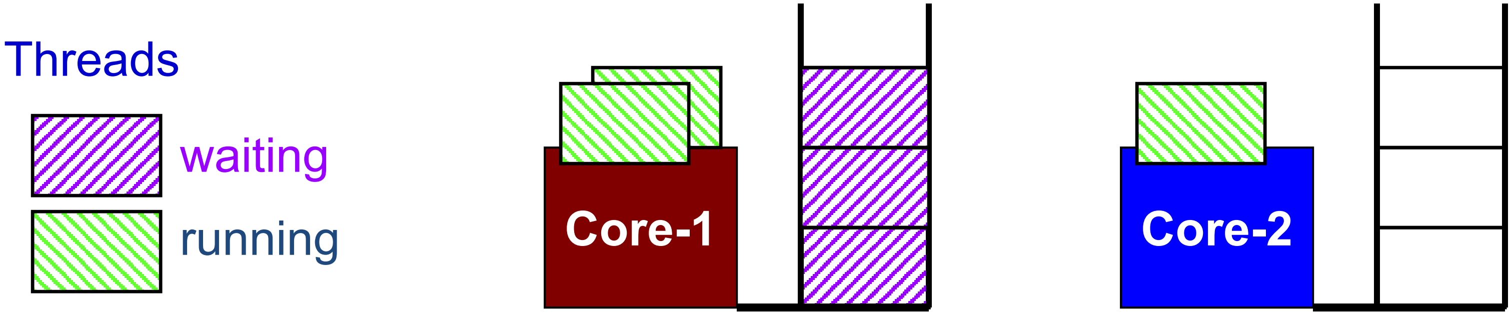
. . .

Load Balancing:

- *Recently run thread:*
  Allocate to the core it ran previously on
- *Otherwise*
  Allocate to the core that has the lowest priority thread
- *Significant imbalance at runtime*
  Balance

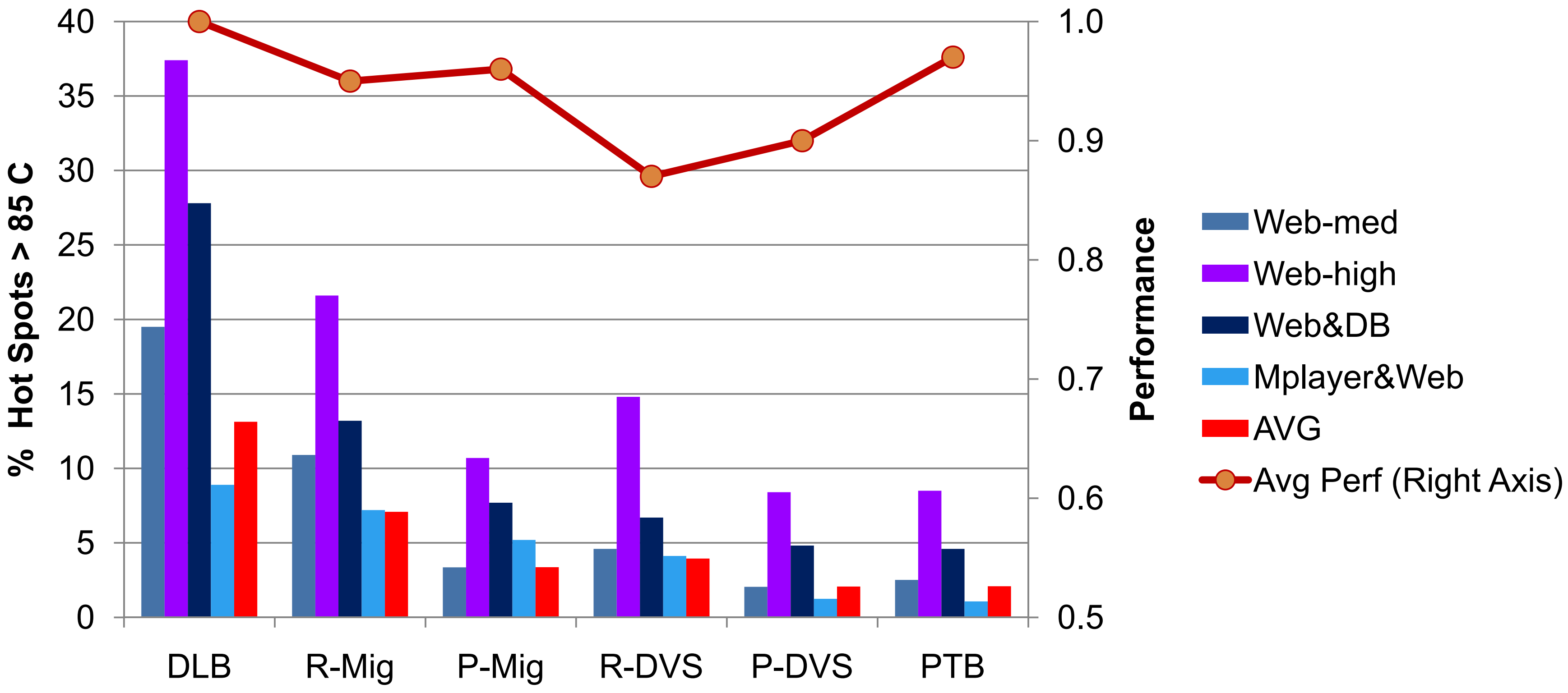
Proactive Temperature Balancing

- Uses principle of locality as in default load balancing policy at initial assignment
- Utilizes ARMA predictor & thermal forecast:
  - A core is projected to have a hot spot **OR**
  - $\Delta T_{spatial}$ is projected to be large
  - → ***Migrate threads to balance temperature***
  - → ***Move "waiting" threads***



Proactive vs. Reactive: Hot Spots

- Proactive Balancing (PTB) achieves similar hot spot reduction with P-DVS while improving performance by ~8%
- PTB reduces hot spots 60% over reactive migration



Power and Thermal Management

- Power management can achieve large energy savings by exploiting variations in workload
    - TISMDP DPM/DVS policy optimized for stationary workloads
        - Implementable in hardware
    - Machine learning to optimally select among individual DPM/DVS policies
- Minimizing power consumption does not always lead to optimal thermal profiles both in terms of hot spots and temperature gradients
- Thermal management:
    - Very low overhead policies minimize hot spots and thermal gradients
    - Online learning performs significantly better than any individual policy
    - Proactive thermal management further reduces hotspots by 60% with practically no overhead

Power management

MODELING & ANALYSIS

- Novel memory systems – DRAM & PCM
- Cycle-accurate simulation of energy consumed by CPU, memory hierarchy, interconnect, power conversion system and battery
- Energy software profiler
- Software optimization to minimize the energy consumption using complex instruction mapping
    - speech recognition, multimedia

- DAC’09
- IEEE D&T ‘04
- DSD’04, GLVLSI’04
- IEEE TCAD’03
- DAC’02, DATE’02
- ICASSP ’02
- IEEE TVLSI’01
- DATE’00, DAC’99
- ISLPED’99, ISSS’00
- CODES’99

STOCHASTIC POLICIES

- Statistical models of workload and devices in computing systems
- Optimal power management algorithms using Time-Indexed Semi-Markov decision processes

- Book: “The best papers in 10 years of DATE”, ‘07
- ESTIMedia’03, DATE’02
- IEEE TCAD’01
- DAC’01, MOBICOM’00
- DATE’00, ISLPED ’00
- ISSS’99

ONLINE LEARNING

- Adaptively selects among a set of policies

- IEEE TCAD’09
- USENIX-HotPower’08
- ISLPED’07, ICCAD’06

Thermal management

MODELING AND ANALYSIS

- Fine-grained reliability modeling of multicore systems
- Fast architecture-level simulation framework
- Large scale modeling of system reliability and power
- Modeling and analysis methodologies for 3D circuits
- Thermal estimation based on a limited set of sensors
- Sensor placement for accurate thermal measurement

- SIGMETRICS’09
- DATE’09
- GLSVLSI’08
- ISQED’08
- IEEE TVLSI’07
- Journal of LPE’06
- GLSVLSI’06
- PATMOS’05, DSD’04

TEMPERATURE-AWARE SCHEDULING

- Optimal scheduling solution for known workloads
- Extremely light-weight dynamic OS-level job scheduler
- Scheduling in 3D coupled with liquid cooling
- Online learning for selecting the best fit policy

- DATE’09
- IEEE TVLSI’08
- ASPDAC’08
- DAC’08
- DATE’07

PROACTIVE MANAGEMENT

- Highly accurate, fully dynamic temperature prediction
- Proactive job allocation to prevent thermal problems

- IEEE TCAD’09
- ICCAD’08
- ISLPED’08

Select recent publications

- **BOOK CHAPTERS & JOURNAL PAPERS**
  - E. Regini, D. Lim, T. Simunic Rosing, “Energy management in heterogeneous wireless sensor networks,” submitted to IEEE TMC, 2009.
  - G. Dhiman, T. Simunic Rosing, “Using online learning for system level power management,” IEEE TCAD 2009.
  - Y. Lu, E. Chung, T. Simunic, L. Benini, G. De Micheli: “Quantitative Comparison of Power Management Algorithms”, in The Most Influential Papers of 10 Years DATE, Edited by Lauwereins, Rudy; Madsen, Jan, Springer-Verilag, 2008.
  - A. Coskun, T. Simunic Rosing, K. Whisnant, K. Gross, “Static and dynamic temperature-aware scheduling for multiprocessor SoCs,” IEEE TVLSI 2008.
  - T. Simunic Rosing, K. Mihic, G. De Micheli, “Power and reliability management of SOCs,” IEEE Transactions on VLSI, April 2007.
  - G. Park, T. Simunic Rosing, M. Todd, C. Farrar, W. Hodgkiss, “Energy Harvesting for Structural Health Monitoring in Sensor Networks,” ASCE Journal, 2007.
  - J. Kim, T. Simunic Rosing, “Power-aware resource management techniques for low-power embedded systems,” in Handbook of Real-Time and Embedded Systems, Edited by S. H. Son, I. Lee, J. Y-T Leung, Taylor-Francis Group LLC, 2006.
  - A. Coskun, T. Simunic Rosing, K. Mihic, G. De Micheli, Y. Leblebici, “Analysis and Optimization of MPSoC Reliability,” Invited paper to Journal of Low-Power Electronics, April 2006.
- **CONFERENCE PAPERS:**
  - P. Aghera, D. Fang, T. Simunic Rosing, K. Patrick, “Energy management in wireless healthcare systems,” IPSN 2009.
  - A. Coskun, R. Strong, D. Tullsen, T. Simunic Rosing, “Job scheduling and power management on chip multiprocessors for improved processor lifetime,” SIGMETRICS 2009.
  - G. Dhiman. T. Simunic Rosing, “Analysis of DVFS in modern processors,” HotPower2008.
  - A. Coskun, T. Simunic Rosing, K. Gross, “Proactive temperature balancing for low cost thermal management in MPSOCs,” ICCAD’08.
  - E. Regini, D. Lim, T. Simunic Rosing, “Distributed scheduling for heterogeneous wireless sensor networks,“ submitted to ISM’08.
  - A. Coskun, T. Simunic Rosing, K. Gross, “Proactive temperature management in MPSOCs,” to appear in ISLPED 2008.
  - A. Coskun, T. Simunic Rosing, K. Gross, “Temperature management in MPSOCs using online learning,” to appear in DAC 2008.
  - A. Coskun, T. Simunic Rosing, “Temperature-aware MPSOC scheduling for reducing hot spots and gradients,” ASPDAC’08.
  - S. Sharifi, T. Simunic Rosing, “Accurate temperature sensing for efficient thermal management,” ISQED’08.
  - G. Dhiman, T. Simunic Rosing, “Dynamic Voltage Scaling using Machine Learning,” ISLPED 2007.
  - D. Musian, K. Lin, T. Simunic Rosing, “An Active Sensing Platform for Structural Health Monitoring Application,” IPSN-SPOTS’07.
  - A. Coskun, T. Simunic Rosing, “Temperature-aware task scheduling,” DATE’07.
  - D. Lim, J. Shim, T. Simunic Rosing, T. Javidi, “Scheduling data delivery in heterogeneous wireless sensor networks,” ISM’06.
  - G. Dhiman, T. Simunic Rosing, “Dynamic Power Management Using Machine Learning,” ICCAD’06
  - A. Coskun, T. Simunic Rosing, “ A Simulation Methodology for Reliability Analysis in Multi-Core SoCs,” GVLSI’06
  - T. Simunic, K. Mihic, G. De Micheli: “Optmization of Reliability and Power Consumption in Systems on a Chip, “ PATMOS’05.
  - T. Simunic, W. Quadeer, G. De Micheli: “Managing heterogeneous wireless environments via Hotspot servers, “ MMCN’05.