Optimal Topology Exploration for Application-Specific 3D Architectures

O. Ozturk, F. Wang, M. Kandemir, and Y. Xie

Pennsylvania State University

Outline

- Introduction
- 3D Thermal Model
- ILP Formulation of Application-Specific 3D Placement
- An Example
- Experimental Evaluation
- Conclusion

Introduction

- 3D ICs: multiple device layers stacked together with direct vertical interconnects tunneling through them
- Advantages:
  - Reduction on global interconnect
  - Higher packing density and smaller footprint
  - Lower interconnect power due to reduction in total wiring length
  - Support for realization of mixed-technology chips

3D ICs

- Thermal Issues:
  - Higher cooling/packaging costs
  - Acceleration of failure mechanisms
  - Performance degradation.
- Thermal issues even more pronounced for 3D
  - Higher packing density
  - Especially for the inner layer of the die
  - A major hindrance for 3D integration
- 3D integration: Need to be a thermal-aware design

Chip Multiprocessors

- Chip multiprocessor (CMP):
  - AMD Opteron, IBM Power5 and Intel Yonah
  - 2 cores now, soon will have 4, 8, 16, or 32
- Promising for embedded systems:
  - Performance: increasingly difficult to obtain more performance out of single-processor
  - Power consumption: lower frequency
  - Scalability: both loop-level and instruction-level parallelism
  - Cost: simpler design and verification
  - Area: better utilization of the available silicon area

3D CMPs: Placement of processors and storage blocks

- Placement of processors and storage blocks
  - Determine the data communication distances
  - Both power and performance depend on data communication distances
  - Frequently accessed data storage blocks should be placed close to the processor
  - Data block shared between two processors should be put close to both

Application-specific Placement in a Customized 3D Design

- Application-specific
  - Each embedded application can require a different placement for achieving the minimum data communication distances
- Our approach:
  - Integer linear programming (ILP) based placement
  - Constraints: thermal bounds
  - Objective: minimize data communication distances

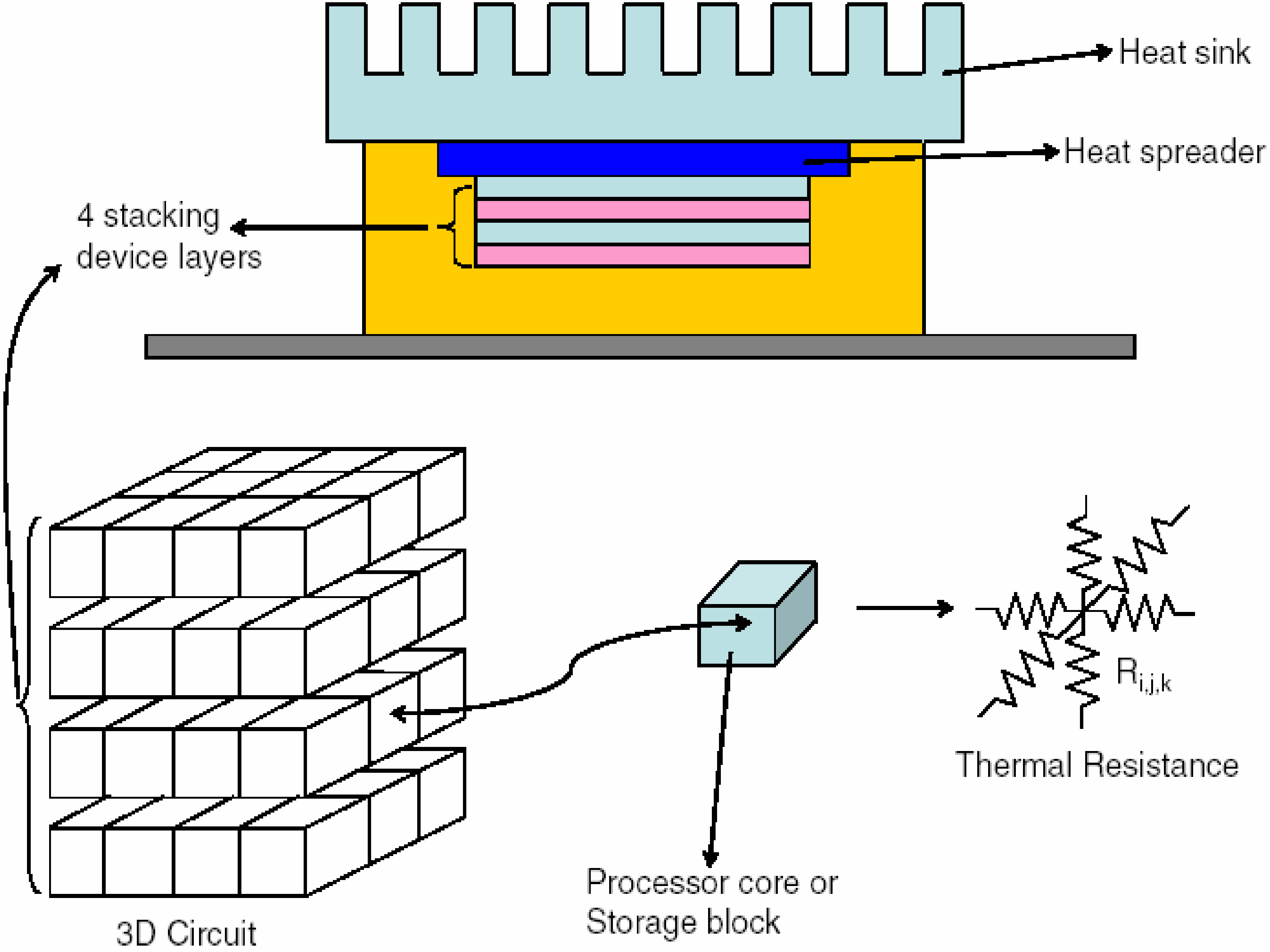
3D Thermal model

- An 3D resistor mesh model
  - Based on Skadron’s Hotspot thermal model (lumped thermal resistances and thermal capacitances)
  - Employs thermal-electrical duality to enable effcient computation of thermal effects at the functional block level
- Transfer thermal resistance $R_{i,j}$ of block i with respect to block j

$$R_{i,j} = \frac{\Delta T_{i,j}}{\Delta P_{i,j}}.$$

- Temperature rise for each block

$$T = R \times P$$


Heat sink
Heat spreader
4 stacking device layers
3D Circuit
Processor core or Storage block
$R_{i,j,k}$
Thermal Resistance

ILP Formulation of Application-Specific 3D Placement

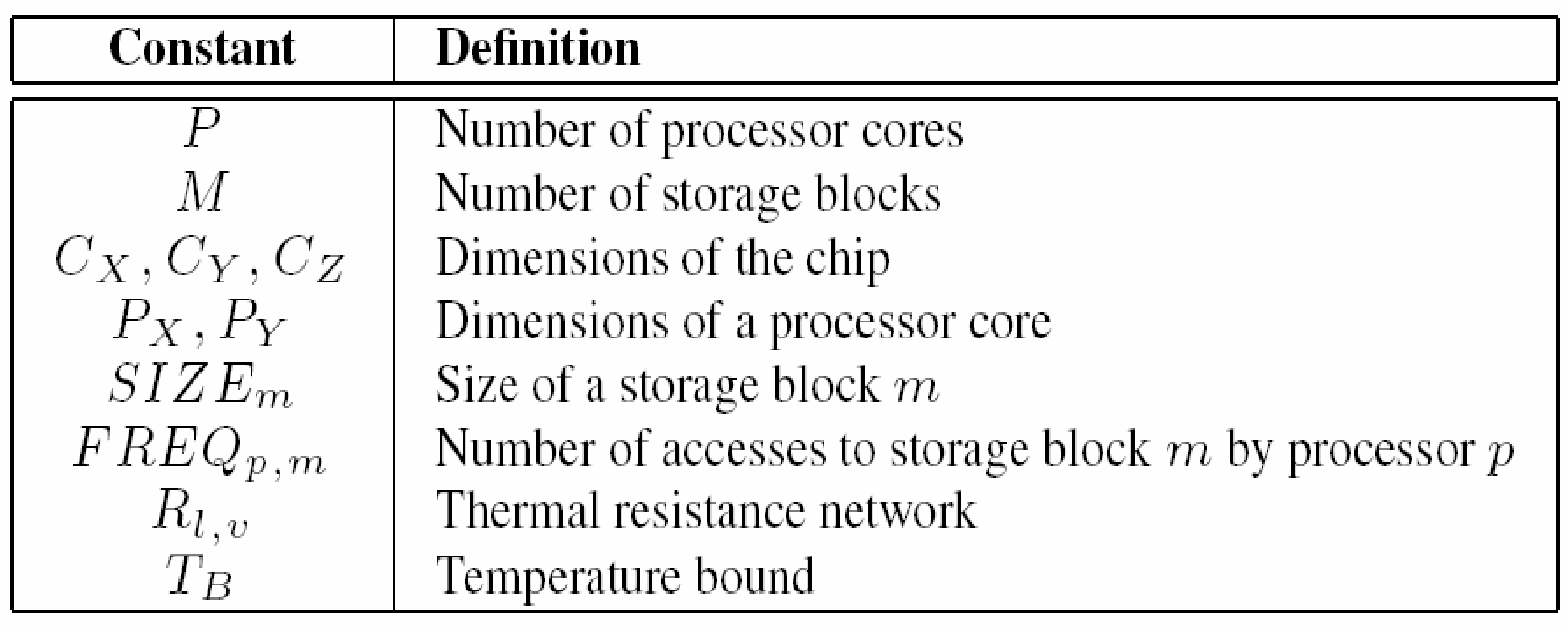
- Problem: Minimize data communication cost of a given application by determining the optimal placement of storage blocks and processor cores under a temperature bound
- A storage block corresponds to a set of consecutive cache lines
    - Data cache assumed to be divided into storage blocks of equal size
- In ILP formulation, we view the chip area as a 3D grid and assign processor cores and storage blocks into this grid

ILP Formulation

- ILP provides a set of techniques that solve optimization problems:
  - Objective function and constraints are linear functions
  - Solution variables restricted to be integers.
- In 0-1 ILP
  - Each (solution) variable is restricted to be 0 or 1.
- 0-1 ILP is used in this work for determining:
  - Storage block placements
  - Processor core placements
  - Under temperature bounds

ILP Formulation of Application-Specific 3D Placement

- Constant terms definition



Constant	Definition
P	Number of processor cores
M	Number of storage blocks
C_X, C_Y, C_Z	Dimensions of the chip
P_X, P_Y	Dimensions of a processor core
$SIZE_m$	Size of a storage block m
$FREQ_{p,m}$	Number of accesses to storage block m by processor p
$R_{l,v}$	Thermal resistance network
T_B	Temperature bound

Major Constraint Functions

- $MC_{m,x,y,z}$: indicates whether storage block m is in (x,y,z)
- $MD_{m,x,y}$: indicates whether storage block m has dimensions of (x,y)

Geometric:

$$\sum_{i=0}^{C_X-1}\sum_{j=0}^{C_Y-1}\sum_{k=0}^{C_Z-1} MC_{m,i,j,k} = 1, \quad \forall m.$$

$$\sum_{i=1}^{C_X}\sum_{j=1}^{C_Y} MD_{m,i,j} = 1, \quad \forall m.$$

$$SIZE_m = \sum_{i=1}^{C_X}\sum_{j=1}^{C_Y} MD_{m,i,j} \times i \times j, \quad \forall m.$$

Thermal:

$$Temp_m = \sum_{j=1}^{P+M+1} R_{m,j} \times Power_j, \quad \forall m.$$

$$Temp_m \leq T_B, \quad \forall m.$$

Objective Function

- Xdist$_{p,m,x}$: indicates whether the distance between processor p and sotrage block m is equal to x on the x-axis

$$X_{Cost} = \sum_{i=1}^{P} \sum_{j=1}^{M} \sum_{k=1}^{C_X - 1} FREQ_{i,j} \times Xdist_{i,j,k} \times k.$$

$$Y_{Cost} = \sum_{i=1}^{P} \sum_{j=1}^{M} \sum_{k=1}^{C_Y - 1} FREQ_{i,j} \times Ydist_{i,j,k} \times k.$$

$$Z_{Cost} = \sum_{i=1}^{P} \sum_{j=1}^{M} \sum_{k=1}^{C_Z - 1} FREQ_{i,j} \times Zdist_{i,j,k} \times k.$$

$$\boxed{\min \quad (\alpha \times (X_{Cost} + Y_{Cost}) + \beta \times Z_{Cost}).}$$

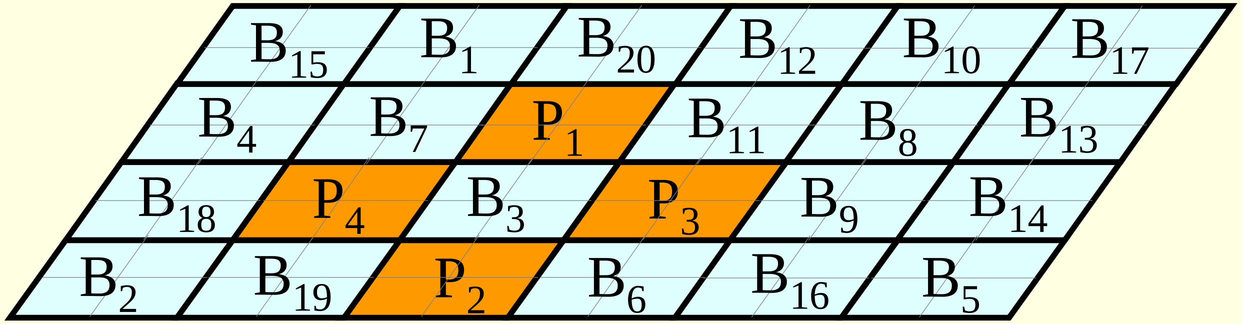
An Example

- 4 processors and 20 storage blocks

Processor	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
1	0.20%	0.18%	61.76%	0.19%	0.17%	0.20%	3.48%	2.86%	0.20%	0.20%
2	0.20%	0.19%	61.86%	0.19%	0.22%	4.07%	2.30%	0.18%	0.19%	0.18%
3	0.18%	0.18%	61.83%	0.18%	0.18%	4.05%	2.34%	0.19%	0.21%	0.19%
4	0.18%	0.22%	61.76%	0.19%	0.18%	4.05%	2.32%	0.18%	0.20%	0.18%
Processor	**B11**	**B12**	**B13**	**B14**	**B15**	**B16**	**B17**	**B18**	**B19**	**B20**
1	22.83%	0.22%	0.20%	0.18%	0.20%	0.19%	0.18%	0.19%	1.83%	4.55%
2	22.64%	0.20%	0.18%	0.18%	0.17%	0.19%	0.18%	1.91%	4.49%	0.28%
3	22.65%	0.20%	0.20%	0.22%	0.19%	0.20%	0.18%	1.89%	4.47%	0.29%
4	22.59%	0.17%	0.18%	0.18%	0.19%	0.21%	0.18%	1.89%	4.49%	0.31%

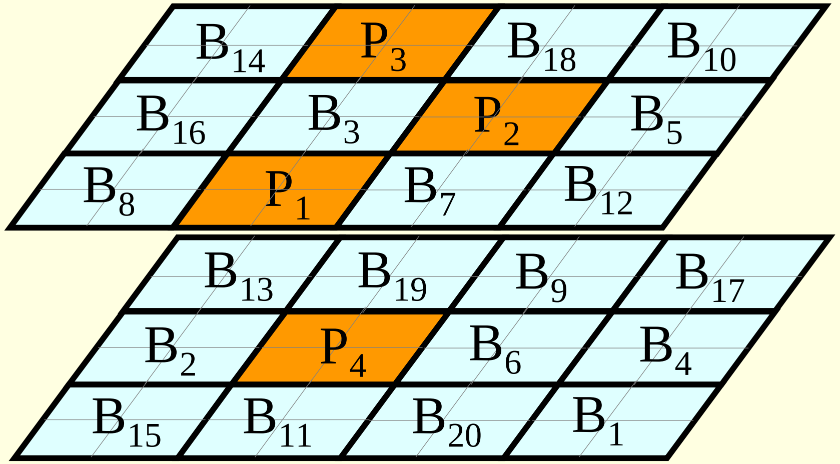
An Example

2D



B_{15}	B_{1}	B_{20}	B_{12}	B_{10}	B_{17}
B_{4}	B_{7}	P_{1}	B_{11}	B_{8}	B_{13}
B_{18}	P_{4}	B_{3}	P_{3}	B_{9}	B_{14}
B_{2}	B_{19}	P_{2}	B_{6}	B_{16}	B_{5}

3D



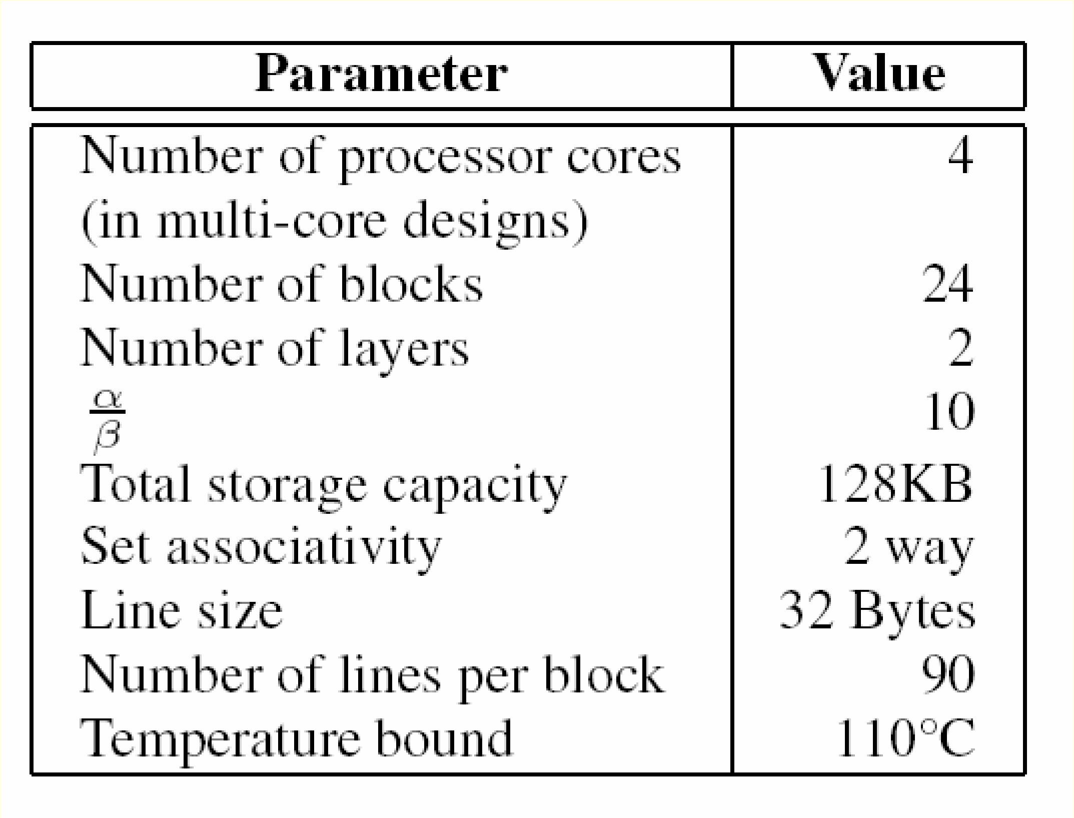
Layer 2

B_{14}	P_{3}	B_{18}	B_{10}
B_{16}	B_{3}	P_{2}	B_{5}
B_{8}	P_{1}	B_{7}	B_{12}

Layer 1

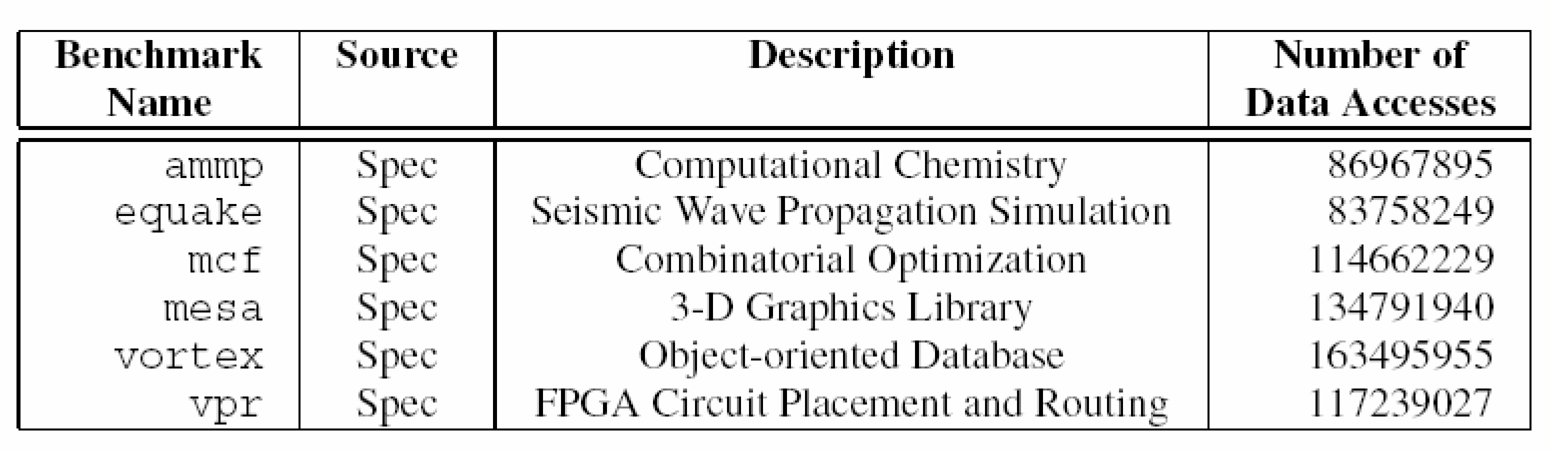
B_{13}	B_{19}	B_{9}	B_{17}
B_{2}	P_{4}	B_{6}	B_{4}
B_{15}	B_{11}	B_{20}	B_{1}

Experimental Parameters

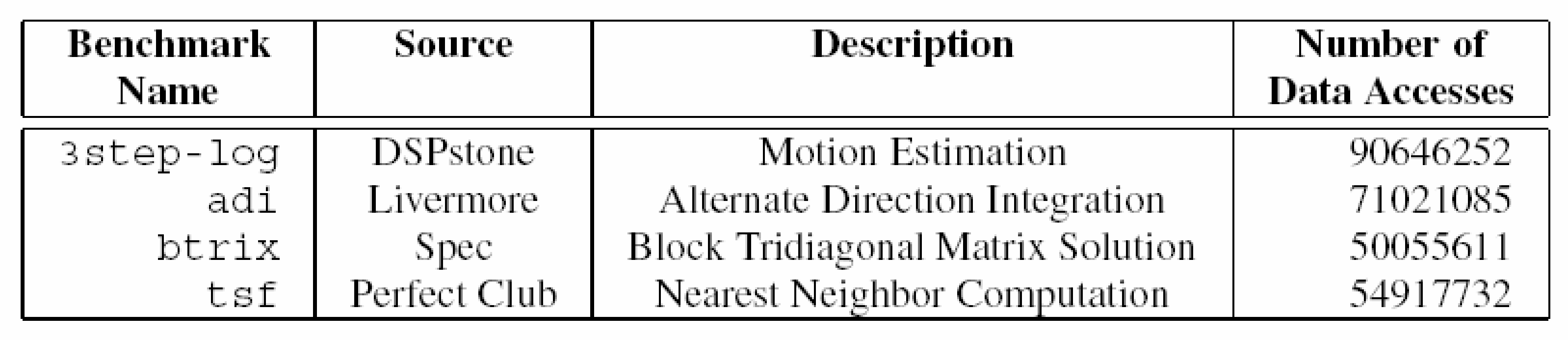


Parameter	Value
Number of processor cores (in multi-core designs)	4
Number of blocks	24
Number of layers	2
$\frac{\alpha}{\beta}$	10
Total storage capacity	128KB
Set associativity	2 way
Line size	32 Bytes
Number of lines per block	90
Temperature bound	110°C

Benchmarks



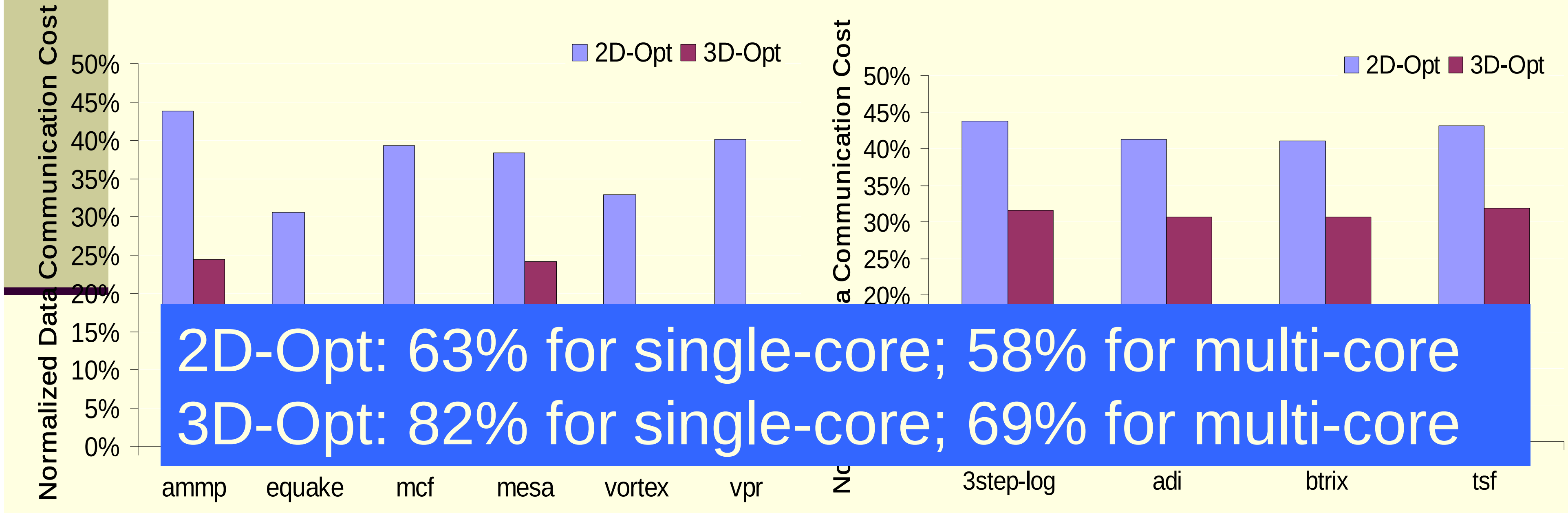
Benchmark Name	Source	Description	Number of Data Accesses
ammp	Spec	Computational Chemistry	86967895
equake	Spec	Seismic Wave Propagation Simulation	83758249
mcf	Spec	Combinatorial Optimization	114662229
mesa	Spec	3-D Graphics Library	134791940
vortex	Spec	Object-oriented Database	163495955
vpr	Spec	FPGA Circuit Placement and Routing	117239027



Benchmark Name	Source	Description	Number of Data Accesses
3step-log	DSPstone	Motion Estimation	90646252
adi	Livermore	Alternate Direction Integration	71021085
btrix	Spec	Block Tridiagonal Matrix Solution	50055611
tsf	Perfect Club	Nearest Neighbor Computation	54917732

Experimental Evaluation

- Normalized data communication cost of 2D-Opt and 3D-Opt w.r.t. 2D-Random



Single-Core

Multi-Core

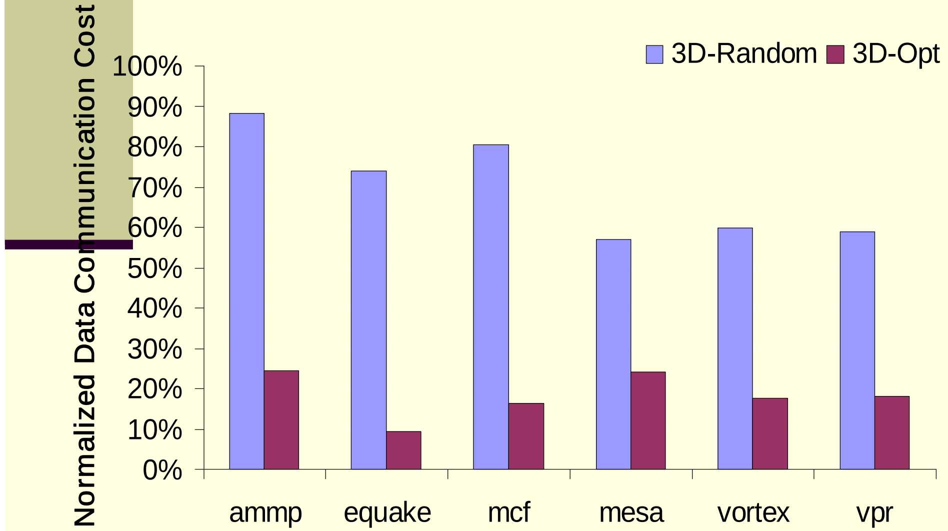
2D-Opt: 63% for single-core; 58% for multi-core

3D-Opt: 82% for single-core; 69% for multi-core

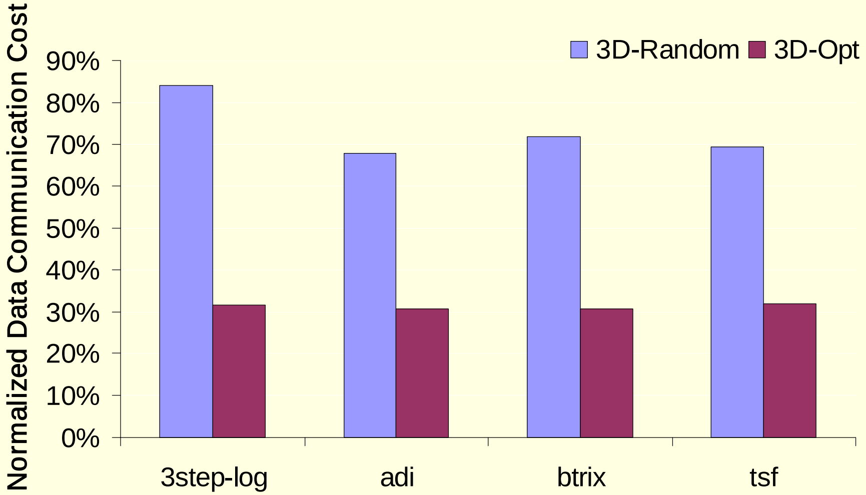
Experimental Evaluation

- Normalized data communication cost of 3D-Opt and 3D-Random w.r.t 2D-Random

Single-Core

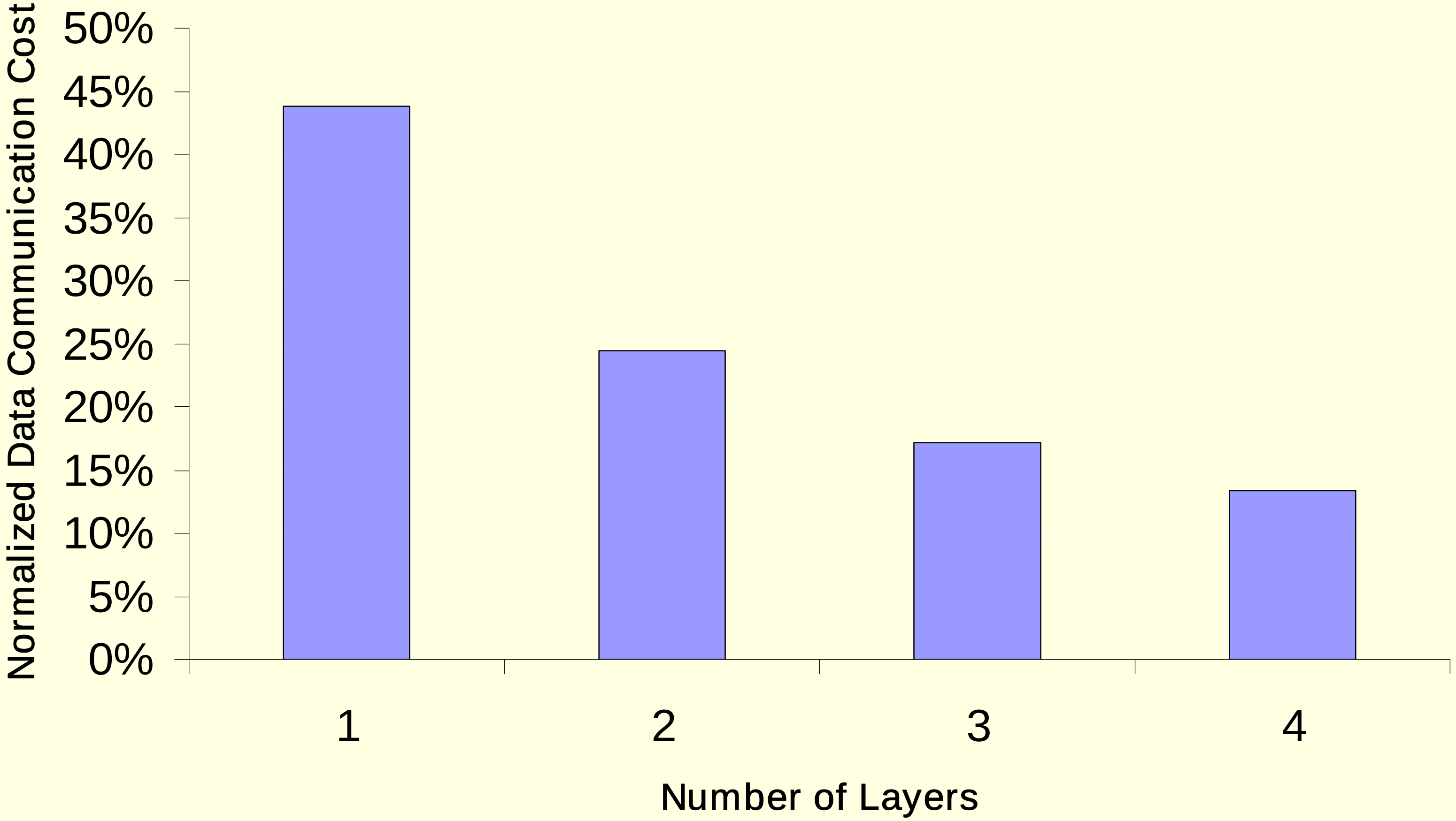


Multi-Core



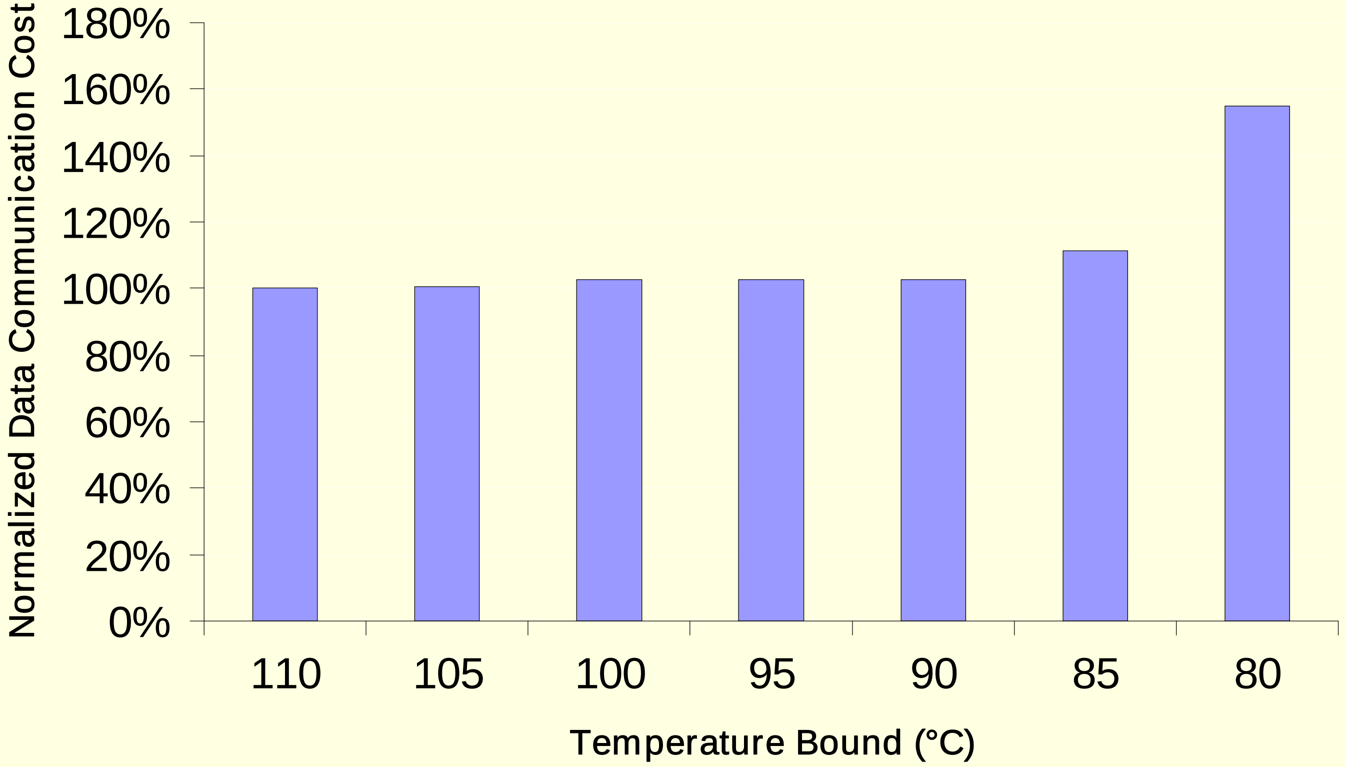
Experimental Evaluation

- Effect of number of 3D layers (ammp)



Experimental Evaluation

- Normalized data communication cost with respect to the temperature bound (ammp)
- Default: 110



Conclusion

- Shrinking process technology and increasing data communication requirements of embedded applications
  - An increasing bottleneck: On-chip interconnects
  - Solution to the global interconnect problem: 3D designs
- Our goal: application-specific placement of processor cores and storage blocks in a customized 3D design
- Formulated using ILP
- Experiments with single-core and multi-core
  - Optimal placement of storage blocks and processor cores is very important in 3D design

Thanks