GDP and More:

Thermal and Power Aware Intelligent Design and Optimization of Integrated Systems

Hai Wang

University of Electronic Science & Technology of China

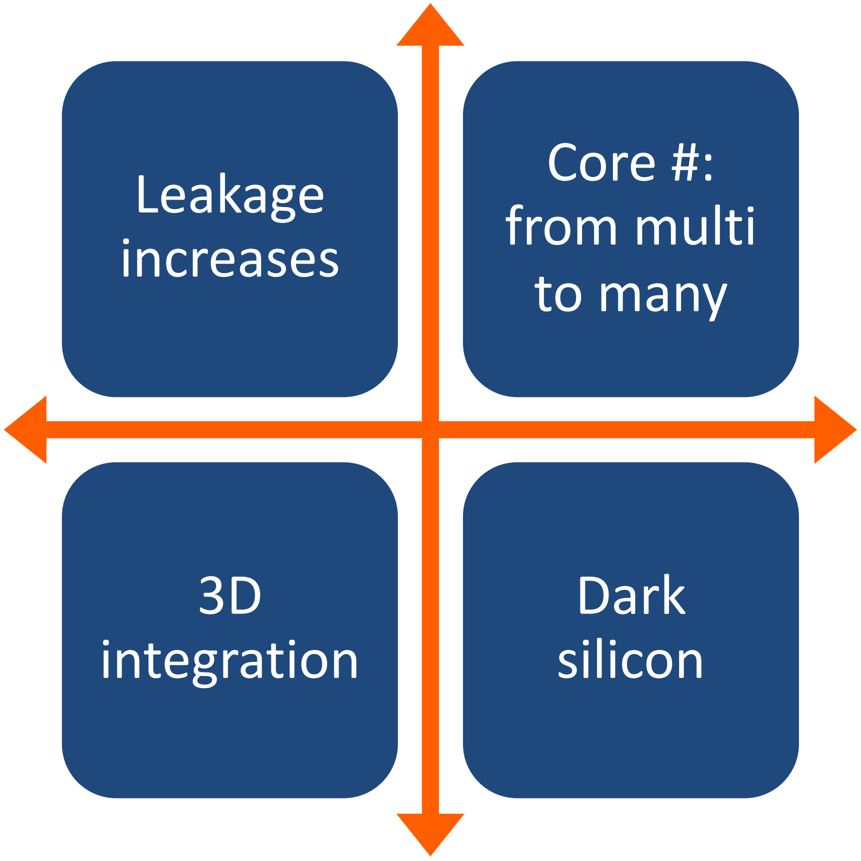
Homepage (English): https://wanghaiuestc.github.io

Homepage (Chinese): http://faculty.uestc.edu.cn/wanghai1

2019

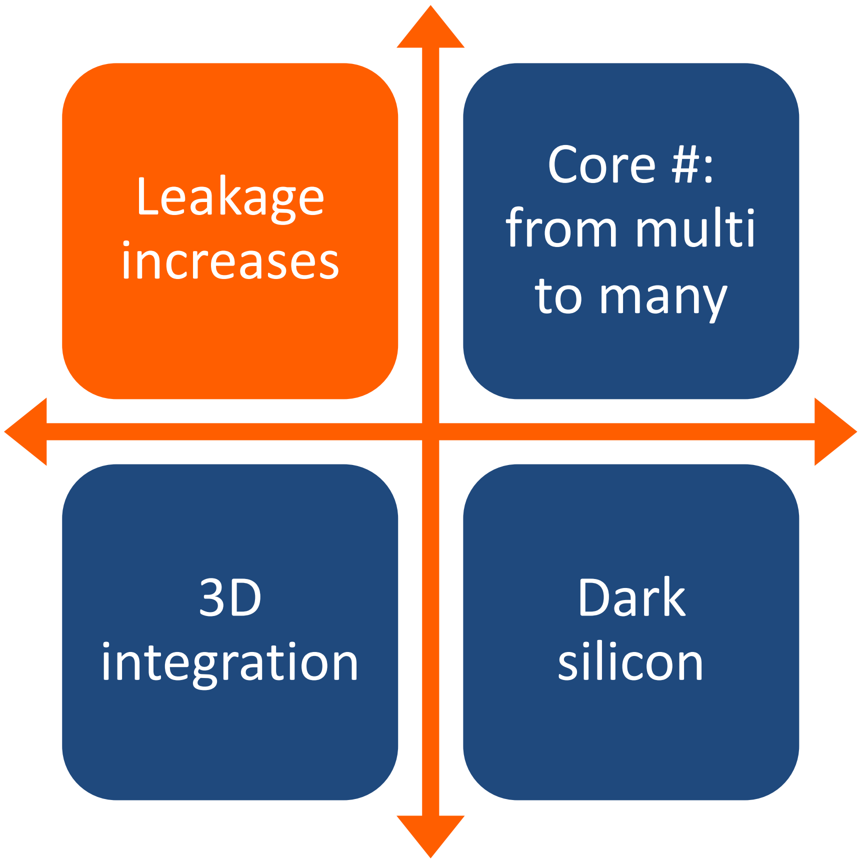
Motivation and Background

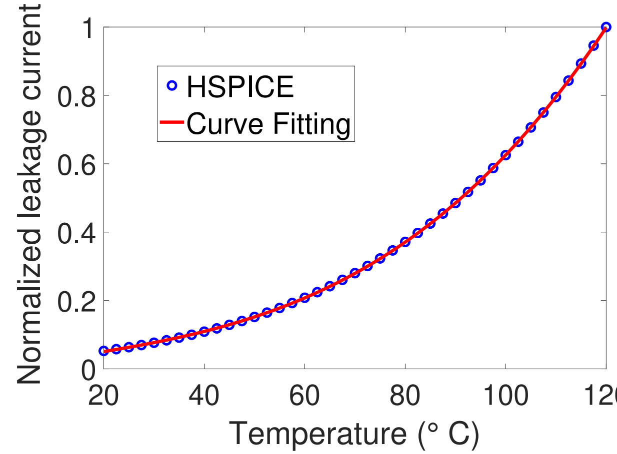
The new challenges in IC industry

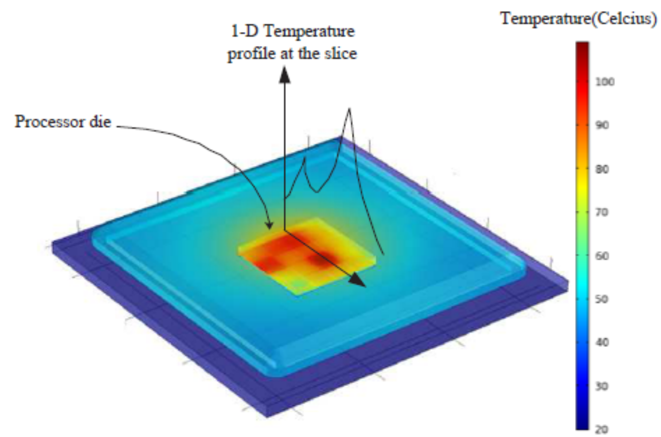


- Scaling causes new challenges in IC industry.
- Solutions needed for new challenges.

The leakage problems

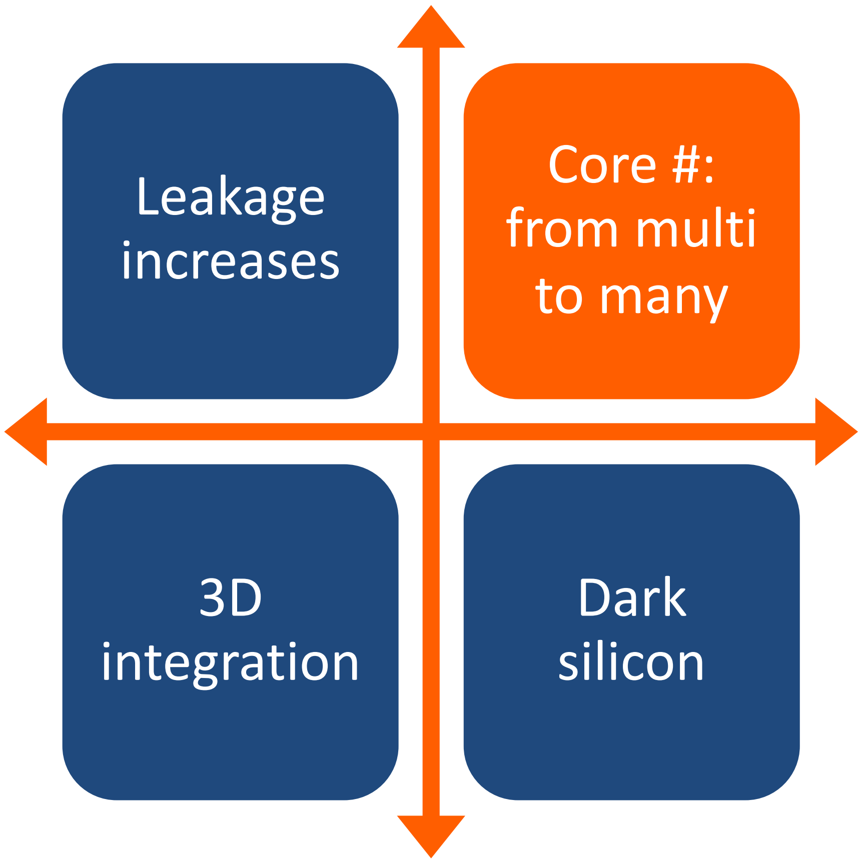


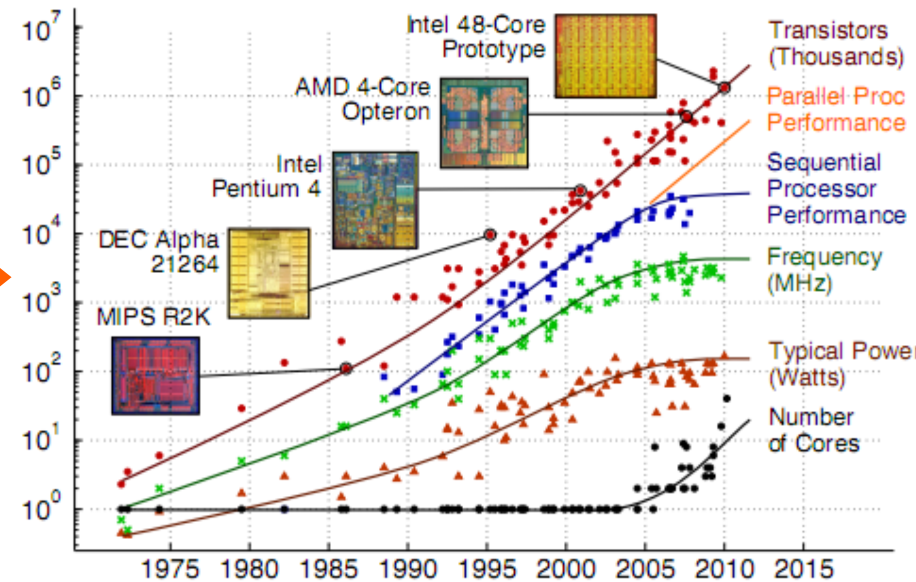




- Leakage power becomes significant.
- Leakage power highly and nonlinearly relates to temperature: dangerous and difficult to model.

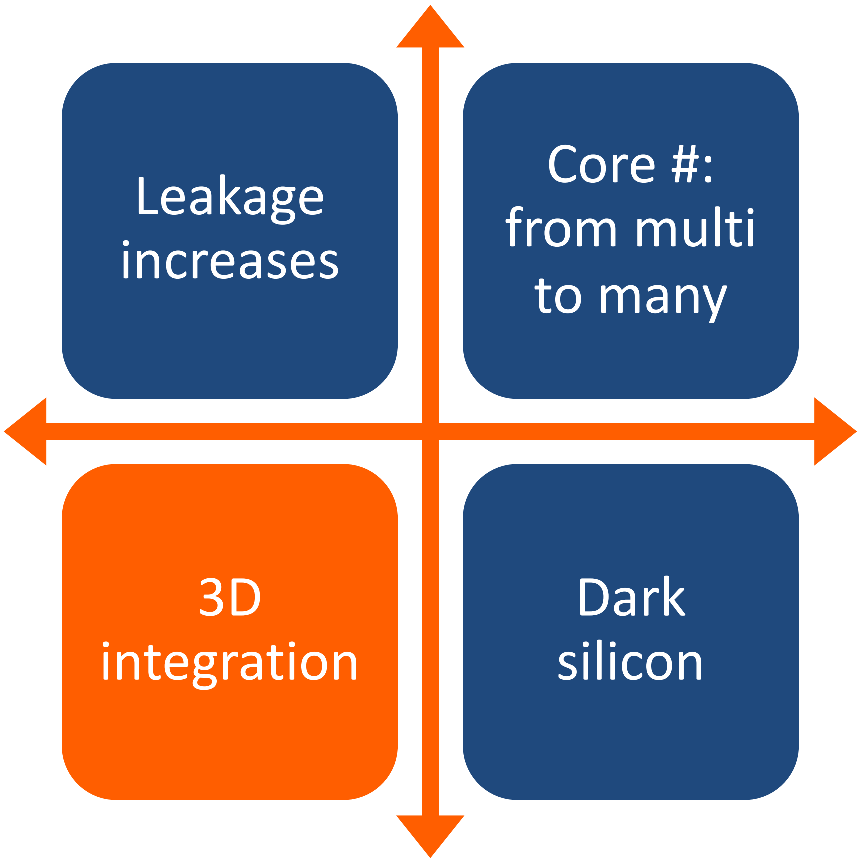
The many-core challenge

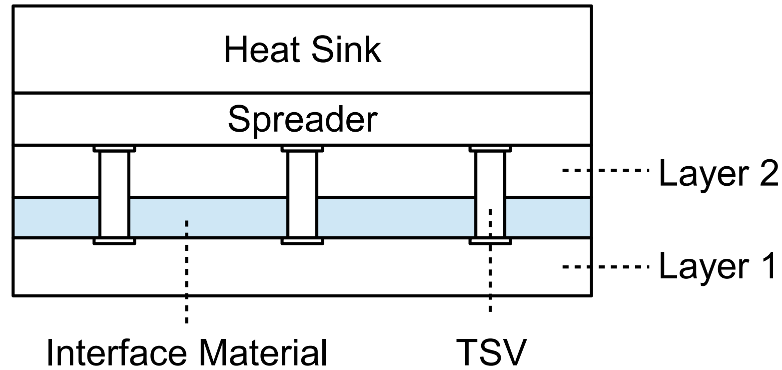


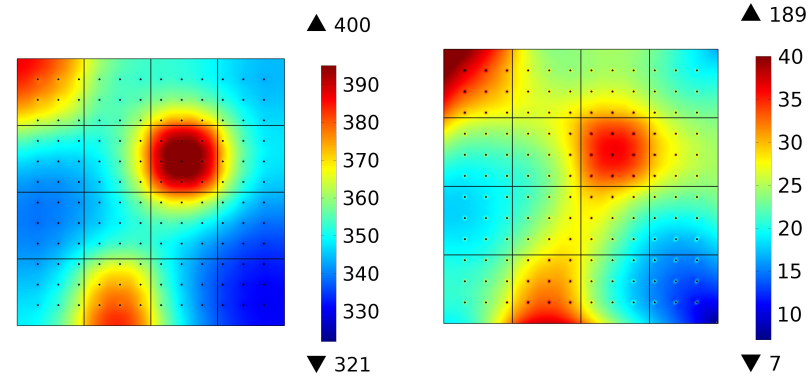


- Core # increases: tens or more cores on a single die.
- Difficult to coordinate cores for best performance under thermal constraint.

The problem of 3D integration





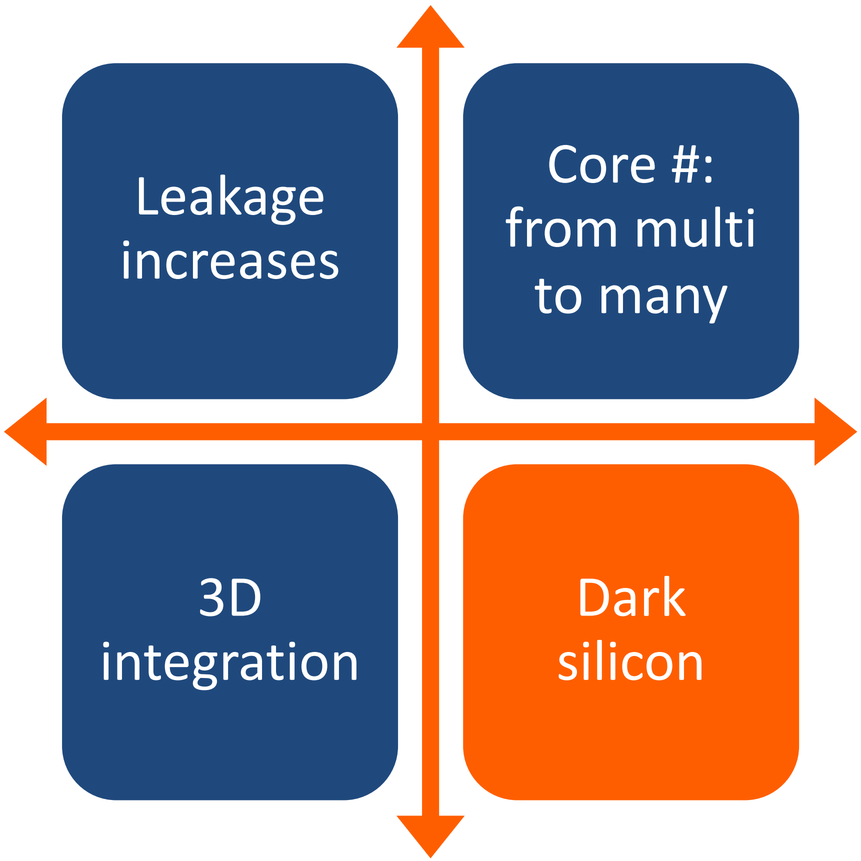


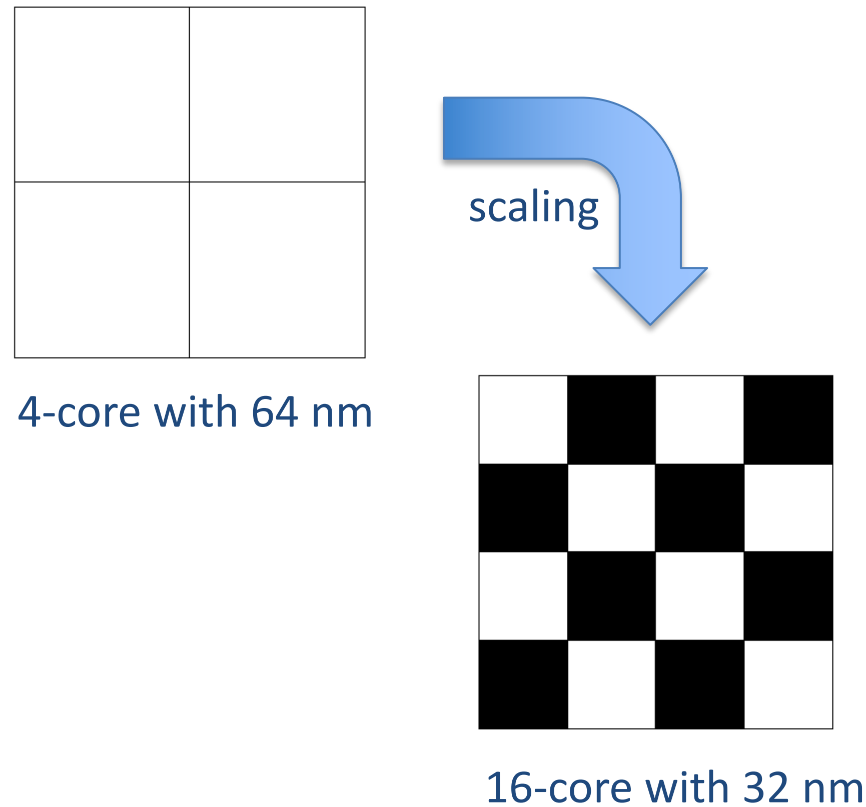
(a) Temperature (K) distribution.

(b) Von Mises thermal stress (MPa) distribution.

- 3D IC: go vertical for higher integration density.
- High power density leads to high temperature, large stress, and reliability issues.

The dark silicon hazard





- Not all cores can be on simultaneously anymore.
- When should which cores be on with how much power to achieve the best performance?

Outline

- Leakage Matters:
  - Leakage-aware thermal estimation
    (IEEE Trans. on Computers, 2018)
  - Leakage-aware thermal management (white-box model)
    (ASP-DAC Best Paper Nomination, 2019)
  - Artificial neural network solution (echo state network)
    (IEEE Trans. on CAD of Integrated Circuits and Systems, 2019)
- Many-Core Solutions:
  - Hierarchical thermal management
    (ACM Trans. on Design Automation of Electronic Systems, 2017)
- 3D Integration:
  - Runtime stress estimation using artificial neural network
    (International Symposium on Quality Electronic Design, 2016)
  - STREAM: Stress-aware reliability management
    (IEEE Trans. on CAD of Integrated Circuits and Systems, 2018)
- Dark Silicon Hazard:
  - GDP: Greedy based dynamic power budgeting
    (IEEE Trans. on Computers 2019)

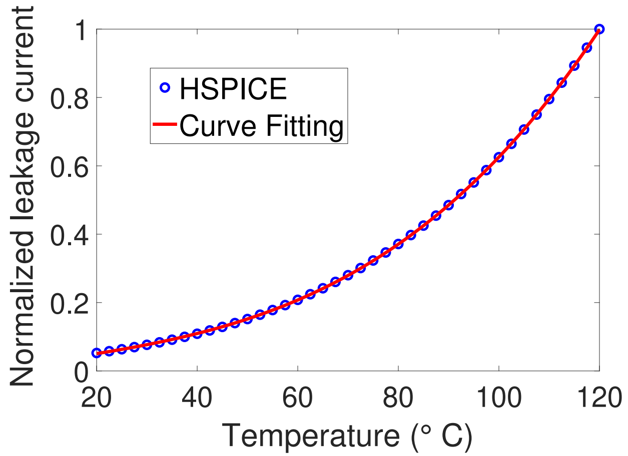
Leakage Matters

- Leakage-aware thermal estimation
  H. Wang, J. Wan, *et al.*, "A fast leakage-aware full-chip transient thermal estimation method", IEEE Trans. on Computers, 2018

- Leakage-aware thermal management
  - White-box model through PWL approximation
    X. Guo, H. Wang, *et al.*, "Leakage-aware thermal management for multi-core systems using piecewise linear model predictive control", ASP-DAC Best Paper Nomination, 2019
  - Artificial neural network (echo state network)
    H. Wang, X. Guo, *et al.*, "Leakage-aware predictive thermal management for multi-core systems using echo state network", IEEE Trans. on CAD of Integrated Circuits and Systems, 2019

Nonlinear leakage problem in thermal estimation

- Leakage power depends on temperature nonlinearly.



- Difficult to compute temperature
  - Initial guess and iteration needed to solve the nonlinear thermal model (white-box model)!

$$GT(t) + C\frac{dT(t)}{dt} = BP(T,t),$$
$$Y(t) = LT(t),$$

Piecewise linear based thermal estimation

- Build local linear thermal models by Taylor expansion

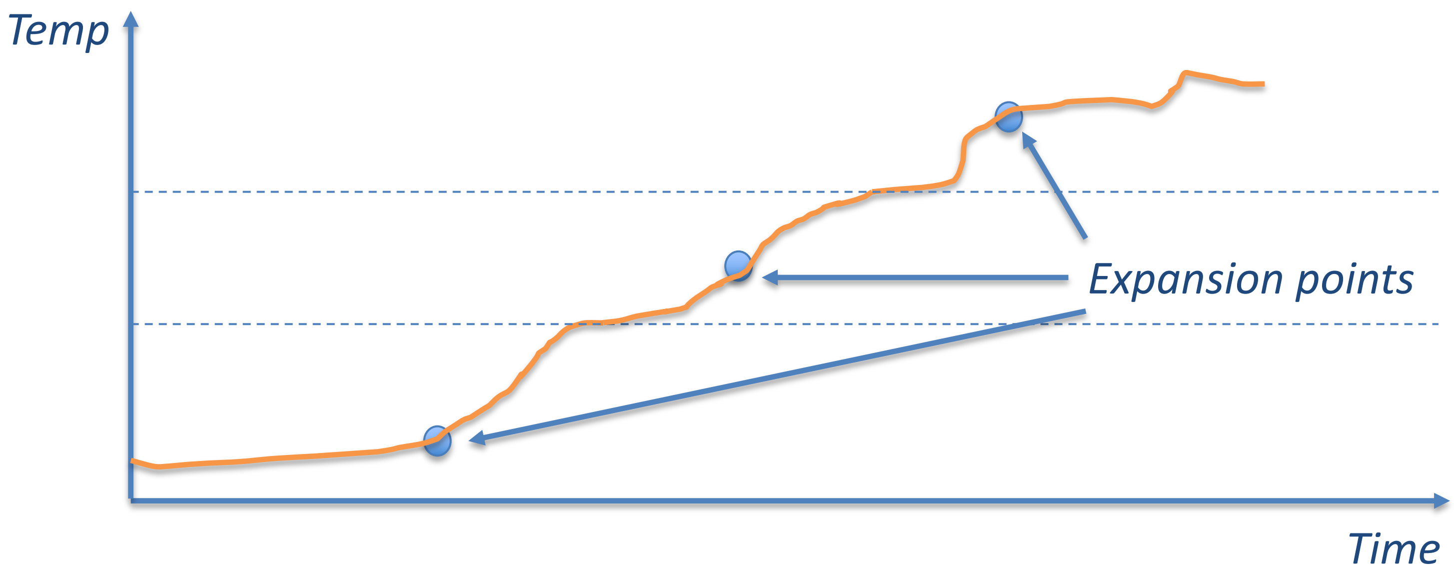
$$P_s = P_0 + A_s T,$$

$$G_l T(t) + C\frac{dT(t)}{dt} = B(P_d(t) + P_0),$$

$$Y(t) = LT(t).$$

$$G_l = G - BA_s$$

- Change Taylor expansion points on the fly



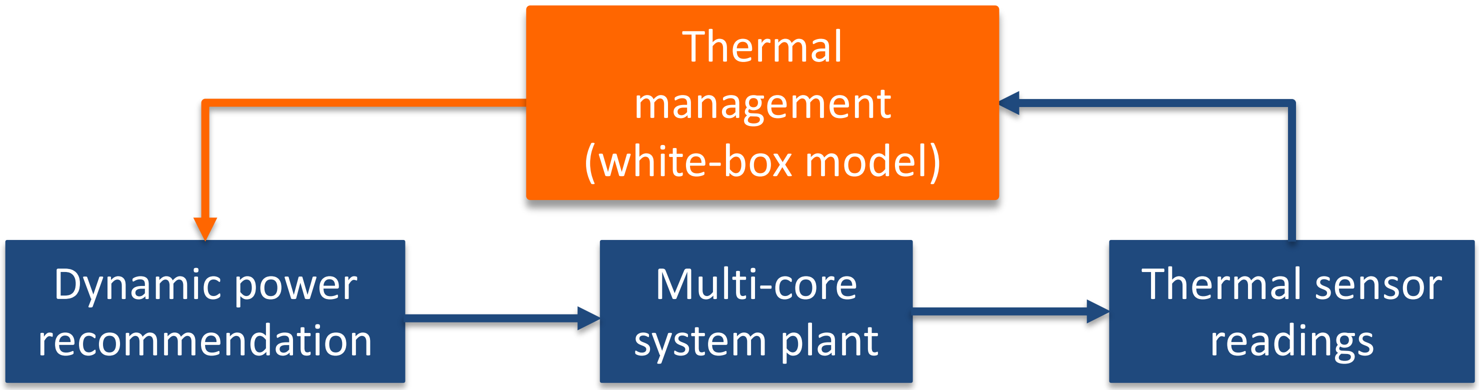
Leakage Matters

- Leakage-aware thermal estimation
  H. Wang, J. Wan, *et al.*, “A fast leakage-aware full-chip transient thermal estimation method”, IEEE Trans. on Computers, 2018

- Leakage-aware thermal management
  - White-box model through PWL approximation
    X. Guo, H. Wang, *et al.*, “Leakage-aware thermal management for multi-core systems using piecewise linear model predictive control”, ASP-DAC Best Paper Nomination, 2019
  - Artificial neural network (echo state network)
    H. Wang, X. Guo, *et al.*, “Leakage-aware predictive thermal management for multi-core systems using echo state network”, IEEE Trans. on CAD of Integrated Circuits and Systems, 2019

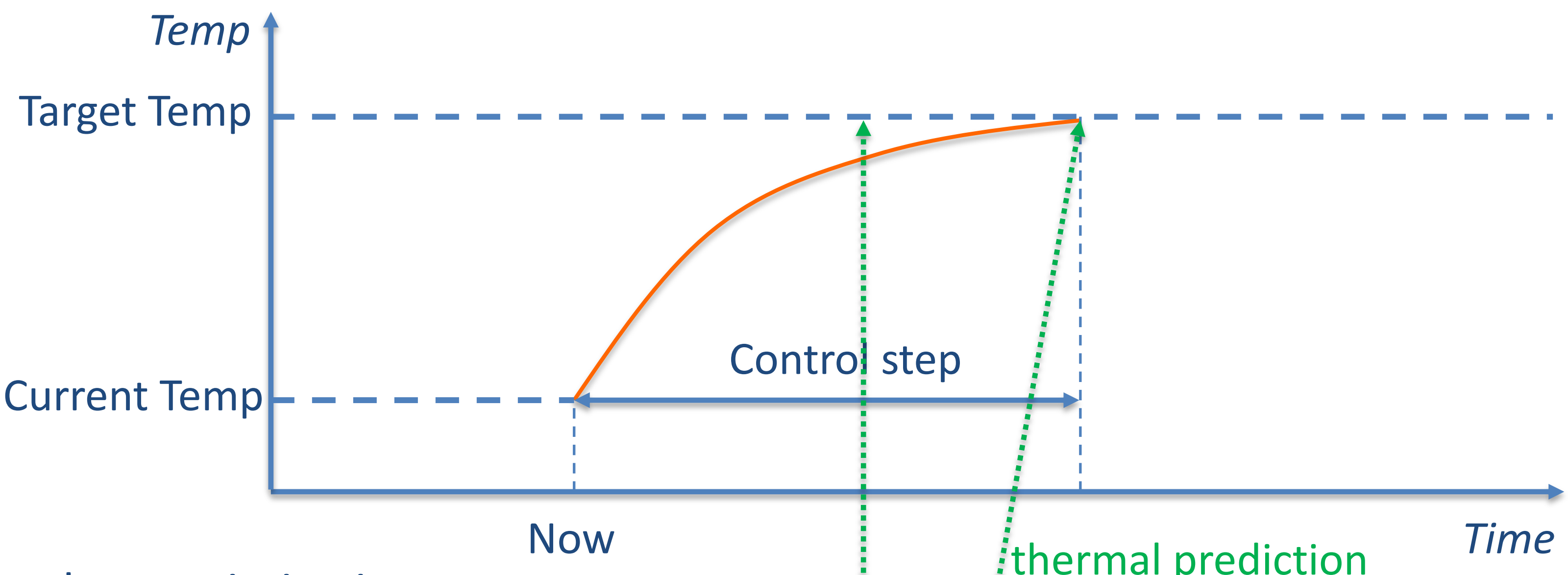
Leakage-aware thermal management problem

- Dynamic power is controllable
  - Change core’s V/f
  - Switch tasks by scheduling
- Leakage power is uncontrollable
  - Depends mainly on temperature
- How to compute the dynamic power recommendation in leakage-aware thermal management?



Basic framework of Predictive DTM

- The basic idea of predictive DTM
  - Compute the dynamic power recommendation P_d, which tracks the given target temperature
  - P_d can be solved by optimization using thermal prediction

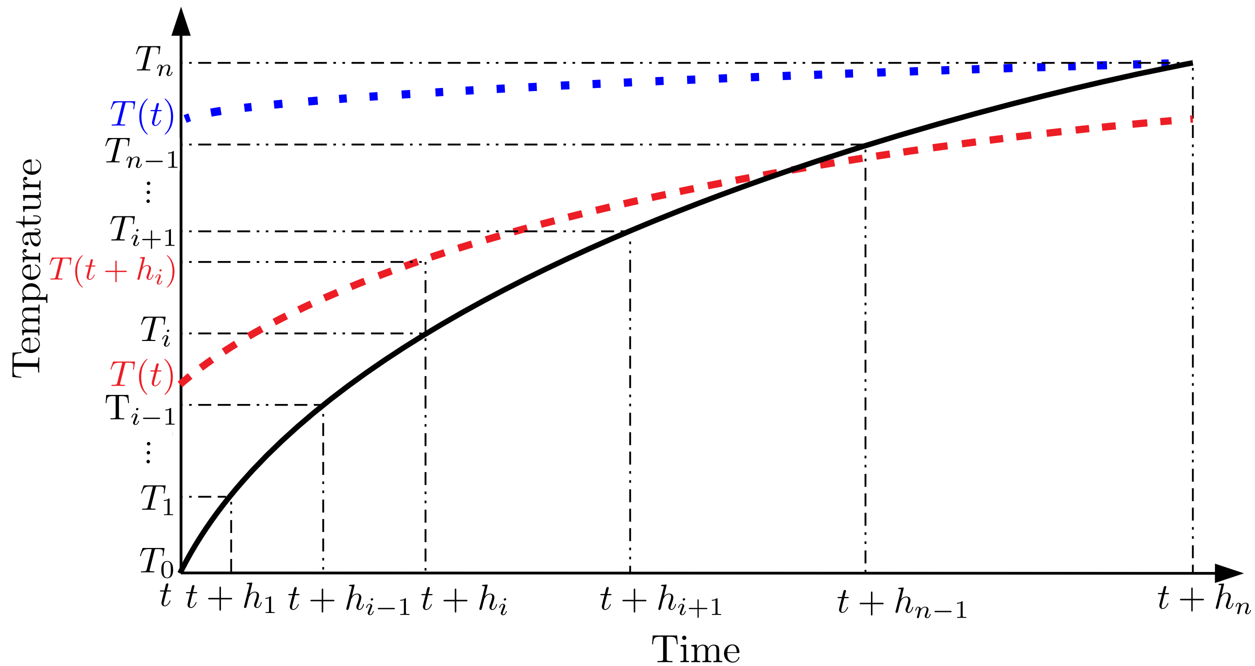


Formulate optimization using white-box thermal model

minimize $\mathcal{J} = (\mathcal{Y}_g - \mathcal{Y})^T(\mathcal{Y}_g - \mathcal{Y})$

Determine expansion points in thermal management

- Build PWL white-box thermal model for DTM
- A systematic way to choose Taylor expansion points
  - Simulate the extreme curve (black) to determine points
  - Normal curves (orange, blue) share the points of the extreme

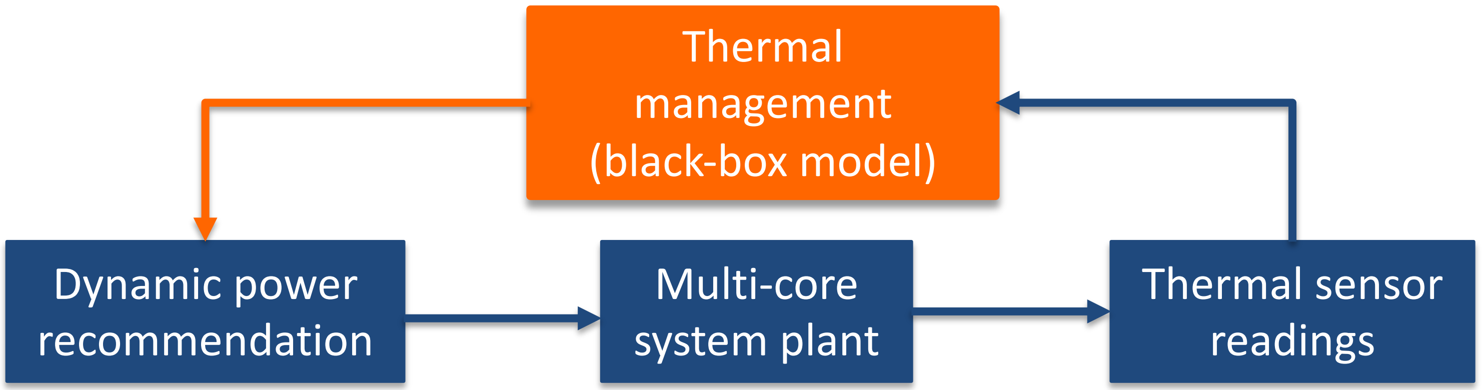


Leakage Matters

- Leakage-aware thermal estimation
  H. Wang, J. Wan, *et al.*, "A fast leakage-aware full-chip transient thermal estimation method", IEEE Trans. on Computers, 2018
- Leakage-aware thermal management
  - White-box model through PWL approximation
    X. Guo, H. Wang, *et al.*, "Leakage-aware thermal management for multi-core systems using piecewise linear model predictive control", ASP-DAC Best Paper Nomination, 2019
  - Artificial neural network (echo state network)
    H. Wang, X. Guo, *et al.*, "Leakage-aware predictive thermal management for multi-core systems using echo state network", IEEE Trans. on CAD of Integrated Circuits and Systems, 2019

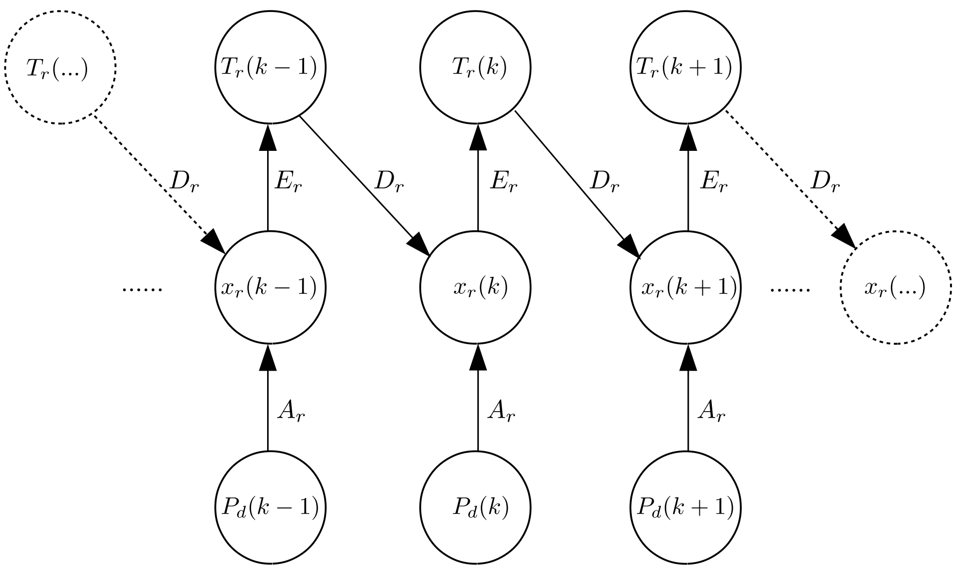
Using black-box model for DTM

- When detailed structure unavailable
  - Build black-box thermal model
  - Training using input (power) and output (temp.) pairs
- Remarks
  - Input should be dynamic power
  - Model should be nonlinear
  - Leakage handled implicitly inside model

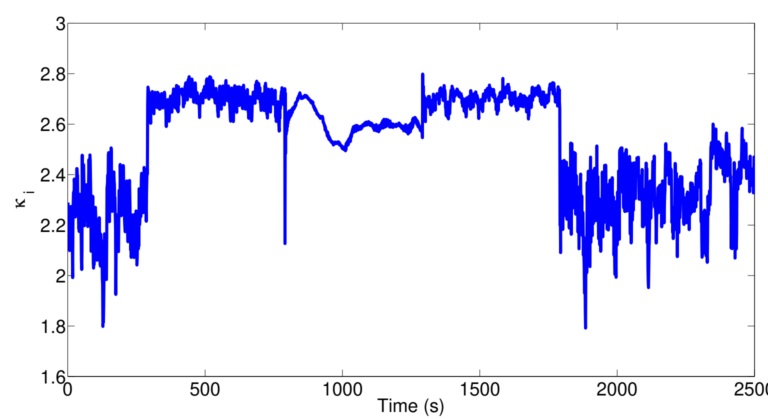


First try (failed): RNN based model

- Using recurrent neural network (RNN)
  - Nonlinear model specially for dynamic system modeling
  - Training using back propagation through time (BPTT)
  - First try failed! Due to exploding gradient in training
  - Large error using RNN



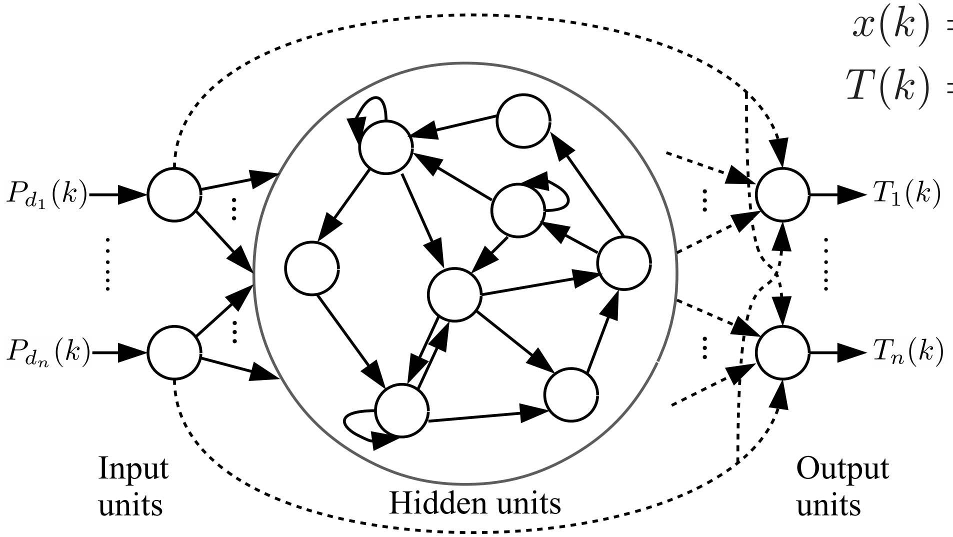
$$x_r(k) = f(A_r P_d(k) + D_r T_r(k-1) + \alpha),$$
$$T_r(k) = E_r x_r(k) + \beta,$$



Singular value > 1: exploding gradient

ESN to avoid exploding gradient

- Echo State Network (ESN) is a special RNN
  - Fixing the recurrent weights in hidden units
  - Only train the input and output weights
  - Training does not propagate through time (vs. BPTT)
  - Good accuracy in leakage-aware thermal modeling



$$x(k) = (1-\gamma)x(k-1) + \gamma f(AP_d(k) + Dx(k-1)),$$
$$T(k) = Ex(k) + HP_d(k),$$

Simple training via least square,
No exploding gradient problem:

$$S = \begin{bmatrix} x(1), x(2), \ldots, x(n_k) \\ P_{tr}(1), P_{tr}(2), \ldots, P_{tr}(n_k) \end{bmatrix}^T$$

$$O = [T_{tr}(1), T_{tr}(2), \ldots, T_{tr}(n_k)]^T$$

$$W_{out} = (S^\dagger O)^T$$

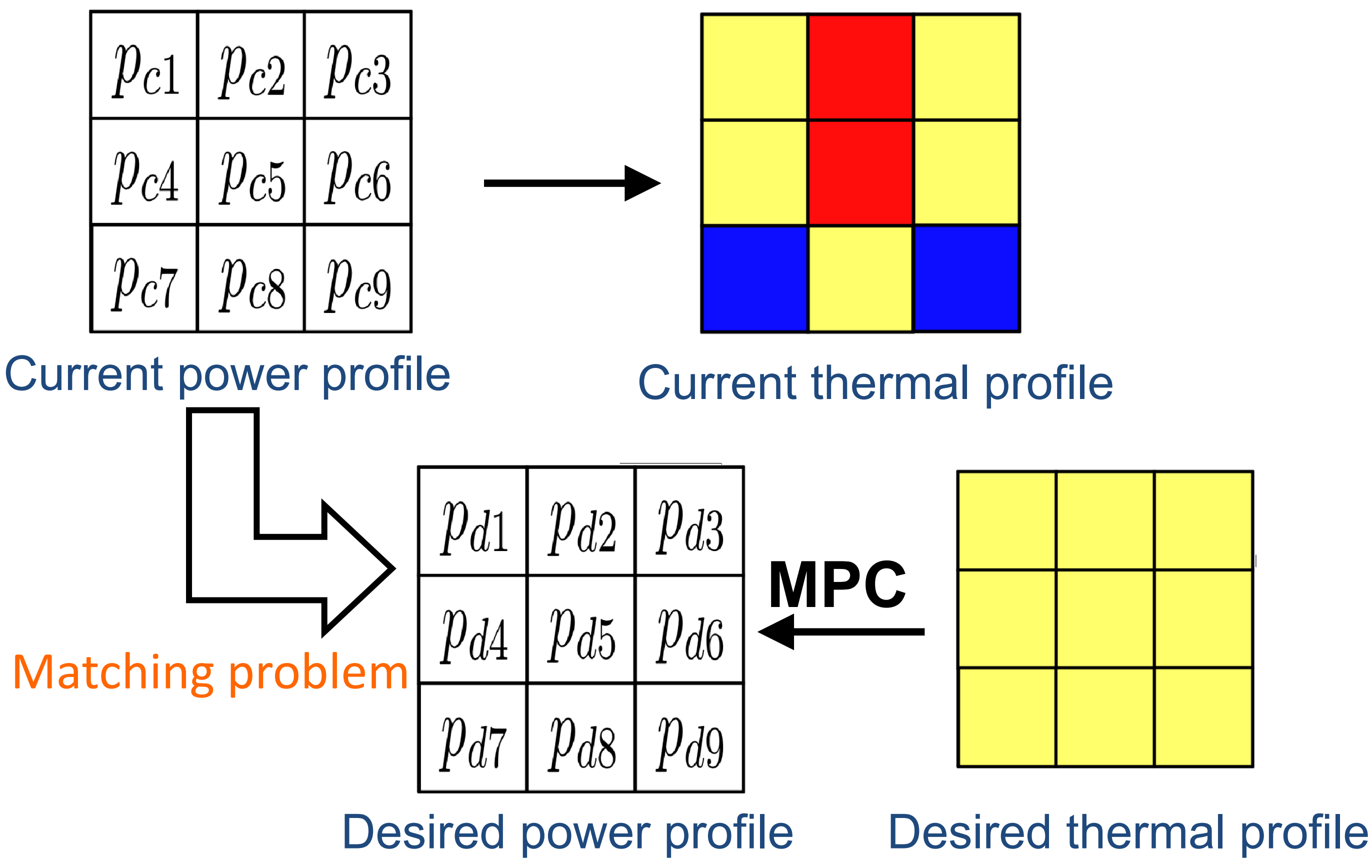
Many-Core Solutions

- Hierarchical thermal management

  H. Wang, J. Ma, *et al.*, “Hierarchical dynamic thermal management method for high-performance many-core microprocessors”, ACM Trans. on Design Automation of Electronic Systems, 2017

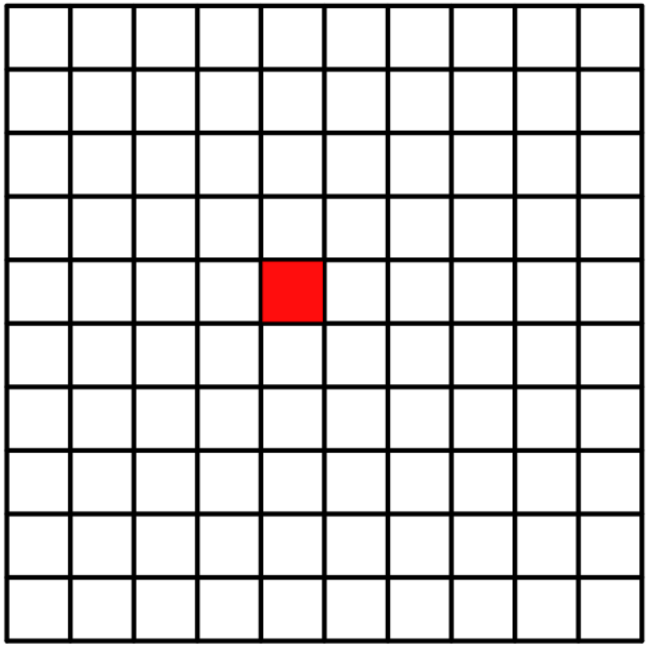
Model predictive control in thermal management

- We want to match the desired power profile using current power profile, by using task migration and DVFS.



The many-core system DTM problem

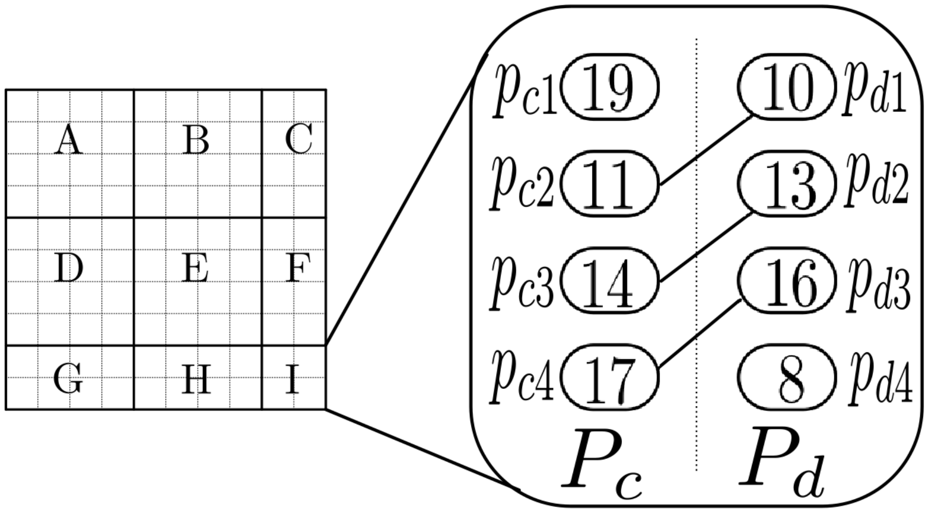
- Computing time increases as core number increases
- Large control delay reduces efficiency



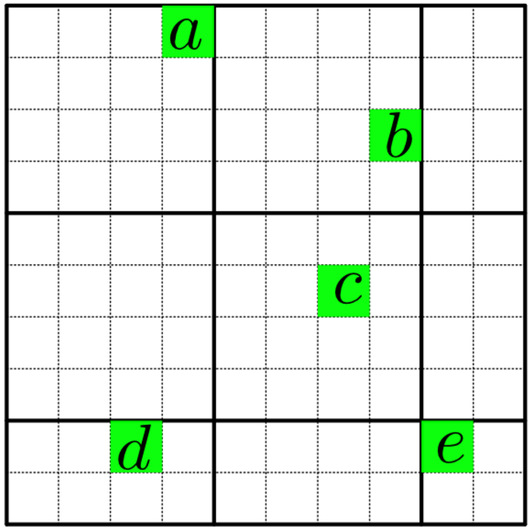
An example of 100-core chip, assuming core in red is in charge of the DTM computing.

Two-level Hierarchical method

- Lower level matching
  - Simply group spatially adjacent cores into blocks.
  - Do matching inside each block (intra block)
- Upper level matching
  - Do Matching using lower level unmatched ones (inter block)



Lower level matching



Upper level matching

3-D Integration

- Runtime stress estimation using ANN

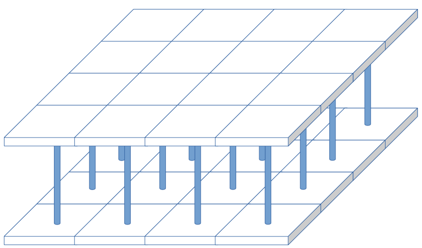
  L. Zhang, H. Wang, S. Tan, “Fast Stress Analysis for Runtime Reliability Enhancement of 3D IC Using Artificial Neural Network”, International Symposium on Quality Electronic Design, 2016

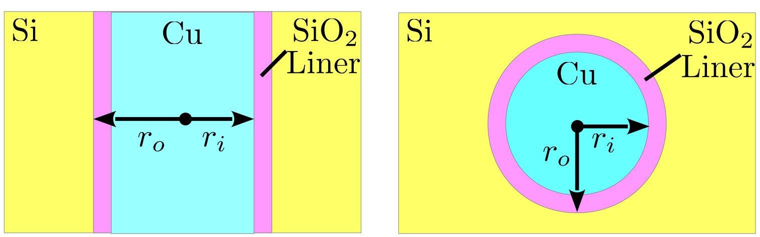
- STREAM: Stress-aware reliability management

  H. Wang, D. Huang, *et al.*, “STREAM: Stress and thermal aware reliability management for 3D-ICs”, IEEE Trans. on CAD of Integrated Circuits and Systems, 2018

Stress problem in 3D IC

- Stress is significant around Through silicon via (TSV)
- Stress changes with temperature in space and time
- Temperature changes significantly in multi-core systems
- Runtime stress estimation needed

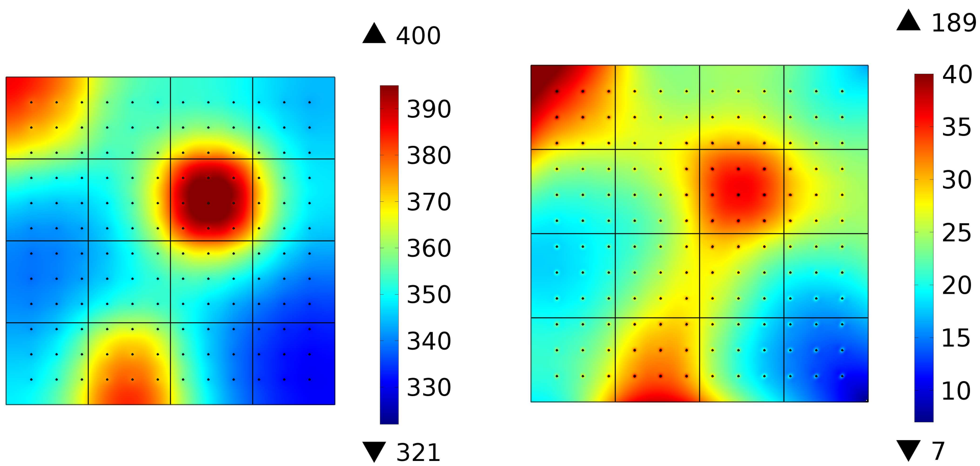




(a) Cross-section view.

(b) Longitudinal-section view.

A 3D IC (up) with its TSV structure (down)



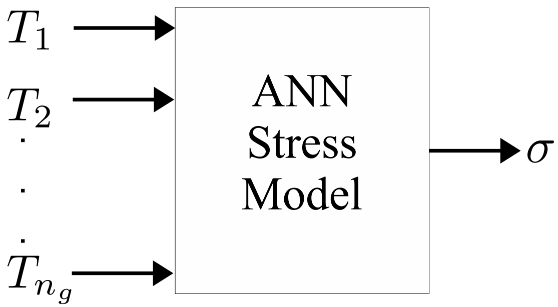
(a) Temperature (K) distribution.

(b) Von Mises thermal stress (MPa) distribution.

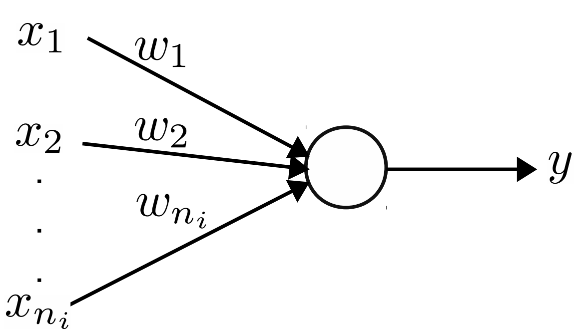
Stress changes with temperature

Framework of ANN stress model

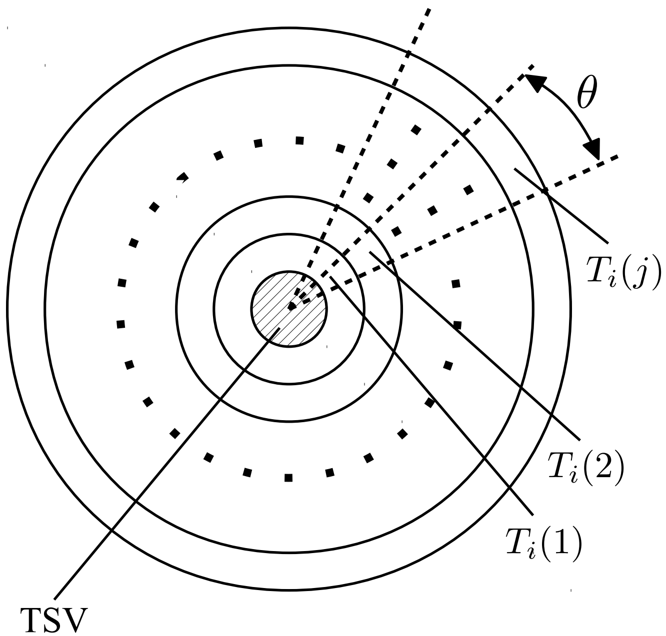
- Input: temperatures around each TSV
- Output: maximum stress
- Inside: neurals with different connections



ANN stress model framework



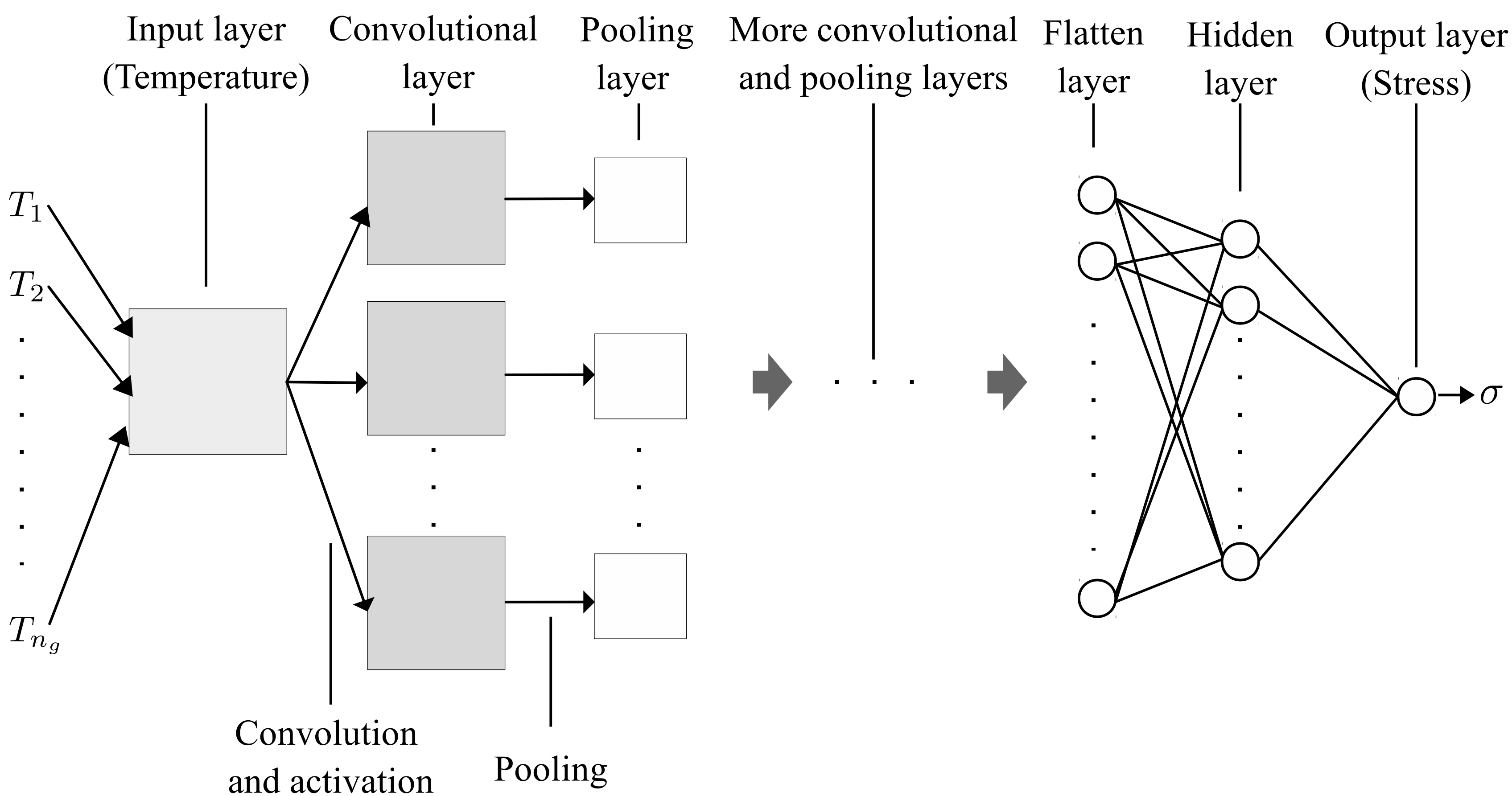
Neural inside ANN stress model



Model input: temperatures around each TSV

Example: CNN stress model

- Different neural connections leads to different models
- CNN stress model works best in our test



3-D Integration

- Runtime stress estimation using ANN

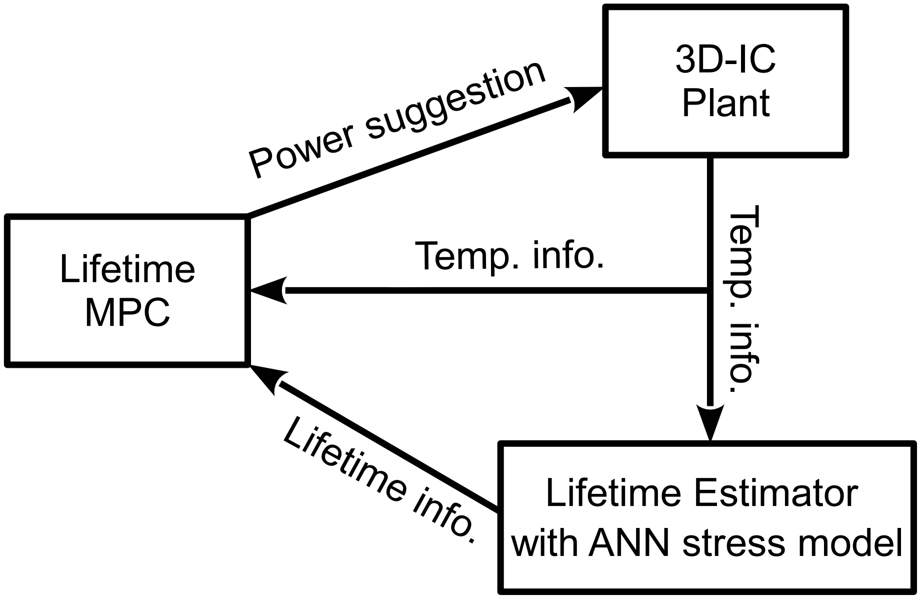
  L. Zhang, H. Wang, S. Tan, "Fast Stress Analysis for Runtime Reliability Enhancement of 3D IC Using Artificial Neural Network", International Symposium on Quality Electronic Design, 2016

- STREAM: Stress-aware reliability management

  H. Wang, D. Huang, *et al.*, "STREAM: Stress and thermal aware reliability management for 3D-ICs", IEEE Trans. on CAD of Integrated Circuits and Systems, 2018

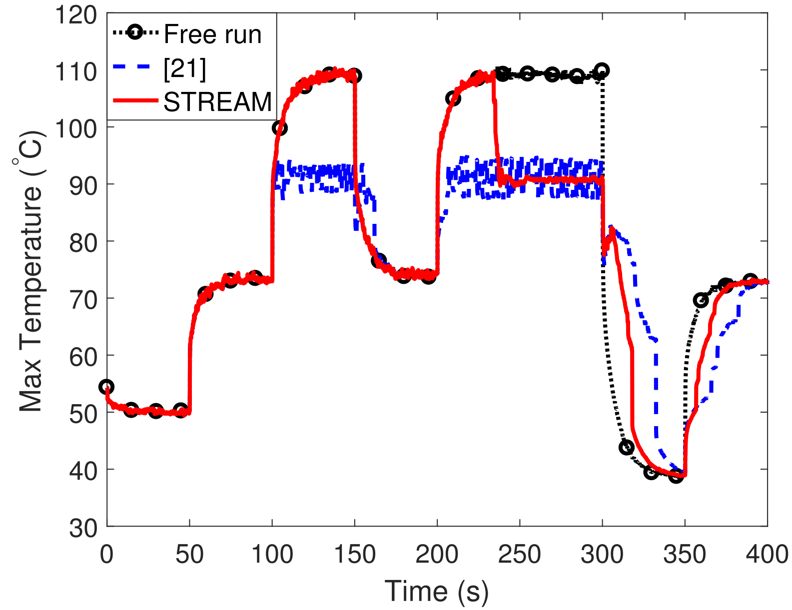
Boost 3D IC performance with ANN stress model

- We can estimate 3D IC lifetime with ANN stress model
- When the expected lifetime is
  - longer than designed: boost performance
  - shorter than designed: limit performance

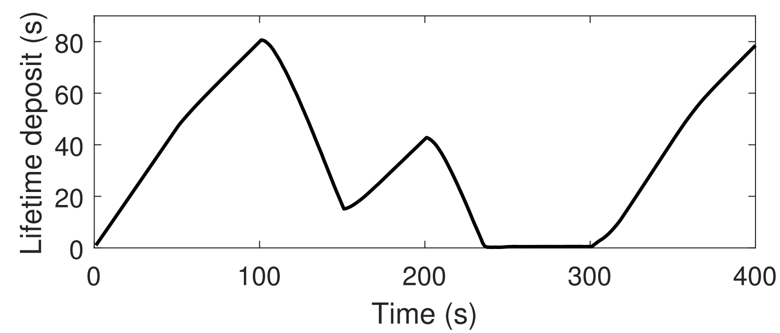


Lifetime banking with lifetime MPC

- Lifetime banking
  - Deposit lifetime
  - Consume lifetime
  - Lifetime deposit should never be negative
- Lifetime model predictive control (MPC)
  - Compute the power recommendation for 3D IC
  - DVFS performed to match the power recommendation



(a) Max temperature of synthetic workload with STREAM, existing method [21] and free run without any reliability management.



(b) Lifetime deposit information of STREAM.

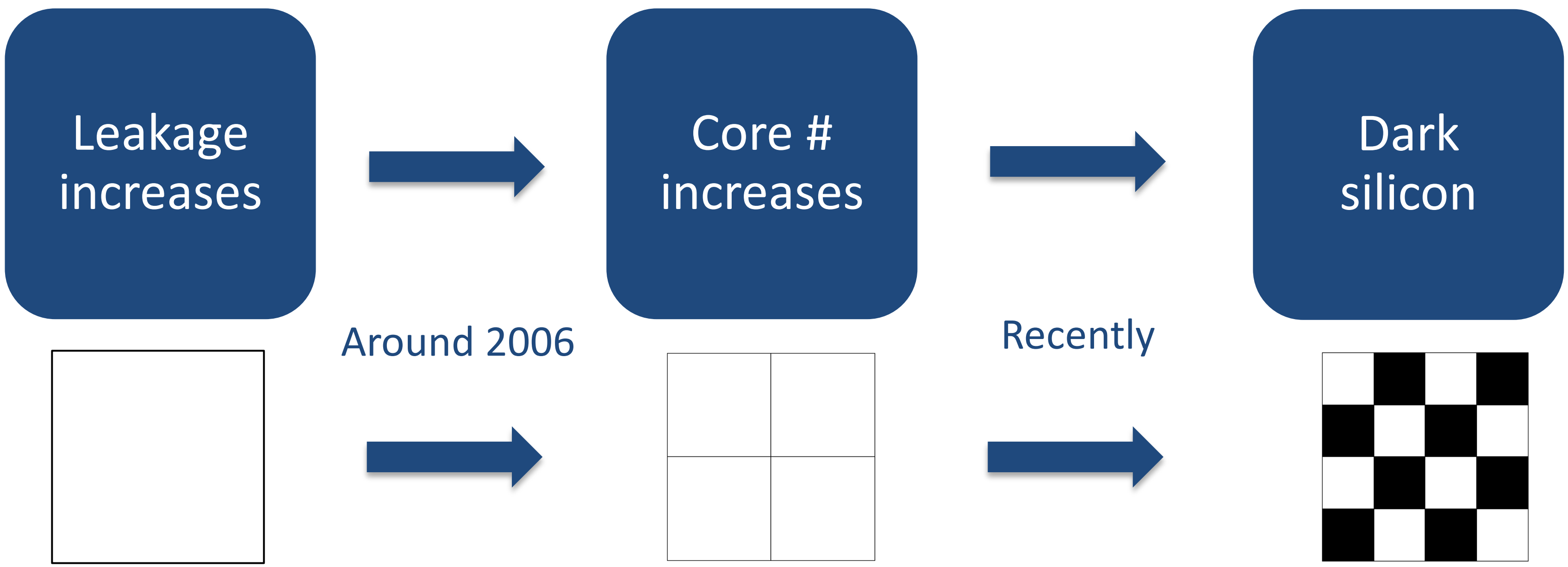
Dark Silicon Hazard

- GDP: Greedy based dynamic power budgeting

  H. Wang, D. Tang, M. Zhang, *et al.*, "GDP: A greedy based dynamic power budgeting method for multi/many-core systems in dark silicon", IEEE Trans. on Computers, 2019

Two battles lost against leakage

- Leakage power does not scale like dynamic power
  - Power density increases with scaling (Dennard scaling lost)
- Power (heat) removal ability remains the same



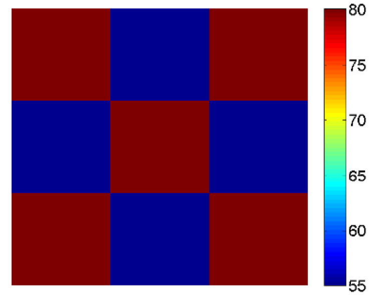
Fix core #
Increase frequency
Best days in performance increase!

Fix frequency
Increase core #

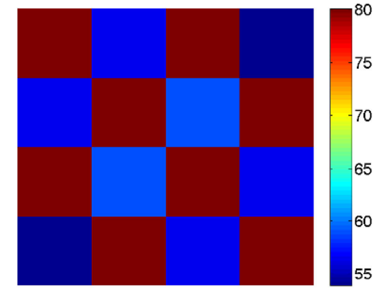
Not all cores operates
@ full freq anymore
We lost Dennard scaling
Solutions needed!

Power budgeting for dark silicon

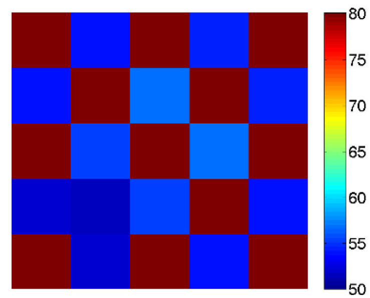
- Activating different cores leads to different power budget
- How to determine the active core distributions and power budget?
- Our solution: Greedy Dynamic Power (GDP)
  - Locate active core positions at runtime
  - Compute power budget for each core



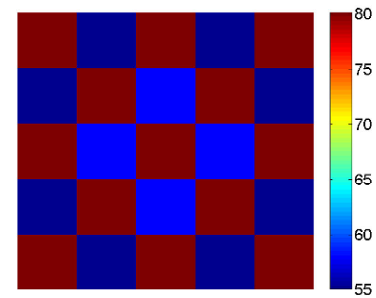
(a) 9-core system with 5 active cores.



(b) 16-core system with 8 active cores.



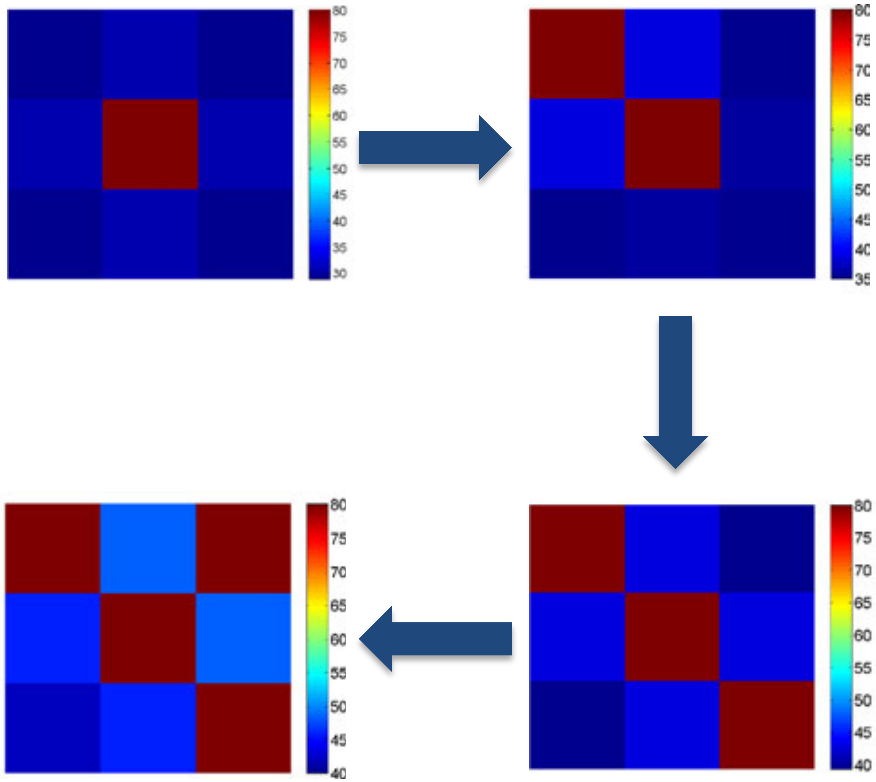
(c) 25-core system with 12 active cores.



(d) 25-core system with 13 active cores.

The greedy iteration in GDP

- Searching for the best distribution is expensive
- Search the local best one instead!
  - Locate the first best one and fix its position
  - Search for the second best one and fix its position
  - Continue this greedy iteration
- Transient temp. effects considered at runtime



9-core system's first 4 GDP iterations

Thank you!