CASL
COMPUTER ARCHITECTURE
AND SYSTEMS LAB

Amdahl’s Law for Lifetime Reliability Scaling in Heterogeneous Multicore Processors

HPCA 2016 – Session 8B
Barcelona, Spain

William Song, Saibal Mukhopadhyay, and Sudhakar Yalamanchili

School of Electrical and Computer Engineering
Georgia Institute of Technology

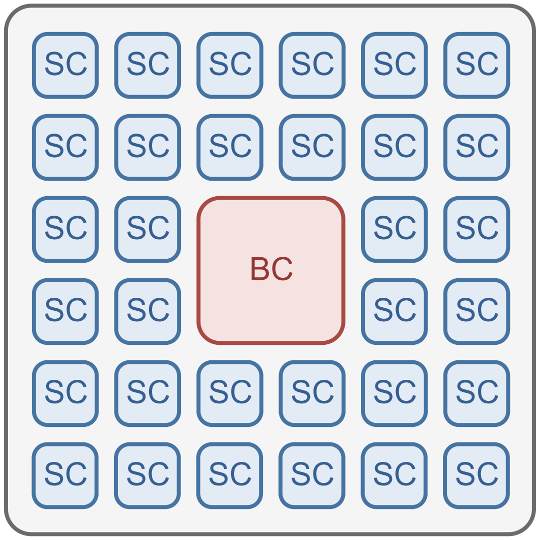
03/16/2016

Outline

- Introduction
- Review of Performance and Energy Models
- Compact Thermal Estimation
- Amdahl’s Law for Heterogeneous Multicore Reliability
- Conclusion

Introduction

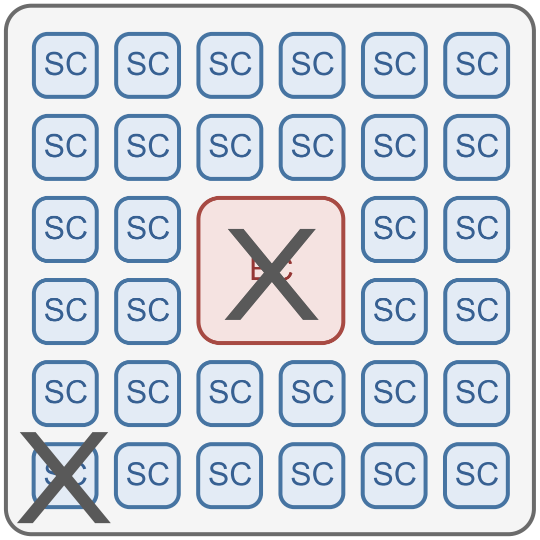
- The paradigm of designing processors has been shifting.
  - Performance → energy efficiency
  - Heterogeneous multicore processors
    - Big core: faster sequential executions
    - Small core: energy-efficient parallel executions
- Amdahl's Law in heterogeneous processors
  - Performance speed-up model from Hill and Marty, Computer (2008)
  - Energy scaling model from Woo and Lee, Computer (2008)



SC: Small Core (i.e., in-order execution)
BC: Big Core (i.e., out-order execution)

Problem Description

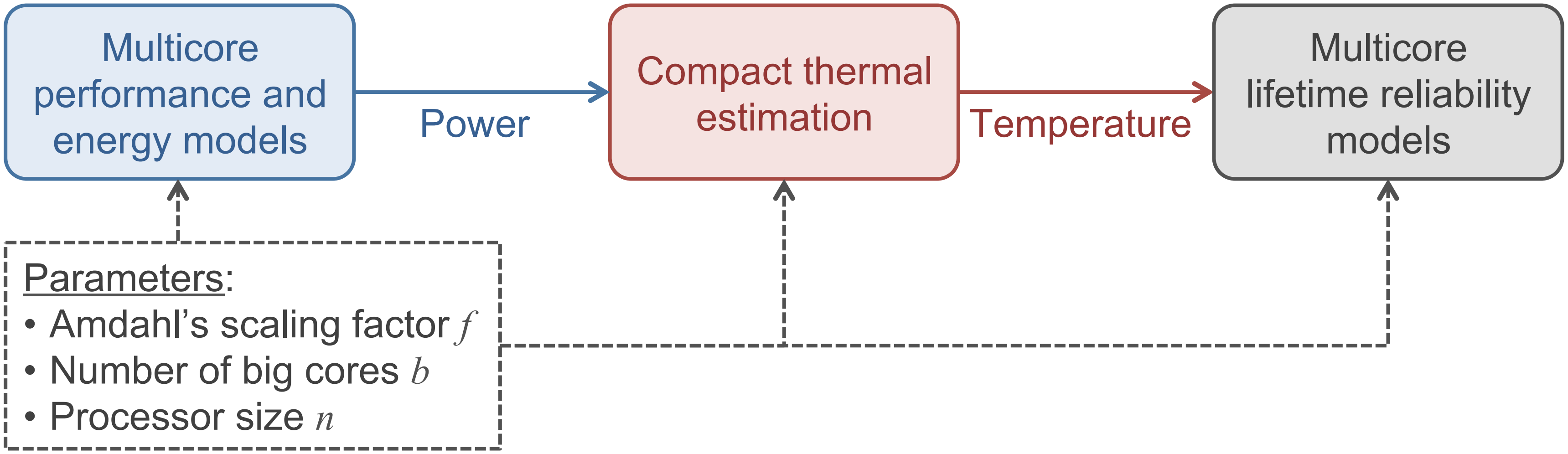
- *Reliability bottleneck* in heterogeneous processors
  - Failure of the only big core:
    - *No other cores* can replace it.
    - Failure of the entire processor or significant performance penalty
  - Failure of a small core:
    - Exploiting *component redundancy*
    - Relatively minor performance loss



SC: Small Core (i.e., in-order execution)
BC: Big Core (i.e., out-order execution)

Modeling Method

- Heterogeneous multicore reliability is a *function of (f, b, n)*:
  - Amdahl's factor f determines stress duration.
  - Multiple big cores b can share serial loads.
  - Processor size n affects total execution time.


Multicore performance and energy models
Power
Compact thermal estimation
Temperature
Multicore lifetime reliability models

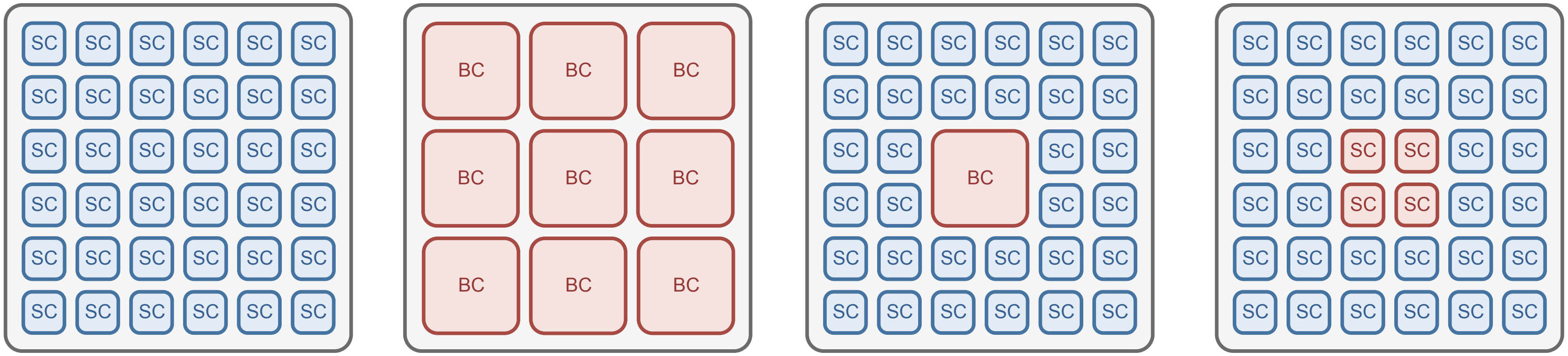
Parameters:
- Amdahl's scaling factor f
- Number of big cores b
- Processor size n

Outline

- Introduction
- **Review of Performance and Energy Models**
- Compact Thermal Estimation
- Amdahl’s Law for Heterogeneous Multicore Reliability
- Conclusion

Revisiting Performance and Energy Models from “Hill and Marty” and “Woo and Lee”

- Multicore models studied in prior work:
  1. Homogeneous processor of small cores
  2. Homogeneous processor of big cores
  3. Heterogeneous processor of one big core and many small cores
  4. Composed processor of small cores

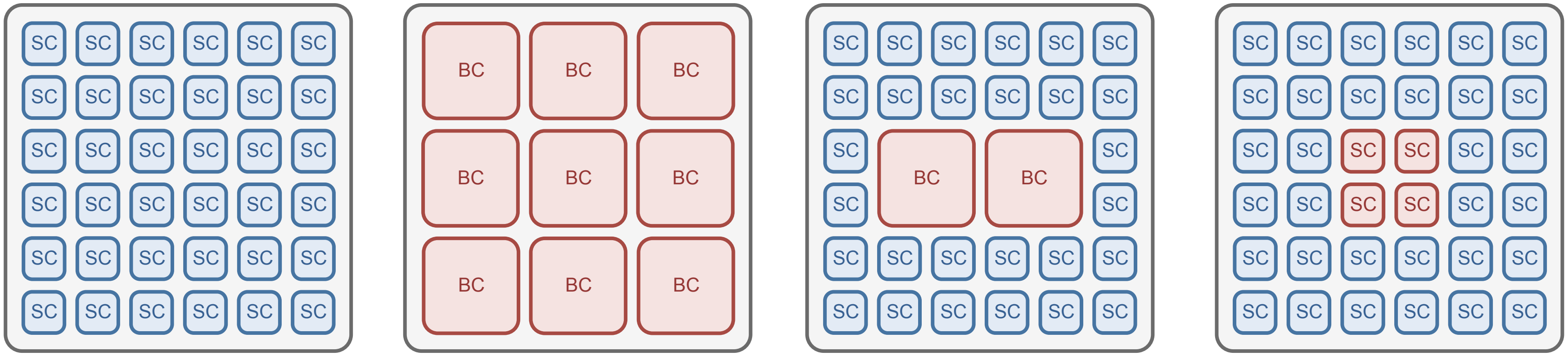


SC: Small Core (i.e., in-order execution)

BC: Big Core (i.e., out-order execution)

Revisiting Performance and Energy Models from “Hill and Marty” and “Woo and Lee”

- Modified assumptions in multicore models:
  - Unused cores are *power-gated*.
  - Heterogeneous processor includes *multiple big cores*.
    - *Maximum scheduling*: big cores take part in parallel executions.
    - *Dynamic scheduling*: big cores are turned off in parallel phases.



SC: Small Core (i.e., in-order execution)

BC: Big Core (i.e., out-order execution)

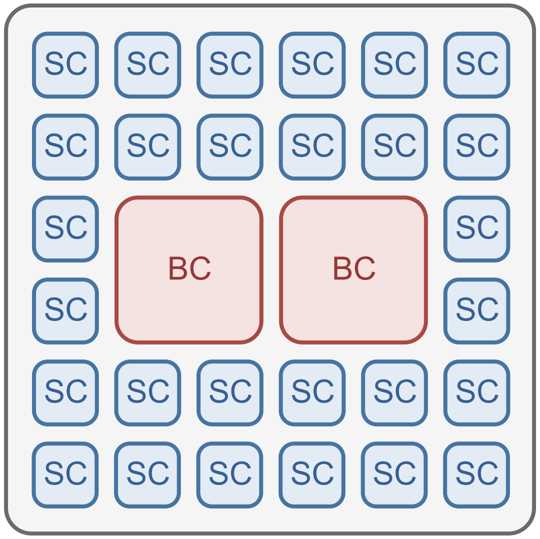
Revisiting Performance and Energy Models from “Hill and Marty” and “Woo and Lee”

- Heterogeneous processor with maximum scheduling
  - Performance speed-up model:

$$Perf_{het:ms} = \frac{1}{\underbrace{\frac{1-f}{s}}_{\text{Serial execution speed-up}} + \underbrace{\frac{f}{b\times s+(n-b\times r)}}_{\text{Parallel execution speed-up}}}$$

  - Energy scaling model:

$$E_{het:ms} = \underbrace{\frac{1-f}{s}p}_{\text{Serial execution energy}} + \underbrace{f\frac{b\times p+(n-b\times r)}{b\times s+(n-b\times r)}}_{\text{Parallel execution energy}}$$



SC: Small Core (i.e., in-order execution)
BC: Big Core (i.e., out-order execution)

f: Amdahl’s factor (parallelizable fraction)

n: processor size (in unit of small cores)

1: small core area, performance, power.

r: big core area.

s: big core performance $s \propto \sqrt{r}$ (Pollack’s Rule, MICRO 1999)

p: big core power $p \propto (\sqrt{r})^{\alpha}$ (Chung’s model MICRO 2010)

i : big core idle-state power

Big and Small-Core Pairs

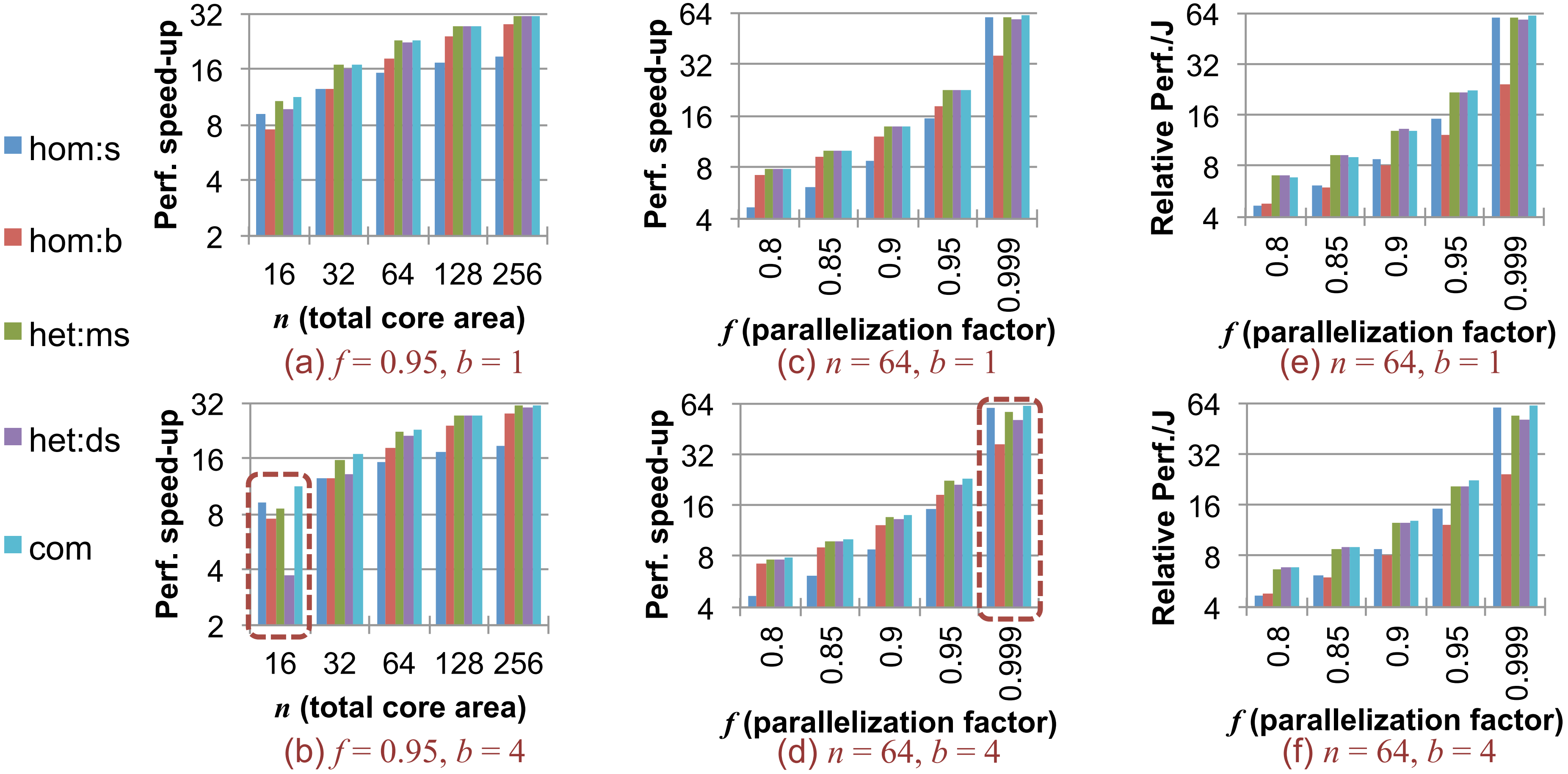
	IBM BlueGene/Q	IBM POWER7	Intel Atom Z520	Intel i7 960
Core execution type	In-order	Out-of-order	In-order	Out-of-order
Technology node	45nm	45nm	45nm	45nm
Estimated die area	360mm^2	567mm^2	26mm^2	263mm^2
Number of cores	18	8	1	4
Cores-to-die area ratio	34%	32%	37%	37%

- *Focus on the cores:* No discernible correlation is found between *cores-to-die area ratio* and *core types*, *number of cores*, or other *uncore configurations*.
- *r = 3 is chosen:* Big and small core area ratio is estimated around r = 2.5-4.4

	IBM POWER7+	Intel i7 2700K	Intel i7 3770K
Core execution type	In-order	Out-of-order	Out-of-order
Technology node	32nm	32nm	22nm
Core area scaling from prev. gen.	0.68x	0.66x	0.66x
Number of cores	8	4	4
Cores-to-die area ratio	37%	37%	37%

Evaluation of Performance and Energy Efficiency Models

- The most energy efficiency and performance speed-up are achieved when a heterogeneous processor includes *one big core (b=1) and many small cores*.
- Including multiple big cores penalizes the performance and energy efficiency of heterogeneous multicore especially for small n or large f.



(a) $f = 0.95$, $b = 1$

(c) $n = 64$, $b = 1$

(e) $n = 64$, $b = 1$

(b) $f = 0.95$, $b = 4$

(d) $n = 64$, $b = 4$

(f) $n = 64$, $b = 4$

Outline

- Introduction
- Review of Performance and Energy Models
- **Compact Thermal Estimation**
- Amdahl’s Law for Heterogeneous Multicore Reliability
- Conclusion

Compact Thermal Estimation

- Reliability and temperature dependency
  - Assuming *uniform temperature or failure rate* is not a reasonable approach.
  - Goal: *thermal difference → reliability difference estimation*
- Thermal difference is created by:
  - *Processor composition* (homogeneous or heterogeneous)
  - *Execution phase* (sequential or parallel)
  - *Thread scheduling* (maximum or dynamic)
- *Baseline*: homogeneous processor of small cores in parallel executions
  - Steady-state temperature difference: $\Delta\mathbf{x}' = \mathbf{C}\Delta\mathbf{u}$
    - $\mathbf{x}'$: steady-state temperature vector
    - $\mathbf{C}$: power input to temperature conversion factors
    - $\mathbf{u}$: power input vector

Compact Thermal Estimation

- Steady-state temperature difference: $\Delta \mathbf{x}' = \mathbf{C}\Delta \mathbf{u}$
- Thermal change of floorplan at j: $\Delta \mathrm{x}'_j = \underbrace{\mathrm{C}_{jj}\Delta \mathrm{u}_j}_{\text{Thermal change due to power input at } j} + \underbrace{\sum_{k\neq j}^{n} \mathrm{C}_{kj}\Delta \mathrm{u}_k}_{\text{Thermal change due to power input at } k\neq j}$
- For unknown floorplanning, a *scalar approximation* is used.

$$\Delta \mathrm{x}'_j = \mathrm{C}_{jj}\Delta \mathrm{u}_j + \underbrace{\bar{\mathrm{C}}_{kj}\Delta \bar{\mathrm{u}}}_{\text{Average of power changes and thermal impact}}$$

- If j belongs to a big core

$$\Delta \bar{\mathrm{u}}_b = \underbrace{\frac{(b-1)r}{n-r}\delta(p_b-1)}_{\text{Average power changes of other big cores at } k\neq j} + \underbrace{\frac{n-b\times r}{n-r}\delta(p_s-1)}_{\text{Average power changes of other small cores at } k\neq j}$$

- If j belongs to a small core:

$$\Delta \bar{\mathrm{u}}_s = \frac{b\times r}{n-1}\delta(p_b-1) + \frac{n-1-b\times r}{n-1}\delta(p_s-1)$$

f: Amdahl's factor (parallelizable fraction)

n: processor size (in unit of small cores)

1: small core area, performance

r: big core area

b: big core count

s: big core performance

p_b: big core power
$p \propto (\sqrt{r})^{\alpha}$ when active, 0 if power-gated

p_s: small core power
1 when active, 0 if power-gated

Accuracy of Compact Thermal Estimation

- Comparison to a detailed thermal model (HotSpot)

	Maximum difference (°C) to HotSpot model			
Processor type	**Sequential execution**		**Parallel execution**	
	Big core	**Small core**	**Big core**	**Small core**
Homogeneous: small cores	N/A	-0.41	N/A	Baseline
Homogeneous: big cores	+0.63	N/A	+0.45	N/A
Heterogeneous: max. sch.	+0.63	Unused	-0.19	-0.01
Heterogeneous: dyn. sch.	+0.63	Unused	Unused	-0.58
Composed: small cores	N/A	-0.32	N/A	Same as Baseline

- Compact thermal estimation has *less than 1°C difference* to a detailed model.

Outline

- Introduction
- Review of Performance and Energy Models
- Compact Thermal Estimation
- **Amdahl’s Law for Heterogeneous Multicore Reliability**
- Conclusion

Lifetime Reliability Model

- Hot carrier injection: $\mathrm{MTTF}_{HCI} = A_{HCI}\ I_{sub}^{-n}\ e^{(E_{a\ HCI}/kT)}$
- Negative bias temperature instability: $\mathrm{MTTF}_{NBTI} = A_{NBTI}\ V_{gs}^{-r}\ e^{(E_{a\ NBTI}/kT)}$
- Exponential model: $\mathrm{MTTF} = 1 / \lambda$
- Sum of failure rates (SOFR)
  - $\lambda = \lambda_{HCI} + \lambda_{NBTI}$ for each core
  - $\lambda_b = r \times \lambda_s$ for the same operating conditions

λ_b: big core failure rate
λ_s: small core failure rate

- λ is obtained from *compact thermal estimation* for each *core type* of different *processor composition* and *execution phase*.

MTTF: Mean Time to Failure

Multicore Lifetime Reliability Model

- Heterogeneous processor with maximum scheduling

$$\lambda_{het:ms} = \underbrace{\frac{1-f}{s}}_{\text{Serial execution time}} \times \underbrace{\frac{\lambda_{b:seq}}{b}}_{\text{Big core failure rate (load sharing) during serial executions}}$$

$$+ \underbrace{\frac{f}{b\times s + (n - b\times r)}}_{\text{Parallel execution time}} \times \{\underbrace{b\times\lambda_{b:par}}_{\text{Big core failure rate during parallel executions}} + \underbrace{(n - b\times r)\lambda_{s:par}}_{\text{Small core failure rate during parallel executions}}\}$$

f: Amdahl's factor (parallelizable fraction)

n: processor size (in unit of small cores)

1: small core area, performance, power

r: big core area

b: big core count

s: big core performance $s \propto \sqrt{r}$

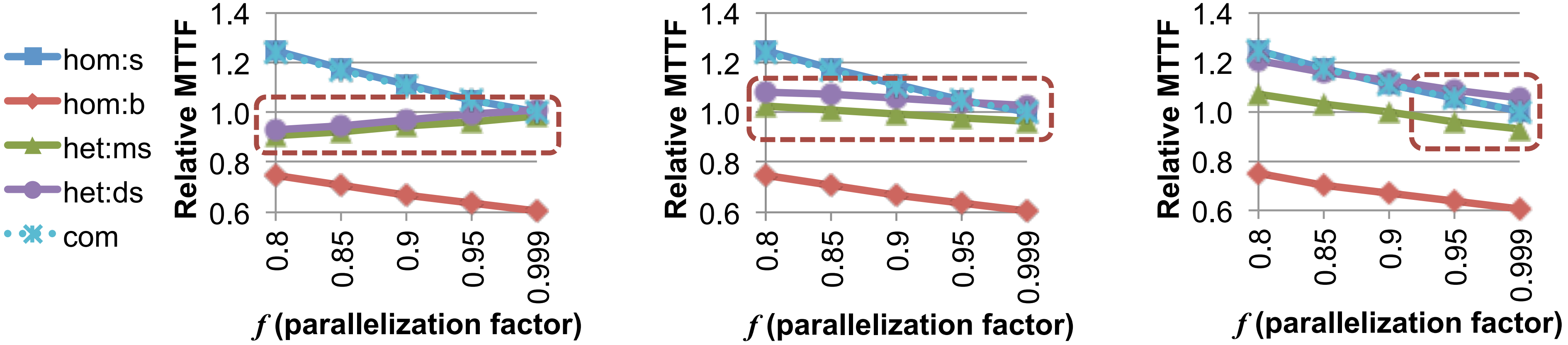
p: big core power $p \propto (\sqrt{r})^{\alpha}$, $\alpha = 1.75$

λ_b: big core failure rate

λ_s: small core failure rate

Lifetime Reliability Evaluation

- Relative lifetime reliability

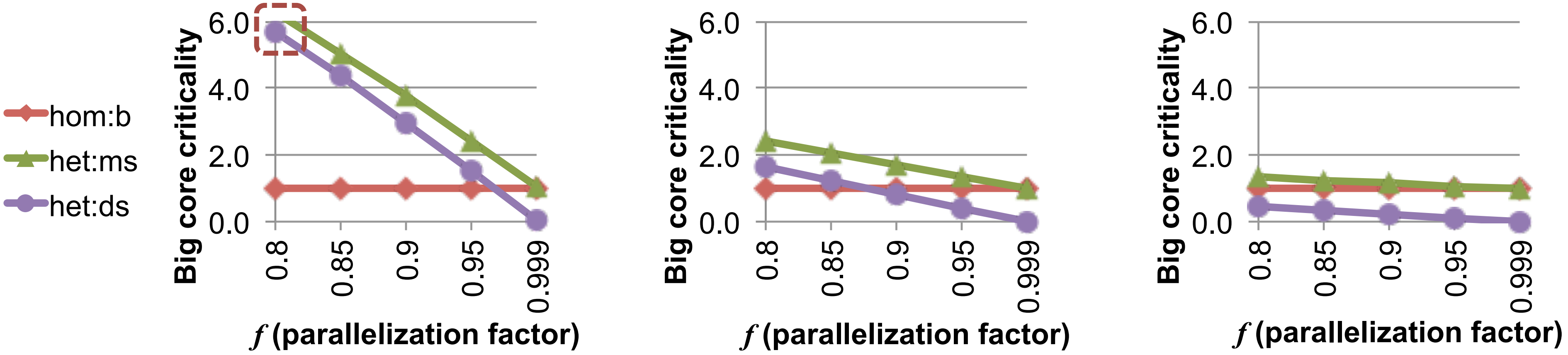


(a) $n = 64, b = 1$

(c) $n = 64, b = 2$

(e) $n = 64, b = 4$

- Big core criticality



(b) $n = 64, b = 1$

(d) $n = 64, b = 2$

(f) $n = 64, b = 4$

Conclusion

- **Contributions**
  1. Extended performance and *energy efficiency models*
  2. *Compact thermal estimation* for reliability modeling
  3. *Heterogeneous multicore reliability models*
  4. *Lifetime reliability assessment* of heterogeneous processors

- **Insights**
  - Reliability bottleneck:
    - *Small b/n ratio* puts biased stress on big cores.
    - *Diminishing parallelization fraction f* shifts stress from small to big cores.
  - Performance and reliability tradeoff:
    - *Increasing b/n ratio* relieves big core criticality.
    - But, *large b/n ratio* i) decreases peak parallel throughput, ii) extends total execution time, and thus iii) has an adverse impact on reliability.