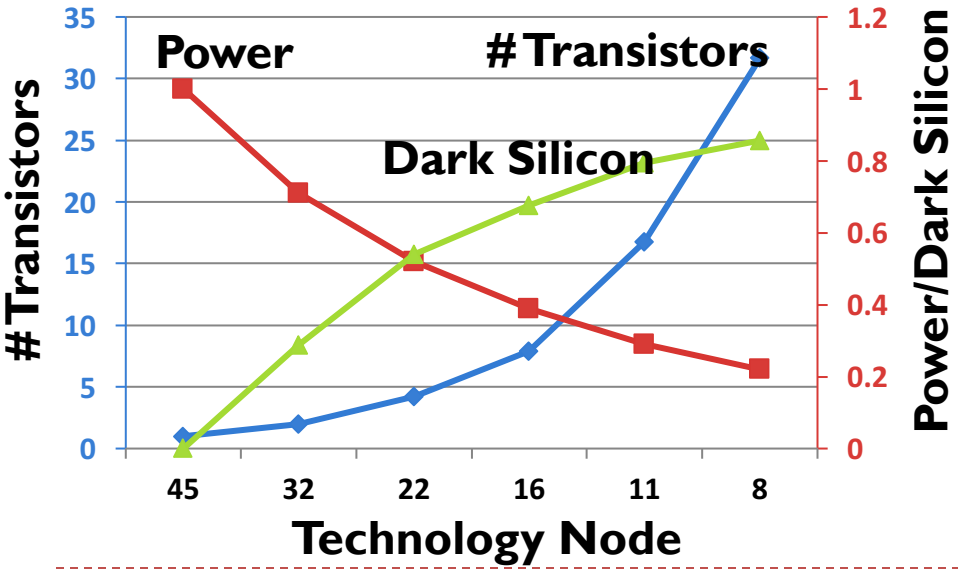
HaDeS: Architectural Synthesis for Heterogeneous Dark Silicon Chip Multi-processors

Yatish Turakhia, Bharathwaj Raghunathan, Siddharth Garg, Diana Marculescu

Dark Silicon Challenge

- **Transistor dimensions scale by a factor of *S***
  - S^2 more transistors in same die area
  - Each transistor dissipates $\approx \frac{1}{S}$ times power
  - Fixed power budget results in **dark silicon**



80% dark silicon at the 8 *nm* technology node

[Esmaeilzadeh *et al.*, ISCA'11]

Progress Hits Snag: Tiny Chips Use Outsize Power

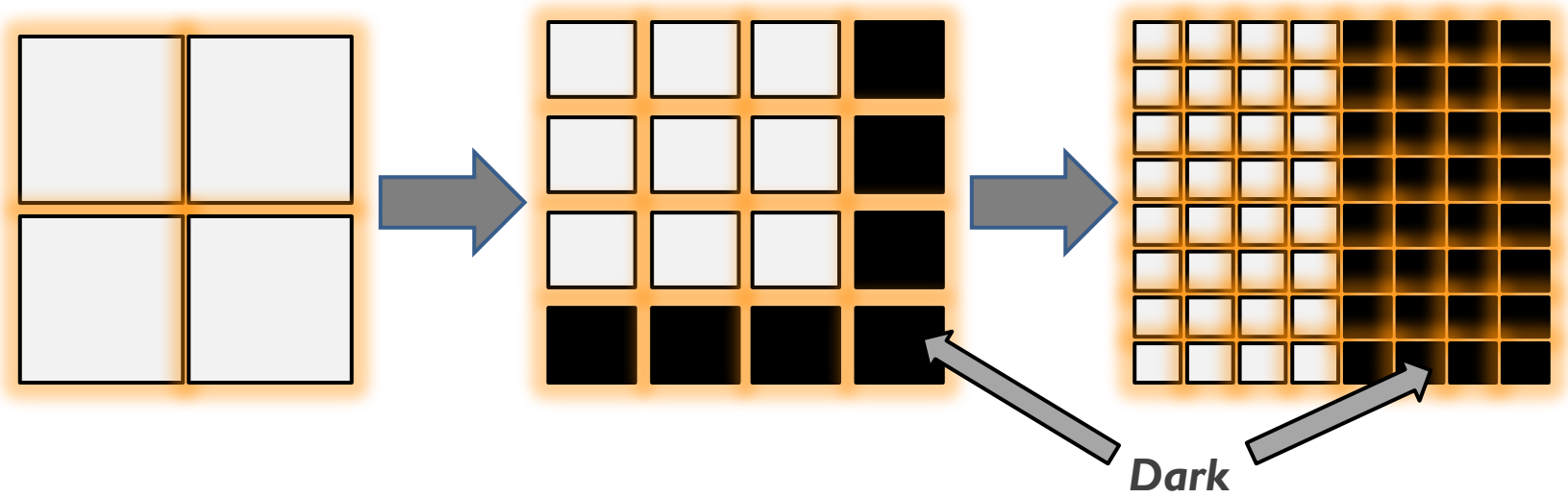
By JOHN MARKOFF
Published: July 31, 2011

For decades, the power of computers has grown at a staggering rate as designers have managed to squeeze ever more and ever tinier transistors onto a silicon chip — doubling the number every two

RECOMMEND
TWITTER
LINKEDIN

The Dark Silicon Era: Demise of Multi-core Scaling?

- Rapid rise in the fraction of dark silicon due to power constraints.



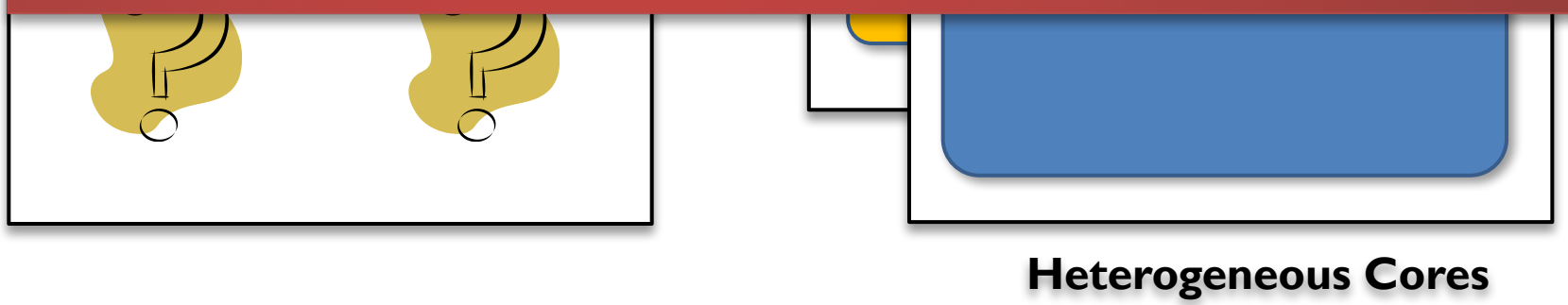
- What's the point of increasing core count if they aren't active??
- Need to think out of the  !!!

Dark Silicon Architectures

- **How to best utilize dark silicon for**

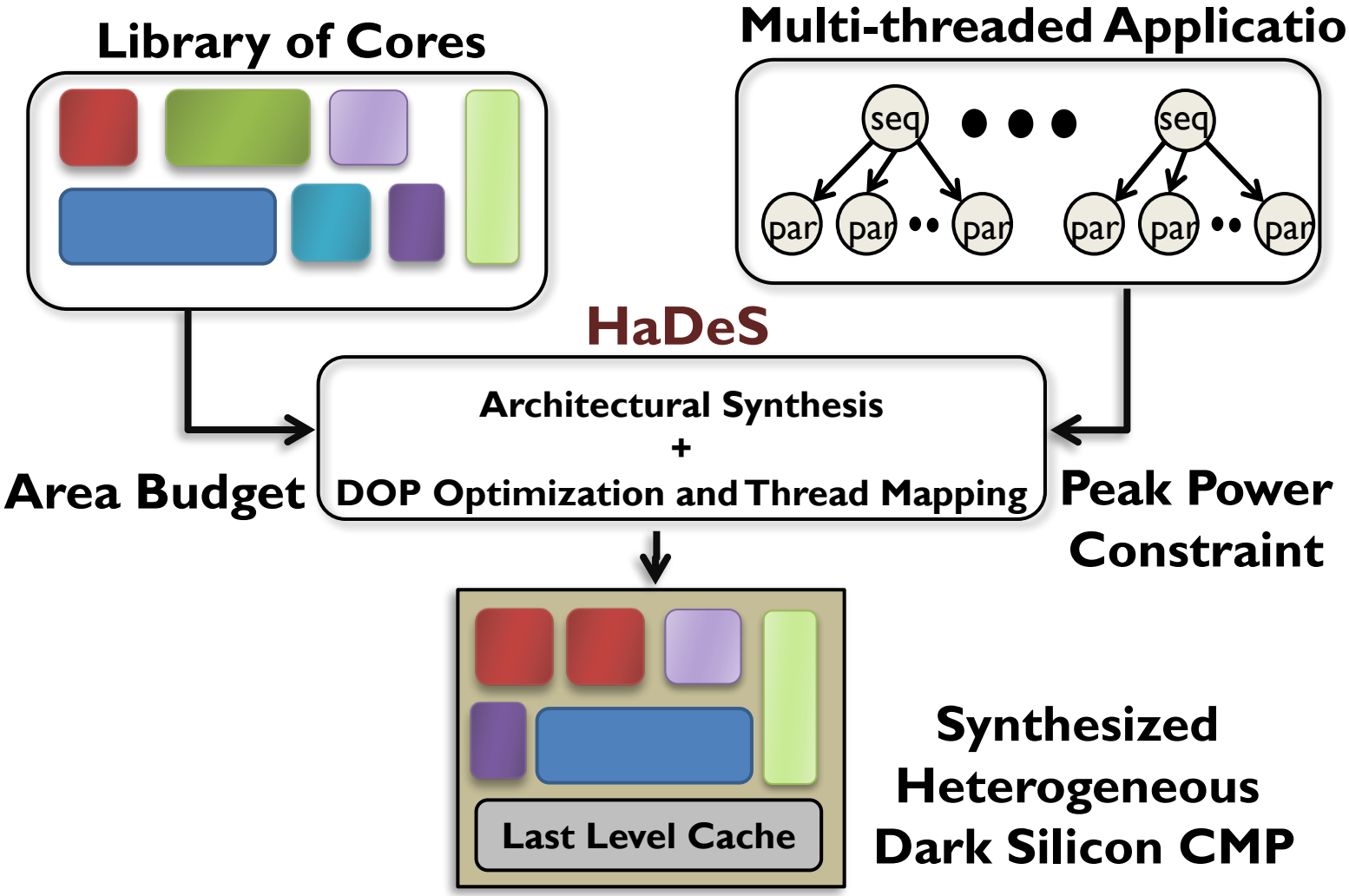
How do we decide?

- What types of cores and how many cores of each type should one provision on a CMP of given area so as to maximize the overall performance benefit?
- Which core/s to turn on without exceeding power budget / how to map threads?
- How much parallelism to exploit?

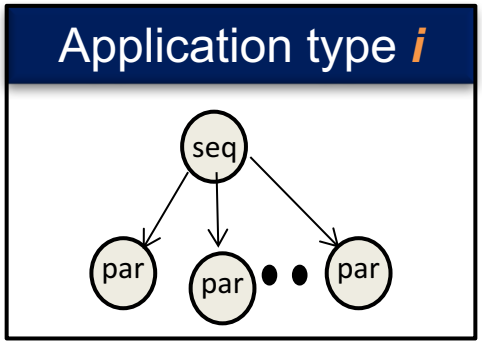


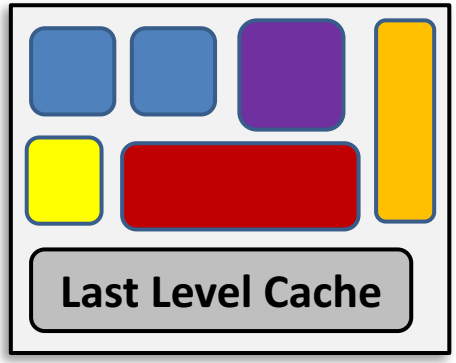
Heterogeneous Cores

HaDeS: Overview



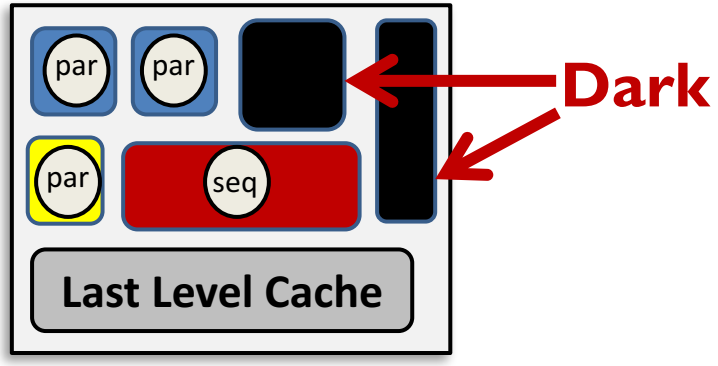
Application Performance Model





Synthesized Heterogeneous Dark Silicon CMP

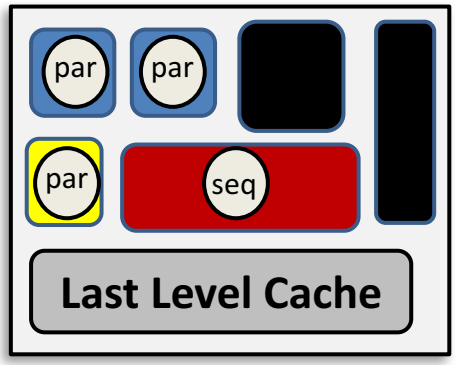
Application Performance Model



Application Performance Model

E_i

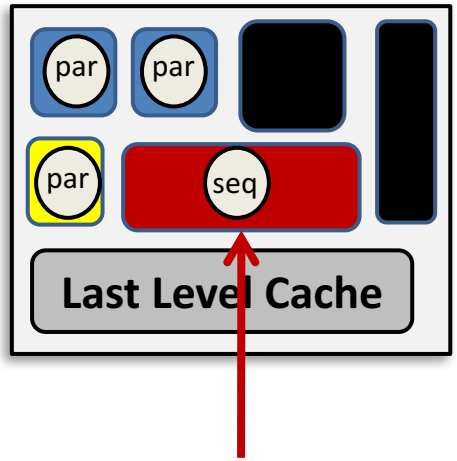
Total execution time of application *i*



Application Performance Model

$$E_i = t^{s}_{im^s}$$

Sequential time of application *i* on core type m^s

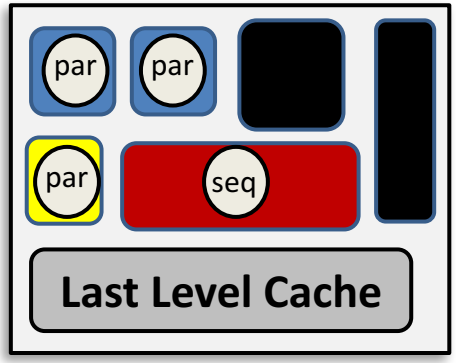


Core type m^s

Application Performance Model

Slowest parallel thread limits performance

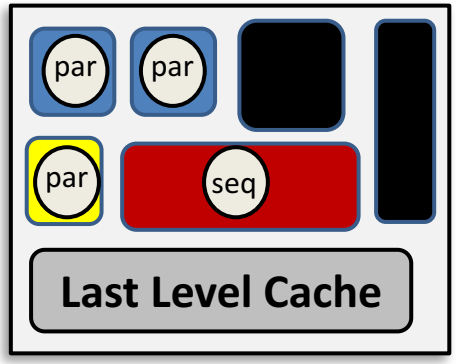
$$E_i = t^s_{im^s} + \max_{v \in [1, DOP_i]} \{$$



Application Performance Model

$$E_i = t^s_{im^s} + \max_{v \in [1, DOP_i]} \{$$

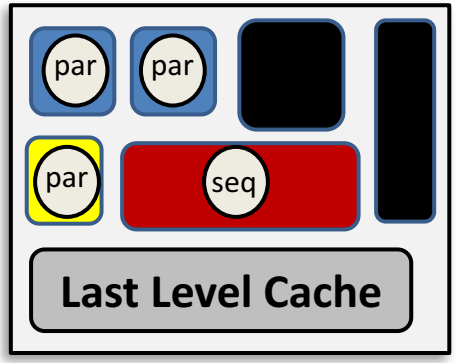
Ranges over all parallel threads



Application Performance Model

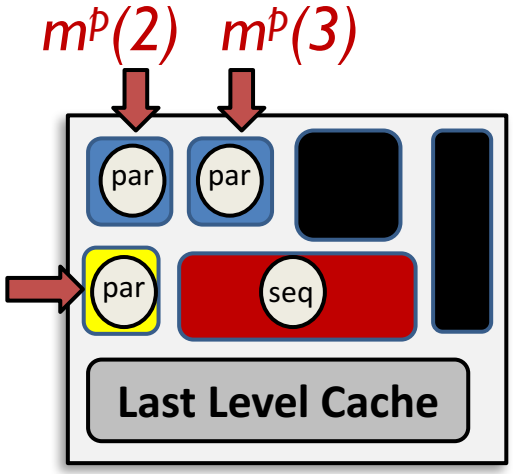
$$E_i = t^s_{im^s} + \max_{v \in [1, DOP_i]} \{$$

Degree of parallelism of application *i*



Application Performance Model

$$E_i = t^{s}_{im^s} + \max_{v \in [1, DOP_i]} \{ \frac{t^{p}_{im^p(v)}}{DOP_i} + \}$$

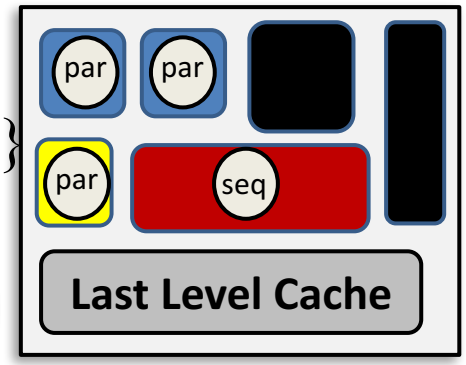


Total execution time of a parallel thread *v* of application *i* depends on the core mapping $m^p(v)$ and is inversely proportional to # parallel threads.

Application Performance Model

$$E_i = t^s_{im^s} + \max_{v \in [1, DOP_i]} \{ \frac{t^p_{im^p(v)}}{DOP_i} + K_{im^p(v)} . DOP_i \}$$

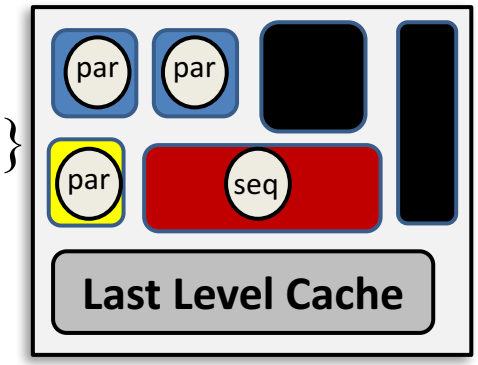
Penalty term for increased resource contention with increasing # parallel threads



Application Performance Model

Model parameters

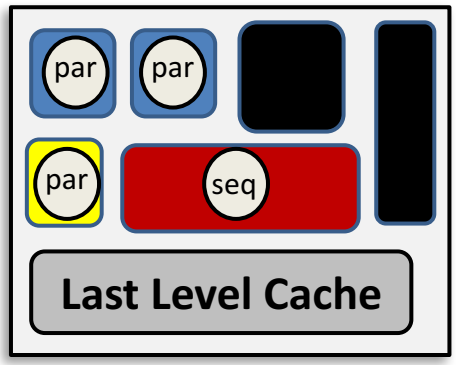
$$E_i = t^s_{im^s} + \max_{v \in [1, DOP_i]} \{ \frac{t^p_{im^p(v)}}{DOP_i} + K_{im^p(v)} . DOP_i \}$$

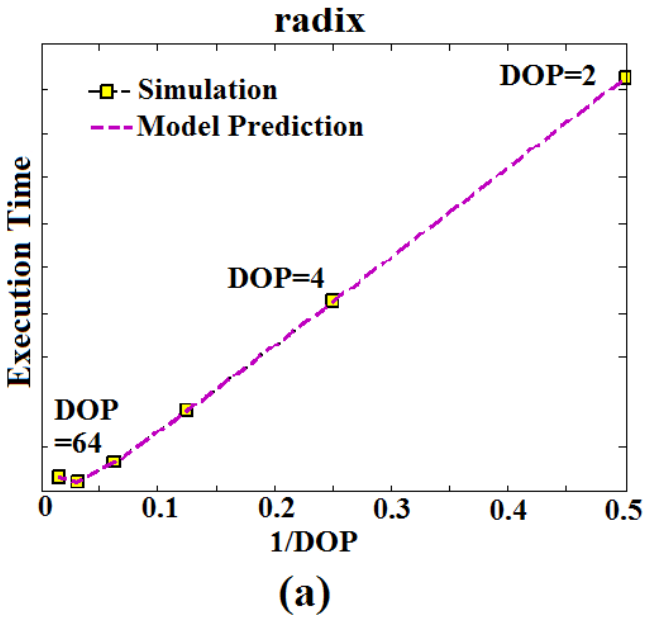


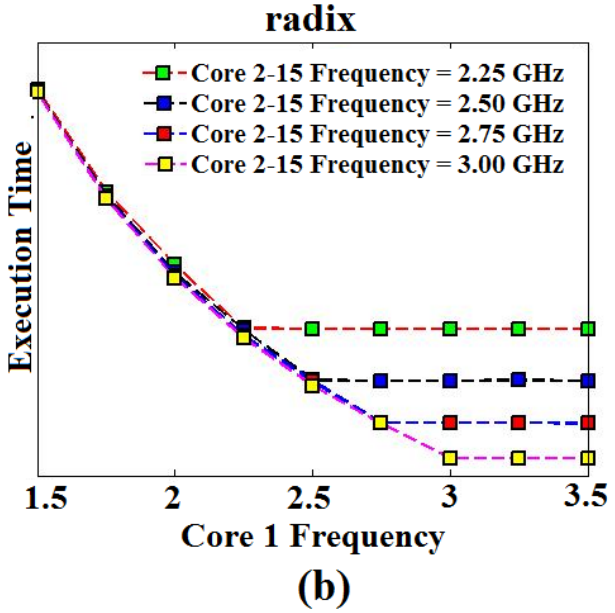
Application Performance Model: Validation

Simple, but surprisingly accurate!!!

$$E_i = t^s_{im^s} + \max_{v \in [1, DOP_i]} \{ \frac{t^p_{im^p(v)}}{DOP_i} + K_{im^p(v)} . DOP_i \}$$







HaDeS Framework

$$\min\left(\sum_{i}^{N} \rho_i \cdot (l_i^s + l_i^p)\right)$$

User defined weights for workload *i*

Goal: Minimize weighted sum of execution time for each application.

HaDeS Framework

$$\min\left(\sum_{i}^{N} \rho_i \cdot (l_i^s + l_i^p)\right)$$

Sequential execution time of workload *i*

Goal: Minimize weighted sum of execution time for each application.

HaDeS Framework

$$\min\left(\sum_{i}^{N}\rho_i.(l_i^s+l_i^p)\right)$$

Parallel execution time of workload *i*

Goal: Minimize weighted sum of execution time for each application.

HaDeS Framework: ILP-OPT

$$\min\left(\sum_{i}^{N}\rho_i.(l_i^s+l_i^p)\right)$$

OBJECTIVE

$$l_i^p \geq b_{ij}\left(\frac{t_{ij}^p}{DOP_i}+K_{ij}.DOP_i\right)\ \forall i\in[1,N]$$

Constraint: Parallel Execution Time

0-1 Indicator variable that is 1 if at least one parallel thread of application i executes on a core of type j

HaDeS Framework: ILP-OPT

$$\min\left(\sum_{i}^{N}\rho_i.(l_i^s+l_i^p)\right)$$

OBJECTIVE

$$l_i^p \geq b_{ij}\left(\frac{t_{ij}^p}{DOP_i}+K_{ij}.DOP_i\right)\ \forall i\in[1,N]$$

Constraint: Parallel Execution Time

Variable DOP introduces non-linearity

HaDeS Framework: ILP-OPT

$$\min\left(\sum_{i}^{N}\rho_i.(l_i^s+l_i^p)\right)$$

OBJECTIVE

$$l_i^p \geq \sum_{w=1}^{DOP_{\max}} b_{ij}\, y_{iw}\left(\frac{t_{ij}^p}{w}+K_{ij}w\right)=\sum_{w=1}^{DOP_{\max}} m_{ijw}\left(\frac{t_{ij}^p}{w}+K_{ijw}\right)$$

$$m_{ijw}\leq b_{ij} \quad \forall i\in[1,\mathrm{N}],\ j\in[1,\mathrm{N}],\ w\in[1,\mathrm{DOP}_{\max}]$$

$$m_{ijw}\leq y_{iw} \quad \forall i\in[1,\mathrm{N}],\ j\in[1,\mathrm{N}],\ w\in[1,\mathrm{DOP}_{\max}]$$

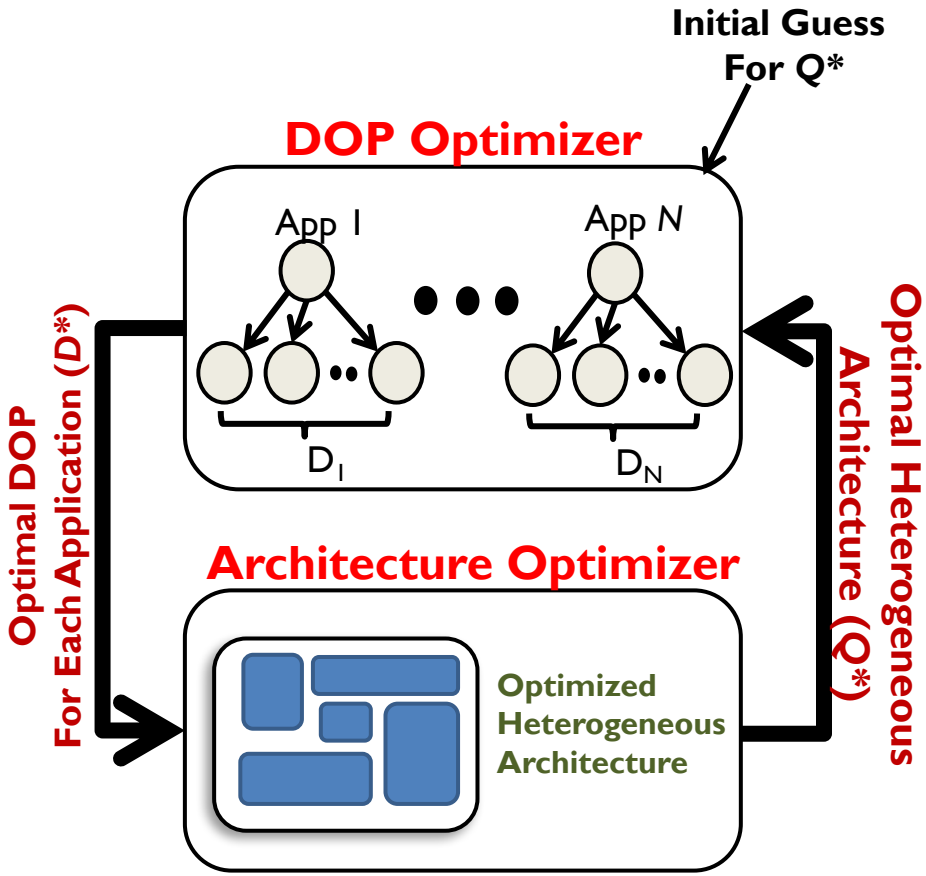
$$m_{ijw}\geq b_{ij}+y_{iw}-1 \quad \forall i\in[1,\mathrm{N}],\ j\in[1,\mathrm{N}],\ w\in[1,\mathrm{DOP}_{\max}]$$

Constraint: Parallel Execution Time

Many Only O(MxN) variables in fixed DOP case an ILP!

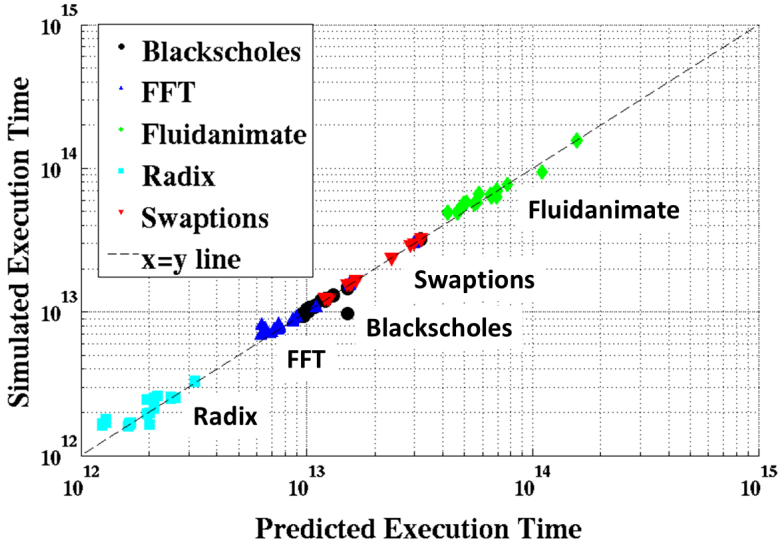
HaDeS Framework: ITER-OPT

- Separates the architectural optimization from DOP optimization.
- Start with initial guess of design vector, compute optimal DOP for this heterogeneous design, re-synthesize heterogeneous design for this DOP and keep iterating till no further improvement in performance.
- Represents **local minima**. Only 2.5% loss of optimality in experiments.
- Orders of magnitude faster (430X speedup)than ILP-OPT.



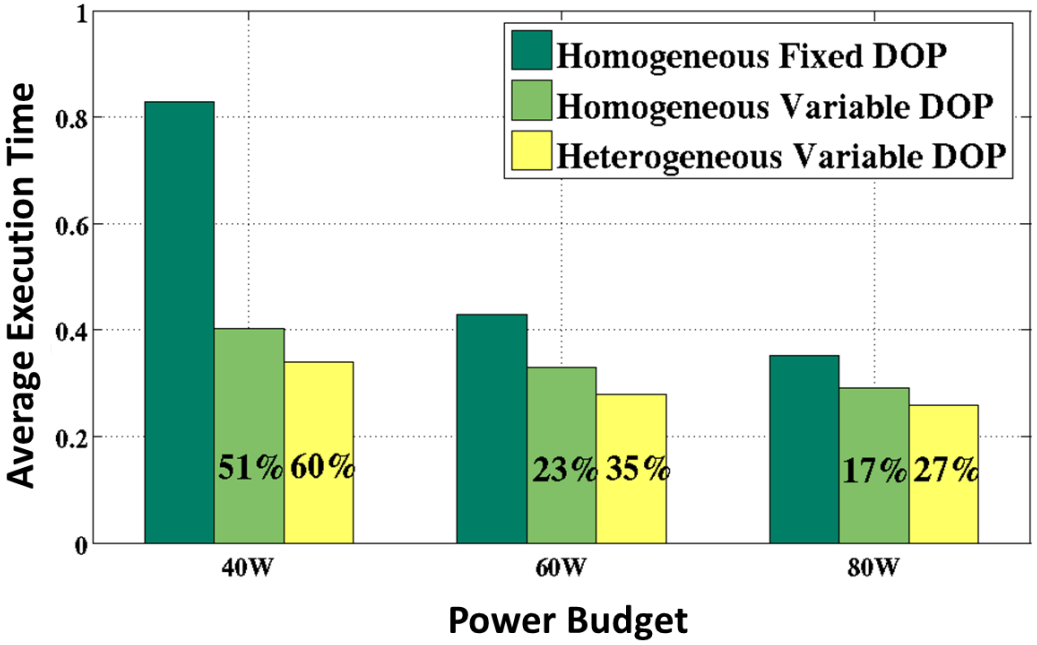
Experimental Results

- Used Snipersim, interval core model based simulator for performance estimates and McPAT for power/area numbers.
- Library of 108 cores with 5 multi-threaded application benchmarks (from Parsec and Splash2 suites). Only 15 cores pareto optimal in area, power or performance.
- Performance model fairly accurate. Only 5.2% average error over 180 experiments with homogeneous and heterogeneous dark silicon CMPs.



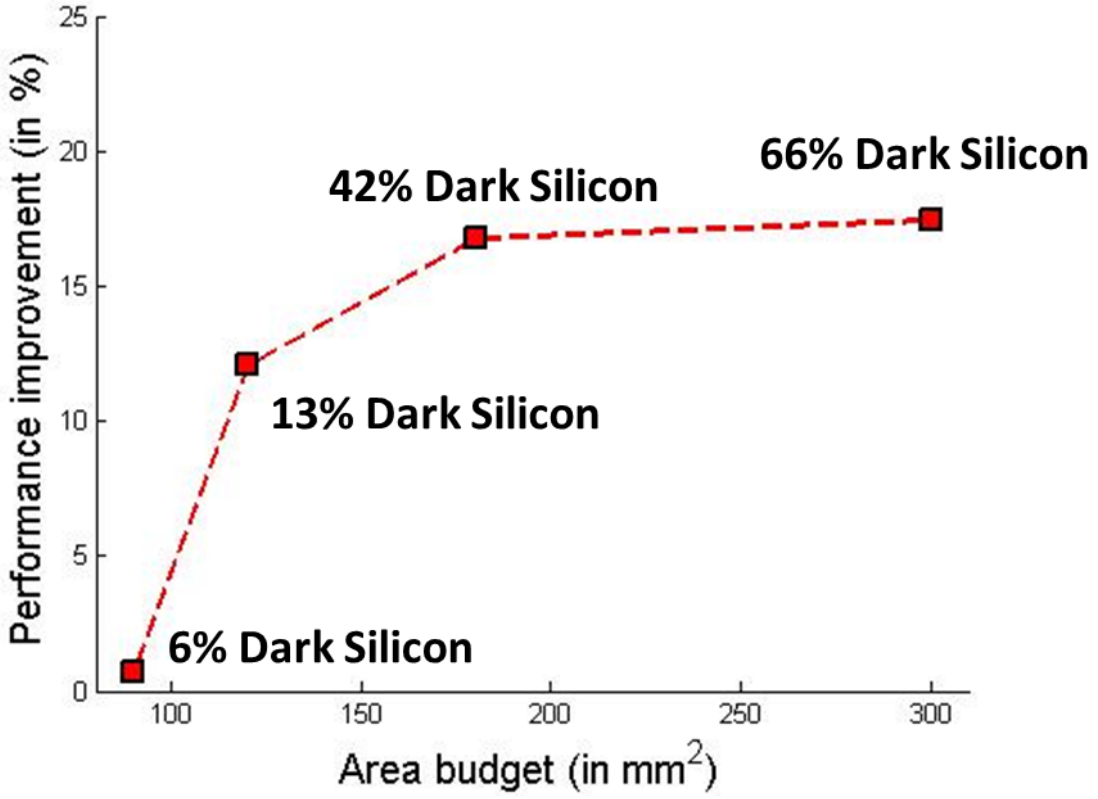
Experimental Results

- Between 19% to 60% performance improvements in heterogeneous dark silicon CMPs over variable and fixed (16 threads) DOP homogeneous design respectively for ~50% dark silicon.



Experimental Results

- Benefits of heterogeneous design begin to saturate beyond certain portion of dark silicon.



Thank You!

Please visit the poster for questions and discussions!