# Data-Driven Scenario-Based Application Mapping For Hetero-geneous Many-Core Systems

Jan Spieck, Stefan Wildermann, Tobias Schwarzer, Jürgen Teich, Michael Glaß*

Hardware-Software-Co-Design

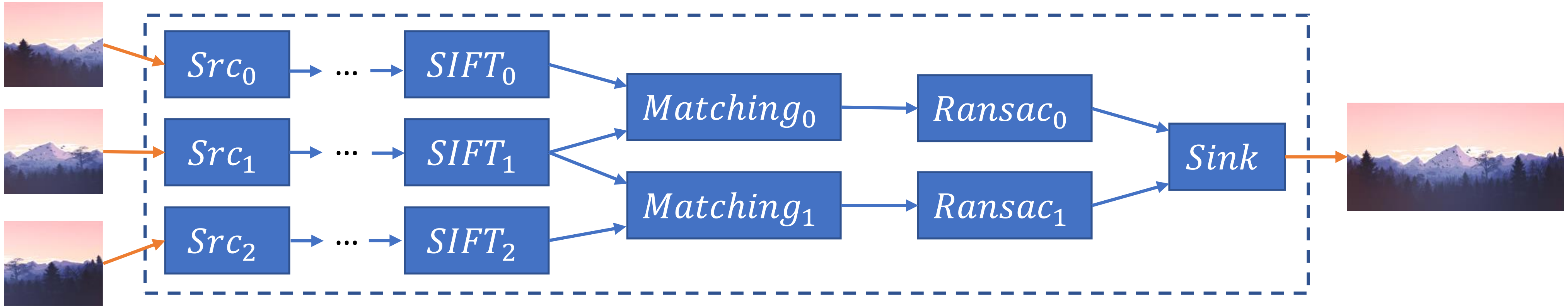
Friedrich-Alexander-Universität Erlangen-Nürnberg

*Universität Ulm

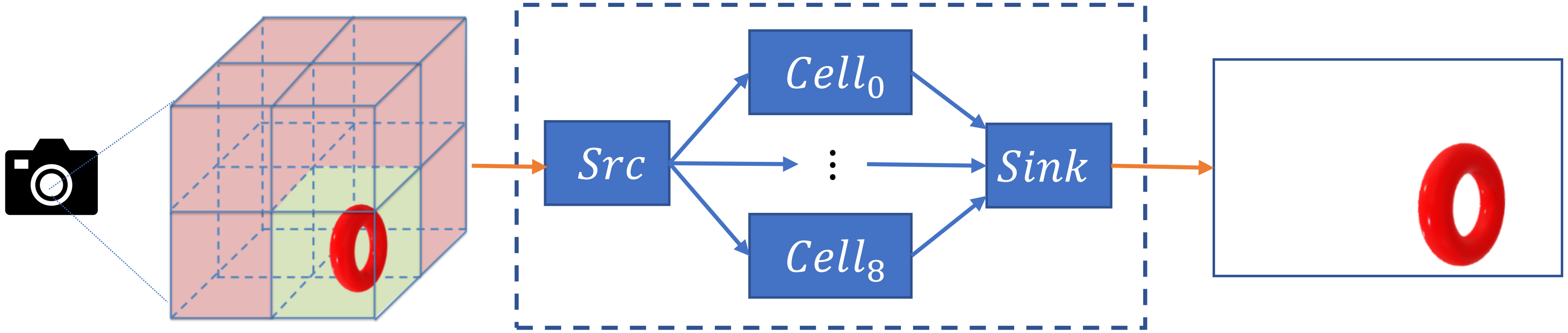
# Motivation: Input-Dependent Workload

Task-based applications with input-dependent workload distribution, e.g.:

Stitching




Ray tracing

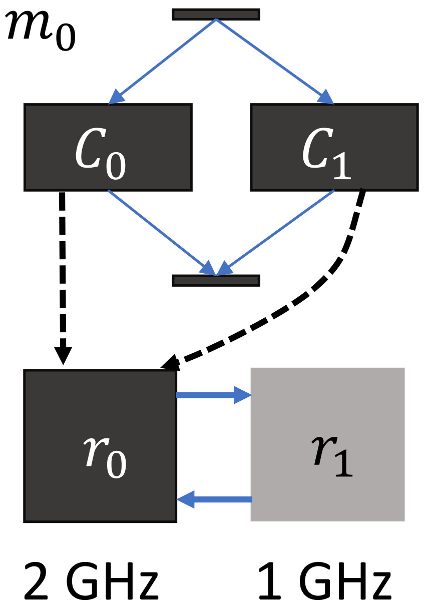


# Motivation: Data-Driven Mapping I

Task: Mapping application tasks onto a heterogeneous architecture

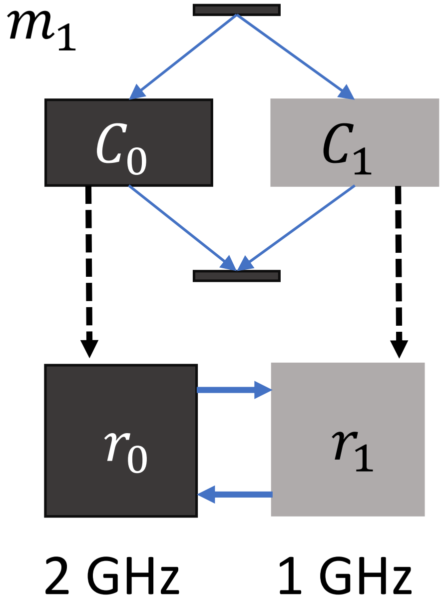
Problem: Single mappings do not exploit specializations of input data

Example: Processing subsequent data $d_0 = \{100, 200\}$ and $d_1 = \{200, 100\}$ triangles




| $d$ | $C_0$ | $C_1$ |
|---|---|---|
| $d_0$ | $50\ ms$ | $100\ ms$ |
| $d_1$ | $100\ ms$ | $50\ ms$ |

| $d$ | End to end |
|---|---|
| $d_0$ | $150\ ms$ |
| $d_1$ | $150\ ms$ |
| | $\sum$: $300\ ms$ |

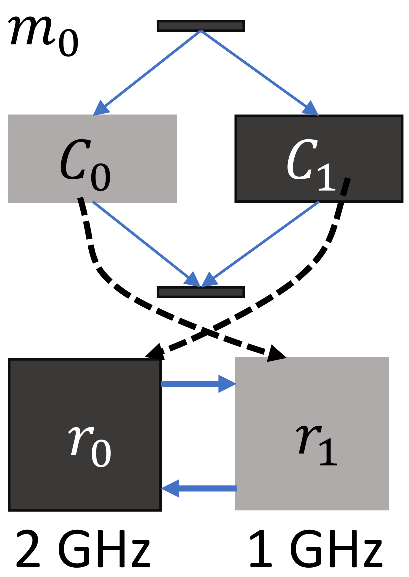


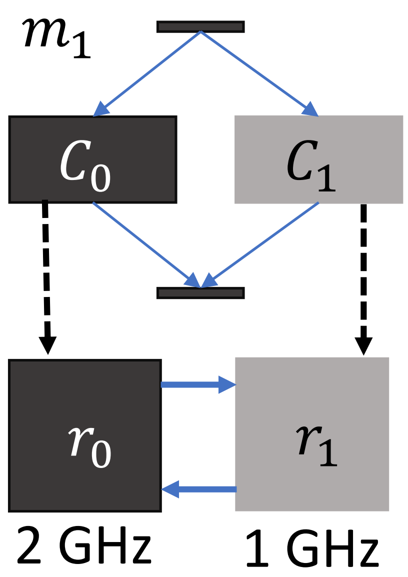

| $d$ | $C_0$ | $C_1$ |
|---|---|---|
| $d_0$ | $50\ ms$ | $200\ ms$ |
| $d_1$ | $100\ ms$ | $100\ ms$ |

| $d$ | End to end |
|---|---|
| $d_0$ | $200\ ms$ |
| $d_1$ | $100\ ms$ |
| | $\sum$: $300\ ms$ |

# Motivation: Data-Driven Mapping II

Solution: Partitioning the data space $D = \bigcup_k D_k$ into scenarios $D_k$




$$D_0 = \{d_1\} \rightarrow m_0$$
$$D_1 = \{d_0\} \rightarrow m_1$$

| $d$ | $C_0$ | $C_1$ |
|---|---|---|
| $d_0$ | — | — |
| $d_1$ | $100\ ms$ | $100\ ms$ |

| $d$ | $C_0$ | $C_1$ | End to end |
|---|---|---|---|
| $d_0$ | $100\ ms$ | $100\ ms$ | $100\ ms$ |
| $d_1$ | — | — | $100\ ms$ |

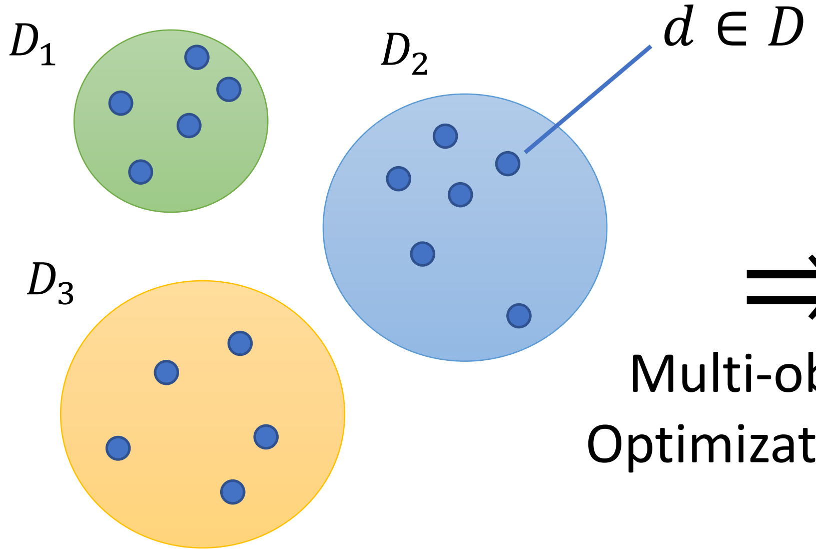
$\Sigma$: $200\ ms$

But how do we determine these scenarios and scenario-optimized mappings?

# Multi-Objective Optimization Problem I

Problem 1: find scenario-optimized mappings $m \in M$

Given: scenario distribution $S = (D_1, \dots, D_n)$




$\Rightarrow$

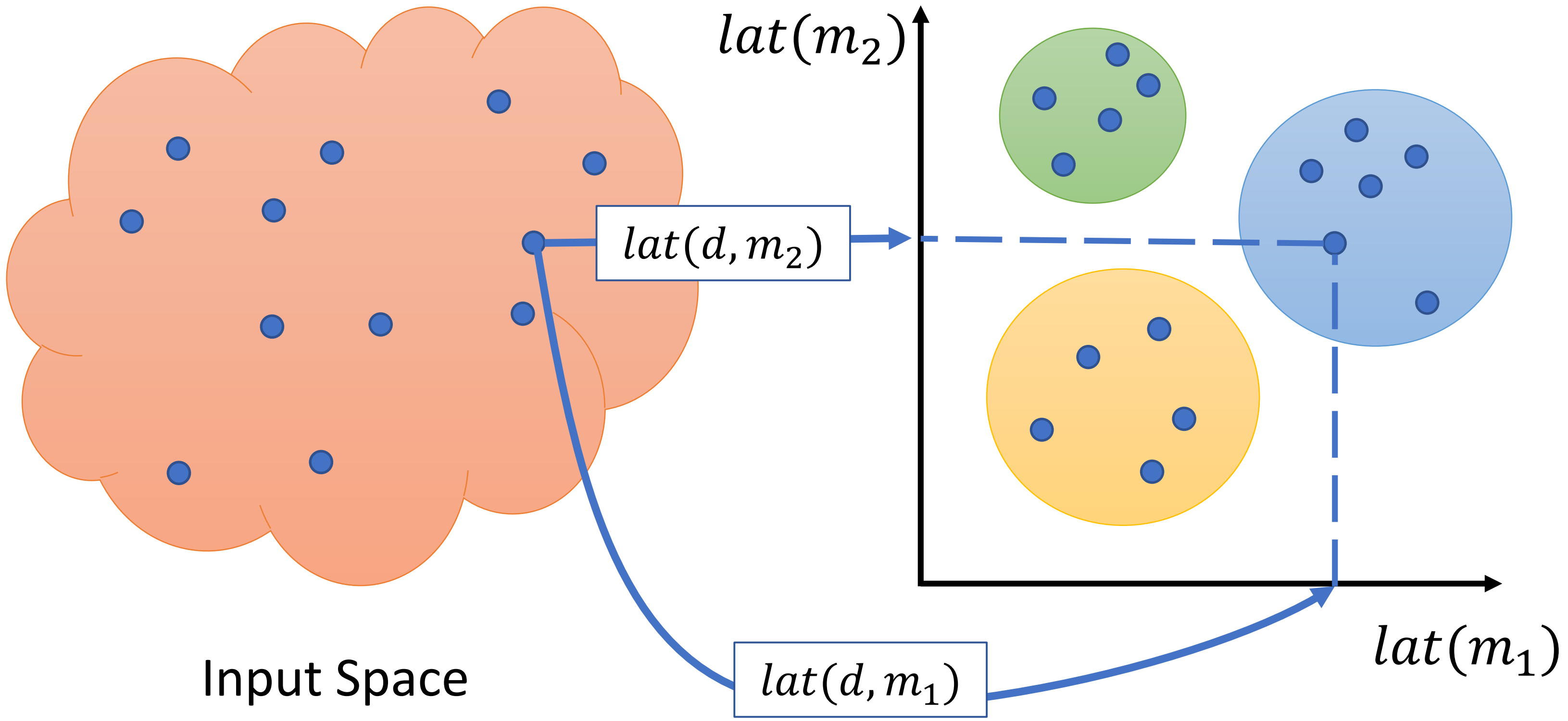
Multi-objective Optimization (DSE)

| | Performance | | | Minimal Resources | |
|---|---|---|---|---|---|
| $\boldsymbol{m}$ | $\boldsymbol{p(D_1)}$ | $\boldsymbol{p(D_2)}$ | $\boldsymbol{p(D_3)}$ | $\#(\boldsymbol{R_1})$ | $\#(\boldsymbol{R_2})$ |
| $m_1$ | 10 | 2 | 2 | 4 | 1 |
| $m_2$ | 8 | 6 | 6 | 3 | 0 |
| $m_3$ | 4 | 10 | 5 | 4 | 2 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

Pareto-optimal mappings
$M = \{m_1, m_2, m_3, \dots\}$

# Multi-Objective Optimization Problem II

Problem 2: Given a set of optimized mappings $M = \{m_1, \ldots\}$ find a scenario distribution $S \in S^{\circ}$




Scenarios

S= $\{D_1, D_2, D_3\}$

Scenario $D_i \subseteq D$:
All data which performs equally/similarly for different mappings $m_j$

# Circular Dependency

Given: scenario distribution $S$
Find: optimized mappings $M$

Given: optimized mappings $M$
Find: optimal scenario distribution $S$

# Design-Time Optimization

Solution: Iterative scenario-based design space exploration

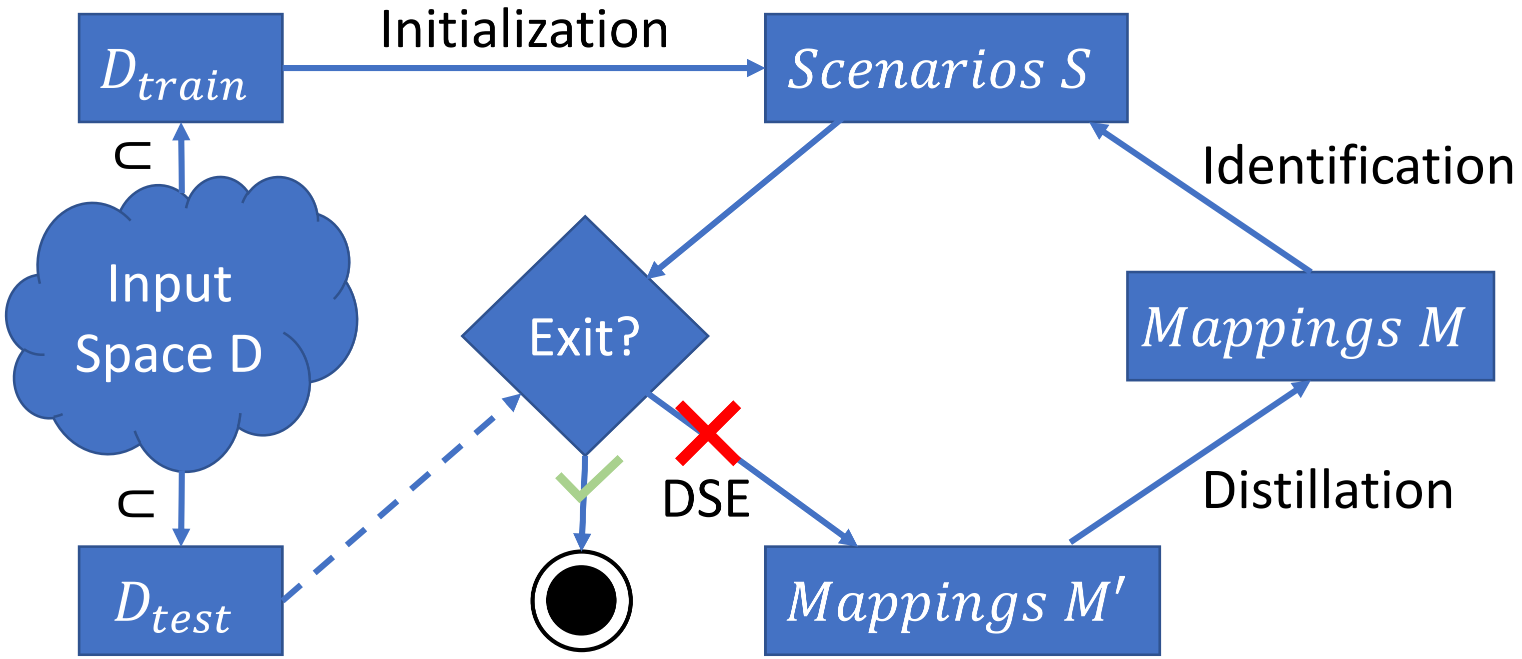
Steps:

1. Input Generation
2. Scenario Initialization

Loop:

3. Design space exploration
4. Distillation
5. Scenario Identification

6. Termination



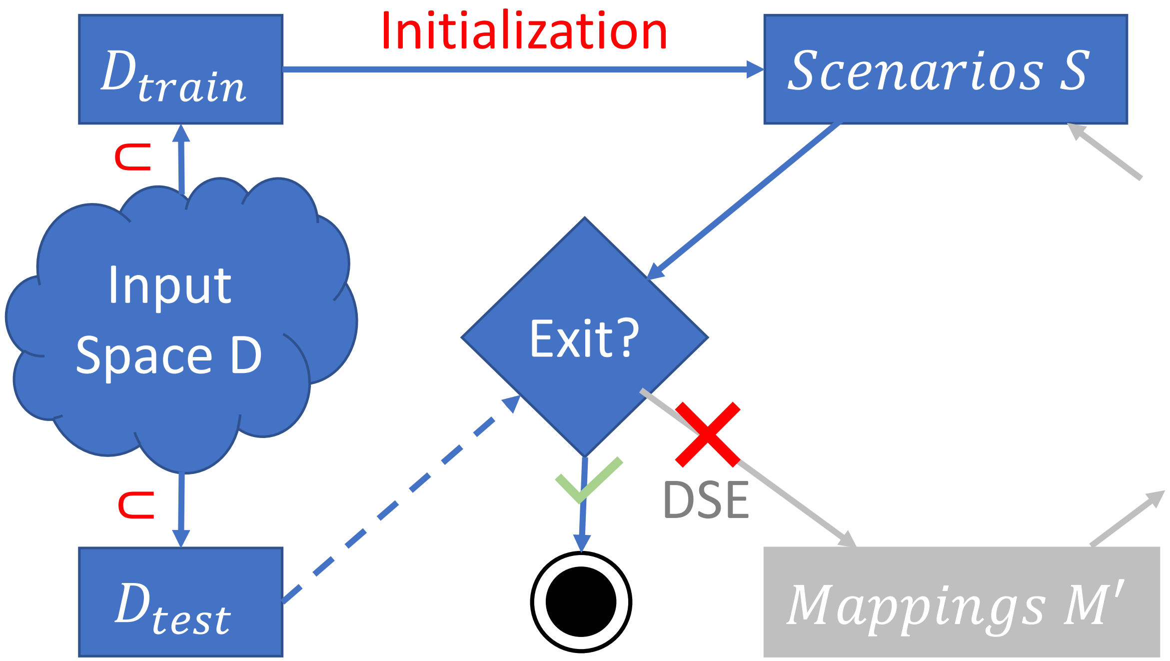
# Iterative Optimization Loop I

Input Generation: select representative subset of data

- $D_{train} \subset D, D_{test} \subset D$
- $D_{train} \cap D_{test} = \emptyset$

Scenario Initialization:

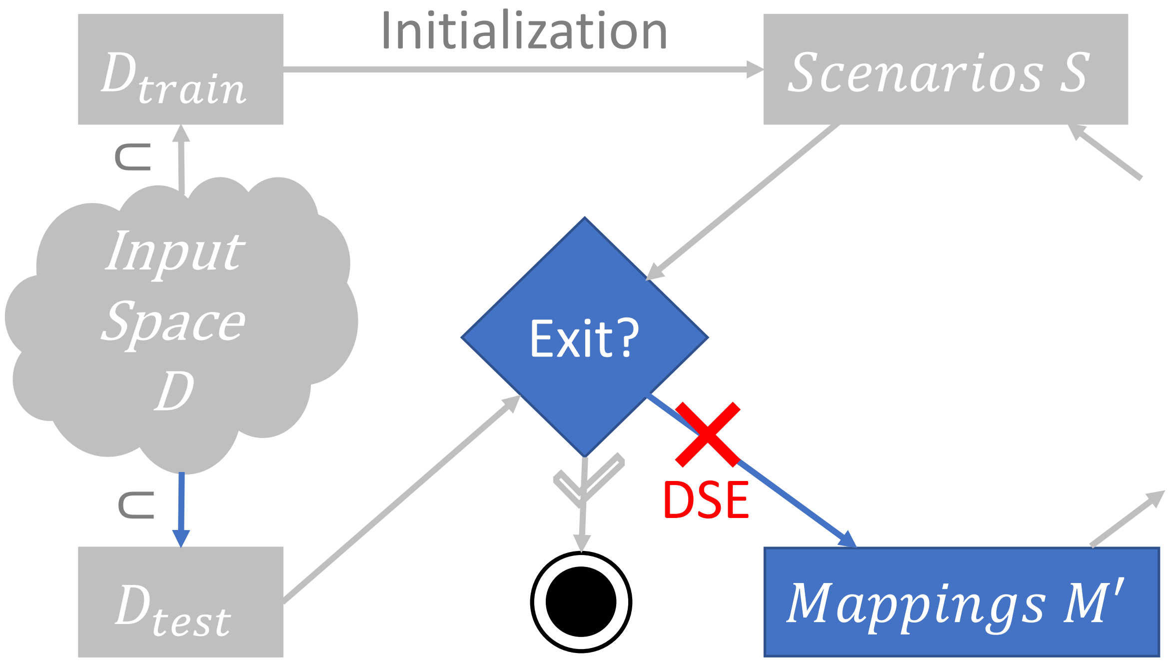
- Random scenario distribution
- Clustering on default mappings



# Iterative Optimization Loop II

Design space exploration (DSE):

- $$minimize \begin{pmatrix} p(D_1, m) \\ \vdots \\ p(D_n, m) \\ |R_1(m)| \\ \vdots \\ |R_u(m)| \end{pmatrix}$$
- Using evolutionary algorithms



# Iterative Optimization Loop III

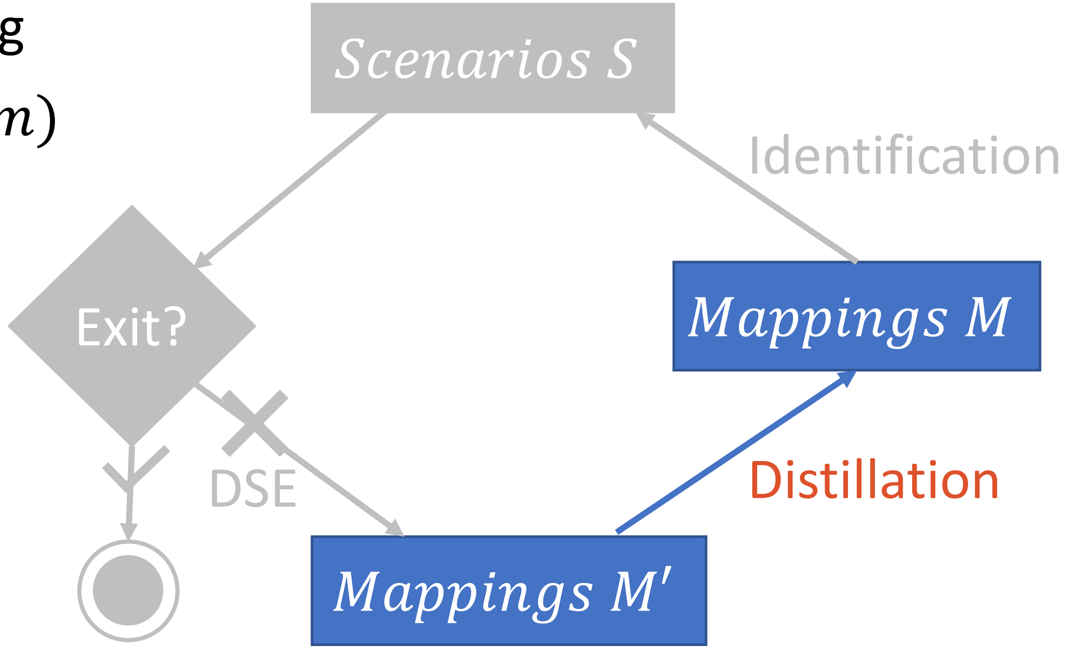
Distillation: Reduce resulting set $M'$ to a (smaller) set $\mathrm{M} \subseteq M'$

- Improves identification step
- Option 1: clustering over mappings and sampling
- Option 2: based on a weighted sum over $p(D_k, m)$

Example: $p(D_k, m) = (latency, energy)$

$w_p = latency + 0.5 \cdot energy \quad |M| = 2$

| $M'$ | $latency$ | $energy$ | $w_p$ |
|---|---|---|---|
| $m_1$ | $10\ ms$ | $20\ mJ$ | $20$ |
| $m_2$ | $20\ ms$ | $10\ mJ$ | $25$ |
| $m_3$ | $30\ ms$ | $5\ mJ$ | $32.5$ |



# Iterative Optimization Loop IV

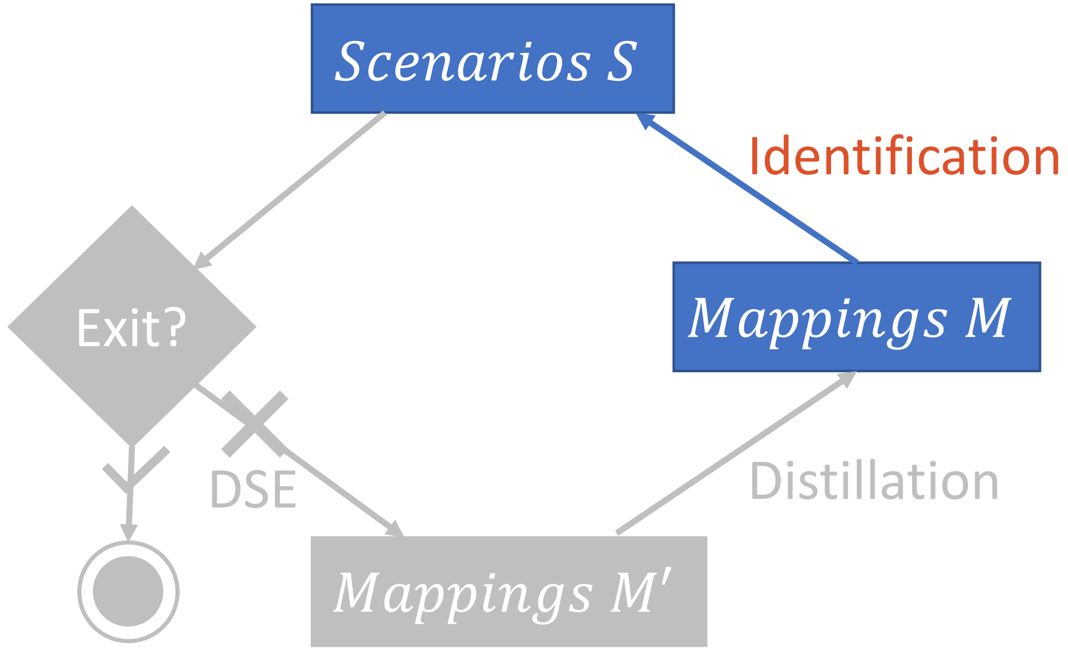
Scenario Identification: $S = \arg \min_{(D_1,\dots,D_n) \in S^\circ} \sum_{k=1}^{n} dist(D_k, M)$

Option 1: Clustering (e.g., K-Means)

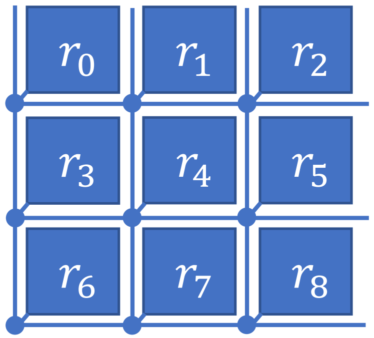
- Performance per mapping $m_i \in M$
- $v(d) = [p(d, m_1) \dots p(d, m_l)]^T$
- $dist(D_k, M) = \sum_{d \in D_k} ||v(d) - \mu_k||^2$

Option 2: Performance Optimization

- $dist(D_k, M) = \min_{m_k \in M_i} \{\sum_{d \in D_k} p(d, m_k)\}$
- Best suited for low-dimensional $p(d, m)$



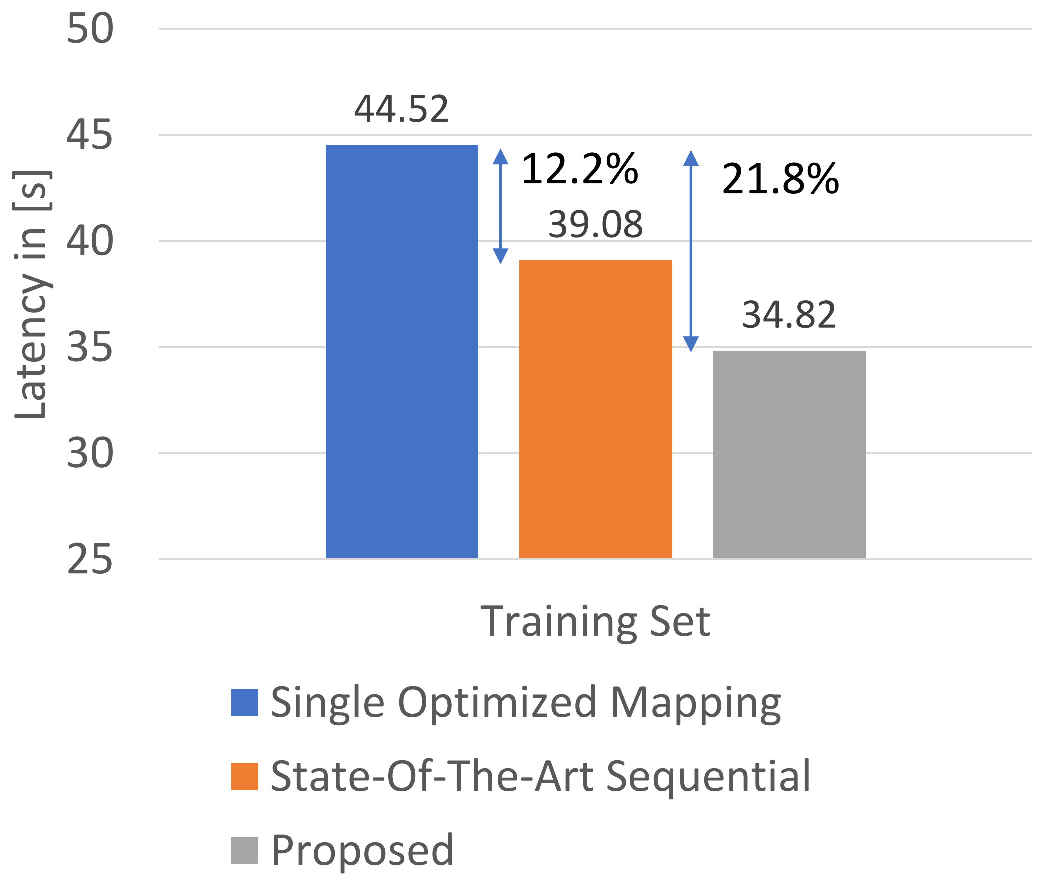
# Evaluation Setup

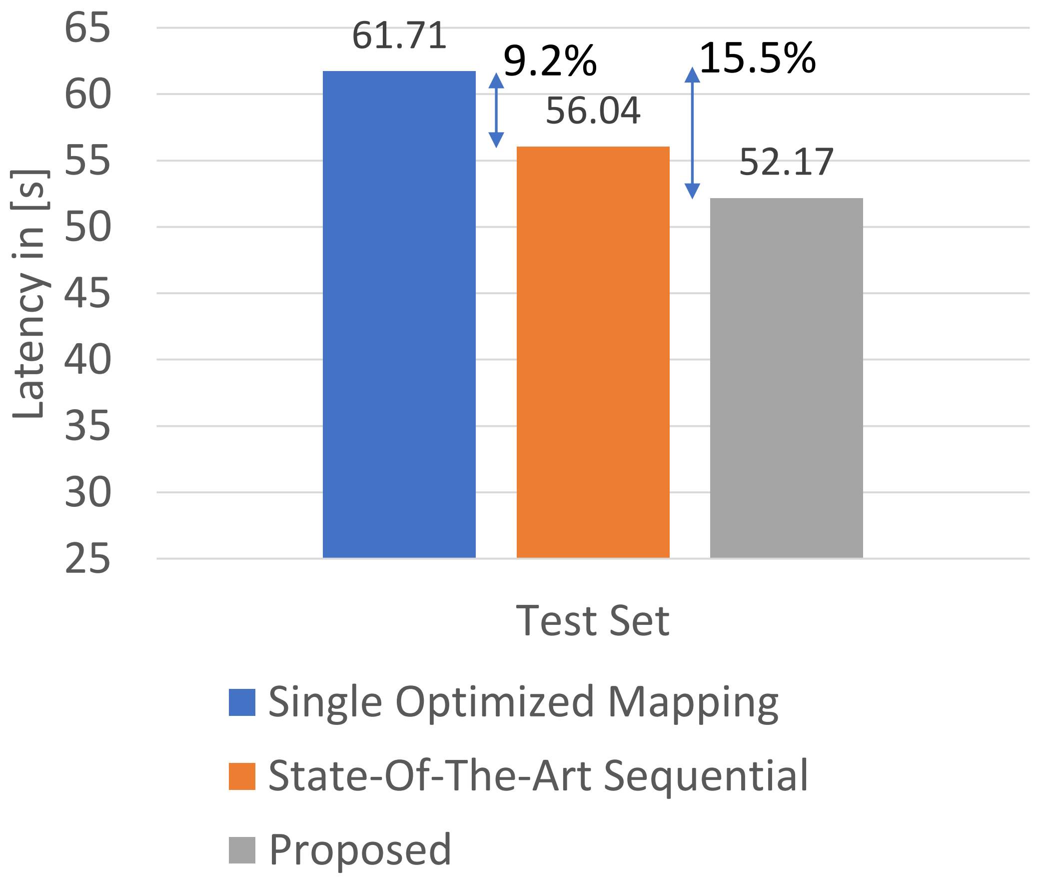



- Applications (Data):
  - Ray tracing (virtual 3D-scenes)
  - Stitching (partial images of panoramas)
- Architecture: heterogeneous 3x3 NoC mesh
- Data: split into training and test set
- Each test data is executed in the best-suited scenario $D_i \in S$
- Goal: minimal latency for processing the total scenario distribution

# Eval.: Ray Tracing Latency

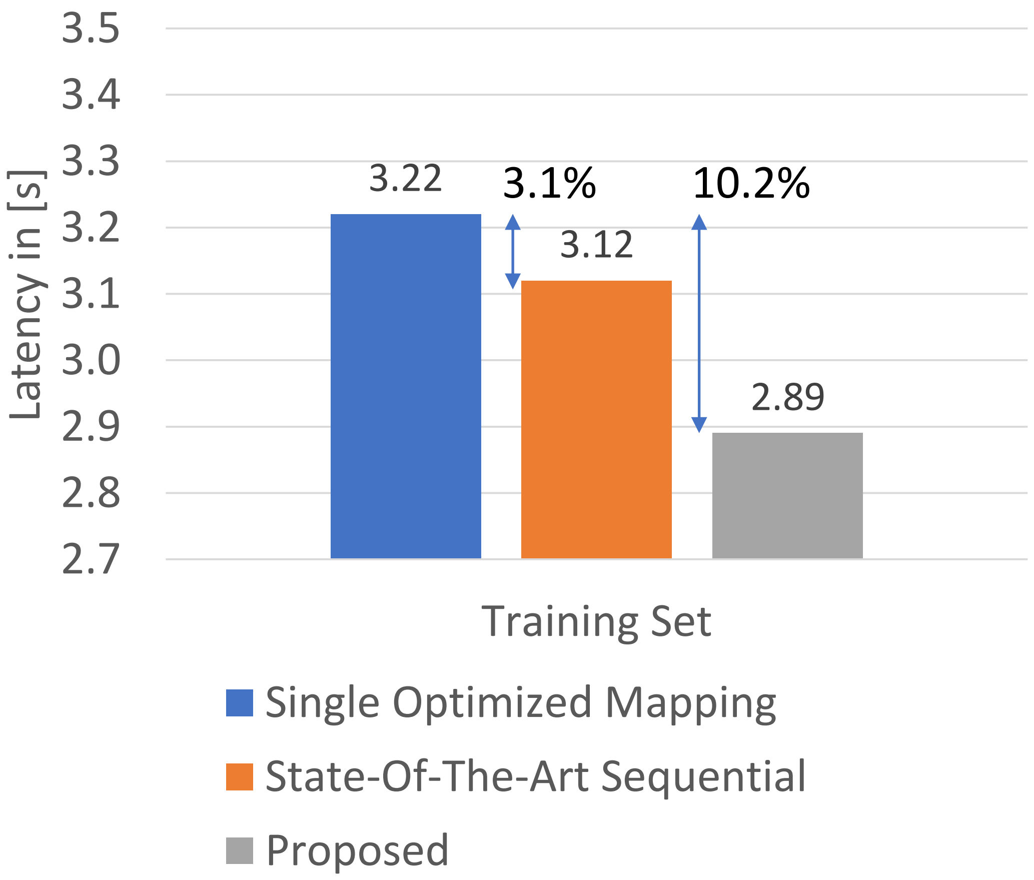
Latency for different optimization approaches (test set with bigger scenes)

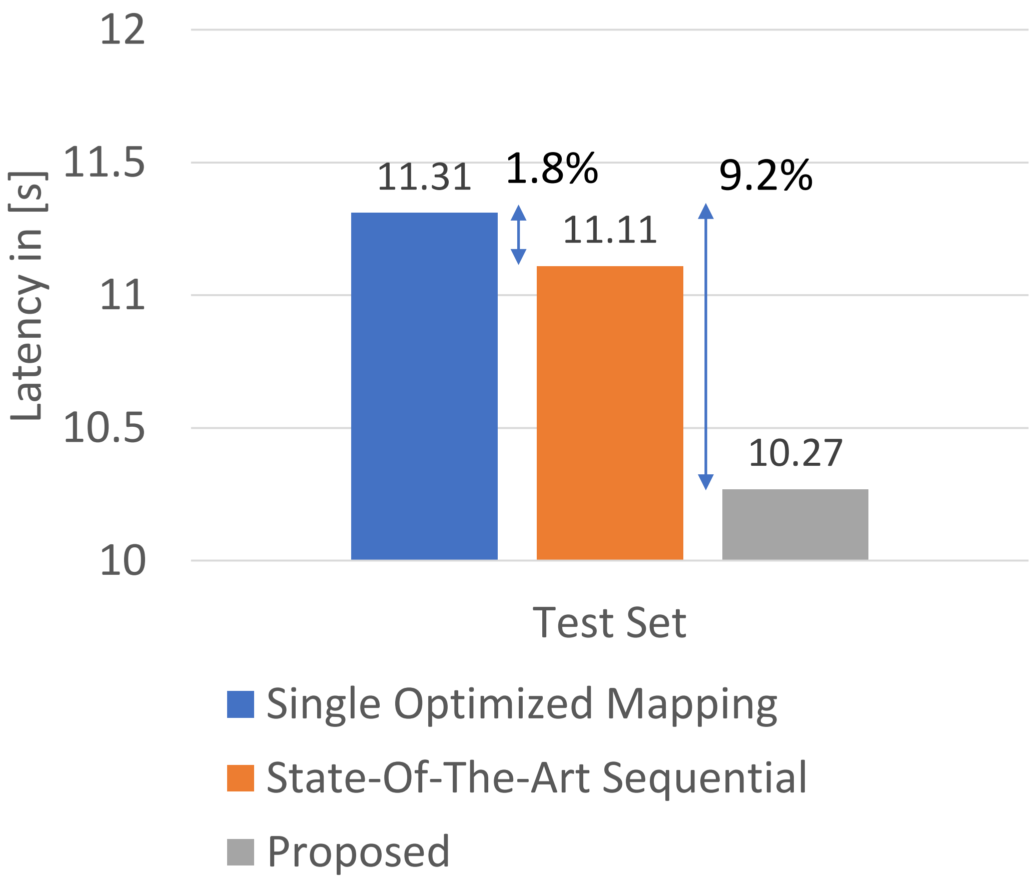


# Eval.: Stitching Latency

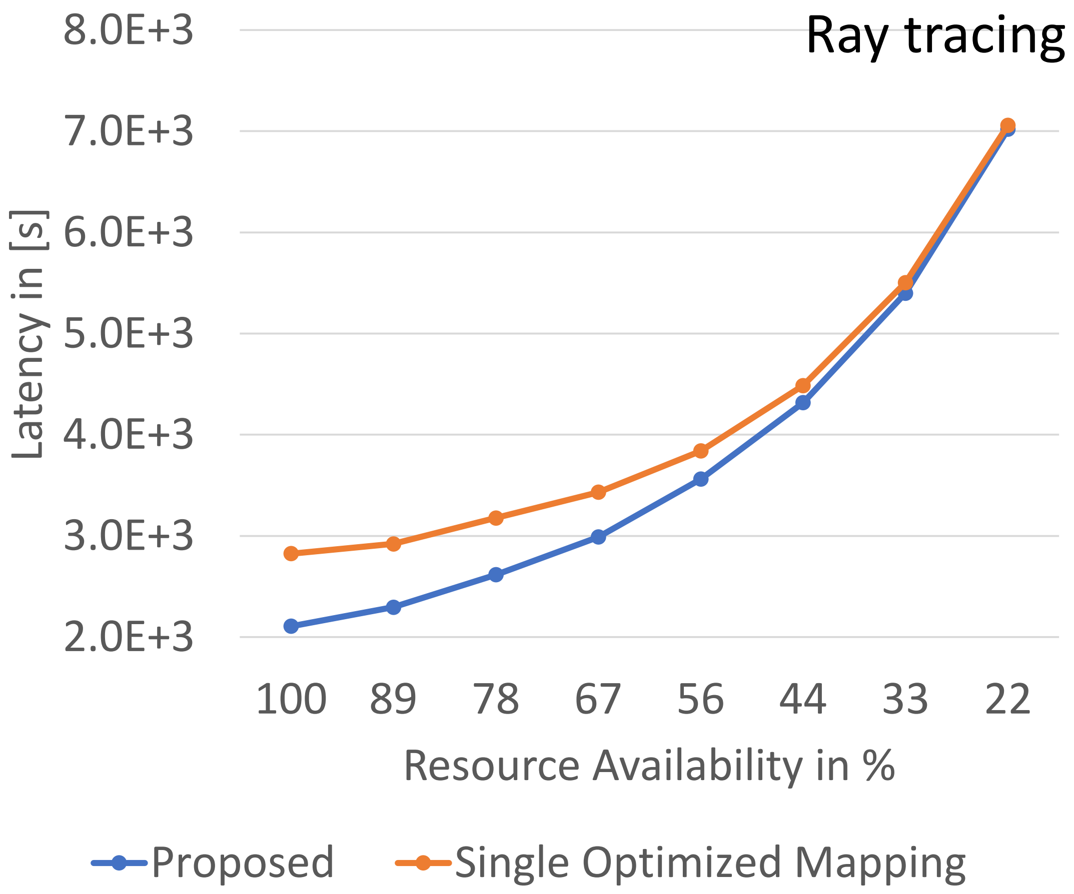
Latency for different optimization approaches (test set with bigger images)

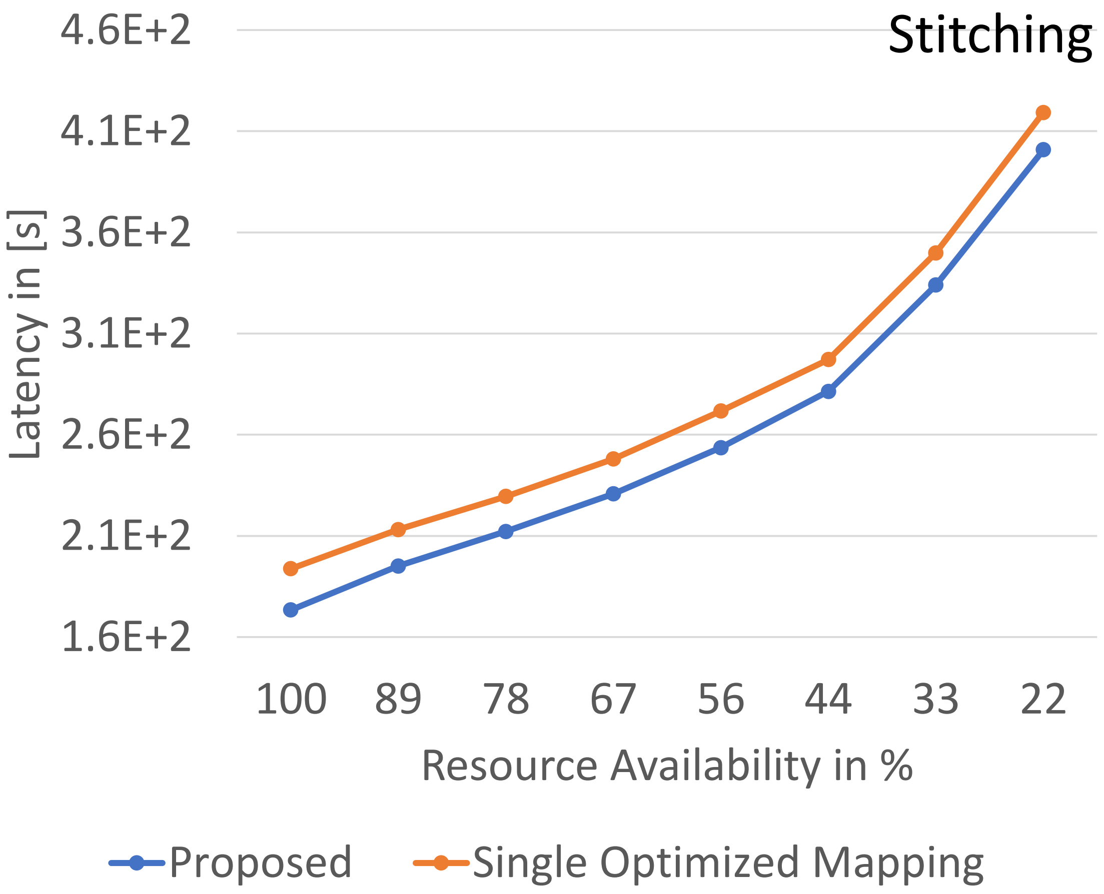


# Eval.: Hybrid Application Mapping

Latencies for different resource availability (considered during DSE by $R_i(m)$):



# Conclusion

- Mapping optimization for applications with input-dependent task workload onto heterogeneous architectures
- Scenario-based design space exploration

1. Input Generation
2. Scenario Initialization
3. Design space exploration
4. Distillation
5. Scenario Identification
6. Termination

- Significant speedup compared to a single optimized mapping for the average-case (15 % ray tracing, 10 % stitching (test set))

# Thanks for listening!

This work was supported by the German Research Foundation (DFG) as part of the Transregional Collaborative Research Centre „Invasive Computing“ (SFB/TR 89)

Are there any questions?

# Run-Time Manager

At run time: Optimize latency of data mappings

Given: Sequence of data with unknown scenario affiliation