Physical-Aware System-Level Design for Tiled Hierarchical Chip Multiprocessors

Jordi Cortadella, Javier de San Pedro,
Nikita Nikitin and Jordi Petit

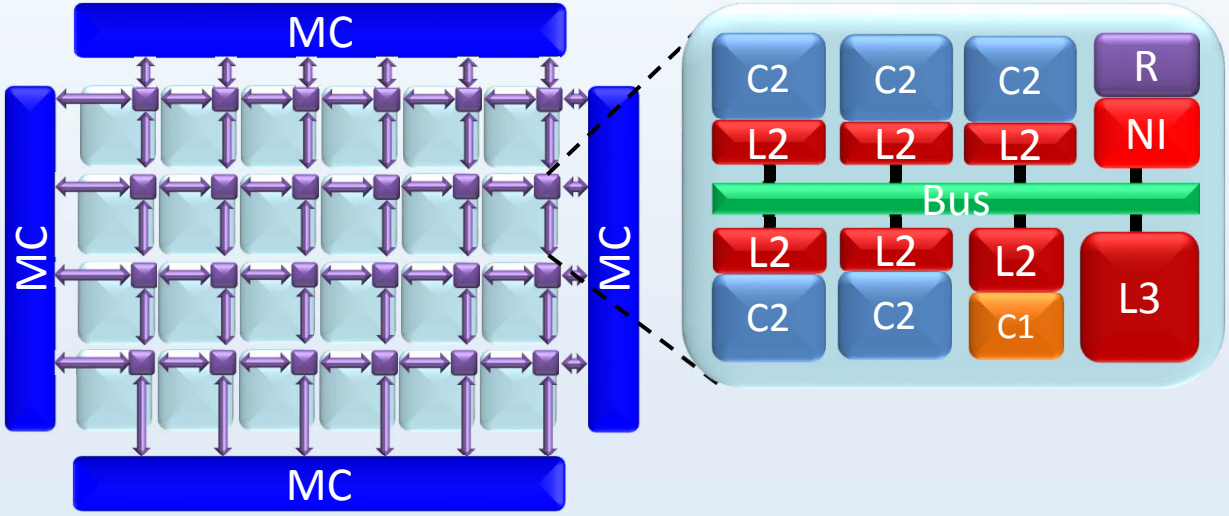
Universitat Politècnica de Catalunya (Barcelona)

Project funded by Intel Corp.

Designing a Chip Multiprocessor

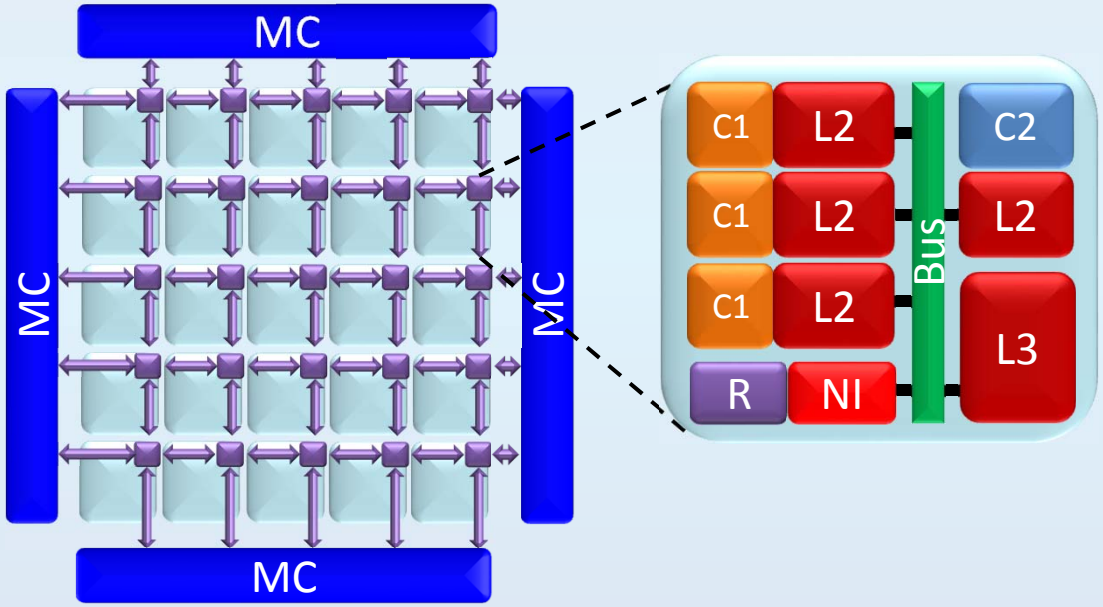
- How many cores?
- How much L2/L3 on-chip cache?
- Interconnect: mesh/ring/bus?
- How many memory controllers?

What is architectural exploration?



- 6x4 mesh, 24 clusters
- total 144 cores
- 6 cores/cluster
  - 1 C1, 128K L1, 256K L2
  - 5 C2, 64K L1, 96K L2
- 146 Mb total shared L3

Throughput = 85.71 IPC



- 5x5 mesh, 25 clusters
- total 100 cores
- 4 cores/cluster
  - 3 C1, 128K L1, 1M L2
  - 1 C2, 128K L1, 1M L2
- 130 Mb total shared L3

Throughput = 103.26 IPC

Library of models

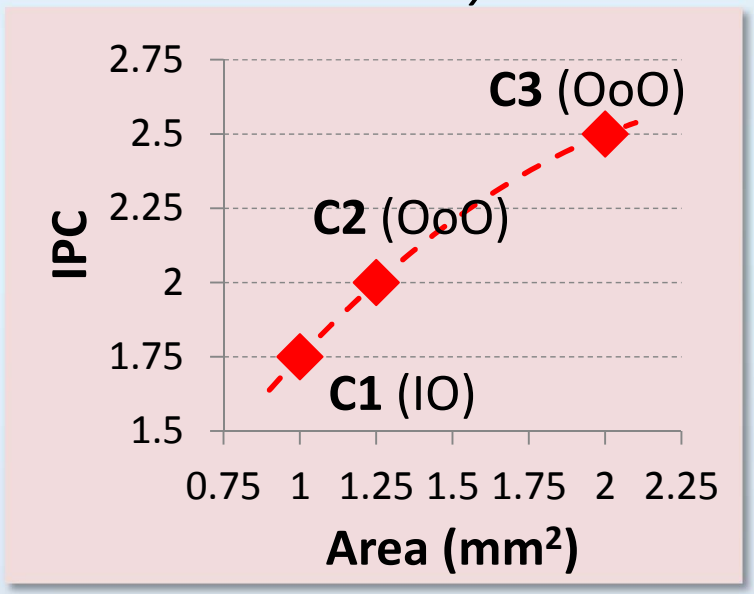
CMP configuration

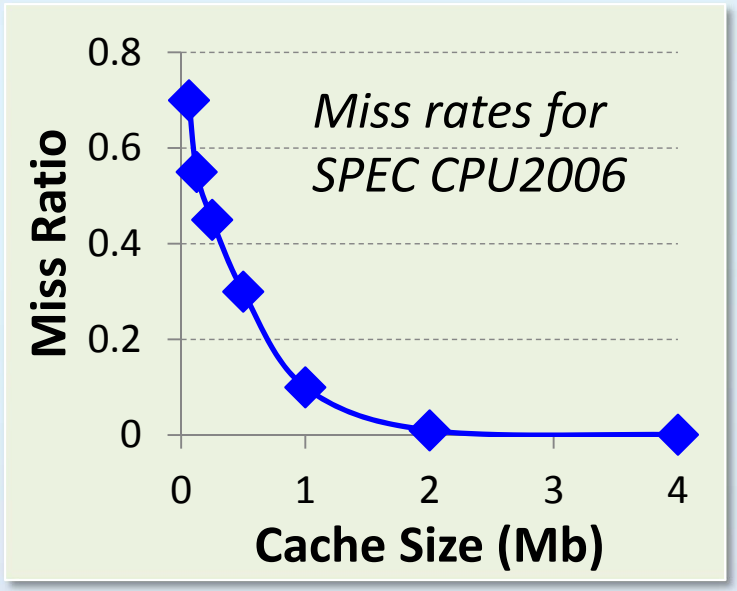
Parameter	Value
Chip area	350 mm^2
Mesh dimensions	2x2 to 16x16
Mem. Cntrl. latency	200 cycles
Interconnect	Bus, uni- / bi-ring
Link width	256 – 1024 bits
Workload MPI	0.5
Workload MLP	1.25

Cache models

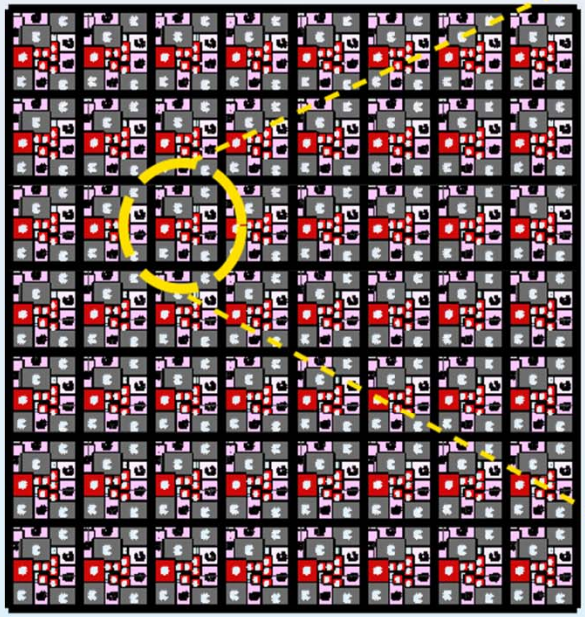
Cache Size	Area (mm^2)	Latency (cycles)
64Kb	0.063	2
128Kb	0.125	3
256Kb	0.25	4
…	…	…
8Mb	8.0	9

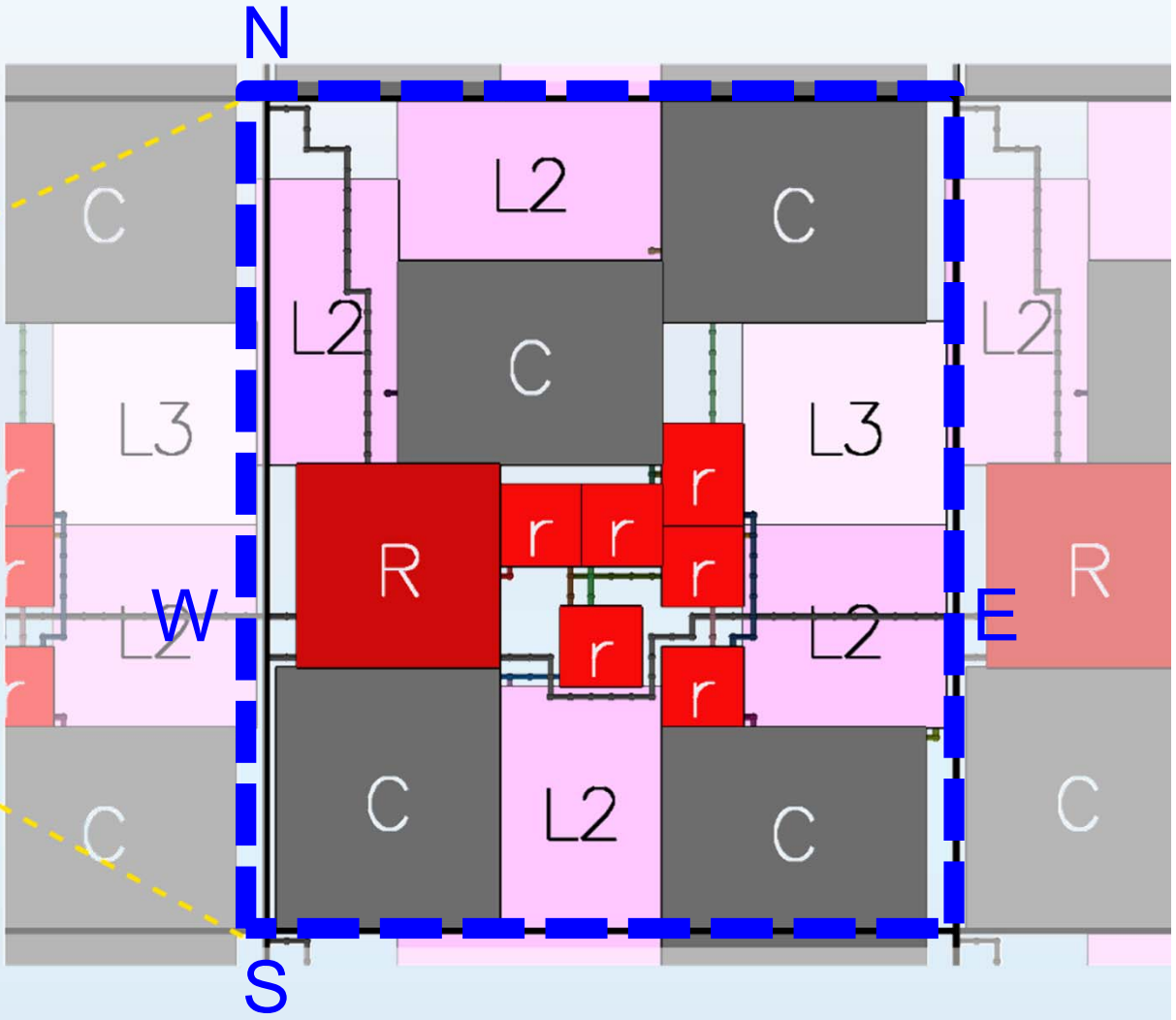
Core library





Physical planning for tiled CMPs

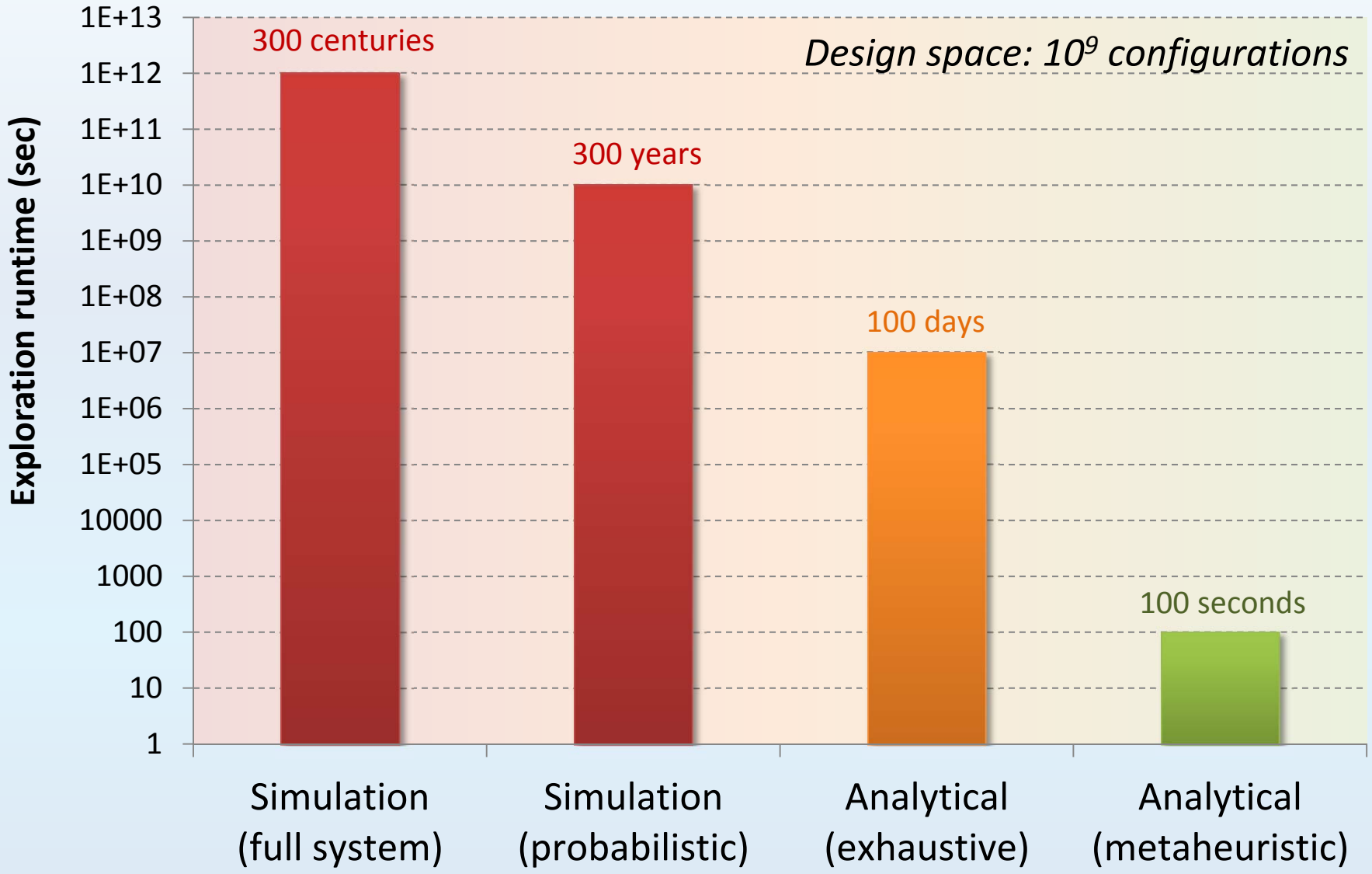




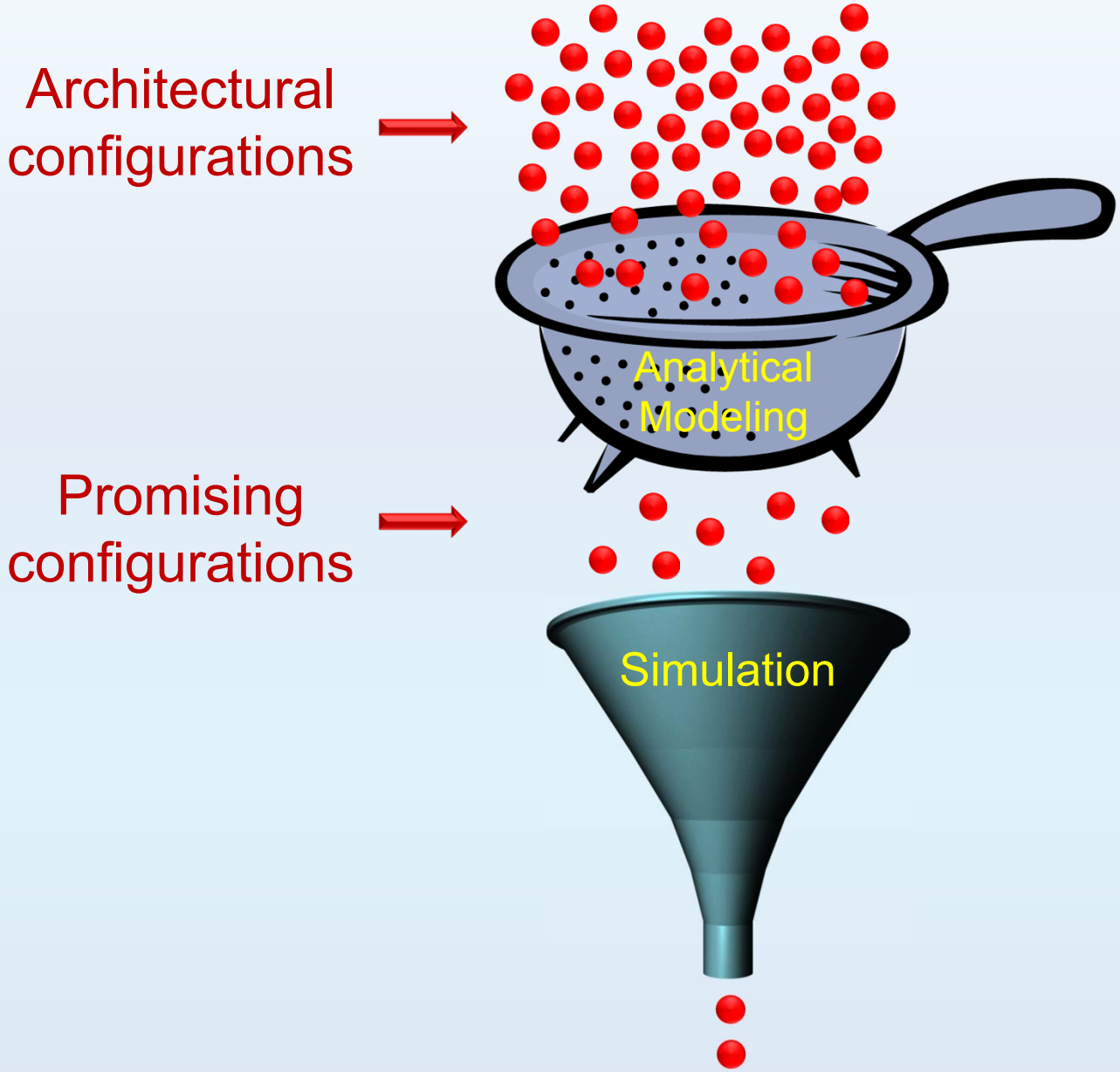
Outline

- **Architectural exploration**
  - **The cost of exploration**
  - **Exploring with metaheuristics**
  - **Analytical models**
- Physical planning for tiled CMPs
- Current work: regular floorplanning

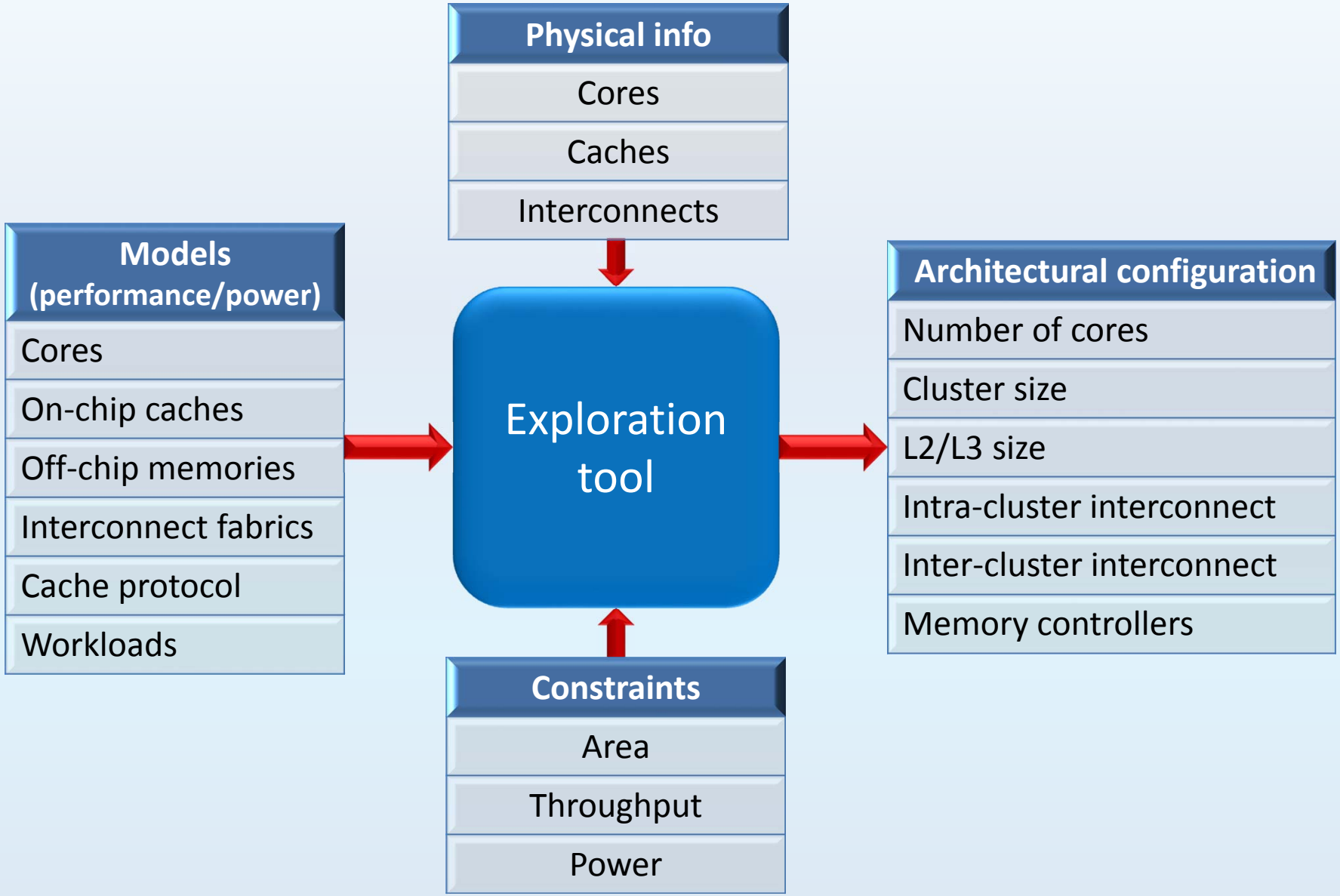
Exploration engines



Scalable exploration

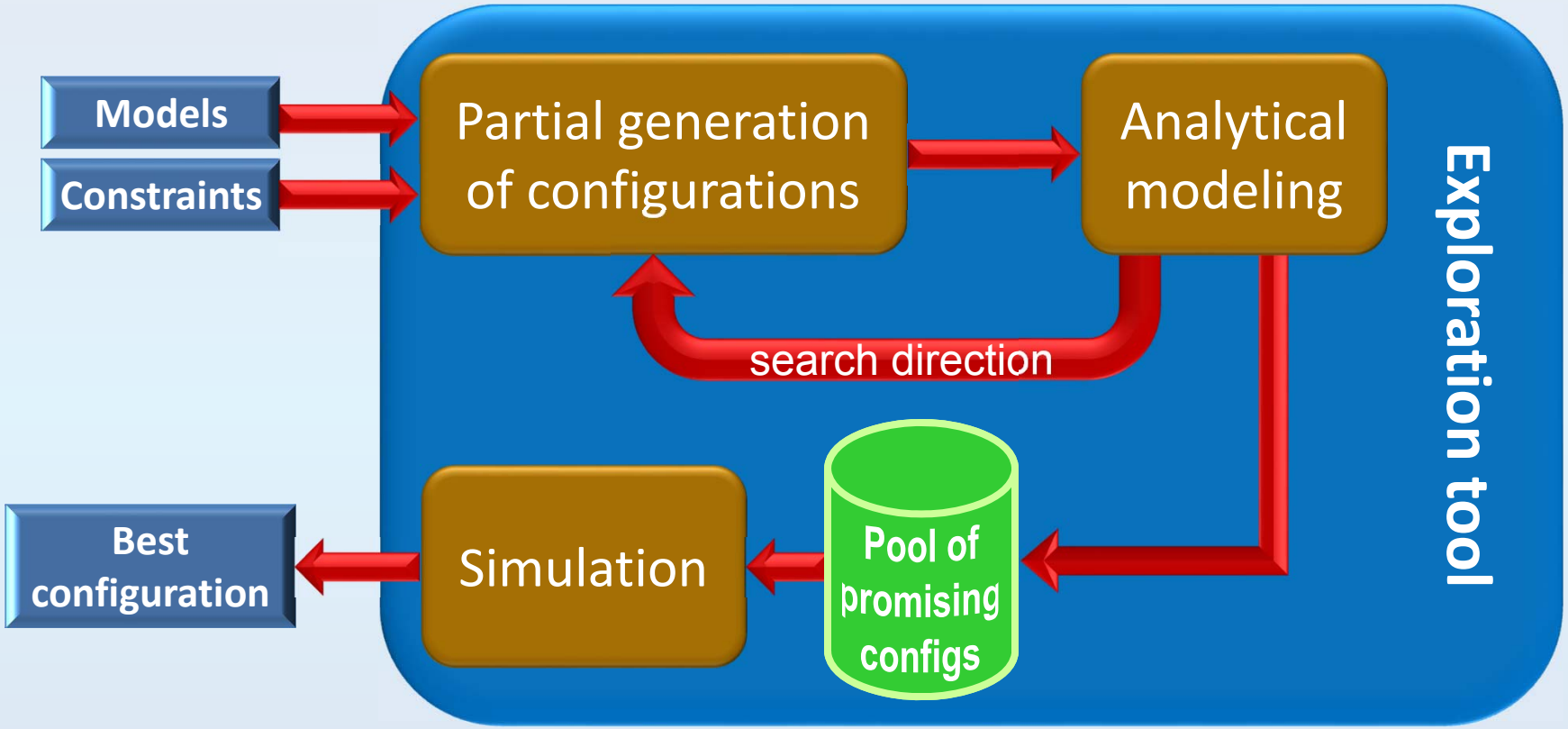


Automated exploration



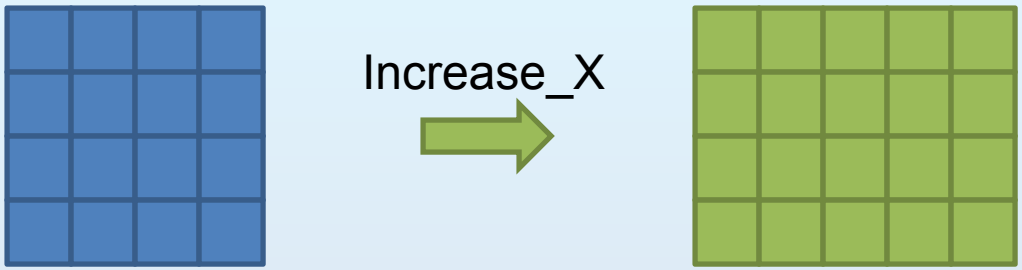
Exploration engine: metaheuristics

- Explore huge design spaces efficiently
- Our proposal:
  - Simulated Annealing (Kirkpatrick et al., 1983)
  - Extremal Optimization (Boettcher et al., 1999)

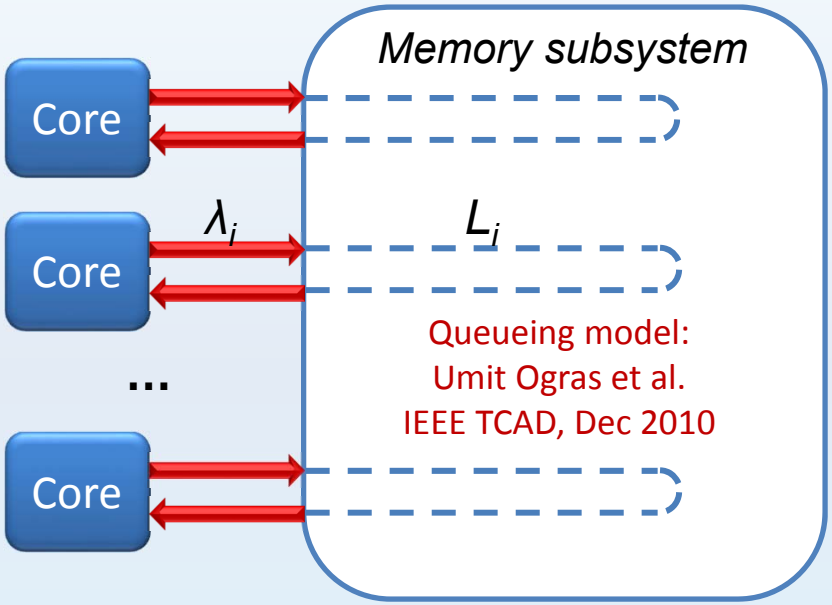


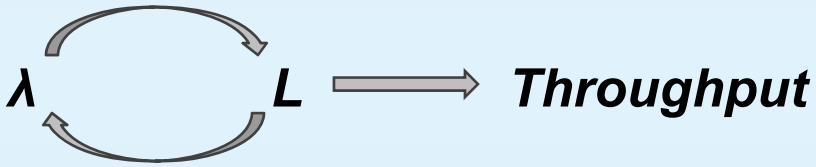
Generation of configurations

- Generate neighbors by applying transformations
  - Increase/Decrease
    - mesh dimensions
    - core count per cluster
    - L1, L2 size
  - Change interconnect type (bus/uni-ring/bi-ring)
  - Complex updates (increase mesh/decrease core count)
- Example: Increase_X(mesh 4x4) => mesh 5x4



Analytical performance model for CMPs





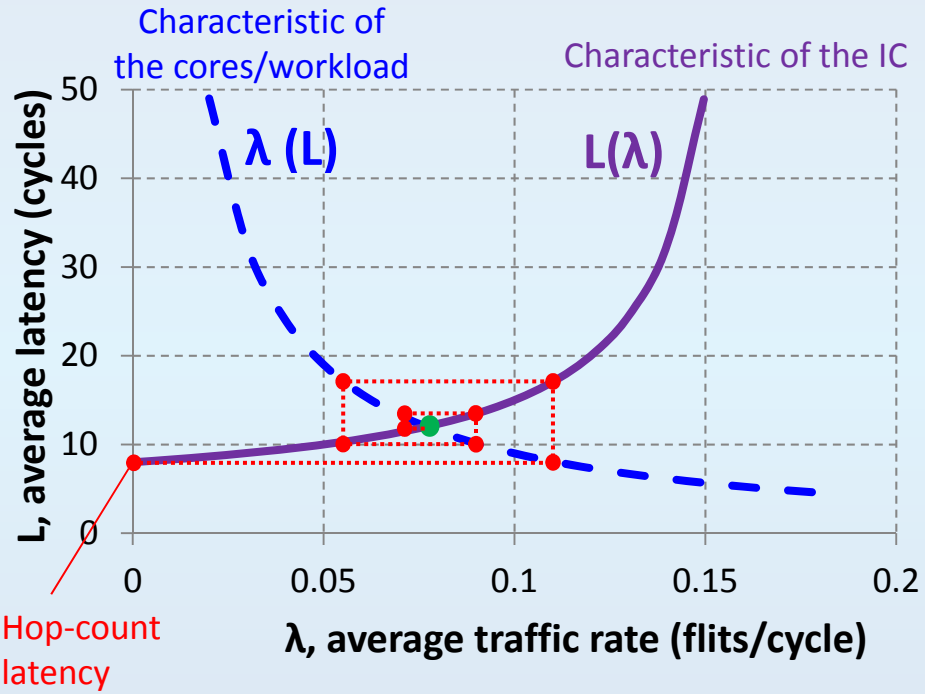
λ↑ → L↑ → λ↓ → L↓ → λ↑ → •••

Nonlinear analytical models

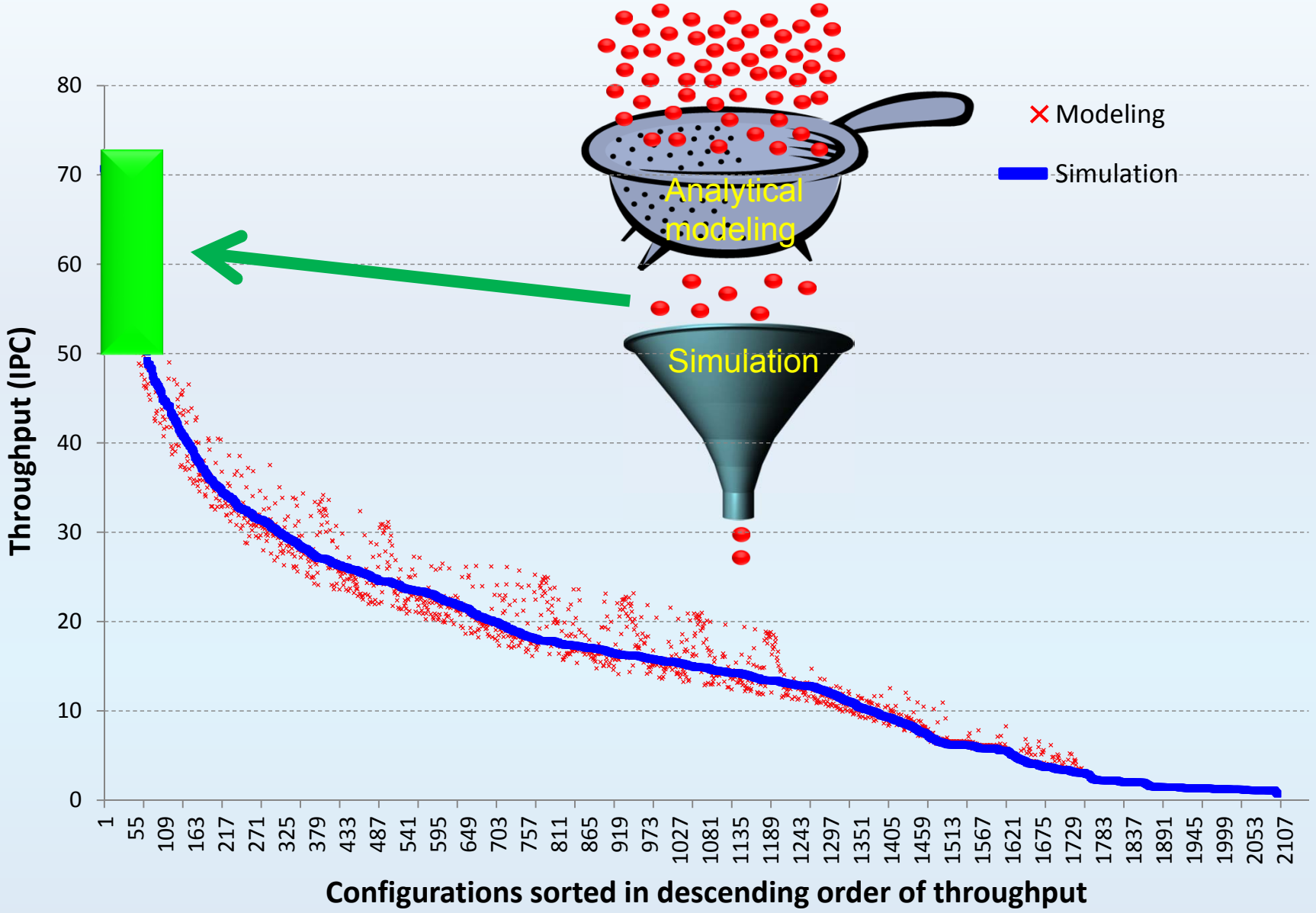
Throughput model: $CPI = CPI_0 + MPI \cdot L$

Traffic model: $\lambda = \frac{MPI}{CPI_0 + MPI \cdot L}$

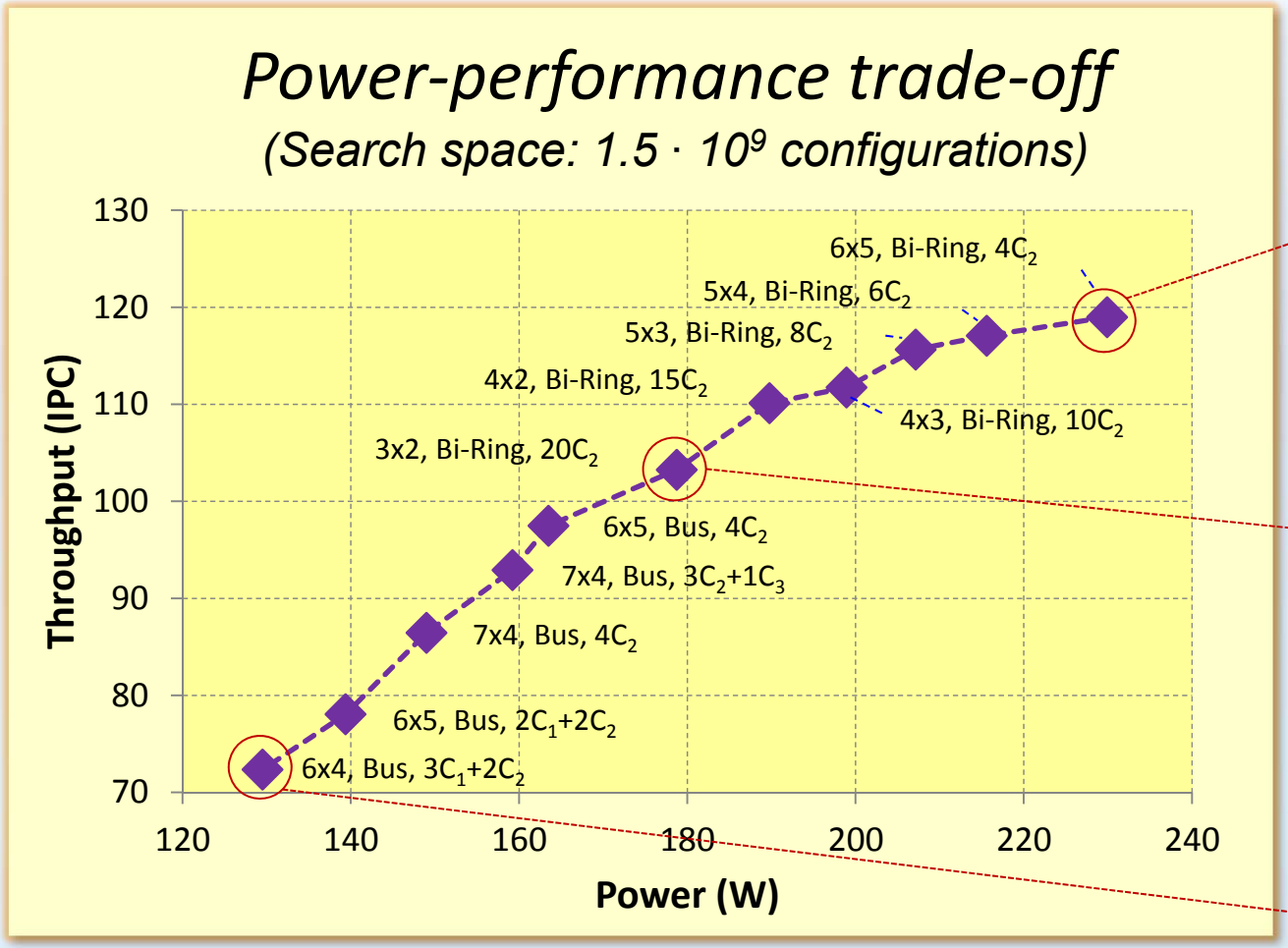
Latency model: $L = L_{hop\text{-}count} + L_{contention}$

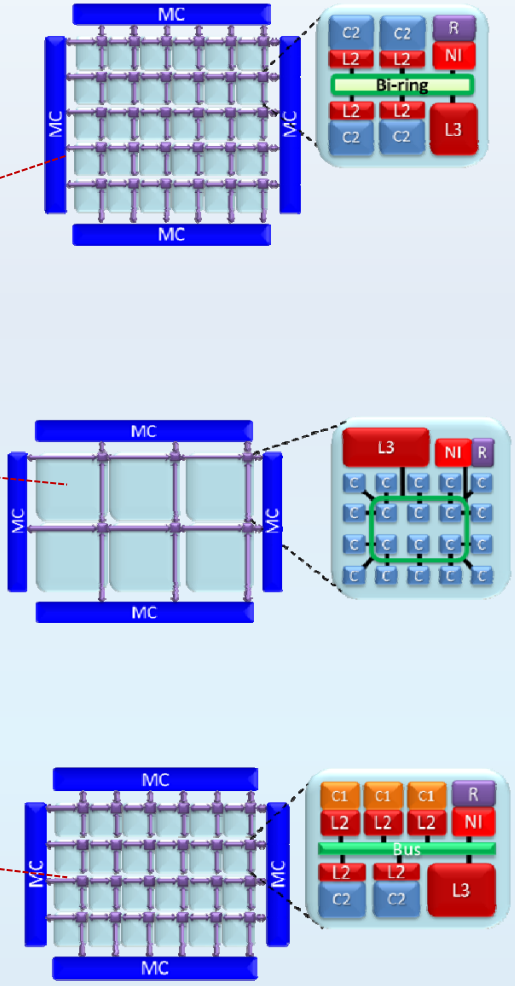


Analytical model vs. simulation



Case Study: Power-performance exploration

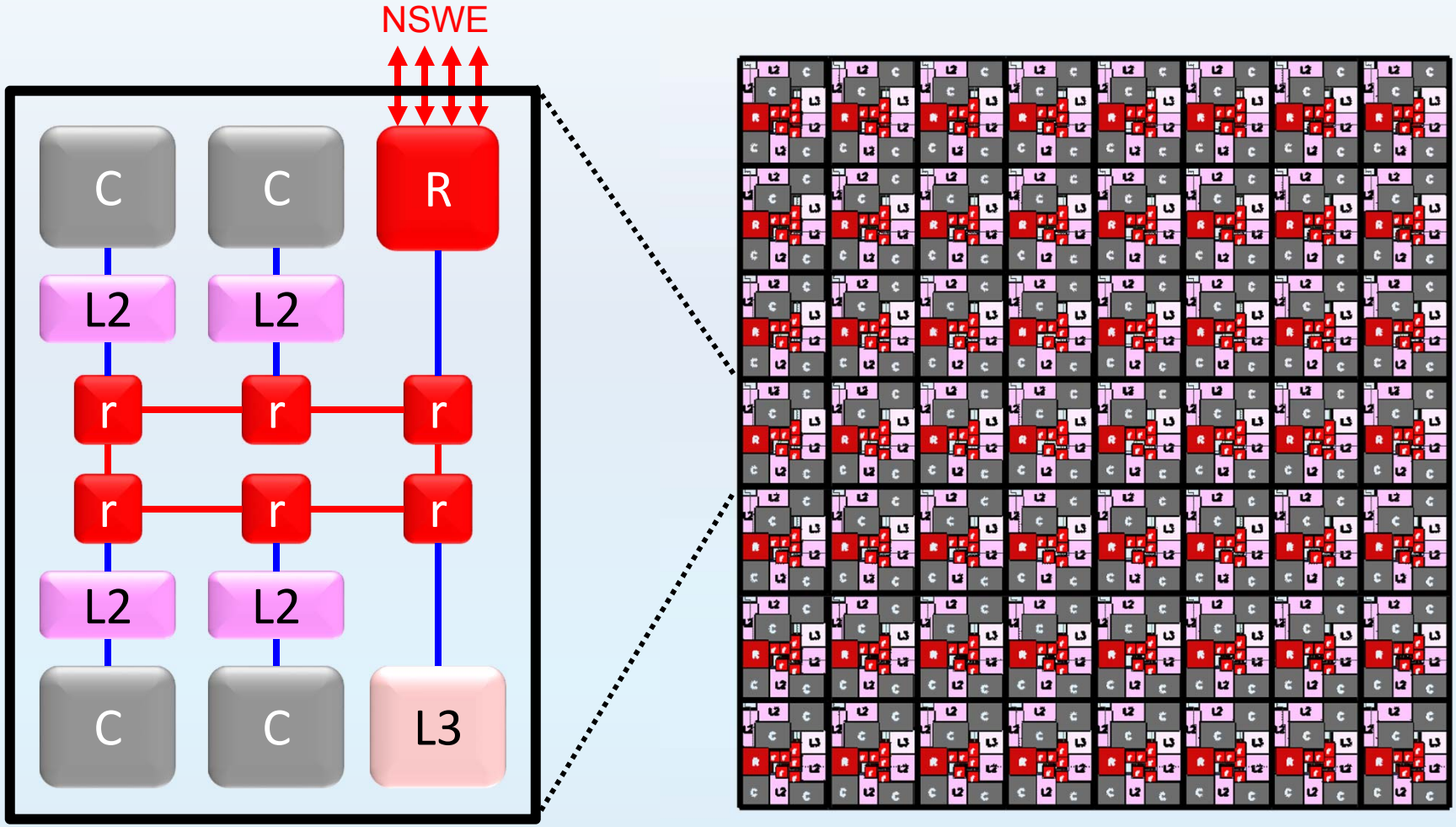




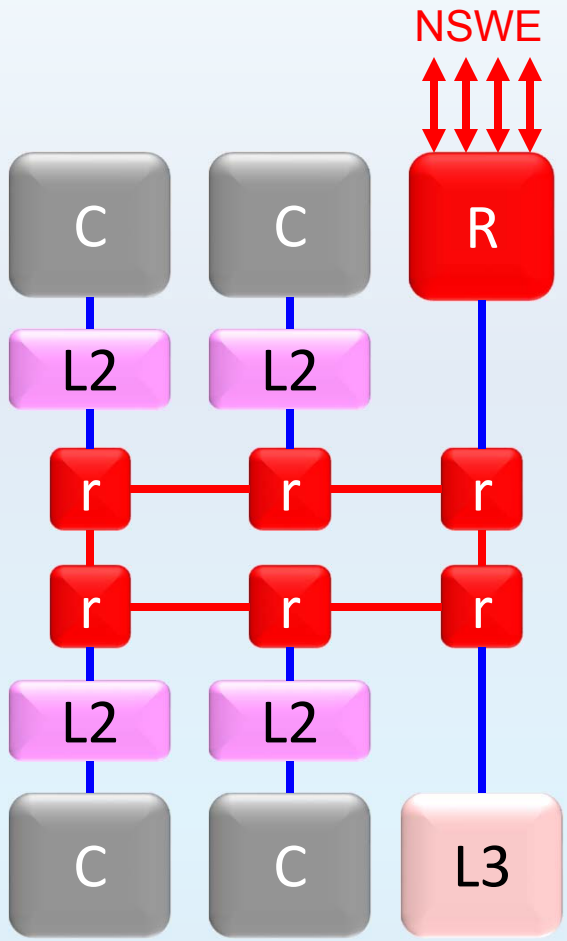
Outline

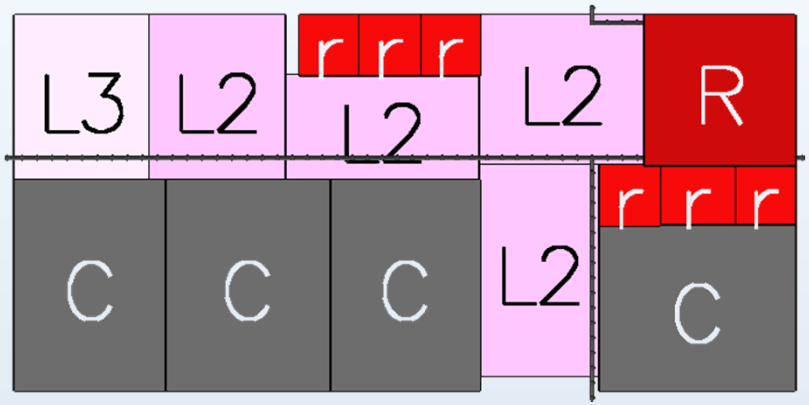
- Architectural exploration
- **Physical planning for tiled CMPs**
  - **Impact of physical planning**
  - **Floorplanning**
  - **Wire planning**
- Current work: regular floorplanning

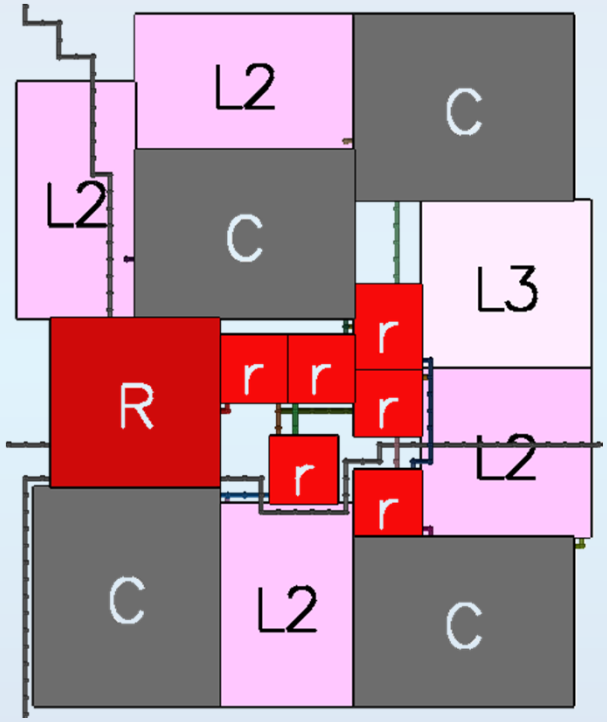
Physical planning



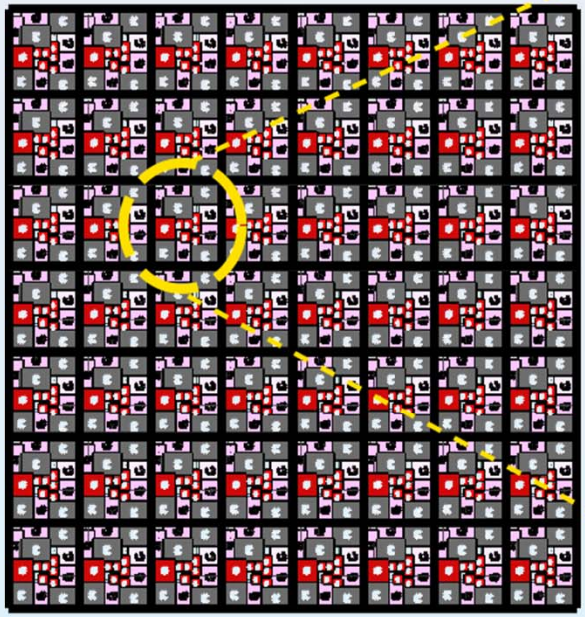
The impact of physical planning

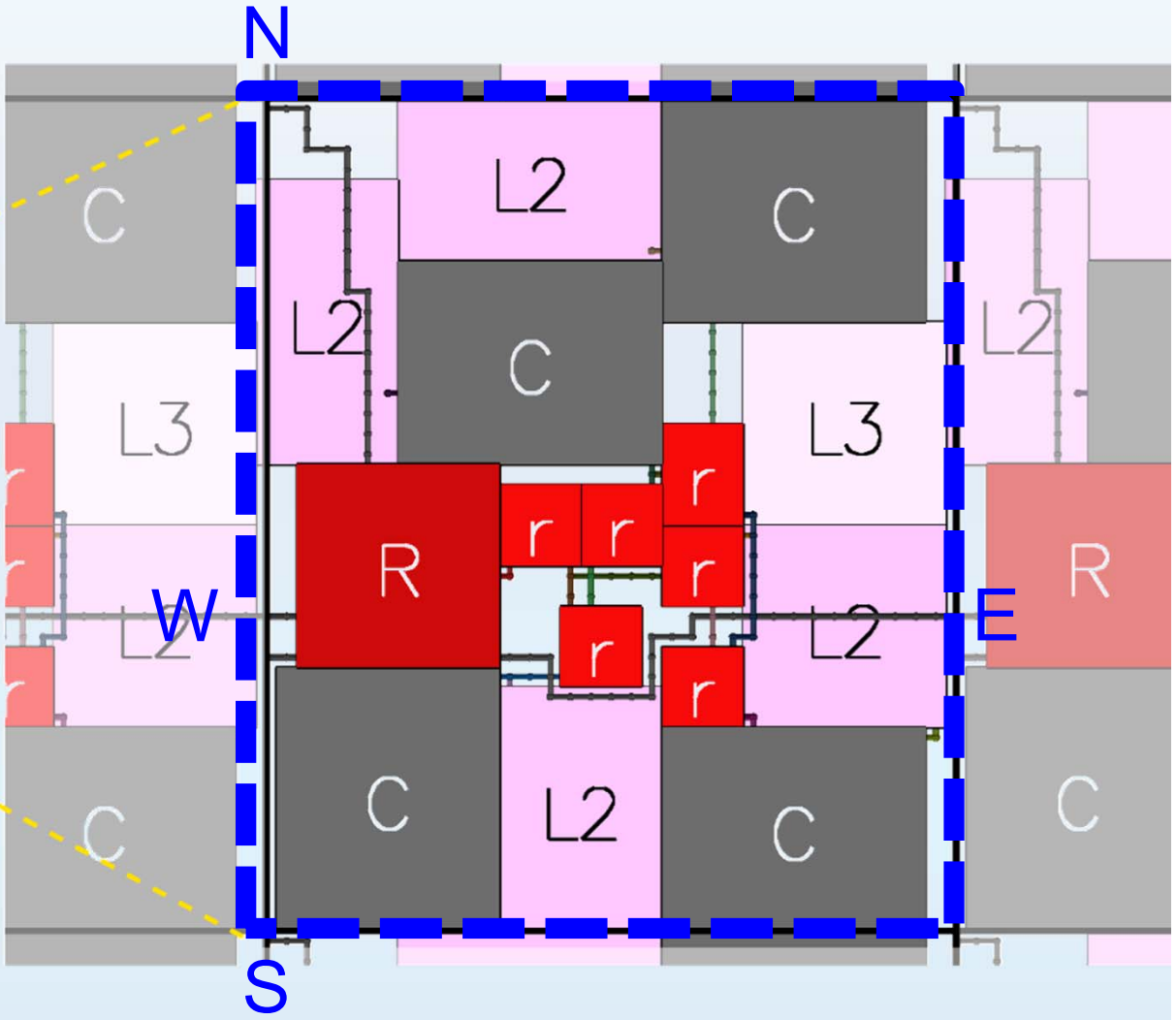




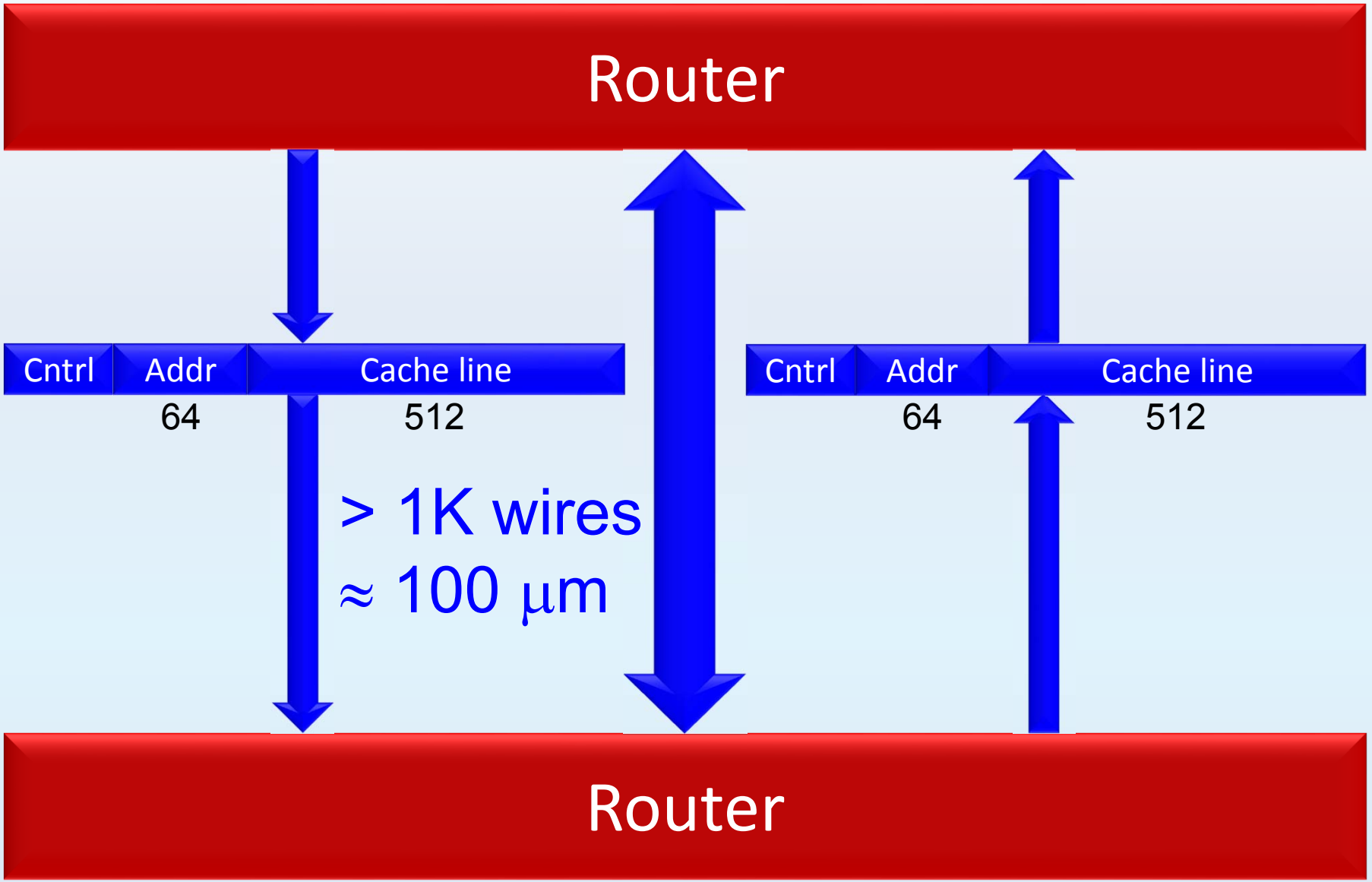


Physical planning for tiles

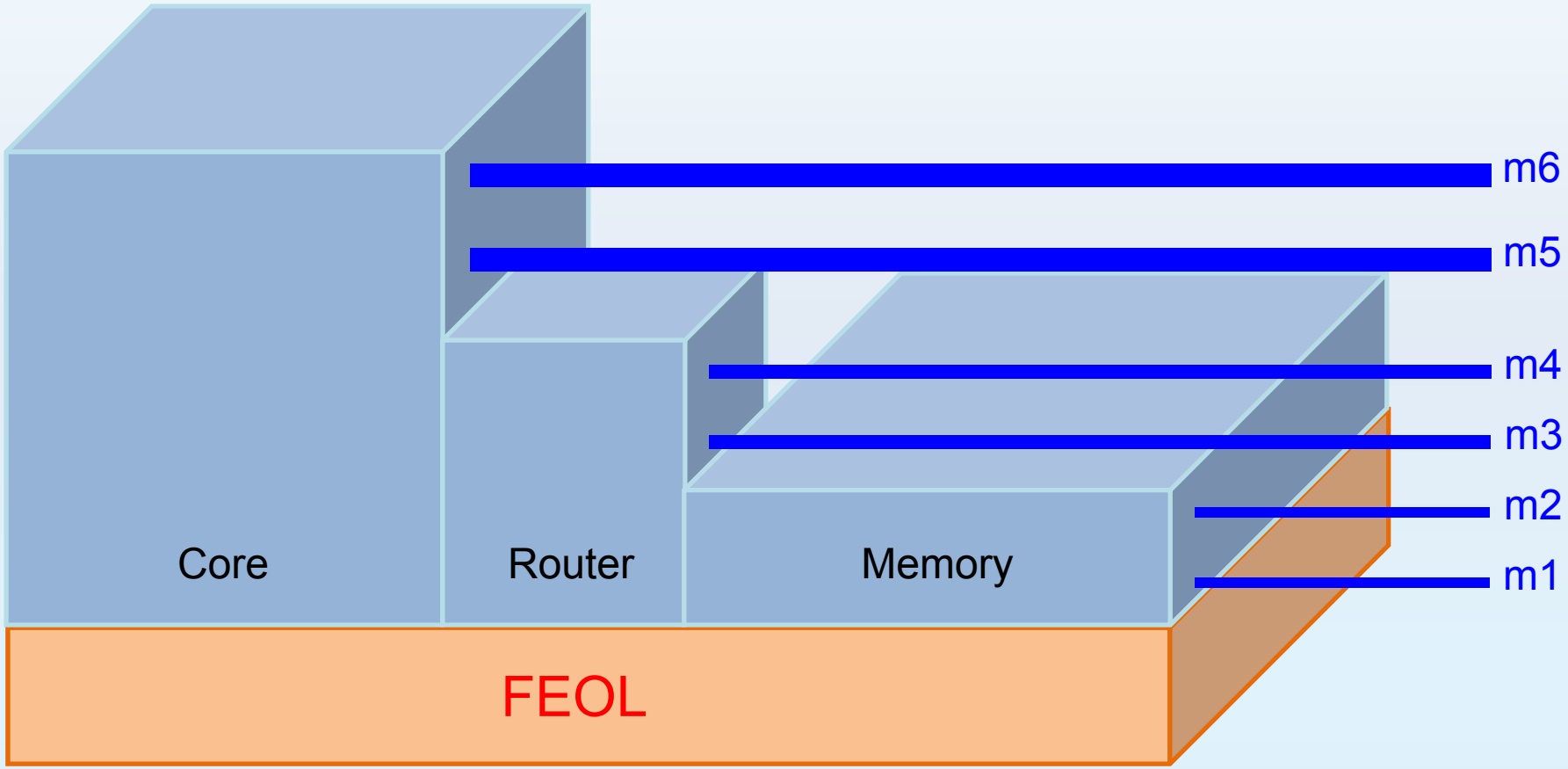




Link width: how many wires?

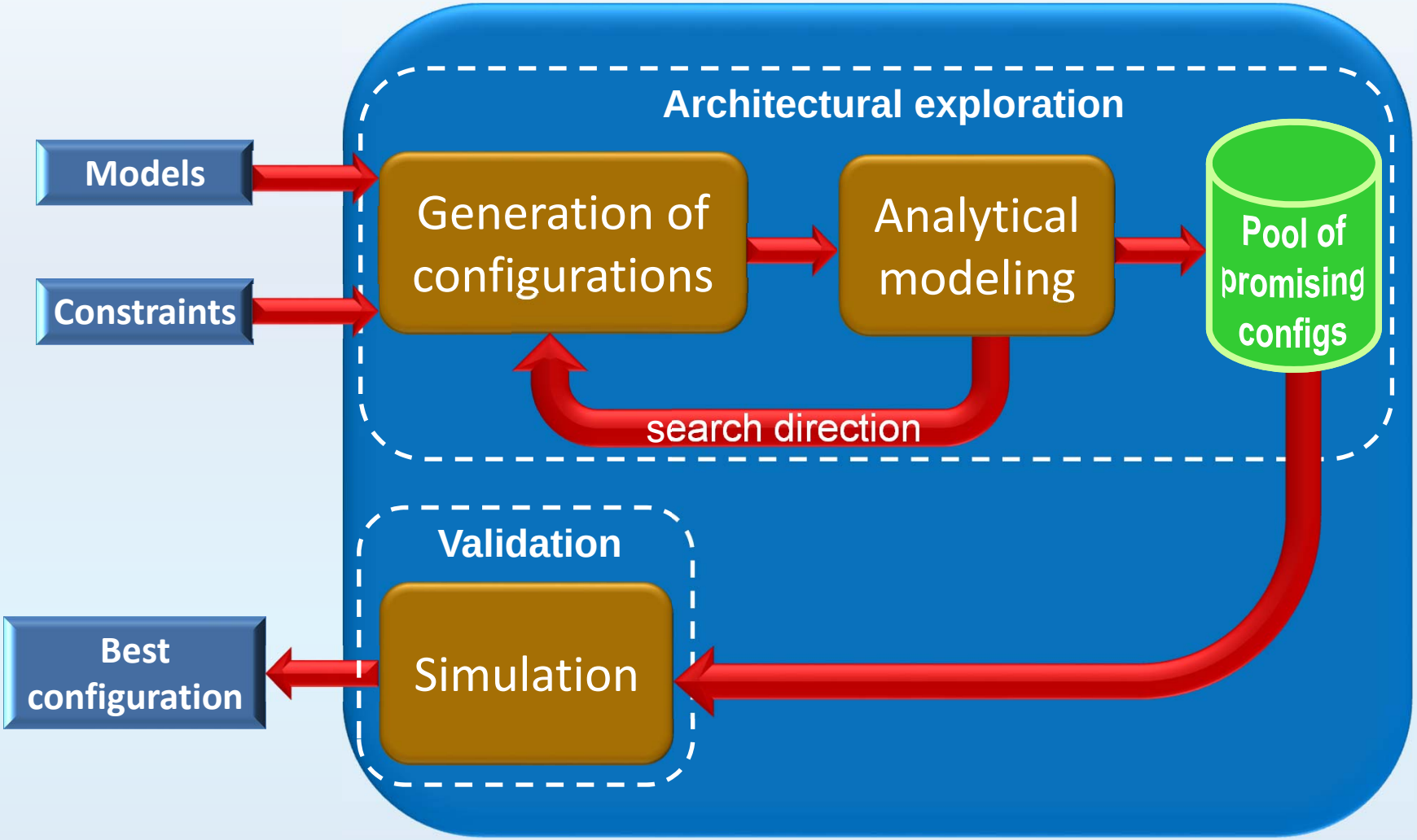


3D Wire Planning

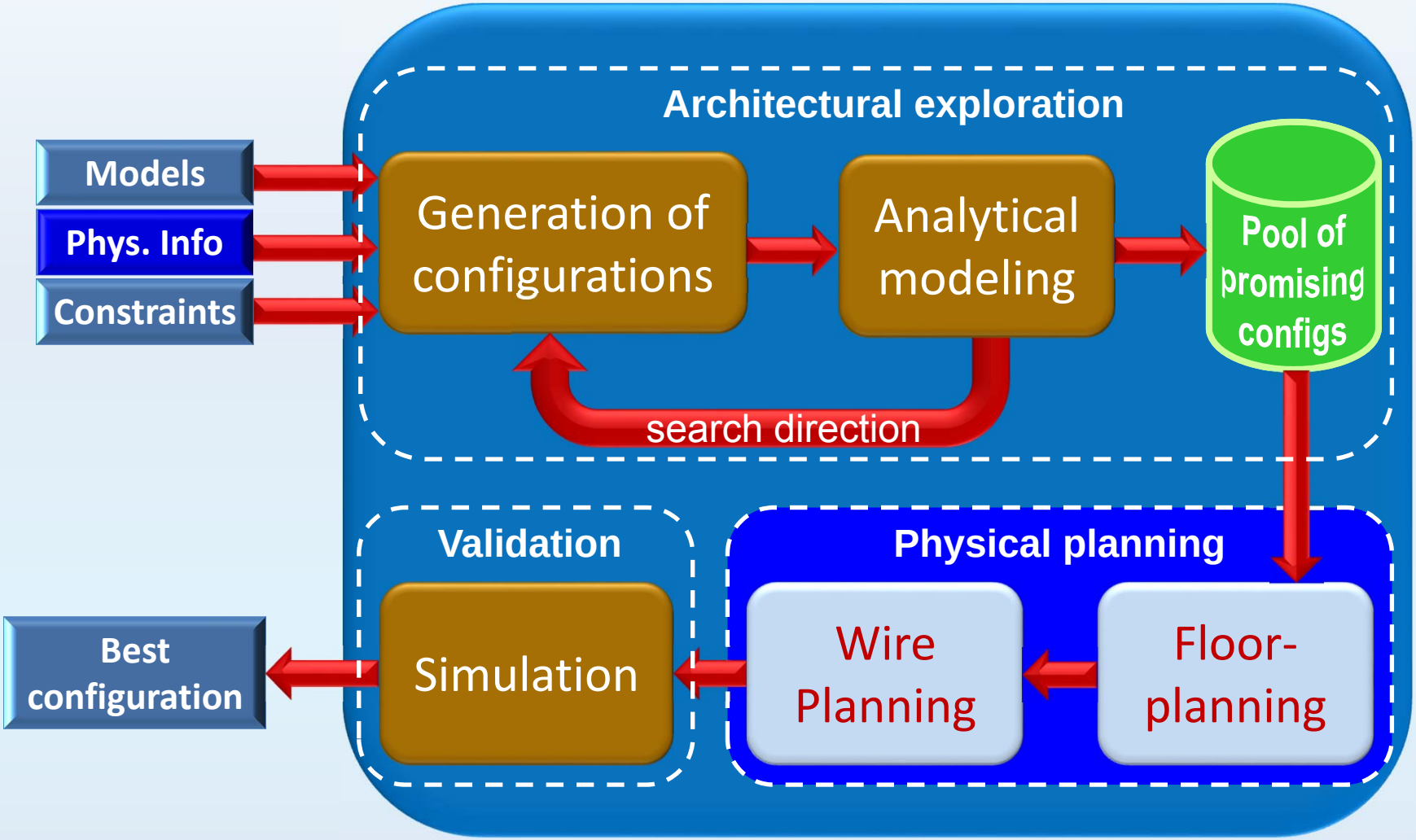


In systems where memory bandwidth is the bottleneck,
the physical resources providing the bandwidth are critical

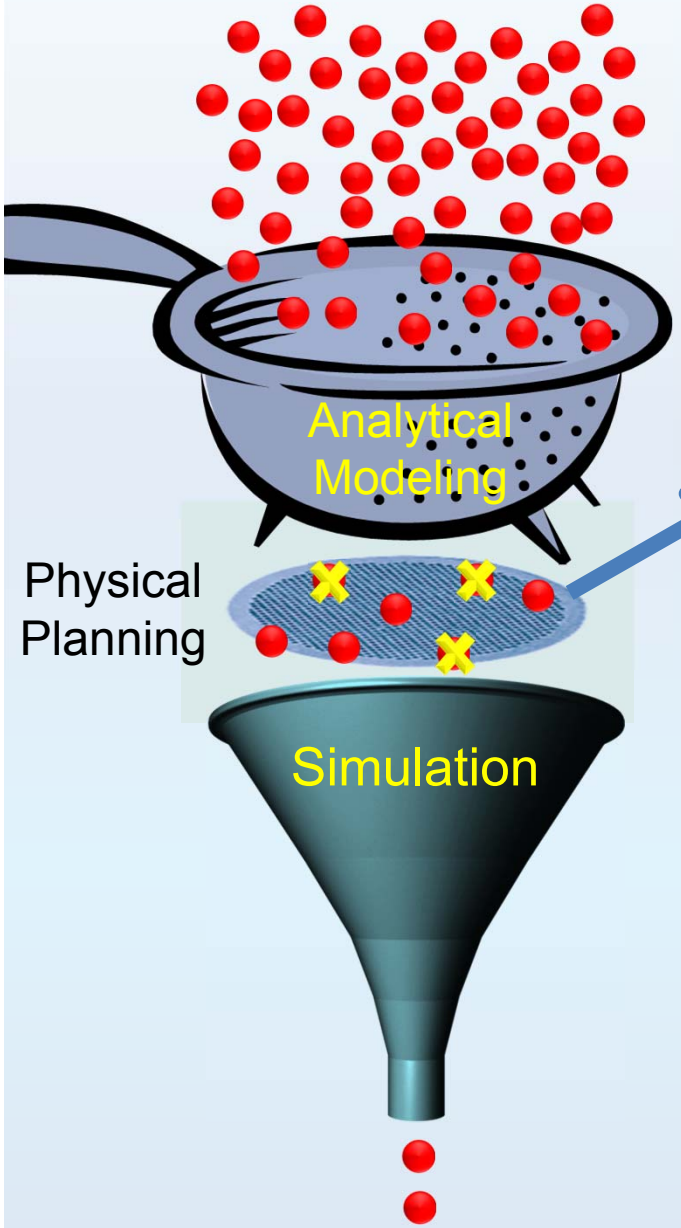
Exploration without physical planning

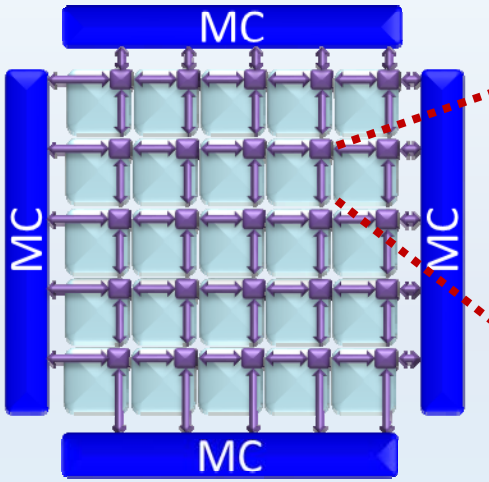


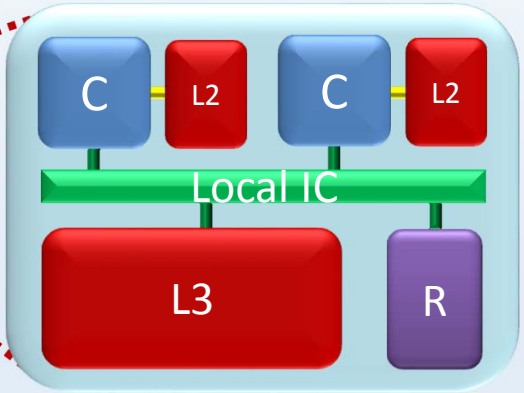
Exploration with physical planning



Physical planning

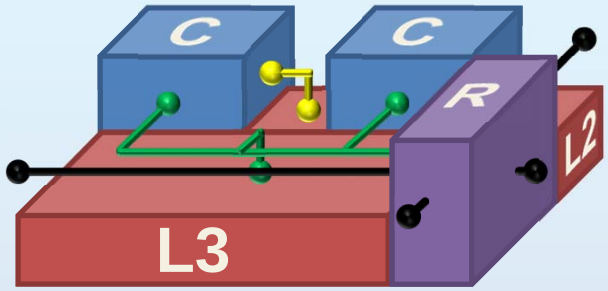






Estimations:

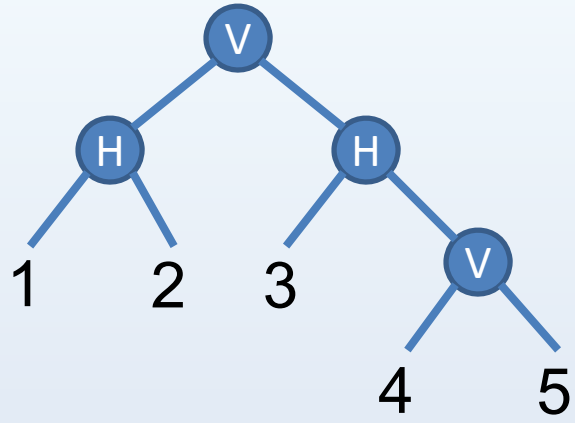
- Area
- Wirelength
- Routability

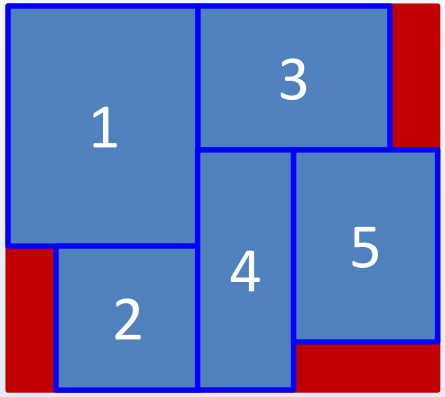


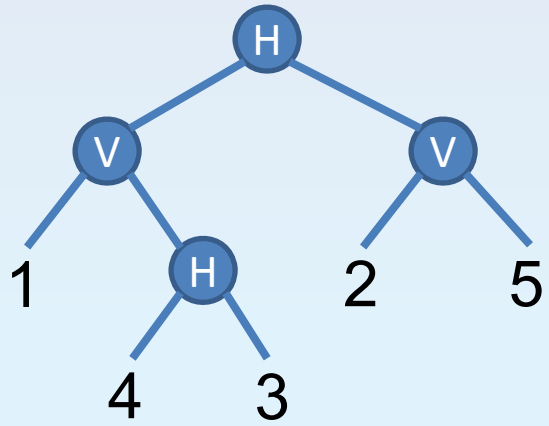
Physical planning technology

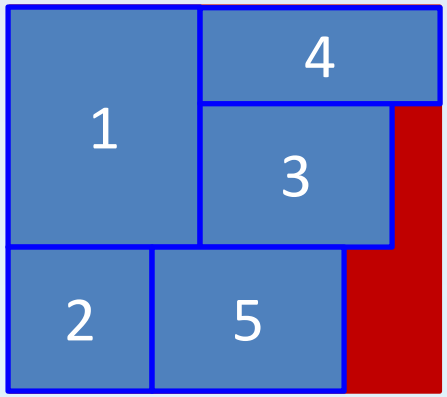
- Floorplanning
  - Slicing structures & Simulated Annealing
  - Lightweight 3D maze router
  - Constraints:
    - Adjacency (Core $\leftrightarrow$ L2)
    - Balanced links (rings)
- Wire planning
  - SAT-based 3D global routing
  - Boolean constraints

Slicing structures



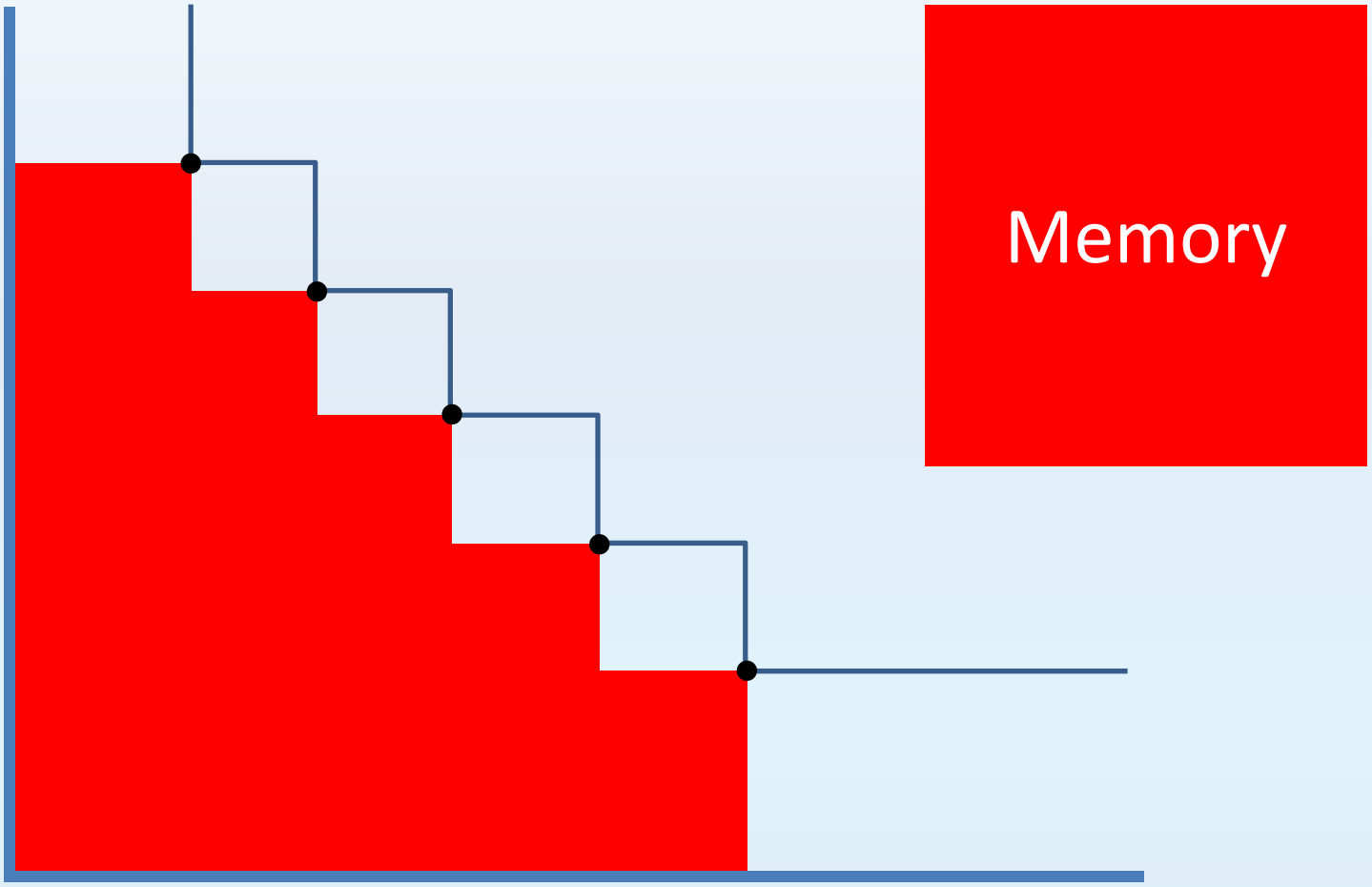






D.F. Wong and C.L. Liu,
"A New Algorithm for Floorplan Design"
DAC, 1986, pages 101-107.

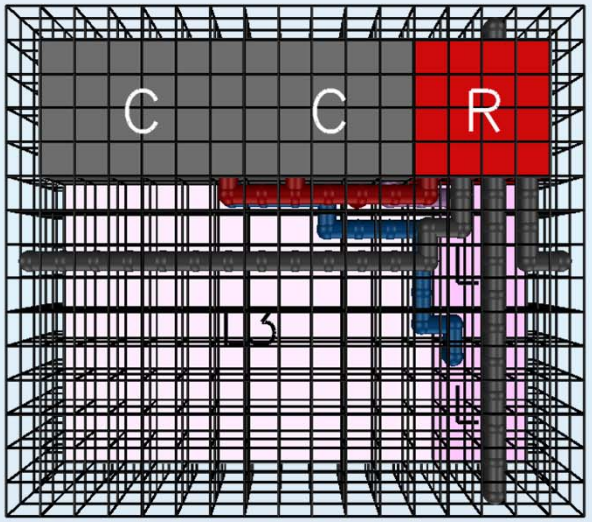
Bounding curves



L. Stockmeyer, 1983,
Optimal Orientation of Cells in Slicing Floorplan Designs

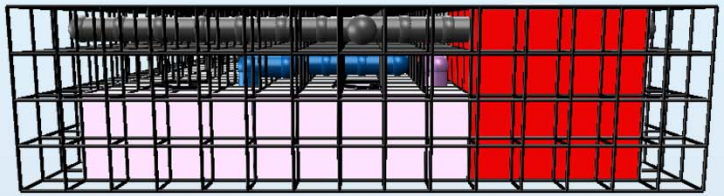
Wire planner

- SAT-based approach for gridded routing
- Grid unit: link width (≈ 500 - 1000 wires)
- Support for floating terminals
- Customizable for any type of Boolean-encoded constraints (symmetry, 1D/2D routing, …)

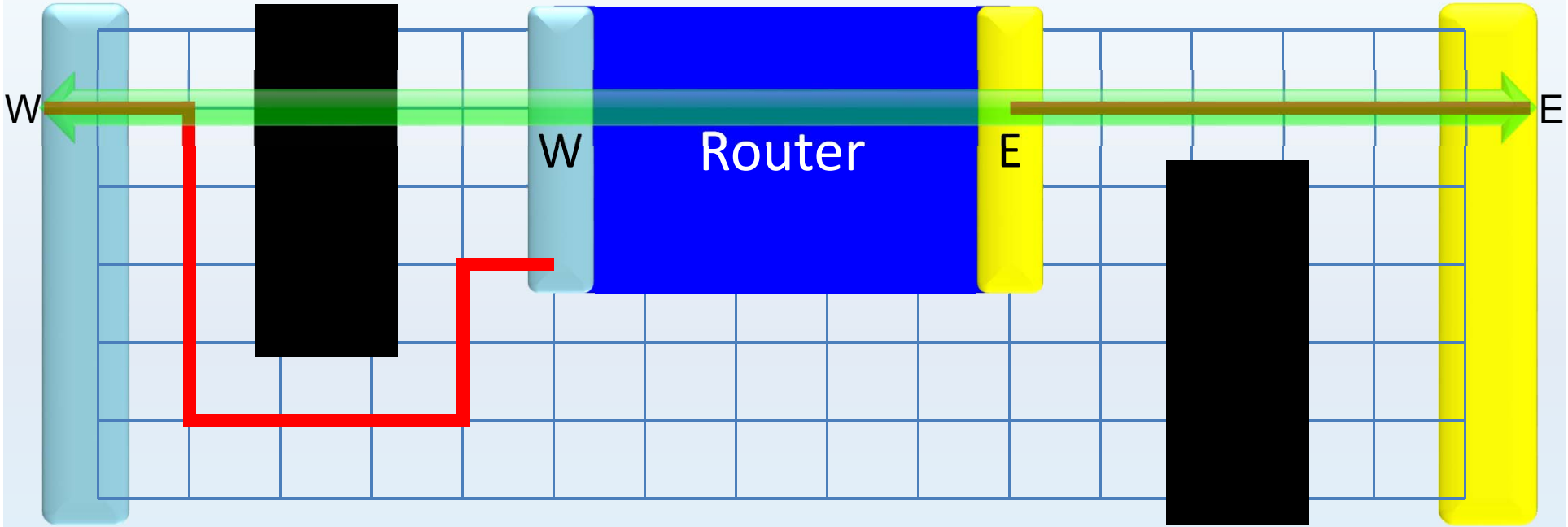


⇦ Top view

Cross-section view

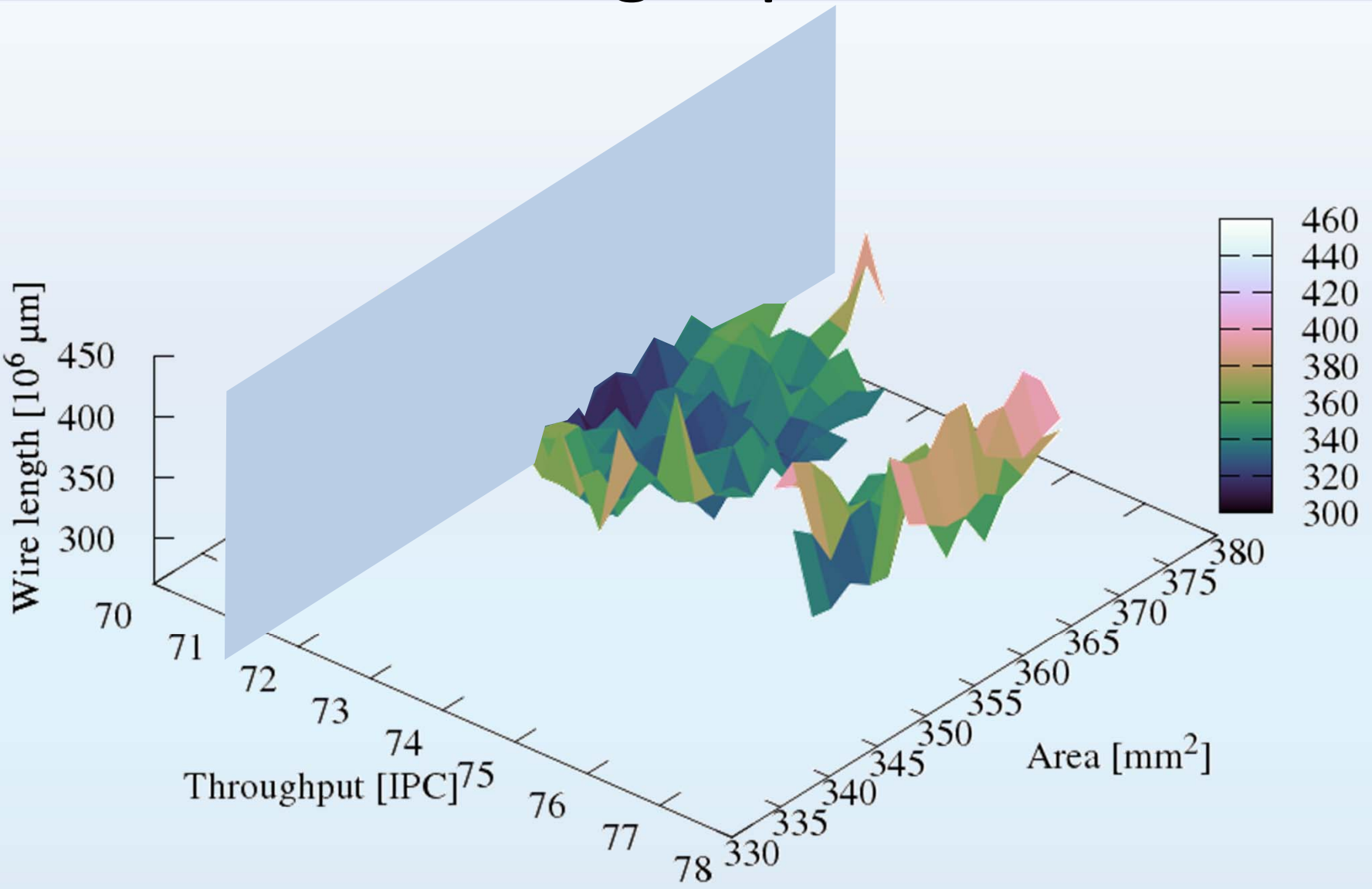


Wire planner

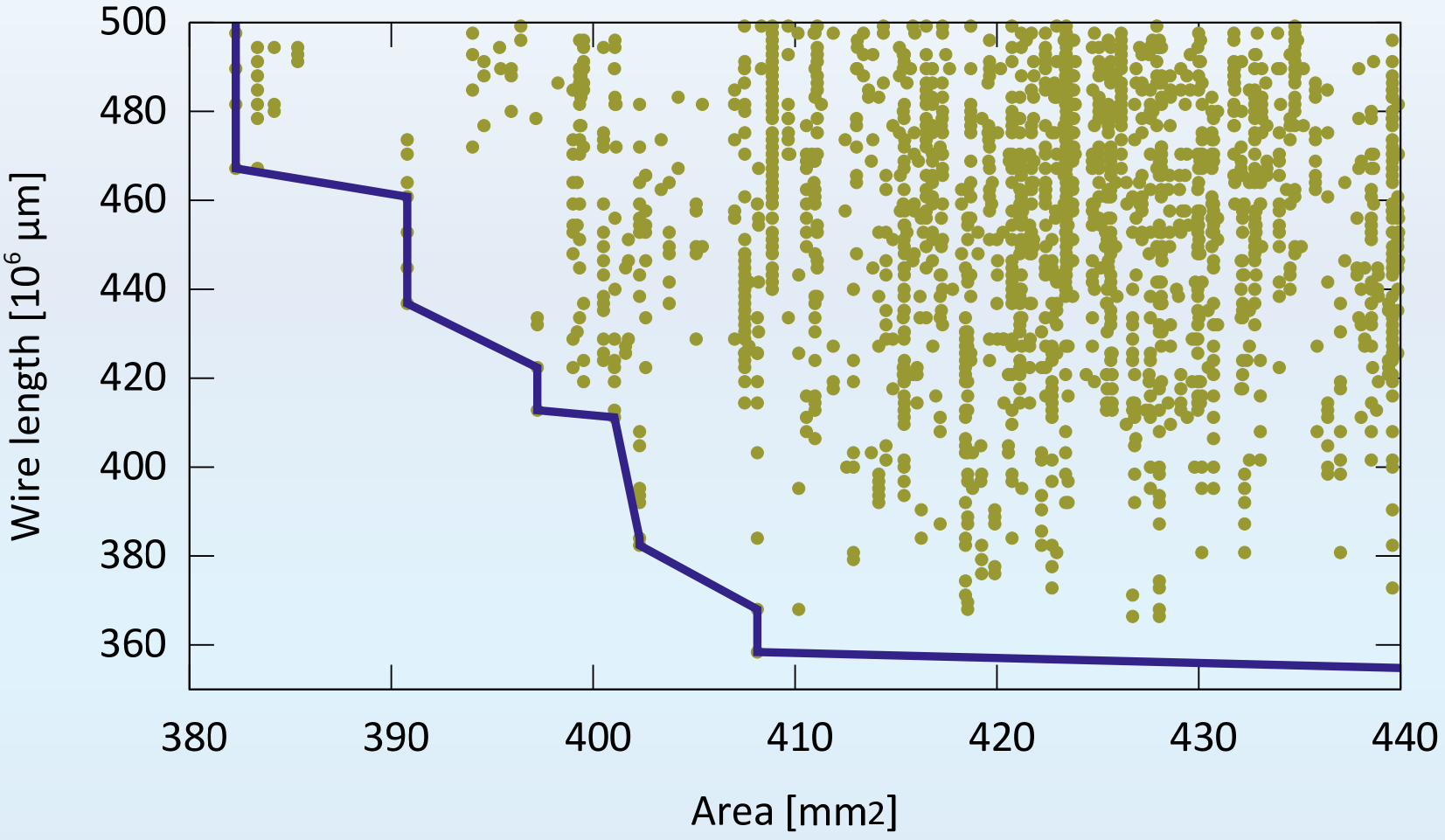


- Concurrent routing: all nets simultaneously
- Using Euler's theory to find legal routes
- SAT: a route is always found if it exists
- ILP-based route optimization

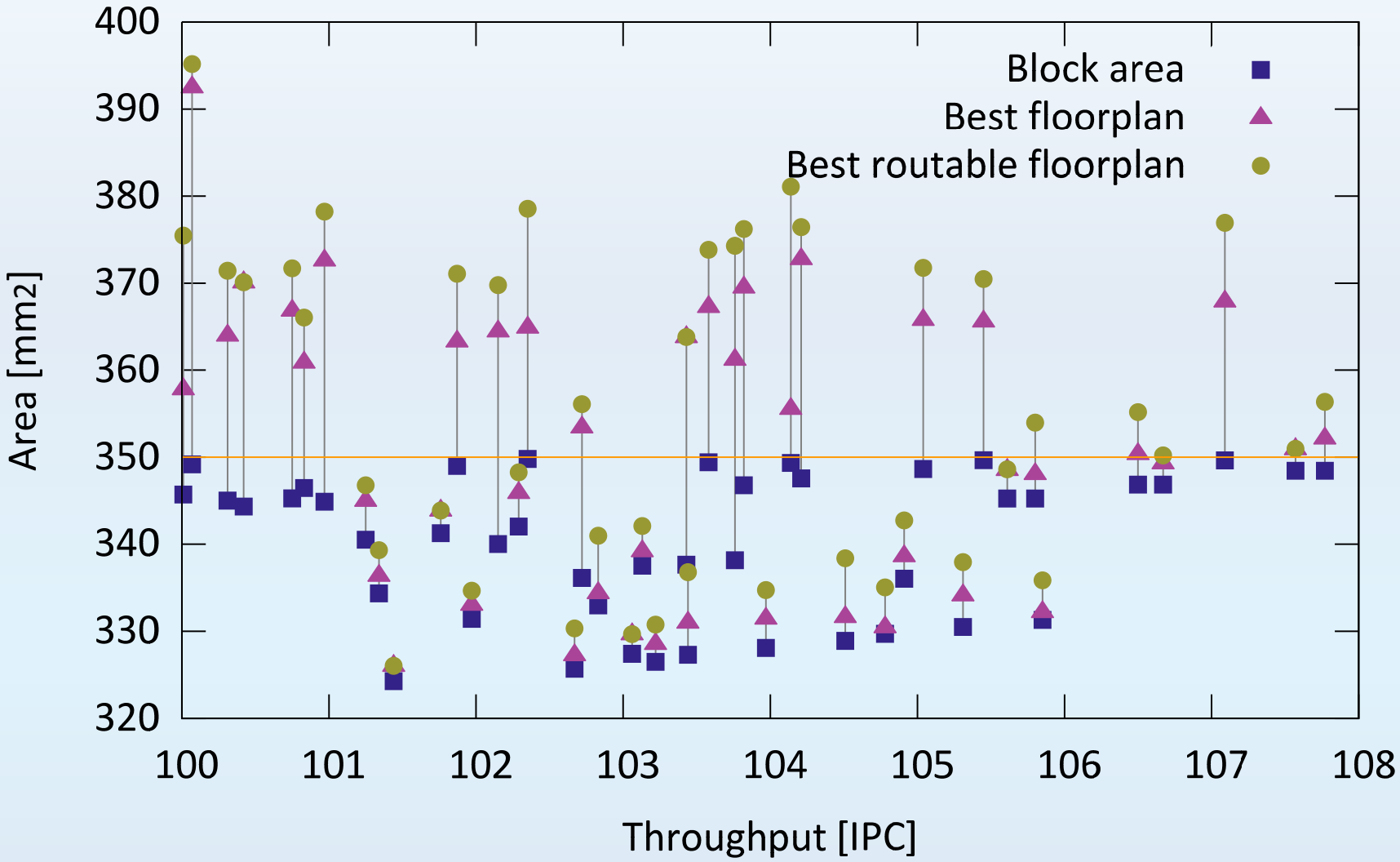
Design space



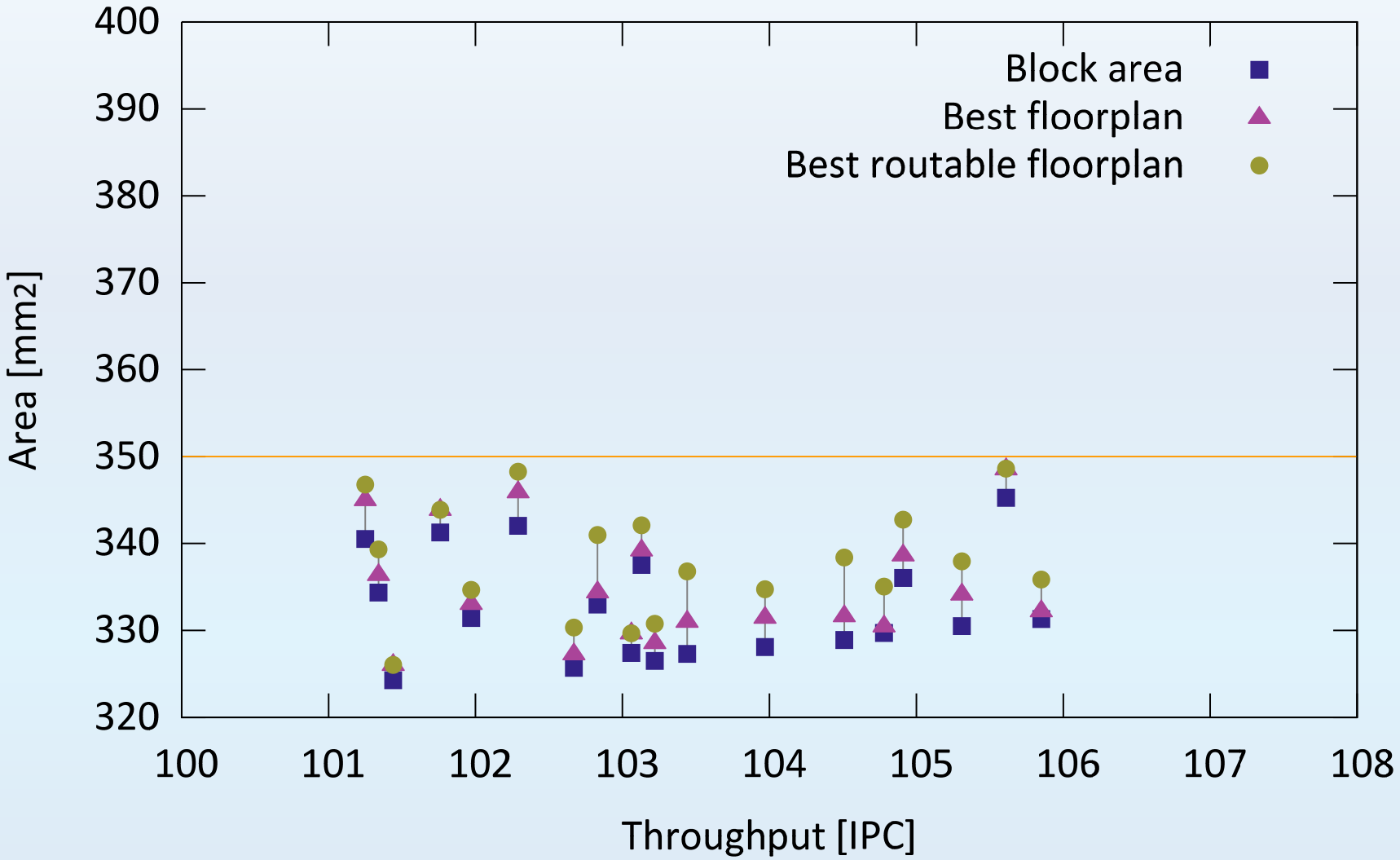
Design space



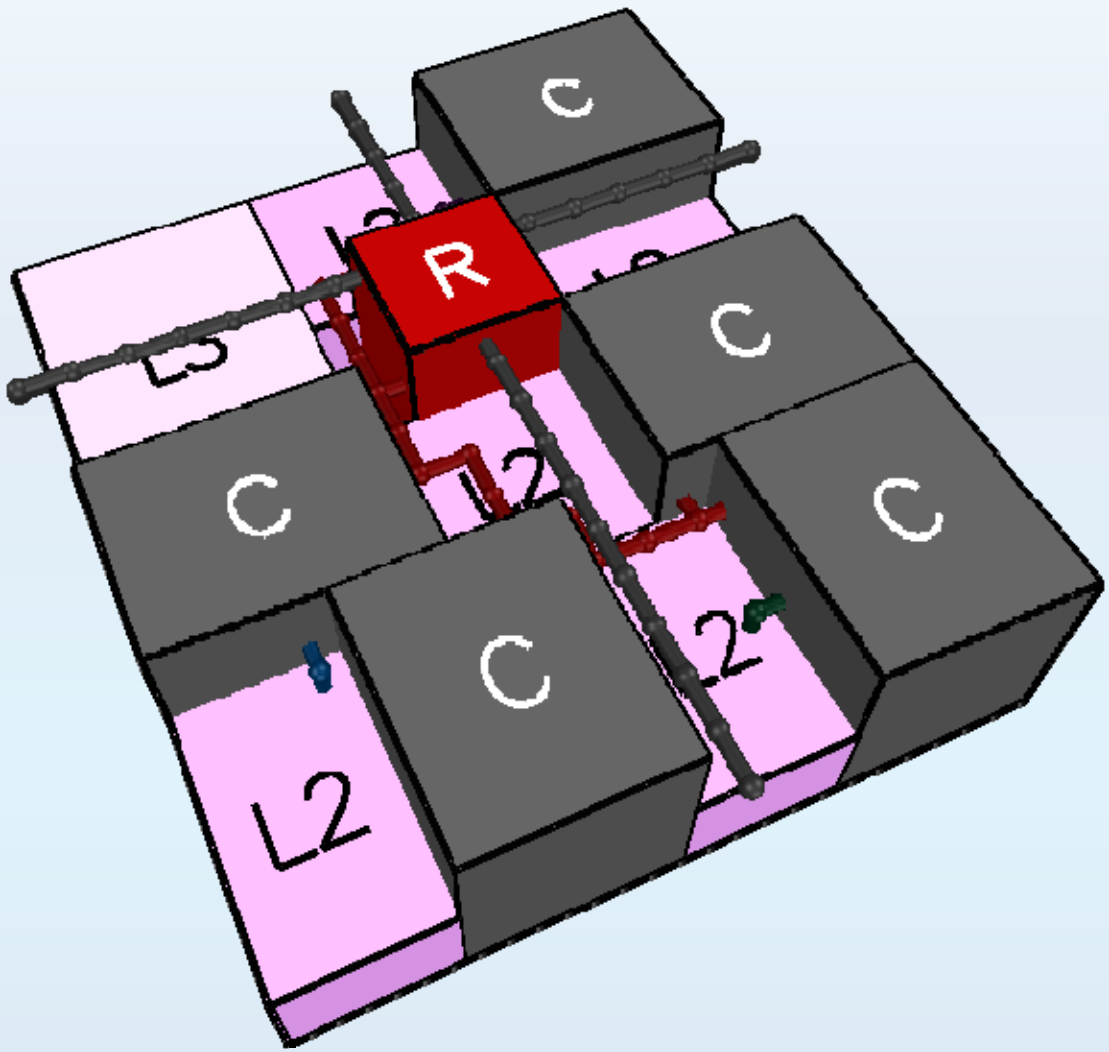
Filtering floorplans



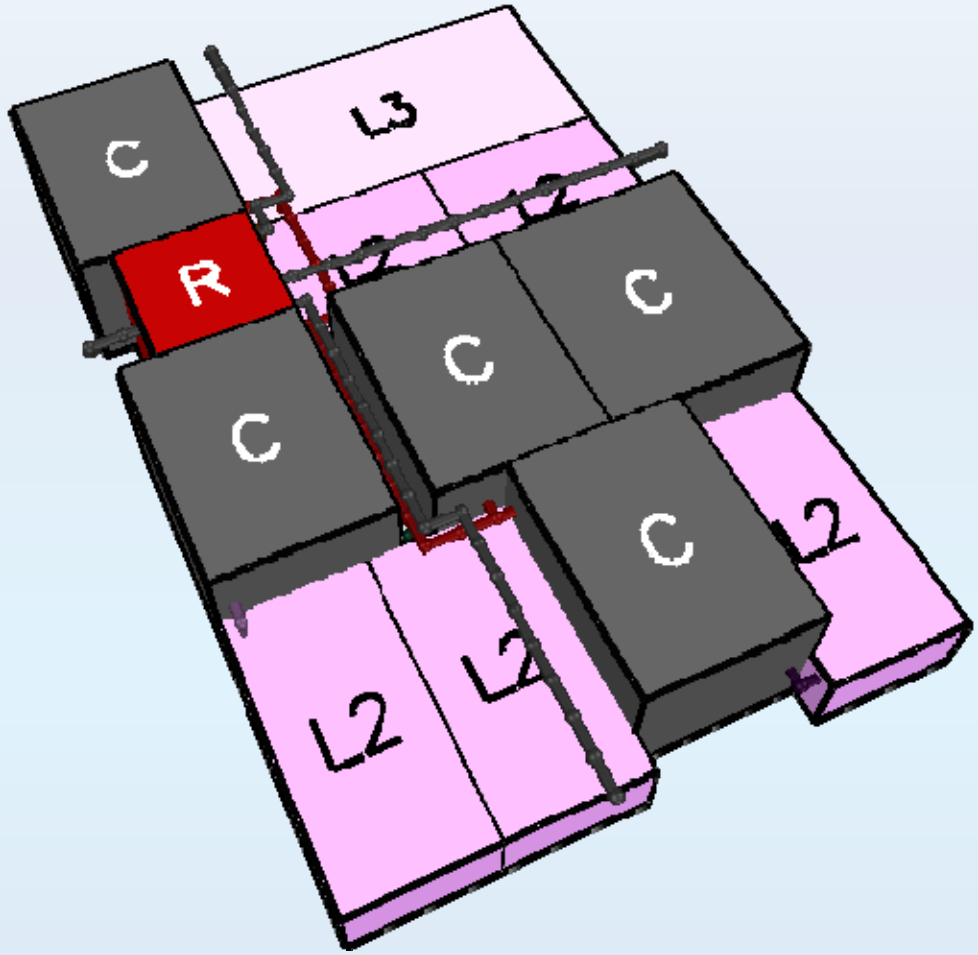
Filtering floorplans



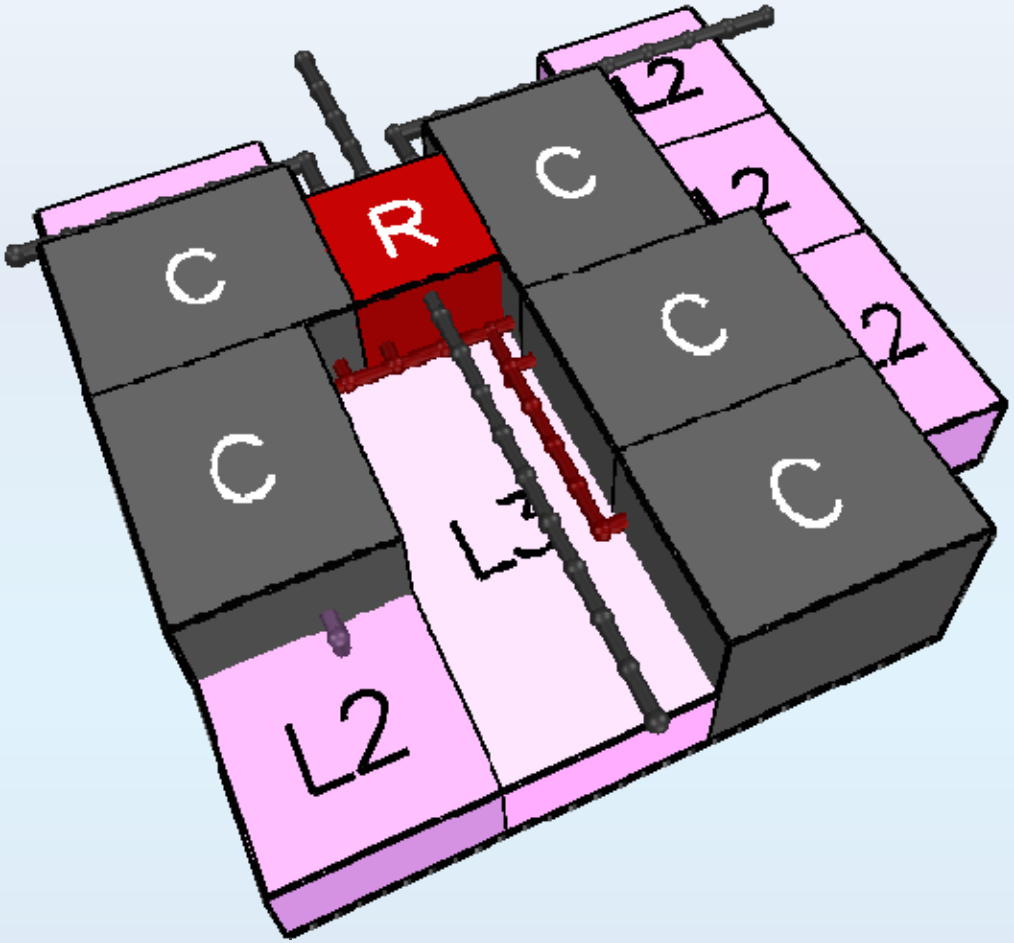
After physical planning



After physical planning



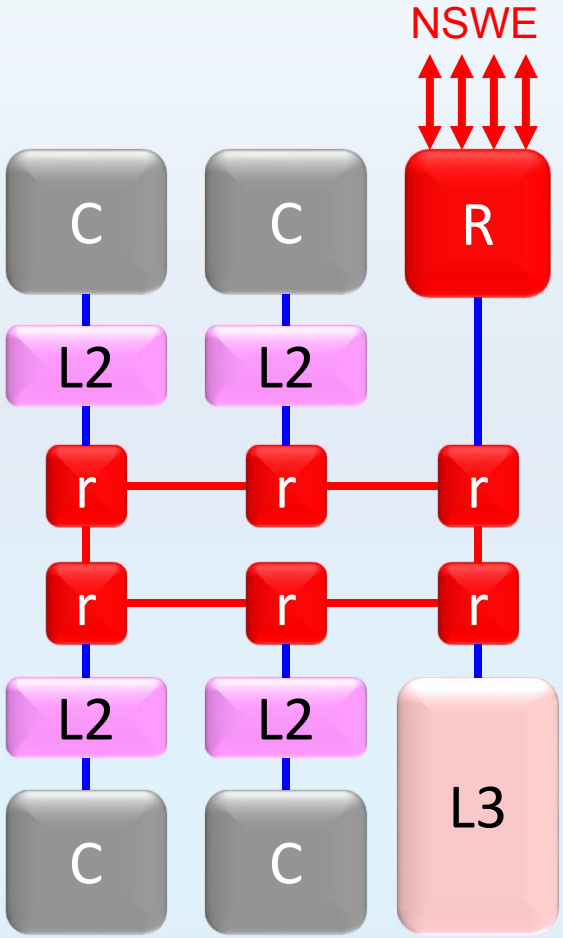
After physical planning

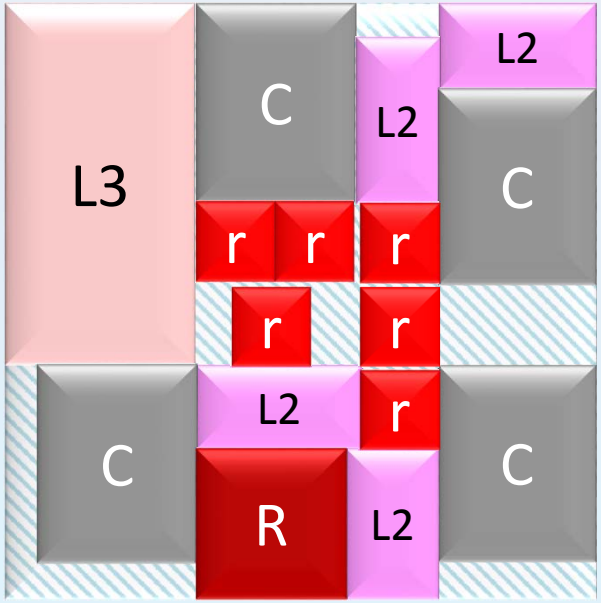


Outline

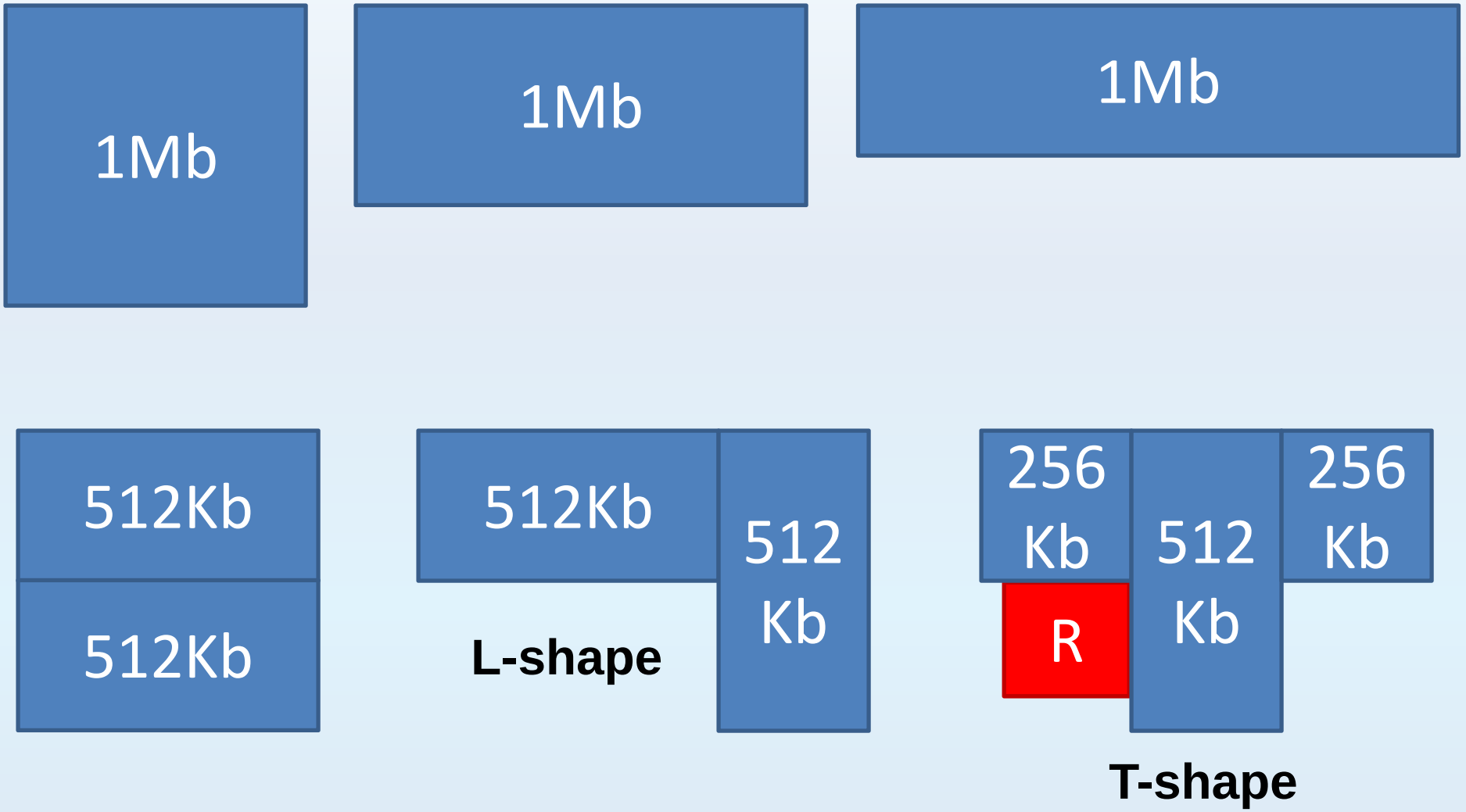
- Architectural exploration
- Physical planning for tiled CMPs
- **Current work: regular floorplanning**
  - **Memory floorplanning**
  - **Regularity extraction**

Min-area floorplan

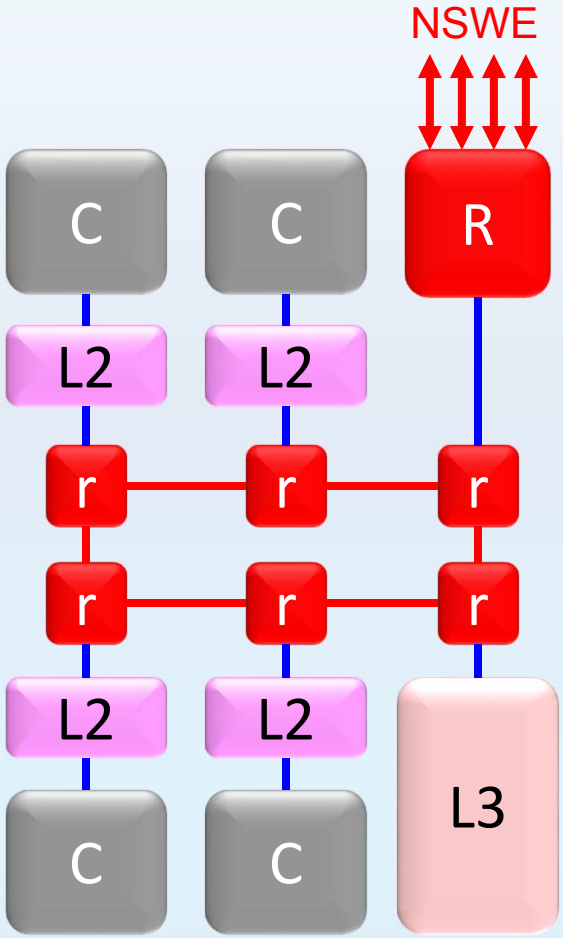


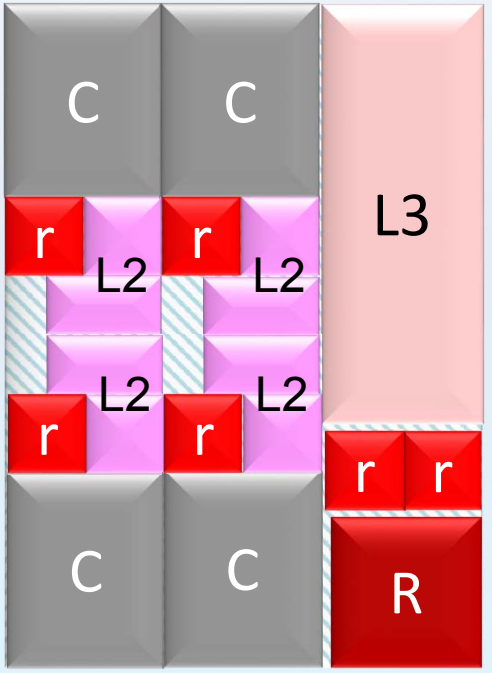


Integrated memory floorplanner

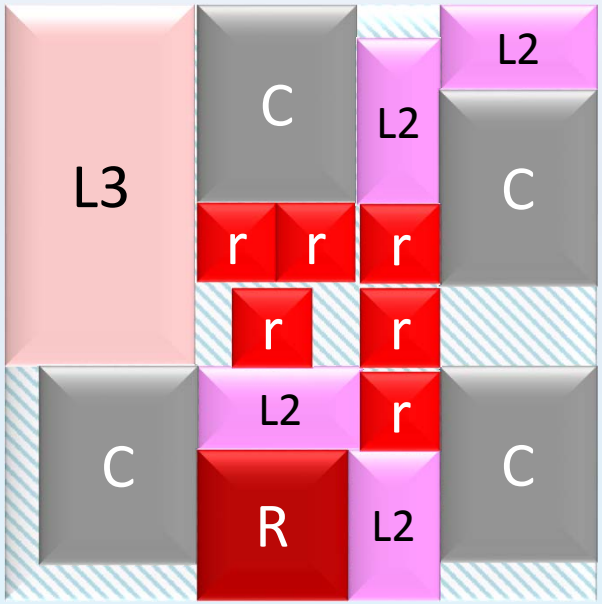


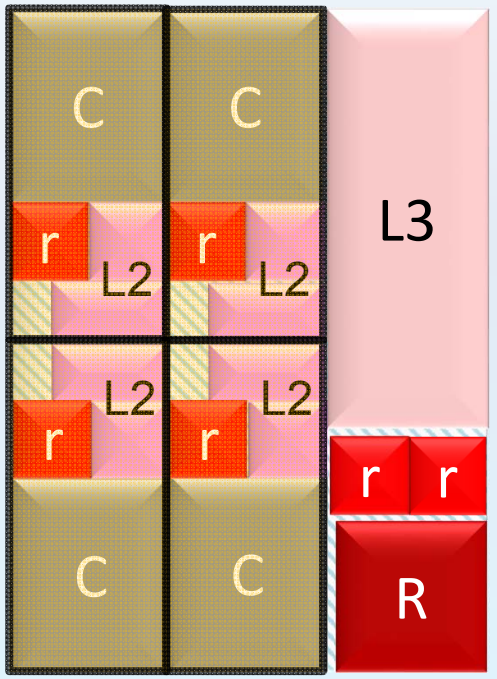
Regular floorplan





Regular floorplan

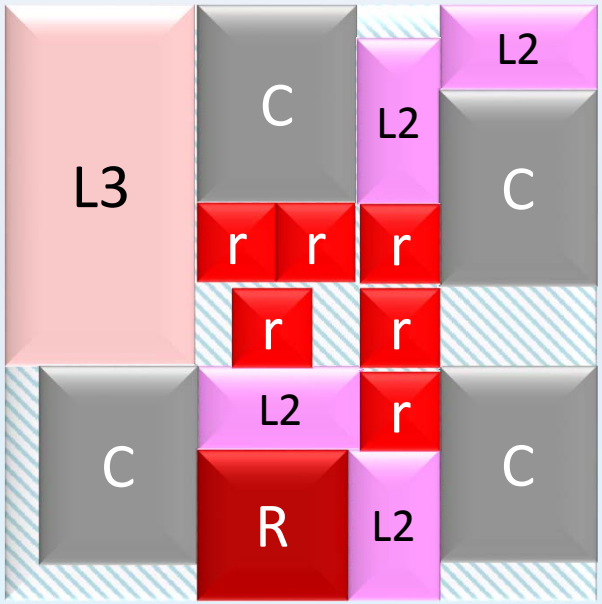


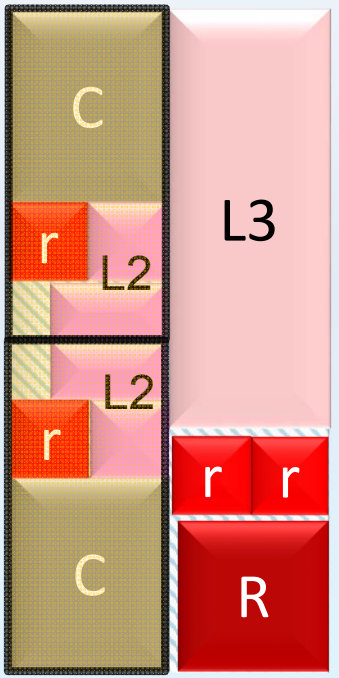


Regularity:

- Smaller design effort
- Efficient timing closure
- Choppability

Regular floorplan





Regularity:

- Smaller design effort
- Efficient timing closure
- Choppability

Exploration:

- Graph based knowledge discovery
- Hierarchical slicing structures
- Simulated Annealing

Conclusions

- Physical information is essential for the architectural exploration of tiled CMPs
- Approach: divide & conquer
  - 1-cluster floorplan + abutment constraints
  - Build tiled CMPs (scalable and choppable)
- Ongoing work:
  - Multi-module memory floorplanning
  - Regularity