

# Adaptive Embedded Processing with RISPP

by Jörg Henkel  
together with Lars Bauer, Muhammad Shafique, ...

Chair for Embedded Systems (CES)

University of Karlsruhe, Germany



J. Henkel

Talk @ MPSoC'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>

2

## Development of Embedded Systems

### □ Typical:

- **Static analysis of hot spots**
- **Building tightly optimized system**

### □ Nowadays:

- **Increasing complexity**
- **More functionality**

### □ Problem:

- **Statically chosen design point** has to match all requirements
- **Typically inefficient** for individual components (e.g. tasks or hot spots)



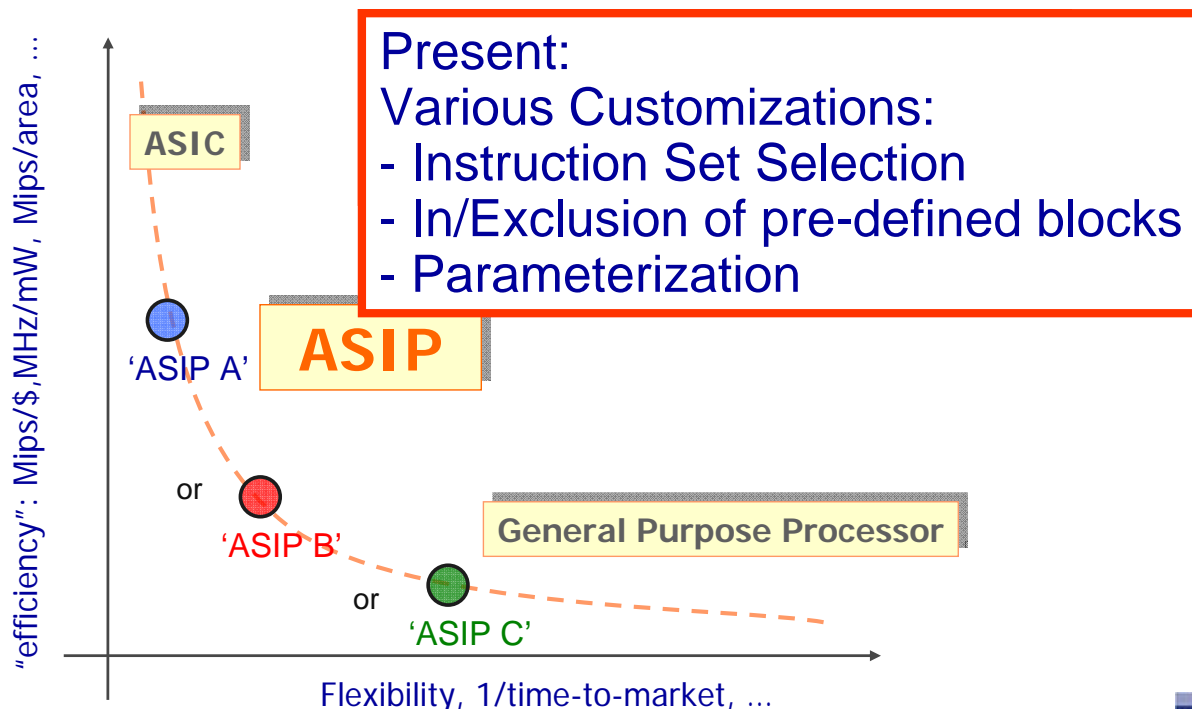
J. Henkel

Talk @ MPSoC'08, June 23rd

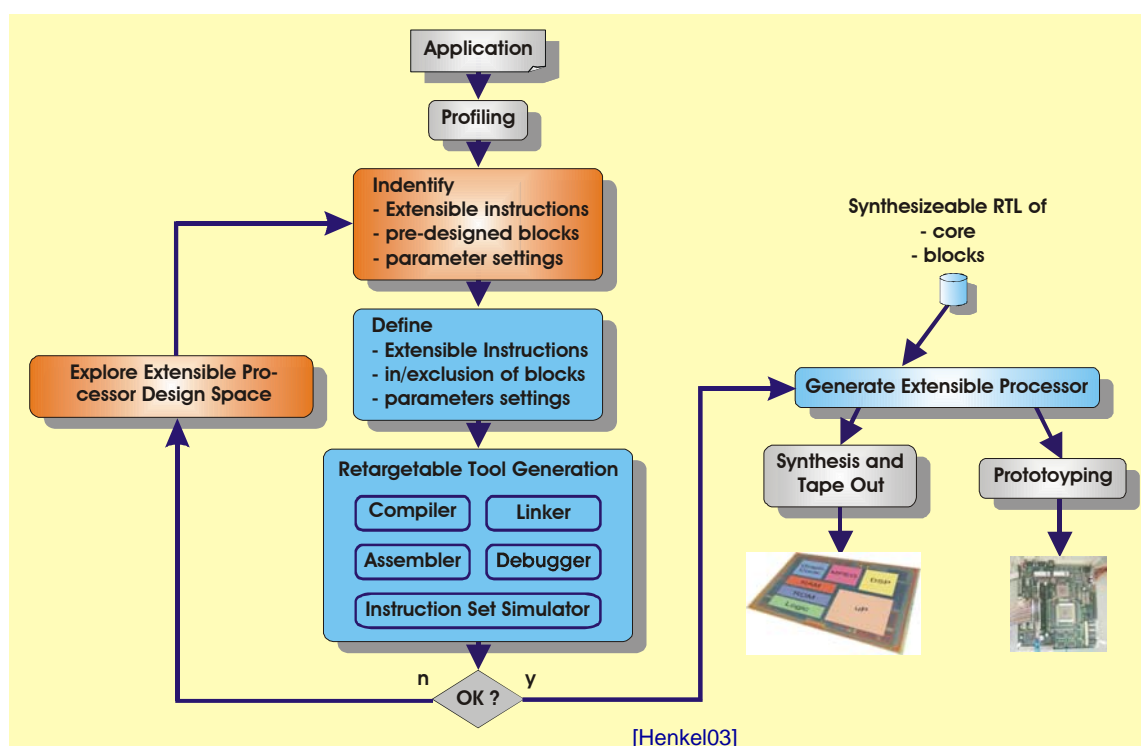
<http://ces.univ-karlsruhe.de/RISPP>



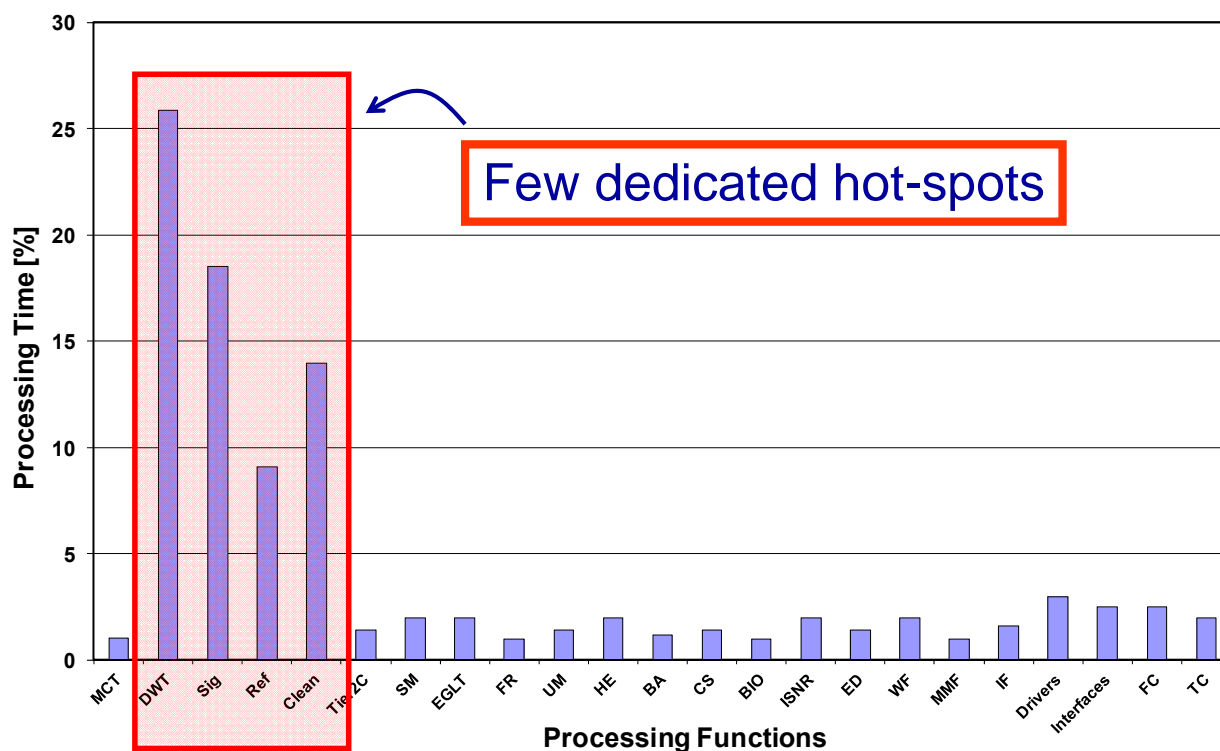
# The place of ASIPs: from past to present



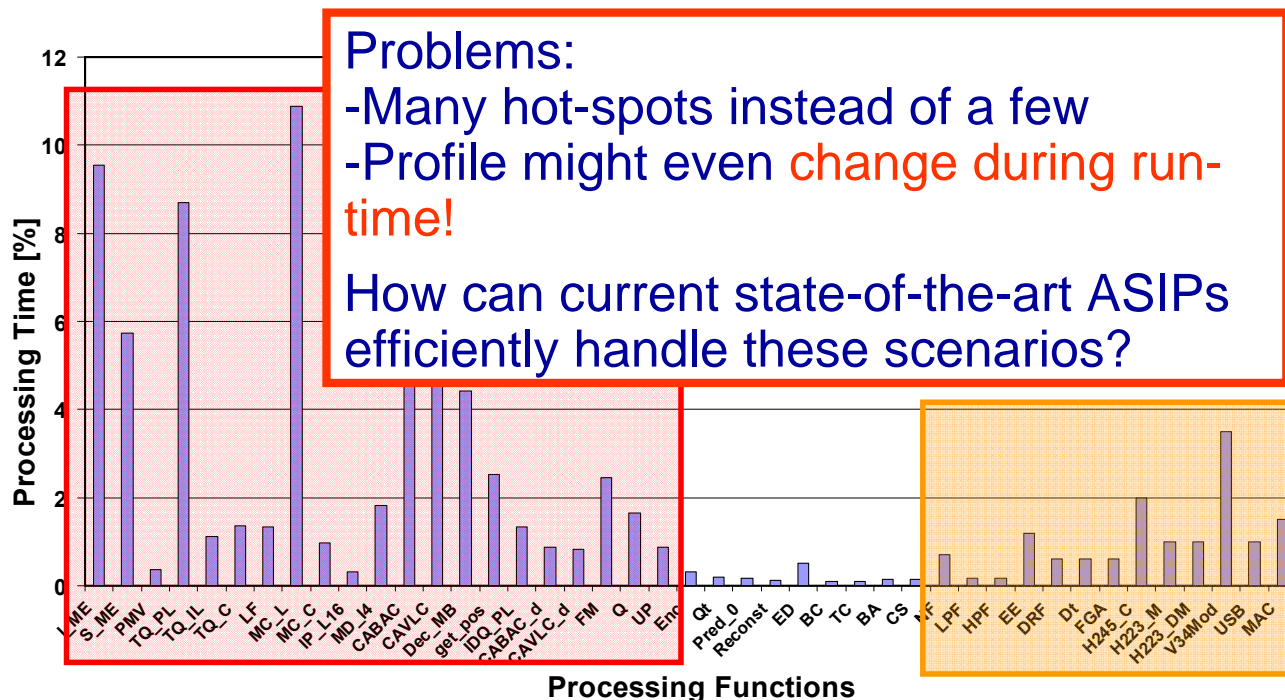
## State-of-the-Art ASIP Design Flow



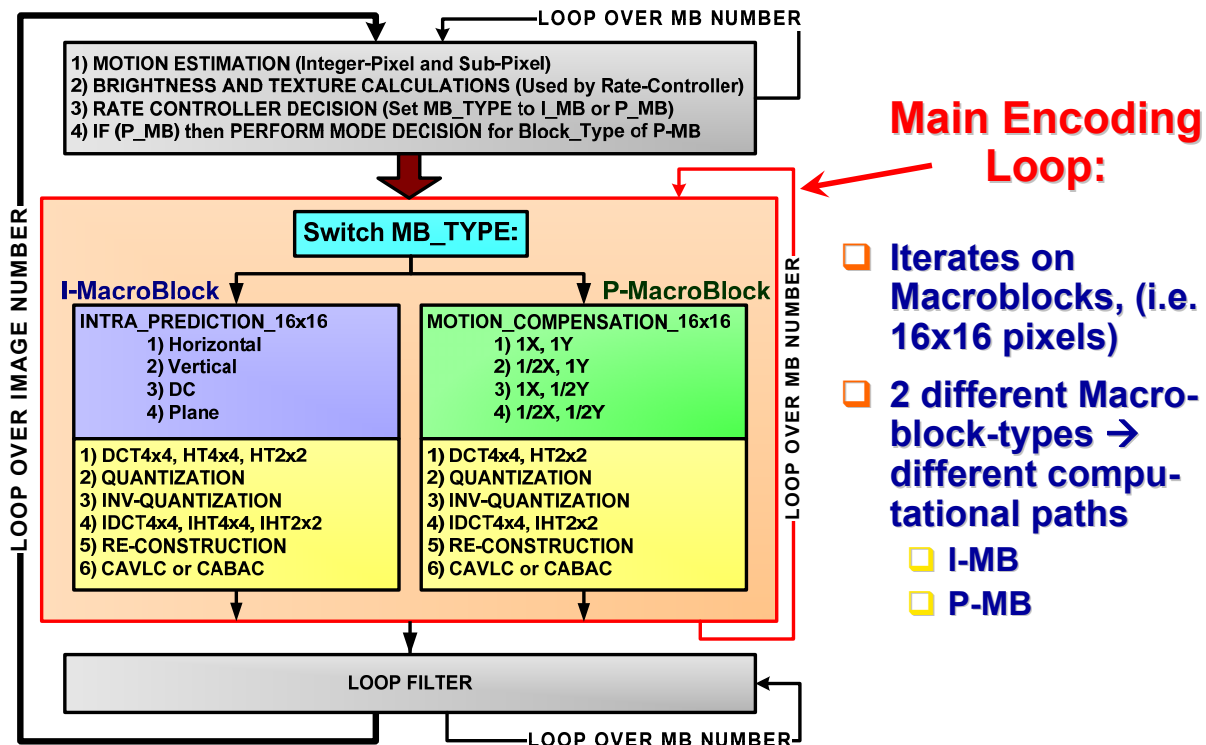
# Typically ...



# But what if ...



# Example: Execution Flow of an H.264 Encoder



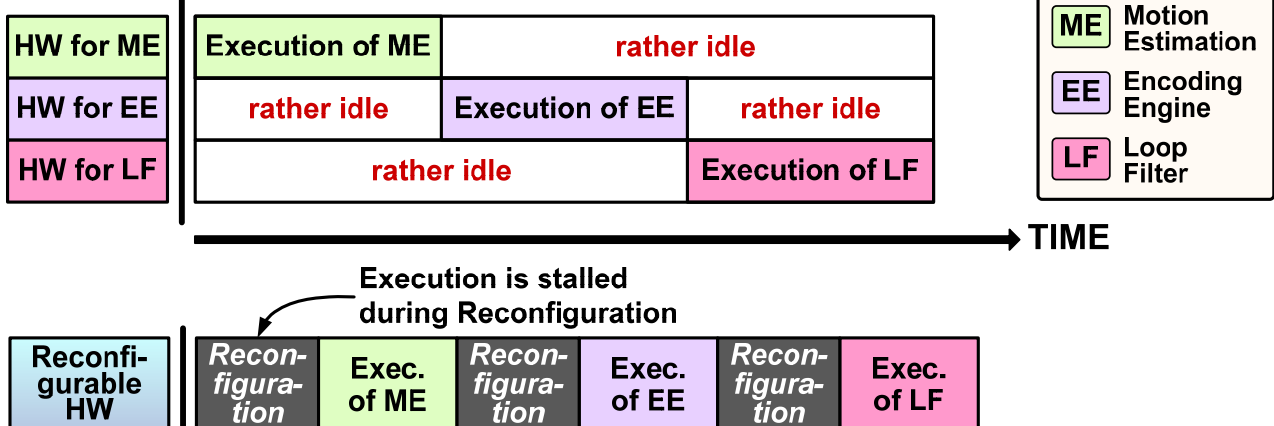
J. Henkel

Talk @ MPSoc'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>

## ASIP Efficiency problems when targeting multiple hot spots

Rather huge accelerators (to exploit parallelism) targeting full hot spots



### Reconfigurable Computing:

- + Efficient use of hardware
- Potentially performance degradation due to reconfiguration time & FPGA fabric

J. Henkel

Talk @ MPSoc'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>



# Related Work

## ❑ Extensible Processors:

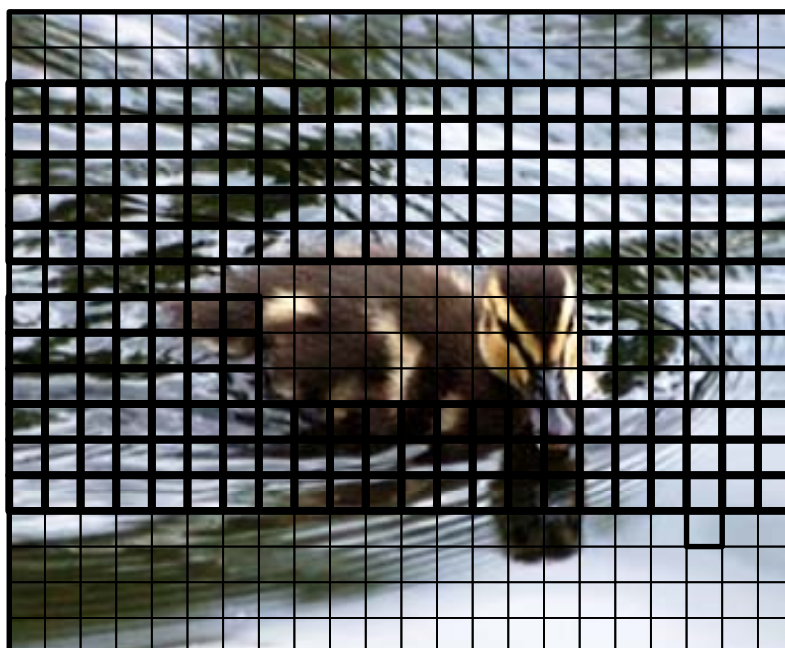
- ❑ Automatic detection & synthesis of Special Instructions
- ❑ Automatic generation of Tool-Suits (simulator, compiler, ...)
- ❑ Early estimations to guide through the design space

## ❑ Reconfigurable Computing Systems:

- ❑ Basic technique that enables the hardware to adapt to different requirements
- ❑ Data-Flow driven systems
- ❑ Control-Flow driven (typically CPU extended by a reconfigurable Co-Processor)

# Uncertainty in computation effort

## ❑ Distribution of I- and P- Macroblocks: Prediction from previous frame or surrounding MBs?



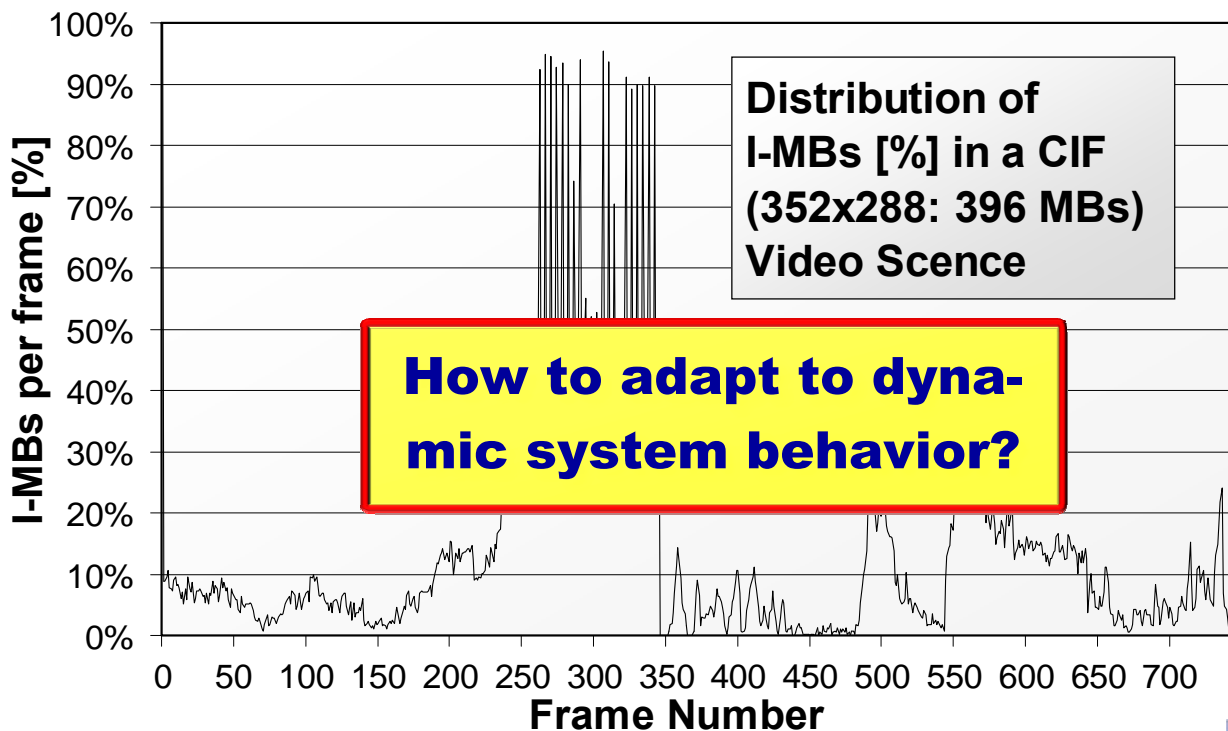
**Macroblock (MB):**  
16x16 pixel Block

**Common Intermediate Format (CIF):**  
Image with 352x288 pixel size. 1CIF = 396 MBs

❑ **I-MB:** Blocks with thick border

❑ **P-MB:** Blocks with thin border

# Input-Dependent Dynamic Application Behavior



## Dynamic System Behavior

- ❑ **Extensible Processor:** selecting points in design space at **design time**
- ❑ **Reconfigurable Computing:** typically fix at **compile time** when and how to deploy reconfigurable hardware

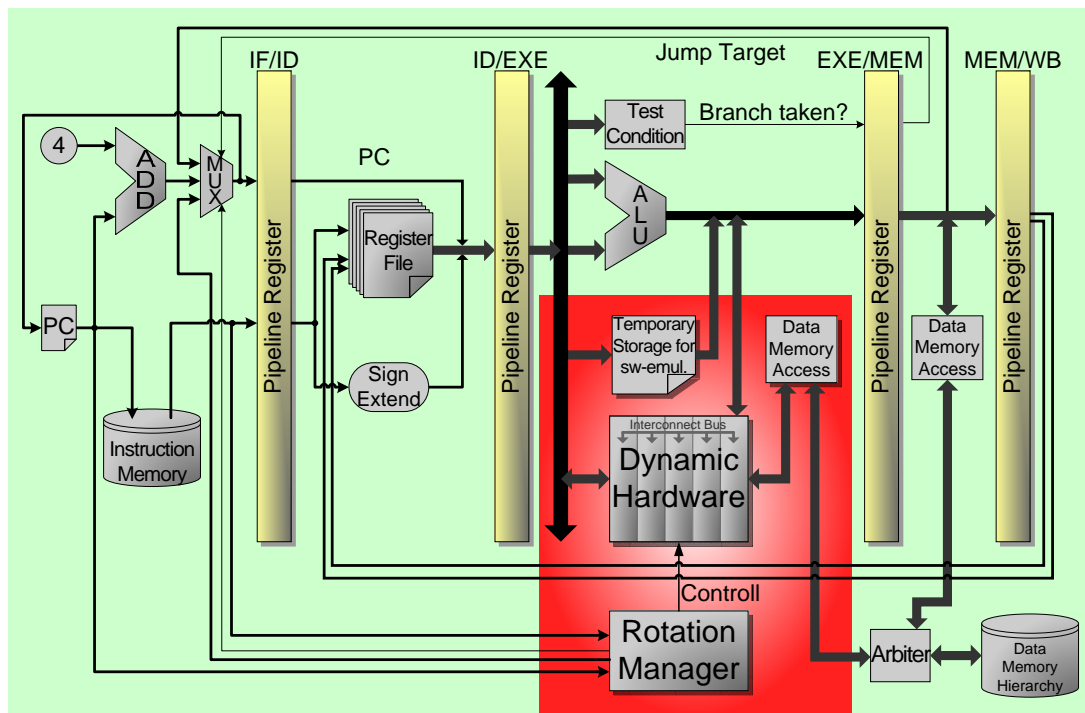
**How to handle situations that are unknown at design- & compile time?**

- ❑ Depending on **input data** (e.g. different computational paths in a video encoder)

# Our New Concept: RISPP

## Basic Idea and Overview

13



J. Henkel

Talk @ MPSoc'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>

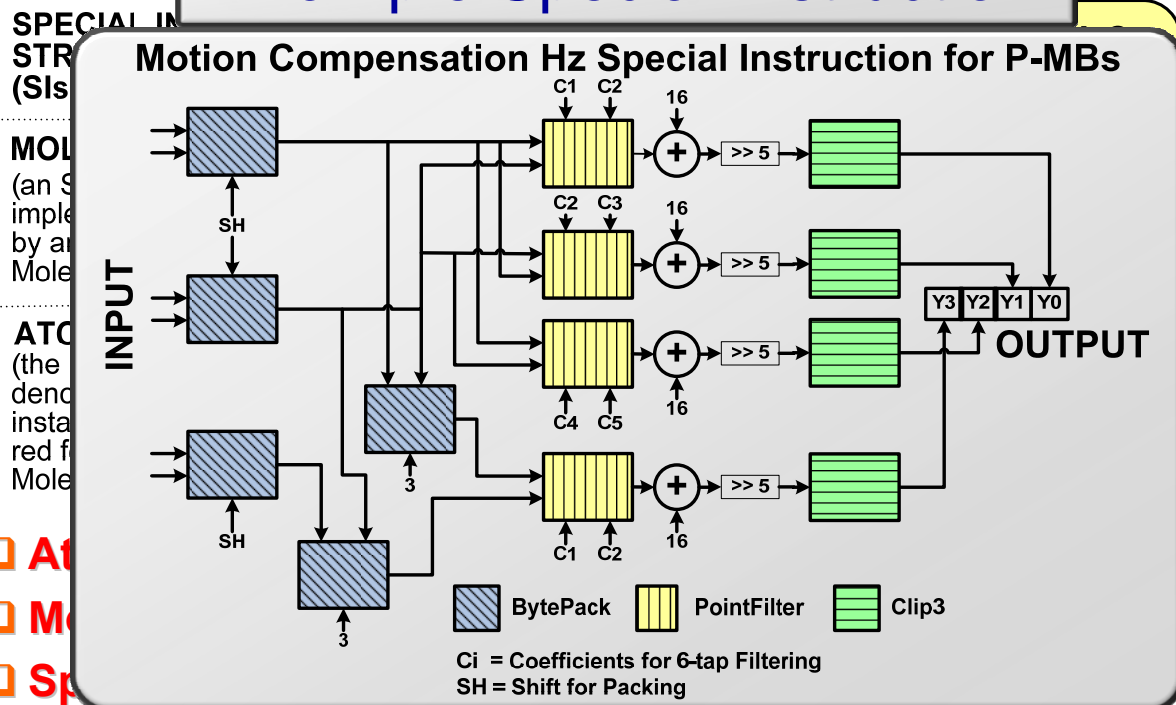


# Fundamental Processor Extension:

## Atom / Molecule Model

14

### Example Special Instruction



J. Henkel

Talk @ MPSoc'08, June 23rd

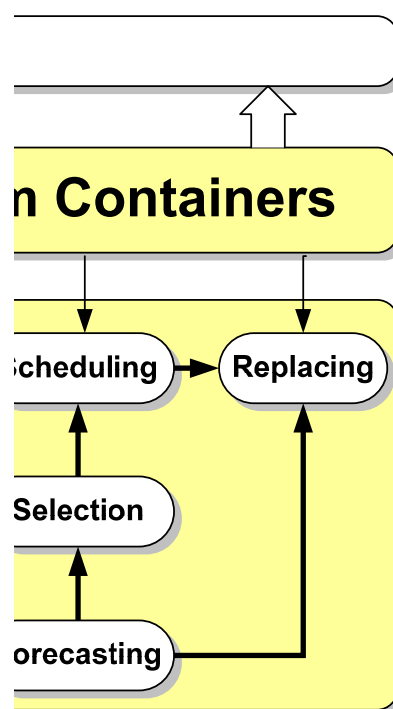
<http://ces.univ-karlsruhe.de/RISPP>



# Topics in RISPP

## Details

- ❑ Formal Atom/Molecule Model
- ❑ Automatically Inserting Forecast Points
- ❑ Overview of the Run-Time System (State-Transition)
- ❑ Fine-tuning the Forecasts
- ❑ Selecting Molecules
- ❑ Scheduling Atom loading sequence
- ❑ Efficiency Comparison
- ❑ Hardware Prototype



## Fix at Design/Compile Time — Adapt at Run Time

### Design Time

- ❑ Fix the available reconfigurable hardware resources
- ❑ Fix the algorithms for the run-time system

### Compile Time

- ❑ Determine Special Instruction composition
- ❑ Add forecasts to the application

### Run Time

- ❑ Adapt the forecasts
- ❑ Control SI execution
- ❑ Choose implementations (i.e. Molecules) for SIs

- ❑ Compile-time analysis: DAC'07
- ❑ Adaptation of the reconfiguration: SASO'07
- ❑ Atom/Molecule scheduling: DATE'08
- ❑ Molecule selection: DAC'08

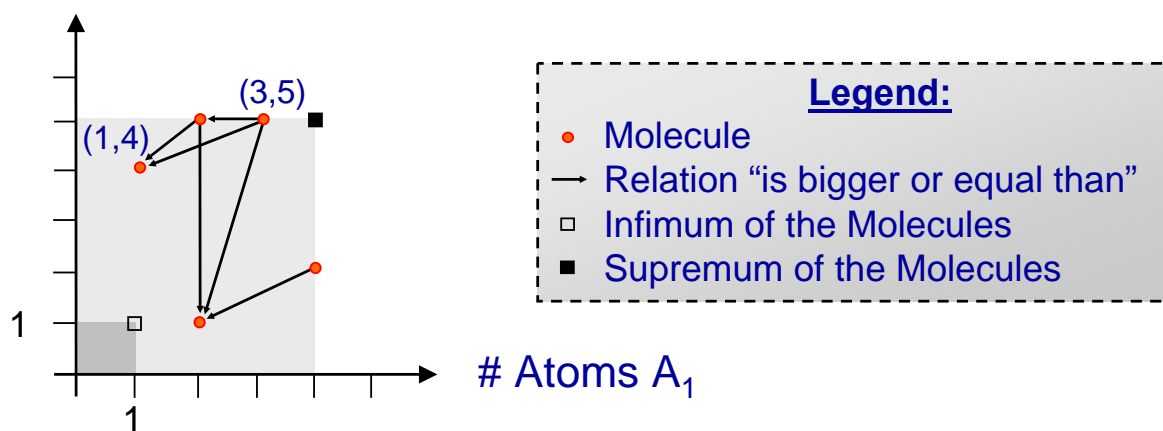


# Formal Atom/Molecule Model



## Formal Atom / Molecule Model: Example

# Atoms  $A_2$  (in general: n-dimensional)



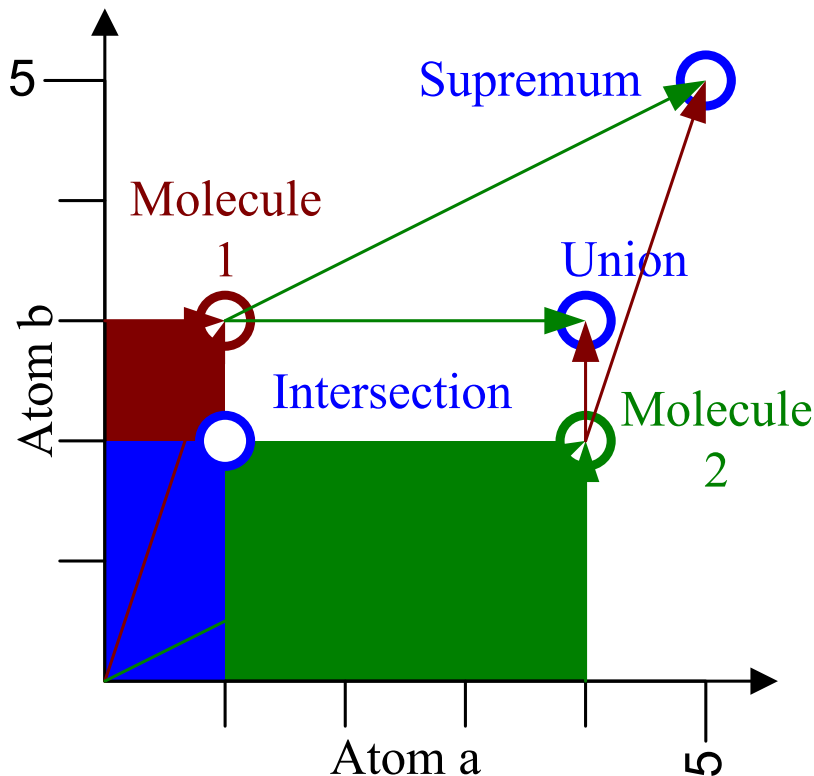
□ Molecule relations are e.g. needed when Molecules comprise each other

□ In such cases we can first configure the smallest possible Molecule with required functionality and then *upgrade* to faster implementations





# Example for base operators



## Formal Atom / Molecule Model: Details

$$(\mathbb{N}^n, \cup, \cap, \leq)$$

- **Main data structure:**  
**Set of all Molecules**

$$\begin{aligned} \cup: \mathbb{N}^n \times \mathbb{N}^n &\rightarrow \mathbb{N}^n \\ \vec{m} \cup \vec{o} &:= \vec{p} \\ p_i &:= \max\{m_i, o_i\} \end{aligned}$$

- **Meta-Molecule** to implement two Molecules, such that they can be **executed consecutively**, i.e. temporal domain (Abelian Group)

$$\begin{aligned} \cap: \mathbb{N}^n \times \mathbb{N}^n &\rightarrow \mathbb{N}^n \\ \vec{m} \cap \vec{o} &:= \vec{p} \\ p_i &:= \min\{m_i, o_i\} \end{aligned}$$

- **Meta-Molecule** for the **common Atoms** (indicator for *compatibility*)

$$\vec{m} \leq \vec{o} \quad \forall i \in [1, n]: m_i \leq o_i$$

- **Relation (Complete Lattice)**, with

$$\sup M := \bigcup_{\vec{m} \in M} \vec{m}$$

- **Supremum:** Meta-Molecule that is needed to implement all Molecules

$$\inf M := \bigcap_{\vec{m} \in M} \vec{m}$$

- **Infimum:** Meta-Molecule that is collectively needed for all Molecules



# Formal Atom / Molecule Model: Details

- **Determinant:** number of Atoms needed to implement

$$|\vec{m}| := \sum_{i \in [1, n]} m_i$$

- **Upgrading:** Atoms that are additionally needed to

$$\triangleright: \mathbb{N}^n \times \mathbb{N}^n \rightarrow \mathbb{N}^n; \quad \vec{m} \triangleright \vec{o} := \vec{p}, \quad p_i := \begin{cases} o_i - m_i, & \text{if } o_i - m_i \geq 0 \\ 0 & , \text{ else} \end{cases}$$

## Dynamic Special Instruction (SI) Forecasting

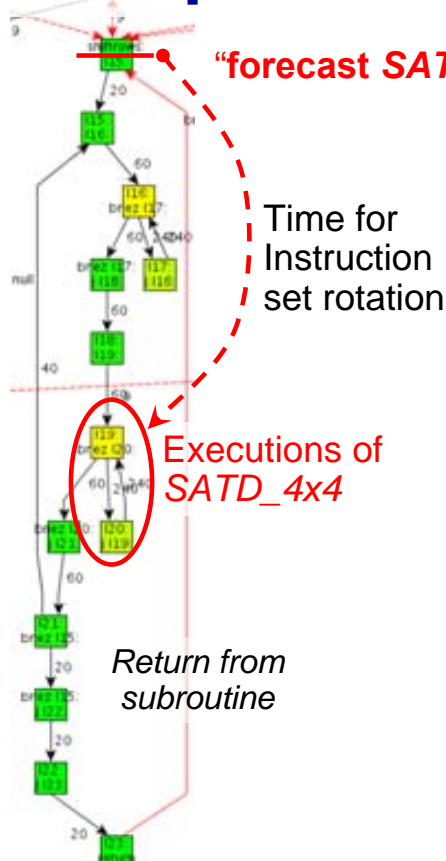
# Dynamic Special Instruction (SI) Forecasting

## Why?

- ❑ Rotation (i.e. re-configuring Special Instructions) takes long time (ms)
  - ❑ => start early to prepare rotation (=> conflicts)
  - ❑ => possibility of SW execution exists if run-time system did not manage to rotate
- ❑ Many hot spots
  - ❑ Would require too many Special Instructions
  - ❑ Only a subset is available at a certain time (context-dependent)
  - ❑ => cannot determine at design time which Special Instruction to execute at which time during execution



## Simplified Forecasting Example



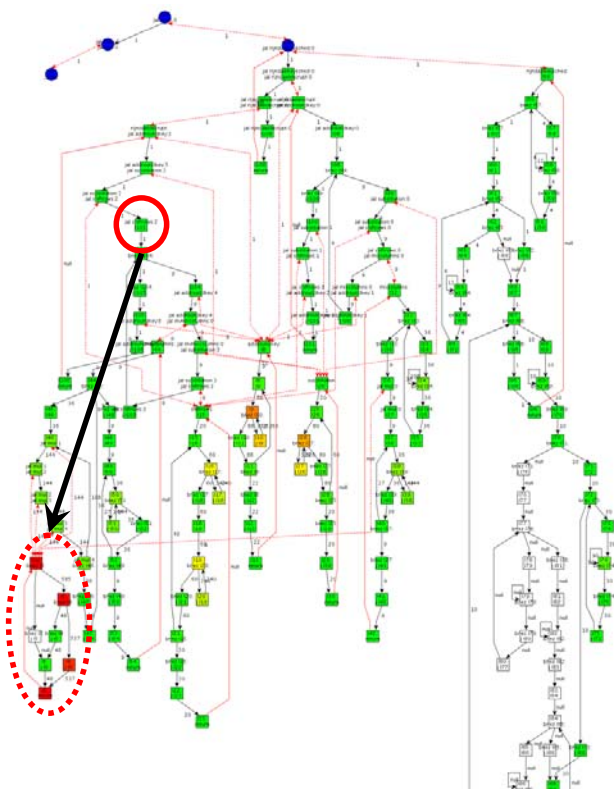
- ❑ Control-flow graph
  - ❑ Each node is a Base-Block (BB)
- ❑ At compile time:
  - ❑ Determine points to forecast a SI
  - ❑ Add Forecast Instructions with forecast values (about the SI importance) to these points
- ❑ At run time:
  - ❑ Use the Forecasts to determine the Instruction Set rotation
  - ❑ Dynamically update the importance of the forecasted SIs



# Inserting Forecast Points (FCs): General Idea of Algorithm

- I. **Pre-computations** from profiling data for each Special Instruction (SI)
- II. For every SI **determine Forecast Candidates**
- III. **Optimize list of FC-Candidates and select final forecasts**

## I. Pre-Computations



☐ **Pre-computations** are done on control-flow graph using profiling-information

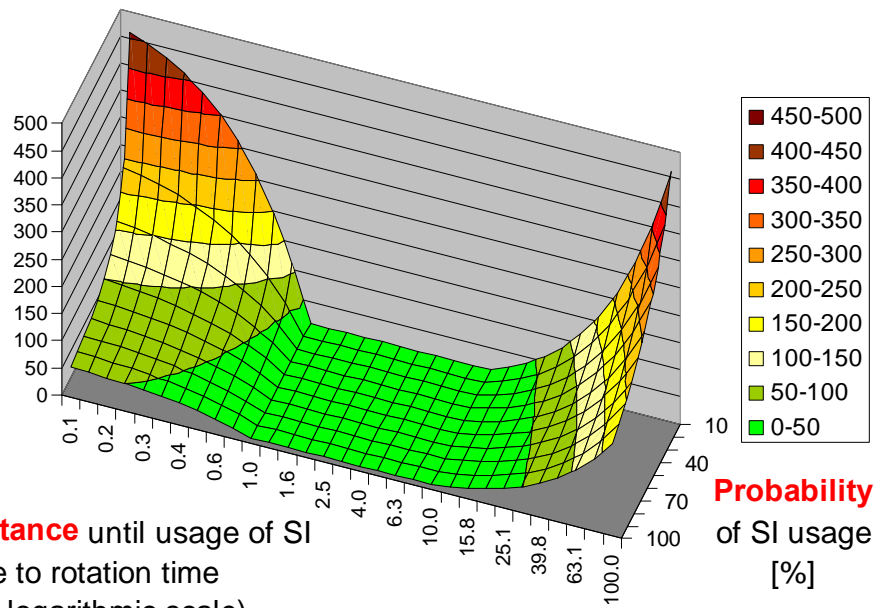


☐ **Temporal Distance** from Base Block to SI execution

☐ **Probability** that the SI executions are reached

☐ **Number of executions** of this SI (if it is executed)

## J. . . . .



<http://ces.univ-karlsruhe.de/RISPP>

J. Henkel

Talk @ MPSoC'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>

<http://ces.univ-karlsruhe.de/RISPP>

- ❑ **T: Time (Rot: for Rotation; SW: For SW Execution)**
- ❑ **p: Probability**
- ❑ **E: Energy**
- ❑ **alpha: Parameter for Energy vs. Speedup fine-tuning**



### III. Optimize list of FC Candidates

//  $S_1, \dots, S_k$  are the SIs of the FC Candidates in this BB

#### General Idea:

**While** the forecasted SIs in a Base Block consume *too many* area  
**Remove** the forecast with the worst

**Achieved Speedup**  
**Exclusively used Atoms**

```

15.     } else {
16.         break;
17.     }
18. }
```

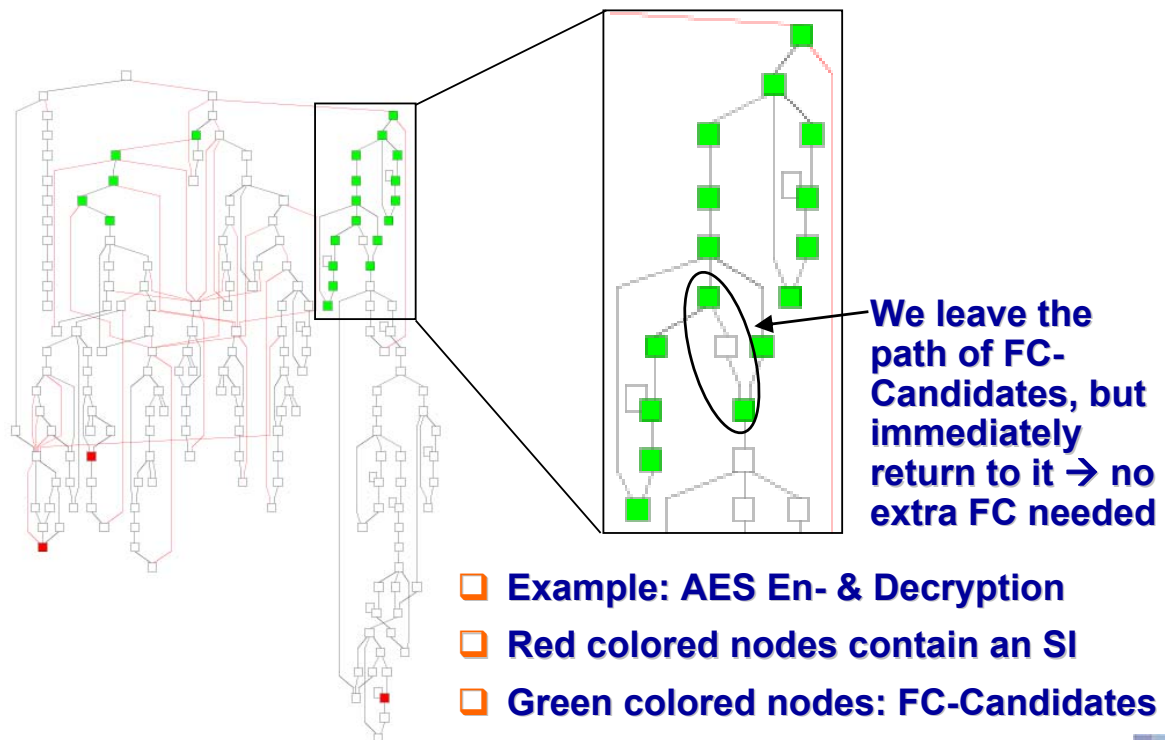


### Select final Forecasts

- ❑ Prerequisite: SI-Termination is already added (i.e. FCs, that a SI is no longer needed)
- ❑ Optimization goals
  - ❑ As few FCs as possible, as many as needed
  - ❑ Choose FCs with a good trade-off between 'sufficiently early' and a 'high execution probability'
- ❑ For each FC-Type T start a Depth-First-Search on the transposed Base Block graph (i.e. all edges reversed)
  - ❑ No recursive relegation, if a node is not a FC-Candidate for the current FC-Type T
  - ❑ The current Node N becomes a Forecast IFF
    - ❑ A successor S of Node N is not a FC-Candidate for Type T  
AND
    - ❑ The Path beginning with S is not soon (in terms of ms, not BBs) reaching another FC-Candidate  
AND
    - ❑ We don't have added a FC a few nodes previously

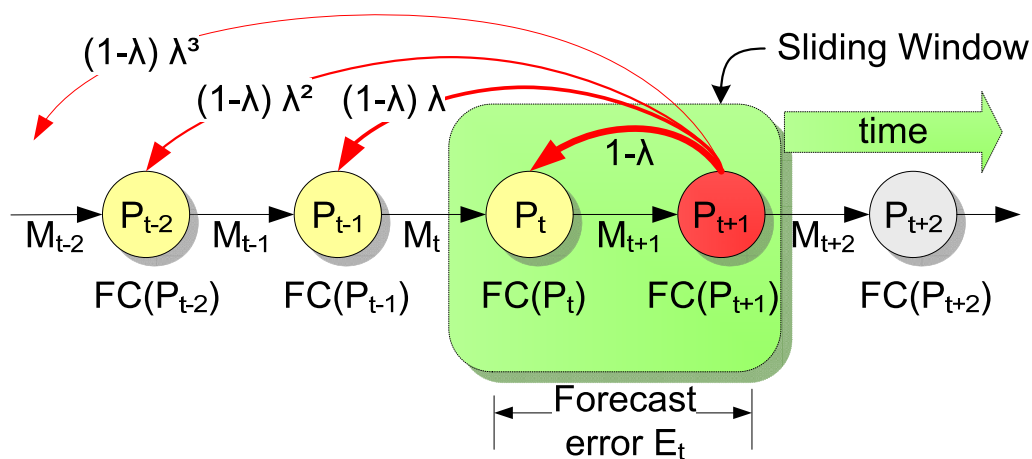


# Example: Choosing FCs



## Forecast fine-tuning

# Adapting the Special Instruction Forecasts



- ❑ **Computing the Error:**  

$$E_t = M_{t+1} + \gamma FC(P_{t+1}) - FC(P_t)$$
- ❑ **Back-Propagating the Error:**  

$$FC(P_t) = FC(P_t) + \alpha E_t$$

## Legend

**P:** Forecast Point  
**FC:** Forecast Value in the Point  
**M:** Monitored number of SI executions

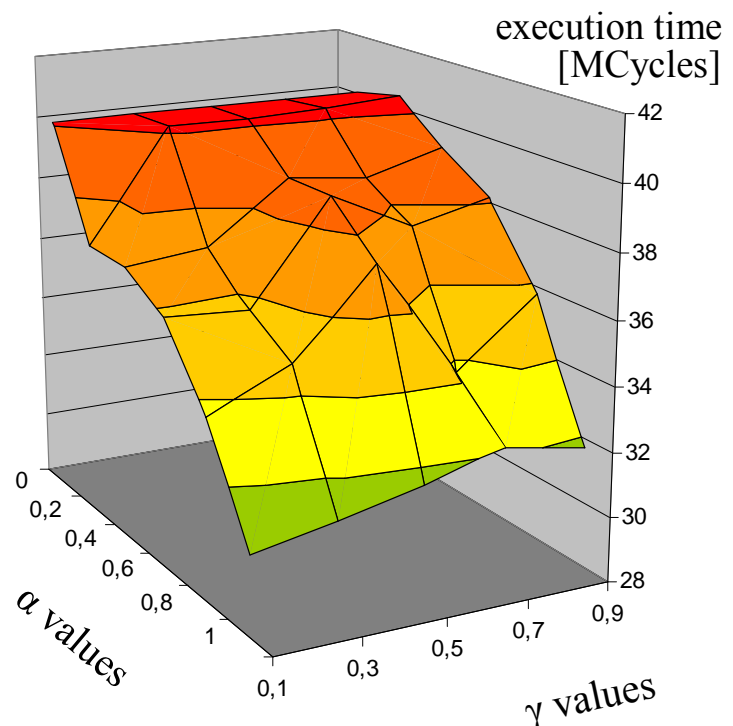
# Parameter Evaluation

- ❑  $\lambda > 0$  rapidly increases overhead
  - ❑ Max speedup vs.  $\lambda=0$ : **1.10x**
  - ❑  $\lambda=0$  is a good performance vs. overhead compromise

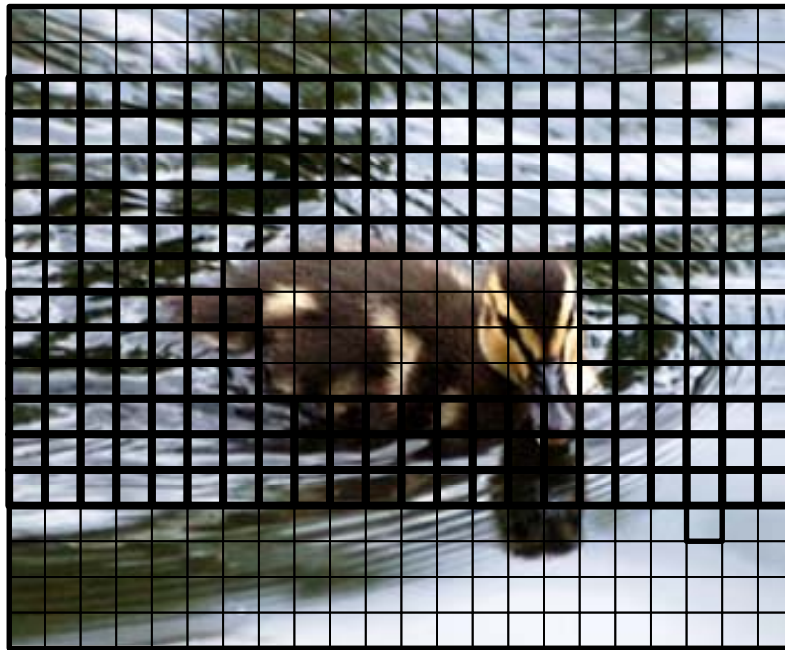
	$\lambda=0.6$	$\lambda=1.0$
Min.* speedup	1.01x	1.01x
Avg.* speedup	1.03x	1.03x
Max.* speedup	1.08x	1.10x

\*comparing with  $\lambda=0$

## Evaluation for $\lambda=0$



# Forecasting I- / P-MBs in practice



**Macroblock (MB):**  
16x16 pixel Block

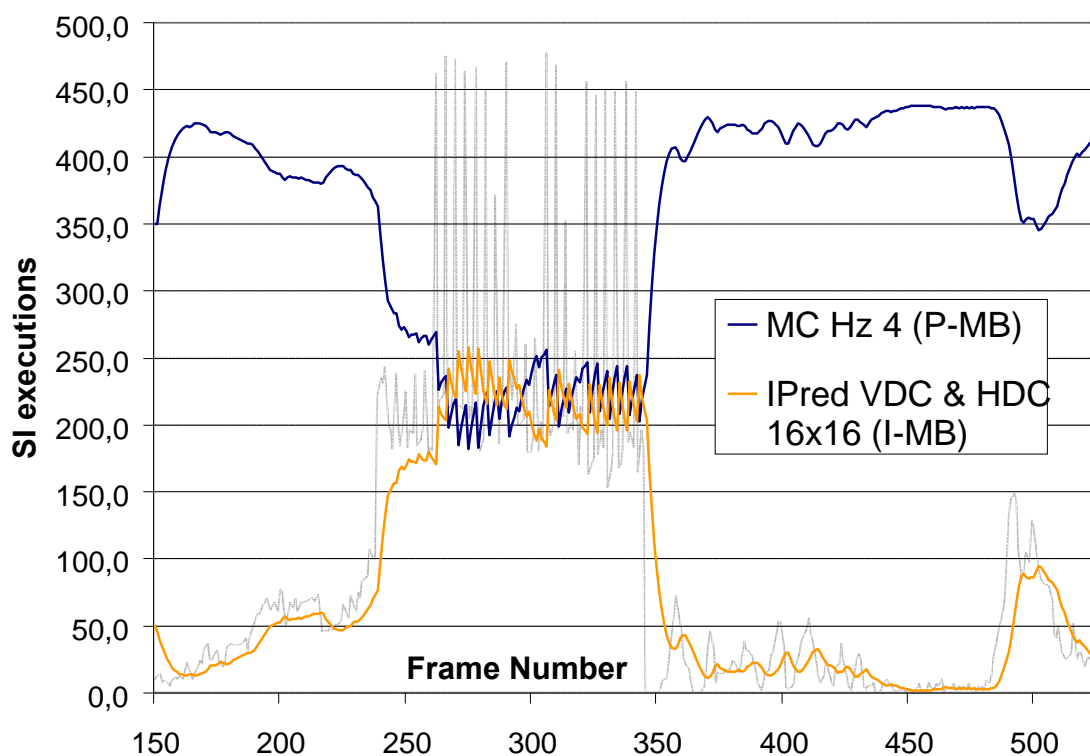
**Common Intermediate Format (CIF):**  
Image with 352x288 pixel size. 1CIF = 396 MBs

☒ **I-MB:** Blocks with thick border

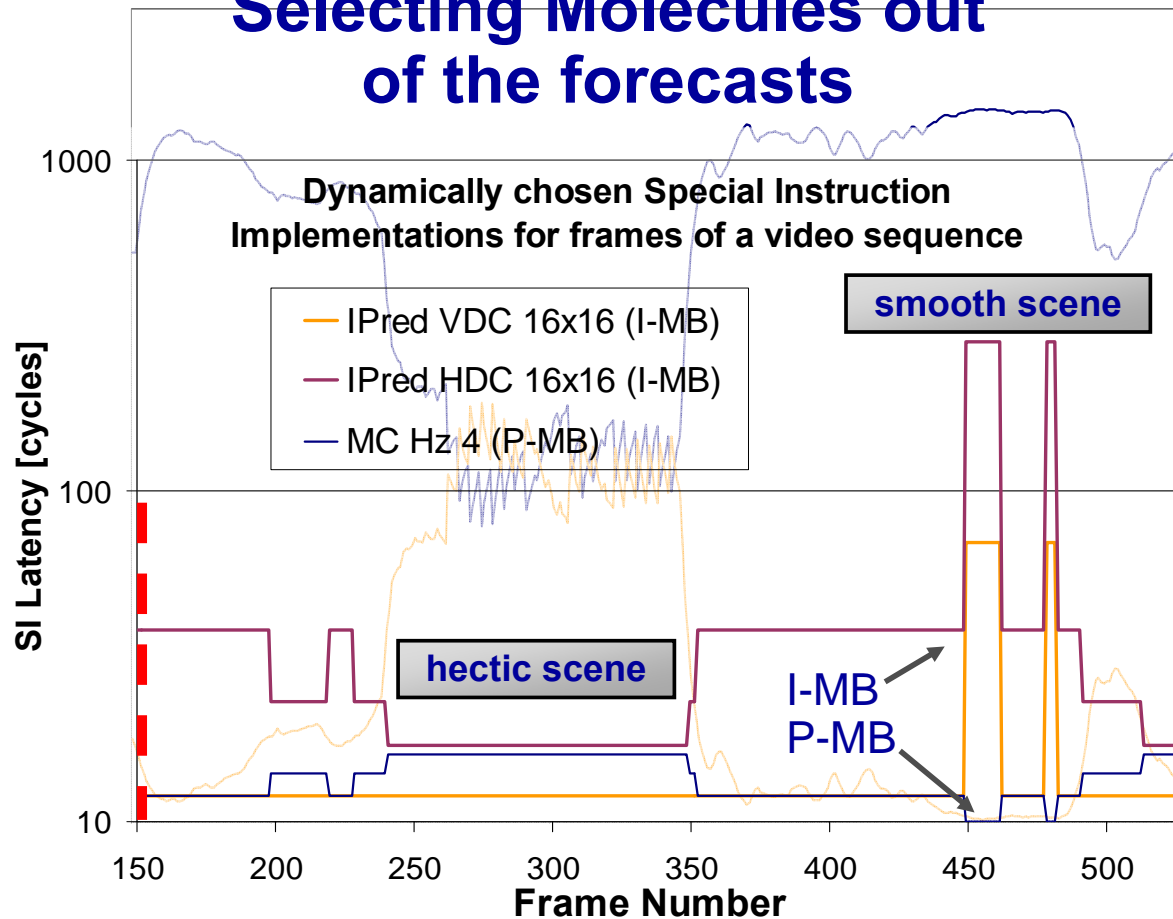
☐ **P-MB:** Blocks with thin border

src: <http://www.copyright-free-pictures.org.uk>

## Dynamically Adapting the Forecasts to the actual application requirements <sup>36</sup>



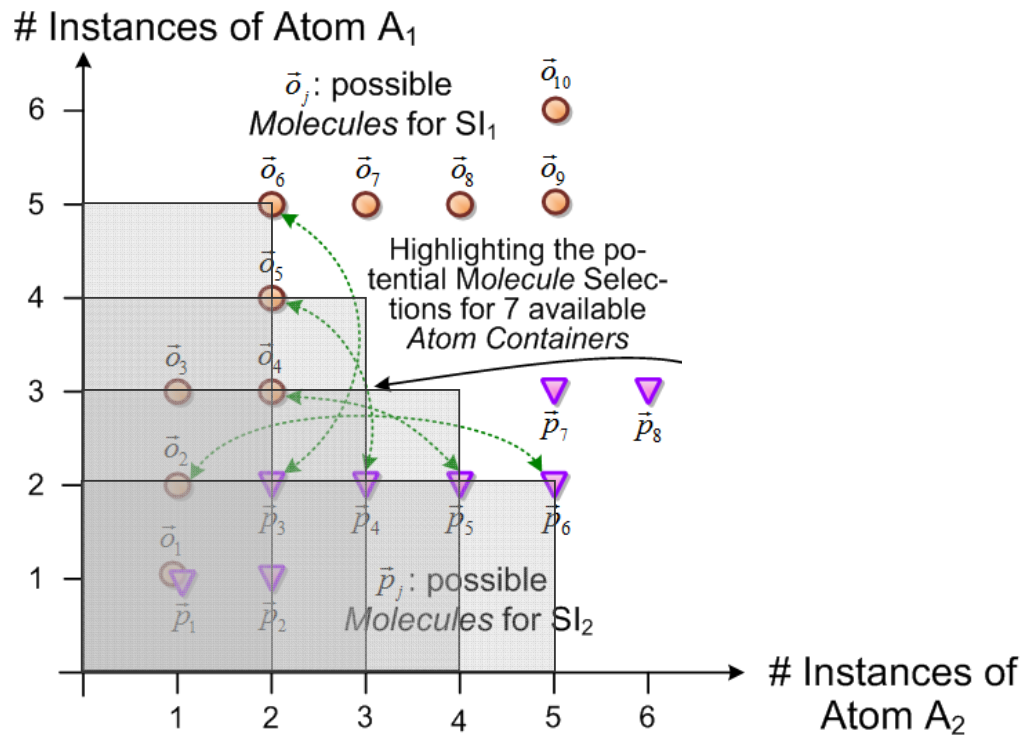
# Selecting Molecules out of the forecasts



## Molecule Selection



# Molecule Selection



## Formal Selecting Molecules for SIs

### Formalized as a Knapsack Problem

→ NP-Hard

### Input to the selection:

$$F = \{(M_i, f_i, t_i)\} \quad i \text{ index of a certain SI}$$

### Chose exactly one Molecule to implement a SI:

$$\forall i : |S \cap M_i| = 1$$

### Stay within the capacity of the Knapsack:

$$\left| \bigcup_{\vec{o} \in S} \vec{o} \right| \leq N$$

# Selecting Molecules for SIs (cont'd)

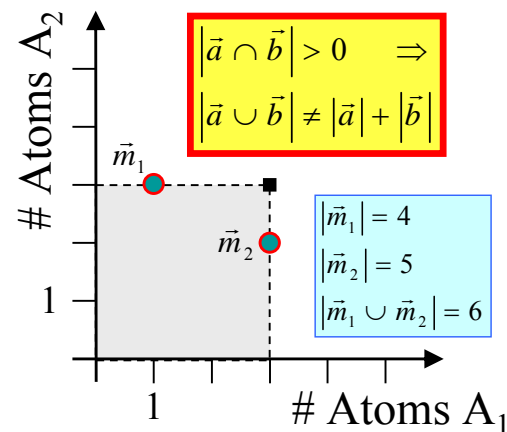
- Unlike traditional Knapsack: The weight of a Molecule is not constant!

- Two Molecules might share some Atoms:

- Special Version: Set-Union Knapsack Problem

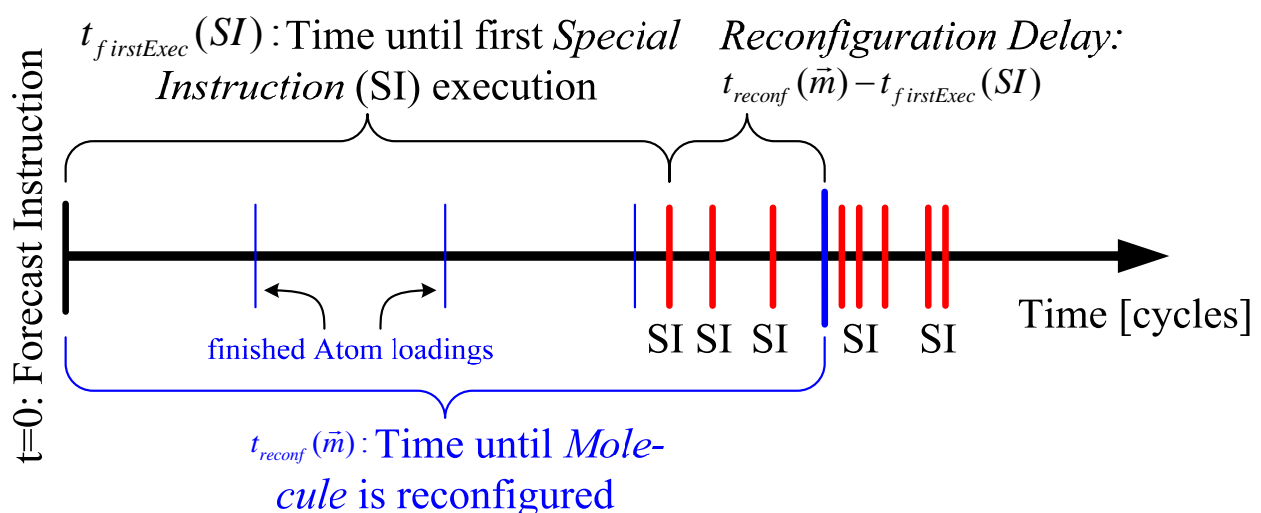
- Profit function (optimization goals):

$$\text{maximize } \sum_{\forall i: \vec{o} \in M_i \cap S} \text{profit}(\vec{o}, f_i, t_i)$$



## Relevant Selection Parameters

- Depending on the selection, the reconfiguration may finish far too late

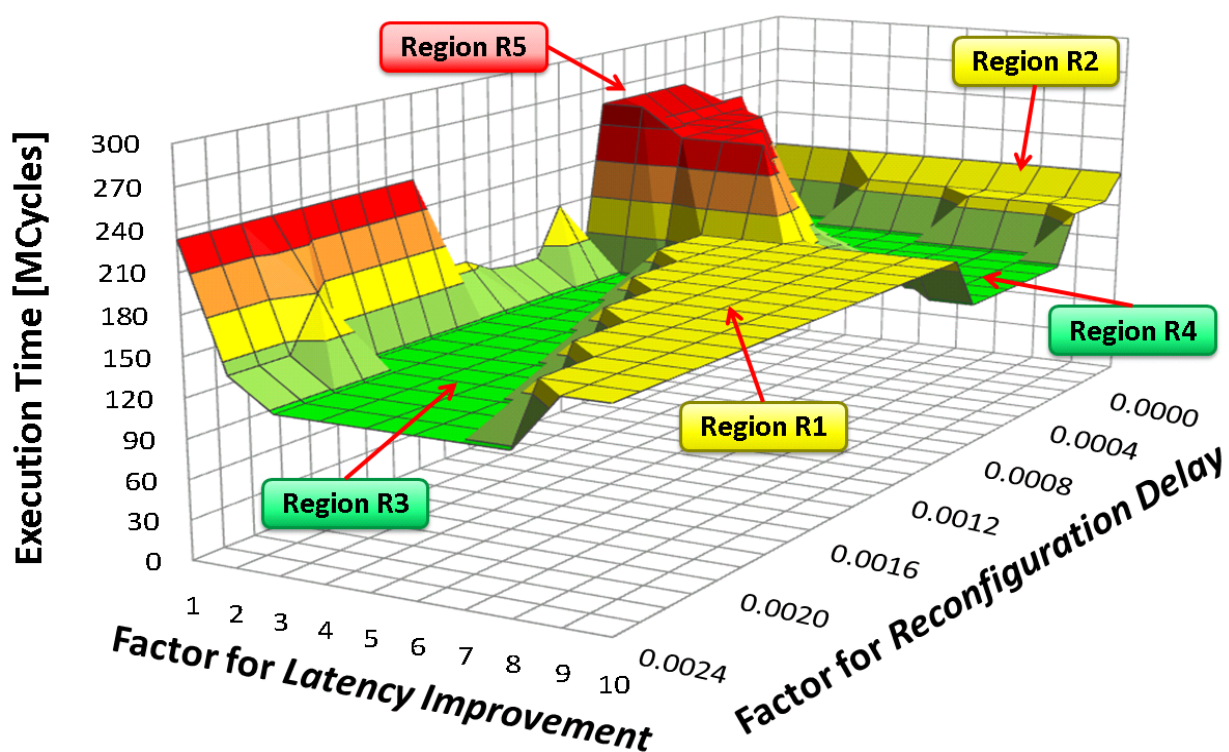


# Profit function for a Molecule

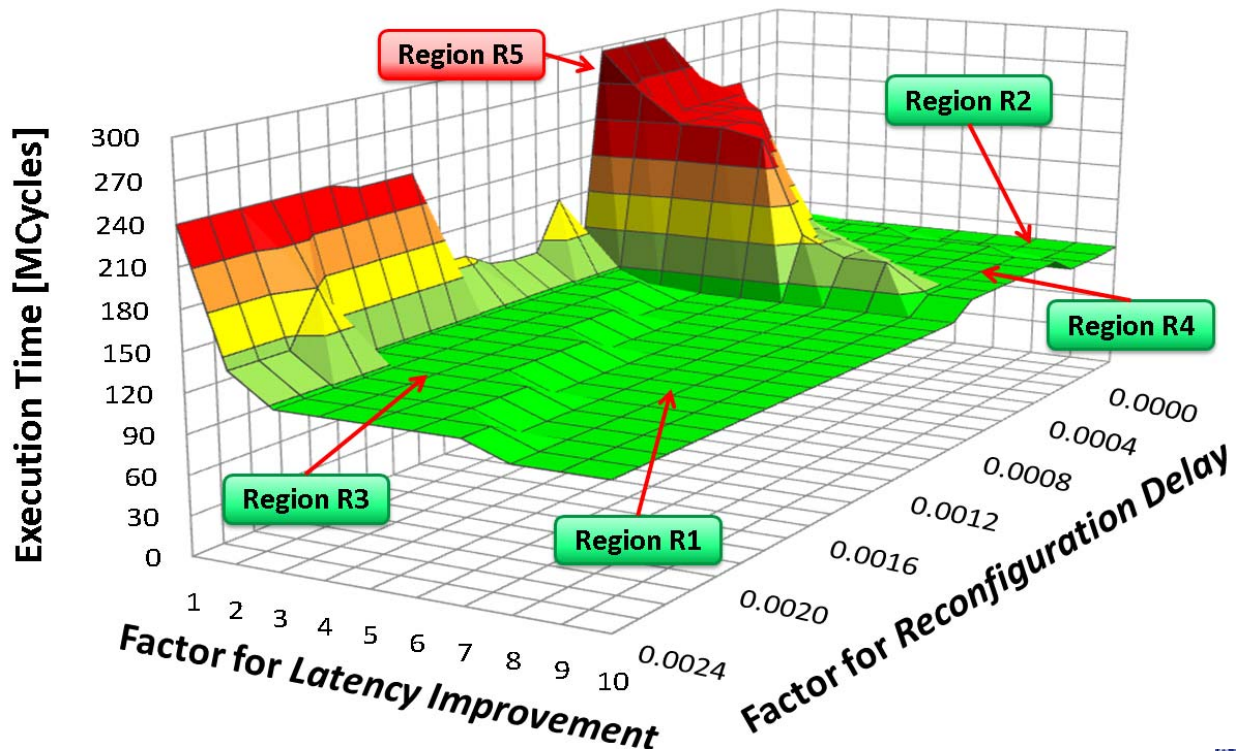
- ❑ Selection parameters L & R are used to scale the parameters
  - ❑ Latency Improvement
  - ❑ Reconfiguration Delay
- ❑ Additional parameter:
  - ❑ Expected SI execution frequency

$$profit(\vec{m}, f_i, t_i) := f_i \cdot \left( L \cdot \begin{pmatrix} latency(\vec{m}_{i\_sw}) \\ -latency(\vec{m}_{ij}) \end{pmatrix} - R \cdot \max \left( 0, \begin{pmatrix} t_{reconf}(\vec{m}) \\ -t_i(SI) \end{pmatrix} \right) \right)$$

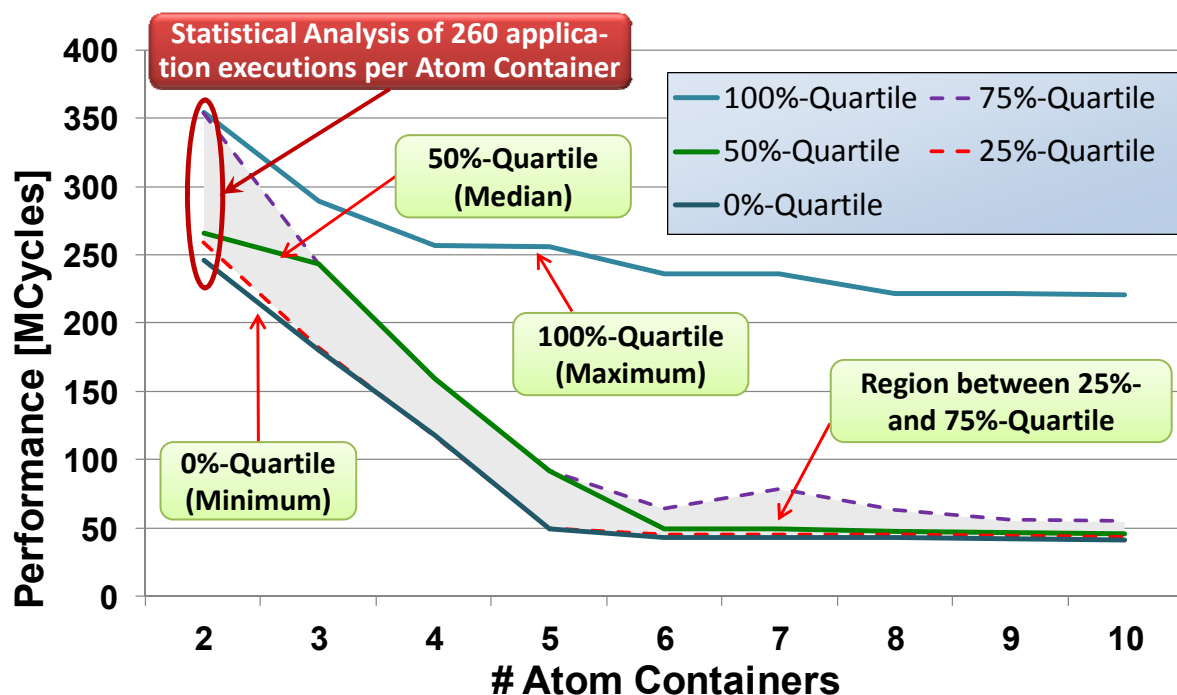
## Evaluation of the profit function (greedy)



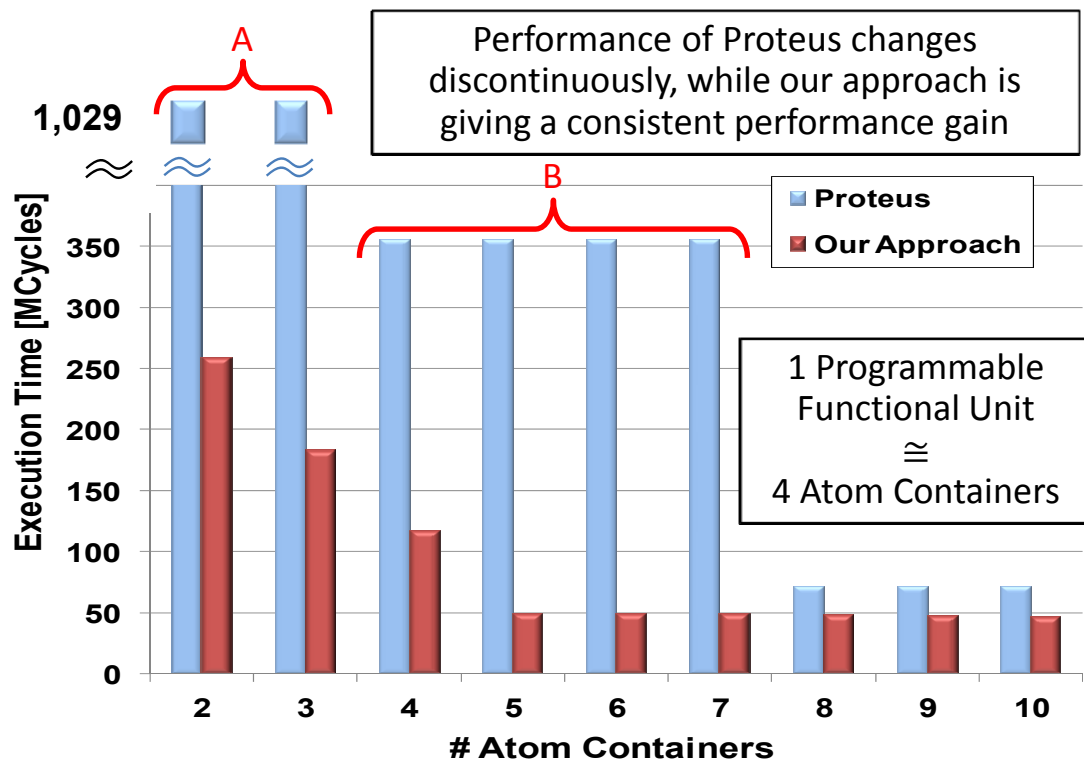
# Evaluation of the profit function (optimal)



# Statistical Analysis of the profit function parameters for the greedy implementation



# Comparing Reconfiguration on SI-level (Proteus) with Atom-level (RISPP)

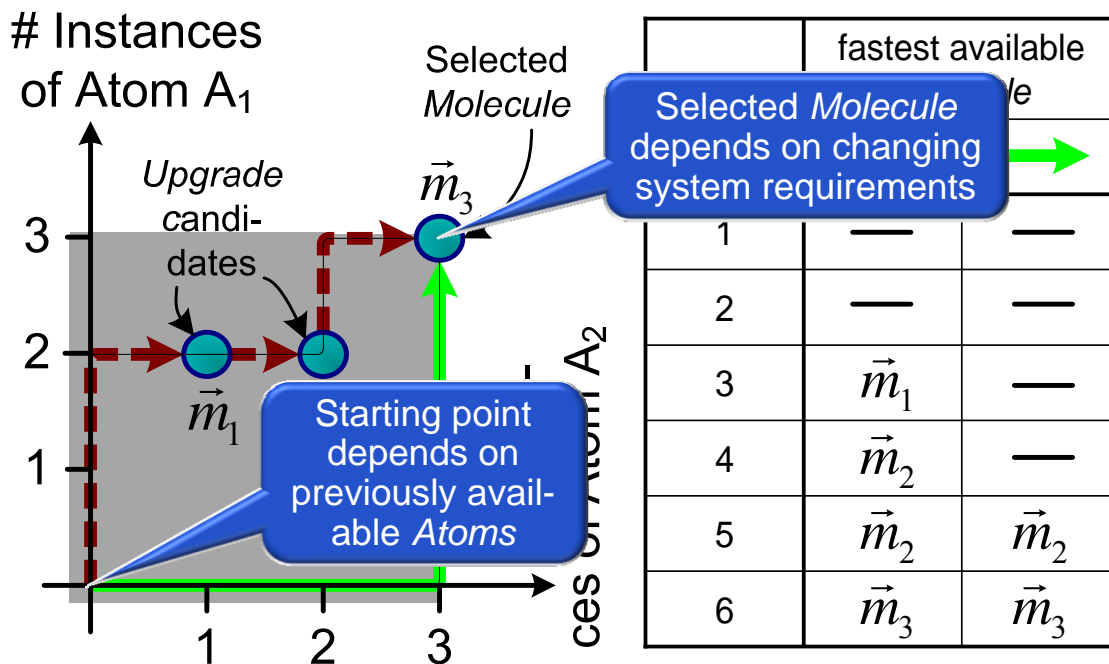


## Atom/Molecule Scheduling



# Determining Atom loading sequence

- ❑ **Problem: Reconfiguration is slow**
- ❑ **Constraint: At most one reconfiguration at a time**



J. Henkel

Talk @ MPSoC'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>

## Overview: Implemented SIs

	Special Instruction	# utilized Atom-types	# different possible Molecules
Motion Estimation (ME)	SAD	1	3
	SATD	4	20
Encoding Engine (EE)	(I)DCT	3	12
	(I)HT_2x2	1	2
	(I)HT_4x4	2	7
	MC_4	3	11
	IPred_HDC	2	4
	IPred_VDC	1	3
	LF_BS4	2	5
Filter (LF)	LF_BS4	2	5

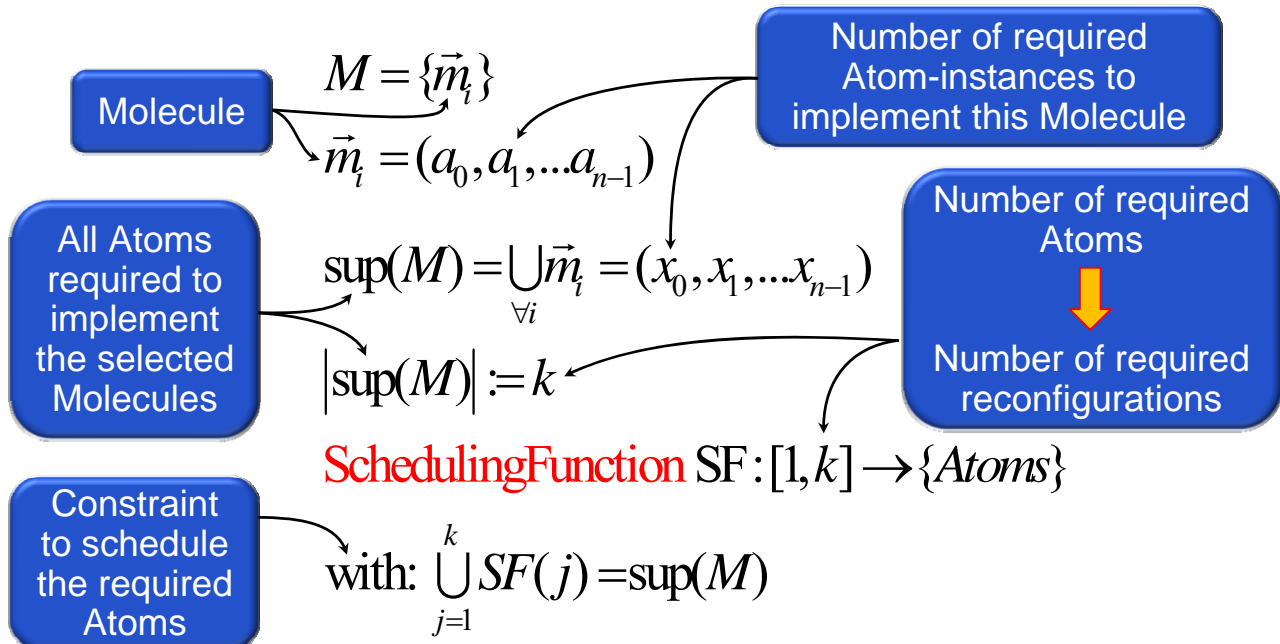
J. Henkel

Talk @ MPSoC'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>

# Defining the Scheduling Problem

## Input: Set of selected Molecules $M$



# Schedule Molecules, not Atoms

## Idea: The SI performance changes, **when a new Molecule becomes available**

- To reduce the complexity and to become target oriented: determine the Molecule loading sequence

$$M = \{\vec{m}_i\}$$

### Consider upgrade candidates (once):

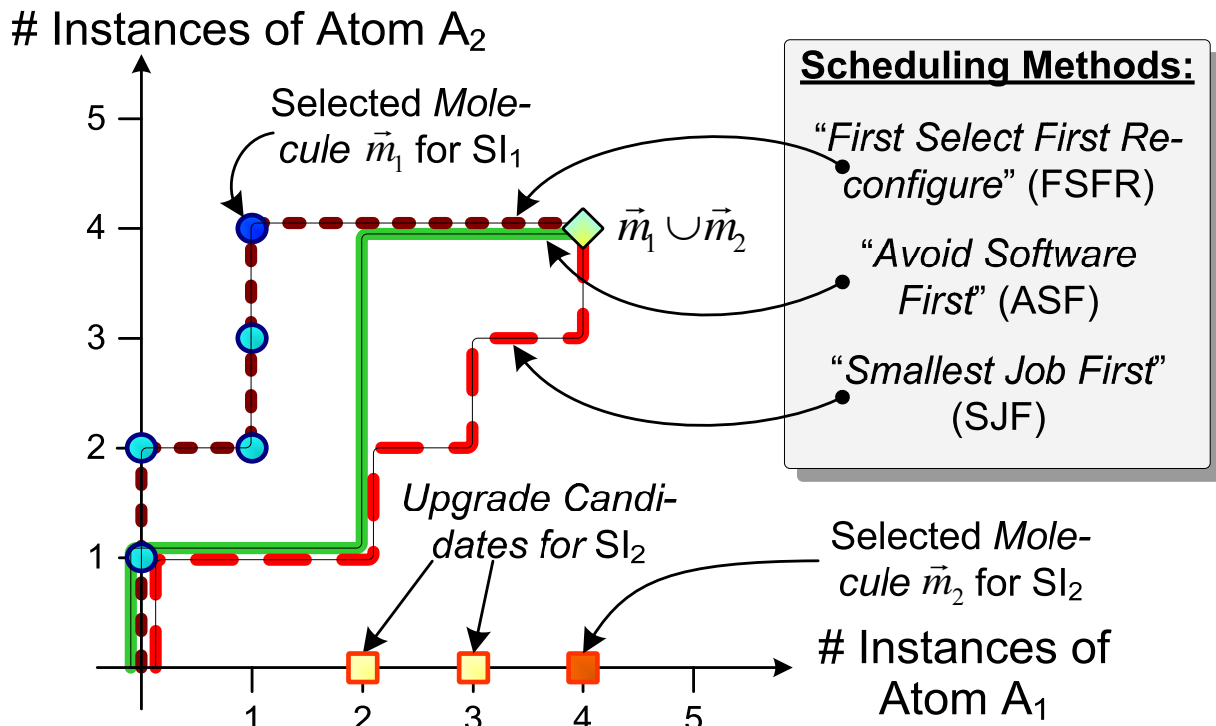
$$M' = \bigcup_{\vec{m} \in M} \{ \vec{o} : \vec{o} \leq \vec{m} \wedge \vec{o}.getSI() = \vec{m}.getSI() \}$$

### Trim candidates (iteratively):

$$M'' = \left\{ \vec{m} \in M' : \left( |\vec{a} \triangleright \vec{m}| > 0 \wedge \vec{m}.getLatency() < \vec{m}.getSI().getFastestAvailableMolecule(\vec{a}).getLatency() \right) \right\}$$

# Comparing different Scheduling Methods for 2 Selected SIs

53



J. Henkel

Talk @ MPSoc'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>



## Our Scheduling Approach: Highest Efficiency First (HEF)

54

```

1. //
2. bestBenefit ← 0
3.
4. ∀  $\vec{o} \in M'$  {
5.
6. //
7.    $\vec{o}.getSI().getExpectedExecutions() * (\vec{o}.getSI().bestLatency - \vec{o}.getLatency()) / |\vec{a} \triangleright \vec{o}|$ 
8.   benefit ←
9.
10.  if (benefit > bestBenefit) {
11.    bestBenefit ← benefit
12.     $\vec{m} \leftarrow \vec{o}$ 
13.  }
14.
15. }
16.
17. }
18.
19. }
20.
21. }
22. bestBenefit ← benefit;  $\vec{m} \leftarrow \vec{o}$ ;
23. }
24. }
25. // schedule the chosen Molecule
26. ∀ Atoms  $a_i \in (\vec{a} \triangleright \vec{m})$  scheduledList.push( $a_i$ );
27.  $\vec{a} \leftarrow \vec{a} \cup \vec{m}$ ;
28.  $\vec{m}.getSI().bestLatency \leftarrow \vec{m}.getLatency()$ ;
29. }
30. return scheduledList;

```

SI execution frequency

Latency improvements (per SI execution)

Additionally required Atoms

3. Extract the additionally needed Atoms

4. Trim the candidates

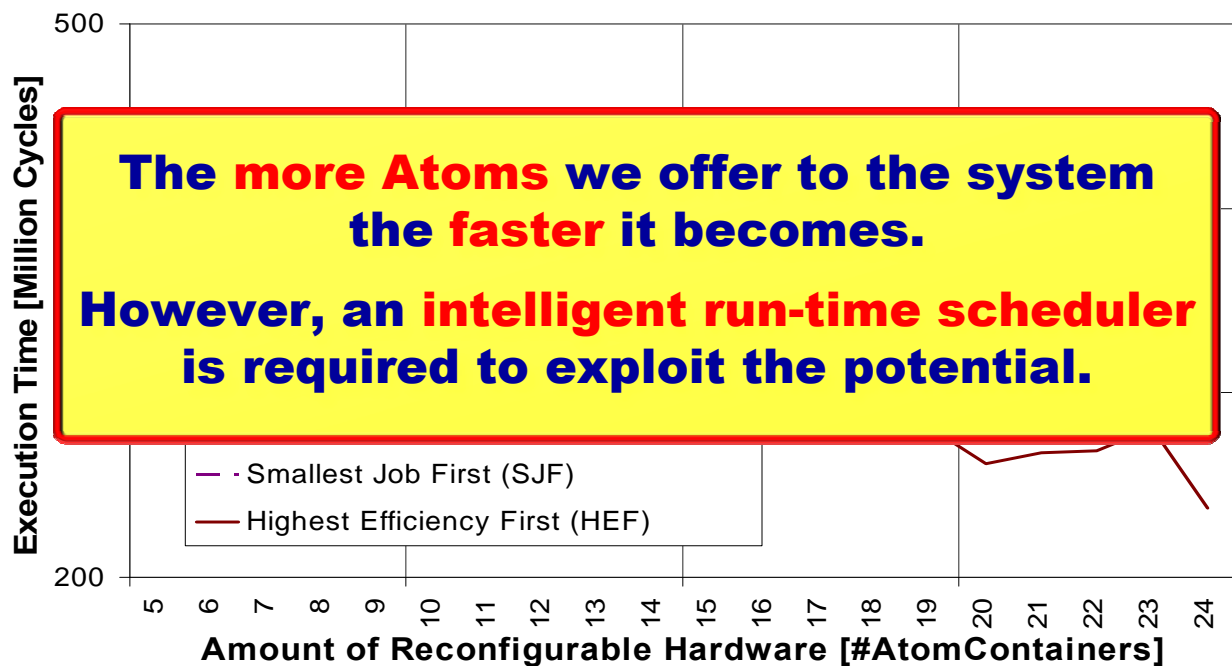
J. Henkel

Talk @ MPSoc'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>



# Comparing our proposed scheduling scheme



Encoding 140 frames (352x288 resolution) with H.264



## Speedup due to our scheduler and due to our architecture

- Comparing state-of-the-art approaches (Molen) with our architecture (ASF: simple but nontrivial Scheduler)
- Evaluating, what our proposed Scheduler (HEF) achieves on top

Speedup	#ACs	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
	HEF vs. ASF	1.00	1.04	1.04	1.06	1.05	1.08	1.06	1.06	1.13	1.18	1.21	1.26	1.36	1.48	1.45	1.52	1.51	1.39	1.26	1.52

2.38x

Speedup

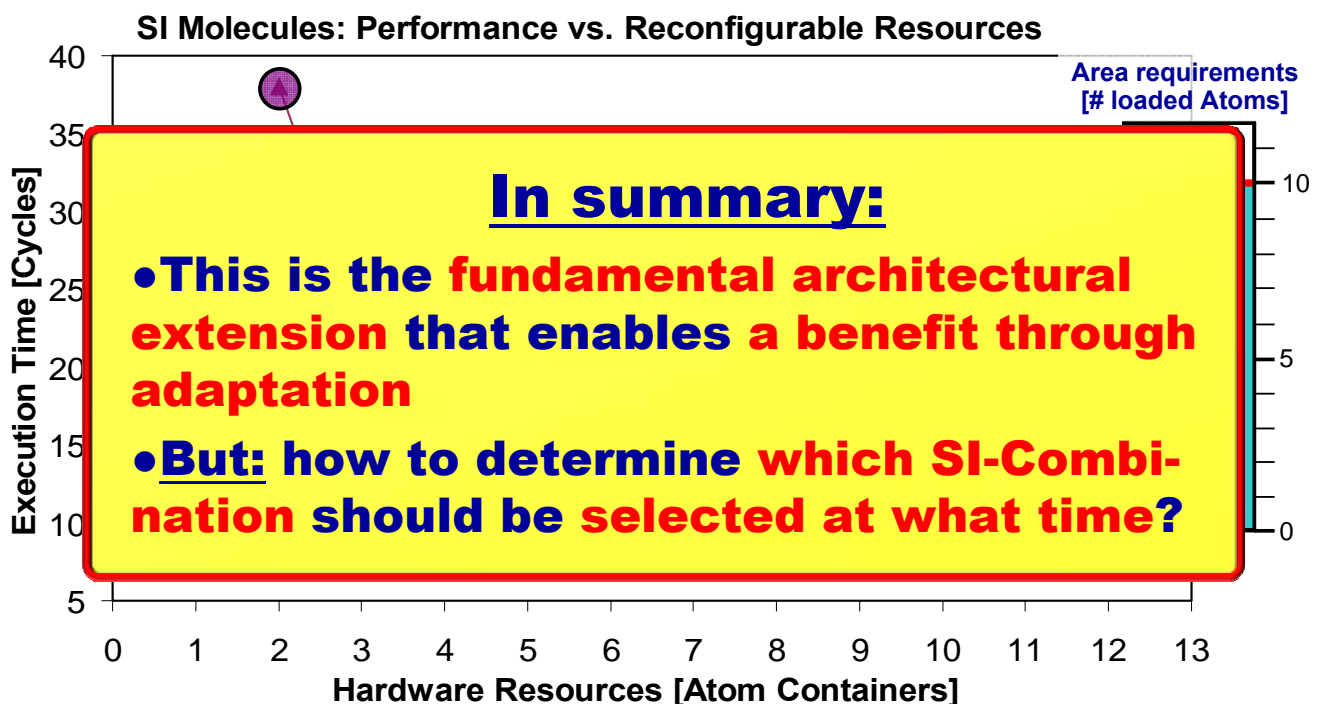
1x



# Efficiency / Utilization

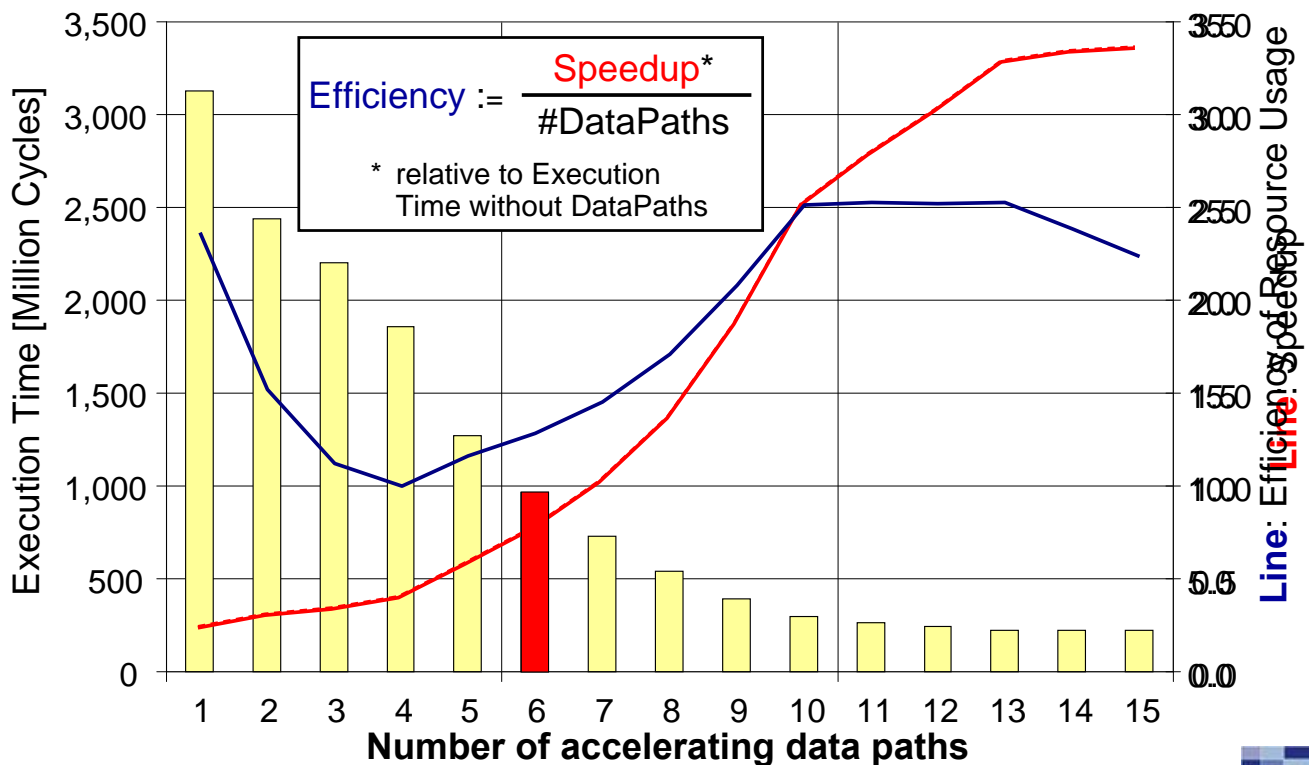


## Adaptivity Through Dynamic Performance vs. Area Trade-off



# Analyzing ASIPs for varying amount of accelerating data paths

59



J. Henkel

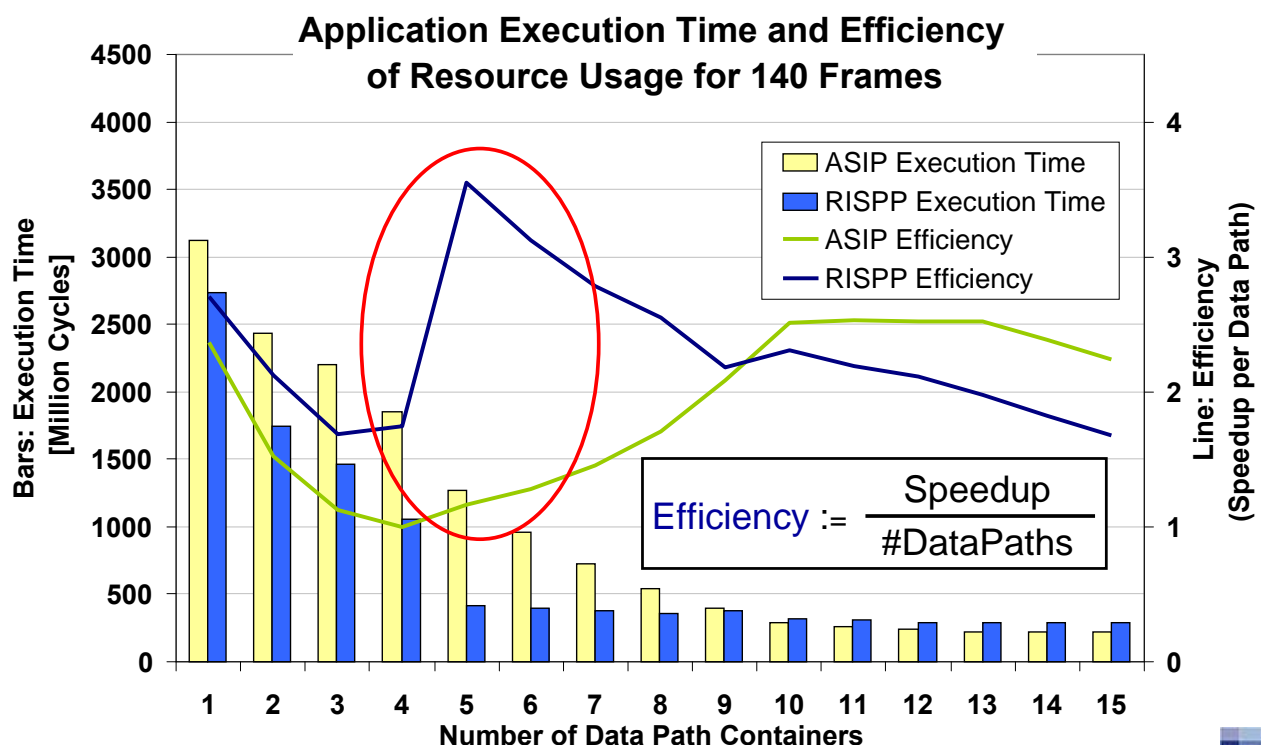
Talk @ MPSoc'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>



## Results: execution time and efficiency of resource usage for ASIP and RISPP

60



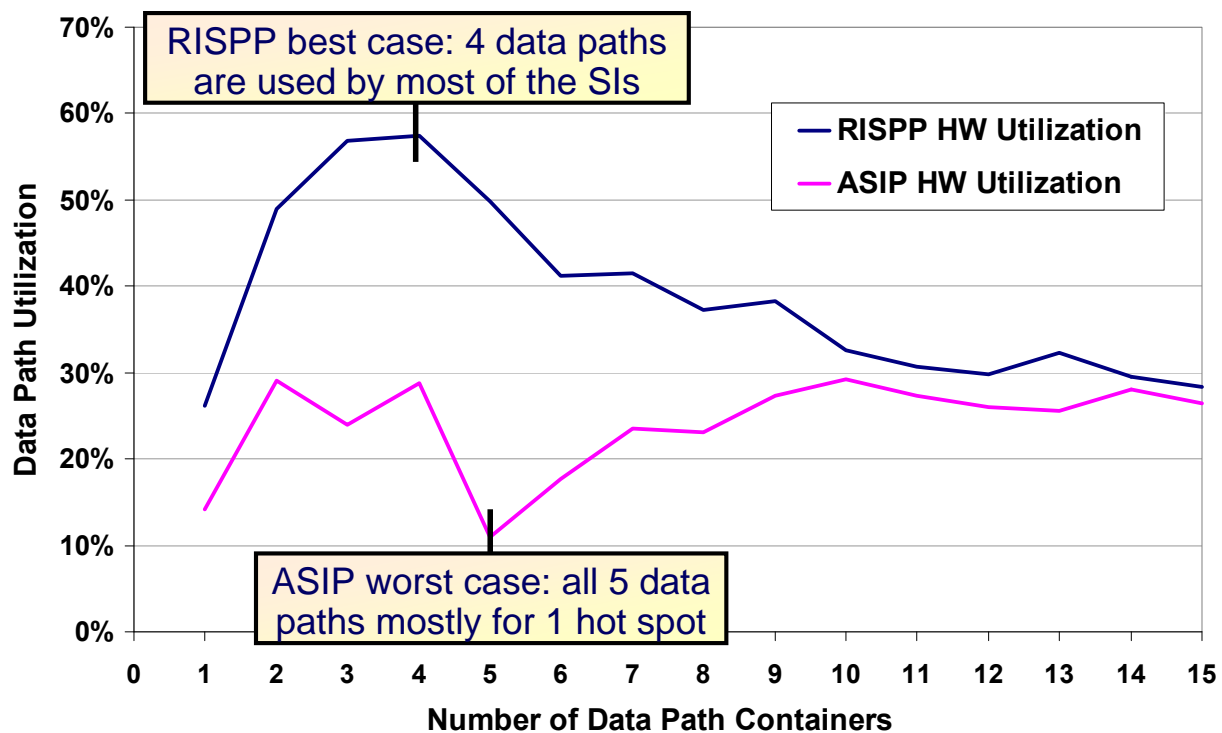
J. Henkel

Talk @ MPSoc'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>



# Data path utilization during hot spot execution



## Summary of comparison of ASIP and RISPP

	ASIP			RISPP		
	Min	Avg	Max	Min	Avg	Max
<b>Execution Time [MCycles]</b>	220.6	999.6	3126	288.3	715.1	2734
<b>Speedup vs. GPP</b>	2.4	16.8	33.6	2.7	17.6	25.7
<b>Efficiency</b>	1.0	1.9	2.5	1.7	2.3	3.6

- ❑ **Measurements for different numbers of data paths**
  - ❑ Avg. & max. consider many data paths/containers
- ❑ **Only for huge number of data paths ASIP is better**





# Hardware Prototype



## Hardware Prototyping Board

**Using a 10-bit fixed-point arithmetic and a state machine (3 states for the profit computation, 8 altogether)**

Serial connection to e.g. initialize the memories

Touch Screen LCD to allow interaction with the application

Characteristics	Selection	Avg. Atom
# Slices	556	421
# LUTs	1012	839
# FFs	231	45
#MULT18X18	11	0
Clock delay [ns]	14.229	5.465

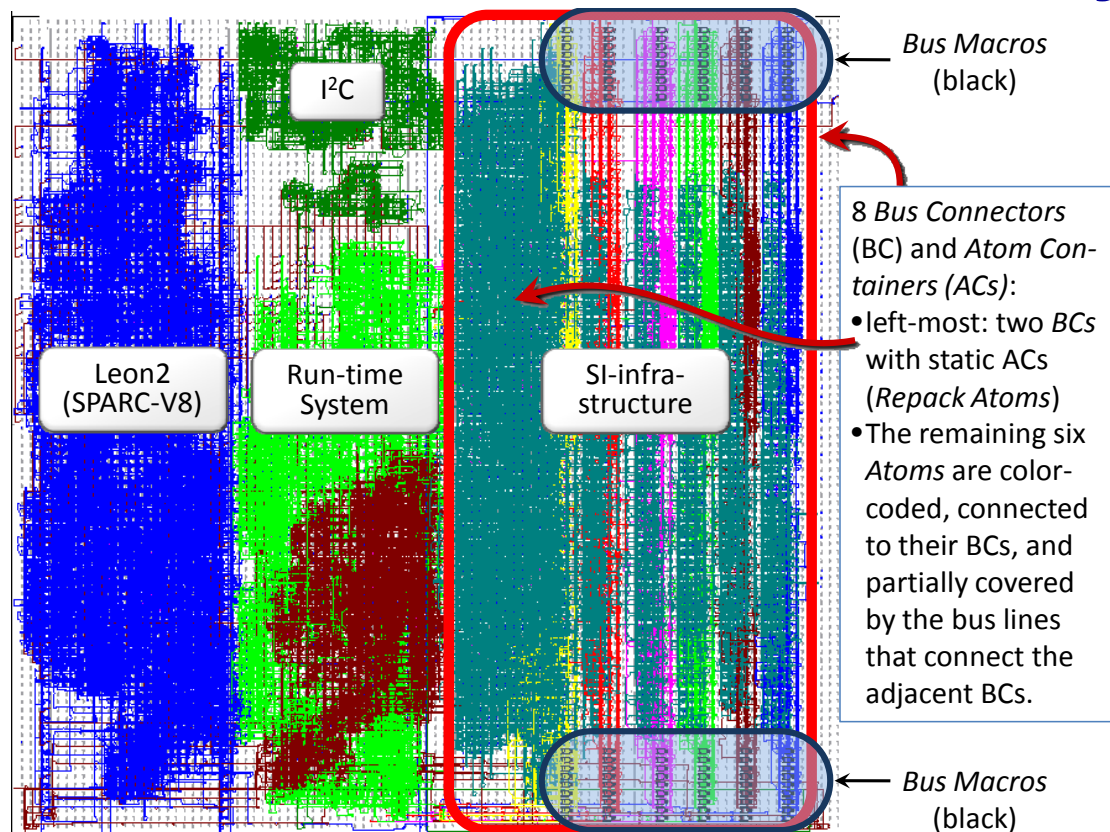
Flash memory for reconfiguration data

SRAM for Instruction- and Data-memory

Socket for an Virtex-II FPGA with FF1152 package



# Latest Version of Hardware Prototype



J. Henkel

Talk @ MPSoC'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>

## RISPP vs. ASIP

- ❑ At a fixed area **ASIPs typically cannot implement all necessary special instructions** in hardware. Alternative of software execution => may result in slow processing
- ❑ RISPP uses the a **rotational concept** instead to reconfigure the hardware for fore-coming Special Instructions
  - ❑ The rotational delay is reduced using dynamic forecasting techniques
- ❑ The rotational delay may be amortized => **improved performance** while still **keeping the cost of hardware below ASIP level**

J. Henkel

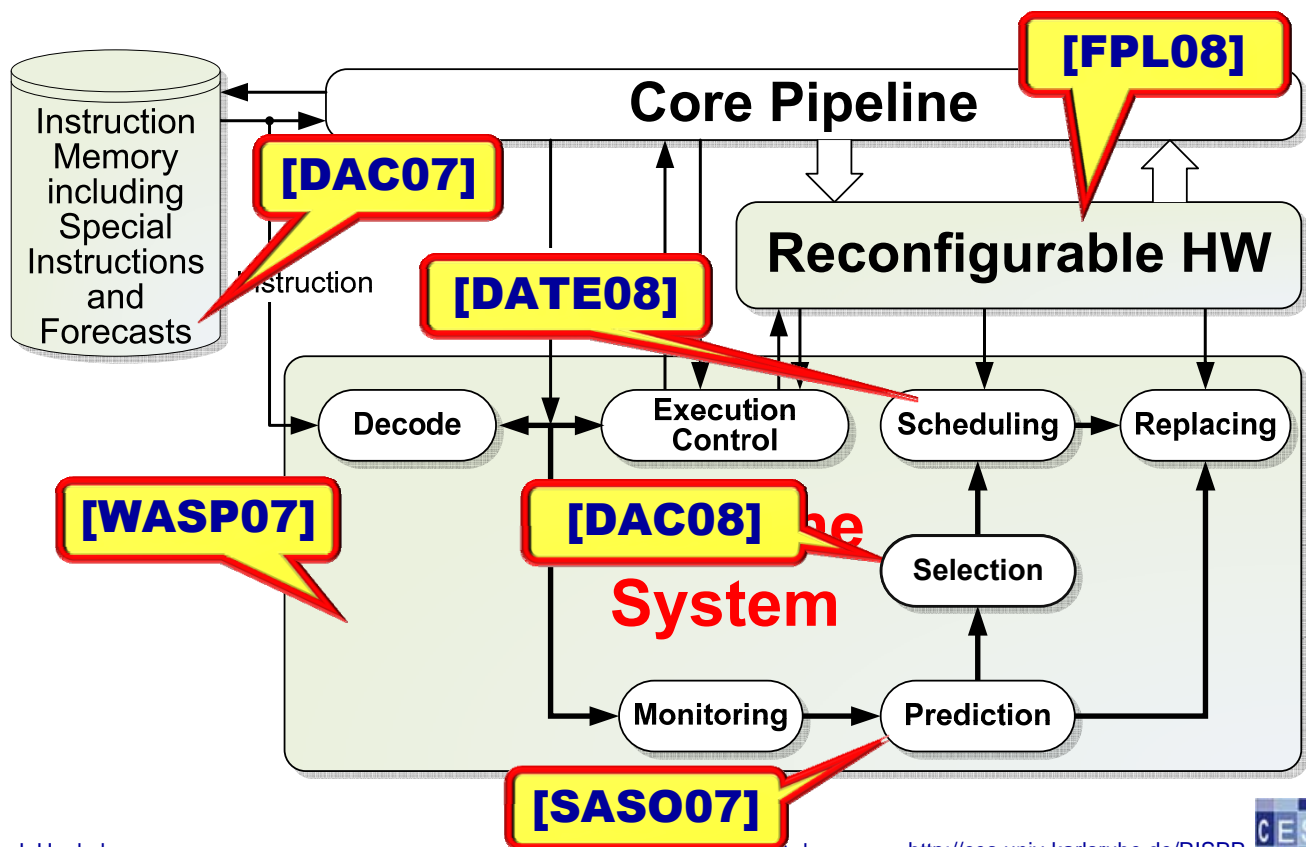
Talk @ MPSoC'08, June 23rd

<http://ces.univ-karlsruhe.de/RISPP>

# RISPP vs. ASIP (cont'd)

- ❑ The **atom/molecule model** reduces the possibilities for rotation with a more fine-grained and reusable vision of **Special Instructions**
- ❑ Typical **ASIP** is very specialized => **does not systematically re-use hardware components**
- ❑ **RISPP** requires **lesser hardware (in time-multiplex)** than ASIPs due to Atom/Molecule model and rotational concepts
- ❑ **However:** fabric to execute on (e.g. FPGA) is less efficient (performance/power)

## Conclusion



# Conclusion

## ❑ Current Project Status:

- ❑ Almost all major parts implemented in HW
- ❑ Running FPGA demonstrator
- ❑ In addition a complete simulation environment